CSE574LECC: Intro to Machine Learning

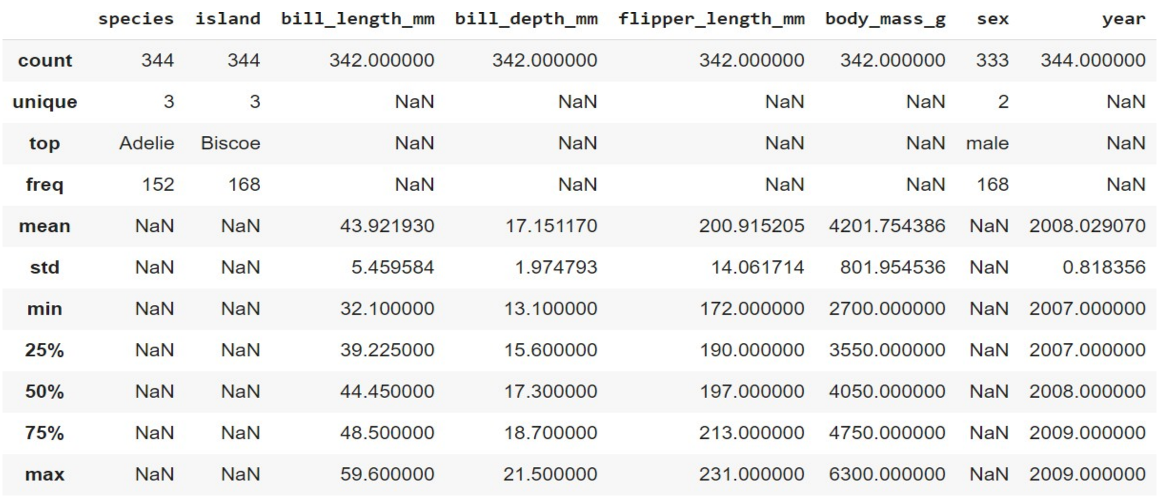
Penguin Species & Flight Price Prediction Datasets

Part 1: Logistic Regression

Dataset Details

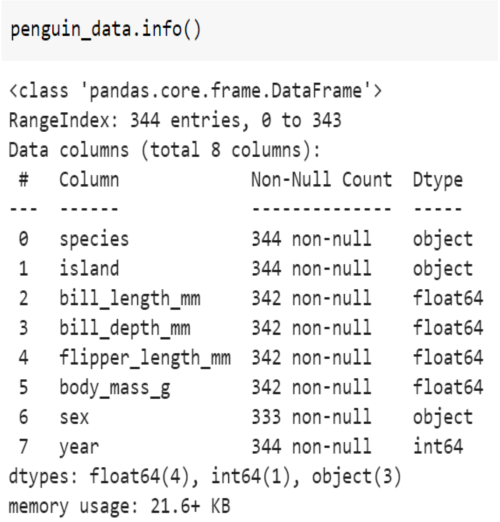
The dataset given for this task is the Penguins dataset, which includes 344 instances and 7 attributes. The dataset comprises of 4 columns with numerical values, and 3 columns with categorical values representing the species, island, and sex.

Data Statistics



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
count	344	344	342.000000	342.000000	342.000000	342.000000	333	344.000000
unique	3	3	NaN	NaN	NaN	NaN	2	NaN
top	Adelie	Biscoe	NaN	NaN	NaN	NaN	male	NaN
freq	152	168	NaN	NaN	NaN	NaN	168	NaN
mean	NaN	NaN	43.921930	17.151170	200.915205	4201.754386	NaN	2008.029070
std	NaN	NaN	5.459584	1.974793	14.061714	801.954536	NaN	0.818356
min	NaN	NaN	32.100000	13.100000	172.000000	2700.000000	NaN	2007.000000
25%	NaN	NaN	39.225000	15.600000	190.000000	3550.000000	NaN	2007.000000
50%	NaN	NaN	44.450000	17.300000	197.000000	4050.000000	NaN	2008.000000
75%	NaN	NaN	48.500000	18.700000	213.000000	4750.000000	NaN	2009.000000
max	NaN	NaN	59.600000	21.500000	231.000000	6300.000000	NaN	2009.000000

Checking for null values



```
penguin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   bill_length_mm     342 non-null    float64
 3   bill_depth_mm      342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                333 non-null    object 
 7   year               344 non-null    int64  
dtypes: float64(4), int64(1), object(3)
memory usage: 21.6+ KB
```

Null values are present in bill length, bill_depth,flipper_length,body_mass and sex

Data Modelling

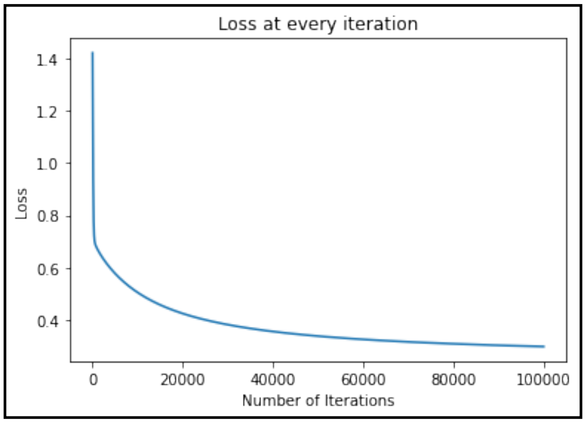
After dividing the dataset into training and testing sets and applying Logistic Regression on the training set, we conducted hyperparameter tuning to improve the model's performance. Specifically, we experimented with different values for the learning rate and number of iterations and generated three different results.We then evaluated each of these three models using the testing set and compared their accuracies to determine the best-performing model.

1.Best accuracy.

We received an accuracy of 89.855 with 0.45 learning rate and 10,000 iterations.

2.Include loss graph and provide a short analysis of the results.

Smaller learning rate and more number of iterations results in reaching the optimal value although the step size will be small without over estimating the gradient.



3.*Explain how hyperparameters influence the accuracy of the model. Provide at least 3 different setups with learning rate and #iterations and discuss the results along with plotting of graphs.*

Learning Rate	Number of iterations	Model Accuracy
0.001	100000	88.40
0.45	10000	89.855
0.009	100000	91.30

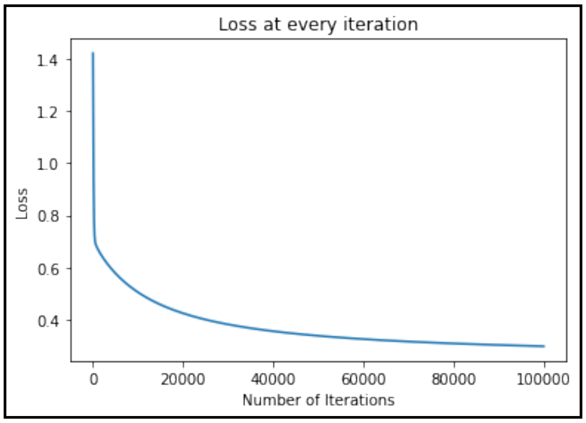


Figure 1:Learning rate 0.009 and iterations 100000

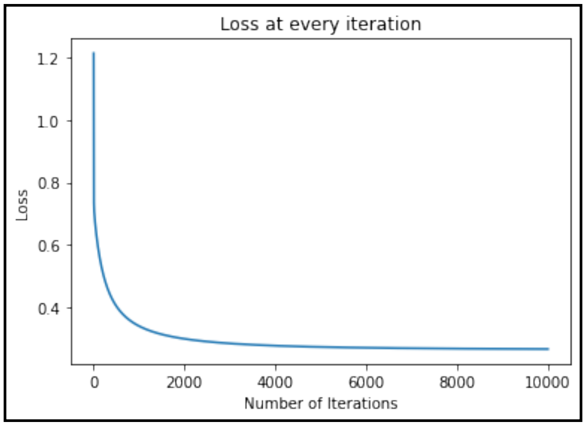


Figure 2:Learning rate 0.45 and iterations 10000.

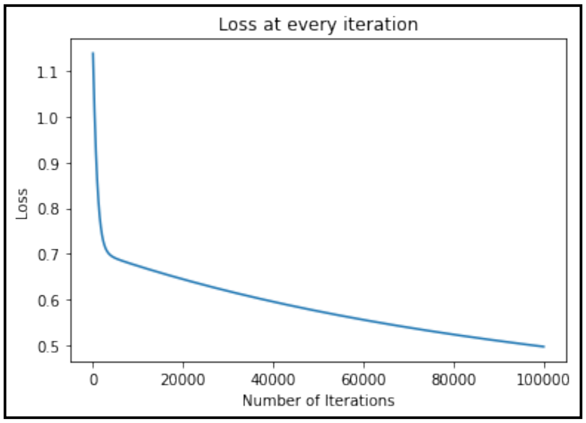


Figure 3:Learning rate 0.001 and iterations 100000

Observations:

- ➢ If we maintain constant number of iterations ie.100000 and the learning rate is increasedfrom 0.001 to 0.009, the accuracy of model is increased.
  - ○ Reason: Increasing the learning rate can lead to faster convergence and a better optimization of the model. As a result, the model is better able to fit the training data, leading to improved accuracy.
- ➢ As the learning rate is increase the accuracy is increased
  - ○ Reason Increasing the learning rate too much can lead to overshooting the optimalsolution, leading to decreased performance. However, if the learning rate isincreased within a reasonable range, the model can converge faster and lead to higher accuracy.
- ➢ As the number of iterations is decreased from 100000 to 10000and learning rate is increased from 0.001 to 0.45, the accuracy is increased.
  - ○ Reason: Decreasing the number of iterations reduces the amount of training timerequired for the model to converge.
  - ○ Increasing the learning rate allows the model to converge faster during each iteration, making up for the reduction in the number of iterations.
  - ○ This combination of higher learning rate and fewer iterations can lead to faster convergence and a better optimization of the model, resulting in higher accuracy. However, it is important to note that decreasing the number of iterations too muchcan lead to underfitting the data, which would result in decreased accuracy.

4. Discuss the benefits/drawbacks of using a Logistic Regression model.

Benefits

a. Simple algorithm
b. Easy to update.
c. Provides probabilities for each class.
d. Performs well data needs to be linearly separable.

Drawbacks

a. Chances of overfitting of data for high dimensional datasets
b. Cannot predict continuous outcomes.
c. Presence of multicollinearity between independent variables can impact the modelperformance.
d. Sensitive to outliers

Part 2:
Linear Regression

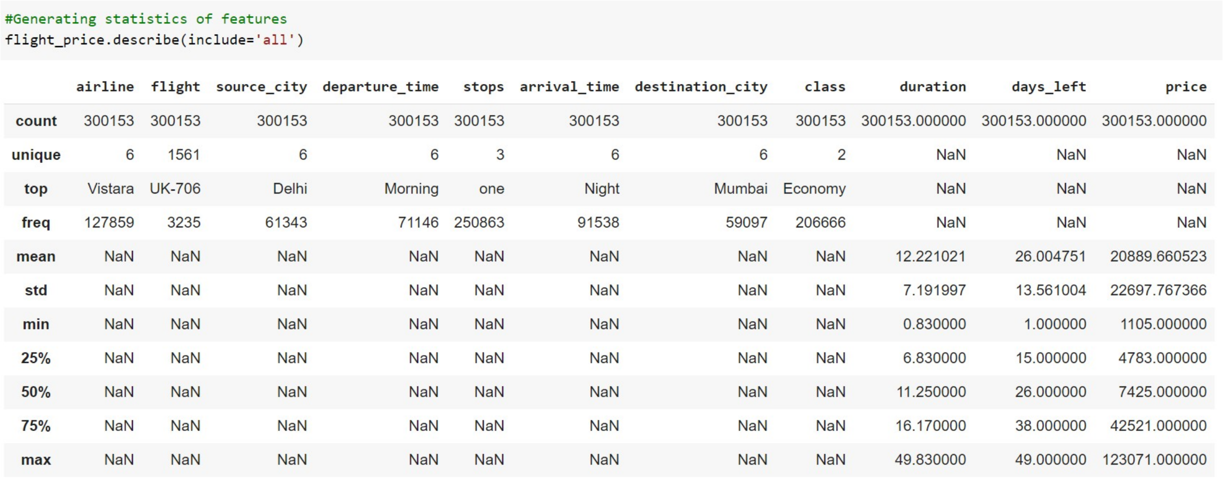
1. Dataset Details

The dataset we have taken for this task is Flight_price_prediction data. It has 11 features, and300153 samples present with both numerical and categorical columns.

Data Description:

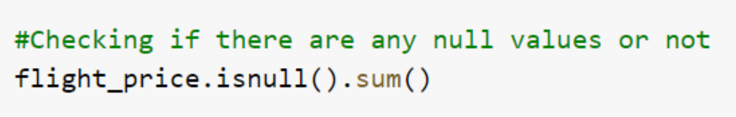
- Airline: Type of airlines
- Flight: flight name
- Source_city: city from which flight takes off.
- Destination_city: city where flight lands
- Departure time: at which part of the day the flight takes off.
- Arrival time: at which part of the day the flight arrives in destination_city
- Stops: Whether there is a stop in between or not
- Class: type of class seating in flight
- Duration: time taken to reach destination_city
- Days_left: time left for booking a ticket
- Price: cost of the ticket -Target variable

2. Statistics of dataset

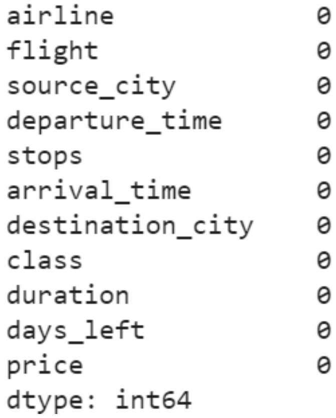


```
#Generating statistics of features
flight_price.describe(include='all')
```

	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
count	300153	300153	300153	300153	300153	300153	300153	300153	300153.000000	300153.000000	300153.000000
unique	6	1561	6	6	3	6	6	2	NaN	NaN	NaN
top	Vistara	UK-706	Delhi	Morning	one	Night	Mumbai	Economy	NaN	NaN	NaN
freq	127859	3235	61343	71146	250863	91538	59097	206666	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	12.221021	26.004751	20889.660523
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	7.191997	13.561004	22697.767366
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.830000	1.000000	1105.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6.830000	15.000000	4783.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	11.250000	26.000000	7425.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	16.170000	38.000000	42521.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	49.830000	49.000000	123071.000000



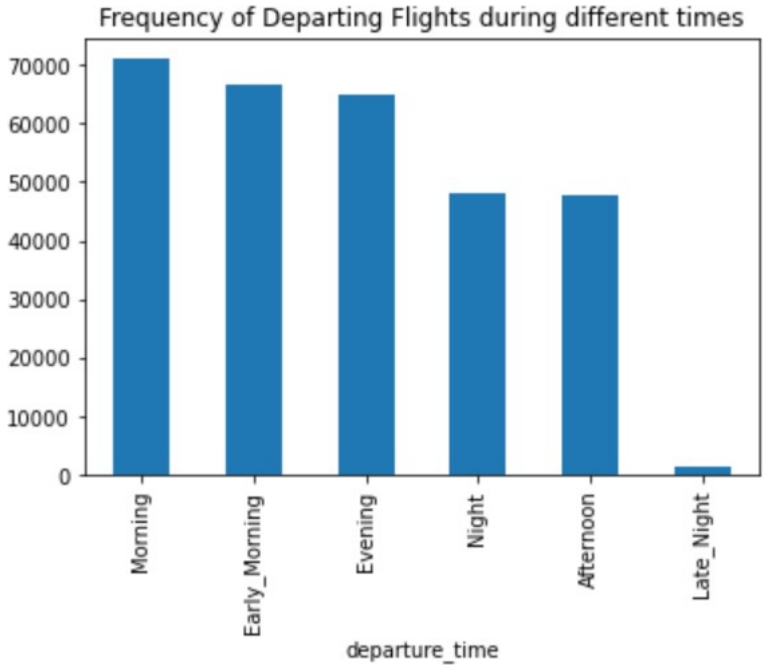
```
#Checking if there are any null values or not
flight_price.isnull().sum()
```



```
airline             0
flight              0
source_city         0
departure_time      0
stops               0
arrival_time        0
destination_city    0
class               0
duration            0
days_left           0
price               0
dtype: int64
```

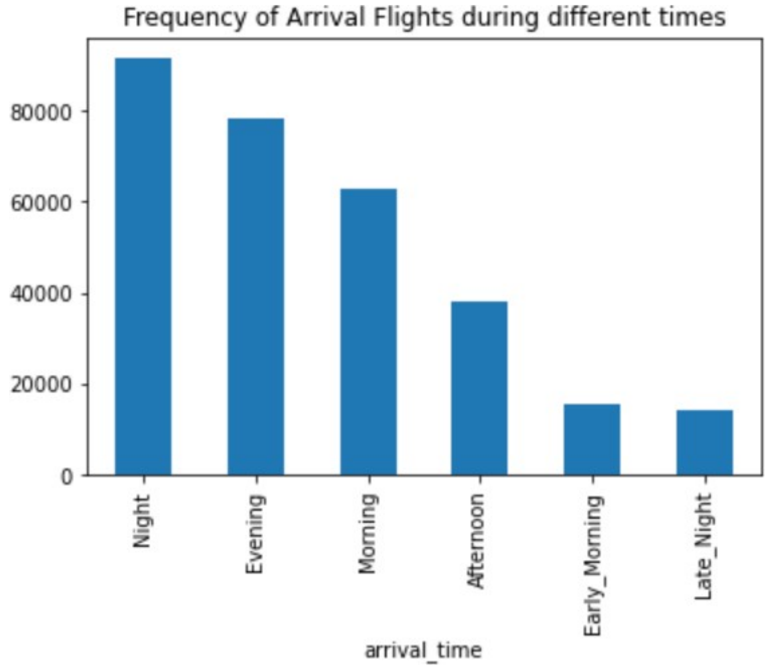
3. Data Visualization

1. No of flights flying during different times of the day from source city



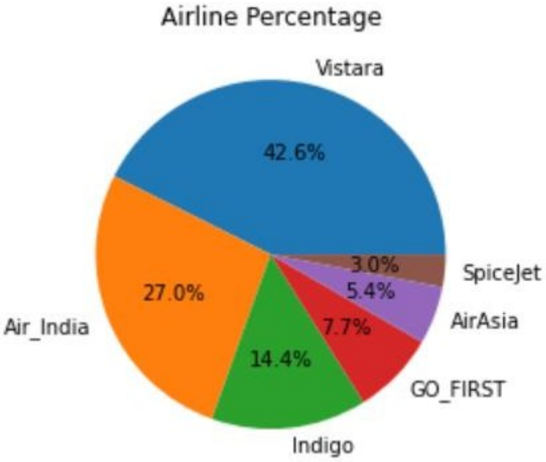
It can be observed that a greater quantity of flights takes off during the morning hours in comparison to other times, while there is a notably lower amount of flights departing late at night

No of flights arriving at different times of the day to destination city



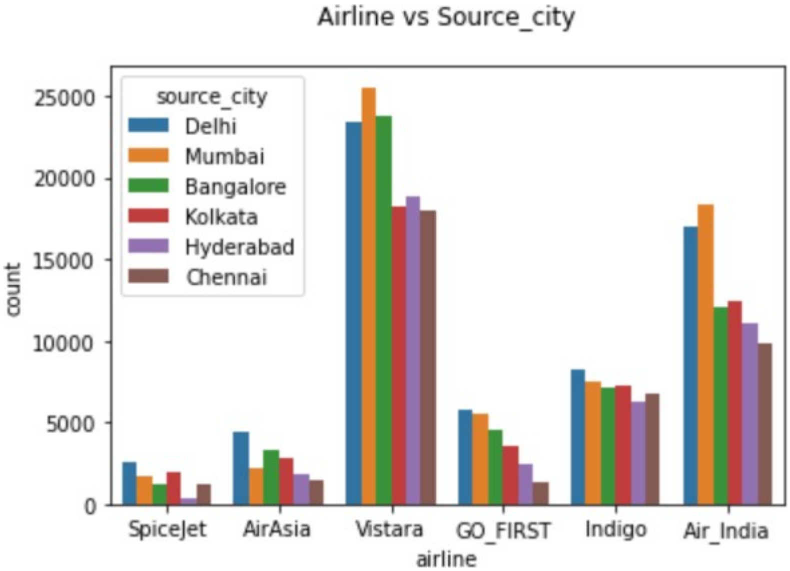
It can be observed that most flights arrive at the destinations during night times.

2. Percentage of number of each airline present in the whole dataset



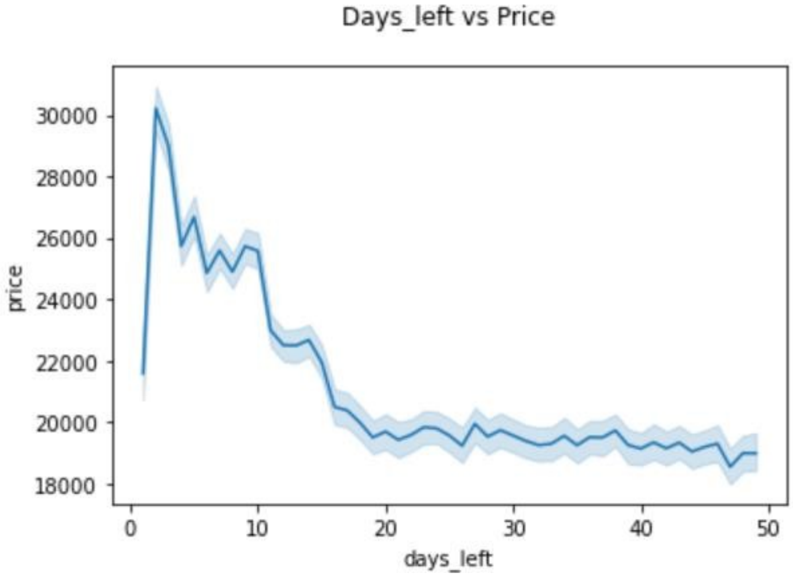
Based on the information depicted in the pie chart, Vistara appears to be the most prevalent flight company, while SpiceJet has the least representation.

3. Number of Airlines per Source City



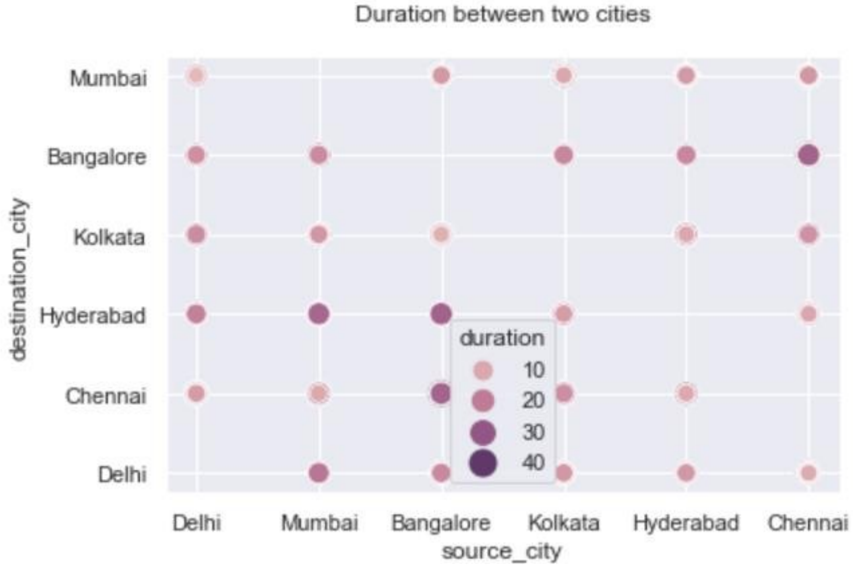
The bar graph shows that SpiceJet and AirAsia have a high frequency of flights departing from Delhi, Vistara flies frequently from Mumbai, and Go_First, Indigo, and Air India have a high number of flights departing from Delhi. Based on this information, it can be inferred that most major airlines have a significant presence in Delhi in terms of flight operations

4. Price variation as the number of days decrease



Initially, when there were 45 days remaining until the flight, the ticket price was approximately 19,000. However, as the days passed and the departure date approached, the price of the ticketgradually increased, ultimately reaching a cost of around 30,000 when only 5 days were left untilthe flight.

5. Duration of journey between two cities

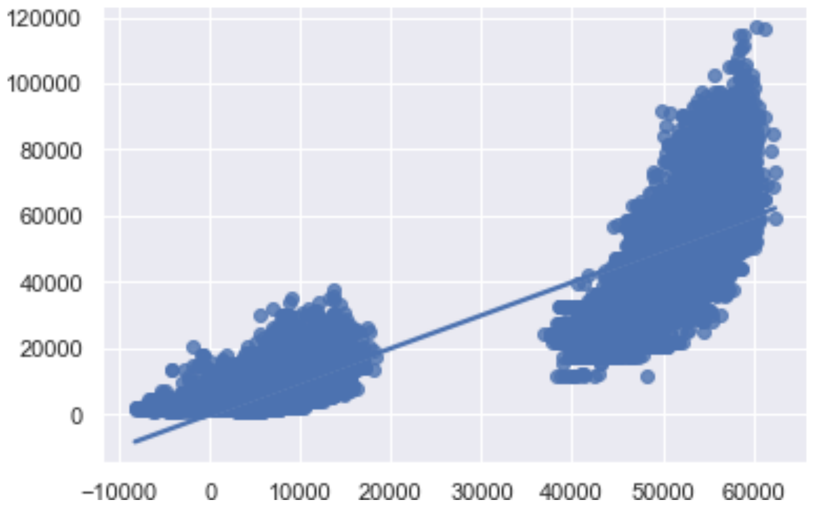


The duration of the flights between Hyderabad and Mumbai, Hyderabad and Bangalore, and Bangalore and Chennai are approximately 40 minutes, which is the longest flight duration compared to other flight routes between different cities.

4. Provide your loss value

MSE Loss for train data =
45451763.27047183MSE Loss for test data =
46280193.48798858

5. Show the plot comparing the predictions vs the actual test data

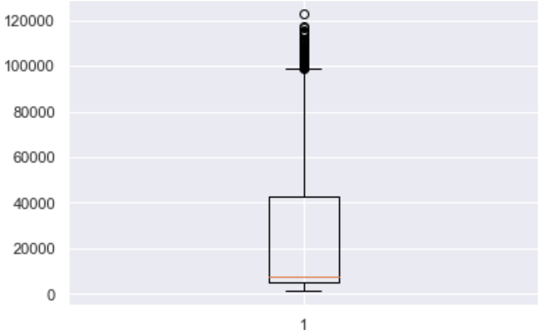


The difference between predicted price and actual price values i.e., loss is very high and there are some negative price values in the predictions. This is due to following reasons:

1. The range of values of target i.e., price is very high when compared to range of numerical independent columns i.e., duration and days_left.

2. The data has a greater number of categorical columns than numerical columns.

3. The target has many outliers, and the median is less than 10000 whereas the range is 120000. The price column has high variance.



6.Discuss the benefits/drawbacks of using OLS estimate for computing the weights.

Benefits:

1. OLS method is simple, efficient and easy to compute.

2. It is used to get maximum information from small datasets.

DrawBacks:

1. It is sensitive to outliers, which can have significant effect on computed weights.

2. OLS assumes that the relation between independent and dependent variables is linear and most of the real time data is non-linear.

3. If there is multicollinearity in the data, it can lead to reduced accuracy of model and unstable values of weights.

4. It may perform poorly on datasets which have single independent variable and multiple dependent variables.

7. Discuss the benefits/drawbacks of using a Linear Regression model.

Benefits:

1. It is easy to implement and computationally inexpensive.

2. Linear regression results can be easily interpreted.

3. It performs well if the relationship between dependent and independent variables is linear.

Drawbacks:

1. Sensitive to outliers.

2. Prone to overfitting and there is no way to control it.

3. It assumes the relationship between dependent and independent variables is linear and most of the real time data is non-linear.

4. Unstable if there is multicollinearity in the data.

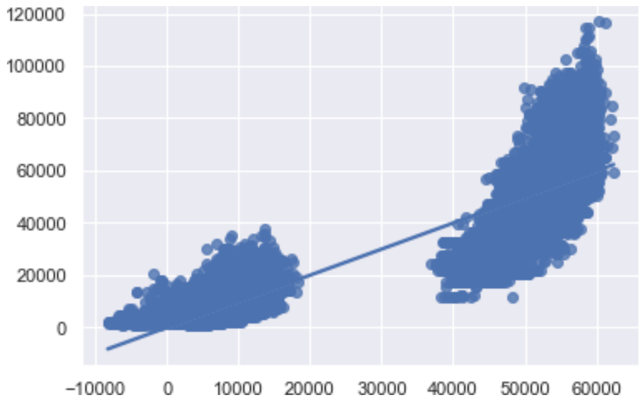
5. It is not suitable to model the datasets which have more number of categorical columns.

Part 3:
Ridge Regression

1. Provide your loss value

MSE Loss for train data =
50535490.228481755MSE Loss for test data =
51363920.17872904

2. Show the plot comparing the predictions vs the actual test data



The difference between predicted price and actual price values i.e., loss is very high and there are some negative price values in the predictions. The reasons are same as mentioned in the OLS method.

3. Discuss the difference between Linear and Ridge regressions. What is the mainmotivation for using l2 regularization?

The main difference between linear and ridge regressions is that ridge regression adds a penaltyterm to loss function, which helps to prevent overfitting. The main motivation for using ridge regression is to improve the generalization of the model and it can deal with multicollinearity i.e.,reduces impact of highly correlated features.

4. Discuss the benefits/drawbacks of using a Ridge Regression model.

Benefits:
1. Reduces Overfitting.
2. Makes the model more generalizable.
3. Handles Multicollinearity.

Drawbacks:

1. It cannot perform feature selection as lasso regression. It retains the features even though some of them are not important.
2. Requires Hyperparameter Tuning i.e., lambda value and the performance depends on the choice of the parameters.

References:

Lecture Slides by Prof.Alina