

**2570505 - EMRE KALE**

**2561587 - ENES YANAÇ**

**2657344 - FUAD İBRAHİMLİ**

**2643609 - FURKAN ÇELİK**

**2561207 - SADIK DİNLEMEZ**

**TABLE OF CONTENTS**

[1 ABSTRACT 3](#_Toc156055178)

[2 INTRODUCTION 3](#_Toc156055179)

[2.1 DESCRIPTIVE STATISTICS 4](#_Toc156055180)

[3 DATA TIDYING AND STEPS OF IT 5](#_Toc156055181)

[4 EXPLORATORY DATA ANALYSIS QUESTIONS 6](#_Toc156055182)

[4.1 Is there any relationship between “Price” and “The Number of Passengers”? 6](#_Toc156055183)

[4.2 How does the price of tickets change in 4 years period? 7](#_Toc156055184)

[4.3 How does the price vary based on cabin types? 7](#_Toc156055185)

[4.4 How do delay minutes change between flight companies? 8](#_Toc156055186)

[4.5 According to the data, what is the most preferable flight routes? 9](#_Toc156055187)

[5 SUMMARY 9](#_Toc156055188)

[6 REFERENCES 10](#_Toc156055189)

**TABLE OF FIGURES**

[Figure 1 **The Number of Passenger vs Price** 6](file:////Users/furkancelik/Desktop/sonnnnnnnnnn.docx#_Toc156056957)

[Figure 2 **Distribution of Prices between Cabin Types** 6](file:////Users/furkancelik/Desktop/sonnnnnnnnnn.docx#_Toc156056958)

[Figure 3 **Distribution of Prices between Cabin Types** 7](file:////Users/furkancelik/Desktop/sonnnnnnnnnn.docx#_Toc156056959)

[Figure 5 **Average Delay Minutes by Airline** 8](file:////Users/furkancelik/Desktop/sonnnnnnnnnn.docx#_Toc156056960)

[Figure 6 **Airport Relationship Heatmap** 9](file:////Users/furkancelik/Desktop/sonnnnnnnnnn.docx#_Toc156056961)

# **ABSTRACT**

Commercial aviation is the most growing industry of the world and due to this reason companies are always creating data of the airline companies for improving the quality of trips. It is because airline companies want to attract passengers to their companies. In this analysis, after cleaning the data, with the sales information of airline companies, the relationships between the categories of the cabin type, price, day of the week and the number of the passengers were examined by using descriptive statistics and exploratory data analysis techniques. From the analysis, the following information were found:

* First Class is the most preferred seat place plan.
* Standard cabin type is the least preferred cabin type by passengers.
* There is a positive linear relationship between distance (km) and baggage fee.
* Most expensive tickets are sold by Emirates airline.
* Duration hours have little effect on delay minutes.
* The least avia trips had been in 2021.

# **INTRODUCTION**

This data set contains information details such as duration time, ticket price, cabin type and so on about avia-trips. There are 144 flight information details. The dataset has fourteen variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Type | Level of Measurment | Data Examples |
| FLIGHT ID | Categorical | Nominal | 1,2, 3,…10 |
| DATE | Categorical | Nominal | 12.01.2024 |
| AIRLINE | Categorical | Nominal | Delta Airlines, Air France |
| SOURCE AIRPORT | Categorical | Nominal | HND, LHR, CDG |
| DESTINATION AIRPORT | Categorical | Nominal | LAX, CDG, HND |
| DURATION HOURS | Numerical | Ratio | 7.5, 6, 12 |
| SEAT PLACE | Categorical | Nominal | Business, Economy |
| PRICE | Numerical | Ratio | 1200, 800, 600 |
| DAY OF WEEK | Categorical | Nominal | Monday, Tuesday |
| DISTANCE KM | Numerical | Ordinal | 8785, 5834 |
| THE NUMBER OF PASSENGERS | Numerical | Ratio | 150, 120, 180 |
| DELAY MINUTES | Numerical | Ratio | 15, 10, 30 |
| BAGGAGE FEE | Numerical | Ratio | 25, 20, 35 |
| CABIN TYPE | Categorical | Nominal | Standard, Economy |

## **DESCRIPTIVE STATISTICS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Mean | Minimum | Maximum | Q3- Q1 | IQR | Outlier |
| | **DURATION HOURS** | | --- | | 7.60 | 2 | 13.5 | 9 - 5.5 | 3.5 | NONE |
| PRICE | 1300.72 | 400 | 2800 | 2000-700 | 1300 | NONE |
| | **DISTANCE KM** | | --- | | 5032.93 | 341 | 9582 | 8785- 2232 | 6553 | NONE |
| | **THE NUMBER OF PASSENGERS** | | --- | | 140.07 | 110 | 180 | 160-120 | 40 | NONE |
| | **DELAY MINUTES** | | --- | | 18.138686 | 5 | 30 | 25 - 15 | 10 | NONE |
| BAGGAGE FEE | 25.912409 | 15 | 35 | 30 - 20 | 10 | NONE |

We have shown the descriptive statistics. We calculated the mean, minimum, maximum, third quartile (Q3), first quartile (Q1), interquartile range (IQR), and outlier for the numeric values in our dataset. The mean duration hours is 7.6 hours, the minimum value of it is 2 hours, and the maximum is 13.5 hours. Q3 and Q1 for the duration hours are 9 and 5.5 hours, respectively. IQR, which is 3.5 hours for duration hours, was founded by substructing Q1 from Q3. Outlier for hours could not be found. The mean price is $1300.72, the minimum value is $400, and the maximum is $2800. Q3 and Q1 for the price are $2000 and $700, respectively. IQR, which is 1300 for the price, was founded by substructing Q1 from Q3. An outlier for the price could not be found. The mean distance of km is 5032.93, the minimum value of it is 341 km, and the maximum is 9582 km. Q3 and Q1 for the distance km are 8785 km and 2232 km, respectively. IQR, which is 6553 for the distance km, was founded by substructing Q1 from Q3. The outlier for the distance km could not be found. The mean number of passengers is 140.07, the minimum value of it is 110, and the maximum is 180. Q3 and Q1 for the number of passengers is 160 and 120, respectively. IQR, which is 40 for the number of passengers, was founded by substructing Q1 from Q3. The outlier for the number of passengers could not be found. The mean of the delay minutes is 18.138686 minutes, the minimum value is 5 minutes, and the maximum one is 30 minutes. Q3 and Q1 for the delay minutes are 25 and 15, respectively. IQR, which is 10 for the delay minutes, was founded by substructing Q1 from Q3. The outlier for the delay minutes could not be found. The mean baggage fee is $25.912409, the minimum value of it is $15, and the maximum is $30. Q3 and Q1 for the baggage fee are $30 and $20, respectively. IQR, which is $10 for the baggage fee, was founded by substructing Q1 from Q3. The outlier for the baggage fee could not be found.

# **DATA TIDYING AND STEPS OF IT**

1. Pandas, NumPy, Seaborn, Plotly, and Matplotlib libraries were imported, and the data was read in the notebook.
2. There are 14 variables presented in the given dataset. 6 of them have object data type, 6 of them have integer data type, 1 of them datetime, and 1 one of them float data type.
3. Objects are “Airline, Source Airport, Destination Airport, Seat Place, Day of Week, Cabin Type”, integers are “Price, Distance Km, The Number of Passengers, Delay Minutes, Baggage Fee, Flight Id”, float is “Duration Hours” ,datetime is “Date”
4. Duplicates are checked, we have one duplicated observation. We removed the duplicated row.
5. As a next step, NA lines with at least one element missing were dropped.
6. The head and tail of the data frame were examined. For any separation argument problem (“;” or “,”) of the data were checked. Incorrect special characters ' " ' and ' ## ' between data names have been removed.
7. Column headers, multiple variables, multiple types of observational units were checked. Column names were fixed. Unnecessary columns were checked, but we couldn’t find it.
8. False and inappropriate values were replaced with correct ones. Some values have been replaced with their abbreviations to be compatible with all values.
9. The absolute value of the values was taken in case there was any negative value in the numeric data.
10. The spaces in the DISTANCE column were filled with the mean of distance.
11. The spaces in the DURATION HOURS column were filled with the mode of duration hours.
12. The spaces in the PRICE column were filled with the mean of the price.
13. The spaces in the NUMBER OF PASSENGERS column were filled with the mean of the number of passengers.
14. The spaces in the CABIN TYPE column were filled with the mode of cabin types.
15. DATE column was converted to datetime type and only the date part was taken from the DATE column.
16. The names of the columns were renamed with their correct and appropriate form.
17. All strings were made in the same format.
18. Outlier analysis was performed, and no outlier was found.
19. The data was reviewed one final time to ensure it was clean. Clean data was transferred to a new excel file.

# **EXPLORATORY DATA ANALYSIS QUESTIONS**

## **Is there any relationship between “Price” and “The Number of Passengers”?**

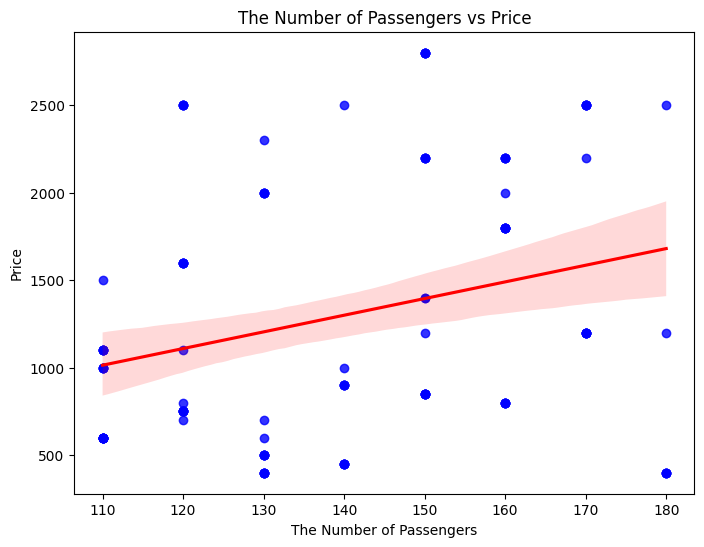
Scatter plots are used to detect relationships between two variables. However, a scatter plot is not enough to learn the relationship between two variables. Using trend lines in scatter plot shows the relationship which is called correlation coefficient (r) in statistics. Correlation coefficient enables us to understand the relation between two numeric variables which are Price and The Number of Passengers in this graph. The trend line shows that the price of the tickets affects the number of passengers. The correlation coefficient is approximately equal to 0.4, which means that there is a positive correlation between these two variables. According to this graph, when the price of the ticket is higher, it is clear that airline companies sell more tickets. Briefly, the price of the tickets affects the number of passengers in a positive way.

Figure 1 **The Number of Passenger vs Price**

## **How does the price of tickets change in 4 years period?**

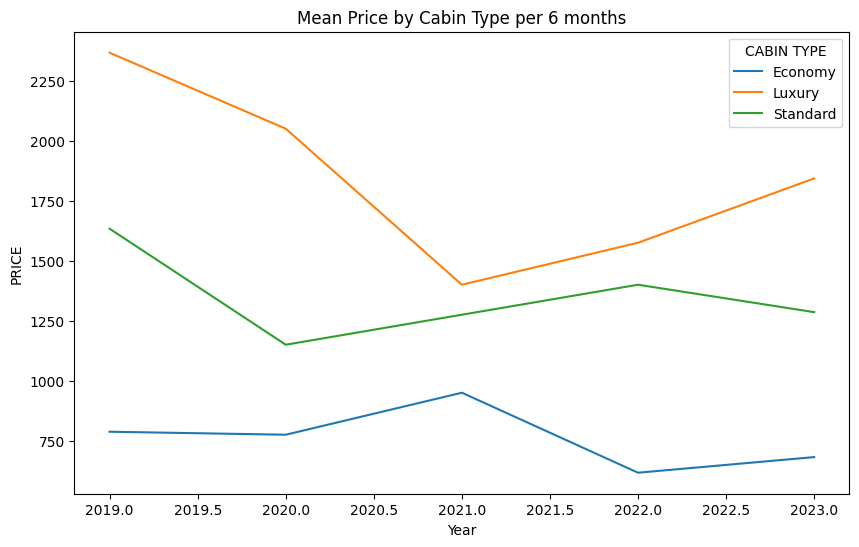
Showing the price change of products is crucial for understanding the inflation rate in a default country or industry. Line charts are used to visualize price changes in a given interval date. Tickets have different prices depending on cabin types such as standard, economy, luxury. There are three lines because each cabin type has different prices. Economic tickets were 750$ in the first period of 2019, and after 4-years, airline companies sold economy tickets for approximately 675$. It means that the price of economy tickets decreased in a 4-year period. Standard cabin-type tickets were sold for 1650$ in 2019, and after four years, their price was reduced to 1300$. Luxury cabin-type tickets were sold for 2300$ in 2019, and in 4 years, their price decreased to 1850$. In conclusion, it is clear that the prices of tickets fell in a 4-year period. According to the graph, the prices of tickets will likely increase in the future.

Figure 2 **Distribution of Prices between Cabin Types**

## **How does the price vary based on cabin types?**

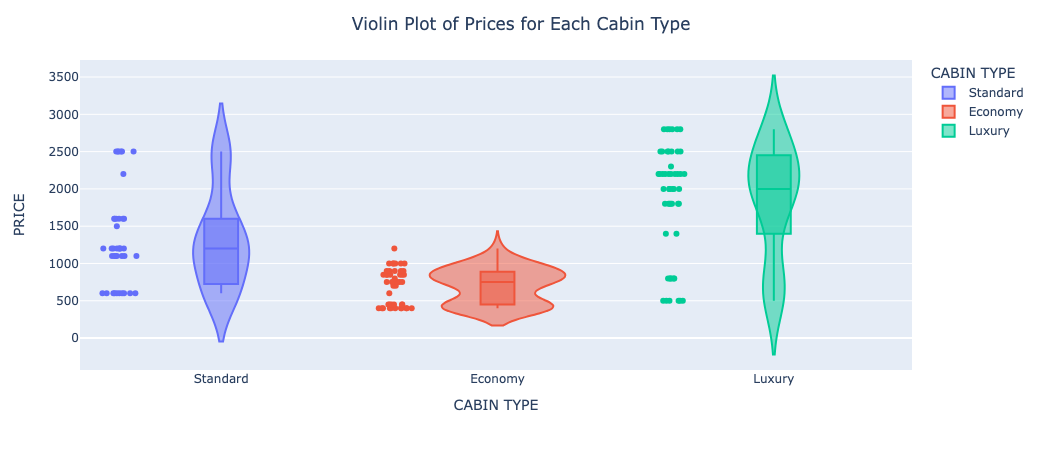
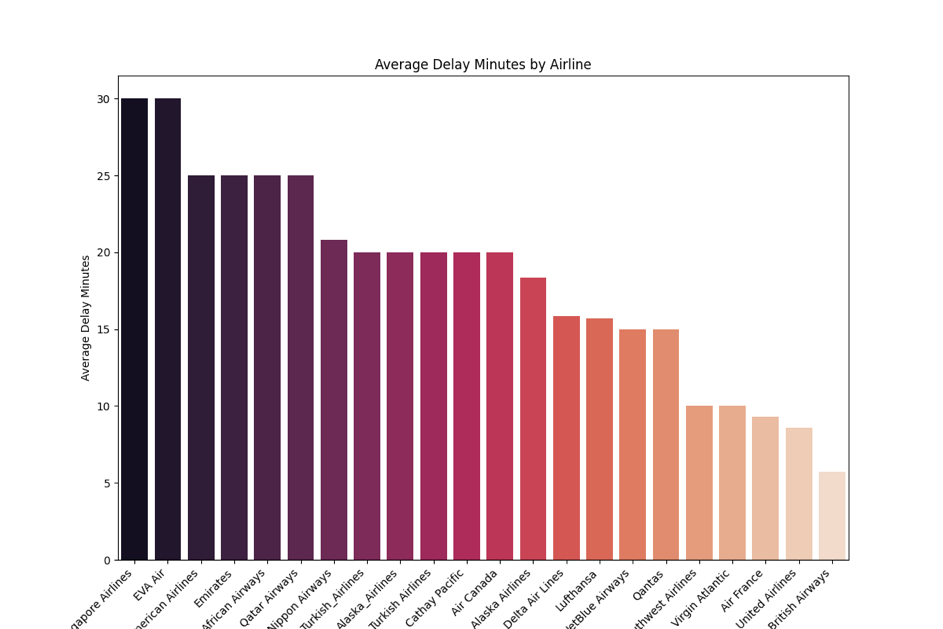
Violin plot chart is an effective type of statistical chart which is used to explain the distribution of the data set visually. This chart helps us understand the patterns in the data set by simultaneously showing the central tendency, spread and possible intensities of the data distribution. The distribution of the price by cabin type in this data set that has been examined. According to this graph, it can be seen that the price of the standard cabin type is concentrated in the range of 1250 $, for economy cabin type is around 800 $ and the price of the luxury type is concentrated between 2000 $ and 2500 $. Observing the graph, it becomes evident that the luxury cabin type experiences the most significant variation in prices, while the economy cabin type showcases the least fluctuation. In summary, this analysis sheds light on the different pricing dynamics between cabin types, highlighting the diversity in price variations among them.

Figure 3 **Distribution of Prices between Cabin Types**

## **How do delay minutes change between flight companies?**



The pareto chart provides a comprehensive overview of the average minutes of delay across multiple airlines, providing insights into their on-time performance. Each bar on the chart represents a different airline; The x-axis shows the airline names, and the y-axis shows the average delay time in minutes. Specifically, airlines higher on the chart tend to experience longer average delays, while airlines lower on the chart exhibit shorter delays. This visual exploration serves as a valuable tool for stakeholders to quickly compare and evaluate the relative punctuality of different airlines within the data set. The results contribute to a detailed understanding of airline performance, facilitating informed decision-making and contributing to potential areas for improvement in managing flight delays.

Figure 5 **Average Delay Minutes by Airline**

## **According to the data, what is the most preferable flight routes?**

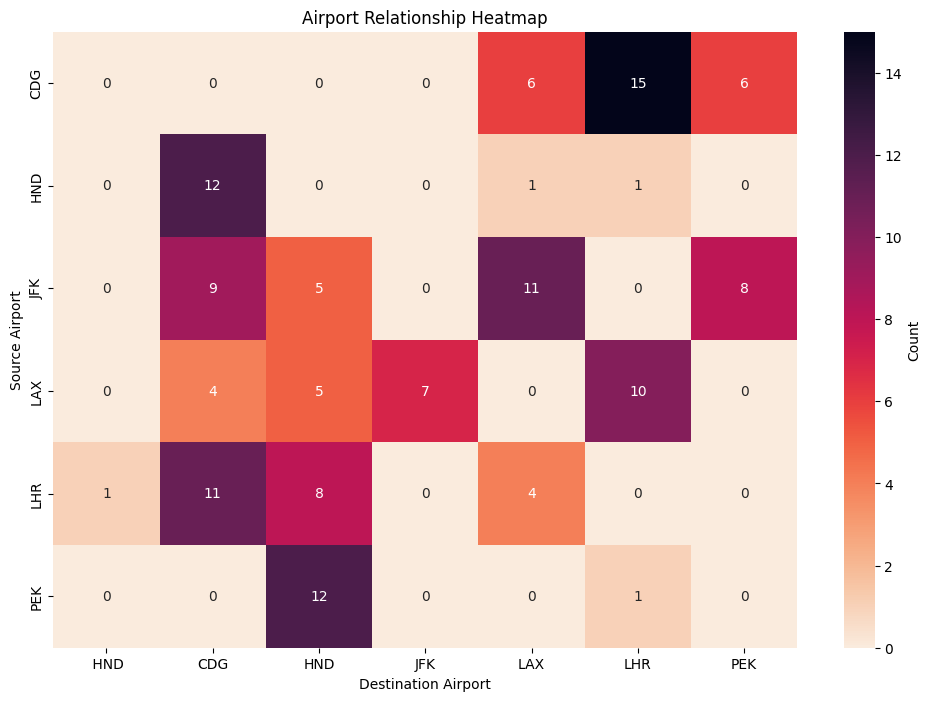
One of the most effective charts for comparing the number of two categorical data, which are destination airport and source airport, is the waffle chart. As can be seen from this graph, the colors are at a certain intensity in each frame, and it can be said that the common number between the data increases as you go from light to dark. There are many squares with a value of 0 which means those rotations aren’t preferred, on the other hand, we can say that the highest value in the graph is between CDG and LHR and its value is 15. According to this graph flight companies can take a measurement for cancelling those rotations with 0 value. In addition, if flight companies increase their routes which is highly preferred, they will likely decrease their busy routes. By doing this, they can increase their income.

Figure 6 **Airport Relationship Heatmap**

# **SUMMARY**

In conclusion, first of all we cleaned our data set by using NumPy and pandas. Then, in order to understand our data set, we created five research questions and plotted relevant graphs by using seaborn and matplotlib. As a result, we investigated how prices change over 4 years. From this, we found that prices decreased in these 4 years. Moreover, it is observed that the luxury cabin type is the most expensive among these three cabin types, while the economy cabin type is the least costly. To investigate this, we compared the cabin type and price of airline companies and found that there is a strong relationship. Later, when we observed the delay minutes between flight companies variables, we observed that passengers of Singapore airlines experienced delay minutes the more than passengers of other companies. British airline company gives best performance to passengers about delay minutes. Then, we investigated the relationship between price of tickets and number of passengers and found that there is a linear relationship.

# REFERENCES

***GitHub Links***

* Our Phyton codes that includes data cleaning steps, data analysis parts and finally data visualization processes is here:
  + [**https://github.com/celikfrkn/finalproject/blob/main/dataprocessingandvisualization.ipynb**](https://github.com/celikfrkn/finalproject/blob/main/dataprocessingandvisualization.ipynb)
* Cleaned data frame is here**:**
  + [**https://github.com/celikfrkn/finalproject/blob/main/cleaned\_dataframe.xlsx**](https://github.com/celikfrkn/finalproject/blob/main/cleaned_dataframe.xlsx)

***Sources***

* DataCamp. (n.d.). *Intermediate Data Visualization with Seaborn*.

<https://www.datacamp.com/courses/intermediate-data-visualization-with-seaborn>

* *Get started with GitHub documentation - GitHub Docs*. (n.d.). GitHub Docs.

<https://docs.github.com/en/get-started>

* Yozgatlıgil, C. (2023) *Numpy.* [PowerPoint slides].ODTÜClass.
* Yozgatlıgil, C. (2023) *Pandas.* [PowerPoint slides].ODTÜClass.
* Yozgatlıgil, C. (2023) *Data Cleaning With Pandas And Numpy.* [PowerPoint slides].ODTÜClass.
* Yozgatlıgil, C. (2023) *Data Visualization with Python.* [PowerPoint slides].ODTÜClass.
* Yozgatlıgil, C. (2023) Data Visualization Example and Bad Visualizations. [PowerPoint slides].ODTÜClass.