

**Ain Shams University**

**Faculty of Computer and Information Science**

**Scientific Computing department**

**Ain shams university**

**Faculty of computer and information science**

**Bioinformatics department**

**Project Title**

**Airline Ticket Price Prediction**

**By**

**Nour Mohammed Hussein Kamaly**

**Nourhan Abdel-Karim Khalaf Abdel-Hafez**

**Ahmed Mohammed Samy Mahmoud**

**Mohammed Nour-Elden Abbas Ismael**

**Abdul-Rahman Sayed Ali Mohammed**

**Under the supervision of**

**Dr. Dina Khattab**

**Scientific Computing Department,**

**Faculty of computer and Information Science**

**Ain Shams University**

We can find data here [Airline-Price](../Data/airline-price-prediction.csv)

We can find the description of data here [Data-Description](../Data/Data-Description.txt)

**Data Analysis phase**

**We can suggest some questions to analysis this data such that:**

Q1-What is the airline that has the most expensive ticket (business and economy)?

Table, Excel

Description automatically generatedAns:

the most expensive ticket for business class is: Vistara Airline

the most expensive ticket for economy class is: India Airline

**Table 1**

Q2-What is the airline that has the cheapest ticket (business and economy)?

Table

Description automatically generatedAns:

the most expensive ticket for business class is: Vistara Airline

the most expensive ticket for economy class is:

{Asia Airline, Go First Airline, Indigo Airline}

**Table 2**

Q3-What is the most expensive 'business type' ticket in each airline and what is its duration and date?

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
| Airline | The Most Expensive price | Duration | Date |
| Vistara | 99677 | 36 | February |
| Air India | 89257 | 35 | February |

Q4-What is the cheapest 'business type' ticket in each airline and what is its duration and date?

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
| Airline | Cheapest price | Duration | Date |
| Vistara | 100 | 36 | February |
| Air India | 12000 | 35 | February |
| Air India | 12000 | 36 | March |

Q5-What is the most expensive 'economy type' ticket in each airline and what is its duration and date?

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
| Airline | The Most Expensive price | Duration | Date |
| Air Asia | 31917 | 19 | February |
| Go First | 31773 | 20 | February |
| Indigo | 31952 | 15 | February |
| SpiceJet | 34158 | 28 | February |
| Air India | 42349 | 45 | February |
| Vistara | 37646 | 41 | February |
| Star Air | 9682 | 2 | March |
| TruJet | 4844 | 3 | February |

Q6-What is the cheapest 'economy type' ticket in each airline and what is its duration and date?

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
| Airline | Cheapest price | Duration | Date |
| Air Asia | 1105 | 19 | February |
| Air Asia | 1105 | 18 | March |
| Go First | 1105 | 20 | February |
| Go First | 1105 | 19 | March |
| Indigo | 1105 | 14 | March |
| SpiceJet | 1106 | 28 | March |
| Air India | 1526 | 42 | March |
| Vistara | 1714 | 41 | February |
| Vistara | 1714 | 36 | March |
| Star Air | 2000 | 2 | March |
| TruJet | 3124 | 3 | February |
| TruJet | 3124 | 3 | March |

Chart, pie chart

Description automatically generatedTable

Description automatically generated

**Figure 1**

**Table 3**

Q7-How much did the longest flight cost in business and economy types?

Table

Description automatically generatedAns:

the longest flight in business type cost: 52446.90

the longest flight in economy type cost: 6572.47

**Table 4**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedQ8-What time of the day has the most expensive (business and economy) ticket in each airline?

**Figure 2**

Ans:

the time of the day that have the most expensive business ticket is: 12:00

the time of the day that have the most expensive economy ticket is: 20:00

Table

Description automatically generatedQ9-What is the day-average ticket price (business and economy) in each day of the week?

Ans:

**Table 5**

Text

Description automatically generated with medium confidenceQ10-What is the month-average ticket price (business and economy) in each month of the year?

Ans:

**Table 6**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedQ11-What is the most departure and arrival time for planes during the day?

**Figure 3**

Ans:

the most departure time at 20:00

the most arrival time at 10:00

Q12-Who is the most frequent source which takes off planes from it?

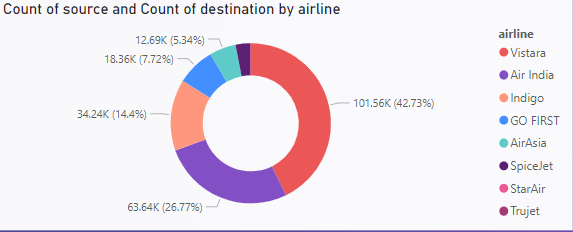
Ans:

the most frequent source which takes off planes from is: Vistara Airline

Q13-Who is the most frequent destination for planes?

Ans:

the most frequent destination which takes off planes from is: Vistara Airline



**Figure 4**

Q14-What is the most popular day and month for flight?

Ans:

the most popular day is Monday

the most popular Month is March

Chart, line chart

Description automatically generated

**Figure 5**

Q15-How much did the longest flight take?

Ans:

The longest Flight takes 50 hours

Graphical user interface, application

Description automatically generated

**Figure 6**

Q16-What is the most airline used?

Chart, funnel chart

Description automatically generatedAns:

The most airline used Vistara Airline

**Figure 7**

Q17-What is the most number of stops during the flight?

Chart, pie chart

Description automatically generatedAns:

The most number of stops during the flight is 2 hours

Q18-What is the least frequent source for airplanes? (Look Figure 4)

Ans:

The Least frequent source for airplanes is Trujet Airplane

Q19-What is the least frequent destination? (Look Figure 4)

Ans:

The Least frequent Destination for airplanes is Trujet Airplane

Q20-What is the busiest duration of the day for flights? (Look Figure 6)

Ans:

the busiest duration of the day for flights is 2 Hours

Q21-What is the Freest duration of the day for flights? (Look Figure 6)

Ans:

the Freest duration of the day for flights are: 42,46,48,45,47,50 Hours

**Preprocessing phase**

1. Preprocessing techniques
2. Features used / discarded

* Hypothesis testing using p value was used in feature selection, our null hypothesis is that the model is learning from the feature, so if the p value exceeded 0.05 then my null hypothesis failed, and the model is not making use of the feature
* Features used are (p value < 0.05): type, flight month, number of stops, distance between 2 countries, airline, source, destination, Saturday, Sunday, Thursday, Tuesday, departure time, and arrival time.
* Features discarded were: flight day, number of hours taken, one stop in, Friday, Monday, Wednesday.
* The above features were discarded based on hypothesis testing, but these features were discarded as they were in the wrong format (they were fixed and given a new name) : date, time taken, stop, route, price.
* Ch code was dropped as there was a 100% dependency between it and the airline as the ch code is a code for the airline, so using both would cause redundancy
* Num code was dropped as there were so many value counts each had a low frequency between the observations.

1. Sizes of training and testing sets

* As this is a time series data, choosing a random train test would lead to data leakage as we can’t let the model train on new data and test on old or shuffled data so :

The train set consists of the first 80% of the sorted dataset (sorted by date)

The test set consists of the last 20% of the sorted dataset (newer dates)

1. Techniques used to improve results