**Assignment 7- Data Cleaning and Processing**

Dataset - NYCFlights

1. List all flights scheduled for 1 Jan
2. Find flights that were not delayed more than 2 hours
3. Find all flights
   1. Had an delay of 2 or more hours
   2. Flew to Houston (IAH or HOU)
   3. Were operated by Unites, American or Delta
   4. Departed in summer(March, April , May)
   5. Arrived more than 2 hours late but didn’t leave late
   6. Departed between midnight and 6.00 am
4. Sort flights to find most delayed flights
5. Which flights travelled longest?
6. What are different ways to select dept\_time, dept\_delay,arr\_time and arr\_delay from flights
7. Convert dep\_time and sched\_dep\_time to number of minute from midnight for further calculation

Q2. This data is for building binary predictive model to predict the possibility of patient having heart attack or not.

Dataset - HeartDieases

1. Load csv and **set value -9 to NA**

1. Processing will be focused on first 58 columns.
   1. Take subset of first 58 columns
2. Values 58 column defines how many times a patient had a heart attack
   1. Since for building predictive model, let’s focus on only 2 values: 0 means no attack, 1: possibility of attack. Therefore values > 1 should be reset to 1
3. Since its a binary predictive model, we will process only numeric data
   1. Check data types of each column
   2. Drop columns which are having strings or objects
4. Drop columns which have all null values
   1. Check columns which have all null values // is.na, all()

V5', 'V6', 'V7', 'V45', 'V46', 'V47', 'V48', 'V49', 'V52', 'V53

* 1. Drop the columns which have all null data

1. Find rows with missing values greater than 50%
2. Fill missing values with mode value for all columns which are float or int
   1. Print columns that have minimum one NaN value
   2. For each value, Update Nan values with mode value of that column

Note: character vectors and factors are imputed with the mode. Numeric and integer vectors are imputed with the median.

* 1. Verify that no more Na values

1. Remove columns which have same value in all rows // since such column can not contribute to prediction
2. Check summary of data set for value range inspection
3. Save the cleaned data set