IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Radzi Aziz

2 February 2025

Outline

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

Executive Summary

- This project analyzes SpaceX Falcon 9 launches to predict first-stage landing success, aiding cost estimation and mission planning. Data was collected via API and web scraping, processed using feature engineering, and explored through visualization and SQL analysis.
- Key findings include KSC LC-39A having the highest success rate (76.9%), a payload mass threshold of 6000 kg affecting success rates, and SVM achieving the highest classification accuracy (0.88). The insights support data-driven decision-making for the space industry.

Introduction

- SpaceX has revolutionized the space launch industry with its Falcon 9 rocket, notable for its reusable first stage. This reusability significantly reduces launch costs compared to traditional expendable rockets. SpaceX advertises launches for around $62 million, while competitors charge upwards of $165 million, primarily due to this difference. The ability to reuse the first stage is a key factor in SpaceX's cost advantage.
- For companies looking to compete with SpaceX or for potential SpaceX customers, accurate cost estimation is crucial. A key factor in determining the cost of a Falcon 9 launch is whether the first stage will successfully land.
- Therefore, the problem this project addresses is: Can we accurately predict the successful landing of the Falcon 9 first stage? Answering this question allows for more precise launch cost predictions, which is vital for competitive bidding and overall business planning in the space launch market.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

API, Webscrapping

Perform data wrangling

Data cleaning, Handle missing values, Data transformation, Feature engineering

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Preprocess data > Select features > Train models > Evaluate performance > Tune hyperparameters > Select best model

Data Collection

The data collection process for the SpaceX Falcon 9 launch data involved two primary method

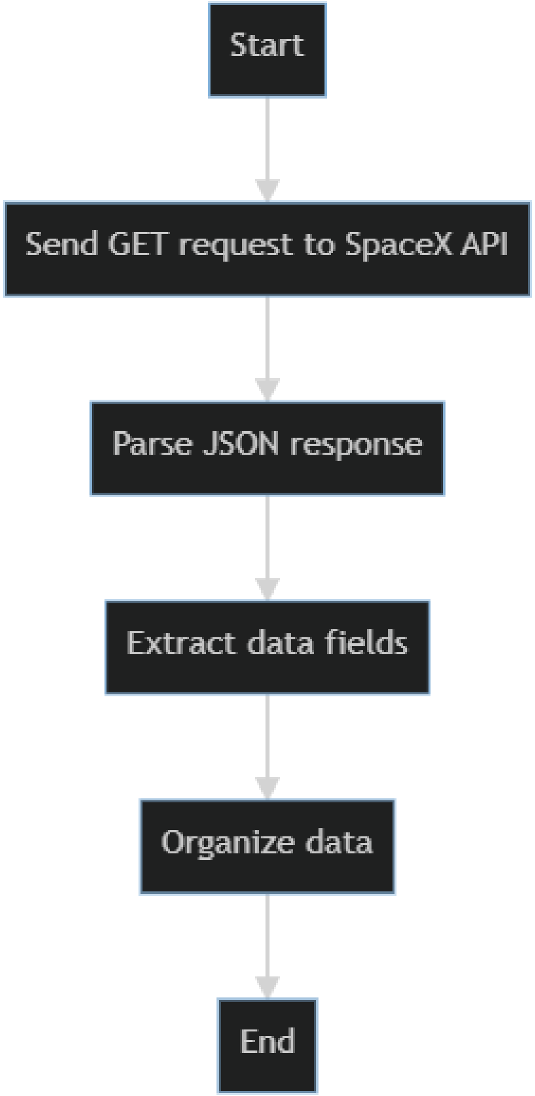
SPACEX API

WEB SCRAPING WIKIPEDIA PAGE

Data Collection – SpaceX API

- Process Flow:
  - A GET request is sent to the SpaceX API endpoint to retrieve past launch data.
  - The API response, in JSON format, is parsed.
  - Relevant data fields (e.g., flight number, payload, launch date, landing success) are extracted from the JSON.
  - The extracted data is organized into a Pandas DataFrame for further processing and analysis.

- Link: https://github.com/radziaziz/ds-spacex/blob/main/01-jupyter-labs-spacex-data-collection-api.ipynb



Start

Send GET request to SpaceX API

Parse JSON response

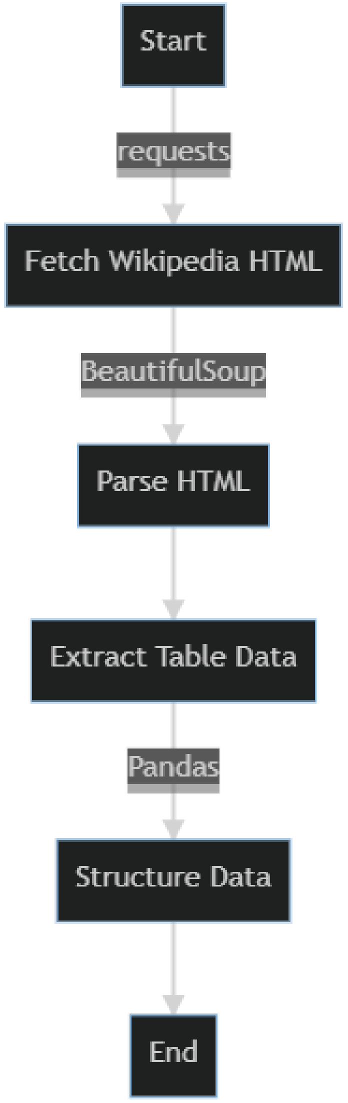
Extract data fields

Organize data

End

Data Collection - Scraping

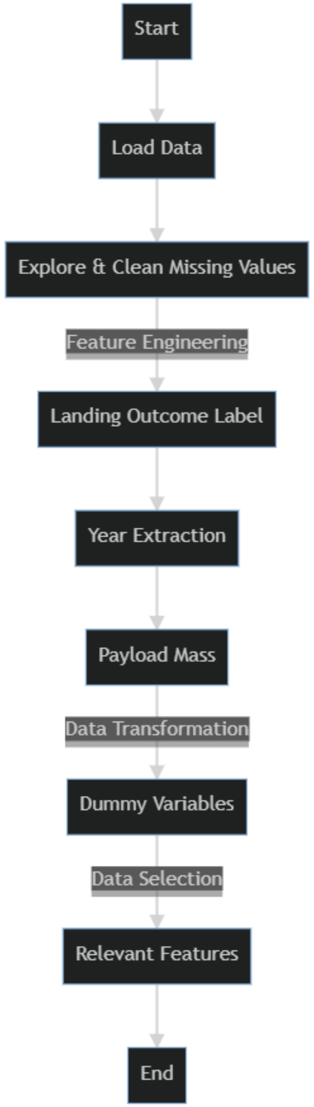
- Process Flow:
  - The requests library is used to fetch the HTML content of the Wikipedia page listing Falcon 9 launches.
  - BeautifulSoup is used to parse the HTML and identify the relevant table containing launch records.
  - The table data (launch date, payload, launch site, etc.) is extracted.
  - The extracted data is structured into a Pandas DataFrame.
- Link: https://github.com/radziaziz/ds-spacex/blob/main/02_data_collection_webscrapping.ipynb



Data Wrangling

- The process started by loading SpaceX launch data using pandas. Missing values were handled with dropna(), and feature engineering created a 'Class' label (0 = failure, 1 = success) from landing outcomes. The year was extracted from 'Date', and payload mass was converted to numeric. Categorical features like 'Orbit' and 'LaunchSite' were transformed using pd.get_dummies(). Finally, relevant features were selected, preparing the dataset for analysis and machine learning.

- Link: https://github.com/radziaziz/ds-spacex/blob/main/03_data_wrangling.ipynb



EDA with Data Visualization

Several charts were plotted to analyze SpaceX launch data. Different types were chosen to explore relationships between variables, understand distributions, and identify patterns in the SpaceX launch data.

Scatter Plot: This chart displayed the relationship between FlightNumber and PayloadMass, with the launch outcome indicated. It illustrate that as the number of flights increased, the likelihood of a successful first-stage landing also increased. Additionally, it showed that heavier payloads were less likely to result in successful landings.

Bar Chart: A bar chart was used to show the number of launches for each LaunchSite. This visualization highlighted which launch sites were most frequently used.

Pie Chart: This chart depicted the distribution of different Orbit types. It provided a clear view of the most common orbits targeted by SpaceX launches.

Histogram: A histogram was plotted to show the distribution of PayloadMass. This helped in understanding the range and frequency of payload masses in the dataset.

Heatmap: A heatmap was created to display the correlation between various numerical features in the dataset. It assisted in identifying which factors might influence the success of a launch.

Link: https://github.com/radziaziz/ds-spacex/blob/main/05_eda_panda.ipynb

EDA with SQL

Summary the SQL queries performed

1. Retrieving basic launch data with `SELECT` statements.
2. Filtering launches by specific criteria like launch success, failure, or certain launch sites using `WHERE`.
3. Aggregating data to find total launches, successes, or failures using `COUNT` and `GROUP BY`.
4. Joining tables to combine information, such as merging launch details with payload data.
5. Calculating statistics like average payload mass with `AVG`.
6. Ordering results to find top-performing or most frequently used launch sites using `ORDER BY`.

Link: https://github.com/radziaziz/ds-spacex/blob/main/04_eda_sql.ipynb

Build an Interactive Map with Folium

The following Folium map objects were added to visualize SpaceX launch activities, showing key sites, their proximities, and relationships to successful landings.

Markers: Plotted at SpaceX launch sites to indicate their locations.

Circles: Used to highlight specific launch areas and their influence.

Lines/Polylines: Represented trajectories or distances between launch sites and landing zones.

Link: https://github.com/radziaziz/ds-spacex/blob/main/06_viz_folium.ipynb

Build a Dashboard with Plotly Dash

Pie Chart - Displays total successful launches across all sites. If a specific site is selected, it shows the success vs. failure rate for that site.

Scatter Plot - Shows the relationship between payload mass and launch success, color-coded by booster version.

Dropdown Menu - Allows filtering data by launch site.

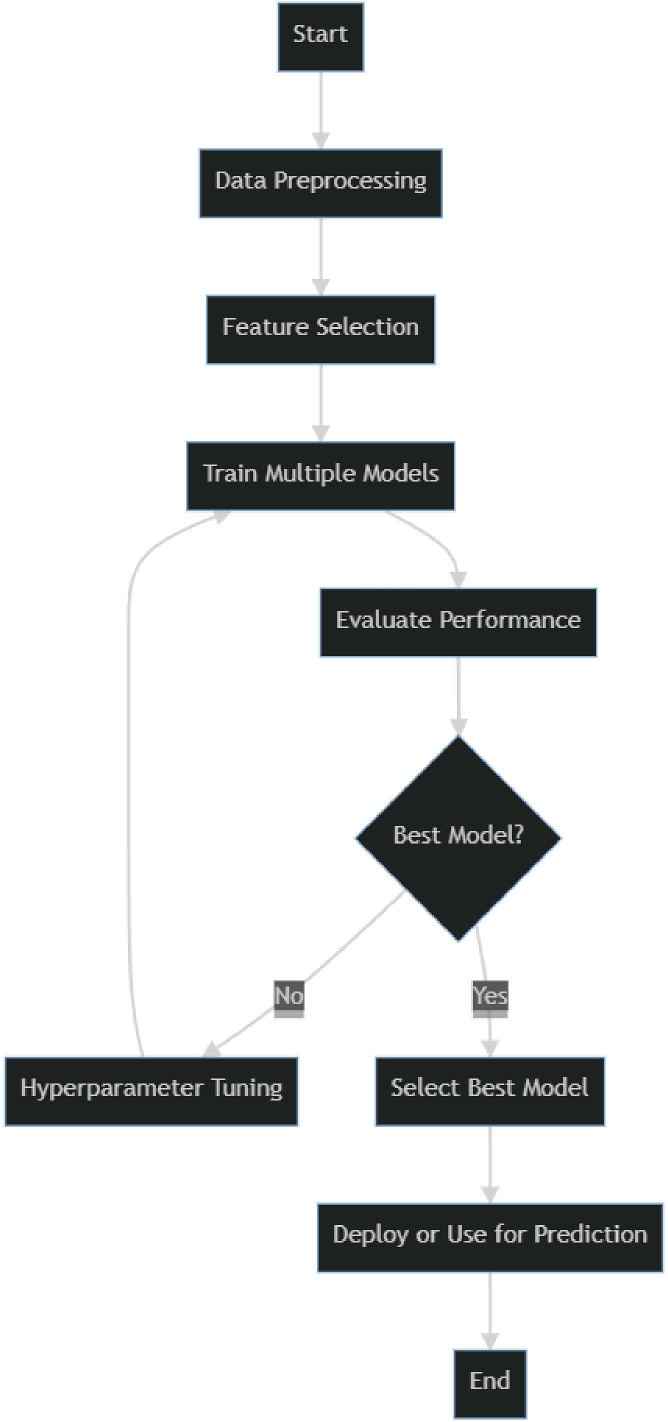
Range Slider - Adjusts the payload range for filtering data in the scatter plot.These interactions help visualize launch success rates and payload impact dynamically.

Link: https://github.com/radziaziz/ds-spacex/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)

The process started with data preprocessing to clean and prepare the dataset, followed by selecting relevant features. Multiple classification models, including Logistic Regression, Decision Tree, SVM, and KNN, were trained and evaluated based on accuracy, precision, recall, and F1-score. Hyperparameter tuning using GridSearchCV was applied to improve model performance. Finally, the best-performing model was selected based on evaluation metrics to optimize prediction accuracy for SpaceX launch success.

Link: https://github.com/radziaziz/ds-spacex/blob/main/08_predictive.ipynb



Results

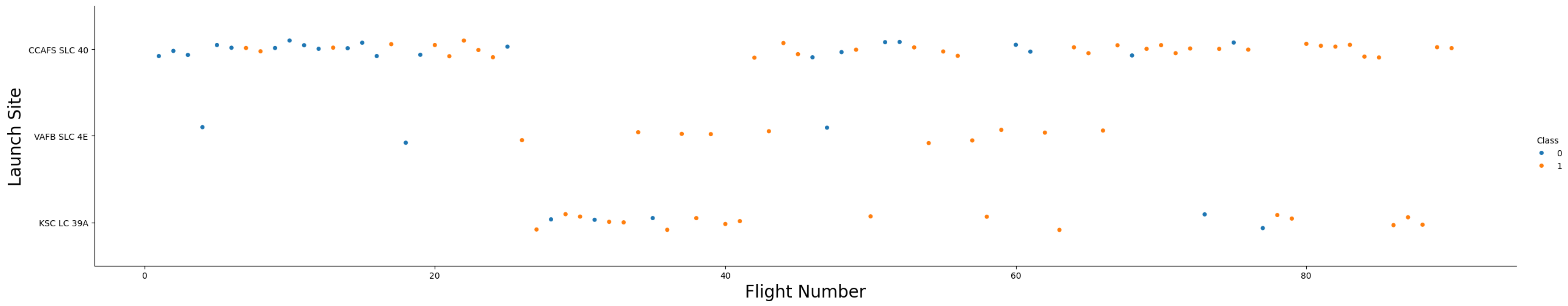
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

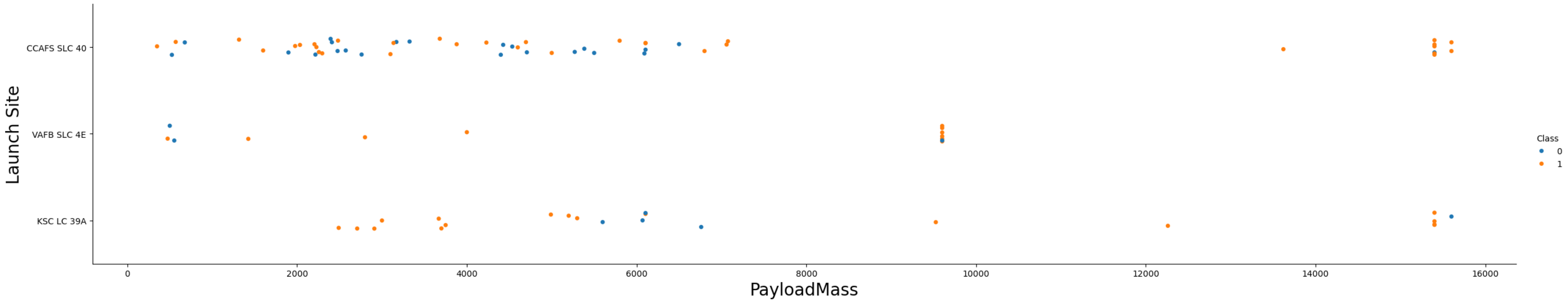
Flight Number vs. Launch Site

- Distribution of SpaceX launches by Flight Number and Launch Site.
- As Flight Numbers increase, there's a noticeable shift towards KSC LC 39A. This implies that KSC became the primary launch site for later missions



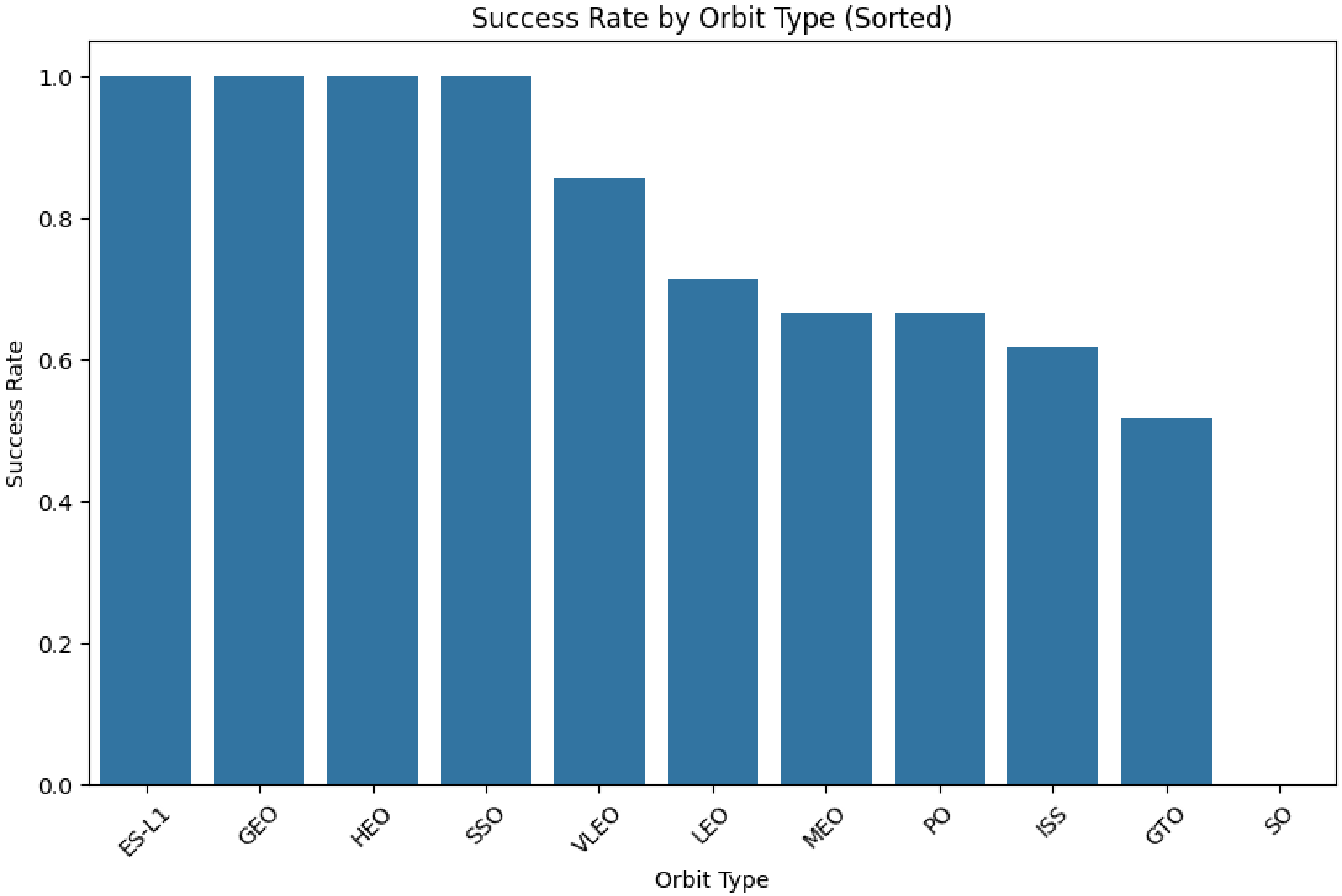
Payload vs. Launch Site

- This plot provides valuable insights into the relationship between payload mass and launch site in SpaceX missions. The observed patterns suggest potential correlations with landing success and highlight the distinct operational profiles of the different launch facilities. Further analysis, combining this visualization with statistical methods and external data, can lead to a deeper understanding of the factors influencing launch site selection and mission outcomes.



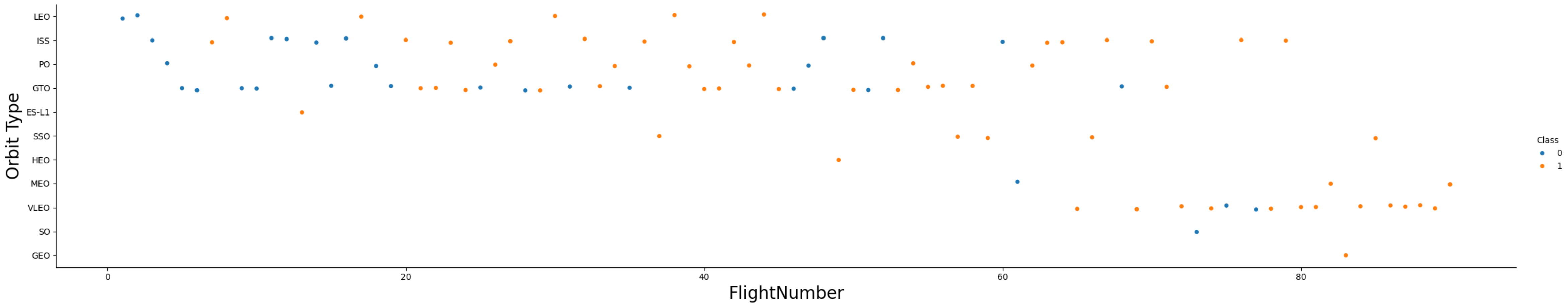
Success Rate vs. Orbit Type

- The type of orbit targeted can influence the likelihood of a successful launch and mission.
- The high success rates across most orbits demonstrate the advancement and reliability of SpaceX's launch technology and operational procedures.
- The slight variations in success rates for certain orbits (GTO, ISS, MEO) likely reflect the unique technical challenges associated with achieving and operating in those specific orbital environments.



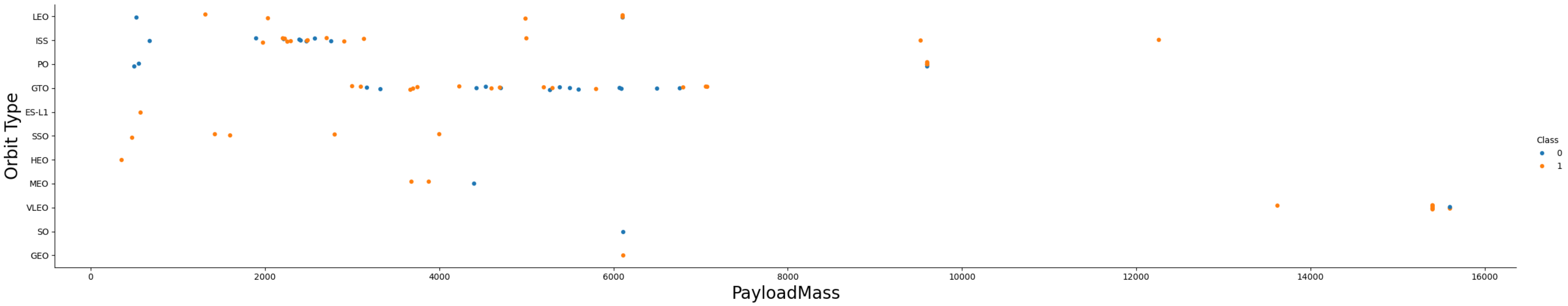
Flight Number vs. Orbit Type

- The lower flight numbers (earlier missions) seem to be concentrated in certain orbit types. For example, we might observe a cluster of early launches in LEO (Low Earth Orbit). This could suggest initial focus and development in achieving these specific orbits.
- As flight numbers increase, we see a diversification of orbit types. Later missions explore a wider range of orbits, indicating expanding capabilities and mission objectives.
- Some orbits appear to be more frequently visited than others. There might be clusters of launches in certain orbits, suggesting their importance for SpaceX's overall mission architecture.



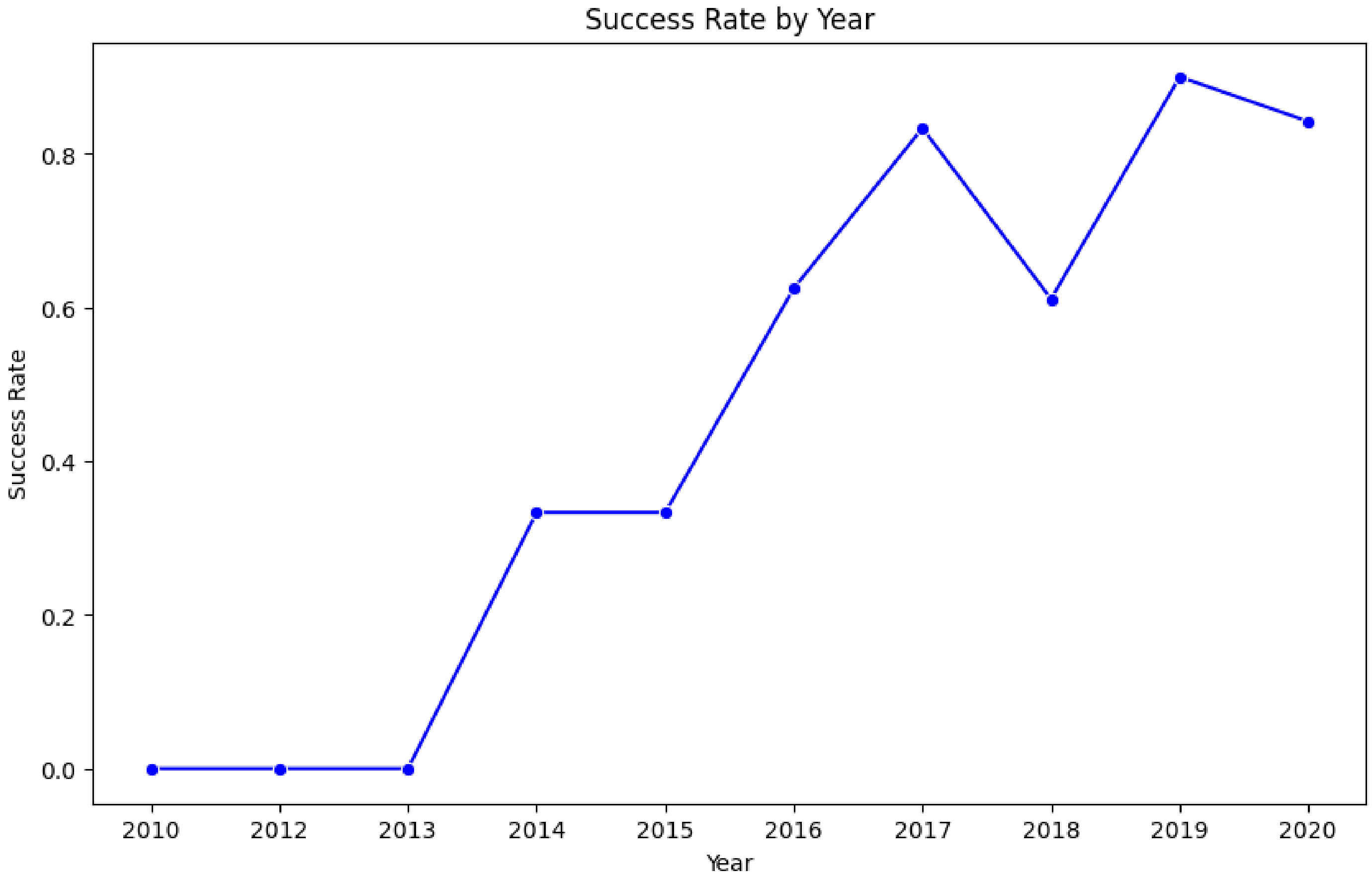
Payload vs. Orbit Type

- LEO appears to accommodate a wide range of payload masses, including both lighter and heavier payloads.
- GTO launches seem to favor heavier payloads compared to some other orbits.
- Other orbits might show more specific or limited ranges of payload masses.



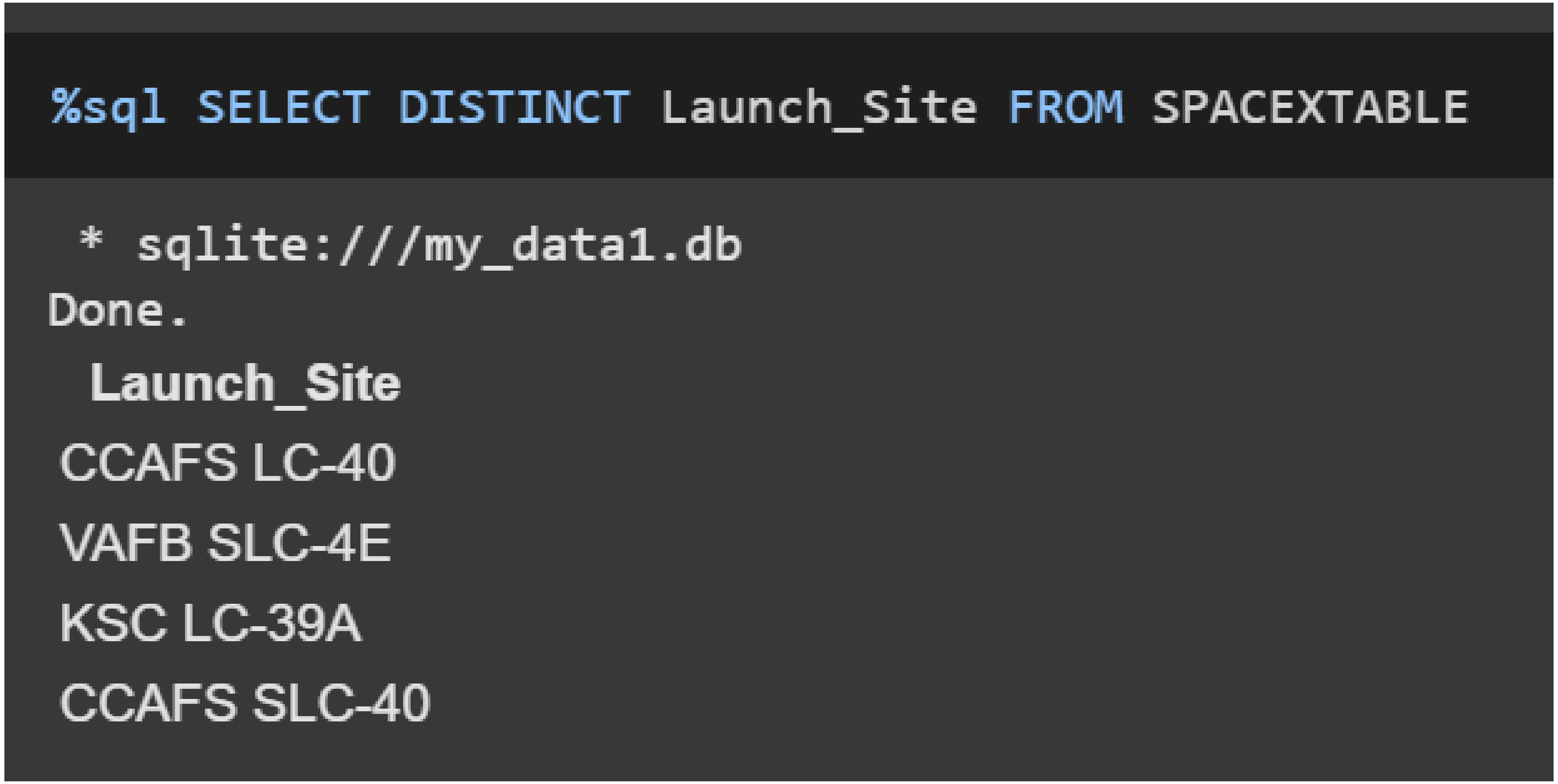
Launch Success Yearly Trend

- A sharp increase in the success rate is observed around 2014. This indicates a potential breakthrough in technology or operational procedures, marking a turning point in SpaceX's launch reliability.
- From 2014 onwards, the success rate generally trends upwards, with some fluctuations. This suggests continuous improvements and refinements in the launch program, leading to greater consistency and reliability.
- The success rate remains high and relatively stable in the later years, hovering around 0.8 and above. This demonstrates a mature and well-established launch program capable of consistent performance.



All Launch Site Names

This query used to extract unique name of launch sites from Space X dataset

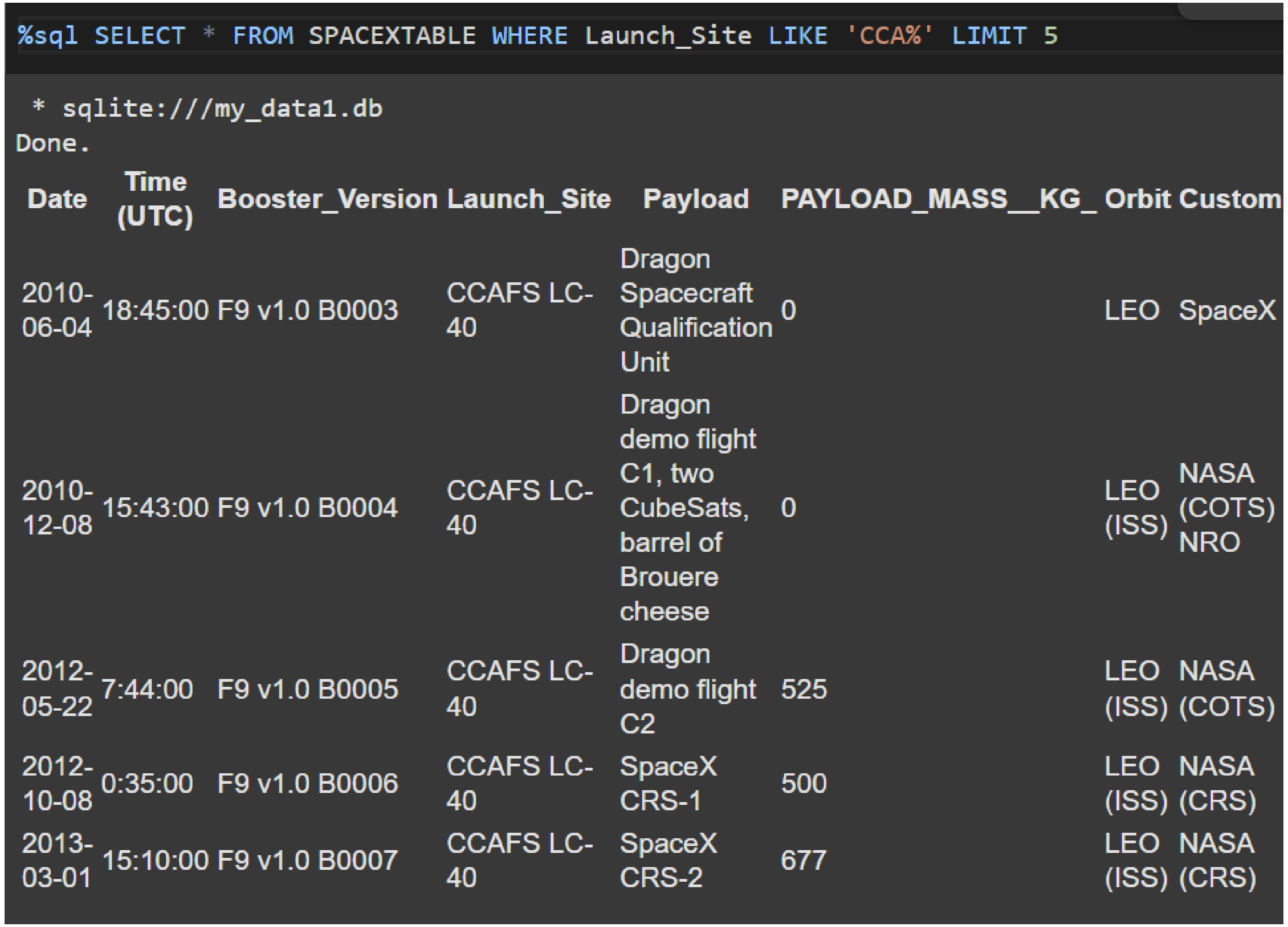


```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

```
 * 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'

- This query returns 5 records which launch site names begins with ‘CCA’
- % symbol use as wildcard



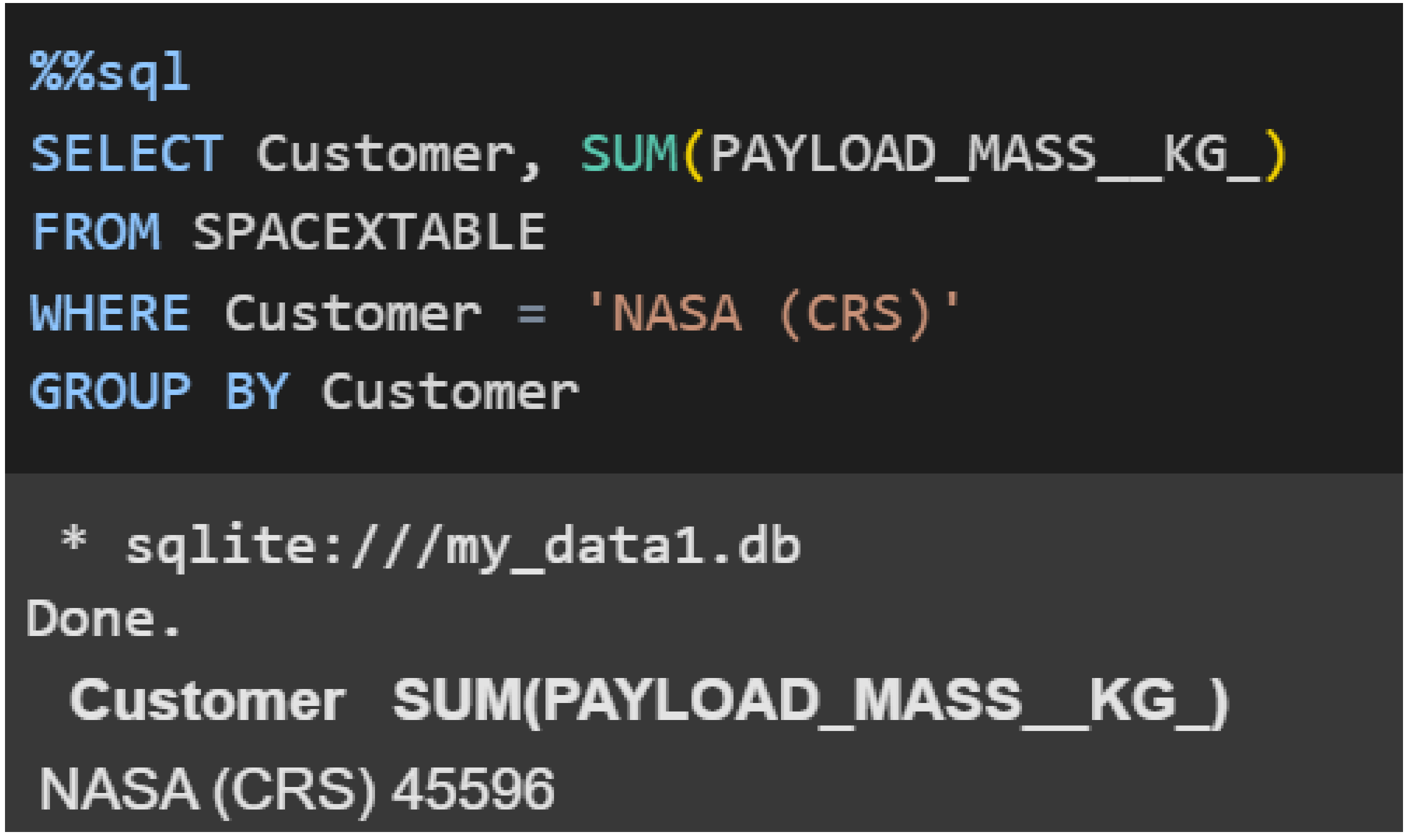
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Custom
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

- This query used to calculate the total payload carried by boosters from NASA
- We can se the total is 45,596 kg

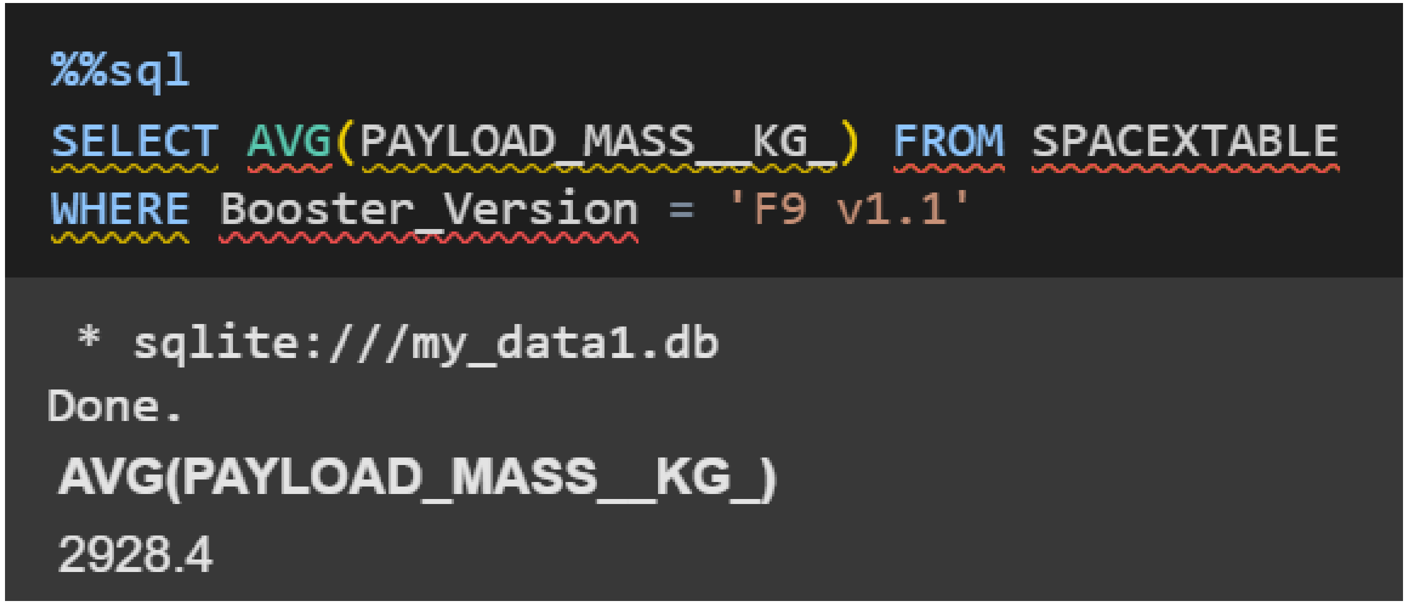


```
%%sql
SELECT Customer, SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)'
GROUP BY Customer
```

```
 * sqlite:///my_data1.db
Done.
 Customer   SUM(PAYLOAD_MASS__KG_)
NASA (CRS) 45596
```

Average Payload Mass by F9 v1.1

- This query used to calculate the average payload mass carried by booster version F9 v1.1
- The average load is 2,928.4 kg

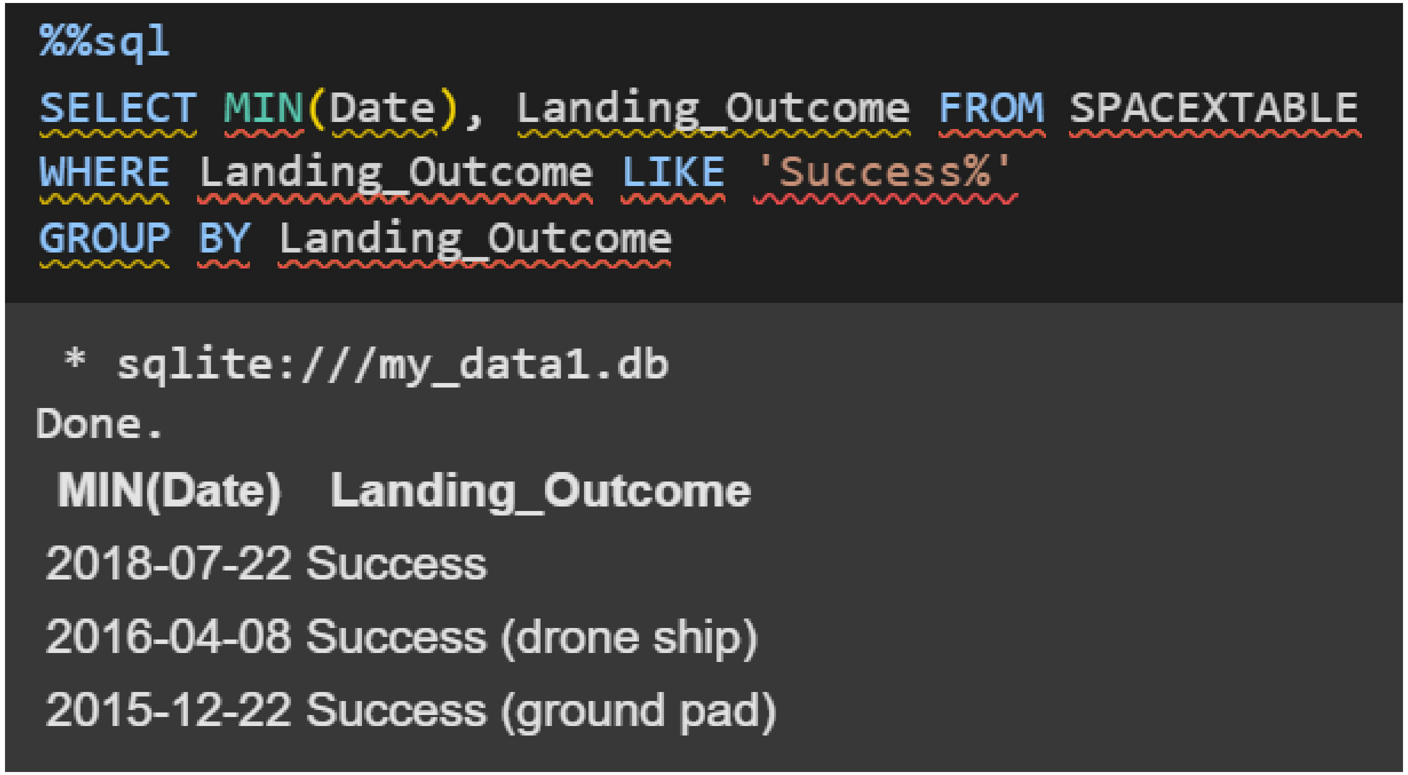


```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
AVG(PAYLOAD_MASS__KG_)
2928.4
```

First Successful Ground Landing Date

- From this query, we can see there are 3 types of success outcome.
- For ground landing, the first successful date is 2015-12-22

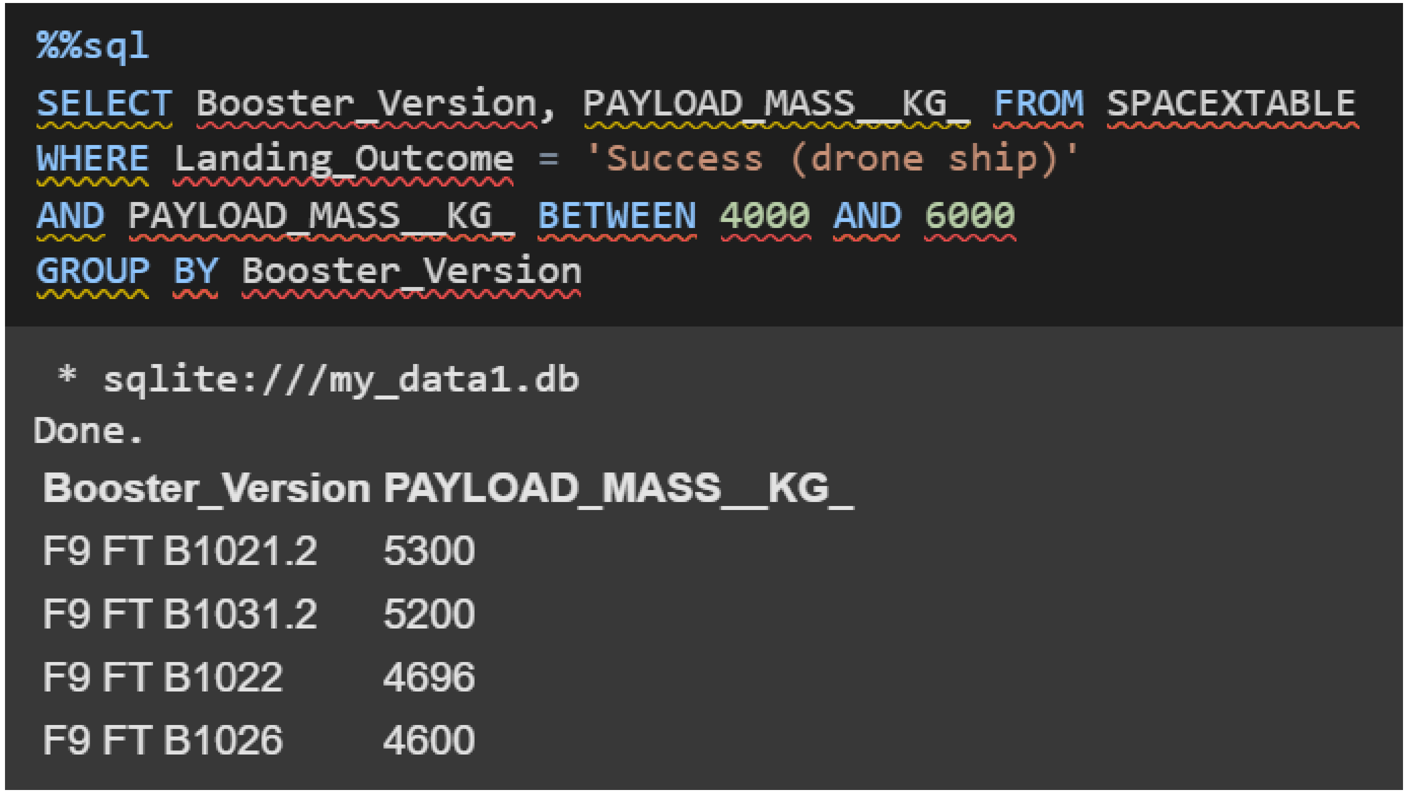


```
%%sql
SELECT MIN(Date), Landing_Outcome FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success%'
GROUP BY Landing_Outcome
```

```
 * sqlite:///my_data1.db
Done.
 MIN(Date)   Landing_Outcome
 2018-07-22 Success
 2016-04-08 Success (drone ship)
 2015-12-22 Success (ground pad)
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- This SQL query retrieves the booster version and payload mass for successful drone ship landings where the payload mass is between 4000 and 6000 kg.
- The results are grouped by booster version, showing the specific payload masses associated with each distinct booster version that met the criteria



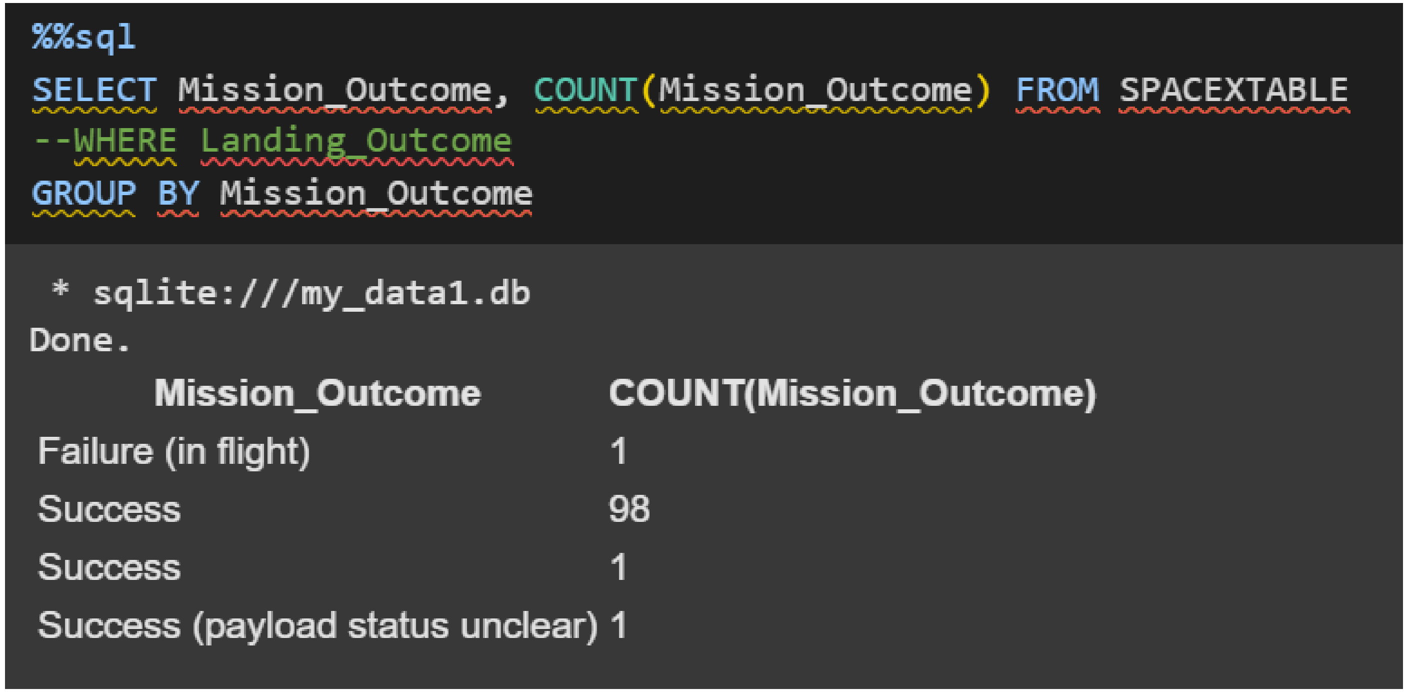
```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
GROUP BY Booster_Version
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1021.2	5300
F9 FT B1031.2	5200
F9 FT B1022	4696
F9 FT B1026	4600

Total Number of Successful and Failure Mission Outcomes

- This SQL query counts the occurrences of each distinct mission outcome in the "SPACEXTABLE". The output shows the number of missions for each outcome: 1 failure in flight, 98 successes, 1 success, and 1 success with unclear payload status
- The ratio of mission outcome for success vs failure is 100:1

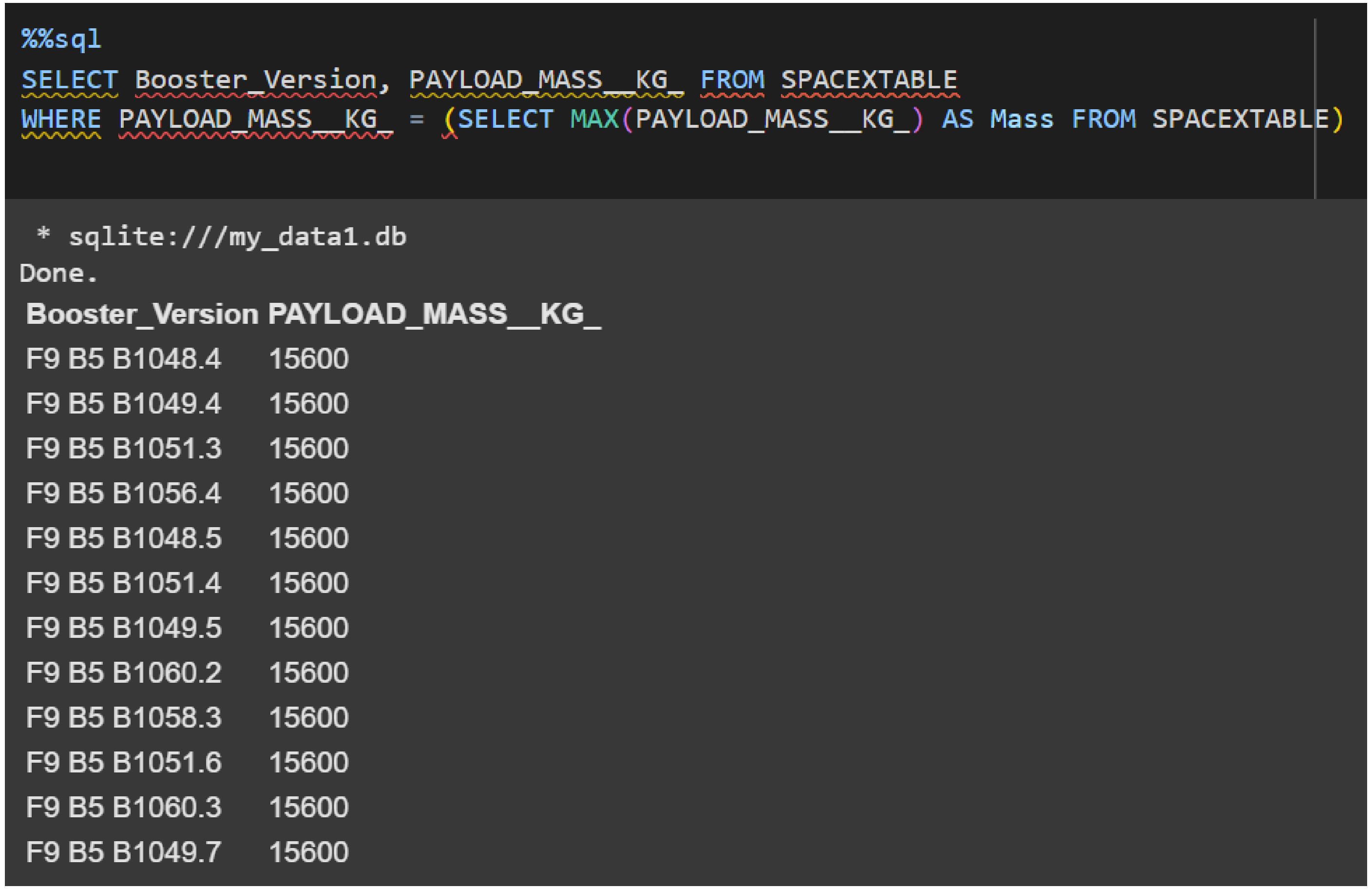


```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE
--WHERE Landing_Outcome
GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
         Mission_Outcome           COUNT(Mission_Outcome)
Failure (in flight)                1
Success                            98
Success                            1
Success (payload status unclear)   1
```

Boosters Carried Maximum Payload

- This SQL query retrieves the booster version and payload mass for the launches with the heaviest payload. It first determines the maximum payload mass using a subquery (`SELECT MAX(PAYLOAD_MASS_KG)`) and then selects the booster versions and their corresponding payload masses where the payload mass equals this maximum value. The output lists 12 different booster versions that carried the heaviest payload, which was 15600 kg.

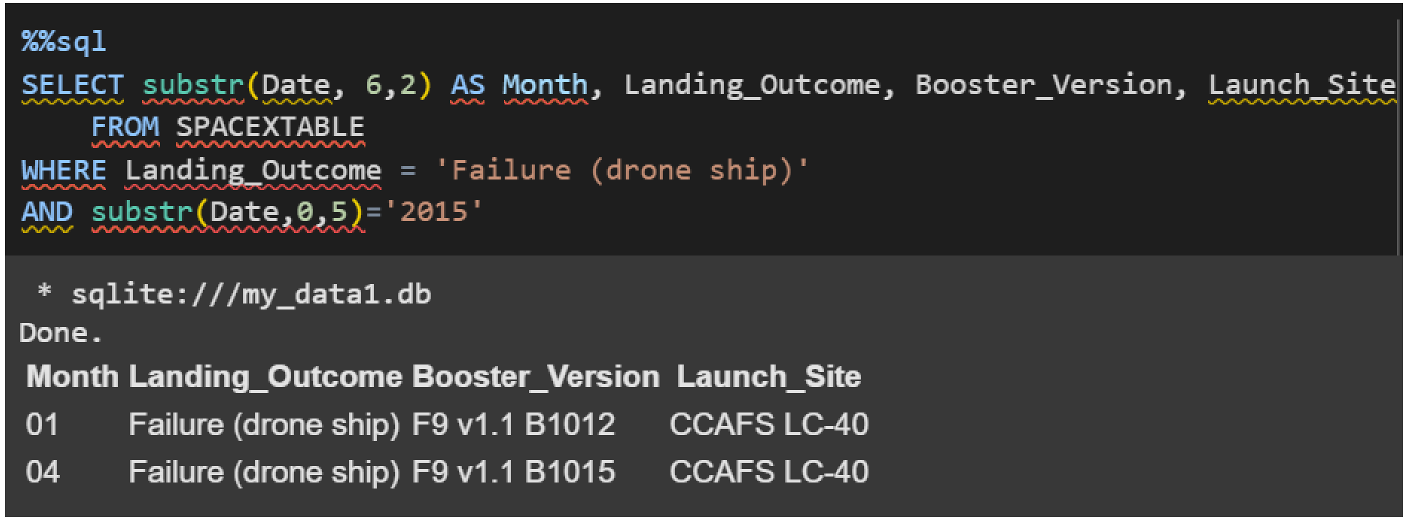


```
%%sql
SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) AS Mass FROM SPACEXTABLE)
```

```
 * 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

- This SQL query list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The output shows two such failures, occurring in January (01) and April (04) of 2015, with their respective booster versions and launch sites.

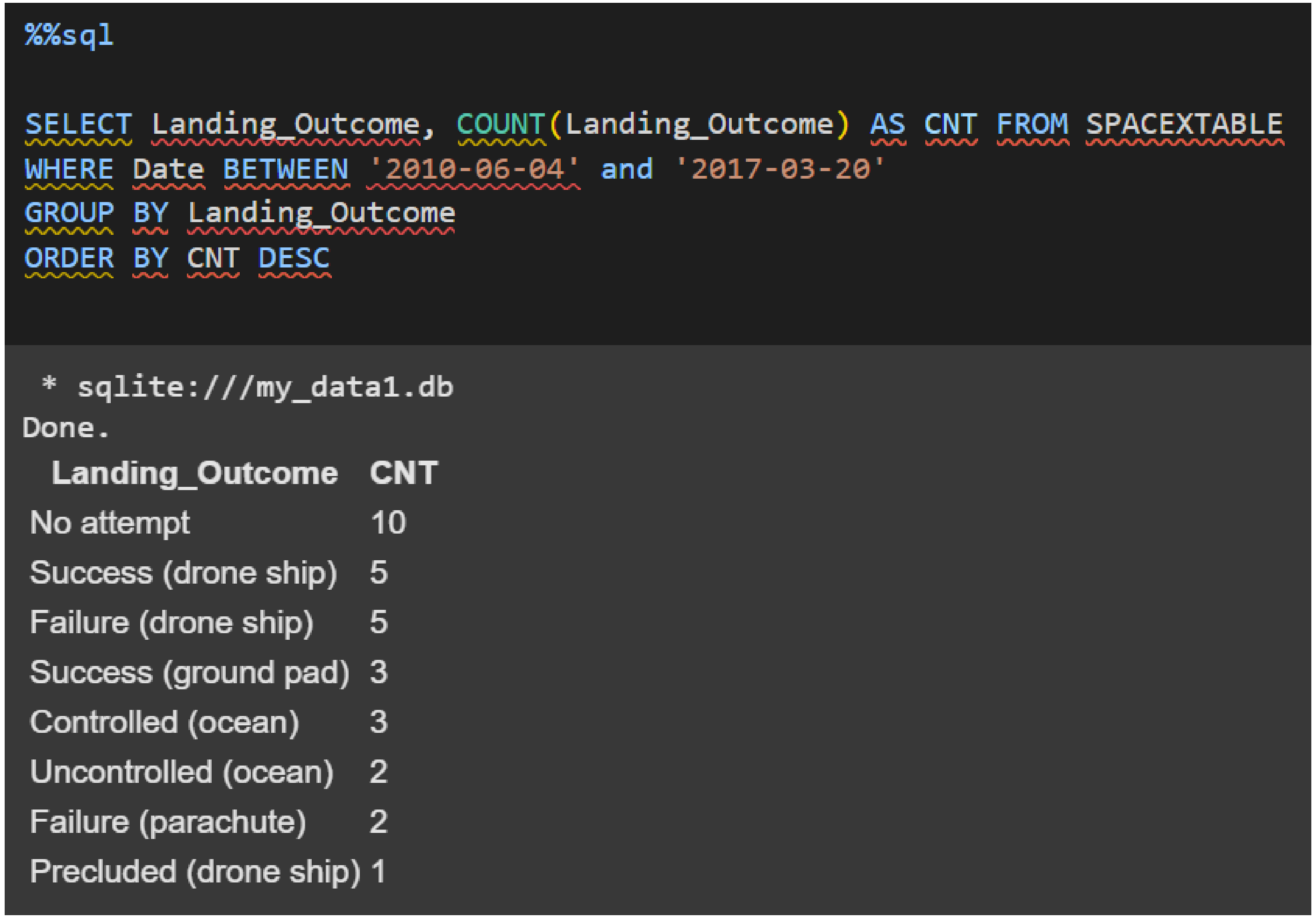


```
%%sql
SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site
    FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
AND substr(Date,0,5)='2015'
```

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

- This SQL query counts the occurrences of each distinct landing outcome for SpaceX missions between June 4, 2010, and March 20, 2017. The results are ordered by count in descending order, showing the most frequent outcomes first. The output reveals "No attempt" as the most common outcome (10 times), followed by various success and failure types.



```
%%sql

SELECT Landing_Outcome, COUNT(Landing_Outcome) AS CNT FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' and '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY CNT DESC
```

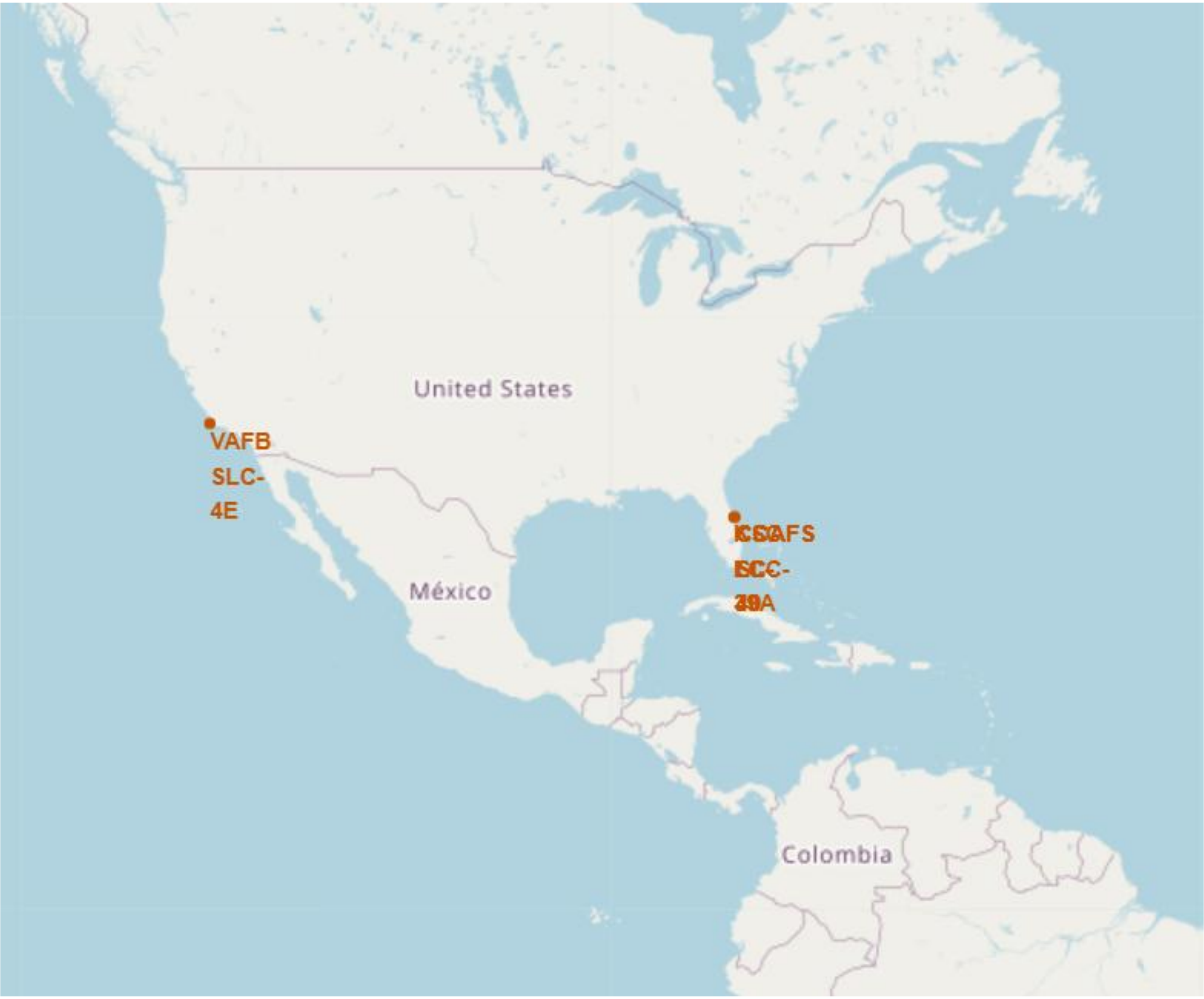
```
 * sqlite:///my_data1.db
Done.
  Landing_Outcome        CNT
 No attempt              10
 Success (drone ship)    5
 Failure (drone ship)    5
 Success (ground pad)    3
 Controlled (ocean)      3
 Uncontrolled (ocean)    2
 Failure (parachute)     2
 Precluded (drone ship)  1
```

Section 3

Launch Sites Proximities Analysis

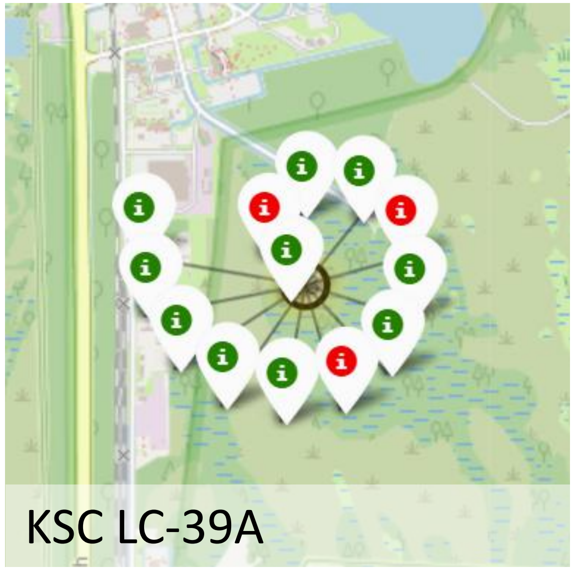
Space X Launch Sites

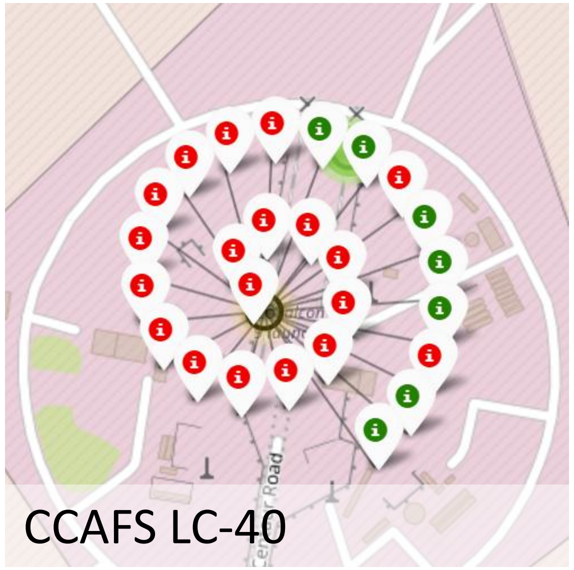
- The map displays the locations of four launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E
- The sites are spread across the southern United States, with CCAFS LC-40 and CCAFS SLC-40 are located very close to each other

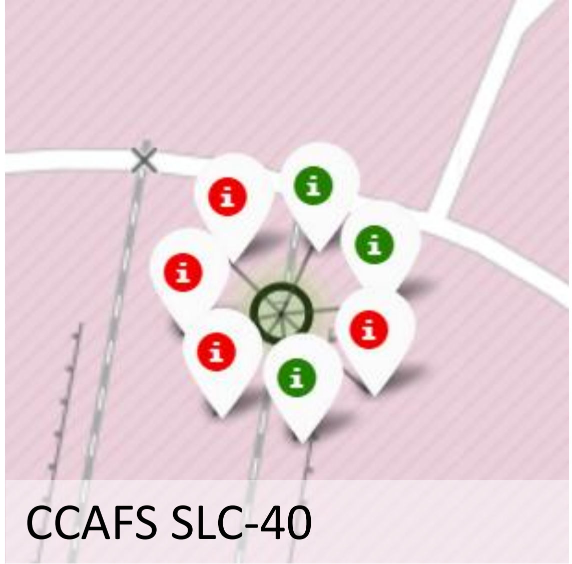


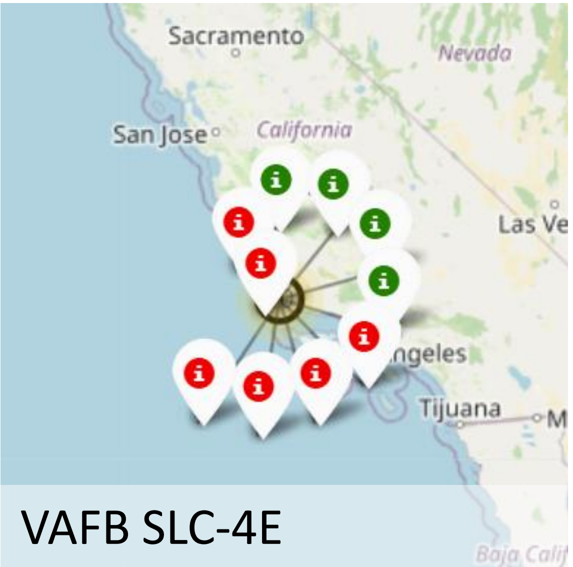
Launch Sites vs Outcome

- This map shows the launch outcomes (green for success, red for failure) for different launch sites.
- KSC LC-39A demonstrates a high success rate, while VAFB SLC-4E and CCAFS LC-40 show a mix of successes and failures.


KSC LC-39A


CCAFS LC-40

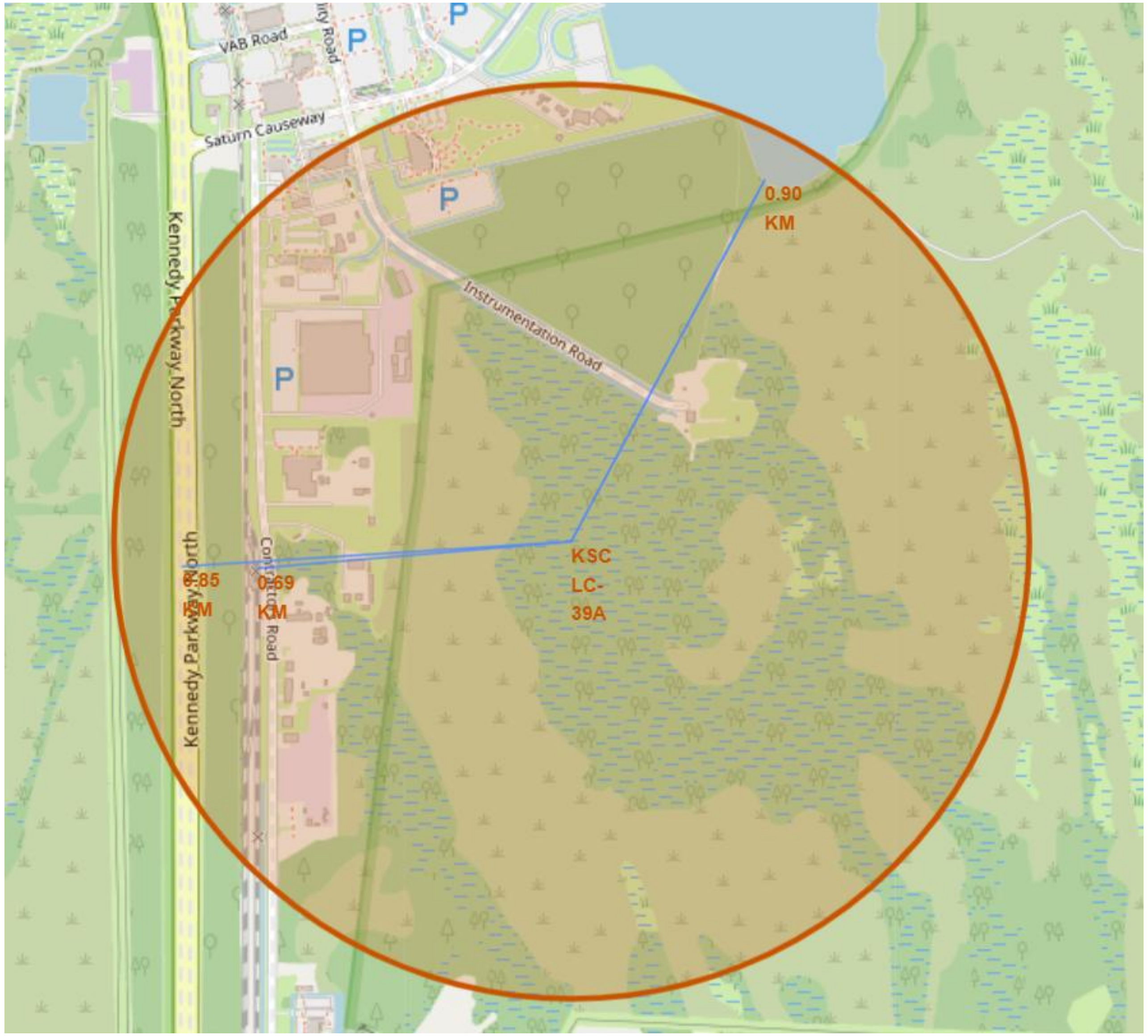

CCAFS SLC-40

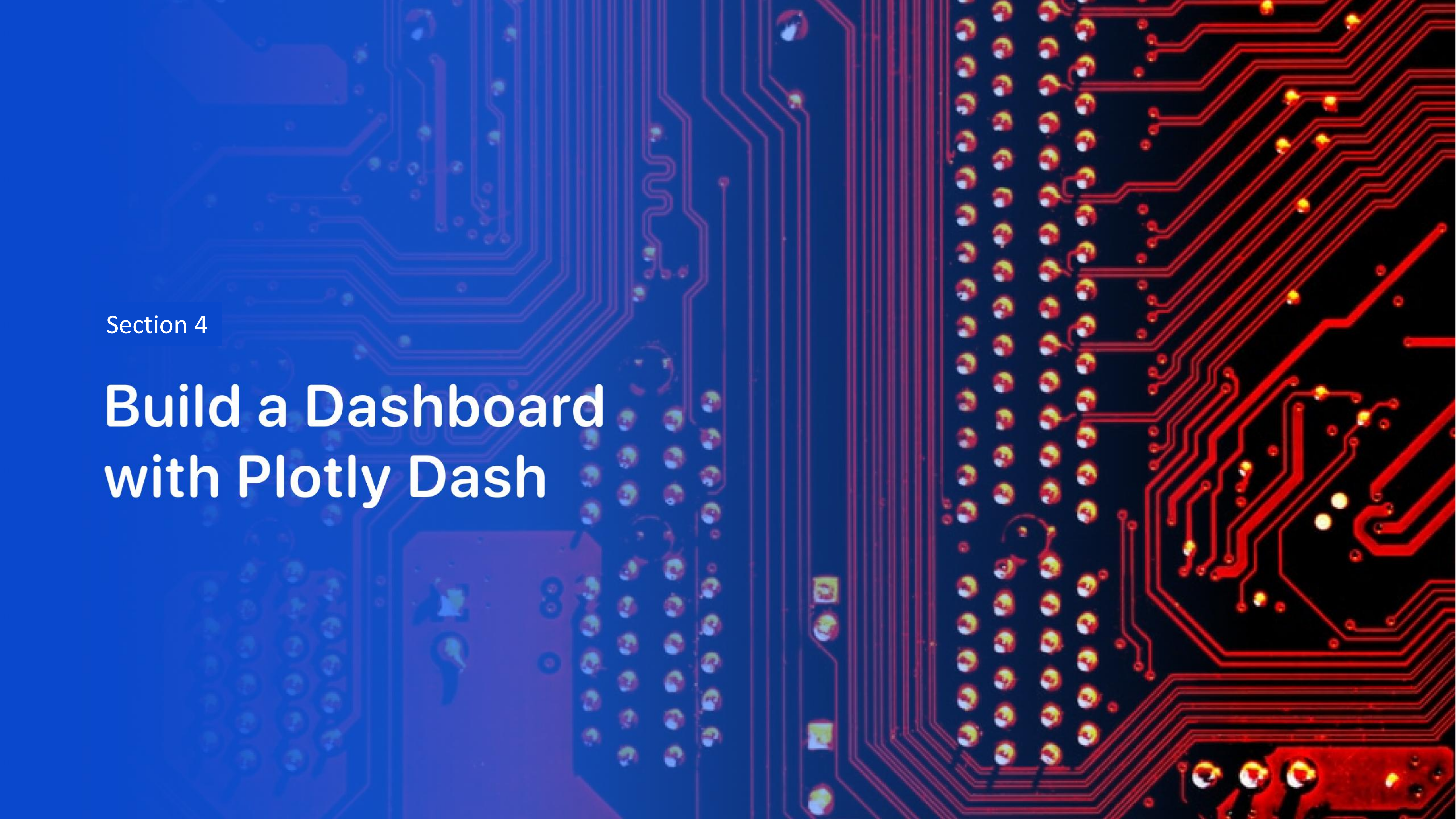


VAFB SLC-4E

Connectivity of KSC LC-39A Launch Site

- The map highlights the importance of infrastructure connectivity and geographical context for a launch site like KSC LC-39A. Direct road access, potential rail links, and proximity to the coastline are all critical factors for successful launch operations. The marked distance emphasizes the relevance of precise spatial measurements in this environment.



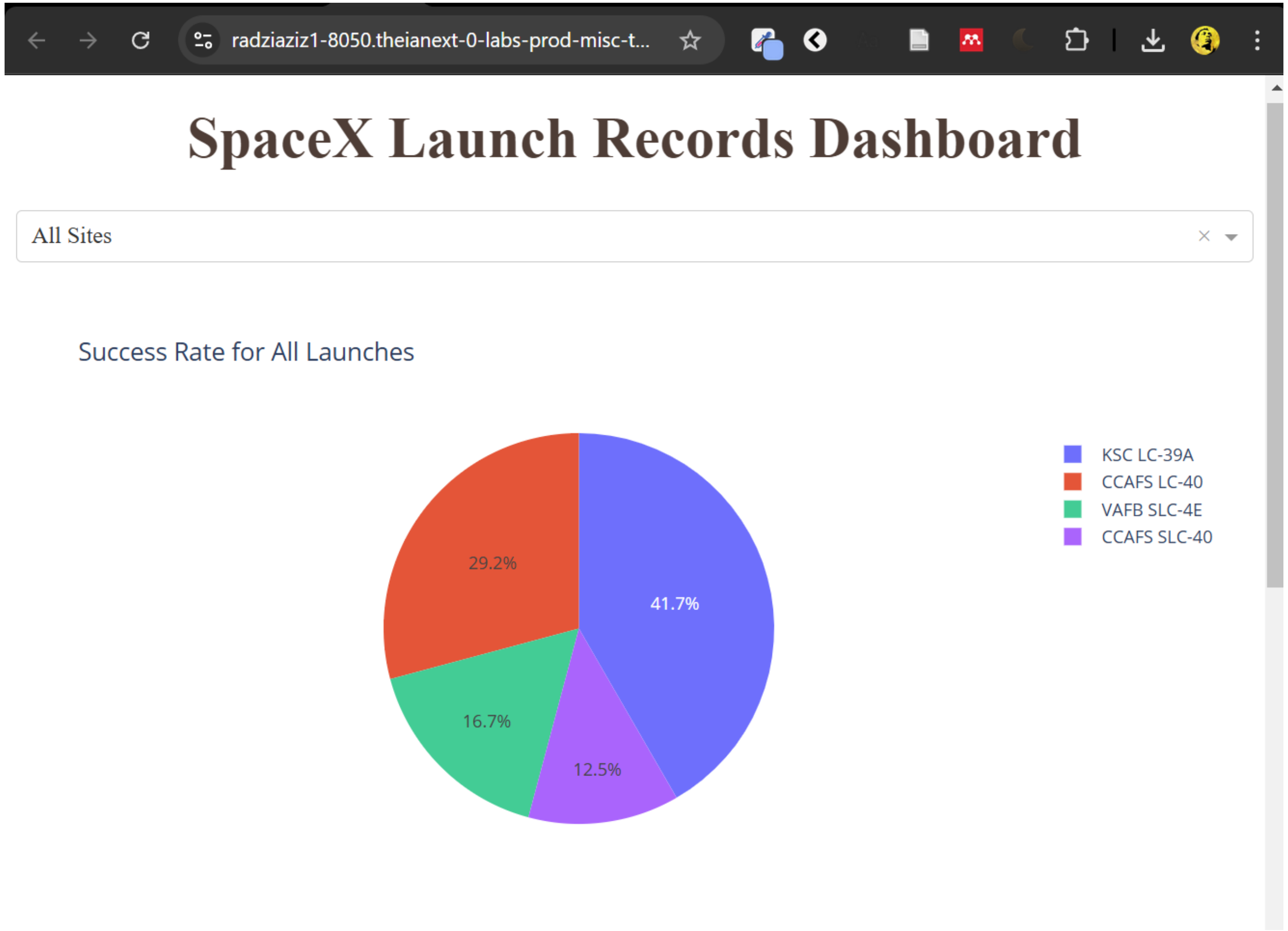


Section 4

Build a Dashboard with Plotly Dash

Visualization using Dashboard

- The dashboard provides visual comparison of launch success rates across different sites. The findings suggest KSC LC-39A has a largest success rates as compared to other sites



radziaziz1-8050.theianext-0-labs-prod-misc-t...

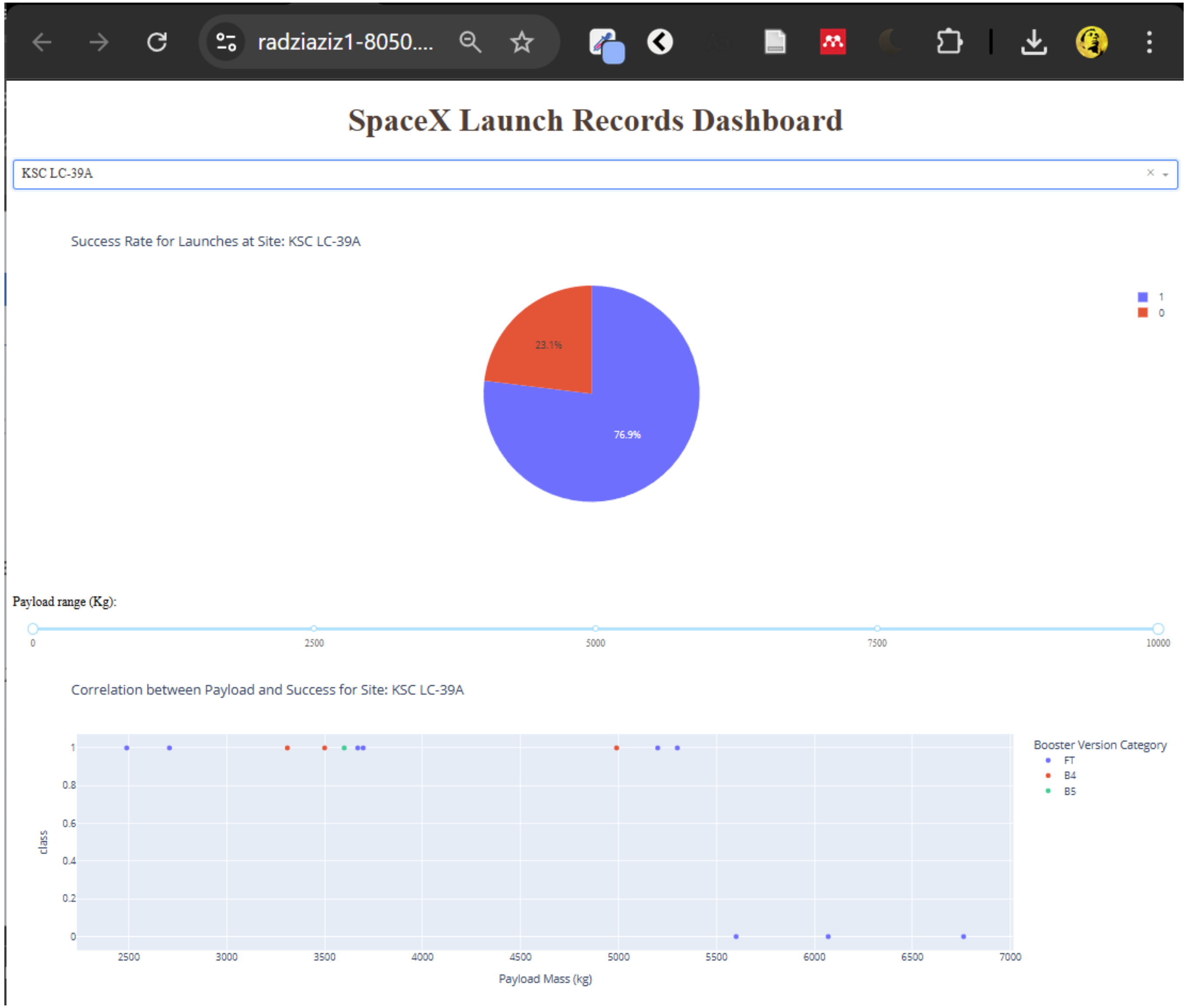
SpaceX Launch Records Dashboard

All Sites

Success Rate for All Launches

KSC LC-39A Performance Analysis

- KSC LC-39A shows a high success rate of 76.9%
- The plot strongly suggests a relationship between payload mass and launch success at KSC LC-39A. There's a clear "sweet spot" for payload mass where launches are highly successful. The data hints at a potential payload mass threshold beyond which the risk of failure increases considerably.



radziaziz1-8050....

SpaceX Launch Records Dashboard

KSC LC-39A

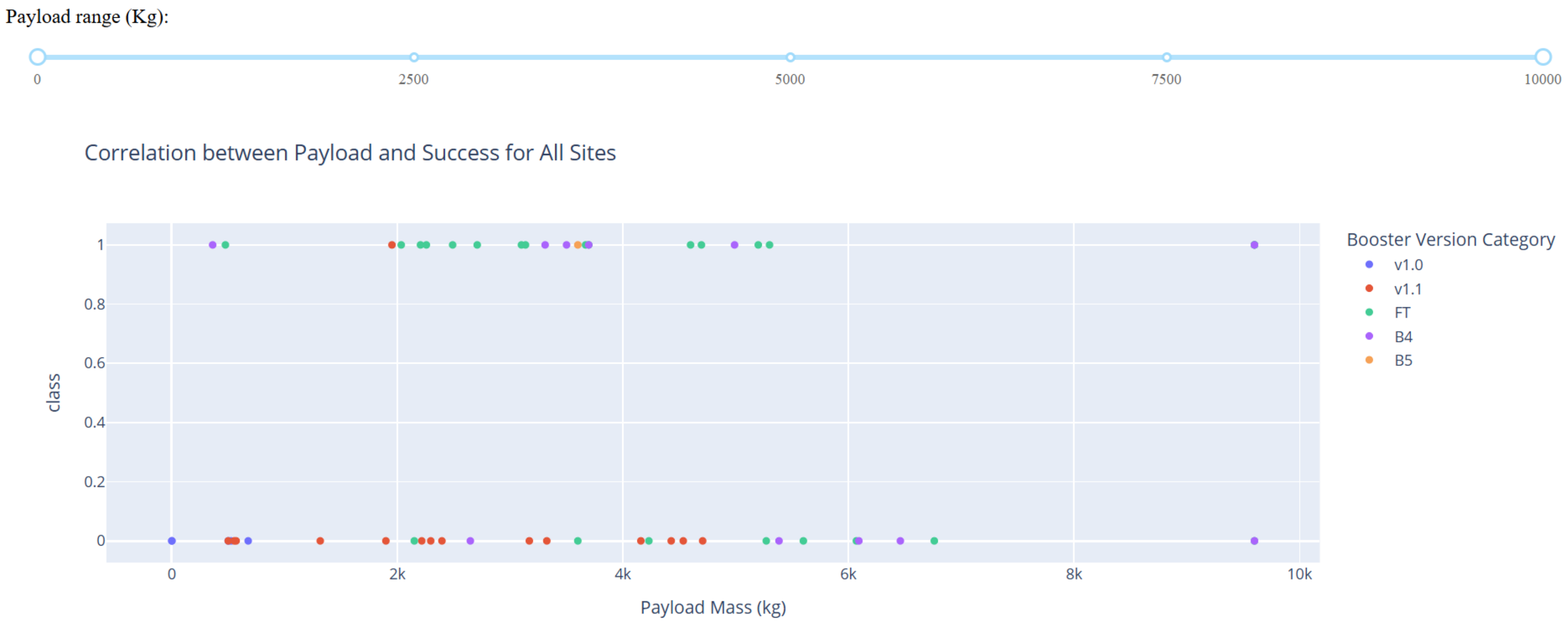
Success Rate for Launches at Site: KSC LC-39A

Payload range (Kg):

Correlation between Payload and Success for Site: KSC LC-39A

Payload Mass and Booster Version Analysis

- There might be a potential threshold effect for payload mass, where exceeding a certain 6000 kg increases the risk of failure
- Booster FT shows overall highest success rates, while booster version v1.1 show the poorest success rates

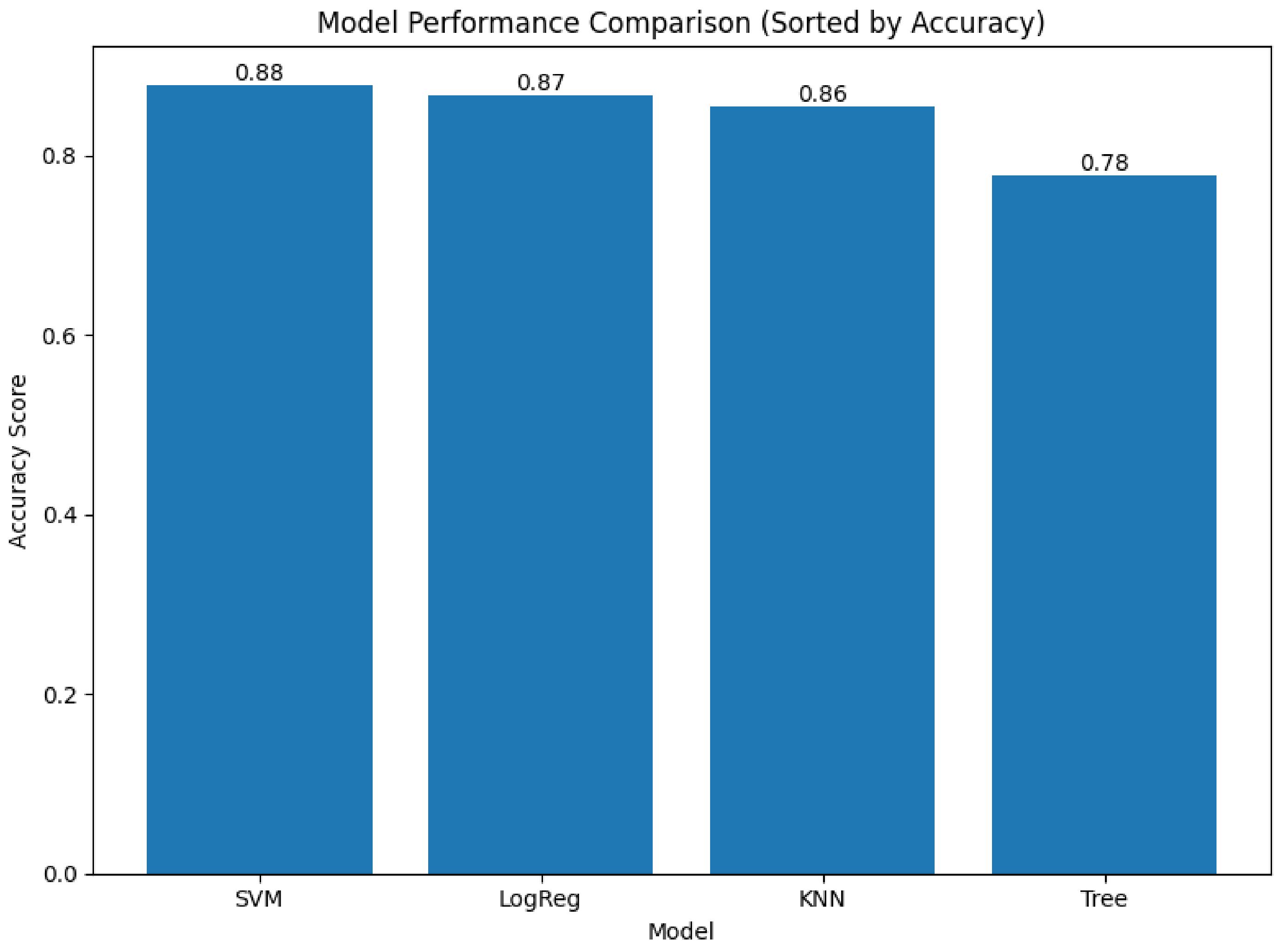


Section 5

Predictive Analysis (Classification)

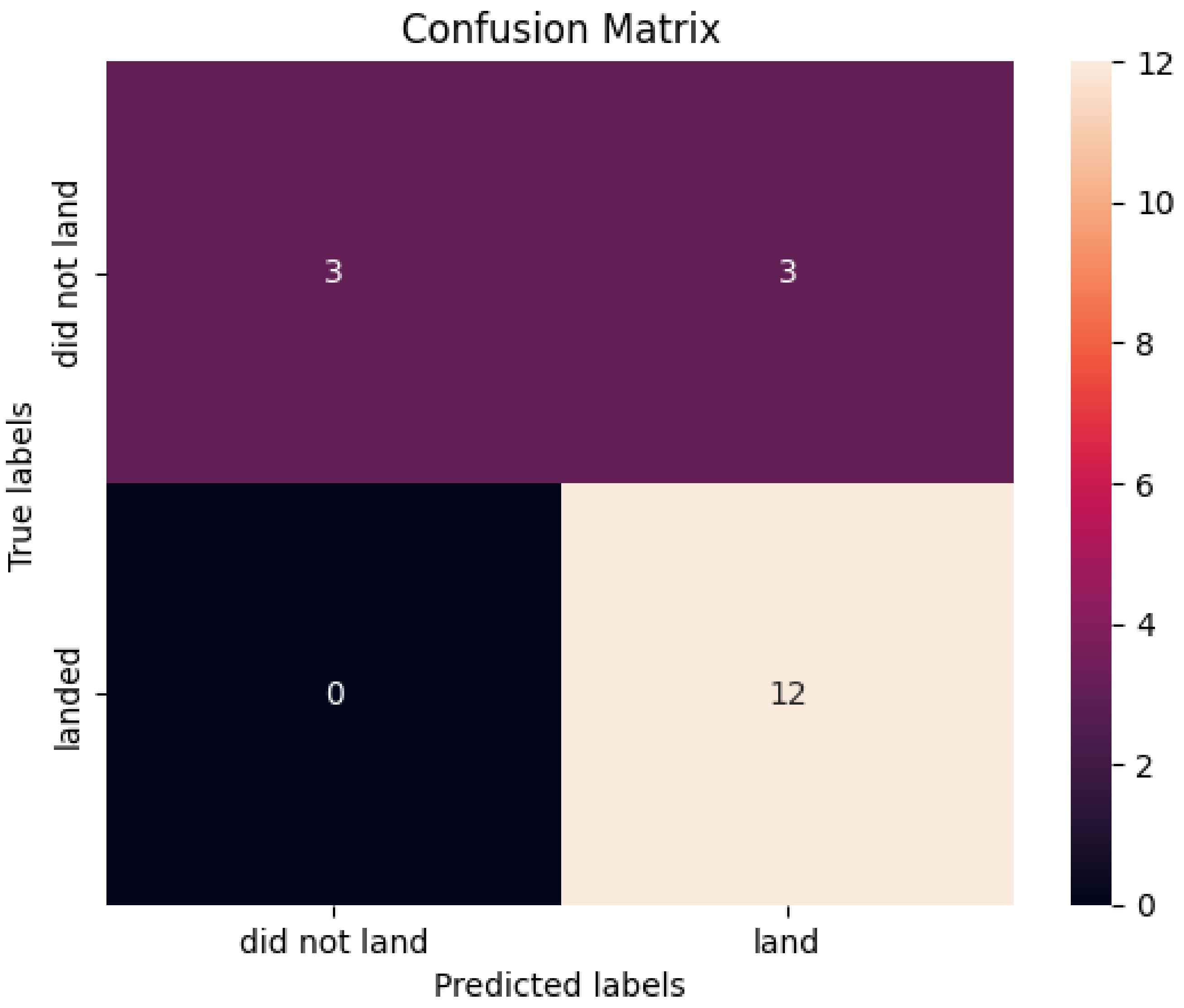
Classification Accuracy

- The chart shows SVM with the highest accuracy score of 0.88, followed closely by LogReg (0.87) and KNN (0.86). Tree has the lowest accuracy at 0.78



Confusion Matrix

All models perform the same in terms of the confusion matrix



Conclusions

- Accurate landing prediction improves launch cost estimation and business planning.
- EDA findings show correlations between payload, launch site, and success rate.
- Machine learning models effectively classify landing success, with SVM achieving the highest accuracy (0.88).
- Interactive dashboards enhance data-driven decision-making.
- Future work includes refining models with additional features such as weather conditions and booster flight history.

Appendix

- The appendix includes supporting materials used in the project, such as:
  - Python code snippets from data collection, wrangling, and model development.
  - SQL queries used for exploratory data analysis.
  - Charts and figures summarizing key insights from the analysis.
  - Notebook outputs showcasing intermediate results.
  - Datasets retrieved and processed during the study.
- All materials can be found at https://github.com/radziaziz/ds-spacex

Thank you!