IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Desirée Santos Rivera
August 2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection: Gathered and cleaned Falcon 9 landing data using RESTful API, web scraping, and Pandas.
- EDA & Visualization: Conducted EDA with Matplotlib, Seaborn, and SQL queries to identify key success factors. Developed interactive visualizations with Plotly Dash and Folium.
- Predictive Modeling: Built and optimized machine learning models (SVM, Classification Trees, Logistic Regression) to predict landing success.

Summary of all results

- Key Insights: Identified payload mass, launch site, and booster version as crucial for landing success.
- Interactive Tools: Dashboards and maps provided actionable insights on launch.
- Best Model: DecisionTreeClassifier achieved the highest accuracy, with 87.50 % on the test set.

Introduction

The Falcon 9 rocket, developed by SpaceX, is renowned for its cost-effective approach to space launches, primarily due to the reusability of its first stage. With a launch cost of $62 million compared to over $165 million for the competition, the ability to predict first stage landing success can significantly impact the overall cost and competitiveness of space missions. Therefore, understanding and predicting landing success can provide a strategic advantage in the aerospace industry.

Our primary focus is on predicting landing success and uncovering valuable patterns from historical landing data.

Section 1

Methodology

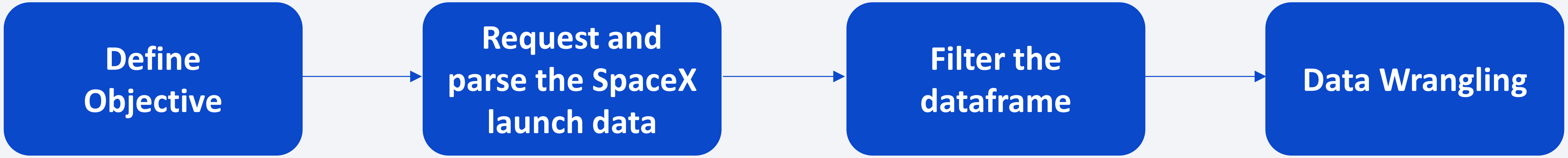
Methodology

Executive Summary

- Data collection methodology:
    - Accessed SpaceX public API for historical launch data. Extracted JSON data and converted it into a structured DataFrame.
- Perform data wrangling
    - Cleaned and filtered the data, removing missing values and organizing key variables such as flight number, date, rocket version, payload mass, launch site, and landing outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Implemented models including KNN, Decision Tree, SVM, and Logistic Regression to predict landing success.
    - Developed models, optimized hyperparameters, and evaluated performance using accuracy and confusion matrices to enhance predictive accuracy.

Data Collection

To predict Falcon 9 landing success, we first defined the target and accessed the SpaceX public API to obtain historical launch data. We converted this JSON data into a structured format in a DataFrame, filtering out relevant information such as flight details, rockets, launch pads, and payloads. We performed data cleaning to remove missing values and ensured data quality. Finally, we organized the data into a final DataFrame that included flight number, date, rocket version, payload mass, launch site, and landing outcome.



Data Collection – SpaceX API

For a detailed view of the completed SpaceX API calls and data processing, you can refer to the GitHub repository:

SpaceX API Data Collection Notebook

- ✓ Define Objective
- Access SpaceX API (GET Endpoinit)
- Data Extraction (Parse JSON)
- Data Conversion (JSON to DataFrame)
- Data Cleaning and Filtering
- Data Preparation

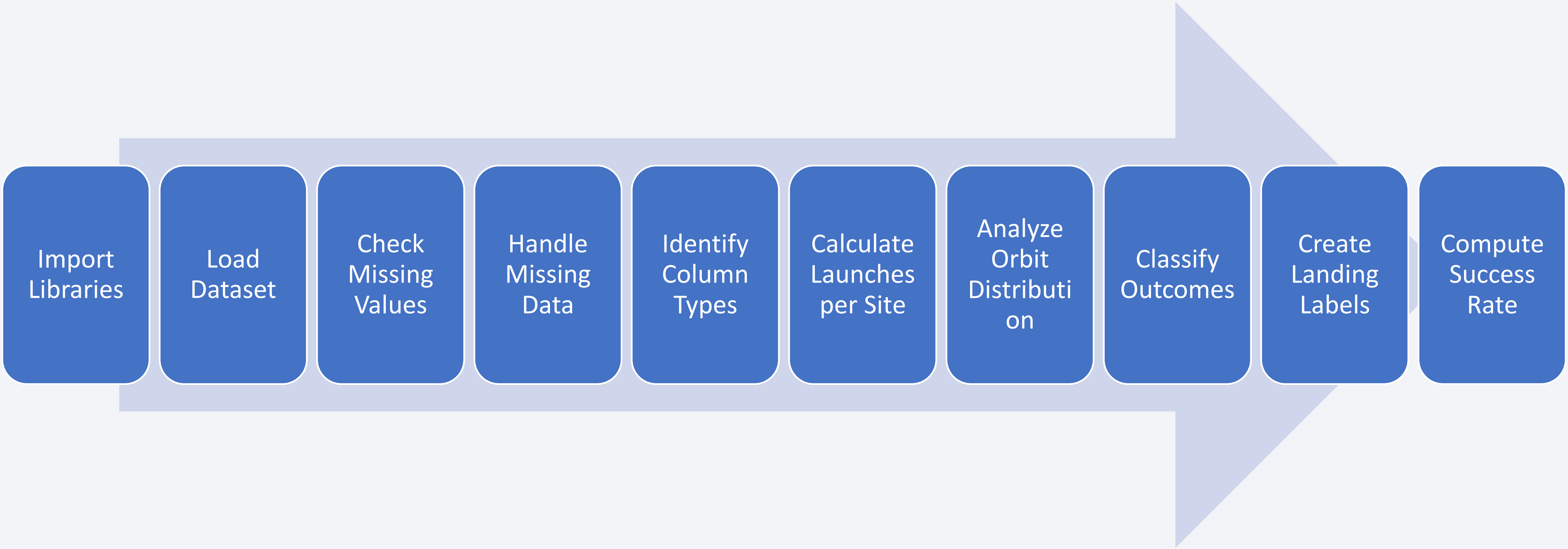
Data Collection - Scraping

For a complete view of the web scraping notebook, including code and results, please visit the GitHub repository:

SpaceX Falcon 9 Web Scraping Notebook

- Request HTML Page (using Requests library)
- Create BeautifulSoup Object (Parse HTML)
- Identify Target Table
- Extract Table Elements
- Extract Column Names
- Parse Table Rows
- Clean an Organize Data

Data Wrangling



Import Libraries

Load Dataset

Check Missing Values

Handle Missing Data

Identify Column Types

Calculate Launches per Site

Analyze Orbit Distributi on

Classify Outcomes

Create Landing Labels

Compute Success Rate

For detailed implementation and code, refer to our GitHub repository:
[SpaceX Falcon 9 Data Wrangling Notebook]

EDA with Data Visualization

Flight Number vs. Payload Mass (Catplot): Examine if flight attempts and payload mass affect landing success.

Flight Number vs. Launch Site (Catplot): Identify how different launch sites impact landing success over flight numbers.

Payload Mass vs. Launch Site (Catplot): Determine if payload mass varies by launch site and its effect on success.

Success Rate by Orbit Type (Bar Chart): Compare landing success rates across different orbit types.

Flight Number vs. Orbit Type (Scatterplot): Assess the relationship between flight numbers and orbit types on success rates.

Payload Mass vs. Orbit Type (Scatterplot): Investigate how payload mass impacts success rates for each orbit type.

Yearly Trend of Launch Success (Line Chart): Track changes in success rates over the years.

Feature Engineering - One-Hot Encoding and Data Preparation: Convert categorical data into numeric for modeling and ensure data consistency.

For the full EDA and visualizations, visit: [SpaceX Falcon 9 EDA Notebook]()

EDA with SQL

Displayed unique launch sites.

Showed 5 records where launch sites begin with 'CCA'.

Calculated total payload mass for NASA (CRS).

Computed the average payload mass for booster version F9 v1.1.

Listed the date of the first successful landing outcome on a ground pad.

Identified boosters with successful drone ship landings and payload mass between 4000 and 6000 kg.

Counted the number of successful and failed mission outcomes.

Found booster versions carrying the maximum payload mass using a subquery.

Displayed records from 2015 with failure landings on a drone ship, including month names and relevant details.

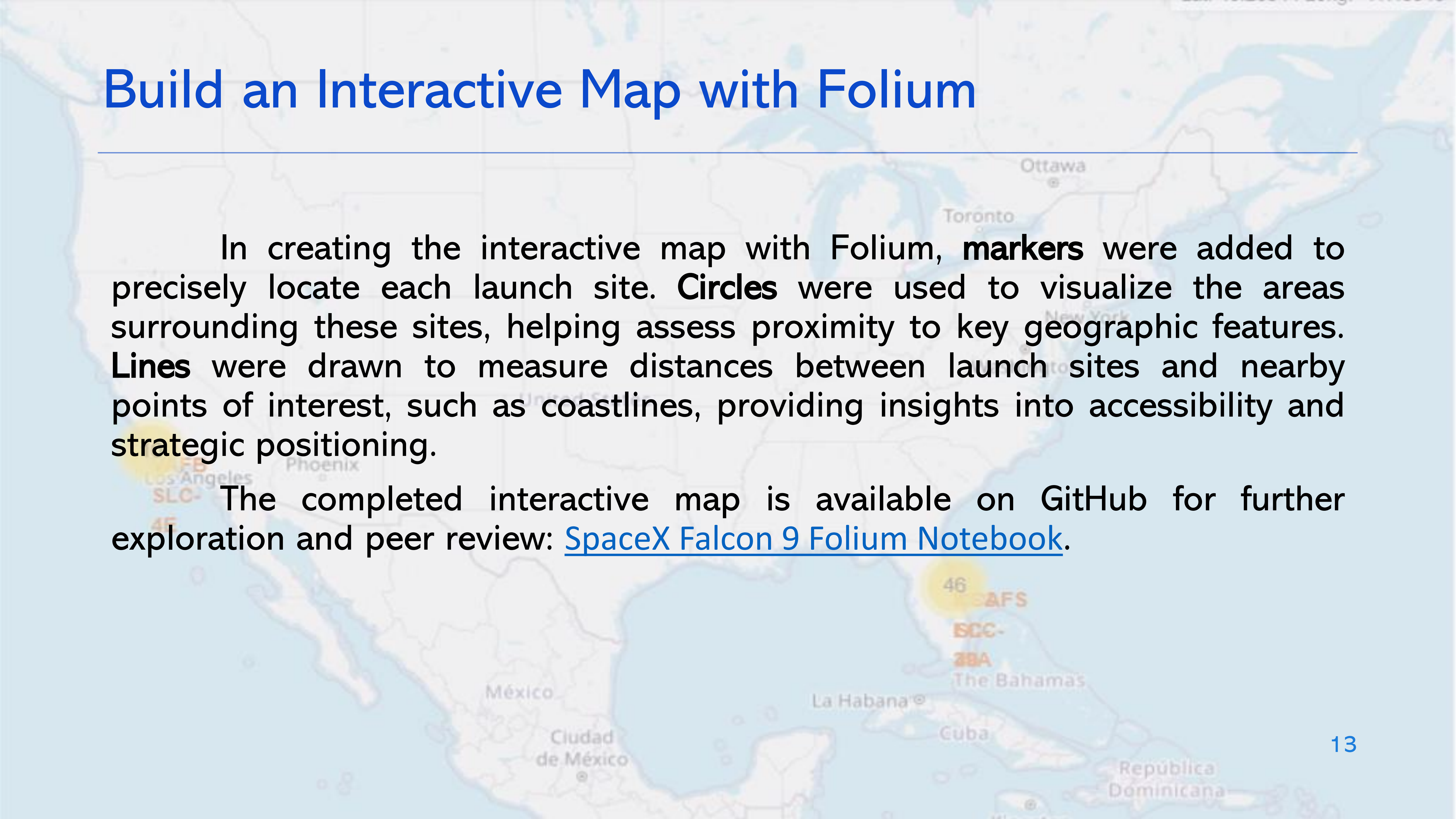
Ranked landing outcomes between 2010-06-04 and 2017-03-20 by count in descending order.

For the full EDA with SQL, visit: SpaceX Falcon 9 EDA with SQL Notebook

Build an Interactive Map with Folium

In creating the interactive map with Folium, **markers** were added to precisely locate each launch site. **Circles** were used to visualize the areas surrounding these sites, helping assess proximity to key geographic features. **Lines** were drawn to measure distances between launch sites and nearby points of interest, such as coastlines, providing insights into accessibility and strategic positioning.

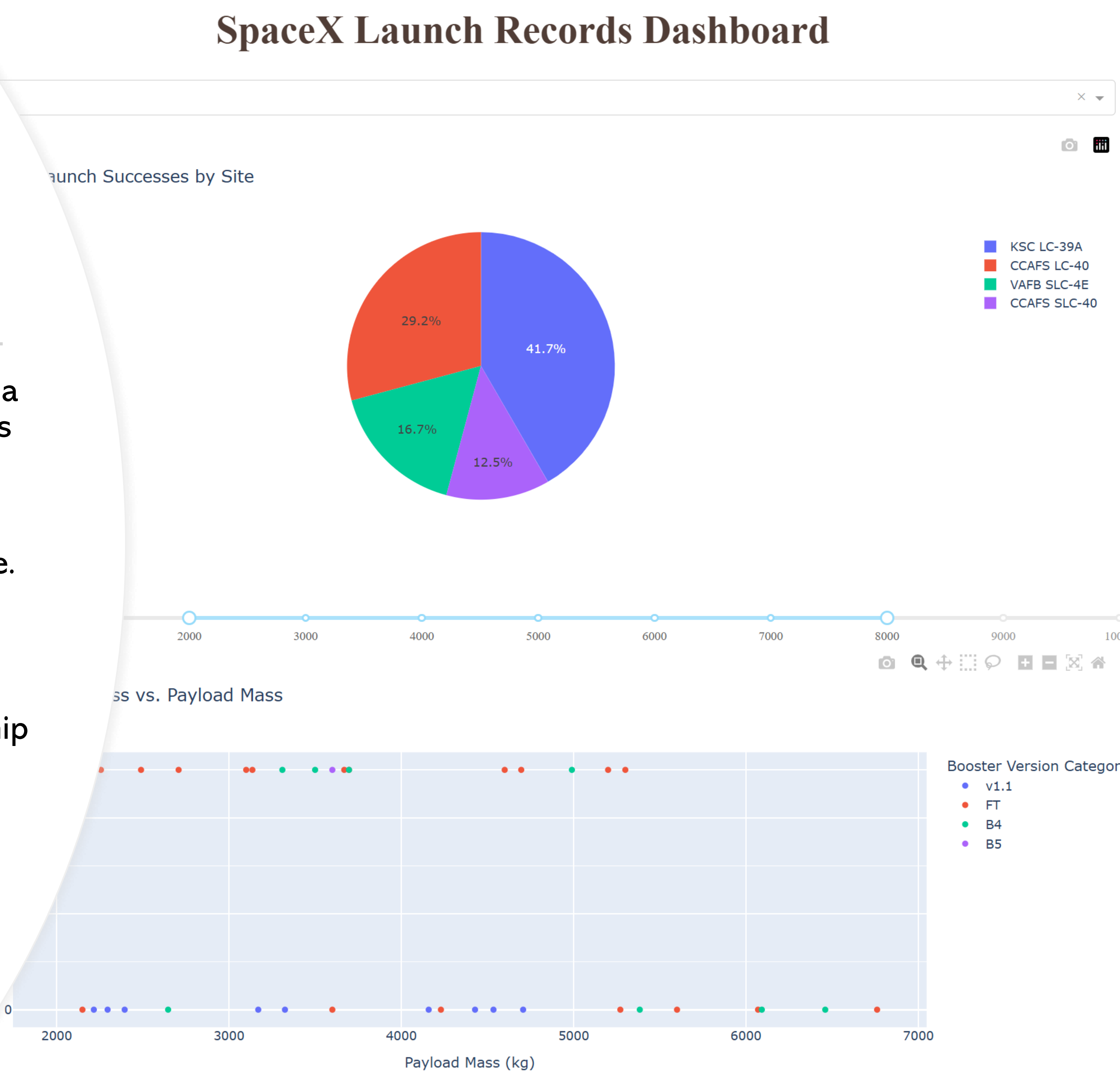
The completed interactive map is available on GitHub for further exploration and peer review: SpaceX Falcon 9 Folium Notebook.



Build a Dashboard with Plotly Dash

- Launch Site Dropdown Menu: Allows users to select a specific SpaceX launch site or view data from all sites collectively.
- Success Pie Chart: Visualizes the proportion of successful versus failed launches for the selected site.
- Payload Range Slider: Lets users filter and analyze launches based on payload mass.
- Success-Payload Scatter Plot: Displays the relationship between payload mass and launch outcomes, with points color-coded by Booster Version.

- Access the completed dashboard and code: [SpaceX Falcon 9 Plotly Dash Notebook](#).



SpaceX Launch Records Dashboard

Predictive Analysis (Classification)

- Data Collection: Gathered and preprocessed data for training and testing.
- Feature Selection: Identified relevant features and engineered new features to enhance model performance.
- Model Building: Implemented various classification algorithms, including KNN, Decision Tree, SVM, and Logistic Regression.
- Model Evaluation: Used metrics like accuracy, confusion matrix, and cross-validation to assess model performance.
- Model Improvement: Tuned hyperparameters and optimized models to enhance accuracy and reduce misclassifications.

For the full Predictive Analysis, visit: [SpaceX Falcon 9 Predictive Analysis Notebook]

Results

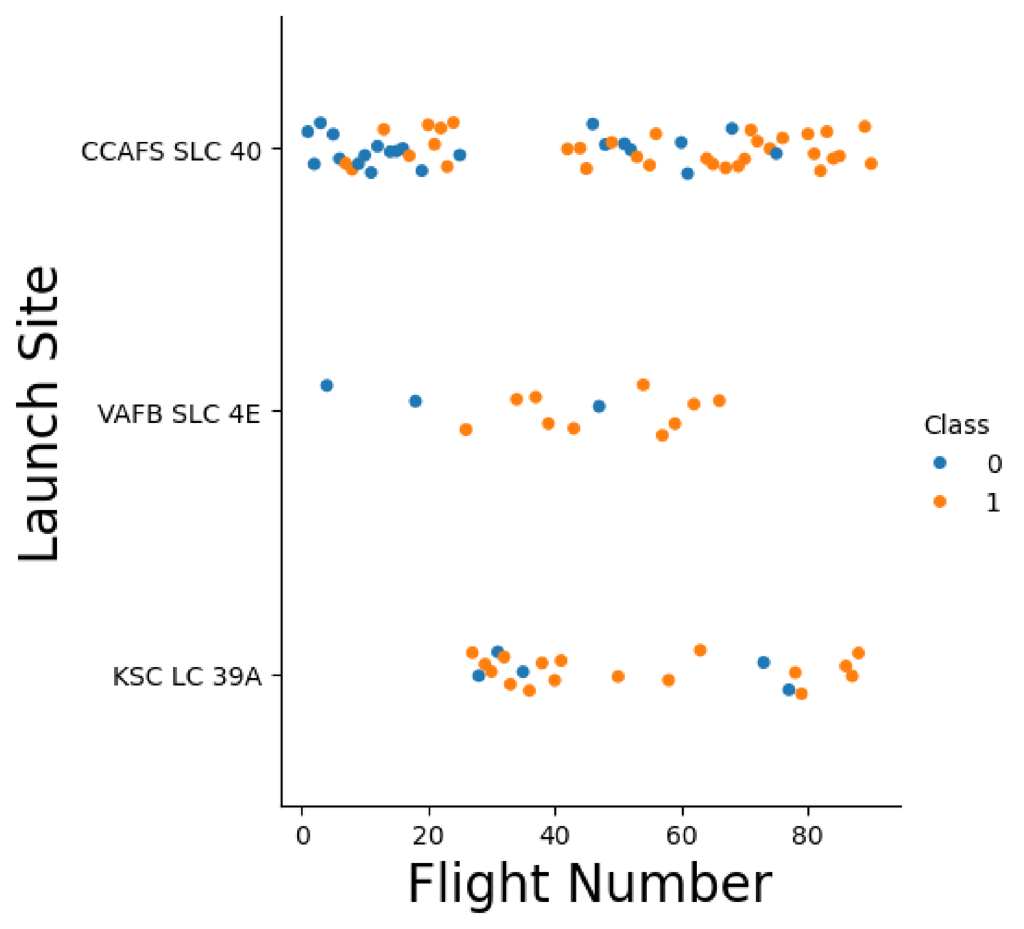
- Exploratory Data Analysis
  - Payload Mass Impact: Lighter payloads generally have higher landing success rates.
  - Launch Site Variation: Success rates differ by launch site, with Cape Canaveral showing higher success.
  - Orbit Type Analysis: Certain orbits (e.g., Low Earth Orbit) have higher success rates.
  - Trend Over Time: Improved success rates observed over the years.
- Interactive Analytics Demo: Map with launch sites and key geographic features, and visualizations of payload mass vs. success rates.
- Predictive Analysis: Achieved 87.50% accuracy.

Section 2

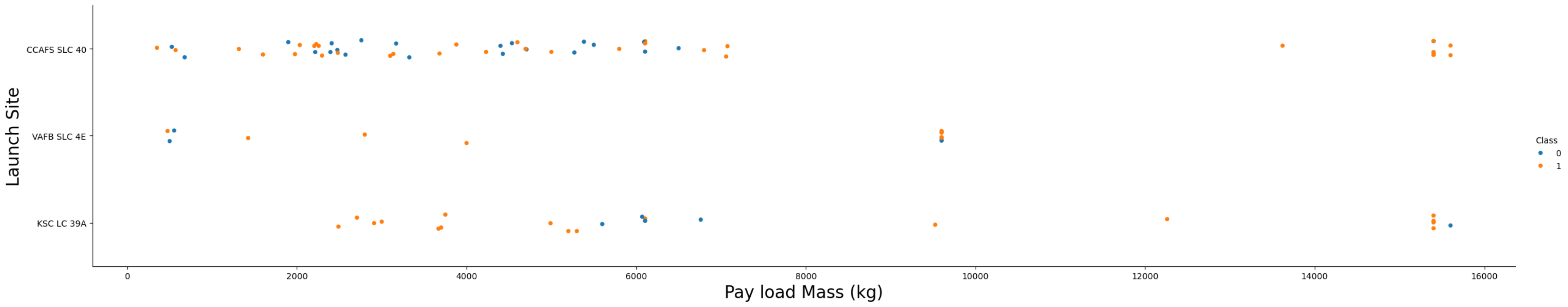
Insights drawn from EDA

Flight Number vs. Launch Site

- Flight Number is plotted on the x-axis, representing the sequence of flights.
- Launch Site is on the y-axis, with each point representing a specific launch site for the corresponding flight number.
- The plot shows how flights are distributed across different launch sites.



Payload vs. Launch Site



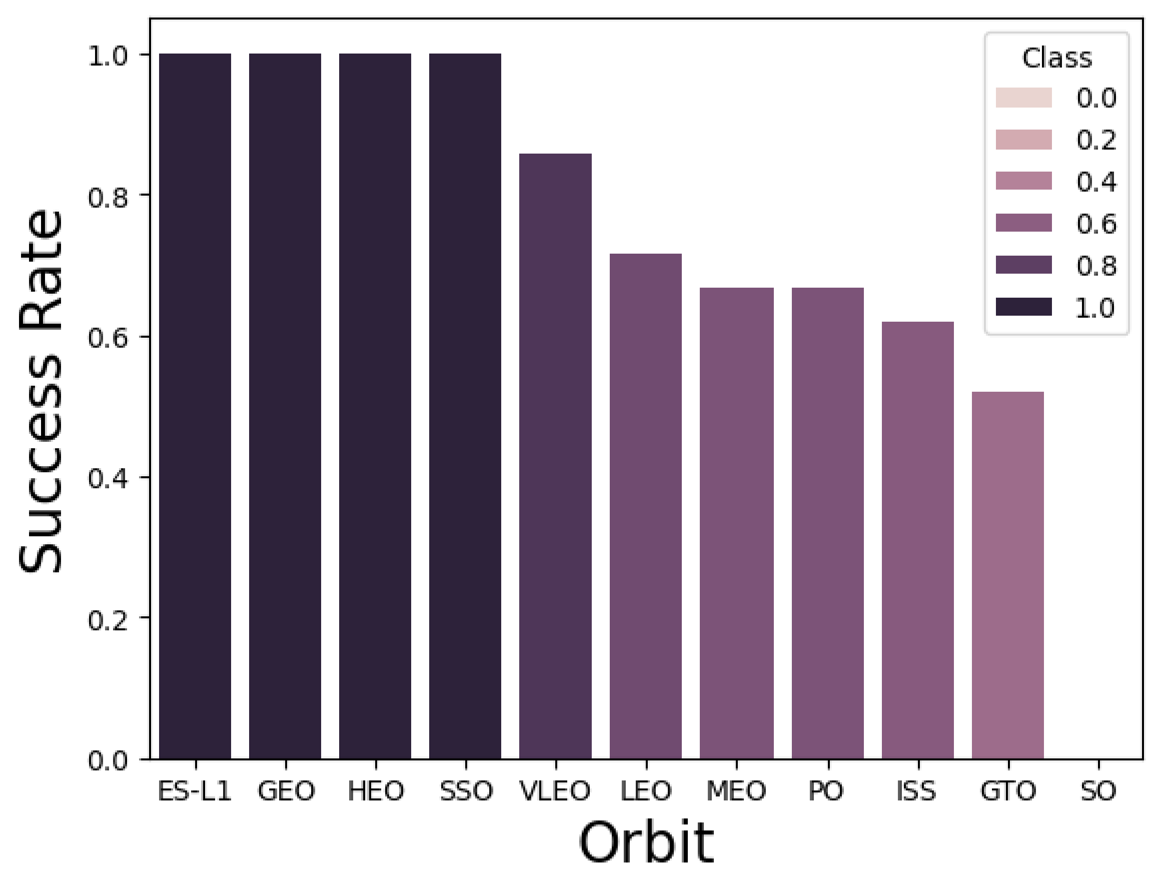
- Payload (kg) is plotted on the x-axis, representing the mass of the payloads for different flights.

- Launch Site is on the y-axis, showing where each flight with its corresponding payload was launched.

- Each point represents a specific flight, with its payload mass and the launch site used.

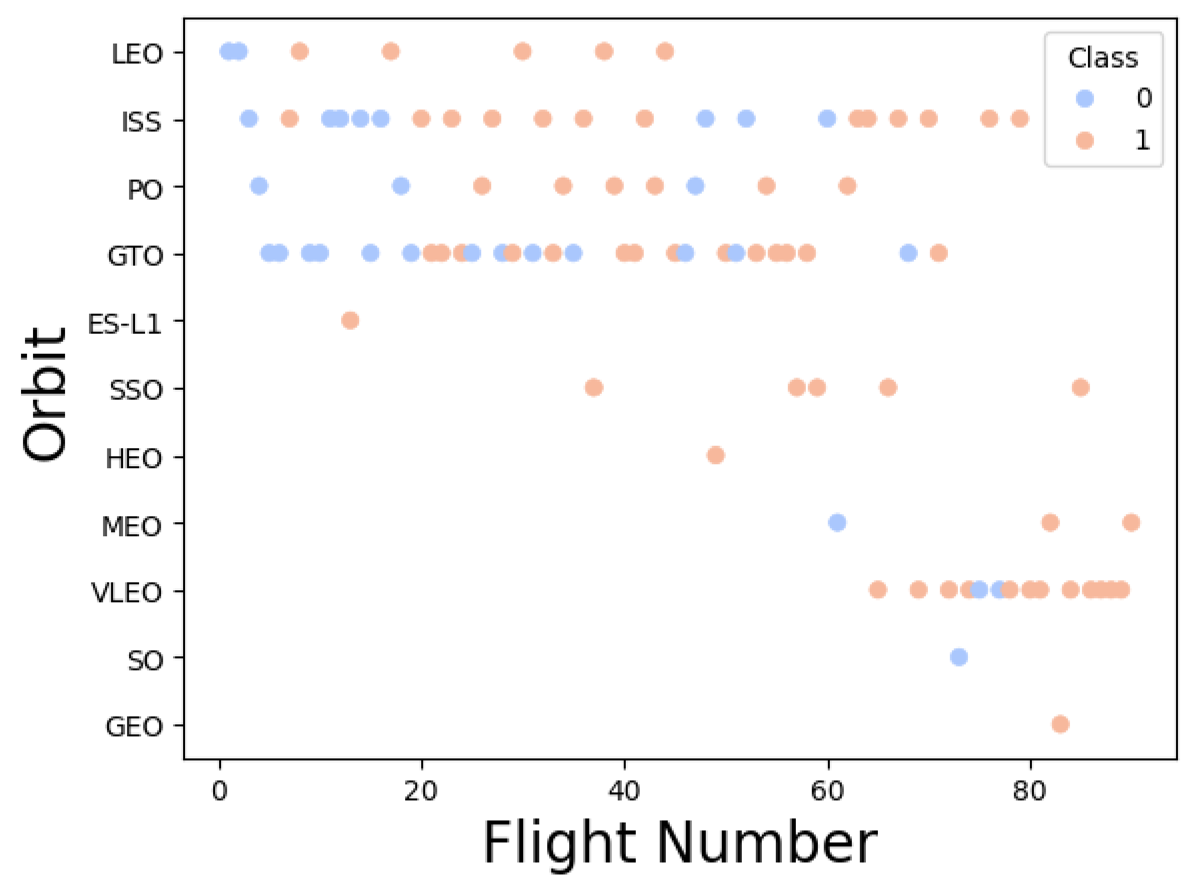
Success Rate vs. Orbit Type

- Orbit Type is plotted on the x-axis, representing different types of orbits (e.g., LEO, GTO, MEO, etc.).
- Success Rate (%) is on the y-axis, representing the percentage of successful missions for each orbit type.
- Each bar shows the success rate for a particular orbit type, calculated as the ratio of successful missions to the total number of missions.



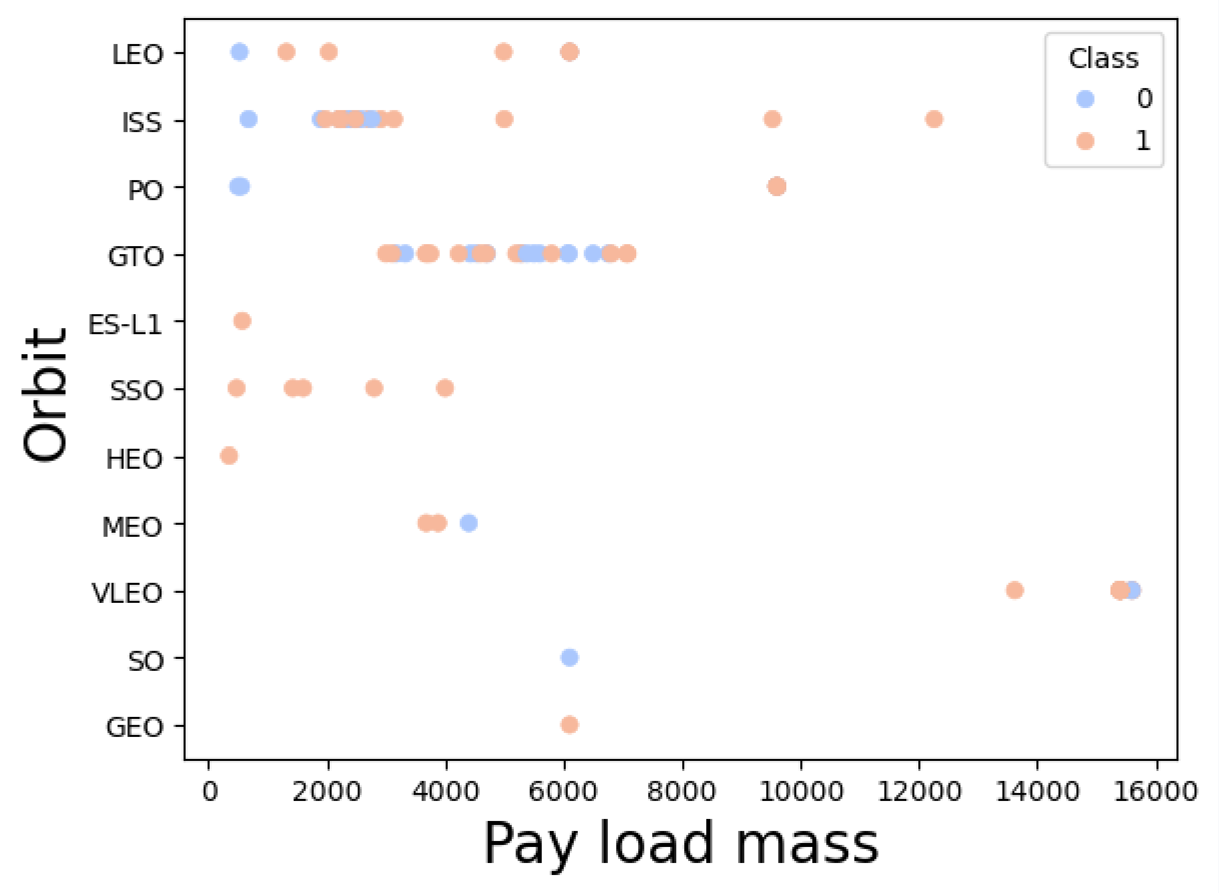
Flight Number vs. Orbit Type

- Flight Number would be plotted on the x-axis, representing the sequence of flights (e.g., 1, 2, 3, ...).
- Orbit Type would be displayed on the y-axis, where each point represents a specific flight with its corresponding orbit type (e.g., LEO, GTO, MEO, etc.).
- Each point on the scatter plot represents a flight and the orbit type it was assigned to.



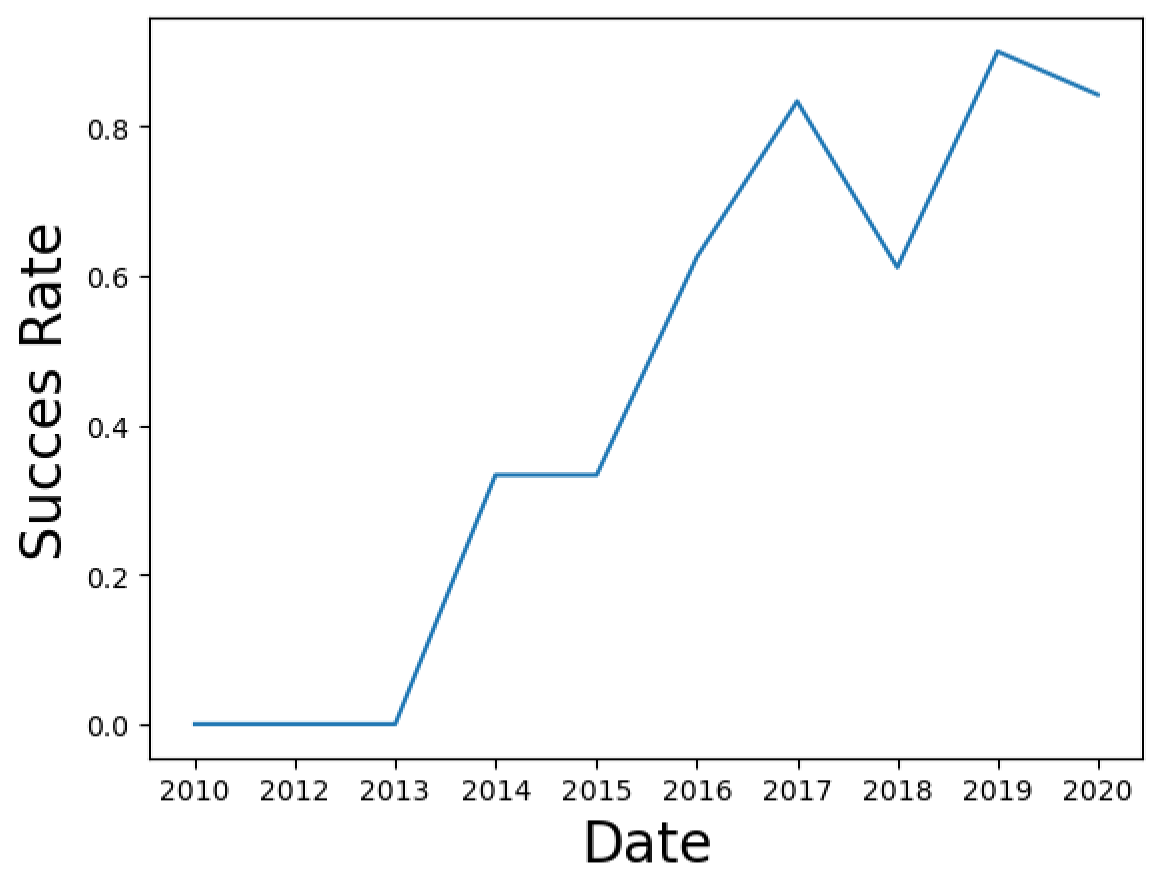
Payload vs. Orbit Type

- Payload (kg) would be plotted on the x-axis, representing the mass of the payloads for various flights.
- Orbit Type would be displayed on the y-axis, where each point represents a specific payload mass and the corresponding orbit type (e.g., LEO, GTO, MEO, etc.).
- Each point on the scatter plot represents a flight, with its payload and orbit type.



Launch Success Yearly Trend

- Year would be plotted on the x-axis, representing different years over a certain period.
- Average Success Rate (%) would be plotted on the y-axis, showing the average success rate of missions for each year.
- The line on the chart would connect data points representing the yearly average success rates, allowing you to see trends over time.

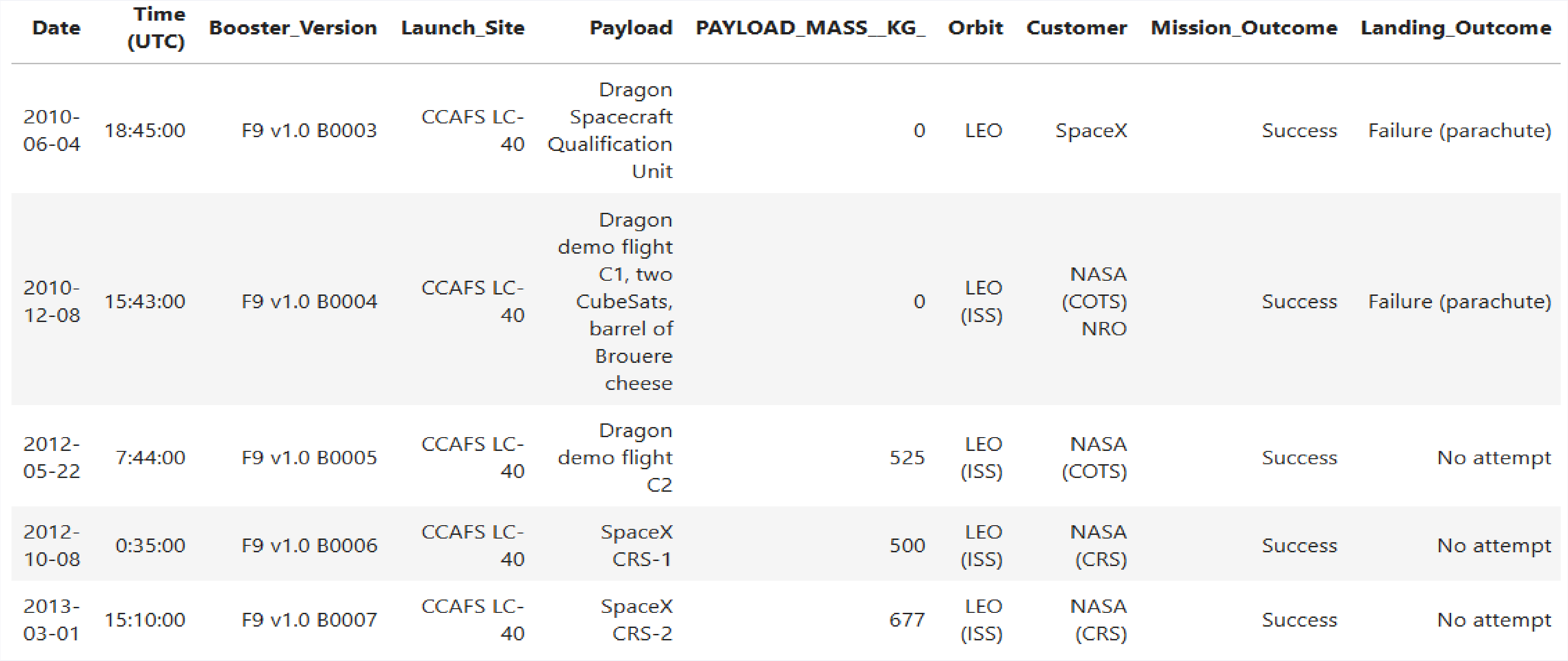


All Launch Site Names

- CCAFS LC-40: Cape Canaveral Air Force Station Launch Complex 40
- VAFB SLC-4E: Vandenberg Air Force Base Space Launch Complex 4E
- KSC LC-39A: Kennedy Space Center Launch Complex 39A
- CCAFS SLC-40: Cape Canaveral Air Force Station Space Launch Complex 40

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

Launch Site Names Begin with 'CCA'



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

- Launch_Site: All records have launch sites beginning with CCA, specifically CCAFS LC-40, which is Cape Canaveral Air Force Station Launch Complex 40.
- Other Columns: Include details such as the date of the launch, the booster version used, the payload, its mass, the orbit, the customer, mission outcome, and landing outcome.

Total Payload Mass

The total payload mass carried by boosters for missions where the customer is NASA (CRS) is 45,596 kilograms.

45,596 kilograms

Average Payload Mass by F9 v1.1

The average payload mass carried by the F9 v1.1 booster version is approximately 2,534.67 kilograms.

2,534.67 kilograms

First Successful Ground Landing Date

The earliest date on which a successful landing on the ground pad occurred is December 22, 2015.

December 22, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

The boosters that successfully landed on a drone ship and had a payload mass between 4000 kg and 6000 kg are:

- F9 FT B1032.1
- F9 B4 B1040.1
- F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

SUCCESS: THERE WERE 61 SUCCESSFUL LANDING OUTCOMES.

FAILURE: THERE WERE 10 FAILURE LANDING OUTCOMES.

Boosters Carried Maximum Payload

- The boosters listed are those that have carried the maximum payload mass recorded in the dataset.
- Multiple booster versions are shown because they all share the maximum payload mass.

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

Date	Mount	Landing_Outcome	Booster_Version	Launch_Site
2015-01-10	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The query results show the records from 2015 where the landing outcome was a failure on the drone ship.
- For each record, the date of the event, month, landing outcome, booster version, and launch site are provided.

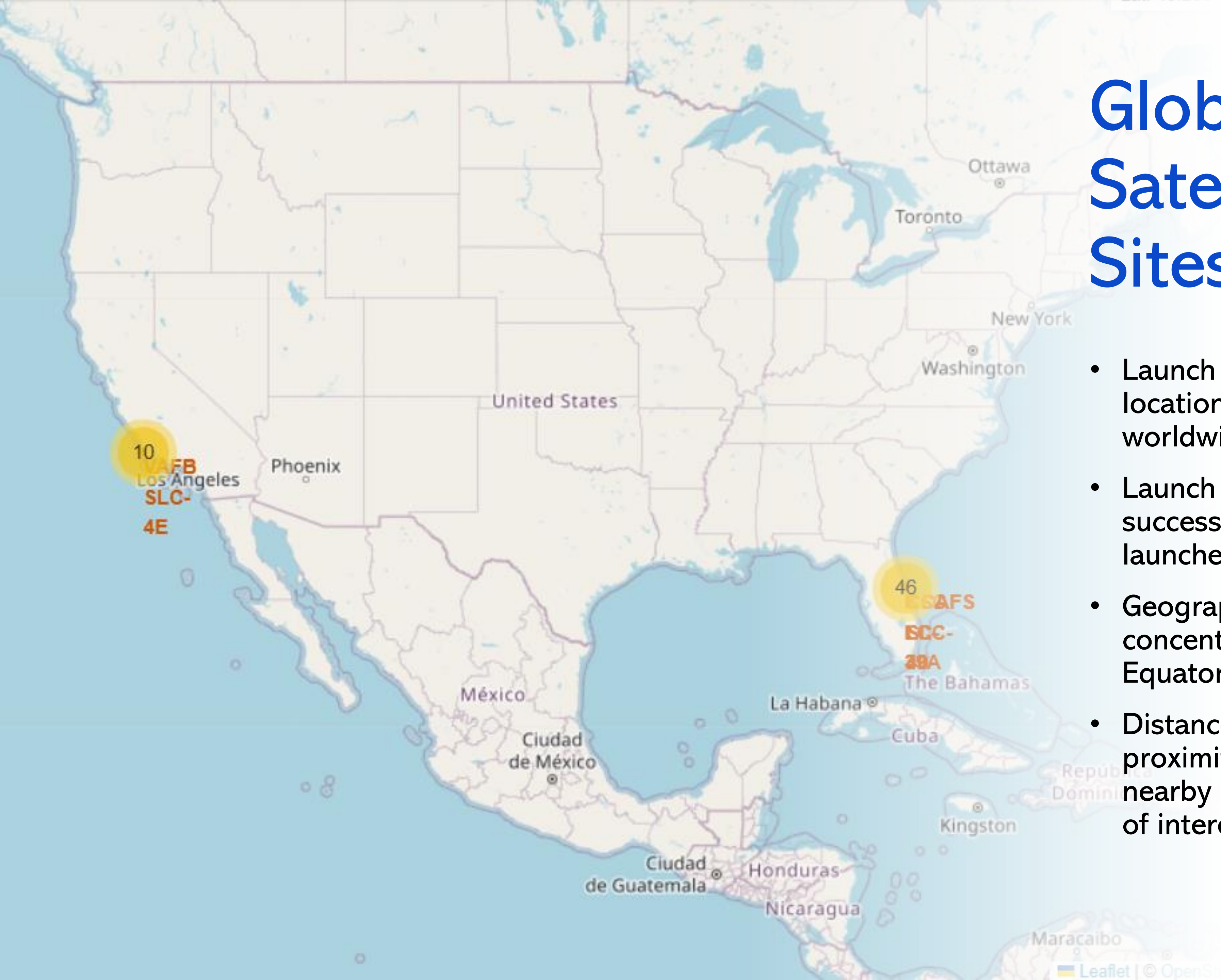
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The result shows the number of occurrences for each type of landing outcome within the specified date range.
- No attempt has the highest count, indicating that there were many missions where no landing attempt was made.
- Outcomes like Success (drone ship) and Failure (drone ship) have fewer occurrences compared to No attempt, showing these outcomes are less frequent.
- The results provide a clear picture of how often different landing outcomes occurred during the specified period.

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

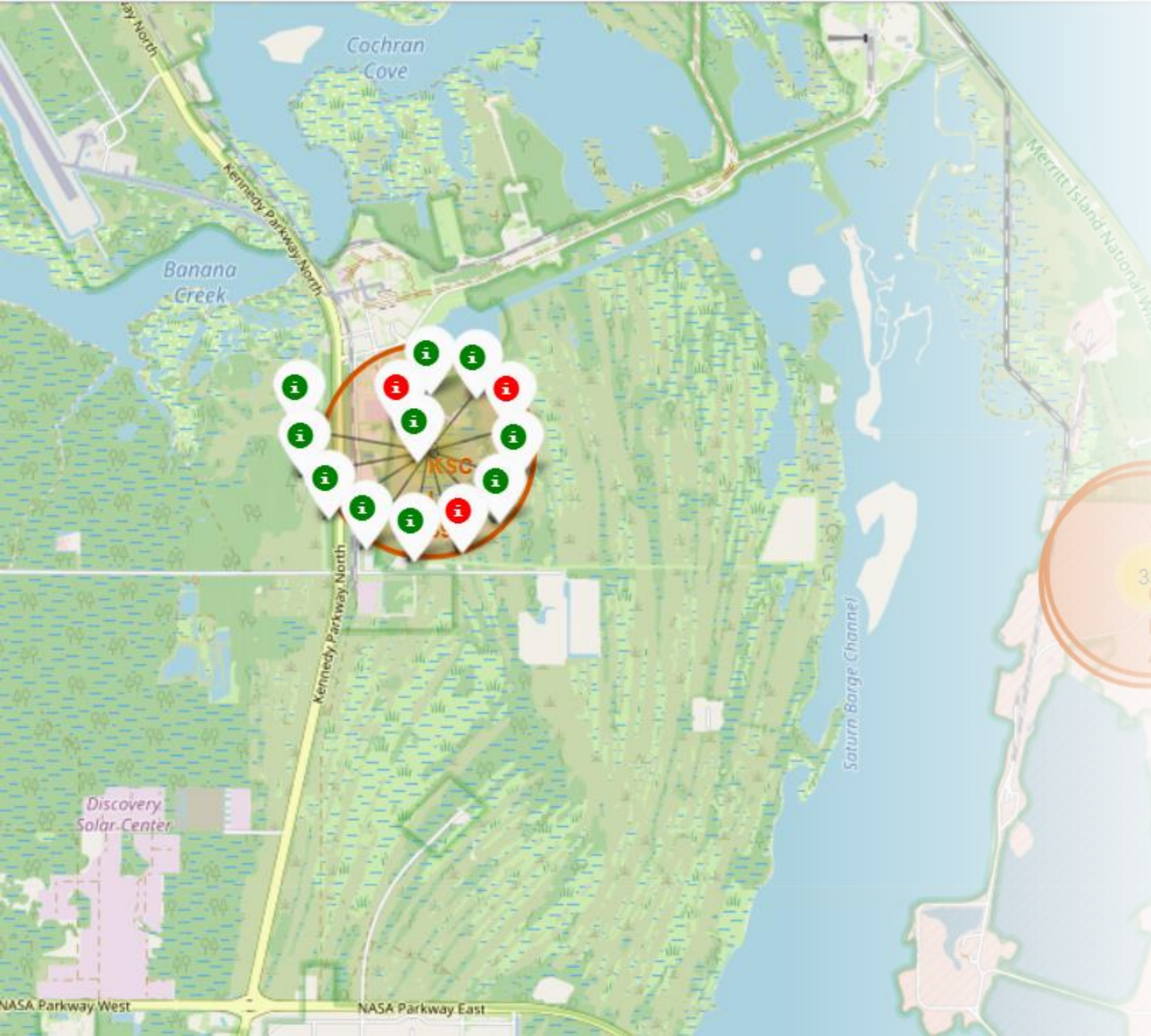
Section 3

Launch Sites Proximities Analysis



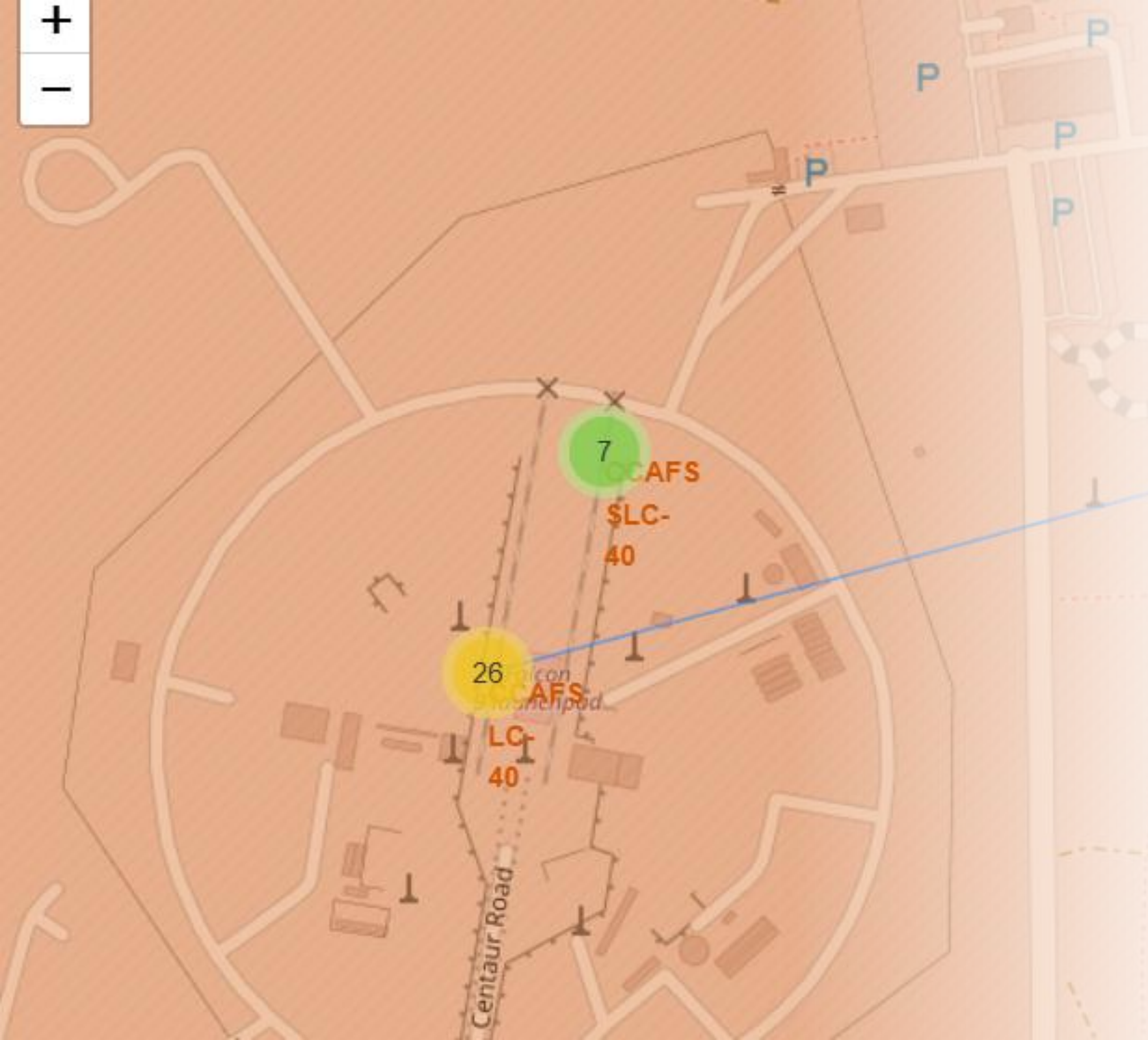
Global Map of Satellite Launch Sites

- Launch Site Markers: Identifies the locations of satellite launch sites worldwide with markers.
- Launch Outcomes: Displays successful (green) and failed (red) launches using marker clusters.
- Geographic Patterns: Reveals concentration of sites near the Equator and coastal areas.
- Distance Analysis: Measures proximity from launch sites to nearby coastlines and other points of interest.



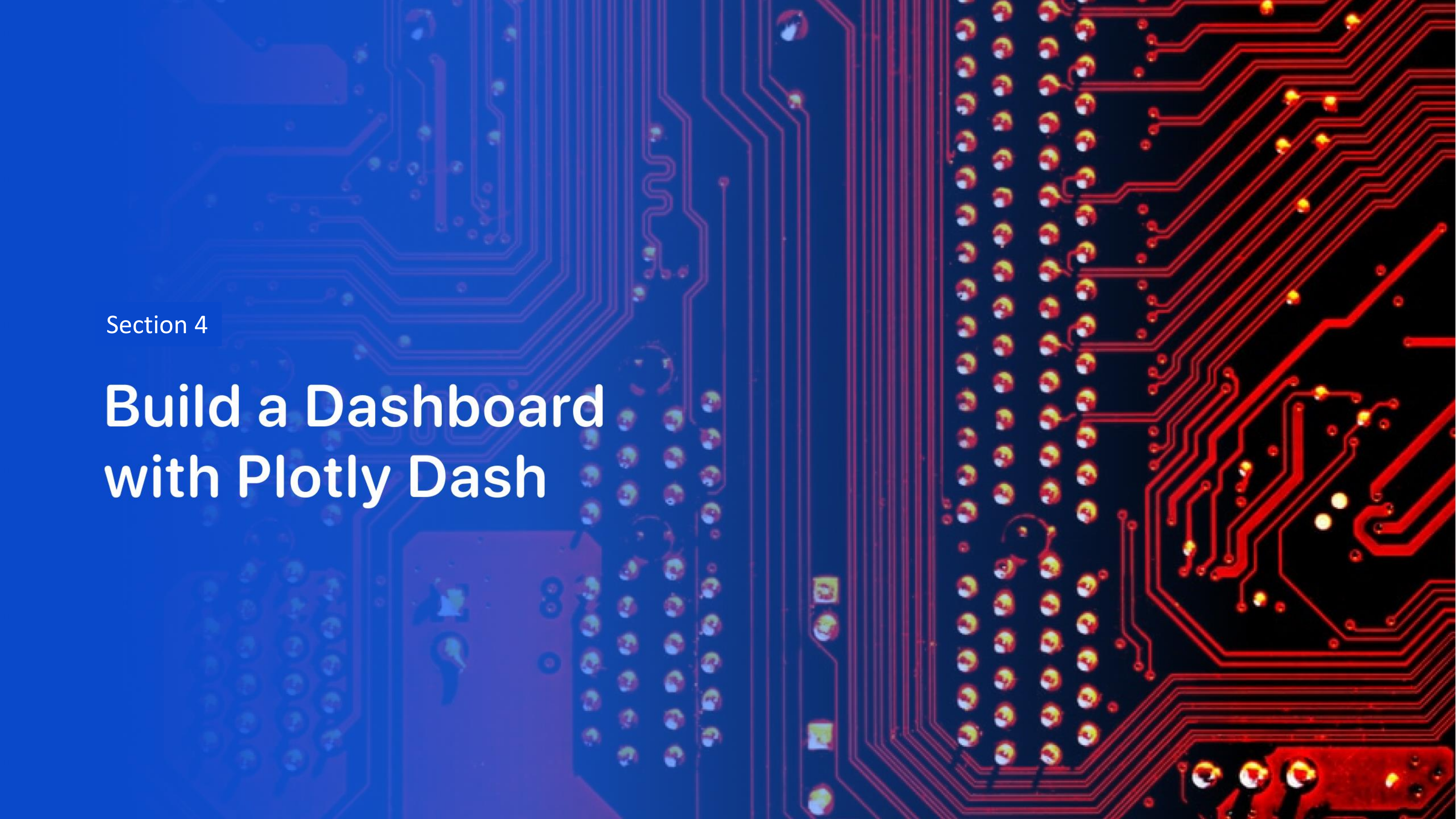
Color-Coded Launch Outcomes on Satellite Launch Sites Map

- Color-Labeled Markers:
  - Green Markers: Represent successful launches.
  - Red Markers: Indicate failed launches.
- Marker Clustering: Uses clusters to manage overlapping markers and provide a clearer view of launch outcomes.
- Launch Site Analysis: The map shows which sites have higher success rates based on the color of the markers.
- Geographical Context: Visualizes launch outcomes in relation to the geographic distribution of sites.



Proximity Analysis of Selected Launch Site

- Highlighted Launch Site: Clearly marked on the map with its precise location.
- Nearby Features:
  - Railway, Highway, Coastline: Identified and marked to show proximity.
- Distance Measurement:
  - Distance Line: Displays the measured distance from the launch site to the nearest coastline.
  - Distance Label: Shows the distance in kilometers.
- Contextual Insight: Demonstrates the spatial relationship between the launch site and surrounding infrastructure or natural features.

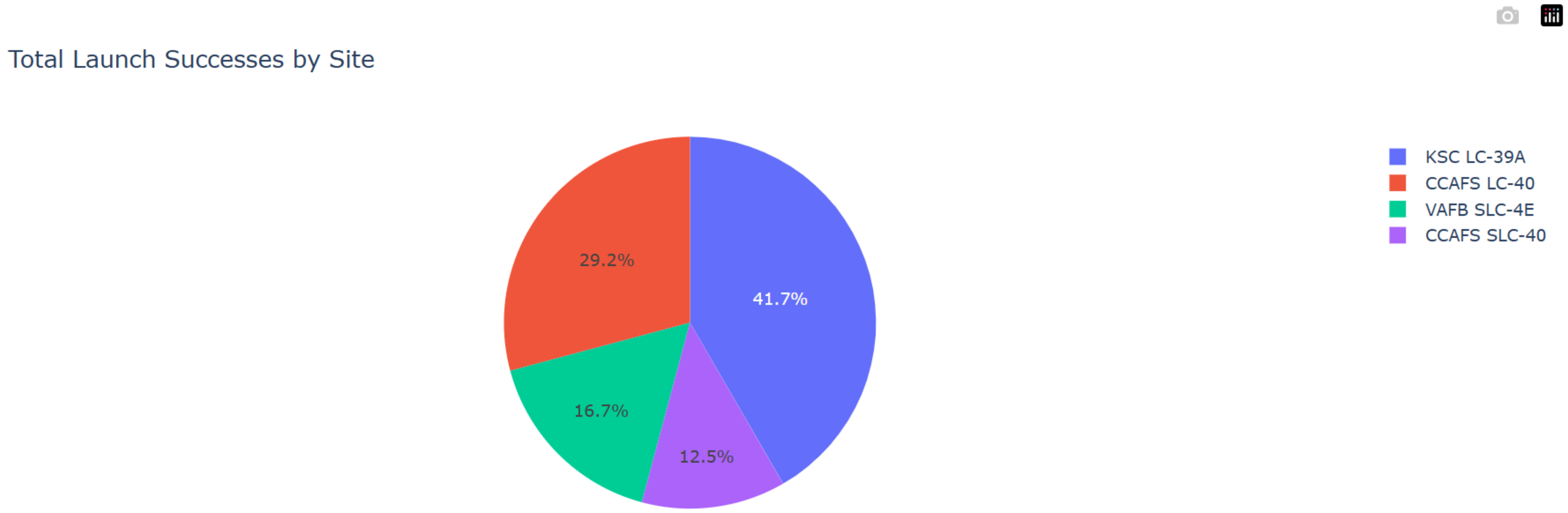


Section 4

Build a Dashboard with Plotly Dash

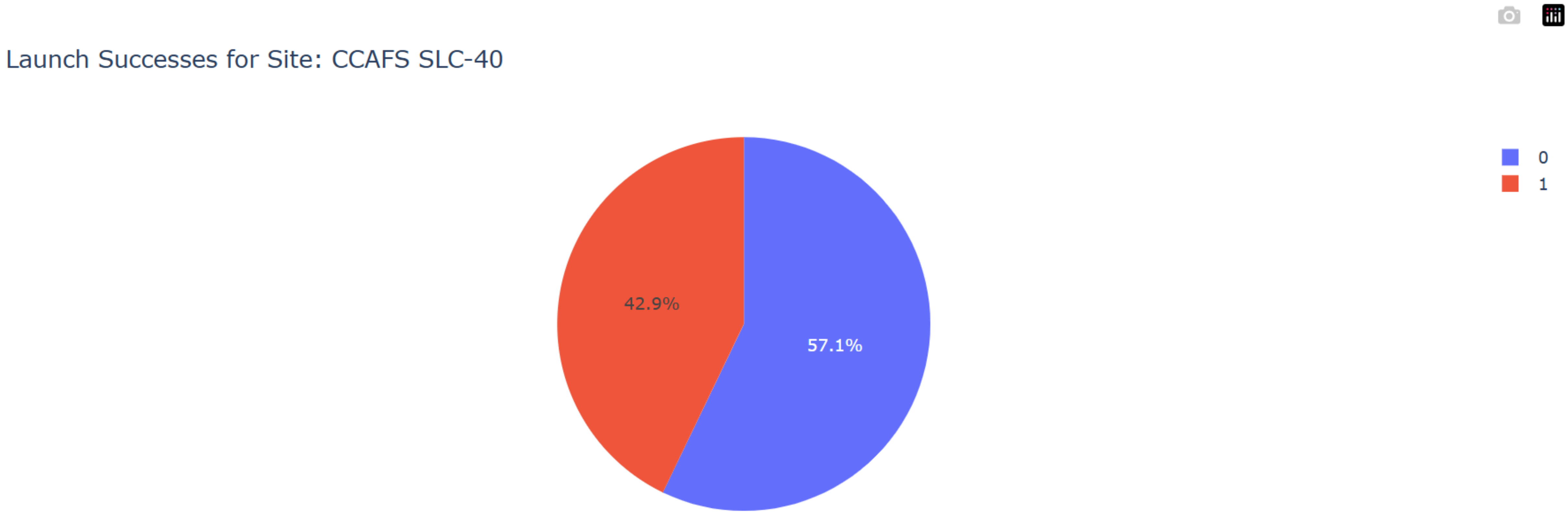
Launch Success Count for All Sites

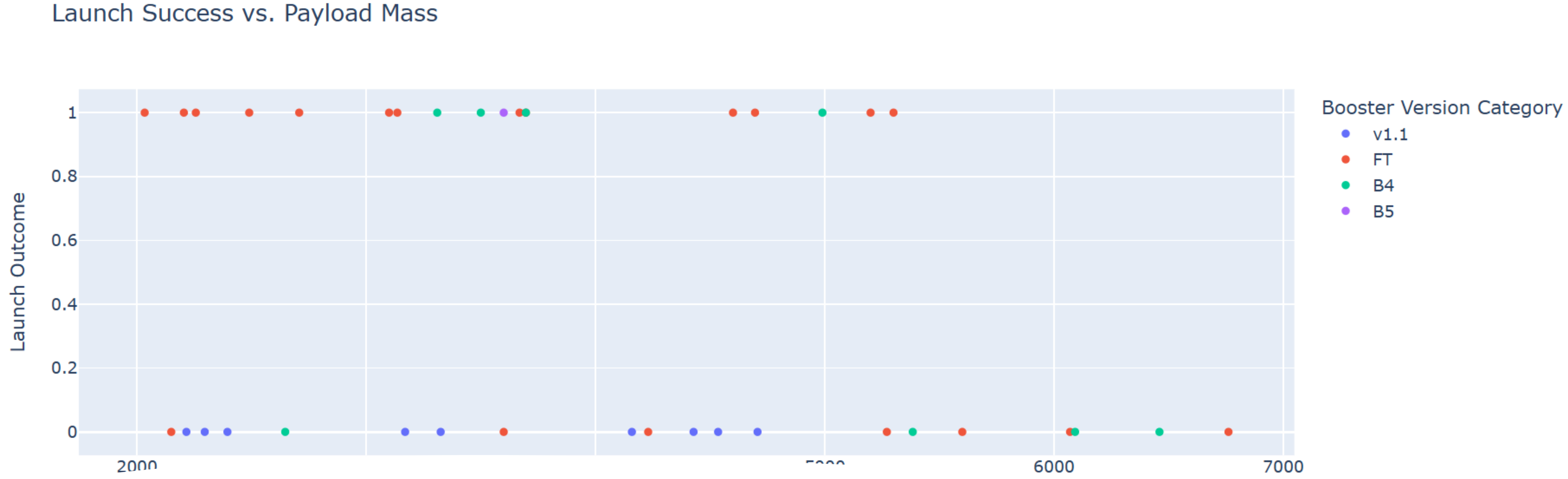
- The chart shows the overall success and failure rates aggregated across all sites.
- The proportion of successful launches compared to failed ones can be observed here.
- This visualization helps in understanding the general success rate of SpaceX missions irrespective of individual sites.



Launch Success Ratio for the Top Site

- The chart highlights the success rate of the site with the most favorable launch outcomes.
- This visualization provides a detailed breakdown of successful vs. failed launches for this site alone, offering insights into its performance.
- Helps identify the best-performing site in terms of launch success rate.





Payload vs. Launch Outcome Scatter Plot

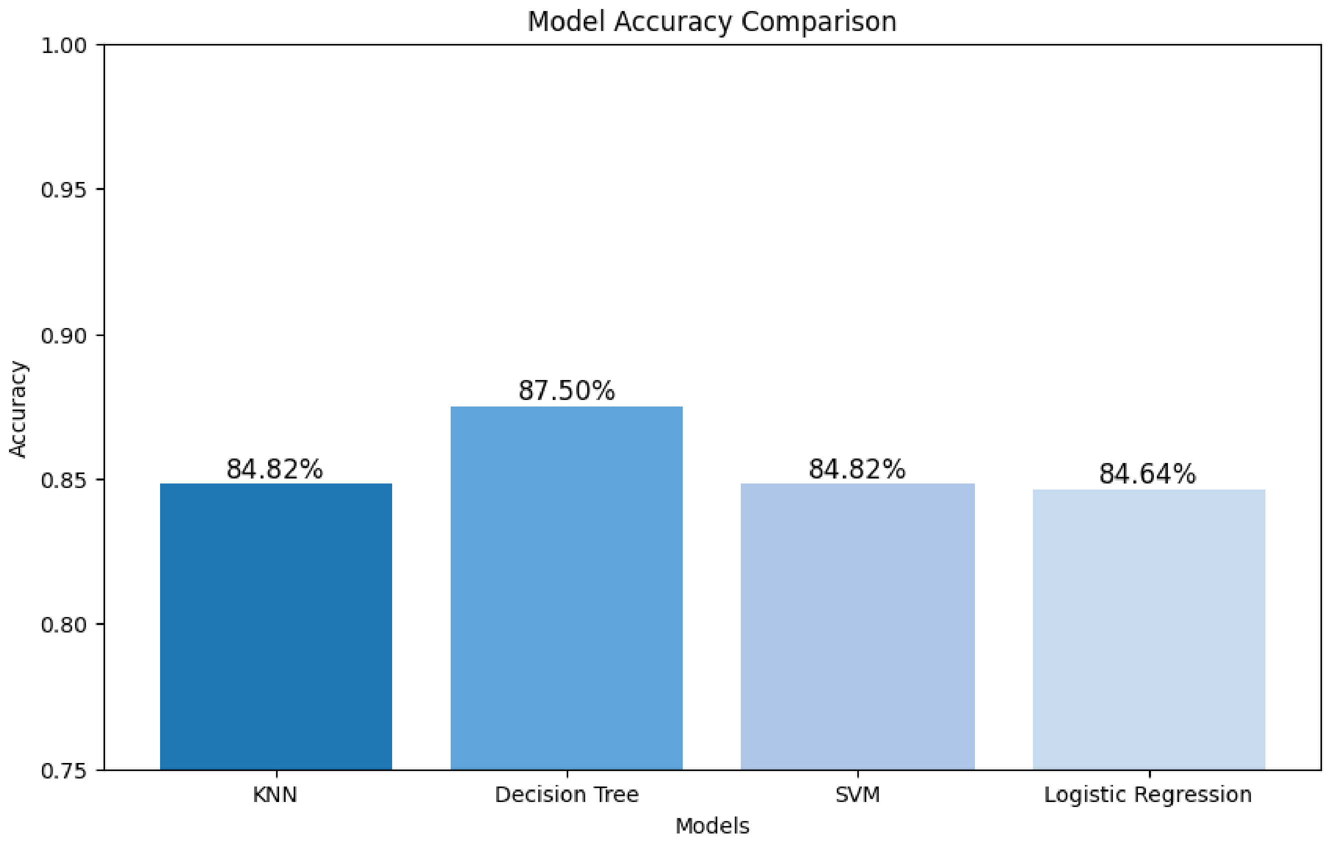
- Payload Ranges: By adjusting the range slider, you can observe how payload mass correlates with launch outcomes across different ranges.
- Booster Version Analysis: The color-coding by Booster Version allows analysis of how different versions impact launch success rates.
- Patterns: Identify payload ranges with the highest and lowest success rates, and analyze if certain booster versions correlate with higher or lower success rates.

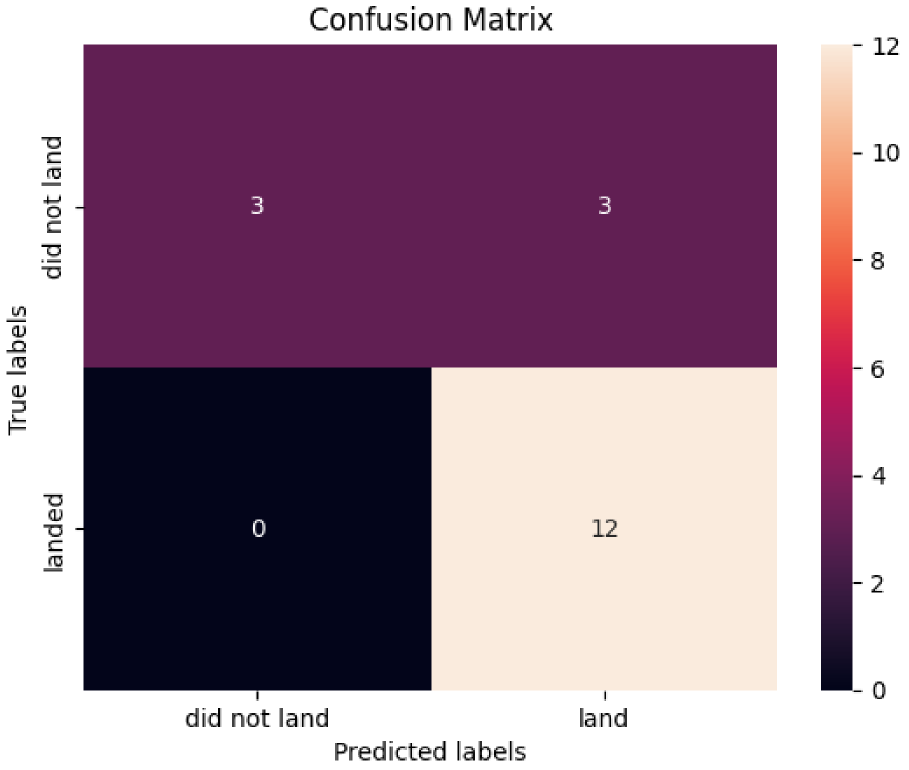
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree model has the highest accuracy at 87.50%.
- The other models (KNN, SVM, Logistic Regression) have similar accuracies around 84.64% and 84.82%.





Confusion Matrix

- High True Positive Rate: The model correctly identifies positive samples with a TP of 3.
- Moderate False Positive Rate: The model incorrectly labels some negative samples as positive (FP = 3).
- No False Negatives: The model perfectly identifies all positive samples without missing any (FN = 0).
- High True Negative Rate: The model correctly identifies most negative samples (TN = 12).
- This confusion matrix suggests the model performs well in identifying positives with no missed positives, but there is some misclassification of negatives.

Conclusions

- The Decision Tree model demonstrates the highest classification accuracy among all tested models, achieving an accuracy of 87.50%. This indicates that the model performs well in correctly predicting outcomes.
- The confusion matrix highlights that the model is effective in identifying positive samples without missing any, though it does misclassify a few negative samples as positive.
- The model excels in achieving high accuracy and accurately identifying all positive cases (0 FN). The high number of true negatives (12) also reflects its strong performance in recognizing negative cases.
- Despite the high accuracy, the model has a moderate rate of false positives (3), where negative samples are incorrectly predicted as positive. Addressing this issue could enhance the model's reliability and precision.

Appendix

GitHub: IBM-Data-Science/Capstone · dessanriv

Thank you!