Lesson 3.5 - February 23, 2021

Ashish J.

Box Plots

- Box Plots can give you a five-number summary of a specific dataset
- Box plots basically give you a synopsis of what the dataset "looks" like without giving too much information either
- Box Plots are like 2-dimensional histograms as theey show the outliers and can help you compare multiple datasets
- There is a calculation to arbitrarily find which values are and are not outliers
- Whiskers of box plots determine the most extreme values that are still not classified as outliers by this arbitrary calculation
- The five numbers included in a five-numer summary
  - Minimum non-outlier value
  - Maximum non-outlier value
  - Median value ($Q2$)
  - Quartile 1 ($Q1$)
  - Quartile 3 ($Q3$)
- Dots after the maximum or before the minimum are considered outliers from the dataset

---

Comparing Two Box Plots

- Using a box plot, we can compare two given datasets by simply seeing them next to each other
- A single box plot can give context to another plot, especially if they graph the same data
- You can compare each one of the number from the five-number summary to get a holistic understanding of the two datasets and how they compare

---

Scatter Plots

- Scatter plots help visualize variability in a dataset

- Scatter plots help identify relations between two given variables
- This is not looking at cause and effect, but just association
- Association: as one value increases or decreases, what do we expect will happen to the other value that we are measuring?
- The main idea of scatter plots is trying to notice trends in the dataset
- Sometimes, the pattern that we find may be linear and may work quite well with incremented values, but in other cases, the pattern may be different, oftentimes including bell curves

Correlation Coefficient

- The correlation coefficient is found by graphing the best-fit line
- The correlation coefficient is on a scale between 0 and 1, or in other words, $c \in [0, 1]$
- A correlation coefficient of 1 signifies the data follows a strictly linear pattern
- A correlation coefficient of 0 signifiees that the data has no particular linear pattern
- In essence, the correlation coefficient tells us how close the best fit line is to being a perfectly straight line

Nuances

- When a problem asks to describe the relationship or trend that is shown in a particular graph or dataset, you should include the trend (positive/negative), the shape (linear/nonlinear), and the strength (how spread out the data is)
- The strength refers to the strength of the pattern or relation that is being referred to
- In certain cases, there appears to be no trend, in which case, the correlation coefficient will be 0