# Week 1 and Week2

# Data Pre-Processing

* Read default directory name
* Concatenate directory with files
* Use concatenated string to read .csv file using panda data frame
* Identify categorical variables
* Convert those variables into category
* Show dimension and basic stat of the data set

# Missing value

## Visualize Missing Value

* Find column wise number of missing values
* Convert into percentage term
* Print missing value count
* Draw the graph showing missing value count
* Try out package missingno (optional)
* Identify columns automatically with missing value percentage greater than 30% and Drop those columns

## Impute missing value by mean or by median or by most frequent KNN imputation

* Impute missing values by mean/median/most frequent technique
* Do not impute missing value of categorical variables

# Feature Engineering

* Create additional columns based on existing columns

# Encoder

* Encode categorical variables using
  + Ordinal encoder
  + target encoding

# Scale data

* Scale numeric columns of the data frame

# Outlier

## visualize

* Visualize outlier of each column

## Detect outlier

* by ZSore
* by IQR
* by clustering
  + DBScan
    - Identify outlier row numbers
    - Visualize outliers
  + K Means
    - Draw within-groups sums of squares against the number of clusters graph
    - Find the number of clusters from the graph
    - Visualize cluster graphically

# Week 3, Week 4 & Week 5

# Feature Engineering

* Write function to create binned column from existing numeric columns automatically
* Write function to create a numeric column from categorical column to show % distribution of each category
* Write a function which converts a numeric column into categorical column if number of unique values for that column is less than 1%

# Visualization

## SNS/ggplot/panda

* Categorical scatter plot with hue
  + strip
  + Swarm
* Categorical distribution plots
  + Box
  + violen
  + Boxen
* Categorical estimate plots
  + Point
  + Bar
  + Count
* Joinplot
* Displot
* Pairplot
* rugplot
* Line plot
* Numerical scatterplot
* Multiple relationships with facets

# PCA

* Split dataset into train/test
* Scale data
* Apply PCA

# Week 6, Week 7 & Week 8

# Modelling

* + Linear regression
  + Random forest