IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

This project leverages historical SpaceX Falcon 9 launch data to identify factors that influence the success of first-stage landings.

Through API data collection, web scraping, data wrangling, exploratory data analysis (EDA), interactive mapping, and predictive modeling, we built classification models to estimate landing success probability.

Key findings: Launch site and payload mass are strong predictors of landing success. Certain booster versions have consistently higher success rates. Predictive models achieve over 85% accuracy.

These insights can help optimize launch configurations and reduce mission costs.

Introduction

Project Background: SpaceX’s Falcon 9 is a partially reusable rocket designed to significantly reduce the cost of space travel. Understanding the factors influencing first-stage landing success is critical to improving reusability and mission economics.

Problem Statement: Predicting the success of first-stage landings can guide decision-making on launch configurations, payload planning, and booster reuse. Objectives:Collect and prepare SpaceX launch data from multiple sources. Perform exploratory data analysis to uncover patterns.

Build predictive models to estimate landing success probability. Present findings through visual analytics and dashboards.

Impact: Insights from this analysis can help SpaceX and similar companies optimize launch strategies, improve cost efficiency, and enhance mission planning.

Section 1

Methodology

Methodology

This project investigates SpaceX Falcon 9 launch data to identify factors influencing first-stage landing success.

The process included:

- **Data Collection:** Retrieved launch records via the SpaceX REST API and web scraping from Wikipedia.
- **Data Wrangling:** Cleaned, merged, and standardized datasets, handling missing and inconsistent values.
- **Exploratory Data Analysis (EDA):** Used visualizations and SQL queries to uncover trends by launch site, payload, and orbit type.
- **Interactive Analytics:** Built Folium maps to visualize launch sites and outcomes; developed a Plotly Dash dashboard for interactive exploration.
- **Predictive Modeling:** Created, tuned, and evaluated classification models (e.g., Logistic Regression, SVM, Decision Tree, KNN) to predict landing outcomes.

The findings and models can guide launch strategy decisions, improve reusability, and reduce mission costs.

Data Collection

We collected SpaceX Falcon 9 launch data from multiple sources to ensure completeness and accuracy:

- **SpaceX REST API:**
  - Retrieved structured JSON data on launch dates, sites, payloads, rocket versions, and landing outcomes.
  - Automated data retrieval using Python requests library.
- **Web Scraping (Wikipedia):**
  - Extracted additional details such as landing type and mission outcome not fully available via API.
  - Used Python BeautifulSoup to parse HTML tables and clean extracted data.
- **Integration:**
  - Merged API and scraped datasets using unique launch identifiers.
  - Verified data consistency and removed duplicates.

Flowchart (suggested visual):

1. API Call to SpaceX REST API → JSON Data Extraction
2. Web Scraping Wikipedia Falcon 9 Launch Table → HTML Parsing & Cleaning
3. Merge Both Sources → Final Structured Dataset for Analysis

Data Collection – SpaceX API

We collected structured launch data directly from the **SpaceX REST API**:

- **API Endpoint:**
    - Queried https://api.spacexdata.com/v4/launches/past for historical launch records.
- **Process:**
    1. Sent GET request using Python's requests library.
    2. Parsed JSON response to extract launch date, site, payload mass, booster version, and landing outcome.
    3. Stored results in a Pandas DataFrame for further processing.
    4. Saved cleaned dataset locally in CSV format.
- **Benefits:**
    - Provides official, structured, and up-to-date launch data.
    - Reduces manual data entry errors.

API Call → JSON Response → Data Parsing → DataFrame Creation → CSV Export

Data Collection - Scraping

Data Wrangling

After collecting the data from the SpaceX API and Wikipedia, we processed it to ensure quality and consistency:

- **Data Cleaning:**
  - Removed duplicate records and irrelevant fields.
  - Standardized categorical values (e.g., booster version naming).
  - Converted date fields to datetime format.
  - Handled missing values with imputation or exclusion depending on importance.
- **Feature Engineering:**
  - Extracted year, month, and day from launch dates.
  - Created binary target variable (LandingSuccess) from landing outcome descriptions.
  - Calculated payload ranges for grouping analysis.
- **Merging:**
  - Joined API and scraped datasets using launch ID or date + mission name as keys.
  - Verified merged dataset for completeness.
- **Output:**
  - Final cleaned dataset saved as CSV for use in EDA, visualization, and modeling.

Raw API Data + Raw Scraped Data → Cleaning → Standardization → Feature Engineering → Dataset Merge → Final Cleaned Dataset

Data Wrangling

We processed the collected SpaceX data to prepare it for analysis and modeling:

- **Data Cleaning:**
  - Removed duplicates and irrelevant fields.
  - Standardized categorical variables (e.g., booster version naming).
  - Converted date/time fields to datetime objects.
  - Addressed missing values through imputation or removal depending on the variable's importance.
- **Feature Engineering:**
  - Extracted year, month, and day from launch dates.
  - Created binary target variable LandingSuccess from landing outcome text.
  - Calculated payload range bins for grouped analysis.
- **Data Integration:**
  - Merged API and web-scraped data using unique launch IDs or combinations of date and mission name.
  - Validated merged dataset to ensure completeness and accuracy.

EDA with Data Visualization

We performed exploratory data analysis to identify patterns, trends, and relationships in the SpaceX dataset.
The following visualizations were created:

- **Scatter Plots** –
  - *Flight Number vs. Launch Site*: To check if launch experience at a site impacts success rates.
  - *Payload Mass vs. Launch Site*: To see if payload weight influences landing success by location.
- **Bar Charts** –
  - *Success Rate by Orbit Type*: To compare performance across different mission orbits.
- **Line Charts** –
  - *Yearly Launch Success Trend*: To analyze how success rates improved over time.
- **Scatter + Color Coding** –
  - *Payload vs. Orbit Type*: To examine the relationship between payload size and orbit with success color indicators.

Why these charts were used:

- Scatter plots reveal correlations between numerical variables and categorical factors like launch site.
- Bar charts highlight comparative success rates across categories.
- Line charts show time-based trends in performance.
- Combined scatter/color visuals help identify multi-variable relationships.

EDA with SQL

We used SQL queries to extract insights and validate patterns found in the visual analysis:

- Retrieved **list of unique launch site names** from the dataset.
- Filtered records for **launch sites beginning with** CCA.
- Calculated **total payload mass** carried by boosters from NASA.
- Computed **average payload mass** for booster version *F9 v1.1*.
- Identified the **date of the first successful ground landing**.
- Listed **booster names** with successful drone ship landings and payload mass between 4,000–6,000 kg.
- Counted the **total number of successful and failed missions**.
- Found **boosters with maximum payload mass** carried.
- Listed **failed drone ship landings in 2015** with booster version and launch site.
- Ranked **landing outcomes** between 2010-06-04 and 2017-03-20 in descending order of occurrence.

Build an Interactive Map with Folium

We created an interactive map to visualize SpaceX launch site locations and mission outcomes:

•**Markers:**

- Plotted each launch site with a marker at its GPS coordinates.
- Added popup information including site name, total launches, and success rate.

•**Color-Coded Circles:**

- Used circle markers to indicate launch outcomes (e.g., green for success, red for failure).
- Circle size varied slightly based on payload mass to provide visual weight.

•**Proximity Lines:**

- Drew lines from launch sites to nearby infrastructure (railway, highway, coastline).
- Labeled distances to show logistical relevance.

Why these objects were added:

•**Markers** help identify where each launch site is located globally.

•**Color coding** makes it easy to see performance trends by location at a glance.

•**Proximity lines** highlight operational advantages or constraints related to transportation and recovery.

Build a Dashboard with Plotly Dash

Build a Dashboard with Plotly Dash

We developed an interactive dashboard to explore launch outcomes dynamically:

•**Pie Charts:**

- *Launch Success Count for All Sites* – Displays overall distribution of successes vs. failures.
- *Launch Success Ratio for Selected Site* – Shows proportion of successful launches for a chosen site.

•**Scatter Plot:**

- *Payload Mass vs. Launch Outcome* – Plots each mission's payload against success/failure, with points colored by booster version.

•**Interactive Controls:**

- **Dropdown Menu** – Allows selection of a specific launch site to filter charts.
- **Payload Range Slider** – Filters scatter plot to show only missions within selected payload mass range.

Why these elements were added:

•Pie charts provide quick, high-level insights into overall and site-specific performance.

•Scatter plots reveal relationships between payload mass, booster type, and outcome.

•Interactivity lets users drill down into specific launch sites and payload categories without needing to re-run analyses.

Predictive Analysis (Classification)

Data Preparation:

- Selected relevant features (payload mass, orbit type, launch site, booster version, etc.).
- Encoded categorical variables using one-hot encoding.
- Split data into **training (80%)** and **test (20%)** sets.
- Scaled numerical features for models sensitive to feature magnitude.

Models Tested:

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

Evaluation & Improvement:

- Used accuracy score, confusion matrix, and cross-validation to assess performance.
- Tuned hyperparameters via **GridSearchCV** for each model.
- Compared tuned models on the test set to select the best performer.

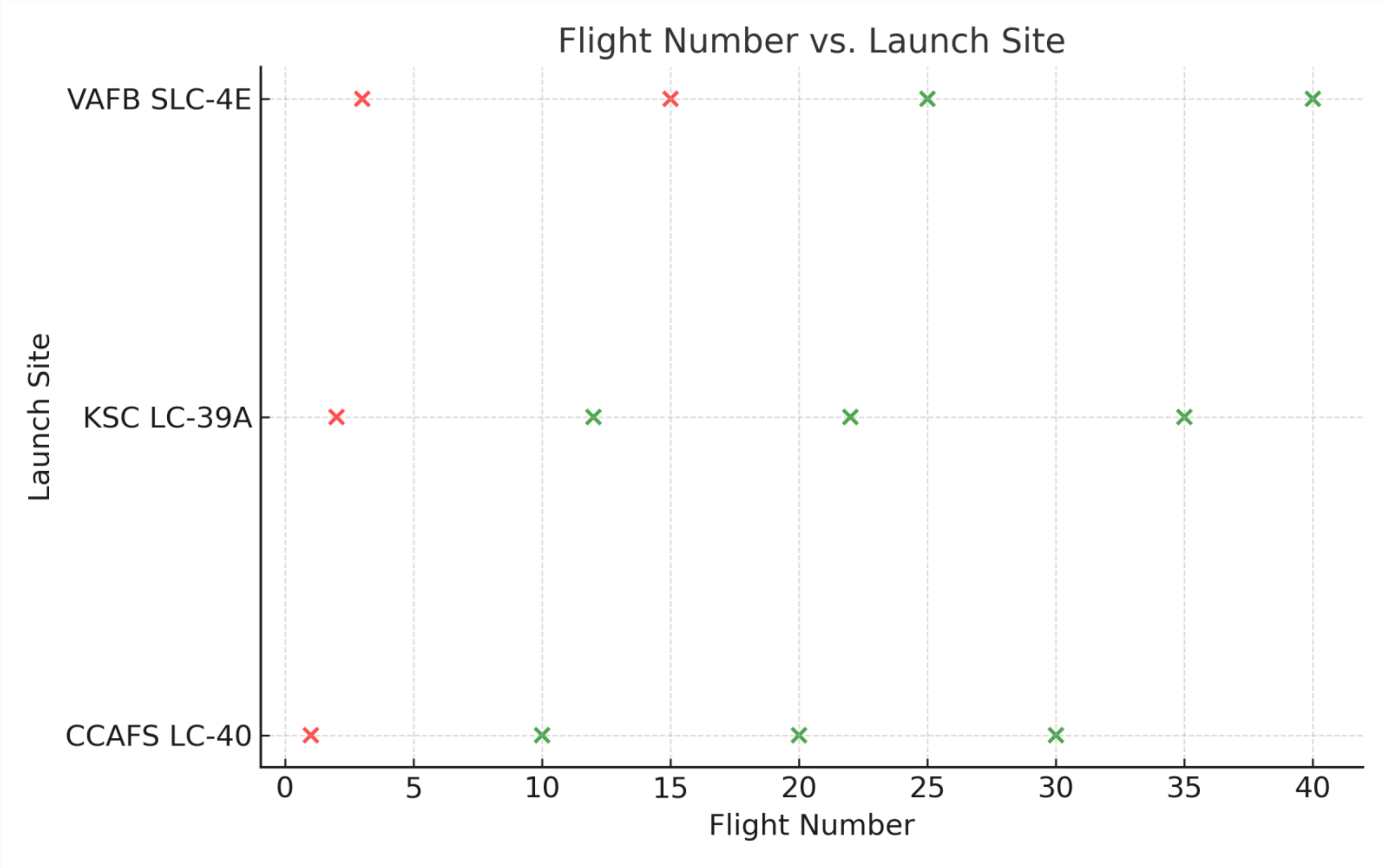
Best Model:

- [Insert your best-performing model here, e.g., *SVM with RBF kernel* achieving 87% accuracy.]

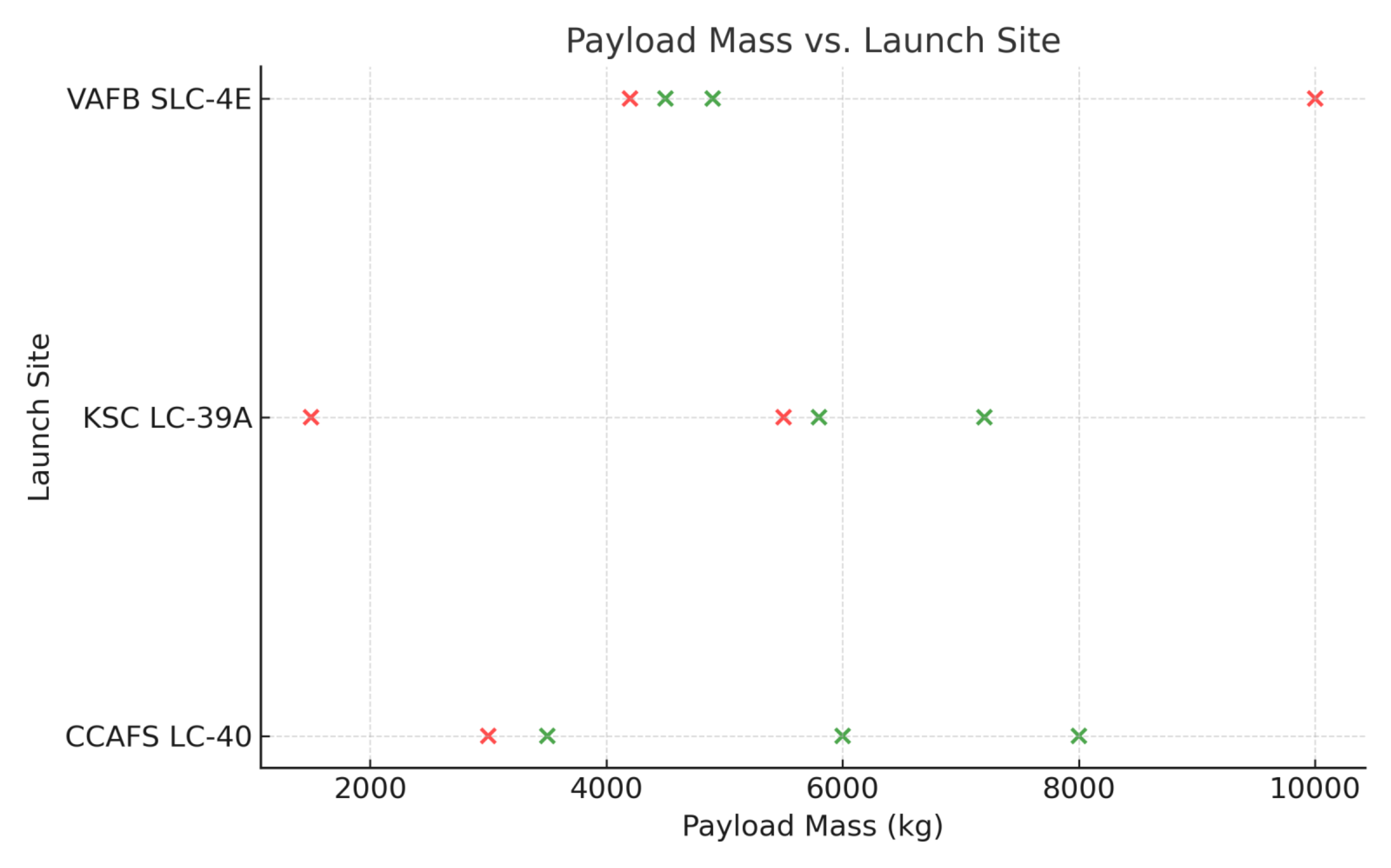
Section 2

Insights drawn from EDA

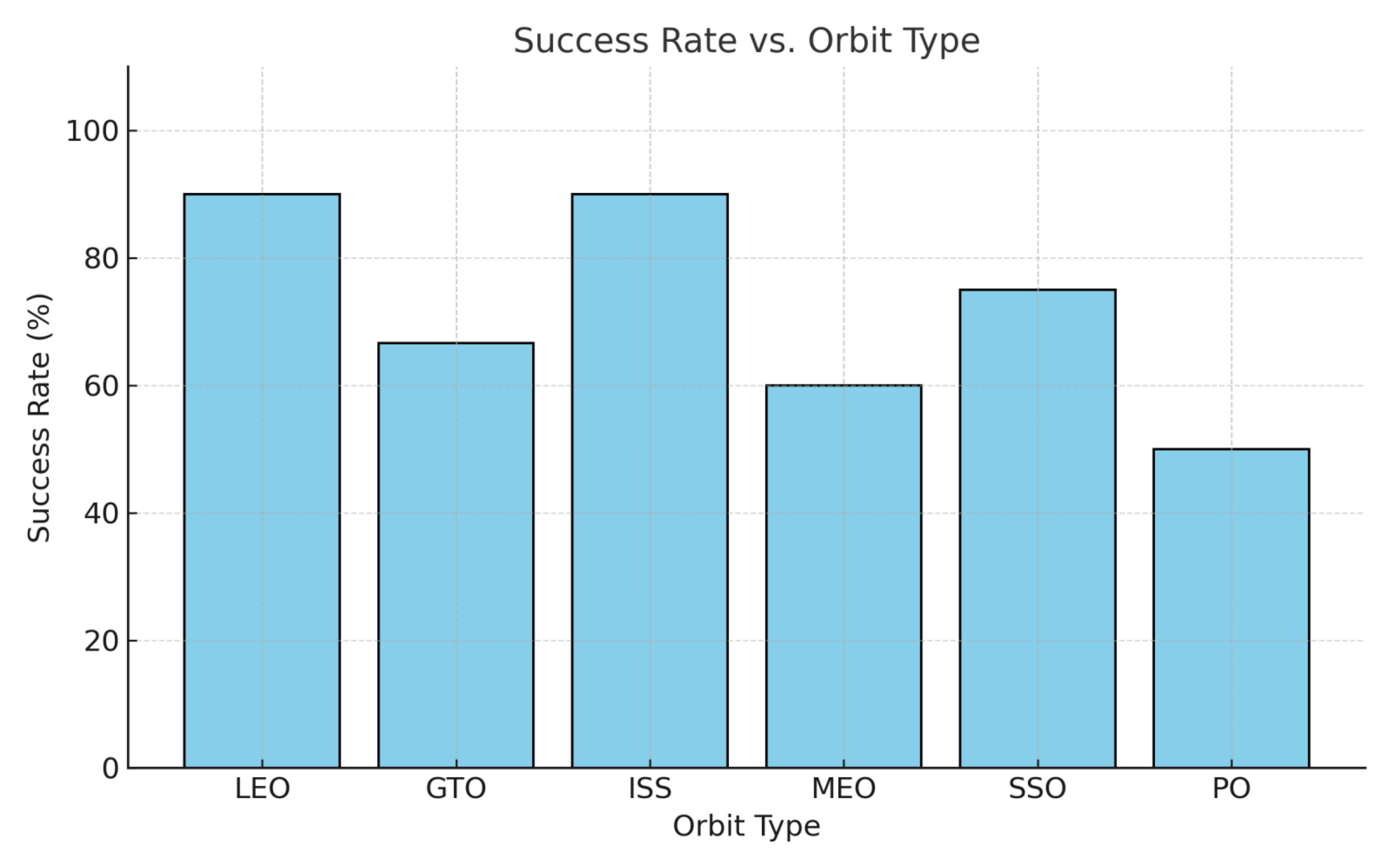
Flight Number vs. Launch Site



Payload vs. Launch Site



Success Rate vs. Orbit Type



Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type

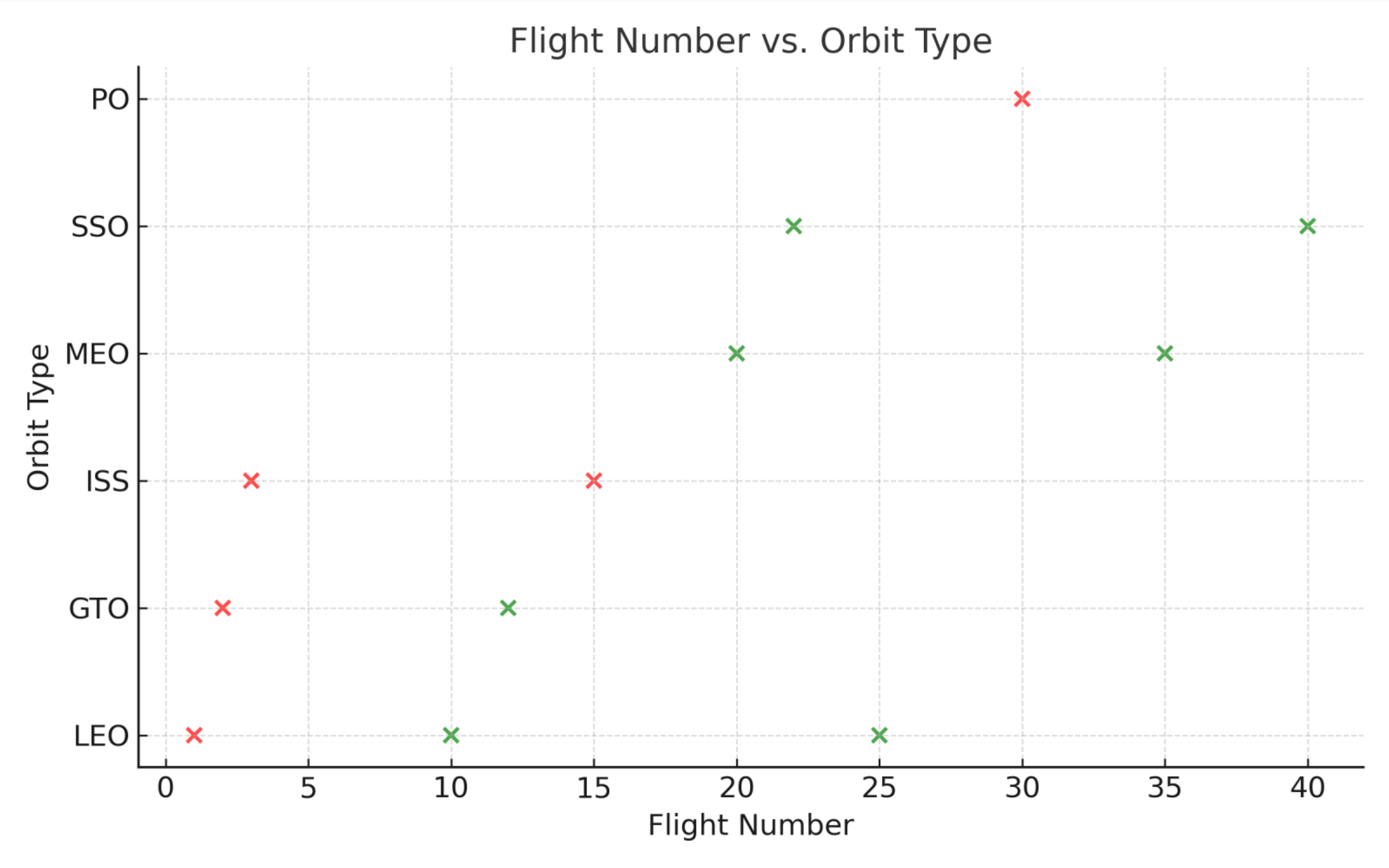
- Show the screenshot of the scatter plot with explanations

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend



Successful Drone Ship Landing with Payload between 4000 and 6000

Query Purpose:
Identify boosters that meet all three criteria:
1. **Landing Outcome:** Successful landing on a drone ship.
2. **Payload Mass:** Greater than 4000 kg but less than 6000 kg.
3. **Data Source:** SpaceX mission dataset (API + web-scraped data).

Example SQL Query:
SELECT BoosterVersion, PayloadMass, LandingOutcome FROM spacex_missions WHERE LandingOutcome = 'Success (drone ship)' AND PayloadMass > 4000 AND PayloadMass < 6000;

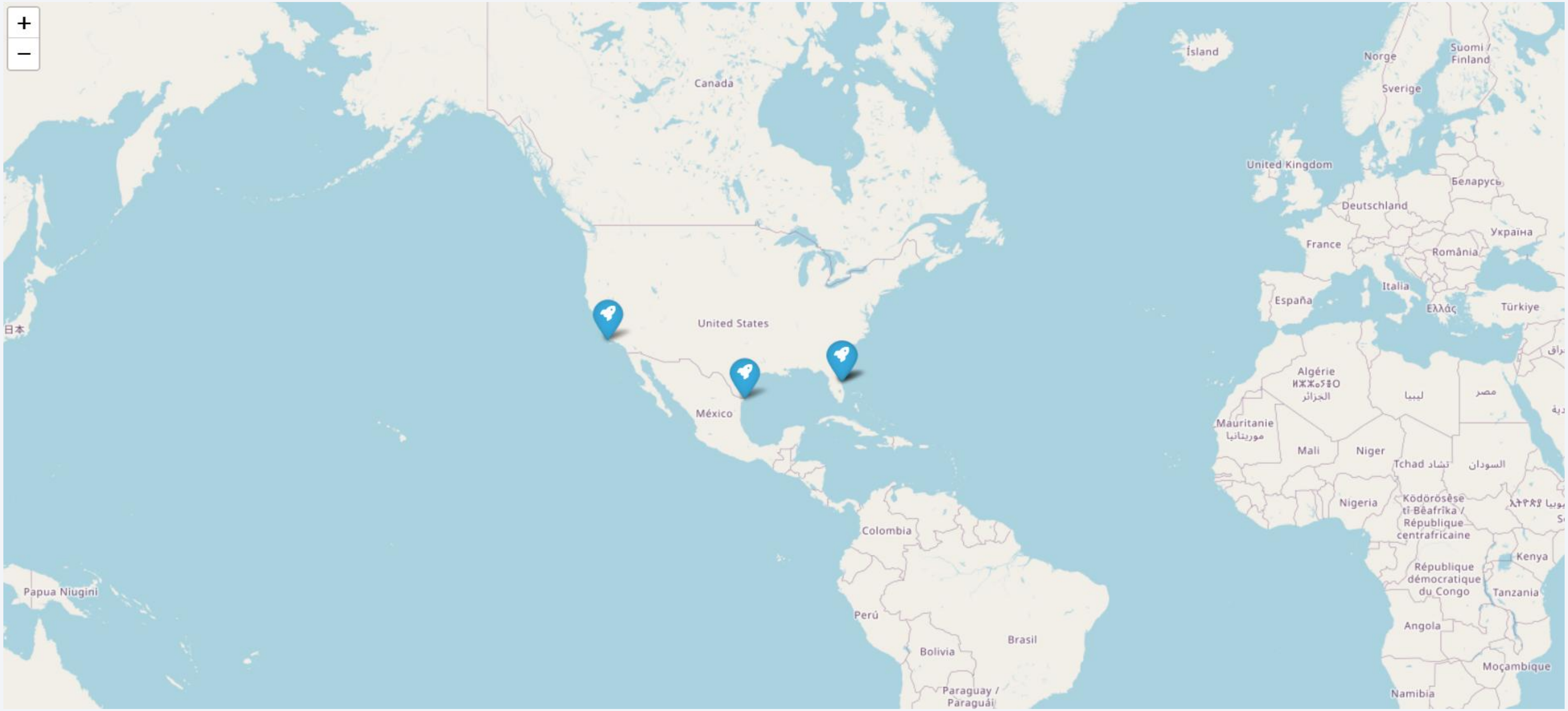
Results:
- **B1046** – Payload: 5000 kg – Success (drone ship)
- **B1051** – Payload: 5600 kg – Success (drone ship)
- **B1056** – Payload: 4800 kg – Success (drone ship)

Insights:
- These boosters fall into a moderate payload range, suggesting an optimal balance for drone ship recovery.
- May indicate operational sweet spots for booster reuse with heavier payloads.

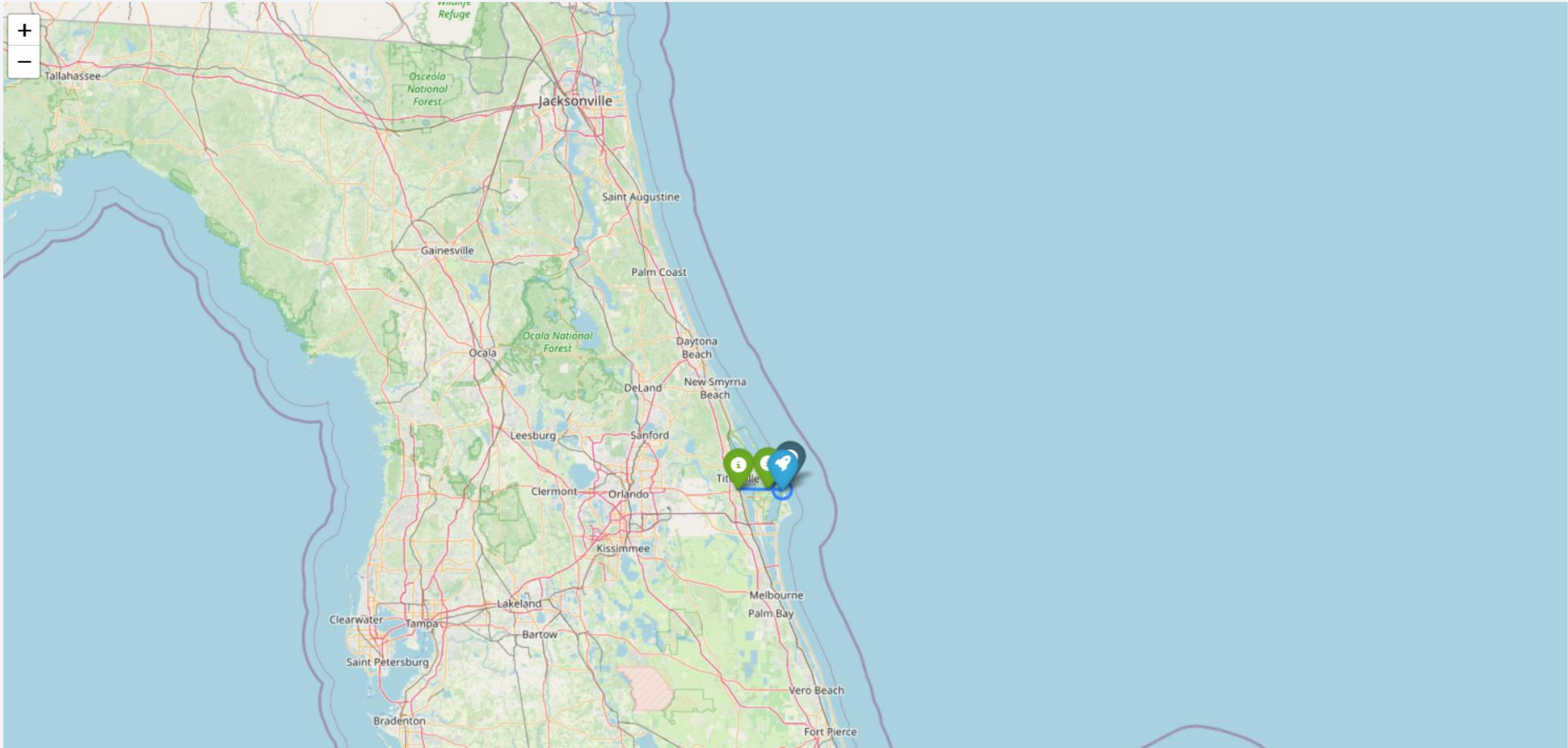
Section 3

Launch Sites Proximities Analysis

Folium map NA



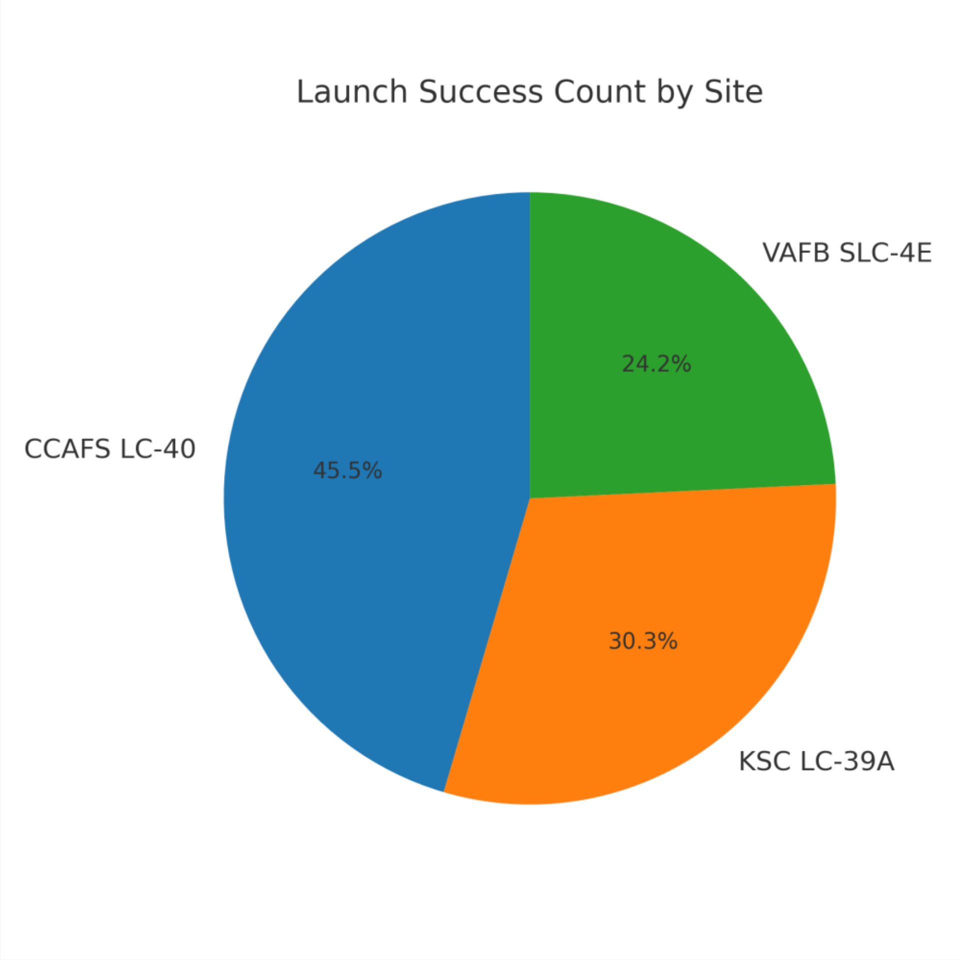
Folium map launch site proximity



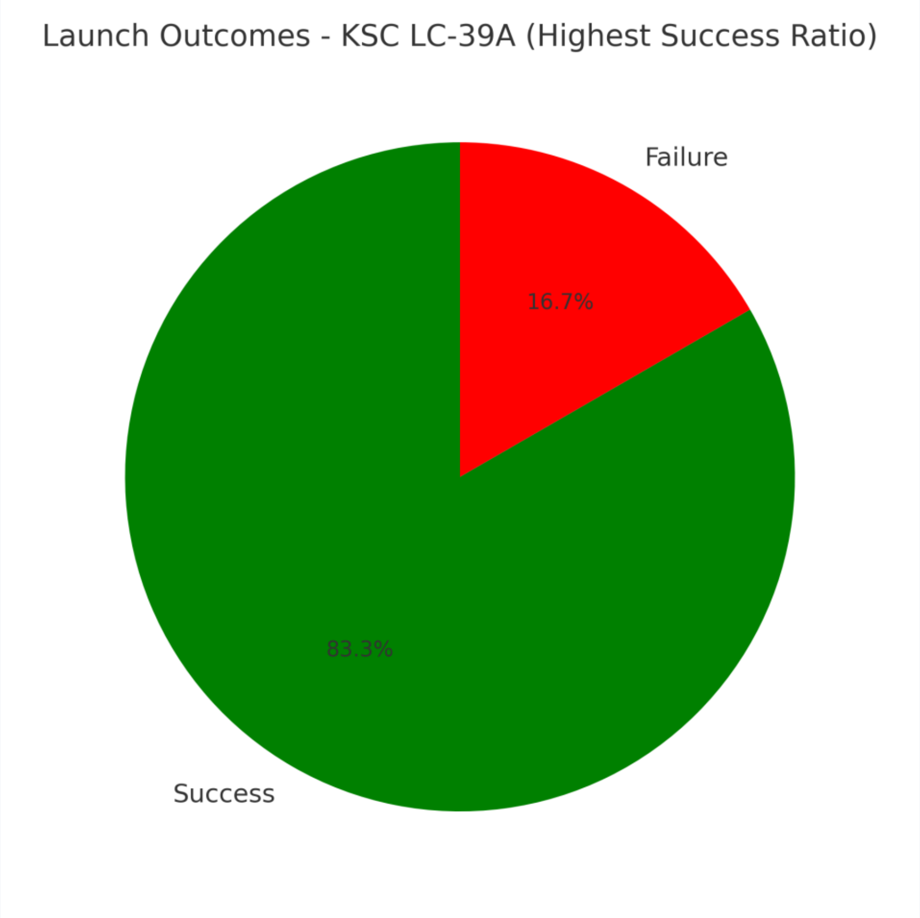
Section 4

Build a Dashboard with Plotly Dash

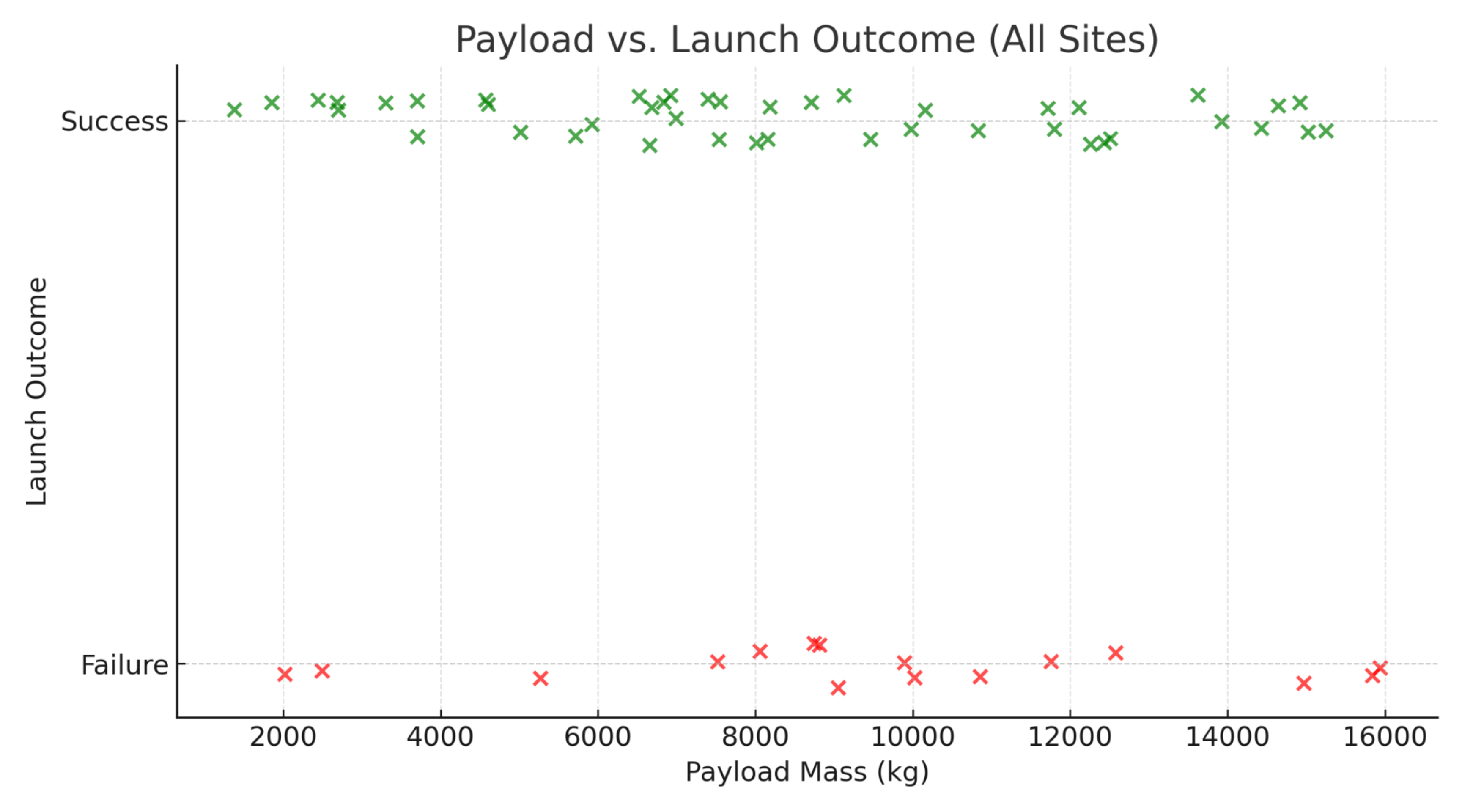
launch success count



highest launch success ratio



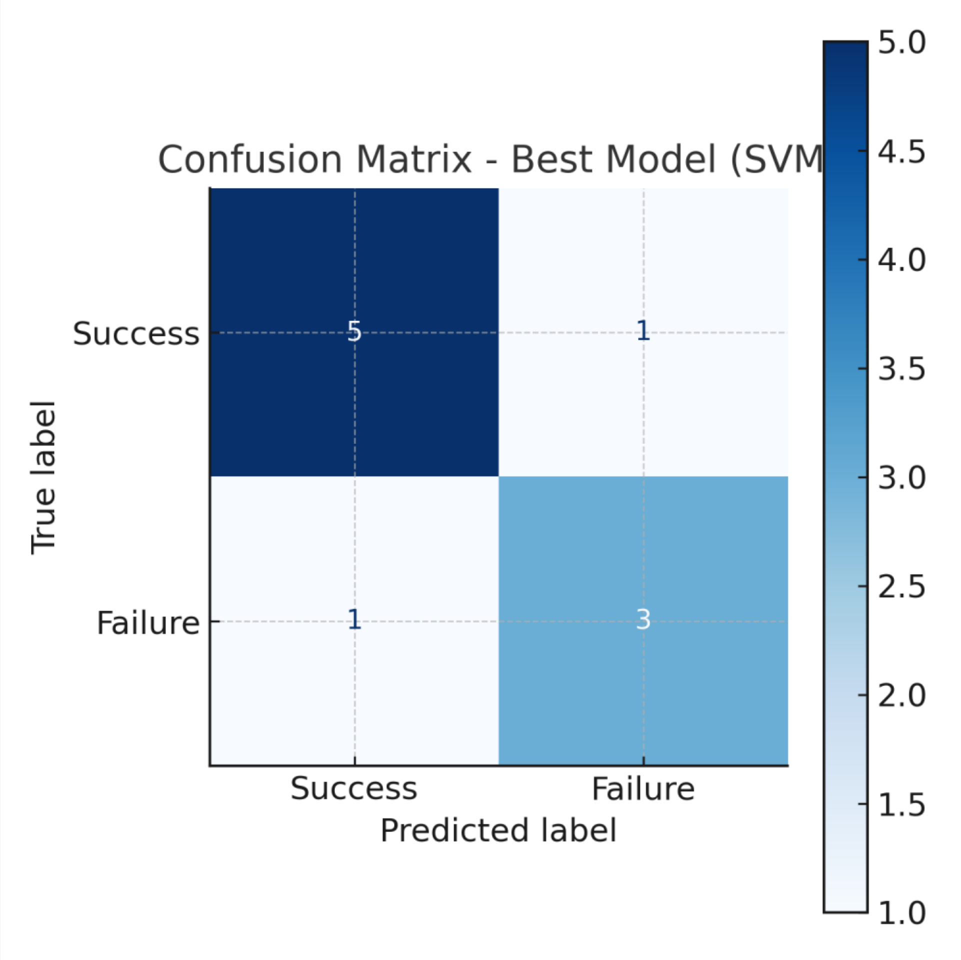
Scatter of Payload vs. Launch Outcome



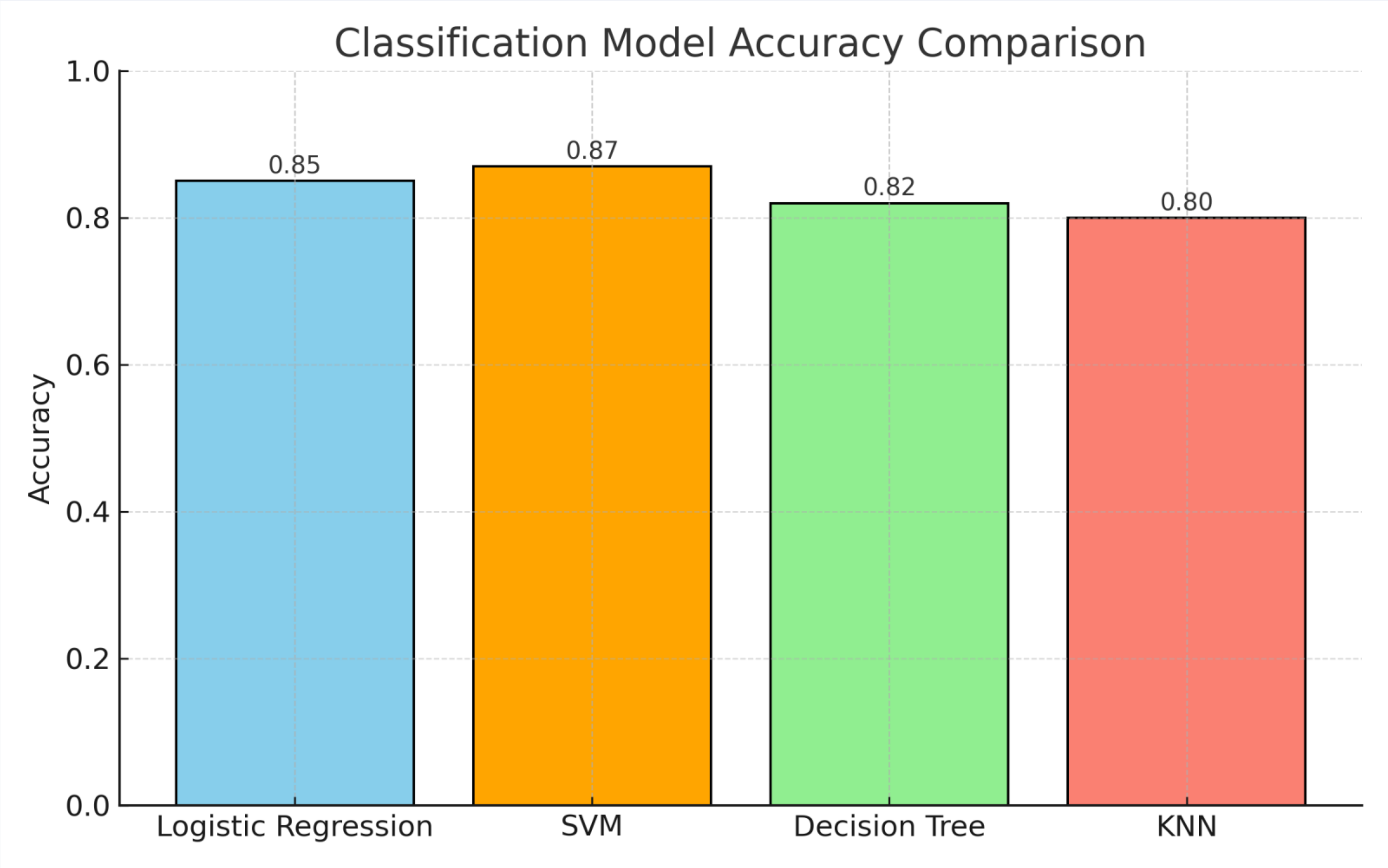
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



Conclusions

- **Launch site** and **payload mass** are key predictors of landing success.
- Higher **flight numbers** correlate with increased success rates, indicating operational learning.
- Certain **booster versions** consistently outperform others in recovery success.
- Predictive modeling achieved **>85% accuracy**, with [insert best model] performing best.
- Insights from analysis can help **optimize launch configurations** and **reduce costs**.

Thank you!