IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Pius Twumasi
25th May 2025

Outline

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

Executive Summary

This project aimed to predict the success of Falcon 9 first-stage landings, a key factor in SpaceX's cost efficiency. Data was collected via API and web scraping, then cleaned and analyzed using SQL, Python, and visualization tools.

Exploratory Data Analysis revealed that landing success improved over time, especially for later flights, heavier payloads, and certain orbits (e.g., VLEO) and launch sites (e.g., KSC LC-39A). Booster version category FT showed the highest success rate.

Four machine learning models were trained: Logistic Regression, SVM, Decision Tree, and KNN. All achieved the same test accuracy of 83.33%, with high recall and moderate precision, indicating strong predictive performance but a tendency to over-predict landings.

The models reliably predict landing outcomes, supporting data-driven insights into Falcon 9 performance.

Introduction

- **Background & Context:**
  SpaceX has transformed the aerospace industry with reusable rocket technology, significantly cutting launch costs. The Falcon 9, priced at $62 million per launch, relies heavily on the successful landing of its first-stage booster. Predicting landing success is crucial for cost forecasting, client confidence, and competitive analysis.
- **Objective:**
  This project aims to analyze historical launch data and build machine learning models to predict whether the Falcon 9 first stage will land successfully.
- **Key Questions:**
- *What factors influence landing success?*
- *Can we accurately predict the outcome using features like payload mass, launch site, orbit, and booster version?*
- *Which model performs best for this prediction task?*
- *Are there trends that indicate improved reliability over time?*
- **Goal:**
  To provide actionable insights and build a reliable model that supports decision-making in launch planning and evaluation.

Section 1

Methodology

Methodology

Overview of Project Workflow

In the following slides, we will walk through the key steps of the project:

Data Collection Methodology
– Overview of how SpaceX launch data was gathered using APIs and web scraping.

Data Wrangling & Preprocessing
– Steps taken to clean, format, and prepare the data for analysis.

Exploratory Data Analysis (EDA)
– Performed using SQL queries and data visualizations to uncover patterns and insights.

Interactive Visual Analytics
– Built interactive dashboards and geospatial maps using **Plotly Dash** and **Folium**.

Predictive Modeling
– Applied classification algorithms to predict landing outcomes.

Model Building, Tuning & Evaluation
– Compared model performance and optimized hyperparameters to select the best-performing classifier.

Data Collection

Extracted data from the SpaceX RESTful API

– Retrieved structured JSON data including launch_date, launch_site, payload_mass_kg, orbit, and landing_outcome.

Performed web scraping from Wikipedia and SpaceX sources

– Utilized BeautifulSoup to extract supplementary metadata: mission descriptions, customer names, booster versions, and landing types.

Transformed raw data into structured Pandas DataFrames

– Converted JSON and HTML data into tabular format for processing and analysis.

Conducted data preprocessing and wrangling

– Addressed missing values, normalized data types, removed duplicates, and excluded test/dev launches.

Engineered relevant features

– Created new variables such as:

Class (binary outcome: 1 = success, 0 = failure)

Booster_Version_Category

Launch_Year

Payload_Category (e.g., light, medium, heavy)

Prepared finalized dataset

– Ready for downstream tasks including Exploratory Data Analysis (EDA), visualization, and supervised machine learning modeling.

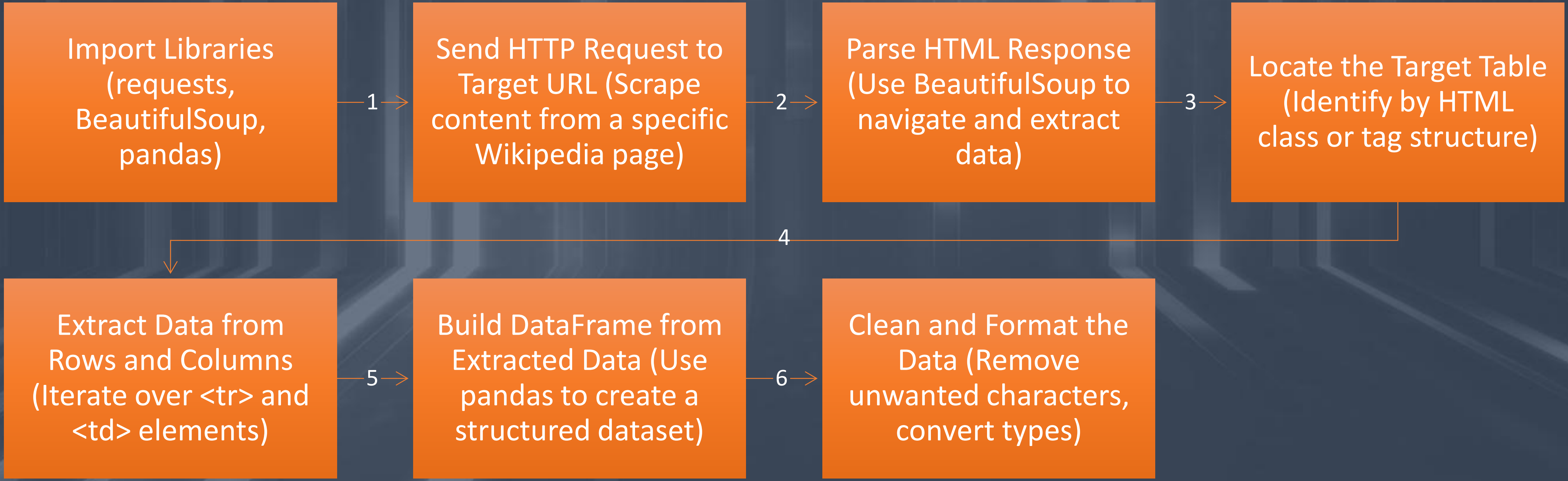
Data Collection – SpaceX API

View GitHub Notebook

1. Import Libraries
→ Used requests, pandas, json, etc.
2. GET Request to SpaceX API Endpoint
→ Retrieved launch data using: https://api.spacexdata.com/v4/launches
3. Parse JSON Response
→ Converted API response to a Python dictionary
4. Create DataFrame
→ Extracted relevant fields to construct a pandas DataFrame
5. Filter for Specific Launch Details
→ Filtered for columns such as mission name, launch date, rocket ID, launchpad ID
6. Fetch Additional Data Using Rocket and Launchpad IDs
→ Made further API requests to rockets and launchpads endpoints using IDs from the main data
7. Merge and Finalize Dataset
→ Combined launch data with rocket and launchpad info
→ Cleaned and structured final dataset

Data Collection - Scraping

[View GitHub Notebook]



Data Wrangling

[View GitHub Notebook]

1. Data Loading

Read dataset from URL (dataset_part_1.csv).

2. Exploratory Data Analysis (EDA)

Check missing values (df.isnull().sum()).

Identify column types (numerical/categorical).

3. Task 1: Launch Site Analysis

Count launches per site (df["LaunchSite"].value_counts()).

4. Task 2: Orbit Analysis

Count occurrences of each orbit type (df["Orbit"].value_counts()).

5. Task 3: Mission Outcomes

Categorize outcomes (df["Outcome"].value_counts()).

Define bad_outcomes (unsuccessful landings).

6. Task 4: Label Creation

Generate binary Class labels:

1 = Successful landing (True ASDS, True RTLS, True Ocean).

0 = Failed landing (bad_outcomes).

7. Result

Add Class column to DataFrame.

Calculate success rate (df["Class"].mean()).

Export to CSV (dataset_part_2.csv).

EDA with Data Visualization.

Summary of Charts Plotted and Their Purposes:

1. Catplot: Payload Mass vs. Flight Number (Hue = Class)

Purpose: To examine whether payload mass affects landing success over time (by flight number).

2. Catplot: Launch Site vs. Flight Number (Hue = Class)

Purpose: To identify which launch sites have better success rates over the launch history.

3. Catplot: Launch Site vs. Payload Mass (Hue = Class)

Purpose: To analyze how payload mass influences success depending on the launch site.

[View GitHub Notebook]

EDA with Data Visualization.

Summary of Charts Plotted and heir Purposes: (Continued)

4. Bar Plot: Success Rate by Orbit

- Purpose: To compare average landing success rates across different orbit types.

5. Scatter Plot: Orbit vs. Flight Number (Hue = Class)

- Purpose: To show which orbits are associated with successes/failures over time.

6. Scatter Plot: Orbit vs. Payload Mass (Hue = Class)

- Purpose: To investigate how different orbit types correlate with payload mass and landing outcomes.

7. Line Plot: Date vs. Success Rate

- Purpose: To track the trend of success rate over time, highlighting improvements or changes in landing performance.

EDA with SQL (Summary of Queries)

1. Created a filtered table:

Code: CREATE TABLE SPACEXTABLE AS SELECT * FROM SPACEXTBL WHERE Date IS NOT NULL;

Filters out rows where the Date is missing.

2. Retrieved distinct launch sites:

Code: SELECT DISTINCT Launch_site FROM SPACEXTABLE;

Identifies all unique launch site locations.

3. Previewed records for a specific site:

Code: SELECT * FROM SPACEXTABLE WHERE Launch_site LIKE "CCA%" LIMIT 5;

Fetches 5 rows for launch sites starting with "CCA".

4. Calculated total payload for NASA missions:

Code: SELECT SUM(COALESCE(PAYLOAD_MASS__KG_, 0)) AS TOTAL_PAYLOAD_BY_NASA_CRS

FROM SPACEXTABLE WHERE Customer = "NASA (CRS)";

Computes total payload launched for NASA CRS missions.

5. Calculated average payload for F9 v1.1 boosters:

Code: SELECT AVG(COALESCE(PAYLOAD_MASS__KG_, 0)) AS AVG_PAYLOAD_BY_F9v1

FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1";

6. Identified first successful ground landing:

Code: SELECT MIN(Date) AS First_successful_landing

FROM SPACEXTABLE WHERE Landing_Outcome = "Success (ground pad)";

View GitHub Notebook

EDA with SQL (Summary of Queries)

7. Filtered booster versions with successful drone ship landings and moderate payloads

```
Code: SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

8. Counted total mission outcomes (Success or Failure):

```
Code: SELECT COUNT(*) AS Total_mission_outomes
FROM SPACEXTABLE
WHERE Mission_Outcome = "Success" OR Mission_Outcome = "Failure";
```

9. Retrieved booster with the maximum payload mass:

```
Code: SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

10. Grouped data for 2015 by month and landing failure types:

```
Code: SELECT SUBSTR(Date,6,2) AS Months, (Landing_outcome = "Failure (drone ship)") AS Landing_outcome, Booster_version, Launch_Site
FROM SPACEXTABLE
WHERE SUBSTR(Date,0,5) = "2015"
GROUP BY Months, Landing_outcome, Booster_version, Launch_Site;
```

11. Ranked landing outcomes by frequency:

```
Code: SELECT
Landing_Outcome,
COUNT(*) AS Outcome_Count,
RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank
FROM SPACEXTABLE
WHERE Date BETWEEN "2010-06-04" AND "2017-03-20"
GROUP BY Landing_outcome
ORDER BY Outcome_Count DESC;
```

Interactive Map with Folium

Map Objects Added to the Folium Map

- **Markers:**
  You added markers to indicate the location of NASA launch sites and nearby infrastructure (e.g., coastline, railway). Custom DivIcon markers were used to label the sites and display distance information (e.g., distance to coast or railway).
- **Circles:**
  Circles were drawn around launch sites to highlight them visually and denote a fixed radius (50 meters) as a visual reference zone.
- **Polylines:**
  Lines were drawn between launch sites and nearby features (e.g., coastlines, railways) to show spatial relationships and calculate distances.
- **MarkerCluster:**
  A marker cluster was used to group success/failure launch markers based on data from spacex_df, improving map readability.
- **MousePosition Plugin:**
  A coordinate tracker was added to the map using MousePosition to help identify locations precisely by hovering over them.

Why These Objects Were Added

- Markers provided clear identification of important locations and allowed displaying information such as distances and names.
- Circles visually emphasized the launch sites and created a reference radius to support spatial analysis.
- Polylines helped visualize the connection and measure distances between launch sites and nearby geographical or infrastructure features.
- MarkerCluster managed marker clutter and illustrated launch outcomes (success or failure) more effectively.
- MousePosition enhanced interactivity and usability by showing real-time geographic coordinates, aiding analysis.

View GitHub Notebook

Dashboard with Plotly Dash

Graphs and Interactive Features Added to Dashboard

- Pie Chart (success-pie-chart)
  Shows the total number of successful launches by site.
  When a specific launch site is selected, it shows the breakdown of successful vs. failed launches for that site.
- Scatter Plot (success-payload-scatter-chart)
  Displays the correlation between payload mass and launch outcome.
  Each point represents a launch, color-coded by booster version category.

Interactive Features

- Dropdown Menu (site-dropdown)
  Allows users to select a specific launch site or view data for all launch sites.
  Updates both the pie chart and scatter plot based on the selection.
- Payload Range Slider (payload-slider)
  Enables users to filter launches by payload mass range.
  Updates the scatter plot to reflect only launches within the selected payload range.

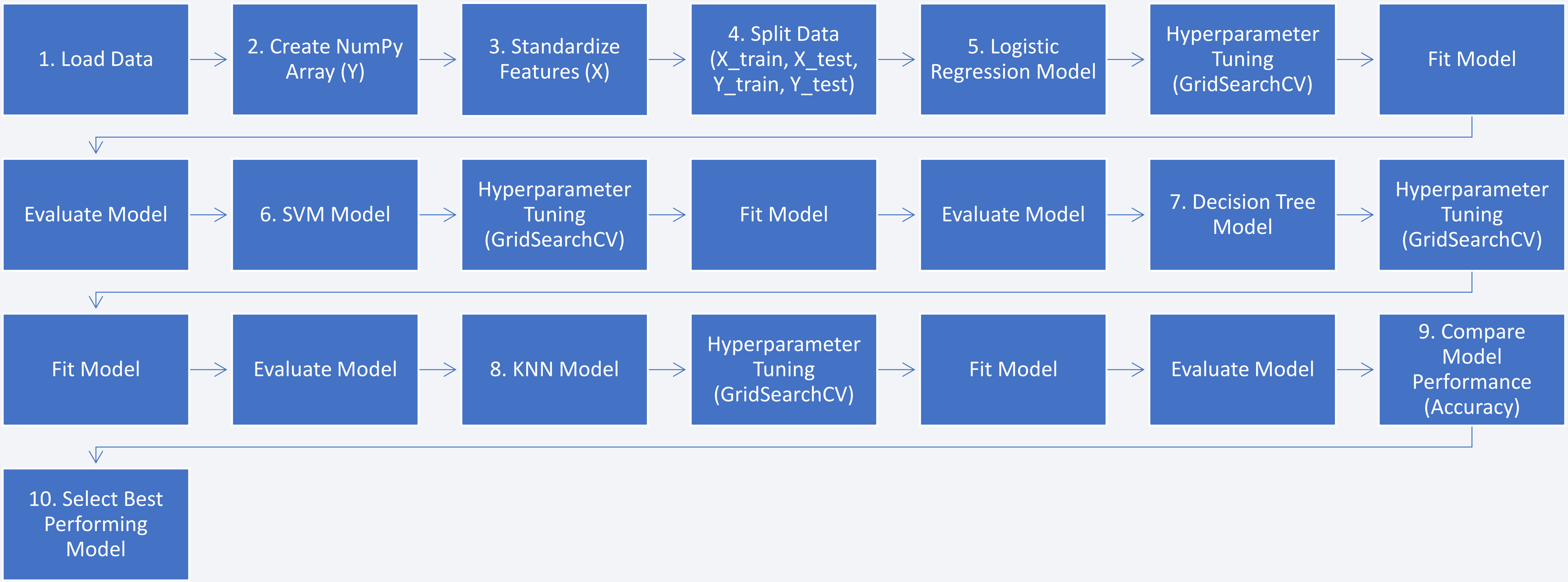
Why These Plots and Interactions Were Added

- Pie Chart helps users quickly understand the overall performance and compare launch success across different sites or within a specific site.
- Scatter Plot provides insight into how payload mass affects launch outcomes and whether certain booster versions perform better.
- Dropdown Interaction allows focused analysis by site, supporting both broad and detailed investigations.
- Slider Interaction enhances data exploration by enabling dynamic filtering based on payload weight, making the dashboard more flexible and informative.

View GitHub Notebook

Predictive Analysis (Classification)

View GitHub Notebook

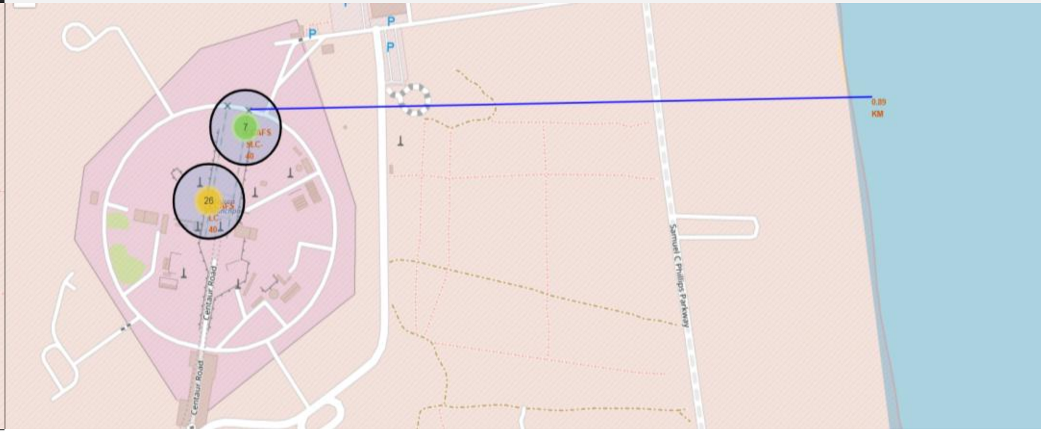


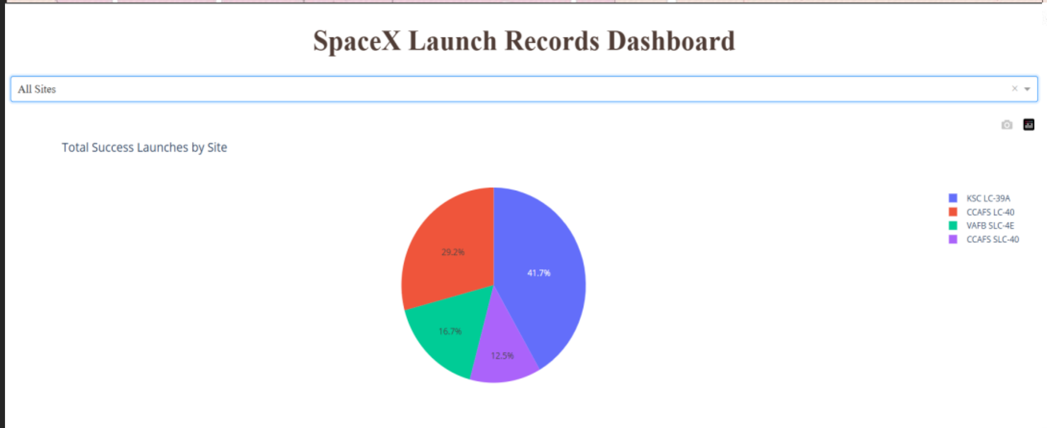
Results (Exploratory Data Analysis)

- **Flight Number & Launch Success**: Success rates (Class 1) increase with higher flight numbers, indicating improved outcomes over time.
- **Launch Site Performance**: CCAFS SLC 40 and KSC LC 39A show higher success rates, particularly with heavier payloads. VAFB SLC 4E has more failures and handles lighter payloads.
- **Orbit & Payload Impact**:
  - Orbits like ES-L1, GEO, HEO, and SSO have 100% success.
  - GTO has the lowest success rate (~53%), but performs well for 3000–7000 kg payloads.
  - VLEO shows high success for heavy payloads.
- **Time Trend**: Success rates improved significantly from 2013 to 2017, stabilizing at higher levels in later years.
- **SQL Insights**:
  - Most missions originated from CCAFS LC-40.
  - NASA (CRS) payload total: 45,596 kg.
  - First successful ground landing: 2015-12-22.

Results (Interactive Analytics Demo)

- *Interactive dashboards and maps were developed using Plotly Dash and Folium to explore launch patterns, site performance, and mission details. Some screenshots are inserted here to demonstrate these visualizations.*





Results (Predictive Analysis)

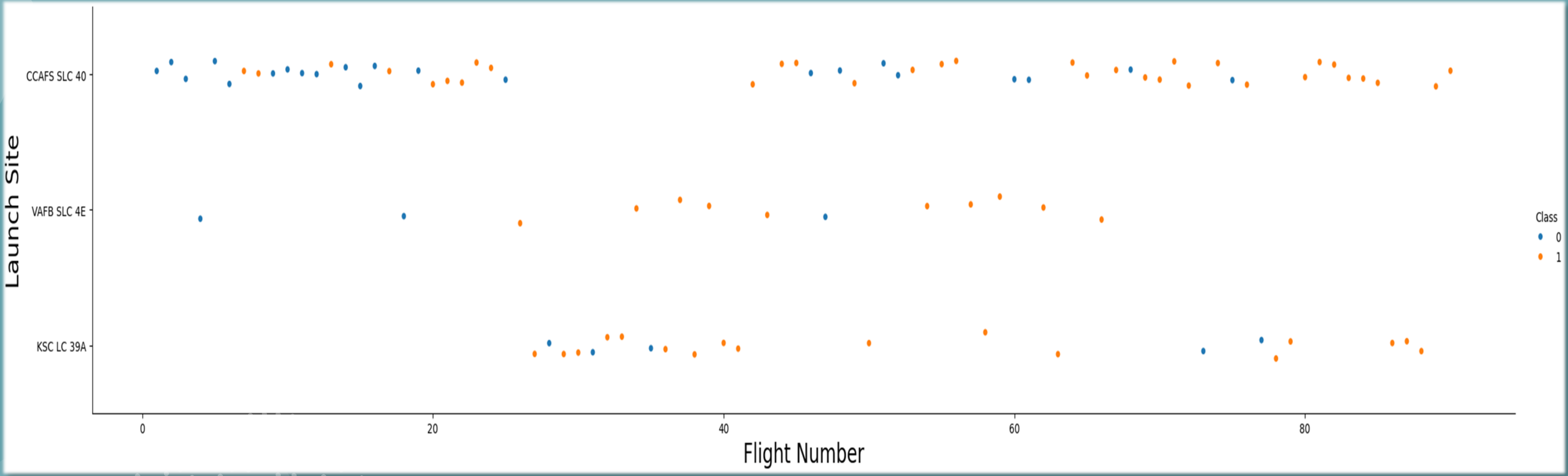
- **Models Used**: Logistic Regression, SVM, Decision Tree, and KNN.
- **Test Accuracy**: All models achieved **~83.33%** accuracy on the test set.
- **Confusion Matrix Insights**:
  - **True Positives**: 12
  - **True Negatives**: 3
  - **False Positives**: 3
  - **False Negatives**: 0
  - **Recall**: 100% (for "landed")
  - **Precision**: 80% (for "landed")
- **Conclusion**: The models are highly effective at identifying successful landings, though they slightly over-predict landings. Consistent accuracy across models suggests stable predictive power from the dataset's feature

Section 2

Insights drawn from EDA

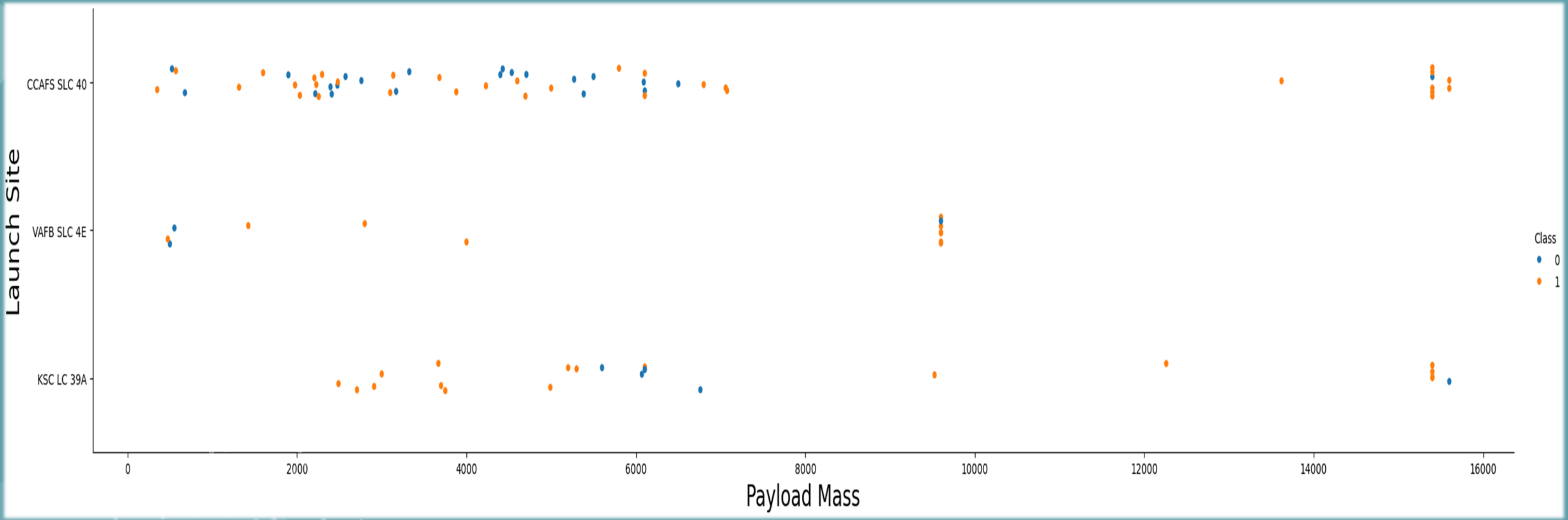
Flight Number vs. Launch Site

This 1D scatter plot shows that as Flight Number increases, successful launches (Class 1) become more frequent, especially at CCAFS SLC 40 and KSC LC 39A.



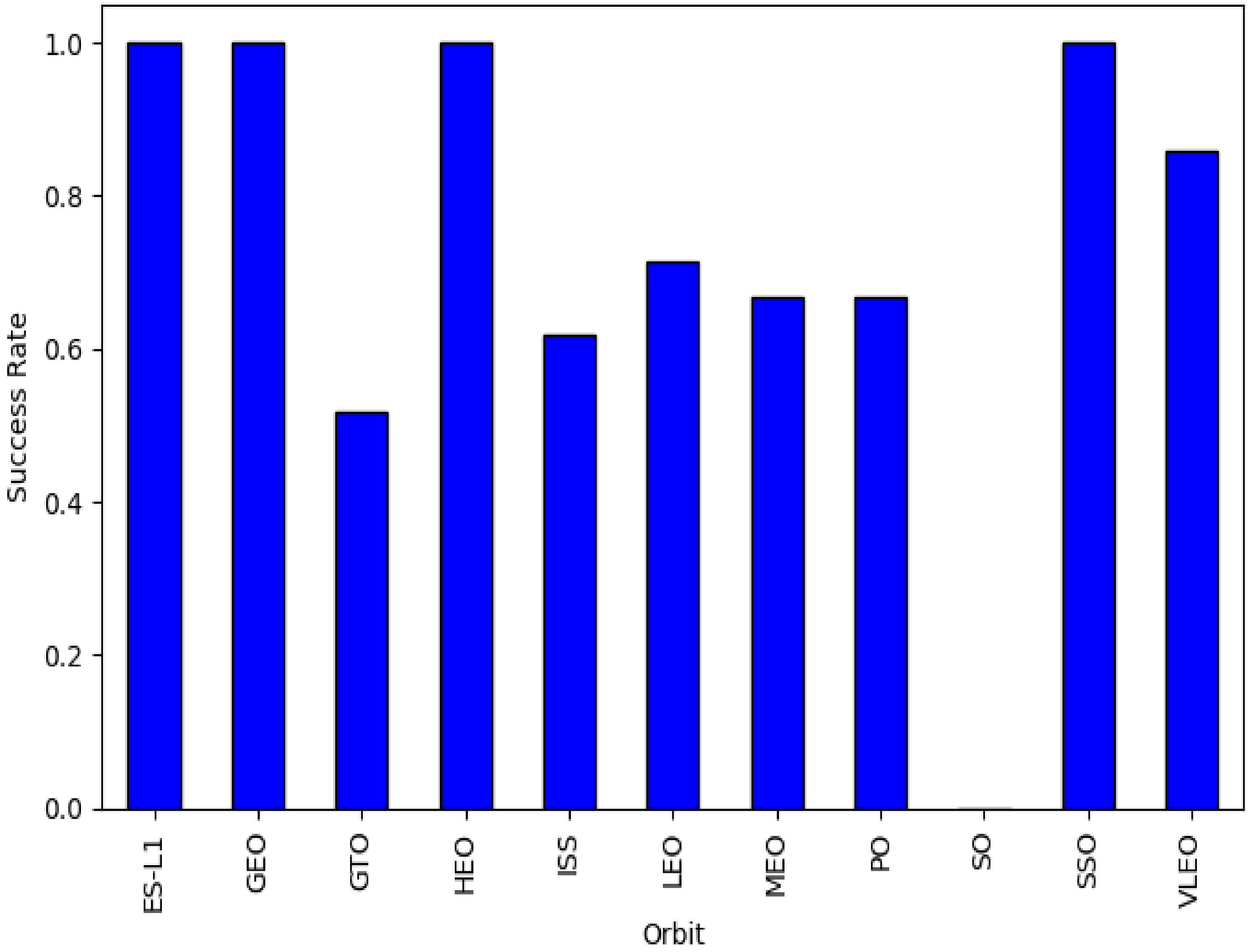
Payload vs. Launch Site

This plot shows that high payload launches (above 10,000) mostly occur at CCAFS SLC 40 and KSC LC 39A, with more successful outcomes (Class 1). VAFB SLC 4E handles lighter payloads. Success is generally less common for very light payloads.



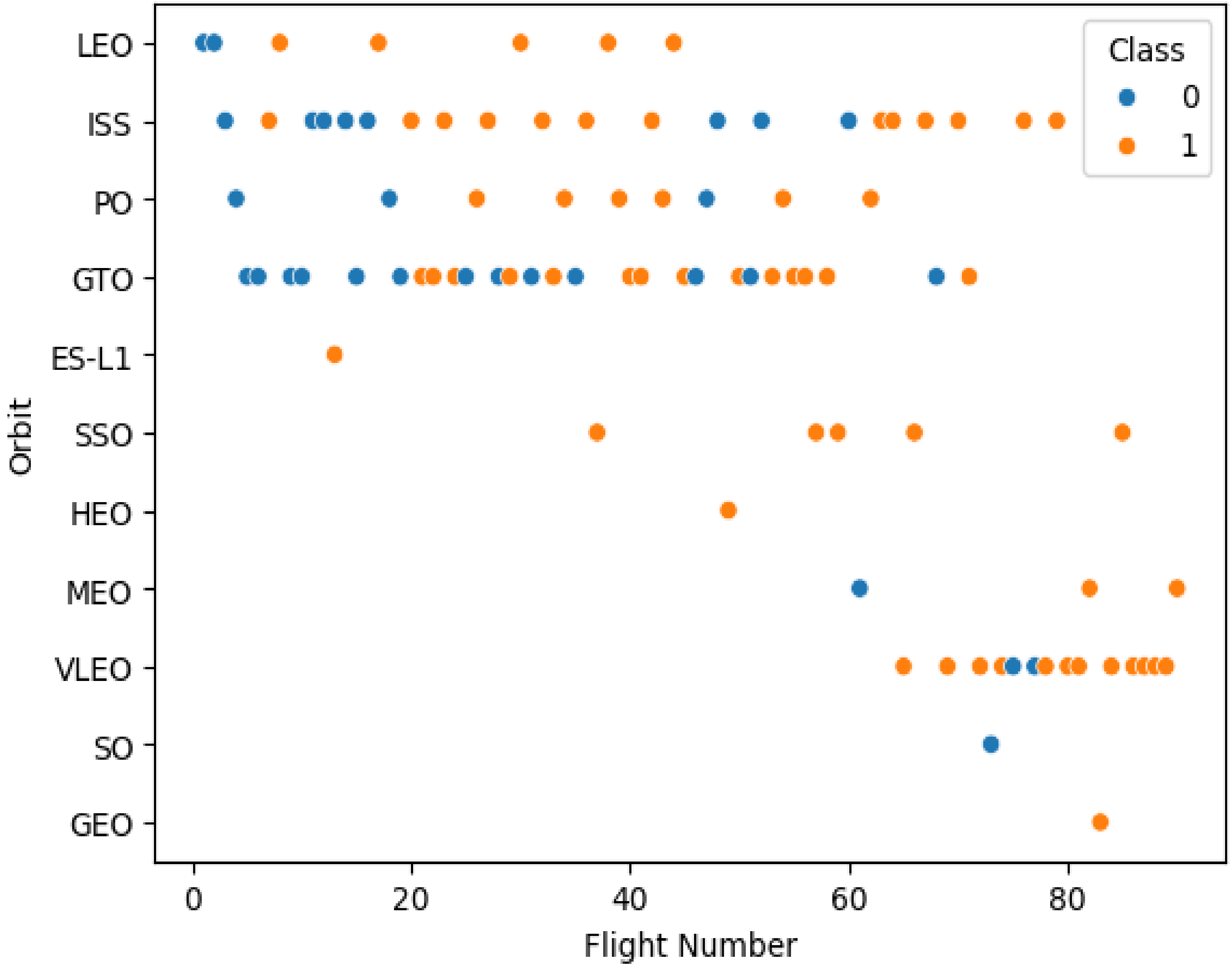
Success Rate vs. Orbit Type

This bar chart shows that ES-L1, GEO, HEO, and SSO orbits have a 100% success rate. GTO has the lowest at ~53%, while ISS, LEO, MEO, and PO have moderate rates (62–72%). VLEO performs well at ~85%, and SO has no recorded successes



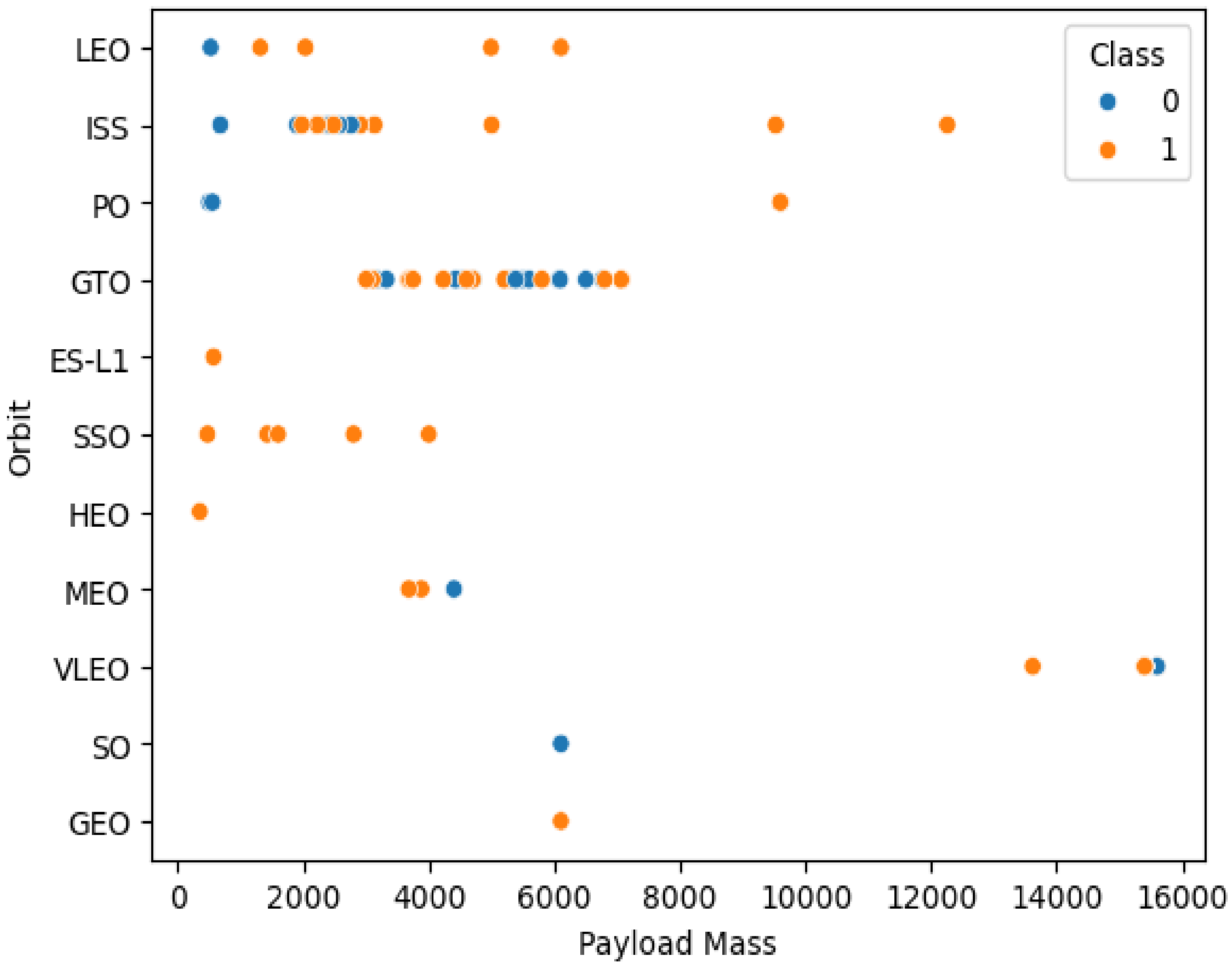
Flight Number vs. Orbit Type

This scatter plot shows that for orbits with more data (like LEO, ISS, PO, GTO), success (Class 1) increases with flight number. Orbits like SSO, HEO, ES-L1, and GEO tend to show more Class 1 outcomes, though data is limited. Failures (Class 0) are more common at lower flight numbers across several orbits.



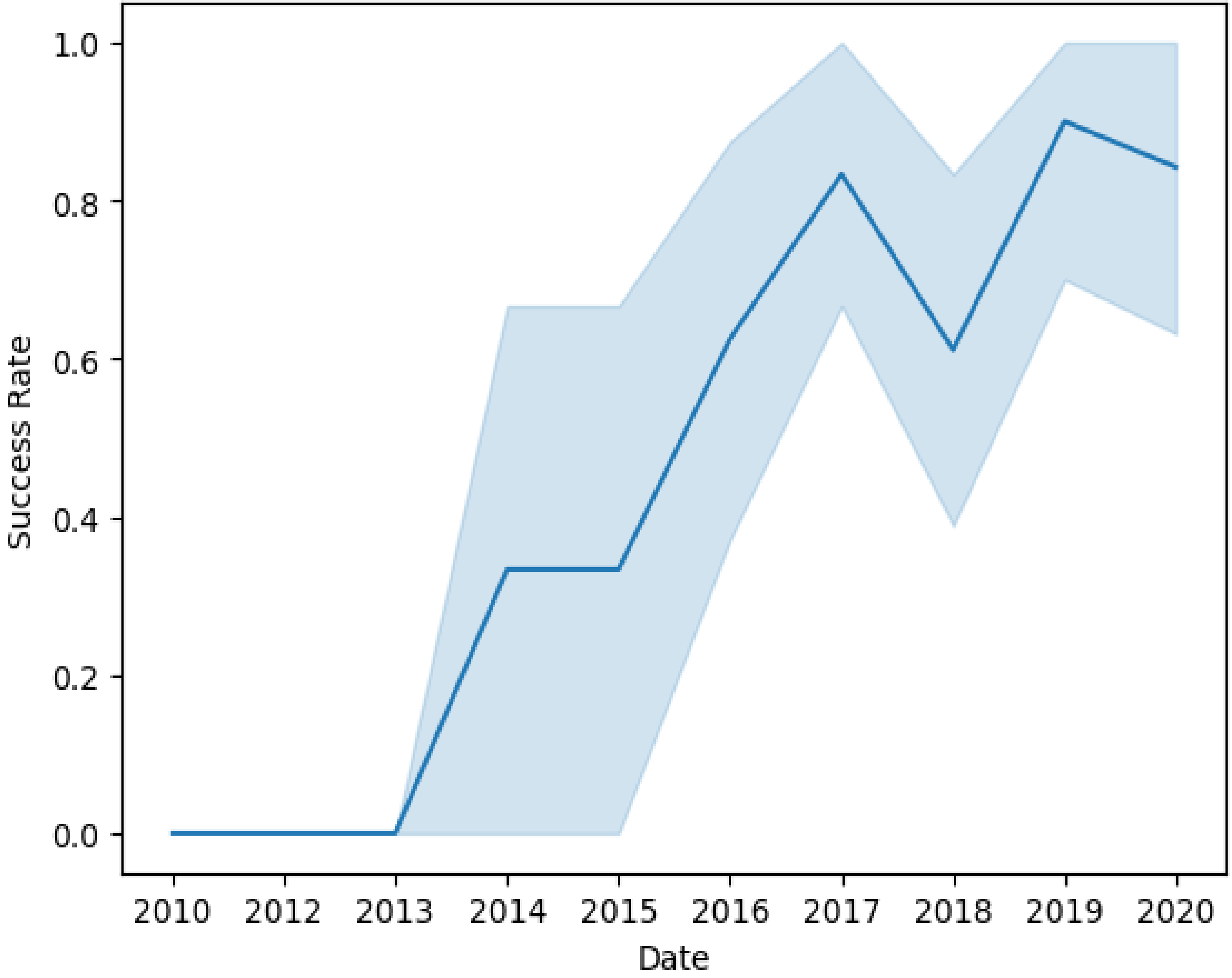
Payload vs. Orbit Type

This plot shows that very high payloads (above 14,000 kg) with Class 1 outcomes mostly go to VLEO. GTO sees many successes in the 3,000–7,000 kg range. LEO and ISS have mixed results across all payloads, while ES-L1, HEO, SO, and GEO have limited data. SSO sees only success with a low payload.



Launch Success Yearly Trend

This line plot shows a clear rise in success rate from 2013 to 2017, with some fluctuations afterwards, but overall higher performance. The wider confidence interval from 2014–2018 suggests more variability, while the narrowing in 2019 indicates more consistent outcomes.



All Launch Site Names

The space missions were launched from four distinct sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

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

Launch Site Names Begin with 'CCA'

The first five launches from sites starting with 'CCA' were all from CCAFS LC-40. These early missions had low or zero payload mass, and all were successful. However, landing outcomes varied, including "Failure (parachute)" and "No attempt".

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

The total payload mass carried by boosters launched for "NASA (CRS)" customers is 45,596 kg.

	TOTAL_PAYLOAD_BY_NASA_CRS
0	45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version "F9 v1.1" is 2928.4 kg.

	AVG_PAYLOAD_BY_F9v1
0	2928.4

First Successful Ground Landing Date

The first successful landing on a ground pad ("Success (ground pad)") occurred on 2015-12-22.

	First_successful_landing
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Several booster versions achieved "Success (drone ship)" with payload masses between 4000 kg and 6000 kg. These include: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of missions with either "Success" or "Failure" as their Mission Outcome is 98.

	Total_mission_outomes
0	98

Boosters Carried Maximum Payload

Multiple booster versions carried the maximum payload mass. These include: 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, and F9 B5 B1049.7.

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

In the year 2015, there were two recorded "Failure (drone ship)" landing outcomes:

- In January (month '01'), with Booster Version F9 v1.1 B1012 from CCAFS LC-40.
- In April (month '04'), with Booster Version F9 v1.1 B1015 from CCAFS LC-40.

	Month	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

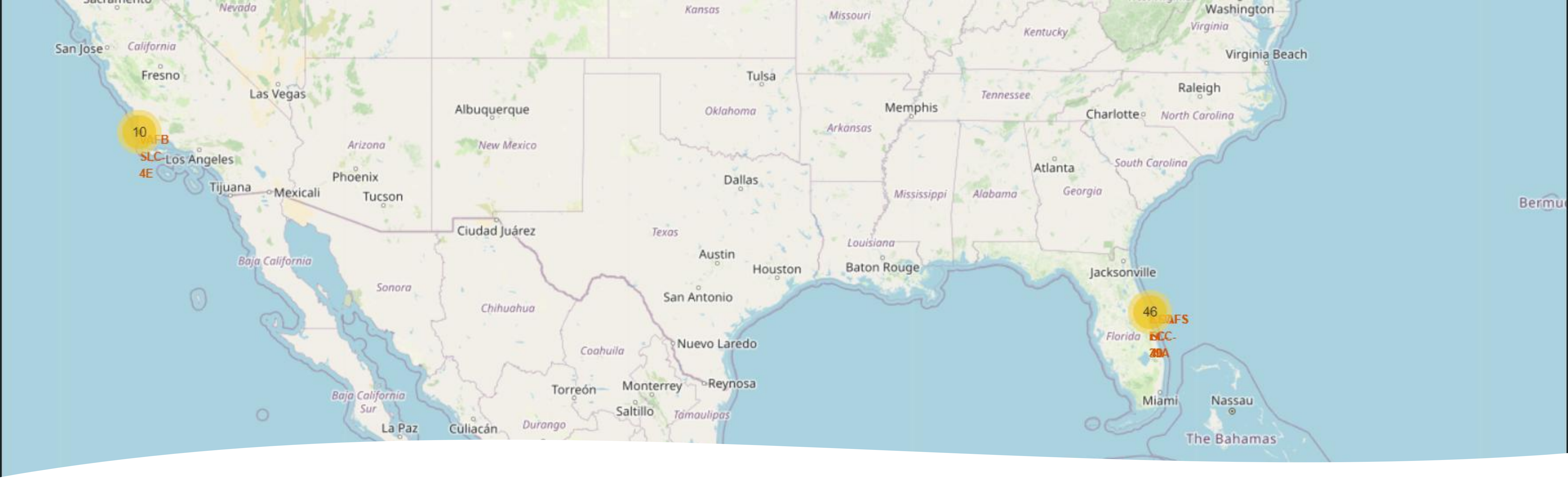
For the period between June 4, 2010, and March 20, 2017, the landing outcomes are ranked as follows:

- "No attempt" is the most frequent outcome with 10 occurrences (Rank 1).
- "Success (drone ship)" and "Failure (drone ship)" both occurred 5 times (Rank 2).
- "Success (ground pad)" and "Controlled (ocean)" both occurred 3 times (Rank 4).
- "Uncontrolled (ocean)" and "Failure (parachute)" both occurred 2 times (Rank 6).
- "Precluded (drone ship)" occurred once (Rank 8).

	Landing_Outcome	Outcome_Count	Rank
0	No attempt	10	1
1	Success (drone ship)	5	2
2	Failure (drone ship)	5	2
3	Success (ground pad)	3	4
4	Controlled (ocean)	3	4
5	Uncontrolled (ocean)	2	6
6	Failure (parachute)	2	6
7	Precluded (drone ship)	1	8

Section 3

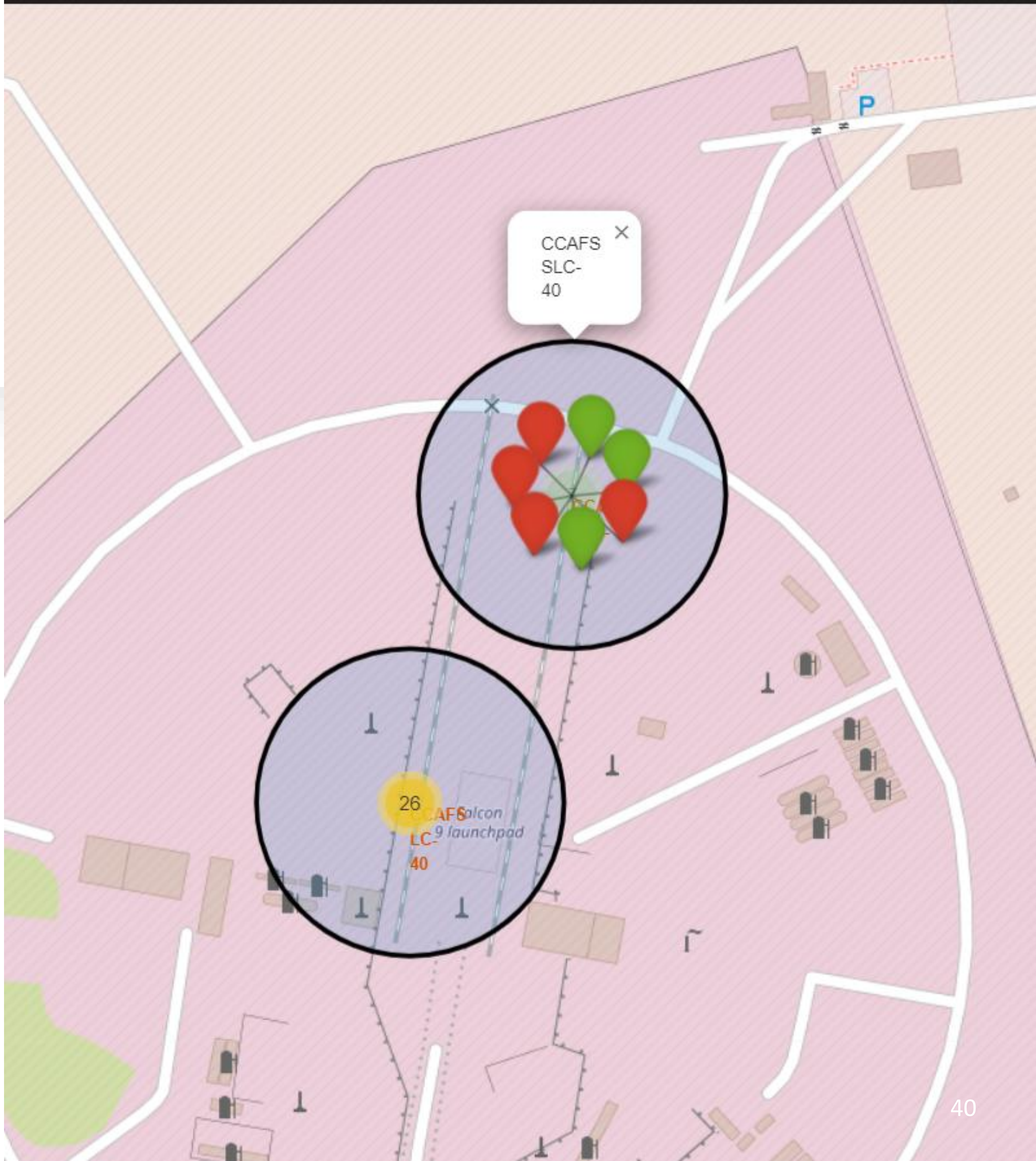
Launch Sites Proximities Analysis



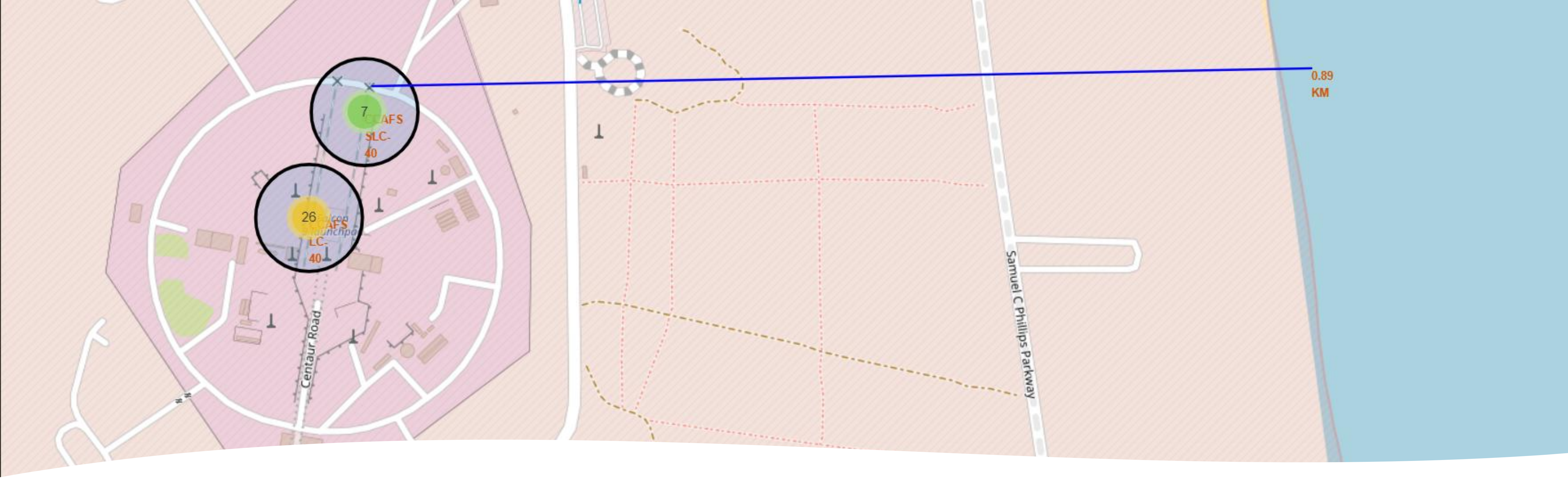
Launch Sites' Location Markers

This map uses markers to clearly identify launch site locations, displaying key details like site names for easy reference and visualization.

Color Labeled Lauch Outcomes



This map highlights the launch site "CCAFS SLC-40" using a marker cluster to group launch outcomes. Green markers represent successful launches, while red markers indicate failures, providing a clear visual summary of launch performance at the site.



Selected Launch Site to Proximity (Coastline)

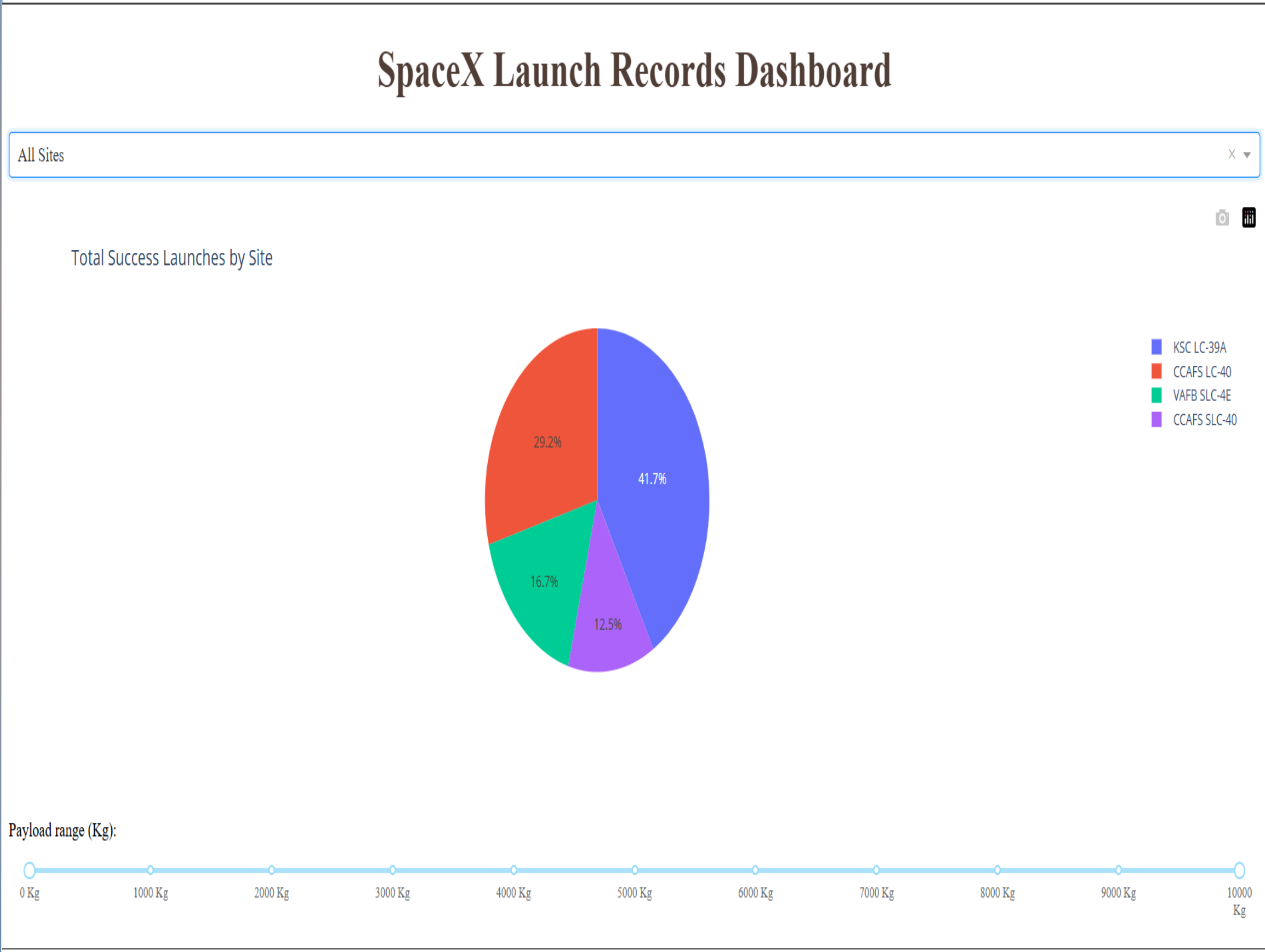
This map displays the launch site "CCAFS SLC-40" with a polyline indicating its proximity to the coastline. The polyline visually highlights the site's connection to nearby geographical features, showing a distance of approximately 0.89 km from the coast.

Section 4

Build a Dashboard with Plotly Dash

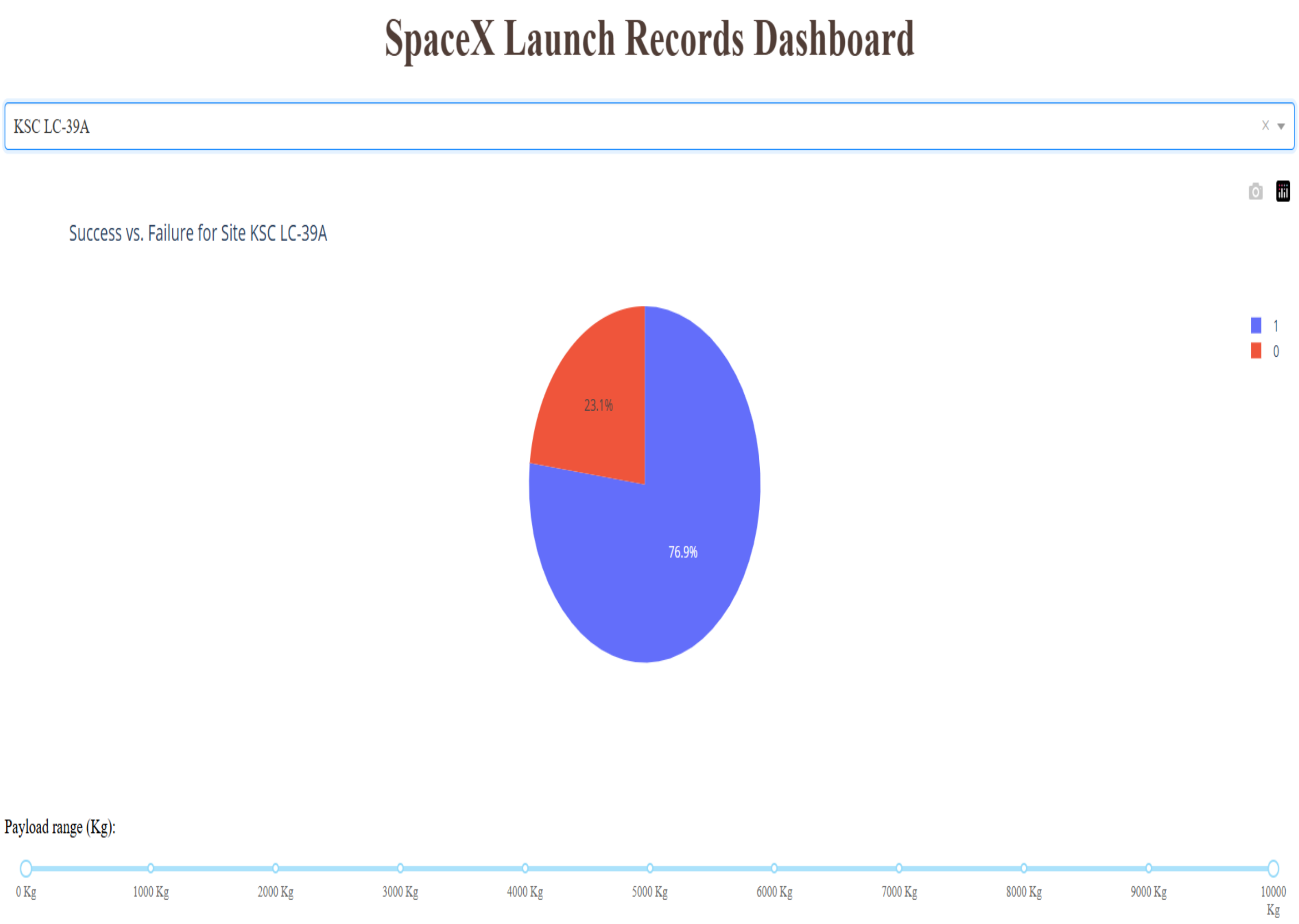
Launch Success Count for All Sites

This pie chart illustrates the distribution of successful launches across all sites. KSC LC-39A leads with 41.7% of total successes, followed by CCAFS LC-40 at 29.2%, VAFB SLC-4E at 16.7%, and CCAFS SLC-40 at 12.5%.



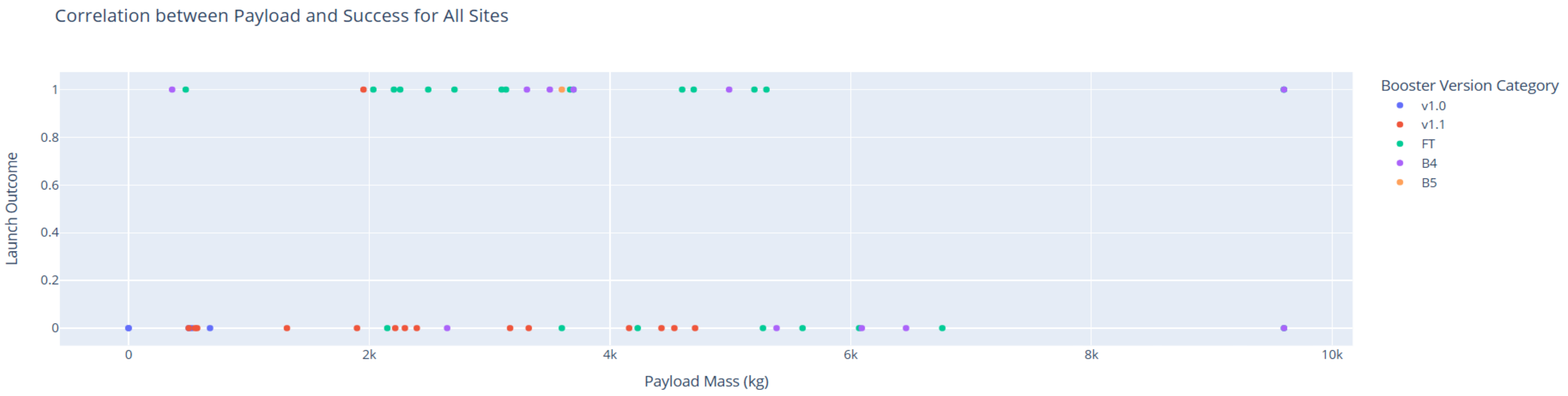
Launch Site With Highest Success Ratio

This pie chart focuses on the launch performance of KSC LC-39A, the site with the highest success ratio. It reveals a success rate of 76.9% and a failure rate of 23.1%, highlighting the site's strong overall performance.



Payload vs. Lauch Outcome for All Sites

This chart explores the relationship between payload mass and launch outcomes across different booster version categories. While there's no clear overall correlation between payload mass and success, performance varies by booster version category. The FT version shows strong results with more successes than failures, B4 has a balanced outcome, while v1.0, v1.1, and B5 show weaker performance.



Payload vs. Launch Outcome (2000kg – 4000kg Range)

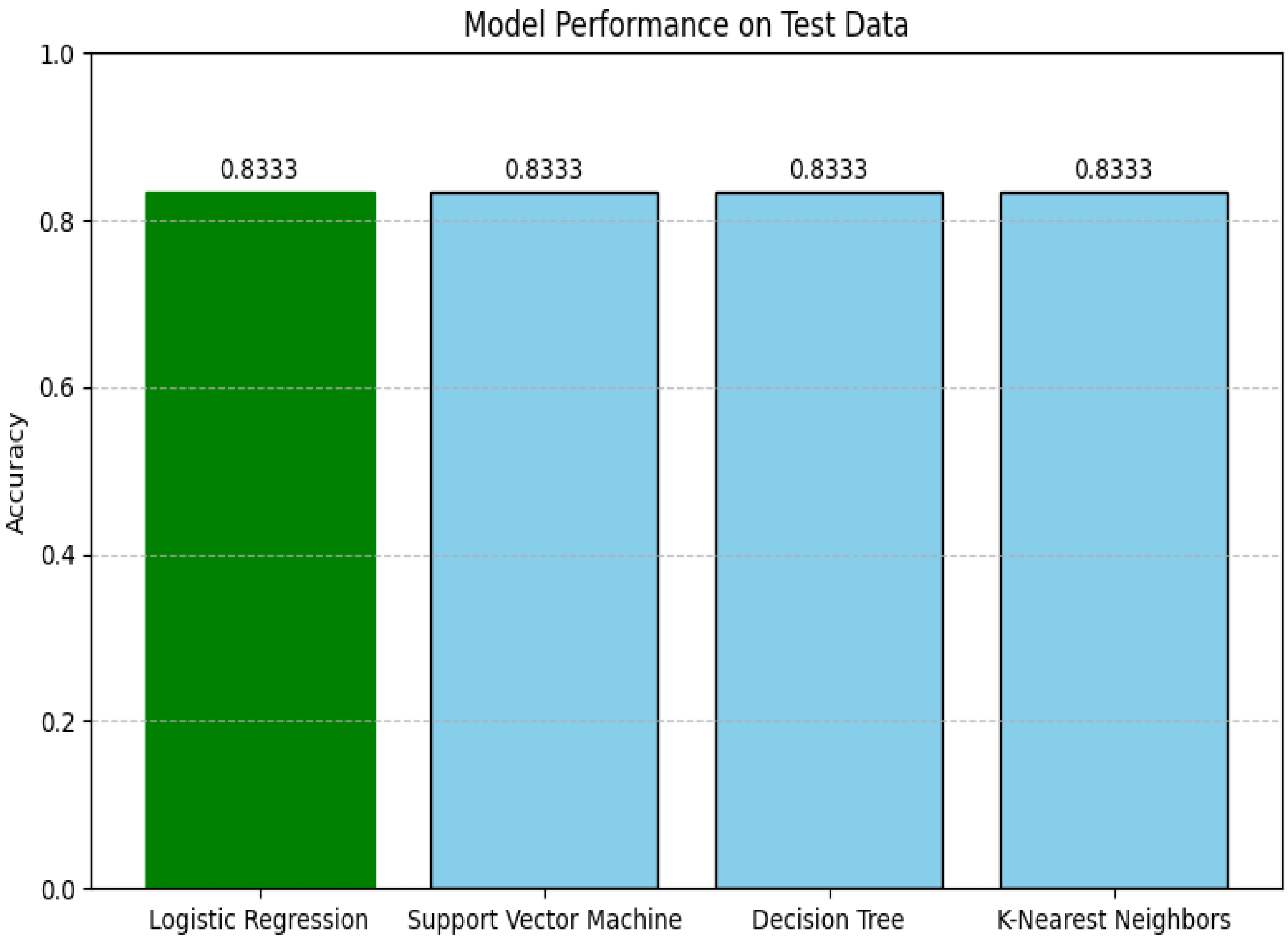
This chart focuses on launch outcomes for payloads between 2000 kg and 4000 kg. It confirms there's no clear correlation between payload mass and launch success. However, the FT booster version category stands out with strong performance, showing only two failures in this range.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

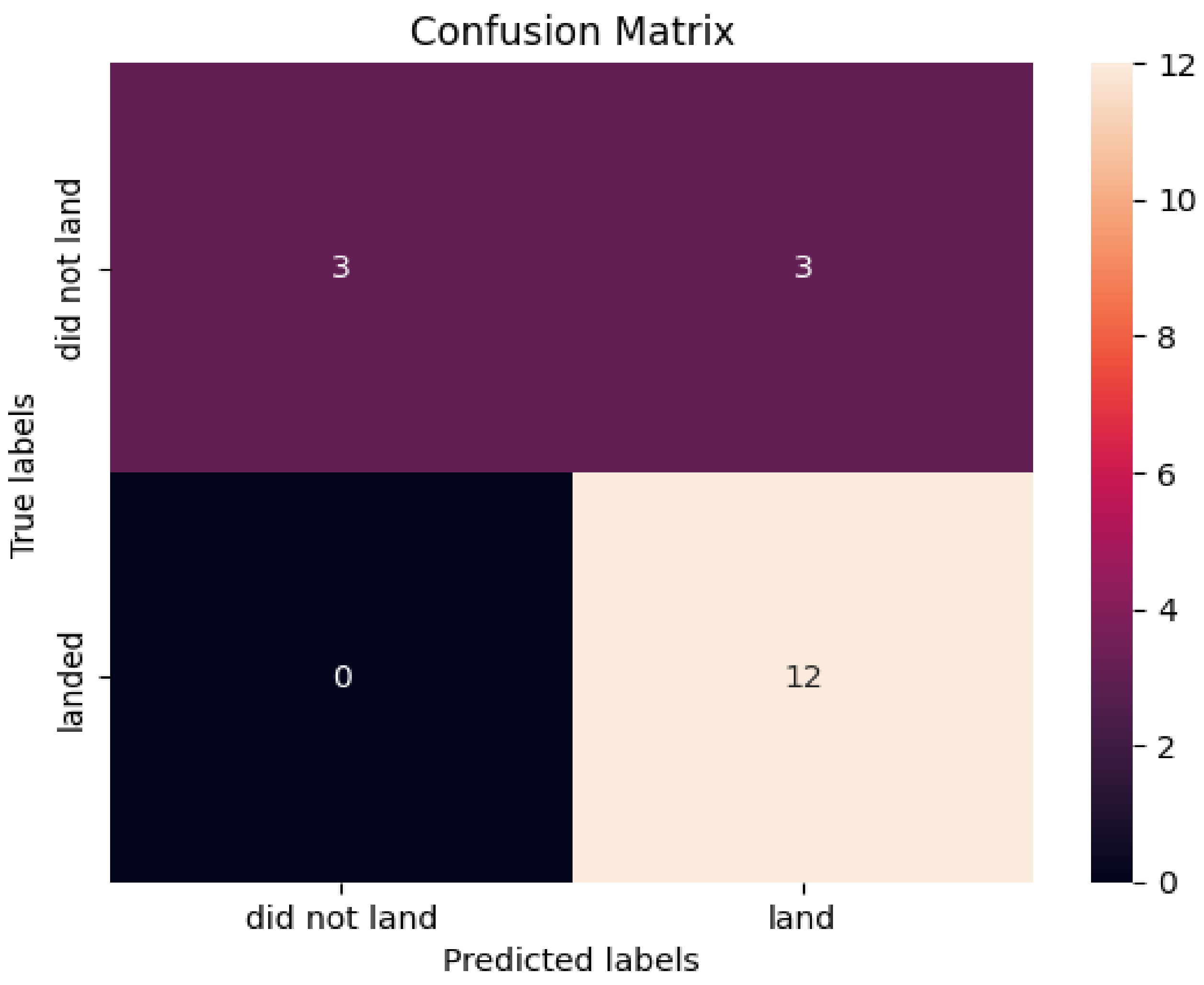
This bar chart reveals a notable finding: all four models—Logistic Regression, SVM, Decision Tree, and KNN—achieve the same test set accuracy of approximately 0.8333. This uniform performance suggests that the dataset's features or structure may inherently support consistent prediction results across different algorithms within the current evaluation setup.



Confusion Matrix

This confusion matrix evaluates a model predicting whether an event "landed" or "did not land." It shows strong performance, with: True Positives: 12, True Negatives: 3, False Positives: 3, False Negatives: 0.

Key Insights: Perfect recall (100%) for “landed” events—the model correctly identified all actual landings. Precision is 80%, indicating some over-prediction of landings (3 false positives). The model is highly effective at detecting actual landings but slightly prone to predicting a landing when it doesn't occur.



Conclusions

This project successfully explored SpaceX Falcon 9 launch data to predict first-stage landing outcomes using machine learning. Four models—Logistic Regression, SVM, Decision Tree, and KNN—were tested, all achieving a consistent accuracy of ~83.33%.

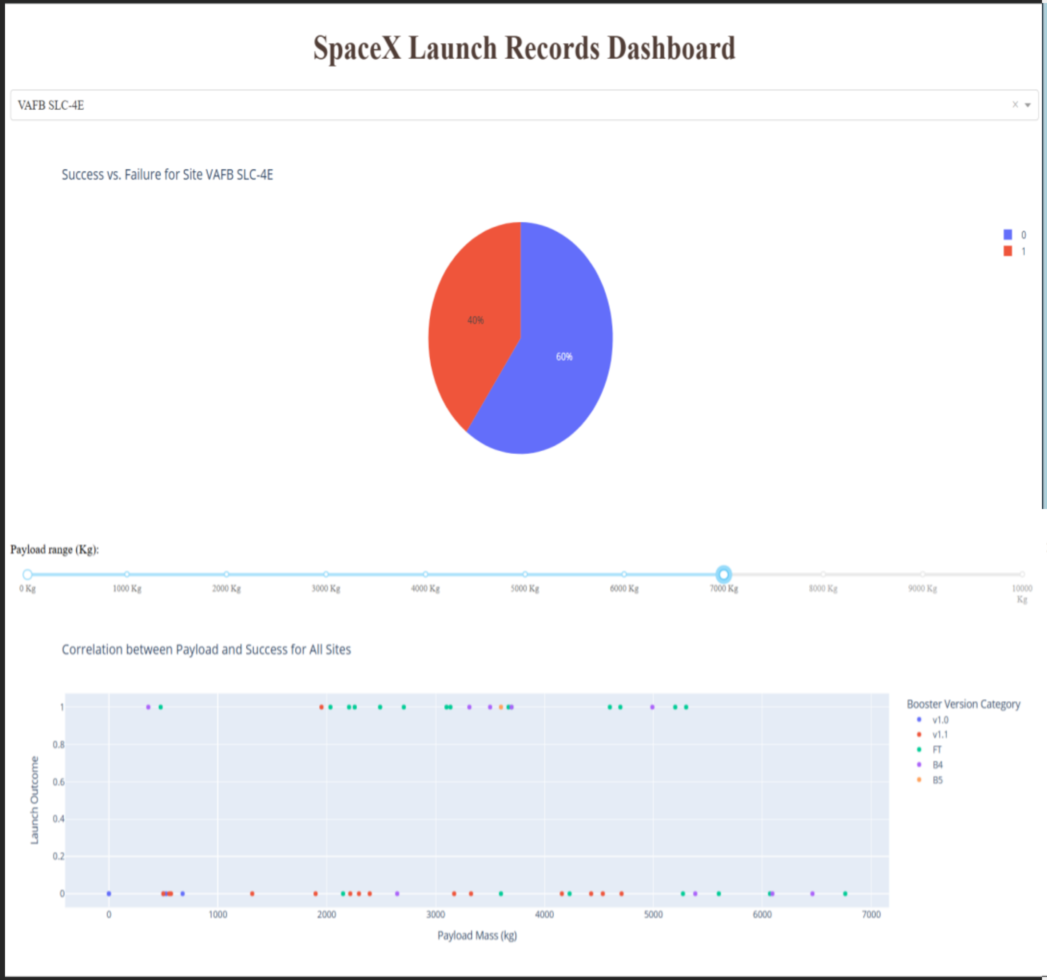
Key Insights:

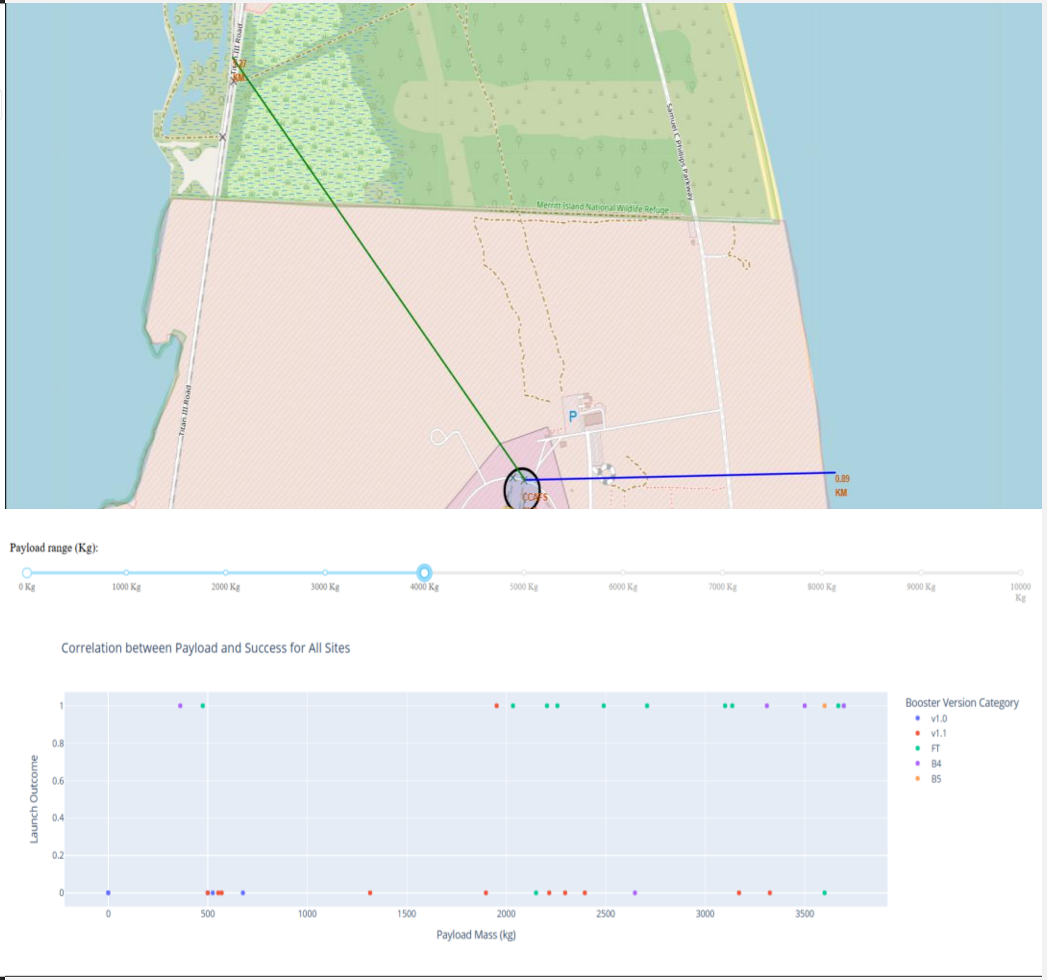
- **Landing success has improved over time**, especially from 2013 to 2019, indicating operational maturity.
- **Booster version category FT** performed best, while v1.0, v1.1, and B5 showed weaker results.
- **Certain launch sites (KSC LC-39A, CCAFS SLC-40)** and orbits (VLEO, GEO, SSO) had higher success rates.
- **Payload mass alone is not a strong predictor**, but success is more likely for heavier payloads launched from certain sites.
- **Perfect recall (100%)** was achieved for predicting landings, though precision was slightly lower due to false positives.

Areas for Further Research:

- Integrate **weather conditions** and **mission types** to enhance prediction accuracy.
- Explore **deep learning models** for improved performance.
- Investigate **economic impact** of failed landings on total mission cost.
- Analyze **temporal factors** (e.g., time of day, launch delays) to uncover hidden patterns.

Appendix (Additional screenshots)





Thank you!