IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chloe Drummond
2/19/2025

Outline

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

Executive Summary

Methodology

- Collected data from SpaceX Rest API as well as web scraping tables from Wiki pages.
- Wrangled data for future analysis
- Performed EDA using SQL, Pandas, & Matplotlib
- Created interactive Visuals using Folium and Plotly Dash
- Built ML pipeline to predict if first stage of Falcon 9 lands successfully

Results

- Was able to see early analysis information like success rate
- EDA gave insightful relationships like the launch success trend over the years
- Interactive visualizations allowed better pattern visualization and allowed optimal launch site selection
- Predictive models all performed relatively well with ~83% accuracy. Small test sample is likely main cause of the lack of differentiation.

Introduction

Background & Context

- The new age of space is here. SpaceX is one of the more successful companies aiming to make space travel more affordable. It does this by being able to resuse its first stage of launch. If we can determine if the first stage will successfully land then we can calculate the cost of a launch.

Purpose

- The purpose of this project is to determine the price of each launch for a competing company SpaceY. This includes determining if SpaceX will reuse the first stage and predicting if the first stage will land successfully.

Section 1

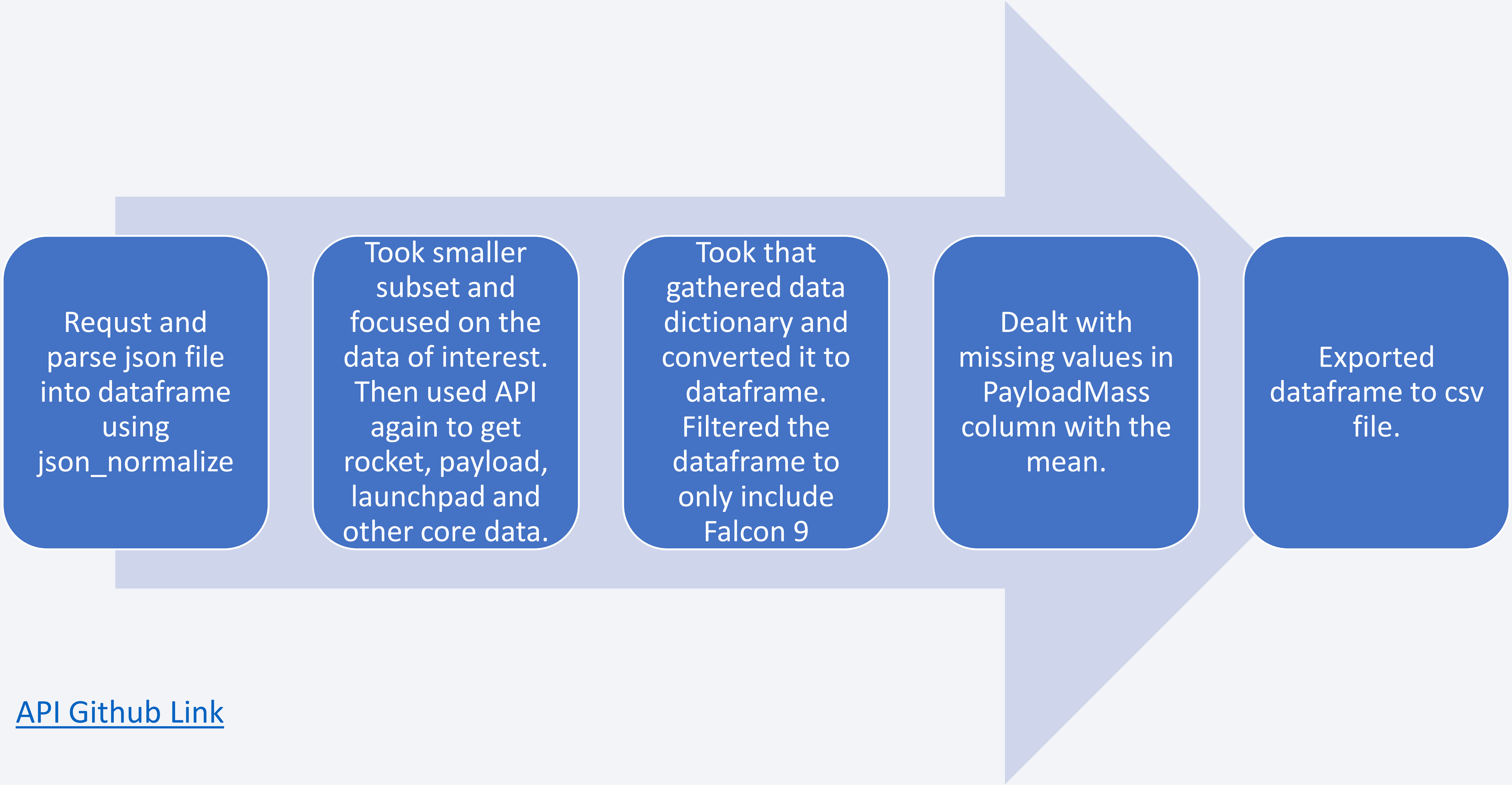
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from API GET requests and Web-scraping from Wiki page.
- Perform data wrangling
  - Used two gathered tables to create class column of successful/failed landings.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split into 80/20 split
  - Performed GridSearchCV for logistic regression, SVC, Decision Tree, KNN
  - Created confusion matrix and found accuracy to determine best model.

Data Collection – SpaceX API



Requst and parse json file into dataframe using json_normalize

Took smaller subset and focused on the data of interest. Then used API again to get rocket, payload, launchpad and other core data.

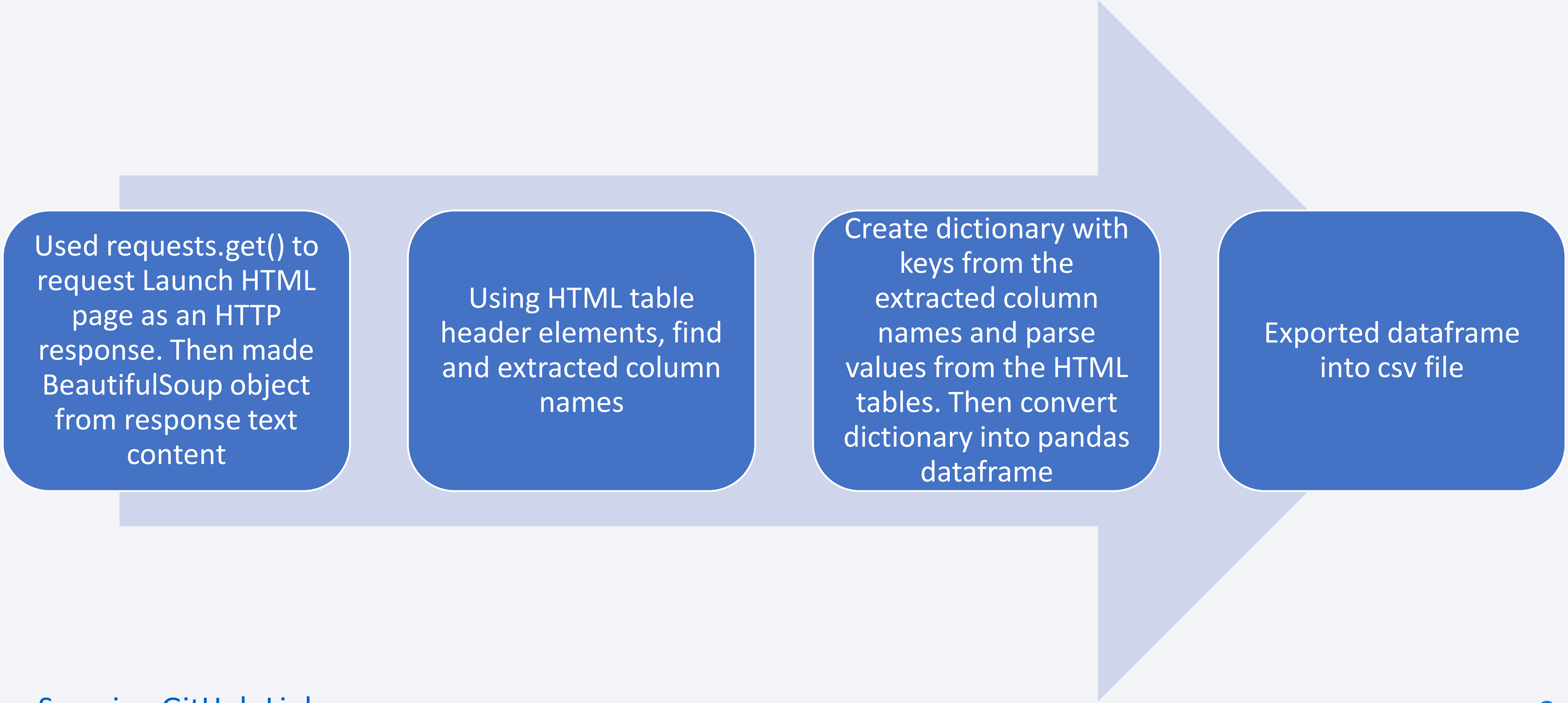
Took that gathered data dictionary and converted it to dataframe. Filtered the dataframe to only include Falcon 9

Dealt with missing values in PayloadMass column with the mean.

Exported dataframe to csv file.

API Github Link

Data Collection - Scraping



Used requests.get() to request Launch HTML page as an HTTP response. Then made BeautifulSoup object from response text content

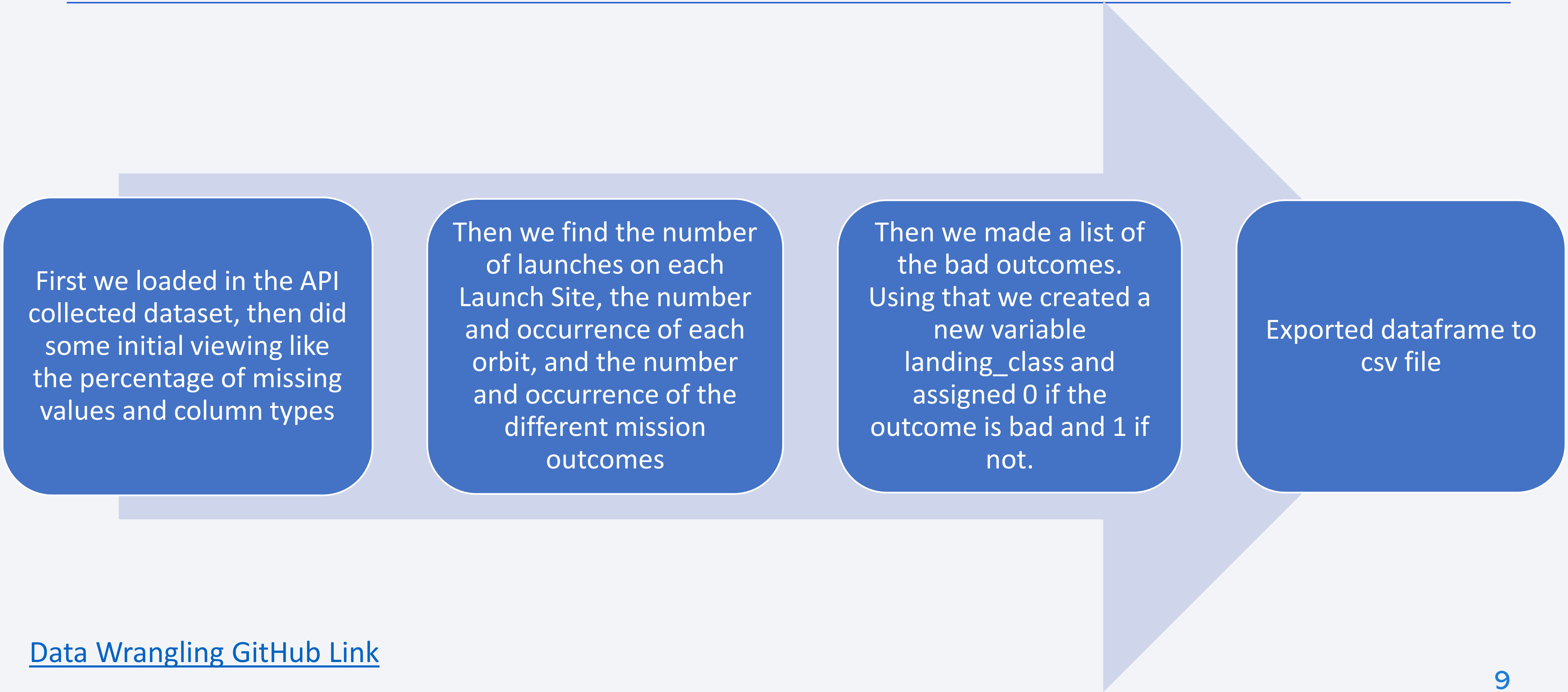
Using HTML table header elements, find and extracted column names

Create dictionary with keys from the extracted column names and parse values from the HTML tables. Then convert dictionary into pandas dataframe

Exported dataframe into csv file

Scraping GitHub Link

Data Wrangling



First we loaded in the API collected dataset, then did some initial viewing like the percentage of missing values and column types

Then we find the number of launches on each Launch Site, the number and occurrence of each orbit, and the number and occurrence of the different mission outcomes

Then we made a list of the bad outcomes. Using that we created a new variable landing_class and assigned 0 if the outcome is bad and 1 if not.

Exported dataframe to csv file

Data Wrangling GitHub Link

EDA with Data Visualization

Github Link

- Scatterplot of Payload Mass and Flight Number colored by Class
  - Shows the spread of the payload mass by success.
- Scatterplot of Launch Site and Flight Number colored by Class
  - See spread of different flights at each site and their outcome
- Scatterplot of Payload and Launch site colored by Class
  - Observe relationship between sites and mass.
- Bar Chart for success rate of each orbit
  - See relationship of success rate for each orbit.
- Scatterplot of Flight number vs Orbit and Payload vs Orbit
  - See relationship between two variables
- Line Chart of Year vs Success Rate
  - View yearly trend of success

EDA with SQL

- Display unique launch site names from SpaceXtable
- Display first 5 records with launch sites that start with CCA
- Display total payload mass carried for the customer NASA
- Display average payload mass for booster version F9 v1.1
- List the date when first successful landing in ground pad happened
- List the booster names which have success in drone ship and have payload mass between 4000 and 6000 kg
- List the total number of successful and failure mission outcomes
- Use subquery to list names of booster versions that have carried the max payload mass.
- List the records which display months, failure landings in drone ship, booster versions and launch site for months in 2015
- List each landing outcome type and occurrence between 2010-06-04 and 2017-03-20 in descending order of number of occurrences

Github Link

Build an Interactive Map with Folium

- Center location is Johnson Space Center in Houston, Texas
- Added circle object for each launch site with name as a popup label
  - Launch sites are mainly by the coasts and are somewhat close to equator line
- MarkerCluster object created to add markers for all launch records in colors related to their class.
  - Marker cluster used since many launch records will have same coordinates
- Mouse position to get any coordinate on a map, create marker to show distance between launch site and mouse coordinate.
  - Used to show distance between launch site and nearest coastline
- PolyLine added between launch site and selected coastline point

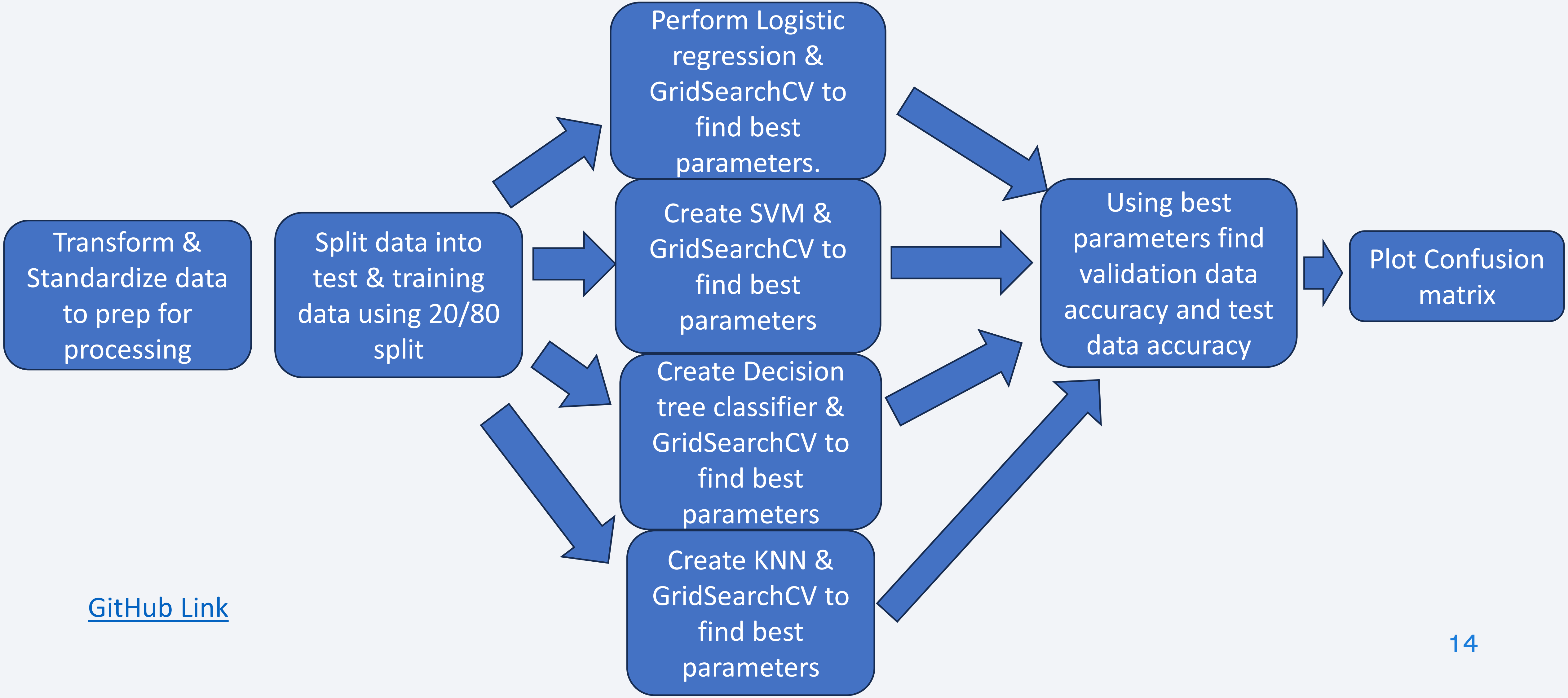
Github Link

Build a Dashboard with Plotly Dash

- Dropdown menu to chose between site specific information or total for all sites
- Pie chart for all sites shows success rate for each site. Site specific pie chart shows percentage of success and failures for that site
- Scatterplot of Payload vs Success for all sites or one specific site based on dropdown selection. Colors show different booster versions
- Payload range slider to filter different ranges of mass on the scatterplot.

GitHub Link

Predictive Analysis (Classification)



Transform & Standardize data to prep for processing

Split data into test & training data using 20/80 split

Perform Logistic regression & GridSearchCV to find best parameters.

Create SVM & GridSearchCV to find best parameters

Create Decision tree classifier & GridSearchCV to find best parameters

Create KNN & GridSearchCV to find best parameters

Using best parameters find validation data accuracy and test data accuracy

Plot Confusion matrix

GitHub Link

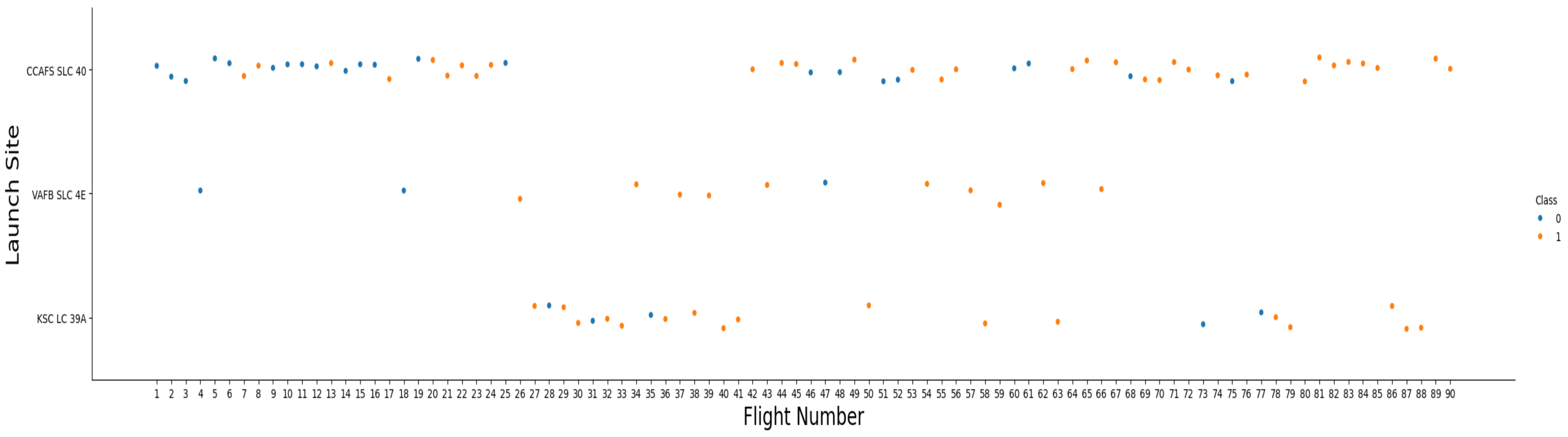
Results

- EDA shows that time and repetition is beneficial to the success chances of landing the first stage since there is a positive correlation between success rate and flight number and the yearly trend shows a positive increase in success rate over increasing years.
- Payload mass does seem important as well as the smaller payloads seem to be more successful
- Interactive plots show that site KSC LC-39A was most successful. This site was less than 8 KM away from any coast, highway, and railway. This could lead to lower costs in materials transportation so that funds can be better allocated since the site is so close to transportation sites.
- Decision Tree gave the best validation accuracy of 0.875 while all of the models gave the same testing accuracy of 0.8333 since we had a small testing data size.

Section 2

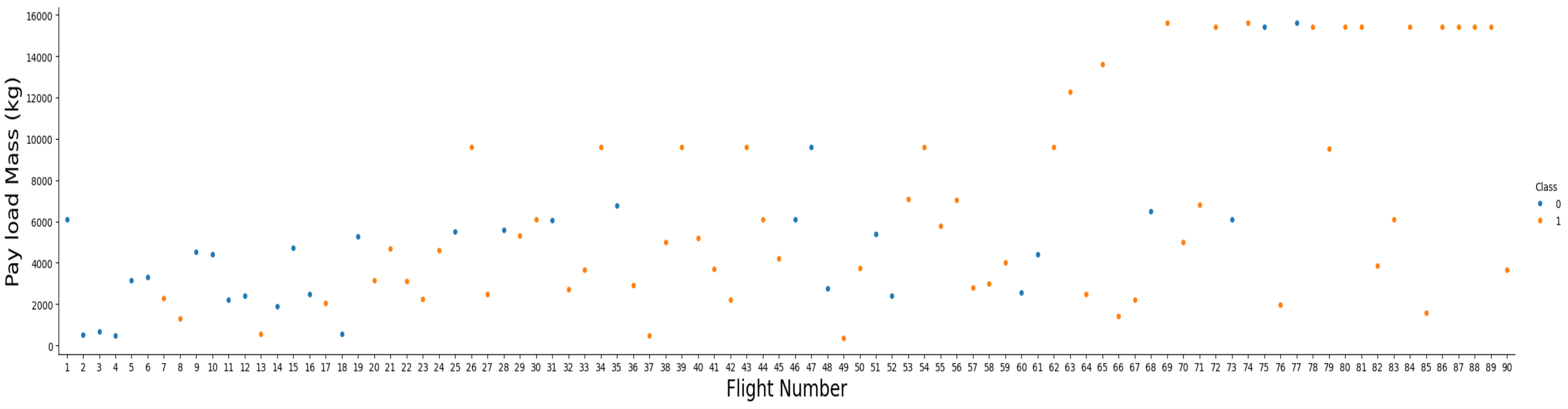
Insights drawn from EDA

Flight Number vs. Launch Site



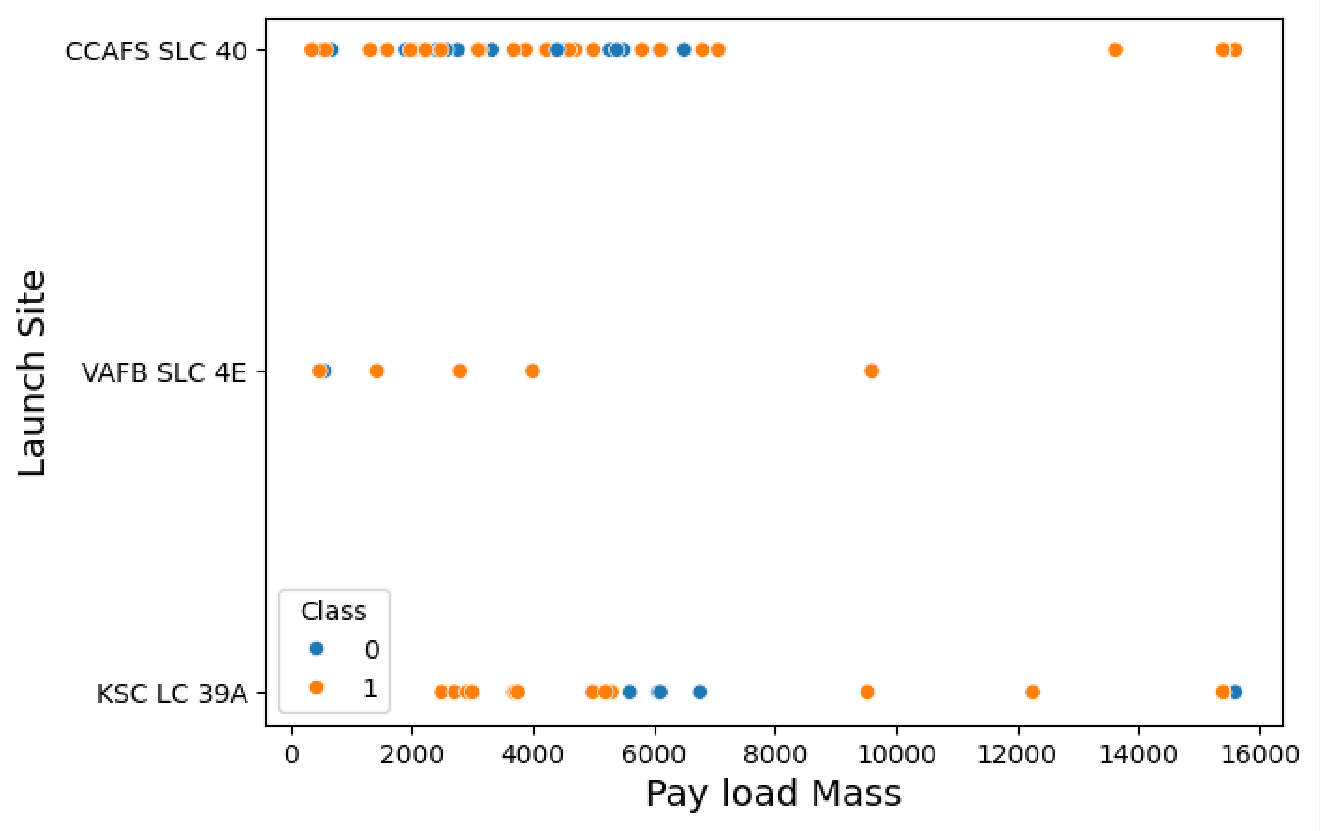
Least amount of flights at VAFB SLC 4E while most at CCAFS SLC 40. Later flights more successful

Flight Number vs. Payload Mass



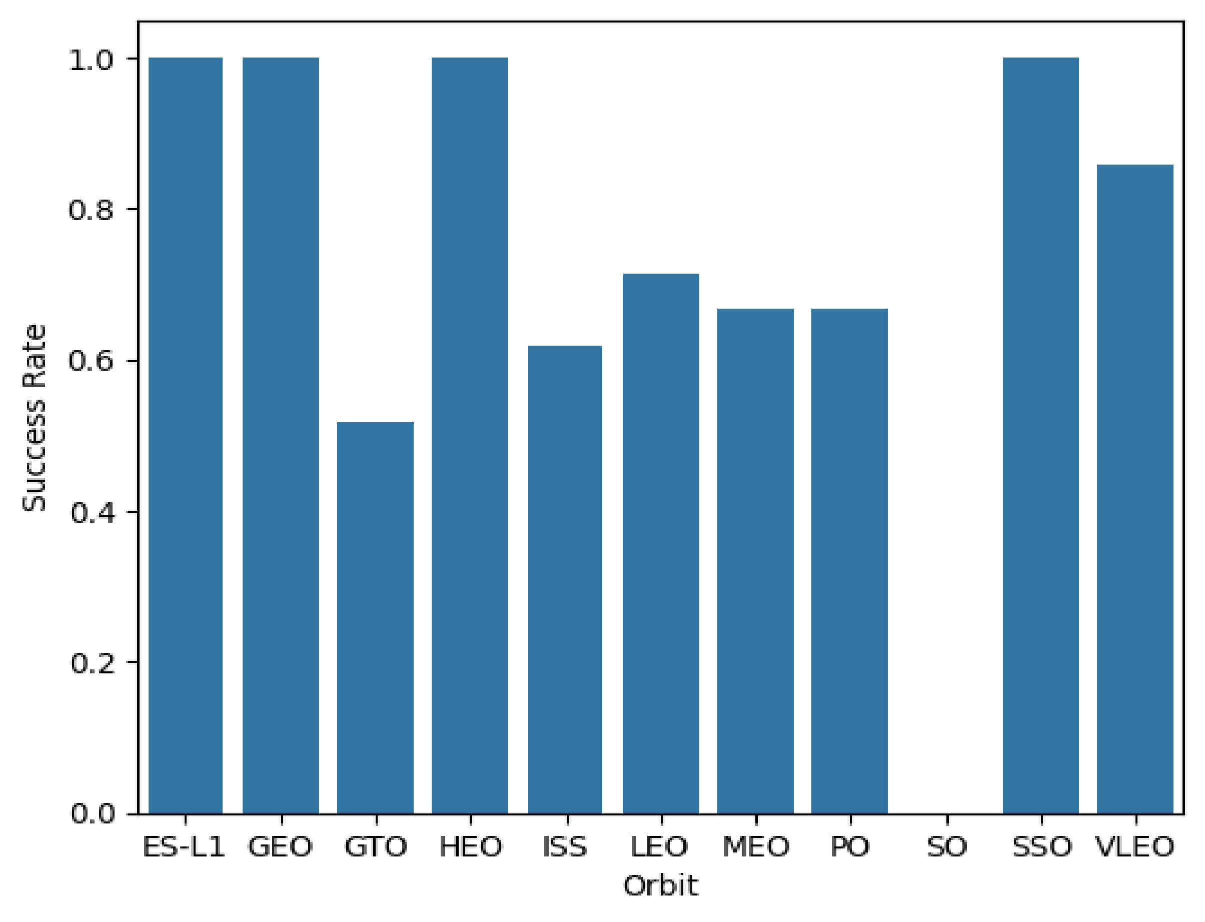
It seems that the more massive the payload the less likely the first stage is to return.

Payload vs. Launch Site



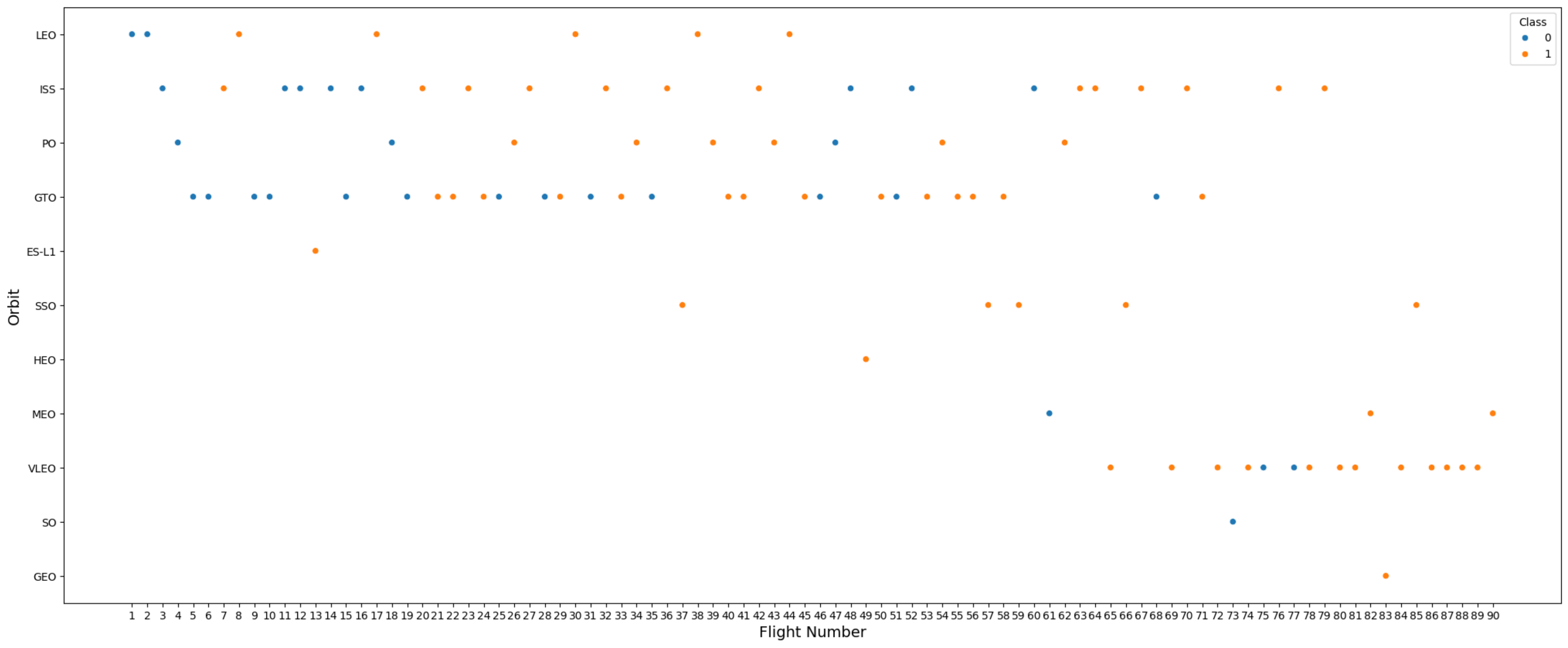
Site VAFB SLC 4E had no launches for mass greater then 10000. Site CCAFS SLC 40 had 3 success with mass greater than 12000

Success Rate vs. Orbit Type



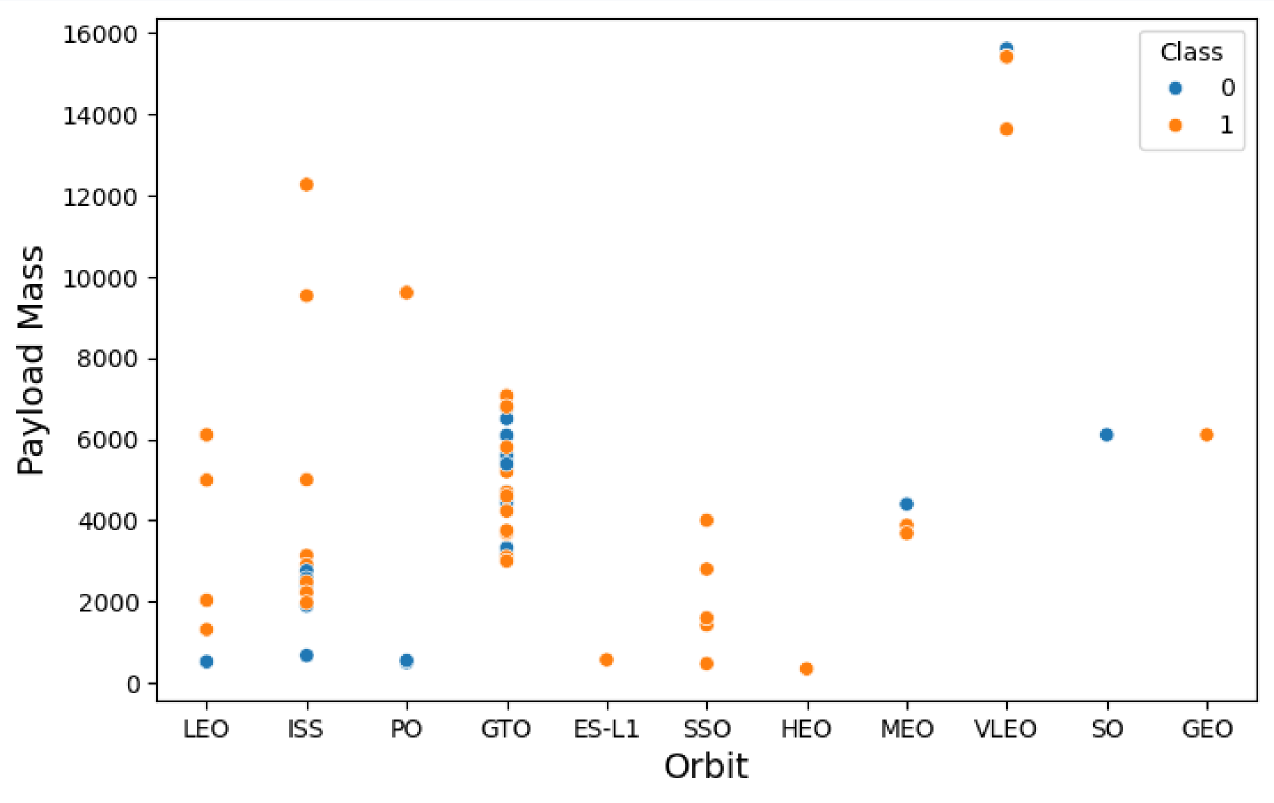
Orbits ES-L1,EO, HEO, SSO had highest success rate

Flight Number vs. Orbit Type



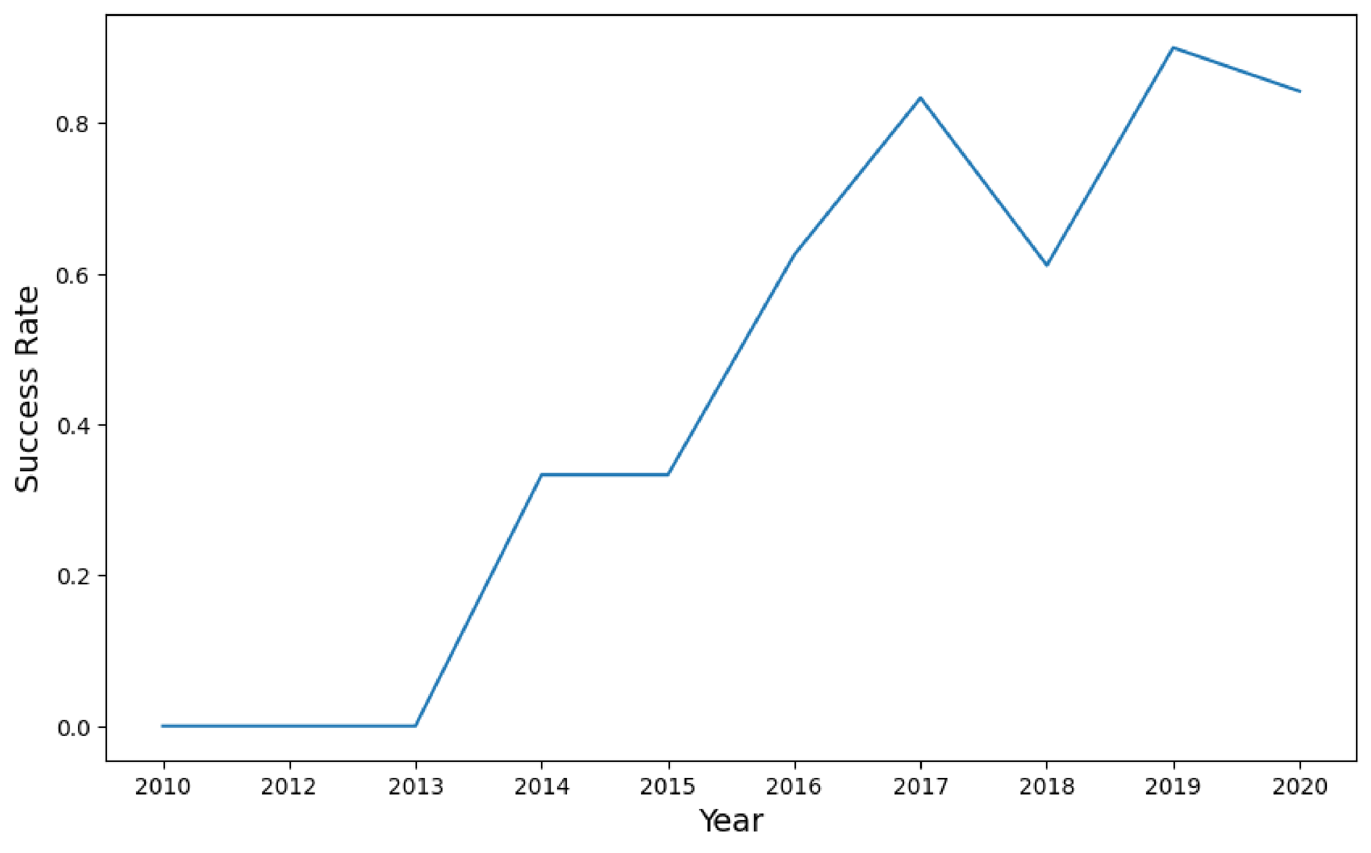
Leo orbit Success is related to the flight number, while no relationship in other orbits.

Payload vs. Orbit Type



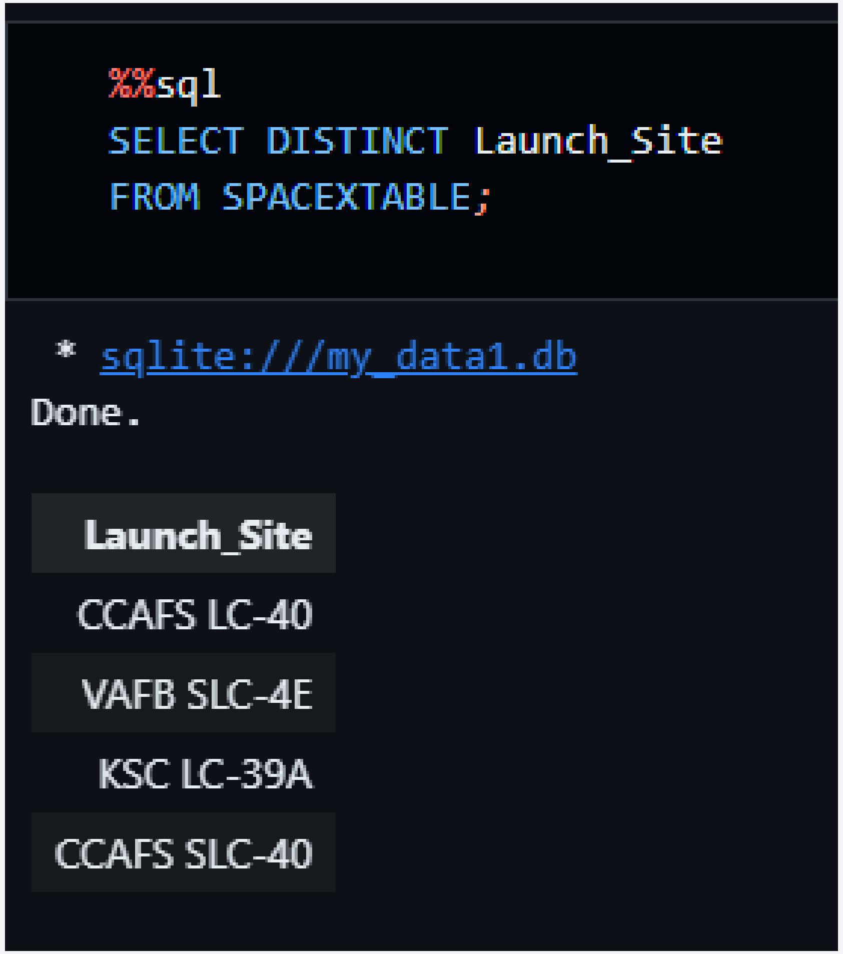
Heavy mass have positive success rate more for Polar, LEO, ISS

Launch Success Yearly Trend



Success increases since 2013 with some leveling out in 2014 and a small drop in 2017.

All Launch Site Names



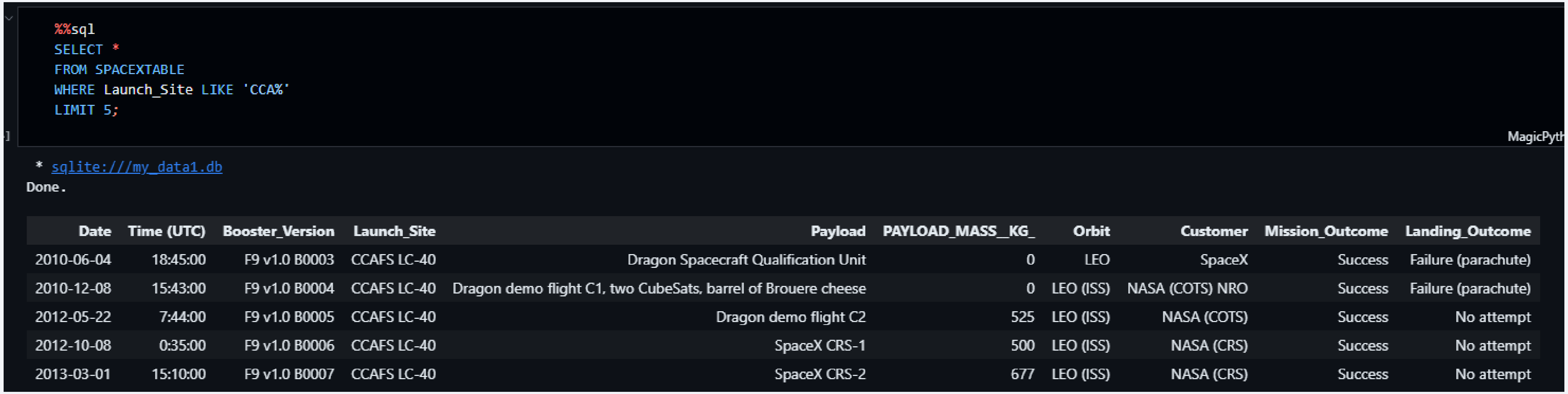
```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Extracted the unique launch site names from the table

Launch Site Names Begin with 'CCA'



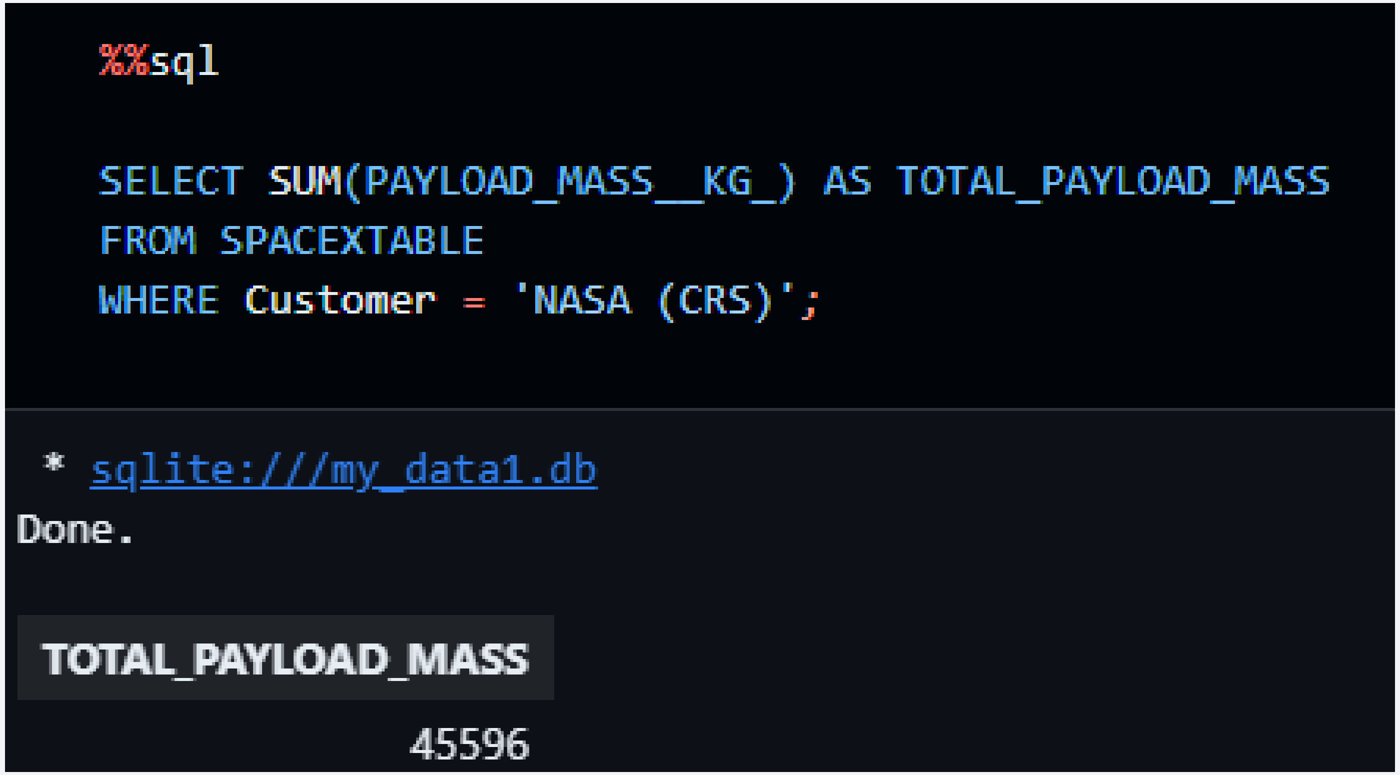
```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Used like to find the launch sites from the table that begin with CCA but have anything afterwards with the %. The limit 5 shows only the first 5 entries.

Total Payload Mass

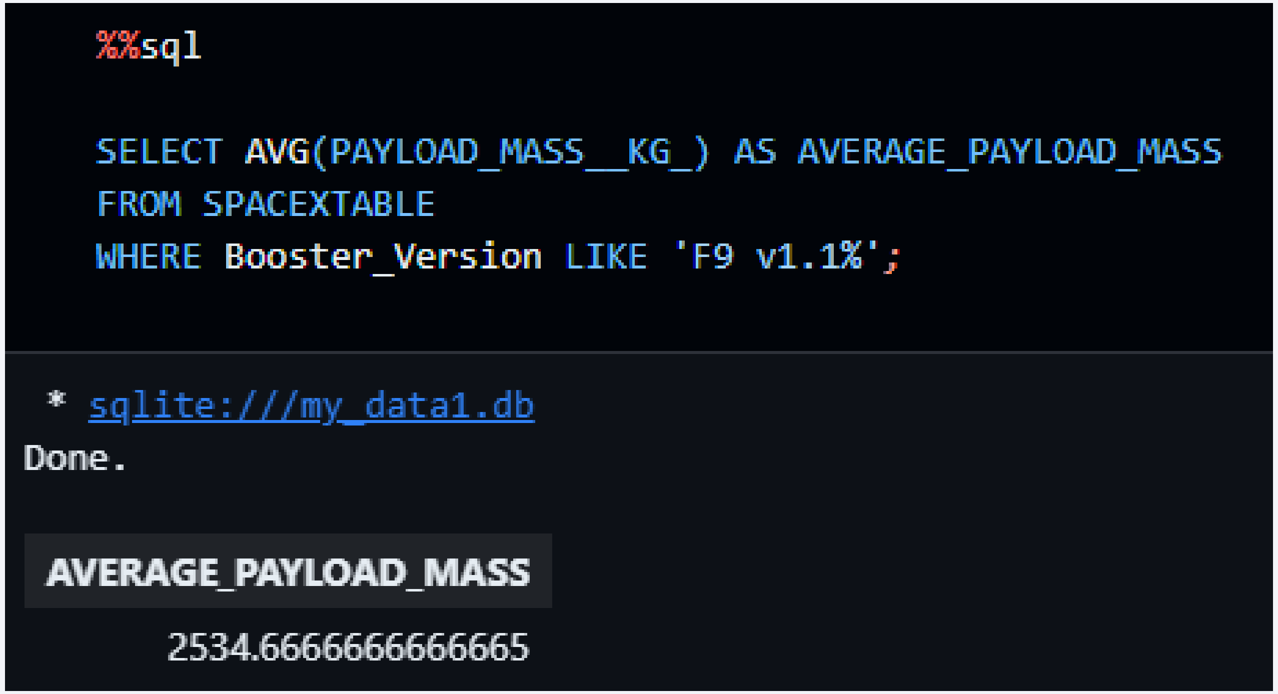


```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.

TOTAL_PAYLOAD_MASS
             45596
```

Finds the sum of the payload mass from the flights where the customer is Nasa. Assigns it to Total_payload_mass.

Average Payload Mass by F9 v1.1



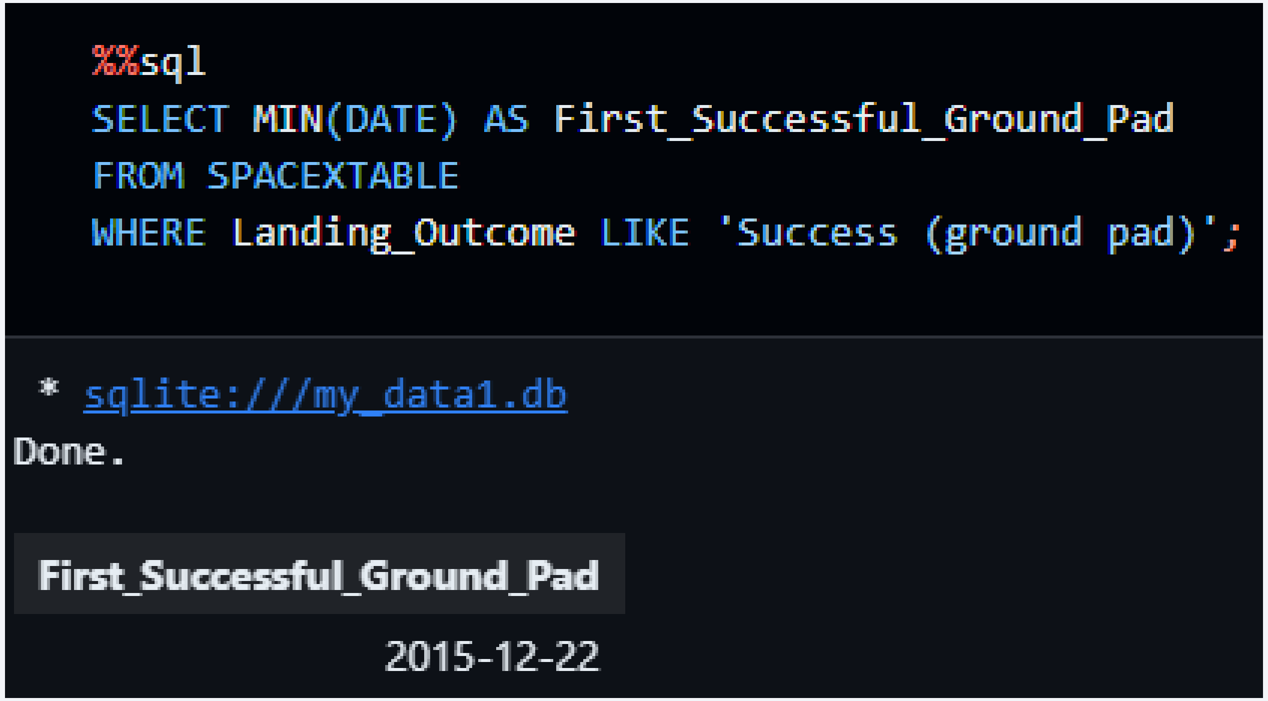
```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS
FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
```

AVERAGE_PAYLOAD_MASS
2534.6666666666665

Found average of the payload mass where the booster version was F9 v1.1.

First Successful Ground Landing Date



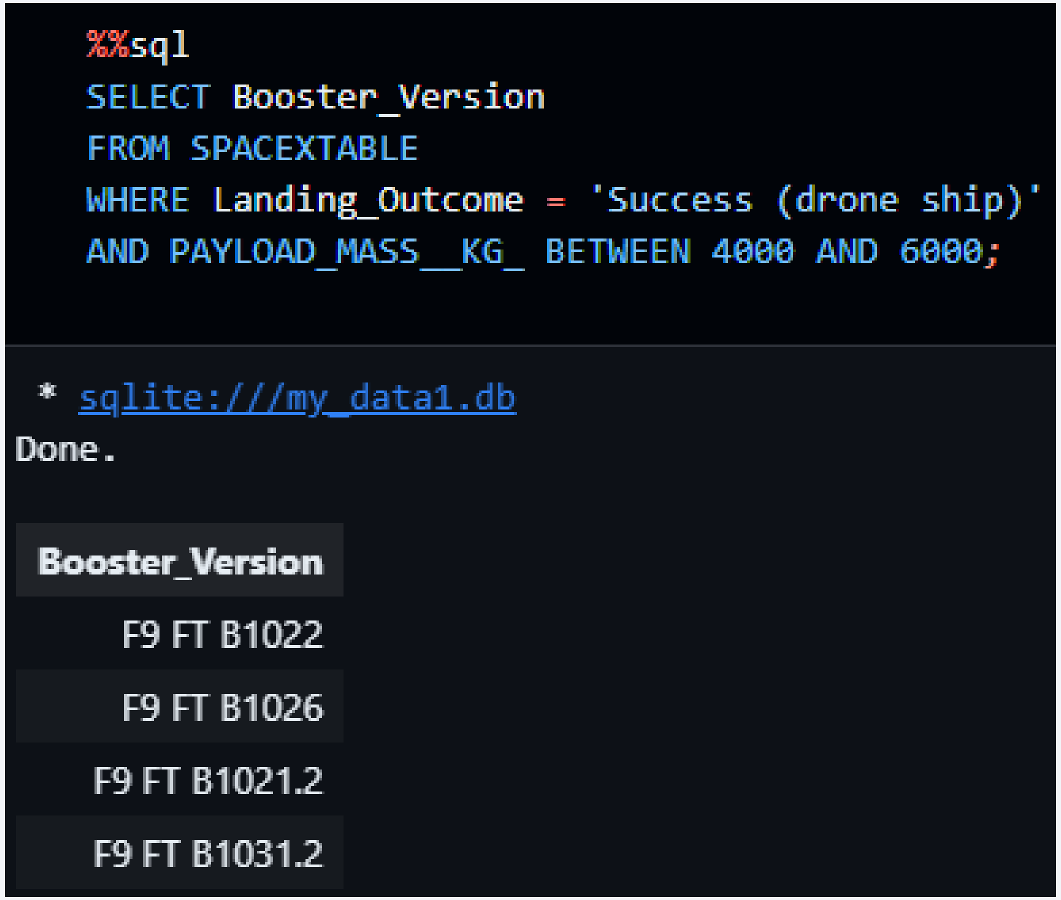
```
%%sql
SELECT MIN(DATE) AS First_Successful_Ground_Pad
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

First_Successful_Ground_Pad
2015-12-22

Using the min function on date it finds the first date where the landing outcome is success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000



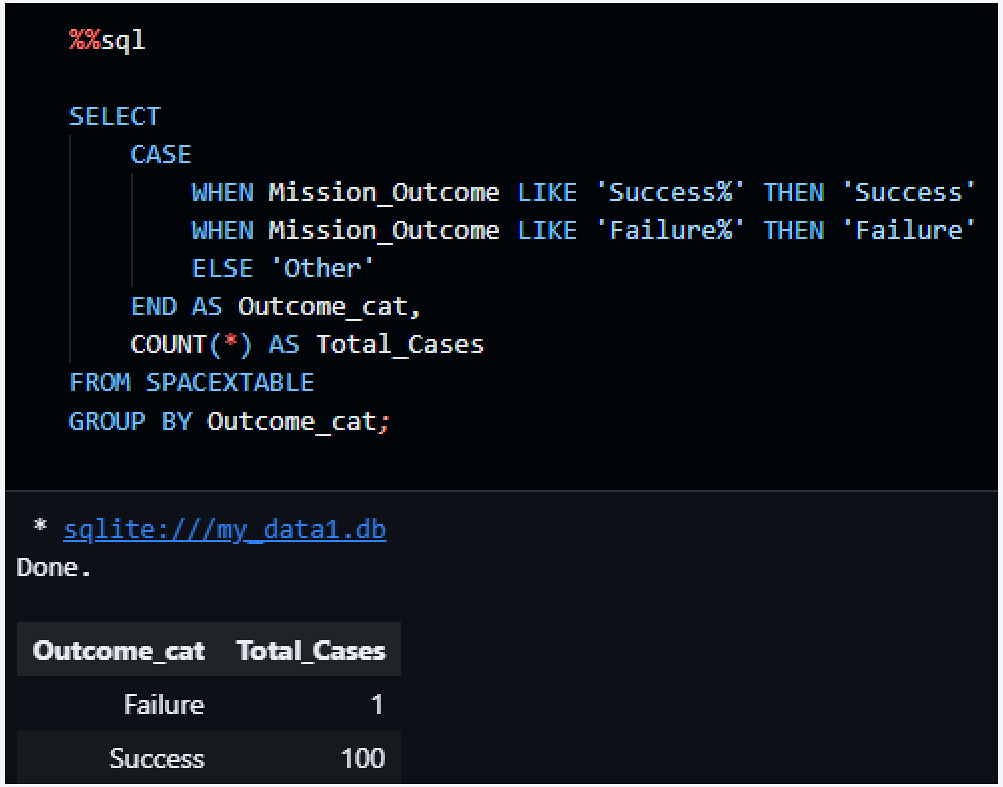
```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

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

These Booster versions had a successful drone ship landing with payload masses between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes



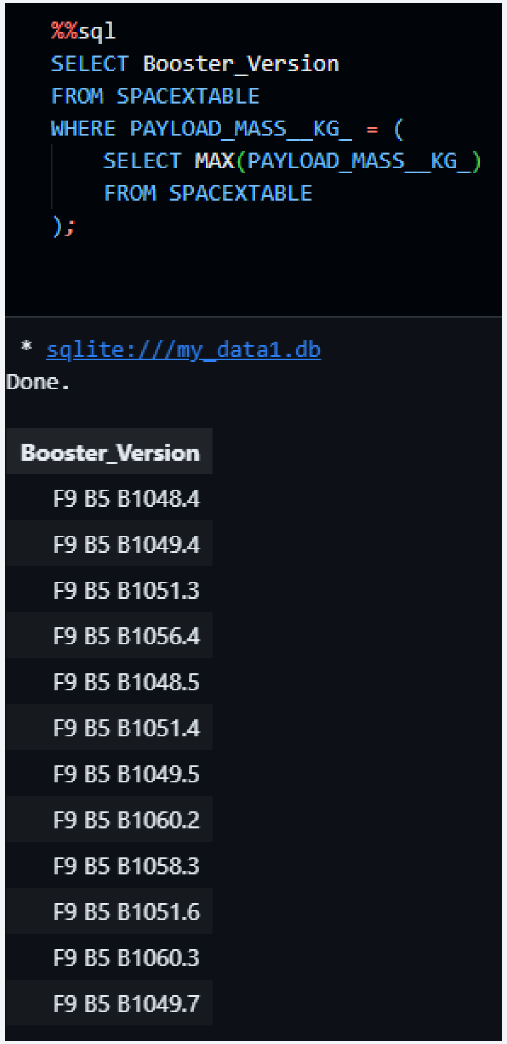
```
%%sql

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

```
 * sqlite:///my_data1.db
Done.
```

Outcome_cat	Total_Cases
Failure	1
Success	100

Only 1 failure and 100 successes

Boosters Carried Maximum Payload



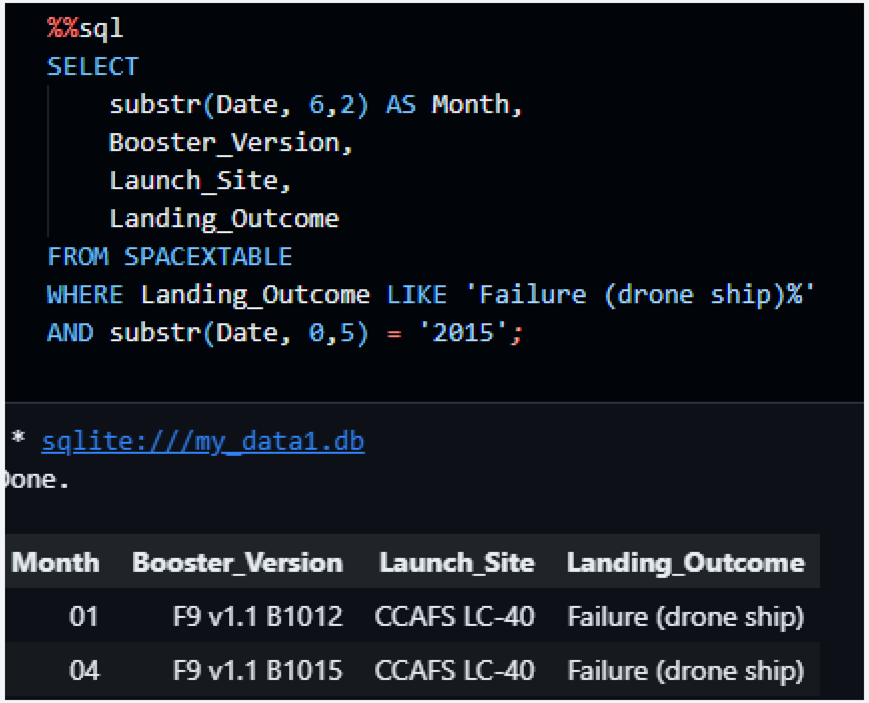
```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

```
 * sqlite:///my_data1.db
Done.
```

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

Suquery of select max() to find booster versions that carry the max payload mass

2015 Launch Records



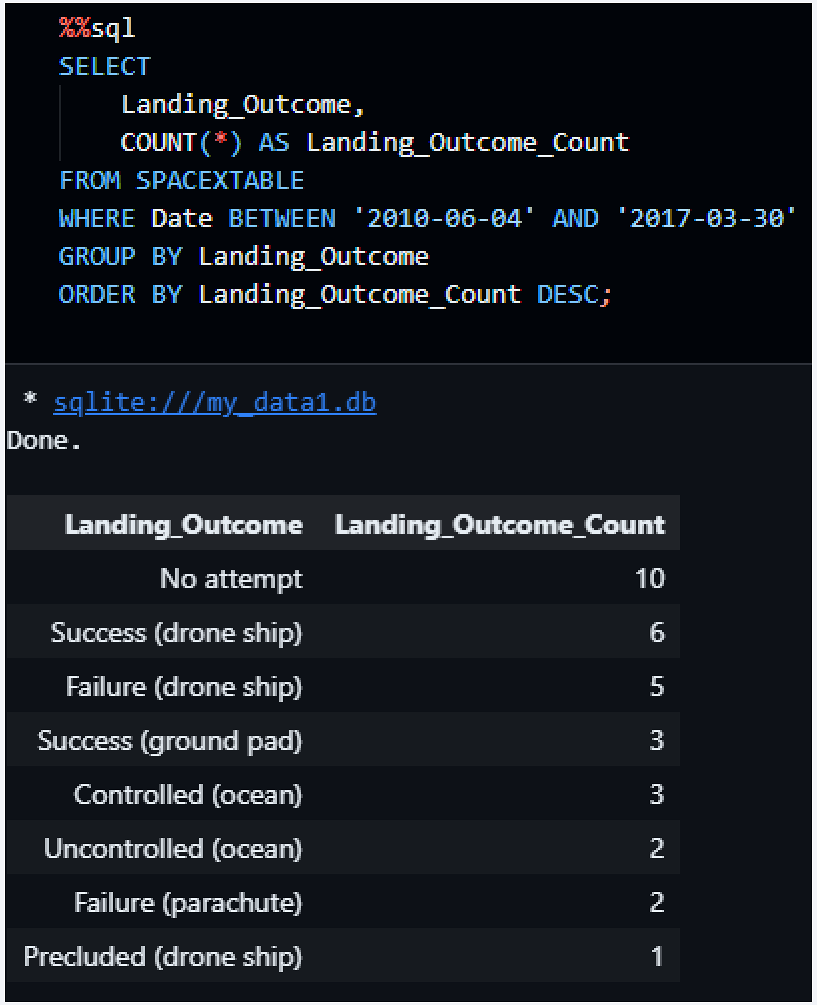
```
%%sql
SELECT
    substr(Date, 6,2) AS Month,
    Booster_Version,
    Launch_Site,
    Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure (drone ship)%'
AND substr(Date, 0,5) = '2015';
```

```
* sqlite:///my_data1.db
one.
```

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

Months January and April have the two records where there was a failure drone ship landing.

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



```
%%sql
SELECT
    Landing_Outcome,
    COUNT(*) AS Landing_Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-30'
GROUP BY Landing_Outcome
ORDER BY Landing_Outcome_Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

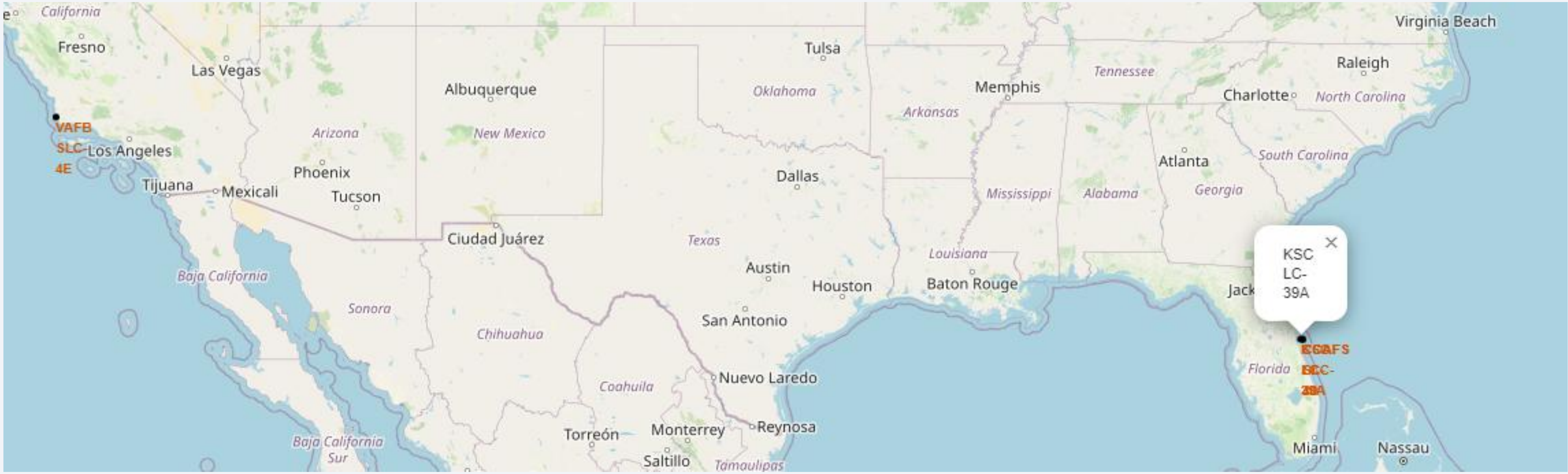
Landing_Outcome	Landing_Outcome_Count
No attempt	10
Success (drone ship)	6
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

No attempt at landing had the most occurrences while Precluded (drone ship) only had 1

Section 3

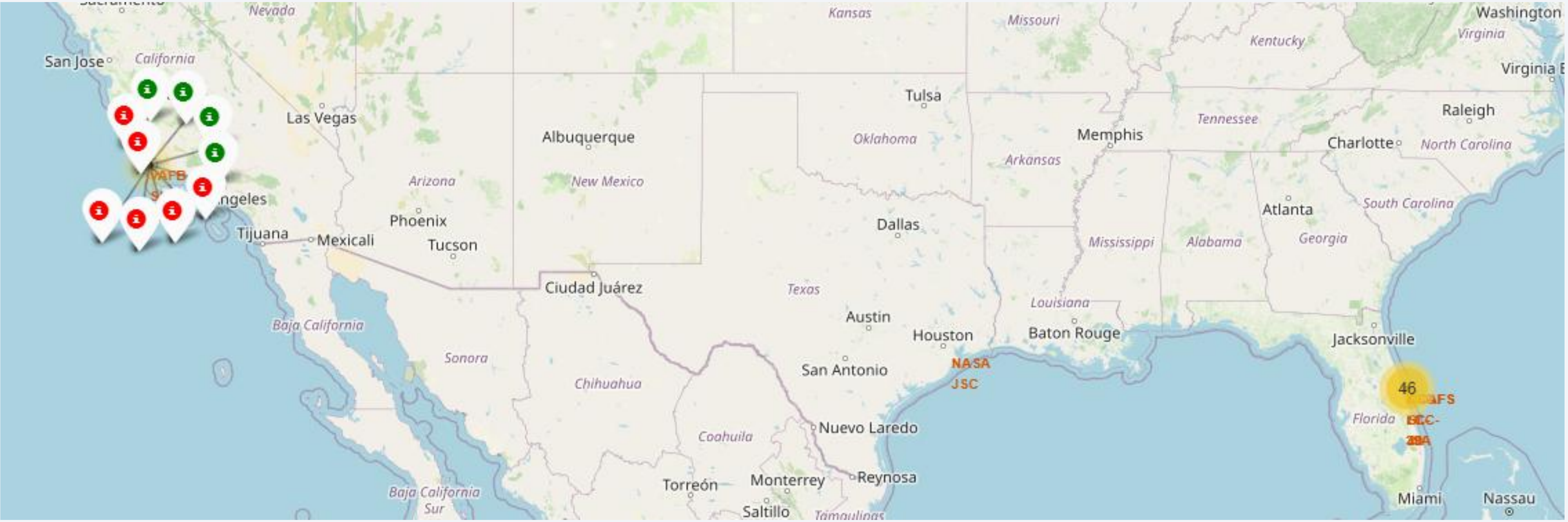
Launch Sites Proximities Analysis

Launch Site Markers



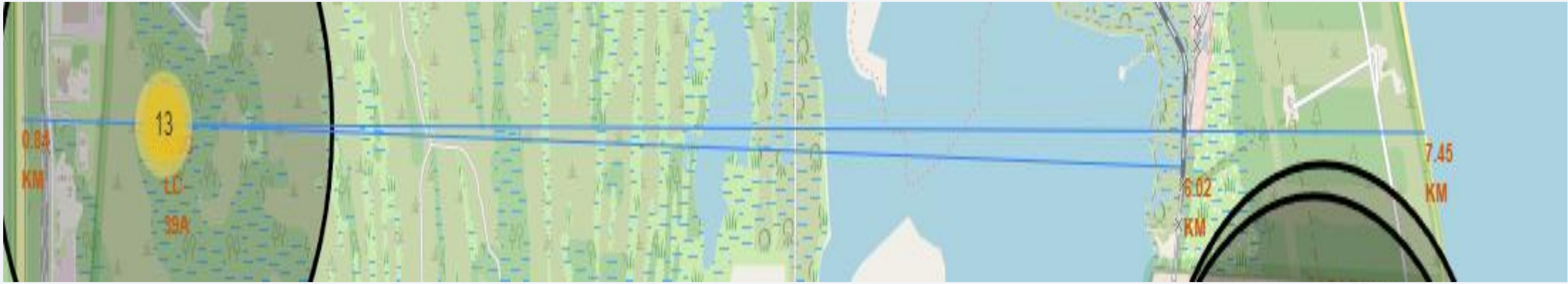
Launch sites are along the coast close to the equator while staying in the USA

Successful and Failed Landings at Launch Sites



Green is successful landing and red is a failed landing

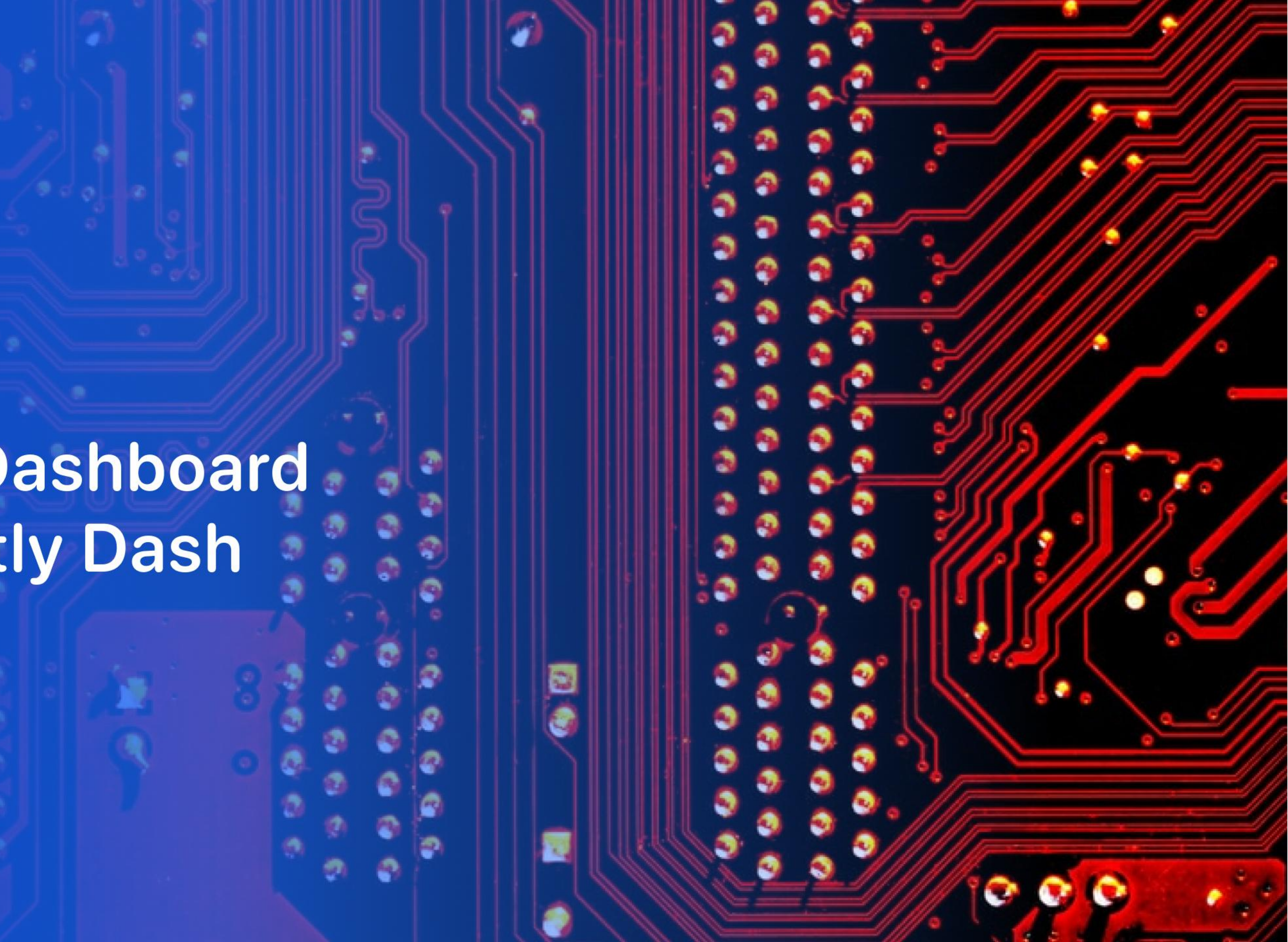
Distance to Nearest Highways, Railroads, Coasts



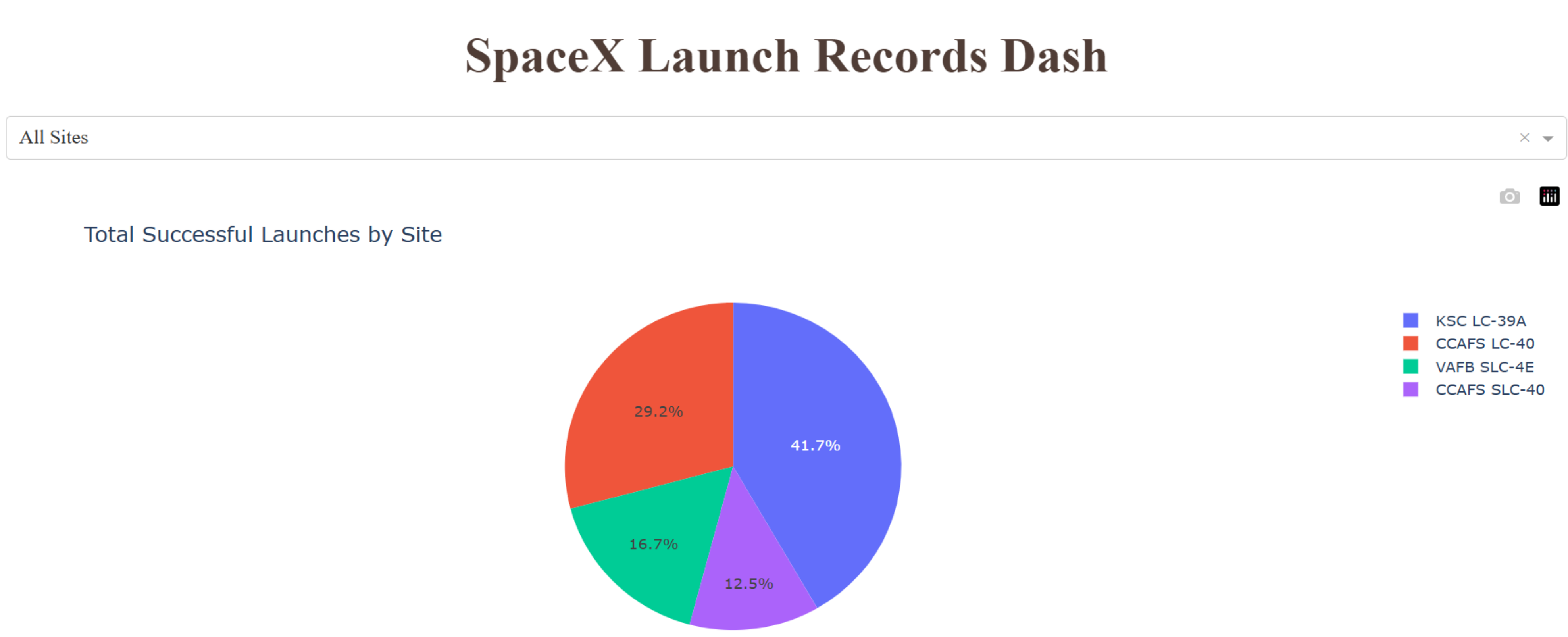
Chose site KSC LC-39A since it has the best success rate. It was no more than 8 KM away from any major point.

Section 4

Build a Dashboard with Plotly Dash



Successful Launches by Site



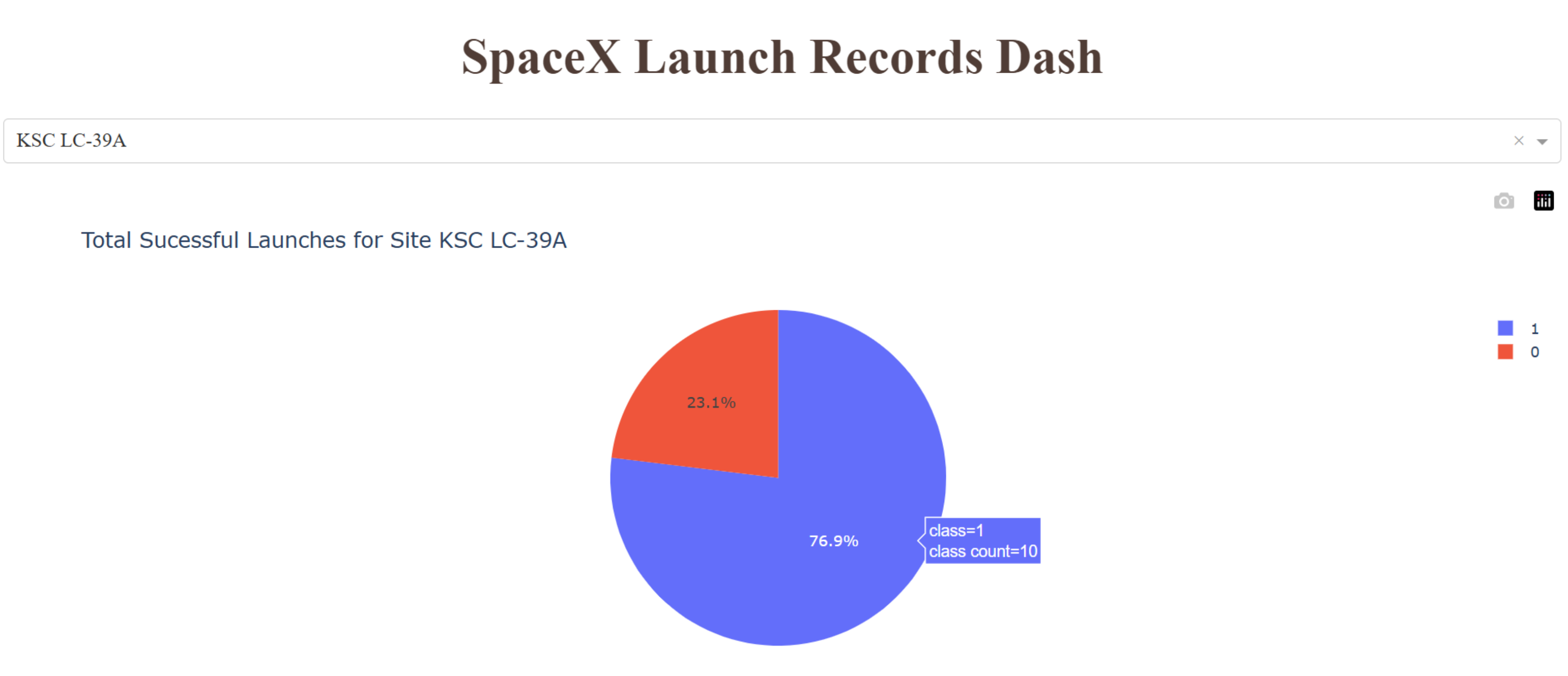
SpaceX Launch Records Dash

All Sites

Total Successful Launches by Site

- The launch site that was most successful is the KSC LC-39A site while the least successful site is the CCAFS SLC-40 site.

KSC LC-39A Breakdown



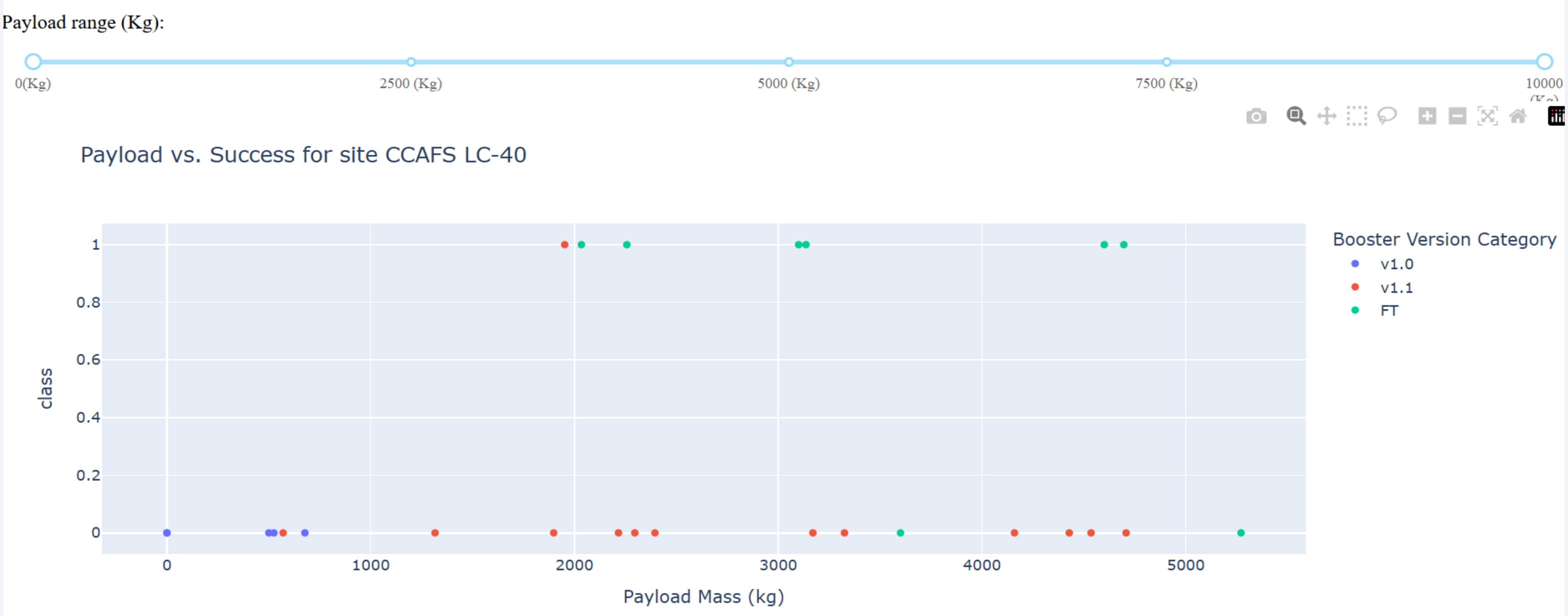
SpaceX Launch Records Dash

KSC LC-39A

Total Sucessful Launches for Site KSC LC-39A

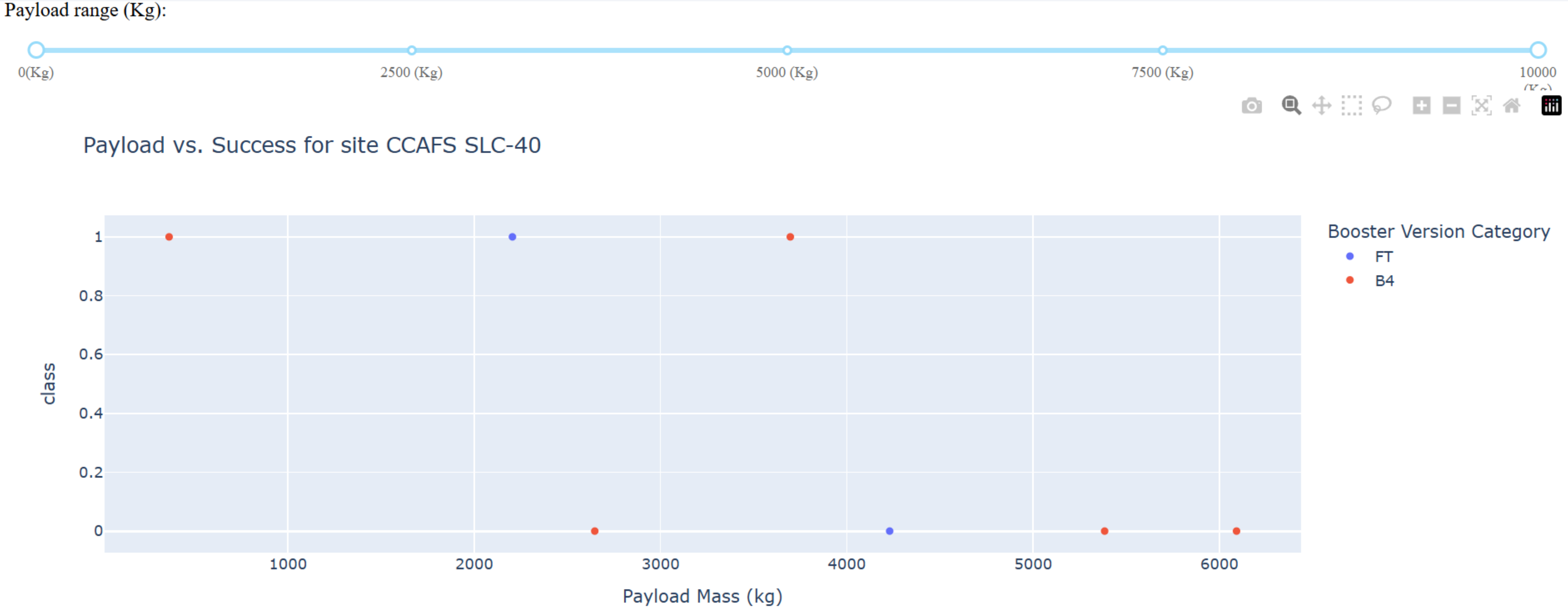
- Out of 13 total launches site KSC LC-39A had 10 successful launches and 3 failures

Payload Vs Success site CCAFS LC-40



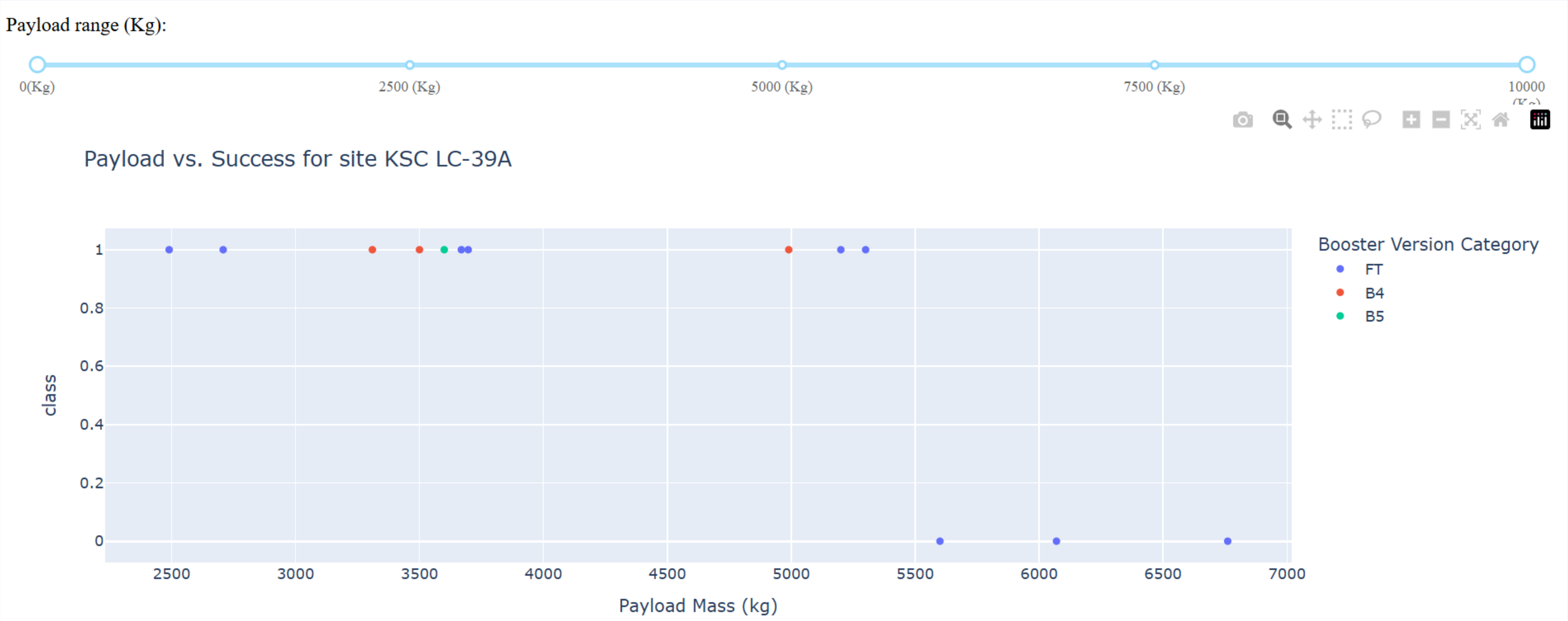
- For this site, the FT booster version had the most success. Version v1.0 failed in all attempts here and never had a payload larger than 1000KG while v1.1 had one success around 2000KG it failed most of its attempts at a wide variation of payloads

Payload Vs Success site CCAFS SLC-40



- For this site, there were only 3 of 7 successful attempts. The booster version FT only had 2 attempts, 1 success and 1 failure and ranged between a payload of 2000KG and 4500KG. Booster version B4 had 5 attempts 3 failures and 2 successes across many different payloads from under 1000KG and a little over 6000KG

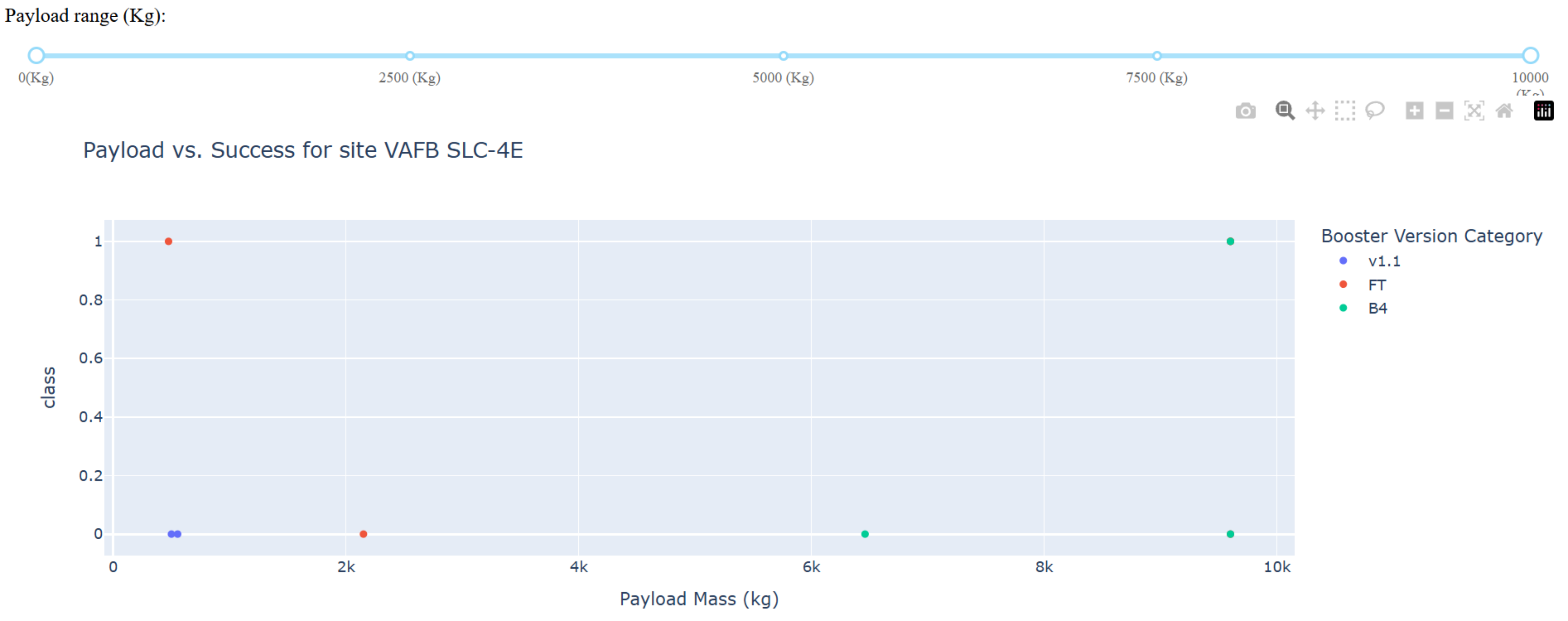
Payload Vs Success site KSC LC-39A



Payload range (Kg):

- For this site, there were 10 successes and 3 fails. Booster version FT had the most attempts with 9 total, with 6 successes and 3 fails, across payloads from just below 2500KG to around 6800KG. Version B4 had 3 attempts all successes and version B5 had 1 successful attempt.

Payload Vs Success site VAFB SLC-4E



Payload range (Kg):

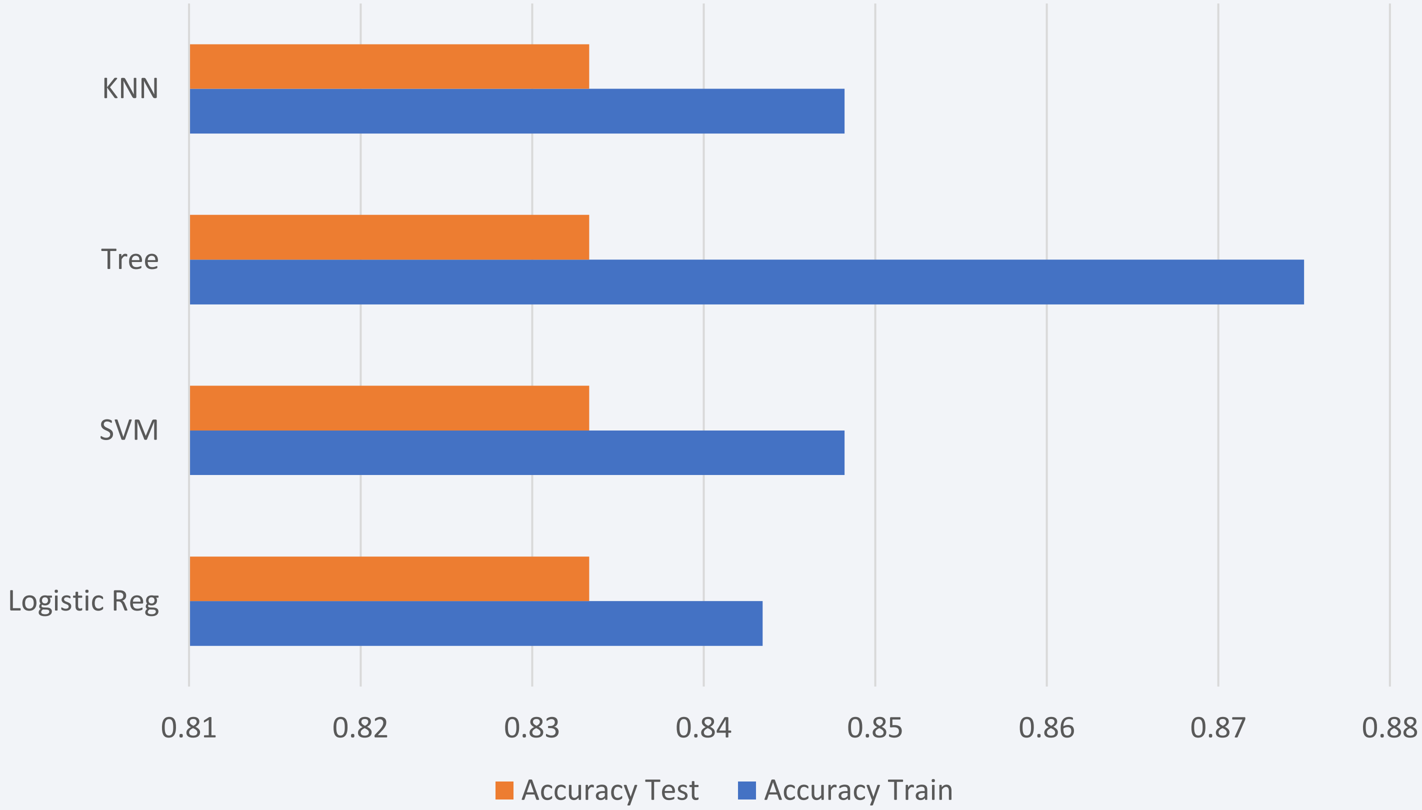
Payload vs. Success for site VAFB SLC-4E

- For this site, there was only two successful attempts with widely variating payloads. Both of version's v1.1 attempts were failures. Version FT had success at the lower payload around 500KG but failed around 2100KG. Version B4 failed twice and succeeded once with heavier payloads than the other two versions.

Section 5

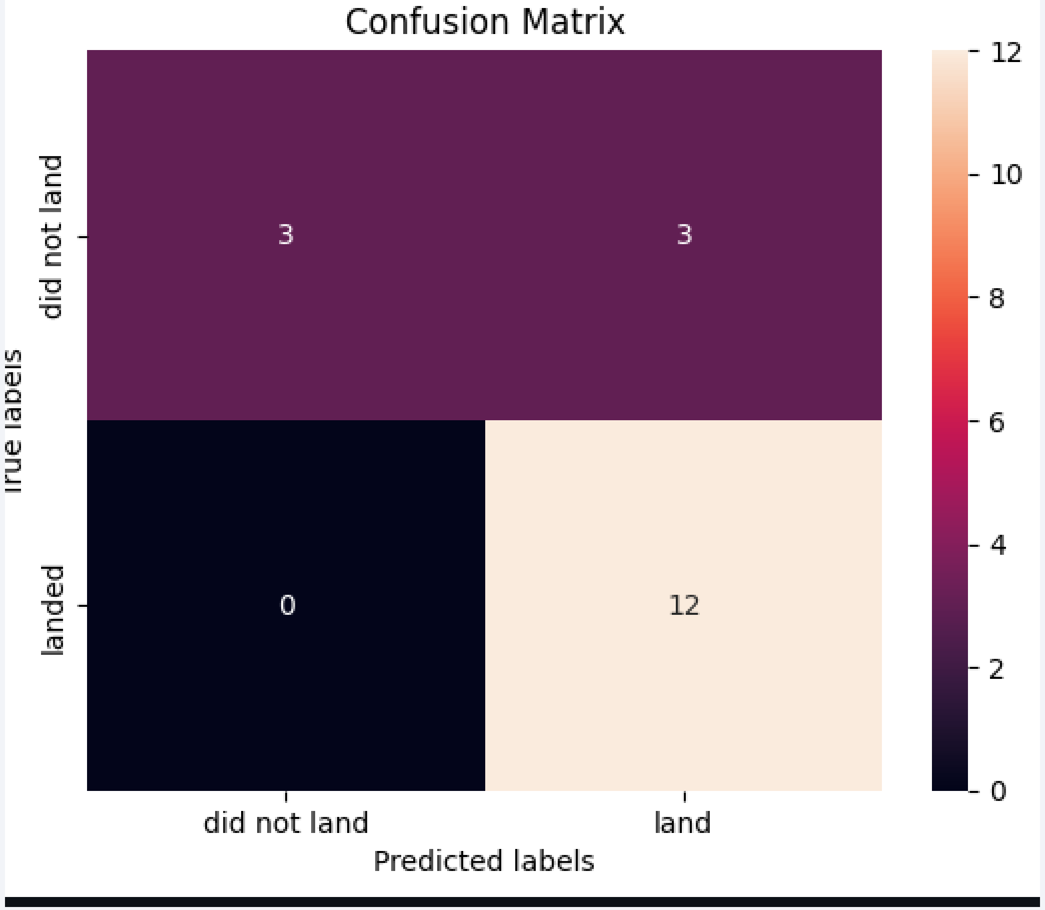
Predictive Analysis (Classification)

Classification Accuracy



- Accuracy on the test data was the same for all different models at around 83.333%. This is mainly due to the small test sample size of only around 18 observations.

Confusion Matrix

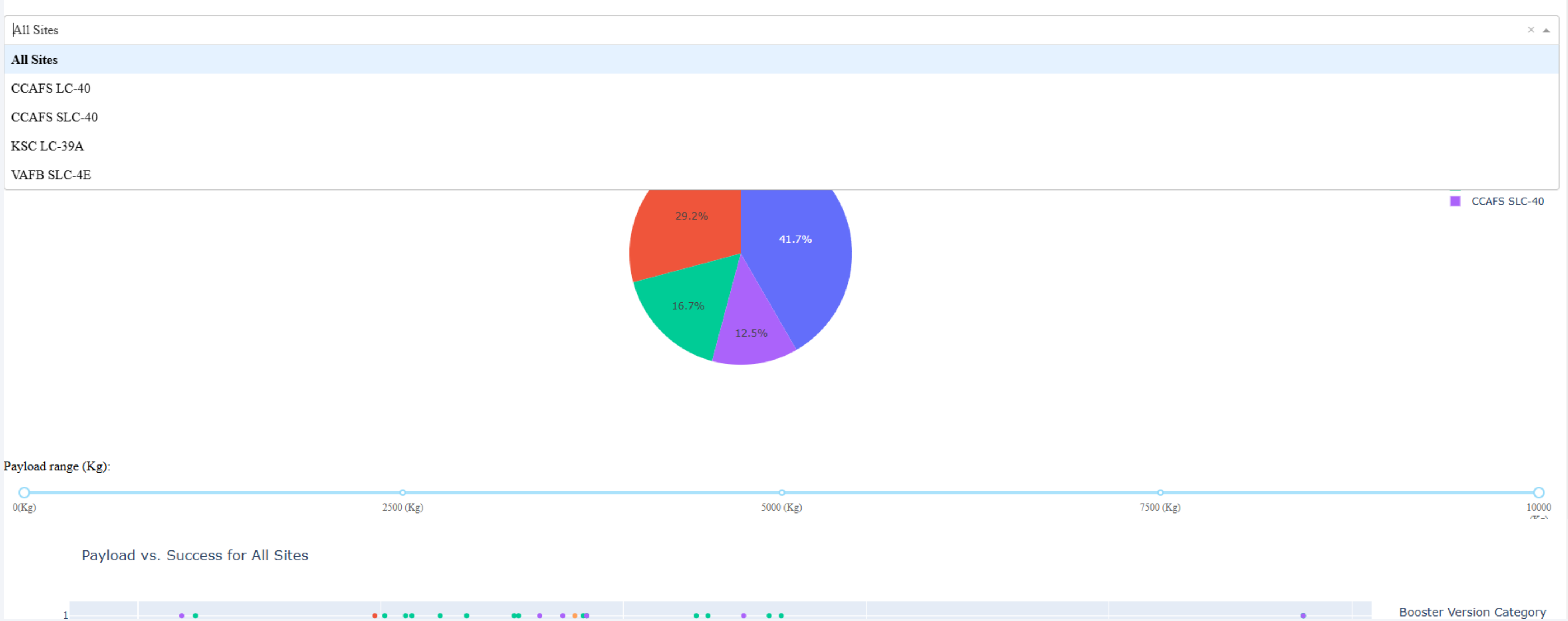


Confusion matrix was the same for all models as well as the accuracy on the test data for each model. Since the test data size was so small. It gave 3 false positives where it predicted landed when in truth it was not landed.

Conclusions

- If SpaceY is a new high-risk it could need some time to see positive results as is shown by the positive correlation with flight number and success rate as well as the success rate increase in the yearly trend.
- If new launch sites are to be used, they should be on the coast of an ocean and close to the equator.
- Payload is important and it seems like the smaller the payload is better to land successfully and thus reuse the first stage.
- Also having the launch site be close to a railway and highway allows for easy transportation of materials which limits costs. Launch sites also tend to be away from cities which makes sense to limit any potential risk from a failed landing.
- Decision trees gave the best training accuracy. However, I would suggest a larger testing data size before a solid opinion can be made since the testing accuracy was the same across all models and the confusion matrix showed the error came from false positives which can lead to some potentially fatal or high-risk situations.

Appendix



Shows dropdown menu options, and the full slider displayed. Also, a peak of the scatterplot to show that it is plotting for all sites as the dropdown menu selects

Appendix

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Wikipedia Link used for Webscraping data collection

https://github.com/cjdrummond/Falcon9-Capstone

Link to GitHub Repository

Thank you!