Winning Space Race with Data Science

Austin Kingsley
January 16, 2024

Outline

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

Executive Summary

- API calls and web scraping was executed to retrieve the SpaceX launch information, python and SQL were used for data wrangling and analysis, Plotly and Dash allowed for interactive dashboards, and
- False positives were a common issue for predictive models, but the collected data contains various attributes that lead to predictive success .

Introduction

- SpaceX has gained a lot of national attention due to their success as a private company in launching rockets into orbit. The goal is to be able to predict the success of their launches based on obtainable data.
- Can the success of the SpaceX launches be determined based upon publicly available information?
- What factors are important in determining successful launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected through API calls to the official SpaceX resources and webscraping from Wikipedia.
- Perform data wrangling
  - Standardization and mean values were used to clean a relatively complete dataset.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logisitic Regression, SVM, Decision Tree, and KNN models were tested using a train-test split of 0.2.

Data Collection

Data used in this exercise was retrieved from API calls to the official SpaceX API and web scraping performed on the SpaceX Wikipedia page.

API calls were REST calls which retrieved information in JSON format, which was reformatted using the Pandas and Numpy library, and saved as a CSV file.

Web Scraping was performed using the BeautifulSoup library, parsing information stored in HTML tables and converted into a CSV file.

Data Collection – SpaceX API

- Various REST call were made to https://api.spacexdata.com, receiving Launch Site, Booster Version, and more launch specific information in JSON format. This information was stored in CSV files for further use.

- https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- Using BeautifulSoup, web scraping was performed on the SpaceX Wikipedia page to retrieve information stored in the HTML based tables regarding launch information.

- https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

The "Outcome" feature, which contained information on the success/failure of the mission as well as location. This feature was simplified into a binary "Class" feature, which signified successful launches with a "1" value and failed launches with a "0" value.

https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Scatter plots were used to display the relationship "Flight Number", "Payload Mass", and "Launch Site". This allows us to see correlation between the displayed features, while also observing distribution and the result through color coding points on the plot.

A bar graph was used to observe the percentage of success for each "Orbit Type", taking advantage of the categorical nature of the "Orbit Type" feature.

A line graph was employed to observe the success rate of launches by year, allowing us to better observe trends over time.

EDA: https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Data Visualization: https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Display unique "Launch Site" values
- Display records where "Launch Site" begins with "CCA"
- Display total "Payload Mass"
- Display average "Payload Mass"
- Display first successful launch date
- List the boosters with successful launches with payloads between 4000 and 6000
- Display the total number of successful and unsuccessful launches
- Display list of boosters which have carried the maximum payload
- Display information regarding launches with a failed outcome involving a drone ship in 2015
- Rank the count of "Outcome" feature between June 4, 2010 and March 20, 2017

https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

On the Folium map, multiple types of markers were added:

- Marker Cluster: This marker allows one to display multiple markers that share the same coordinate, which when clicked can still display the information conveyed through each individual data point.
- Longitude-Latitude Tracker: This allows us to provide accurate location information based on mouse location, allowing greater interactivity.
- Line Marker: Lines were placed on the graph to display distance between launch sites and the coast, highways, and railways to show precautions for the launches.

https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

A pie graph was used to display the percentage of successful launches. This pie graph took the input of a dropdown menu, allowing one to observe the percentage of successful launches for each launch site, or what percentage of all successful launches came from which launch sites.

A scatterplot was used to display the correlation between payload and success for the selected launch site or all launch sites. This graph also color coded data points for booster type and allowed for a selection of the payload range observed.

https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Using a train-test split of 20%, four different models were trained: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

Each of these models were trained using GirdSearchCV, which explores a matrix of provided values to find the best hyperparameters for the model used.

https://github.com/akingsley319/IBM_Data_Science_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

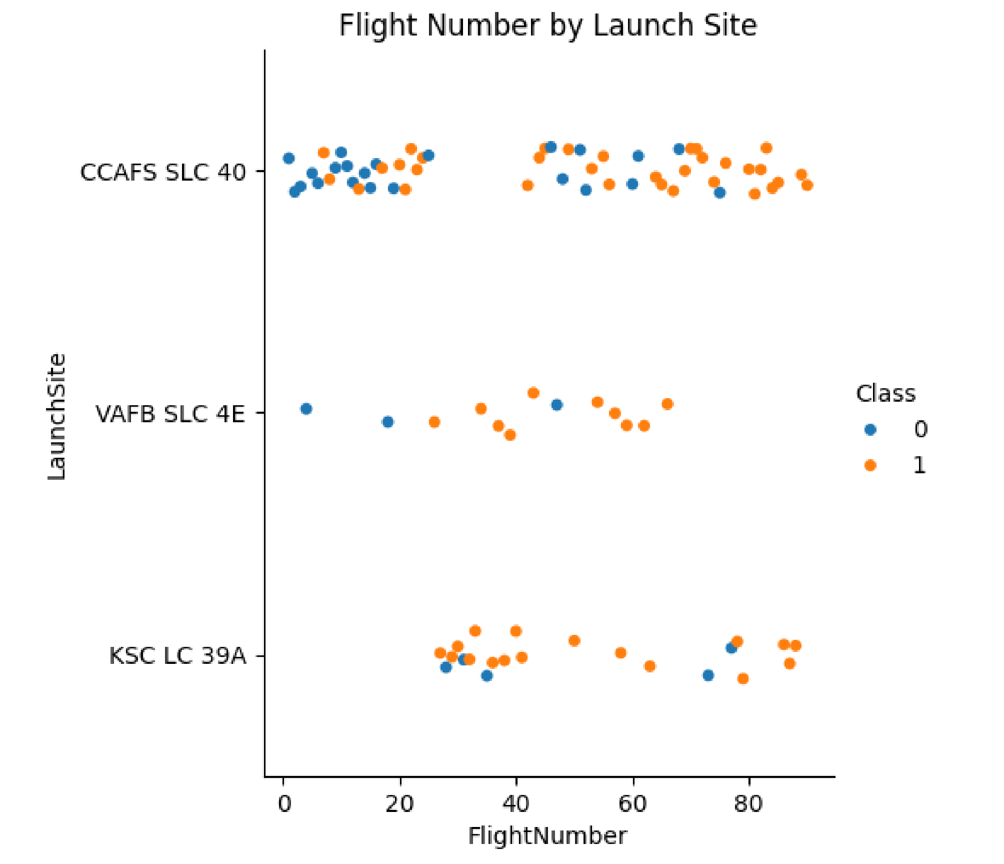
- Success Rates have improved over the years due to better booster versions and procedure at launch sites.
- Lower payloads and safer orbit types have resulted in better success rates.
- Certain launch sites have found greater amount of success, but it is not indicative of the individual launch site success rates.
- All models performed similarly, with False Positives being the source of error in the models.

Section 2

Insights drawn from EDA

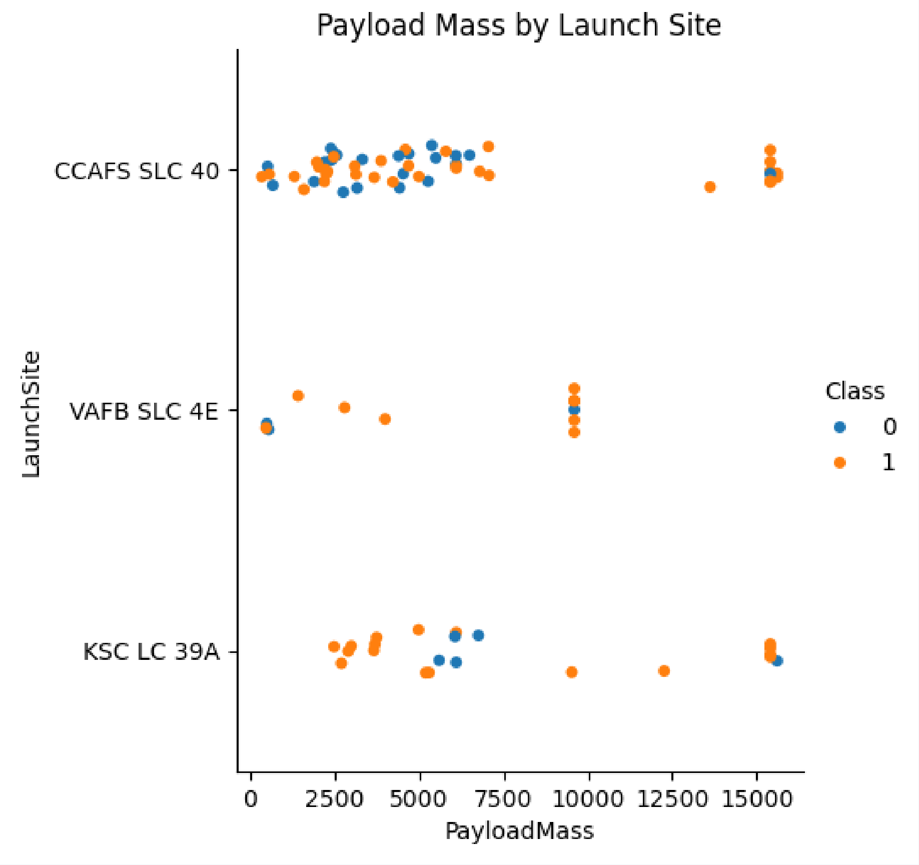
Flight Number vs. Launch Site

- "KSC LC 39A" involves higher "Flight Number", while "CCAFS SLC 40" is much more evenly distributed.
- Higher "Flight Number" results in more success.
- "VAFB SLC 4E" has only a few launches performed.



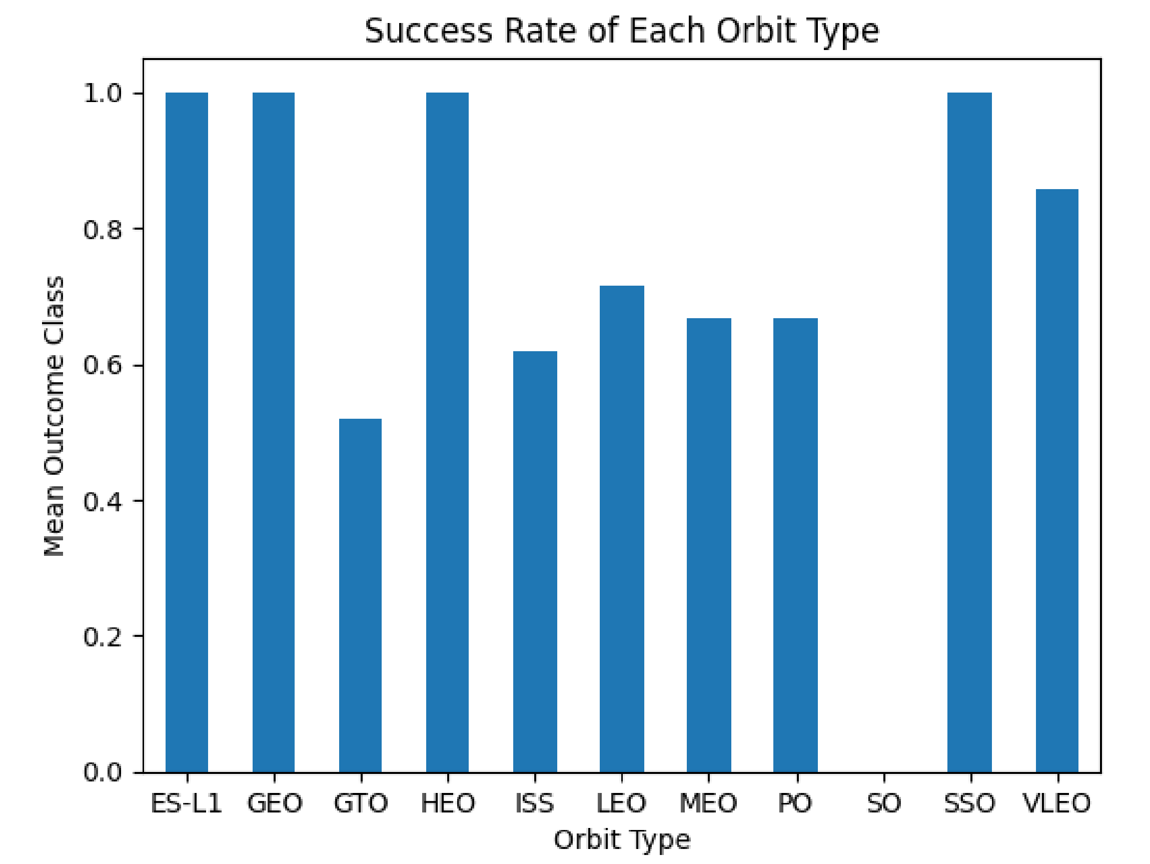
Payload vs. Launch Site

- Launch sites have what might be a payload maximum, which a number of tests are performed at, then the rest are performed at a range well below that maximum.
- Success seems to be well spread out, and cannot be determined by this graph.
- "VAFB SLC 4E" has the lowest payload test, while the other two sites are comparable in their launches.



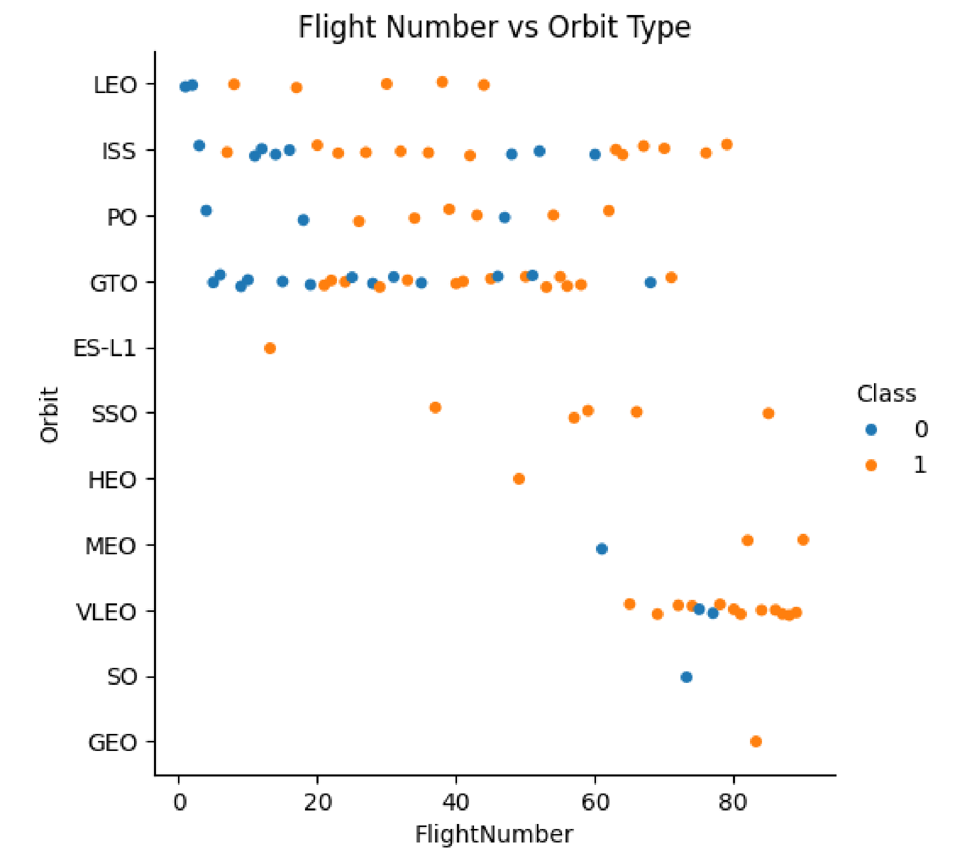
Success Rate vs. Orbit Type

- "SO" has no successes, which might be from lack of attempts or bad results.
- There are four orbit types which have a perfect success rate, which might be from lack of attempts or good results.
- Other orbit types sit between 50% and 70% success rate.



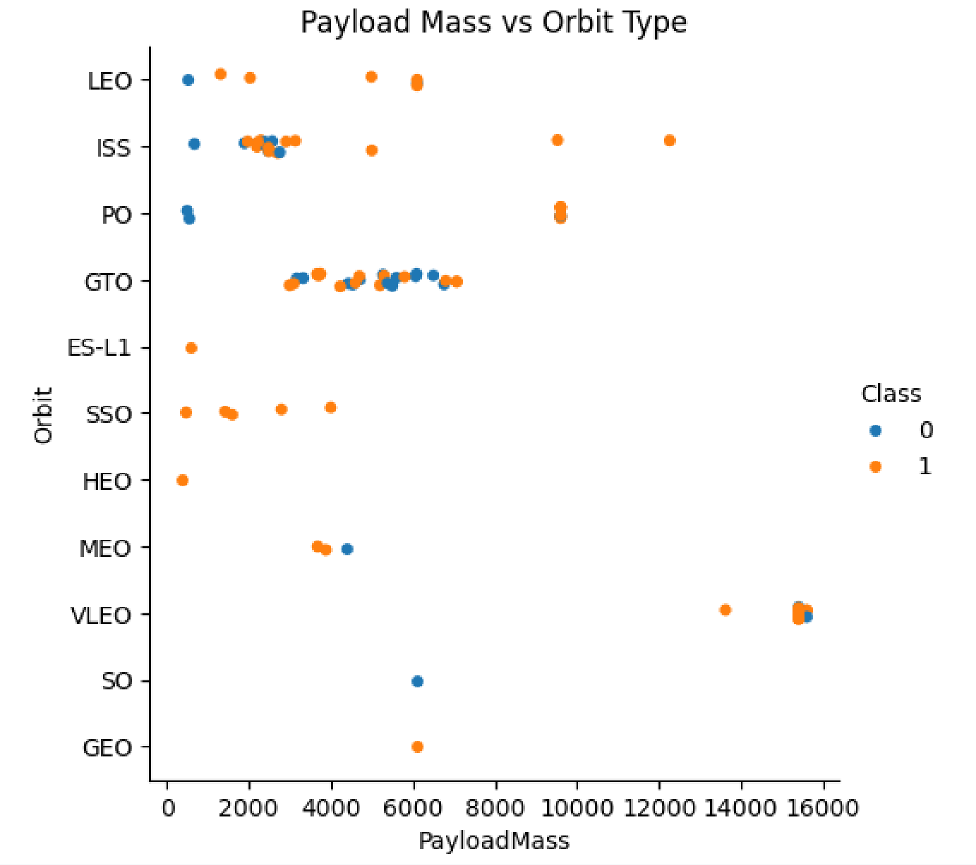
Flight Number vs. Orbit Type

- Higher Flight Number results in more success.
- Certain orbit types are more conducive to higher Flight Number



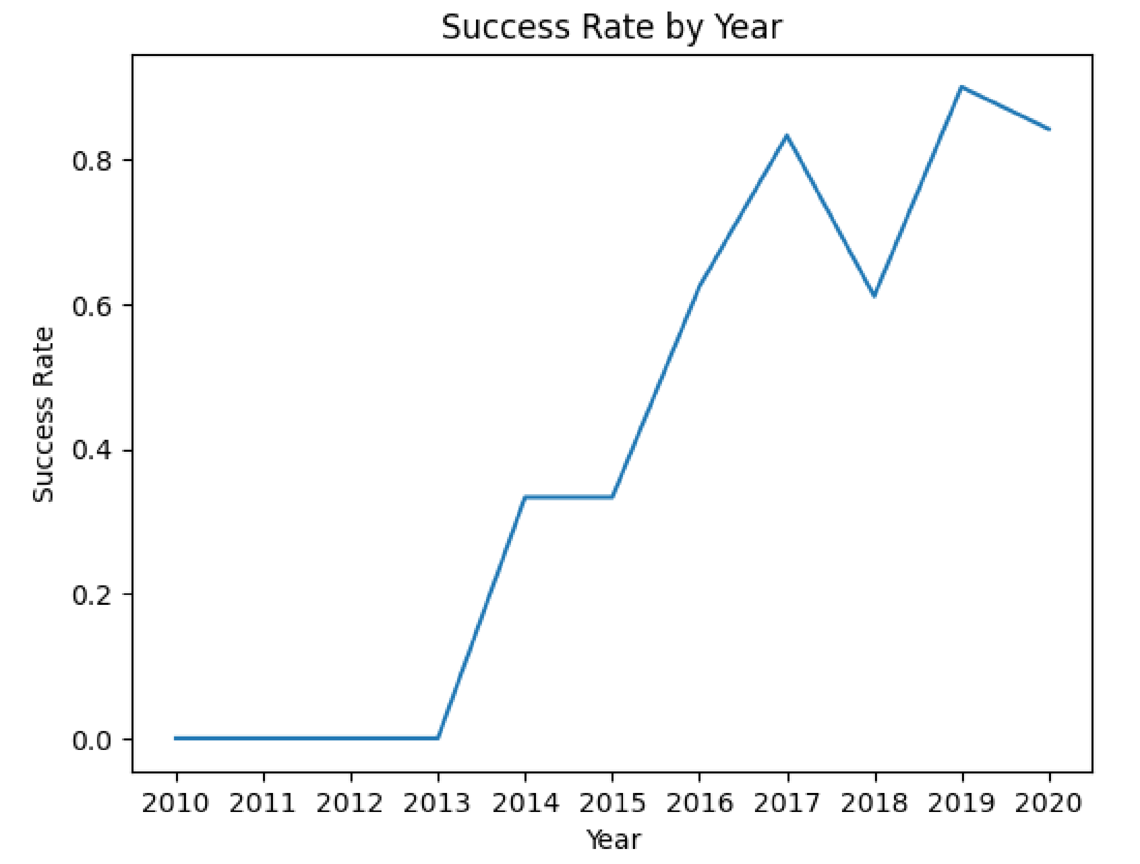
Payload vs. Orbit Type

- It is hard to determine correlation due to lack of data points.
- "VLEO" orbit has the highest payload mass used, while all other orbit types exist in the same range of payload mass.



Launch Success Yearly Trend

- More success is being found in each year.
- There is a dip in 2018, which would be worth looking into.



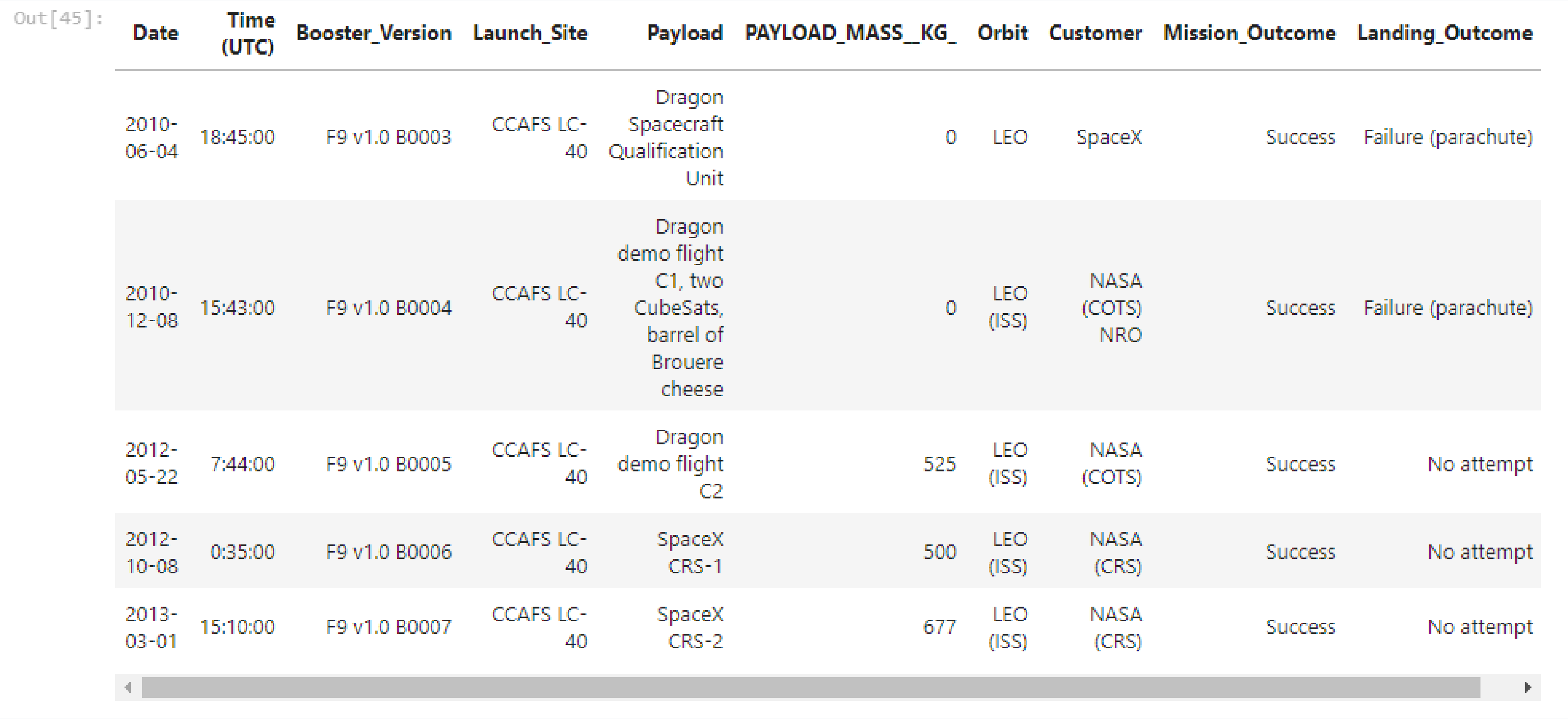
All Launch Site Names

The list of unique launch sites are:

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

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`



Out[45]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by boosters from NASA was 45596kg

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 was 2534.667kg.

First Successful Ground Landing Date

The first successful landing outcome on ground pad was performed on June 4, 2010.

Successful Drone Ship Landing with Payload between 4000 and 6000

The list of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

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

Boosters Carried Maximum Payload

The list of the boosters which have carried the maximum payload mass are:

- F9 B5 1048.4
- F9 B5 1048.5
- F9 B5 1049.4
- F9 B5 1049.5
- F9 B5 1049.7
- F9 B5 1051.3
- F9 B5 1051.4
- F9 B5 1051.6
- F9 B5 1056.4
- F9 B5 1058.3
- F9 B5 1060.2
- F9 B5 1060.3

Total Number of Successful and Failure Mission Outcomes

There were 100 Successful Outcomes and 1 Failure Outcomes, due to success not always meaning a successful launch.

2015 Launch Records

The list of failed landing_outcomes in drone ship, which includes their booster versions and launch site names for in year 2015 are:

- F9 v1.1 B1012, CCAFS LC-40, in January
- F9 v1.1 B1015, CCAFS LC-40, in April

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

The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 are:

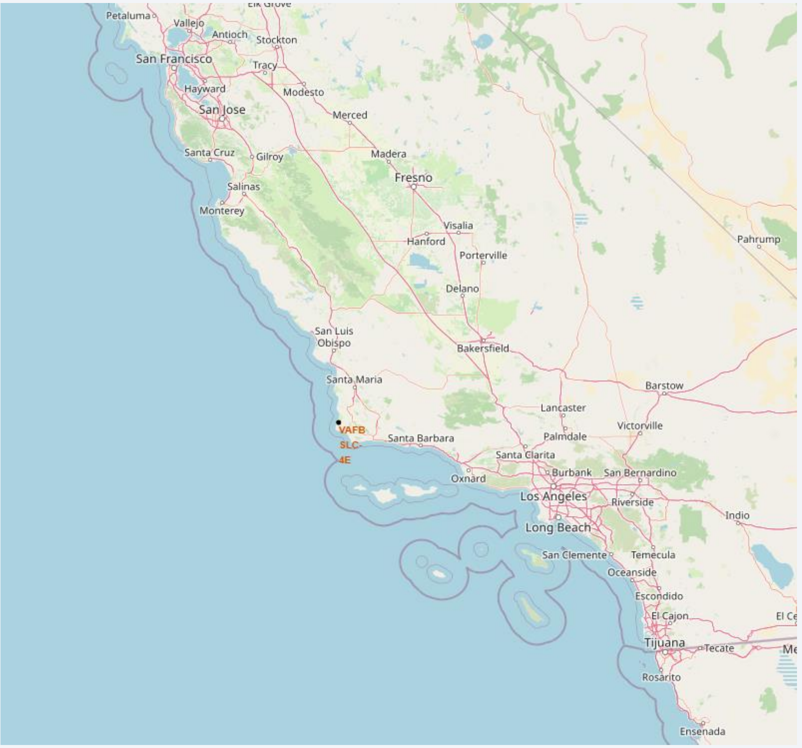
- No attempt: 10
- Success (drone ship): 5
- Failure (drone ship): 5
- Success (ground pad): 3
- Controlled (ocean): 3
- Uncontrolled (ocean): 2
- Failure (parachute): 2
- Precluded (drone ship): 1

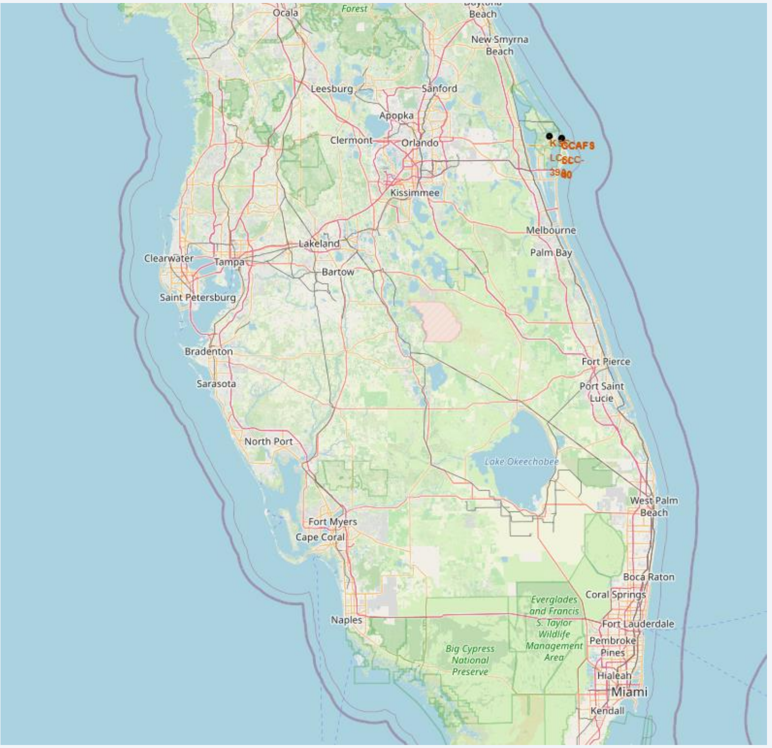
Section 4

Launch Sites Proximities Analysis

Launch Site Locations

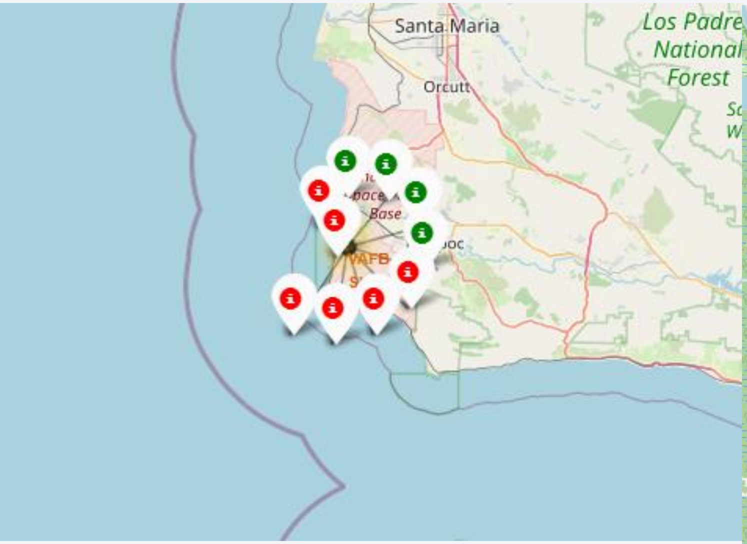
There is one launch site located in California, while 3 launch sites are located in the same area in Florida.

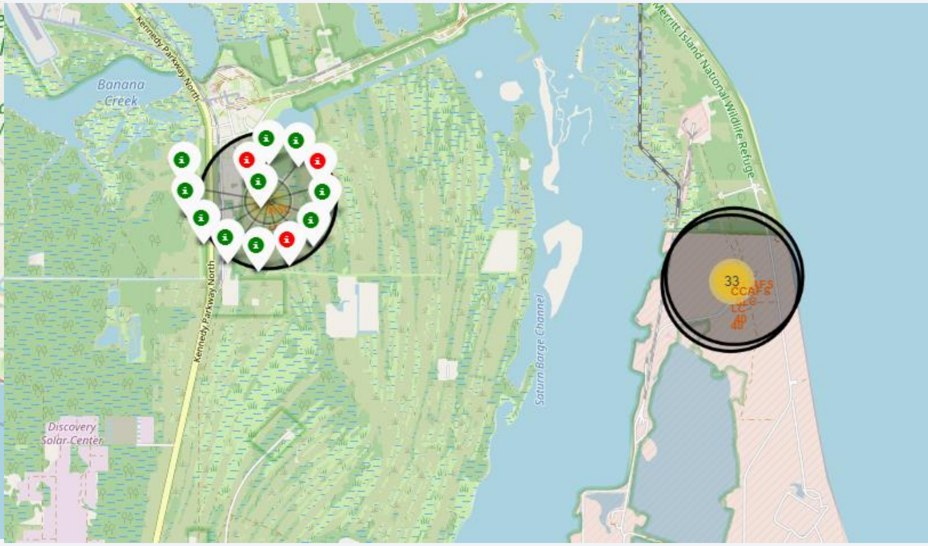




Marker Clusters

Using MarkerClusters, we can still display the result of launches at each location without stacking the markers. This allows the map to remain easy to read, interactive, and informative.

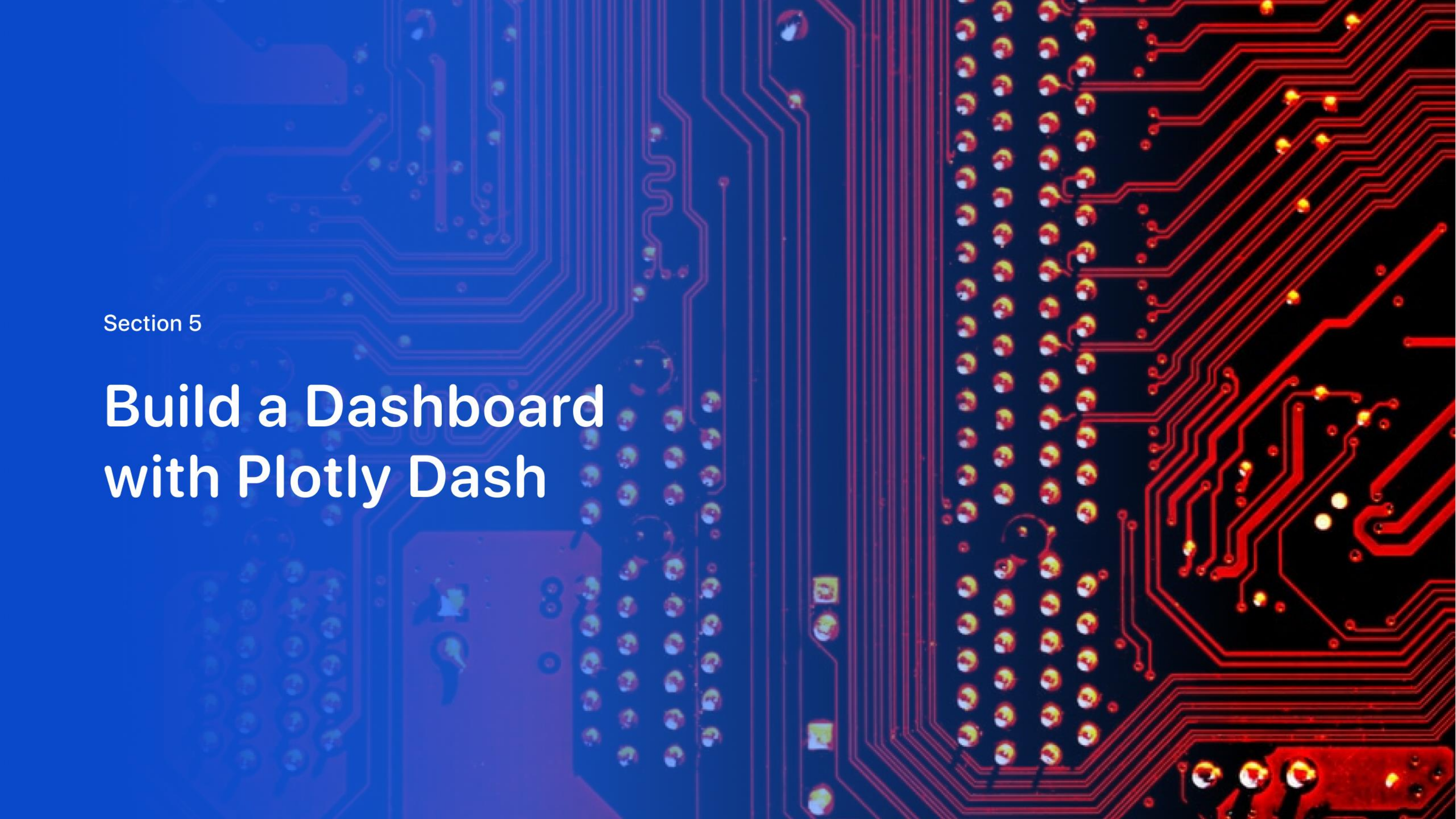




26
Centaur Road

Distance from Landmarks

I was unable to get a working map for this.

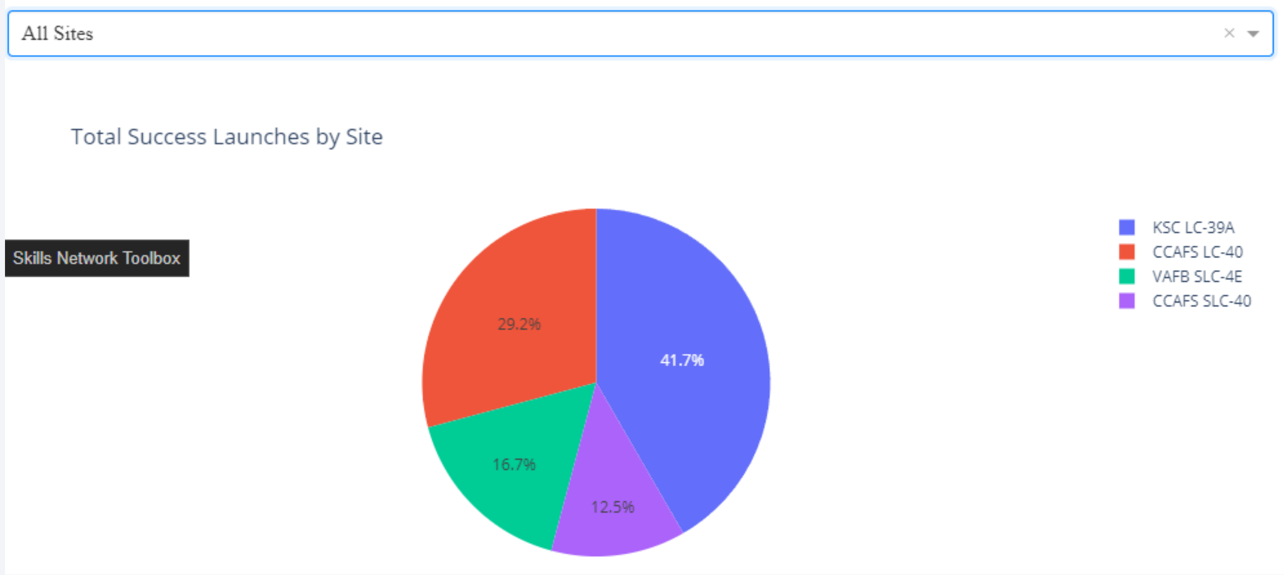


Section 5

Build a Dashboard with Plotly Dash

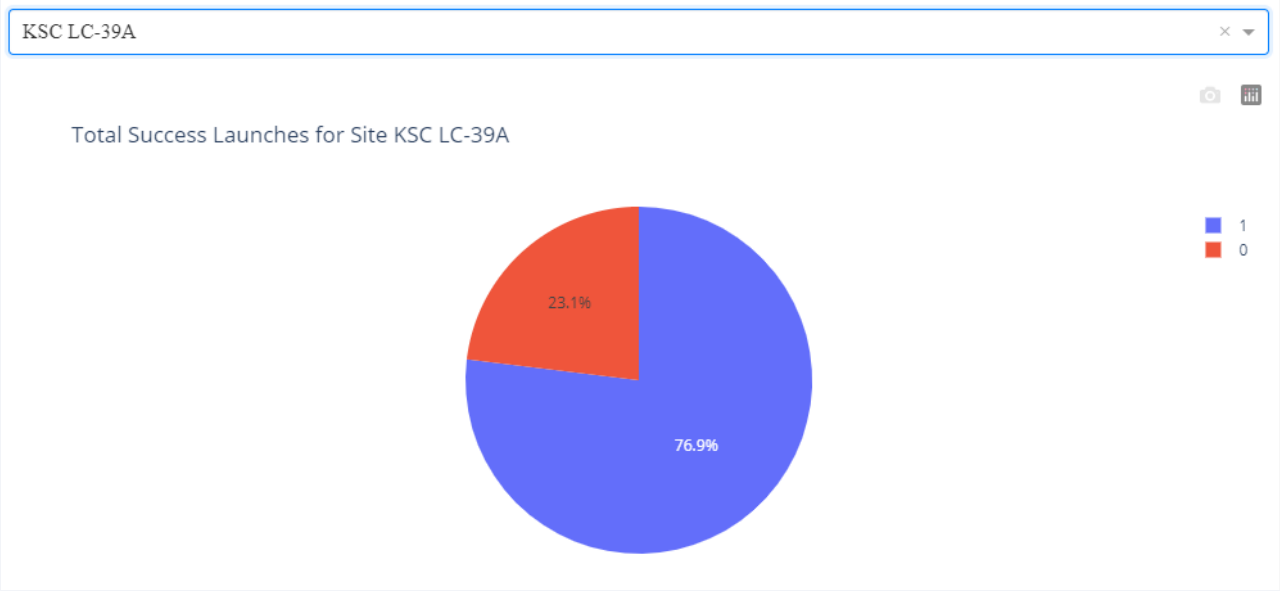
Total Successful Launches by Launch Site

By using a pie chart, we can see that most of the successful launches occurred at launch site "KSC LC-39A", while the least came from site "CCAFS SLC-40". However, we must keep in mind that this is not indicative of launch success rates.



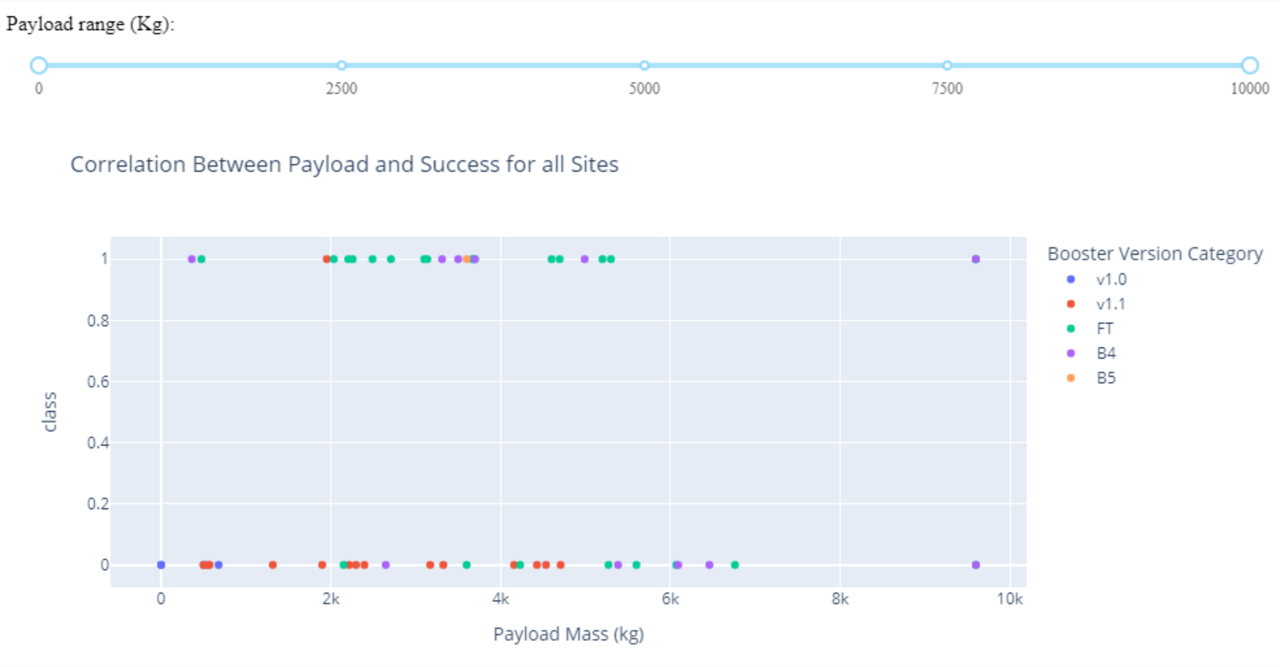
Highest Launch Success Rate

Launch site "KSC LC-39A" had the highest success rate of all launch sites at 76.9%, while "VAFB SLC-4E" had the lowest success rate at 40%.



Correlation Between Payload and Success for all Sites

We can see that the greatest success rates appear to come from payload masses between 2000kg and 4000 kg. FT boosters and B4 boosters also appear to have the most success of the booster versions.

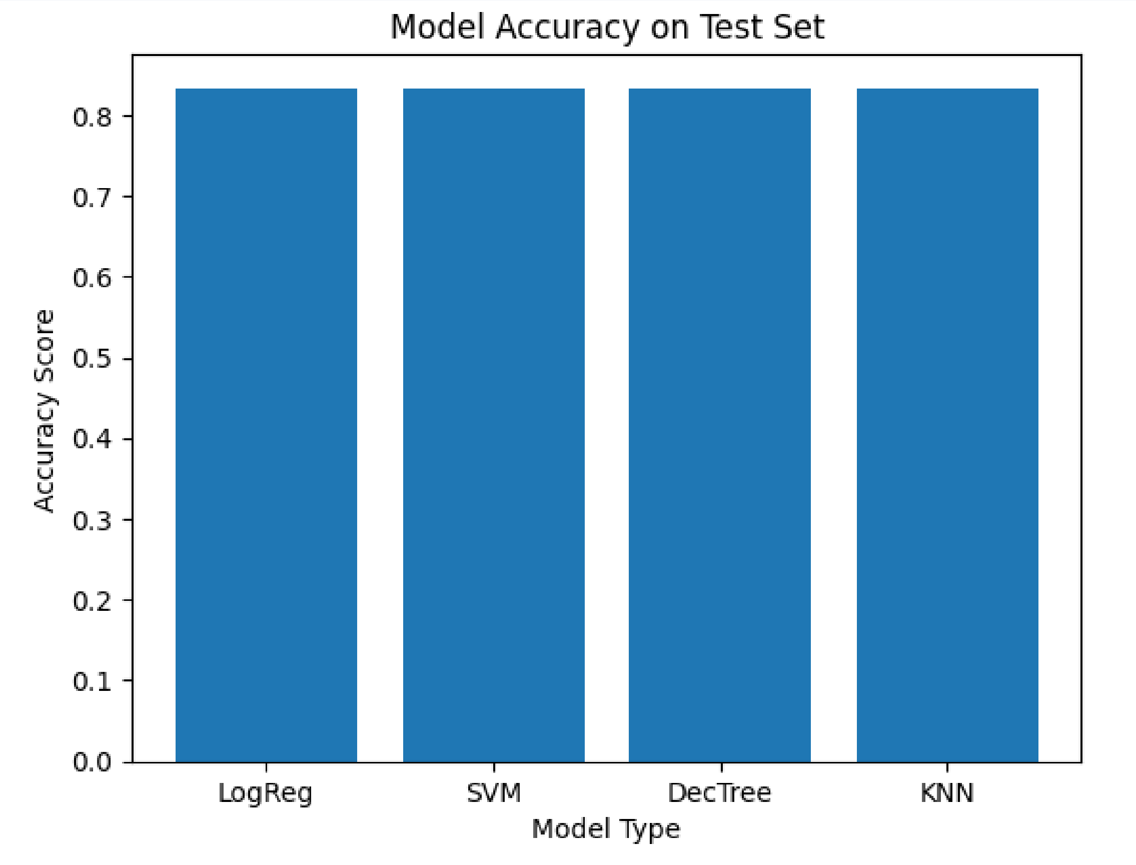


Section 6

Predictive Analysis (Classification)

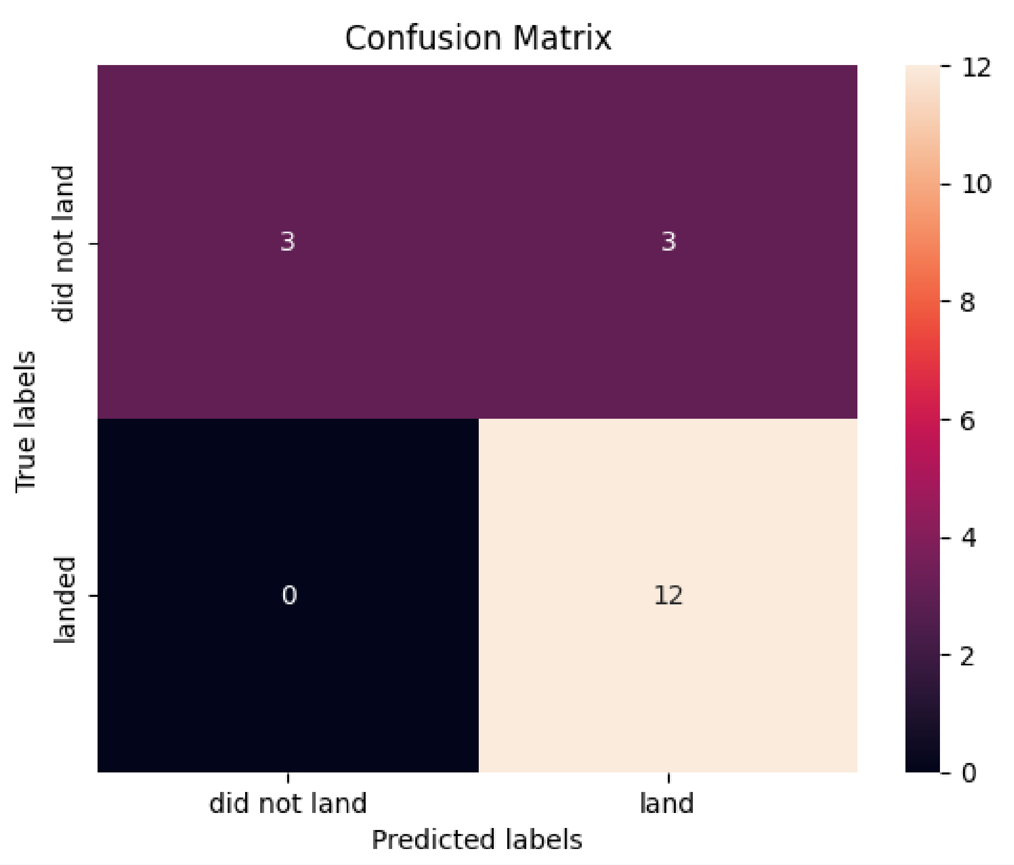
Classification Accuracy

- All models achieved the same accuracy on the test set.



Confusion Matrix

- Each model had the same looking confusion matrix, with problems involving False Positives.



Conclusions

- Payload, Orbit Type, Booster Version, and Flight Number have an effect on the success of the launch.
- While the Launch Site does appear to have effects on the success on launches, this might be as a result of other factors that are a result of resources or location.
- Current predictive models have issues with false positives, which will be a focus of improvement in future study.

Appendix

The full GitHub can be found at:

https://github.com/akingsley319/IBM_Data_Science_Capstone/

Thank you!