**Statistics Note**

**Types of Statistics**

**Statistics**

(Statistics is the science of collecting, organizing, presenting, analyzing and interpreting numerical data to assist in making more effective decisions.)

* **Descriptive Statistics**

(Descriptive stats describes the data set that’s being analyzed but doesn’t allow us to draw any conclusions or make any inferences about the data.

* + - Measures of **Central** **Tendency**
    - Measures of Dispersion
    - Tables & Graphs).
* **Inferential Statistics**

(Inferential Stats is a set of methods that is used to draw conclusions or inferences about the characteristics of population based on the data from sample.

* + - * **Estimation**
      * **Hypothesis**)

**Business Stats**

**Exploration**

* **Univariate**
  + Categorical
    - Count, Count%
    - Pie Chart, Bar Chart
  + Numerical
    - Min, Max, Mean, Median, Mode
    - Range, Quantiles, Variance, Standard Deviation, Coefficient of Variation
    - Skewness, Kurtosis
    - Histogram & Box Plot
* **Bivariate**
  + Categorical & Categorical
    - Chi2(Sq) Test
    - Bar Chart, 2-Y axis plot
  + Numerical & numerical
    - Correlation
    - Scatter Plot
  + Categorical & Numerical
    - Z Test, T Test, ANOVA
    - Bar & Line Chart, 2 –Y Axis Plot

**Modeling**

* **Classification**
  + Frequency Tables
    - Zero R
    - One R
    - Naïve Bayesian
    - Decision Tree
  + Covariance Matrix
    - Linear Discriminant Analysis
    - Logistic Regression
  + Similarity Functions
    - K Nearest Neighbors
  + Others
    - Artificial Neural Network
    - Support Vector Machine
* **Regression**
  + Frequency Tables
    - Decision Tree
  + Covariance Matrix
    - Multiple Linear Regression
  + Similarity Functions
    - K Nearest Neighbors
  + Others
    - Artificial Neural Network
    - Support Vector Machine
* **Clustering**
  + Hierarchical
    - Agglomerative
    - Divisive
  + Partitive
    - K-Means
    - Self-Organizing Map
* **Association Rules**

**Types of data**

* **Categorical Data**

(Is non numeric, can be observed but not measured e.g. favorite color, place of birth etc.)

* **Quantitative Data**

(Quantitative data is numeric data which can be measured.)

* + **Continuous**

(Random variable which takes any value in its range of variation..e.g. height of a person.

* + **Discrete**

**(**Random variables which takes only isolated values in its range of variation..e.g. number of heads in 10 tosses of a coin.)

* + - **Nominal**

**(**Value do not have ordering…e.g. categorical variables. Like color, nationality and so on)

* + - **Ordinal**

**(**Values are ordered…e.g. satisfaction score)

**Measures of Central tendency**

* **MEAN**

It is just the average of the data computed as the sum of the data points divided by the number of points.

**Pros:** It is the easiest metric to understand and communicate

**Cons:** Mean is prone to presence of outliers

* **MEDIAN**

It is the value in the middle of the data set, when the data points are arranged from smallest to largest. Tricky circumstances: If there is an even number of data points, you will need to take the

Average of the two middle values.

* MODE