Research Proposal

Lokesh Bokkisam

Pavani Samala

Shumel Siraj

The research topic that our group chose to explore is data collected from people traveling domestically in India via different airlines. The variables that are included in the dataset focus on airline names, airfare price, departure and arrival times, travel duration, number of stops during the trip, and the class of travel. This dataset was created by merging economy and business datasets, aggregating about 30,000 observations. Our SMART question asks: is there a relationship between the price of airfare with the convenience provided by the airlines such as short travel time and fewer stops, or are there other factors influencing airfare price. The modeling methods that are currently being proposed are linear regression and logistic regression data visuals such as histograms, boxplots, heatmaps, pair plots, and stripplots. The information extracted from the data can be used to advise airlines on how to design flight routes to make the most profit.

The link to the source of the data is provided below:
https://www.kaggle.com/code/borandabak/flight-price-data-analysis/data

The link to the github repository is provided below:
https://github.com/psamala99/6103-project.git

Our SMART question asks: do the same factors affect business class airfare and economy class airfare (i.e is the luxury of business class the only reason the prices are higher)?

Economy	Business
Air India	Air India
Vistra	Vistra
SpiceJet	
AirAsia	

GO_FIRST	
Indigo	

Make some scatter plots but its hard to spot any trends so we need to make linear models
Make linear models for economy and business with duration, stops, and days left (not great)

Cross validation score
Regression tree, knn

Graphs needed:
-histogram: days left (numerical value) and price
-histogram: days left (categorical value) and price
Objectives

- Can the price of airline tickets be predicted by the number of stops, duration of flight, and day left until take off, or is it better predicted by adding categorical variables such as departure and arrival location?

Introduction/Data Used

- Where we got data
- How many variables-what each one is
- describe() to min, max, other stats

Data Preparation

- Dropped na values
- Changed categorical to numerical

Data Analysis

- Normality test
- Because the dataset did not included luxury to distinguish between the facilities of both class we decided to divide between class
- grid for source and destination cities, econ and buz
- histogram: days left (numerical value) and price
- Piechart of days left (categorical)
- boxplot/stripplot

Model Building

- stops, duration, days-left
    - Linear regression for

    - Tree regression
    - knn
- stops, source_city, desination_city
    - Linear regression for
    - Tree regression
    - knn

Conclusion

Rows: 300153, Columns: 11

1. **Airline:** this column will have all the types of airlines like Indigo, Jet Airways, Air India, and many more.

2. **Flight:** The unique identification number of the flight on which the passenger will travel.

3. **Source City:** This column holds the name of the place from where the passenger's journey will start.

4. **Destination City:** This column holds the name of the place to which passengers want to travel.

5. **Arrival Time:** The arrival time is the time at which the passenger will arrive at his or her destination.

6. **Class:** In our data set we only have two classes: "economy class" and "business class."

7. **Departure Time:** The departure time is the time at which the passenger will board the plane from the origin city.

8. **Duration:** The total amount of time it takes a flight to travel from one city to another.

9. **Days Left:** the number of days remaining between the purchase of a flight ticket and the flight's departure date.

10. **Stops:** This tells us how many times a flight will stop during its entire journey.

11. **Price:** The price of the flight for a complete journey, including all the expenses before onboarding.