IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Noura Alhuthali
5/6/2025

Outline

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

Executive Summary

- **Methodologies Summary**
- Collected and cleaned launch data using Pandas
- Preformed exploratory data analysis (EDA) with visual tools
- Built interactive visuals using **Folium** and **Plotly Dash**
- Used **SQL** to query launch data for key insights
- Built a classification model to predict launch success
- **Results Summary**
- Most launches occurred at **CCAFS LC-40**
- Higher success rates for payloads under **10,000 kg**
- **Folium maps** show site locations and outcomes
- Predictive model identified key features affecting success

Introduction

SpaceX has conducted numerous rocket launches from various sites, each with different outcomes. This project aims to analyze launch performance and identify key factors influencing mission success. Using data science techniques such as EDA, mapping, and classification models, we explore which launch sites are most successful, how payload size and booster versions impact outcomes, and whether it's possible to predict launch success based on mission features.

Section 1

Methodology

Methodology

This project investigates SpaceX launch records to uncover patterns and predict mission outcomes. The data was collected from publicly available SpaceX datasets and supplemented with geolocation information. After initial data wrangling to clean and prepare the dataset, we conducted exploratory data analysis (EDA) using visualizations and SQL queries to identify key trends. Interactive visual analytics were developed using Folium maps to visualize launch site locations and success, and a Plotly Dash dashboard to allow dynamic exploration of launch performance. For predictive analysis, we applied classification models to estimate launch success based on features such as payload mass and

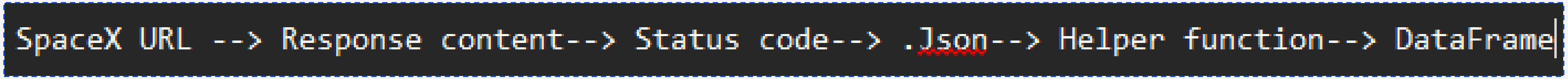
Data Collection

- The dataset collected from the official SpaceX API and supplementary CSV file from the IBM course The main dataset included launch records with details such as launch site, payload mass, booster version, and mission outcome additional dataset was obtained to provide geographic coordinates (latitude and longitude) for each launch site. These files were downloaded directly using Python tools ensuring data was loaded into the workspace in real-time for analysis.

- **data collection process**

1. Access SpaceX public dataset (CSV/API)
2. Retrieve launch records
3. Download geolocation dataset
4. Load dataset into python using pandas
5. Merge datasets

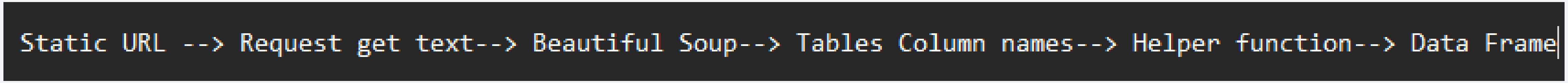
Data Collection – SpaceX API



```
SpaceX URL --> Response content--> Status code--> .Json--> Helper function--> DataFrame
```

- Noura-DS/IBM_FINAL_PROJECT

Data Collection - Scraping



```
Static URL --> Request get text--> Beautiful Soup--> Tables Column names--> Helper function--> Data Frame
```

- Noura-DS/IBM_FINAL_PROJECT

Data Wrangling

The cleaned dataset was imported, and missing values were checked, especially in the LaunchingPad column which contained nulls when no launchpad was used.
Data types were reviewed and included int64, float64, object, and bool.
Basic analysis was performed on columns like LaunchSite, showing that Cape Canaveral SLC 40 had the highest number of launches.
A new feature called class was created from the Outcome column, where any result containing "False" or "None" was marked as failed (0), and others as successful (1).

- https://github.com/Noura-DS/IBM_FINAL_PROJECT

EDA with Data Visualization

I used Seaborn and Matplotlib to create visualizations that helped explore patterns in the data.
One of the key charts was a scatter plot, which I used to visualize the relationship between Flight Number and Launch Site, along with other related features.
I also created a bar chart to show the relationship between the success rate and each orbit type.

- https://github.com/Noura-DS/IBM_FINAL_PROJECT

EDA with SQL

- **SELECT with WHERE and DISTINCT**: Queried data for specific launch outcomes and Extracted unique values such as successful or failed landings.
- **GROUP BY with COUNT()**: Counted the number of launches per site or per booster version.
- **ORDER BY and LIMIT**: Restricted the number of rows returned and Sorted results.
- **JOINs**: Merged data from different tables to analyze combined information like booster versions and landing outcomes.
- **Aggregation**: Calculated average, maximum, or minimum payload mass or number of successful launches.
- **Filtering using LIKE**: Filtered text patterns such as payloads containing "Starlink".
- **Subqueries**: Used to filter or calculate values based on nested SELECT statements.

- https://github.com/Noura-DS/IBM_FINAL_PROJECT

Build an Interactive Map with Folium

Map Objects Added:

- **Base Map:** Centered around the average coordinates of SpaceX launch sites.
- **Markers:** Placed at each launch site with popups showing site name and launch success stats.
- **Circles:** Colored by success rate (e.g., green for high, red for low) to highlight performance visually.
- **Circle Markers:** Represented individual launches; color-coded by success/failure.
- **Popups:** Displayed additional data like payload mass, orbit, and outcome on click.
- **Layer Control:** Enabled toggling between launch sites and individual launches for better interactivity.

We add this because we want to to give users a geographic overview of all SpaceX launch sites and make the map interactive and informative with minimal clutter.

- https://github.com/Noura-DS/IBM_FINAL_PROJECT

Build a Dashboard with Plotly Dash

- **Pie Chart:** Shows the proportion of successful vs. failed launches.
- **Bar Chart:** Displays the number of launches per site with filters by success outcome.
- **Scatter Plot:** Visualizes correlation between payload mass and launch success, color-coded by booster version.
- **Dropdown Menus:** Allow users to select a launch site or filter by payload range.
- **Range Slider:** Lets users filter data by payload mass range dynamically.

These where added to allow users to explore launch performance from different angles and support interactive analysis making insights visually accessible for technical and non-technical stakeholders.

- https://github.com/Noura-DS/IBM_FINAL_PROJECT

Predictive Analysis (Classification)

Model Development Process

1. **Data Preprocessing**
    1. Cleaned data (missing values, outliers)
    2. Standardized and normalized features
2. **Model Selection**
    1. Tried several classification models such as **Logistic Regression**, **Decision Trees**, **Random Forest**, and **Support Vector Machines (SVM)**
    2. Used **GridSearchCV** for hyperparameter tuning
3. **Model Training**
    1. Split data into **training** and **testing** sets (80% training, 20% testing)
    2. Trained multiple models on the training set
4. **Model Evaluation**
    1. Used evaluation metrics such as **accuracy**, **precision**, **recall**, **F1 score**, and **ROC-AUC**
    2. Compared model performance to select the best-performing model
5. **Model Improvement**
    1. Fine-tuned the best model using cross-validation and optimized hyperparameters
    2. Implemented **feature engineering** to improve model accuracy

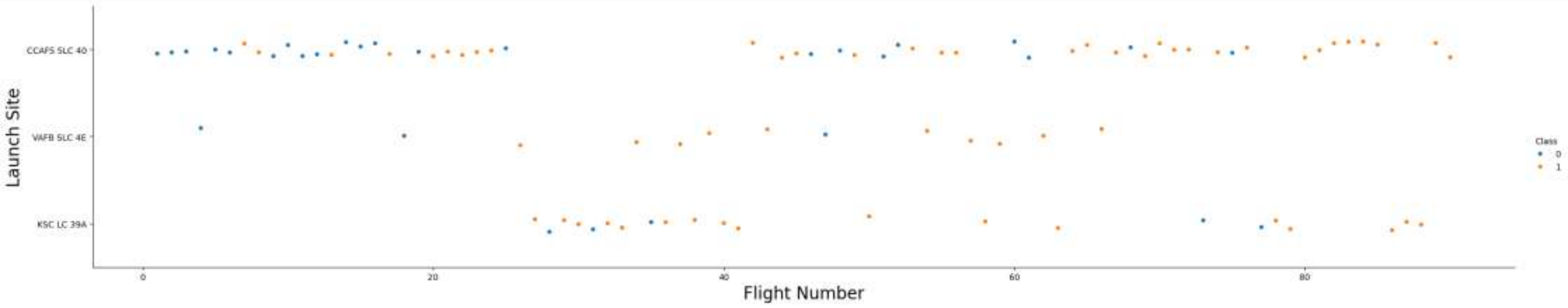
Results

The classification models were trained and evaluated using a train-test split approach. GridSearchCV was used to find the optimal hyperparameters for each model. After comparing the performance of several classifiers, the DecisionTreeClassifier achieved the best results with a training accuracy of 88.9% and a test accuracy of 94%, along with an F1-score of 88.21%. This model was selected as the final model due to its strong balance between accuracy and interpretability, making it a suitable choice for further analysis or deployment.

Section 2

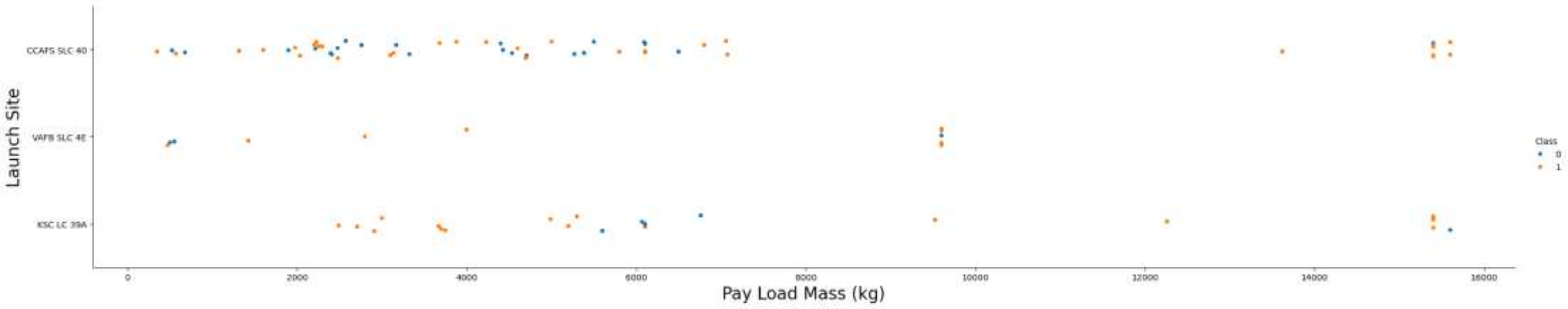
Insights drawn from EDA

Flight Number vs. Launch Site



The plot shows that launch sites with a higher number of flights tend to have higher success rates. For example, CCAFS SLC 40 shows more successes as the flight number increases, while VAFB SLC 4E has fewer successful launches at higher flight numbers.

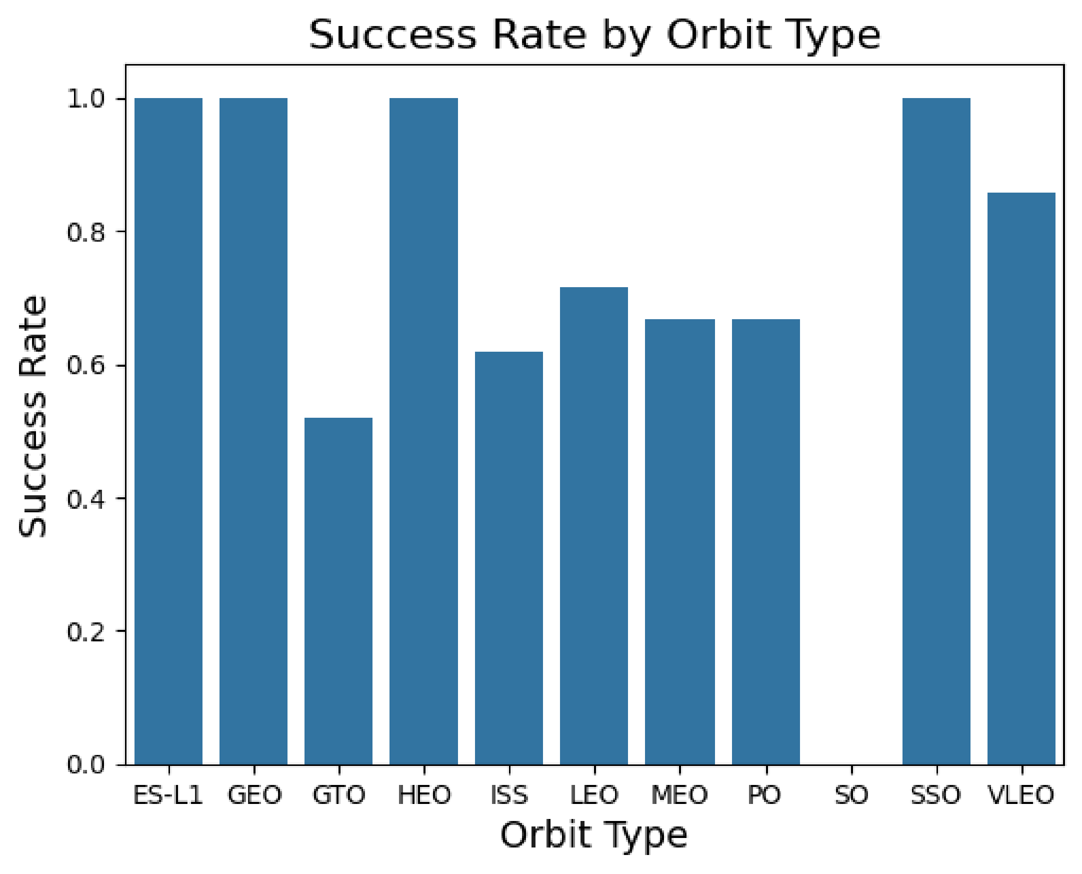
Payload vs. Launch Site



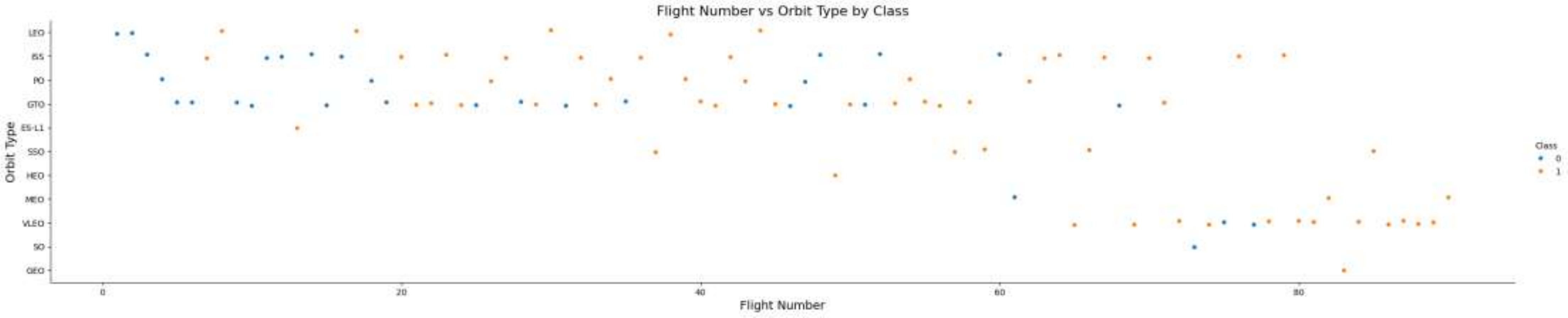
The plot shows that as the flight number increases, the success rate improves, especially for CCAFS SLC 40. In early flights, this site had more failures, but later flights show more consistent success. Also, higher payload mass at CCAFS SLC 40 is associated with a higher success rate.

Success Rate vs. Orbit Type

From the plot, we observed that four orbit types **ES-L1**, **SSO**, **HEO**, and **GEO** achieved a **100% success rate** these orbits showed the most consistent successful launches compared to others.

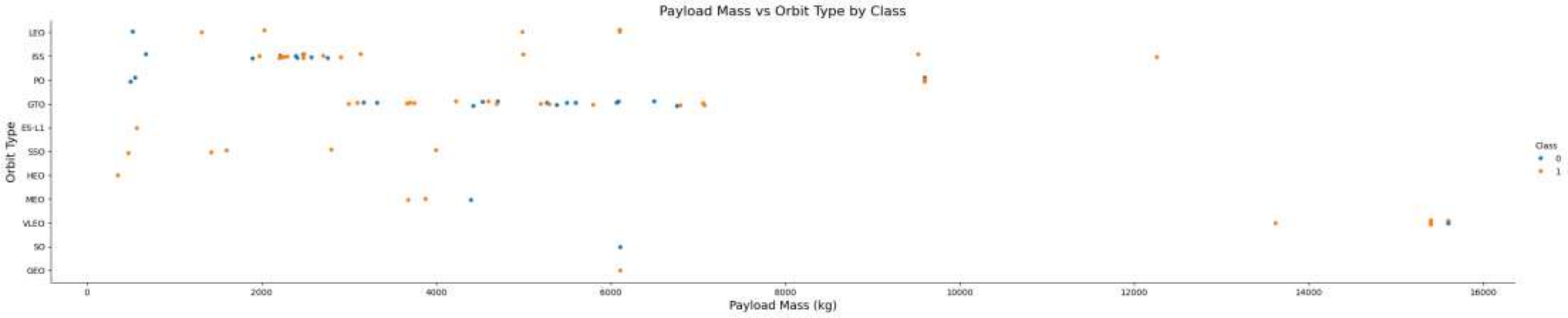


Flight Number vs. Orbit Type



From the plot, we can see that in the LEO and VLEO orbits, the success rate increases with higher flight numbers — especially after flight 60 for VLEO. However, in the GTO orbit, there is no clear relationship between flight number and success rate.

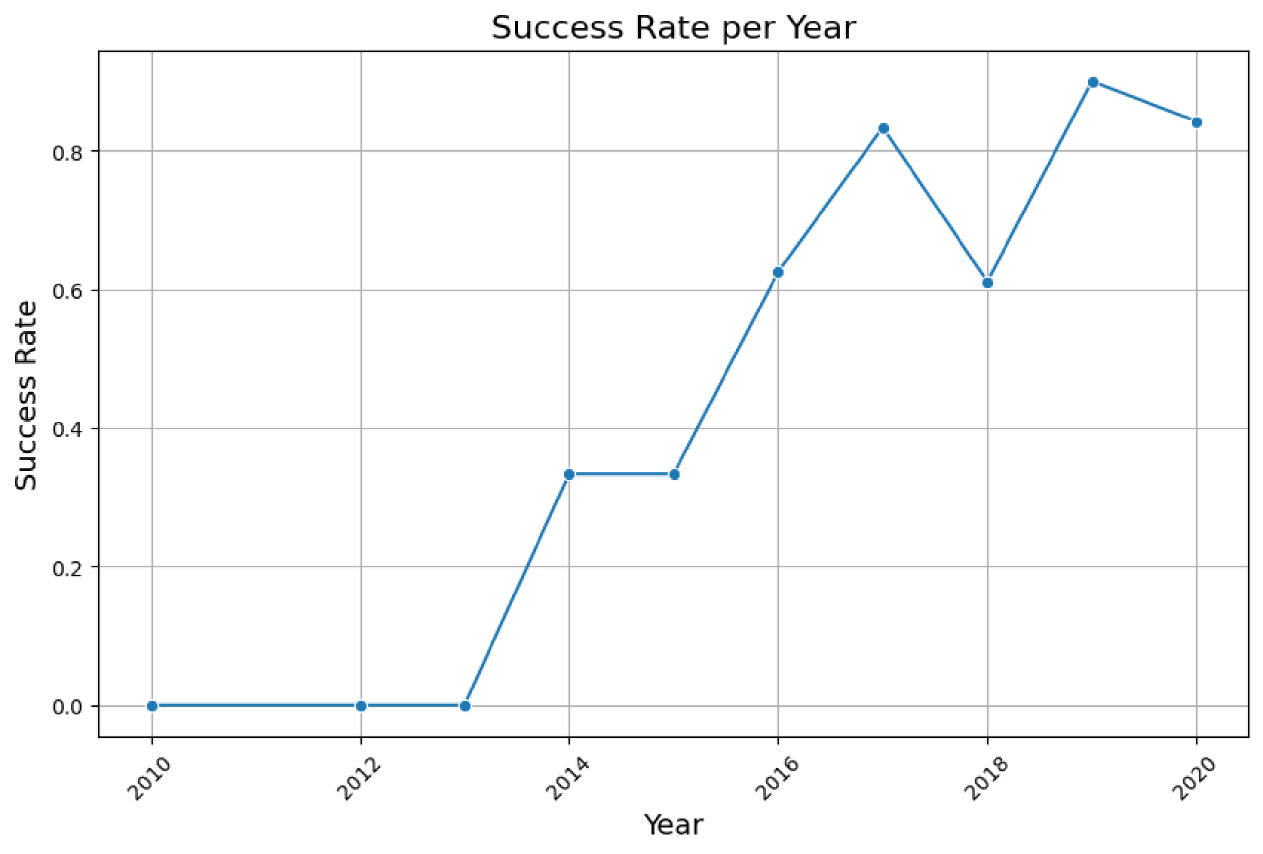
Payload vs. Orbit Type



The plot shows that heavier payloads are associated with more successful landings in the Polar, LEO, and ISS orbits. However, in the GTO orbit, both successful and unsuccessful landings are observed, indicating no clear trend for landing success with heavy payloads.

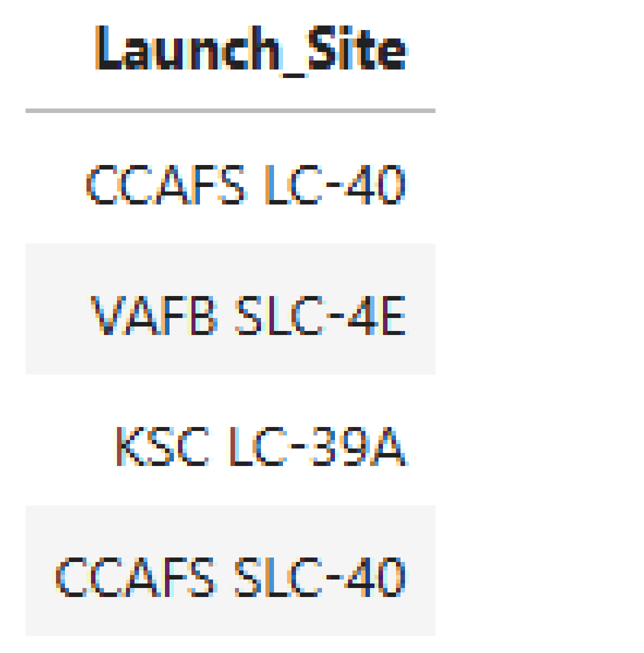
Launch Success Yearly Trend

The plot shows that the yearly success rate of Falcon 9 landings has been steadily increasing since 2013 and continued to rise through 2020.



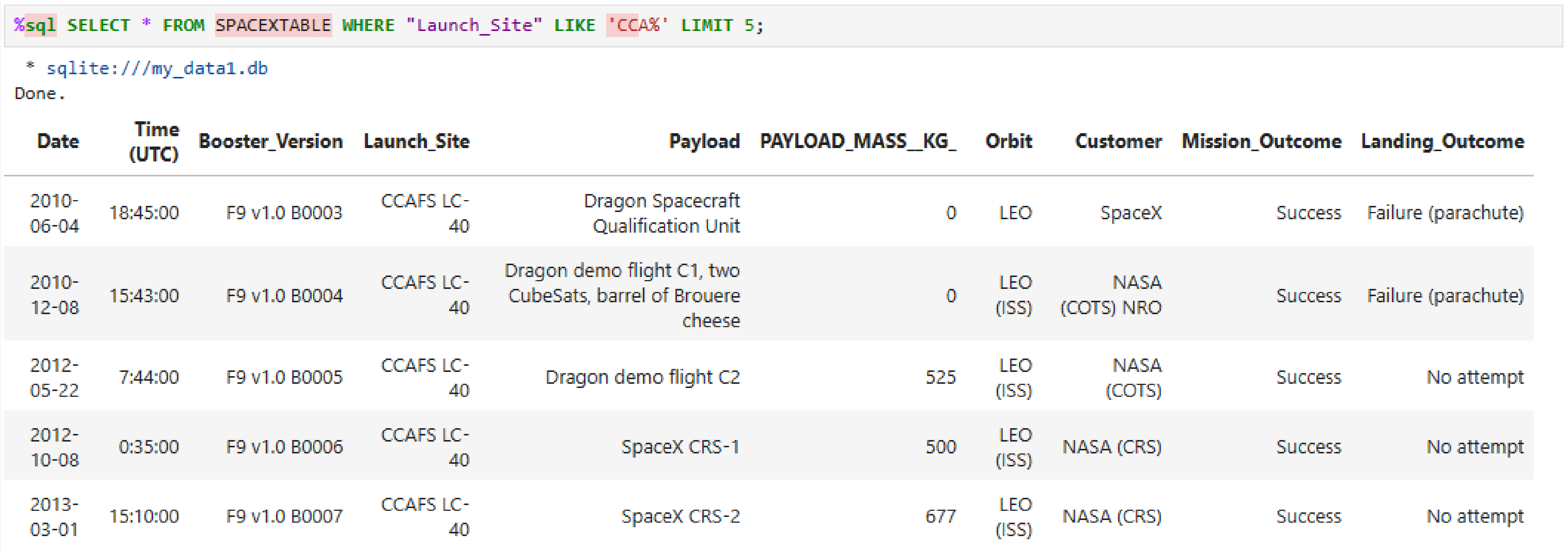
All Launch Site Names

- We use DISTINCT to show only unique launch sites



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

Launch Site Names Begin with 'CCA'

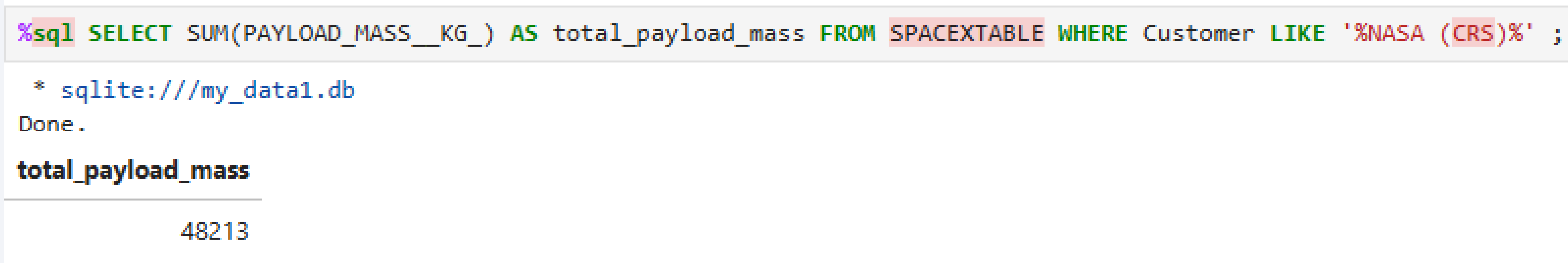


```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

```
 * 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

Total Payload Mass

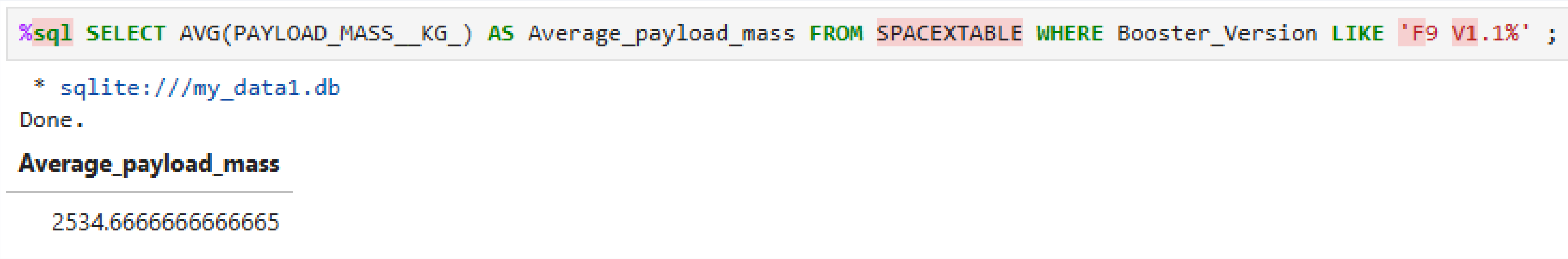


```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%' ;
```

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

total_payload_mass
48213

Average Payload Mass by F9 v1.1

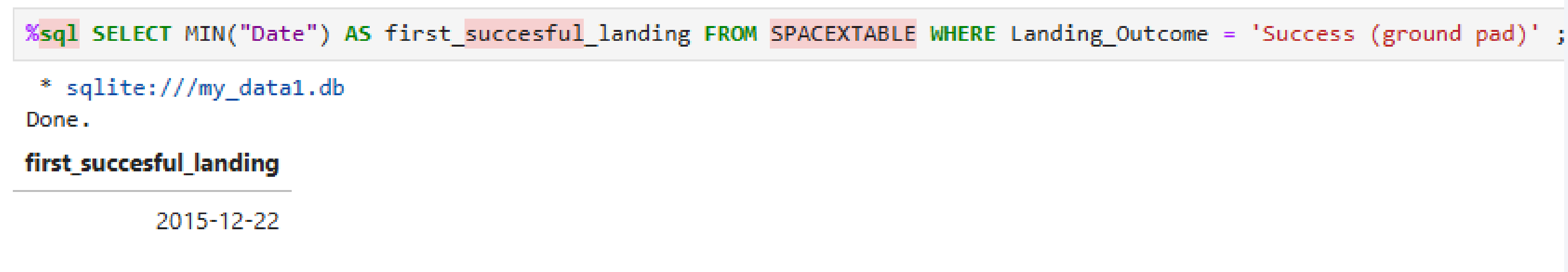


```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_payload_mass FROM SPACEXTABLE WHERE Booster_Version LIKE 'F9 V1.1%' ;
```

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

Average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

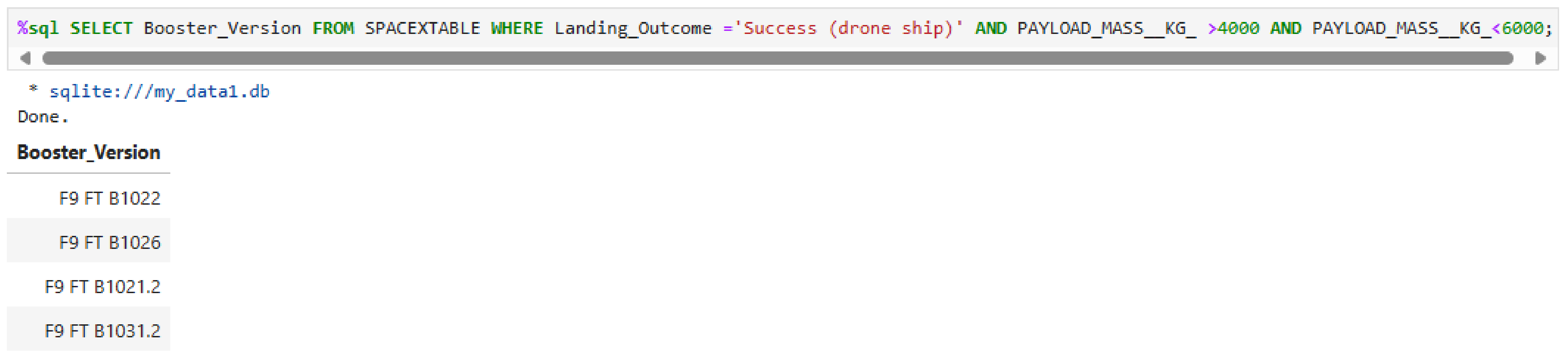


```
%sql SELECT MIN("Date") AS first_succesful_landing FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)' ;
```

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

first_succesful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

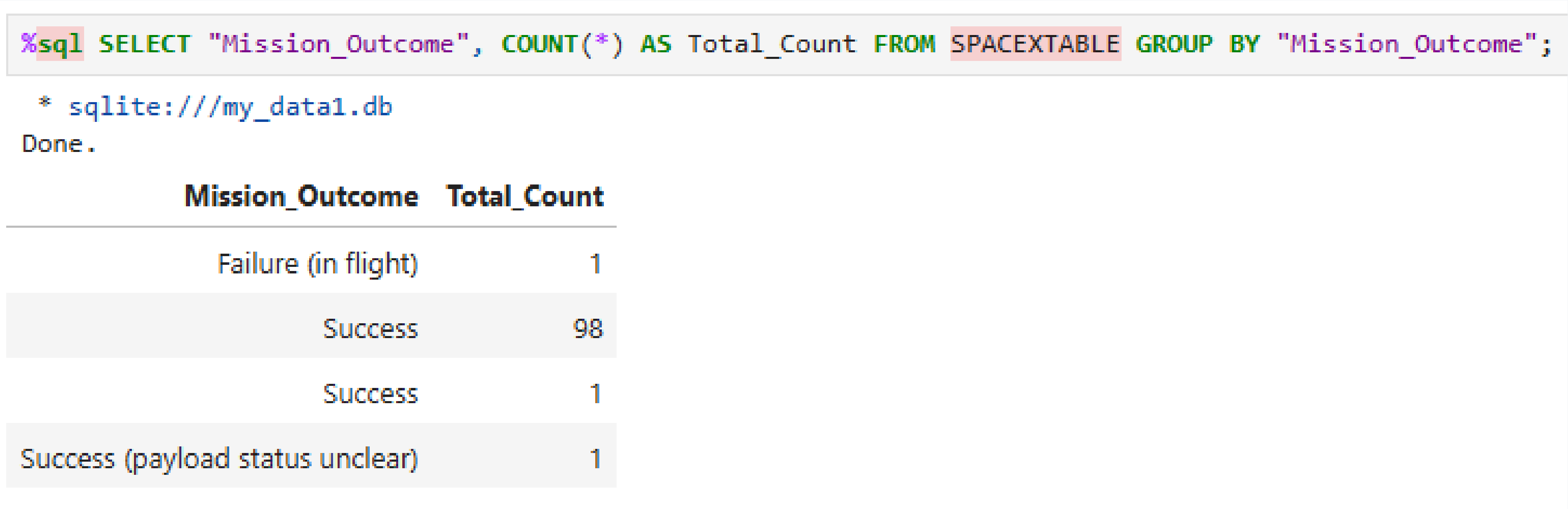


```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome ='Success (drone ship)' AND PAYLOAD_MASS__KG_ >4000 AND PAYLOAD_MASS__KG_<6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

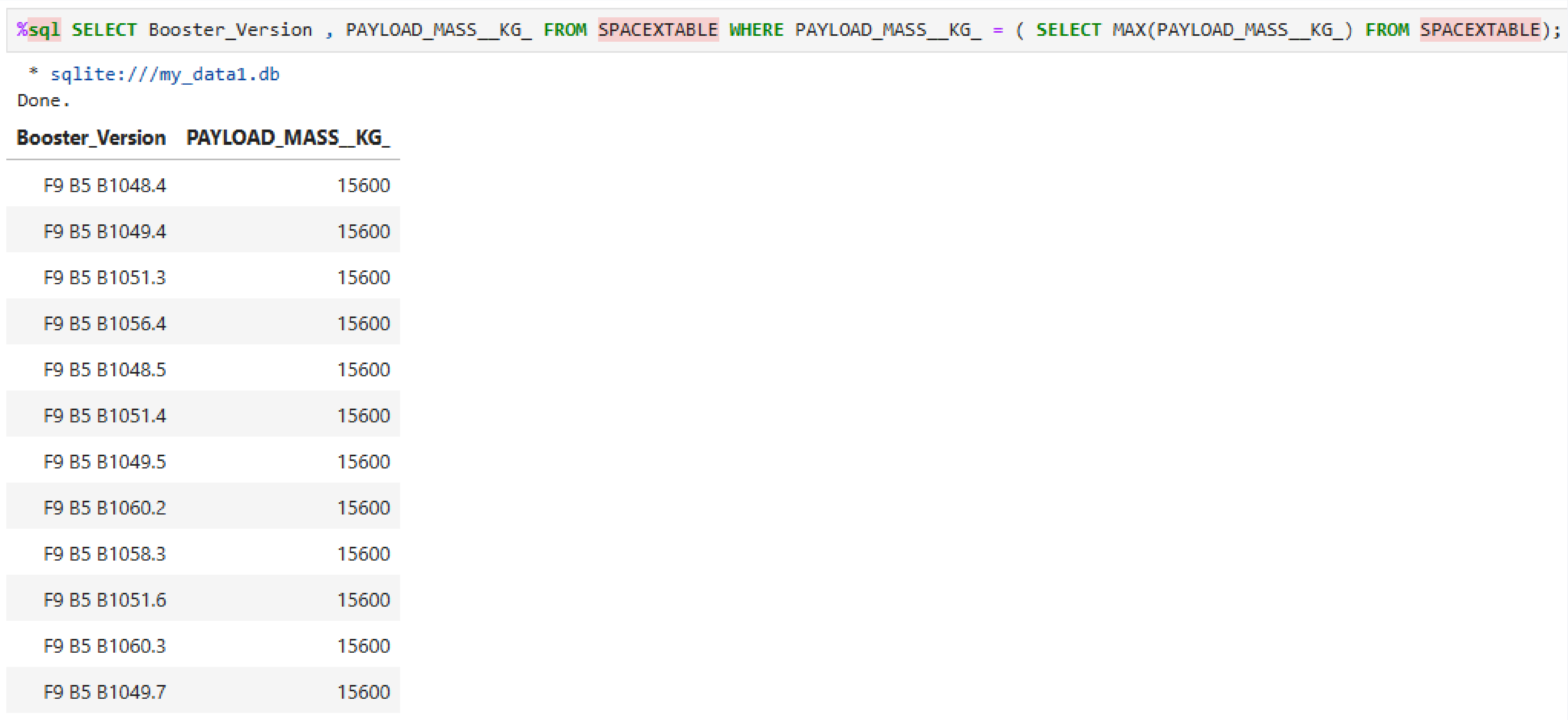


```
%sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

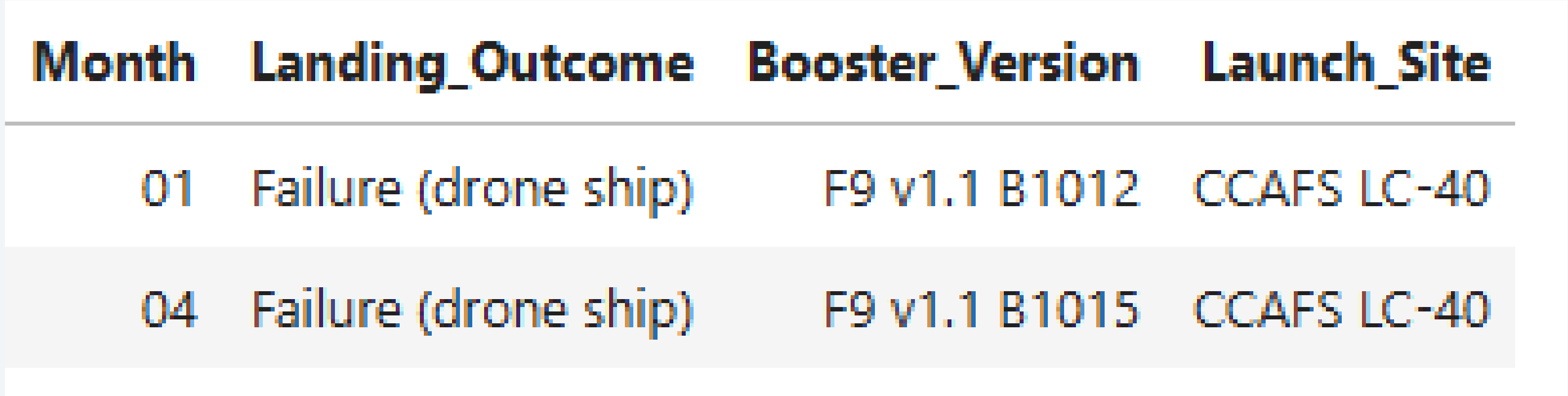


```
%sql SELECT Booster_Version , PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = ( SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

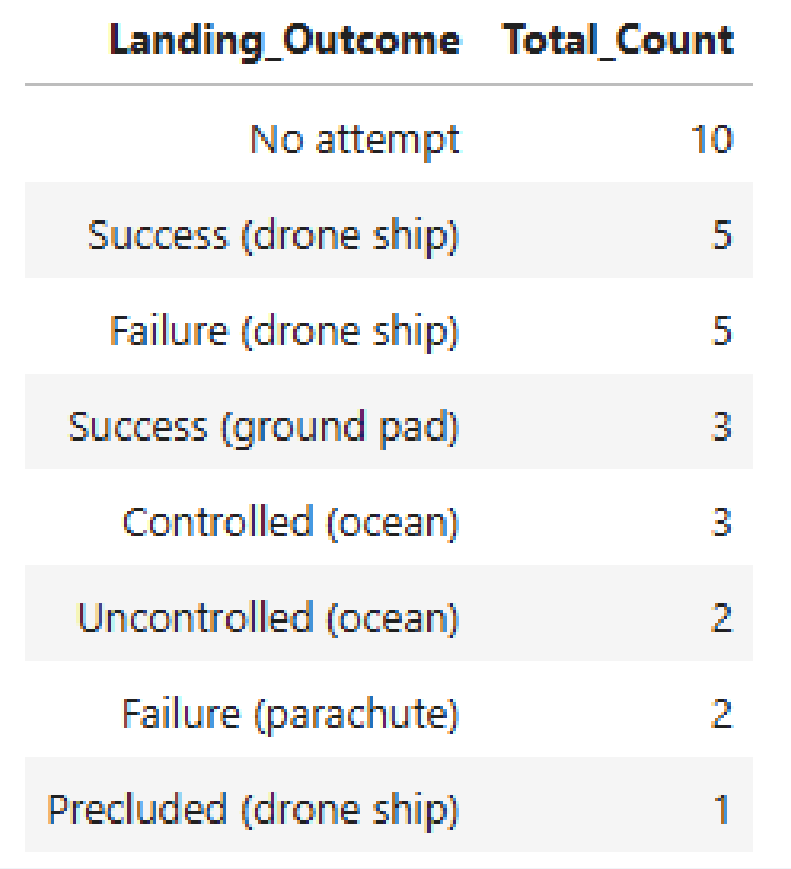
Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records



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

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

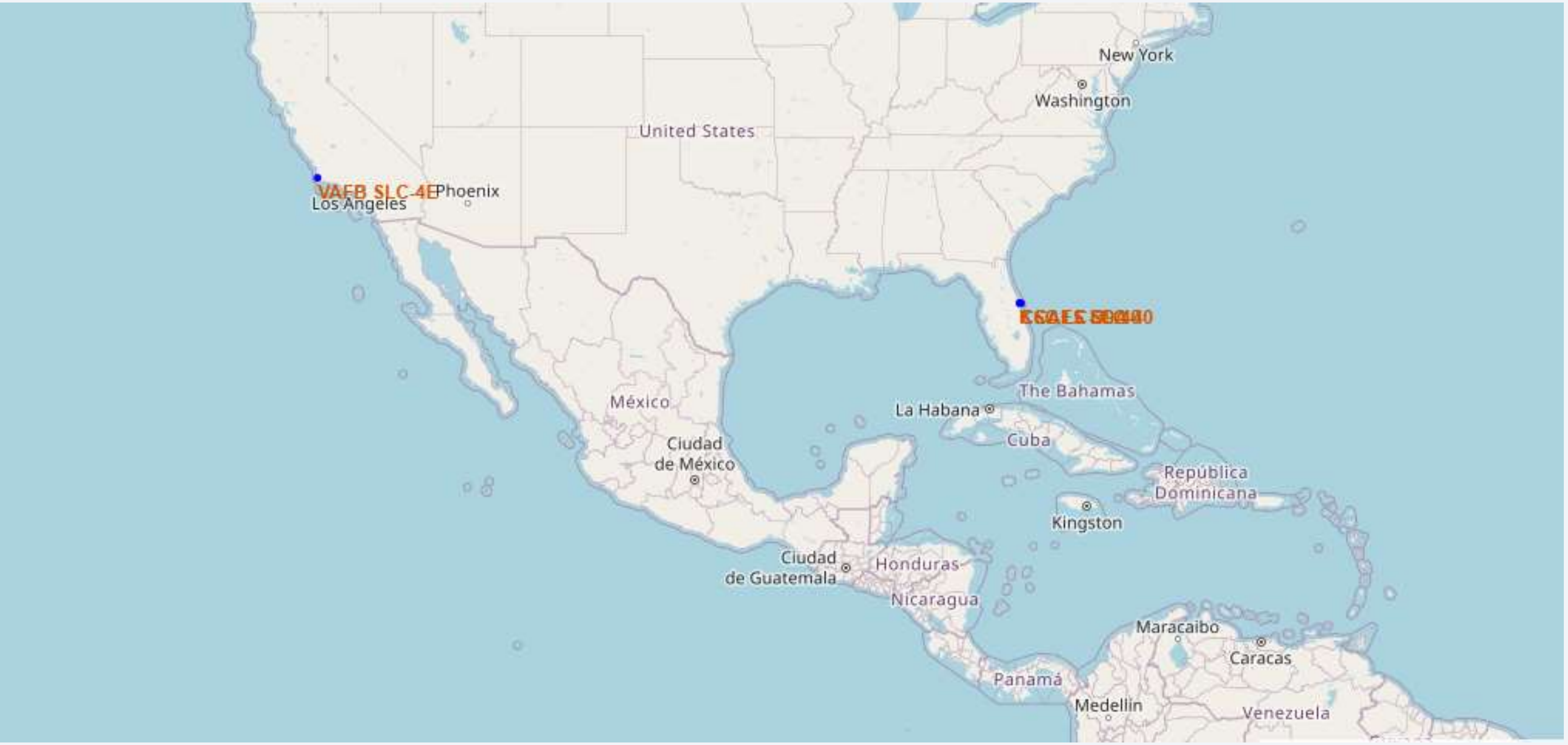


Landing_Outcome	Total_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

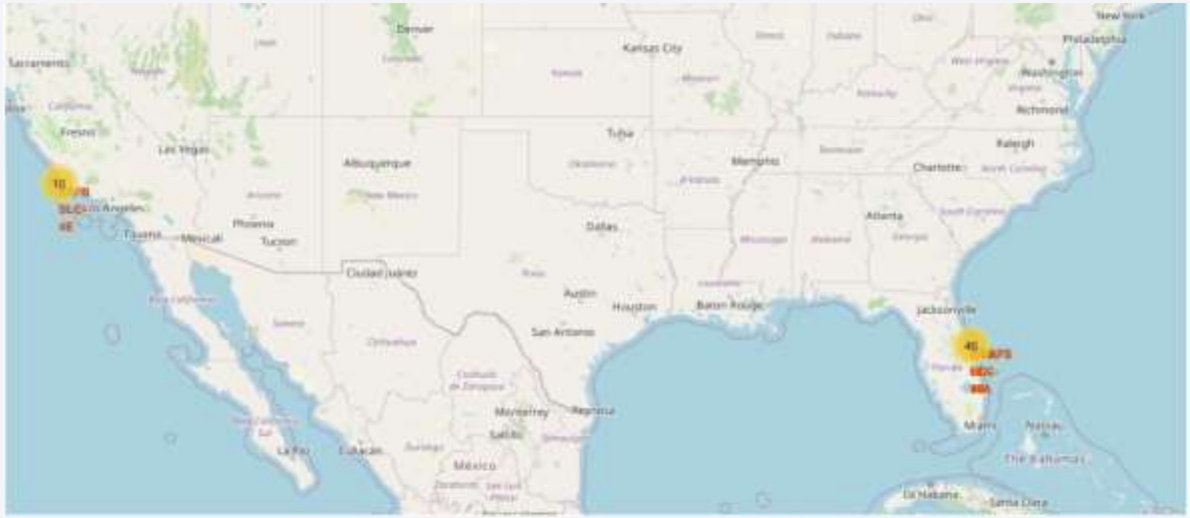
Section 3

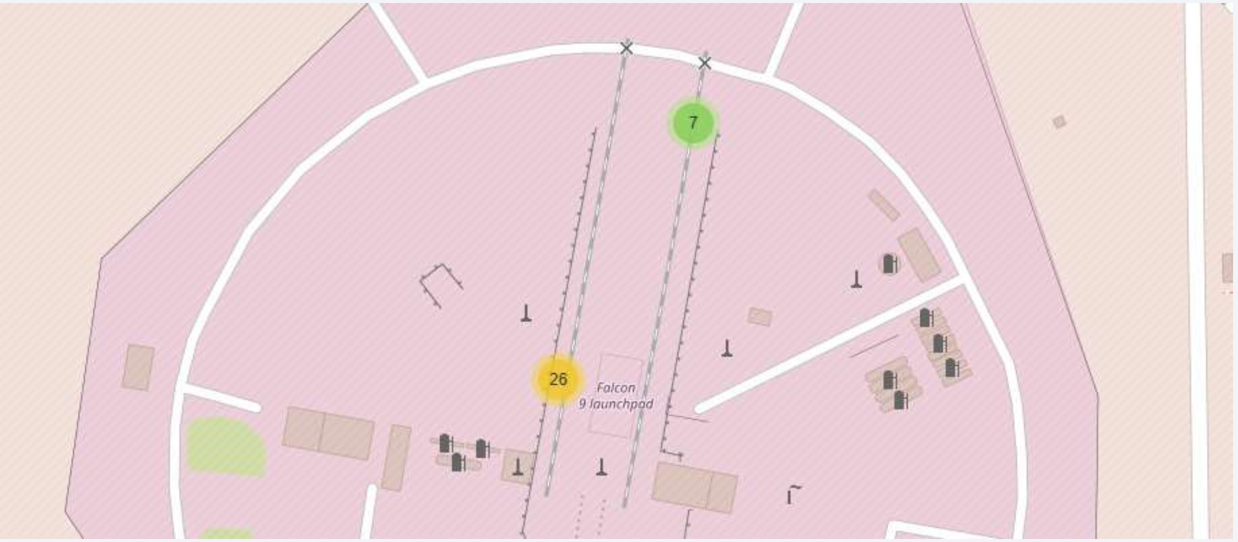
Launch Sites Proximities Analysis

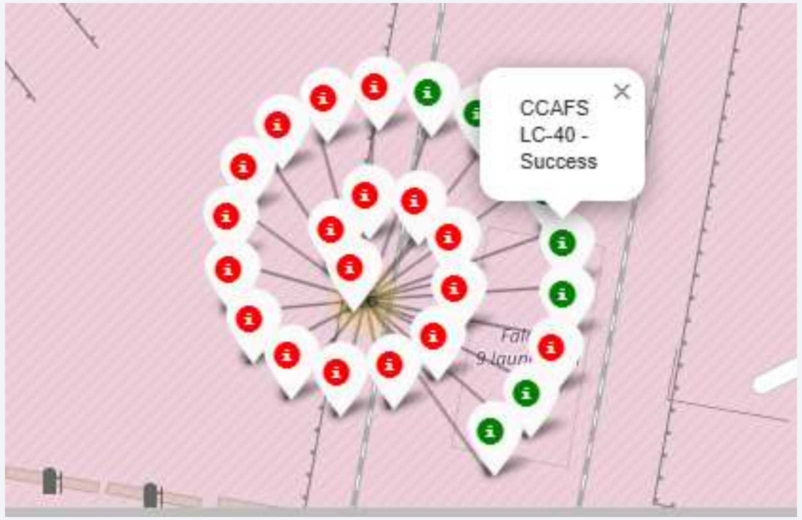
all launch sites on a map



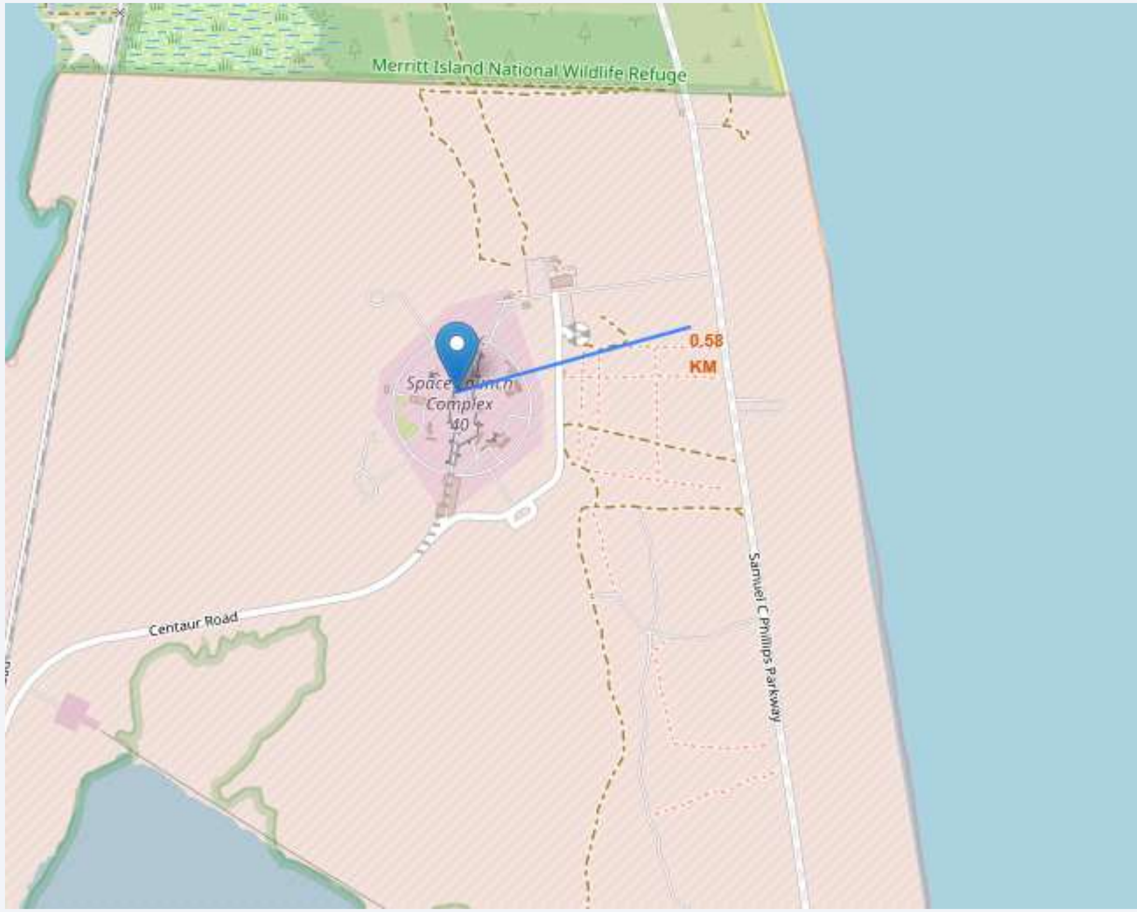
success/failed launches





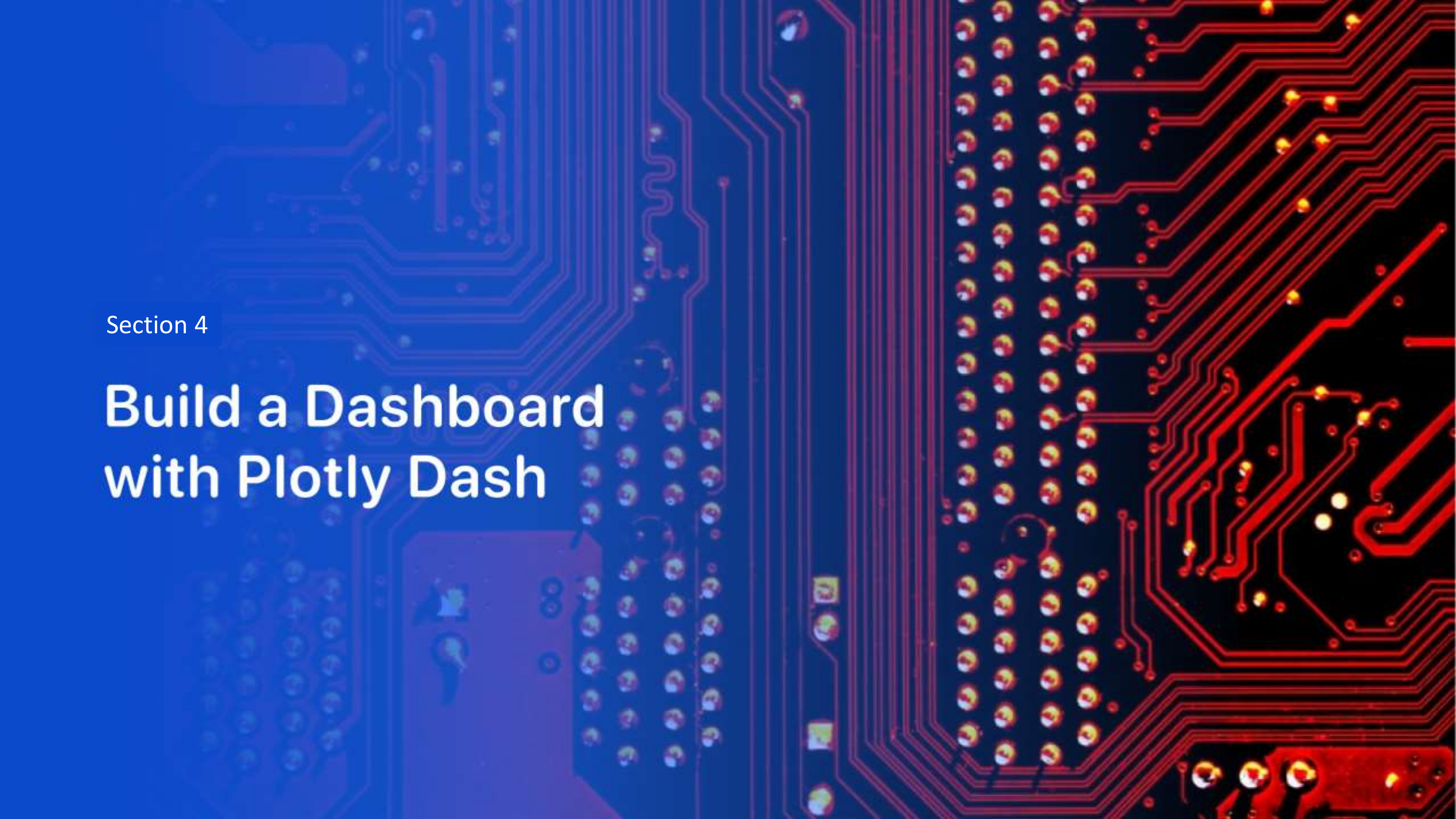


distances between a launch site

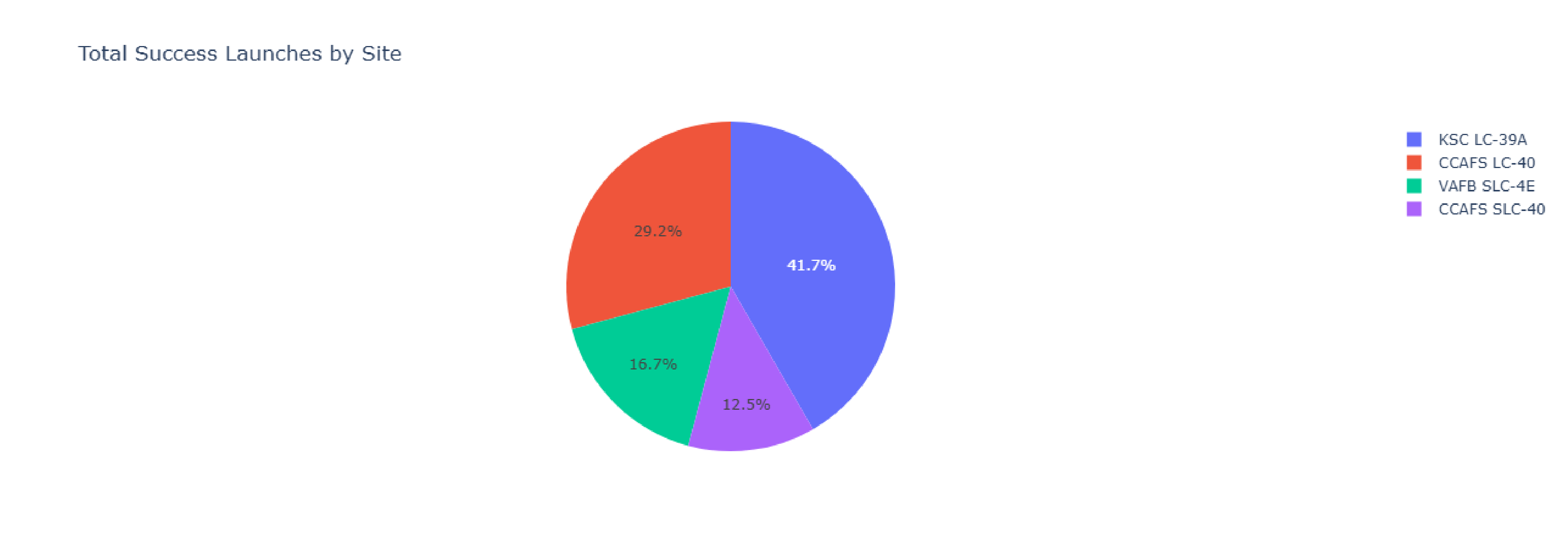

Merritt Island National Wildlife Refuge
Space Launch Complex 40
0.58 KM
Centaur Road
Samuel C Phillips Parkway

Section 4

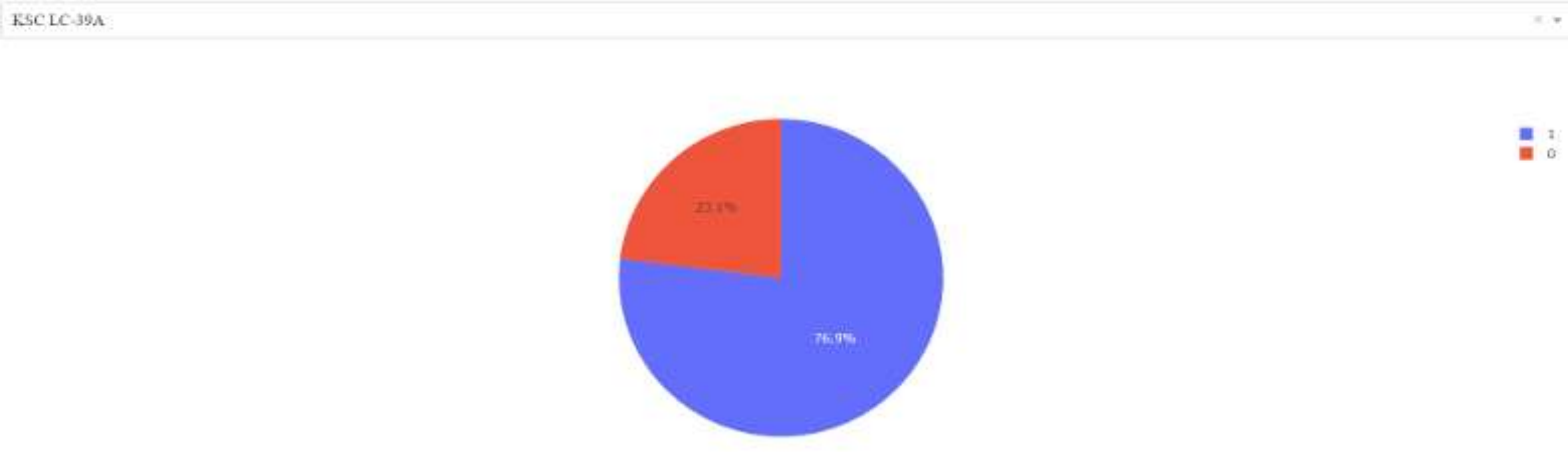
Build a Dashboard with Plotly Dash



Total Success By All Sites

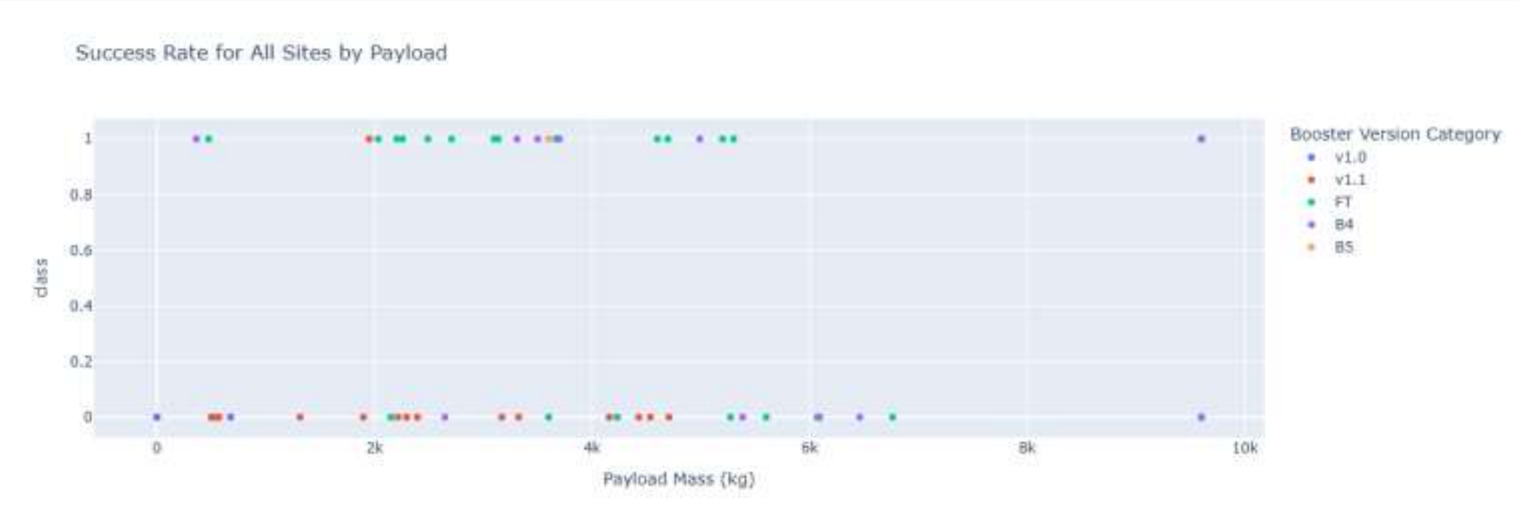


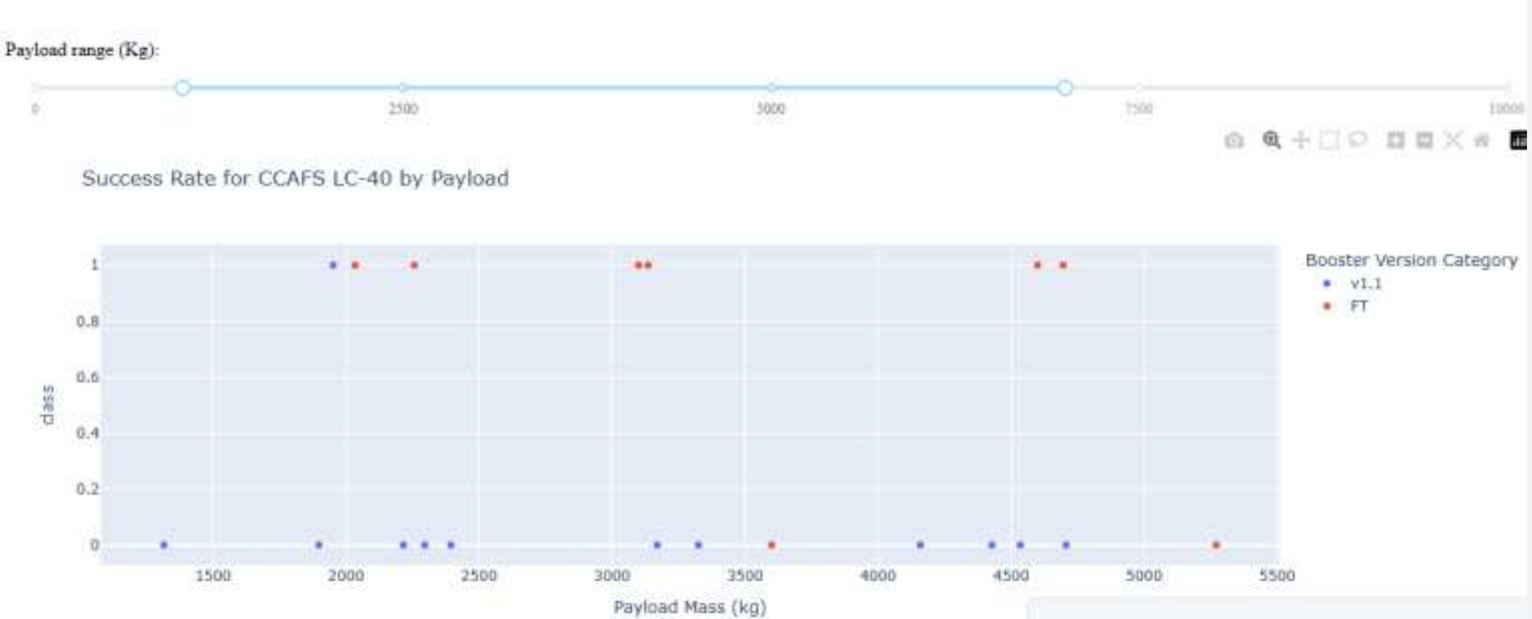
Pie chart for the launch site with highest launch success



KSC LC-39A

Launch Outcome for all sites, with different payload selected

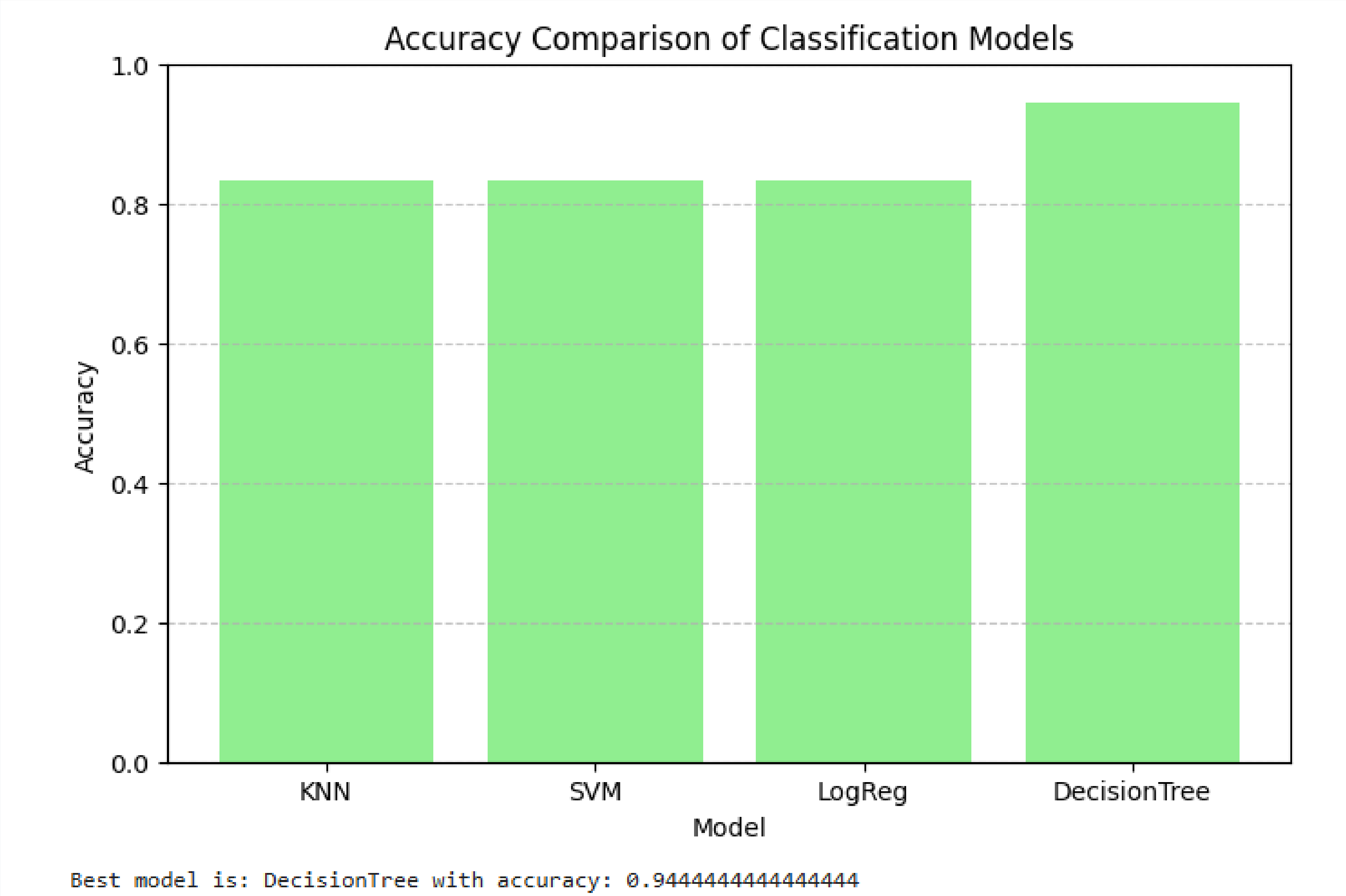




Section 5

Predictive Analysis (Classification)

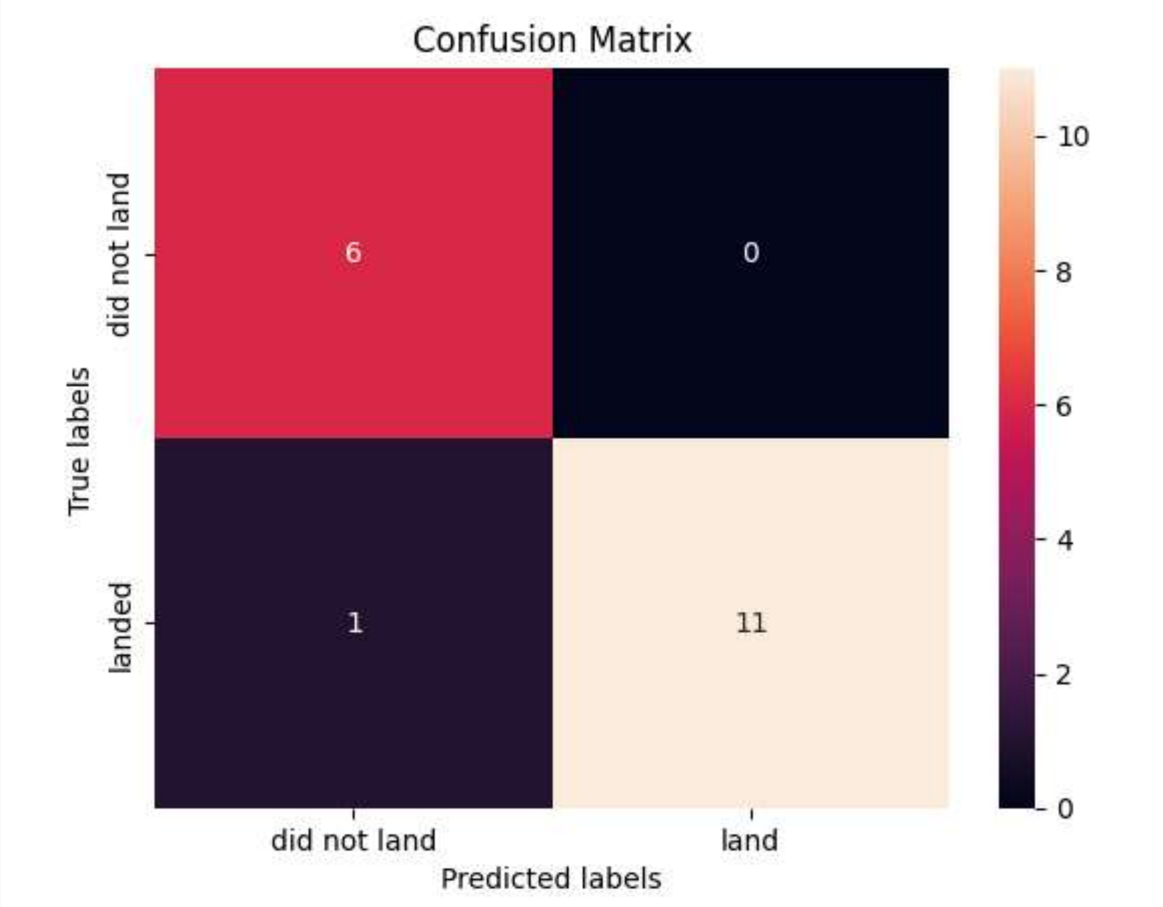
Classification Accuracy



```
Best model is: DecisionTree with accuracy: 0.9444444444444444
```

Confusion Matrix

- This confusion matrix is for a Decision Tree model predicted 17 out of 18 cases correctly.
- It got 94% accuracy.
- 6 correct “did not land”
- 11 correct “landed”
- 1 wrong “did not land” (it actually landed)
- No false “landed” predictions.



Conclusions

•Launch success rates have steadily increased from 2010 to 2020, with a notable rise starting in 2013.

•Launch sites with a higher number of flights, such as CCAFS SLC 40 and KSC LC-39A, showed greater success rates.

•Orbits like ES-L1, SSO, GEO, and HEO had a 100% landing success rate, making them the most reliable.

•There is a positive correlation between payload mass and launch success.

•Among all the classification models tested, the **Decision Tree Classifier** performed best with a **training accuracy of 88.9%** and a **test accuracy of 94%**.

•This high accuracy demonstrates the model's effectiveness in predicting Falcon 9 landing outcomes, supporting better mission planning in the future.

Appendix

- All relevant assets included in this presentation can be found GitHub
- https://github.com/Noura-DS/IBM_FINAL_PROJECT

Thank you!