**Week 4 – LLO**

LO1 Use scatterplots to describe the relationship between two numerical variables making sure to note the direction, form and strength of the relationship as well as any unusual observations that stand out

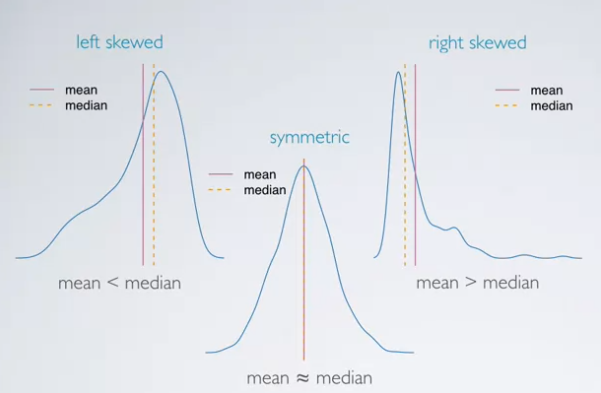
* Explanatory variable on the x axis
* Response variable on the y axis
* When evaluating the relationship on a scatterplot, think about whether the
  + Direction is positive or negative
  + Shape is linear, curved, some other shape
  + Strength is strong (little scatter, points are close together) or weak
  + Any outliers (either a point or a cluster of points)

LO2 When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations

LO3 Note that there are three commonly used measures of center and spread:

Center: mean, median, mode

* If these are calculated from a sample, they are called a sample statistic
  + A sample statistic is a point estimate of the population parameter
  + It is an ESTIMATE
  + But if the sample is good (at least, the sample is representative of the population), we can say that it is a good guess

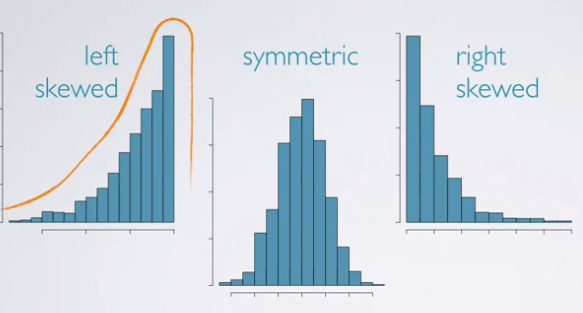


Spread: standard deviation, range, interquartile range

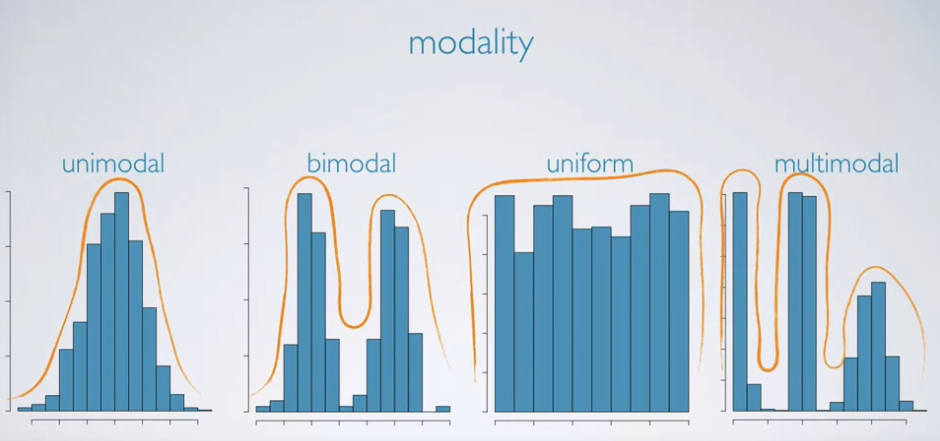
* Statistics that tell us about the variability of the data
* Variance:
  + The average square deviation from the mean
* Range: difference between thee max and min
  + Not reliable, depends on the end points of the values
* Interquartile range: range of the middle 50% of the data (25th to 75th percentile of data)

LO4 Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodal, multimodal, or uniform

* Skewness: distributions are skewed to the side of the long tail



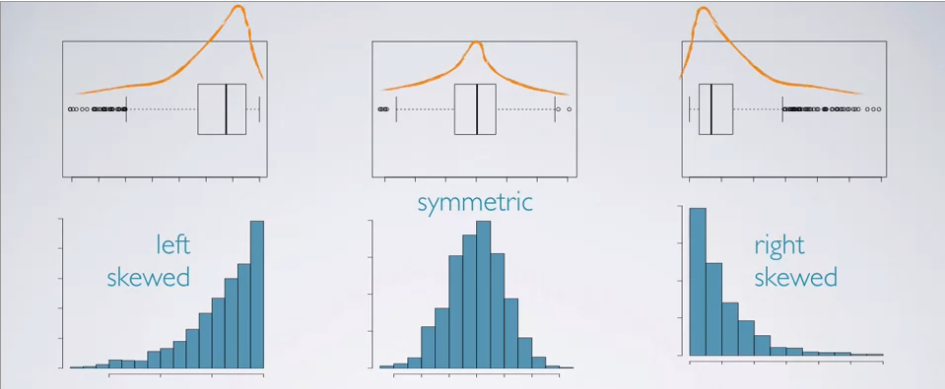
* Modality: related to where the peaks of the distribution are



* Most common is generally a symmetric, unimodal distribution i.e. the normal distribution (also called the bell curve)

LO5 Use histograms and boxplots to visualize the shape, center, and spread of numerical distributions, and intensity maps for visualizing the spatial distribution of data

* Histograms
  + provide a view of the data density
  + Used for looking at the distribution of the data
  + Note: important to choose an appropriate bin width for the histograms
* Boxplot
  + Displays the median, interquartile range, min and max and outliers
  + Good for highlighting outliers
  + Give a general idea of skewness



LO6 Define a robust statistic as a statistic that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics

* Robust statistics are measures on which extreme observations have little effect
* This means that adding an extreme point will not change the statistic much if it is robust
* Median is a more robust than the mean
* IQR is a more robust statistic than the variance and range
* Robust is better for describing symmetric distributions

LO7 Recognize when transformations can make the distribution of data more symmetric, and hence easier to model

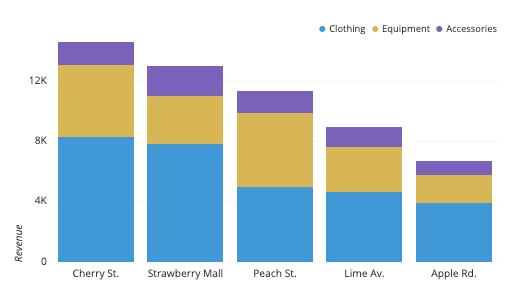
* Useful tricks to make certain types of data easier to model
* A rescaling of data using a function
* Good for when data is very skewed
* Common transformations
  + Log transformation
    - Applied when much of the data cluster near zero relative to the larger values in the data set and all observations are positive
    - Makes the relationship between variables more linear and hence easier to model with simple methods
  + Square root transformation
    - Plot the square root of the variable
  + Inverse transformation
    - Divide 1 by the variable
* Goals of transformations:
  + See the data structure differently
  + Reduce skew assist in modelling
  + Straighten a non-linear relationship in a scatterplot

LO8 Use frequency tables and bar plots to describe the distribution of one categorical variable

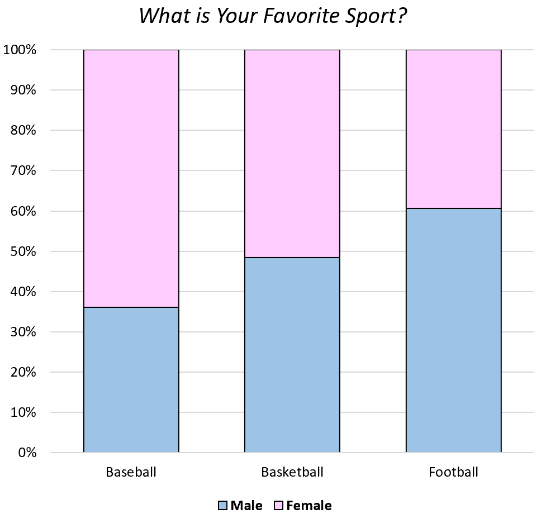
* Bar plots are used to look at the distribution of categorical variables rather than numerical variables
* X-axis is interchangeable
* Generally, a bar plot gives us the most information as opposed to a pie chart

LO9 Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables

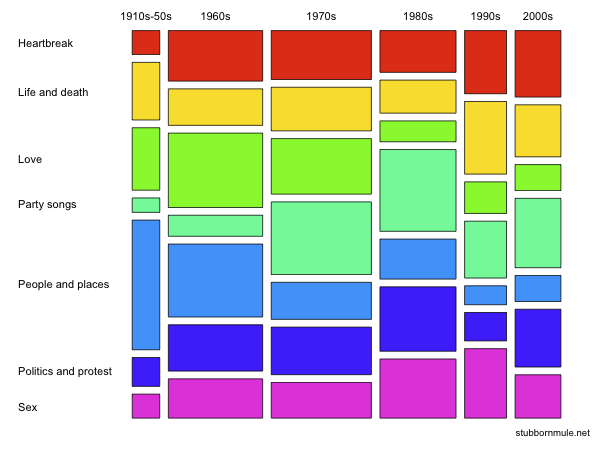
* Segmented bar plots are useful for visualizing conditional frequency distributions



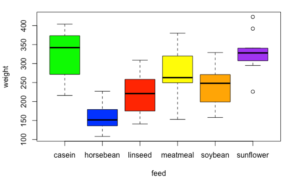
* Can also use segmented bar plots to visualize relative frequency



* A mosaic plot is also a good way to do so



LO10 Use side by side boxplots for assessing the relationship between numerical and a categorical variable



**Hypothesis Testing Framework**

