IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

James, Friday Emmanuel

October, 2022

Outline

Executive Summary

Problem Definition

CBS news has it that the most competitive businesses are those that effectively transform data relating to competitors into decision-relevant insights. Also, understanding a competitor will be easier if we develop a keen sense of market awareness and product quality.

SpaceX designs, manufactures and launches rockets and spacecrafts and dominate the market because of the reusability of its rockets. However, not all rockets launched land successfully. The problem is how we can leverage this unsuccessful landing and develop methodologies to predict if a rocket would land successfully or not.

Significance

The machine learning models would enable our company, SpaceY to determine the success of rocket landing. This is particularly significant as a failed landing could lead to huge financial losses. The prediction would help the cost analysis and make better offers than SpaceX.

Method Used:

We collected data on SpaceX's Falcon 9 rocket launches and used it to train Machine Learning models that will predict whether the first stage of the rocket will land successfully. This was done through the following ways:

- Harvesting data on SpaceX's Falcon 9 rocket launches using API requests and Web Scraping
- Exploratory Data Analysis
- Building predictive models using machine learning algorithms

Introduction

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. However, not all of its launches land successfully.

The objective of this project is to allow our company, SpaceY to compete with SpaceX. In order to achieve this goal, it is necessary to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.

Ultimately, an accurate prediction of the likelihood of the first stage rocket landing successfully will help us to determine the cost of a launch. With the help of the Exploratory Data Analysis and predictive models, we can make more informed decisions and compete with SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Make requests to the SpaceX API.
  - Perform web scraping to collect Falcon 9 historical launch records on the Wikipedia page titled: List of Falcon 9 and Falcon Heavy launches
- Perform data wrangling
  - This involves cleaning the data and removing unwanted features in order to make the data fit for modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Python’s Folium, Plotly and Dash Libraries.
- Perform predictive analysis using classification models
  - We build Machine Learning classification algorithms to determine whether the first state of the Falcon 9 rocket launch would land successfully or not.
  - We use different hyperparameters to in Grid Search to train the best model and determine its accuracy.

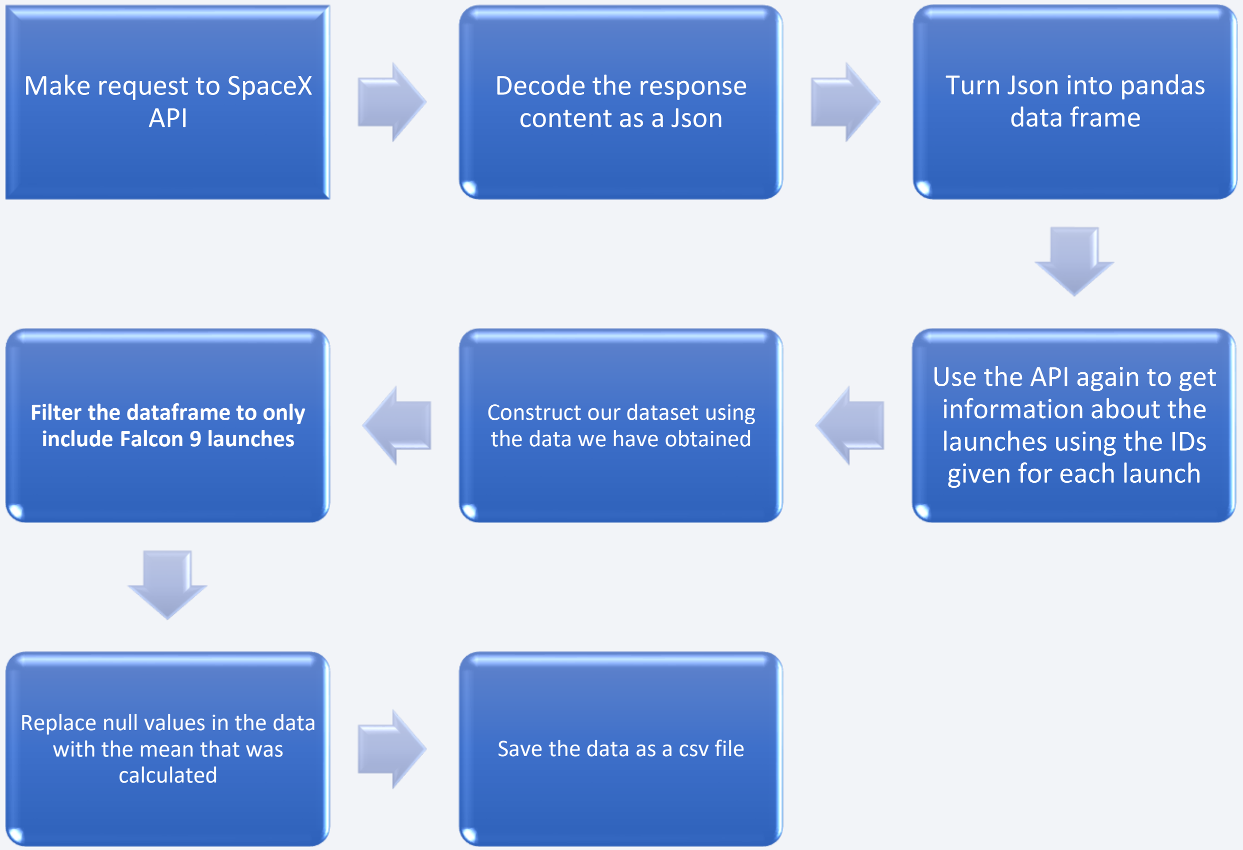
Data Collection

Having defined the objective and significance of the project, we now have a clear direction of the nature and method(s) of collecting the relevant data viz:

- Making an Application Program Interface (API) request to SpaceX
- Scraping SpaceX data from their website data repository

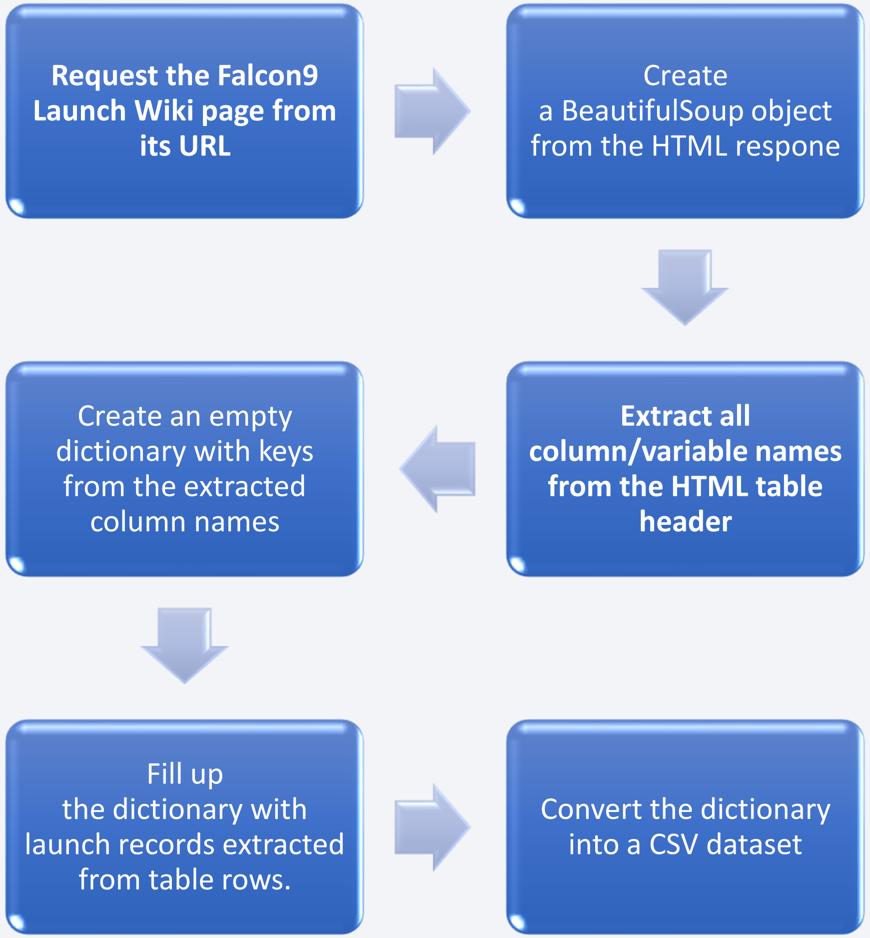
Data Collection – SpaceX API

- Make a request to SpaceX API
- Perform data wrangling, dealing with missing data and putting data in proper format.
- Lastly, save the data as a csv file
- URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/SpaceX-data-collection-api.ipynb



Data Collection - Scraping

- Using BeautifulSoup Python library, we perform web scraping on the wikipedia page with title: [List of Falcon 9 and Falcon Heavy launches]
- Store the launch records in an HTML table.
- Parse the table, save the data as a csv file
- URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/SpaceX-data-collection-web-scraping.ipynb



Data Wrangling

The data is cleaned, and null values replaced with the mean of the respective columns.

We notice several instances where the rocket did not land successfully. For example, ***True RTLS*** means the rocket successfully landed on a ground pad while ***False RTLS*** means the rocket unsuccessfully landed on a ground pad.

Those outcomes were converted into Training Labels whereby ***1*** means the rocket landed successfully while ***0*** means it was unsuccessful.

URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/Spacex_Data%20wrangling.ipynb

Determine the number of launches on each facility

↓

Calculate the number and occurrences of each orbit

↓

Calculate the number and occurrences of mission outcome per orbit type

↓

Create a landing outcome training label and loop through all the landing outcomes

↓

Create a "Class" column containing the information from the outcome label

↓

Transform the data frame into a CSV dataset.

EDA with Data Visualization

Exploratory Data Analysis through visualizations depicts the distribution of the data, identifies any outliers and patterns in the data. It also explores the relationship between the features and the target label. Here, we used:

- Cat plots and scatter plots to view the relationships of categorical variables like *Launch Site* and *Orbit*.
- A bar chart was used to visualize the success rate of each orbit type.
- A line chart was used to visualize the launch success yearly trend.

URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/Exploratory_Data_Analysis_with_Matplotlib_Pandas.ipynb

EDA with SQL

Summary of SQL queries that were used:

- Display the names of the unique launch sites in the space mission
- Compare the payload mass with boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the total number of successful and failure mission outcomes
- Determine the dates of different landing outcomes

URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/Exploratory_Data_Analysis_with_SQL.ipynb

Build an Interactive Map with Folium

- Folium Circles were used to highlight circle area of launch sites.
- Folium Markers were used to show the SpaceX launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to their nearest land marks.
- In order to mark the success/failed launches for each site, marker clusters were used on the map. Whereby Red represents rocket launch failures while Green represents the successes.
- URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/Launch%20Site%20Location.ipynb

Build a Dashboard with Plotly Dash

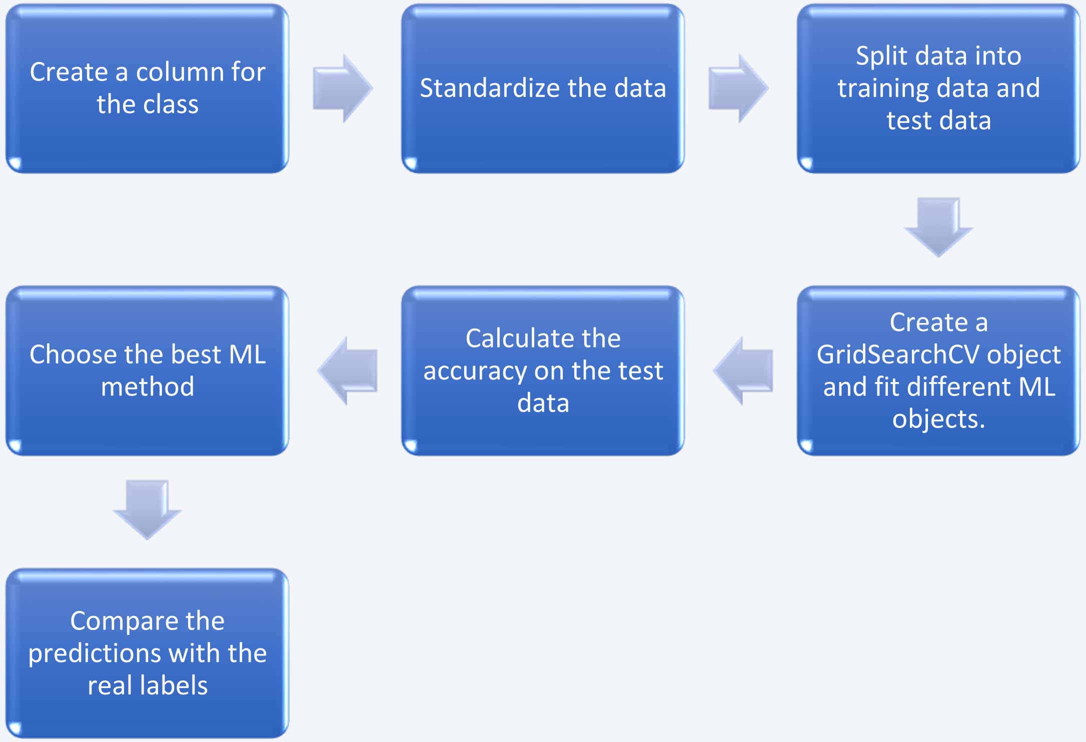
- Pie charts and scatter charts were used to visualize the launch records of SpaceX.
- The plots and interactions give us insight as to the rocket launch success rate per launch site and mass of the payload.
- Successful launches were represented by 1 while failures were represented by 0.
- URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/Building%20Interactive%20Dashboards.ipynb

Predictive Analysis (Classification)

Python's Machine Learning library Scikit-learn is for the predictive analysis. We used it to:

- Build a machine learning pipeline to predict whether the first stage will land successfully
- Using *GridSearchCV*, found the best parameters for each classification algorithm
- Apply the model on the test data to get accuracy scores for each model.

URL link: https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction/blob/main/SpaceX_Machine%20Learning%20Predictions.ipynb



Results

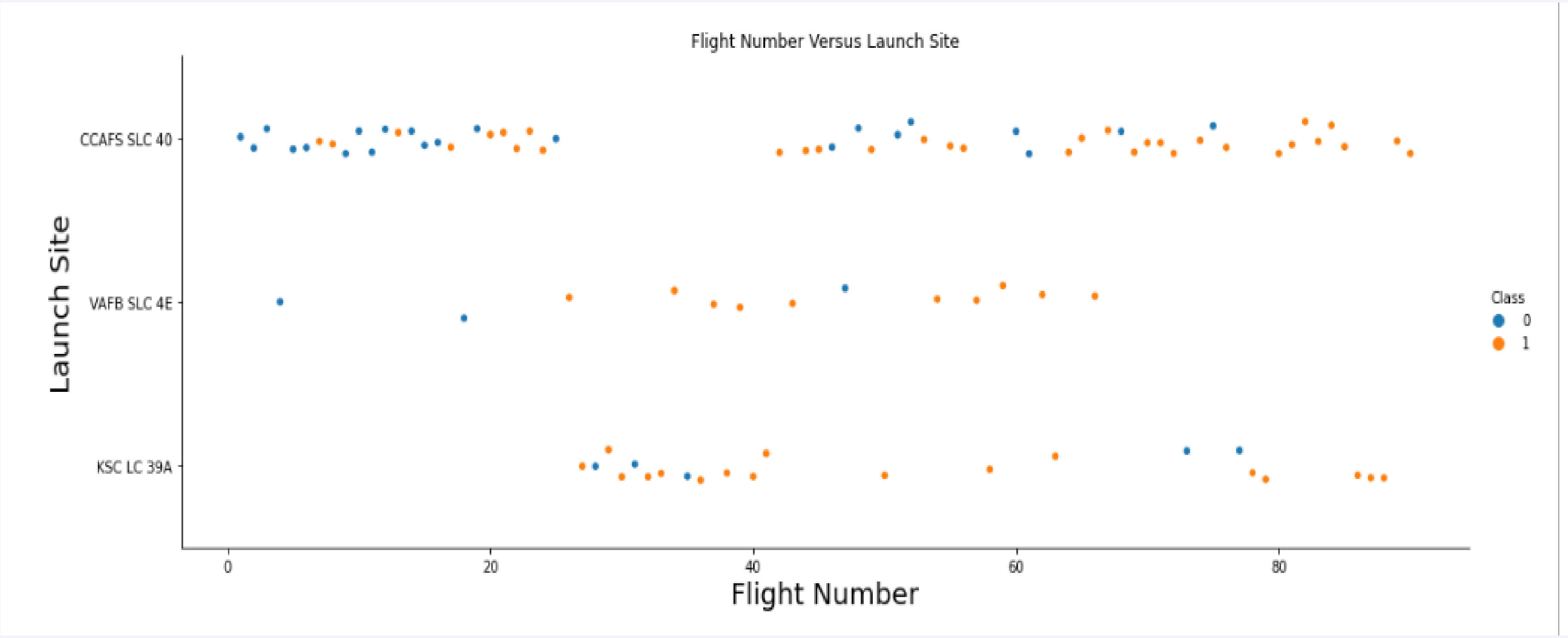
- The EDA shows that:
  - Strong correlation between the landing outcome and the flight number.
  - Payload mass in kg is inversely correlated with the landing outcomes
  - It was also apparent that successful landing outcomes have had a significant increase since the year 2015.
- Spatial visualizations show that:
  - All launch sites are located near the coastline
  - The launch sites are also located near highways and railways.
- The machine learning algorithms classifies rockets as successfully/unsuccessfully landing with an accuracy of 83.33%.

Section 2

Insights drawn from EDA

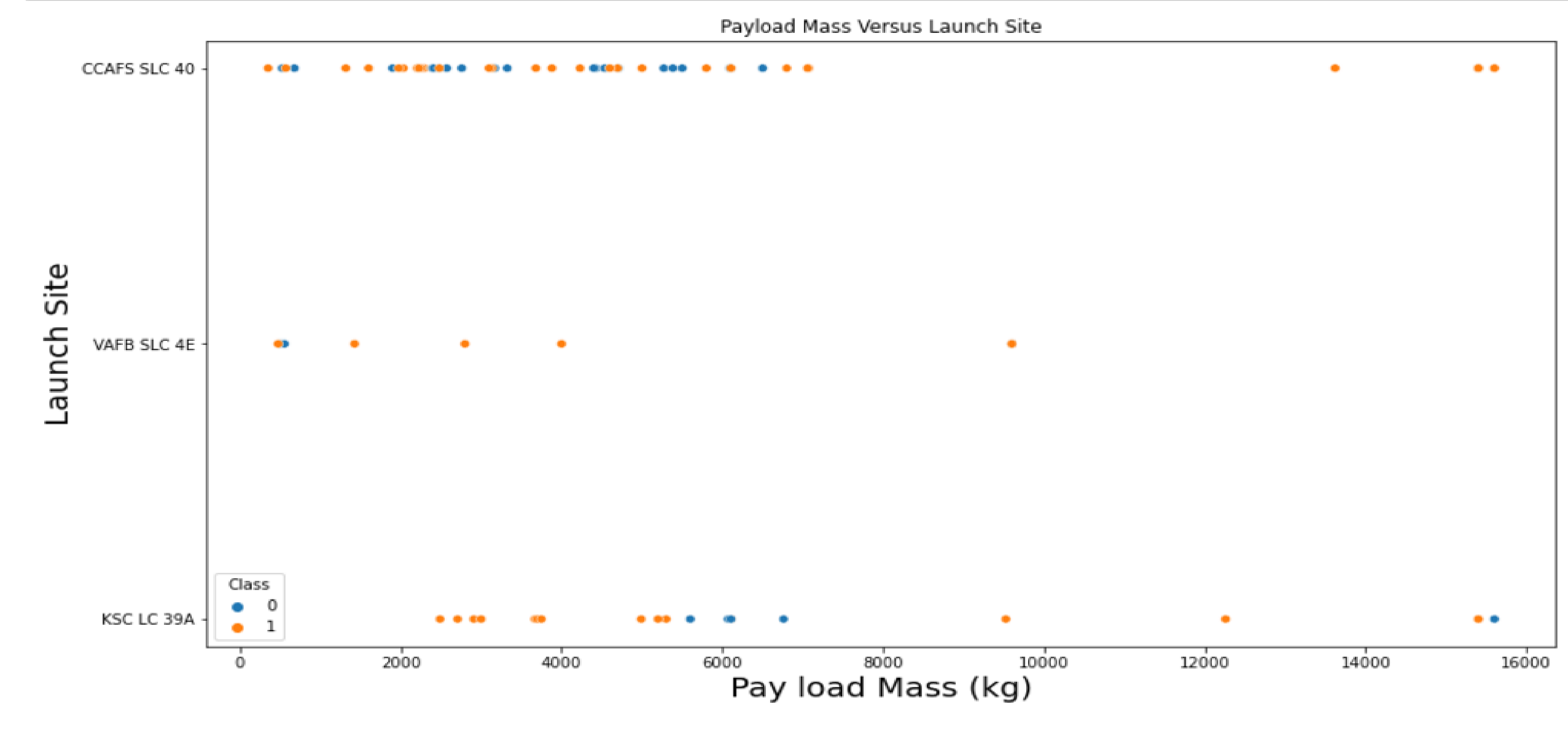
Flight Number vs. Launch Site

- The visualization depicts that the flight number is directly proportional to the number of successful landings.
- The launch site **CCAFS SLC 40** had the most number of landing attempts while the site **VAFB SLC 4E** had the least number of attempts.



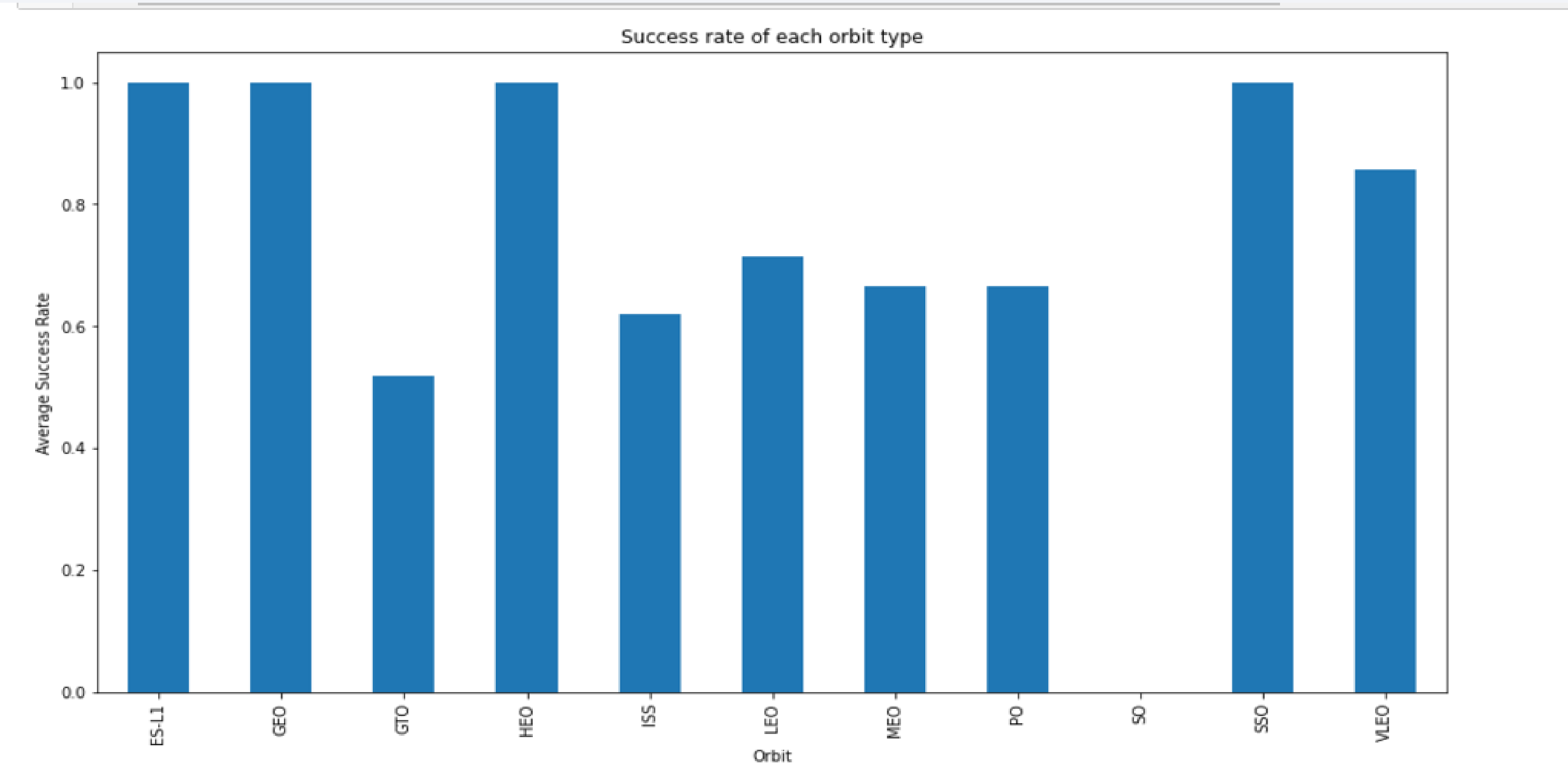
Payload vs. Launch Site

- The launch sites are sensitive to the mass of the payload in that rockets with payload mass greater than 10000kg are most unlikely to be launched at the sites.



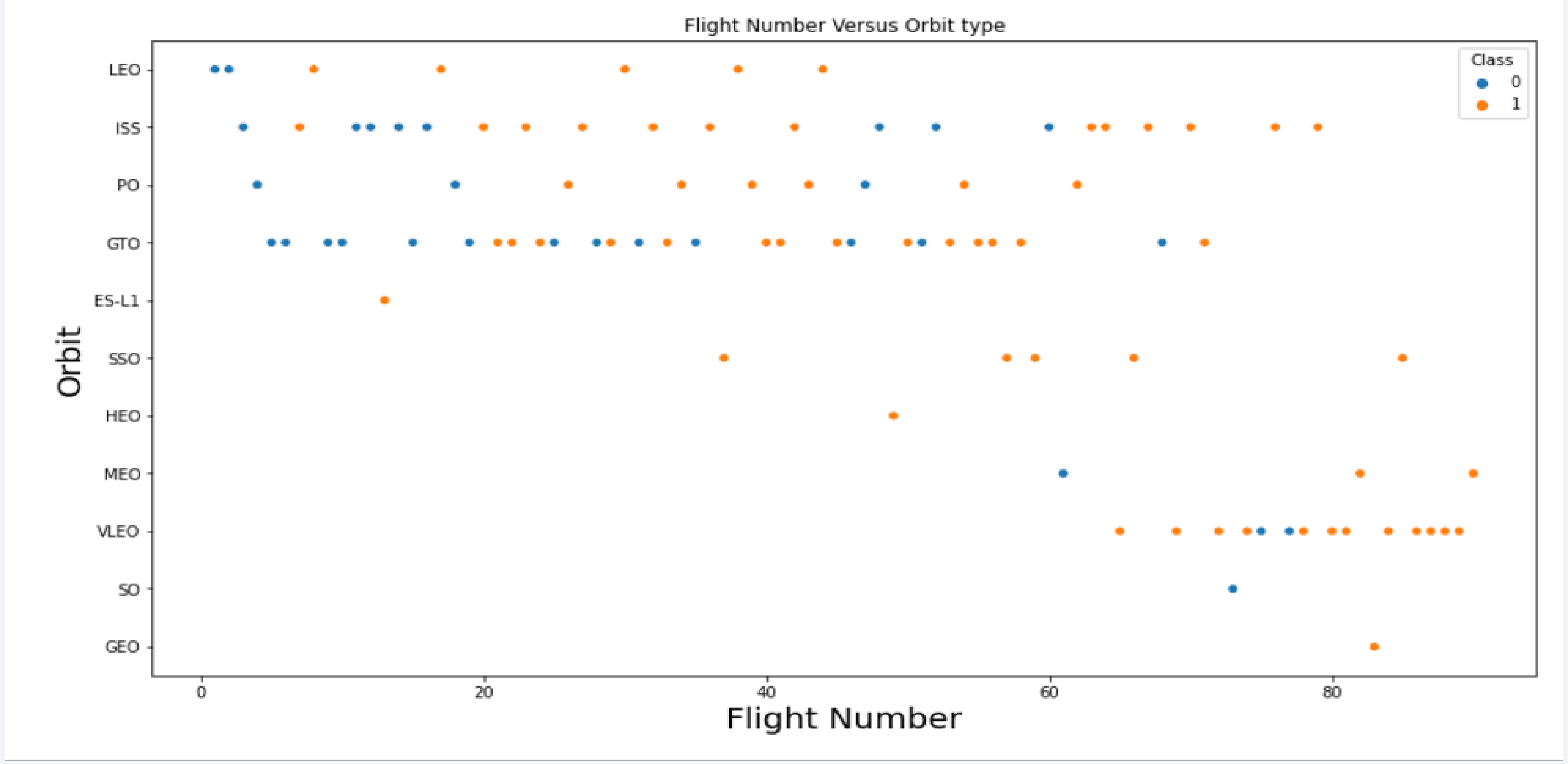
Success Rate vs. Orbit Type

- Four (4) of the Orbits (ESL1, GEO, HEO and SSO) records the most success rate on the average. This should be a crucial factor to consider when launching a rocket



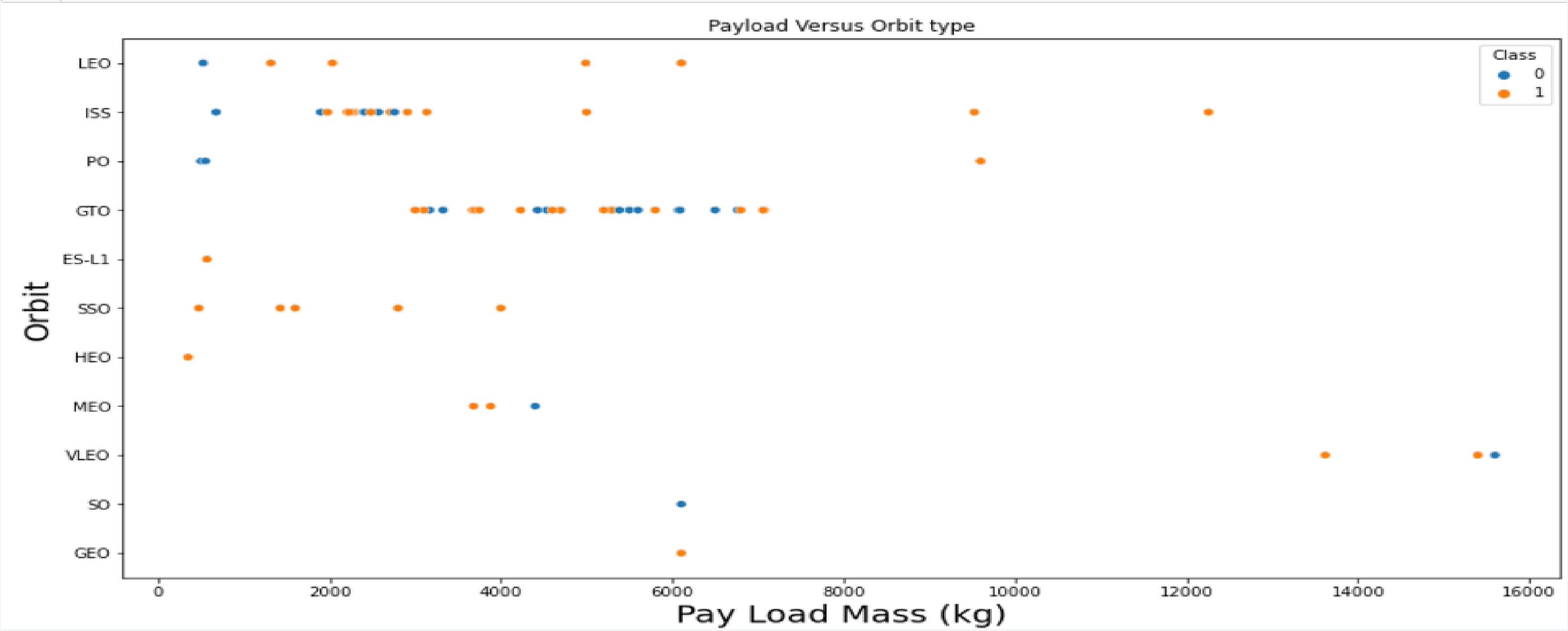
Flight Number vs. Orbit Type

- LEO, SSO and VLEO appears to have strong correlation with the flight number in recording landing successes.



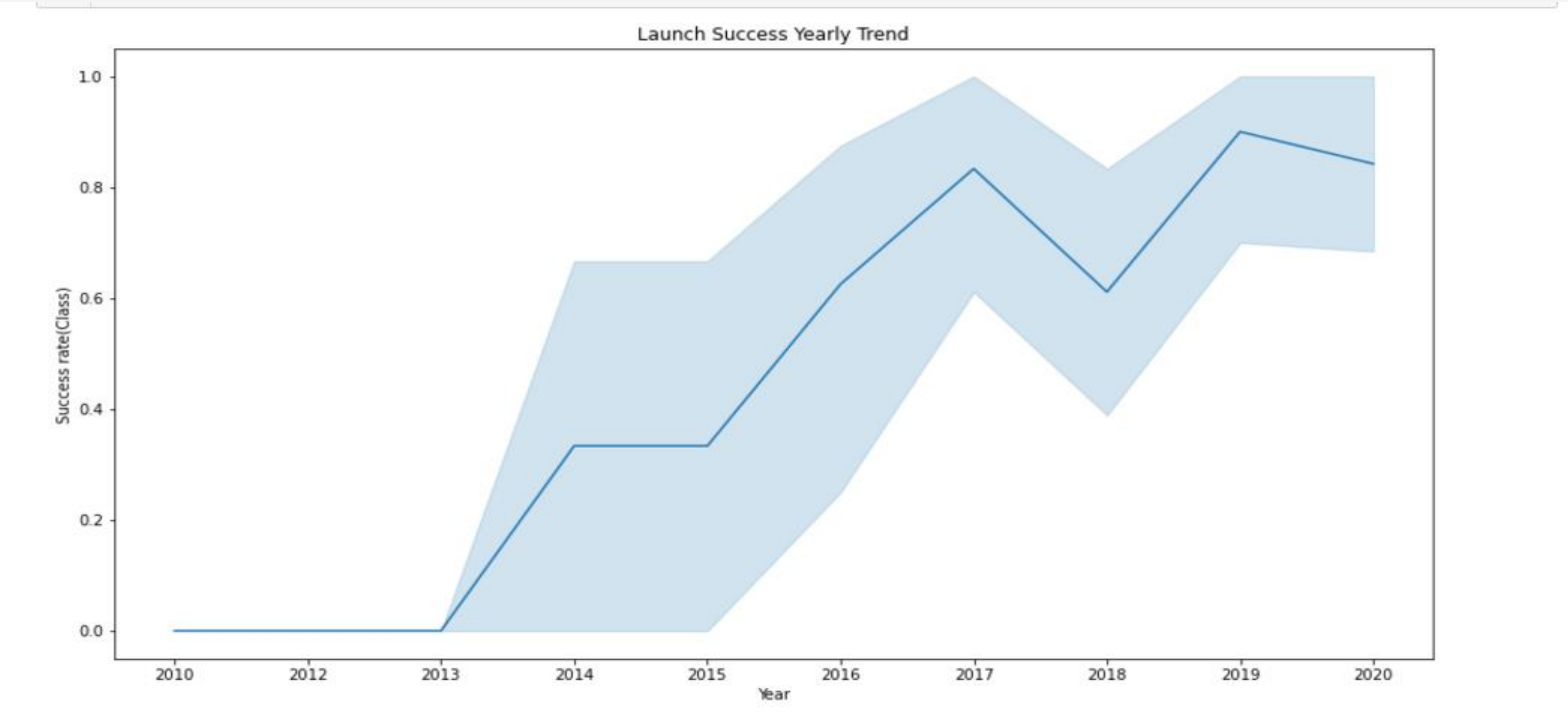
Payload vs. Orbit Type

- There are more successful launches for LEO and SSO Orbits when the payload mass is relatively small compared to GTO and ISS where the effect of the payload mass is not easily noticeable. GEO, ESL1 and GEO all had successful landing with low payload masses.



Launch Success Yearly Trend

- There has been a successful launch of the SpaceX Falcon 9 rocket from 2013. Albeit there was a slight decline in 2018 but the trend goes upward thereafter



All Launch Site Names

There are four (4) distinct launch sites gathered from the data viz:

- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Result for 5 launch sites that begin with ‘CCA’ shows that NASA also launched their rockets on these sites.

Date	Launch_Site	Orbit	PayloadMass (Kg)	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	CCAFS LC-40	LEO	0	SpaceX	Success	Failure (parachute)
08-12-2010	CCAFS LC-40	LEO (ISS)	0	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	CCAFS LC-40	LEO (ISS)	525	NASA (COTS)	Success	No attempt
08-10-2012	CCAFS LC-40	LEO (ISS)	500	NASA (CRS)	Success	No attempt
01-03-2013	CCAFS LC-40	LEO (ISS)	677	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload mass carried by boosters launched by NASA .
- *NASA (CRS)* had a relatively high total payload mass than the other NASA customers

Customer	Total_Payload_Mass
NASA (CRS)	45596
NASA (CCDev)	12530
NASA (CCP)	12500
NASA (CCD)	12055
NASA (CTS)	12050
NASA (CRS), Kacific 1	2617
NASA / NOAA / ESA / EUMETSAT	1192
NASA (LSP) NOAA CNES	553
NASA (COTS)	525
NASA (LSP)	362
NASA (COTS) NRO	0

Average Payload Mass by F9 v1.1

- The average payload mass carried by F9 v1.1 was 2928.4 kg.

Average_Payload_Mass (kg)	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad was December 22, 2015

Date	Landing_Outcome
22-12-2015	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- This confirms that low payload mass have a positive impact on the landing outcome

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes

Mission_Outcome	Outcomes
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Twelve (12) Boosters carried the maximum payload and all from the same manufacturers.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- In the months of January and April of the year 2015, two versions of the F9 v1.1 rockets failed to land successfully at the CCAFS LC-40 site.

Date	Launch_Site	Booster_Version	Landing_Outcome
10-01-2015	CCAFS LC-40	F9 v1.1 B1012	Failure (drone ship)
14-04-2015	CCAFS LC-40	F9 v1.1 B1015	Failure (drone ship)

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

- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 shows an increase in successful landing from 2015

date	Landing_Outcome	Outcomes
2016-04-08	Success (drone ship)	14
2015-12-22	Success (ground pad)	9
2015-06-28	Precluded (drone ship)	1
2015-01-10	Failure (drone ship)	5
2014-04-18	Controlled (ocean)	5
2013-09-29	Uncontrolled (ocean)	2
2012-05-22	No attempt	22

Section 3

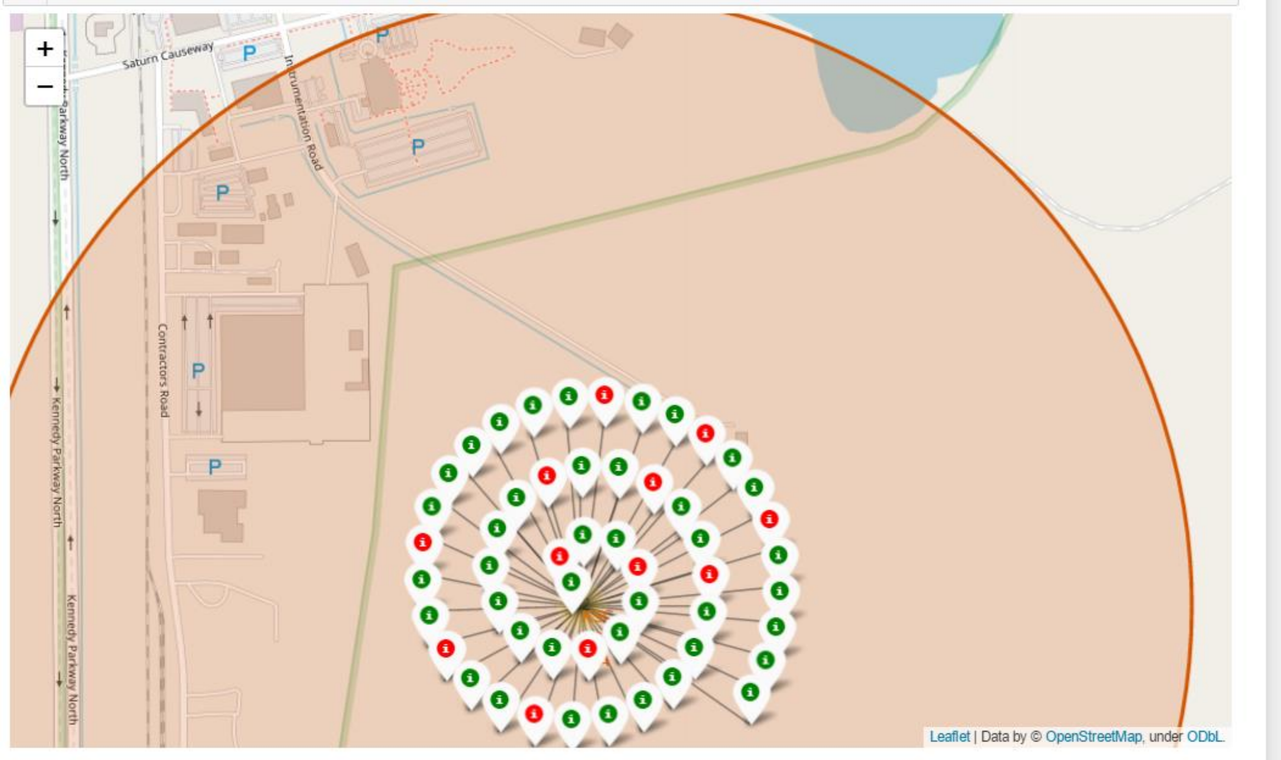
Launch Sites Proximities Analysis

Launch Site Locations

- Majority of the launch sites are around the coastal regions of Florida

Success Rate by Launch Sites

- The figure below shows the success rate distribution at the **KSC LC-39A** launch site which happens to have the most successful landing outcomes compared to other launch sites.
- The successful launches are represented by a green marker while the red marker represents failed rocket launches.



Launch Site Proximities

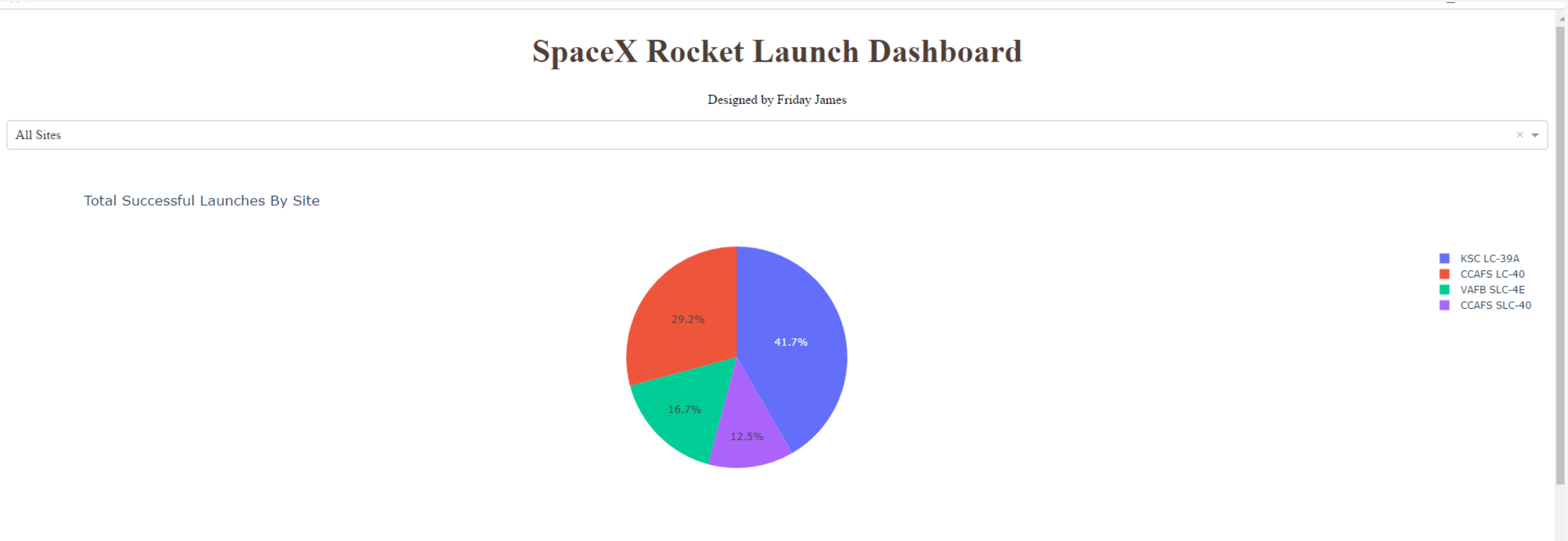
- Launch sites are far away from metropolitan cities in order to avoid damages to human and materials
- They are also close to the coastlines as it is more appropriate to test the rocket landing on large water masses

Section 4

Build a Dashboard with Plotly Dash

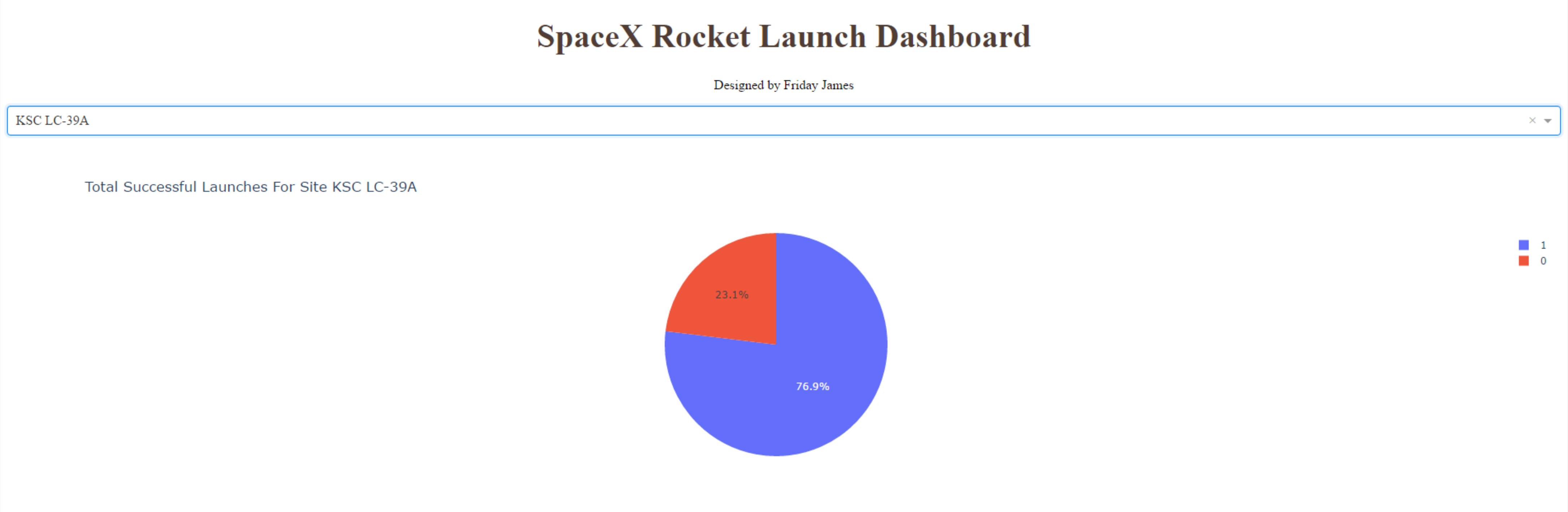
Launch Success Count for All Sites

- Displaying the distribution of the total successful launches by sites.
- Site **KSC LC-39A** has the most successful launch compared to the rest launch sites



Highest Launch Success Ratio

- Site **KSC LC-39A** has the most successful launch compared to the rest launch sites



Launch Outcome With Respect to Payload

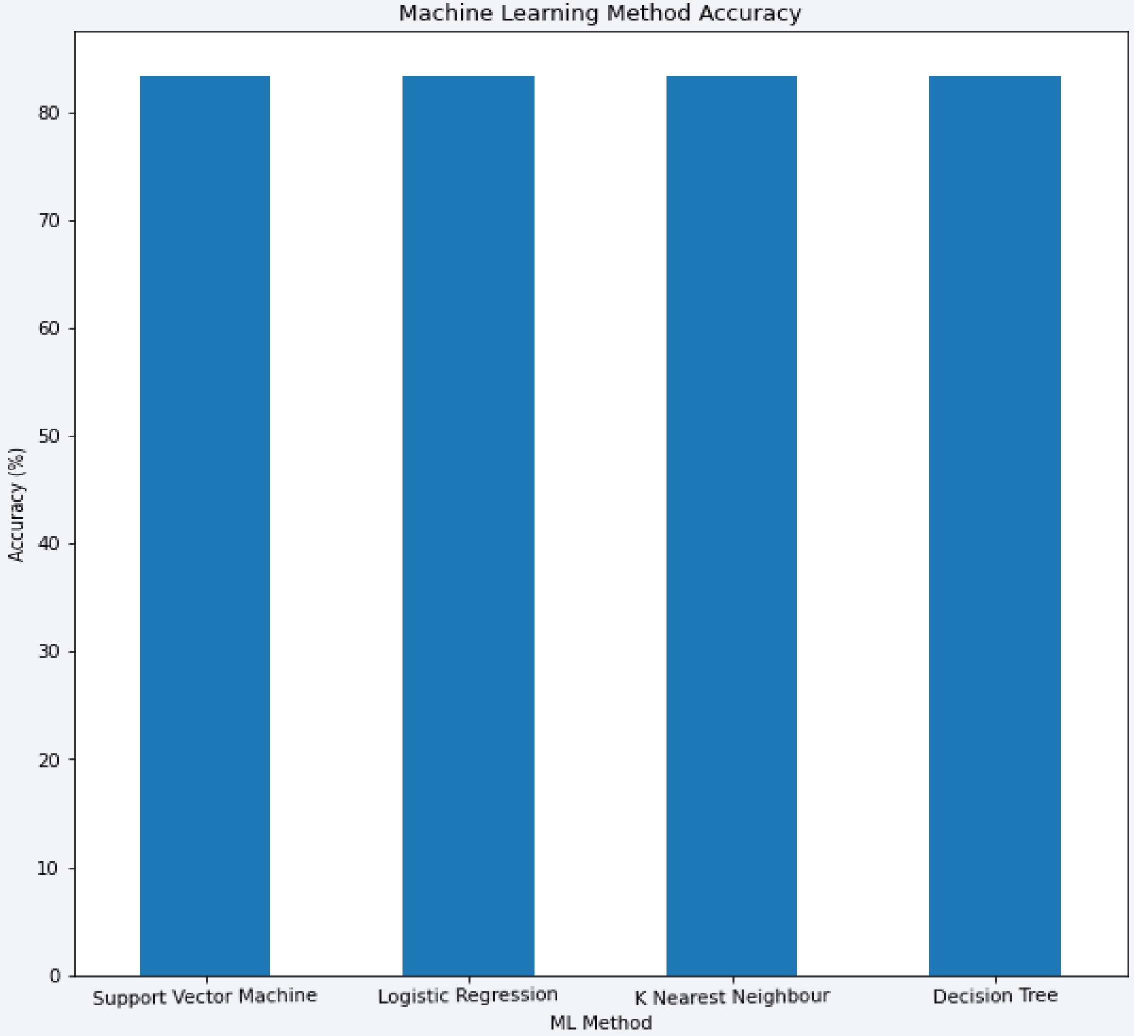
- A payload range between 3000 kg and 4000 kg has the highest success rate (approximately 72.73%).
- The booster version **FT**, seems to have a higher success rate than other booster versions

Section 5

Predictive Analysis (Classification)

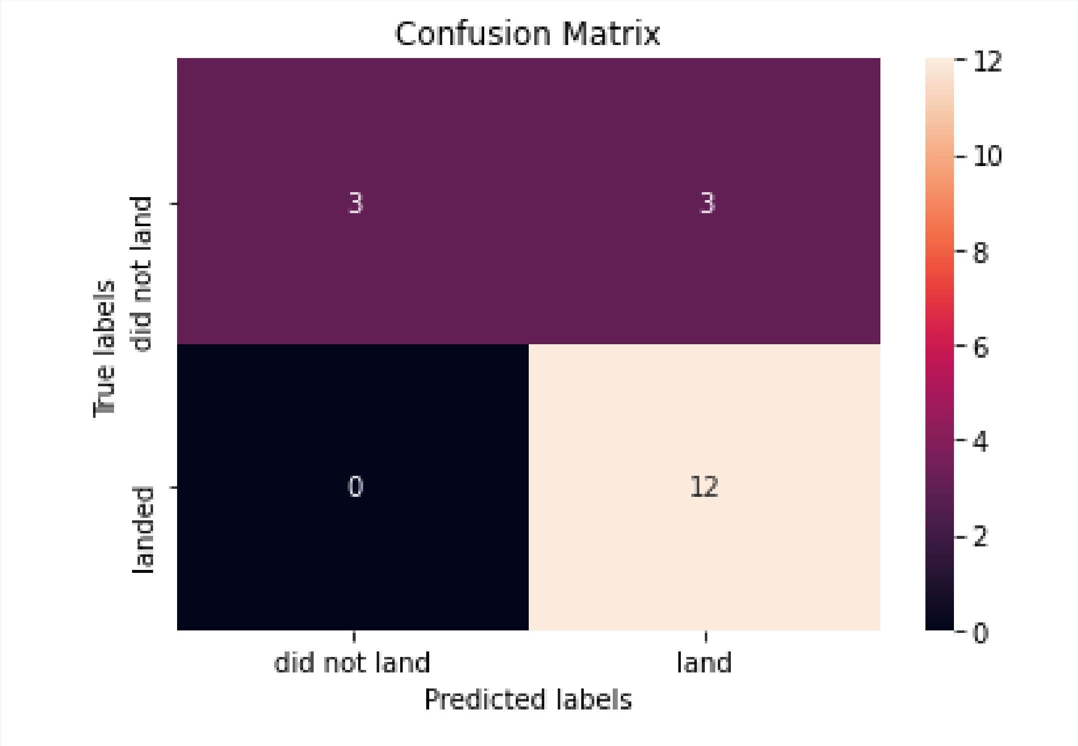
Classification Accuracy

- Each machine learning algorithm has equal accuracy score of 83.33%.



Confusion Matrix

- The chart shows the confusion matrix of the Decision Tree
- The model only failed to accurately predict 3 labels.



Conclusions

From the Exploratory Data Analysis and interactive visualizations on the SpaceX dataset, the following conclusions can be drawn with respect to the landing outcomes and relationships among the features that determine the landing outcomes of the Falcon 9 rocket

- Payload mass (in kg) is inversely correlated with the landing outcomes. The higher the mass of the payload, the less successful it is for the rocket to land successfully
- Site KSC LC-39A had the highest launch success rate out of all the launch sites.
- Most of the launch sites are located close to the coastlines and far away from metropolitan cities.
- It was also apparent that successful landing outcomes had an upward trend since the year 2015
- There is a possibility to predict whether a rocket launch would be successful or not with an 83.33% accuracy.
- With these, our company, SpaceY would be able to compete with SpaceX since we can channel resources to rockets with less payload mass, target the KSC LC-39A and increase the flight number.

Appendix

The Python codes used for this project can be found in my GitHub repository at:

https://github.com/Friday-J/SpaceX_Falcon-9_Rocket_Landing_Prediction

Thank you!