IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Praful Aggarwal
21-Sep-2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Collect data for SpaceX Falcon 9 rockets using API and Web Scraping
  - Data wrangling and EDA is performed using Pandas and SQL
  - Launch site visualization is performed using Folium
  - An interactive dashboard is built using Plotly Dash
  - Finally, machine learning prediction is used to determine if the Falcon 9 will land successfully or not.

- **Summary of all results**
  - Overall, the SpaceX data is good with minimal cleaning to be done.
  - We observe interesting relationships between Payload and the Launch Site.
  - Launch Site success class shows how well the landing performs at the different sites.
  - All machine learning algorithms perform equally while distinguishing between the different classes.

Introduction

- **Project background and context**
  - I am a Data Scientist for Space Y, a competitor for SpaceX.
  - My goal is to use SpaceX data to analyze the Falcon 9 rocket landing.
  - This data and project will allow my team to make more informed bids for rocket launch against SpaceX.
  - In this project, I'll be using Python to collect, wrangle, visualize the SpaceX data.
  - I will then use machine learning approaches to build and evaluate models for rocket landing predictions.
- **Problems you want to find answers**
  - How is the SpaceX Falcon 9 rocket launch data quality?
  - Identify any correlation between the Launch Site, Payload and landing success?
  - Predict the successful landing of the first stage of the Falcon 9 rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected using Space X API and web-scraping
- Perform data wrangling
  - Data was analyzed and cleaned (as needed) using Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four models were used, Logistic Regression, SVM, Decision Tree and KNN.
  - The analysis was done using the scikit-learning library.
  - The models were evaluated using different parameters to find the best hyperparameter and accuracy scores.
  - Confusion matrix was used to visualize the prediction.

Data Collection

Describe how data sets were collected.

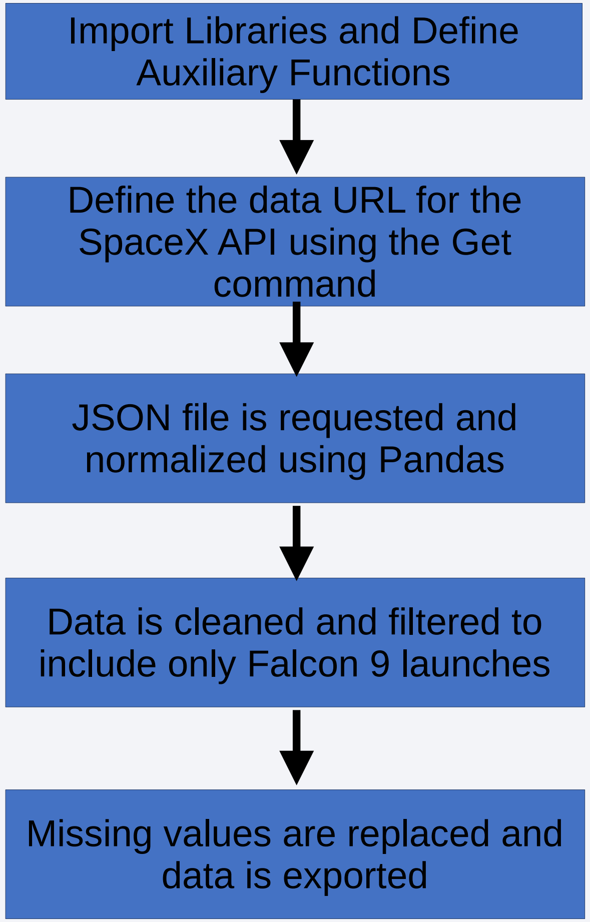
- Requests were made to the SpaceX API.
- Used web scraping to extract records HTML table from Wikipedia

Data Collection – SpaceX API

- SpaceX REST API was used to collect the data. The key steps are shown in the flowchart

- GitHub URL

  https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/Data%20Collection%20API.ipynb



Import Libraries and Define Auxiliary Functions

Define the data URL for the SpaceX API using the Get command

JSON file is requested and normalized using Pandas

Data is cleaned and filtered to include only Falcon 9 launches

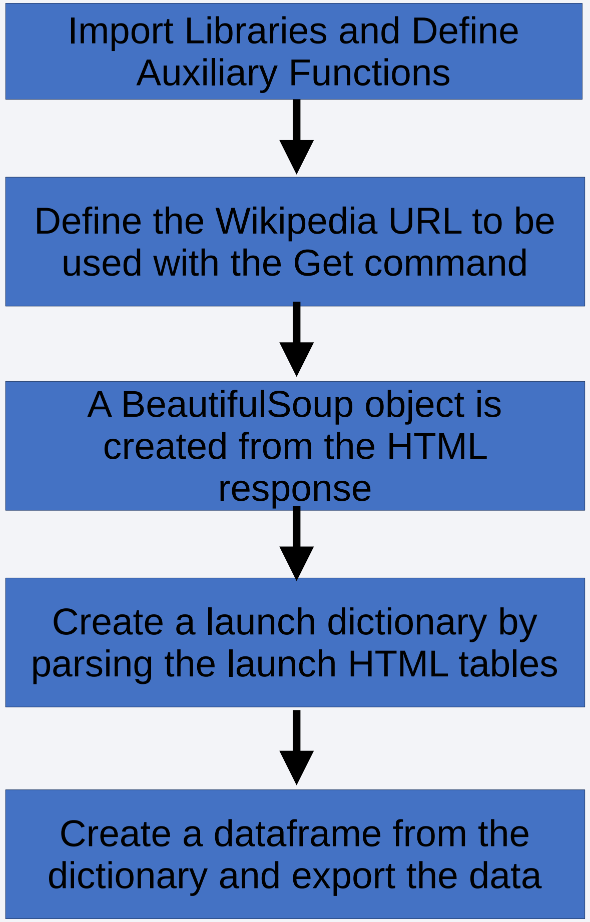
Missing values are replaced and data is exported

Data Collection - Scraping

- Web scraping is done to collect historical launch data from Wikipedia

- GitHub URL

  https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/Data%20Collection%20web%20scraping.ipynb



Import Libraries and Define Auxiliary Functions

Define the Wikipedia URL to be used with the Get command

A BeautifulSoup object is created from the HTML response

Create a launch dictionary by parsing the launch HTML tables

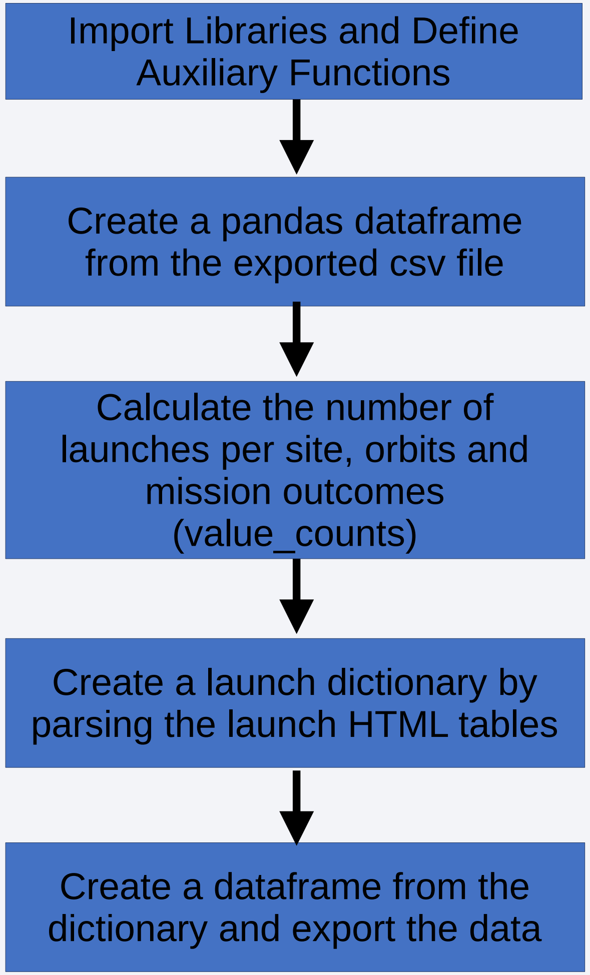
Create a dataframe from the dictionary and export the data

Data Wrangling

- Data wrangling and cleaning is done using the Pandas library

- GitHub URL

  https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/Data%20Wrangling.ipynb



Import Libraries and Define Auxiliary Functions

Create a pandas dataframe from the exported csv file

Calculate the number of launches per site, orbits and mission outcomes (value_counts)

Create a launch dictionary by parsing the launch HTML tables

Create a dataframe from the dictionary and export the data

EDA with Data Visualization

Summary

- Scatter plots are created for Payload vs. Flight Number and Launch Site vs. Flight Number.
  - These plots were generated to visualize the relationship between these features.
- Barplot was created on the categorical Orbit variable to compare the success rate between the different Orbits.
- Line chart was created between Success rate and Launch year to observe the improvement in success rate with increasing launch years.

Github URL:

https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/EDA%20with%20Data%20visualization.ipynb.jupyterlite.ipynb

EDA with SQL

Summary

- SQL queries to identify and filter launch site names.
- Queries to perform calculations (total, average) on payload mass.
- Queries were run to look at launch success and failures including first launch, boosters used, and total number of launches.
- Finally queries were run to identify the landing outcomes and perform calculations on them.

Github URL:

https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

Summary

- Folium circles with markers were created for the unique launch sites .
- Marker clusters were created for each site highlighting the success and failure launches
- A polyline was created for a proximity point from a launch site
- These objects were added for improved visualization of the launch sites and their success/failure rate.

Github URL:

https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/Interactive%20visual%20analytics%20with%20Folium.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Summary

- Create a dashboard using Dash to generate a drop down list, a pie chart, a slider and a correlation scatter plot.
- The pie chart is an efficient way to visualize category wise fraction of a parameter.
- The scatter plot is the preferred graph to visualize correlation between multiple variables.

Github URL

https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/SpaceY_Dash_app.py

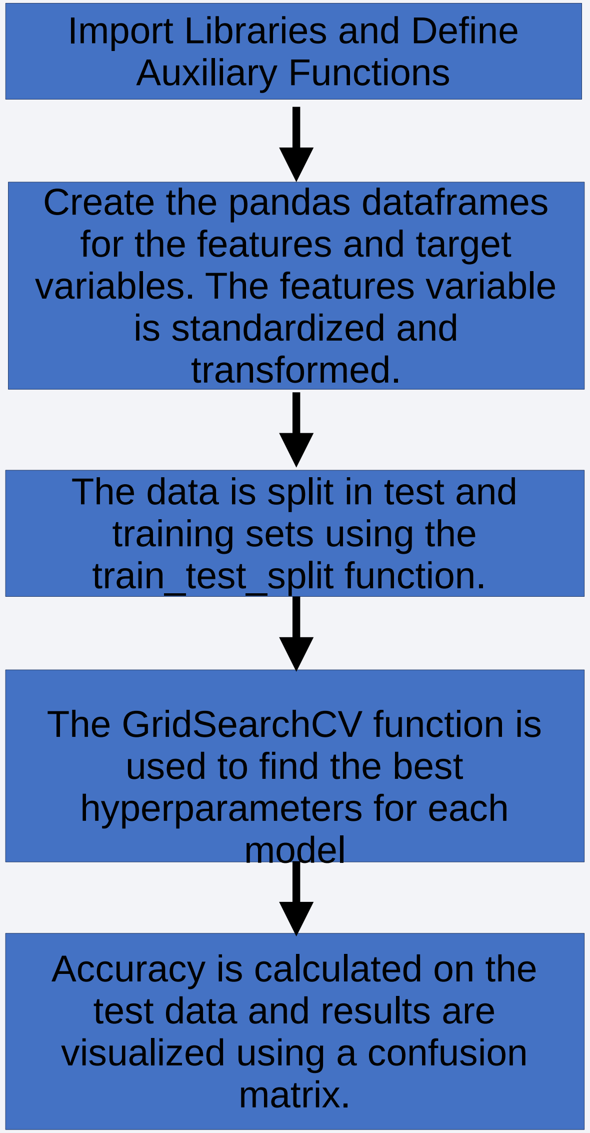
Predictive Analysis (Classification)

Summary

- Predict the success of the Falcon 9 landing
- Use predictive modeling (Logistic Regression, SVM, Decision Tree & KNN)
- Find the best hyperparameters and model
- The models were built using the different algorithms in the scikit-learn library

GitHub URL

https://github.com/aggp11/SpaceY-Capstone/blob/Final-assignment-submission/Machine%20Learning%20Prediction.ipynb



Import Libraries and Define Auxiliary Functions

Create the pandas dataframes for the features and target variables. The features variable is standardized and transformed.

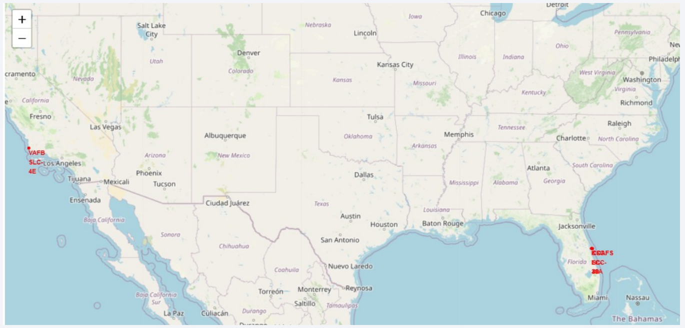
The data is split in test and training sets using the train_test_split function.

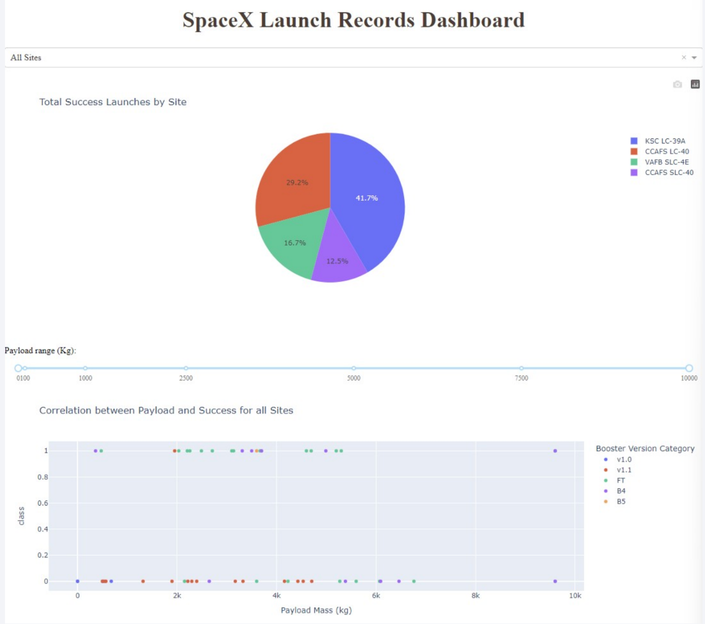
The GridSearchCV function is used to find the best hyperparameters for each model

Accuracy is calculated on the test data and results are visualized using a confusion matrix.

Results

- Exploratory data analysis results
  - Observed trends and correlation between the different attributes
  - SQL was used to query the data to answer specific questions
- Interactive analytics demo in screenshots
  - See dashboard and folium examples in the right panel
- Predictive analysis results
  - All four models performed equally with the accuracy of 83.333%




SpaceX Launch Records Dashboard

All Sites

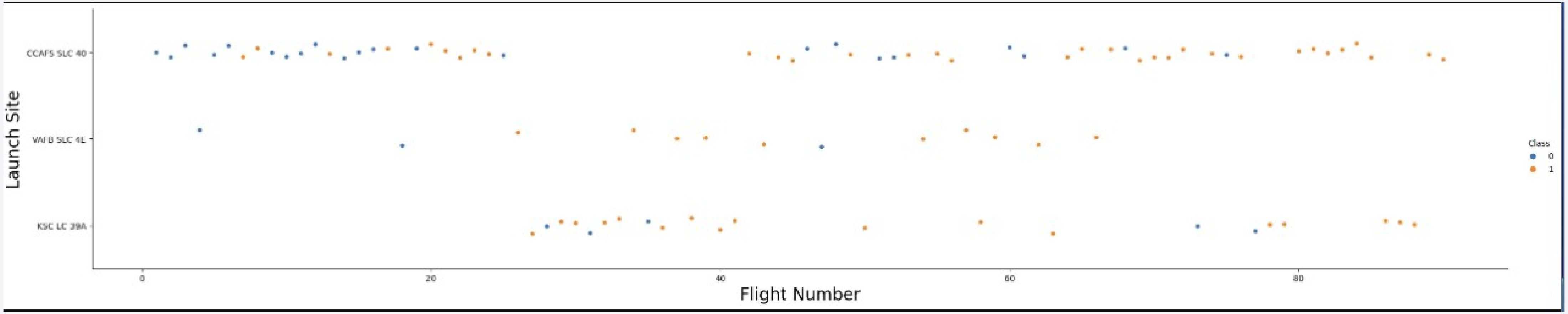
Payload range (Kg):

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

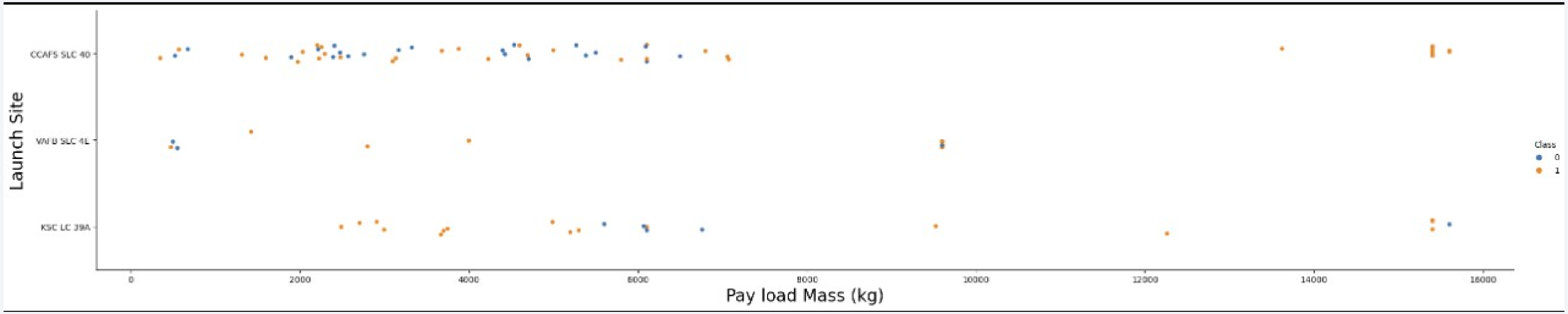
- Scatter plot for Launch Site vs. Flight Number



- Launch Site CCAFS SLC 40 has the largest number of launches
  - The success rate for the launches improved as more flights were launched

Payload vs. Launch Site

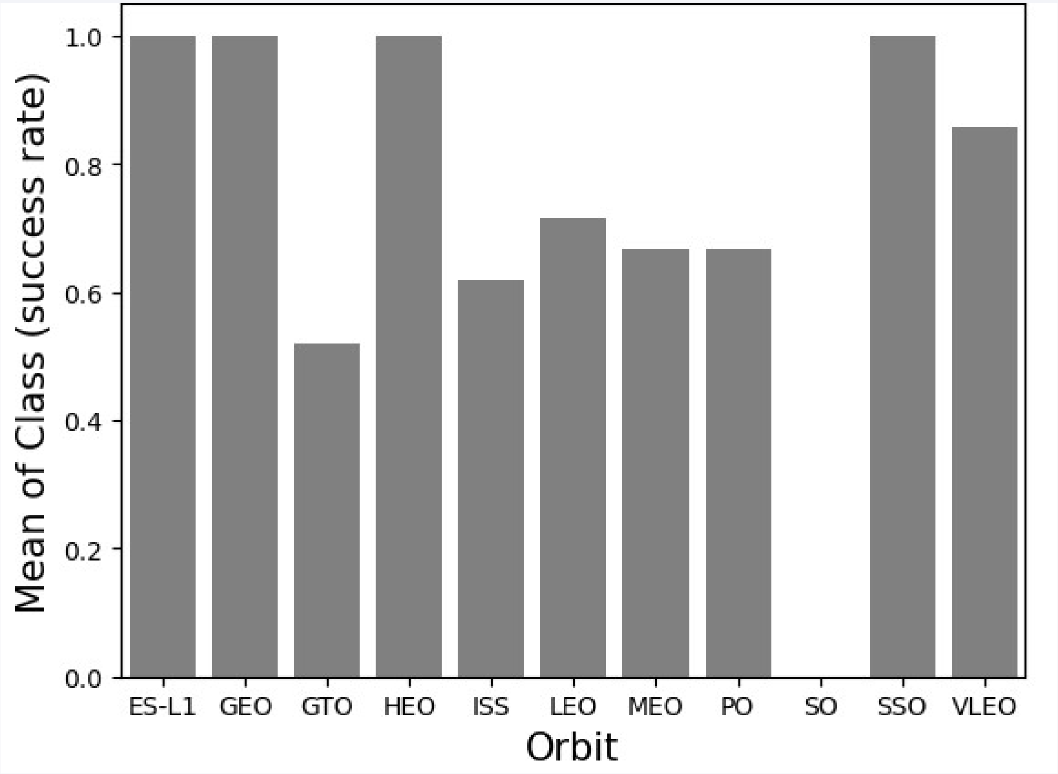
- Scatter plot for Launch Site vs. Pay load Mass (kg)



- Majority of the flights carried under 8,000 kg payloads
- Almost all the payload flights from the different sites were all successful

Success Rate vs. Orbit Type

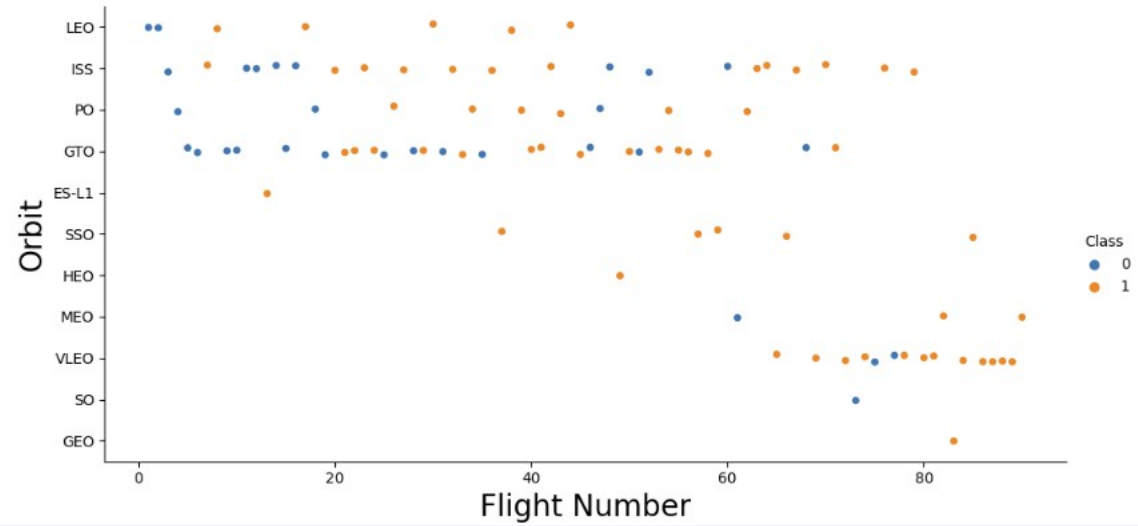
- Barchart for Success Rate vs. Orbit type



- Orbit types ES-L1, GFO, HEO and SSO have perfect success rate.
- While Orbit SO had no successful outcomes

Flight Number vs. Orbit Type

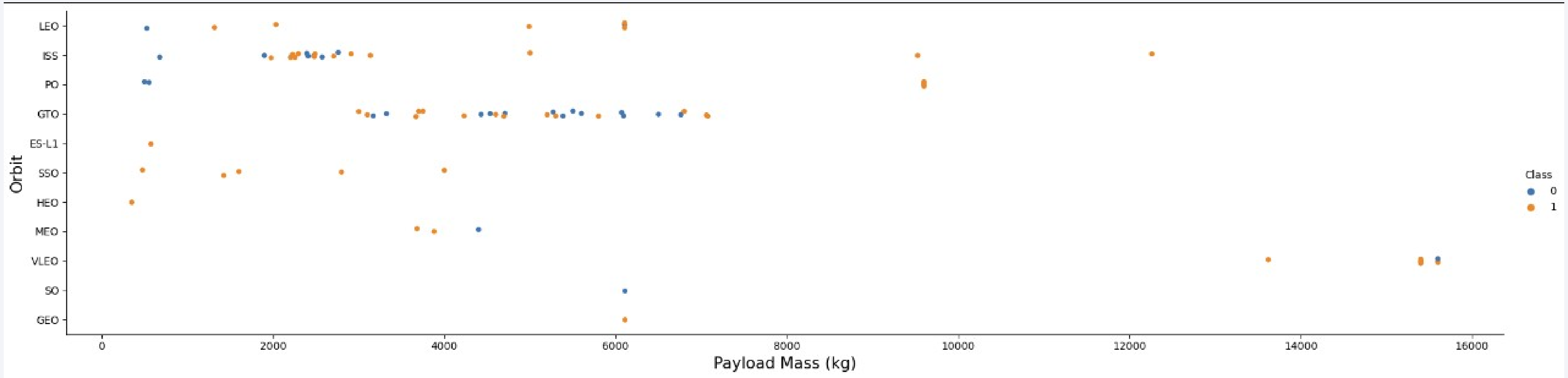
- Scatter plot for Orbit type vs. Flight number



- Later flights (higher flight numbers) were mainly sent to the VLEO orbit.

Payload vs. Orbit Type

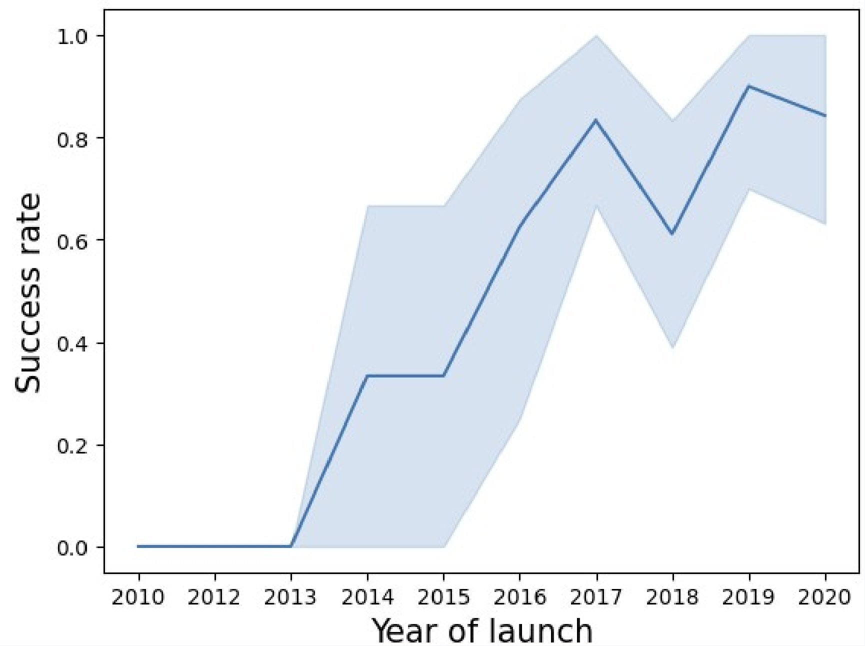
- Scatter plot for Orbit type vs. Pay load Mass (kg)



- High payload flights were primarily sent to the VLEO orbit

Launch Success Yearly Trend

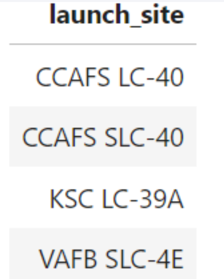
- Line chart with yearly success trend



- Since 2013, the success rate has mainly increased with a flat line in 2015 and slight decreases in 2018 and 2020.

All Launch Site Names

Launch site names



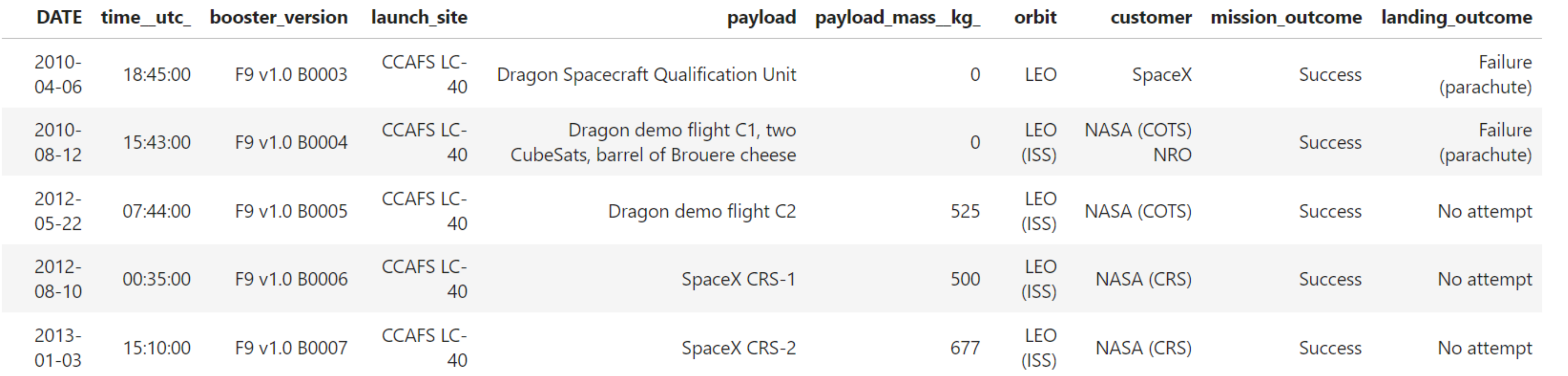
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

There are four unique launch sites. Used Distinct function.

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`



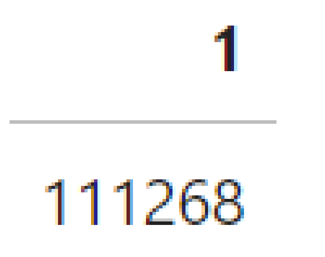
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

These are five records where the launch site begins with `CCA`. Used Like and Limit commands.

```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

The total payload mass (kg) carried by boosters launched by NASA (CRS) is 111,268 kg



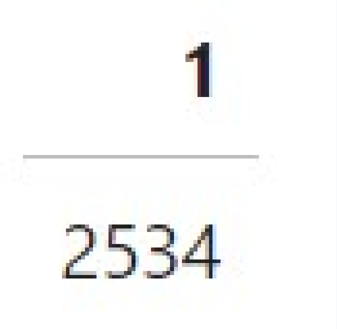
1
111268

Used Like and Sum functions

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE PAYLOAD LIKE '%CRS%';
```

Average Payload Mass by F9 v1.1

The average payload carried by booster version F9 v1.1 is 2,534 kg



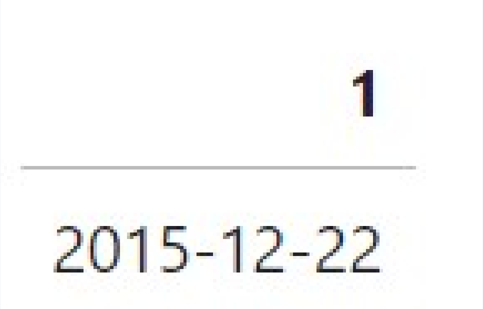
1
2534

Used Avg and Like functions

%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION LIKE '%F9 v1.1%';

First Successful Ground Landing Date

The first successful landing outcome on ground pad happened on 22-Dec-2015



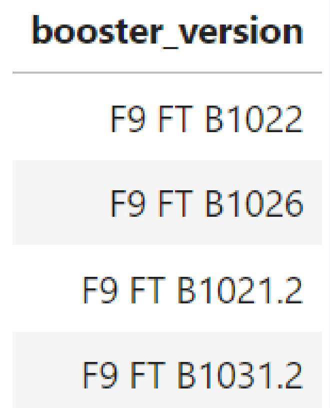
1
2015-12-22

Used Min and Like functions

%sql SELECT MIN(DATE) FROM SPACEX WHERE LANDING_OUTCOME LIKE '%Success%ground%';

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



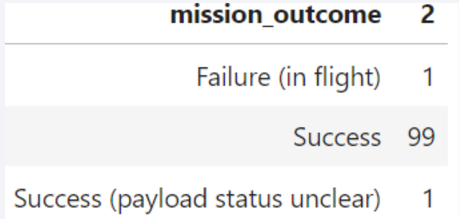
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Four booster versions have successfully landed on drone ship with payloads between 4000 and 6000kg. Used the between function to be inclusive of the weight limits.

```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE (PAYLOAD_MASS__KG_ BETWEEN 4001 AND 5999) AND LANDING_OUTCOME LIKE 'Success%drone%';
```

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes



mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

All but one mission were successful.

Used the count function on mission outcome to calculate the number of successful and failed missions.

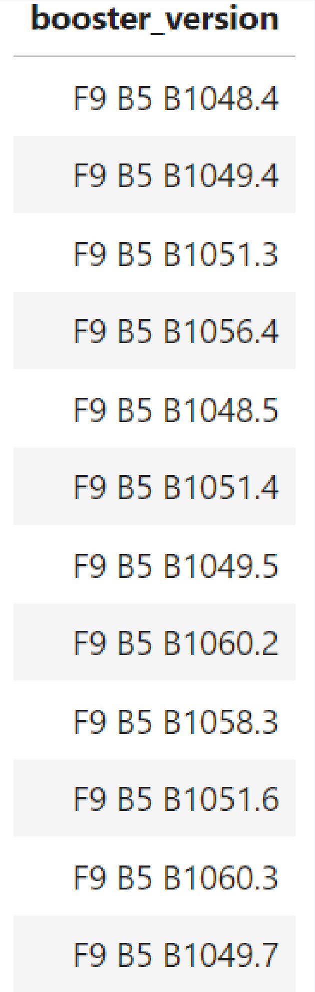
%sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEX GROUP BY MISSION_OUTCOME;

Boosters Carried Maximum Payload

There are 12 booster versions that carried maximum payload mass

A sub-query was used to calculate the max payload mass

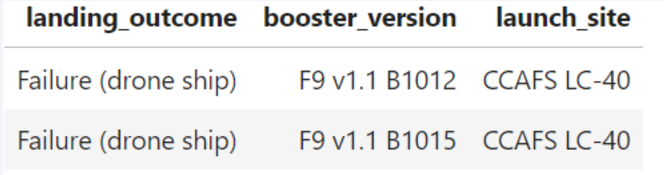
```
%sql SELECT BOOSTER_VERSION FROM SPACEX
WHERE PAYLOAD_MASS__KG_ =(SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```



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

2015 Launch Records

Two failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed below



landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Year function on date was used to filter results

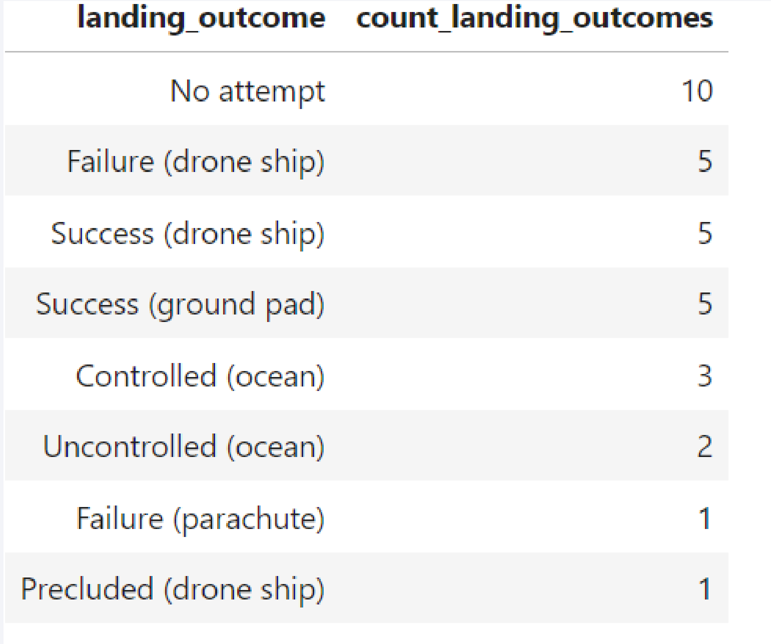
```
%sql SELECT LANDING_OUTCOME, BOOSTER_VERSION,
LAUNCH_SITE FROM SPACEX WHERE YEAR(DATE)='2015' AND
LANDING_OUTCOME LIKE 'Failure%drone%';
```

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

Ranked landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are shown in the table on the right.

Group by, order by and desc functions were used to extract these results.

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS
"COUNT_LANDING_OUTCOMES" FROM SPACEX
WHERE (DATE(DATE) BETWEEN '2010-06-04' AND
'2017-03-20') GROUP BY LANDING_OUTCOME
ORDER BY COUNT_LANDING_OUTCOMES DESC;
```



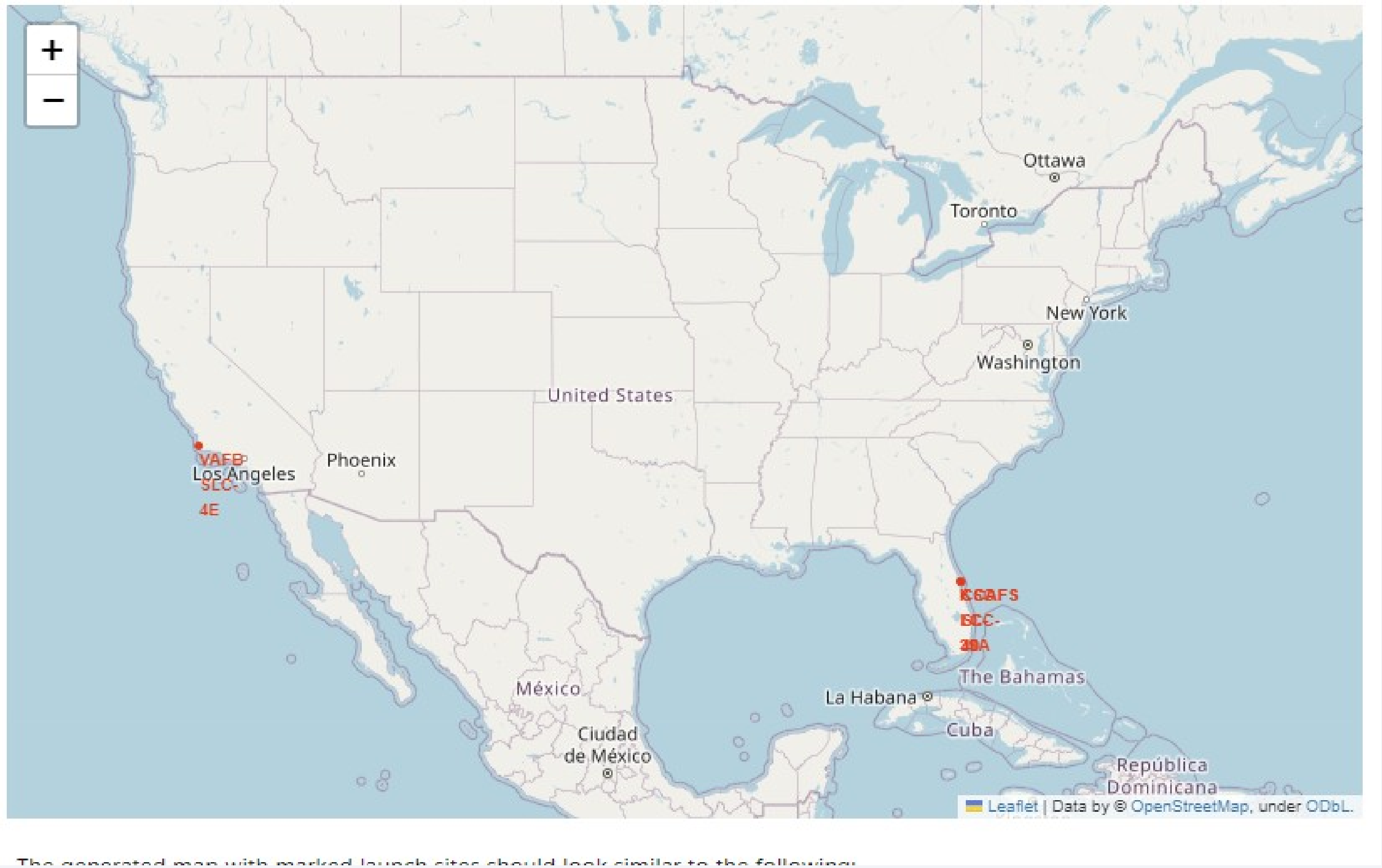
landing_outcome	count_landing_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Success (ground pad)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	1
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch site locations on a global map

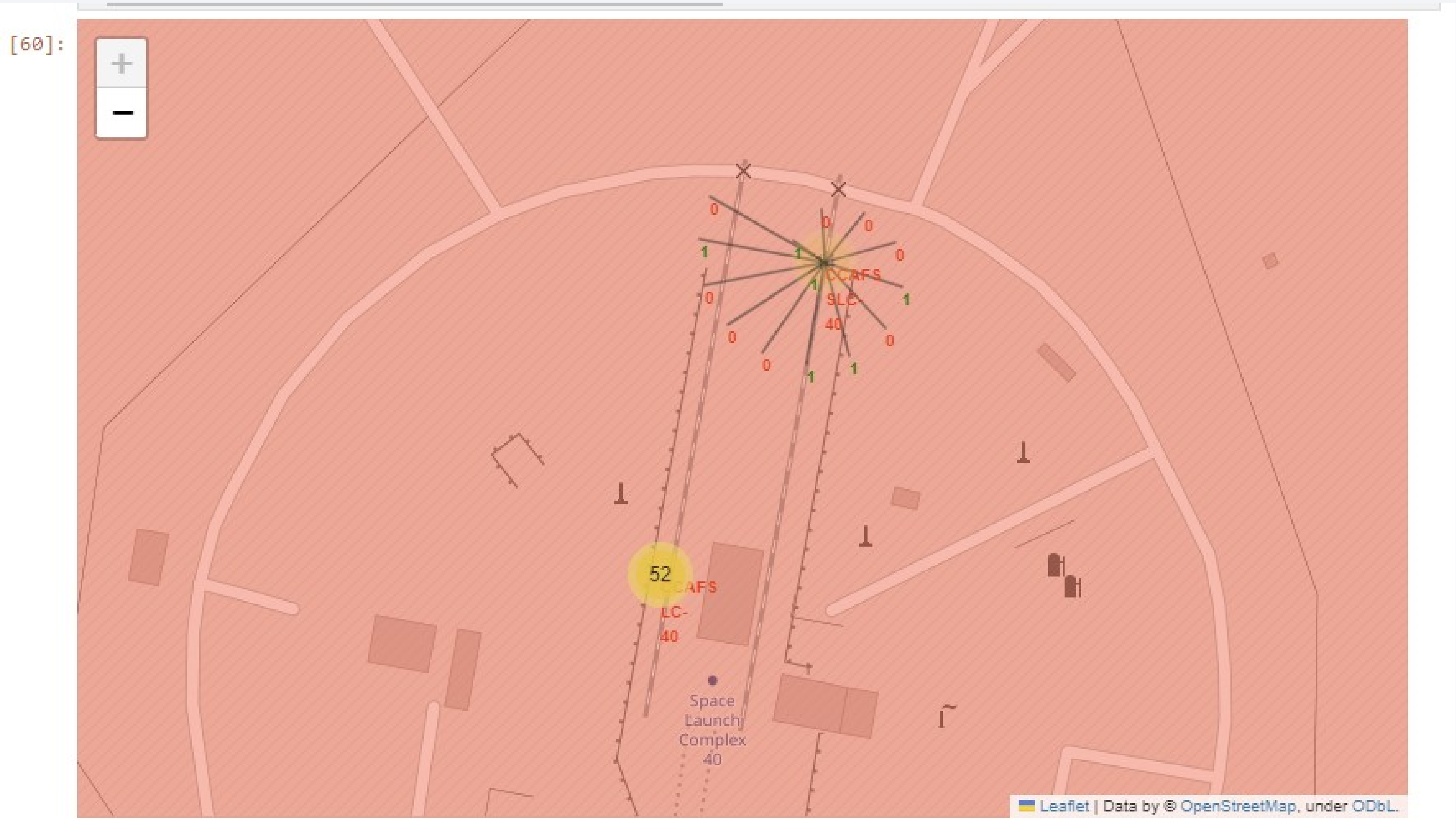
Circles and markers (with icon) are used to show all the launch sites



The generated map with marked launch sites should look similar to the following:

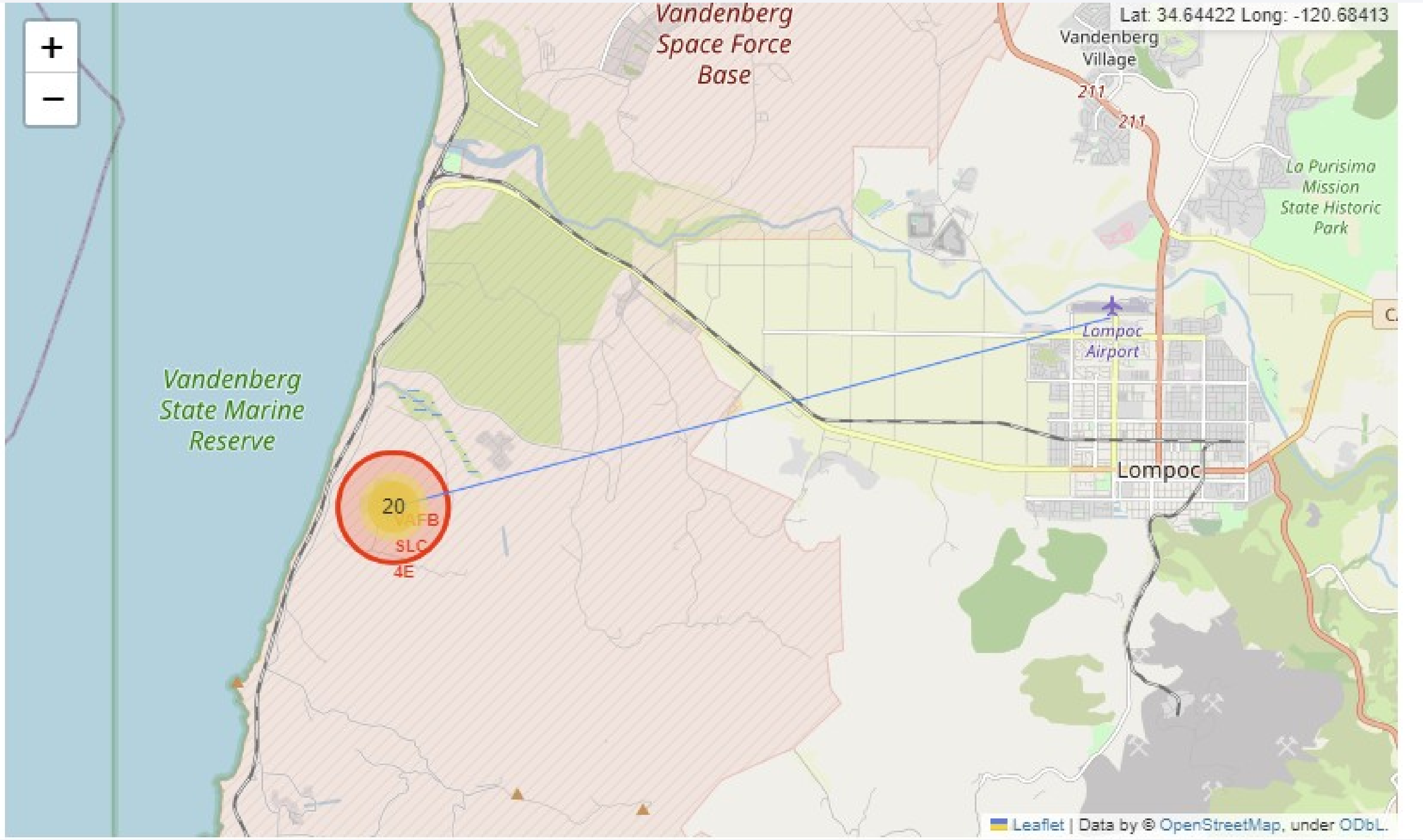
Successful (green) and failed (red) launch outcomes

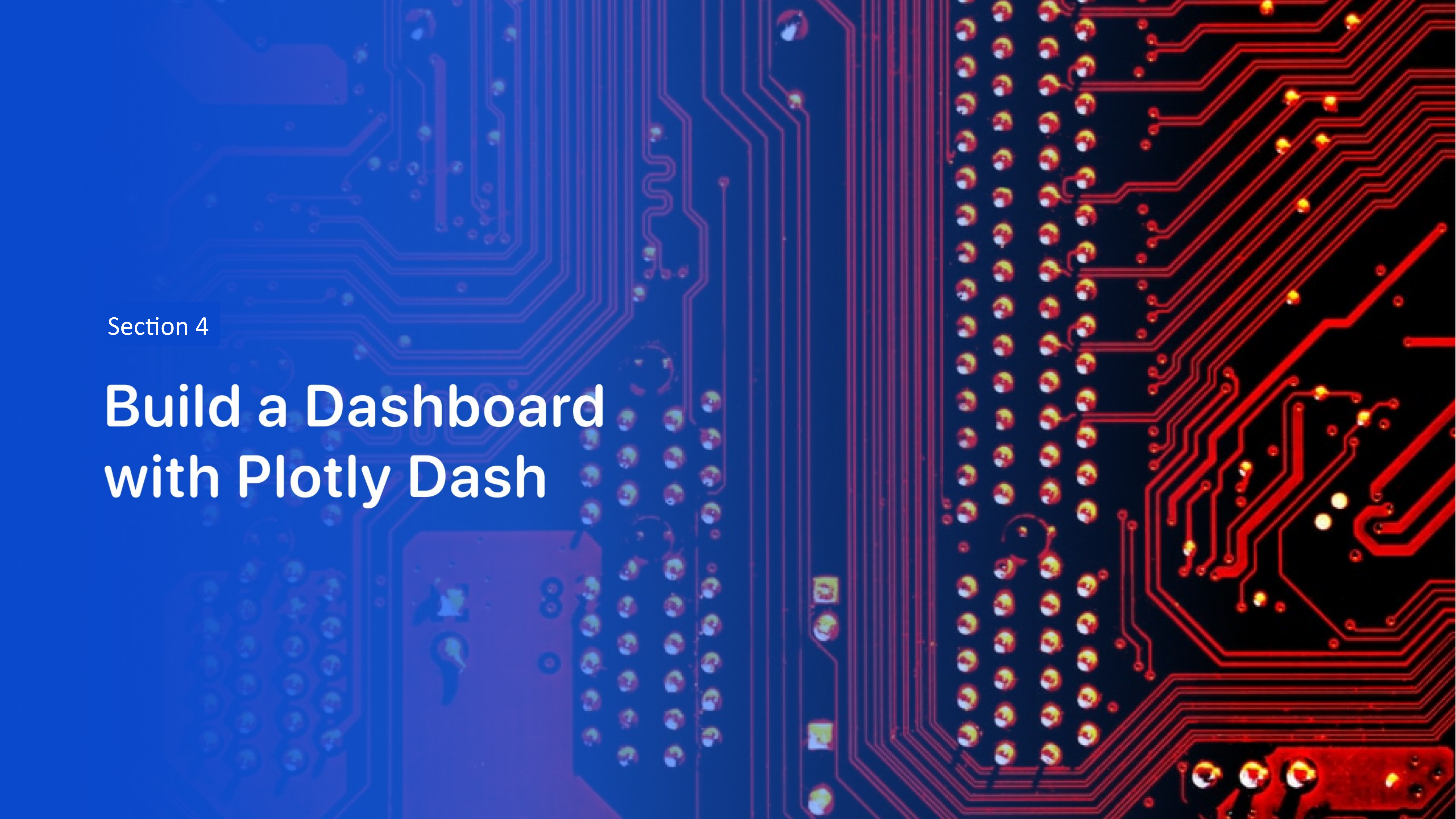
Marker clusters and outcome class are used to generate this visual



Proximity distance line plot

A polyline is shown to the nearest airport to site VAFB SLC 4E


+
−
Lat: 34.64422 Long: -120.68413
Vandenberg Space Force Base
Vandenberg Village
211
211
La Purisima Mission State Historic Park
Lompoc Airport
Lompoc
Vandenberg State Marine Reserve
20
VAFB SLC 4E
Leaflet | Data by © OpenStreetMap, under ODbL.



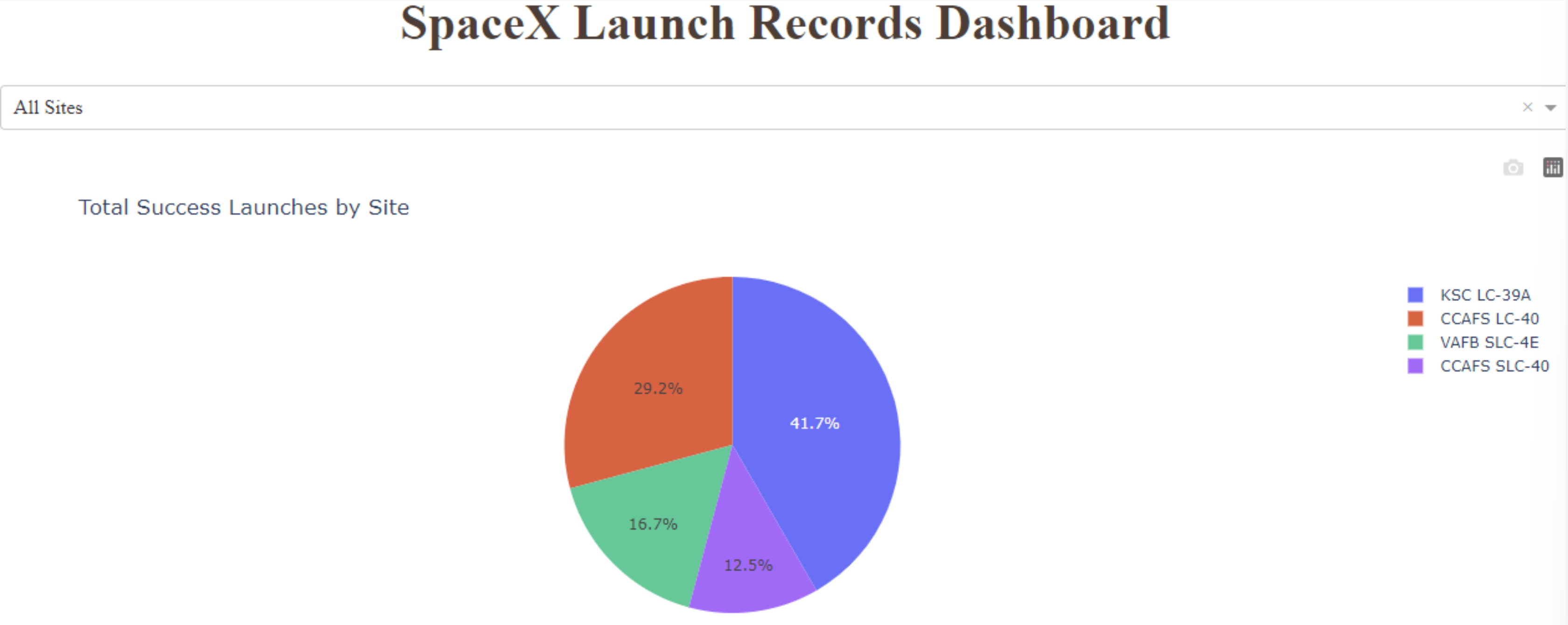
Section 4

Build a Dashboard with Plotly Dash

All sites pie-chart

The dropdown menu has “All sites” selected

Site KSC LC-39A has the highest proportion of successful launches



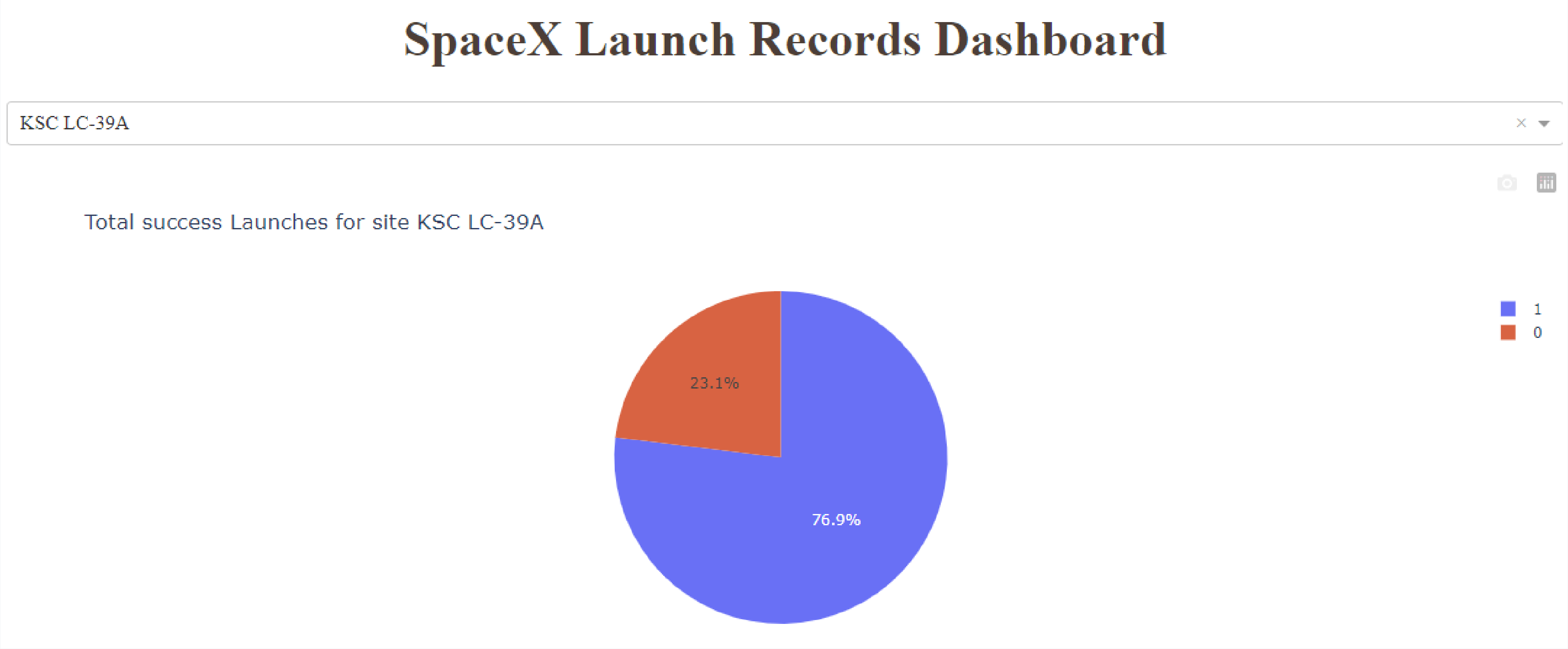
SpaceX Launch Records Dashboard

All Sites

Piechart for site KSC LC-39A

The pie chart shows the proportion of successful and failed launches at this site.

Over 3/4th of the launches were successful at this site



SpaceX Launch Records Dashboard

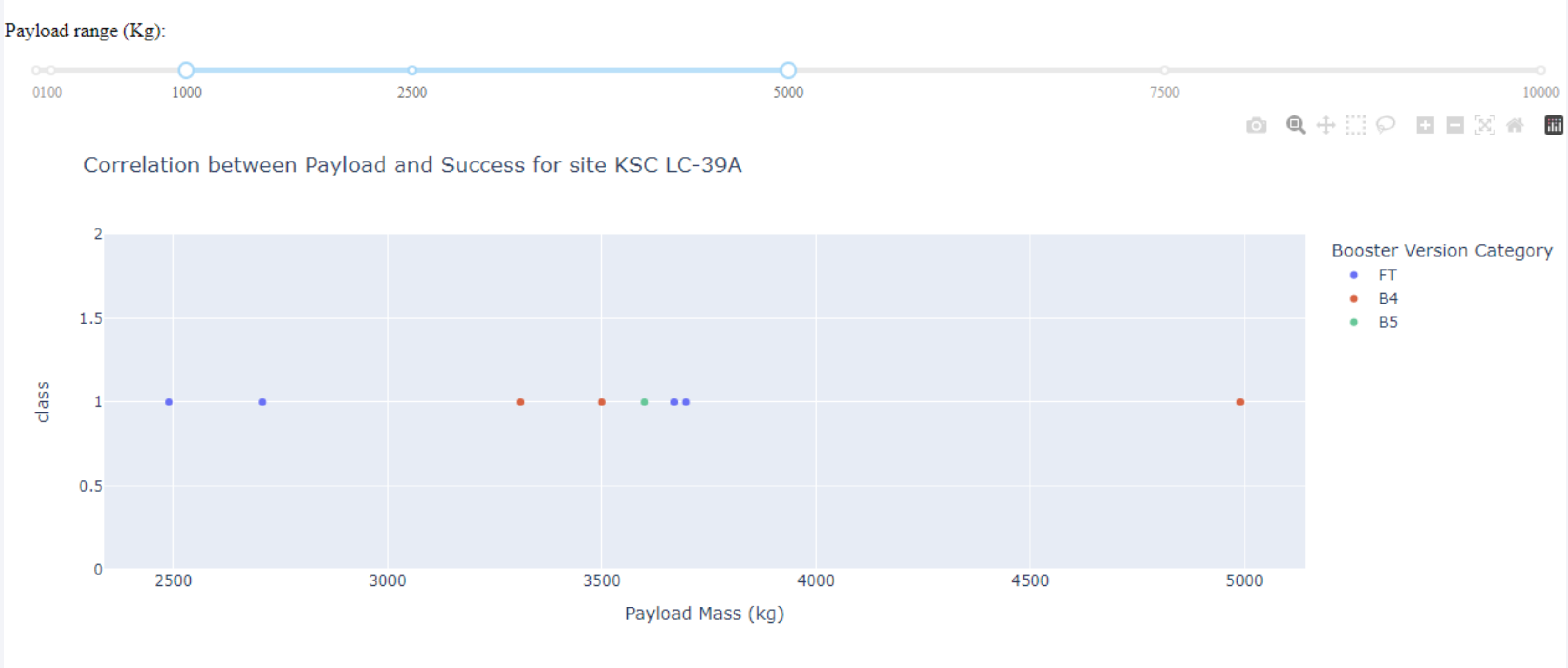
KSC LC-39A

Total success Launches for site KSC LC-39A

Launch outcome vs. Payload

Used the payload range slider and drop down menu to show launch outcome vs. payload scatter plot.

This plot shows that all payloads at site KSC LC-39A with payload between 1000 and 5000kgs were successful

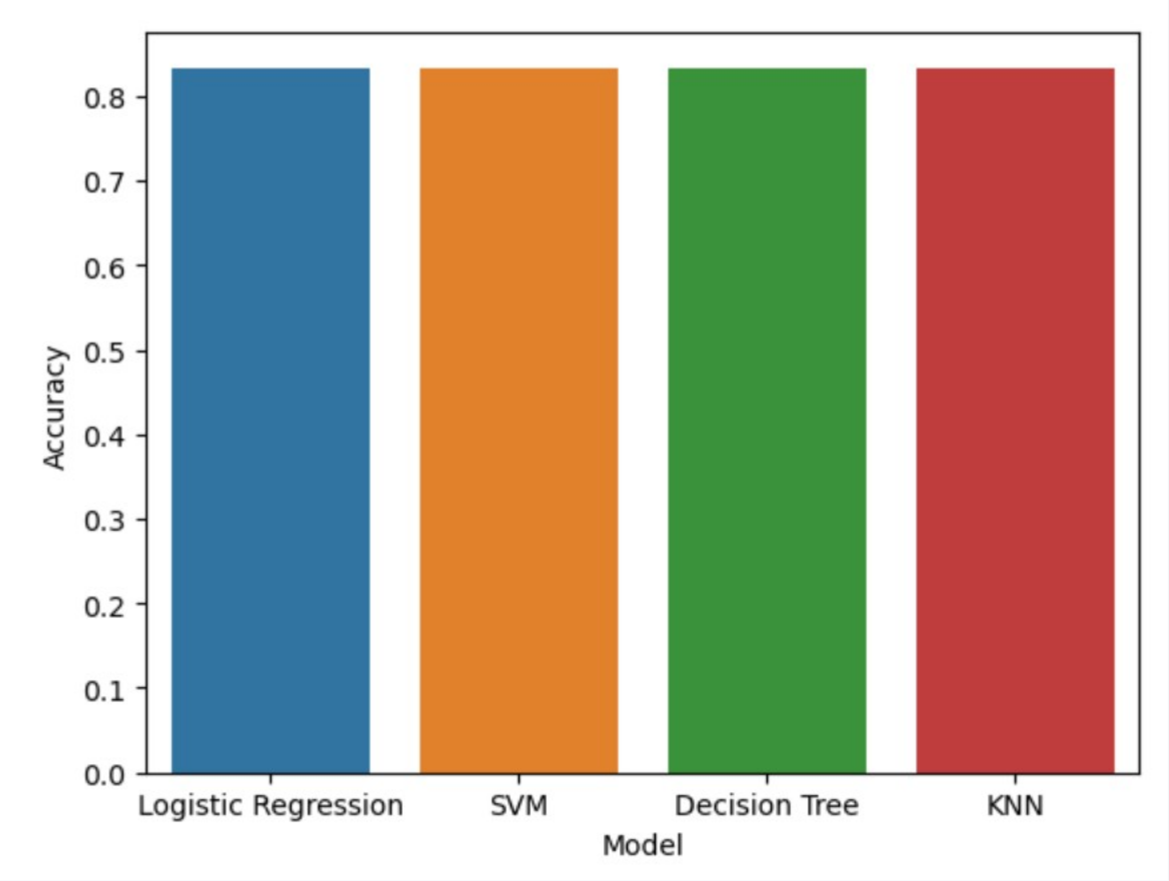


Section 5

Predictive Analysis (Classification)

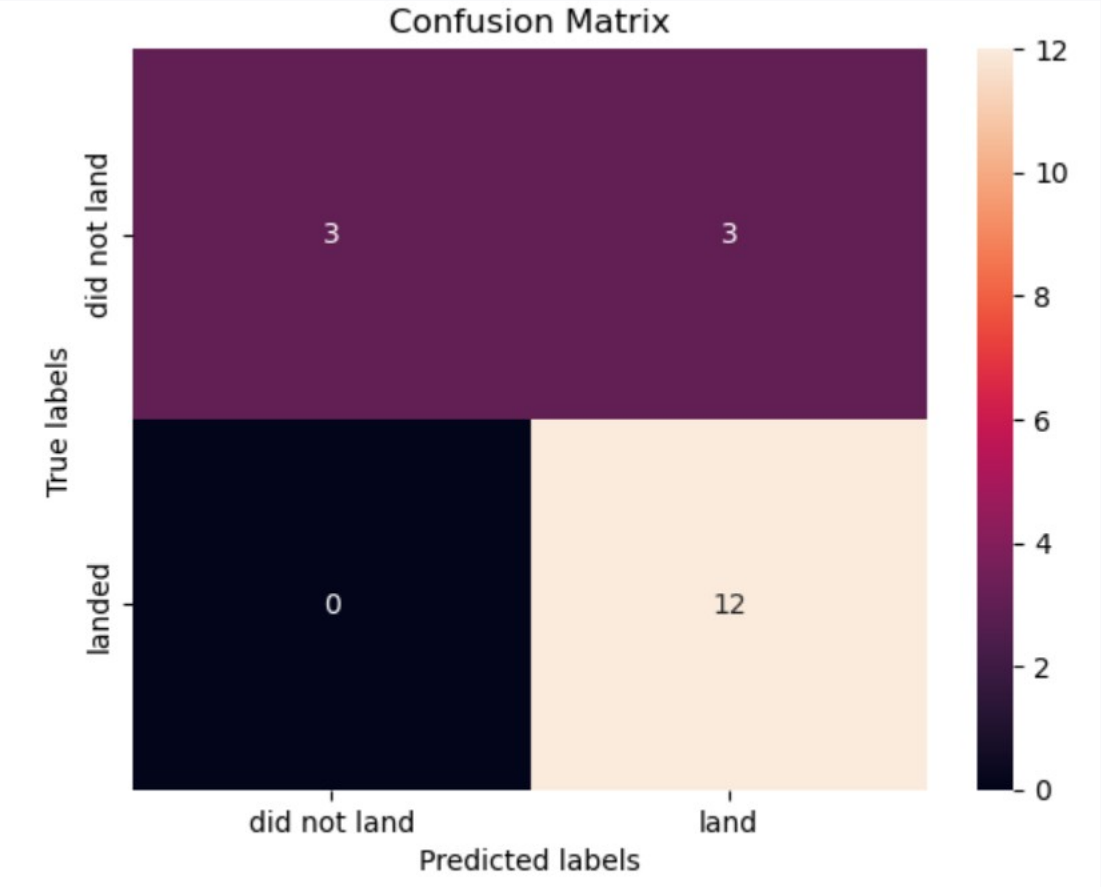
Classification Accuracy

- For this dataset, all models show the same prediction accuracy, 83.33%



Confusion Matrix

- This is the confusion matrix of the SVM model. It shows that the model can distinguish between the different classes (success and failure)



Conclusions

- Data acquisition and wrangling helped gather and clean the relevant data.
- EDA showed that some of the launch sites have highly successful launch rates.
- The yearly success rate increased, which was also shown as SpaceX took more flights
- Majority of the high payload flights were sent to the VLEO orbit.
- The visual dashboard and folium maps help better visualize the data.
- All predictive models tested show similar high accuracy (>80%), highlighting that the classes can be differentiated with high confidence

Appendix

Github URL for the repository:

https://github.com/aggp11/SpaceY-Capstone/tree/Final-assignment-submission

Thank you!