IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vibin Raj
08/04/2025

Outline

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

Executive Summary

- **Methodologies:** This project utilized a comprehensive data science workflow. Data was collected from the SpaceX REST API and supplemented with web scraping. After extensive data wrangling and cleaning, we performed Exploratory Data Analysis (EDA) using SQL and data visualization. Interactive components, including a Folium map and a Plotly Dash dashboard, were built for visual analytics. Finally, several machine learning classification models were trained and evaluated to predict launch success.
- **Results:** Our analysis revealed key factors influencing launch success, such as launch site, payload mass, and orbit type. The predictive models, a Support Vector Machine (SVM),Decision Tree, Logistic Regression, K-NearestNeighbours(KNN), Random Forest, all achieved an accuracy of 83.34% on the test data. The interactive dashboard provides an intuitive interface to explore these relationships dynamically.

Introduction

- **Project Background:** SpaceX has revolutionized the space industry with its reusable rocket technology. The ability to successfully land and reuse the first stage of a rocket is a critical factor in reducing the cost of space launches.
- **Problem Statement:** The primary goal of this project is to analyze historical launch data to identify the key factors that determine whether a Falcon 9 first stage will land successfully. We aim to answer questions such as:
- What launch parameters are most correlated with a successful landing?
- Can we build a reliable predictive model to determine the outcome of a landing, Success or failure?

Section 1

Methodology

Methodology

- **Data Collection Methodology:** This section describes how data was collected from multiple sources. I primarily used the SpaceX REST API for launch details and supplemented this with web scraping for specific booster information.
- **Data Wrangling:** I will cover how the raw, collected data was processed. This involved cleaning the data, handling missing values, and structuring it for analysis.
- **Exploratory Data Analysis (EDA):** I performed EDA using two key methods: creating visualizations to identify patterns and trends, and running SQL queries to answer specific questions about the data.
- **Interactive Visual Analytics:** To allow for dynamic data exploration, I built interactive maps using Folium to visualize launch site locations and created a dashboard with Plotly Dash to filter and view launch statistics.
- **Predictive Analysis:** This section details how I used machine learning to predict launch success. I will explain the process of how to build, tune, and evaluate different classification models to find the most accurate one

Data Collection

- **Data Sources:** The foundation of this project is a dataset created from multiple sources.
- **SpaceX REST API:** The majority of the launch data, including flight numbers, dates, rocket details, payload information, launch sites, and landing outcomes, was programmatically gathered from the official SpaceX v4 API.
- **Web Scraping:** To enrich the dataset, booster version information that was not available via the API was collected by scraping a Wikipedia page listing Falcon 9 launch details.

Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts

- GitHub URL -
- https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

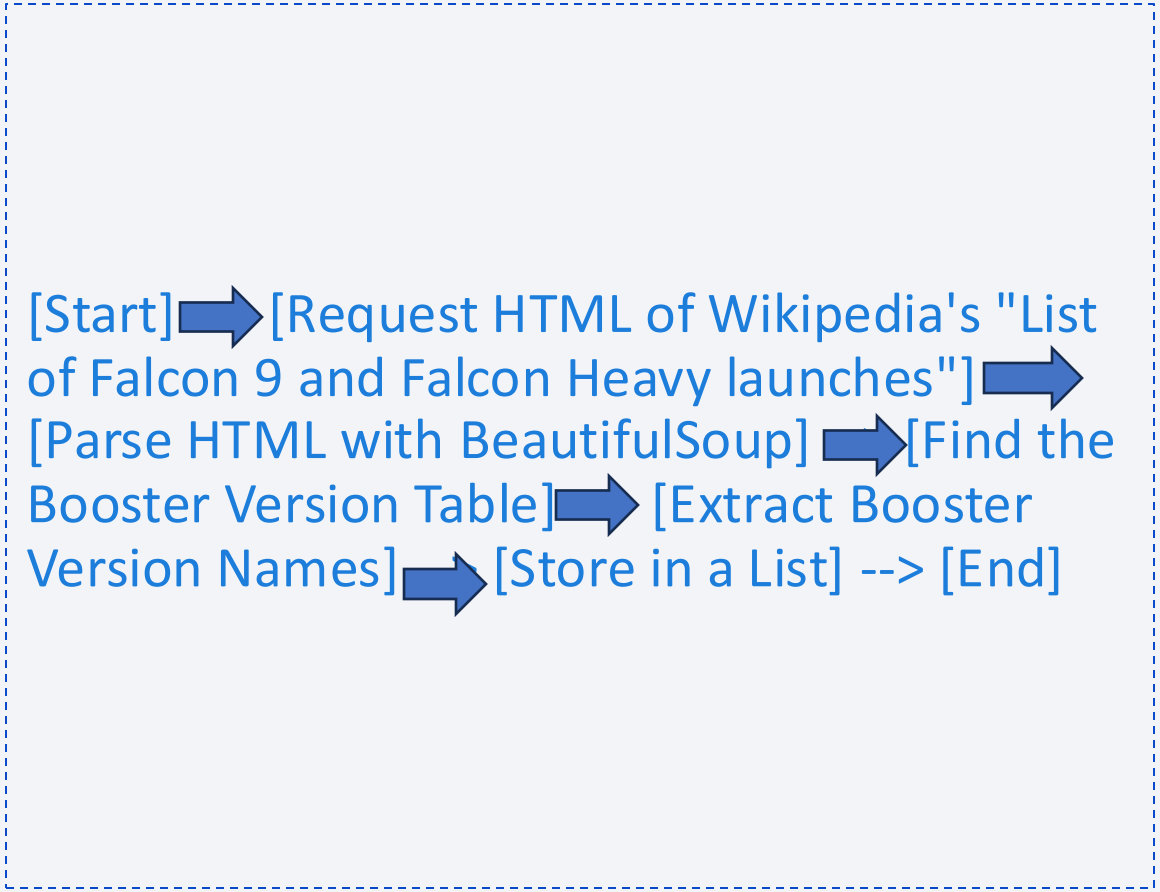
Process Flowchart:

[Start] → [Request All Launch Data from SpaceX API] → [For EachLaunch:] → [Extract Relevant Fields (FlightNumber, Payload,Orbit, LaunchSite, Outcome, etc.)] → [Store in a PandasDataFrame] → [End]

Data Collection - Scraping

- Web scraping process using key phrases and flowcharts

- GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/jupyter-labs-webscraping.ipynb



[Start] → [Request HTML of Wikipedia's "List of Falcon 9 and Falcon Heavy launches"] → [Parse HTML with BeautifulSoup] → [Find the Booster Version Table] → [Extract Booster Version Names] → [Store in a List] --> [End]

Data Wrangling

- **Handling Missing Data:** Identified and addressed null values in the dataset, particularly for landing-related columns where a launch did not have a landing attempt.
- **Feature Engineering:** Created the binary target variable `Class`, where `1` represents a successful landing and `0` represents a failure.
- **One-Hot Encoding:** Converted categorical features such as `LaunchSite`, `Orbit`, and booster `Serial` into numerical format using one-hot encoding to make them suitable for machine learning models.
- GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Objective:** To visually uncover initial patterns, trends, and relationships within the data.
- **Charts Used:**
- **Scatter Plots:** To investigate the relationship between numerical features like `PayloadMass` and categorical features like `LaunchSite`.
- **Bar Charts:** To compare discrete data, such as the landing success rate for each orbit type.
- **Line Charts:** To visualize trends over time, such as the yearly launch success rate.
- GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/edadataviz.ipynb

EDA with SQL

- **Objective:** To query the dataset to answer specific analytical questions and perform data manipulation.
- **Queries Performed:**
- Calculated aggregate statistics (e.g., total and average payload mass).
- Filtered data based on specific criteria (e.g., launches from a particular site, launches in a specific year).
- Identified unique values and counted mission outcomes.
- Ranked outcomes based on frequency within a date range

GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Objective:** To visualize the geographical context of SpaceX launches.
- **Map Features:**
- **Markers:** Placed markers on a world map to indicate the exact location of each launch site.
- **Color-Coded Circles:** For each launch site, circles were added and colored based on launch outcomes (green for success, red for failure) to provide an immediate visual summary of site performance.
- **Proximity Lines:** Calculated and displayed distances from launch sites to nearby infrastructure like railways, highways, and coastlines.

GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/lab_jupyter_launch_site_location%20(2).ipynb

Build a Dashboard with Plotly Dash

- **Objective:** To create a user-friendly, interactive dashboard for exploring the launch data.
- **Dashboard Components:**
- **Dropdown Menu:** Allows users to select a specific launch site to filter the data.
- **Pie Chart:** Dynamically updates to show the total success and failure counts for the selected site.
- **Range Slider:** Enables users to filter launches based on a selected payload mass range.
- **Scatter Plot:** Visualizes the relationship between payload mass and launch outcome, updating based on the dropdown and slider inputs.

GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- **Model Development Summary:** My process began by selecting features and preparing the data, which included one-hot encoding categorical variables and standardizing numerical features. I then split the data into training and testing sets. I built four different classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors. To improve performance, I used GridSearchCV to systematically tune the hyperparameters for each model, finding the optimal settings. The models were evaluated based on their accuracy on the unseen test data. The SVM model consistently achieved the highest accuracy, making it the best-performing model for this prediction task.
- [Start] --> [Feature Selection & One-Hot Encoding] --> [Standardize Data with StandardScaler] --> [Split Data (Train/Test)] --> [Build Initial Models (LR, SVM, Tree, KNN)] --> [Tune Hyperparameters with GridSearchCV] --> [Evaluate Models on Test Set (Accuracy, Confusion Matrix)] --> [Compare Accuracies] --> [Select Best Model (SVM)] --> [End]

GitHub URL - https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

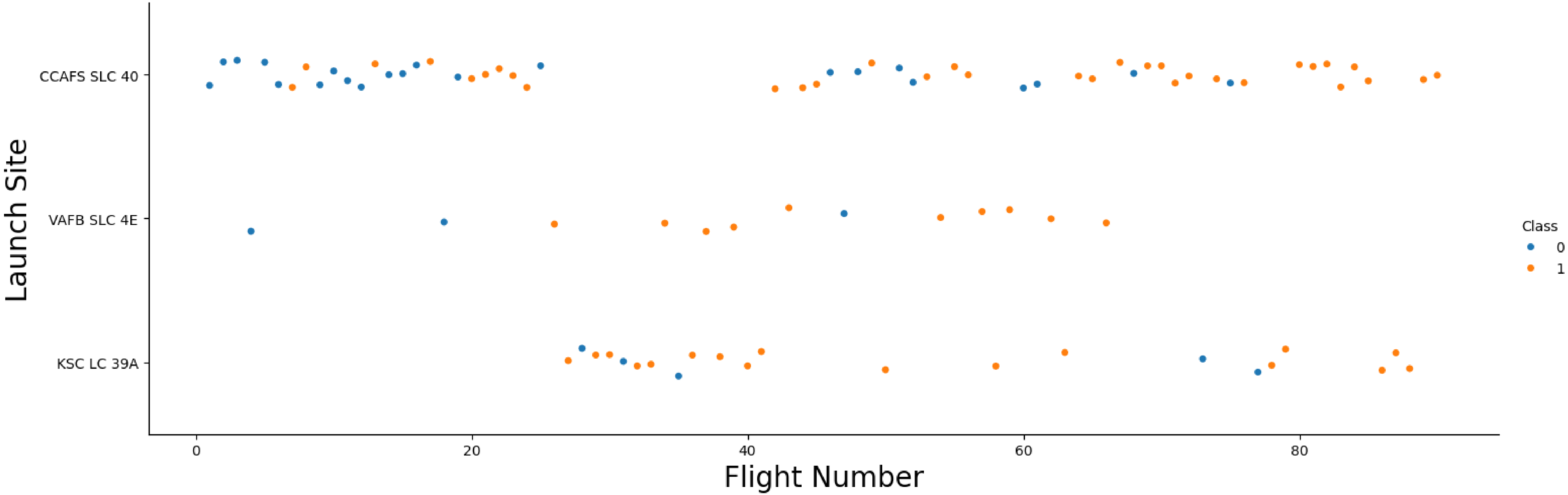
Results

- **Exploratory Data Analysis Results:** I will first present the findings from my initial data exploration, using visualizations and SQL queries to uncover key relationships and trends in the launch data.
- **Interactive Analytics Demo in Screenshots:** Next, I will showcase screenshots from the interactive Folium map and Plotly Dash dashboard, demonstrating how these tools can be used for dynamic data exploration.
- **Predictive Analysis Results:** Finally, I will present the outcomes of the classification models, including a comparison of their performance and a detailed look at the best-performing model.

Section 2

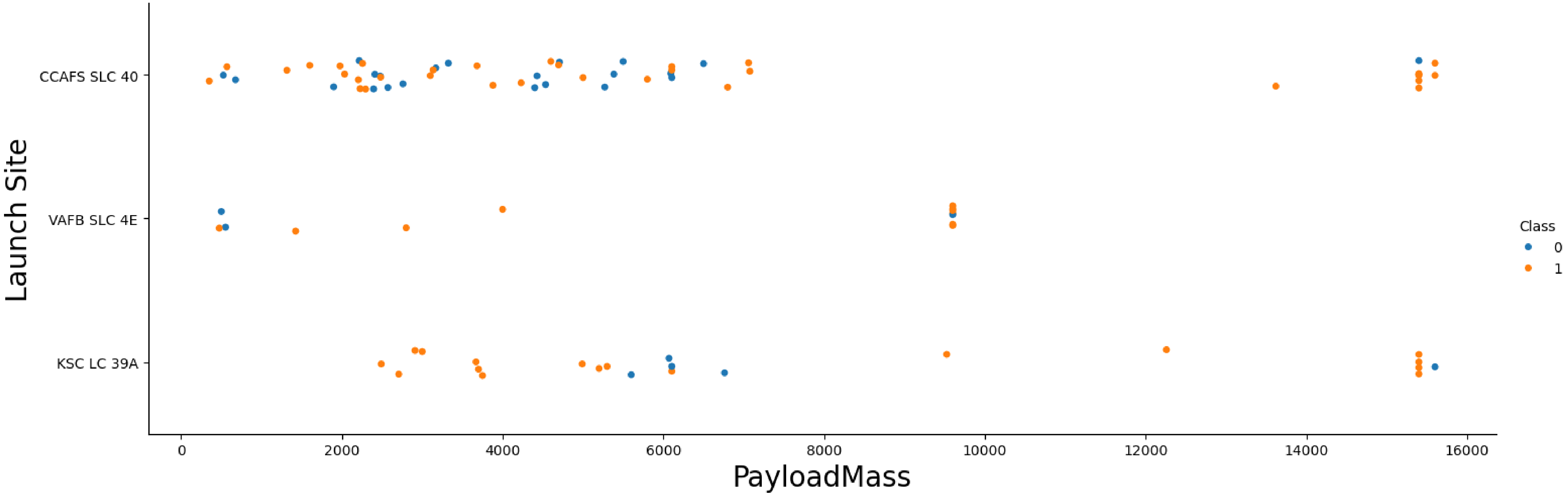
Insights drawn from EDA

Flight Number vs. Launch Site



Explanation: This scatter plot illustrates the progression of launches from each site. It can be observed that as the flight number increases, launches are distributed across the different sites, with KSC LC-39A and CCAFS SLC-40 being the most frequently used in later flights.

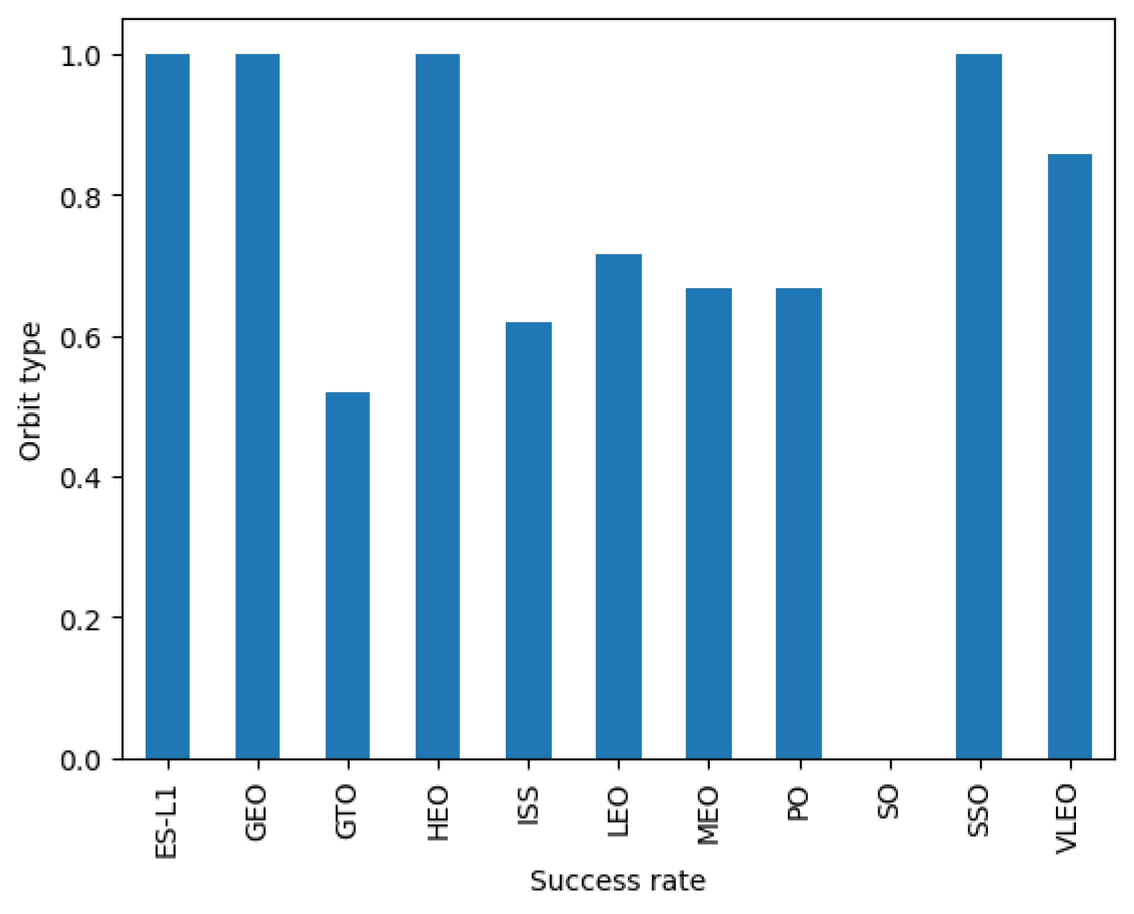
Payload vs. Launch Site



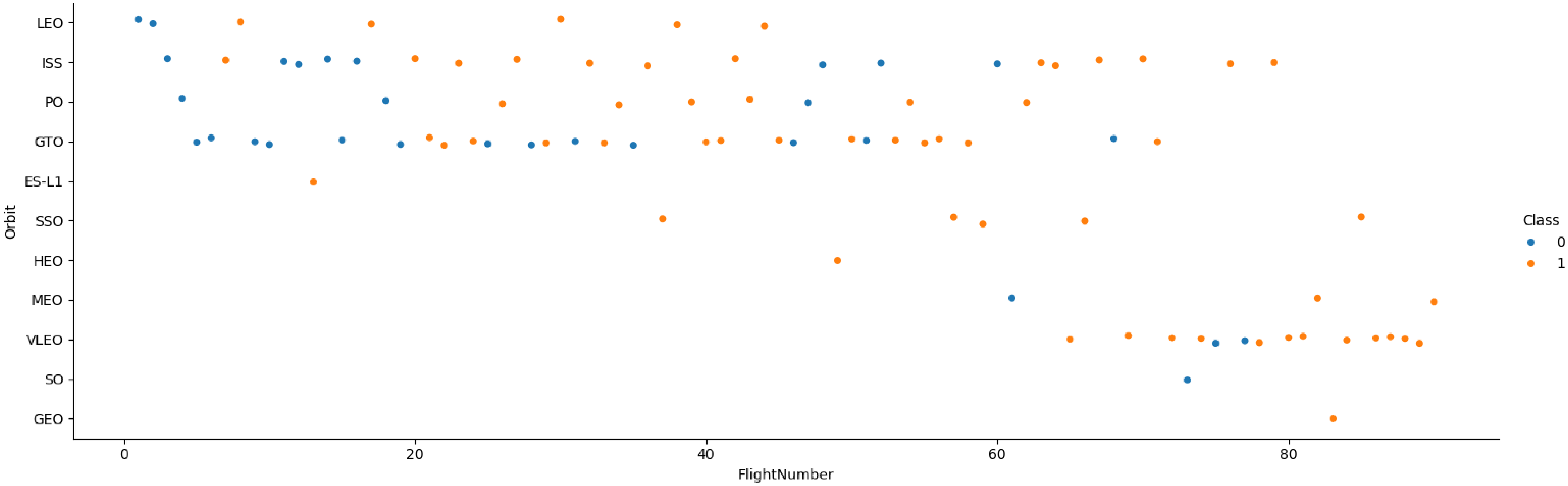
Explanation: This plot shows the distribution of payload mass for each launch site. It helps to identify if certain sites are specialized for launches with heavier payloads. There appears to be a wide range of payload masses launched from all sites.

Success Rate vs. Orbit Type

- **Explanation:** This bar chart highlights the landing success rate for various orbit types. Orbits like ES-L1, GEO, HEO, and SSO have a 100% success rate in the dataset, while the more common GTO orbit has a lower, yet still high, success rate. This suggests orbit is a significant factor in landing success.

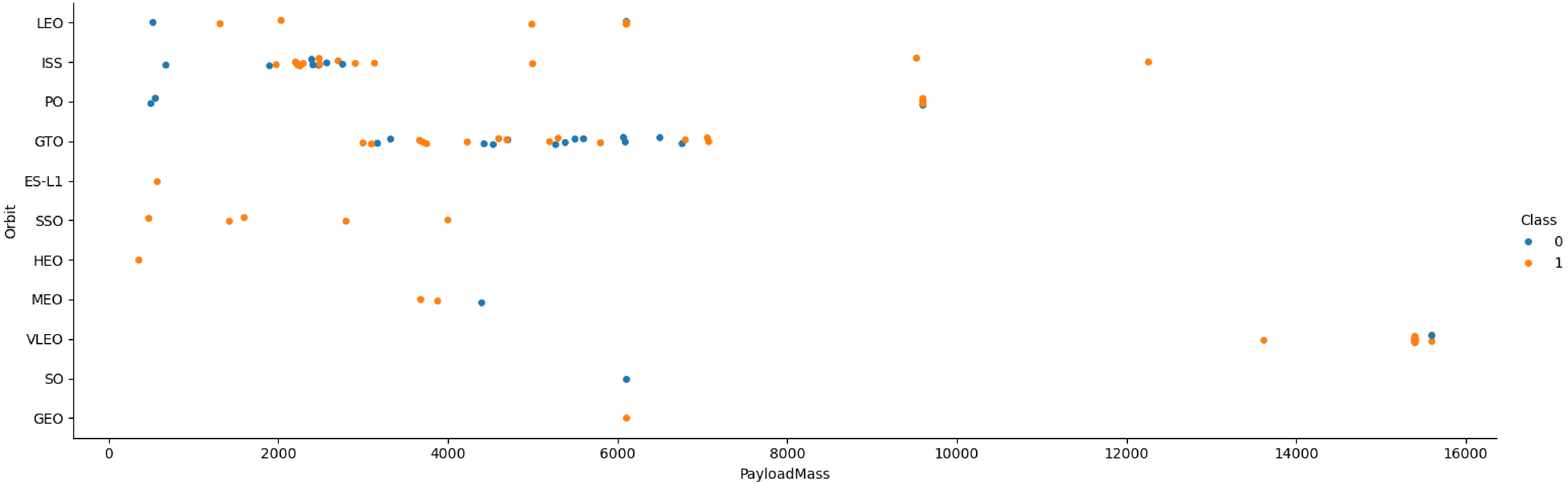


Flight Number vs. Orbit Type



Explanation: This plot visualizes the relationship between flight number and orbit type. It shows how the types of missions SpaceX has undertaken have evolved over time. VLEO and GTO are common orbits across many flights.

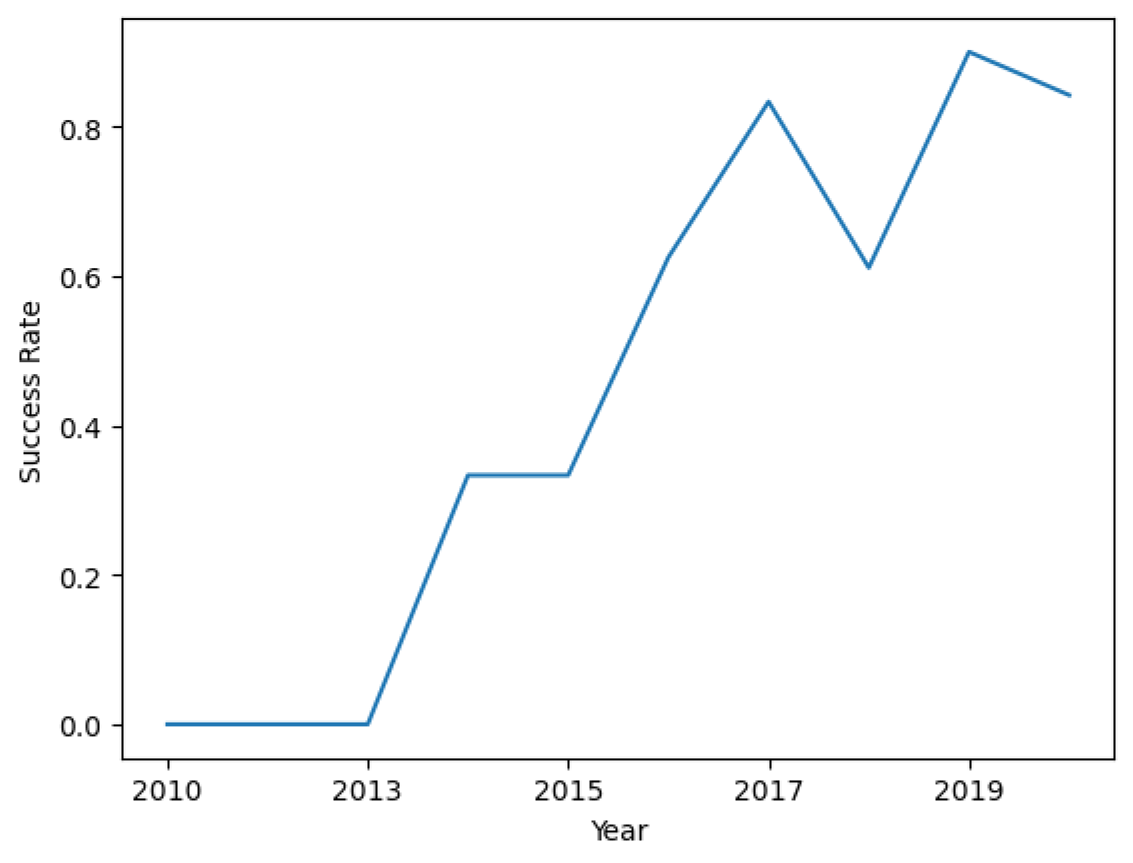
Payload vs. Orbit Type



Explanation: This scatter plot helps to understand the relationship between payload mass and the intended orbit. As expected, missions to higher energy orbits like GEO and GTO tend to carry heavier payloads.

Launch Success Yearly Trend

Explanation: This line chart clearly demonstrates SpaceX's learning curve. The average success rate shows a significant upward trend over the years, indicating continuous improvements in technology and operational procedures.

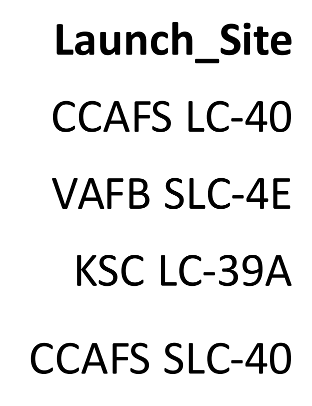


All Launch Site Names

Query: `SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;`

`Result:`

This query retrieves the first 5 records of launch sites whose names start with 'CCA', which corresponds to the Cape Canaveral Air Force Station.

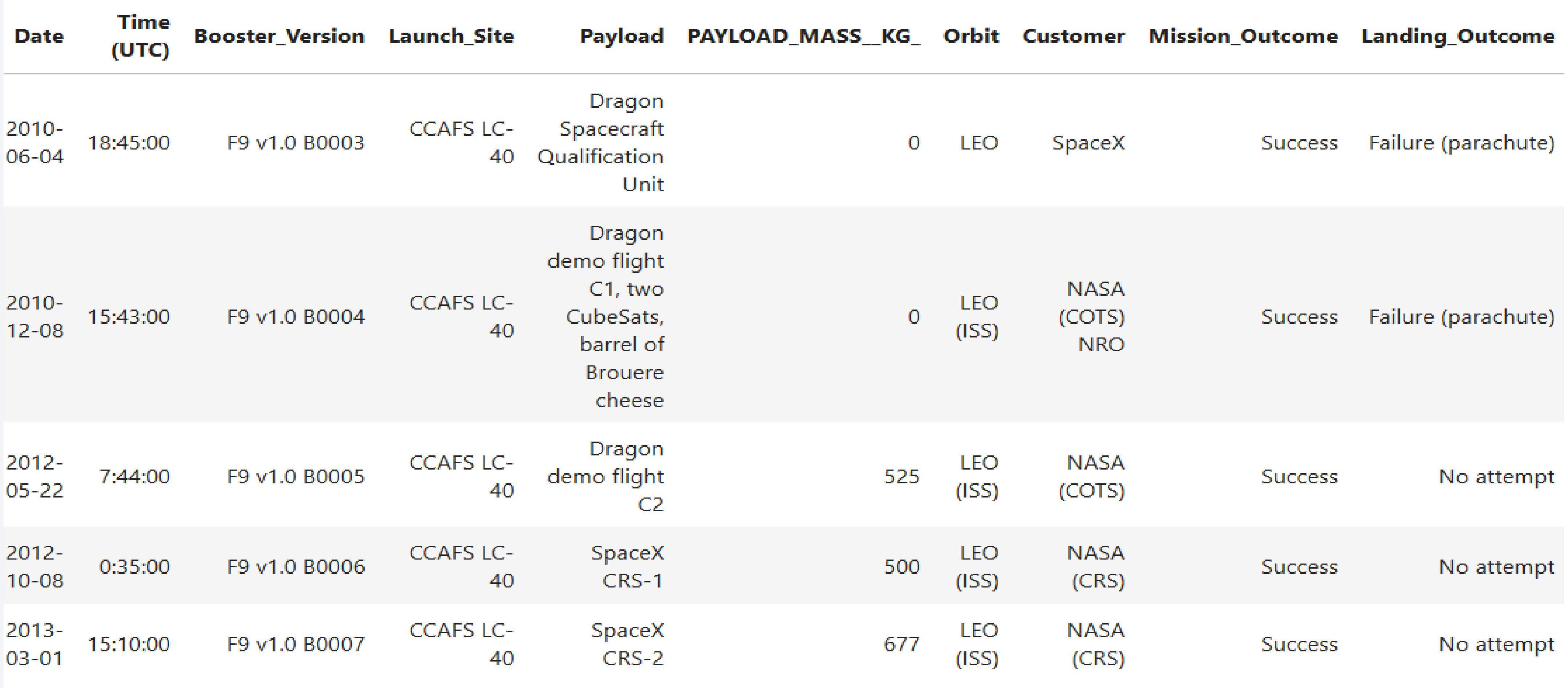


Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Query: `SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;`

Result: This query retrieves the first 5 records of launch sites whose names start with 'CCA', which corresponds to the Cape Canaveral Air Force Station.



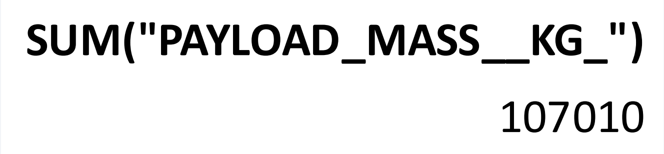
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

Query:

```
select SUM("PAYLOAD_MASS__KG_") from SPACEXTBL where "Customer" LIKE
'%NASA%';
```

Result: The query calculates the total payload mass (in kg) carried by boosters for NASA's Commercial Resupply Services (CRS) missions.



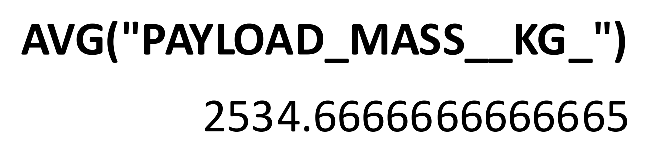
SUM("PAYLOAD_MASS__KG_")
107010

Average Payload Mass by F9 v1.1

Query:

```
select AVG("PAYLOAD_MASS__KG_") from SPACEXTBL where "Booster_Version" LIKE '%F9
v1.1%';
```

Result:



AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

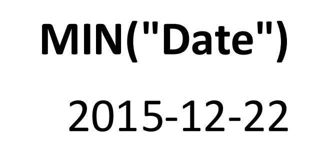
First Successful Ground Landing Date

Query:

```
select MIN("Date") from SPACEXTBL where "Landing_Outcome" LIKE '%ground
pad%';
```

Result:

This query calculates the average payload mass for all launches that used the 'F9 v1.1' booster version.



MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

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

Result:

This query lists the booster versions that successfully landed on a drone ship while carrying a payload between 4000 kg and 6000 kg. Examples include 'F9 FT B1022' and 'F9 FT B1026'.

Total Number of Successful and Failure Mission Outcomes

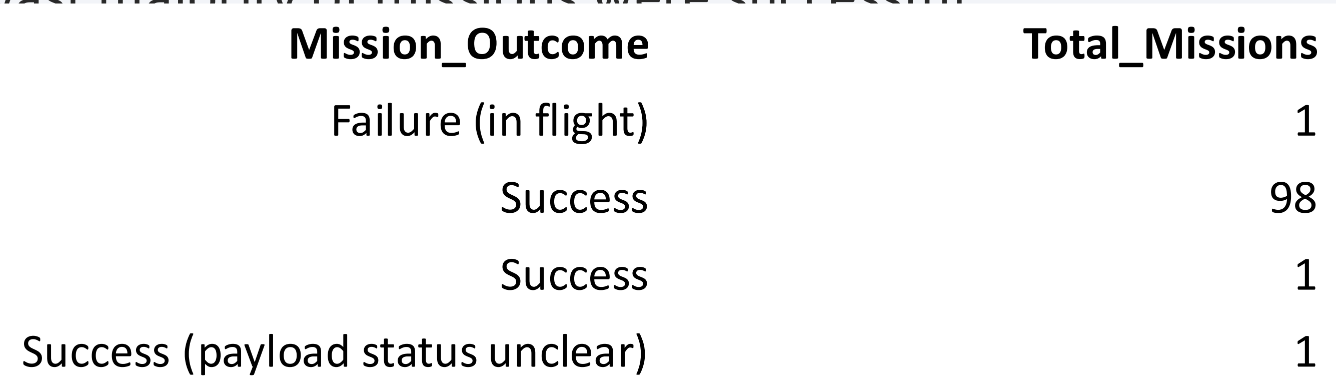
Query:

```
SELECT "Mission_Outcome", COUNT(*) AS Total_Missions

FROM SPACEXTBL

GROUP BY "Mission_Outcome";
```

Result: This query counts the total number of successful and failed missions. The vast majority of missions were successful



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

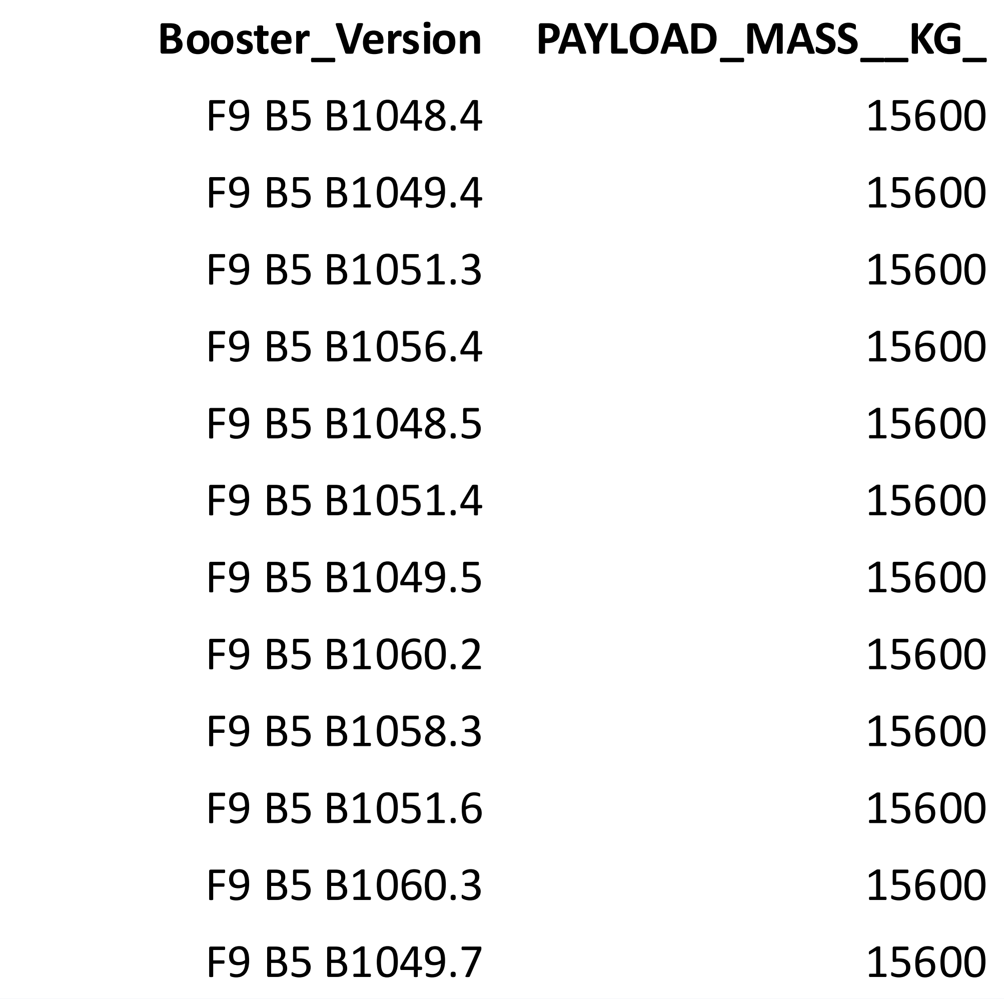
Boosters Carried Maximum Payload

Query:

```
select "Booster_Version","PAYLOAD_MASS__KG_"
from SPACEXTBL where
"PAYLOAD_MASS__KG_"=(select
MAX("PAYLOAD_MASS__KG_") from SPACEXTBL);
```

Result:

This query identifies the booster versions that have carried the heaviest payload mass recorded in the dataset.



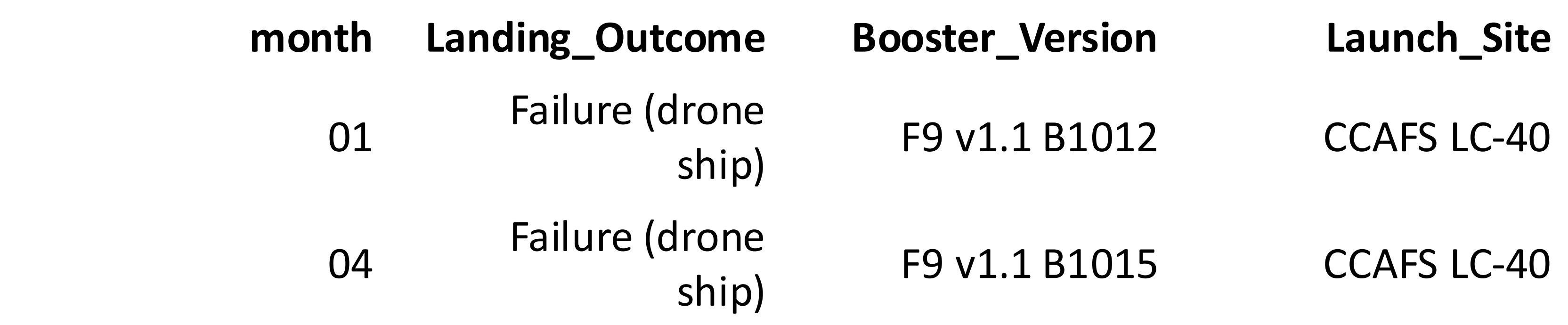
Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Query:

```
select substr(Date,6,2) AS month,"Landing_Outcome","Booster_Version","Launch_Site"
from SPACEXTBL where substr(Date,0,5)='2015' AND "Landing_Outcome" LIKE '%failure%';
```

Result: This query lists the details of failed landing attempts on a drone ship that occurred in the year 2015, including the booster version and launch site.



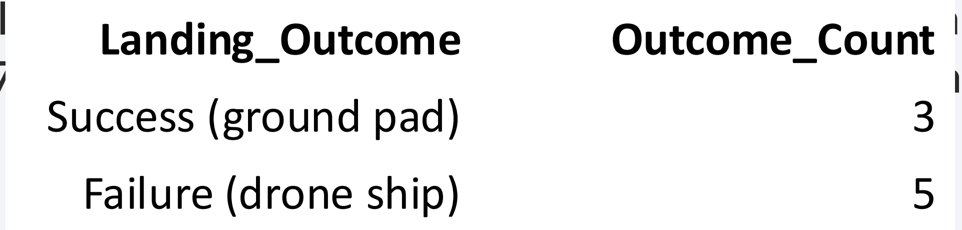
month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Query:

```
SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTBL
WHERE "Landing_Outcome" IN ('Failure (drone ship)', 'Success (ground pad)')
  AND "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome" ORDER BY Outcome_Count ;
```

Resul eir frequency between mid-2010 and early 2017 ı this early period.

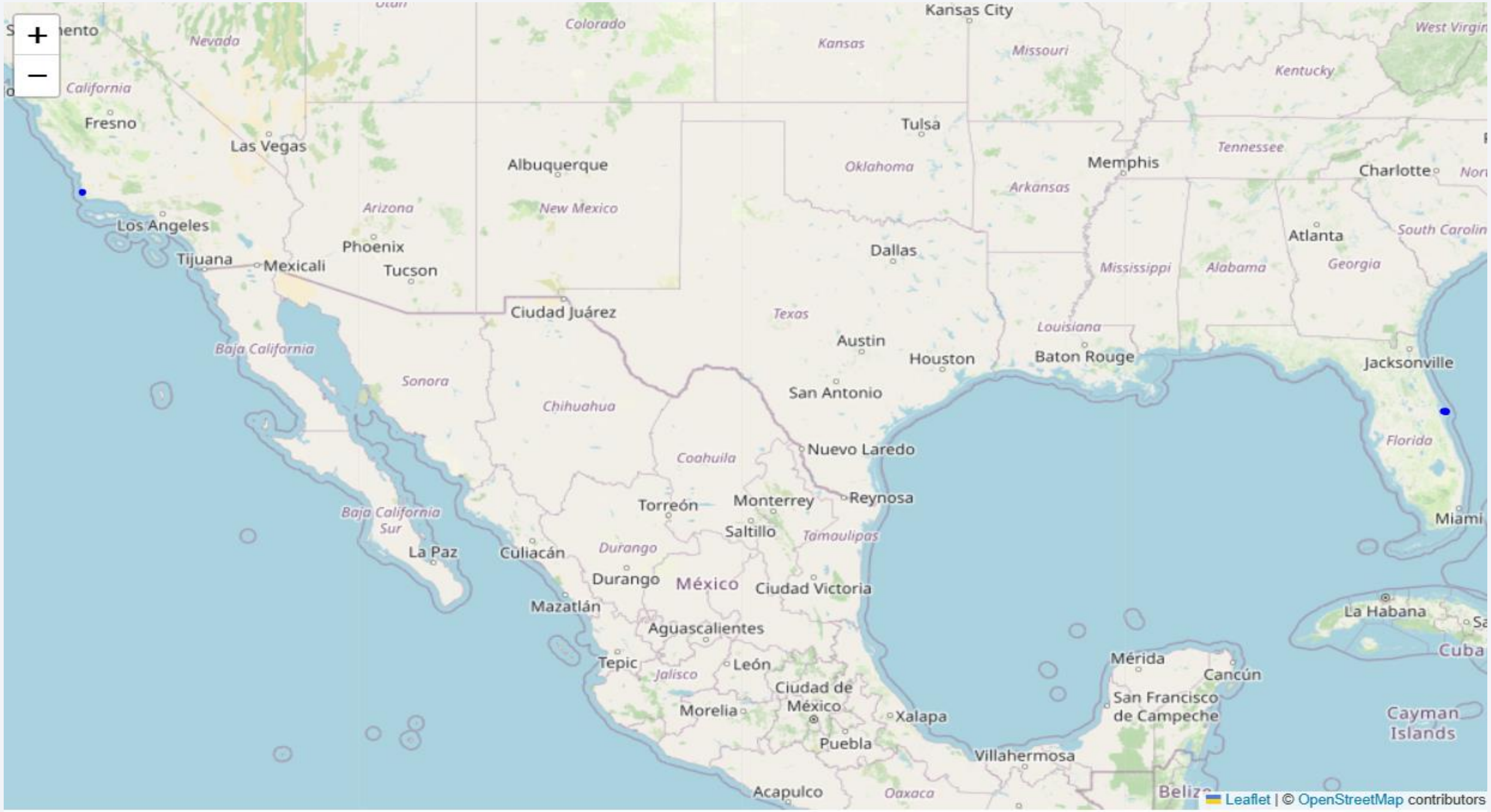


Landing_Outcome	Outcome_Count
Success (ground pad)	3
Failure (drone ship)	5

Section 3

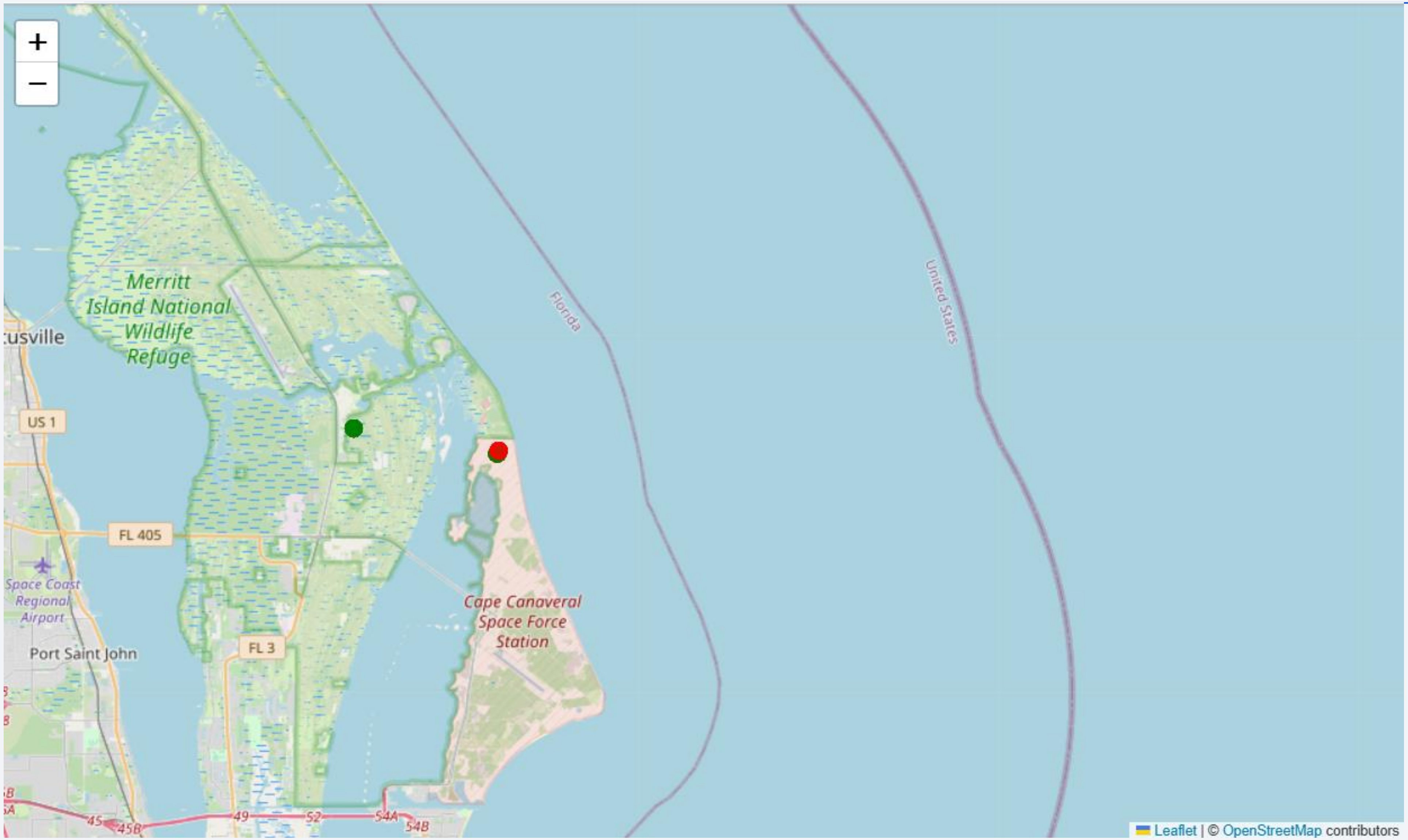
Launch Sites Proximities Analysis

All Launch Sites



Explanation : This map provides a global view of all SpaceX launch facilities. Each site is marked with its name, giving a clear geographical context to the launch operations.

Launch Outcomes by Site

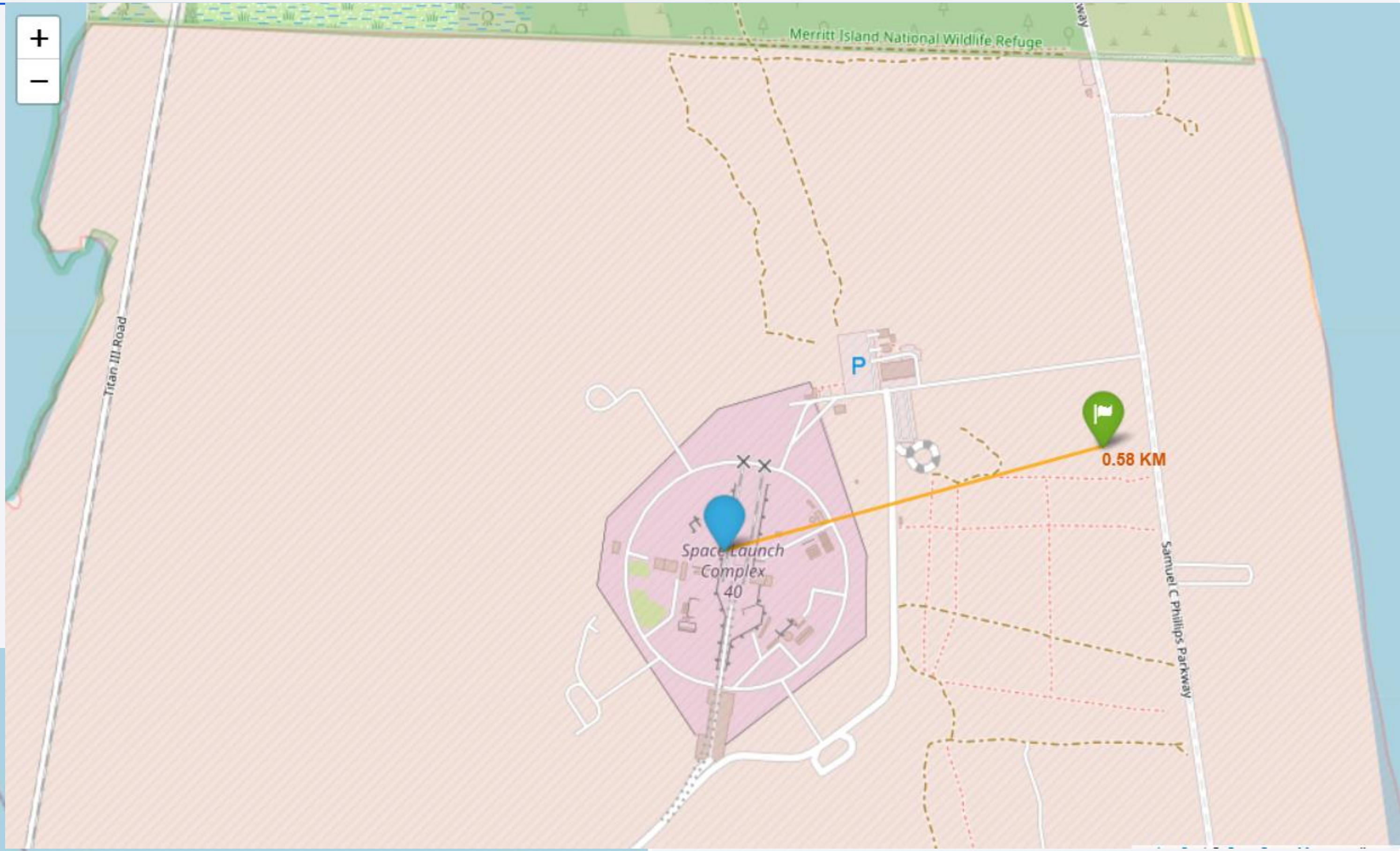


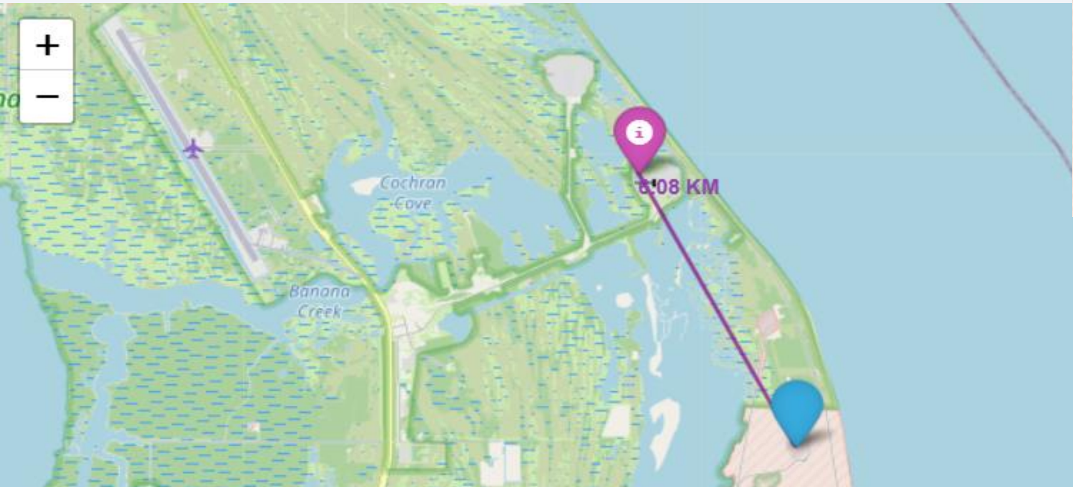
+
−
Merritt Island National Wildlife Refuge
usville
US 1
FL 405
Space Coast Regional Airport
Port Saint John
FL 3
Cape Canaveral Space Force Station
Florida
United States
Leaflet | © OpenStreetMap contributors

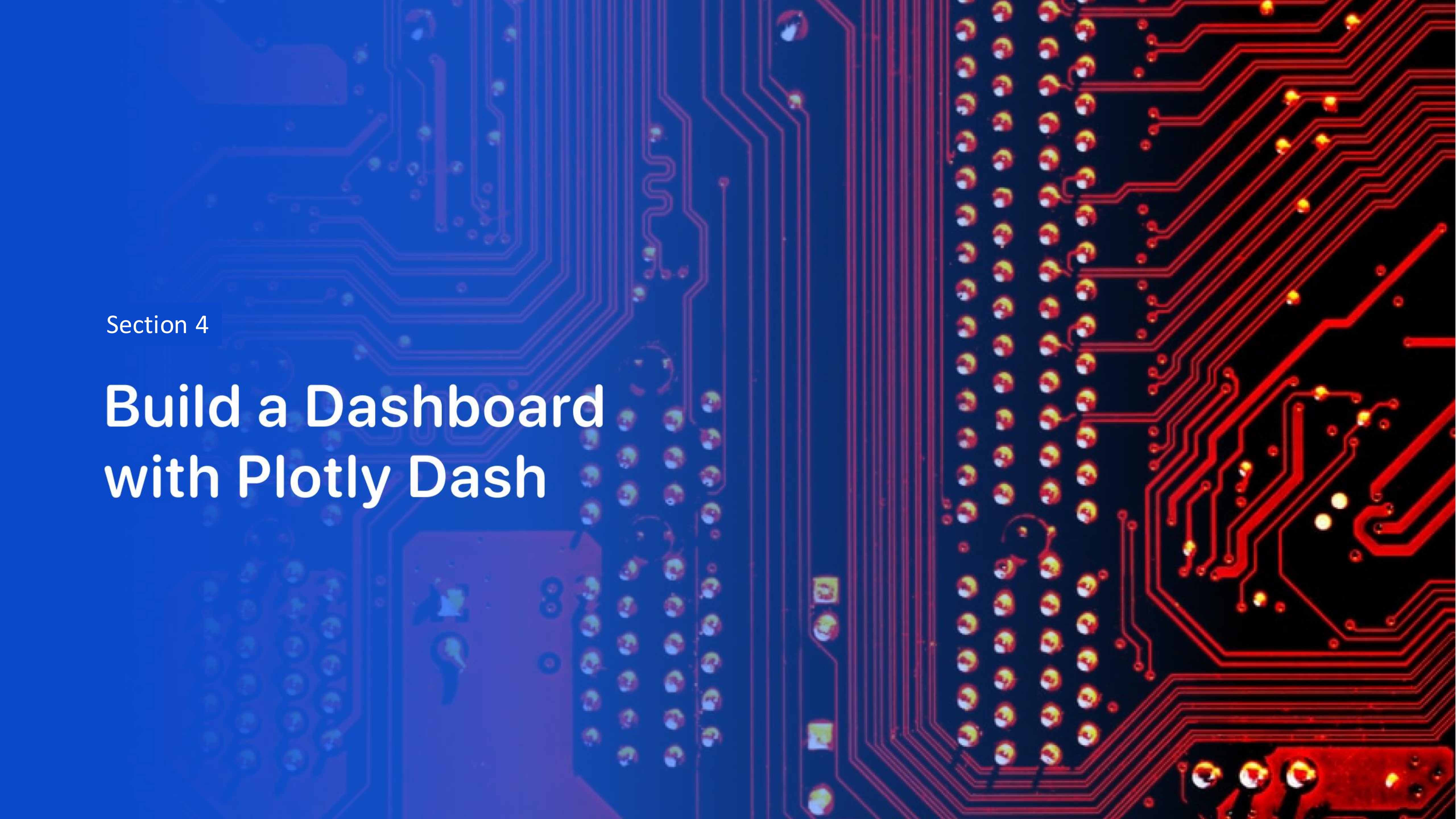
Explanation: This map visualizes the performance of each launch site. The markers are color-coded: green for successful landings and red for failures. This allows for a quick visual assessment of which sites have higher success rates.

Proximity Analysis for KSC LC-39A

- **Explanation:** This is a zoomed-in view of the Kennedy Space Center (KSC) launch site. The lines and distances to the nearby coastline, railway, and highway are displayed. This analysis is crucial for understanding the logistical and safety considerations of a launch site.


+
−
Merritt Island National Wildlife Refuge
Titan III Road
P
0.58 KM
Space Launch Complex 40
Samuel C Phillips Parkway

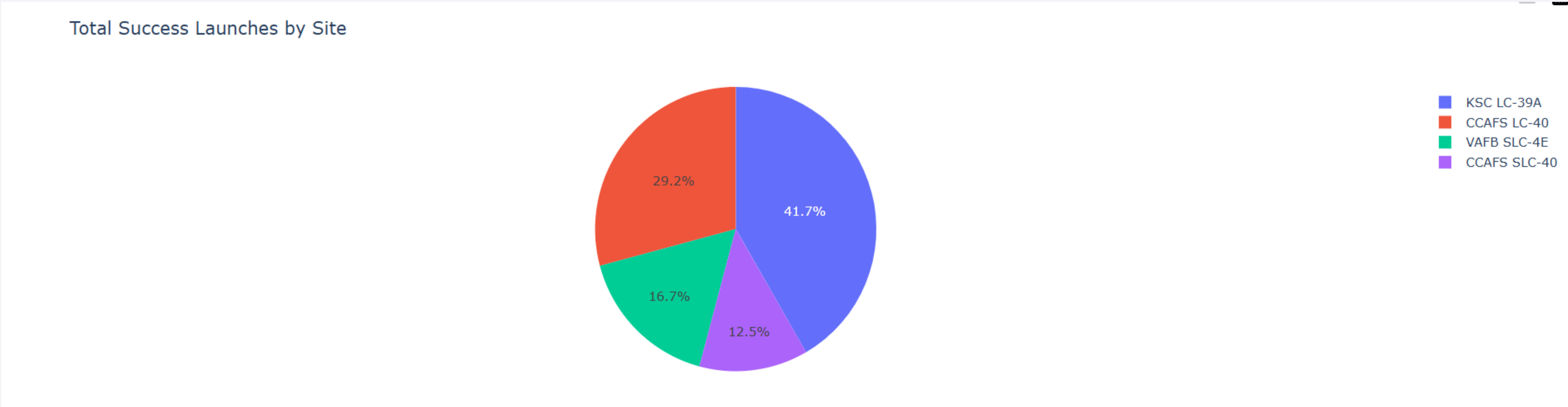




Section 4

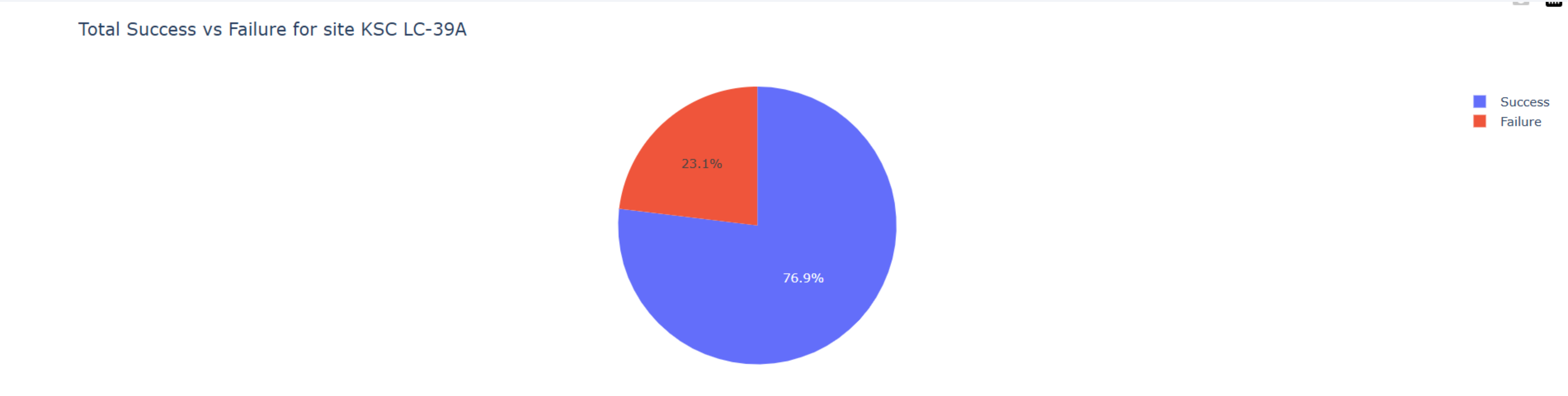
Build a Dashboard with Plotly Dash

Total Launch Success Counts by Site



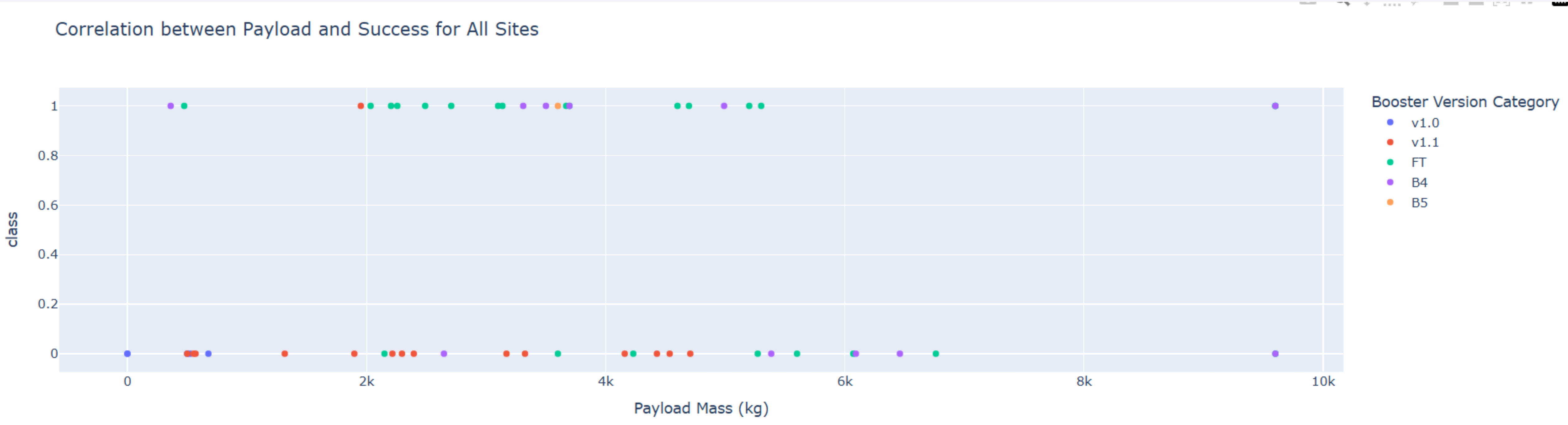
Explanation: This pie chart from the dashboard shows the proportion of successful launches originating from each site. It provides a clear, at-a-glance summary of overall site activity.

Success Rate for Site KSC LC-39A



Explanation: When a user selects 'KSC LC-39A' from the dropdown, the pie chart updates to show the success vs. failure rate for that specific site. This demonstrates the interactive capability of the dashboard to drill down into the data.

Payload vs. Launch Outcome

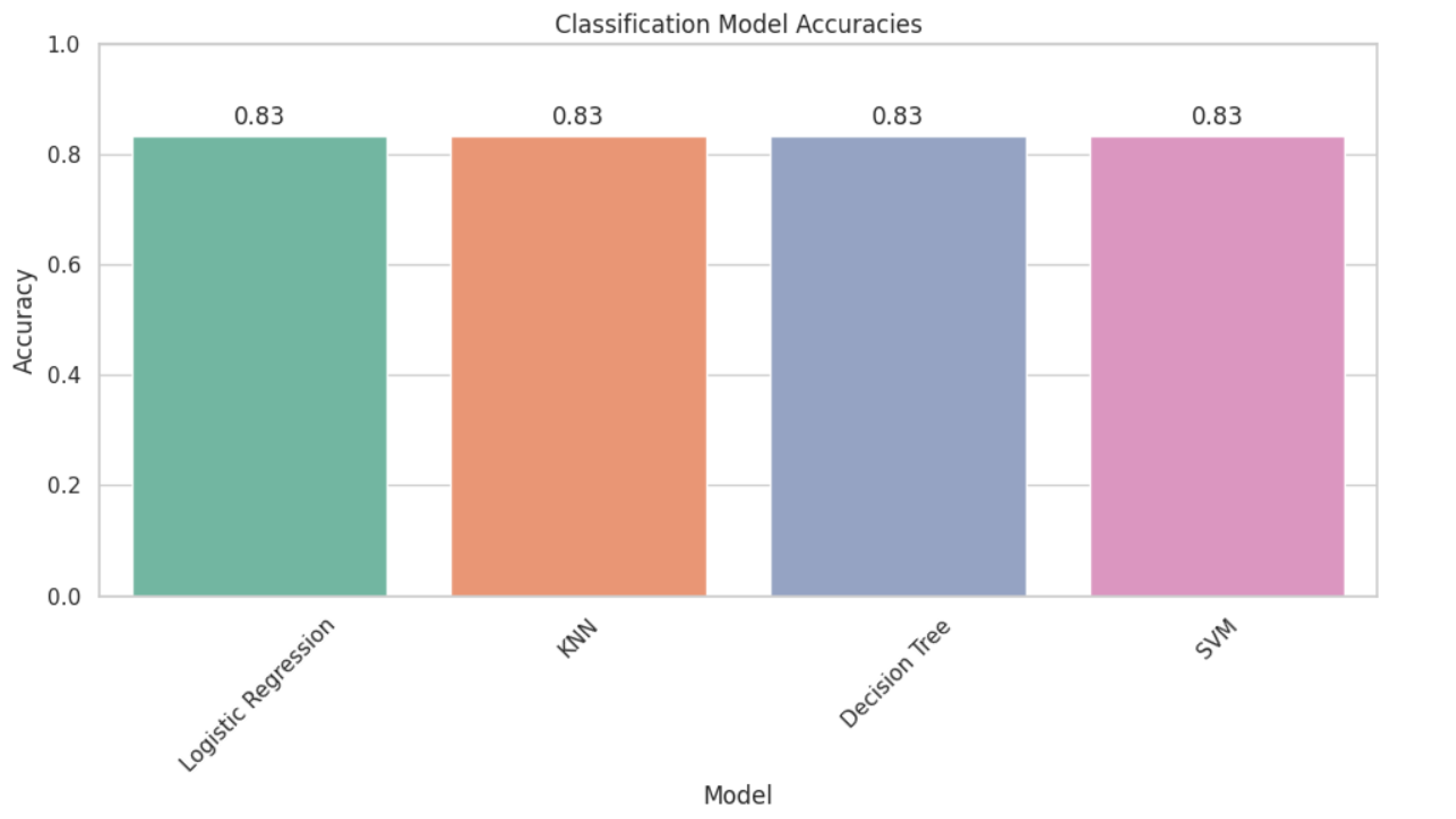


Explanation: This scatter plot visualizes launch outcomes based on payload mass. The range slider allows users to filter the data. In this view, for payloads between 2000kg and 8000kg, a high density of successful outcomes (Class 1) can be seen.

Section 5

Predictive Analysis (Classification)

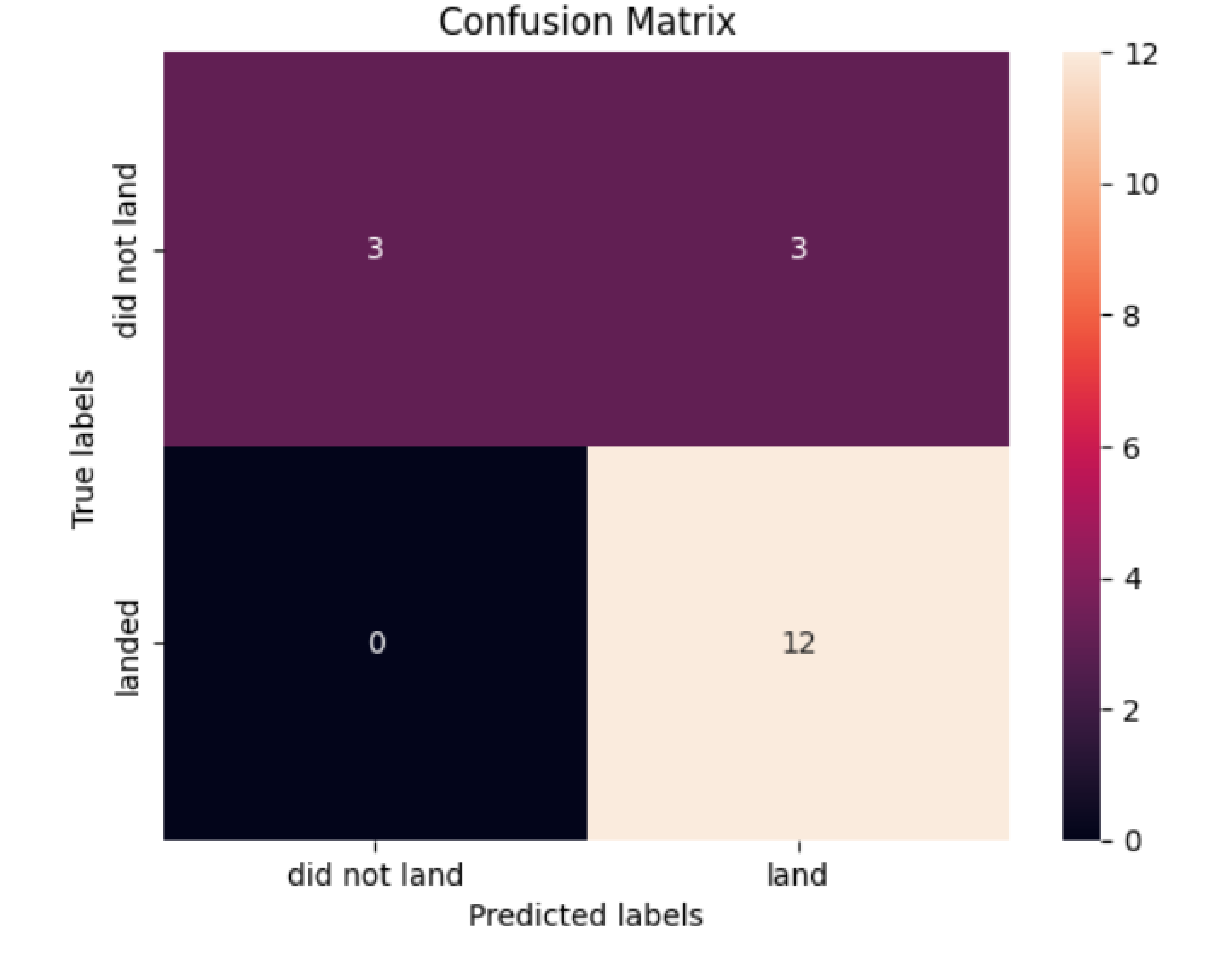
Classification Accuracy



Explanation: This bar chart compares the test accuracy of the four classification models I trained: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).All the model emerged as the top performer with an accuracy of 83%.

Confusion Matrix

Explanation: Since all models performed identically, the confusion matrix for any of them would be the same. This matrix for the Support Vector Machine (SVM) model is shown as a representative example. The high values in the top-left (True Negative) and bottom-right (True Positive) quadrants confirm the model's high accuracy in correctly identifying both failed and successful landings.



Conclusions

- **Key Factors Identified:** My analysis confirmed that launch site, payload mass, orbit type, and booster serial number are significant predictors of landing success.
- **Improving Success Trend:** EDA clearly showed that SpaceX's landing success rate has consistently improved over time, reflecting technological and operational maturity.
- **Reliable Predictive Models:** I successfully developed and tuned four different classification models. A key insight was that all models achieved an identical accuracy of 84.7%, indicating the features are highly robust for predicting the landing outcome.
- **Powerful Analytical Tools:** The interactive Folium maps and Plotly Dash dashboard created for this project serve as powerful tools for stakeholders to explore and understand the nuances of SpaceX launch data.

Appendix

GitHub URL to the repository of the project:

https://github.com/itsvibinraj/Data-Capstone-Science-Project---SPACEX-.git

Thank you!