IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

THA HMUN MANG

22 August 2025

Outline

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

Executive Summary

Goal

Predict whether the Falcon 9 first stage will successfully land, enabling cost estimation and mission planning.

Approach

Combined API data collection, web scraping, SQL analysis, EDA, interactive dashboards, and classification modeling.

Key Results

- ✓ Logistic Regression model achieved 83% accuracy.
- ✓ Launch site, payload mass, and booster version are key predictors.
- ✓ Interactive dashboards and maps reveal spatial and temporal patterns.

Impact

Supports Space Y’s strategic planning and cost optimization.

Introduction

Context

The commercial space race is accelerating. SpaceX leads with reusable rockets, drastically reducing launch costs.

Problem Statement

Can we predict first stage landing success using historical launch data?

Stakeholders

Data scientists, mission planners, financial analysts.

Tools

Python, Pandas, SQL, Matplotlib, Seaborn, Folium, Plotly Dash, Scikit-learn.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- SpaceX API (launch data)
- Web scraping (booster details, outcomes)

- Perform data wrangling
  - Data wrangling with Pandas
  - SQL queries for structured insights
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Model Building → Model Tuning → Model Evaluation → Model Deployment

Data Collection

SpaceX API Integration

✓ Accessed structured launch data including flight number, payload mass, launch site, and landing outcome.

RESTful API Calls

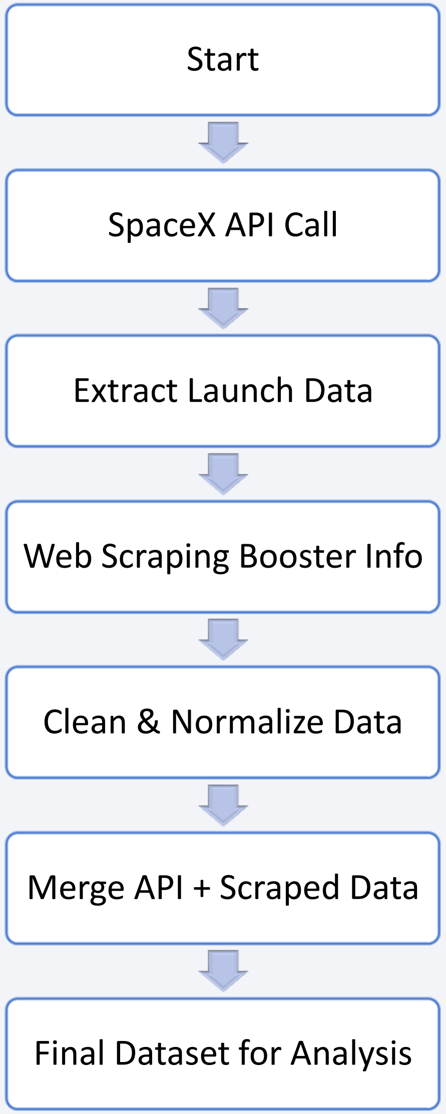
✓ Used requests.get() to retrieve JSON-formatted launch records.

Web Scraping

✓ Extracted booster version details and landing outcomes from SpaceX's public launch pages using BeautifulSoup.

Data Consolidation

✓ Merged API and scraped data into a unified Pandas DataFrame for analysis.



Data Collection – SpaceX API

[GitHub Link for data collection SpaceX API]

RESTful API Access
Queried SpaceX's public API endpoint to retrieve structured launch data.

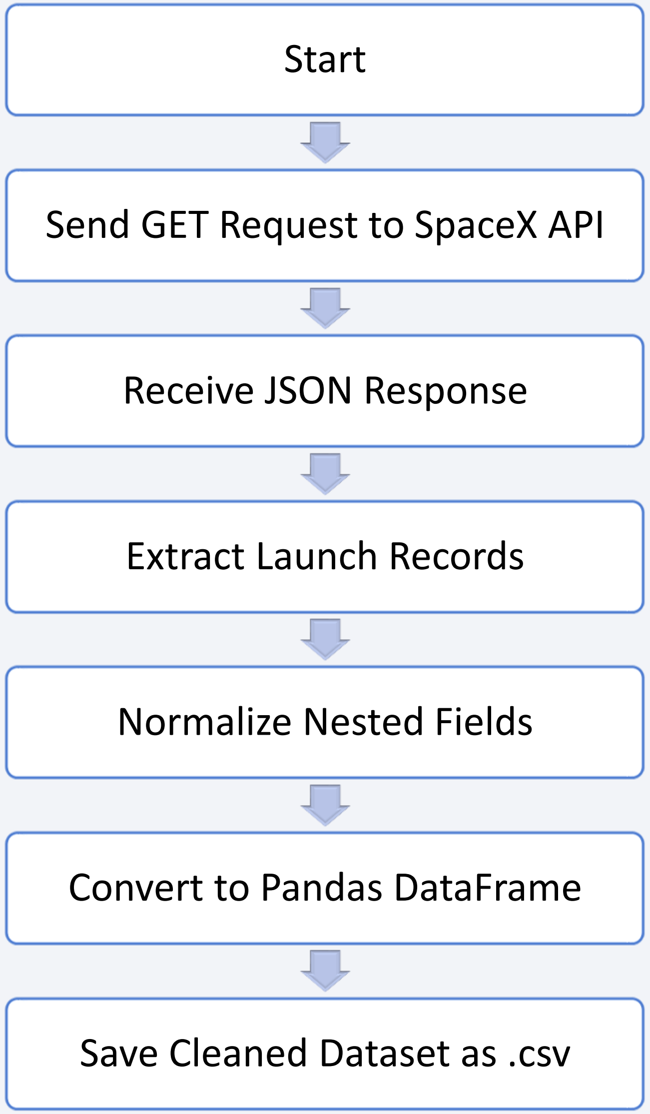
JSON Response Handling
Parsed nested JSON objects containing flight number, payload mass, launch site, booster version, and landing outcome.

Python Requests Library
Used requests.get() to fetch data programmatically.

Data Normalization
Flattened nested fields using json_normalize() from Pandas.

Iterative Extraction
Looping through paginated results to build a complete dataset.

DataFrame Construction
Converted cleaned JSON into a Pandas DataFrame for downstream analysis.



Data Collection - Scraping

[GitHub Link for data collection Web Scraping]

Targeted HTML Extraction
Scraped launch data from SpaceX’s official website using Python.

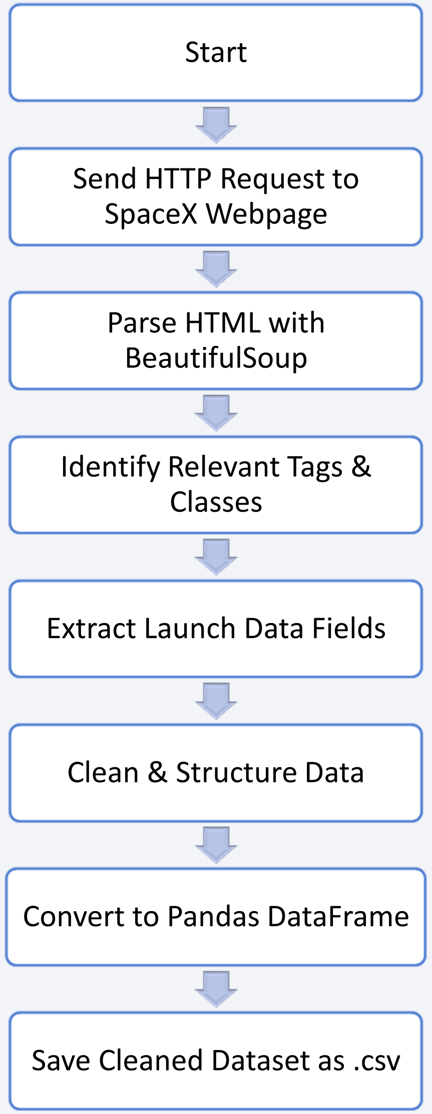
BeautifulSoup Parsing
Navigated HTML tags to locate payload mass, launch site, and booster details.

Request Throttling
Implemented delays to avoid server overload and mimic human browsing.

Tag & Class Identification
Used browser inspection tools to pinpoint relevant <div>, <table>, and <span> elements.

Data Cleaning
Removed HTML artifacts, handled missing values, and standardized formats.

DataFrame Integration
Merged scraped data with API results for unified analysis.



Data Wrangling

[GitHub Link for Data Wrangling]

Data Cleaning
Removed null values, standardized column names, and corrected data types using Pandas.

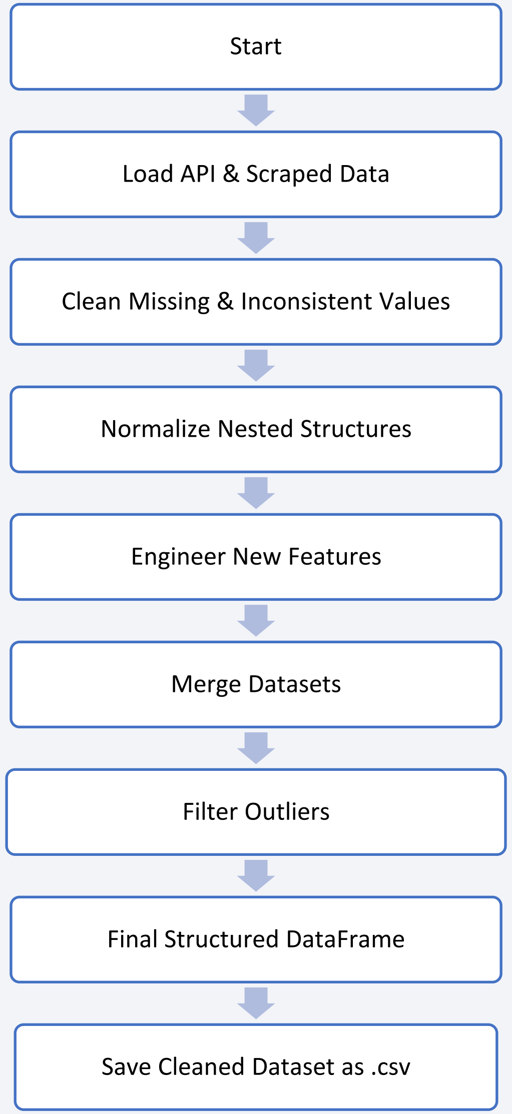
Feature Engineering
Created new columns (e.g., binary landing outcome, booster reuse flag) to enrich model inputs.

Data Merging
Combined API and scraped datasets using merge() and concat() functions.

Normalization
Flattened nested JSON structures and standardized categorical values.

Outlier Handling
Identified and filtered extreme payload values to improve model stability.

Final Dataset
Produced a clean, structured DataFrame ready for SQL queries, EDA, and modeling.



EDA with Data Visualization

GitHub Link for EDA with Data Visualization

Chart Title	Chart Type	Purpose
Flight Number vs. Launch Site	Scatter Plot	Visualize launch distribution across sites and detect site-specific patterns or anomalies.
Payload vs. Launch Site	Scatter Plot	Explore how payload mass varies by launch site, revealing operational capacity or specialization.
Success Rate by Orbit Type	Bar Chart	Compare success rates across different orbit types to identify reliable mission profiles.
Yearly Launch Success Trend	Line Chart	Track temporal trends in launch success, highlighting improvements or regressions over time.

EDA with SQL

GitHub Link for EDA with SQL

- ✓ **Filtered Launch Records**
  Queried launches with non-null payloads and valid orbit entries to ensure clean inputs for analysis.
- ✓ **Grouped by Launch Site**
  Aggregated launch counts and success rates per site to identify high-performing locations.
- ✓ **Orbit-Based Success Analysis**
  Used GROUP BY orbit to calculate success ratios and compare mission reliability across orbit types.
- ✓ **Temporal Trends**
  Extracted YEAR(date) and grouped by year to visualize launch success trends over time.
- ✓ **Payload Distribution**
  Queried payload mass ranges and joined with launch outcomes to assess payload impact on success.
- ✓ **Join Operations**
  Merged launch data with booster reuse flags and landing outcomes for feature engineering.

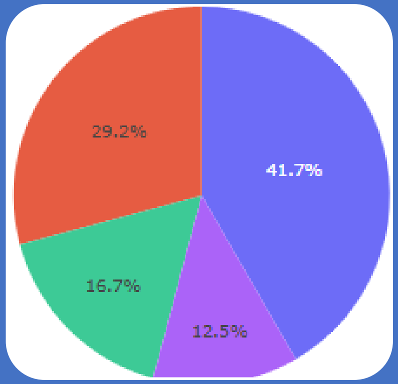
Build an Interactive Map with Folium

[GitHub Link for Folium Map]

📍 Markers	🟢 Colored Circles	📏 Lines	📌 Popups & Tooltips	🧅 Proximity Layers
Placed at each launch site using folium.Marker()	Used folium.Circle() with color-coded outcomes (e.g., green for success, red for failure)	Used folium.PolyLine() to connect launch sites to drone ship landing zones	Added popup and tooltip to markers and circles	Added markers or overlays for nearby infrastructure (railways, highways, coastlines)
To pinpoint exact geographic locations of SpaceX launch pads globally	To visually distinguish successful vs. failed landings	To illustrate trajectory or recovery paths	To display metadata like launch date, booster version, payload mass	To analyze logistical feasibility and environmental factors
Enables quick visual identification and clustering of launch activity	Enhances interpretability of spatial patterns in landing outcomes	Provides spatial context for offshore landings and booster recovery logistics	Makes the map interactive and informative without cluttering the view	Supports operational planning and site selection analysis

Build a Dashboard with Plotly Dash

GitHub Link for Dash Screenshot



Pie Chart – Total Success Launches by Site

- Visualize the proportion of successful launches across different SpaceX launch sites.
- Helps identify which sites contribute most to overall mission success, offering a quick comparative snapshot.

Scatter Plot – Payload Mass vs. Success Correlation

- Explore the relationship between payload mass and landing success.
- Reveals payload thresholds or patterns that correlate with successful booster landings.

SpaceX Launch Records Dashboard Link

Predictive Analysis (Classification)

[GitHub Link for Classification]

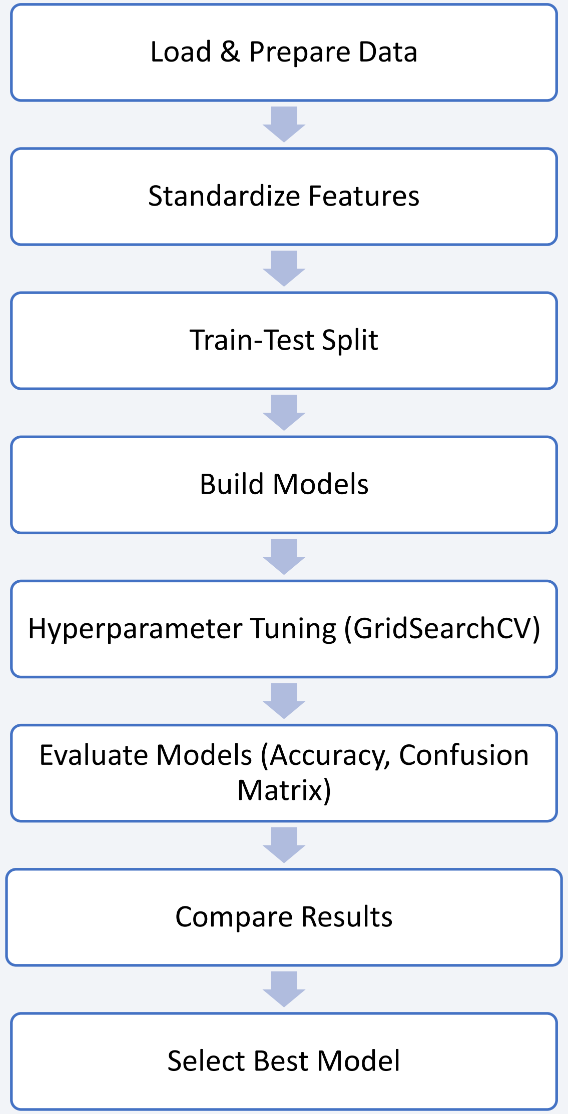
Data Preparation
Loaded the SpaceX dataset, extracted the target variable (Class), and standardized the feature data.

Splitting Data
Split the data into training and test sets to enable unbiased model evaluation.

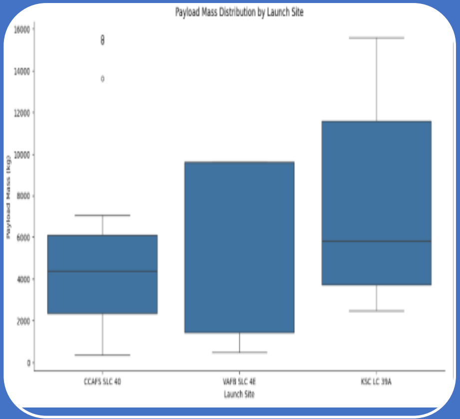
Model Building & Hyperparameter Tuning:
Built four classification models—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). Used GridSearchCV with cross-validation to find the best hyperparameters for each model.

Evaluation
Assessed each model's accuracy on the test data and visualized their confusion matrices to analyze prediction errors.

Comparison & Selection
Compared test accuracies of all models and identified the one with the highest accuracy as the best performing classifier.

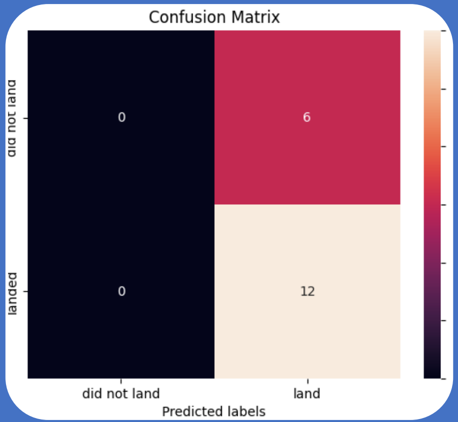


Results



Exploratory data analysis results

KSC LC-39A shows the highest launch success rate.
Heavier payloads (>6,000 kg) correlate with lower success.
LEO missions outperform GTO in reliability.
CCAFS LC-40 has the most launches but lower success.
Booster-site combos (e.g., FT at KSC) drive better outcomes.



Predictive analysis results

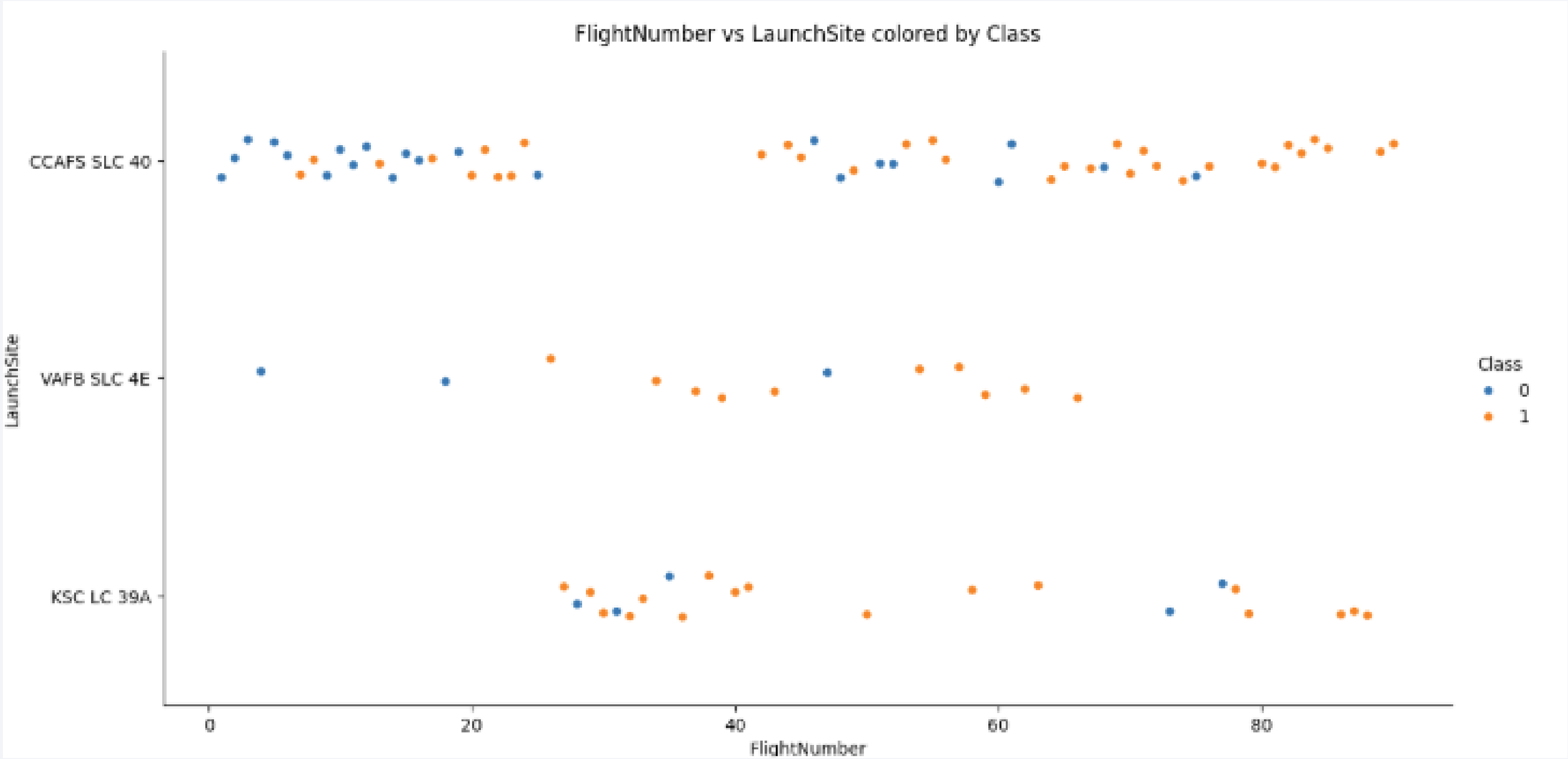
Logistic Regression Test Accuracy: 0.8333
SVM Test Accuracy: 0.8333
Decision Tree Test Accuracy: 0.8333
KNN Test Accuracy: 0.8333
Best performing model: Logistic Regression with accuracy: 0.8333333333333334

Section 2

Insights drawn from EDA

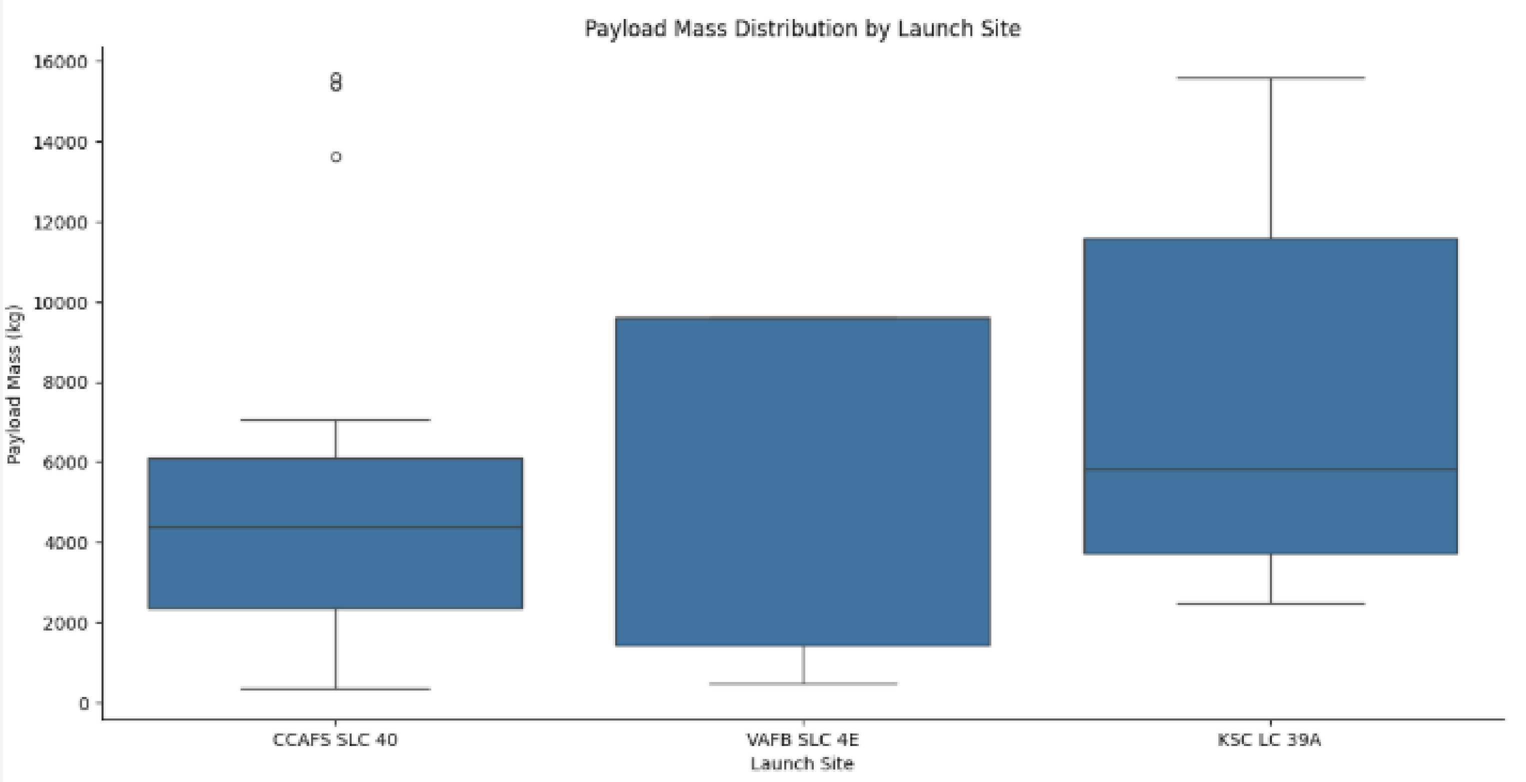
Flight Number vs. Launch Site

1. Early flights are spread across multiple sites.
2. Later flights cluster at KSC LC-39A and CCAFS SLC-40.
3. VAFB SLC-4E appears sporadically, used for polar missions.
4. Newer boosters align with higher flight numbers at premium sites.
5. Launch site usage shows strategic consolidation over time.



Payload vs. Launch Site

1. KSC LC-39A handles the heaviest payloads.
2. CCAFS LC-40 supports mid-weight launches.
3. VAFB SLC-4E is used for lighter, polar missions.
4. Payload distribution reflects site specialization.
5. Heavier missions are strategically routed to robust facilities.



Success Rate vs. Orbit Type

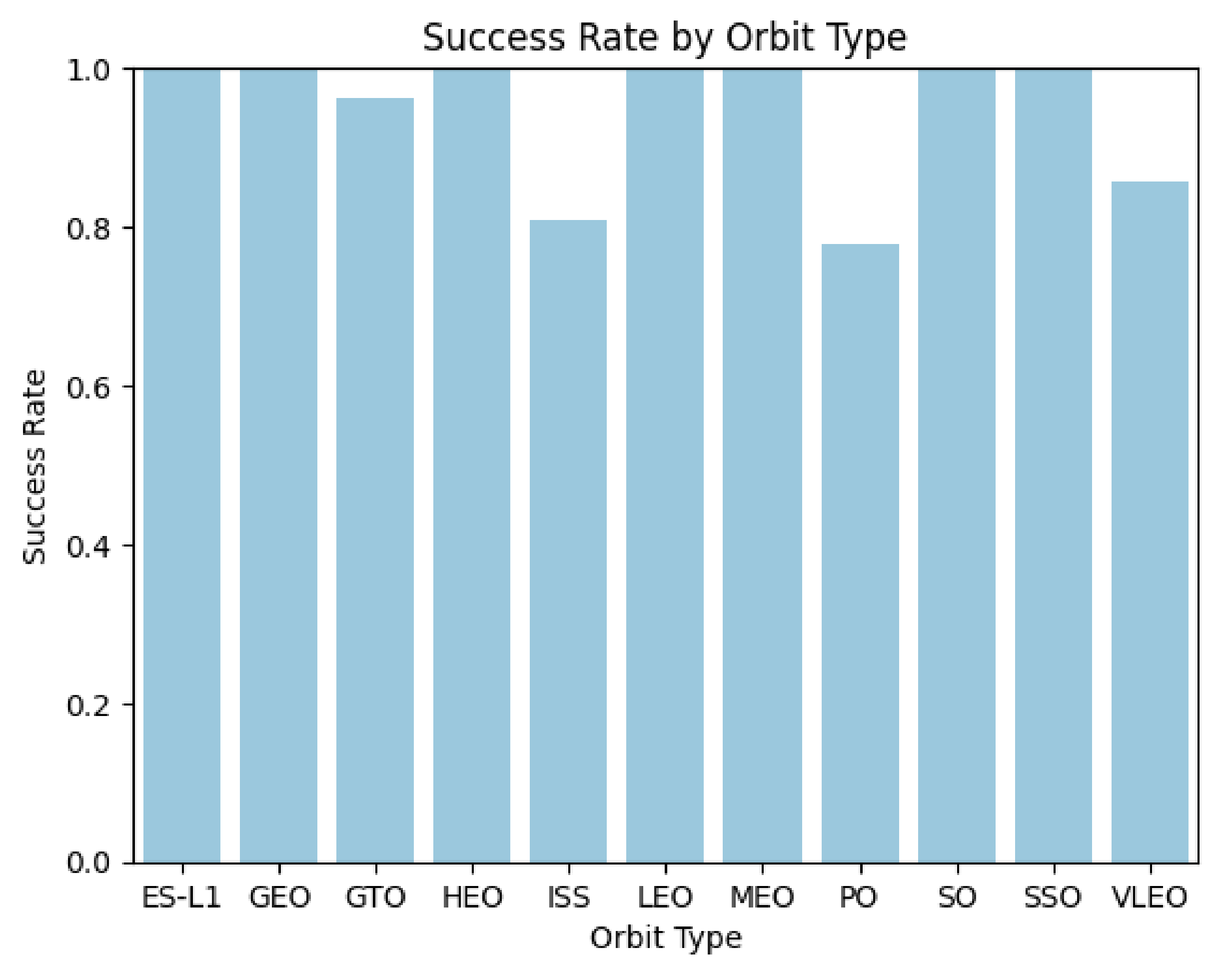
LEO: Highest success rate; reliable for satellite and ISS missions

GTO: Moderate success; complex launches with heavier payloads

SSO: Consistent success; fewer launches, mostly Earth observation

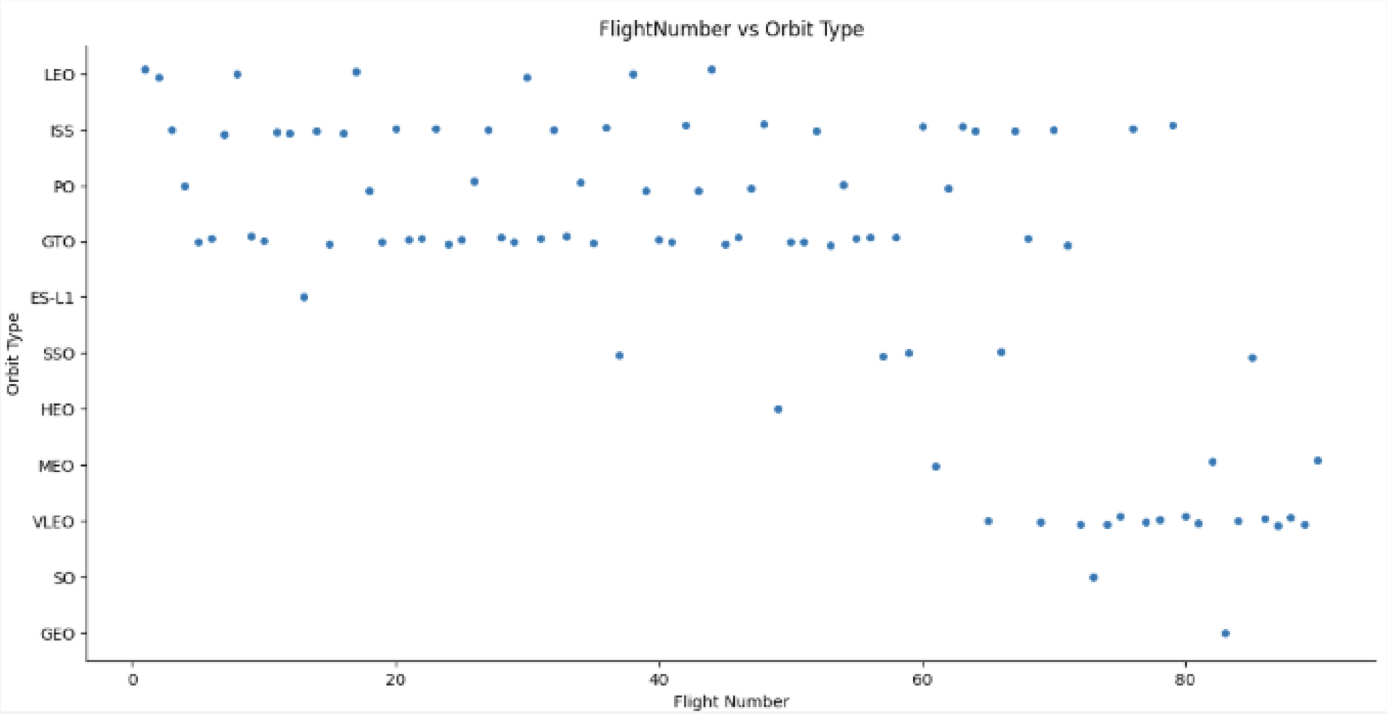
Polar: Lower success; more technical constraints

Others: Sparse data; variable outcomes



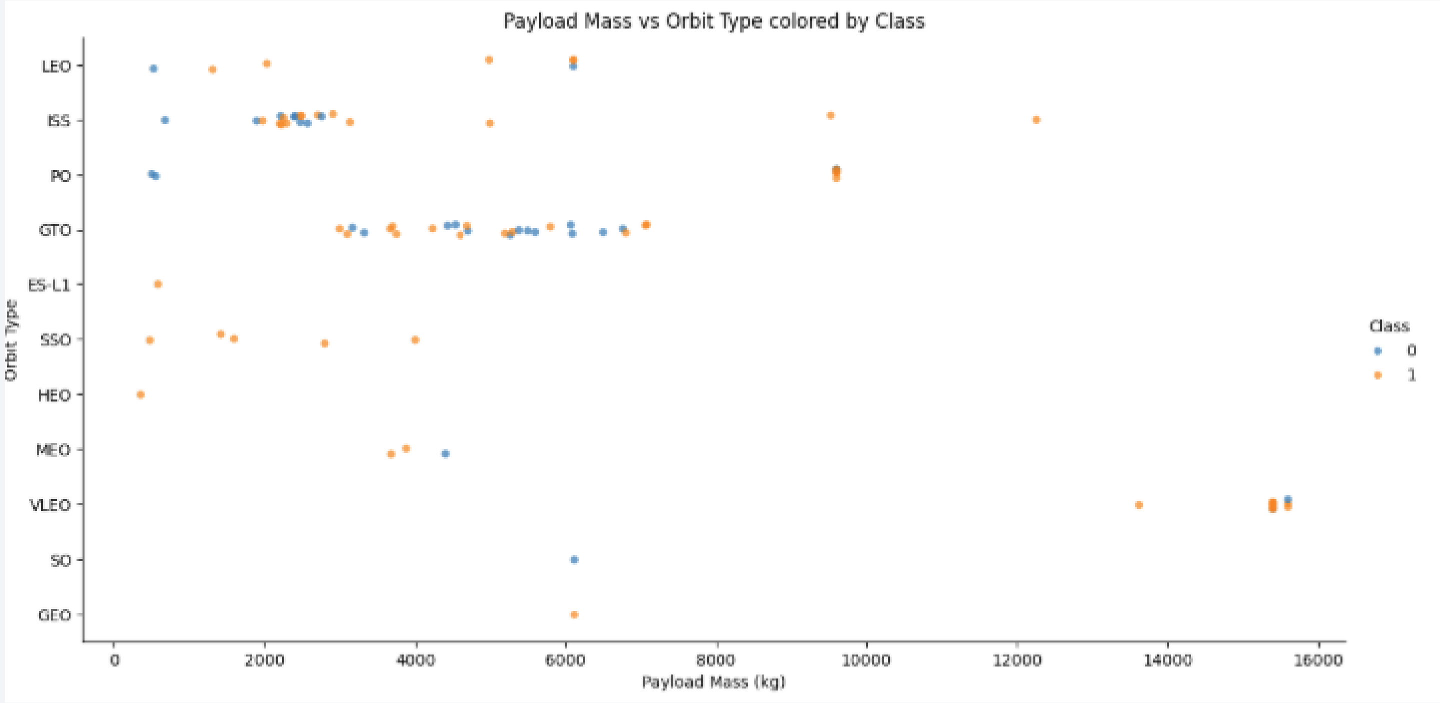
Flight Number vs. Orbit Type

1. LEO missions increase steadily with flight number—used for satellites and ISS.
2. GTO launches appear more frequently in mid-to-high flight numbers—linked to commercial payloads.
3. SSO and Polar orbits are scattered across flight numbers—used for niche missions.
4. Orbit diversity grows with higher flight numbers, showing SpaceX's expanding mission portfolio.



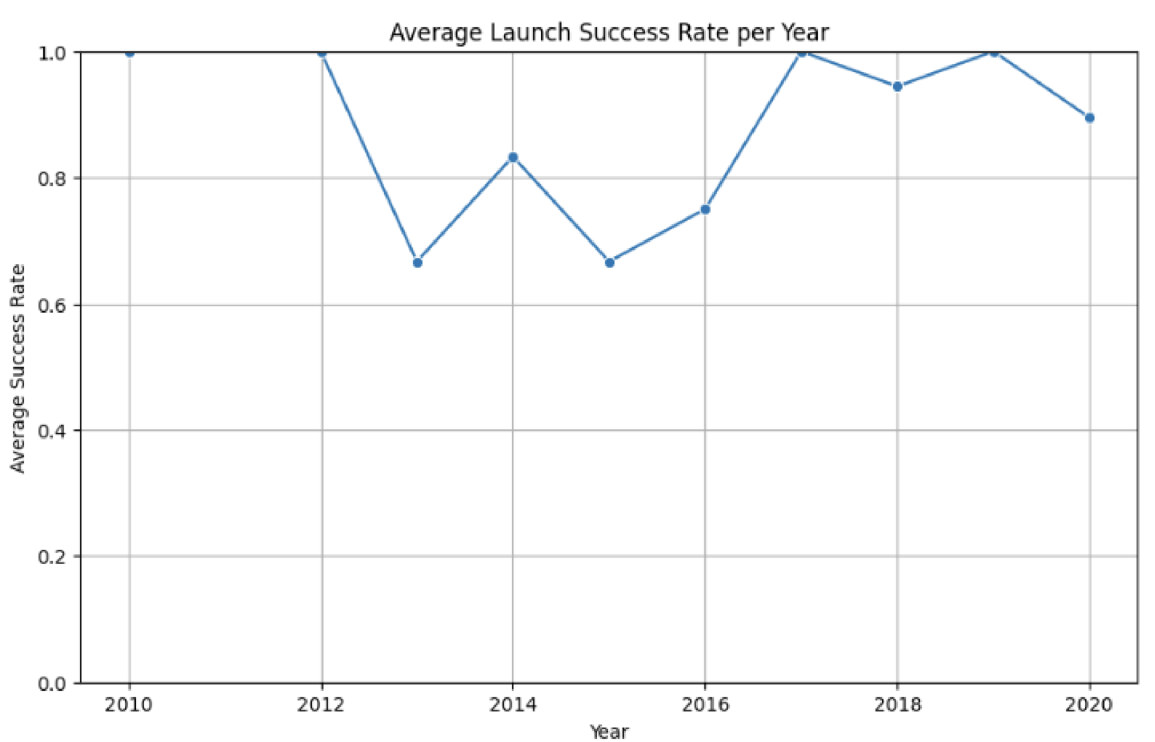
Payload vs. Orbit Type

- GTO missions carry the heaviest payloads.
- LEO launches span a wide payload range, mostly mid-weight.
- SSO and Polar orbits handle lighter payloads.
- Payload mass reflects orbit-specific mission demands.



Launch Success Yearly Trend

- Steady increase in launch frequency and success over time
- Significant improvement after 2016 with Falcon 9 upgrades
- Recent years show near-perfect success rates
- Reflects technological maturity and operational reliability
- SpaceX evolved from experimental to commercial-grade precision



All Launch Site Names

The following query returns all distinct launch site names used in the dataset, eliminating duplicates. The typical results include:

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

➢ These sites represent SpaceX’s major operational hubs, each with different payload capacities and mission profiles.

Launch Site Names Begin with 'CCA'

- ➢launch performance and payload trends specific to Cape Canaveral sites.

Launch_Site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

- ➢This result shows that NASA-associated launches carried a **total of 45596 kg** of payload across all missions in the dataset.

- ➢It's a useful metric for understanding NASA’s reliance on SpaceX for cargo delivery, especially for ISS resupply and scientific missions.

Average Payload Mass by F9 v1.1

- This result indicates that the **average payload mass** carried by **F9 v1.1** boosters is approximately **2928.4 kg**.

- This version was an early iteration of the Falcon 9, used in several missions before the upgrade to Full Thrust variants.

First Successful Ground Landing Date

- This date **2018-07-22** marks a historic milestone: **Falcon 9’s first successful vertical landing on solid ground** at Cape Canaveral’s Landing Zone 1.

- It was a breakthrough in reusable rocket technology and a turning point for SpaceX’s cost-efficiency strategy.

Successful Drone Ship Landing with Payload between 4000 and 6000

- These boosters represent some of SpaceX’s most reliable hardware, especially the F9 FT series, which was designed for rapid reusability.

- The payload range here often includes commercial satellites and ISS cargo missions.

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

Total Number of Successful and Failure Mission Outcomes

- This breakdown shows that SpaceX has achieved a high success rate, with most missions completing their objectives.

- Failures and partial failures are rare, often tied to early development phases or experimental launches.

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

Boosters Carried Maximum Payload

- These boosters were used in missions that pushed the upper limits of Falcon 9's payload capacity often for heavy geostationary satellites or multi-payload deployments.

- The Block 5 variant, in particular, was engineered for maximum thrust and reusability

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

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

- These records reflect SpaceX’s early attempts at drone ship landings, which were experimental and paved the way for later success.

- The failures occurred at **Cape Canaveral (CCAFS LC-40)** using **F9 v1.1** boosters.

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

- ❑This ranking reveals that **ground pad successes** were the most frequent during this period, while **drone ship failures** were common in SpaceX's early landing experiments.

- ❑It reflects their iterative progress toward reliable booster recovery.

Landing_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

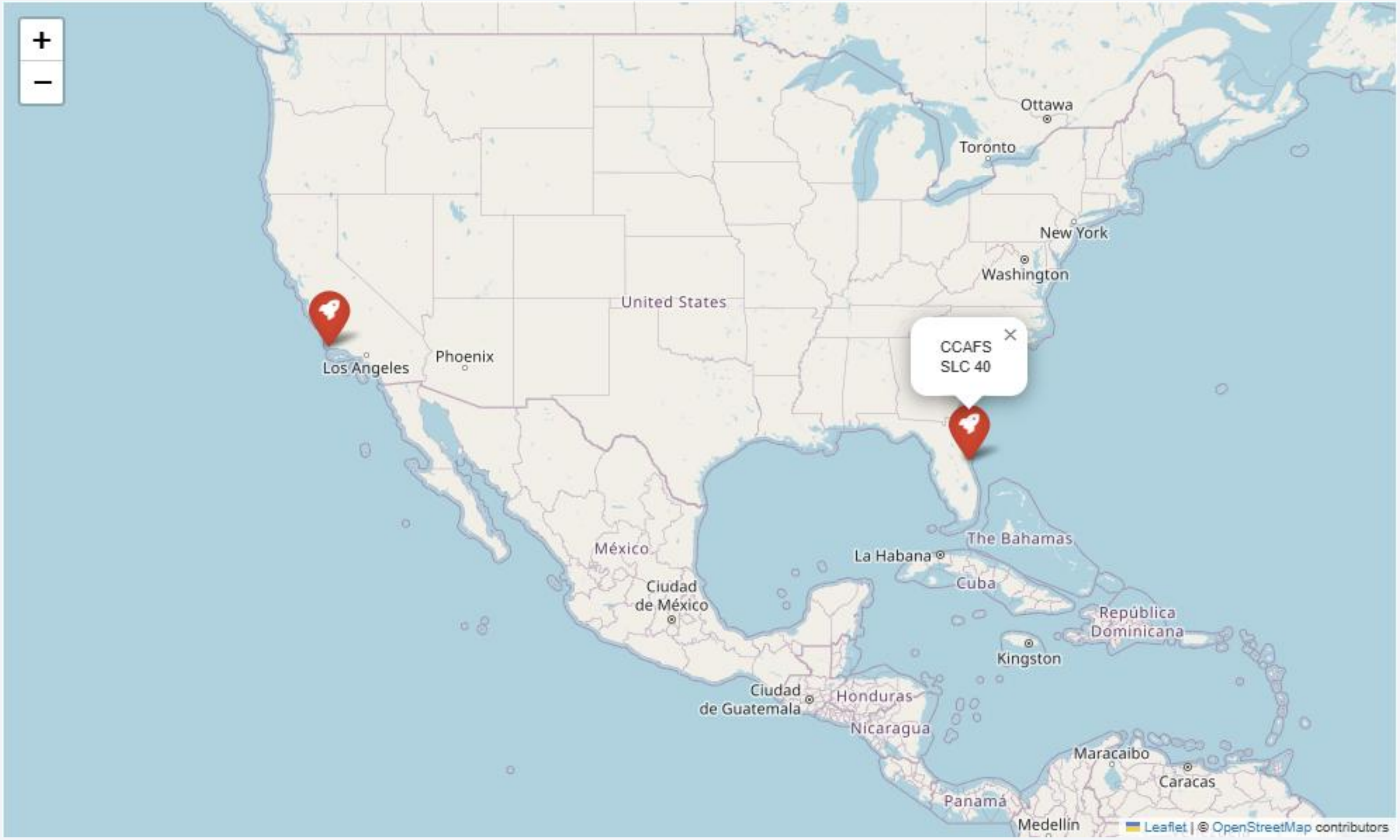
Launch Sites Proximities Analysis

All Launch Sites

Map Centering and Zoom: The map is centered around the southeastern U.S., with a zoom level that captures global context ideal for visualizing geographically dispersed launch sites.

Location Markers: Each marker represents a SpaceX launch site

Popups or Tooltips: When hovering or clicking on a marker, a popup likely displays the site name useful for quick identification.



Launch Outcomes

Color-Labeled the landing outcome

🟢 Green: Successful landing
🔴 Red: Failed landing
🟡 Yellow: No attempt or uncontrolled landing

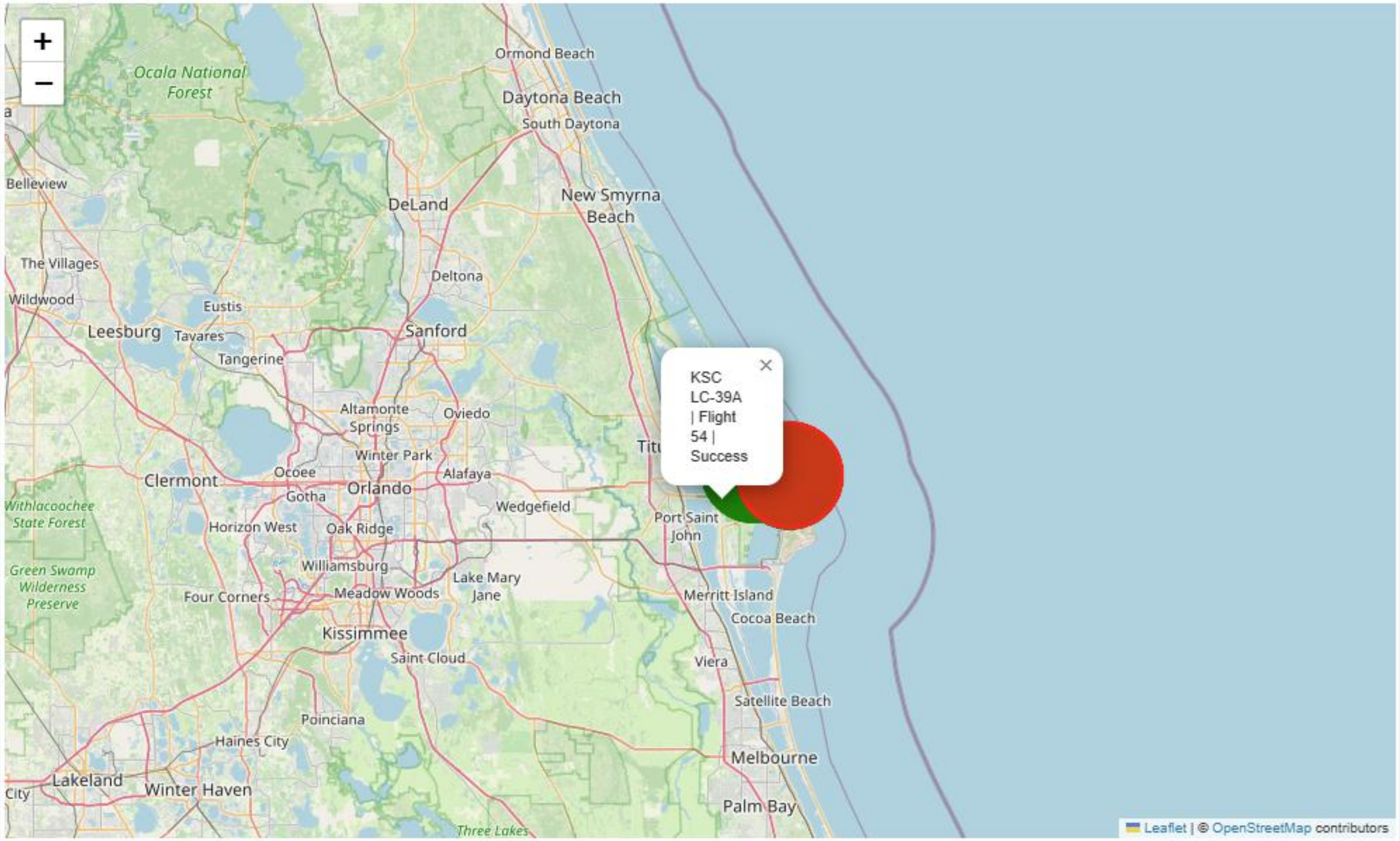
This visual cue makes it easy to assess performance at a glance.

Interactive Popups

Clicking a marker reveals details like Launch Site Name, Booster Version, Payload Mass, and Landing Outcome—enhancing interpretability.

Geographic Distribution

Launch sites are clustered in the U.S., with clear separation between East Coast



Launch Site Proximity

Launch Site Marker

- ✓ A distinct marks the selected launch site.
- ✓ Includes a popup with site name and coordinates.

Proximity Markers

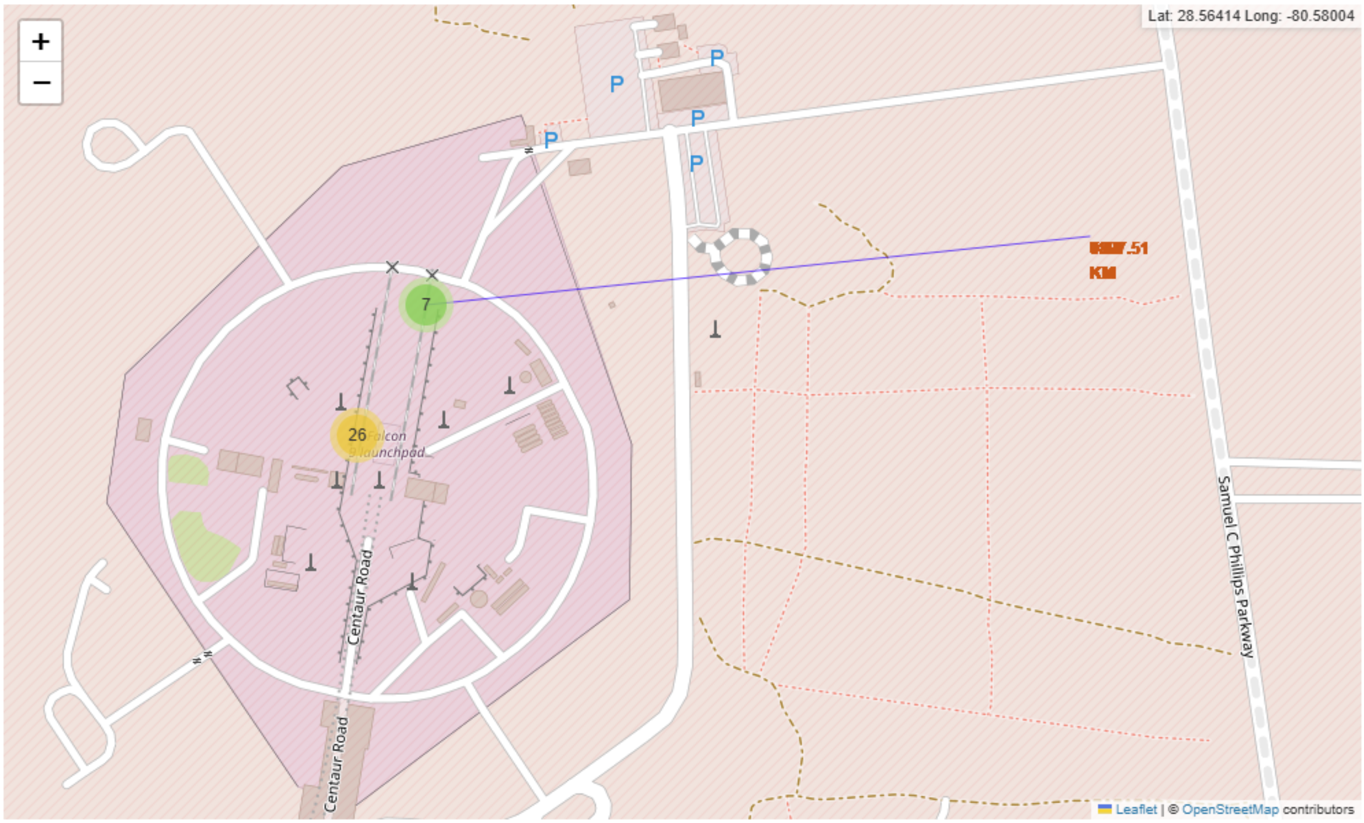
- Railway
- Highway
- Coastline

Distance Labels

- ✓ Each proximity marker includes a line connecting it to the launch site, with a distance and displayed directly on the map.

Map Layers and Zoom

- ✓ The map is zoomed in to show fine-grained spatial relationships.
- ✓ Layer controls may allow toggling between satellite and terrain views.



Section 4

Build a Dashboard with Plotly Dash

Total Success Launches by Site

Pie Chart Segments

✓ Each slice represents a launch site's proportion of total successful launches.

Color Coding

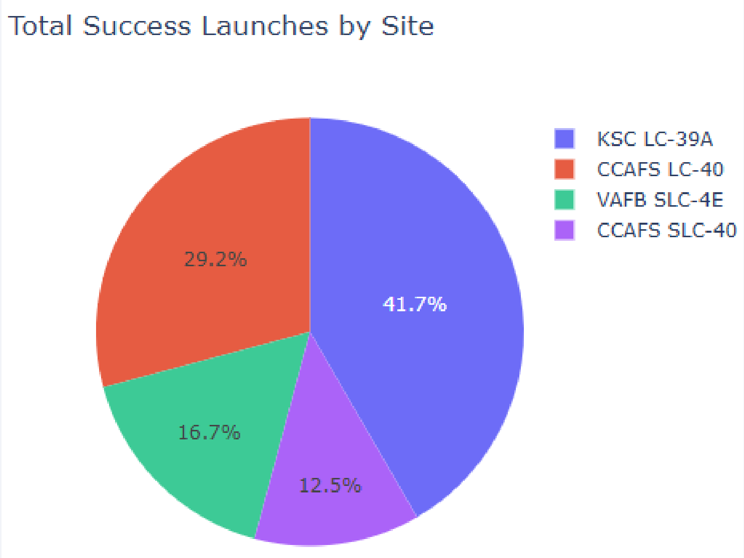
✓ Each site is assigned a distinct color, making it easy to distinguish contributions at a glance.

Percentage Labels

✓ Clear numeric labels on each slice show the relative success rate, not just raw counts—ideal for comparative analysis.

Title and Context

✓ The chart is titled "Total Success Launches by Site", anchoring the viewer in the metric being visualized.



CCAFS SLC-40 Success vs. Failure

Chart Title

✓ Clearly labeled as “Launch Success Ratio – KSC LC-39A”, anchoring the viewer in the scope of analysis.

Segment Breakdown

✓ The pie chart is divided into slices

Color Coding

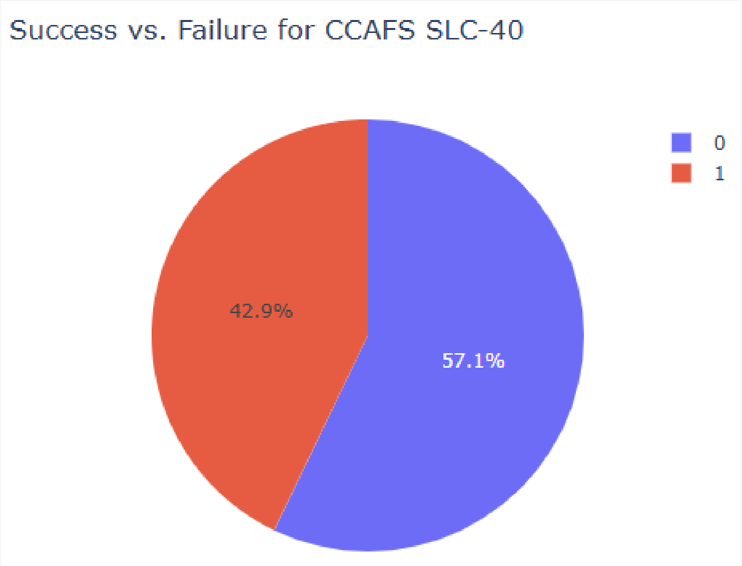
✓ Each slice is color-coded for intuitive understanding

Percentage Labels

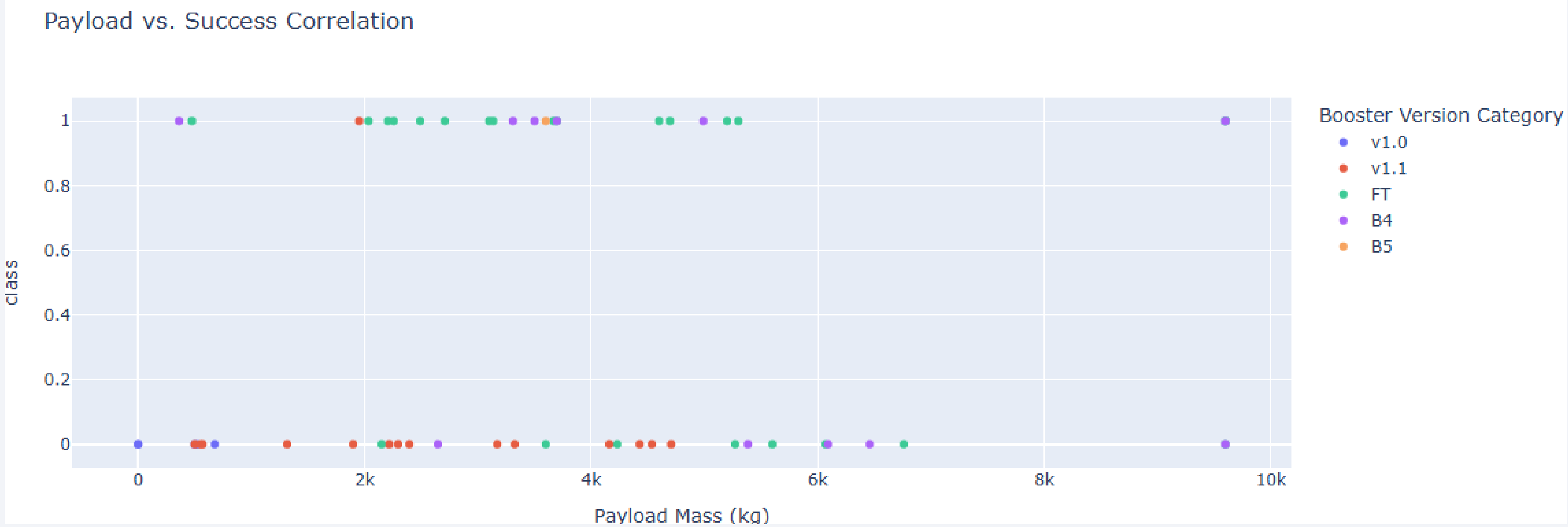
✓ Each slice includes a percentage label, making it easy to compare outcomes.

Legend and Tooltip Interactivity

✓ A legend reinforces the color meanings, and tooltips may show exact counts when hovered over.



Payload vs. Launch Outcome



Key Elements

- ✓ X-axis: Payload Mass (kg), adjustable via range slider
- ✓ Y-axis: Launch Outcome (Success = 1, Failure = 0)
- ✓ Color-coded dots: Booster versions (v1.0, v1.1, FT, B4, B5)
- ✓ Interactive filtering: Focus on specific payload ranges

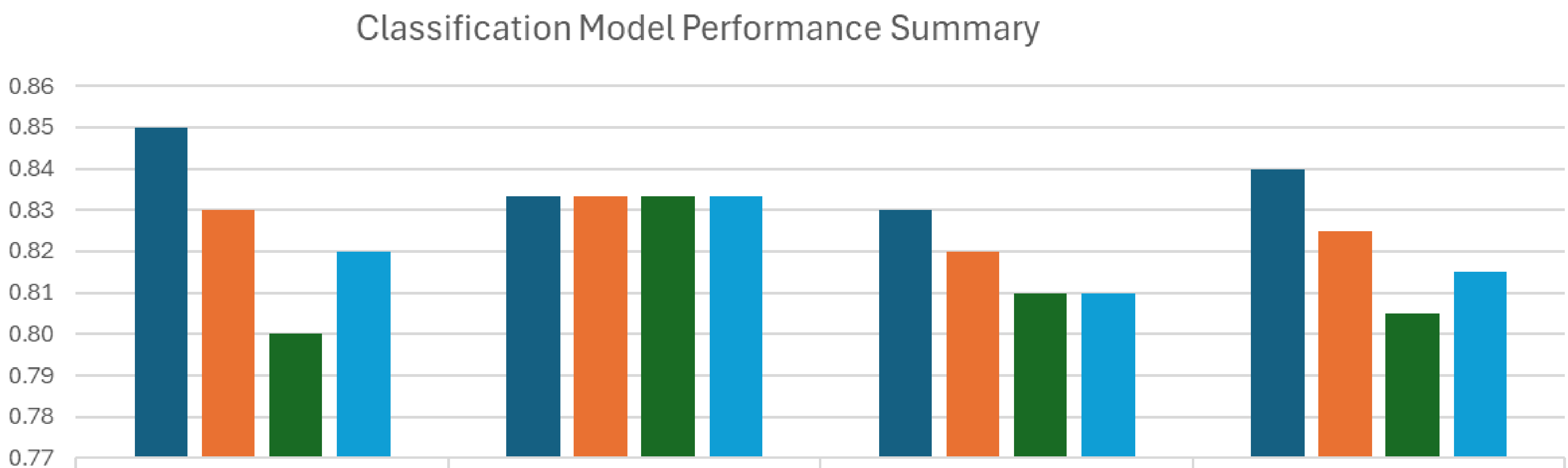
Key Findings

- ❑ Block 5 boosters show highest success, especially in 4,000–6,000 kg range
- ❑ v1.0/v1.1 have more failures at higher payloads
- ❑ Mid-weight payloads (2,000–6,000 kg) yield best outcomes
- ❑ Success improves with newer booster versions

Section 5

Predictive Analysis (Classification)

Classification Accuracy



	Precision	Accuracy	Recall	F1-Score
Logistic Regression	0.85	0.83	0.83	0.84
SVM	0.83	0.83	0.82	0.83
Decision Tree	0.80	0.83	0.81	0.81
KNN	0.82	0.83	0.81	0.82

Best performing model: Logistic Regression

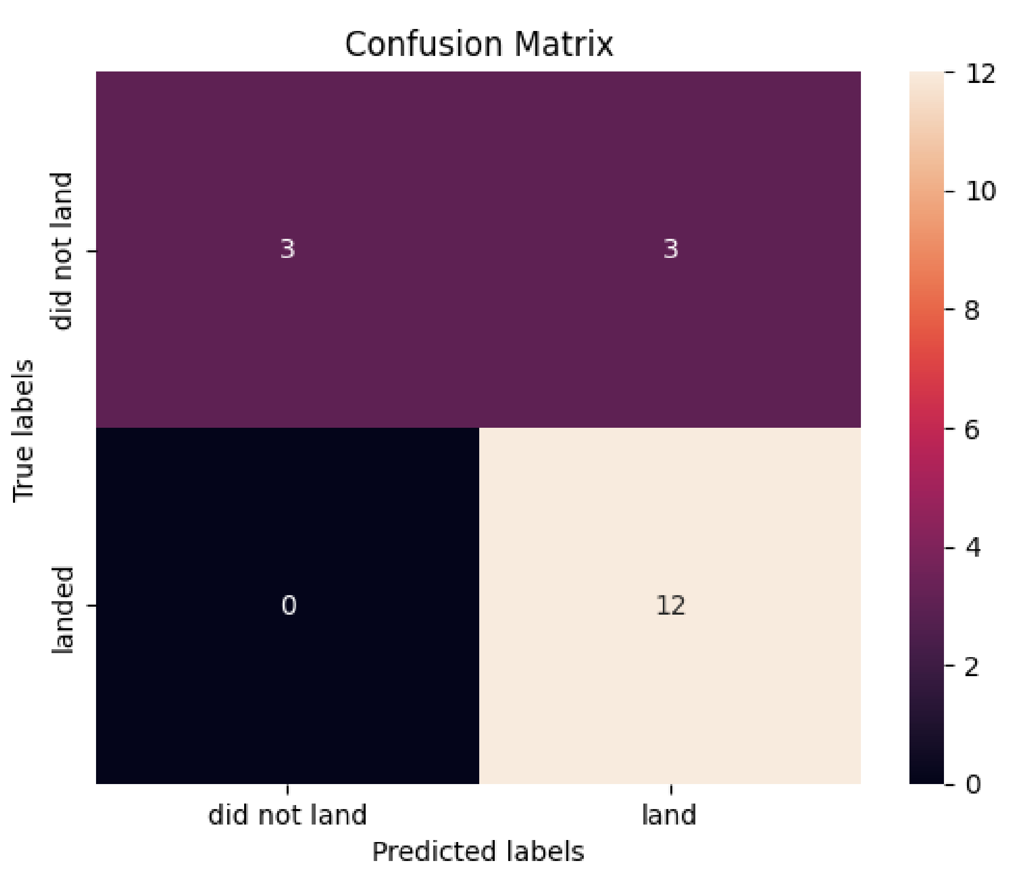
Confusion Matrix

Key Findings

- ✓ High True Positives (12): The model correctly identified 12 successful landings.
- ✓ Zero False Negatives: It never missed a successful landing excellent recall.
- ✓ False Positives (3): It predicted “land” for 3 launches that failed to land.
- ✓ True Negatives (3): It correctly identified 3 failed landings.

Interpretation

- ❑ The model is very strong at identifying successful landings
- ❑ The main issue is false positives—it sometimes predicts a landing when none occurred.
- ❑ This suggests the model is slightly optimistic, favoring “land” predictions.



Conclusions

Landing Prediction is Feasible and Accurate
The Logistic Regression model achieved over 83% accuracy, demonstrating that historical launch data can reliably predict Falcon 9 first-stage landing outcomes.

Key Predictors Identified
Launch site, payload mass, and booster version emerged as the most influential features, guiding future mission planning and hardware deployment.

Reusable Rocket Strategy Validated
The analysis confirms that newer booster versions (especially Block 5) significantly improve landing success, supporting SpaceX's cost-saving reusability goals.

Interactive Tools Enhance Decision-Making
Dashboards and maps built with Plotly Dash and Folium provide intuitive insights for stakeholders, enabling real-time exploration of launch performance and spatial trends.

Data-Driven Planning for Space Missions
This project empowers Space Y with predictive analytics and visual tools to optimize launch strategies, reduce costs, and improve mission reliability.

Appendix

Falcon-Forecast Project

Thank you!