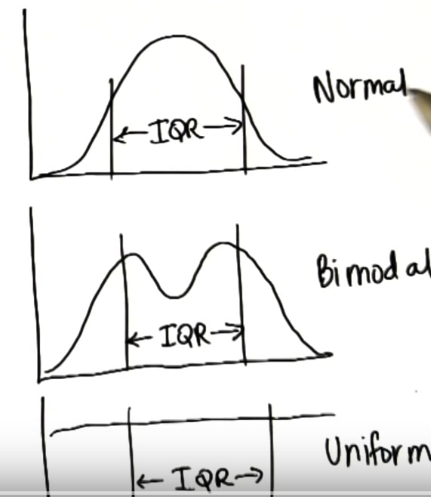
***Udacity Data Analyst Track***

**I. Into to Descriptive Stats**

4. Variability

* **Measures of Center** 🡪 *typically* describe the center of a distribution
* **Mean** = numerical average and is heavily affected by outliers = not **robust**
* X\_ = sample
* Mu (u) = population
* Good for normal distributions, pulled to right w/ positive skew, pulled to left w/ negative skew
* **Median** = data point directly in the middle of a data set (If 2 #’s are in the middle, the median is the average of the 2) which IS robust to outliers = outlier will not affect of median value
* Good for skewed distributions
* Even # of observations 🡪 (X(n/2) + X(n+1/2) / 2
* Odd # of observations 🡪 X(n+1/2)
* **Mode**. = most frequent data point and it robust to outliers as well.
* Data doesn't need to be numeric to find a mode 🡪 can also compute mode for categorical data, which would be the category/variable with the highest count
* Mode occurs on the X-axis, so you are looking for whatever *value* has the highest frequency.
* 7,000 and 1,000 are actual frequencies and the mode flavor would be "Plain” (7k)
* So the mode can be used to describe what type of data we have, numerical or categorical
* Normal distribution 🡺 Mean = Median = Mode 🡪 all the same
* Skewed left 🡪 mean < median < mode
* Skewed right 🡪 mean > median > mode
* **Outlier** = < Q1 – 1.5(IQR) or > Q3 + 1.5(IQR)
* Mean is not always w/in IQR (strong outliers)
* IQR doesn’t tell us all we need to know about a dataset (different distributions could have same IQR)
* 
* **Variability** = sum of differences between each data point + mean
* **Variance** = sum of differences between each data point + mean square = bad b/c not in same units as original data
* Find square root of variance = **standard deviation** = same units = square root of average squared deviations
* In general, samples *underestimate* variability in a population b/c samples tend to be values from the middle of distribution, especially for a normal distribution (where most values are in the middle of the distribution = are similar)
* Can mitigate this w/ **Bessel’s Correction =** *use n-1 for denominator for variance and SD rather than n*
* Makes them bigger
* Used if given a sample and want to estimate population SD
* *f*