IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aaron Pratt
June 2025

Outline: In This Presentation

GitHub Project URL: https://github.com/aaronpratt1981/DataScienceCapstone/tree/main/

Executive Summary

- Data collection methodology
  Publicly available SpaceX data was obtained through the REST API and by scraping the SpaceX Wikipedia page
- Perform data wrangling
  Data was confined to a particular date range and naming was reclassified
- 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 was split into an 80/20 testing set and testing using various models. Results were captured and visualized with confusion matrices.

Key Conclusions

Although a small dataset was used for testing, a successful Falcon 9 launch can be predicted with some level of certainty. A larger data set would likely lead to better predictions.

The data available through SpaceX is generally clear and requires little wrangling.

The success of launches varies based on payload weight and booster version. Further exploration of these variables effects on launches should be explored.

Launch success has generally improved for SpaceX over time.

Introduction

The intent of this capstone project is to 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 could be used, for example, if an alternate company wants to bid against SpaceX for a rocket launch.

Through data acquisition, wrangling, visualization, and classification, we attempt to determine:

- What makes a Falcon 9 rocket launch successful or unsuccessful?
- Are there any patterns that would help us better compete against SpaceX?
- Are there any other areas we should explore in the future?

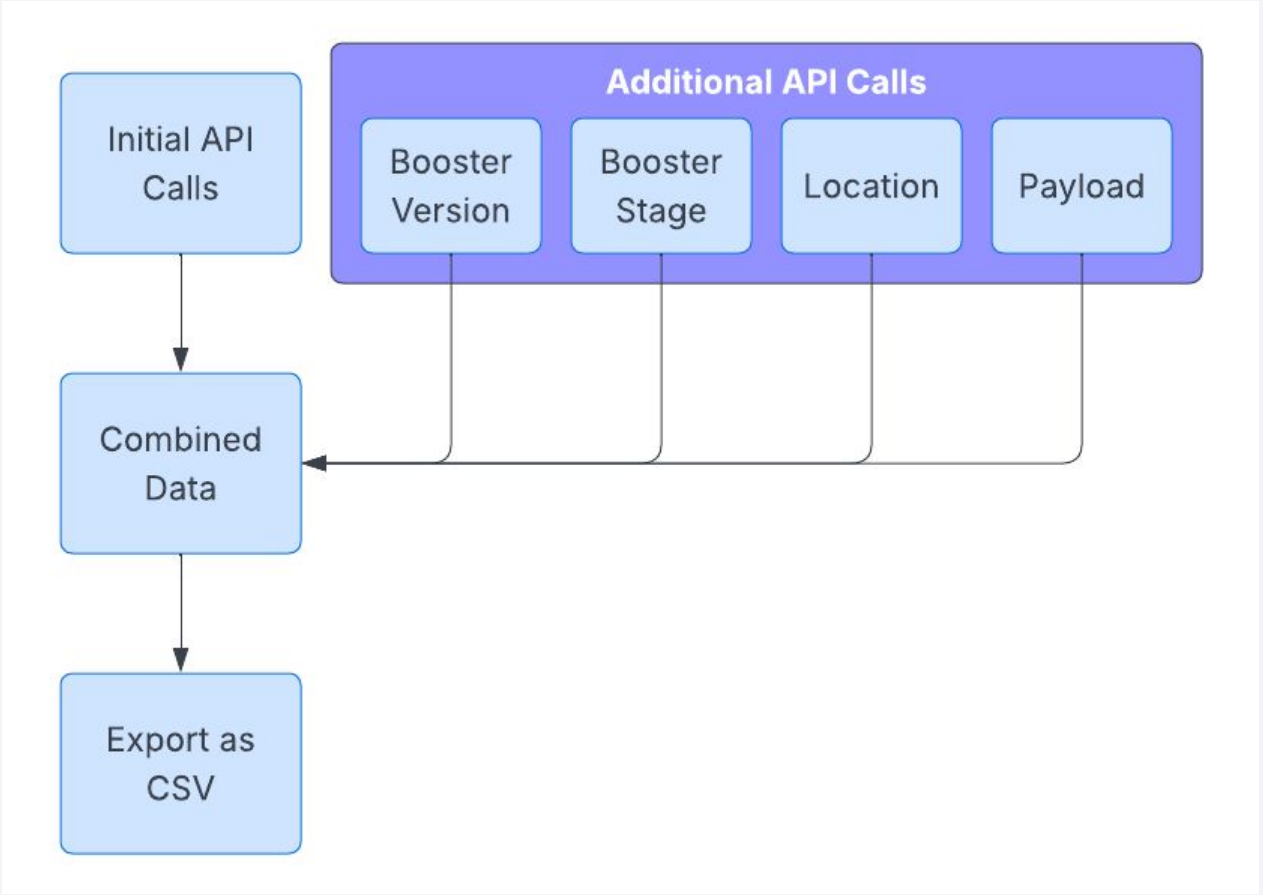
Section 1

Methodology

Data Collection – SpaceX API

Using python, data is collected through REST API calls to the SpaceX API.

1. Obtain launch data
2. Enrich launch data
   a. Booster Version
   b. Location
   c. Payload
   d. Booster Stage
3. Export as CSV



GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/1-api-collection.ipynb

Data Collection - Scraping

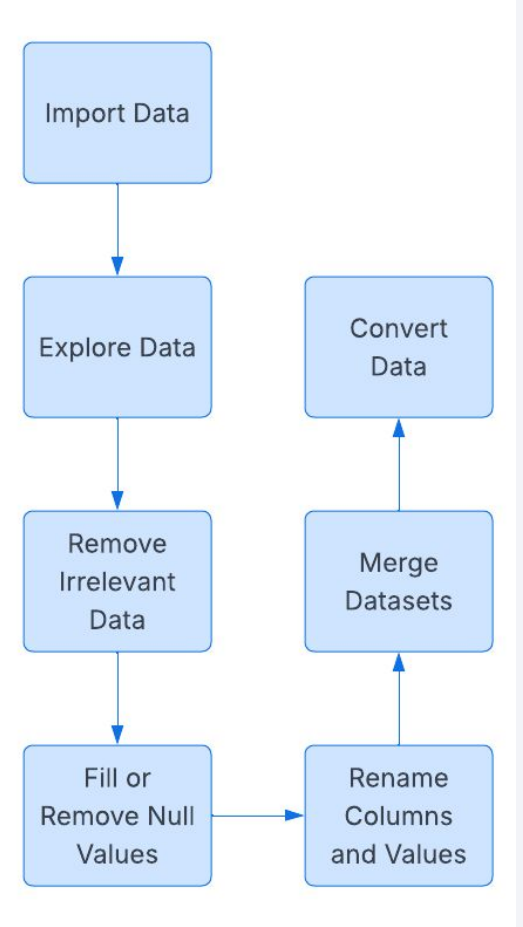
Using beautifulsoup4 and other python libraries, an HTML page similar to a Wikipedia entry was loaded, the target data was identified, it was scraped row-by-row, and exported as a new CSV file.

GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/2-web-scraping.ipynb

Data Wrangling

Data was provided in a CSV file from a public URL. The data was imported into a pandas dataframe for exploration and exported to a slightly modified CSV file limited to a defined data range. Because this data was part of a class project, it perhaps required less wrangling than is usually required. Typically data wrangling would include:

- Dropping missing or irrelevant columns
- Filling or removing nulls
- Renaming columns
- Merging datasets
- Data type conversion



GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/3-data-wrangling.ipynb

EDA with Data Visualization

Charts plotted:

- **Scatter Plots:** Because a scatter plot can be used to display relationships between three variables, they are a useful way to quickly explore relationships between multiple metrics in a visual manner. Relationships were examined through scatter plots comparing Flight Number vs. Launch Site; Payload vs. Launch Site; Flight Number vs. Orbit Type; and Payload Vs. Load Type (all colored by class).
- **Bar Chart:** Bar charts are best for displaying information categorically so a bar chart was used to explore the relationship between Success Rate and Orbit Type.
- **Line Chart:** Line charts are well-suited for displaying a change in a metric over time, so a line chart was used to show how launch success has changed by year.

GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/5-eda-viz.ipynb

EDA with SQL

The following SQL queries were performed on the data set:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass. Use a subquery.
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/4-eda-sql.ipynb

Build an Interactive Map with Folium

Folium Map Objects Added:

a. **Circle:** Used to show launch locations on the map in a highlighted circle based on latitude and longitude.

b. **Marker:** Used to show the launch location in the center of the highlighted circles.

c. **Marker Clusters:** Used to show successful (green) and unsuccessful (red) markers within the highlighted circle for each launch location.

d. Polylines: Used to show the distance between launch sites and key features (such as highways, railways, cities, and coastlines).

GitHub URL: https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/6-folium.ipynb

Build a Dashboard with Plotly Dash

Charts added:

a. **Pie chart** showing the number of successful launches (with ability to select all launch sites or specific launch sites). The pie chart was chosen because it is ideal for showing ratios for a specific measurement.
b. **Scatter plot** showing the payload vs. launch success by booster version (with the ability to set the minimum and maximum payload in kg). The scatter plot was chosen because it allows the comparison of three variables in an understandable, visual manner.

GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/7-dashboard.py

Predictive Analysis (Classification)

1. Data Preparation

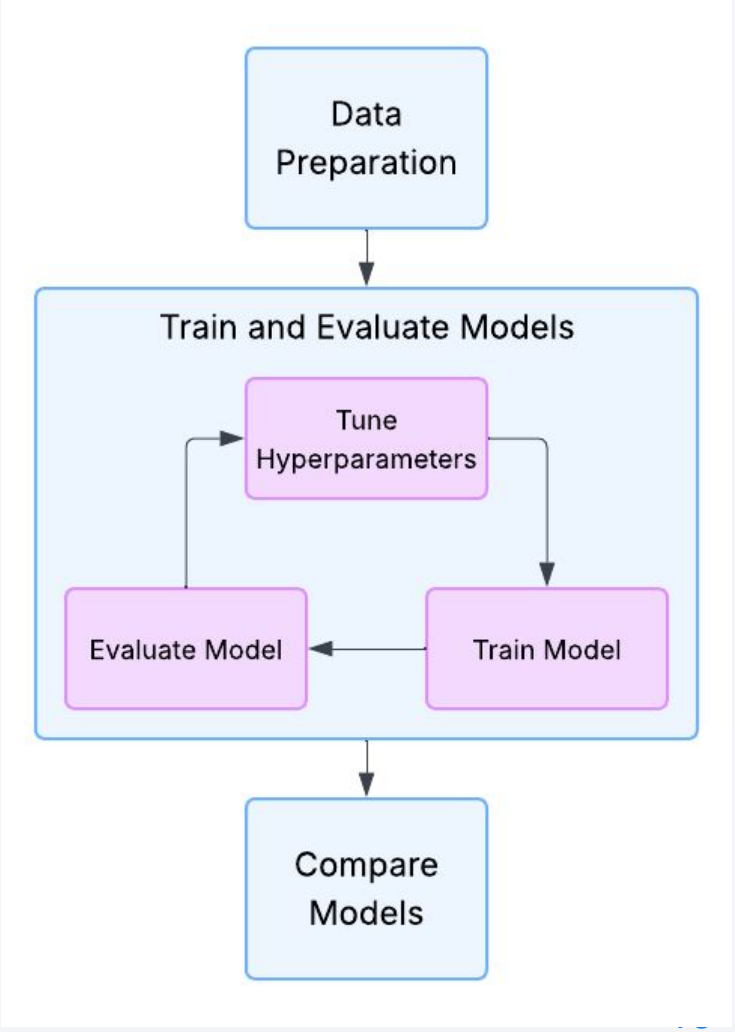
- Loaded datasets from cloud URLs using pandas
- Standardized data and split into test and training sets

2. Train and Evaluate Models

Built and compared four classification models

3. Compare Models

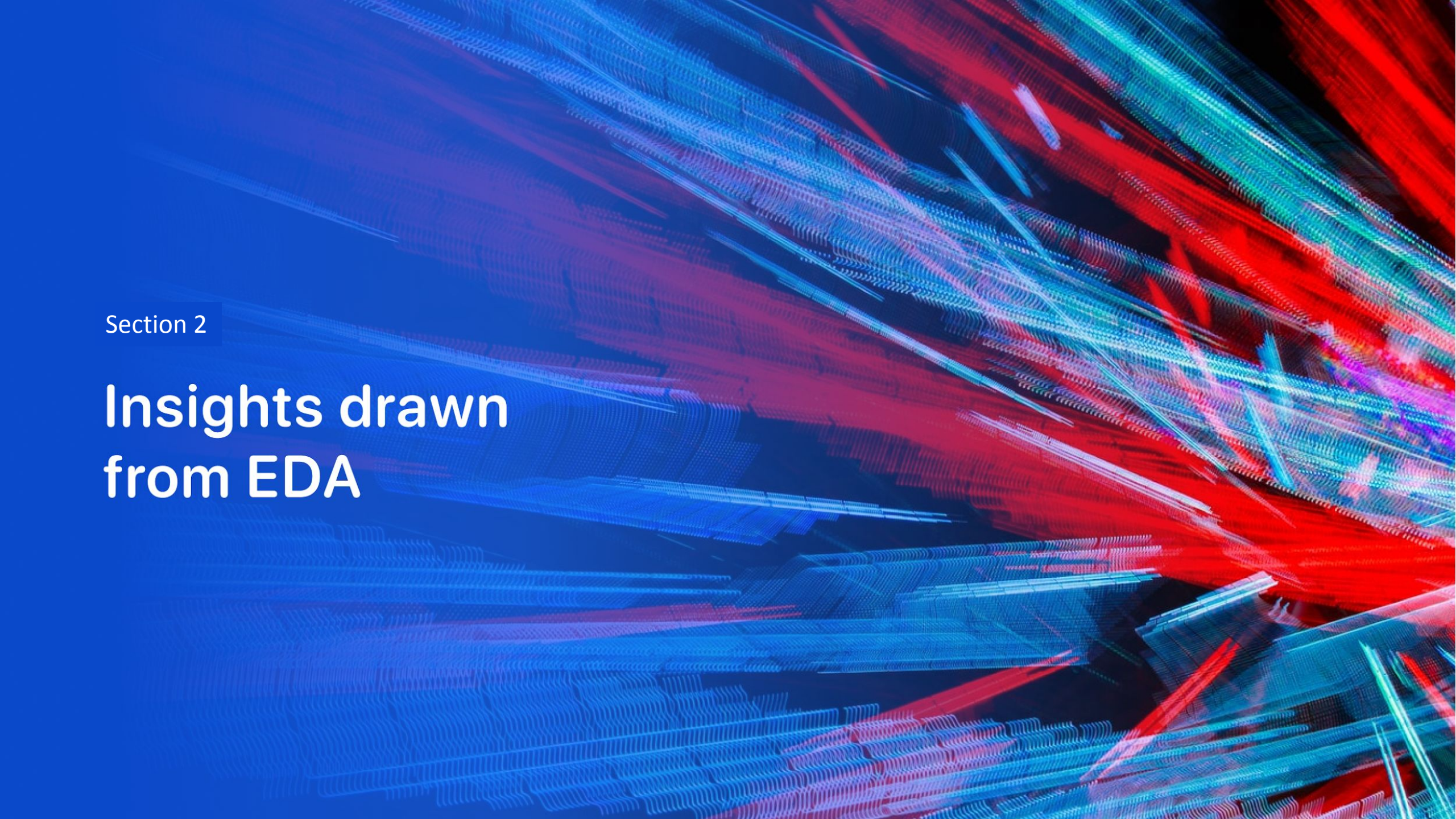
- Review cross-validation accuracy and test accuracy, as well as confusion matrices
- **Logistic Regression, SVM, and KNN** performed consistently with ~83% accuracy
- **Decision Tree** showed overfitting (high CV accuracy, low test accuracy)



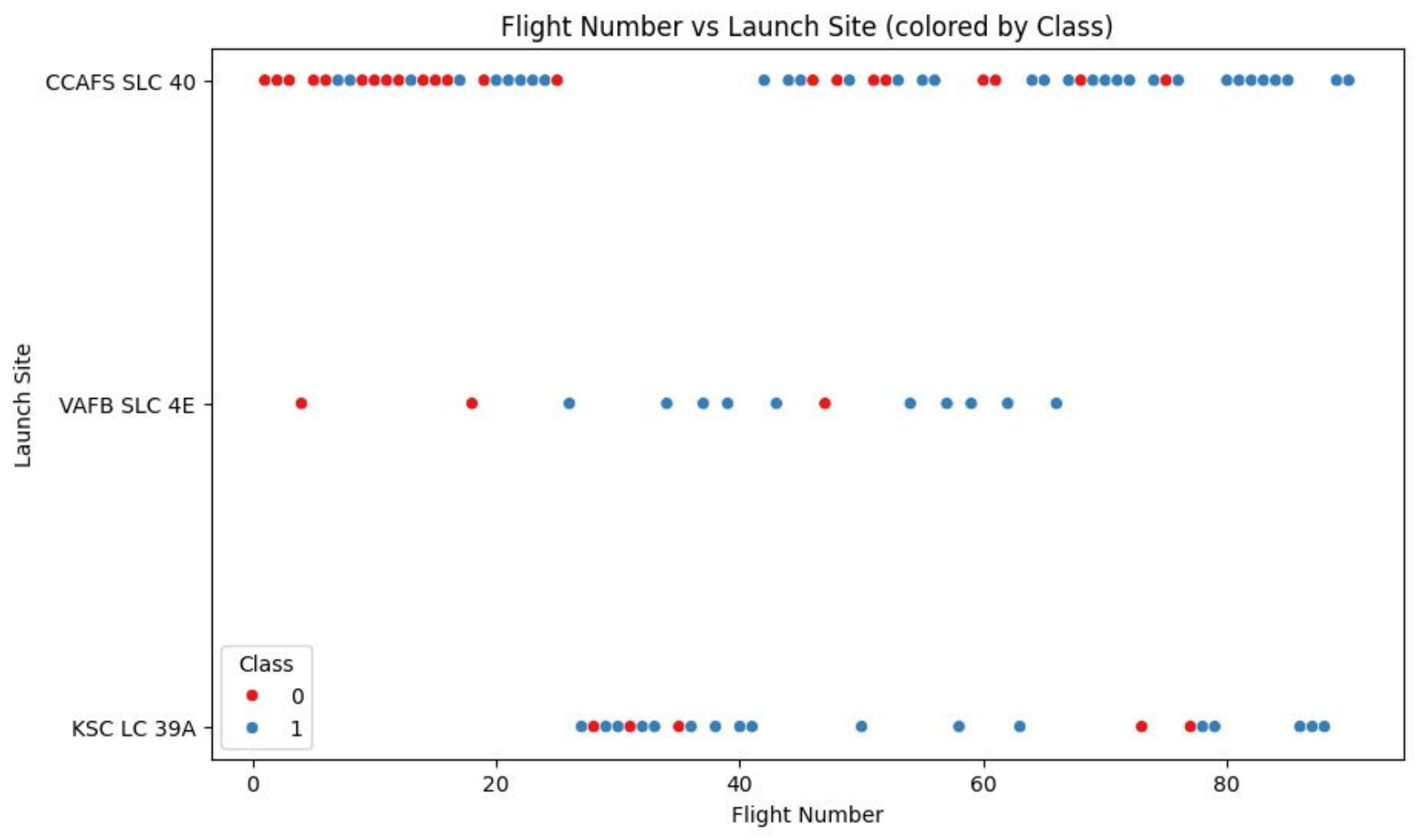
GitHub URL:
https://github.com/aaronpratt1981/DataScienceCapstone/blob/main/8-training.ipynb

Section 2

Insights drawn from EDA

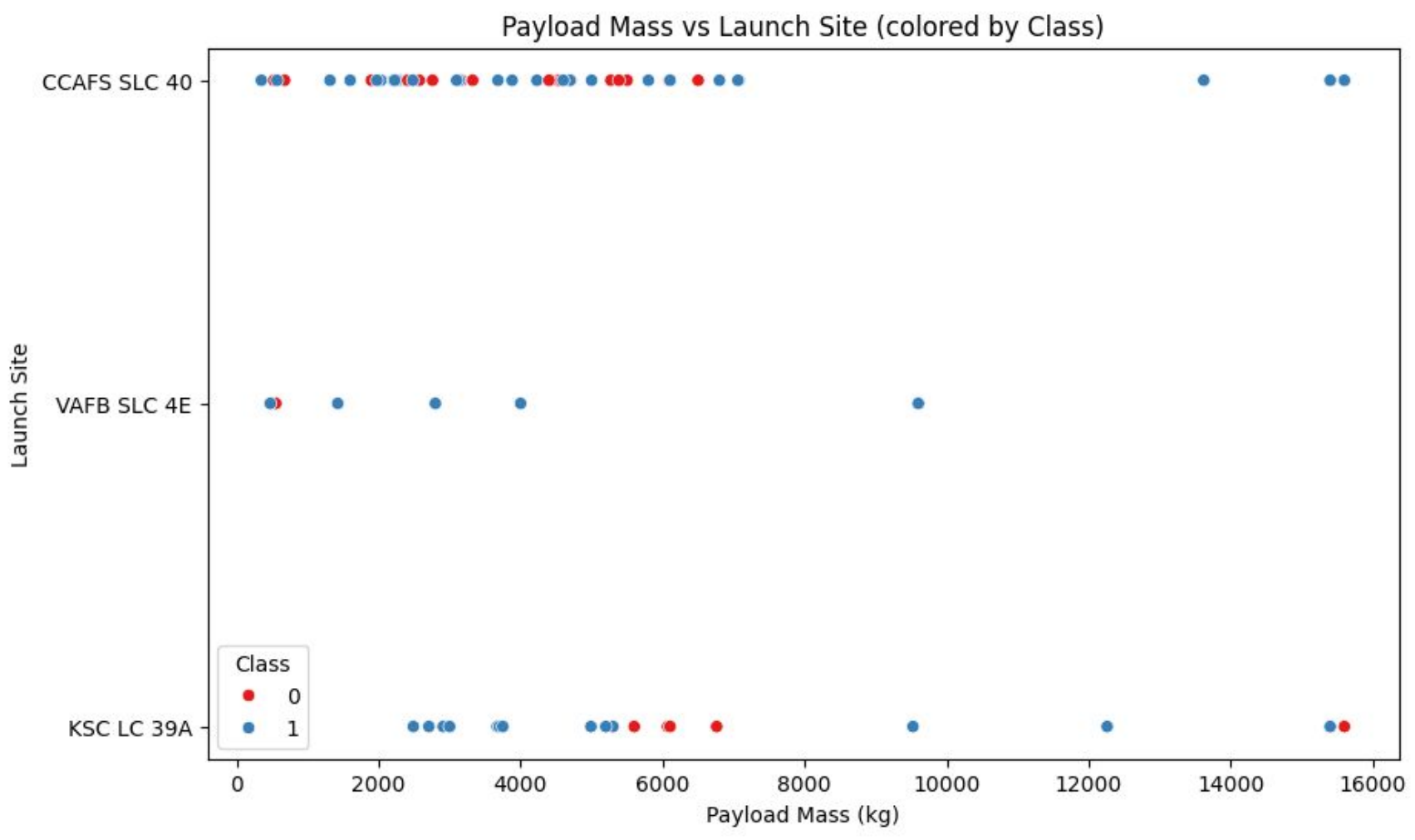


Flight Number vs. Launch Site



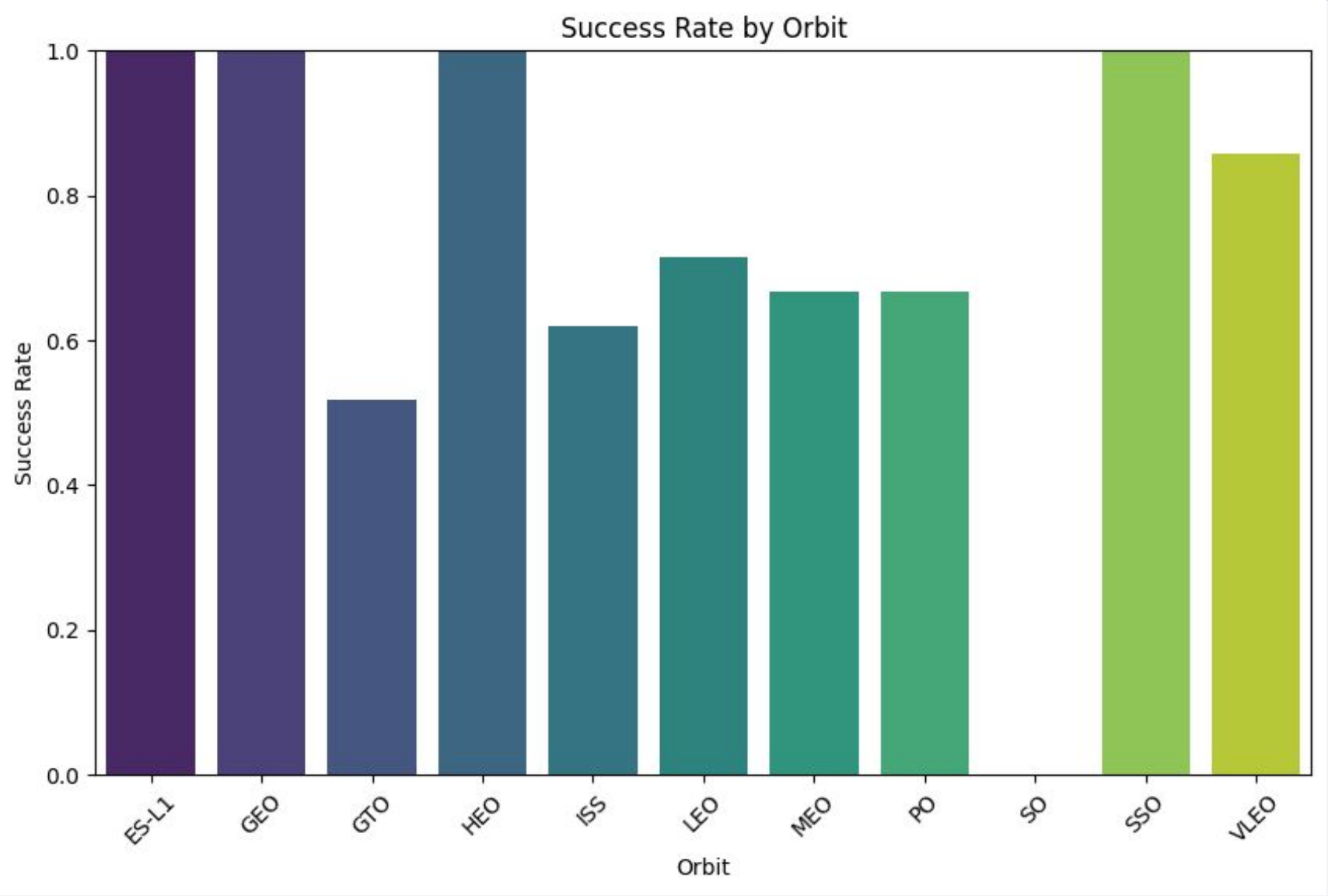
Launch site CCAFS had the majority of launches, but if flight numbers correspond to time, there was a period where KSC seemed to be used instead of CCAFS. Additionally most early flights were Class 0 (unsuccessful).

Payload vs. Launch Site



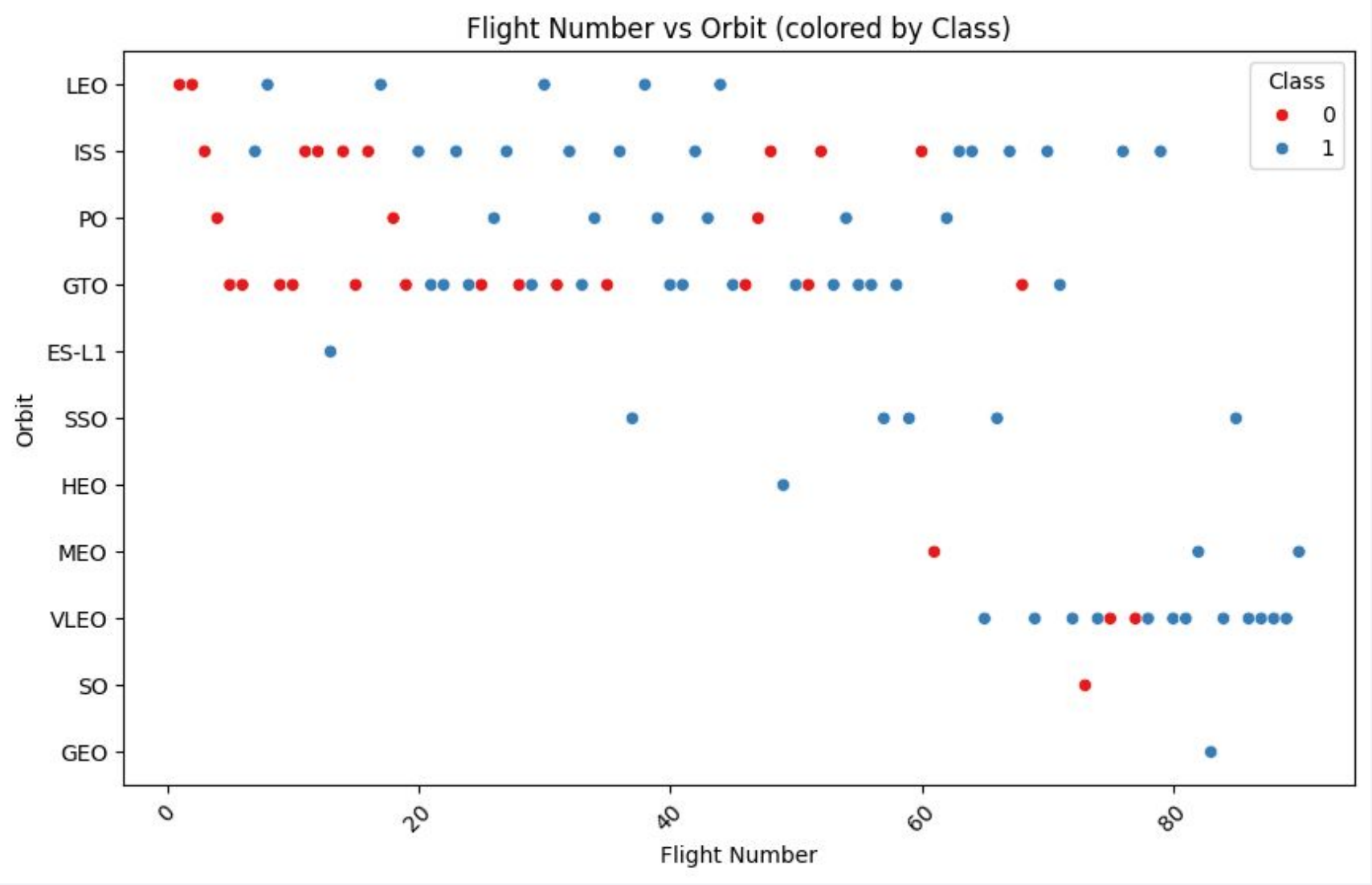
For the VAFB-SLC launchsite, there are no rockets launched with a payload greater than 10,000 kg (although a large majority of payloads were below 10,000 kg's regardless of launch site).

Success Rate vs. Orbit Type



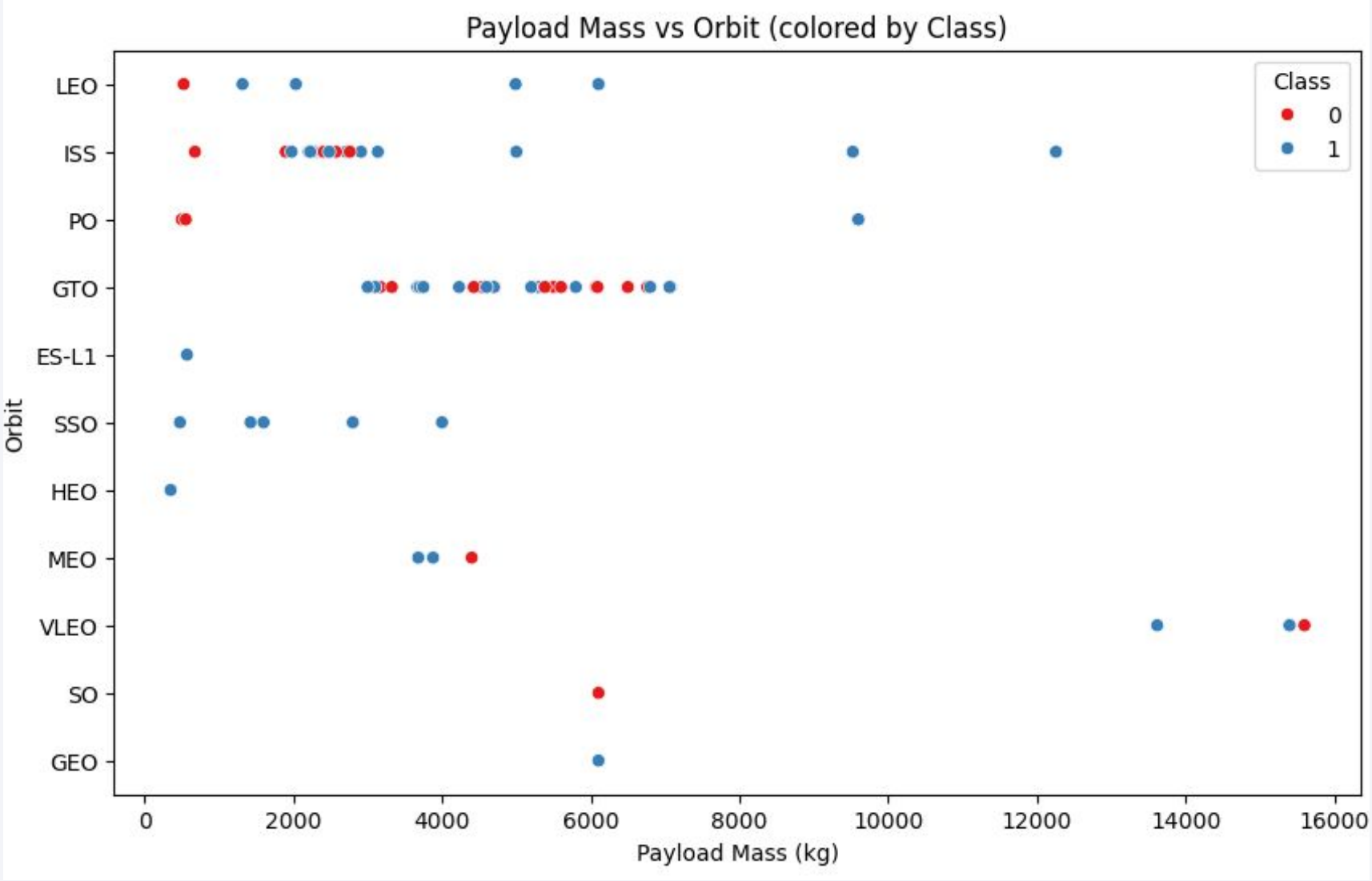
Orbits ES-L1, GEO, HEO, and SSO had the highest success rates. Others had a success rate roughly from 50% - 70% with the exception of a 0% success rate for SO.

Flight Number vs. Orbit Type



In the LEO orbit, success seems to be related to the number of flights. But in the GTO orbit, there appears to be no relationship between flight number and success.

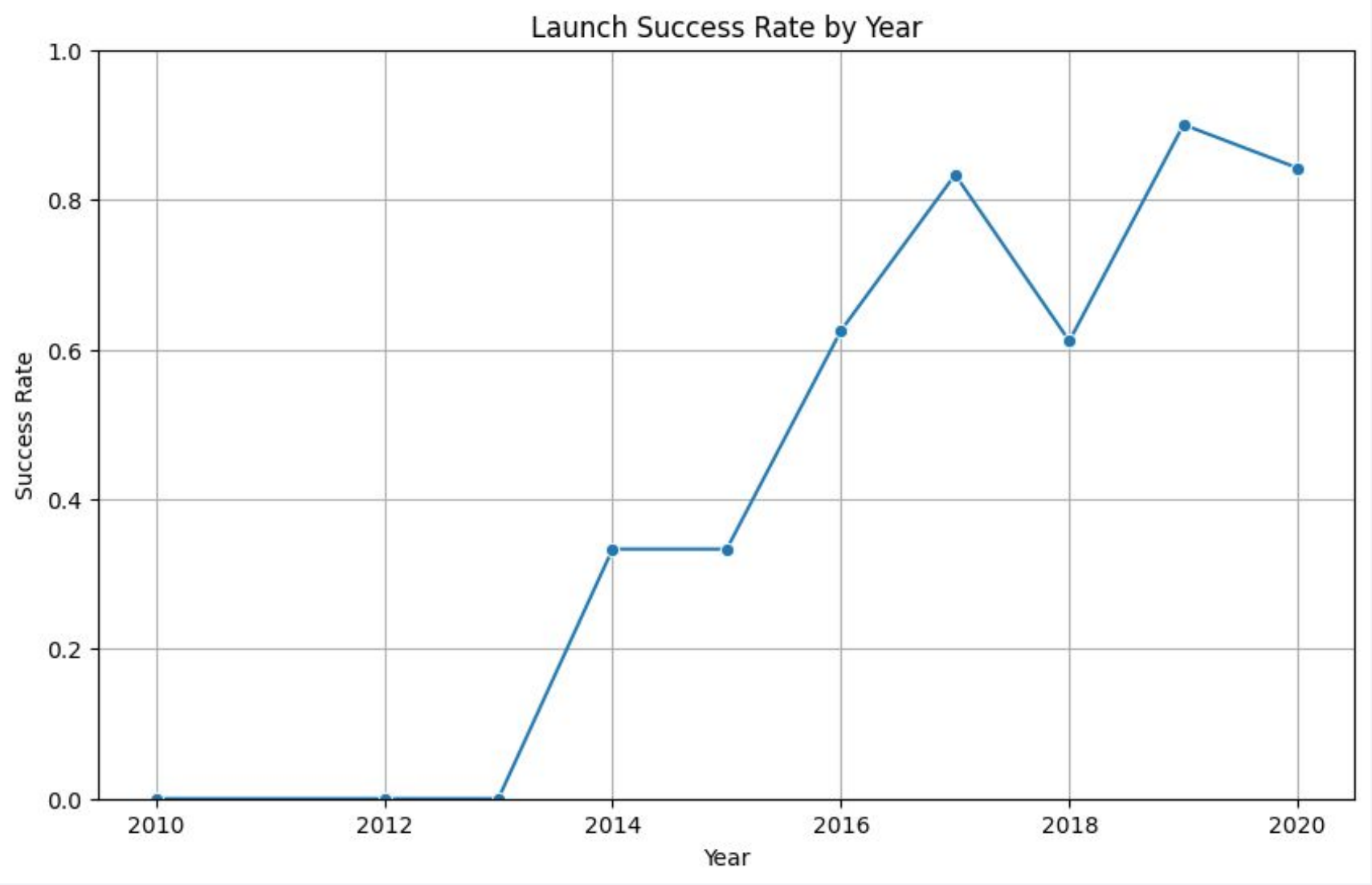
Payload vs. Orbit Type



With heavy payloads there are more successful for Polar, LEO and ISS.

It's difficult to distinguish between successful and unsuccessful landings for GTO as both outcomes are present.

Launch Success Yearly Trend



Success rate has shown general improvement since 2013 through 2020 with some noted drops in 2018 and 202. However, since 2016, success rate has continued to be above 60%.

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Querying with DISTINCT ensures that only unique values are returned in a SQL query.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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	**CCA**FS LC-40	Dragon Spacecraft …	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight…	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight...	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1…	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2…	677	LEO (ISS)	NASA (CRS)	Success	No attempt

This query returns all fields of the first five records where the launch site begins with “CCA”.

Total Payload Mass

```
%sql SELECT SUM(Payload_Mass__kg_) AS Total_Payload_Mass FROM SPACEXTABLE
WHERE Customer LIKE '%NASA (CRS)%';
```

Total_Payload_Mass
48213

This query adds the Payload Mass (in kg) together for each NASA launch.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(Payload_Mass__kg_) AS Avg_Payload_Mass FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

Avg_Payload_Mass
2928.4

Similar to the previous query, this query averages the Payload Mass (in kg) together for each F9 v1.1 booster.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS First_Ground_Pad_Landing FROM SPACEXTABLE WHERE
Landing_Outcome = 'Success (ground pad)';
```

First_Ground_Pad Landing
2015-12-22

This query returns the oldest (MIN()) date that a rocket successfully landing on a ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success
(drone ship)' AND Payload_Mass__kg_ > 4000 AND Payload_Mass__kg_ < 6000;
```

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

This query returns the boosters that successfully landed on the drone ship and had a payload mass between 4000 and 6000. If the query needed to be inclusive of 4000 and 6000, BETWEEN could have been used instead of < and >.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT CASE WHEN Mission_Outcome LIKE '%Success%' THEN
'Success' WHEN Mission_Outcome LIKE '%Failure%' THEN 'Failure'
ELSE 'Other' END AS Outcome_Category, COUNT(*) AS Total_Count
FROM SPACEXTABLE GROUP BY Outcome_Category;
```

Outcome_Category	Outcome_Total
Failure	1
Success	100

This query counts the number of records that contain Success or Failure (or neither), and then counts and displays the number of the launch successes and failures.

Boosters Carried Maximum Payload

```
%sql SELECT DISTINCT
Booster_Version FROM SPACEXTABLE
WHERE Payload_Mass__kg_ = (SELECT
MAX(Payload_Mass__kg_) FROM
SPACEXTABLE);
```

This query first finds the highest Payload Mass from the table then returns the unique Booster Versions that carried a payload equal to that mass.

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

```
%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome,
Booster_Version, Launch_Site FROM SPACEXTABLE WHERE
Landing_Outcome LIKE '%Failure%' AND Landing_Outcome LIKE
'%drone ship%' AND substr(Date, 1, 4) = '2015';
```

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

This query gets drone ship failures that occurred in 2015. SQLLite does not support month names so str() is used to get the correct month and year values.

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

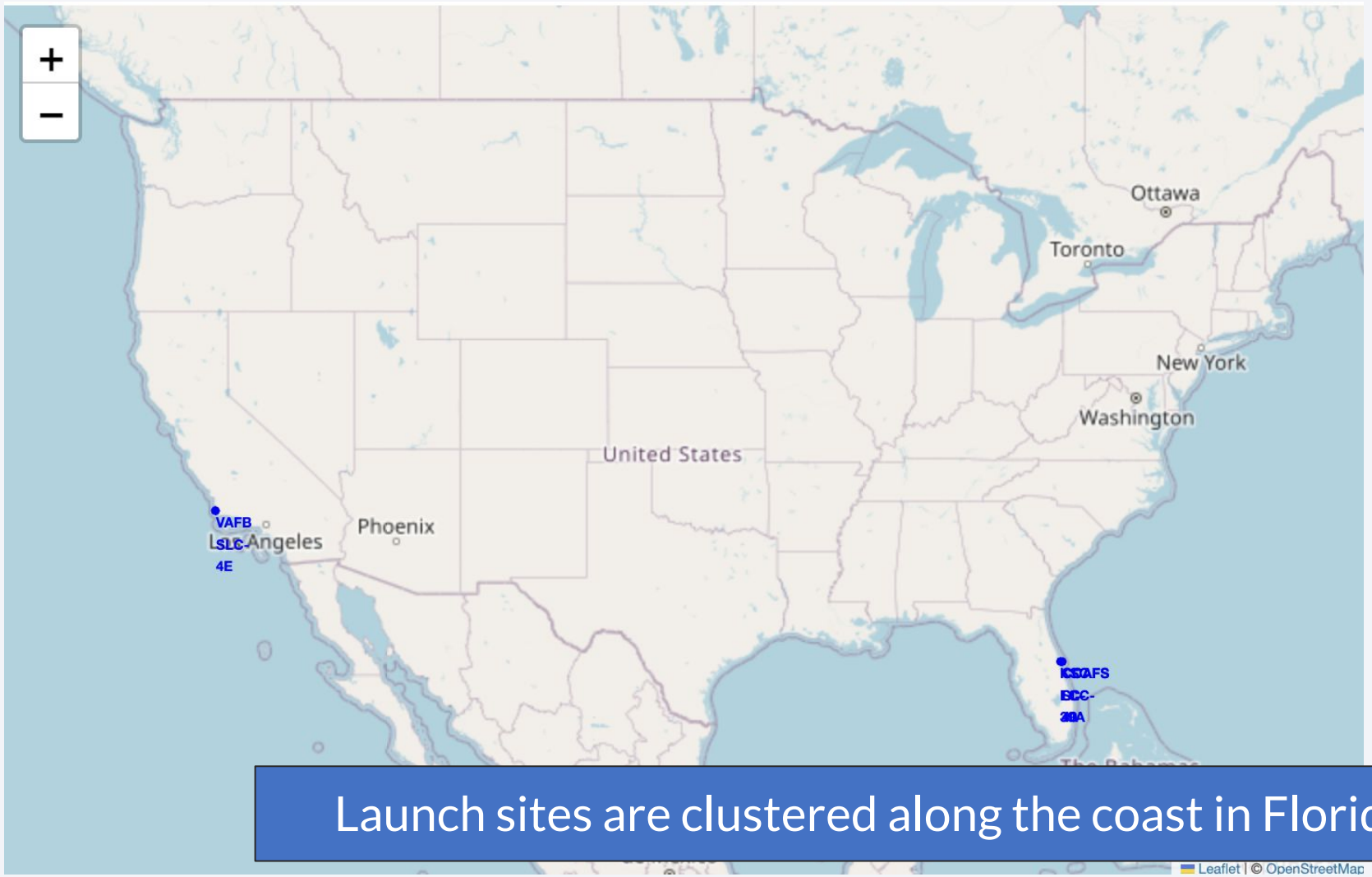
```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;
```

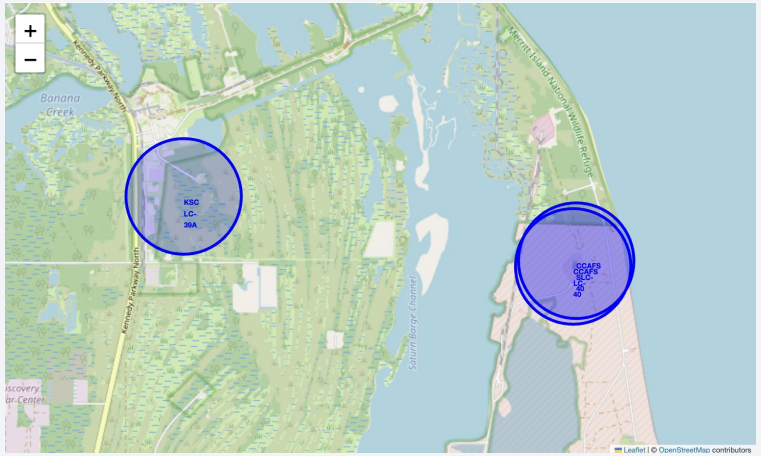
This gets all landings between 6/4/2010 and 3/20/2017, groups them by landing outcome, and returns a count of every landing outcome that occurred during that time period, from highest to lowest.

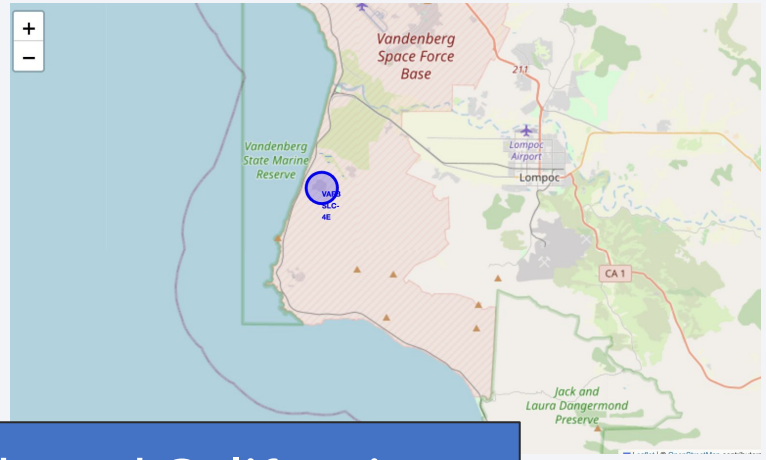
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Folium Map Launch Site Locations

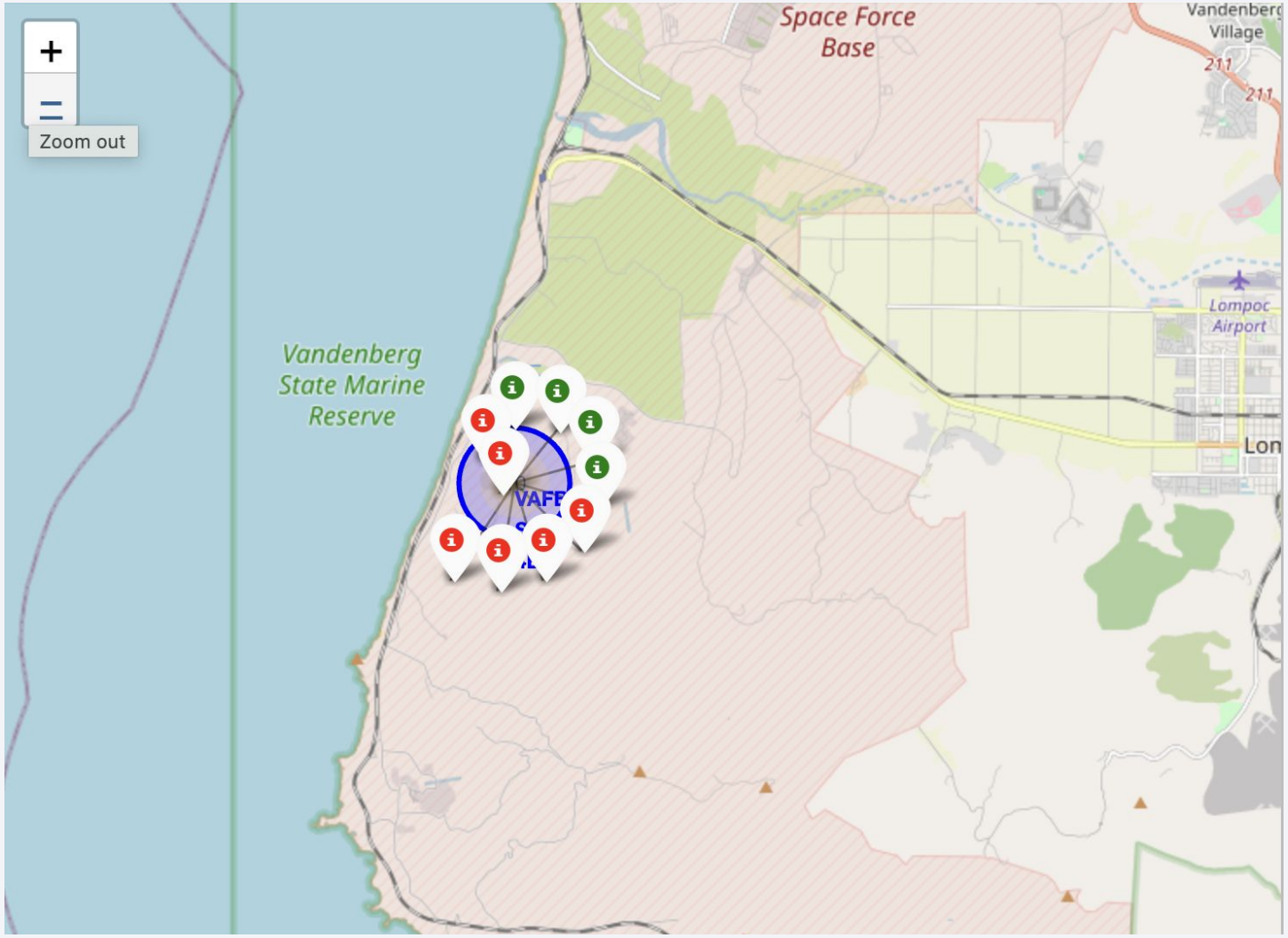






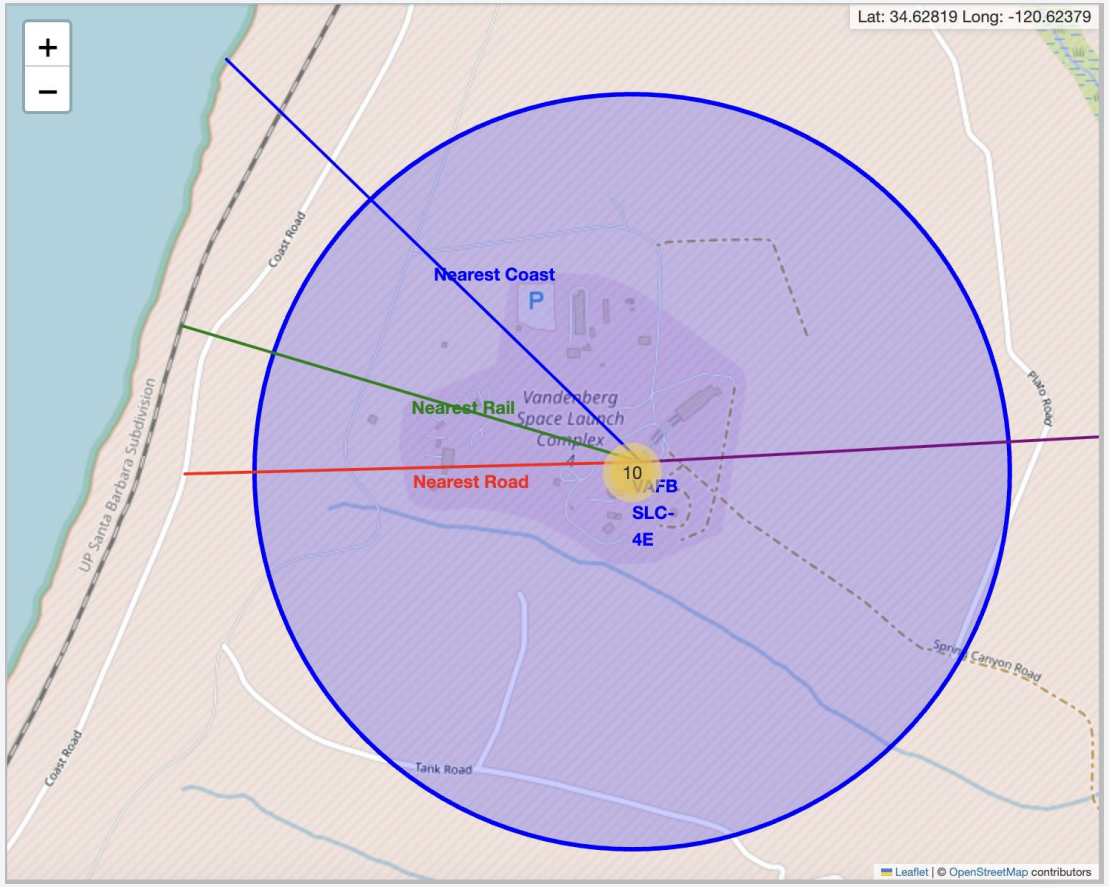
Launch sites are clustered along the coast in Florida and California.

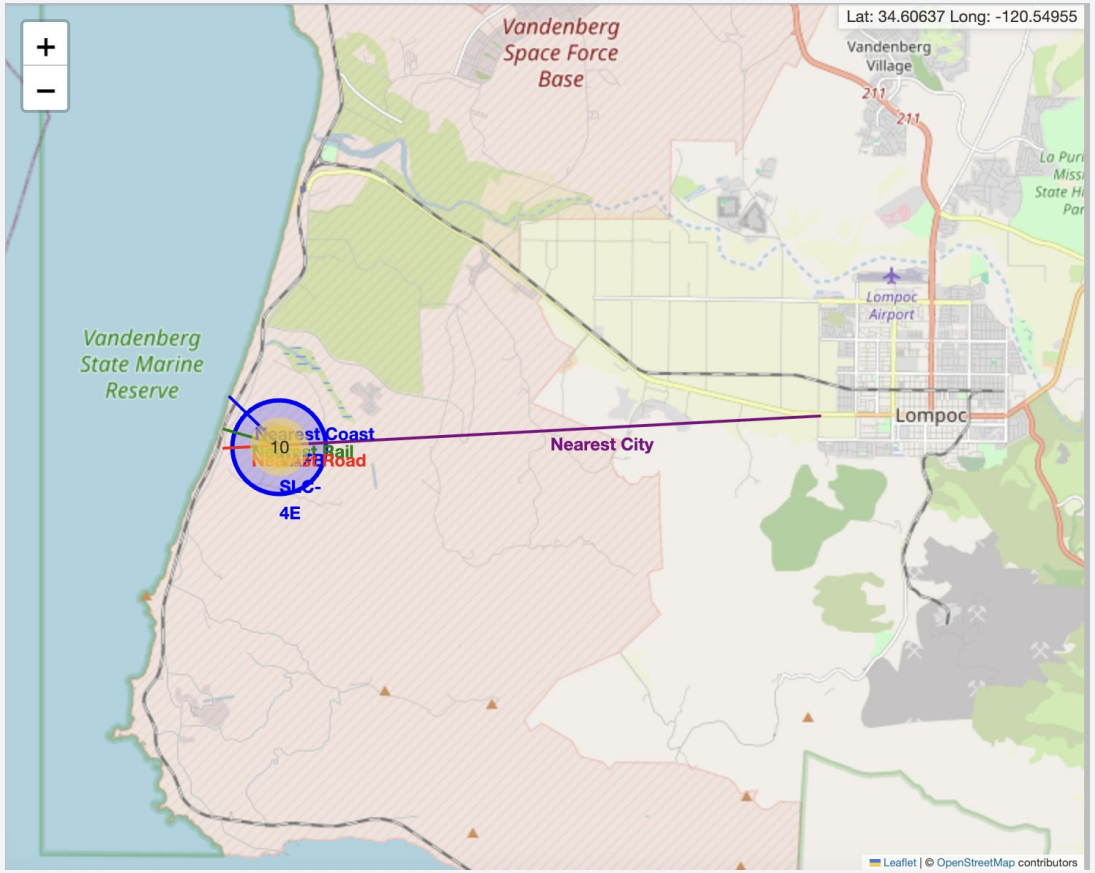
Marking Launch Successes and Failures on the Folium Map



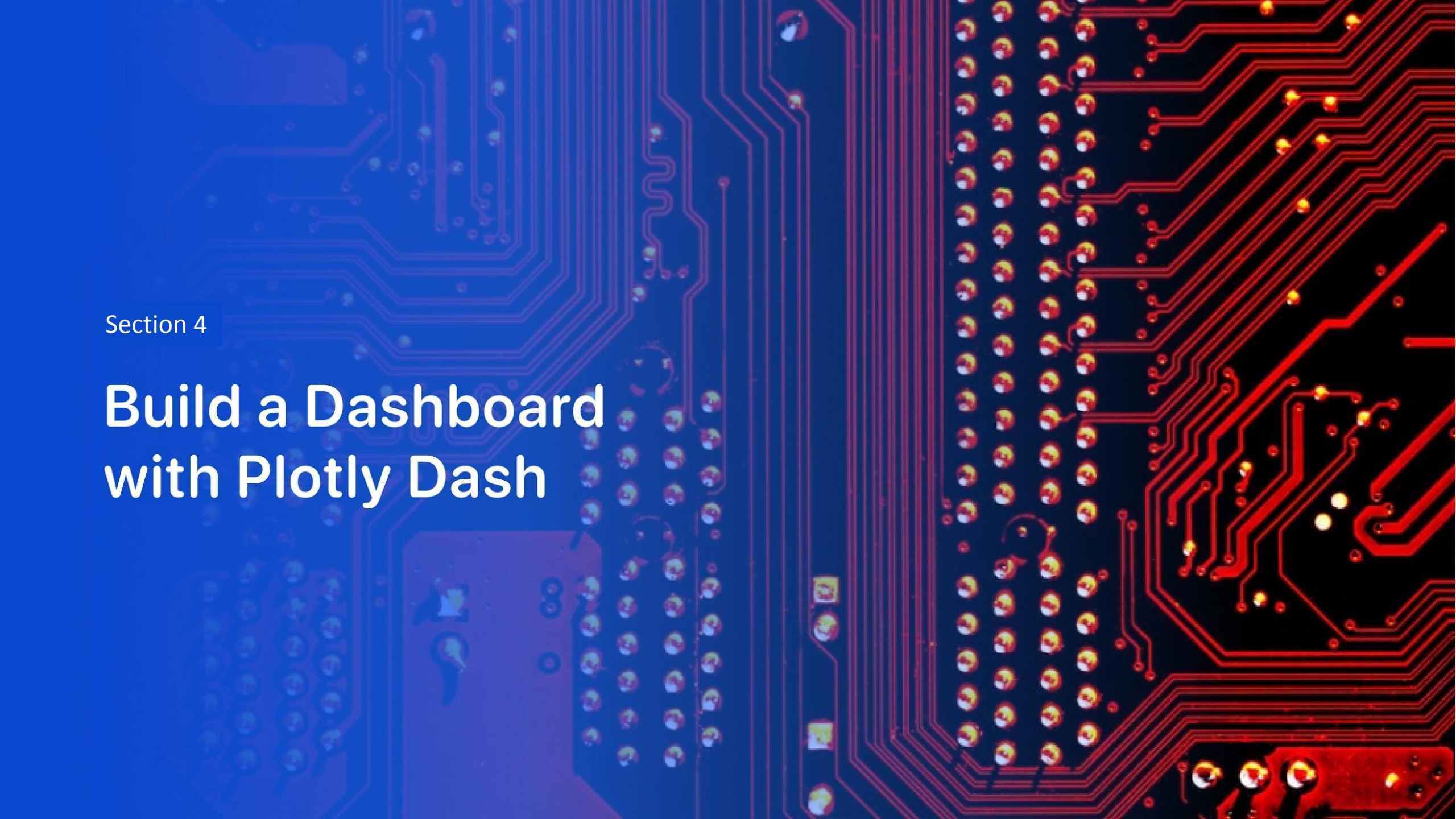
Successful launches are marked in green, with failures marked in red. This allows a quick, visual way to see if a launch site has a pattern of successful launches.

Launch Site Proximity to Key Features

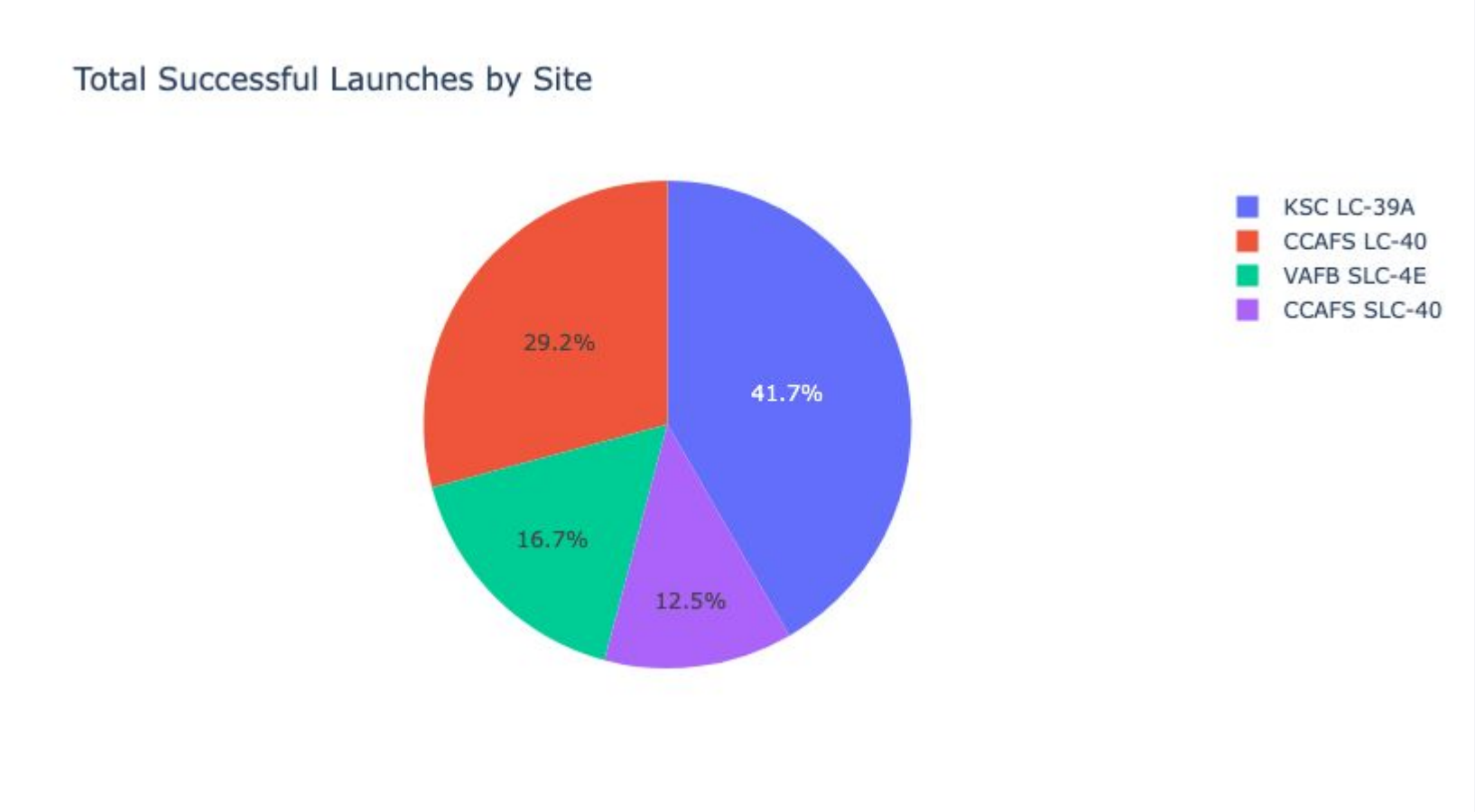




Launch sites are relatively close to a rail line, the cost, and roads (although not necessarily a highway). The nearest city is usually further away.

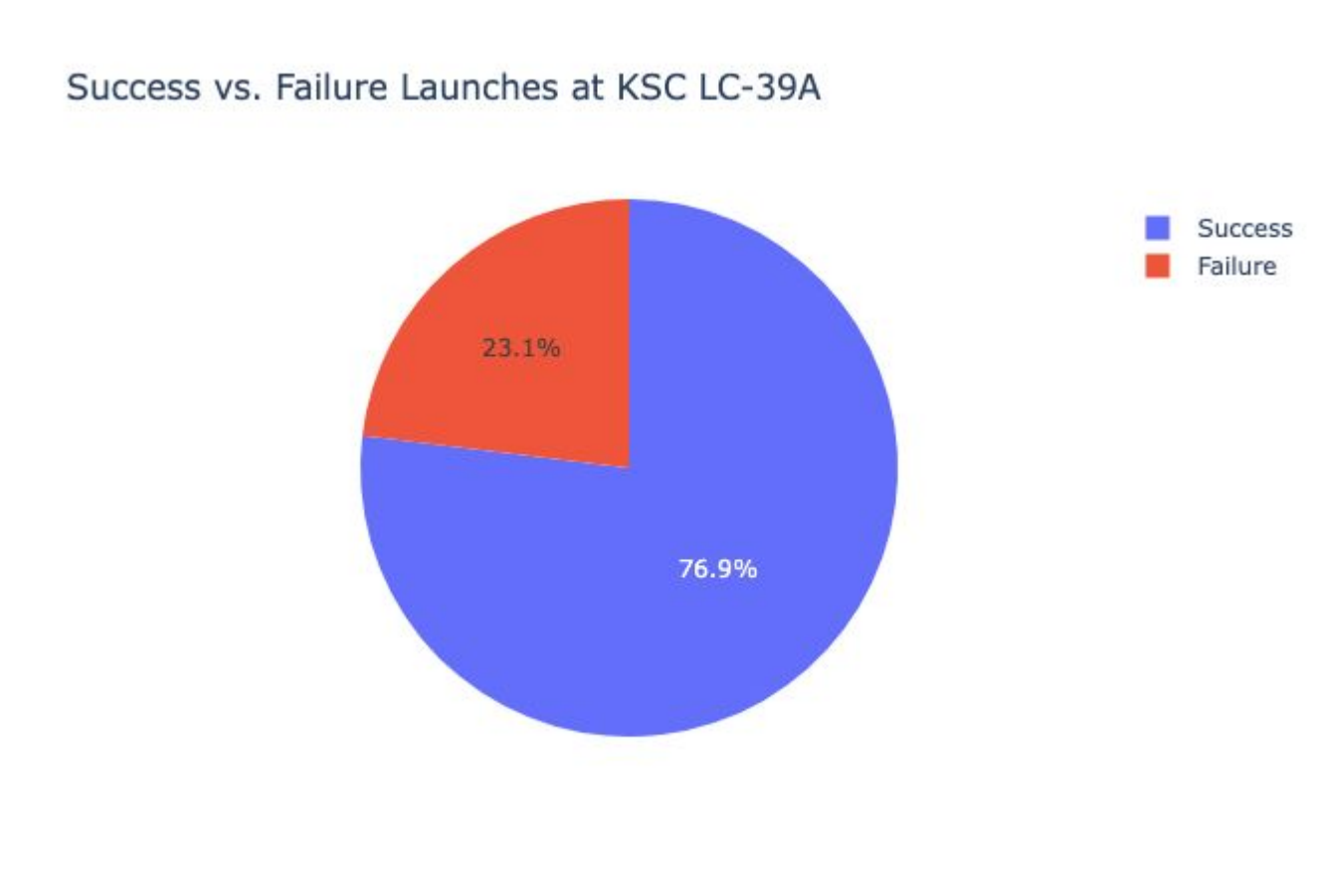

Section 4
Build a Dashboard with Plotly Dash

Launch Success for All Sites, Pie Chart



Of the four sites, KSC LC-39A accounted for more than 40% of the successful launches across all sites, whereas CCAFS SLC-40 accounted for the least at 12.5%.

Launch Site with Highest Launch Success Ratio



Not only did KSC LC-39A account for the largest percentage of successful launches across all sites, but it had the best ratio of successful to unsuccessful launches. Nearly 77% of launches were successful. That means ten launches were successful and three launches were unsuccessful.

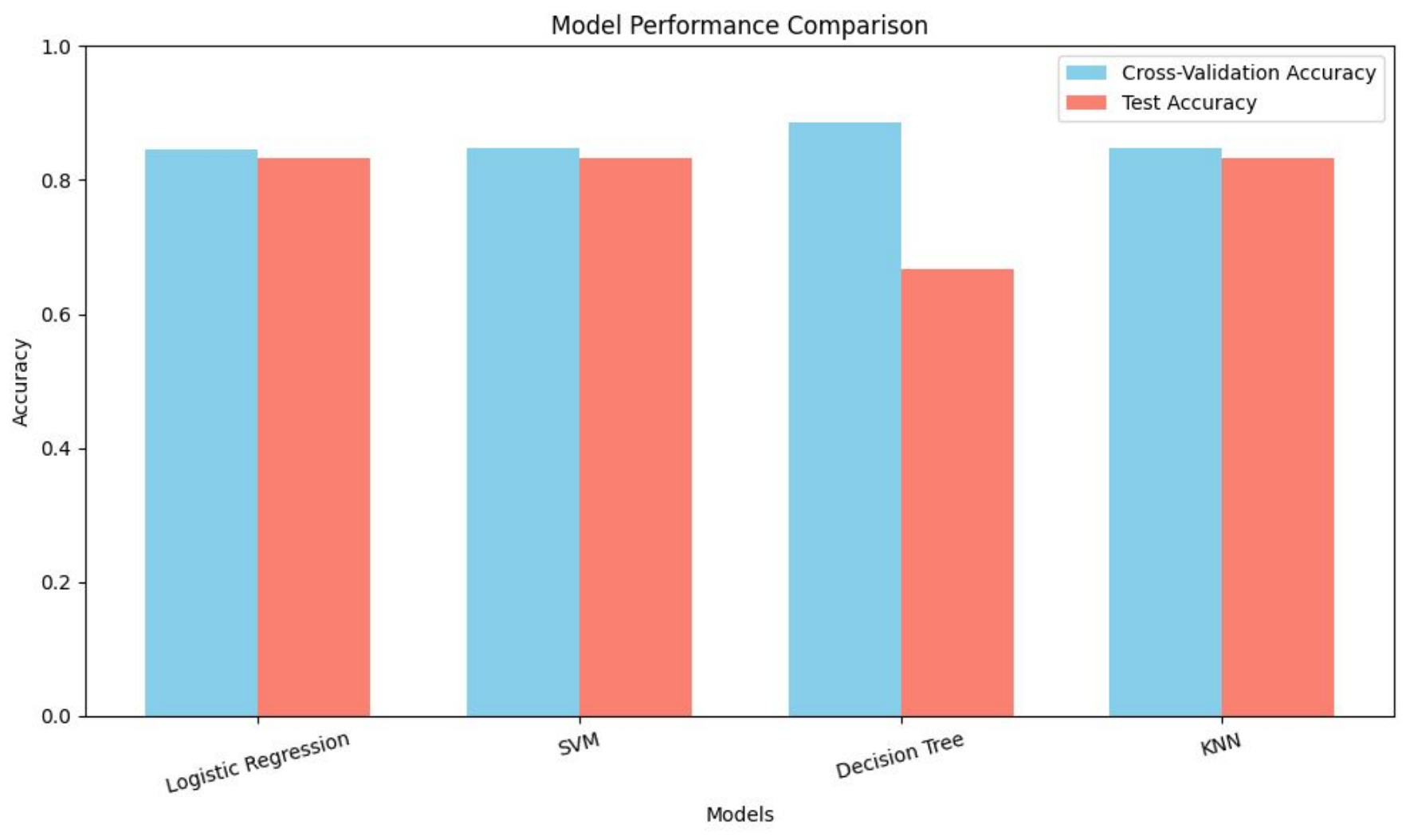
Payload vs. Launch Outcome Scatter Plot

For the largest half of payloads (chosen in the slider), the FT booster had more successful launches (2) when compared to the B4 (1). However, both had the same number of failures (4).

Section 5

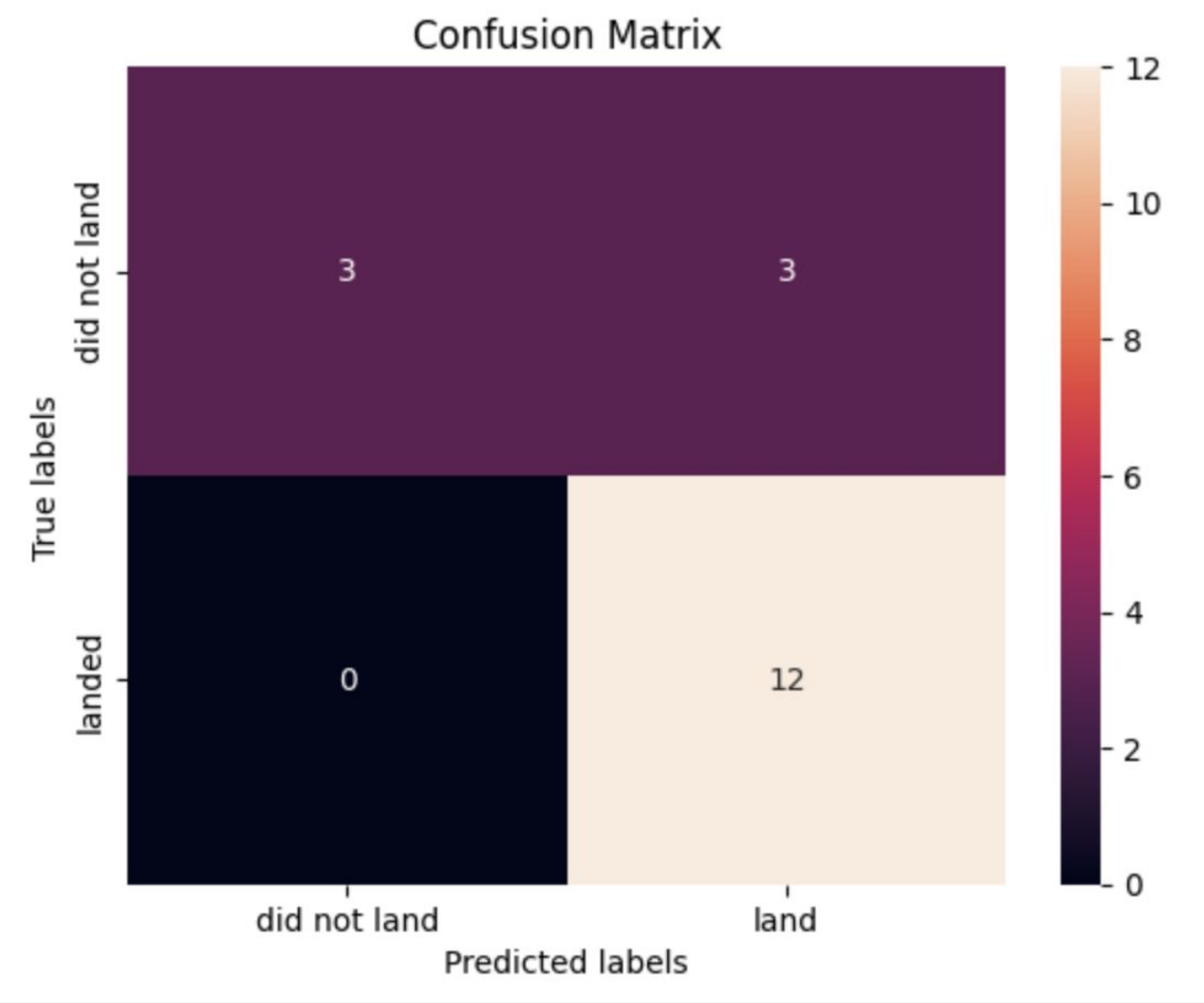
Predictive Analysis (Classification)

Classification Accuracy



Logistic Regression, SVM, and KVN have nearly identical accuracy. A larger test set may show greater variation in model performance.

Confusion Matrix



The Confusion matrix for the Logistic Regression, SVM, and KVN models are all identical. There are a few false positives for all three models. It's not pictured here, but the Decision Tree model overfits its results.

Again, a larger test set may show greater variability in models.

Conclusions

1. Although a small dataset was used for testing, a successful Falcon 9 launch can be predicted with some level of certainty. A larger data set would likely lead to better predictions.

2. The data available through SpaceX is generally clear and requires little wrangling.

3. The success of launches varies based on payload weight and booster version. Further exploration of these variables effects on launches should be explored.

4. Launch success has generally improved for SpaceX over time.

Appendix

Related Data Sets

SpaceX API information

SpaceX Wikipedia page

Thank you!