IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

MRHIZOU Abdessamad
02/11/2023

Outline

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

Executive Summary

- Summary of methodologies

  - Data collection

  - Data wrangling

  - Exploratory Data Analysis with Data Visualization

  - Exploratory Data Analysis with SQL

  - Building an interactive map with Folium

  - Building a Dashboard with Plotly Dash

  - Predictive analysis (Classification)

- Summary of all results

  - Exploratory Data Analysis results

  - Interactive analytics

  - Predictive analysis results

Introduction

- **Project background and context**

SpaceX's cost-effective space travel is driven by reusing Falcon 9 rocket first stages, drastically reducing launch expenses. The objective is to predict first-stage landing success based on variables like payload mass, launch site, flights, and orbits. Evaluating Space Y's competitiveness is crucial.

- **Problems you want to find answers**

- How do payload mass, launch site, flight count, and orbits impact first stage landing success?

- Does the rate of successful landings increase over time?

- What's the optimal classification algorithm for predicting first stage landings?

- How can we estimate launch costs by factoring in first stage success?

- What's the ideal launch location?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data
  - Dealing with missing values
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluation of classification models to ensure the best results

Data Collection

- Data sets were collected from both the SpaceX REST API and Wikipedia using a combination of API requests and web scraping techniques to ensure a comprehensive dataset for detailed analysis.
- Data Sources:

SpaceX REST API: Data collected from SpaceX's official API

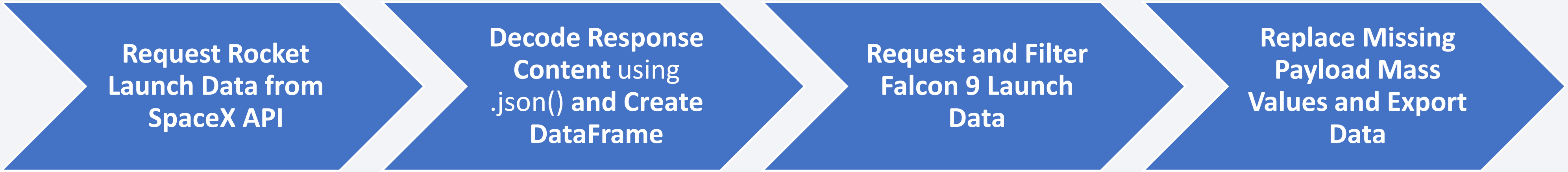
(https://api.spacexdata.com/v4/rockets/).

Wikipedia: Data scraped from SpaceX's Wikipedia page

(https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- SpaceX provides a publicly accessible API for retrieving data, which is subsequently utilized as per the aforementioned flowchart, and the obtained data is then stored or persisted.
- Bellow the flowchart representing data collection with SpaceX REST calls :

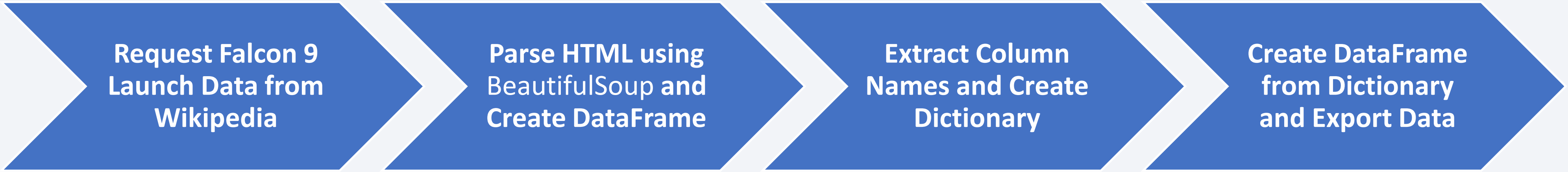

Request Rocket Launch Data from SpaceX API
Decode Response Content using .json() and Create DataFrame
Request and Filter Falcon 9 Launch Data
Replace Missing Payload Mass Values and Export Data

- GitHub URL of the completed SpaceX API calls notebook (Collecting the data):

https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied Data Science Capstone/Week1-Introduction/data-collection-api-spacex.ipynb

Data Collection – SpaceX API

- Data related to SpaceX launches can also be obtained from Wikipedia using an HTML parser like BeautifulSoup.
- Bellow the flowchart representing web scraping process :



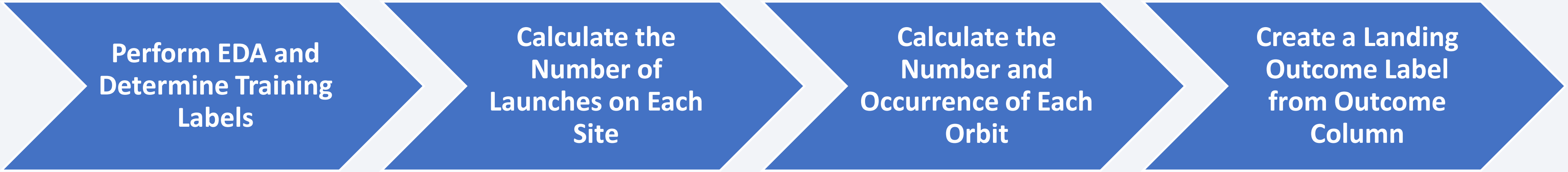
Request Falcon 9 Launch Data from Wikipedia
Parse HTML using BeautifulSoup and Create DataFrame
Extract Column Names and Create Dictionary
Create DataFrame from Dictionary and Export Data

- GitHub URL of the completed Web Scraping notebook (WebScraping):

https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied%20Data%20Science%20Capstone/Week1-Introduction/webscraping.ipynb

Data Wrangling

- The dataset contains various cases of unsuccessful booster landings, categorized as "True" or "False" for different scenarios, such as "Ocean," "RTLS," and "ASDS." These outcomes were simplified into training labels: "1" for success and "0" for failure.
- Initial Exploratory Data Analysis (EDA) was performed on the dataset.
- Launch counts per site, orbit type occurrences, and mission outcomes per orbit type were summarized.
- Landing outcome labels were derived from the "Outcome" column.



Perform EDA and Determine Training Labels

Calculate the Number of Launches on Each Site

Calculate the Number and Occurrence of Each Orbit

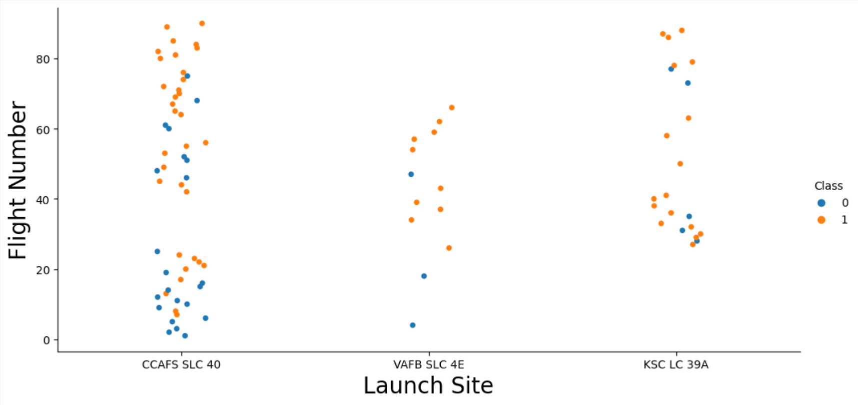
Create a Landing Outcome Label from Outcome Column

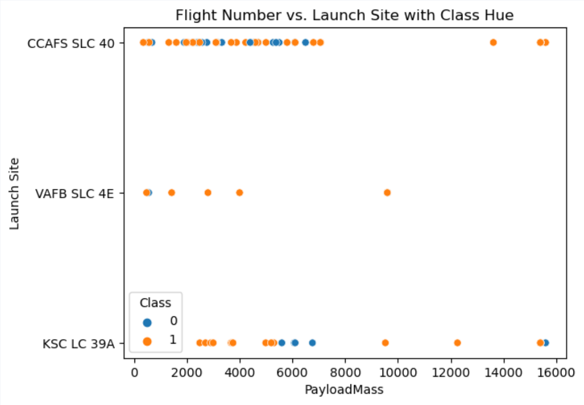
- GitHub URL of the completed data wrangling related notebooks (Data Wrangling):

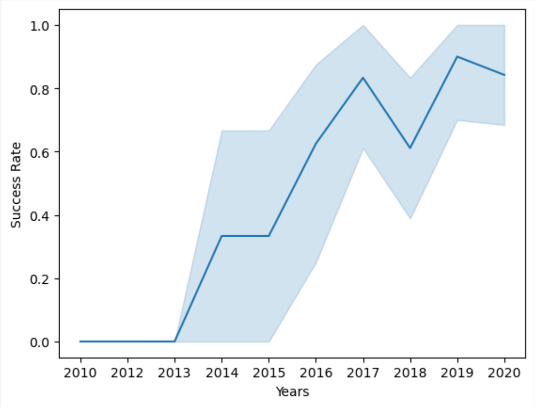
https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied%20Data%20Science%20Capstone/Week1-Introduction/Data%20wrangling-spacex.ipynb

EDA with Data Visualization

- We plotted various charts to provide insights into our data. These included Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Success Rate Yearly Trend. These charts were selected to visually explore relationships, trends, and comparisons within our dataset.




Flight Number vs. Launch Site with Class Hue



- GitHub URL of the completed EDA with data visualization notebook ([EDA visualization]()):

https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied%20Data%20Science%20Capstone/Week2-Exploratory_Data_Analysis/EDA_with_Visualization.ipynb

EDA with SQL

Here are the SQL queries formatted as bullet points:

- Names of the unique launch sites in the space mission.
- Top 5 launch sites whose names begin with the string 'CCA'.
- Total payload mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1.
- Date when the first successful landing outcome on a ground pad was achieved.
- Names of the boosters that have had success on a drone ship with payload mass between 4000 and 6000 kg.
- Total number of successful and failed mission outcomes.
- Names of the booster versions that have carried the maximum payload mass.
- Failed landing outcomes on a drone ship in 2015, along with their booster versions and launch site names.
- Rank of the count of landing outcomes (such as Failure on a drone ship or Success on a ground pad) between the dates 2010-06-04 and 2017-03-20.

GitHub URL of the completed EDA with sql notebook (EDA sql):

https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied%20 Data%20Science%20Capstone/Week2-Exploratory_Data_Analysis/Eda_sql_sqllite.ipynb

Build an Interactive Map with Folium

In creating the Folium map, various map objects were used to enhance the visualization of geographical data related to the space mission launch sites. These objects include:

- Markers: Used for key locations like launch sites and NASA Johnson Space Center.
- Circles: Highlight important areas and proximity around specific coordinates.
- Marker Clusters: Group events at a single coordinate, aiding visualization of event distribution.
- Lines: Show distances between coordinates, revealing spatial relationships and proximity.

The addition of these objects was aimed at improving the understanding of the space mission's launch sites and their geographical context.

- GitHub URL of the completed Map with Folium notebook (MAP Folium):

https://github.com/samadmrh/IBM_Data_Science/blob/bcd0ffee666bd32508487ca4e618ba9e53b8250d/Applied%20Data%20Science%20Capstone/Week3-Interactive_Visual_Analytics_and_Dashbord/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

In the dashboard, we incorporated the following for data visualization and interaction:

1. Percentage of Launches by Site: Shows site distribution.
2. Payload Range: Analyzes payloads and launch site relationships.

- Interactive features:

3. Launch Sites Dropdown: Select specific data.
4. Pie Chart for Success Launches: Displays success counts.
5. Payload Mass Range Slider: Filters data.
6. Scatter Chart for Payload vs. Success Rate: Visualizes payload-success correlation.

These elements enhance user engagement and data exploration for informed decision-making in space missions.

- GitHub URL of the completed Plotly Dash lab (Ploty Dash): (check pictures)

https://github.com/samadmrh/IBM_Data_Science/blob/96dc13dd70e60d2e3a50544f5bedf9aeec12a8c2/Applied%20Data%20Science%20Capstone/Week3-Interactive_Visual_Analytics_and_Dashbord/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb

Predictive Analysis (Classification)

- Model Comparison: Compared four classification models: logistic regression, support vector machine, decision tree, and k nearest neighbors.
- Data Preparation: Created a NumPy array from the "Class" column, standardized the data with StandardScaler, and split it into training and testing sets.
- Hyperparameter Tuning: Used GridSearchCV with cv = 10 to optimize model parameters.
- Model Evaluation: Calculated accuracy, examined confusion matrices, and assessed F1_score metrics to identify the best-performing model.
- This streamlined process systematically developed, evaluated, and improved classification models to determine the best performer.

Model Comparison

Data Preparation

Data Splitting

Hyperparameter Tuning

Model Evaluation

Model Evaluation

- GitHub URL of the completed Predictive Analysis lab (Classification): https://github.com/samadmrh/IBM_Data_Science/blob/96dc13dd70e60d2e3a50544f5bedf9aeec12a8c2/Applied%20Data%20Science%20Capstone/Week4-Predictive_Analysis_classification/_Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

Results

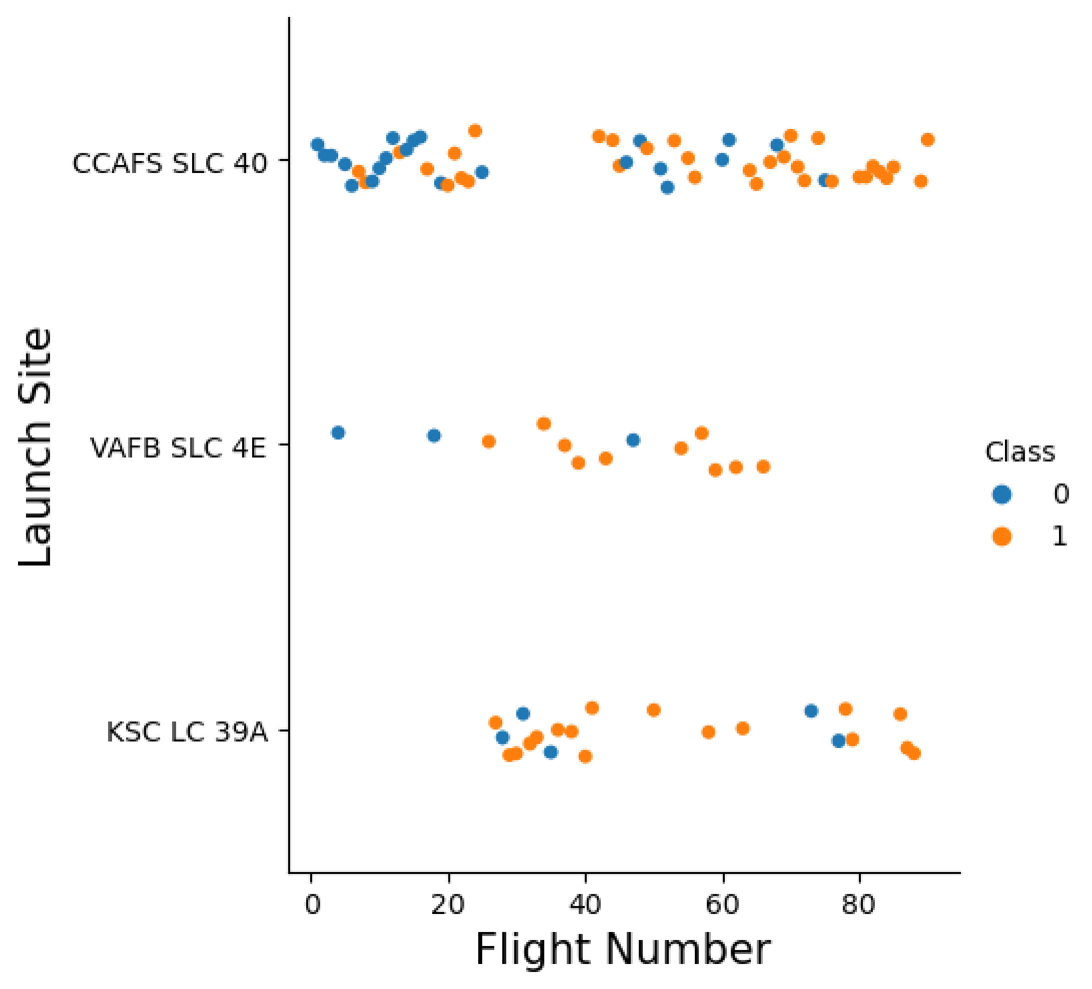
- Exploratory data analysis results
- Interactive analytics
- Predictive analysis results

Section 2

Insights drawn from EDA

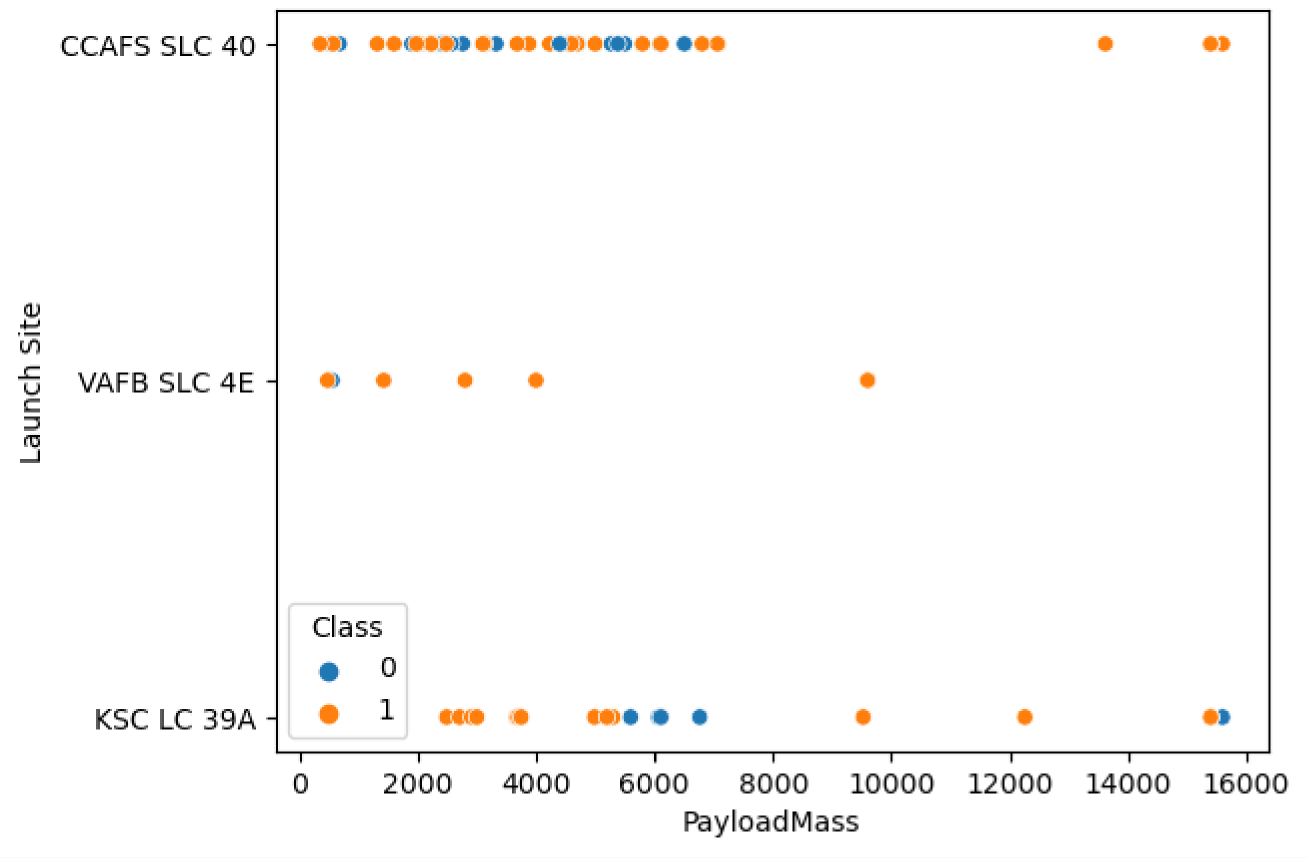
Flight Number vs. Launch Site

- The plot shows a trend of increasing success over time, with early flights failing and later flights succeeding.
- CCAFS SLC 40 is the most frequently used launch site, while VAFB SLC 4E and KSC LC 39A have higher success rates.
- Success rates appear to improve with each new launch, indicating ongoing advancements.
- Notably, CCAFS SLC 40 is the most successful recent launch site, followed by VAFB SLC 4E and KSC LC 39A.
- Overall, there is a noticeable improvement in the general success rate over time, reflecting SpaceX's evolving capabilities.



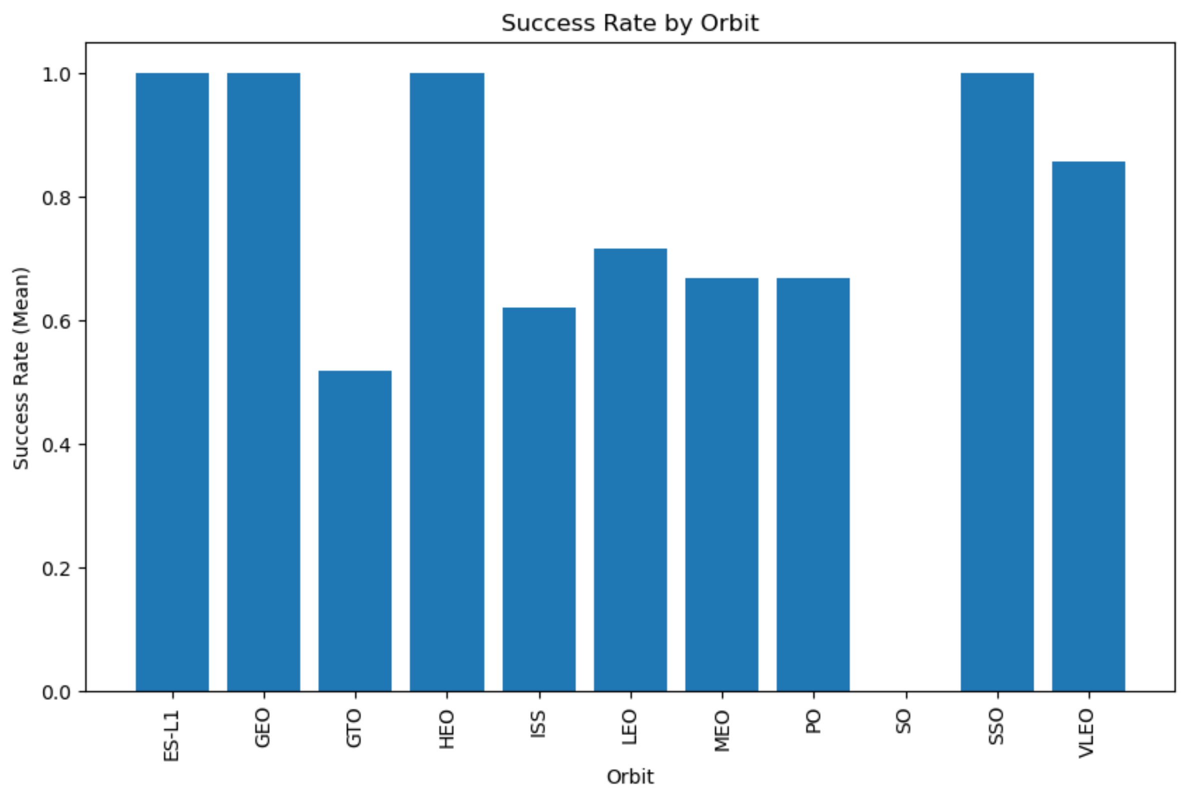
Payload vs. Launch Site

- Payloads over 9,000kg tend to have a high success rate.
- Payloads exceeding 12,000kg are feasible primarily at CCAFS SLC 40 and KSC LC 39A.
- Higher payload mass generally leads to a higher success rate across launch sites.
- Most launches with payload masses over 7,000 kg are successful.
- KSC LC 39A has a remarkable 100% success rate for payloads under 5,500 kg.



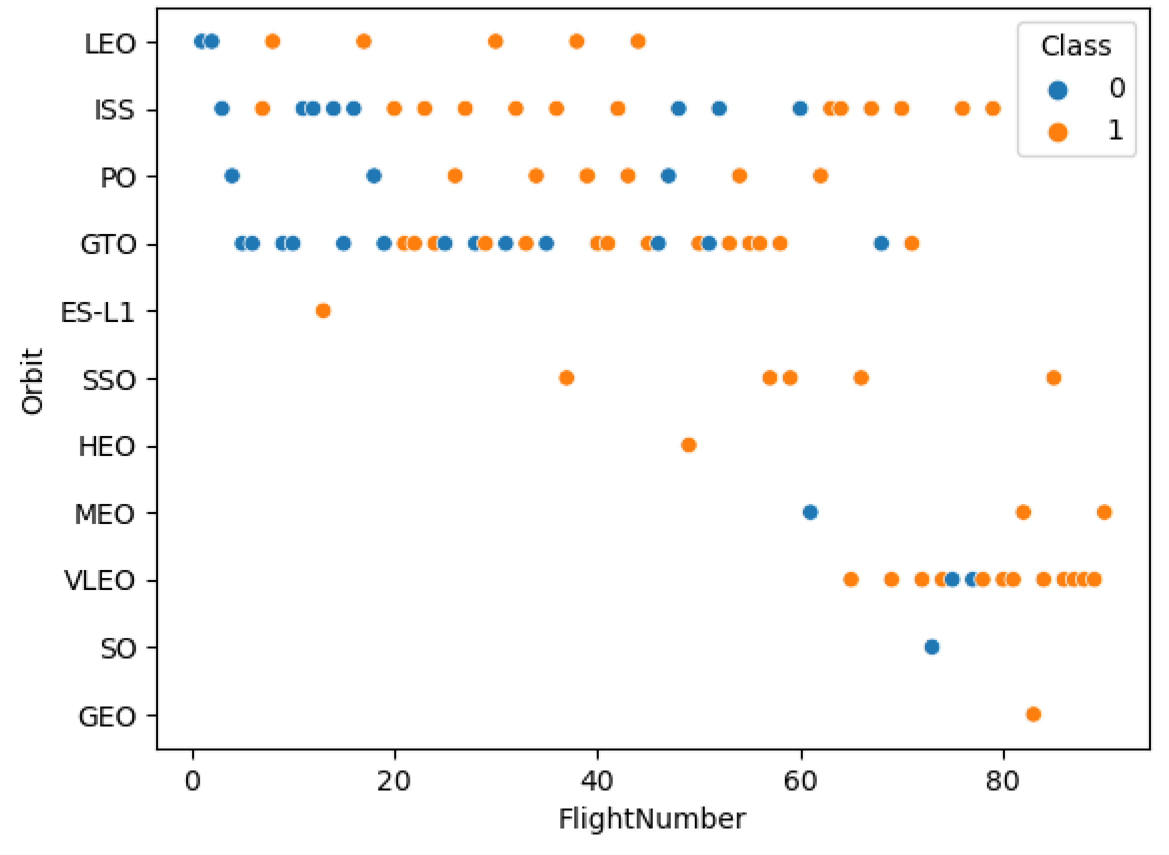
Success Rate vs. Orbit Type

- High Success Orbits (100%): Orbits like ES-L1, GEO, HEO, and SSO consistently achieve perfect success rates, signifying reliability.
- Low Success Orbit (0%): The "SO" orbit has a 0% success rate, indicating significant challenges for missions to this orbit.
- Moderate Success Orbits (50-85%): Orbits like GTO, ISS, LEO, MEO, and PO show varying success rates, likely influenced by specific factors related to each orbit.



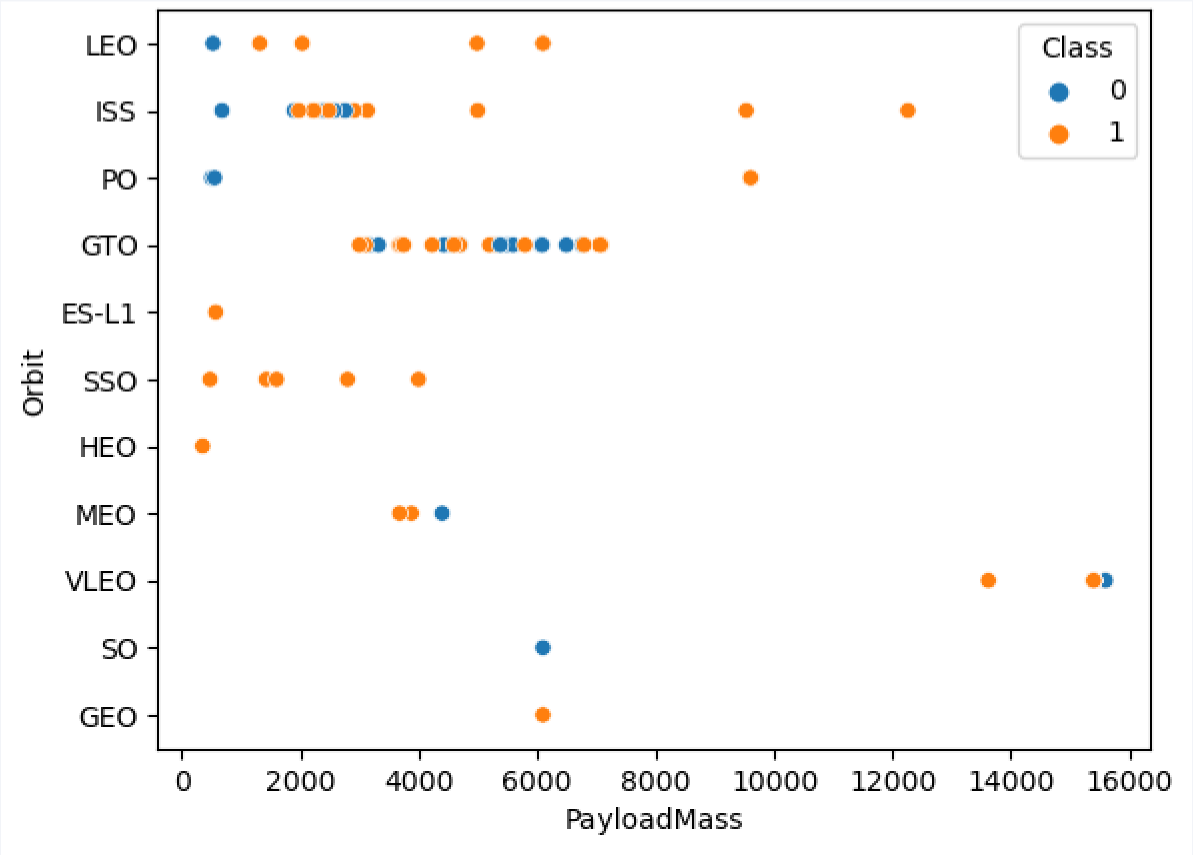
Flight Number vs. Orbit Type

In the LEO orbit, success appears to be influenced by the number of flights, whereas in the GTO orbit, flight number doesn't seem to have a significant impact on success.



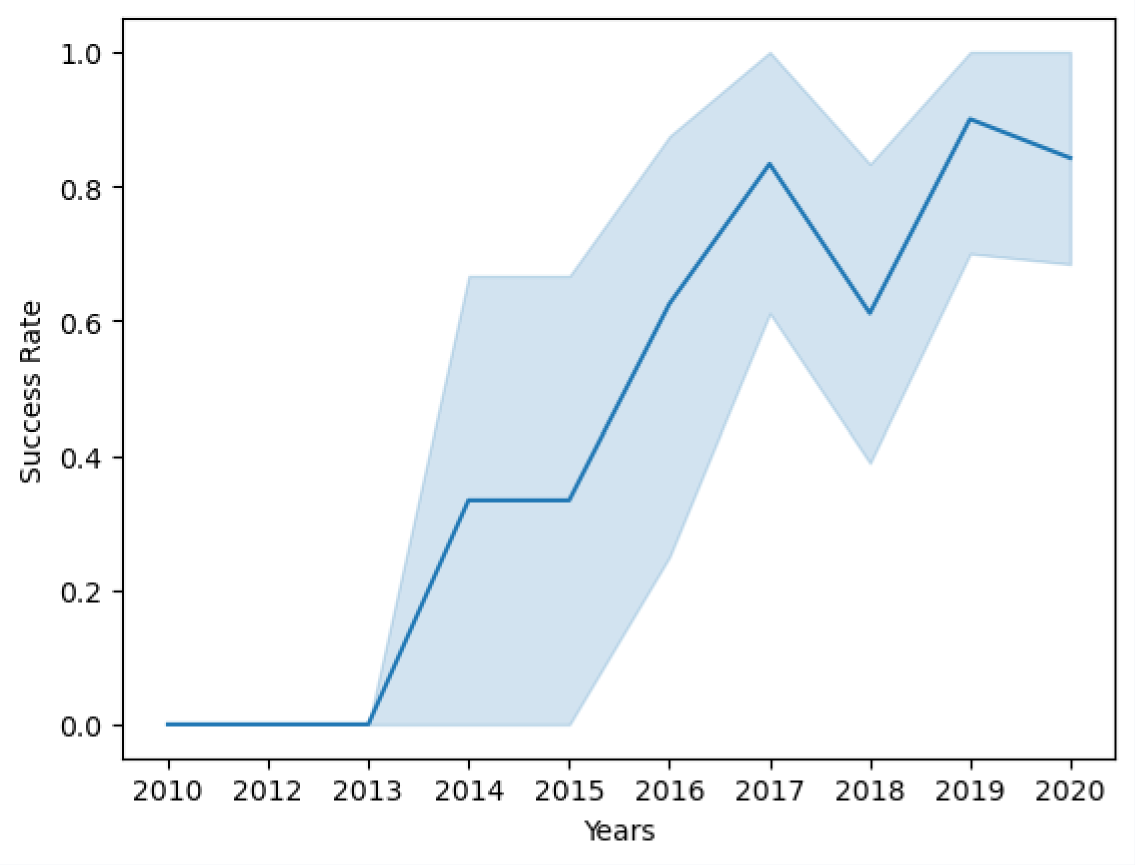
Payload vs. Orbit Type

- There seems to be no clear correlation between payload and success rate for the GTO orbit, while the ISS orbit exhibits a broad range of payload variation with a favorable success rate, and there are only a limited number of launches to the SO and GEO orbits.

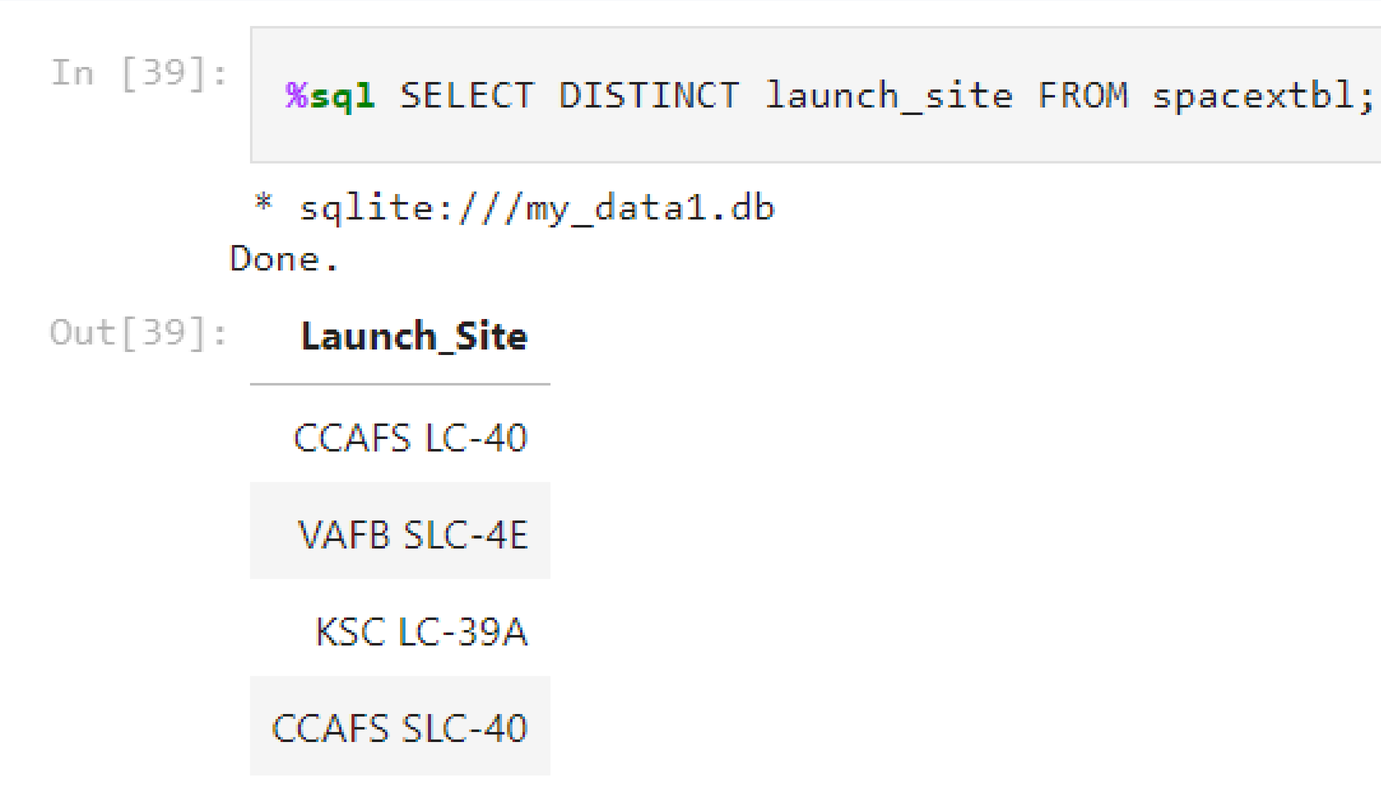


Launch Success Yearly Trend

- The success rate since 2013 kept increasing till 2020



All Launch Site Names



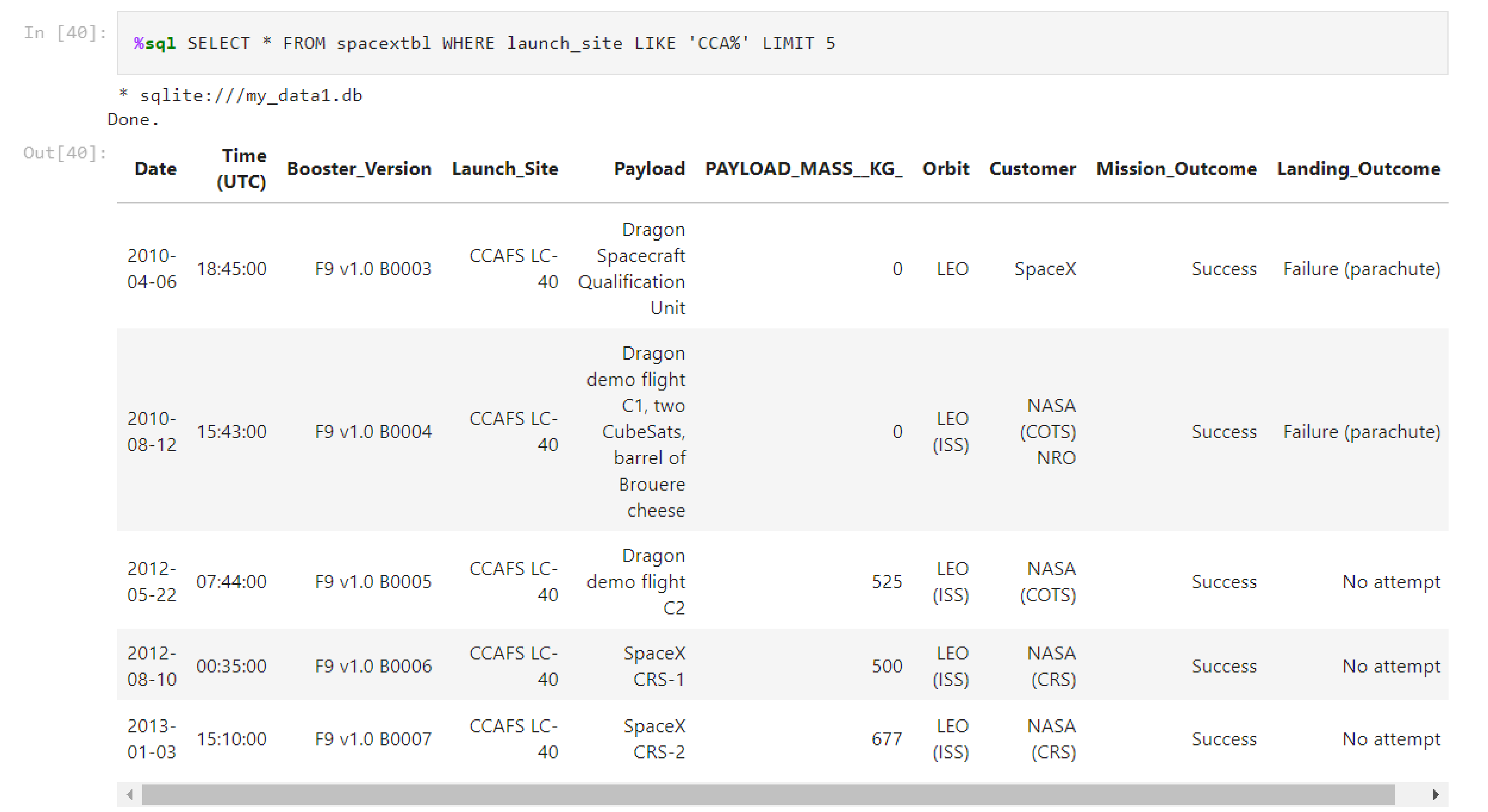
```
In [39]: %sql SELECT DISTINCT launch_site FROM spacextbl;

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

Out[39]:

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

Launch Site Names Begin with 'CCA'



In [40]:

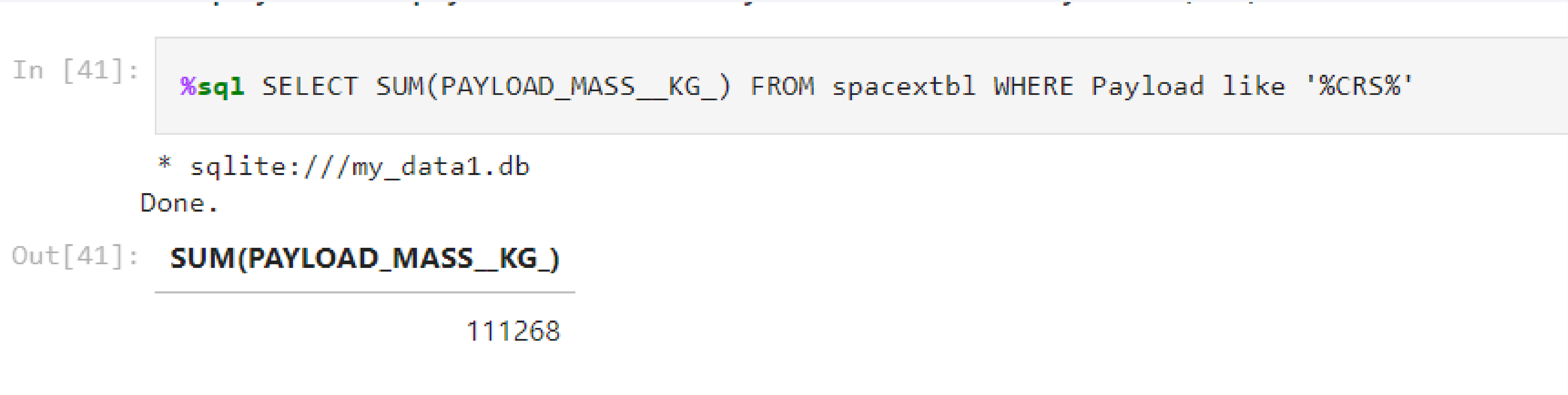
```
%sql SELECT * FROM spacextbl WHERE launch_site LIKE 'CCA%' LIMIT 5
```

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

Out[40]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass



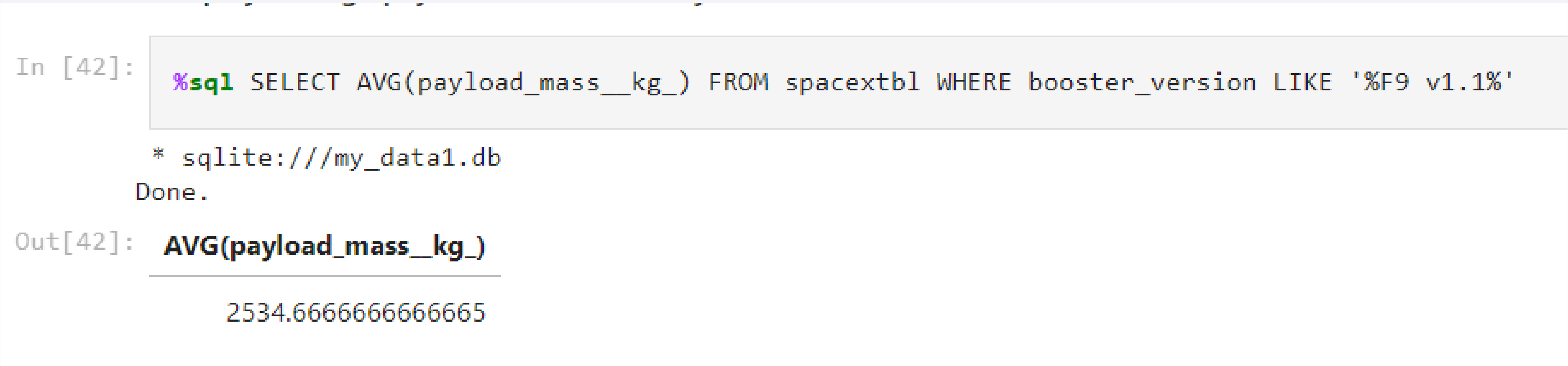
```
In [41]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM spacextbl WHERE Payload like '%CRS%'

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

Out[41]:

SUM(PAYLOAD_MASS__KG_)
111268

Average Payload Mass by F9 v1.1

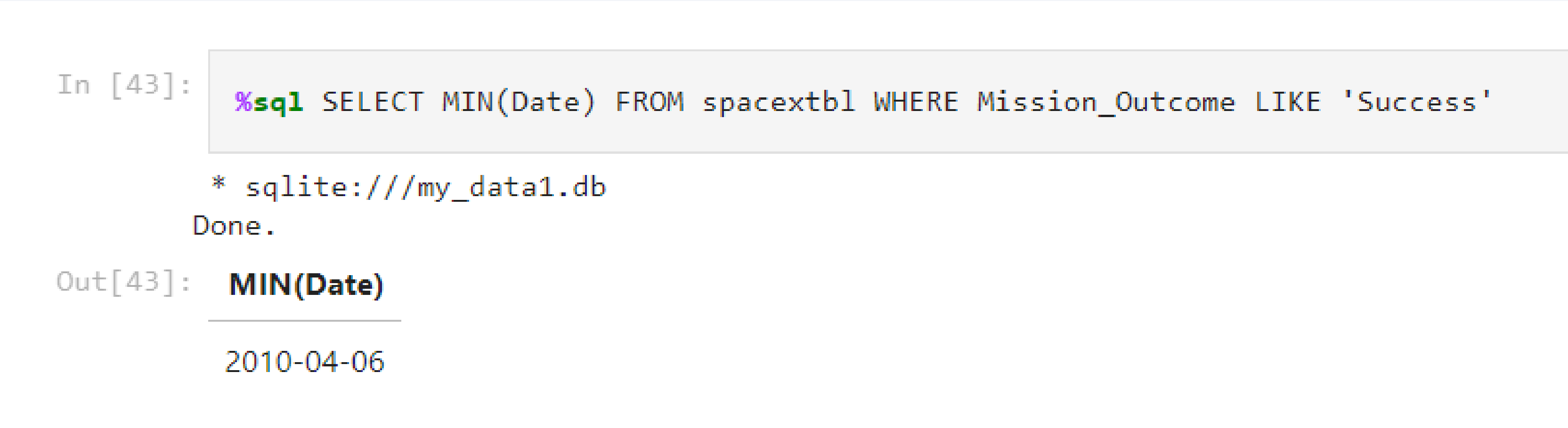


```
In [42]: %sql SELECT AVG(payload_mass__kg_) FROM spacextbl WHERE booster_version LIKE '%F9 v1.1%'

 * sqlite:///my_data1.db
Done.

Out[42]: AVG(payload_mass__kg_)
         2534.6666666666665
```

First Successful Ground Landing Date

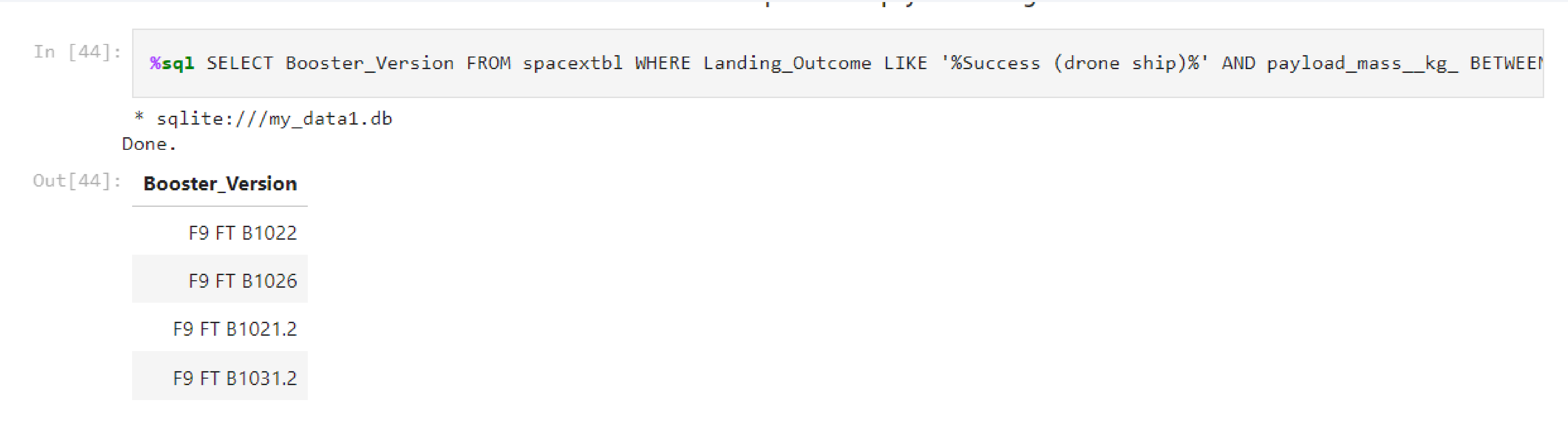


```
%sql SELECT MIN(Date) FROM spacextbl WHERE Mission_Outcome LIKE 'Success'
```

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

MIN(Date)
2010-04-06

Successful Drone Ship Landing with Payload between 4000 and 6000



In [44]:

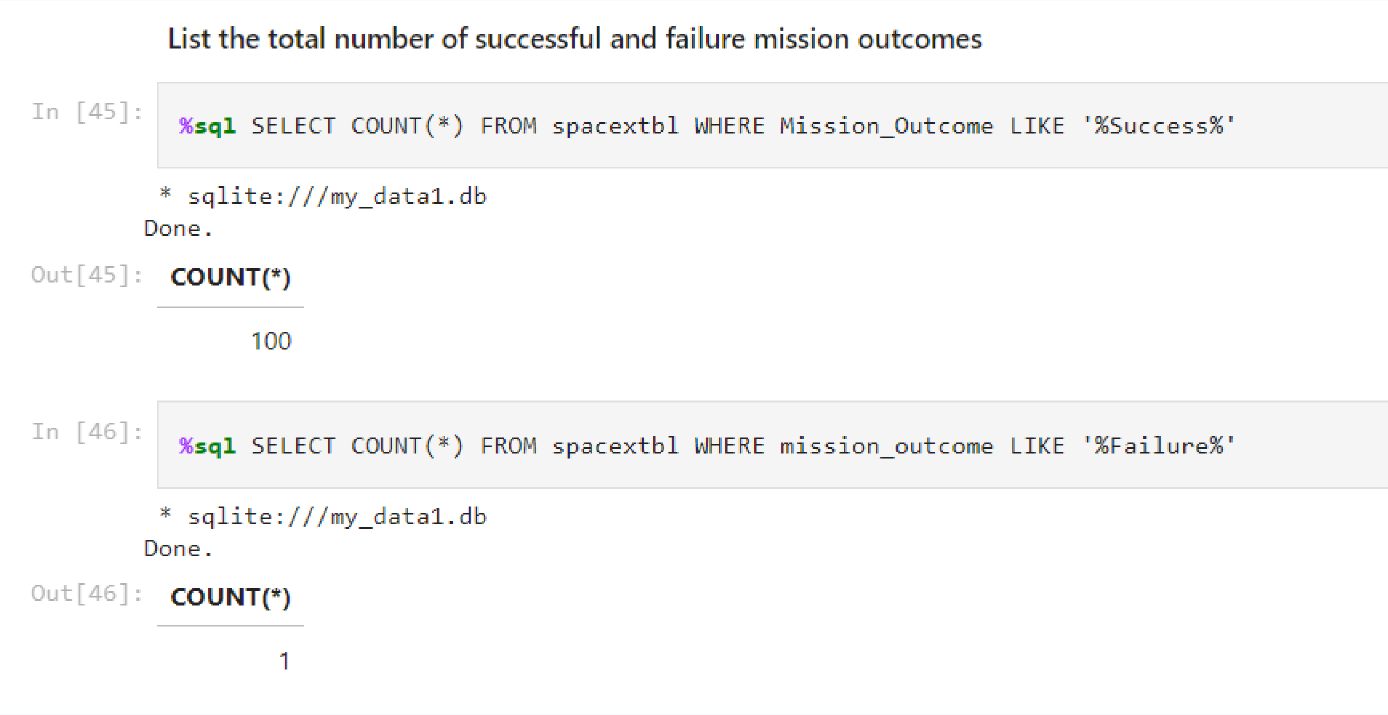
```
%sql SELECT Booster_Version FROM spacextbl WHERE Landing_Outcome LIKE '%Success (drone ship)%' AND payload_mass__kg_ BETWEEN
```

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

Out[44]:

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

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

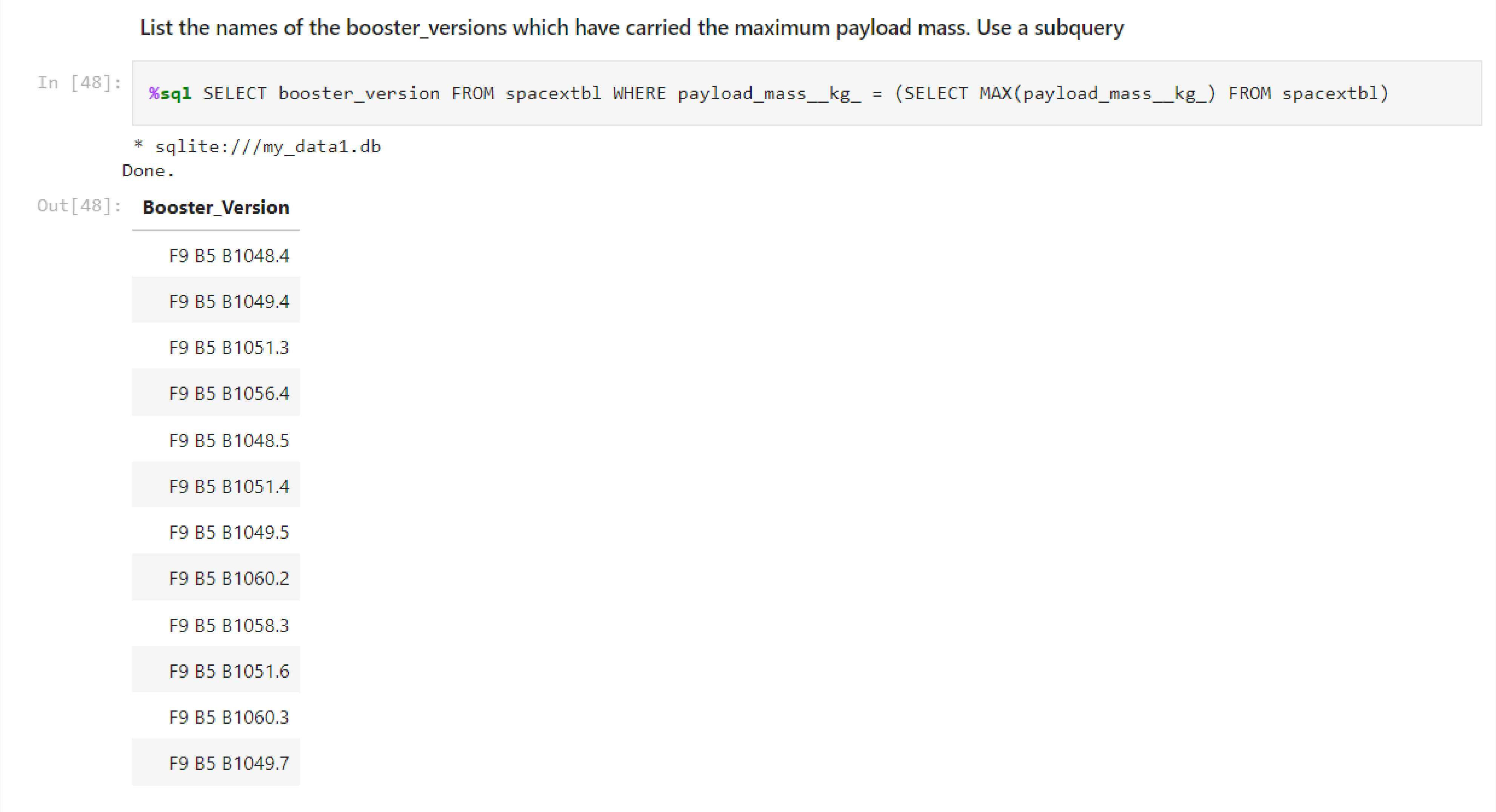
```
In [45]: %sql SELECT COUNT(*) FROM spacextbl WHERE Mission_Outcome LIKE '%Success%'

 * sqlite:///my_data1.db
Done.
Out[45]: COUNT(*)
              100
```

```
In [46]: %sql SELECT COUNT(*) FROM spacextbl WHERE mission_outcome LIKE '%Failure%'

 * sqlite:///my_data1.db
Done.
Out[46]: COUNT(*)
                1
```

Boosters Carried Maximum Payload



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

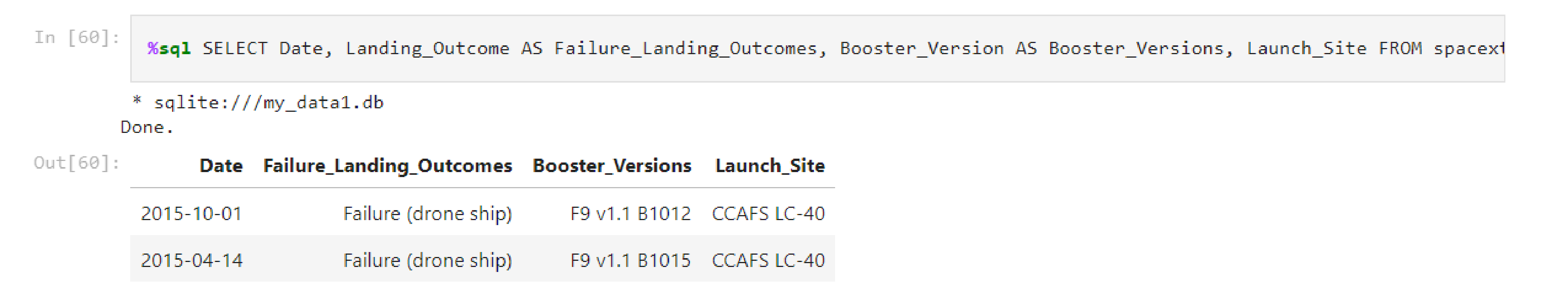
```
In [48]: %sql SELECT booster_version FROM spacextbl WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM spacextbl)
```

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

Out[48]:

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



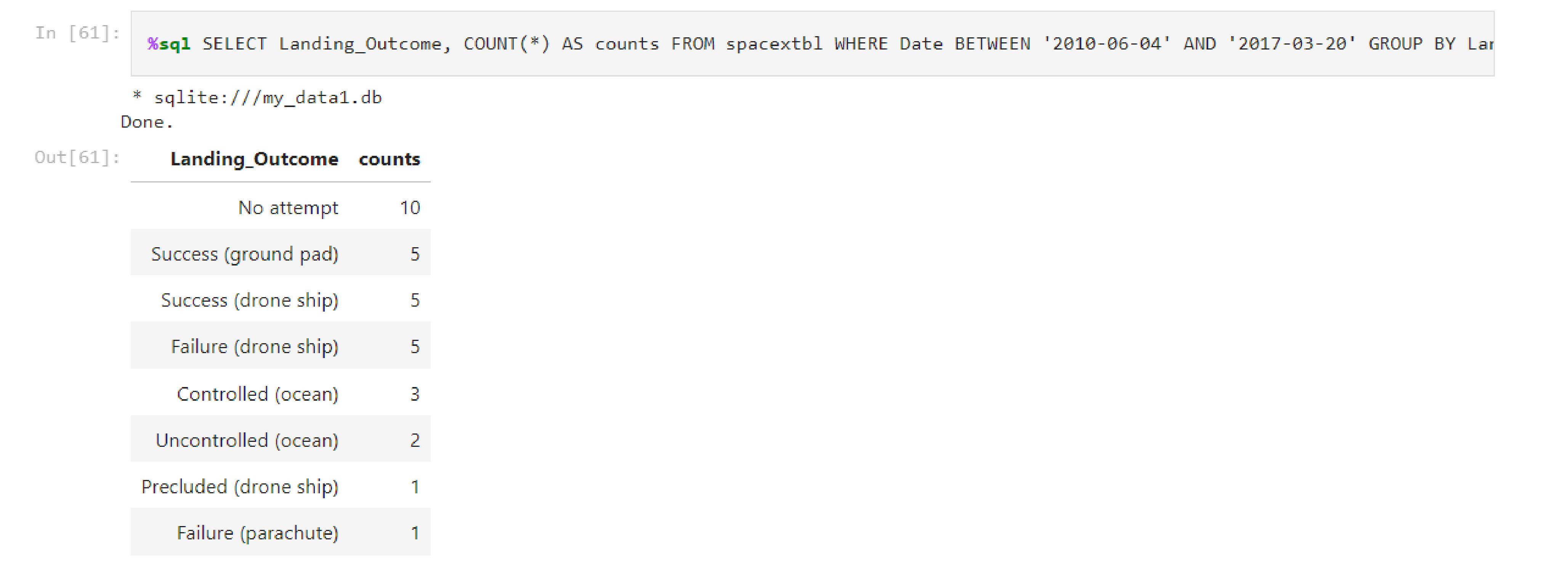
```
In [60]: %sql SELECT Date, Landing_Outcome AS Failure_Landing_Outcomes, Booster_Version AS Booster_Versions, Launch_Site FROM spacext
```

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

Out[60]:

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

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



```
In [61]: %sql SELECT Landing_Outcome, COUNT(*) AS counts FROM spacextbl WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Lar
```

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

Out[61]:

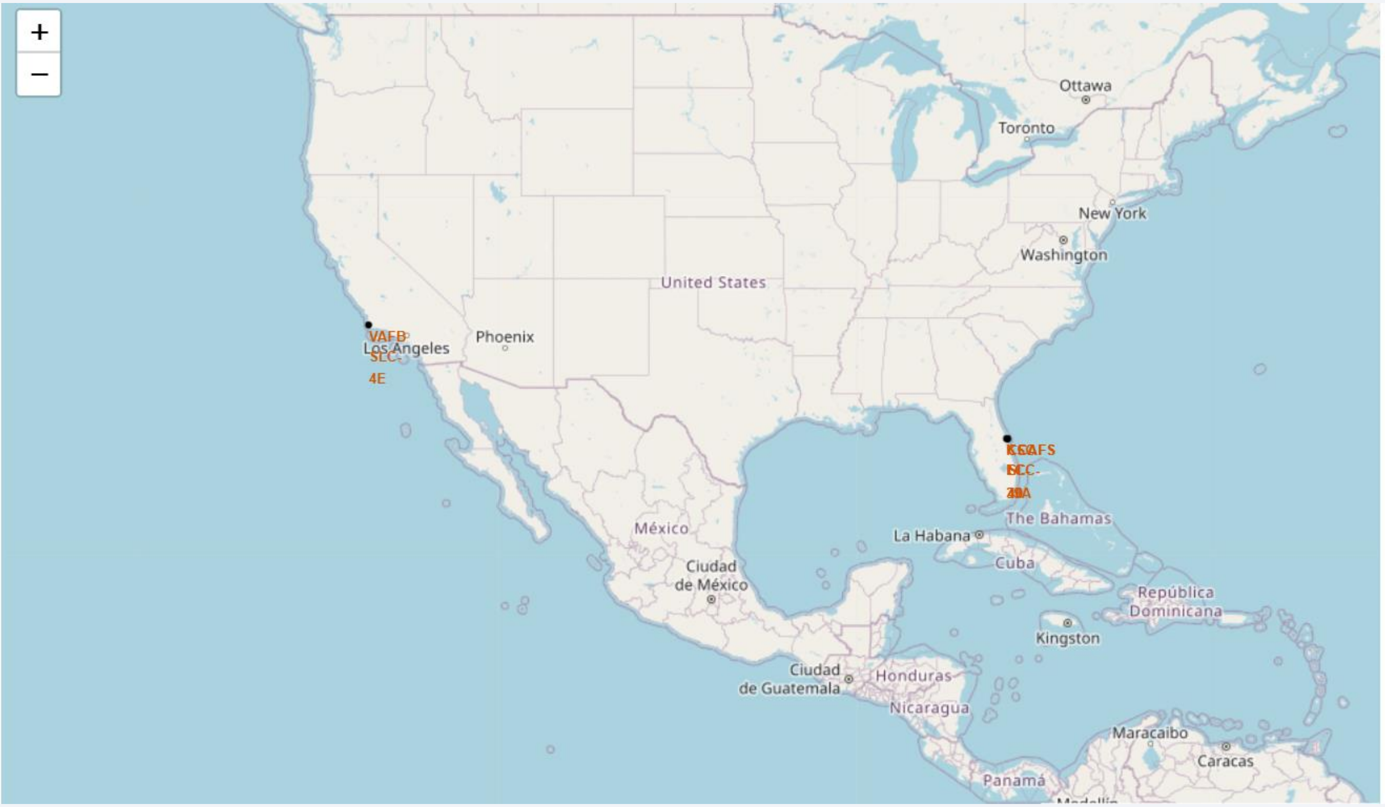
Landing_Outcome	counts
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

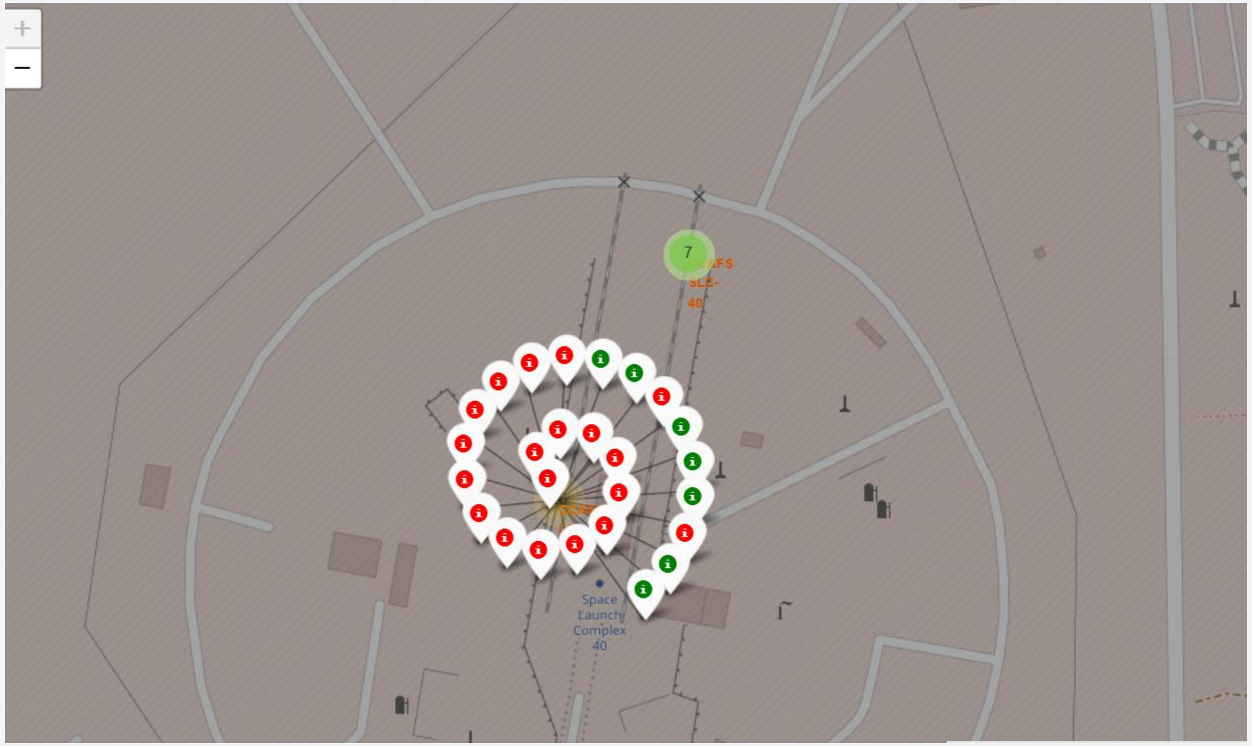
All launch sites

Most launch sites are located near the Equator, where the Earth's surface moves at a high-speed of 1670 km/hour due to its natural rotation. When a spacecraft is launched from the Equator, it retains this speed thanks to inertia, which helps it maintain the necessary velocity for staying in orbit. Additionally, all launch sites are situated close to coastlines to minimize the risk of debris falling or exploding near populated areas.



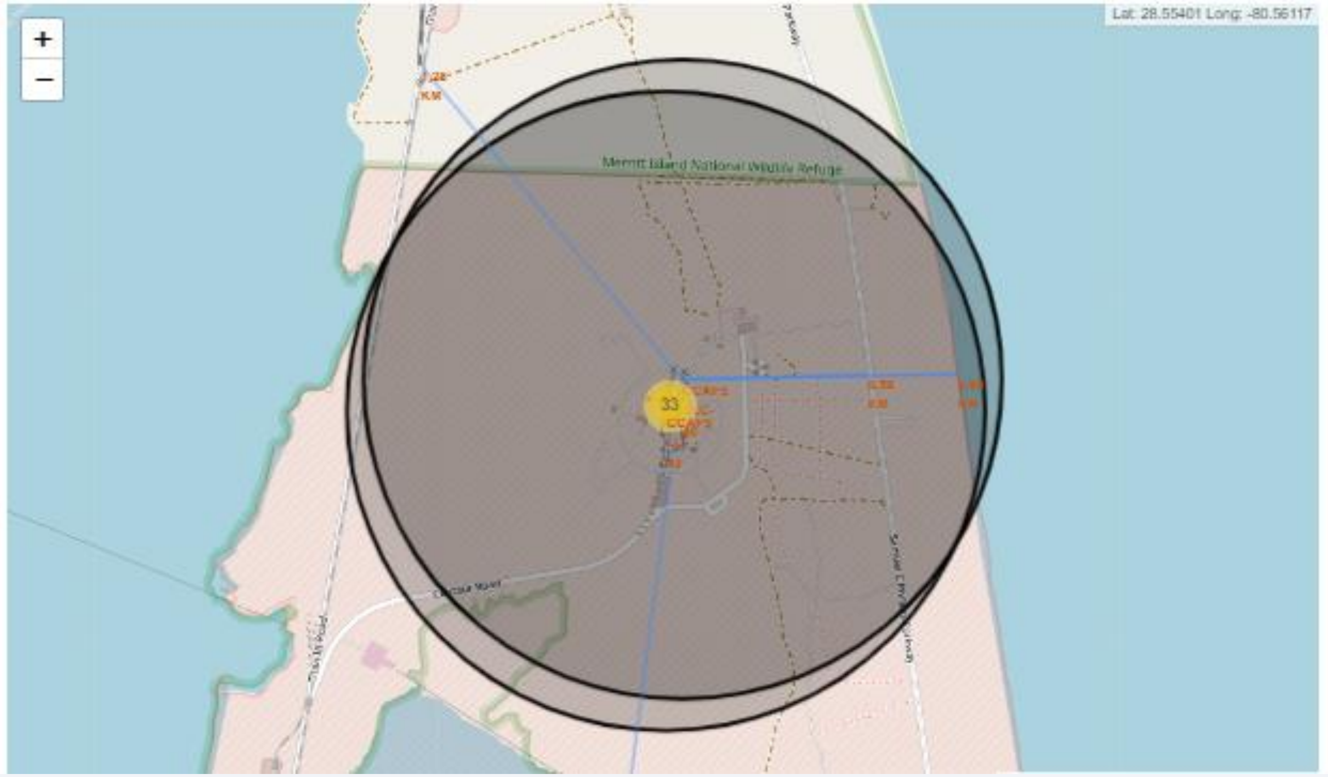
Map with Color-Labeled Launch Records

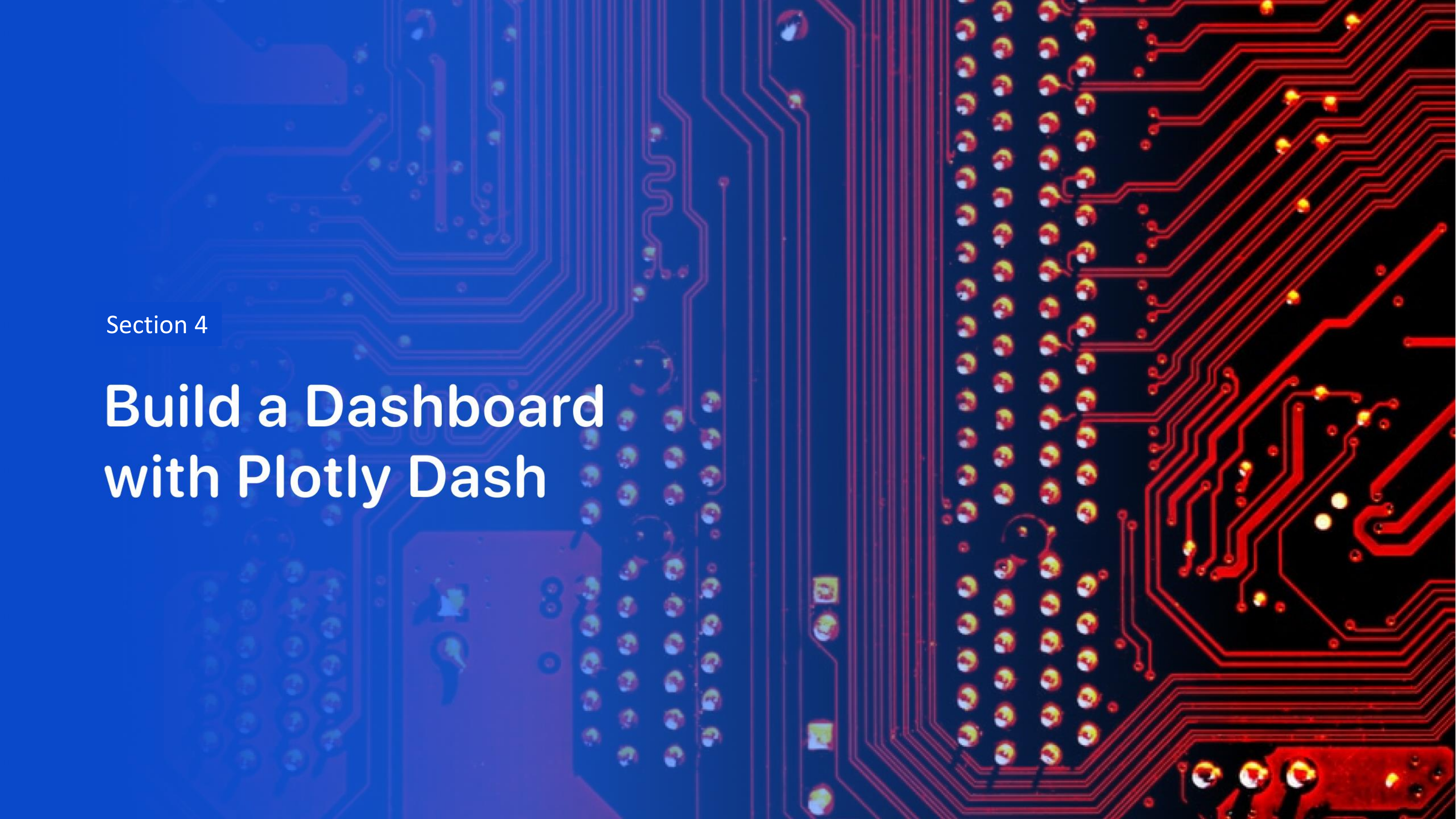
- The color-coded markers simplify the identification of launch sites with varying success rates:
- Green markers indicate successful launches.
- Red markers represent failed launches.



Logistics

Launch site with good logistics aspects, being near railroad and road and relatively far from inhabited areas.

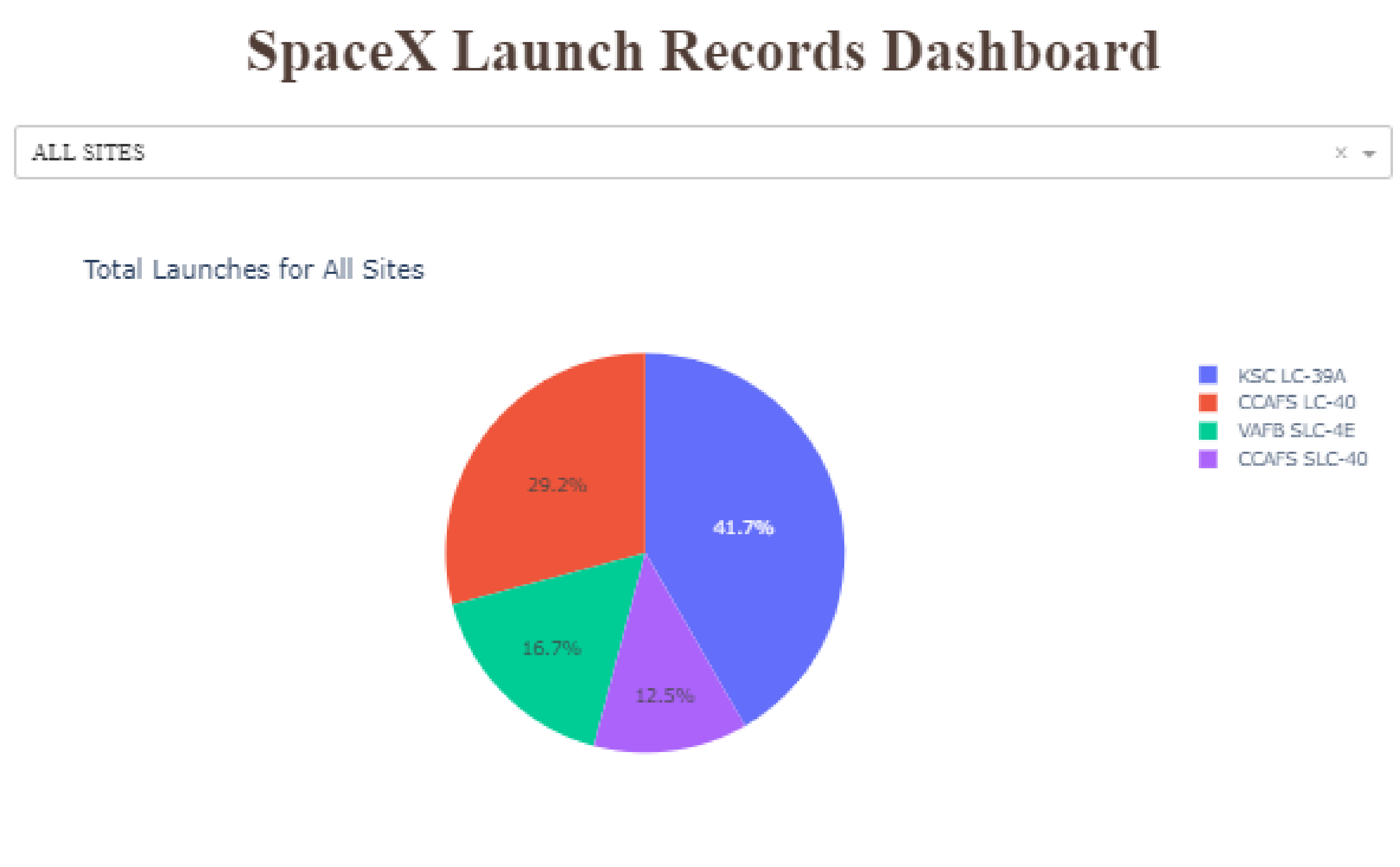

Lat: 28.55401 Long: -80.56117
+
−
Merritt Island National Wildlife Refuge
33



Section 4

Build a Dashboard with Plotly Dash

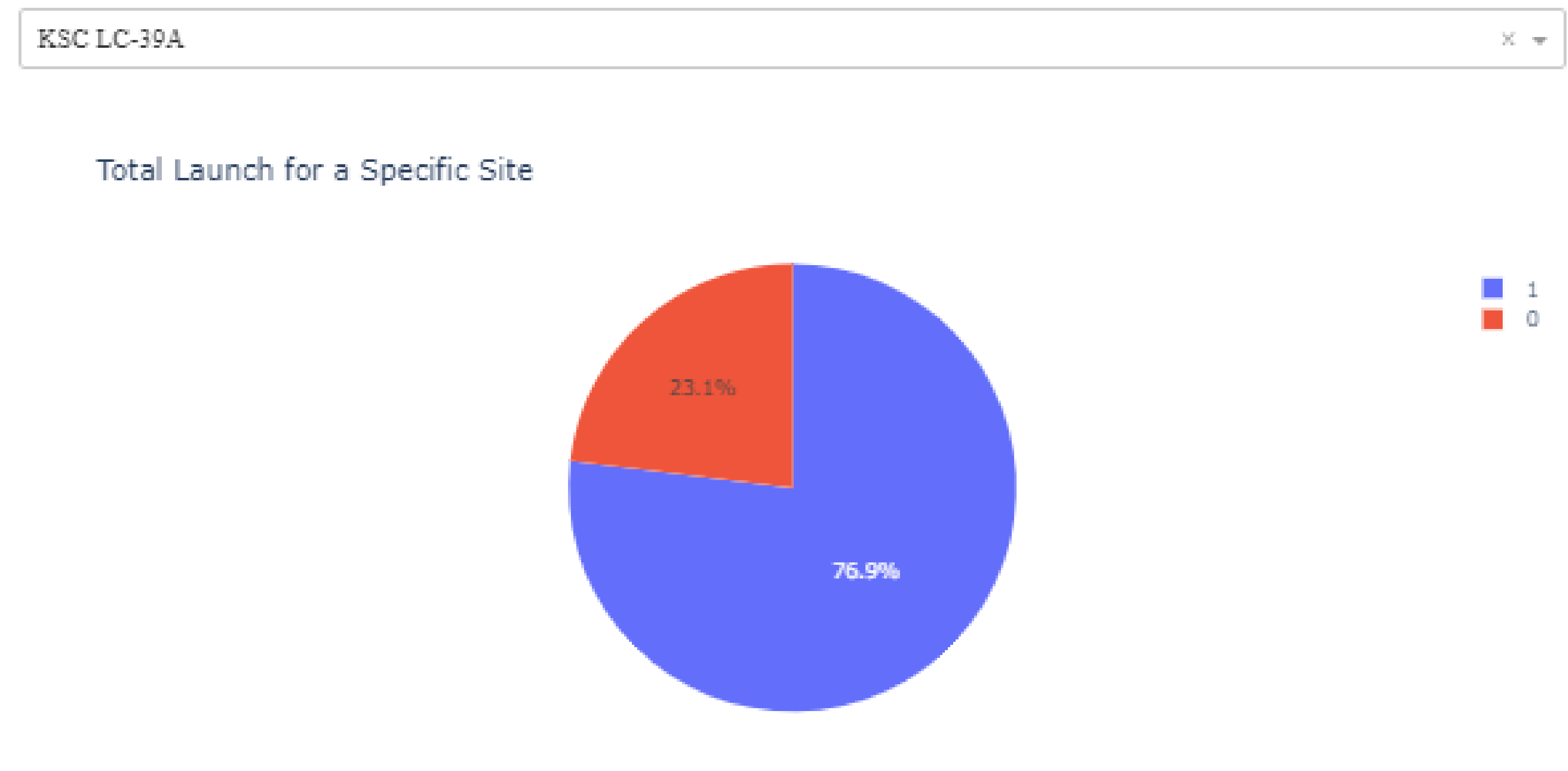
Successful Launches by Site



SpaceX Launch Records Dashboard

The launches sites are very important factor of success of missions.

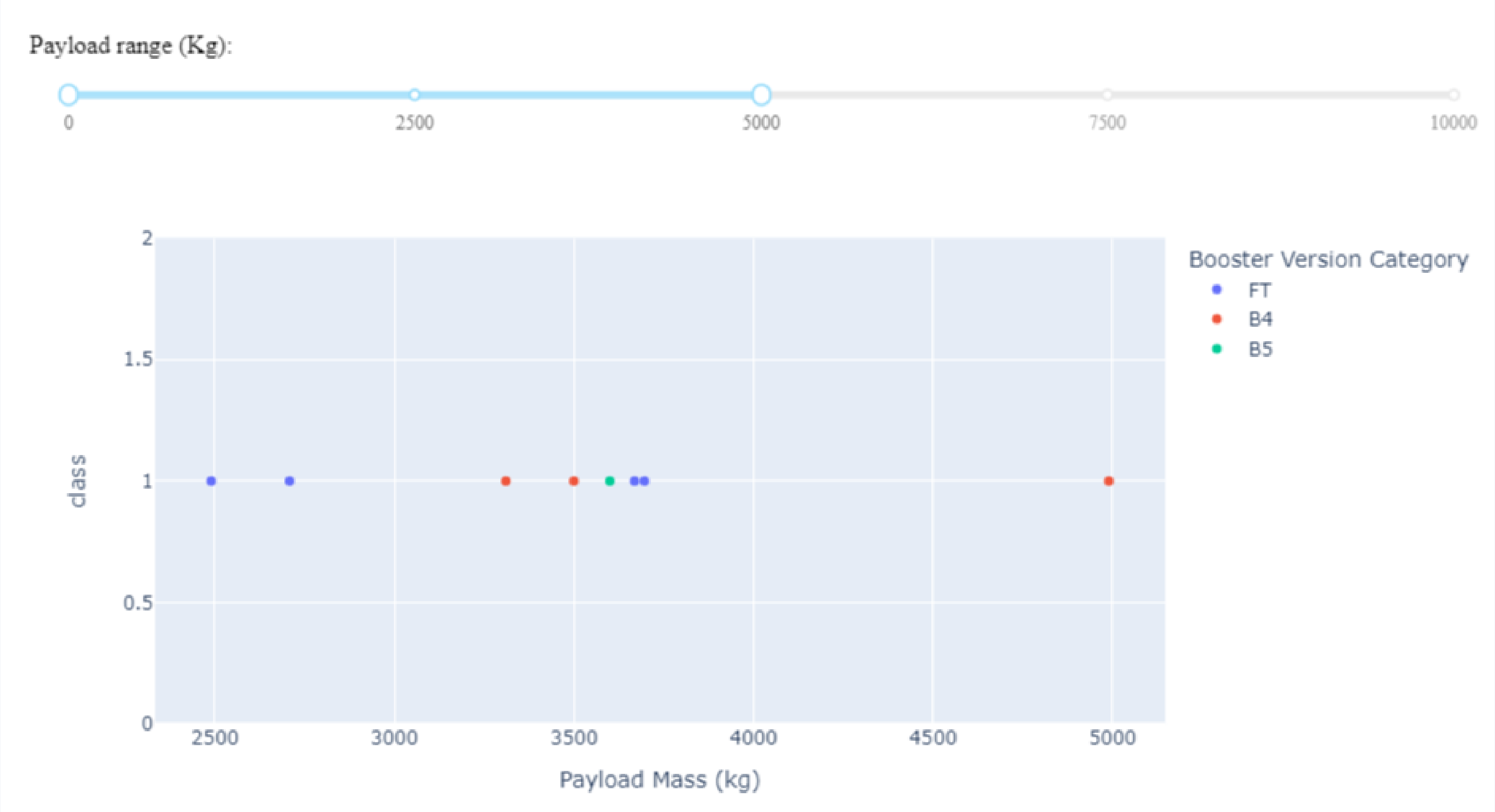
Highest launch success ratio : KSC LC-39A



KSC LC-39A

76.9% of launches are successful in this site.

Payload vs. Launch Outcome



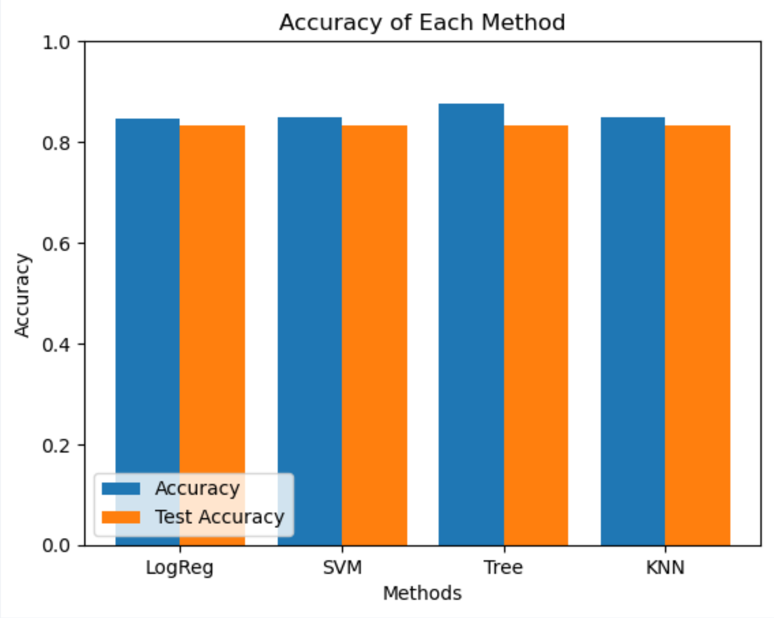
Payloads of 5000 Kg and FT boosters are the most successful combination.

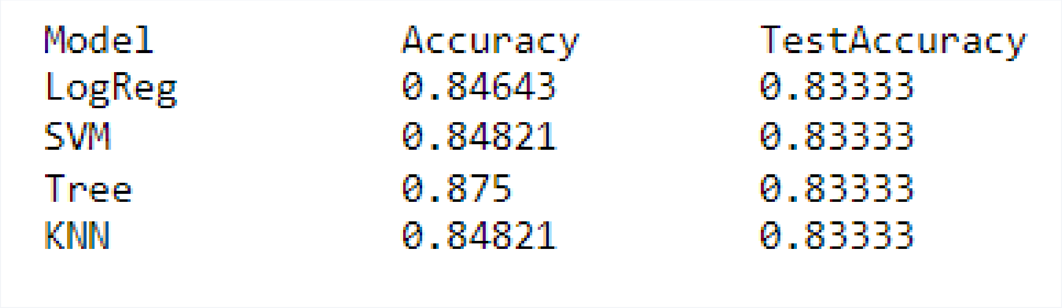
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted below.

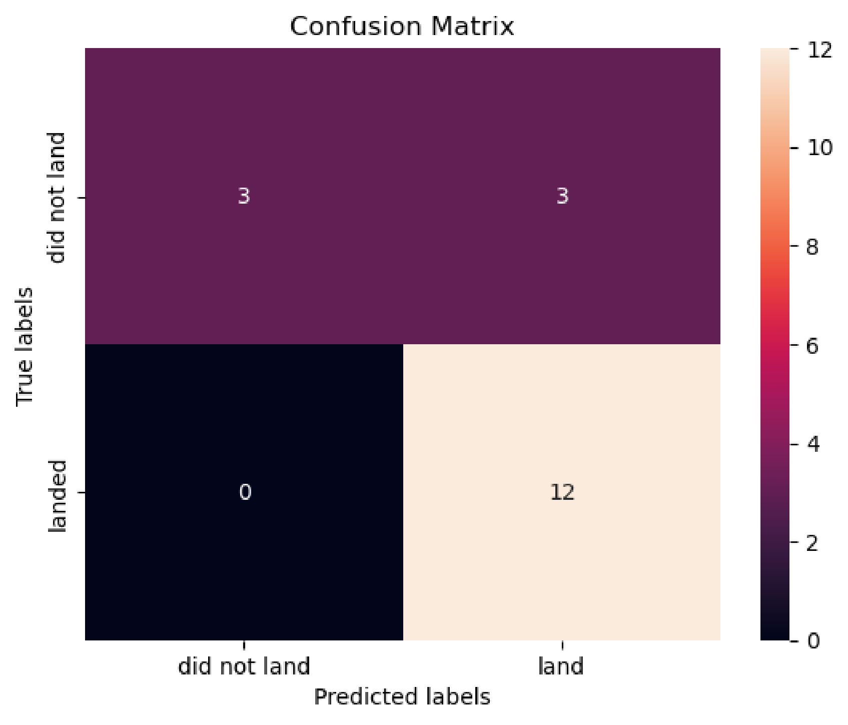




```
Model          Accuracy       TestAccuracy
LogReg         0.84643        0.83333
SVM            0.84821        0.83333
Tree           0.875          0.83333
KNN            0.84821        0.83333
```

- The models are close in terms of testing, but the best training model is Decision Tree Classifier, which has accuracies over than 87%.

Confusion Matrix



Confusion matrix of Decision Tree Classifier with big numbers of true positive compared to the false ones.

Conclusions

- The Decision Tree Model emerges as the most effective algorithm for this dataset, facilitating more accurate predictions.
- Launches with lower payload masses tend to exhibit higher success rates, a vital insight for mission planning.
- The strategic placement of launch sites near the Equator and coast enhances both operational efficiency and safety.
- A promising upward trend in launch success rates over the years indicates ongoing advancements in processes and rocket technology.
- Notably, KSC LC-39A boasts the highest success rate among all launch sites, underlining its reliability.
- Specific orbits, including ES-L1, GEO, HEO, and SSO, consistently achieve a remarkable 100% success rate, reinforcing their dependability.
- These findings offer valuable guidance for decision-making in the space mission domain, and the Decision Tree Classifier stands out as a reliable tool for predicting successful landings, contributing to increased profitability in the space launch industry.

Appendix

- I used my local Jupyter Lab for many labs due to issues with Coursera tools
- You can view the dashboard pictures in the GitHub link provided in that section.
- Special thanks for IBM and Coursera for the course.

Thank you!