Submission by: Pavana Lakshmi Venugopal

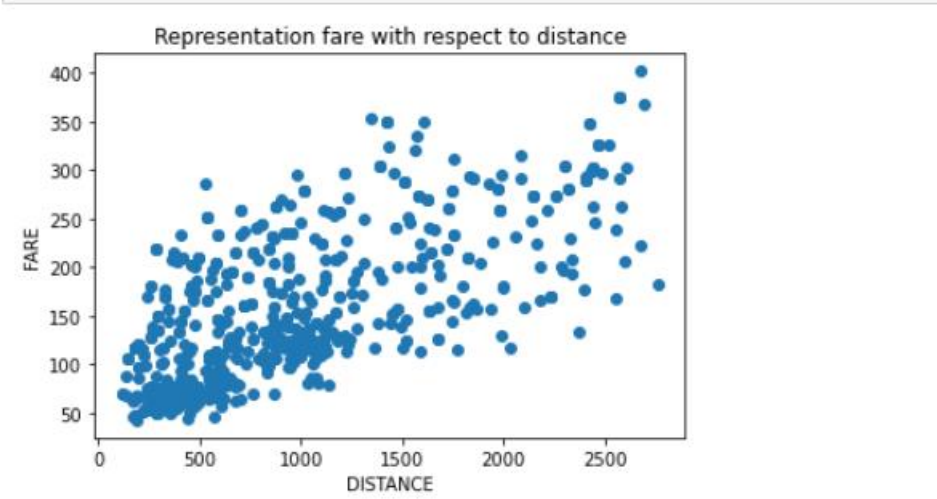
UBIT ID : 50464513

Part 1: Basic exploratory analysis

Airport congestion was one of the problems that took place in the late 1990s. This was because fares and routes were freed from regulation. Few carriers started nonstop service on routes. Here the firm is expecting a model to be designed to predict airfares on new routes. We are given the Airfares csv file to achieve this task. The steps to achieve this would be explained in this writeup. There are 638 rows and 17 columns in the dataset. The following columns are of the datatype int: NEW, S_POP, DISTANCE, PAX. During analysis of the dataset, I found that it does contain NULL values. The columns that have NA values are associated with S_INCOME, E_INCOME, E_POP. Chicago appears more frequently in the dataset for starting city and New York/Newark appears more frequently in the dataset for ending city. On analyzing the column SW I was able to conclude that the number of routes South west served were 194.

Part 2: Data visualization

1) I am creating a scatter plot to compare the results of two variables Distance and Fare.



On observing the plot, we see that as the distance increases the fare is increasing, goes high as 400$. A-lot of data is seen for a closer distance which shows that many people are travelling within 1000 miles. The distance given in the csv is in miles and hence the comparison. Only few people travel even with a higher fare. By looking at the scatter plot we can say that only few people spend more than 150 dollars on the Airfare. In general, travelling is more within 1250 miles when compared to miles greater than 1250. There are even lesser people travelling with fares greater than 300$.

2) Plotting a bar graph for vacation route.

The majority of the routes within the dataset is not vacation routes. From the bar graph we see that only around 150 routes are vacation routes and around 475 is non vacation routes. This is derived from the definition of the variable given where yes means it is a vacation route and No means it is not a vacation route.

Part 3: Data preprocessing
PART 3.1:

S_CODE column can be dropped because it is redundant when other columns are present. Such as S_CITY that would give the information required. With the same assumption I have decided to drop E_CODE as E_CITY would serve my purpose. Processing times hence would comparatively be reduced when dealing with additional data that might be added to our dataset in the future. Also, with the help of python we have analyzed that many values are NULL for both the columns. Hence concluded that I can drop these variables.

PART 3.2:

Since machine understands only numbers, we would have to deal with NA values. We have two choices either to drop them as in remove them entirely or impute them with appropriate values as we see fit. Here S_INCOME column of NA values has been given values as the mean. While for E_INCOME and E_POP columns we were directed to drop those specifics rows have NA. S_INCOME had 6 NA values, E_INCOME had 2 NA values, E_POP had 5 NA values. If given an option, I would decide to drop the NA values as the number is quite low so it should not impact my result. On the other hand, if I was given a very small data set say one which is having 100 rows then I would have decided to impute. After performing these operation 631 rows are left.

PART 3.3:

After flagging the categorical variables (S CITY, E CITY, VACATION, SW, SLOT, and GATE) columns have increased to 136. The get_dummies function converts the categorical variable into dummy/indicator variables. Dummy variables have 2 possible values 0 or 1. 1 encodes the presence of a category and 0 encodes the absence of a category. Since the machine understands only numerical values this is one of the best ways to convert categorical variables.

PART 3.4:

We do normalization for data to appear similar across all records and fields. It leads to cleansing, lead generation, segmentation, and higher quality data. It is one of the most important steps to get rid of errors. Overall, a reliable prediction is provided for air fares