IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Vivek Nakul Beera>
20-06-2025

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 demo in screenshots
- Predictive analysis results

Introduction

Project background and context

SpaceX has transformed the space industry by drastically reducing launch costs through reusable rockets—specifically the Falcon 9, which costs around $62M per launch compared to $165M+ from competitors. The key driver of this cost efficiency is the successful landing and reuse of the rocket's first stage. This project aims to predict whether the first stage will land successfully using public data and machine learning. Accurate predictions can help estimate launch costs, assess mission risks, and support more efficient planning in commercial spaceflight.

Problems you want to find answers

- ➢ How do factors like payload mass, launch site, number of previous flights, and orbit type influence the success of the first stage landing?
- ➢ Has the success rate of first stage landings improved over time?
- ➢ Which machine learning algorithm works best for this kind of binary classification problem?

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 collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

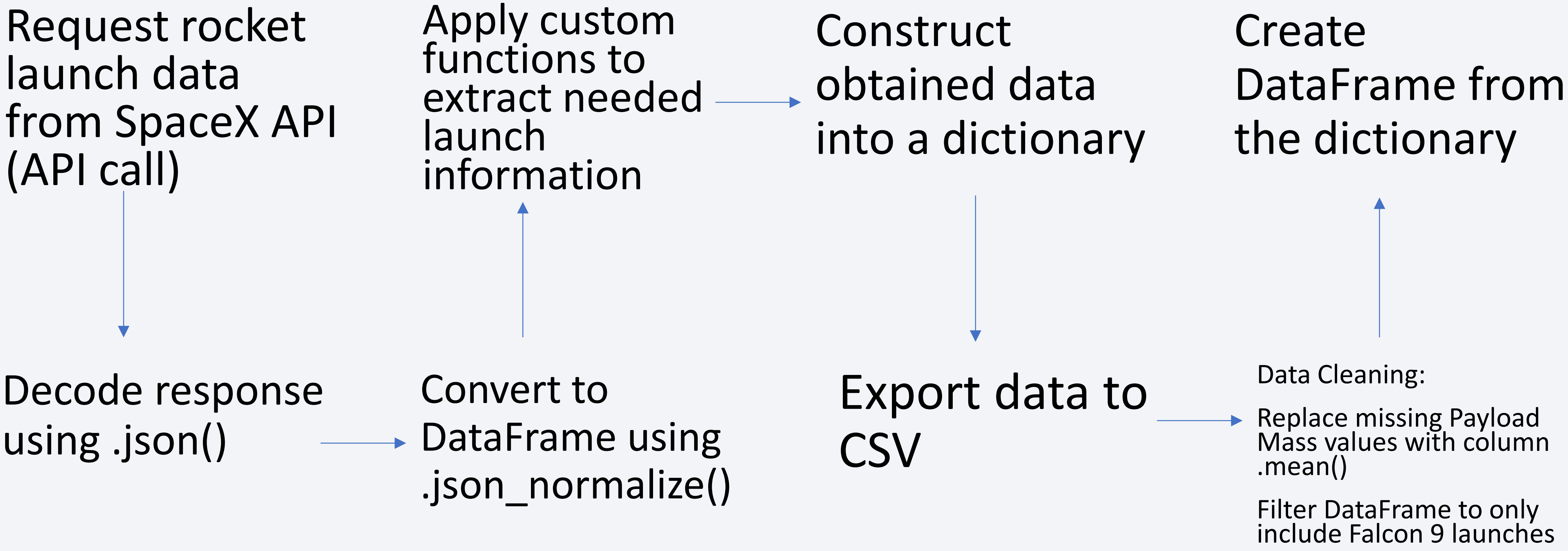
Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



Request rocket launch data from SpaceX API (API call)

Decode response using .json()

Convert to DataFrame using .json_normalize()

Apply custom functions to extract needed launch information

Construct obtained data into a dictionary

Export data to CSV

Data Cleaning:

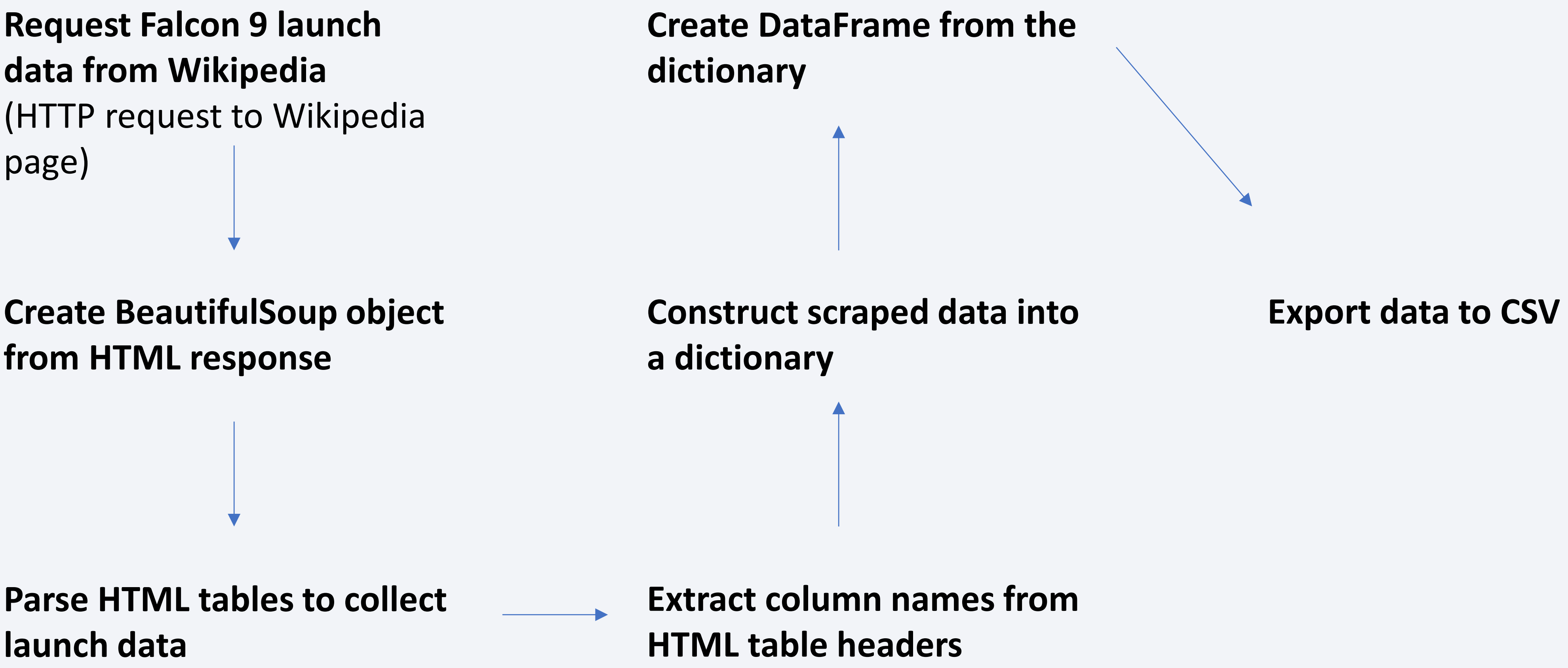
Replace missing Payload Mass values with column .mean()

Filter DataFrame to only include Falcon 9 launches

Create DataFrame from the dictionary

Link: Data Collection

Data Collection - Scraping



Request Falcon 9 launch data from Wikipedia (HTTP request to Wikipedia page)

Create BeautifulSoup object from HTML response

Parse HTML tables to collect launch data

Extract column names from HTML table headers

Construct scraped data into a dictionary

Create DataFrame from the dictionary

Export data to CSV

[Link: Web scraping]

Data Wrangling

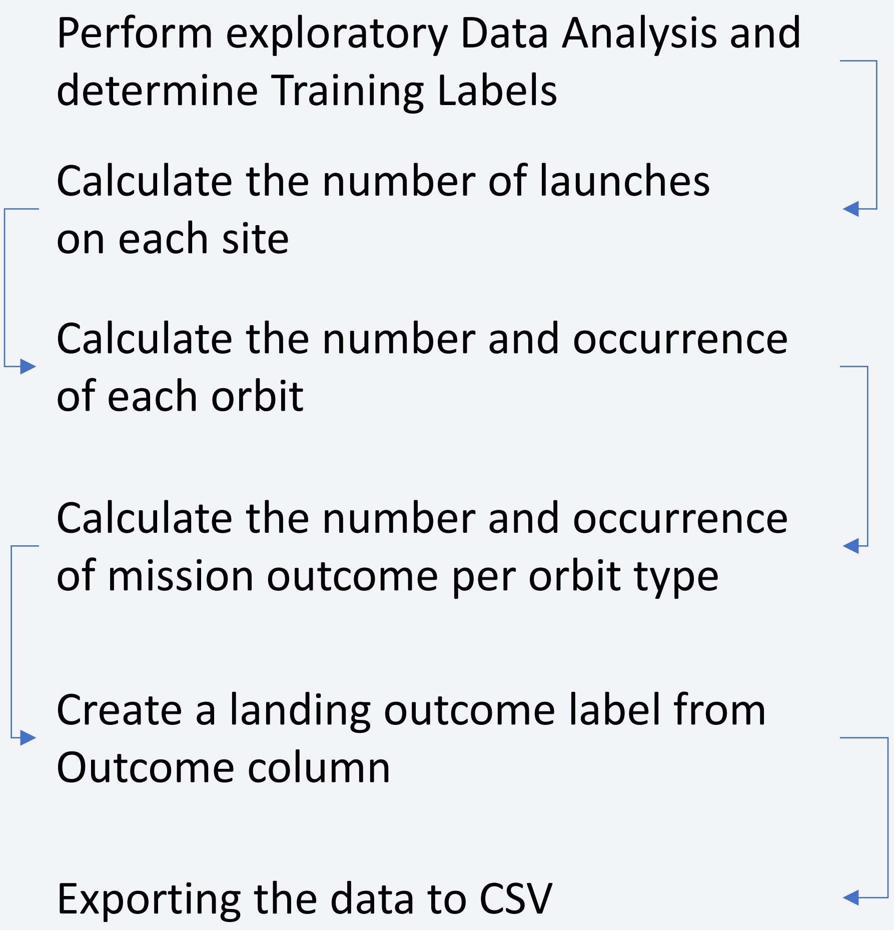
Understanding Landing Outcomes in the Dataset

The dataset includes various scenarios where SpaceX's booster rockets failed to land successfully. Let me break down what these different cases mean:

- **Ocean Landings**
    - True Ocean: The rocket successfully landed in a designated ocean area.
    - False Ocean: The rocket failed to land safely in the ocean (it crashed or was lost).
- **Ground Landings (RTLS - Return to Launch Site)**
    - True RTLS: The rocket made it back safely to the landing pad on ground.
    - False RTLS: The rocket failed to land properly on the ground pad.
- **Drone Ship Landings (ASDS - Autonomous Spaceport Drone Ship)**
    - True ASDS: The rocket landed successfully on the floating drone ship.
    - False ASDS: The rocket missed or crashed on the drone ship.

For our analysis, we've simplified these outcomes:

- "1" = Successful landing (any True case)
- "0" = Unsuccessful landing (any False case)



Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

[Link : Data Wrangling]

EDA with Data Visualization

- Charts were plotted:

  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

- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

- Line charts show trends in data over time (time series)

Link: EDA wit Dataviz

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

Link : EDA with SQL

Build an Interactive Map with Folium

Interactive SpaceX Launch Site Map

I built an interactive map using Folium to visualize SpaceX launch sites and their landing outcomes. Here's what it includes:

1. Launch Site Markers

   - Started by plotting NASA Johnson Space Center (as a reference point) with a labeled marker and circle overlay.
   - Added all SpaceX launch sites with markers, popup labels, and text to show their exact locations—highlighting how close they are to the equator and coastlines.

2. Success vs. Failure Clusters

   - Used color-coded markers (green = success, red = failure) grouped in clusters to quickly see:
   - Which launch sites have the highest success rates.
   - Patterns in failed landings (e.g., if certain sites struggle more than others)

3. Proximity Analysis

- Drew colored lines from Kennedy Space Center (KSC LC-39A) to nearby landmarks like:
  - Railways
  - Highways
  - Coastlines
  - The nearest city

This helps visualize how infrastructure and geography might influence launch operations. The map makes it easy to explore SpaceX's launch geography, success trends, and site logistics—all in one interactive tool!.

link : visual analytics with Folium

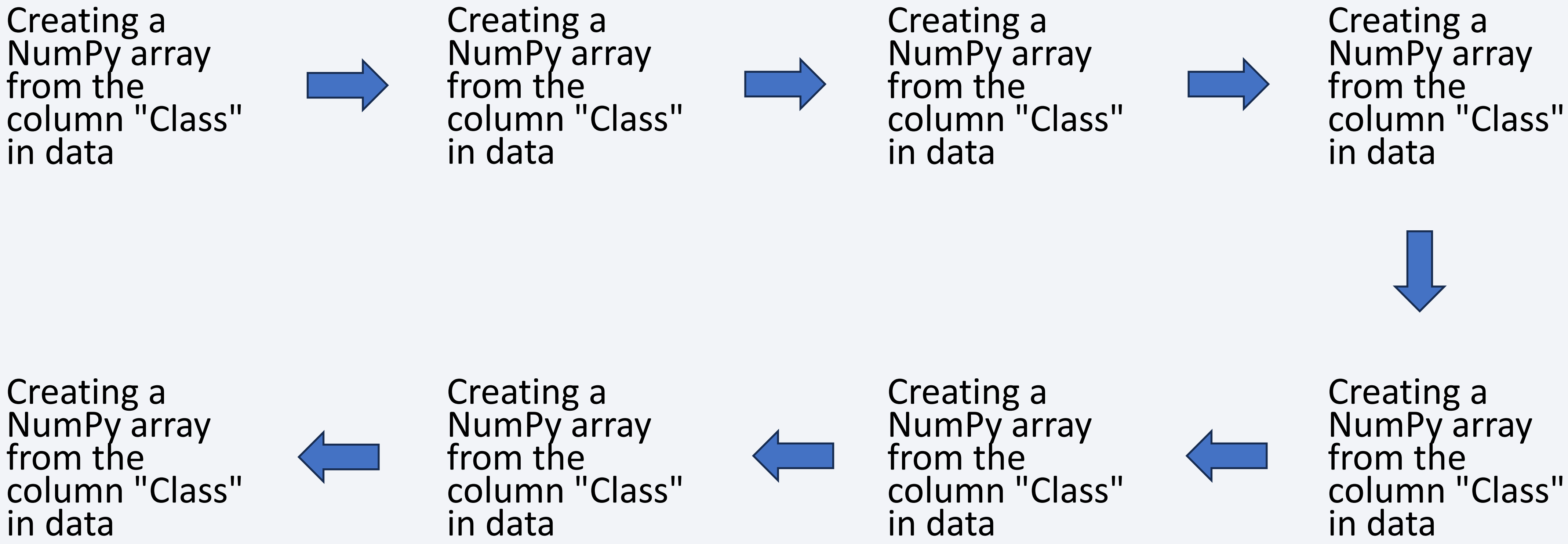
Build a Dashboard with Plotly Dash

Launch Sites Dropdown List

- Pie Chart showing Success Launches (All Sites/Certain Site)
- Slider of Payload Mass Range
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions

Link : Dashboard with Plotly Dash

Predictive Analysis (Classification)



Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Creating a NumPy array from the column "Class" in data

Link : ML predictive analysis

Results

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

Section 2

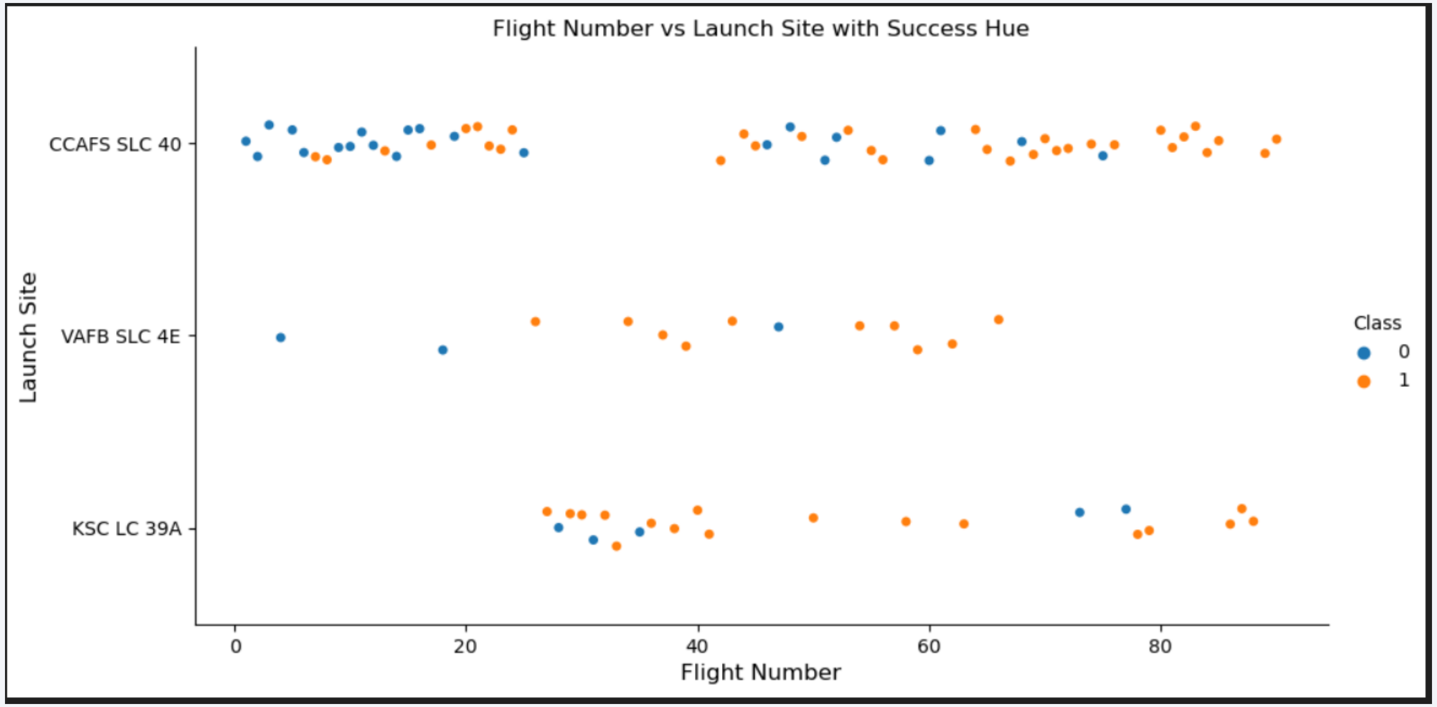
Insights drawn from EDA

Flight Number vs. Launch Site

KSC LC 39A (historic site for crewed missions) should show high success rates in later flights due to rigorous safety standards.

CCAFS SLC 40 may have early failures but improves over time

Early flights (0–20) show sparse launches, likely reflecting initial testing phases with higher risk of failures.
Mid-range flights (20–60) exhibit increased launch frequency, suggesting operational maturity. Success rates likely improve here.
Later flights (60–80) indicate sustained operations, with possible clustering at **KSC LC 39A** or **CCAFS SLC 40** due to reusable rocket recovery infrastructure.



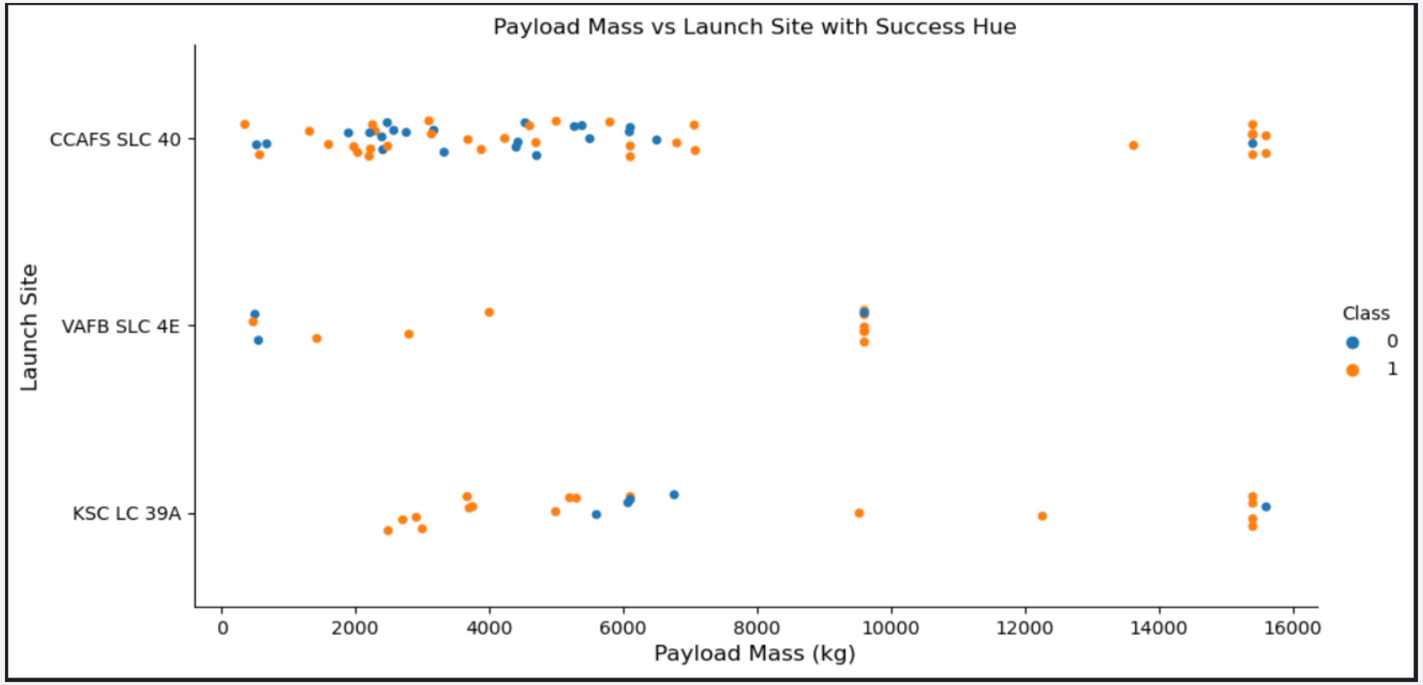
Payload vs. Launch Site

Success Dominates Mid-Mass Range 4,000–12,000 kg across sites.

Heavy Payload Expertise at KSC.

Failures at Extremes.

VAFB’s Light-Payload Niche



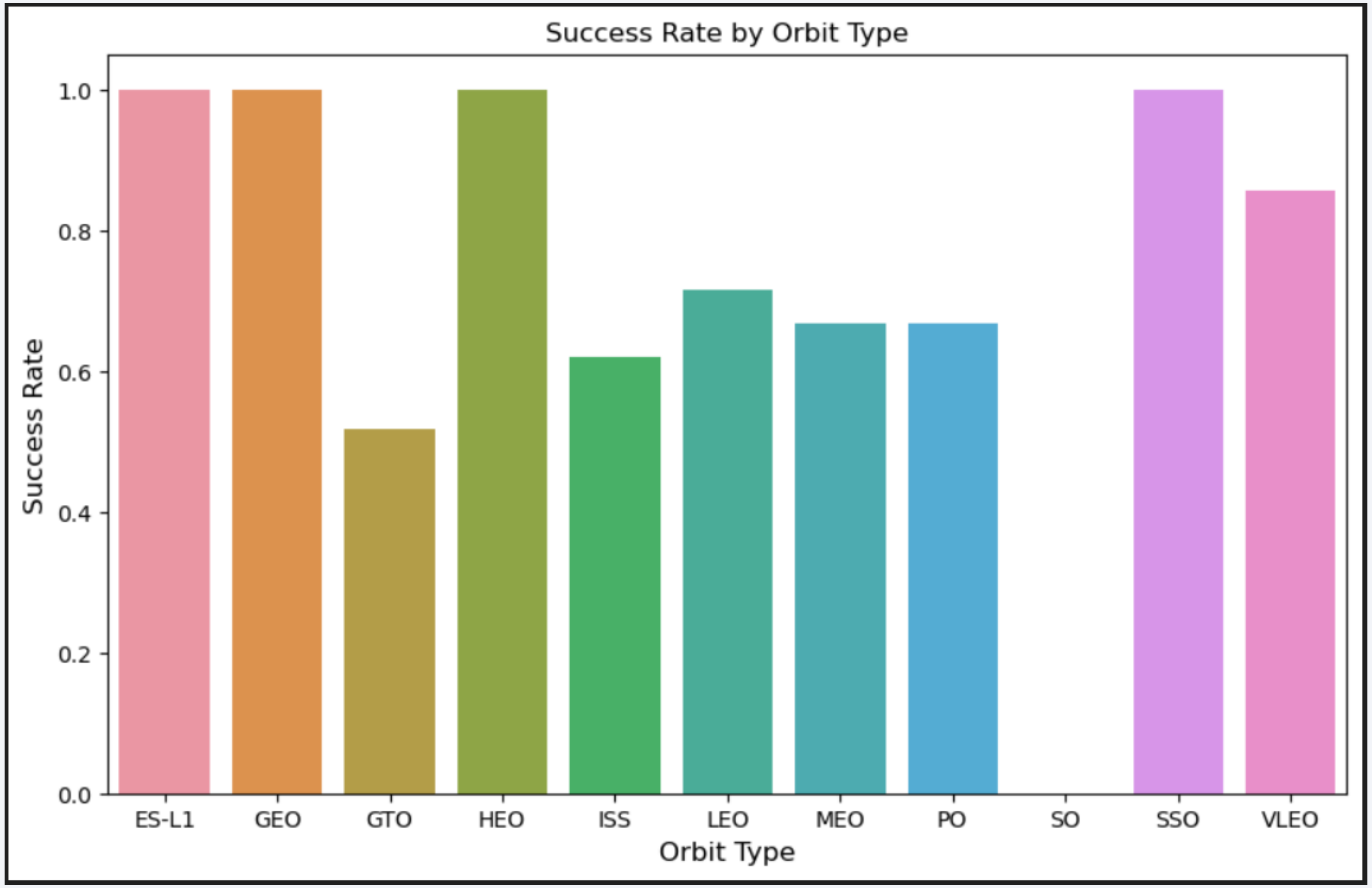
Success Rate vs. Orbit Type

ISS (resupply) and LEO (satellite deployments) show near-perfect success rates, reflecting mature launch protocols and frequent missions.

GTO and HEO exhibit lower success rates.

E5-L1 (deep space) and VLEO (atmospheric drag) show the lowest success, highlighting technical extremes.

PO and SSO (Earth observation orbits) achieve strong success, driven by standardized smallsat launch systems.



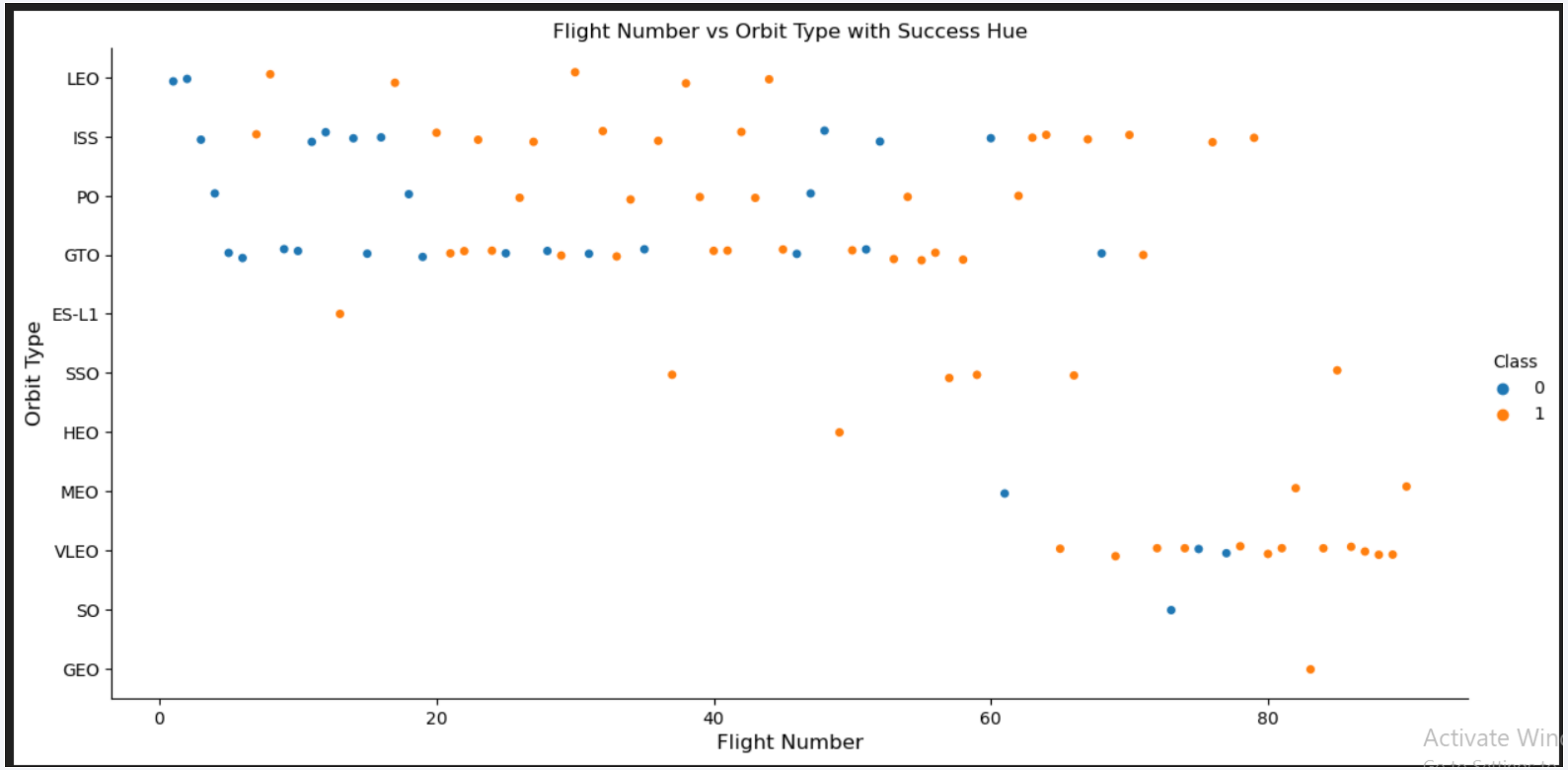
Flight Number vs. Orbit Type

- High failure density in **flights <20** for complex orbits (e.g., **GTO, HEO, ES-L1**), reflecting initial technical challenges.

- **LEO** and **ISS** orbits achieve near-perfect success (orange) after ~Flight 15, indicating quick operational maturity for low-Earth missions.

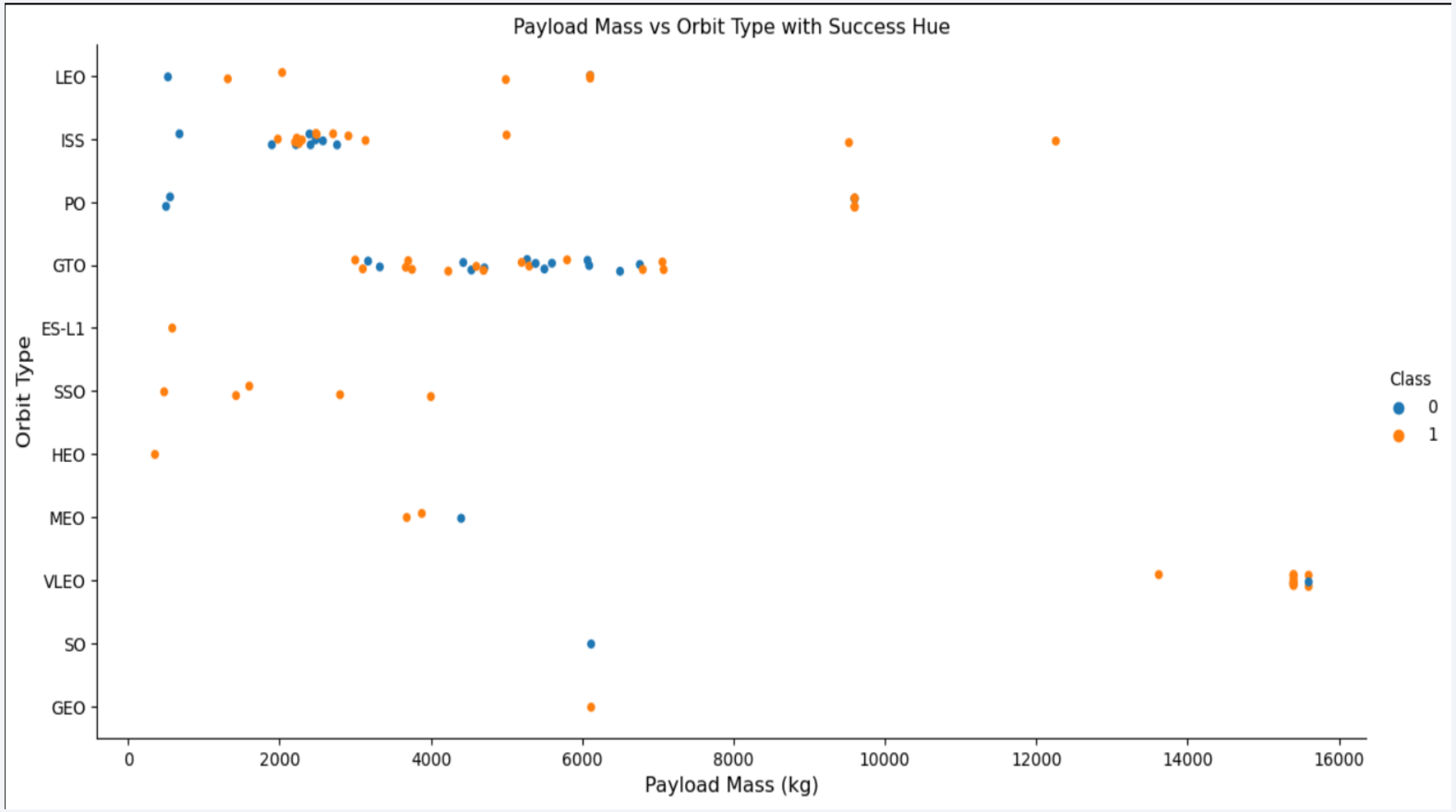
- GTO and HEO show intermittent failures (blue) even beyond Flight 40, highlighting enduring risks in high-energy orbit injections.

- ES-L1 (deep space) and VLEO suffer failures in later flights (>60), stressing unique challenges in extreme environments.



Payload vs. Orbit Type

- LEO (Low Earth Orbit) has the highest number of payloads across a wide range of masses, with both successful and unsuccessful launches.
- VLEO (Very Low Earth Orbit) and GEO (Geostationary Orbit) have fewer launches, with VLEO showing a mix of successes and failures at higher payload masses, while GEO has minimal activity.
- GTO (Geostationary Transfer Orbit) and ISS (International Space Station) orbits show a moderate number of launches, with a noticeable proportion of successful missions at lower to mid-range payload masses.
- Successful launches tend to dominate across most orbit types, especially at lower payload masses, while unsuccessful launches are more scattered and less frequent.



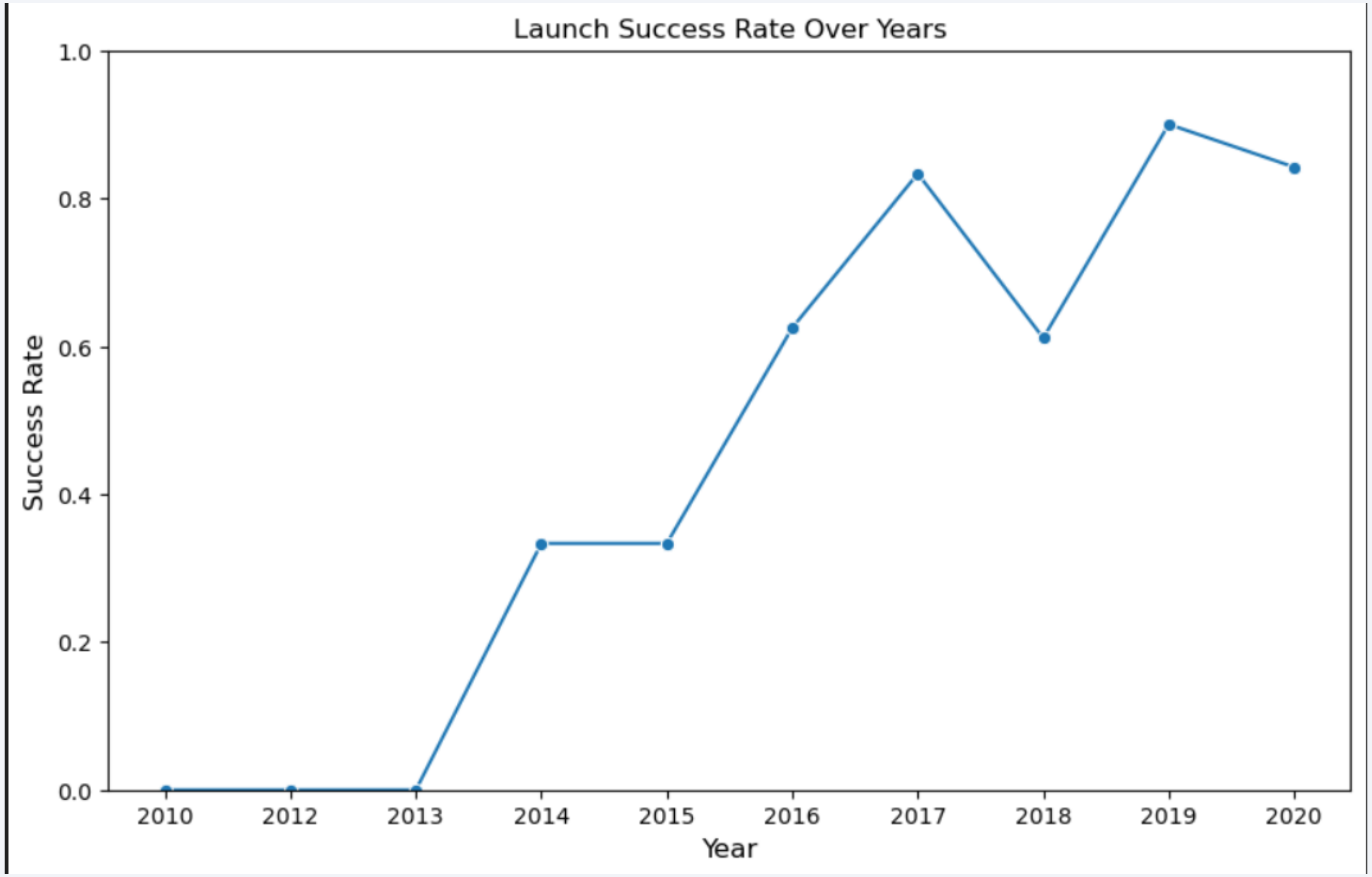
Launch Success Yearly Trend

- The launch success rate remained near 0 from 2010 to 2012, indicating very few or no successful launches.

- A significant increase in success rate began around 2013, rising steadily to around 0.4 by 2014-2015.

- The success rate peaked at approximately 0.9 in 2018, showing a high level of reliability during that year.

- After 2018, the success rate fluctuated, dropping slightly to around 0.8-0.9 by 2020.

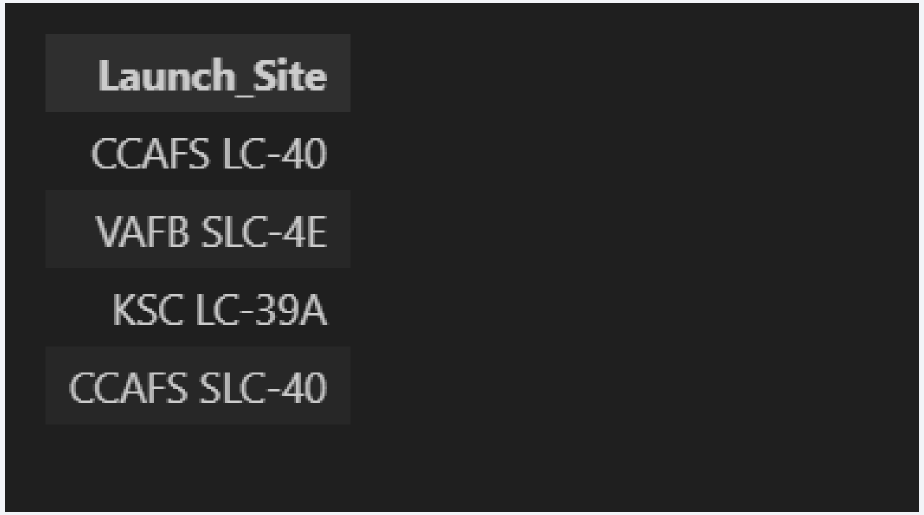


All Launch Site Names

Displaying the names of the unique launch sites in the space mission.

Code :

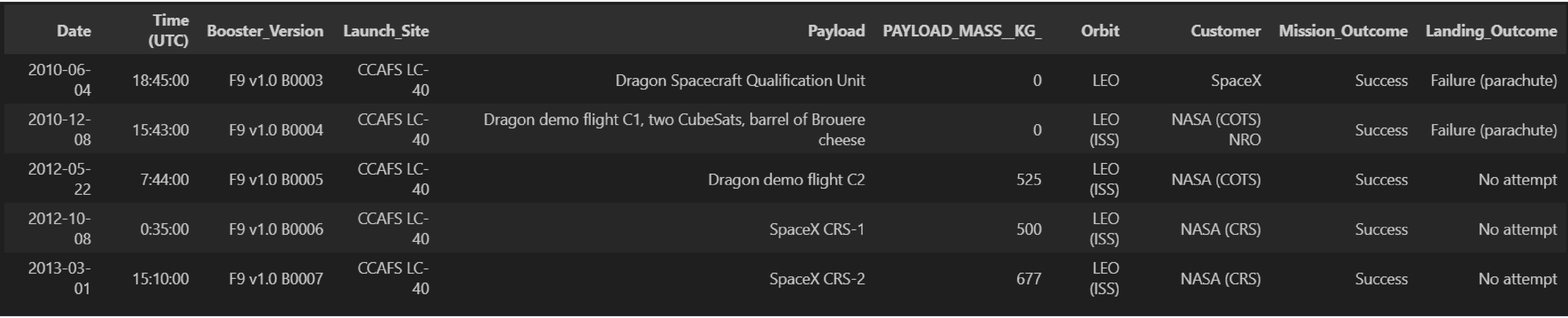
```
%sql SELECT DISTINCT "Launch_Site" FROM
SPACEXTABLE;
```



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

Launch Site Names Begin with 'CCA'

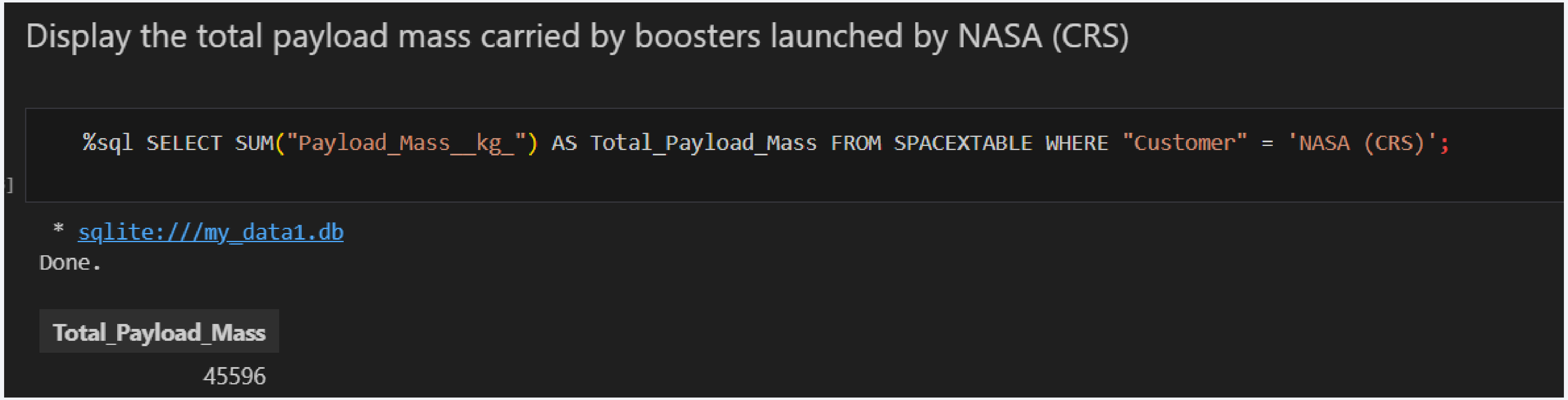
- All missions from 2010 to 2013 were successful in reaching their intended orbits, despite initial landing failures.
- The first two missions (2010) carried no significant payload mass (0 kg) and involved qualification or demo flights, with landing failures due to parachute issues.
- Starting in 2012, payload masses increased (e.g., 525 kg for Dragon demo flight C2), and no landing attempts were made, indicating a focus on orbit success.
- NASA was the primary customer, supporting SpaceX's development for ISS resupply missions (CRS program).



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Displaying the total payload mass carried by boosters launched by NASA (CRS)



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

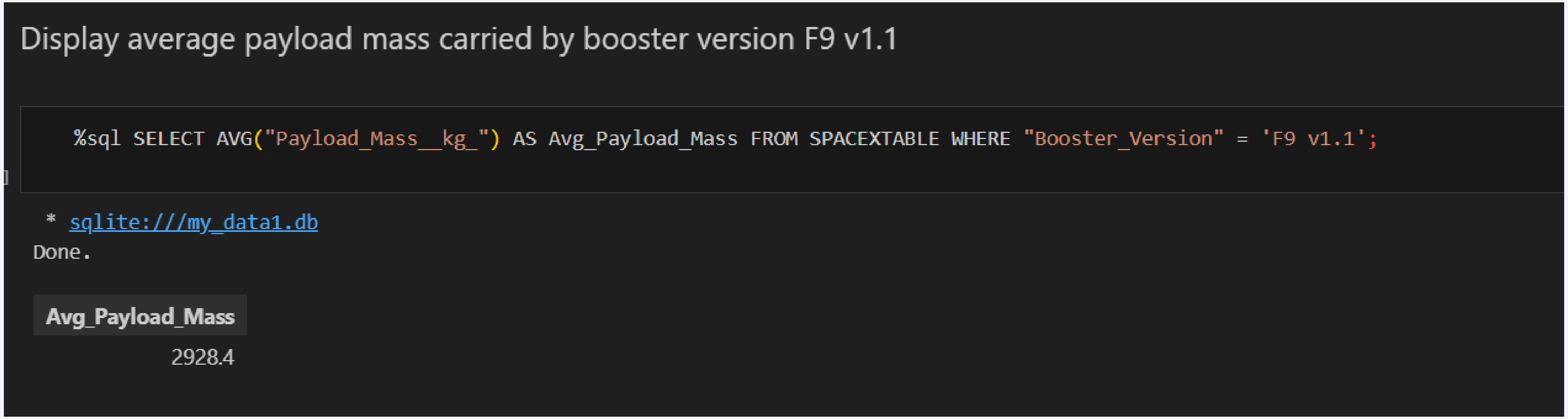
```
%sql SELECT SUM("Payload_Mass__kg_") AS Total_Payload_Mass FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

Displaying average payload mass carried by booster version F9 v1.1.



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

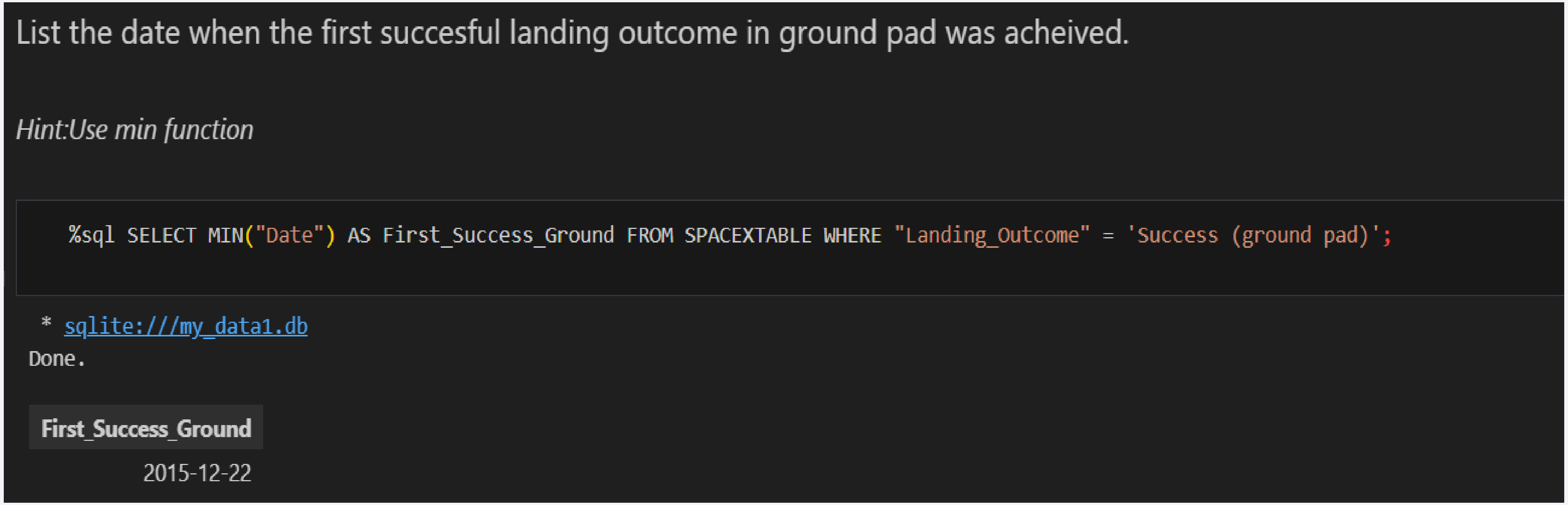
```
%sql SELECT AVG("Payload_Mass__kg_") AS Avg_Payload_Mass FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

the date when the first successful landing outcome in ground pad was achieved



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

Hint:Use min function

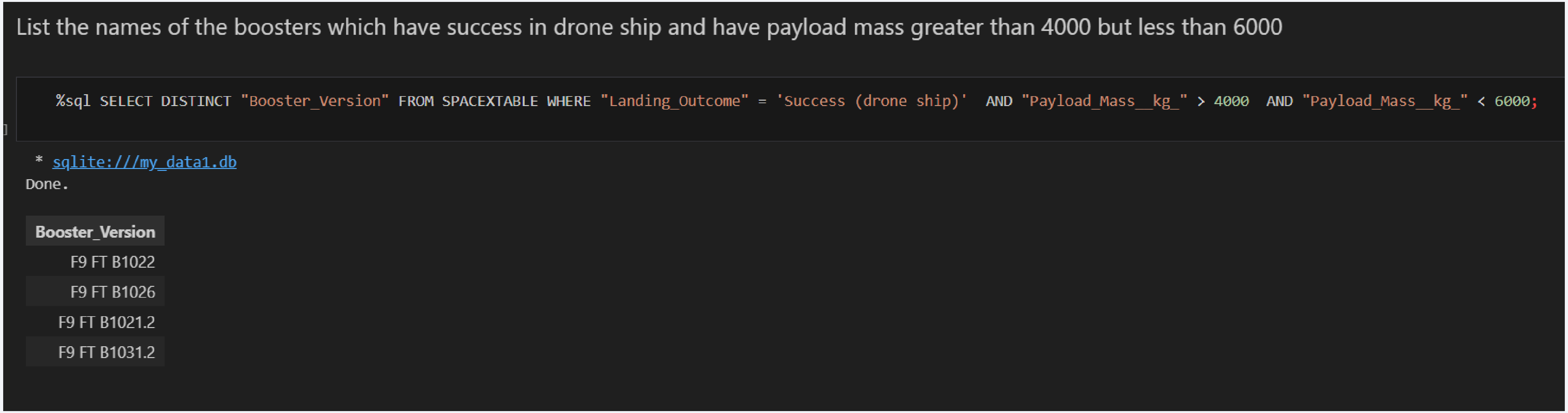
```
%sql SELECT MIN("Date") AS First_Success_Ground FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

First_Success_Ground
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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



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

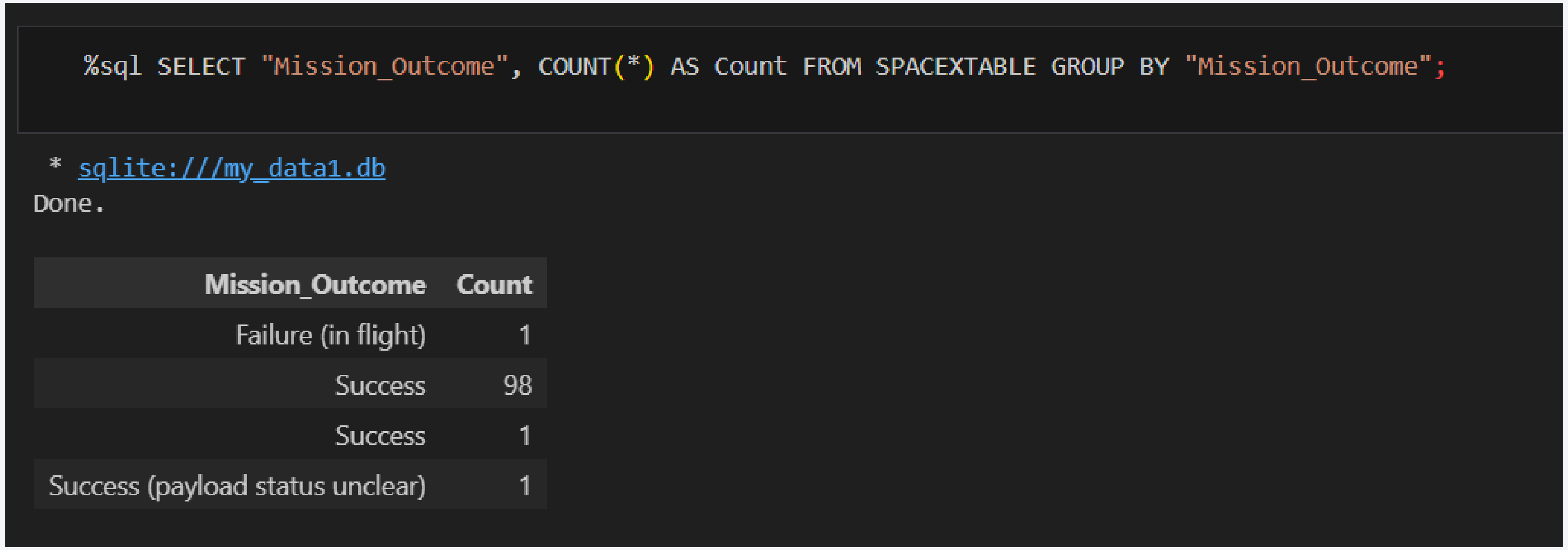
```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)'  AND "Payload_Mass__kg_" > 4000  AND "Payload_Mass__kg_" < 6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes



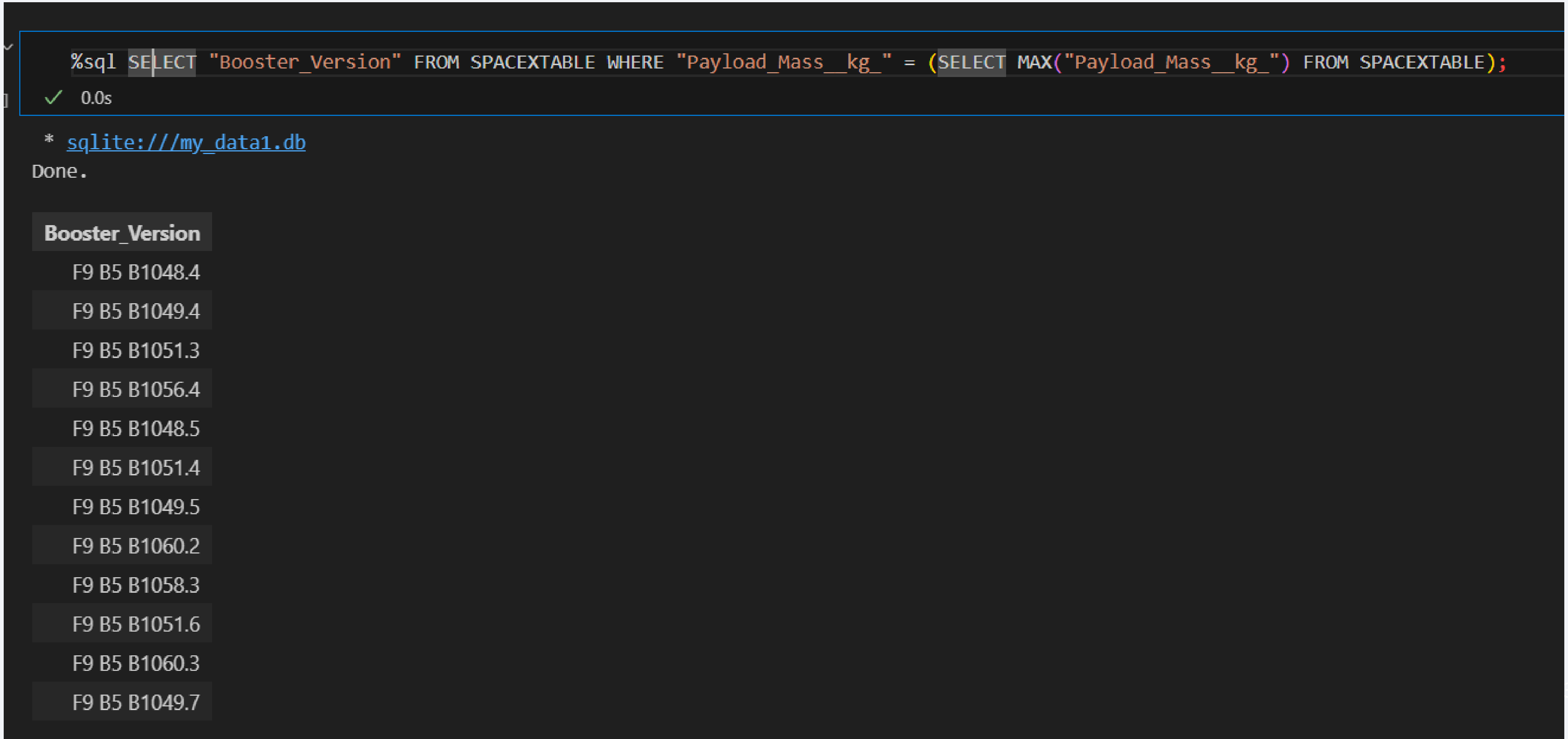
```
%sql SELECT "Mission_Outcome", COUNT(*) AS Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

The names of the booster versions which have carried the maximum payload mass



```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);
```

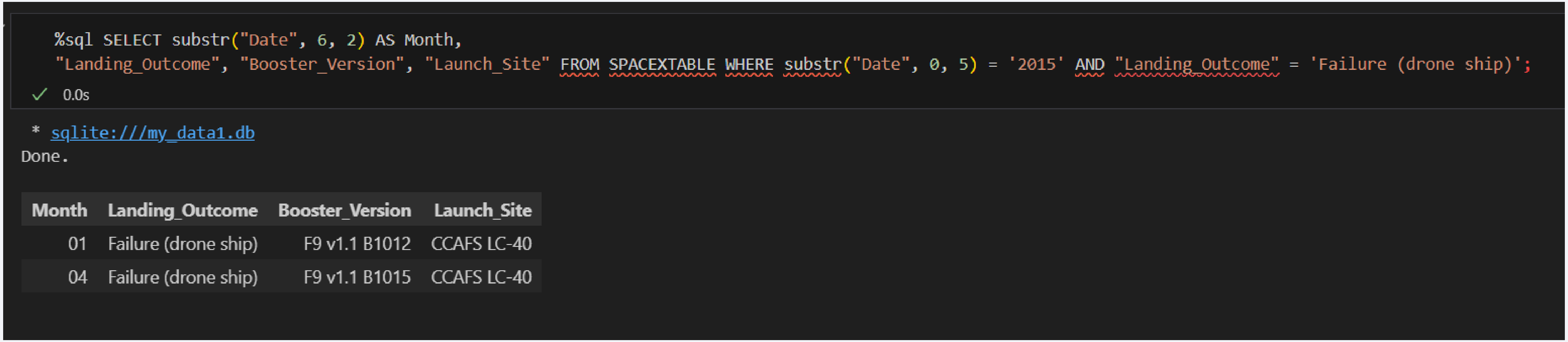
✓ 0.0s

```
 * 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

The failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015



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

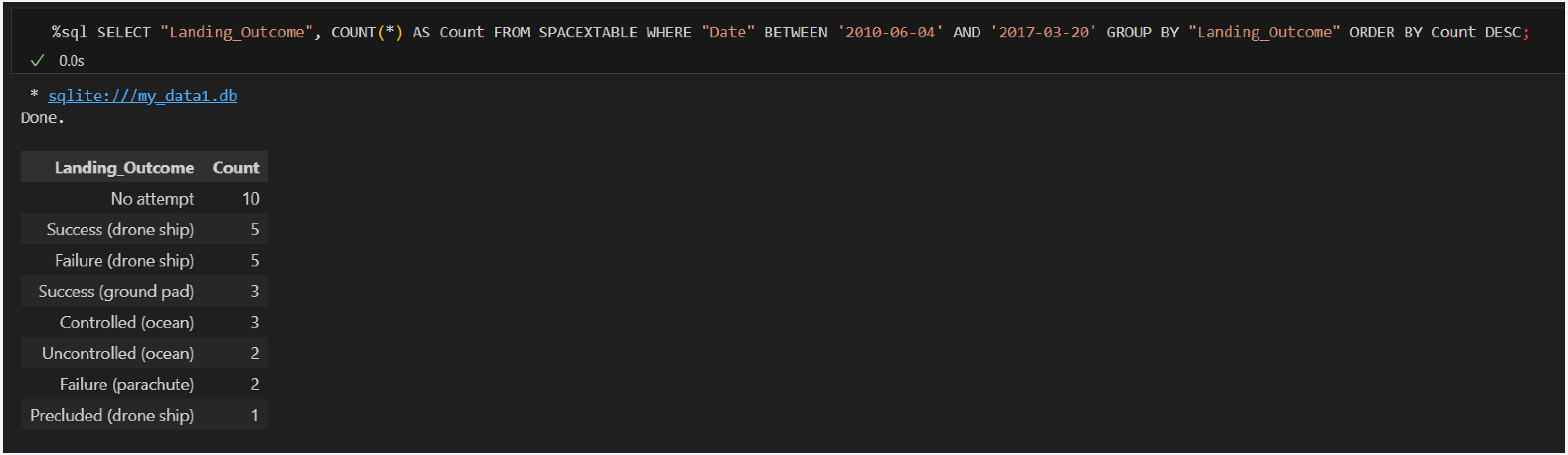
✓ 0.0s

```
 * 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

Ranking 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 FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY Count DESC;
```

✓ 0.0s

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

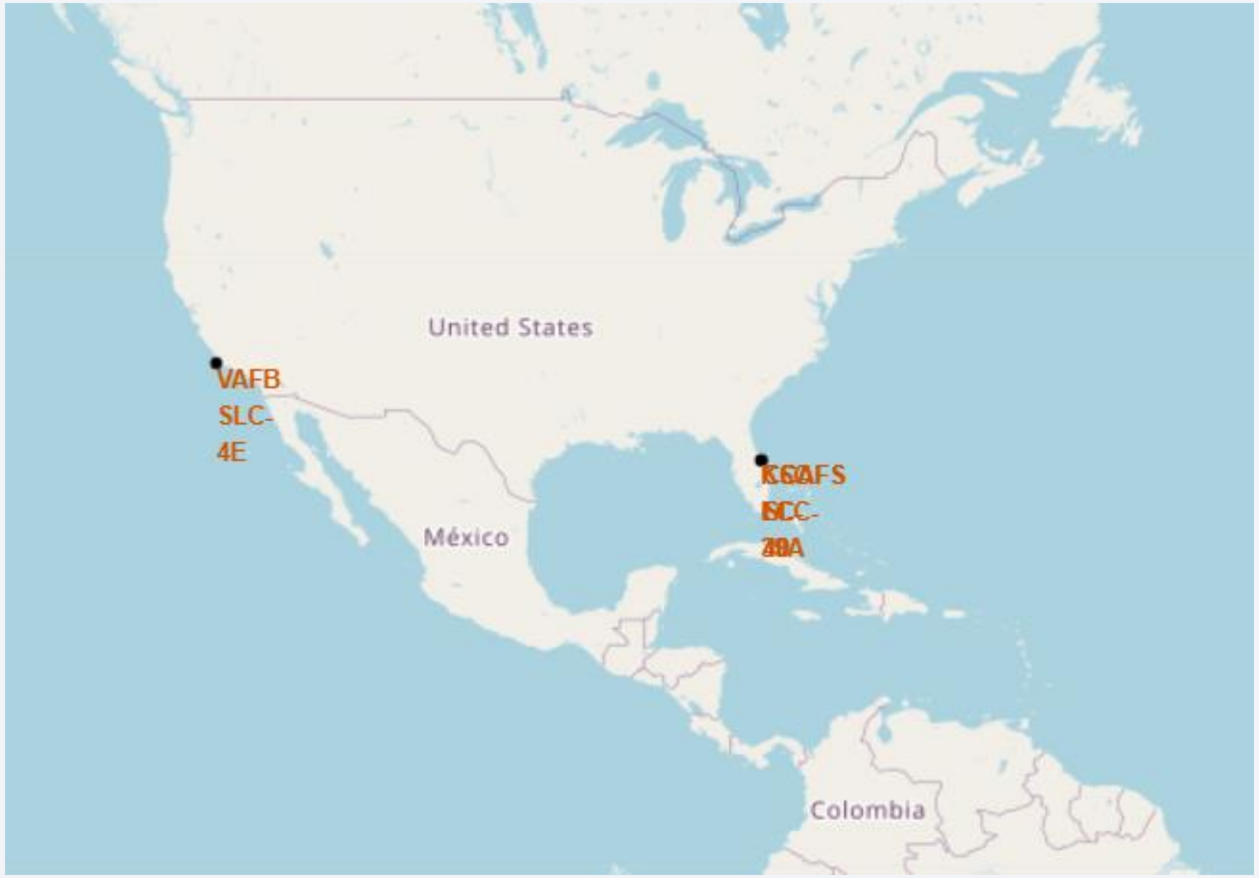
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

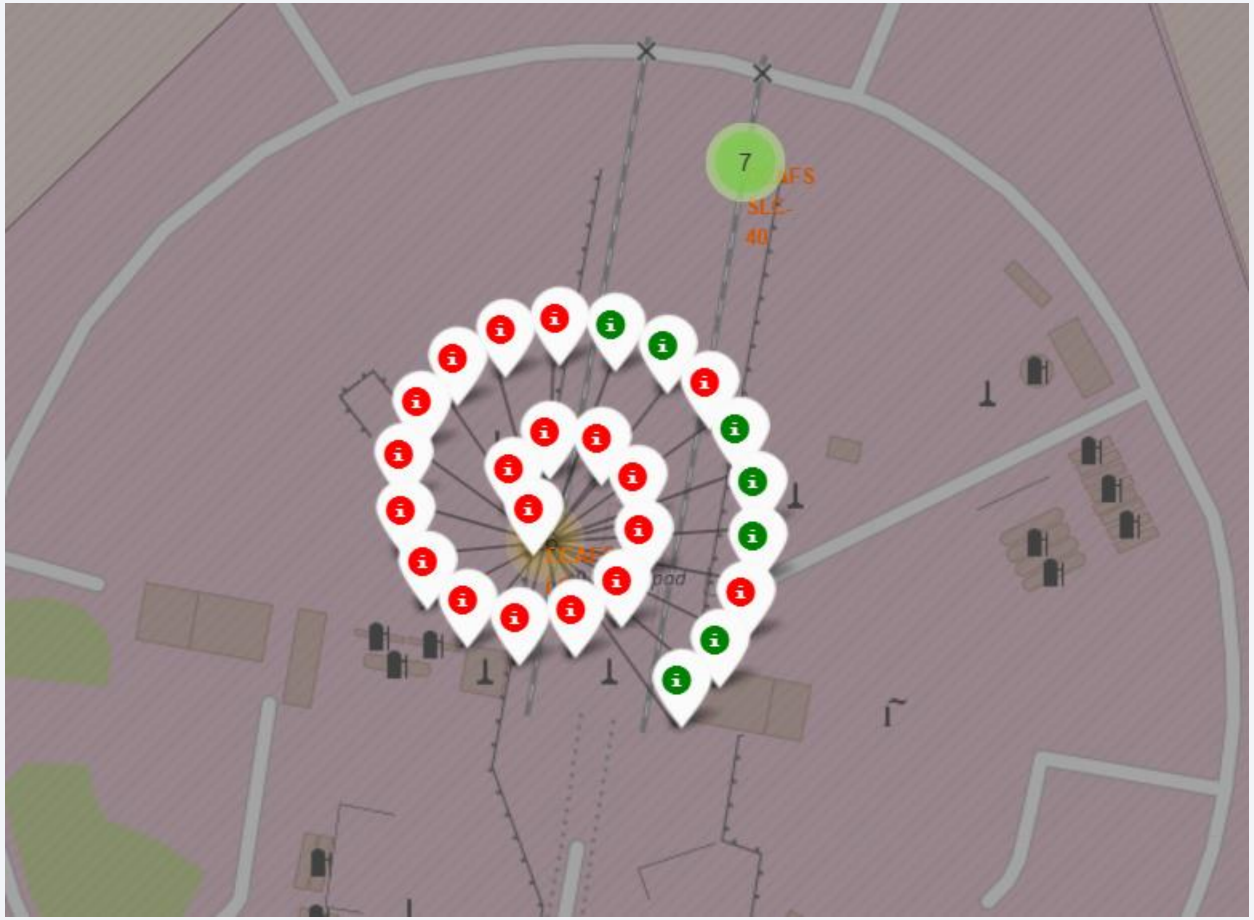
All launch sites’ location markers on a global map

- Most launch sites are located near the Equator, where the Earth's surface moves fastest at 1670 km/hour.
- This equatorial speed, due to inertia, assists spacecraft in achieving the velocity needed to stay in orbit.
- Launching from the Equator provides an initial boost, aiding the rocket's ascent into space.
- Proximity to the coast allows rockets to be launched over the ocean, reducing the risk of debris falling on land.
- Coastal locations minimize potential hazards to populated areas during launch operations.



Colour-labeled launch records on the map

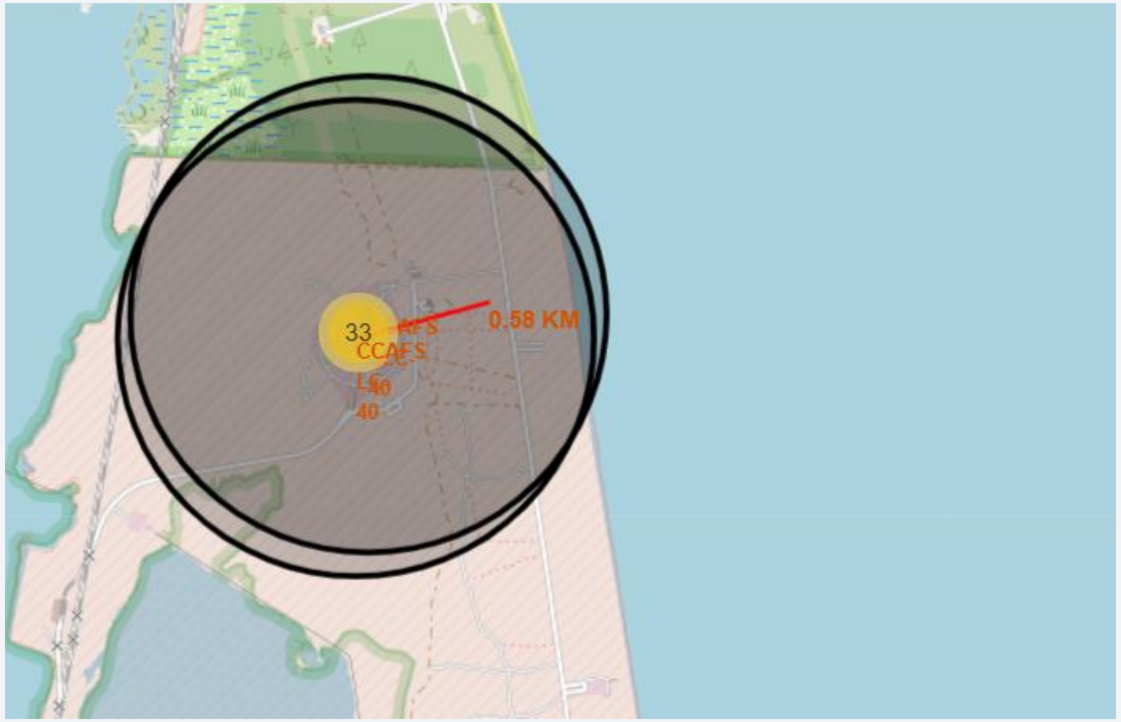
- From the colour-labeled markers we can identify which launch sites have relatively high success rates.
- Green Marker = Successful
- Launch - Red Marker = Failed Launch
- Launch Site CCAFS LC-40 has a very high Success Rate

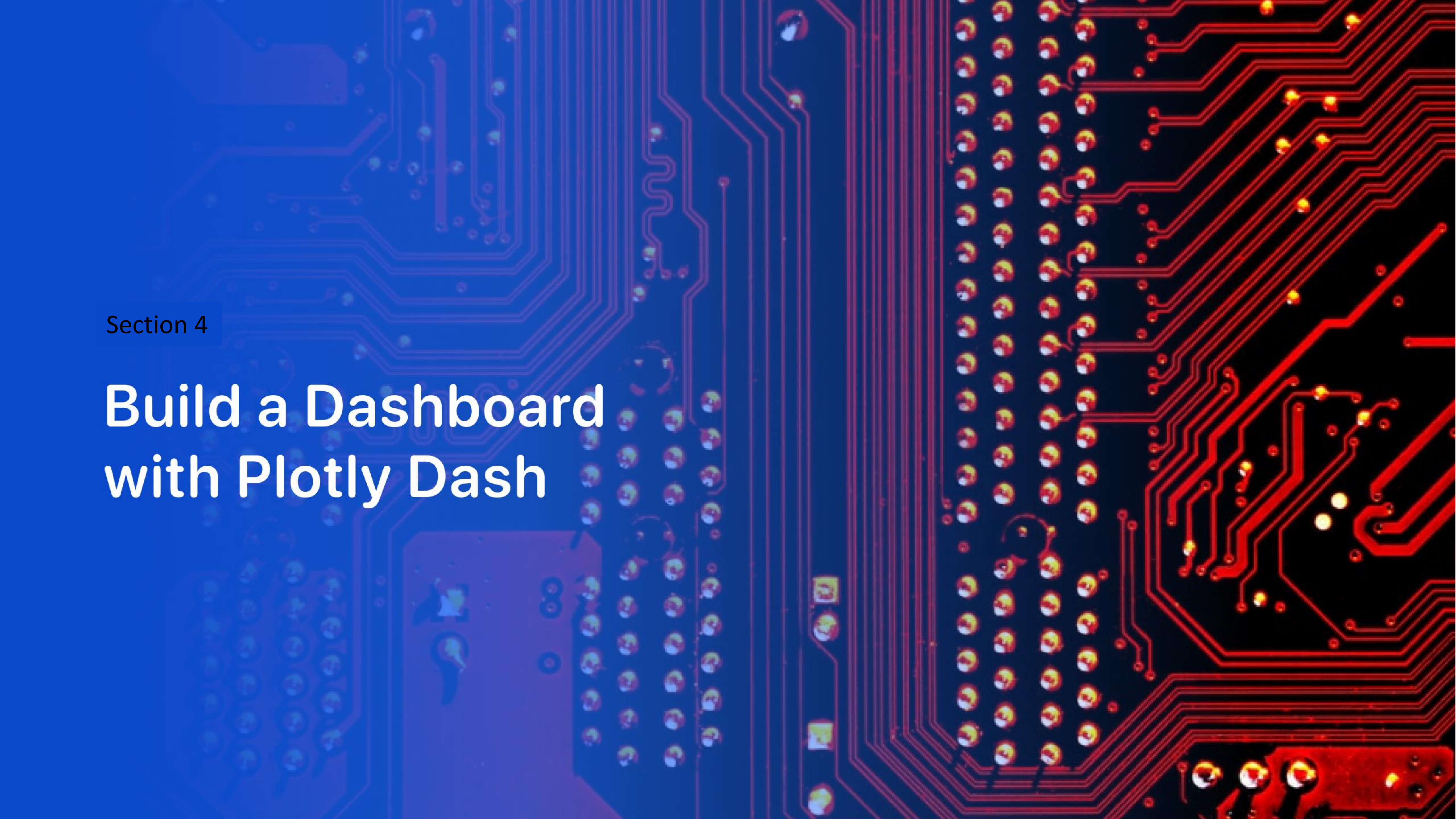


Distance from the launch site CCAFS LC-40 to its proximities

From the visual analysis of the launch site CCAFS LC-40 we can clearly see that it is:

relative close to coastline 0.58 km



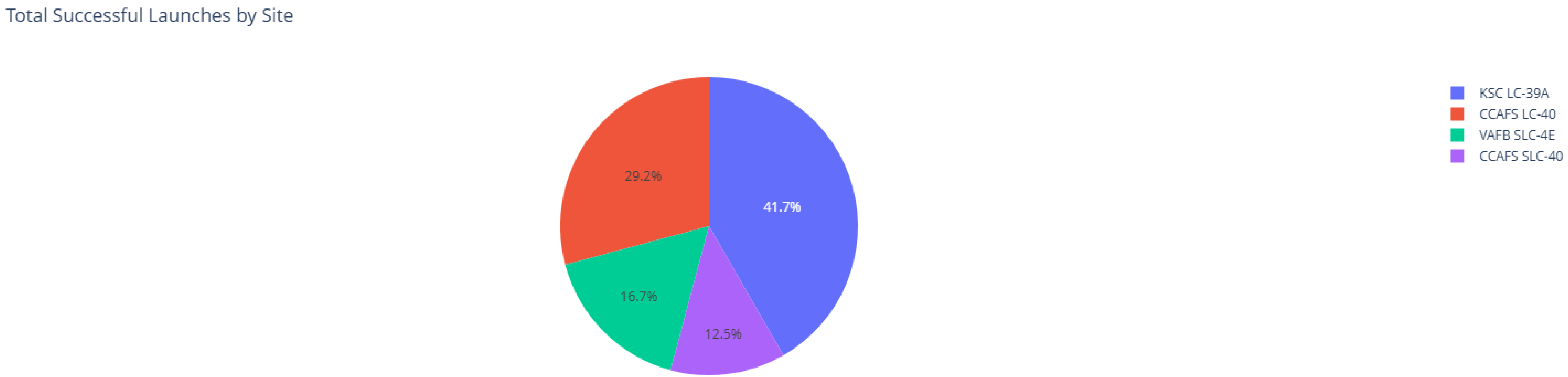


Section 4

Build a Dashboard with Plotly Dash

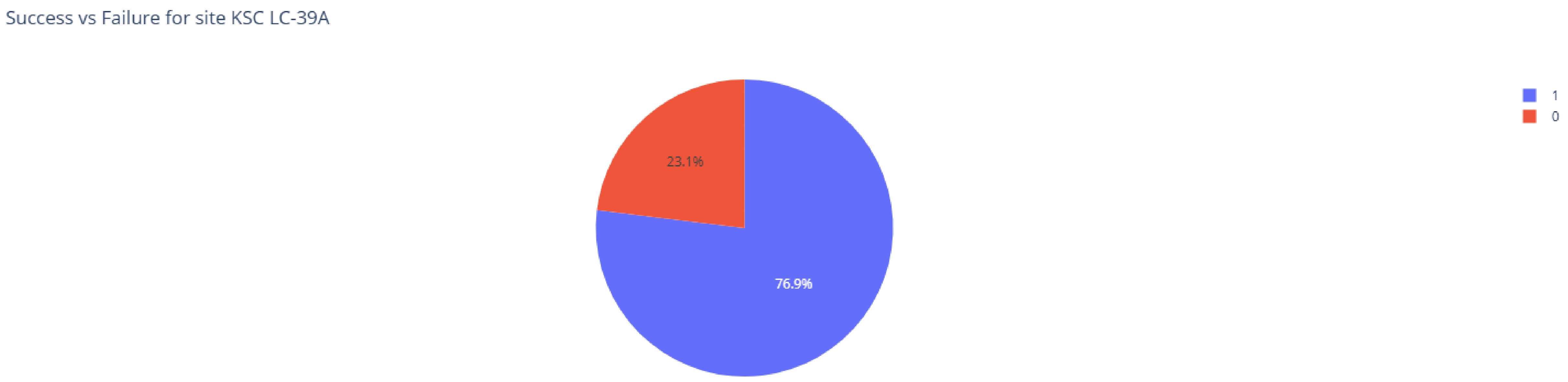
Launch success rate of all sites

The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.



the launch site with highest launch success ratio

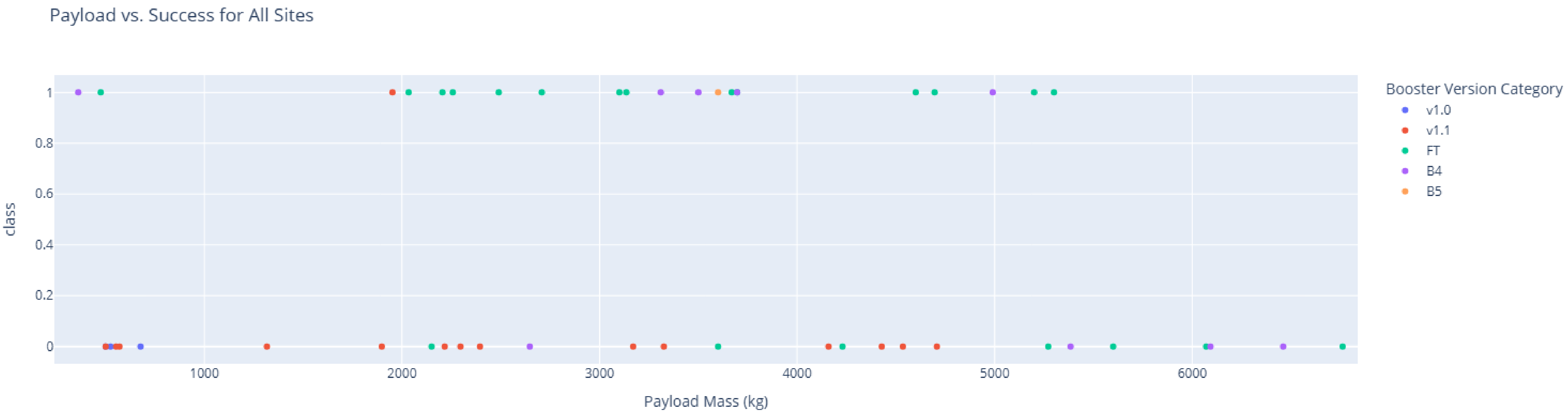
KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Success vs Failure for site KSC LC-39A

Payload Mass vs Launch Outcome for all sites

The charts show that payloads between 500 and 6800 kg have the highest success rate

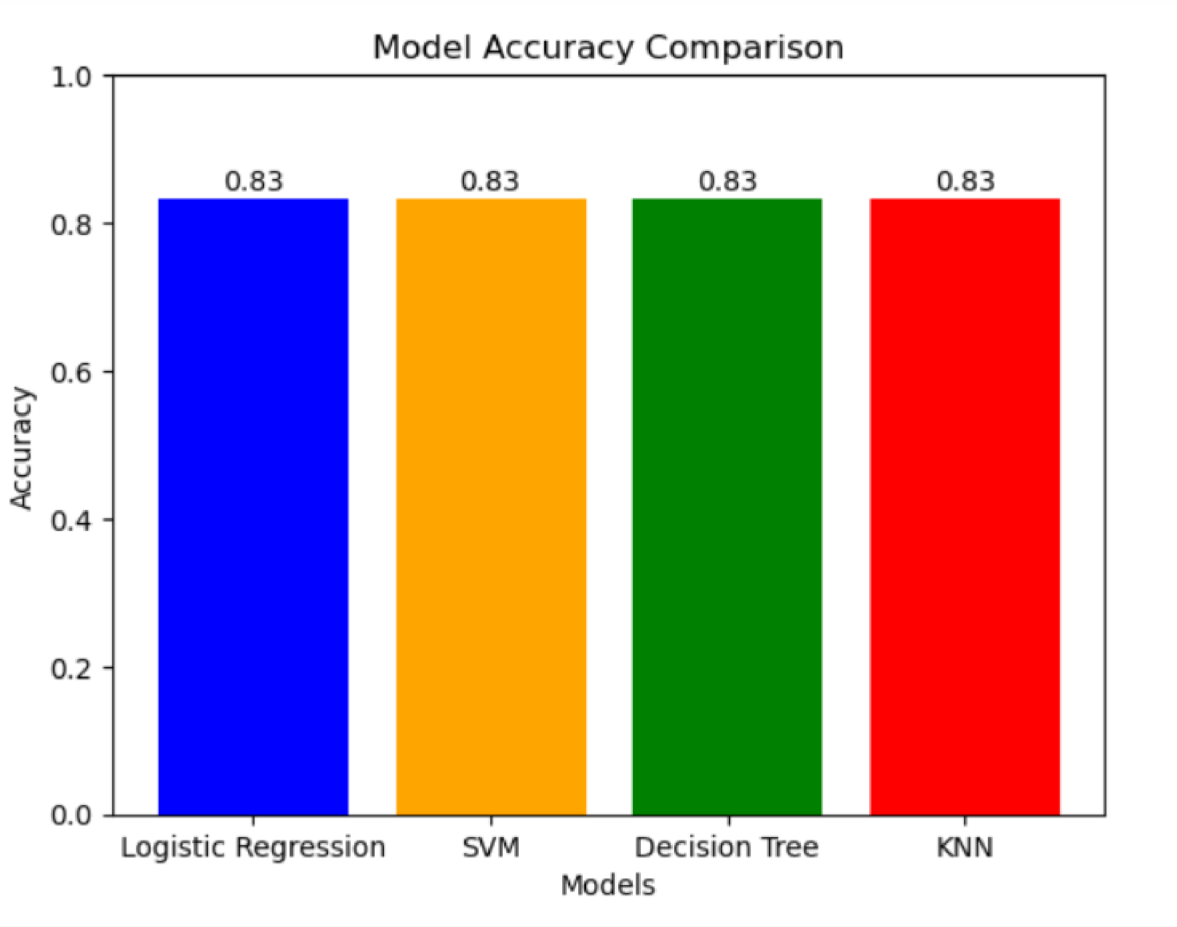


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size



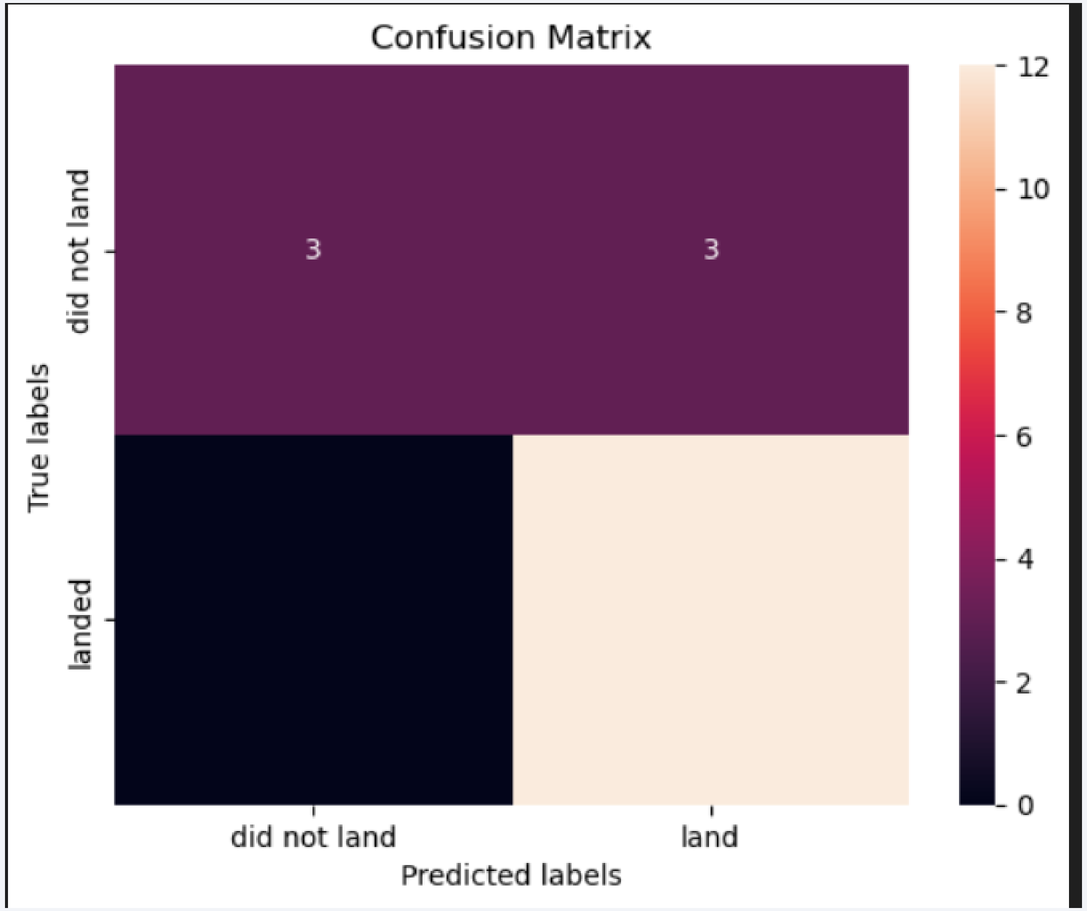
Confusion Matrix

- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the problem is false positives.

Overview:

- True Positive - 12 (True label is landed, Predicted label is also landed)

- False Positive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- Launches with a low payload mass show better results than launches with a larger payload mass.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.

Appendix

Instructor : Coursera IBM

Links :

- SpaceX launches
- IBM Datasets

Thank you!