THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Introduction to Data Mining Midterm Project

Domestic Indian Airlines

Group 3

Lokesh Bokkisam

Pavani Samala

Shumel Siraj

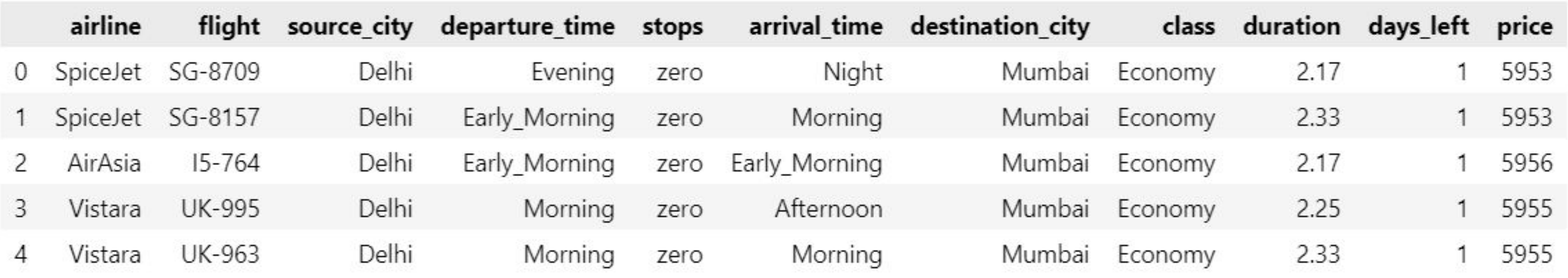
SMART Question

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?

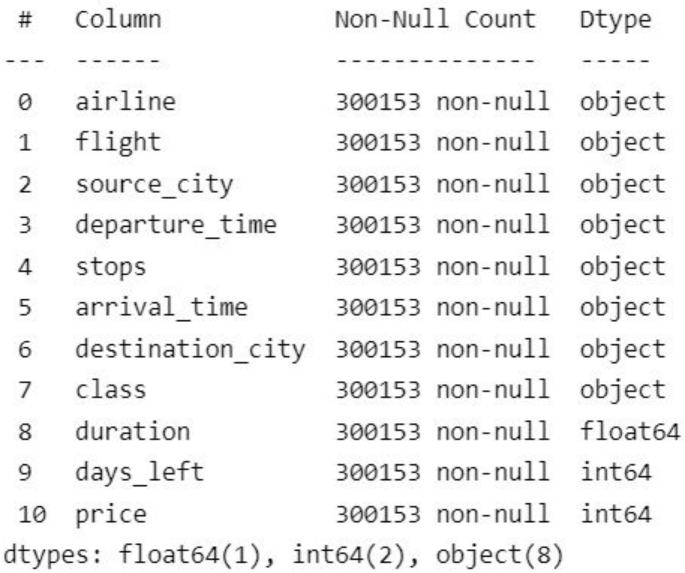
About the Dataset

- Domestic Airline that travel across India
- 30,015 observations/11 variables
- variables:
  - price
  - departure + arrival times
  - departure + arrival cities
  - number of stops
  - duration of flight
  - days left until take off
  - Airlines + flights number

Data: Before Preprocessing

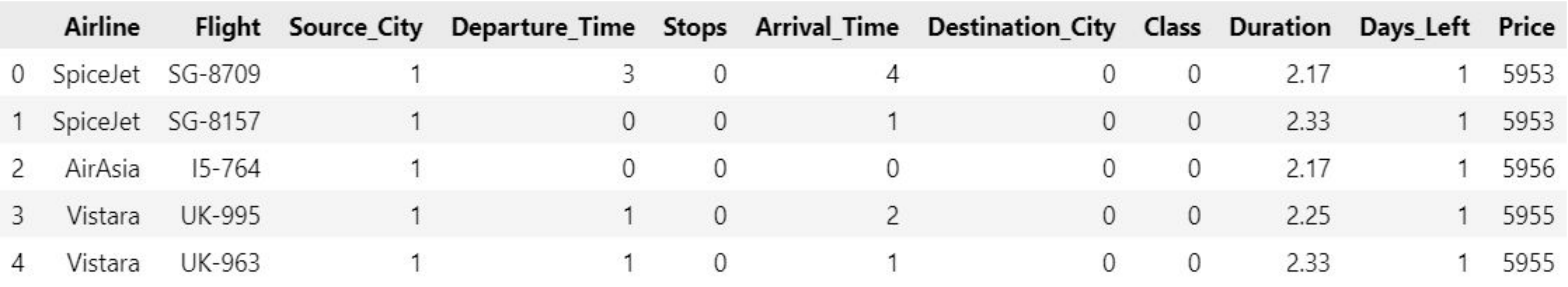


	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

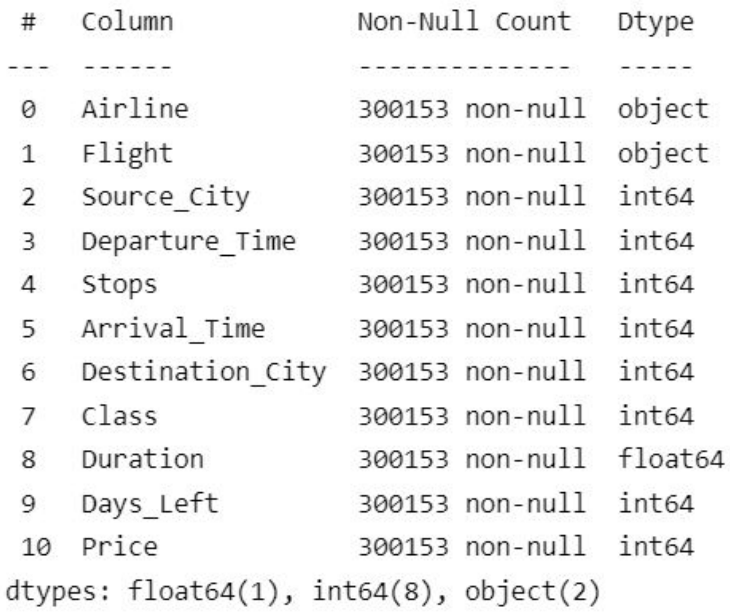


```
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   airline           300153 non-null  object
 1   flight            300153 non-null  object
 2   source_city       300153 non-null  object
 3   departure_time    300153 non-null  object
 4   stops             300153 non-null  object
 5   arrival_time      300153 non-null  object
 6   destination_city  300153 non-null  object
 7   class             300153 non-null  object
 8   duration          300153 non-null  float64
 9   days_left         300153 non-null  int64
 10  price             300153 non-null  int64
dtypes: float64(1), int64(2), object(8)
```

Data: After Preprocessing

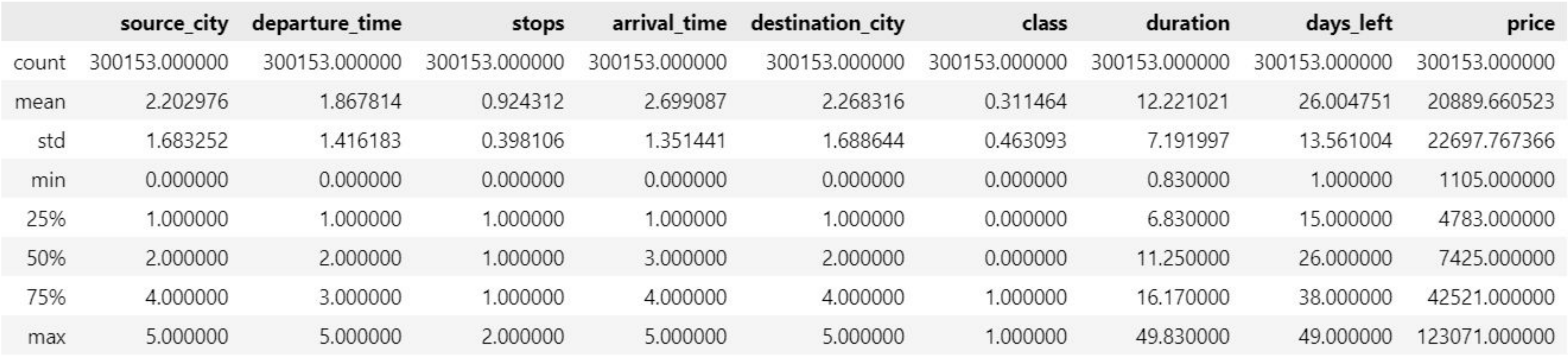


	Airline	Flight	Source_City	Departure_Time	Stops	Arrival_Time	Destination_City	Class	Duration	Days_Left	Price
0	SpiceJet	SG-8709	1	3	0	4	0	0	2.17	1	5953
1	SpiceJet	SG-8157	1	0	0	1	0	0	2.33	1	5953
2	AirAsia	I5-764	1	0	0	0	0	0	2.17	1	5956
3	Vistara	UK-995	1	1	0	2	0	0	2.25	1	5955
4	Vistara	UK-963	1	1	0	1	0	0	2.33	1	5955



```
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Airline           300153 non-null  object 
 1   Flight            300153 non-null  object 
 2   Source_City       300153 non-null  int64  
 3   Departure_Time    300153 non-null  int64  
 4   Stops             300153 non-null  int64  
 5   Arrival_Time      300153 non-null  int64  
 6   Destination_City  300153 non-null  int64  
 7   Class             300153 non-null  int64  
 8   Duration          300153 non-null  float64
 9   Days_Left         300153 non-null  int64  
 10  Price             300153 non-null  int64  
dtypes: float64(1), int64(8), object(2)
```

Summary of Dataset



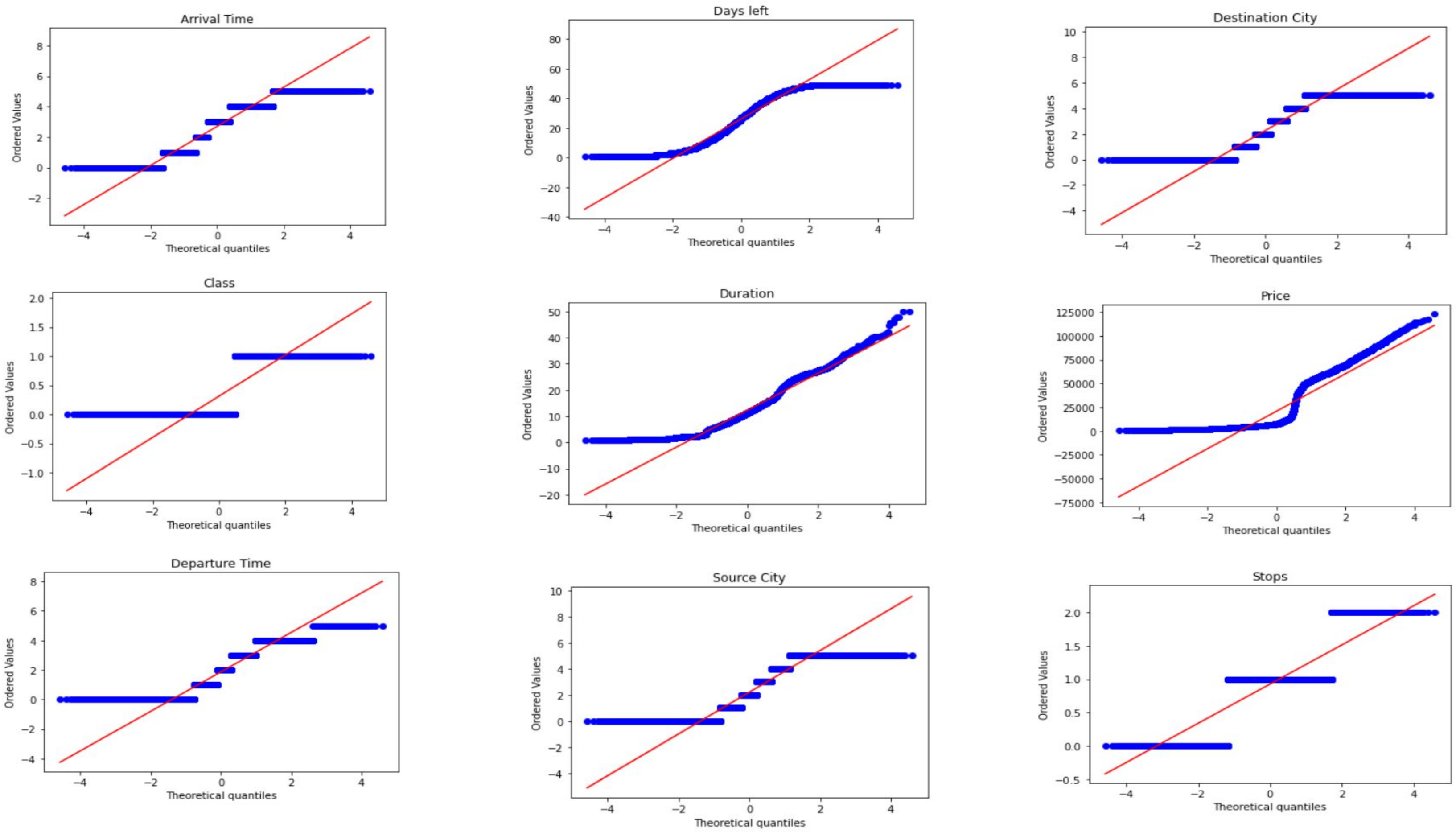
	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
count	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000	300153.000000
mean	2.202976	1.867814	0.924312	2.699087	2.268316	0.311464	12.221021	26.004751	20889.660523
std	1.683252	1.416183	0.398106	1.351441	1.688644	0.463093	7.191997	13.561004	22697.767366
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.830000	1.000000	1105.000000
25%	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	6.830000	15.000000	4783.000000
50%	2.000000	2.000000	1.000000	3.000000	2.000000	0.000000	11.250000	26.000000	7425.000000
75%	4.000000	3.000000	1.000000	4.000000	4.000000	1.000000	16.170000	38.000000	42521.000000
max	5.000000	5.000000	2.000000	5.000000	5.000000	1.000000	49.830000	49.000000	123071.000000

Normality Test

Shapiro-Wilk Test:

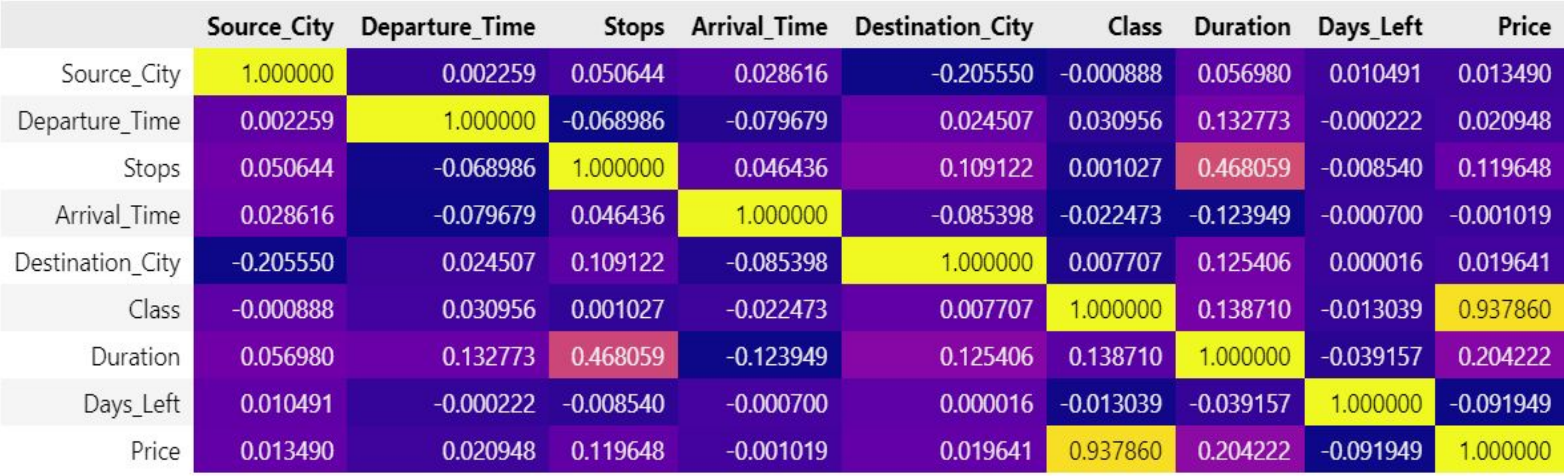
Variables	Statistics	P-Value
Source City	0.903	0.00
Departure Time	0.888	0.00
Stops	0.543	0.00
Arrival Time	0.896	0.00
Destination City	0.905	0.00
Class	0.583	0.00
Duration	0.956	0.00
Days Left	0.959	0.00
Price	0.752	0.00

QQ Plots


Arrival Time
Ordered Values
Theoretical quantiles
8
6
4
2
0
−2
−4
−2
0
2
4
Days left
Ordered Values
Theoretical quantiles
80
60
40
20
0
−20
−40
−4
−2
0
2
4
Destination City
Ordered Values
Theoretical quantiles
10
8
6
4
2
0
−2
−4
−4
−2
0
2
4
Class
Ordered Values
Theoretical quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−4
−2
0
2
4
Duration
Ordered Values
Theoretical quantiles
50
40
30
20
10
0
−10
−20
−4
−2
0
2
4
Price
Ordered Values
Theoretical quantiles
125000
100000
75000
50000
25000
0
−25000
−50000
−75000
−4
−2
0
2
4
Departure Time
Ordered Values
Theoretical quantiles
8
6
4
2
0
−2
−4
−4
−2
0
2
4
Source City
Ordered Values
Theoretical quantiles
10
8
6
4
2
0
−2
−4
−4
−2
0
2
4
Stops
Ordered Values
Theoretical quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−4
−2
0
2
4

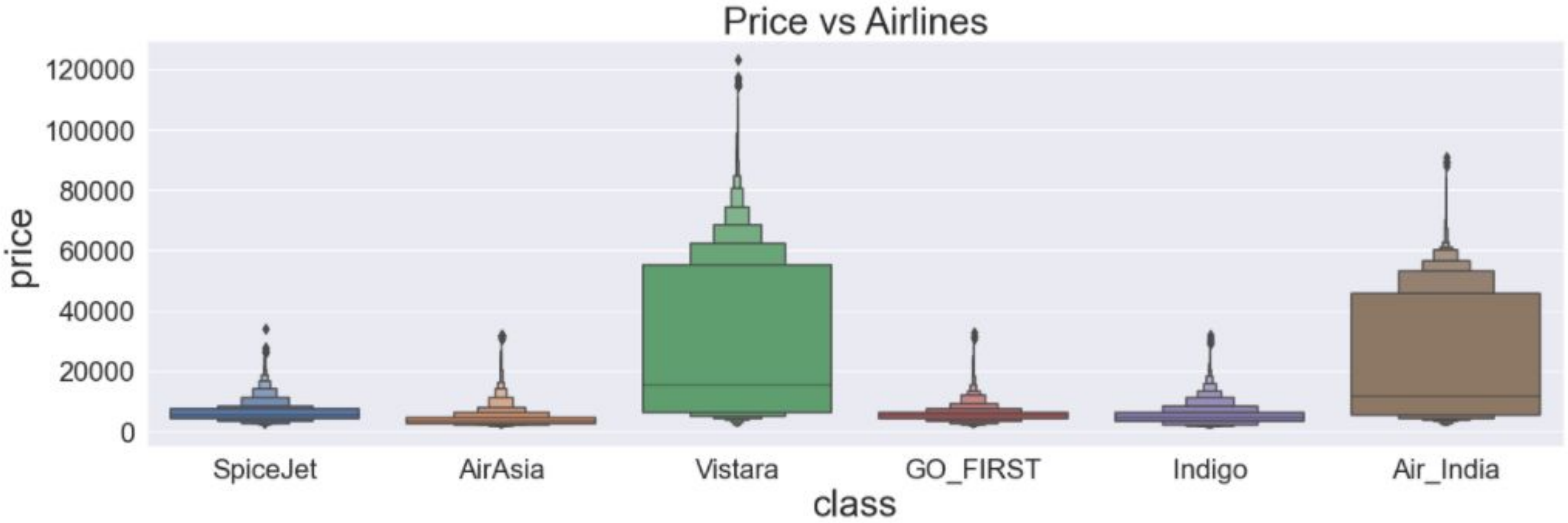
Exploratory Data Analysis

Correlation Matrix

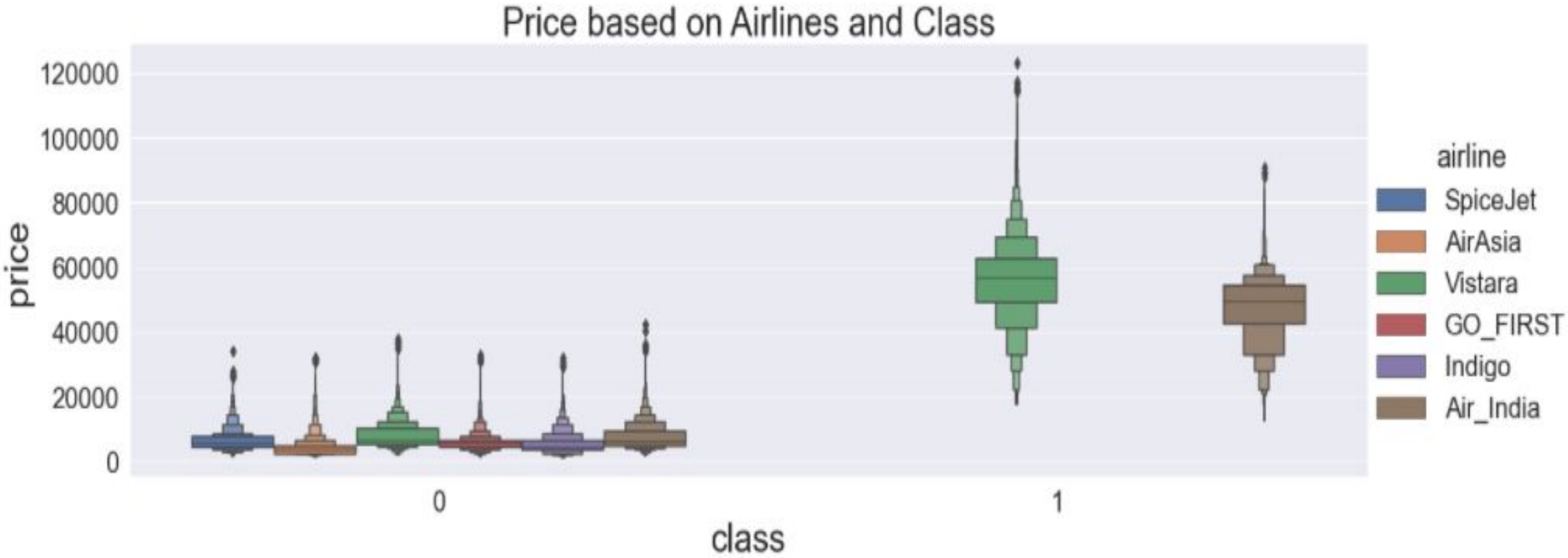


	Source_City	Departure_Time	Stops	Arrival_Time	Destination_City	Class	Duration	Days_Left	Price
Source_City	1.000000	0.002259	0.050644	0.028616	-0.205550	-0.000888	0.056980	0.010491	0.013490
Departure_Time	0.002259	1.000000	-0.068986	-0.079679	0.024507	0.030956	0.132773	-0.000222	0.020948
Stops	0.050644	-0.068986	1.000000	0.046436	0.109122	0.001027	0.468059	-0.008540	0.119648
Arrival_Time	0.028616	-0.079679	0.046436	1.000000	-0.085398	-0.022473	-0.123949	-0.000700	-0.001019
Destination_City	-0.205550	0.024507	0.109122	-0.085398	1.000000	0.007707	0.125406	0.000016	0.019641
Class	-0.000888	0.030956	0.001027	-0.022473	0.007707	1.000000	0.138710	-0.013039	0.937860
Duration	0.056980	0.132773	0.468059	-0.123949	0.125406	0.138710	1.000000	-0.039157	0.204222
Days_Left	0.010491	-0.000222	-0.008540	-0.000700	0.000016	-0.013039	-0.039157	1.000000	-0.091949
Price	0.013490	0.020948	0.119648	-0.001019	0.019641	0.937860	0.204222	-0.091949	1.000000

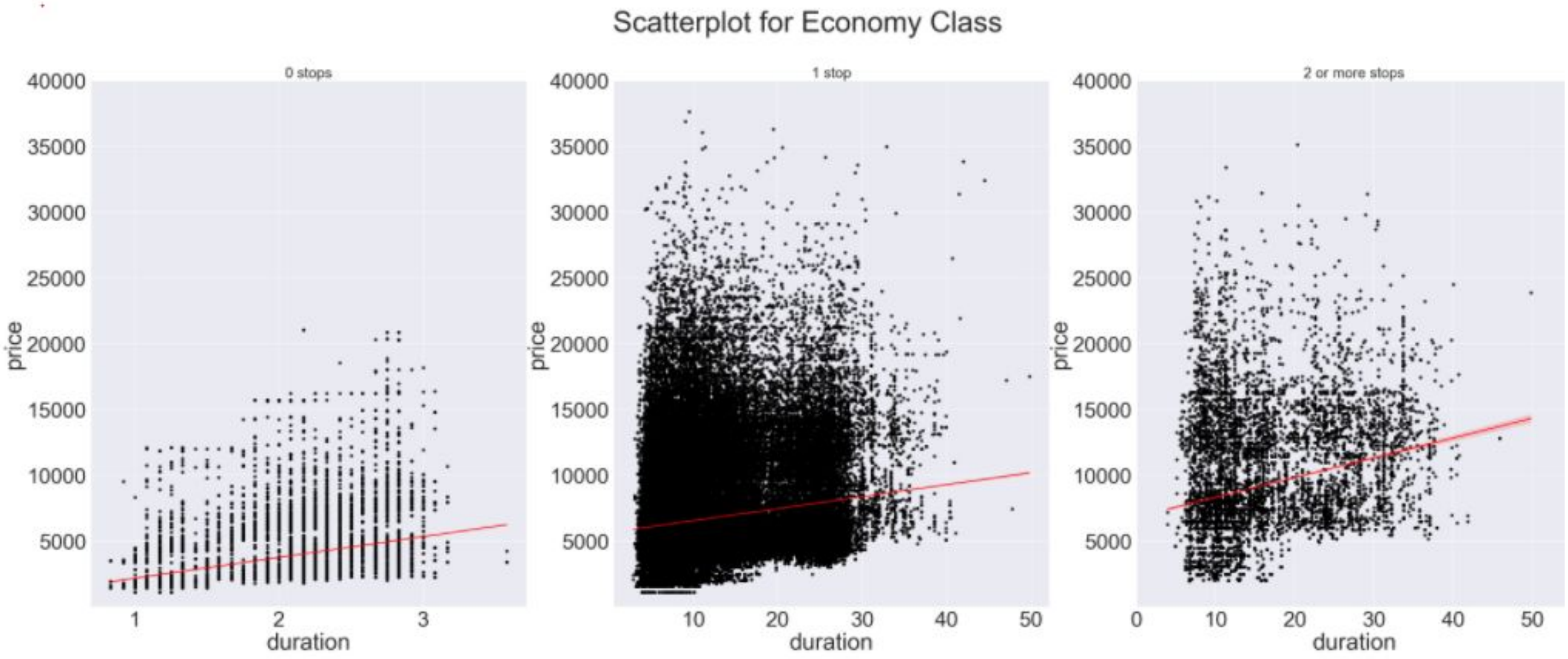
Price for Airlines



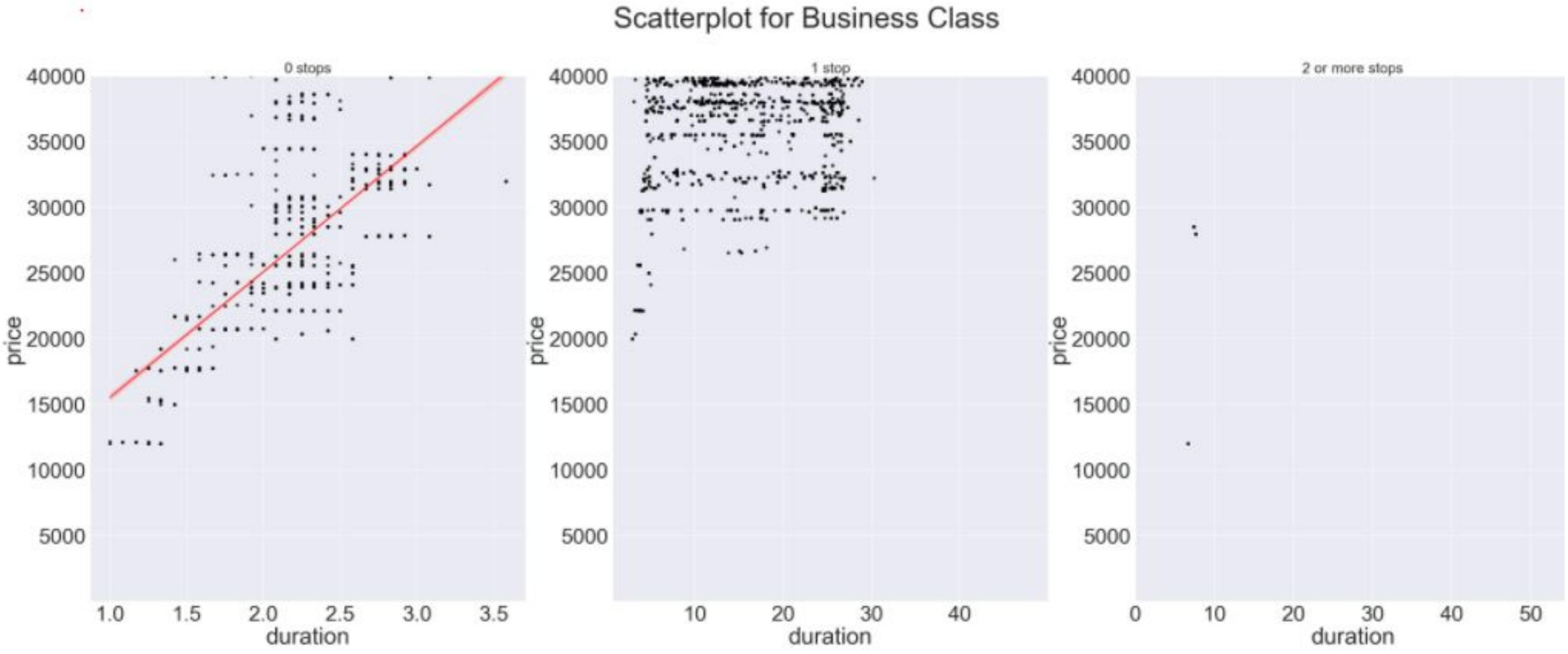
Price based on Airlines and Class



Price vs. Duration (Economy)



Price vs. Duration (Business)



Linear Regression

variables	economy		business	
	p-values	R^2	p-values	R^2
stops+duration+days_left	all<0.05	0.432	all<0.05	0.385
stops+duration+days_left+source city+destination city	source city 1, destination city 1,stops >0.05	0.437	source city 2 & 5, destination city 2, stops>0.05	0.432

Regression Trees

max_depth=8, min_samples_leaf =1, random_state=50

variables	economy		business	
	MSE	R^2	MSE	R^2
stops+duration+days_left	~5414276	0.59	~93013576	0.452
stops+duration+days_left+ source city+destination city	~5405555	0.6166	~85186687	0.497

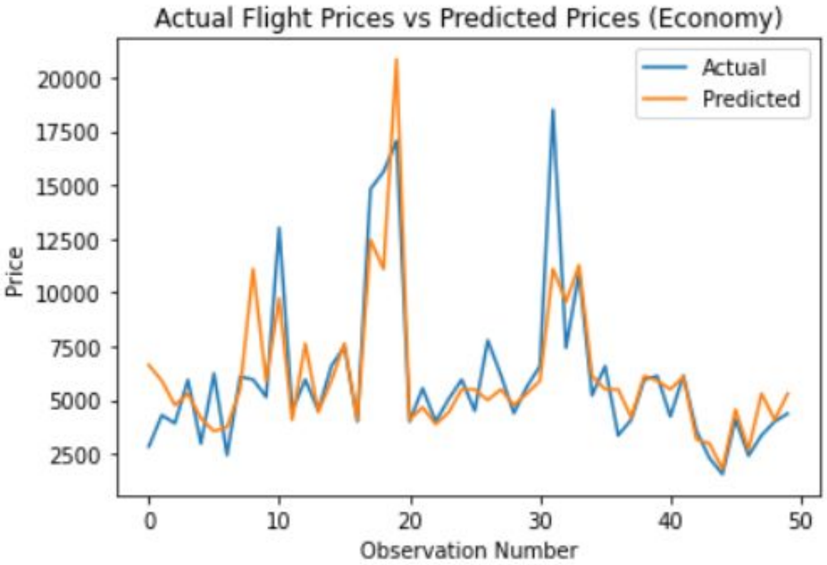
KNN

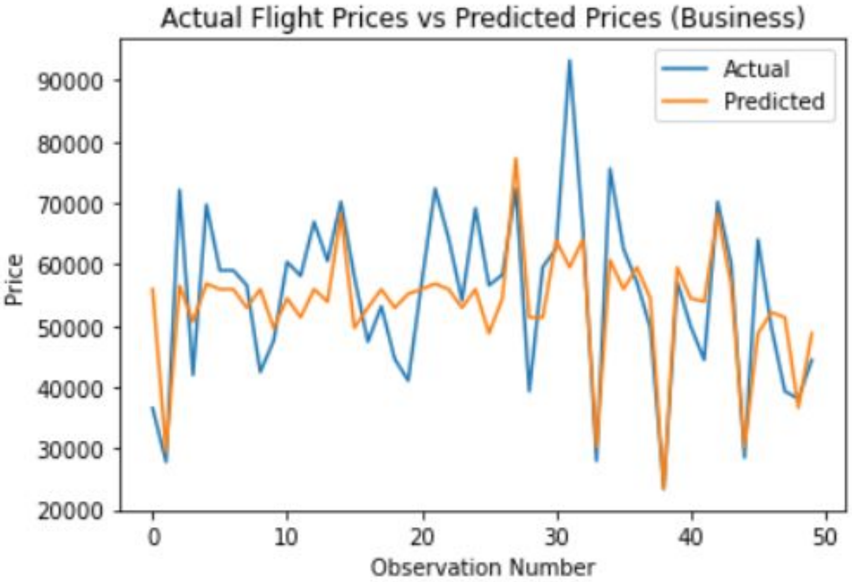
neighbor=100

variables	economy		business	
	score	R^2	score	R^2
stops+duration+days_left	0.034	0.216	0.0211	0.122
stops+duration+days_left+source city+destination city	0.178	0.329	0.195	0.212

Conclusion

- Better models when we included source city and destination city
- Regression Trees best predict price
  - explained 61.7% of economy class
  - explained 49.7% of business class





Looking Forward

- Neural Network or Gradient Boosted Model
- Find additional data on distance of flight and locations of stops, and different airport fees

References

- Flight Price-Data Analysis | Kaggle