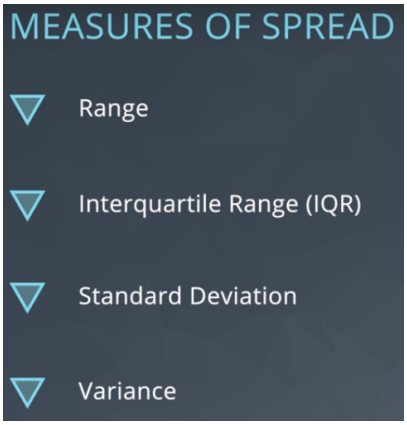
What are Measures of Spread?

Measures of Spread are used to provide us an idea of how spread out our data are from one another.

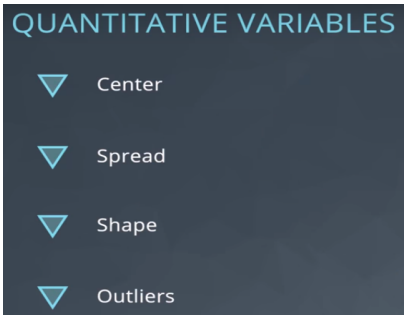
Common Measures of Spread:-

1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**

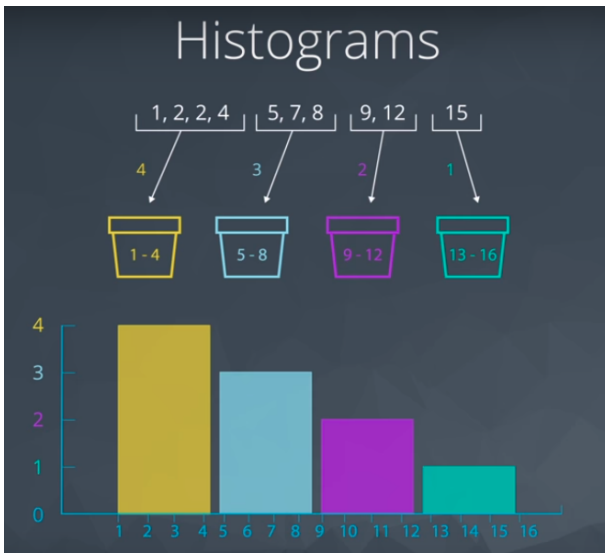


Histogram

Histograms are super useful to understanding the different aspects of quantitative data. In the upcoming concepts, you will see histograms used all the time to help you understand the four aspects.

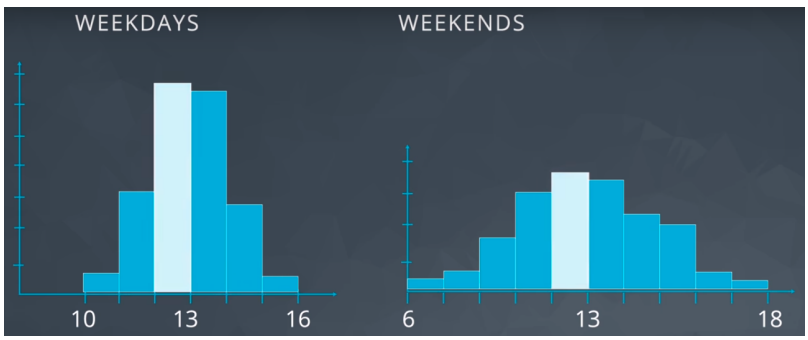


How Histograms constracted?

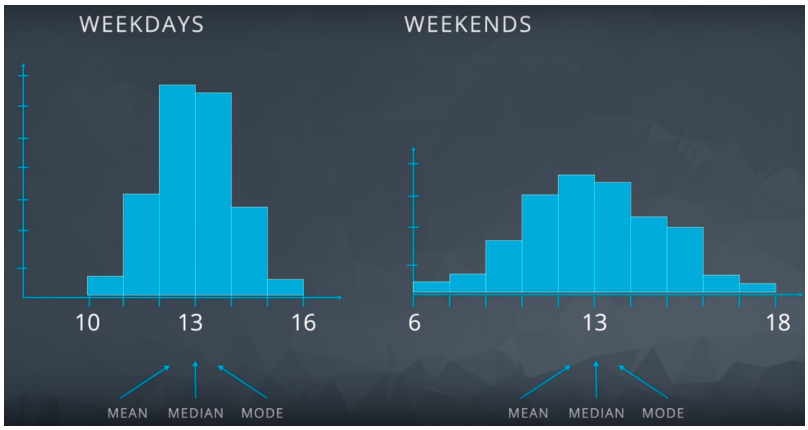


For Example:-

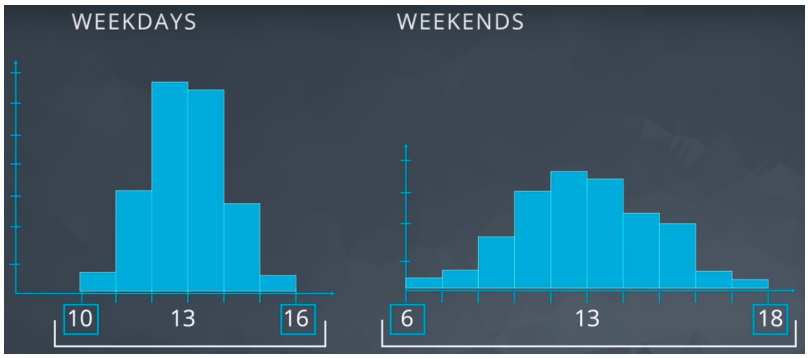
- **Here, are two histograms comparing The number of dogs was seen on the weekends and week days. You will notice that the tallest bins for both weekdays and weekends are associated with 13 dogs.**



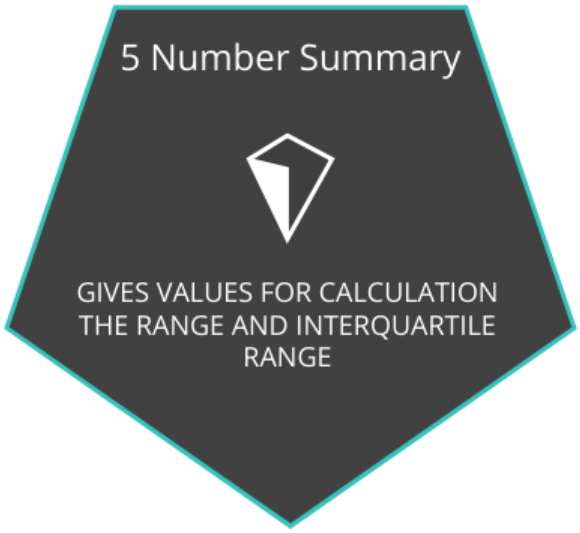
- **So the number of dogs I expect to see are essentially the same for weekdays as on weekends. And the measures of center were the same on both have a mean, median, and mode that are about 13 dogs.**



- **But something is different about these two distributions. So what's the difference? Well, the difference is how spread out the data are for each group. You can see that the number of dogs I see on weekdays, ranges from 10-16. While on weekends, it ranges from 6-18.**

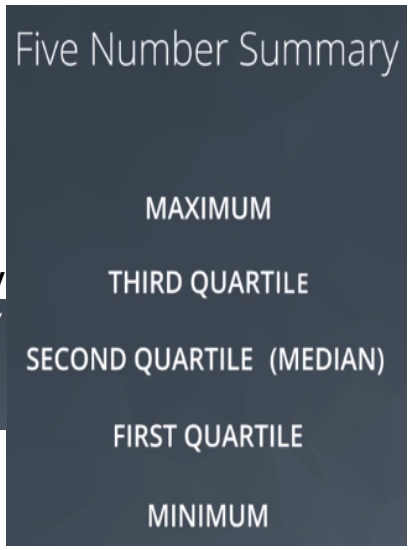


Five Number Summary



1. **Minimum:** The smallest number in the dataset.
2. **Q1:** The value such that 25% of the data fall below.
3. **Q2:** The value such that 50% of the data fall below.
4. **Q3:** The value such that 75% of the data fall below.

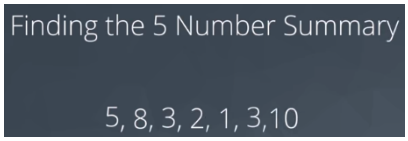
5. **Maximum:** The largest value in the dataset.

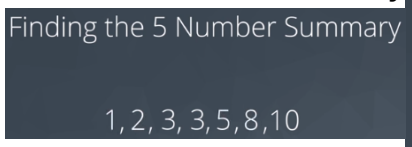


For Examples:-

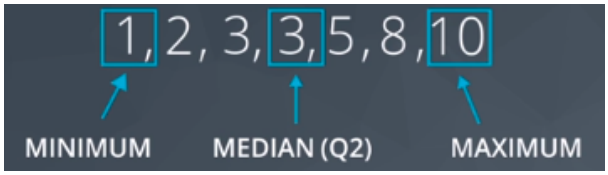
When n is Odd number

- **First we need to order the data to calculate Five Number summary**

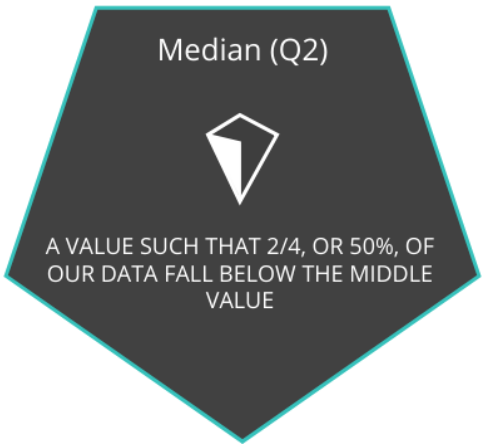




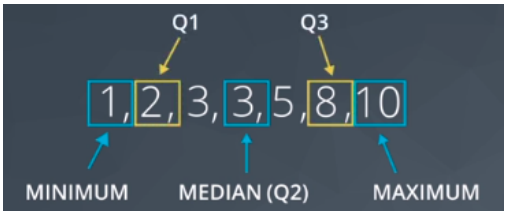
- **the minimum and the maximum values are easy to identify as the smallest and largest values. As we calculate it in the section on measures of center, the median is the middle value in our dataset.**



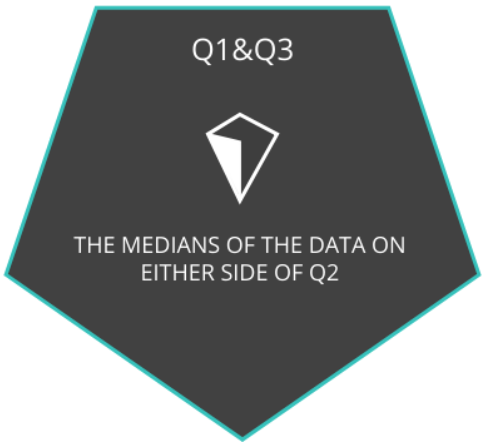
- **We also call this Q2 or the second quartile because 50% of our data or two quarters fall below this value.**



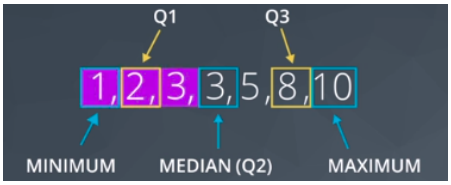
- **The remaining two values to complete the Five Number Summary are Q1 and Q3.**

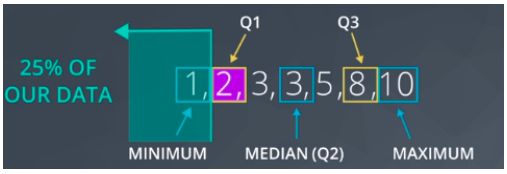


- **These values can be thought of as the medians of the data on either side of Q2**

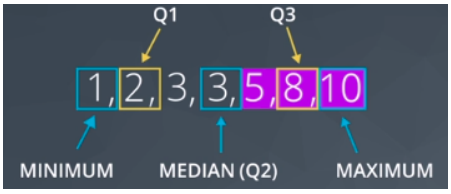


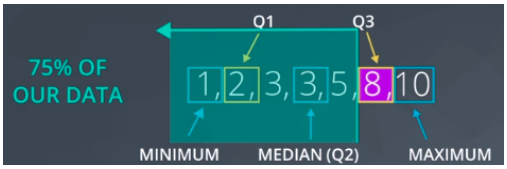
- **That is the median of these data points is Q1, this value is such that 25% of our data fall below it.**



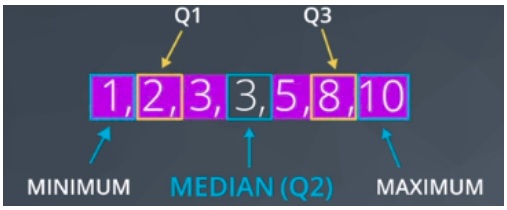


- **the median of these data points is Q3 or the third quartile, This value is such that 75% of our data fall below this mark.**

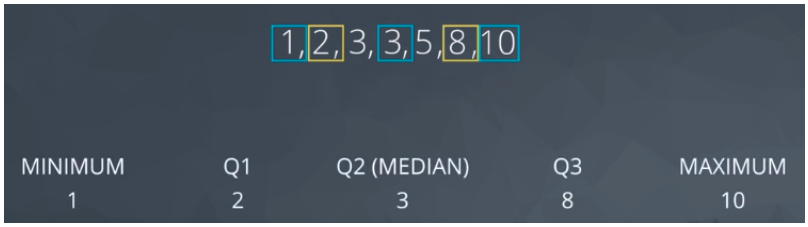




- **Notice, Q2 was not an either a set of these points used to calculate Q1 or Q3.**

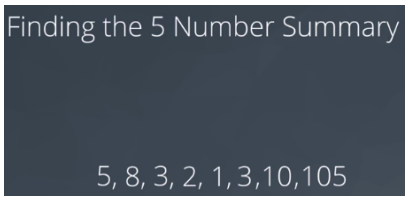


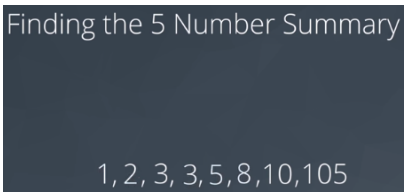
- **This provides our Five Number Summary as the following.**



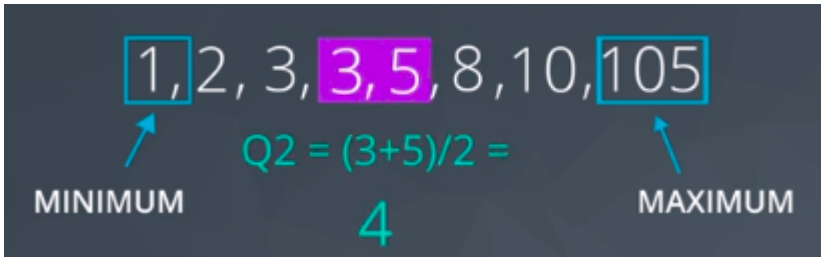
When n is Even number

- **First we need to order the data to calculate Five Number summary**

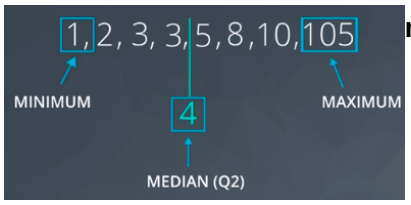


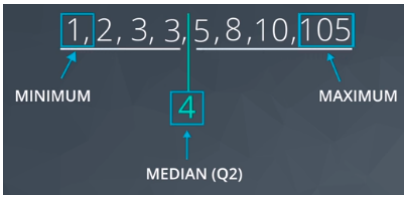


- **We can quickly identify the maximum and the minimum. Remember, with an even number of values, the median or Q2 is given as the mean of these two values here.**

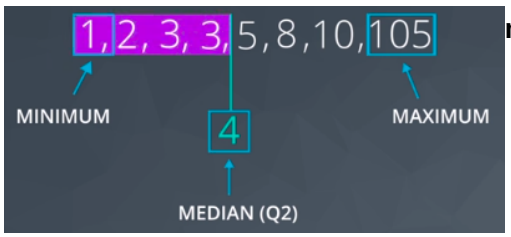


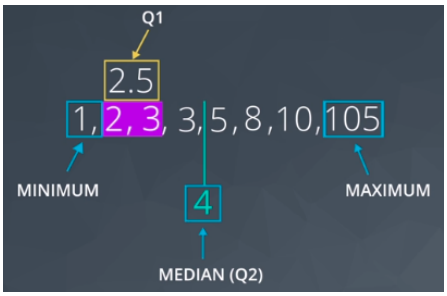
- **In order to find Q1 and Q3, we divide our dataset between the two values we use to find the median. This provides these two datasets.**



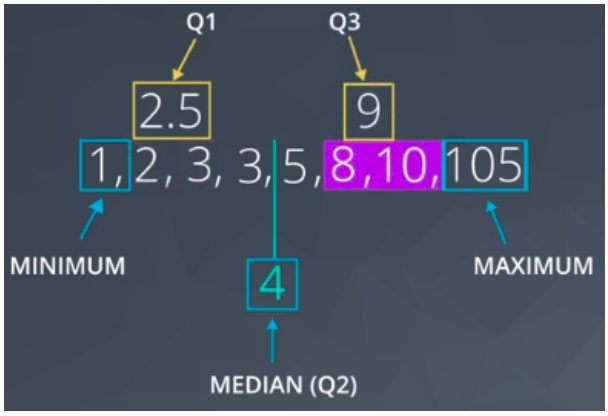


- **Finding the median of each of these will provide Q1 and Q3. For this dataset, Q1 will be the mean of these two values.**





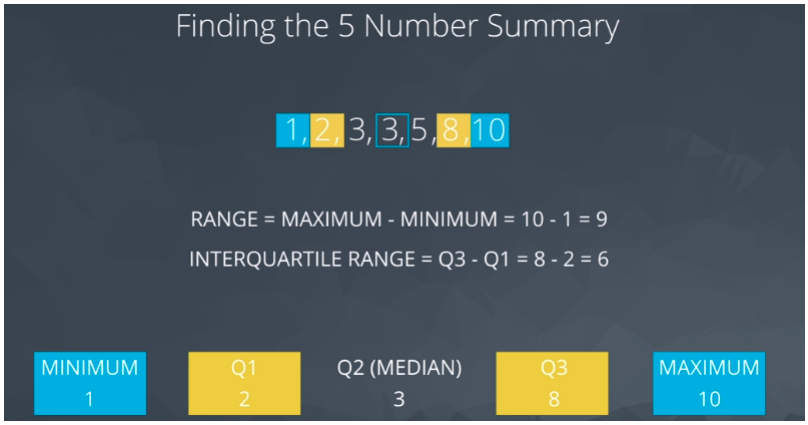
- **Q3, will be the mean of these two values.**



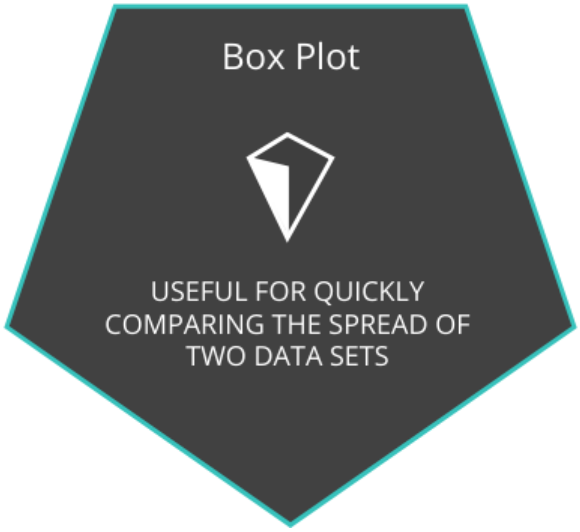
- **This provides our Five Number Summary as the following.**
- **Q3, will be the mean of these two values.**

MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2.5	4	9	105

- **Range: The range is then calculated as the difference between the maximum and the minimum.**
- **IQR: The interquartile range is calculated as the difference between $\bold{Q_3}$ and $\bold{Q_1}$.**
- **Once we've calculated all the values for the Five Number Summary, finding the range and interquartile range is no problem. For the first dataset, the range is calculated as the maximum minus the minimum. For the first dataset, this was 10 minus 1 or 9. And the interquartile range is calculated as Q3 minus Q1, which is 8 minus 2 or 6.**

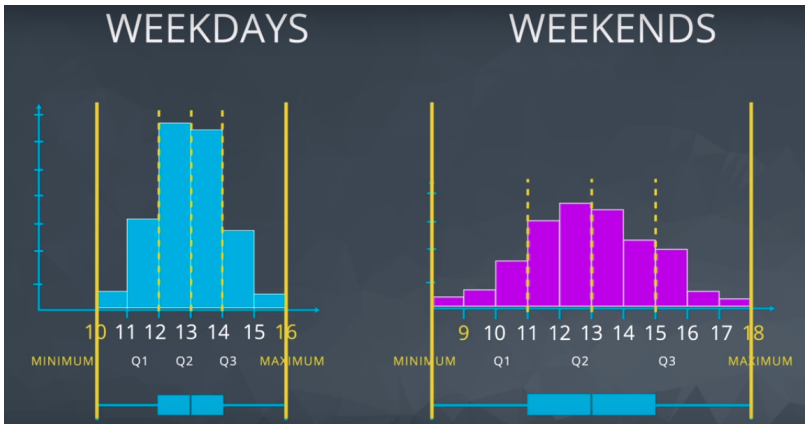


Box Plot

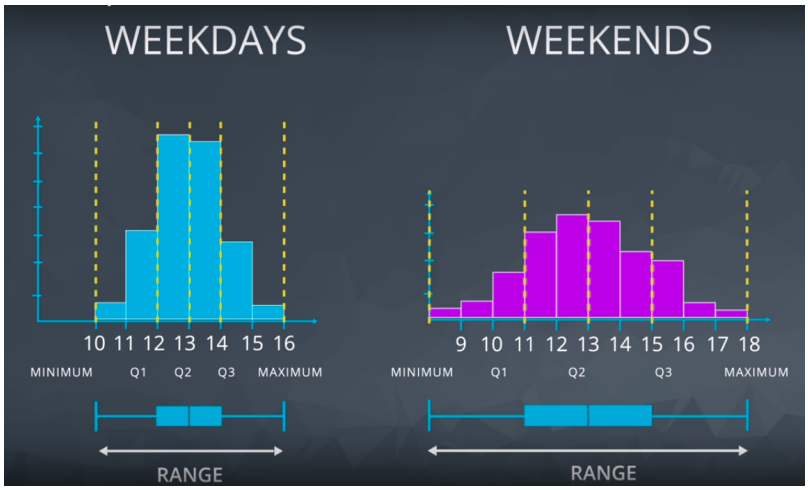


For Example:-

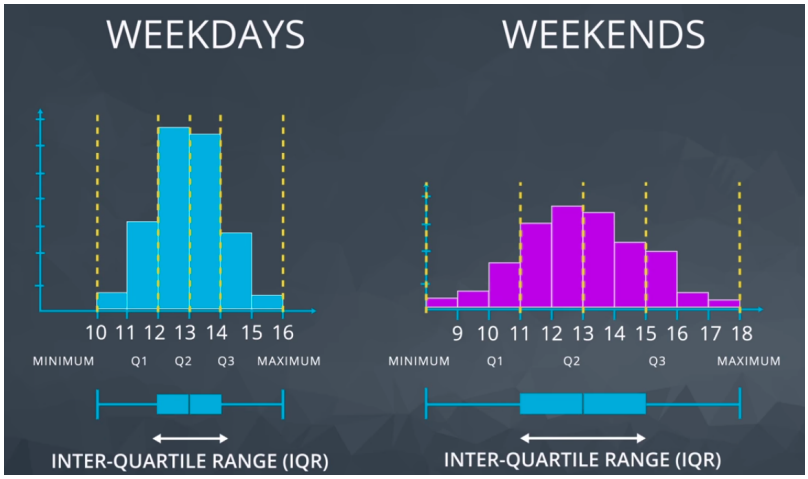
- **We can quickly see that the number of dogs I see on weekends varies much more than the number of dogs I see on weekdays.**



- **We can also visualize the distance from here to here as the range**

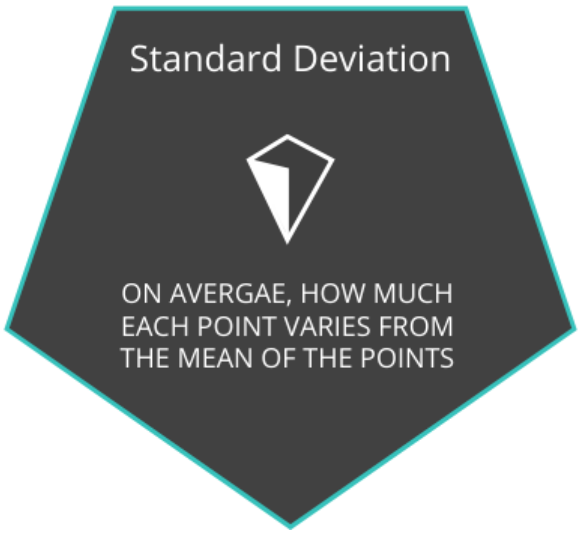


- **While the distance from here to here is the interquartile range**



Standard Deviation and Variance

- **The standard deviation is one of the most common measures for talking about the spread of data. It is defined as the average distance of each observation from the mean.**

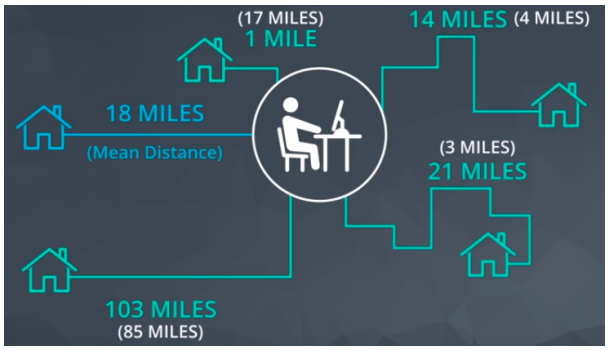


For Example:-

- **Imagine we wanted to know how far employees were located from their place of work. One person might be 15 miles, another 35, another only one mile, and another might be remote and is 103 miles. We could aggregate all of these distances together to show that the average distance employees are located from their work is 18 miles.**

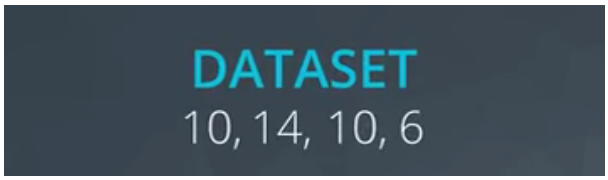
- **we want to know how the distance to work varies from one employee to the next. We could use the five number summary as a description. But if we wanted just one number to talk about the spread, we'd probably choose the standard deviation. For this example this is, how much on average the distance each person is from work differs from the average distance all of them are from work. So, this one is three miles farther from work than the average while this individual is four miles closer to work than the average.**

- **The standard deviation is how far on average are individuals located from this mean distance. So, it is like the average of all of these distances.**

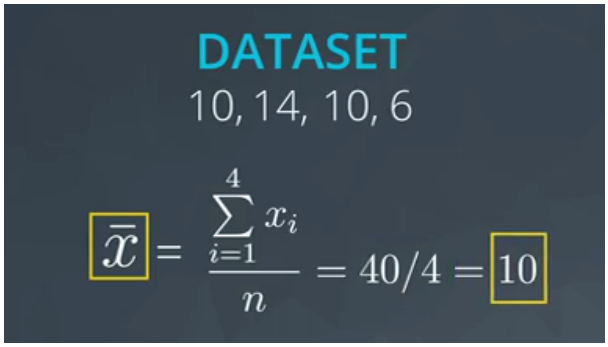


Standard Deviation Calculation

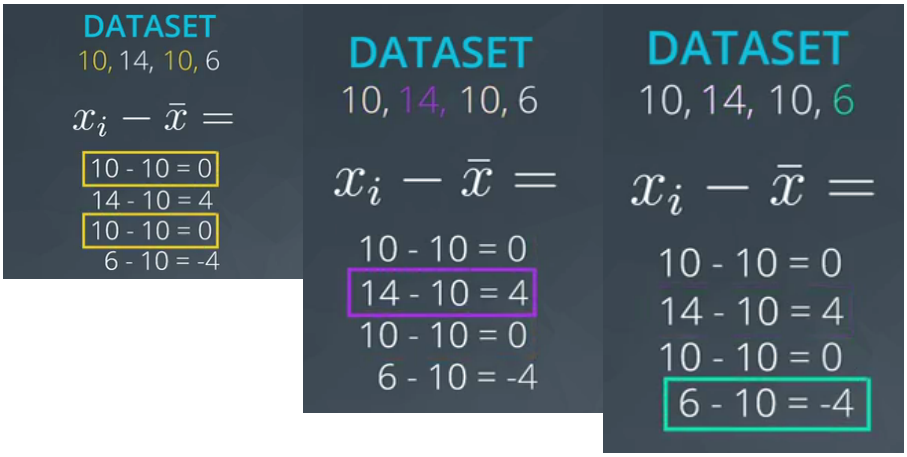
- **Imagine we have a data set with four values, 10, 14, 10, and 6.**



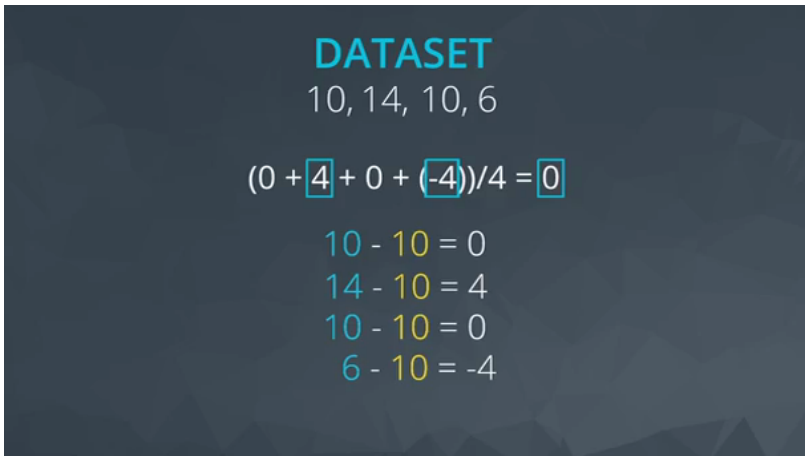
- **The first thing we need to do to calculate the standard deviation is to find the mean. In notation, we have this as X bar. For our values, the sum is 40, and we have four numbers. So the mean is 40 over 4 or 10.**



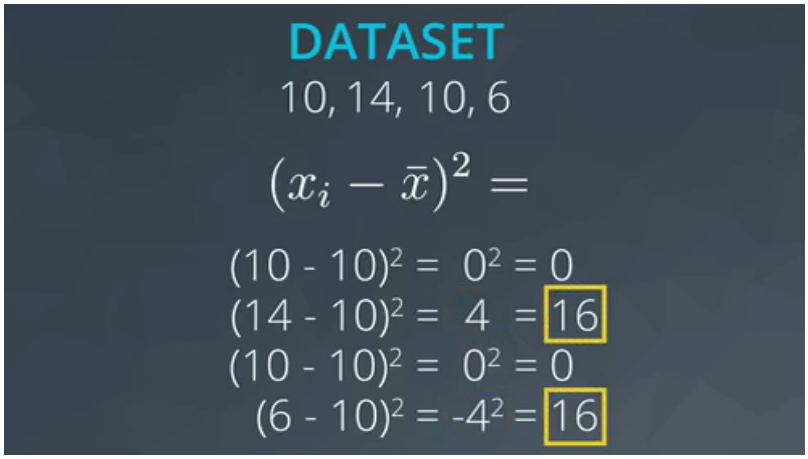
- **Then we want to look at the distance of each observation from this mean. Two of these observations are exactly equal to the mean. So the distance here is zero. One is 4 larger the 14, while the other is 4 smaller the 6.**



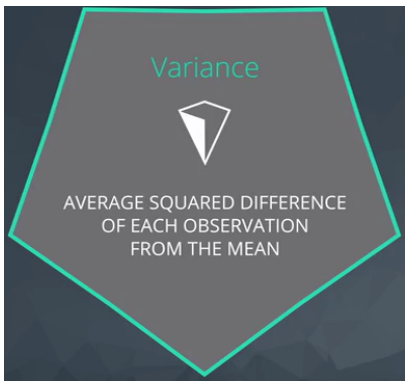
- **In notation, each of these is XI minus X bar. Then, if we were to average these distances, the positive would cancel with the negative value. And the value of zero isn't a great measure of the spread here. Zero would suggest that all the values are the same or that there's no spread.**

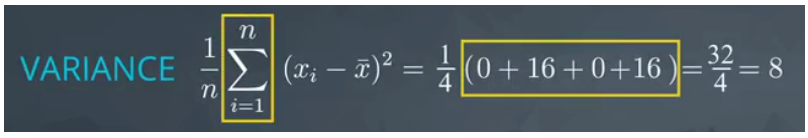


- **So instead, we need to make all of these values positive. The way we do this when calculating the standard deviation is by squaring them all. If we do that here, our negative and positive 4 values will become 16s.**

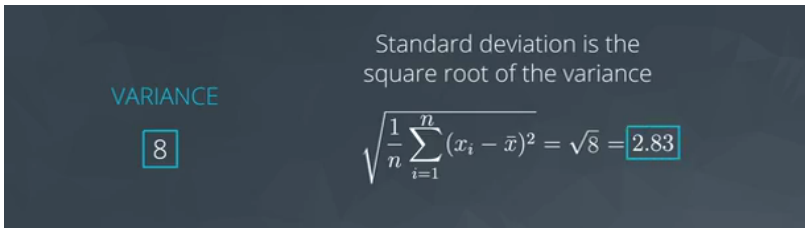


- **Now, we could average these to find the average squared distance of each observation from the mean. This is called the variance. Finding the average, just as we did before, means adding all of these values and dividing by how many there are. In our case, we had 0, 16, 0, 16 and we divide by 4 because we have four observations.**





- **However, this is an average of squared values which we only did to get positive values in the first place. So to get our standard deviation, we take the square root of this ending value. Here, our standard deviation is 2.83.**



Important Final Points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.

2. When comparing the spread between two datasets, the units of each must be the same.

3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.

4. The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

Use in the World

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

Measures of Center and Spread Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either `categorical` or `quantitative`. We then learned, we could identify quantitative variables as either `continuous` or `discrete`. We also found we could identify categorical variables as either `ordinal` or `nominal`.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: `lab` and `not lab`. We might say, 32% of the dogs were `lab` (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe `quantitative` variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

We looked at calculating measures of `Center`

1. **Means**
2. **Medians**
3. **Modes**

We also looked at calculating measures of `Spread`

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

Calculating Variance

We saw that we could calculate the `variance` as:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

You will also see:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

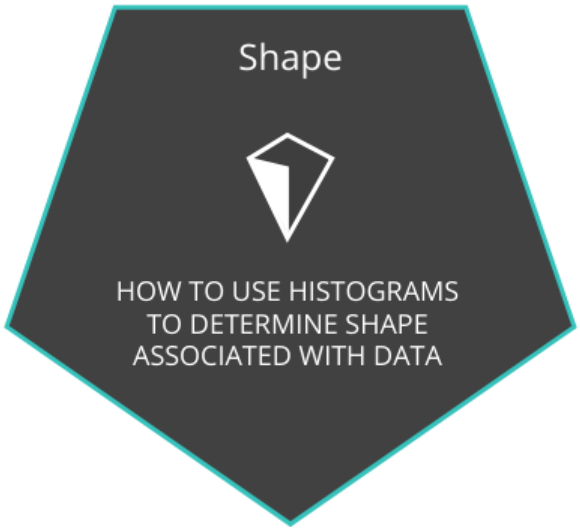
The reason for this is beyond the scope of what we have covered thus far, but you can find an explanation here.

You can commonly find answers to your questions with a quick Google search. Now is a great time to get started with this practice! This answer should make more sense at the completion of this lesson.

Standard Deviation vs. Variance

The standard deviation is the square root of the variance. In practice, you usually use the standard deviation rather than the variance. The reason for this is because the standard deviation shares the same units with our original data, while the variance has squared units.

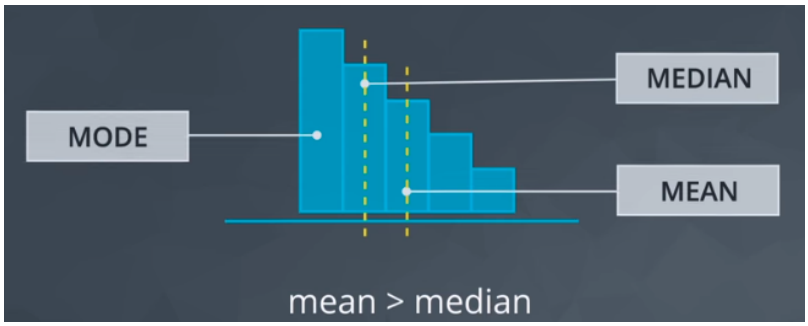
Shape



Symmetric Shape:-

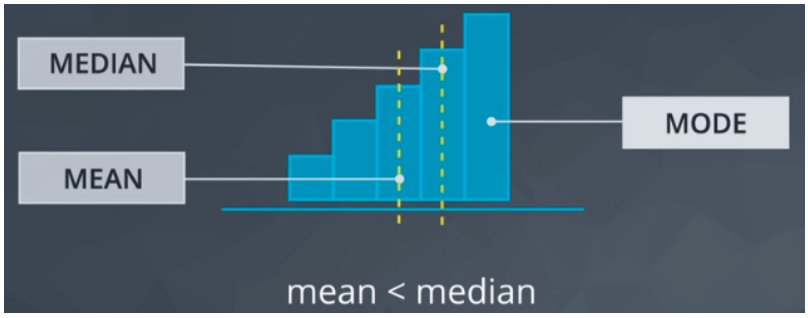
- **It also called Normal Distribution"The Bill Curve", when the data be in that shape (mean=median=mode) and the mode is the tallest bar in the histogram.**

Right Skewed Shape:-



- **The data longed to the right and the median close to the mode and after the median is the mean, mean > median.**

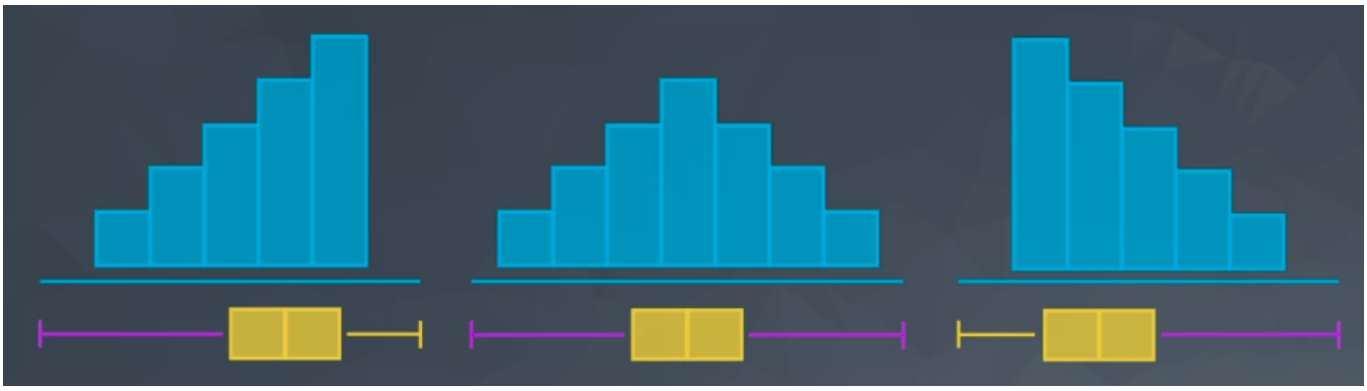
Left Skewed Shape:-



- **The data longed to the left and the median close to the mode and after the median is the mean, mean < median.**

Relate the shapes to the 5 Number summary:-

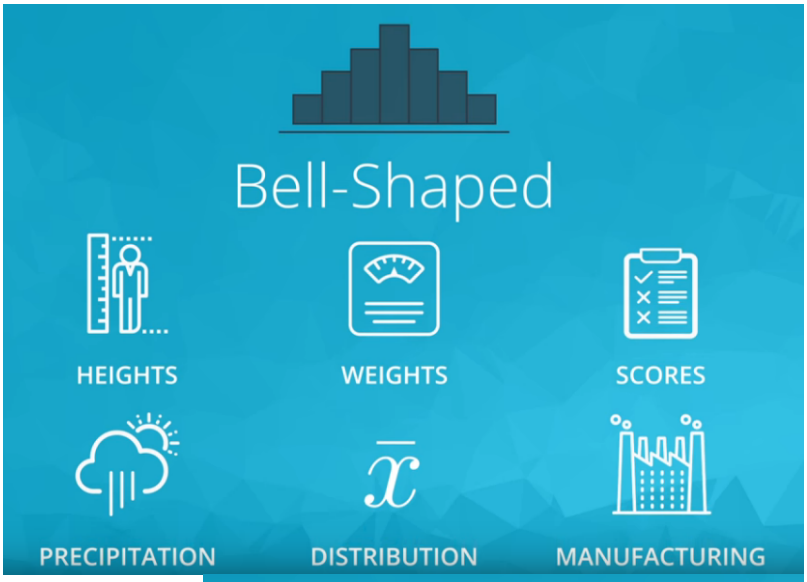
- **In order to relate this to the visual of a histogram, back to the five number summary, here are the corresponding box plots below each histogram. Notice how the whiskers stretch in the direction of the skew, for each of the skewed distributions. That is the longer whisker is on the left, for a left skewed distribution and it's on the right, for the right skewed distribution.**



The Shape For Data In The World

Some Examples of Bell Shaped Data:-

- **Heights**
- **Weights**
- **Scores**
- **Precipitation**
- **Distribution**
- **Manufacturing**



Examples Which Follow Left Skewed Shape:-

- **GPA**
- **Age OF Death**
- **Price**

Examples Which Follow Right Skewed Shape:-

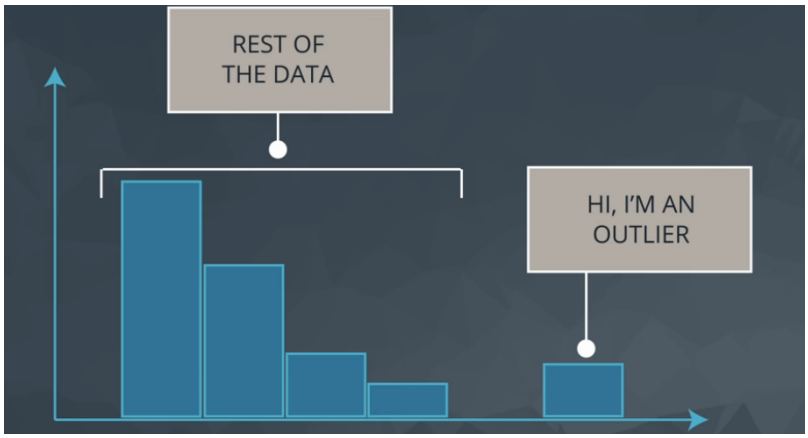
- **Blood**
- **Distribution OF Wealth**
- **Athletic Abilities**

#Shape and Outliers

- **outliers are points that fall very far from the rest of our data points. This influences measures like the mean and standard deviation much more than measures associated with the five number summary.**

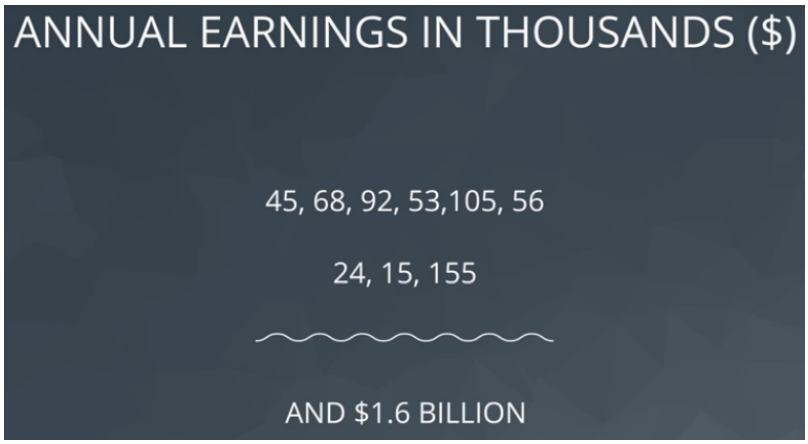
How to detect outliers:-

- **We could use Histogram to detect it.**

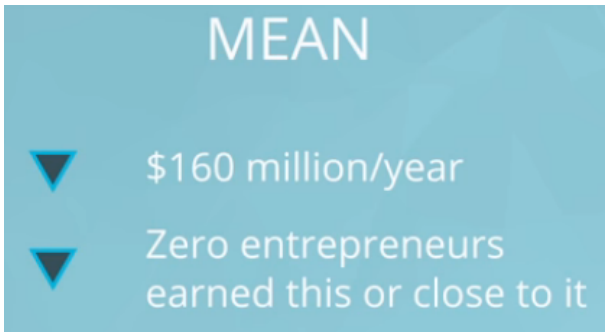


Effect of Outliers on five numbers summary:-

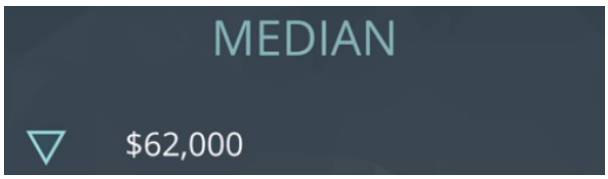
- **let's consider the salaries of entrepreneurs. Imagine I select ten entrepreneur earnings and I pull these nine values here as earnings in thousands of dollars, and the tenth is the CEO of Facebook.**



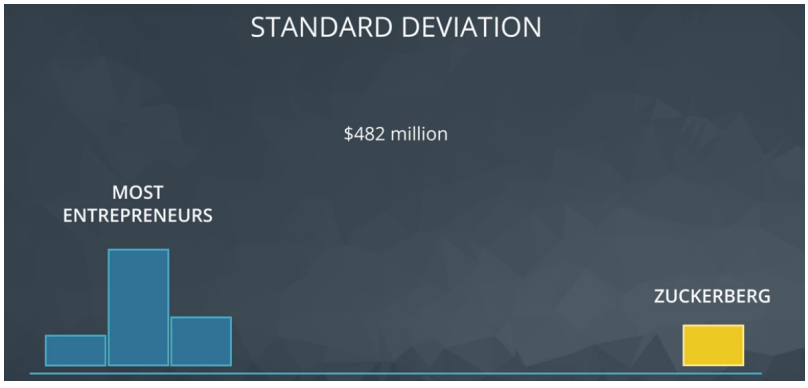
- **we can calculate the mean of these salaries for entrepreneurs based on this data to be approximately 160 million dollars. This is incredibly misleading. Literally zero of the entrepreneurs earned this salary. None of the ten salaries are even close to this amount.**



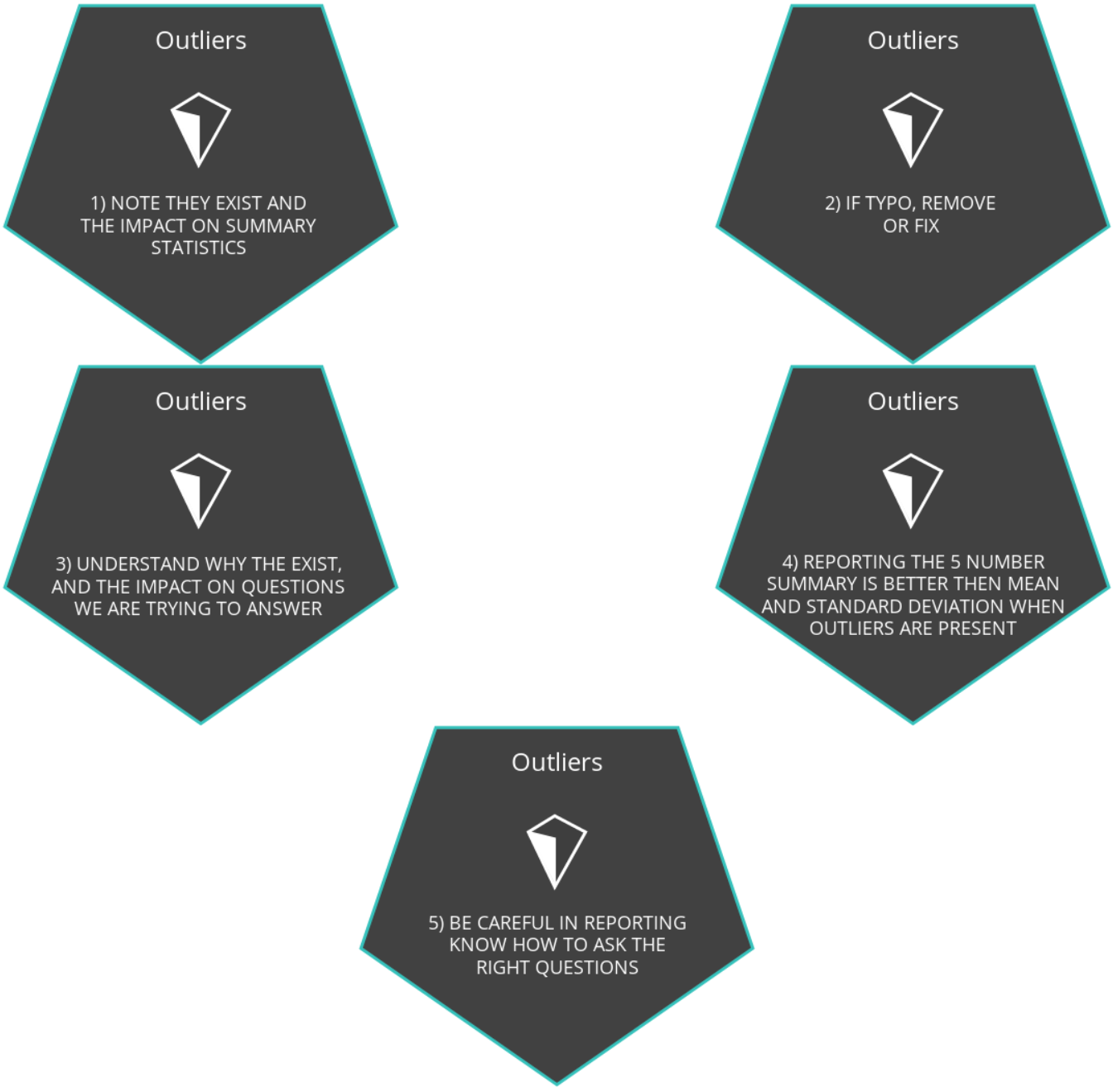
- **A better measure of center would certainly be the median. The median here is calculated at 62,000 dollars a year and is a better indication of what an entrepreneur is likely to earn based on our data.**



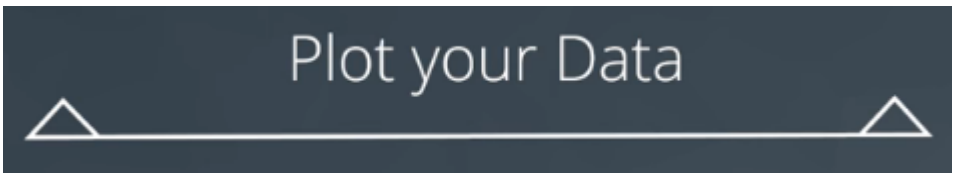
- **Our standard deviation is also not a great measure in this case. At approximately 482 million dollars, all this suggests is that our earnings for entrepreneurs are really spread out, but that really isn't fair either. Just one point is really far from the rest. Like really really far.**



Working With Outliers

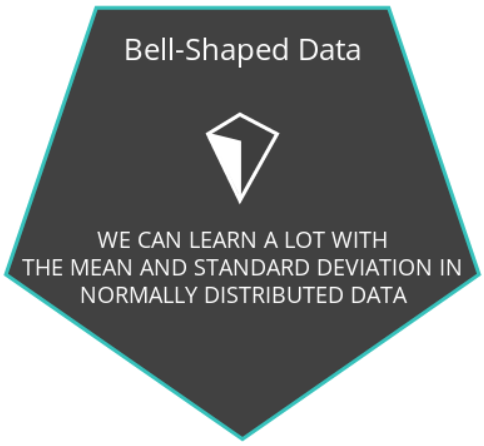


Other Advices:-

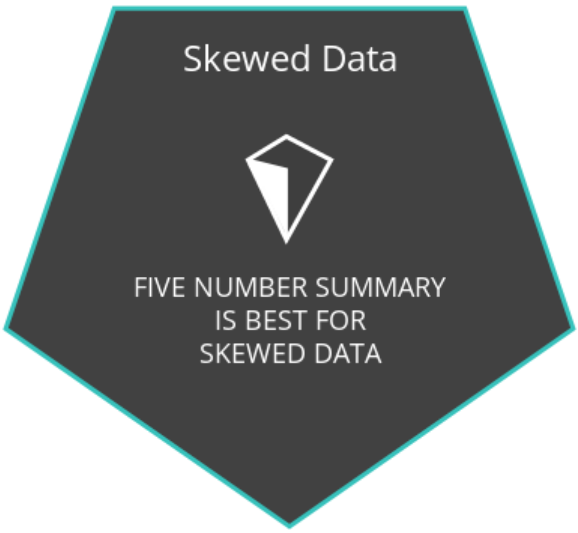
1. 

2. 

3. **If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.**

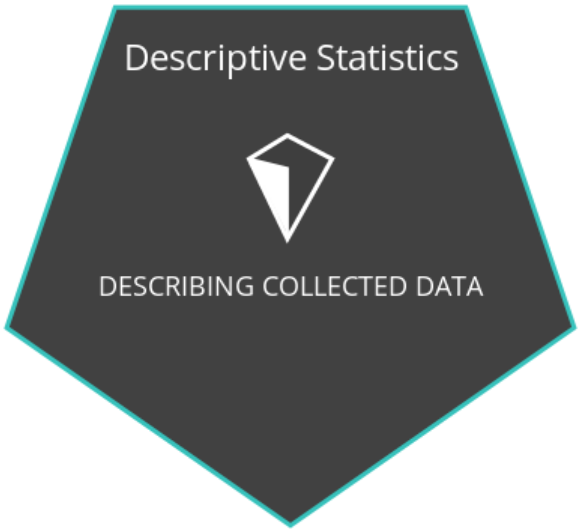


4.

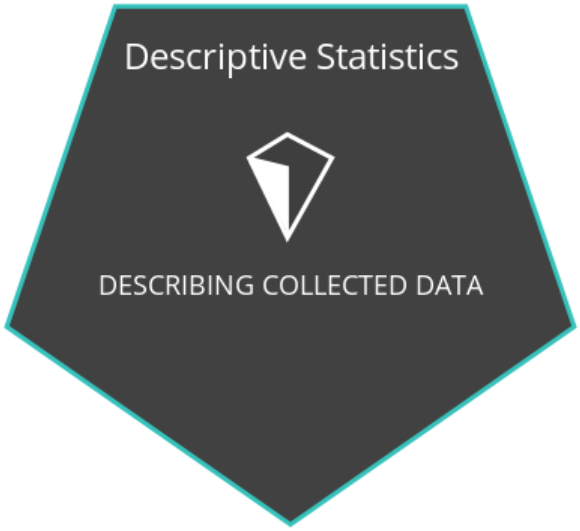


Descriptive vs. Inferential Statistics

Descriptive Statistics:-



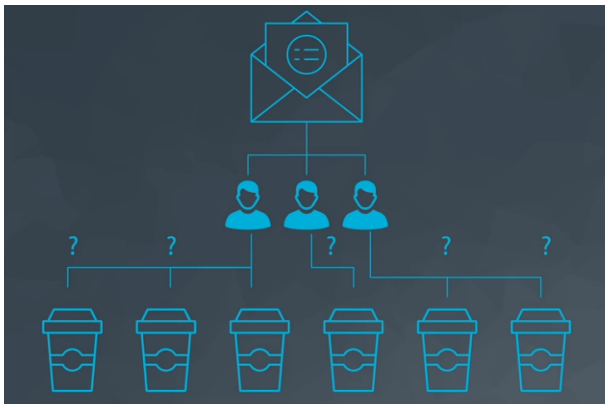
Inferential Statistics:-



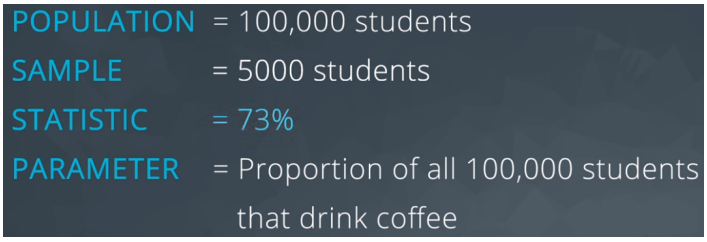
Example:-

- **what proportion of all Udacity students drink coffee. in order to get projects in on time, assume that you almost drink a ton of coffee.**

- **I send out an email to all Udacity alumni and current students asking the question, do you drink coffee?**

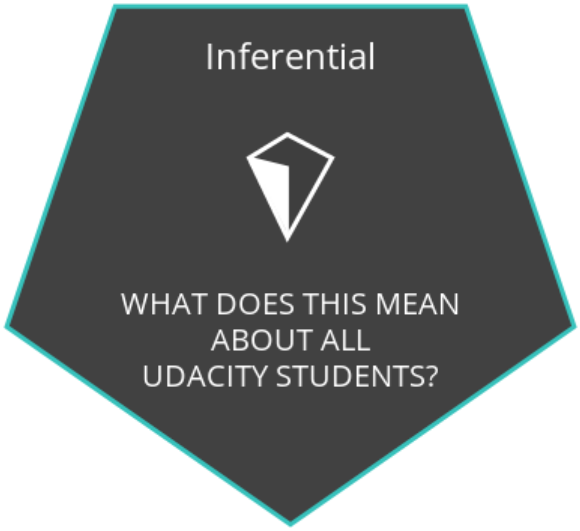


- **let's say the list contained 100,000 emails. Unfortunately, not everyone responds to my email blast. Some of the emails don't even go through. Therefore, I only receive 5,000 responses. I find that 73% of the individuals that responded to my email blast, say they do drink coffee.**



- **Descriptive statistics is about describing the data we have. That is, any information we have and share regarding the 5,000 responses is descriptive.**

- **Inferential statistics is about drawing conclusions regarding the coffee drinking habits of all Udacity students, only using the data from the 5,000 responses.**



- **The general language associated with this scenario is as shown here. We have a population which is our entire group of interest. In our case, the 100,000 students. We collect a subset from this population which we call a sample. In our case, the 5,000 students. Any numeric summary calculated from the sample is called a statistic. In our case, the 73% of the 5,000 that drink coffee. This 73% is the statistic. A numeric summary of the population is known as a parameter. In our case, we don't know this value as it's a number that requires information from all Udacity students.**