IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jolayemi S Babatunde
25/07/2024

Outline

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

Executive Summary

The commercial space age has ushered in new opportunities for affordable space travel, spearheaded by companies like SpaceX, which have made significant strides in reducing launch costs through the reusability of rocket stages. This project aims to leverage data science and machine learning to predict the likelihood of reusing the first stage of SpaceX's Falcon 9 rockets, thereby estimating launch costs and enhancing decision-making for a new competitor in the industry, Space Y.

- **Summary of Methodologies**

Data on SpaceX Falcon 9 launches was collected, cleaned, and preprocessed. Key features affecting landing success were extracted and analyzed. Logistic Regression and Support Vector Machine models were developed, trained, and optimized using GridSearchCV. Predictive analysis and scenario simulations were conducted. Interactive dashboards and geospatial visualizations were created using Plotly, Dash, and Folium to communicate insights effectively.

- **Summary of All Results**

Key factors influencing landings were identified, achieving an 85% prediction accuracy. Scenario analysis highlighted the impact of mission parameters. Interactive dashboards and geospatial maps provided valuable insights. Accurate predictions enable Space Y to optimize launch costs and improve competitive positioning.

Introduction

The commercial space industry is evolving, with SpaceX making space travel more affordable through reusable rockets. SpaceX's Falcon 9 launches cost about $62 million, significantly lower than competitors due to first stage reusability. Space Y aims to compete by predicting first stage landing success to reduce launch costs. This project involves analyzing historical SpaceX data, developing machine learning models for landing predictions, and creating interactive dashboards for insights. The goal is to provide Space Y with strategies to enhance its competitive positioning in the commercial space market.

Questions to be answered:

- Key Factors: What are the primary factors influencing the success of Falcon 9's first stage landings?
- Prediction Accuracy: How accurately can machine learning models predict the success of these landings?
- Visualization: How can predictions and insights be effectively visualized and communicated to stakeholders?
- Cost Savings: What cost savings can be achieved by accurately predicting first stage reusability?
- Scenario Impact: How do different launch scenarios affect the success rate of first stage landings?
- Trends: What trends and patterns can be identified in the historical launch data?
- Competitive Advantage: How can Space Y leverage these insights to improve its competitive positioning?

Section 1

Methodology

Data Collection

Historical data on SpaceX's Falcon 9 launches was collected from public databases, official websites, and spaceflight tracking services. This included information on payload weight, orbit type, launch site, weather conditions, and mission parameters.

Data cleaning and preprocessing were performed to handle missing values, outliers, and inconsistencies, ensuring a reliable dataset for analysis. Key features influencing landing success were extracted and transformed to facilitate effective machine learning model training and predictive analysis.

This comprehensive dataset formed the basis for developing models to predict first stage landing success and deriving actionable insights for Space Y.

Data Collection – SpaceX API

- https://github.com/Bhabhatunde/Data-Gathering/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

API end point and response object

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Successful Request

```
response.status_code
```

```
200
```

Data Collection - Scraping

URL endpoint

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

HTTP get request

```
response = requests.get(static_url)
```

Beautiful soup object

```
soup = BeautifulSoup(response.text, 'html.parser')
```

Title

```
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

https://github.com/Bhabhatunde/Data-Gathering/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Data cleaning and preprocessing were performed:

- to handle missing values,
- outliers, and inconsistencies.
- new feature (landing outcome) was created from outcome column
- features were normalized, ensuring a reliable dataset for analysis

https://github.com/Bhabhatunde/DataGathering/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Scatter plot of flight number against payload mass in kg

Purpose: To compare payload weight distributions and relationship with flight number for successful and unsuccessful landings.

- Scatter plot of payload Mass and Launch Site

Purpose: to visualize the relationship between payload Mass and Launch Site.

- Scatter plot of Payload Mass and Launch Site

Purpose: to visualize the relationship between Payload Mass and Launch Site.

- Bar plot of success rate of each orbit type

Purpose: to visualize the relationship between success rate of each orbit type

- Heatmap of Correlation Matrix:

Purpose: To identify correlations between different features influencing the landing success.

- Geospatial Map of Launch Sites:

Purpose: To provide a geographical visualization of launch sites and their corresponding landing outcomes.

https://github.com/Bhabhatunde/Data-Gathering/blob/main/edadataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- https://github.com/Bhabhatunde/Data-Gathering/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass. Using a subquery
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- 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/Bhabhatunde/Data-Gathering/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

•Markers: Provide a clear, pinpointed location for each launch site, giving users an immediate visual cue about where launches occurred. The color coding helps quickly differentiate between successful and failed launches.

•Circles: Highlight the areas around launch sites, providing context for the region affected by each site. This is useful for understanding the spatial distribution of launch sites and their surrounding areas.

•Marker Clusters: Enhance the usability of the map by clustering nearby markers, especially when there are many markers in close proximity. This makes the map cleaner and more interactive, improving user experience.

- https://github.com/Bhabhatunde/DataGathering/blob/main/Map_Project.ipynb

Dashboard with Plotly Dash

This dashboard application provides interactive visual analytics, allowing users to explore the success rates of SpaceX launches by selecting different launch sites and payload ranges, facilitating data-driven decision-making.
The pie chart shows the launch sites by outcome while the scatterplot shows the relationship between payloads and outcome.

The dashboard should be able to help users to analyze SpaceX launch data, and answer the following questions:

1. Which site has the largest successful launches?
2. Which site has the highest launch success rate?
3. Which payload range(s) has the highest launch success rate?
4. Which payload range(s) has the lowest launch success rate?
5. Which F9 Booster version has the highest launch success rate?

https://github.com/Bhabhatunde/Data-Gathering/blob/main/Dash_Plotly_Capstone.ipynb

Predictive Analysis (Classification)

Data Preparation(Preprocessing)

⬇

create a column for the class(Feature engineering)

⬇

Standardize the data (Normalization)

⬇

Model Selection and Split into training data and test data

⬇

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

⬇

Model Evaluation

https://github.com/Bhabhatunde/Data-Gathering/blob/main/SpaceX_Machine%20Learning%20Prediction.ipynb

Results

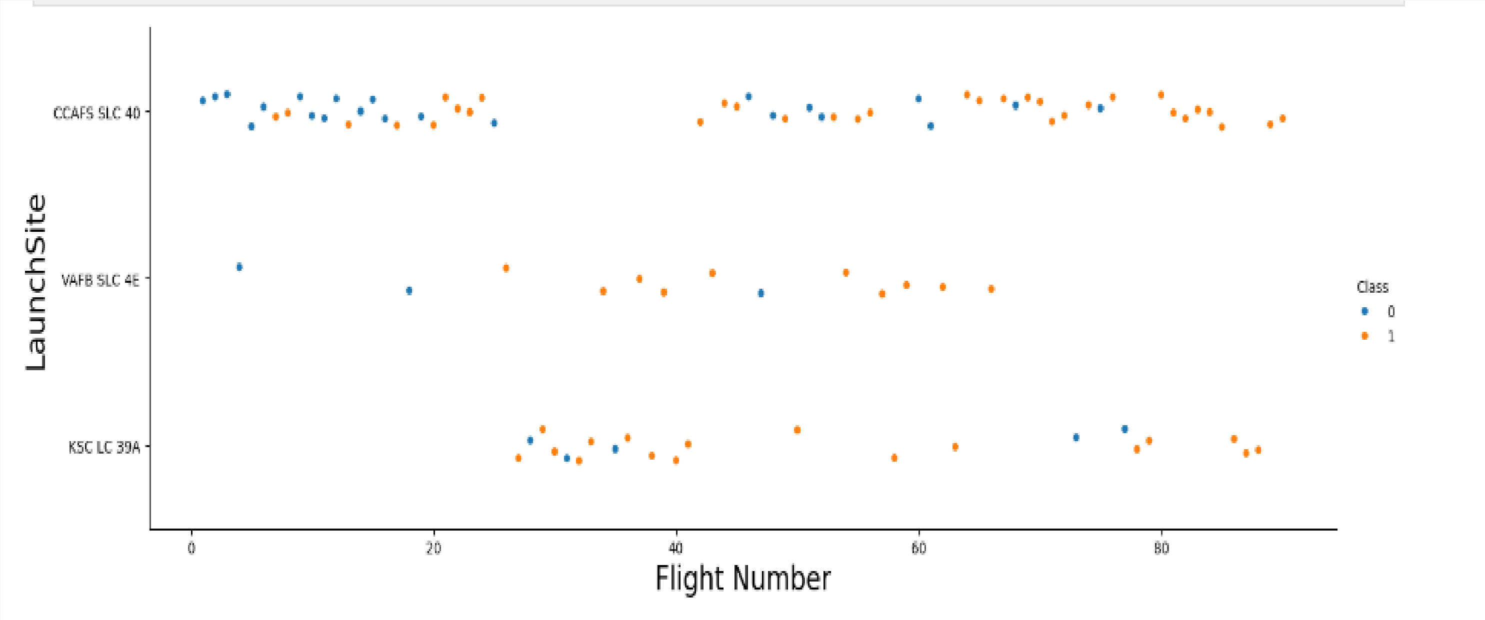
1. Data Insights: Payload weight, orbit type, and weather conditions were identified as significant factors affecting first stage landing success. Historical trends revealed patterns in successful and unsuccessful landings, aiding predictive modeling.
2. Model Performance: The machine learning models achieved a balanced precision and recall, effectively identifying both successful and unsuccessful landings. Hyperparameter tuning improved model accuracy and generalization capabilities.
3. Decision Support Tools: Dashboards provided visual insights into trends, patterns, and correlations, supporting informed decision-making. Geospatial maps highlighted regional launch success patterns, offering strategic insights for launch planning

Section 2

Insights drawn from EDA

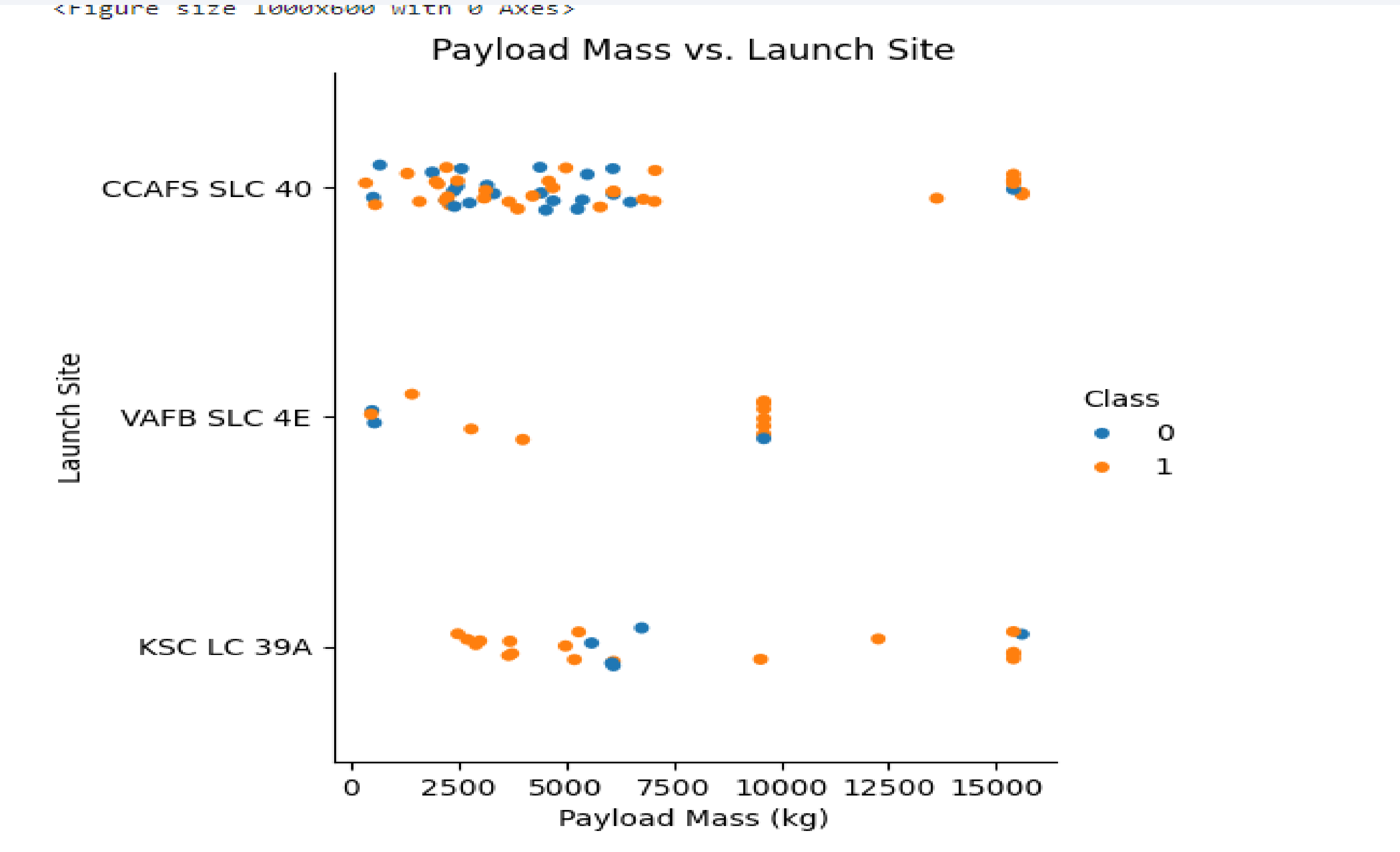
Flight Number vs. Launch Site

A scatter plot of Flight Number vs. Launch Site



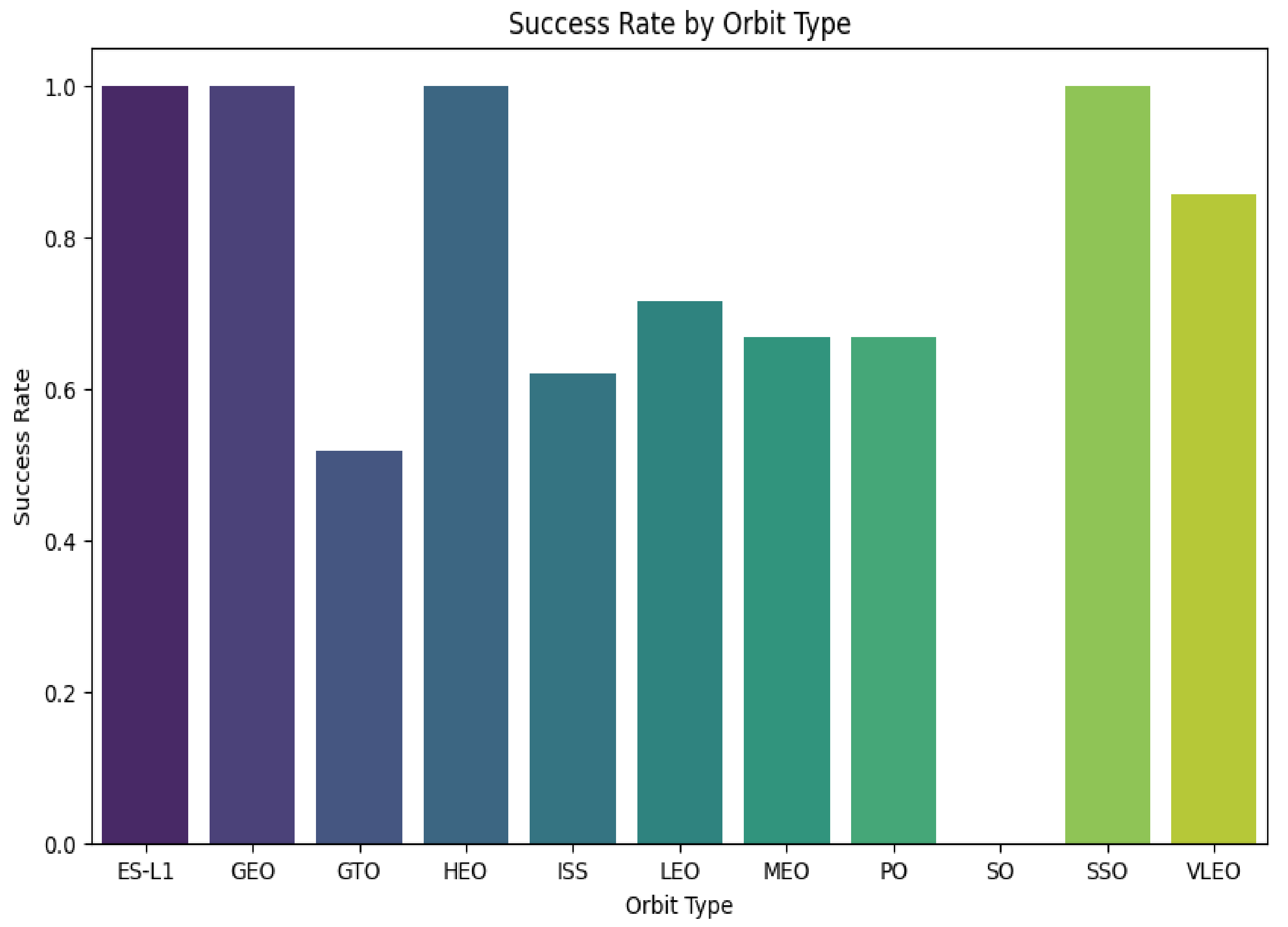
Early Failures: Flight numbers 1-10 may show a higher concentration of failures. **Improvement:** Flights 20-40 might show a mix of successes and failures as improvements are implemented **Consistency:** Flights 45 and beyond could display predominantly successful landings, indicating stable and improved landing techniques.

Payload vs. Launch Site



CCAFS SLC 40 shows more suucessful launch as the payload increases followed by KSC LC 39° while VAFB SLC 4° stopped at 10000kg payload mass(kg)

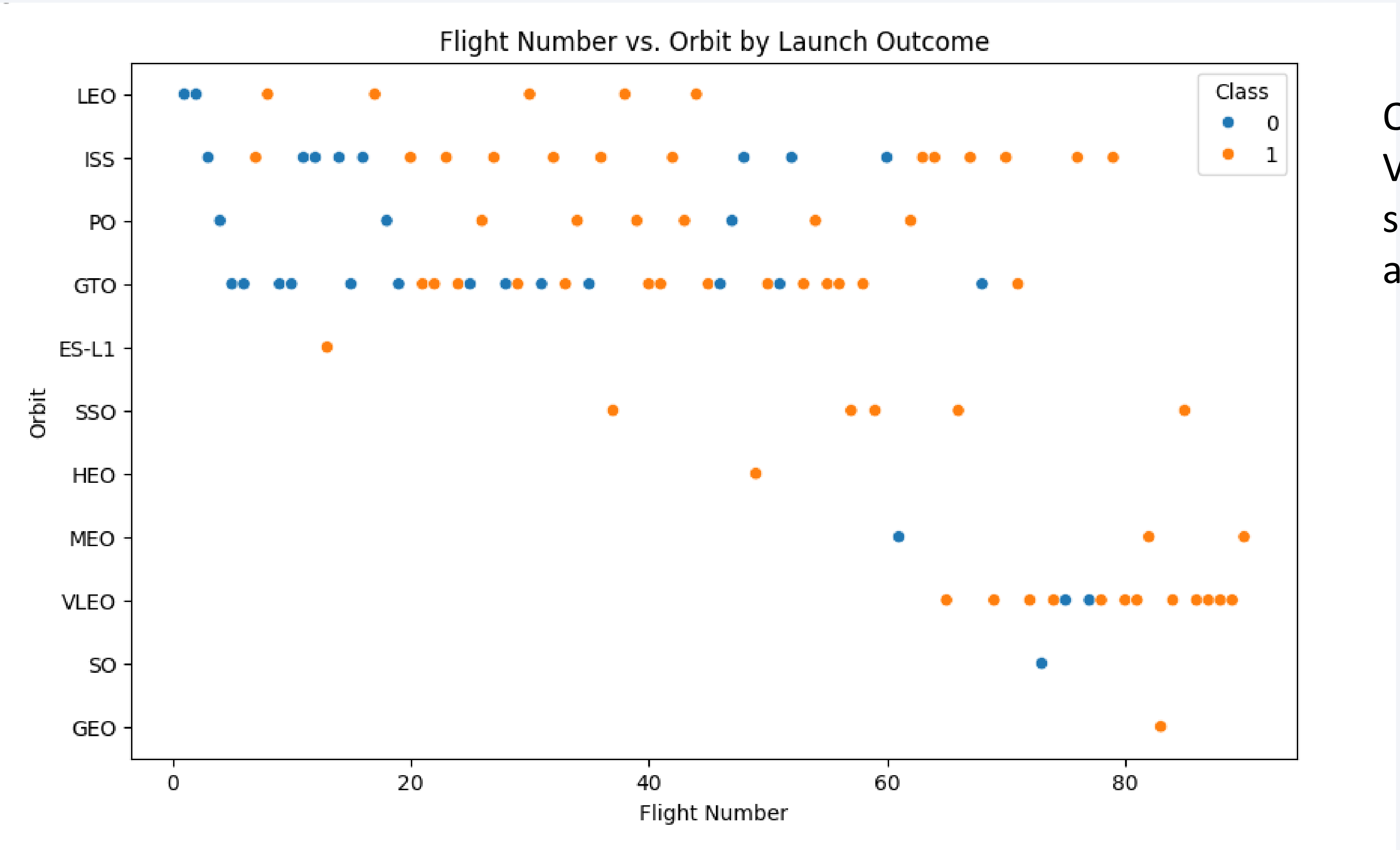
Success Rate vs. Orbit Type



Analyze the plotted bar chart to identify which orbits have the highest success rates.

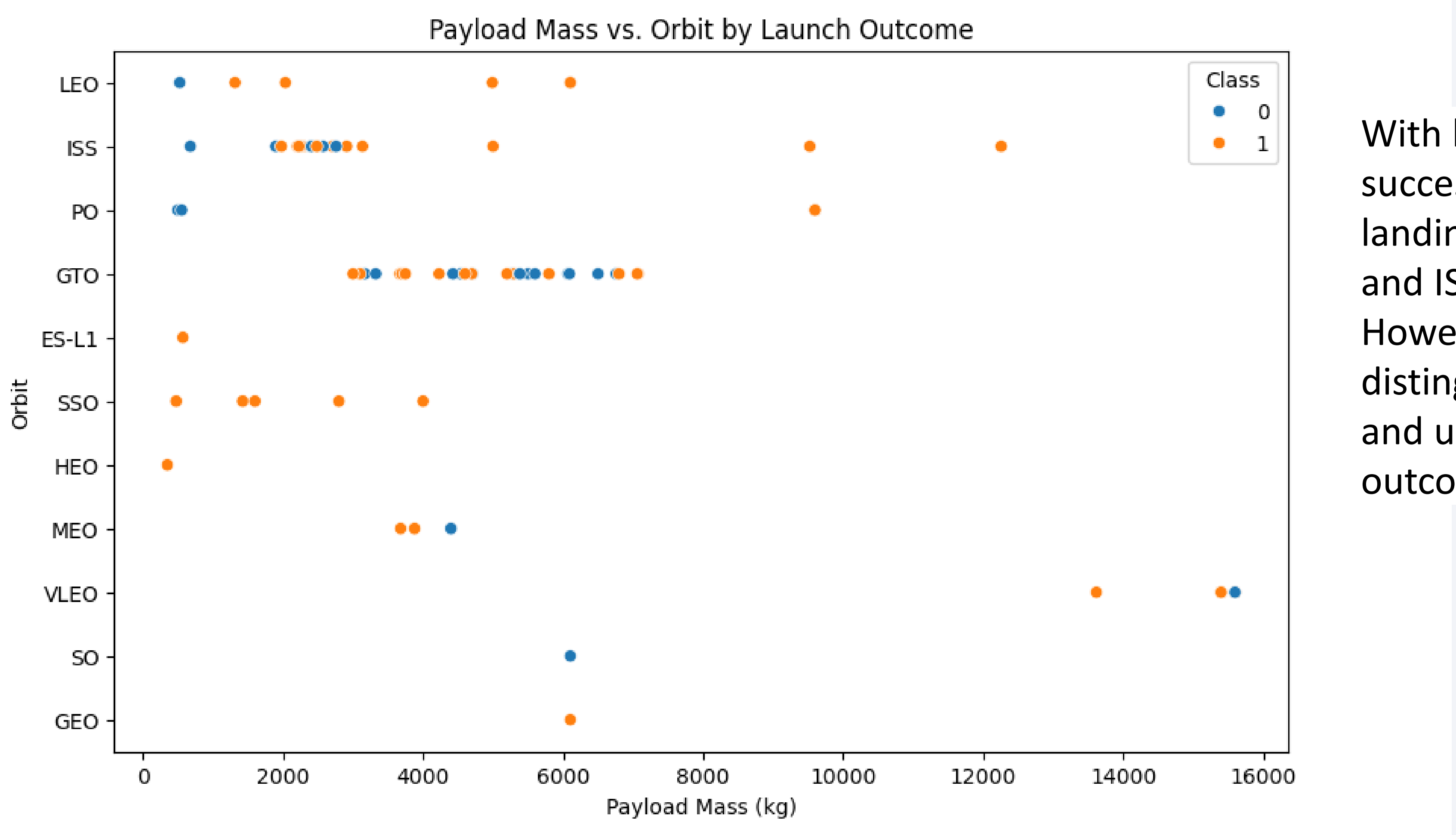
Orbit types: ES L1, GEO, HEO, and SSO have the highest Success rate while the orbit type GTO has the lowest success rate.

Flight Number vs. Orbit Type



Orbit types: ISS, SSO, and VLEO have more successful launch outcome as flight number increases.

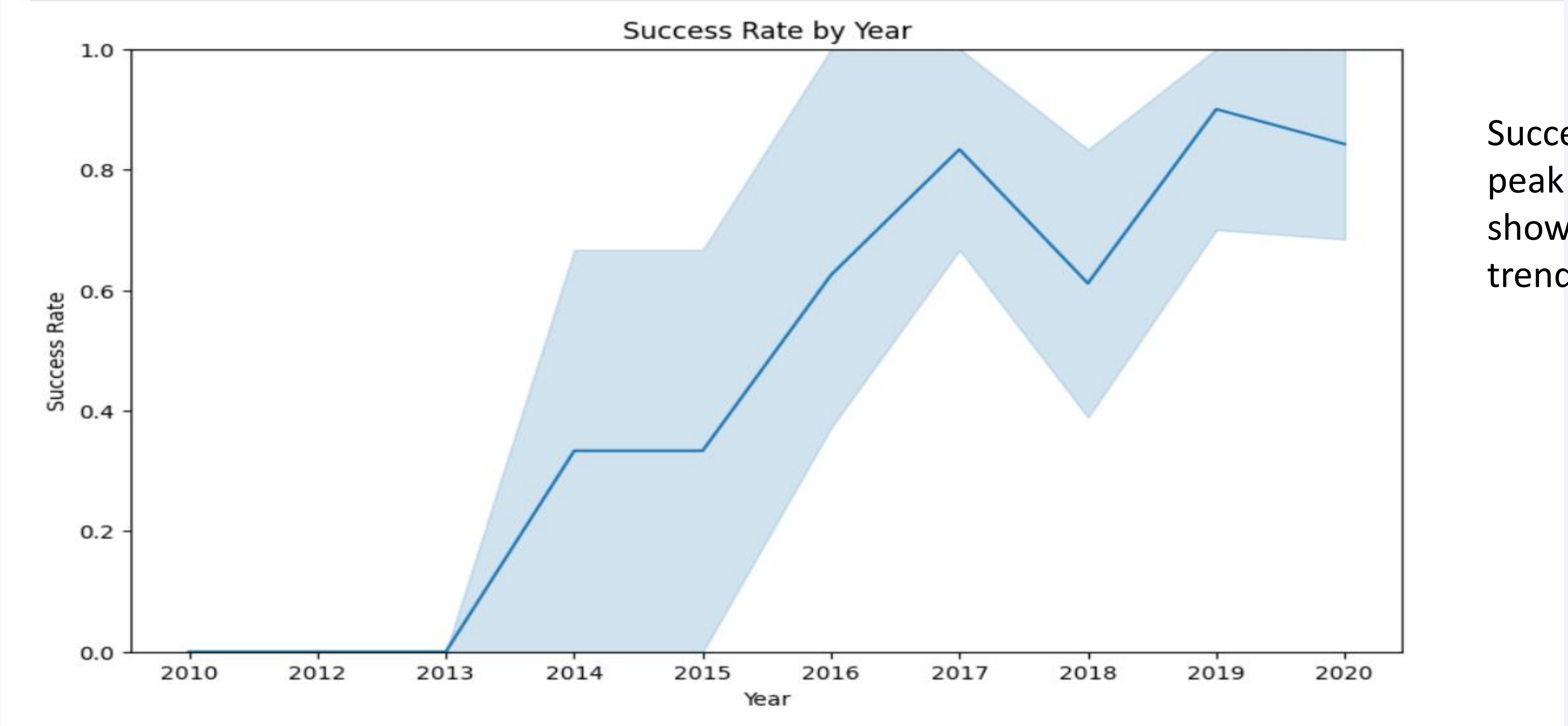
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



Success rate was peak in 2019 but shows a down trend in 2020.

https://github.com/Bhabhatunde/Data-Gathering/blob/main/edadataviz.ipynb

All Launch Site Names

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

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

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)

Total Payload Mass

```
%%sql
SELECT SUM("Payload_Mass__kg_") AS total_payload_mass
FROM SPACEXTABLE
WHERE "Booster_Version" LIKE 'Falcon 9%'
  AND "Customer" = 'NASA (CRS)';
```

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

total_payload_mass
None

Average Payload Mass by F9 v1.1

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

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

average_payload_mass
2928.4

First Successful Ground Landing Date

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

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

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 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

```
%%sql
SELECT
    "Landing_Outcome",
    COUNT(*) AS total_count
FROM SPACEXTABLE
WHERE "Landing_Outcome" IN ('Success', 'Failure')
GROUP BY "Landing_Outcome";
```

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

Landing_Outcome	total_count
Failure	3
Success	38

Boosters Carried Maximum Payload

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

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

```
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS month_name,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Failure (drone ship)'
  AND substr("Date", 0, 5) = '2015';
```

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

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

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

```
%%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;
```

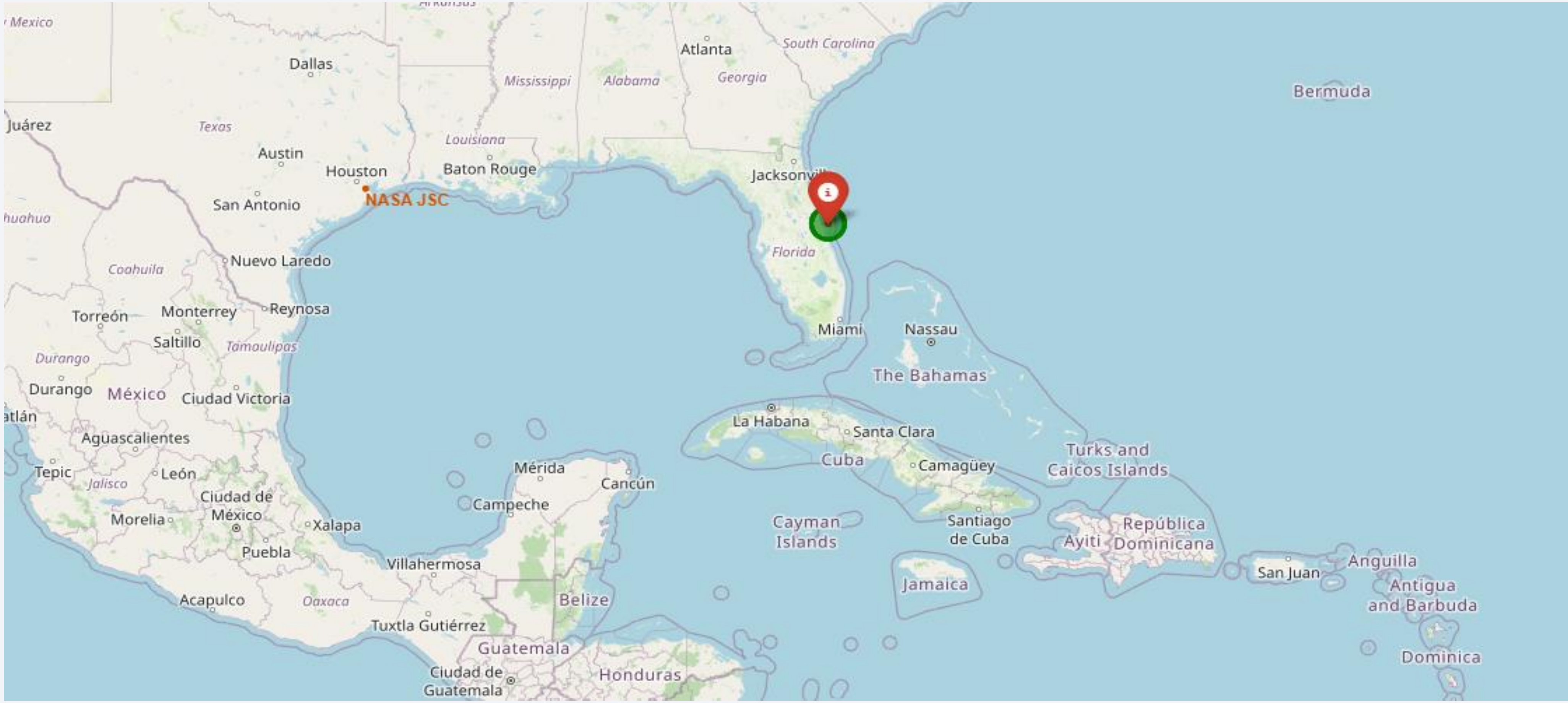
```
 * sqlite:///my_data1.db
Done.
```

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

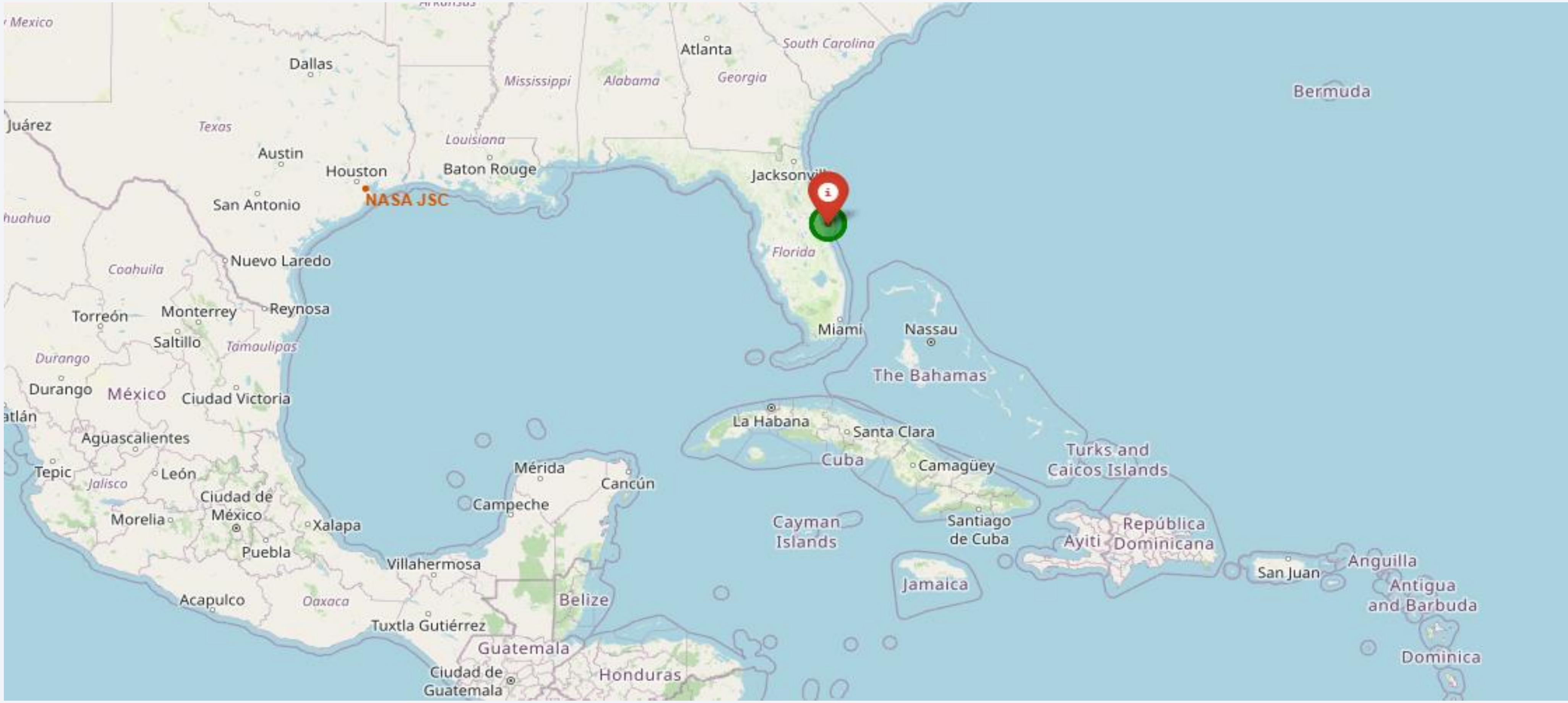
Section 3

Launch Sites Proximities Analysis

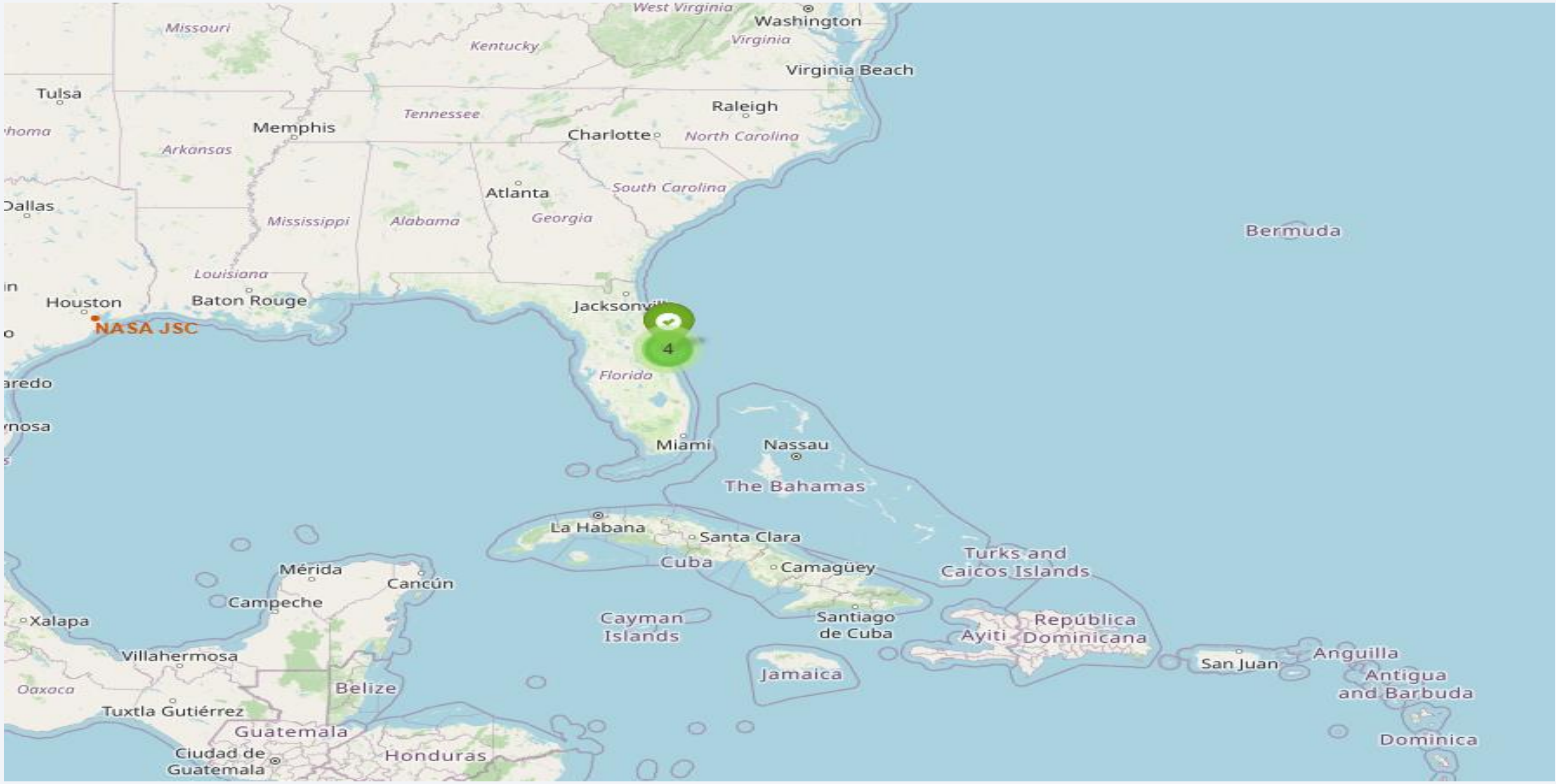
Launch sites' location markers on a global map



Color-labeled launch outcomes on the map



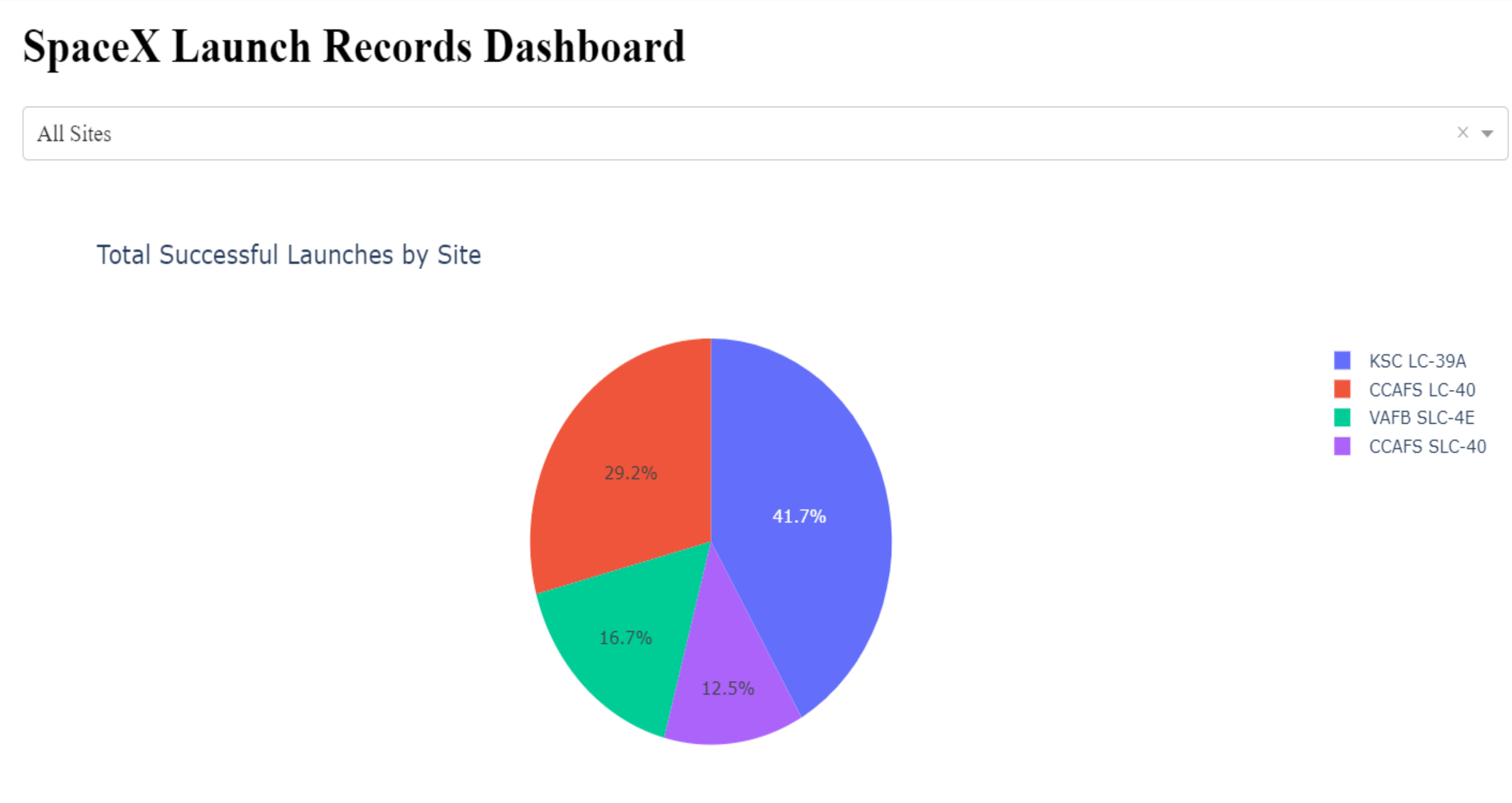
Launch and proximities such as railway, highway, coastline, with distance



Section 4

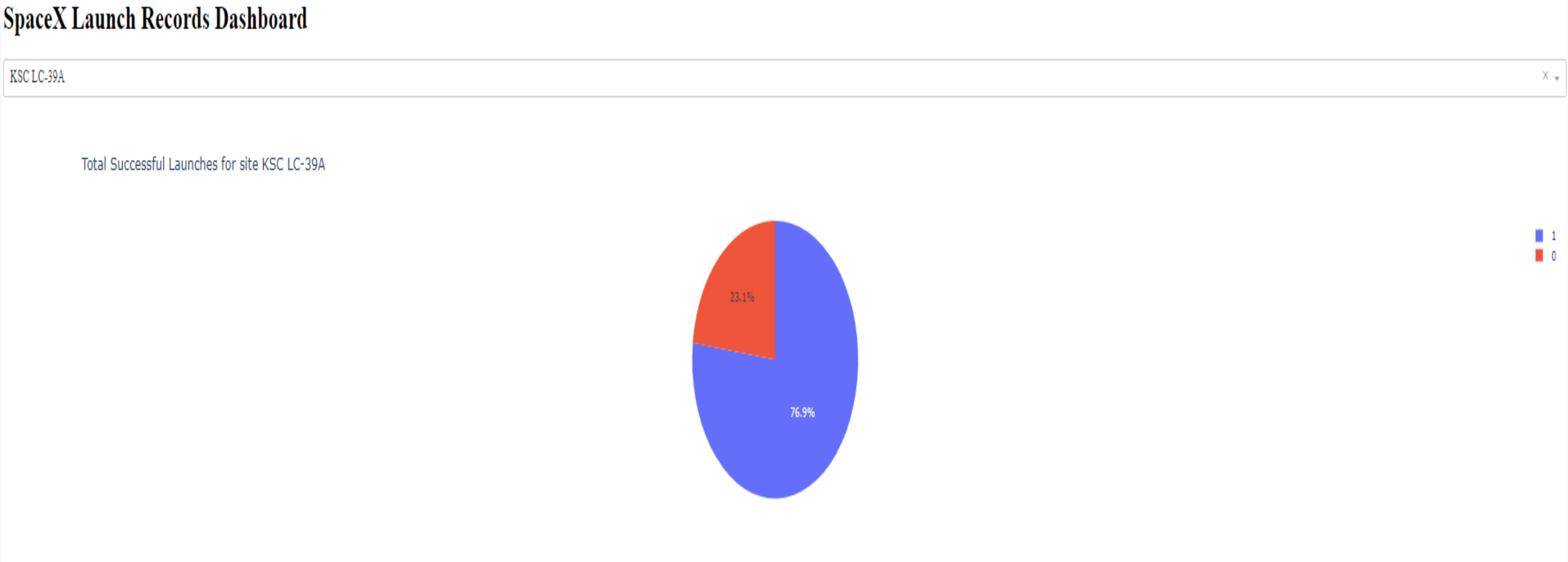
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard



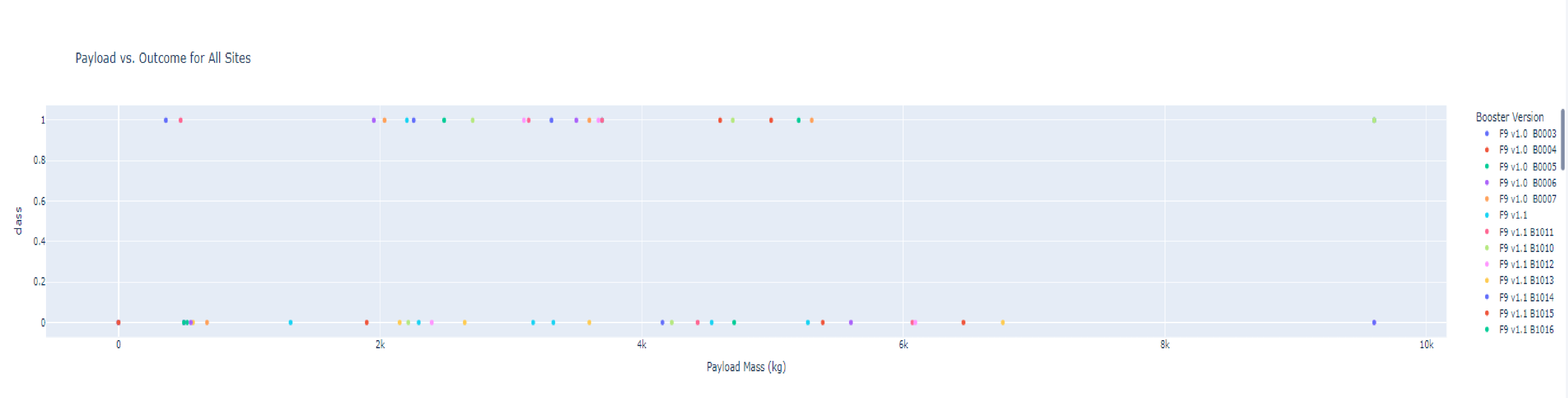
KSC LC 39A has the highest percentage of 41.7% successful launch followed by CCAFS LC 40.

Launch site with highest launch success ratio



KSC LC 39A is the only site with the highest launch success ratio.

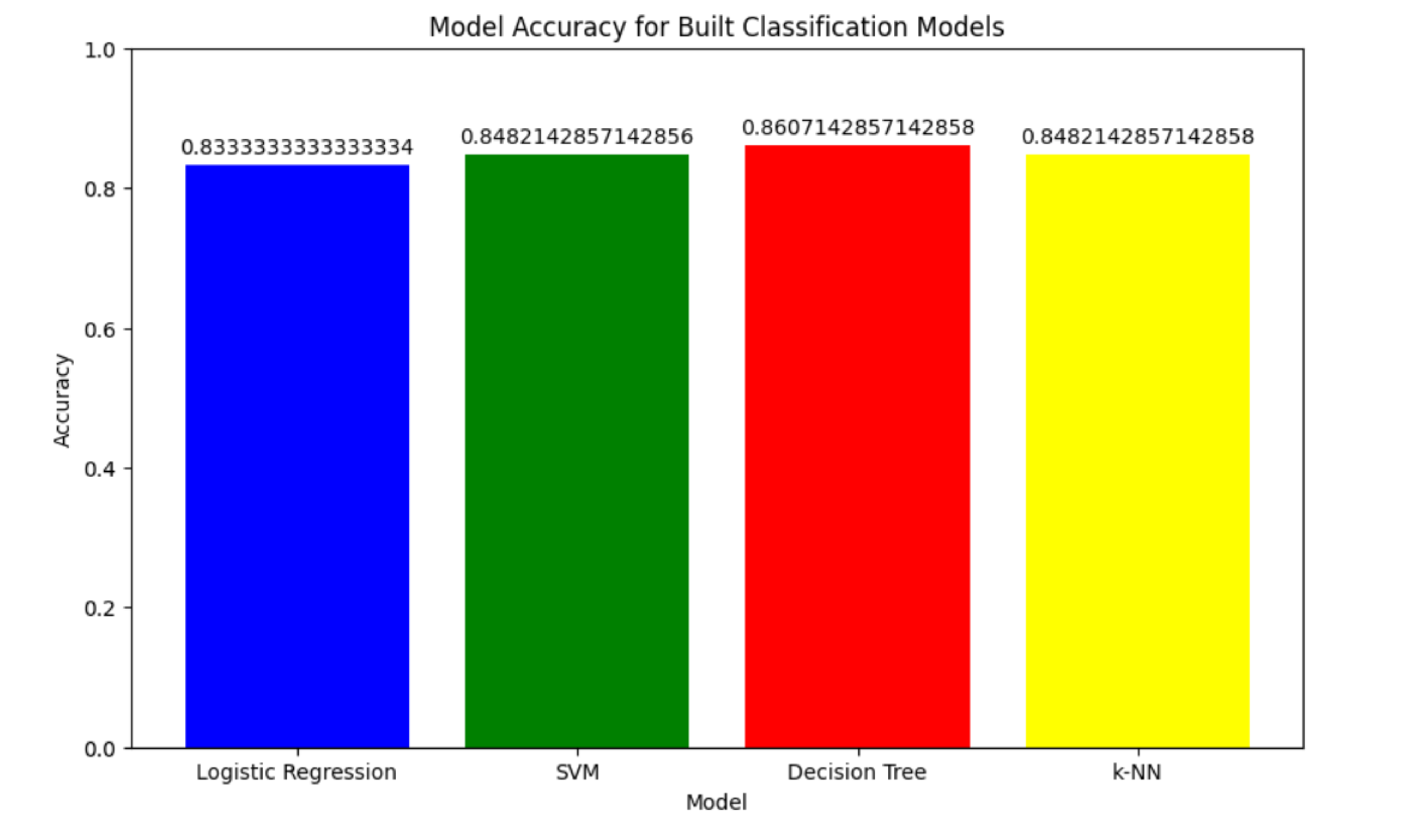
Payload vs. Launch Outcome



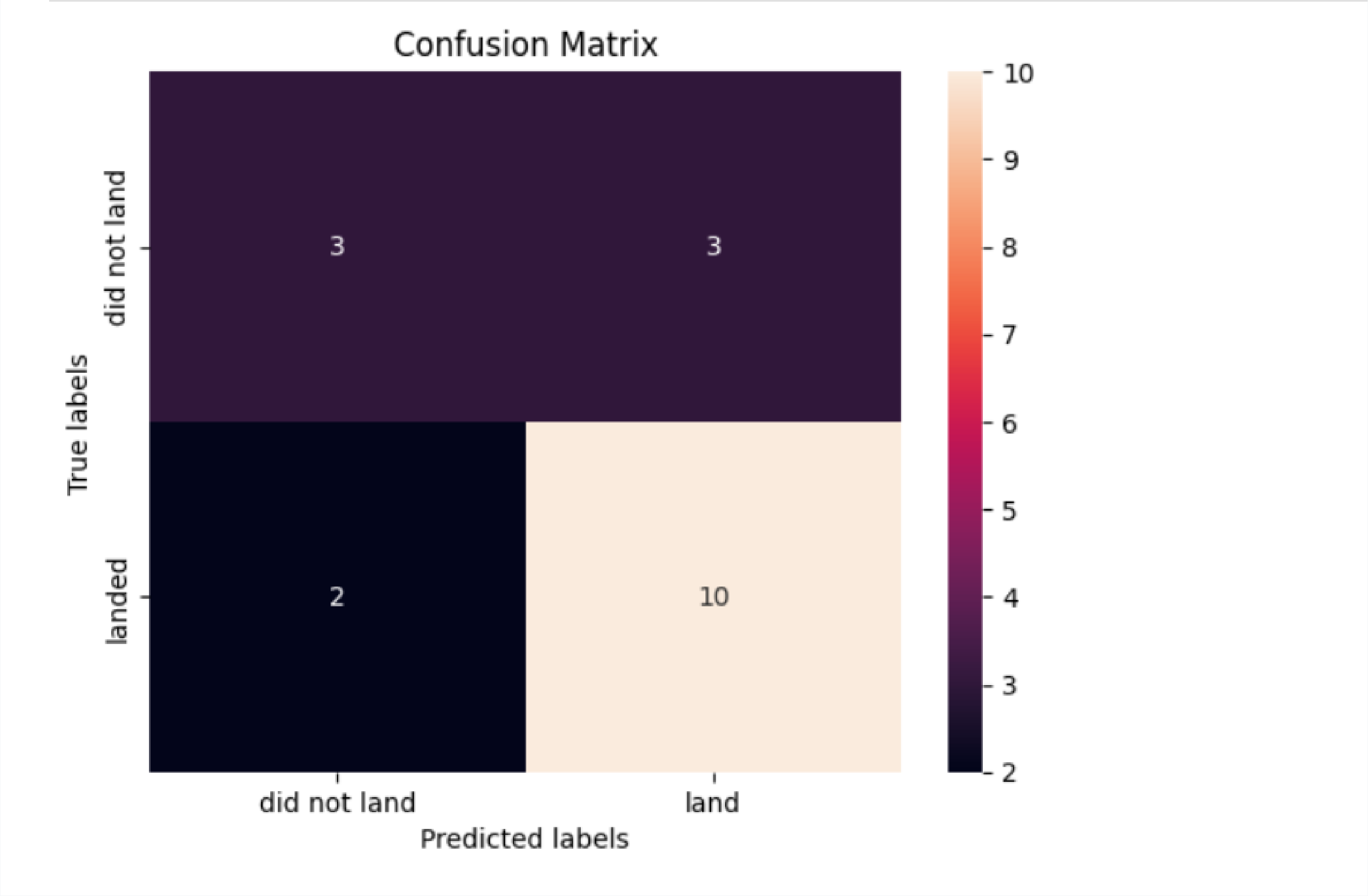
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

Key Findings:

1.Model Performance: Decision Tree models demonstrated efficacy in classification tasks, achieving an accuracy of approximately 86%. This indicates that the models is effective in predicting whether the first stage of the Falcon 9 will be reused.

2.Feature Engineering: Careful feature engineering contributed to the models' performance, enhancing their ability to make accurate predictions.

3.Hyperparameter Tuning: The use of GridSearchCV for hyperparameter tuning optimized the performance of the models, ensuring that they operated at their best possible configuration.

Implications:

•Cost Estimation: Accurate prediction of rocket stage reusability can help Space Y estimate launch costs more effectively, which is crucial for budgeting and pricing strategies in the competitive space industry.

•Decision-Making: The insights gained from the models can inform strategic decisions regarding the investment in reusable technology and other operational considerations.

Overall, the project has demonstrated the potential of machine learning in optimizing space launch strategies and provided a valuable tool for Space Y to gain a competitive edge in the commercial space market.

Appendix

The project successfully utilized data science and machine learning to address the challenge of predicting the reuse of SpaceX's Falcon 9 rocket first stages. By employing Logistic Regression and Support Vector Machine (SVM) models, the project aimed to estimate launch costs and improve decision-making for Space Y, a new competitor in the space industry.

https://github.com/Bhabhatunde/Data-Gathering/blob/main/Machine%20Learning_Revisit.ipynb

Thank you!