IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kihoon Kim
12/10/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection API and Data Collection with Web Scrapping
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Visualization
  - Interactive Visual Analytics and Dashboard
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis Results
  - Interactive Visual Analytics Results
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch
- Problems you want to find answers
  - How do different launch sites impact the success rate of Falcon 9 rocket landings?
  - Does the success rate of Falcon 9 rocket landings improve over time?
  - What is the relationship between payload mass and the landing success rate of the rocket?
  - Which machine learning model performs best in predicting Falcon 9 first stage landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data was collected from Space X API and Web Scraping from Wikipedia. It included diverse information like orbit type, launch site, landing outcomes and so on.
- Perform data wrangling
    - Filtering and preprocessed the data.
    - Cleaned and preprocessed the data to handle missing value.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Built classification models and tuned hyperparmeters

Data Collection

For data collection, I utilized the SpaceX API and Web Scraping from Wikipedia

- **SpaceX API**
  - The SpaceX API was used to gather data such as rocket launches, mission details, and payloads.
    The API provides structured data in JSON format, making it easy to process and analyze.
- **Web Scraping – Wikipedia**
  - Web scraping was performed on Wikipedia to extract necessary information. Tools like BeautifulSoup and HTTP requests were used to parse and collect data from the HTML content.

Data Collection – Web Scraping Process

1. Identify and Fetch Web Page Content – HTTP GET request
2. Parse HTML Content – BeautifulSoup
3. Extract Required Data
4. Clean and Structure Data
5. Filter and Save Final Data

Data Collection – SpaceX API

Step1: Request to the SpaceX API

- **HTTP GET request**

Step2: Parse API Response into DataFrame

- **json_normalize()**

Step3: Extract Relevant Details

Step4: Clean and Combine Data Frame

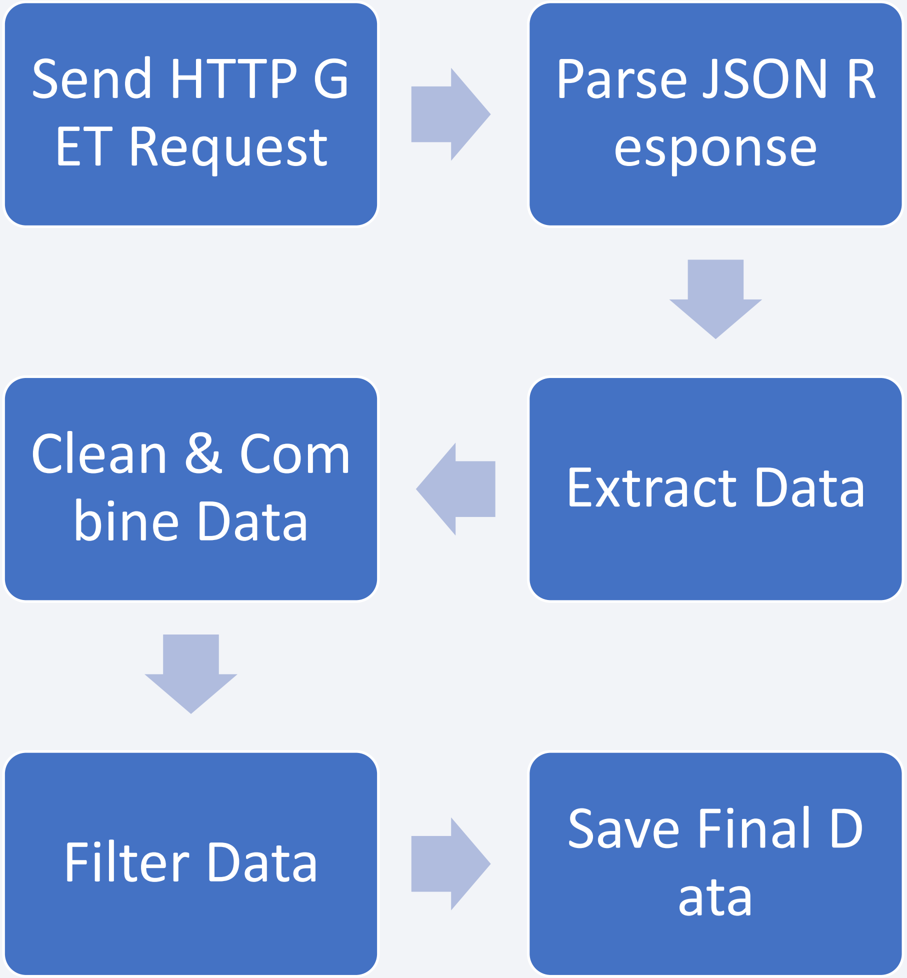
- **Remove unnecessary fields**

Step5: Final Data Wrangling

Step6: Filter and Save Final Data

- **CSV Format**

Link to Github



Data Collection - Scraping

Step1: Identify and Fetch Web Page Content

- **HTTP GET request**

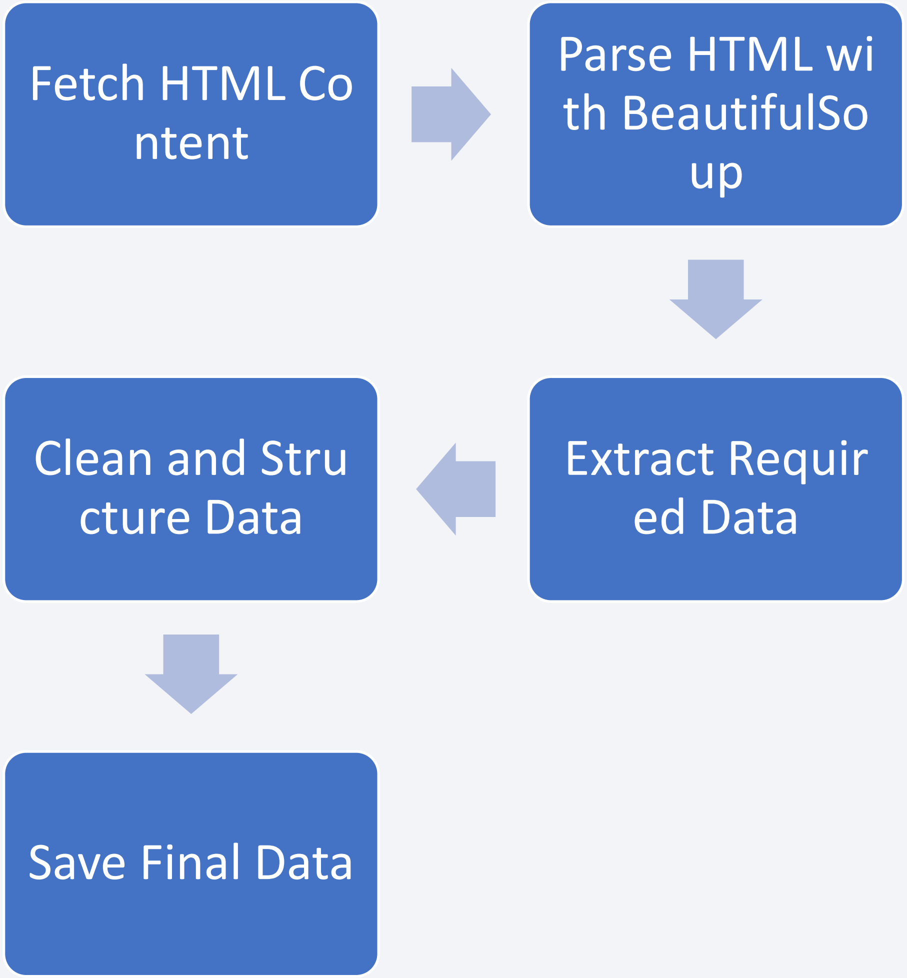
Step2: Parse HTML Content

- **BeautifulSoup**

Step3: Extract Required Data

Step4: Clean and Structure Data

Step5: Filter and Save Final Data



[Link to Github]

Data Wrangling

Step1:Data Loading and Exploration

- **Loaded CSV dataset using pands**
- **Performed EDA to identify missing value**

Step2: Handling missing value

- **Calculate percentage of missing value**

Step3: Feature Analysis

- value_counts() for LaunchSite

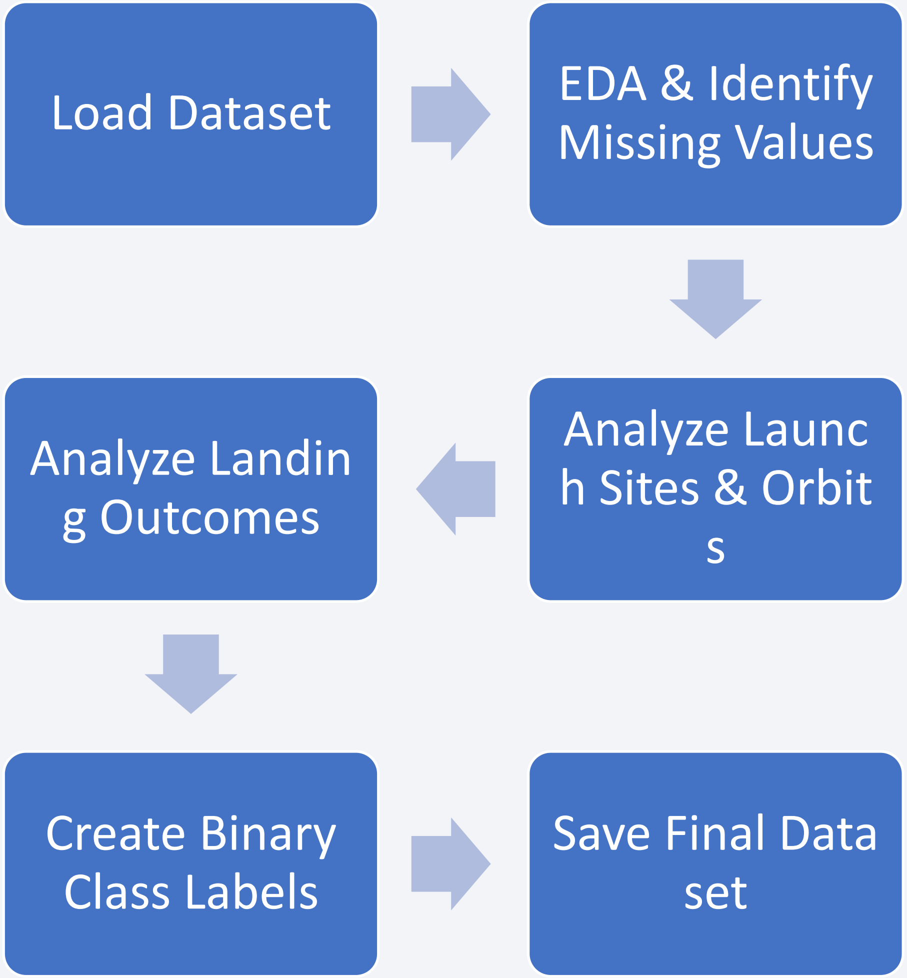
Step4: Outcome Analysis

- **Determine landing outcomes(Success of Unsuccess)**

Step5: Label Creation 0 and 1

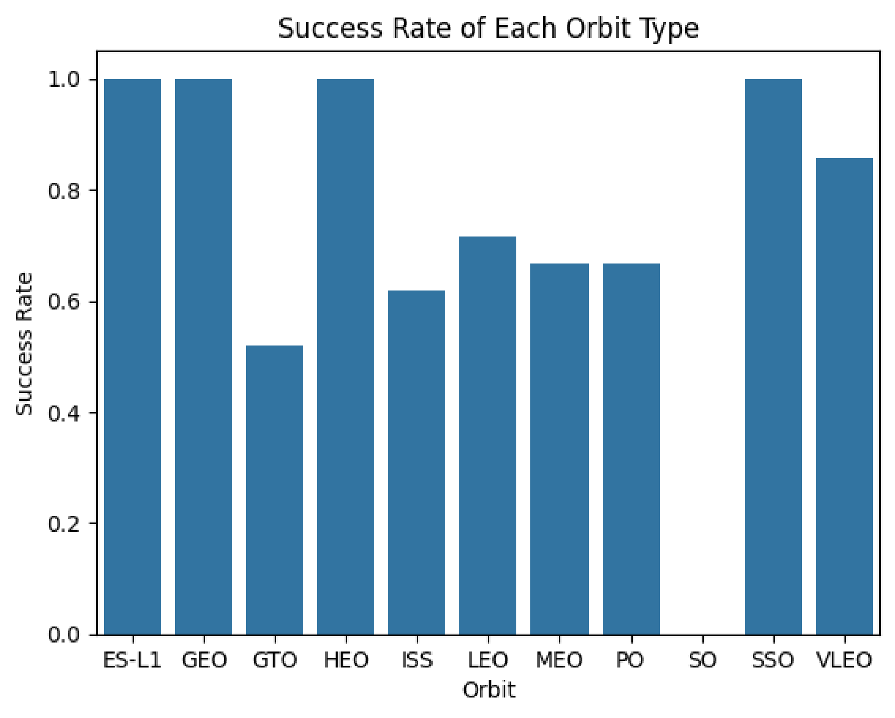
Step6: Export and Save Final Data

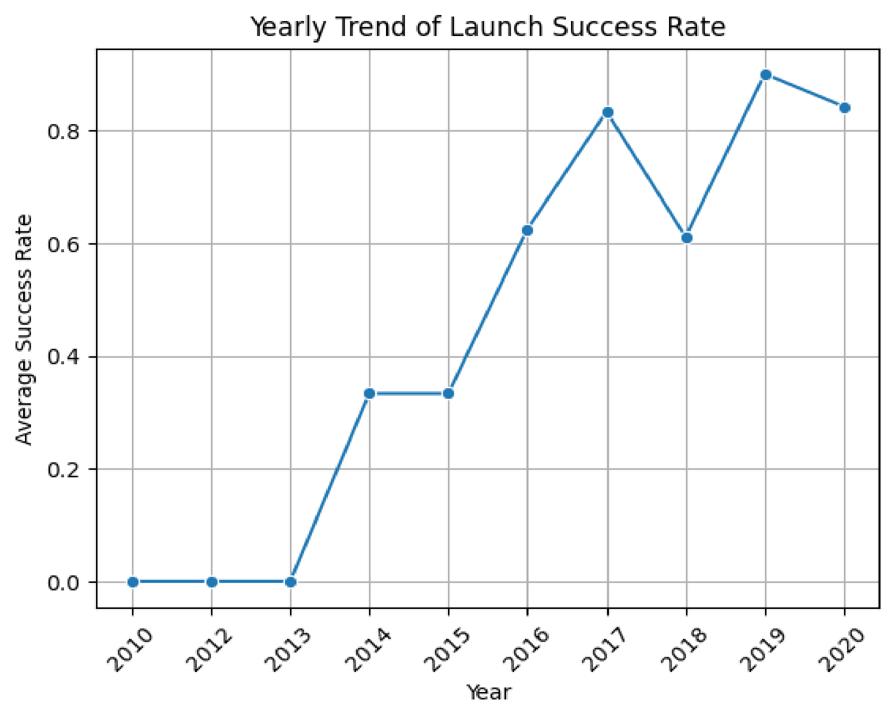
- **CSV Format**


Load Dataset
EDA & Identify Missing Values
Analyze Launc h Sites & Orbit s
Analyze Landin g Outcomes
Create Binary Class Labels
Save Final Data set

Link to Github

EDA with Data Visualization





- Chart Type: Line Chart
- Purpose: To show the yearly success rate trend
- **The steady upward trend from 2013 onward highlights SpaceX's continuous advancements in reusability and landing technologies, ultimately leading to cost efficiency and operational success.**

- Chart Type: Bar Chart
- Purpose: To compare success rates across different orbit types
- **Certain orbits (e.g., ES-L1, SSO) show high reliability, while others like GTO face challenges, likely due to orbit-specific complexities**

[Link to Github]

EDA with SQL

- The purpose of this EDA with SQL was to analyze the SpaceX launch dataset to gain insights into factors influencing mission success and payload performance
- Purpose of SQL Queries
  - To display all unique launch sites in the dataset.
  - To calculate the total payload mass carried by boosters launched for NASA (CRS)
  - To calculate the average payload mass for the booster version F9
  - To find the date of the first successful landing on a ground pad
  - To count the total number of successful and failed mission outcomes
  - To find the booster versions that carried the maximum payload mass
  - To rank landing outcomes by count

Link to Github

Build an Interactive Map with Folium

- Map objects
  - Markers: to place markers at each launch site and outcomes
    - Green Markers(Success) and Red markers(Fail)
  - Circles: to highlight the location of each launch site with radius of 1000 m
  - Marker Clusters: to group markers with the same coordinates
  - Mouses Position: to display real time Lat and Lon of any point on the map
  - Distance Lines: to draw between launch sites and other place like coastline, highway, city and railway
  - Distance Markers: to display calculated distances

Link to Github

Build a Dashboard with Plotly Dash

- Purpose
  - Explore launch success rates
  - Analyze the impact of payload mass
  - Identify pattern and insight

- Plots and Interaction
  - Pie Chart
    - Selected All Sited: Shows total successful launches for all sites
    - Selected Specific Sites: Display the success and failure for that site
  - Scatter Plot
    - Plots the relationship between "Payload Mass" and "Launch Success"
    - To identify patterns and correlations between payload, booster versions and launch success
  - Dropdown selection
    - To interact exploration of success and failure data for all sites or individual sites
  - Range Slider
    - To analysis how payload mass affects launch outcomes dynamically

Link to Github

Predictive Analysis (Classification)

Step1. Data Loading and Preparation

- Load datasets
- Extract target variable and features
- Standardize features and split data

Step2. Model Building and Hyperparameter Tuning

- Logistic Regression, SVM, Decision Tree, KNN
- Hyperparameter tuning ensured optimal performance

Step3. Model Evaluation

- Test Accuracy, Confusion Matrices, Comparing model performance

Step4. Best Model Selection

- KNN is the best performing model (=83.3%)

Through systematic data preparation, model building, and evaluation, **KNN** emerged as the best model for predicting the success of Falcon 9 first stage landings

Link to Github

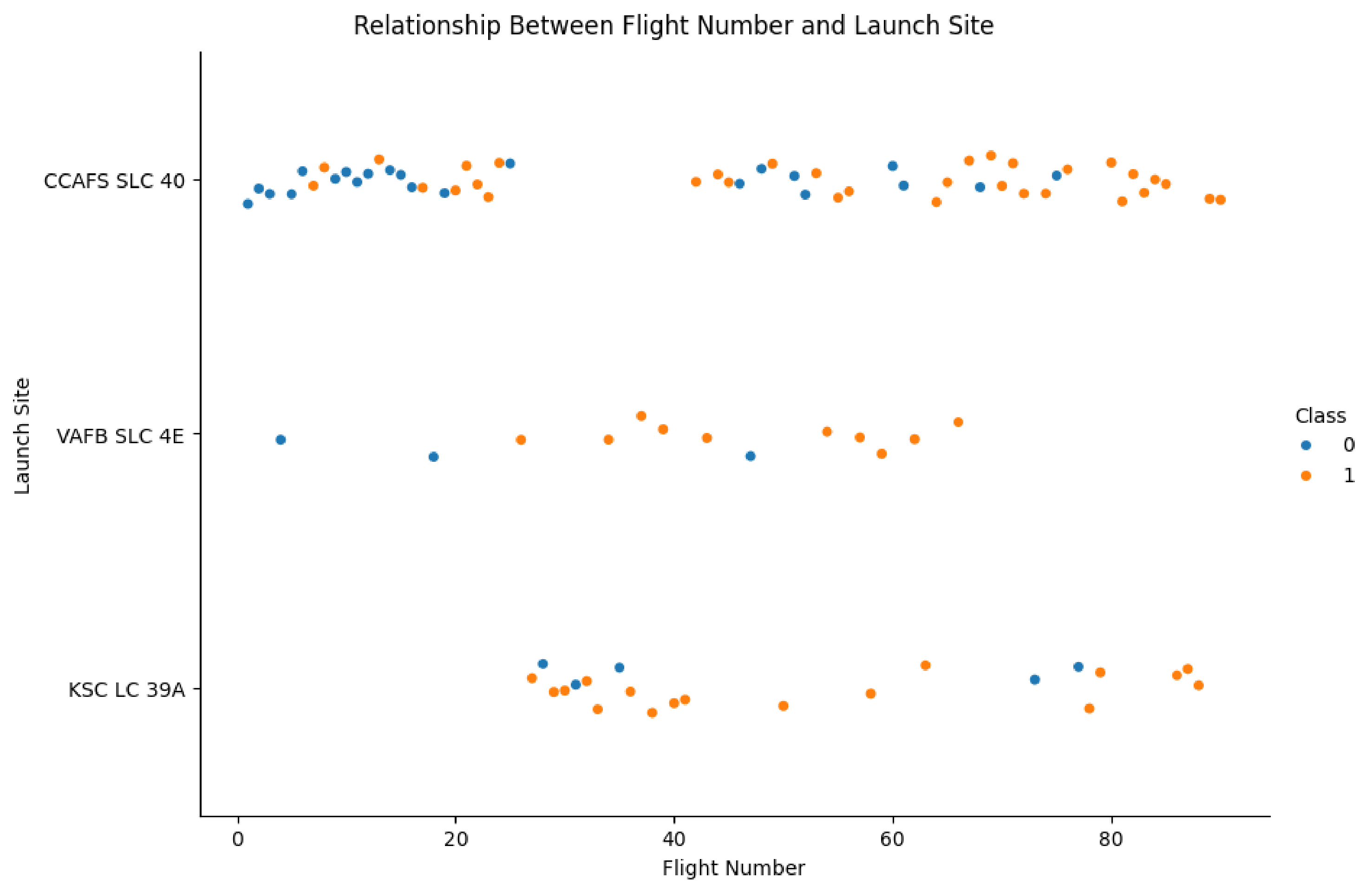
Results

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

Section 2

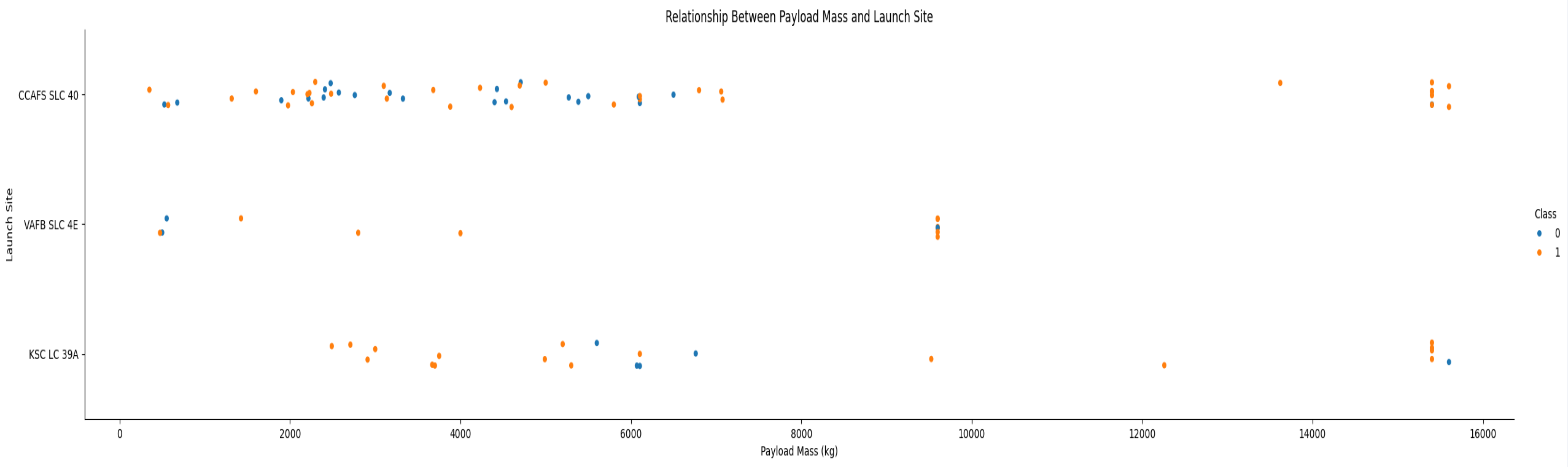
Insights drawn from EDA

Flight Number vs. Launch Site



- Success Rate Increases: Higher flight numbers show more successful landings, indicating improvement over time.
- CCAFS SLC 40: More failures early on but improves later.
- KSC LC 39A: Shows consistent success in later launches.
- VAFB SLC 4E: Fewer launches, but relatively successful.

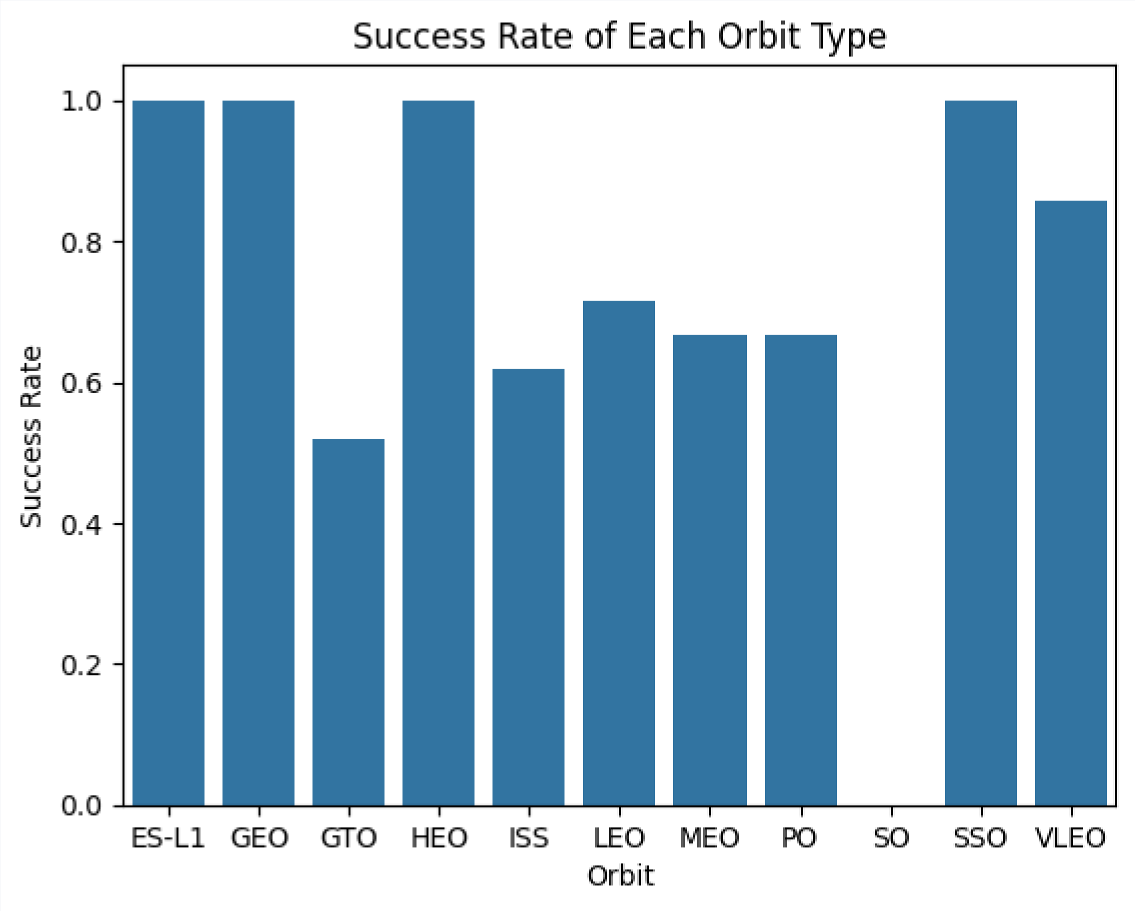
Payload vs. Launch Site



- Higher payloads tend to succeed, especially at KSC LC 39A

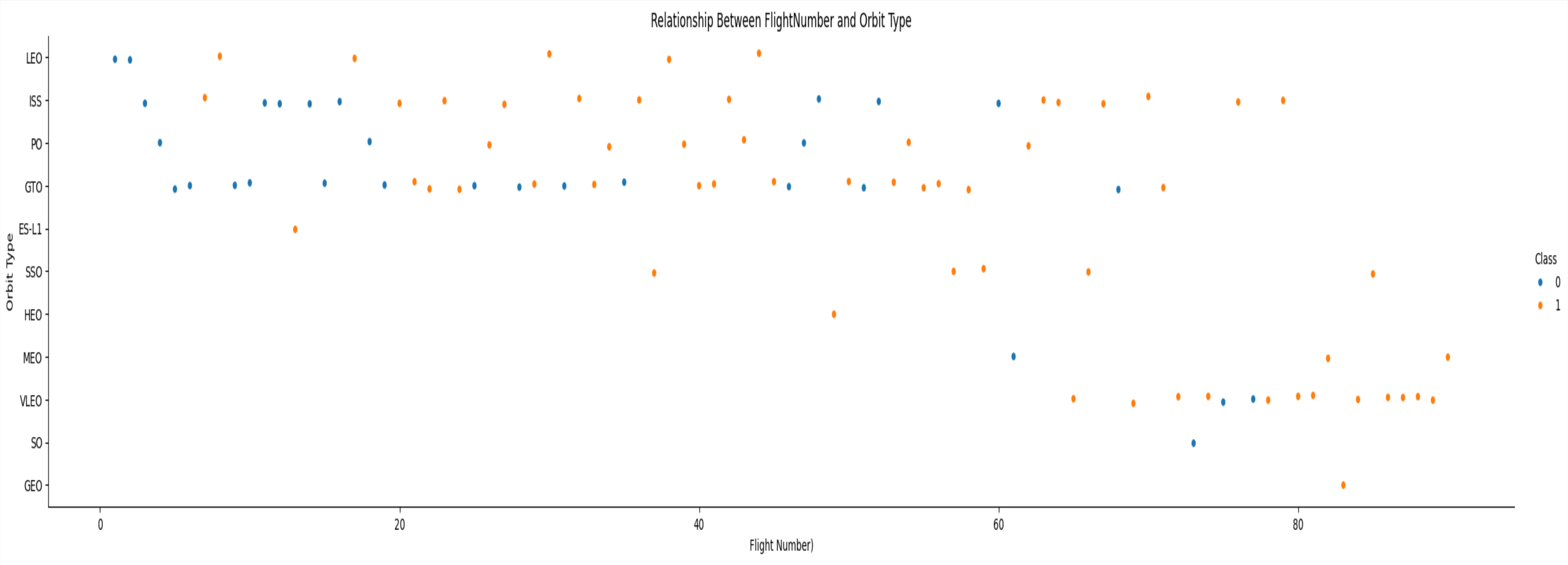
- CCAFS SLC 40: Handles a wide range of payloads; success increases with heavier payloads.
- KSC LC 39A: Successful landings even with high payload mass (over10,000 kg).
- VAFB SLC 4E: Limited launches with smaller payloads; success rate varies.

Success Rate vs. Orbit Type



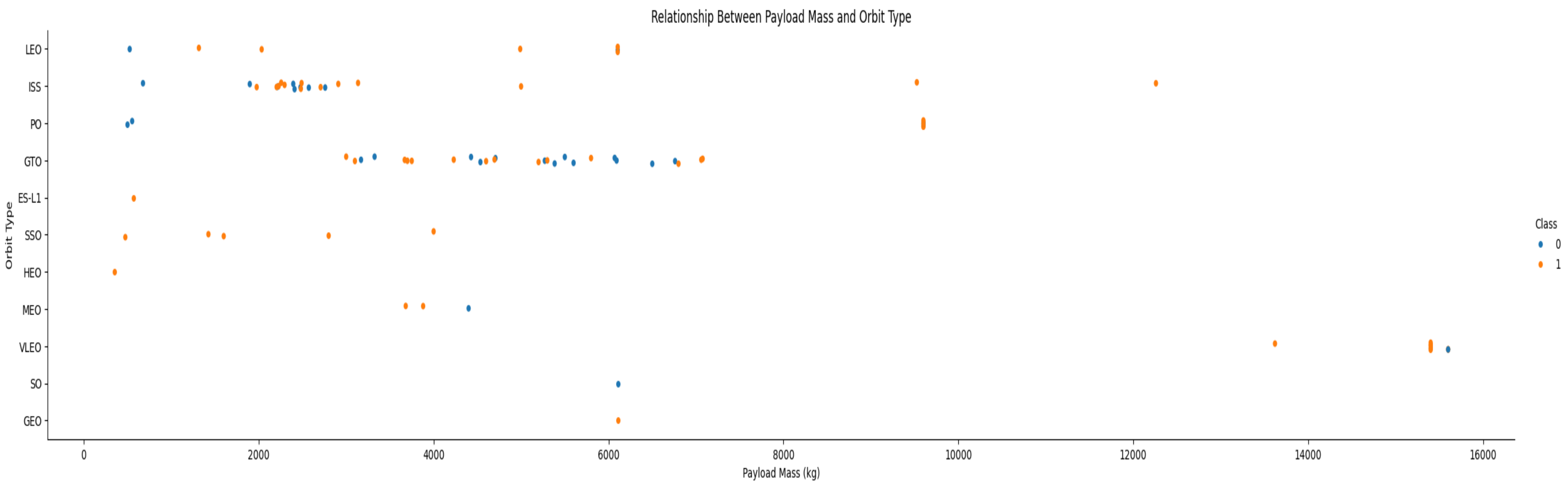
- High Success Rates: Orbits ES-L1, GEO, HEO, and SSO achieved a around 100% success rate.
- Low Success Rate: GTO shows the lowest success rate.

Flight Number vs. Orbit Type



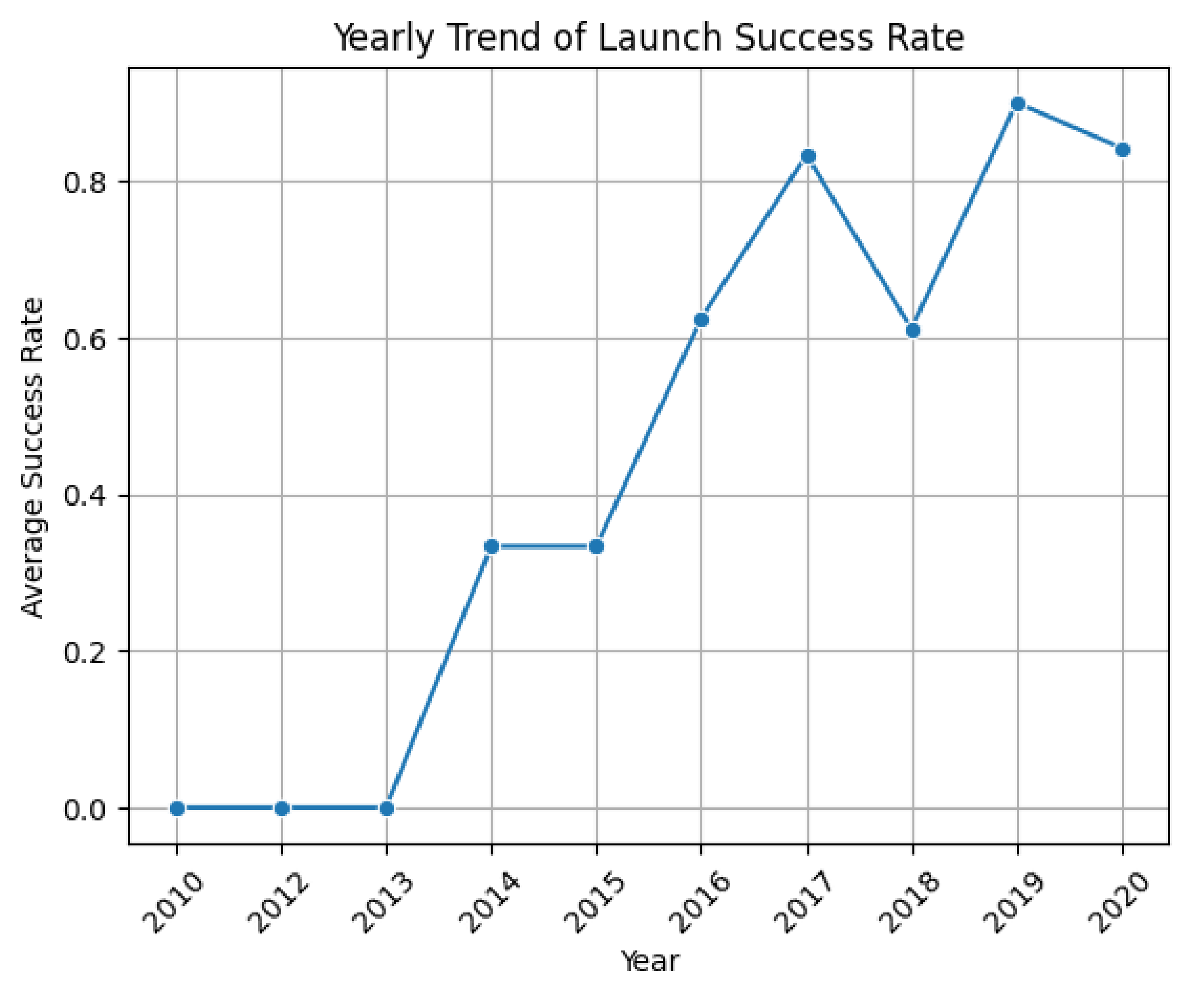
- LEO and ISS orbits show an improvement in success rates as flight numbers increase.
- GTO orbit consistently shows lower success rates, while orbits like SSO and ES-L1 maintain high success rates.

Payload vs. Orbit Type



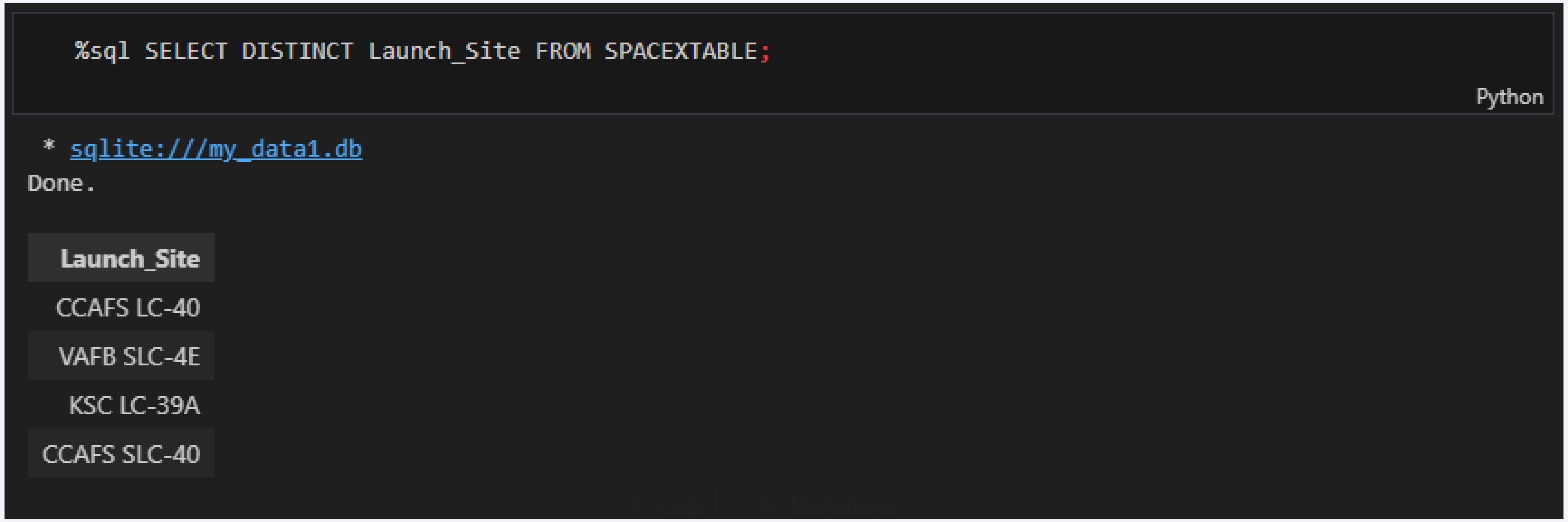
- LEO and ISS Orbits show higher success rates (Class 1) even with heavier payloads.
- GTO Orbit displays a lower success rate, with a mix of successful and unsuccessful landings, especially as payload mass increases.
- Heavier payloads (over 10,000 kg) are primarily successful in specific orbits like ISS and LEO, while GTO struggles with success at varying payload levels.

Launch Success Yearly Trend



- The success rate remained 0% until 2013, but after 2013, there is a significant upward trend in success rates.
- Success rates peaked in 2019 at approximately 90%.
- The success rate shows continuous improvement, highlighting advancements in landing technology and operational efficiency.

All Launch Site Names



```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

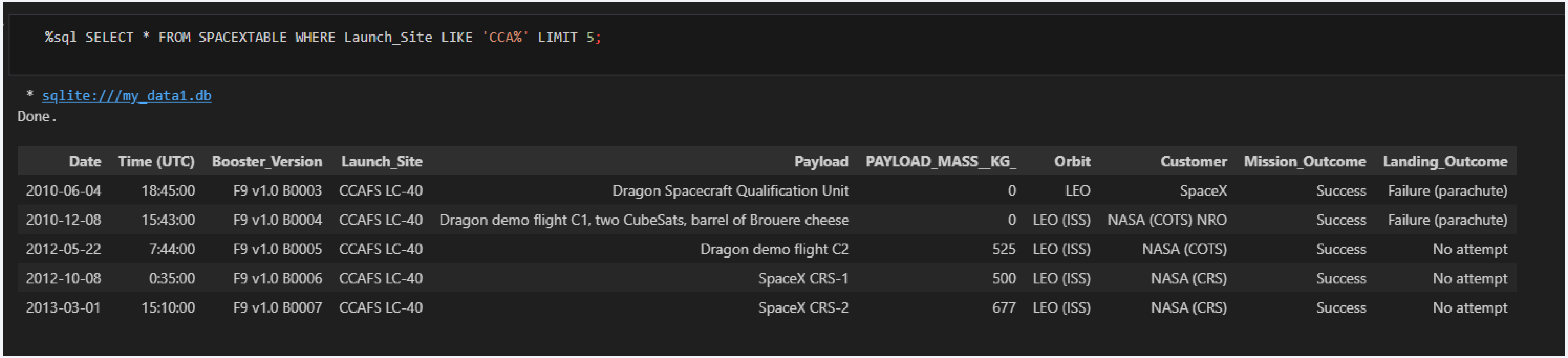
```
 * sqlite:///my_data1.db
Done.
```

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

- The query uses DISTINCT to fetch unique values from the Launch_Site column in the SPACEXTABLE.
- The result shows four distinct launch sites: CCAFS LC-40, VAFB SLC-4E, and KSC LC-39A

The query retrieves the first 5 records from the SPACEXTABLE where the Launch_Site starts with 'CCA' using the LIKE clause

Launch Site Names Begin with 'CCA'



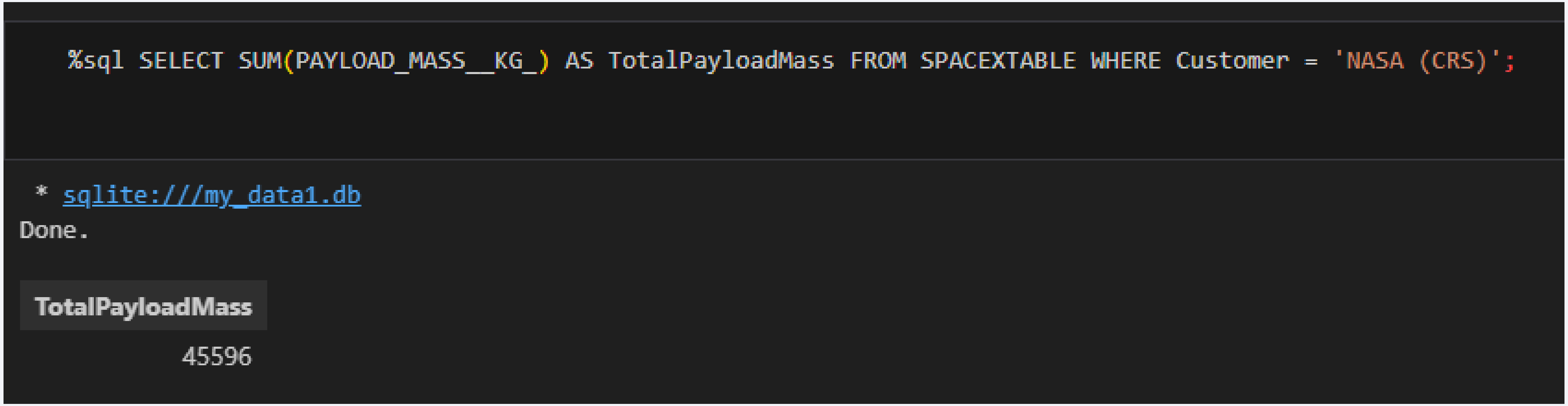
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

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

- The query retrieves the first 5 records from the SPACEXTABLE where the Launch_Site starts with 'CCA' using the LIKE clause

Total Payload Mass



```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TotalPayloadMass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

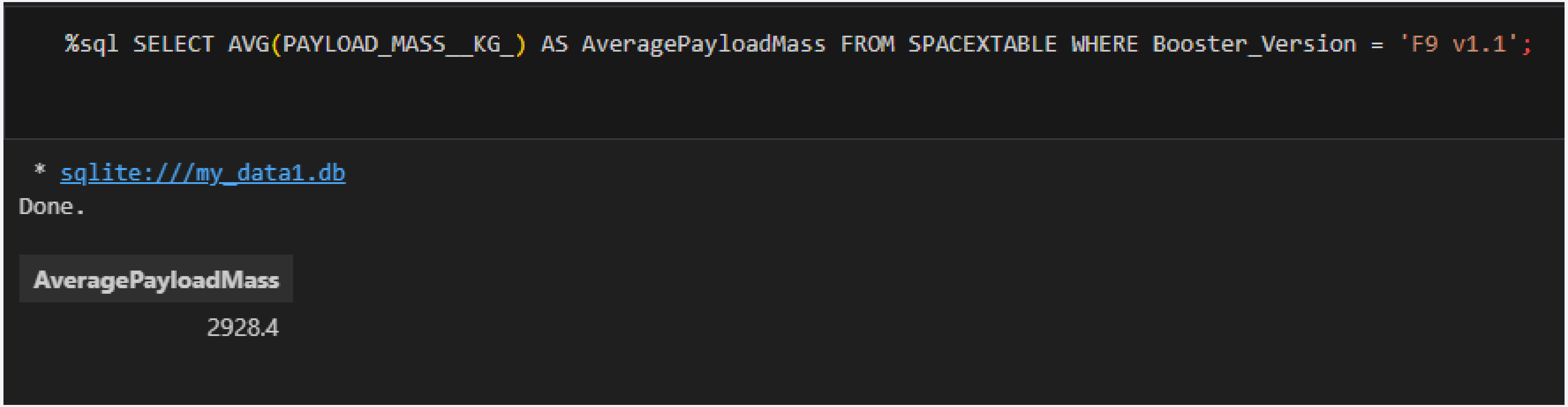
```
 * sqlite:///my_data1.db
Done.

TotalPayloadMass
           45596
```

- The query calculates the total payload mass for missions where the Customer is 'NASA (CRS)' using the SUM function.
- The result shows a total payload mass of 45,596 kg

The query calculates the average payload mass for missions using the Booster_Version 'F9 v1.1' with the AVG function. The result shows an average payload mass of 2928.4

Average Payload Mass by F9 v1.1



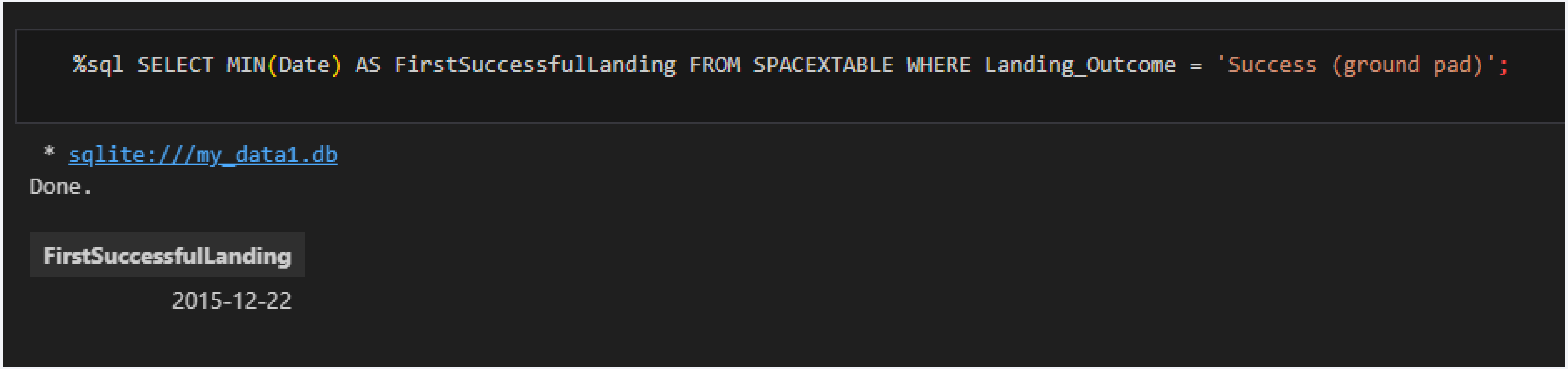
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AveragePayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

AveragePayloadMass
            2928.4
```

- The query calculates the average payload mass for missions using the Booster_Version 'F9 v1.1' with the AVG function.
- The result shows an average payload mass of 2928.4 kg

First Successful Ground Landing Date



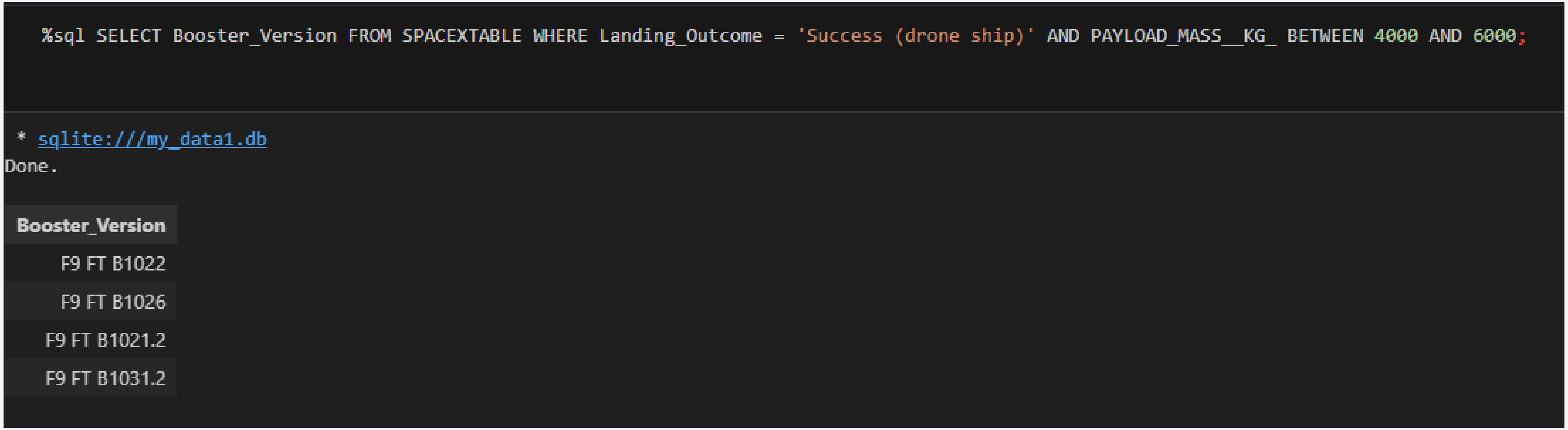
```
%sql SELECT MIN(Date) AS FirstSuccessfulLanding FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

FirstSuccessfulLanding
            2015-12-22
```

- The query retrieves the earliest successful landing date on a ground pad using the MIN function. The result shows the first successful landing occurred on 2015-12-22.

Successful Drone Ship Landing with Payload between 4000 and 6000



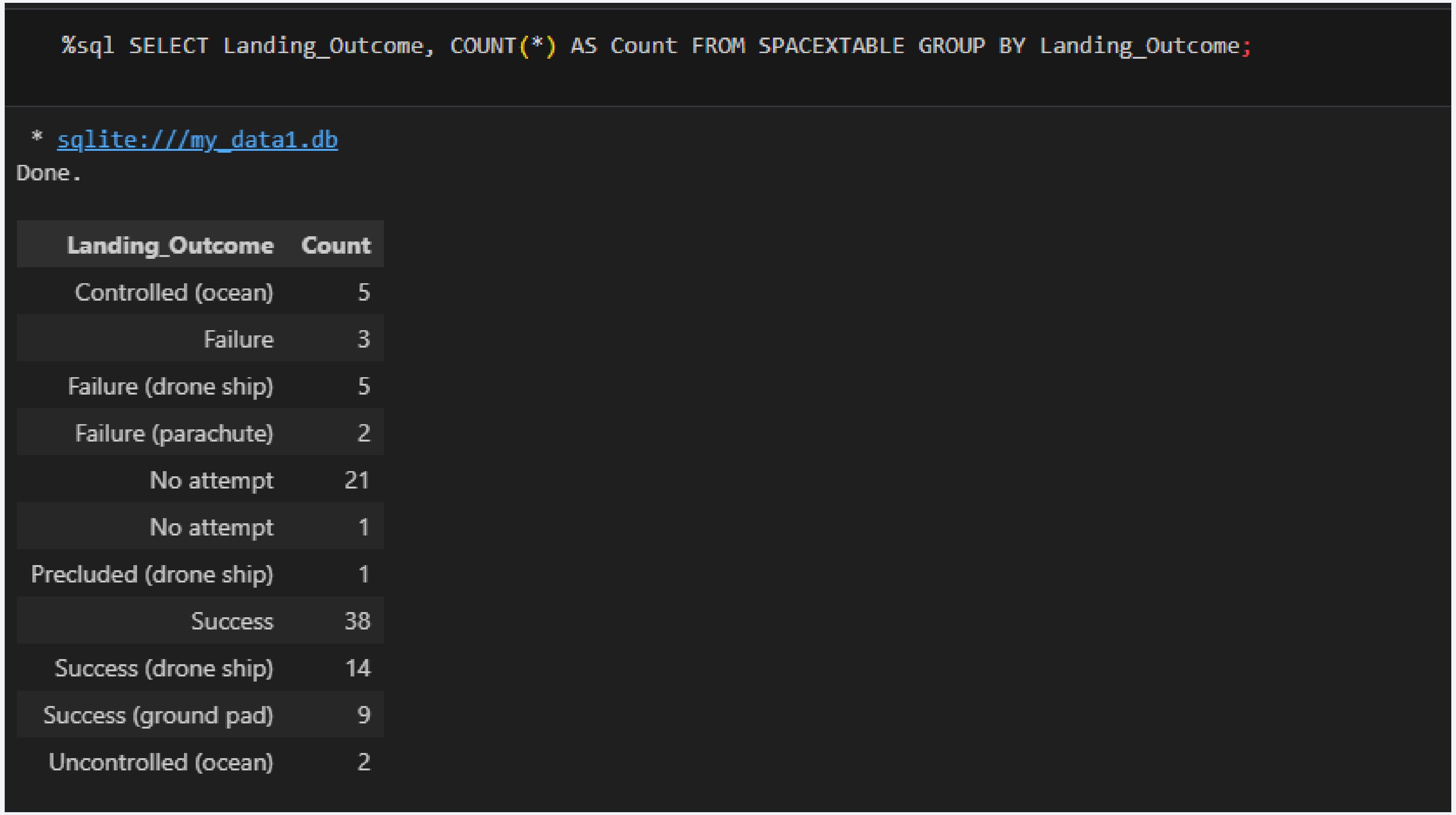
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

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

- This query retrieves the Booster_Version for successful landings on a drone ship where the PAYLOAD_MASS__KG_ is between 4000 and 6000 kg.
- The result lists four booster versions that meet these conditions.

Total Number of Successful and Failure Mission Outcomes



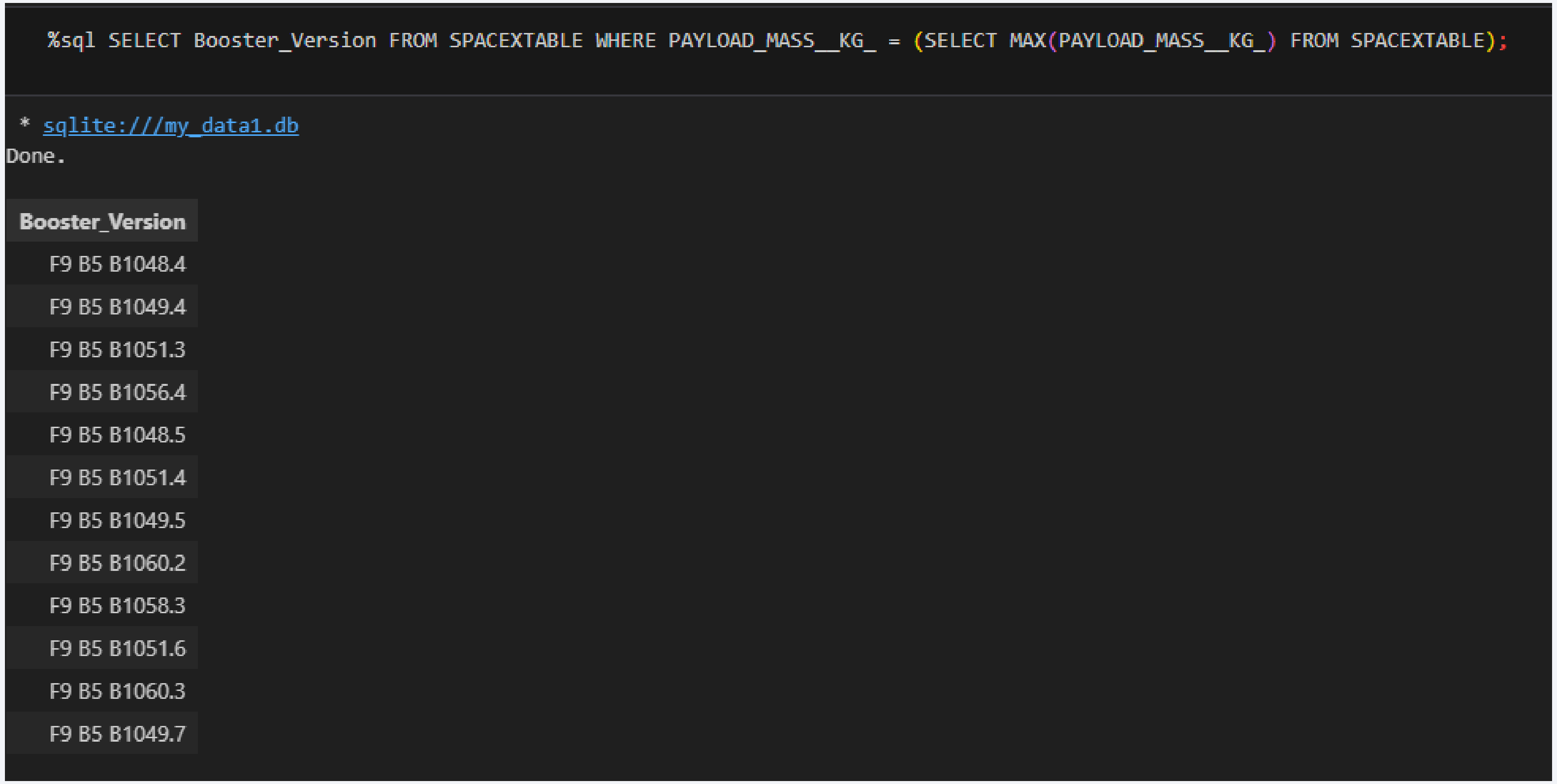
```
%sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE GROUP BY Landing_Outcome;
```

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

Landing_Outcome	Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

- This query groups records by Landing_Outcome and counts the occurrences for each category.
- The results show outcomes, with "Success" (38) being the most frequent.

Boosters Carried Maximum Payload



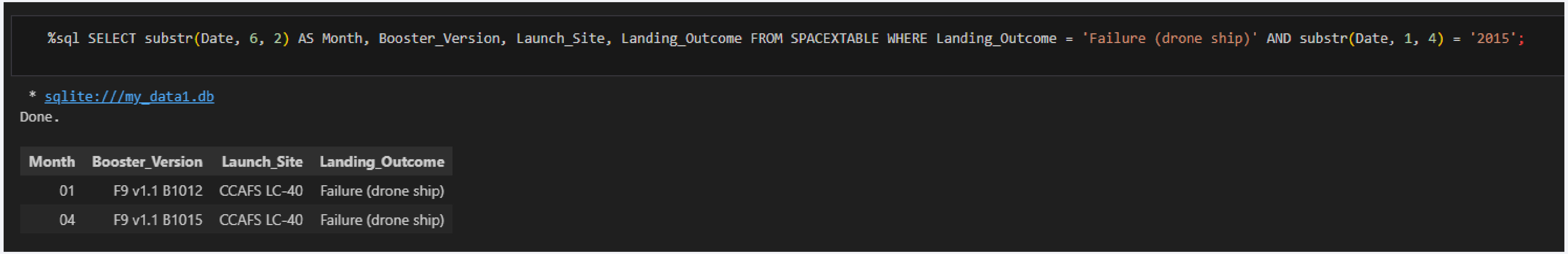
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

- This query retrieves all Booster_Version entries where the PAYLOAD_MASS__KG_ is equal to the maximum payload mass in the table.
- The result lists multiple boosters that achieved the heaviest payloads.

2015 Launch Records



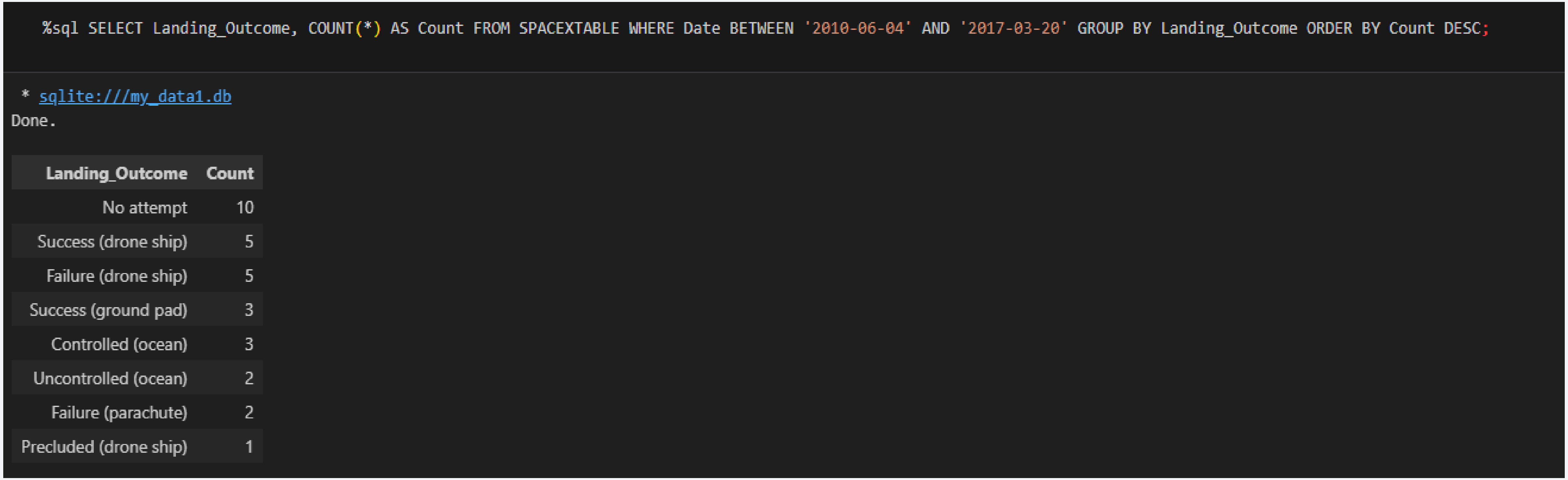
```
%sql SELECT substr(Date, 6, 2) AS Month, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTABLE WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date, 1, 4) = '2015';
```

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

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

- This query finds drone ship landing failures in 2015, showing two failures in January and April from CCAFS LC-40 with Booster Versions F9 v1.1 B1012 and B1015.

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



```
%sql SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Count DESC;
```

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

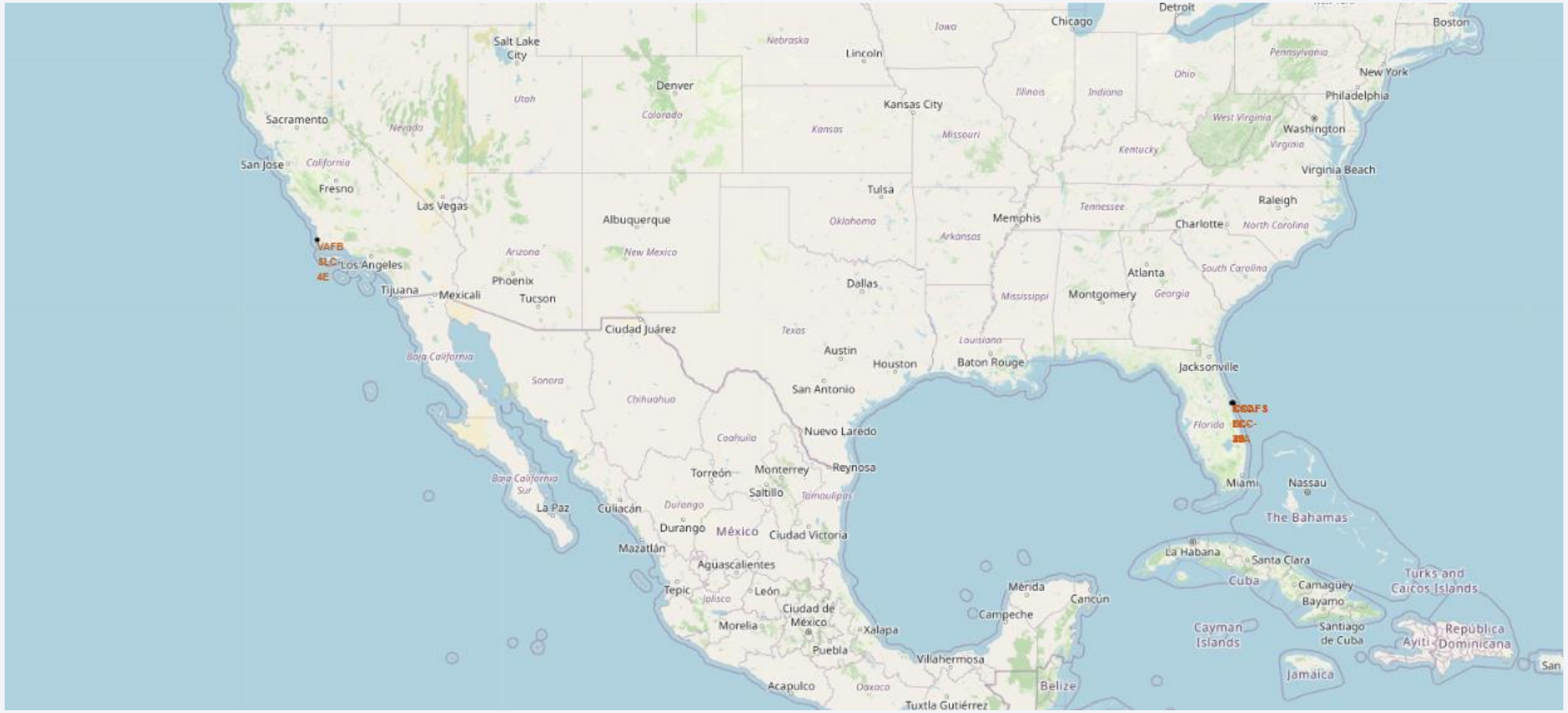
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query groups and counts landing outcomes between 2010-06-04 and 2017-03-20, showing "No attempt" as the most frequent outcome (10), followed by Success (drone ship) and Failure (drone ship) with 5 occurrences each.

Section 3

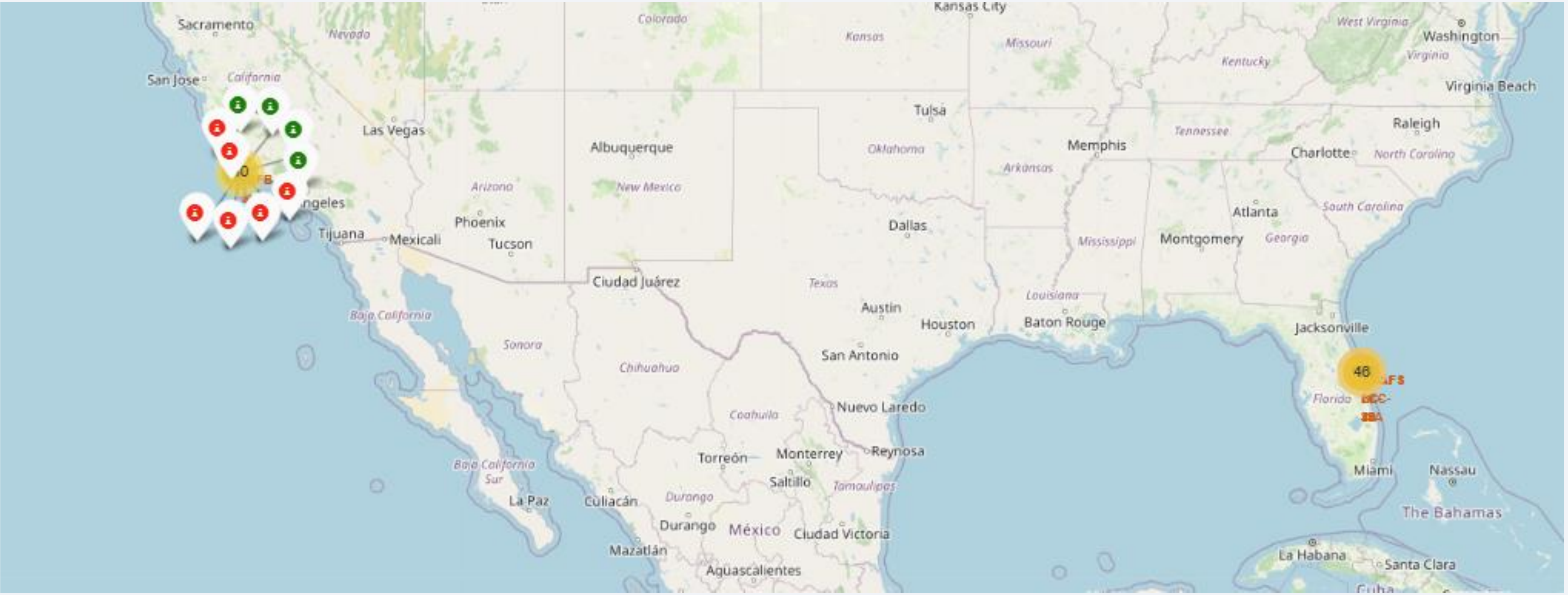
Launch Sites Proximities Analysis

Mark all launch sites on a map



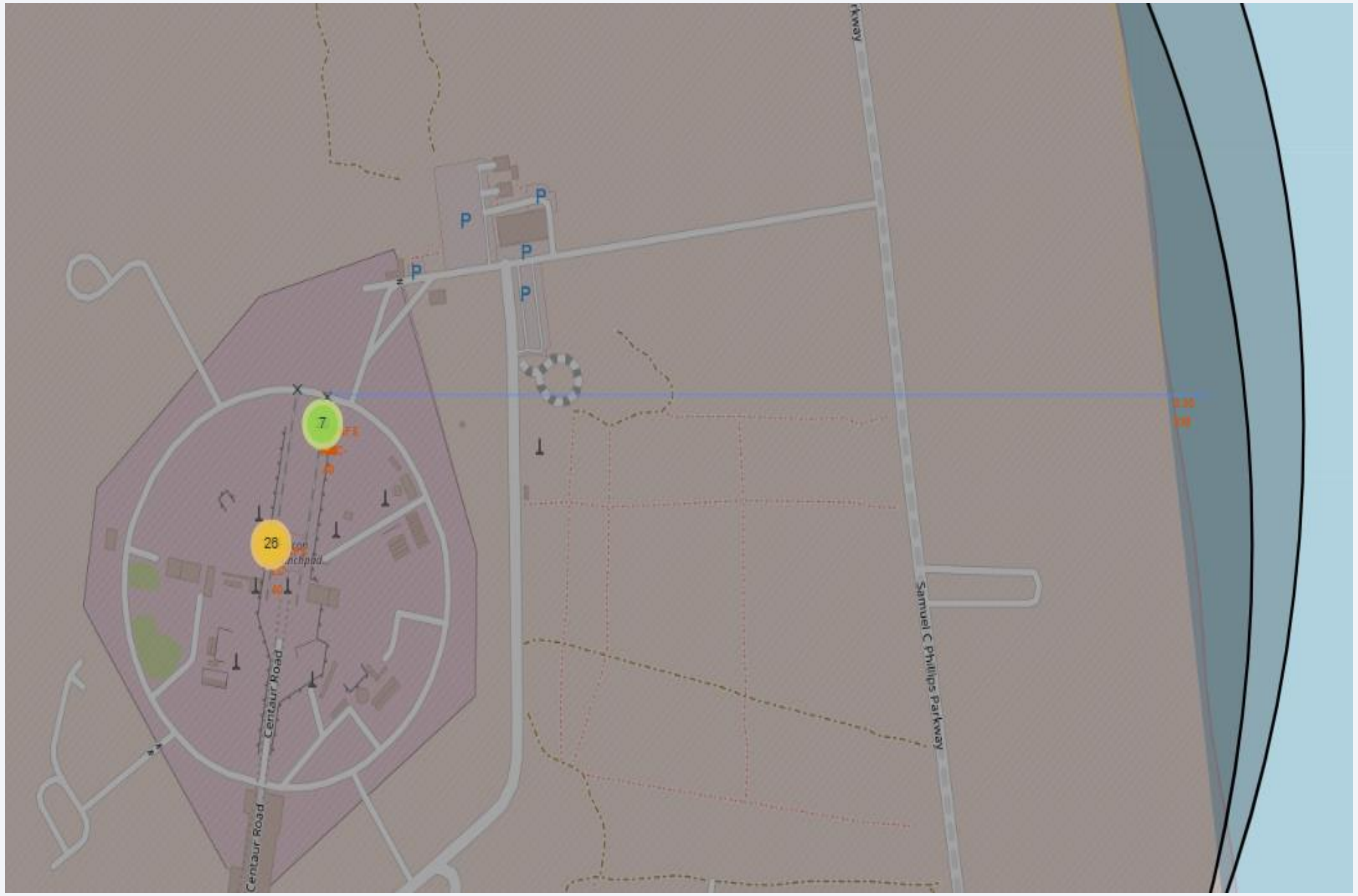
- I can guess that launch sites are strategically positioned near coastlines, infrastructure, and geographically advantageous areas to optimize launch success and minimize risks

MarkerCluster to current Site Map



- Green markers indicate successful launches / Red markers indicate failed launches.
- The California launch site (VAFB SLC-4E) experienced 6 failures and only 4 successes, indicating a higher failure rate. This suggests that the site has faced challenges with launch success

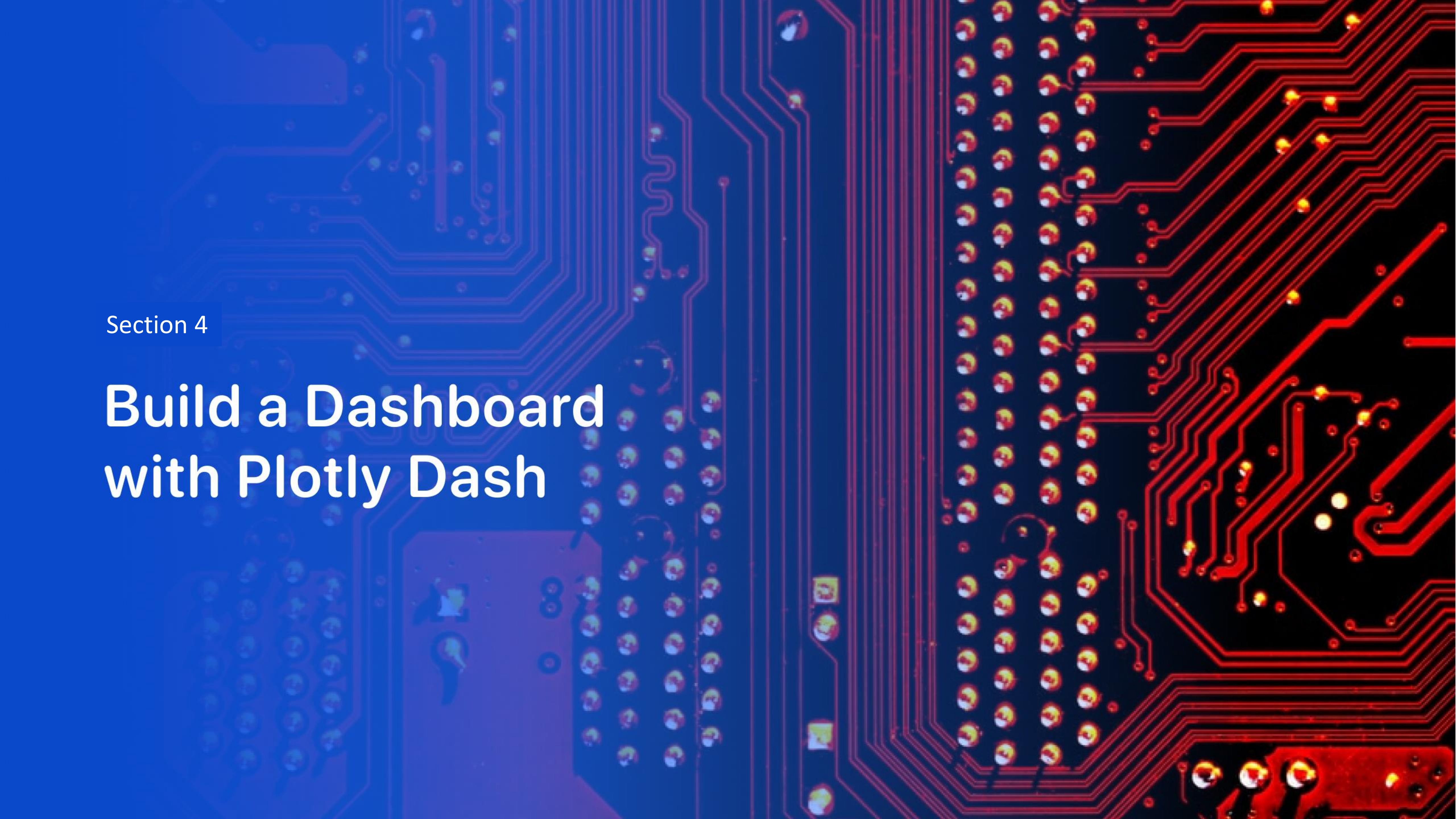
<Folium Map Screenshot 3>



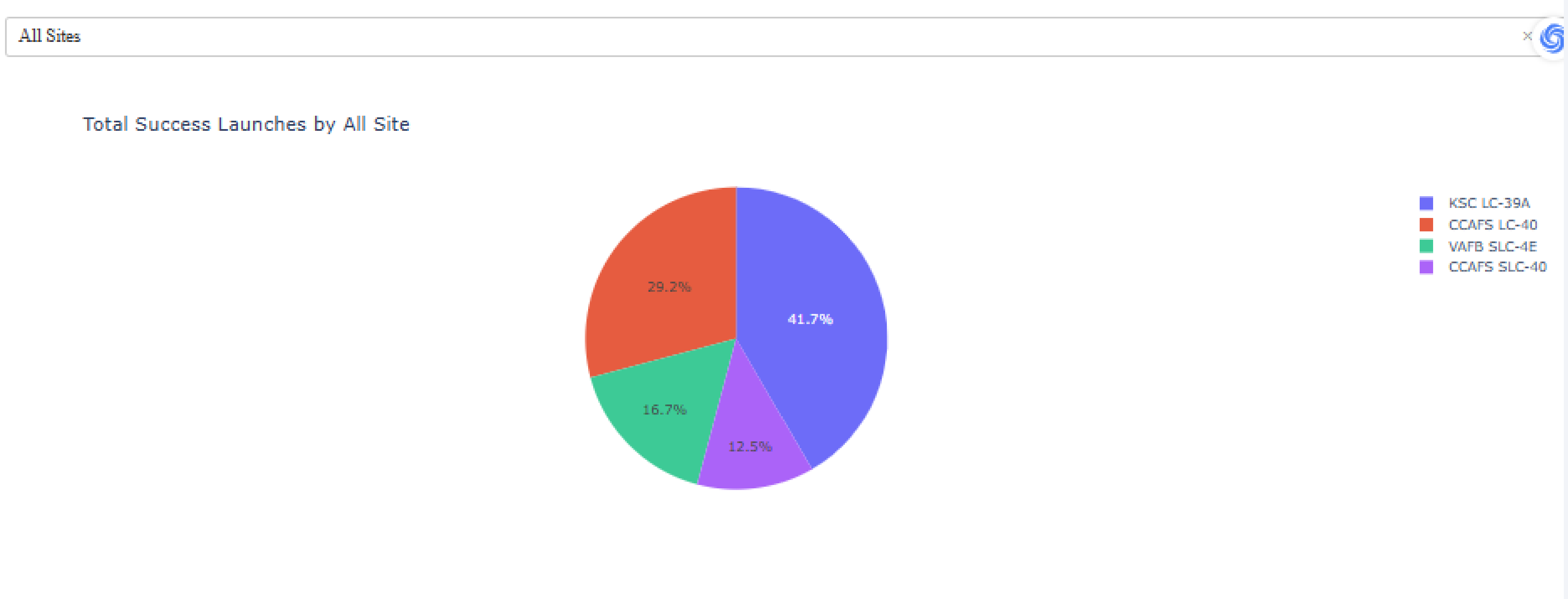
- Using the MouseOver tool, the coordinates of the closest coastline were identified, and the distance from the launch site to the coastline was calculated as 9.0 KM. This indicates that the site is in close proximity to the coast, which is optimal for safety and trajectory planning

Section 4

Build a Dashboard with Plotly Dash



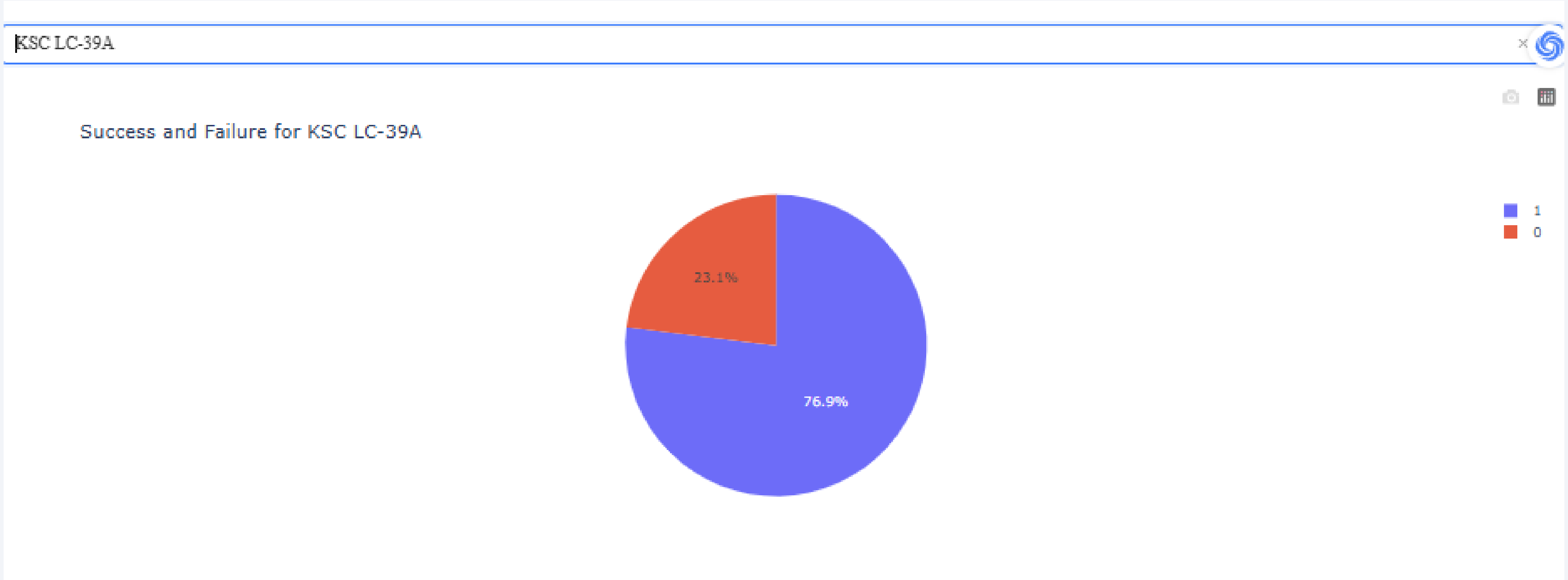
Total Success Launches by All Site



All Sites

- This pie chart represents the success rate of launches across all launch sites
- KSC LC-39A has the highest success rate, accounting for 41.7% of all successful launches.
- CCAFS LC-40 follows with 29.2%, while VAFB SLC-4E and CCAFS SLC-40 have the lowest success rates at 16.7% and 12.5%, respectively

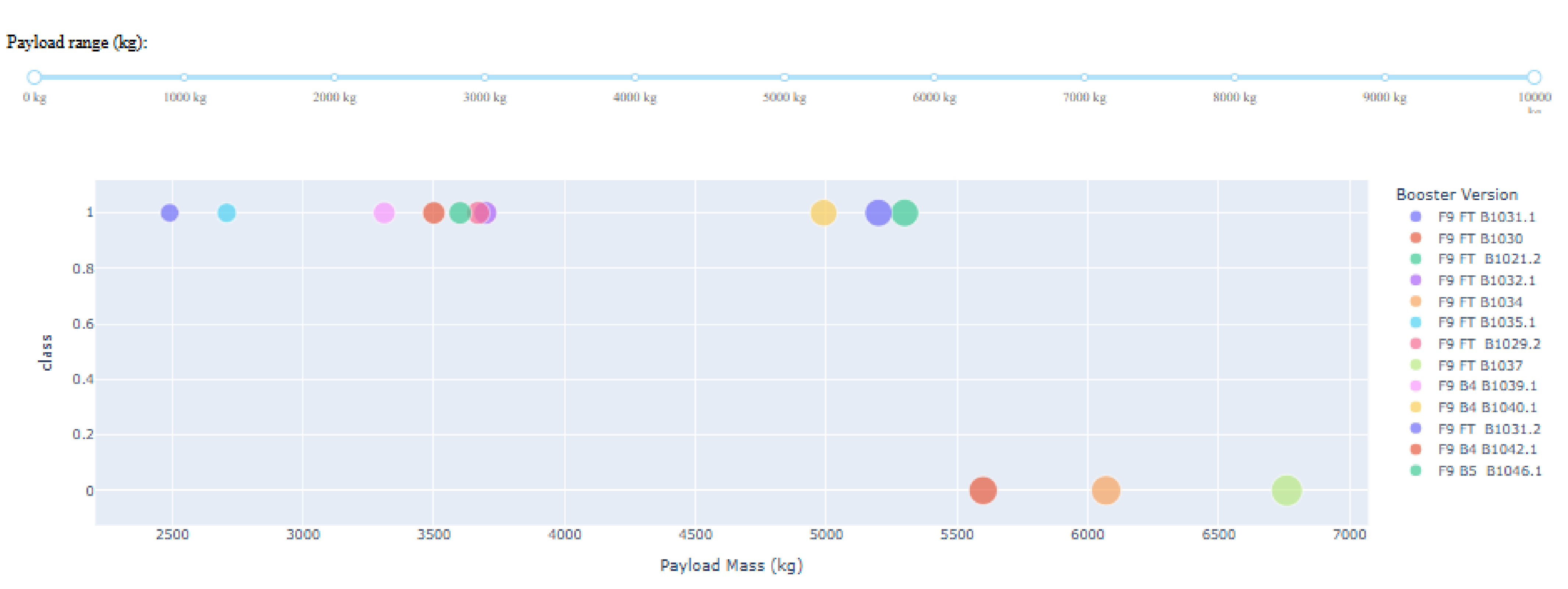
Highest Launch site



KSC LC-39A

- This pie chart shows the success and failure rate for the launch site KSC LC-39A:
- Success (Blue): 76.9%
    - KSC LC-39A has a high success rate, indicating strong reliability.

Impact of Payload Mass on Launch Success

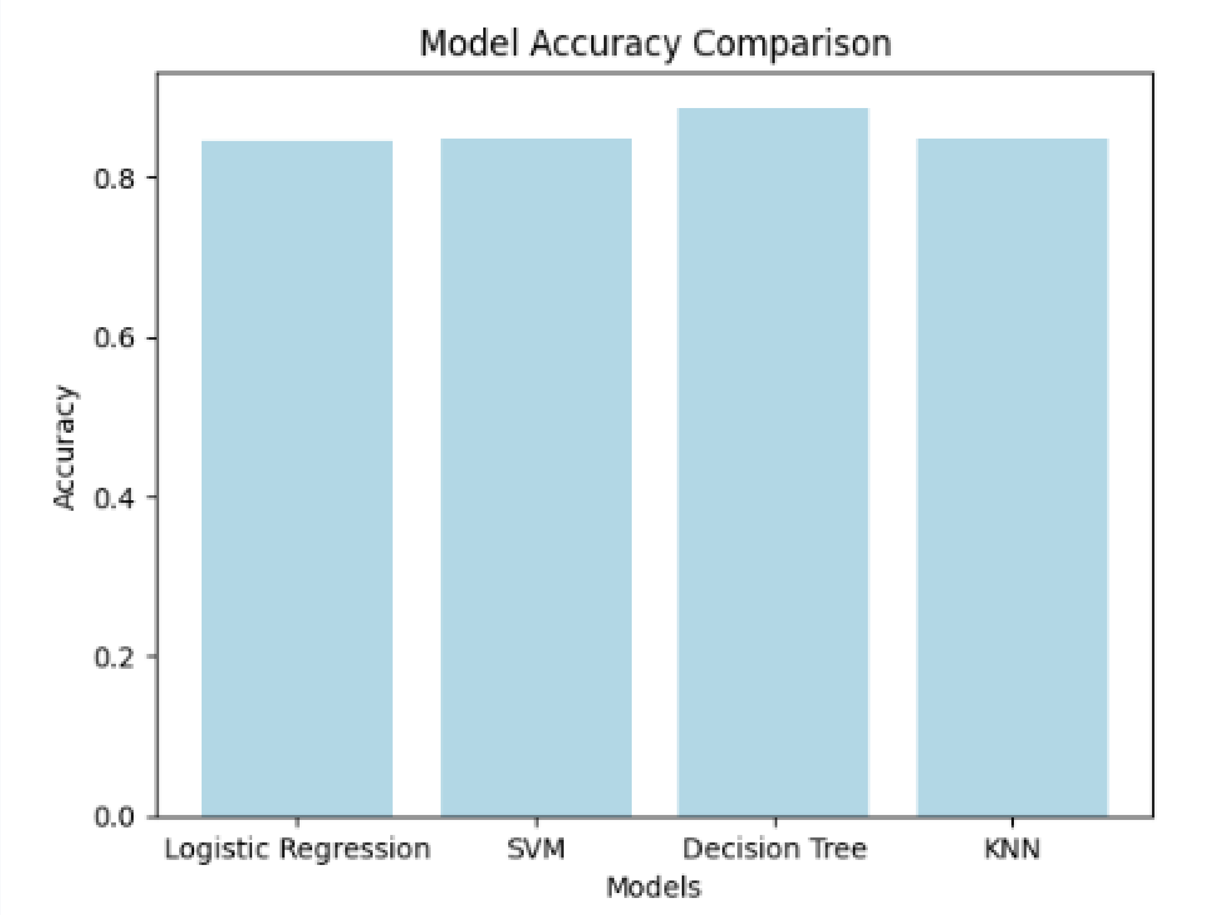


- This scatter plot shows the relationship between Payload Mass (kg) and Launch Success (class) for different Booster Versions
- Payload mass has a significant impact on success rates, with heavier payloads showing a higher risk of failure

Section 5

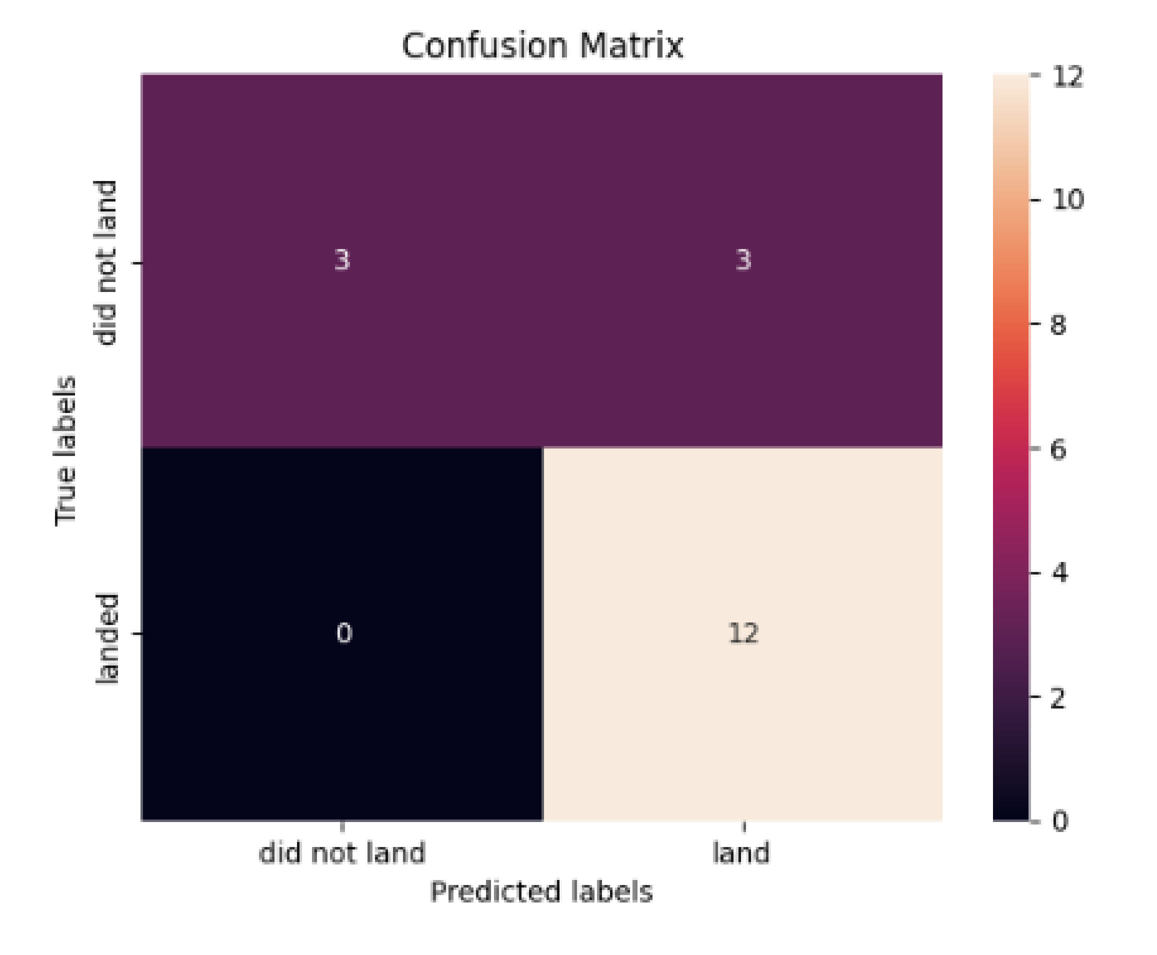
Predictive Analysis (Classification)

Classification Accuracy



- This bar chart compares the accuracy of four machine learning models:
- Decision Tree has the highest accuracy at 88.75%.
- SVM, Logistic Regression, and KNN perform similarly, each around 84.6% - 84.8%.

Confusion Matrix



- The model correctly predicts landings well, with no false negatives, but it has a few false positives, indicating minor misclassification for "did not land" cases. Overall, the performance is strong.

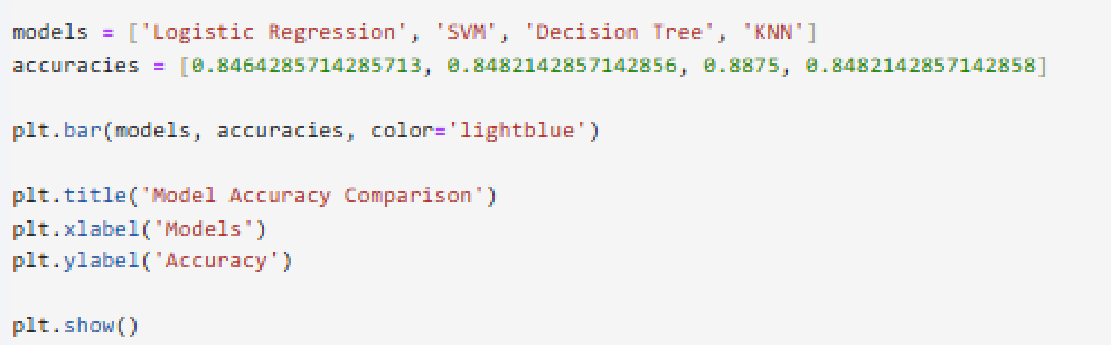
Conclusions

- Improved Launch Success Rate
  - Over the years, SpaceX's launch success rate has steadily increased due to technological advancements and accumulated experience.
- Success Rate by Launch Site
  - KSC LC-39A recorded the highest success rate (76.9%), making it the most reliable and key launch site compared to others.
- Payload and Success Relationship
  - Medium payload ranges (3000kg to 5000kg) demonstrated relatively higher success rates, indicating that payload mass plays a significant role in launch success.
- Success Rate by Orbit
  - Certain orbits, such as LEO (Low Earth Orbit) and SSO (Sun-Synchronous Orbit), showed higher success rates compared to others, highlighting the importance of target orbit selection.
- Best Performing Model
  - The Decision Tree model achieved the highest accuracy (88.75%) among other machine learning models (Logistic Regression, SVM, KNN), making it the most effective model for predicting launch success.

Appendix

- Data Set
  - Collection_API
  - Web_Scrapping
  - Dash_data

- Comparing Machine Learning accuracy Code



```
models = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']
accuracies = [0.8464285714285713, 0.8482142857142856, 0.8875, 0.8482142857142858]

plt.bar(models, accuracies, color='lightblue')

plt.title('Model Accuracy Comparison')
plt.xlabel('Models')
plt.ylabel('Accuracy')

plt.show()
```

Thank you!