**Data Science**

-Identifying a hypothesis and trying to validate that hypothesis with your data (Academia)

-it’s data and curiosity

-the science and art of uncovering insights and trends that are hiding behind data

-using those insights to make strategic decisions

-using your data to find answers to the questions that you’re exploring

-the analysis of diverse data

-to get the most insightful and compelling answer to your research question

-use your tools to find order, meaning and value in the data

-Data Science provides insight and competitive advantage (Business)

NUMERICAL (Quantitative)

**Discrete** (Cardinal): discrete data is counted

**Continuous**: continuous data is measured

CATEGORICAL (QUALITATIVE)

- describes characteristics

**-** no mathematical meaning

- can assign numbers to them but can’t compare them mathematically (e.g., can’t be averaged)

ORDINAL

-the position that something is in a list)

-they can be averaged

**Data Visualization (Exploratory Data Analysis)**

-Critical first step

-Allows you to communicate your message to others who may not understand your data

-Allows you to get to know your data, to get a feel for it

-what’s the distribution…

-Allows you to explore your data before modeling it

-pay attention for hidden or unassumed details/patterns that may appear

-check for anomalies, unusual distributions and errors

-the data can make suggestions to you on how to approach your problem

**Exploring data (EDA):**

-you want to see what your data set is like

**Numerical** (quantitative exploration):

-df.describe()

-are there errors or missing data

-df.info()

-data types and quantity

-df.corr()

-correlations among the variables

-see if any assumptions are met

-you can manipulate the data (feature engineering, transformations)

**Dealing with Outliers**

-Trimmed Mean (to deal with outliers)

-you can trim a certain percentage (5-10%) of the values from each end

-Winsorized Mean (to deal with outliers)

-you move the outliers to the minimum or maximum value

**Fixed Skewed Data**

Tukey’s Ladder of Powers

-a log transformation can reduce the impact of outliers and enable us to make a more nuanced inspection of the data

-can undo the skew to get back to a more central distribution

**Visual exploration**:

-information dense

-best way to check shape, gaps, outliers

-can give the best overall impression

-can give better precision

-note outliers

-very quick way of getting an understanding of the features /variables in your dataset

-**Univariate** (distributions)

-histogram (shape of distribution)

-bar chart (compare quantity, by Category)

- box plot (observe a quantitative variable)

-**Bivariate** (joint distributions):

-scatterplot (2 quantitative continuous variables, strength of association)

-you can see correlations

-you can see clusters

-pair plot/matrix (paired quantitative variables)

-grouped box plot (Categorical vs. Quantitative)

-**Multivariate** (distributions):

Get multiple perspectives on your data

This sets the stage for the modeling of your data

**Scaling the data**:

-gets the data in the same range

Normalization (min-max scaling) tries to get the values closer to the mean

Standardization (z-score) is better with outliers that you want to capture

**Skewness**:

-an asymmetry in the data distribution

-the Square root is weaker than the log transformation.

-if there are negative values, the cubed root is more appropriate

**Hypothesis testing**: (Null Hypothesis. Alternative Hypothesis)

-Evaluating if there is a difference between groups

-significance level (alpha)

-a T-test is a way to compare the means of 2 different populations

-we can test if the means are different between 2 groups

-most of these tests assume that the data conforms to a certain distribution

(so you shouldn’t use them blindly)