Data Mining & Predictive Modelling

Lab Assign-03- Data Pre-processing & Data Cleaning Functions

* Read the datasets "ToyotaCorolla.csv" and "ToyotaCorolla-Dirty.csv" as per the requirements.
* Test following data pre-processing and data cleaning functions
* You need to install following packages and use following libraries
  + install.packages("tidyverse") install.packages("Hmisc")
  + library(tidyverse) library(dplyr) library(ggplot2) library(Hmisc)
* Save your R command file.
* Save your code and results in a word file and submit it on the VOLP.
* FUNCTIONS
  + rename() - for rename columns
  + select() - for selecting columns
  + slice() - for selecting rows
  + sample\_n() -- for sampling rows
  + arrange() -- for arranging rows by column values
  + filter() -- for filtering rows
  + mutate() -- for transforming values in a field
  + is.na() --- find missing data
  + sum(is.na()) -- find number of missing values
  + mean(is.na()) -- find % of missing values
  + na.omit() -- deleting incomplete observations
  + mean(<column>, na.rm = TRUE) -- find mean leaving records with missing data
  + complete.cases() --- to detect rows with complete data
  + clean=na.omit() --- to remove incomplete rows or data
  + glimpse(data) -- summarise dataset
  + boxplot.stats() - for detecting outlier stats
  + boxplot() - for comparing boxplots
  + boxplot.stats()$out -- for setting outlier detection limit

**Numerical imputation**

* + dirty$column[is.na(column)] <- mean(column, na.rm = TRUE) --- mean imputation
  + dirty$column <- impute(dirty$column, fun = mean) # mean imputation
  + dirty$column <- impute(dirty$column, fun = median) # median imputation

**You are required to pre-process 1 more data-set of your choice**

\*\* \*\* \*\*