

## List of Practical:

### 1. Find the below data set and perform the following operations:-

**Dataset name:** -mtcars\_DataDescription

1. Read the dataset.
2. Find the head of the dataset.
3. Find the Datatype of Dataset (each column).
4. From the given dataset 'mtcars.csv', plot a histogram to check the frequency distribution of the variable 'mpg'(Miles per gallon).
5. Find the highest frequency of interval.
6. Which can be inferred from scatter plot of 'mpg'(Miles per gallon) vs 'wt' (Weight of car) from the dataset mtcars.csv.

### 2. Find the below data set and perform the following operations:-

**Dataset name:** -Churn\_DataDescription

1. Find the no. of duplicate records in the churn dataframe based on the customerID column.
2. In the churn dataframe, what are the total no. of missing values for the variable TotalCharges?
3. From the churn dataframe, what is the average monthly charge paid by a customer for the services he/she has signed up for?
4. In the churn dataframe, under the variable Dependents how many records have
5. Find the data type of the variable tenure from the churn dataframe.

### 3. Find the below data set and perform the following operations:-

**Dataset name:** -Diamond\_DataDescription

1. Plot a boxplot for "price" vs "cut" from the dataset "diamond.csv" Which of the categories under "cut" have the highest median price?
2. Create a frequency table (one-way table) for the variable "cut" from the dataset "diamond.csv" What is the frequency for the cut type "Ideal"?
3. Show the subplot of the diamond carat weight distribution.
4. Show the subplot of diamond depth distribution.
5. Build the Model using linear regression and find the accuracy.

Reference link:-<https://www.kaggle.com/shrutisaxena0617/exploring-diamonds-dataset>

### 4. Use the dataset named "People Charm case.csv" that deals with HR analytics and answer the following questions:-

1. Which of the variables have missing values?

2. What is the third quartile value for the variable “lastEvaluation”
3. Construct a Crosstable for the variables “dept” and “salary” and find out which department has highest frequency value in the category low salary.
4. Generate a boxplot for the variable “numberOfProjects” and get the median value for the number of projects where the employees have worked on.
5. Plot a histogram using the variable “ avgMonthlyHours” and find the range in which the number of employees worked for 150 hours per month?
6. Generate a boxplot for the variables “lastEvaluation” and “numberOfProjects”

**5. Use the dataset named “People Charm case.csv” that deals with HR analytics and answer the following questions:-**

1. Build a Logistic Regression model using all the variables. Use 75% of the data as the training set  
and fix the random state as 2. The accuracy score for the predicted model is?
2. Build a Logistic Regression model using all the variables. Use 75% of the data as the training set  
and fix the random state as 2 and find out how many samples are misclassified?
3. Build a k-Nearest Neighbors model using all the variables. Use 75% of the data as the training set, fix the random state as 0 and the k value as 2. The accuracy score for the predicted model is?

**6. Problem Description:**

Data from an online microlending platform has been collected. This data contains details of the purpose for which the loans would be used and how the loan is funded. Additional information on the country of loan recipient and the poverty levels of the country are also given.

It is to be seen whether a loan would be funded or not based on the available data.

**Variable Description:**

Parameter	Description
activity	Activity for which loan was requested
borrower_genders	Gender of the borrowers
country	Country in which loan was disbursed
country_code	ISO country code
currency_policy	The currency policy in which loan was disbursed
distribution_model	Loan disbursed through field partner or not
lender_count	the total number of lenders that contributed to this loan
original_language	language of the original loan application
loan_amount	The amount disbursed by the field agent to the borrower(USD)
repayment_	interval interval between payments
sector	High level category
status	The status of a loan : whether funded,not funded

term_in_months	The duration for which the loan was disbursed in months
rMPI	Multiple Poverty Index

1. How many columns are of “object” data type? Read the given data “lendingdata.csv” and save it as a dataframe called data, and answer the questions below:-
2. Find the total number of missing values in the data set?
3. Identify which of the columns contain redundant information and can be dropped from the dataframe.
4. What is the third quartile value of the variable “loan\_amount”
5. What is the percentage split of the different categories in the column “ repayment\_interval” after dropping the missing values?
6. What is the minimum loan amount disbursed in the Agriculture sector?

**7. Identify what the web page is about using NLTK in Python.**

Reference link:-<https://towardsdatascience.com/gentle-start-to-natural-languageprocessing-using-python-6e46c07addf3>

**8. Detecting the Spam or Ham using the NLP Programming.**

Reference link:-<https://towardsdatascience.com/spam-or-ham-introduction-to-naturallanguage-processing-part-2-a0093185aebd>

**9. Simple application of sentiment analysis using natural language processing techniques.**

Reference link:-<https://dzone.com/articles/simple-sentiment-analysis-with-nlp>