**Note: Data Analysis with Python**

**Week 1:**

1. Defining Problem Statement
   1. Setting goals and objectives
   2. Analytical Approach
2. Data Requirements & Data Collection

**Week 2:**

**Data Wrangling:**

Data wrangling is the process of converting data from the initial format to a format that may be better for analysis.

1. Identify and handle missing values
2. Data Formatting
3. **Data Normalization**: (centering/scaling):
   1. Data normalization can be defined as a process designed to facilitate a more cohesive form of data entry, essentially ‘cleaning’ the data. When you normalize a data set, you are reorganizing it to remove any unstructured or redundant data to enable a superior, more logical means of storing that data.
   2. The main goal of data normalization is to achieve a standardized data format across your entire system. This allows the data to be queried and analyzed more easily, which can lead to better business decisions.
   3. Examples
      1. Miss ANNA will be written Ms. Anna
      2. 4158488400 will be written 415-848-8400
      3. 37 buttercup AVE will be written 37 Buttercup Avenue
      4. Amazon will be written Amazon.com, Inc.
      5. VP product will be written Vice President of Product
      6. Price range should be between 5000 – 10000. Over/under these values, column won’t store the values.
4. **Data Binning :** Data binning, also called data discrete binning or data bucketing, is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values which fall into a given small interval, a bin, are replaced by a value representative of that interval, often a central value.
5. **Turning Categorical values into numeric variables**

**Data Normalization Methods/Techniques:**

This technique is very important to understand the data in data pre-processing process.

When it comes to build a model that predict the outcome of variables, if needed, the data normalization is very important and it can be achieved through several techniques such as;

1. **Simple Feature Scaling**

In this method, simply divide each value by maximum value of the feature and formula would be:

**Xnew = Xold / Xmax**

1. **Min-Max Method**

In this method, we simply divide the difference of Xold and Xmin to the difference of Xmax and Xmin.

**Xnew = (Xold – Xmin) / (Xmax – Xmin)**

1. **Z-score or Standard Score**

In this method, we divide the difference of Xold and Xavg with the standard deviation of the feature variable.

**Xnew = (Xold – Xavg) / Xstd**

**Or** ( **Note**: The resulting values hover around zero, and typically range between -3 to +3 but can be higher or lower.)

**Xnew = (Xold – Xavg) / standard deviation**

**Note:** All these methods result in the providing the Xnew value between 0 to 1.

**Data Binning:**

**Data Standardization**

Data is usually collected from different agencies in different formats. (Data standardization is also a term for a particular type of data normalization where we subtract the mean and divide by the standard deviation.)

**What is standardization?**

Standardization is the process of transforming data into a common format, allowing the researcher to make the meaningful comparison.

**Example**

Transform mpg to L/100km:

In our dataset, the fuel consumption columns "city-mpg" and "highway-mpg" are represented by mpg (miles per gallon) unit. Assume we are developing an application in a country that accepts the fuel consumption with L/100km standard.

We will need to apply **data transformation** to transform mpg into L/100km.

## Data Normalization

**Why normalization?**

Normalization is the process of transforming values of several variables into a similar range. Typical normalizations include scaling the variable so the variable average is 0, scaling the variable so the variance is 1, or scaling the variable so the variable values range from 0 to 1.

**Example**

To demonstrate normalization, let's say we want to scale the columns "length", "width" and "height".

**Target:** would like to normalize those variables so their value ranges from 0 to 1

**Approach:** replace original value by (original value)/(maximum value)

**Binning**

**Why binning?**

Binning is a process of transforming continuous numerical variables into discrete categorical 'bins' for grouped analysis.

**Example:**

In our dataset, "horsepower" is a real valued variable ranging from 48 to 288 and it has 59 unique values. What if we only care about the price difference between cars with high horsepower, medium horsepower, and little horsepower (3 types)? Can we rearrange them into three ‘bins' to simplify analysis?

We will use the pandas method 'cut' to segment the 'horsepower' column into 3 bins.

**Indicator Variable (or Dummy Variable)**

**What is an indicator variable?**

An indicator variable (or dummy variable) is a numerical variable used to label categories. They are called 'dummies' because the numbers themselves don't have inherent meaning.

**Why we use indicator variables?**

We use indicator variables so we can use categorical variables for regression analysis in the later modules.

**Example**

We see the column "fuel-type" has two unique values: "gas" or "diesel". Regression doesn't understand words, only numbers. To use this attribute in regression analysis, we convert "fuel-type" to indicator variables.

We will use pandas' method 'get\_dummies' to assign numerical values to different categories of fuel type.