**Stats & Excel Day 3**

**Gathering Data** often means finding data from various sources. Some examples include:

* Databases
* Files
* API;s
* Data Scraping tools
* Other online platforms

Data can be structured and unstructured.

**Transform and Clean** – in this step, we transform the data into a format that can be analyzed. This process can include:

* Cleaning (scrubbing)
* Preprocessing
* Combining data
* Feature Engineering – is a buzzword in ML, creating columns to structure data
* Data Reduction: cutting away – dimensionality reduction

Data cleaning is an important step in handling data. It ensures that the data is accurate, complete and consistent. The quality of data will directly impact the quality of insights and results.

**Explore** – in we explore the data to gain insights and understand the relationship between different variables. Exploratory data analysis includes:

* Visualizing the data
* Calculating summary statistics
* Identifying any apparent patterns or trends in data

This step will often loop back to transforming and cleaning the data. As we explore and make decisions/glean insights about the data.

**Analyze and Build Models** – we use statistical and ML methods to analyze the data and build predictive models. The goal of this step is to find relationships and patterns in the data, using these to provide valuable insights and inform/improve your decision making.

**Communicate Results** – we communicate the results of the analysis to the stakeholders. This can be done in a variety of ways:

* Visualizations
* Reports
* Presentations
* Dashboards

It is important to effectively communicate the results in a way that is accessible, easy to understand, and actionable for whatever audience you are presenting to.

**Descriptive Statistics** – the branch of stats that deals with the collection, analysis, interpretation, presentation, and organization of data. The goal of descriptive stats is to summarize and describe the main features of data set, including:

* Central Tendency
* Variability
* Distribution

**Normal Distribution** – Bell curve for central tendency Mean=Median=Mode.

**Skew** – data going in a direction, positive or negatively.

**Central Tendency** – is where the ‘meat’ of the data is. It is a statistical measure that describes the central value of a set of data. Mean, Median, and Mode.

**Variability** – Degree of spread, or the dispersion, of the data in the dataset. It is a measure of how much the values in the dataset differ from each other and from the average or mean values of that dataset. Variability is generally measured using range, variance, and the standard deviation.

**Outliers** – values that are unusual in our data set, typically falling outside the boundaries of normal distribution. Outliers can be troublesome when trying to perform analysis.

**Multimodal Distribution** – a distribution that has two or more peaks in it’s distribution. This occurs when we are representing different groups of data points with different ranges or values. They do not have a well-defined central tendency.

**Subpopulations** – a portion of segment of a larger population that has specific characteristics or attributes that differentiate it from the rest of the population. (i.e. marathon finish times)