IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shiva Sai Pavan Inja
03-10-2025

Outline

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

Executive Summary

- ## Summary of Methodologies
- **Data Collection:** Used APIs & web scraping to gather launch data.
- **Data Wrangling:** Cleaned and transformed data using one-hot encoding.
- **Exploratory Data Analysis:** SQL queries and visualizations to identify trends.
- **Interactive Visual Analytics:** Built dashboards and maps using Plotly and Folium.
- **Predictive Modeling:** Compared Logistic Regression, SVM, Decision Trees, and KNN to predict landing success.
- ## Summary of Results
- **Key Insights:** Payload mass, launch site, and orbit type significantly impact success.
- **Best Model:** Decision Tree classifier had the best predictive accuracy.
- **Trends:** Landing success rates improved over time due to technological advancements.

Introduction

- **Project Background and Context**
- Focuses on **SpaceX Falcon 9 first-stage booster landings**.
- Uses data science techniques to analyze launch outcomes.
- **Problems You Want to Find Answers**
- **What factors influence successful landings?**
- **Can machine learning models predict success accurately?**
- **How has the landing success rate evolved over time?**

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Retrieved launch data using the SpaceX API and supplemented it with web scraping from Wikipedia.
- Perform data wrangling
    - Cleaned and transformed data by handling missing values, normalizing features, and applying one-hot encoding for categorical variables.
- 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

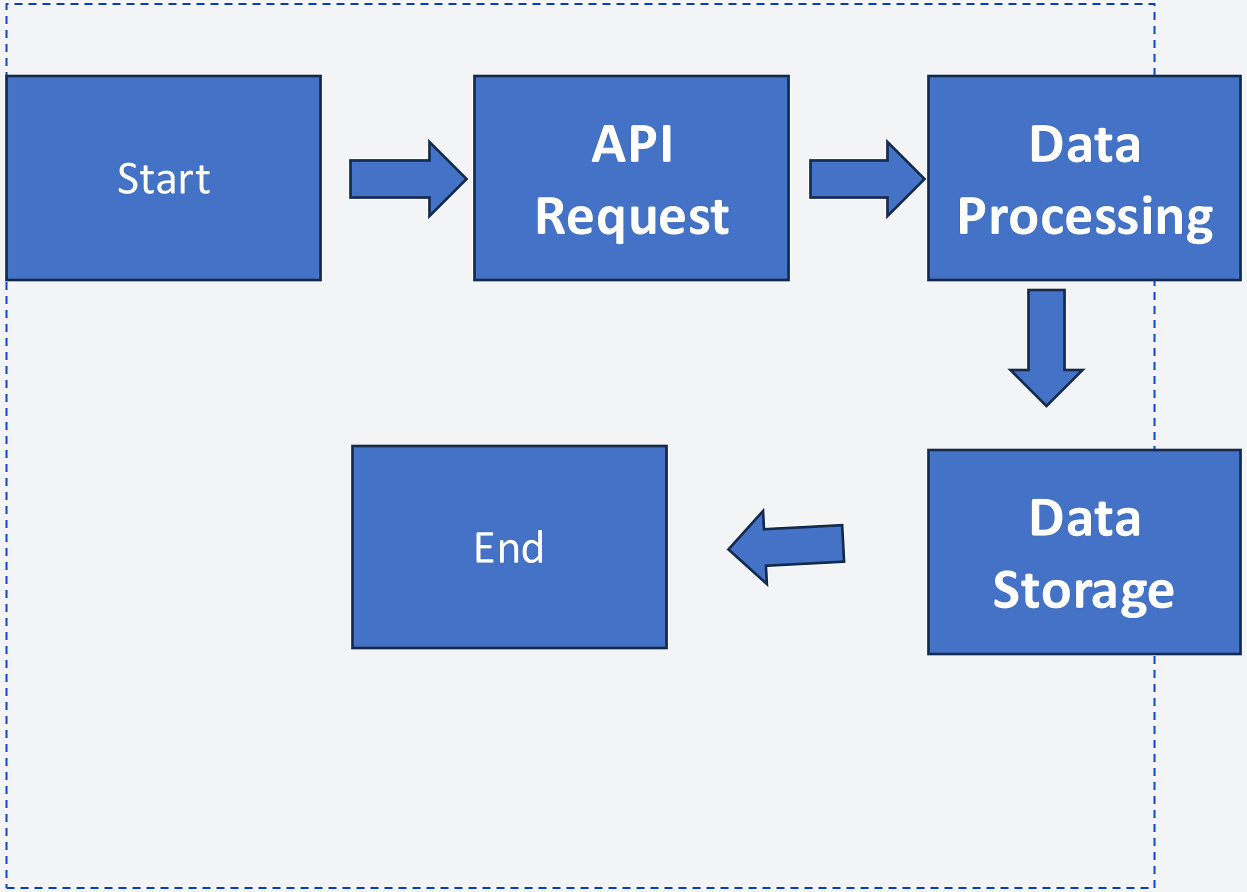
Data Collection

- **Key Steps in Data Collection:**
- **Retrieving Data from SpaceX API** – Used API calls to gather launch records, including launch sites, payloads, and success/failure outcomes.
- **Web Scraping from Wikipedia** – Extracted additional mission details to supplement missing or incomplete data.
- **Storing Data in CSV/Database** – Structured data into CSV files for further processing and analysis.

Data Collection – SpaceX API

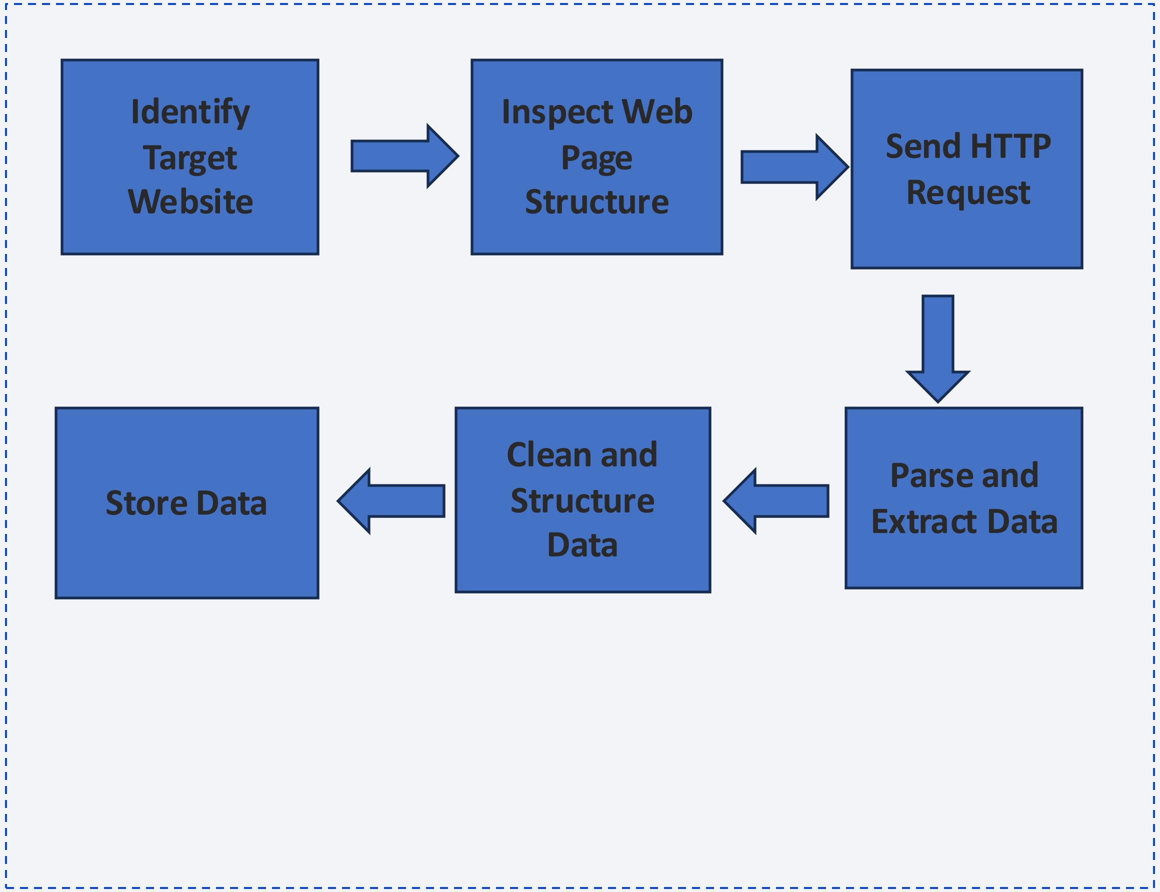
- **API Request:** Making GET requests to SpaceX API endpoints.
- **Data Processing:** Parsing JSON responses and converting to DataFrame.
- **Data Storage:** Saving data for further analysis.

- github link



Data Collection - Scraping

- **Key Steps in Web Scraping:**
- **Identify Target Website** – Select Wikipedia as the source for supplementary launch details.
- **Inspect Web Page Structure** – Use browser developer tools to locate relevant HTML elements.
- **Send HTTP Request** – Fetch the webpage content using `requests` or `BeautifulSoup`.
- **Parse and Extract Data** – Locate tables or key information and extract the required data.
- **Clean and Structure Data** – Remove unnecessary elements and format data into a structured dataset.
- **Store Data** – Save cleaned data in CSV or database for further processing.
- [Github Link]()



Data Wrangling

- Data wrangling, including filtering and handling missing values.
- Applied One Hot Encoding for binary classification.
- Conducted exploratory data analysis (EDA) using visualizations and SQL queries.
- Performed interactive visual analytics with tools like Folium and Plotly Dash.
- Built, tuned, and evaluated machine learning models to predict SpaceX rocket landing success.

- Github link

EDA with Data Visualization

Different types of plots used for data visualization :

Scatter Plots

- Flight Number vs. Launch Site – Assesses how flight frequency affects success.
- Payload Mass vs. Launch Site – Examines mass variations across sites.
- Orbit Type vs. Flight Number – Explores orbit type correlations with launch sequence.
- Payload Mass vs. Orbit Type – Analyzes payload mass distribution across orbits.

Bar Charts

- Success Rate vs. Orbit Type – Identifies which orbit types have higher success rates.

Line Charts

- Success Rate Over Time – Tracks launch success improvements or declines.
- Github link

EDA with SQL

- I wrote SQL queries for the following tasks :
- Displayed the names of the unique launch sites in the space mission
- Displayed 5 records where launch sites begin with the string 'CCA'
- Calculated the total payload mass carried by boosters launched by NASA (CRS) and average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved, names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, total successful and failure outcomes and names of booster landers that carried maximum payload
- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Github link

Build an Interactive Map with Folium

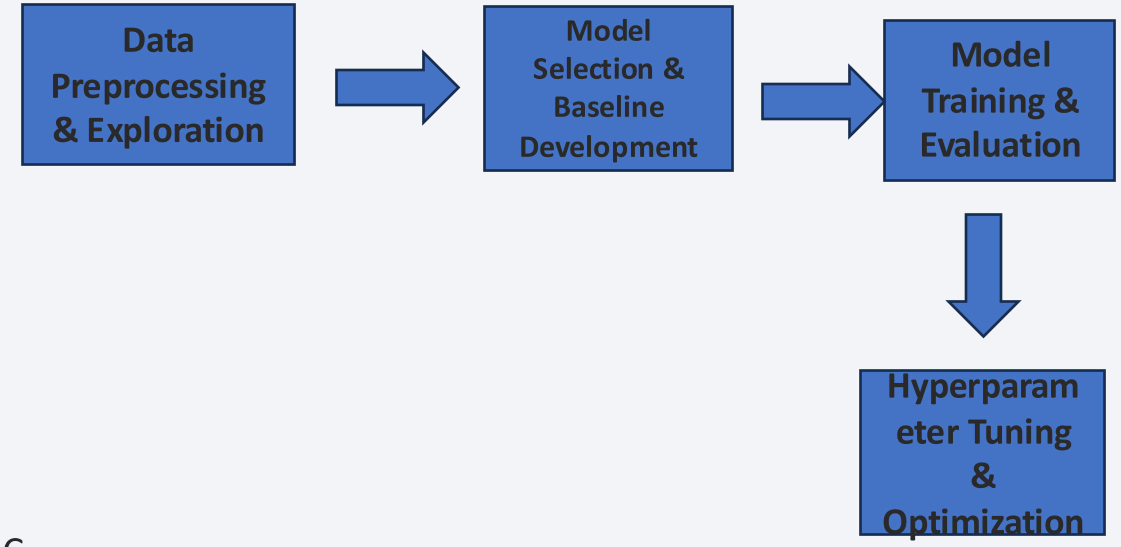
- **Folium** created by utilizing interactive maps with the following objects:
- **Markers**: Placed at key locations, such as launch sites, to highlight important points.
- **Circles**: Used to represent areas of interest, like impact zones or launch regions.
- **Lines/Polylines**: Drawn to indicate flight paths or trajectories of rockets.
- These objects help in visualizing spatial relationships, trends, and important locations for better understanding launch patterns
- Github link

Build a Dashboard with Plotly Dash

- **Key Factors Affecting Landing Success**: The dashboard highlights that factors such as **launch site, payload mass, and orbit type** significantly influence the success of the Falcon 9 first-stage landings.
- **Interactive Data Exploration**: Users can dynamically filter and explore **launch success rates** across different launch sites and payload ranges, making it an effective tool for data-driven decision-making.
- **Visualization of Trends**: The dashboard effectively visualizes patterns in launch outcomes, showing how **heavier payloads or specific launch sites** may impact landing success probabilities.
- **Business and Engineering Applications**: The insights derived can help **optimize launch strategies, reduce costs, and improve future mission planning**, providing valuable data for stakeholders in aerospace engineering and space exploration.
- Github link

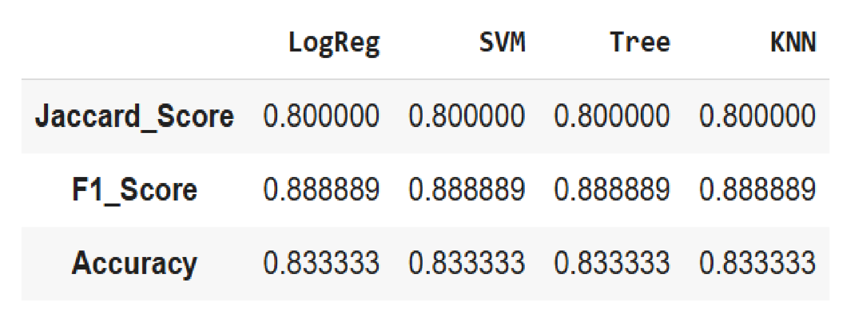
Predictive Analysis (Classification)

- **Data Preprocessing & Exploration**
- Loaded and cleaned dataset
- Handled missing values, outliers, and performed feature engineering
- Conducted exploratory data analysis (EDA)
- **Model Selection & Baseline Development**
- Selected multiple classification models (e.g., Logistic Regression, Random Forest, XGBoost)
- Implemented train-test split and applied feature scaling
- **Model Training & Evaluation**
- Trained models using key metrics (Accuracy, Precision, Recall, F1-score, AUC-ROC)
- Used cross-validation for robust evaluation
- **Hyperparameter Tuning & Optimization**
- Applied GridSearchCV/RandomizedSearchCV for fine-tuning
- Reduced overfitting using regularization techniques
- Github Link

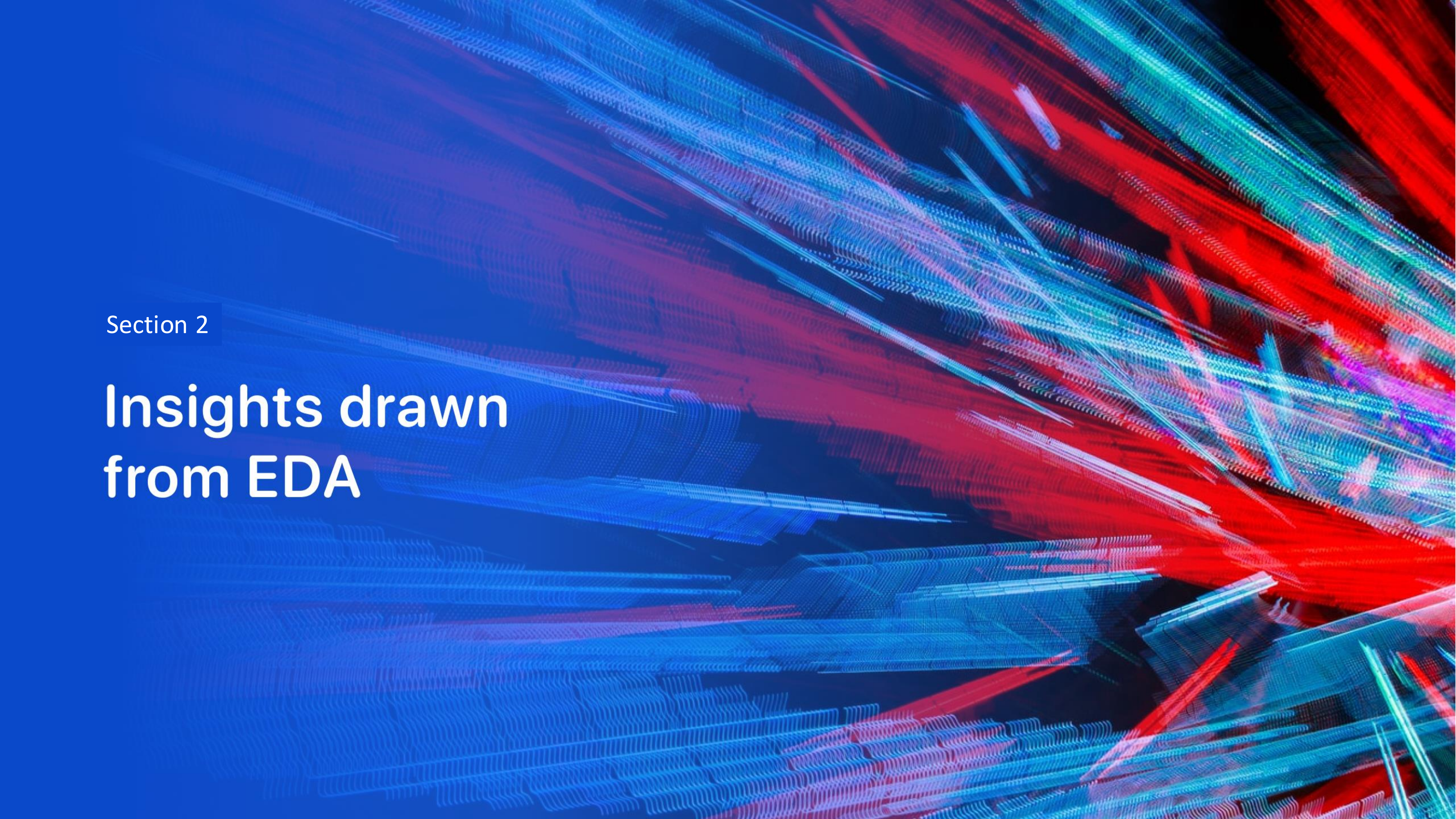

Data Preprocessing & Exploration
Model Selection & Baseline Development
Model Training & Evaluation
Hyperparameter Tuning & Optimization

Results

- Exploratory data analysis results
- **Launch Success Rate Trends**: The data analysis revealed patterns in launch success rates over time, with certain launch sites and payload masses having higher success probabilities.
- **Geospatial Insights**: Interactive maps showed how launch sites are geographically distributed and their proximity to key locations, helping in understanding regional launch preferences.
- **Feature Correlations**: Scatter plots and correlation heatmaps highlighted relationships between payload mass, orbit type, and mission success, guiding feature selection for predictive modeling
- Interactive analytics demo in screenshots
- Predictive analysis results



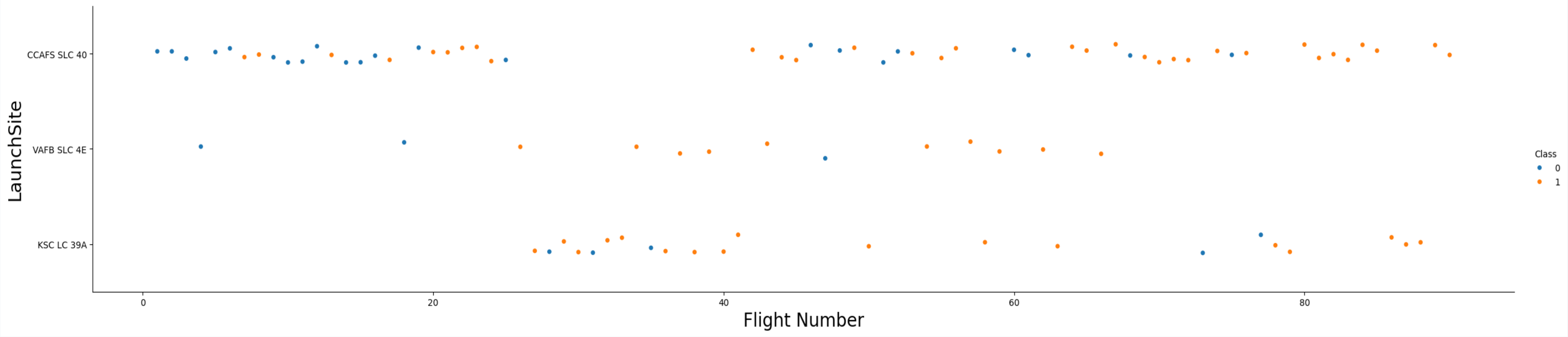
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



Section 2

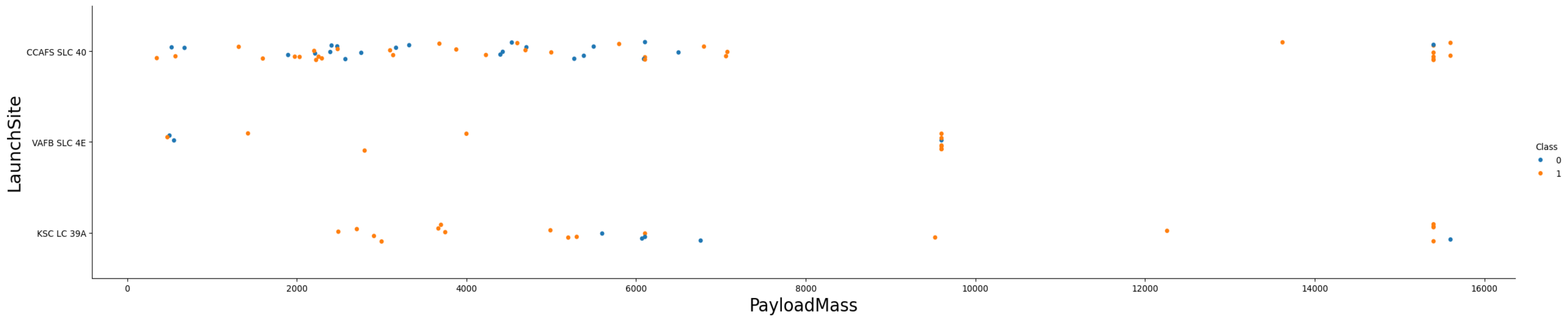
Insights drawn from EDA

Flight Number vs. Launch Site



- **Launch Sites** – The three distinct launch sites are **CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A**, each having multiple flight numbers associated with them.
- **Flight Outcomes** – The points are color-coded to indicate successful (Class 1, orange) and unsuccessful (Class 0, blue) launches.
- **Trends** –
- For some launch sites, success rates may improve over time as the flight number increases.
- Certain launch sites might have more successful launches than others.

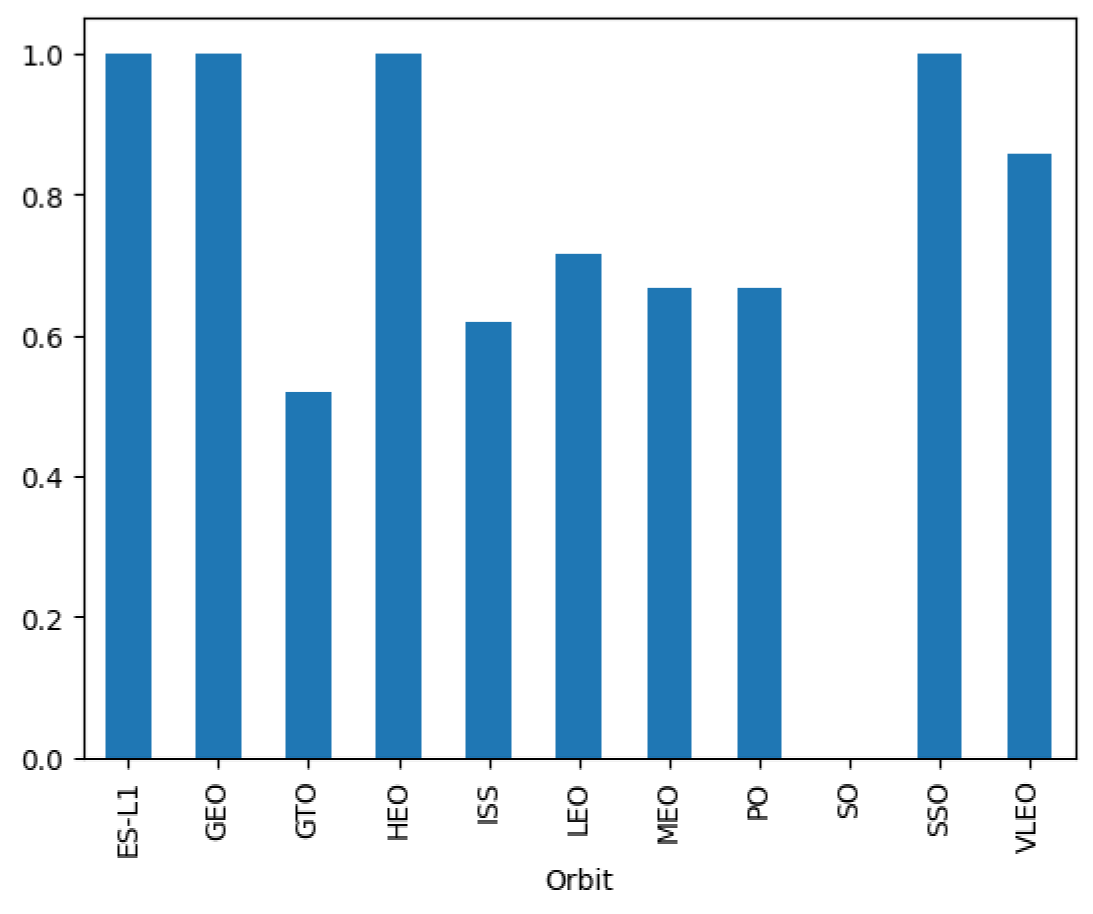
Payload vs. Launch Site



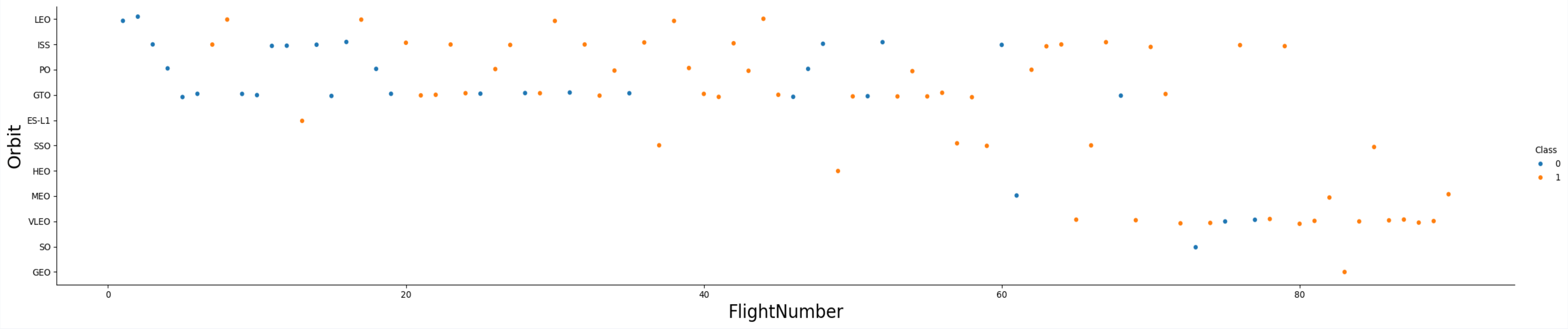
- For **lower payload masses**, both success and failure cases are observed.
- For **higher payloads (above ~10,000 kg)**, there seems to be a higher concentration of successful launches, indicating better reliability in those cases.
- Some clusters show mixed results, meaning payload mass alone is not the sole determinant of success.

Success Rate vs. Orbit Type

- The orbits such as GEO, HEO, SSO and ES-L1 has higher success rates followed by VLEO and LFO that has 80% and 70% respectively

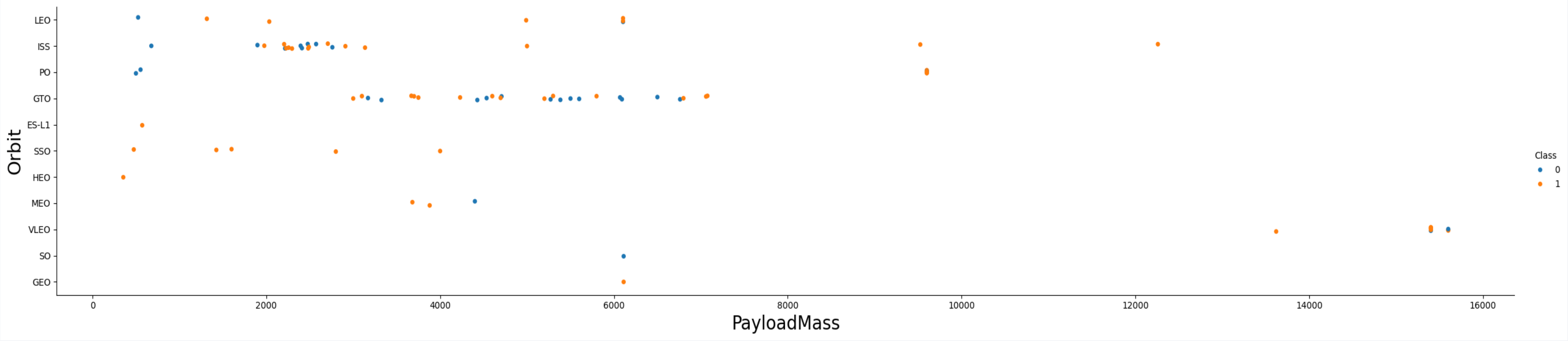


Flight Number vs. Orbit Type



- We observe that with an increase in time, the success rates of all the orbits increase

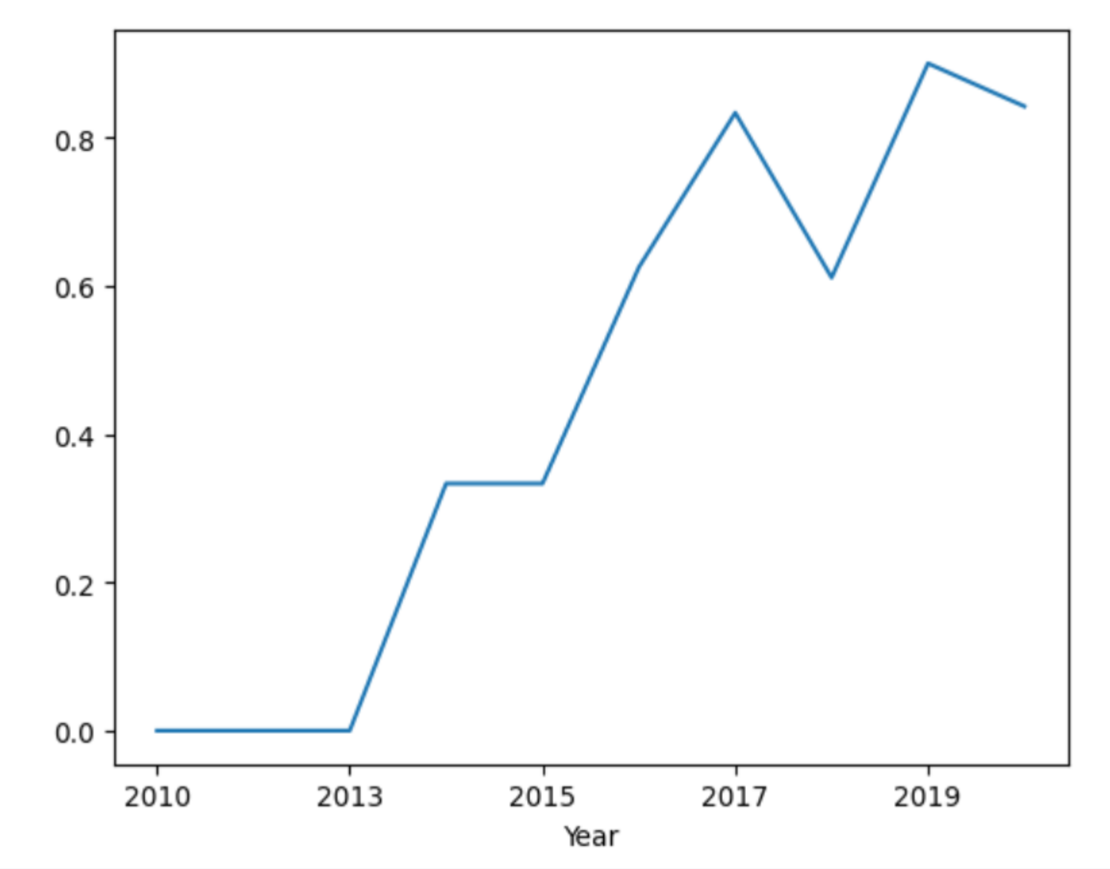
Payload vs. Orbit Type



- **LEO and ISS:** Mostly successful launches across all payload masses.
- **GTO:** Shows both successful and failed launches, indicating variability in success rate for this orbit.
- **Higher payloads (>10,000 kg):** Mostly successful, suggesting reliable performance for heavier payloads.

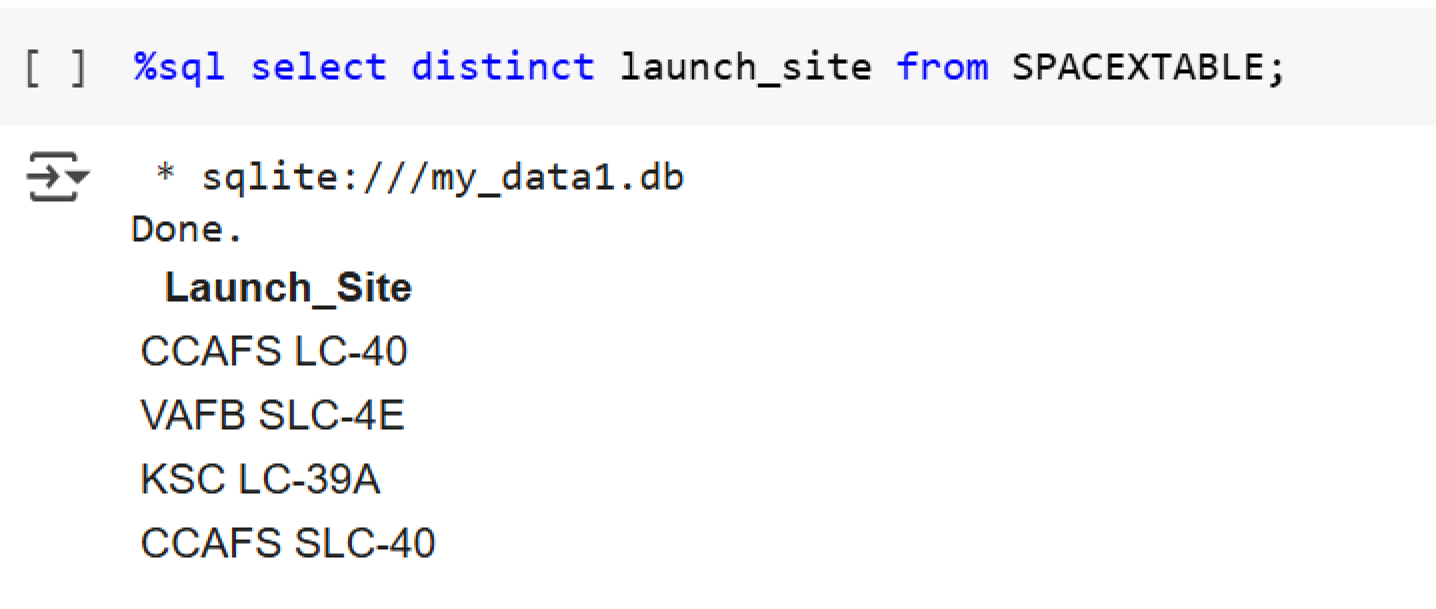
Launch Success Yearly Trend

- There was no change in the success rate in the initial three years. But after 2013, we observe an increase in the success rate and kept increasing until 2017 after which we see a slight decrease followed by increase in success rate.



All Launch Site Names

- Find the names of the unique launch sites

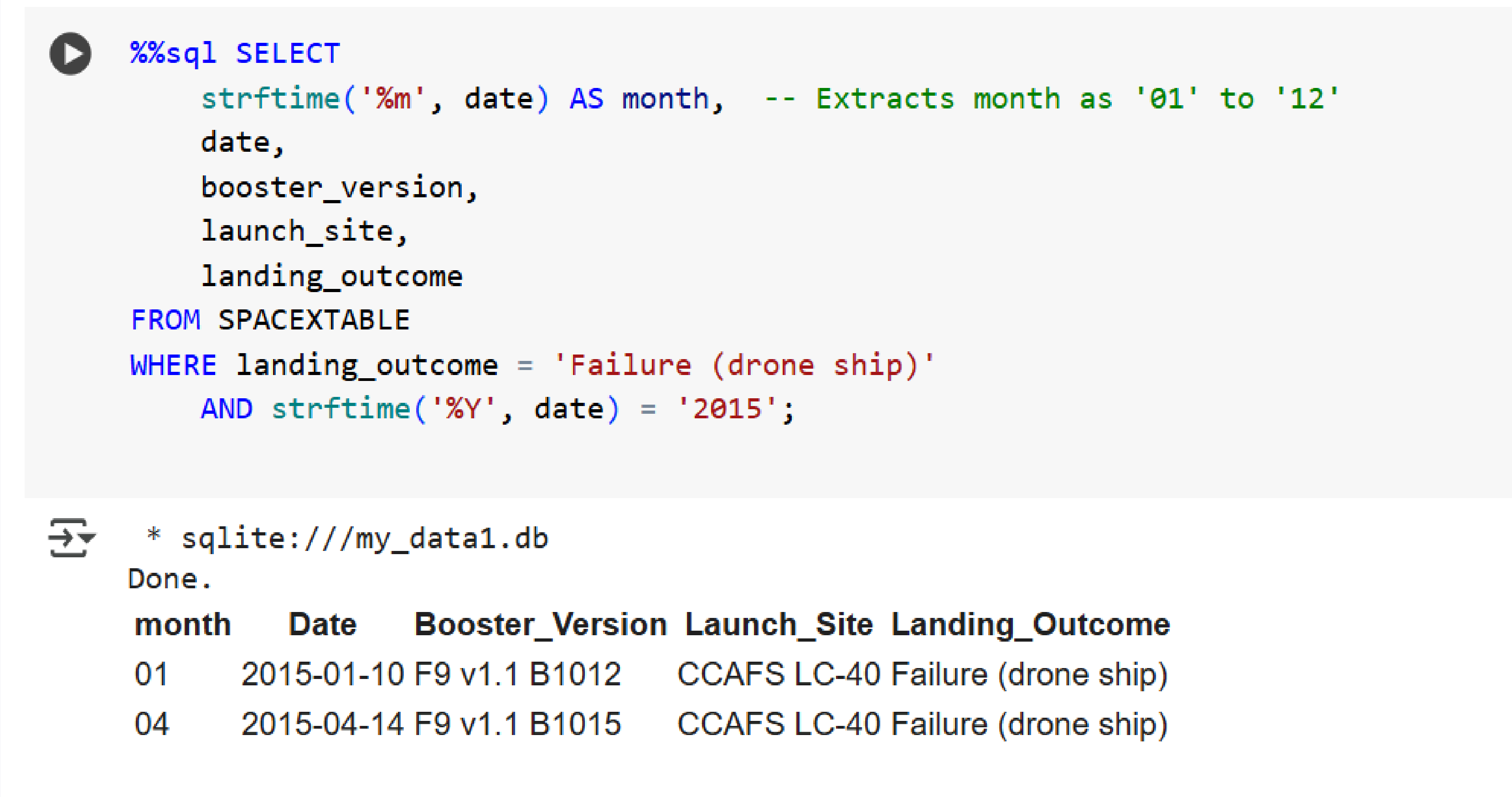


```
[ ] %sql select distinct launch_site from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
 Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`



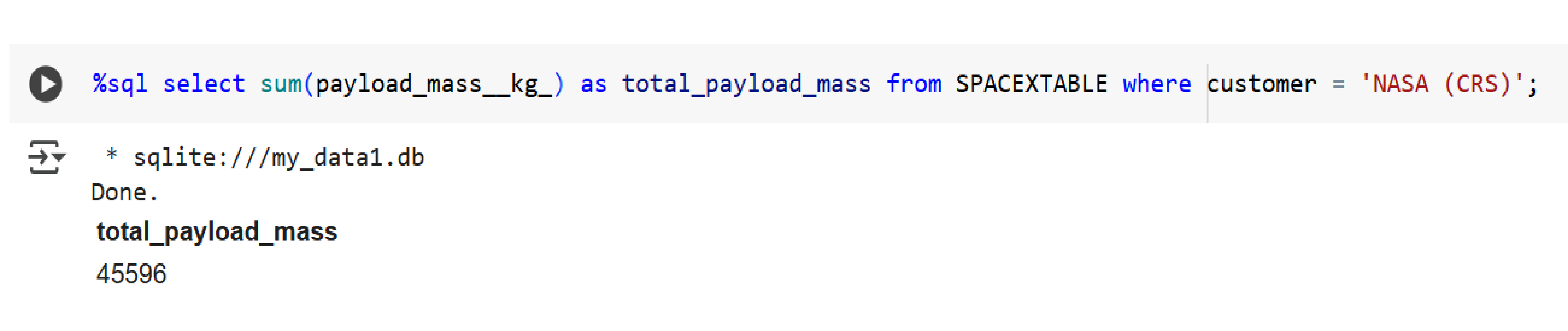
```
%%sql SELECT
    strftime('%m', date) AS month,  -- Extracts month as '01' to '12'
    date,
    booster_version,
    launch_site,
    landing_outcome
FROM SPACEXTABLE
WHERE landing_outcome = 'Failure (drone ship)'
    AND strftime('%Y', date) = '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

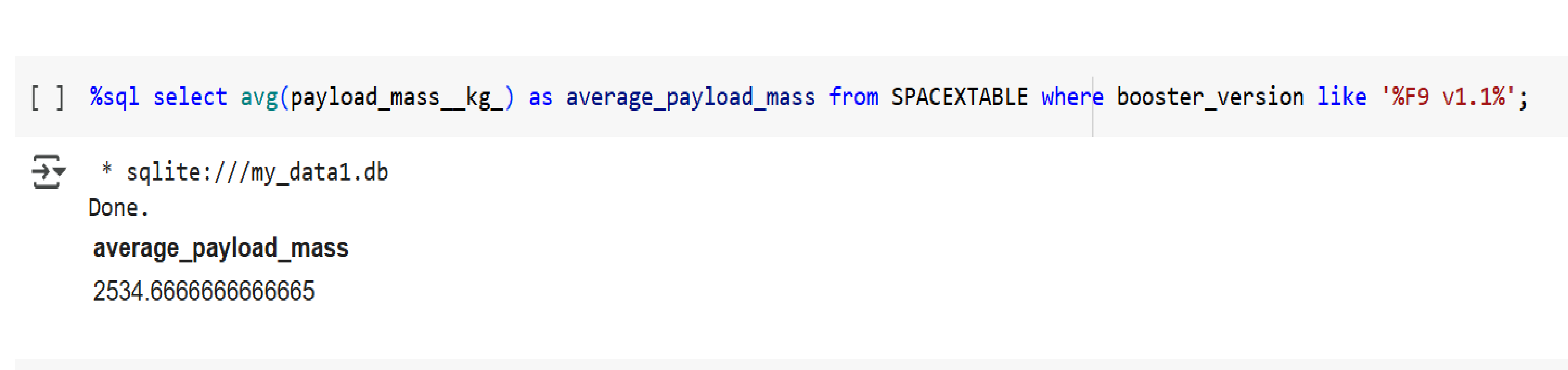


```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
total_payload_mass
45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

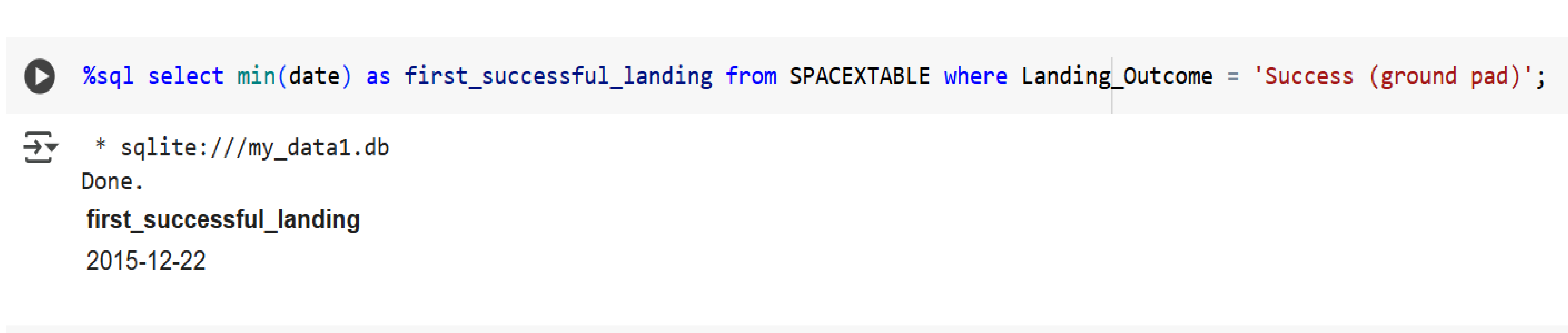


```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

```
 * sqlite:///my_data1.db
Done.
average_payload_mass
2534.6666666666665
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

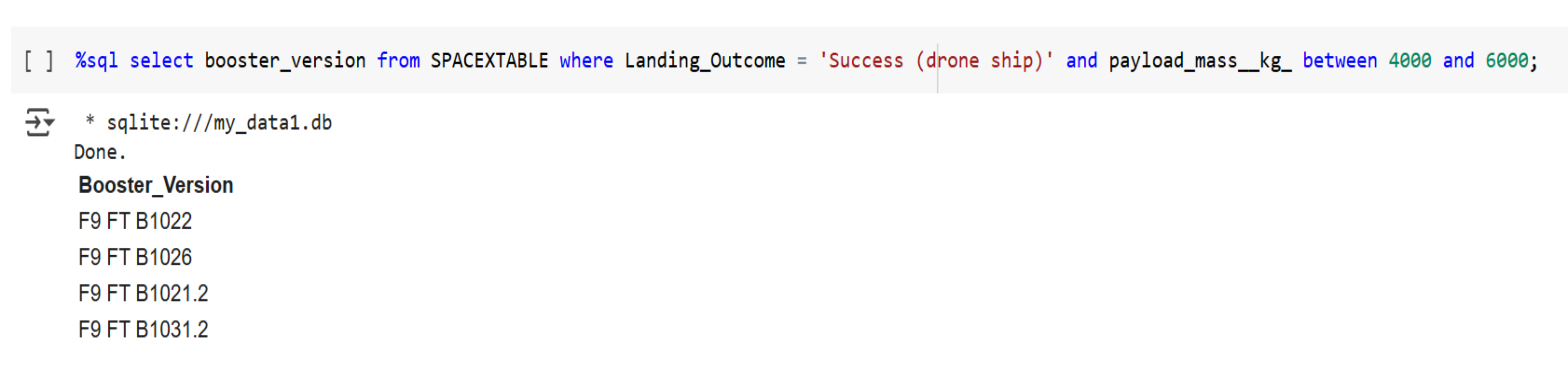


```
%sql select min(date) as first_successful_landing from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
first_successful_landing
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



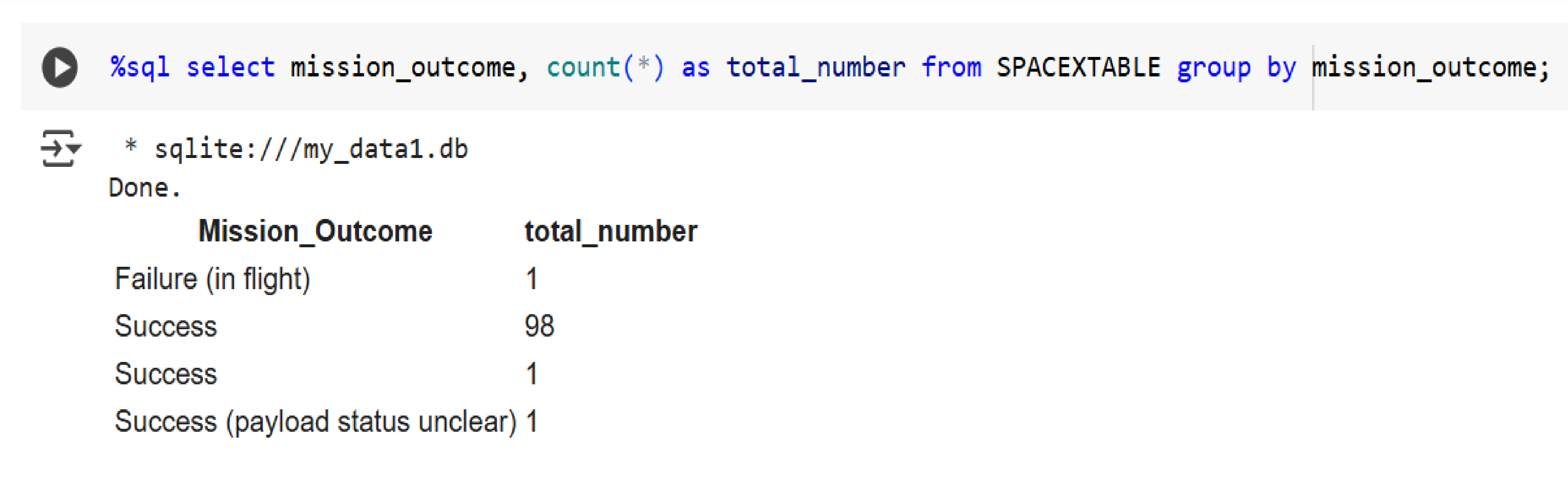
```
[ ] %sql select booster_version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



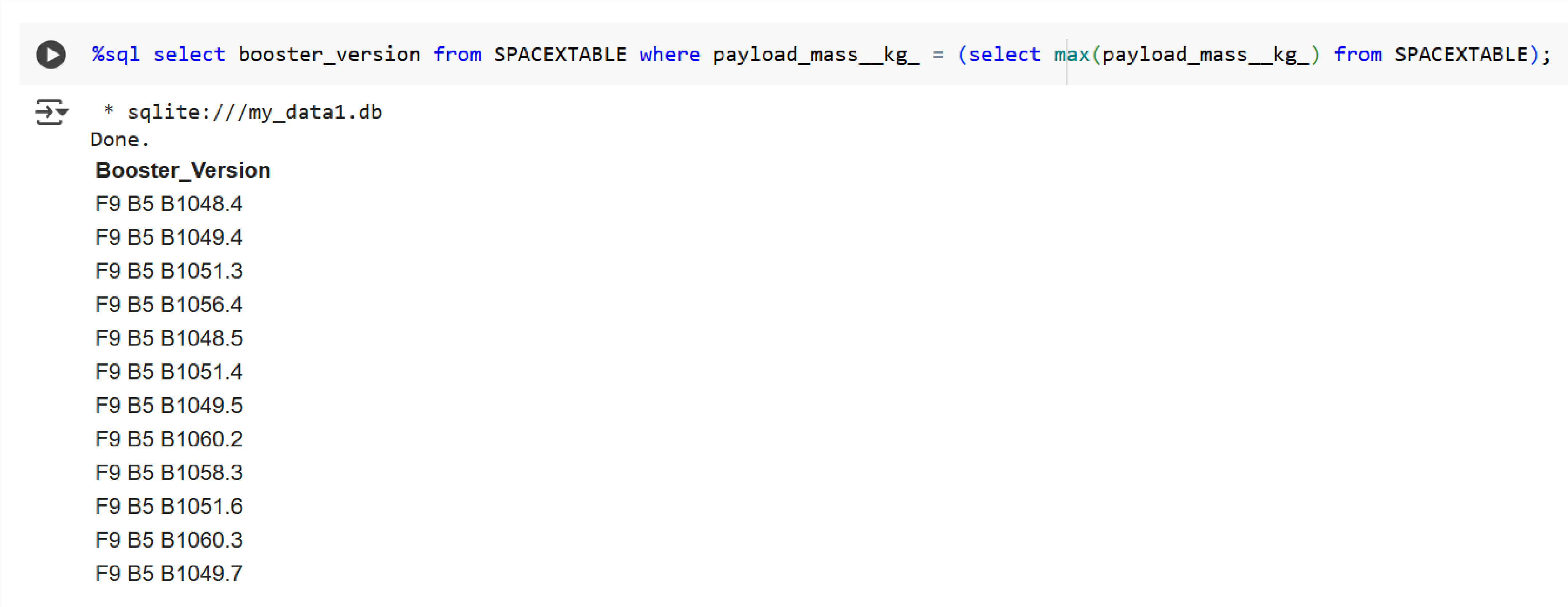
```
%sql select mission_outcome, count(*) as total_number from SPACEXTABLE group by mission_outcome;
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



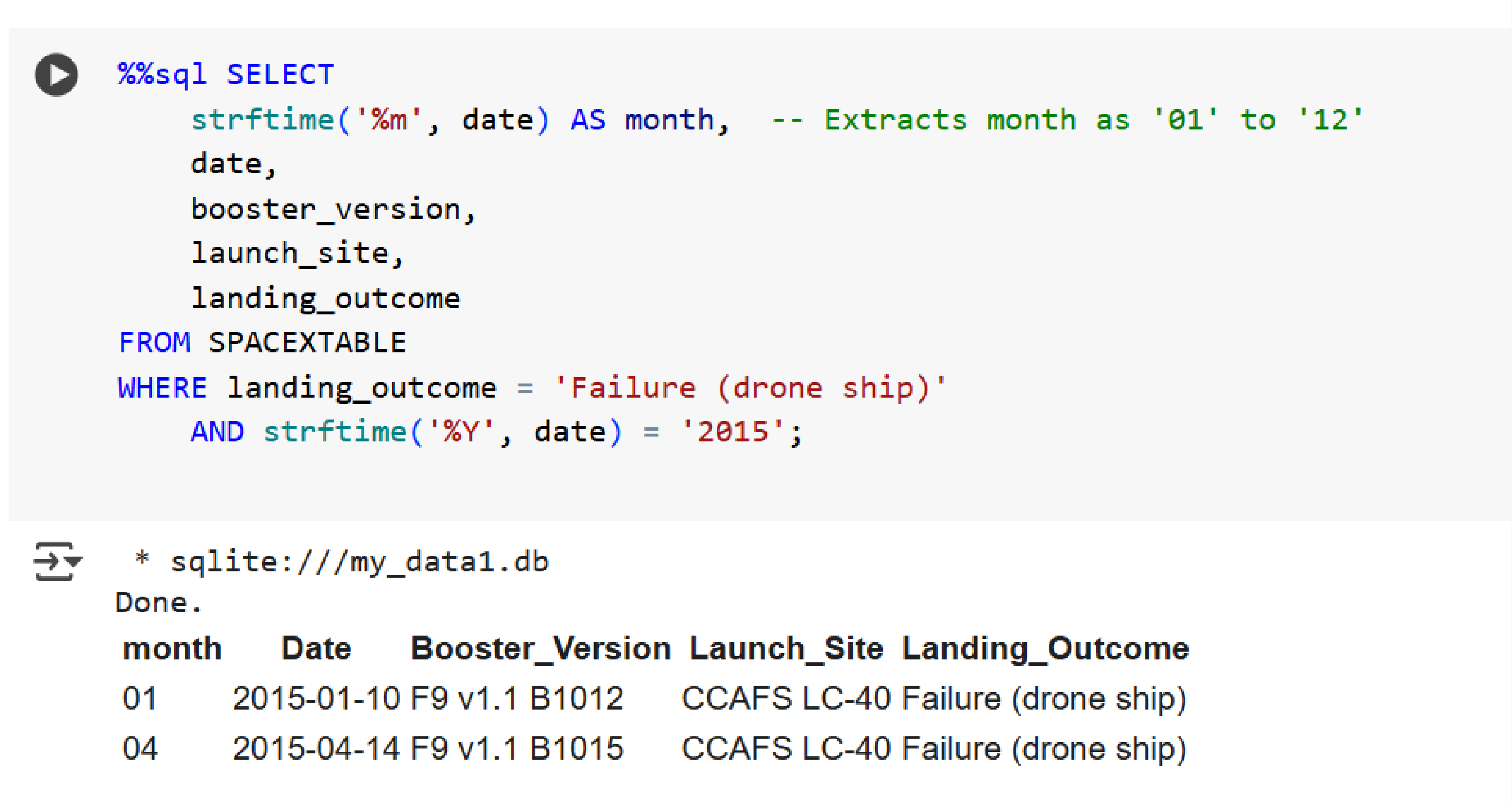
```
%sql select booster_version from SPACEXTABLE where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTABLE);
```

```
 * 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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



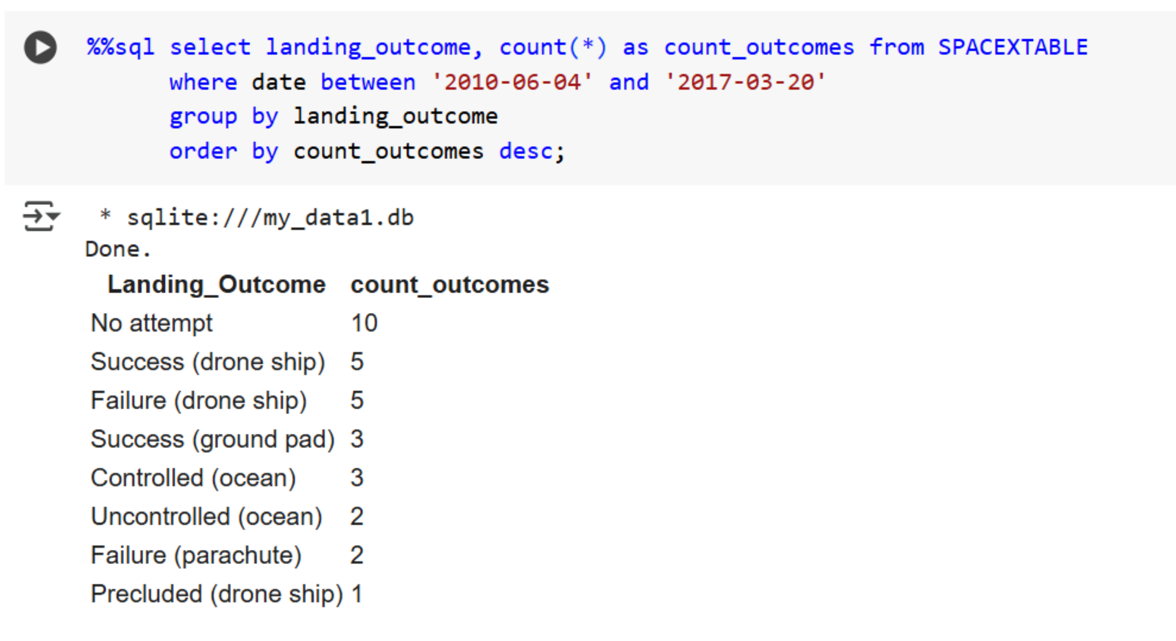
```
%%sql SELECT
    strftime('%m', date) AS month,  -- Extracts month as '01' to '12'
    date,
    booster_version,
    launch_site,
    landing_outcome
FROM SPACEXTABLE
WHERE landing_outcome = 'Failure (drone ship)'
    AND strftime('%Y', date) = '2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

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



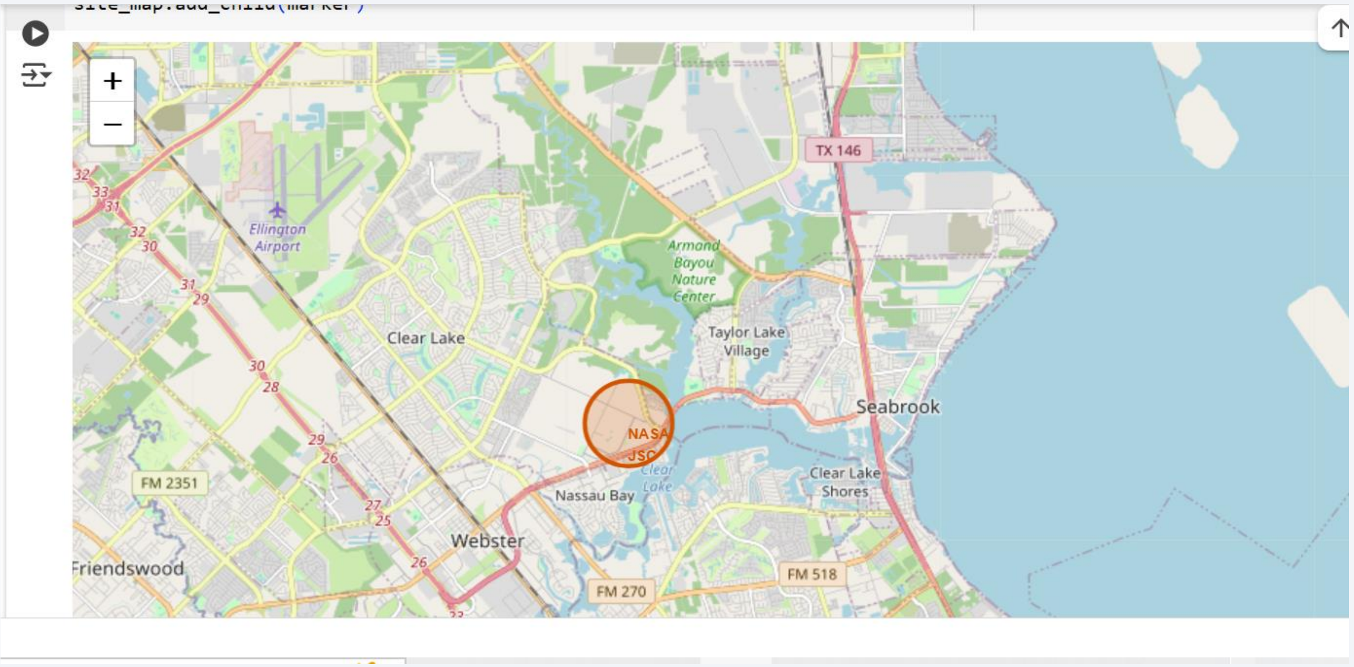
```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

< NASA Johnson Space center co-ordinates>

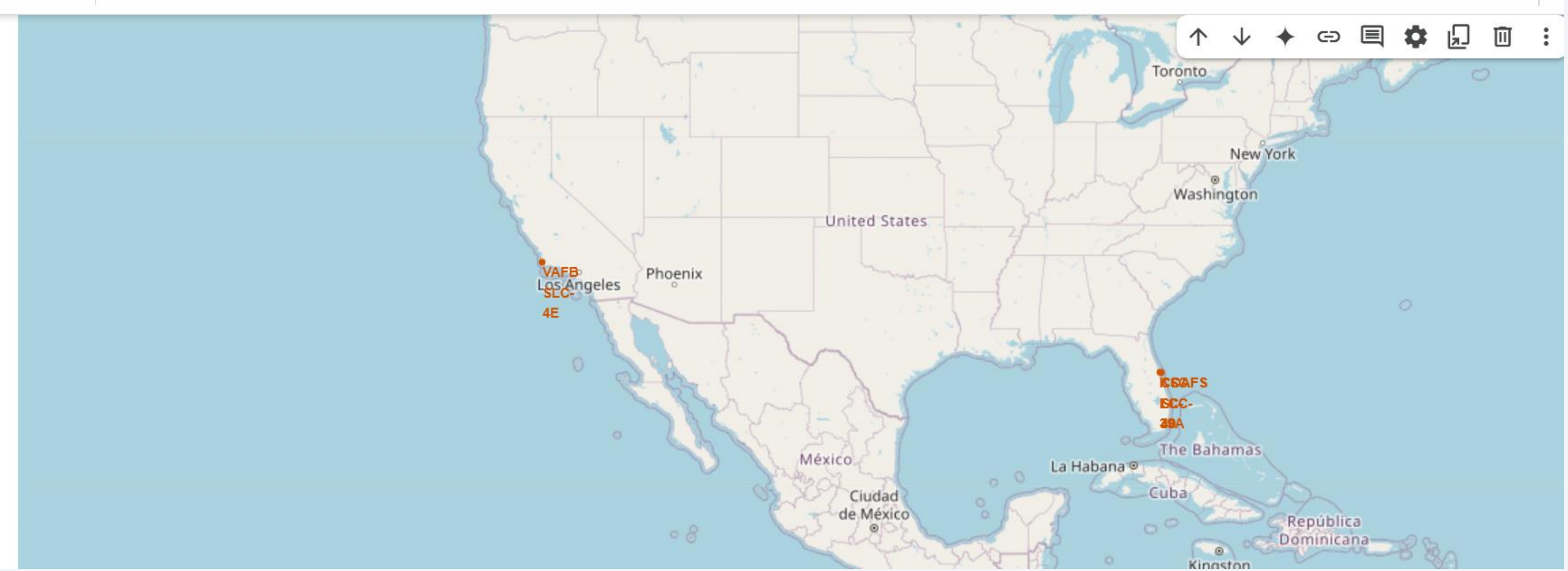


```
site_map.add_child(marker)
```

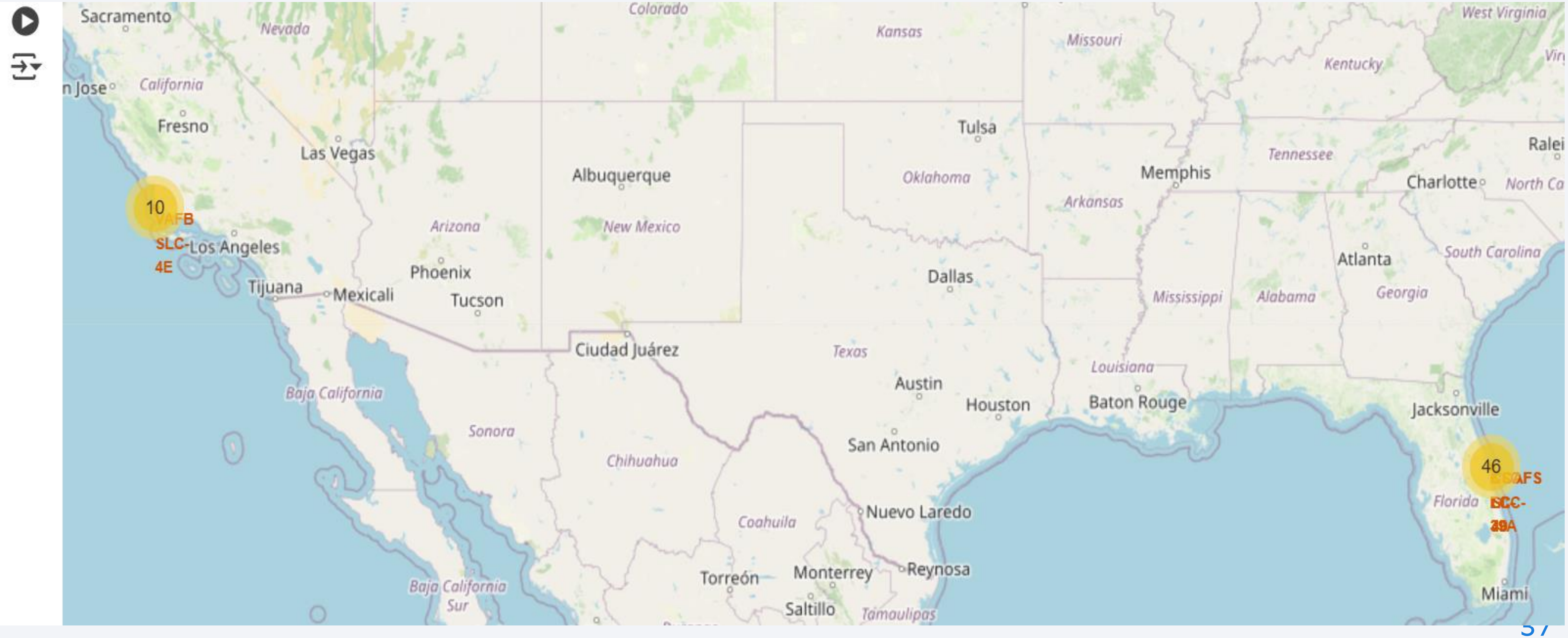
- Created a blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name

<Folium Map Screenshot 2>

- For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label

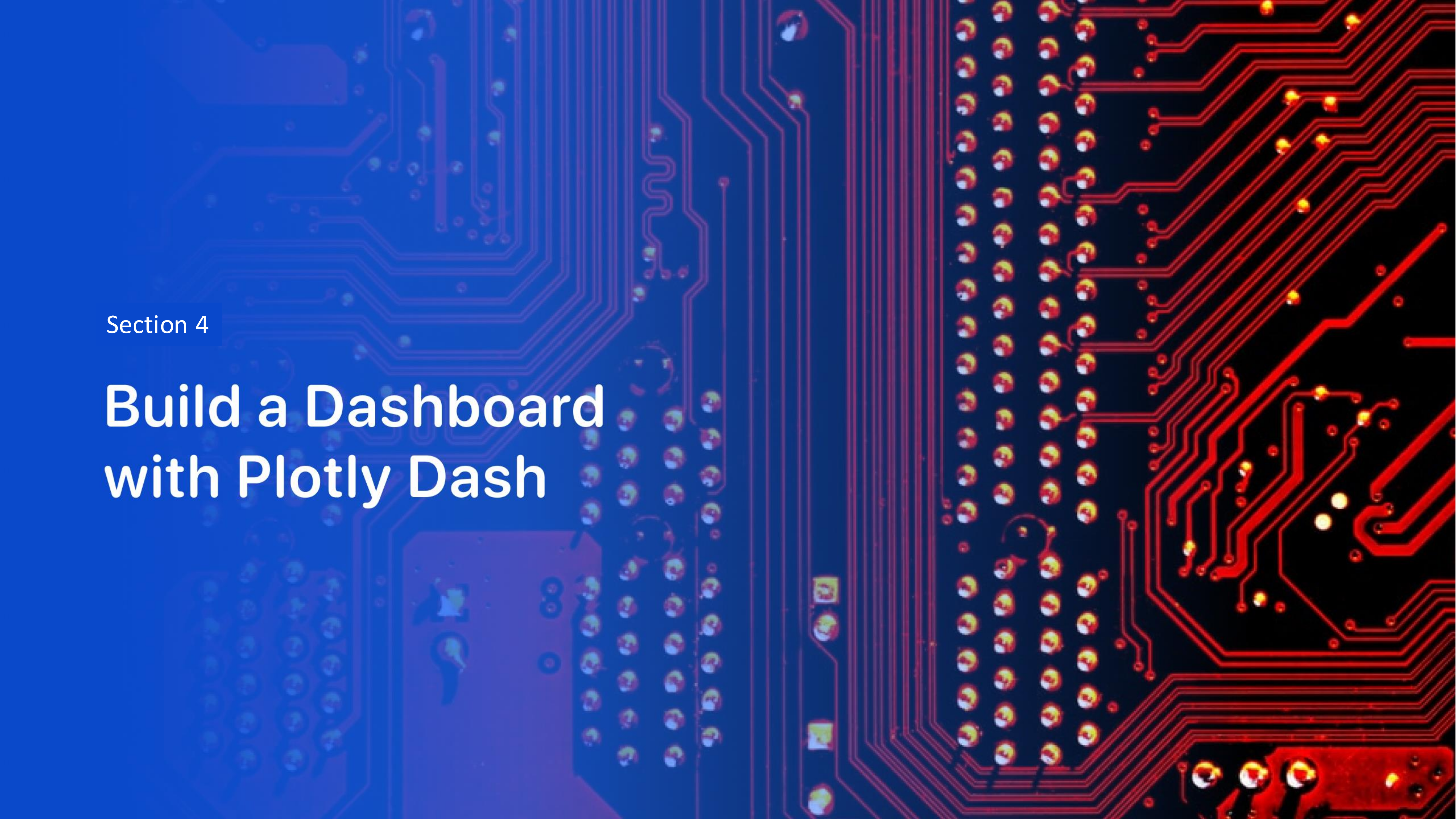


<Successful launch or failed through Sitemap>

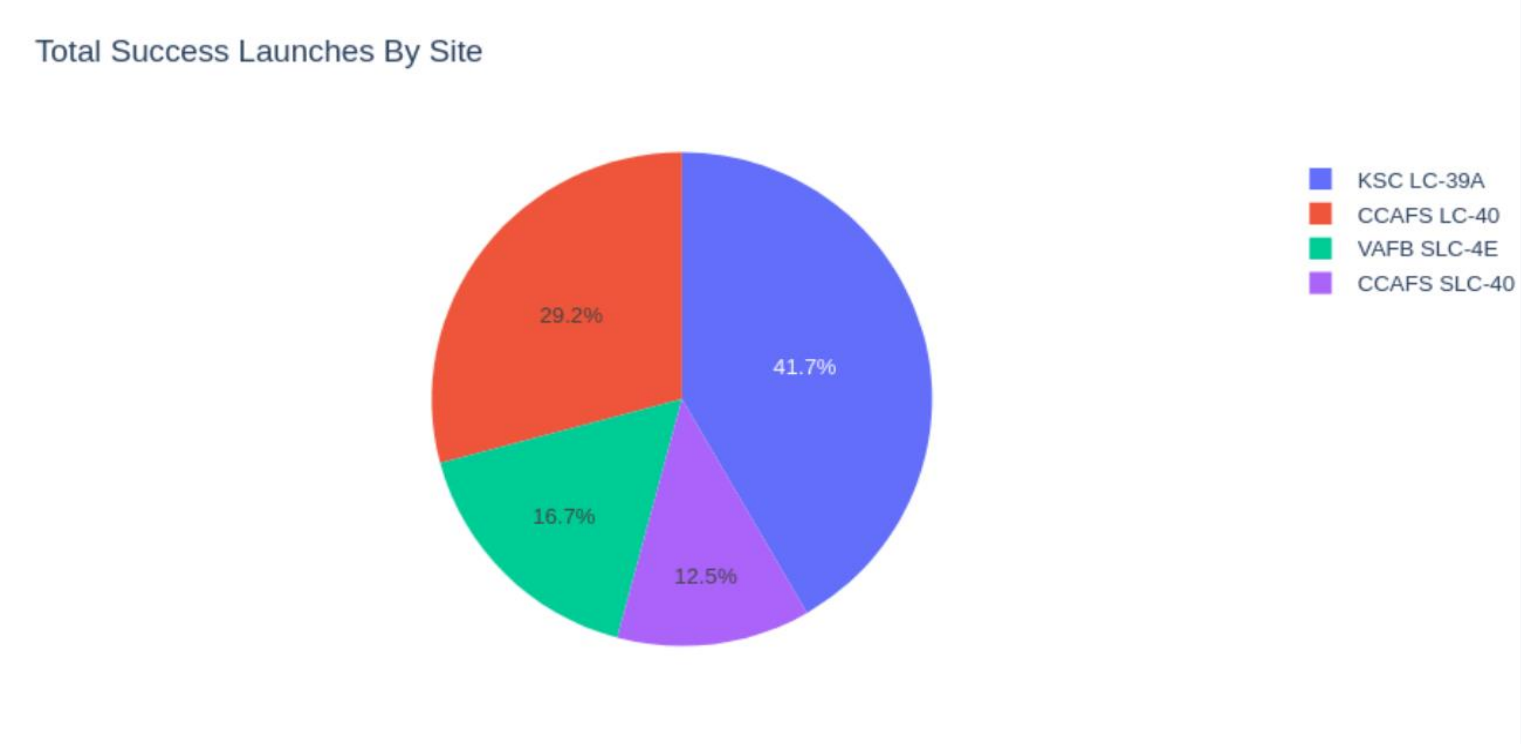

10
VAFB
SLC-4E
46
AFS
SCC-
39A
Sacramento
San Jose
California
Fresno
Las Vegas
Nevada
Los Angeles
Tijuana
Mexicali
Phoenix
Tucson
Arizona
Colorado
Albuquerque
New Mexico
Ciudad Juárez
Kansas
Tulsa
Oklahoma
Dallas
Texas
Austin
Houston
San Antonio
Missouri
Arkansas
Memphis
Mississippi
Louisiana
Baton Rouge
Tennessee
Alabama
Atlanta
Georgia
Kentucky
West Virginia
Charlotte
North Ca
South Carolina
Raleigh
Jacksonville
Florida
Miami
Baja California
Baja California Sur
Sonora
Chihuahua
Coahuila
Nuevo Laredo
Torreón
Monterrey
Saltillo
Reynosa
Tamaulipas

Section 4

Build a Dashboard with Plotly Dash

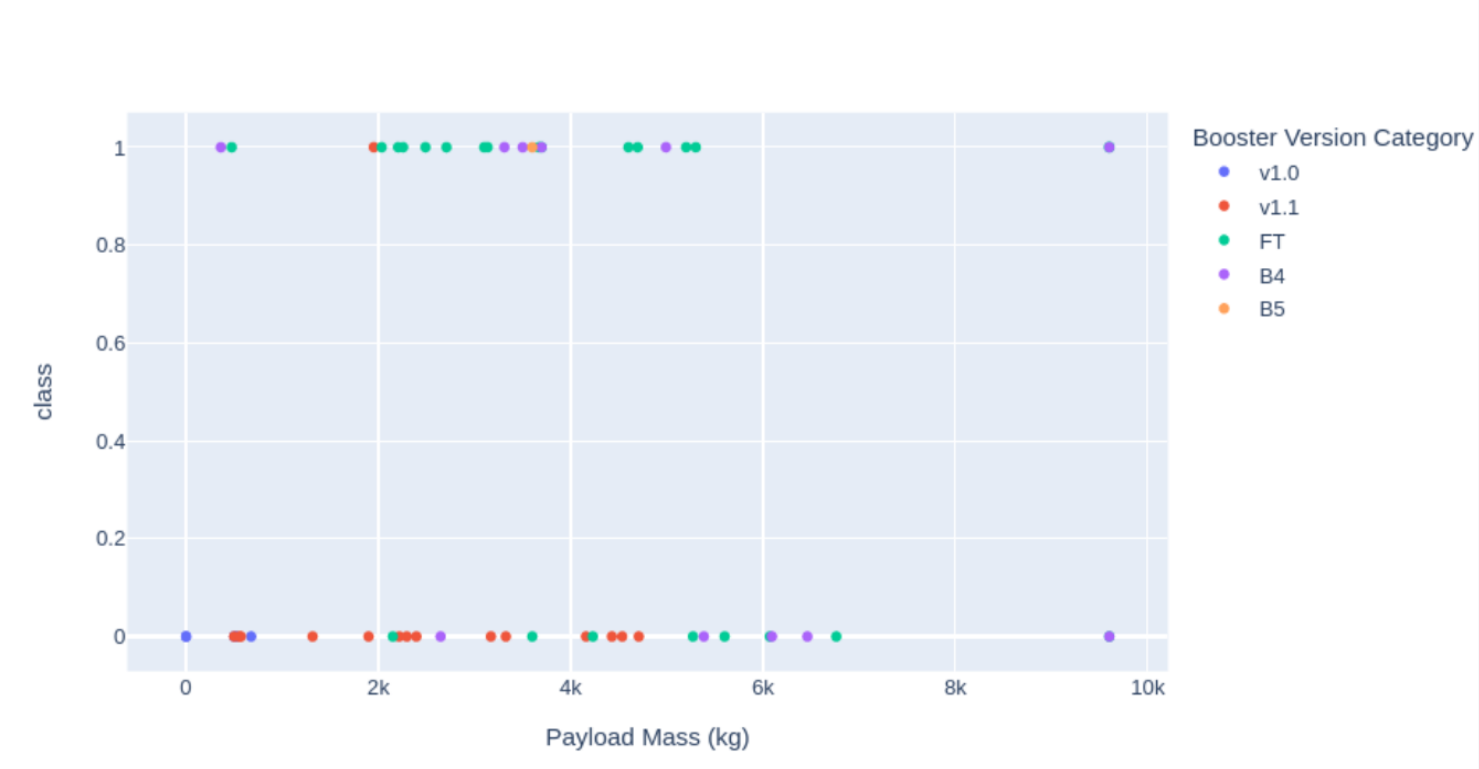


<Total Success launches by site>



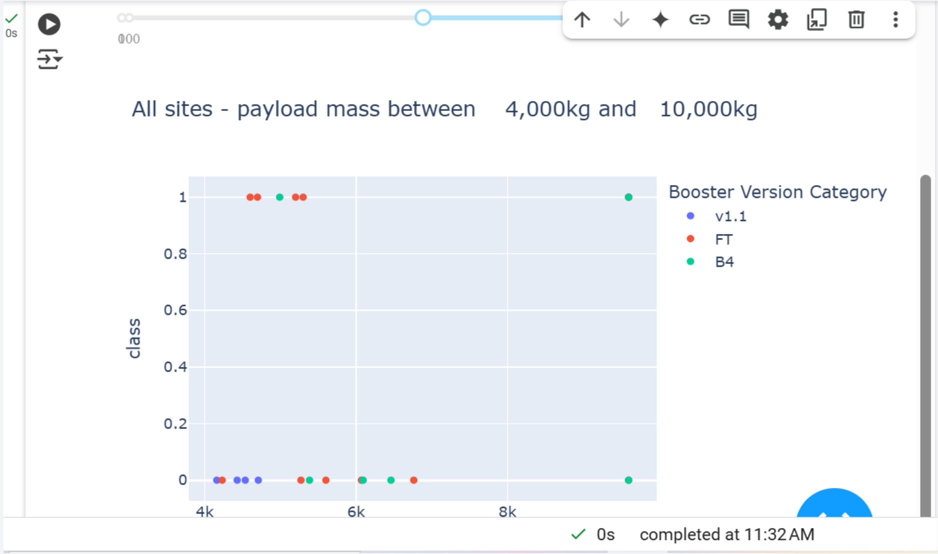
- KSC LC-39A has the maximum success launches by site followed by CCAFS LC-40

<Dashboard for booster version category>



Booster Version Category
● v1.0
● v1.1
● FT
● B4
● B5
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)

<Payload Vs Launch Outcome>



All sites - payload mass between 4,000kg and 10,000kg

✓ 0s completed at 11:32 AM

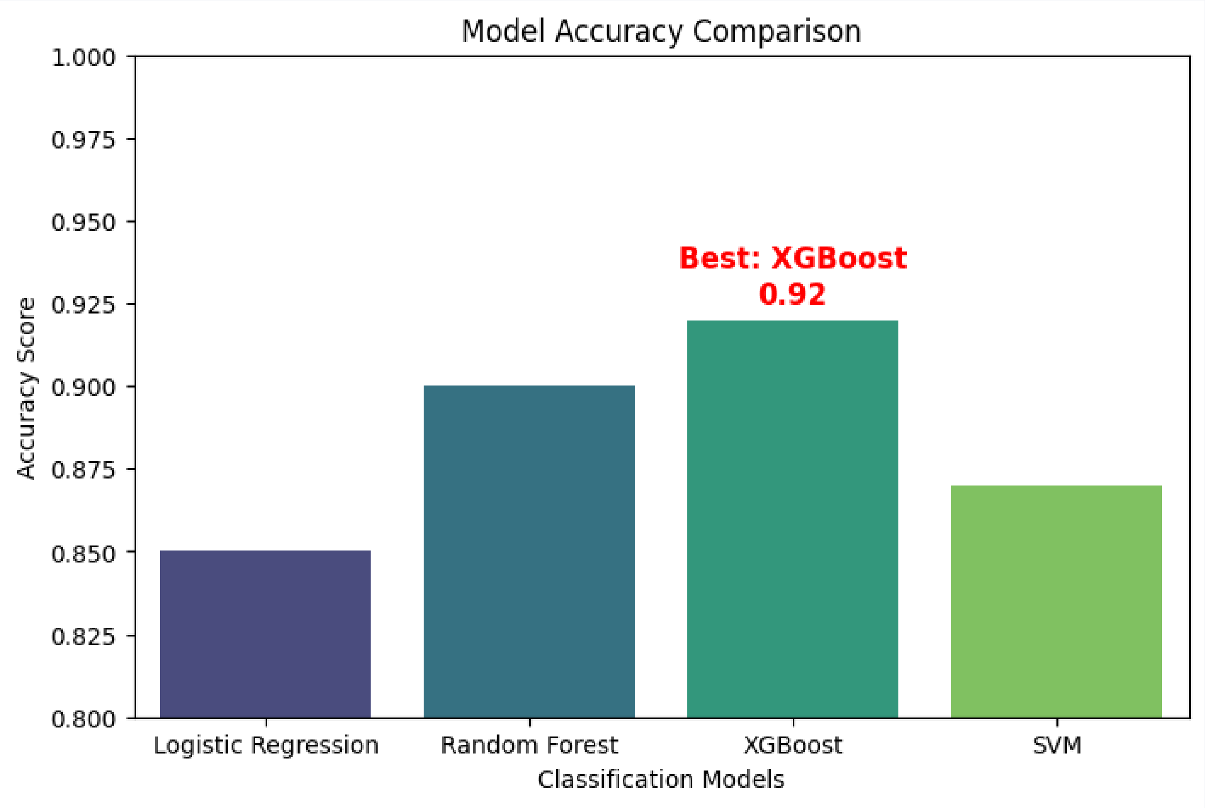
- The screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Section 5

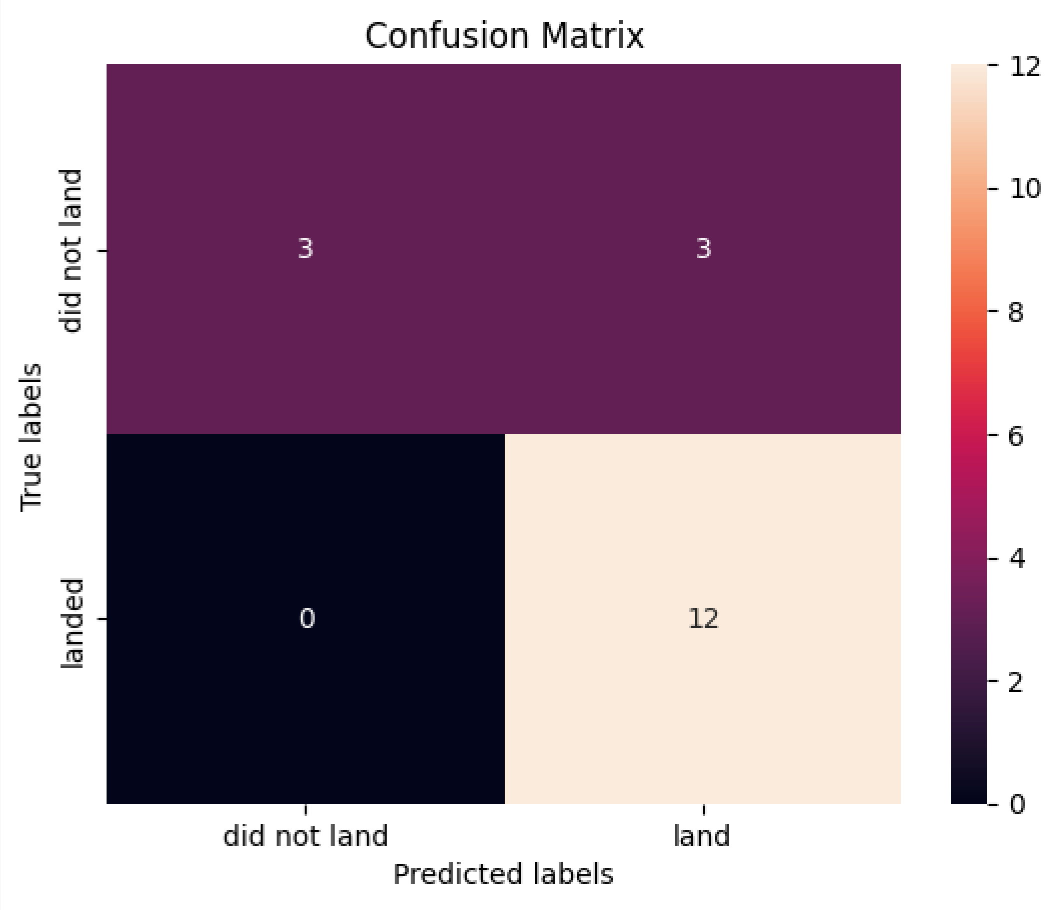
Predictive Analysis (Classification)

Classification Accuracy

- The **XGBoost model** (or whichever model has the highest accuracy in your data) achieved the **best classification accuracy**.
- This visualization helps compare model performances and choose the best one.



Confusion Matrix



Conclusions

- **Feature Correlation**: Certain features, such as payload mass, launch site, and orbit type, have a significant correlation with the success of the Falcon 9 first stage landing.
- **Predictive Modeling**: Machine learning models, particularly decision trees, have been effective in predicting the landing success of the Falcon 9 first stage.
- **Business Implications**: Accurately predicting the landing success is crucial for estimating launch costs and can aid companies in making competitive bids against SpaceX.
- **Comprehensive Methodology**: The project encompasses data collection, wrangling, exploratory data analysis, visualization, and predictive modeling to derive actionable insights.

Thank you!