IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jorge Israel Frómeta Moya
November 1st, 2024

Outline

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

Executive Summary

Summary of methodologies

Data Collection:

- Accessed data from SpaceX API and Wikipedia including web scraping.

Data Processing:

- Collected, filtered, and cleaned data from separate sources and stored in CSV files.

Exploratory Data Analysis (EDA):

- Generated visualizations to examine launch site performance, orbit success rates, and success trends over time, also perform SQL queries to analyze launch site usage, NASA payload contributions, and landing outcomes.

Interactive Visualization Tools:

- Developed interactive maps and charts using Folium and Dash for dynamic data exploration.

Predictive Modeling:

- Trained multiple models (Logistic Regression, SVM, Decision Tree, KNN) with hyperparameter tuning using GridSearchCV, identified the most accurate model and analyzed key features influencing performance.

Executive Summary

Summary of all results

Launch Site Success:

- Identified certain launch sites with consistently higher success rates, suggesting location impact on mission outcomes.

Orbit and Success Correlation:

- Found specific orbits with higher success rates, providing insights for optimal mission planning.

Improvement Over Time:

- Observed a positive trend in success rates, indicating improvements due to technological advancements.

Predictive Model Performance:

- Best Model: Decision Tree Classifier with 0.9444 accuracy and 0.96 F1 score.
- Top Features: Legs and ReusedCount emerged as key indicators of success, emphasizing the role of reusability in outcomes.

Introduction

Project background and context

- Overview of SpaceX's Success: SpaceX with reusable rockets, particularly through the Falcon 9 model, can reach significantly lower launch costs (estimated at $62 million per launch, compared to $165 million from other providers). A large portion of these savings comes from reusing the first stage of the rocket.

Problems you want to find answers

Primary question:

- Can we accurately predict the likelihood of the first-stage landing success, given key mission parameters?

Secondary question:

- What factors (payload, orbit type, etc.) most strongly influence the success of first-stage landings?

Section 1

Methodology

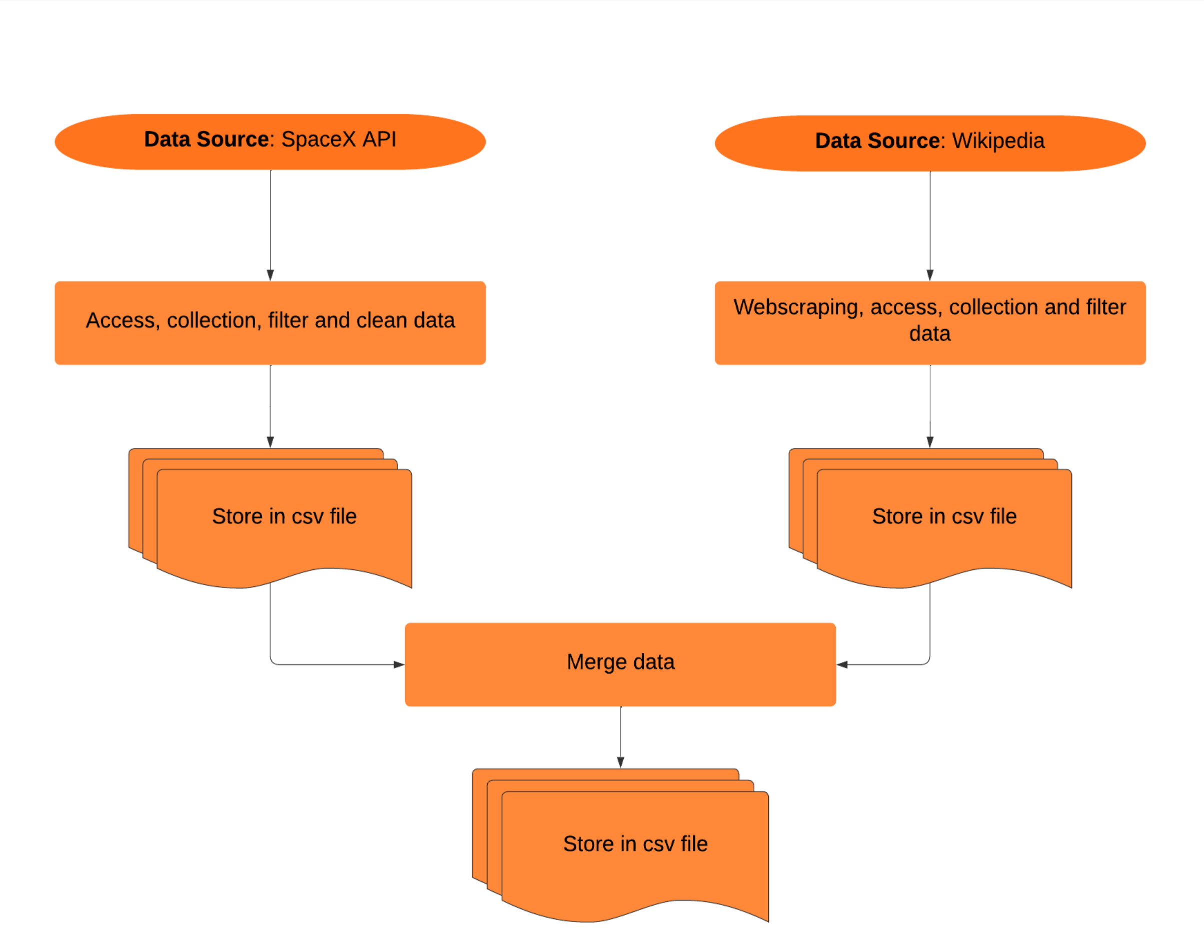
Methodology

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

- Data access to sources: SpaceX API and Wikipedia.
- Collection, filter and clean data from separated sources (including webscraping).
- Store in separated csv files and then merged on a single csv file.

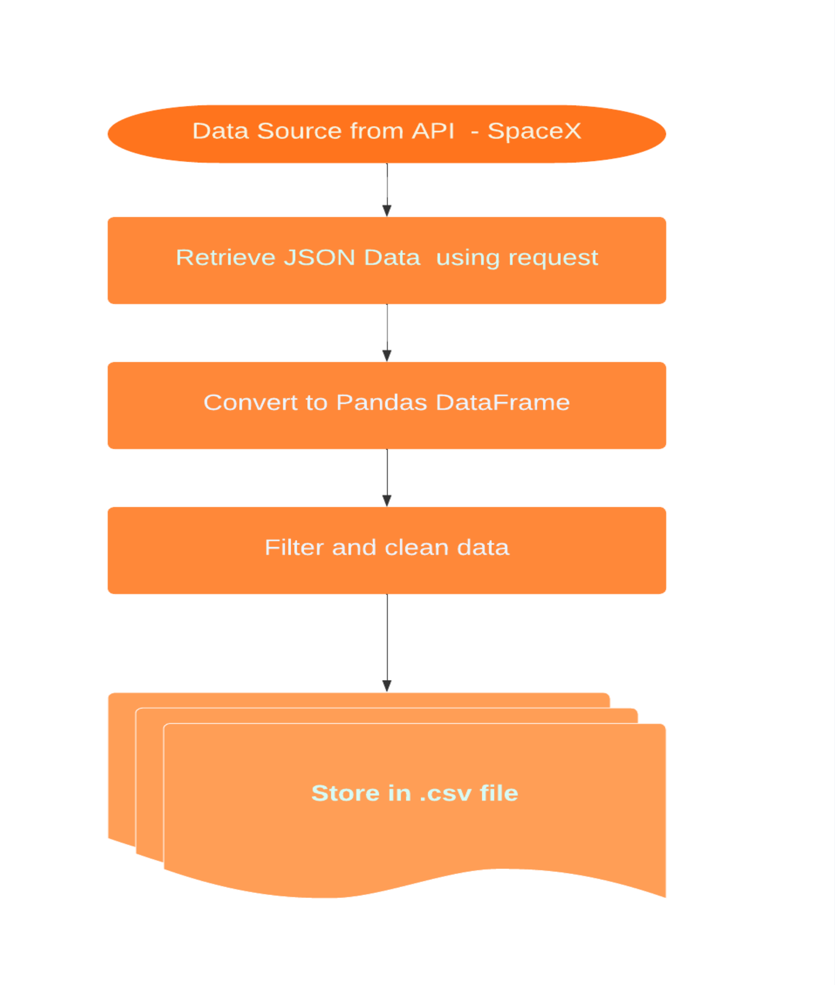

Data Source: SpaceX API
Data Source: Wikipedia
Access, collection, filter and clean data
Webscraping, access, collection and filter data
Store in csv file
Store in csv file
Merge data
Store in csv file

Data Collection – SpaceX API

1. Data collected from SpaceX API (https://api.spacexdata.com/v4/launches/past) using Python's requests module.
2. Use Python requests module to access data.
3. Convert JSON data to Pandas DataFrame for processing.
4. Filter relevant columns (e.g., payload, launch date, orbit type), impute missing values.
5. Store the cleaned data as a CSV file for further analysis.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb



Data Source from API - SpaceX

Retrieve JSON Data using request

Convert to Pandas DataFrame

Filter and clean data

Store in .csv file

Data Collection - Scraping

1. Data access from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches) using Python's requests module.

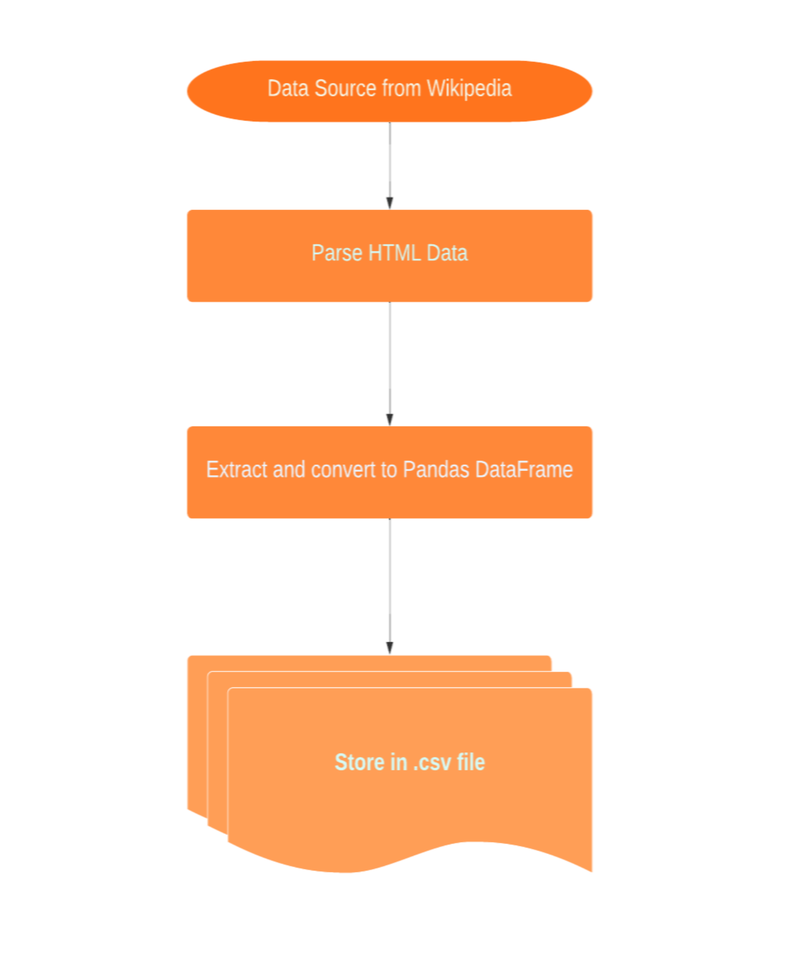
2. Parse HTML with BeautifulSoup

3. Extract relevant tables and convert to Pandas DataFrame.

4. Store the data as a CSV file for further analysis.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/jupyter-labs-webscraping.ipynb

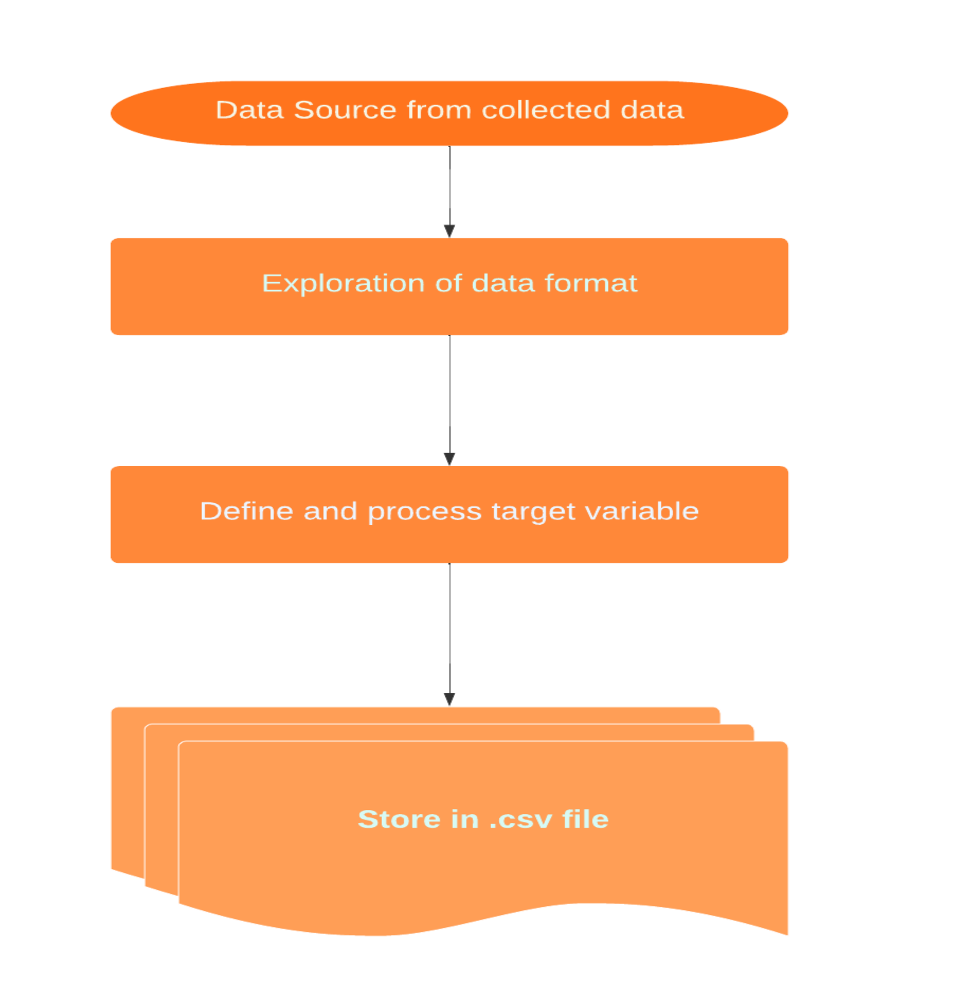

Data Source from Wikipedia
Parse HTML Data
Extract and convert to Pandas DataFrame
Store in .csv file

Data Wrangling

1. Data access from collected data.
2. Exploration of data format using pandas module.
3. Define and process the target variable as a binary variable for further analysis.
4. Store the data as a CSV file for further analysis.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb


Data Source from collected data
Exploration of data format
Define and process target variable
Store in .csv file

EDA with Data Visualization

Charts used and their Purpose:

- **Flight number vs Launch site and succes or failed outcome (categorical plot)**: to identify patterns of success across launch sites and analyze if certain sites have consistently higher success rates

- **Payload mass vs Launch site and succes or failed outcome (categorical plot)**: to determine if heavier or lighter payloads have a higher success rate at specific launch sites.

- **Orbit vs Success rate (bar plot)**: to assess whether certain orbits are associated with higher or lower success rates, informing planning for mission type and destination.

- **Flight number vs Orbit and succes or failed outcome (categorical plot)**: to examines if particular orbit types tend to have more successful landings, especially for rockets with higher flight numbers.

EDA with Data Visualization

Charts used and their purpose:

- **Payload mass vs Orbit and succes or failed outcome (categorical plot)**: to help identify any correlations between payload mass and success rates for specific orbits, guiding payload planning.
- **Average success rate over time (line plot)**: to track overall trends in SpaceX's success rate over time, showing improvements and identifying periods with significant advancements.

Data preparation steps:

- **Dummy variables**: Created for categorical columns to enable model compatibility.
- **Data type casting**: Numeric columns cast to float64 for consistency in numerical operations.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

SQL queries used and their Purpose:

- **Unique Launch Sites**: Retrieved distinct launch site names used in space missions to understand site distribution.
- **Launch Sites Beginning with 'CCA'**: Queried the first 5 records of launch sites starting with "CCA" for location-specific analysis.
- **Total Payload Mass for NASA (CRS)**: Calculated the total payload mass for missions launched by NASA (CRS), focusing on NASA's payload contributions.

EDA with SQL

SQL queries used and Their Purpose:

- **Date of First Successful Ground Pad Landing**: Identified the date of the first successful landing outcome on a ground pad to mark milestones in reusability.
- **Boosters with Successful Drone Ship Landings (Payload 4000-6000)**: Listed booster names that successfully landed on drone ships with payloads between 4000 and 6000 kg.
- **Count of Successful and Failed Missions**: Summed up the total number of successful and failed mission outcomes, giving an overview of mission reliability.
- **Boosters with Maximum Payload**: Retrieved booster versions that carried the maximum payload mass, identifying high-capacity boosters.

EDA with SQL

SQL queries used and Their Purpose:

- **Failed Drone Ship Landings in 2015**: Displayed records with month names, landing outcomes (failures on drone ships), booster versions, and launch sites specifically for 2015.
- **Ranked Landing Outcomes (2010-06-04 to 2017-03-20)**: Ranked landing outcome counts (e.g., Failure on drone ship, Success on ground pad) in descending order over a specific date range for outcome analysis.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Maps used and their purpose:

- **Marking launch sites**: Plotted all SpaceX launch sites on an interactive Folium Map with Markers and Circle objects, providing a spatial overview of launch locations.
- **Visualizing launch outcomes**: Marked each launch at its respective site on an interactive Folium Map with MarkerCluster object, color-coded by outcome (success or failure). Enabled quick visual assessment of site performance and success rates.
- **Distance calculation to nearest coast**: Calculated and displayed the distance from each launch site to the nearest coastline on an interactive Folium Map with Markers and Polyline objects. Important for understanding proximity to water, which is relevant for safety and recovery planning.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Dashboard elements added:

- **Total and single site successful rate for launches (pie chart)**: Displays the overall count or rate of successful and failed launches across all launch sites or a selected single site. Provides a clear overview of launch site performance, helping identify high-performing locations and areas for improvement.
- **Payload mass vs. Launch success or fail and booster version (scatter chart):** Plots payload mass against launch success or fail and booster version, showing the correlation between the variables. Helps analyze the impact of payload mass on launch success, guiding decisions on payload limits and mission planning.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Key Steps in Model Development Process

1. Data Preprocessing:

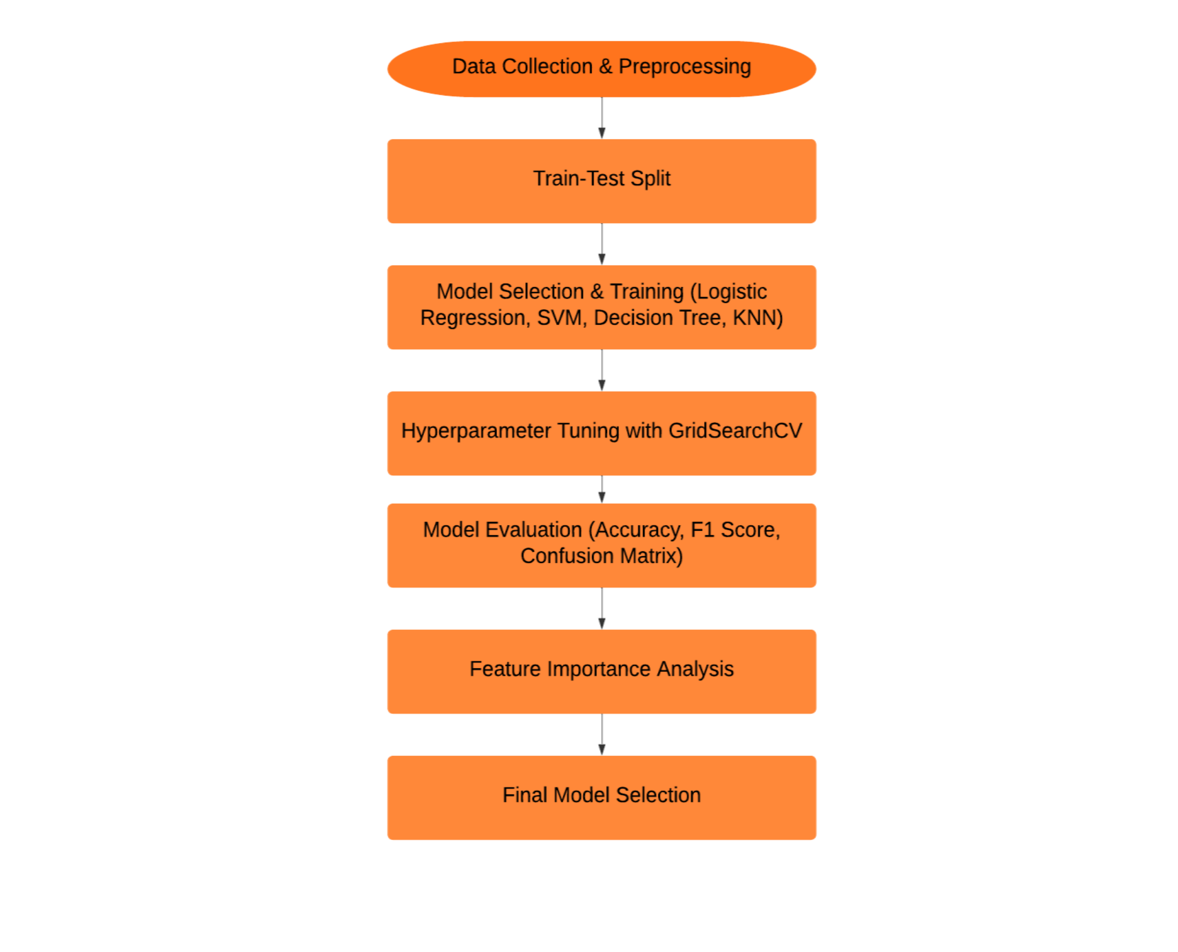
- Loaded and standardized dataset.
- Split data into training and testing sets.

2. Model Selection and Initial Training:

- Trained multiple classification models: Logistic Regression, SVM, Decision Tree, KNN.
- Used GridSearchCV for hyperparameter tuning with cross-validation.

3. Model Evaluation:

- Evaluated each model's performance on test data using accuracy and F1 score.
- Plotted confusion matrix for each model to assess prediction quality.



Data Collection & Preprocessing
Train-Test Split
Model Selection & Training (Logistic Regression, SVM, Decision Tree, KNN)
Hyperparameter Tuning with GridSearchCV
Model Evaluation (Accuracy, F1 Score, Confusion Matrix)
Feature Importance Analysis
Final Model Selection

Predictive Analysis (Classification)

Key Steps in Model Development Process

4. Feature Importance Analysis:

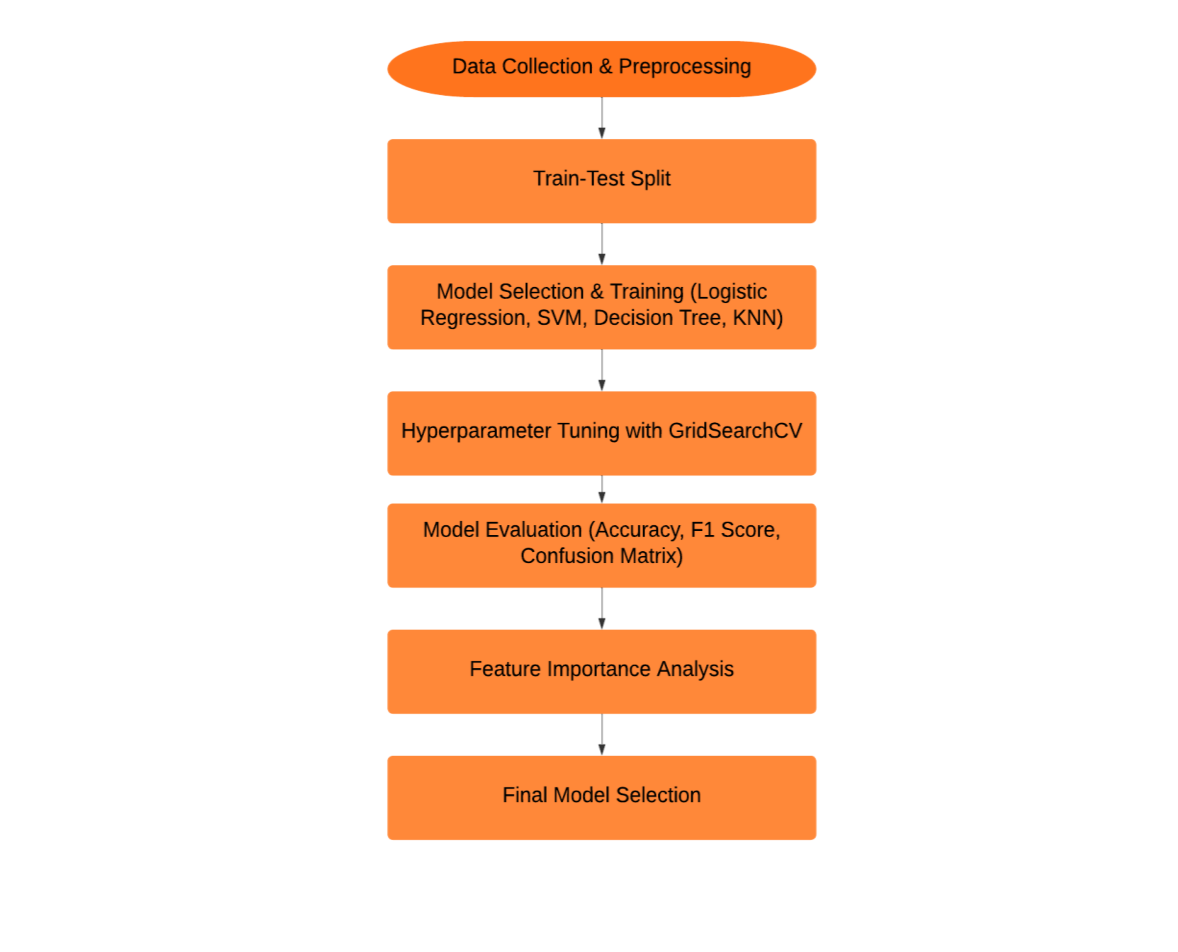
- Identified and visualized top features affecting model performance.
- Compared feature importance for Logistic Regression and Decision Tree models.

5. Model Comparison and Selection:

- Compared models based on cross-validation score, test accuracy, and F1 score.
- Selected best-performing model for deployment.

Code and process details available in GitHub Notebook:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb


Data Collection & Preprocessing

Train-Test Split

Model Selection & Training (Logistic Regression, SVM, Decision Tree, KNN)

Hyperparameter Tuning with GridSearchCV

Model Evaluation (Accuracy, F1 Score, Confusion Matrix)

Feature Importance Analysis

Final Model Selection

Results

Exploratory data analysis results

Visualization for Key Patterns:

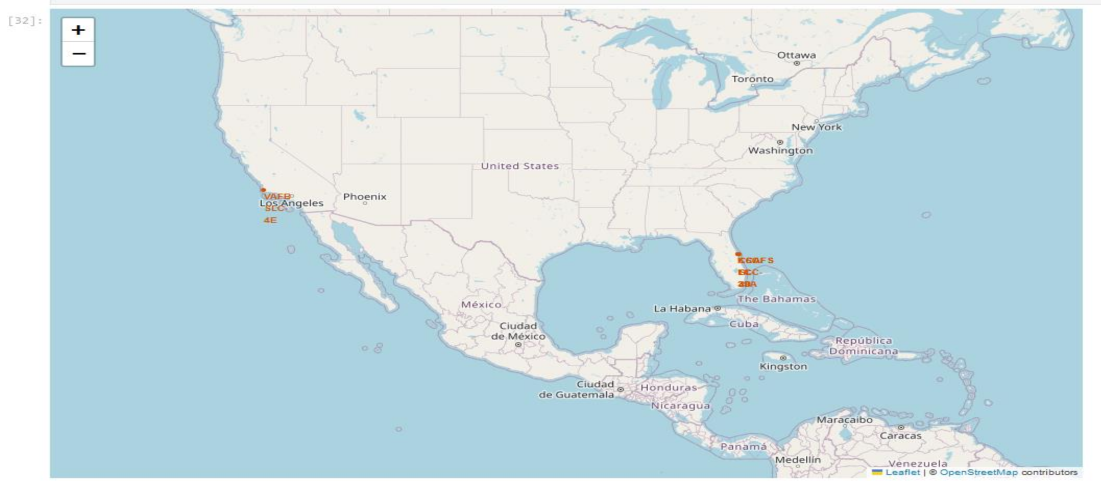
- Launch Sites and Success: Compared launch sites and payloads to see which sites consistently perform better.
- Orbit Success Rates: Identified orbits with higher success rates to inform mission planning.
- Success Over Time: Tracked trends in SpaceX's success rates to spot improvements.

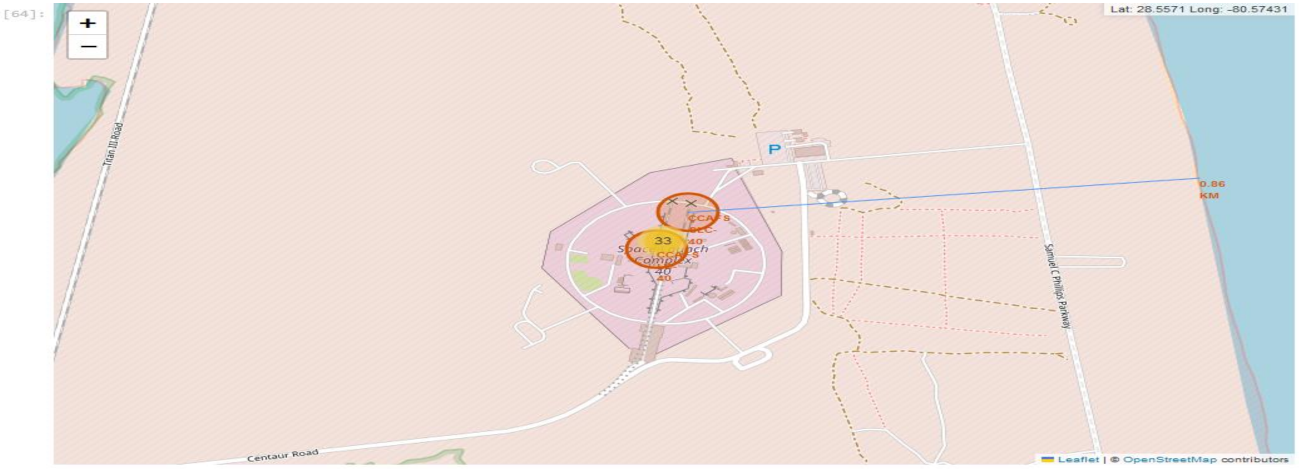
SQL Queries for Targeted Insights:

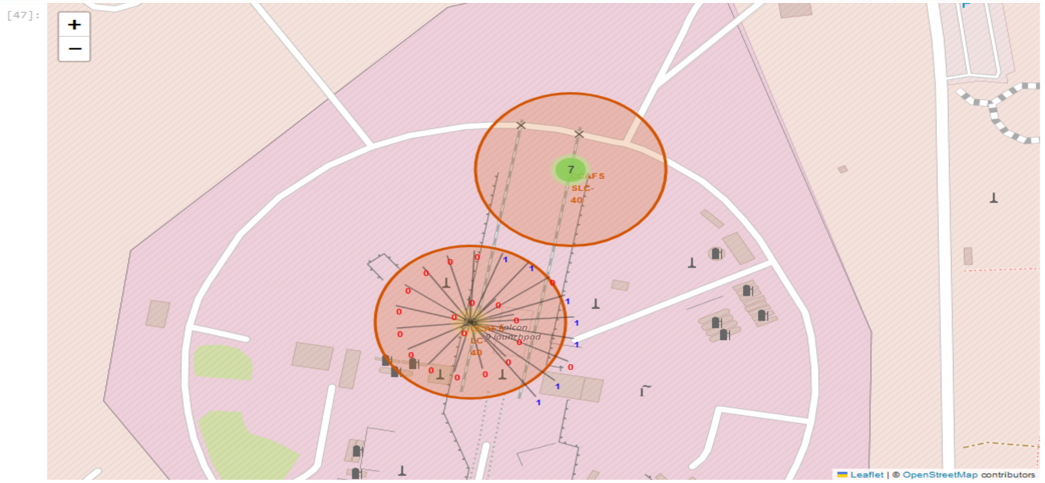
- Site and Payload Distribution: Mapped out launch site use and NASA's payload contributions.
- Landing and Outcome Patterns: Analyzed milestones in reusability, success/failure counts, and patterns in failed drone landings.

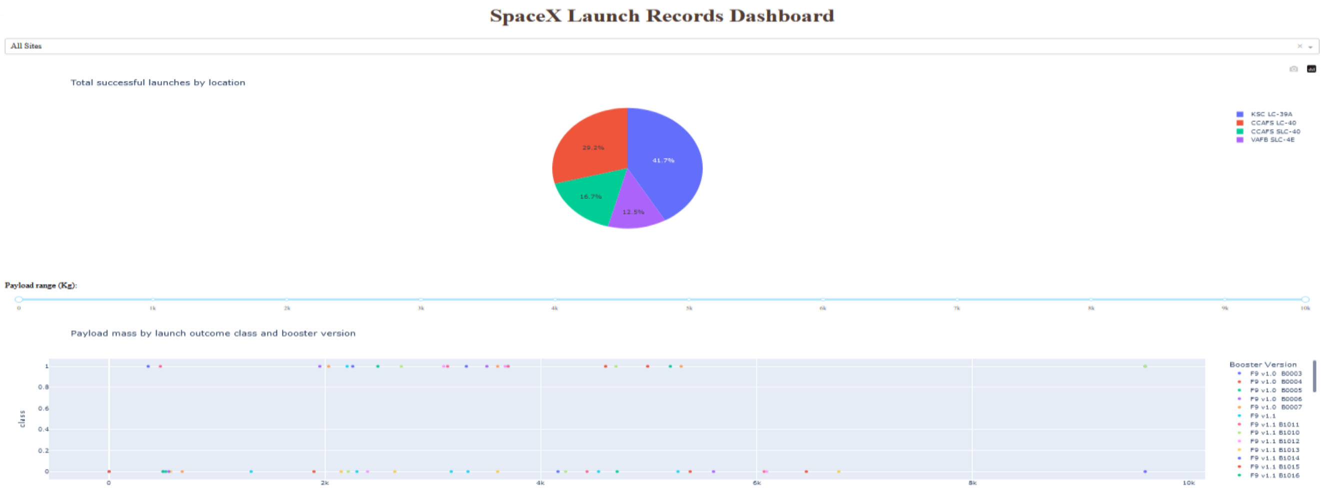
Results

- **Interactive analytics demo in screenshots**




[64]:
+
−
Lat: 28.5571 Long: -80.57431
P
0.86 KM
CCAFS SLC 40
33
Space Launch Complex 40
Samuel C Phillips Parkway
Centaur Road
Leaflet | © OpenStreetMap contributors





SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

Results

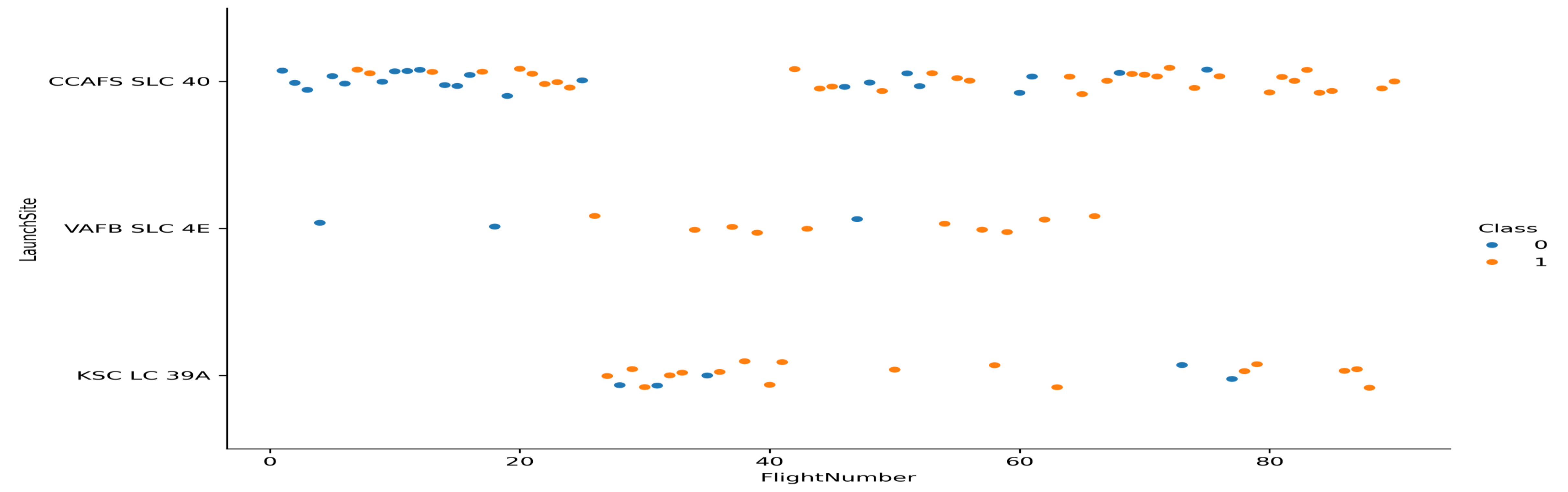
Predictive analysis results

- Trained multiple classification models: Logistic Regression, SVM, Decision Tree, KNN using GridSearchCV for hyperparameter tuning.
- Decision Tree Classifier was the best model with the highest classification accuracy (0.9444) and F1 score of 0.96.
- Identified and visualized top features affecting model performance for Logistic Regression and Decision Tree models.
- Legs and ReusedCount appears to be the most importance features among others.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

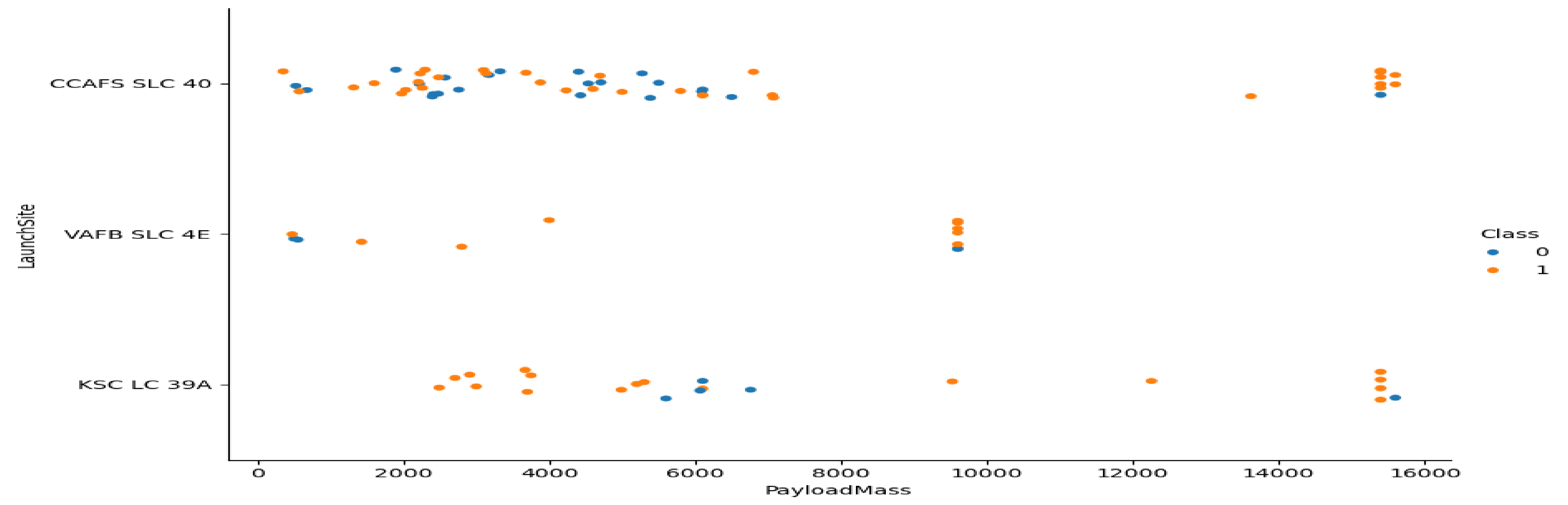


- The plot suggests that CCAFS SLC 40 and KSC LC 39A are the primary sites for launches, while VAFB SLC 4E is used less frequently.

- Despite fewer launches, VAFB SLC 4E shows a strong trend of success, indicating that this site might be used selectively.

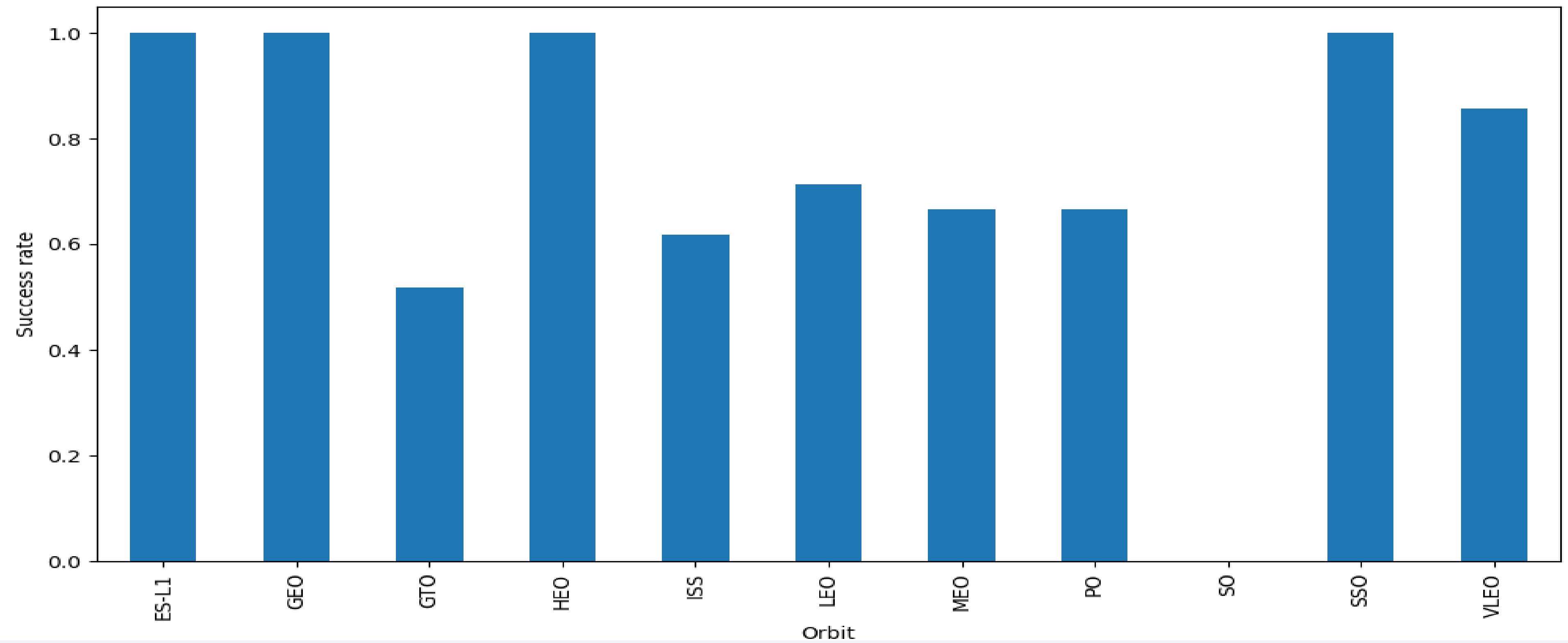
- Failures are more common in the early flight numbers, especially for launches from CCAFS SLC 40. This trend suggests a learning curve

Payload vs. Launch Site



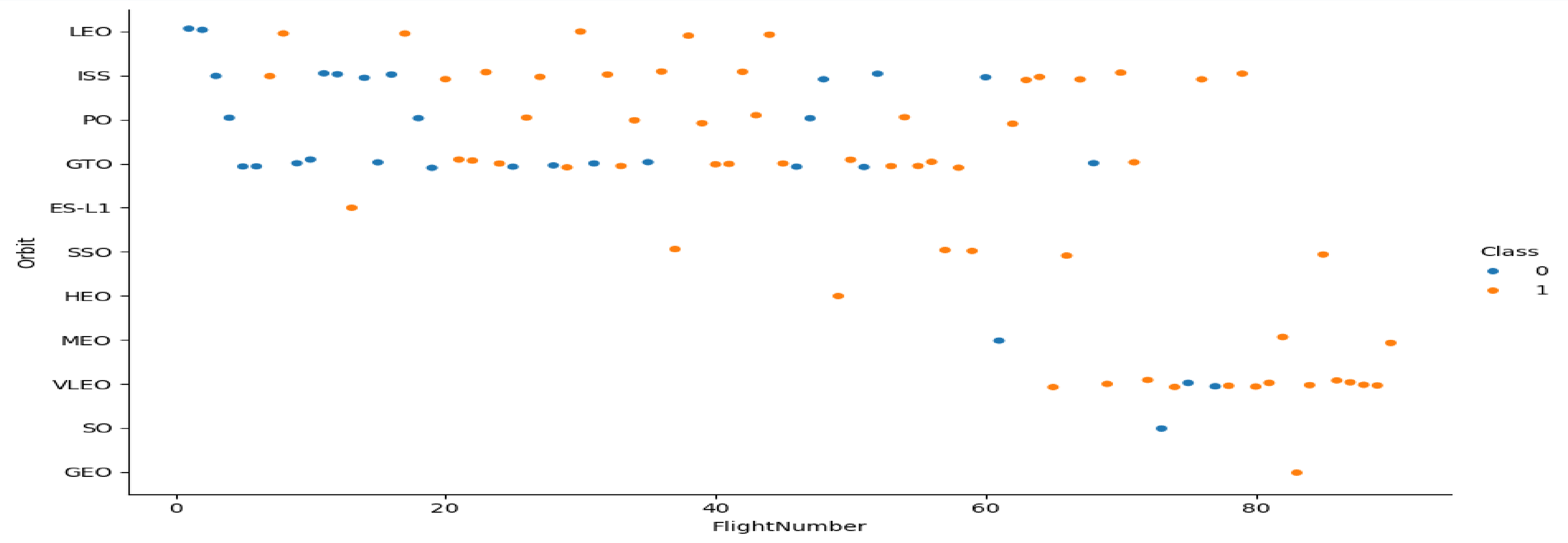
- This chart highlights that heavier payloads (above 8000 Kg.) tend to have a higher success rate, possibly due to added caution, better planning, or more reliable rocket configurations.
- There is not clear additional patterns remarking difference across launch sites.

Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO are the most successful orbits (100% or near).
- GTO presents the poorest success rate (under 60%).

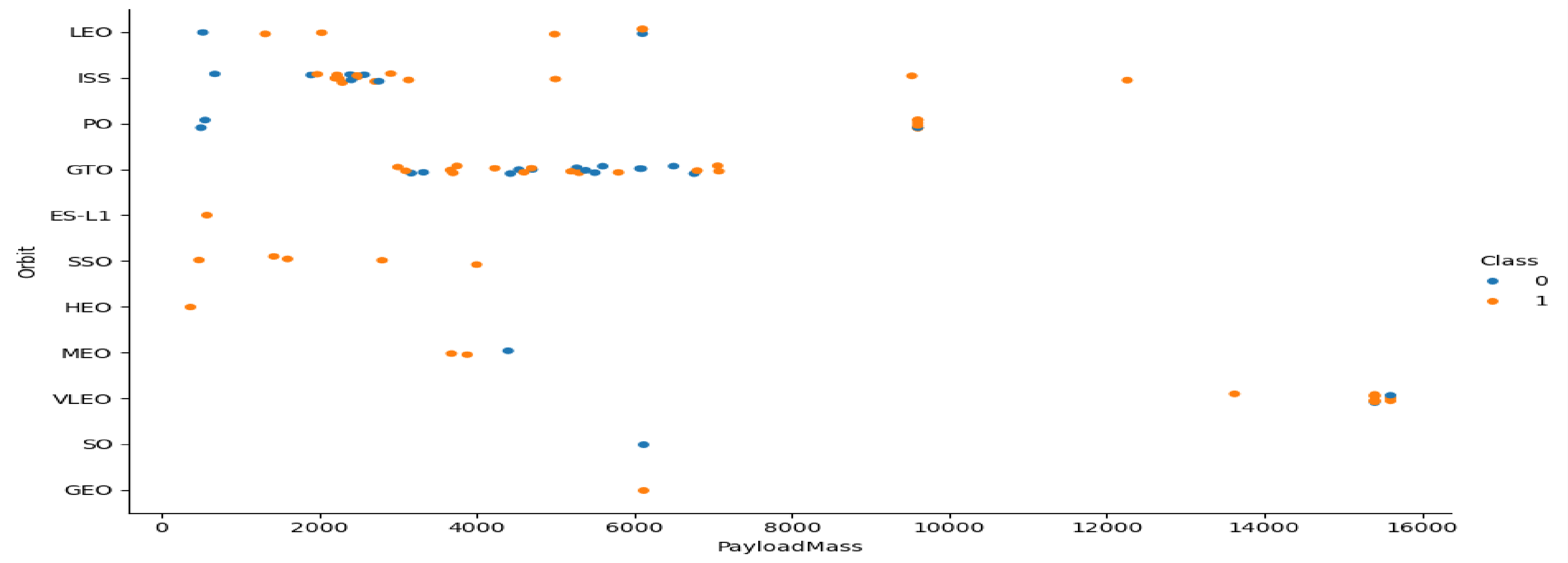
Flight Number vs. Orbit Type



- ISS and GTO present more flights than others with a mix of success and failures, specially in GTO.

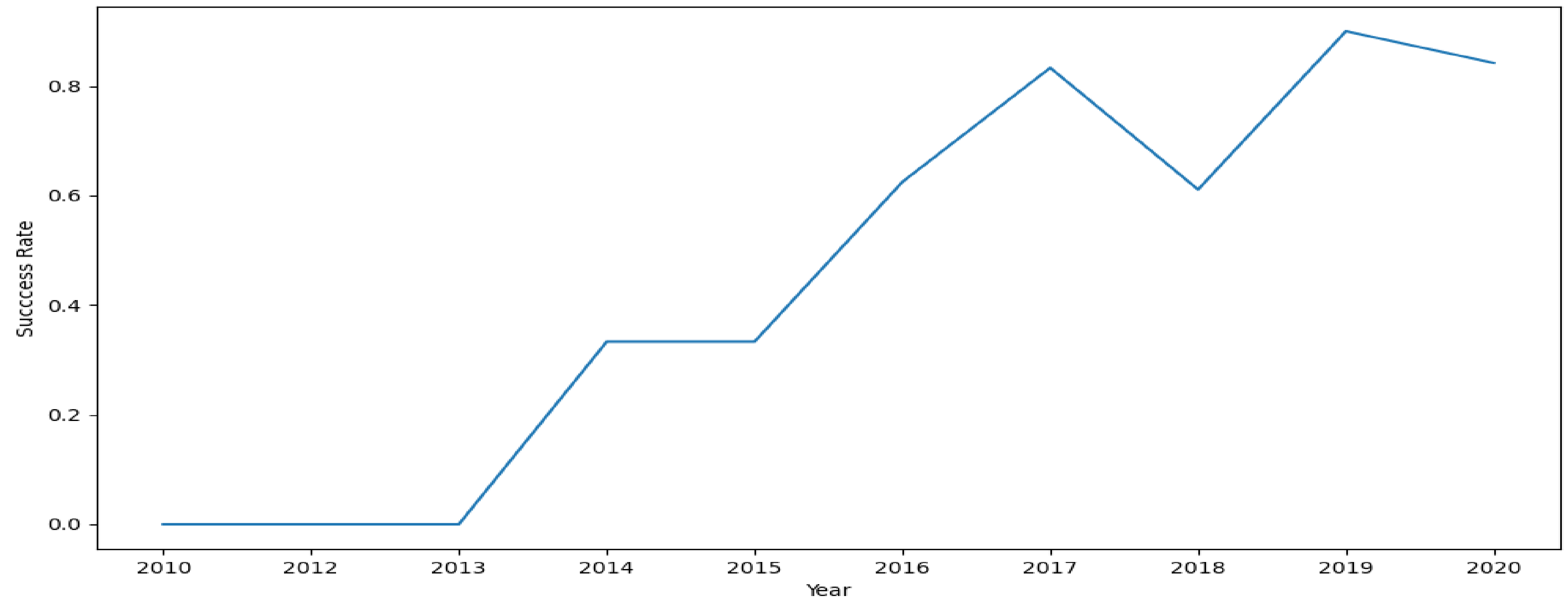
- In LEO orbit the Success appears related to the number of flights.

Payload vs. Orbit Type



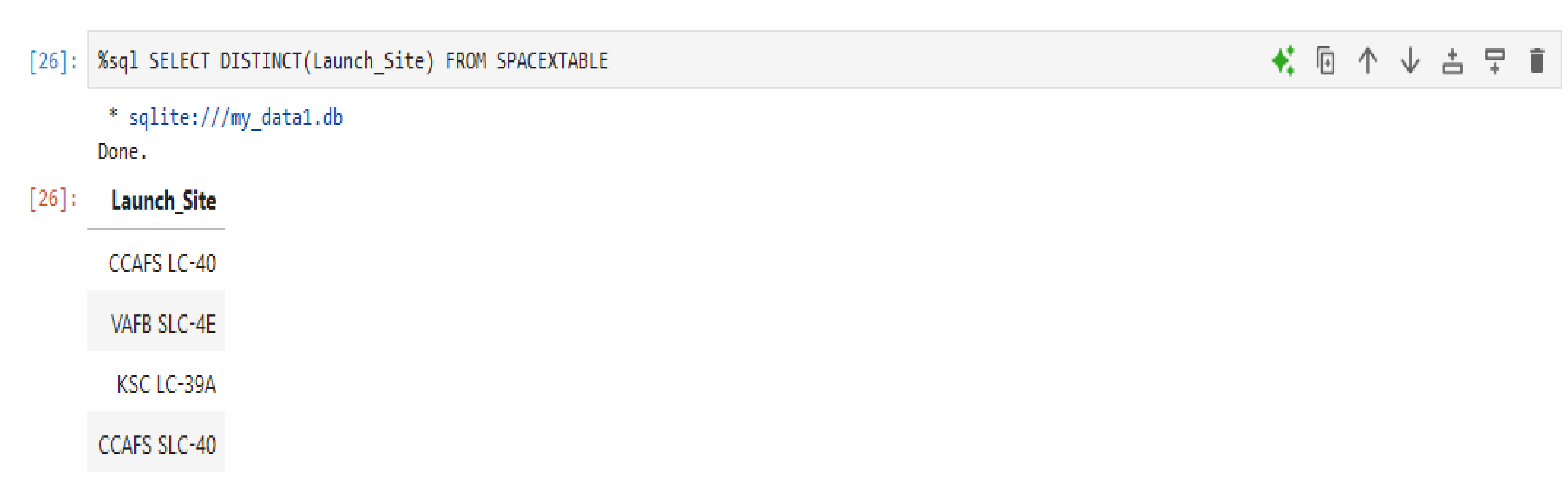
- With the increase of payload mass the successful landing or positive landing rate are more for Polar, LEO and ISS, although not great evidence is provided.
- Not clear patterns for the others orbits.

Launch Success Yearly Trend



- There is a clear trend to get better Launch Success rate over the years, occasionally decreases like in 2018 and 2020.

All Launch Site Names



```
[26]: %sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

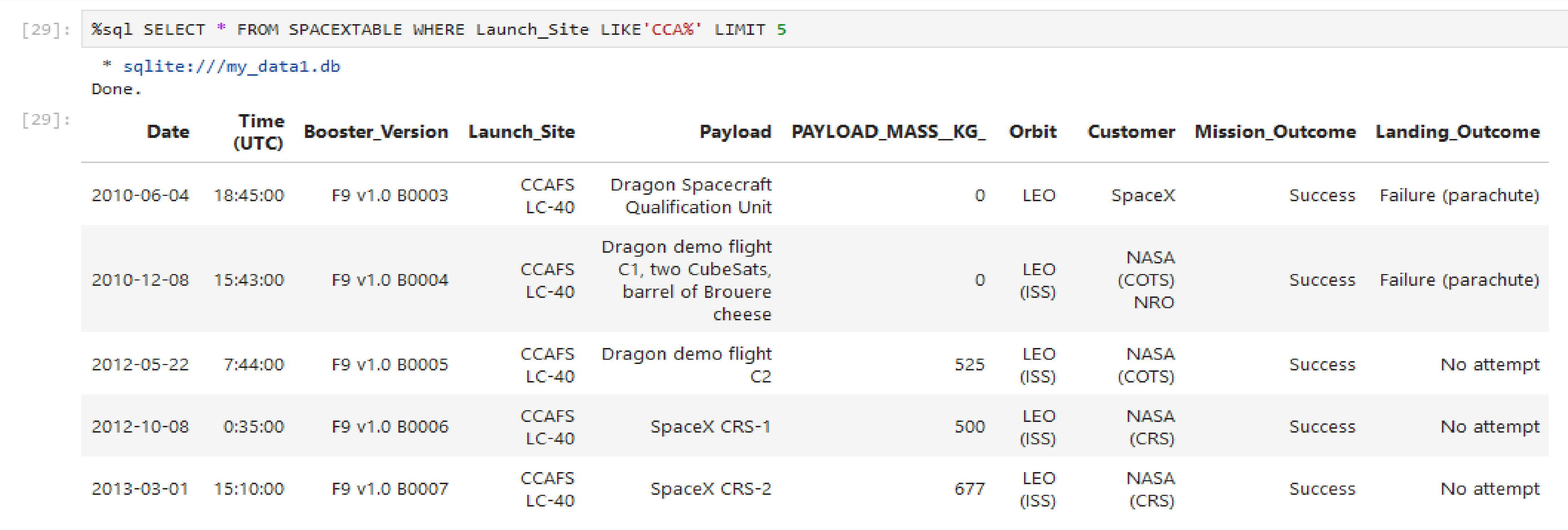
```
 * sqlite:///my_data1.db
Done.
```

[26]:

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

- There are 4 unique launch sites as appear in the picture.

Launch Site Names Begin with 'CCA'



```
[29]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE'CCA%' LIMIT 5
```

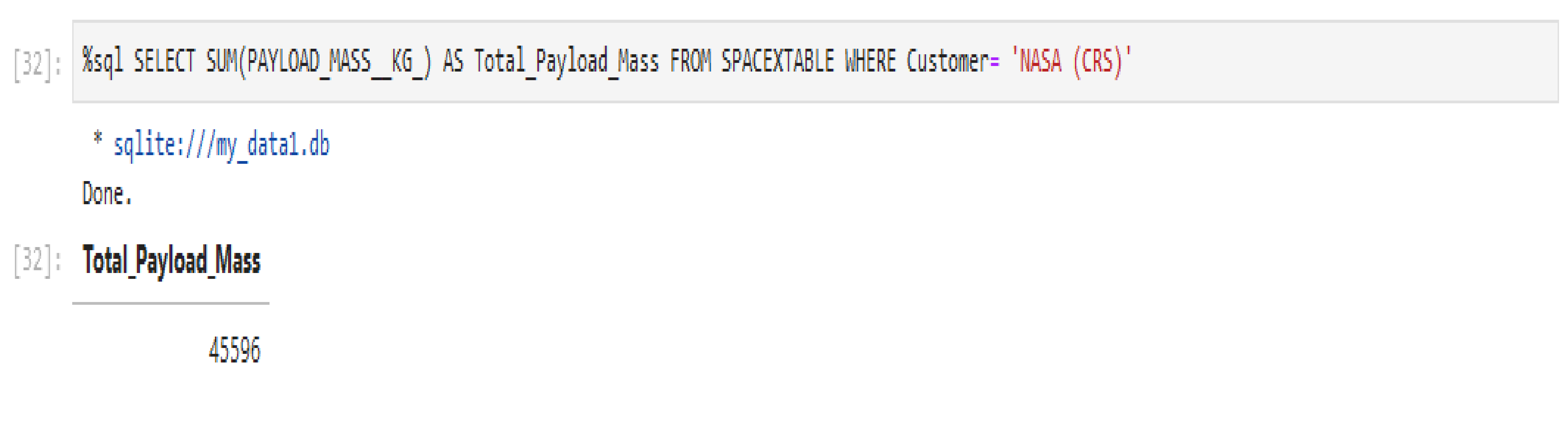
```
 * sqlite:///my_data1.db
Done.
```

[29]:

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

- Only 1 launch site begin with `CCA`.
- The query retrieves 5 records from launch site CCAFS LC-40.

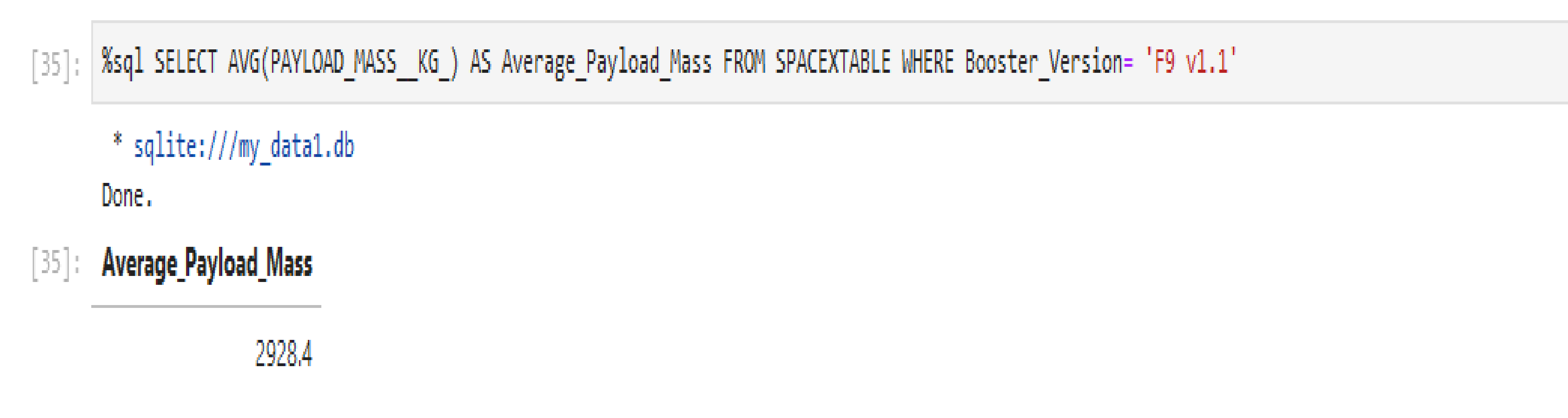
Total Payload Mass



```
[32]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer= 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

[32]: Total_Payload_Mass
                   45596
```

- The total payload carried by boosters from NASA was 45596 Kg.

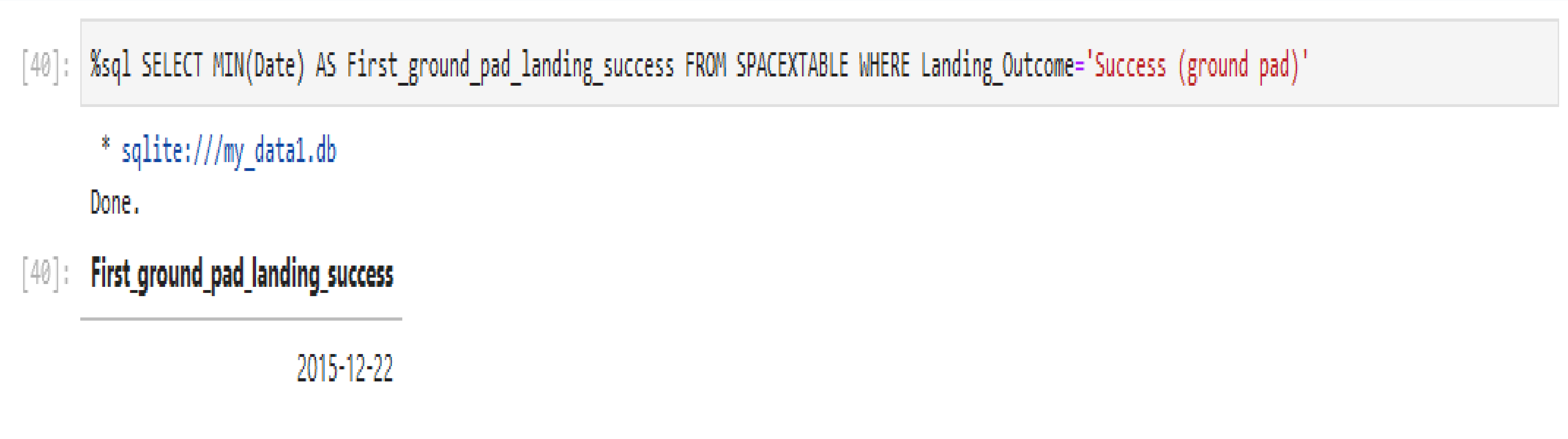
Average Payload Mass by F9 v1.1



```
[35]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version= 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

[35]: Average_Payload_Mass
                    2928.4
```

- The average payload mass carried by booster version F9 v1.1 was 2928.4 Kg.

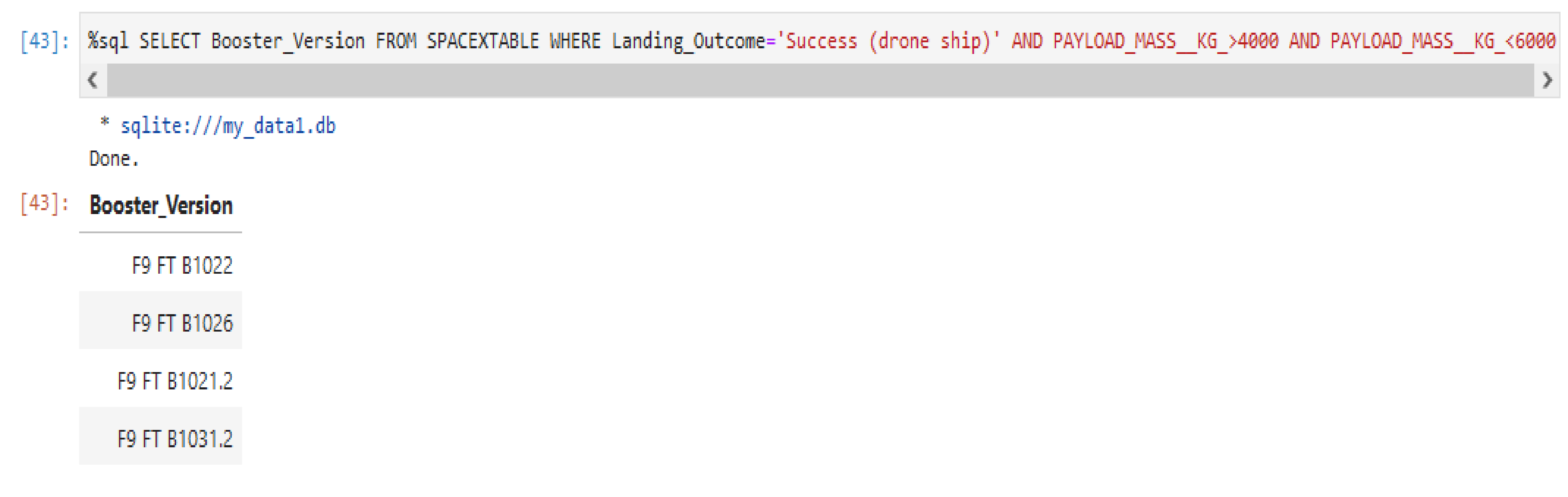
First Successful Ground Landing Date



```
[40]: %sql SELECT MIN(Date) AS First_ground_pad_landing_success FROM SPACEXTABLE WHERE Landing_Outcome='Success (ground pad)'

 * sqlite:///my_data1.db
Done.

[40]: First_ground_pad_landing_success
      2015-12-22
```

- December 22, 2015 was the date of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000



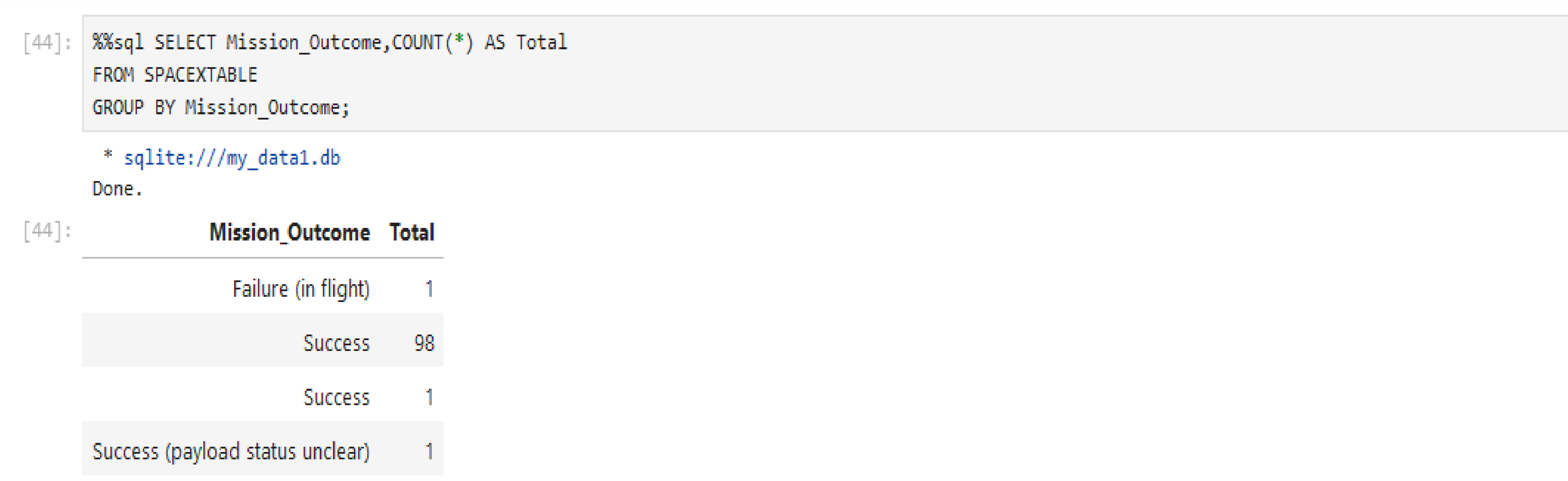
```
[43]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000

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

[43]:

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

- 4 boosters versions have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 as shown in picture.

Total Number of Successful and Failure Mission Outcomes



```
[44]: %%sql SELECT Mission_Outcome,COUNT(*) AS Total
      FROM SPACEXTABLE
      GROUP BY Mission_Outcome;
```

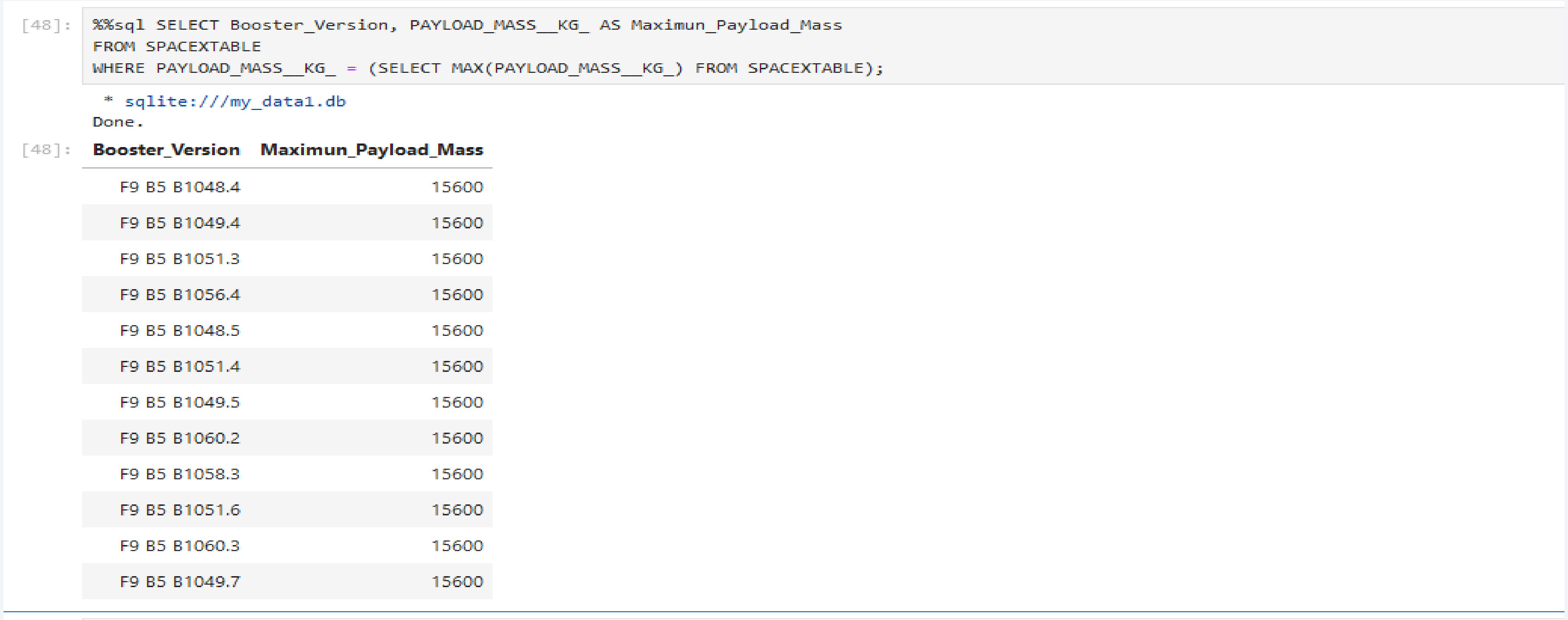
```
 * sqlite:///my_data1.db
Done.
```

[44]:

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

- The total number of successful mission outcomes was 100 and just 1 failure mission outcome.

Boosters Carried Maximum Payload



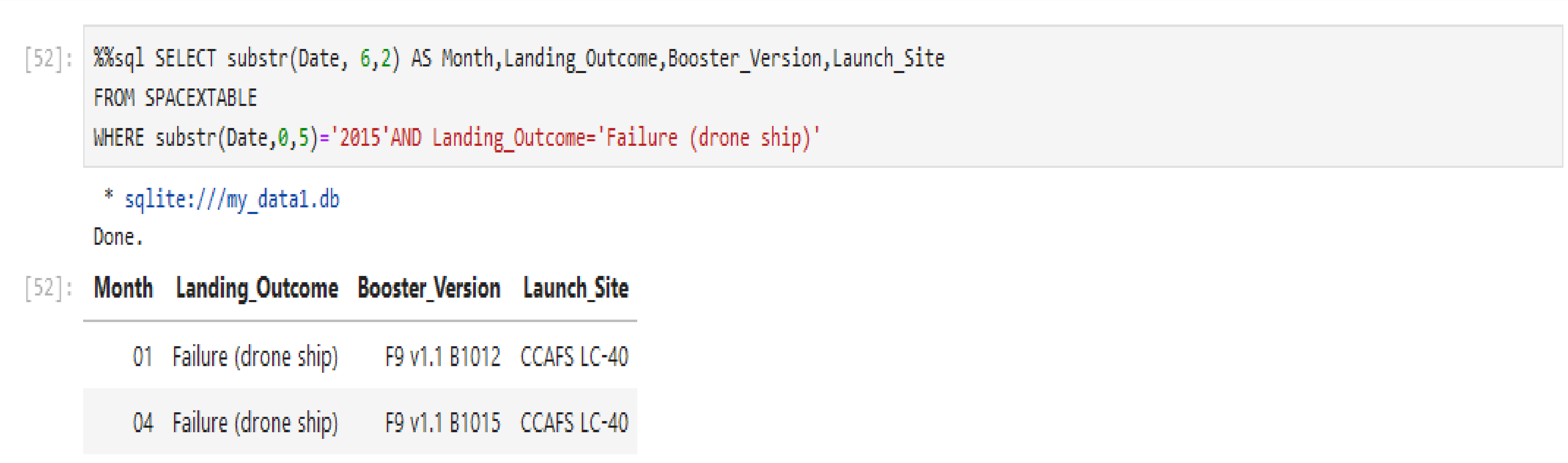
```
[48]: %%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ AS Maximun_Payload_Mass
      FROM SPACEXTABLE
      WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

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

[48]:

Booster_Version	Maximun_Payload_Mass
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

- The maximum payload mass was 15600 Kg., carried by 12 different booster versions.

2015 Launch Records



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

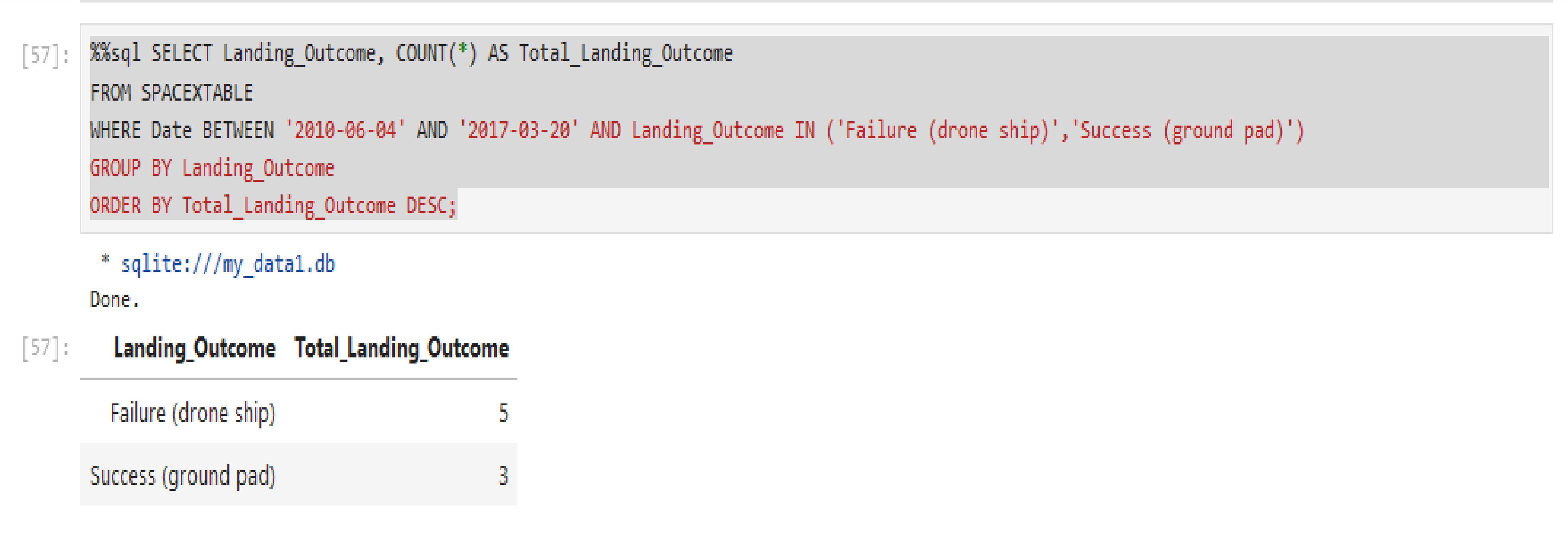
```
 * sqlite:///my_data1.db
Done.
```

[52]:

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

- 2 different failed landing_outcomes in drone ship occurs in year 2015, the query retrieves the booster versions, and launch site name.

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



[57]:
```
%%sql SELECT Landing_Outcome, COUNT(*) AS Total_Landing_Outcome
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' AND Landing_Outcome IN ('Failure (drone ship)','Success (ground pad)')
GROUP BY Landing_Outcome
ORDER BY Total_Landing_Outcome DESC;
```

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

[57]:

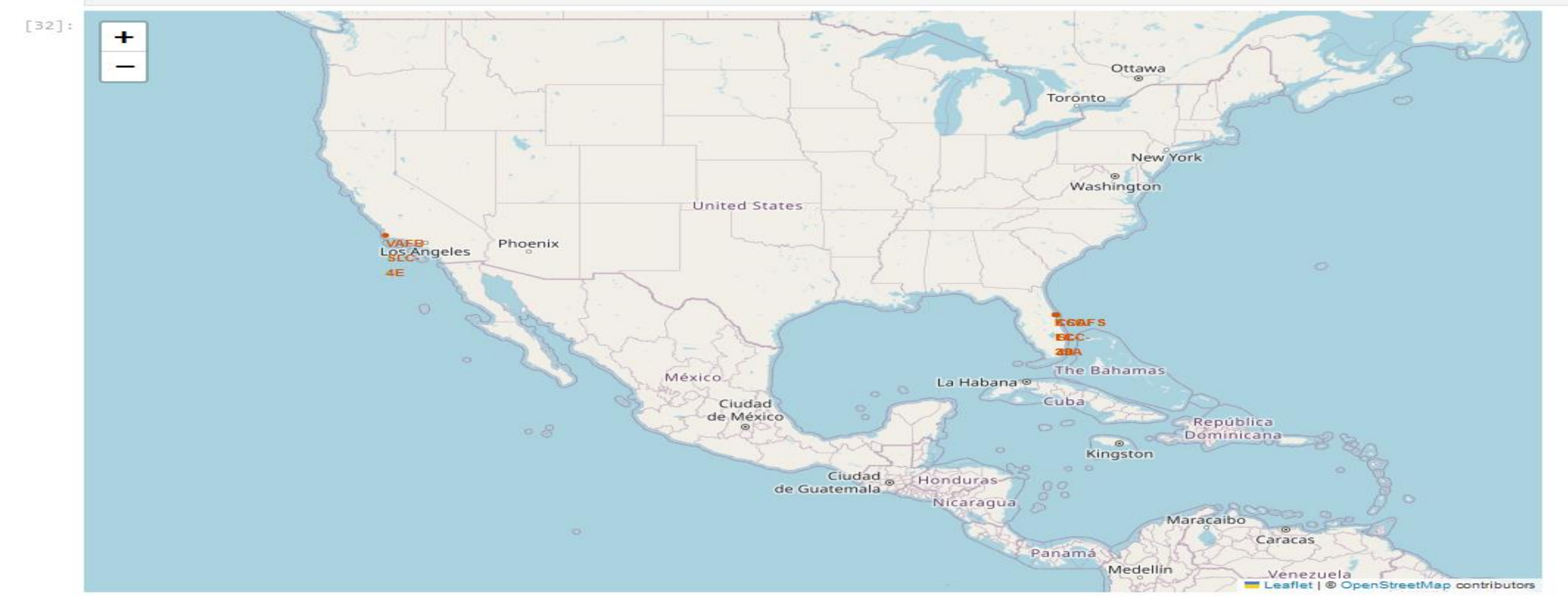
Landing_Outcome	Total_Landing_Outcome
Failure (drone ship)	5
Success (ground pad)	3

- Failure (drone ship) occurs 5 times and Success (ground pad) occurs 3 times between the date 2010-06-04 and 2017-03-20.

Section 3

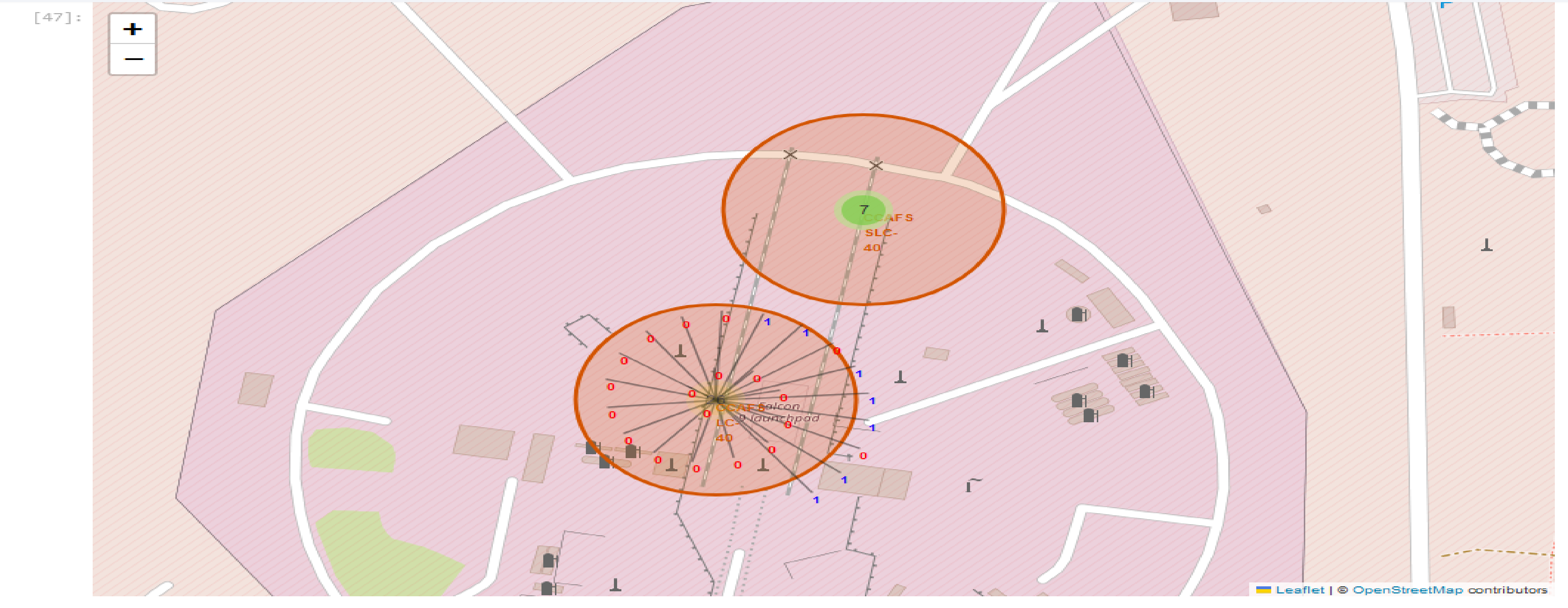
Launch Sites Proximities Analysis

Launch Sites Locations Map



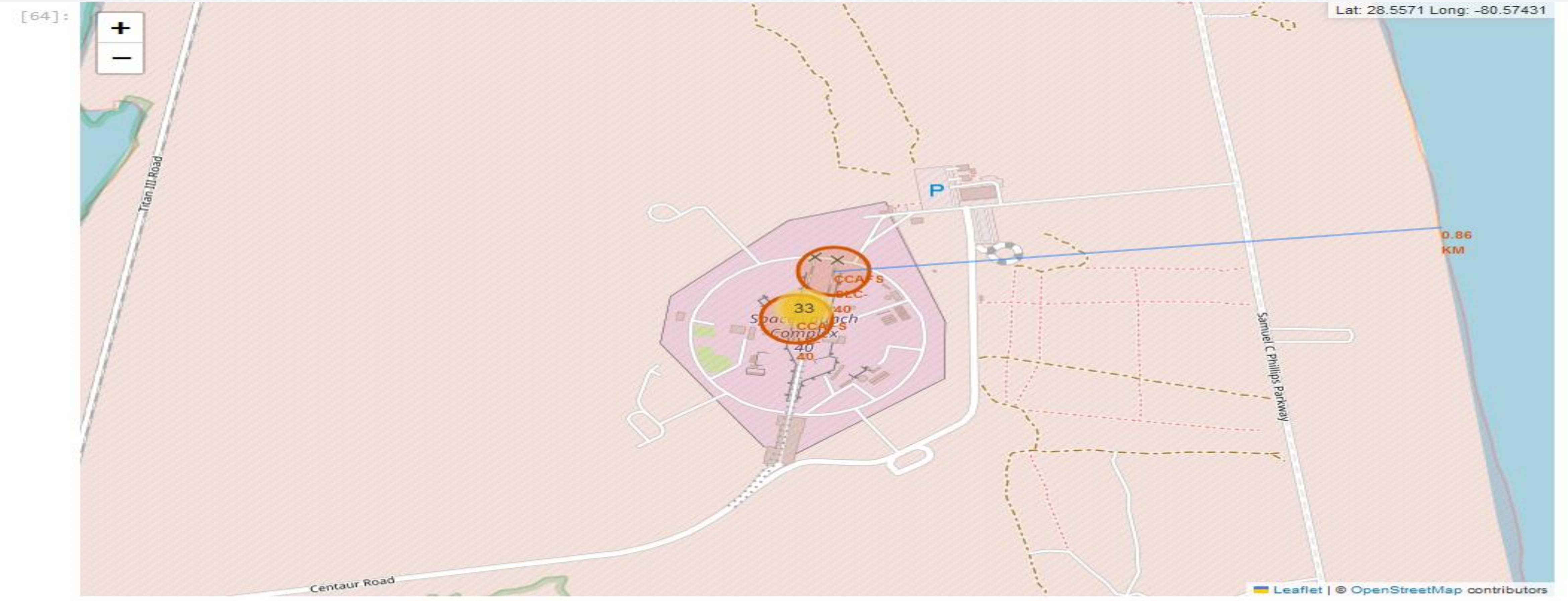
- The launch sites are located on the Pacific and Atlantic coasts of United States of America. Specifically, in the states of Florida and California.

Successed or failed launch according location

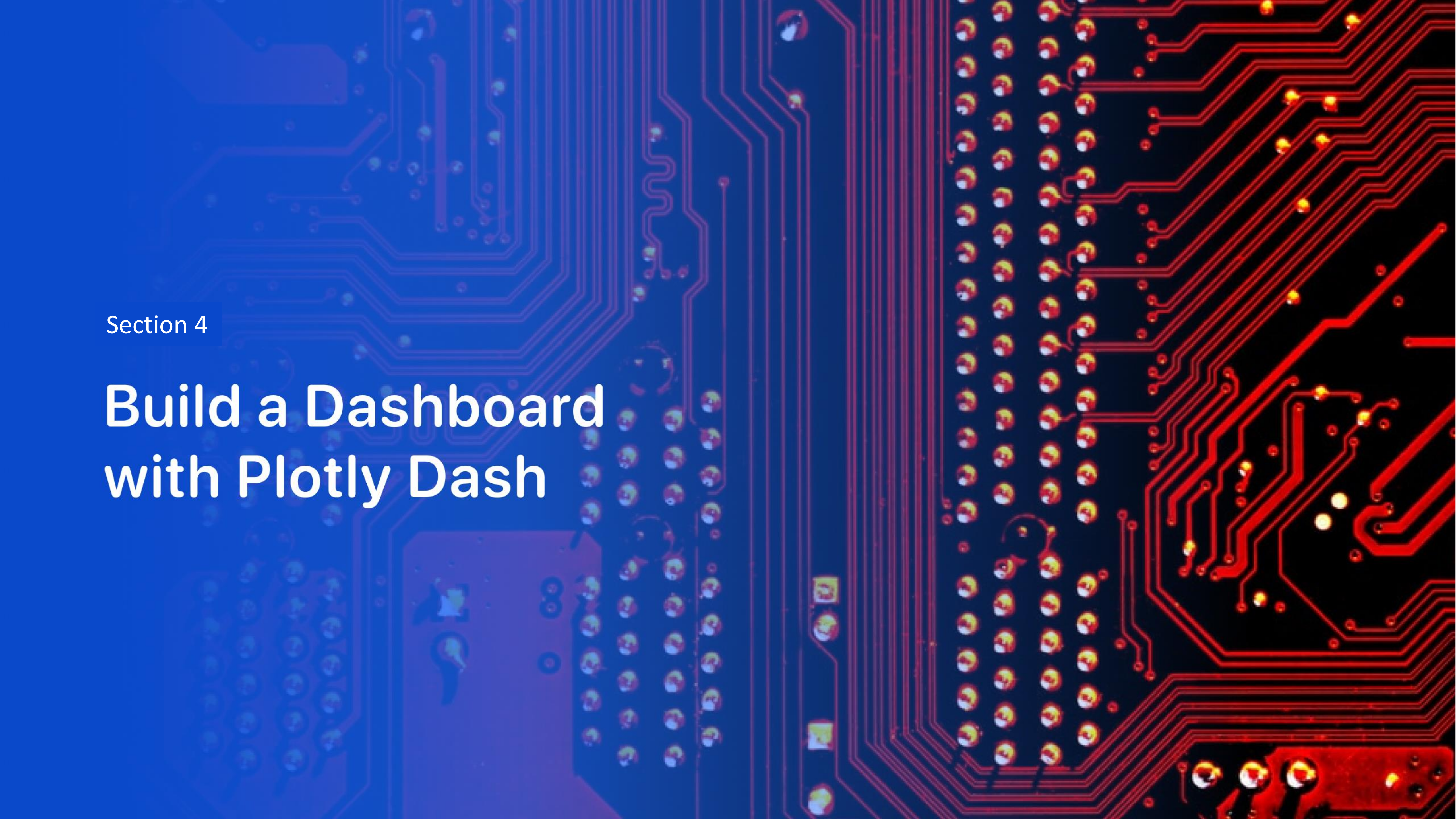


- It can be easily identify which launch sites have relatively high or low success rates from the color-labeled markers.

Distance to nearest coast according launch location



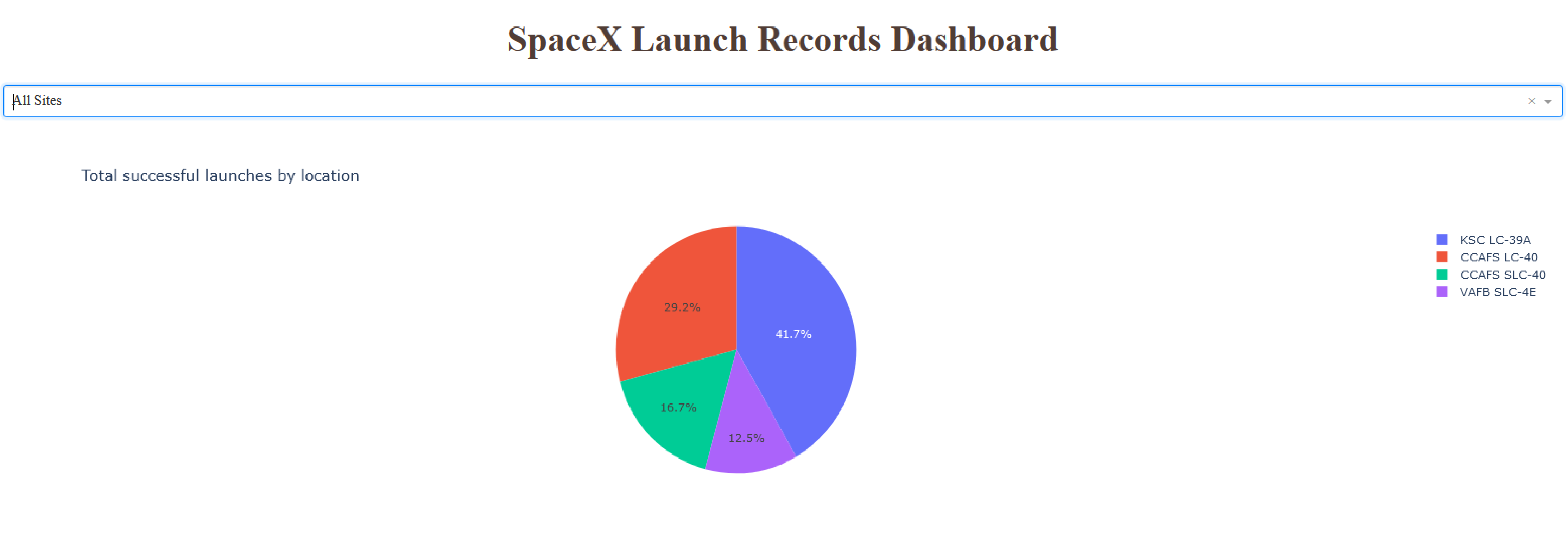
- Distance to the coast is less than 1 kilometer, which is relevant for safety and recovery planning.



Section 4

Build a Dashboard with Plotly Dash

Successful launches count for all sites



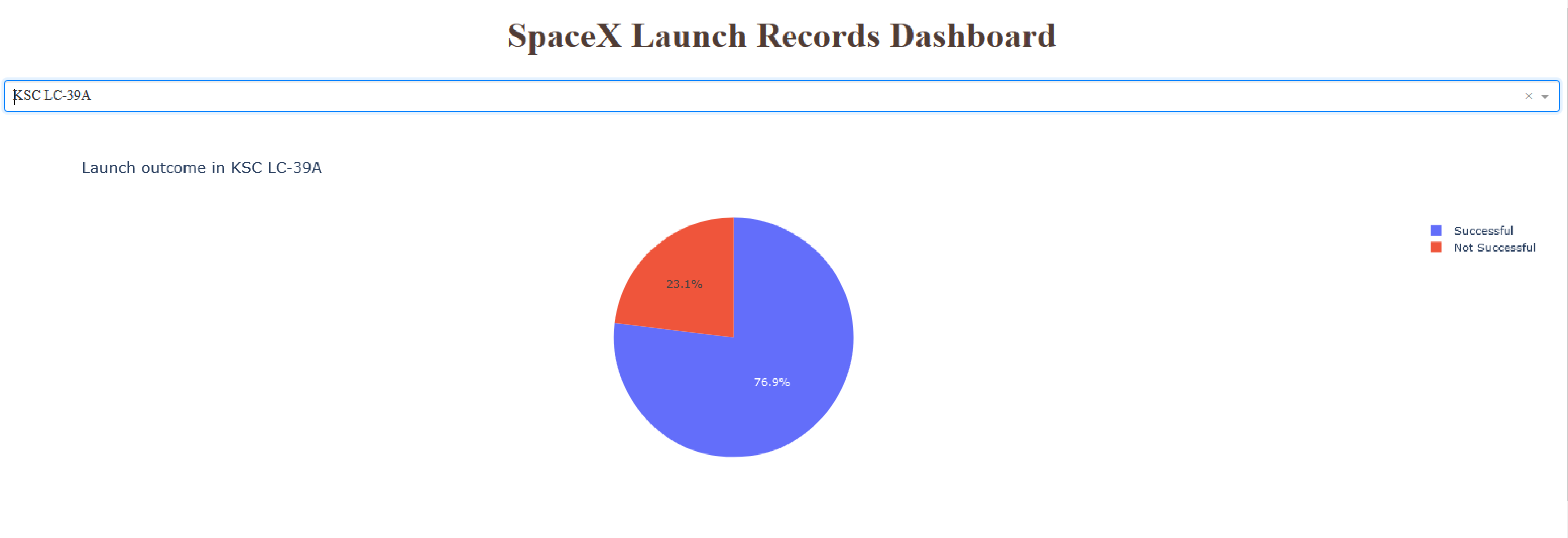
SpaceX Launch Records Dashboard

All Sites

Total successful launches by location

The data suggests KSC LC-39A is the most frequently used site for successful launches, highlighting its significance in SpaceX's launch operations.

Success vs. Failed counts for a specific site



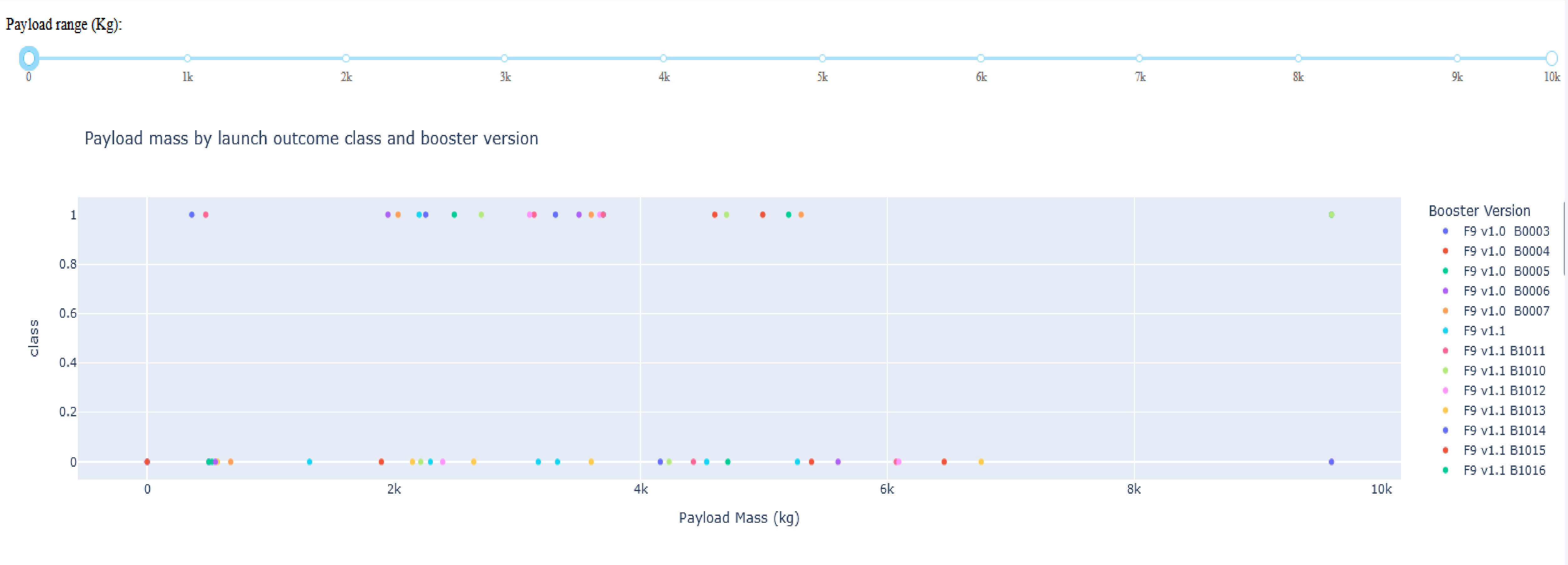
SpaceX Launch Records Dashboard

KSC LC-39A

Launch outcome in KSC LC-39A

The data suggests that, KSC LC-39A is the most frequently used site for successful launches because its higher success ratio.

Relation between Payload Mass, Success or Fail and Booster version



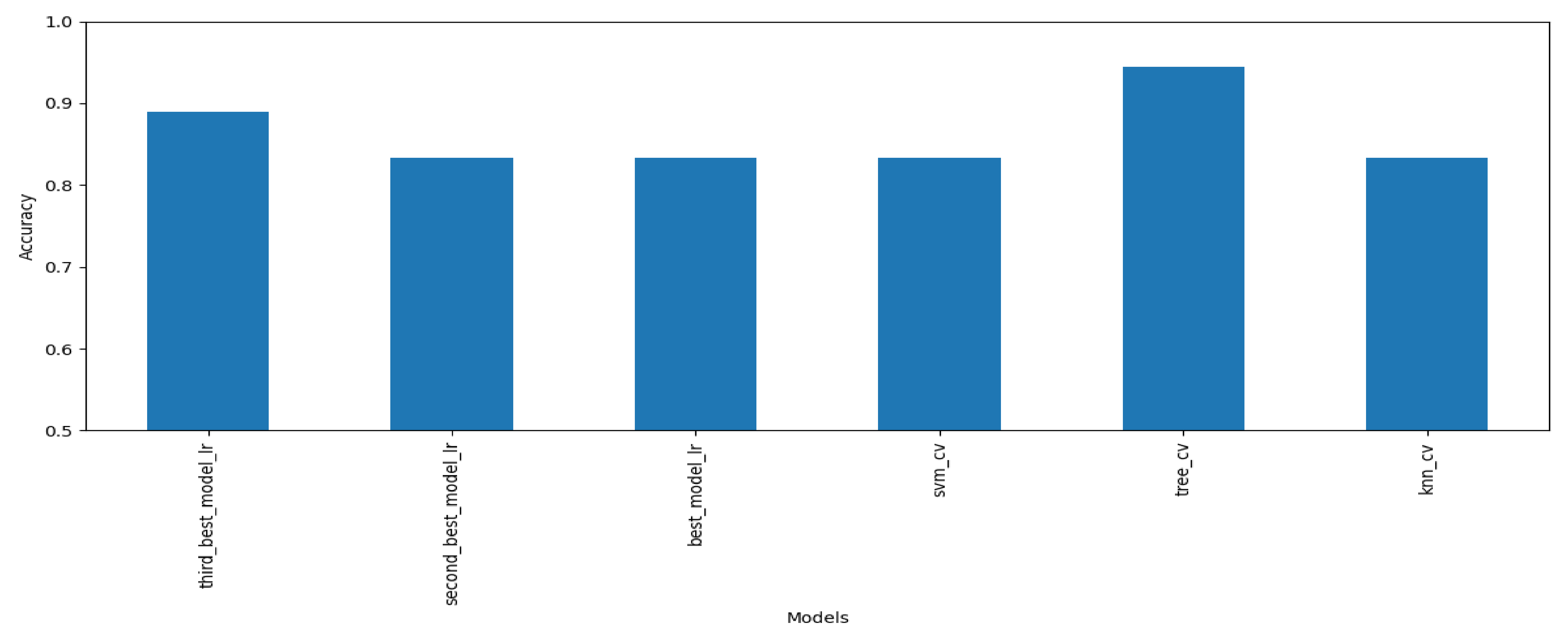
Payload range (Kg):

A clear correlation between the variables represented in the chart is not readily apparent.

Section 5

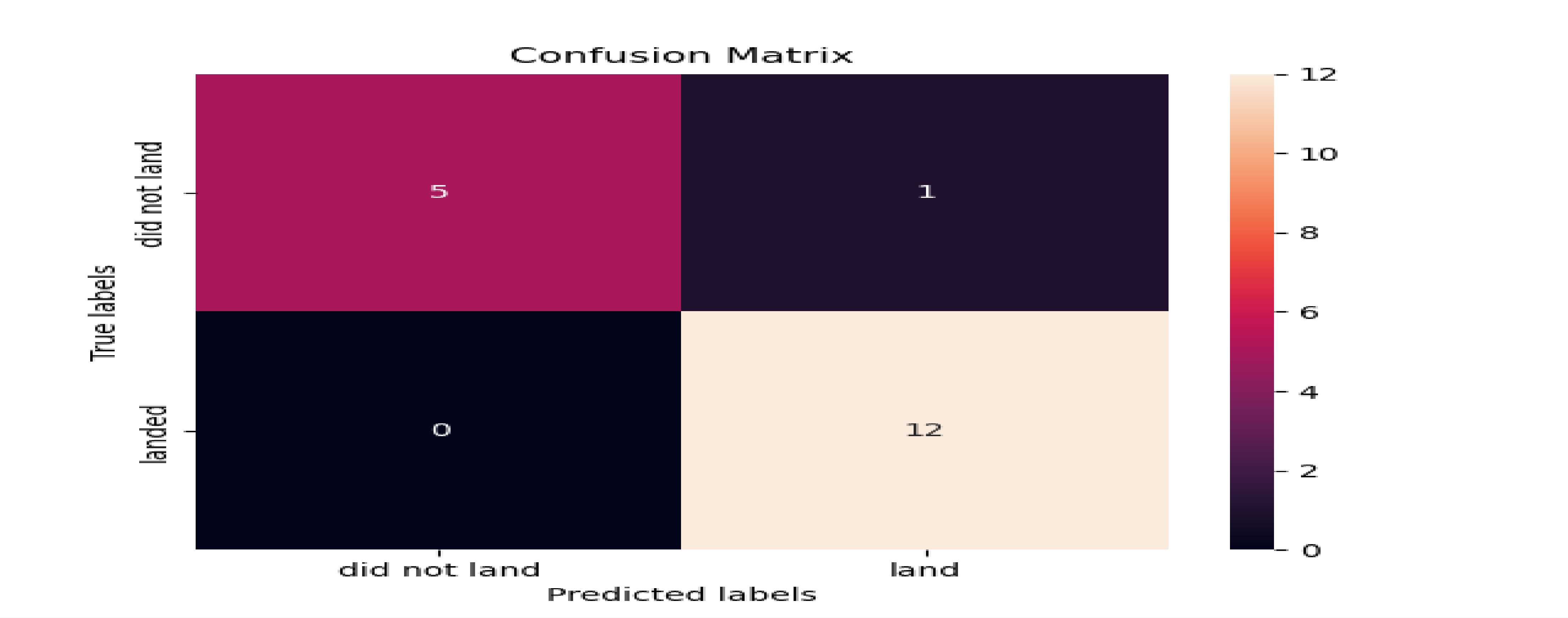
Predictive Analysis (Classification)

Classification Accuracy



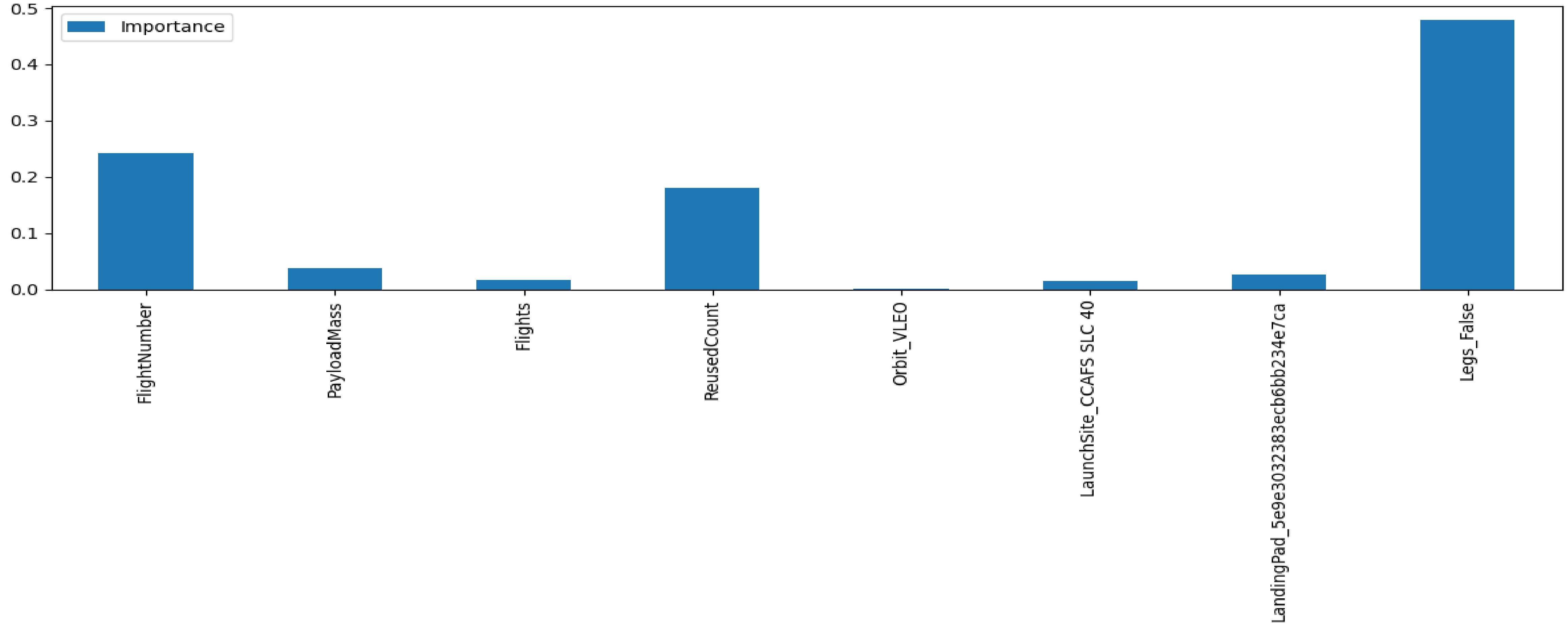
Decision Tree Classifier was the best model with the highest classification accuracy (0.9444).

Confusion Matrix



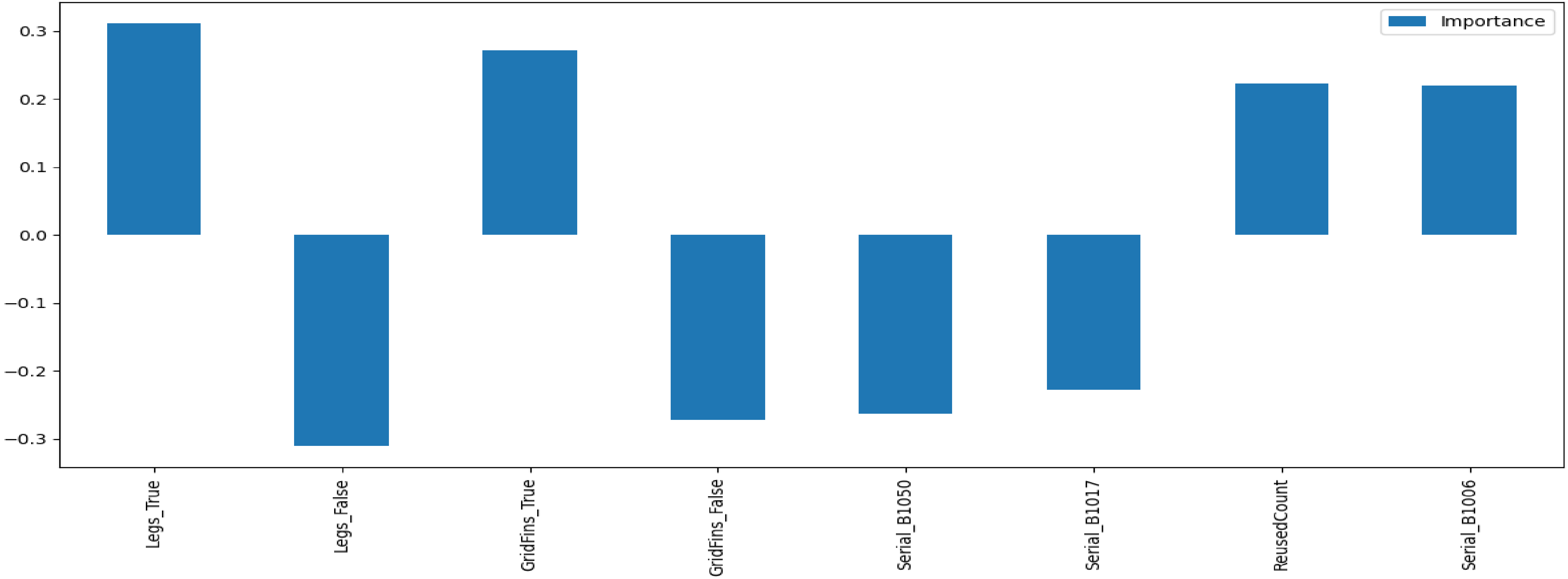
The confusion matrix shows good performance of the model Decision Tree Classifier on unseen data, even with the class imbalance in the target variable, as demonstrated by an F1 score of 0.96.

Features importance (Decision Tree Classifier)



The analysis of the features importance in the Decision Tree Classifier model reveals in the chart the most significant features, with variables Legs, FlightNumber and ReusedCount standing out.

Features importance (Logistic Regression)



The analysis of the features importance in the Logistic Regression model (as the second best model) reveals in the graph the most significant features according to coefficient values, with variables Legs and ReusedCount matching coincidence respect the Decision Tree Classifier.

Conclusions

Key Analysis Findings

1. Exploratory Analysis Highlights:

- Launch Site Success Rates: Certain launch sites perform consistently better, highlighting site-specific factors that may impact success.
- Orbit Success Patterns: Higher success rates were found for specific orbits, guiding optimal mission destination planning.
- Improvement Over Time: Clear upward trend in success rates, showing the impact of technological advancements.

2. SQL-Based Insights:

- Landing and Outcome Patterns: Identified reusability milestones and failure patterns, especially in drone landings, to refine future landing approaches.

Conclusions

3. Interactive Visualization for Deeper Insights

- Created dynamic maps and analytics dashboards with Folium and Dash to allow interactive data exploration.

4. Predictive Modeling Results

- Models Trained: Logistic Regression, SVM, Decision Tree, KNN (using GridSearchCV for tuning).
- Best Model: Decision Tree Classifier with accuracy of 0.9444 and F1 score of 0.96.
- Key Features Identified: Legs and ReusedCount among others emerged as the most influential features for model performance.

Conclusions

The analysis revealed actionable insights on launch site effectiveness, optimal orbits and other factors influencing success. Interactive tools and predictive models, especially the Decision Tree, provided a robust framework for improving mission planning and enhancing reusability strategies.

Appendix

- All the code, charts and stored data created during this project is available in:

https://github.com/jfrometa88/Applied-Data-Science-Capstone-IBM

Thank you!