Winning Space Race with Data Science

Dorisa Tabaku

05/05/2023

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

- **Methodology**

- Data Collection: API, Web Scraping
- EDA: Data Visualization, SQL
- Interactive Map: Folium
- Dashboard: Plotly Dash
- Predictive Analysis

- **Main Results**

- Summary of Findings
- EDA Results
- Interactive Maps and Dashboards
- Predictive Outcomes

Introduction

- Project background and context

This project aims to predict the successful landing of the Falcon 9 first stage, whichis a key factor in determining the cost of a rocket launch.

On its website, SpaceX promotes Falcon 9 rocket launches at a cost of 62 million dollars, which is significantly lower compared to other providers who charge at least 165 million dollars for each launch.

This cost advantage stems from the fact that SpaceX has the ability to reuse the first stage of the rocket.

By analyzing the characteristics of successful and failed landings and identifying the impact of rocket variables on the landing outcome, the project seeks to determine the conditions that can lead to the highest landing success rate.

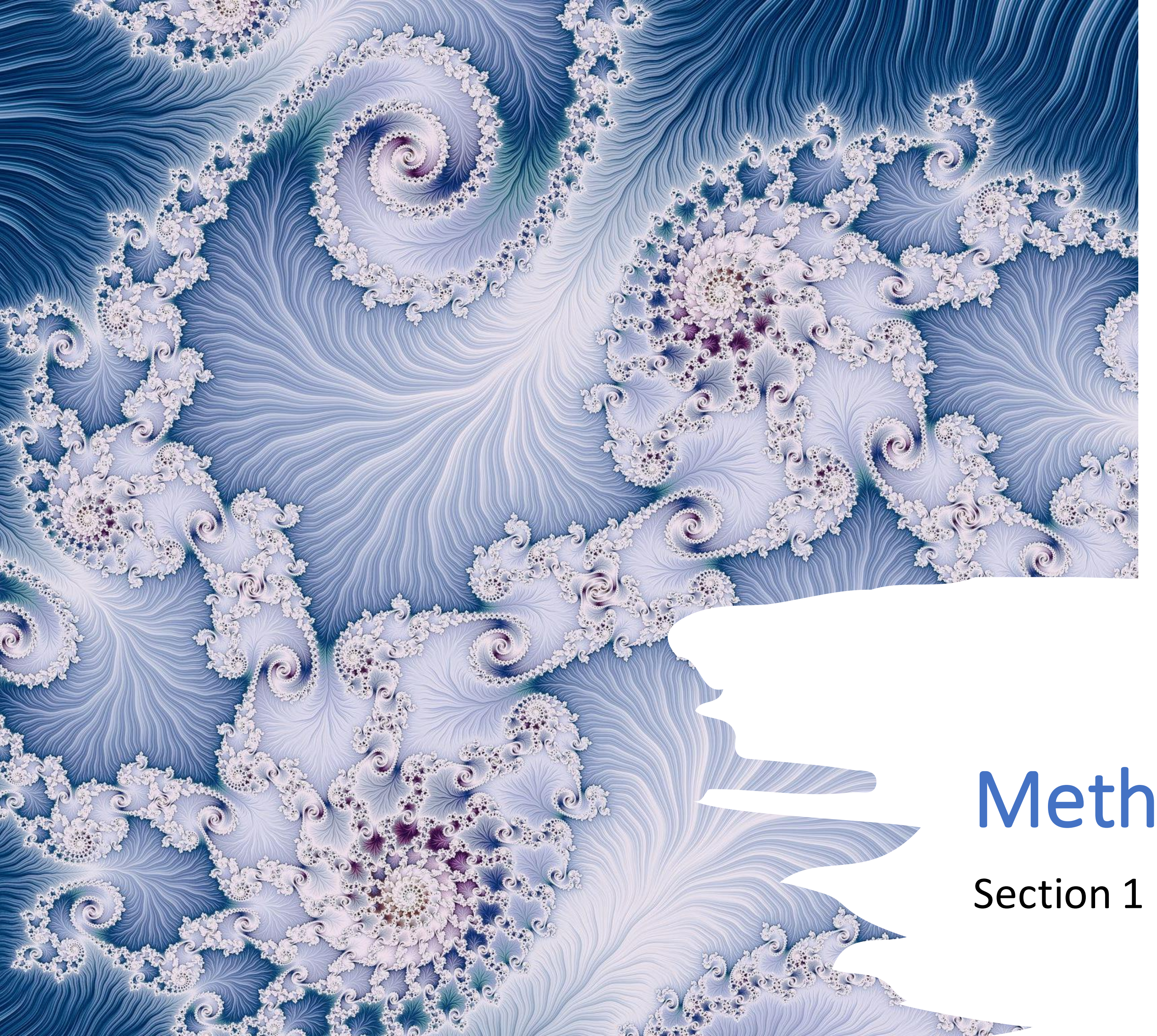
The findings can be useful for companies interested in competing with SpaceX in the rocket launch market.

- The project seeks to answer the following questions:

  - What are the main characteristics of a successful or failed landing?

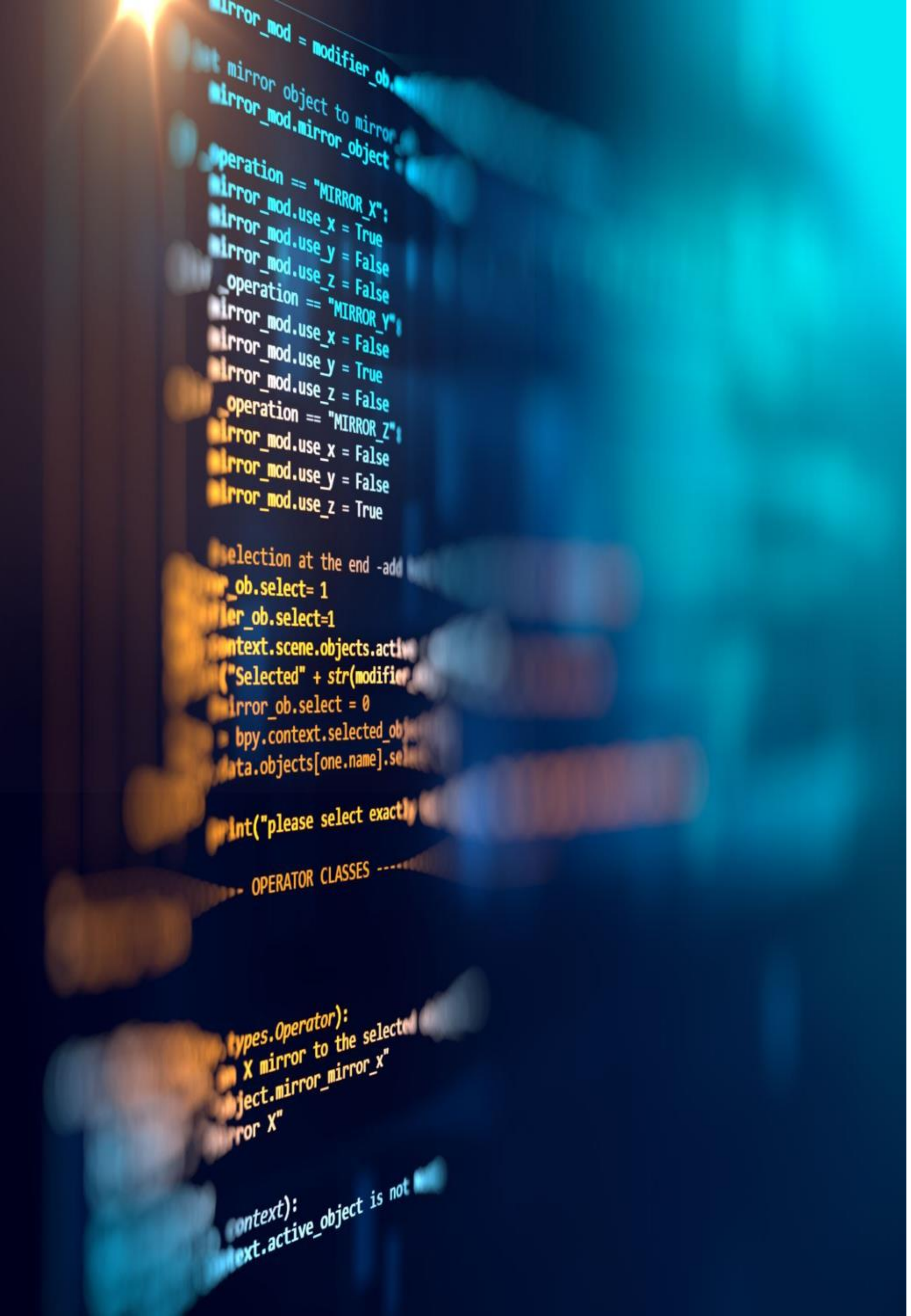
  - How do different rocket variables affect the success or failure of a landing?

  - What are the conditions that can help SpaceX achieve the highest landing success rate?



Methodology

Section 1



Methodology

- Data collection methodology:

  - Using SpaceX Rest API and web scraping from Wikipedia

- Performed data wrangling

  - Filtering the data

  - Dealing with missing values

  - Using One Hot Encoding to prepare the data for binary classification

- Performed exploratory data analysis (EDA) using visualization and SQL

- Performed interactive visual analytics using Folium and Plotly Dash

- Performed predictive analysis using classification models:

  - Building, tuning, and evaluating classification models to ensure the best results

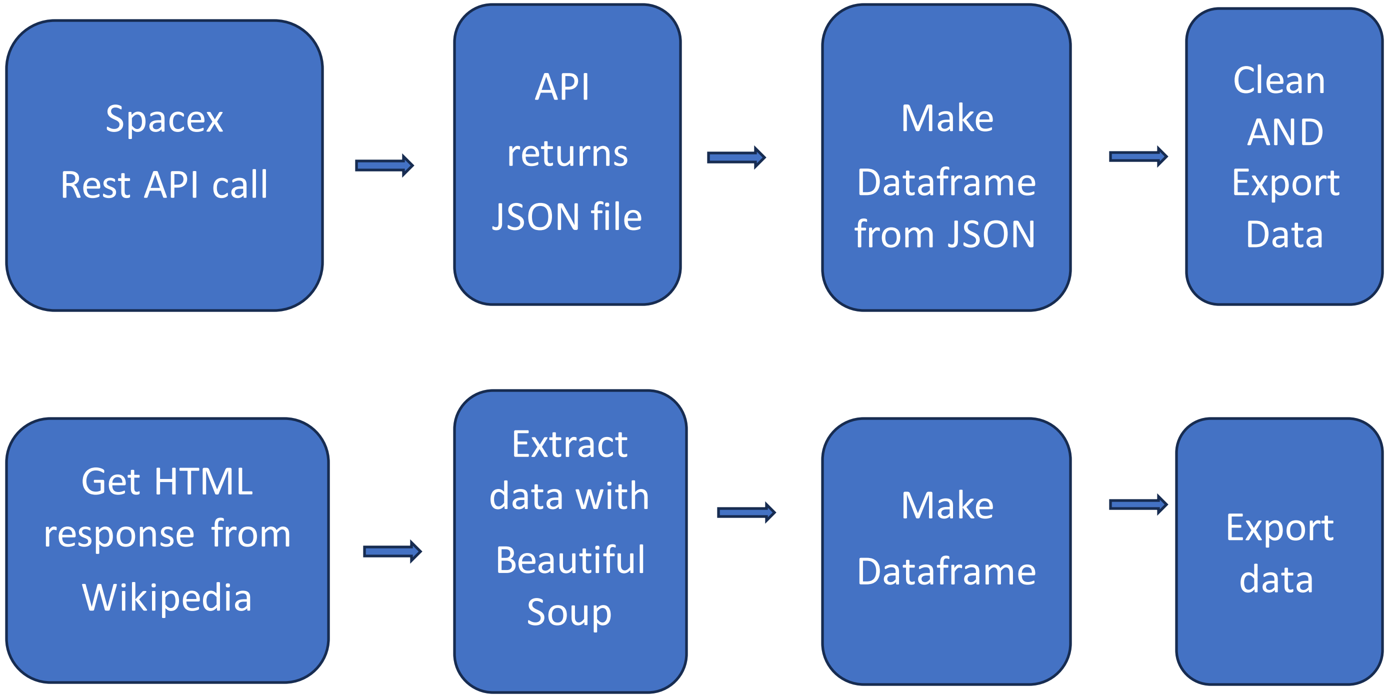
Data Collection

- To ensure a comprehensive analysis of the launches, the data collection process utilized a combination of API requests from the SpaceX REST API URL is api.spacedata.com/v4/ and web scraping data from a table in SpaceX's Wikipedia entry URL is https://en.wikipedia.org/w/index.php?title=List.

  This ensured that we obtained all the necessary data columns, such as:

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude from the Space REST API,
- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time from Wikipedia web scraping.

Flowchart of SpaceX API calls and web scraping



GitHuburl:
[SpaceX API]

[Webscraping]

Data Wrangling

The dataset includes several instances where the booster failed to land successfully.

In some cases, landing was attempted but failed due to accidents.

A mission outcome is labeled as True Ocean when the booster successfully landed in a specific region of the ocean, whereas False Ocean indicates an unsuccessful landing in the ocean.

True RTLS signifies a successful landing on a ground pad, whereas False RTLS represents an unsuccessful landing on a ground pad.

True ASDS means a successful landing on a drone ship, while False ASDS means an unsuccessful landing on a drone ship.

Our primary goal is to convert these outcomes into training labels, where a label of 1 indicates a successful landing, and a label of 0 signifies an unsuccessful landing.

EDA with Data Visualization

- Visualize the relationship between success rate of each orbit type
- Visualize the relationship between Payload and Launch Site
- Visualize the relationship between FlightNumber and Orbit type
- We can plot out the FlightNumber vs. PayloadMass and overlay the outcome of the launch.
- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.
- Observe that the sucess rate since 2013 kept increasing
- [Eda with datavs.link]

EDA with SQL

- The following SQL queries were executed as part of the exploratory data analysis:

Retrieve the names of all unique launch sites involved in the space mission.

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

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

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

Identify the date when the first successful landing was achieved on a ground pad.

List the names of boosters that successfully landed on a drone ship and carried a payload mass between 4000 and 6000.

Count the total number of successful and failed mission outcomes.

Determine the booster versions that have carried the maximum payload mass.

List all failed landing outcomes on a drone ship along with their booster versions and launch site names for the months in the year 2015.

Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates of June 4, 2010, and March 20, 2017, in descending order.

EDA with SQL

Build an Interactive Map with Folium

- The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- Finding an optimal location for building a launch site certainly involves many factors and we can discover some of the factors by analyzing the existing launch site locations.
- The following markers were added to the map for all launch sites:
- A marker with a circle, popup label, and text label was added for NASA Johnson Space Center, using its latitude and longitude coordinates as the starting location.
- Markers with circles, popup labels, and text labels were added for all launch sites to show their geographical locations, as well as their proximity to the Equator and coastlines.
- Additionally, markers were colored to represent launch outcomes for each launch site:
- Colored markers were added using Marker Cluster to show success (green) and failed (red) launches. This helps to identify launch sites that have relatively high success rates.
- To show the distances between a launch site and its proximities, colored lines were added to the map:
- As an example, colored lines were added to show the distance between KSC LC-39A launch site and its proximities such as railway, highway, coastline, and the closest city.
- Now, we can easily show all launch sites, their surroundings and the number of successful and unsuccesful landings.
- githublink:Folium

Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, rangeslider and scatter plot components
- A dropdown list was added to enable the selection of a Launch Site.
- A pie chart was added to show the total count of successful launches for all sites, as well as the Success vs. Failed counts for a specific Launch Site if selected.
- A slider was added to select the Payload Mass range.
- A scatter chart was added to show the correlation between Payload and Launch Success for different Booster Versions
- Plotly Dash app

Predictive Analysis (Classification)

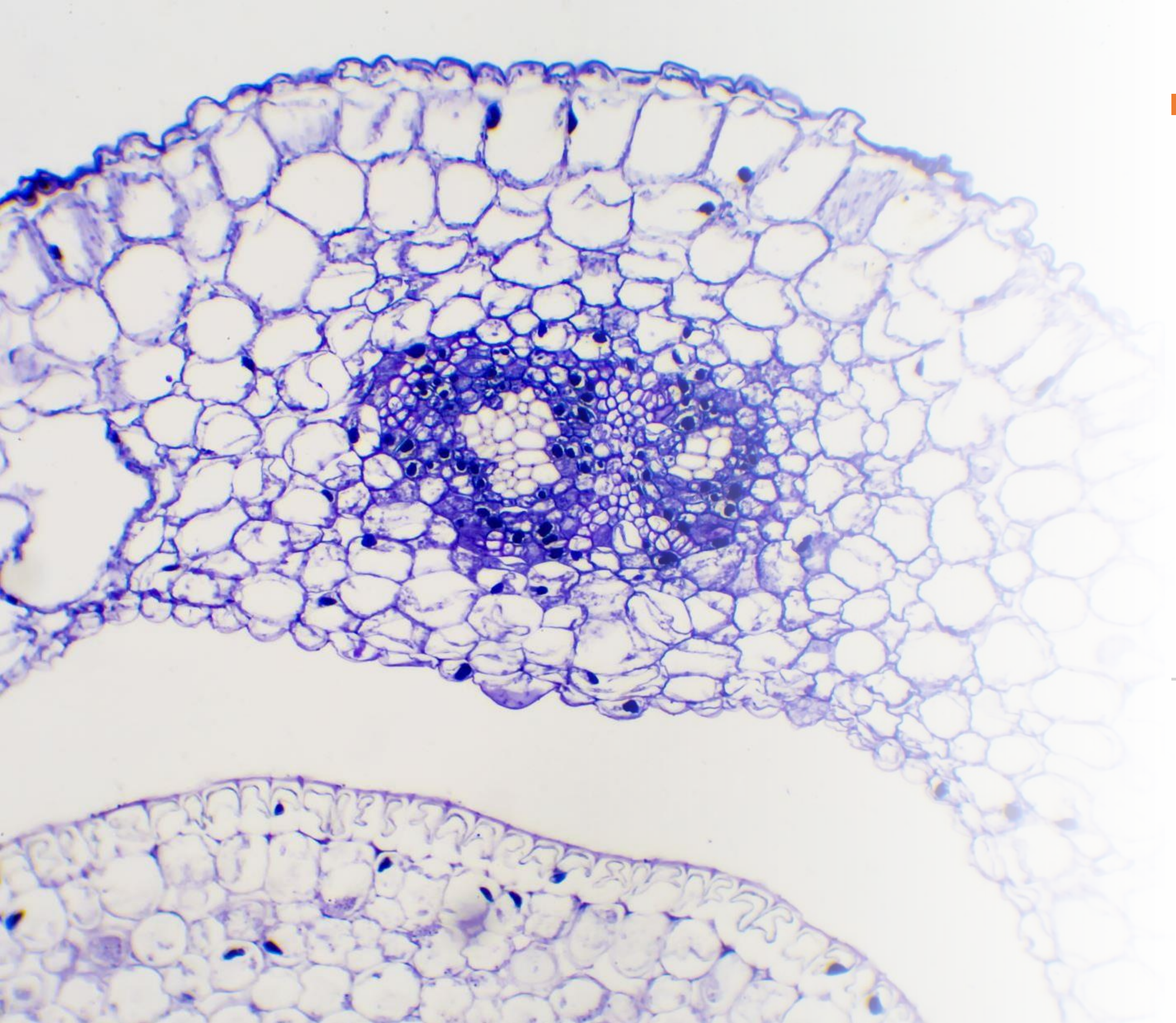
- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model Evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen Prediction

Results

Exploratory data analysis results

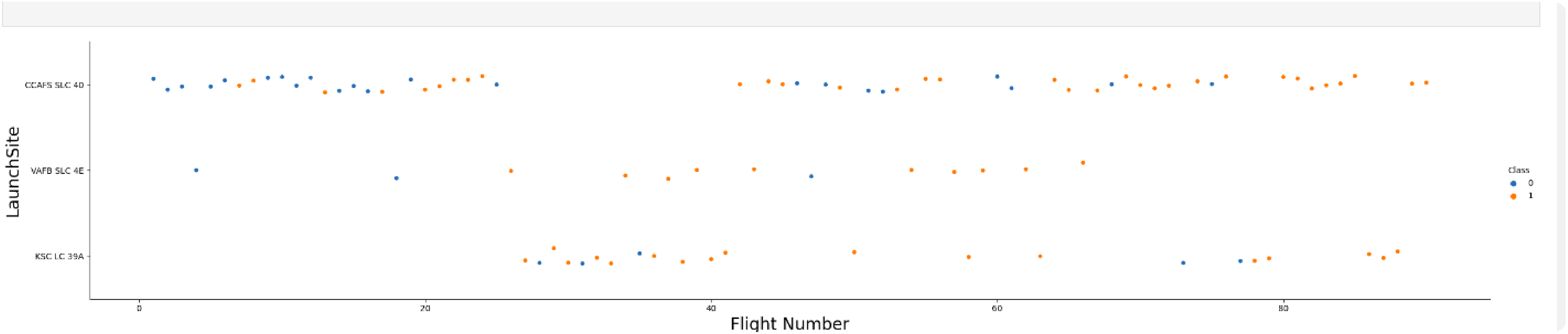
Interactive analytics demo in screenshots

Predictive analysis results



Insight Drawn from EDA

Section2



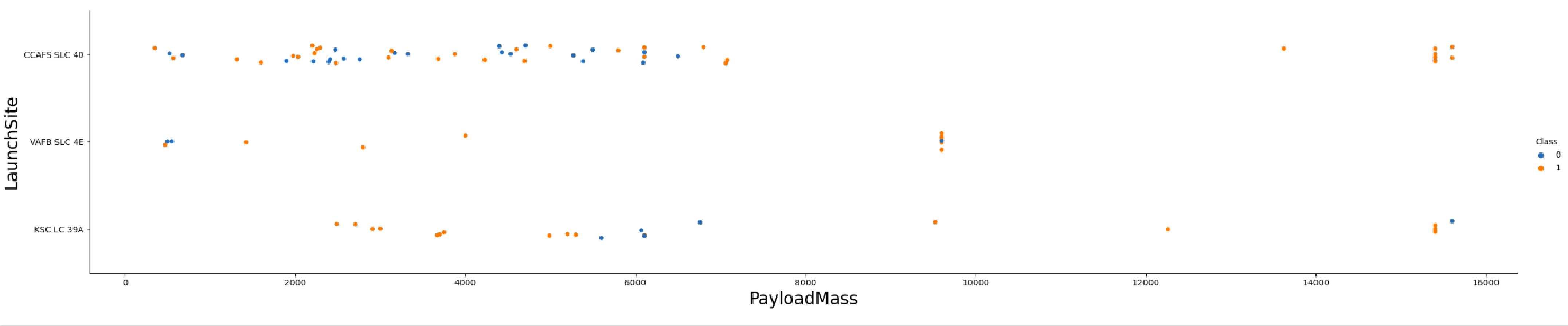
Flight Number vs Launch Site

According to the analysis, the initial space flights had a higher rate of failure, while the most recent launches have mostly been successful.

The CCAFS SLC 40 launch site has been the most frequently used location for space missions, accounting for about half of all launches.

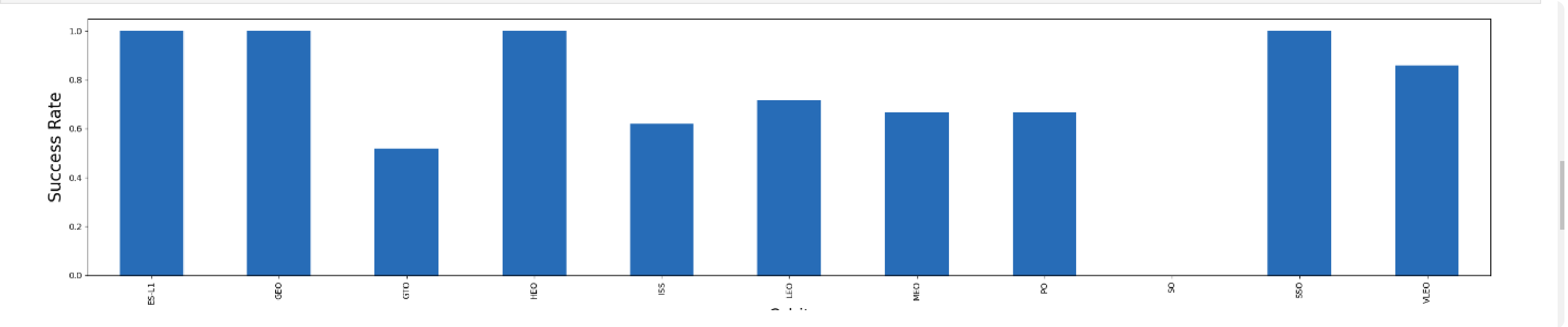
VAFB SLC 4E and KSC LC 39A have higher success rates, suggesting that they are well-equipped and optimized for successful launches.

We observe the trend of increasing success rates with each new launches.



Payload vs. Launch Site

- There is a positive correlation between the payload mass and the launch success rate for every launch site. The higher the payload mass, the higher the success rate, suggesting that there may be a relationship between the launch site's capabilities and its ability to handle higher payload masses.
- Most launches with a payload mass over 7000 kg had a successful outcome. This may indicate that the launch site's infrastructure and technology are optimized for handling larger payloads.
- KSC LC 39A had a 100% success rate for payload mass under 5500 kg.It suggests that this launch site is particularly well-suited for handling smaller payloads.



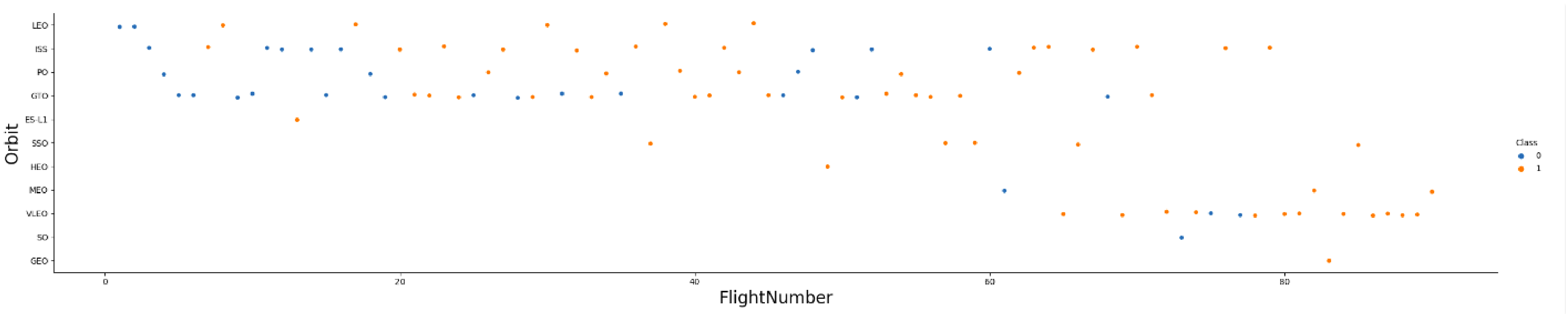
Success Rate vs. Orbit Type

Orbits with 100% success rate:

- ES-L1, GEO, HEO, SSO

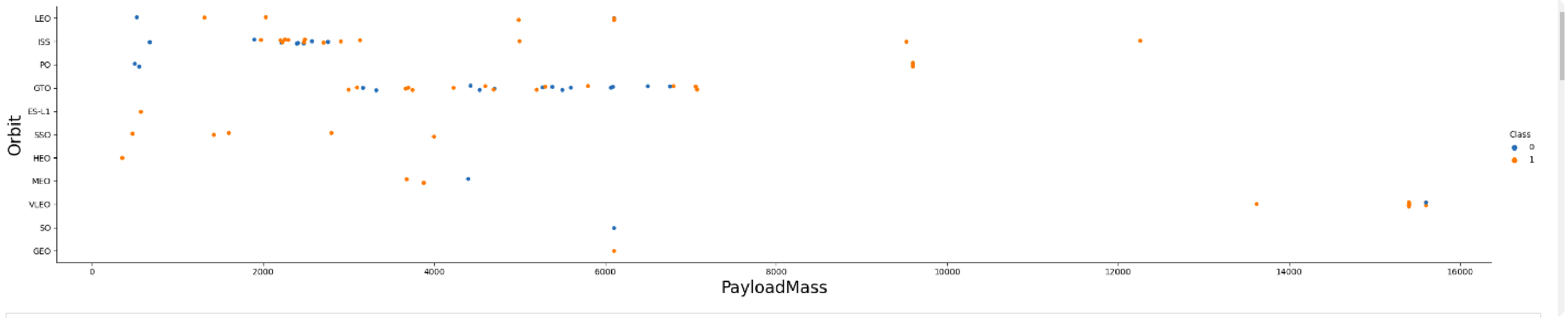
Orbits with 0% success rate:

- SO



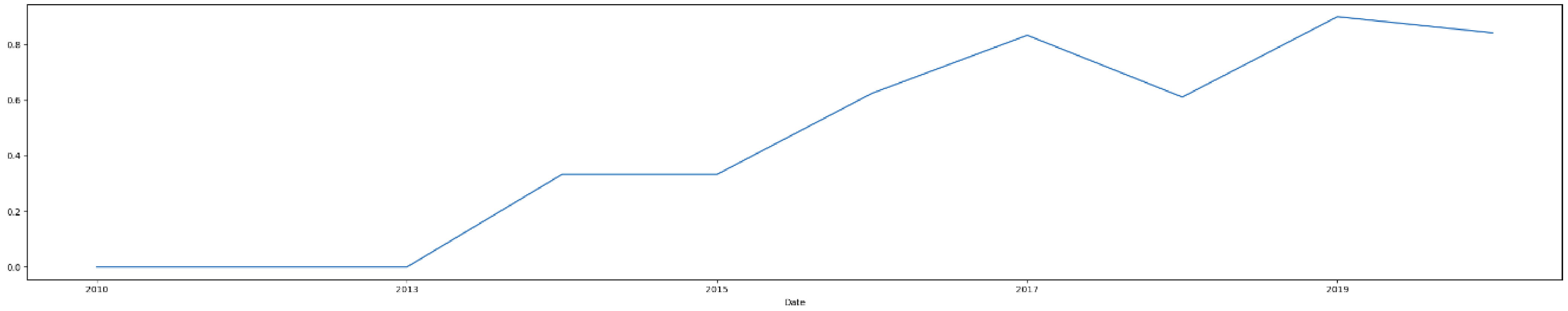
Flight Number vs. Orbit Type

We can observe that there is a correlation between the number of flights and the success rate for the LEO orbit, indicating that as more flights are conducted, the success rate tends to increase. However, for other orbits like GTO, there doesn't seem to be any relation between the number of flights and success rate. Despite this, it's possible that the high success rate for orbits like SSO or HO is due to knowledge gained from previous launches of different orbits, suggesting that experience and learning from past missions can improve success rates for future launches.



Payload vs. Orbit Type

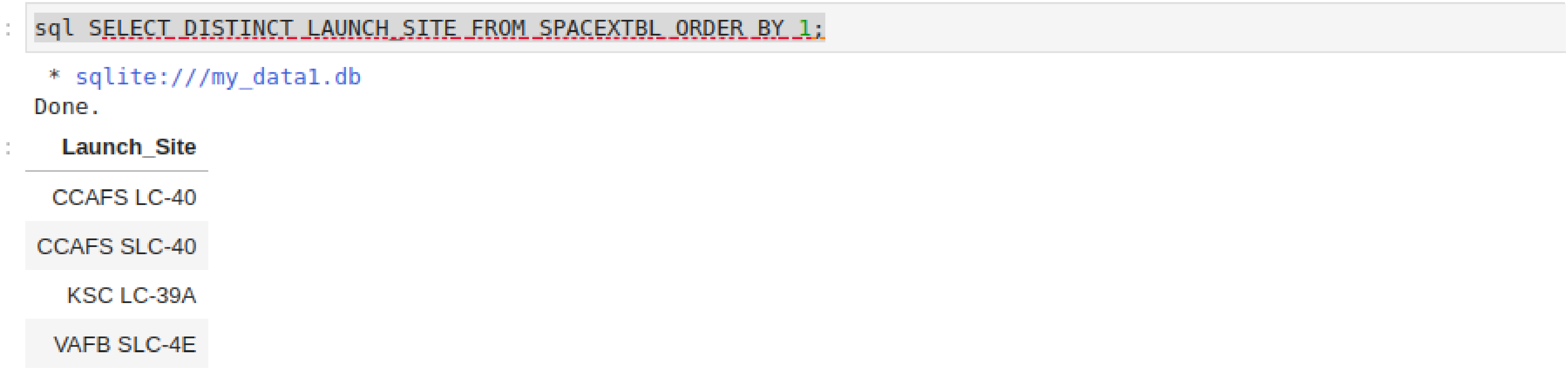
The weight of the payloads can have a great influence on the success rate of the launches in certain orbits. For example, heavier payloads improve the success rate for the LEO orbit. Another finding is that decreasing the payload weight for a TO orbit improves the success of a launch.



Launch Success Yearly Trend

- The success rate kept increasing from 2013 to 2020

All Launch Site Names



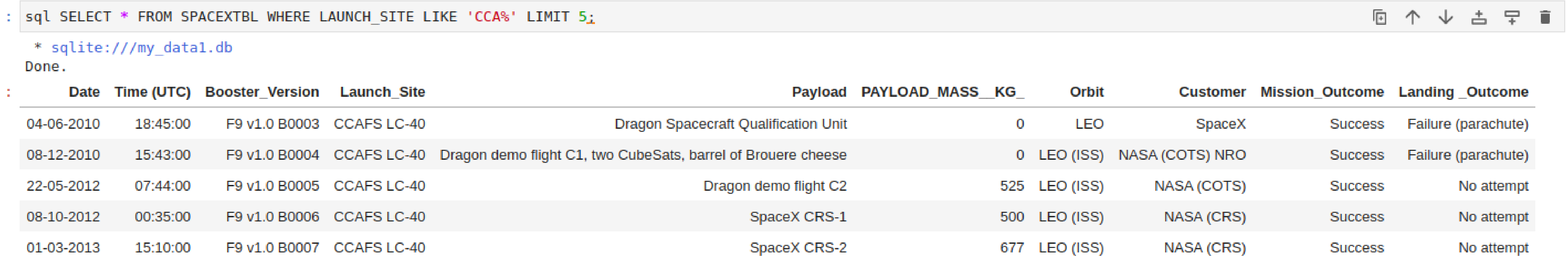
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

- Four distinct launch sites:

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



```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

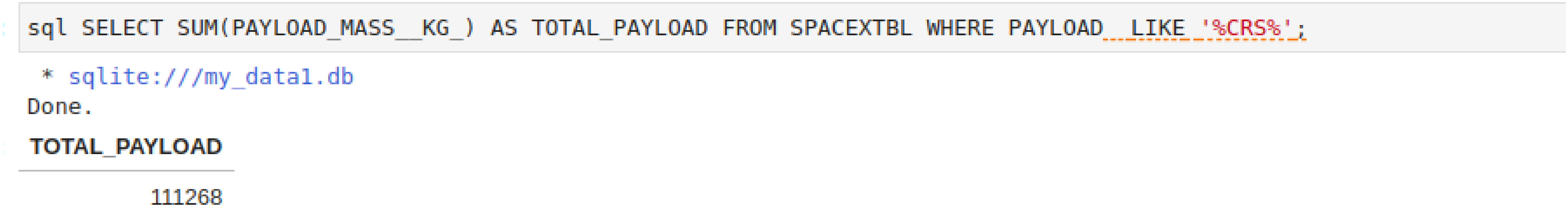
```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- CCAFS LC-40

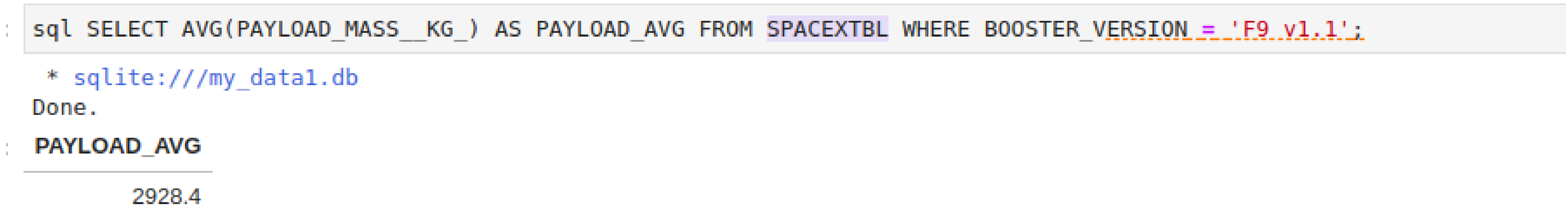
Total Payload Mass



```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

```
 * sqlite:///my_data1.db
Done.
TOTAL_PAYLOAD
       111268
```

- The total payload mass(kg) is 111268(kg).



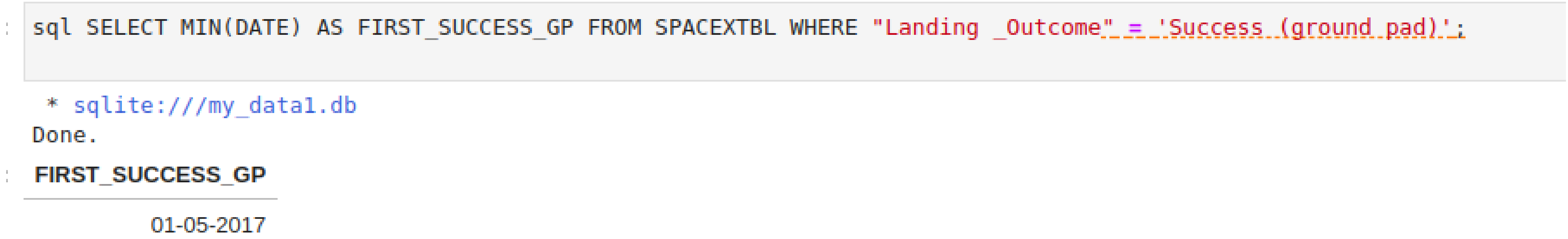
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS PAYLOAD_AVG FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

```
PAYLOAD_AVG
     2928.4
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4(kg).



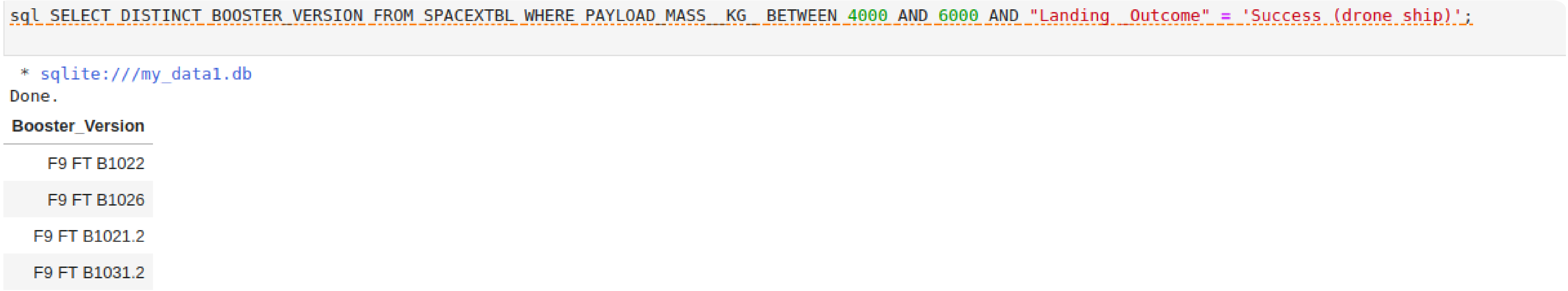
```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
01-05-2017

First Successful Ground Landing Date

- The oldest successful ground landing is on 01.05.2017.



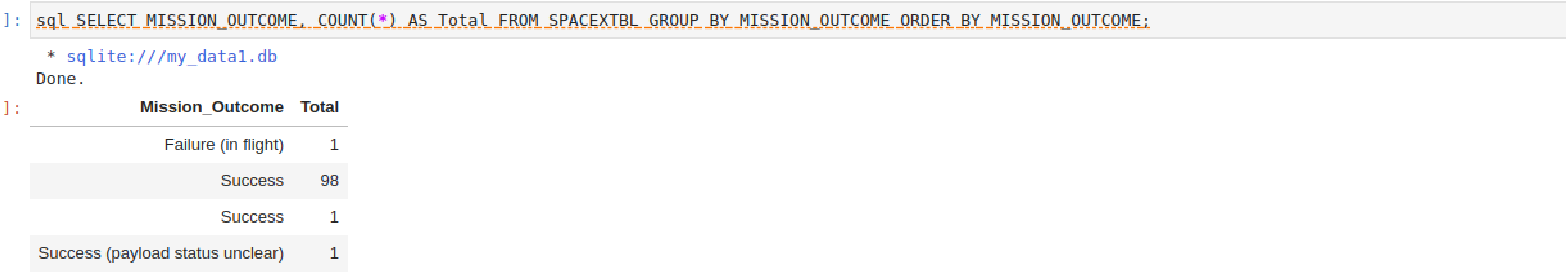
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND "Landing _Outcome" = 'Success (drone ship)';
```

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

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- The most succesful returns the booster version where landing was succesful and payload mass is between 4000 and 6000.



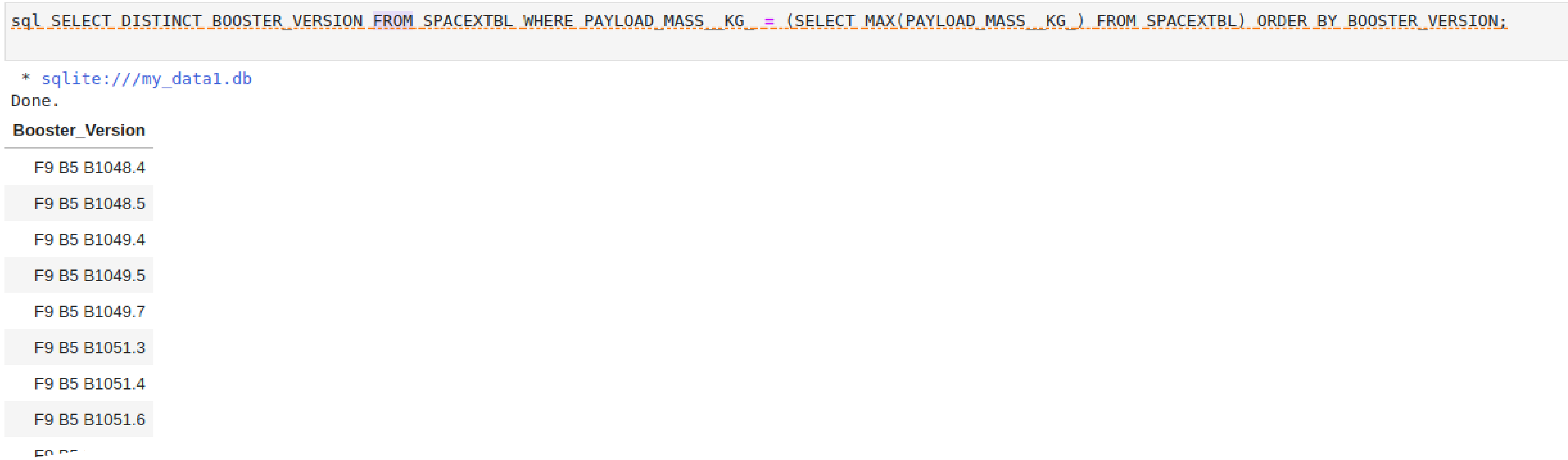
```
]: sql SELECT MISSION_OUTCOME, COUNT(*) AS Total FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

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

]:

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

Total Number of Successful and Failure Mission Outcomes

- 100 successful missions and 1 failure.



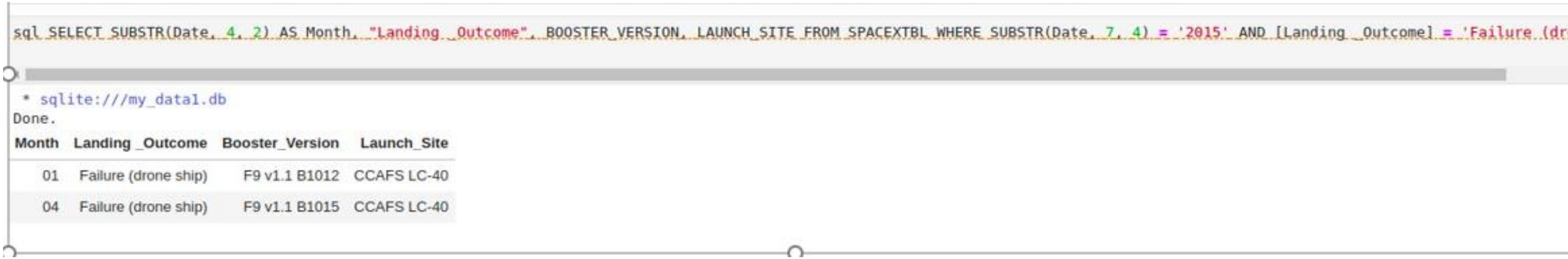
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5

Boosters Carried Maximum Payload

- Diplays a list of boosters carried max payload mass(kg)



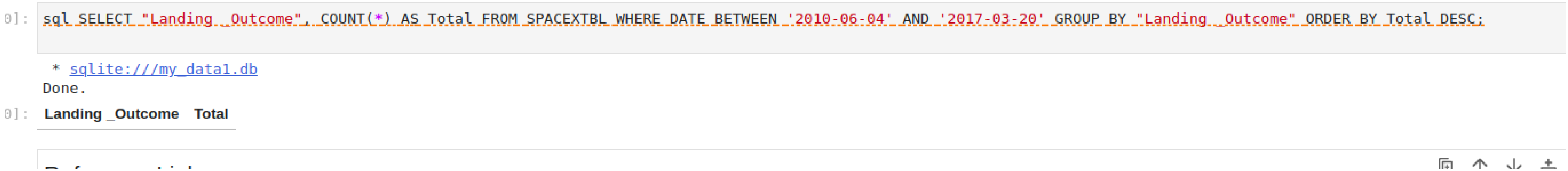
```
sql SELECT SUBSTR(Date, 4, 2) AS Month, "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE SUBSTR(Date, 7, 4) = '2015' AND [Landing _Outcome] = 'Failure (dr
```

```
 * 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

2015 Launch Records

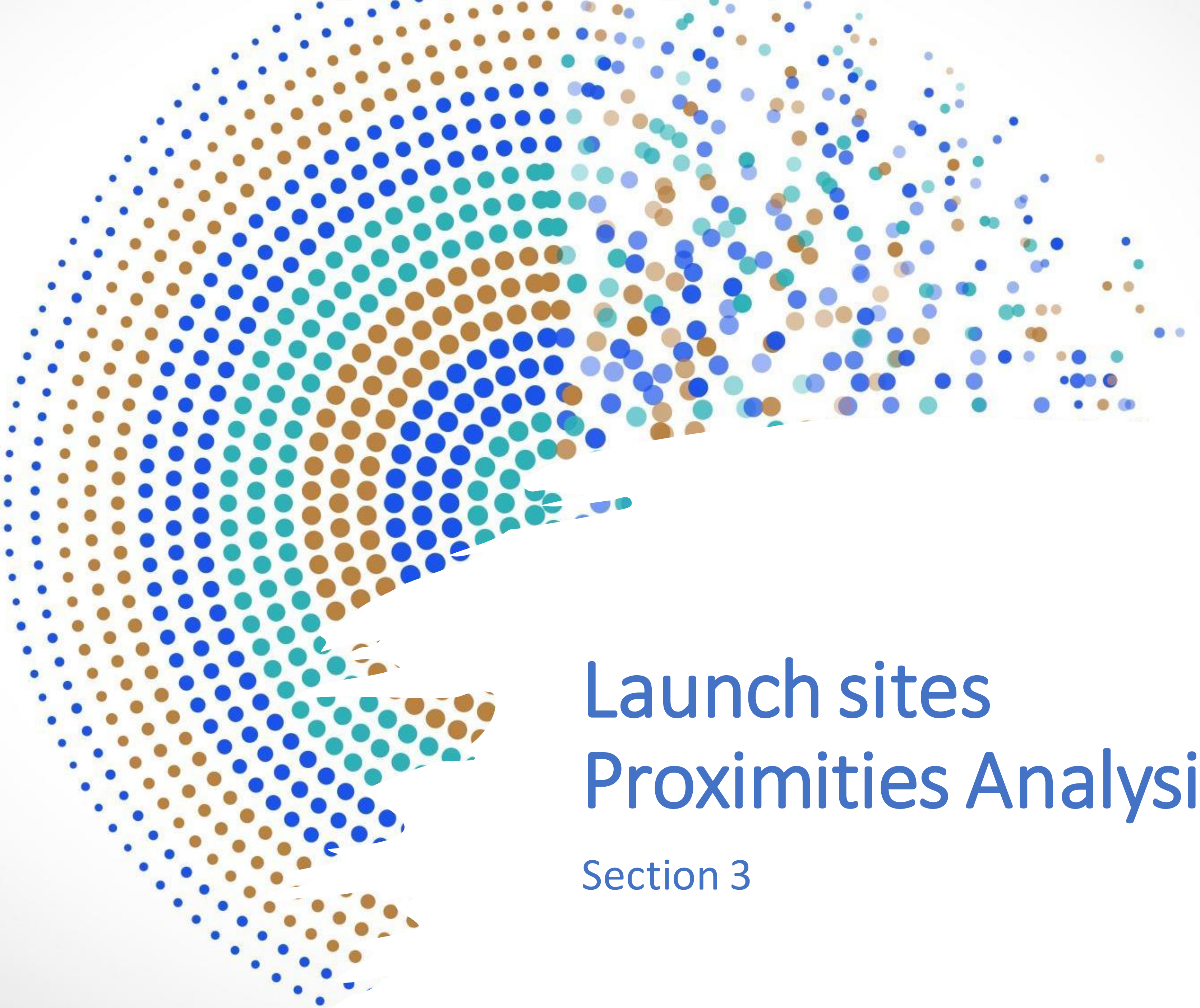
- Displays the month of the failure, failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



```
0]: sql SELECT "Landing _Outcome", COUNT(*) AS Total FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing _Outcome" ORDER BY Total DESC;

   * sqlite:///my_data1.db
  Done.
0]: Landing _Outcome   Total
```

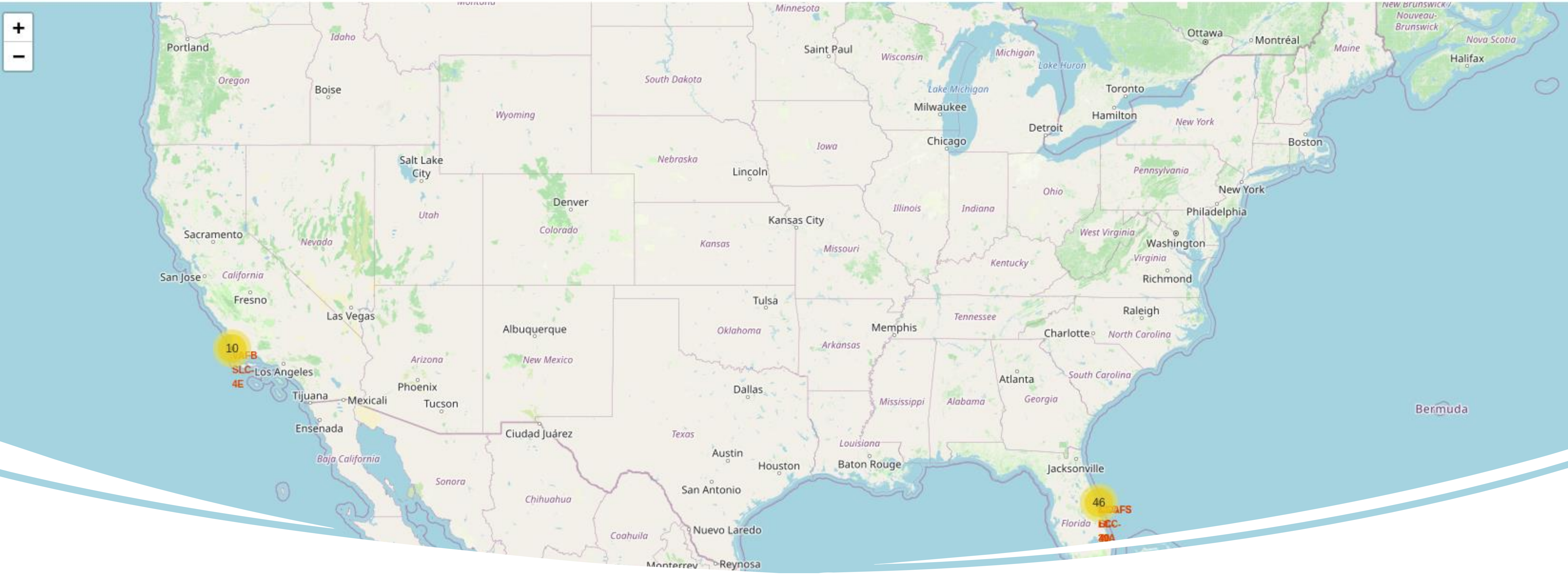
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Displays the landing outcome and the year (none) so we didn’t have any landing in this period.



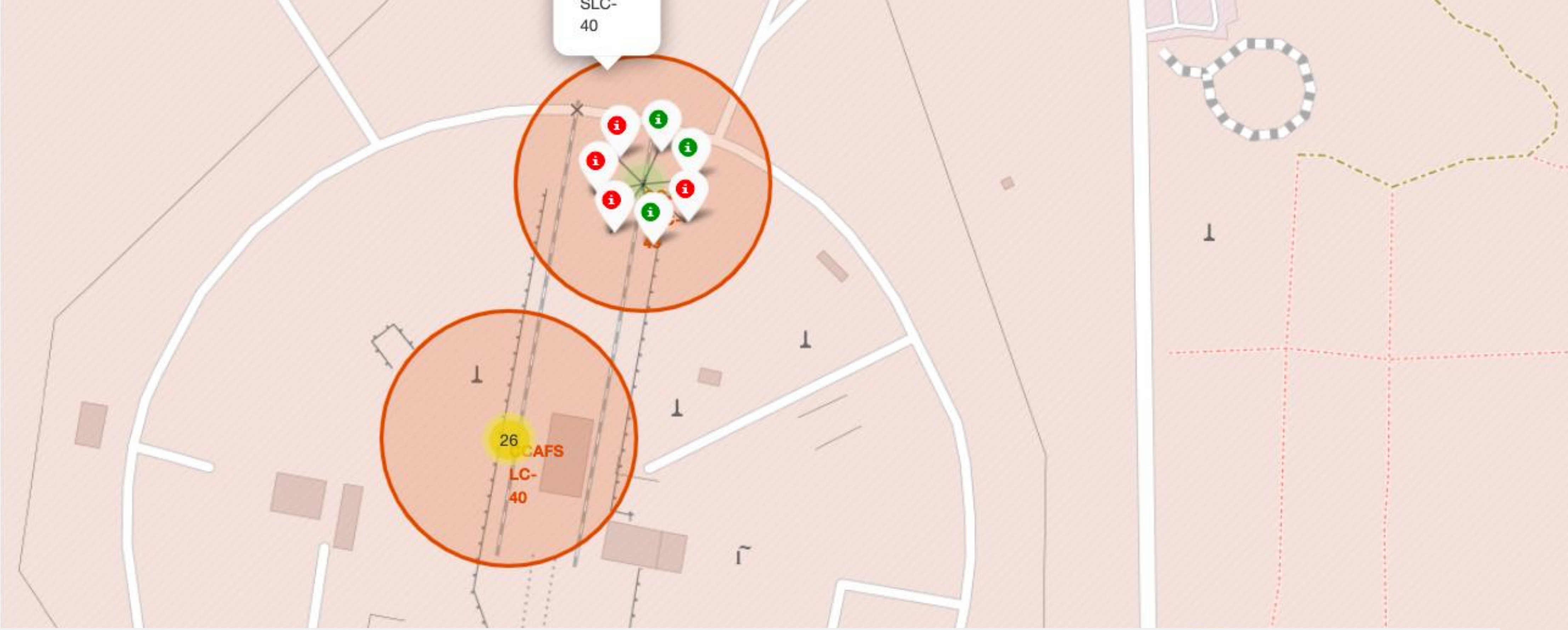
Launch sites Proximities Analysis

Section 3



Creating Interactive Maps with Folium

- Mark all launch sites on the map



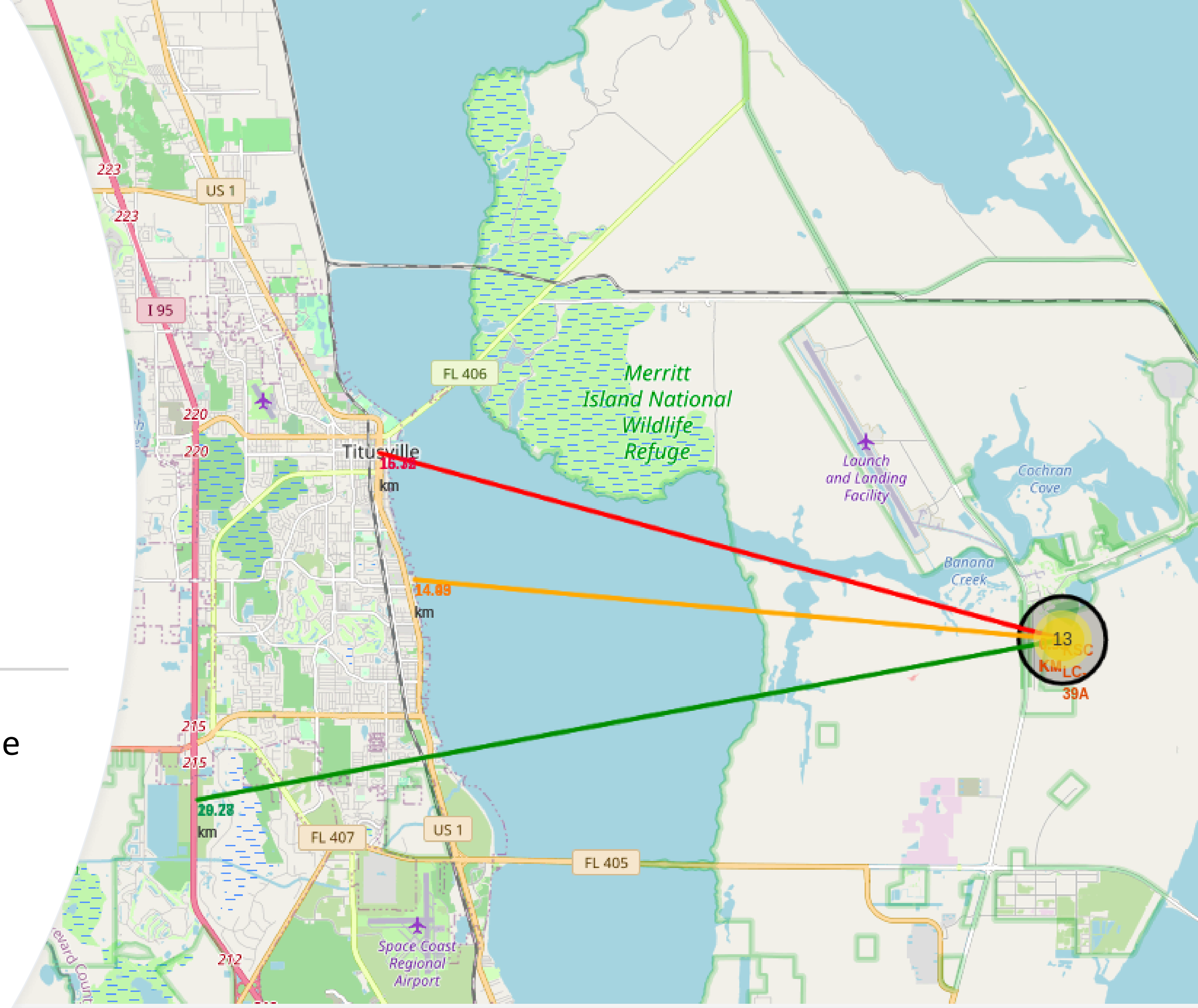
From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

Folium Map Markers

- Green in success, red in failure
- From the color-labeled markers in marker clusters, we are able to identify which launch sites have relatively high success rates.

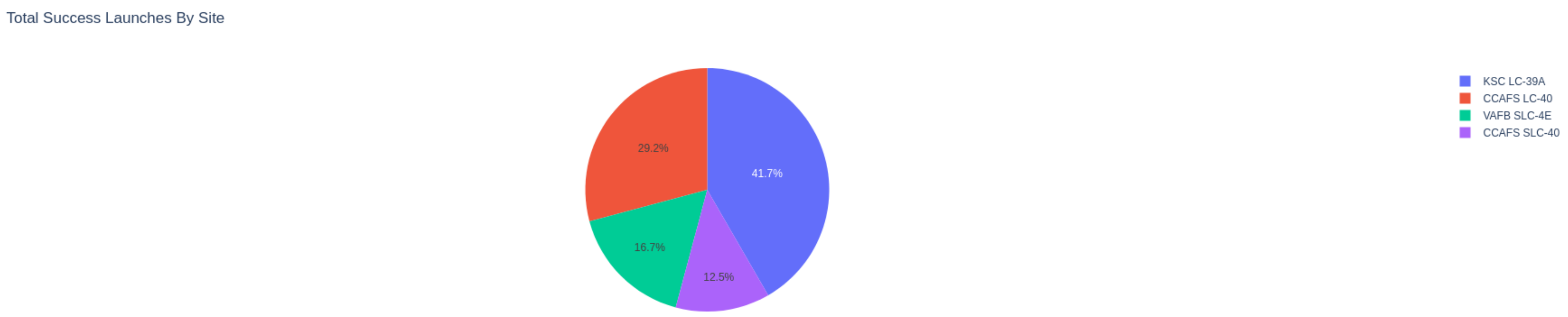
Distance between KSC LC-39A and it proximities

As we can see KSC LC-39A its close to its proximities as railways, highways and cities.



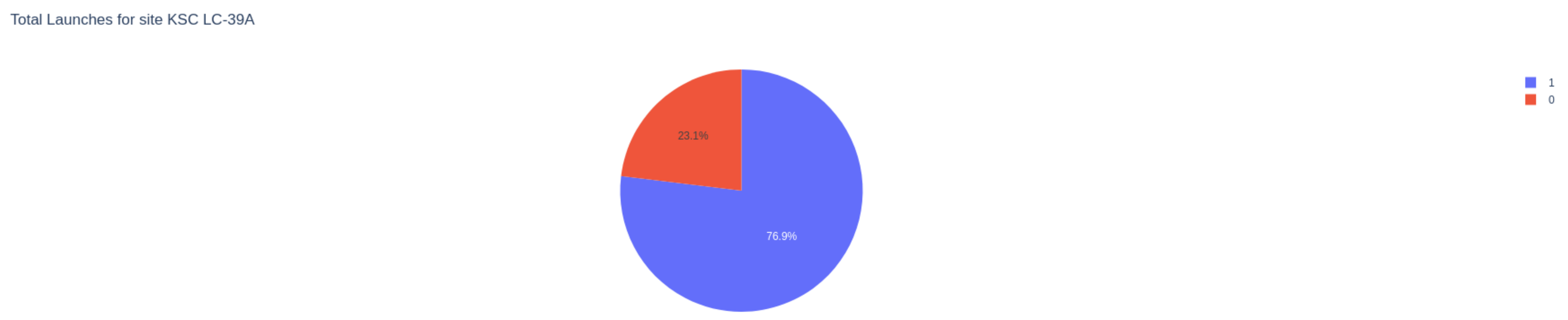
Build a Dashboard with Plotly

Section 4

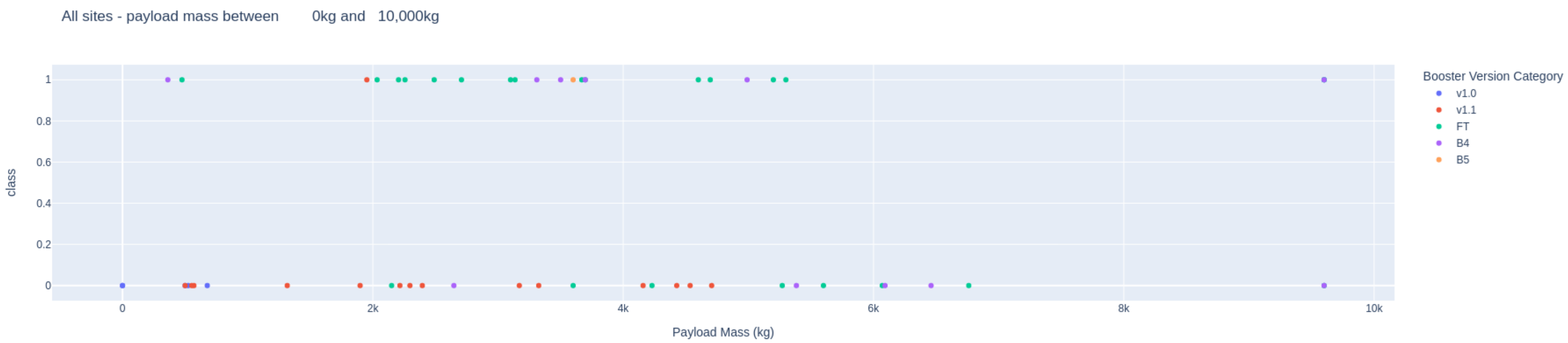


Successful Launches by Site

KSC LC-39A has the most succesful launches



KSC LC-39A Successful launches



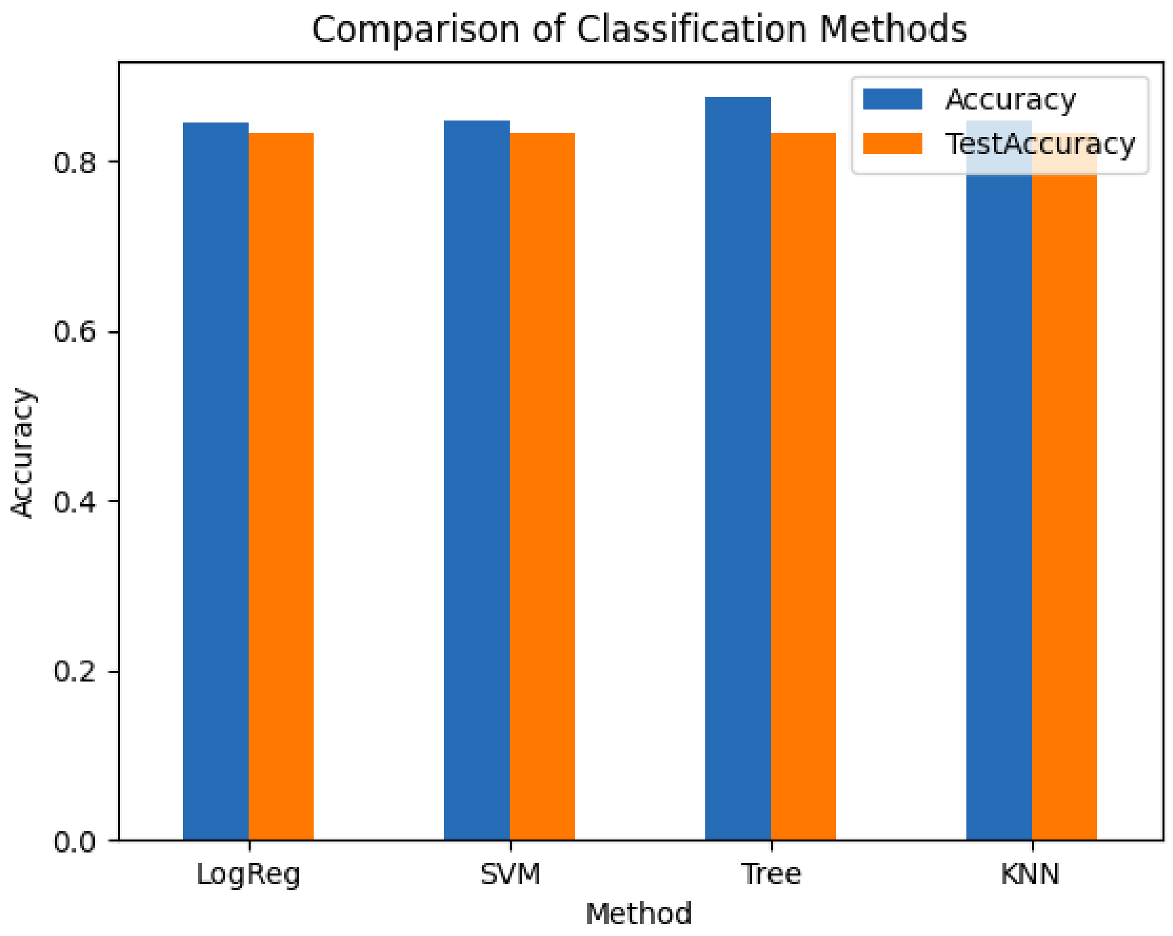
Payload vs Success rate for all sites

Predictive Analysis Classification

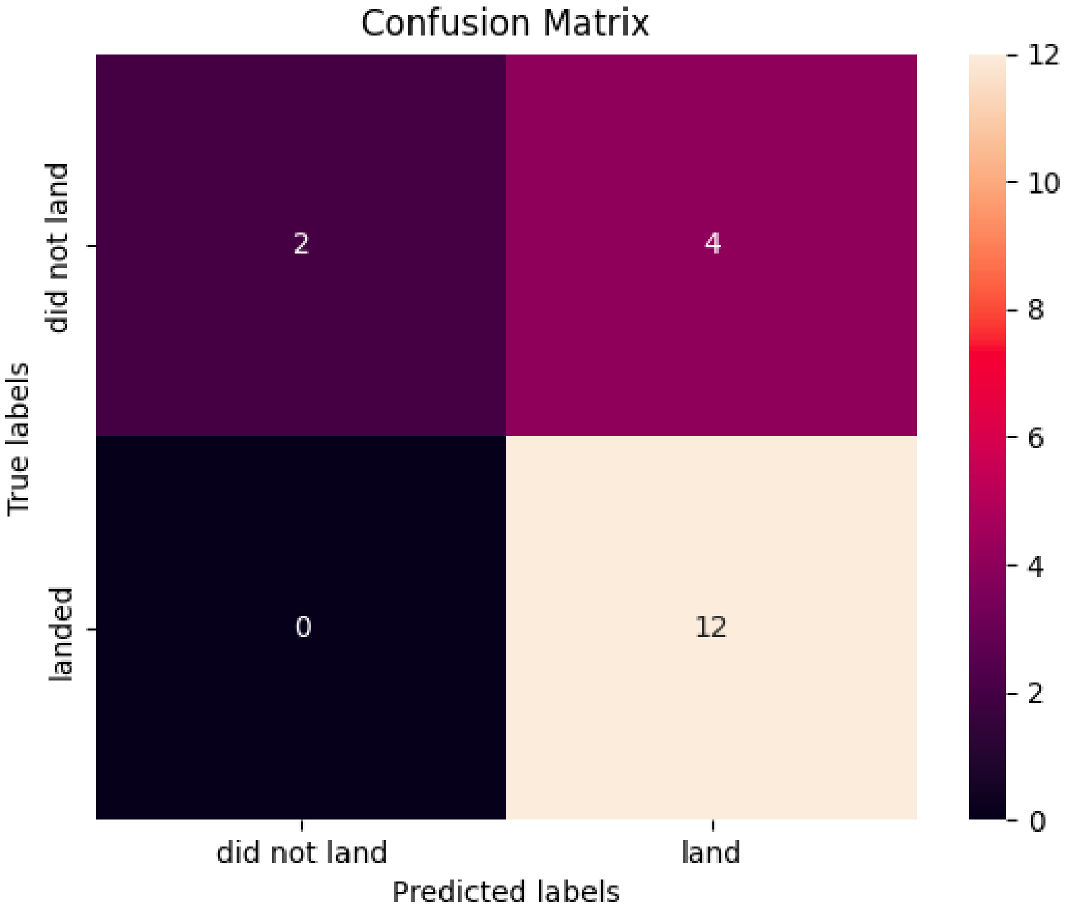
Section4

Comparasion of Classification Methods

- The method that performs best using the test data is Decision Tree with an accuracy of 0.87

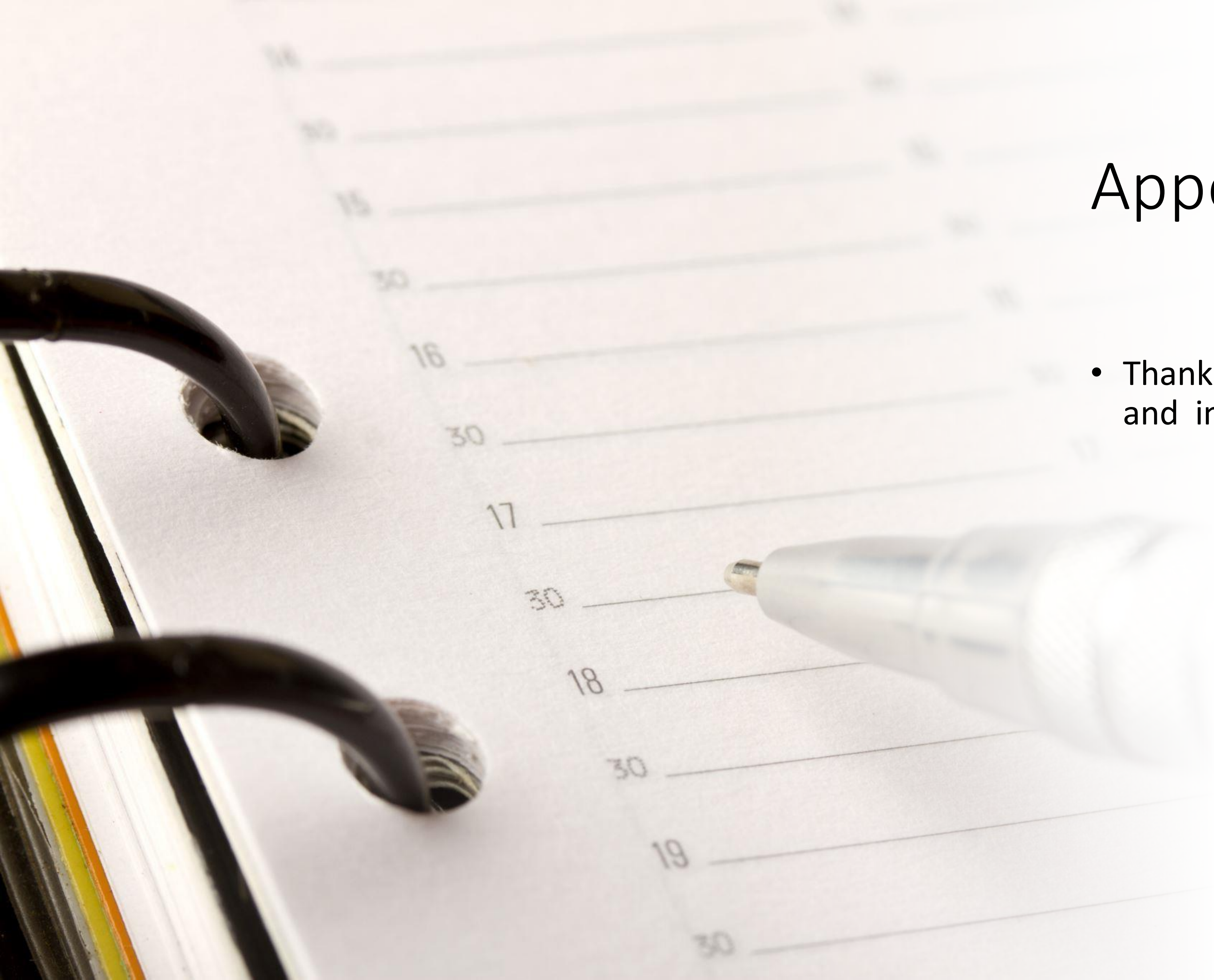


Confusion Matrix of Decision Tree



Conclusions

- The decision tree model is the best algorithm for predicting the success of space launches, which can help improve the accuracy of launch predictions and inform decision-making for future missions.
- Launches with lower payload mass tend to have better success rates, which could be a potential strategy to increase launch success.
- Launch sites located close to the Equator and the coast are preferred due to the reduced fuel requirement and favorable trajectory for achieving orbit.
- The success rate of space launches has increased over time due to advancements in technology and improved understanding of the risks associated with space launches.
- KSC LC-39A is the most reliable launch site with the highest success rate, which could be useful for space agencies in selecting a launch site for future missions.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate, suggesting that they are relatively safe and preferred for future missions.



Appendix

- Thank you to Coursera and instructors for the course.

Thank you