K2's Data Analytics For Accountants And Auditors

Continuing Professional Education Series

K2 Enterprises
www.k2e.com

K2E Canada Inc.
www.k2e.ca

K2 Enterprises
1905 W. Thomas St., Suite D #363
Hammond, LA 70401-2901
985.542.9390

K2E Canada Inc.
174 Brewer Court
Burlington, Ontario, L7L4G4
905-746-1581

You may email requests to reproduce or transmit to info@k2e.com or caroline@k2e.ca. You may also obtain additional information on the web at www.k2e.com or www.k2e.ca.

CONTRIBUTORS: The shareholders, associates, and staff of K2 Enterprises and K2 Canada Inc. contributed to producing these materials.

Table of Contents

Introduction

Three Rules of this Seminar

1. **There Are No Dumb Questions.** Please ask questions if you need further explanation. We welcome your questions and will do our best to provide accurate and meaningful answers.

2. **Obtain the Information for Which You Came.** If you have a specific interest in a topic, please let us know. If your topic is not part of the course materials, we will do our best to accommodate your request. Your instructor is anxious to meet your needs.

3. **Share and Share-Alike.** Your instructor is here to share relevant information about the course. We hope that you will be willing to do the same. If you have helpful tips or experiences, please share them with the other participants and us.

Course Development

The shareholders, associates, and staff of K2 Enterprises and K2E Canada Inc. developed this course and the related course materials. We aim to deliver high-quality educational practices that help you use and understand the tools and concepts discussed in this session more effectively.

We designed our course presentation style for busy professionals. Through numerous short case studies and feature reviews, we deliver concise, easy-to-understand, easy-to-absorb information.

To maximize your time and the value of attending this course, we have carefully filtered out trivial and obscure features. Our goal is not to cover every keystroke and menu option, but to provide a fast-paced course that focuses on state-of-the-art information technology and how it can benefit you and your staff.

About K2 Enterprises

For over three decades, our team members have delivered more than 45,000 technology-focused seminars, webinars, and conferences worldwide. To stay on top of technology, our team members read numerous technology trade magazines, attend conferences, and speak with technology companies each year to stay current with the latest information for our courses. Also, we experience everything we teach and recommend to our audiences. Because our shareholders and associates are based in different states and provinces, using technology efficiently and appropriately is essential to our business's success and survival.

Our Mission Statement

K2 aims to develop and deliver the highest-quality technology seminars and conferences for business professionals. We work cooperatively with professional organizations such as state CPA societies, associations of Chartered Professional Accountants, and technology vendors. We make every effort to maintain high integrity, family values, and friendship among all involved.

Our Instructional Team

For a complete listing of instructors and bios, please visit:

K2 Enterprises

Instructors - K2 Enterprises

K2E Canada Inc.

Instructors - K2E Canada Inc

The K2 Seminar Curriculum

Full-day Seminars

- K2’s Advanced Excel
- K2’s Business Intelligence, Featuring Microsoft’s Power BI Tools
- K2’s Case Studies In Fraud And Technology Controls
- K2’s Excel Best Practices
- K2’s Excel Essentials For Staff Accountants
- K2’s Excel PivotTables For Accountants
- K2’s Excel Tips, Tricks, And Techniques For Accountants
- K2’s Next Generation Excel Reporting
- K2’s QuickBooks For Accountants
- K2’s Small Business Internal Controls, Security, And Fraud Prevention And Detection
- K2’s Technology For CPAs – Don’t Get Left Behind

Half-day Seminars

- K2’s Artificial Intelligence For Accounting And Financial Professionals
- K2’s Best Word, Outlook, And PowerPoint Features
- K2's Better Productivity Through Artificial Intelligence and Automation Tools
- K2’s Case Studies In Fraud And Technology Controls
- K2’s Data Analytics For Accountants And Auditors
- K2’s Ethics And Technology
- K2’s Excel Charting And Visualizations
- K2’s Implementing Internal Controls In QuickBooks Online Environments
- K2’s Improving Productivity With Microsoft 365Cloud Applications And Services
- K2’s Mastering Advanced Excel Functions
- K2’s Microsoft Teams
- K2’s Securing Your Data: Practical Tools For Protecting Information
- K2’s Technology Update
- K2’s Top PDF Features You Should Know

Besides the seminars listed above, K2 also produces numerous one and two-day technology conferences annually. You can find a complete listing of our course offerings, including dates, locations, and course descriptions, on our website at:

K2 Enterprises
Technology Conference - K2 Enterprises

K2E Canada Inc.
Technology Conference - K2E Canada Inc.

On-Site Training

Working cooperatively with state CPA organizations and associations of Chartered Professional Accountants, we are pleased to offer all our seminars and conference sessions as on-site training opportunities. On-site training provides companies of all sizes with numerous advantages, including the following.

- ***Customization***. K2 can customize the content of all on-site training events to meet your organization's specific needs. For example, we can combine selected parts of existing courses into a focused learning experience to meet your team's needs. Likewise, we can develop custom content from the ground up and deliver it to your team members, ensuring we meet your training objectives.

- ***Cost savings***. For organizations of all sizes, on-site training can provide significant cost savings compared to public training venues. In addition to potentially reduced registration fees, these cost savings materialize in the form of reduced travel costs and reducing the amount of time team members spend away from the office.

- ***Convenience***. Because you select your on-site training event's date, time, and location, on-site training is more convenient for you and your team than traditional training options. Many organizations schedule on-site training with staff meetings, retreats, and customer/client events to leverage the value of all participants' time.

Customization, ***cost savings***, and ***convenience*** are the three C's of "training to go." For more information on how you and your organization can experience these benefits or schedule an on-site training event, please visit:

K2 Enterprises
On-site Training - K2 Enterprises

caroline@k2e.com

K2E Canada Inc.
Seminars - K2E Canada Inc

caroline@k2e.ca

Web-Based Training

K2 offers web-based CPE and CPD to accounting and business professionals throughout the United States and Canada. K2's award-winning instructors deliver two-, four-, and eight-hour webinars on numerous technology-focused topics.

Additionally, K2 provides on-demand training. On-demand courses offer you the advantage of receiving top-quality training at a time and location that fits your busy schedule. You can start a class today and complete it in the future. For more information, please visit:

For more information on either platform, please visit:

K2 Enterprises
Webinars - K2 Enterprises

K2E Canada Inc.
Webinars - K2E Canada Inc

K2 Internet Sites

K2 maintains multiple websites containing information and links relevant to CPAs and other business professionals as a service to the profession. Further, the content of these sites can enrich your experience during and after our seminars. In addition, our pages contain links to some of the top accounting and content management software solutions and even to a few sites that are just fun places to visit. If you have any comments or questions regarding our web pages, please email *webmaster@k2e.com*.

K2 has two primary websites. You can view information regarding CPE and CPD opportunities, access a library of technology tips and articles, and link to K2-related websites.

www.k2e.com

www.k2e.ca

www.cpafirmtech.com

CPA Firm Technology contains technology-related information explicitly targeted at public accounting firms. We include product reviews, hardware and operating system discussions, affinity programs for CPA firms, and Cloud computing resources.

www.accountingsoftwareworld.com

Accounting Software World provides information related to accounting and financial management solutions. Here, you can find software and service reviews, product comparisons, white papers, and other material of interest to those seeking information about accounting software and service solutions.

www.totallypaperless.com

For information about document management and paperless processes, please visit **Totally Paperless**. In addition, you will find reviews on hardware and software used by all types of businesses in document imaging and management processes.

What Is Data Analytics?

Chapter 1

Data analytics is one of today's fastest-growing disciplines. Spurred by the emergence of "big data," business professionals across disciplines – especially in accounting, auditing, and finance – are turning to data analytics to answer many questions about business performance. In this chapter, you will learn about the fundamentals of data analytics – including the four types of data analytics – and how you and your organization can benefit from data analytics.

Learning Objectives

Upon completing this chapter, you should be able to:

- Differentiate data analytics from business intelligence;
- List the four types of data analytics; and
- Identify how each of the four types of data analytics can be useful.

Differentiating Data Analytics From Business Intelligence

Many professionals use *data analytics* and business intelligence (BI) interchangeably. In fact, these are two different, albeit complementary, disciplines.

What Is Business Intelligence?

Countless definitions for BI exist, and for purposes of this course, we borrow from many of these definitions to create the following.

> *Business Intelligence is a set of procedures, tools, and technologies working in concert to transform raw data – potentially vast quantities of raw data – into meaningful information. Managers can use this information to help make decisions and prescribe courses of action that optimize organizational performance.*

Without question, "big data" relates to BI. Over time, organizations accumulate an ever-increasing cache of data on sales, purchases, human resources activities, customer satisfaction scores, market demographics, and related data. Collecting data for the mere sake of having access to the information provides no competitive advantages. However, organizations that leverage this data in their BI initiatives often find they can improve efficiency and profitability as a direct outcome of these efforts.

BI efforts can be helpful in virtually all departments of all types of organizations, including governmental units and not-for-profits. However, until recently, many believed BI was available only to large, enterprise-class organizations. The reasoning behind this myth is the level of resources required to implement necessary procedures, tools, and technologies successfully. However, the advent of Cloud-based BI platforms and desktop-based BI applications such as **Power BI**, **Qlik**, **Tableau**, and others means that organizations of all sizes can benefit from BI without necessarily investing significant resources. Further, because the cost of entry can be meager, your BI efforts' return on investment (ROI) can prove extraordinary.

In particular, the ROI on BI can be exceptionally high in four specific areas.

1. Addressing the need for more timely and improved decisions
2. Automating repetitive reports
3. Identifying trends and opportunities
4. Motivating team members and encouraging teamwork

What Is Data Analytics?

Many agree that data analytics is a subset of business intelligence, as pictured below in **Figure 1**.

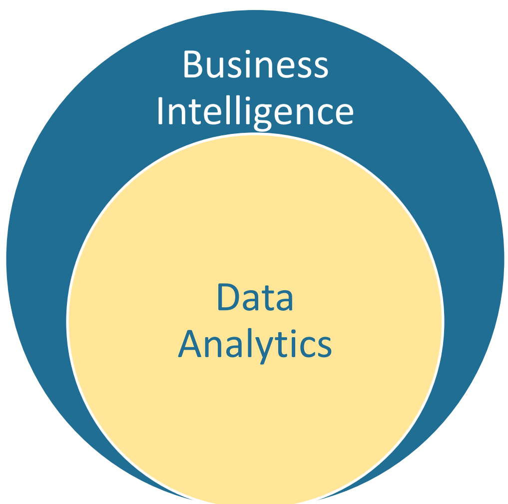


Figure 1 - The Relationship Between Business Intelligence And Data Analytics

More specifically, business professionals use the information gathered through their data analytics efforts to inform their decision-making in business intelligence. In other words, data analytics is a tool used by BI. For example, business professionals use data analytics tools to build analyses, forecasts, and projections. Additionally, they use BI tools to present this information in reports, dashboards, and other presentation formats.

Data analytics refers to the tools, procedures, and techniques business professionals use to transform raw data into meaningful information. To achieve this objective, data analytics attempts to identify patterns and relationships in data and answer the following questions:

- What happened?
- Why did we attain a specific set of results?
- What will happen in the future?
- How can we achieve better performance?

For accountants and auditors, the emergence of data analytics has profound implications, as many are now required to assume a data analyst role, explicitly or implicitly. Given the impact of technology on the profession, many tasks formerly performed by these professionals are now automated or will soon be. Further, the amount of data that operations systems and processes gather has exploded over recent years and will continue to grow for the foreseeable future. The explosive growth of data, coupled with data management and mining tools, has led many organizations to use data analytics to turn this information into a valuable asset.

Four Types of Data Analytics

As organizations mine and manage these increasing volumes of data, they determine how to use the data to add value to the organization. There is widespread agreement that there are four types of data analytics, each serving a different purpose. These four classifications are *descriptive*, *predictive*, *prescriptive*, and *diagnostic analytics*.

Descriptive Analytics

As its name implies, descriptive analytics is useful for determining what happened in the past. For example, most reports generated by accounting, ERP, and similar systems provide descriptive data. For instance, we can create reports that identify sales by salesperson, thus describing each salesperson's performance. However, these reports do not explain *why* one salesperson struggled and another had great success. To answer these questions, we would need to focus on *diagnostic analytics*.

Diagnostic Analytics

As inferred above, diagnostic analytics help answer the question "why." For example, with diagnostic analytics, we might be able to identify how much of an increase in sales was attributable to each of the following three factors:

- An increase in advertising expenses,
- Decrease in interest rates, and
- Paying special incentives to salespersons to generate higher sales.

Predictive Analytics

Closely related to diagnostic analytics is *predictive analytics*. As the name implies, predictive analytics seeks to identify patterns in data to produce reasonably accurate forecasts of future activity. To illustrate, an auditor or business analyst might use *regression analysis* to forecast future factory payroll expenditures based on estimated amounts of inventory production.

Accountants and auditors can use predictive analytics to establish a “baseline” for measuring achieved results.

Prescriptive Analytics

Prescriptive analytics is the fourth type of data analytics and the group’s most complex. Prescriptive analytics uses models to predict outcomes and then suggests the best course of action to optimize them. Prescriptive analytics answers the question, “What should happen?” To illustrate, a data analyst could create an analytical model forecasting revenue of $20 million in the upcoming fiscal quarter. Then, the analyst might use a model based on prescriptive analytics to specify the best course of action to generate $20 million in revenue at the least expense. In this example, prescriptive analytics surface the steps required to optimize results.

A Simple Example Of Applied Data Analytics

An auditor might be interested in analyzing credit memos to ensure that all issued credit memos are valid and approved by appropriate levels of management. As part of the audit, the auditor might generate a report showing the total amount of credit memos compared to sales over the preceding five years; this is an example of *descriptive analytics* because it answers the question, “What happened?”

The auditor might then use the data generated by descriptive analytics to run a regression analysis to forecast future credit memos based on seasonality and total sales volume. Of course, this action represents *predictive analytics*. The auditor could then evaluate the forecasted data against future results to determine if the credit memos issued were consistent with expectations. If so, no additional audit work may be necessary. On the other hand, if the results did not align with expectations, more actions could be required.

If more audit work is necessary, the auditor might utilize various *diagnostic analytics* approaches to answer the question, “Why did we have unexpectedly high write-offs?” In this case, perhaps the auditor begins examining macroeconomic trends. For instance, an increase in unemployment may be driving higher write-offs. Possibly significant inflation is causing customers to become cash-strapped and unable to pay their bills. Perhaps changes in manufacturing processes created quality issues with the products sold to consumers, and they are resistant to paying for merchandise that did not meet their expectations. Of course, the purpose of the diagnostic analytics procedures is to answer the question, “Why did we write off more than was forecast?”

A Very Big Caveat!

Sometimes, it can become too easy to rely on data analytics to answer questions about what happened, why it happened, what will happen, and how we should optimize. Therefore, exercise caution when using data analytics, particularly when considering correlation and causation. Just because two items are tightly correlated does not mean a cause-and-effect relationship exists. **Correlation does not equate to causation**!

Consider the curious relationship between ice cream sales and thunderstorm activity. In North America, ice cream sales surge in the summer months. As the weather gets warmer, many like to cool down with a refreshing scoop or two of their favorite ice cream flavor, a milkshake, a banana split, or another ice cream-based treat. Also, during the summer, thunderstorms increase throughout most of North America. Because ice cream sales and thunderstorm activity increase simultaneously, can we say that one drives the other? Does the amount of ice cream sold depend on the thunderstorm activity? Conversely, is the number of thunderstorms somehow dependent upon the amount of ice cream sold? Of course, the answer to these questions is unequivocally and emphatically "**NO!**"

Of course, the outdoor temperature is the "driver" of increased ice cream sales and thunderstorm activity. Warmer weather is likely responsible for increased ice cream sales and thunderstorm activity. This example illustrates that just because two variables move in tandem does not mean a cause-and-effect relationship exists. As accountants, auditors, and business professionals, we must carefully examine all related variables to differentiate between correlation and causation.

Data Analytics Tools And Techniques In Excel

Chapter 2

Excel is the "tool of choice" for most accountants, auditors, and other business professionals performing many reporting and data analysis activities. We tend to be quite comfortable working in the friendly confines of Excel and know how to make the most of its functions and features. Therefore, it should not surprise anyone that Excel is often the application of first resort when data analytics is needed. This chapter examines the vast array of Excel features and functions that support data analytics applications.

Learning Objectives

Upon completing this chapter, you should be able to:

- List at least five descriptive statistics generated using formulas and Excel's Analysis ToolPak;
- Identify the process of using Excel's Analyze Data feature to recognize relationships and anomalies in data sets;
- Distinguish between Excel's regression analysis tools and functions and how these can be useful in data analytics exercises;
- Identify the steps required to analyze the quality of data using Power Query; and
- List examples of Excel add-ons that are useful when performing data analytics tasks.

The Role Of Statistics In Data Analytics

Statistical measures play an essential role in data analytics. Analysts and auditors commonly use statistical measures to identify outliers, assess risk, and forecast expected results. For example, data analysts and auditors can use regression analysis as a predictive analytics tool to predict future outcomes from historical data. Likewise, they can use standard deviations to identify "outliers" in data sets that may require follow-up to confirm the validity of the data point.

Among the many potential uses of Excel is generating statistical calculations. Further, Excel generally provides multiple options for calculating these statistical measures – formulas and the ***Analysis ToolPak***. Let us begin by reviewing the functionality in the Analysis ToolPak.

Analysis ToolPak

The Analysis ToolPak is an Excel add-in that, by default, is not activated. However, if you see **Data Analysis** in the Analysis group on the Ribbon's **Data** tab, the Data Analysis ToolPak is active and available. If the option is not on the Ribbon, activate it by accessing Excel's **Options**, choosing **Add-Ins**, and clicking **Go** next to **Manage Excel Add-ins** to open the **Add-Ins** dialog box. Then, as shown in **Figure 2**, check the boxes next to **Analysis ToolPak** and **Analysis ToolPak – VBA**.

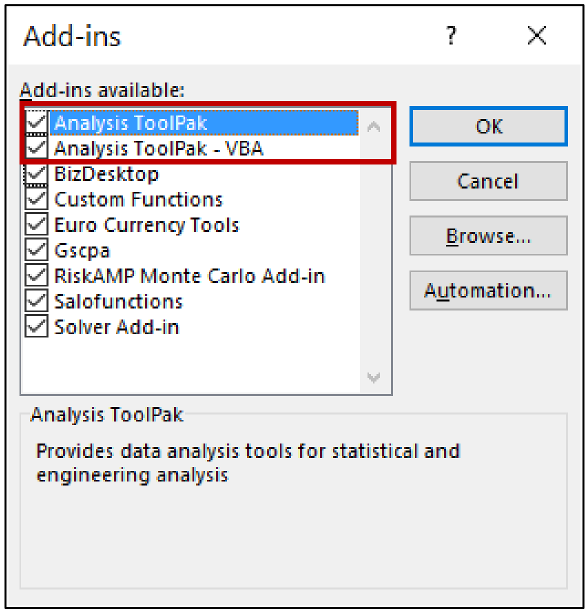


Figure 2 - Activating the Data Analysis ToolPak

The Analysis ToolPak contains a considerable number of statistical analysis procedures. Among them are ANOVA, simple and multiple linear regression, covariance, descriptive statistics, random number generation, *t*- and *z*-tests, sample selection, and more. Auditors and data analysts can use the calculations generated by the Analysis ToolPak to identify numbers that fall outside expected values.

To calculate the statistics displayed in **Figure 3**, select **Data Analysis** from the Ribbon's **Data tab**. Then, select **Descriptive Statistics** in the **Data Analysis** dialog box, specify the input and output ranges, and click **OK** to calculate the statistics from samples or large data sets.

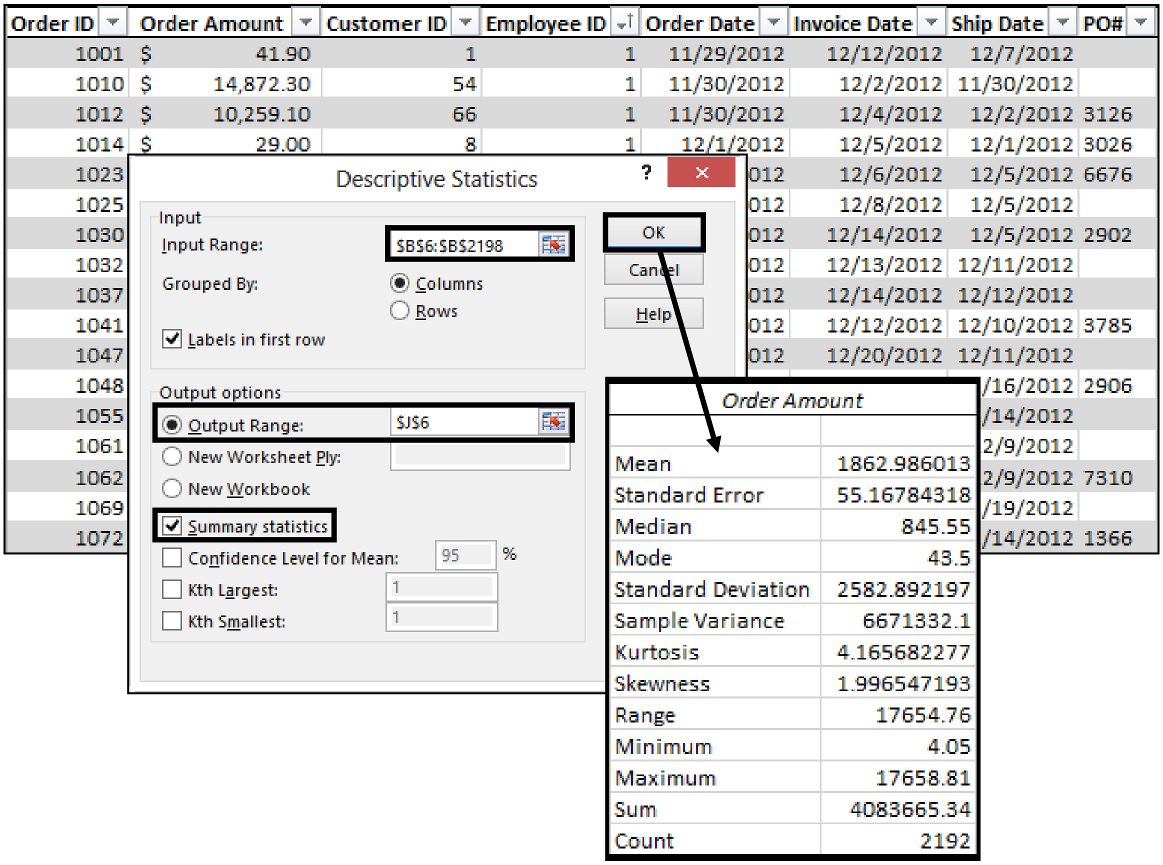


Figure 3 - Calculating Descriptive Statistics with the Analysis ToolPak

Notably, you can use functions to generate all the calculations displayed above; however, by using the Analysis ToolPak for the task, you avoid the task of creating formulas in Excel containing the appropriate statistical function.

Histograms

Data analysts and auditors can use *histograms* to group data into "buckets," or "bins." For example, this form of data analytics can help identify risk concentrations, enabling audit and examination efforts to be focused and efficient.

You can use the **Histogram** tool in the Data Analysis ToolPak to create a histogram. First, open the **Data Analysis ToolPak**, click **Histogram**, and then **OK** to construct a histogram. Next, enter the **Input Range** and the **Bin Range** in the Histogram dialog box and select the desired outputs, as shown in **Figure 4**.

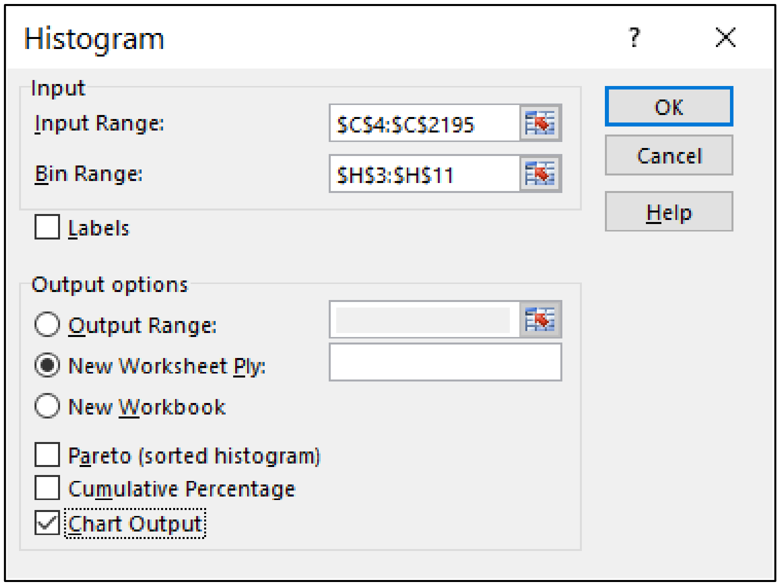


Figure 4 - Creating a Histogram for Benford's Law Analysis

Excel creates a histogram summarizing the data when you click **OK**, as shown in **Figure 5**. From this output, you can begin to draw conclusions about your data.

Bin	Frequency	Bin	Frequency
1	472	1	472
2	293	2	293
3	238	3	238
4	230	4	230
5	221	9	223
6	180	5	221
7	127	8	208
8	208	6	180
9	223	7	127
More	0	More	0

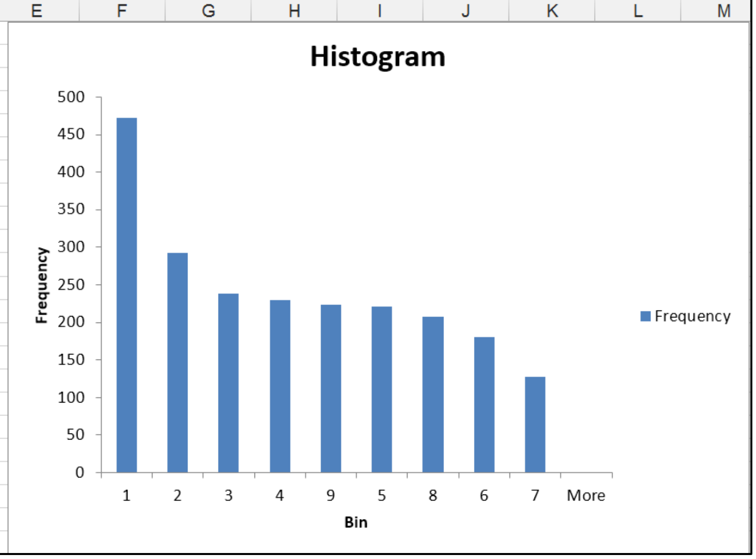


Figure 5 - Histogram Created by Data Analysis ToolPak

Benford's Law Analysis

In 1881, Astronomer Simon Newcomb noticed that the pages of his logarithm books with smaller numbers were much more worn than those with larger numbers. Newcomb postulated that the reason for this was that numbers beginning with smaller digits – such as "1" and "2" – occurred more frequently than numbers beginning with higher digits – such as "8" and "9." Therefore, he and others using the logarithm books would have used more pages with numbers beginning with smaller digits than those starting with larger digits. Newcomb published his findings in *The American Journal of Mathematics* in 1881. Subsequently, in 1938, Frank Benford made essentially the same observation and reached the same conclusions as Newcomb while working as a physicist at GE Research Laboratories. Benford then tested and validated his theory, which became known as Benford's Law despite Newcomb's work a half-century earlier, against twenty different data lists with over 20,000 observations.

Benford's Law is a powerful tool for detecting fraud and other data anomalies because it allows examiners to test data against recognized norms. In doing so, you can identify and investigate significant deviations from these expected values to determine if there is evidence of fraud. In addition, Benford's Law is useful for identifying fraudulent transactions and detecting those that have been manipulated to avoid authorization and approval.

First Digit Probabilities

The following formula serves as the foundation for Benford's Law.

$$P\{d\} = \frac{\ln\left[1 + \frac{1}{d}\right]}{\ln(10)}$$

Based on this formula, one can project that the probability distribution of the first digits would approximate the following for a given population of numbers.

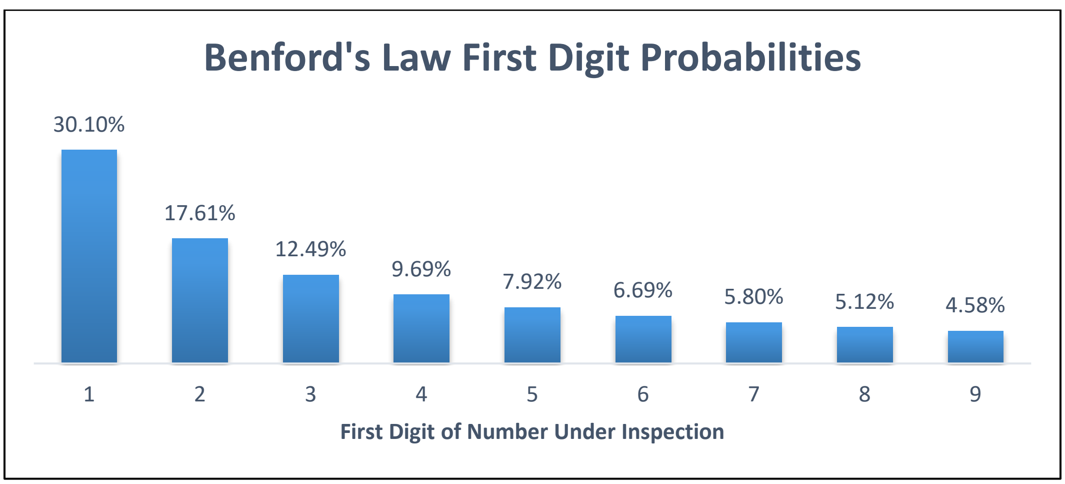


Figure 6 - Benford's Law: First Digit Probabilities

As shown in **Figure 6**, numbers that begin with "1" occur approximately seven times more frequently than numbers that start with 9. Understanding this pattern, suppose an auditor discovered that the digit "9" began, for example, 35% of the values in a population of transactions. Knowing that Benford's Law predicts that the number should be approximately

4.6%, the auditor should likely perform some testing to identify the cause of such a significant deviation. Perhaps the numbers are legitimate and there are no errors or irregularities. On the other hand, maybe someone recorded fraudulent transactions—for instance, bogus accounts payable invoices. We cannot use Benford's Law to conclude with certainty that there is a problem; at best, it indicates that more work may be necessary.

Second Digit Probabilities

Just as Benford's Law can predict first-digit probabilities, it also predicts second-digit probabilities. **Figure 38** displays the second-digit probabilities predicted by Benford's Law without becoming bogged down in formulae.

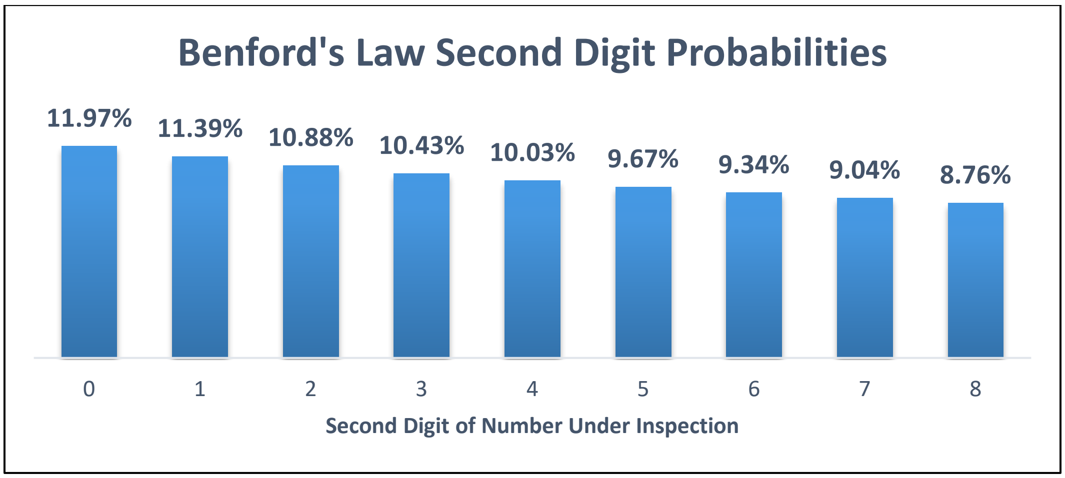


Figure 7 - Benford's Law: Second Digit Probabilities

While the variation in the percentages associated with second-digit probabilities is not as significant as with first-digit probabilities, there remains a distinct pattern of lower digits occurring more frequently than higher ones. Like the first digit probabilities shown in Figure 6, you can use this pattern to identify potentially fraudulent or erroneous transactions.

First Two-Digit Probabilities

Combining probabilities for the first digit and second digit Benford's Law tests, you can construct a set of probabilities for the first two digits of a number. **Figure 8** displays these probabilities.

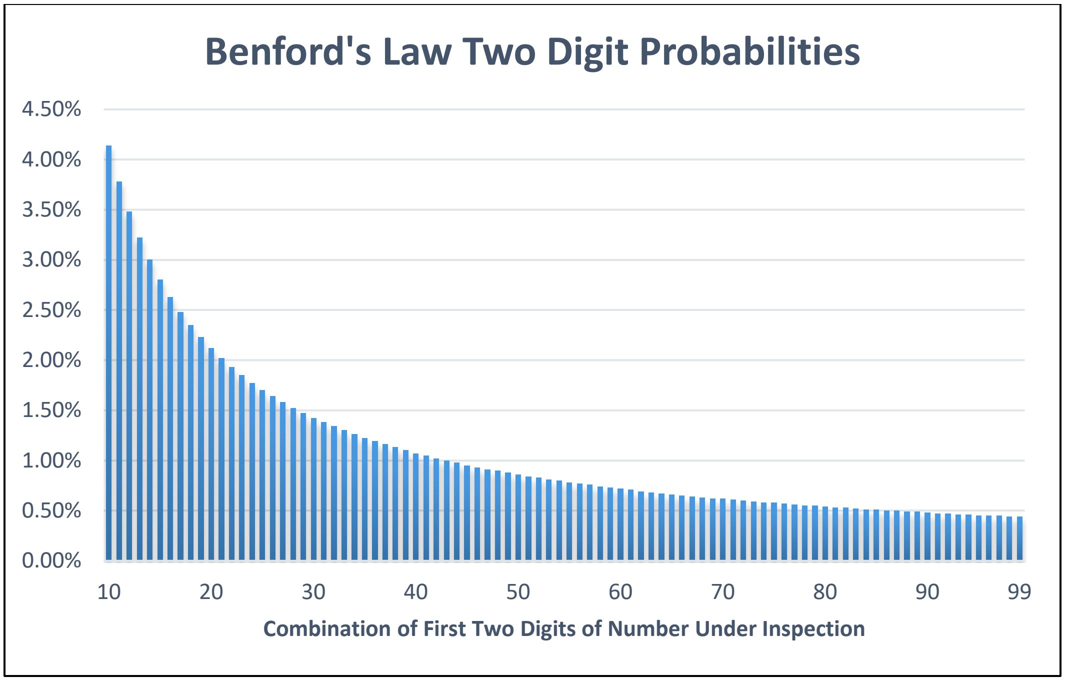


Figure 8 - Benford's Law: First Two Digits Test

Using this information, an auditor can see that numbers—invoice amounts, for example—beginning with "10" and "11" should occur much more frequently in a sample or population than numbers beginning with "98" or "99."

Utilizing Benford's Law

Benford's Law provides accountants and auditors with a powerful tool for detecting potential fraud. To see how this tool can help in a fraud-detection scenario, consider the case of *The State of Arizona vs. Wayne James Nelson*. As reported by Mark Nigrini in the May 1999 issue of the *Journal of Accountancy*, Nelson, a manager in the Arizona State Treasurer's office, was found guilty in 1993 of defrauding the state of approximately $2 million by issuing bogus checks to fictitious vendors. Nelson, of course, diverted the checks to his personal accounts.

As shown in **Table 1**, the amounts of the checks Nelson caused to be issued did not conform to the expected distribution under Benford's Law. Of the twenty-three disbursements issued, only one began with the digit "1;" likewise, only one check began with the digit "2." According to Benford's Law, 30.10% of the payments (approximately seven) should have led with "1." Also, Benford's Law indicates that 17.61% of the checks (about four) should have started with "2."

Check Date	Amount
October 9, 1992	$1,927.48
October 9, 1992	27,902.31
October 14, 1992	86,241.90
October 14, 1992	72,117.46
October 14, 1992	81,321.75
October 14, 1992	97,473.96
October 19, 1992	93,249.11
October 19, 1992	89,658.17
October 19, 1992	87,776.89
October 19, 1992	92,106.83
October 19, 1992	79,949.16
October 19, 1992	87,602.93
October 19, 1992	96,879.27
October 19, 1992	91,806.47
October 19, 1992	84,991.67
October 19, 1992	90,831.83
October 19, 1992	93,766.67
October 19, 1992	88,338.72
October 19, 1992	94,639.49
October 19, 1992	83,709.28
October 19, 1992	96,412.21
October 19, 1992	88,432.86
October 19, 1992	71,552.16
Total	**$1,878,688.58**

Table 1 - Fraudulent Checks in the Case of the State of Arizona vs. Wayne James Nelson

The fact that 21 of the 23 checks began with the digits “7,” “8,” or “9” does not align with Benford’s Law. According to Benford’s Law, only approximately four checks should have led with these digits.

Benford's Law provides an effective means of highlighting transactions that do not conform to expected norms. This tool has obvious value for accountants and auditors who face the suspicion of bogus or illegal transactions. However, be aware that if the perpetrators are aware of Benford's Law, it is likely they will arrange fraudulent transactions to fit the expected distributions defined by Benford's Law. In those cases where there is no suspicion of fraudulent transactions, applying Benford's Law can still be helpful as a quick check to identify any suspicious pattern in recorded transactions.

Applying Benford's Law in Excel

First Digit Test in Excel

With smaller data sets like the one presented in Table 1, applying Benford's Law requires only a quick visual observation and inspection of the data. However, when applying Benford's Law to hundreds or thousands of transactions, a more efficient analysis is required, and Excel provides this. Using familiar Excel commands such as **LEFT** and **COUNTIFS**, you can construct a Benford's Law test in Excel for any single-digit or multiple-digit probabilities.

In the example below, an Excel worksheet contains credit memo data issued by a company. An auditor concerned about fictitious credit memos wants to run the first digit Benford's Law test on the population. Before a modification to include the Benford's Law test, the worksheet containing the credit memo data appears as shown in **Figure 9**.

	A	B
1	Check Number	Invoice Amount
2	1001	41.90
3	1002	5,060.28
4	1003	986.87
2188	3188	99.90
2189	3189	31.00
2190	3190	53.90
2191	3191	91.70
2192	3192	959.70
2193	3193	23.80

Figure 9 - Worksheet Containing Credit Memo Data to Analyze with Benford's Law

The first step in creating a first-digit Benford's Law test is extracting the first digit from the *Invoice Amount* field. You can accomplish this by creating the following formula in cell C2 and copying the formula down through the data set.

=LEFT(B2,1)

As entered, the formula instructs Excel to extract the first character of the entry in cell B2. Thus, upon entering and copying the formula, the worksheet resembles that shown in **Figure 10**.

	A	B	C
1	Check Number	Invoice Amount	First Digit
2	1001	41.90	4
3	1002	5,060.28	5
4	1003	986.87	9
2188	3188	99.90	9
2189	3189	31.00	3
2190	3190	53.90	5
2191	3191	91.70	9
2192	3192	959.70	9
2193	3193	23.80	2

Raw I ...

Figure 10 - Credit Memo Worksheet with First Digit Extracted

Next, it is necessary to count each digit's number of occurrences in column B and compare that to the expected norms defined by Benford's Law. To perform this step, add a second worksheet to the workbook; a sample of that worksheet appears in **Figure 11**.

	A	B	C	D	E	F
1	Digit	Count	Expected Proportion	Actual Proportion		
2	1		30.10%			
3	2		17.61%			
4	3		12.49%			
5	4		9.69%			
6	5		7.92%			
7	6		6.69%			
8	7		5.80%			
9	8		5.12%			
10	9		4.58%			
11	Total		100.00%			

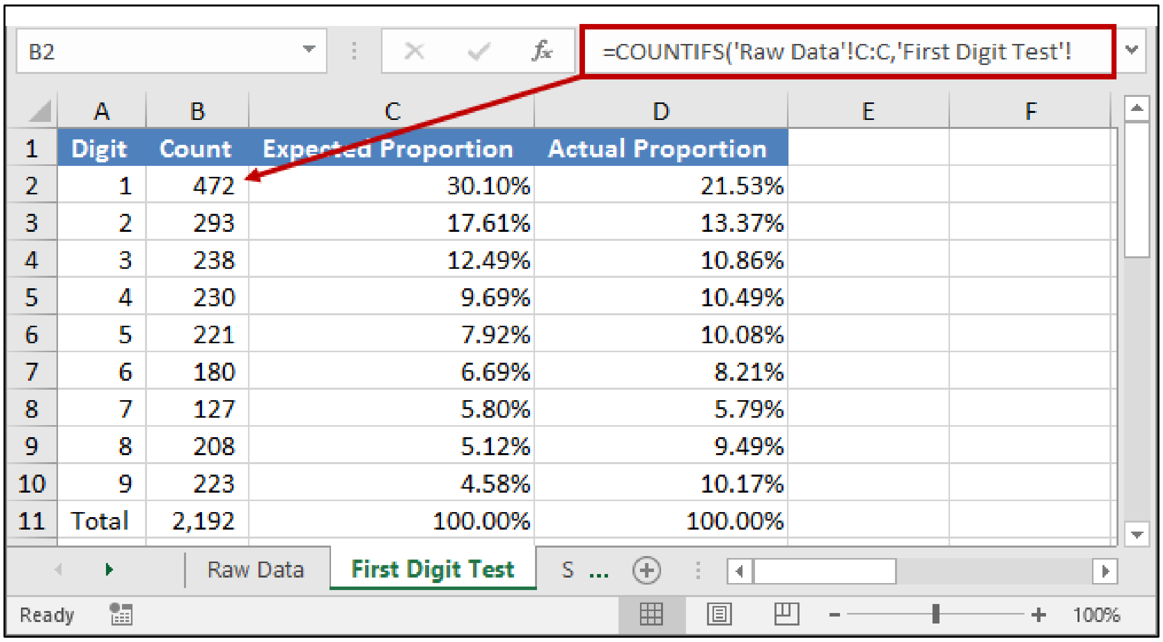
Raw Data | First Digit Test | S ...

Figure 11 - Data Analysis Worksheet for Benford's Law Test

In cell B2 of the First Digit Test worksheet shown in **Figure 12**, enter the following formula to count the number of instances of each digit in the Raw Data worksheet.

=COUNTIFS('Raw Data'!C:C,' First Digit Test'!A2)

Using Excel's **COUNTIFS** function, the worksheet counts and summarizes all occurrences of the digit in cell A2 of the First Digit Test worksheet found in column C of the Raw Data worksheet. Copy this formula down through cell B11 of the First Digit Test worksheet. Adding **SUM** formulas to cells B11, C11, and D11 and percentage calculations to column D transform the First Digit Test worksheet to that shown in **Figure 12.**



Digit	Count	Expected Proportion	Actual Proportion
1	472	30.10%	21.53%
2	293	17.61%	13.37%
3	238	12.49%	10.86%
4	230	9.69%	10.49%
5	221	7.92%	10.08%
6	180	6.69%	8.21%
7	127	5.80%	5.79%
8	208	5.12%	9.49%
9	223	4.58%	10.17%
Total	2,192	100.00%	100.00%

Figure 12 - Data Analysis Worksheet with First Digits Counted

In the first-digit Benford's Law example presented above, far too many transactions begin with the digits "8" and "9". Additionally, there seem to be too few transactions starting with "1" and 2." **Figure 13** illustrates this anomaly, comparing expected to actual percentages derived from the first-digit Benford's Law test. Whether these findings represent errors or fraud is not determined solely by applying Benford's Law; however, Benford's Law helps identify areas that may require additional testing and evaluation.

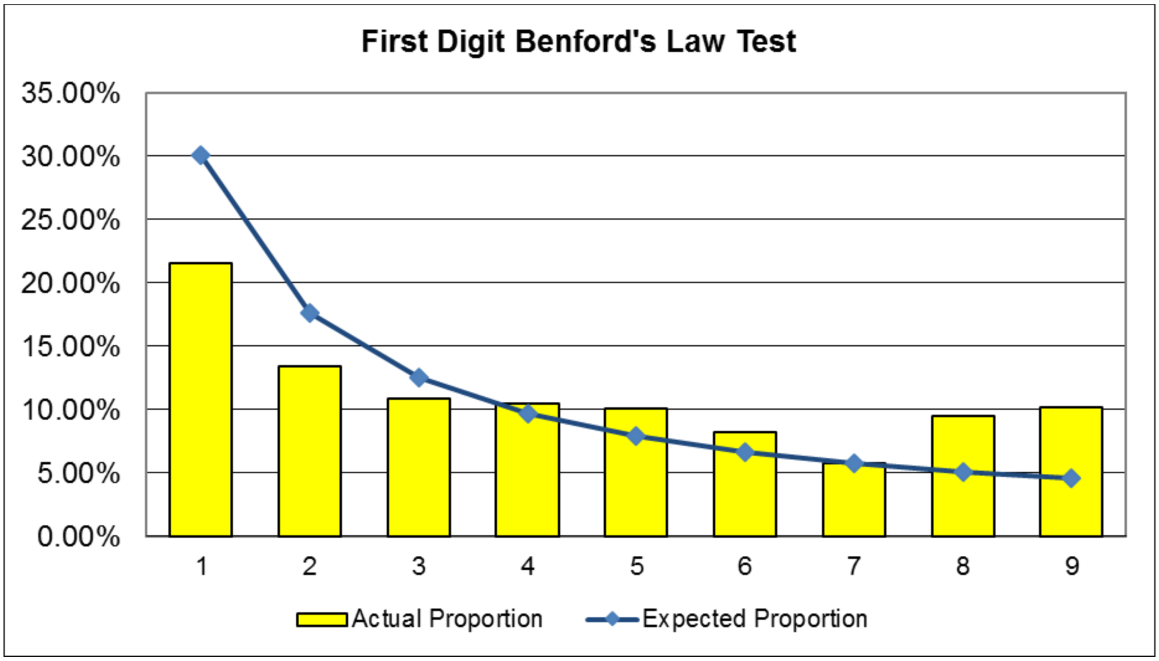


Figure 13 - Expected Versus Actual Percentages Graph

Second Digit Test in Excel

In addition to creating a first-digit Benford's Law test in Excel, you can create a second-digit test relatively easily. The steps for generating a second-digit test closely resemble those for the first-digit test; however, we will use Excel's MID function instead of LEFT to identify the second digits.

As shown in **Figure 14**, you can use the following formula to identify the second digit in the *Invoice Amount* column.

=MID(B2,2,1)

This formula instructs Excel to return the contents of cell B2, beginning with the second character and continuing for one space. Excel ignores the dollar sign applied to the data in column B because it is only a format.

	A	B	D
1	Check Number	Invoice Amount	Second Digit
2	1001	41.90	1
3	1002	5,060.28	0
4	1003	986.87	8
2188	3188	99.90	9
2189	3189	31.00	1
2190	3190	53.90	3
2191	3191	91.70	1
2192	3192	959.70	5
2193	3193	23.80	3

Raw D ...

Figure 14 - Identifying the Second Digit in Excel Using the MID Function

Copying the formula across the dataset generates a list of all the second digits. These can be summarized and analyzed using the same approach as when conducting a first-digit test, with one notable exception: when performing a second-digit test, you must consider the numeral zero as a valid second digit. When conducting a first-digit test, you can ignore "zero" because it is an invalid first digit.

	A	B	C	D
1	Digit	Count	Expected Proportion	Actual Proportion
2	0	220	11.97%	10.04%
3	1	157	11.39%	7.16%
4	2	213	10.88%	9.72%
5	3	281	10.43%	12.82%
6	4	201	10.03%	9.17%
7	5	225	9.67%	10.26%
8	6	231	9.34%	10.54%
9	7	201	9.04%	9.17%
10	8	237	8.76%	10.81%
11	9	226	8.50%	10.31%
12	Total	2,192	100.00%	100.00%

Figure 15 - Analyzed Data in a Second Digit Benford's Law Test

Finally, when viewed graphically (see **Figure 16**), the second-digit Benford's test shows relatively significant deviations from expected norms with the digits 0, 1, 2, 3, 8, and 9. Again, this does not necessarily indicate erroneous or fraudulent transactions – only that additional audit work may be necessary.

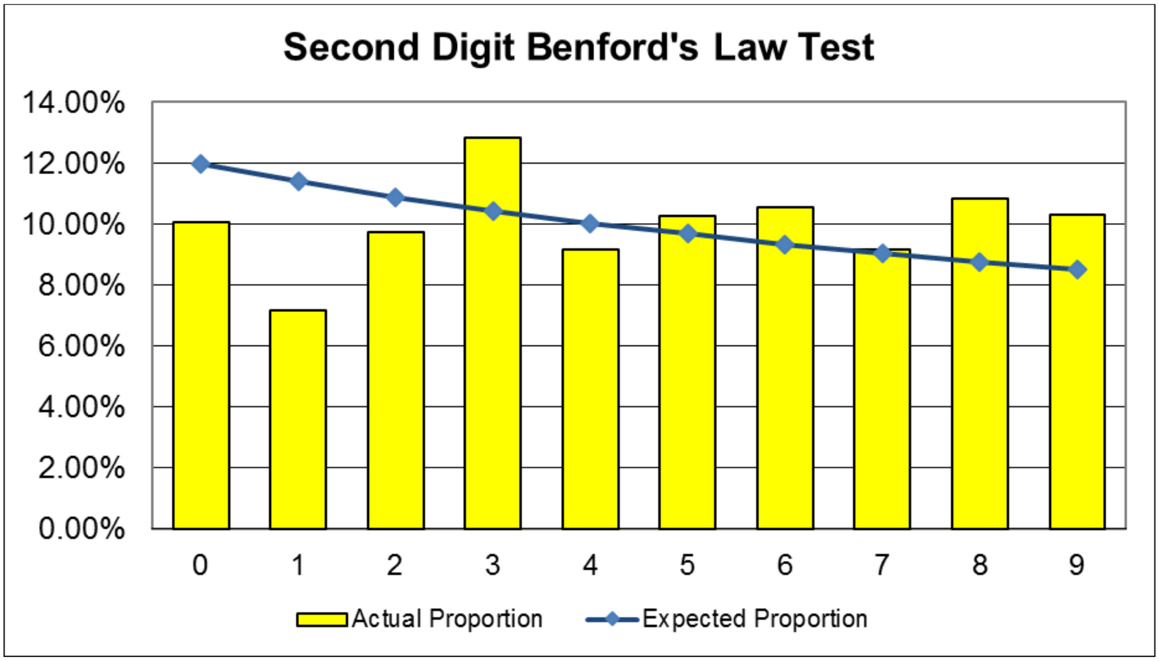


Figure 16 - Expected Versus Actual Percentages

First Two-Digit Test in Excel

Just as you can use Excel for a single-digit test – either first or second, as shown previously – you can also test the conformity of the first two digits to Benford's Law. The procedure will be like that of the first-digit test. However, we will modify the formula containing the LEFT function slightly. Specifically, changing the LEFT formula to match the one shown below will return the first two characters of the cell instead of just the first.

=LEFT(B2,2)

Once we identify the two left characters of the cells under consideration, as shown in **Figure 17**, we can analyze the data to determine its conformity with the results predicted by Benford's Law.

	A	B	E
1	Check Number	Invoice Amount	First Two Digits
2	1001	41.90	41
3	1002	5,060.28	50
4	1003	986.87	98
2188	3188	99.90	99
2189	3189	31.00	31
2190	3190	53.90	53
2191	3191	91.70	91
2192	3192	959.70	95
2193	3193	23.80	23

Raw D: ...

Figure 17 - Identifying Characters for First Two-Digit Benford's Law Test

Using the same analysis techniques employed with the first and second-digit tests, you can easily create a chart like that in **Figure 18**, providing a pictorial representation of the data values that do not fit the expected distribution of numbers according to Benford's Law.

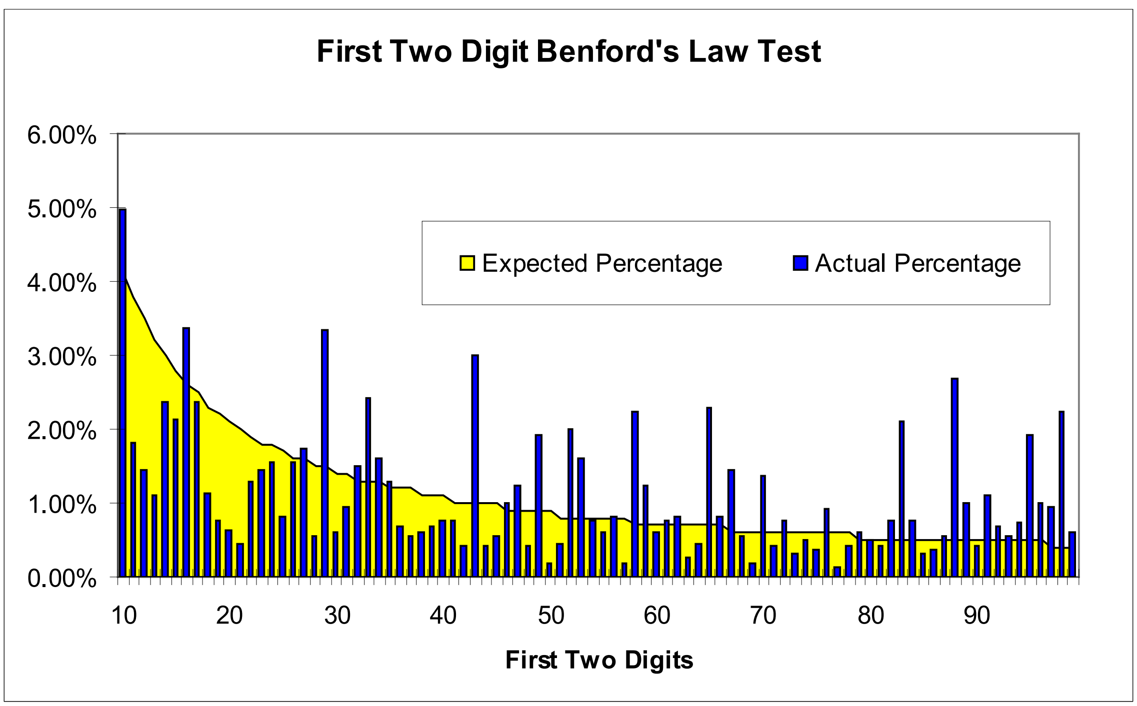


Figure 18 - Results of First Two Digits Benford's Law Test

Analyze Data (formerly known as Ideas)

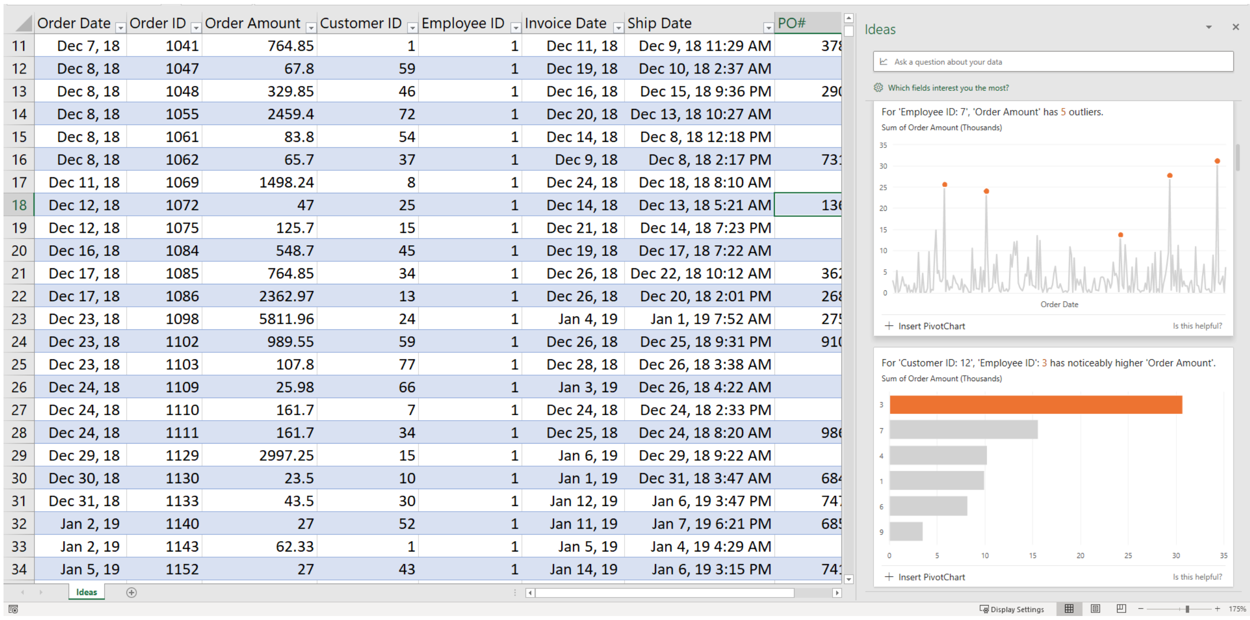
Excel's **Analyze Data** (formerly known as **Ideas**) feature offers significant opportunities to automatically summarize and analyze data. With Analyze Data, you no longer need to write formulas to identify outliers and get answers to the questions you have about your data – the tool does the work for you!

A Few Details about Analyze Data

Unfortunately, not all Excel users will have access to Analyze Data. Presently, Analyze Data is available only in versions of Excel provided through Office 365/Microsoft 365 subscriptions or Excel Online. Therefore, if you are using a perpetual license of Excel, you will not have access to this feature. Second, the tool does not work with a dataset containing more than 1.5 million cells. Finally, Analyze Data cannot analyze all types of data. For instance, if you have merged cells in your dataset, you must unmerge them before using the tool. Also, if your dataset contains dates stored as strings, Analyze Data will treat them as text strings. Therefore, before running Analyze Data, you should use Excel's **DATE** or **DATEVALUE** functions to convert the date strings into "true" Excel dates.

Summarizing Data With Excel's Ideas Feature

To quickly summarize your data, click anywhere in the dataset. Next, click **Analyze Data** on the **Home** tab of the Ribbon. Excel will then generate a series of PivotTables and PivotCharts that automatically summarize your data. You will see these summaries in a pane on the right side of your screen. If desired, you can insert one or more of these summarizations into your workbook. To do so, click **Insert PivotTable** or **Insert Chart** in the lower left corner of the summarization. **Figure 19** below illustrates the types of automated analyses created by Analyze Data.



	Order Date	Order ID	Order Amount	Customer ID	Employee ID	Invoice Date	Ship Date	PO#
11	Dec 7, 18	1041	764.85	1	1	Dec 11, 18	Dec 9, 18 11:29 AM	378
12	Dec 8, 18	1047	67.8	59	1	Dec 19, 18	Dec 10, 18 2:37 AM	
13	Dec 8, 18	1048	329.85	46	1	Dec 16, 18	Dec 15, 18 9:36 PM	290
14	Dec 8, 18	1055	2459.4	72	1	Dec 20, 18	Dec 13, 18 10:27 AM	
15	Dec 8, 18	1061	83.8	54	1	Dec 14, 18	Dec 8, 18 12:18 PM	
16	Dec 8, 18	1062	65.7	37	1	Dec 9, 18	Dec 8, 18 2:17 PM	731
17	Dec 11, 18	1069	1498.24	8	1	Dec 24, 18	Dec 18, 18 8:10 AM	
18	Dec 12, 18	1072	47	25	1	Dec 14, 18	Dec 13, 18 5:21 AM	136
19	Dec 12, 18	1075	125.7	15	1	Dec 21, 18	Dec 14, 18 7:23 PM	
20	Dec 16, 18	1084	548.7	45	1	Dec 19, 18	Dec 17, 18 7:22 AM	
21	Dec 17, 18	1085	764.85	34	1	Dec 26, 18	Dec 22, 18 10:12 AM	362
22	Dec 17, 18	1086	2362.97	13	1	Dec 26, 18	Dec 20, 18 2:01 PM	268
23	Dec 23, 18	1098	5811.96	24	1	Jan 4, 19	Jan 1, 19 7:52 AM	275
24	Dec 23, 18	1102	989.55	59	1	Dec 26, 18	Dec 25, 18 9:31 PM	910
25	Dec 23, 18	1103	107.8	77	1	Dec 28, 18	Dec 26, 18 3:38 AM	
26	Dec 24, 18	1109	25.98	66	1	Jan 3, 19	Dec 26, 18 4:22 AM	
27	Dec 24, 18	1110	161.7	7	1	Dec 24, 18	Dec 24, 18 2:33 PM	
28	Dec 24, 18	1111	161.7	34	1	Dec 25, 18	Dec 24, 18 8:20 AM	986
29	Dec 29, 18	1129	2997.25	15	1	Jan 6, 19	Dec 29, 18 9:22 AM	
30	Dec 30, 18	1130	23.5	10	1	Jan 1, 19	Dec 31, 18 3:47 AM	684
31	Dec 31, 18	1133	43.5	30	1	Jan 12, 19	Jan 6, 19 3:47 PM	747
32	Jan 2, 19	1140	27	52	1	Jan 11, 19	Jan 7, 19 6:21 PM	685
33	Jan 2, 19	1143	62.33	1	1	Jan 5, 19	Jan 4, 19 4:29 AM	
34	Jan 5, 19	1152	27	43	1	Jan 14, 19	Jan 6, 19 3:15 PM	741

Figure 19 - Sample Output Available With Excel's Idea Feature

Getting What You Need from Analyze Data

In addition to generating default summarizations and analysis, Analyze Data supports natural language queries. These queries enable you to extract the information you need from your data. For example, referring to the dataset shown in Figure 19, you could type "*sum of order amount for employee id 4*" into the Question box at the top of the pane. In turn, Analyze Data would create a PivotTable that summarizes the Order Amount field for the designated team member. Further, you could refine your query by entering "*sum of order amount for employee id 4 in December,*" and Analyze Data would filter the results to meet that criterion.

Excel's Analyze Data feature is a significant step toward enabling artificial intelligence in Excel. However, it is also a practical tool for uncovering trends, outliers, and other vital facts that we might miss with the human eye. Further, because of the tool's natural language query capabilities, we can "ask" Excel precisely what we want to know, and it will respond accordingly. Excel's Analyze Data feature has much to offer, and if you have access to it, you should try it soon.

Regression Analysis Fundamentals

Regression analysis is a statistical method used to estimate and understand relationships between a dependent variable and one or more independent variables. Linear regression models the relationship between two or more variables by "fitting" a linear equation to observed data.

One or more independent variables predict a single dependent variable in a regression analysis. For example, a data analyst may want to investigate the effect of advertising expenditures on sales revenue. The following linear regression equation explains relationships.

$$\mathbf{Y = \alpha + \beta} \boldsymbol{x}$$

Y is the dependent variable (sales revenue).
α is the intercept (the value of **Y** when ***x*** = 0).
β is the slope of the line.
x is the independent variable (advertising expenses).

The correlation coefficient quantifies the strength of the association between the independent and dependent variables. The correlation coefficient always takes a value between minus one and one, with one or minus one indicating perfect correlation (with all points along a straight line). A positive correlation indicates a positive association between the variables. Increasing values in one variable correspond to increasing values in the other variable. A negative correlation indicates a negative association between the variables. Increasing values in one variable correspond to decreasing values in the other variable. A correlation value of zero indicates no association between the variables.

R^2 is the square of the correlation coefficient. It shows the percentage of the dependent variable's variation explained by variation in the independent variable. For example, when the correlation between the variables sales revenue and advertising expense is 0.80, the linear regression model estimates that 64% (0.80^2) of the variability in the data is predictable. The major conceptual limitation of regression analysis is that it allows one to ascertain relationships without knowing the underlying causes.

For example, a strong positive relationship may exist between the number of overtime hours worked by production personnel and the number of goods produced; however, a regression analysis cannot guarantee that the quantity of goods produced caused overtime. Other factors, such as machinery breakdowns or production line inefficiencies, may have caused it. Thus, although regression analysis can quantify the relationship between overtime hours worked and the quantity of goods produced, it cannot explain the cause of the relationship.

Regression Analysis in Excel

Excel includes several tools helpful in conducting regression analysis. For example, the regression analysis tool in the Analysis ToolPak add-in facilitates estimating a single dependent variable from up to 16 independent variables. Further, regression analysis is not the only tool available in the ToolPak. Tools are also available to compute descriptive statistics, calculate moving averages and ANOVAs, generate random numbers, determine ranks and percentiles, and more.

In our first example, a sales executive wants to determine if a cause-and-effect relationship exists between monthly revenue and advertising expenditures incurred in the previous month. Regression analysis will evaluate the strength of the relationship between these variables and help to forecast future sales revenues.

Consider the data shown in **Figure 20**, which summarizes monthly revenue and advertising expenditures for the past thirty months. Notably, the data indicate that as advertising expenses in the prior month increase, sales revenues in the current month also increase. Likewise, as advertising expenses for the *preceding* month decrease, sales revenues for the *current* month decrease. The scatter chart in **Figure 21** shows this relationship, and a scatter chart suggests a significant relationship between monthly sales revenue and advertising expenses incurred in the previous month.

	A	B	C
1	Monthly Sales	Ad Expense 1 Month Prior	Ad Expense 2 Months Prior
2	$ 1,700,000	$ 20,000	$ 20,000
3	1,650,000	19,000	20,000
4	1,725,000	20,000	19,000
5	1,750,000	21,000	20,000
6	1,725,000	20,000	21,000
7	1,750,000	21,000	20,000
8	1,750,000	20,500	21,000
9	1,775,000	21,000	20,500
10	1,775,000	21,000	21,000
11	1,800,000	21,000	21,000
12	1,800,000	21,500	21,000
13	1,775,000	20,500	21,500
14	1,800,000	21,000	20,500
15	1,825,000	21,000	21,000
16	1,825,000	21,500	21,000
17	1,900,000	22,000	21,500
18	1,900,000	22,000	22,000
19	1,875,000	21,500	22,000
20	1,875,000	21,750	21,500
21	1,850,000	21,750	21,750
22	1,850,000	21,500	21,750
23	1,900,000	22,000	21,500
24	1,875,000	21,750	22,000
25	1,775,000	20,500	21,750
26	1,850,000	21,500	20,500
27	1,900,000	22,000	21,500
28	1,925,000	22,500	22,000
29	1,925,000	22,500	22,500
30	1,850,000	21,750	22,500
31	1,900,000	22,000	21,750

Figure 20 - Regression Analysis Data Set

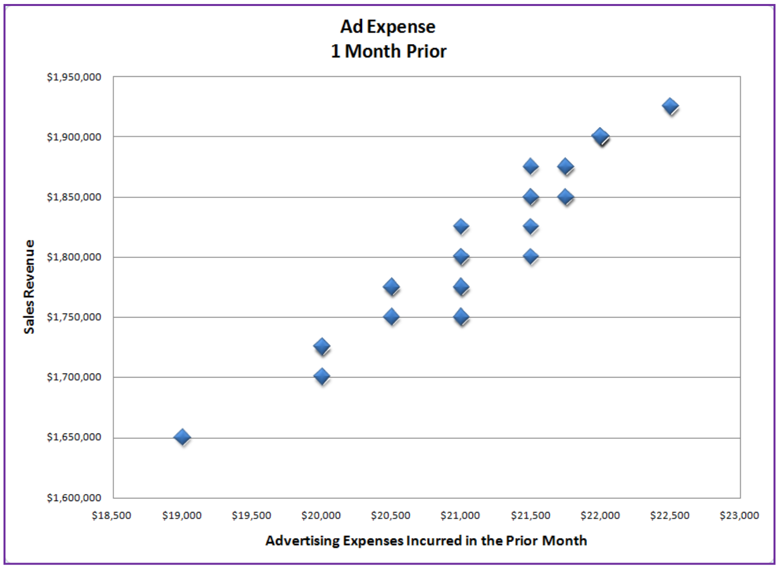


Figure 21 - Scatter Chart of the Regression Analysis Data Set

Click **Data Analysis** on the **Data** tab to begin the regression estimation process. Then, select **Regression** in the **Analysis Tools** list of the **Data Analysis** dialog box and click **OK**, as shown in **Figure 22**.

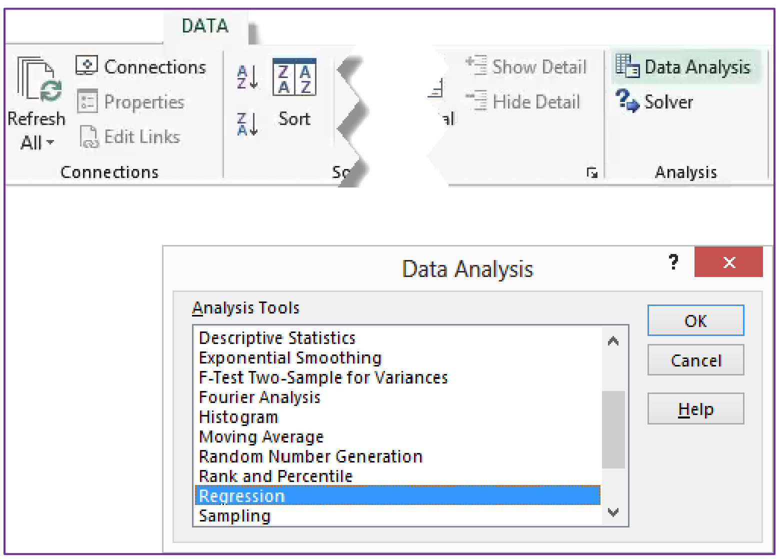


Figure 22 - Selecting Regression in the Data Analysis Dialog Box

In the **Regression** dialog box shown in **Figure 23**, click the **Collapse Dialog** button to the right of the **Input Y Range** box and select the range of cells containing the dependent variable. In this case, select the range **A2:A31**, Monthly Sales. Next, click the **Collapse Dialog** button to the right of the **Input X Range** box and choose the containing the independent variable. In this case, select the range **B2:B31**, Ad Expense 1 Month Prior. Note that the number of observations in the X and Y ranges must be equal.

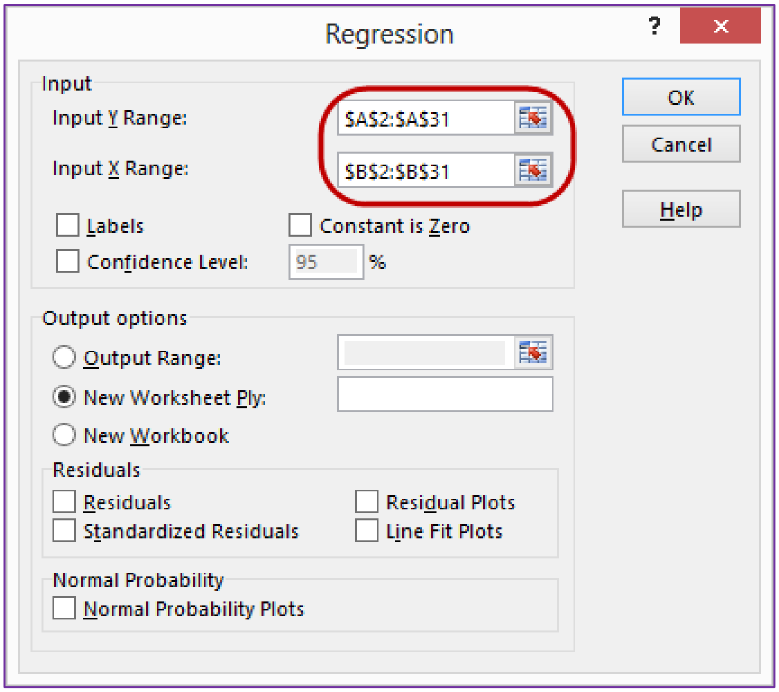


Figure 23 - Specifying the X and Y Ranges in a Simple Linear Regression

Click **OK** to calculate the regression and display the results on a new worksheet, as shown in **Figure 24**.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.954919307							
5	R Square	0.911870882							
6	Adjusted R Square	0.908723414							
7	Standard Error	21880.19817							
8	Observations	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	1.38699E+11	1.38699E+11	289.7156508	2.6595E-16			
13	Residual	28	13404806016	478743072					
14	Total	29	1.52104E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	10538.36094	106333.4823	0.099106704	0.921759697	-207275.9036	228352.6255	-207275.9036	228352.6255
18	X Variable 1	85.17872711	5.00432109	17.02103554	2.6595E-16	74.92784005	95.42961417	74.92784005	95.42961417

Figure 24 - Linear Regression Analysis Based on a Single Independent Variable

When interpreting regression results, key components include R-Squared, the **Intercept Coefficient**, and the **X Variable 1 Coefficient**. Recall that R Square measures the "fit" of the data – it measures how well the computed linear equation fits the data. Generally, an R Square value of 0.80 or higher or -0.80 or lower indicates a strong relationship between the independent variable(s) and the dependent variable. In this example, the R Square is 0.9118, which shows a strong relationship between the variables and that the linear equation fits the data very well.

The Intercept Coefficient and the X Variable 1 Coefficient are part of the equation for estimating or predicting new values of the dependent variable based on selected independent variable values. In this case, the equation for estimating monthly sales as a function of advertising expenses incurred in the prior month is as follows.

Monthly Sales = $10,538 + (85.179 * Ad Expense 1 Month Prior)

To forecast the sales revenue for a future month based on anticipated advertising expenses of $25,000 in the prior month, substitute the anticipated advertising expenses into the equation and solve for monthly revenue.

Monthly Revenue = $10,538 + (85.179 * $25,000)

$2,140,013 = $10,538 + (85.179 * $25,000)

In this case, the regression equation yields an estimated sales revenue of $2,140,013.

The regression analysis tool can regress up to sixteen independent (predictor) variables on a single dependent variable. To include more independent variables in the regression, expand the X Input Range to include the observations of the additional variables.

For example, many experienced marketing executives argue that advertising expenses have a lasting impact on sales. Therefore, to include advertising expenses incurred two months before the measurement of sales revenue, expand the X Input Range to include the column that contains the additional observed advertising expenses. Thus, in this example, the **Input X Range** incorporates two months of advertising expenses – range B2:C31, as shown in **Figure 25**.

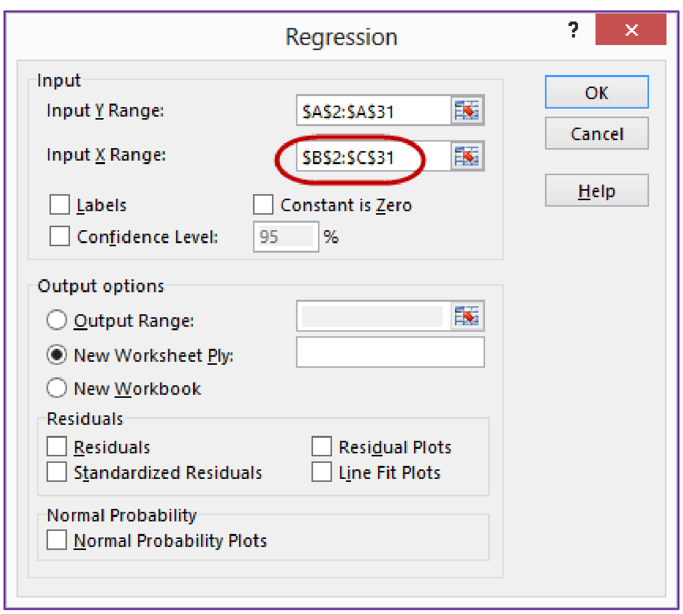


Figure 25 - Specifying the X and Y Ranges in a Multiple Linear Regression

The multiple linear regression performed in this example yields the calculations displayed in **Figure 26**.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.965000261							
5	R Square	0.931225503							
6	Adjusted R Square	0.926131096							
7	Standard Error	19683.5027							
8	Observations	30							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1.41643E+11	70821639574	182.7936937	2.01962E-16			
13	Residual	27	10460887519	387440278.5					
14	Total	29	1.52104E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-91683.19386	102594.5494	-0.893645856	0.379408192	-302189.8187	118823.431	-302189.8187	118823.431
18	X Variable 1	73.03751936	6.298189624	11.59658945	5.40485E-12	60.11470184	85.96033688	60.11470184	85.96033688
19	X Variable 2	17.00881257	6.170402417	2.756515933	0.010342499	4.34819274	29.6694324	4.34819274	29.6694324

Figure 26 - Linear Regression Analysis Based on Multiple Independent Variables

The results of the multiple linear regression suggest a better fit of the data since R Square has increased from 0.9118 in the simple linear regression to 0.9312 in the current analysis. The results support the assertion that advertising expenditures have a lasting effect—at least in this example. The summary output displays a second X-variable coefficient that corresponds to the second independent variable. Therefore, as shown below, you must revise the formula for predicting sales revenue based on advertising expenditures to two months' expenses.

Monthly Sales = $-91,683 + (73.038 * Ad Expense 1 Month Prior) + (17.009 * Ad Expense 2 Months Prior)

Adding a Trendline to a Scatter Chart

There is an alternative to using the regression analysis tool to generate a regression equation to forecast future revenues or expenses. That alternative involves adding a linear trendline with the equation and R Square to a scatter chart of the data points, as shown in **Figure 27**.

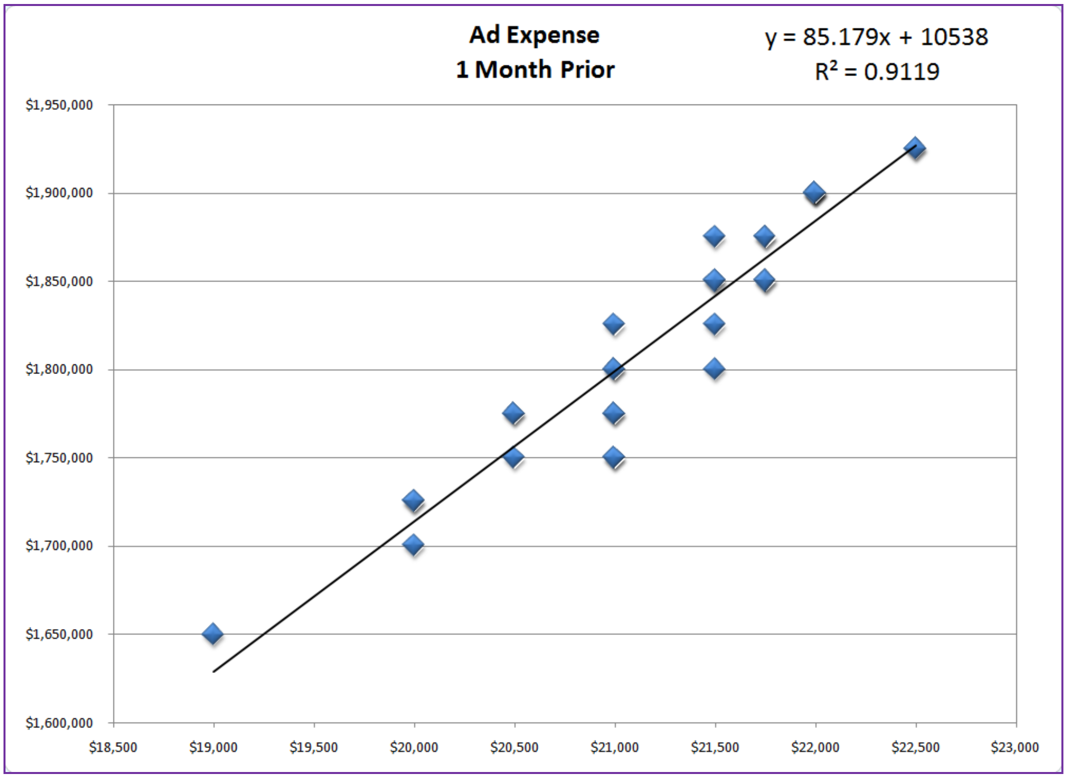


Figure 27 - Adding a Linear Trendline with R-Squared and Equation to a Scatter Chart

Adding a linear trendline uses the same least-squares process to calculate the results. Note that the R-Squared value and regression equation are the same as those produced using the regression tool in the Analysis ToolPak.

To add a trendline to a scatter chart, right-click on one of the data points in the chart and select **Add Trendline**. Then, in the **Format Trendline** dialog box, click **Linear** as the **Trend/Regression Type** and check **Display Equation on Chart** and **Display R-squared value on chart**, as shown in **Figure 28**.

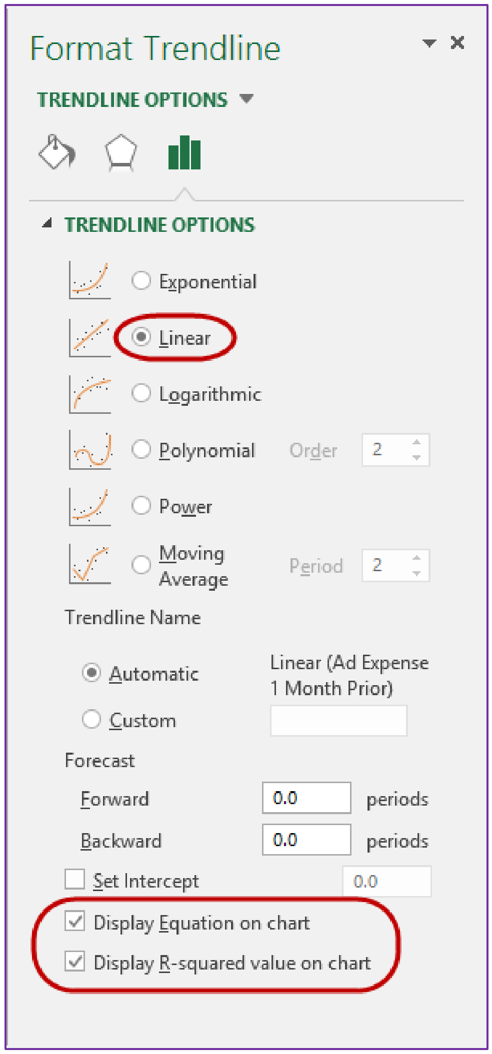


Figure 28 - Adding a Linear Trendline to a Scatter Chart

Using Regression Functions

In addition to adding trendlines to charts, Excel contains numerous statistical functions for calculating the components produced by regression analysis. These functions allow users to construct simple formulas from observed data to create regression equations that help forecast revenues or expenses.

- **FORECAST**
- **SLOPE**
- **INTERCEPT**
- **TREND**
- **GROWTH**
- **LINEST**

FORECAST can calculate or predict a value based on existing values. Using the sample data as an example, the formula **=FORECAST(25000,A2:A31,B2:B31**) predicts the monthly sales associated with advertising expenses of $25,000 one month prior, assuming a linear relationship between the independent variable in range B2:B31 and the dependent variable in range A2:A31.

The FORECAST function generates the linear regression equation to predict monthly sales from advertising revenue. However, this method does not calculate or display R Square, nor does it show the intercept or slope of the linear equation. Nevertheless, FORECAST is an excellent alternative for predicting values using simple formulas to estimate revenues and expenses when relationships exist and the data fit is good.

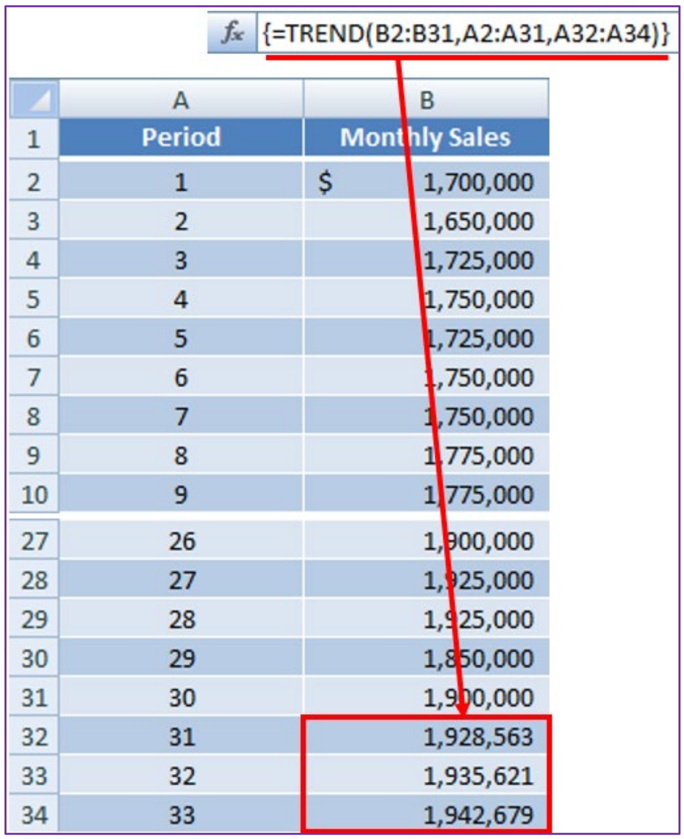
SLOPE calculates the slope of the linear regression line, given the independent and dependent variables' values. For example, using the sample data, the formula **=SLOPE(C2:C32, B2:B32)** returns a slope of 85.179, the same value computed by the simple linear regression analysis discussed earlier.

INTERCEPT calculates the intercept of the regression line, given the values of the independent and dependent variables. Again, using the sample data, the formula **=INTERCEPT(C2:C32,B2:B32)** returns an intercept of $10,538, the same value computed by the simple linear regression analysis.

TREND generates values along a linear trendline and can predict multiple future dependent values based on the relationship between existing dependent and independent variables. Note that TREND formulas are array formulas, and you must use a **SHIFT+CTRL+ENTER** keystroke sequence to enter them in non-dynamic array-aware versions of Excel.

GROWTH is very similar to TREND, except that GROWTH provides an exponential calculation rather than a linear calculation as provided by the TREND function. Again, like TREND, the GROWTH function must be entered as an array formula in older versions of Excel.

LINEST can produce an array formula to calculate the slope and intercept of a linear regression without using the Analysis ToolPak. It can also be used to generate descriptive statistics for linear regressions, including R-squared. As shown in **Figure 29**LINEST calculates the slope and intercept from the dependent and independent variables in an array formula. Before entering the formula, select two adjacent cells – such as cells E2 and F2 – to produce the displayed output. To modify the LINEST formula to generate descriptive statistics, highlight a range two columns wide by five rows deep and construct the array formula shown in **Figure 29**.


{=TREND(B2:B31,A2:A31,A32:A34)}

	A	B
1	Period	Monthly Sales
2	1	$ 1,700,000
3	2	1,650,000
4	3	1,725,000
5	4	1,750,000
6	5	1,725,000
7	6	1,750,000
8	7	1,750,000
9	8	1,775,000
10	9	1,775,000
27	26	1,900,000
28	27	1,925,000
29	28	1,925,000
30	29	1,850,000
31	30	1,900,000
32	31	1,928,563
33	32	1,935,621
34	33	1,942,679

Figure 29 - Using LINEST to Calculate the Regression Intercept and Slope

Additional Forecasting Features

For those who engage in budgeting and forecasting activities or other activities where predicting with some degree of accuracy what will happen in the future is necessary, the **FORECAST.ETS** function added beginning with Excel 2016 is noteworthy. Before FORECAST.ETS, the **FORECAST** function provided only linear forecasting capabilities. However, with FORECAST.ETS, you can generate forecasts with exponential smoothing, as pictured in **Figure 30**.

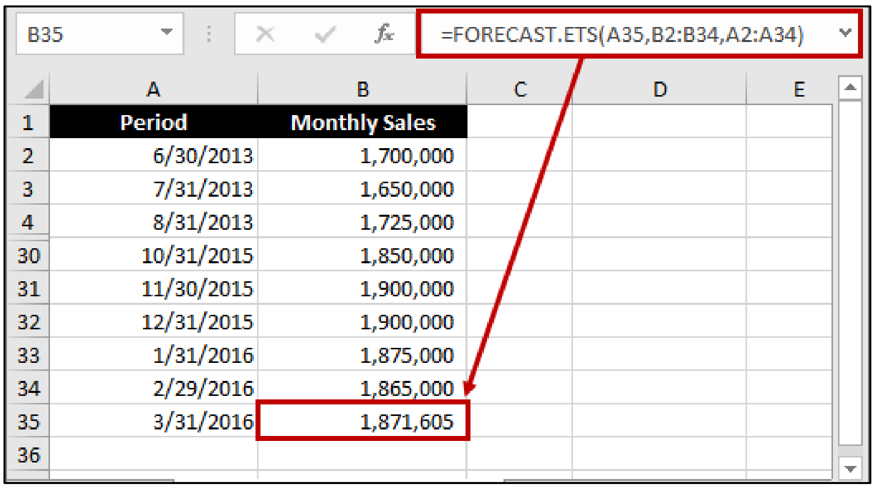
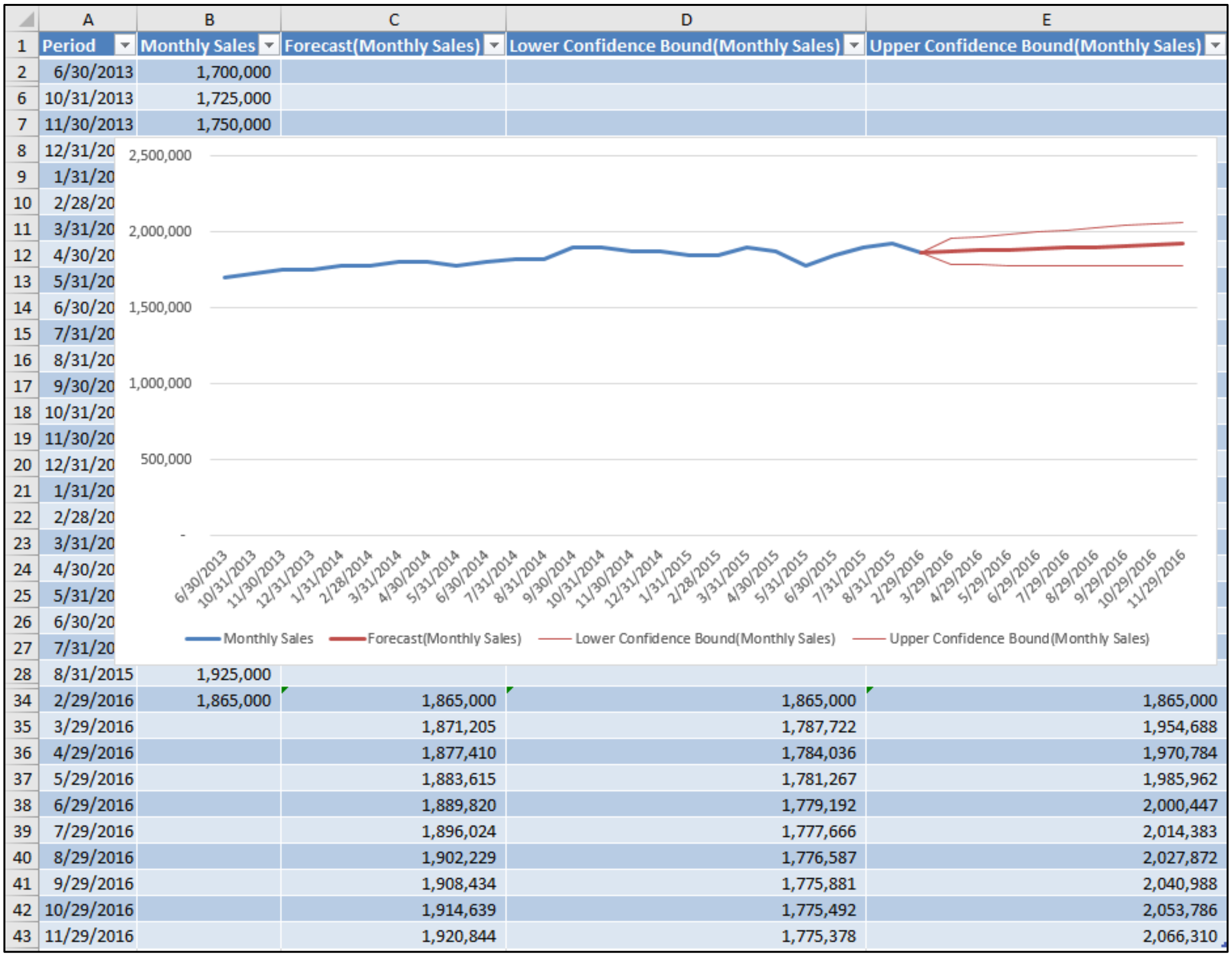


Figure 30 - Using FORECAST.ETS in Excel

Additionally, you can use the **Forecast Sheet** option on the Ribbon's **Data** tab to create a visualization of your forecast and modify parameters, including seasonality.



	A	B	C	D	E
1	Period	Monthly Sales	Forecast(Monthly Sales)	Lower Confidence Bound(Monthly Sales)	Upper Confidence Bound(Monthly Sales)
2	6/30/2013	1,700,000			
6	10/31/2013	1,725,000			
7	11/30/2013	1,750,000			
8	12/31/20				
9	1/31/20				
10	2/28/20				
11	3/31/20				
12	4/30/20				
13	5/31/20				
14	6/30/20				
15	7/31/20				
16	8/31/20				
17	9/30/20				
18	10/31/20				
19	11/30/20				
20	12/31/20				
21	1/31/20				
22	2/28/20				
23	3/31/20				
24	4/30/20				
25	5/31/20				
26	6/30/20				
27	7/31/20				
28	8/31/2015	1,925,000			
34	2/29/2016	1,865,000	1,865,000	1,865,000	1,865,000
35	3/29/2016		1,871,205	1,787,722	1,954,688
36	4/29/2016		1,877,410	1,784,036	1,970,784
37	5/29/2016		1,883,615	1,781,267	1,985,962
38	6/29/2016		1,889,820	1,779,192	2,000,447
39	7/29/2016		1,896,024	1,777,666	2,014,383
40	8/29/2016		1,902,229	1,776,587	2,027,872
41	9/29/2016		1,908,434	1,775,881	2,040,988
42	10/29/2016		1,914,639	1,775,492	2,053,786
43	11/29/2016		1,920,844	1,775,378	2,066,310

Figure 31 - Completed Forecast Sheet Generated in Excel

Solver

One of the critical considerations for anyone preparing forecasts in the field of prescriptive analytics is maximizing a specific measure – such as profit – given available resources and operational and financial constraints. Constraints on profitability may include distribution restrictions, production capacity limits, limited inventory, or shortages of working capital due to growth in accounts receivable or inventory. Regardless of the constraints, every financial manager aims to find solutions that maximize organizational value. Excel's **Solver** add-in can assist in that process.

Solver identifies optimal solutions to *linear programming models*. For example, you can use Solver to maximize gross profits by calculating the optimal mix of products sold, minimize interest expense by finding the optimal capital structure, or maximize production output by calculating the most efficient production runs. The Solver feature optimizes a value in a single workbook cell by adjusting various input cells, subject to user-specified constraints on those cells.

In this example, the CFO of a medical technology lab seeks to maximize the gross profit generated by a piece of testing equipment. The equipment can perform three tasks with varying billing rates, marginal costs, and gross profits. Further, the maximum number of units the lab can sell for each procedure varies. **Table 2** displays the relevant data.

	Task A	Task B	Task C
Billing Rate	$350	$250	**$200**
Marginal Cost	$175	$50	**$100**
Gross Profit Per Unit Sold	$175	$200	**$100**
Max Number Of Sales Per Year	**500 minus 1/2 of the instances of Tasks B and C sold for the year**	**500 minus 1/2 of the instances of Task C sold for the year**	**500**

Table 2 - Relevant Data for a Solver Equation

In addition to the data presented in Table 2, the lab can sell no more than 750 procedures during the year. Additionally, they must sell at least 50 units of each procedure each year. Given these constraints, the CFO must calculate the optimal mix of each task to maximize overall gross profit. Using Solver, planners can quickly calculate a solution to this complex revenue-maximization model.

First, build the basic structure of the worksheet, as shown in **Figure 32**. The second panel in the figure displays the underlying formulas.

	A	B	C	D	E	F
1		Task A	Task B	Task C		
2	Maximum Number Available	450	475	500	Total Sold Cannot Exceed	750
3	Billing Rate	$350.00	$250.00	$200.00		
4	Marginal Cost	$175.00	$ 50.00	$100.00		
5	Gross Profit Per Unit	$175.00	$200.00	$100.00		
6	Number Sold	50	50	50	Total Sold	150
7	Total Gross Profit	$ 8,750	$10,000	$ 5,000	Total Gross Profit	$ 23,750

	A	B	C	D	E	F
1		Task A	Task B	Task C		
2	Maximum Number Available	=500-(0.5*(D6+C6))	=500-(0.5*D6)	500	Total Sold Cannot Exceed	750
3	Billing Rate	350	250	200		
4	Marginal Cost	175	50	100		
5	Gross Profit Per Unit	=B3-B4	=C3-C4	=D3-D4		
6	Number Sold	50	50	50	Total Sold	=SUM(B6:D6)
7	Total Gross Profit	=B5*B6	=C5*C6	=D5*D6	Total Gross Profit	=SUM(B7:D7)

Figure 32 - Constructing a Worksheet with the Appropriate Structure and Formulas

To find a solution, click **Solver** on the **Data** tab to open the **Solver Parameters** dialog box shown in **Figure 33**. In the **Solver Parameters** dialog box, specify the target cell to be optimized, in this case, cell **F7**. The variable cells to be adjusted are the range **B6:D6**. The table above identifies the relevant constraints. Next, click **Solve**, and Solver will attempt to find a solution based on the input cells and the limitations identified. When Solver completes the calculations, select **Keep Solver Solution** in the **Solver Results** dialog box and click **OK**.

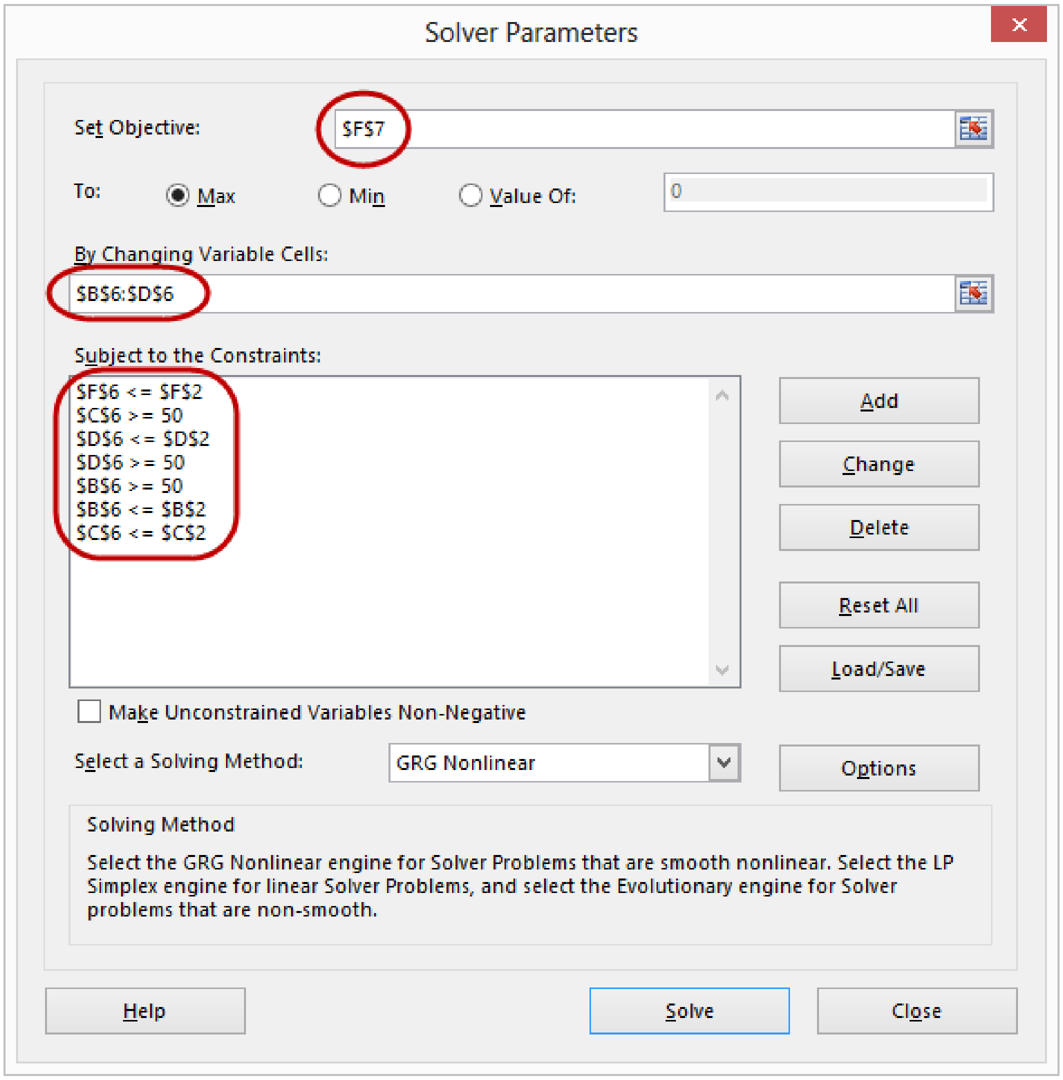


Figure 33 - Specifying the Parameters for a Solver Optimization

Figure 34 illustrates the suggested solution identified by Solver. The lab will generate maximum profits when it sells 225 units of Task A, 475 units of Task B, and 50 units of Task C. Closer inspection reveals that the models satisfy all identified constraints.

	Task A	Task B	Task C		
Maximum Number Available	238	475	500	Total Sold Cannot Exceed	750
Billing Rate	$ 350.00	$ 250.00	$ 200.00		
Marginal Cost	$ 175.00	$ 50.00	$ 100.00		
Gross Profit Per Unit	$ 175.00	$ 200.00	$ 100.00		
Number Sold	225	475	50	Total Sold	750
Total Gross Profit	$ 39,375	$ 95,000	$ 5,000	Total Gross Profit	$ 139,375

Figure 34 - Results of Solver Produced Profit Maximization

In addition to Solver's computational efficiency, data analysts can benefit from its documentation features, including several reports. **Figure 35** illustrates the **Solver Answer Report**.

Microsoft Excel 15.0 Answer Report
Worksheet: [Figure 074 Through Figure 077 - Solver.xlsx]GP Maximization
Report Created: 3/10/2013 5:43:57 PM
Result: Solver found a solution. All Constraints and optimality conditions are satisfied.
Solver Engine
Engine: GRG Nonlinear
Solution Time: 0.016 Seconds.
Iterations: 0 Subproblems: 0
Solver Options
Max Time 100 sec, Iterations 100, Precision 0.000001
Convergence 0.0001, Population Size 100, Random Seed 0, Derivatives Forward, Require Bounds
Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 5%, Solve Without Integer Constraints

Objective Cell (Max)

Cell	Name	Original Value	Final Value
$F$7	Total Gross Profit	$ 139,375	$ 139,375

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$B$6	Number Sold Task A	225	225	Contin
$C$6	Number Sold Task B	475	475	Contin
$D$6	Number Sold Task C	50	50	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$F$6	Total Sold	750	$F$6<=$F$2	Binding	0
$B$6	Number Sold Task A	225	$B$6<=$B$2	Not Binding	12.5
$C$6	Number Sold Task B	475	$C$6<=$C$2	Binding	0
$D$6	Number Sold Task C	50	$D$6>=50	Binding	-
$D$6	Number Sold Task C	50	$D$6<=$D$2	Not Binding	450
$C$6	Number Sold Task B	475	$C$6>=50	Not Binding	425
$B$6	Number Sold Task A	225	$B$6>=50	Not Binding	175

Figure 35 - Answer Report Generated with a Solver Solution

In addition to **Answer** reports, Solver can also generate **Sensitivity Analysis** and **Limits** reports. Further, users can save Solver solutions as **Scenarios** for later recall in the **Scenario Manager**.

Analyzing Quality of Data with Power Query

First added to Excel with the 2010 release, Power Query offers many options for linking external data from virtually any common source into Excel. Further, Power Query provides many tools for transforming data, so individual users do not waste time manually modifying data. Importantly, once you establish a query using Power Query and incorporate any necessary transformations, you can "refresh" your queries to get updated data into your analysis.

One of Power Query's key features is the option to quickly analyze your data for potential errors, omissions, and other "quality" issues. To use this feature, query the data you wish to explore into Excel. Complete this task by clicking **Get Data** on the **Data** tab of the Ribbon, then selecting your data source. In this example, choose the **Xtreme** Access database provided and select the **Orders Detail** table. Then, click the **Transform Data** button near the window's lower right corner to complete your query, as shown in **Figure 36**.

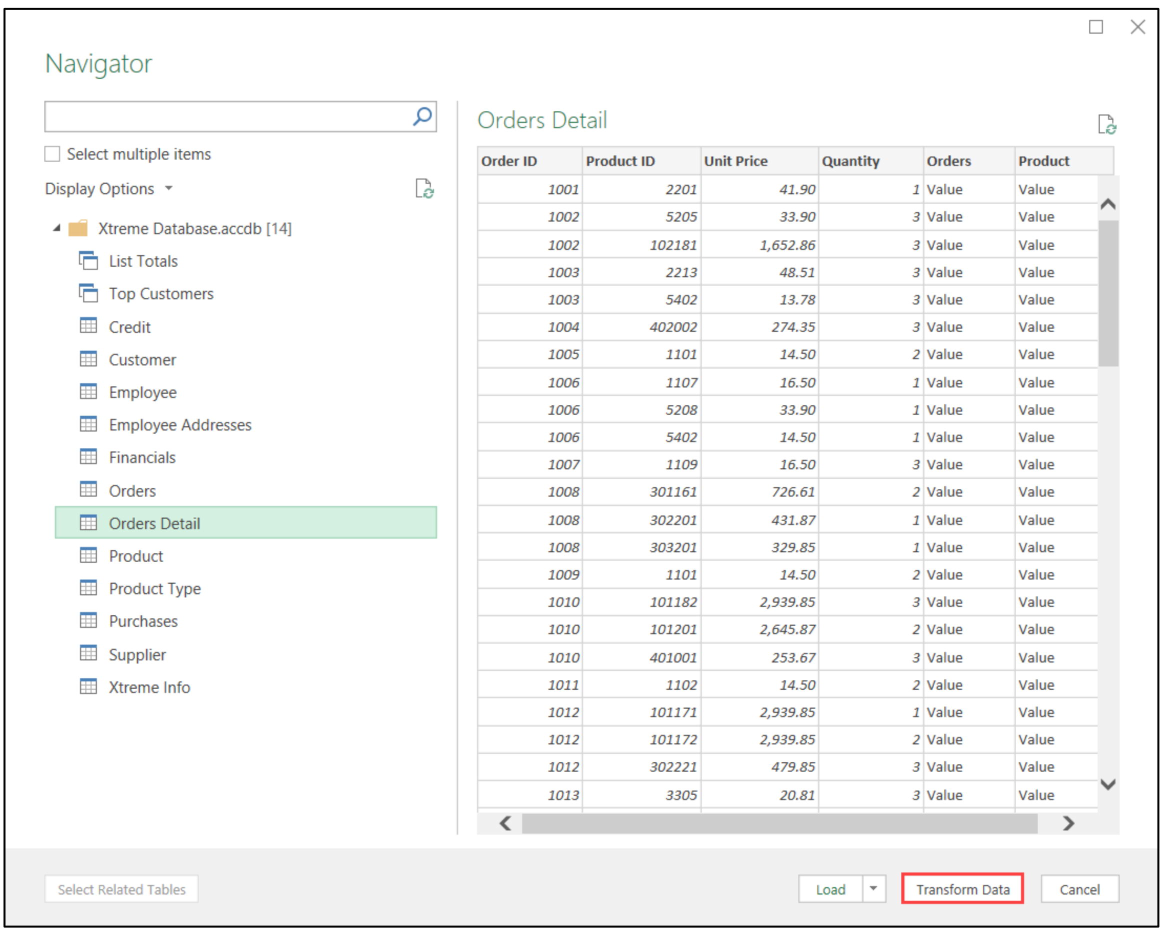


Figure 36 - Querying Data Using Power Query

The previous actions will load the data into Excel's data model and open the Power Query Editor. Next, click the **View** tab and check the boxes next to the following three labels in that window.

- **Column quality**. This feature tests the data for valid, erroneous, and empty cells.
- **Column distribution**. This feature examines the data and profiles it to identify distinct and unique values. It also includes an option to remove duplicates.
- **Column profile**. This feature computes various statistics and shows the data distribution.

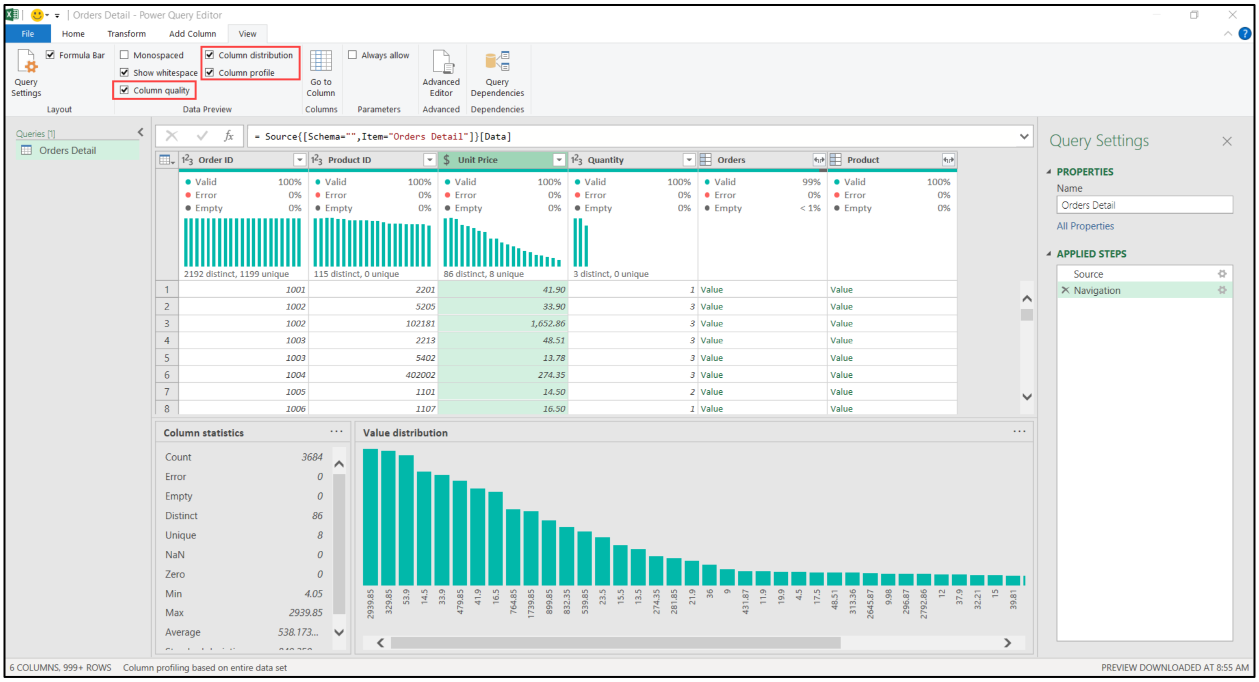


Figure 37 - Sample Output from Power Query's Data Quality Tests

Examples of Useful Third-Party Excel Add-Ins

Although Excel offers over 500 functions and valuable tools such as the PivotTables, Power Query, and Analysis ToolPak, it is not a complete tool that satisfies all our data analytics needs. This section examines two examples of add-in solutions that can enhance Excel's capabilities when performing data analytics tasks.

RiskAMP

Although tools like Solver (discussed previously) are helpful in many environments, they have a particular limitation – they do not account for uncertainty and risk in the estimates used in spreadsheet models. Instead, they rely on single-point assumptions that yield single-point results, virtually guaranteed to be incorrect. Yet, these models remain in use in practically all prescriptive analytics environments.

Some estimates will likely be too conservative in the single-point estimate environment, while others will be too optimistic. The combined errors in each calculation often lead to a real-life result that is significantly different from the one anticipated. Based on the expected result from a simple Excel model, a decision may be wrong —one that would never be made if a complete picture of all possible outcomes were available. That's where Monte Carlo simulation comes in. Monte Carlo simulations generate models that account for all possible outcomes, based on the likelihood of each assumption. Further, Monte Carlo simulations identify the inputs that most affect the result of a decision. Although Excel does not provide built-in Monte Carlo simulation capabilities, you can use inexpensive add-ins such as **RiskAMP** to facilitate this type of analysis.

To set up a project cash flow model to use Monte Carlo simulation with RiskAMP, you must first specify all assumption input cells as distributions instead of point estimates. Next, select **Insert Distribution** from the **Monte Carlo** tab of the **Ribbon** to open the **Insert Random Distribution** dialog box. Next, choose a distribution type and then fill in the requested ranges. For example, **Figure 38** illustrates an assumption input of expected unit sales as a **Triangular Distribution** with **Minimum**, **Most Likely**, and **Maximum** values of 3200, 3500, and 3800 units, respectively. Once you update the model to replace all point estimates with appropriate distributions, you are ready to set up a results sheet and run a simulation.

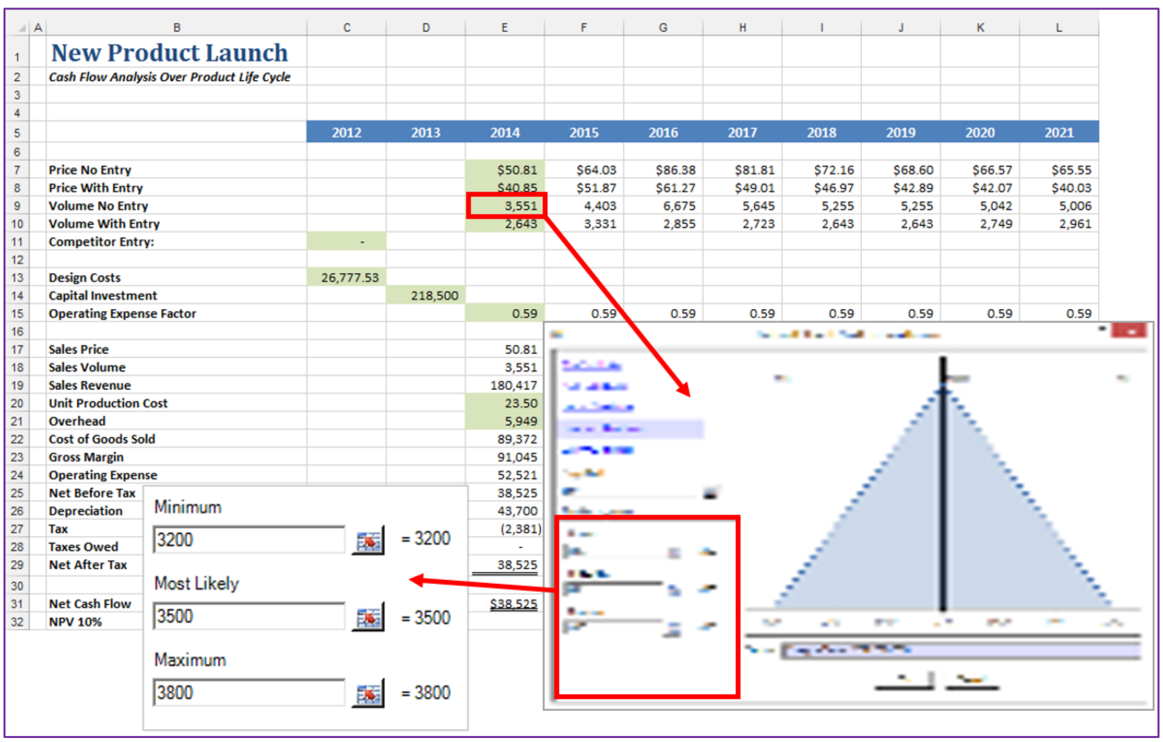
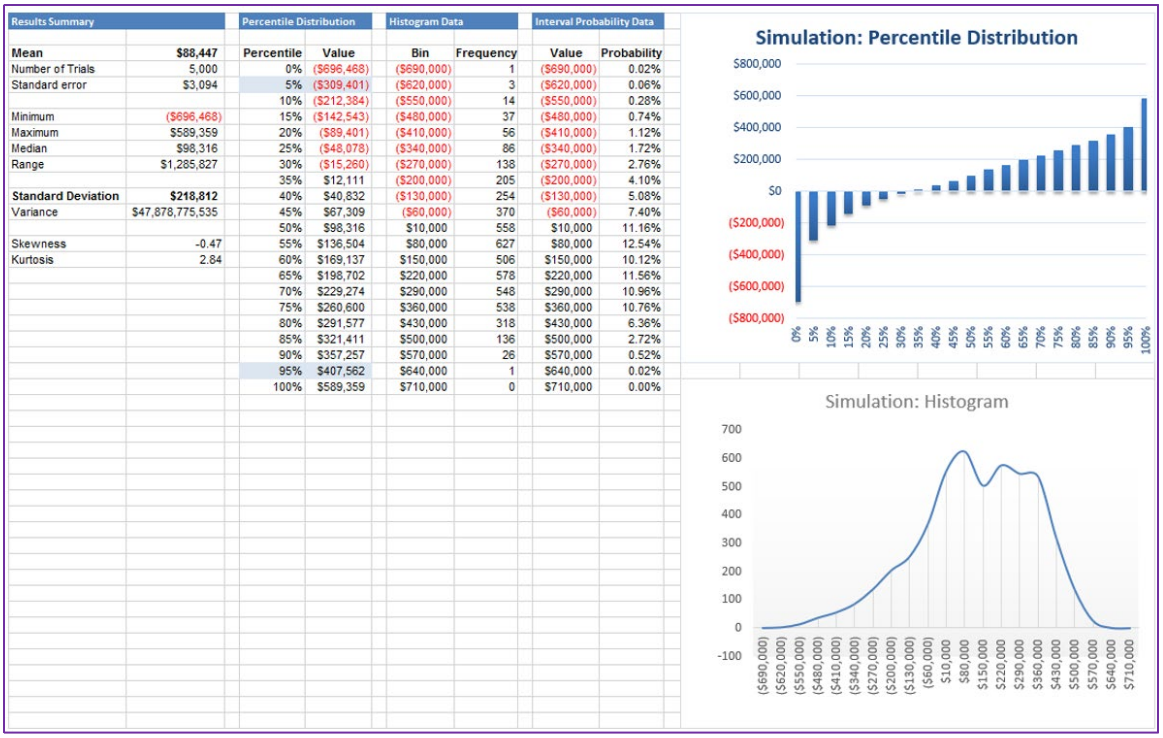


Figure 38 - Specifying an Assumption as a Distribution in RiskAMP

To set up a results sheet, select **Histogram & Chart Wizard** from the **Monte Carlo** tab of the Ribbon. Next, specify the results cell of your model - in this case, the cell containing the estimated NPV for the new product introduction - and then choose to **Create a new Simulations Results**

Sheet. Next, click **Run a Monte Carlo Simulation**, then select **Chart the results table data** to produce the output shown in **Figure 39**.



Results Summary	
Mean	**$88,447**
Number of Trials	5,000
Standard error	$3,094
Minimum	($696,468)
Maximum	$589,359
Median	$98,316
Range	$1,285,827
Standard Deviation	**$218,812**
Variance	$47,878,775,535
Skewness	-0.47
Kurtosis	2.84

Percentile Distribution	
Percentile	**Value**
0%	($696,468)
5%	($309,401)
10%	($212,384)
15%	($142,543)
20%	($89,401)
25%	($48,078)
30%	($15,260)
35%	$12,111
40%	$40,832
45%	$67,309
50%	$98,316
55%	$136,504
60%	$169,137
65%	$198,702
70%	$229,274
75%	$260,600
80%	$291,577
85%	$321,411
90%	$357,257
95%	$407,562
100%	$589,359

Histogram Data		Interval Probability Data	
Bin	**Frequency**	**Value**	**Probability**
($690,000)	1	($690,000)	0.02%
($620,000)	3	($620,000)	0.06%
($550,000)	14	($550,000)	0.28%
($480,000)	37	($480,000)	0.74%
($410,000)	56	($410,000)	1.12%
($340,000)	86	($340,000)	1.72%
($270,000)	138	($270,000)	2.76%
($200,000)	205	($200,000)	4.10%
($130,000)	254	($130,000)	5.08%
($60,000)	370	($60,000)	7.40%
$10,000	558	$10,000	11.16%
$80,000	627	$80,000	12.54%
$150,000	506	$150,000	10.12%
$220,000	578	$220,000	11.56%
$290,000	548	$290,000	10.96%
$360,000	538	$360,000	10.76%
$430,000	318	$430,000	6.36%
$500,000	136	$500,000	2.72%
$570,000	26	$570,000	0.52%
$640,000	1	$640,000	0.02%
$710,000	0	$710,000	0.00%

Figure 39 - Simulation Results of 5,000 Iterations in RiskAMP

From the results, you can quickly determine the NPV of the proposed product ranges between -$696,468 and $589,359, a range of $1,285,827, with a mean result of $88,447. From the percentile distribution, you can be 90% confident that the NPV will range from -$309,401 to $407,562. Also, note that there is more than a 30% chance that the project will generate a negative NPV. As you can see, the evaluative information provided by simulation results is far superior to that from simple point-estimate models.

A formal risk assessment requires correlating the individual assumptions with the target output. Notwithstanding that simulations produce superior information for decision-making than single-point estimate models, the simulation results do not identify which inputs have the greatest impact on NPV. RiskAMP calculates this information but does not automatically present it; however, analysts can examine these relationships using the custom functions provided with the RiskAMP add-in. The two most relevant functions for this analysis are **SimulationCorrelation** (**SCorr**) and **SimulationRSquare** (**SRSq**). SCorr returns the correlation coefficient for the relationship between the simulation output (in this case, NPV) and a specified simulation assumption input. Thus, it measures the strength of the relationship between the results and a single variable. SRSq is simply the square of the correlation coefficient. It measures the percentage of output variability explained by the variability of an input. You can determine which inputs significantly impact the results by creating simple formulas. **Figure 40** presents just such an analysis, along with an explanatory chart.

Risk Analysis

Assumption Outputs	Cell
Project NPV	C32

Regression Sensitivity to Project NPV			
Assumption Inputs	Cell	Correlation	RSquare
Unit Production Cost	E20	-0.710	0.504684
Competitor Entry:	C11	-0.621	0.385303
Overhead	E21	0.453	0.205222
Operating Expense Factor	E15	-0.288	0.082692
Capital Investment	D14	-0.084	0.00705
Price No Entry	E7	0.072	0.005238
Price With Entry	E8	0.029	0.00085
Volume No Entry	E9	0.053	0.002823
Volume With Entry	E10	0.008	0.000064
Design Costs	C13	-0.061	0.003665

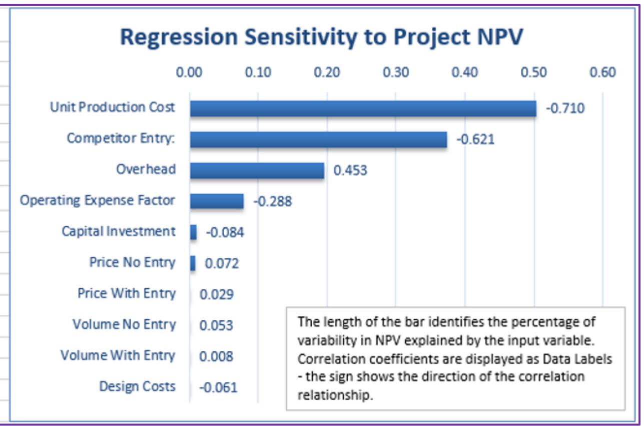


Figure 40 - Sensitivity of Project NPV to Assumption Inputs

From a prescriptive analytics perspective, it should be clear that single-point estimate models are severely lacking. Accordingly, add-ins such as RiskAMP can help data analysts provide better, more accurate results in less time than performing Monte Carlo simulations manually.

Active Data

ActiveData is an Excel add-in that offers over 100 features – many of which can be helpful when analyzing data in Excel workbooks. More specifically, ActiveData provides tools for performing advanced data analysis in Excel workbooks. To illustrate, consider the Professional Edition for accounting, audit, banking, and other finance-related professions. This application allows users to perform data stratification, aging, sampling (random, stratified, and monetary/PPS), digital analysis using Benford's Law, and produce cross-tabulations of large datasets.

Account and Transaction Analysis

In this first example, an auditor must ensure that a spreadsheet includes all checks (or any other transaction document). This type of testing is called "gap analysis." **Figure 41** depicts performing such a test using ActiveData. For example, gaps in the sequenced numbers of checks, invoices, and bills may indicate simple input errors. However, they could also show potentially fraudulent activity. To begin identifying missing documents, select **Gaps** on the **ActiveData** tab. Then, in the **Gap Analysis** dialog box, select the column of data to analyze. This example examines the **Num** column (the check number field). Next, click **Preview** to display the gaps detected in the **Preview** pane. If desired, click **Finished** to add the results to a new worksheet.

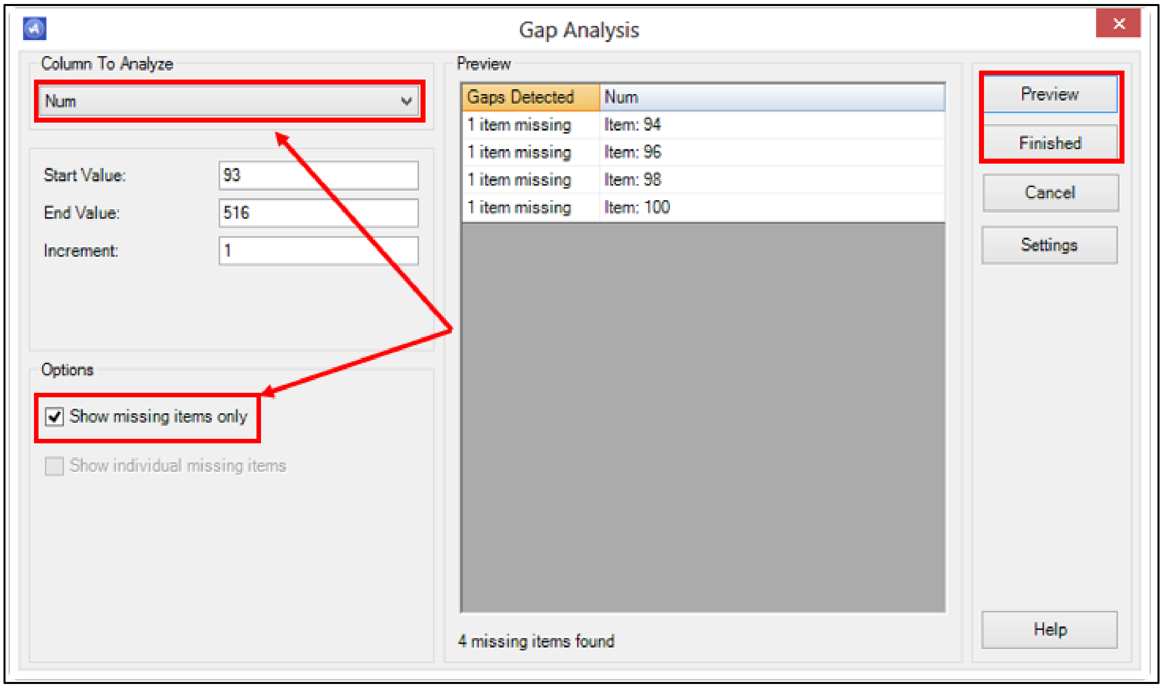


Figure 41 – Gap Analysis Using ActiveData

Merging Worksheets

You can perform the following tasks with ActiveData's worksheet commands, shown in **Figure 42**.

1. Merge, compare, and split worksheets;
2. Combine, duplicate, and sort data to new worksheets;
3. Remove blank rows, split and combine columns;
4. Remove hidden characters from data;
5. Query worksheets by formula or by example; and
6. Tag rows by multiple criteria.

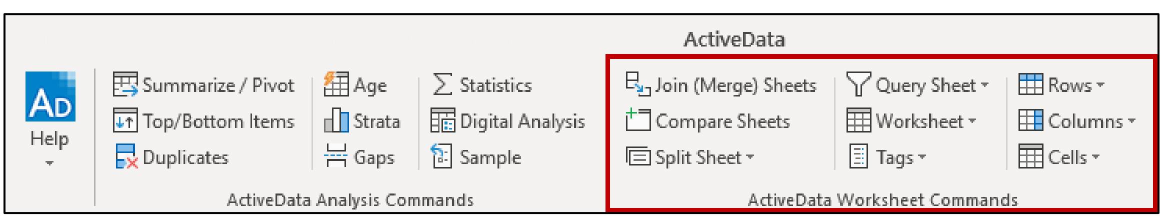


Figure 42 - ActiveData's Worksheet Commands

Frequently, users have data on multiple sheets that they must merge to complete an analysis or report. A conventional way to do this is to use VLOOKUP, XLOOKUP, or a similar Excel function to join the data on a common column. However, ActiveData can merge the required columns into a single table for analysis with a few clicks without building or copying complex formulas.

To merge data from multiple worksheets, do the following.

- Click one of the worksheets to merge and select **Join (Merge) Sheets** on the **ActiveData** tab.

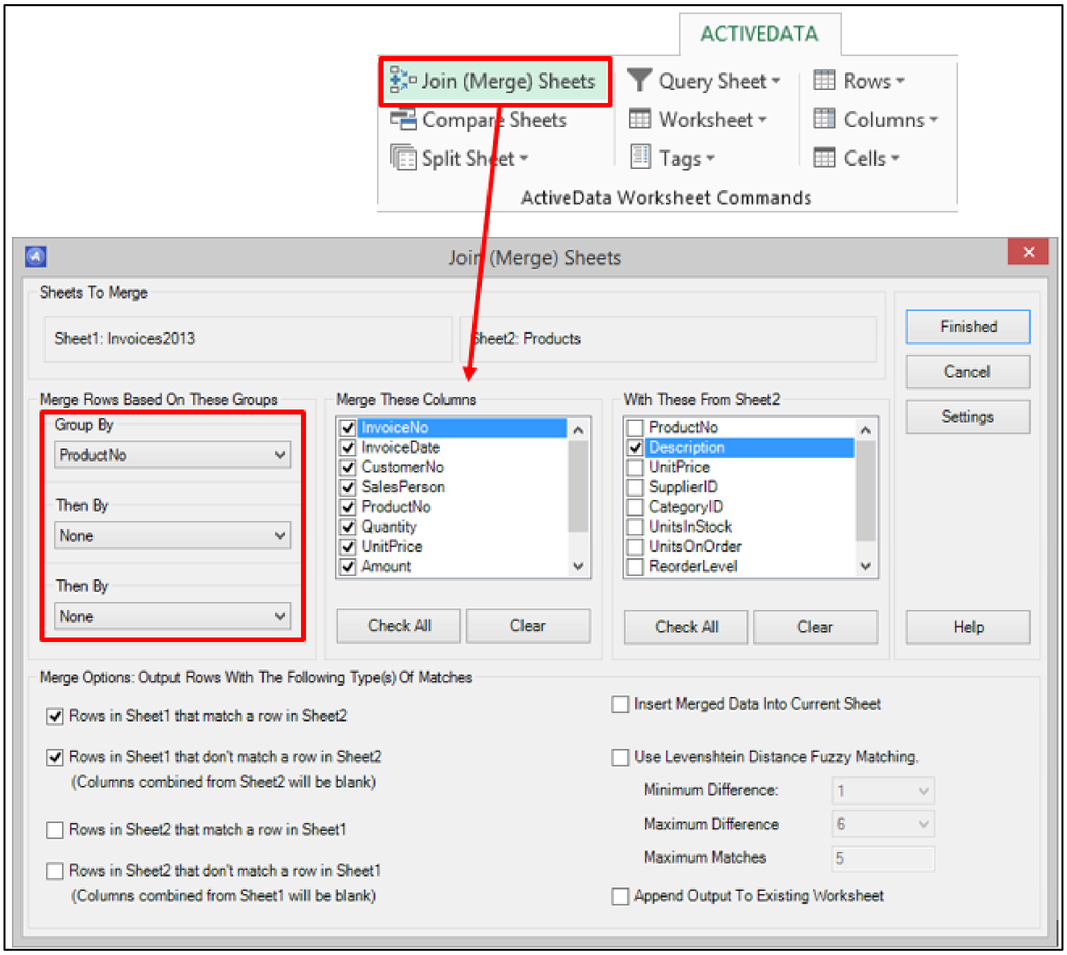


Figure 43 – Joining Sheets on a Common Column

- **Select a Sheet to Merge With** dialog box, select a worksheet to merge into the open worksheet, and click **Select**. Note that the worksheets to join need not be in the same workbook.

- In the **Join (Merge) Sheets** dialog box, select **ProductNo** in the **Group By** drop-down menu, as shown in **Figure 43**. In this case, ProductNo represents the *common column* between the sheets.

- In the **Merge These Columns** box on the left, check the columns from the initial worksheet to include in the merged data. Then, check the columns to merge in the **With These From Sheet 2** box on the right. Finally, select any desired **Merge Options** and click **Finished** to complete the process.

Similarly, you can split data by dates – year, month, week, day, or day of the week. Select **Split Sheet**, **Split Sheet By Date** from the ActiveData tab, and follow the prompts to divide a sheet by dates. ActiveData will place each date group on a separate worksheet.

Stratifying Data

Stratifying data is often helpful for identifying the areas of greatest risk. ActiveData can be used to easily stratify data. Select **Strata** from the **ActiveData** tab, define the stratification bands, and click **Stratify**. Click **Finished** to add the results and a summarizing chart to a new sheet in the workbook, as shown in **Figure 44**. Having created the stratification, you can determine if the data aligns with expectations. If it does not, further testing may be warranted to determine the reason for the deviation.

Strata #	>=	<	Count	% of Total	OrderAmount	% of Total
1.00	-	1,000.00	1,225	55.89	340,281.08	8.33
2.00	1,000.00	20,000.00	967	44.11	3,743,384.26	91.67
< exceptions			0	0.00	0.00	0.00
>= exceptions			0	0.00	0.00	0.00
Totals			**2,192**	**100.00**	**4,083,665.34**	**100.00**

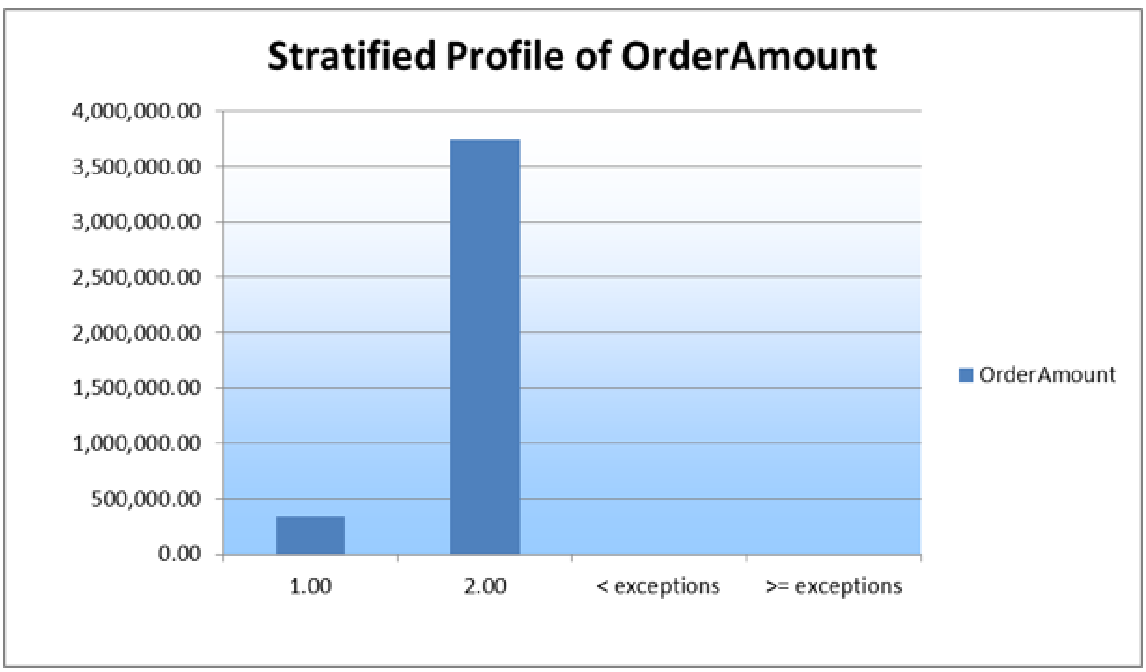


Figure 44 – Stratification Results and Chart in ActiveData

Benford Law Analysis

ActiveData provides an automated process for applying Benford's Law. Begin the analysis by selecting **Digital Analysis** from the **ActiveData** tab to open the **Digital Analysis Using Benford's**

Law dialog box. Choose **InvoiceAmount** as the **Column to Analyze** in the dialog box, and set the **Confidence Level** to **95.00%**. Next, check **Chart Results**, **First Digit Test**, **Second Digit Test**, and **First 2 Digits Test**, as shown in **Figure 45**. Click **Finished**, and ActiveData will complete and display the analysis on a new worksheet. **Figure 46** depicts the results of the **First Digit Analysis**.

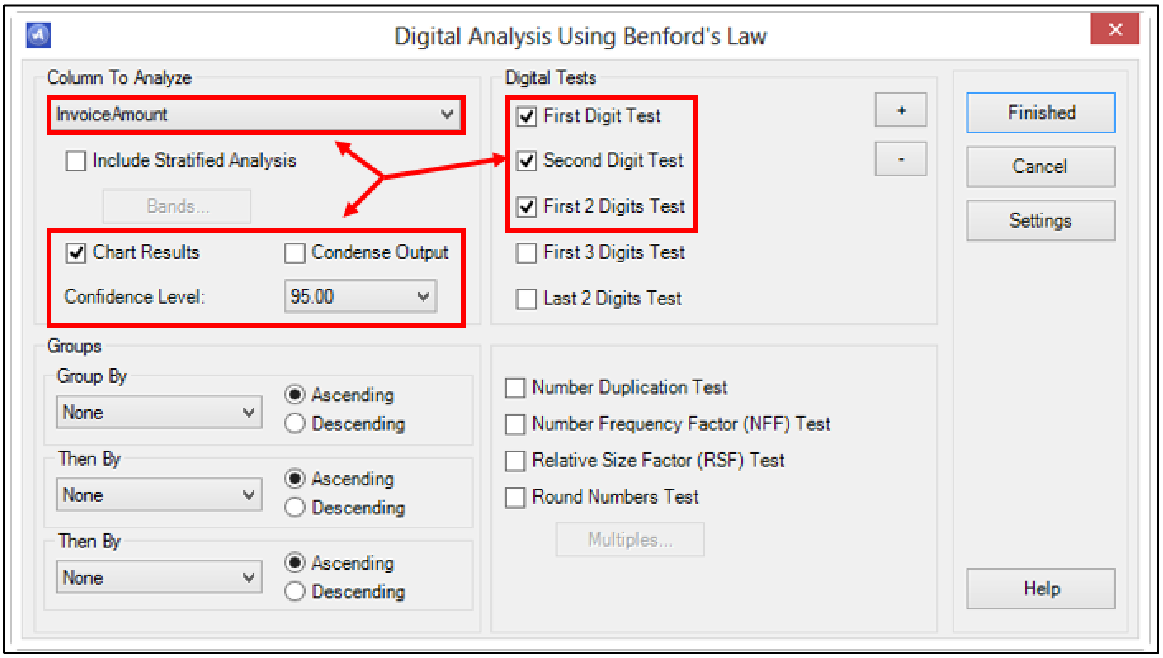


Figure 45 – Setting Options for Benford's Analysis in ActiveData

Note that "**8**" and "**9**" are oversampled significantly, as indicated by the magnitude of the **Z Statistic**. In addition, the results of the **First 2 Digits Test** (not shown) suggest that "**83**," "**88**," "**95**," and "**97**" are significantly oversampled also. Under these circumstances, an auditor should likely examine those transactions where invoice amounts begin with oversampled numbers.

	A	B	C	D	E	F	G	H	I
1	**First Digit Test**	**First Digit**	**Count**	**Proportion**	**Benford Proportion**	**Difference**	**Upper Bound**	**Lower Bound**	**Z Statistic**
2	**Analysis For: InvoiceAmount**	1	472	0.21533	0.30103	0.0857015	0.32046	0.28160	8.72403
3		2	293	0.13367	0.17609	0.0424234	0.19227	0.15992	5.18651
4		3	238	0.10858	0.12494	0.0163621	0.13901	0.11087	2.28452
5		4	230	0.10493	0.09691	0.0080170	0.10952	0.08430	1.23267
6		5	221	0.10082	0.07918	0.0216399	0.09071	0.06765	3.71258
7		6	180	0.08212	0.06695	0.0151700	0.07764	0.05626	2.79903
8		7	127	0.05794	0.05799	0.0000540	0.06800	0.04798	-0.03488
9		8	208	0.09489	0.05115	0.0437380	0.06060	0.04170	9.24648
10		9	223	0.10173	0.04576	0.0559761	0.05473	0.03678	12.49076
11		Total	2,192		**Mean Absolute Difference**	**0.0321202**			
12									

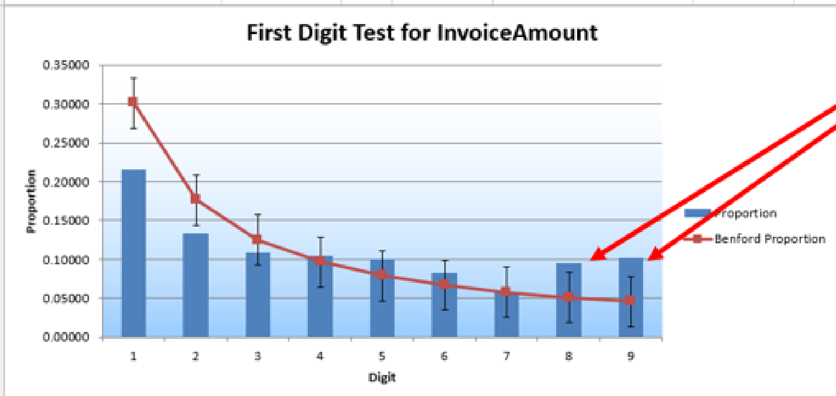


Figure 46 – Results of First Digit Analysis Performed in ActiveData

ActiveData Summary

ActiveData is a widely respected tool for testing and auditing Excel workbooks. The Excel version of the application is available in two editions: Professional and Business. The Professional Edition ($299) targets auditors and accountants, whereas the Business Edition ($149) targets general business users. You can find more information about ActiveData at www.informationactive.com.

Using Power BI For Data Analytics

Chapter 3

Microsoft's Power BI platform empowers auditors and data analysts to analyze data using many of the same measures and calculations as Excel. However, for many, Power BI provides superior options for presenting these calculations and, in some cases, even creating these measures. In this chapter, you will learn how to use Power BI for data analytics.

Learning Objectives

Upon completing this chapter, you should be able to:

- Differentiate between Power BI Desktop and Power BI Pro;
- Utilize Power BI's Quick Insights feature for analyzing data;
- Generate statistical measures in Power BI;
- Identify "outliers" using Power BI;

Power BI Fundamentals

Power BI is a set of Microsoft tools that you can use with or without Excel to quickly and easily summarize large volumes of data. With Power BI, you can create rich, interactive visualizations of your data, drill down on results to see supporting details, and share the data in dashboards. You can use Power BI's capabilities to generate analytics for all data types – financial, sales, manufacturing, human resources, engineering, operations, etc. In addition, Power BI's flexibility allows users to create Key Performance Indicators (KPIs) that can be updated quickly and filtered to spot trends over time, regions, budgets, or goals.

What Are the Components of Power BI?

Microsoft has created *two* parallel development environments for business intelligence and data analytics. As discussed in the preceding chapter, one uses Excel for data manipulation and report development. The other, more modern approach uses the Power BI platform to facilitate business intelligence and data analytics. **Figure 47** provides a visual representation of these environments and their components.

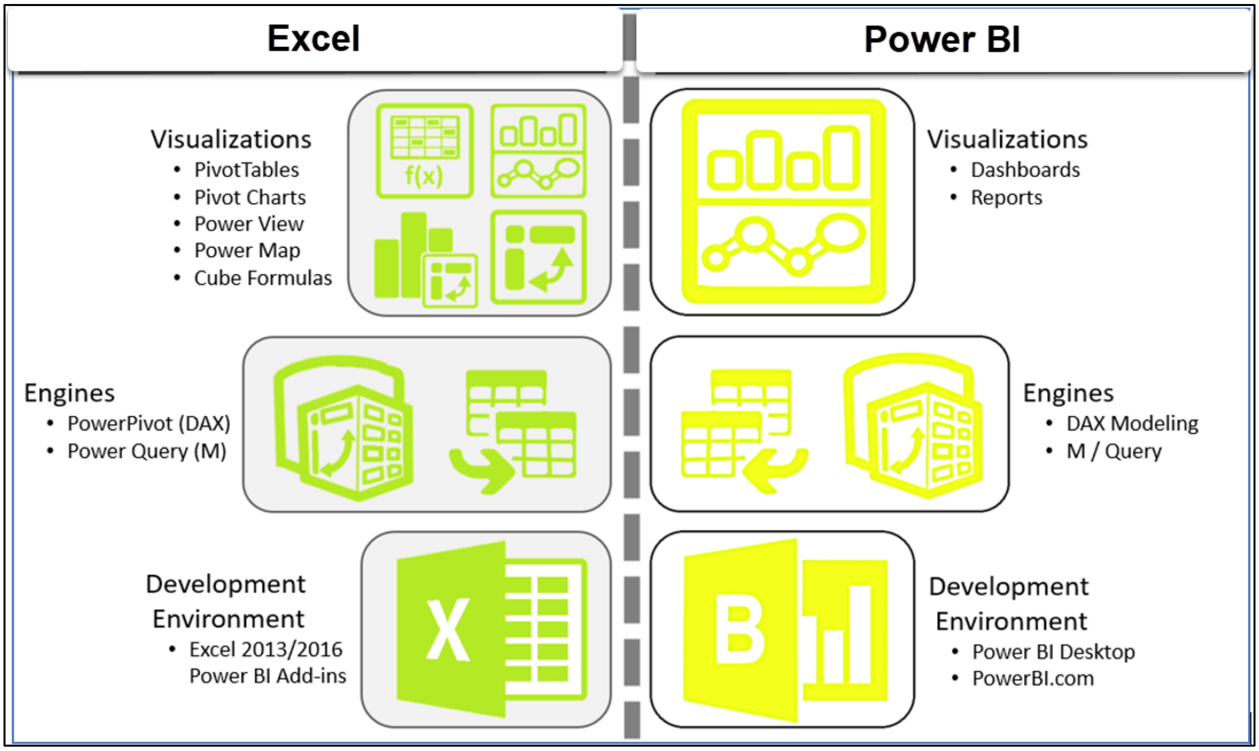


Figure 47 – Power BI Providing Two Parallel Development Environments [1]

The first environment, pictured on the left in Figure 47, works with Excel. More specifically, it works with Excel included with most business-oriented Microsoft 365 subscriptions, with perpetual licenses of Office 2013 and newer, or with the stand-alone version of Excel 2013 or newer.

However, subscribing to Power BI adds functionality not otherwise available with Excel. With a Power BI subscription, users can view and share reports and dashboards through Power BI.com, access BI reports and dashboards on mobile devices, and interact with reports using natural language queries. The second development environment, pictured on the right in Figure 47, uses the Cloud-based data manipulation and dashboard reporting tools of Power BI.com and the free companion application Power BI Desktop. Power BI Desktop addresses many of the issues and limitations of using Excel as a development environment – memory resource constraints associated with 32-bit Excel and large datasets, difficulty publishing Excel workbooks to mobile devices, and the long development cycle of delivering new features to Excel in a rapidly changing marketplace. Further, in this environment, you do not need an Excel license. Freedom from Excel enables Mac users to use Power BI's Cloud-based tools.

Power BI Desktop provides almost identical functionality to the primary components of an Excel-based BI environment – Power Pivot, Power Query, and DAX formulas. As a result, these parallel

[1] Figure adapted from one originally published at *powerpivotpro.com* by Rob Collie, July 2015.

development environments provide analysts, business professionals, and information consultants with exceptional flexibility in meeting the needs of information consumers.

Power BI Versions and Pricing

Microsoft licenses Power BI for use in two primary versions: 1) Power BI Desktop, which is free, and 2) Power BI subscriptions, which begin at $14 per month per user.[2] Power BI subscriptions are web-based services accessible through supported browsers—including Microsoft Edge, Chrome, Safari, or Firefox—for creating, viewing, and sharing dashboards and reports.

Both versions provide all the core functionality. A free Power BI license is sufficient if your report links to data in Excel workbooks, Power BI Desktop, or CSV files, or to built-in content packs. A Pro license is required to create or view a Power BI dashboard or report in the following situations.

1. Data results from a direct query dataset, such as tabular data from SQL Server Analysis Services, Azure SQL Database, Azure SQL Data Warehouse, or Apache Spark for HDInsight.
2. You want to schedule automatic refreshes of your dataset.
3. A dataset connects to on-premises data using a Data Management Gateway.
4. You created your dashboard or report from a custom organizational content pack.
5. A dashboard, report, or dataset resides in a group workspace.
6. A dashboard receives data from a streaming source at a rate greater than 10K rows per hour.

Table 3 identifies the functionality offered by each version.

	Power BI *Free*	**Power BI Pro** *$14 per month*
General		
Create, view, and share your personal dashboards and reports with other Power BI users	√	√
Author content with the Power BI Desktop	√	√
Explore data with Natural Language Queries	√	√

[2] Microsoft also makes a server-based version of Power BI available. This product is typically used by larger organizations with 500 or more users. Coverage of the server-based solution is beyond the scope of this session.

	Power BI *Free*	**Power BI Pro** *$14 per month*
Access your dashboards on mobile devices using native apps for iOS, Windows, and Android	√	√
Consume curated content packs for services like Dynamics Salesforce and Google Analytics	√	√
Import data and reports from Excel CSV and Power BI Desktop files	√	√
Data Refresh		
Consume content that is scheduled to refresh	Daily	Hourly
Consume streaming data in your dashboards and reports	10K rows/hour	1M rows/ hour
Consume live data sources with full interactivity		√
Access on-premises data using the Data Connectivity Gateways (Personal and Data Management)		√
Collaboration		
Collaborate with your team using Office 365 Groups in Power BI		√
Create, publish, and view organizational content packs		√
Manage access control and sharing through Active Directory groups		√
Share data queries through the Data Catalog		√

Table 3 - Comparing Power BI to Power BI Pro

Power BI Desktop is a free application for developing Power BI datasets and reports on a local PC. It comes in 32-bit (x86) and 64-bit (x64) versions, which require users to match the architecture to the task at hand. If you plan on using Power BI Desktop with 32-bit applications or ODBC drivers, install the 32-bit version. The 64-bit version will deliver the best performance and overcome common memory management issues when performing sophisticated analysis of extensive datasets. A paid license is not required to use the desktop application, but is required to publish or share a dashboard or report on Power BI.com. You can save and exchange Power BI

Desktop files (*.pbix*) without a paid license. The only requirement is that all users install the Power BI Desktop application. However, remember that reports created with the desktop application may result in large file sizes. Sharing files via a network location will probably work fine in most cases, but large files will likely prevent sharing via email. Power BI Desktop is available for free download from the following site:

https://www.microsoft.com/en-us/power-platform/products/power-bi/desktop.

Power BI Desktop Interface

Refer to **Figure 48** as we acquaint ourselves with the application's interface. The familiar Ribbon, from which users can enter data or run queries, is at the top of the window. Along the left is a navigation bar containing buttons for the three work views – **Report**, **Data**, and **Model**.

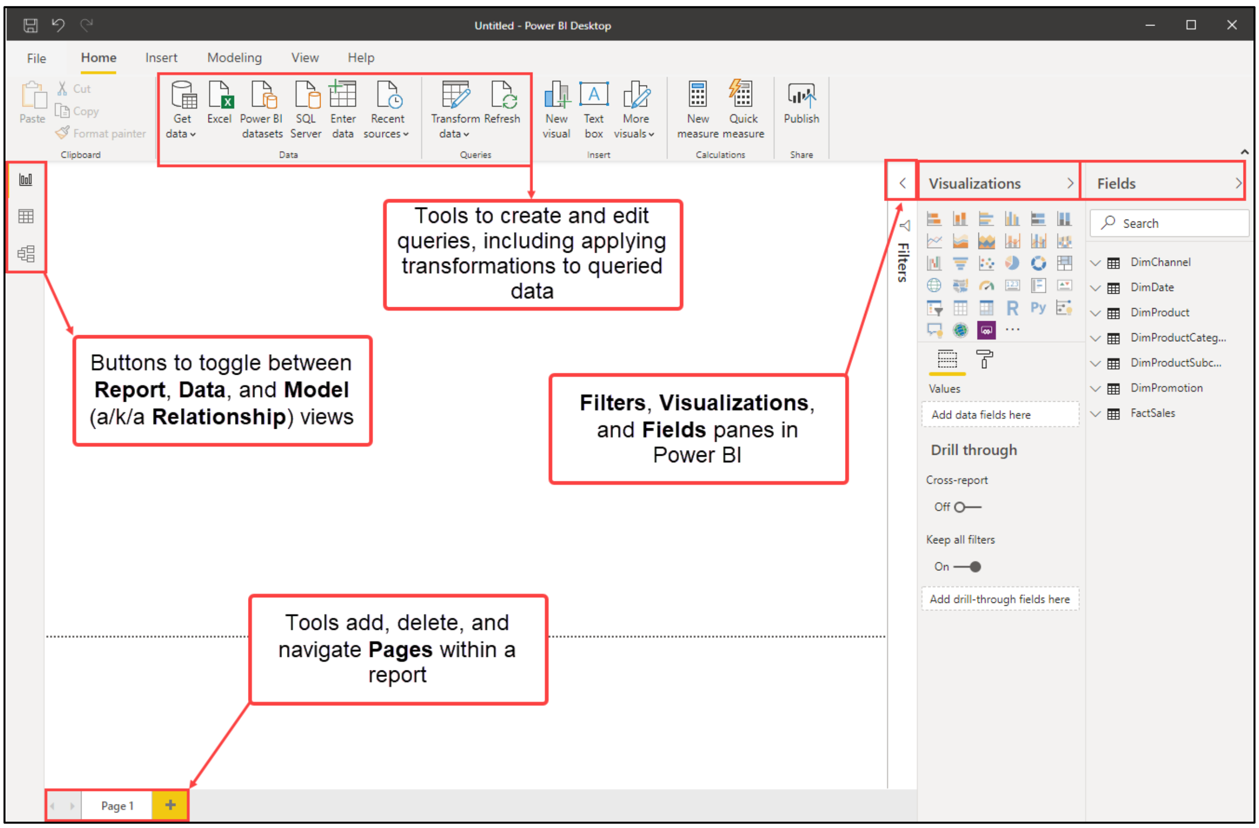


Figure 48 – Power BI Desktop GUI

On the right are the **Fields** list and the **Visualizations** pane. The latter contains a gallery for applying and formatting data visualizations (tiles), as well as the Filters area for filtering reports and report elements. You can dock the Fields list and Visualizations pane along the right margin to expand the main viewing area. You can use buttons at the bottom to add or navigate report pages.

Building the Dataset

To begin building a new report, click **Get Data, More…** on the **Home** tab. In the **Get Data** window, select **Database** in the navigation pane on the left and choose **Access Database** in the list on the right. Next, click **Connect**, browse to and select the **ContosoSales** database, and click **Open**.

In the **Navigator** window, check the desired tables in the list. In this case, check **DimChannel**, **DimDate**, **DimProduct**, **DimProductSubcategory**, **DimPromotion**, and **FactSales**, as shown in **Figure 49**. You can add or remove columns and formats, or filter the data imported from individual tables. To do so, select the table of interest and click **Transform Data** to open the Query Editor. In this case, click **Load** to import the data into the dataset. The imported tables will now be visible in the **Data** view.

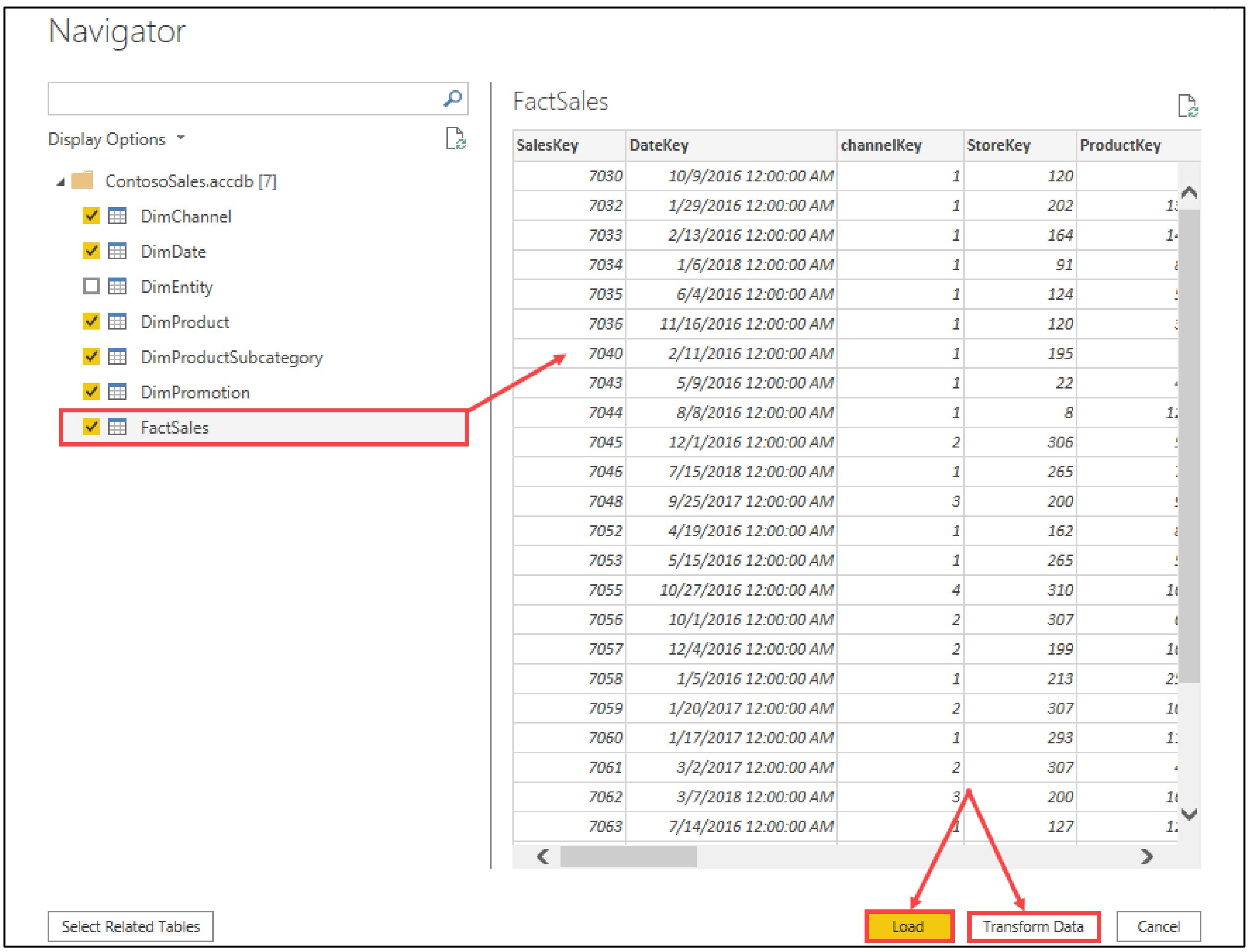


Figure 49 – Importing Tables into the Power BI Desktop Dataset

Continue building your dataset by importing data from two Excel workbooks—DimStores and DimGeography—containing information about the stores and locations where the company earns sales revenue. Click **Get Data**, **Excel**, browse to and select the **DimStores** workbook, and click **Open** from the Ribbon. Select the **DimStores** table in the **Navigator** window and click **Load**. Repeat the process to import the **DimGeography** workbook.

Defining Relationships Using Design View

Power BI Desktop attempts to recognize and maintain relationships if you import multiple related tables from the same data source. Additionally, Power BI Desktop attempts to identify and relate tables from *different* data sources by matching column names and data types. In cases where relationships are not recognized or maintained, you may establish them manually in the **Relationships** view, shown in **Figure 50**, using a simple click-and-drag process.

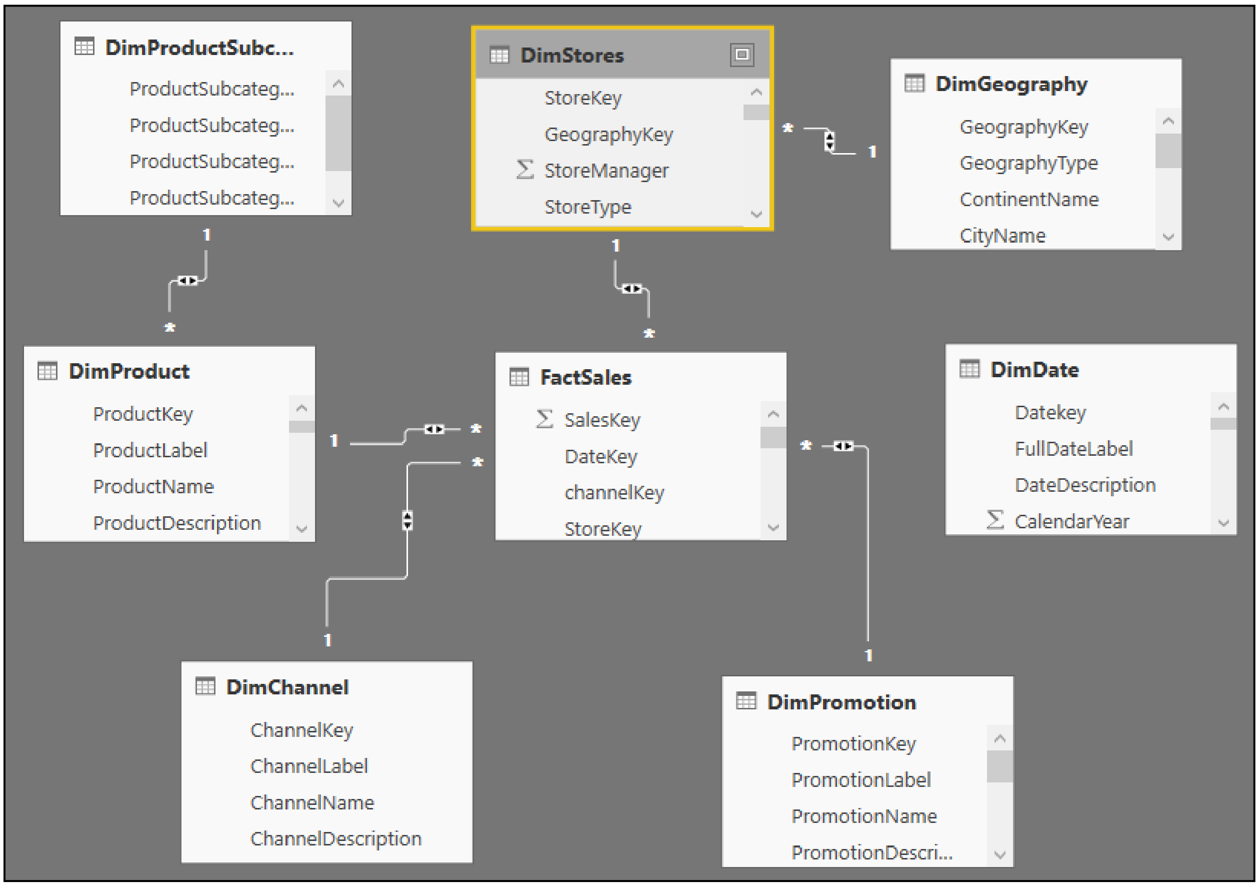


Figure 50 – The Relationships View in Power BI Desktop

In this case, Power BI Desktop recognized and established the relationships between the two imported Excel workbooks and the tables imported from Access. However, the dataset did not recognize the relationship between the **FactSales** and DimDate tables. To create a relationship manually, drag **DateKey** from the **DimDate** table and drop it on **DateKey** in the **FactSales** table. Once you establish a relationship, double-click a join line to examine or edit the relationship. The process creates an *inner join* between the related tables by default. Also, note that the diagram in the **Relationships** view indicates the relationship type (one-to-one, one-to-many, or many-to-one) and the cross-filter direction.

Building Matrix Reports

You can create a Power BI Desktop report after importing data and relating the tables. Power BI reports consist of a dataset and one or more visualizations. Our report will contain a **Matrix** visualization of sales revenue summarized by state. Note that a Matrix provides functionality similar to an Excel-based PivotTable.

In the **Fields** list, expand the **DimGeography** table and check **StateProvinceName** to add a list of states and provinces to the canvas. Power BI adds it to the canvas as a two-dimensional **Map** tile by default. With the map tile selected, click **Matrix** in the Visualization pane's visualization gallery to convert it to a matrix. Note that the list contains states and provinces from around the world. To filter tiles on the page, drag the **RegionCountryName** field to the **Page level filters** box in the **Filters** area of the **Visualizations** pane. Expand the list of regions and countries, and then check **United States** to filter the list of states and provinces to just those in the United States, as shown in **Figure 51**. Note that you may use fields as visual (individual tiles), page (affecting a specific page), or report (affecting all pages in a report) filters. Make sure to drop field filters into the appropriate filter box.

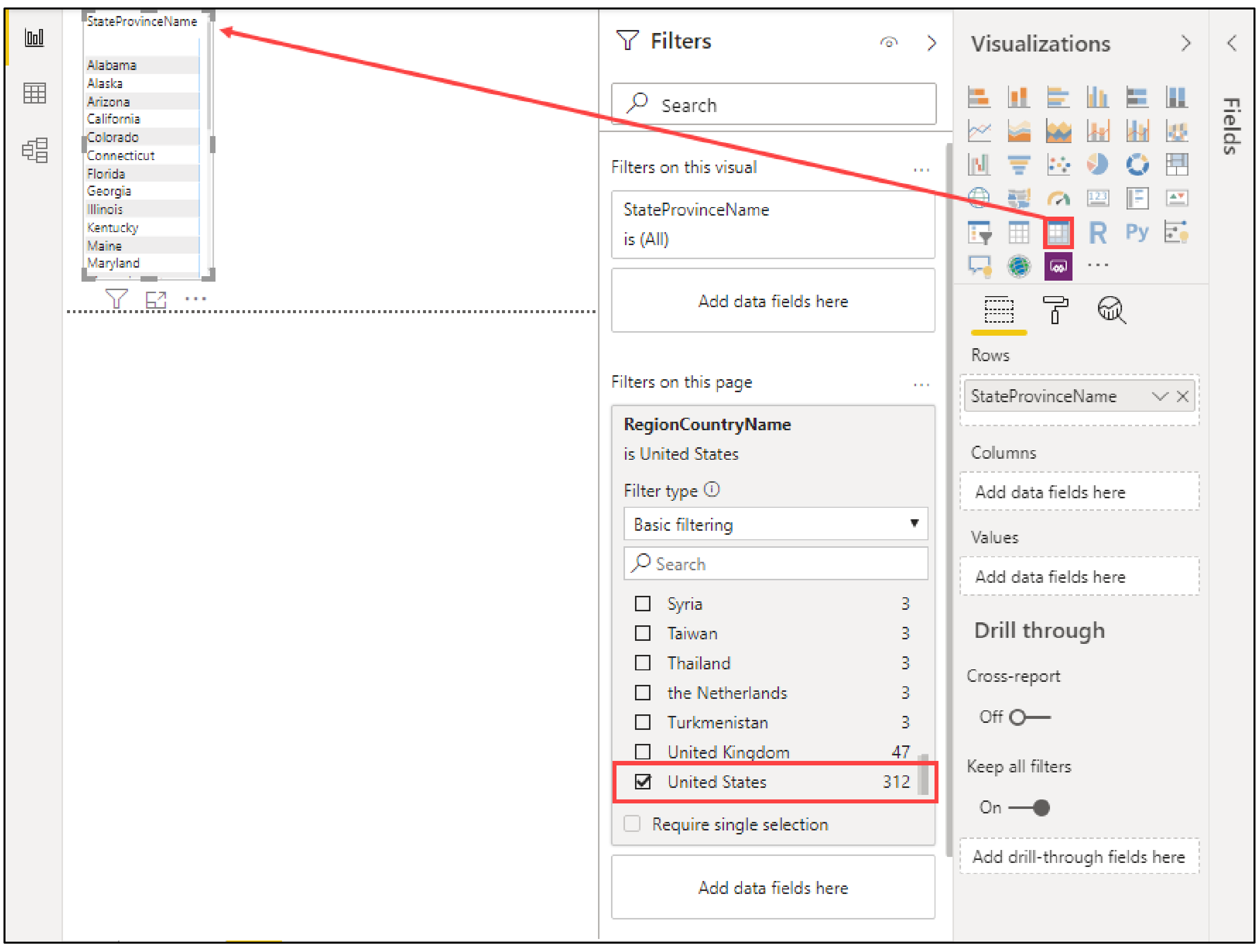


Figure 51 – Adding and Filtering a Field in a Report as a Matrix Tile

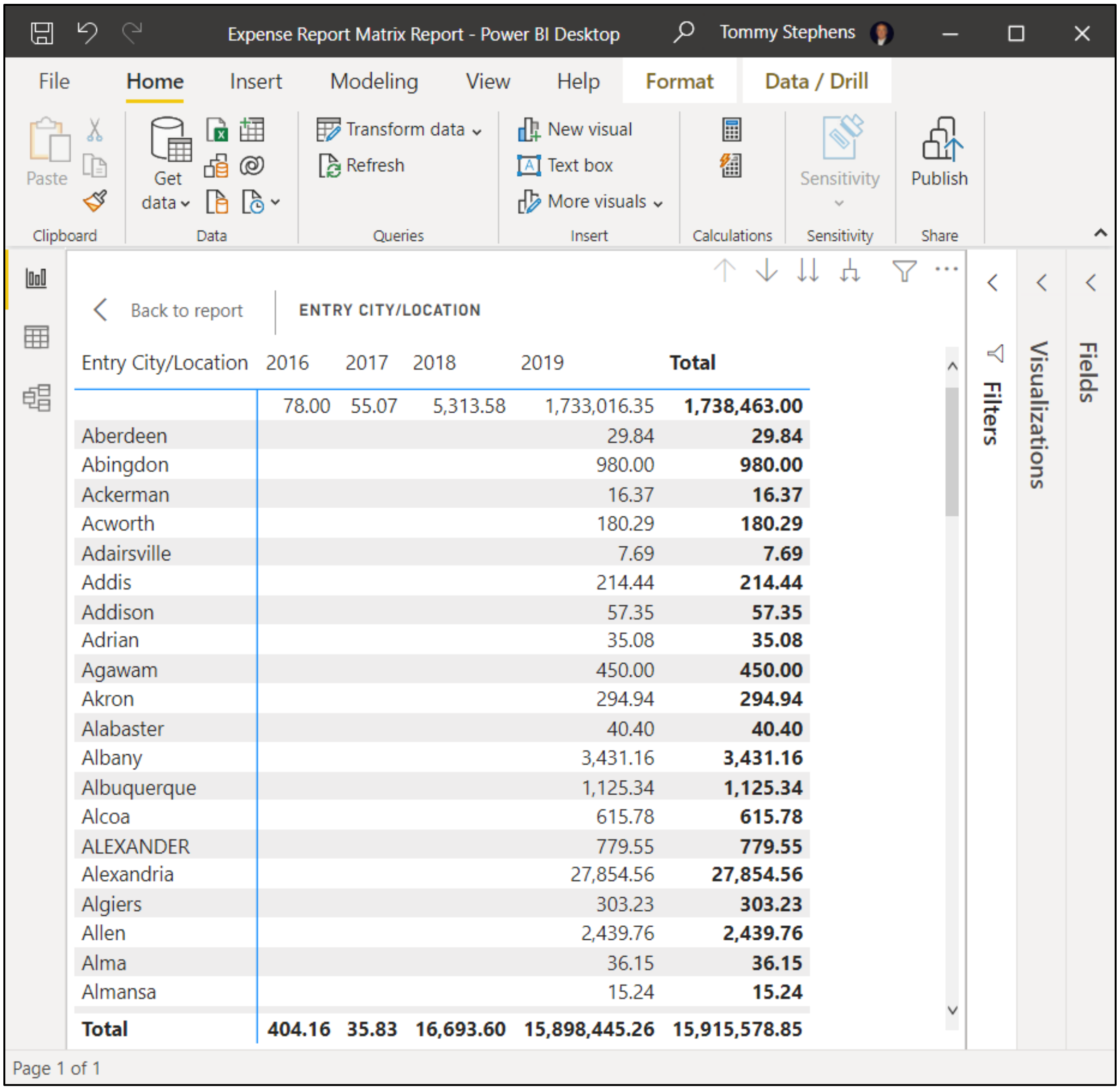
Because we changed the tile to a Matrix, the **Visualizations** pane includes **Rows**, **Columns**, and **Values** drop-downs, similar to the drop-downs in the PivotTable Field List in Excel. Drag the **Sales Amount** field from the **FactSales** table to the **Values** drop area, the **FiscalYearLabel** field from the **DimDate** table to the **Columns** drop area, and the **StoreName** field from the **DimStores** table to the **Rows** drop area below **StateProvinceName**. Click and drag the tile border to resize it and display all columns. These actions complete the base Matrix visualization, which resembles and functions like an Excel PivotTable, but lacks drill-down capability.

Now, let's complete the report. If the **SalesAmount** is incorrectly formatted, click its field name in the **Fields** list. Next, on the **Column Tools** contextual tab of the Ribbon, click the **Dollar Sign ($)** to apply the accounting number format with dollar signs and use the spinner control to enter two decimal places. Next, select **Sum** as the **Default Summarization**. If the columns in the table are not wide enough to display all the data, click and drag in the column headings area to widen or narrow the columns accordingly. Note that there are no visual cues regarding where the column boundaries reside. Move your mouse in the column heading area until the double-arrow cursor appears, and then click and drag the column border to the desired position. Note also that the value column widths work in unison. Adjusting the width of one value column causes all the value columns to change to the same width. To sort on a column, click its heading, with each additional click reversing the sort's order.

Quick Insights

With a fundamental knowledge of the Power BI platform and how to create a Power BI report, let us now focus on generating data analytics using Power BI. For those with a Power BI Pro subscription, the tool's Quick Insights feature offers an excellent way to quickly and easily gain insights into your data.

To generate Quick Insights into your data, create a Power BI report. In this example, we will use the sample matrix report pictured in **Figure 52**.



Entry City/Location	2016	2017	2018	2019	Total
	78.00	55.07	5,313.58	1,733,016.35	**1,738,463.00**
Aberdeen				29.84	**29.84**
Abingdon				980.00	**980.00**
Ackerman				16.37	**16.37**
Acworth				180.29	**180.29**
Adairsville				7.69	**7.69**
Addis				214.44	**214.44**
Addison				57.35	**57.35**
Adrian				35.08	**35.08**
Agawam				450.00	**450.00**
Akron				294.94	**294.94**
Alabaster				40.40	**40.40**
Albany				3,431.16	**3,431.16**
Albuquerque				1,125.34	**1,125.34**
Alcoa				615.78	**615.78**
ALEXANDER				779.55	**779.55**
Alexandria				27,854.56	**27,854.56**
Algiers				303.23	**303.23**
Allen				2,439.76	**2,439.76**
Alma				36.15	**36.15**
Almansa				15.24	**15.24**
Total	**404.16**	**35.83**	**16,693.60**	**15,898,445.26**	**15,915,578.85**

Figure 52 – Sample Matrix Report

Next, publish the report to your Power BI Pro subscription service. Upon doing so, access your workspace in your Power BI subscription, select the report, and choose **Quick Insights**, as shown in **Figure 53**.

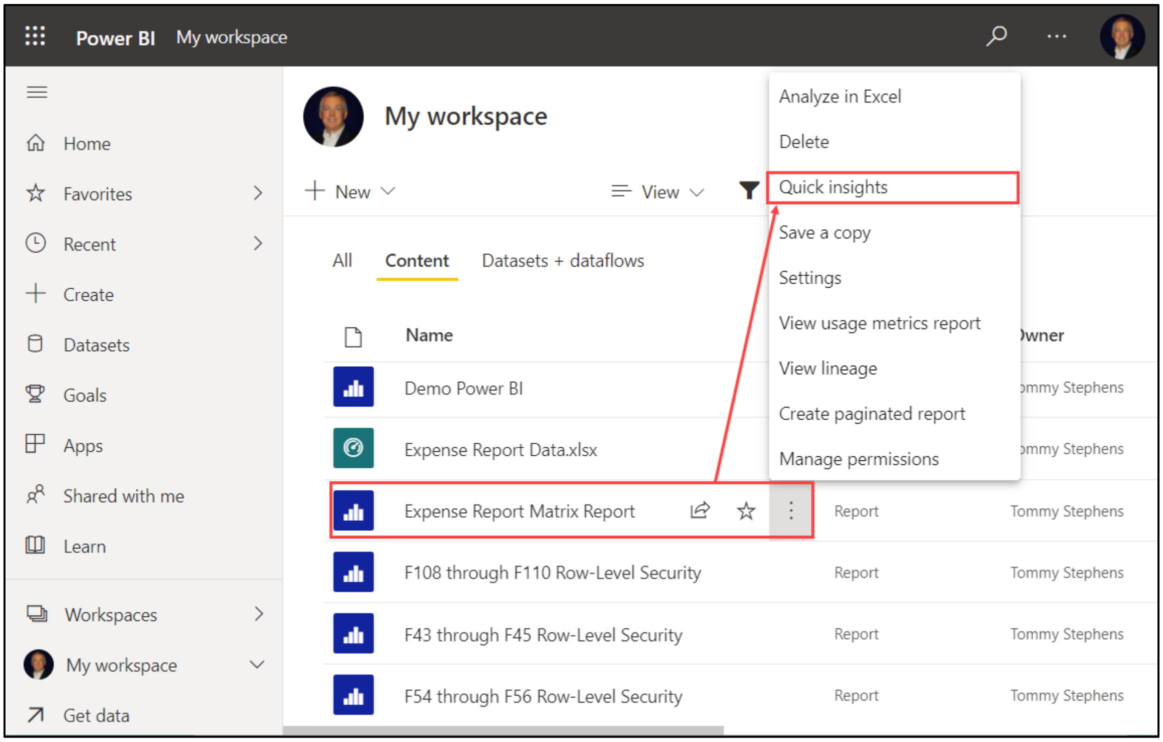


Figure 53 - Generating Quick Insights in Power BI

After analyzing your data, Power BI responds with a set of analyses similar to that pictured in **Figure 54**. Of course, the volume and nature of the insights provided will vary depending on the data analyzed. Nonetheless, for those with a Power BI subscription, using Quick Insights is a best practice for high-level data analytics to discover trends, anomalies, and other key insights from your data.

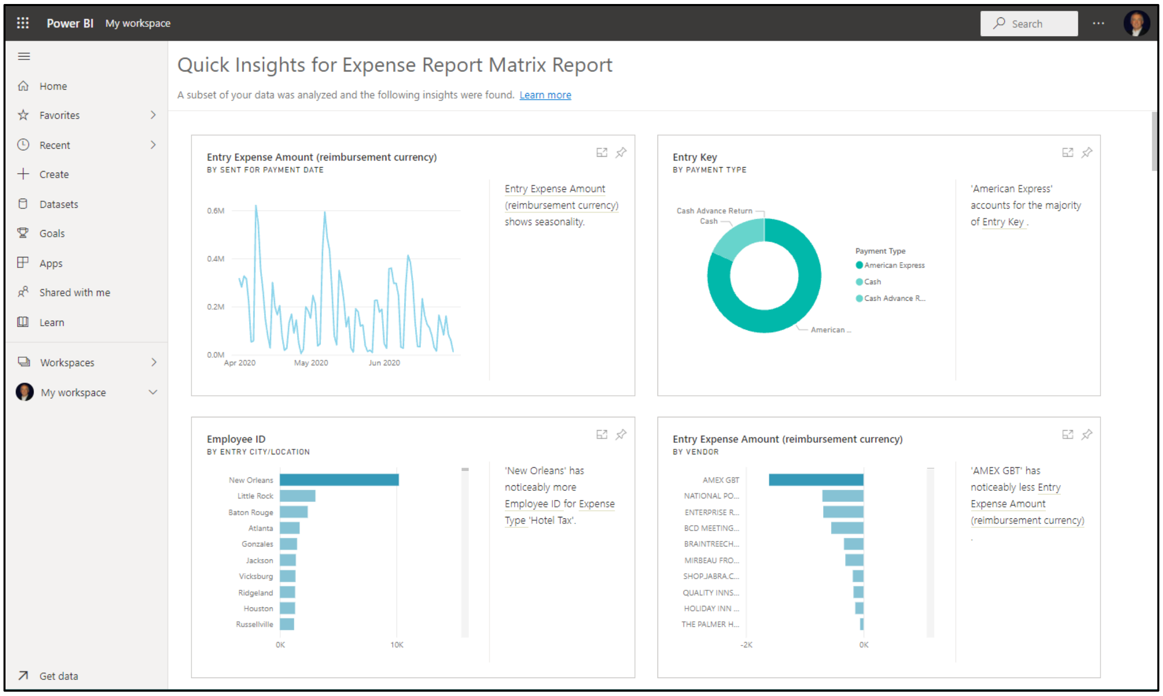


Figure 54 - Sample Quick Insights in Power BI

Statistical Analysis With Power BI

Power BI also offers numerous data analytics options that do not require access to a paid Power BI subscription. Among these options are generating statistical measures in the desktop application.

By default, when building a report in Power BI, the application creates a summing calculation of any field you add as a Value. However, you can change the measure to any of the following:

- Average
- Minimum
- Maximum
- Count (Distinct)
- Count
- Standard Deviation
- Variance
- Median

To do so, as shown in **Figure 55**, click the **Values** field and choose your desired summarizing calculation. Further, you can add the summarizing field to your report multiple times and select a different measure for each field instance.

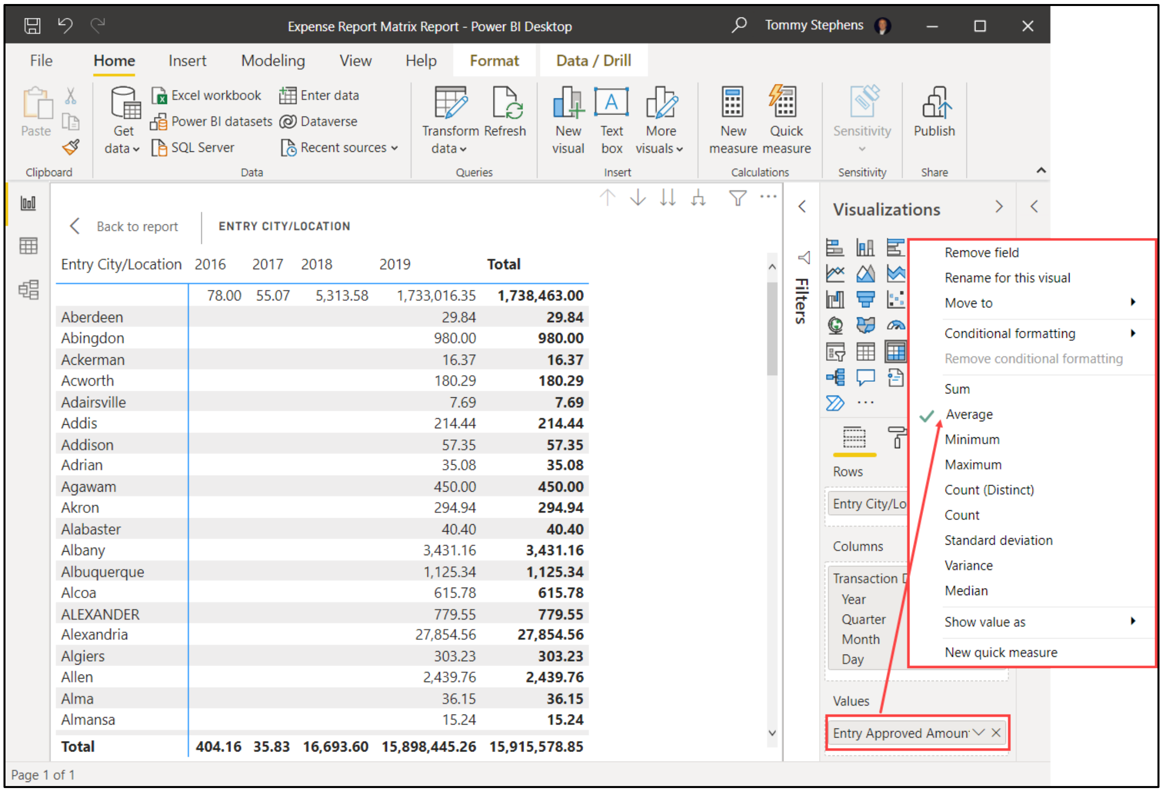


Figure 55 - Changing the Summarization Measure in Power BI Desktop

In addition to changing the summarizing calculations generated by Power BI, you can change how your report presents these calculations. For example, returning to Figure 55, clicking **Show value as** provides the following presentation options.

- Percent of grand total,
- Percent of column total, and
- Percent of row total.

To illustrate, changing the presentation to **Percent of Grand Total** updates the visualization to that shown in **Figure 56**.

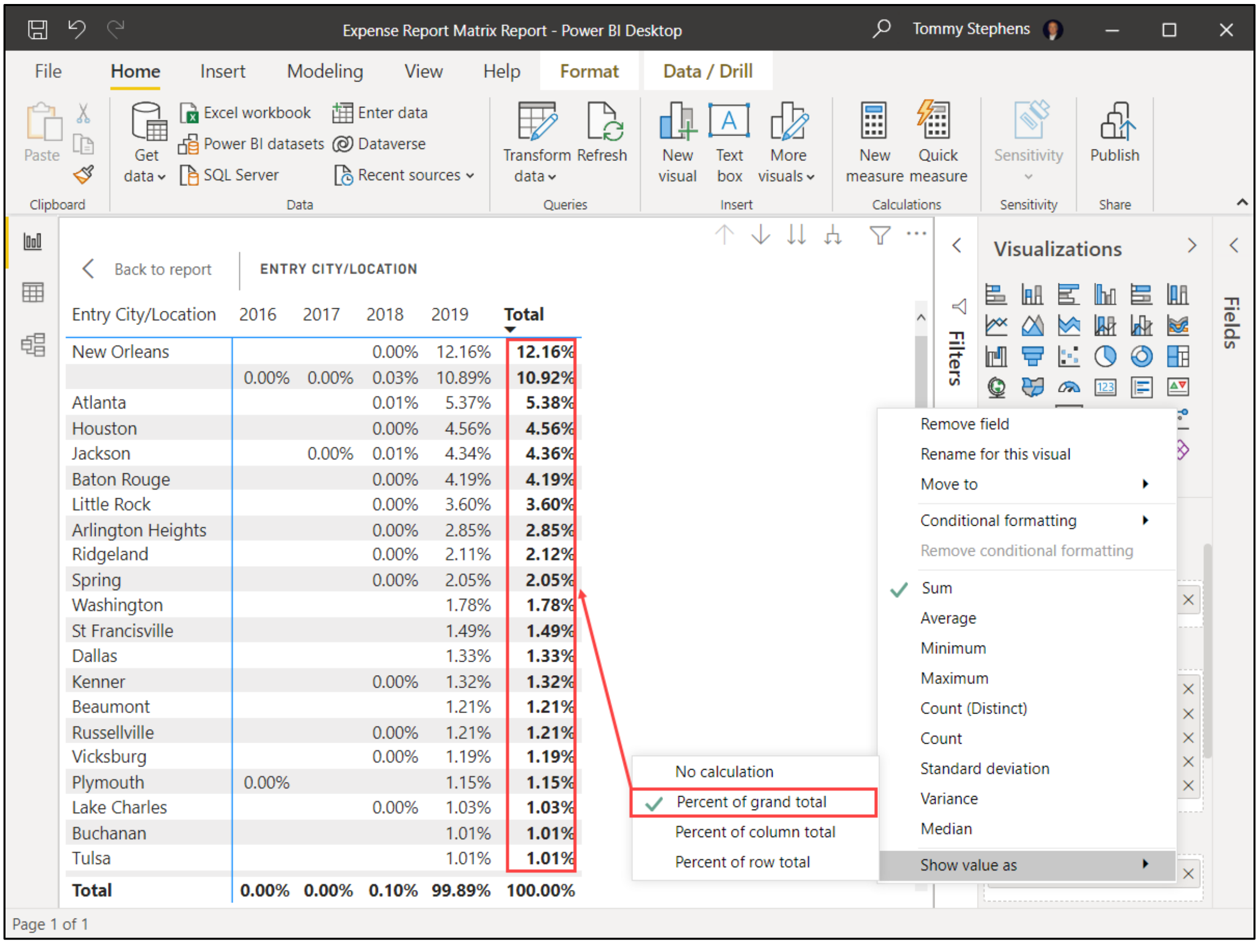


Figure 56 - Displaying the Summarized Data as a Percent of the Grand Total

Grouping Data and Bin Analysis

Previously, you learned how to group Excel data into multiple "bins" and use the bins to perform analyses such as those related to Benford's Law. Power BI Desktop also contains options for grouping data.

While pressing the **CTRL** key, click the data points you wish to group in Power BI. With the desired data points selected, right-click and choose **Group data**, as shown in **Figure 58**. You can add other groups as necessary. In addition, you can edit groups by accessing them from the **Fields** pane and selecting **Edit Groups**.

You can also generate data bins in Power BI, like creating data groups. In this example, the data analyst wants to summarize all transactional data in a dataset into ten groupings. Therefore, the analyst can create the groups to do so, as shown in **Figure 57**.

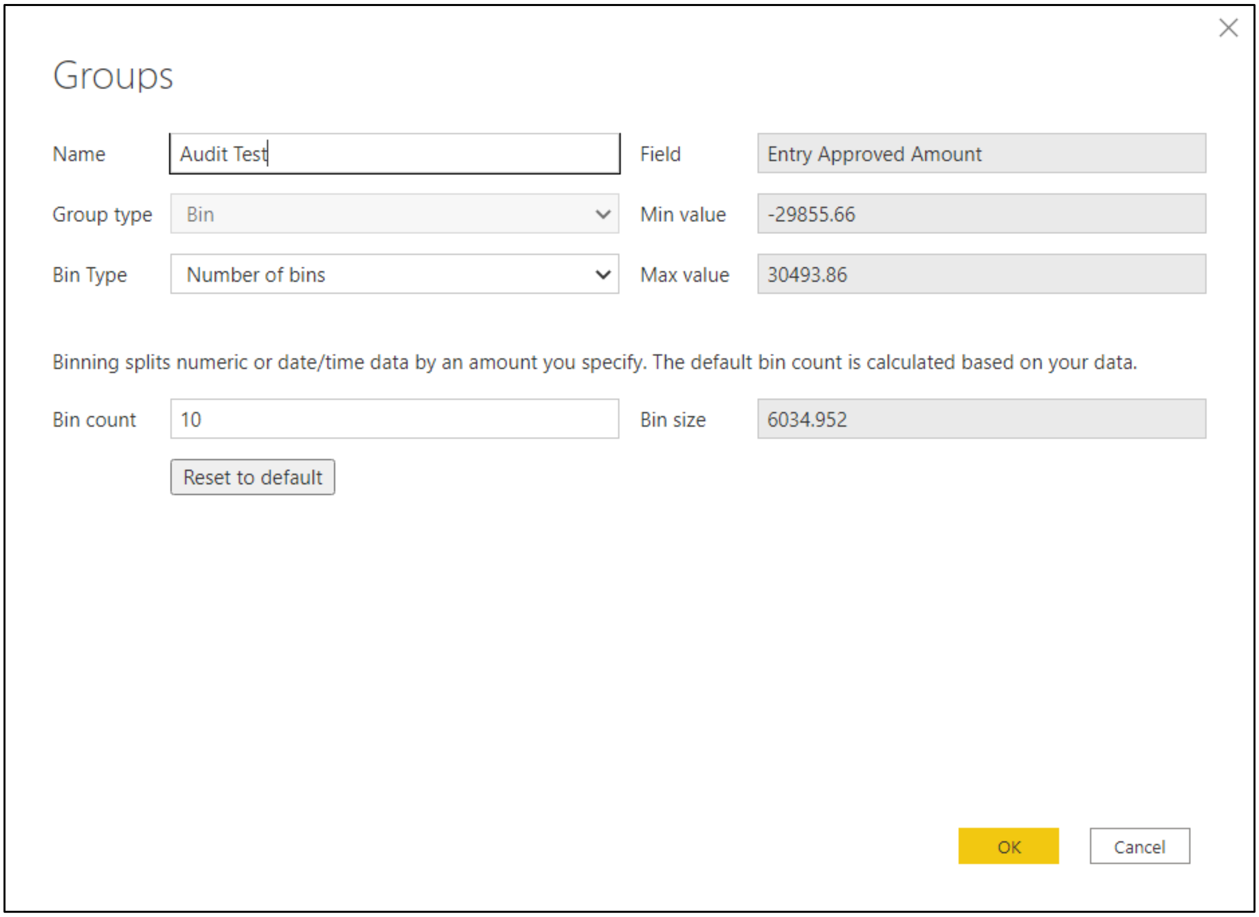


Figure 57 - Creating Data Bins in Power BI

The analyst can then create a Power BI report that groups the data by the bins created.

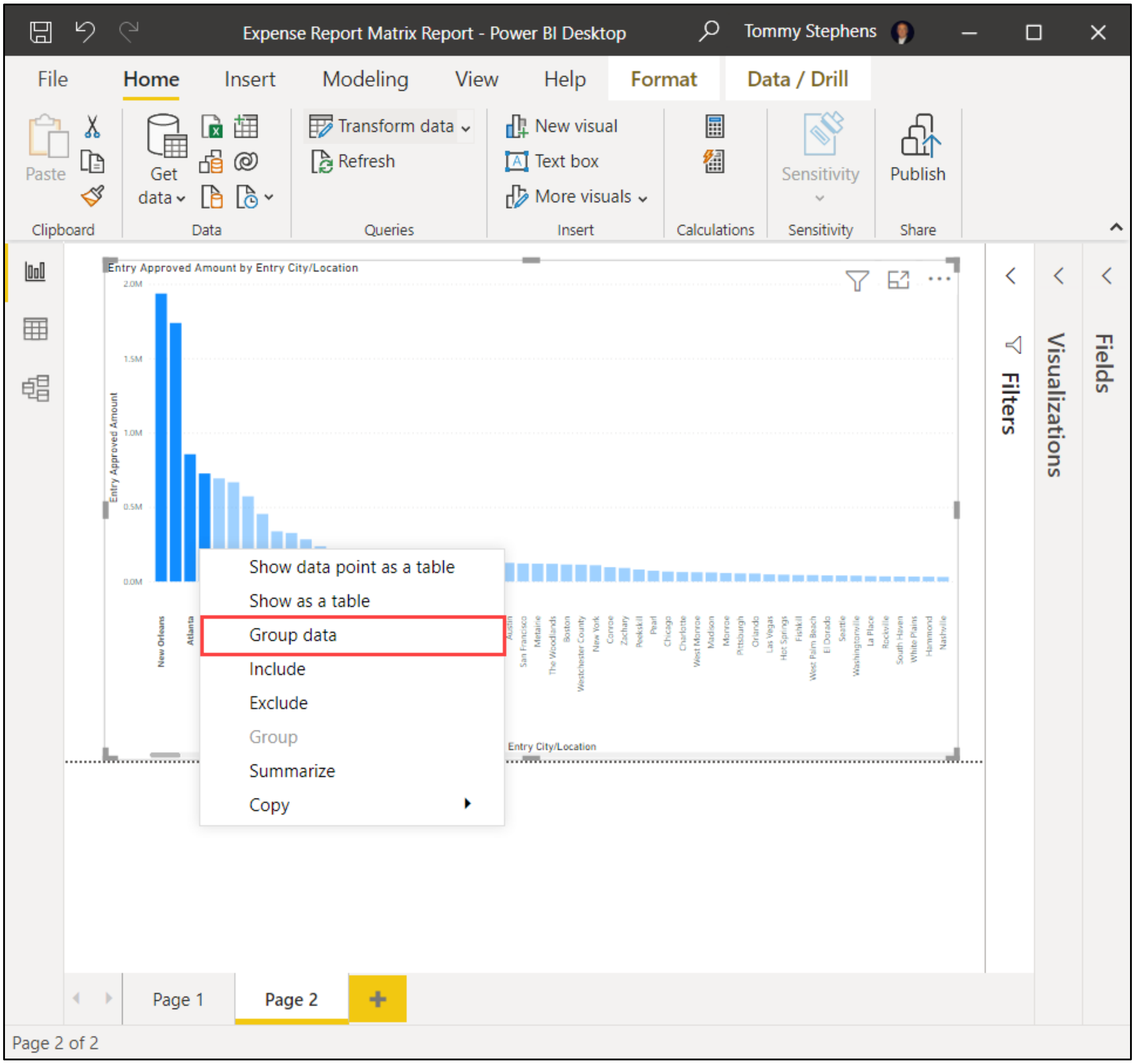


Figure 58 - Grouping Data in Power BI

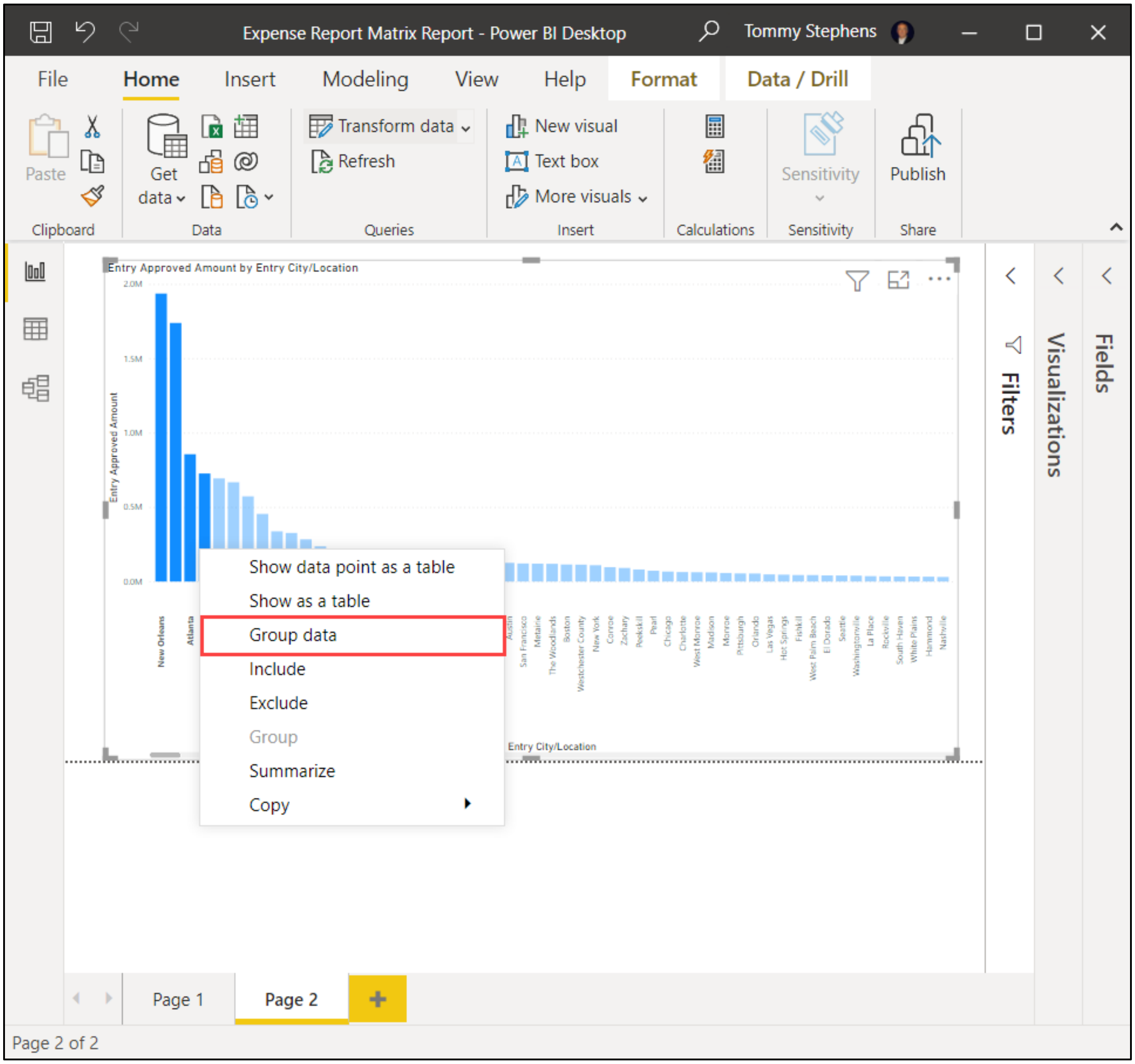


Figure 59 - Creating Groups in Power BI Desktop

Identifying Outliers

Identifying **outliers** in datasets is an essential consideration in data analytics. Outliers are data points that are out of line with expectations. Often, these expectations emanate from historical data or projected data, such as budgets.

In Power BI, you can use two methods to identify outliers. The first involves creating formulas based on Data Analysis Expressions (DAX). DAX is a collection of approximately 300 functions in Power BI, Power Pivot, and SQL Analysis Services. Some of these functions duplicate tools you can use in Excel, but most DAX functions have no counterparts in Excel. Thus, using DAX-based formulas to identify outliers and create other calculations could require a modest to significant effort to learn more about DAX – a task beyond this course's scope.

On the other hand, a simple **scatter chart** in Power BI can be handy for identifying outliers, as it depicts the relationship between two numerical values and is very easy to create. For example, the scatter chart illustrated in **Figure 60** identifies the relationship between Discounts, Returns, and Total Sales by product.

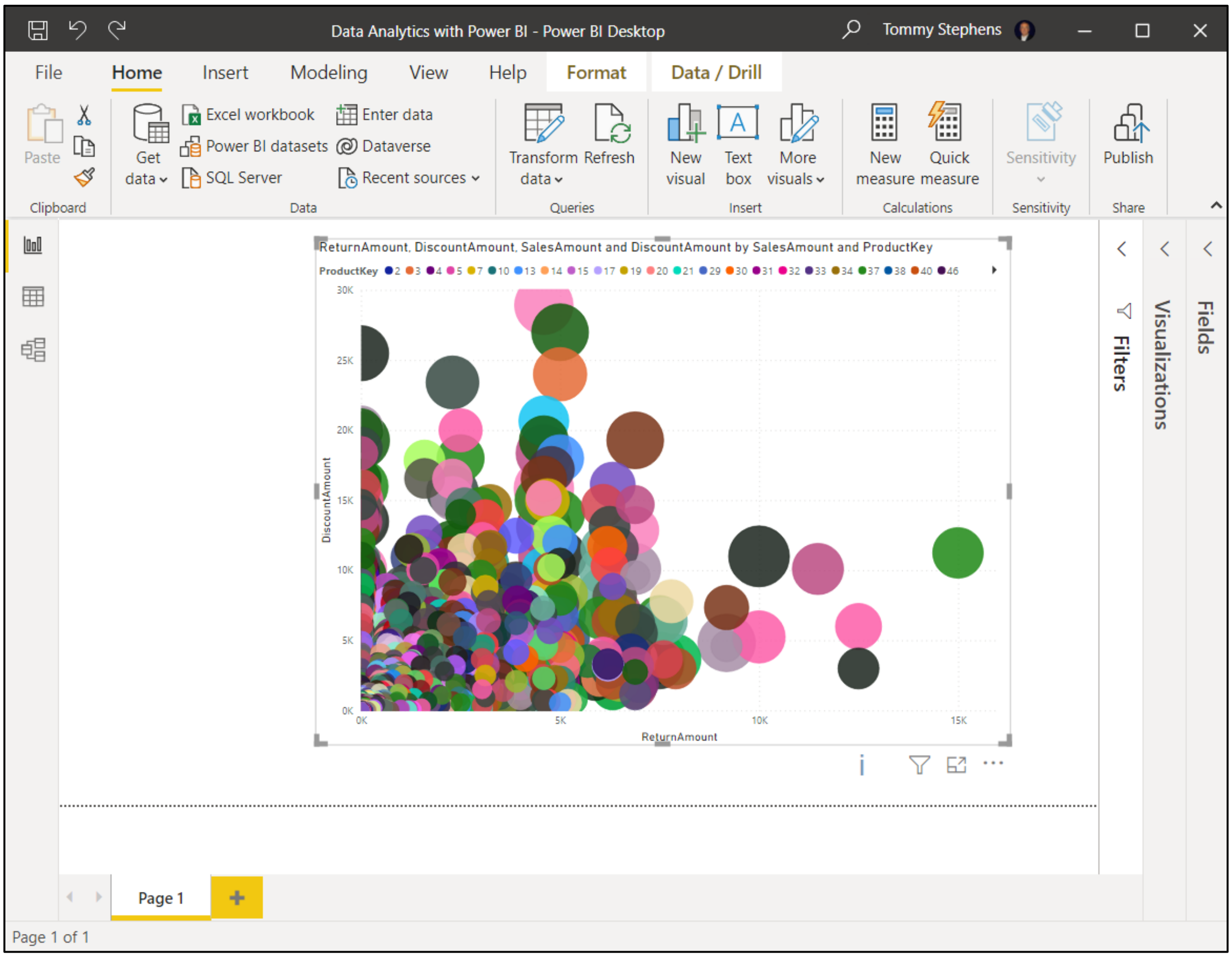


Figure 60 - Sample Scatter Chart in Power BI

One of the strengths of Scatter Plots – and other Power BI visualizations – is the option to drill into the details presented by the chart. For instance, right-clicking on any data point in the scatter plot and choosing the **Show as Table** option adds what is effectively a PivotTable to the report, as shown in **Figure 61**. Of course, adding the Table allows you to see and analyze the data in greater detail.

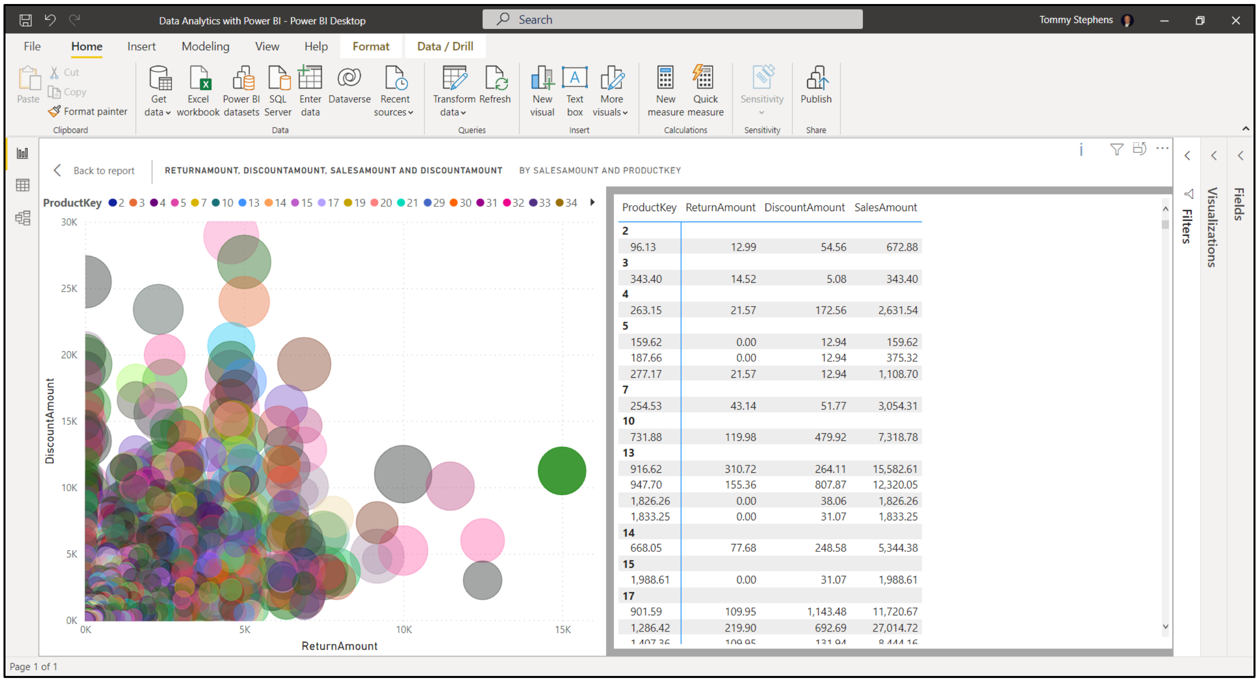


Figure 61 – Scatter Plot Chart with Accompanying Table Report

Power BI's Analyze Feature

Power BI's **Analyze** feature can be handy when identifying why a data distribution depicted in a Power BI visualization appears the way it does. Analyze helps you run automated analyses of your data. To use the feature, right-click on any data point in your visualization. Then, depending on the visualization and the insights you desire, choose between "*Explain the increase*" or "*Find where this distribution is different*." Upon doing so, Power BI responds with a visualization similar to that pictured in **Figure 62**. In the example presented, this analytical tool helps the reader to understand which filter causes the approved amount of employee expense reports to change the most.

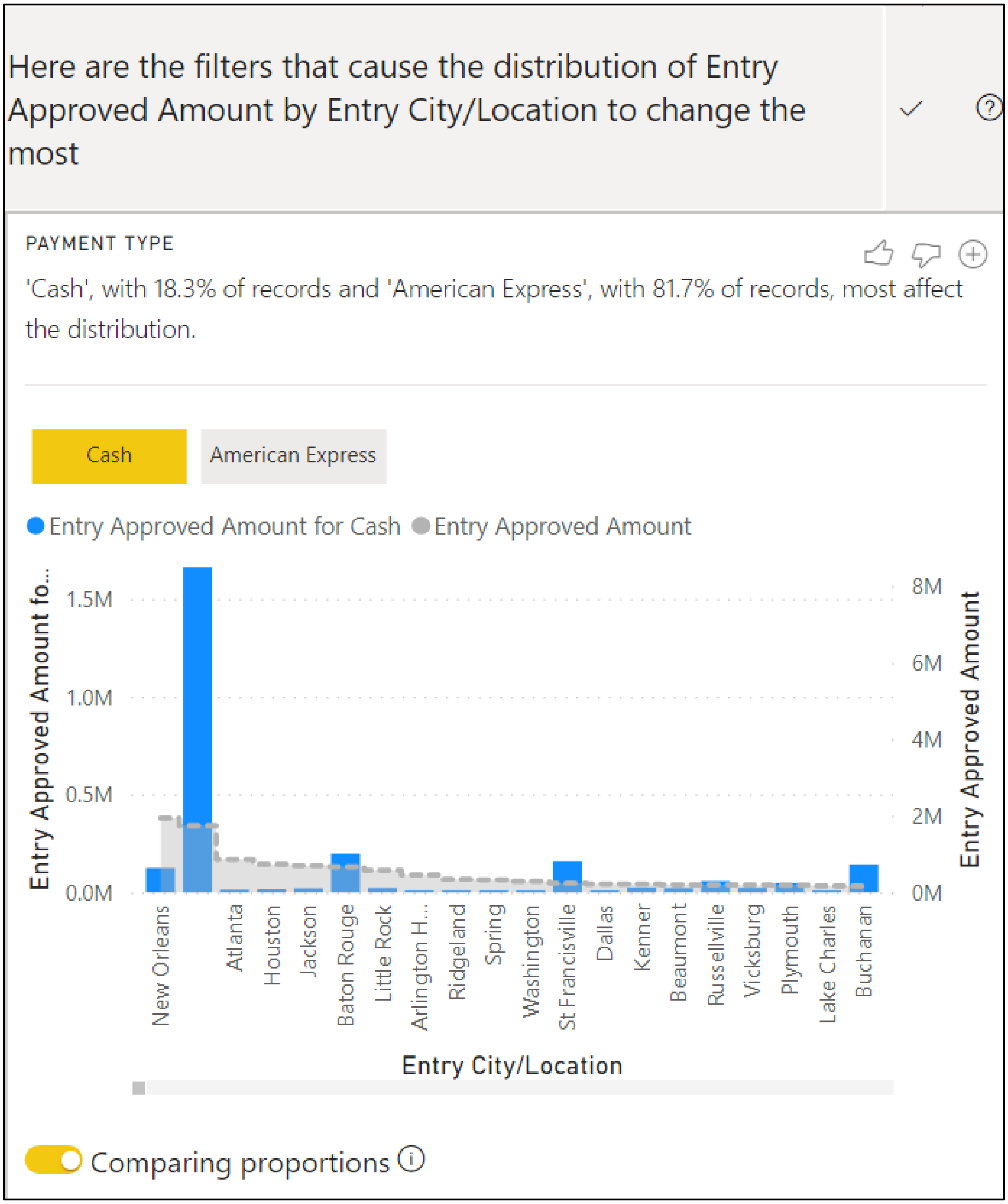


Figure 62 - Power BI's Analyze Data Feature

Forecasting with Power BI

Power BI provides a limited but effective set of forecasting tools that you can use to forecast future results. These options focus on using linear regression models to produce forecasts from date-oriented data. To that point, consider the simple Power BI-based line chart in **Figure 63**. The graph illustrates the number of passengers cleared through airport security check stations in the

United States from January 1 through October 9, 2021. An analyst would like to use this data to forecast the number of passengers likely to board commercial flights during the next 45 days.

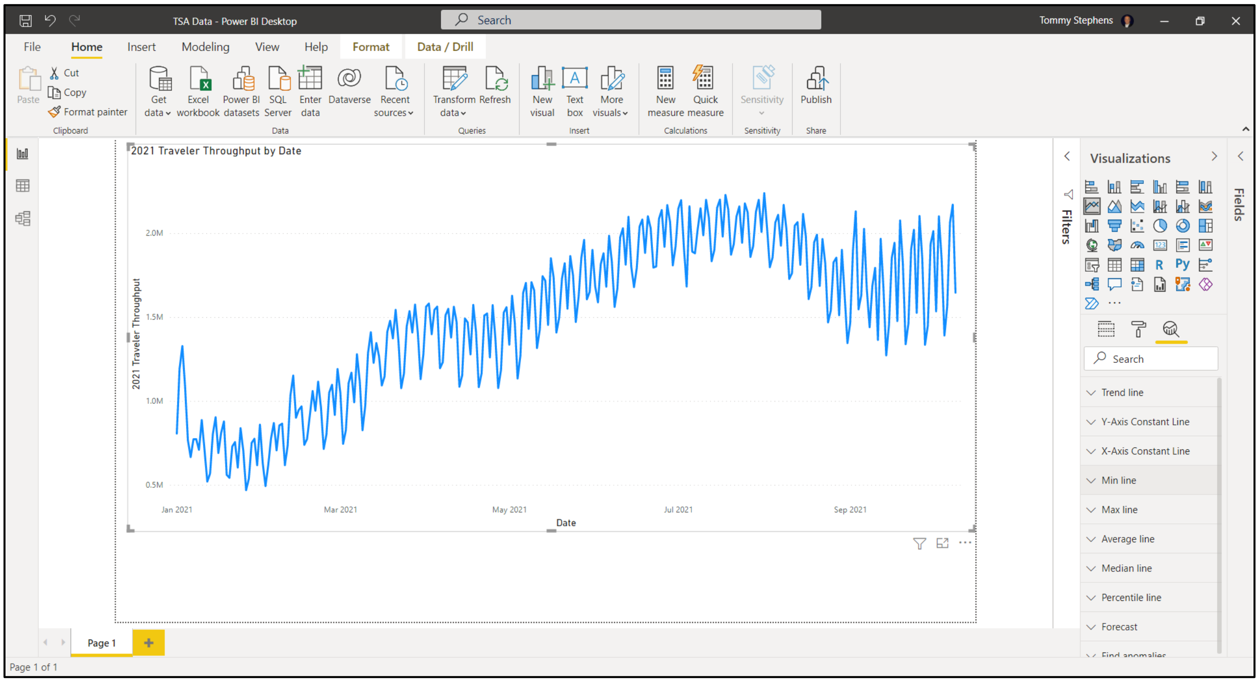


Figure 63 - Simple Power BI Line Chart Based on TSA Passenger Screening Data

The analyst clicks **Forecast**, **Add**, **45 Days**, and **Apply** to create a forecast, as shown in **Figure 64**. Upon doing so, Power BI generates a projection based on the inputs specified. Note that one of the inputs is the ***Confidence interval***. The higher the amount you choose for this setting, the greater the forecast variance; conversely, the lower the amount you select, the narrower the variance. A red box outlines the forecasted data in the example below to make it easier to identify.

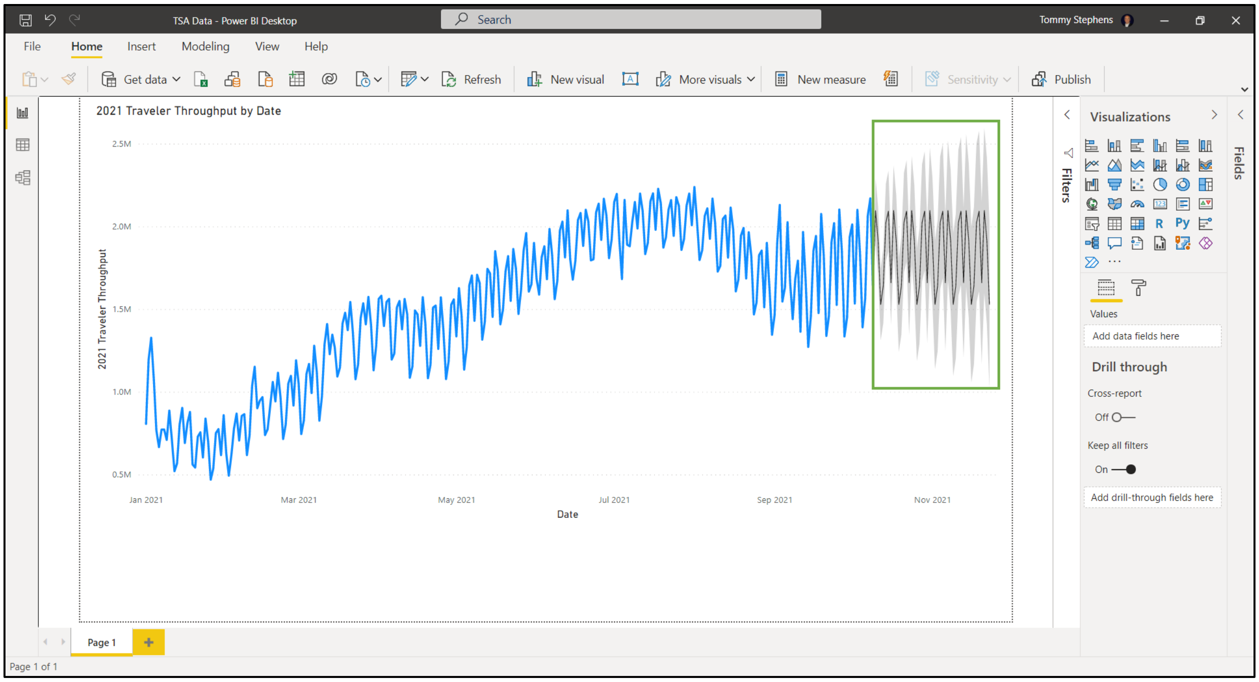


Figure 64 - Creating Forecasted Data in Power BI

Power BI Summary

Power BI provides an incredibly robust ecosystem. While many professionals view it purely as a "reporting" platform – and it is excellent at fulfilling that use case – Power BI is also a powerful tool for generating business analytics. Power BI offers tremendous data query and transformation capabilities. In addition, it provides rich visualizations, powerful calculation capabilities, and artificial intelligence capabilities to help data analysts, auditors, and other business professionals achieve better results in less time. Further, most of these options are available in the free Power BI Desktop environment, so no out-of-pocket investment is necessary. Given these benefits, the obvious question is, "What's keeping you from using Power Query to assist with your data analytics needs?"