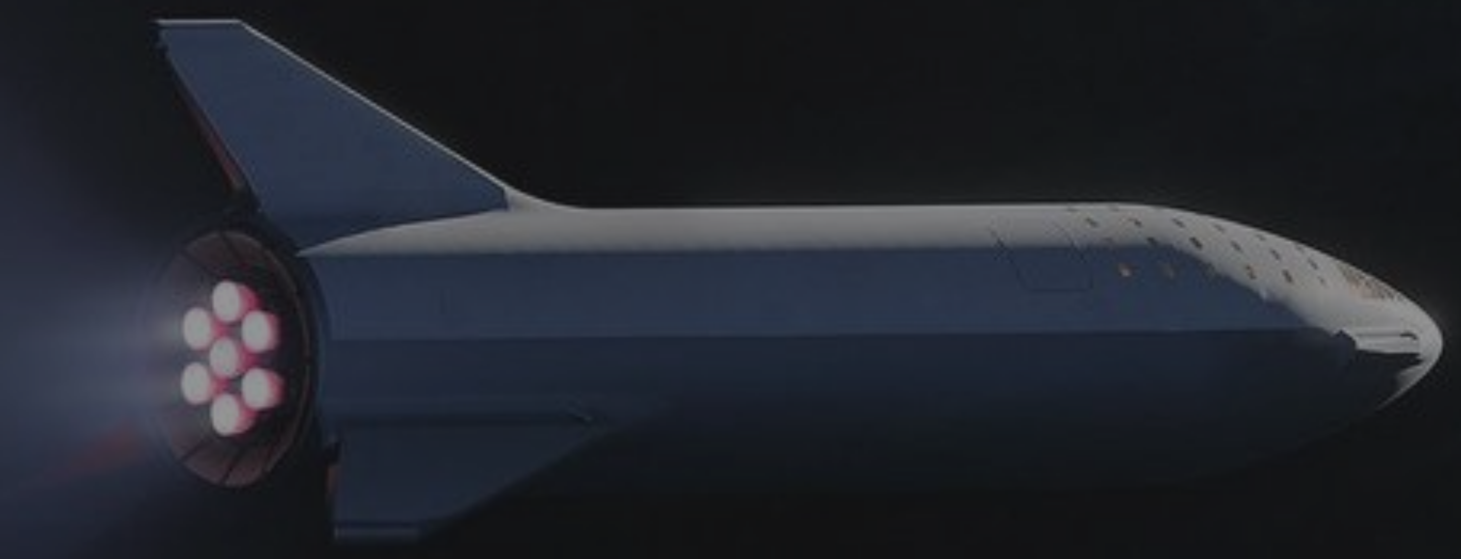
Applied Data Science Capstone

DATA-DRIVEN ANALYSIS OF SPACEX LAUNCHES

PREDICTING FALCON 9 FIRST-STAGE LANDING SUCCESS



EXECUTIVE SUMMARY

- Title: Predicting SpaceX Falcon 9 First Stage Landing Success
- Objective: Develop a predictive model to determine the success of the Falcon 9 first stage landing.
- Methodology: EDA, Interactive Visual Analytics, Predictive Modeling
- Key Findings:
  - Diverse Mission Capabilities: SpaceX has launched a wide range of payloads to various orbits, indicating a broad mission capability.
  - Improving Success Rates: The success rate of first-stage landings has generally improved over the years, reflecting technological and operational advancements.
  - Orbit-Specific Success Variability: Different orbits exhibit varying success rates, which may correlate with mission and payload specifics.
  - Model Performance: The Decision Tree model showcased the highest accuracy among evaluated models, though further validation is necessary to ensure it's not overfitting

INTRODUCTION

SpaceX has been working hard to make space travel more sustainable and cost-effective. However, to achieve this goal, it is crucial to accurately predict the success of Falcon 9's first stage landing. The use of machine learning models can help in generating these predictions, which can ultimately enhance strategic planning, reduce risks, and optimize resource allocation in future missions. In this analysis, we will train, evaluate, and compare multiple machine learning models, such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The comparative study aims to scrutinize the predictive abilities of these models using a range of evaluation metrics. The goal is to determine the model that outperforms others in accurately predicting the landing success of Falcon 9's first stage. This approach can lead to more data-driven, insightful, and strategically sound decision-making in SpaceX's future missions.

METHODOLOGY

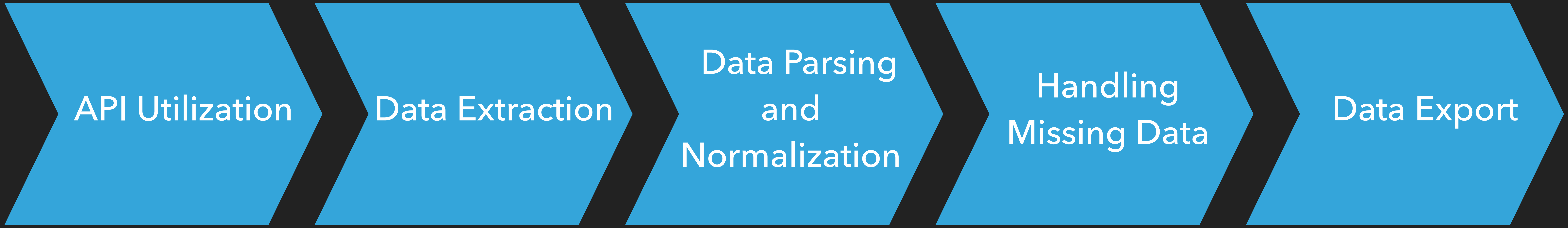
- Data collection
- Data wrangling
- EDA
- Interactive visual analytics
- Predictive analytics

DATA COLLECTION METHODOLOGY

The process of predicting the landing of SpaceX Falcon 9 first stage involves collecting data from two sources:

1. **The SpaceX API**
   - Link: https://api.spacexdata.com/v4/launches/past
   - Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.
2. **Web scraping from Wikipedia**
   - Link: List of Falcon 9 and Falcon Heavy launches https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
   - Columns: Flight No., Date, Time, Version Booster, Launch site, Payload, PayloadMass, Customer, Launch outcome, Booster Landing

DATA COLLECTION API METHODOLOGY



API Utilization

Data Extraction

Data Parsing and Normalization

Handling Missing Data

Data Export

Utilize Python's requests library to send GET requests to the API.

Defined Python functions to extract detailed information

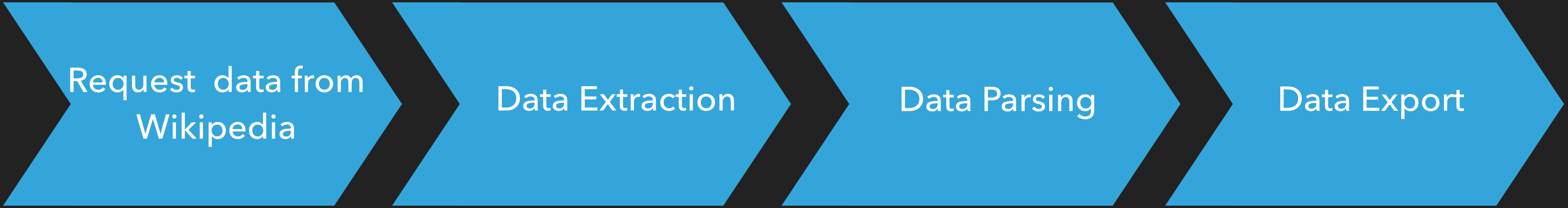
Utilized json_normalizeto flatten the JSON data and transform it into tabular format.

Checked for missing values. Calculated the mean of the PlayloadMass column and replaced NaN values with this mean value

Data is exported to a CSV file

GitHub : [Data Collection API]

DATA COLLECTION WITH WEB SCRAPPING METHODOLOGY



Request data from Wikipedia

Data Extraction

Data Parsing

Data Export

Retrieve the HTML content of the Falcon9 Launch Wiki page. Utilize the BeautifulSoup class to parse the HTML content.

Use find_all(table) to fetch all table elements. Extract relevant column names from the HTML table.

Initialize an empty dictionary lanch_dict with keys as column names. Create a structured DataFrame by extracting and parsing launch records.

Convert lanch_dict into a Pandas DataFrame

GitHub : [Data Collection with Web Scrapping]()

DATA WRANGLING METHODOLOGY

Data wrangling aims to perform Exploratory Data Analysis (EDA) to discover patterns and determine suitable labels for supervised model training.

- Illustrate various landing scenarios (e.g., True Ocean, False Ocean) and related visual content are provided to showcase the different possible outcomes of booster landings.
- Create a landing outcome label that will classify the outcome of each launch based on the Outcome column. A value of zero means the first stage did not land successfully, while a value of one means the first stage landed successfully; assign this to the Class column.
- The success rate of the landings is calculated using the mean of the Class column, returning approximately 0.67.

GitHub : [Data Wrangling]()

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

EDA WITH SQL

- This step aims to understand the SpaceX dataset, which contains records for each payload carried during SpaceX missions to outer space.
- Application of SQL Queries
  - Load the SpaceX dataset into a corresponding table in a Db2 database.
  - Execute SQL queries to derive insights and answers to specific questions related to the assignment or analysis.

GitHub : EDA with SQL

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

EDA WITH VISUALIZATION

- The process of exploring relationships between variables can help in understanding and summarizing the main characteristics of a dataset, often visualizing these features for easier interpretation.
- Various visualizations, such as scatter plots, bar plots, and line plots, were performed to provide insight into the factors affecting the success of Falcon 9 landings and how different variables interrelate.

GitHub : EDA with Visualization

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

- Previous EDAs hinted at potential correlations between launch sites and success rates.
- Utilized Folium for interactive mapping, enabling detailed visual exploration of launch sites' geographical and environmental contexts.
- Investigate the influence of launch site geographical locations on SpaceX launch success rates using interactive visual analytics.

GitHub : Interactive Visual Analytics with Folium

EDA AND INTERACTIVE VISUAL ANALYTICS METHODOLOGY

DASHBOARD WITH PLOTLY DASH

- Engaging Dashboards: Develop user-friendly dashboards to allow stakeholders to delve into the data, exploring various facets interactively, and deriving actionable insights.
- Core Features of the Dashboard
  - Dynamic Input Components: Features like dropdown lists and range sliders facilitate user-driven interactions with the data visualizations.
  - Interactive Visualizations: Including pie charts and scatter plots, providing comprehensive views and insights into the SpaceX launch data.

GitHub : Interactive Dashboard with Ploty Dash

PREDICTIVE ANALYSIS METHODOLOGY

PREDICTIVE ANALYSIS

- Develop a structured machine learning pipeline to predict the success of SpaceX's Falcon 9 first stage landing.
    - **Determine the target variable:** Create a class column that indicates whether the landing was successful or not.
    - **Data Standardization:** Standardize feature variables to ensure consistent scaling and improve model training efficacy.
    - **Train-Test Data Splitting:** Split the dataset into training and test subsets to facilitate model training and validation.
    - **Hyperparameter Tuning for Multiple Models:** Employ techniques GridSearchCV to optimize hyperparameters and enhance model performance.
    - **Model Evaluation and Selection:** Evaluate the performance of each model on the test data, considering metrics like accuracy, precision, and recall. Identify the model that demonstrates the highest predictive accuracy and robustness.

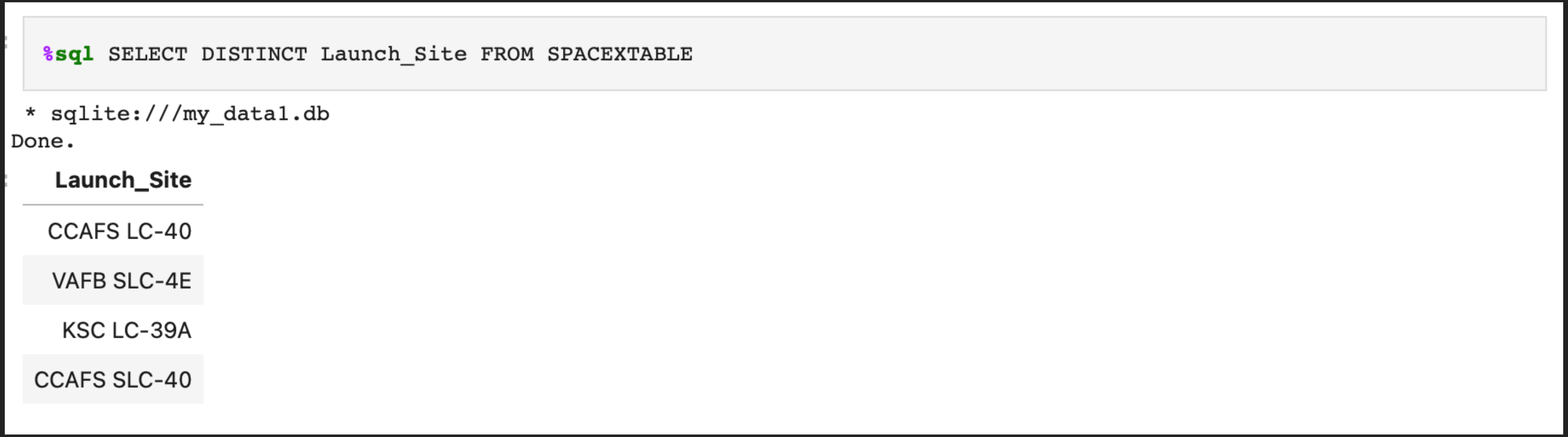
GitHub : Machine Learning Prediction

RESULTS

- EDA with visualization
- EDA with SQL
- Interactive map with Folium
- Plotly Dash dashboard
- Predictive analysis

EDA WITH SQL RESULTS (1)

1. Display the names of the unique launch sites in the space mission



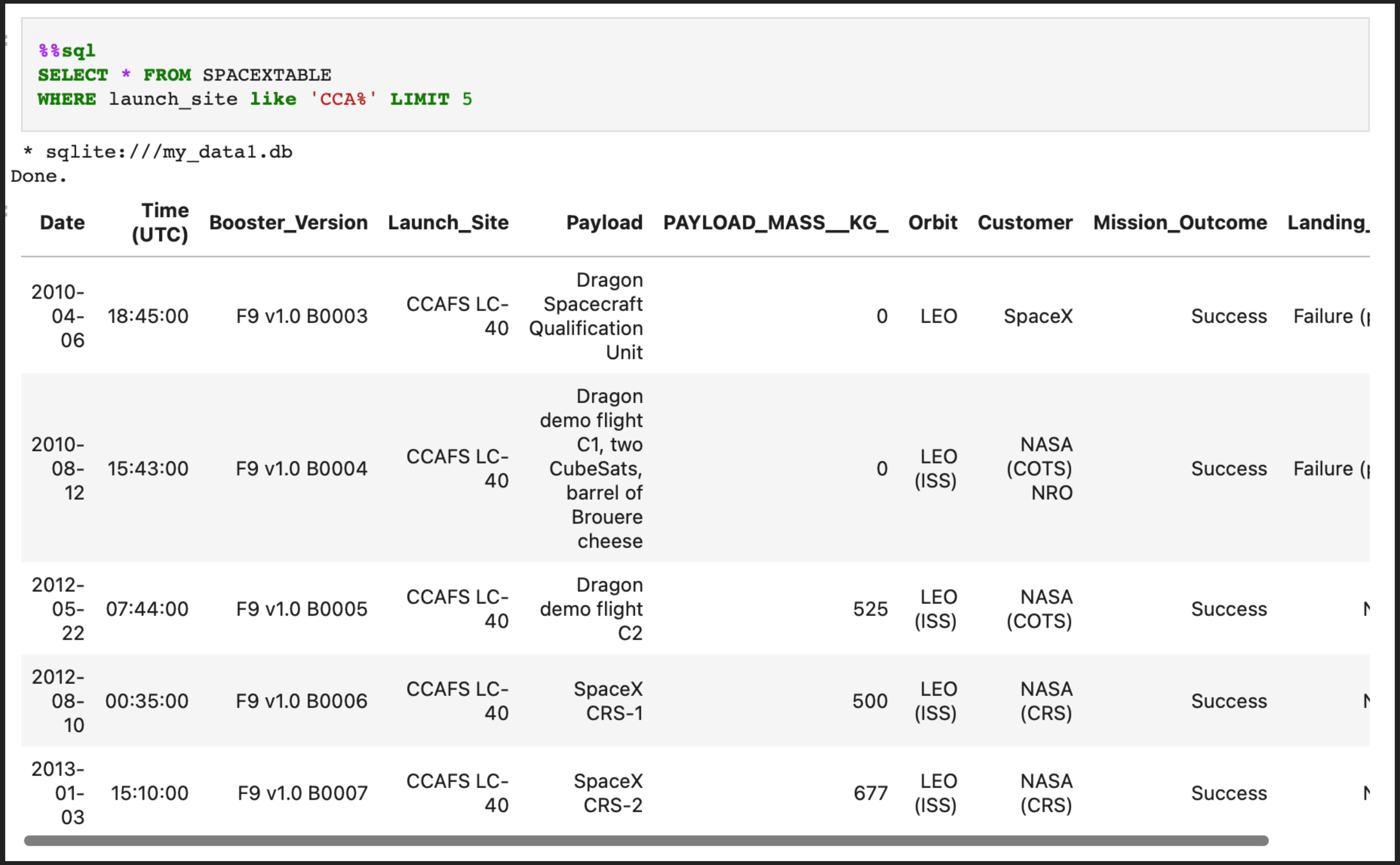
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

EDA WITH SQL RESULTS (2)

2. Display 5 records where launch sites begin with the string 'CCA'



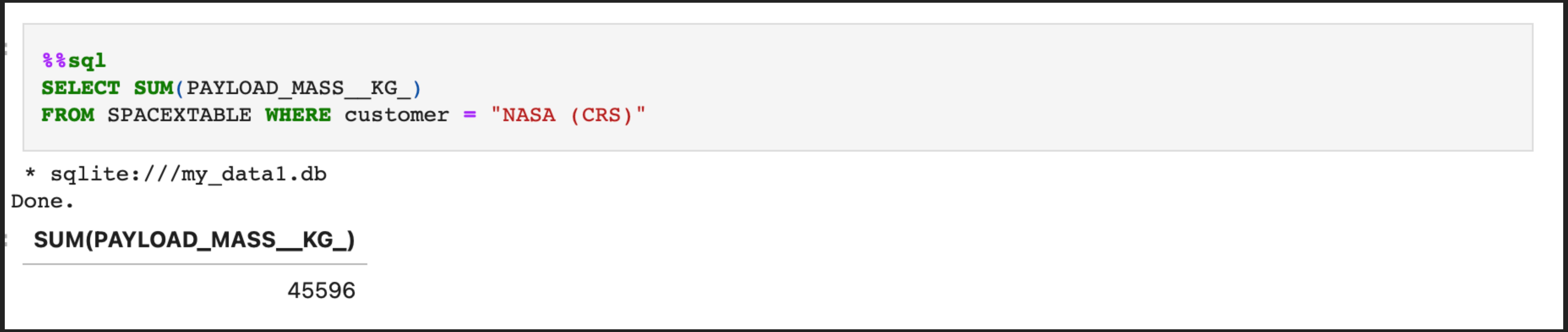
```
%%sql
SELECT * FROM SPACEXTABLE
WHERE launch_site like 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (
2010-08-12	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 (
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

EDA WITH SQL RESULTS (3)

3. Display the total payload mass carried by boosters launched by NASA (CRS)



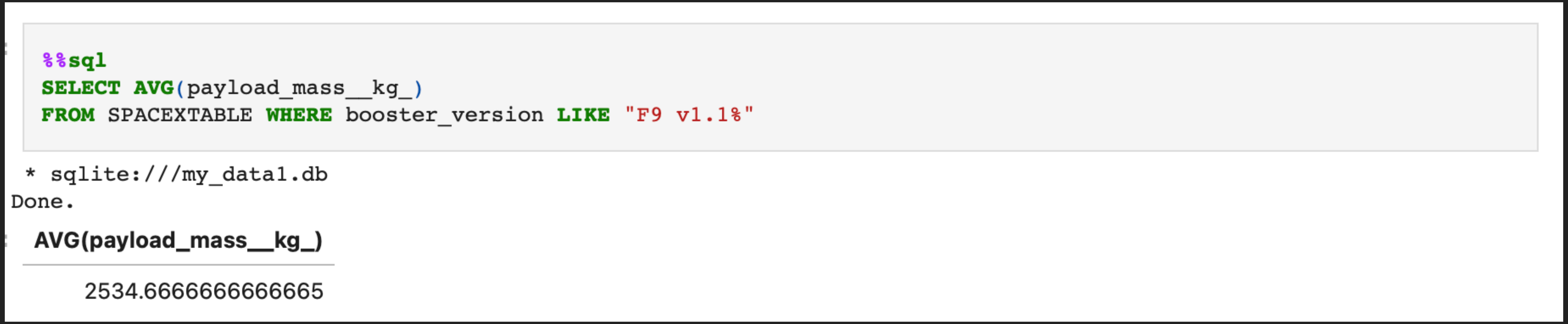
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE WHERE customer = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

EDA WITH SQL RESULTS (4)

4. Display average payload mass carried by booster version F9 v1.1



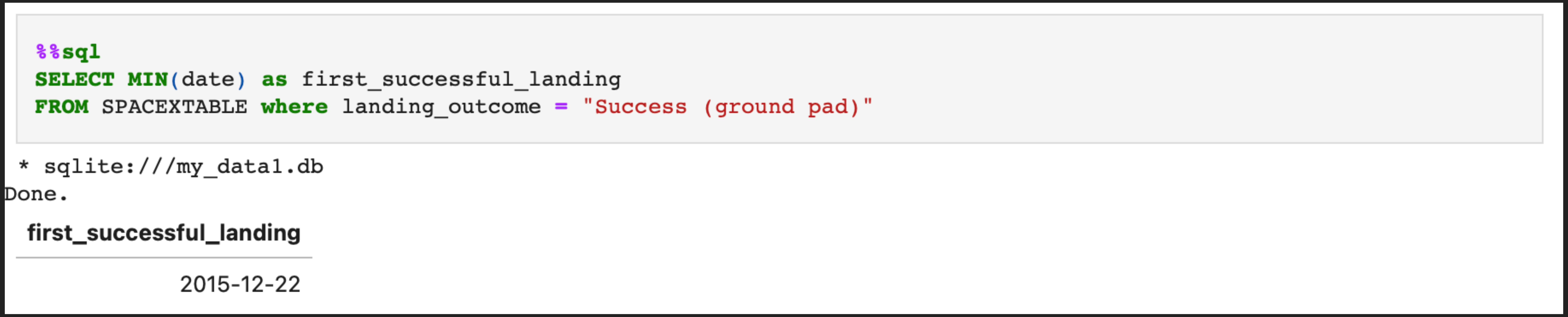
```
%%sql
SELECT AVG(payload_mass__kg_)
FROM SPACEXTABLE WHERE booster_version LIKE "F9 v1.1%"
```

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

AVG(payload_mass__kg_)
2534.6666666666665

EDA WITH SQL RESULTS (5)

5. List the date when the first succesful landing outcome in ground pad was acheived



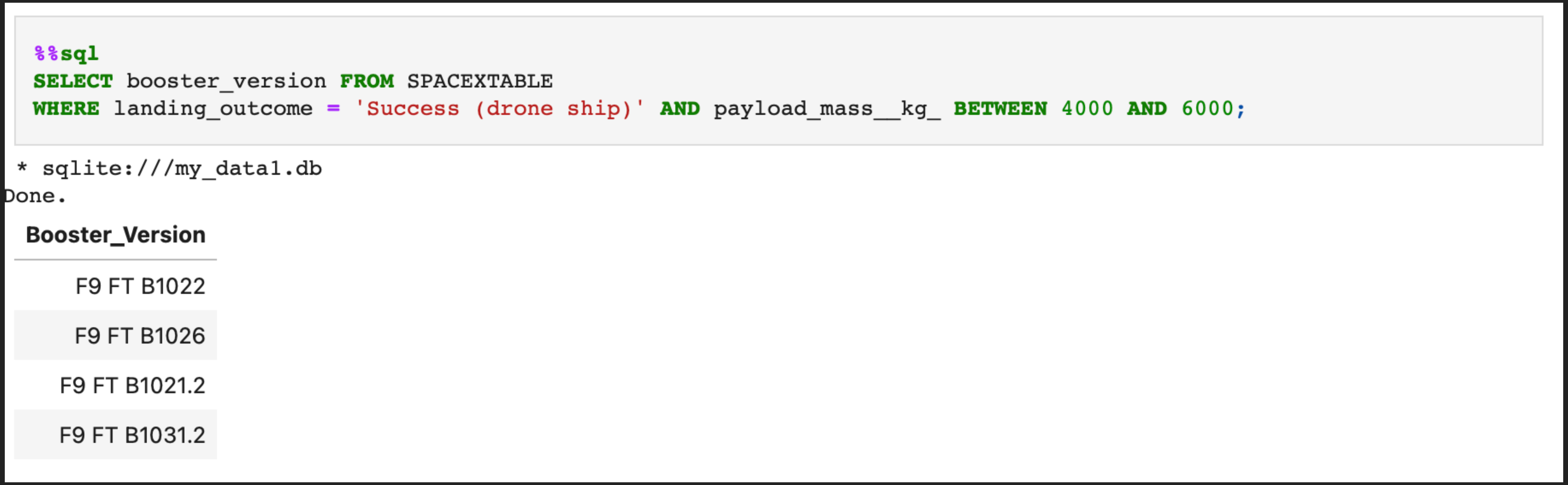
```
%%sql
SELECT MIN(date) as first_successful_landing
FROM SPACEXTABLE where landing_outcome = "Success (ground pad)"
```

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

first_successful_landing
2015-12-22

EDA WITH SQL RESULTS (6)

6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000



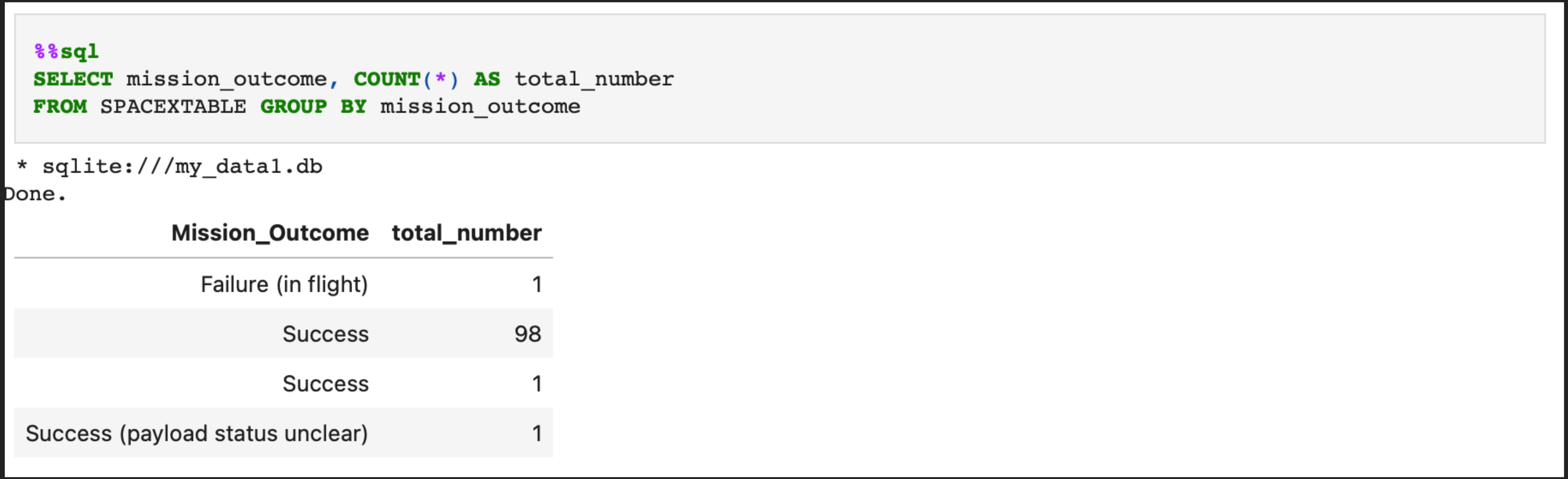
```
%%sql
SELECT booster_version FROM SPACEXTABLE
WHERE landing_outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
```

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

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

EDA WITH SQL RESULTS (7)

7. List the total number of successful and failure mission outcomes



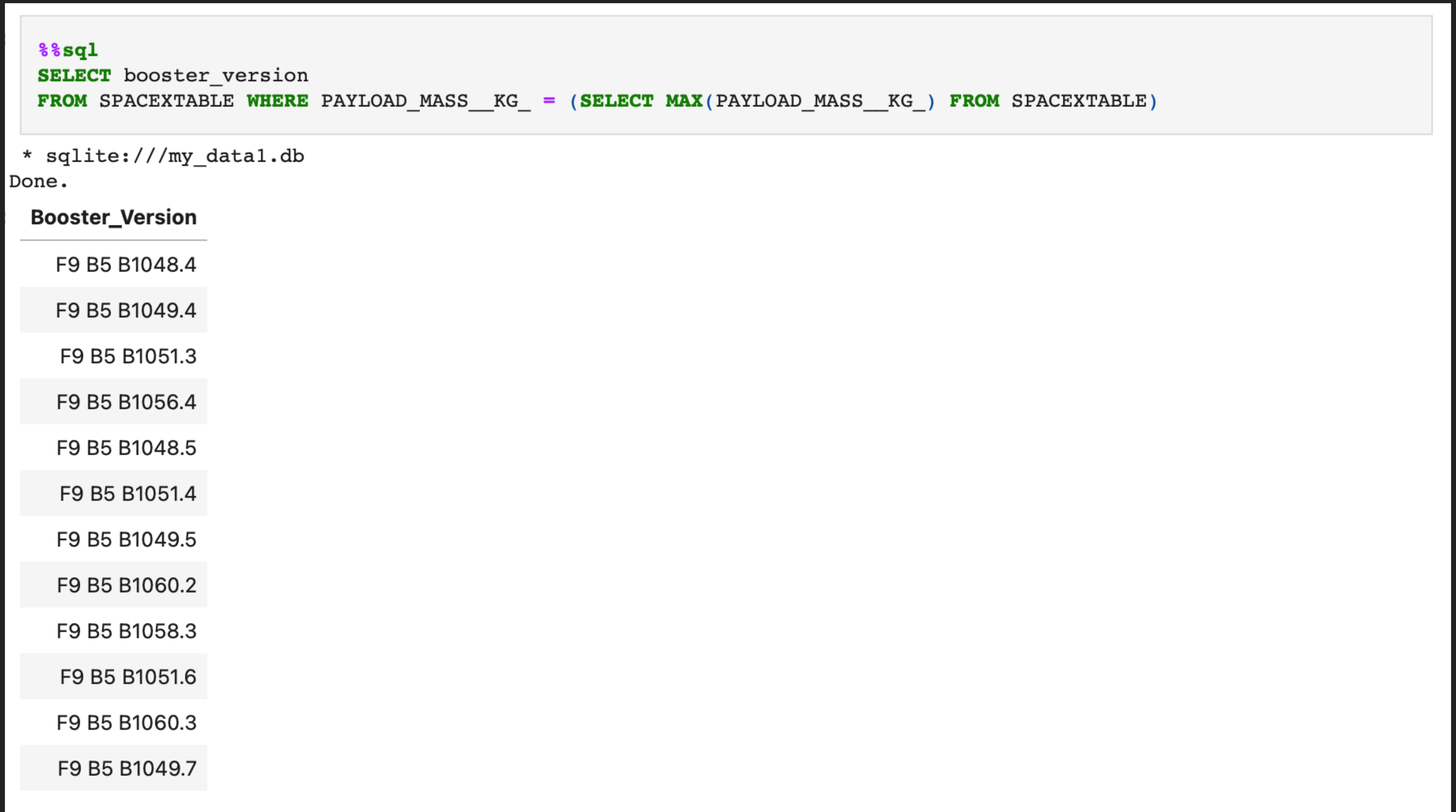
```
%%sql
SELECT mission_outcome, COUNT(*) AS total_number
FROM SPACEXTABLE GROUP BY mission_outcome
```

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

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

EDA WITH SQL RESULTS (8)

8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery



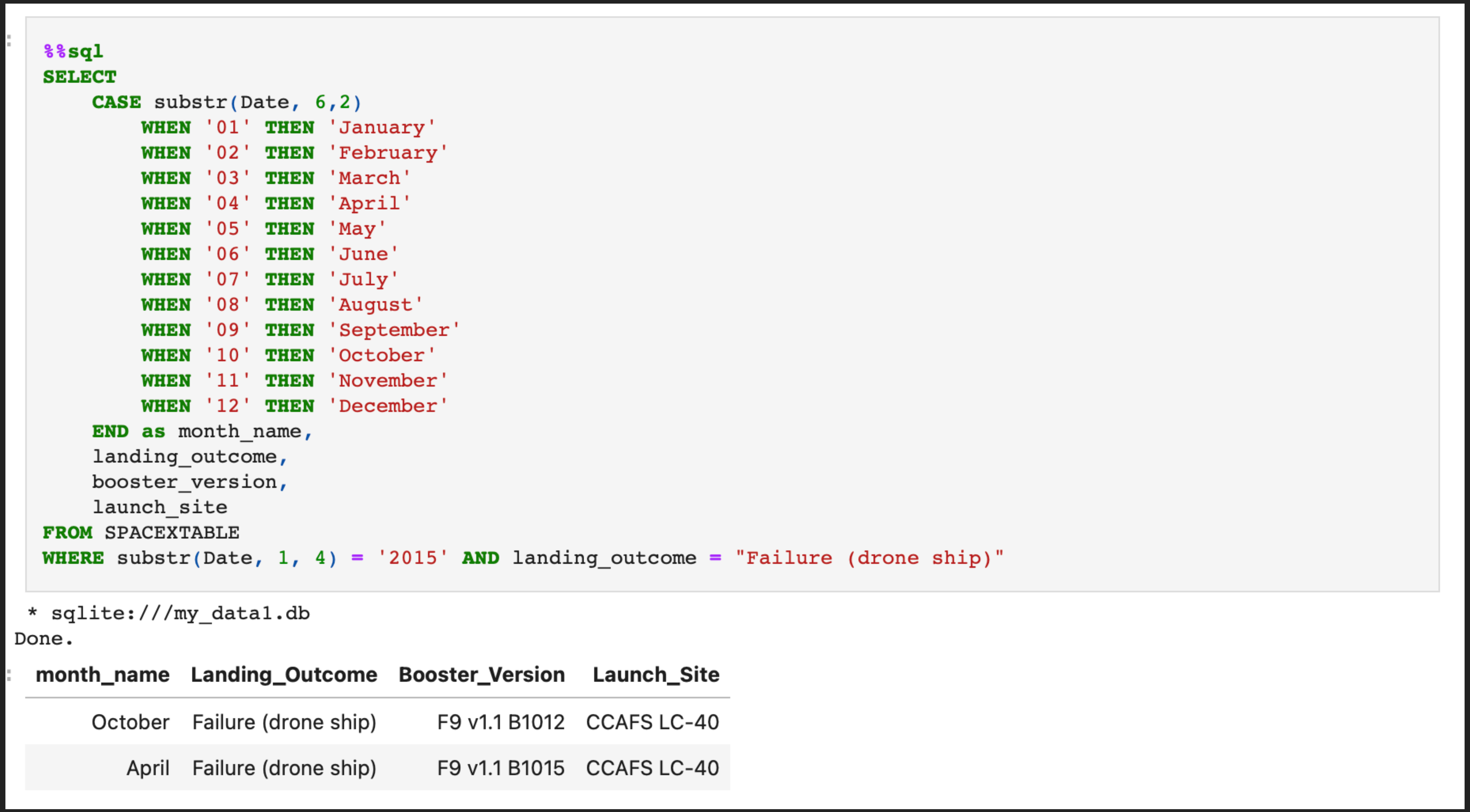
```
%%sql
SELECT booster_version
FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

```
 * 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

EDA WITH SQL RESULTS (9)

9. List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015



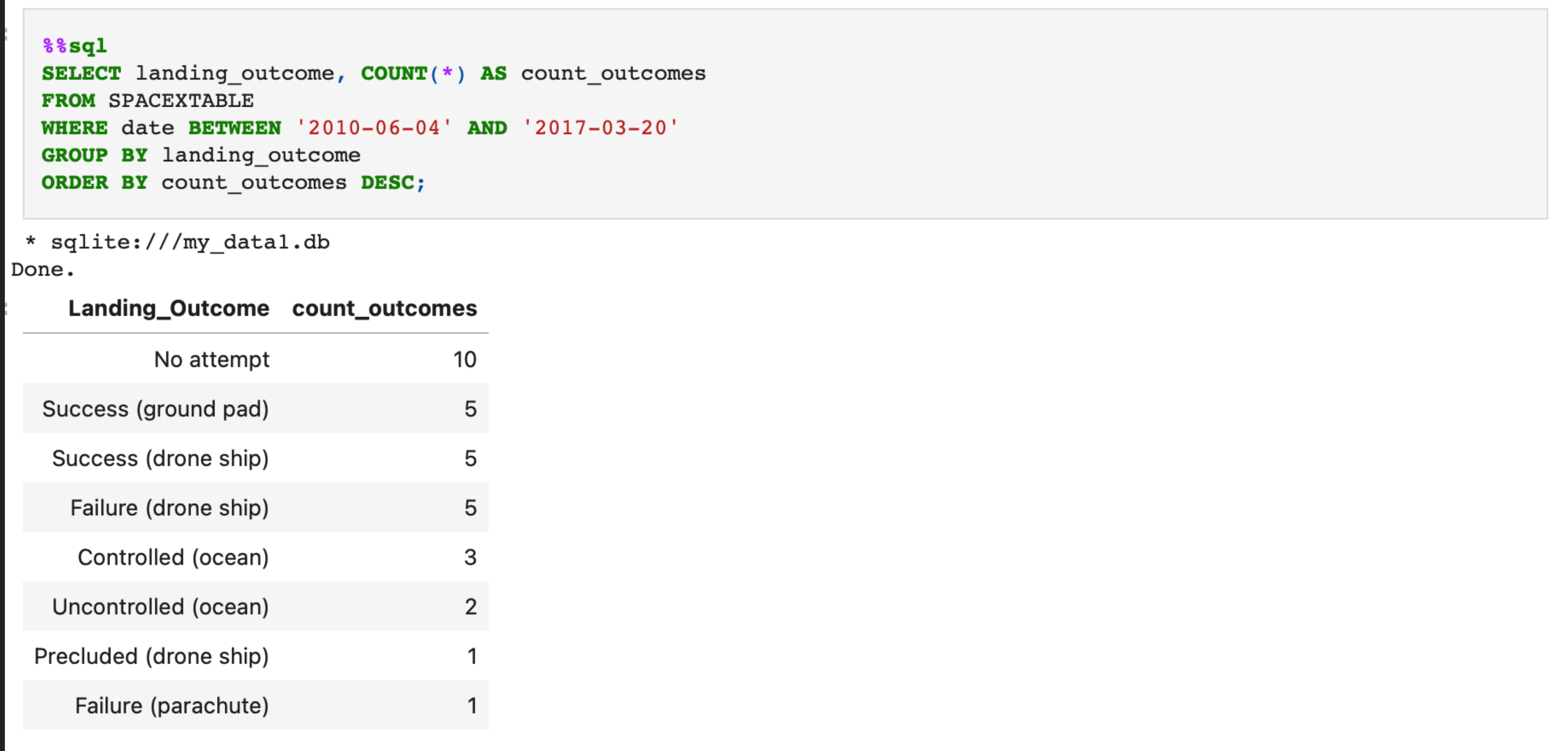
```
%%sql
SELECT
    CASE substr(Date, 6,2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END as month_name,
    landing_outcome,
    booster_version,
    launch_site
FROM SPACEXTABLE
WHERE substr(Date, 1, 4) = '2015' AND landing_outcome = "Failure (drone ship)"
```

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

month_name	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

EDA WITH SQL RESULTS (10)

10. 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

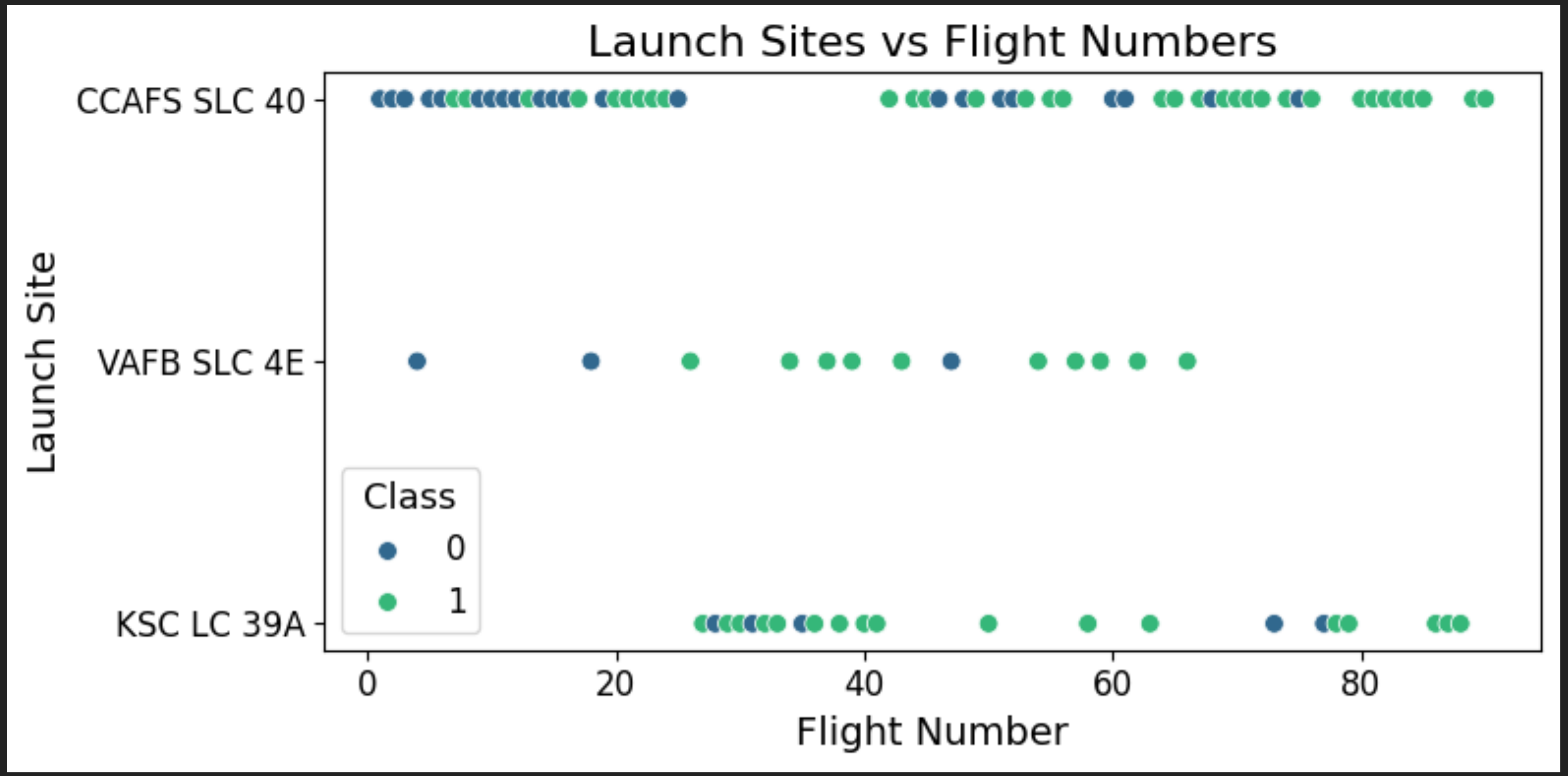


```
%%sql
SELECT landing_outcome, COUNT(*) AS count_outcomes
FROM SPACEXTABLE
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing_outcome
ORDER BY count_outcomes DESC;
```

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

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

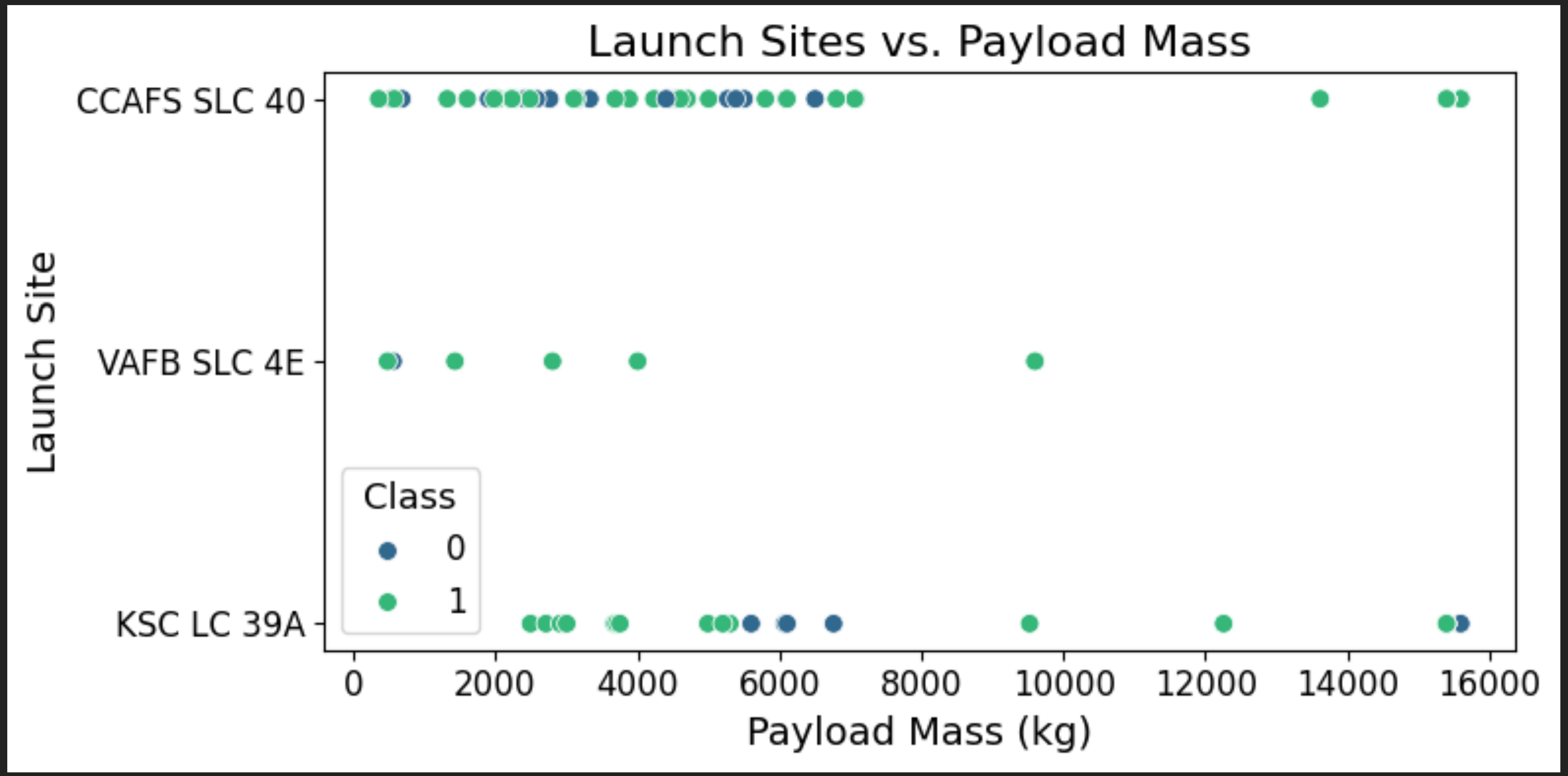
EDA WITH VISUALIZATION RESULTS (1)



LAUNCH SITE VS. FLIGHT NUMBER

- CCAFS SLC 40 has hosted the majority of flights, followed by VAFB SLC 4E and KSC LC 39A.
- The CLAFS SLS 40 launch site is more frequently used for earlier flights while KSC LC 39A was introduced later.
- There seems to be an improvement over time, with more successful launches (depicted by the color-coded classes).

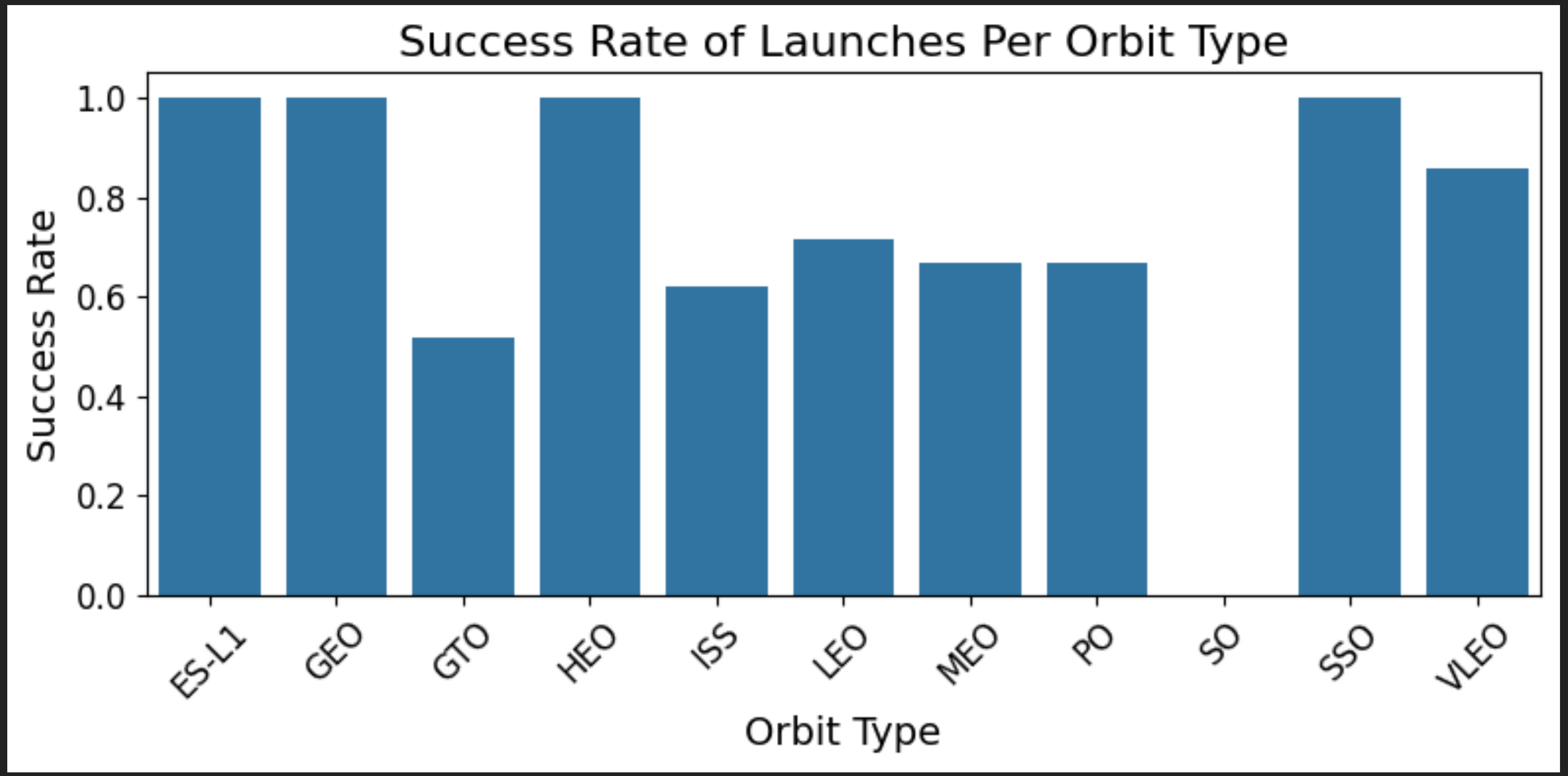
EDA WITH VISUALIZATION RESULTS (2)



LAUNCH SITE VS. PAYLOAD MASS

- CCAFS SLC 40: This site has handled a wide range of payload masses and has experienced both successful and unsuccessful launches, which are distributed across various payload sizes.
- VAFB SLC 4E: Primarily focused on lighter payloads, with a notable number of successful launches. The unsuccessful launches do not show a clear pattern relative to payload mass.
- KSC LC 39A: While handling a broad range of payload masses, this site tends to cater to heavier payloads with a high success rate.

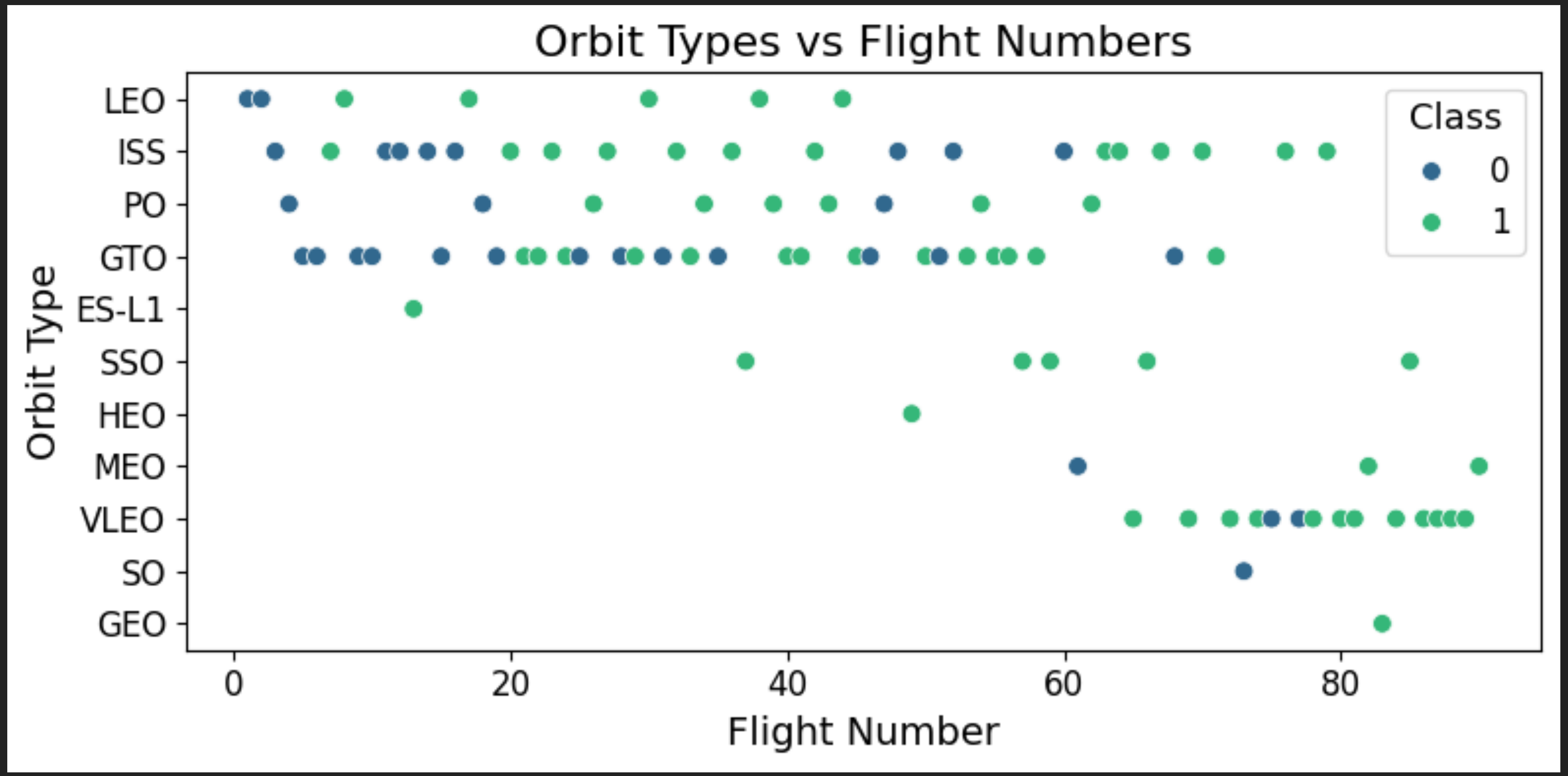
EDA WITH VISUALIZATION RESULTS (3)



SUCCESS RATE OF EACH ORBIT

- ES-L1, GEO, HEO, and SSO have a success rate of 100%, meaning all launches aiming for these orbits have been successful.
- VLEO has a notably high success rate of approximately 86%.
- LEO, MEO, and PO orbits have success rates around or above 66%.
- ISS and GTO orbits have success rates of approximately 62% and 52% respectively.
- SO (Sun-Synchronous Orbit) has not seen success in the available data, with a success rate of 0%.

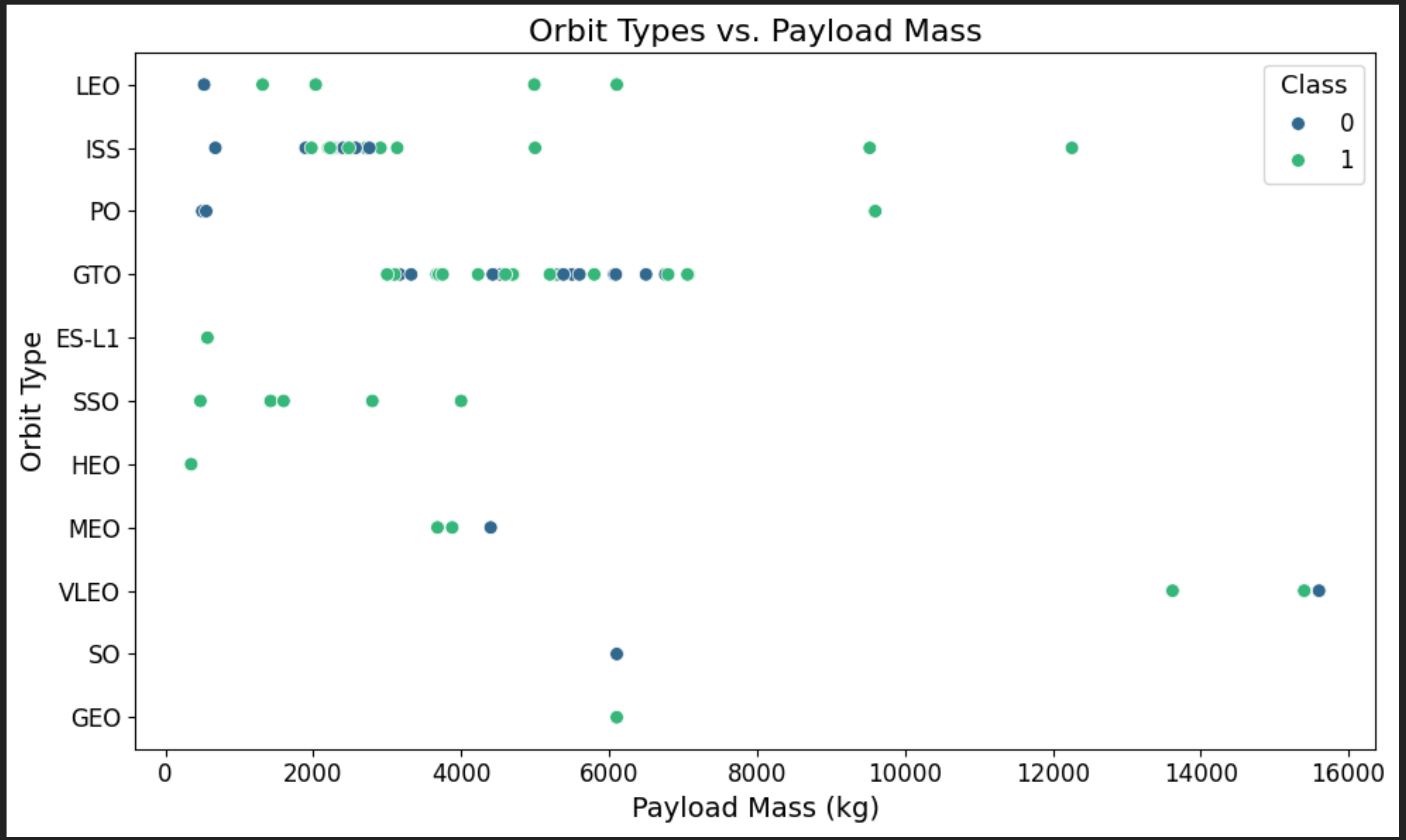
EDA WITH VISUALIZATION RESULTS (4)



ORBIT VS FLIGHT NUMBER

- LEO, GTO, and ISS orbits were the primary targets in earlier flights.
- There's an apparent progression in targeting different orbit types over time, with MEO, HEO, SSO, VLEO, and ES-L1 being targeted in later flights.
- Launches targeting GTO, LEO, and ISS orbits exhibit varied success, with both successful and unsuccessful instances throughout.
- Other orbit types like SSO, HEO, and ES-L1 show consistent success but are also less frequently targeted.

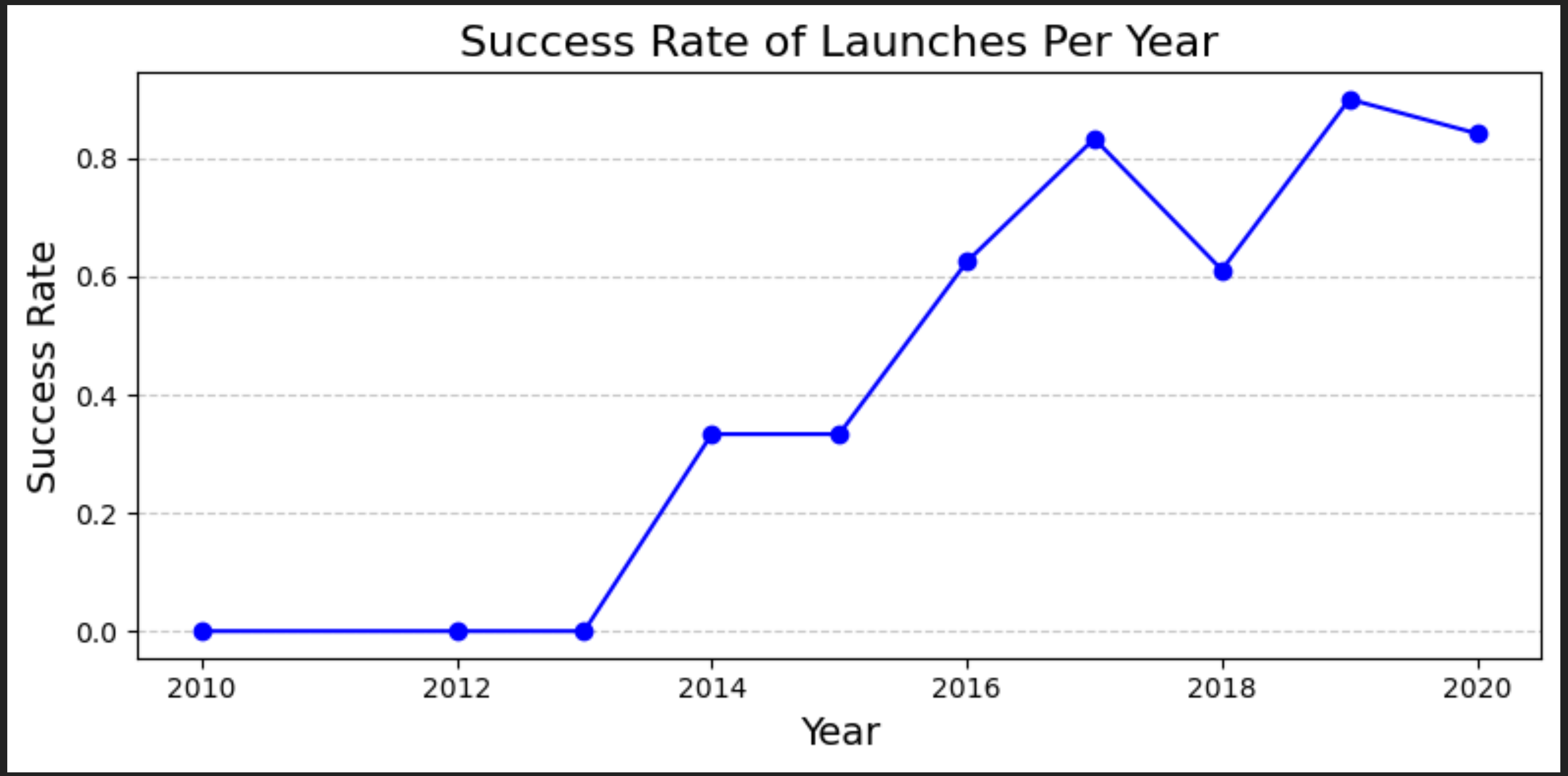
EDA WITH VISUALIZATION RESULTS (5)



ORBIT VS PAYLOAD MASS

- GTO tends to have a wide range of payload masses and shows both successful and unsuccessful launches.
- LEO and ISS also handle a variety of payload masses but have a notable number of unsuccessful launches at lower payload masses.
- VLEO and MEO seem to manage heavier payloads with a relatively high success rate.
- SSO, HEO GEO, and ES-L1 have experienced success with specific payload masses

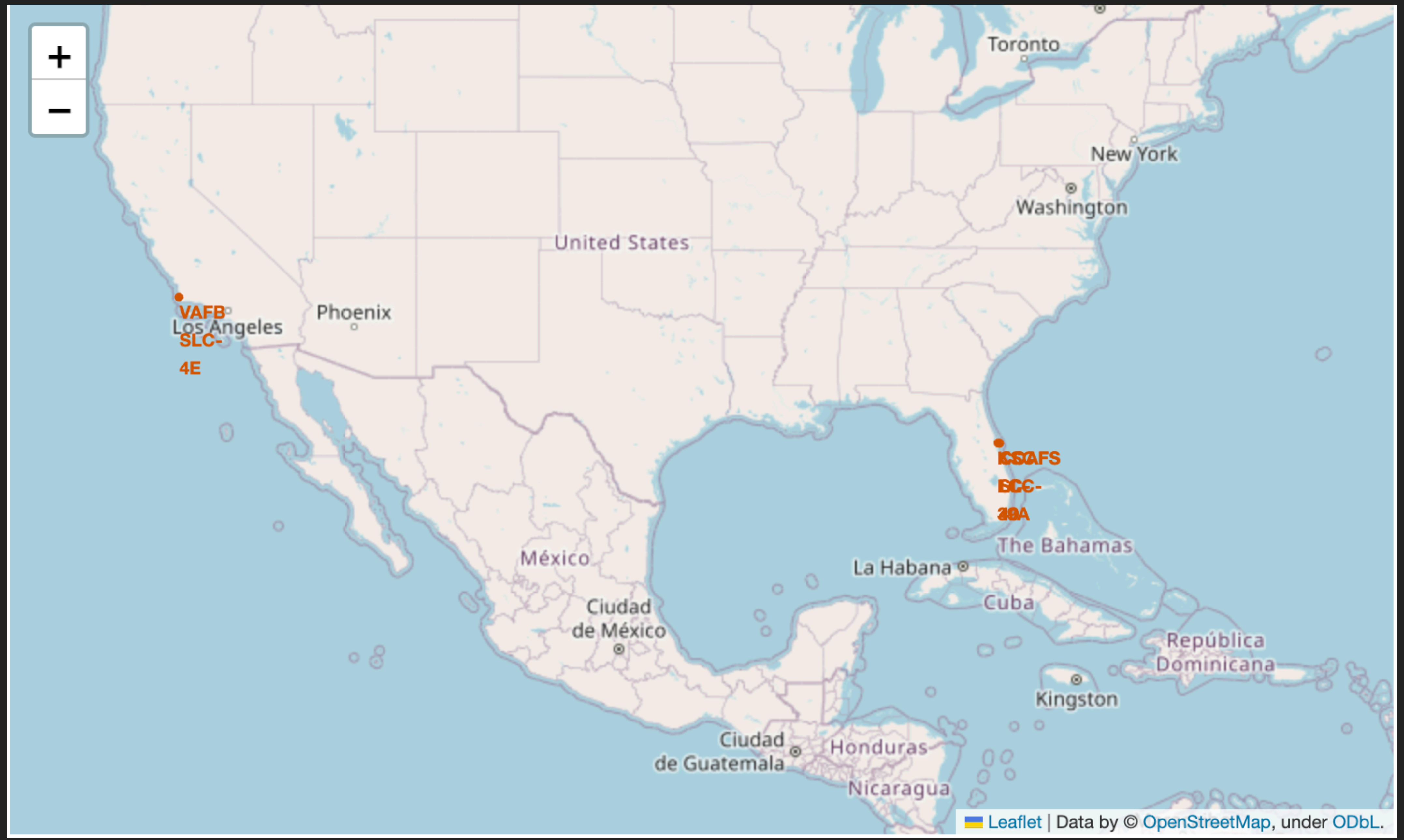
EDA WITH VISUALIZATION RESULTS (6)



AVERAGE SUCCESS RATE OVER TIME

- The early years (2010-2013) show a 0% success rate.
- A gradual increase in success rate is observed from 2014, peaking notably in 2019 with a 90% success rate.
- The years 2017, 2019, and 2020 demonstrate particularly high success rates, above 80%.

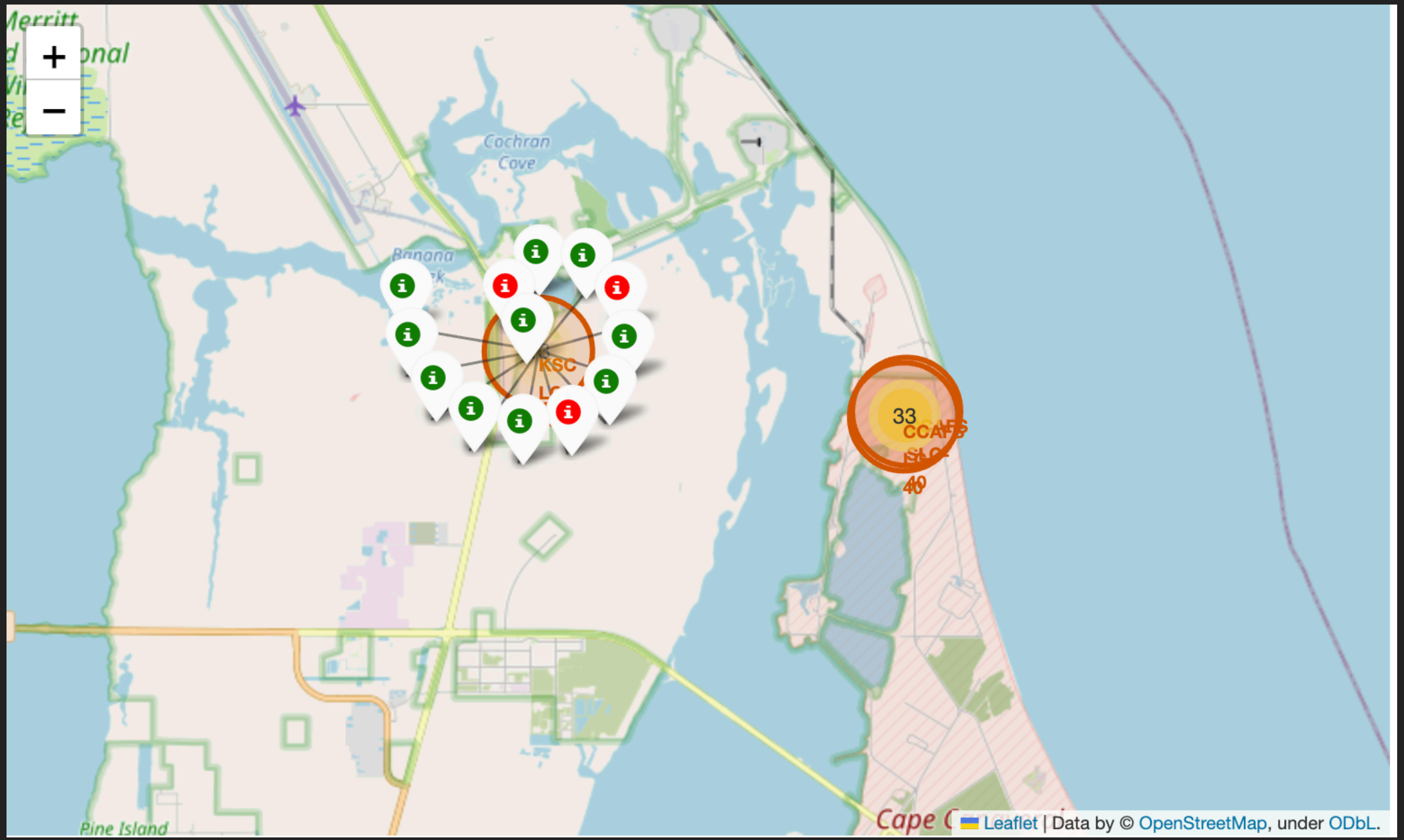
INTERACTIVE MAP WITH FOLIUM RESULTS (1)



- CCAFS SLC 40 and KSC LC 39A are relatively close to each other in terms of geographical location, both situated in Florida, USA.

- VAFB SLC 4E is located at a different longitude, being on the west coast of the USA in California.

INTERACTIVE MAP WITH FOLIUM RESULTS (2)

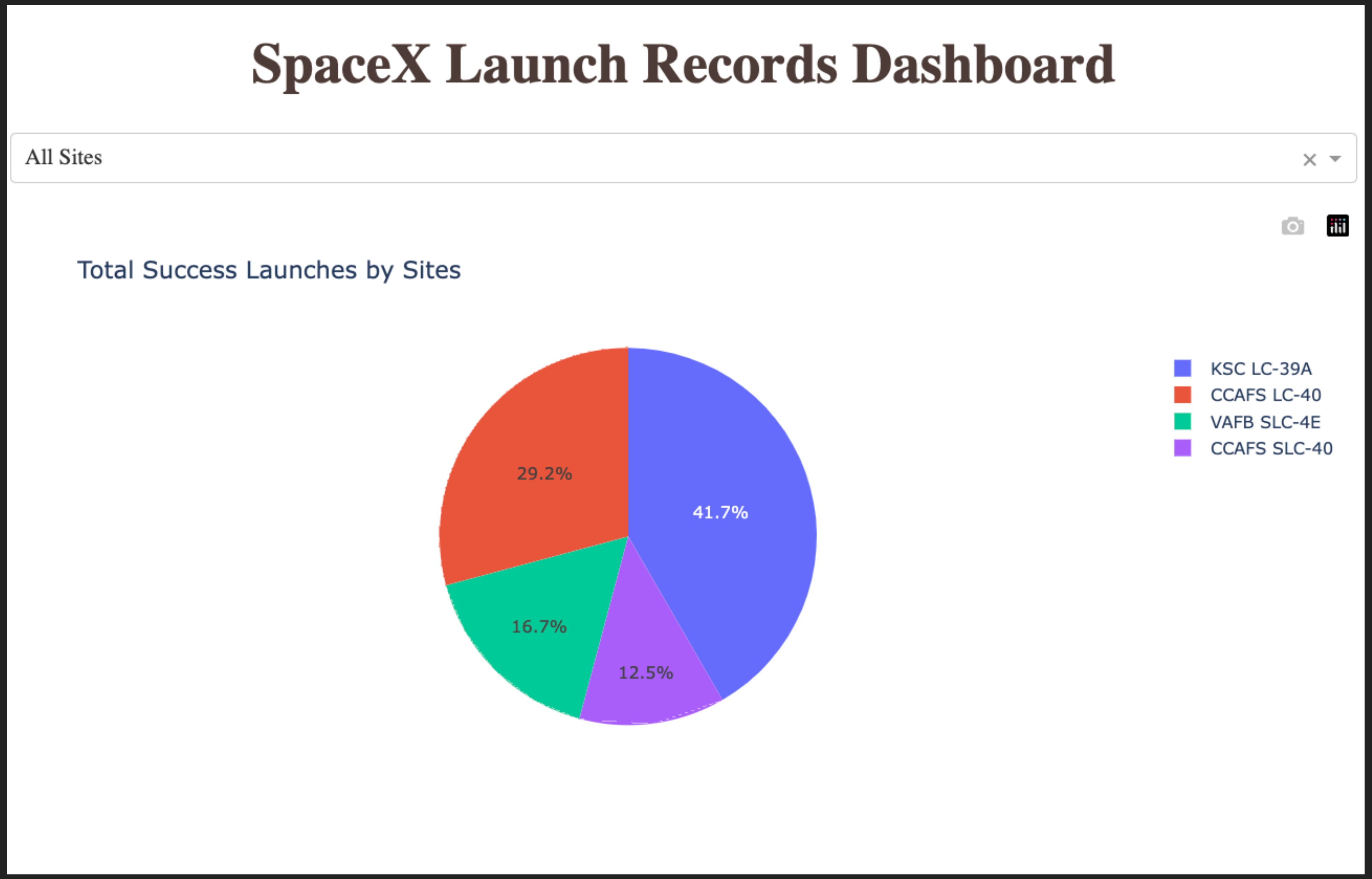


- The markers labeled with different colors in marker clusters help easily identify launch sites with high success rates.

  Green = Success, Red = Fail

- KSC LC 39A has the highest success rate among the launch sites

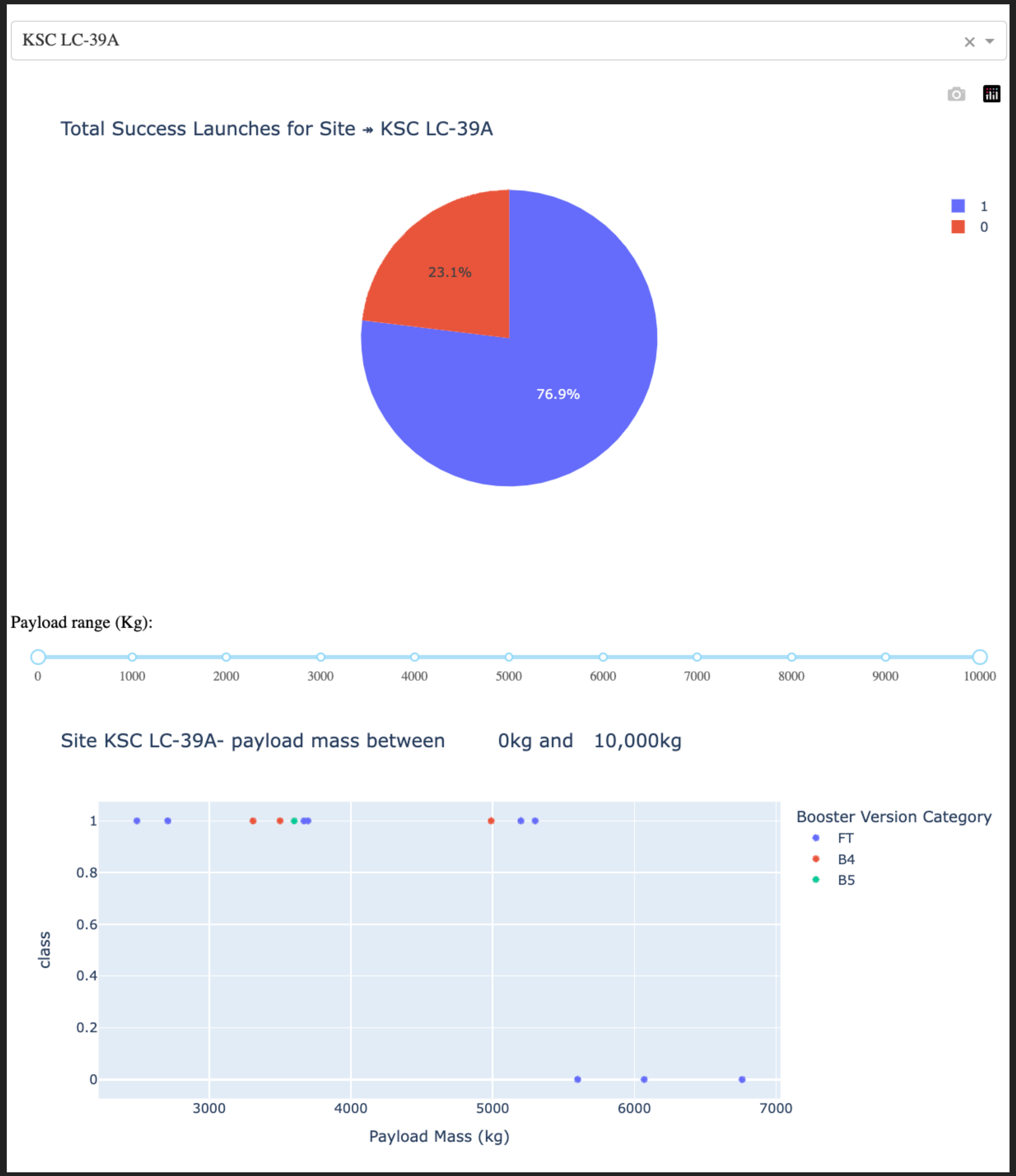
PLOTLY DASH DASHBOARD RESULTS (1)



SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Sites



KSC LC-39A

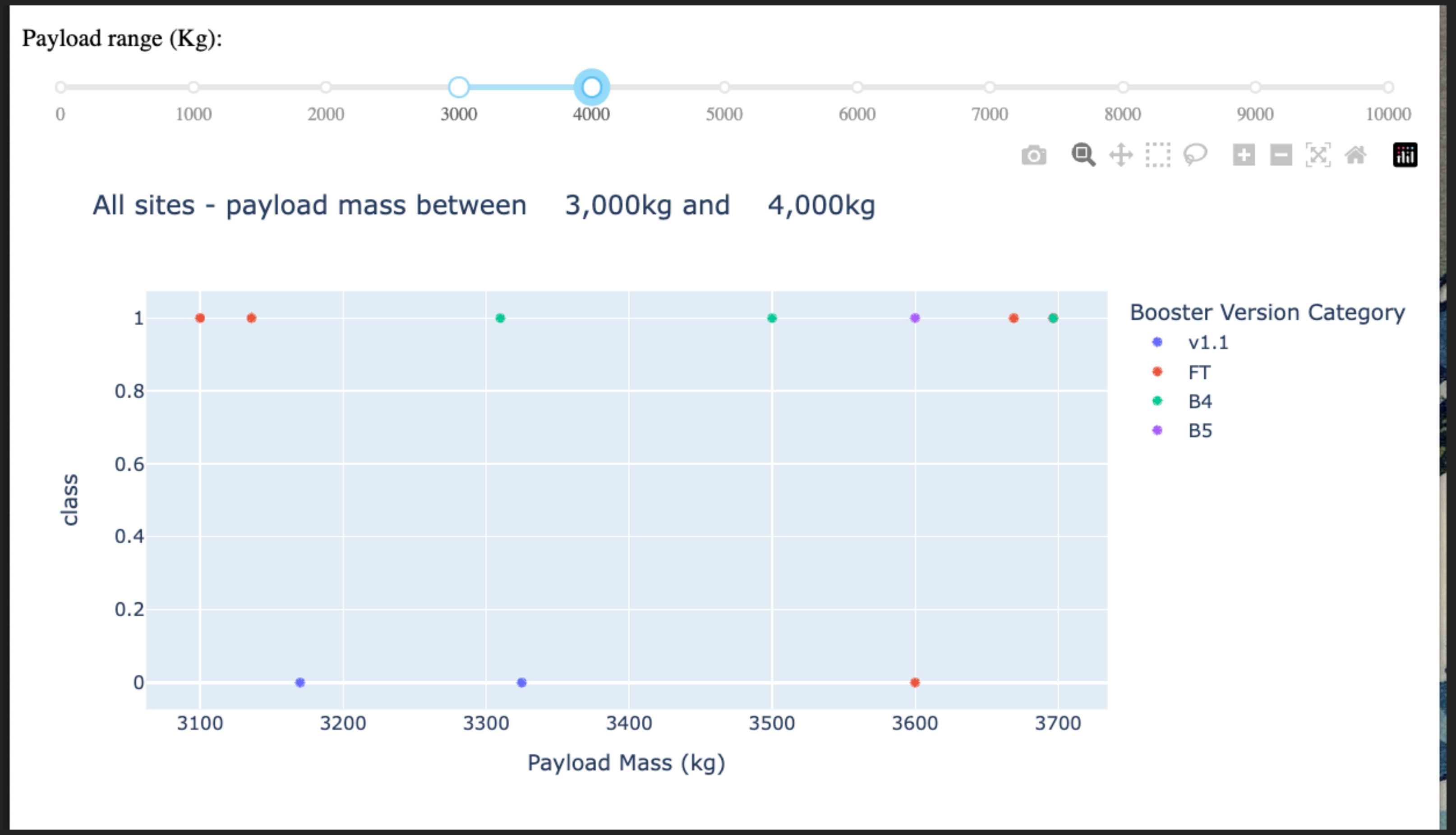
Total Success Launches for Site → KSC LC-39A

Payload range (Kg):

Site KSC LC-39A- payload mass between 0kg and 10,000kg

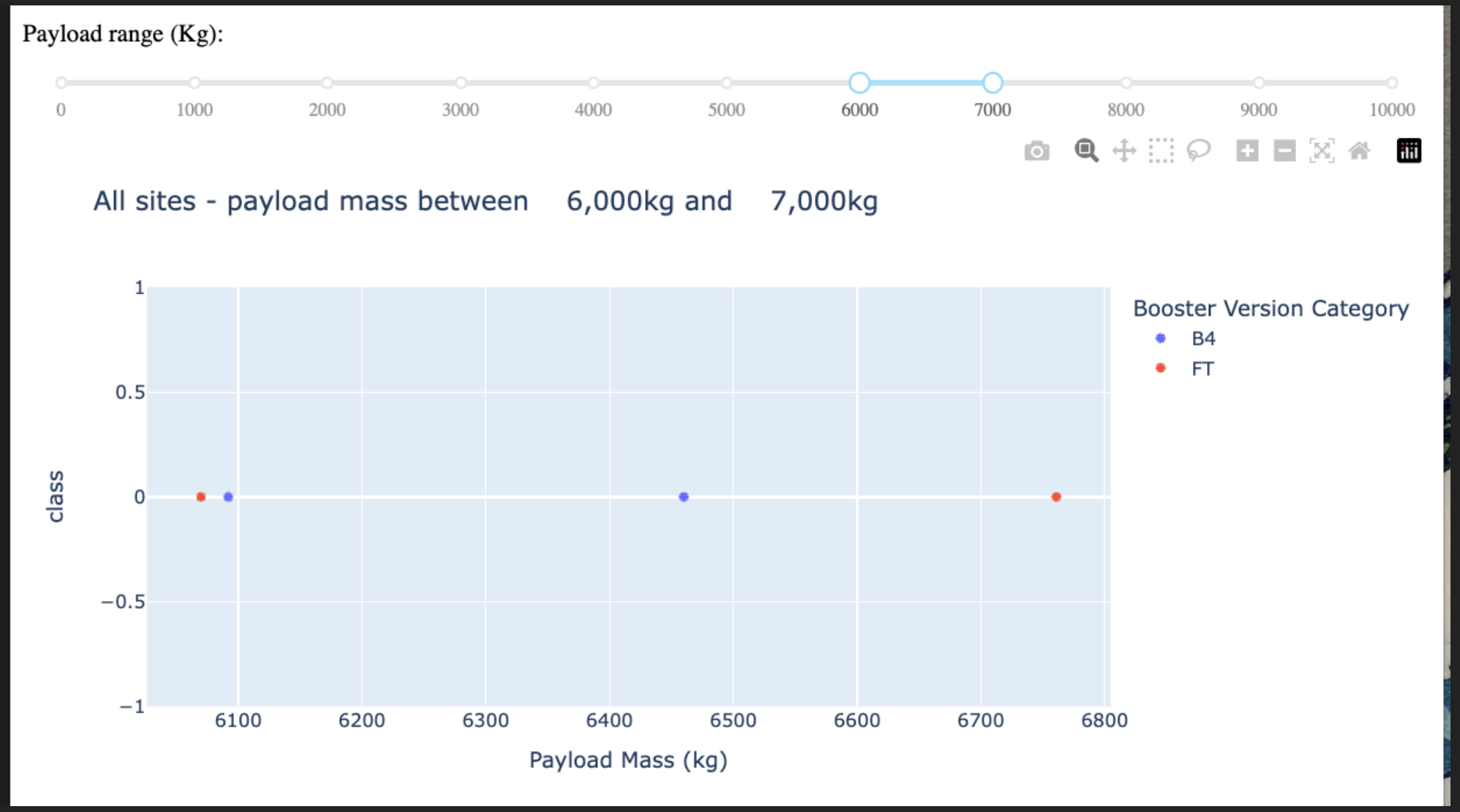
- The launch site with the highest launch success rate is KSC LC 39A

PLOTLY DASH DASHBOARD RESULTS (2)



Payload range (Kg):

All sites - payload mass between 3,000kg and 4,000kg

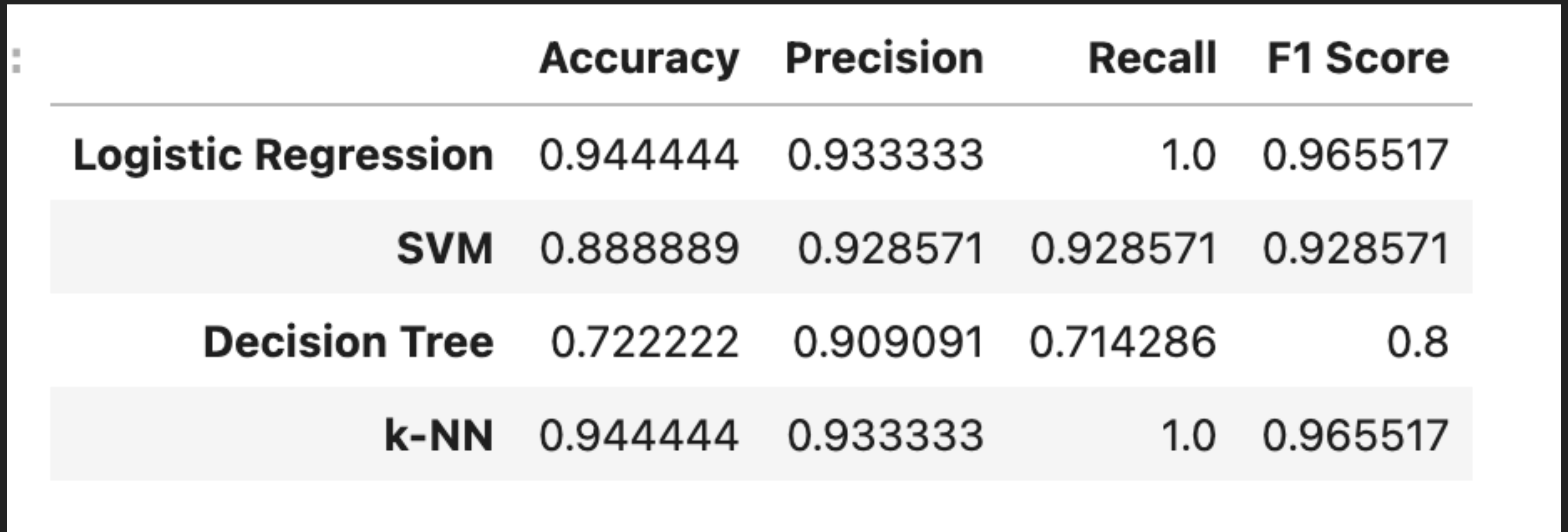


Payload range (Kg):

All sites - payload mass between 6,000kg and 7,000kg

- Payload Mass range with the Highest Launch Success Rate: 3000kg - 4000kg
- Payload Mass range with the Lowest Launch Success Rate: 6000kg - 7000kg

PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS (1)



	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.944444	0.933333	1.0	0.965517
SVM	0.888889	0.928571	0.928571	0.928571
Decision Tree	0.722222	0.909091	0.714286	0.8
k-NN	0.944444	0.933333	1.0	0.965517

- A classification model consisting of logistic regression, SVM, decision tree, and KNN was performed.
- Among the four models, logistic regression and KNN have the highest accuracy with the current train-test split.

PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS (2)

- While accuracy is a straightforward metric, it might not be the best metric
- Precision and Recall:
  - Precision helps us determine how many successful landings our model predicts out of all the instances where it predicts a landing will succeed
  - Recall refers to the number of successful landings our model correctly predicts. It is essential to consider the financial implications of false positives and false negatives, which are inaccuracies in predicting successful or failed landings.
  - Depending on the cost involved, we can optimize for either precision or recall.

CONCLUDING REMARKS

Key Takeaways

- Model Diversity: Employed a spectrum of machine learning models (Logistic Regression, Decision Tree, SVM, and KNN) to ascertain the most proficient predictor for Falcon 9 first-stage landing success.
- Metric-Oriented Evaluation: Leveraged metrics (Accuracy, Precision, Recall, F1 Score) to meticulously evaluate and compare the models, ensuring a holistic view of their predictive capabilities.

Insights & Implications

- Model Efficacy: Logistic Regression and KNN exhibited superior predictive performance, signaling it as a potential tool for future predictive endeavors.
- Data-Driven Planning: The incorporation of predictive analytics into SpaceX's operational planning could potentially amplify the success rates of future missions by providing foresight into landing outcomes.

RECOMMENDATIONS

- Model Optimization: Continuous fine-tuning and optimization of the selected model should be pursued to enhance its predictive accuracy over time and under varied scenarios.
- Feature Engineering: Delve deeper into feature engineering and exploration to potentially uncover additional predictors that could enhance the model’s predictive capabilities.
- Real-Time Adaptation: Explore possibilities of integrating real-time data into the model, facilitating dynamic predictions that adapt to real-world, real-time scenarios and variations.