IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ponto Gift Malahlela

29 July 2022

Outline

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

Executive Summary

- Data Collection with SpaceX API as well as with Web Scrapping using BeautifulSoup package in Python.
- Data cleaning (e.g., removing NULLs) and data preprocessing (e.g., rescaling data, One Hot encoding etc.)
- Exploratory data analysis using SQL, Folium, as well as with Seaborn and Matplotlib package in Python.
- Useful Feature selection for predictive analysis.
- Predictive analysis using various Machine Learning models.

Introduction

SpaceX is an American spacecraft manufacturer, space launch provider and a satellite communications company. The company was founded in 2002 by Elon Musk, with aim of reducing space transportation costs to its client by using reusable rockets. SpaceX Falcon 9 rocket launches cost 62 million dollars, while other providers cost up to 165 million dollars each, i.e. because SpaceX can reuse the first stage.

We will use classification modeling to predict whether the Falcon 9 first stage will land successfully. And find factors that are important for first stage landing. This will be useful for an alternative company that wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

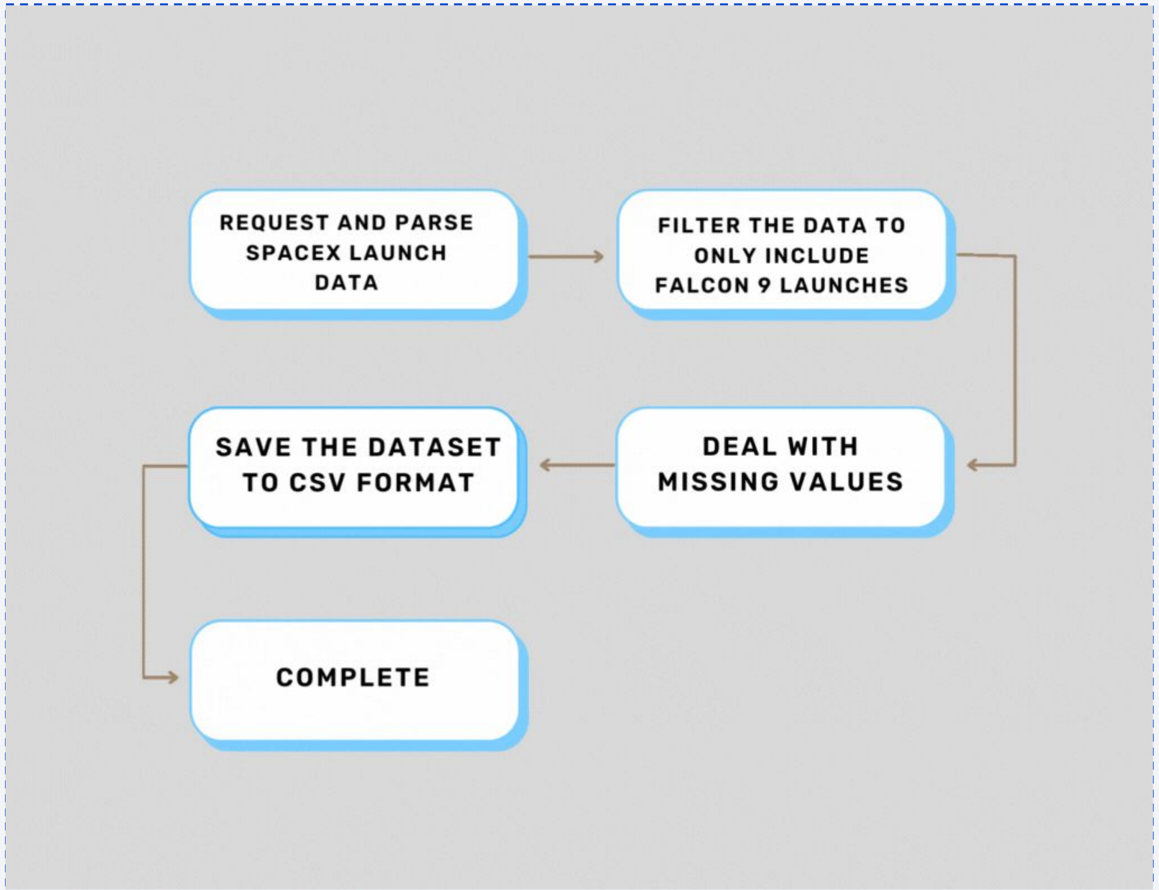
Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and using web scraping
- Perform data wrangling
  - Removing undefined values in the dataset, performing data scaling, and One Hot encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using different classification models, optimizing the hyper-parameters, and comparing the accuracy the models used.

Data Collection – SpaceX API

- Used Requests package to extract data from the SpaceX API

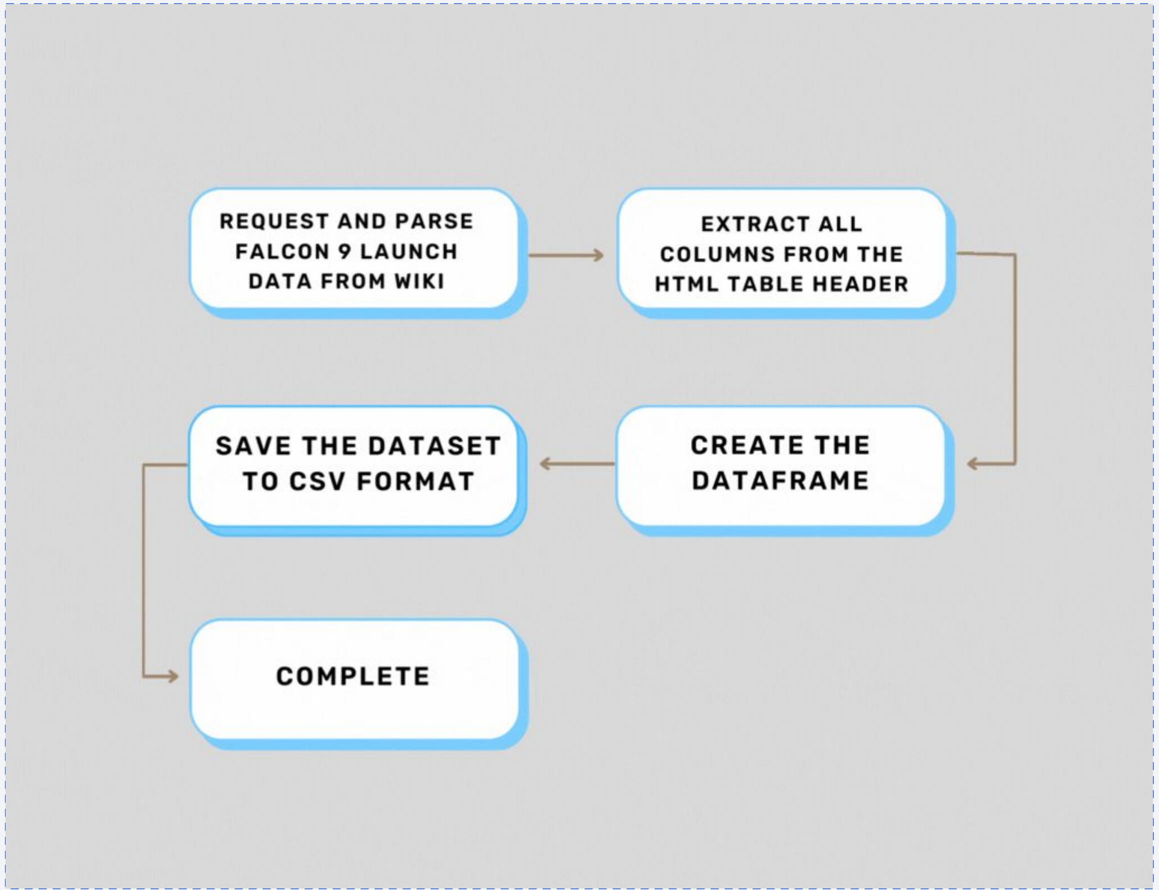
[Data Collection API Notebook]


REQUEST AND PARSE SPACEX LAUNCH DATA
FILTER THE DATA TO ONLY INCLUDE FALCON 9 LAUNCHES
DEAL WITH MISSING VALUES
SAVE THE DATASET TO CSV FORMAT
COMPLETE

Data Collection - Scraping

- Used Beautifulsoup to extract data from a wikipedia web page

[Data Collection with Webscrapping Notebook]


REQUEST AND PARSE FALCON 9 LAUNCH DATA FROM WIKI
EXTRACT ALL COLUMNS FROM THE HTML TABLE HEADER
CREATE THE DATAFRAME
SAVE THE DATASET TO CSV FORMAT
COMPLETE

Data Wrangling

- Checking the type/class of data, and changing the data types as required.
- Counting the Null values present in different columns in the data set.
- Removing the Nulls values from the dataset.
- Renaming different columns in the dataset with the correct column names.
- The link to the complete notebook is below:

  Data Wrangling Notebook

EDA with Data Visualization

- Visualized the relationship between:
    - Flight Number and Launch Site
    - Flight Number and Payload Mass
    - Payload Mass and Launch Site
    - success rate of each orbit type
    - Flight Number and Orbit type
    - Payload Mass and Orbit type
- Visualized launch success yearly trend.

EDA with Data Visualization Notebook

EDA with SQL

- Evaluated names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begin with the string 'CCA’.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1.
- Displayed the date when the first successful landing outcome in the ground pad was achieved.
- Displayed the names of the boosters which have success in drone ships and have payload mass greater than 4000 but less than 6000.
- Displayed the total number of successful and failed mission outcomes.
- Displayed the names of the booster versions which have carried the maximum payload mass.
- Displayed the failed landing outcomes in drone ships, their booster versions, and launch site names for the year 2015.
- Ranked the count of landing outcomes

EDA with SQL Notebook

Build an Interactive Map with Folium

- Marked all the launch sites on an interactive map.
- Added map objects such as markers, circles, and lines to mark the success or failure of launches for each site.
- Assigned the launch outcomes (failure or success) to class 0 and 1, i.e., 0 for failure and 1 for success.
- Color-labeled marker clusters to identify which launch sites have relatively high success rate.
- Show which site is nearest to highway road, rail, city, and a coastline using line

SpaceX Map Location Notebook

Build a Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly dash.
- Plotted pie charts showing the total launches for all sites and specific sites.
- Generated a scatter plot depicting the relationship between Outcome and Payload Mass (Kg) for the different booster versions.

SpaceX Dash App

Predictive Analysis (Classification)

- Used Linear Regression, SVM, KNN, Decision Tree models to predict success or failure of a mission.
- Data scaling using StandardScalar() from sklearn library in Python.
- Data splitting into testing (20%) and training (80%) datasets.
- GridSearchCV with 10-fold cross-validation for hyperparameter tuning to achieve the best model accuracy.
- Listed the model providing the best accuracy.

SpaceX ML Prediction Notebook

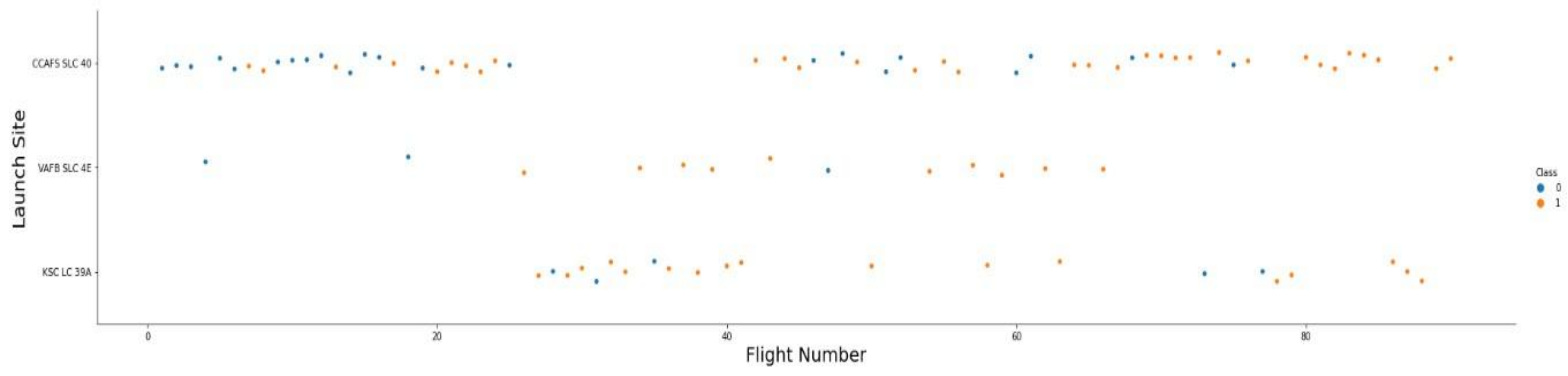
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

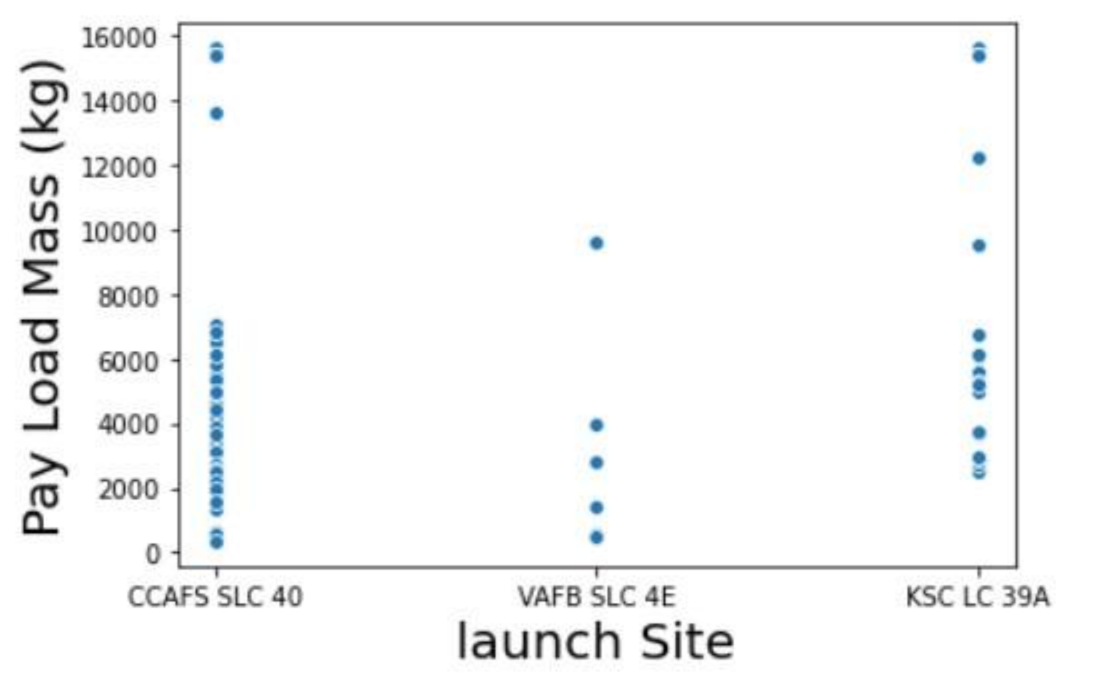
Insights drawn from EDA

Flight Number vs. Launch Site



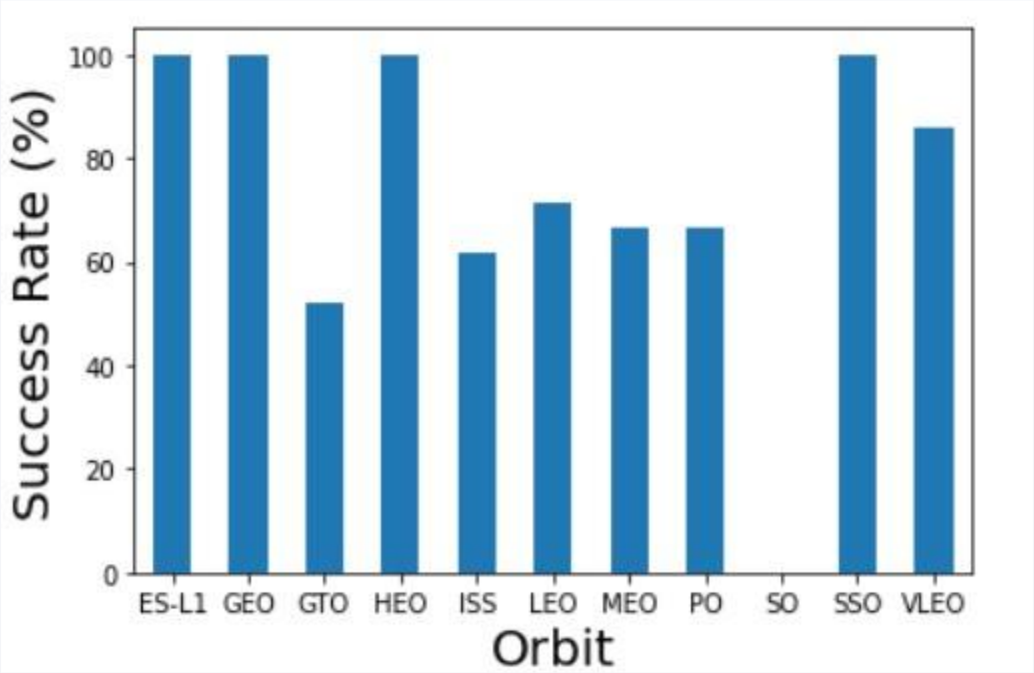
Launch Site CCAFS SLC 40 has experienced more failed launches than any other launch sites, but got more success launches as number of flights increases like all the other sites.

Payload vs. Launch Site



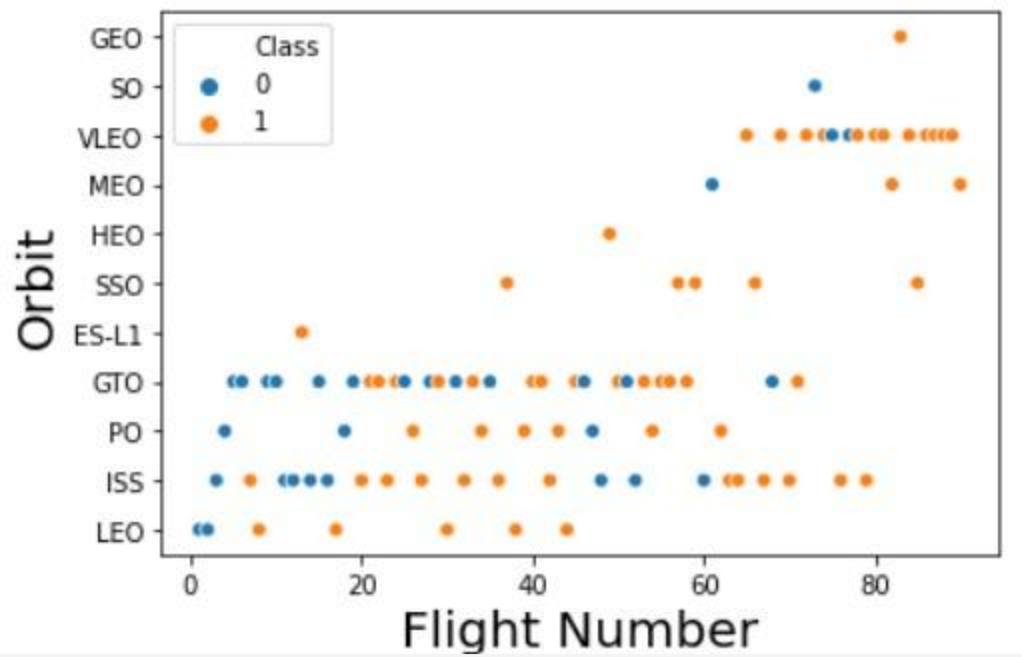
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



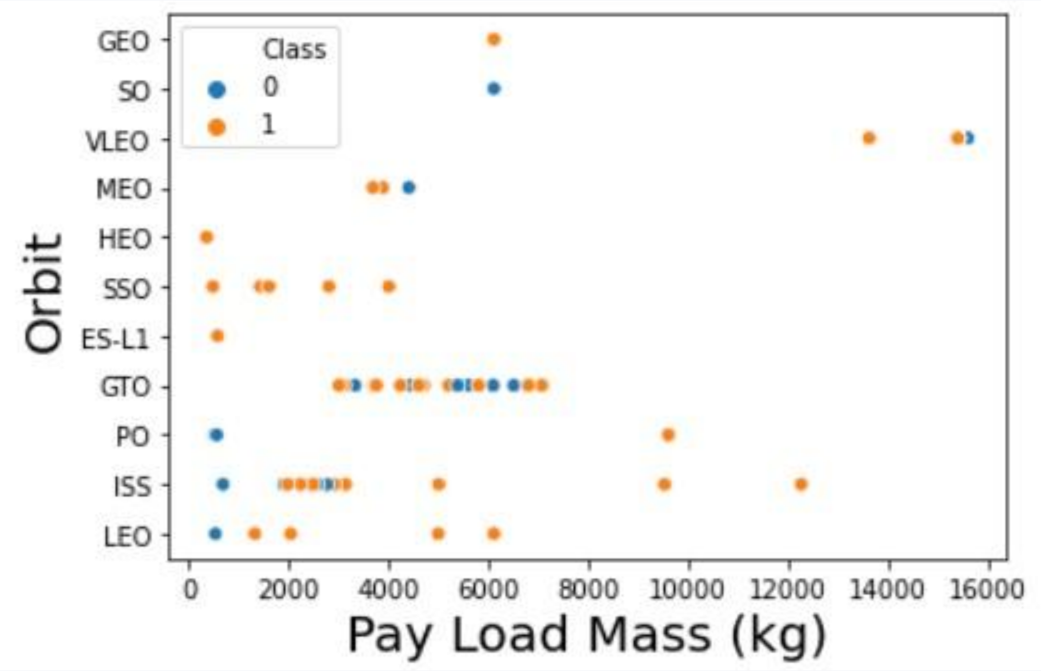
Observe that there are 4 Orbits that achieved 100% Success Rate namely: ES-LI, GEO, HEO, and SSO. The Orbit with the lowest Success Rate is GTO 50-60%

Flight Number vs. Orbit Type



You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

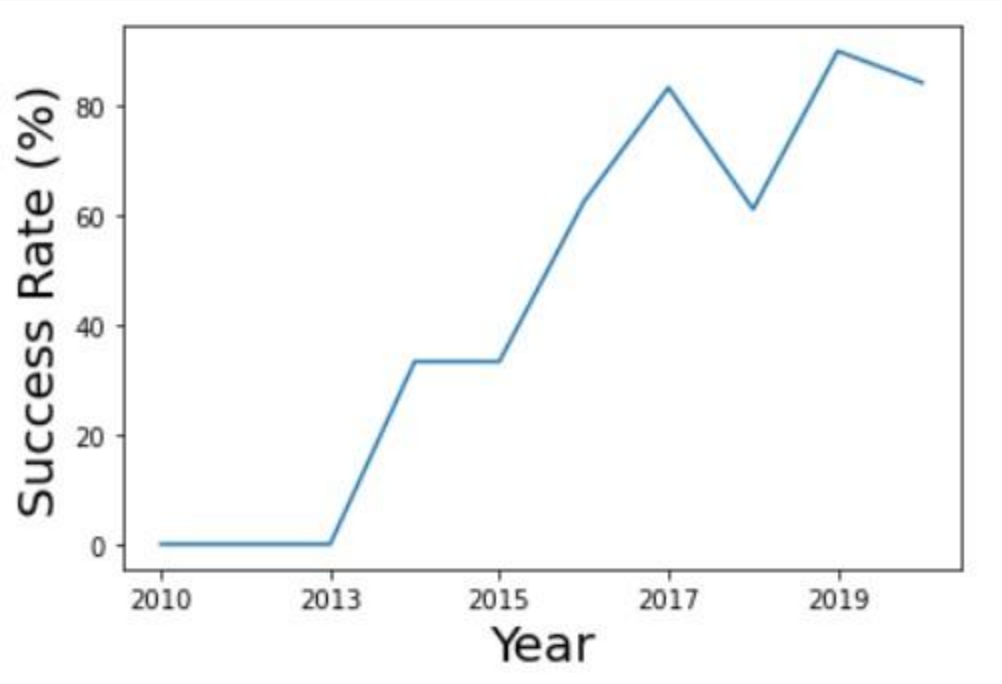
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

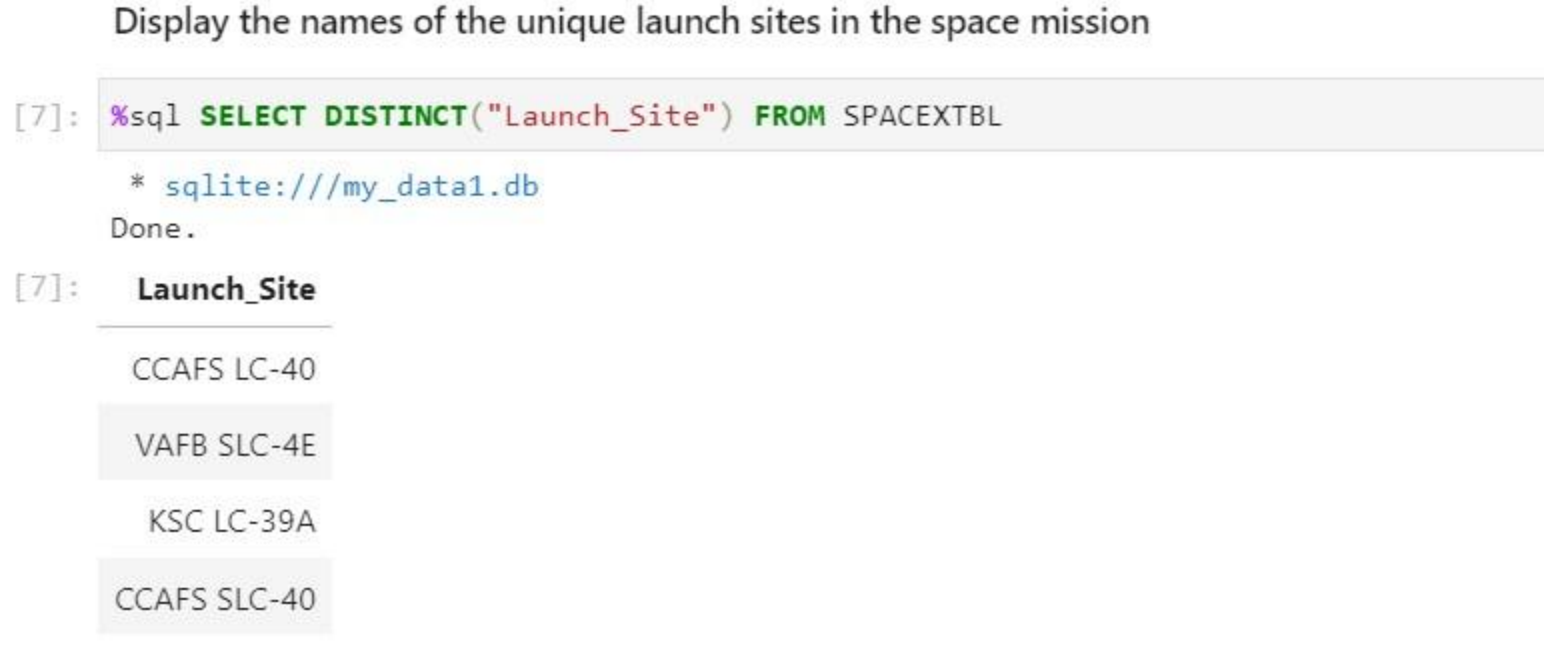
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names



Display the names of the unique launch sites in the space mission

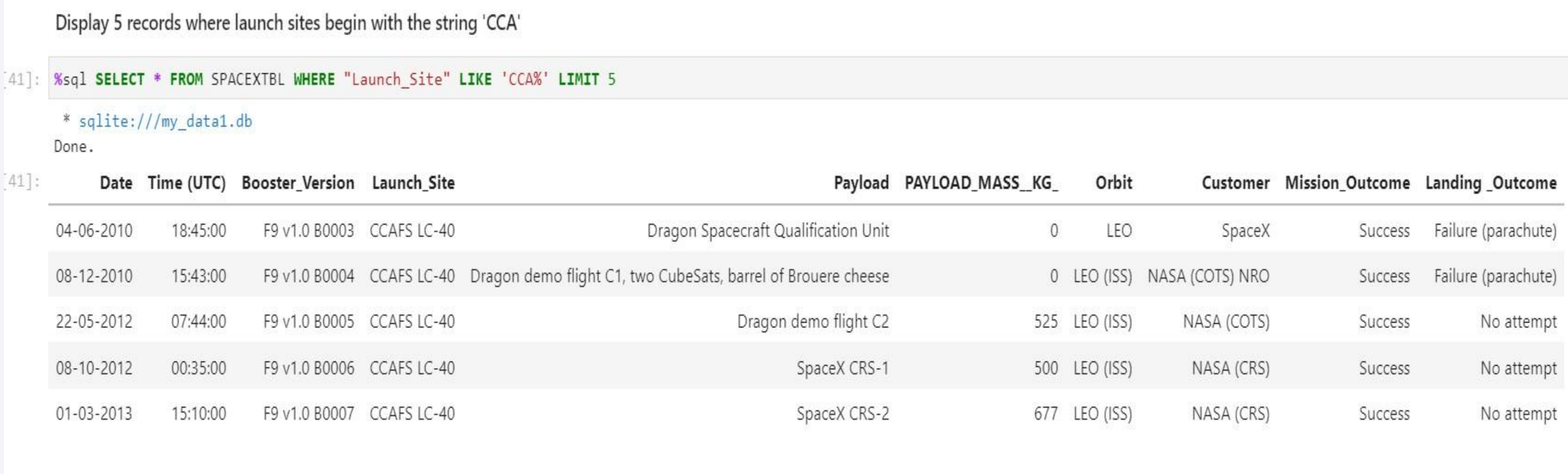
```
[7]: %sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL
```

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

[7]:

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

Launch Site Names Begin with 'CCA'



Display 5 records where launch sites begin with the string 'CCA'

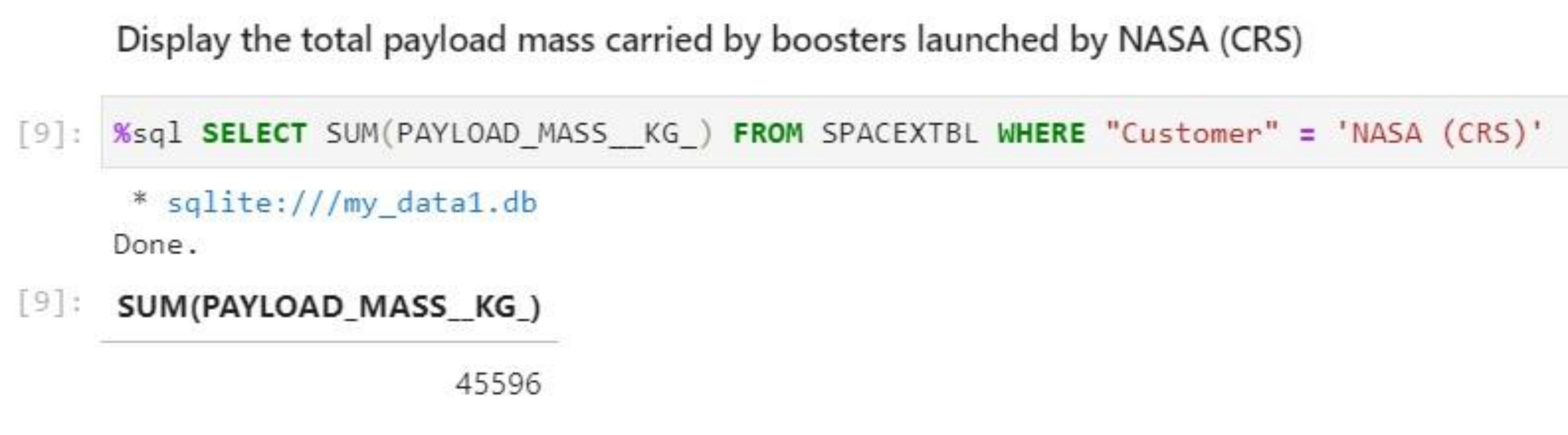
```
[41]: %sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

[41]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass



Display the total payload mass carried by boosters launched by NASA (CRS)

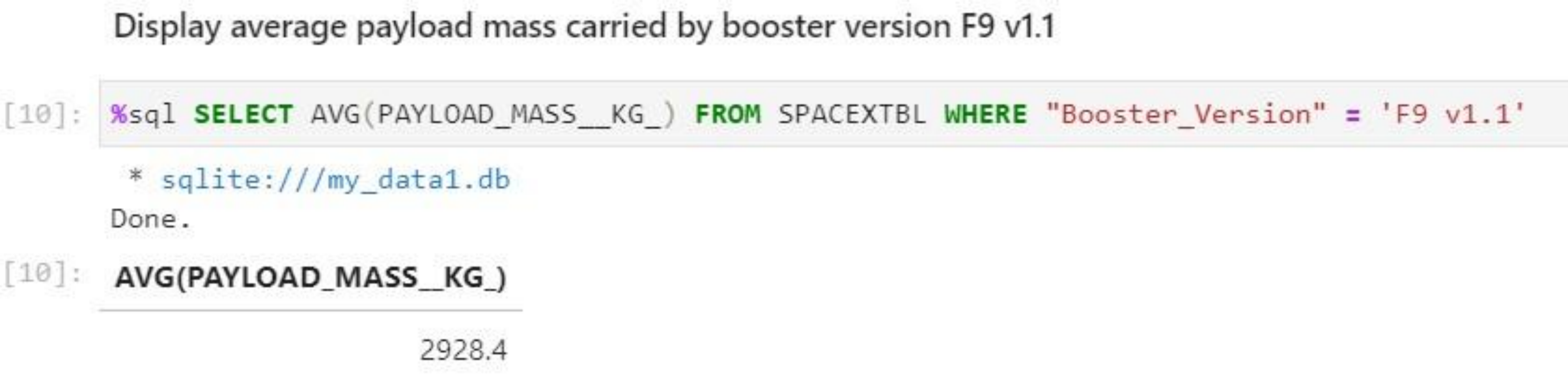
```
[9]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Customer" = 'NASA (CRS)'
```

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

[9]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1



Display average payload mass carried by booster version F9 v1.1

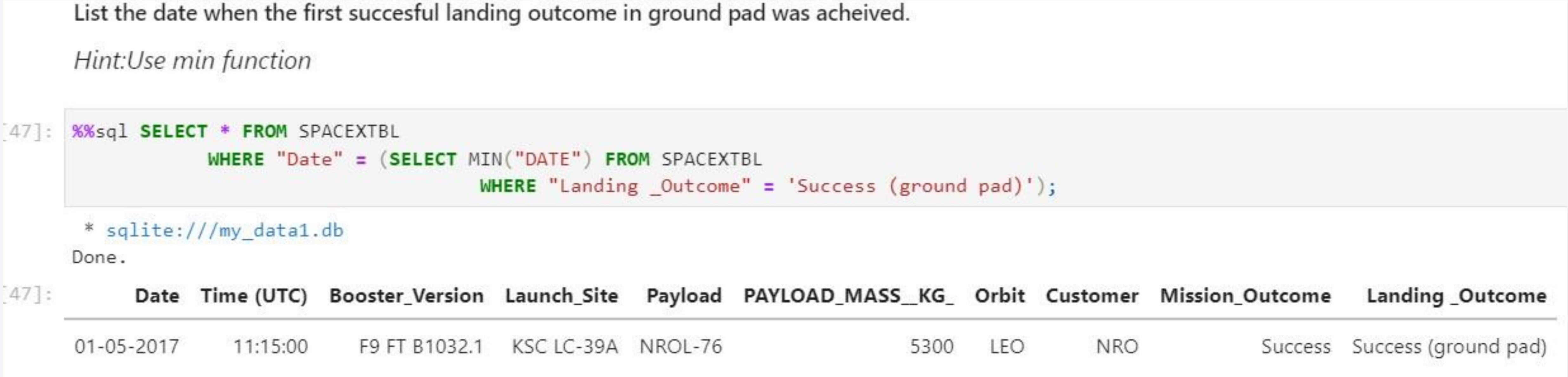
```
[10]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version" = 'F9 v1.1'
```

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

[10]:

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date



List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

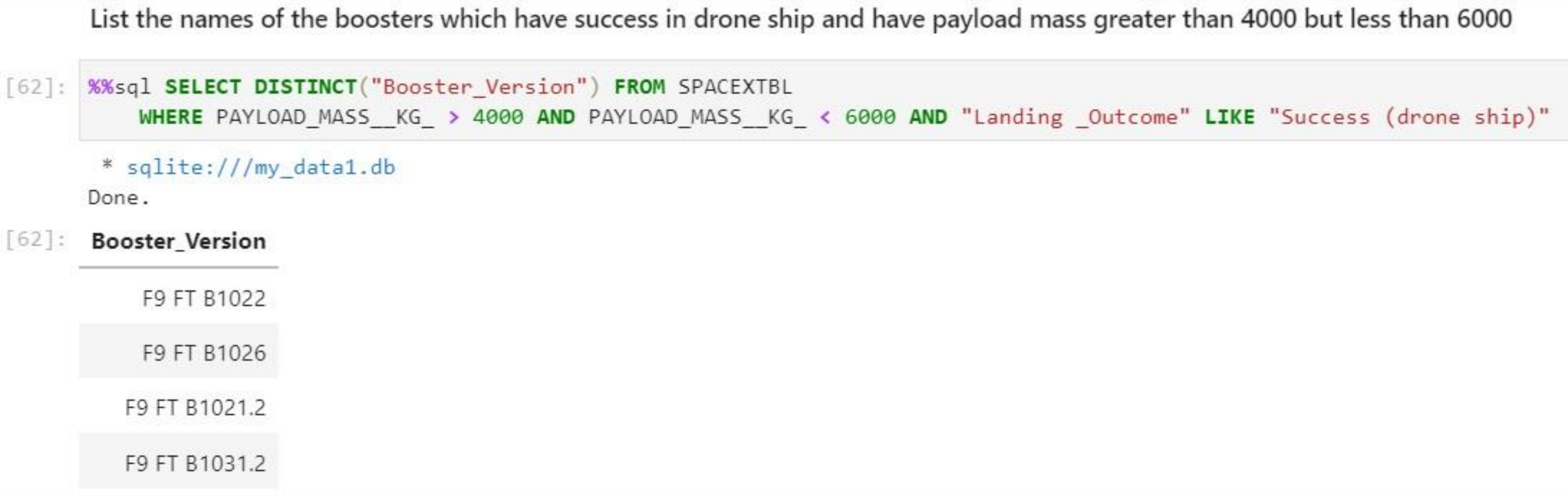
```
[47]: %%sql SELECT * FROM SPACEXTBL
              WHERE "Date" = (SELECT MIN("DATE") FROM SPACEXTBL
                                          WHERE "Landing _Outcome" = 'Success (ground pad)');
```

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

[47]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

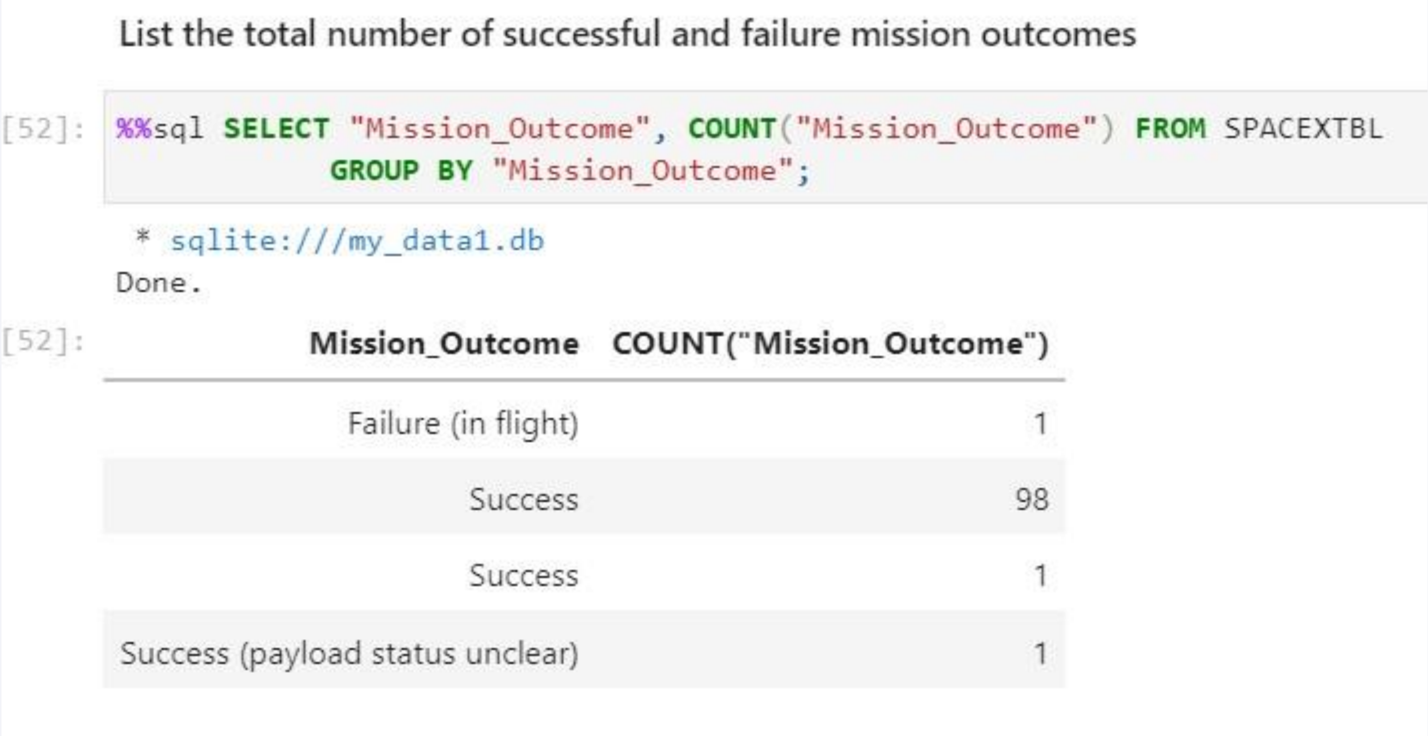
```
[62]: %%sql SELECT DISTINCT("Booster_Version") FROM SPACEXTBL
          WHERE PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000 AND "Landing _Outcome" LIKE "Success (drone ship)"
```

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

[62]:

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

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

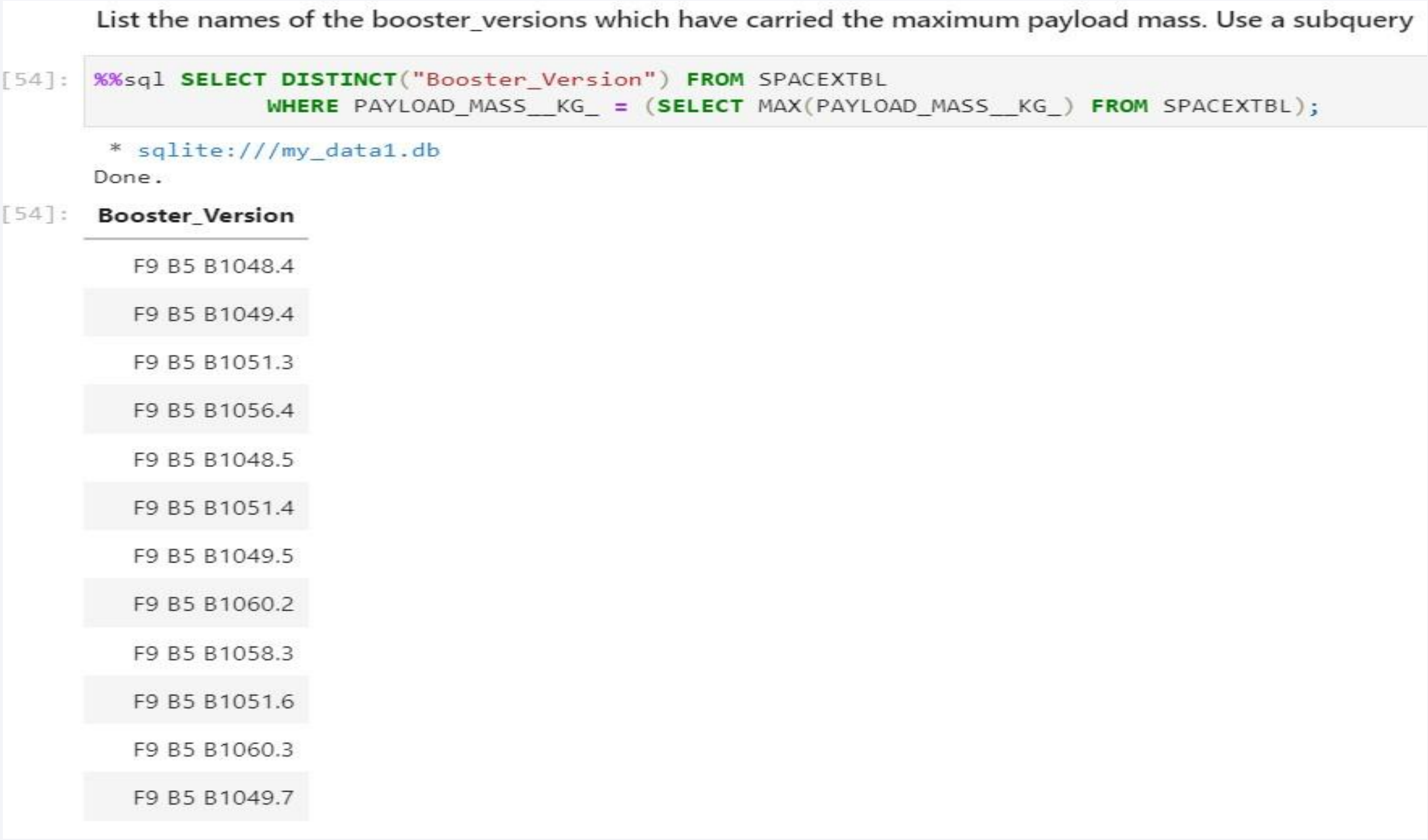
```
[52]: %%sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTBL
                 GROUP BY "Mission_Outcome";
```

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

[52]:

Mission_Outcome	COUNT("Mission_Outcome")
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

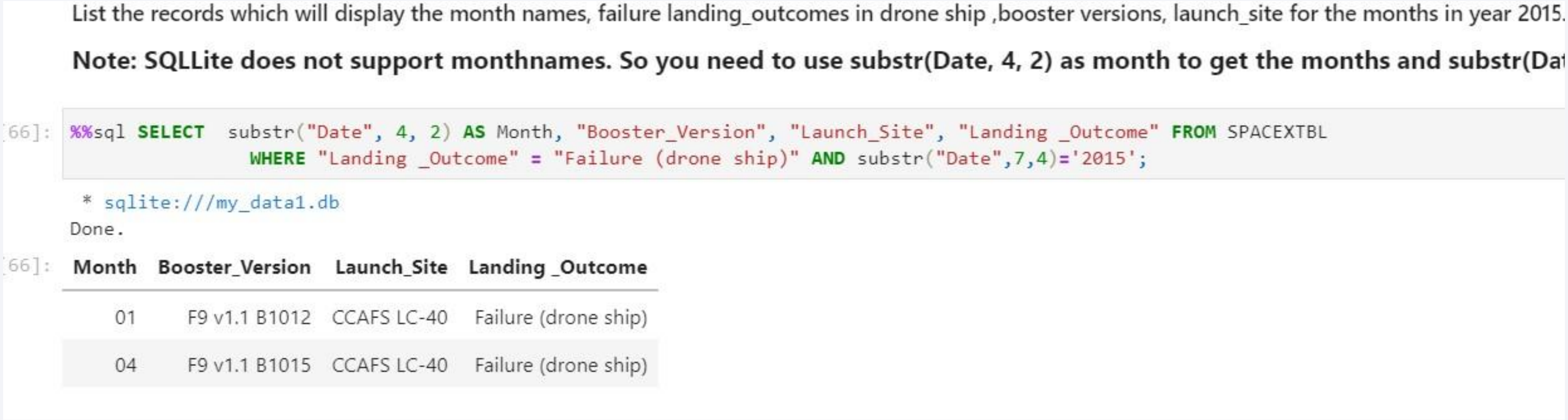
[54]:
```
%%sql SELECT DISTINCT("Booster_Version") FROM SPACEXTBL
            WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

[54]:

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



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.

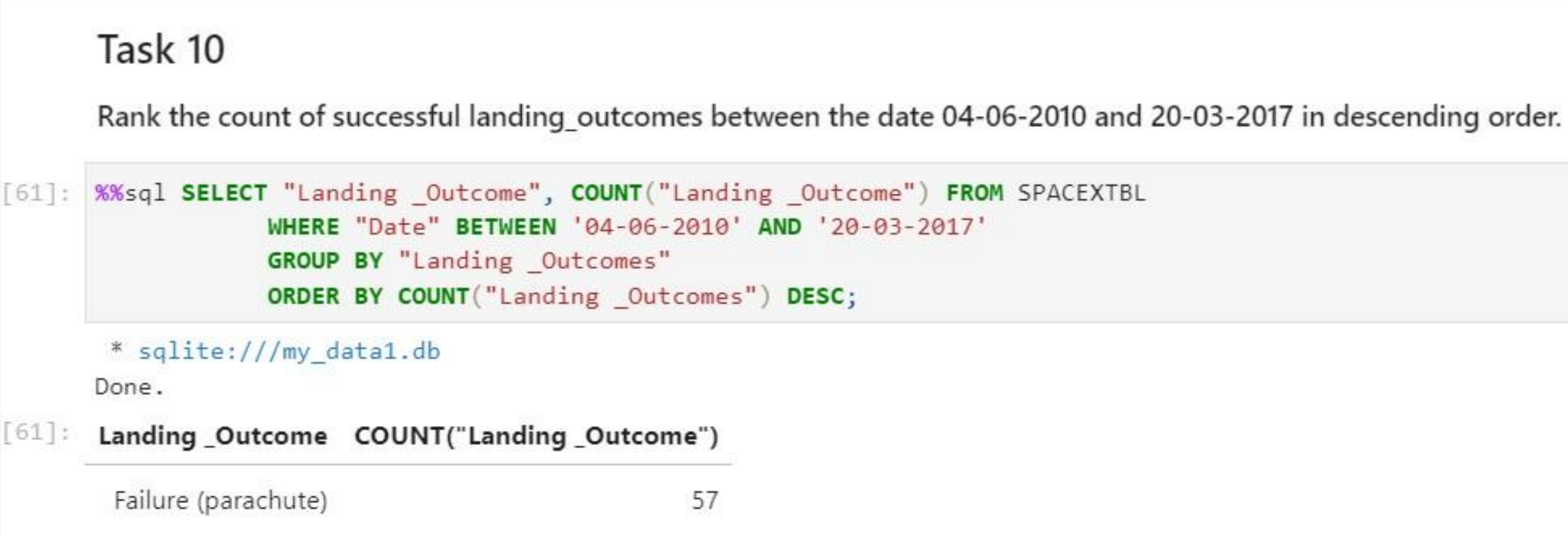
Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Dat

```
%%sql SELECT  substr("Date", 4, 2) AS Month, "Booster_Version", "Launch_Site", "Landing _Outcome" FROM SPACEXTBL
          WHERE "Landing _Outcome" = "Failure (drone ship)" AND substr("Date",7,4)='2015';
```

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

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)

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



Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
[61]: %%sql SELECT "Landing _Outcome", COUNT("Landing _Outcome") FROM SPACEXTBL
              WHERE "Date" BETWEEN '04-06-2010' AND '20-03-2017'
              GROUP BY "Landing _Outcomes"
              ORDER BY COUNT("Landing _Outcomes") DESC;
```

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

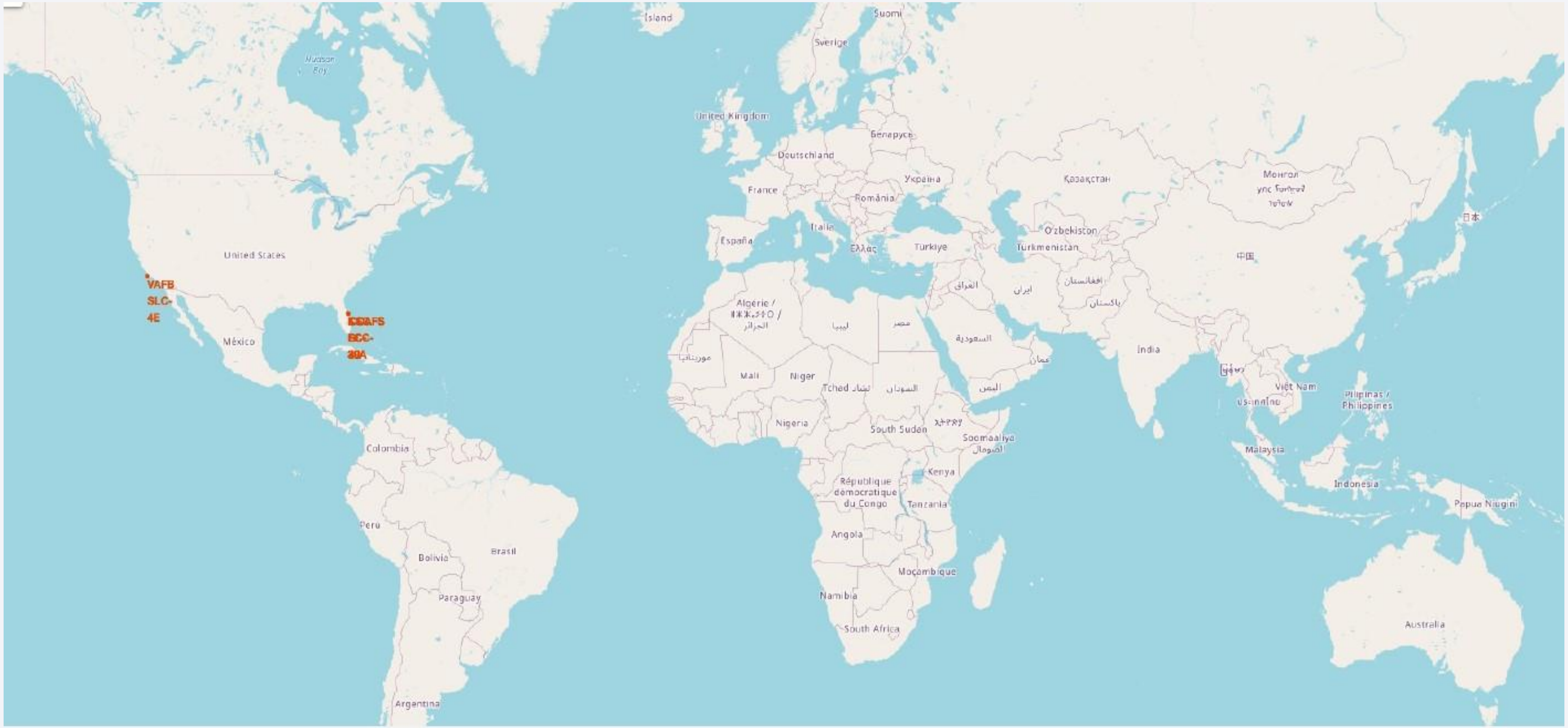
[61]:

Landing _Outcome	COUNT("Landing _Outcome")
Failure (parachute)	57

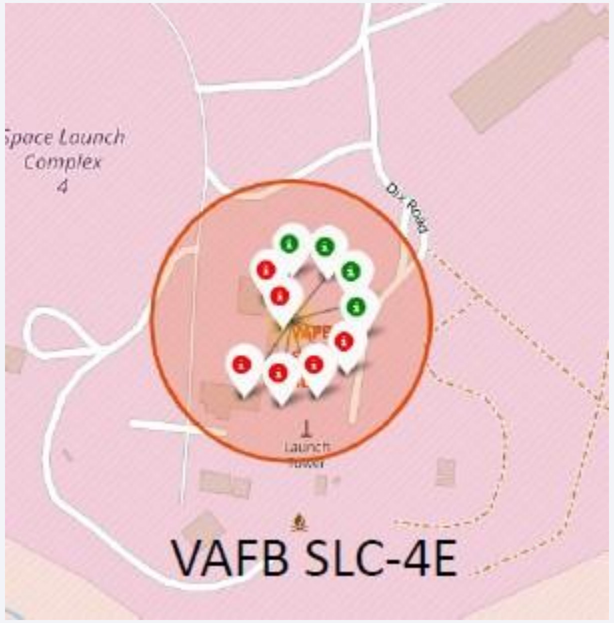
Section 3

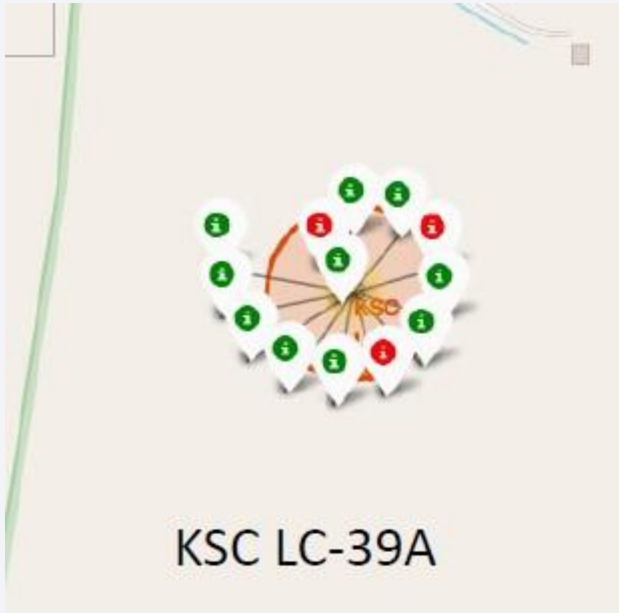
Launch Sites Proximities Analysis

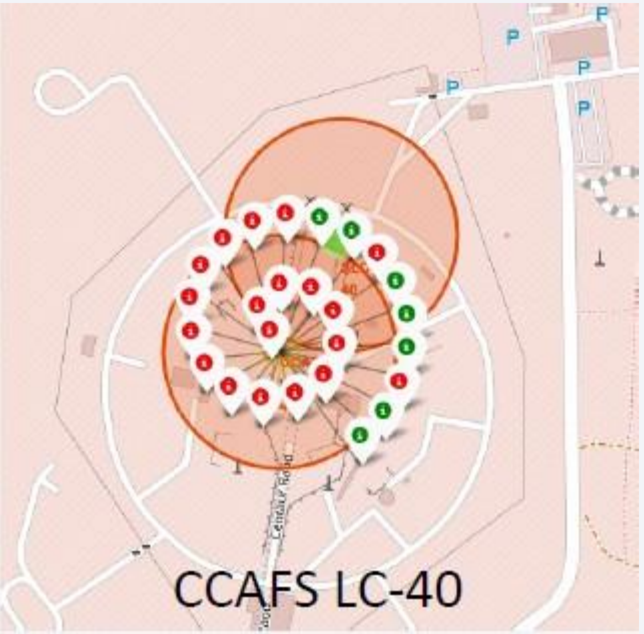
Launch Site Locations

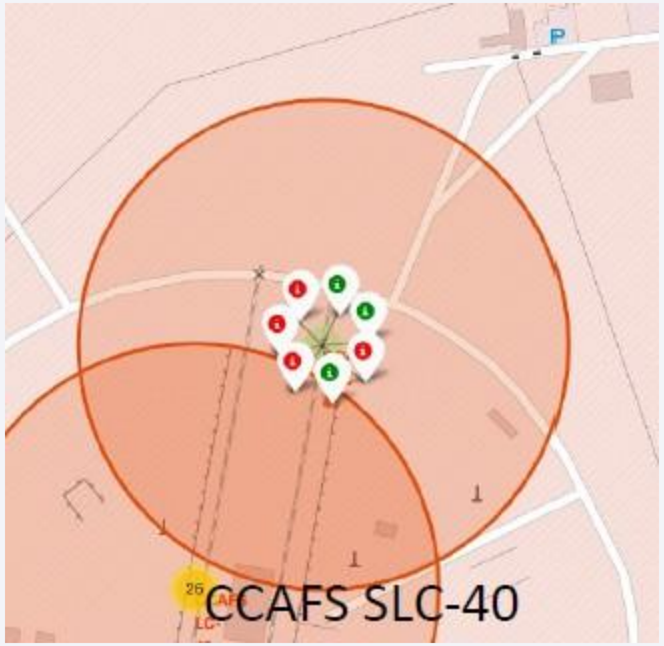


Launch Sites with Coloured Markers



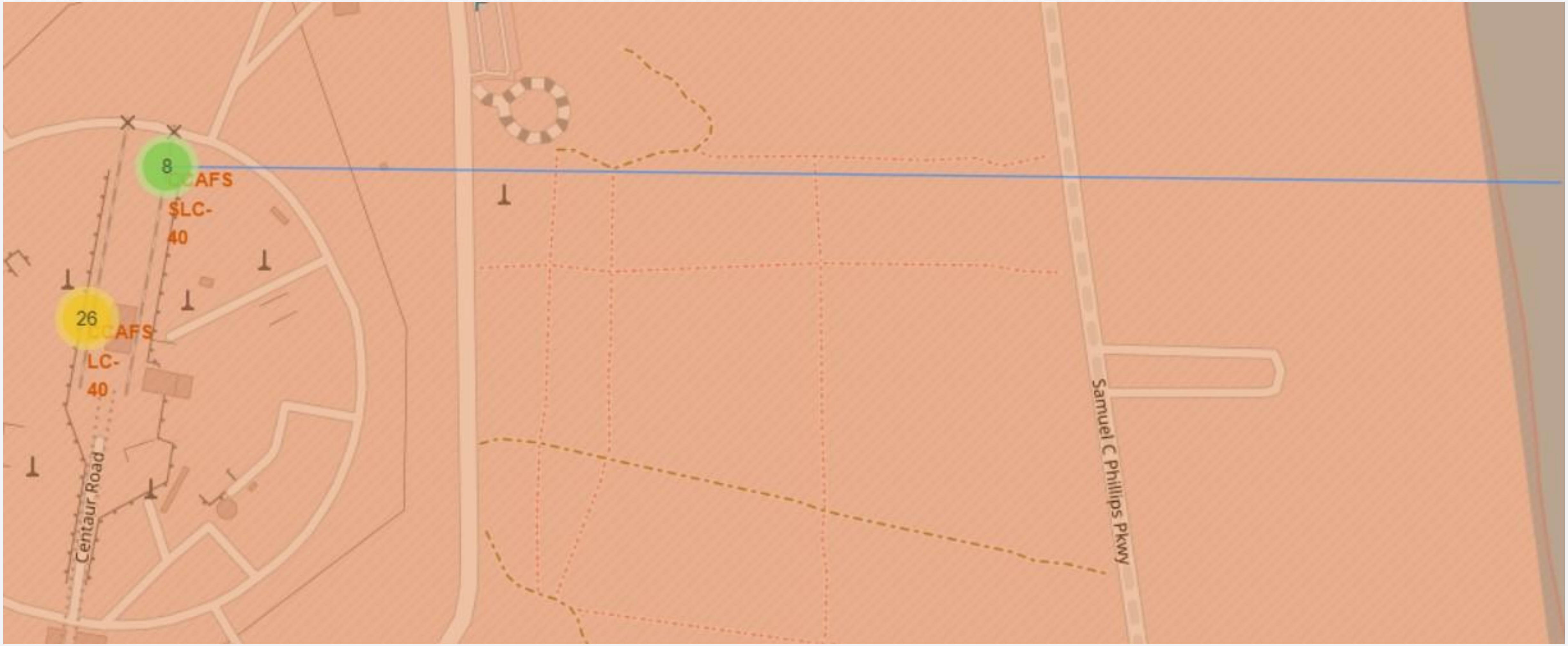




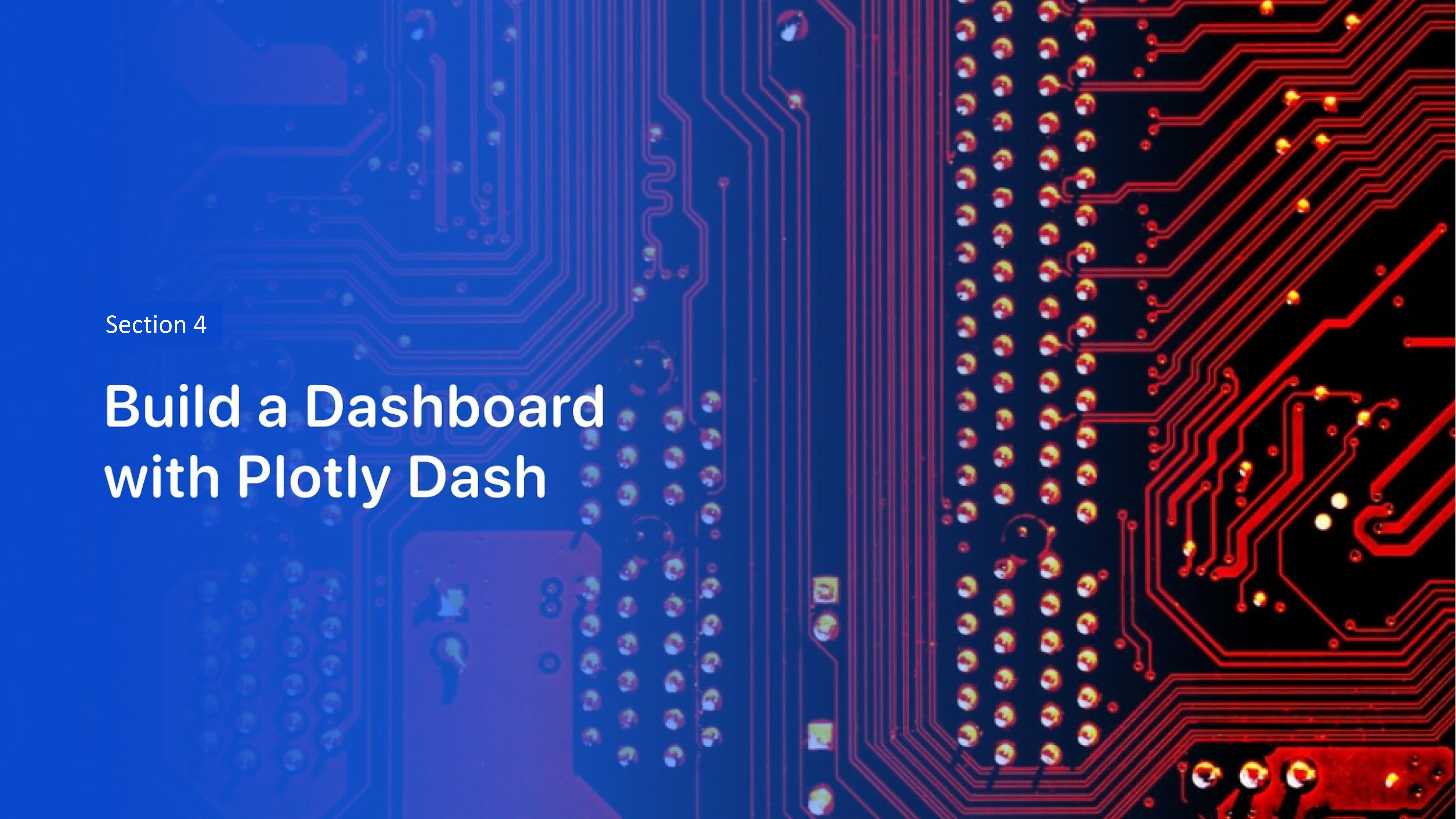


Green marker indicates success, while a red marker indicates a failure

A Launch Site Closest to a Coastline



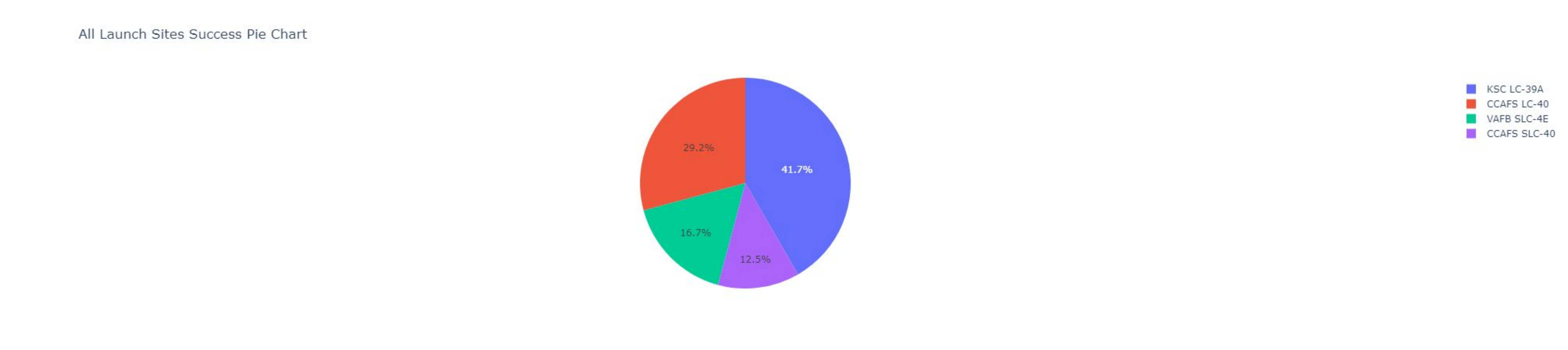
i.e. CCAFS SLC 40



Section 4

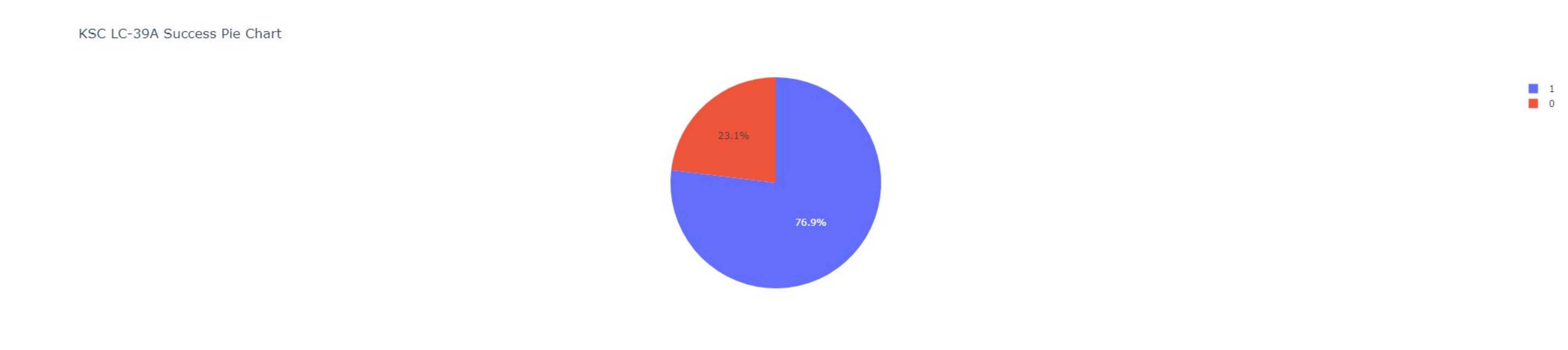
Build a Dashboard with Plotly Dash

Launch Site Success for All Sites



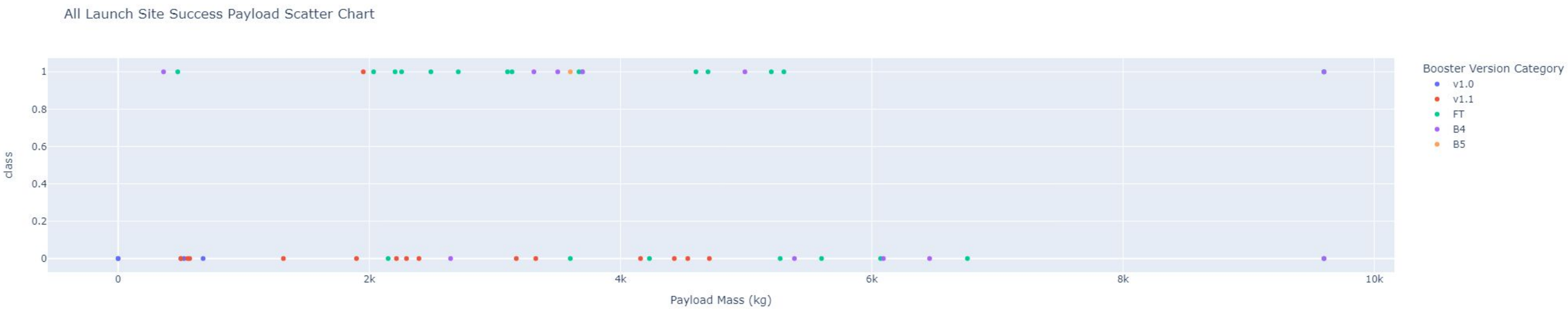
The launch site KSC LC-39A has the highest success rate among all the orbits.

KSC LC-39A Launch Site Success Rate



KSC LC-39A has the success rate of 76.9 %.

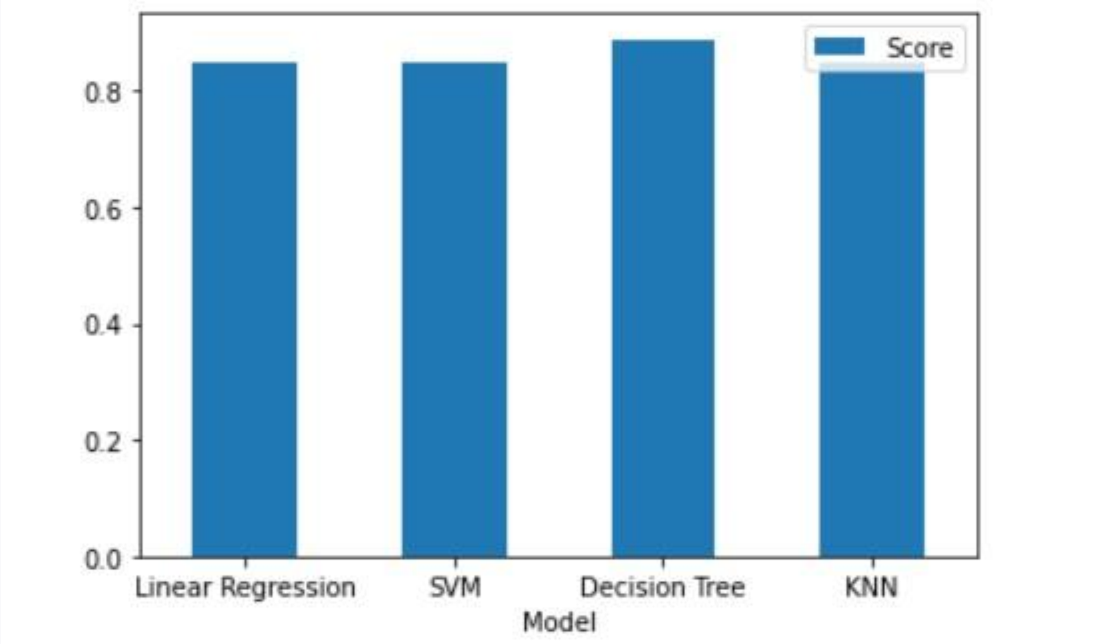
Correlation Between Payload and Success For All Sites



Section 5

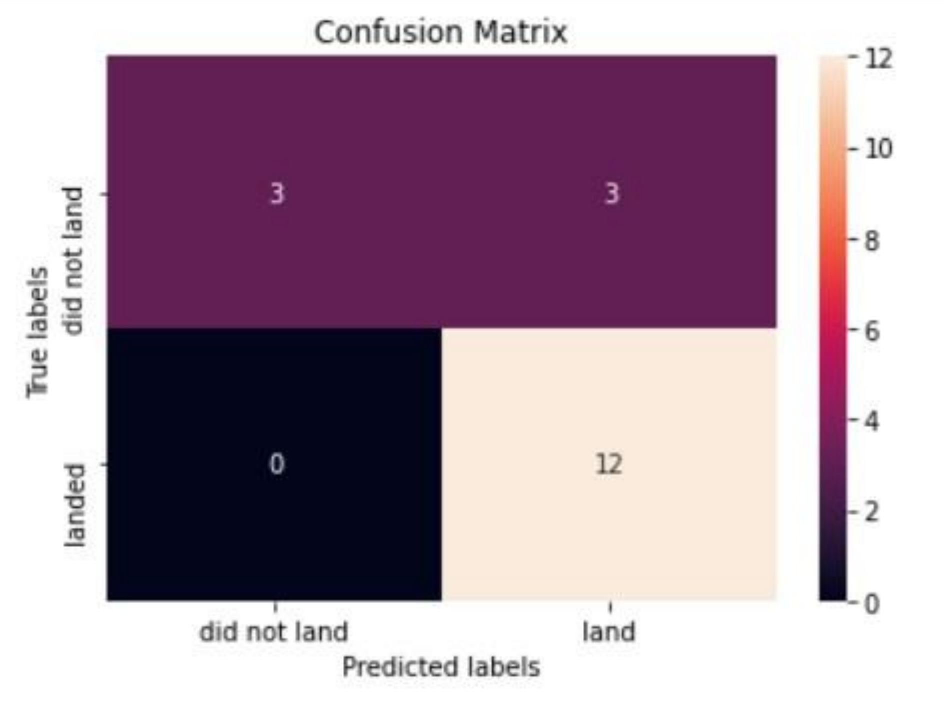
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree Model has the best accuracy of 0.889

Confusion Matrix



The decision tree model can distinguish between the different classes with high accuracy. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the model.

Conclusions

- Through EDA, we have studied several features that influence the success rate of the launch.
- Through predictive modeling using different Machine Learning models, we were able to predict the launch outcome with high accuracy.
- The Decision Tree Model has the best accuracy compared to other models.
- To conclude, we performed predictive analysis to predict the success rate of the SpaceX launches.

Thank you!