**REPORT**

In this submission, a dataset named tips has been used. Tips dataset has the information various variables which are used to analyze data in it. There are 7 variables/columns in this dataset. The first variable ‘total\_bill’ denotes the total bill of the customer spent in the restaurant. Tips denotes the amount of tip paid by the customer. Sex and smoker variable shows the gender and smoking habit respectively. Time of the arrival is shown by the day and time variable and finally size shows the number of people accompanying the customer. The observations in the variable ‘total\_bill’ are floating number which contains missing values. Same goes with tips also. But the other variables like smoker day and time are strings, and finally size with integer values. We can observe that there are missing values in the dataset.

We shall use pandas to implement this dataset. Firstly, we create a data frame df to copy the tips.csv file into it.

**COMPREHENSIVE EDA OF THE DATASET**

EDA refers to Exploratory Data Analysis which performs various operations on the dataset to analyze the data. To determine EDA, we shall calculate the following for each variable.

* Is it categorical or continuous?
* If continuous variable then find Min, Max, and Average value.
* Find out the missing values.
* Handle missing values

According to the code used in the notebook, we have concluded the following.

**Total\_bill** : Continuous variable with 2 missing values.

**Tip**: Continuous variable with 2 missing values.

**Sex**: Categorical variable with no missing values.

**Smoker**: Categorical variable with no missing values.

**Day**: Categorical variable with no missing values.

**Time**: Categorical variable with no missing values.

**Size**: Categorical variable with no missing values.

There are 2 missing values in the variables total\_bill and tip.

**HANDLING OF MISSING VALUES**

A missing value can be handled by the following ways

* Replacing the observation with mean of the variable.
* Remove/drop the whole variable and its values.
* Creating machine learning algorithm to predict the missing value.

In this section we have replaced the observation with the mean value.

We now have a data frame with no missing values. Hence data is clean.

**UNSUPERVISED MACHINE LEARNING K-MEANS**

K means can be done only on variables with number but not on strings. Hence we shall drop variables having strings as observations. In such case, we shall get only 3 variables, total\_bill, tip and size. We shall now perform machine leaning on it. A scaler variable is used here which is used to display the array. The following code is used to determine K.

def find\_k (X, k\_range, sample\_percent=1):

"""

k\_range: a list of possible k values

X, the data we're clustering on

"""

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

import numpy as np

from scipy.spatial.distance import cdist

from sklearn.metrics import pairwise\_distances

N = X.shape[0]

sampleSize = X.shape[0] \* sample\_percent

if sampleSize > 0:

index = np.arange(np.shape(X)[0])

np.random.shuffle(index)

X = X[index, :]

mean\_distortions=[]

for k in k\_range:

#cluster using k, then calculate the mean distortion (average distance to closest centroid)

kmeans\_model = KMeans(n\_clusters=k, init='k-means++', n\_jobs=-1).fit(X)

mean\_distortions.append(sum(np.min(pairwise\_distances(X, kmeans\_model.cluster\_centers\_,

metric='euclidean'),axis=1)) / X.shape[0])

#visualize results

plt.plot(k\_range, mean\_distortions)

plt.xlabel("K Value")

plt.ylabel("Mean Distortion")

plt.title("Elbow Graph for Mean Distortion per K")

plt.show()