IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Agnieszka Haja
1.09.2023

Outline

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

Executive Summary

Summary of Methodologies:

I gathered data from both a public SpaceX API and scraped SpaceX's Wikipedia page. After preprocessing and labeling, I stored it in a DB2 SQL database. I utilized various machine learning algorithms, resulting in an 83% accuracy rate for predicting successful Stage 1 landings.

Summary of All Results:

The machine learning model I developed, with an 83% accuracy rate, is a valuable tool for SpaceY. It assists in evaluating the likelihood of successful launches, potentially saving up to $100 million USD per launch. While my visualization dashboard enhances decision-making, I recommend ongoing data collection to further refine and enhance model accuracy.

Introduction

- In this capstone project, my main goal is to predict the successful landing of the Falcon 9 first stage, a critical aspect of SpaceX's cost-efficient rocket launches. SpaceX advertises Falcon 9 launches for $62 million, significantly cheaper than competitors who charge over $165 million due to SpaceX's innovative first stage reuse. Accurate landing predictions can determine launch costs and support competitive bidding. I'll collect and format data from the SpaceX API to develop this model.

In this project I want to find answers to questions:

Can I create a precise model to predict Falcon 9 first stage landings, crucial for SpaceX's cost-effective launches?

How can I use this model to estimate launch costs based on first stage landing outcomes?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Colected data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classifying true landings as successful and unsuccessful

Data Collection

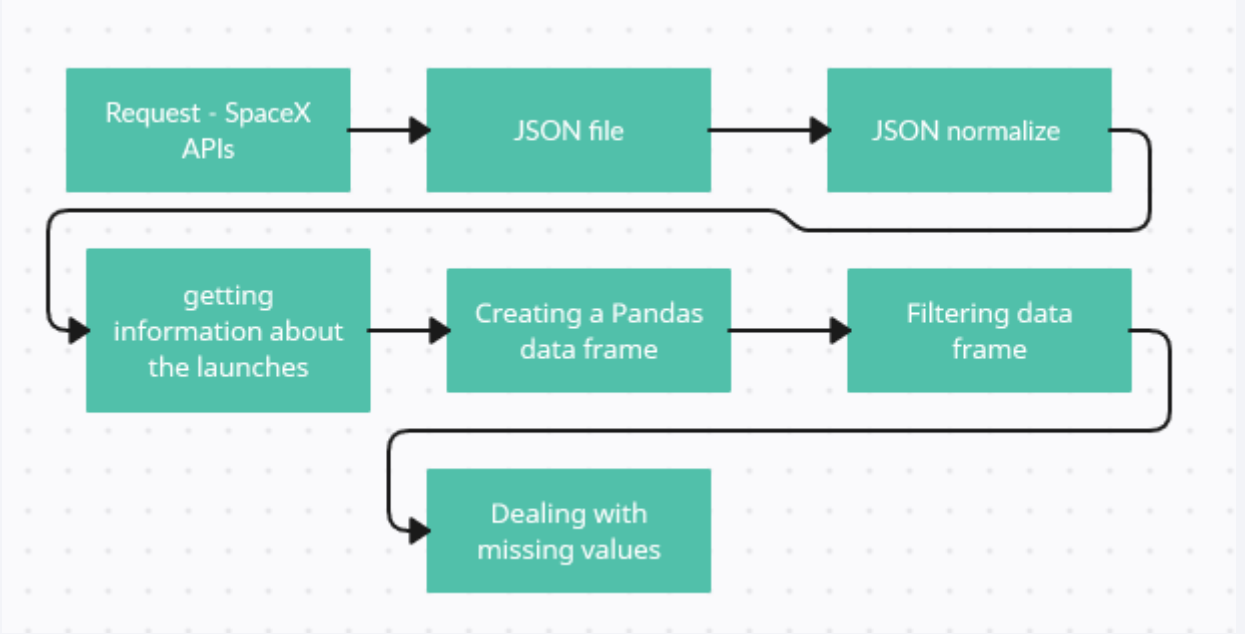
The data acquisition process involved a combination of obtaining information through API requests from SpaceX's public API and extracting data by web scraping a table found in SpaceX's Wikipedia page.

Next I will illustrate the data collection process from the API, while the subsequent slide will depict the data collection procedure through web scraping.

Columns obtained from SpaceX's API data include:
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

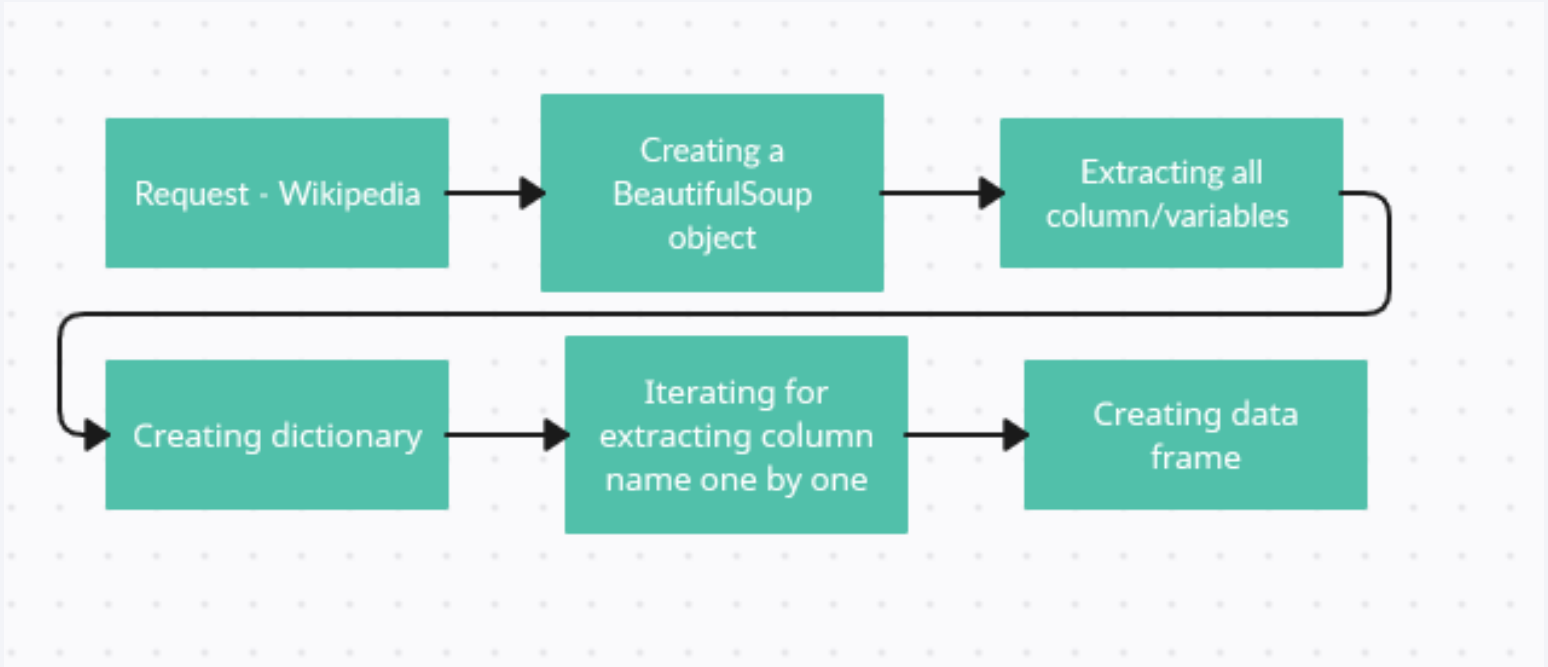
In contrast, the data columns extracted through web scraping from SpaceX's Wikipedia page encompass:
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection – SpaceX API



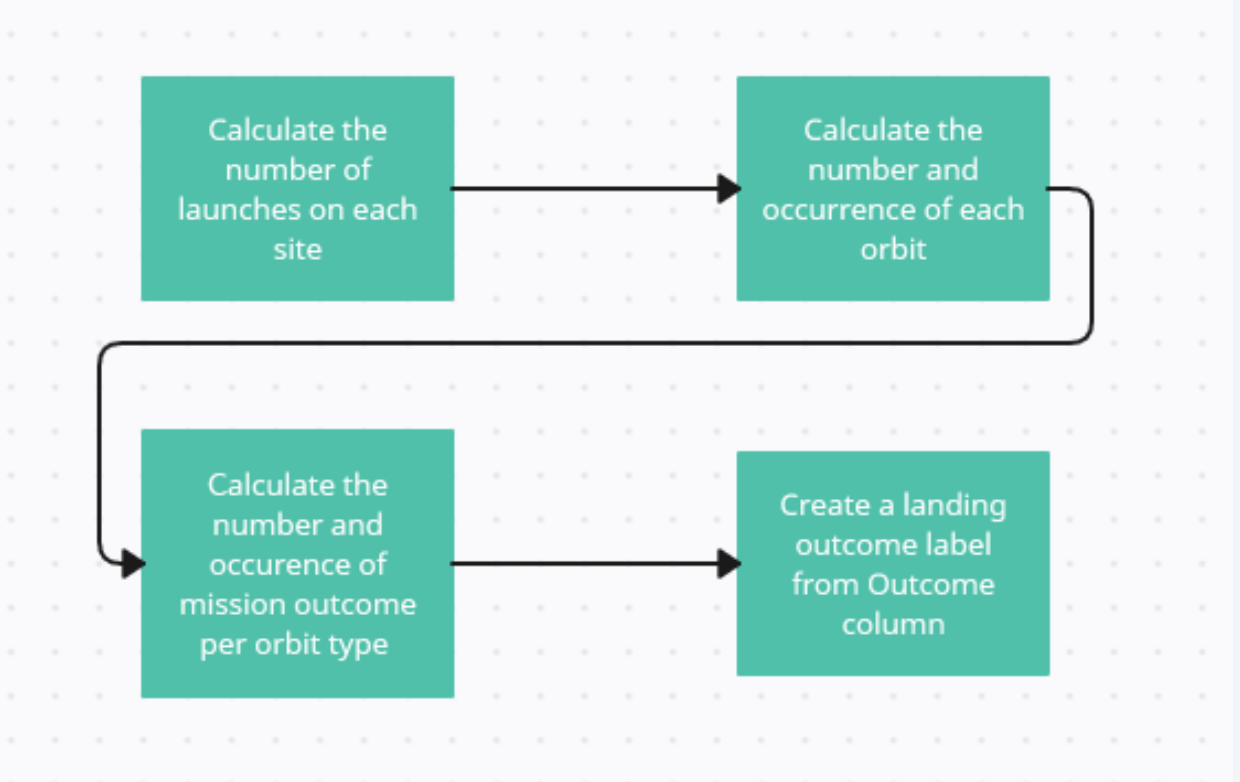
https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling



https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Perform exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib
- Exploratory Data Analysis
- Preparing Data Feature Engineering

Plots: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, Success Yearly Trend
I employed scatterplots, line charts and bar graphs to explore connections between variables, assessing their interrelationships to determine their suitability for inclusion in the machine learning model's training dataset.

https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

1. Understand the Spacex DataSet
2. Load the dataset into the corresponding table in a Db2 database
3. Execute SQL queries to answer assignment questions:

- names of the unique launch sites in the space mission
- 5 records where launch sites 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 succesful landing outcome in ground pad was achieved
- names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- total number of successful and failure mission outcomes
- names of the booster_versions which have carried the maximum payload mass
- records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
- 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

https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Folium maps are utilized to pinpoint the precise locations of launch sites, distinguish between successful and unsuccessful landings, and illustrate proximity to significant landmarks, including railways, highways, coastlines, and urban centers. This visual approach offers insights into the rationale behind the selection of launch site locations and provides a clear visualization of the geographical distribution of successful landings.

https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

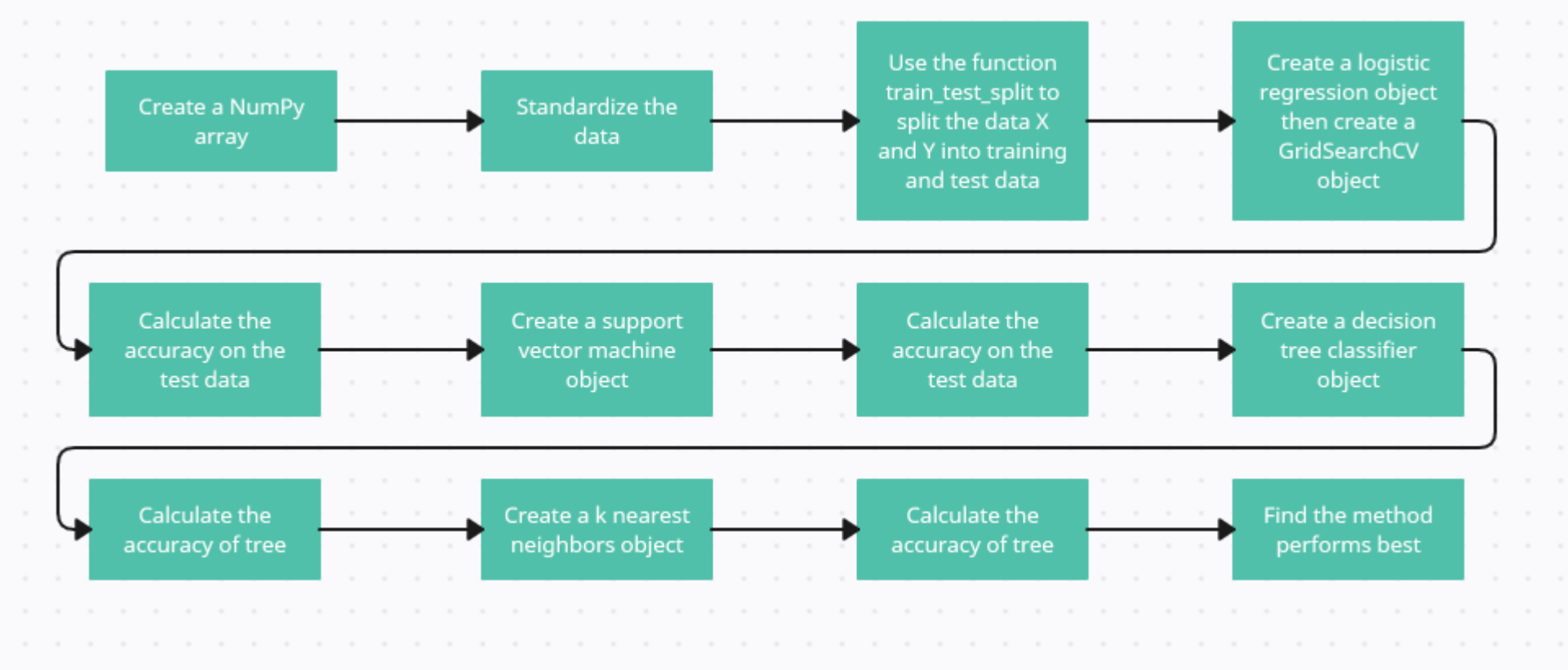
I've created a SpaceX Launch Records Dashboard using Dash, featuring

- a dropdown menu for launch site selection, a pie chart showing success rates by site, a payload mass range slider for data filtering, and a scatter plot illustrating the relationship between payload mass and launch success.

These interactive components enhance the user experience by allowing me to explore success rates, analyze payload mass influence, and filter data according to my preferences. The dashboard is designed for a comprehensive analysis of SpaceX launch records.

https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

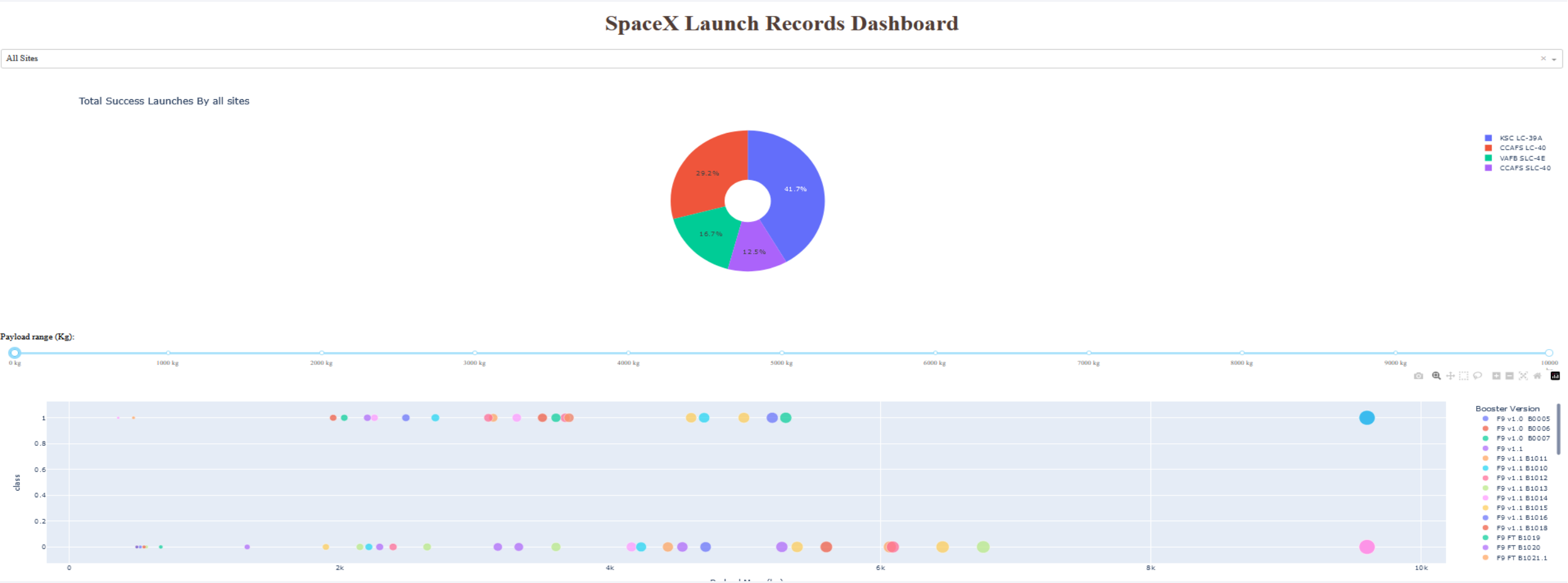


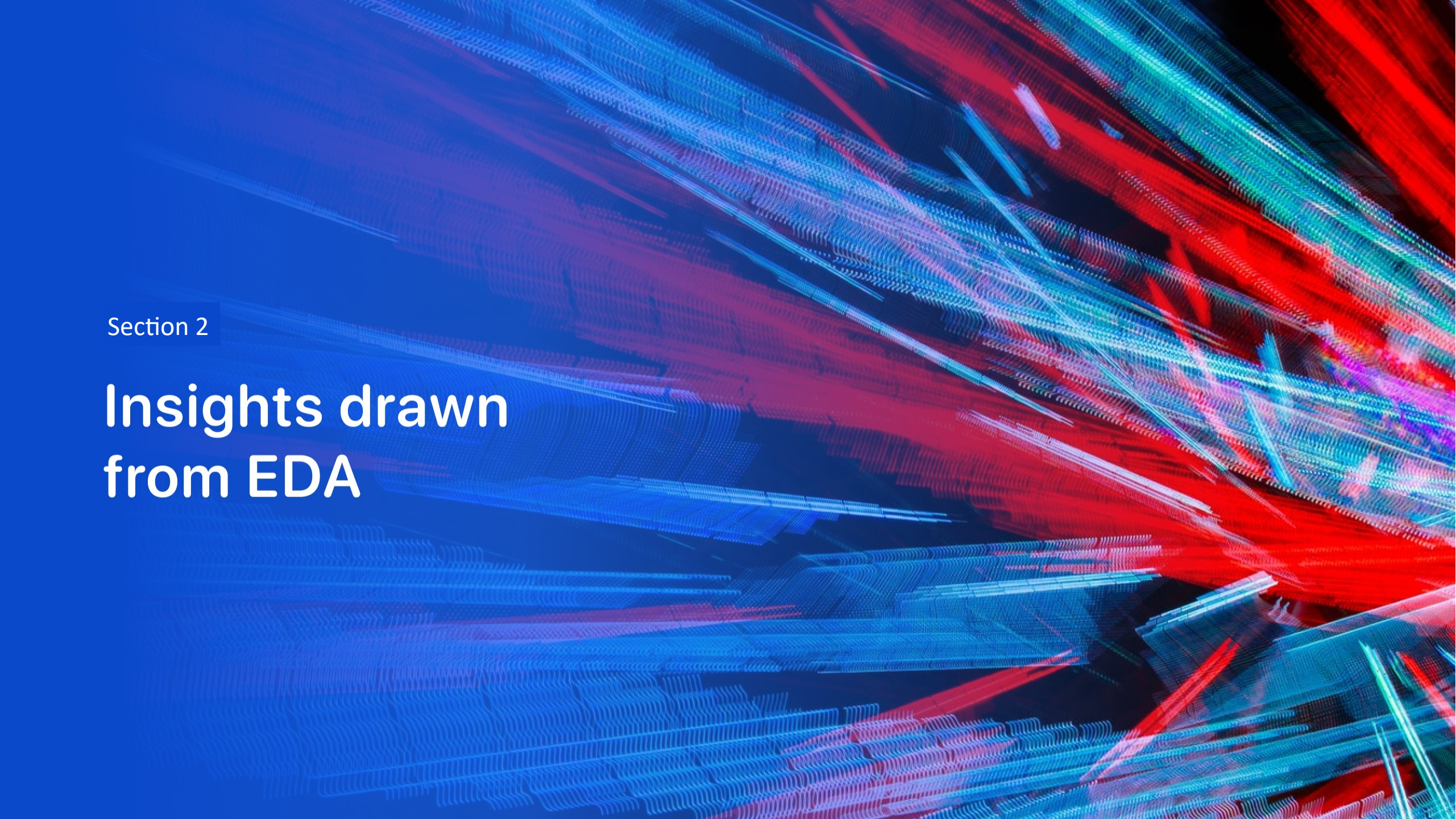
https://github.com/miraylin/Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

The provided code used there primarily focuses on exploratory data analysis and interactive visualization.

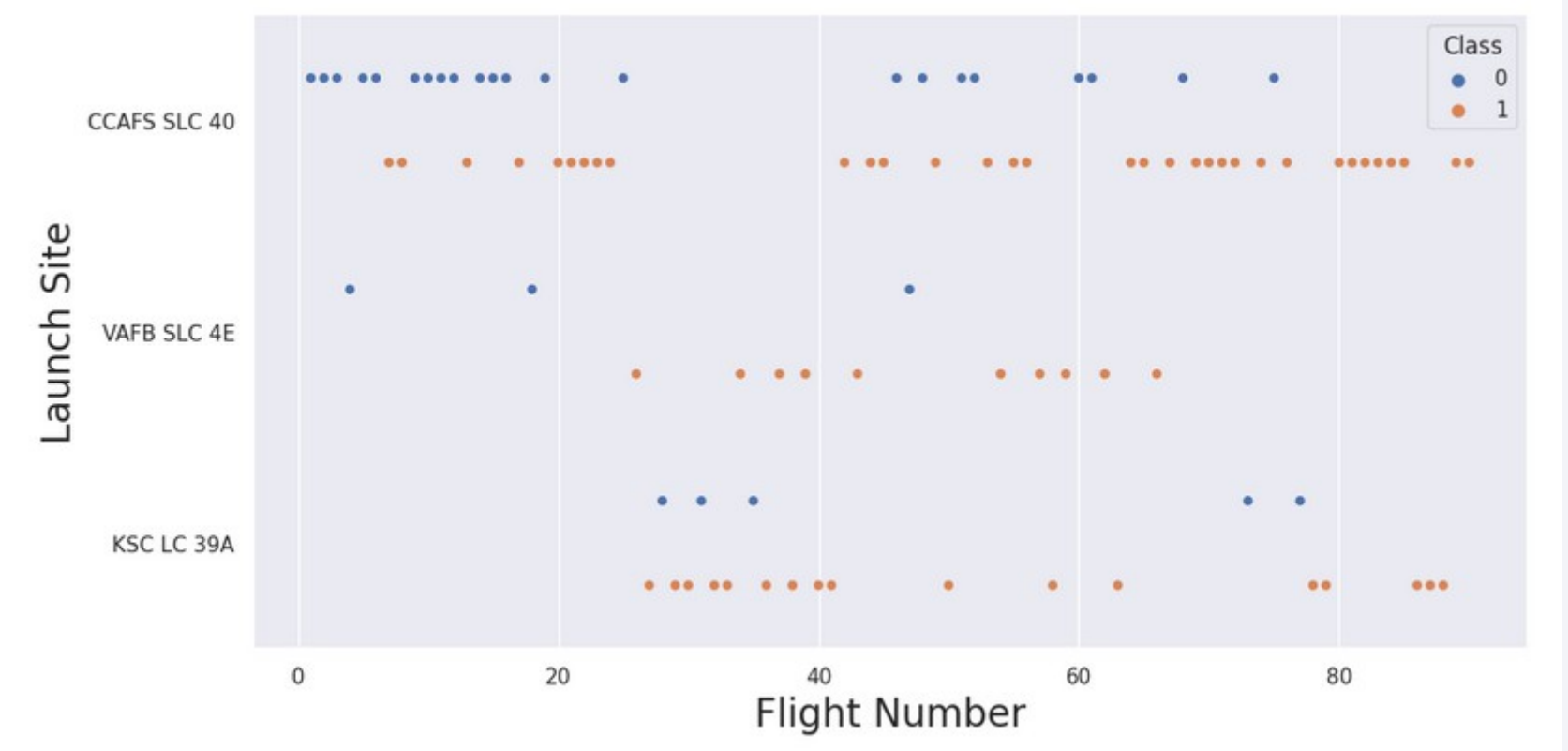
This is a preview of a SpaceX Launch Records Dashboard:




Section 2
Insights drawn from EDA

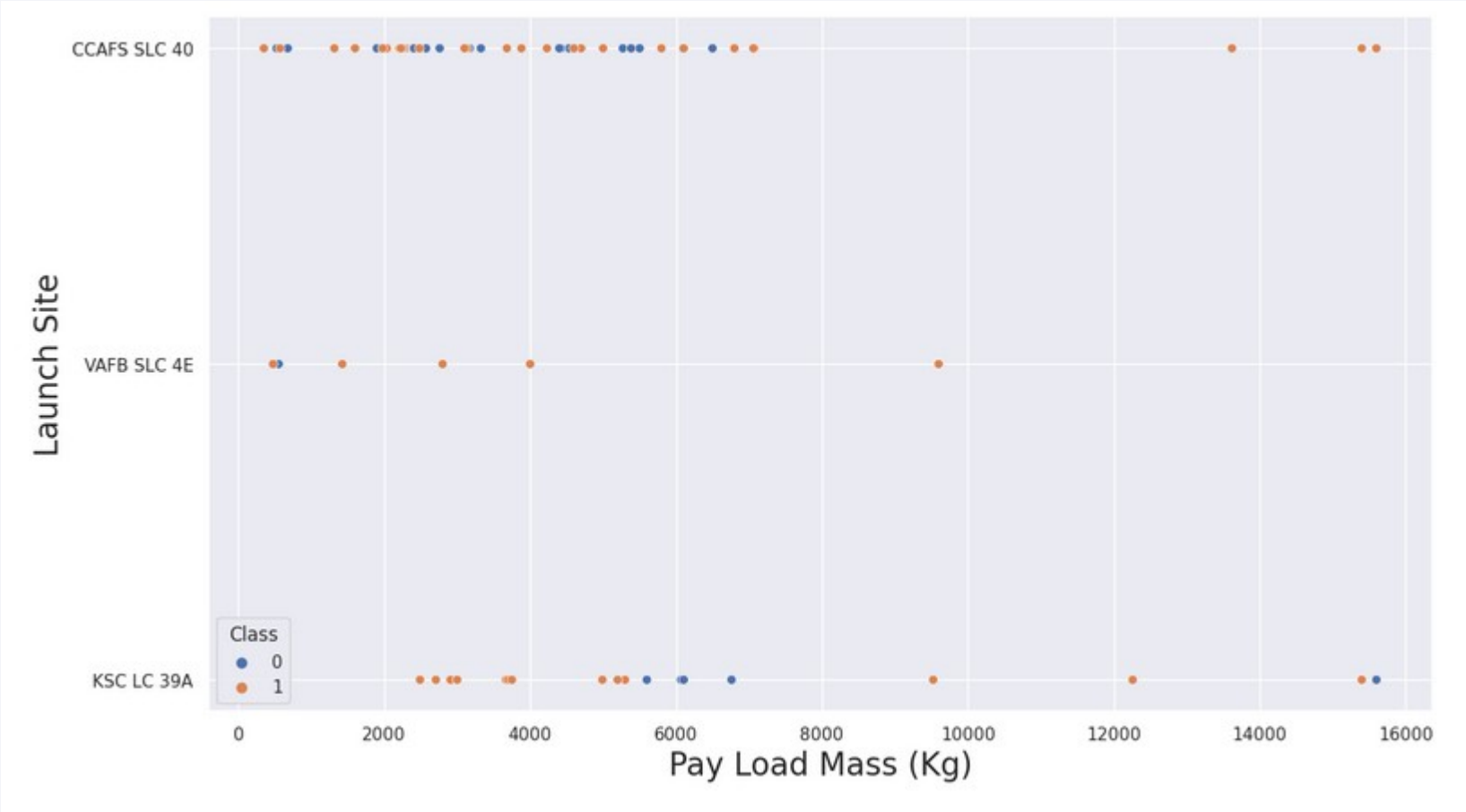
Flight Number vs. Launch Site

- In simpler terms, the scatter plot indicates that when a launch site can accommodate a larger payload mass, the likelihood of a successful launch is higher.



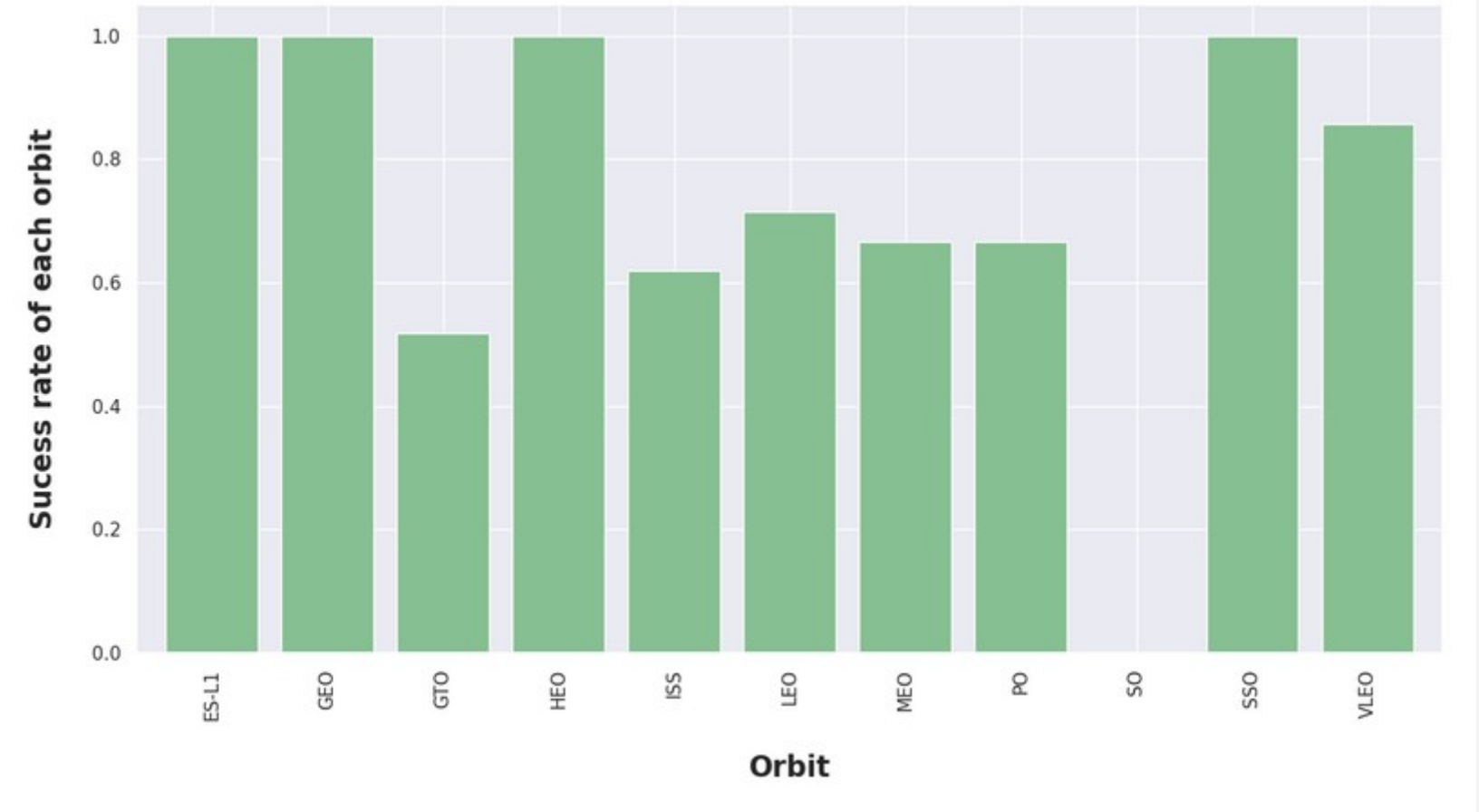
Payload vs. Launch Site

It seems that the majority of payload masses fall within the range of 0-6000 kilograms. Additionally, various launch sites appear to have different preferences when it comes to the payload mass they handle



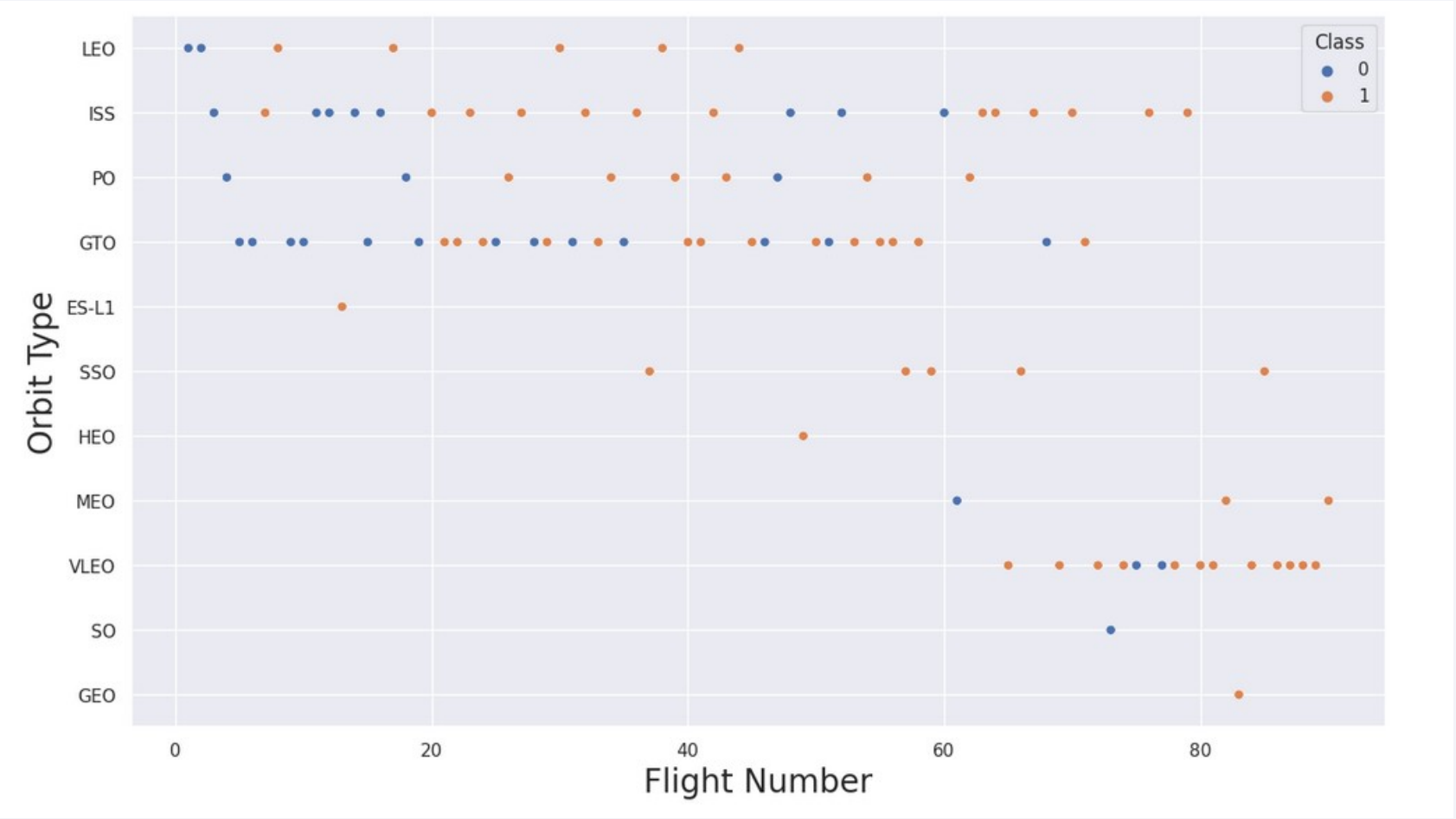
Success Rate vs. Orbit Type

- ES-L1, GEO, and HEO, each with a sample size of one, have a 100% success rate. SSO, with a sample size of five, also boasts a perfect 100% success rate. VLEO, which has seen 14 attempts, has a respectable success rate, indicating a noteworthy level of success. In contrast, SO, with a sample size of one, has a 0% success rate, while GTO, with the largest sample size of 27, maintains a success rate of approximately 50%, making it the most extensively sampled category.



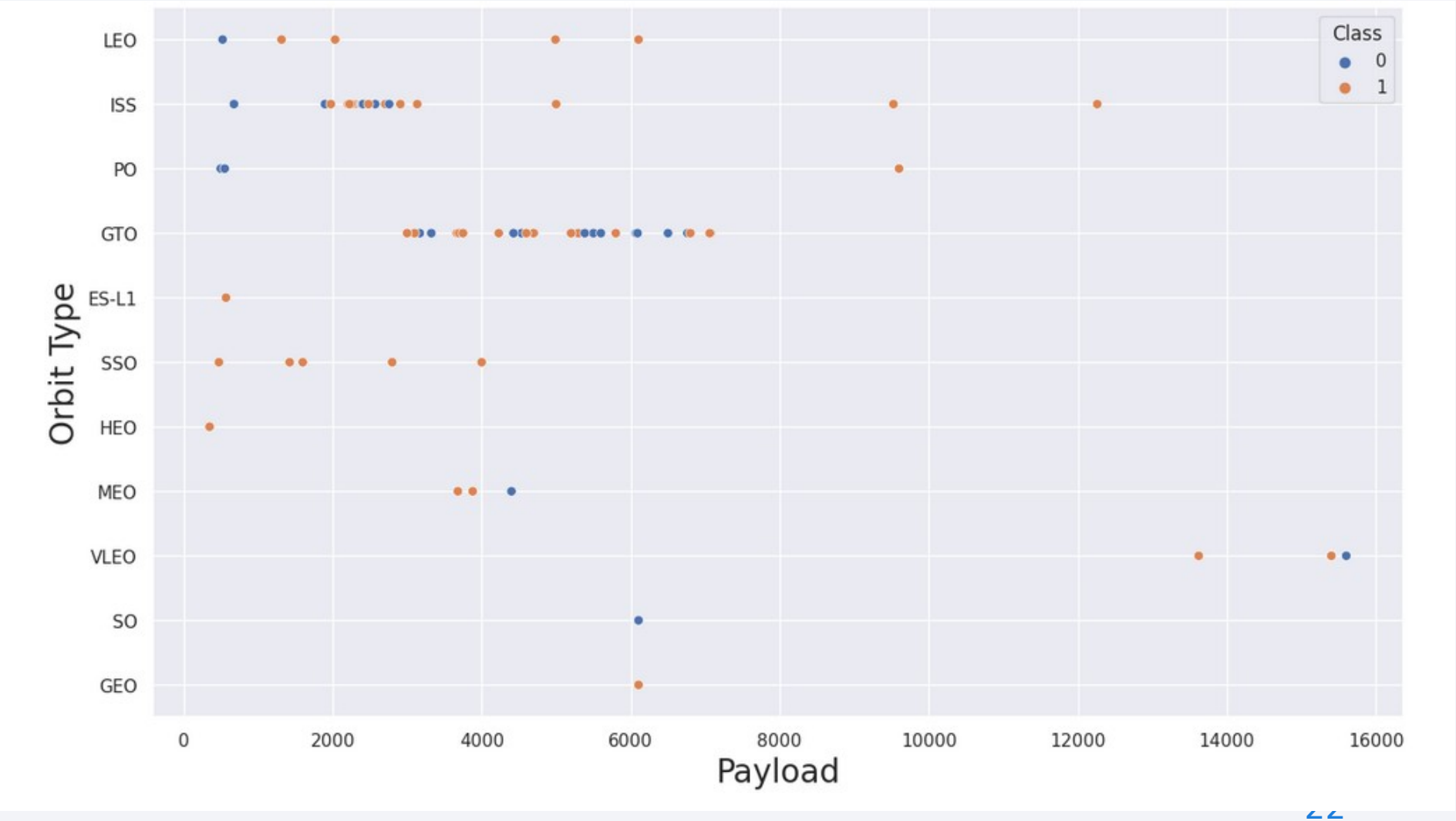
Flight Number vs. Orbit Type

Over the course of their flight numbers, SpaceX's launch orbit preferences have evolved. They began with a focus on Low Earth Orbits (LEO), which yielded moderate success rates. Subsequently, they transitioned back to Very Low Earth Orbits (VLEO) in recent launches. Notably, SpaceX's performance appears to be more favorable in lower orbits or Sun-synchronous orbits, as there seems to be a correlation between launch outcomes and orbit selection.



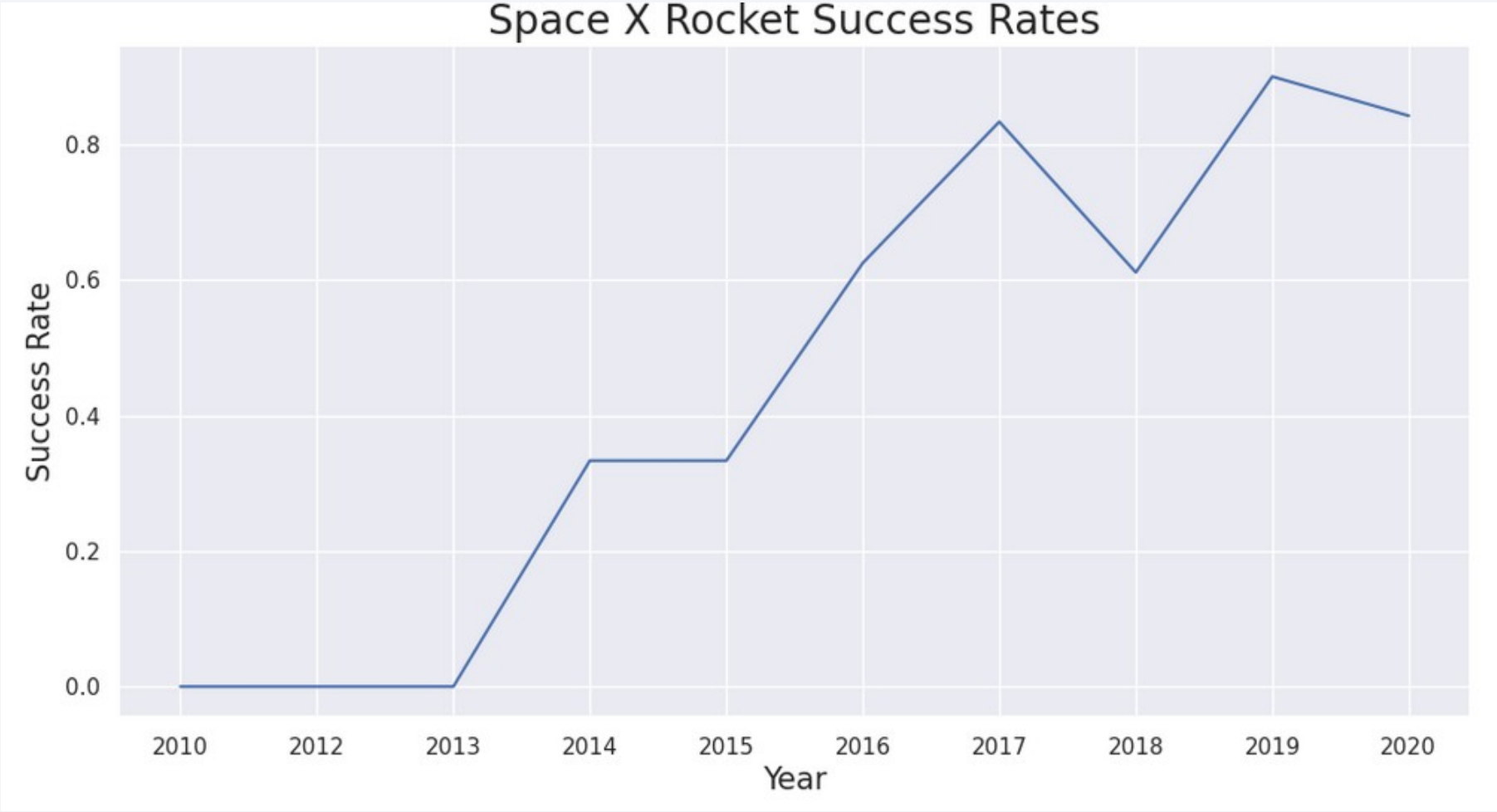
Payload vs. Orbit Type

There appears to be a correlation between payload mass and launch orbit. Orbits such as Low Earth Orbit (LEO) and Sun-synchronous orbit (SSO) tend to have relatively lower payload masses. In contrast, the most successful orbit, Very Low Earth Orbit (VLEO), primarily features payload mass values on the higher end of the spectrum. This suggests that the choice of orbit is associated with the payload mass, with VLEO being more suitable for heavier payloads, while LEO and SSO are preferred for lighter payloads



Launch Success Yearly Trend

The trend in launch success has generally been on the rise since 2013, with a minor decrease observed in 2018. In recent years, success rates have stabilized at approximately 80%, indicating a relatively consistent level of successful launches.



All Launch Site Names

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

```
In [8]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Out[8]:

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

Launch Site Names Begin with 'CCA'

Displaying 5 records where launch sites begin with the string 'CCA'

In [9]:
```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[9]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

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

```
In [10]: %sql SELECT SUM (PAYLOAD_MASS__kg_) FROM SPACEXTBL WHERE CUSTOMER = 'NASA(CRS)' ;
 * sqlite:///my_data1.db
Done.
Out[10]: SUM (PAYLOAD_MASS__kg_)
                           45596
```

Average Payload Mass by F9 v1.1

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

```
In [13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
Out[13]: AVG(PAYLOAD_MASS__KG_)
         2928.4
```

First Successful Ground Landing Date

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

```
In [15]: %sql SELECT MIN(DATE) AS "First Successful Landing" FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

Out[15]:

First Successful Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [16]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AI
```

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

Out[16]:

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 of the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

In [23]:
```
%sql SELECT COUNT(MISSION_OUTCOME) AS "succesful mission "FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

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

Out[23]:

succesful mission
100

In [24]:
```
%sql SELECT COUNT (MISSION_OUTCOME) AS "failure mission " FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Fail%';
```

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

Out[24]:

failure mission
1

Boosters Carried Maximum Payload

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

In [25]:
```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Out[25]:

Booster Versions which carried the Maximum Payload Mass
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

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

In [36]:
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND  LANDING_OUTCOME = 'Failure (drone
```

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

Out[36]:

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

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

In [38]:

```
%sql SELECT LANDING_OUTCOME as "Landing Outcome", COUNT(LANDING_OUTCOME) AS "Total Count" FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING_OUTCOME \
ORDER BY COUNT(LANDING_OUTCOME) DESC ;
```

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

Out[38]:

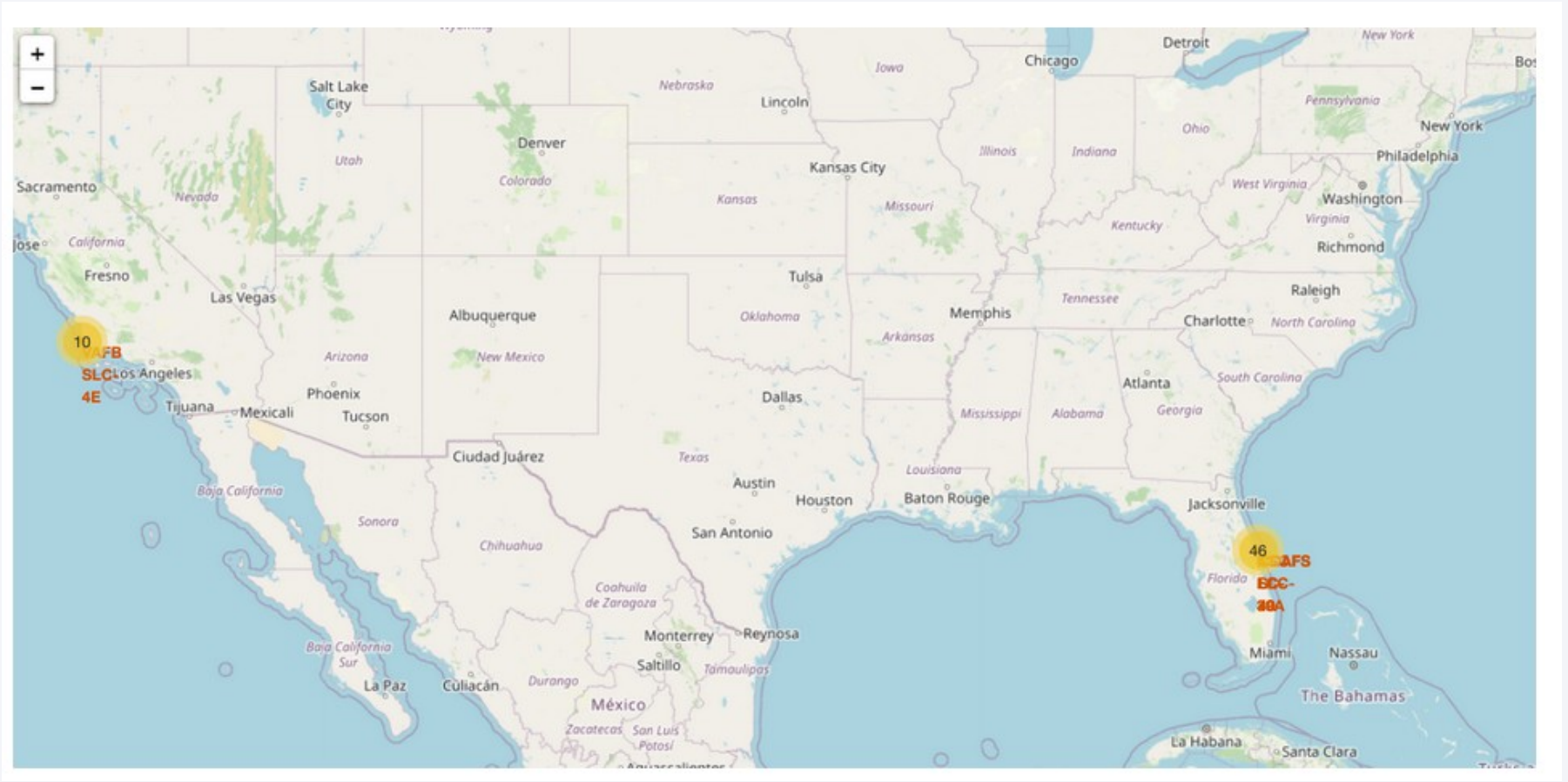
Landing Outcome	Total Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

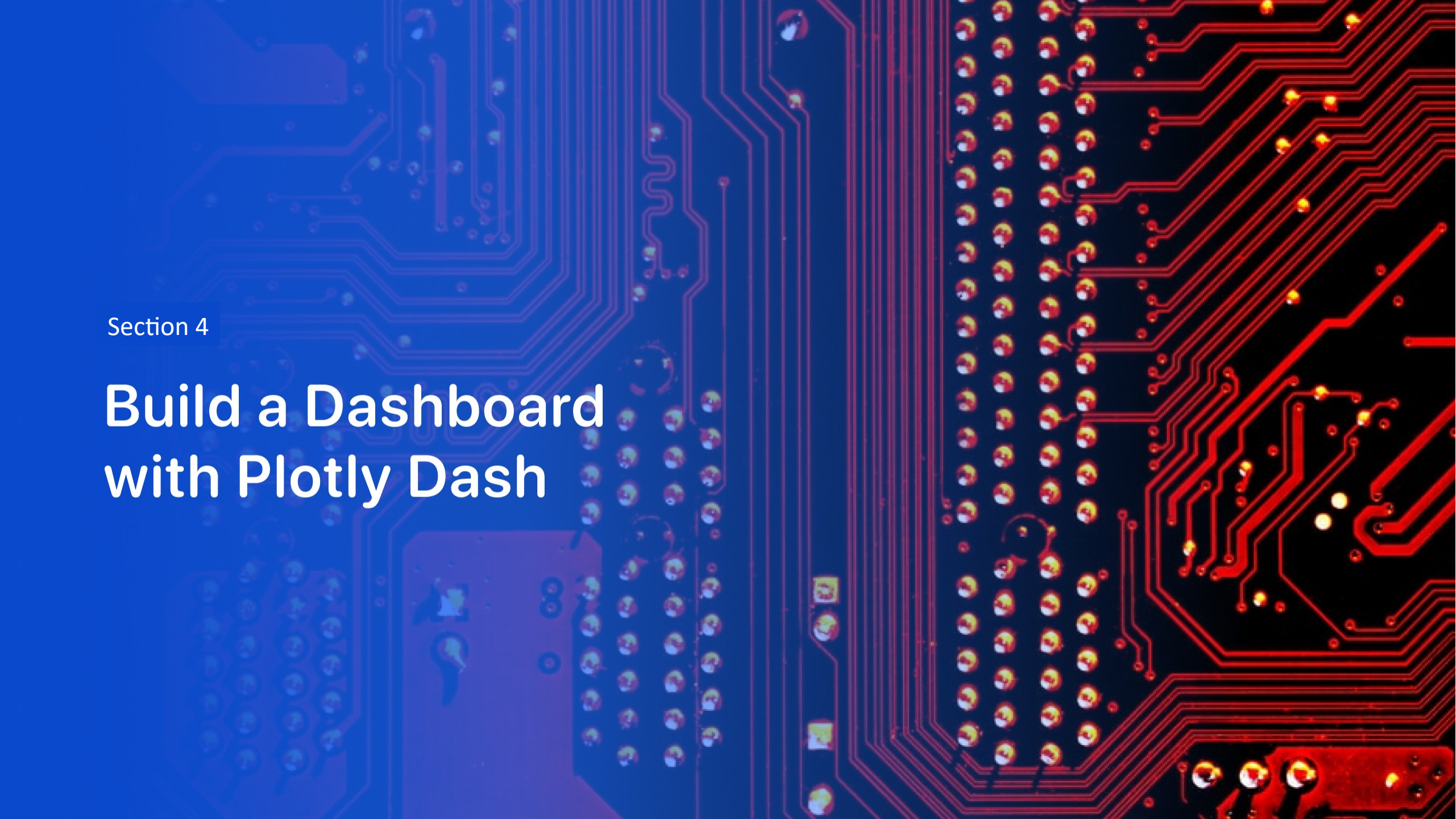
Launch Site Locations

All launch sites on a map.



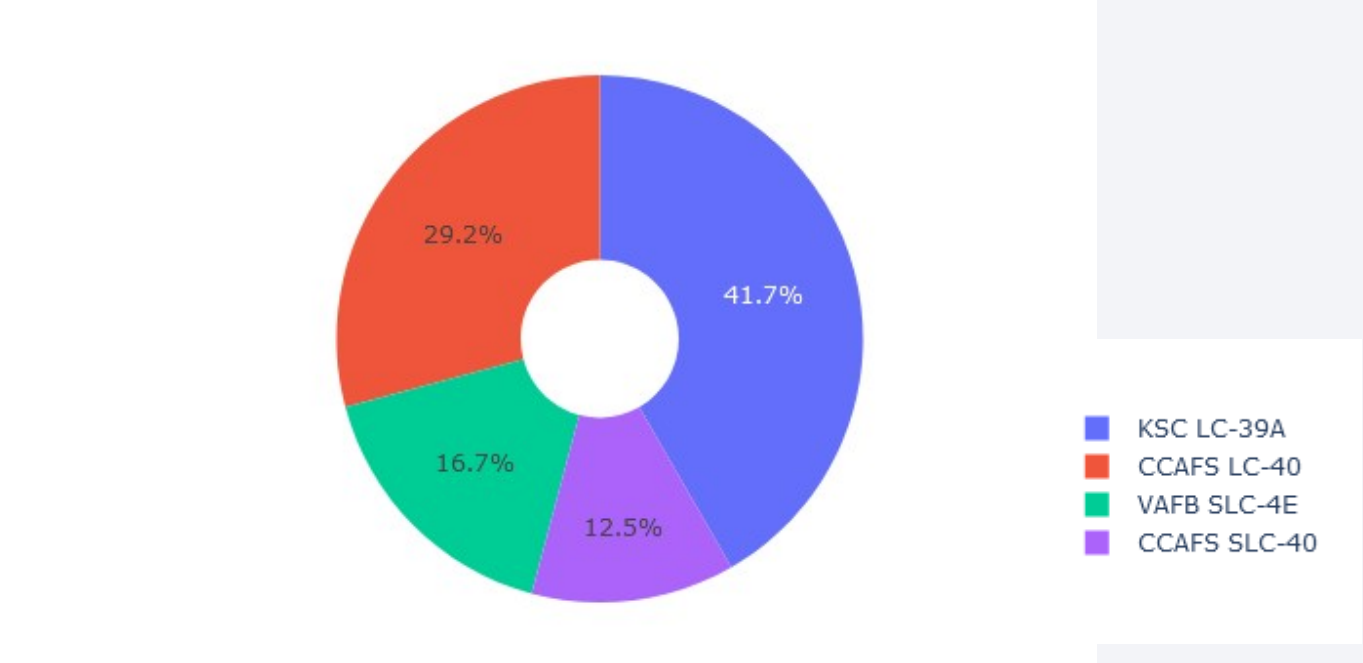
Color-Coded Launch Markers

Green represents successful landing (there are 3) and unsuccessful as a red markers (there are 4 red).


Section 4
Build a Dashboard with Plotly Dash

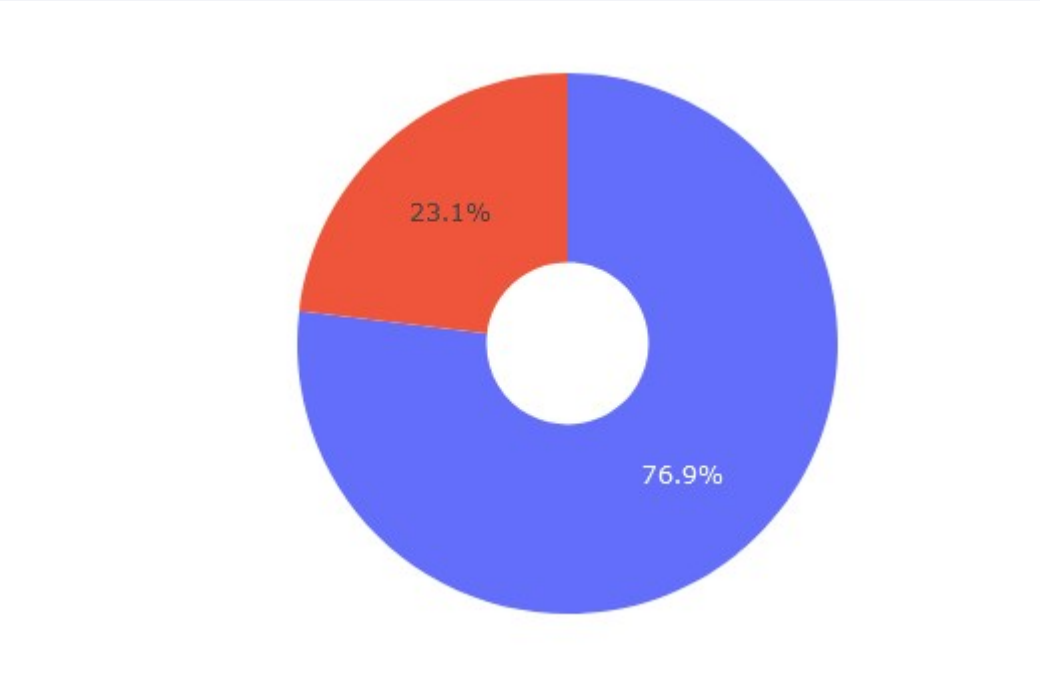
Total Success Launches by all sites

This data represents the distribution of successful rocket landings at various launch sites. It's worth noting that CCAFS LC-40 and CCAFS SLC-40 are essentially the same location, so CCAFS and KSC share an equal number of successful landings. However, a significant portion of these successes occurred before the name change from CCAFS LC-40 to CCAFS SLC-40. On the other hand, VAFB has the lowest percentage of successful landings, possibly influenced by a smaller sample size and the increased challenges associated with launching from the West Coast.



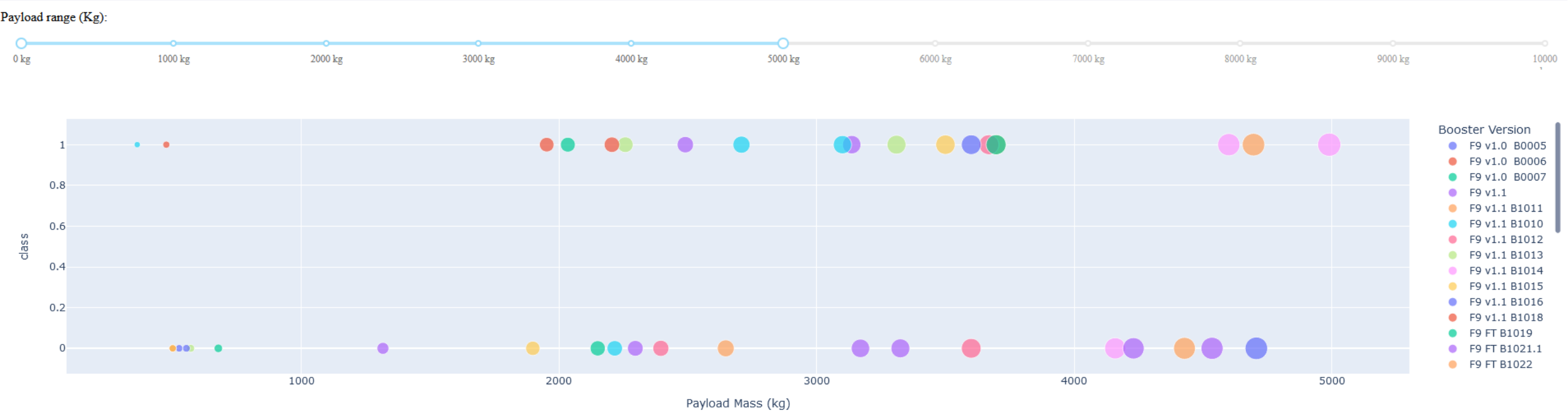
Total Success Launches for site KSC LC-39A

KSC LC-39A has the highest success rate.
Blue means Success, Red failed launches.



Payload Mass vs. Success vs. Booster Version

In the Plotly dashboard, the Payload range selector is currently configured with a range from 0 to 10,000. The "Class" variable is used to signify successful landings (1) and failures (0). The scatter plot additionally incorporates booster version categories for color representation and the number of launches for point size.

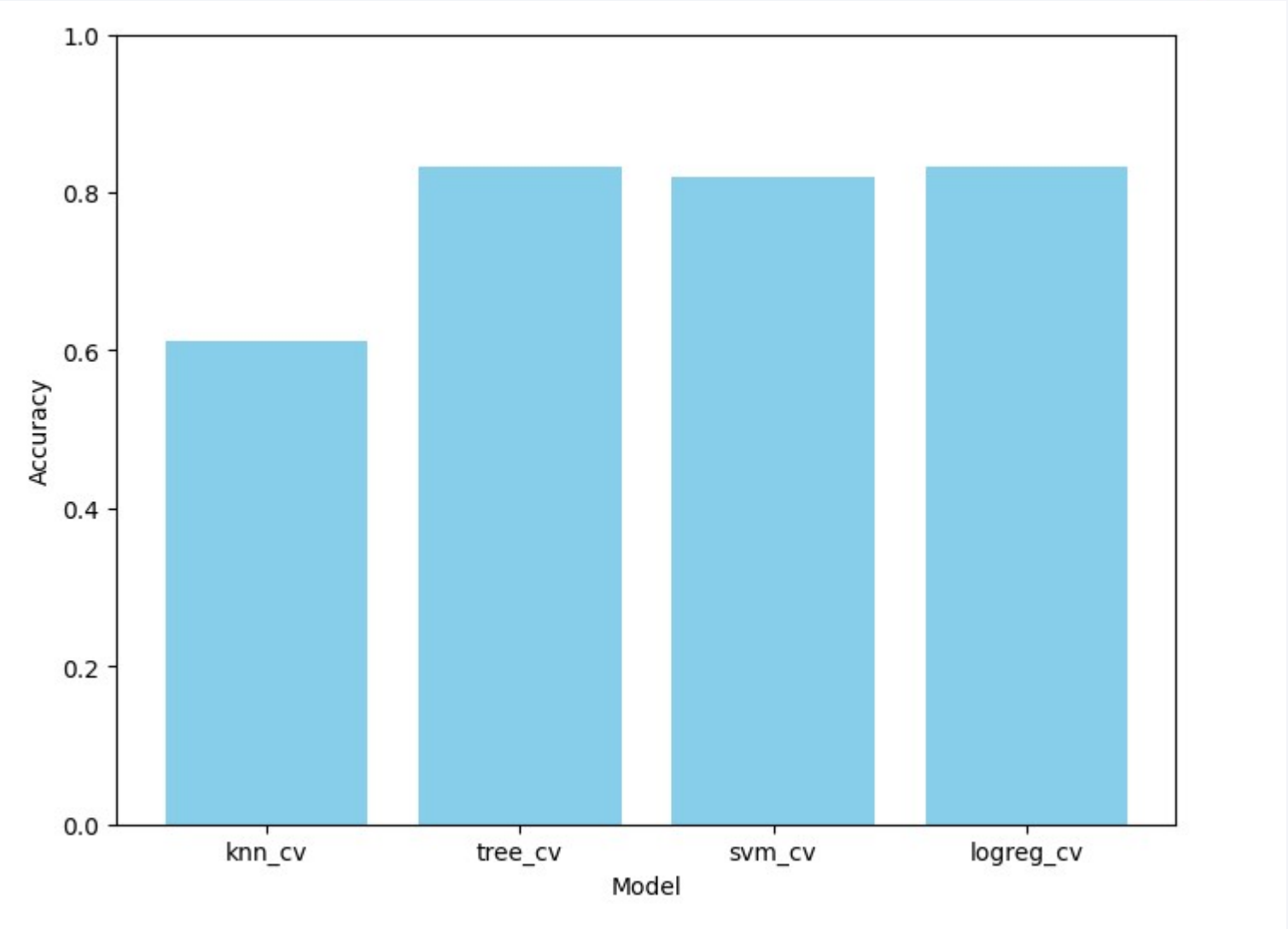


Section 5

Predictive Analysis (Classification)

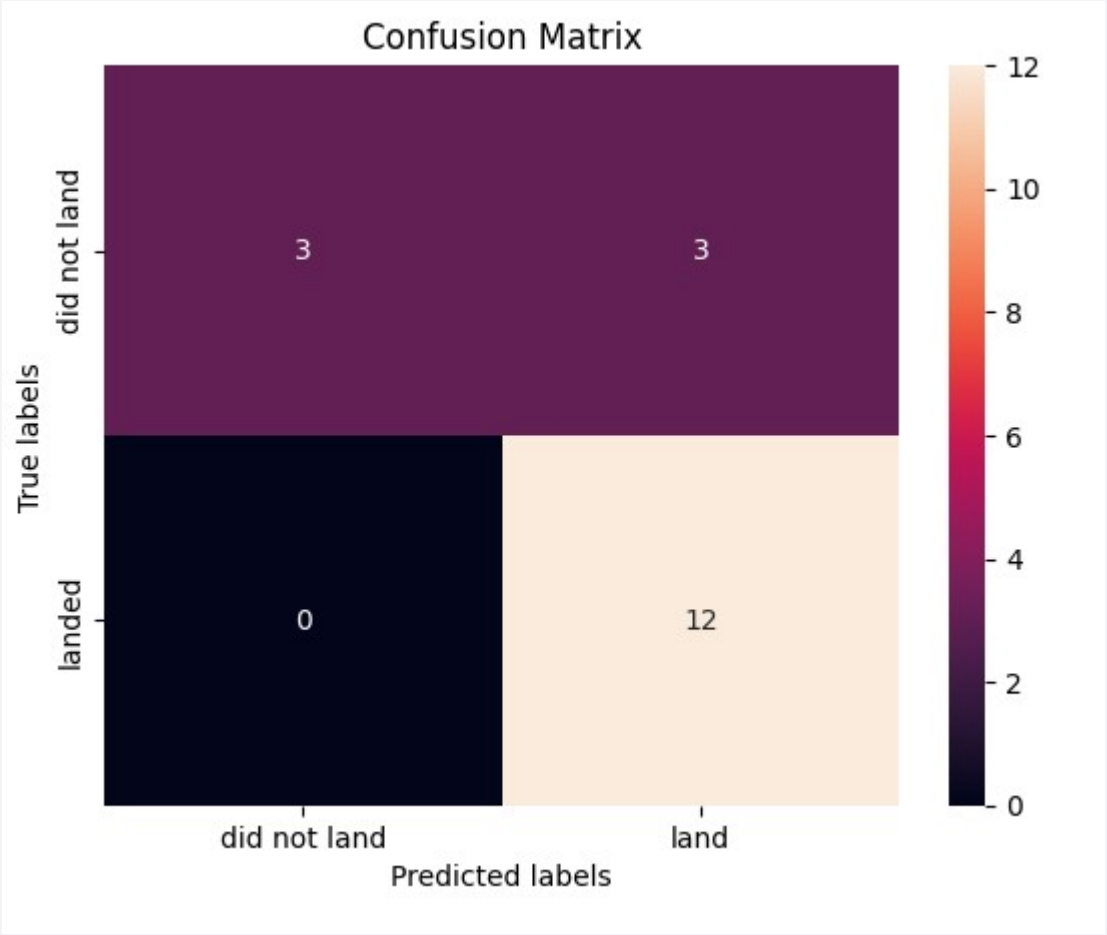
Classification Accuracy

The models "tree_cv" and "logreg_cv" exhibit the highest likelihood of success in the analysis, indicating their potential as strong candidates for further exploration and utilization in the given context.



Confusion Matrix

Since two models performed equally on the test set, their confusion matrices were identical. They correctly predicted 12 successful landings and 3 unsuccessful ones. However, they also predicted 3 successful landings when they were actually unsuccessful (false positives), indicating a tendency to overpredict successful outcomes.



Conclusions

Our task was to develop a machine learning model for SpaceY, a potential competitor to SpaceX. The primary goal was to predict the successful landing of Stage 1 rockets, which could potentially save up to $100 million USD per launch.

I collected data from a public SpaceX API and scraped information from SpaceX's Wikipedia page. Subsequently, I labeled this data and stored it in a DB2 SQL database to facilitate further analysis.

I created a visualization dashboard to explore and understand the data more effectively. The machine learning model I built achieved an impressive accuracy rate of 83%, showcasing its potential usefulness.

This model could be a valuable asset for SpaceY, as it enables them to predict, with relatively high accuracy, whether a launch will result in a successful Stage 1 landing before proceeding. To further enhance the model's accuracy, I recommend collecting additional data, which could help determine the most effective machine learning model for this task.

Thank you!