PROJECT REPORT

Submitted By

Aswini Arun Patil

NAME OF THE PROJECT

Flight Price Prediction

ACKNOWLEDGMENT:

- Primarily I would like to thank God to being able to complete this project with success. Then I would like to express my special thanks of gratitude to my SME,

- And I am thankful I am part of flib rob technology of employee, who given me the golden opportunity to do this wonderful project on the given topic which is also help me in doing a lot of research and I came to know about so many new things, I am really thankful to flip robo.

DATE:11/07/2022

ASWINI A. PATIL
Data Science course
Institute: Data trained education
Internship: Flip Robo technology
@Bangalore

INTRODUCTION:

- ## Business Problem Framing

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics which they call "revenue management" or "yield management". The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on –

1. Time of purchase patterns (making sure last-minute purchases are expensive)

2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

• Conceptual Background of the Domain Problem:

Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements.

According to a report, India's civil aviation industry is on a highgrowth trajectory. India aims to become the third-largest aviation market by 2020 and the largest by 2030. Indian domestic air traffic is expected to cross 100 million passengers by FY2017, compared to 81 million passengers in 2015, as per Centre for Asia Pacific Aviation (CAPA). According to Google Trends, the search term - "Cheap Air Tickets" is most searched in India. Moreover, as the middle-class of

India is exposed to air travel, consumers hunting for cheap prices increases.

• Review of Literature:

The airline implements dynamic pricing for the flight ticket. According to the survey, flight ticket prices change during the morning and evening time of the day. Also, it changes with the holidays or festival season. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not.

It is very difficult for the customer to purchase a flight ticket at the minimum price. For this several techniques are used to obtain the day at which the price of air ticket will be minimum. Most of these techniques are using sophisticated artificial intelligence(AI) research, also known as Machine Learning.

Utilizing AI models to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 85.3% precision. I contemplated the exhibition of Ridge Regression ,

Elastic Net, SGD Regression, Decision Tree Regression, K-Neighbors Regression, Random Forest Regression and Gradient Boosting Regression models in anticipating air ticket costs.

The data was collected from major travel journey booking website yatra.com. Additional data were also collected and are used to check the comparisons of the performances of the final model.

• Motivation for the Problem Undertaken:

Objective – To Scrape the data from website (I have scrapped the data from yatra.com) and then build a machine learning model to predict the price of the flights.

Analytical Problem Framing:

• Data Sources and their formats

Data Collection is one of the most important aspects of this project. There are various sources of airfare data on the Web, which I could use to train our models. A multitude of consumer travel sites supply fare information for multiple routes, times, and airlines. I tried various sources ranging from many APIs to scraping consumer travel websites like yatra.com.

	Date	Airline	Source	Destination	Dep_time	Arr_time	Durtion	Route	stop	price
0	25-Oct	GO FIRST	Bangalore	Delhi	6:00	8:40	02h 40m	BLR--->DEL	non-stop	7,487
1	25-Oct	GO FIRST	Bangalore	Delhi	21:15	23:55	02h 40m	BLR--->DEL	non-stop	7,487
2	25-Oct	SpiceJet	Bangalore	Delhi	2:10	4:50	02h 40m	BLR--->DEL	non-stop	7,488
3	25-Oct	Indigo	Bangalore	Delhi	12:55	15:40	02h 45m	BLR--->DEL	non-stop	7,488
4	25-Oct	Indigo	Bangalore	Delhi	14:35	17:20	02h 45m	BLR--->DEL	non-stop	7,488
...	...	...	...	...	...	...	...	...	...	...
7694	31-Oct-21	IndiGo	Chennai	Delhi	9:00	14:50	5:50	MAA-->DEL	1-stop	21055
7695	31-Oct-21	Air India	Chennai	Delhi	6:20	12:15	5:55	MAA-->DEL	1-stop	21213
7696	31-Oct-21	Vistara	Chennai	Delhi	7:00	16:45	9:45	MAA-->DEL	2-stop	21318
7697	31-Oct-21	Vistara	Chennai	Delhi	7:00	16:30	9:30	MAA-->DEL	2-stop	22368
7698	31-Oct-21	Vistara	Chennai	Delhi	7:00	17:05	10:05	MAA-->DEL	2-stop	24941

7699 rows × 10 columns

• Data Preprocessing Done:

A few basic cleaning and feature engineering looking at the data. A lot of data preparation needs to be done according to the model and strategy we use, but here are the basic cleaning I did initially to understand the data better. There were not many, but a few repetitions in the data collected. There were no missing values in the dataset and since all the feature variables are of object data type, we need not check for it’s skewness, outliers or distribution.

• Data Inputs- Logic- Output Relationships:

We can see this graphically that Jet Airways are more costly and Spice Jet are very affordable. Also the flights that provide free meal and additional facilities are more costly as compared to the direct flights.

• Hardware and Software Requirements and Tools Used:

Python was the most popular technology for implementing machine learning ideas, owing to the fact that it has a large number of built-in algorithms in the form of bundled libraries. The following are some of the most important libraries and tools we used in our project:

1. Numpy:
   NumPy is a Python module for array processing. It includes a highperformance multidimensional array object as well as utilities for manipulating them. It is the most important Python module for scientific computing. NumPy may be used as a multi-dimensional container of general data in addition to its apparent scientific applications. Numpy allows any data types to be created, allowing NumPy to connect with a broad range of databases cleanly and quickly.

2. SciPy:
   SciPy is a Python library for scientific and technical computing that is free and open-source. Optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, and other activities used in research and engineering are all covered by SciPy modules. SciPy is based on the NumPy array object, and it's part of the NumPy stack, which also contains Matplotlib, pandas, and SymPy, as well as a

growing number of scientific computing libraries. Other apps with comparable users to NumPy include MATLAB, GNU Octave, and Scilab. The SciPy stack is occasionally used interchangeably with the NumPy stack. The SciPy library is now available under the BSD licence, with an open community of developers sponsoring and supporting its development.

3. Scikit-Learn
   Scikit-learn offers a standard Python interface for a variety of supervised and unsupervised learning techniques. It is provided under several Linux distributions and is licenced under a liberal simplified BSD licence, promoting academic and commercial use. The library is being constructed.

4. Jupyter Notebook
   Jupyter Notebook is an open-source online software that lets you create and share documents with live code, equations, visualisations, and narrative prose. Data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and more are all included. Jupyter Notebook is an open-source online software that lets you create and share documents with live code, equations, visualisations, and narrative text. Data cleansing and transformation, numerical simulation, statistical modelling, data visualisation, machine learning, and more are all included.

5. Chrome driver

Chrome driver is used to web scrape the data and automate the process. I have scraped the data using Selenium python.

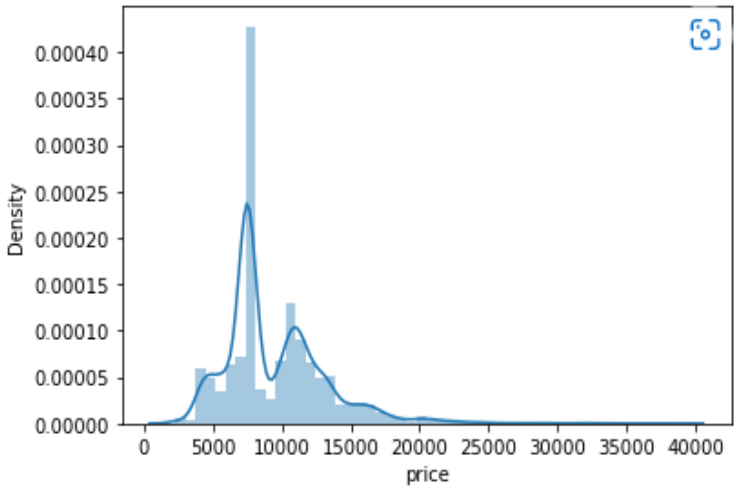
Model/s Development and Evaluation

• Identification of possible problem-solving approaches (methods)

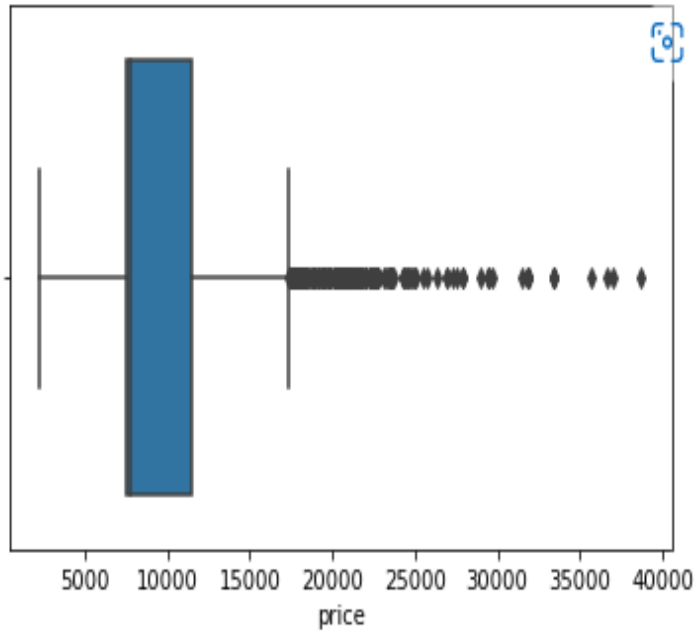
We go over many techniques and datasets that were used to create this module. The model will be trained using a dataset comprising over 1500 tuples. The price of a flight is determined by factors such as the number of stops, duration, the kind of facilities provided, etc. I created regressor methods and compared all on different flight models because this is a regression problem. Anaconda seeks to address Python's dependency well, where distinct projects have various dependency versions, so that project dependencies do not require separate versions, which might conflict.

• Visualizations:

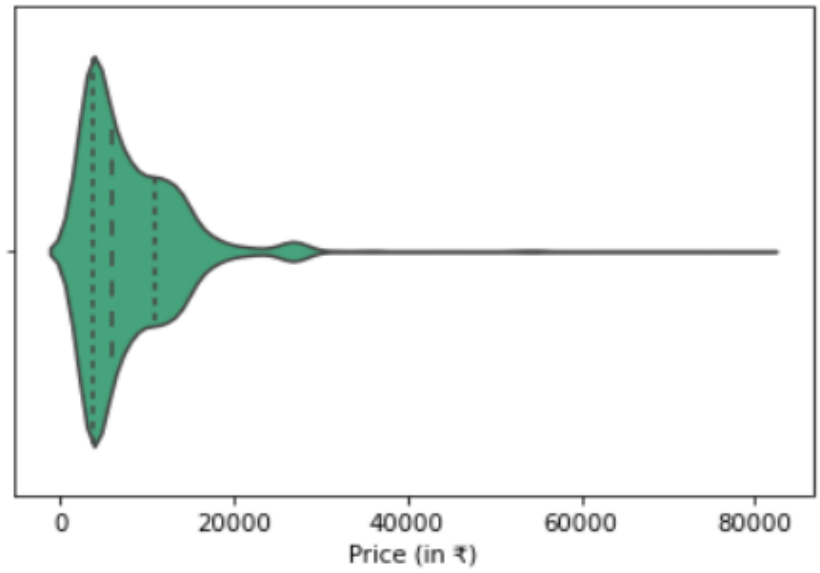
1. Distribution plot:



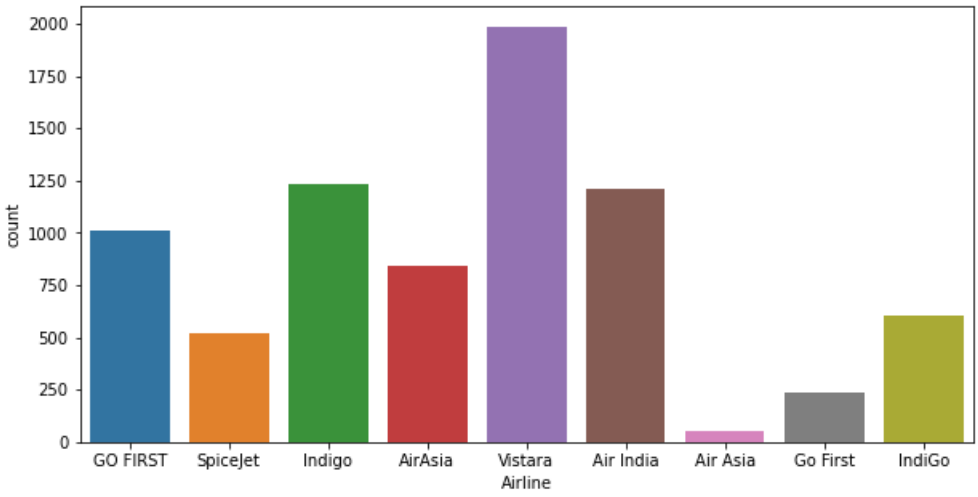
2. Box plot:

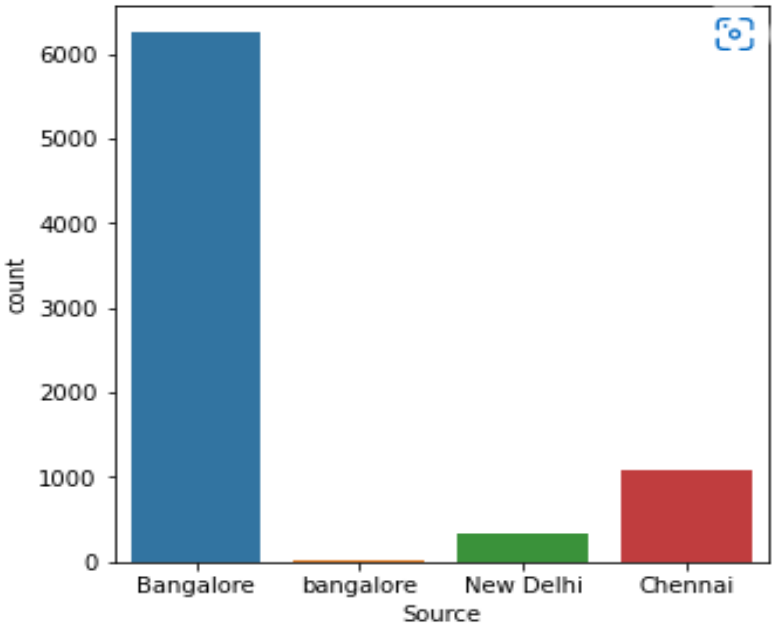


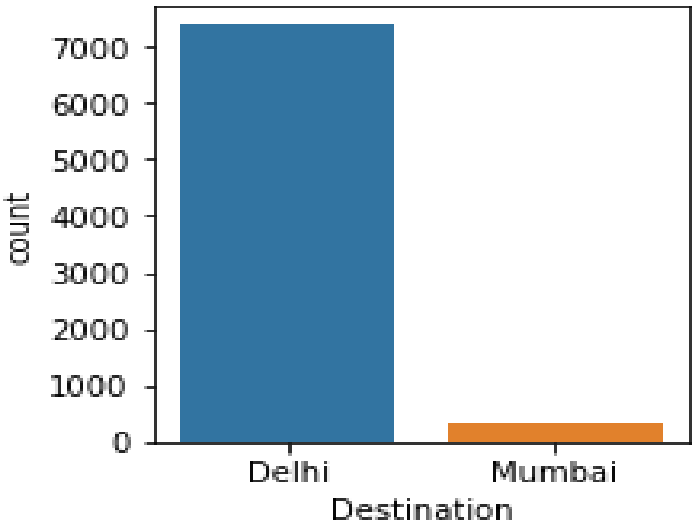
3. Violin plot:

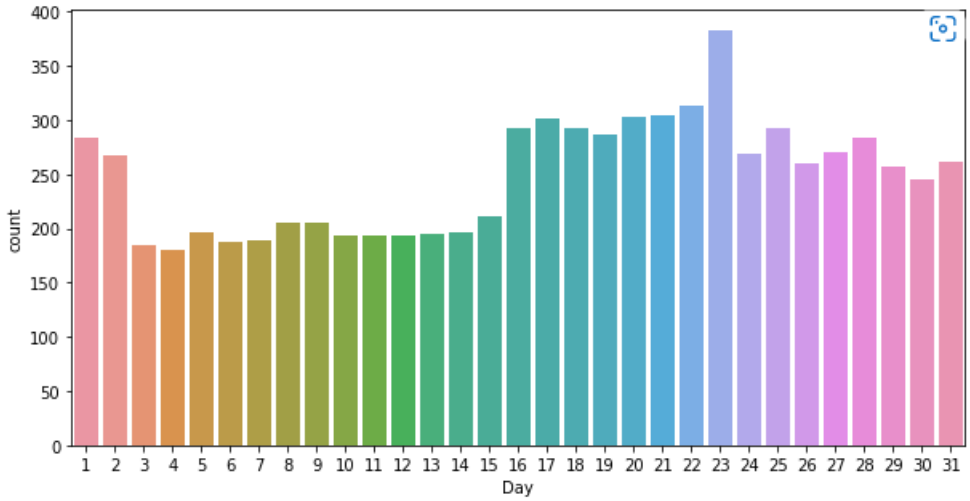


5.Count plot:

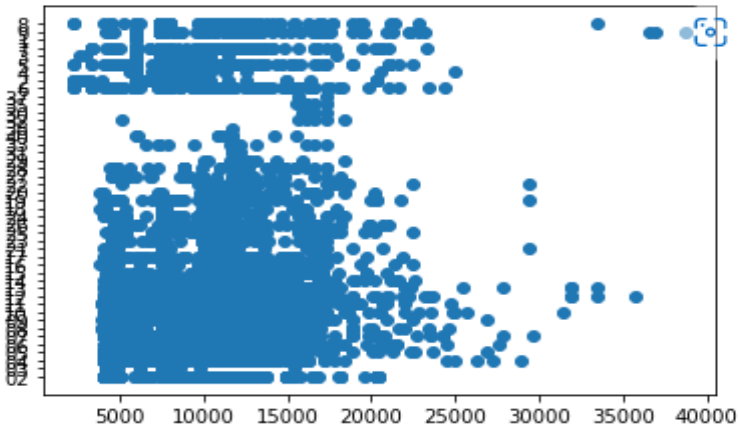

2000
1750
1500
1250
1000
750
500
250
0
count
GO FIRST
SpiceJet
Indigo
AirAsia
Vistara
Air India
Air Asia
Go First
IndiGo
Airline

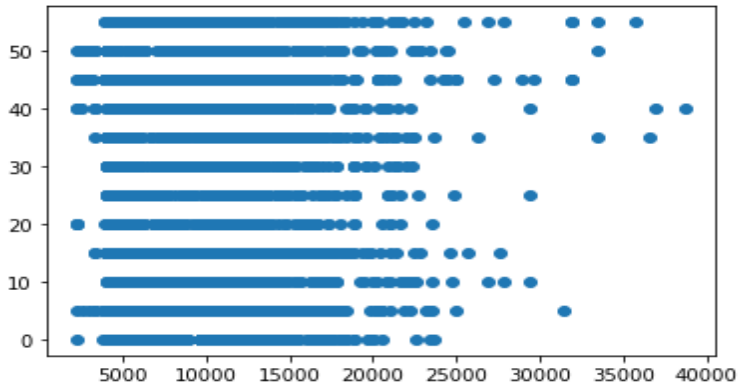

6000
5000
4000
3000
2000
1000
0
count
Bangalore
bangalore
New Delhi
Chennai
Source

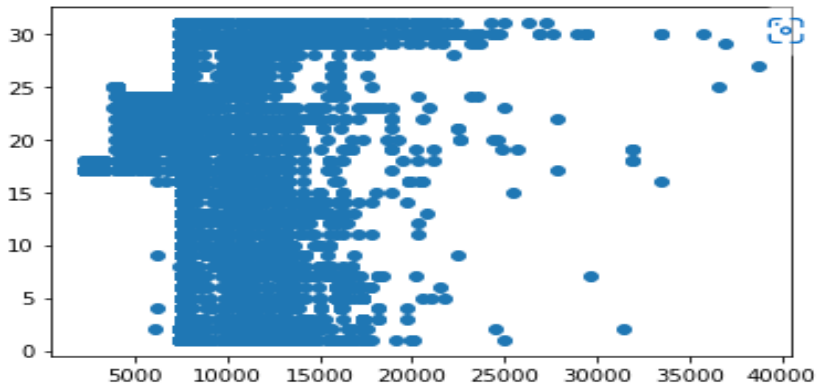




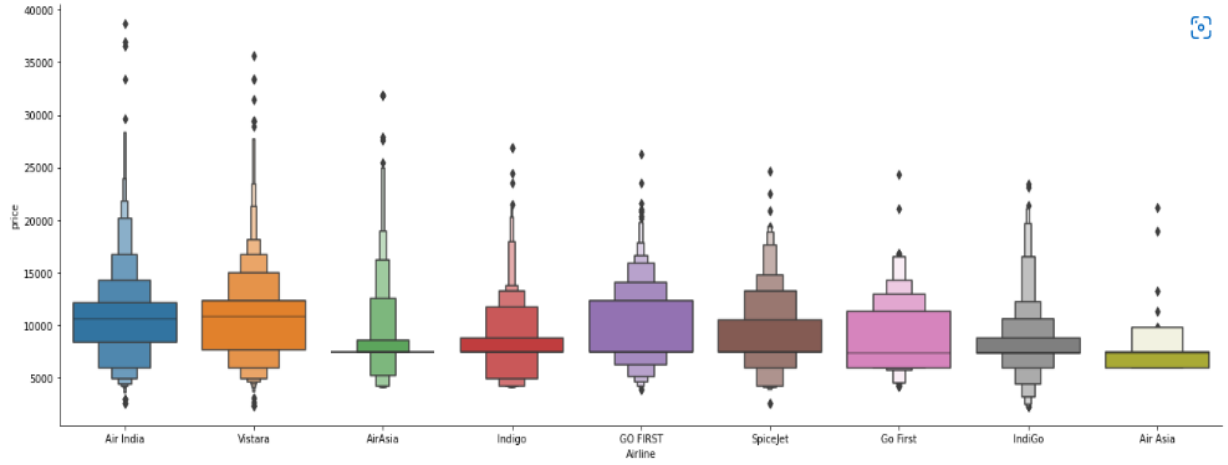
6. scatter plot:




50
40
30
20
10
0
5000
10000
15000
20000
25000
30000
35000
40000


30
25
20
15
10
5
0
5000
10000
15000
20000
25000
30000
35000
40000

7.catplot:


price
40000
35000
30000
25000
20000
15000
10000
5000
Air India
Vistara
AirAsia
Indigo
GO FIRST
SpiceJet
Go First
IndiGo
Air Asia
Airline

7. Heat map:

• Interpretation of the Results:

The Jet Airway Airlines are more costly than others whereas SpiceJet and IndiGo are quite affordable. Flights from metro cities are more in number and hence few are in budget and few are way too expensive. The expensive flights usually come with layover(long/short), free meal and some other additional facilities as well.

CONCLUSION

• Key Findings and Conclusions of the Study:

The trend of flight prices vary over various months and across the holiday. There are two groups of airlines: the economic group and the luxurious group. Spicejet, AirAsia, IndiGo, Go Air are in the economical class, whereas Jet Airways and Air India in the other. Vistara has a more spread out trend.

• Learning Outcomes of the Study in respect of Data Science:

Collected and analysed data for 6 routes which spanned across business & tourist routes in India

Some of the routes had non-decreasing prices and thus the model suggested to buy the ticket always

Implemented algorithms like Decision Tree, Random Forest , Gradient Boosting and statistical analysis

• Limitations of this work and Scope for Future Work:

I can also consider days of weeks and months as well to predict the price more efficiently.

To make the prediction accurate, time of booking should also be considered (as in when a person is booking tickets, how many days prior to the journey?, etc.)