IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sven Malama
03/16/2024

Outline

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

Executive Summary

Summary of methodologies:

- Acquisition of data via Application Programming Interfaces (APIs)
- Gathering data through systematic web data extraction techniques
- Data cleaning and preprocessing for analysis readiness
- Conducting initial data investigation to uncover patterns and insights
- Visual exploration of data to identify trends and anomalies
- Leveraging Folium for dynamic data visualization on maps
- Application of machine learning models for predictive insights

Summary of all results:

- Key findings from the exploratory analysis of the dataset
- Presentation of interactive visual data representations, exemplified through screenshots
- Outcome of employing predictive analytics through machine learning algorithms

Introduction

Project Context and Objectives:

- Examination of cost efficiency in space launch markets, highlighting SpaceX's reusable Falcon 9's competitive price point of $62 million.
- Comparative analysis of market alternatives pricing over $165 million for non-reusable launch systems.
- The pivotal role of first-stage reusability in launch cost reduction, influencing competitive bidding scenarios.
- Development of a predictive model to assess the likelihood of successful first-stage landings, impacting financial and strategic planning for potential SpaceX competitors.

Research Challenges and Queries:

- Identification of critical variables that predict successful first-stage recovery in rocket launches.
- Analysis of feature interplay that correlates with the likelihood of first-stage landing success.
- Elucidation of the essential conditions that underpin a robust and repeatable first-stage landing process for rockets.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using web scraping from Wikipedia’s list of Falcon 9 launches
- Perform data wrangling
  - Launch outcomes were classified and analyzed by location to prepare a dataset for predictive success modeling.
- 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

Data Collection:

- Data sourced from SpaceX API and Wikipedia.
- Key data includes launch records, payload mass, orbit details, and landing outcomes.

Data Processing:

- Used requests to fetch data from the API.
- Normalized JSON responses into pandas DataFrames.
- Handled missing values, particularly in PayloadMass.
- Extracted and formatted date information for time series analysis.

Feature Engineering:

- Generated features like BoosterVersion, PayloadMass, Orbit, LaunchSite from the API data.
- Created lists to hold parsed data for constructing a comprehensive DataFrame.

Data Wrangling:

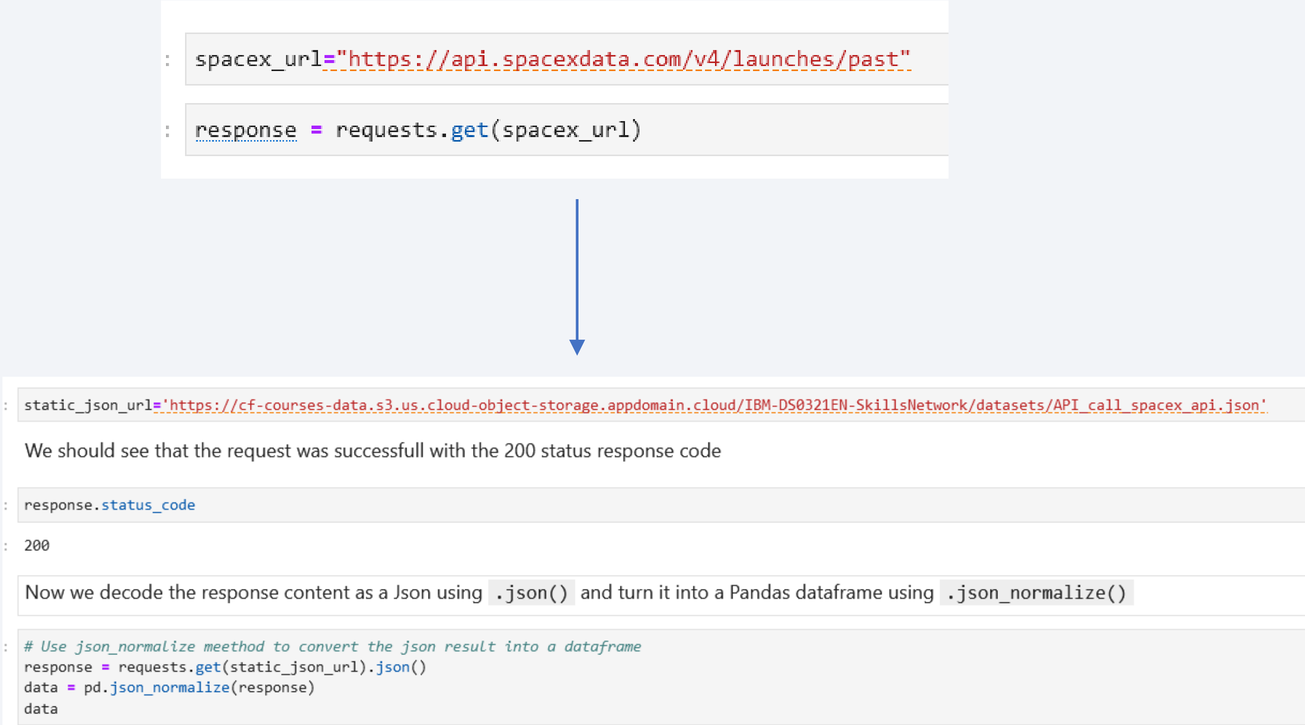
- Cleaned data to filter out irrelevant launches, focusing solely on Falcon 9.
- Imputed missing PayloadMass data with the mean value.
- Preserved 'None' values for LandingPad to indicate no landing pad usage.

Exported Data:

- Finalized dataset exported as CSV for predictive modeling in subsequent phases of the project.

Data Collection – SpaceX API

- We utilized HTTP GET requests to fetch data from the SpaceX API, performed data cleaning, and executed fundamental data wrangling and formatting operations.
- Github Link



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

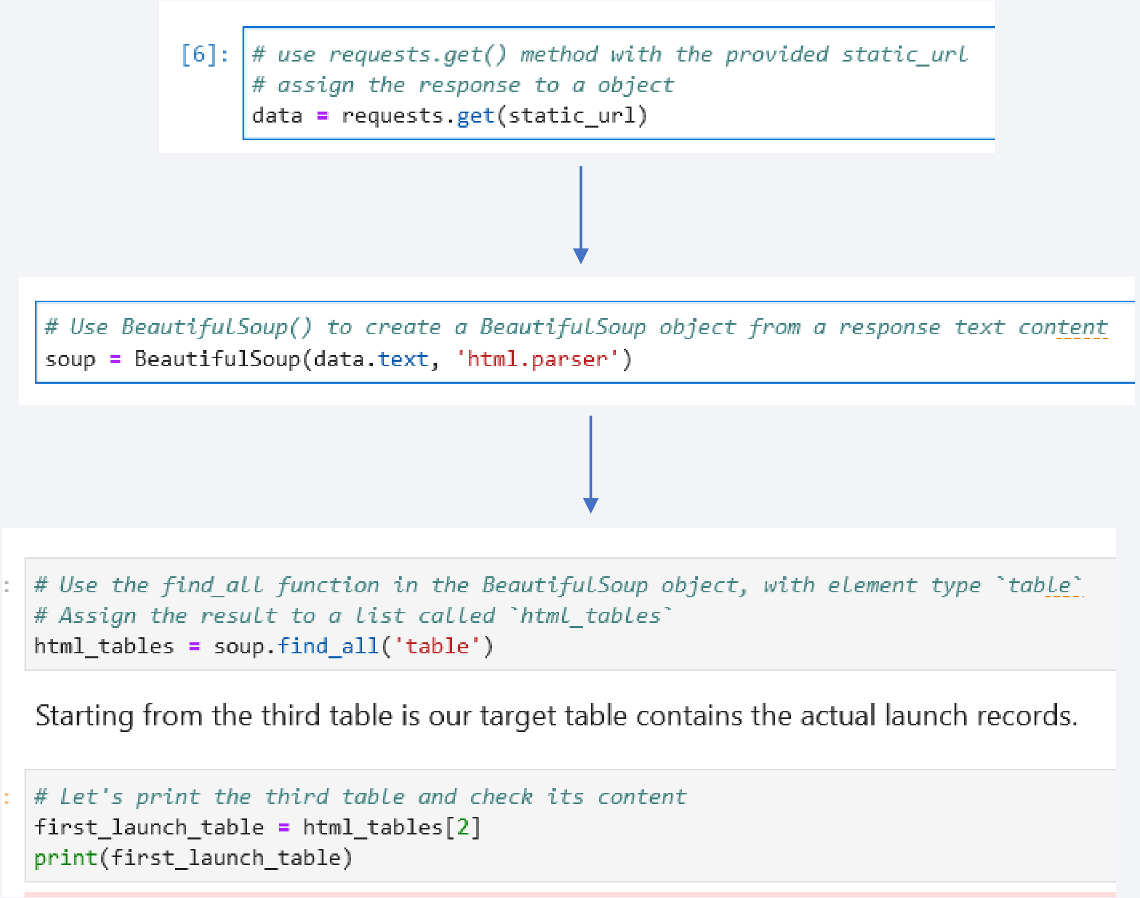
```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
data
```

Data Collection - Scraping

- Conducted web scraping to extract historical Falcon 9 launch records from Wikipedia for landing prediction analysis.
- Utilized BeautifulSoup for data extraction, parsed HTML tables into a Pandas DataFrame, and exported the data to a CSV file.
- Github Link



```
[6]: # use requests.get() method with the provided static_url
     # assign the response to a object
     data = requests.get(static_url)
```

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(data.text, 'html.parser')
```

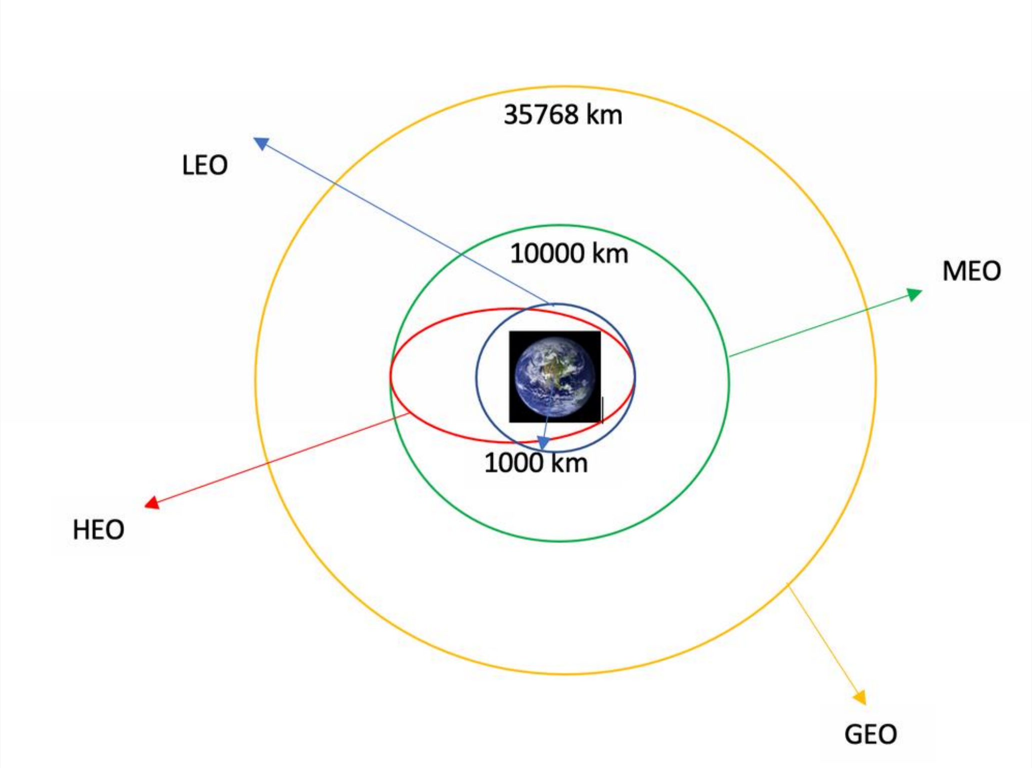
```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
```

Starting from the third table is our target table contains the actual launch records.

```
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

Data Wrangling

- Examined various launch outcomes, including successful and unsuccessful landings, distinguishing them based on the landing attempt location, such as a drone ship, ocean, or ground pad.
- Converted these outcomes into binary training labels: '1' for successful and '0' for unsuccessful landings.
- Conducted data analysis to explore the distribution of launches across different sites, orbits, and outcomes, utilizing Python libraries like Pandas and NumPy.
- Created a new column 'Class' in the dataframe to represent the binary labels for the success of the first stage landing, calculated the overall success rate, and prepared the dataset for further analysis by exporting it.
- Github Link

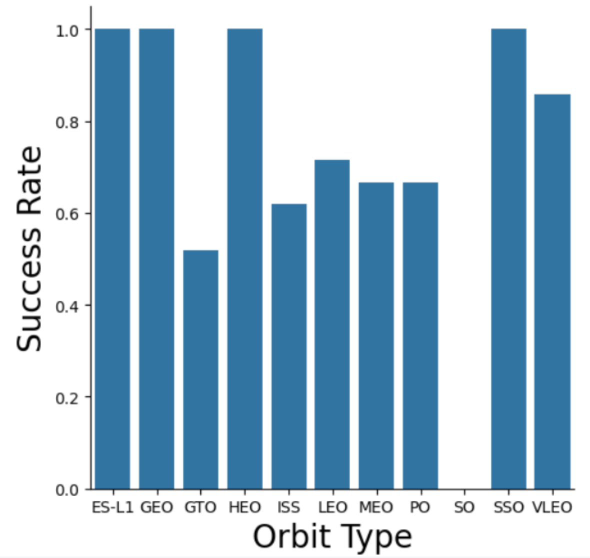


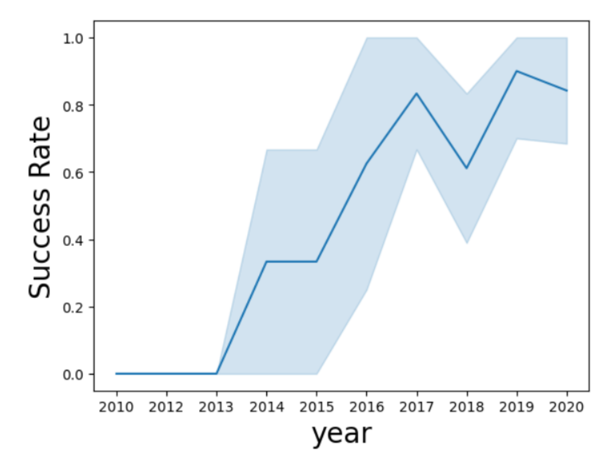
EDA with Data Visualization

Steps and Chart:

- Plotted a scatter chart of Flight Number vs. Payload Mass to investigate the effect of mission sequence and payload size on landing success.
- Created categorical plots for Launch Site against Flight Number and Payload Mass to identify launch site preferences and outcomes.
- Developed bar charts to compare success rates across different orbits, revealing which had higher successful landing probabilities.
- Utilized scatter plots to correlate Orbit Type with Flight Number and Payload Mass, exploring operational patterns and success correlations.
- Compiled a yearly trend line chart to visualize the evolution of landing success rates over time.
- Performed one-hot encoding on categorical data to prepare for machine learning model input

- Github Link





EDA with SQL

- Retrieved unique launch sites from the SpaceX dataset.
- Displayed records where launch sites start with 'CCA'.
- Calculated the total payload mass for missions contracted by NASA (CRS).
- Determined the average payload mass carried by booster version F9 v1.1.
- Found the date of the first successful landing on a ground pad.
- Listed boosters with successful drone ship landings and specific payload mass criteria.
- Counted the total number of successful missions versus failed missions.
- Identified booster versions that carried the maximum payload mass.
- Extracted records of failed drone ship landings along with booster versions and launch sites for 2015.
- Ranked the count of various landing outcomes within a specified date range in descending order.

- Github Link

Build an Interactive Map with Folium

Summary of Map Objects Created

- **Launch Site Markers:** Indicate exact locations of SpaceX launch sites.
- **Highlight Circles:** Visually emphasize the launch site areas.
- **Launch Outcome Markers:** Show success (green) or failure (red) at each site for quick visual assessment.
- **Proximity Lines:** Measure distances from sites to key local features (coastline, railways, etc.).
- **Distance Markers:** Provide numerical distance data from launch sites to important infrastructures.

Reasons for Adding Map Objects

- To analyze geographical advantages for launch success.
- To evaluate launch site safety in relation to nearby populations and structures.
- To assist in operational planning and potential future site selection.

Github Link

Build a Dashboard with Plotly Dash

Dashboard Components and Interactions:

- **Launch Site Dropdown**: Allows users to select a launch site or view data for all sites.
- **Success Pie Chart Callback**: Visualizes the success rate of launches per selected site, enhancing decision-making for launch planning.
- **Payload Range Slider**: Enables filtering of launch data based on payload mass, facilitating analysis of payload correlation with launch success.
- **Scatter Plot Callback**: Displays success vs. payload mass, with color-coded booster versions, for in-depth trend analysis.

Reasons for Adding Plots and Interactions

•**Strategic Analysis**: Dropdown menus and sliders allow users to sift through data to identify trends and inform strategic decisions.

•**Visual Correlation**: Pie charts and scatter plots provide immediate visual insights into success rates and potential payload correlations.

•**User Engagement**: Interactive elements engage users, encouraging exploration and discovery of data-driven insights.

•**Comprehensive Overview**: The combination of filters and visual representations offers a multifaceted view of launch outcomes.

Github Link

Predictive Analysis (Classification)

Model Development and Evaluation Summary:

- **Data Preparation**: Loaded the SpaceX launch dataset and standardized the features to create a consistent scale.
- **Splitting Data**: Divided the dataset into training and testing sets to evaluate the model's performance.
- **Model Selection**: Applied different classification algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors.
- **Hyperparameter Tuning**: Utilized GridSearchCV to find the optimal hyperparameters for each model.
- **Evaluation**: Assessed each model's accuracy on the test set and used a confusion matrix to understand their performance in detail.
- **Best Model Identification**: Compared the accuracy of all models to select the best-performing one.

Github Link

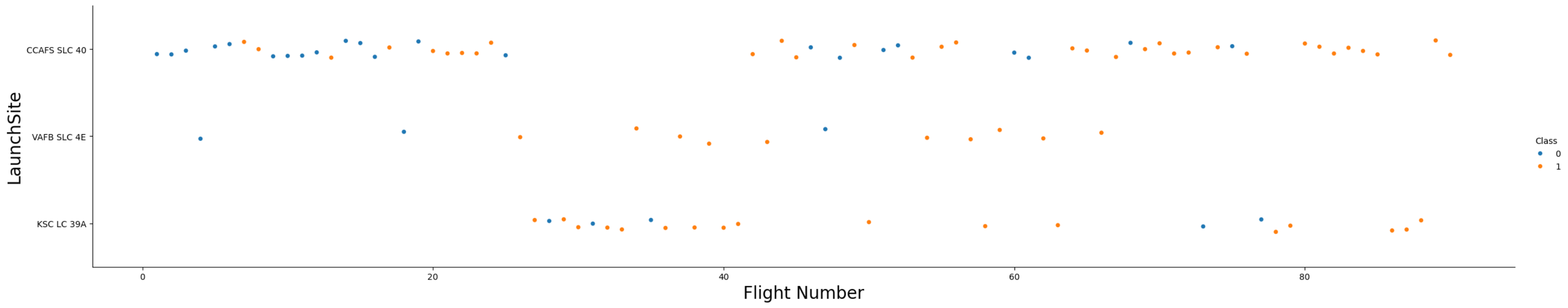
Results

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

Section 2

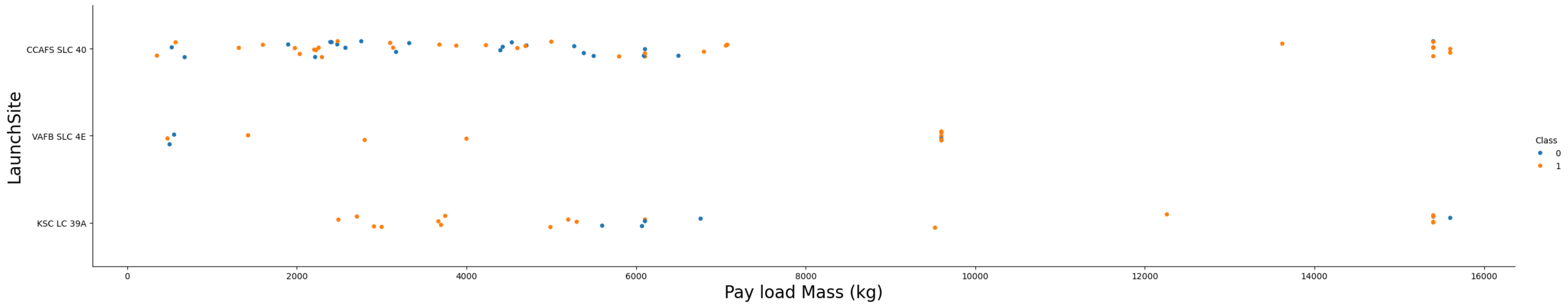
Insights drawn from EDA

Flight Number vs. Launch Site



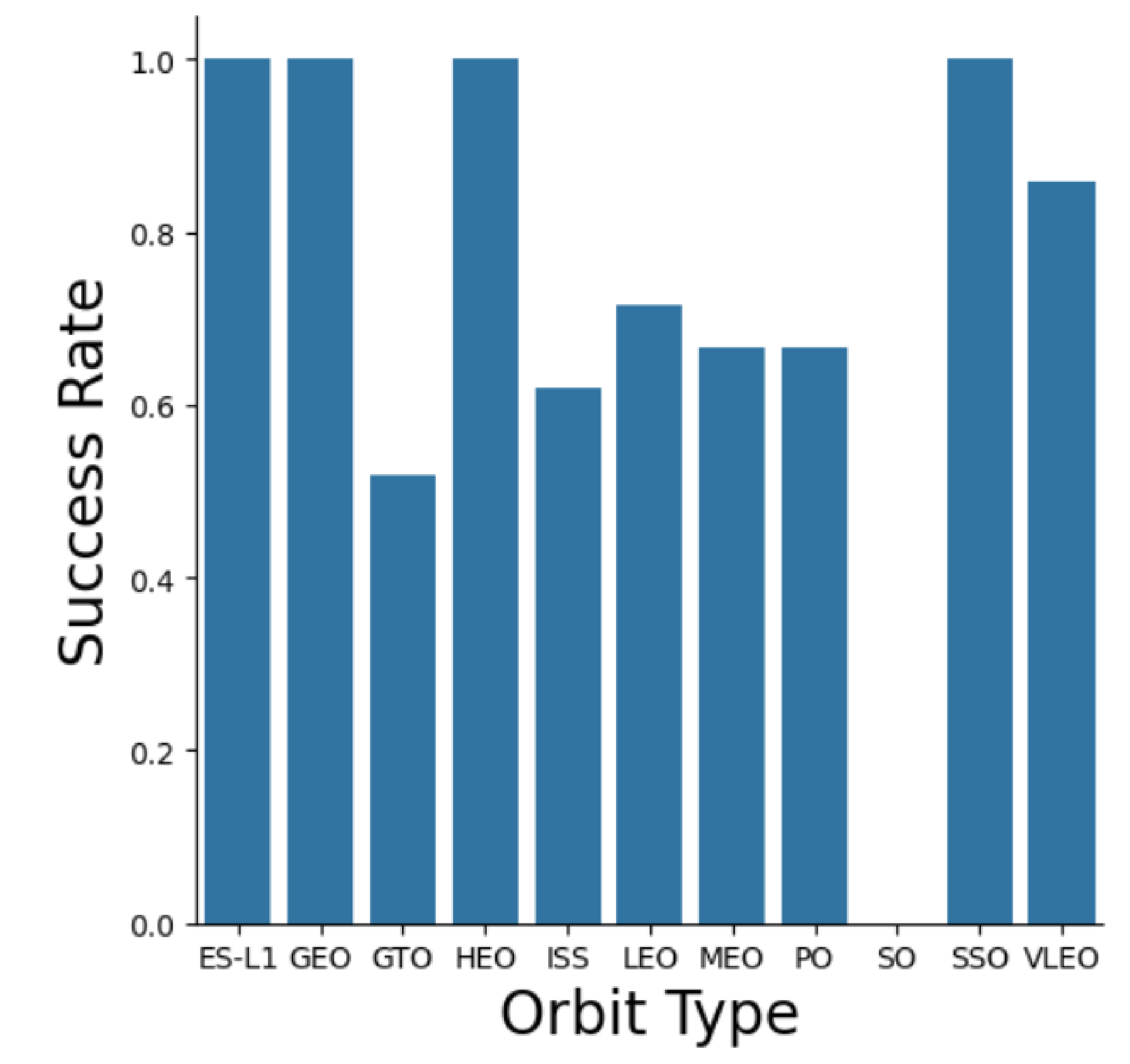
Insight: The scatter plot suggests that the success rate (class 1, orange points) of launches doesn't show a clear correlation with the flight number across different launch sites. Each launch site has both successful and unsuccessful launches (class 0, blue points) distributed throughout the flight history.

Payload vs. Launch Site



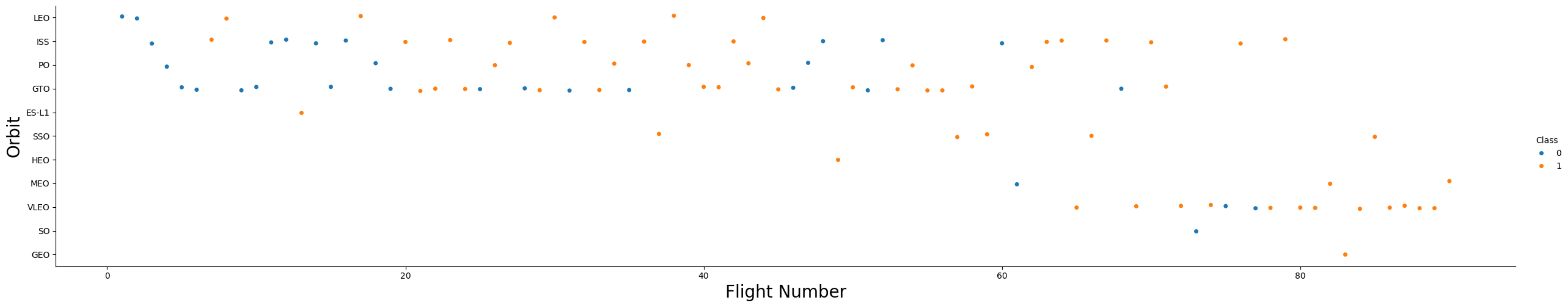
Insight: The scatter plot indicates that successful launches (class 1, orange points) are spread across a range of payload masses, with no clear pattern of success linked to the payload mass. Both successful and unsuccessful launches (class 0, blue points) are observed at various payload sizes for each launch site.

Success Rate vs. Orbit Type



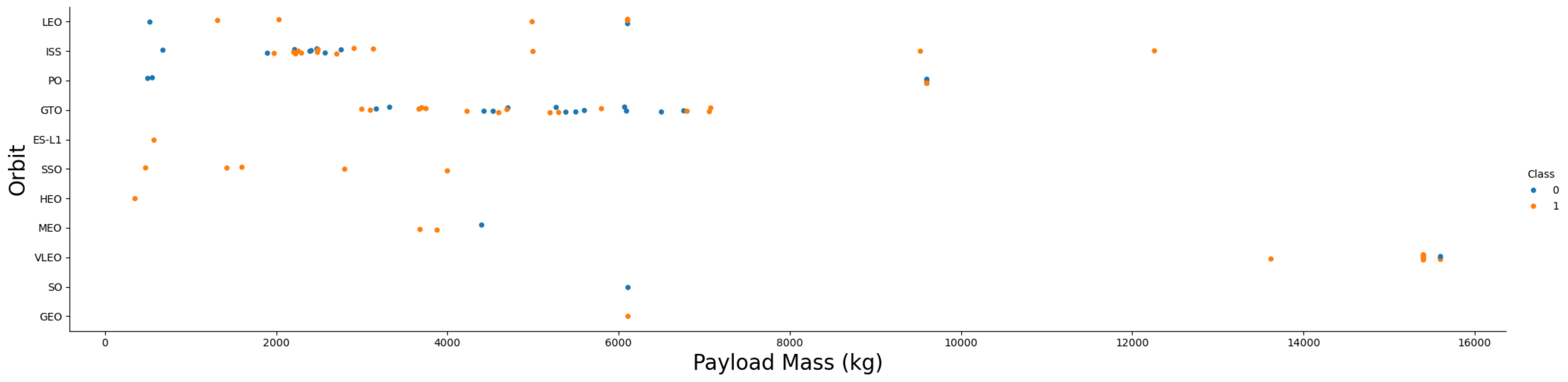
Insight: The bar chart shows the success rate of launches for different orbit types. ES-L1, SSO, and VLEO orbits have the highest success rates, while GTO and MEO have lower success rates compared to others. This suggests that certain orbit types may present more challenges or may have different mission profiles affecting their success rates.

Flight Number vs. Orbit Type



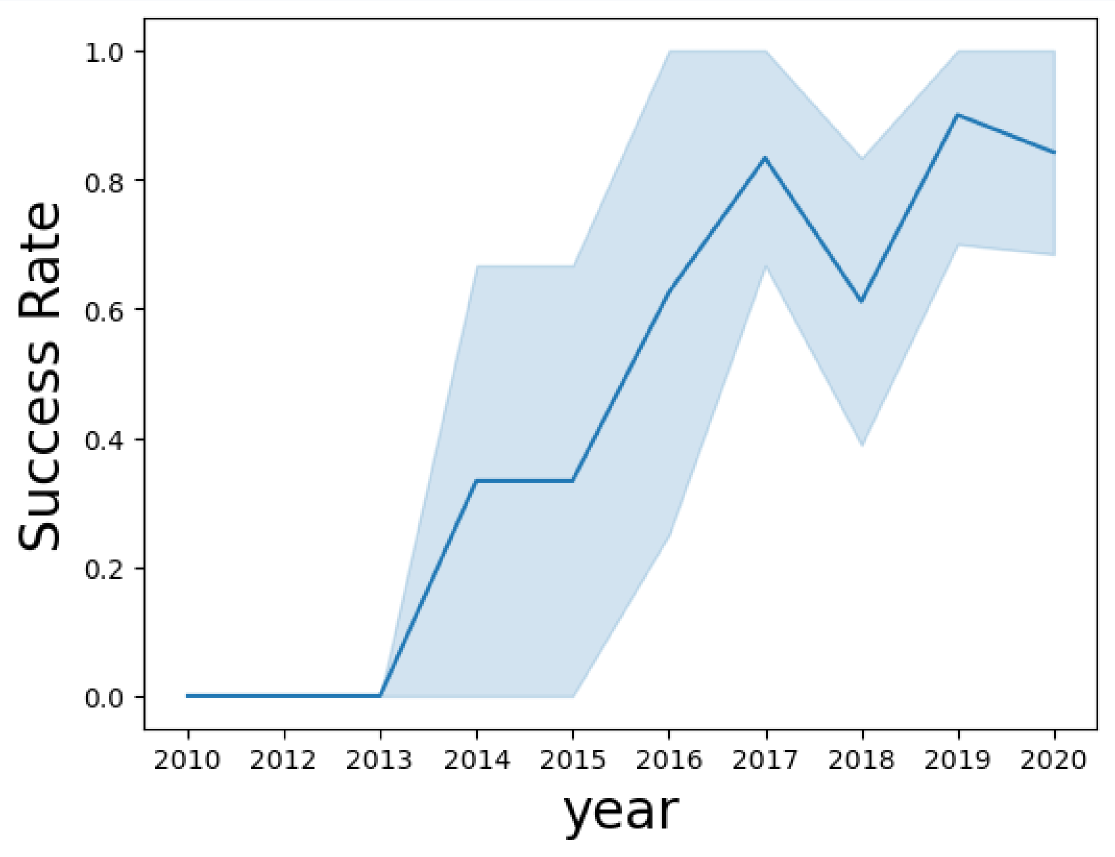
Insight: The scatter plot displays launch outcomes (success or failure) across different orbit types as the flight number increases. Blue points indicate failures (Class 0), and orange points indicate successes (Class 1). The plot suggests that the success rate generally increases with the flight number, indicating possible improvements in launch technology or processes over time. It also shows that launches to certain orbits like ISS have a high success rate, while others like SO have experienced failures, which might suggest that specific orbit types are more challenging to reach.

Payload vs. Orbit Type



Insight: The scatter plot shows launch outcomes across various orbit types with respect to payload mass. Blue dots represent failed launches (Class 0), while orange dots represent successful launches (Class 1). It indicates that most successful launches carry payloads less than 10,000 kg, with some exceptions. Launches to GTO (Geostationary Transfer Orbit) seem to have a broad range of payload masses, indicating flexibility in launch capabilities for this orbit type. There also appears to be a cluster of successful launches with higher payload masses, which may suggest advancements in launch vehicle capabilities or mission profiles that enable heavier payloads to be launched successfully.

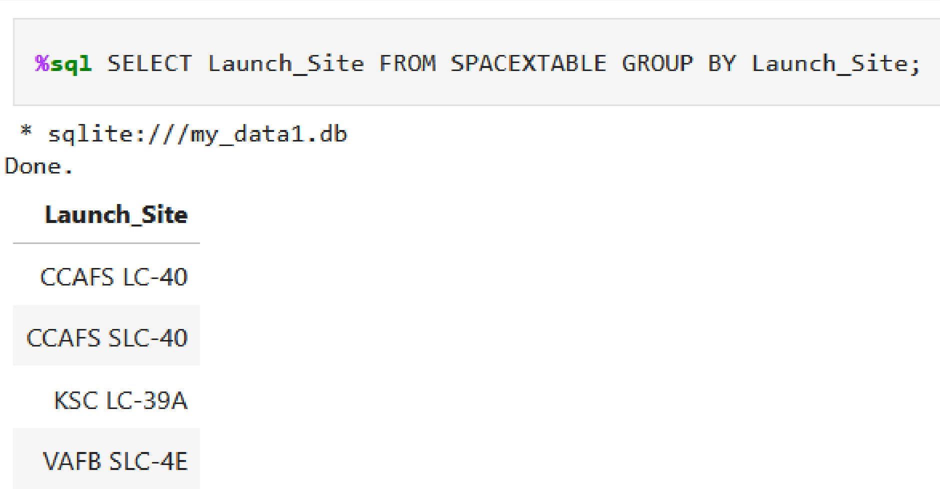
Launch Success Yearly Trend



Insight: The line graph depicts the success rate of launches from 2010 to 2020. There is a clear upward trend in success rate over time, particularly from 2013 to 2017. After 2017, the success rate fluctuates but generally remains high, indicating that over the years, the reliability of the launches has improved. The shaded area might represent the confidence interval or variance in the success rate data year over year.

All Launch Site Names

Following, the unique Launch Site names are displayed



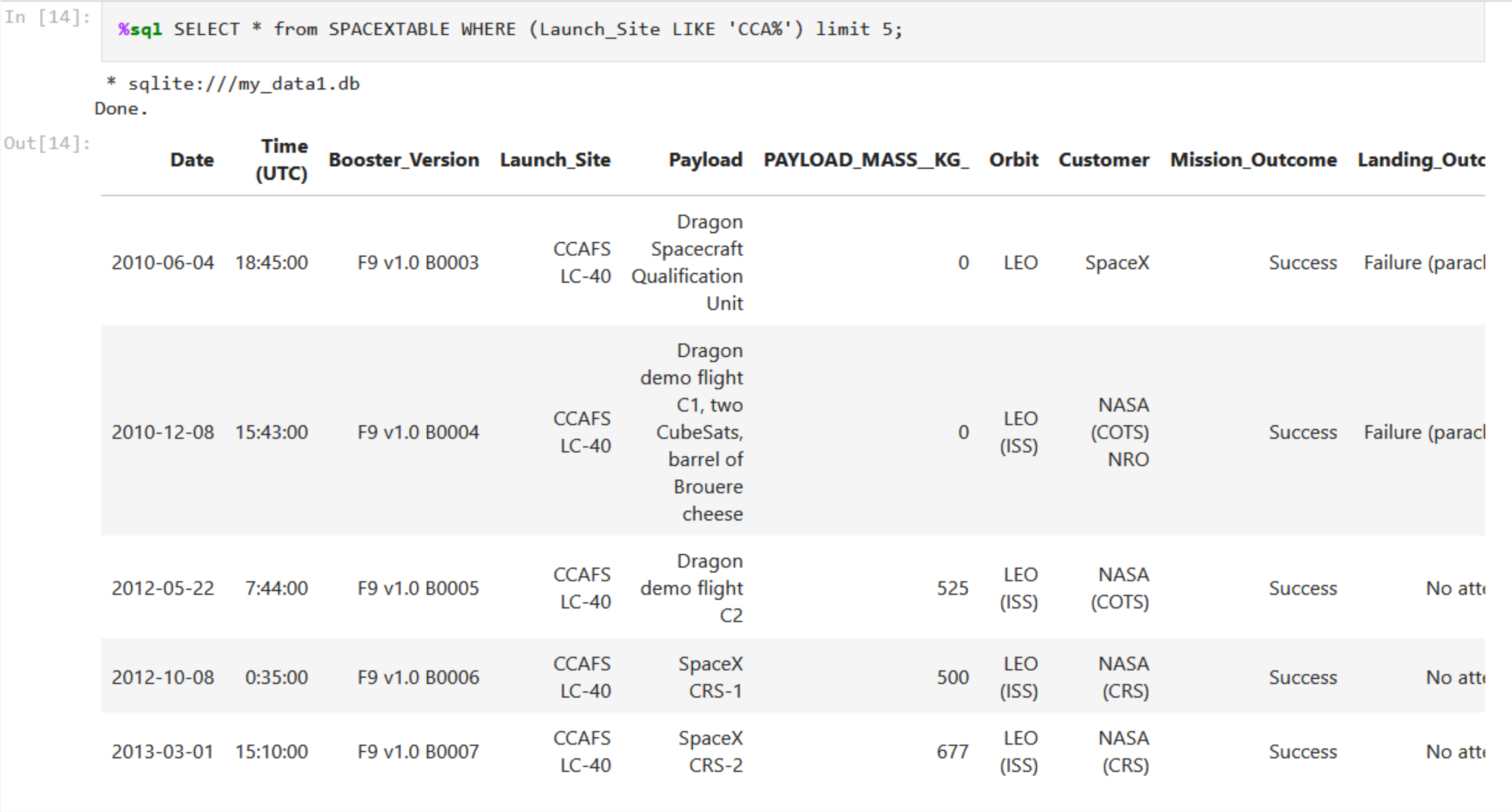
```
%sql SELECT Launch_Site FROM SPACEXTABLE GROUP BY Launch_Site;
```

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

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

Launch Site Names Begin with 'CCA'

Following, the first 5 records where launch sites begin with `CCA` are displayed



```
In [14]:  %sql SELECT * from SPACEXTABLE WHERE (Launch_Site LIKE 'CCA%') limit 5;
```

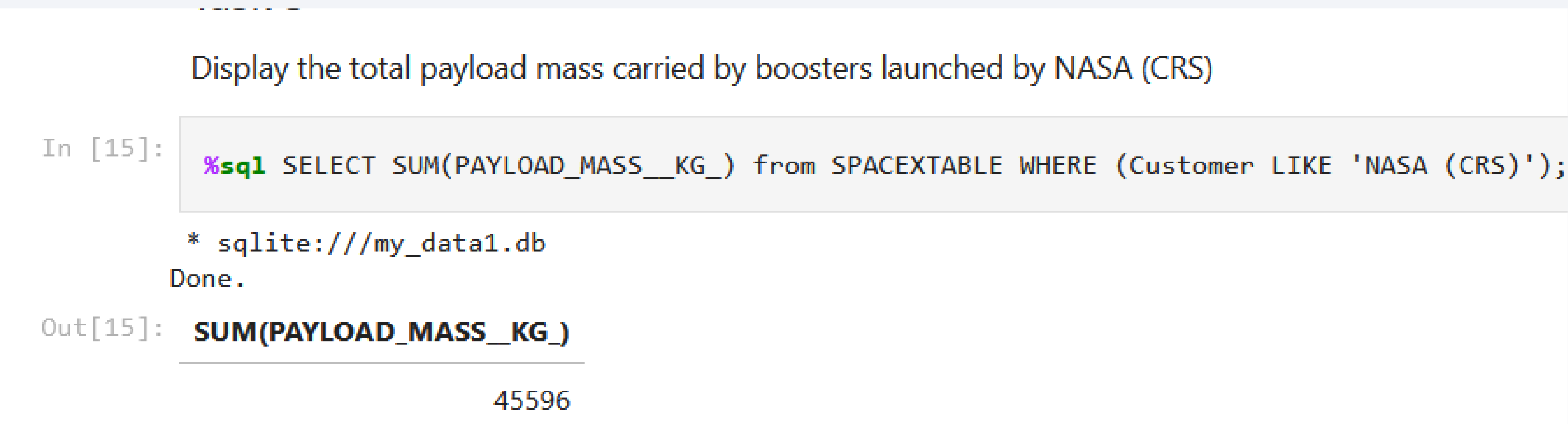
```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No att
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No att
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No att

Total Payload Mass

Following, the total payload carried by boosters from NASA is displayed

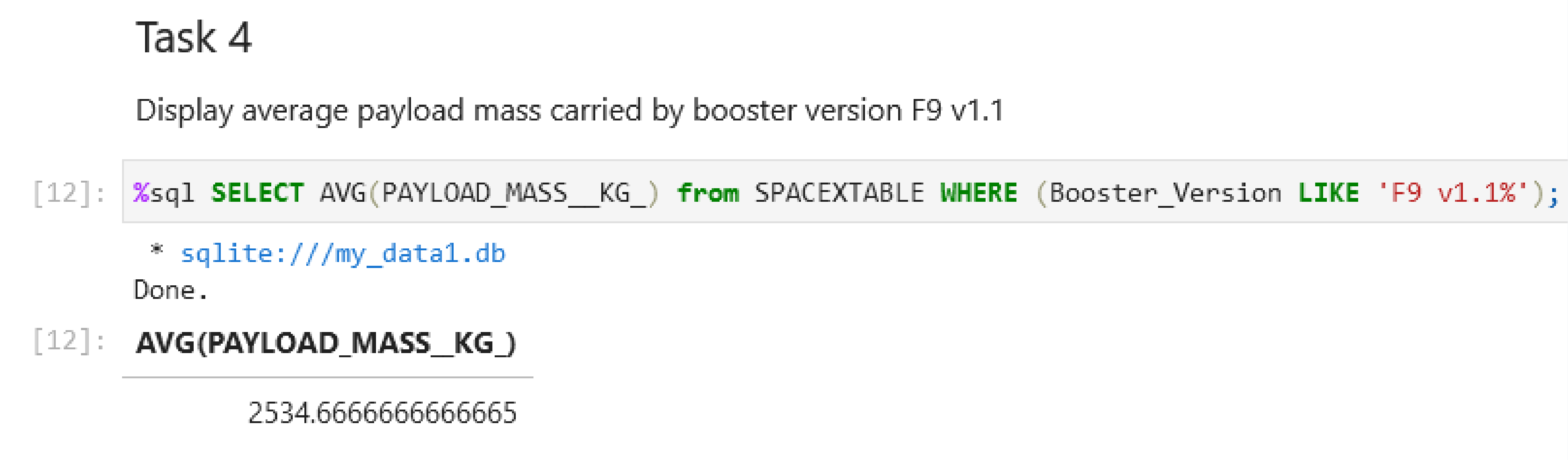


Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [15]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE (Customer LIKE 'NASA (CRS)');

 * sqlite:///my_data1.db
Done.

Out[15]:  SUM(PAYLOAD_MASS__KG_)
                          45596
```

Average Payload Mass by F9 v1.1

Following, the average payload mass carried by booster version F9 v1.1 is displayed



Task 4

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

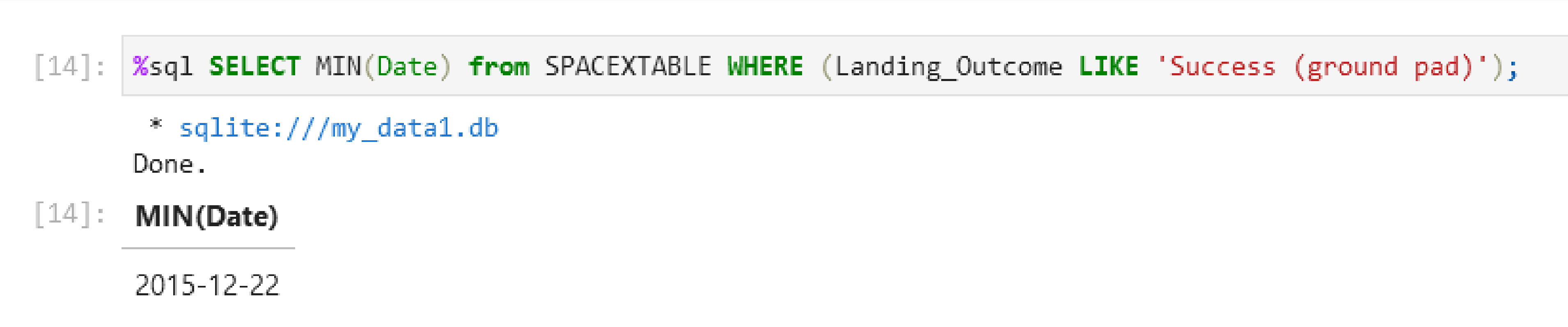
```
[12]: %sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE (Booster_Version LIKE 'F9 v1.1%');
```

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

```
[12]:  AVG(PAYLOAD_MASS__KG_)
           2534.6666666666665
```

First Successful Ground Landing Date

Following, the date of the first successful landing outcome on ground pad is displayed



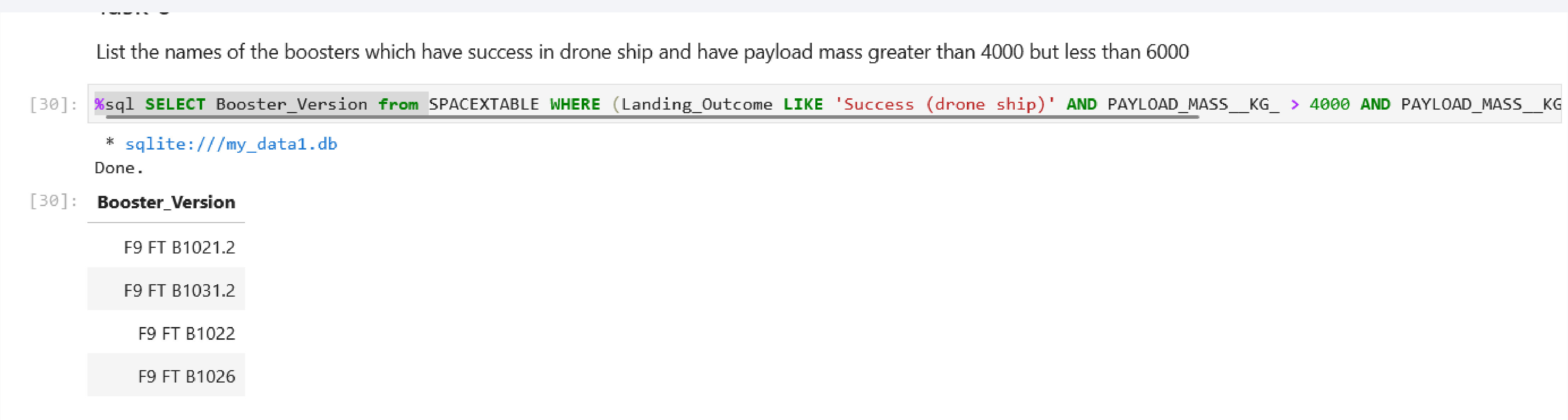
```
[14]: %sql SELECT MIN(Date) from SPACEXTABLE WHERE (Landing_Outcome LIKE 'Success (ground pad)');

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

[14]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Following, the list of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are displayed



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

```
[30]: %sql SELECT Booster_Version from SPACEXTABLE WHERE (Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG
```

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

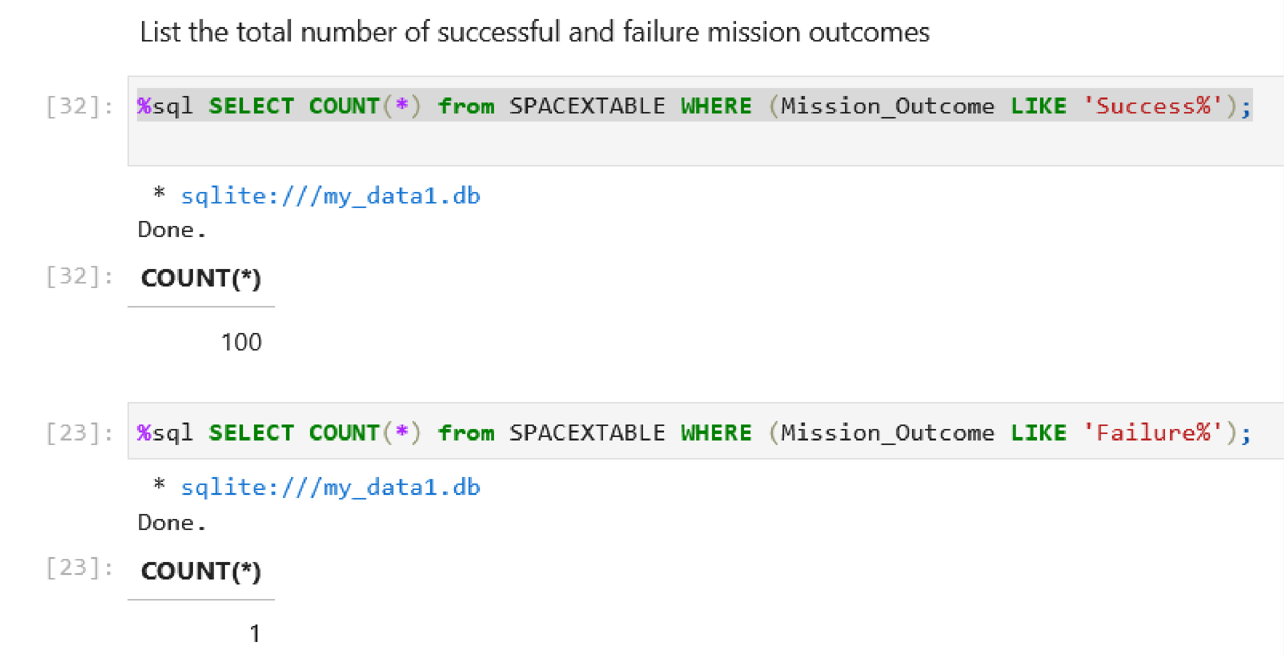
[30]:

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

Total Number of Successful and Failure Mission Outcomes

Successful Missions: 100

Failed Missions: 1



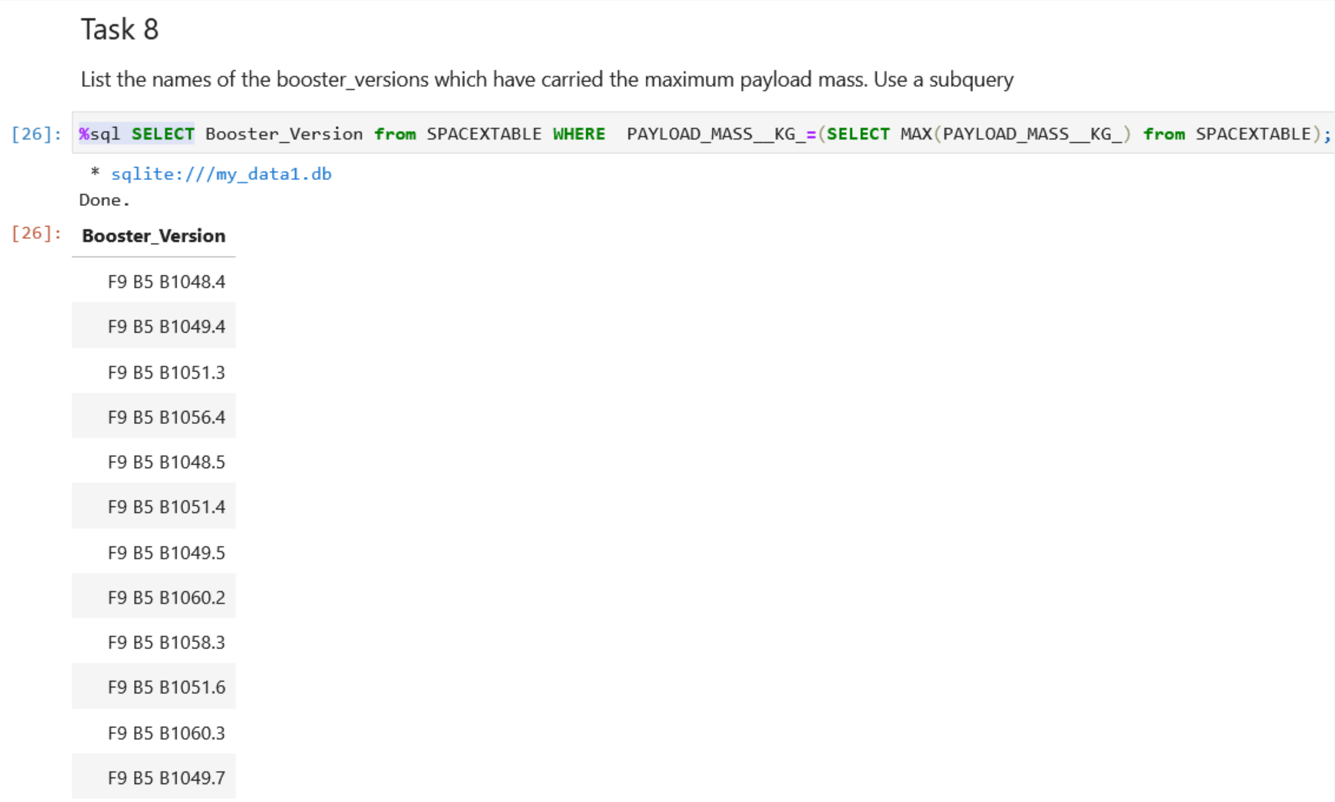
List the total number of successful and failure mission outcomes

```
[32]: %sql SELECT COUNT(*) from SPACEXTABLE WHERE (Mission_Outcome LIKE 'Success%');

 * sqlite:///my_data1.db
Done.
[32]: COUNT(*)

          100
```

```
[23]: %sql SELECT COUNT(*) from SPACEXTABLE WHERE (Mission_Outcome LIKE 'Failure%');

 * sqlite:///my_data1.db
Done.
[23]: COUNT(*)

            1
```

Boosters Carried Maximum Payload

Following, the names of the booster which have carried the maximum payload mass are displayed



Task 8

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

```
[26]: %sql SELECT Booster_Version from SPACEXTABLE WHERE  PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

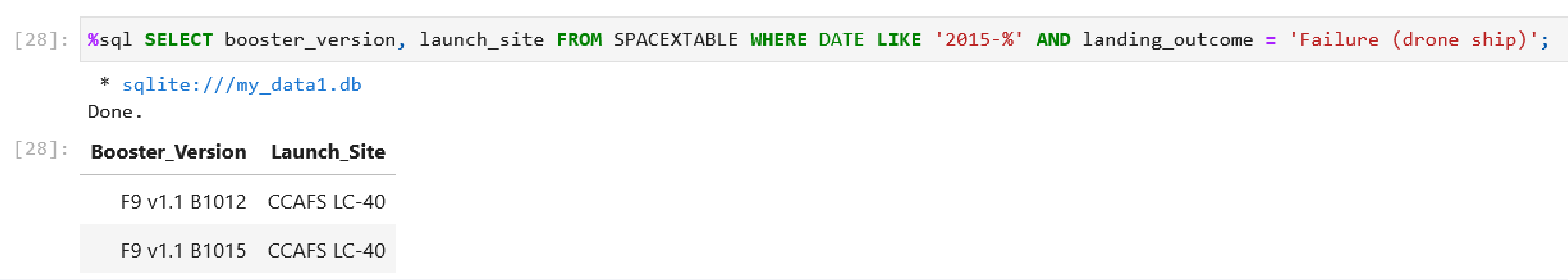
```
 * sqlite:///my_data1.db
Done.
```

[26]:

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

Following, the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are displayed



```
[28]: %sql SELECT booster_version, launch_site FROM SPACEXTABLE WHERE DATE LIKE '2015-%' AND landing_outcome = 'Failure (drone ship)';
```

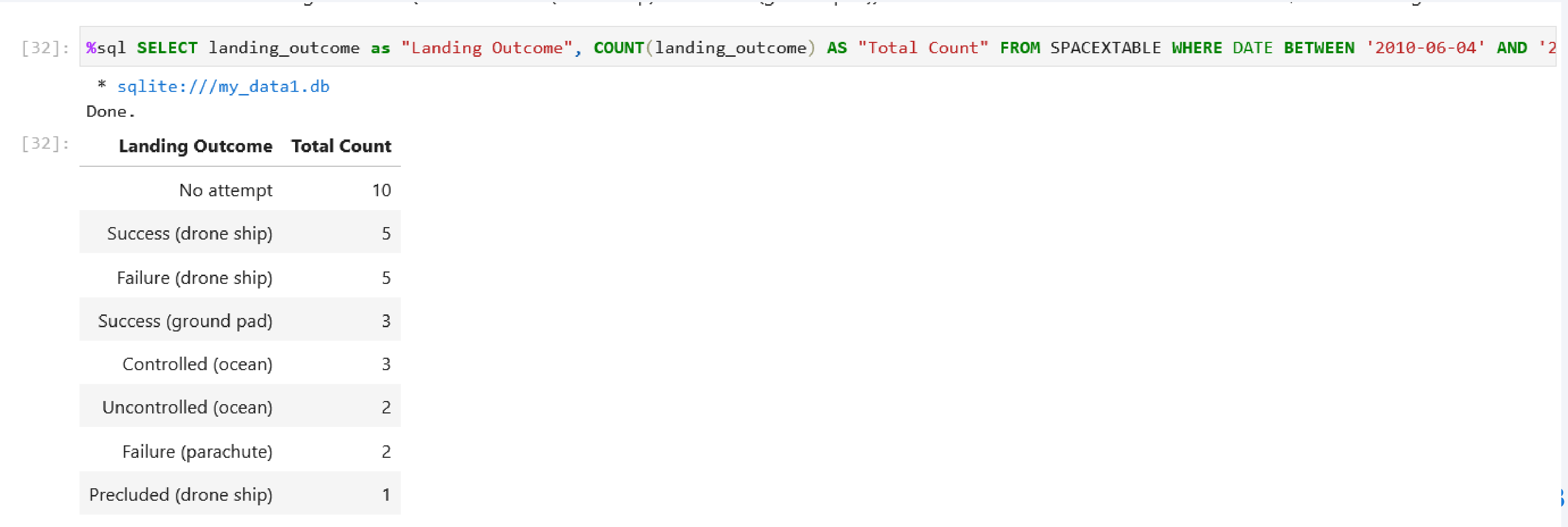
```
 * sqlite:///my_data1.db
Done.
```

[28]:

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

Following, the Rank of counts of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 are displayed in descending order



```
[32]: %sql SELECT landing_outcome as "Landing Outcome", COUNT(landing_outcome) AS "Total Count" FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2
       * sqlite:///my_data1.db
      Done.
```

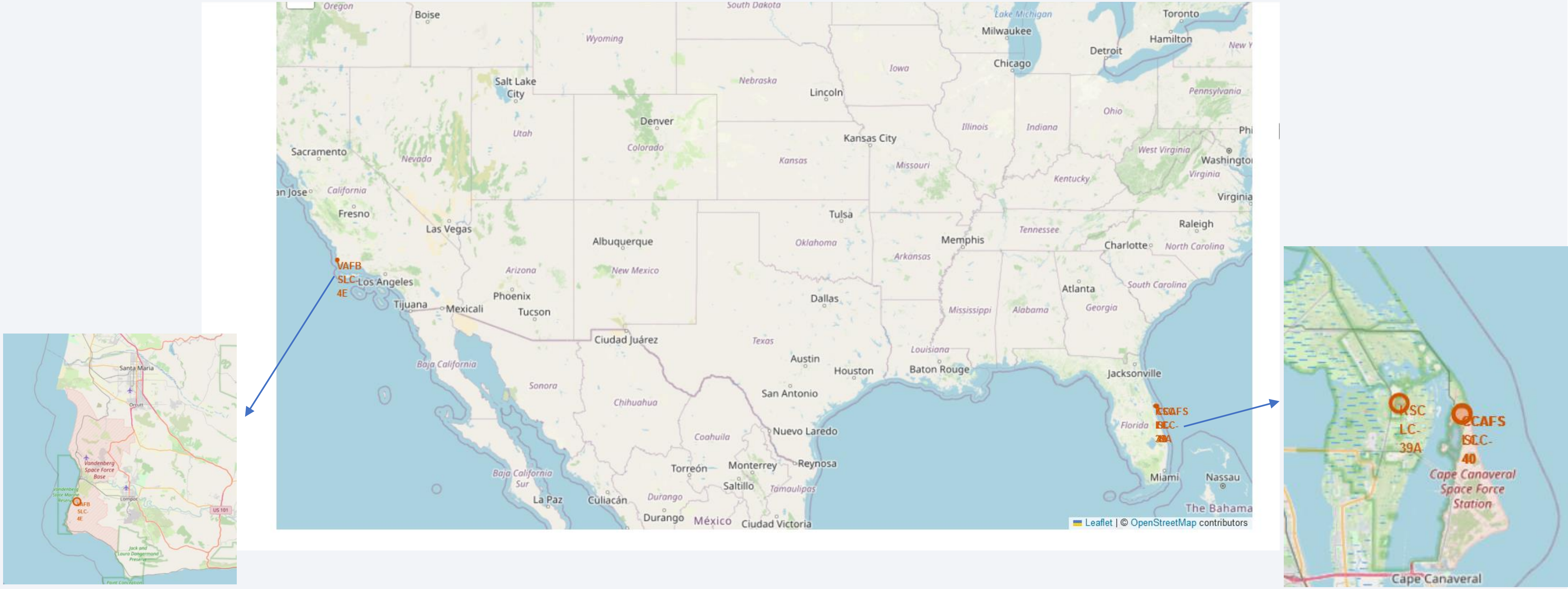
[32]:

Landing Outcome	Total 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

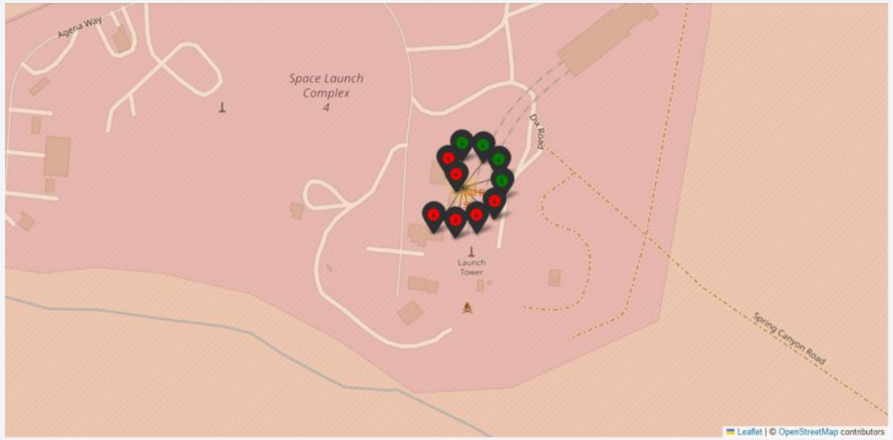
All Launch Sites



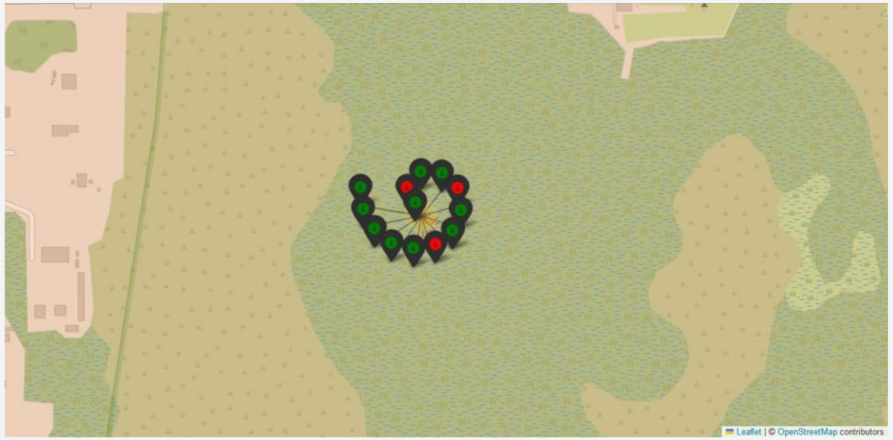
Launch Sites are located at the west and east coast of the US.

Launch sites and launches with color labels

VAFB SLC-4E



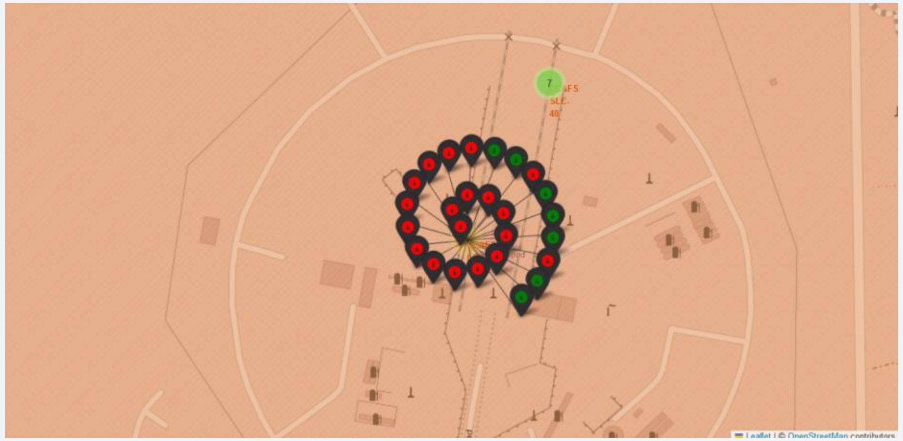
KSC LC-39A



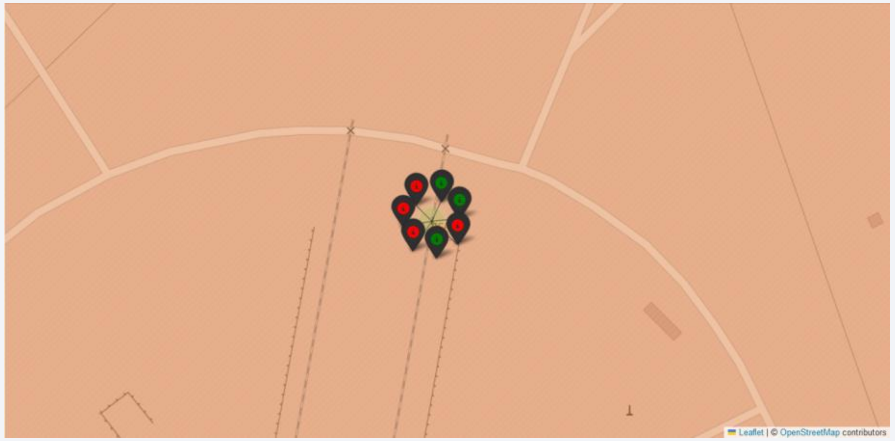
🟩 = Launch Success

🟥 = Launch Failure

CCAFS LC-40

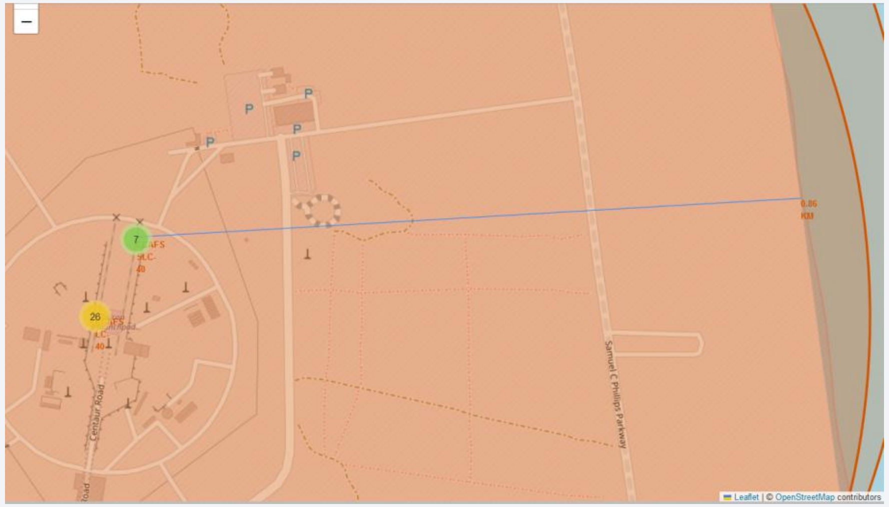


CCAFS SLC-40



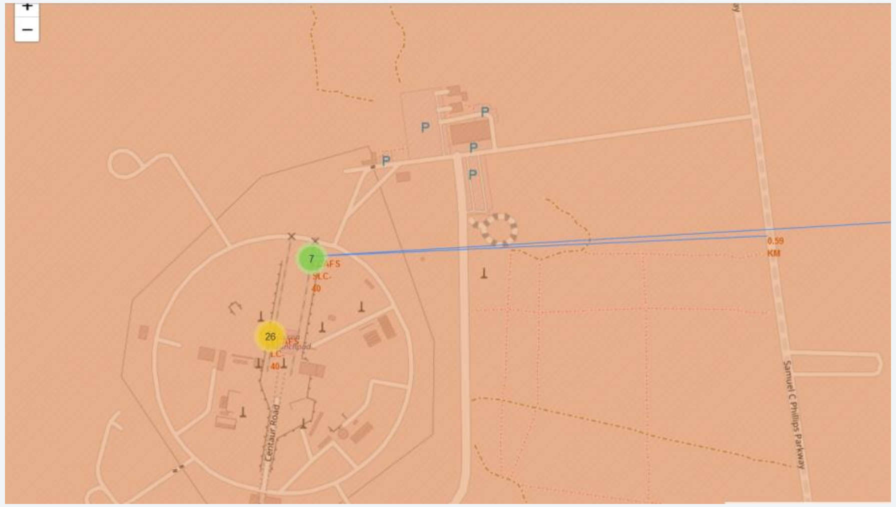
Distances CCAFS SLC-40

Distance to Coast 0.86km

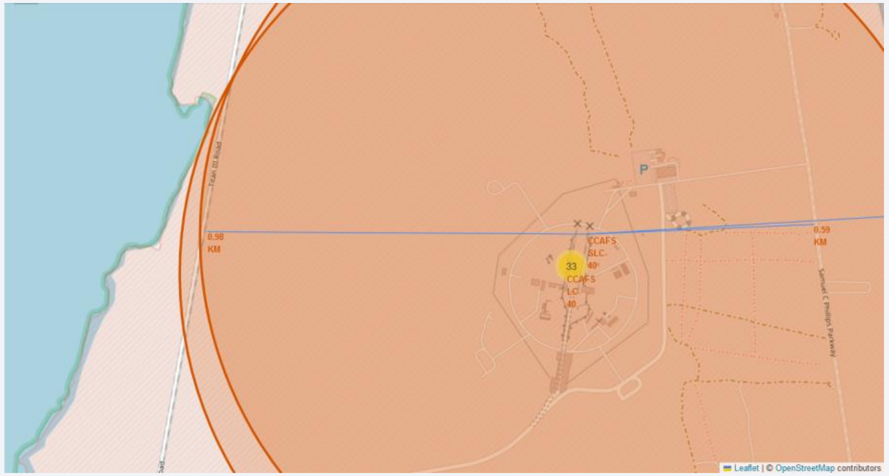


7
26
0.86 KM
Centaur Road
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors

Distance to Highway 0.59km

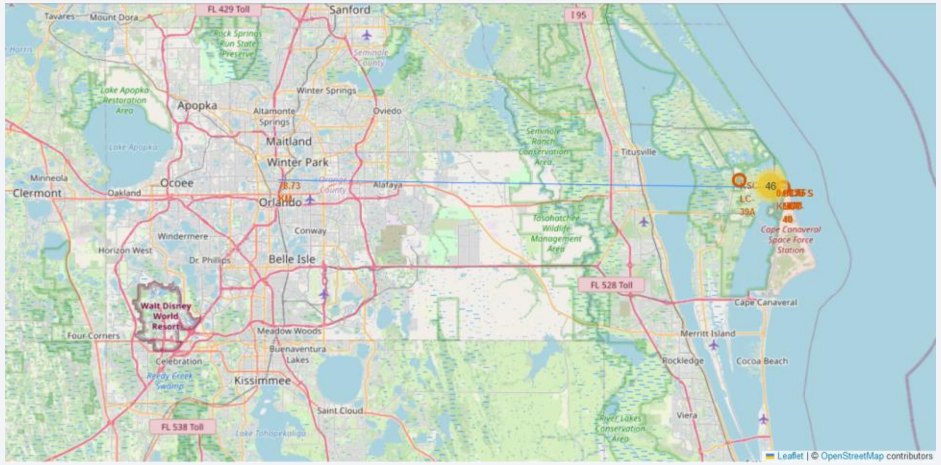


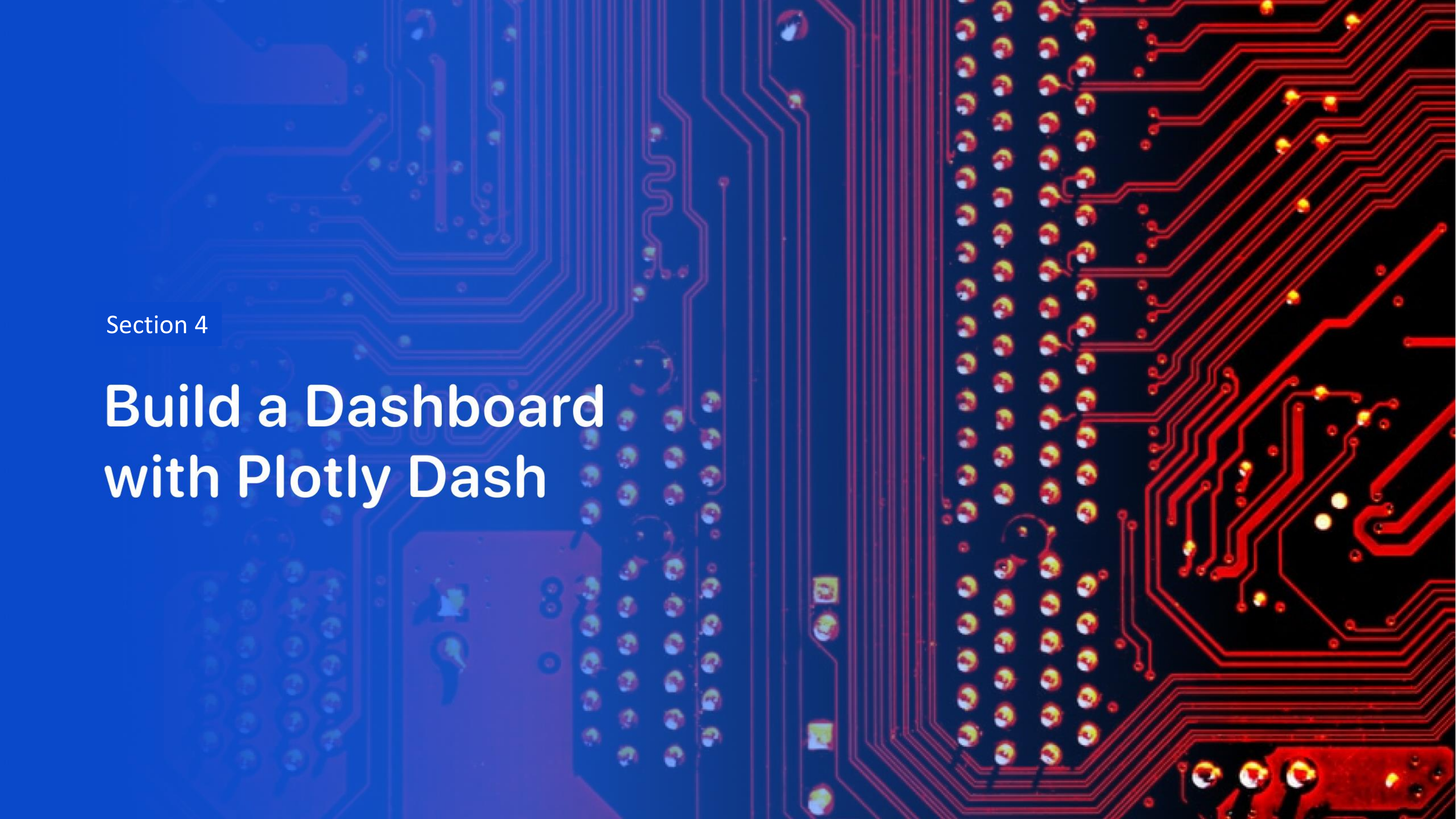
Distance to Railway 0.98km



0.98 KM
33
CCAFS SLC-40
0.59 KM
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors

Distance to Orlando 78.73km

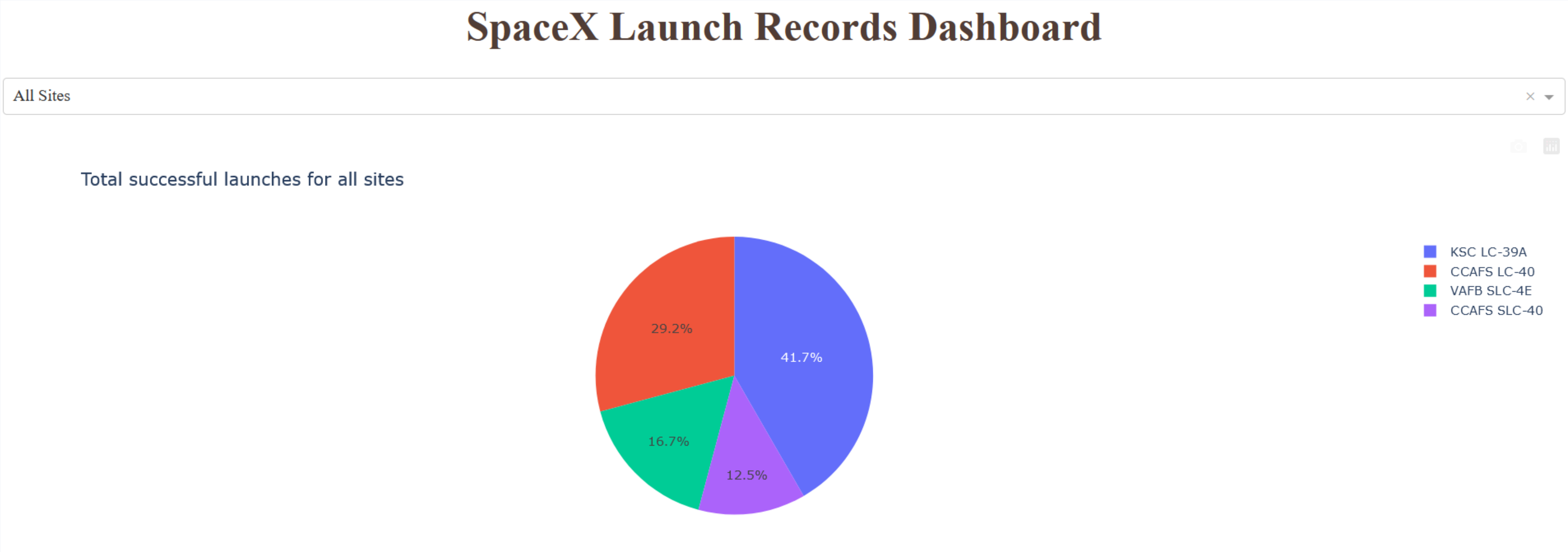




Section 4

Build a Dashboard with Plotly Dash

Success percentage by each launch site



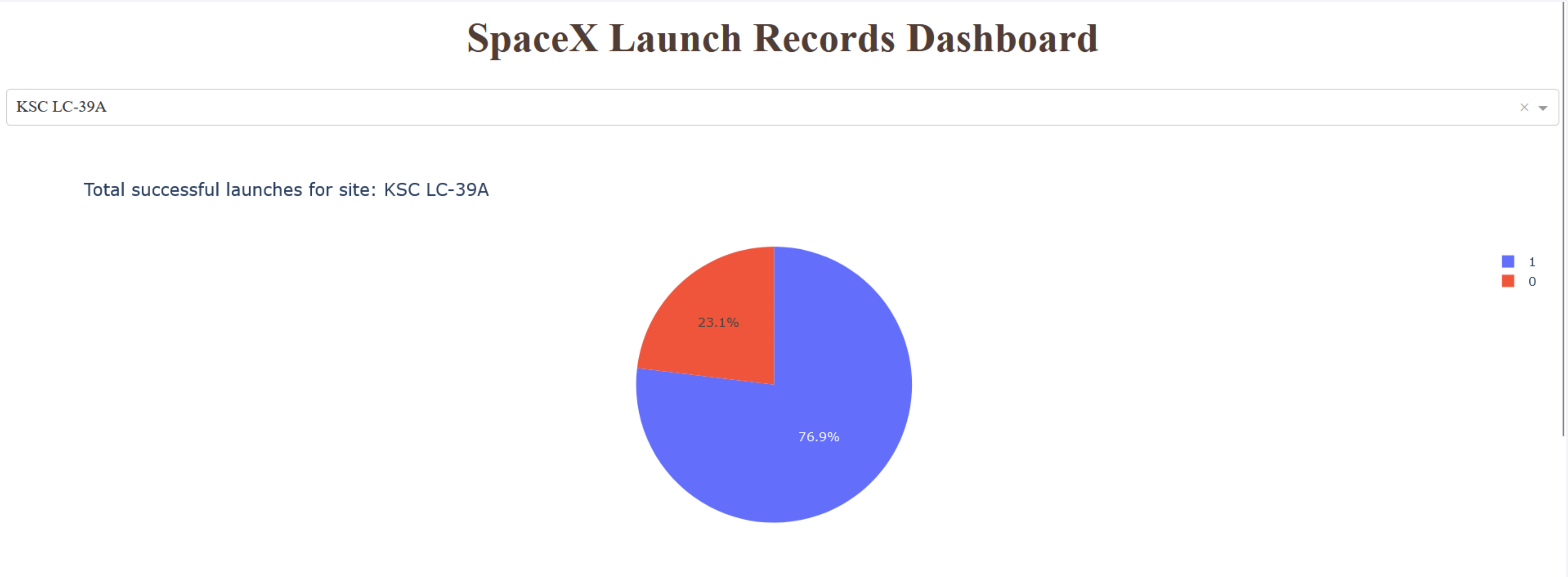
SpaceX Launch Records Dashboard

All Sites

Total successful launches for all sites

KSC LC-39A has the most successful launches

KSC LC-39A Launch Success Rate



SpaceX Launch Records Dashboard

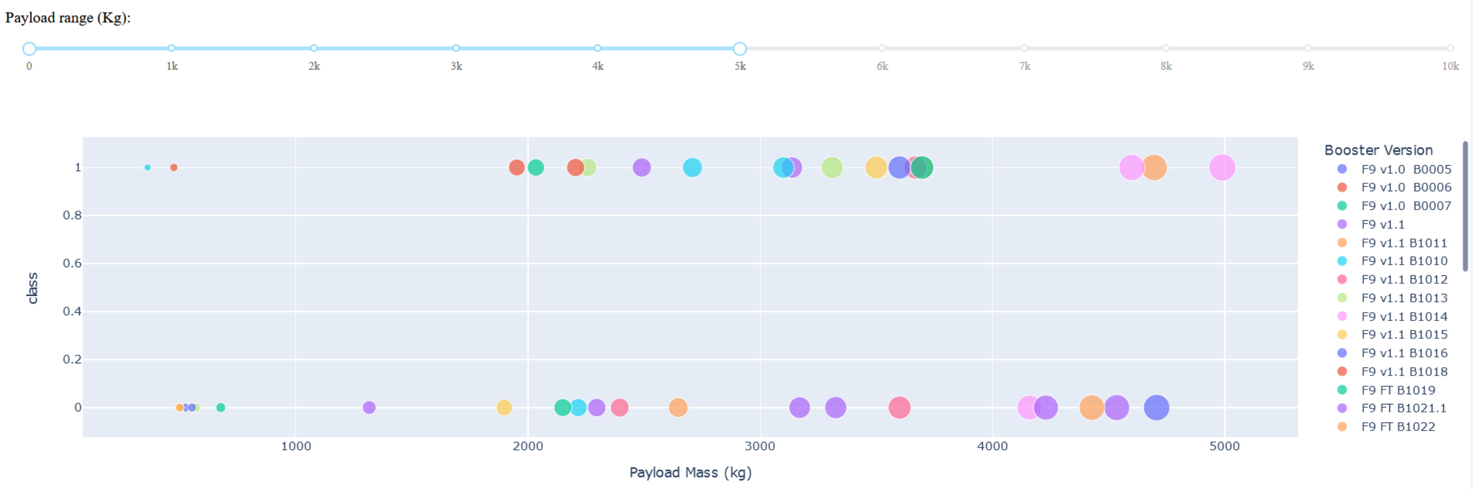
KSC LC-39A

Total successful launches for site: KSC LC-39A

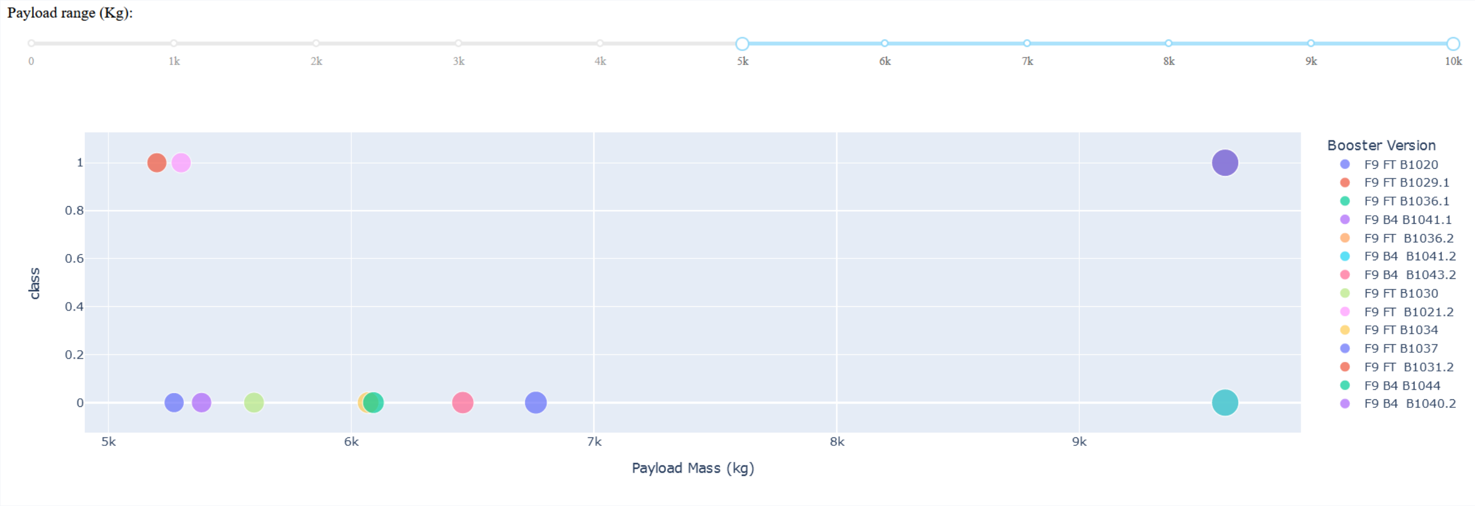
KSC LC-39A has a 76.9% success rate.

<Dashboard Screenshot 3>

Payload: 0 – 5000 kg



Payload: 5000 – 10000 kg

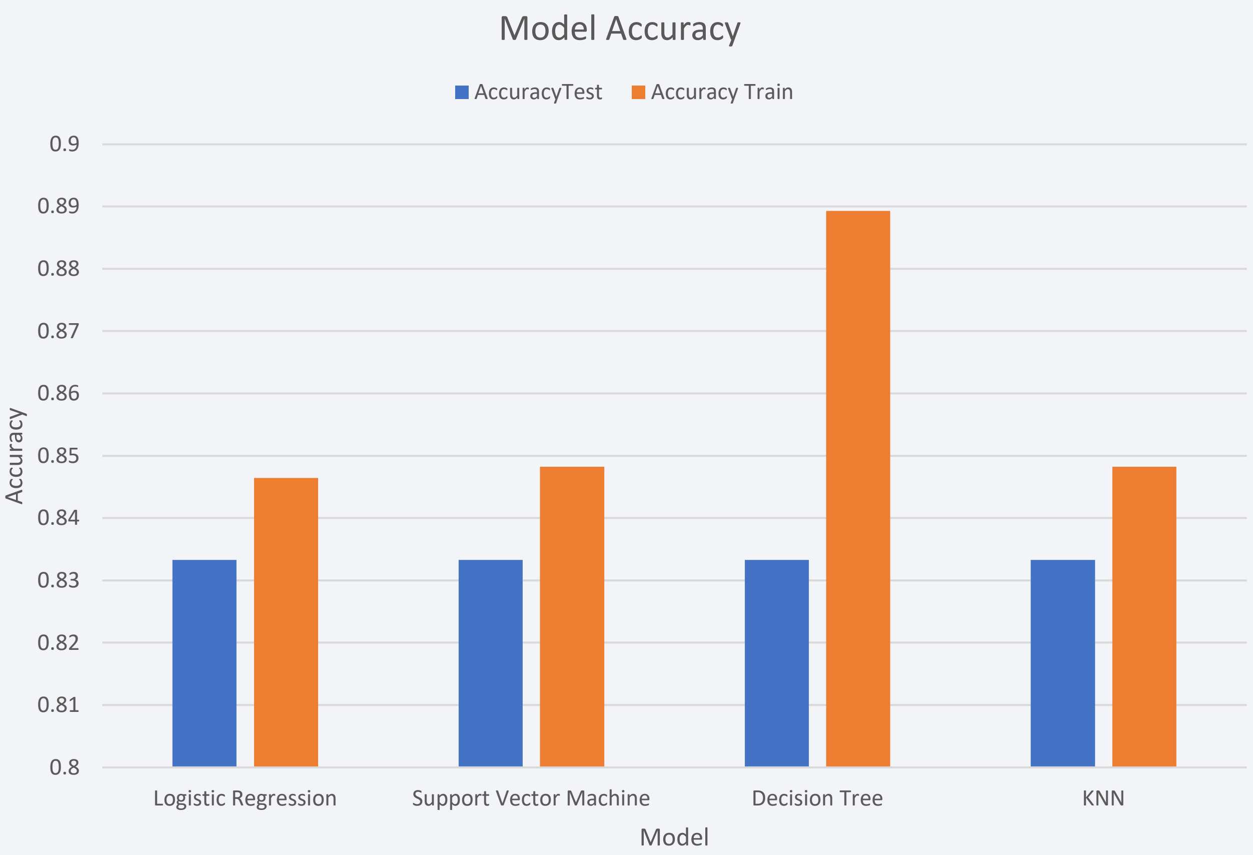


Success rate and number of launches are higher for lower payloads.

Section 5

Predictive Analysis (Classification)

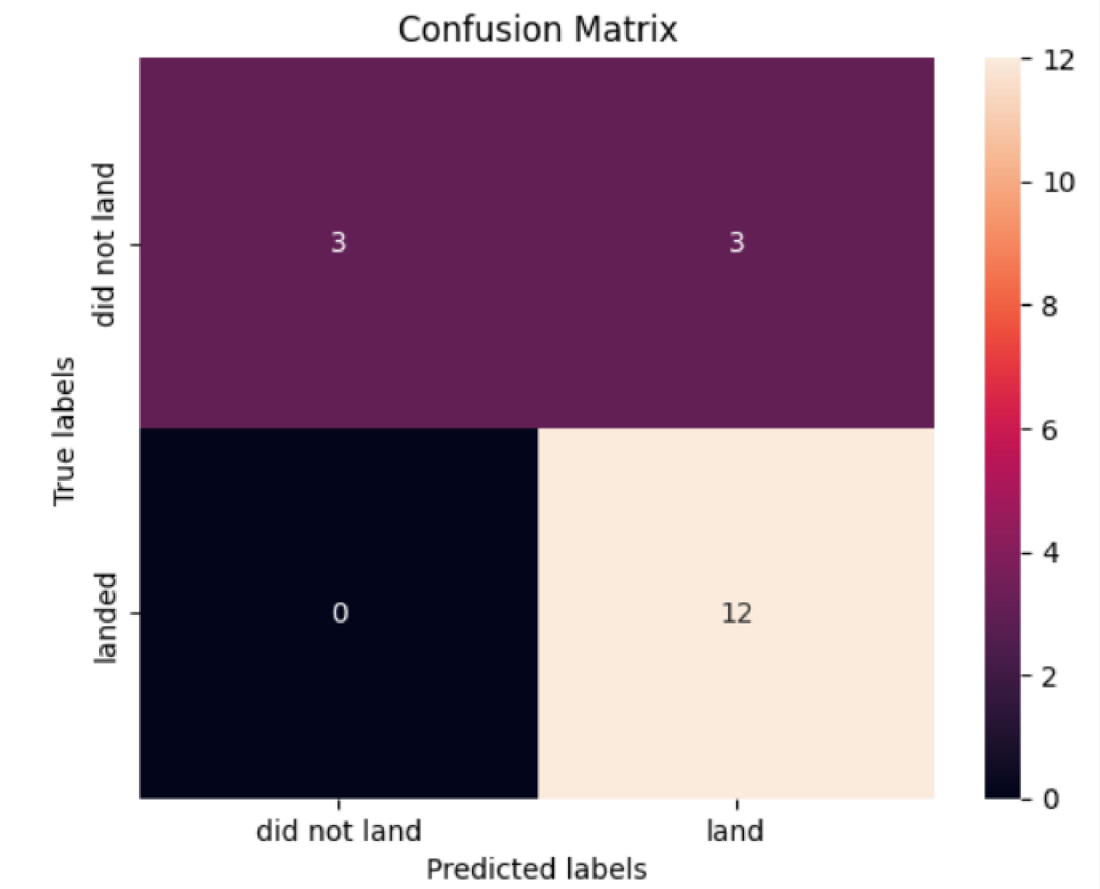
Classification Accuracy



Insights:

- Decision Tree has the highest accuracy on training data (0.889)
- All models have the same accuracy on testing data (0.8333)

Confusion Matrix Decision Tree



True Positives: 12
False Positives: 3

True Negatives: 3
False Negatives: 0

Accuracy: 0.8333

Conclusions

- Launch site location impacts success rate, with certain sites showing higher success rates.
- Heavier payloads correlate with a lower chance of successful landing.
- Some orbits like ES-L1 and SSO have higher success rates, while SO orbits perform the worst.
- Success rate has been on the rise from 2013 to 2020, reflecting improvements in technology and procedures.
- KSC LC-39A has the highest number of successful launches.
- The Decision Tree Classifier is the most accurate model, with an 84% success prediction rate.

Innovative Insights

- Launch outcomes tend to improve with operational experience, as indicated by the positive trend in success rate over time.
- The likelihood of a successful landing might not solely depend on the launch site but also on the technological improvements and operational learning over time.
- Payload mass presents a complex relationship with success rates, suggesting that there may be an optimal range for payload mass that balances the economic and technical aspects of the missions.

Thank you!