IBM Developer
SKILLS NETWORK

Shanece Grant

June 24, 2025

Outline

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

Executive Summary

I collected and analyzed SpaceX Falcon 9 launch data to uncover key factors influencing first-stage landing success — a major driver of cost savings in modern spaceflight.

Methodologies

- Collected data via API access and web scraping (Wikipedia)
- Cleaned and structured the data for analysis
- Performed exploratory data analysis using SQL and Python
- Built interactive launch maps with Folium
- Trained and tuned ML models (Logistic Regression, SVM, Decision Tree, KNN)

Results

- Key influencing factors: payload mass, orbit type, and launch site
- SVM model achieved the highest predictive accuracy
- Visualizations showed geographic and operational trends in launch success

Introduction

Project Background

SpaceX reduces launch costs by reusing Falcon 9 first-stage boosters. These landings are critical to operational efficiency and competitive advantage. Identifying the factors that contribute to successful landings provides valuable insights for both stakeholders and potential competitors.

Data Collection

Historical Falcon 9 launch data was collected through web scraping (Wikipedia) and API access. The dataset includes launch sites, payload mass, booster versions, orbit types, and mission outcomes.

Problem Statements

- What patterns and variables are most associated with successful Falcon 9 first-stage landings?
- Can we use historical data to build predictive models that estimate landing success?

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Acquired Falcon 9 launch records using web scraping (Wikipedia) and API access.
  - Dataset included launch sites, payloads, orbits, booster versions, and landing outcomes.
- Perform data wrangling
  - Cleaned missing values and standardized column formats.
  - Created binary labels for landing success (1 = success, 0 = failure).
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Queried launch data using SQL to uncover trends by site, payload, and orbit type.
  - Used Python (Pandas, Matplotlib) to visualize correlations and distributions.

Methodology

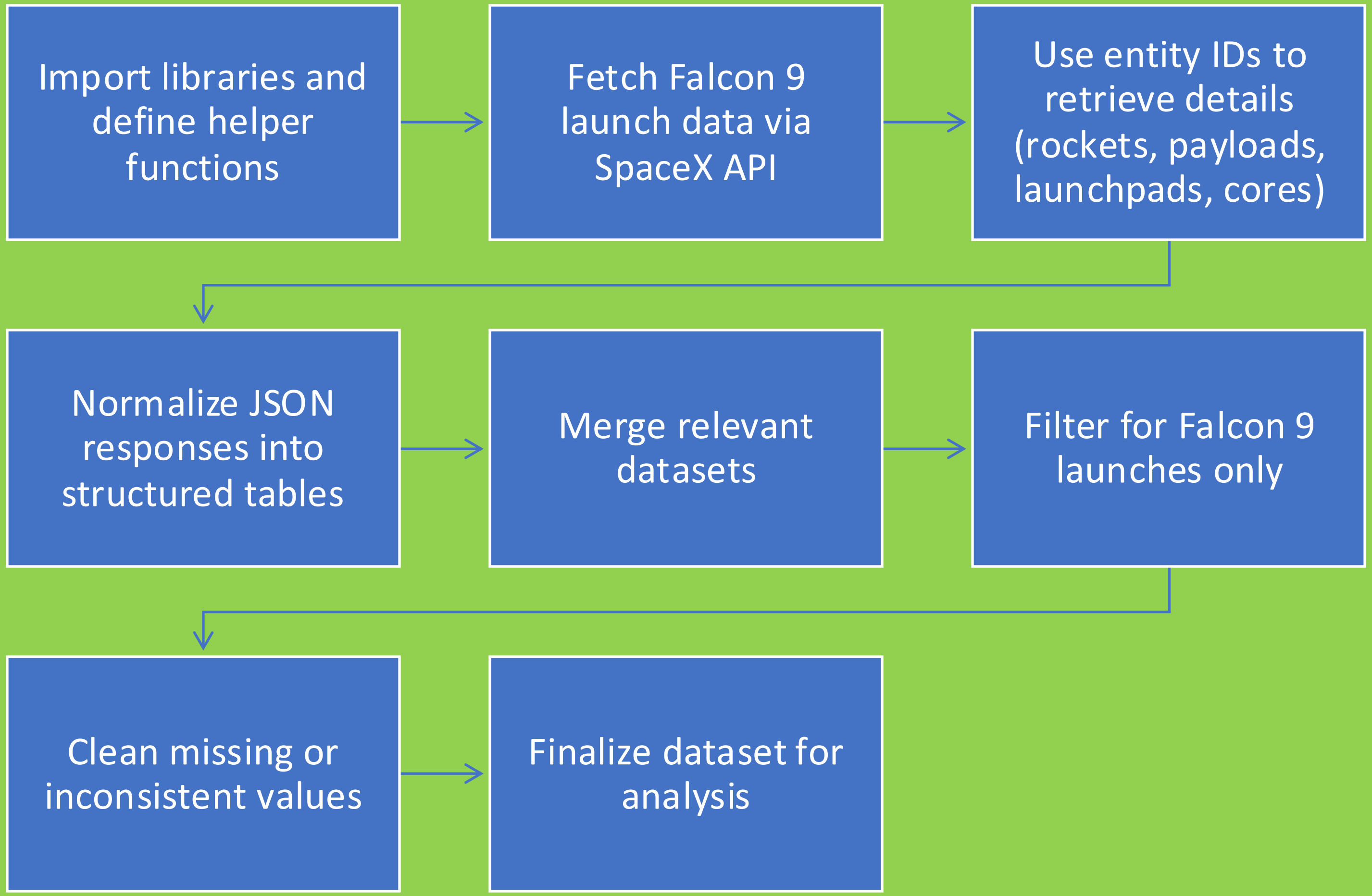
Executive Summary Cont'd

- Perform interactive visual analytics using Folium and Plotly Dash
  - Built interactive maps with Folium to display launch sites and their proximity.
  - Created a Plotly Dash dashboard with dropdowns and filters for dynamic visual exploration.
- Perform predictive analysis using classification models
  - Trained and fine-tuned classification models: Logistic Regression, SVM, Decision Tree, and KNN.
  - Used GridSearchCV for hyperparameter tuning and train_test_split for evaluation.
  - Compared models based on accuracy and validation scores.

Data Collection

Historical Falcon 9 launch data was collected using the SpaceX REST API and web scraping from Wikipedia. I retrieved and normalized details on rockets, payloads, launchpads, and booster cores using entity IDs, filtered for Falcon 9 missions, and prepared a clean dataset for analysis.

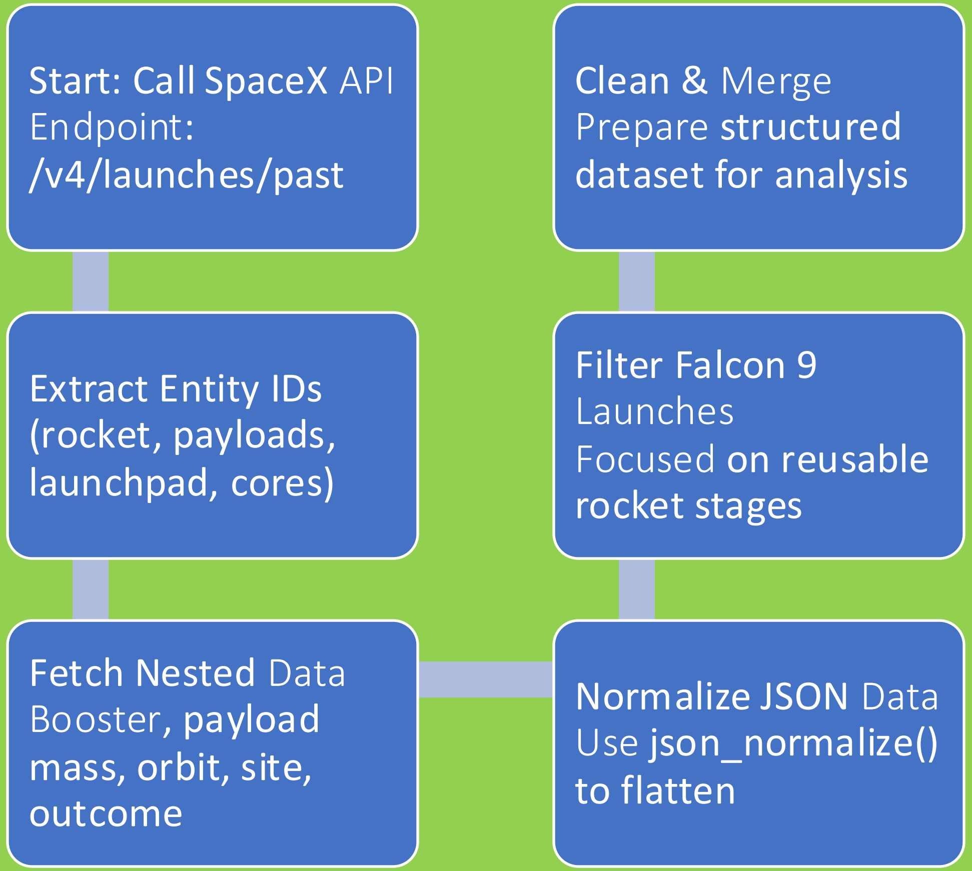
Data Collection (Flowchart)



Data Collection – SpaceX API

SpaceX API Data Collection Summary

- Accessed past launch data from the SpaceX REST API
- Queried launch-specific fields via entity IDs (rocket, payloads, launchpad, cores)
- Parsed JSON and normalized nested data using json_normalize()
- Retrieved key details: booster version, payload mass, orbit, launch site, and landing outcome
- Filtered for Falcon 9 launches only
- Cleaned and merged data into structured format for analysis

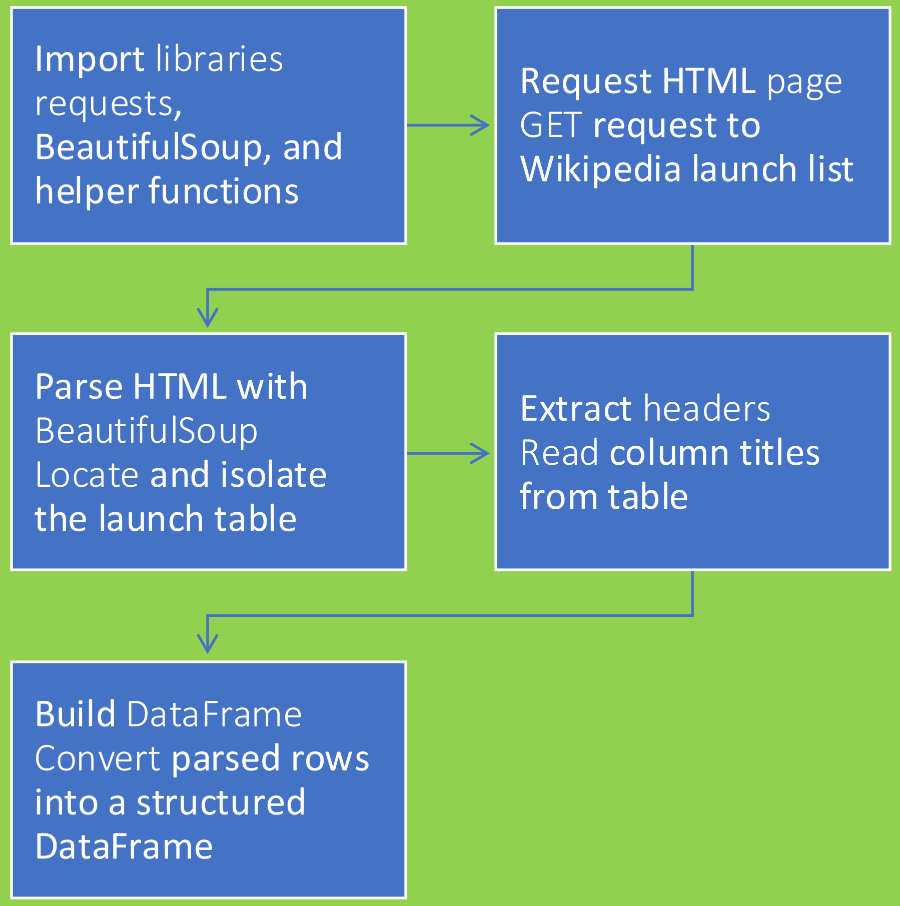


GitHub Notebook: SpaceX API Calls Notebook

Data Collection - Scraping

🔍 Web Scraping Overview

Falcon 9 launch records were scraped from Wikipedia’s List of Falcon 9 and Falcon Heavy launches. Using Python’s requests and BeautifulSoup libraries, we extracted launch tables from the page, parsed HTML content, and converted it into a structured DataFrame for further analysis.



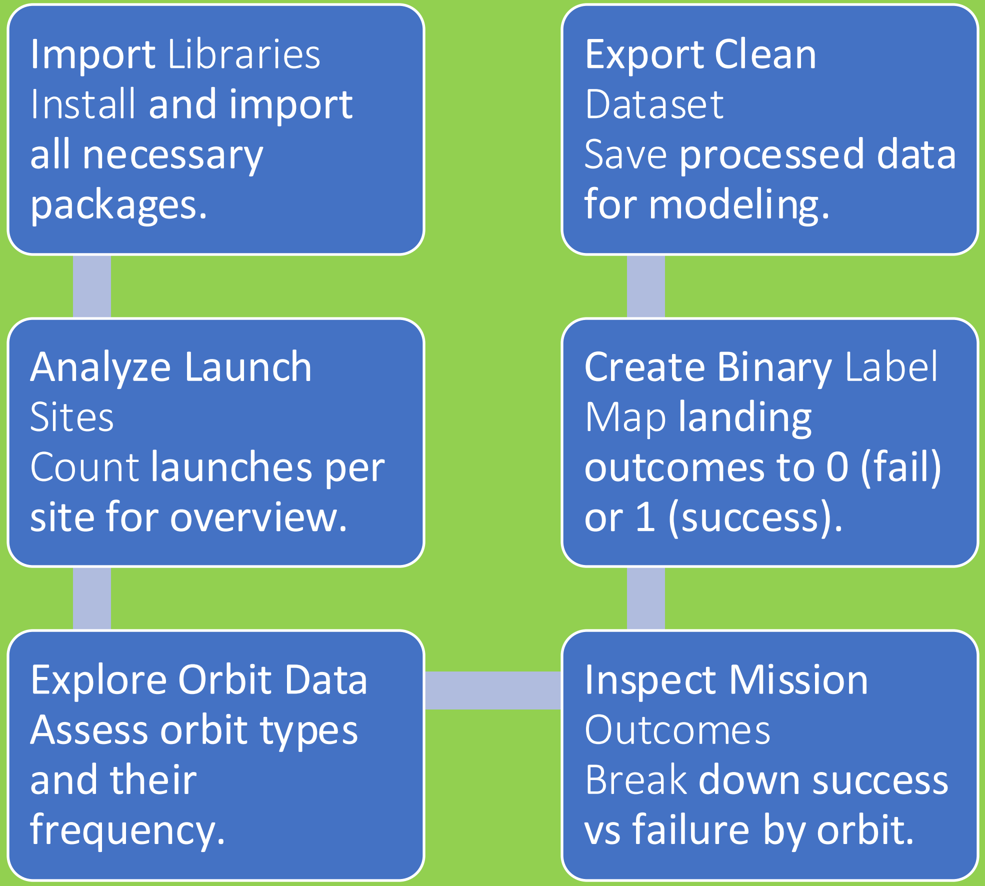
📎 GitHub Notebook: SpaceX Wikipedia Web Scraping Notebook

Data Wrangling

🔍 Data Wrangling Overview

To prepare the SpaceX launch data for machine learning, I performed exploratory analysis and transformed raw outcome labels into a binary classification target. Outcomes like True ASDS, False RTLS, etc., were recoded to indicate landing success (1) or failure (0). I also analyzed launch sites, orbits, and mission results to uncover patterns and cleaned the data for modeling.

📎 GitHub Notebook: Data Wrangling Notebook



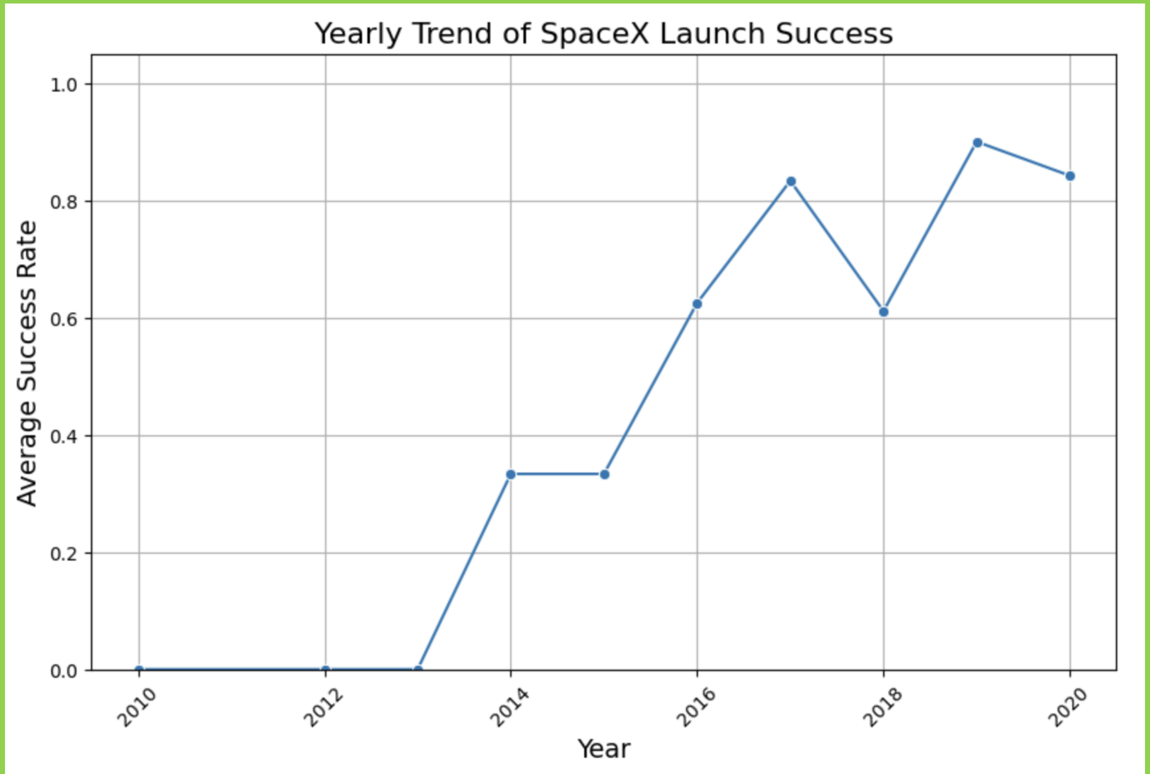
EDA with Data Visualization

🔍 **Purpose of EDA Visuals:**

- Flight Number vs. Payload Mass (Scatter Plot)
  → Shows effect of launch experience and payload on success
- Launch Site Comparisons (Scatter Plots)
  → Examines how site location impacts outcomes
- Success Rate by Orbit (Bar Chart)
  → Highlights orbits with better landing success
- Flight Number & Payload vs. Orbit (Scatter Plots)
  → Detects patterns across orbit types
- Launch Success Over Time (Line Chart)
  → Reveals improvement trends
- Feature Encoding + Data Type Conversion
  → Prepared data for modeling (OneHot, float64)

📎 GitHub Notebook: EDA with Visualization Notebook

Line chart showing SpaceX launch success rate increasing from 2013-2020



EDA with SQL

🔍 SQL-Based Exploratory Data Analysis

- Cleaned table and removed empty rows
- Retrieved unique launch sites
- Filtered for sites starting with 'CCA'
- Calculated total payload launched by NASA (CRS): 45,596 kg
- Computed average payload for F9 v1.1: 2,928.4 kg
- Identified date of first successful ground pad landing
- Listed boosters with drone ship landings & payloads between 4000–6000 kg
- Counted successful vs failed mission outcomes
- Found boosters with max payload via subquery
- Ranked landing outcomes between 2010–2017 by frequency

📎 GitHub Notebook: Exploratory Data Analysis with SQL Notebook

SpaceX landing outcomes between (2010-06-04 to 2017-03-20) in descending order

Landing_Outcome	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

Build an Interactive Map with Folium

Map Objects Created:

- 📍 Launch Site Markers – Plotted all Falcon 9 launch sites using folium.Marker with labels.
- ✅ ❌ Outcome Markers – Colored green for successful landings and red for failed ones, based on the class value.
- 🔵 Circles – Used folium.Circle to highlight launch areas and proximity zones.
- 📏 Polylines – Added folium.PolyLine to draw distances from launch sites to nearby coastlines, highways, and railroads.
- 📌 Text Labels – Applied DivIcon to label key proximity features on the map.

Purpose of Map Objects:

- Help visualize spatial relationships between launch outcomes and geographic features.
- Support analysis of potential environmental or logistical factors (e.g., distance to infrastructure) influencing launch success.
- Enhance interactivity for clearer exploration of trends and site characteristics.

🔗 GitHub Notebook: Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

Interactive Components & Plots:

- 📍 Launch Site Dropdown – Allows users to filter data by individual launch site or view all sites combined.
- 🎯 Success Pie Chart – Updates dynamically to show success vs failure rates based on selected site.
- 📦 Payload Range Slider – Filters launches by payload mass for detailed trend analysis.
- ✨ Scatter Plot – Displays correlation between payload mass and launch success; highlights patterns by booster version and site.

Purpose of Interactions:

- Enables real-time, user-driven analysis of SpaceX launch outcomes.
- Helps visualize how different factors (like payload and site) affect success rates.
- Provides intuitive, visual insights to explore key patterns in the data.

🔗 GitHub Notebook: Plotly Dash App Notebook

Predictive Analysis (Classification)

Model Development Process:

- 🧪 Created training labels from the Class column
- ⚖️ Standardized feature data
- 🔀 Split data into training (80%) and test (20%) sets
- 🔍 Tuned hyperparameters with GridSearchCV (cv=10) for:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

- ✅ Evaluated test accuracy to select the best-performing model

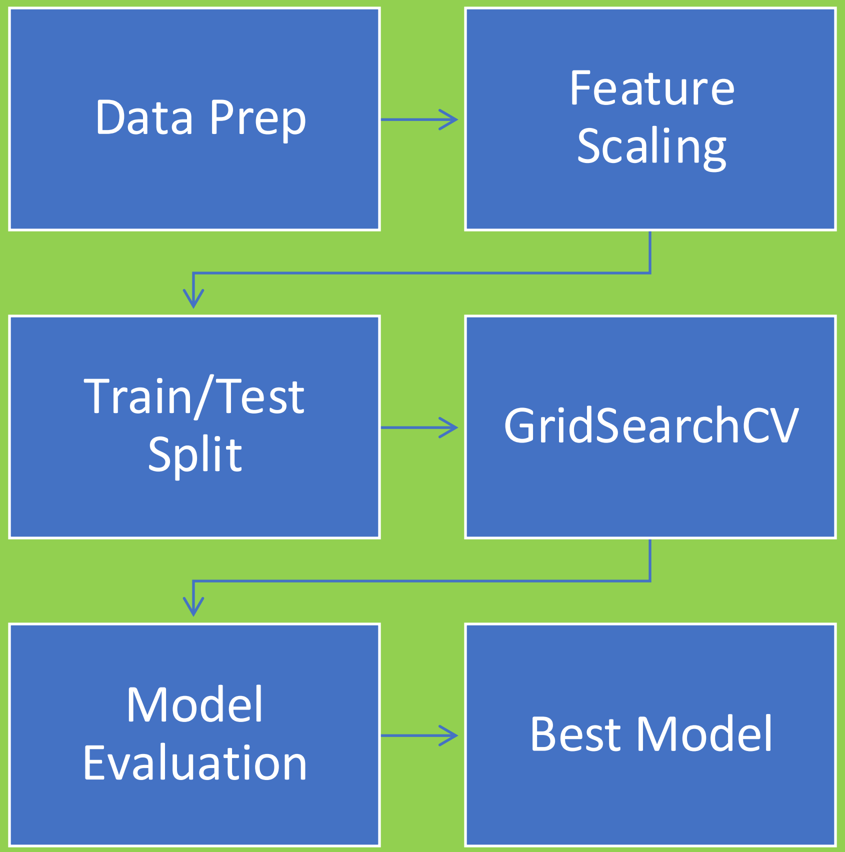
🏆 Best Model:

Logistic Regression

Test Accuracy: 83.3%

🔗 GitHub Notebook: Predictive Analysis Notebook

🔄 Development Process Flowchart



Exploratory Data Analytics Results

🧪 **EDA Highlights**

- Success rates improved with higher flight numbers
- Lower payload mass associated with higher landing success
- Orbit type and launch site were strong predictors
- Success rate trends increased year over year

Interactive Analytics Results

<u>HŌPÑǾMŃPÒQÑ ÌMŐŒ</u>

- Visualized launch outcomes by location
- Showed proximity to railroads, highways, and coastlines
- Marker clusters enabled pattern discovery at each site

Predictive Analysis Results

🤖 Predictive Model

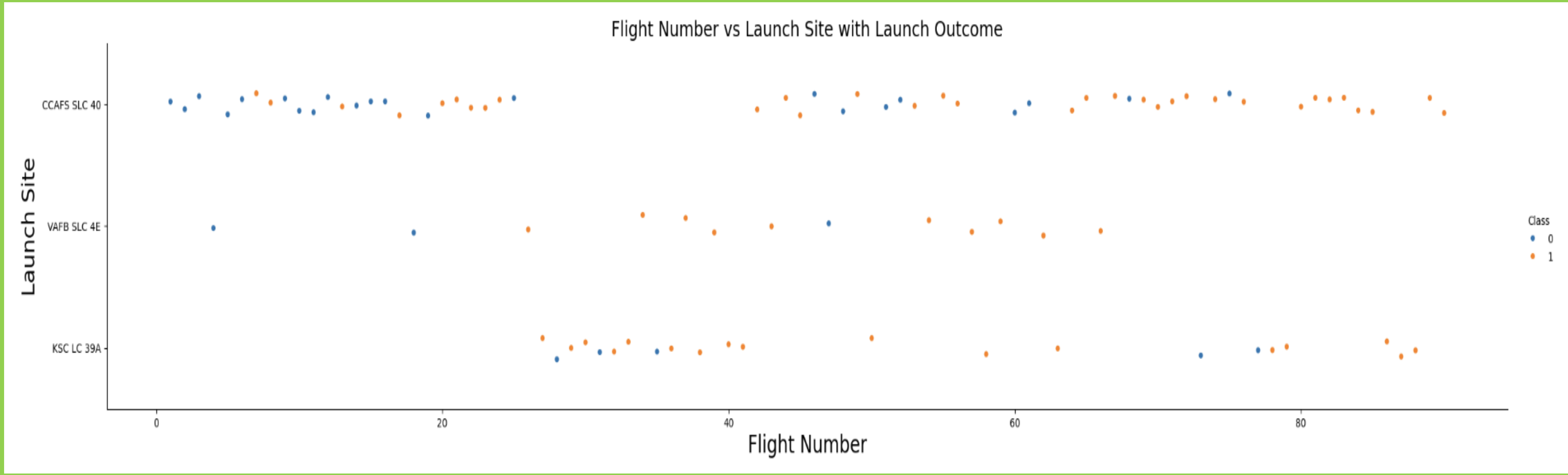
- Best model: Logistic Regression
- Accuracy: 83.3% on test data
- Key predictive features: payload mass, orbit, launch site

Exploratory Data Analysis (EDA) Insights

Flight Number vs. Launch Site

🔍 Key Insight:

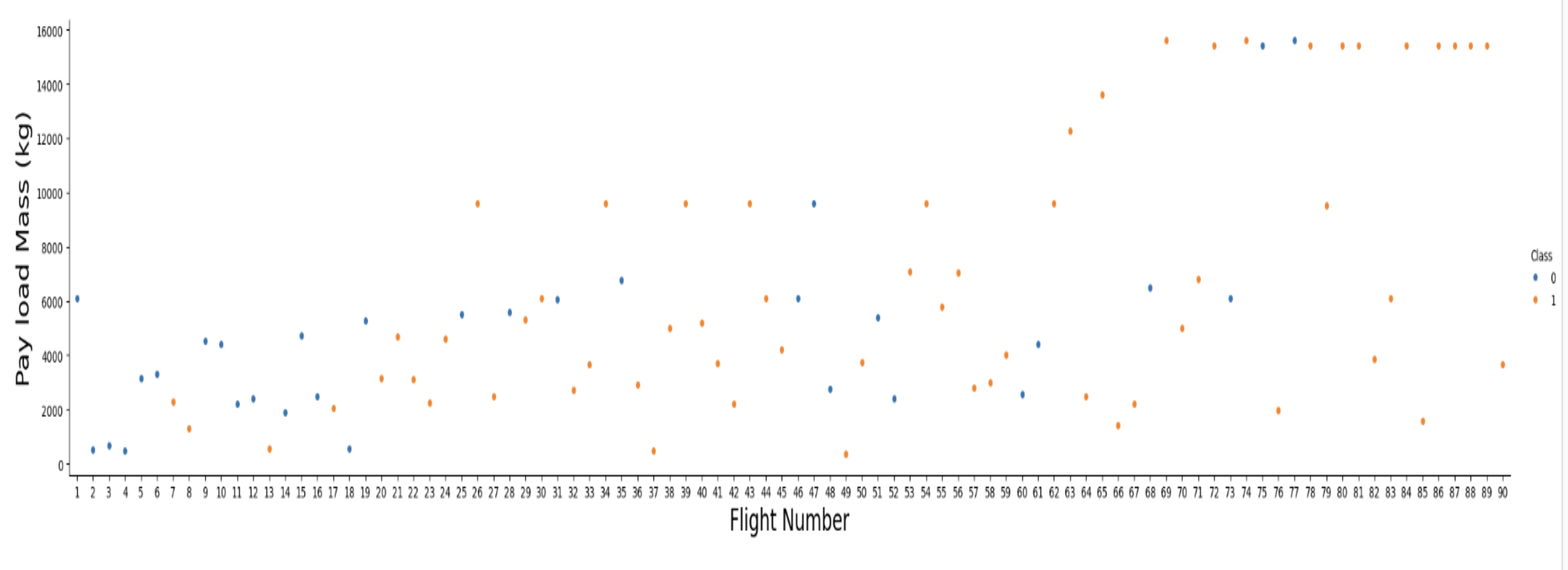
Newer launch sites and earlier flights had more failures, suggesting SpaceX improved launch success over time through iteration and experience.



Payload vs. Launch Site

🔍 **Key Insight:**

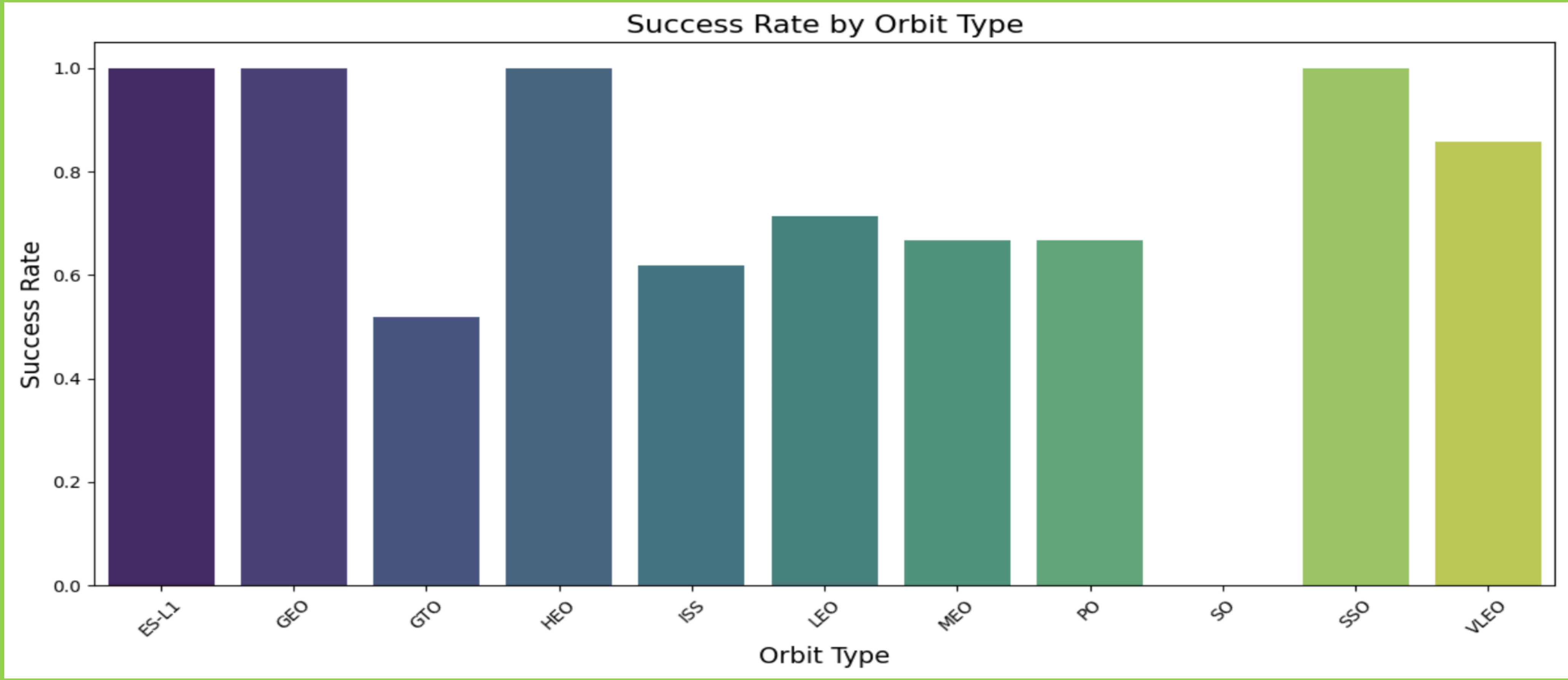
Heavier payloads from certain launch sites experienced more failures. This may indicate launch site capability differences or limitations tied to payload weight.



Success Rate vs. Orbit Type

🔍 **Key Insight:**

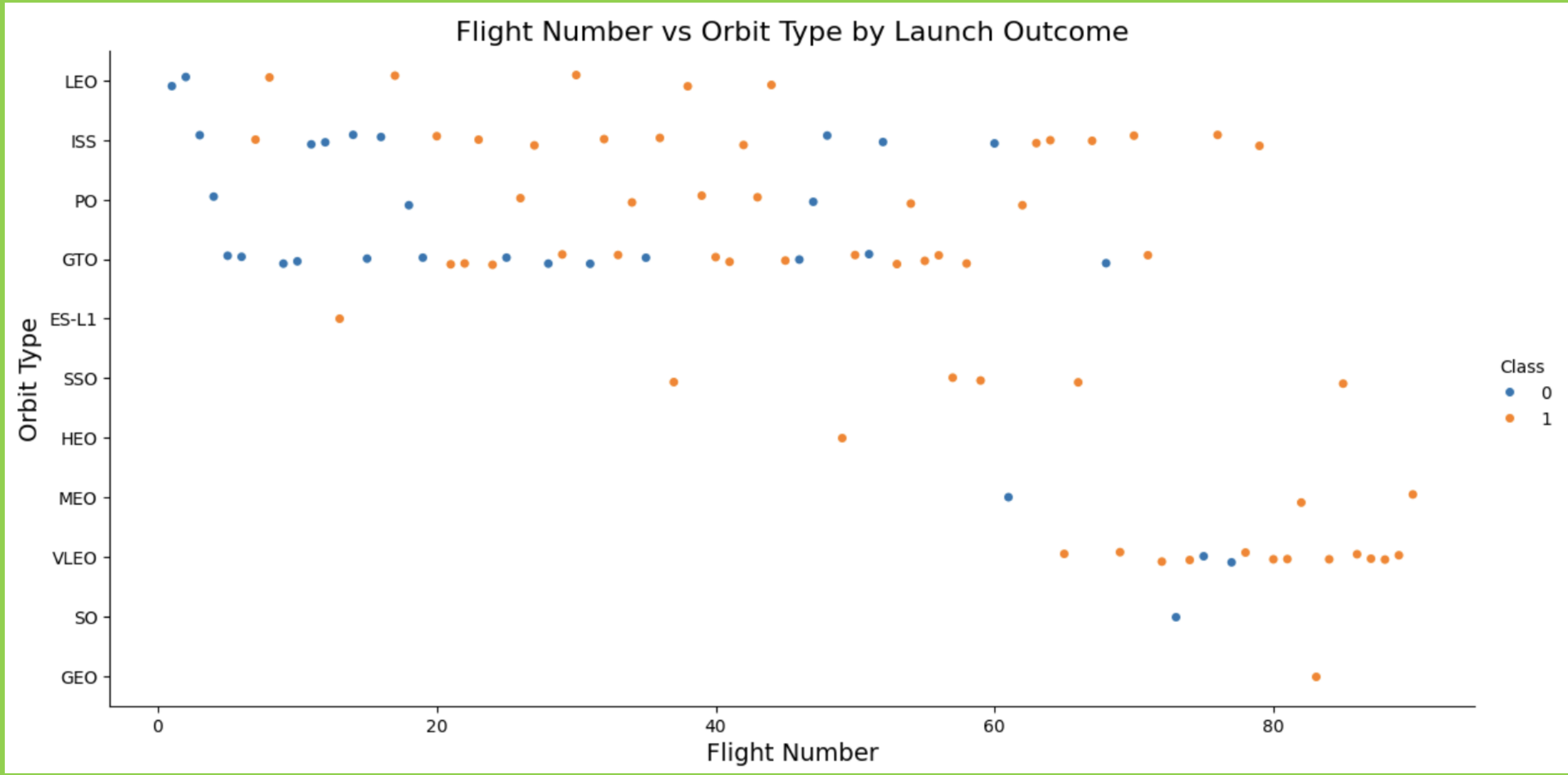
Success rates varied by orbit type, with certain orbits showing higher reliability. This may reflect mission complexity or trajectory demands.



Flight Number vs. Orbit Type

🔍 Key Insight:

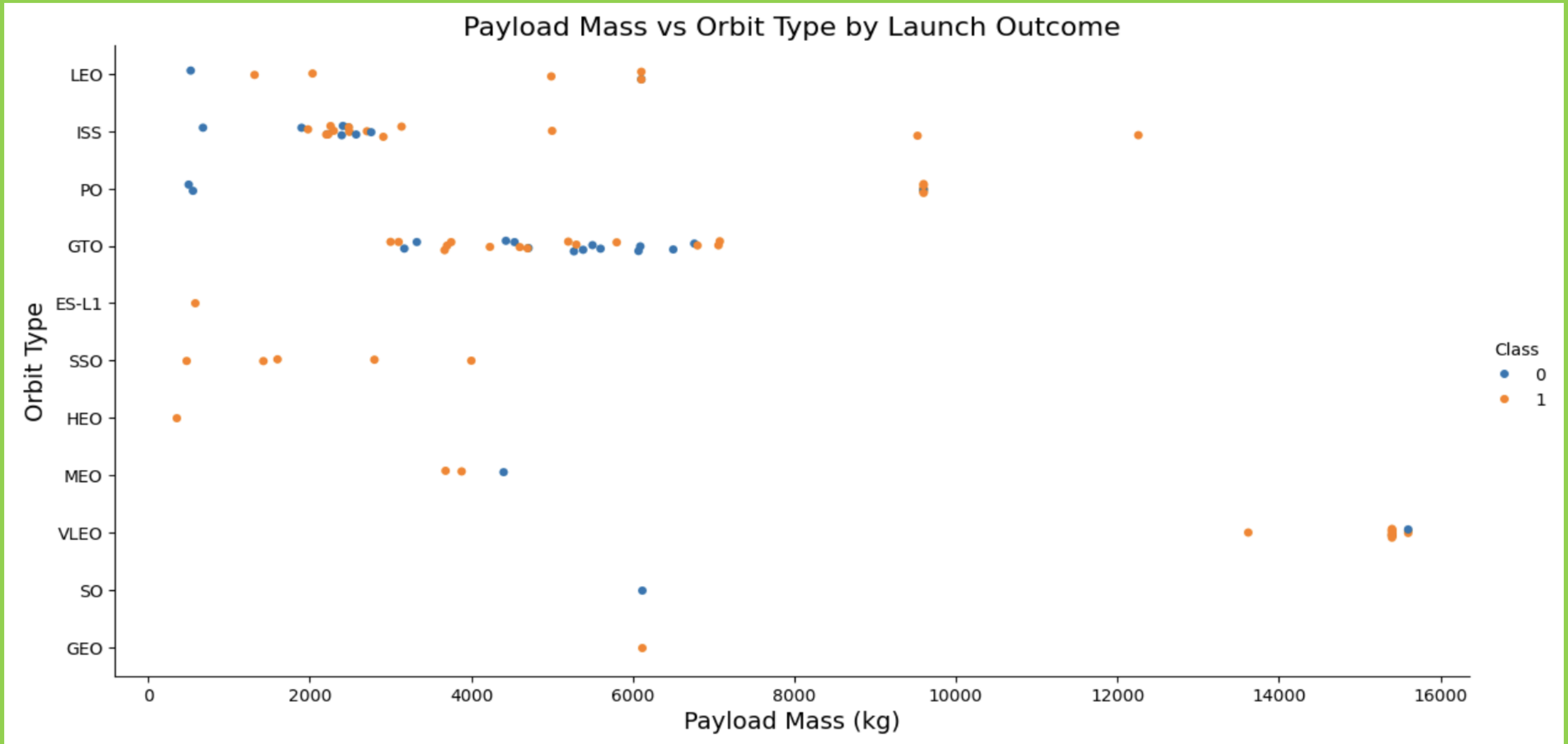
Early missions favored specific orbit types (e.g., LEO), while later missions diversified. Flight number correlates with orbit evolution and mission maturity.



Payload vs. Orbit Type

🔍 **Key Insight:**

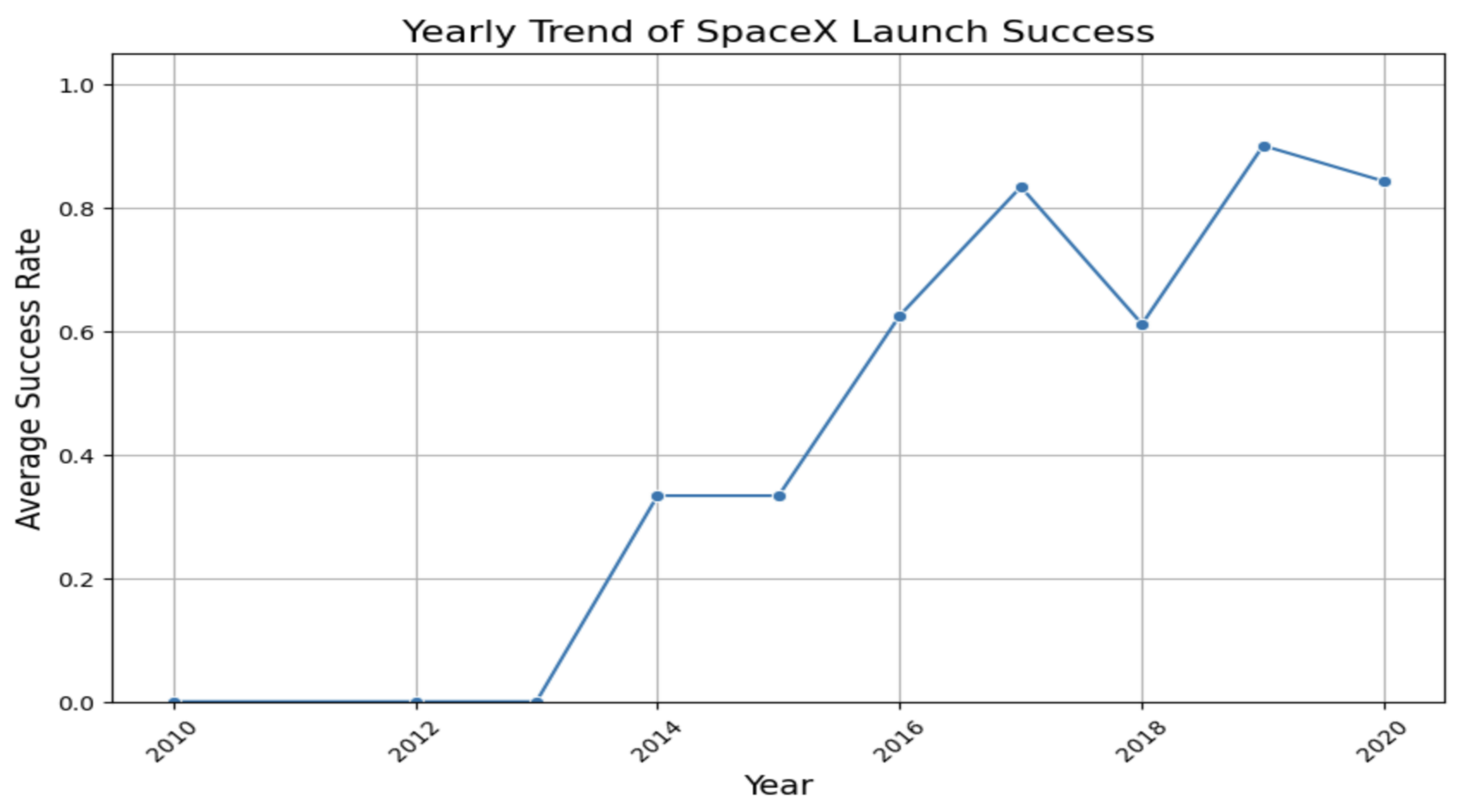
Payload mass clusters differ across orbit types. Heavier payloads are typically destined for higher orbits like GTO, which also show a varied success rate.



Launch Success Yearly Trend

🔍 **Key Insight:**

Success rates steadily improved year over year, demonstrating SpaceX's technological refinement and operational growth.



All Launch Site Names

🔍 Explanation:

The SQL query retrieved all unique launch site names from the SpaceX dataset.

Identifying each distinct site helps establish how SpaceX distributes missions across locations, which may influence mission success and logistics.

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

Launch Site Names Begin with 'CCA'

🔍 **Explanation:**

The query filters the dataset to only include records where the launch site name begins with "CCA", referring to Cape Canaveral launch facilities. It provides insight into SpaceX’s activity at Cape Canaveral and can help analyze how performance may vary by site.

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

🔍 Explanation:

The query calculates the total payload mass delivered by SpaceX boosters for NASA's Commercial Resupply Services (CRS) missions.

The result—45,596 kilograms—highlights the significant cargo volume launched in partnership with NASA, underscoring the operational importance of these missions in the dataset.

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

🔍 Explanation

This query determines the average payload mass carried by the Falcon 9 v1.1 booster version.

The result—2,928.4 kilograms—provides insight into the typical cargo capacity during this era of launches.

It helps establish performance baselines for comparison with newer booster versions.

Avg_Payload_Mass
2928.4

First Successful Ground Landing Date

🔍 **Explanation**

This query identifies the earliest date SpaceX achieved a successful landing on a ground pad.

The milestone on December 22, 2015, marked a major achievement in rocket reusability and significantly contributed to reducing launch costs.

First_Ground_Pad_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

🔍 **Explanation:**

The query filters for booster versions that successfully landed on a drone ship and had a payload mass between 4000 kg and 6000 kg.

This helps identify boosters capable of handling mid-range payloads while achieving successful drone ship landings — important for assessing performance under specific conditions.

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

Total Number of Successful and Failure Mission Outcomes

🔍 **Explanation:**

This query counts the number of times each unique mission outcome appears in the dataset. It helps us quickly see how many launches were successful versus how many failed (or had other outcomes). This kind of summary is useful for understanding SpaceX's overall launch performance and reliability over time.

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

Boosters Carried Maximum Payload

🔍 **Explanation:**

This query identifies the booster(s) that carried the maximum payload mass recorded in the dataset.

It helps highlight which booster versions handled the heaviest missions, providing insight into SpaceX's heavy-lift capabilities.

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

🔍 **Explanation:**

The query lists all 2015 launch records where the landing outcome was a failure on the drone ship. It includes the month (converted from date), landing outcome, booster version, and launch site to provide insight into specific conditions and equipment during unsuccessful drone ship landings in 2015.

Month	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

🔍 **Explanation:**

The query ranks all landing outcomes between June 4, 2010, and March 20, 2017, by the number of occurrences, from most to least frequent.

It helps identify which landing results were most common during this period, providing insights into SpaceX's landing performance trends.

Landing_Outcome	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

Launch Site Proximities Analysis

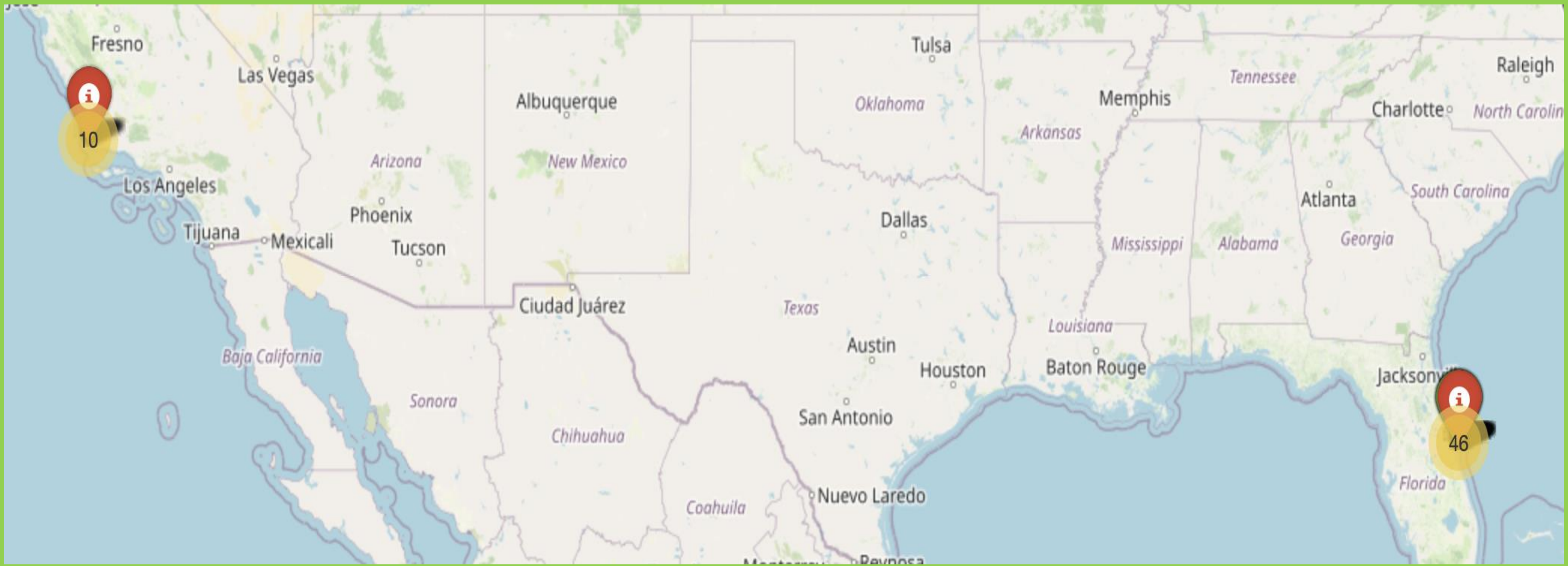
Folium Map: Launch site location markers on a global map

All recorded SpaceX launch site locations using Folium markers. Each marker includes the site name as a popup and is placed according to its real-world coordinates.

Important observations:

- All launch sites are located within the United States.
- The primary sites include KSC LC-39A, VAFB SLC-4E, CCAFS LC-40, and CCAFS SLC-41.
- The map helps visually confirm launch distribution and geographic clustering on the U.S. east and west coasts.

The map was created using Folium and Python, combining real-world coordinate data with interactive visualization features.



Folium Map: Cape Canaveral Launch Outcomes – Success vs. Failure

This map visualizes launch outcomes from Cape Canaveral using color-coded markers. Each marker represents a single launch, where:

🟢 Green = Successful ; 🔴 Red = Failed

By focusing on Cape Canaveral, we can observe:

- A higher concentration of successful launches over time.
- Some failures still occurred, helping to identify areas for operational improvement.
- The map reflects both the evolution of landing technology and the learning curve over repeated launches.

This visualization helps evaluate performance by location, supporting deeper spatial analysis of outcomes.

Folium Map: KSC LC-39A Distance to Railroad

This interactive map highlights the proximity between Kennedy Space Center Launch Complex 39A (KSC LC-39A) and the nearest railroad.

- The blue marker represents the launch site, while the other marker indicates the nearest railroad.
- A straight polyline connects the two locations, and the calculated distance (0.70 KM) is displayed on the map.

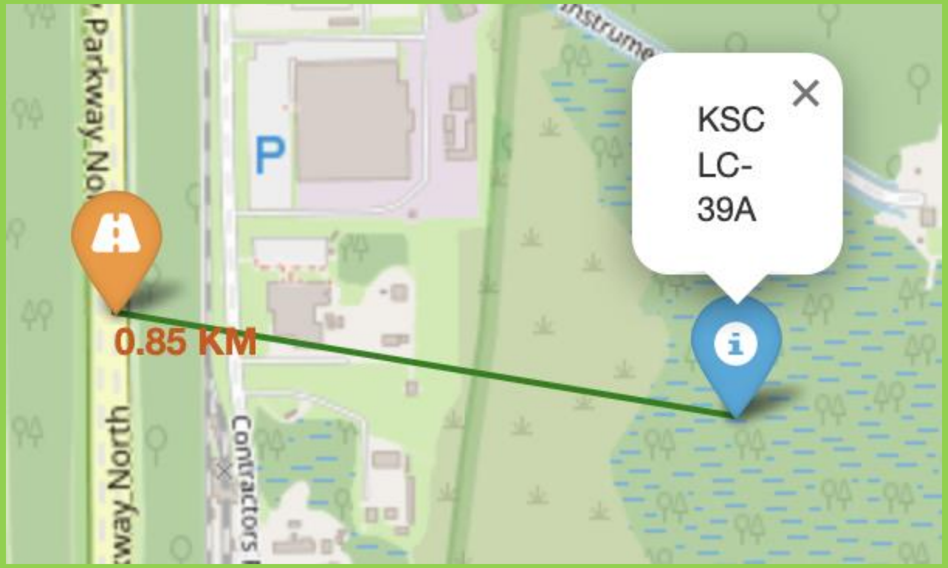
This information is crucial in analyzing logistics support for launch operations, especially for the transport of heavy equipment or boosters.

The proximity to a railroad enhances supply chain efficiency and may factor into site selection for future missions.

Folium Map: KSC LC-39A Distance to Highway

This map visualizes the distance between Kennedy Space Center Launch Complex 39A (KSC LC-39A) and the nearest highway.

- The blue marker represents the launch site, and the orange marker indicates the nearest highway access point.
- A straightpolyline connects the two, and the distance is calculated and displayed on the map (0.85KM).
- This visualization helps assess how easily heavy launch equipment and personnel can reach the launch complex via road.

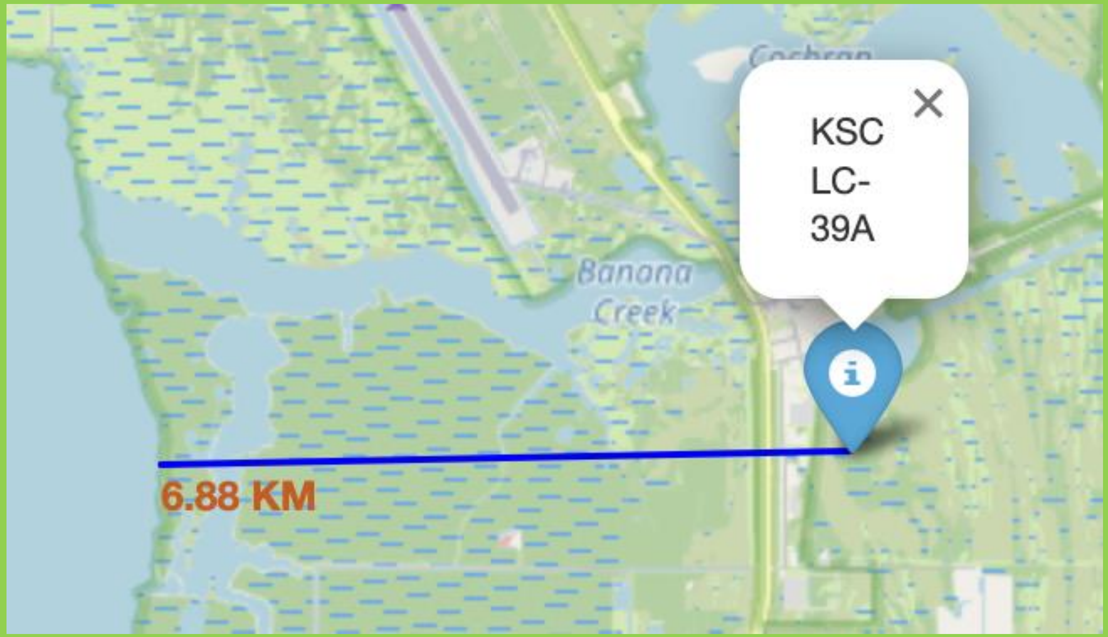


The relatively short distance supports efficient transportation logistics and contributes to the site's suitability for frequent launch operations.

Folium Map: KSC LC-39A Distance to Coastline

This interactive map shows the distance between Kennedy Space Center Launch Complex 39A and the nearest coastline.

- The launch site is marked in blue, and the nearest point on the coastline is marked in green.
- A blue polyline visually connects the two points.
- The distance (6.88 KM) is calculated and labeled directly on the map.



Understanding this distance is important for launch safety, particularly in the event of an early flight termination or booster landing in the ocean. It also demonstrates how geographic proximity to the coastline supports safer, more flexible launch trajectories for Falcon 9 missions.

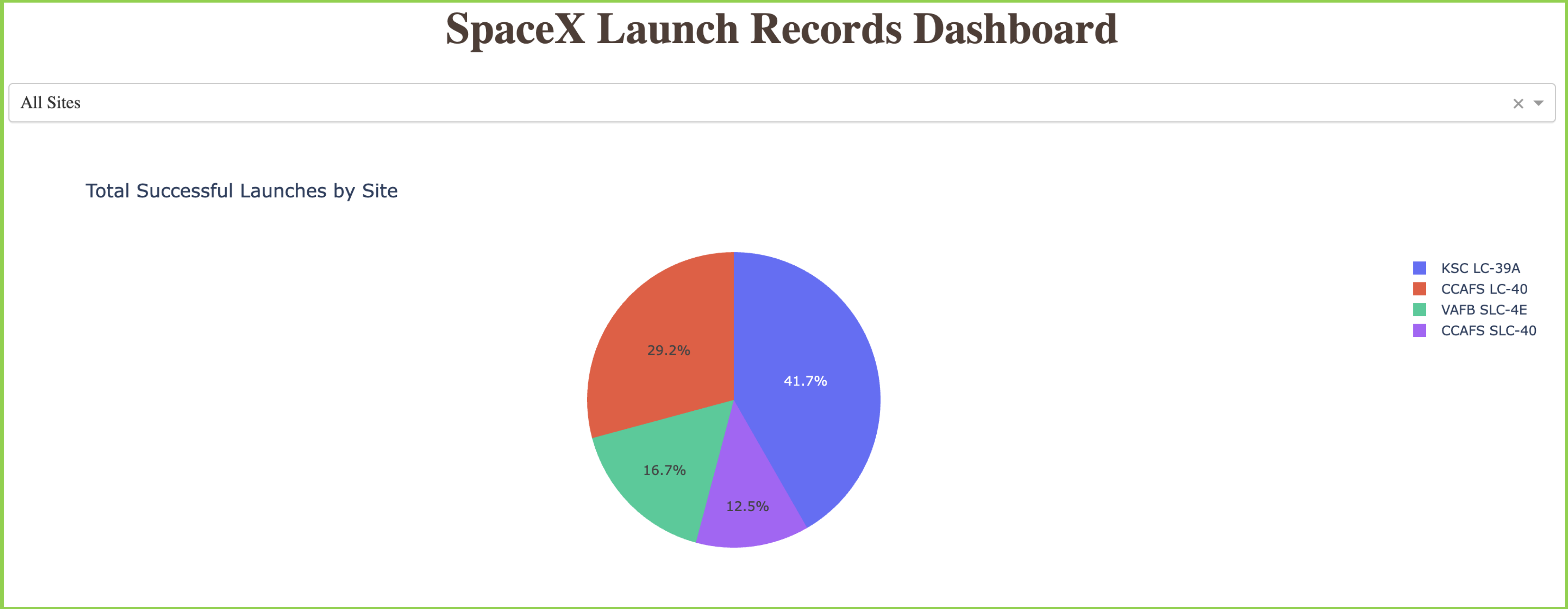
SpaceX Launch Dashboard with Plotly Dash

SpaceX Launch Success For All Sites

This pie chart visualizes the distribution of successful launches across all SpaceX launch sites.

Each slice represents a launch site, and the size of the slice reflects the number of successful launches conducted there.

This helps identify which launch sites SpaceX relies on most for successful missions and may indicate preferred or more reliable locations.

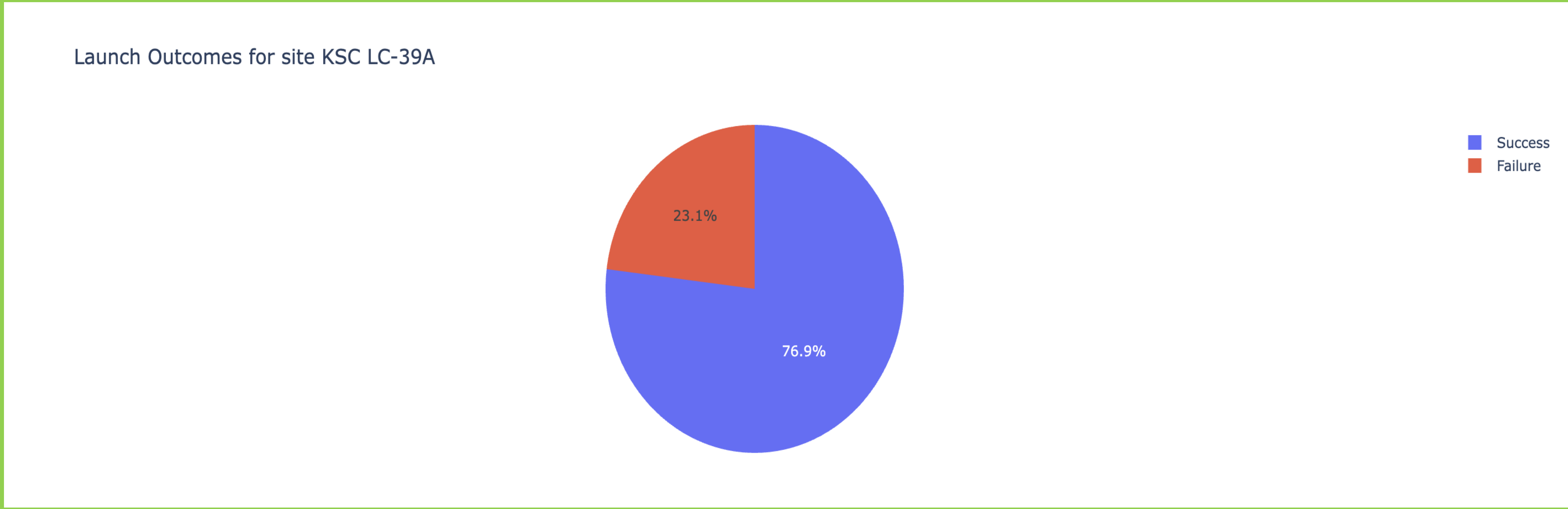


Site with highest launch success ratio: KSC LC-39A

This pie chart shows the launch success ratio for the KSC LC-39A launch site.

- The chart is divided into two segments: Successful and Failed launches.
- Based on the visual, this site has the highest proportion of successful launches compared to its total attempts.

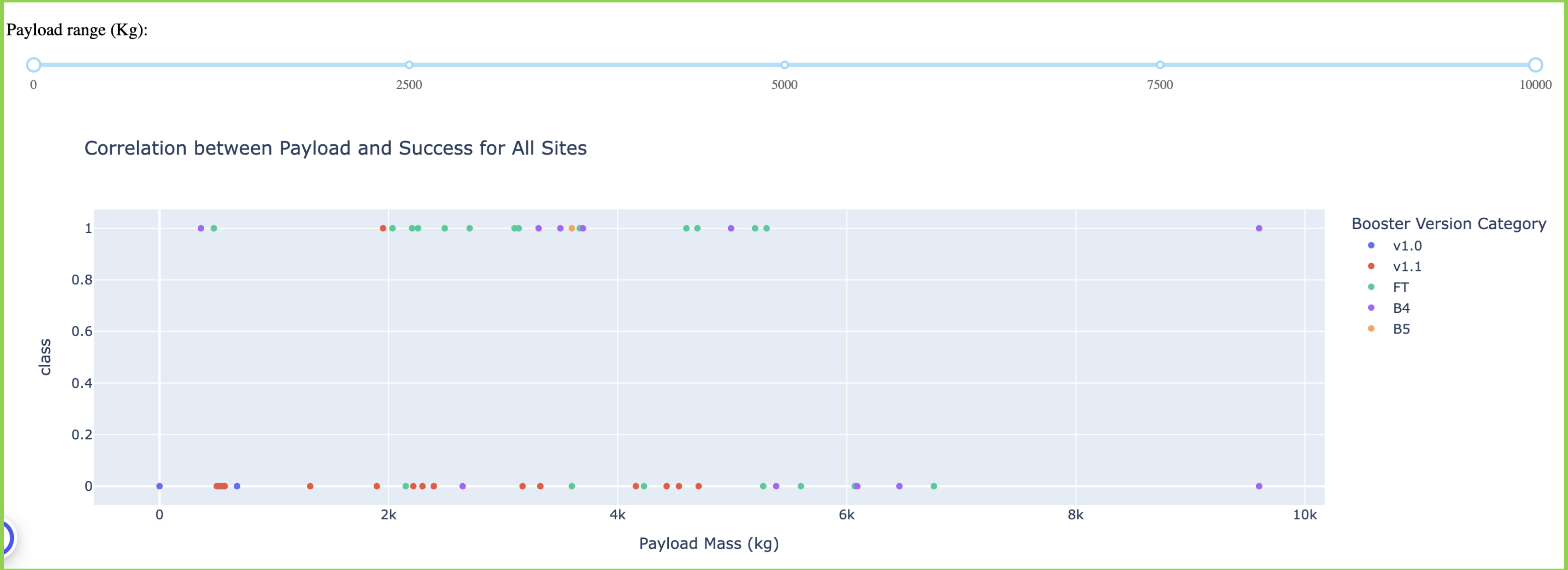
- This insight suggests KSC LC-39A is one of SpaceX's most reliable and efficient launch pads, making it a strong candidate for future missions or for cost-effective launch planning.



Payload vs. Launch Outcome: All Sites (Payload Range: 0–10,000 kg)

Scatter plot showing launch outcomes (1 = success, 0 = failure) vs. payload mass (0–10,000 kg) across all SpaceX sites.

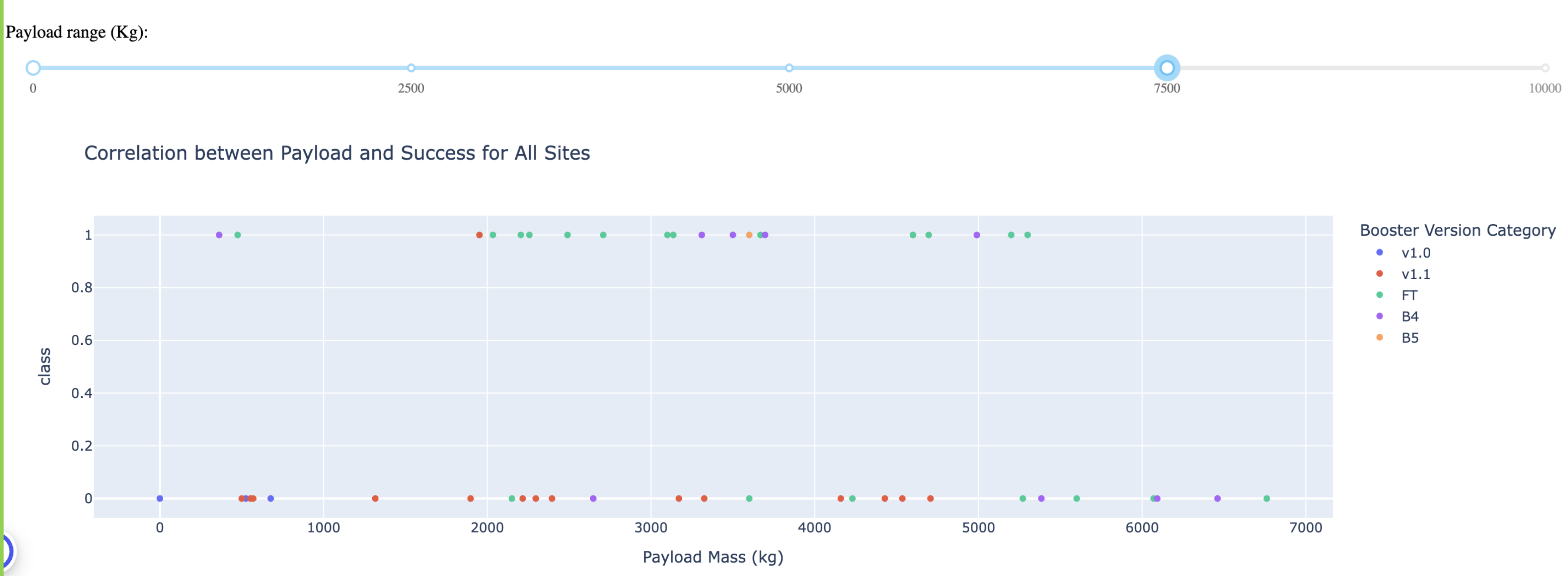
- Older boosters (v1.0, v1.1) show mixed results.
- Falcon 9 FT shows most consistent success, highlighting improved reliability.
- Few successes at extreme payload values.



Launch Outcomes vs. Payload Mass (0–7,500 kg)

Scatter plot shows launch outcomes (1 = success, 0 = failure) by payload mass and booster version.

- More launches within this range are successful, especially between 2,000–6,000 kg.
- Falcon 9 FT continues to show strong performance.
- Older versions (v1.0 = blue, v1.1 = red) display a higher mix of failures.



Payload vs. Launch Outcome scatter plot for all sites: Payload range (2,500-5,500)

This focused payload range shows the highest cluster of successful launches (class = 1).

- Falcon 9 FT (red) dominates the success region, showing strong reliability.
- Failures are minimal in this range, regardless of booster version.
- Suggests that 2,500–5,500 kg is a payload sweet spot for SpaceX missions.

Payload vs. Launch Outcome scatter plot for all sites: Payload range (0-2,500)

Launches in this lighter payload range show mixed outcomes.

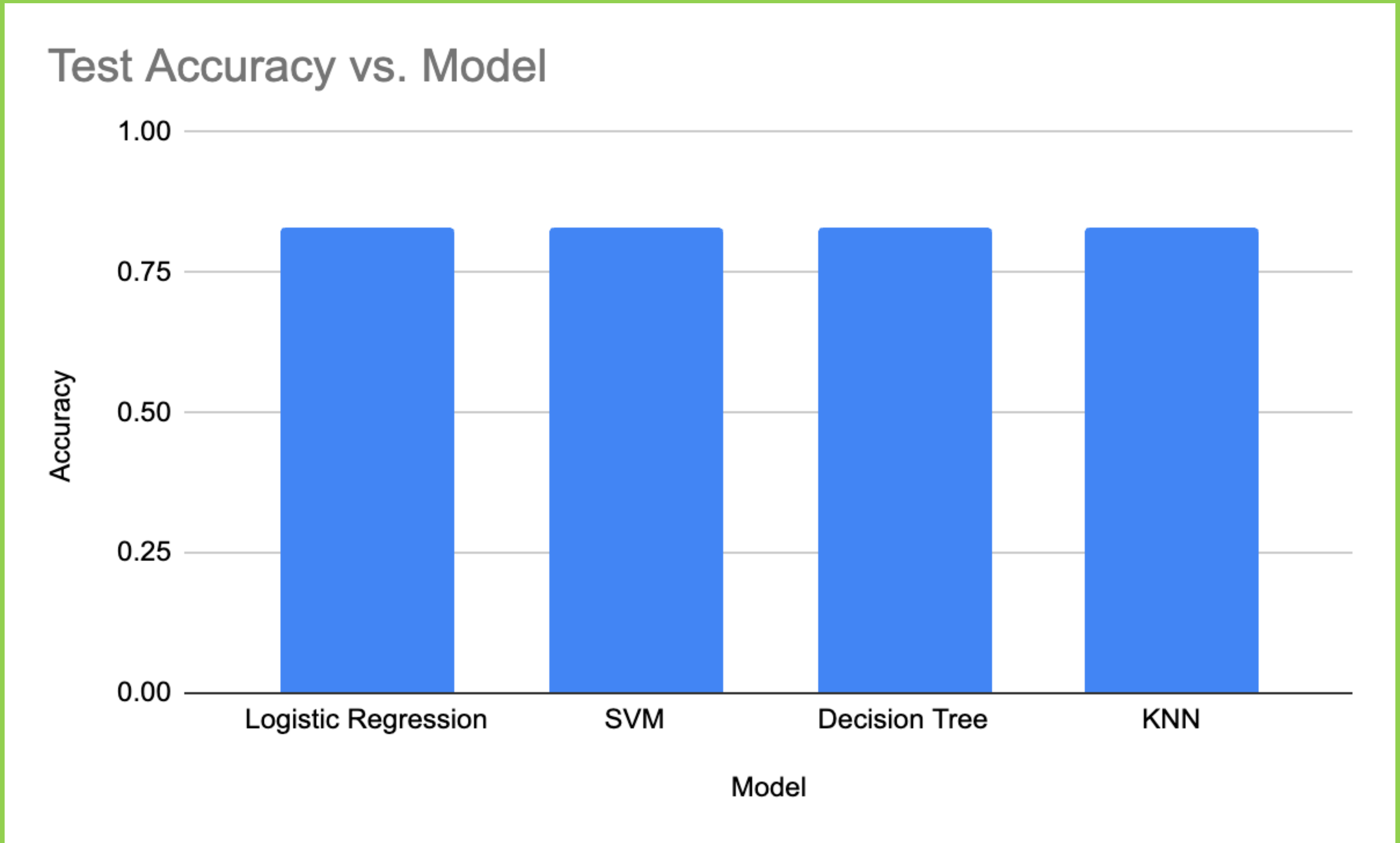
- Failure rates are higher compared to the mid-range (2,500–5,500 kg).
- Booster versions v1.0 and v1.1 (blue, red) are more prominent and scattered across both success and failure.
- Newer versions (FT, B5) are less frequent in this range, suggesting SpaceX shifted to heavier payloads over time.

Predictive Analysis (Classification)

Classification Accuracy

All models — Logistic Regression, SVM, Decision Tree, and KNN — achieved the same accuracy of 0.833 (83.3%) on the test set.

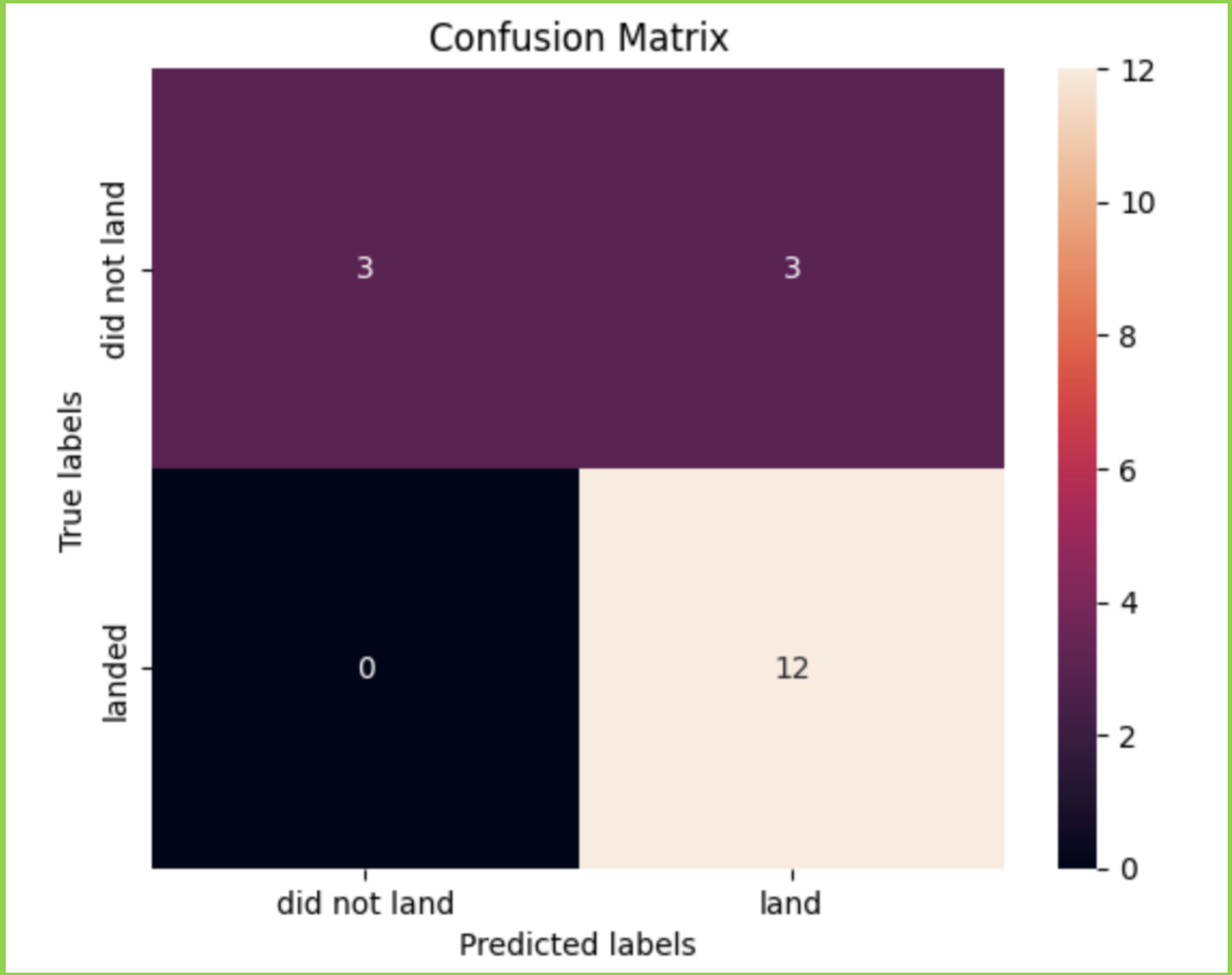
Since their performance is identical based on accuracy alone, no model clearly outperforms the others.



Confusion Matrix for Model with Highest Accuracy: Logistic Regression

Out of all models tested — each reaching an accuracy of 0.833 (83.3%) — Logistic Regression surfaced as the leading performer in the selection process.

However, due to the dataset's size, additional metrics or more data may be needed to confirm its advantage.



Conclusions

- ✅ Launch success rates vary by site — Cape Canaveral showed a high number of successful missions.
- 🚀 Booster version matters — Falcon 9 Full Thrust (FT) showed the most consistent success across payload ranges.
- 📦 Payload mass affects outcomes — Extreme low and high payloads were less likely to result in successful launches.
- 🌍 Proximity to infrastructure like railroads and highways can inform future launch site selection.
- 📊 All classification models performed similarly, each achieving 83.3% accuracy — suggesting limited differentiation with the current dataset.
- 🤖 Logistic Regression was algorithmically selected as top performer, but more data or metrics (like precision/recall) are needed for stronger conclusions.

Appendix

This is the query used to determine the first successful ground pad landing.

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

This is the query used to find drone ship failures in 2015

```
%%sql
SELECT
    CASE substr("Date", 6, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS Month,
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE substr("Date", 1, 4) = '2015'
AND "Landing_Outcome" LIKE 'Failure (drone ship)';
```

Appendix (Cont'd)

SpaceX launch site coordinates

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

The distance of each launch site to NASA (in Km)

	Launch Site	Distance_to_NASA_km
0	CCAFS LC-40	1413.328592
1	CCAFS SLC-40	1413.366602
2	KSC LC-39A	1406.434271
3	VAFB SLC-4E	2462.753721

Appendix (Cont'd)

This snippet shows the input used to determine the best performing model

```
# Calculate test accuracies
logreg_acc = logreg_cv.score(X_test, Y_test)
svm_acc = svm_cv.score(X_test, Y_test)
tree_acc = tree_cv.score(X_test, Y_test)
knn_acc = knn_cv.score(X_test, Y_test)

# Store in a dictionary
model_scores = {
    'Logistic Regression': logreg_acc,
    'SVM': svm_acc,
    'Decision Tree': tree_acc,
    'KNN': knn_acc
}

# Find the model with the highest accuracy
best_model = max(model_scores, key=model_scores.get)
best_score = model_scores[best_model]

print("Best model:", best_model)
print("Accuracy:", best_score)
```

Thank You For Your Attention