IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anderson Borges
16/11/2024

Outline

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

Executive Summary

The research aims to identify the factors contributing to a successful rocket landing. To make this determination, the following methodologies were employed:

- Data collection using the SpaceX REST API and web scraping techniques;
- Data processing to create a success/failure outcome variable;
- Data exploration with visualization techniques, considering factors such as payload, launch site, flight number, and yearly trend;
- Data analysis with SQL, calculating statistics such as total payload, payload range for successful launches, and the total number of successful and failed outcomes;
- Investigation of the success rates of launch sites and their proximity to geographical markers;
- Visualization of the most successful launch sites and successful payload ranges;
- Building models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and k-nearest neighbor (KNN).

Introduction

SpaceX, one of the leading companies in the space sector, aims to make travel to space accessible to everyone.

Among its achievements are the sending of spacecraft to the International Space Station, the launch of a constellation of satellites that offer internet access and the carrying out of manned missions. SpaceX is able to do this because its rocket launches are relatively low cost ($62 million per launch) thanks to the innovative reuse of the first stage of its Falcon 9 rocket. On the other hand, other vendors who are unable to reuse the first stage have costs exceeding US$165 million per launch.

To determine the launch price, it is essential to know whether the first stage will be recovered. Using public data and machine learning models, we can predict whether SpaceX or a competing company will be able to reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

- Request data from the SpaceX API
- Decode the response using .json() and convert it to a dataframe using .json_normalize()
- Fetch information about the launches from the SpaceX API using custom functions
- Create a dictionary from the data
- Generate a dataframe from the dictionary
- Filter the dataframe to include only Falcon 9 launches
- Replace missing values of Payload Mass with the calculated mean using .mean()
- Export the data to a CSV file

Data Collection - Scraping

1. Obtain Falcon 9 launch data from Wikipedia

2. Generate a BeautifulSoup object from the HTML response

3. Retrieve column names from the table headers in the HTML

4. Gather data by parsing the HTML tables

5. Construct a dictionary from the collected data

6. Build a dataframe from the dictionary

7. Save the data to a CSV file

Data Wrangling

Data Collection

- Request rocket launch data from the SpaceX API.
- Decode the JSON response and convert it to a dataframe using json_normalize().

Data Fetching

- Utilize custom functions to fetch additional launch information from the SpaceX API.
- Create a dictionary to store the fetched data.

Data Preparation

- Convert the dictionary to a dataframe.
- Filter the dataframe to include only Falcon 9 launches.
- Handle missing values by replacing them with the calculated mean of the Payload Mass.

Data Export

- Export the cleaned dataframe to a CSV file for further analysis.

EDA with Data Visualization

- **Charts**
- Relationship between Flight Number and Payload
- Relationship between Flight Number and Launch Site
- Comparison of Payload Mass (kg) by Launch Site
- Comparison of Payload Mass (kg) by Orbit Type
- **Exploratory Data Analysis with Visualization**
- Utilize scatter plots to examine relationships. If a correlation exists, these variables may be valuable for machine learning.
- Use bar charts to compare discrete categories. Bar charts illustrate the relationships among categories and a measured value.

EDA with SQL

Display:

- Unique launch site names
- Five records where the launch site starts with 'CCA'
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by the booster version F9 v1.1

List:

- Date of the first successful landing on a ground pad
- Names of boosters that successfully landed on a drone ship and have a payload mass between 4,000 and 6,000
- Total number of successful and failed missions
- Names of booster versions that have carried the maximum payload
- Failed landing outcomes on a drone ship, including their booster versions and launch sites, for the months in the year 2015
- Count of landing outcomes between June 4, 2010, and March 20, 2017, in descending order

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- A blue circle was added at the NASA Johnson Space Center's coordinates, with a popup label displaying its name using its latitude and longitude.
- Red circles were added at all launch site coordinates, with popup labels showing their names using their latitude and longitude.

Map with Folium

- **Colored Markers of Launch Outcomes**
- Colored markers were added to indicate successful (green) and unsuccessful (red) launches at each launch site, highlighting the success rates of the launch sites.
- **Distances Between a Launch Site to Proximities**
- Colored lines were added to show the distance between the CCAFS SLC-40 launch site and its proximity to the nearest coastline, railway, highway, and city.

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Enable users to select either all launch sites or a specific launch site.

Dashboard with Plotly Dash

Slider for Payload Mass Range

- Allow users to choose a range for the payload mass.

Pie Chart Displaying Successful Launches

- Provide users with the ability to view successful and unsuccessful launches as a percentage of the total.

Scatter Chart Displaying Payload Mass vs. Success Rate by Booster Version

- Allow users to observe the relationship between payload mass and launch success.

Predictive Analysis (Classification)

- Generate a NumPy array from the 'Class' column.

- Standardize the data using StandardScaler. Fit and transform the data.

- Divide the data using train_test_split.

- Construct a GridSearchCV object with cv=10 for parameter tuning.

- Utilize GridSearchCV on various algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), and K-Nearest Neighbor (KNeighborsClassifier()).

- Compute accuracy on the test data using the .score() method for all models.

- Evaluate the confusion matrix for each model.

- Determine the best model based on Jaccard_Score, F1_Score, and Accuracy.

Results

Exploratory Data Analysis

- Launch success rates have increased over time.
- KSC LC-39A has the highest success rate among all landing sites.
- Orbits such as ES-L1, GEO, HEO, and SSO have a 100% success rate.

Visual Analytics

- The majority of launch sites are located near the equator and are all situated close to the coast.
- Launch sites are strategically placed to be distant enough from cities, highways, and railways to minimize the risk of damage from failed launches, while still being accessible for bringing in personnel and materials to support launch operations.

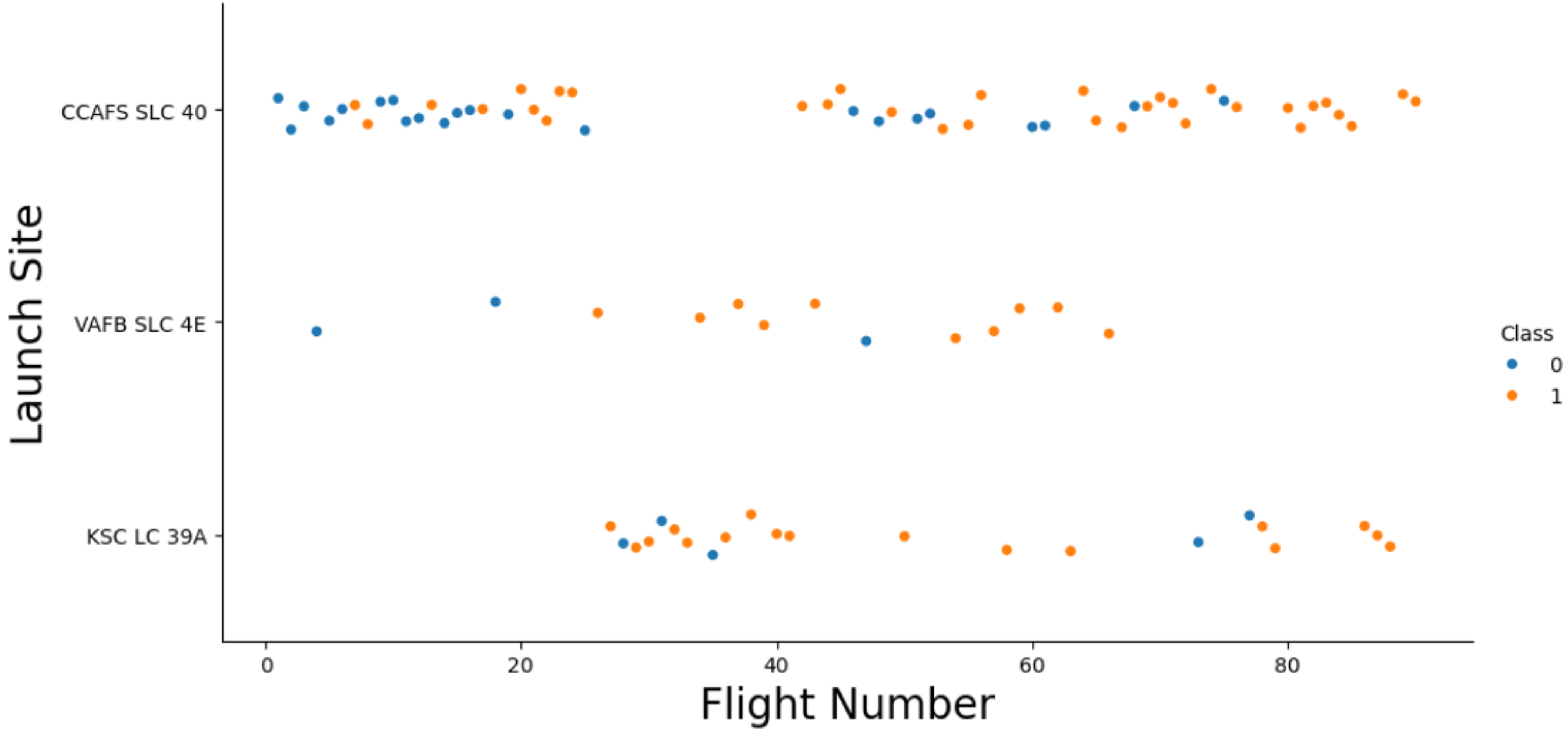
Predictive Analytics

- The Decision Tree model has been identified as the most effective predictive model for this dataset.

Section 2

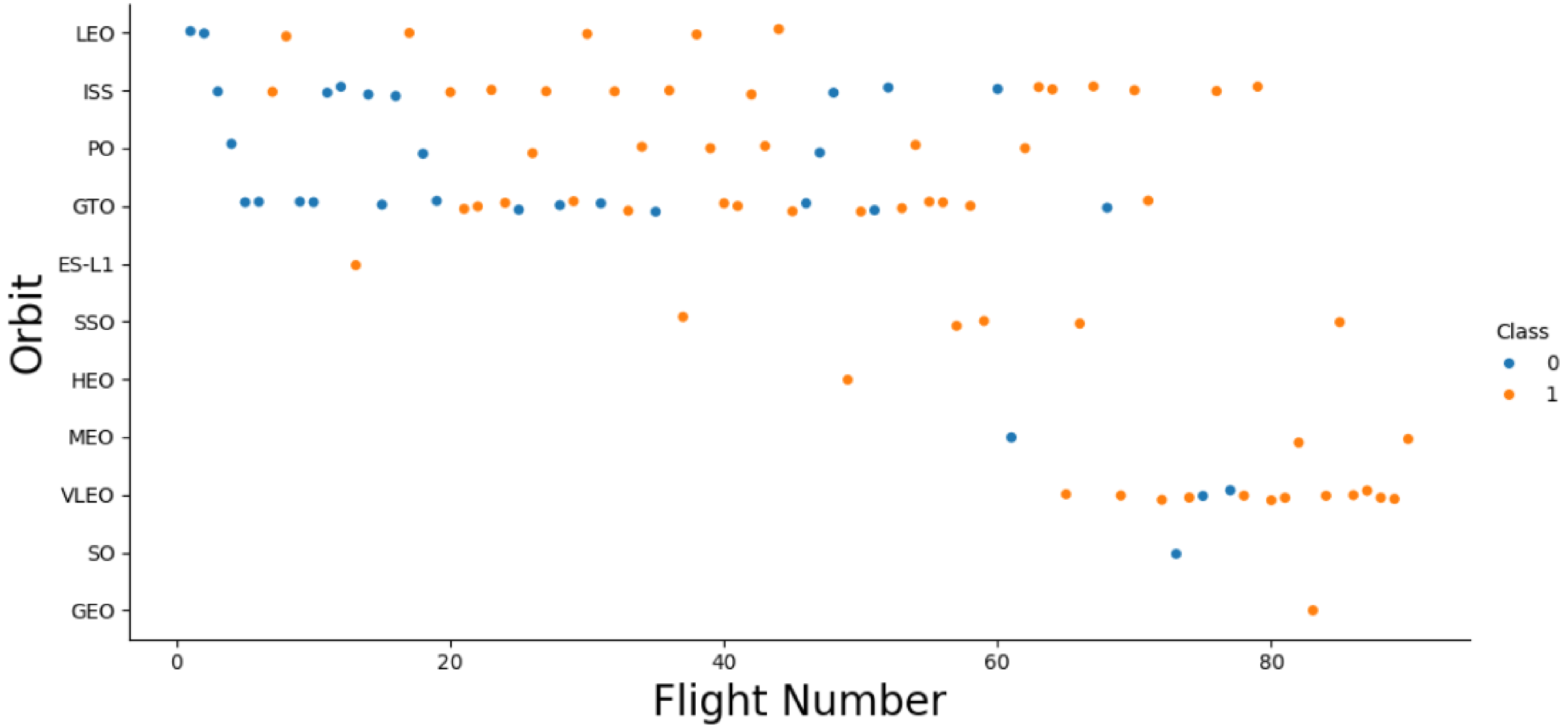
Insights drawn from EDA

Flight Number vs. Launch Site



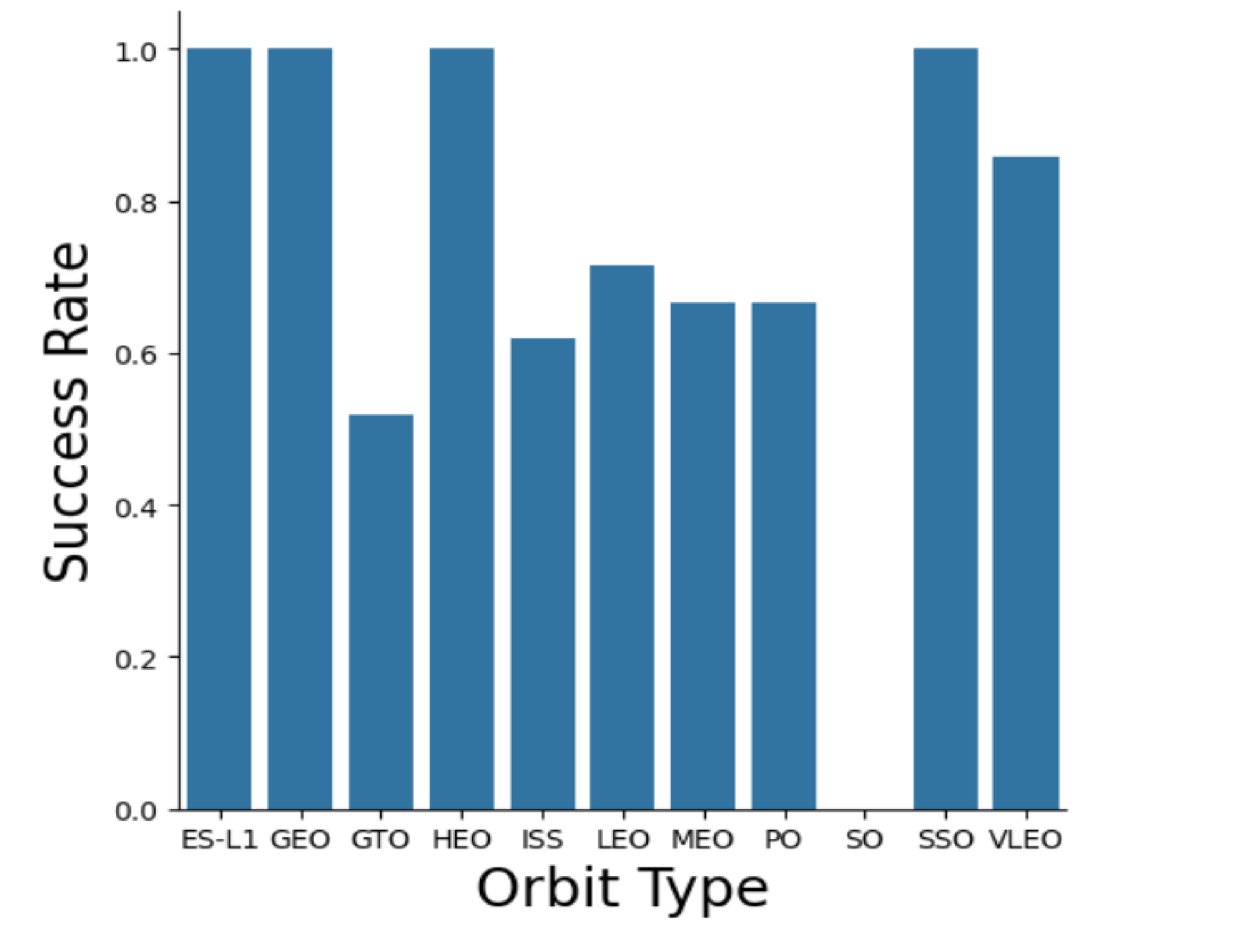
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Payload vs. Launch Site

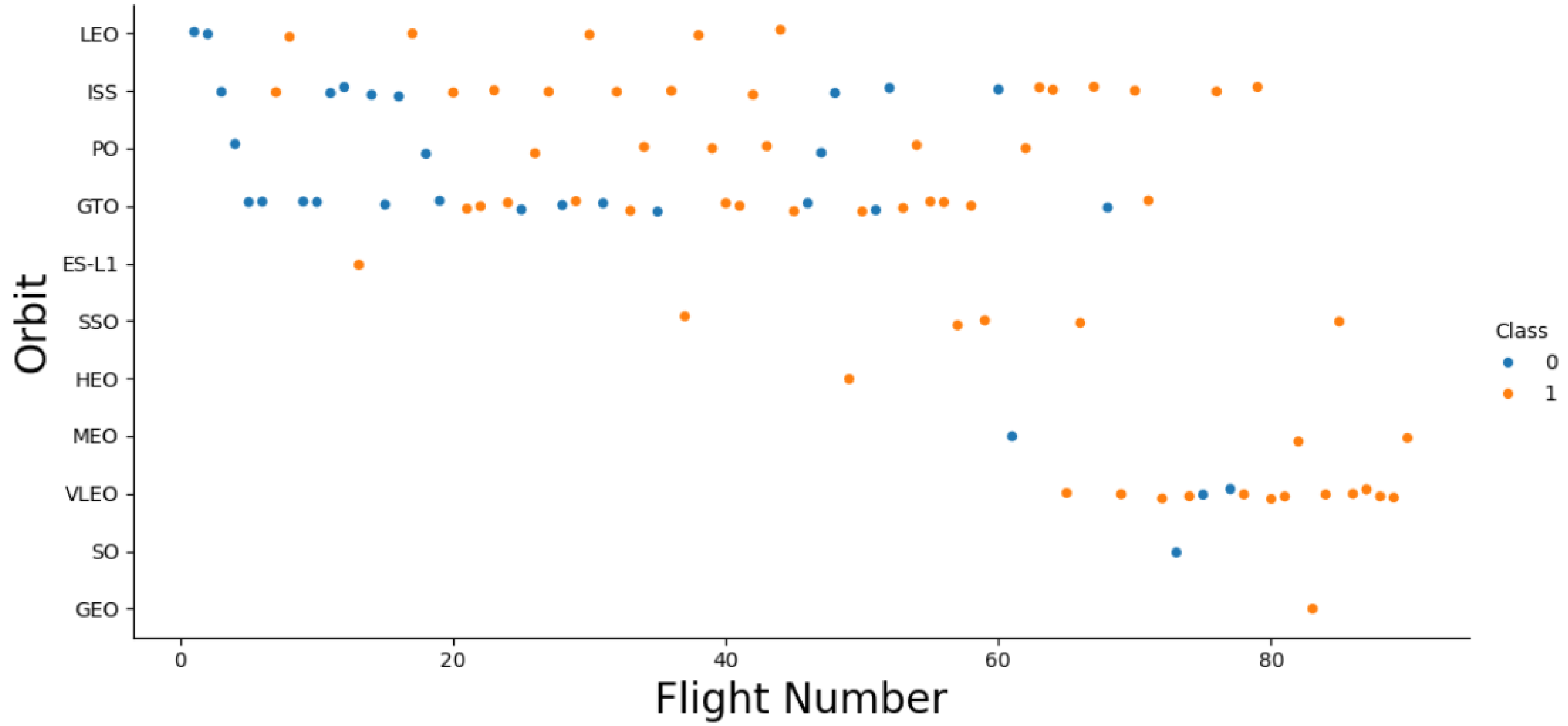


You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Success Rate vs. Orbit Type

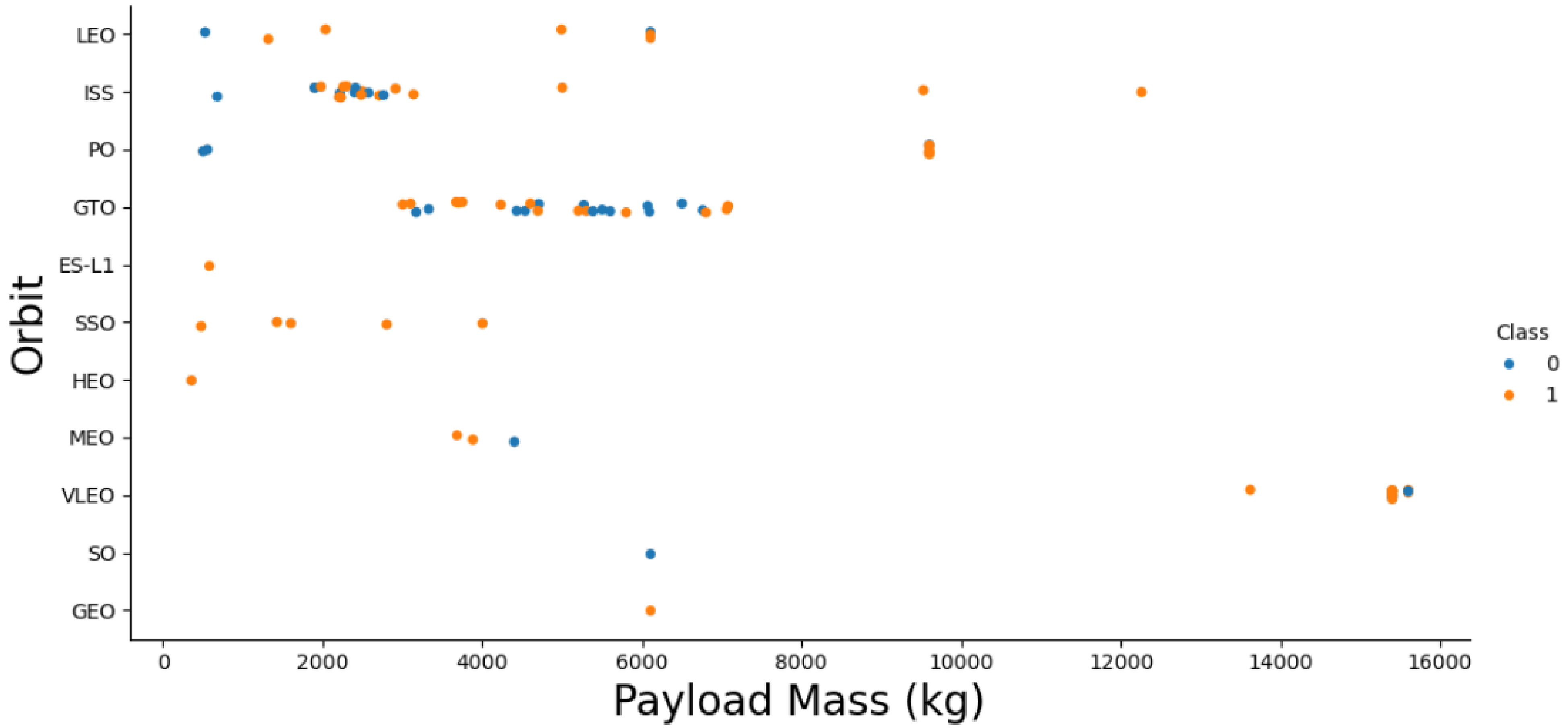


Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

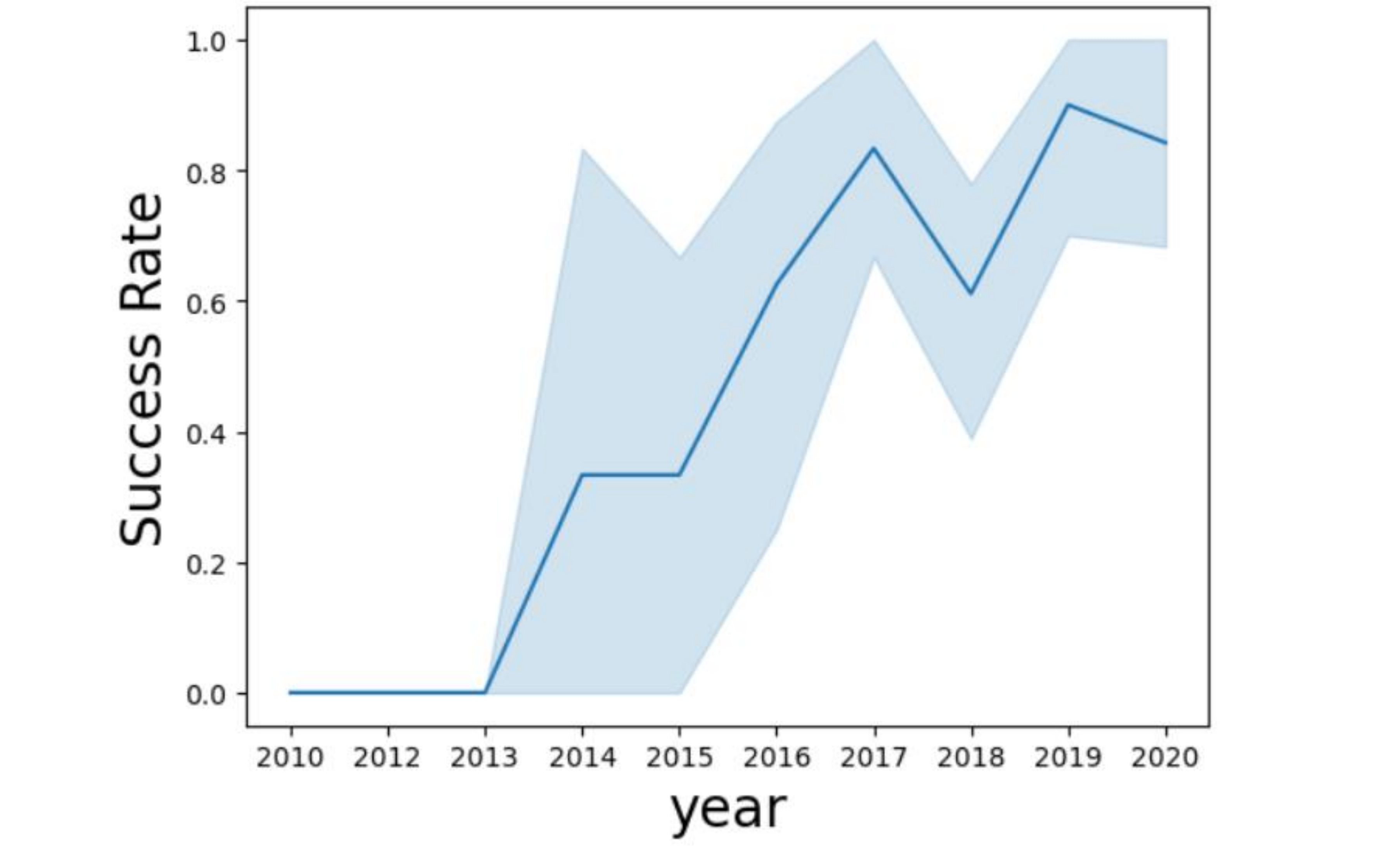
Payload vs. Orbit Type



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

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

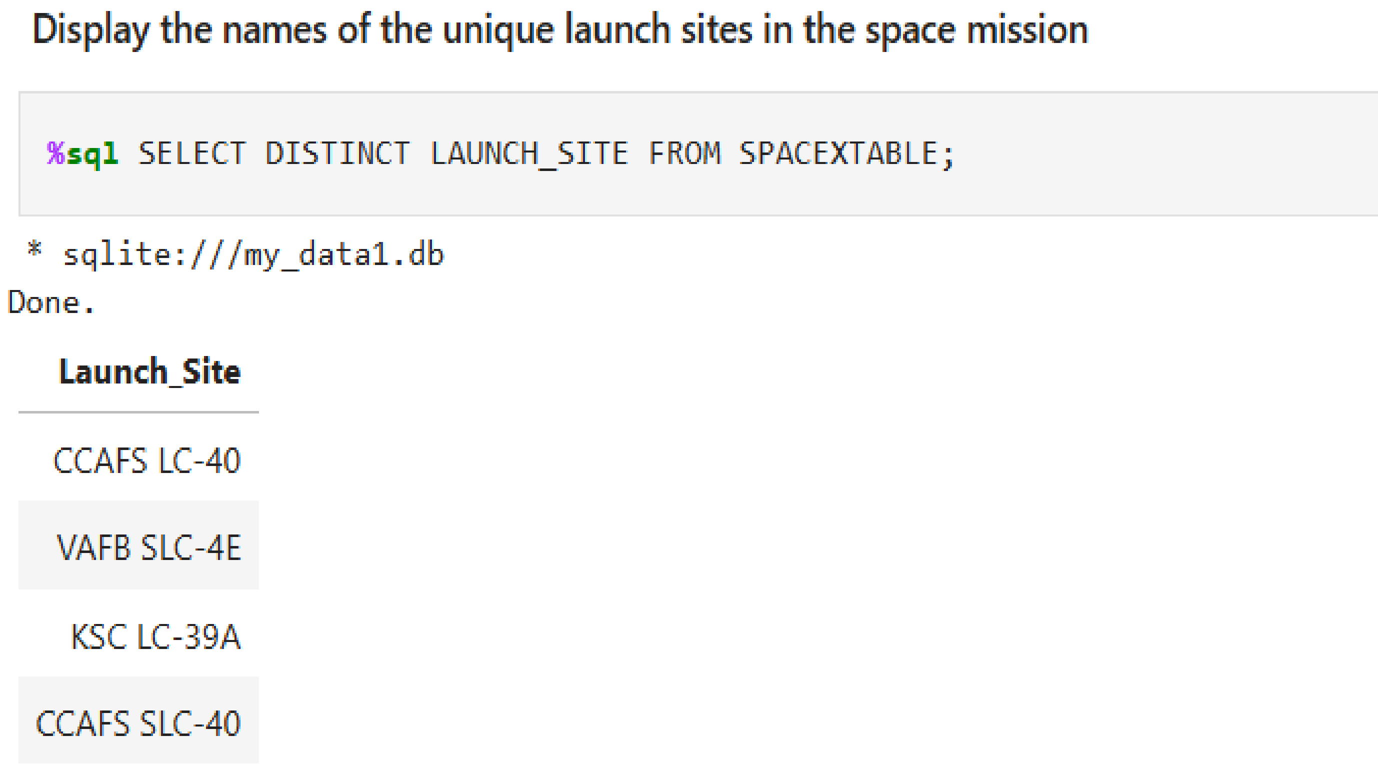
Launch Success Yearly Trend



All Launch Site Names

Launch Site Names:

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



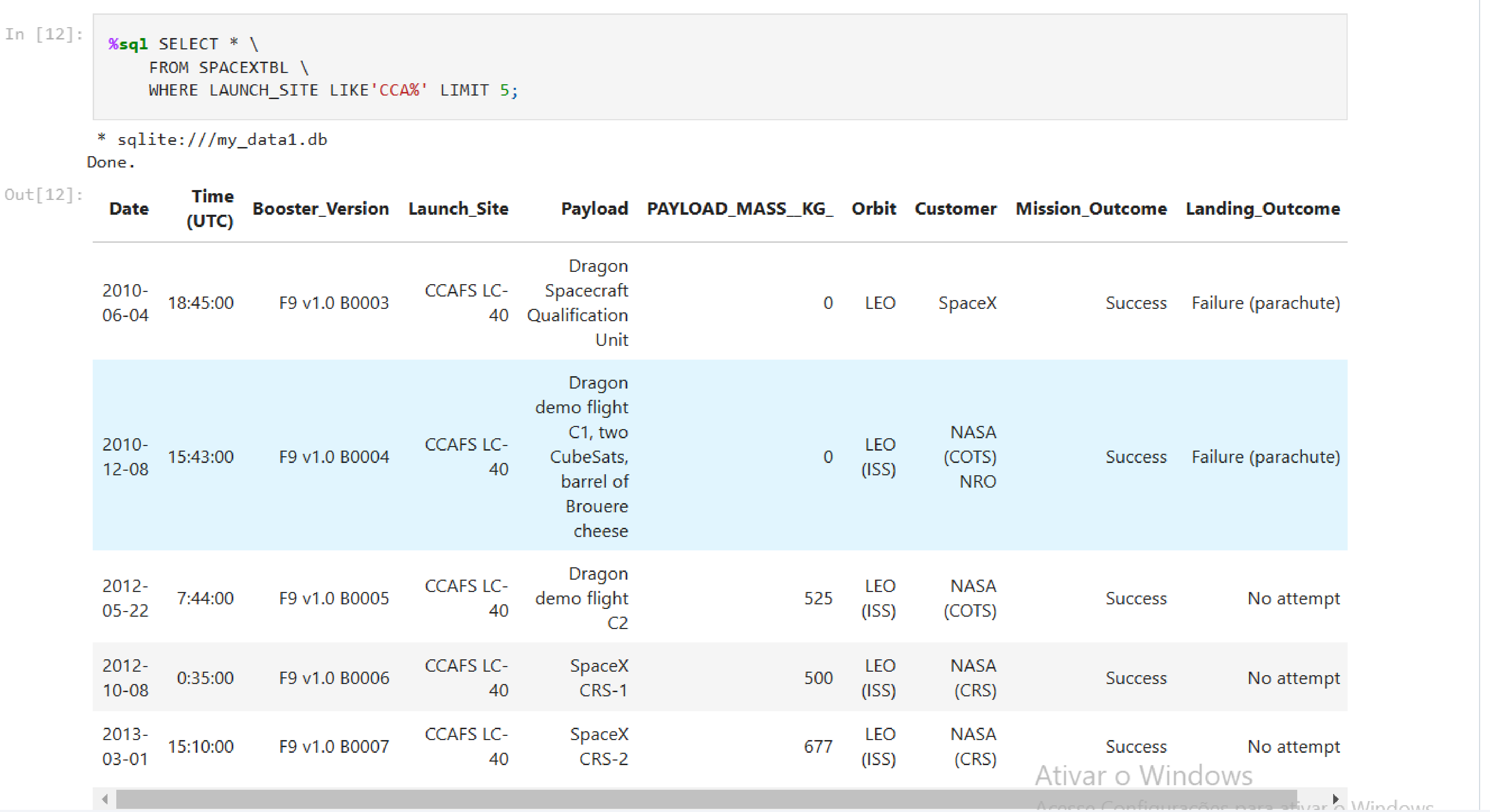
Display the names of the unique launch sites in the space mission

```
%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

Launch Site Names Begin with 'CCA'



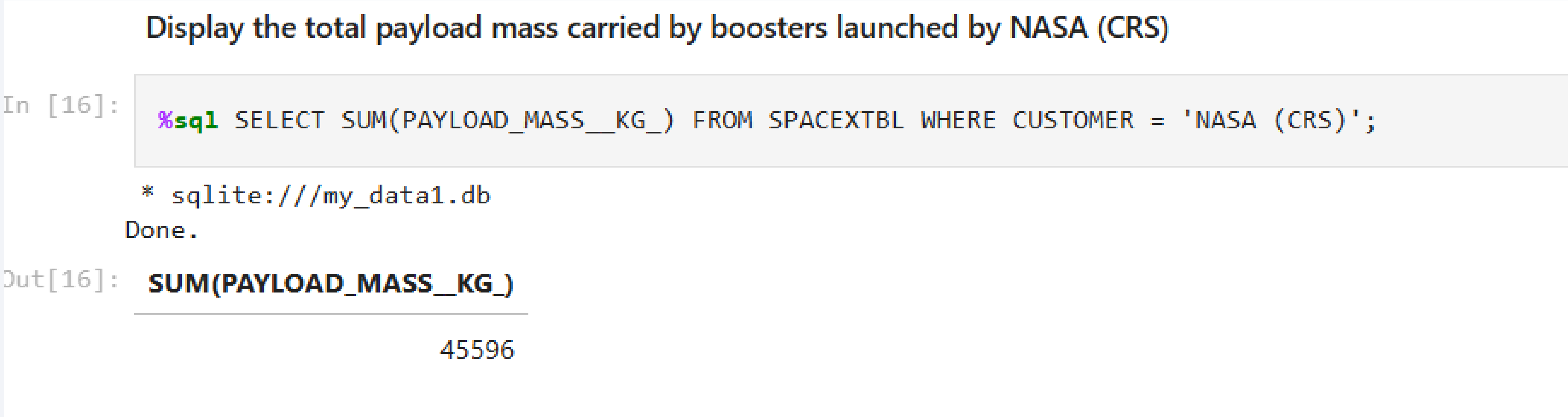
```
In [12]: %sql SELECT * \
             FROM SPACEXTBL \
             WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;

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

Out[12]:

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



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

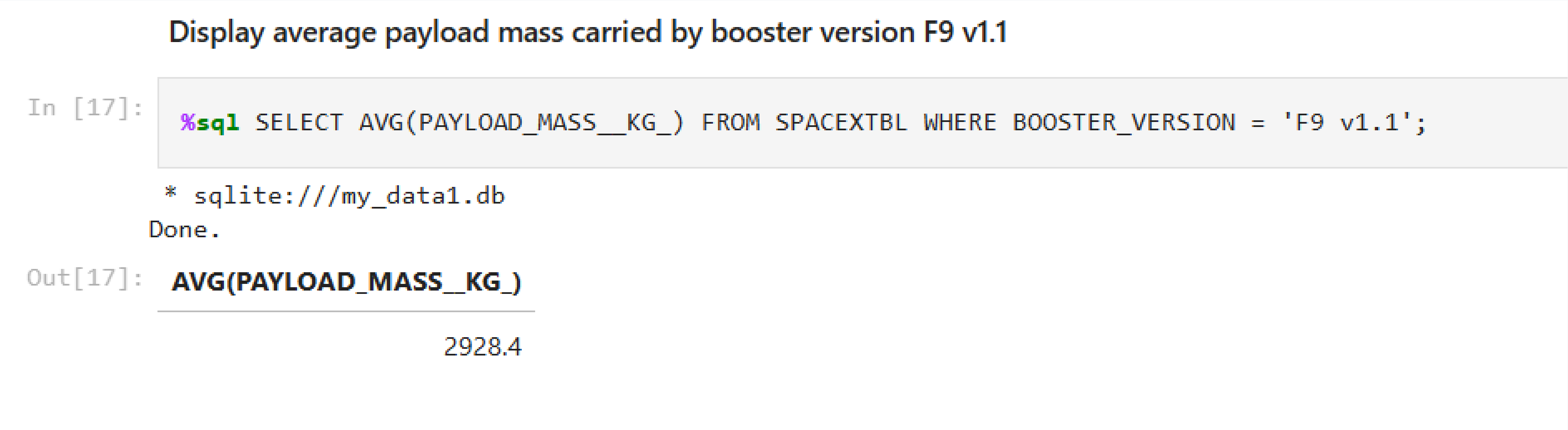
```
In [16]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

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

Out[16]:

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1



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

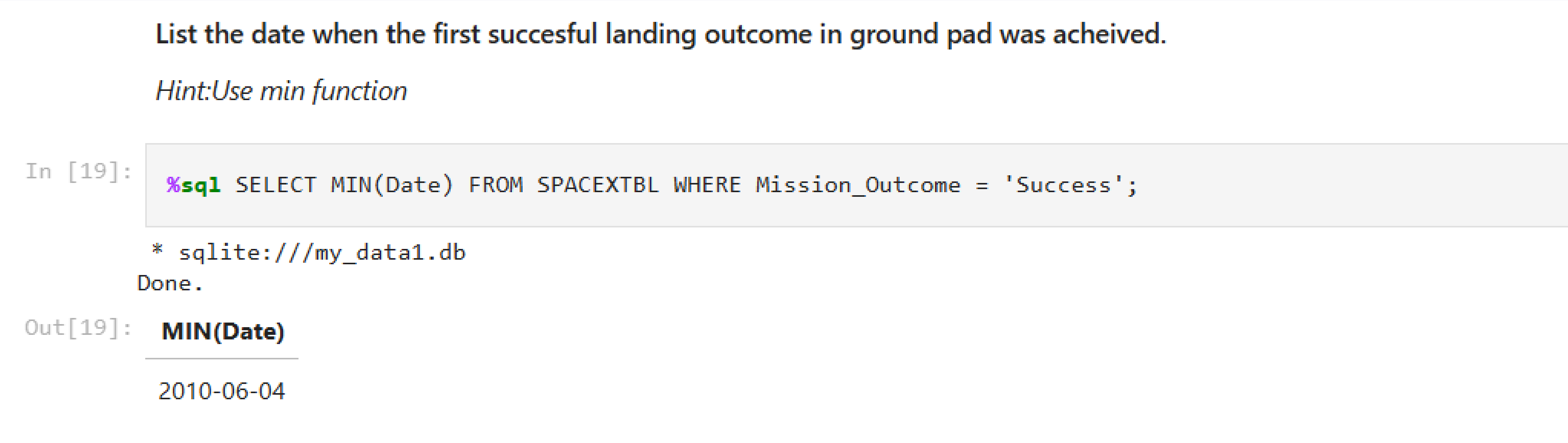
```
In [17]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[17]:

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

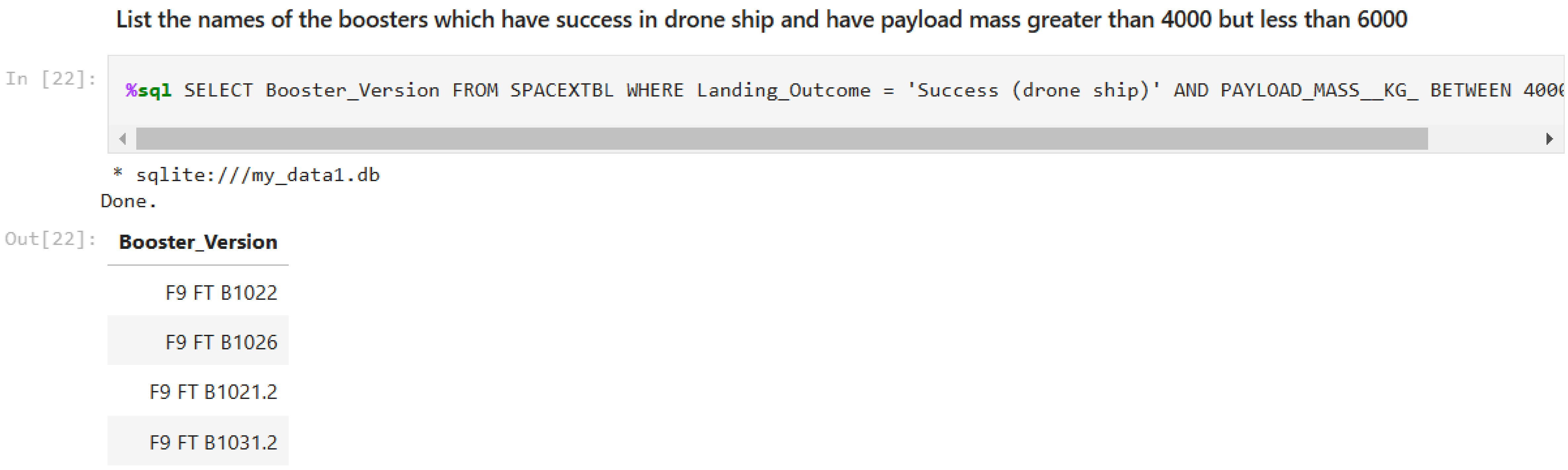
```
In [19]: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Mission_Outcome = 'Success';
```

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

Out[19]:

MIN(Date)
2010-06-04

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

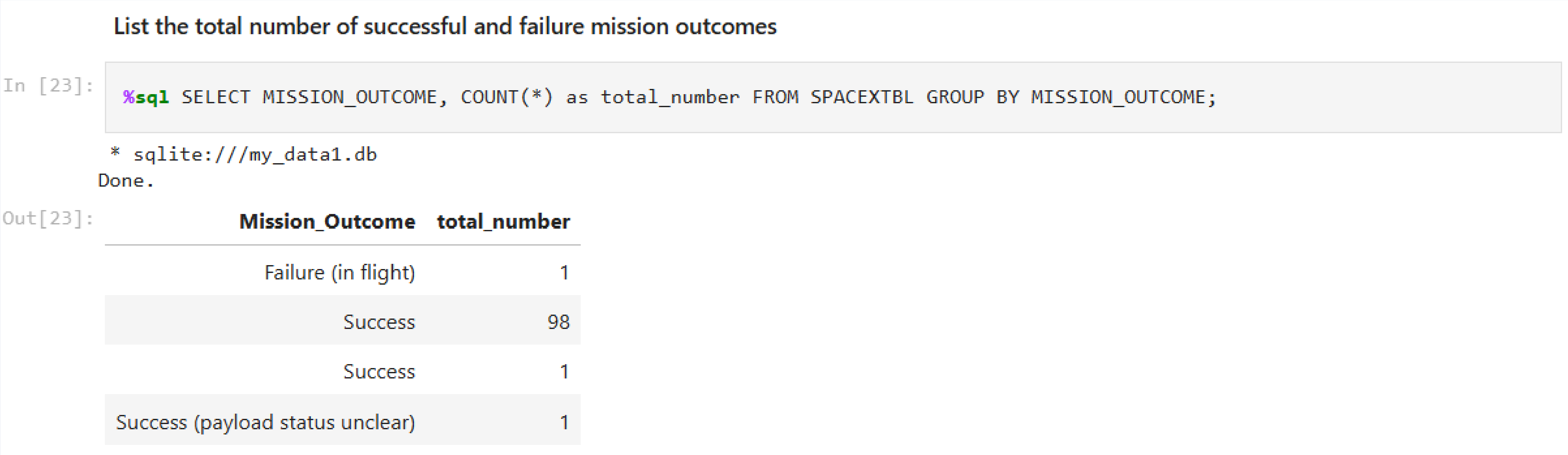
```
In [22]:  %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 400
```

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

Out[22]:

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

In [23]:

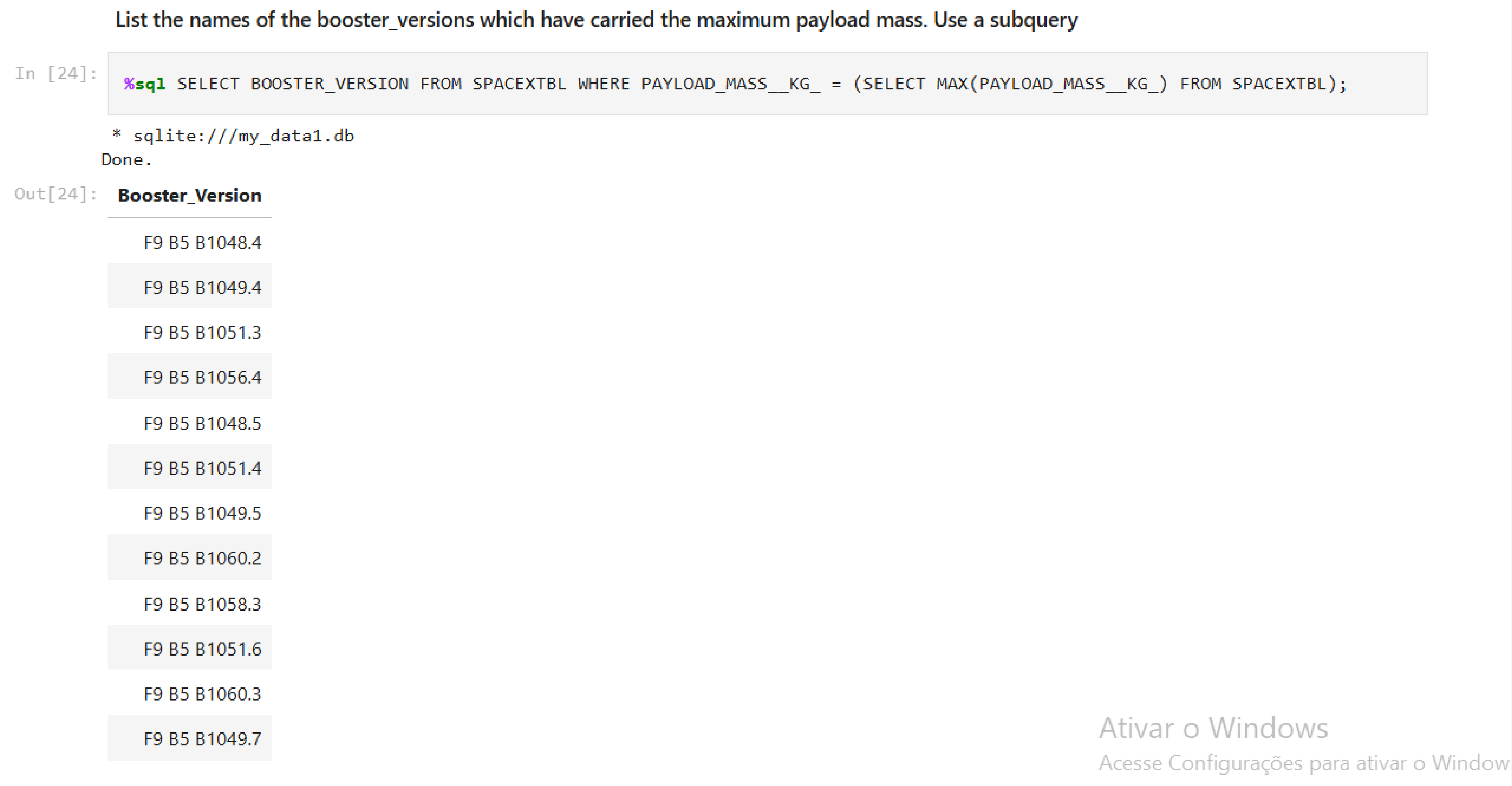
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

Out[23]:

Mission_Outcome	total_number
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

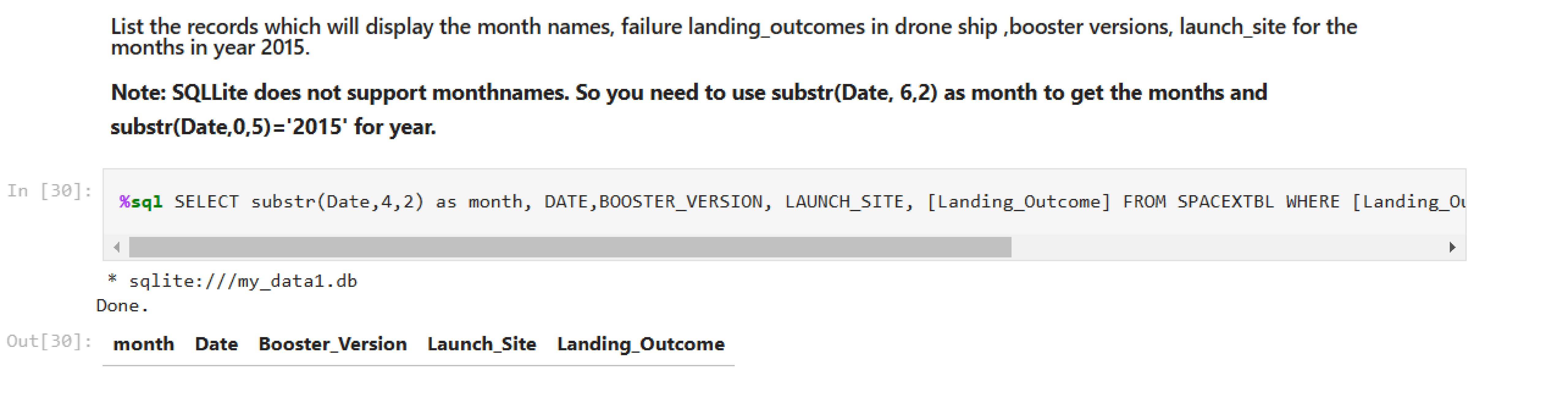
```
In [24]: %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[24]:

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, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

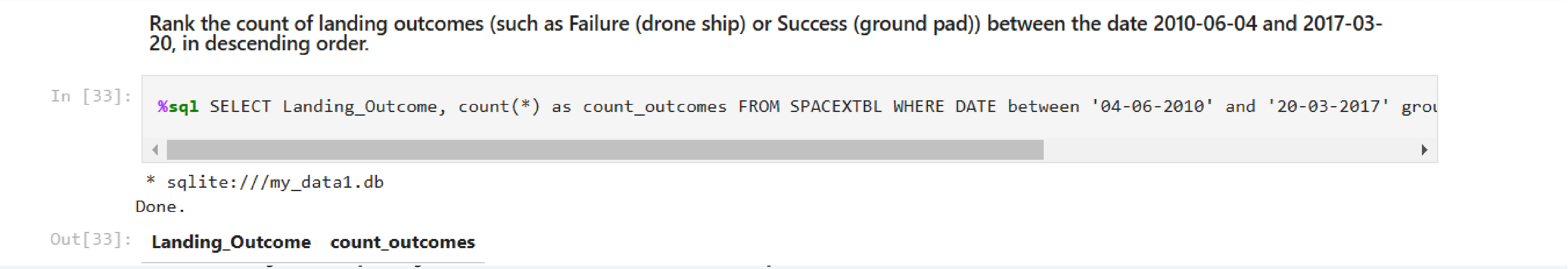
```
In [30]: %sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] FROM SPACEXTBL WHERE [Landing_Ou
```

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

Out[30]:

month	Date	Booster_Version	Launch_Site	Landing_Outcome

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



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.

```
In [33]: %sql SELECT Landing_Outcome, count(*) as count_outcomes FROM SPACEXTBL WHERE DATE between '04-06-2010' and '20-03-2017' grou
```

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

Out[33]:

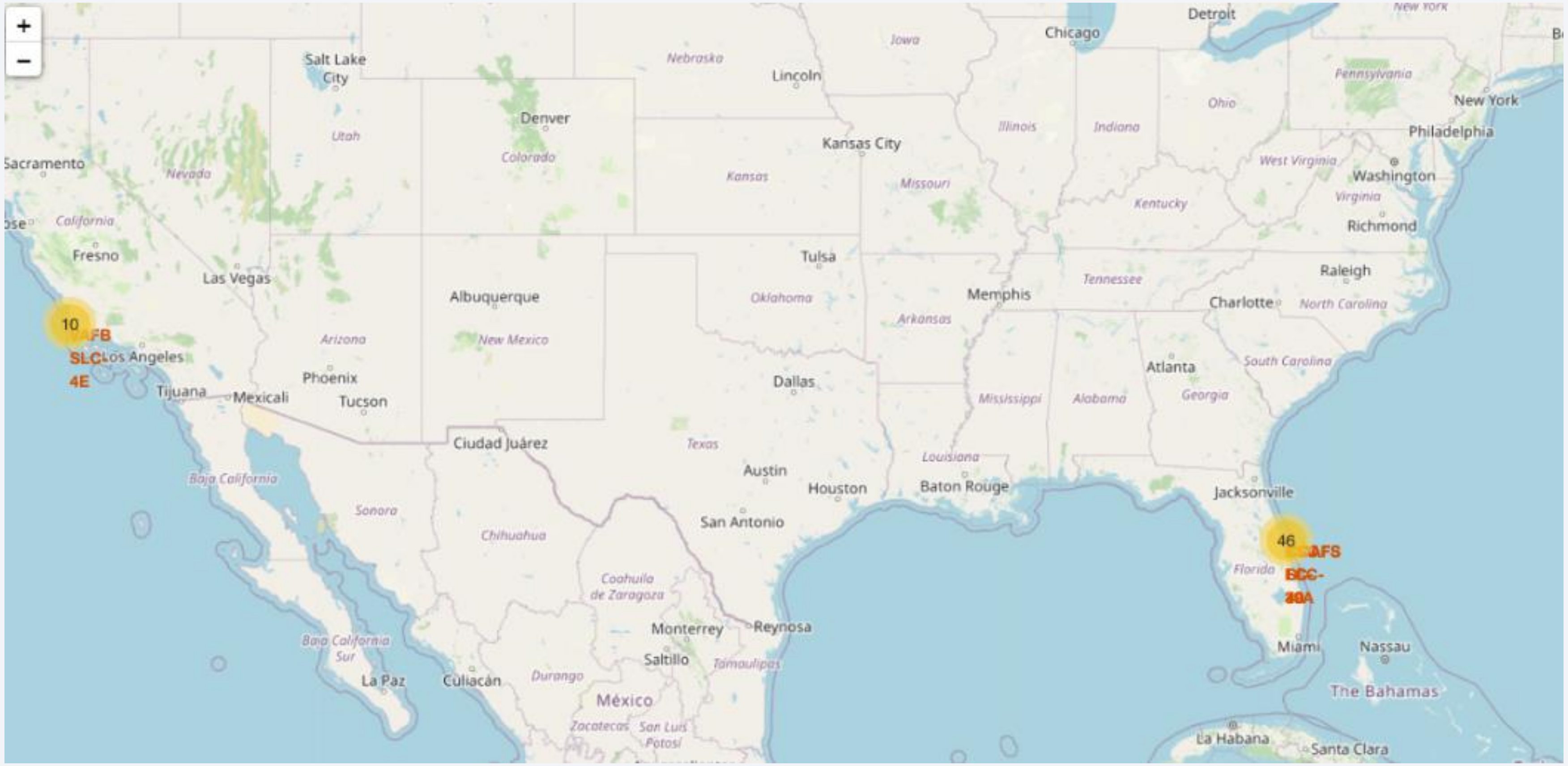
Landing_Outcome	count_outcomes

Section 3

Launch Sites Proximities Analysis

<Launching Sites>

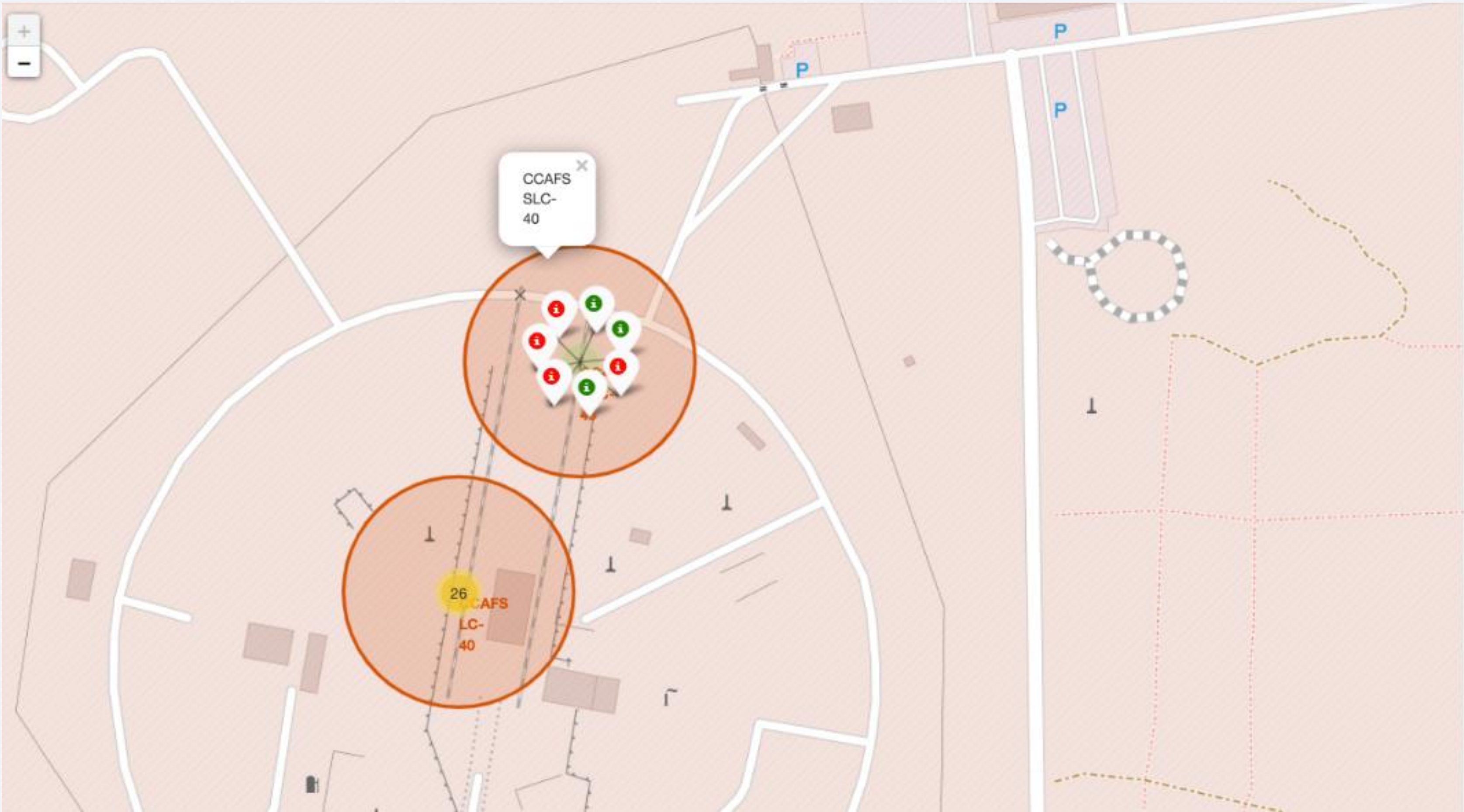
Launching near the equator offers significant advantages. The proximity to the equator simplifies launches to equatorial orbits, benefiting from the Earth's rotation for prograde orbits. Rockets launched from these sites receive an extra natural boost thanks to the Earth's rotational speed, which helps reduce the need for additional fuel and boosters, ultimately lowering launch costs.



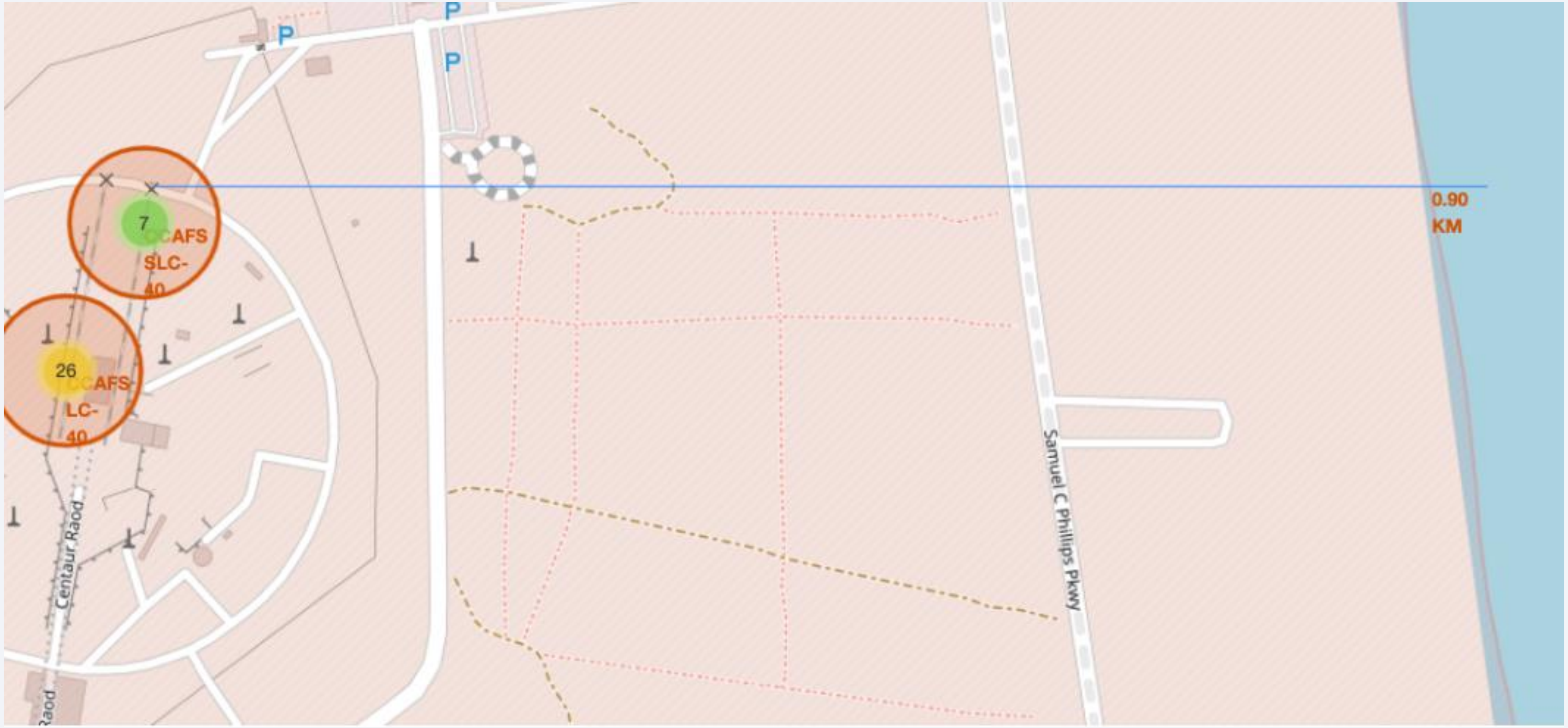
<Launch Outcomes>

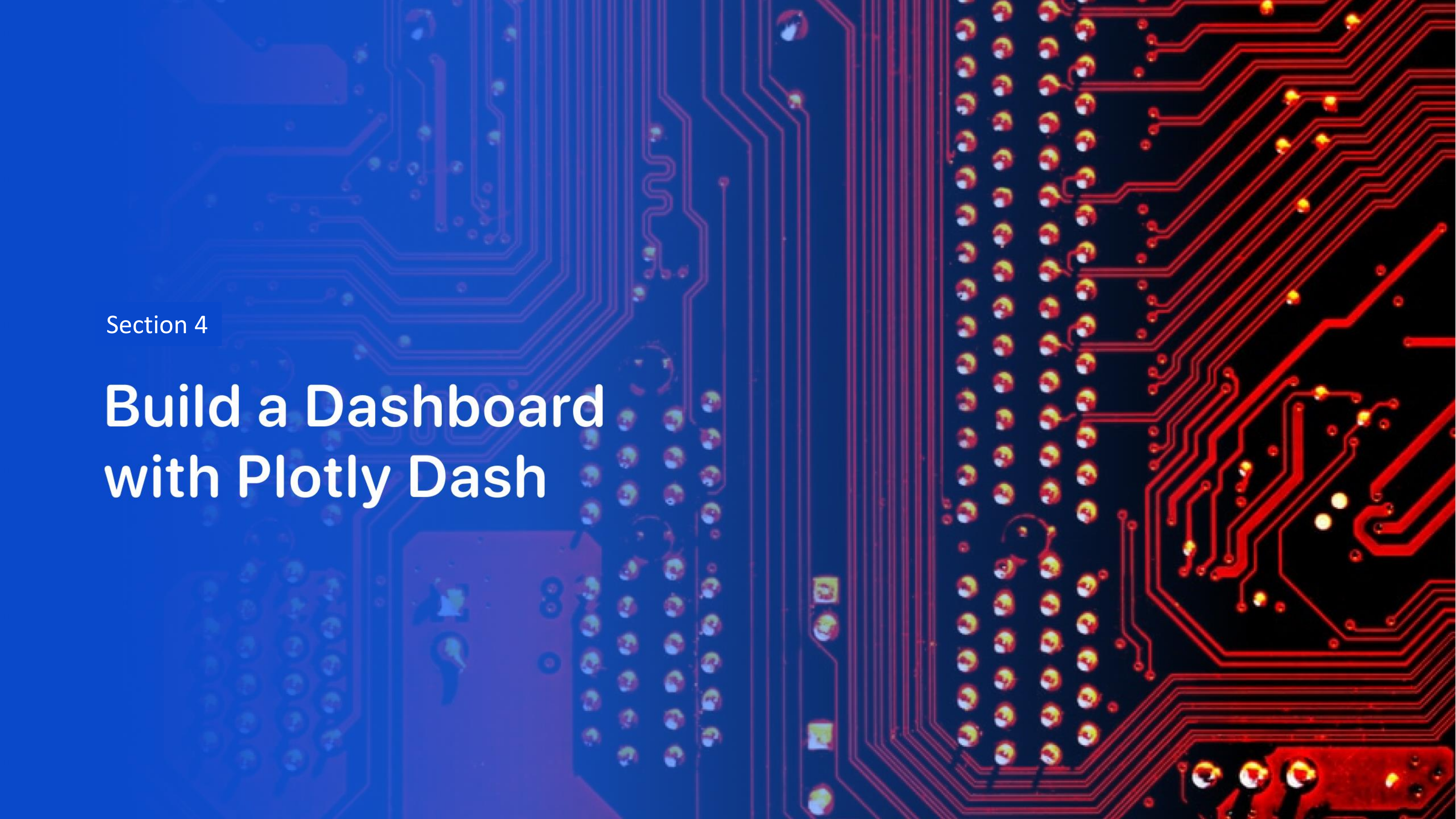
Outcomes:

- Green markers indicate successful launches.
- Red markers indicate unsuccessful launches.
- Launch site CCAFS SLC-40 has achieved a success rate of 3 out of 7 launches, which equates to 42.9%



<Distance to Proximities>


7
CCAFS SLC-40
26
CCAFS LC-40
0.90 KM
Samuel C Phillips Pkwy
Centaur Raod
Raod
P
P
P



Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

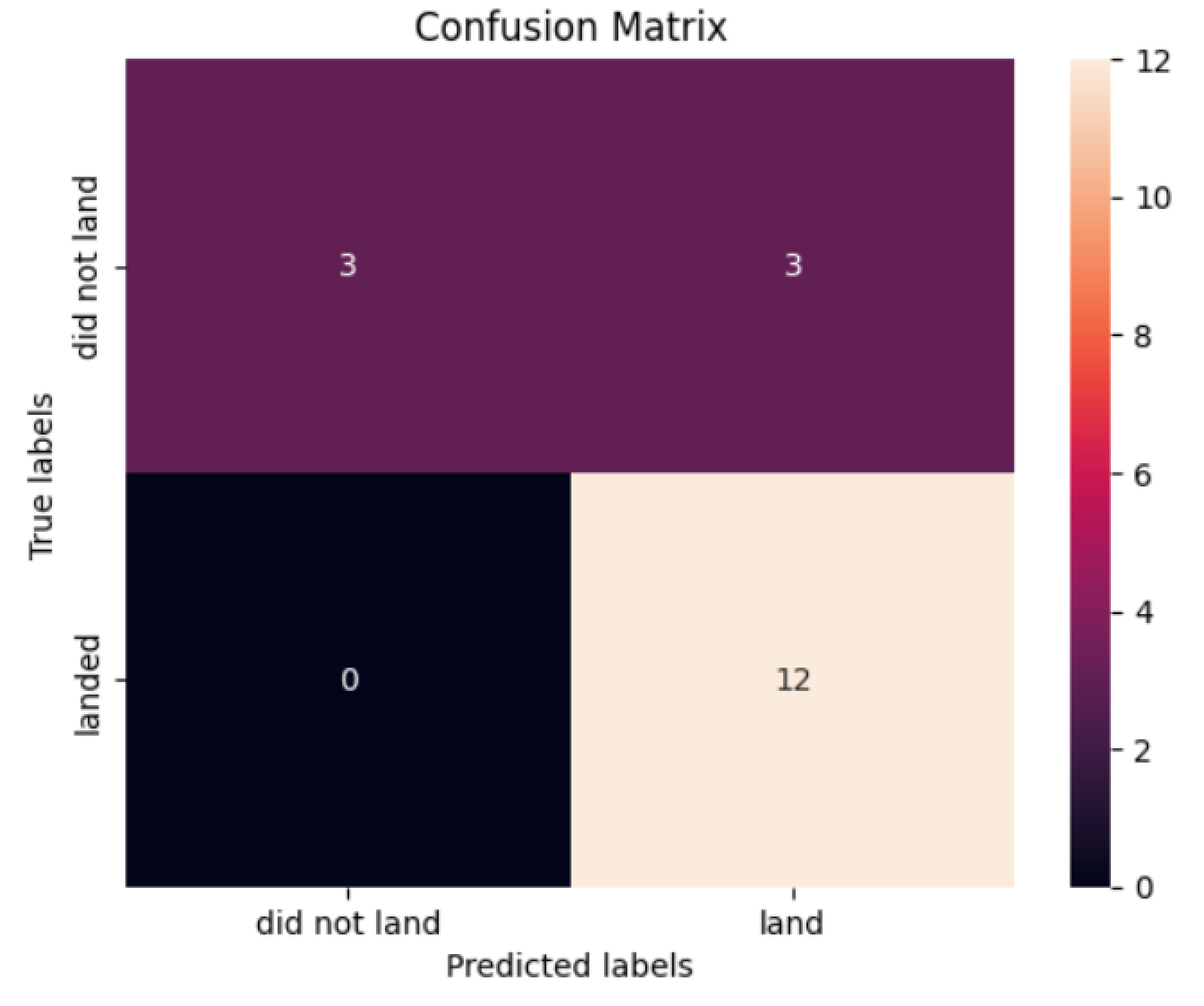
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix



Conclusions

The analysis of SpaceX API data provides valuable insights into the factors influencing the success of space launches. We have observed that the success rate of launches has improved over time, with KSC LC-39A exhibiting the highest success rate among landing sites. Additionally, launches to specific orbits such as ES-L1, GEO, HEO, and SSO have a 100% success rate.

The proximity of launch sites to the equator and the coast is a strategic advantage, leveraging Earth's rotation to save fuel. Launch sites are carefully positioned to minimize risk of damage in case of failures, while remaining accessible for logistical support.

Predictive analysis identified the decision tree model as the most effective for predicting launch success, emphasizing the importance of selecting appropriate models to improve prediction accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!