IBM Developer
SKILLS NETWORK

Data Science Capstone project.

SATYENDRA G 30/06/2024

Executive Summary	1
Introduction	2
Methodology	3
Results	4
Conclusion	5
Appendix	6

OUTLINE

Executive Summary

Summary of Methodologies

- **Data Collection:** Public SpaceX API and Wikipedia.
- **Data Wrangling:** Created label column ‘class’; one-hot encoding; data standardization.
- **Exploratory Data Analysis (EDA):** SQL queries, visualizations, folium maps, dashboards with Plotly Dash.
- **Predictive Analysis:** Logistic Regression, Support Vector Machine, Decision Tree Classifier, K Nearest Neighbors; GridSearchCV for parameter tuning; model accuracy visualization.

Summary of Results

- **EDA Insights:** Uncovered patterns and trends; interactive maps and dashboards for deeper understanding.
- **Model Performance:** All models achieved ~83.33% accuracy; over-prediction of successful landings.
- **Conclusion:** Further data needed for improved accuracy and generalization; solid framework established for ongoing analysis.

Introduction

Project Background and Context:

SpaceX has made significant strides in the commercial space industry by offering cost-effective space travel solutions. Their Falcon 9 rocket launches are priced at $62 million, compared to $165 million from other providers, due to SpaceX's innovative ability to reuse the rocket's first stage. SpaceY, aiming to compete in this market, seeks to achieve similar cost savings by successfully recovering the first stage of their rockets.
To compete with SpaceX, SpaceY has tasked us with developing a machine learning model to predict the likelihood of a successful Stage 1 recovery. Accurately predicting these landings will enable SpaceY to reduce launch costs and enhance their competitiveness in the commercial space sector.

Questions to be Answered

1. Which algorithm provides the best performance for binary classification in predicting successful first stage recoveries?
2. Is there an observable trend in the success rate of landings over time?

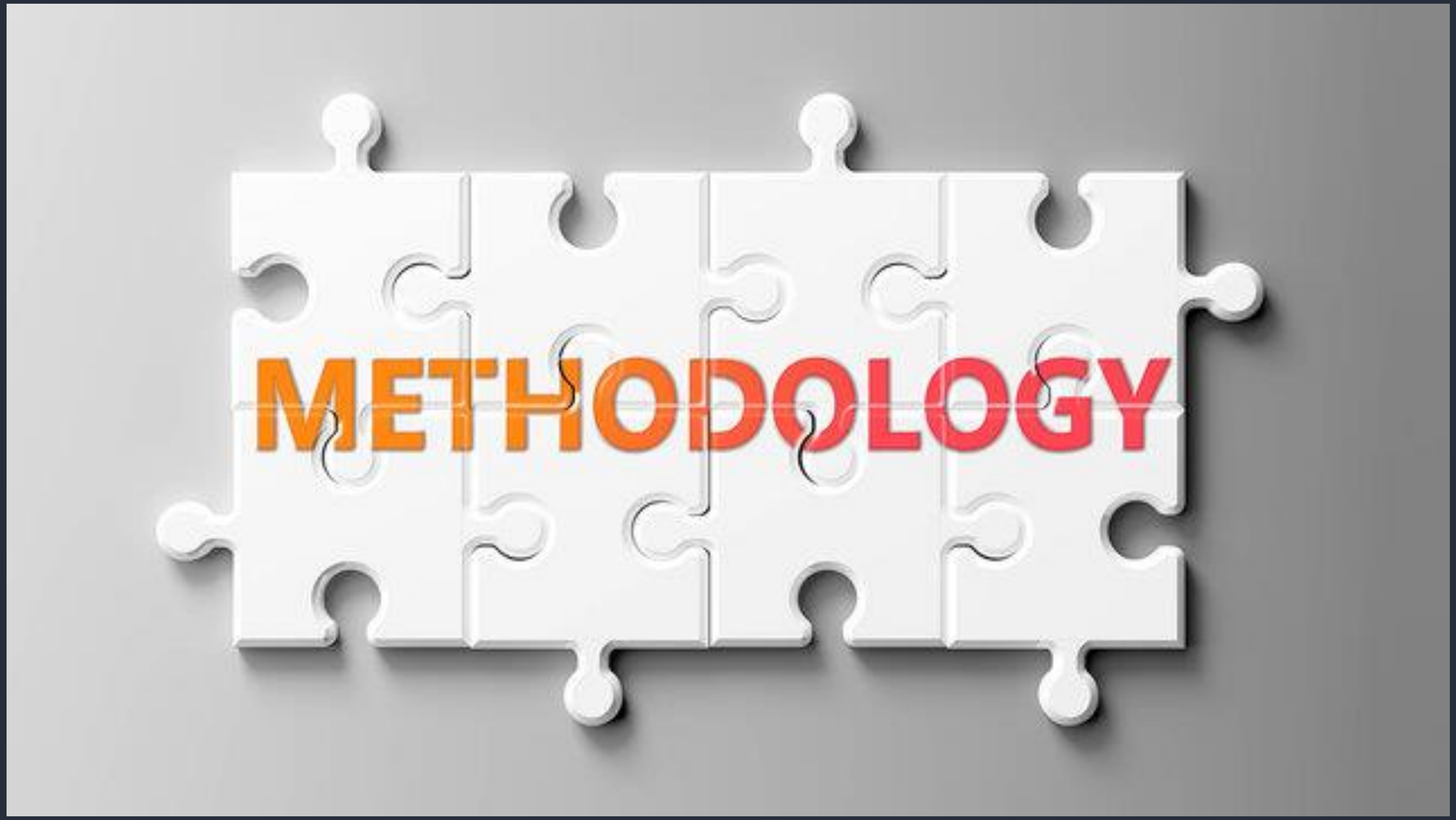
Data collection methodology:

- Acquired data using SpaceX's public API and conducted web scraping for comprehensive data gathering.
- Performed data wrangling:

  Cleaned and preprocessed data to ensure quality:

  - Removed duplicates and handled missing values.
  - Applied feature engineering to enhance model performance.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Model Construction: Developed classification models based on preprocessed data.
  - Parameter Tuning: Utilized GridSearchCV to optimize model hyperparameters.
  - Performance Evaluation: Evaluated models using metrics like accuracy.

Methodology



Data Collection

SpaceX API Data Collection:

•**Utilized SpaceX's public API to gather data on:**

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

•**Flowchart of SpaceX API Data Collection**: (Include a flowchart depicting the process of API requests and data retrieval)

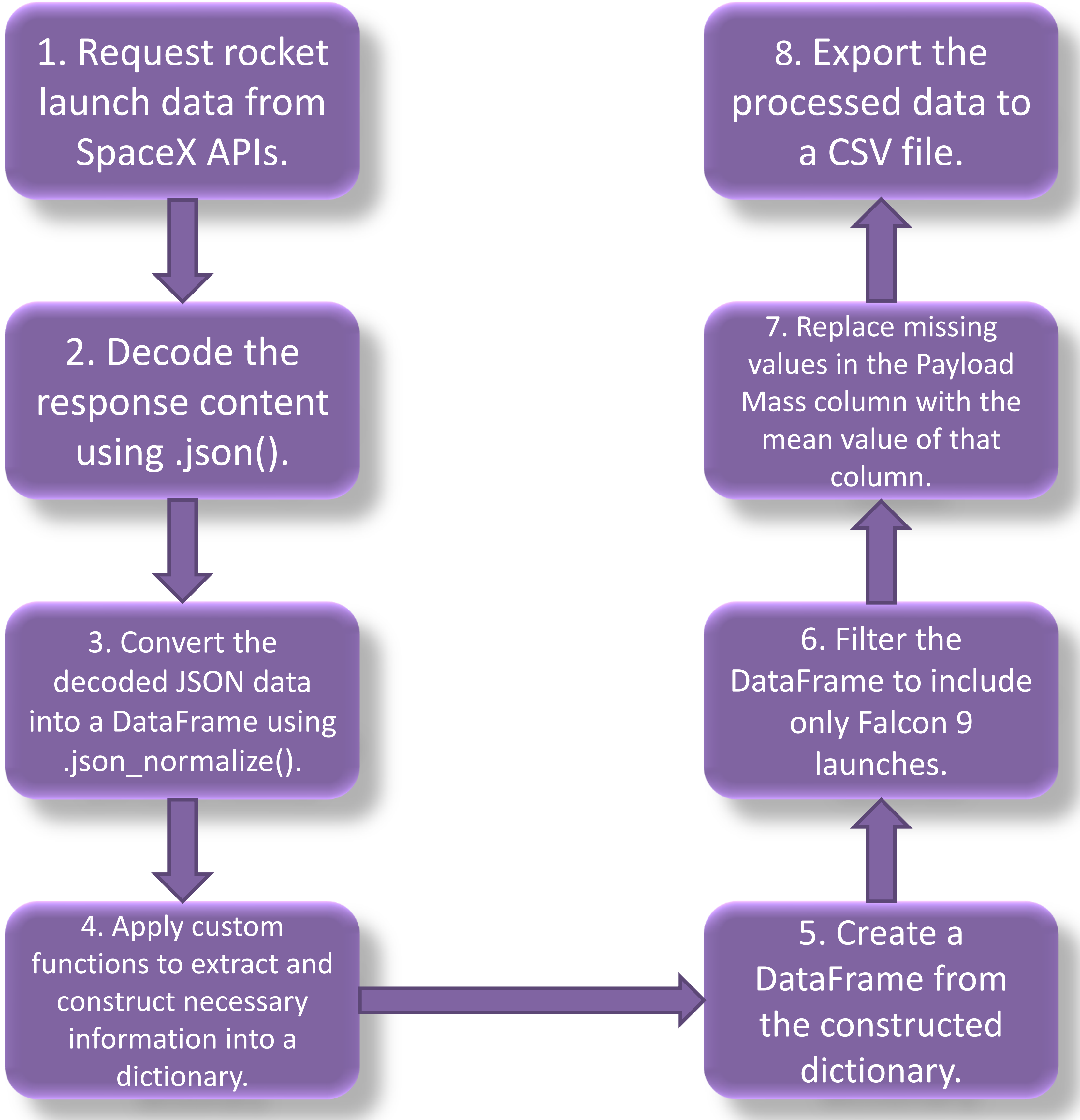
Wikipedia Web Scraping:

•**Scraped data from SpaceX's Wikipedia entry table for:**

- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.

•**Flowchart of Wikipedia Web Scraping Data Collection**: (Include a flowchart showing the steps involved in web scraping and data extraction)

Data Collection –SpaceX API



1. Request rocket launch data from SpaceX APIs.

2. Decode the response content using .json().

3. Convert the decoded JSON data into a DataFrame using .json_normalize().

4. Apply custom functions to extract and construct necessary information into a dictionary.

5. Create a DataFrame from the constructed dictionary.

6. Filter the DataFrame to include only Falcon 9 launches.

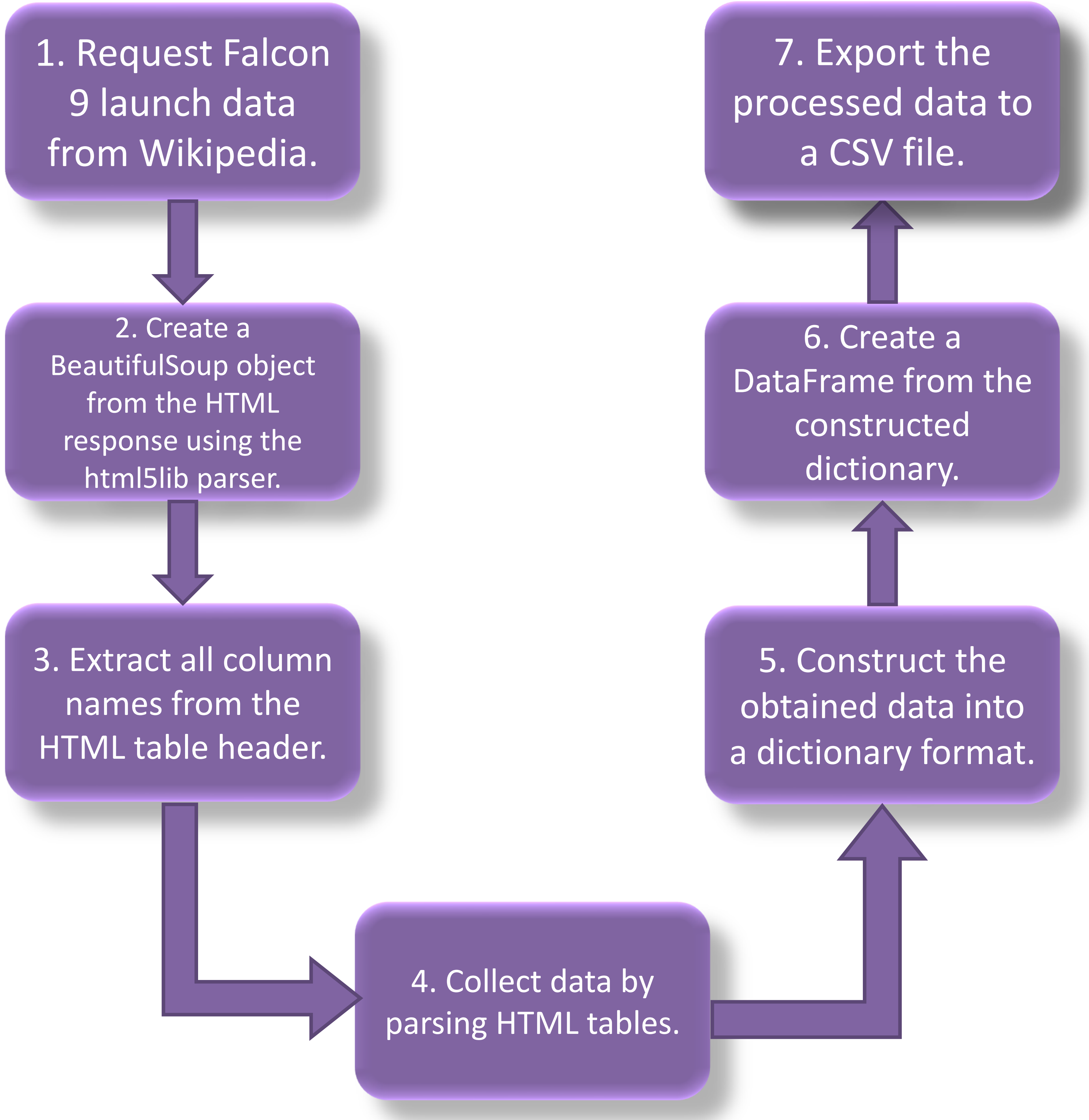
7. Replace missing values in the Payload Mass column with the mean value of that column.

8. Export the processed data to a CSV file.

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/Data-collection-api.ipynb

Data Collection - Scraping



1. Request Falcon 9 launch data from Wikipedia.

2. Create a BeautifulSoup object from the HTML response using the html5lib parser.

3. Extract all column names from the HTML table header.

4. Collect data by parsing HTML tables.

5. Construct the obtained data into a dictionary format.

6. Create a DataFrame from the constructed dictionary.

7. Export the processed data to a CSV file.

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/Data%20Collection%20%26%20Webscraping.ipynb

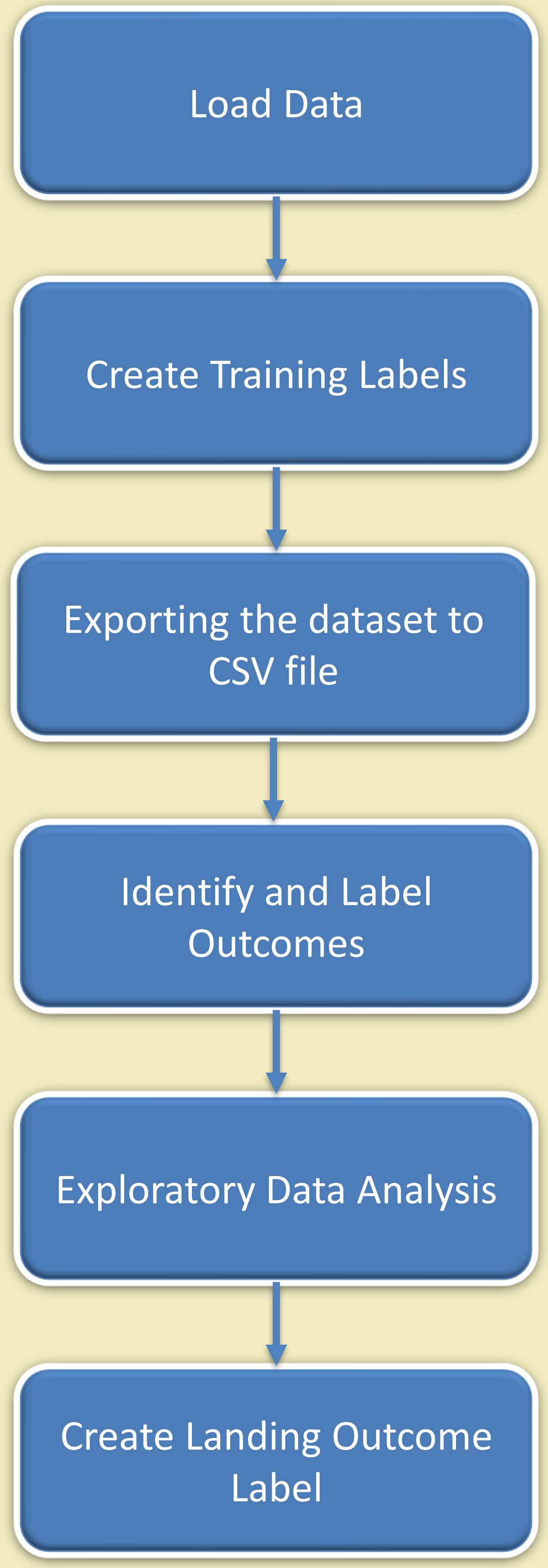
Github Link: https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb

Data Wrangling

Data Wrangling Process

1. Create Training Labels:
   - Create a new column class in the dataset.
   - Map mission outcomes to training labels:
     - True ASDS, True RTLS, True Ocean → 1
     - None None, False ASDS, None ASDS, False Ocean, False RTLS → 0
2. Identify and Label Outcomes:
   - Successful landings:
     - True Ocean → Landed successfully in the ocean.
     - True RTLS → Landed successfully on a ground pad.
     - True ASDS → Landed successfully on a drone ship.
   - Unsuccessful landings:
     - False Ocean → Unsuccessful landing in the ocean.
     - False RTLS → Unsuccessful landing on a ground pad.
     - False ASDS → Unsuccessful landing on a drone ship.
3. Exploratory Data Analysis:
   - Calculate the number of launches at each site.
   - Calculate the number and occurrence of each orbit.
   - Determine the number and occurrence of mission outcomes per orbit type.
4. Create Landing Outcome Label:
   - Generate a new class column indicating successful (1) or unsuccessful (0) landings based on the mission outcome.
5. Export Data:
   - Export the processed dataset to a CSV file for further analysis and modeling.

Flowchart:



Load Data

Create Training Labels

Exporting the dataset to CSV file

Identify and Label Outcomes

Exploratory Data Analysis

Create Landing Outcome Label

EDA with Data Visualization

Key Plots and Their Purpose:

1. Scatter Plots:
   - Flight Number vs. Payload Mass:
     - Correlation between flight number and payload mass.
   - Flight Number vs. Launch Site:
     - Distribution of flights across launch sites.
   - Payload Mass vs. Launch Site:
     - Payload capacities at different launch sites.
   - Flight Number vs. Orbit Type:
     - Patterns in orbit types over flights.
   - Payload Mass vs. Orbit Type and Success Rate:
     - Relationship between payload mass, orbit type, and mission success rate.
2. Bar Charts:
   - Orbit Type vs. Success Rate:
     - Comparison of success rates across orbit types.
3. Line Charts:
   - Success Rate Yearly Trend:
     - Trends in mission success rates over time.

Variables Analyzed:

Flight Number
Payload Mass
Launch Site
Orbit
Class (Success/Failure)
Year

Chart Types:

Scatter Plots: Identify potential predictors for ML models.
Bar Charts: Compare categorical data.
Line Charts: Show time series trends.

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/EDA%20with%20DV.ipynb

EDA with SQL

- **Loading Data:**
  - IBM DB2 Database using SQL Python integration.
- **Key Queries:**
  - Unique launch site names.
  - 5 records where launch sites begin with 'CCA'.
  - Total payload mass by NASA (CRS) boosters.
  - Average payload mass by F9 v1.1 boosters.
  - Date of first successful ground pad landing.
  - Boosters with successful drone ship landings and payload mass 4000-6000.
  - Total successful and failed mission outcomes.
  - Booster versions with maximum payload mass.
  - Failed drone ship landings, booster versions, and launch site names for 2015.
  - Ranking landing outcomes (2010-06-04 to 2017-03-20).

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/EDA-SQL.ipynb

Build an Interactive Map with Folium

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/IVA%20with%20Folium.ipynb

Markers of All Launch Sites:

- **Purpose:** Show geographical locations and proximity to the Equator and coasts.
- **Details:** Added markers with circles, popup labels, and text labels for each launch site.

Marker for NASA Johnson Space Center:

- **Purpose:** Serve as a start location.
- **Details:** Added a marker with a circle, popup label, and text label using latitude and longitude coordinates.

Colored Markers for Launch Outcomes:

- **Purpose:** Identify success rates of launch sites.
- **Details:** Added green markers for successful launches and red markers for failed launches using Marker Cluster.

Distances to Proximities:

- **Purpose:** Illustrate why launch sites are located where they are.
- **Details:** Added colored lines showing distances from Launch Site KSC LC-39A to nearby railway, highway, coastline, and city.

Build a Dashboard with Plotly Dash

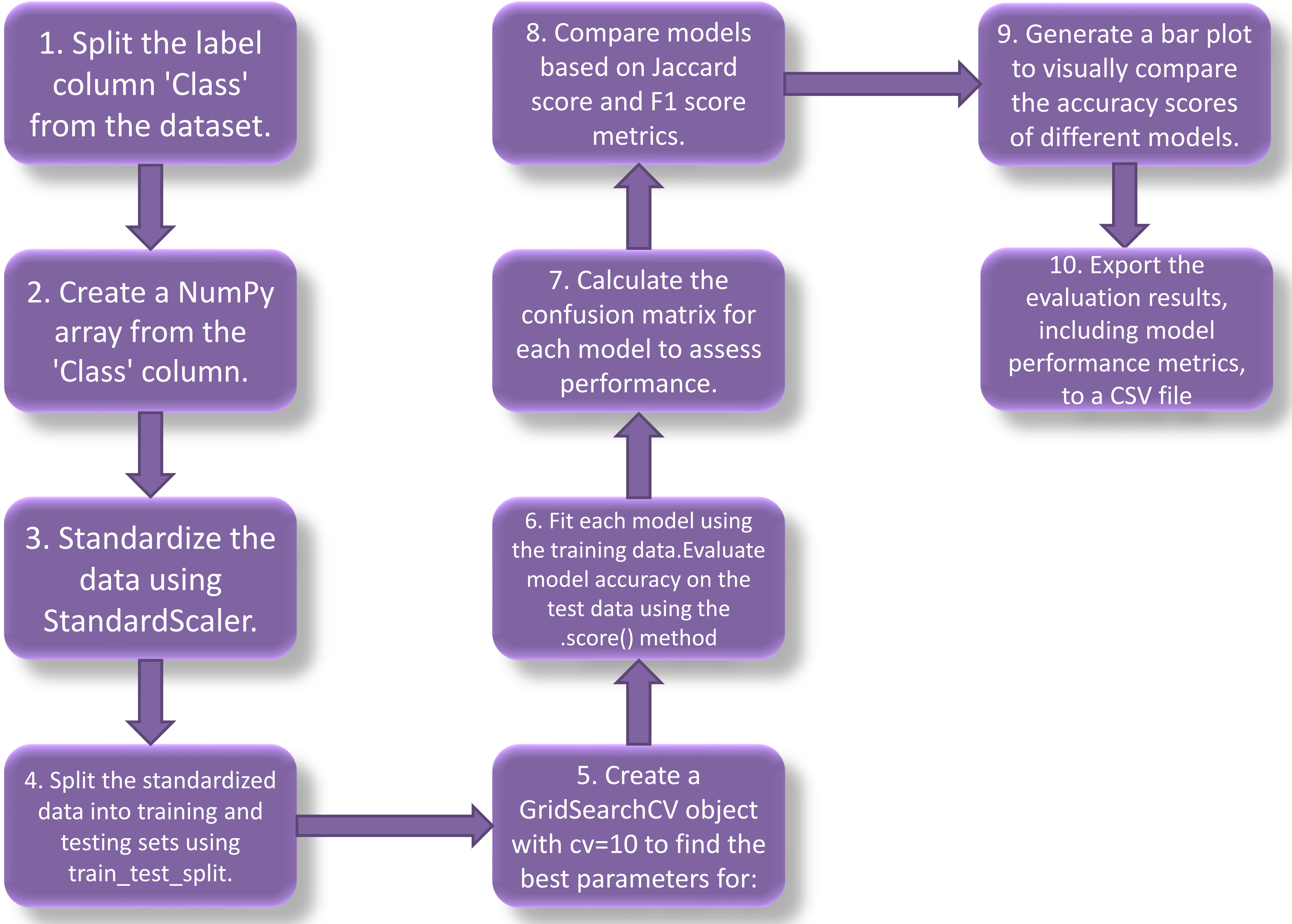
Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Added Plots/Graphs and Interactions:

1. Launch Sites Dropdown List:
   - o Purpose: Select specific launch sites to filter data.
2. Pie Chart:
   - o Purpose: Show success rates.
   - o Interaction:
     - Display success counts for all sites.
     - Show Success vs. Failed counts for selected sites.
3. Payload Mass Slider:
   - o Purpose: Filter by payload mass (0-10000 kg).
4. Scatter Chart (Payload Mass vs. Success Rate):
   - o Purpose: Show correlation between payload mass and launch success.
   - o Interaction:
     - Filter by all sites or selected sites.
     - Adjust based on payload mass.

Predictive Analysis (Classification)



1. Split the label column 'Class' from the dataset.

2. Create a NumPy array from the 'Class' column.

3. Standardize the data using StandardScaler.

4. Split the standardized data into training and testing sets using train_test_split.

5. Create a GridSearchCV object with cv=10 to find the best parameters for:

6. Fit each model using the training data.Evaluate model accuracy on the test data using the .score() method

7. Calculate the confusion matrix for each model to assess performance.

8. Compare models based on Jaccard score and F1 score metrics.

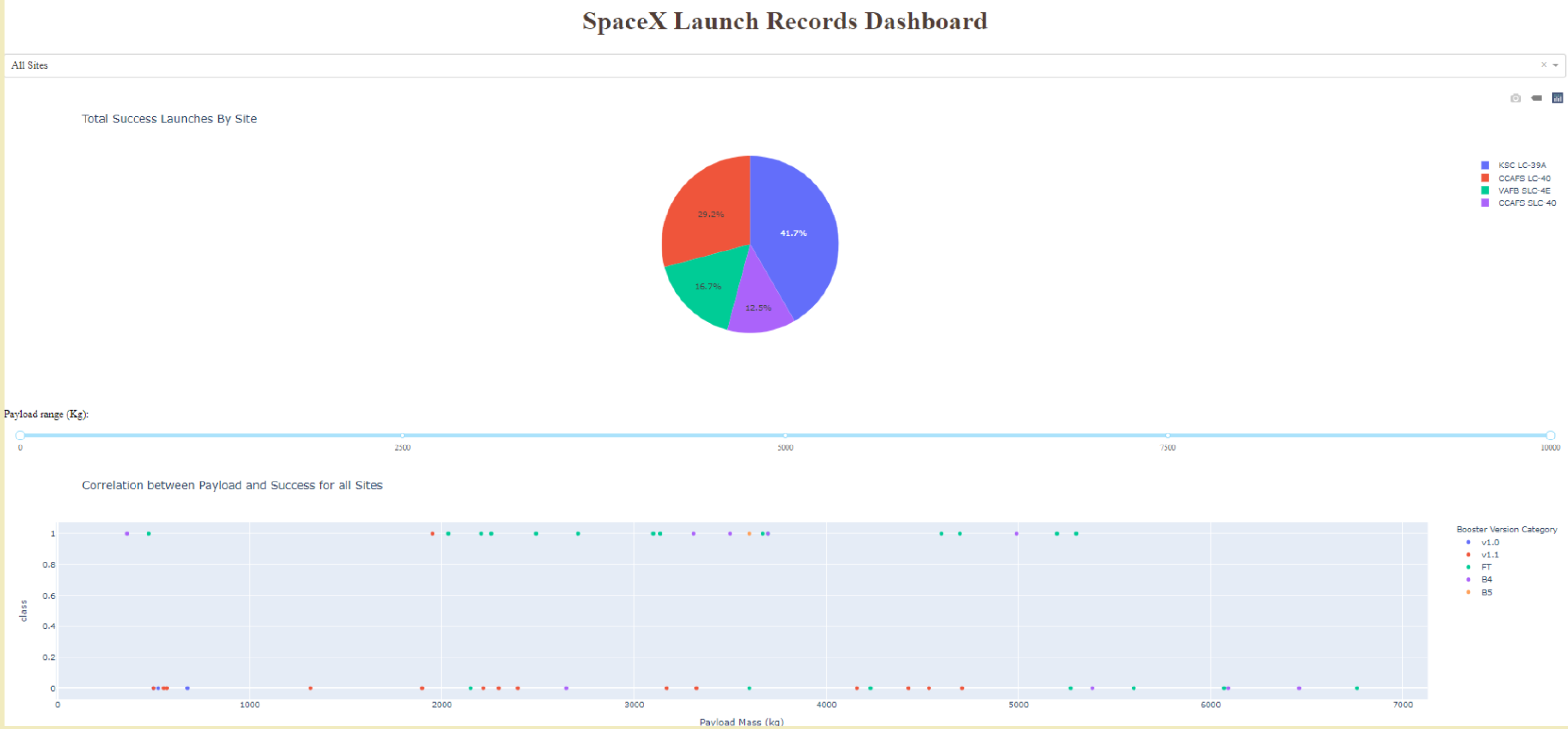
9. Generate a bar plot to visually compare the accuracy scores of different models.

10. Export the evaluation results, including model performance metrics, to a CSV file

Github Link

https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction.ipynb

Results:



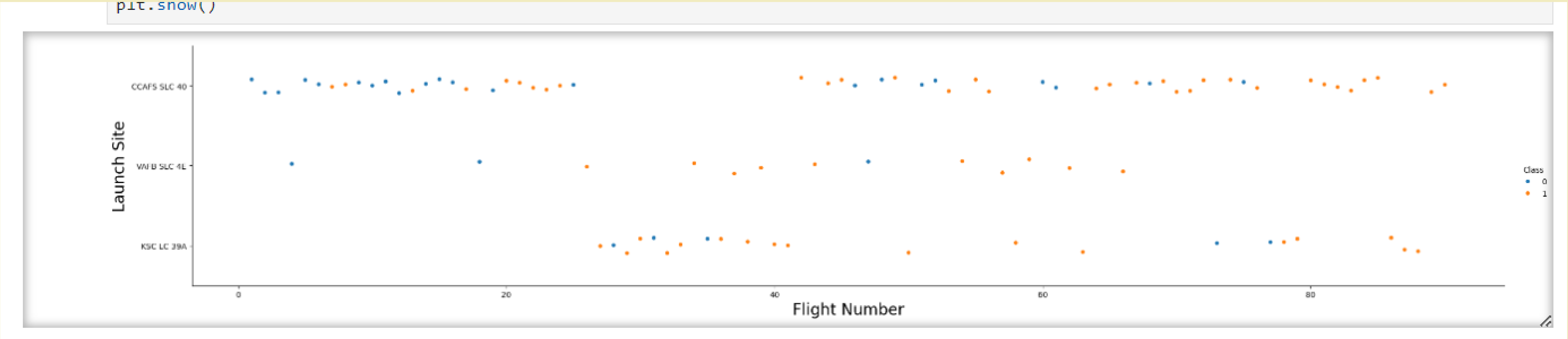
Plotly dashboard

In the following slides:

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

EDA with Visualization

Flight Number vs. Launch Site



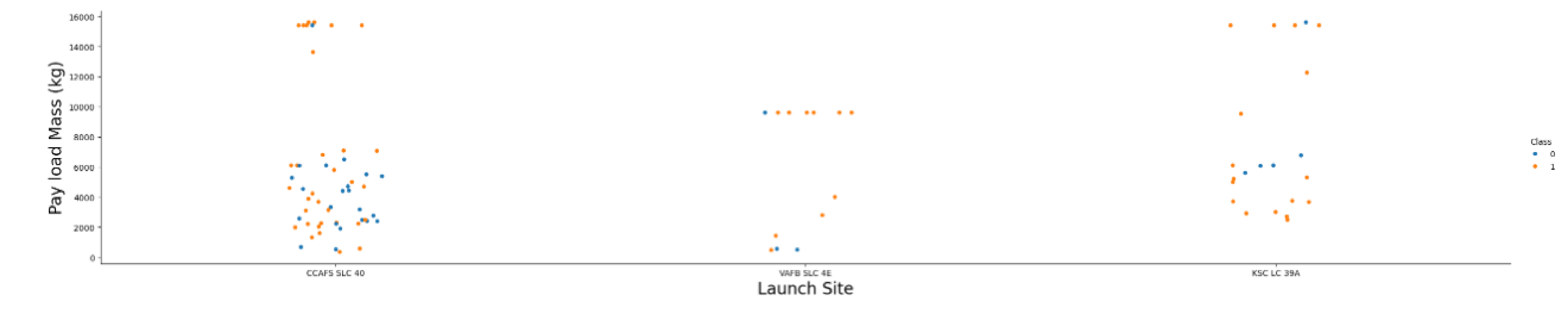
```
plt.show()
```

Explanation: The graphic shows a significant increase in success rates over time, particularly around flight number 20, indicating a major breakthrough. Early flights all failed, while recent flights all succeeded, reflecting technological improvements.

- **CCAFS SLC 40:** Handles most of all launches.
- **VAFB SLC 4E and KSC LC 39A:** These sites have higher success rates.

Overall, each new launch shows higher success, highlighting continuous advancements in launch capabilities.

Payload vs. Launch Site



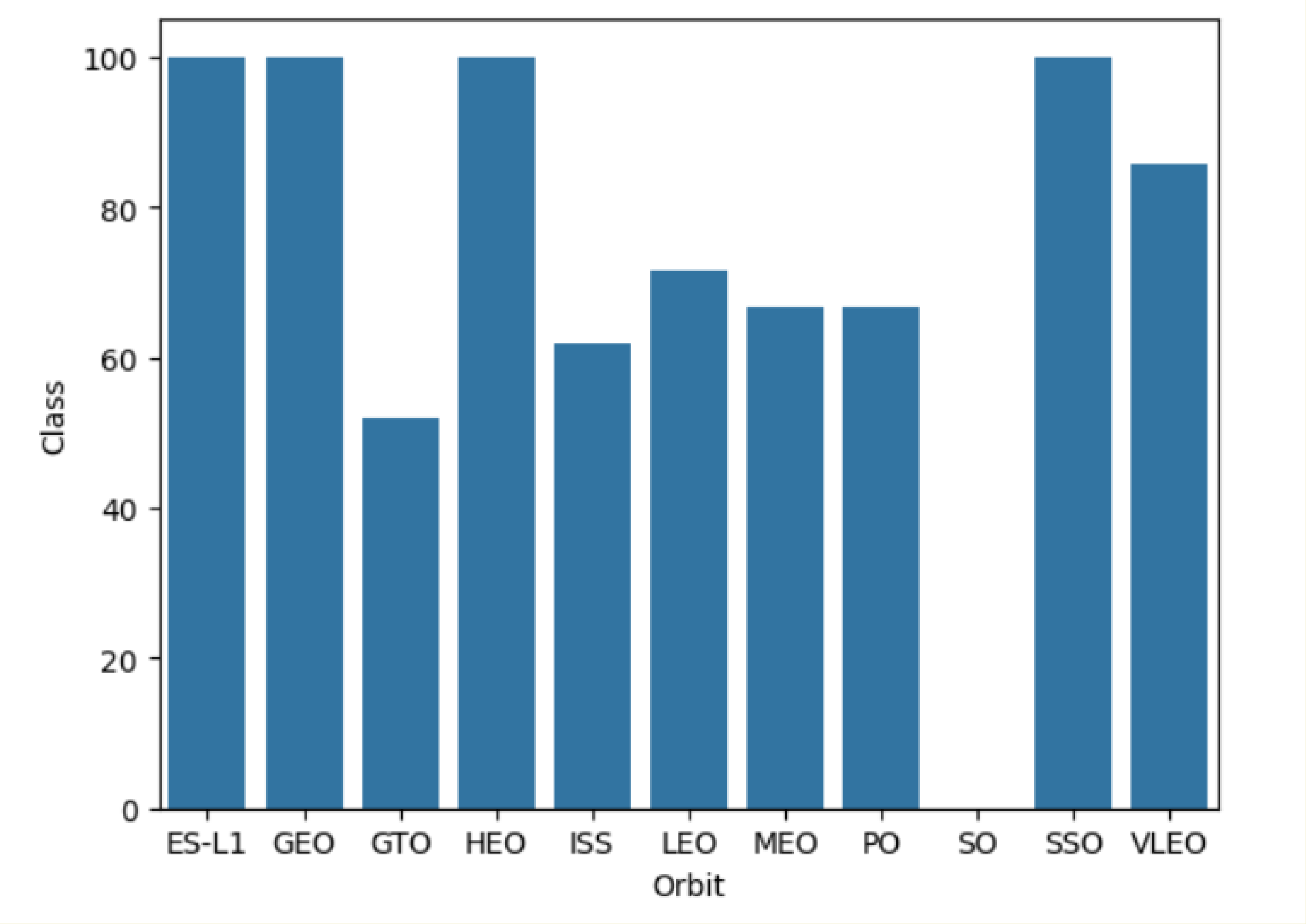
Explanation:

- •The graphic indicates that payload mass mostly falls between 0-6000 kg, with different launch sites using varying payload masses.
- •Higher Payload Mass Correlation: Generally, higher payload mass correlates with higher success rates.
- •Overall, different launch sites show varying payload capacities and success rates, with higher payloads generally indicating better outcomes.

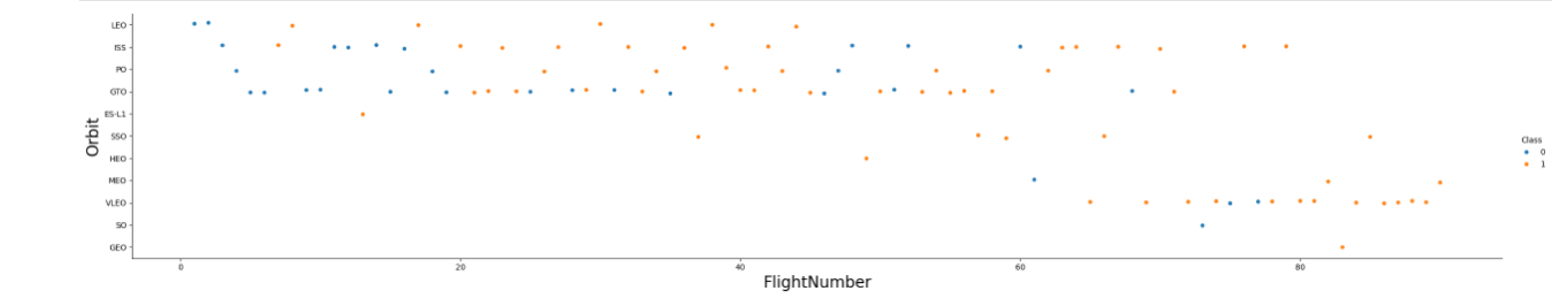
Success Rate vs. Orbit Type

The graphic shows success rates for different orbit types:

- **100% Success Rate:**
  - ES-L1, GEO, HEO, SSO
- **0% Success Rate:**
  - SO
- **Decent Success Rate (Around 85%):**
  - VLEO
- **Success Rate Between 50% and 75%:**
  - GTO, ISS, LEO, MEO, PO



Flight Number vs. Orbit Type

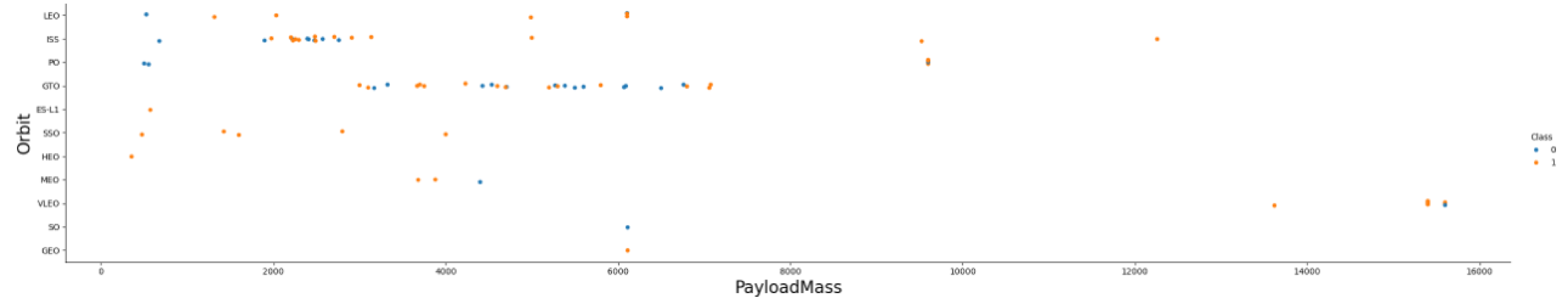


The graphic reveals:

- Higher success rates with increased flight numbers in LEO orbit.
- No clear link between flight number and success rate in GTO Orbit.
- Orbit preferences over time shifted with flight numbers, impacting outcomes.
  - Early launches targeted LEO orbits, achieving moderate success.
  - Recent launches focused on VLEO, showing improved results.
  - SpaceX exhibits better performance in lower or Sun-synchronous orbits (SSO).

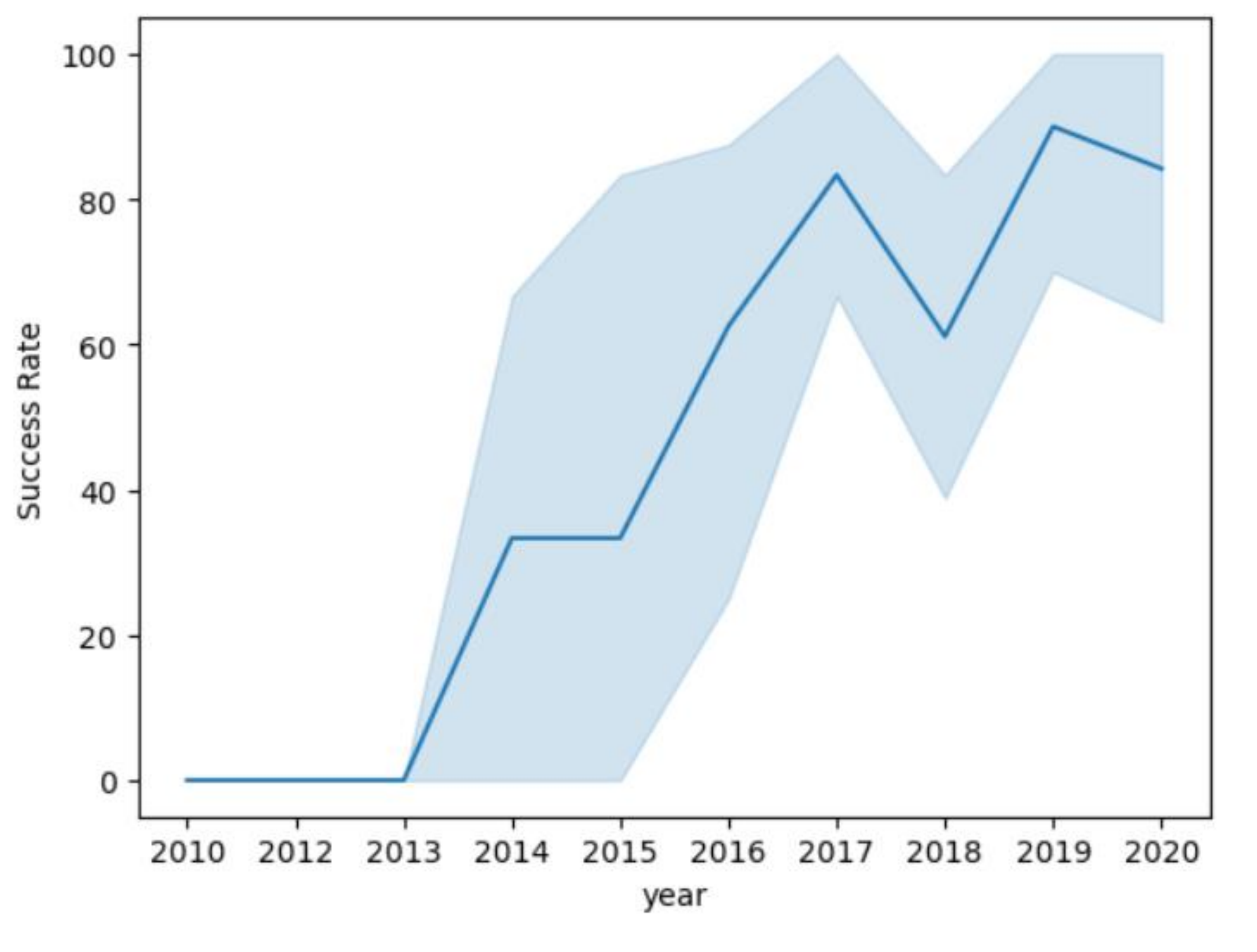
This suggests that SpaceX's success rates increase with more flights in LEO, and their overall success is greater in lower orbits.

Payload vs. Orbit Type



The graphic indicates:
•Payload mass is linked to orbit type.
•LEO and SSO Orbits typically have lower payload masses.
•VLEO Orbit shows high success with payload masses at the upper end of the range.
Overall, heavy payloads tend to have a positive influence on GTO and Polar LEO (ISS) orbits but can negatively affect GTO orbits. LEO and SSO typically handle lighter payloads, while VLEO manages heavier ones successfully.

Launch Success Yearly Trend

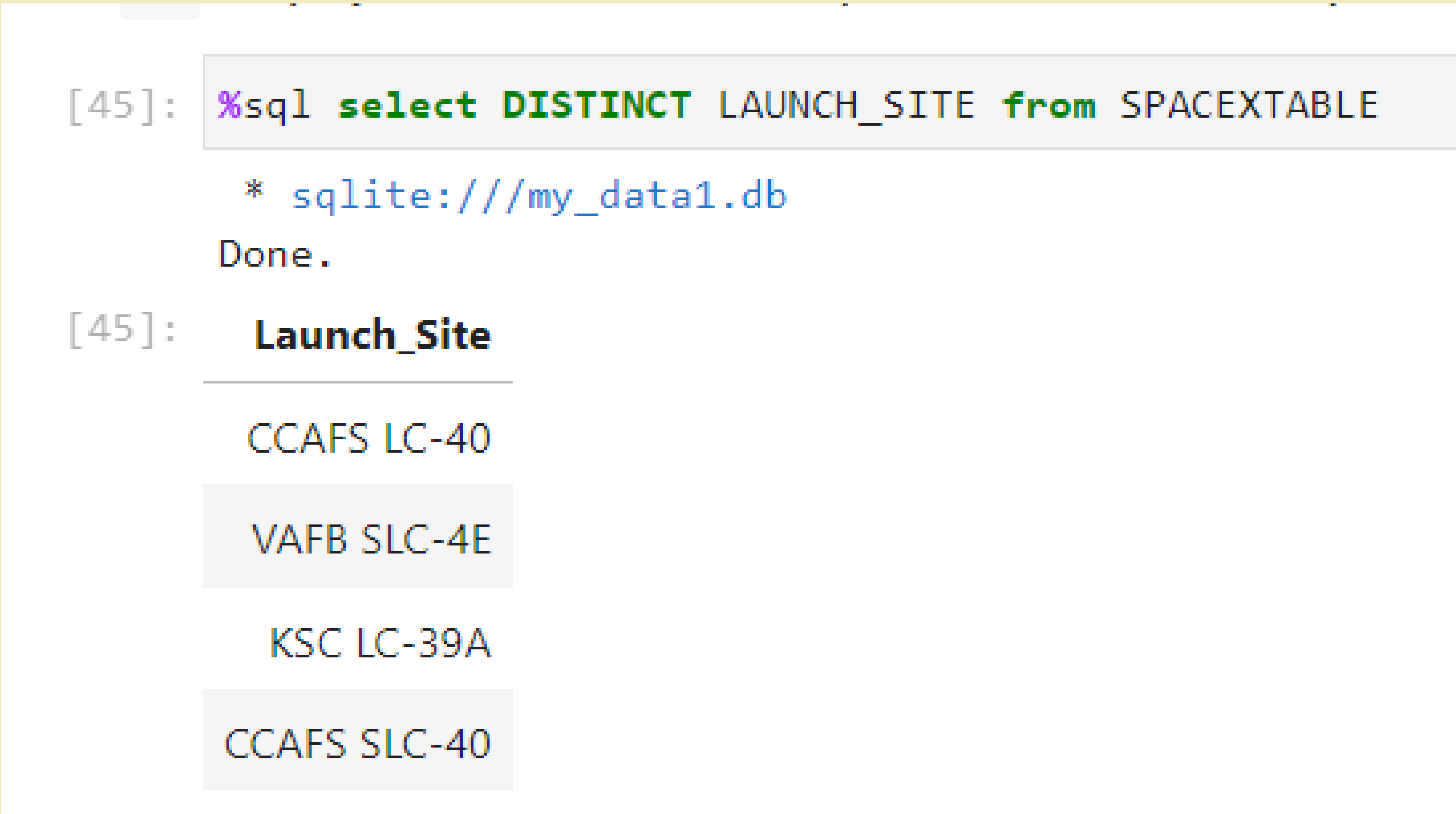


The graphic shows:

- Success rates generally increased over time since 2013.
- A slight decrease in success was observed in 2018.

EDA with SQL

All Launch Site Names



```
[45]: %sql select DISTINCT LAUNCH_SITE from SPACEXTABLE
```

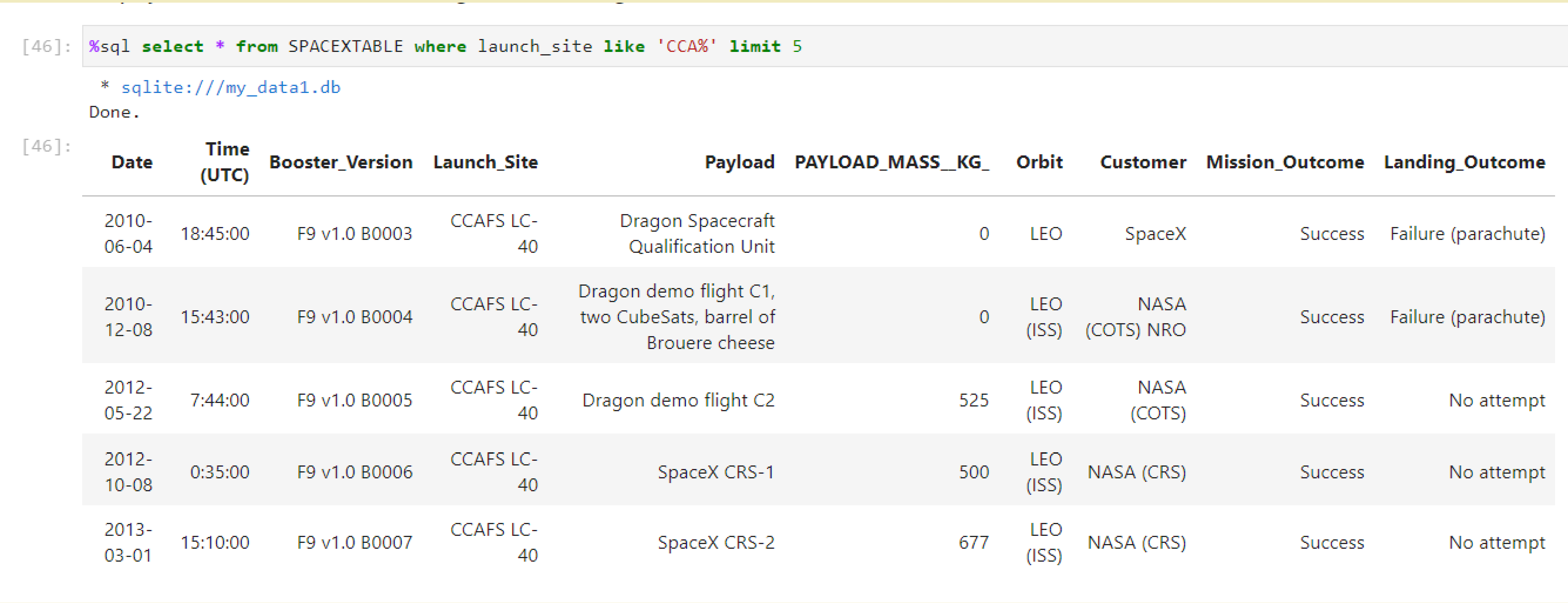
```
 * sqlite:///my_data1.db
Done.
```

[45]:

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

This query displays the names of the unique launch sites in the space mission.

Launch Site Names Beginning with ‘CCA’

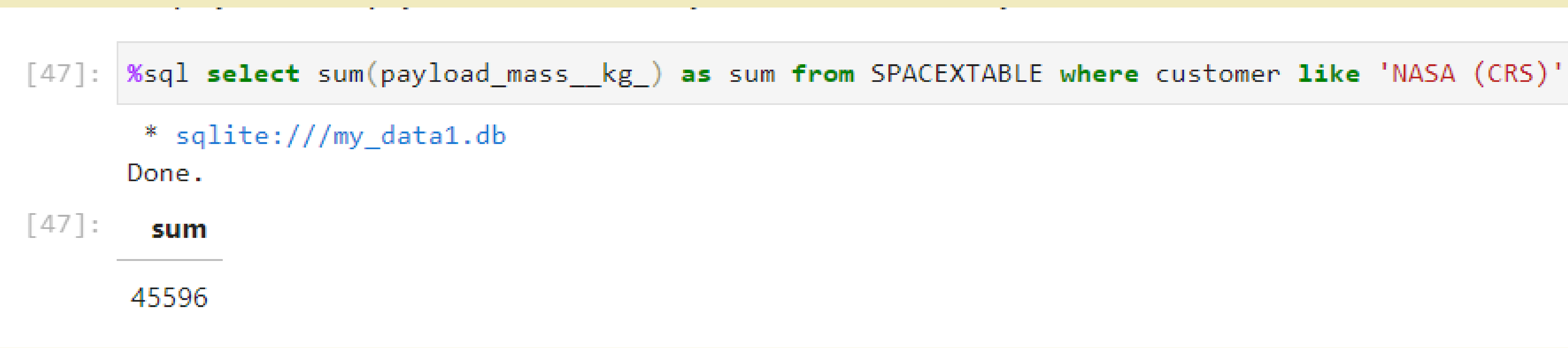


```
[46]: %sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
[46]:
```

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

This query displays the first five records in the database where the Launch Site name starts with ‘CCA’.

Total Payload Mass



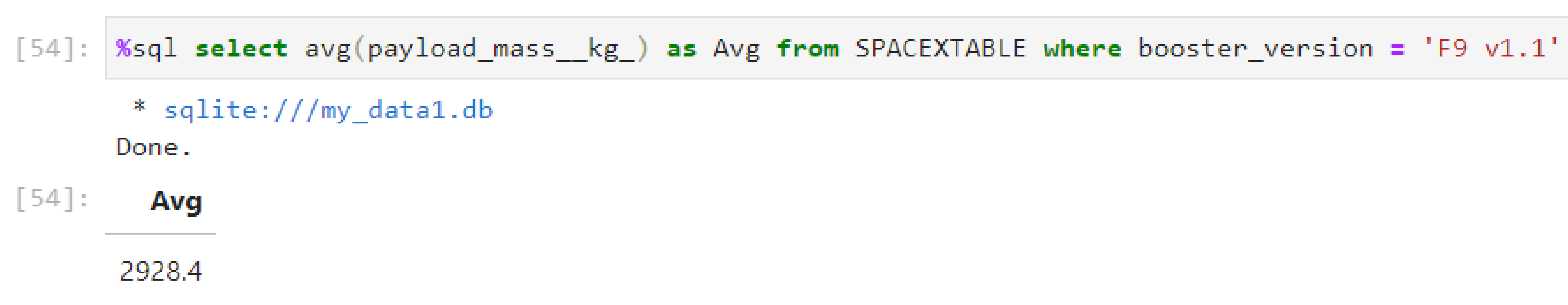
```
[47]: %sql select sum(payload_mass__kg_) as sum from SPACEXTABLE where customer like 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.
[47]:  sum

       45596
```

This query calculates the total payload mass in kilograms carried by boosters launched by NASA (CRS) which is **45596 kg** in this case.

Average Payload Mass by F9 v1.1



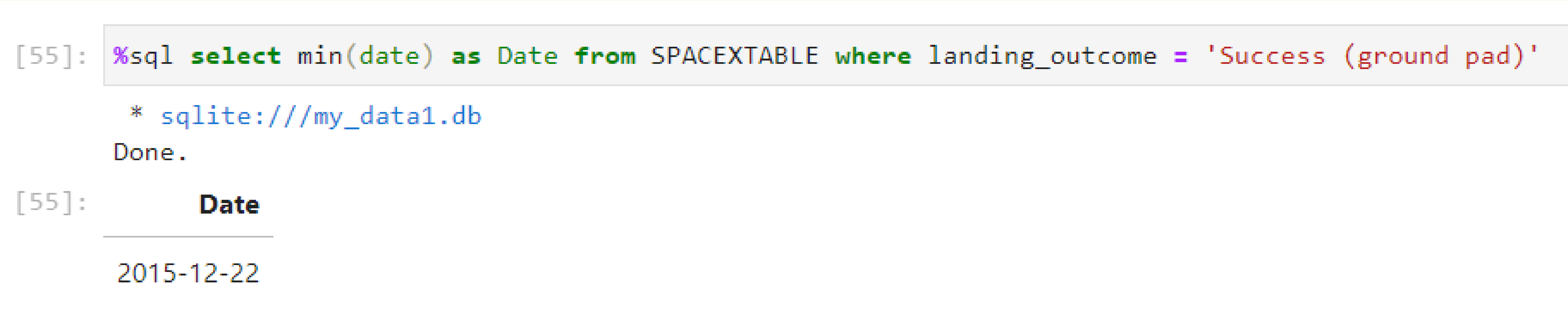
```
[54]: %sql select avg(payload_mass__kg_) as Avg from SPACEXTABLE where booster_version = 'F9 v1.1'

 * sqlite:///my_data1.db
Done.

[54]:   Avg
      ------
      2928.4
```

Calculates the average payload mass for launches using booster version F9 v1.1. which is **2928.4 kg** in this case

First Successful Ground Landing Date



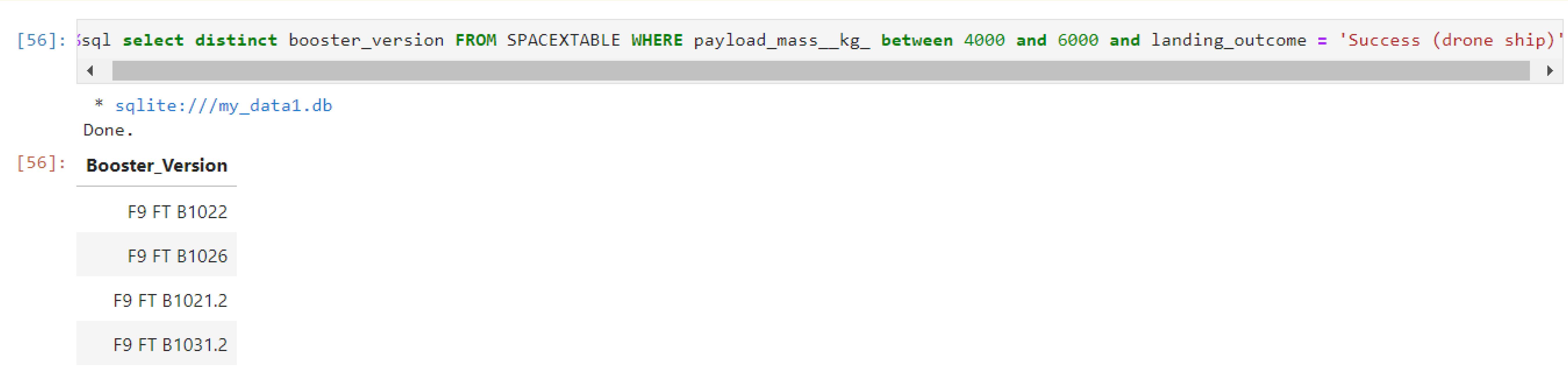
```
[55]: %sql select min(date) as Date from SPACEXTABLE where landing_outcome = 'Success (ground pad)'

 * sqlite:///my_data1.db
Done.
[55]:       Date

      2015-12-22
```

This query displays the date of the first successful ground pad landing, which occurred on the 22nd **of December in the year 2015.**

Successful Drone Ship Landing with Payload between 4000 and 6000



```
[56]: ;sql select distinct booster_version FROM SPACEXTABLE WHERE payload_mass__kg_ between 4000 and 6000 and landing_outcome = 'Success (drone ship)'
```

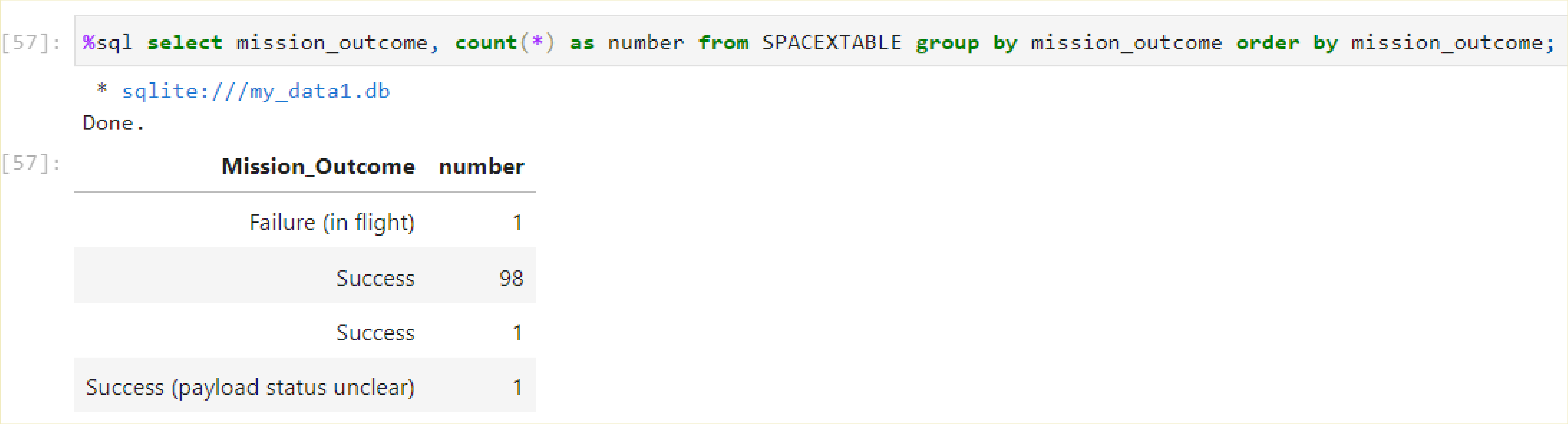
```
 * sqlite:///my_data1.db
Done.
```

[56]:

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

This query displays the four booster versions that had successful drone ship landings and a payload mass between 4000 kg and 6000 kg.

Total Number of Successful and Failure Mission Outcomes



```
[57]: %sql select mission_outcome, count(*) as number from SPACEXTABLE group by mission_outcome order by mission_outcome;
```

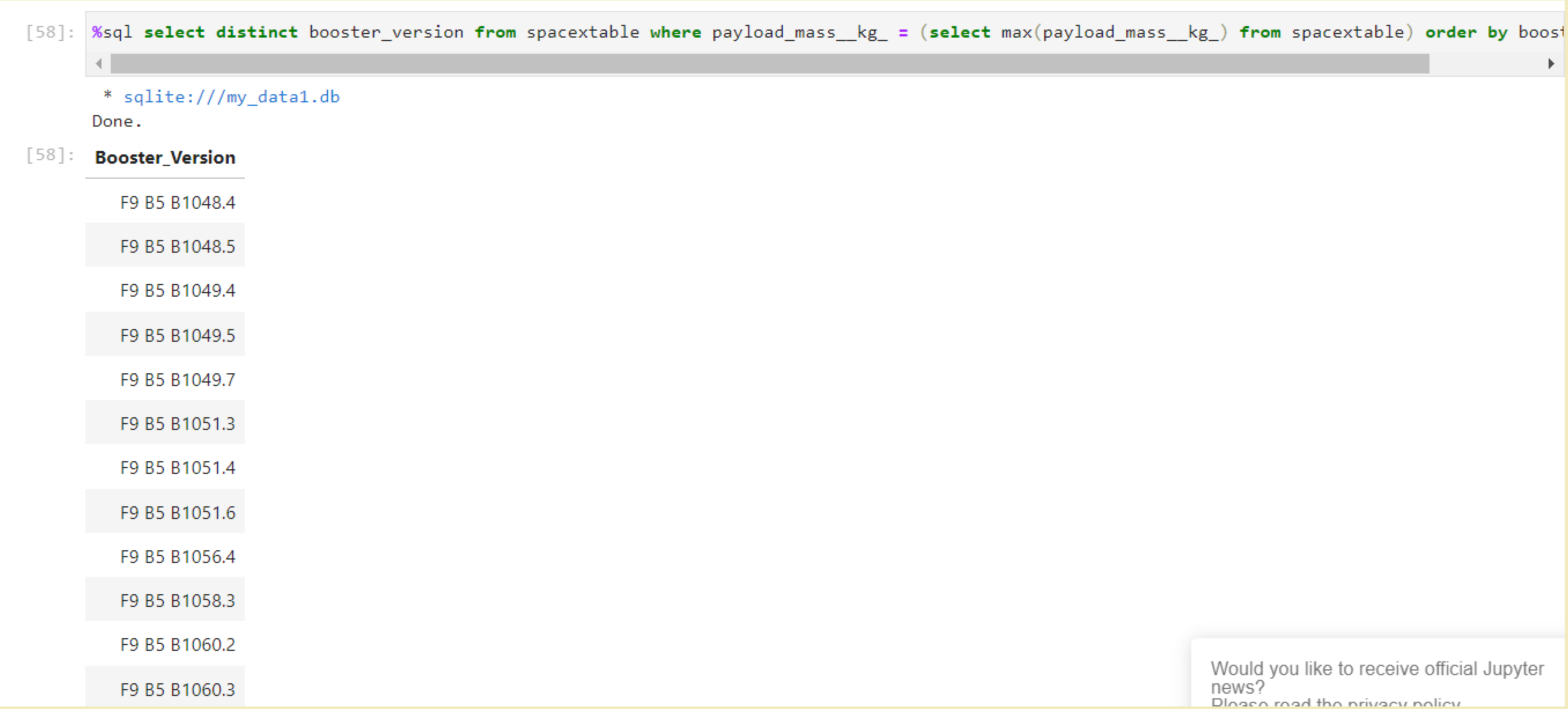
```
 * sqlite:///my_data1.db
Done.
```

[57]:

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

This query counts the total number of successful and failed mission outcomes, showing SpaceX achieves its mission outcome nearly 99% of the time.

Boosters which have carried minimum load



```
[58]: %sql select distinct booster_version from spacextable where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextable) order by boost
```

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

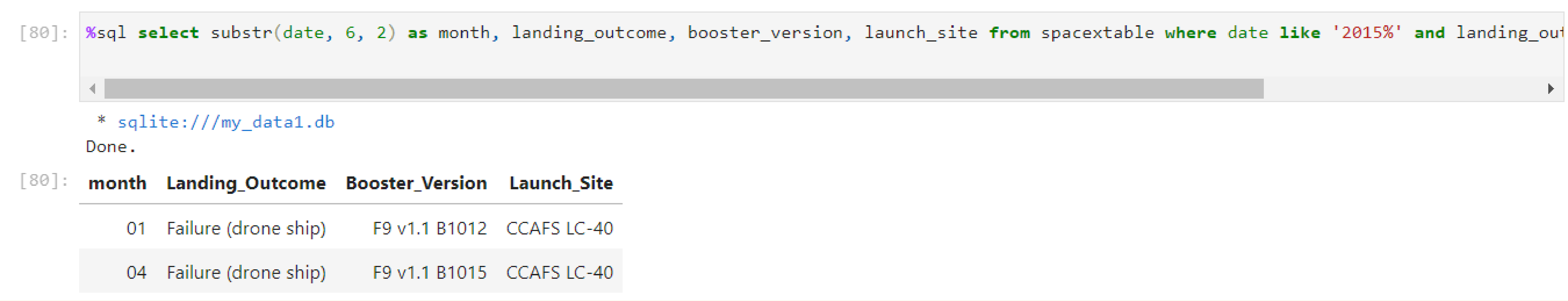
[58]:

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

Would you like to receive official Jupyter news?
Please read the privacy policy

This query displays booster versions carrying the highest payload mass,

2015 Launch Records



```
[80]: %sql select substr(date, 6, 2) as month, landing_outcome, booster_version, launch_site from spacextable where date like '2015%' and landing_out
```

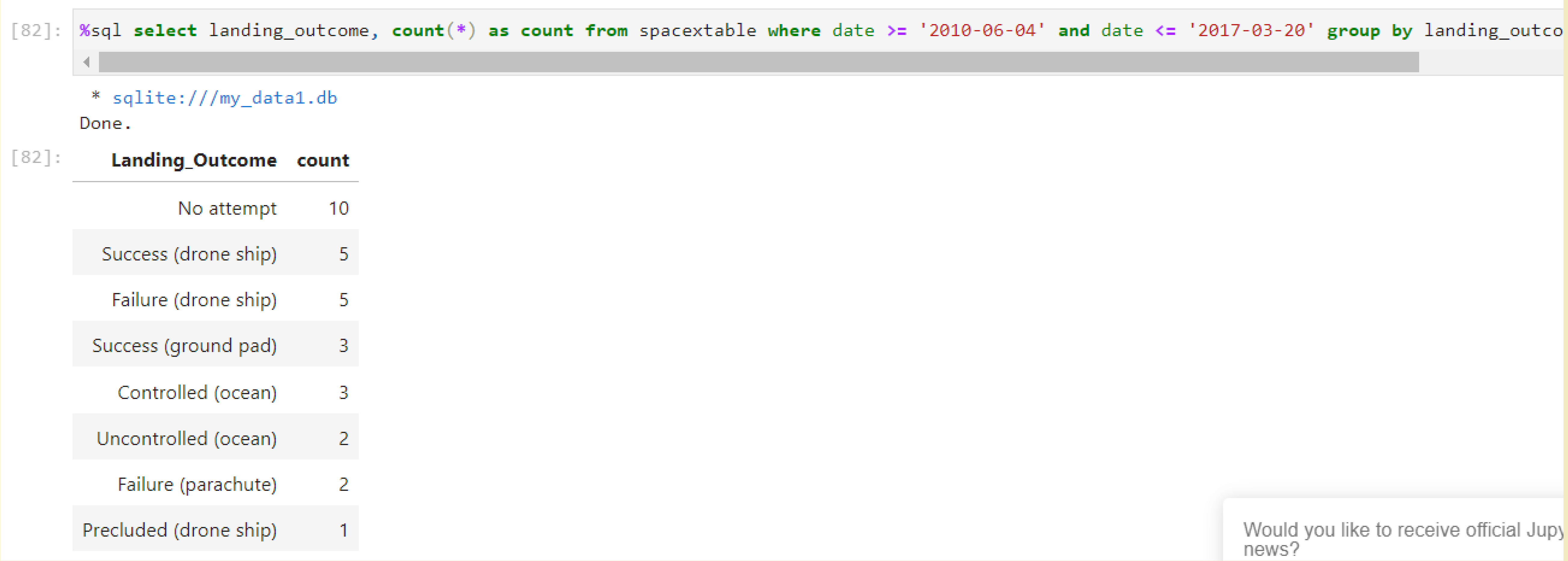
```
 * sqlite:///my_data1.db
Done.
```

[80]:

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

This query displays the two 2015 launches where stage 1 failed to land on a drone ship, showing the month, landing outcome, booster version, payload mass (kg), and launch site.

Rank success count between 2010-06-04 and 2017-03-20



```
[82]: %sql select landing_outcome, count(*) as count from spacextable where date >= '2010-06-04' and date <= '2017-03-20' group by landing_outco
```

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

[82]:

Landing_Outcome	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

Would you like to receive official Jupy
news?

This query sorts the number of landing outcomes between June 4, 2010, and March 20, 2017, in descending order.

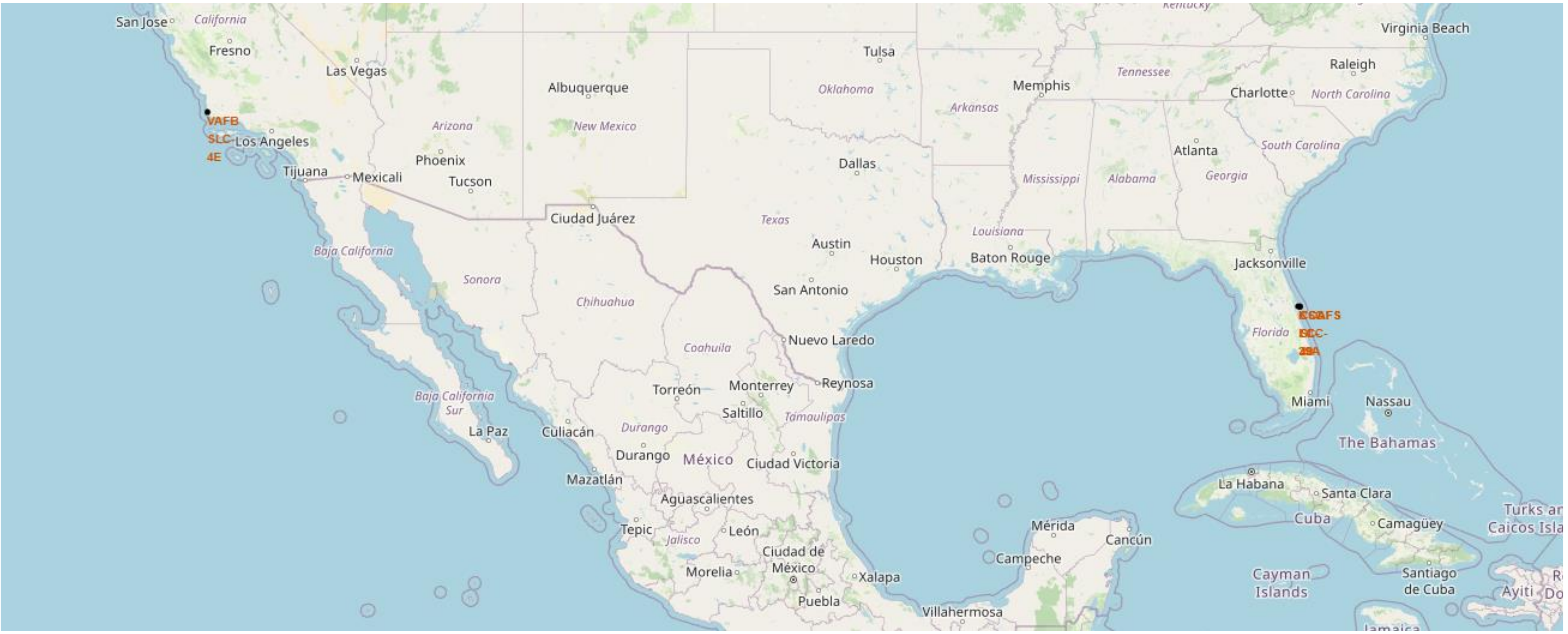
Interactive map with Folium

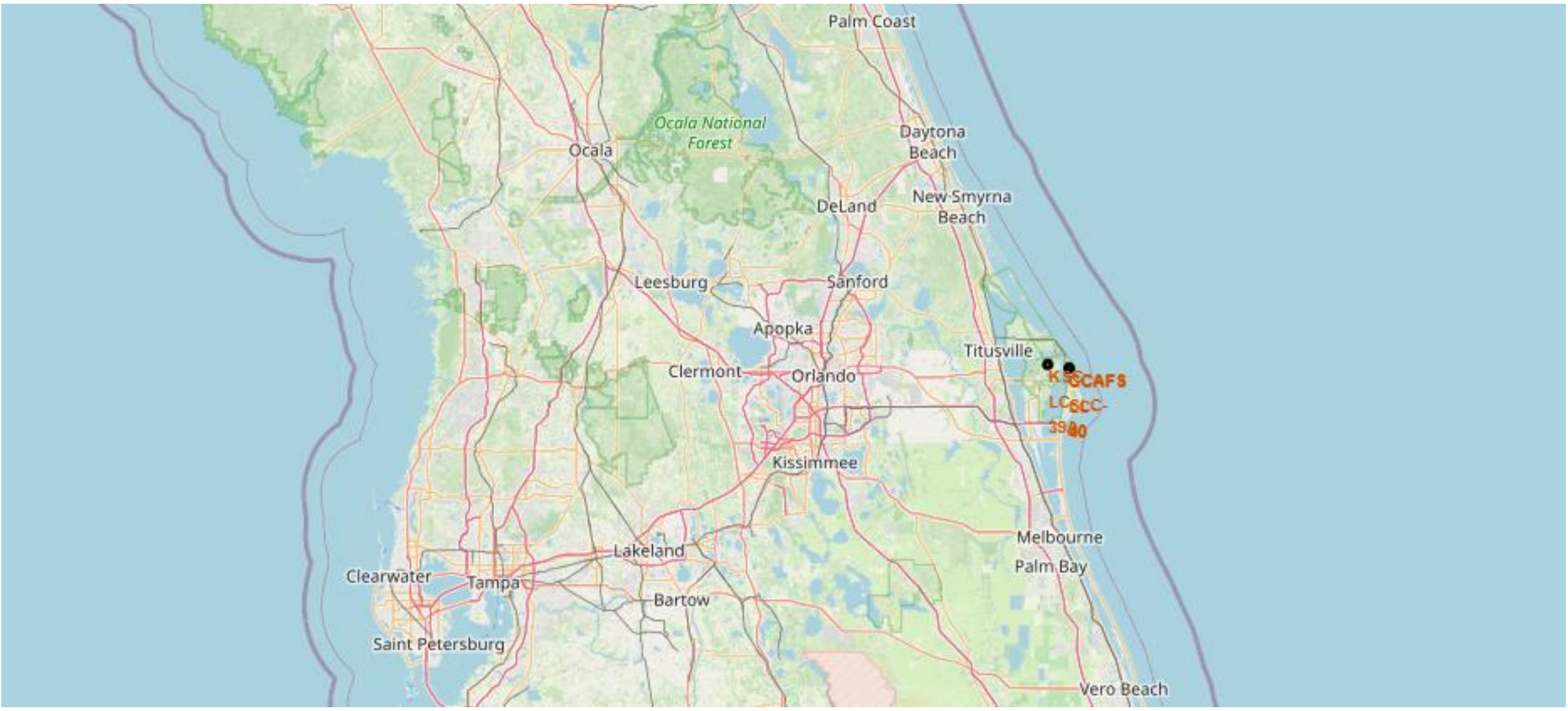
Folium

All launch site locations

Explanation:

The launch sites are typically situated near coastlines to minimize risks associated with debris and ensure safety by directing rocket trajectories over the ocean.

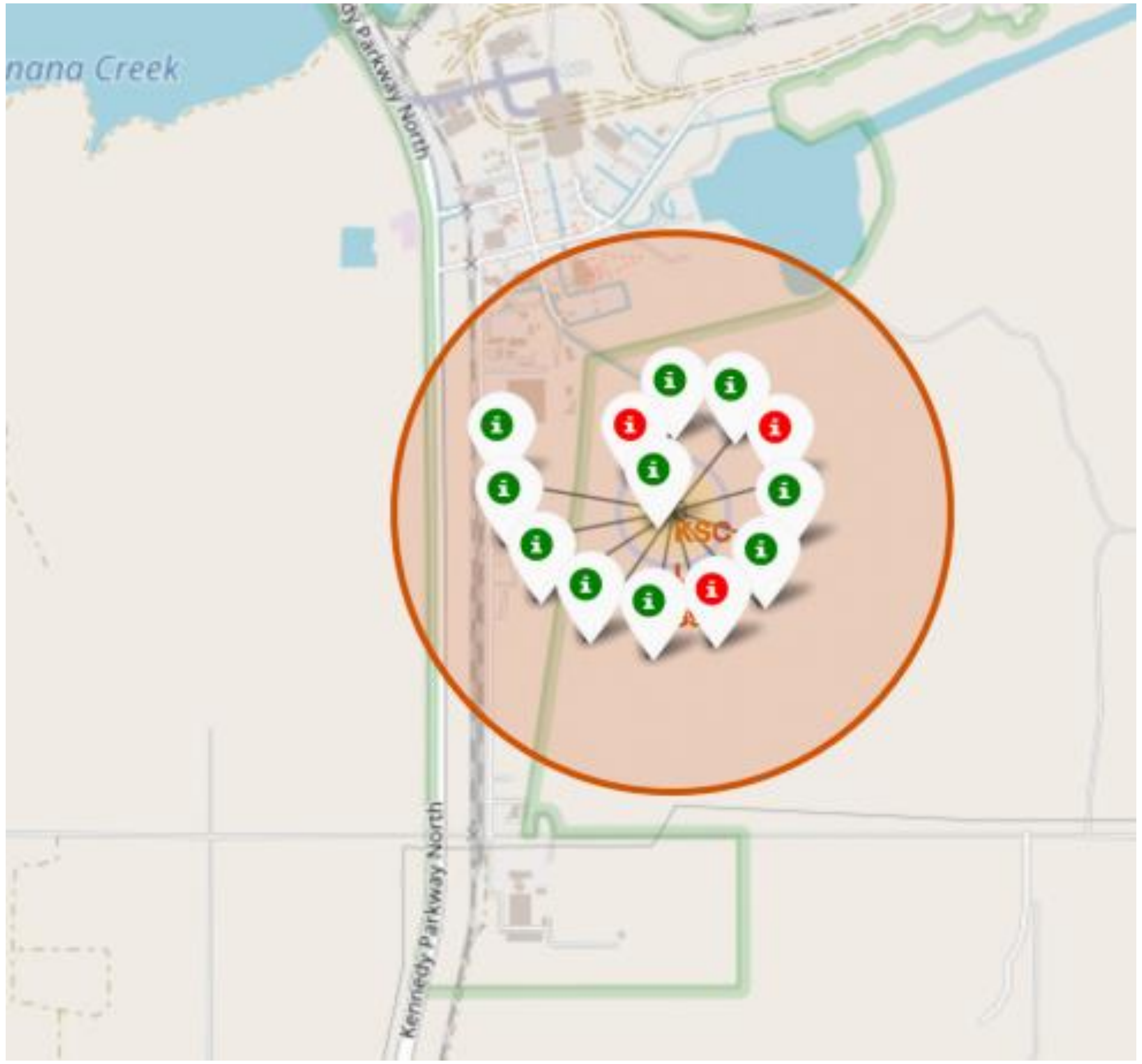




Colour-labelled launch Outcomes

Explanation:

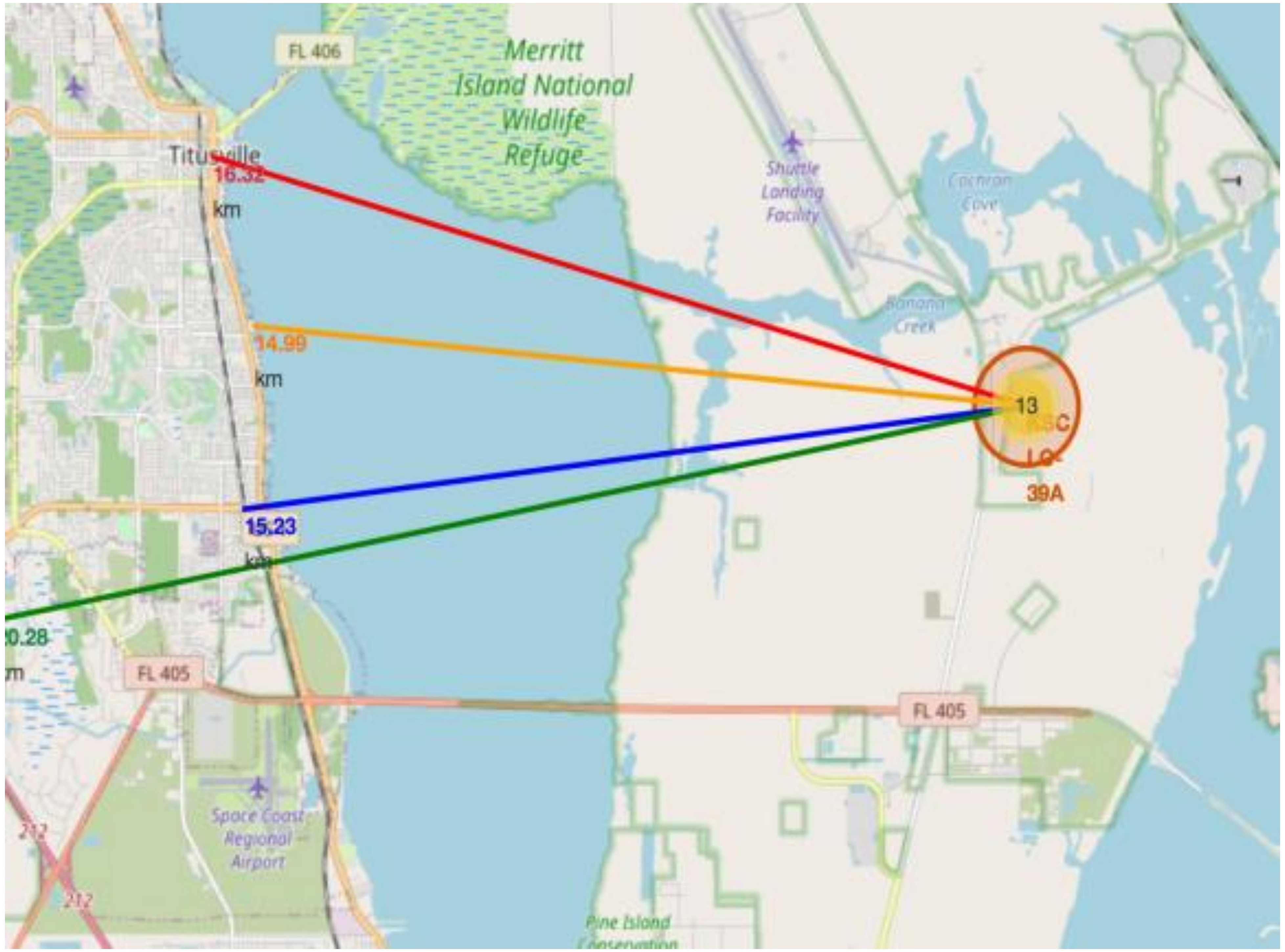
The Folium map displays clusters that can be clicked to show successful (green icon) and failed (red icon) landings for each launch site. Green markers indicate high success rates, such as KSC LC-39A.



Distance from the KSC LC-39A launch site to its proximities

Explanation:

KSC LC-39A is strategically positioned:
Close to railway (15.23 km) and highway (20.28 km) for logistical support.
Near coastline (14.99 km) to ensure launch failures can safely land in the ocean.
Proximity to Titusville (16.32 km) poses a potential risk from high-speed rocket failures, emphasizing safety measures.



Build a Dashboard with Plotly Dash

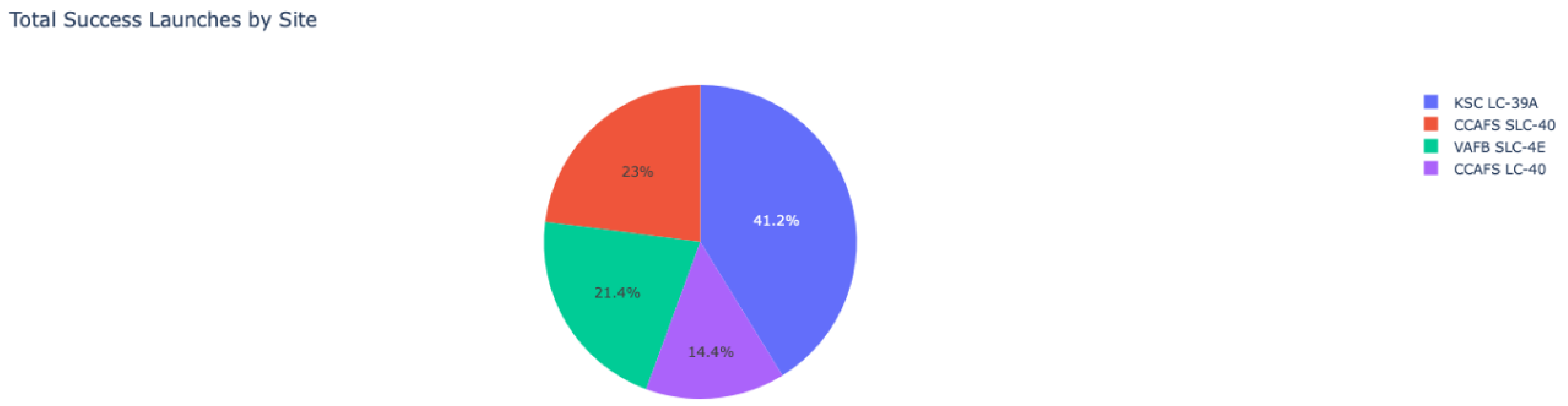
Dash

by plotly

Launch success count for all sites

Explanation:

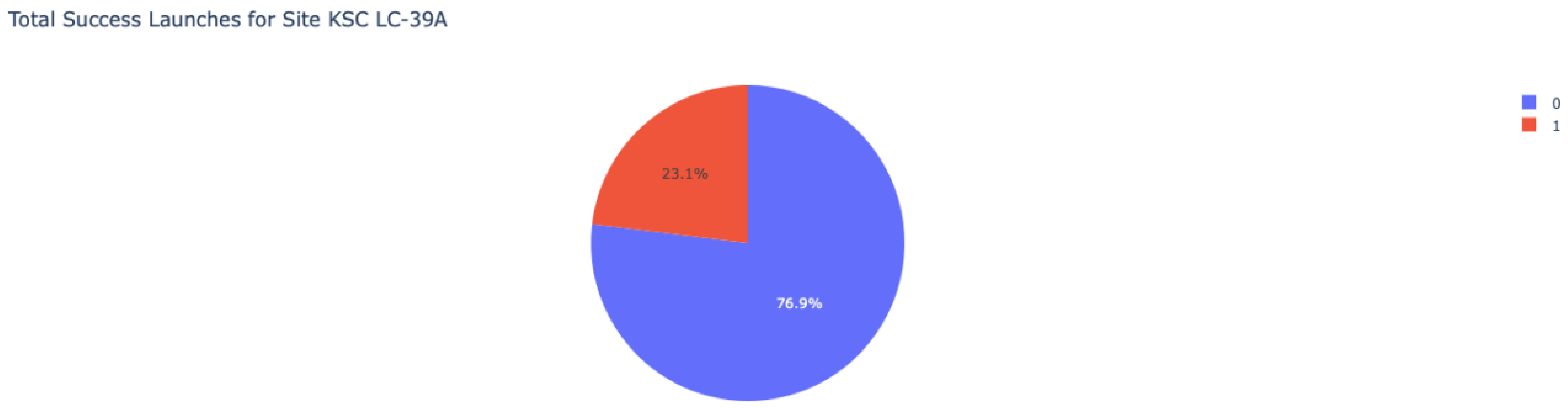
- KSC LC-39A stands out with the highest number of successful launches among all sites.
- CCAFS SLC-40, and VAFB SLC-4E have almost an equal number of successful landings.
- CCAFS LC-40 has the smallest share of successful landings.


Total Success Launches by Site

Launch site with highest launch success ratio

Explanation:

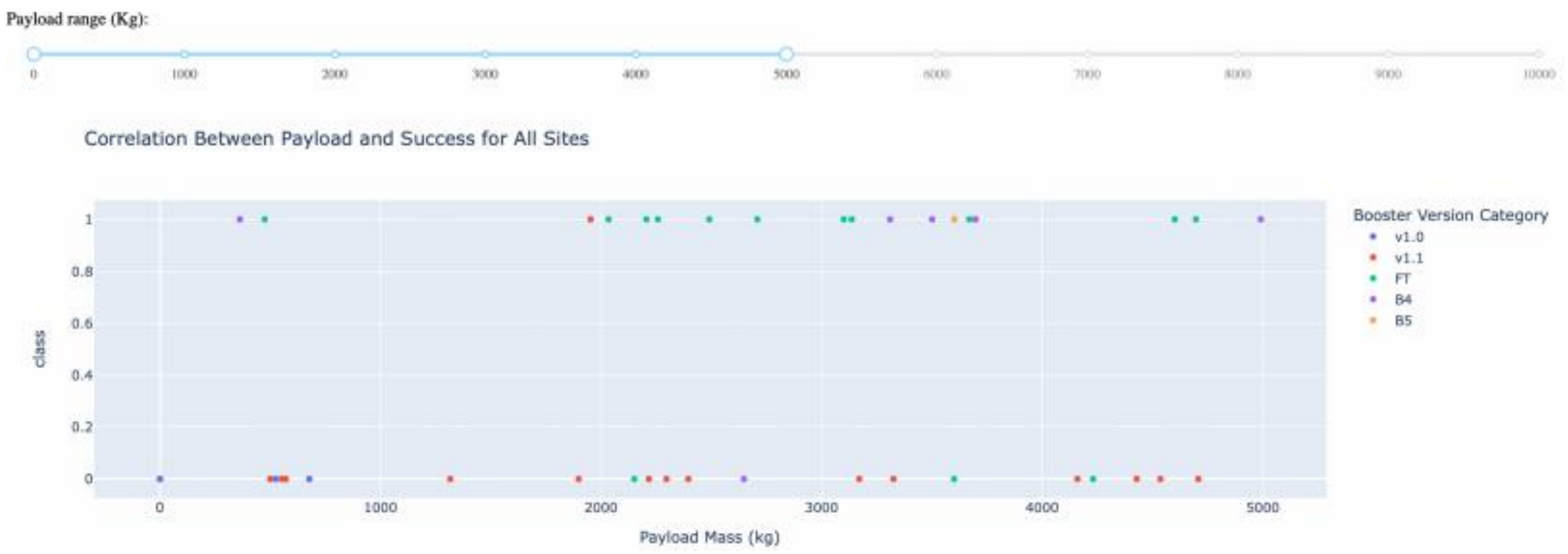
KSC LC-39A achieves the highest launch success rate, with 10 successful landings out of 13 attempts, marking a success rate of 76.9%.



Payload Mass vs. Launch Outcome for all sites

Explanation:

The scatter plot indicates launch outcomes, with successful and failed landings differentiated by color and the size of data points reflecting the number of launches. Notably, payloads between 2000 and 5500 kg show the highest success rates, despite two failed landings with zero kg payloads observed within the 0-6000 kg range.

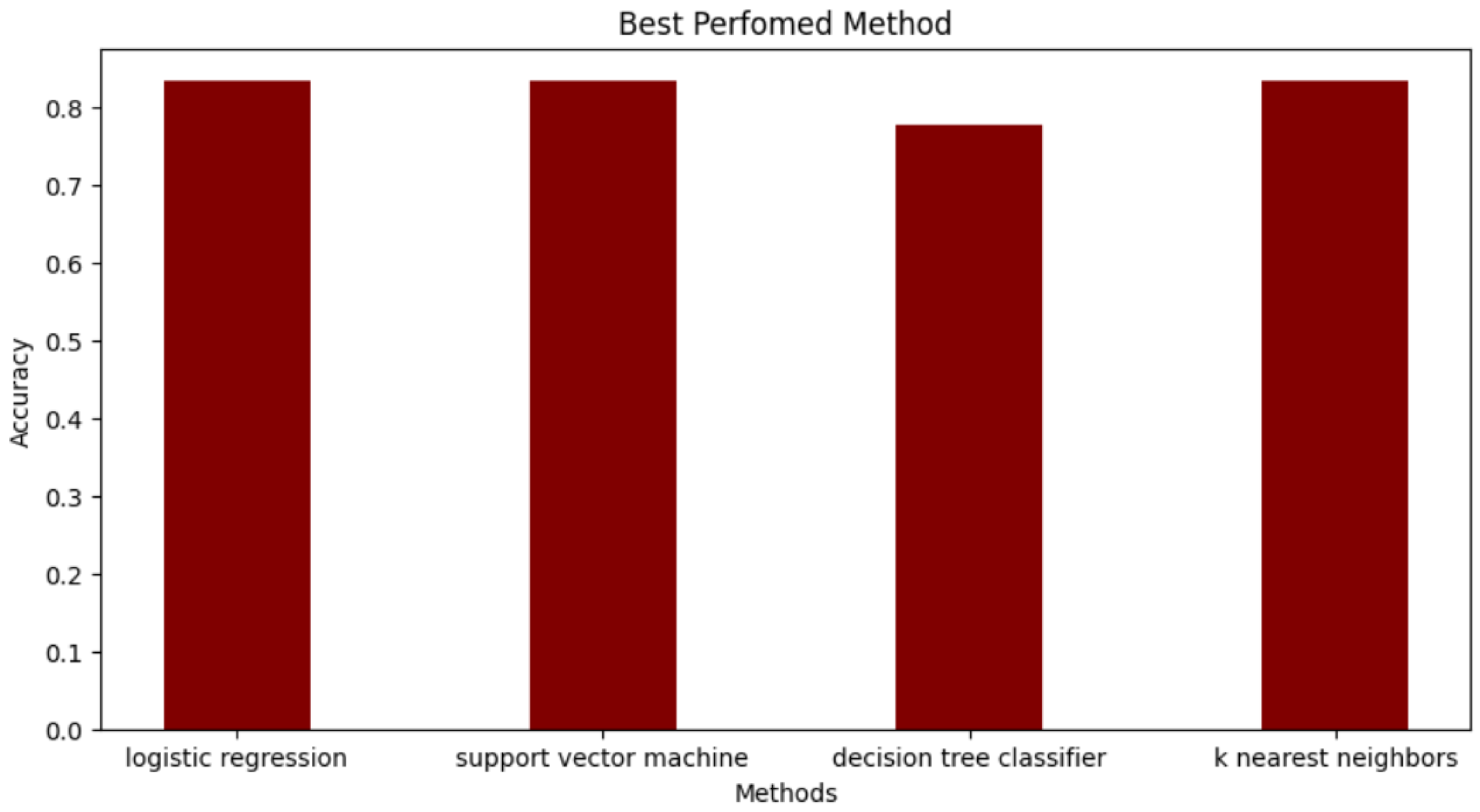


Predictive analysis (Classification)

Classification Accuracy

Explanation

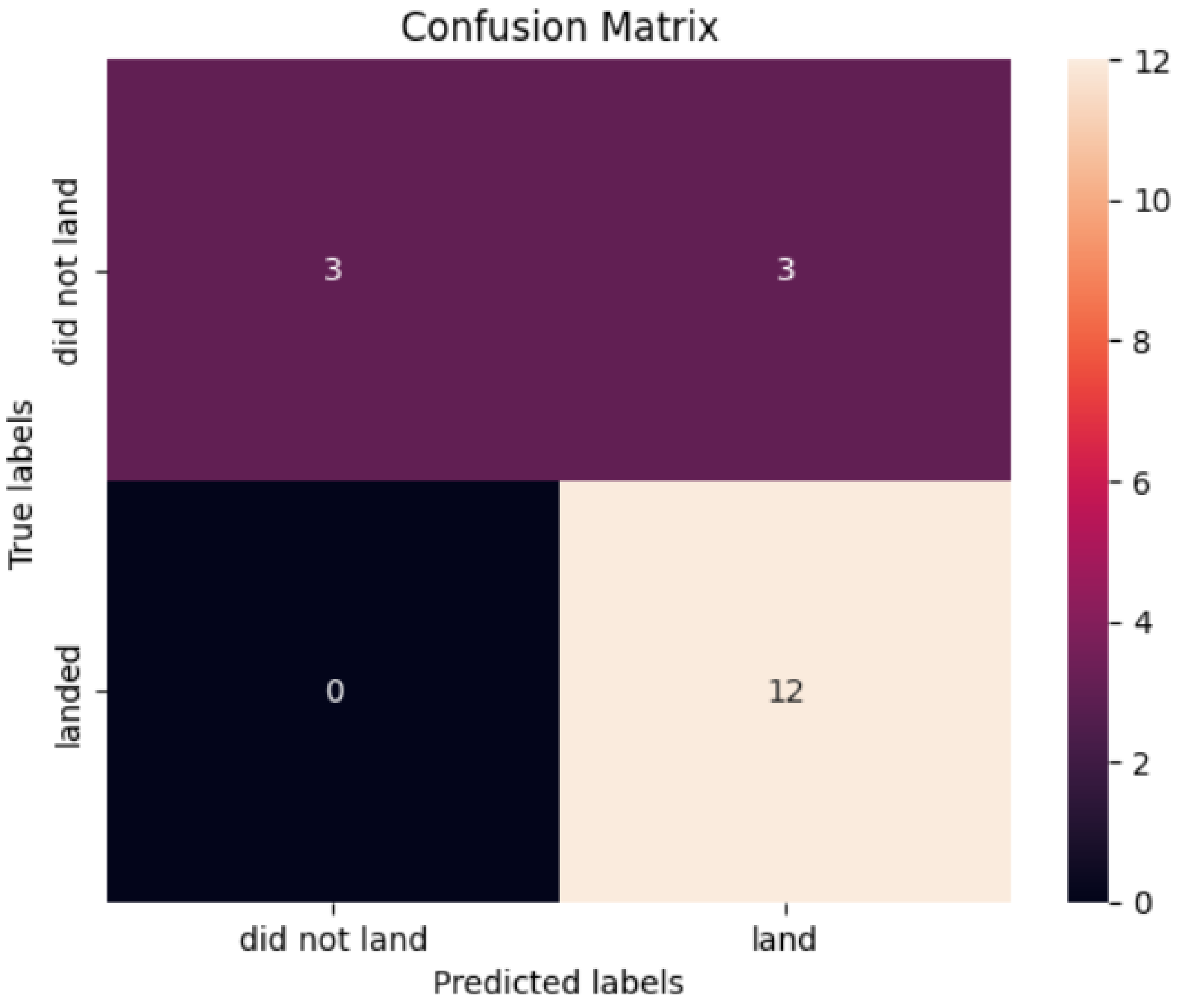
Despite various model architectures employed, all exhibited a comparable accuracy of 83.33% on the test set.
This limited sample size can lead to considerable fluctuations in accuracy metrics, as evidenced particularly in the Decision Tree Classifier across multiple iterations.
To establish robust model selection, a larger dataset would be indispensable for conclusive evaluations.



Explanation

In evaluating various models on a small test set, all demonstrated consistent 83.33% accuracy. Despite this uniform performance, the models exhibited a notable trend of overpredicting successful landings, particularly evident in false positives where they incorrectly identified successful outcomes when they were actually unsuccessful.
The models correctly predicted 12 successful landings and 3 unsuccessful landings based on their true labels. Additionally, they incorrectly predicted 3 successful landings when the true label indicated unsuccessful landings.

Confusion Matrix



Conclusions

- Our analysis, utilizing data from a public SpaceX API and scraping SpaceX's Wikipedia page, aimed to develop a machine learning model for Space Y to predict Stage 1 landing success, potentially saving ~$100 million USD per launch.
- The dataset revealed several insights: the Decision Tree Model emerged as optimal, showing superior performance.
- Launches with lower payload masses exhibited higher success rates, and sites near the Equator and coast had predominant usage.
- Over time, launch success rates showed improvement, notably with KSC LC-39A displaying the highest success rate.
- Specific orbits like ES-L1, GEO, HEO, and SSO achieved a perfect 100% success rate.
- Our model achieved an 83% accuracy, enabling reliable predictions for launch decisions.
- Future improvements include gathering more data to enhance model selection and predictive accuracy.

Appendix

Github Repository:
https://github.com/Satyendra2309/IBM-Applied-Data-Science-Capstone

Special thanks to all the INSTRUCTORS.