

# Flight Analysis & Prediction

Final Project DiBimbing  
Data Science Batch 21

**G'aly Rizq Prima**



# Table Of Contents

**01**

**Data Understanding**

**02**

**Data Cleaning**

**03**

**EDA**

**04**

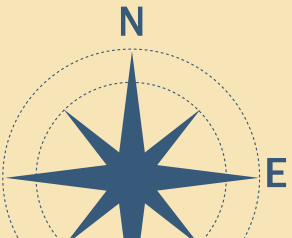
**Data Preprocessing**

**05**

**Modelling**

**06**

**Recommendation**



# Data Understanding



# Data Understanding

Source : <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>


<b><u>airline</u></b>	The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
<b><u>flight</u></b>	Flight stores information regarding the plane's flight code. It is a categorical feature.
<b><u>source_city</u></b>	City from which the flight takes off. It is a categorical feature having 6 unique cities.
<b><u>departure_time</u></b>	This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.
<b><u>stops</u></b>	A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
<b><u>arrival_time</u></b>	This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.

# Data Understanding



Source : <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

<b><u>destination_city</u></b>	City where the flight will land. It is a categorical feature having 6 unique cities.
<b><u>class</u></b>	A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
<b><u>duration</u></b>	A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
<b><u>days_left</u></b>	This is a derived characteristic that is calculated by subtracting the trip date by the booking date.
<b><u>price</u></b>	Target variable stores information of the ticket price.



# Data Cleaning



# Data Cleaning

## Missing Values

```
[9] df.isna().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
```

## Duplicate Values

```
[10] df.duplicated().sum()
```

```
0
```

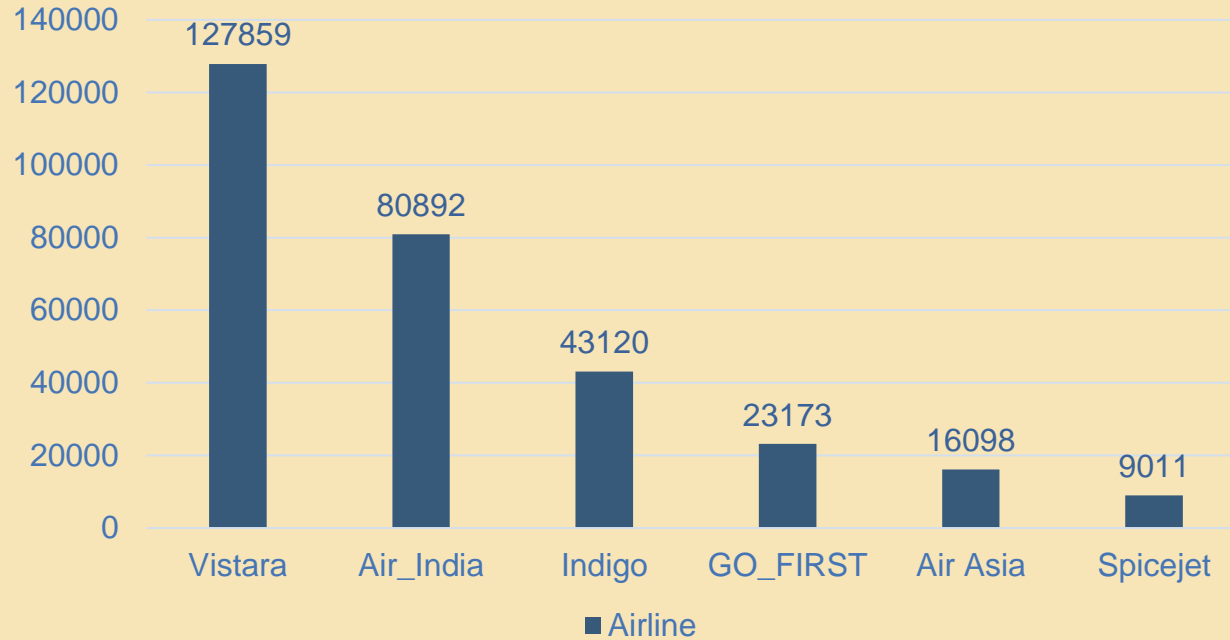
# Exploratory Data Analysis (EDA)





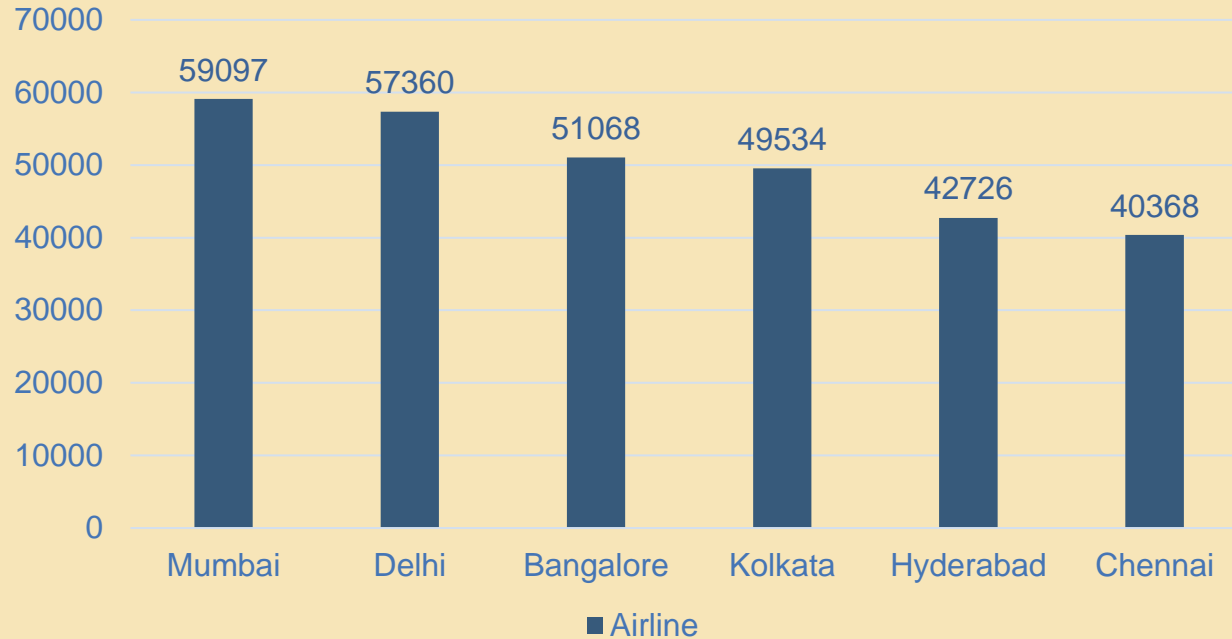
# Airlines

## Airline Frequency



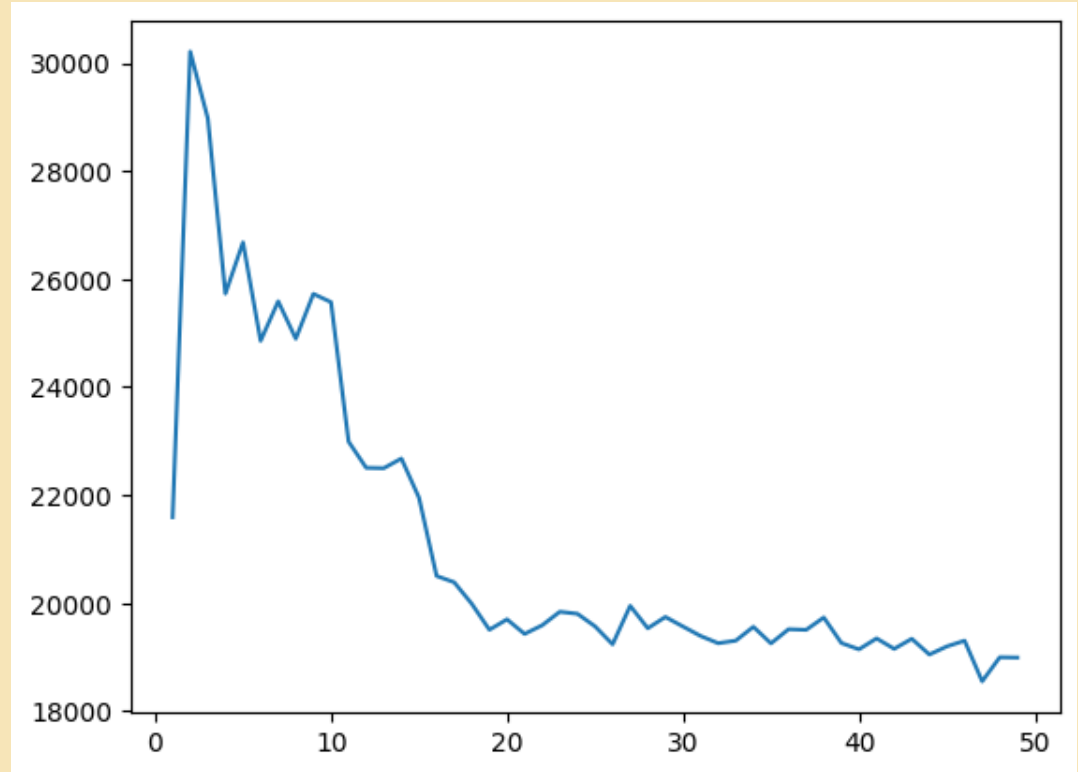
# Flight Destination City

## Destination Frequency



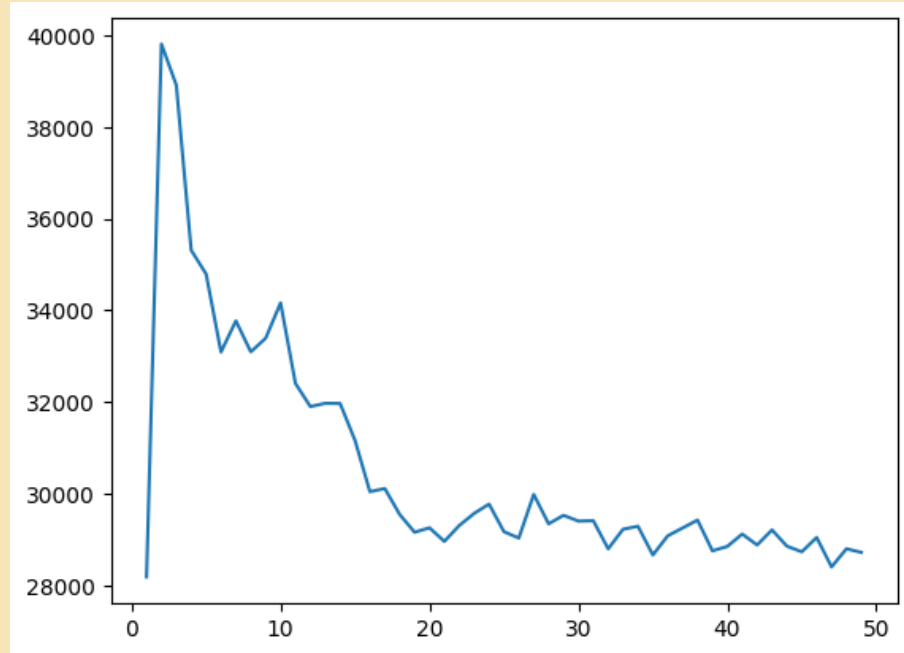
# Flight Price Trends Based On Remaining Days

Days_left	Price
1	21591.867151
2	30211.299801
3	28976.083569
4	25730.905653
5	26679.773368



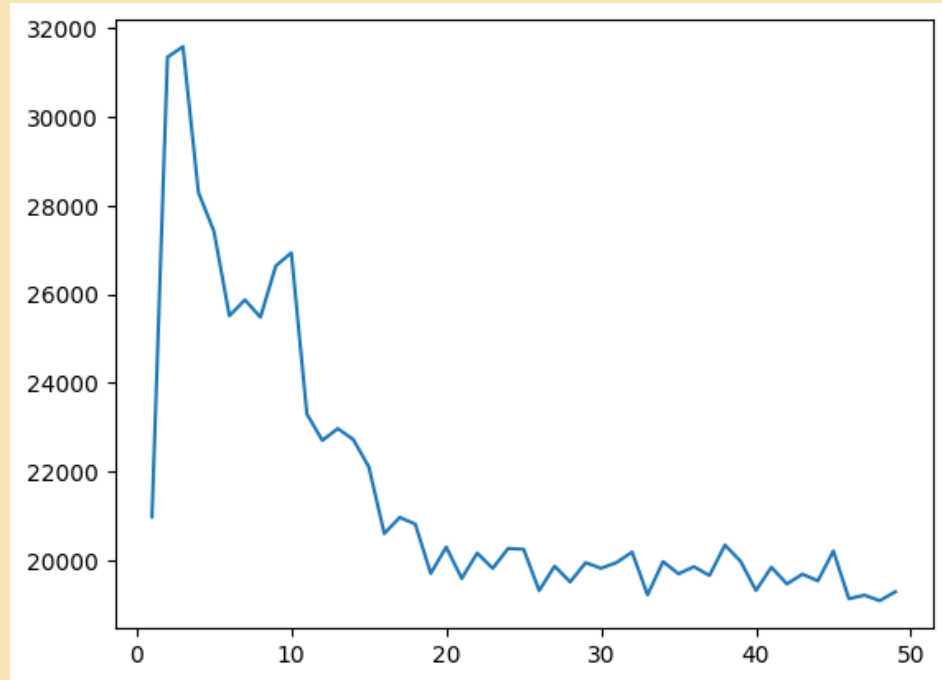
# Flight Price Trends Based On Remaining Days Seen From Vistara Airlines

Days_left	Price
1	28188.201950
2	39808.801423
3	38911.834053
4	35310.988511
5	34793.902605



# Flight Price Trends Based On Remaining Days Seen From Mumbai As Destination City

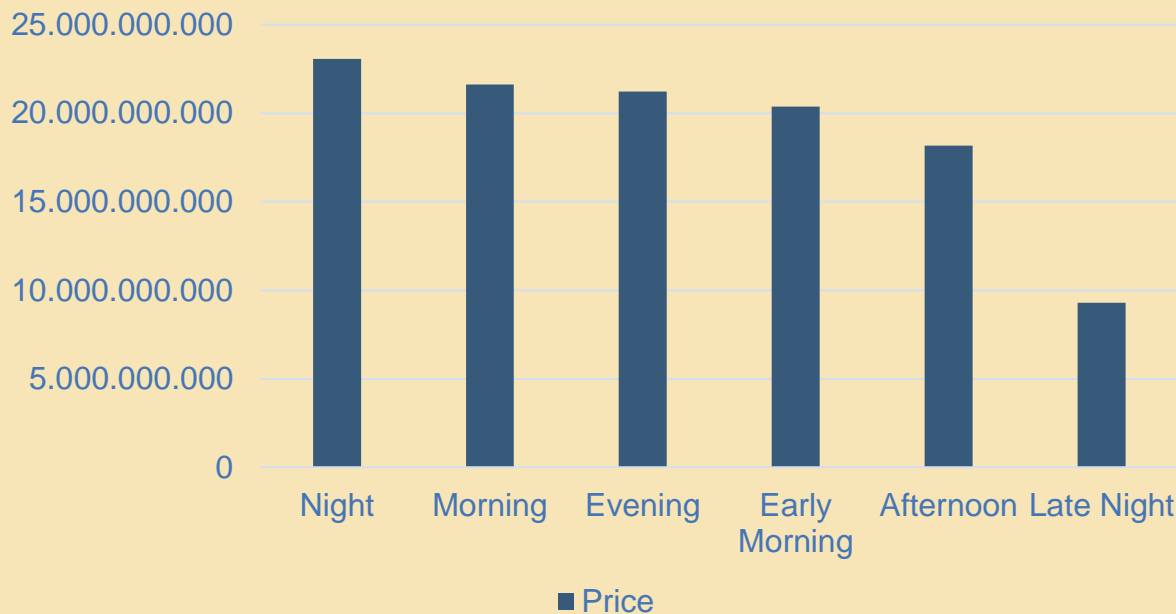
Days_left	Price
1	20982.274725
2	31340.490931
3	31575.331742
4	28288.153770
5	27415.648045



# Flight Price Trends Based On Departure Time

Departure Time	Price
Night	23062.146808
Morning	21630.760254
Evening	21232.361894
Early Morning	20370.676718
Afternoon	18179.203331
Late Night	9295.299387

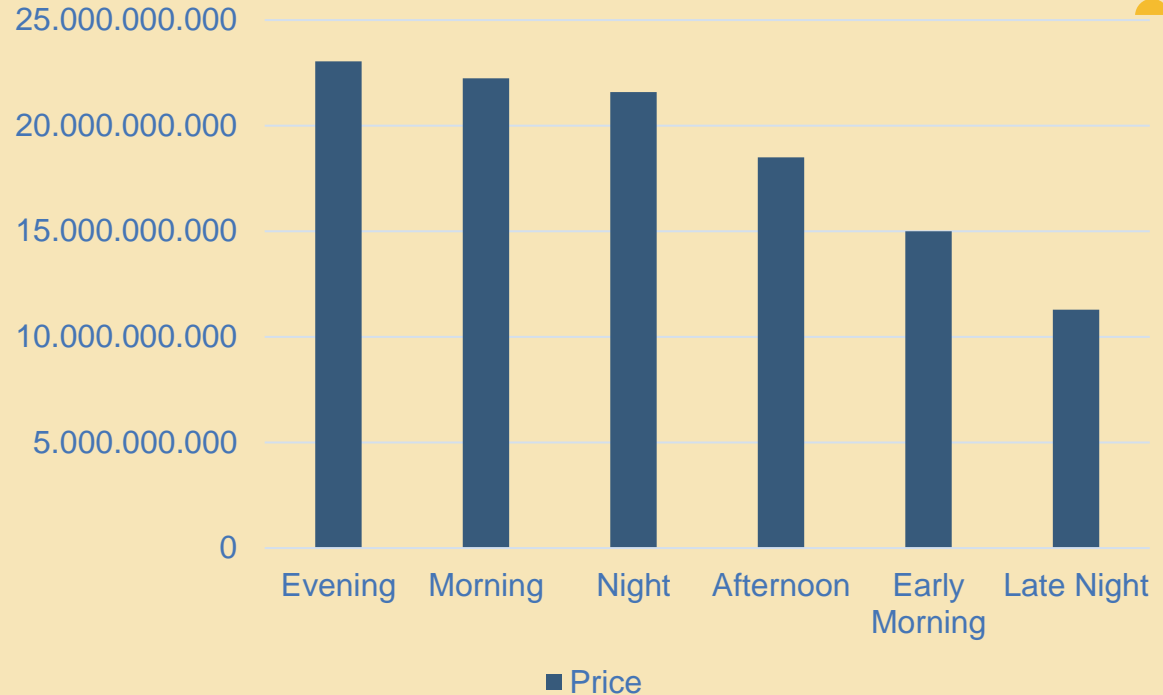
## Departure Time Price Average



# Flight Price Trends Based On Arrival Time

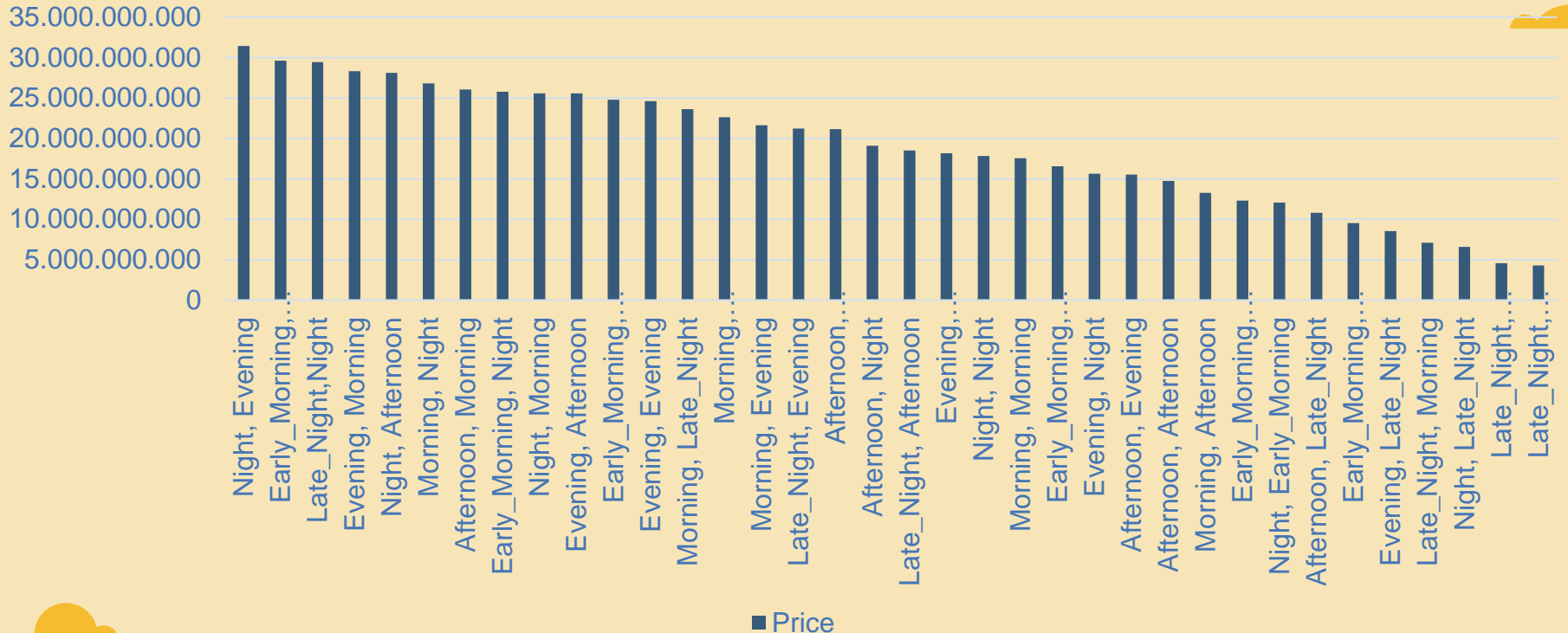
Departure Time	Price
Evening	23044.371615
Morning	22231.076098
Night	21586.758341
Afternoon	18494.598993
Early Morning	14993.139521
Late Night	11284.906078

Artival Time Price Average



# Flight Price Trends Based On Departure And Arrival Time

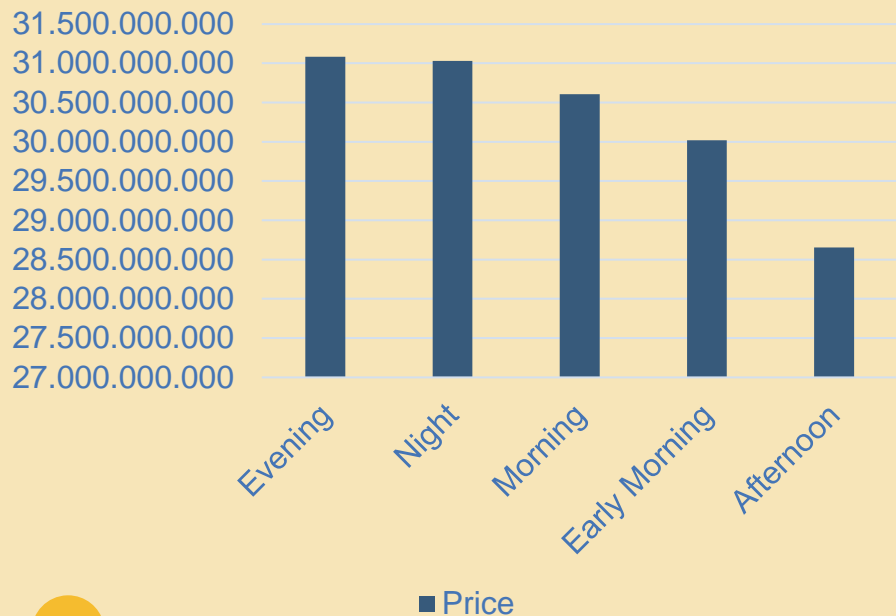
Departure Time, Arrival Time Average Price



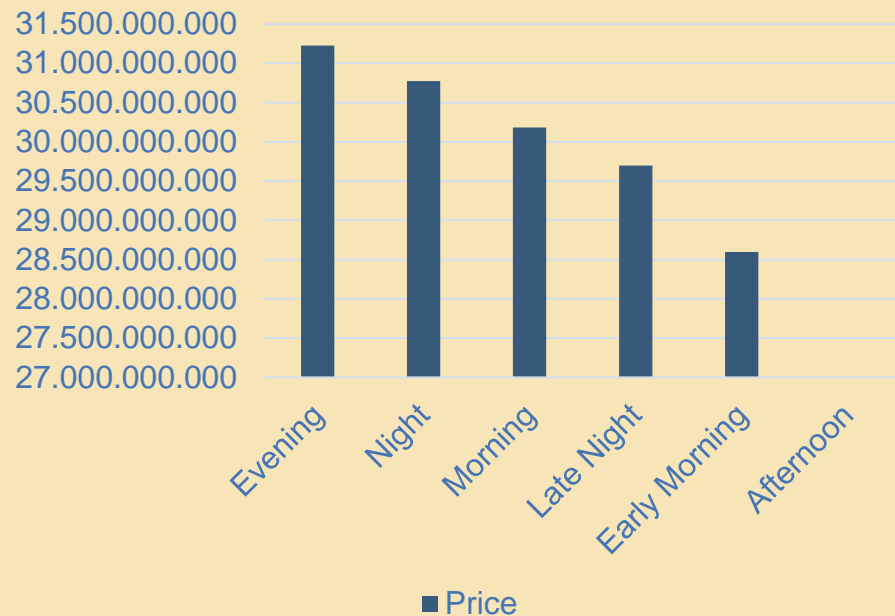


# Flight Price Trends Based On Departure & Arrival From Vistara Airlines

## Departure Price Average



## Arrival Price Average

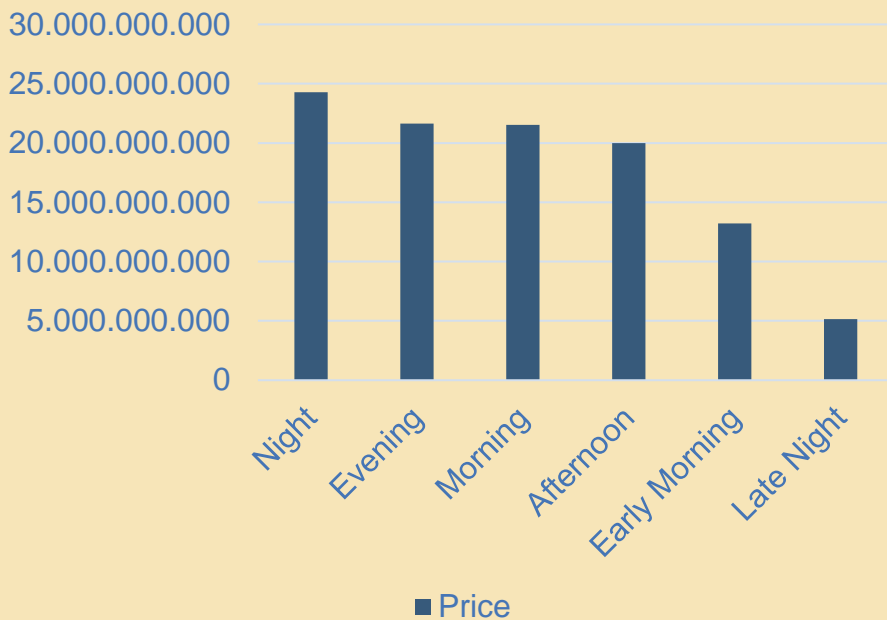


# Flight Price Trends Based On Departure & Arrival From Mumbai City As Destination

## Departure Average Price



## Arrival Average Price



# Flight Price Trends Based On Number Of Stops

Stops	Price
One	22900.992482
Zero	14113.450775
Two/More	9375.938535

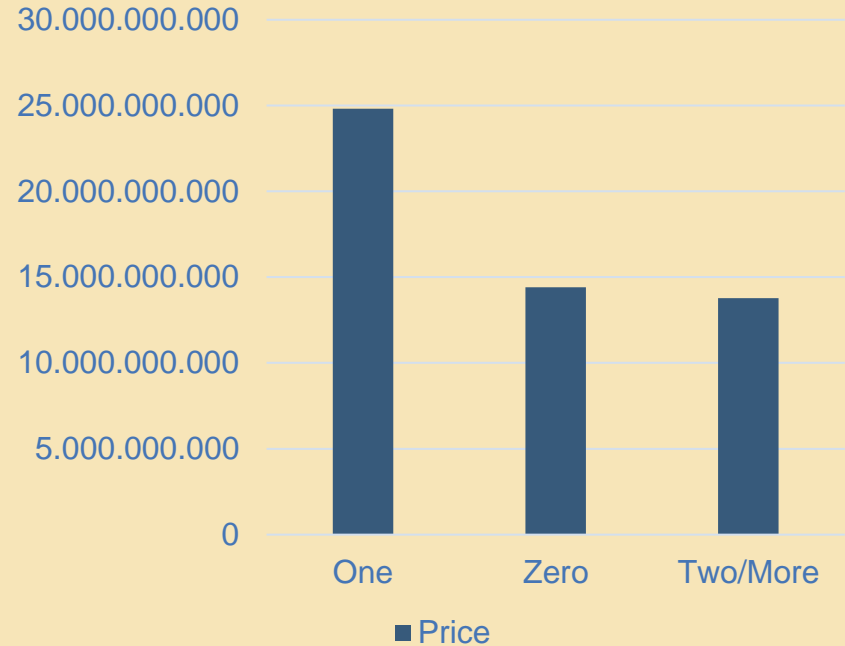


# Flight Price Trends Based On Number Of Stops From Airlines

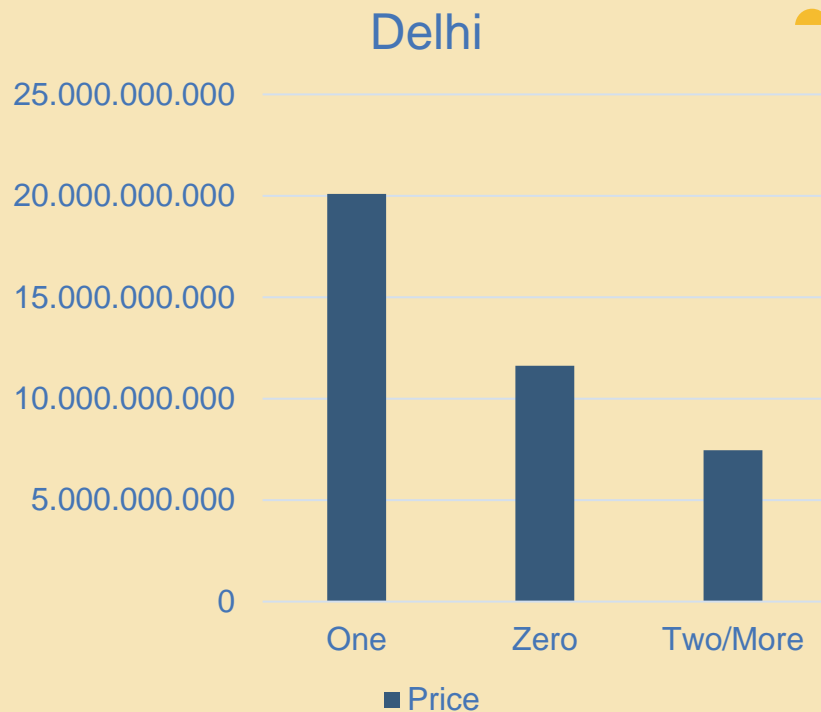
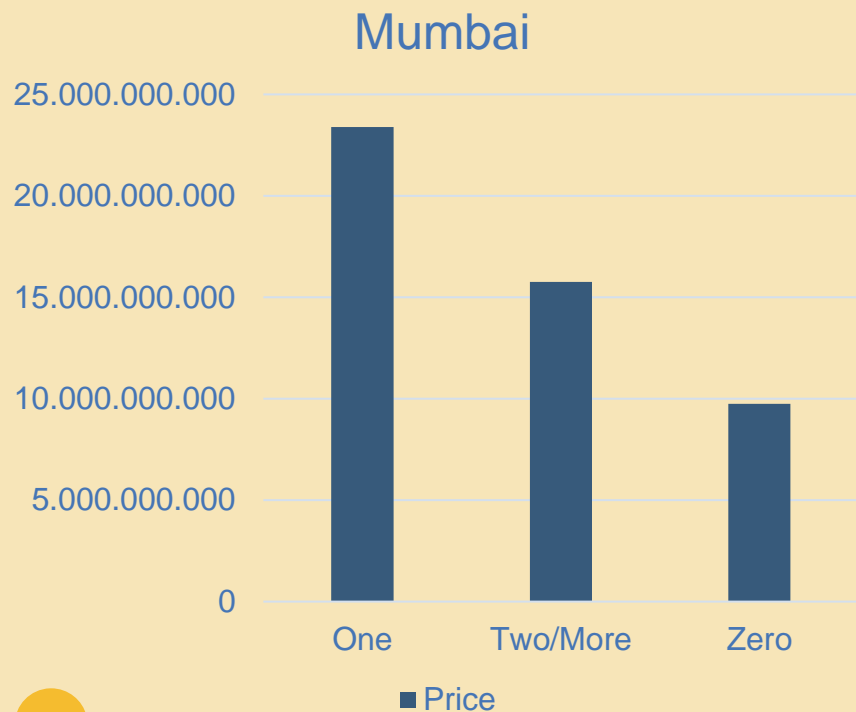
## Vistara



## Air India



# Flight Price Trends Based On Number Of Stops From Destination City



# Data Preprocessing

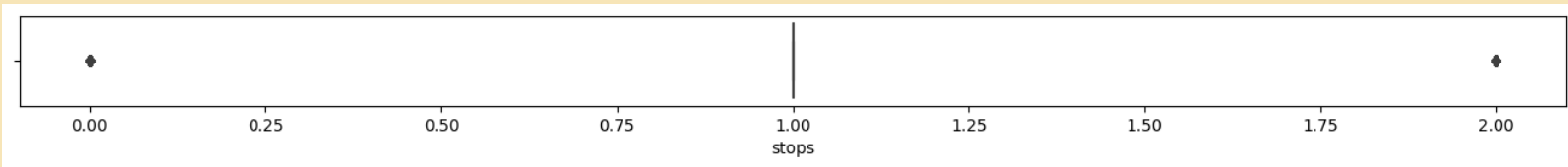
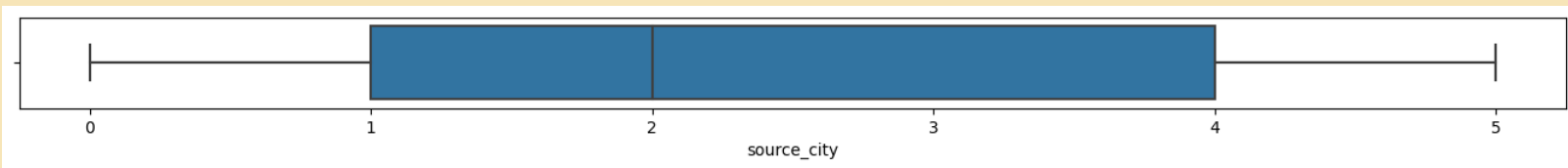


# Encoding

```
category_to_integer = {  
    'airline': {'SpiceJet': 0, 'AirAsia': 1, 'Vistara': 2, 'GO_FIRST': 3, 'Indigo': 4, 'Air_India': 5},  
    'source_city': {'Delhi': 0, 'Mumbai': 1, 'Bangalore': 2, 'Kolkata': 3, 'Hyderabad': 4, 'Chennai': 5},  
    'departure_time': {'Early_Morning': 0, 'Morning': 1, 'Afternoon': 2, 'Evening': 3, 'Night': 4, 'Late_Night': 5},  
    'stops': {'zero': 0, 'one': 1, 'two_or_more': 2},  
    'arrival_time': {'Early_Morning': 0, 'Morning': 1, 'Afternoon': 2, 'Evening': 3, 'Night': 4, 'Late_Night': 5},  
    'destination_city': {'Delhi': 0, 'Mumbai': 1, 'Bangalore': 2, 'Kolkata': 3, 'Hyderabad': 4, 'Chennai': 5},  
    'class': {'Economy': 0, 'Business': 1}  
}
```

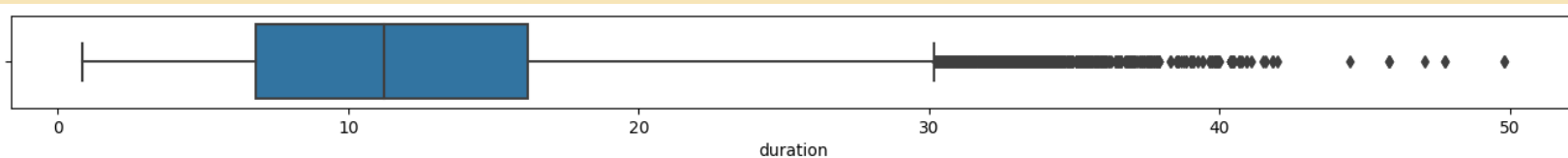
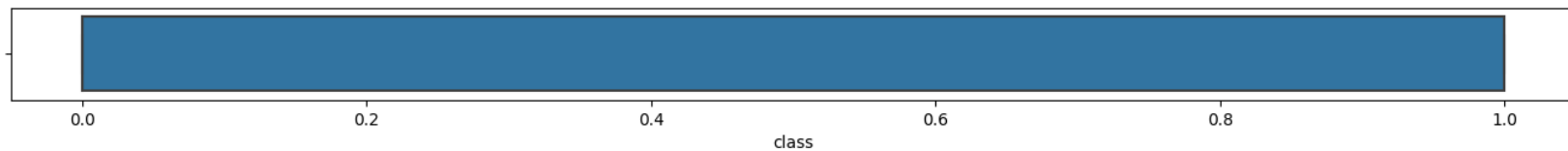
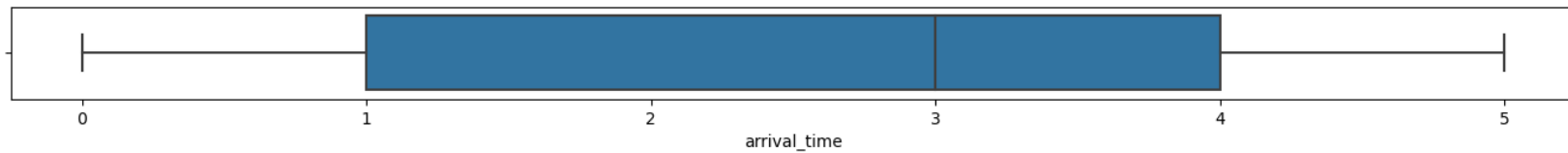
```
for column in df_update.select_dtypes(include=['object']).columns:  
    df_update[column + '_Kode'] = df_update[column].astype('category').cat.codes  
  
df_update.drop(df_update.select_dtypes(include=['object']).columns, axis=1, inplace=True)  
df_update.head()
```

# Outlier

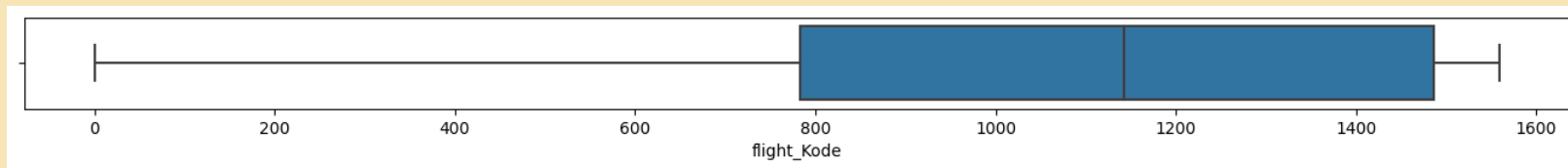
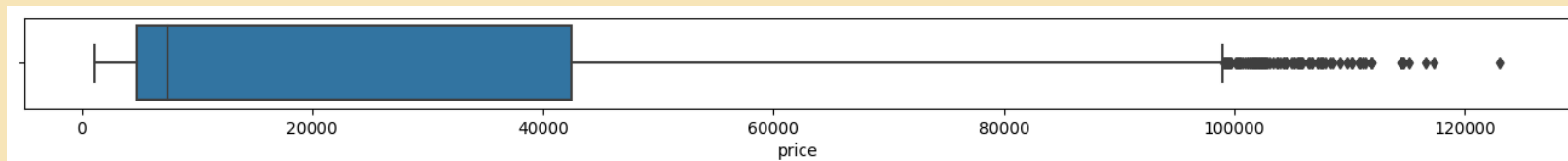
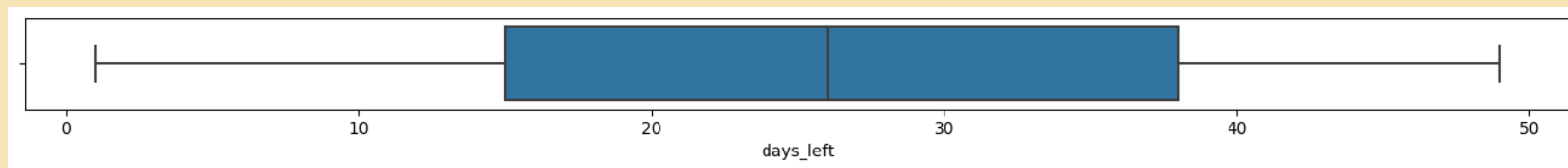




# Outlier

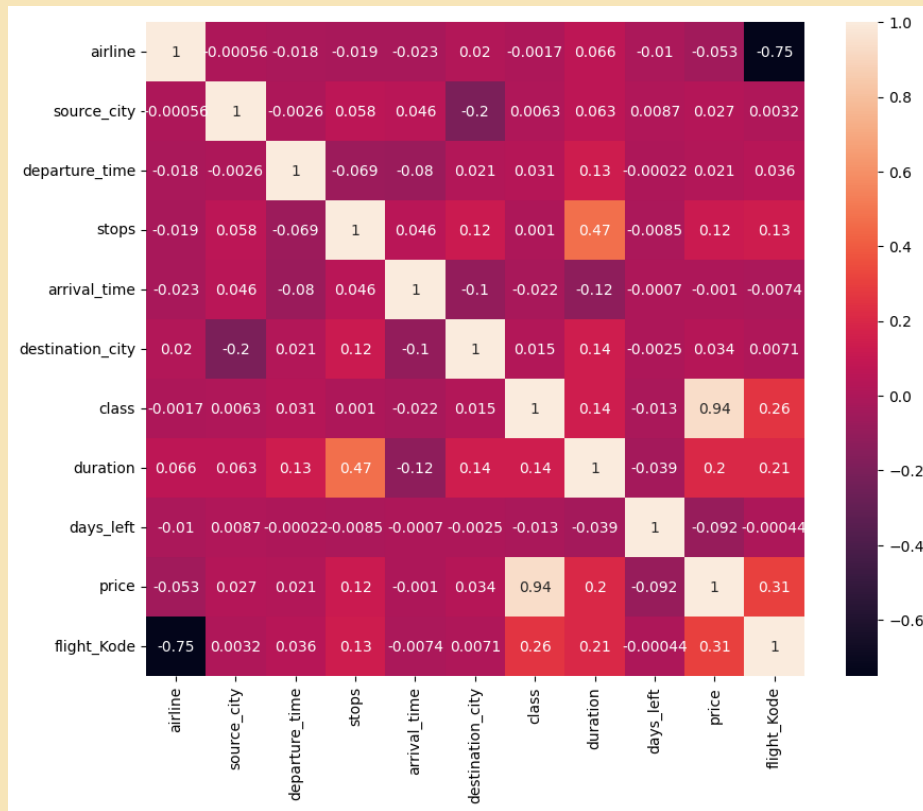


# Outlier



# Correlation

Features	VIF
airline	5.739721
source_city	2.762847
departure_time	2.744334
stops	8.333125
arrival_time	4.844704
destination_city	2.945661
class	1.640480
duration	5.858371
days_left	4.423780
flight_Kode	8.925775



# Correlation

Features	VIF
airline	5.739721
source_city	2.762847
departure_time	2.744334
<del>stops</del>	<del>8.333125</del>
arrival_time	4.844704
destination_city	2.945661
class	1.640480
duration	5.858371
days_left	4.423780
<del>flight_Kode</del>	<del>8.925775</del>



Features	VIF
airline	4.392737
source_city	2.610495
departure_time	2.590079
arrival_time	3.819817
destination_city	2.712053
class	1.465856
duration	3.891989
days_left	3.860021

# Modelling



# Model Test Without Preprocessing

Model	R2 Score	RMSE	MAPE
LinearRegression	0.9052034603234	6995.0620983632	0.4320129203721
Lasso	0.9052026963204	6995.0902862978	0.4318007585726
Ridge	0.9052033692586	6995.0654582108	0.4320109215858
KNeighborsRegre	0.4595352445617	16702.405413370	1.1274907291411
DecisionTreeRegr essor	0.9829334456433	2968.0297929576	0.0608707733374
RandomForestReg ressor	0.9897537249219	2299.7378731021	0.0592587313746
XGBRegressor	0.9835964699975	2909.8057018432	0.1330585952419



# Model Test After Preprocessing

Model	R2 Score	RMSE	MAPE
LinearRegression	0.8972397679329	7282.9590843774	0.3791949514061
Lasso	0.8972395984531	7282.9650901721	0.3791765398946
Ridge	0.8972397060080	7282.9612787880	0.3792056469452
KNeighborsRegre	0.7411215972939	11559.611597927	0.7231749204483
DecisionTreeRegr essor	0.9751404622362	3582.1342361619	0.0785676608484
RandomForestReg ressor	0.9845849844518	2820.7676577184	0.0759097713299
XGBRegressor	0.9762581726524	3500.6799174478	0.1537592994246

# Model

No	Feature	Coefficient
1	Class	0.879552
2	Duration	0.048554
3	Flight_kode	0.030251
4	Days_left	0.017813
5	Destination_city	0.010200
6	Source_city	0.005548
7	Stops	0.002564
8	Arrival_time	0.002424
9	Departure_time	0.002092
10	Airline	0.001001

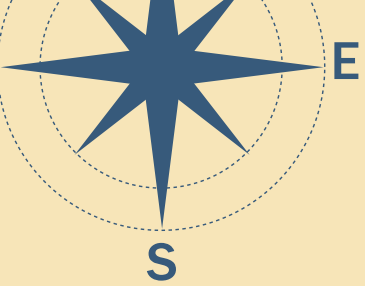


# Recommendation



# Recommendation

1. It is better to purchase flight tickets more than 20 days before the flight
2. Choosing late night and early morning flight times is also recommended
3. Based on modeling results, the determinant of the highest flight price is class and the comparison with other variables is very far. Meanwhile, the time of arrival or departure has an influence on the lowest flight price along with the type of airline.



# Thanks

**Do you have any questions?**

