EXPLORING DATA IN EXCEL

1 Introduction to Descriptive Statistics

1.1 Introduction

In the previous course, we learned to design data visualizations that allowed us to gather insights or identify patterns in our data visually.

But, visualizations are not the only facet of Exploratory Data Analysis (EDA).

How would you calculate how many hours, on average, did learners spend on your favorite online course last week? Or, how many people from France watched the latest Netflix Original so far?

In this course, we will learn to quantitatively summarize our data using **descriptive statistics** to answer such questions. Starting with this lesson, we will discuss what that term means and play around with some of the more basic statistics. But before we jump into that, let's look at the dataset we will be working with!

Throughout this course, we will use a modified version of the Customer Personality Analysis dataset available on Kaggle. Here are the columns you will find in the spreadsheet:

- `ID`: Customer's unique identifier.
- `Age`: Customer's age.
- `BMI`: Customer's body mass index (BMI).
- `Education`: Customer's education level.
- `Marital Status`: Customer's marital status.
- `Income`: Customer's yearly household income.
- `Date Customer`: Date of customer's enrolment with the company.
- `Customer Length`: Number of days a customer was enrolled with the company.
- `Amount Fruit`: Amount spent on fruits in the last 2 years.
- `Num Deals Purchased`: Number of purchases made with discounts.
- `Accepted Campaign`: 1 if customer accepted the offer in the 1st campaign, 0 otherwise.

As a first step, we will take a quick glance at the data and see what we can understand about it.

1.1.1 Instructions

1. Open the `marketing_campaign.xlsx` dataset in Excel and explore it.

1.2 What are Descriptive Statistics?

There are a total of `2197` rows in our dataset. With some relative ease, we can scroll down the spreadsheet and identify this number or we could use `COUNT()` instead.

Counting rows might not seem important, but what if we had tens of thousands of rows? Or more?!

It's a simple and effective way to summarize our dataset that informs us of the number of data points we are working with in a column. We just calculated our first **descriptive statistic**!

Descriptive statistics are a way to summarize or aggregate larger datasets into a single, meaningful number. As the name suggests, they help us *describe* datasets in more meaningful and interpretable ways.

For example, this Kaggle dataset contains information on more than `100,000` patients! We can't always "eyeball" real-life data. But, we can calculate some statistics to describe it, such as:

- The average age of a patient.
- The age distribution for all patients.
- How many patients had diabetes etc.

In our dataset, we can calculate the maximum value in the `Age` column using `MAX()`.

```
=MAX(B2:B2198)
```

ExplainCopy

We now know that the oldest customer(s) was `81` years old. On its own, the number doesn't inform us of much. However, it still presents us with a potential customer segment we could consider in our analysis down the line. For example, would the company benefit from marketing their product to 80-year-olds? What about 80-year-olds with an income of more than `50,000`?

We don’t need immediate answers to such questions. What’s important is that these statistics allow us to ask such questions and explore our data in-depth.

One important point to note is that `Age` is a *quantitative* variable and so the maximum value of the column is interpretable. The same is not true for some categorical variables as we will see below.

1.2.1 Instructions

1. Using MAX() in Excel, answer the following questions.

1.2.2 Questions

1.

What is maximum value for the `Marital Status` column?

None of the above

Single

Together

2.

What are the largest number of purchases made with a discount?

7

15

19

3.

What is the highest `BMI` value?

17.9

32.3

30.4

1.3 Descriptive Statistic: Minimum

As pointed out in the previous screen, we can't calculate the maximum value for every variable. That's why Excel outputs a `0` when we use `MAX()` on the `Marital Status` column. It's a categorical (nominal) column. There's no inherent order to being **Single** or **Married** so there is no need for a maximum value.

Let's look at another summary statistic; the minimum value.

`MIN()` allows us to calculate the minimum value in a column. Similar to `MAX()`, on its own, it might not offer much insight.

```
=MIN(B2:B2198)
```

ExplainCopy

So, the youngest customer(s) in our dataset is `25` years old. Because of these two summary statistics we now know that our customer ages range from `25` to `81`. This doesn't tell us how spread out the ages really are, but we will come back to that in a different lesson.

1.3.1 Instructions

1. Using MIN() or MAX() in Excel, answer the following questions.

1.3.2 Questions

1.

Does the `Education` column have a meaningful minimum value?

Yes, because it's a qualitative variable.

No, because it's a quantitative variable.

No, because it's a qualitative variable.

2.

What is the lowest `BMI` value?

32.3

0

17.9

3.

What is the range of values in `Income`?

[1730, 666666]

[2447, 157243]

[0, 52178]

1.4 Descriptive Statistic: Mean

In the previous screen, the minimum value for the `Education` column was `0`. Just as we saw with `Marital Status`, `Education` is also a categorical column and has no minimum value.

So far, we have looked at the extremes, the upper and lower limit of our columns that can describe the distribution of our data in each such column. At times, we often need a single number that is representative of all the values in a single column.

This descriptive statistic, one you are already familiar with, is called the **mean**. The **mean**, commonly known as the *average* value, is the sum of all values in a given column divided by the number of values in that column.

The mean value can help give us a sense of the typical value in a given distribution. For example, what is the average age of our customers?

```
=AVERAGE(B2:2198)
```

ExplainCopy

The average or mean age of our customers is `52`. That's another useful summary statistic that can give us a sense of the typical value of a distribution.

We will calculate the averages of a few columns next.

1.4.1 Instructions

1. Using AVERAGE() in Excel, answer the following questions.

1.4.2 Questions

1.

What is the average amount customers have spent on fruits? Choose the closest number.

42.35

78.13

26.24

2.

What is the average number of purchases made with discounts by customers? Choose the closest number.

8

2.33

11.21

3.

What is the average of `ID`? Choose the closest number.

0

2481

5600.16

2 Diving Deeper with Descriptive Statistics

2.1 Measure of Central Tendency: Mean

In the previous lesson, we learned about descriptive statistics. One of those statistics was the **mean**, also known as the **average**, which is one of the most common ways to summarize a column of data.

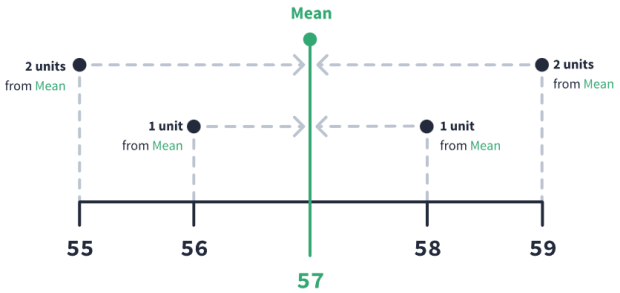
Let's say we have the following list of values containing the ages of some of our customers:

```
[55, 56, 57, 58, 59]
ExplainCopy
```

The average age of the customers would be easy enough to calculate:

$$\frac{55+56+57+58+59}{5} = 57$$

The average age, given the above data, is 57 years old. A visual representation of this would be as follows:



Our calculated mean is right in the center. It tells us where our data is centered.

But, isn't our data above too "perfect"? Ages are increasing in increments of one. We definitely know real-world data is never that simple to work with.

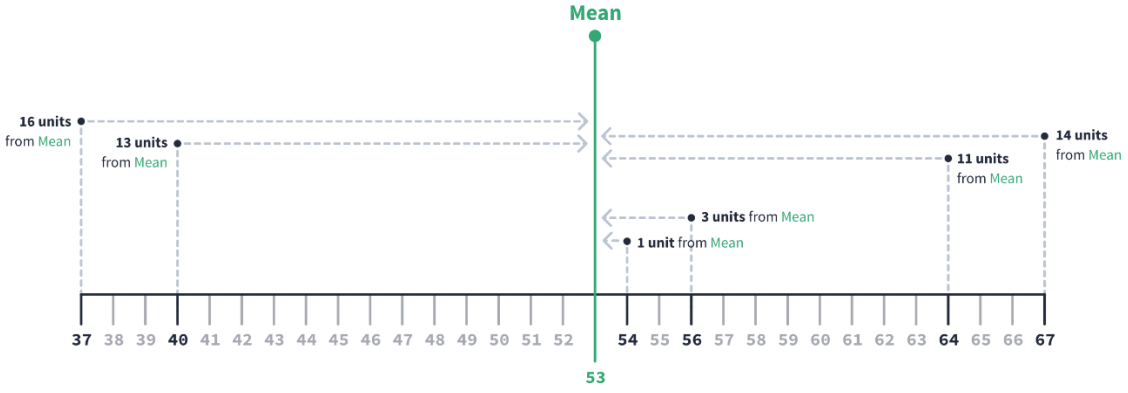
Let’s continue to work with the same dataset that we used in the previous lesson. It’s a modified version of the Customer Personality Analysis dataset available on Kaggle. Here are the columns you'll find in the spreadsheet:

- `ID`: customer's unique identifier
- `Age`: customer's age
- `BMI`: customer’s body mass index (BMI)
- `Education`: customer's education level
- `Marital Status`: customer's marital status
- `Income`: customer's yearly household income
- `Date Customer`: date of customer's enrollment with the company
- `Customer Length`: number of days a customer was enrolled with the company
- `Amount Fruit`: amount spent on fruits in last 2 years
- `Num Deals Purchased`: number of purchases made with a discount
- `Accepted Campaign`: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

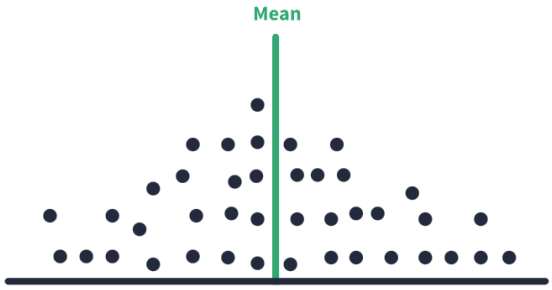
Here are the first six rows of the `Age` column:

```
[64, 67, 56, 37, 40, 54]
```
ExplainCopy

The mean age of our customers given just these values is `53` years old. This is how it would look:



The mean isn't quite in the center anymore, and it can vary depending on the distribution of our data. It's a **measure of central tendency** that gives us a *typical* value, or an *estimate*, for where most of the values in the data are clustered.



We'll discuss **measure of central tendency** further in the upcoming screens.

2.1.1 Instructions

1. Open the file `marketing_campaign.xlsx` in the `Home Folder` under `This PC`.
2. Answer the following questions.

2.1.2 Questions

1.

What is the mean income?

`348.1219845`

`666666`

`52178.251`

`57817`

2.

What is the minimum income?

`666666`

`0`

`2197`

`1730`

3.

What is the maximum income in the dataset?

`687`

`2198`

`52178`

`666666`

2.2 Impact of Outliers on the Mean

On the previous screen, we looked at the `Income` column:

- **Minimum**: `1730`
- **Maximum**: `666666`

- **Mean**: `52178.25171`

Let's say we wanted to target customers for a marketing campaign based on their income. At least one of them has an income of more than `666000`! In fact, if we were to explore the column more, we would notice that the second-highest value in `Income` is `162397`. It's significantly lower than the maximum, thus implying that the maximum value is an *outlier*. We encourage you to confirm this yourself as well using a data visualization.

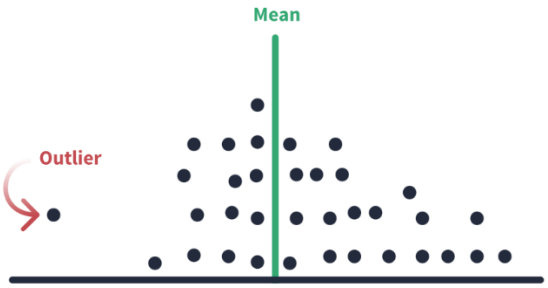
Does the calculated average seem like a typical value that describes `Income`?

Let's quickly calculate the average again after removing the maximum value. For the sake of simplicity, we're assuming that only one customer has such a high income.

For a total of `2197` customers, the average income, if we don't consider the maximum value, would be:

$$\frac{52178.25171*2197-666666}{(2197-1)} = 51898.43$$

The mean value decreased! The difference might seem insignificant, but it's important to note the above change is for only one outlier. For even larger (or much smaller) outliers, our mean may no longer be representative of our data. In the visual below, we can see how the mean will shift to the right if we were to remove the outlier.

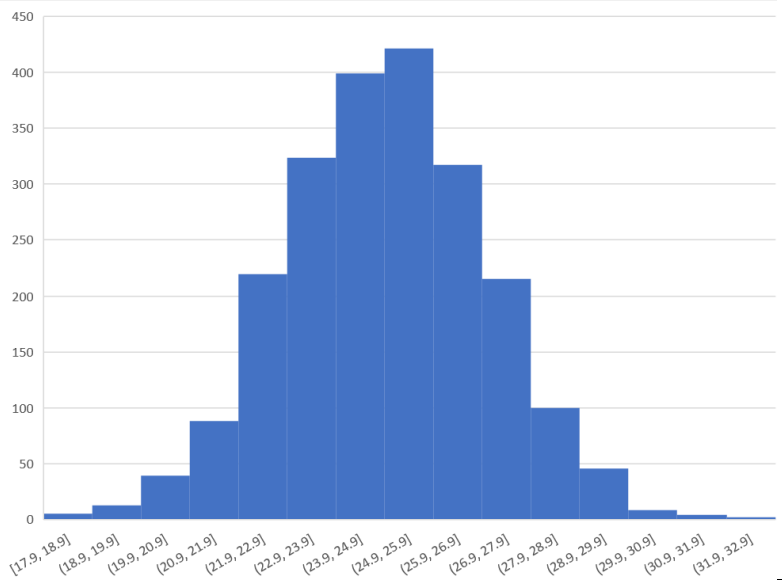


If a handful of customers in our data were executives earning millions of dollars a year, the average income might not accurately describe the rest of the customers. In such a situation, we could no longer recommend using the average value to make a business decision for the majority of the customers.

The mean isn't always the robust summary statistic that we would like. Also, as we have seen through some of the visualizations so far, the average value alone might not give us the entire picture either.

There are alternatives to the mean that we'll explore on the next screen.

2.2.1.1 Instructions



1. Refer to the image above, and answer the following questions.

2.2.1.2 Questions

1.

The histogram shown in the `Instructions` represents a column of data with a mean of about `25`. Based on just a visual check, do you think the mean gives us a typical value for the distribution?

No.

Yes.

2.

Calculate the maximum, minimum, and average number of fruits sold in the past two years. Given those values, can we be certain that the mean gives us a typical value for that column of data?

No.

Yes.

3.

What happens to the mean if you recalculate it after removing a really small outlier?

The mean would increase.

The mean would decrease.

2.3 Measure of Central Tendency: Median

We learned that a small set of outliers can skew the average, which can affect our analysis. A measure of central tendency that is more suitable in such situations is the **median**.

The **median** is the middle value in a ***sorted*** list of values.

Let's look at our sample customer ages from before:

```
[55, 56, 57, 58, 59]
```

ExplainCopy

Fortunately, the above list is already sorted. The middle value is `57`, and that would be our median. If we want consider some customer ages from our dataset like before . . .

```
[64, 67, 56, 37, 40]
```

ExplainCopy

. . . we would first have to sort the above:

```
[37, 40, 56, 64, 67]
```

ExplainCopy

The middle value, the median, is `56`.

Once again, we almost fell prey to some “simple” data in our example. We only looked at how to find the median when we have an *odd* number of values. What would the median be when we have an *even* number of data points?

The process is straightforward. We first sort the values, and then, we calculate the *average* of the middle **two** values.

Let’s try out some simple problems to test our understanding so far.

2.3.1 Instructions

1. Answer the following questions.

2.3.2 Questions

1.

What is the median value for the `Income` column? You can use Excel’s MEDIAN() function.

```
51717
```

```
52178
```

`51898`

`45561`

2.

What is the median value of `[19, 7, 5, 7, 0, 6]`?

`6`

`6 and 7`

`7`

`6.5`

3.

What is the median value of `[80011, 20559, 21994, 7500, 7500]`?

`21994`

`27513`

`20559`

`7500`

2.4 Robustness of Median

On the previous screen, we calculated the median for `Income`: `51717`. Our calculated mean for the same column was `52178.25171`. So, the median is lower than the mean.

Why would that be the case?

As we saw earlier in this lesson, a single outlier of `666666` was affecting our average value. When we removed the outlier, the mean decreased to `51898.43`. Our median value is closer to this updated mean.

Median is a more robust measure of central tendency when our data contains outliers. That's because the median, unlike the mean, doesn't take into account all the values in your data column. It only requires the one(s) in the middle, effectively ignoring any outliers.

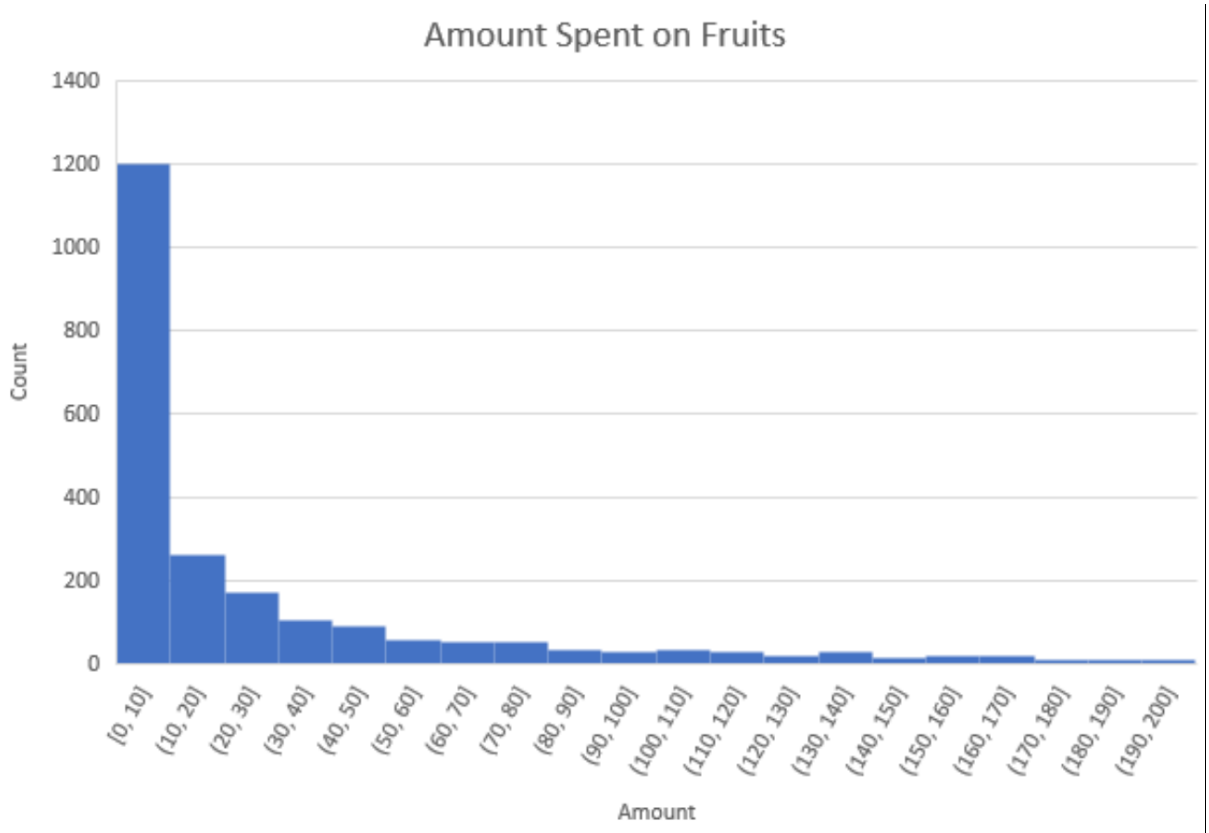
Without outliers, the mean and median values might be similar. But, that's not always the case.

Earlier, we questioned whether the mean of the `Amount Fruit` column was a typical value for the data or not. When we looked at the minimum and maximum values, it didn't seem to be the case. If we calculate the median for that column . . .

```
=MEDIAN(I2:I2198)
```

ExplainCopy

. . . the median is `8`, while the mean is about `26`. That's a significant difference! We might consider attributing the difference to outliers, but that's not necessarily the case every time. The distribution of our data also influences these values. Here is a histogram for the `Amount Fruit` column:



The distribution is right-skewed. The mean of `~26` wouldn't really represent the above distribution, but the median of `8` does when we look at where the bulk of our data is clustered. So, knowing both the mean and the median values helps better summarize our data. We'll discuss this more in a later lesson.

2.4.1 Instructions

1. Answer the following questions.

2.4.2 Questions

1.

What is the median value for `BMI`?

`25`

`17.9`

`24.97`

`18`

2.

The median takes all values in a data column into account, but the mean only requires the middle value(s).

False

True

3.

What is the median value for `Customer Length`?

`313`

`348`

`687`

`348.12`

4.

Would you say that the mean and median `BMI` are about the same?

Yes.

No.

2.5 Measure of Central Tendency: Mode

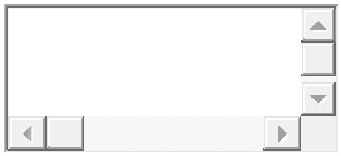
There is one final measure of central tendency that we will discuss: the **mode**.

The **mode** is the most commonly occurring value in a column of data. Let's consider the following list of values:

[5, 2, 2, 9, 9, 9, 1, 3]
ExplainCopy

In the above list, 9 appears three times and would therefore be the mode. It is possible for us to have no mode, a single mode, or multiple modes in a column of data.

The mode is useful in describing a column that contains many repeating values. Let's say we asked our customers to rate our product from 1 to 5, which resulted in the following list of values:



[3, 3, 5, 4, 4, 4, 1, 1, 2]
ExplainCopy

The mode for the above data would be 4. Based on this information, we could communicate that our product is most commonly rated a 4. While technically correct, that statement doesn't necessarily present us with the entire picture.

The most common rating is 4, but the mean (and median) is 3. That difference can affect our perception of customer ratings and any subsequent product decisions.

Let's look at a few questions about mode before moving on.

2.5.1 Instructions

1. Answer the following questions.

2.5.2 Questions

1.

What is the mode for `[19, 27, 18, 9, 8, 12]`?

`18`

No mode.

`8`

`15.5`

2.

What is the most common age of our customers? You can use Excel's MODE() function.

`45`

`51`

`50`

`52`

3.

What is the mode for `[1, 1, 1, 3, 4, 4, 2, 3, 6, 3]`?

(select all that apply)

- [ ] 4
- [ ] 1
- [ ] 3
- [ ] No mode

2.6 Measure of Spread: Range

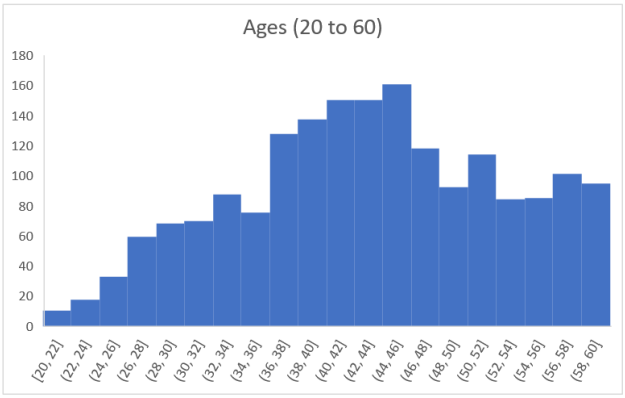
We have learned about some summary statistics, referred to as **measures of central tendency**, which give us a typical value for a distribution. Next, we will look at some summary statistics that measure how much our data values are spread out from one another.

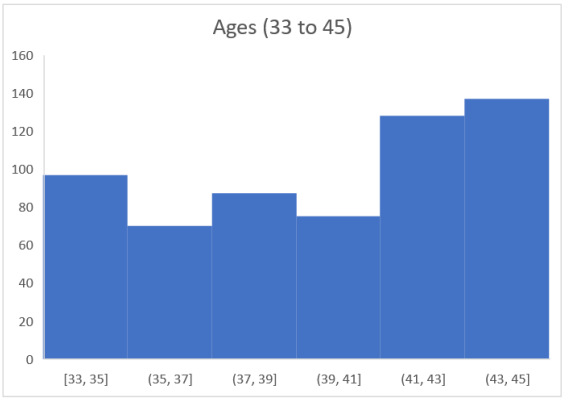
These statistics are referred to as the **measures of spread**.

The first measure of spread we'll learn about is the **range**. Commonly, we would think of **range** based on the minimum and maximum values of a given column of data. However, we calculate the range as a single value because it's a summary statistic:

$$\text{range} = \text{max\_value} - \text{min\_value}$$ range=max_value−min_value

If the maximum age of our customers was 80, and the minimum was 20, the range would be 60. If the maximum age of our customers was 45, and the minimum was 33, the range would be 12.





We can see how the two numbers above give us an idea of how spread out the data can be. If the data was spread out, the range value would be larger. If the data values were closely clustered, the range value would be smaller.

2.6.1 Instructions

1. Answer the following questions.

2.6.2 Questions

1.

What is the range of income of our customers?

664936

45129

51717

666666

2.

What is the range of income of our customers if we remove the outlier (666666)?

51684

```
48154
```

```
160667
```

```
664936
```

3.

Is the range sensitive to outliers?

Yes.

No.

2.7 Measure of Spread: Standard Deviation

One of the most commonly used measures of spread is the **standard deviation**. It tells us how much the values in a given column of data deviate from the mean of that column.

Let's look at an example to understand this better. The following list contains the number of purchases that were made with a discount:

```
[3, 2, 1, 2, 5]
ExplainCopy
```

The mean for the above is `2.6`. We want a single value that describes how far each value in the list above is from `2.6`. These are the steps we take to calculate this:

1. For every value, subtract the mean, and square the result.

$$[(3-2.6)^2, (2-2.6)^2, (1-2.6)^2, (2-2.6)^2, (5-2.6)^2]$$

$$= [0.16, 0.36, 2.56, 0.36, 5.76]$$

2. Add all of the values obtained in the previous step and divide them by **(n− 1)** where **n** is the number of values

$$(0.16 + 0.36 + 2.56 + 0.36 + 5.76)/4 = 2.3$$

3. Take the square root of that average.

$$\sqrt{2.3} = 1.5165$$

There is more than one way to calculate the standard deviation. For this course, we will use the STDEV function in Excel which calculates standard deviation using the above approach.

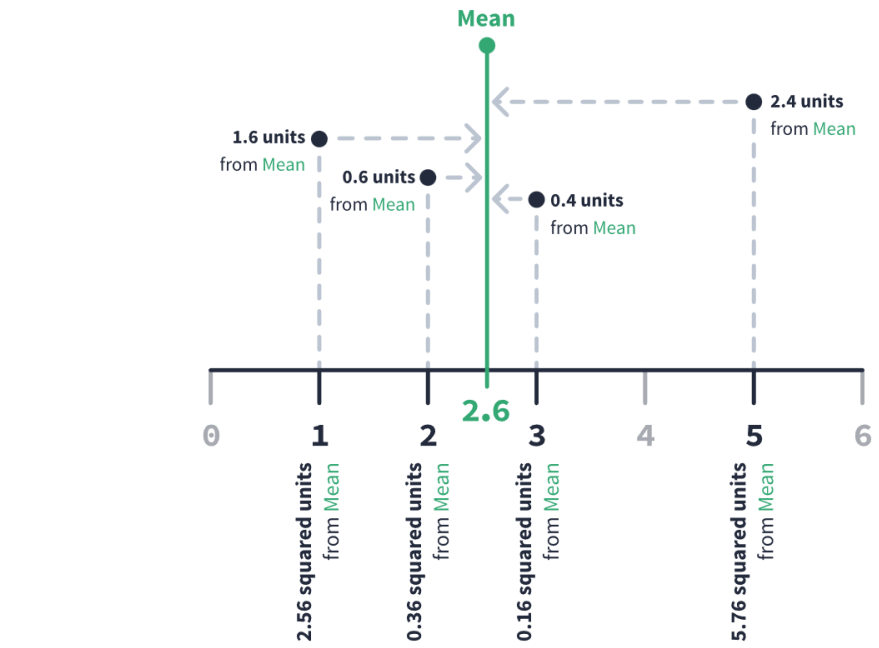
The standard deviation for the data values given above is `1.5165`. But, how do we interpret this?

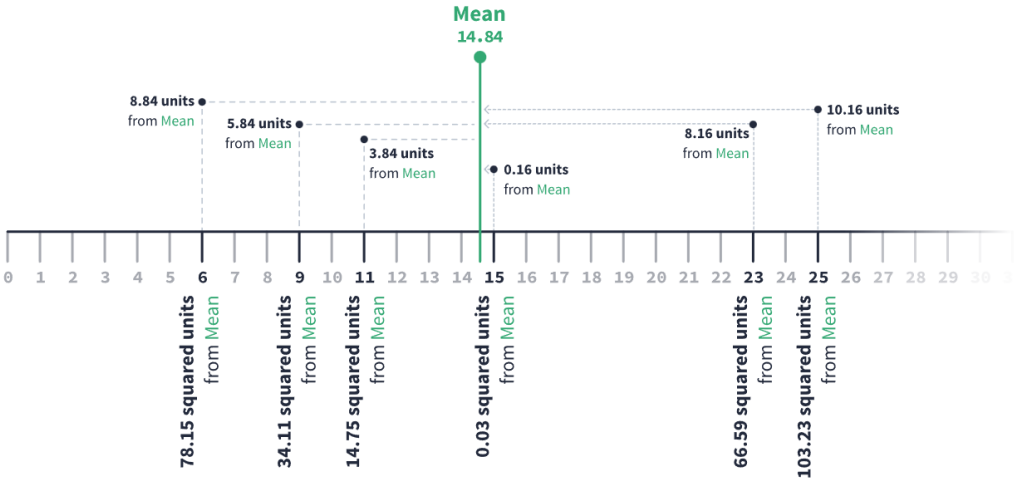
Let's take another list of values:

[15, 9, 11, 23, 6, 25]
ExplainCopy

The standard deviation for the above, which we encourage you to calculate yourself, is `7.705`. And the mean is about `14.83`. That's higher than the value we saw before.

The visuals below can help us better intuit what's happening by looking at these values and their distance from the corresponding means.





In the first image above, the square of the difference between the data point `5` and the mean is `5.76`. In the second image, the square of the difference between the data point `25` and the mean is `103.23`. The closer the point, the smaller the squared difference. The further away the point, the larger that squared difference.

As we saw in our calculations above, the squares of these differences contribute to the calculation of the standard deviation. If more data points are closer to the mean, the smaller the standard deviation. If more points are further away from the mean, the higher the standard deviation.

So, a low standard deviation can imply that the data values are clustered around the mean. Whereas, a high standard deviation can imply that the data values are more spread out.

2.7.1 Instructions

1. Answer the following questions.

2.7.2 Questions

1.

Is the standard deviation sensitive to outliers?

Yes.

No.

2.

If we removed the `25` from `[15, 9, 11, 23, 6, 25]`, would the standard deviation increase or decrease? Try to answer the question without any calculations.

It would decrease.

It would increase.

3.

What is the standard deviation for the `Income` column? You can use Excel's STDEV function.

`21400`

`41925.131`

`14906.21`

```
25096.120
```

2.8 Five-number summary

We have seen how both the range and the standard deviation are sensitive to outliers. We'll now learn about a more robust measure of spread called the **interquartile range (IQR)**.

Before we jump into IQR, let's talk about **percentiles**. If we had a list of values, the Nth percentile would be the value *below* which N% of the data would fall.

Let's consider a real-life example. You recently completed a test and are awaiting your score. You don't know how many students took that test other than you. One day, your teacher comes up to you and congratulates you for being in the 95th percentile of those who took the test. What does that mean?

It means that 95% of the students who took the test did worse than you in the test. Another way of putting it would be that you are in the top 5% of test-takers! Your score on the test would define that threshold - the 95th percentile. If your score was, let's say, 88 then 95% of students scored 88 or below in their test.

In the example above, the only way we could know the number of students below the threshold is if the scores were sorted in ascending order. That is important to note because it leads us to **quartiles**.

For a given list of sorted values, a **quartile** divides that list into **four** parts, or into quarters. There are three quartiles:

1. **First quartile** is the 25th percentile. That means 25% of the data falls below this value.
2. **Second quartile** is the 50th percentile. That means 50% of the data falls below this value. We already know this quartile by a different name — a median!
3. **Third quartile** is the 75th percentile. That means 75% of the data falls below this value.

The three quartiles, along with the minimum and maximum values, help summarize a given column of data. They are collectively known as the **five-number summary**:

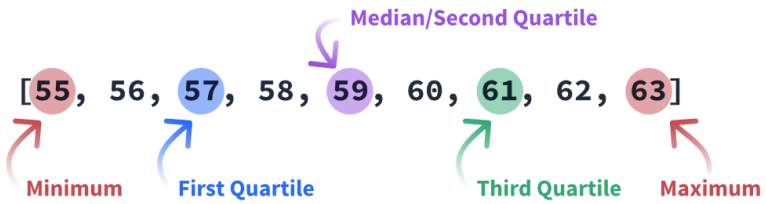
- The minimum.

- The first, or lower, quartile
- The median, or second, quartile
- The third, or upper, quartile
- The maximum

Here's a visual representation of the five-number summary for the following sorted list of customer ages:

```
[55, 56, 57, 58, 59, 60, 61, 62, 63]
```

ExplainCopy



Different tools can use different methods to compute these quartiles. So, we won't discuss those methods in detail. We can easily calculate quartiles in Excel using the QUARTILE() function. For example, to calculate the third quartile of `BMI` we would use:

```
=QUARTILE(C2:C2198, 3)
```

ExplainCopy

Let's answer some questions before discussing IQR.

2.8.1 Instructions

1. Answer the following questions.

2.8.2 Questions

1.

What is the third quartile for the `BMI` column?

19

26.4

23.5

25

2.

What is the first quartile for the `Income` column?

35416

52178

37040

51823

3.

What is the second quartile for the `Age` column?

51

62

```
44
```

```
50
```

2.9 Interquartile Range (IQR)

While the **range** of the data is the difference between the minimum and maximum values, the **interquartile range (IQR)** is the difference between the third and first quartiles. It gives us the range of the middle 50% of our data as opposed to the range for all of our data. Don't forget that to calculate quartiles, we need to sort our data in ascending order.

For the following list of customer ages . . .

```
[55, 59, 50, 70, 56]
ExplainCopy
```

- The **third quartile** is `59`.
- The **first quartile** is `55`.

. . . the **IQR** would be `4`.

If we changed the highest value, `70`, to a really large number, `7777`, the `IQR` would be the same.

Why is that?

On the previous screen, we learned that the quartiles are essentially median values. The second quartile is the median for the entire column of data. While the first quartile is the median (middle value) for the subset of that column — from the minimum to the median.

Since medians effectively ignore outliers, the subsequent calculation of IQR also ignores them. That's why IQR is a robust measure of spread compared to the range or the standard deviation.

Before concluding this lesson, let's answer a few more questions.

2.9.1 Instructions

1. Answer the following questions.

2.9.2 Questions

1.

What is the IQR for the `Age` column?

`56`

`18`

`44`

`22`

2.

For which of these columns is the IQR similar or close to the range?

`Age`

`BMI`

None of the above

`Customer Length`

3.

Without performing any calculations, can you say if the IQR of `Income` will be lower or higher than its range?

IQR will be higher than the range.

IQR will be lower than the range.

3 Applied Descriptive Statistics

3.1 Introduction to PivotTables I

So far, we've looked at the different measures of central tendency and spread. We also learned which measure could be more helpful, depending on the data we're working with. For example, how does removing an outlier impact a particular statistic?

We looked at individual columns and their statistics. But, what do those statistics actually help us explore and analyze in our data?

- How many married customers have an income of more than 60,000?
- What is the median salary of people with a PhD who have been customers for more than 300 days?
- How much money, on average, have customers younger than 40 spent on buying fruit?

Analysts should try to ask such questions to better understand any given dataset and extract relevant insights.

Our dataset, a modified version of the Customer Personality Analysis dataset, contains data on individual customers. By the simple act of asking questions, we can try to identify different customer segments from our data. The company could then use insights from those segments to develop products for those customers specifically.

And the above applies to any kind of data.

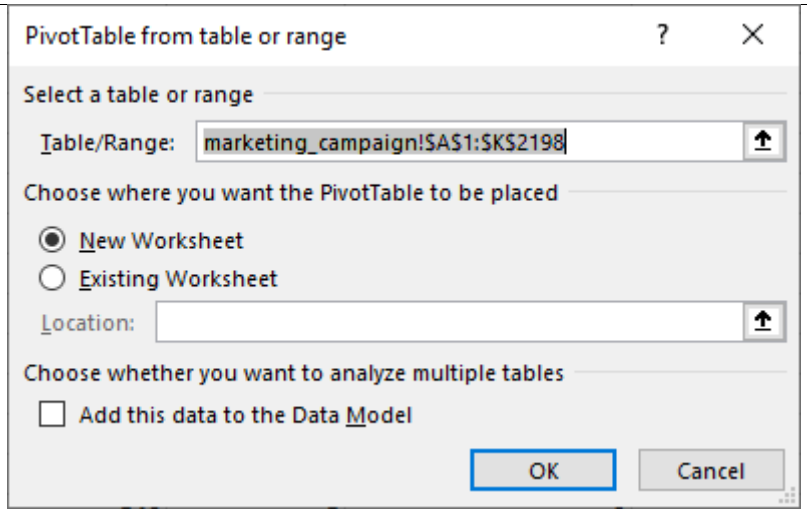
In this lesson, we'll calculate descriptive statistics as we explore and analyze our data further. But for that, we need to learn about PivotTables.

They are best explained in the documentation:

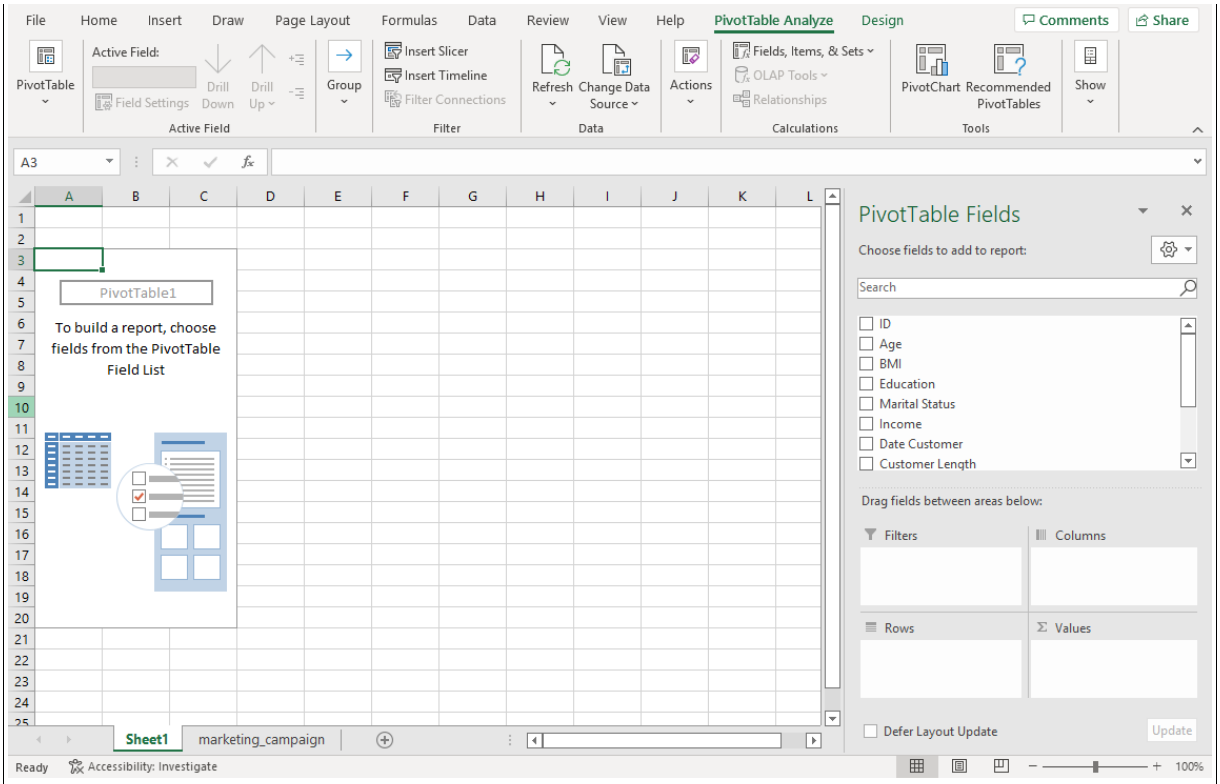
A PivotTable is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data.

We'll use this tool throughout this lesson. We'll start by calculating some basic statistics for the `Date Customer` column using a **PivotTable**.

1. Open the `marketing_campaign.xlsx` file in Excel.
2. Select the entire table containing the data.
3. On the toolbar, select **Insert**.
4. In the **Tables** section, select **PivotTable**.
5. In the **PivotTable from table or range** pop-up window, select **New Worksheet** and click **Ok**.



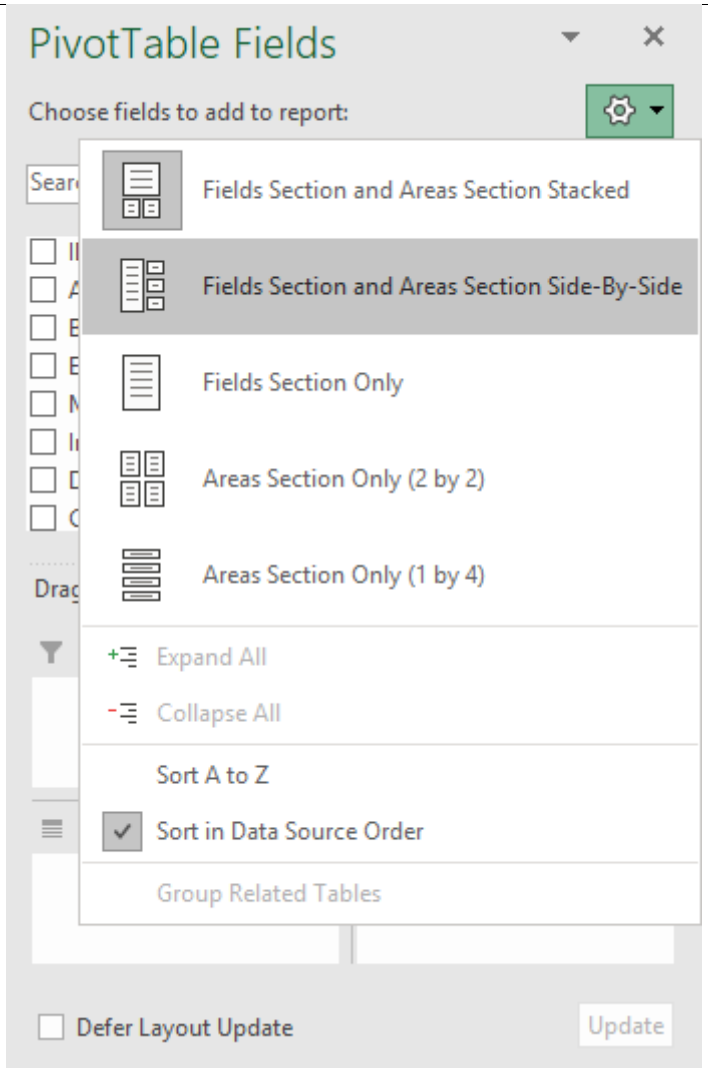
We have successfully added a PivotTable to a new worksheet! But, it's empty right now.



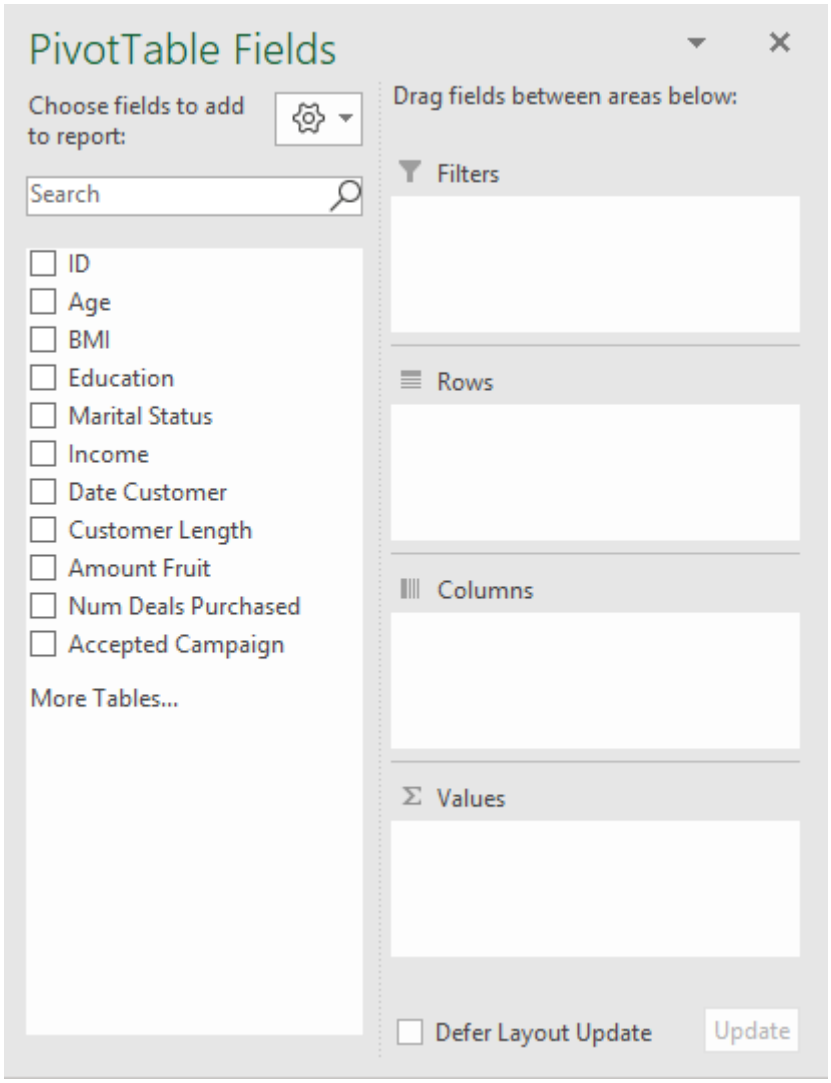
The **PivotTable Fields** task pane on the right side of the screen will allow us to populate our PivotTable. Before moving forward, let's change the layout of that pane.

1. Click on the cog wheel (**Tools**) on the top right of the task pane.

2. From the drop-down menu, select **Fields Section and Areas Section Side-by-Side**.



While changing the layout is a personal preference, it can give us more space to work with in the pane as well. We can also move and resize the task pane as we see fit.



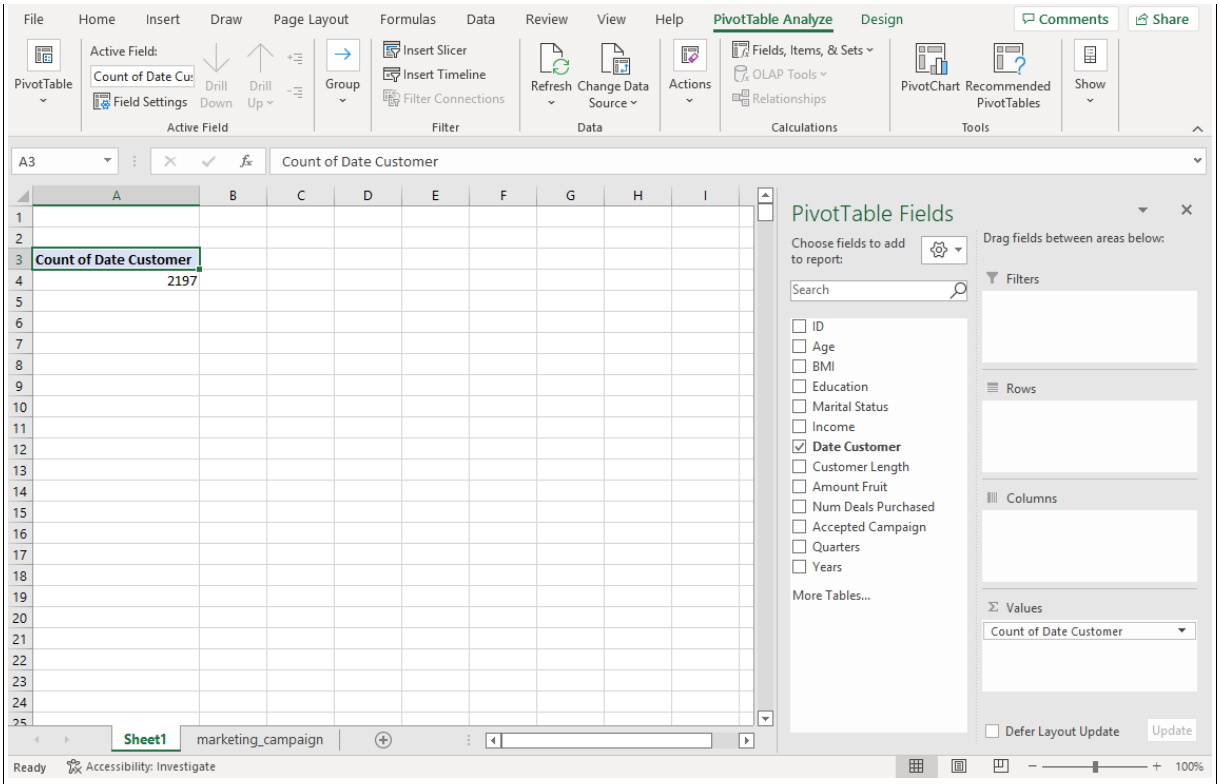
Let's add some statistics to the PivotTable for the `Date Customer` column.

1. In the **PivotTable Fields** task pane:

1.1 Click and drag the `Date Customer` field.

1.2 Drop it under the **Values** area.

Dragging and dropping the field populates our PivotTable. By default, it shows us the count, or the number of rows, of `Date Customer`.



With just a few clicks, we obtained the number of rows (`2197`) in the `Date Customer` column without using any formulas. Next, we will extend our PivotTable and calculate a few more statistics.

3.1.1 Instructions

If you get stuck or would like to compare your answers, we have provided a suggested solution file for this lesson called `marketing_campaign_solution.xlsx` which can be found in the `Home Folder` under `This PC`.

1. Open the `marketing_campaign.xlsx` in Excel.
2. Select the entire table, and insert a PivotTable into a **new** worksheet.
3. In the PivotTable, summarize `Date Customer` values by the following:

- Count
- Min
- Max
- Average

3.2 Introduction to PivotTables II

Previously, we created a PivotTable in a new worksheet and added a statistic for the `Date Customer` column.

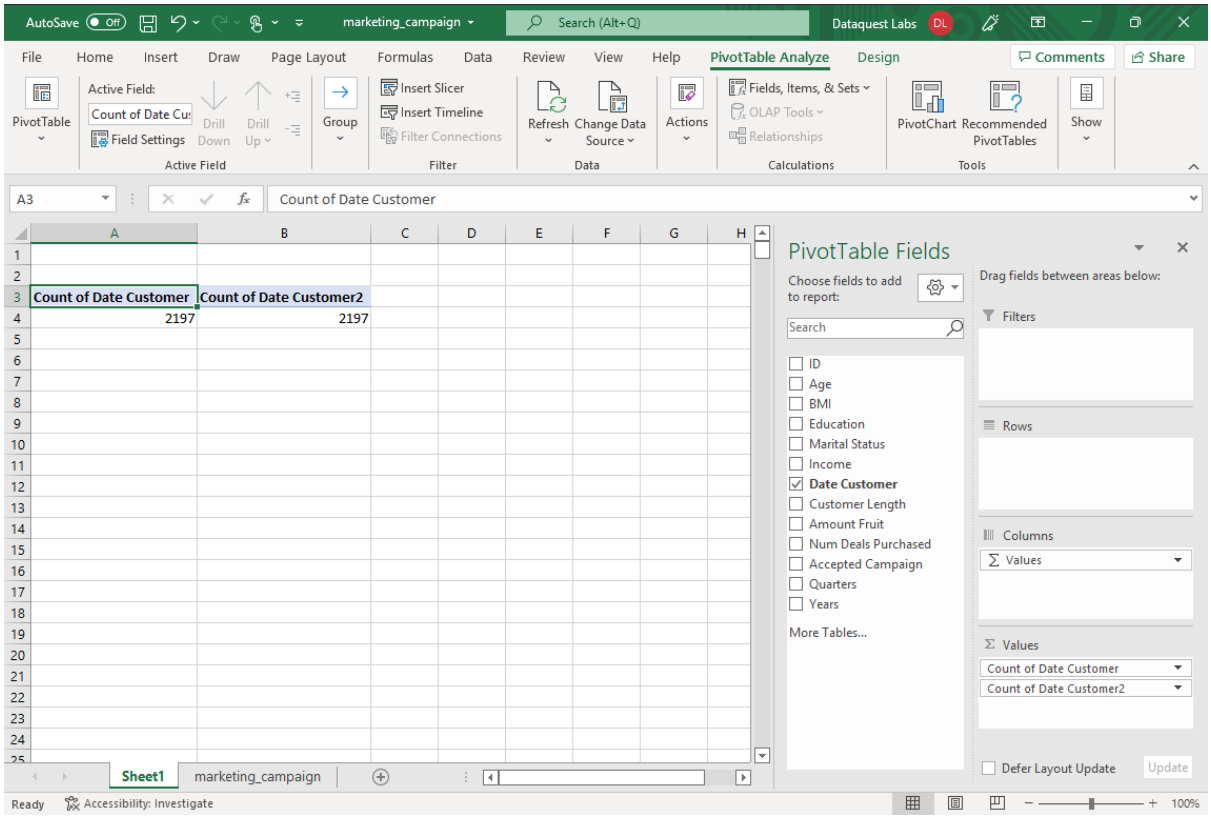
Let's repeat part of the process to calculate the earliest (minimum) date in the same column of data.

1. In the same worksheet, in the **PivotTable Fields** task pane:

1.1 Click and drag the `Date Customer` field.

1.2 Drop it underneath `Count of Date Customer` in the **Values** area.

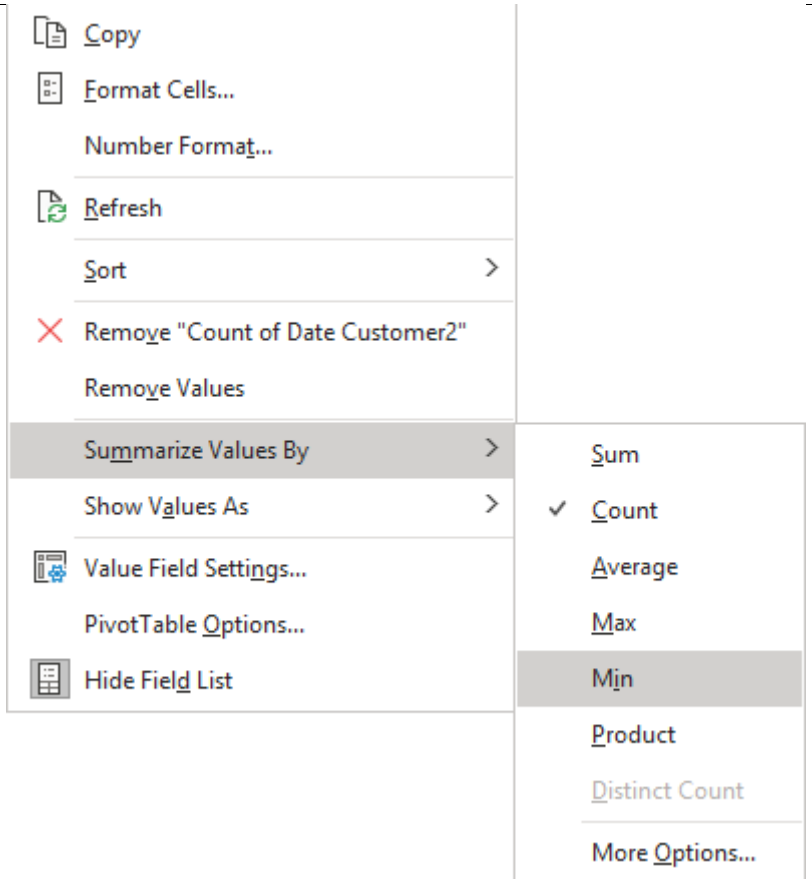
Notice how a new column, `Count of Date Customer2`, has been added to our PivotTable.



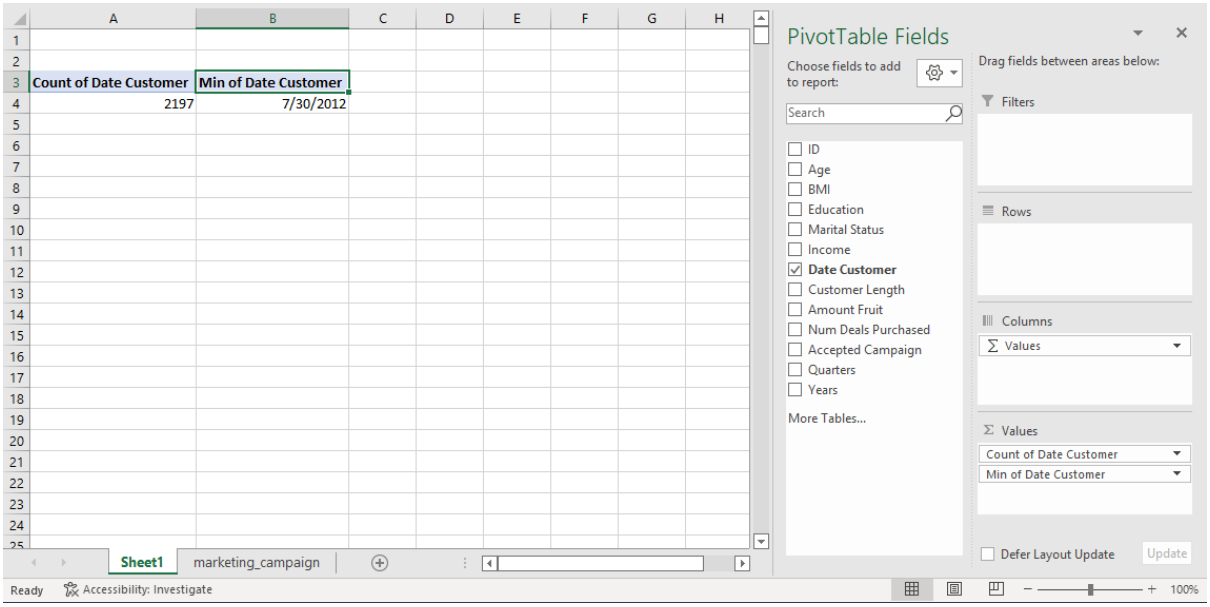
Once a new value has been added to the PivotTable, we can change it to our desired statistic.

1. Right-click on the new column.

2. Under **Summarize Values By**, select **Min**



The new column will automatically update to show us the minimum value - `7/30/2012`. The column name will also update, however, we can rename it ourselves as well.



If you see a different value than the one above, then make sure that the column has the right number format.

1. Right-click on the updated column.

2. Click on **Number Format**.

3. Select **Date** from the pop-up window and click **Ok**.

We have successfully summarized the `Date Customer` column by adding the values for the number of rows (**Count**) and the earliest date a customer enrolled with the company (**Min**).

Let's add a couple more statistics to our PivotTable.

3.2.1 Instructions

1. In the same PivotTable, summarize `Date Customer` by adding the following statistics:

- Max of Date Customer
- Average of Date Customer

2. Answer the questions below.

3.2.2 Questions

1.

What is the range of the `Date Customer` column in days?

686

268

459

314

2.

What is the average for `Date Customer` (mm/dd/yyyy)?

06/16/2014

```
03/14/2015
```

```
09/04/2012
```

```
07/03/2013
```

3.3 PivotTables and Grouping Data

So far, we have calculated multiple statistics for a column using a PivotTable. Calculating some statistics, like the median and mode, in PivotTables can be tricky. We won't include those for now, but we encourage you to check out the **Resources** section of the **Takeaways** screen at the end of this lesson to explore those options.

Now let's focus on the average value of a date. If it was customers' ages, we could easily interpret it to be the average age of a customer. But, what does an average of a date imply?

Our first thought might be to think of it as the date customers enroll with the company, on average. But we also calculated the range value to be `686`.

Is it possible that our dates are too spread out? If yes, then would the average date provide us with any insight?

Probably not. The range value for `Age` was `56`, indicating a smaller spread. If all our customers had ages uniformly spread out, the average age might not have provided us much information. But, as mentioned before, the idea is to be able to ask such questions as we explore our data and iterate upon our analysis.

Since the dates of enrollment are discrete, maybe calculating the mode could be more useful? We encourage you to try that out and come to a conclusion yourself!

The real benefit of PivotTables, however, comes from being able to group data. For example, what if we wanted to know the average salary of customers who have a PhD? What about the average salary of customers who have a PhD and are married?

Grouping data based on categories allows us to access more complex patterns in our dataset that would have been difficult for us to uncover otherwise.

Let's start with summarizing the age of our customers grouped by their marital status.

3.3.1 Instructions

1. Insert a PivotTable into a **new** worksheet.
2. In the PivotTable, drag and drop `Marital Status` into **Rows**.
3. In the PivotTable, summarize `Age` values by the following:

- Count
- Count as a percentage of the grand total
- Right-click on the column, and select **% of Grand Total** under the **Show Values As** option.
- Min
- Max
- Average
- StdDev

4. Answer the following questions.

3.3.2 Questions

1.

Which of the following statuses have a very small count percentage?

(select all that apply)

☐ `Absurd`

☐ `YOLO`

☐ `Widow`

☐ `Alone`

2.

Based on just this PivotTable, if we were to select customers to whom to promote a product, which category would likely be the best choice?

`Divorced`

`Single`

`Married`

`Together`

3.4 Data Cleaning

Some of the categories in our `Marital Status` column have very few values and aren't quite helpful for categorizing our data.

For example:

- `Absurd` could imply that `2` people either don't like the idea of a relationship or maybe they just don't like the idea of being married.
- `YOLO`, meaning "you only live once," could imply `2` people want to live as celibates to experience everything life has to offer. And while a bit unlikely, in this context, it could also mean "you only love once," which could imply someone who used to be in a relationship but is now single and intends to remain as such.
- `Alone` is likely the most direct one to interpret and is the same as being `Single`.

The above is an example of how we can explore our data and still find parts of it that could benefit from cleaning. Our data exploration and analysis doesn't necessarily have to be a linear approach where we only clean our data at the beginning. And, in practice, it frequently isn't.

So, we will quickly discard the rows with `Absurd` and `YOLO` because they can be interpreted in multiple ways. In such a situation, it's a good idea to discuss how the data

was collected, by whom, and if the sources could be contacted for clarifications. Since that's not an option here, we will remove these entries.

And for `Alone`, we'll modify the category to `Single`. There are some Excel functions, or combinations, that allow us to find row numbers corresponding to these categories. But, we don't really need to go for somethting complicated. We can use Excel's **find** feature since we don't have that many rows to modify.

3.4.1 Instructions

1. Modify your data table as follows:

- Delete rows with `ID` values `492`, `4369`, `7734`, and `11133`. These correspond to the `Absurd` and `YOLO` categories.
- Please note the above numbers are the numbers in the `ID` column and *not* the row numbers. The latter change if you delete any row, so it's better to refer the `ID` values.
- Modify all instances of `Alone` to `Single`.

2. Return to your PivotTable from the previous screen and refresh it.
3. Answer the questions below.

3.4.2 Questions

1.

Are single customers younger compared to the rest, on average?

Yes.

No.

2.

What is the average age of a married customer? Select the number closest to the average.

`44`

`54`

```
53
```

```
51
```

3.5 Narrowing Down the Analysis Scope

From our PivotTable, so far, we can see that most of our customers have the following `Marital Status`:

- `Married` (~`39%`)
- `Together` (~`26%`)
- `Single` (~`22%`)

We can also observe that the standard deviation of the ages for these customers isn't too high. So, it's reasonable to assume that the ages don't spread out too much for these segments.

Let's focus our analysis on these three groups for now.

3.5.1 Instructions

1. Filter the PivotTable to only include the above three categories.
2. Remove the following summaries:
   - Count
   - Count percentage
   - Min and Max
3. Add the average and standard deviation values for the `Income` column to the PivotTable.
4. Answer the following questions.

3.5.2 Questions

1.

Which group has the highest average income?

`Together`

`Married`

`Single`

2.

Which group has the highest income spread?

`Married`

`Together`

`Single`

3.6 Exploring Customer Segments

If we were to focus on just one segment, customers categorized under `Together` would seem suitable at first glance. Their average income is the highest. But, as we noticed, their income has a relatively high standard deviation as well.

The relatively high spread in their income might be an issue depending on the business problem we're trying to solve. Let's say, for example, our company priced our product that this particular segment was likely to buy based on the average income. We then ran an ad campaign targeting these customers. It's possible that the more spread out their income, the less effective the campaign might be because the product price doesn't suit enough customers in that segment.

And that's once again a good reminder why asking questions is important. At every step, we have to look at the data, find a pattern, and make sense of it in the context of the problem we want to solve. At times, we might not even have a given problem, but by iterating over our analysis, we can identify a potential solution that positively affects the business and customers.

Let's extend our analysis by adding more values. This time, we will try to look at the spending habits of these customers.

3.6.1 Instructions

1. To the same PivotTable as before, do the following:
1. Add the average for `Amount Fruit`.
2. Add the average of `Num Deals Purchased`.
3. Add the sum of `Num Deals Purchased` as a percentage of the total.
2. Answer the following questions.

3.6.2 Questions

1.

Which segment made the most purchases with a discount?

`Single`

`Married`

`Together`

2.

On average, which customer segment has spent the most on fruits in the past two years?

`Married`

`Together`

`Single`

3.

If we wanted to target customers with a discount coupon to encourage them to purchase our products, which two segments would be most likely to use those coupons?

(select all that apply)

`Married`

`Together`

`Single`

3.7 Filtering by Date

We have unlocked even more insights from our customer segments! We now know the following:

- `Single` and `Married` customers spent the most on fruits, on average.
- `Married` and `Together` customers made the most of their purchases using discounts.
- From our previous observations, we also noticed that `Single` customers do not only spend more on fruits on average, they are also younger than average.

Let's briefly discuss the last point above. `Single` customers are younger on average, spent more on fruits, but also made fewer purchases using discounts. We could argue that makes sense. Customers in a relationship could potentially split their expenses even if the dataset doesn't reflect that. They might also be more likely to try and save more for themselves, hence the use of discounts.

While the above explanation might not add more to our analysis, it helps create a narrative that could be useful. If we were to suggest a customer segment for an ad campaign to sell more fruits while trying to maximize profits, these `Single` customers would be our best option.

Sometimes, we might need to look at previous trends to try to figure out what could be an interesting avenue to explore next. Let's revisit the `Date Customer` column.

We know that the most recent customer enrolled with the company on `06/16/2014` since we already calculated the maximum value for that column early in this lesson. Let's finally look at customers who joined within the last three months of that date.

There could be a lot of date values even after filtering them. To simplify things, we only need to look at the **total row** for each category. Each category has a row at the end of it that summarizes the category as well.

3.7.1 Instructions

1. To the same PivotTable as before, do the following:
1. Add the `Date Customer` field to **Rows**.
- `Year` and `Quarters` will automatically be added to `Rows`. Remove both of these from `Rows`.
2. Filter the dates between `03/16/2014` and `06/16/2014`.
3. Add the count of any column to the PivotTable.
2. Answer the following questions.

3.7.2 Questions

1.

How many married customers enrolled in these 3 months?

`117`

`74`

`102`

`75`

2.

If we wanted to focus on new customers who are likely to spend the most on fruits, which segment would we consider?

`Single`

`Together`

`Married`

3.

If we added the `Divorced` category back to this PivotTable, would it make sense to target that segment in order to sell more fruits?

Yes.

No.

4.

If we had a product that was designed for younger generations, which customer segment could we target?

`Not enough information`

`Single`

`Married`

`Together`

4 Exploring Data with Data Visualization

4.1 Introduction

In a previous lesson, we learned how a single measure doesn't always offer enough insight into our data. We also learned how PivotTables can uncover insights when we start to group different fields.

In this lesson, we'll focus on these aspects once more but from the perspective of visualizing our data. We'll learn how simple visualizations like histograms can help us understand our data without calculating descriptive statistics.

We'll continue with the same dataset from the previous lesson, a modified version of the Customer Personality Analysis dataset. This version of the dataset has had the four records with erroneous `Marital Status` values removed.

Here are the columns you will find in the spreadsheet:

- `ID`: customer's unique identifier
- `Age`: customer's age
- `BMI`: customer's Body mass index (BMI)
- `Education`: customer's education level
- `Marital Status`: customer's marital status
- `Income`: customer's yearly household income
- `Date Customer`: date of customer's enrollment with the company
- `Customer Length`: number of days a customer was enrolled with the company
- `Amount Fruit`: amount spent on fruits in last 2 years
- `Num Deals Purchased`: number of purchases made with a discount
- `Accepted Campaign`: 1 if customer accepted the offer in the 1st campaign; 0 otherwise

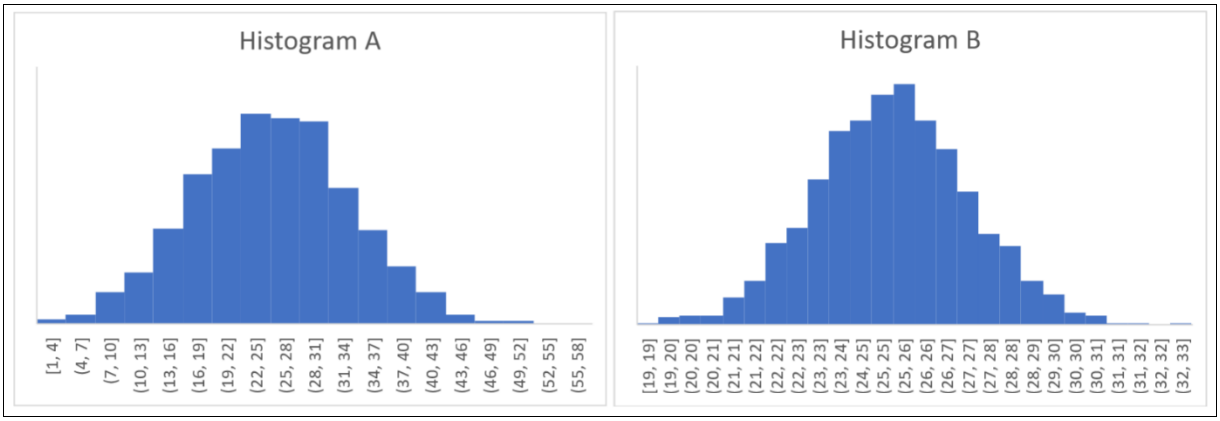
Let's start with a simple question first.

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 `marketing_campaign_solution.xlsx`, which you can find in the `Home Folder` under `This PC`.

1. Answer the question below given the following two histograms.

- Both distributions have the same mean (`25`) and number of data points (`2000`).



4.1.2 Questions

1.

Which distribution has a larger standard deviation?

B

A

4.2 Normal Distributions I

The histograms (or distributions) we saw previously are also called **normal distributions** or **bell-shaped curves** — because they look like a bell. A normal distribution has the following two features:

- The mean, median, and mode will be equal.

- It's symmetric about the mean.

This distribution will often come up in statistics and is, therefore, important to understand. However, it's important to note that when working with real-world data, the above two features, or conditions, might not always hold.

So, should we still refer to them as *normally distributed*?

We'll go over a quick example on the next screen. In the exercise below, we'll also filter values from our `Income` column and calculate some descriptive statistics for the remaining values in the column. For the latter, we'll use the AGGREGATE function in Excel since it allows us to ignore hidden values.

4.2.1 Instructions

1. Open the `marketing_campaign.xlsx` in Excel.
2. Copy the `Income` column to a new worksheet.
- Paste the data as a **link** in the new worksheet.
3. Select the column, and insert a histogram for that column of data.
4. Filter the data to keep only Income values that are below `150000`.
5. Using the AGGREGATE function, calculate the following:
1. Mean
2. Median
3. Standard deviation(STDEV.S)
6. Answer the questions below.

4.2.2 Questions

1.

What is the median of the filtered `Income` column?

`51624.82`

`51497.14`

`51563`

51438

2.

What is the mean of the filtered `Income` column?

51542.88

37013.17

48719.28

42183.56

3.

What is the standard deviation of the filtered `Income` column?

41164.13278

61746.19917

59746.31415

20582.06639

4.

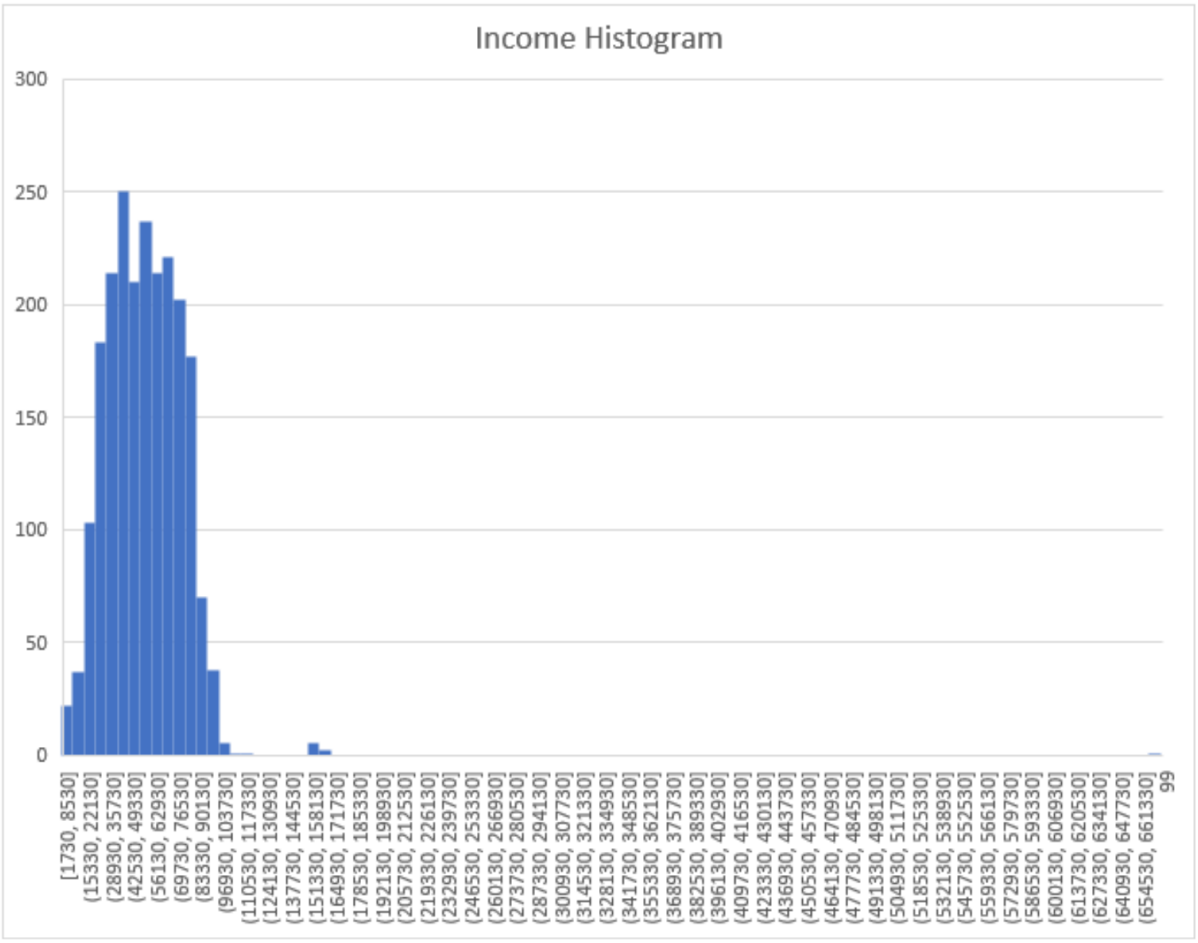
Why do you think we are not calculating the mode for our filtered `Income` column?

(select all that apply)

- [ ] The mode will be the same as the median.
- [ ] The mode is undefined for this column
- [ ] The data in `Income` is continuous, so calculating the mode is not helpful to us.
- [ ] There are better measures of central tendency we could use.

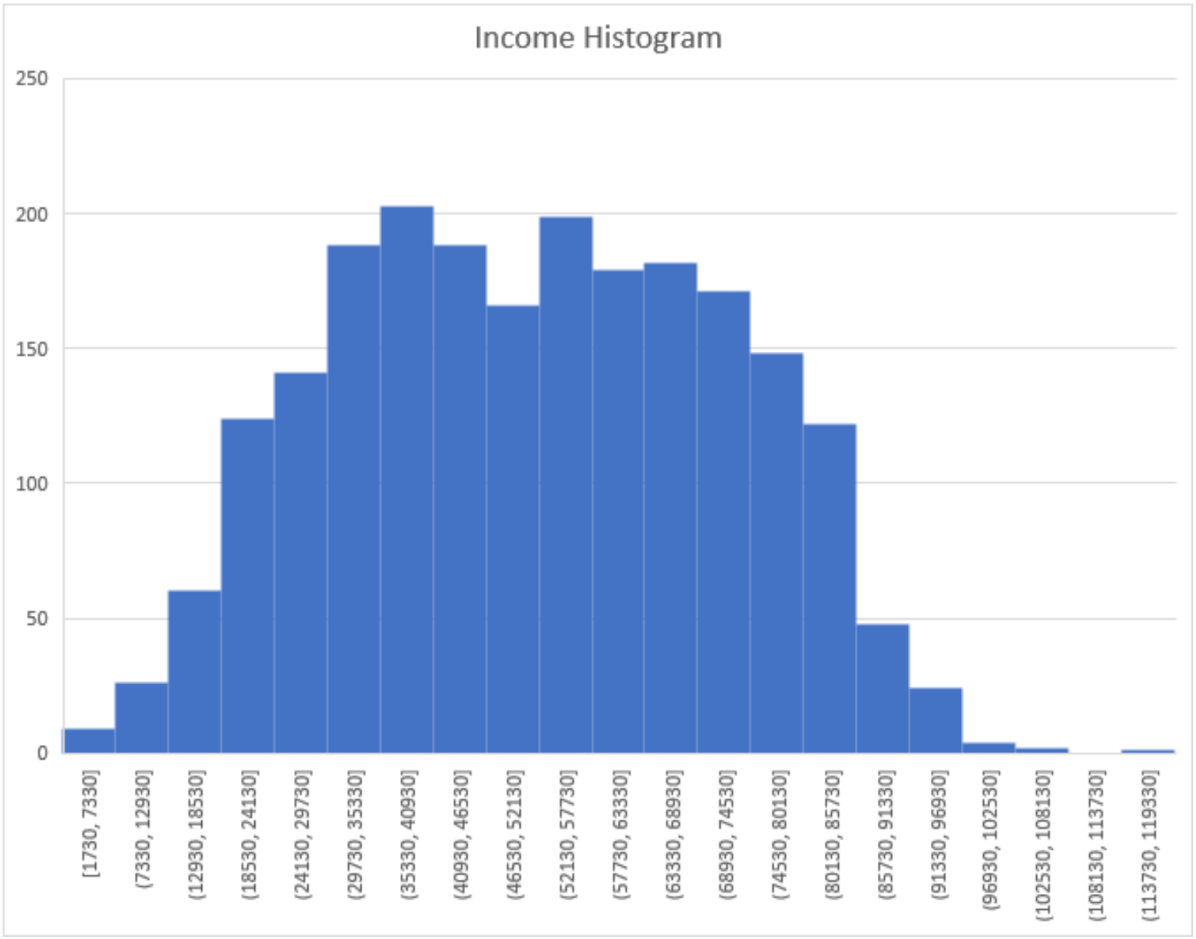
4.3 Normal Distributions II

On the previous screen, when we created the histogram for `Income`, we would have seen a chart like this:



Toward the end, we can see the value we identified as an outlier in a prior lesson. We can also see some values around the `150,000` mark. They don't quite fit with the rest of our data, so we chose to filter them out as well.

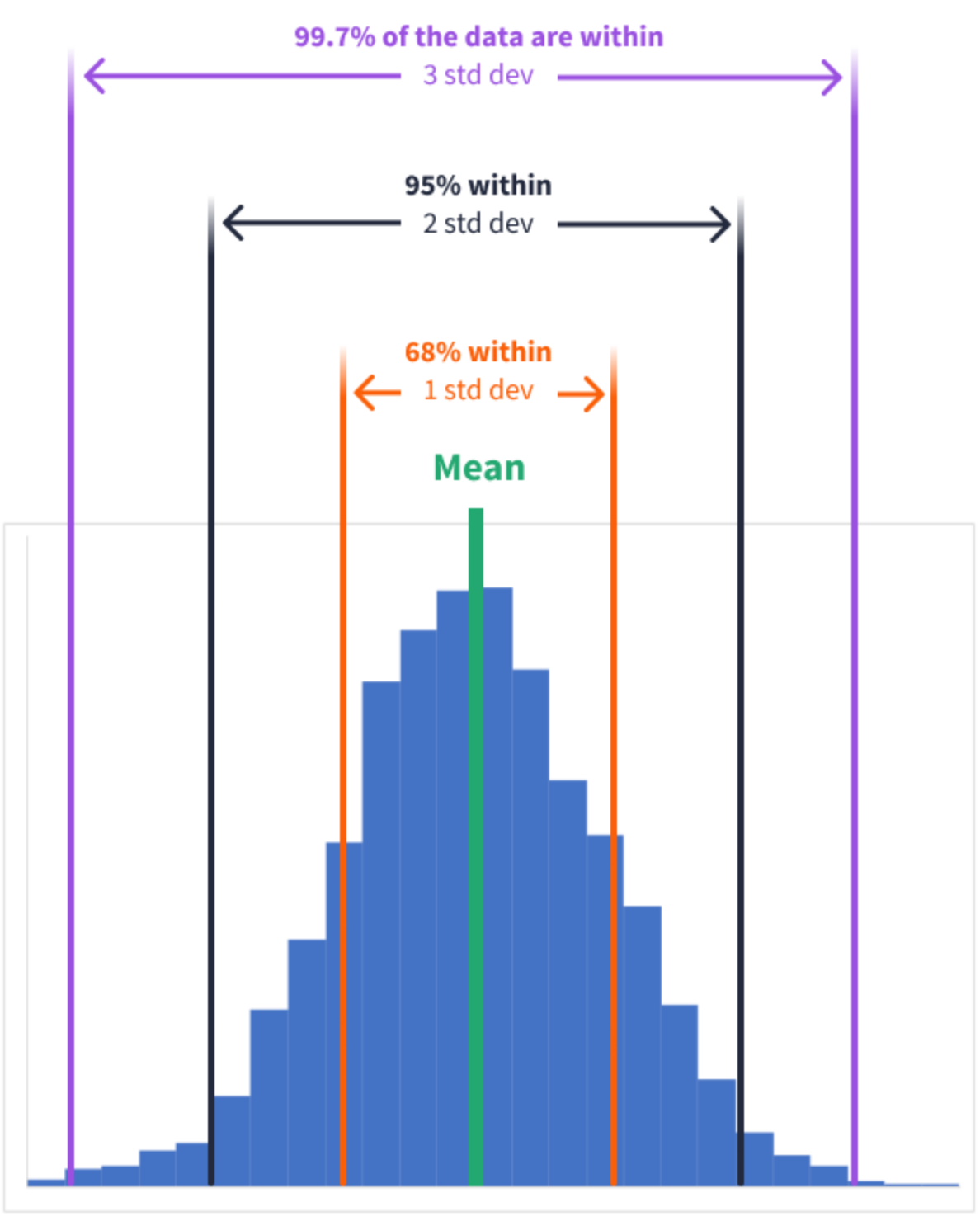
The resulting distribution is closer to a normal distribution:



As we mentioned on the previous screen, real-world data is likely to be close to a normal distribution. For our filtered `Income` column, the median and mean weren't exactly the same, but they were still pretty close in value.

So, how does the normal distribution help us? We'll discuss that in the next course, but for now, it helps us get a better idea about the distribution of our data.

Let's look at a normal distribution in relation to the standard deviation:



In a previous lesson, we learned that the standard deviation tells us how much the values in a given column of data deviate from the mean of that column.

Do you see the pattern in the normal distribution above?

We can see how the bulk of the data is clustered around the mean in the distribution:

- ~68% of the data are within 1 standard deviation of the mean.

- ~`95%` of the data are within 2 standard deviations of the mean.
- ~`99.7%` of the data are within 3 standard deviations of the mean.

The mathematical proof for the above percentages is beyond the scope of this course. However, we can see how straightforward it is to identify an insight if we know our data is distributed normally.

We know that for the filtered `Income` column, the following is true:

- The average income is `51542.88`.
- The standard deviation is `20582.066`.

So, with just these two values and knowing the data is distributed normally, we can make the observation that close to `68%` of our customers have incomes between `30960.814` and `72124.946`. That's a significant percentage of our customers, and this information can be useful when trying to identify relevant customer segments.

Let's look at the histograms of some of the other columns in our dataset next.

4.3.1 Instructions

1. In a **new** worksheet, create histograms for the following columns. We won't filter the data like we did for `Income`.

- `BMI`
- `Customer Length`
- `Amount Fruit`

2. Answer the following questions.

4.3.2 Questions

1.

What percentage of customers have a BMI roughly between `20.85` and `29.1`?

~`68%`

~`95%`

~`50%`

~`99.7%`

2.

Which of the three histograms *look* to be normally distributed?

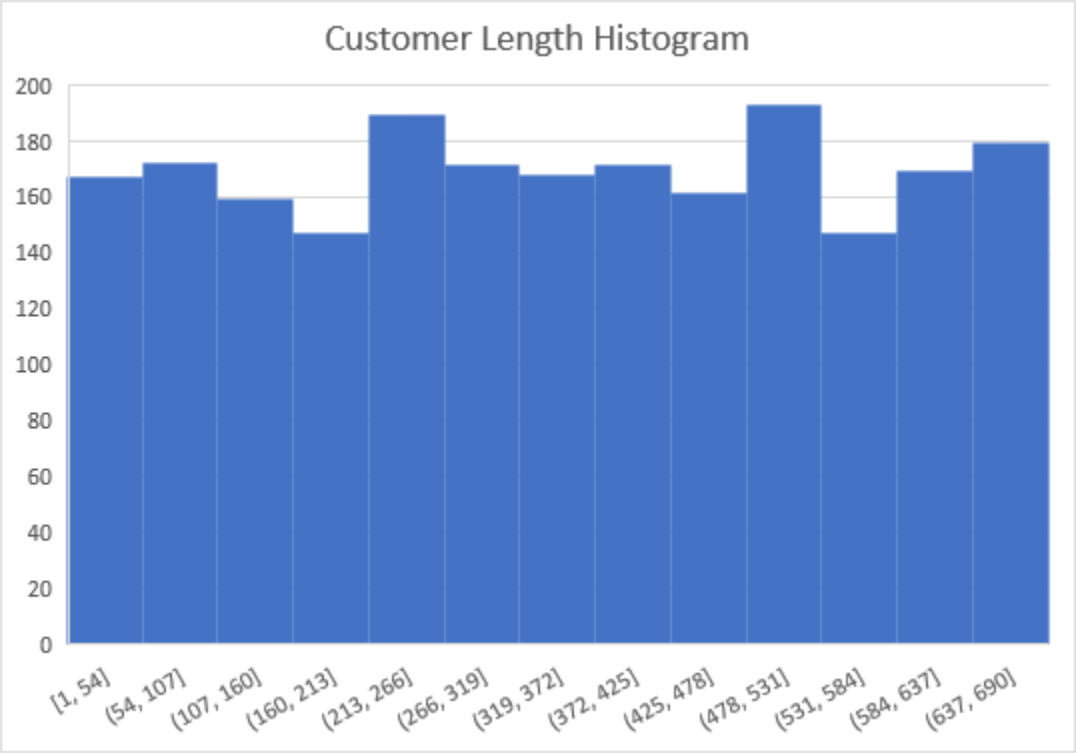
`Customer Length`

`BMI`

`Amount Fruit`

4.4 Uniform and Right-skewed Distributions

Let's look at the histogram for `Customer Length`.



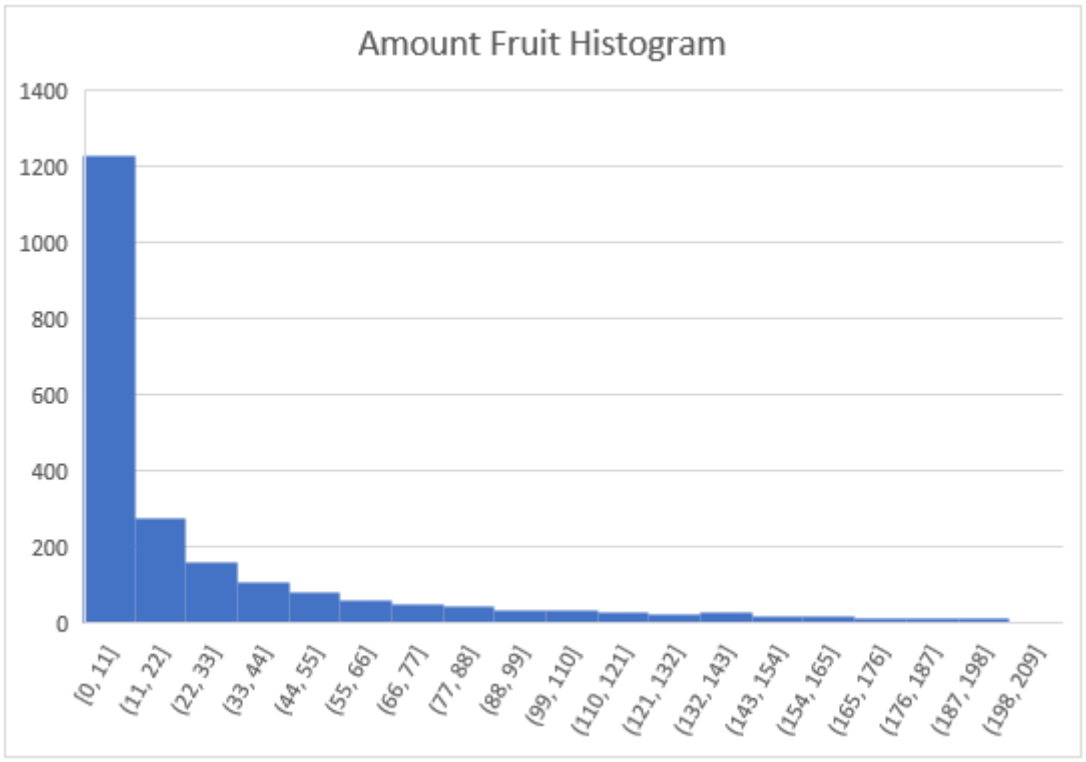
All the values are more or less evenly (or *uniformly*) spread out. We call this a **uniform distribution**. We can make some quick observations about the column by looking at the above chart:

The data is spread out and not centered around the mean.

The mean and median are likely close to each other. (We encourage you to think about why this would be the case based on how the two values are calculated.)

There isn't much of a pattern we can identify here, unfortunately. We can see that there are a couple of spots in the chart where the count is higher than the rest, but not by much. If those two were much higher, then we could focus on those customers more. However, in this case, there isn't much of an insight to extract.

Let's look at the other distribution we saw previously:



In a previous lesson, we referred to the above histogram as *right-skewed*. We can see how the data tapers off to the right. The data is mostly concentrated on the left, with a long tail on the right. We call this a **right-skewed distribution**.

At a glance, can we estimate where the mean and median would be for the above distribution? The questions below well help us answer this question.

4.4.1 Instructions

1. Answer the following questions.

4.4.2 Questions

1.

For a right-skewed distribution, where would the mean be?

Skewed distributions don't have a mean.

Between the median and the mode.

To the left of the median.

To the right of the median.

2.

For a right-skewed distribution, where would the median be?

To the right of the mode.

To the left of the mode.

At the end of the tail.

Skewed distributions don't have a median.

3.

For a right-skewed distribution, where would the mode be?

Right in the center of the plot.

Skewed distributions don't have a mode.

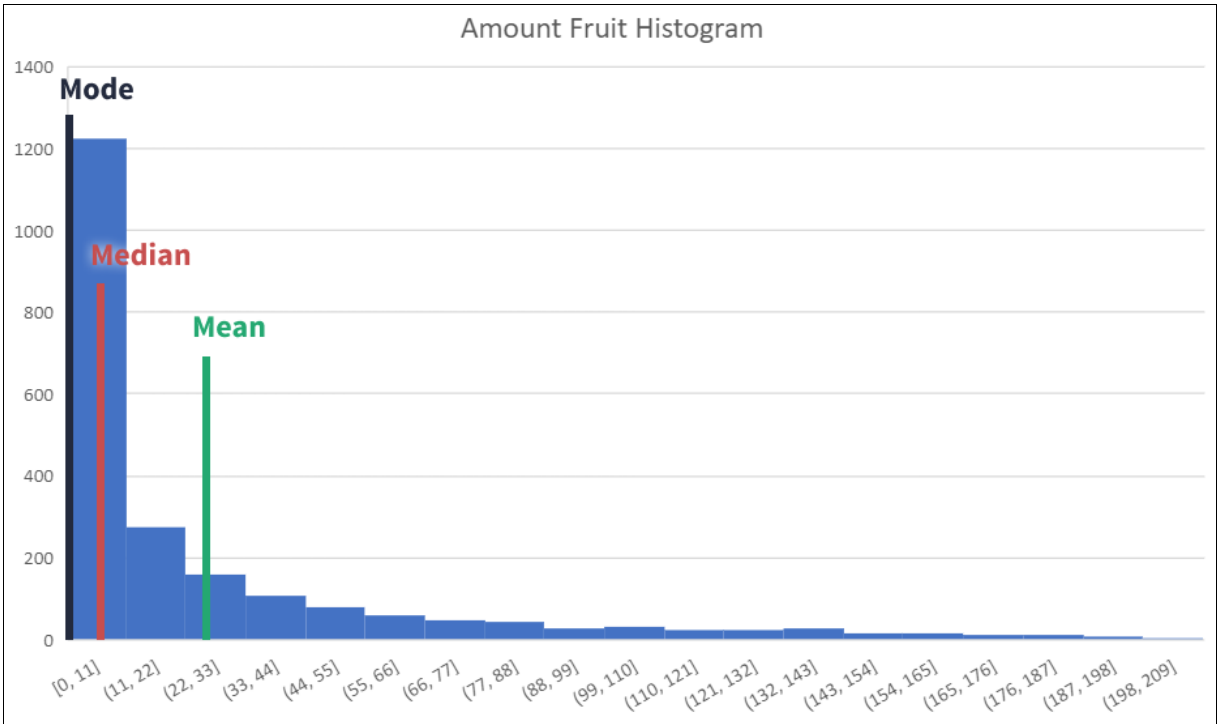
At the end of the tail.

At the peak.

4.5 Left-skewed Distributions

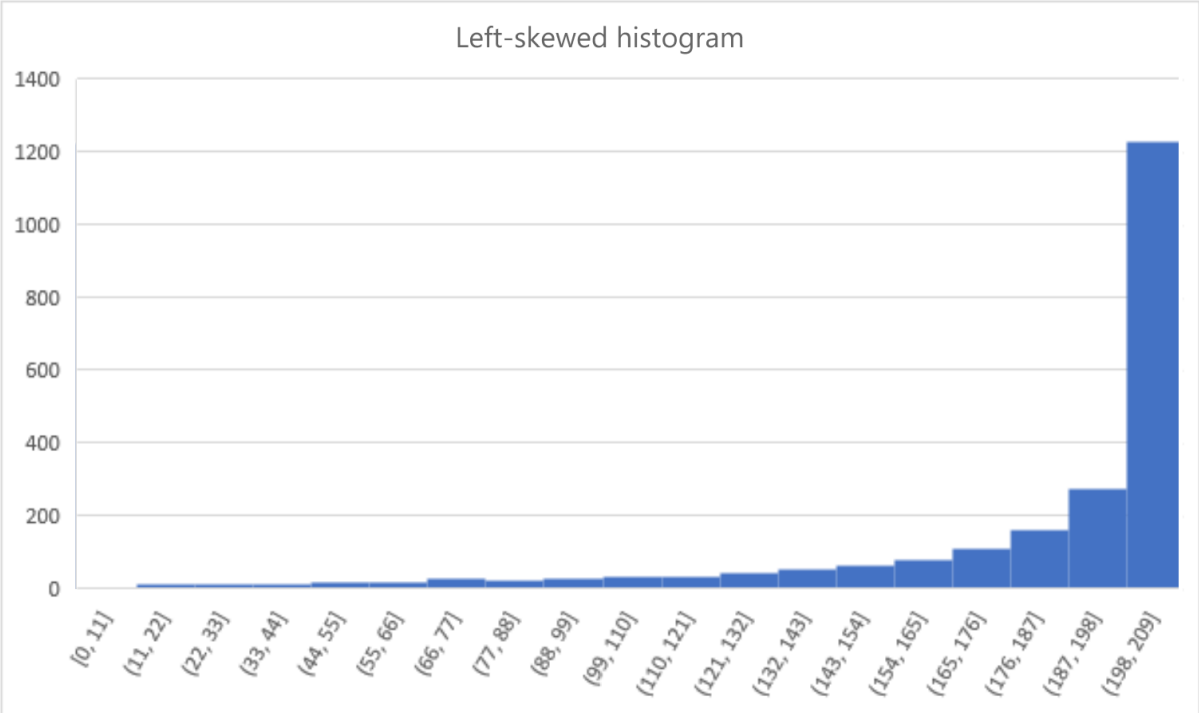
From the questions in the previous screen, we learned that for a **right-skewed distribution**, the following are true:

- The mode is the peak of the distribution.
- The median is to the right of the mode.
- The mean is to the right of the median.



And we didn't even have to calculate the values to draw these conclusions! We didn't focus on the standard deviation because it wouldn't have given us any information on the asymmetry. There are other statistics that can help with that — we'll cover them later in this lesson.

Another distribution we'll often come across is the **left-skewed distribution**. Unfortunately, our dataset doesn't contain a column that follows this distribution. But, this is what it would usually look like:



The data is concentrated on the right with a long tail toward the left.

Given what we have learned about right-skewed distributions, we can make some observations about left-skewed distributions as well.

4.5.1 Instructions

1. Answer the following questions.

4.5.2 Questions

1.

For a left-skewed distribution, where would the mode be?

Right in the center of the plot.

At the end of the tail.

At the peak.

Skewed distributions do not have a mode.

2.

Google the budget for NASA from 1958 to 2017. How has the budget been distributed?

Right-skewed

Uniformly

Normally

Left-skewed

3.

For a left-skewed distribution, where would the median be?

To the left of the mode.

Skewed distributions don't have a median.

To the right of the mode.

At the end of the tail.

4.

For a left-skewed distribution, where would the mean be?

Between the median and the mode.

Skewed distributions don't have a mean.

To the left of the median.

To the right of the median.

5.

Let’s say our dataset included a column containing the age at which our customers retired. Would that data likely have a left-skewed or a right-skewed distribution?

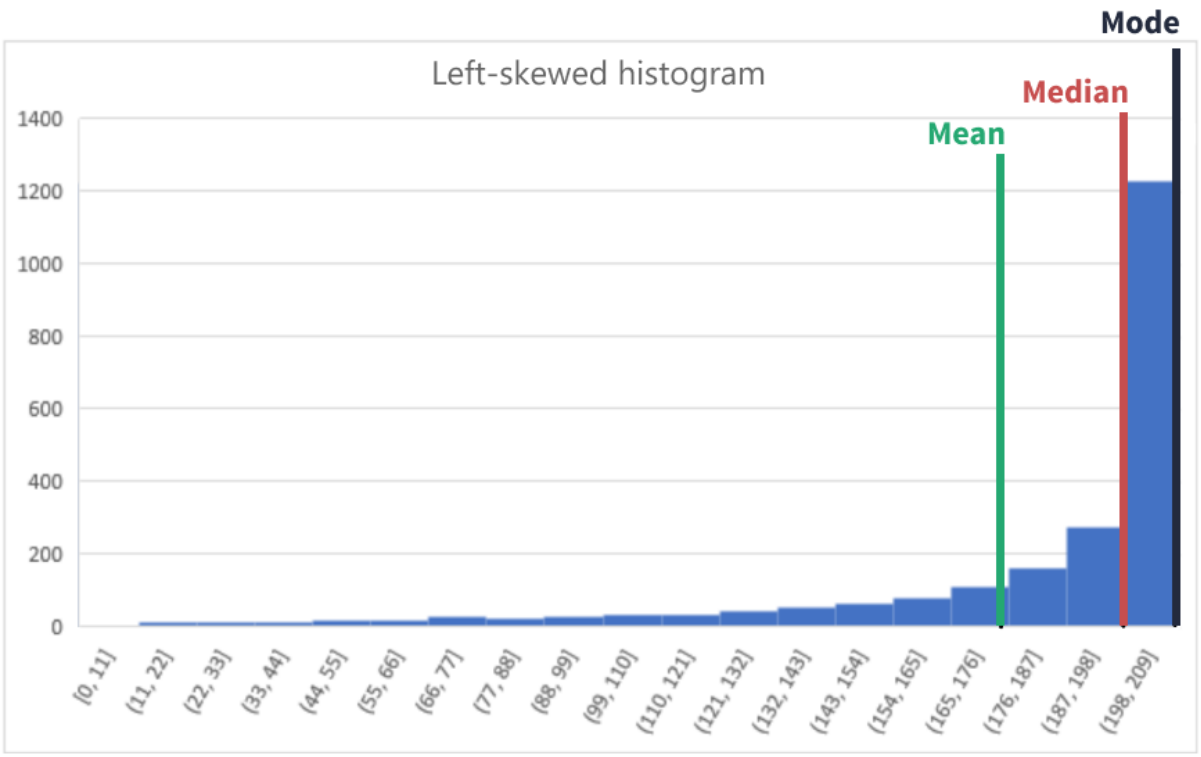
Right-skewed

Left-skewed

4.6 Boxplots I

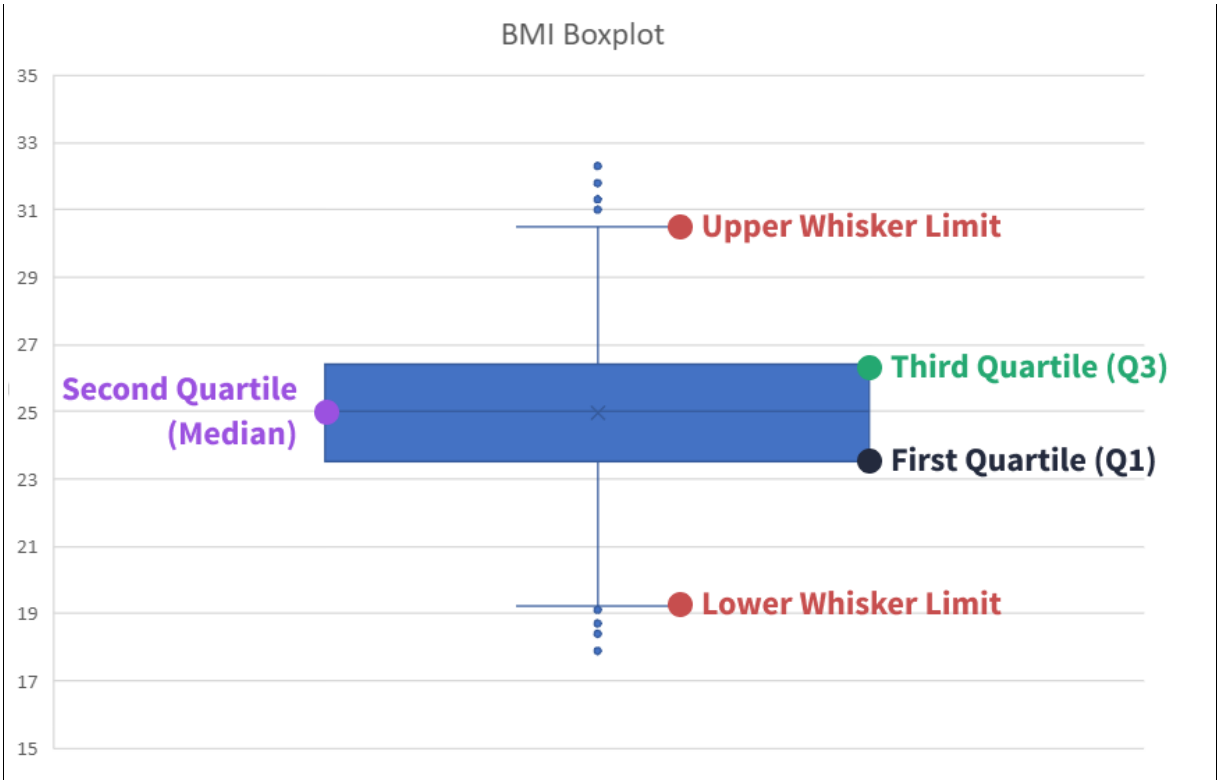
We learned that for a **left-skewed distribution**, the following are true:

- The mode is the peak of the distribution.
- The median is to the left of the mode.
- The mean is to the left of the median.



On the previous screen, we mentioned that the standard deviation isn't very helpful for skewed distributions. What can be helpful, however, is the interquartile range (IQR).

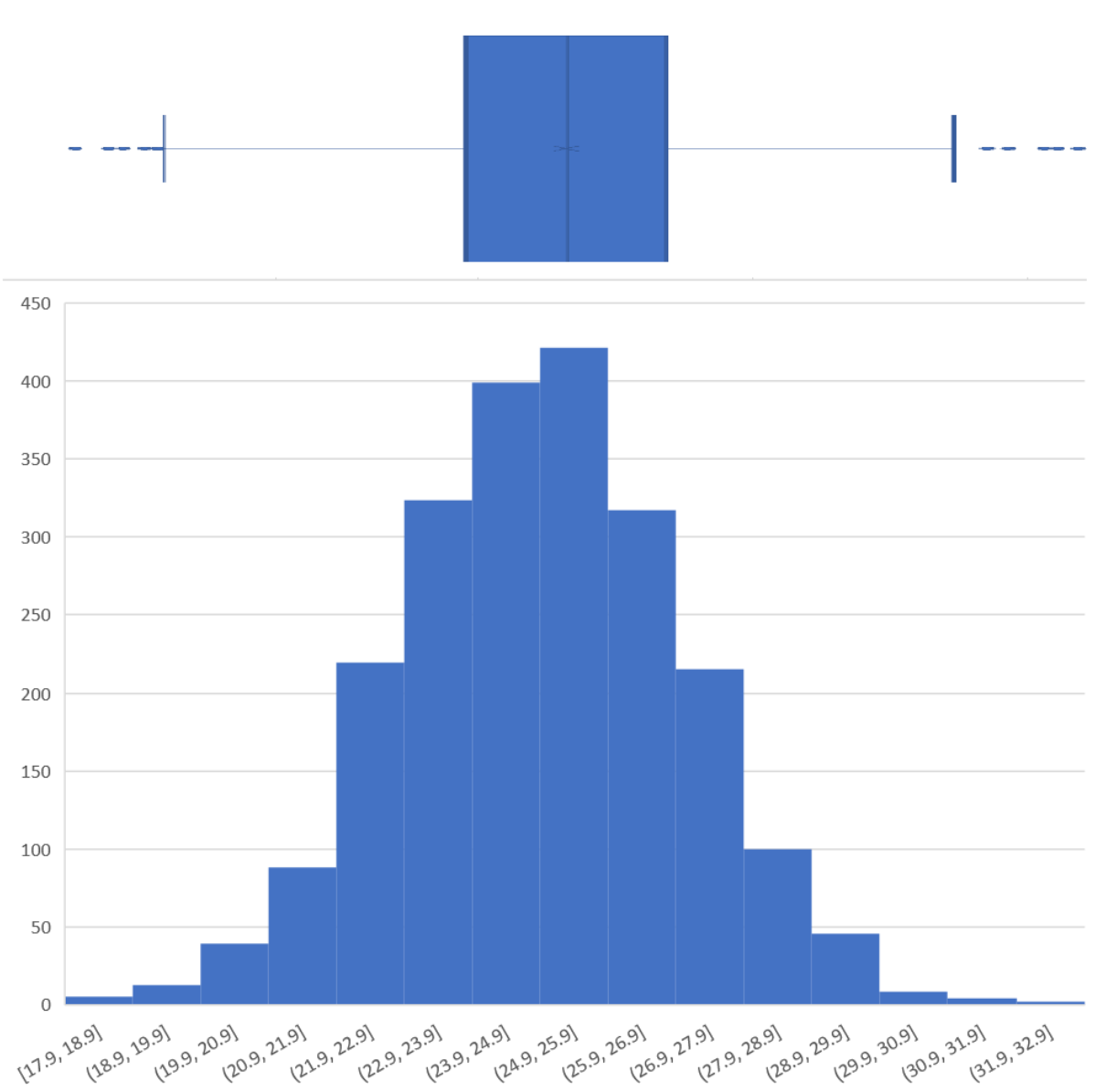
There is one more visualization that we will cover that depicts IQR. We call this the **boxplot**.



The above figure depicts a boxplot of the `BMI` column. It contains the following elements:

- The IQR, which includes the first, second, and third quartiles.
- The `X` in the plot is the mean.
- A whisker on each side of the quartiles.
- Outliers, the dots that lie beyond the whiskers.

We are already familiar with the IQR. However, visually, we can quickly get an idea of the spread of the distribution. If we rotate the boxplot, we can compare it to our `BMI` histogram:



Since, `BMI` is distributed normally, the following are true:

- The median is at the center of the IQR.
- The mean is the same as the median since it's a normal distribution.
- The difference between the median and Q3 is the same as the difference between Q1 and the median, indicating that the distribution is evenly spread out from the center.

The whiskers are a bit tricky to understand. They represent the minimum and maximum value *within* the range of the data points that aren't considered outliers:

- We calculate the upper limit of that range as `Q3 + 1.5*IQR`.
- We calculate the lower limit of that range as `Q1 - 1.5*IQR`.

The mathematics behind the above formula are beyond the scope of this course. However, the above approach is commonly used to identify outliers, which we'll discuss further below.

Once again, to clarify, the above defines the upper and lower limits of that range. In Excel, the values at the end of the whiskers are calculated as follows:

- The data point right *below* the maximum limit.
- The data point right *above* the minimum limit.

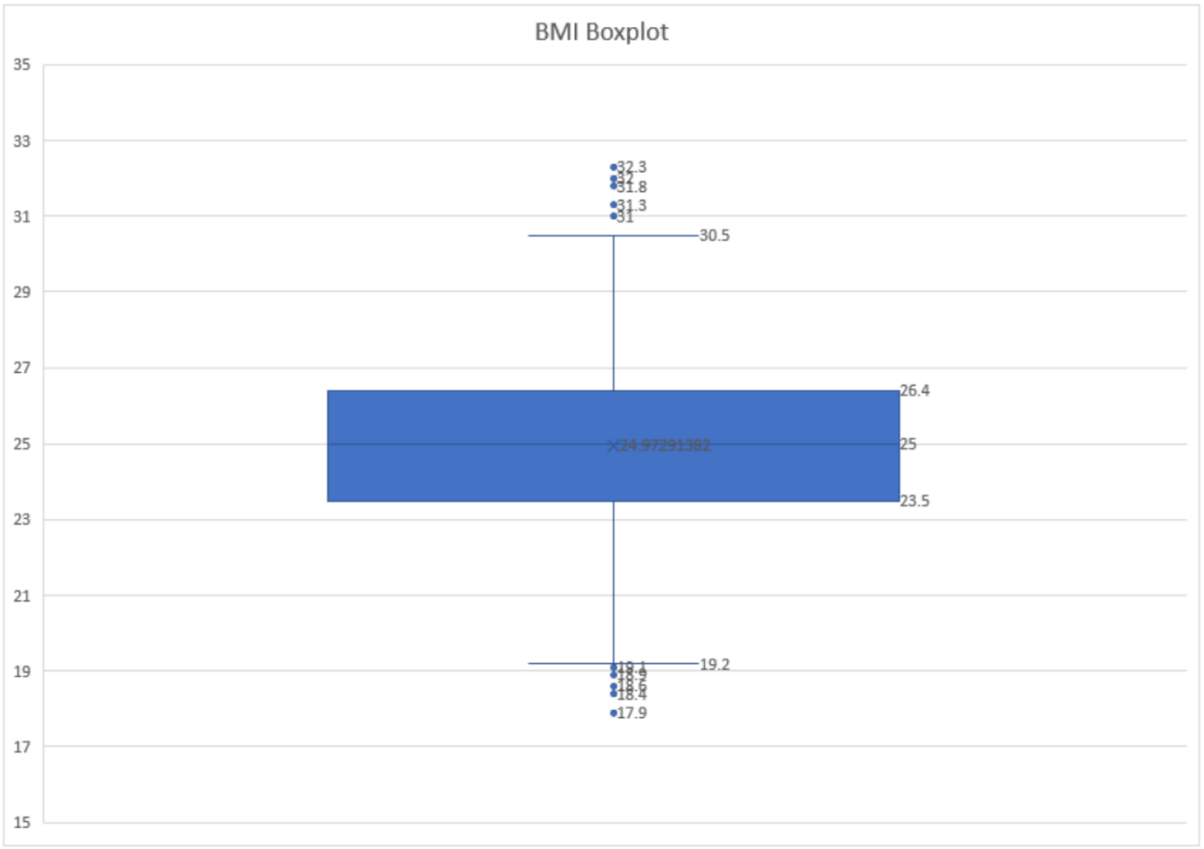
Let's look at the data values we have discussed above for the `BMI` column:

BMI Boxplot Values				
Whisker Minimum	19.2			
Q1	23.5		**IQR**	2.9
Q2	25		**Lower Whisker Limit**	19.15
Q3	26.4		**Upper Whisker Limit**	30.75
Whisker Maximum	30.5			

The **upper whisker limit** is `30.75`. The closest data point in our column *below* that limit is `30.5`. That's our **whisker maximum**.

Similarly, the **lower whisker limit** is `19.15`. The closest data point in our column *above* that limit is `19.2`. That's our **whisker minimum**.

That defines the range of our data when using the boxplot. Every data point *outside* this range is considered an outlier, which we can see as individual dots in the plot. If we add the data labels to our plot, it would look like this:



Please note that the Y-axis lower limit is set to `15`.

We can see how useful boxplots can be. Not only do they give us an idea about the spread of our distribution, they also help us identify outliers!

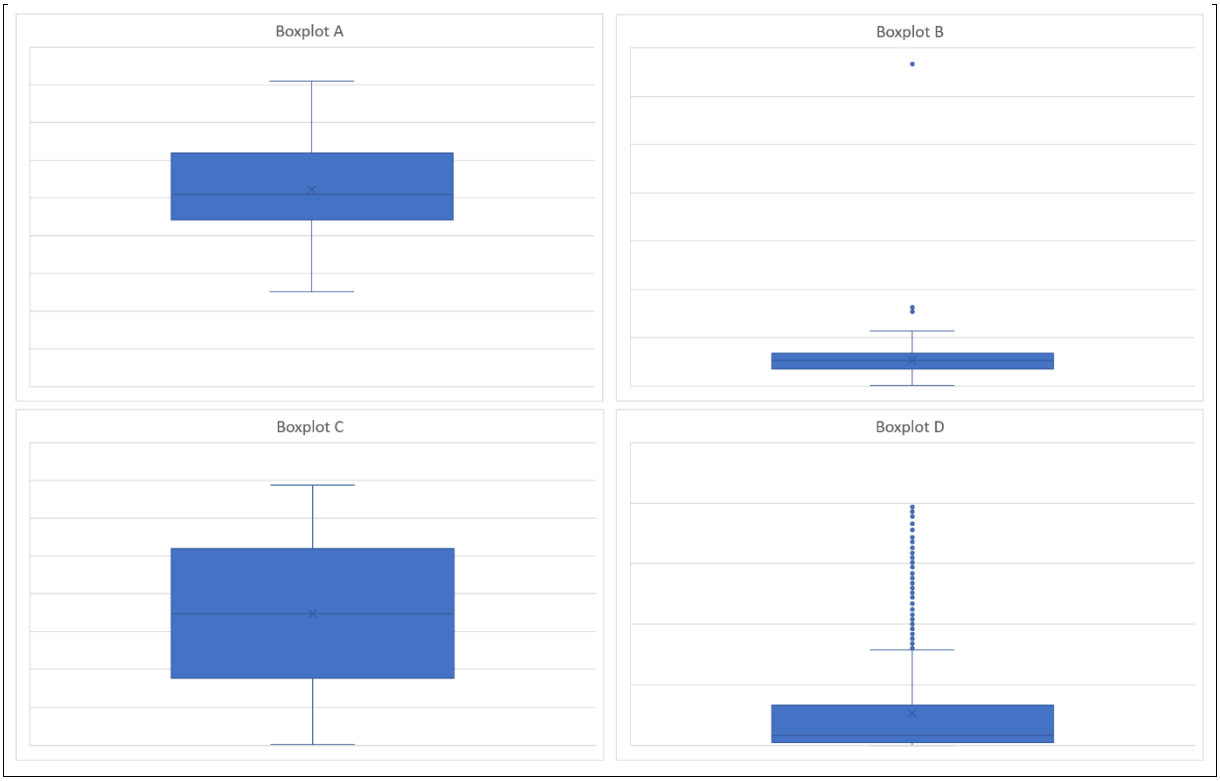
On the next screen, we'll look at some more boxplots corresponding to our dataset.

4.7 Boxplots II

We learned a lot about boxplots on the previous screen. Let's now apply what we learned to some other columns and see what observations we can make.

The questions below are meant to better understand the data by only looking at the plots. We recommend *not* creating the boxplots of the columns before attempting the questions. You can refer to the histograms of each column to make comparisons, however.

4.7.1 Instructions



1. Given the boxplots above, answer the questions below.

4.7.2 Questions

1.

Which boxplot do you think corresponds to the `Amount Fruit` column?

B

A

D

C

2.

Which boxplot do you think corresponds to the `Income` column?

D

B

A

C

3.

Which boxplot do you think corresponds to the `Age` column?

B

C

D

A

4.

Which boxplot do you think corresponds to the `Customer Length` column?

D

A

C

B

4.8 Boxplots by Categories

`Customer Length` has a uniform distribution, and we saw the corresponding boxplot on the previous screen as well. For the boxplot, the following are true:

- The mean and median were the same.
- The boxplot is symmetric.
- There were no outliers.

It does seem eerily similar to a normal distribution's boxplot as well. How can we differentiate the two, then?

Well, the most important bit is that there are no outliers. We know that when data is normally distributed, it follows a bell-shaped curve. There are always likely to be some outliers as a result when creating a boxplot.

However, for a uniform distribution, all data points are evenly spread out. The entire range of the boxplot will contain all of the datapoints.

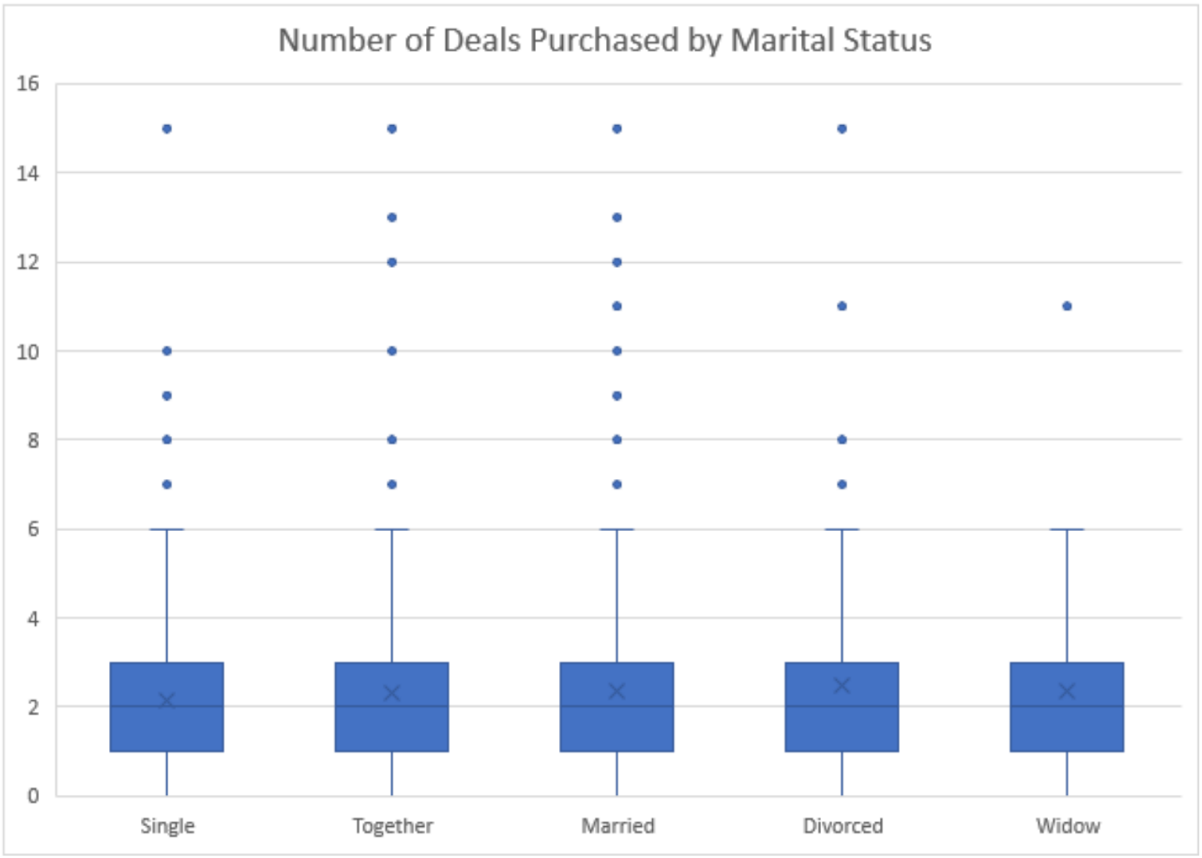
But, does that hold true for every uniform distribution in the real-world? We'll let you figure that out as you explore more datasets in your learning journey!

On this final screen, we'll look at how grouping data by categories affects a boxplot. Unlike what we learned previously, we don't need a PivotTable in this case. Excel makes it fairly simple to create boxplots per category in a single chart:

- Select a category column.

- Select a numerical column.
- Click **Insert**, and select the **Bar and Whisker** option.

This is what boxplots for `Num Deals Purchased` by `Marital Status` would look like:



We can see how, for each status, the data is clustered around the same median value. One thing we can point out is that, on average, divorced customers purchased more deals than the other customers. That is exactly the observation we made in the previous lesson when we used a PivotTable!

Boxplots, like other visualizations, offer multiple advantages when it comes to exploring and analyzing data without calculating any statistics.

We'll end this lesson with a few more questions on categorized boxplots.

4.8.1 Instructions

1. In a **new** worksheet, create boxplots for the following:

    1.1. `Income` categorized by `Education`.

- Filter out any values in `Income` higher than `200000`.

1.2. `Amount Fruit` categorized by `Education`.

1.3. `Num Deals Purchased` categorized by `Education`.

2. Answer the questions below.

4.8.2 Questions

1.

On average (the mean value), customers belonging to which category have purchased the most deals?

`2n Cycle`

`Graduation`

`Master`

2.

Customers in which of the following categories have the highest spread for amount spent on fruits?

`PhD`

`Graduation`

`Master`

3.

Across all three plots you created above, which statement about customers with `Basic` education holds true?

There are likely very few customers with `Basic` education compared to the rest.

These customers earn less on average, compared to the rest, but they spent a lot more on buying fruits and on deals.

4.

What is the education qualification of customers with the lowest median income?

`Master`

`Basic`

`Graduation`