

Flight Analytics: Insights to Reduce Carbon Footprint and Optimize Ticket Pricing

Yousuf Rajput, Tarek Tarif

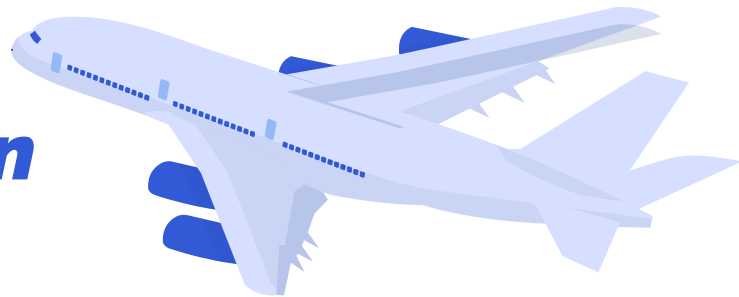




Table of Contents

1

***Problem Definition
and Data Description***

2

***Data Preparation
Process***

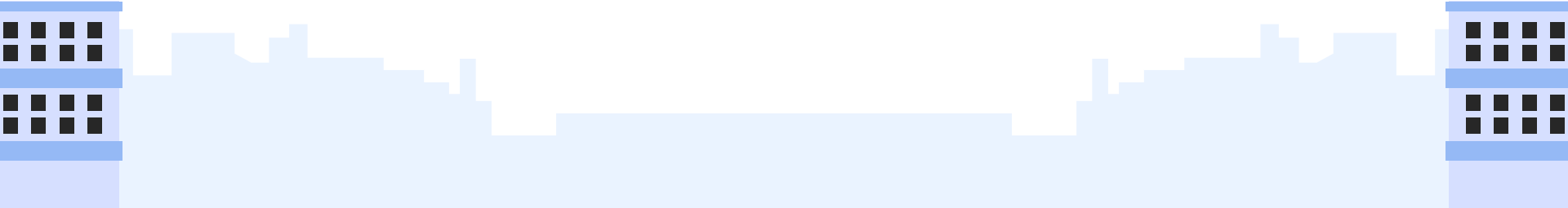


3

Data Analysis

4

Challenges & Insights



An illustration of an airport scene. At the top, there are three stylized blue clouds. In the upper left, a white airplane with blue accents is flying towards the right. In the center, a large orange circle contains a white number '1'. Below this, the text 'Problem Definition and Data Description' is written in a bold, blue, italicized font. At the bottom, there are two airport control towers on the left and right, and a light blue silhouette of a city skyline in the background.

1

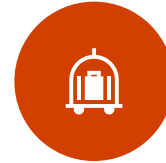
***Problem Definition and
Data Description***

Problem Definition



Definition:

- Problem: Identify if there exists correlations between features in regards to ticket price and origin country based on level of CO2 emission.
- Interests:
 - Interested in utilizing feature set to minimize ticket costs.
 - Find locations with greater carbon footprint



Applications:

- Analysis data could determine existence of correlation between features in dataset
- Reduce emission percentage by identifying airlines that produce a greater co2 emission in comparison with other aircraft
- Data could be made available to predict ticket prices based on selected features

Data Description



The data set consists of **one million** records of flight routes from the following 5 continents:

- Africa
- Asia
- Americas
- Europe

It also includes data related to the environmental impact of the flight:

- CO2 emissions of flight route
- Average CO2 emission for the given route
- CO2 percentage: % difference between CO2 emission and average CO2 emission for given route.

Data Description Cont'd

Columns

Column names and DTypes:

	Column	Type
0	from_country	string
1	dest_country	string
2	aircraft_type	string
3	airline_name	string
4	departure_time	string
5	arrival_time	string
6	duration	string
7	stops	string
8	price	string
9	co2_emissions	string
10	avg_co2_emission_for_this_route	string
11	co2_percentage	string

	0	1	2	3	4	5	6	7
summary	count	mean	stddev	min	25%	50%	75%	max
from_country	998866	None	None	Algeria	None	None	None	India
dest_country	998866	None	None	Algeria	None	None	None	Zurich
aircraft_type	984952	None	None	ATR 42/72 Airbus A320	None	None	None	Boeing 787
airline_name	998866	None	None	[9 Air]	None	None	None	[jetSMART] Wingo TUI Airways Wizz Air]
departure_time	998866	None	None	2022-04-30 00:25:00	None	None	None	2022-08-28 23:59:00
arrival_time	998866	None	None	2022-04-30 04:10:00	None	None	None	2022-09-02 06:55:00
duration	998866	1468.213762406569	705.7996771205948	100	973.0	1410.0	1880.0	999
stops	998866	1.6549256857276151	0.6524852194379935	0	1.0	2.0	2.0	6
price	997513	1763.370116479685	1985.9882093041824	100.00	621.0	1189.0	2127.0	9999.00
co2_emissions	993998	1111010.421550144	987689.1049172259	100000	522000.0	956000.0	1367000.0	9992000
avg_co2_emission_for_this_route	910464	862604.5499876986	522890.61033083015	100000	410000.0	876000.0	1184000.0	999000
co2_percentage	910464	None	None	-1%	None	None	None	None%

Summary Statistics

- Sample count of 998866
- Note string types for columns

Missing Counts in Data

from_country	0
dest_country	0
aircraft_type	13914
airline_name	0
departure_time	0
arrival_time	0
duration	0
stops	0
price	1353
co2_emissions	4868
avg_co2_emission_for_this_route	88402
co2_percentage	88402

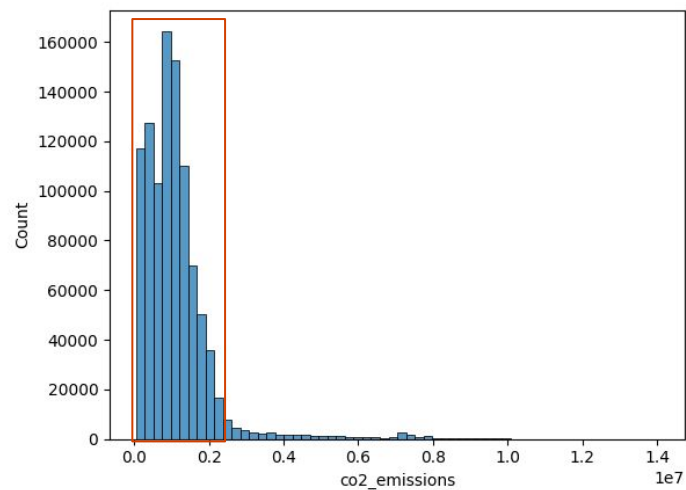
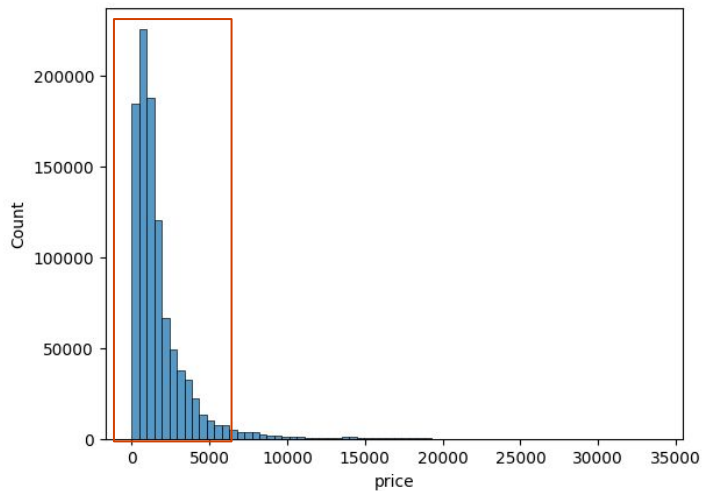
Data Description Cont'd

	0	1	2
from_country	Algeria	Algeria	Algeria
dest_country	Argentina	Argentina	Argentina
aircraft_type	Airbus A318 Canadair RJ 1000 Airbus A330 Airbu...	Airbus A318 Canadair RJ 1000 Boeing 787 Airbus...	Airbus A320 Airbus A321 Boeing 787 Airbus A320
airline_name	[Air France Iberia LATAM]	[Air France Iberia LATAM]	[Air France LATAM]
departure_time	2022-04-30 14:30:00	2022-04-30 14:30:00	2022-04-30 12:45:00
arrival_time	2022-05-01 10:15:00	2022-05-01 10:15:00	2022-05-01 10:15:00
duration	1425	1425	1530
stops	3	3	3
price	1279.00	1279.00	1284.00
co2_emissions	1320000	1195000	1248000
avg_co2_emission_for_this_route	1320000	1320000	1320000
co2_percentage	0%	-9%	-5%

Sample of Data

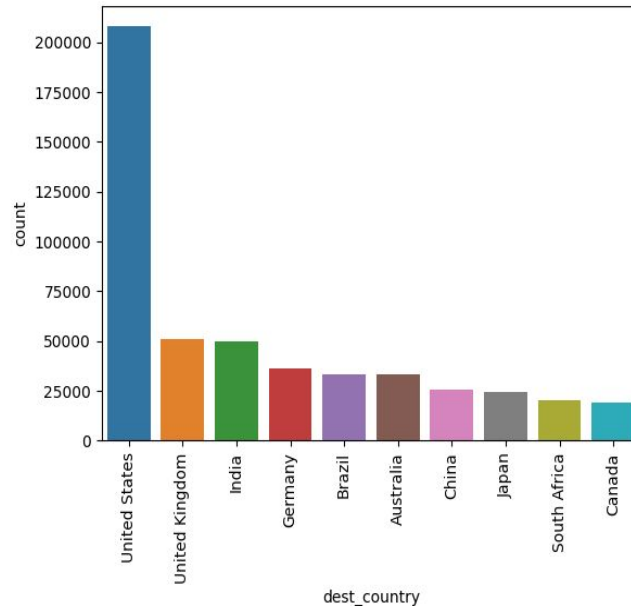
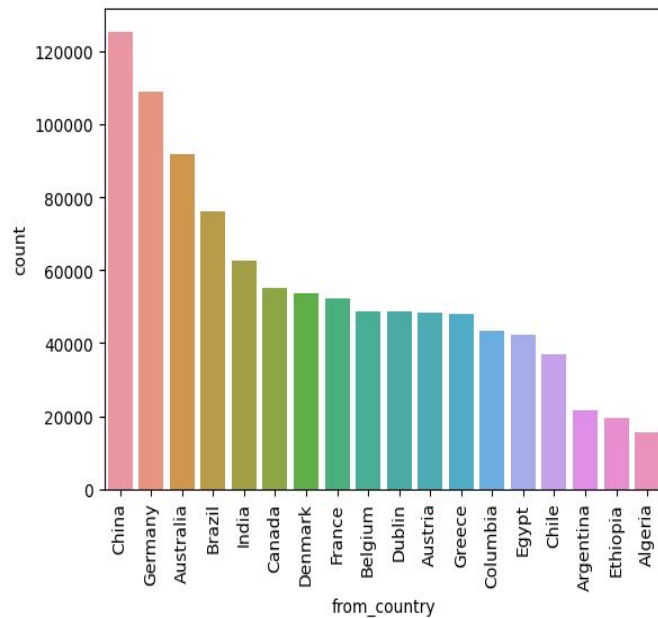


Data Description Cont'd



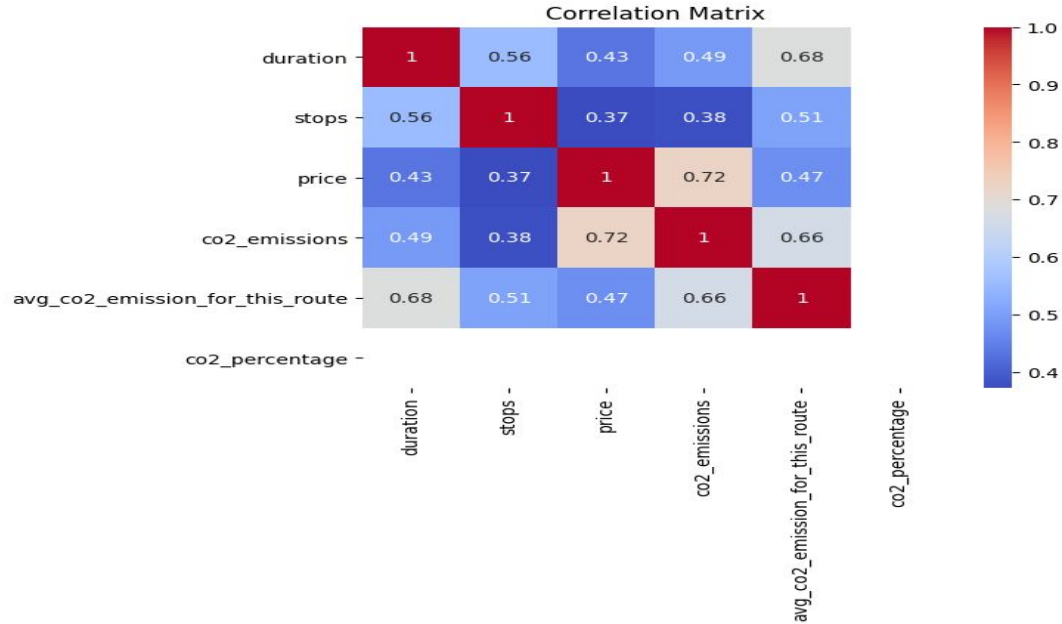
Primary Targets

Data Description Cont'd



Flight Arrivals/Destinations

Data Description Cont'd



Feature Correlation Heatmap



Data Quality Issues

- Airport codes provided are of IATA standard which can be not unique in some cases
- Ticket Prices are inconsistent: There exists scenarios where price is low to negative because of the use of points, vouchers, etc.
- Route duration: Duration of flight does not strictly adhere to in-flight time (layover time included)
- Flight seat categories do not exist, we cannot identify differences in flight cost based upon this factor. Dataset reflects this upon evaluating dataset for outliers
- Data consists of total flight paths with almost 85% of the data containing some number of stops. This leads to inconsistencies with several features such as `airline_name` and `aircraft_type`

IATA - International Air Transport Association

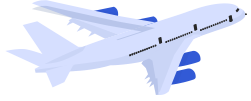


Data Preparation Process

2



Data Cleaning Steps



Dropping Columns

Dropping unnecessary features



Partitioning the Data by Price Range

Create models for split up price range



Excluding Outliers

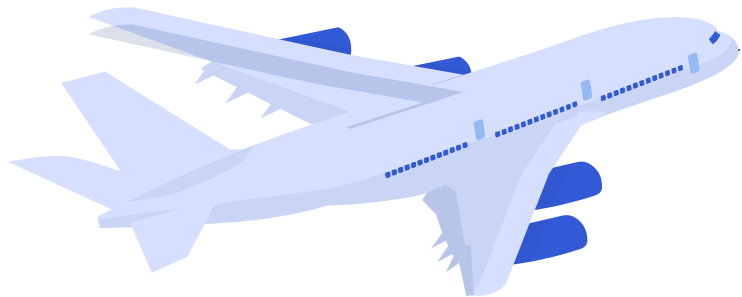
Exclude rows with a price exceeding 15,000 USD and flights with more than 3 stops



Convert Features to Proper DType

Convert features to proper dtype from 'string'

Features Used



from_country

Flight Origin

airline_name

Name of the airline

duration

Total duration of flight

stops

of stops in route

price

Price in USD of flight route(Ticketed)

CO2_emissions, avg_CO2_emissions

CO2 Emission of Flight, Average CO2 of Flight Path

Train/Validation/Test Datasets



Data Partitioning

Used indexers and VectorAssembler from PySpark.ml's feature library to provide labels for column data, then used randomSplit to partition each data set into train and test data.

```
(training_data, testing_data) = data.randomSplit([0.7, 0.3], seed=42)
```

Data Analysis

3



Analysis Task



Ticket Prices

Perform regression analysis w/ various models on price considering the following features:

- Airline Name
- Duration
- Stops
- Price



CO2 Emissions

Implement classification model to predict the origin country by using the following features:

- CO2 percentage
- average CO2 consumption
- destination country

Analysis Approaches



Regression Models

Explore 4 different regression models and determine effectiveness of each model:

- Linear Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient-Boosted Regression



Classification Model

Utilize a logistic regression model for multiclass classification using 3 data sets:

- Price < 5,000
- Price < 10,000
- Price < 15,000

Analysis Inputs



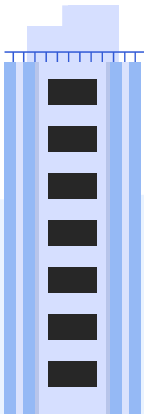
Regression Models

- Cleaned Data set
- Cleaned Data Set with 3 Price Filters
 - 5000, 10000, 15000

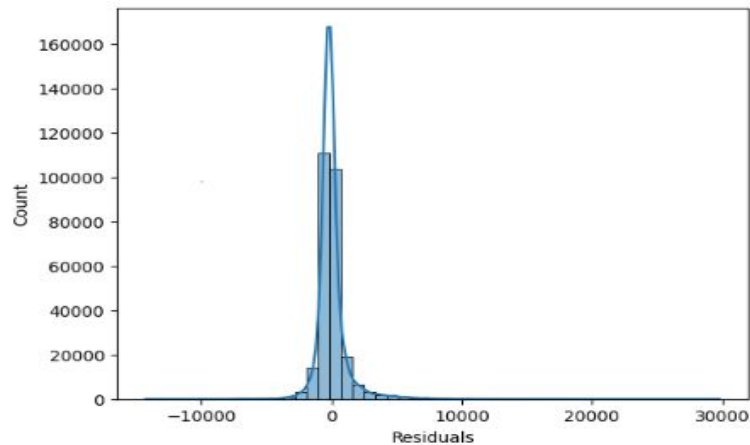


Classification Model

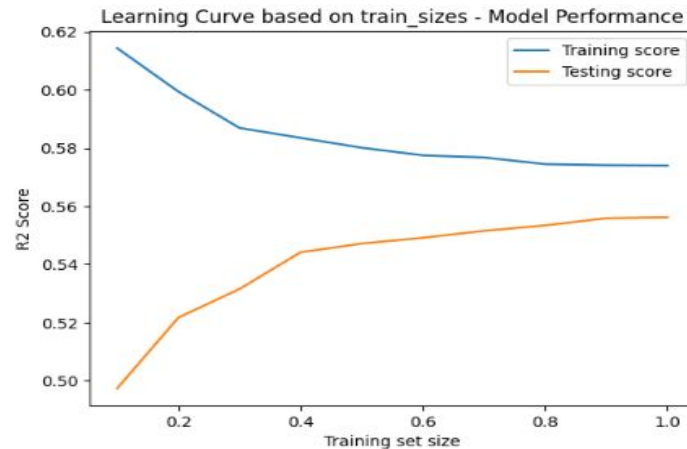
- Cleaned Data Set with 3 Price Filters
 - 5000, 10000, 15000



Analysis Results - Regression Models

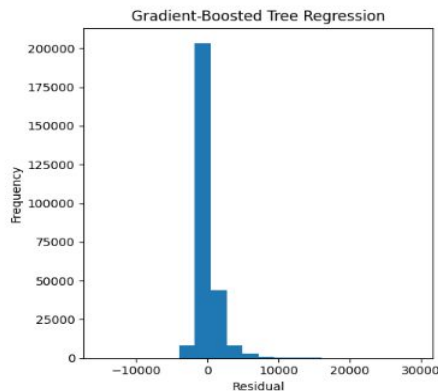
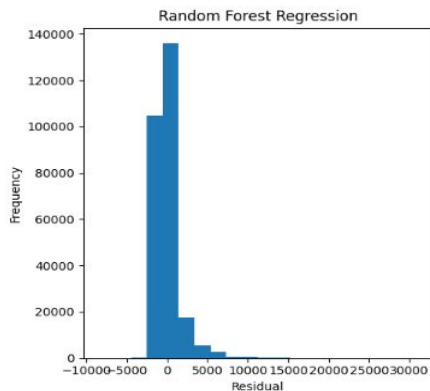
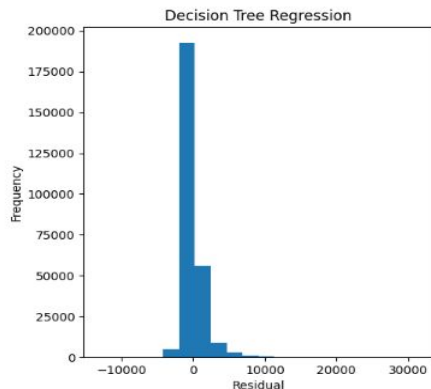
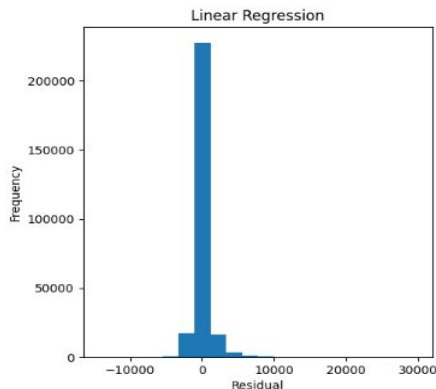


	Metric	Value
0	R2 Score	5.562113e-01

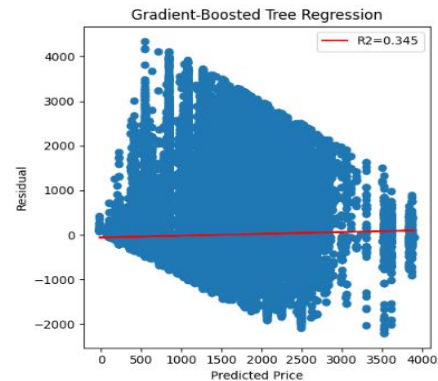
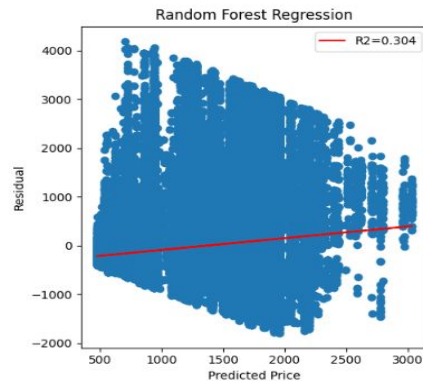
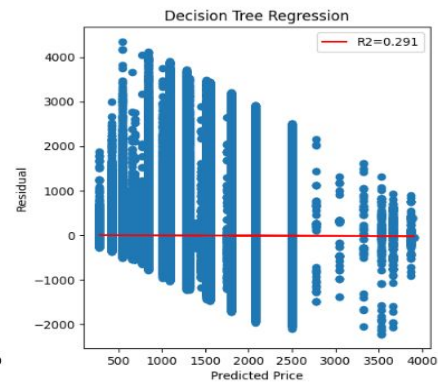
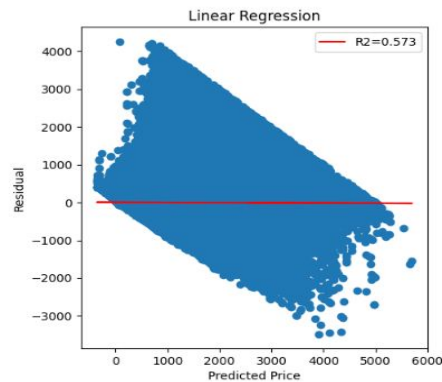
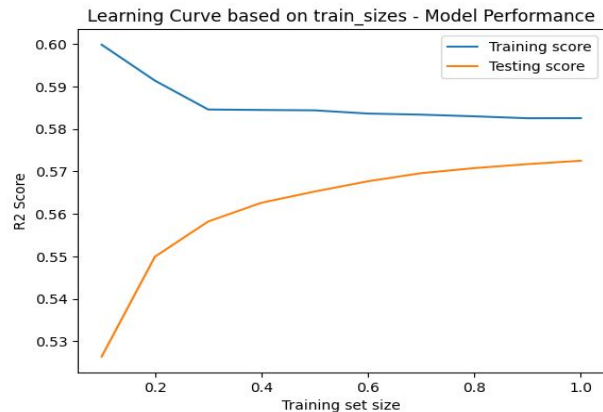


Analysis Results - Regression Models

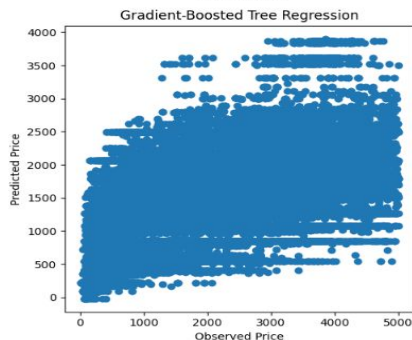
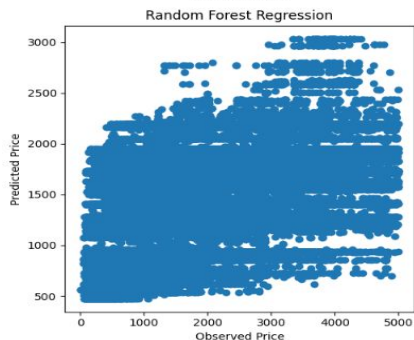
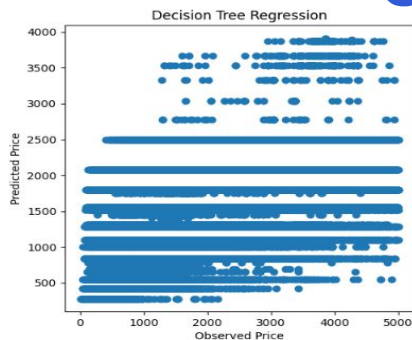
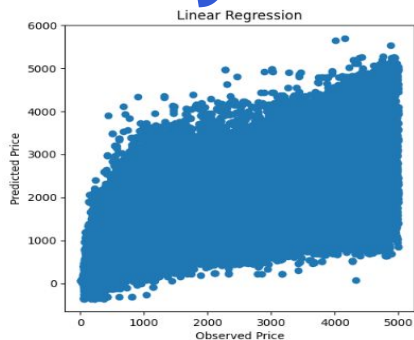
	Model	R2 Score
0	Linear Regression	0.556211
1	Decision Tree Regression	0.237717
2	Random Forest Regression	0.264066
3	Gradient-Boosted Tree Regression	0.316825



Analysis Results - Regression Models < 5k



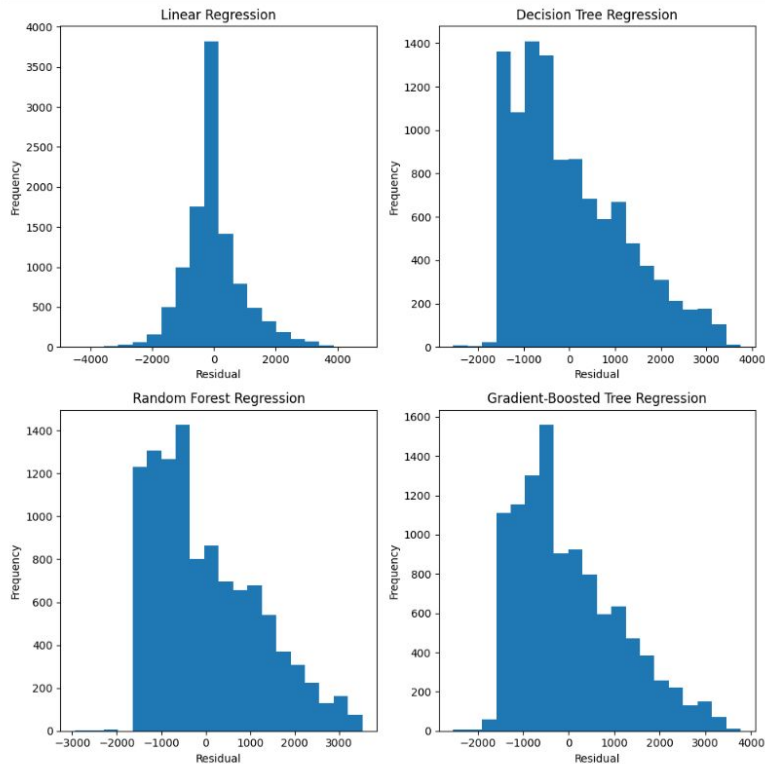
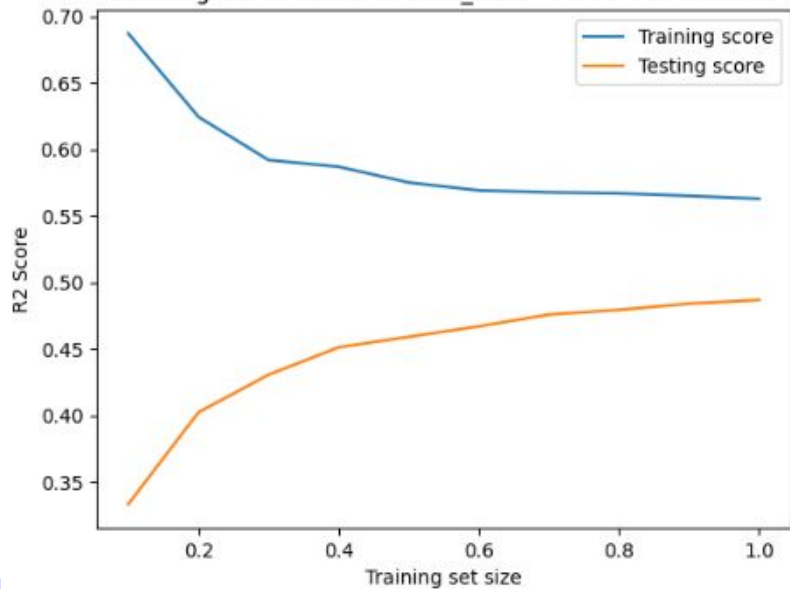
Analysis Results - Regression Models < 5k



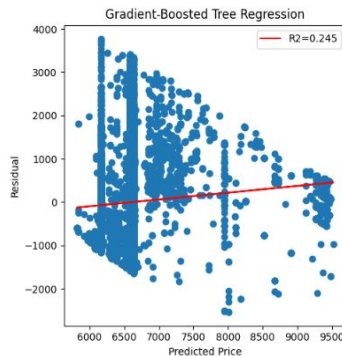
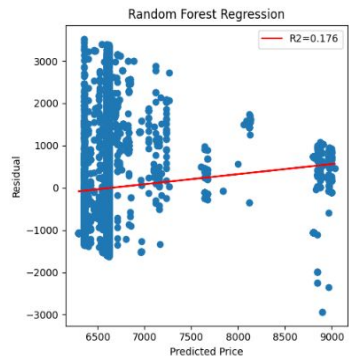
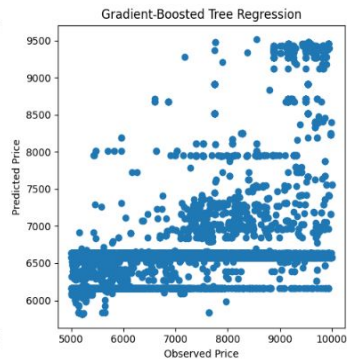
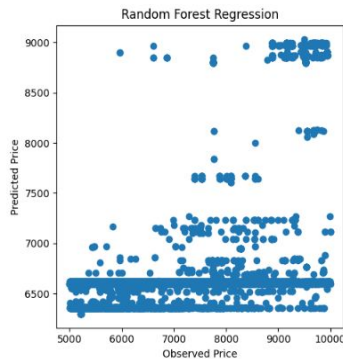
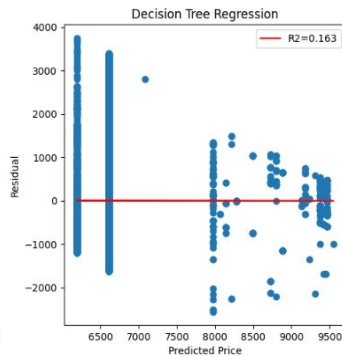
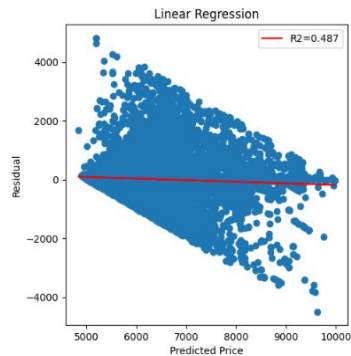
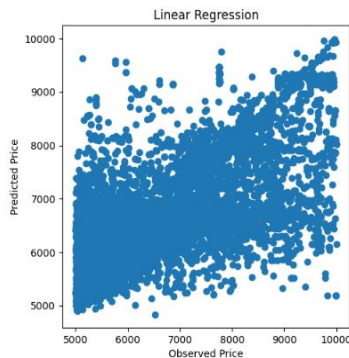
	Model	R2 Score
0	Linear Regression	0.572504
1	Decision Tree Regression	0.291260
2	Random Forest Regression	0.304101
3	Gradient-Boosted Tree Regression	0.344931

Analysis Results - Regression Models < 10k

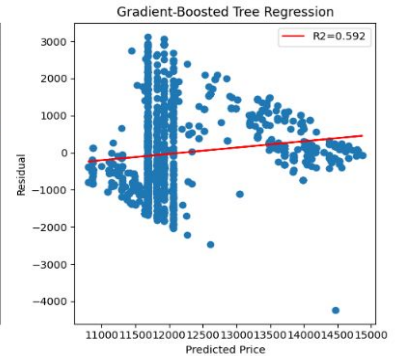
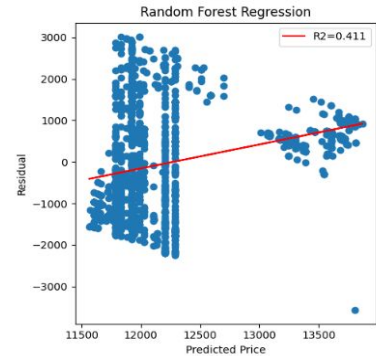
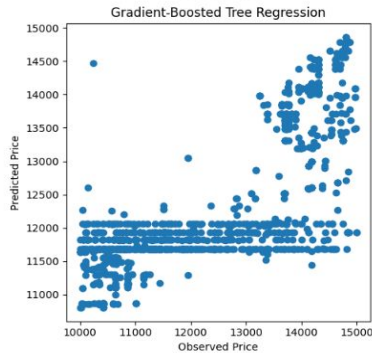
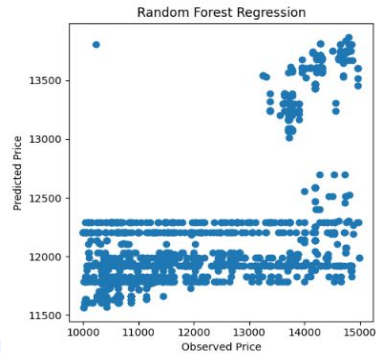
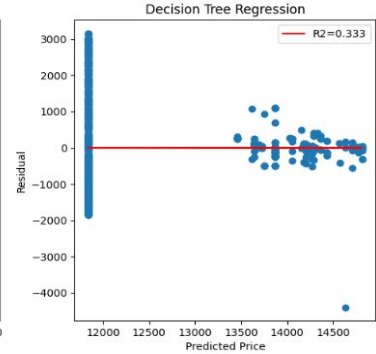
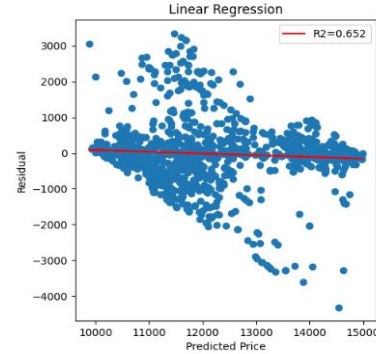
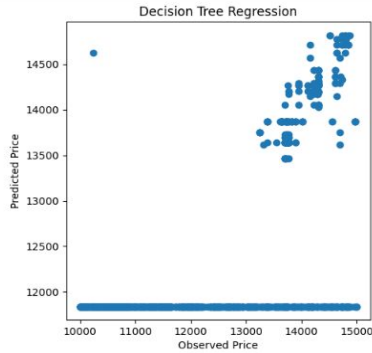
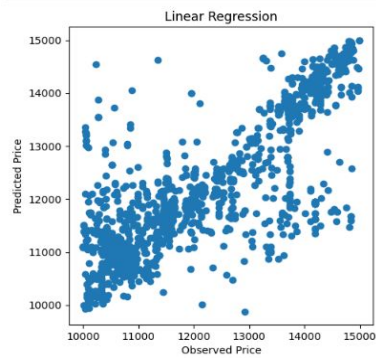
Learning Curve based on train_sizes - Model Performance



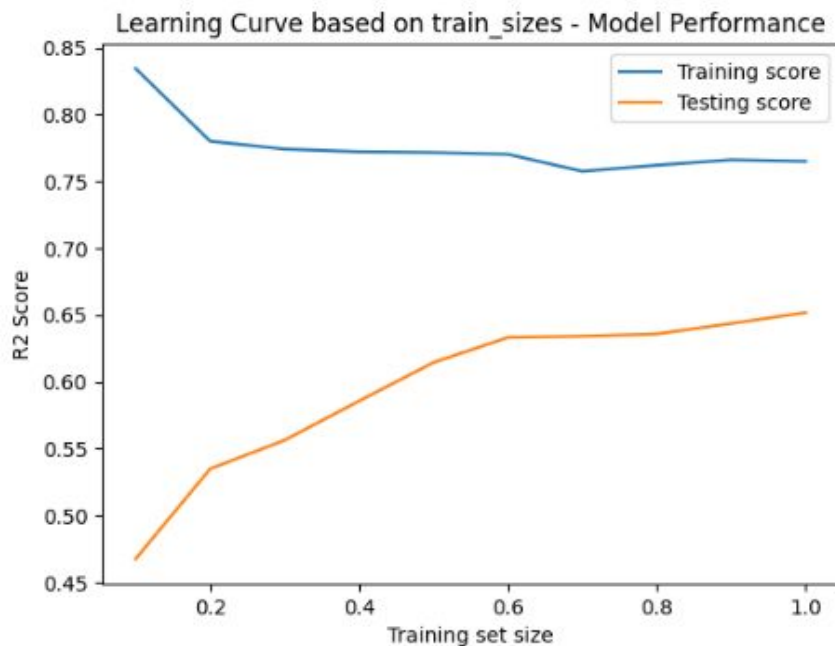
Analysis Results - Regression Models < 10k



Analysis Results - Regression Models < 15k

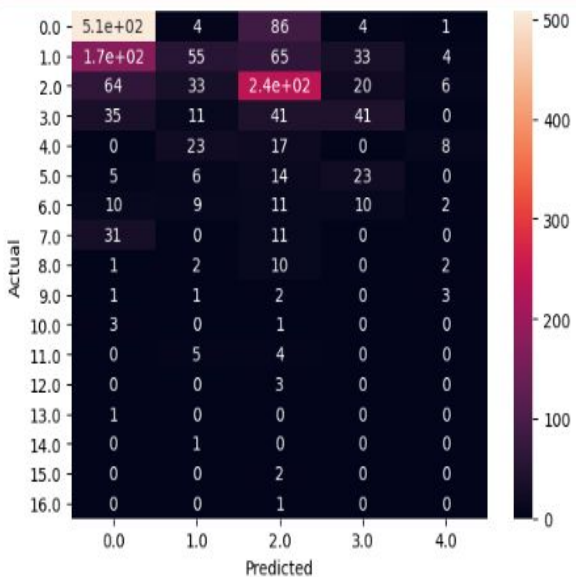


Analysis Results - Regression Models < 15k



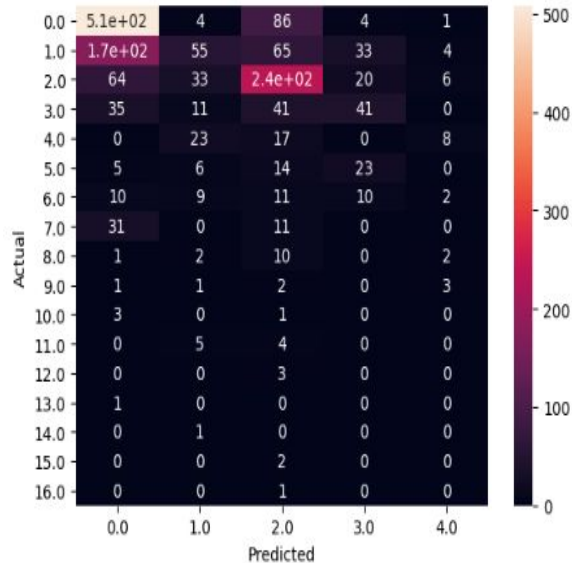
	Model	R2 Score
0	Linear Regression	0.651823
1	Decision Tree Regression	0.332864
2	Random Forest Regression	0.411044
3	Gradient-Boosted Tree Regression	0.592450

Analysis Results - Logistic Class. Model



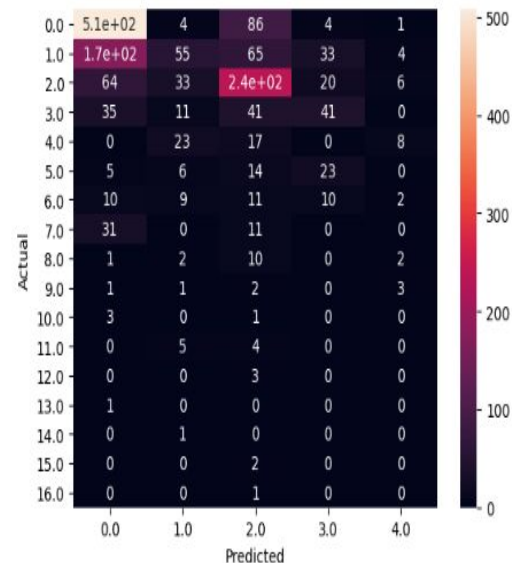
Test Accuracy: .19

5,000



Test accuracy: 0.35

10,000



Test accuracy: 0.52

15,000

Model Accuracy by DataFrames



Analysis Results Discussion



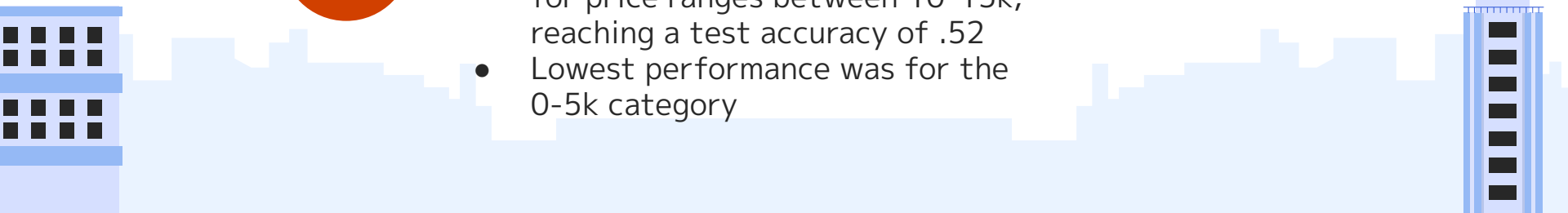


Regression Models

- Determined Linear Regression was the most accurate model for our selected target and features(for all data partitions)
- Model performance was highest for price ranges between 10-15k and lowest for 5-10k



Classification Model

- Model performance was vastly higher for price ranges between 10-15k, reaching a test accuracy of .52
 - Lowest performance was for the 0-5k category
- 
- 
- 

Insights Gained



Data Quality Issues

Price ranges are undeterminable in cases of seat selection(first class vs economy) and time of purchase causing instability in some of our models



Linear Regression

Linear Regression was our best candidate determined by evaluation of R2 scores and histograms in comparison with our different data frames



Model Similarities

Our strongest feature correlations occurred with prices above \$10,000. This can be attributed to the fact that these prices are more indicative of similar aircraft types and/or cabin seat selection

Challenges & Future Insights

4





Challenges



Stops

Variation in # of stops



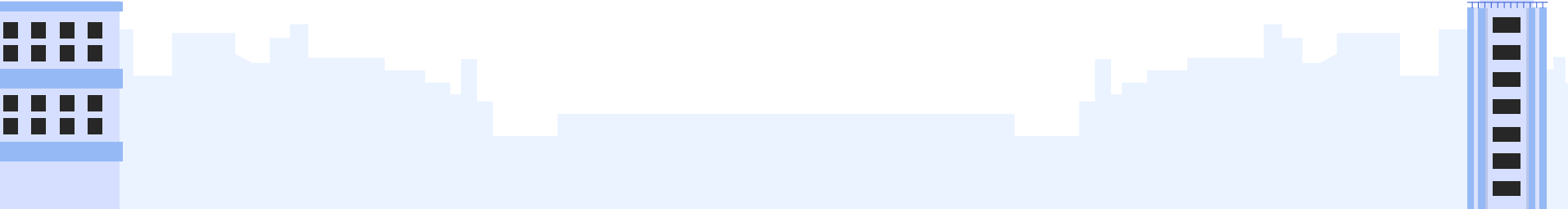
Isolating data to specific Airlines

Multiple airlines,
multiple aircraft types in
majority of rows in
dataset



Predicting/Clustering

CO₂ for aircrafts
and airlines cannot
be specified unless
data is subjected to
using 0 stops.





Solutions



Limitation of Zero Stops

Allows data isolation to specific airline/aircraft but would eliminate majority of data



Stops

Specifying analysis to specific stops, although layover time varies - conducting this allows for higher correlation between locations



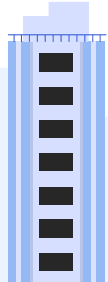
Using diff Features/Targets

Minimally using features that are reliant on number of stops, or congested data(name/type)



Price

Difference in pricing between the 3 classes of flights manipulates our data. Also time of purchase is not available.



Future Work



Select Better Dataset

Finding a dataset with less troubled data/features for a more accurate model



New Features

Create or find new features to test regression methods (distance)