M2-05: Box Plots

Part of the "Exploratory Data Analysis" Learning Badge

Video Walkthrough: *https://discovery.cs.illinois.edu/m2-05/*

Box Plots and Quartiles

Just like histograms, **box plots** are used as a way to _____.

We can divide the data into equal-sized segments called _____:

"We Believe" (12 random responses)	Quartile	Percentiles
35 40 60	First 25% of data	
70 70 75	Next 25% of data	⇐ **Q1** - _____ Percentile
75 80 80	Next 25% of data	⇐ **Q2** - _____ Percentile
80 80 95	Final 25% of data	⇐ **Q3** - _____ Percentile

Review: Q2 is also known as the _____.

Calculating Q1, Q2, and Q3

When we calculate Q2, if it is between two numbers we find the **midpoint** between two numbers. This works because 50% of the data is below Q2 and 50% of the data is above.

When calculating Q1 and Q3, we do not _____!

Instead, we want to make the **best possible estimation** of the **true value** of Q1 (and Q3):

35	40	60	70	70	75	75	80	80	80	80	95
25%			**25%**			**25%**			**25%**		

If a quartile is between two values, we can use a formula to find the best estimation of their true value:

True Location =

Q1 (25%-tile) =

Interquartile Range (IQR)

The interquartile range (IQR) is the range of the middle 50% of the data. Like the SD, it's another measure of the spread:

IQR =

Outliers

Outliers are data points that differ significantly from most of the other points in the dataset. In other words, they "lie outside" most of the data. They are plotted as single dots on a boxplot. You can calculate outliers mathematically using these rules:

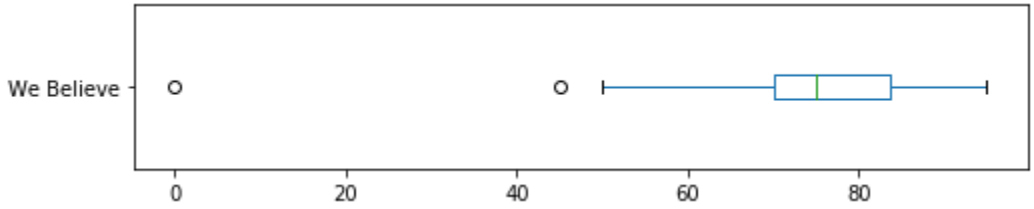
Low outliers:

High outliers:

How do outliers affect our data set?

1. Outliers will cause our _____ to be much greater or smaller than it would otherwise be.
2. Outliers will increase _____.
3. Outliers may affect the median, but if all the other values in the data set are close together, the outlier won't have much of an effect on the median.

Using and Drawing Boxplots:



Initial Code:	`import pandas as pd`
	`df = pd.read_csv(`
	`"https://waf.cs.illinois.edu/discovery/words.csv")`

Boxplot:	`# Makes the visualization a specific size:`
	`import matplotlib.pyplot as plt`
	`plt.figure(figsize=(8, 1.5))`
	
	`df["We Believe"].plot.box(vert=False)`