**DATA VISUALIZATION**

**Name: Anushareddy Ramachandra Reddy**

**Datasets:**

1: Top12\_airlines\_reviews

2: Airline\_top12\_list

**Task1:**

**A screenshot of a computer

Description automatically generated**

**Data preparation:**

Extracted the separate route features by splitting columns as departure and destination using calculations.

Calculations:

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Clean Data:**

Identified the Null values in destination and cabin class variables, removed those values by placing these variables in filters and no null values found in departure variable.

**Visualize data:**

All flights are plotted on a map that clearly shows the geographical distribution, emphasizing hubs and busy routes. The number of flights between destinations is indicated by the size of the markers, which highlights important travel routes. Cabin class color-coded markings provide information on how various travel classes are distributed throughout routes.

**Identify Outliers:**

I have chosen the area lacking flight route from Dubai to East London, compared to other flight routes these have fewer connections than expected. By using the Descriptive Statistics method, I have identified outliers such as count, AGG(flights by Sum), Average, Minimum, Maximum, Median were assigned a value of 1.

**Task2:**

While preparing the data I have included fields in the rows as overall rating, measure values and covid19 mentions, refund mentions are selected columns. Added 2 new binary columns and named them as Covid 19 Mentions and Refund mentions which store the value of either ‘yes’ or ‘No’

**Calculations:**

Calculation to scan review text for covid related key words such as pandemic, covid, coronavirus, virus was created and titled as covid 19 mentions and assigned “Yes” value for columns containing any identified key words and in other cases the mark shall be “No”.

**A screenshot of a computer code

Description automatically generated**

Similarly, we created fields mentioning refunds, reimburse and cancellations.

**A screenshot of a computer code

Description automatically generated**

**Visualization:**

**A screenshot of a computer

Description automatically generated**

To visualize the prepared data, I have used a combination stacked bars and box and Whisker plots, to compare overall rating and aspect ratings. I opted for a range of colors representing each of the values as follows avg AR checking and boarding: Green, Cleanliness: Blue, Legroom: Pink, Value for Money: Yellow.

**Task3:**

**Data preparation:**

**Date Segmentation:** Ensured dataset includes a 'Review Date' field. Created a filed to categorizes reviews into 'Pre-Pandemic' and 'During Pandemic' and titled it as Covid Time Difference.

**A screenshot of a computer

Description automatically generated**

Created 8 calculation aspect rating files such as:

Aspects.Customer service:

**A screenshot of a computer

Description automatically generated**

Aspects.check-in and boarding:

**A screenshot of a computer

Description automatically generated**

Aspects.Cleanliness

**A screenshot of a computer program

Description automatically generated**

Aspects.Food and Beverage:

**A screenshot of a computer

Description automatically generated**

Aspects.In-flight Entertainment:

**A screenshot of a computer

Description automatically generated**

Aspects.Legroom:

**A screenshot of a computer

Description automatically generated**

Aspects.Seat Comfort:

**A screenshot of a computer

Description automatically generated**

Aspects.Value for money**A screenshot of a computer

Description automatically generated**

**Visulaization:**

**A screenshot of a computer

Description automatically generated**

Created a visualization to identify changes in rating before and during the pandemic,

X axis contains a quarter of the review date and Y axis contains 8 aspect ratings which I have created.

**Task 4**

**Variable Selection:**

Plotted the Overall Rating, which shows the rating scores assigned to airlines, on the x-axis. The count of the top 12 airline reviews is plotted on the y-axis to show how many reviews certain airlines have received.

**Data Preparation:** To clean the data set, I filtered variables such as Cabin class and Destination.

**Correlation analysis:** calculated the Pearson correlation coefficient using the calculation below.

**Trend Line description:**

|  |  |
| --- | --- |
| P-value: | < 0.0001 |
| Equation: | Count of Top12\_airlines\_reviews.txt = 0.233582\*Overall Rating + 84.2694 |

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficients | | | |
| Term | Value | StdErr | t-value | p-value |
| Overall Rating | 0.233582 | 0.0034354 | 67.9921 | < 0.0001 |
| intercept | 84.2694 | 310.872 | 0.271074 | 0.791847 |

**P-Value**: "< 0.0001" is the reported p-value for the "Overall Rating" variable. This suggests that the p-value is essentially less than 0.0001.

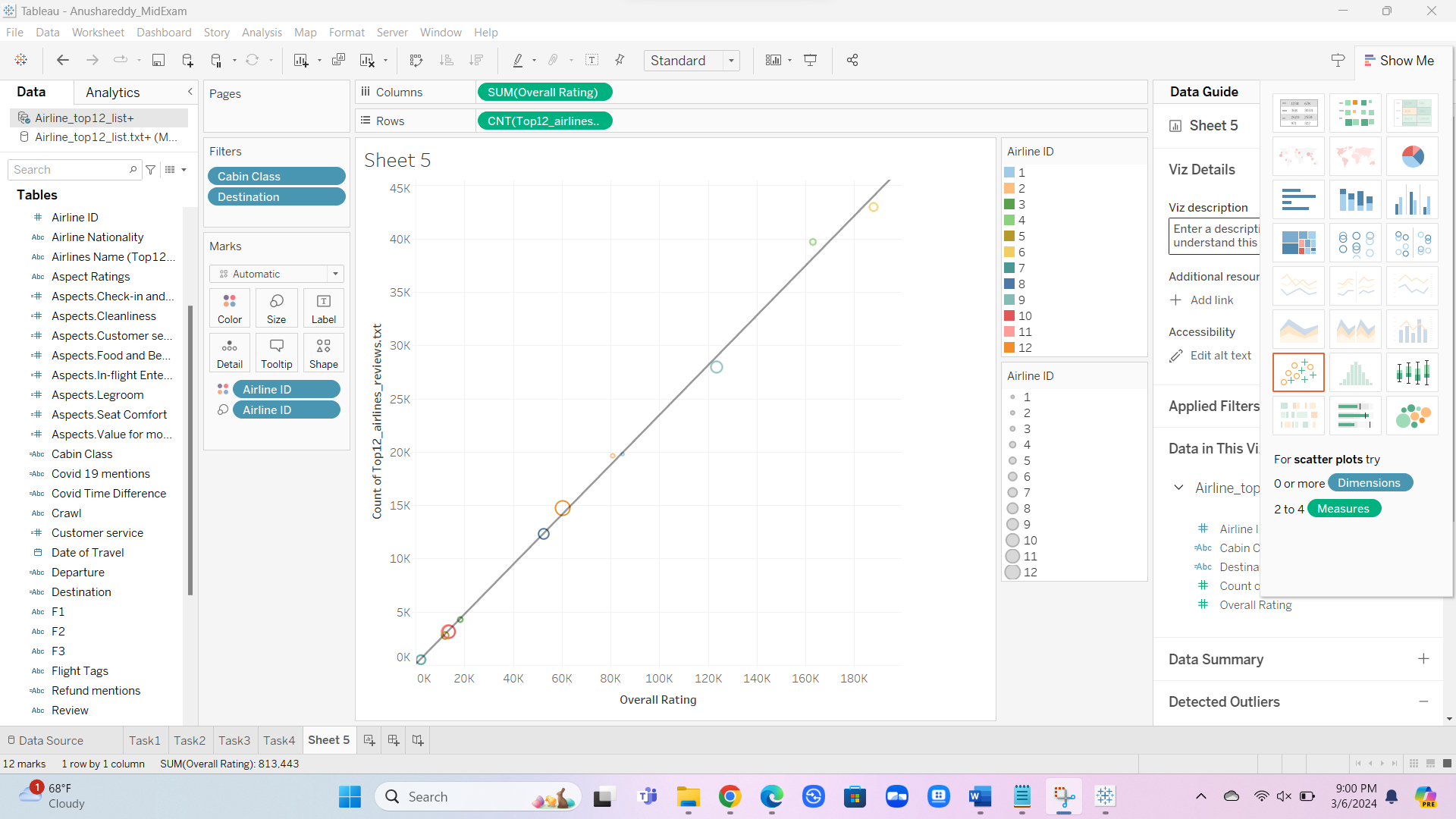
**Coefficient:** The "Overall Rating" variable's coefficient is 0.233582.

**Standard error:** The coefficient of "Overall Rating" has a standard error of 0.0034354. A smaller standard error denotes a more accurate coefficient estimate.

**T value:** 67.9921 is computed using the coefficient ratio.

Count of Top12\_airlines\_reviews.txt" and "Overall Rating" have a highly significant positive relationship, according to the regression analysis. Keeping all other factors equal, it is projected that the "Count of Top12\_airlines\_reviews.txt" will increase by roughly 0.233582 units as the "Overall Rating" rises.

**Visualization:**

****

The above visualization represents the overall Rating and count of top 12 airline reviews, used scatter plots to visualize direction and strength of the relationship between 2 quantitative variables. We have plotted the Overall Rating, which shows the rating scores assigned to airlines, on the x-axis. The count of the top 12 airline reviews is plotted on the y-axis to show how many reviews certain airlines have received. Used color encoding to give each data point a distinct color that matched the Airlines ID to add depth to the image. This makes it possible to distinguish data points according to the specific airlines they represent, which improves our comprehension of each airline's performance within the dataset.

**Task5:**

**A screenshot of a computer

Description automatically generated**

**Identify areas for exploration:** My observations during the pandemic have shown a universal reduction in the overall ratings for all airlines. If given the chance, I would use a variety of visualization approaches to develop plans of action meant to raise these ratings.

**Propose new questions:** What is the overall all satisfaction rating for each aspect rating before pandemic and after pandemic and Which airline is affected.

**Dashboard design:** Looking over the dashboard, I've noticed a general pattern of lower passenger feedback for all airlines throughout the epidemic, with airline 7 having the lowest ratings. To address this issue and improve overall customer happiness, I support the implementation of additional aspect ratings and services that are customized to match changing customer needs.