ANALYZING DATA IN EXCEL

1 VARIANCE ANALYSIS IN EXCEL 3

1.1 EXCEL TABLES 3
1.1.1 Excel Tables 3
1.1.2 Instructions 5
1.1.3 Questions 5
1.2 FINDING THE NUMBER OF CATEGORIES 6
1.2.1 Data hierarchy 6
1.2.2 Instructions 8
1.2.3 Questions 8
1.3 CREATING THE PIVOTTABLE REPORT 8
1.3.1 Instructions 9
1.3.2 Questions 9
1.4 DRILLING DOWN AND ROLLING UP 9
1.4.1 Instructions 10
1.4.2 Questions 10
1.5 CREATING CUSTOM GROUPS 11
1.5.1 Instructions 11
1.5.2 Questions 11
1.6 CREATING CALCULATED FIELDS 12
1.6.1 Instructions 12
1.6.2 Questions 13
1.7 CUSTOMIZING PIVOTTABLE SUMMARIES 14
1.7.1 Instructions 14
1.7.2 Questions 14
1.8 VISUALIZING VARIANCES 15
1.8.1 Instructions 16
1.8.2 Questions 16

2 TREND ANALYSIS IN EXCEL 17

2.1 EXPLORING THE DATA 17
2.1.1 Instructions 19
2.1.2 Questions 19
2.2 VISUALIZING THE DISTRIBUTION OF HOUSING STARTS 20
2.2.1 Instructions 20
2.2.2 Questions 20
2.3 VISUALIZING THE TREND OF HOUSING STARTS 20
2.3.1 Instructions 21
2.3.2 Questions 21
2.4 RESAMPLING THE TREND BY QUARTER 21
2.4.1 Instructions 22
2.4.2 Questions 22
2.5 VISUALIZING SEASONALITY 22
2.5.1 Instructions 23
2.5.2 Questions 23
2.6 ADDING A LINEAR FORECAST 23
2.6.1 Instructions 24

2.6.2 Questions *24*
2.7 Using the Forecast Sheet Feature 25
2.7.1 Instructions *25*
2.7.2 Questions *25*
3 EXPLORATORY DATA ANALYSIS IN EXCEL 27
3.1 Preparing the Data 27
3.1.1 Instructions *27*
3.1.2 Questions *28*
3.2 Creating a Proportion Table 28
3.2.1 Instructions *29*
3.2.2 Questions *29*
3.3 Deriving Descriptive Statistics 30
3.3.1 Instructions *30*
3.3.2 Questions *30*
3.4 Calculating Descriptive Statistics by Group 31
3.4.1 Instructions *31*
3.4.2 Questions *32*
3.5 Visualizing the Distribution of Two Groups 32
3.5.1 Instructions *32*
3.5.2 Questions *33*
3.6 Creating a Correlation Matrix 34
3.6.1 Instructions *34*
3.6.2 Questions *35*
3.7 From Correlation to Regression 35
3.7.1 Instructions *36*
3.7.2 Questions *36*
4 CONFIRMATORY DATA ANALYSIS IN EXCEL 36
4.1 The Law of Large Numbers 36
4.1.1 Instructions *38*
4.1.2 Questions *39*
4.2 Data Distributions 39
4.2.1 Instructions *39*
4.2.2 Questions *40*
4.3 The Central Limit Theorem 40
4.3.1 Instructions *41*
4.3.2 Questions *41*
4.4 The Independent Samples T-test 42
4.4.1 Instructions *43*
4.4.2 Questions *44*
4.5 Adding Confidence Intervals 44
4.5.1 Instructions *45*
4.5.2 Questions *45*
4.6 Bivariate Regression 46
4.6.1 Instructions *46*
4.6.2 Questions *46*
4.7 Evaluating Regression Model Fit 47
4.7.1 Instructions *48*
4.7.2 Questions *48*
5 BUSINESS AND FINANCIAL MODELING IN EXCEL 49
5.1 What Percent of Claims Are Disputed? 49

5.1.1 Instructions *49*
5.1.2 Questions *50*
5.2 Is the Average Disputed Invoice Significantly Different? 50
5.2.1 Instructions *50*
5.2.2 Questions *51*
5.3 How Much Later Are Disputed Payments? 51
5.3.1 Instructions *51*
5.3.2 Questions *51*
5.4 Building a One-Way Data Table 52
5.4.1 Instructions *54*
5.4.2 Questions *54*
5.5 Building a Two-Way Data Table 55
5.5.1 Instructions *56*
5.5.2 Questions *57*
5.6 Finding a Break-Even Point with Goal Seek 57
5.6.1 Instructions *58*
5.6.2 Questions *59*

1 Variance Analysis in Excel

1.1 Excel Tables

Data analysts are often tasked with exploring and explaining differences between budgeted and actual figures. This practice is known as *variance analysis*.

For this lesson, we'll be looking at budgeted versus actual annual costs recorded by the city of Nashville, Tennessee (data source: Nashville.gov).

1.1.1 Excel Tables

Most of the datasets in this course will make use of Excel **Tables**, including the one for this lesson. You can use the same formulas and functions on **Tables** that you've used before, but the syntax might look a little different.

Table syntax uses what's called a structured reference by using a combination of the table name and a column name. For example, if you had a `Sales` table with a `Discount` column, you could refer to it using `Sales[Discount]` rather than using a static reference like `B2:B314`. Referencing a cell within a **Table** will look a little different as well; rather than an absolute cell address like `A7`, the **Table** will use a structured reference that looks like `[@Discount]` when referring to a cell in the `Discount` column.

Recall, to convert a dataset into a named **Table**, do the following:

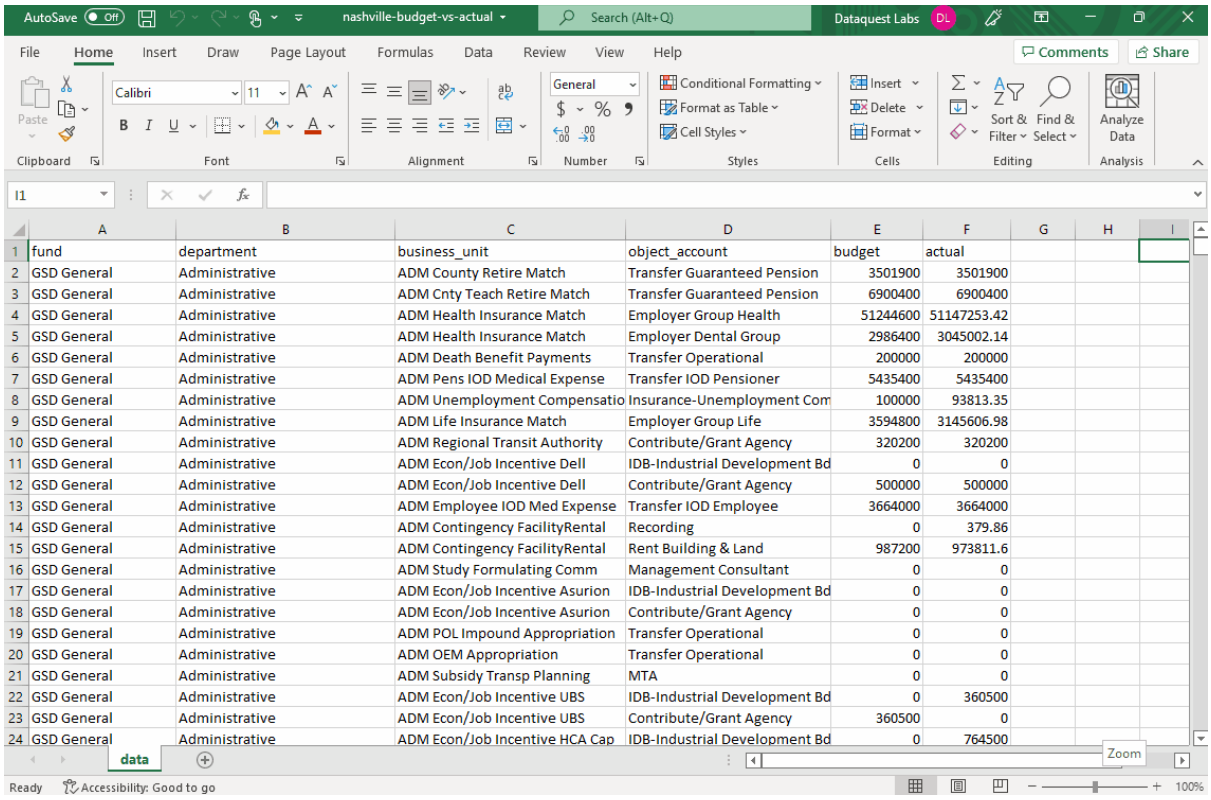
1. Select any cell within the dataset and select **Insert > Tables > Table**.
2. Click `OK` to convert the selected range into a **Table**.

- Make sure the `My table has headers` option has been enabled.

3. Rename the table to `nashville`.

- Under `Table Design`, use the **Table Name** box in the top left of the screen to change `Table1` to `nashville`.

You can see these steps in action here:



	A	B	C	D	E	F	G	H	I
1	fund	department	business_unit	object_account	budget	actual			
2	GSD General	Administrative	ADM County Retire Match	Transfer Guaranteed Pension	3501900	3501900			
3	GSD General	Administrative	ADM Cnty Teach Retire Match	Transfer Guaranteed Pension	6900400	6900400			
4	GSD General	Administrative	ADM Health Insurance Match	Employer Group Health	51244600	51147253.42			
5	GSD General	Administrative	ADM Health Insurance Match	Employer Dental Group	2986400	3045002.14			
6	GSD General	Administrative	ADM Death Benefit Payments	Transfer Operational	200000	200000			
7	GSD General	Administrative	ADM Pens IOD Medical Expense	Transfer IOD Pensioner	5435400	5435400			
8	GSD General	Administrative	ADM Unemployment Compensatio	Insurance-Unemployment Com	100000	93813.35			
9	GSD General	Administrative	ADM Life Insurance Match	Employer Group Life	3594800	3145606.98			
10	GSD General	Administrative	ADM Regional Transit Authority	Contribute/Grant Agency	320200	320200			
11	GSD General	Administrative	ADM Econ/Job Incentive Dell	IDB-Industrial Development Bd	0	0			
12	GSD General	Administrative	ADM Econ/Job Incentive Dell	Contribute/Grant Agency	500000	500000			
13	GSD General	Administrative	ADM Employee IOD Med Expense	Transfer IOD Employee	3664000	3664000			
14	GSD General	Administrative	ADM Contingency FacilityRental	Recording	0	379.86			
15	GSD General	Administrative	ADM Contingency FacilityRental	Rent Building & Land	987200	973811.6			
16	GSD General	Administrative	ADM Study Formulating Comm	Management Consultant	0	0			
17	GSD General	Administrative	ADM Econ/Job Incentive Asurion	IDB-Industrial Development Bd	0	0			
18	GSD General	Administrative	ADM Econ/Job Incentive Asurion	Contribute/Grant Agency	0	0			
19	GSD General	Administrative	ADM POL Impound Appropriation	Transfer Operational	0	0			
20	GSD General	Administrative	ADM OEM Appropriation	Transfer Operational	0	0			
21	GSD General	Administrative	ADM Subsidy Transp Planning	MTA	0	0			
22	GSD General	Administrative	ADM Econ/Job Incentive UBS	IDB-Industrial Development Bd	0	360500			
23	GSD General	Administrative	ADM Econ/Job Incentive UBS	Contribute/Grant Agency	360500	0			
24	GSD General	Administrative	ADM Econ/Job Incentive HCA Cap	IDB-Industrial Development Bd	0	764500			

Using **Tables** has many benefits, including:

- ability to quickly filter data using the header row
- banded rows for easy reading of data
- **calculated columns** automatically fill a column of a **Table** with data based on the value of one cell
- easily add a total row to your data
- more reliable and dynamic than static references

To learn more about **Tables** and **calculated columns**, check out this post from Microsoft.

Now try a couple of formulas that use **Table syntax**!

1.1.2 Instructions

If you get stuck or would like to compare your answers, we've provided a suggested solution file for this lesson called `nashville-budget-vs-actual-solutions.xlsx`, which you can find in the `Home Folder` under `This PC`.

1. Open the file `nashville-budget-vs-actual.xlsx` in the `Home Folder` under `This PC`.
2. Convert the dataset into a **Table** and rename it to `nashville` as shown above.
3. Get the total budget by using the `SUM()` function on the `budget` column using **Table syntax**.
4. In column `G`, create a **calculated column** called `$ to budget`:

- Populate column `G` with the difference between the `budget` column and the `actual` column:
- In cell `G2`, begin your formula with `=`
- Click on `E2`, then press minus (`-`) on your keyboard, and then click on `F2`
- Press `Enter` to automatically fill column `G` with the difference between `budget` and `actual`

1.1.3 Questions

1.

What formula using Table syntax will sum the `budget` column of the `nashville` Table?

`SUM(E:E)`

`SUM(E2:E23371)`

`SUM(nashville[budget])`

`SUM(nashville(budget))`

2.

What formula using Table syntax will create a calculated column that subtracts the `actual` column from the `budget` column?

```
(@budget)-(@actual)
```

```
[@actual]-[@budget]
```

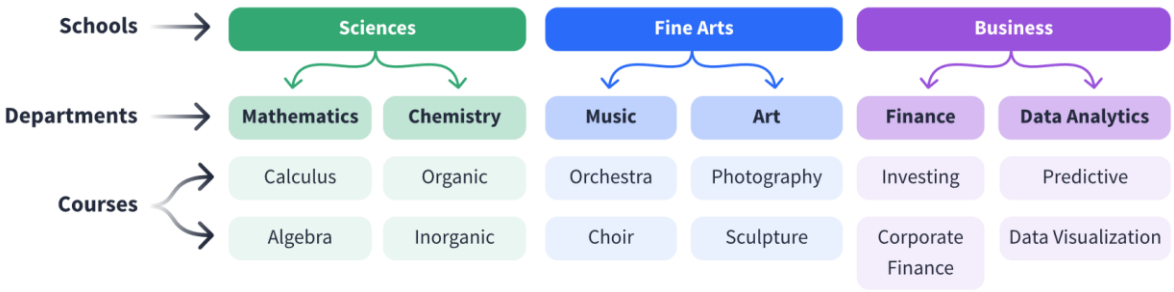
```
E2-F2
```

```
[@budget]-[@actual]
```

1.2 Finding the number of categories

1.2.1 Data hierarchy

Budgets typically exist at varying levels of a hierarchy. For example, most schools are split into academic disciplines which could be summarized by schools, departments, and so forth like so:



This means that analysts must be comfortable checking for variances at different levels of the hierarchy, and presenting their findings in a digestible package.

Our Nashville data is presented with four levels of hierarchy: `fund`, `department`, `business unit`, and `object account`. For our first step, we will find out how many distinct categories can be found in each column. This gives us the relative level of detail we'll get from each.

To do this, we can use Excel's `UNIQUE()` function which lists all distinct values in any range. For exmaple, we can use it to find all the unique values in the `country_of_origin` column of the `products` table in this small dataset:

D2 =UNIQUE(products[country_of_origin])

	A	B	C	D	E
1	sku	country_of_origin			
2	247512	Portugal		Portugal	
3	865367	Ireland		Ireland	
4	533856	Canada		Canada	
5	996416	Sweden		Sweden	
6	431543	Philippines		Philippines	
7	595606	Portugal		Indonesia	
8	357979	Indonesia		Nigeria	
9	184799	Nigeria		India	
10	966164	India			
11	543574	Canada			
12					

If we wrap this inside the `COUNTA()` function, we can find the *number* (or *count*) of distinct values:

D2 =COUNTA(UNIQUE(products[country_of_origin]))

	A	B	C	D	E	F
1	sku	country_of_origin				
2	247512	Portugal		8		
3	865367	Ireland				
4	533856	Canada				
5	996416	Sweden				
6	431543	Philippines				
7	595606	Portugal				
8	357979	Indonesia				
9	184799	Nigeria				
10	966164	India				
11	543574	Canada				
12						

This tells us that there are `8` unique values in the `country_of_origin` column of the `products` table.

1.2.2 Instructions

1. Following the example above, use **Table syntax** to find the number of unique values in the following columns of the `nashville` table (columns `A-D`):

- `fund`
- `department`
- `business unit`
- `object account`

2. You can delete these values from your worksheet after answering the question below.

1.2.3 Questions

1.

Which column contains the most unique values?

`department`

`object_account`

`fund`

`business_unit`

1.3 Creating the PivotTable report

A major objective of variance analysis is to examine budgeted and actual figures at these varying levels of detail. PivotTables will allow us to easily resummarize in order to see these levels of detail.

Generally, it's a good idea to begin the variance analysis at the highest level of detail: in our case, that's the `fund` column.

1.3.1 Instructions

1. Create a PivotTable with `fund` in **Rows**.
2. Add `Sum of actual` and `Sum of budget` to **Values**.
- Format each column as a `currency` with `0` decimal places.

1.3.2 Questions

1.

Which `fund` had the smallest budget?

MNPS Debt Service

USD Debt Service

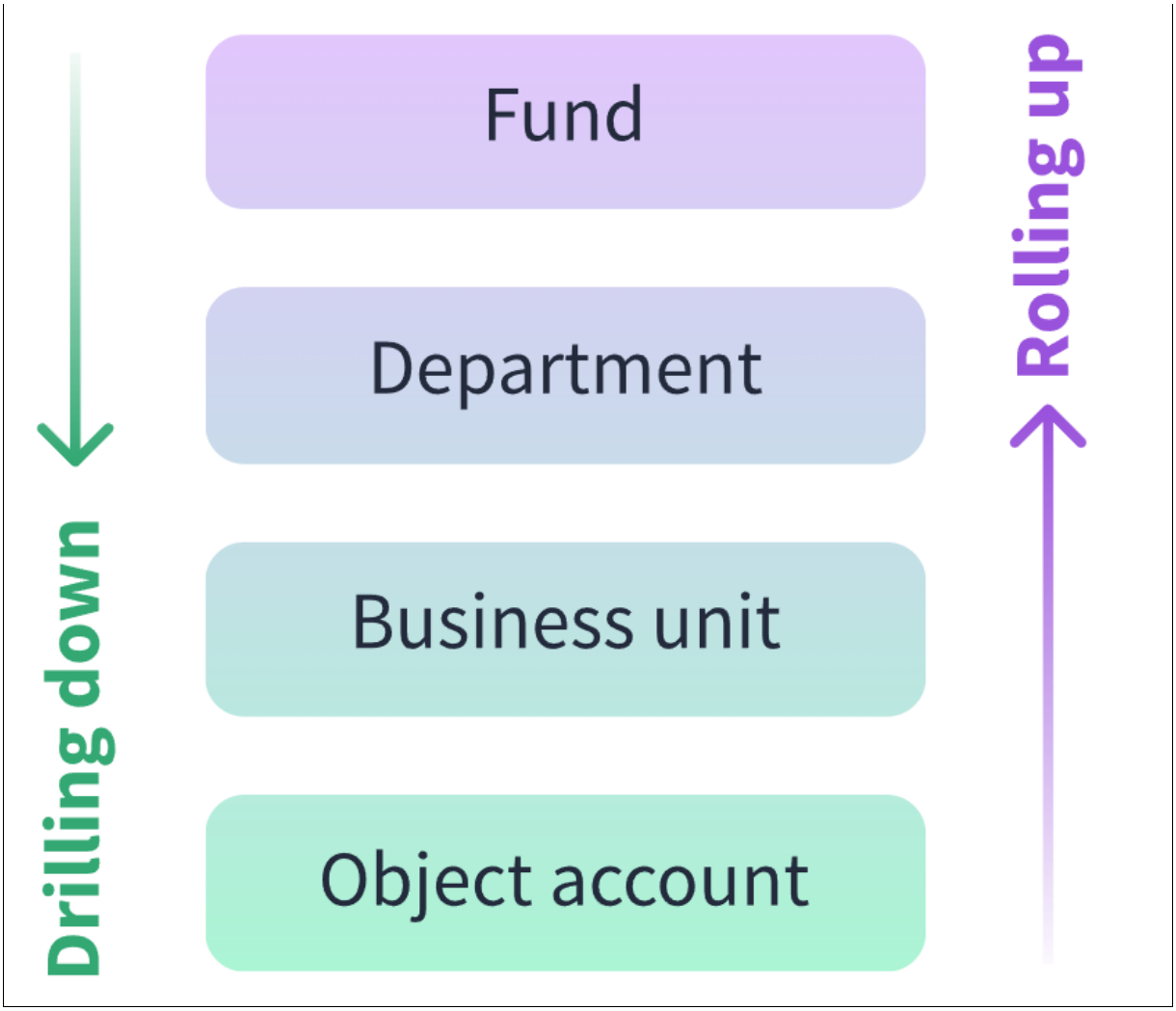
USD General

GSD General

1.4 Drilling down and rolling up

It's one thing to know that *something* was over or under budget, and another to know *why*. By drilling into the results of our report, we can pinpoint specific areas to inquire about with the business.

This process of adding additional levels of detail to the report is called **drilling down**. Conversely, **rolling up** is used to view the data at a less granular level:



1.4.1 Instructions

1. Add an extra layer of hierarchy to your PivotTable by placing `department` in **Rows**.

- You will see that some funds are split into multiple departments while others only have one.

1.4.2 Questions

1.

How many funds consist of only one department?

2

4

0

3

1.5 Creating custom groups

We are getting a better handle on how complex our data's hierarchy is, which will help us pinpoint important variances.

In the previous screen, you saw that four funds consist of only one department. Three of these funds relate to debt service. Let's create a custom grouping of these three funds called "Debt Services."

1.5.1 Instructions

1. Hold down the `Ctrl` key and click the fund labels for `GSD Debt Service`, `MNPS Debt Service`, and `USD Debt Service`.
2. Right-click on the selection and choose **Group**.
3. Rename "Group1" to `Debt Services` by typing over "Group1."

1.5.2 Questions

1.

What is the total budget for your custom Debt Services group?

$285,532,838

$272,569,400

$169,296,200

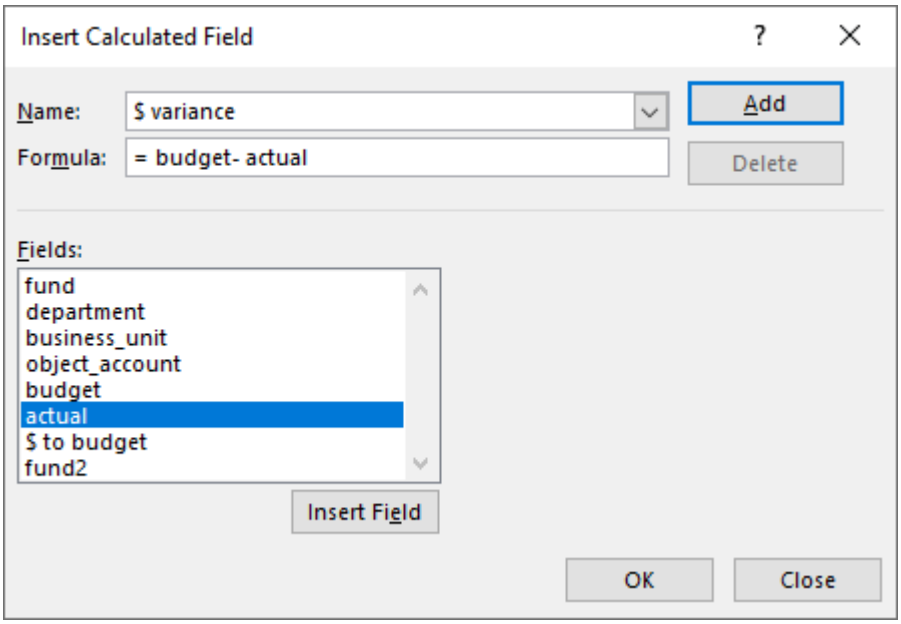
$292,226,700

1.6 Creating calculated fields

So far we've been familiarizing ourselves with the overall hierarchy of the data. But we haven't done much with the real *variance* between `budget` and `actual` amounts. We could use the `$ to budget` column we created earlier but let's use our PivotTable to calculate this value instead. To do that, we will create what's called a **Calculated Field** in our PivotTable.

To create one, click anywhere in the PivotTable, then select **PivotTable Analyze > Calculations > Fields, Items, & Sets > Calculated Field**.

We will create a field called `$ variance` which will be the difference between `budget` and `actual`. Fill out the menu like so and click `OK`:



This variable will then be available in your **PivotTable Fields** so that you can add it to the PivotTable. This will allow you to see exaclty how much a particular fund was *over* or *under* budget.

1.6.1 Instructions

1. Create `$ variance` as show above.

- Format the column as a `currency` with `0` decimal places.

2. Create a new **Calculated Field** called `% variance` for the percentage difference between the `budget` and `actual` figures using the **Formula**: $\frac{(\text{budget} - \text{actual})}{\text{budget}}$ (budget−actual)budget

- Format the column as a `percentage` with `1` decimal place.

1.6.2 Questions

1.

How does the total `$ variance` for the custom group `Debt Services` compare to `$ variance` for `MNPS General Purpose`?

- It's about half of `MNPS General Purpose`
- It's about five times `MNPS General Purpose`
- It's about double that of `MNPS General Purpose`
- It's about the same as `MNPS General Purpose`

2.

Referring to `% variance`, how much *over* budget was the USD Debt Service fund?

- 1.8%
- 2.5%
- 2.3%

2.9%

1.7 Customizing PivotTable summaries

We've learned that with one exception, every fund came in under budget for the year. Not bad! If you were asked to dig into where the positive differences came from, you could easily drill into the PivotTable.

Let's take a different approach and explore the relative magnitude of each fund to the overall total.

To do this, we will place another instance of `actual` in the **Values** section of our PivotTable. As it turns out, we can display these figures in the PivotTable in many different ways.

Then we would right-click on any number in the new `actual` column, select **Show Values As > % of Column Total**. This would give us the total contribution of each fund to the `actual` column. Notice how the percentages sum to 100%.

Let's dig in to understand the contributions of each Debt Service category to its custom group.

1.7.1 Instructions

1. Add another instance of `actual` to the **Values** section of your PivotTable.
2. Display the second `actual` column as `% of Parent Row Total`.

- If necessary, format the column as a `percentage` with `2` decimal places.

1.7.2 Questions

1.

What percent of the overall `Debt Services` custom group does `USD Debt Service` make up?

7.05%

100.00%

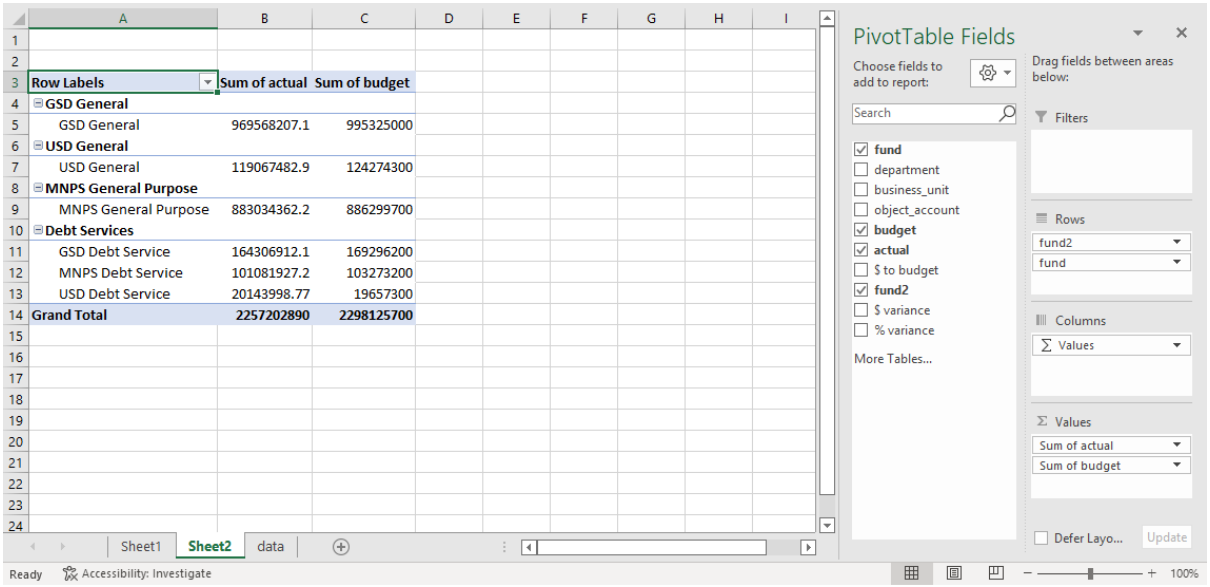
35.40%

12.65%

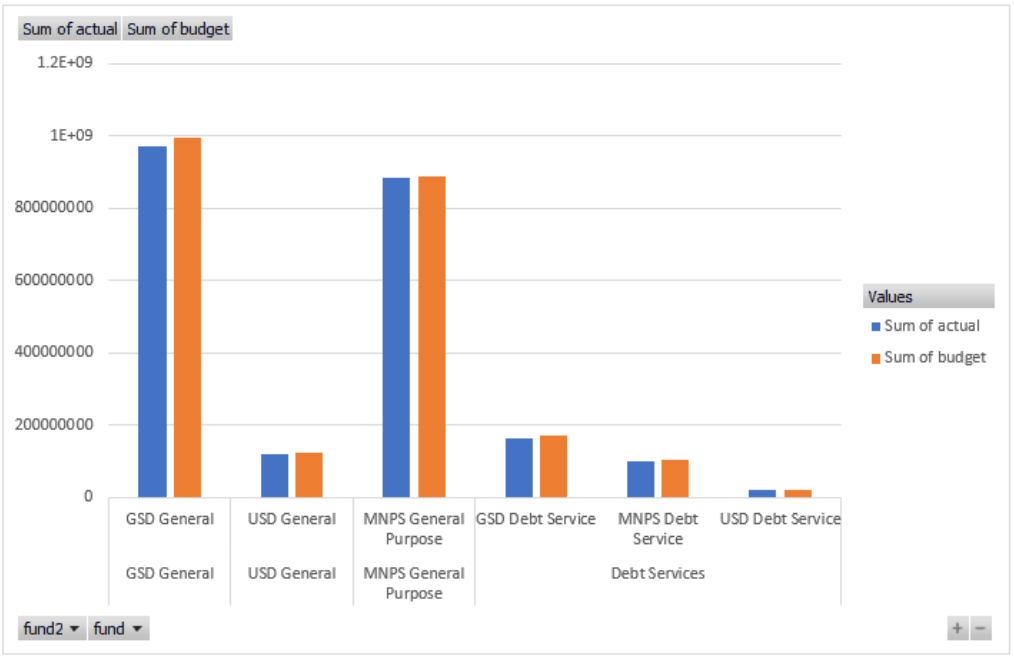
1.8 Visualizing variances

We've done a fantastic job of exploring variances and contributions in this budget data. Now it's time to think about communicating some of these findings to others. As we know, data visualizations are great for that!

We will create a bar chart comparing the budgeted and actual values for each `fund`. We begin by creating a new PivotTable with our special `Debt Services` group in **Rows**. This grouping should be available as a **PivotTable Field** called `fund2`. However, when we place `fund2` into **Rows**, you'll notice our custom name `Debt Services` has reverted to `Group1`. This is because the `fund2` field was created before we changed the name of our group. Go ahead and type over `Group1` with `Debt Services` like we did before. Also, let's add another level of hierarchy by placing `fund` below `fund2` in **Rows**. Finally, place `Sum of actual` and `Sum of budget` inside **Values**. The result should look like this:



Clicking anywhere inside the PivotTable and selecting **Insert > Charts > PivotChart** will bring up the **Insert Chart** window. The default recommended chart will be a column chart. When you click `OK`, you will see that the data will be charted by each individual fund, rather than by our `Debt Services` as a single group. To change that, click on the minus sign in the lower right of the chart to **Collapse Entire Field** so that only four categories are being charted: `GSD General`, `USD General`, `MNPS General Purpose`, and `Debt Services`. You can see that being done here:



1.8.1 Instructions

1. Create the bar chart as described above and answer the True or False question below.
2. Replace the `actual` and `budget` figures in your chart with the `% variance` **Calculated Field** we created ealier and answer the multiple choice question below.

- Remove `Sum of actual` and `Sum of budget` from **Values**
- Add `% variance` to **Values** and format the column to display as a `percentage` with `2` decimal places

1.8.2 Questions

1.

Which of the following statements are true, based on the PivotChart for `% variance`?

(select all that apply)

☐ The `% variance` is about equal for all four categories being charted.

The biggest difference in `% variance` is between `USD General` and `MNPS General Purpose`.

The `% variance` is about equal for `Debt Services` and `GSD General`

Clicking on **Expand Entire Field** in the **PivotChart** shows that `MNPS Debt Service` under `Debt Services` went *over* budget.

2.

True or False: every group displayed on the chart is under budget.

True

False

2 Trend Analysis in Excel

2.1 Exploring the Data

For this lesson, we'll be analyzing trends in the monthly number of housing starts (in thousands) from January 2012 through December 2021. (Data source: Federal Reserve Bank of St. Louis]. The term "housing starts" is an economic indicator that measures the number of privately-owned new houses on which construction has begun during a given month.

The data is stored as a Table called `hstarts` in a single worksheet of the provided `housing-starts.xlsx` Excel file. Open the file to continue along with the lesson. Notice how the `date` column is monthly; all dates are for the first day of each month. The `starts` column records the number of new houses being built that month (measured in thousands).

Working with *time series data*, or data collected from the same observation over multiple periods of time, will require a combination of skills you've picked up so far along with some new ones. Let's get started.

Recall that you can filter for the **Top 10** or **Bottom 10** of a numerical column by clicking on the filter in the column header and selecting **Number Filters > Top 10**, like so:

	A	B	C	D	E	F	G	H	I	J
1	date	starts								
2	1/1/2012	47.2								
3	2/1/2012	49.7								
4	3/1/2012	58								
5	4/1/2012	66.8								
6	5/1/2012	67.8								
7	6/1/2012	74.7								
8	7/1/2012	69.2								
9	8/1/2012	69								
10	9/1/2012	75.8								
11	10/1/2012	77								
12	11/1/2012	62.2								
13	12/1/2012	63.2								
14	1/1/2013	58.7								
15	2/1/2013	66.1								
16	3/1/2013	83.3								
17	4/1/2013	76.3								
18	5/1/2013	87.2								
19	6/1/2013	80.7								
20	7/1/2013	84								
21	8/1/2013	80.4								
22	9/1/2013	78.4								
23	10/1/2013	78.4								
24	11/1/2013	83.8								
25	12/1/2013	67.6								
26	1/1/2014	60.7								

From here, you can customize the number of items you would like to be displayed from either the top or the bottom of the dataset.

2.1.1 Instructions

If you get stuck or would like to compare your answers, we've provided a suggested solution file for this lesson called `housing-starts-solutions.xlsx`, which you can find in the `Home Folder` under `This PC`.

1. After opening the file `housing-starts.xlsx`, filter the `starts` column to find the months with:

- the most housing start.
- the fewest housing starts.

2.1.2 Questions

1.

What month had the fewest housing starts?

June 2021

February 2013

January 2012

December 2021

2.

What month had the most housing starts?

December 2021

June 2021

October 2020

January 2012

2.2 Visualizing the Distribution of Housing Starts

In an earlier course, you used histograms to visualize a variable's distribution. Let's do so again here. Make sure to remove your filter from any previous analysis to capture the entire dataset.

2.2.1 Instructions

1. Create a histogram of `starts` and edit your histogram to have 15 bins.

2.2.2 Questions

1.

How many observations are found in the range between 111.46 and 118.6?

20

22

15

1

2.3 Visualizing the Trend of Housing Starts

Visualizing the distribution of housing starts using a histogram is interesting. But what about the time element of the dataset? Perhaps *that* is giving important information about housing starts that should be included as well? As we have learned, histograms are not good for showing changes over time. Let's try a chart type we know is good for displaying time series data: a line chart.

2.3.1 Instructions

1. Insert a line chart visualizing how `starts` changes by `date`.

2.3.2 Questions

1.

In general, housing starts have ___ over time.

remained the same

decreased

Not enough information to decide

increased

2.4 Resampling the Trend by Quarter

Visualizing the trend of housing starts over time as a line graph tells us something about the data that a histogram didn't: we can see that in general, the number of starts has increased over time.

However, there seems to be quite a lot of variability in the data: it's anything but a straight line from 2012 to 2021.

In cases like this, it can help to *resample* the time series data, or view it at a different level of the date hierarchy like you did in the previous lesson. This is the trend analysis equivalent of rolling up and drilling down into the data to uncover patterns.

2.4.1 Instructions

1. Create a PivotTable, placing `date` along **Rows** and `Sum of starts` in **Values**.
2. Right-click on any of the **Year** labels in the PivotTable under **Row Labels** and select **Group**.
3. In the **Grouping** diaglog box that appears, make sure:

- Only **Quarters** and **Years** are selected
- **Starting at:** is set to `1/1/2012`
- **Ending at:** is set to `12/2/2021`

4. After you click `OK`, the PivotTable will be summarizing `starts` by quarter.
5. Insert a **Line** PivotChart.

2.4.2 Questions

1.

What is the greatest number of housing starts in any Quarter?

209.3

47.2

154.3

435.5

2.5 Visualizing Seasonality

It appears from our investigation thus far that housing starts have generally trended upwards over time. We can see that for most years, **Quarters** 2 and 3 had more starts than **Quarters** 1 and 4. We have seen these types of regular annual changes in the data before; we know it as *seasonality*. But just how regular is it? Let's create a chart to find out.

2.5.1 Instructions

1. Create another PivotTable with:

- `date` along **Columns**
- `Years` in **Rows**
- `Sum of starts` in **Values** with **Show Values As** set to **% of Row Total**.

2. Insert a **100% Stacked Column** PivotChart.

2.5.2 Questions

1.

Rounded to the nearest whole number, what percent of total housing starts in 2012 were performed in Q1?

22%

20%

26%

27%

2.6 Adding a Linear Forecast

Data analysts are commonly tasked with developing forecasts to help the business make plans for the future given what is likely to happen. Forecasting is a very complex topic; we'll just scratch the surface here using a couple of Excel features.

First, let's try adding a **Linear Forecast** to the data using Excel's **Trendline** feature. Recall from a previous course that a *trendline* is a line used to visualize the general statistical trend of the underlying data.

2.6.1 Instructions

1. In a new worksheet, create a line chart showing housing starts by month.
2. Under **Chart Elements**, add a **Trendline > Linear Forecast** .

2.6.2 Questions

1.

Approximately how far off is our Linear Forecast to the actual value for April 2018?

The actual exceeded forecast by about 15 units

The forecast exceeded actual by about 15 units

The actual exceeded forecast by about 50 units

The actual and forecasted values are about the same

2.

Using the Trendline, estimate how many housing starts are anticipated for January 2022.

70

130

120

110

2.7 Using the Forecast Sheet Feature

The problem with using a **Linear Forecast** on our housing starts dataset is that... *the data isn't totally linear*! While there is a consistent increase over time in the number of housing starts, there is consistent variability by month, and this seasonal aspect of the trend is *not* linear.

Fortunately, Excel includes an easy feature for building a forecast that can account for seasonality. Let's take a look.

2.7.1 Instructions

1. Return to your original `hstarts` dataset on the `data` tab, and click inside any cell.
2. Under the **Data** menu on the ribbon, select **Forecast > Forecast Sheet**.

- The **Create Forecast Worksheet** dialog box will appear.
- While there are many options we can set here, let's just make sure:
- **Create line chart** is selected in the top right
- **Forecast End** is set to **6/1/2024**
- After you click `Create`, a new worksheet will appear in your workbook containing a Table and chart.

2.7.2 Questions

1.

Of the options below, which month has the highest forecasted number of housing starts?

April 2023

June 2024

January 2022

April 2024

2.

Rounded to the nearest whole number, what is the forecasted value for February 2022 at the *lower confidence bound*?

88

117

98

137

3.

Is it possible to change the forecasted time window using a Forecast Sheet?

No

Yes

3 Exploratory Data Analysis in Excel

3.1 Preparing the Data

The dataset we will use for this lesson consists of vehicle specifications, including fuel mileage (measured in **mpg**, miles per gallon) for a number of cars from the '70s and early '80s. Data source: University of California, Irvine Machine Learning repository.

Our first task is to explore the data! We want to use a variety of descriptive and visual techniques to see what types of relationships and insights we might find in the data. As you recall, this process is formally known as *exploratory data analysis*, or EDA. Often you'll come to EDA with some initial questions or objectives in mind. Ours will be, *Which variables influence vehicle mileage*?

Let's open the `mpg-eda.xlsx` dataset and get started!

3.1.1 Instructions

If you get stuck or would like to compare your answers, we've provided a suggested solution file for this lesson called `mpg-eda-solutions.xlsx`, which you can find in the `Home Folder` under `This PC`.

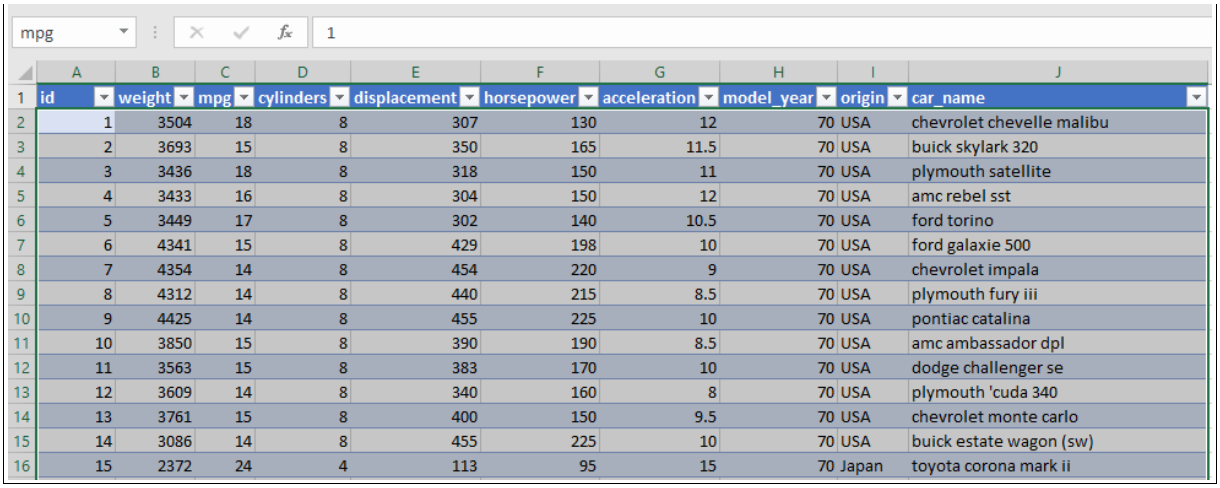
To make the `mpg-eda.xlsx` dataset easier to work with for this lesson, do the following:

1. Insert a new column to the left of `weight`.
2. Name the new column `id`.
3. Manually enter `1`, `2`, and `3` as the first three values of `id`

- Highlight the first three values you just entered under `id`, and double-click the **fill handle** to **flash fill** the entire column with values.
- Confirm that the last row in the dataset has a value of `392` for `id`.

4. Convert the resulting dataset into a Table called `mpg` while making sure `My table has headers` is enbaled.

Your resulting Table should look like this:



mpg | 1

	A	B	C	D	E	F	G	H	I	J
1	id	weight	mpg	cylinders	displacement	horsepower	acceleration	model_year	origin	car_name
2	1	3504	18	8	307	130	12	70	USA	chevrolet chevelle malibu
3	2	3693	15	8	350	165	11.5	70	USA	buick skylark 320
4	3	3436	18	8	318	150	11	70	USA	plymouth satellite
5	4	3433	16	8	304	150	12	70	USA	amc rebel sst
6	5	3449	17	8	302	140	10.5	70	USA	ford torino
7	6	4341	15	8	429	198	10	70	USA	ford galaxie 500
8	7	4354	14	8	454	220	9	70	USA	chevrolet impala
9	8	4312	14	8	440	215	8.5	70	USA	plymouth fury iii
10	9	4425	14	8	455	225	10	70	USA	pontiac catalina
11	10	3850	15	8	390	190	8.5	70	USA	amc ambassador dpl
12	11	3563	15	8	383	170	10	70	USA	dodge challenger se
13	12	3609	14	8	340	160	8	70	USA	plymouth 'cuda 340
14	13	3761	15	8	400	150	9.5	70	USA	chevrolet monte carlo
15	14	3086	14	8	455	225	10	70	USA	buick estate wagon (sw)
16	15	2372	24	4	113	95	15	70	Japan	toyota corona mark ii

3.1.2 Questions

1.

With your data set up for EDA, what is the result of the following formula?

```
=INDEX(mpg, 4, 3)
```

318

304

16

8

3.2 Creating a Proportion Table

As we learned in a previous course, we can broadly categorize variables as either categorical or numerical.

Recall that categorical variables tell us "what kind" or "which type" of something. For example, the `origin` column of our dataset tells us whether each car is from the USA, Japan, or Europe. Therefore, `origin` is a categorical variable.

We can't do math on this variable, such as dividing Europe by Japan. We can, however, count how often we find each unique value in the column in order to create a **Frequency Table**. However, rather than use one of the `COUNT` functions like we've done previously to create a Frequency Table, let's use a PivotTable with percentages to create what's called a **Proportion Table** instead.

A Proportion Table differs from a Frequency Table in that it uses **relative frequencies** instead of **absolute frequencies**. Relative frequencies are expressed as a *percentage*, whereas absolute frequencies are expressed as a *count*. We calculate a relative frequency by dividing how often a value occurs in a column (*count*) by the number of observations in that column (number of rows).

3.2.1 Instructions

Use a PivotTable to create a **Proportion Table** for the `origin` column of `mpg`:

1. Insert a PivotTable, with the following:
- `origin` along **Rows**
- `Count of id` inside **Values**
2. Right-click on the `Count of id` header in your PivotTable, and select **Show Values As > % of Column Total**.
3. Use the resulting PivotTable to answer the True or False question below.

3.2.2 Questions

1.

True or False: the majority of cars found in this dataset originate from the USA.

True

False

3.3 Deriving Descriptive Statistics

The most common way to summarize categorical variables is to count their frequencies or calculate their proportions. By contrast, numerical variables feature an array of descriptive statistics.

Take `weight`, for example. With this variable, we can easily find statistics like the maximum and minimum values. Rather than use the `MIN()` & `MAX()` functions or filtering using **Top/Bottom 10** like we've done previously, let's use Excel's **Analysis ToolPak** to derive this variable's **descriptive statistics**. Descriptive statistics summarize a dataset using measures of central tendency, variability, and some basic summary statistics like minimum, maximum, and range.

The **Analysis ToolPak** is a free add-in that is available for desktop versions of Excel. While we have already loaded it into Excel for you here, you may want to load it on your personal copy of Excel locally. If you would like to do so, please follow these instructions.

3.3.1 Instructions

Generate **descriptive statistics** for the `weight` column:

1. Navigate to **Data > Analyze > Data Analysis**.
2. Select **Descriptive Statistics** from the **Data Analysis** window that appears. The options listed here are in alphabetical order.
3. In the **Descriptive Statistics** window, do the following:

- Select the **Input Range** for the `weight` variable (column `B`) making sure you include the row with the header label.
- Enable the option **Labels in the First Row**.
- Under **Output options**, enable **Summary statistics**.

4. Click `OK` to create a table of summary statisctics on a new worksheet.

3.3.2 Questions

1.

What is the average value of `weight`, rounded to the nearest whole number?

1613

2804

5140

2978

3.4 Calculating Descriptive Statistics by Group

So far, we've looked at the relative frequencies of `origin` and the descriptive statistics for `weight`. Let's combine these variables and see how the descriptive statistics for `weight` varies by `origin`.

3.4.1 Instructions

1. Create a PivotTable for `mpg`, placing the following:

- `origin` along **Columns**
- `Sum of mpg` inside **Values**
- `id` in **Rows**
- By adding the `id` column (unique identifier) to the dataset, each row will be individually added to the PivotTable without grouping (aggregating) it.

2. Turn off the **Grand Total** option for this PivotTable:

- click anywhere inside the PivotTable and navigate to **Design > Layout > Grand Totals > Off for Rows and Columns**

3. Run **descriptive statistics** on all three columns (`Europe`, `Japan`, and `USA`) at the same time:

- don't worry about missing data in the PivotTable; Excel will disregard any blank cells found in the columns.
- in the **Descriptive Statistics** window:
- Set **Input Range** to include all three columns, including header labels
- Enable **Labels in first row**
- Set the **Output Range** to `F5`
- Enable **Summary statistics**

3.4.2 Questions

1.

Which `origin` has the greatest Standard Deviation in `mpg`?

Not enough information

Japan

Europe

USA

3.5 Visualizing the Distribution of Two Groups

We've been looking at how `mpg` varies by `origin`. A histogram would make a great visual to accompany this analysis.

Unfortunately, Excel doesn't include a feature to plot multiple histograms on the same chart. However, we can trick Excel into making one using a PivotChart! We'll do this by creating and customizing a column chart so that it behaves like a histogram.

For this exercise, we'll chart the distribution of `mpg` for cars in `USA` and `Japan` on the same chart.

3.5.1 Instructions

1. Create a new PivotTable, placing the following:

- `mpg` in **Rows**
- `origin` in **Columns**
- `Count of id` in **Values**

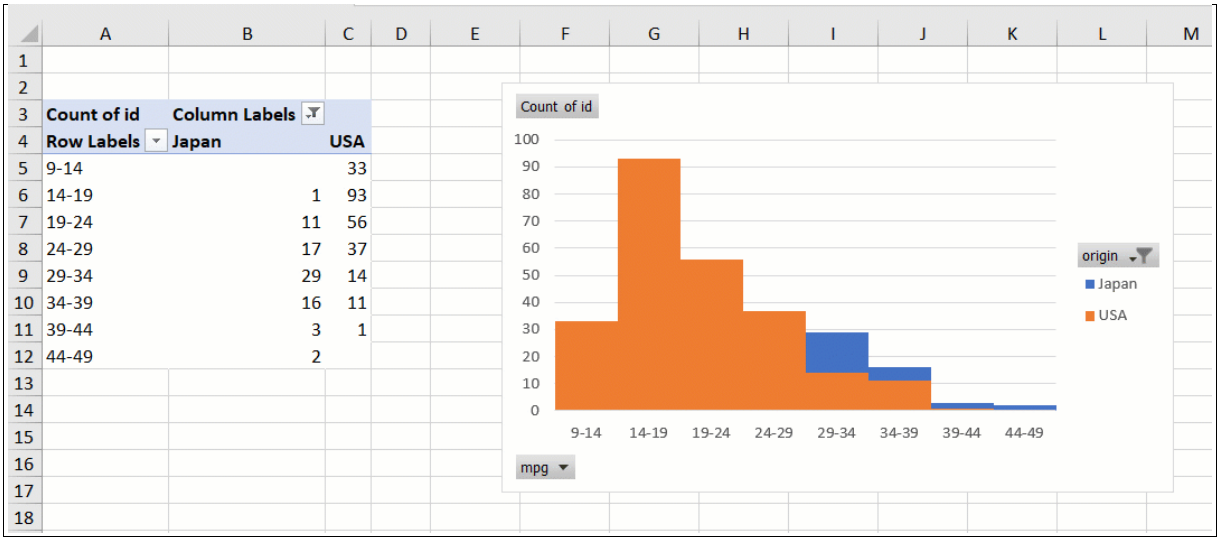
2. Filter on **Column Labels** to include only `USA` and `Japan`.
3. Right-click on any `mpg` value in the first column and select **Group**.

- In the **Grouping** window that appears, set the following:
- **Starting at** to `9` (minimum `mpg` value)
- **Ending at** to `46.6` (maximum `mpg` value)
- **By** to `5`

4. Create a PivotChart by going to **Insert > Charts > Recommended Charts > Clustered Column**.
5. Make the resulting column chart look more like a regular histogram:

- Right-click on any of the bars in the chart and select **Format Data Series**
- Set **Series Overlap** to `100%`
- Set **Gap Width** to `0%`.

6. Make sure that the column for `USA` is displayed before `Japan`. You can change the order by simply typing *USA* over *Japan*, like so:



3.5.2 Questions

1.

How many American cars get between 24 and 29 miles per gallon?

17

29

37

93

3.6 Creating a Correlation Matrix

Congratulations, you just used Excel to compare a quantitative variable against the distinct groups of a categorical variable! Now let's look at analyzing the relationship between two quantitative variables using **correlation**, specifically with the **Analysis ToolPak**.

Since **correlation** is a measure between two quantitative variables, it doesn't make sense to include categorical variables like `origin` and `car_name` in our analysis. It also doesn't make sense to include `id` since it is an arbitrary ID number.

In a previous course, we used **R-squared** to measure the strength of a correlation between two quantitative variables. The **Analysis ToolPak** gives us access to another (more robust) metric for evaluating such relationships: the **correlation coefficient**, also called the **Pearson correlation coefficient**.

Here's how to interpret the **correlation coefficient**:

- It can be any number between `-1` and `1`.
- A correlation of `-1` indicates a perfect negative correlation: as one variable increases, the other decreases.
- A correlation of `+1` indicates the opposite: a perfect positive correlation where as one variable increases, so does the other.
- A correlation of `0` indicates there is no correlation between one variable and the other.

3.6.1 Instructions

1. Return to the `mpg` table on the `data` tab.
2. Navigate to **Data > Analyze > Data Analysis > Correlation**.
3. In the **Correlation** window that appears, set the following:

- **Input Range** to include all quantitative columns (`B:H`) including their labels.

- Enable **Labels in First Row**
- **Output options** to **New Worksheet Ply**

4. Click `OK`. You will now have a **correlation matrix** consisting of the **correlation coefficients** between each quantitative variable.

3.6.2 Questions

1.

Based on the correlation coefficient alone, what happens to `mpg` as the `acceleration` of a car increases?

It decreases

Not enough information given

It increases

It doubles

3.7 From Correlation to Regression

The **correlation matrix** is a powerful way to explore relationships across many variables, but there's a catch: these relationships must be *linear* for the **correlation coefficient** to be a valid measure! This is why it's important to inspect your scatter plots for a linear relationship before using the **correlation coefficient** to interpret the relationship between two variables.

Unfortunately, Excel doesn't include an easy way to create multiple scatter plots at one time, so we'll focus on just one particular relationship: `weight` and `mpg`.

When we add the **Trendline** to our scatter plot, we will also enable the option to **Display Equation on chart**.

This equation takes the form of `y = mx + b`, where `y` is the value of the y-axis, `mx` is the value of `x` times a coefficient `m`, and `b` is the y-intercept. The y-intercept tells us what the value of `y` would be if `x` were `0`. We will explore this equation even further in the upcoming lesson on **confirmatory data analysis**.

3.7.1 Instructions

1. Return to the `mpg` table on the `data` tab, and select the `weight` and `mpg` columns.
2. Navigate to **Insert > Recommended Charts**, and select **Scatter**; it should be the first option.
3. Use **Chart Elements** (`+` icon at the upper-right of the chart) to add a **Trendline**.
4. Format the **Trendline**:

- Make sure **Trendline Options** is set to **Linear** under the **Trendline Options** tab
- Select **Display Equation on chart**

3.7.2 Questions

1.

What is the y-intercept of the equation found on your scatter plot?

-0.83

0.69

-0.0076

46.217

4 Confirmatory Data Analysis in Excel

4.1 The Law of Large Numbers

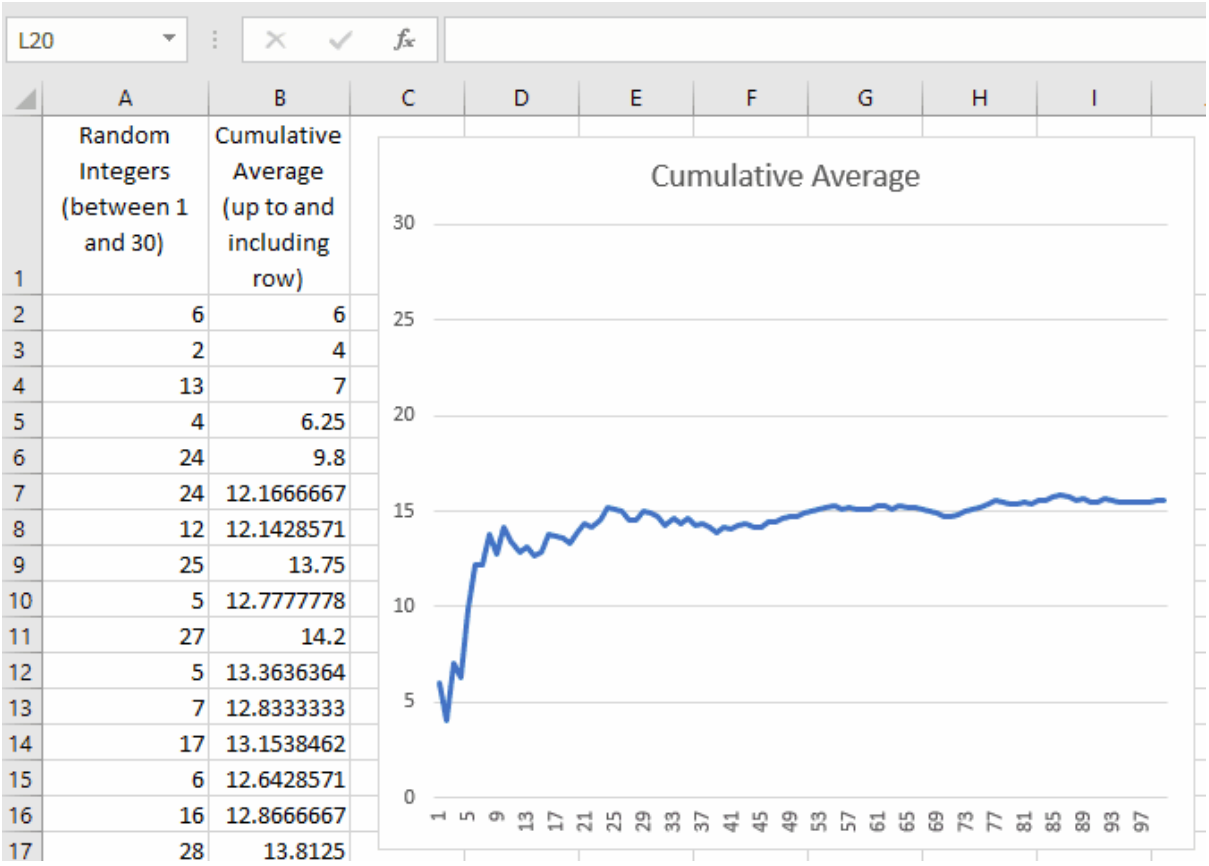
For this lesson, we'll continue working with the fuel mileage dataset we used in the previous lesson. (Data source: University of California, Irvine Machine Learning Repository).

Often, as data analysts, we want to make a claim about an entire population with only a sample of the relevant data. For example, let's say we're interested in properties of the auto market at large, but we only have a few hundred records.

Using a couple of important concepts, we can make probabilistic claims about a population with only data for a sample of the entire population. The first concept we'll look at is **the law of large numbers**.

To show you this law, we're going to need some numbers — and lots of them!
The `RANDARRAY()` function is perfect for creating random data you can experiment with. Specifically, `RANDARRAY()` is used to generate a series of random numbers between any two numbers.

For example, entering `RANDARRAY(100,1,1,30,TRUE)` in cell `A2` will generate `100` random integers between `1` and `30`. Entering `AVERAGE($A$2:A2)` in cell `B2` and using **flash fill**, will calculate the cumulative average of each random integer (i.e., the average of all integers up to and including that integer). Finally, if we create a line chart of the cumulative averages in column `B` against the number of random integers being used to calculate the cumulative average, we can see how different sets of random integers behave visually. By pressing `F9` to repeatedly generate a new random series of integers, we see an interesting pattern emerge in the line chart:



	A	B
1	Random Integers (between 1 and 30)	Cumulative Average (up to and including row)
2	6	6
3	2	4
4	13	7
5	4	6.25
6	24	9.8
7	24	12.1666667
8	12	12.1428571
9	25	13.75
10	5	12.7777778
11	27	14.2
12	5	13.3636364
13	7	12.8333333
14	17	13.1538462
15	6	12.6428571
16	16	12.8666667
17	28	13.8125

So what is happening here?

We can see the random numbers being generated in column `A` change with each press of `F9`. The left side of the line chart tends to be a bit chaotic, whereas the right side of the chart appears to want to settle on a particular number even with different random integers! As we randomly generate more and more numbers between 1 and 30, the cumulative average gets closer and closer to `15.5`. That's because the average of the numbers between 1 and 30 is `15.5`. This is the **expected value**.

A similar thing happens when we sample data from a population — we can expect that as we collect more and more samples, the **sample mean** (the average of the samples) would get closer and closer to the **population mean** (the average of the population). This is **the law of large numbers**.

4.1.1 Instructions

If you get stuck or would like to compare your answers, we've provided a suggested solution file for this lesson called `mpg-cfa-solutions.xlsx`, which you can find in the `Home Folder` under `This PC`.

1. Use the `RANDARRAY()` function and the formula for cumulative average like we did in the **Learn** section above to determine the **expected value** of rolling a six-sided die.

- Try generating `1000` random integers between `1` and `6` instead of just `100`.

4.1.2 Questions

1.

You start rolling a six-sided die repeatedly. What is its expected value?

3.5

1

6

Not enough information given

4.2 Data Distributions

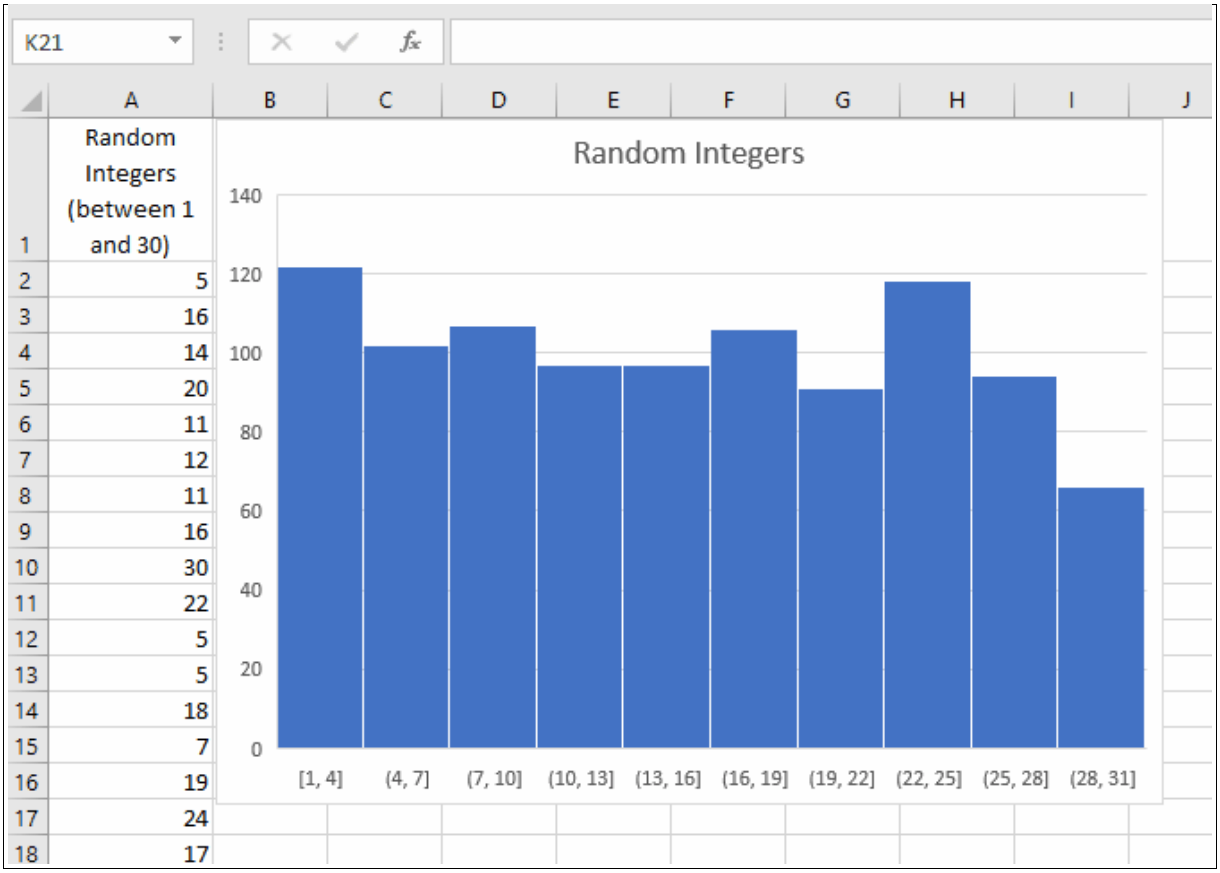
On the previous screen, we learned that due to the law of large numbers, we can expect the sample mean to approach the population mean. But of course they are *not* going to be exactly the same; there is *always* going to be some uncertainty.

Fortunately, you can quantify that uncertainty using the magic of the **Central Limit Theorem**, which we will discuss on the next screen. But first, let's take a look at the distribution of random integers.

4.2.1 Instructions

1. In a new worksheet, create a list of `1000` random integers between `1` and `30` using `RANDARRAY(1000,1,1,30,TRUE)`.
2. Visualize the distribution of these numbers with a histogram.

- Change `Bin width` to `3` by formatting the horizontal axis.
- Try regenerating the random integers a couple of times by pressing `F9`.
- Your histogram should look something like this:



4.2.2 Questions

1.

What is the distribution of the variable you just created based on its histogram?

Normal

Uniform

Right-skewed

Left-skewed

4.3 The Central Limit Theorem

The histogram you created on the previous screen does *not* resemble the classic "bell curve" shape of the normal distribution. And that's because it's not — it's a uniform distribution!

Thanks to the **Central Limit Theorem**, we know that the distribution of **sample means** will be normally distributed, regardless of the variable's distribution, provided the sample size is large enough. How large is large enough? That depends on a few things, but generally if you're working with a properly collected dataset with hundreds of observations, like our `mpg` dataset, you should be in the clear.

Why does all this matter? Keep in mind *we don't know the population mean*, and we can't calculate it because we don't have all that data; we only have data for samples of the population, not the entire population. However, we know that as we get more sample data, we can expect the **sample mean** to approach the **population mean**. Moreover, because we can reliably infer what the distribution of sampling means is, we can consistently give a probabilistic range of values that we *think* would contain the population mean.

Generally, we'll use a threshold of *95% confidence* when making this inference. We'll learn more about what that means on the next screen.

If it's still difficult to believe that the distribution of sampling means of a non-normally distributed variable would be normally distributed, let's build a demonstration in Excel to confirm it!

4.3.1 Instructions

1. In cell `A1` of a new worksheet, create a `50`x`50` matrix of random integers between `1` and `30` using: `RANDARRAY(50,50,1,30,TRUE)`.

- Think of each row of `50` random integers as a **sample**.

2. Create a column of **sample means**:

- In cell `AY1`, use the `AVERAGE()` function to find the average of the first row of `50` random integers.
- Use **flash fill** to automatically calculate the average for all remaining rows of random integers.

3. Plot the distribution of the **sample means** using a histogram.

- Try regenerating the random integers a couple of times by pressing `F9`.
- Observe the shape of the histogram to determine its distribution.

4.3.2 Questions

1.

What kind of distribution do the sample means have?

Left-skewed

Right-skewed

Uniform

Normal

2.

Which of the following is a condition for the central limit theorem to apply?

The variable must be normally distributed

The sample size must be over 1,000

The sample size must be sufficiently large.

None apply.

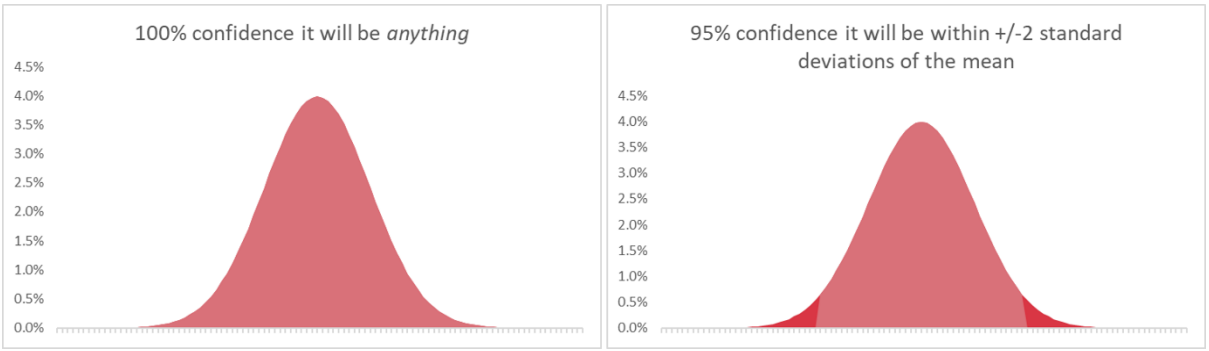
4.4 The Independent Samples T-test

We know a couple of important things so far:

- As our sample size increases, the **sample mean** will get closer to the true **population mean**.

- Our sample means will be normally distributed given a large enough sample size.

We will never truly know what will be found in the population. There's 100% chance that it can be *anything*. At the same time, we know the sample means will be normally distributed, which indicates the possible values follow a consistent shape. We can give a consistent range of values that we'd expect to find the population mean with *95% confidence*:



We'll expect this boundary to tighten up as the sample size increases and our sample mean gets closer to the expected value.

When we conduct statistical tests in most programs, including Excel, we will receive a proability value or **p-value** as a result. If the p-value is less than `0.05` (or 1 - 95%), we will conclude that what we see in our sample is likely true of the population as well.
The `0.05` threshold is often referred to as **alpha**, which we can choose during a test.

Let's try this out in Excel by using what's called a **two-sample independent t-test** to check if there is a significant difference in the average mileage of European and American cars. If the results of our t-test show that **p-value < alpha**, then we will conclude that there is a difference between the average mileage of European and American cars. Conversely, if **p-value > alpha** then we will conclude that there is no statistical difference between the two averages.

4.4.1 Instructions

1. Open the `mpg-cfa.xlsx` dataset.
2. Create a PivotTable with the following:

- `id` along **Rows**
- `Sum of mpg` inside **Values**
- `origin` along **Columns**
- `Japan` filtered out of the **Columns** results
- **Grand Totals** turned **Off for Rows and Columns**

3. Go to **Data > Analyze > Data Analysis > t-Test: Two-Sample Assuming Unequal Variances**.

4. In the **t-Test: Two-Sample Assuming Unequal Variances** window that appears:

- Select the two relevant ranges from your PivotTable for:
- **Variable 1 Range** = mileage for European cars
- **Variable 2 Range** = mileage for American cars
- Enable **Labels**
- Set **Alpha** to `0.05`
- Set **Output Range** to **New Worksheet Ply**

5. Click `OK` to generate the results of the test.

- You will find the resulting **p-value** listed next to `P(T<=t) two-tail` in the results table

4.4.2 Questions

1.

Does there appear to be a significant difference between the means at the 95% confidence level?

No

Yes

4.5 Adding Confidence Intervals

The p-value informs us how likely it is that what we see in the sample will also be true of the population. It does *not* tell us how likely that population difference is going to *be*. However, we can build a 95% **confidence interval** to find an actual range of values for the difference between the two sample means.

Unfortunately, the Analysis ToolPak doesn't include the confidence interval as part of its output, so we will build it on our own. Don't worry too much about the math here — but notice how the results form an upper and lower bound of possible values around our **point estimate** or the difference in means we found in the samples:

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*Europe*	*USA*
4	Mean	27.60294118	20.03346939
5	Variance	43.29879719	41.47854734
6	Observations	68	245
7	Hypothesized Mean Difference	0	
8	df	105	
9	t Stat	8.431121249	
10	P(T<=t) one-tail	9.83149E-14	
11	t Critical one-tail	1.659495383	
12	P(T<=t) two-tail	1.9663E-13	
13	t Critical two-tail	1.982815274	
14			
15	point estimate	-7.56947179	=C4-B4
16	critical value	1.982815274	=B13
17	standard error	0.89780132	=SQRT((B5/B6)+(C5/C6))
18	margin of error	1.780174171	=B16*B17
19	95% confidence interval lower bound	-9.34964596	=B15-B18
20	95% confidence interval upper bound	-5.78929762	=B15+B18
21			

4.5.1 Instructions

1. Enter the labels in cells `A15:A20` of the screenshot above to the cells `A15:A20` of your worksheet containing the results of the **t-Test** from the previous screen.
2. Enter the formulas found in `C15:C20` of the screenshot above to the cells `B15:B20` of your worksheet containing the results of the **t-Test** from the previous screen.

4.5.2 Questions

1.

True or false: with 95% confidence we can say that the difference is between about 9.35 and 5.79 mpg.

True

False

4.6 Bivariate Regression

Many statistical tests exist to evaluate different sorts of relationships in the data. For example, we used the independent samples t-test to check for a difference in means between two groups. We can also check the significance of one continuous variable on another using **linear regression**.

For this example, we will see whether `weight` significantly influences `mpg`. If the p-value of `weight` in our model is less than `0.05`, we will conclude with 95% confidence that `weight` significantly influences `mpg`.

4.6.1 Instructions

1. Return to the `data` tab of `mpg-cfa.xlsx`.
2. Navigate to **Data > Analyze > Data Analysis > Regression**:

- Set the `mpg` values as your **Input Y Range**, including the header
- Set the `weight` values as your **Input X Range**, including the header
- Enable **Labels**
- Select **New Worksheet Ply** under **Output options**

3. Click `OK`:

- You will now have three boxes of summary output
- The final box will include a column for `P-value`

4.6.2 Questions

1.

Does `weight` have a significant influence on `mpg`?

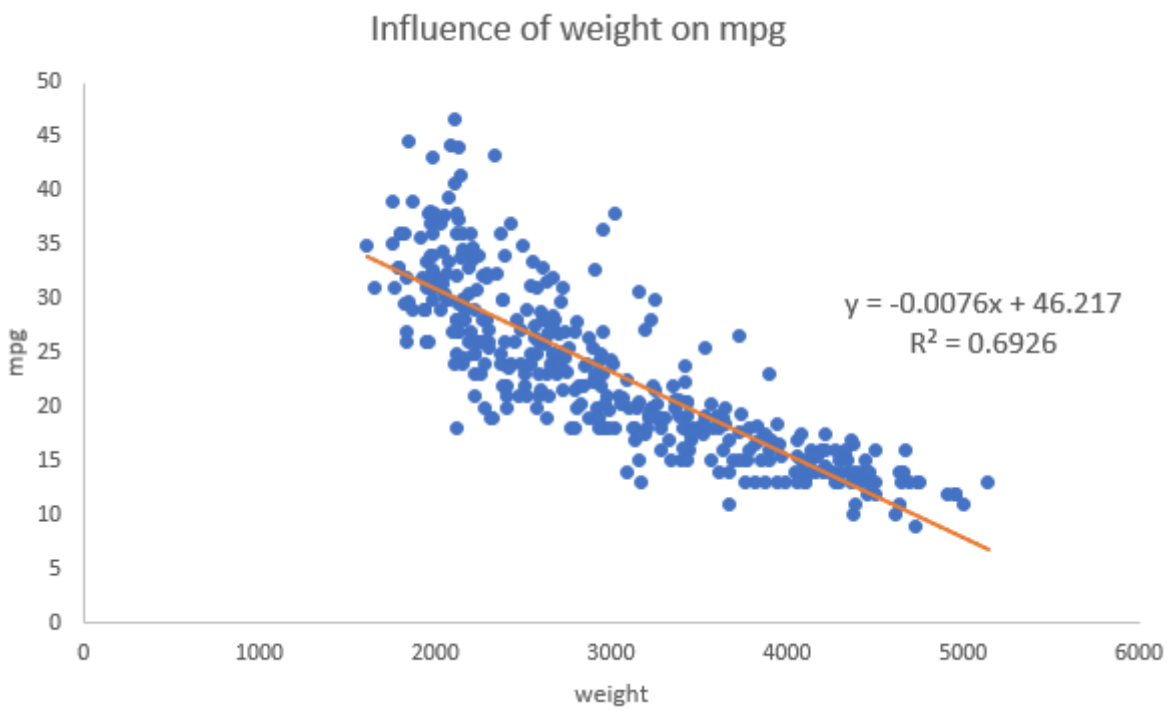
Yes

No

4.7 Evaluating Regression Model Fit

The objective of linear regression is to fit a line summarizing the relationship between the independent and dependent variables. You are familiar with this technique — we can do it in Excel by adding a trendline to a scatter plot.

To do this, we create a scatter plot with `weight` as the independent variable and `mpg` as the dependent variable. Adding a linear trendline allows us to display the **R-squared** value and **Equation** of the line on our chart, like so:



Let's interpret the output for **Equation** and **R-squared** for this chart:

The regression **Equation** provides the slope (`-0.0076`) and y-intercept (`46.217`) of the fit line. You may notice that these are the same values returned by the ToolPak on the previous screen under the `Coefficients` column! However, the Excel chart didn't return the statistical significance of the line. You can use the **Equation** to make predictions about the `mpg` of a vehicle based on its `weight`, like so:

predicted `mpg` = -0.0076 × ×`weight` + 46.217

For example, we would predict that a vehicle with a weight of `3500` would have a `mpg` value somewhere around -0.0076 × ×3500+46.217 = 19.62. Looking at our `data` tab, we can see this estimate of `mpg` is very close to the vehicle with `id`=`1` and a weight of `3504`.

As you may recall, the **R-squared** value can be interpreted as the percentage of variability in the dependent variable that can be explained by the variability of the independent variable. In this example, we can say that 69% of the variability in mileage can be explained by the vehicle's weight. Since it's a percentage, it will always be expressed as a number between `0` and `1`. This value is available both on the chart and as output from performing **Regression** using the ToolPak.

4.7.1 Instructions

1. Create a scatter plot with `horsepower` as the independent variable and `mpg` as the dependent variable:

- Copy `horsepower` and `mpg` to a new worksheet while making sure the `horsepower` column is to the left of `mpg`.
- Insert a scatter plot, and use **Chart Elements** to add a **Trendline**.
- Enable the options **Display Equation on chart** and **Display R-squared value on chart**.

2. Use the **Equation** to predict the `mpg` for a vehicle with `horsepower`=`150`.

4.7.2 Questions

1.

True or False: you can interpret the R-squared value for `mpg` vs `horsepower` as saying 61% of the variability in mileage can be explained by the vehicle's horsepower.

False

True

2.

What do you estimate the `mpg` would be for a vehicle with `horsepower`=`150`?

22.75

23.446

19.62

16.266

5 Business and Financial Modeling in Excel

5.1 What Percent of Claims Are Disputed?

Nowhere is the expression "failing to plan is planning to fail" more true than in business. In this lesson, we'll look at some helpful tools in Excel for modeling business scenarios and measuring the sensitivity of business metrics to various inputs.

We'll be using accounts receivable data from IBM (source: Kaggle) to assess various strategies and outcomes related to gathering late payments. The data is in the file `accounts-receivable.xlsx`.

Here is a description of each variable:

`countryCode`: unique identifier for country
`customerID`: unique identifier for customer
`invoiceNumber`: unique identifier for invoice
`InvoiceDate`: date that invoice was generated
`DueDate`: date that payment for invoice was due
`Disputed`: was the invoice disputed by the customer?
`SettledDate`: date that the invoice was settled (disputed or not)
`DaysToSettle`: how many days did it take to settle the invoice from generation to payment?
`InvoiceAmount`: how much was the invoice?
`DaysLate`: how many total days late was the invoice?

The data consists of 2,466 invoices, some of which are disputed and some are paid. Let's begin by assessing the prevalence of disputed invoices.

5.1.1 Instructions

If you get stuck or would like to compare your answers, we've provided a suggested solution file for this lesson called `accounts-receivable-solutions.xlsx`, which you can find in the `Home Folder` under `This PC`.

1. Open the file `accounts-receivable.xlsx`.
2. Create a PivotTable, placing the following:

- `Disputed` along **Columns**
- `DueDate` along **Rows**
- `Count of invoiceNumber` inside **Values**.

3. Display `Count of invoiceNumber` as a **% of Row Total**.

5.1.2 Questions

1.

What percent of all invoices are disputed, rounded to the nearest whole number?

77%

23%

20%

80%

5.2 Is the Average Disputed Invoice Significantly Different?

At almost `23%` of all invoices, disputed invoices make up enough of a proportion to warrant a closer examination. Let's look at whether or not the typical disputed invoice differs from an undisputed one by using a two-sample t-test.

5.2.1 Instructions

1. Create a PivotTable with **Grand Totals** off, placing:

- `invoiceNumber` along **Rows**
- `Disputed` along **Columns**

- `Sum of InvoiceAmount` inside **Values**

2. Use the **t-Test: Two-Sample Assuming Unequal Variances** from the ToolPak, at the 95% confidence level (`alpha`=`0.05`), to determine if there is a significant difference in the average invoice amount between disputed (`Yes`) and undisputed (`No`) invoices.

5.2.2 Questions

1.

Is there a significant difference between the average disputed and undisputed invoice?

Not enough information

No

Yes

5.3 How Much Later Are Disputed Payments?

So far, we have determined that disputed invoices make up a decent percentage of all invoices (`22.75%`) and that they are greater on average than undisputed ones (`$65.50` vs `$58.25`). Let's quantify the cost of disputed invoices by calculating the average number of additional days it takes to settle disputed invoices compared to undisputed invoices.

5.3.1 Instructions

1. Create a PivotTable to find the average of `DaysToSettle` for disputed vs undisputed invoices.

5.3.2 Questions

1.

How many more days does it take to settle a disputed invoice on average, rounded to the nearest whole number?

13

36

7

24

5.4 Building a One-Way Data Table

Based on their prevalence and relatively high value, it appears that dealing with disputed invoices could be a significant expense. Let's build a model using Excel's **Data Tables** to see how changing the relative mix of disputed invoices might change this total overall cost.

We'll assume the following (roughly based on the data):

- 250 invoices are generated every quarter
- 25% of them are disputed
- It takes about 13 extra days to settle those that are disputed
- There's about $1.00 in variable costs per day for disputed claims
- There is a fixed cost of about $2,000 to resolve disputed invoices

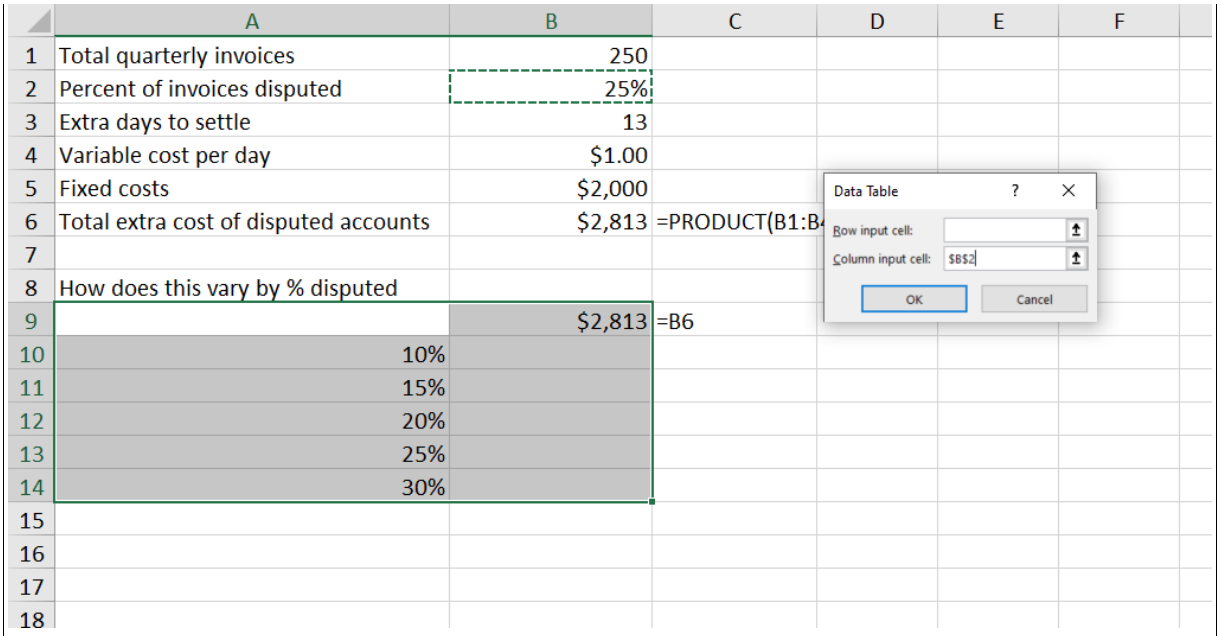
We begin by creating a new worksheet with these inputs and then multiplying them together using the `PRODUCT()` function to find the `Total extra cost for disputed accounts` per quarter, like so:

	A	B	C	D
1	Total quarterly invoices	250		
2	Percent of invoices disputed	25%		
3	Extra days to settle	13		
4	Variable cost per day	$1.00		
5	Fixed costs	$2,000		
6	Total extra cost of disputed accounts	$2,813	=PRODUCT(B1:B4)+B5	
7				

Under the model, we place a series of percentages ranging from 10% to 30% in increments of 5%. The values to the right of these percentages will be populated by Excel's **Data Table**. We will place a cell reference to `B6` above these cells, like so:

	A	B	C	D
1	Total quarterly invoices	250		
2	Percent of invoices disputed	25%		
3	Extra days to settle	13		
4	Variable cost per day	$1.00		
5	Fixed costs	$2,000		
6	Total extra cost of disputed accounts	$2,813	=PRODUCT(B1:B4)+B5	
7				
8	How does this vary by % disputed			
9		$2,813	=B6	
10	10%			
11	15%			
12	20%			
13	25%			
14	30%			
15				
16				

Next, we select the cell range `A9:B14` and navigate to **Data > Forecast > What-if Analysis > Data Table**. Under **Column input cell**, click on cell `B2` like so:



The resulting table shows us how the `Total extra cost of disputed accounts` changes as the `Percent of invoices disputed` variable changes by 5% increments. Since the table is showing how the final result changes as *one variable* changes, we call this a **one-way Data Table**.

5.4.1 Instructions

1. Create the **Data Table** exactly as described in the **Learn** section above.
2. Try changing a couple of inputs to the **Data Table**:

- Change the extra days to settle from `13` to `15`
- Change the variable cost per day from `$1.00` to `$0.50`

5.4.2 Questions

1.

What is the total cost of resolving disputed invoices when the following are true:

- `15%` **of invoices are disputed**
- **It takes** `15` **extra days to settle disputed invoices**
- **Variable cost per day is** `$0.50`**?**

$2,488

$2,563

$2,244

$2,281

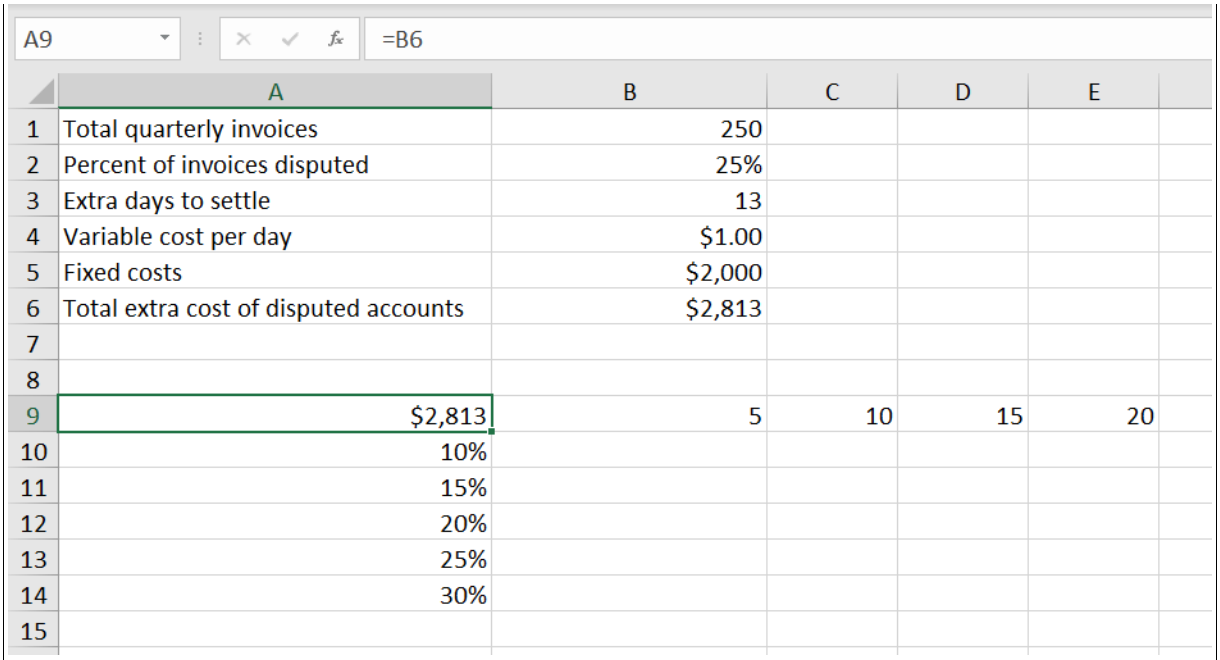
5.5 Building a Two-Way Data Table

On the previous screen, we saw that changes in the percentage of disputed invoices did not change the overall cost structure too much. But this is not our only input that could change. Let's build a **two-way Data Table** to determine how sensitive total costs are to changes in both the overall percentage of disputed invoices and the number of additional days required to settle them.

Let's begin by creating a copy of our **one-way Data Table** worksheet from the previous screen. Next, we do the following:

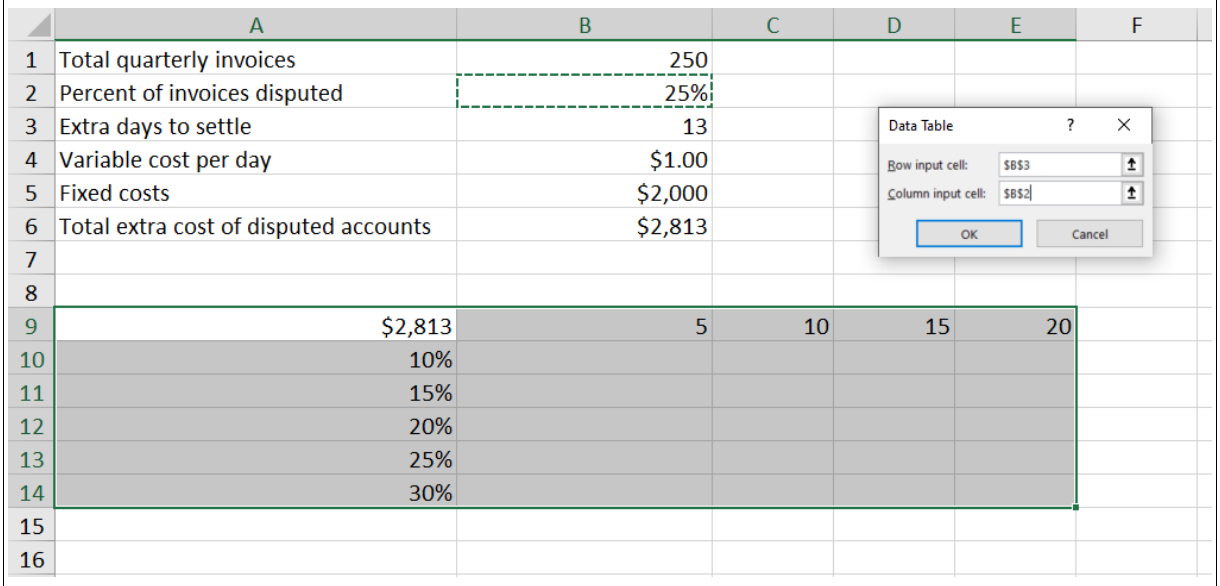
1. Clear out the data in cells `B9:B14`
2. Place the numbers `5` through `20` in increments of `5` along cells `B9:E9`
- These cells may require the use of **Format Cells** to standardize their appearance
3. Refer cell `A9` back to cell `B6`
4. Format the range `B10:E14` as a **Currency** with no decimal places

Our worksheet will look like this:



	A	B	C	D	E
1	Total quarterly invoices	250			
2	Percent of invoices disputed	25%			
3	Extra days to settle	13			
4	Variable cost per day	$1.00			
5	Fixed costs	$2,000			
6	Total extra cost of disputed accounts	$2,813			
7					
8					
9	$2,813	5	10	15	20
10	10%				
11	15%				
12	20%				
13	25%				
14	30%				
15					

Now, we select the range `A9:E14` and go back to **Data > Forecast > What-if Analysis > Data Table**. Select cell `B3` as the **Row input cell** and `B2` as the **Column input cell**.
Our **Data Table** will look like this:



	A	B	C	D	E	F
1	Total quarterly invoices	250				
2	Percent of invoices disputed	25%				
3	Extra days to settle	13				
4	Variable cost per day	$1.00				
5	Fixed costs	$2,000				
6	Total extra cost of disputed accounts	$2,813				
7						
8						
9	$2,813	5	10	15	20	
10	10%					
11	15%					
12	20%					
13	25%					
14	30%					
15						
16						

The resulting table shows us how the `Total extra cost of disputed accounts` changes as the `Percent of invoices disputed` and the `Extra days to settle` variables change by 5 unit increments. Since the table is showing how the final result changes as *two variables* change, we call this a **two-way Data Table**.

5.5.1 Instructions

1. Create the **Data Table** exactly as described in the **Learn** section above.
2. Try changing a couple of the inputs in your **Data Table**:

- in cell `A10`, change `10%` to `1%`
- in cell `B9`, change `5` to `1`

5.5.2 Questions

1.

With all other input variables remaining the same, what is the total overall cost assuming it takes `1` additional day to settle and `1%` of all invoices are disputed?

$2,013

$2,003

$2,125

$2,025

5.6 Finding a Break-Even Point with Goal Seek

Analysts often help business leaders meet quantifiable business objectives. **What-If Analysis** tools can help here. In this example, we'll use **Goal Seek** to determine what is necessary to meet an accounts receivables turnover goal.

Accounts receivable turnover is a metric calculated by dividing total revenue by the average of accounts receivable. A higher turnover is desirable because it means that funds are being collected from customers more quickly.

Let's create a simple model like the one below, by dividing a total revenue of $500,000 by an average accounts receivable of $74,000:

	A	B	C
1	Total revenue	$500,000	
2	Average accounts receivable	$74,000	
3	Receivables turnover	6.756757	=B1/B2
4			
5			

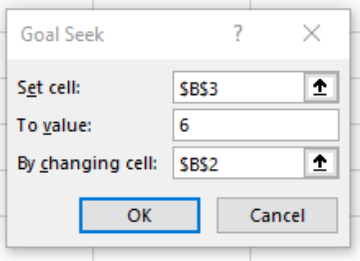
Our current accounts receivables turnover is just about `6.76`. Management would like to know what average of accounts receivable would result in an accounts receivables turnover of `6`. **Goal Seek** allows us to calculate this quickly!

If we go to **Data > Forecast > What-if Analysis > Goal Seek**, we set the following:

- **Set cell**: `$B$3`
- **To value**: `6`
- **By changing cell**: `$B$2`

Here is what that looks like:

	A	B	C	D	E
1	Total revenue	$500,000			
2	Average accounts receivable	$74,000			
3	Receivables turnover	6.756757			
4					
5					
6					
7					
8					



We find that we need an average accounts receivable of around $83,333 to meet the turnover goal of `6`.

5.6.1 Instructions

1. Use **Goal Seek** exactly as demonstrated in the **Learn** section above.
2. Type over the formula in cell `B3` with the current value of that cell, `6`.
3. Change the formula used in the model so that `Total revenue` equals `Average accounts receivable` multiplied by `Receivables turnover`. The only formula should now be in `B1` instead of `B3`.

5.6.2 Questions

1.

Rounded to the nearest whole number, what `Receivables turnover` would be necessary to reach a `Total revenue` of $750,000, given an `Average accounts receivable` of $83,333?

$125,000

$83,333

Not enough information given

9