IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tamas Priksz
01/04/2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies

In this project we have used different techniques to collect the necessary data (API, webscraping) and cleaned and restructured so it could be used in the predictive model.

We applied some nice visuals to identify the connections between different paramters and make the steps and result more end user friendly.

- Summary of all results

Interactive dashboard and map visualisation with the extension of predictive models can help us better predict what will be the outcome of a future mission

Introduction

- Project background and context

We will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

- Problems you want to find answers
  - Identify what are the key circumstances when the Falcon 9 will have a successful landing
  - To do so we need to collect, clean and investigate the relevant data
  - Then buld a prediction model which can calculate the possibilities of different scenarios

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using two different ways
    - Spacexdata.com API calls to pull the launches realted data
    - Using webscraping Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - Both cases we needed to apply data cleaning, data restructing, filtering for the relevant data subets
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.
  - The data was collected using two different ways
    - Spacexdata.com API calls to pull the launches realted data
    - Using webscraping Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - Both cases we needed to apply data cleaning, extracting the right data from the available resources, restructure them, creating new meaningful column names and appropriate data subset for further analytics

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Github link: https://github.com/tp030423/edx/blob/395015f28ed16b796583ca8864d9917cc74a412e/github_spacex-data-collection-api.ipynb

1. Finding the right data source
2. Set up needed cleaning functions
3. Run the API call, create dataframe from the response text
4. Clean up datatable, removing unnecessary columns and creating new, relevant ones
5. Filtering the dataframe to Falcon 9 entries
6. Replacing missing values with column mean

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Github link:
- https://github.com/tp030423/edx/blob/395015f28ed16b796583ca8864d9917cc74a412e/github_webscraping.ipynb

1. Finding the right data source (wiki pageÖ
2. Set up needed cleaning functions
3. Run the webscraping to create dataframe from the response text
4. Clean up datatable, removing unnecessary columns and creating new, relevant ones
5. Creating the final dataframe layout, structure

Data Wrangling

- Describe how data were processed
  - Appropriate datasource was defined
  - Missingness check in each column was applied
  - Checked the main categories of the key columns (launch site, orbit, outcome)
  - Potential outcomes were found and identified
  - Final, classification variable was defined based on the outcome information
  - Success rate was defined, it is 0.66, so in the two thirds the flights will be successful
- Github link: https://github.com/tp030423/edx/blob/395015f28ed16b796583ca8864d9917cc74a412e/github_webscraping.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - We needed to identify what are those features which influences a successful mission
  - So we created scatterplot to visualize the potential connection between:
    - Launch site and successful/unsuccessful missions
    - Connection between the payload and the lauch site
    - How orbit type determines a successful mission
    - The connection between payload and orbit type
    - Overall yearly trend: what was the distribution successful/unsuccessful missions over the years
- GitHub link

https://github.com/tp030423/edx/blob/72d78668c078f0c29083385ae3723095763cdf25/github-eda-dataviz.ipynb

EDA with SQL

- Checking the names of the launch sites
- Checking the payload carried by different boosters to compare options
- Checking the first successful landing in ground pad
- Checking successful booster types depending on the payload volume
- Comparing the number of the sucessful and unsuccessful missions
- Identify the successful booster types with the highest possible payload

- Github link:

https://github.com/tp030423/edx/blob/a94c28153b436a81bf806f0737894806fc8c181c/github_eda_sql.ipynb

Build an Interactive Map with Folium

- There were 3 task to do in this step:
  - Mark all launch sites on a map -> for these i used dots to display the launch sites on the map
  - Mark the success/failed launches for each site on the map -> markers with color codes are added to each launch site to identify the success ratio of the launches per site
  - Calculate the distances between a launch site to its proximities -> additonal feature is added so the user can calculate the proximities of different points from a certain launch site
- Github link:

https://github.com/tp030423/edx/blob/52bf7cd47a7b0ec55c06773571b69a99daf1f813/gitub-folium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Pie chart and scatter plots were created
  - Drop down and range slider features were added so each chart can be dynamically filterable
- Explaination
  - This interactiev dashboard is capable to answers questions regarding launch success rate, payload and successful mission connection.
- Github link:

https://github.com/tp030423/edx/blob/e07b9dd7128605ed0d0fbd0b38fe918453a13947/github_plotly_dashboard.ipynb

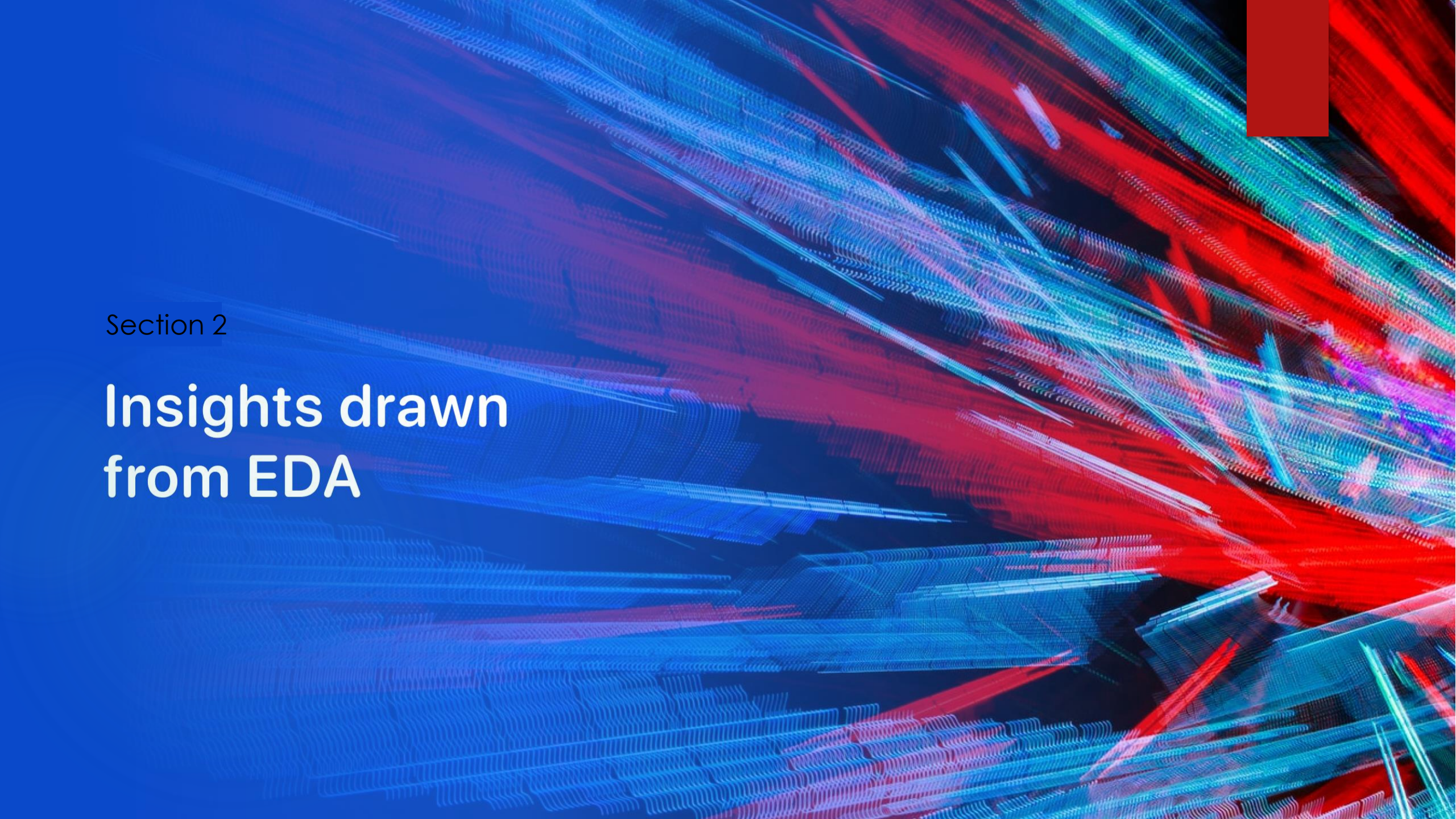
Predictive Analysis (Classification)

- Data used in the analysis was downloaded and cleaned
- Next step is data standardization and splitting the dataset into training and test part
- Various models was applied for prediction and a grid search object was created so we can find the best fitting model easily
- Checking the results of each predicting methord by using confusion matrix
- Compare the model performances
- Based on the results the logistic regression and the SVC method were the most appropriate ones
- Github link:

https://github.com/tp030423/edx/blob/b1725fed9115bb881c35ad3e1214963123c60caa/github_prediction.ipynb

Results

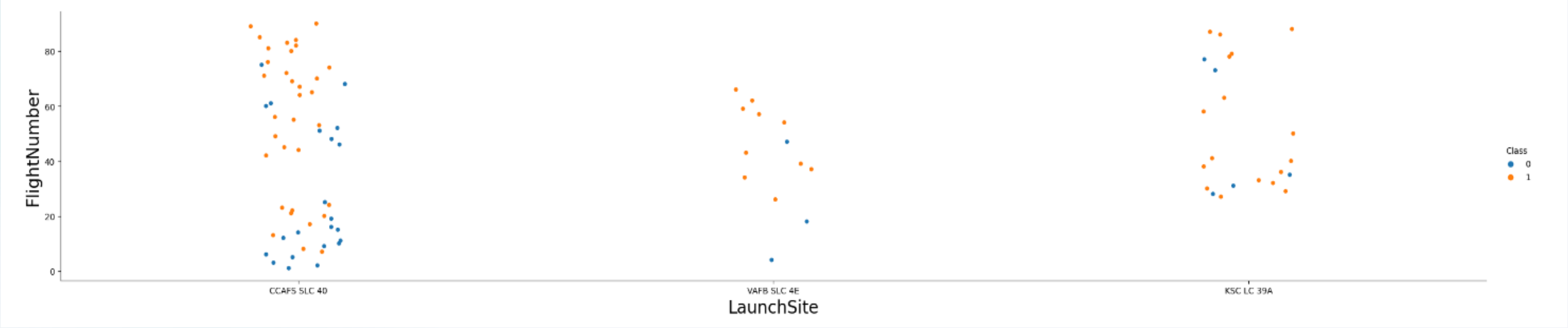
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

Explanation:

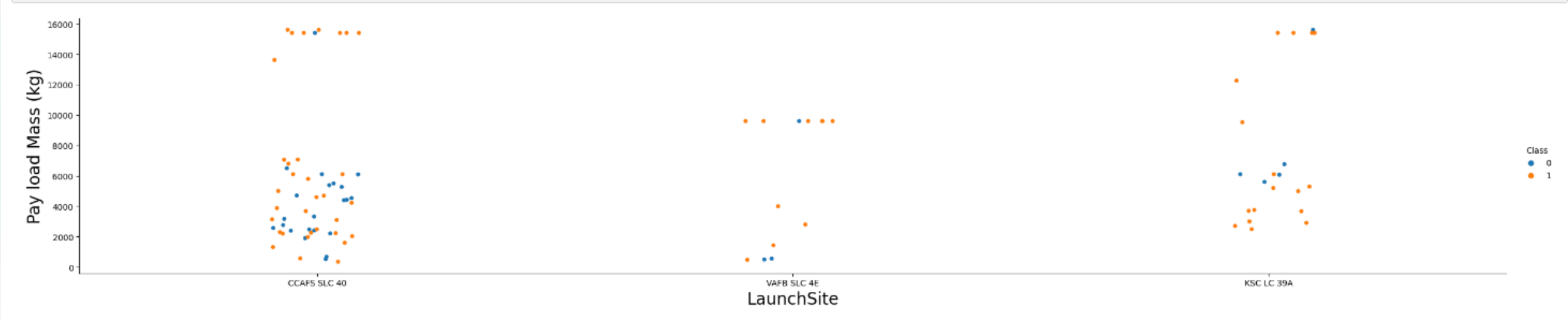
The higher the launch number the higher the chance to have a successful mission, especially in case of CCAFA SLC 40 launch site



Payload vs. Launch Site

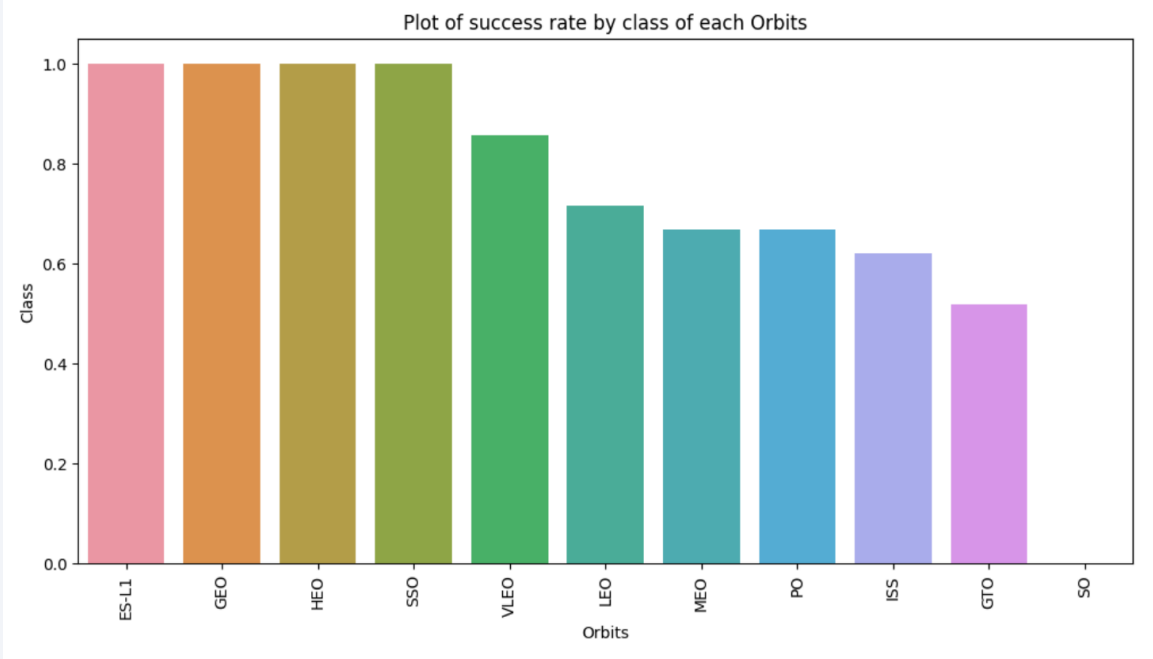
Explanation:

If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).



Success Rate vs. Orbit Type

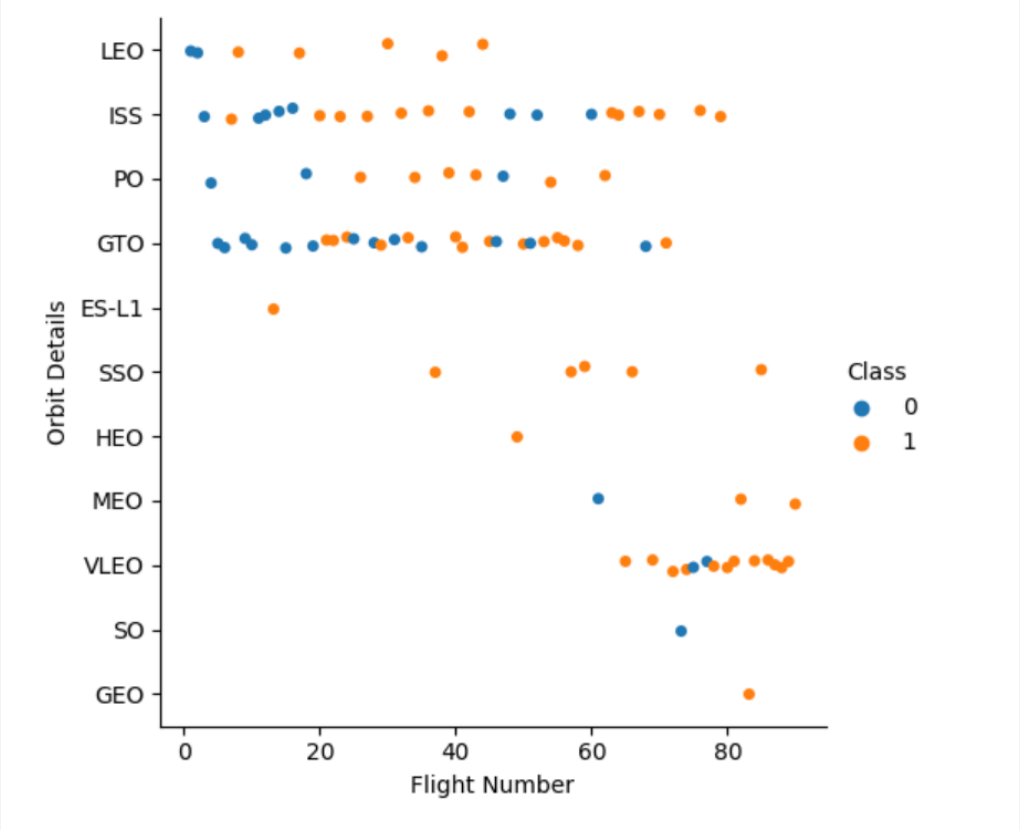
- Show a bar chart for the success rate of each orbit type
- Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

Explanation:

The higher the number of the flight number the higher the chance to have a successful mission regardless of the orbit type. However for SSO and HEO ES-L1 GEO there are only successful missions. But sometimes they only have one mission, so that can be misleading. In case of LEO the lowest number of trials bring good result, almost from the beginning the missions are successful.

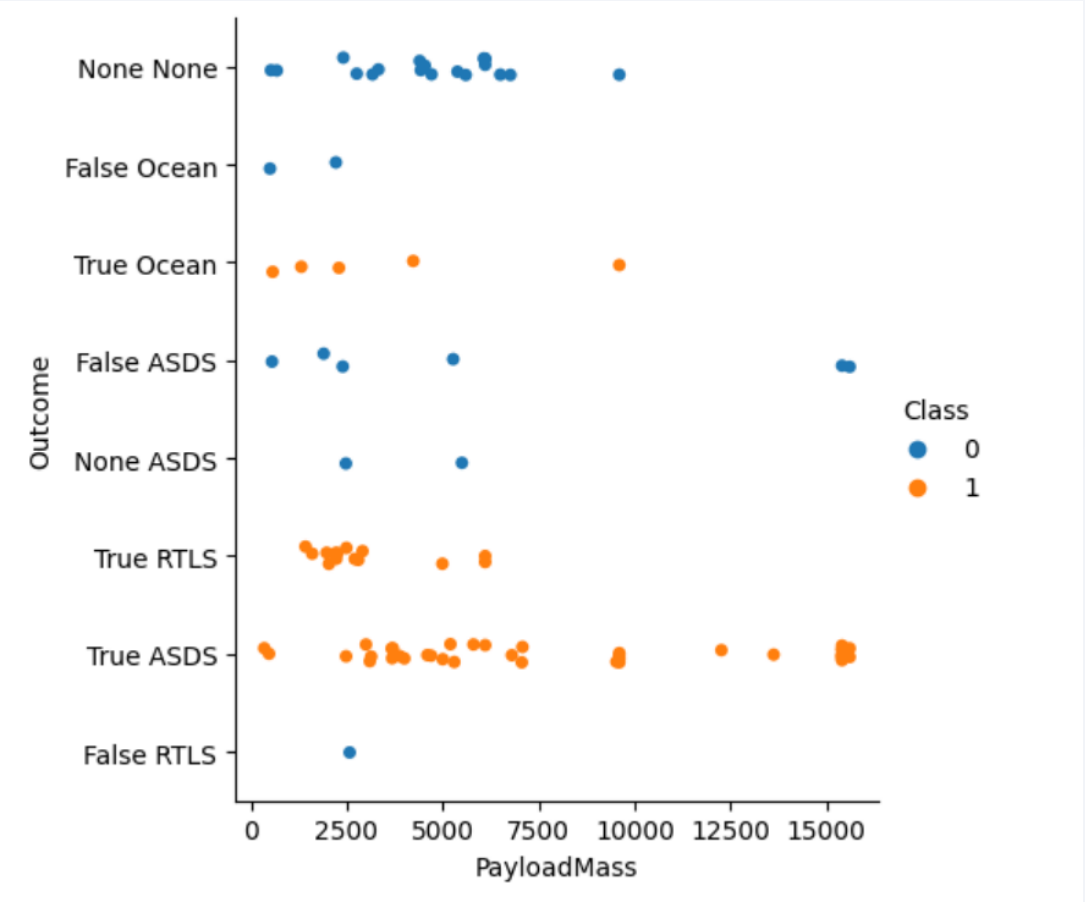


Payload vs. Orbit Type

Explanation:

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

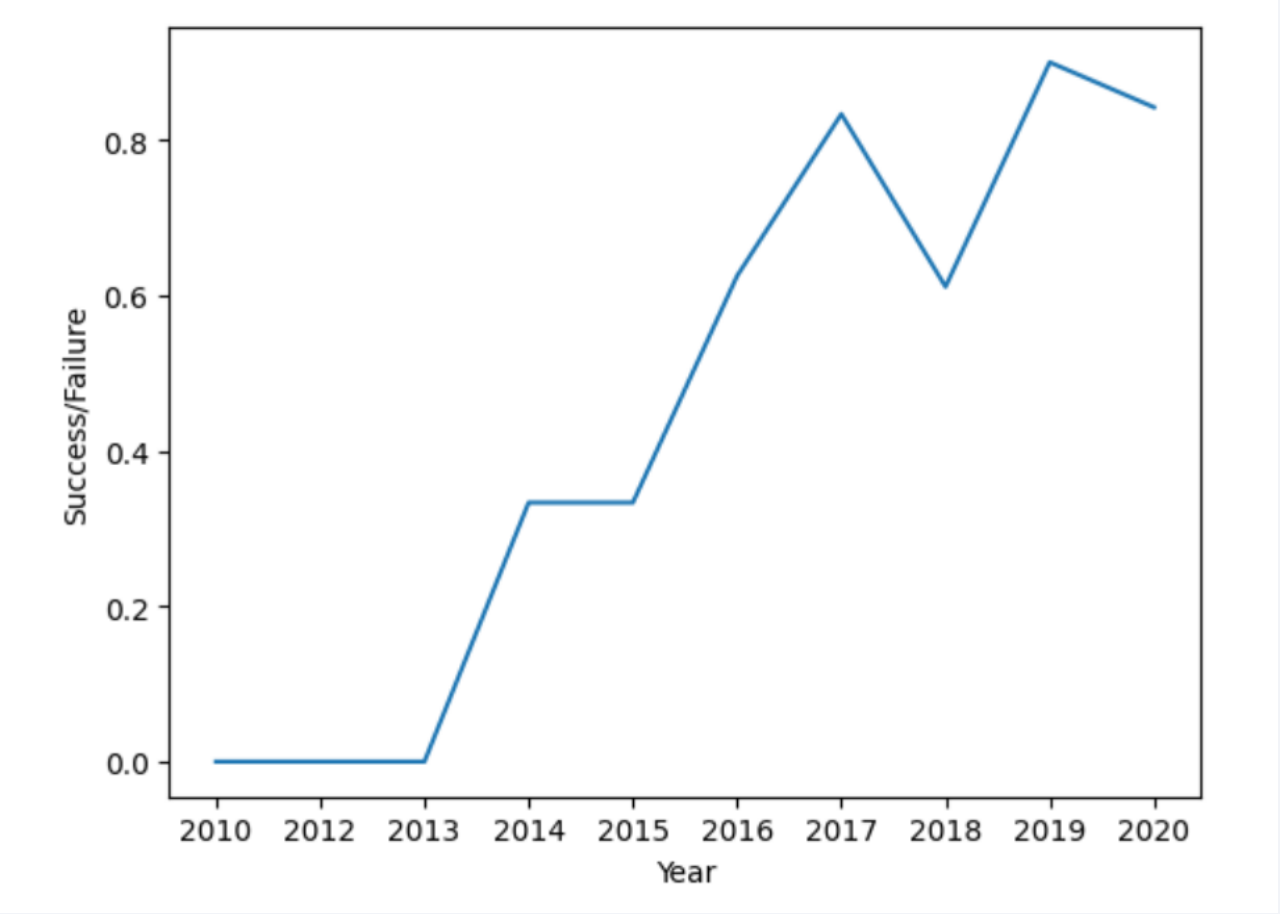
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

Explanation:

Over the years we can observer an increasing trend in case of success rate, with the exception of 2018, all years generally brings higher ratio of successful missions



All Launch Site Names

- Names of the unique launch sites
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- There are 4 different unique launch sites.

Launch Site Names Begin with 'KSC'

5 records where launch sites' names start with `KSC`

The below table displays 5 records where the launch site names starts with KSC.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19/02/2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16/03/2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30/03/2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01/05/2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

Total Payload Mass

Total payload carried by boosters from NASA

- 45596 kg

According to the used database the totapy payload carried by boosters from NASA is 45596 kg.

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

- 2928.4kg

Average payload mass carried by booster version F9 v1.1 is 2928.4kg

First Successful Ground Landing Date

The first successful landing outcome on ground pad was on 22/12/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed with payload mass between 4000 and 6000

F9 FT B1022	F9 FT B1021.2
F9 FT B1026	F9 FT B1031.2

These 4 boosters landed successfully with payload mass between 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

- 101

The total number of successful and failure mission outcomes is 101.

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

F9 B5 B1048.4	F9 B5 B1056.4	F9 B5 B1049.5	F9 B5 B1051.6
F9 B5 B1049.4	F9 B5 B1048.5	F9 B5 B1060.2	F9 B5 B1060.3
F9 B5 B1051.3	F9 B5 B1051.4	F9 B5 B1058.3	F9 B5 B1049.7

There are 12 different boosters which have have carried the maximum payload mass

2015 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2015

month_name	landing__outcome	booster_version	launch_site
DECEMBER	Success (ground pad)	F9 FT B1019	CCAFS LC-40

In 2015 there was only one successful landing outcome. Related details can be found in the table.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

DATE	COUNT
22/12/2015	1
08/04/2016	1
06/05/2016	1
27/05/2016	1
18/07/2016	1
14/08/2016	1
14/01/2017	1
19/02/2017	1

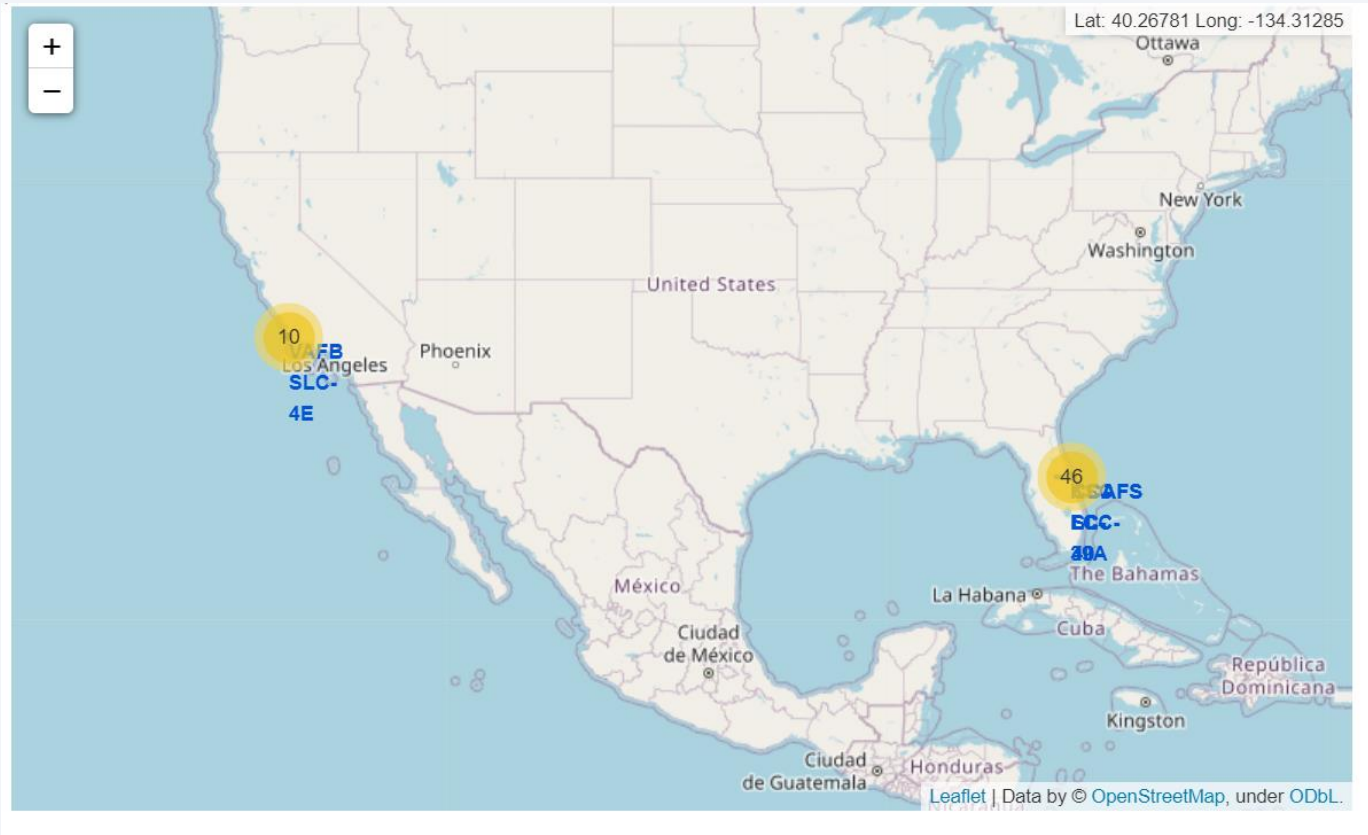
The table represents the successful landing outcomes, on each day there was only one successful landing.

Section 3

Launch Sites Proximities Analysis

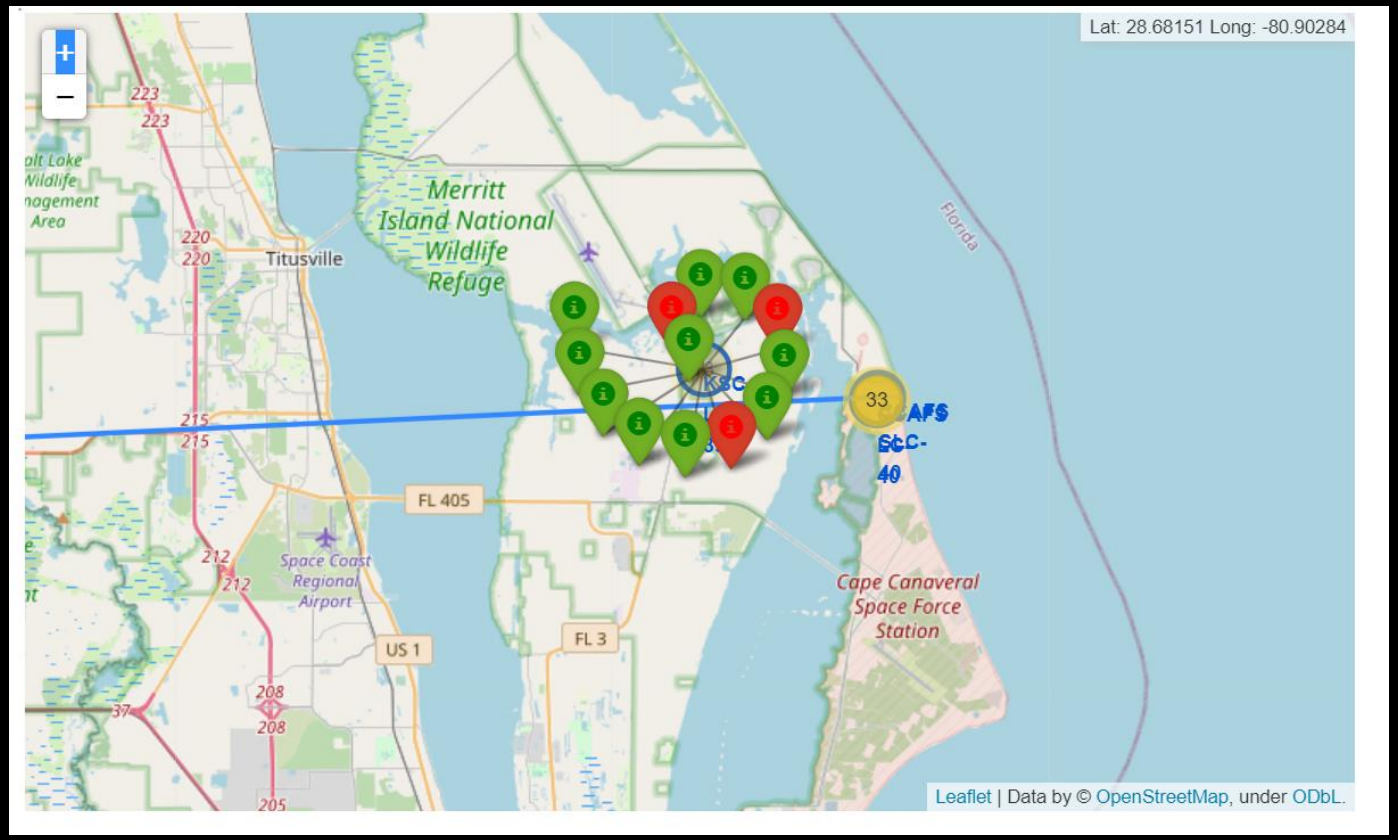
Launch sites

- It can be seen that there are differenc launch sites, west and east side of the US.



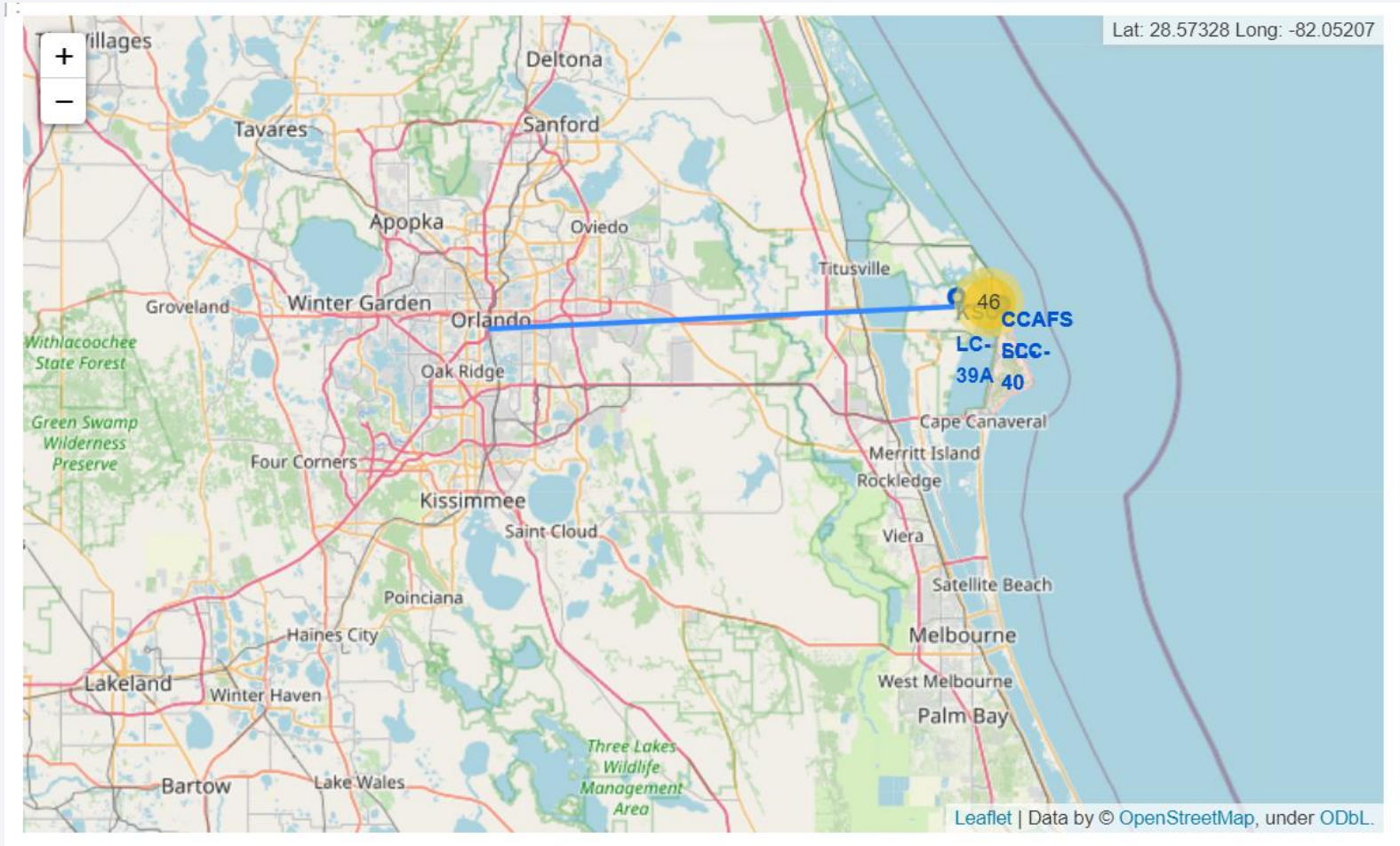
Successful vs. Unsuccessful missions

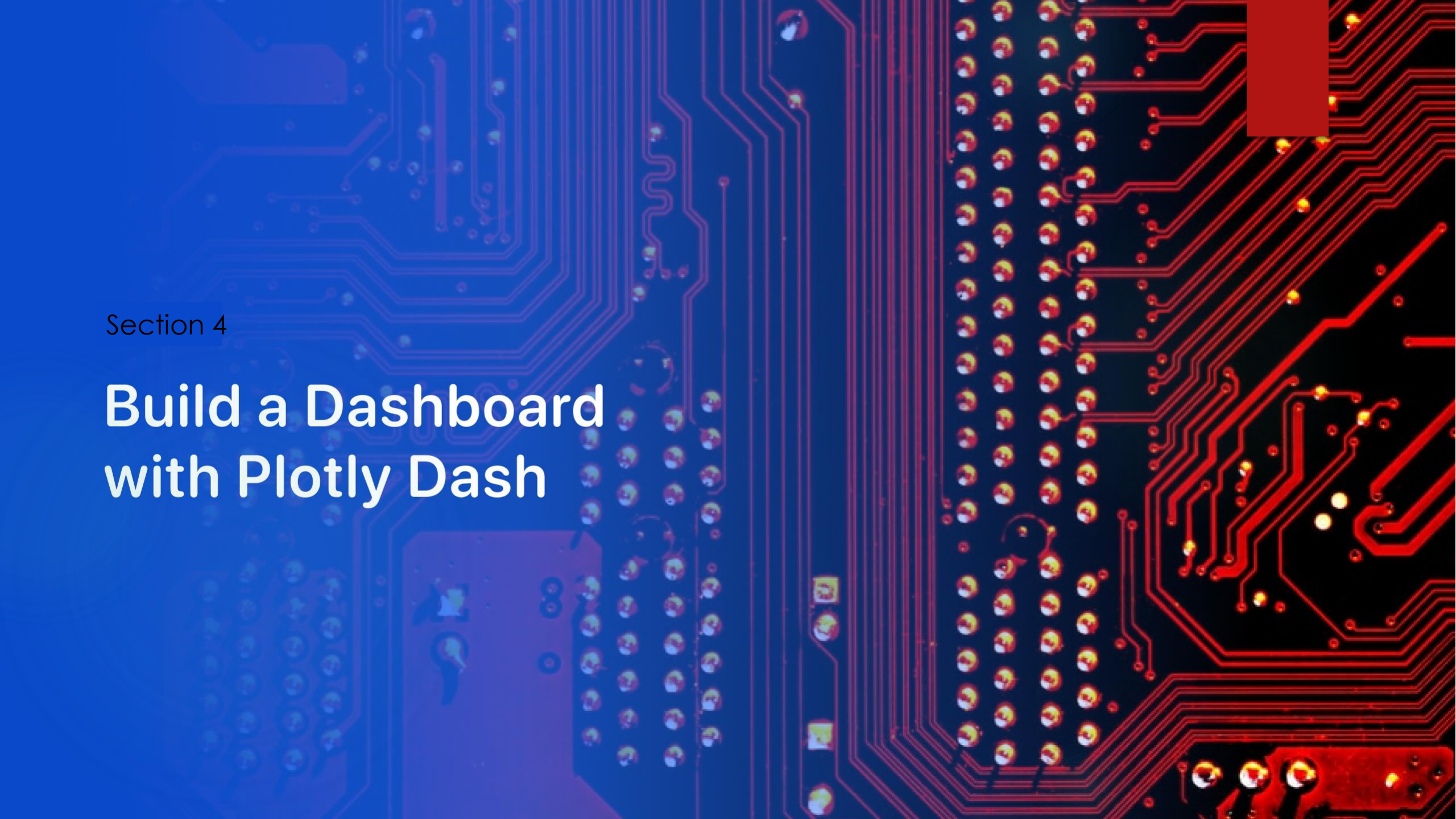
- It can be seen that on a certain launch site how many successful (green) and unsuccessful (red) mission happened.



Proximity

- Proximity displayed with a line can help us calculate the distance from the launch site. This can be useful in many cases, such as unsuccessful landing to see what can be in the nearby.




Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic regression and SVM has the best performance on the data.

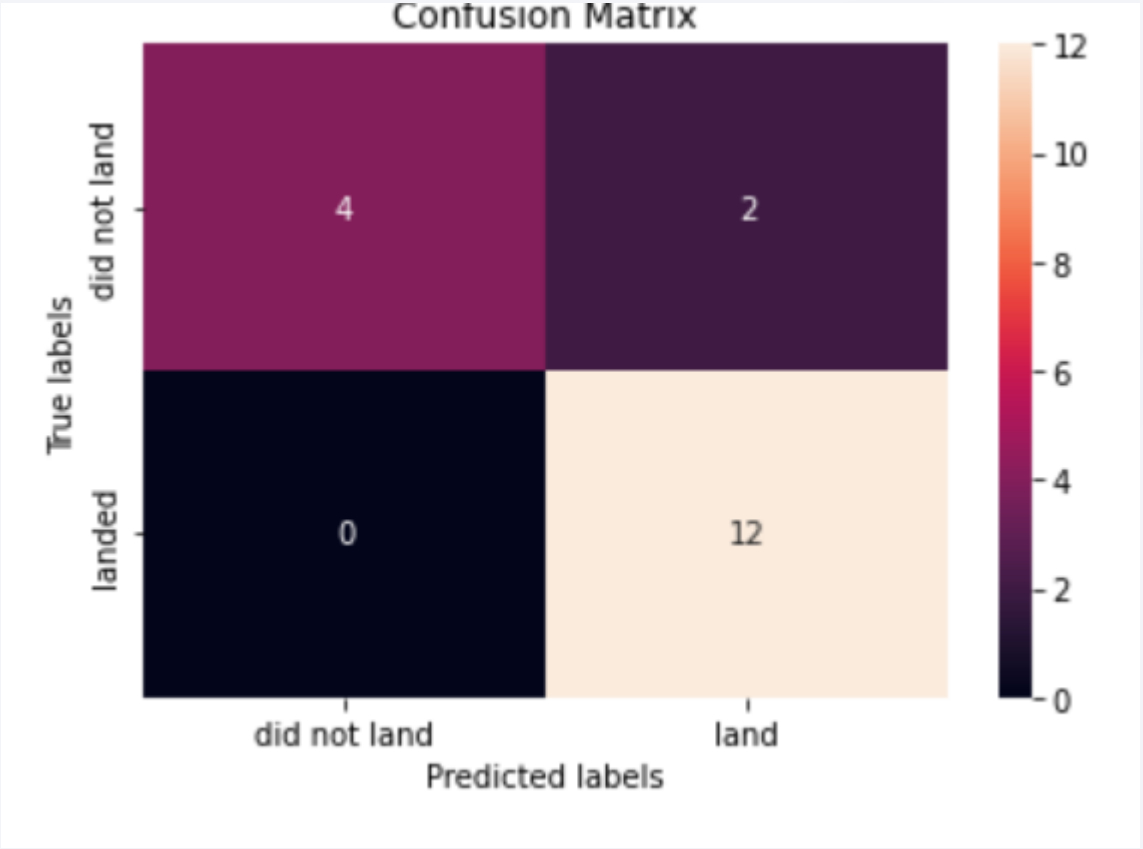
Scores on test data for each method

- Logistic Regression: 0.944
- SVM: 0.944
- Decision Tree: 0.888
- KNN: 0.888

Conclusion: Logistic Regression and SVM deliver the best performance on test data.

Confusion Matrix

- Confusion matrix of the best fitting model. Confusion matrix compares the real and the predicted values. If the fals positive and false negative values are small compated to the true positive and true negative (so the correct predictions) it means the prediction is good. See 0 and 2 values are low compared to the 4 and 12.



Conclusions

- Logistic regression and SVM model has the performance on mission outcome prediction.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!