<u>Assignment-2</u>

IE6600 20430 Computation and Visualization

Report By-

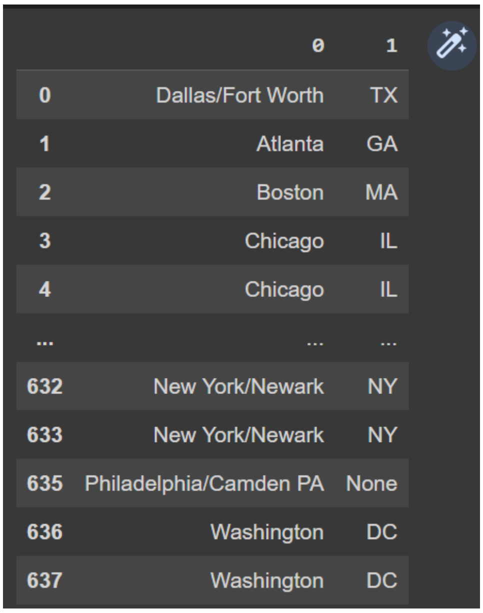
Divya, Nupur, Ranadheer, Varun

Data Description:

The dataset Airfares consists of structured data, consisting of 18 parameters(columns) and 638 entries(rows). Data talks about the average flight fares for various routes in the United States between the third quarter of 1996 and second quarter of 1997. Numerical data such as the average number of coupon where a one coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, distance is the space in miles between two endpoint airports, HI which is Herfindahl index that is measure of market concentration, average income of the starting and ending city, average personal income at the starting and ending city, Number of passengers on the route during the period of data collection and average fare of the route. Also, the non-numerical information such as whether or not they are vacation routes, Southwest Airlines serves that route, endpoint airport is gate constrained or not, endpoint airport is gate controlled or not. A measure for congestion in the destination airports is facilitated by the parameters slot control and gate constraint. Number of new carriers operating in the existing routes during this time period is also laid out in a separate column.

Thought Process:

1. We began the process by Data Cleansing as it ensures that we have the most recent files and important documents, so as to find them easily when required or needed. We bifurcated the states and cities and different tuples to have a profound understanding of the data.



	0	1
0	Dallas/Fort Worth	TX
1	Atlanta	GA
2	Boston	MA
3	Chicago	IL
4	Chicago	IL
...	...	...
632	New York/Newark	NY
633	New York/Newark	NY
635	Philadelphia/Camden PA	None
636	Washington	DC
637	Washington	DC

2. Then in next step we explored the data and meticulously paid attentions to the trends and variations of the Dataset and tried to find anomalies and patterns. We observed that under E_CODE, few of the records went missing which made it anonymous etc. Through this we came across few visualizations as stated below for the airfares dataset.

Visualization Description:

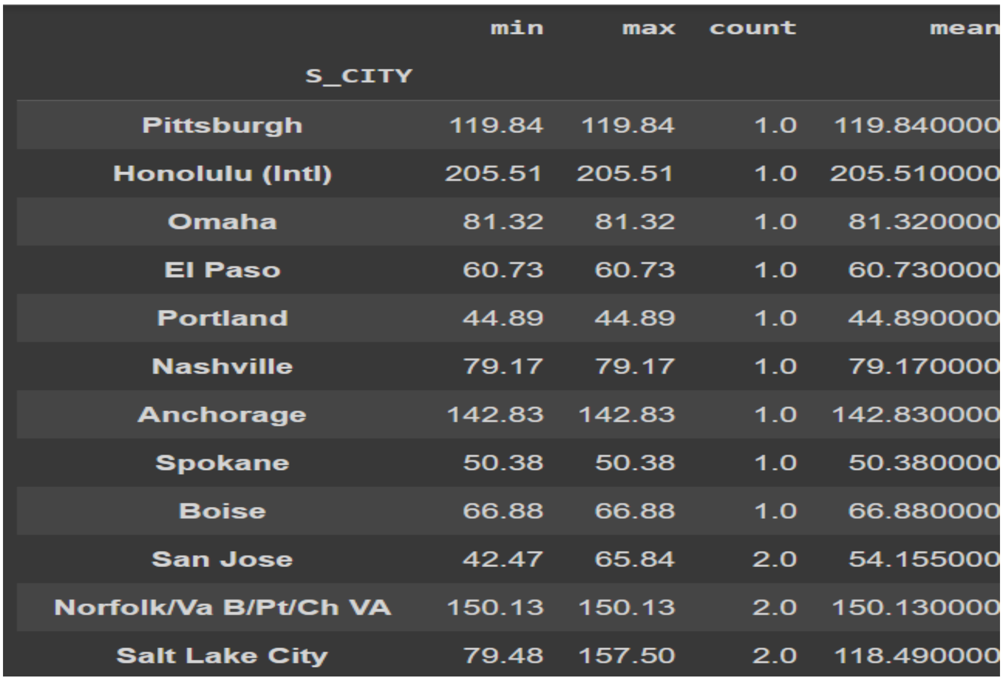
1. We took PAX i.e., the number of passengers on that route during the period of data collection on the **Histogram** for visualization. In this we have taken the number of passengers on x-axis against the frequency of data frame on the y-axis. It is graphically representing the organised number of passengers into the specified frequency. The histogram is negatively skewed as initially the data is longer or fatter at the tail towards left side of plot referring directional distribution of the data of passengers. Here skewness has provided the degree of the asymmetry of given number of passengers from the normal number of passengers in the Airfares dataset. Also we are well informed with the outliers of data because of left skewness of the histogram plot under given constraints.

2. We have used a **Scatter Plot** for this visualization titled Average Fare. We have taken average as univariate attribute under the distribution. In this relationship is suspected between the selected pairs of data. The graph has been plotted with the average fare represented on the X - axis and the index of the data frame that is airfares on the Y - axis. The dots on the graph represent the intersection of the values on both the axes. This scatter plot helps us understand the trend in the average fares between various flight routes in the data set depicted by their respective indices of dataset. The shift is real dynamic between data points of the average fare against the indices of the dataset, no such pattern and hotspots are identified in the plot. Similar analysis for distance as a parameter for **Scatter plot** is used for another visualization.

3. The next visualization is Population Trend under **Line Plot**. In this plot we have taken the starting city's population that is S_POP on the x-axis and ending city's population that is E_POP on the y-axis. Through this plot we are trying to identify the population's variation over the arrival and departure cities. After scrutinizing the plot, connecting the data points that is markers and observing the line we found that initial trends are very scattered having no symmetry, being just random. While observing the trends at the right-hand side of the graph a pattern and gradual slopes are observed in the plot making a relation between the starting and ending city of the passengers in the airfare dataset.

4. In the next visualization we have created a data frame and have visualized it. The **Global Stats** data frame is used for the statistical analysis of the starting city, using fare as the base for finding the minimum, maximum and mean of it and count with the starting city. In the visualization, Minimum is represented with blue line, Maximum is represented with orange, Count is represented with green line while Mean is represented with the red line. Picking a peculiarity, the Minimum,

Maximum and Mean for Pittsburgh, Honolulu, Omaha, El Paso etc is constant that results in overlapping lines. While considering the case of the Sacramento Austin, Atlanta, Chicago etc, clear variations of mathematical identifiers like min, max, count and mean.



	min	max	count	mean
S_CITY				
Pittsburgh	119.84	119.84	1.0	119.840000
Honolulu (Intl)	205.51	205.51	1.0	205.510000
Omaha	81.32	81.32	1.0	81.320000
El Paso	60.73	60.73	1.0	60.730000
Portland	44.89	44.89	1.0	44.890000
Nashville	79.17	79.17	1.0	79.170000
Anchorage	142.83	142.83	1.0	142.830000
Spokane	50.38	50.38	1.0	50.380000
Boise	66.88	66.88	1.0	66.880000
San Jose	42.47	65.84	2.0	54.155000
Norfolk/Va B/Pt/Ch VA	150.13	150.13	2.0	150.130000
Salt Lake City	79.48	157.50	2.0	118.490000

Conclusion:

On a broad vision various patterns are observed from the dataset of Airlines, but the trends presented through the tuples of Distance, Fare, S_Income, E_Income, Pax are significant. These are the few tuples that we improvised in our visualization and made the importance count. Histogram, Scatter plot, Line plot, and graphically representing the dataset etc are few visualizations that we have worked on to examine the data and identify the patterns.