IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Hill
2025-06-21

Outline

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

Executive Summary

Business Problem:
SpaceX advertises Falcon 9 rocket launches at $62 million while competitors cost upward of $165 million. Much of the savings comes from SpaceX's ability to reuse the first stage. If we can predict first stage landing success, we can determine launch costs and enable competitive bidding against SpaceX.

Key Methodologies Applied:

- Data Collection: SpaceX REST API calls and web scraping
- Data Wrangling: landing outcome classification and feature engineering
- Exploritory Data Analysis: Visualization and SQL analysis
- Interactive Analytics: Folium maps and Plotly dashboard
- Predictive Modelling: Classification algorithms training and tuning

Summary of Results:

- Identified key factors influencing landing success rates across launch sites
- Built predictive models achieving 83% accuracy
- Launch site location and payload mass used as predictive features

GitHub Repository: https://github.com/danielmh111/imb-ds-cert-capstone

Introduction

Project Background and Context:

- SpaceX has revolutionized space launches through first stage reusablity, dramatically reducing costs compared to traditional disposable rocket stages. This is their key competative advantage

Poblems We Want to Find Answers To:

- What factors determine first stage landing success?
- Can we predict landing outcomes with high accuracy?
- Which launch sites have the highest success rates?
- How do payload characteristics affect landing probability?
- What is the best model for cost prediction?

Business Impact:

Accurate landing predictions enable competitor cost modelling and strategic bidding decision in the commercial launch market.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Combined SpaceX API integration with web scraping to create a dataset covering rocket specifications, launch sites, payloads, and outcomes
- Perform data wrangling
    - Transformed landing outcomes into binary class labels (1=success, 0=failure) and engineered features from categorical variables including launch site, booster version, and payload characteristics
- 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 and compared multiple classification algorithms (Logistic Regression, SVM, Decision Tree, KNN) with hyperparameter tuning and cross-validation to achieve optimal prediction accuracy

Data Collection

SpaceX REST API Integration:

- Collected launch data using GET requests to SpaceX API endpoints
- Retrieved rocket specifications, launch sites, payloads, and outcomes
- Processed JSON responses into structured pandas DataFrames

Web Scraping Supplement:

- Used BeautifulSoup for additional launch information
- Extracted tabular data from SpaceX mission pages
- Integrated scraped data with API results for comprehensive dataset

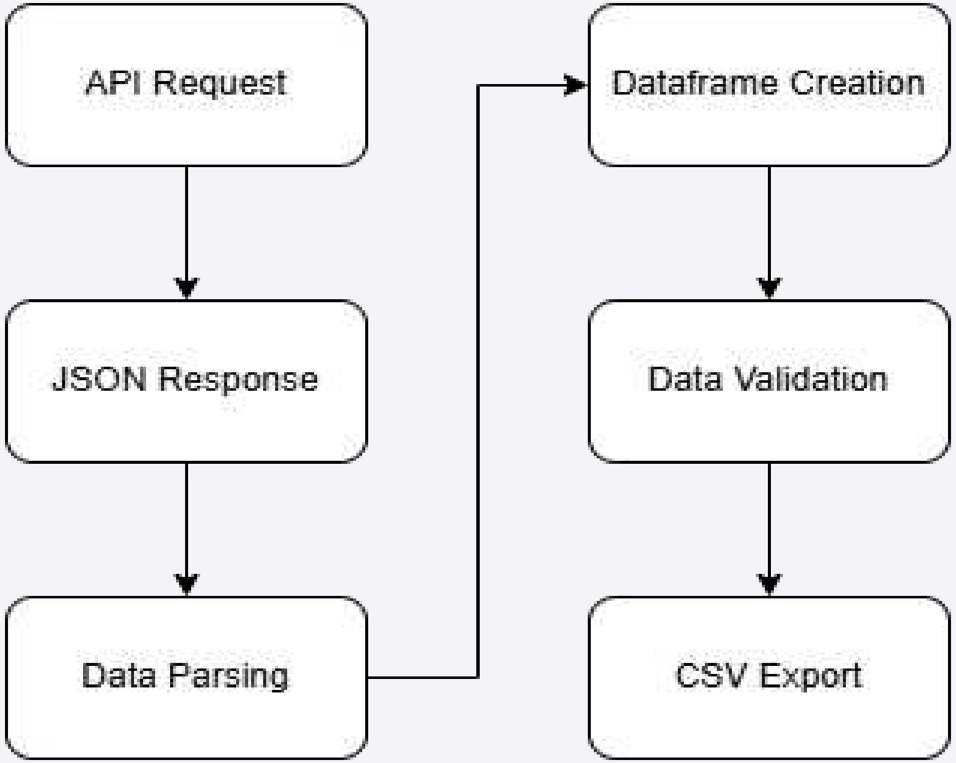
Data Collection – SpaceX API

Process Overview:

- Used SpaceX Rest API to collect launch data
- Extracted rocked specs, launch sites, payloads, and outcomes
- Implemented data validation for quality checks

Key Data Points Collected:

- Flight numbers and dates
- Booster versions
- Launch sites and coordinates
- Payload mass and orbits
- Landing outcomes

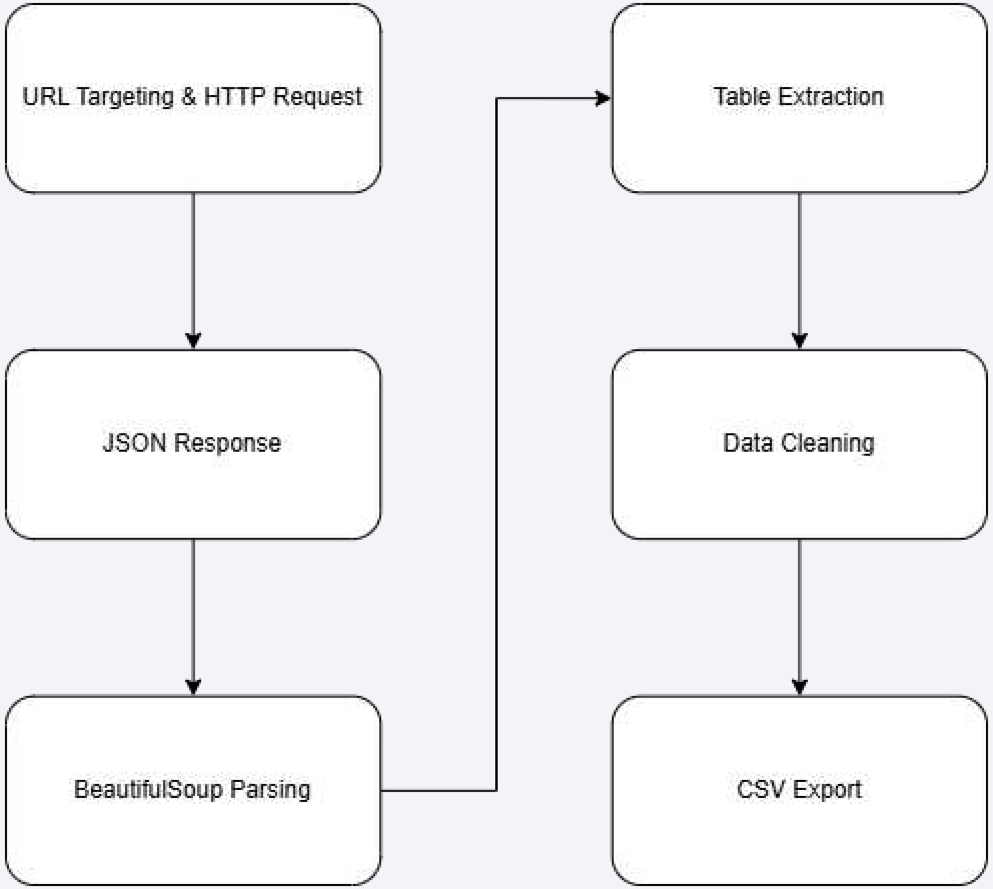


GitHub URL: https://github.com/danielmh111/imb-ds-cert-capstone/blob/main/spacex-data-collection-api.ipynb

Data Collection - Scraping

Process Overview:

- Augmented API data with additional launch information
- Used beautifulsoup for html parsing
- Implemented error handling and data cleaning



Key Data Points Collected:

- Launch Dates
- Booster Versions
- Customers
- Payloads
- Landing Success

GitHub URL:
https://github.com/danielmh111/imb-ds-cert-capstone/blob/main/spacex-data-webscraping.ipynb

Data Wrangling

Approach:

- Created binary classification target variable
- Handled missing values in landing pad data (28.9% missing)
- Engineered features from categorical variables
- Standardized booster version classifications
- Computed overall success rate: 66%

GitHub Reference:

https://github.com/danielmh111/imb-ds-cert-capstone/blob/main/spacex-data-wrangling.ipynb

EDA with Data Visualization

Visualisations:

- Scatter plots to the relationship between Flight Number and Launch Site and between Payload Mass and Launch Site
- Bar chart for showing the success of each of each orbit
- Line graph for showing launch success yearly trend

GitHub Reference: https://github.com/danielmh111/imb-ds-cert-capstone/blob/main/eda-viz.ipynb

EDA with SQL

Query Summary:

1) Unique Launch Sites:
   - 4 primary launch sites identified (CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40)
2) CCAFS Launch Records:
   - CCAFS sites handle majority of commercial and ISS missions
3) NASA CRS Payload Analysis:
   - Total NASA CRS payload mass = 45,596 kg across multiple missions
4) Booster Version Performance:
   - F9 v1.1 average payload capacity demonstrated incremental improvements

Build an Interactive Map with Folium

Map Objects Created:

- Markers: Colour-coded markers for each launch site with custom labels showing site names
- Circles: Highlighted areas around launch sites for geographic reference
- Lines: Distance measurement lines connecting launch sites to nearby features
- Marker Clusters: Grouped markers for better visualization when zoomed out
- MousePosition Plugin: Interactive coordinate tracking for precise location identification

Why These Objects Were Added:

- Success/Failure Markers: Enable quick visual identification of which launch sites have higher success rates
- Proximity Lines: Reveal geographic advantages - all sites positioned near coastlines for safety, with varying distances to infrastructure
- Distance Measurements: Support analysis of whether launch site location correlates with success rates based on proximity to transportation and population centers

GitHub Reference: https://github.com/danielmh111/imb-ds-cert-capstone/launch_site_locations.ipynb

Build a Dashboard with Plotly Dash

Plots and Interactions Added:

- Dropdown Menu: Site selection filter (All Sites, CCAFS LC-40, etc.)
- Pie Charts: Success distribution visualization that updates based on site selection
- Range Slider: Payload mass filter (0-10,000 kg) with 1000 kg increments
- Interactive Scatter Plot: Payload mass vs. launch success with booster version color coding

Why These Features Were Added:

- Site-Specific Analysis: Dropdown enables comparison of individual launch site performance vs. aggregate performance
- Payload Correlation: Scatter plot reveals payload ranges for mission success
- Real-time Filtering: Combined interactions allow users to explore specific payload ranges at specific sites to identify performance patterns
- Booster Evolution Tracking: Colour coding shows how different booster versions perform across payload ranges

GitHub Reference: https://github.com/danielmh111/imb-ds-cert-capstone/spacex-dash-app.py

Predictive Analysis (Classification)

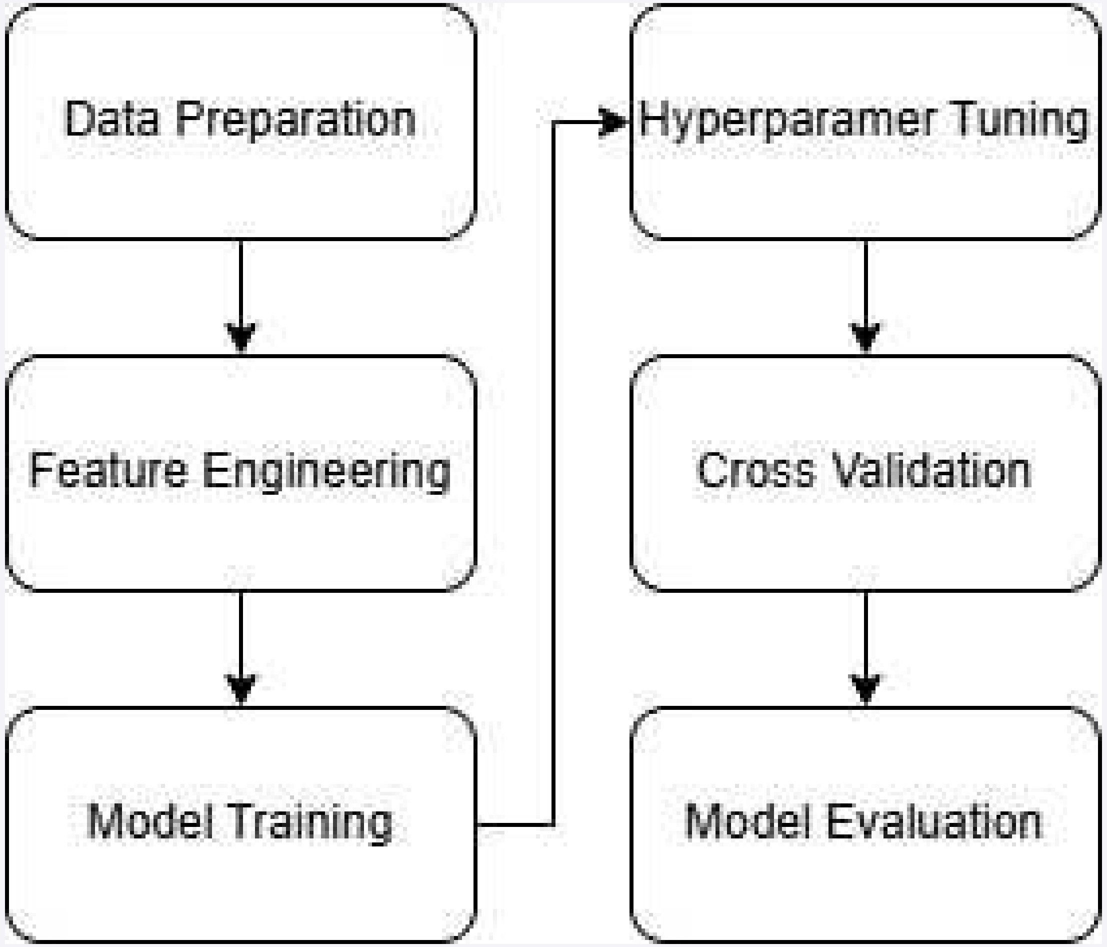
Process Summary:

- Built: Four classification algorithms
- Evaluated: 10-fold cross-validation for robust assessment
- Improved: hyperparameter optimisation for each algorithm to maximize accuracy
- Selected: Decision Tree as best performer (87.52% CV accuracy) despite identical test set performance

Key Process Steps:

- Feature Engineering: Binary encoding of categorical variables
- Data Splitting: 80/20 train-test split with stratification
- Hyperparameter Tuning: Grid search across multiple parameter combinations
- Performance Validation: Cross-validation to ensure model generalisability
- Final Testing: Hold-out test set evaluation for unbiased performance assessment

GitHub Reference:
https://github.com/danielmh111/imb-ds-cert-capstone/
SpaceX_Machine_Learning_Prediction_Part_5.ipynb



Results

Exploratory Data Analysis Results:

- Launch Site Performance: KSC LC-39A achieved highest success rate (41.7%)
- Payload Impact: Best performance in 3k-5k kg range; high failure rates above 6,000kg
- Booster Evolution: Clear progression from experimental v1.0/v1.1 (high failure rates) to operational FT/B4/B5 (consistent success)

Interactive Analytics Demo Results:

- Geographic Insights: All launch sites positioned near coastlines, Florida sites leverage Earth's rotation for orbital efficiency
- Dashboard Discoveries: Payload mass and launch location shown as key factors for mission success
- Performance Patterns: Site-specific analysis revealed VAFB optimized for polar orbits while Florida sites handle commercial/ISS missions

Predictive Analysis Results:

- Model Accuracy: 83.33% prediction accuracy achieved across all four classification models
- Business Impact: Perfect recall (100%) for successful landings ensures no missed cost-saving opportunities
- Key Limitation: Small test dataset (18 samples) prevented meaningful model differentiation despite cross-validation differences

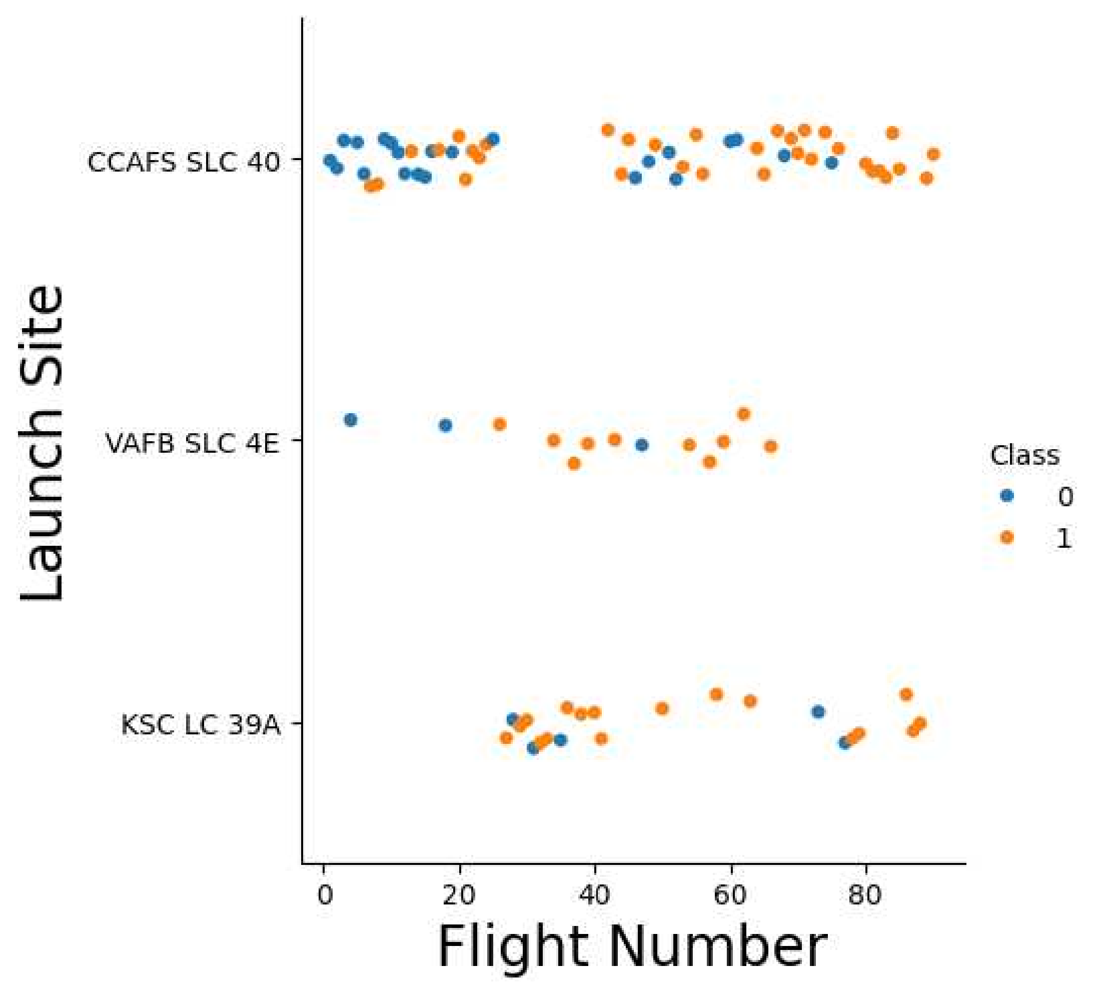
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site

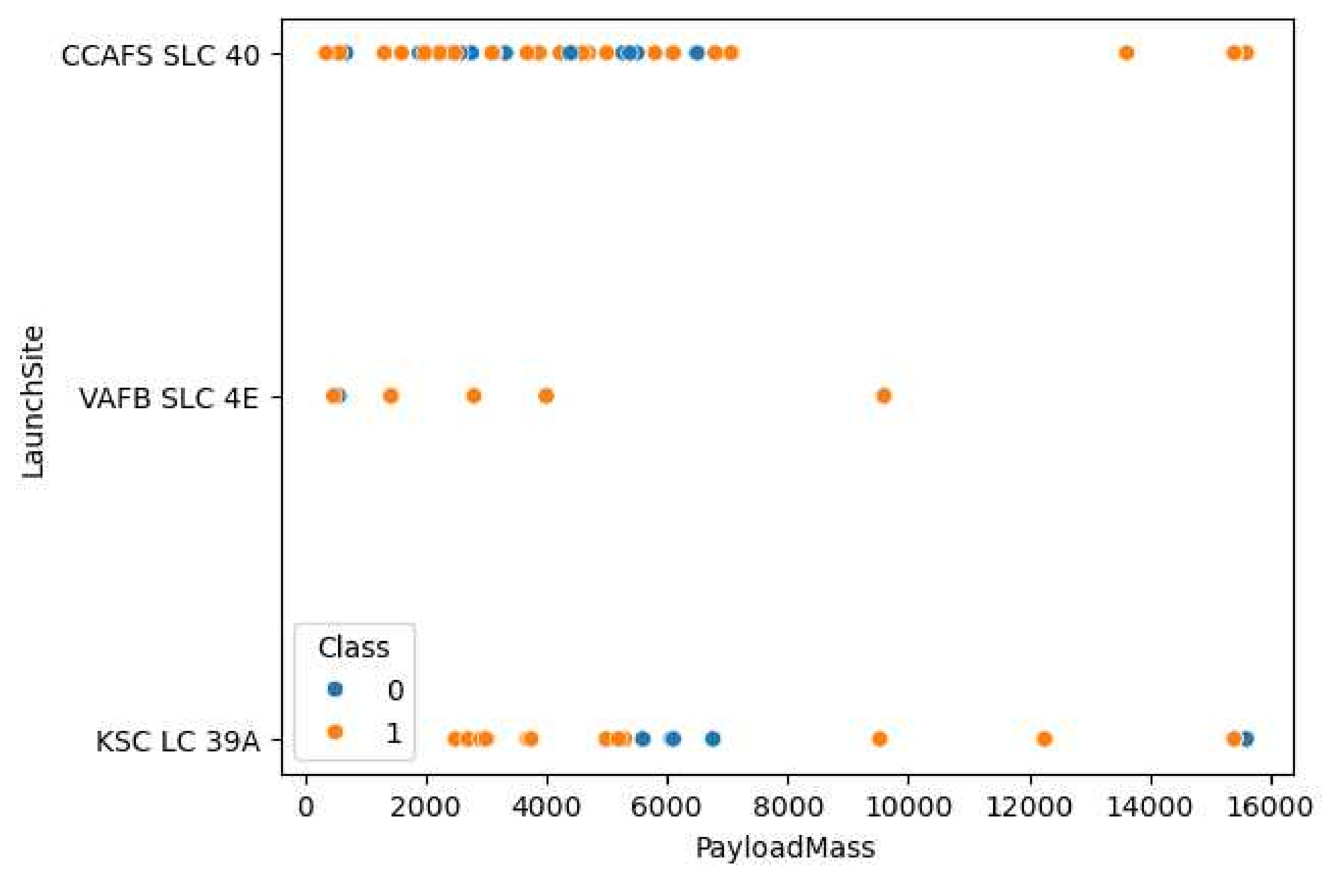
- Show the screenshot of the scatter plot with explanations



Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

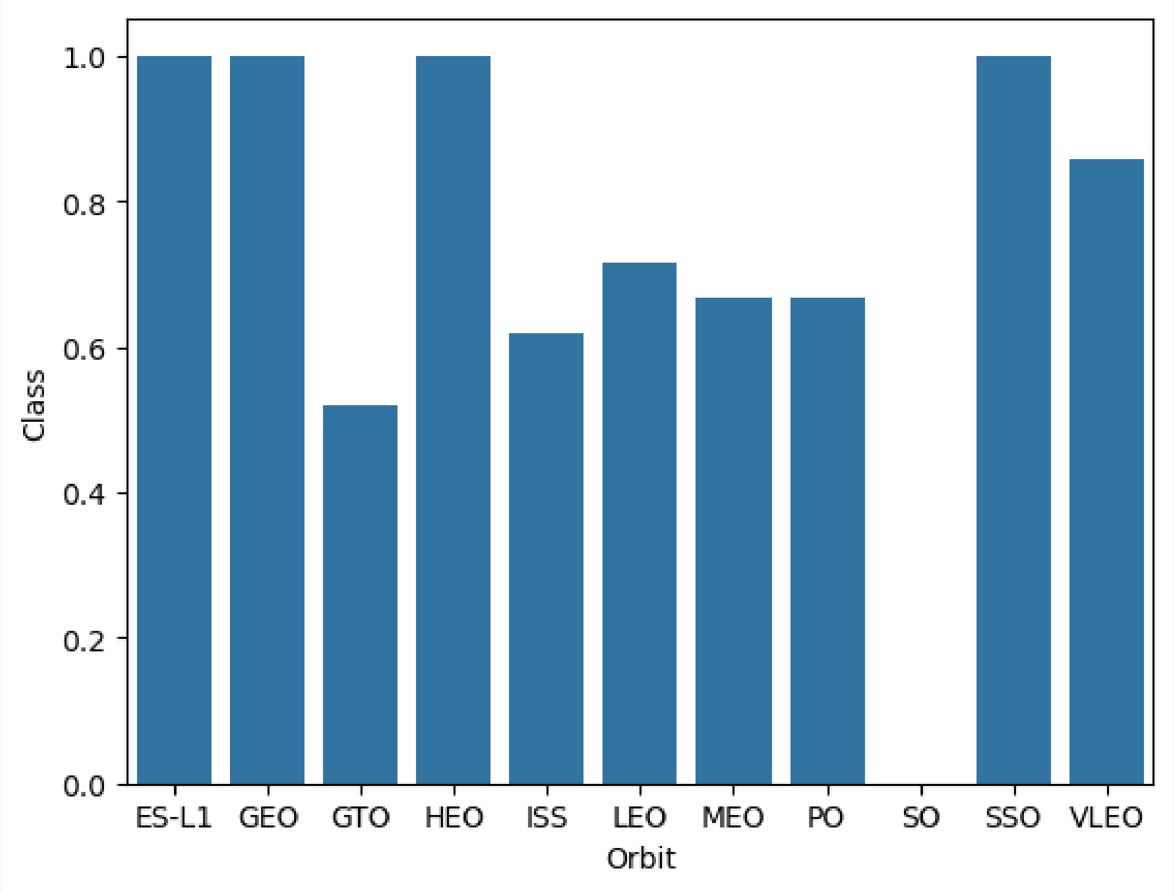
- Show the screenshot of the scatter plot with explanations



Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

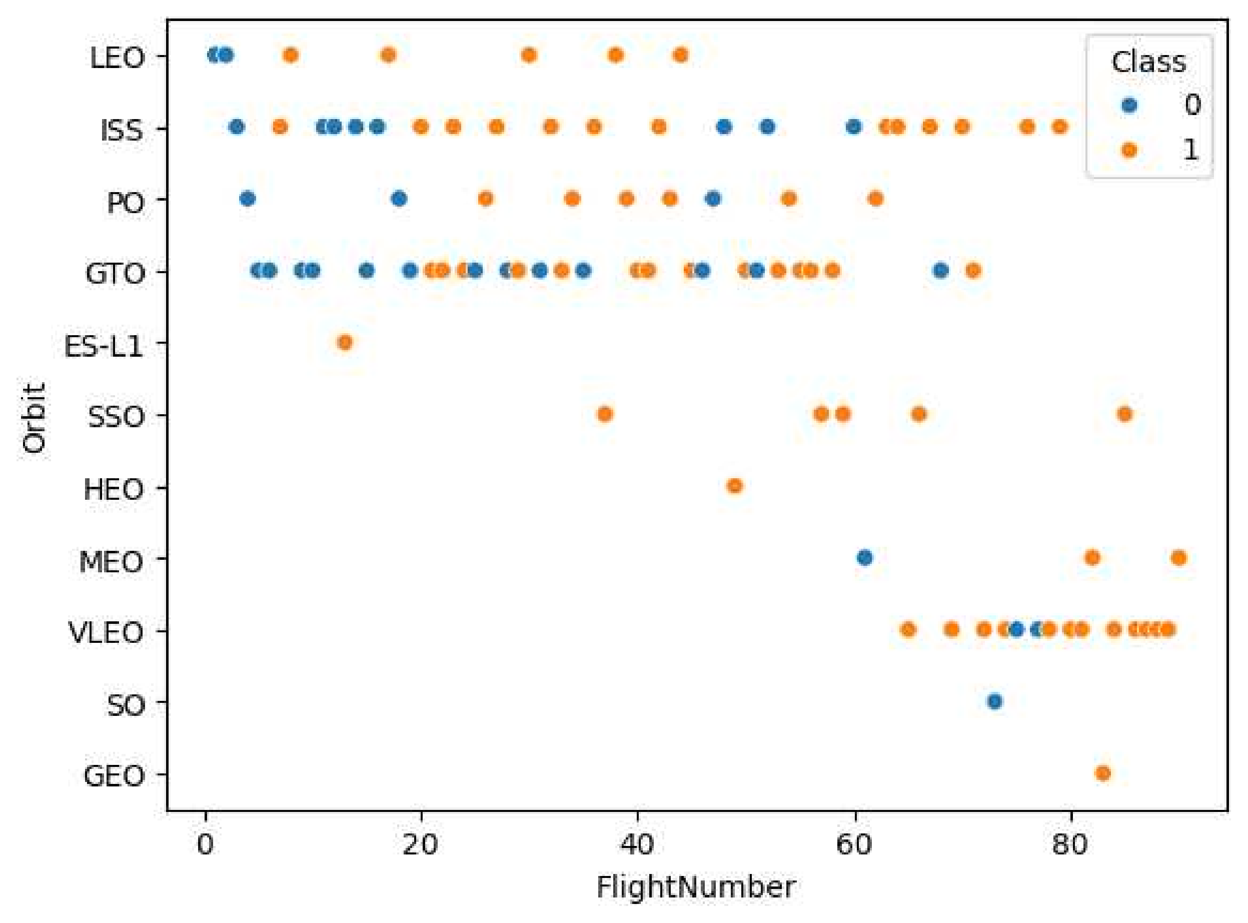
- Show the screenshot of the scatter plot with explanations



Flight Number vs. Orbit Type

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

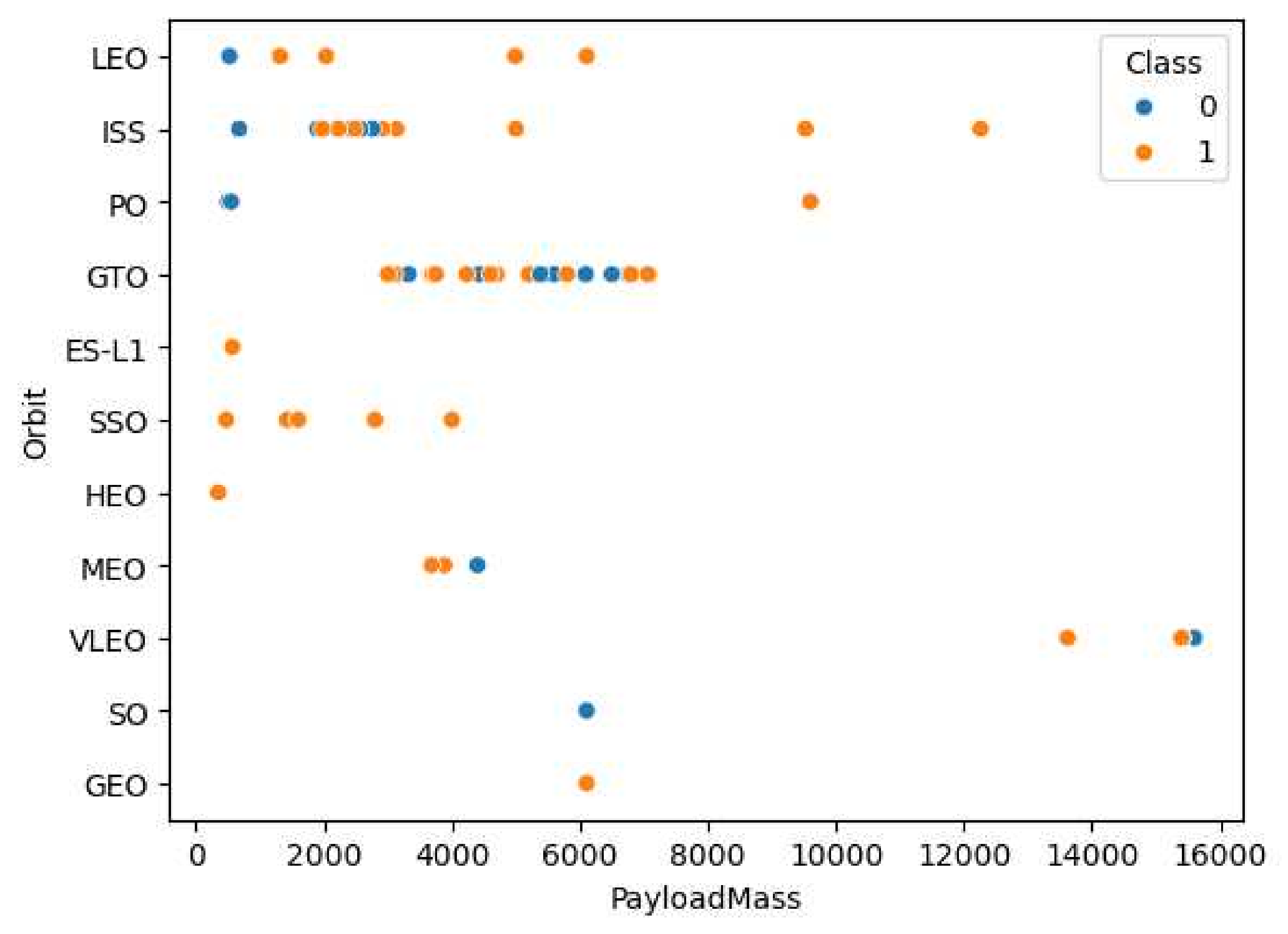
- Show the screenshot of the scatter plot with explanations



Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

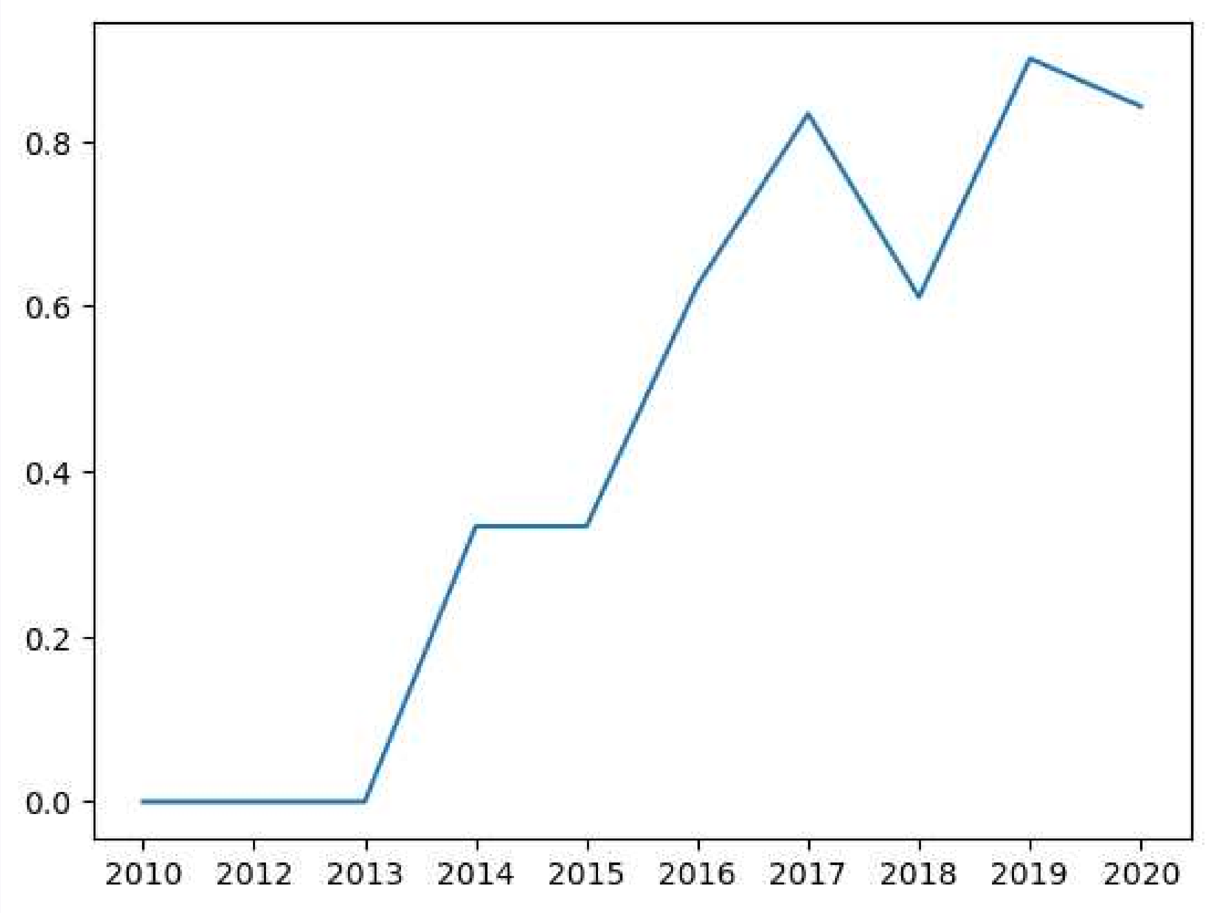
- Show the screenshot of the scatter plot with explanations



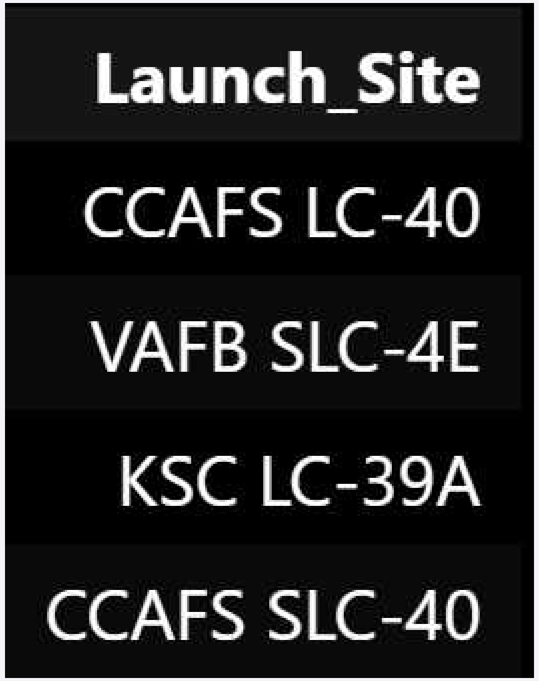
Launch Success Yearly Trend

- Show a line chart of yearly average success rate

- Show the screenshot of the scatter plot with explanations



All Launch Site Names



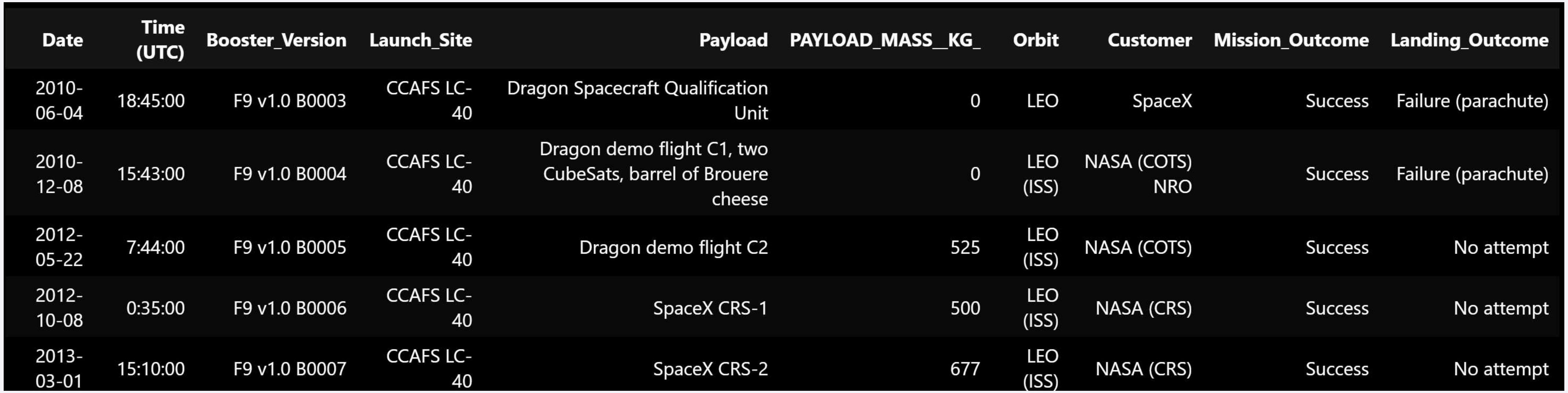
Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Four primary launch sites identified: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Geographic distribution reveals strategic positioning - Florida sites (CCAFS/KSC) leverage Earth's eastward rotation for efficient orbit insertion, while California site (VAFB) enables polar and sun-synchronous orbits. This site diversity provides SpaceX operational flexibility across mission types.

Launch Site Names Begin with 'CCA'

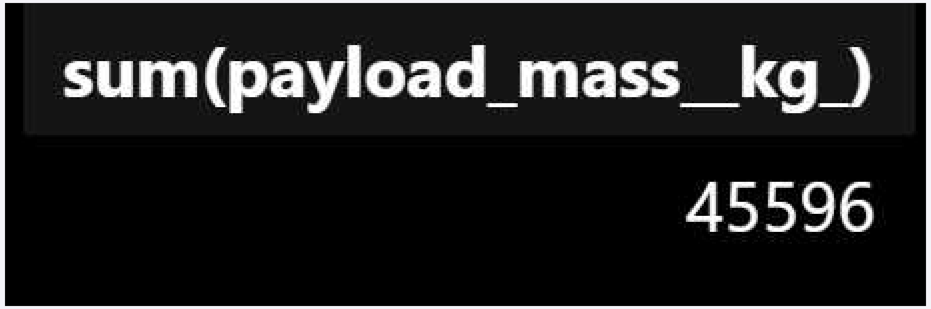
Cape Canaveral Air Force Station (CCAFS) operates two launch complexes: LC-40 and SLC-40, handling multiple mission types from commercial satellites to ISS resupply.



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

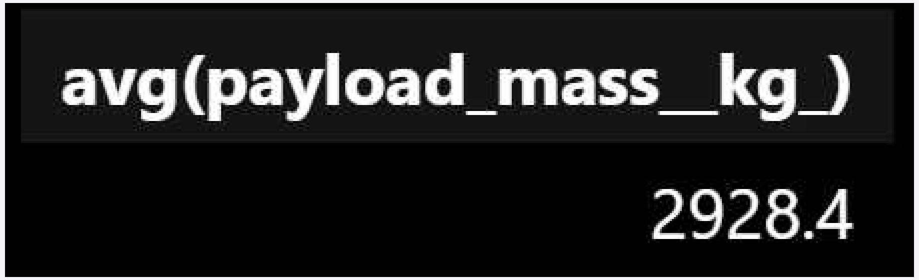
NASA CRS (Commercial Resupply Services) missions delivered 45,596 kg total payload mass across multiple flights. This represents significant NASA partnership value.



sum(payload_mass__kg_)
45596

Average Payload Mass by F9 v1.1

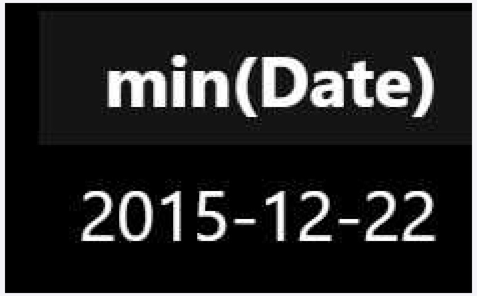
Falcon 9 version 1.1 carried an average payload mass of 2928 kg, representing improved capability over earlier versions.



avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

22nd of December 2015 marked SpaceX's first successful Return-to-Launch-Site (RTLS) landing, a historic aerospace achievement.

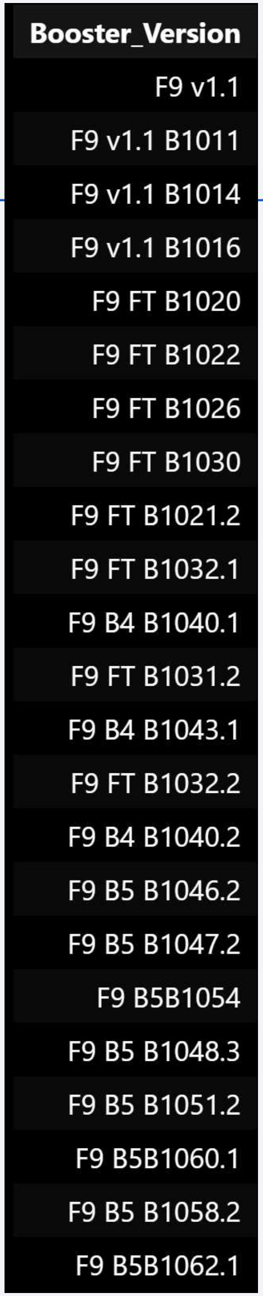


min(Date)
2015-12-22

Successful Landing with Payload between 4k and 6k

Multiple boosters successfully completed drone ship landings within the 4000-6000kg payload range, including various Falcon 9 versions (v1.1, FT, B4, B5).

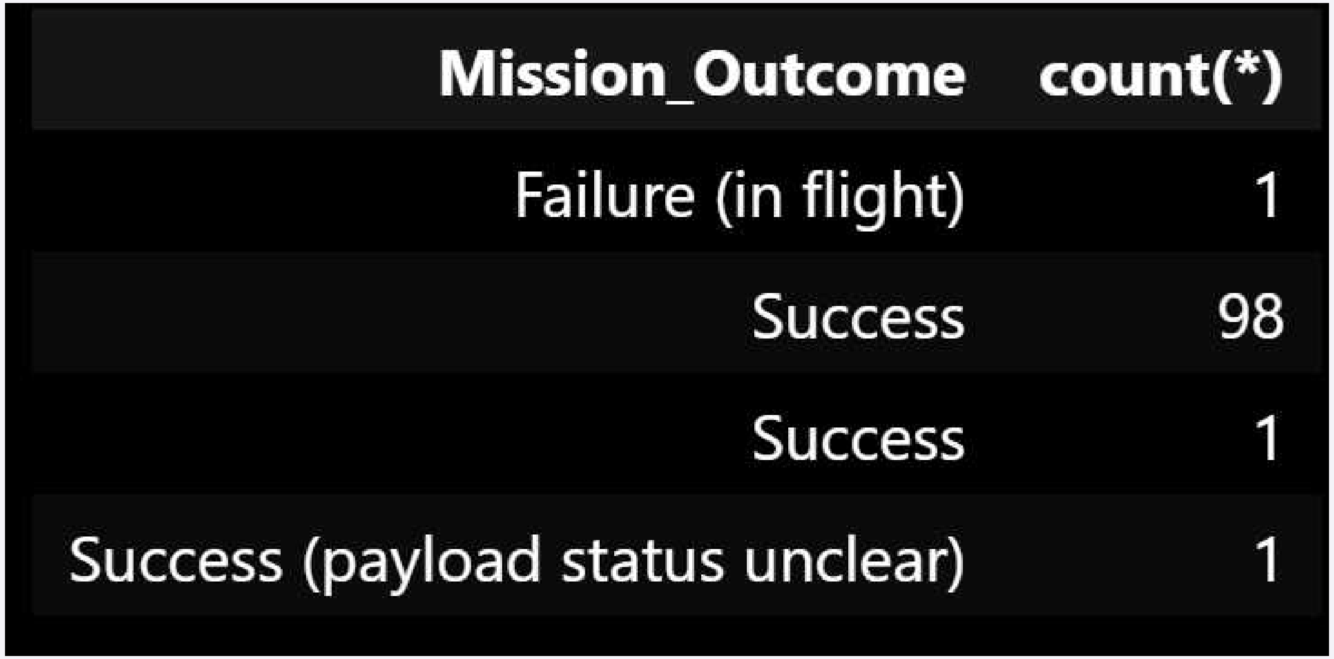
This range of payload masses is significant because it is a sweet spot for providing mission value and managing fuel for the landing.



Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

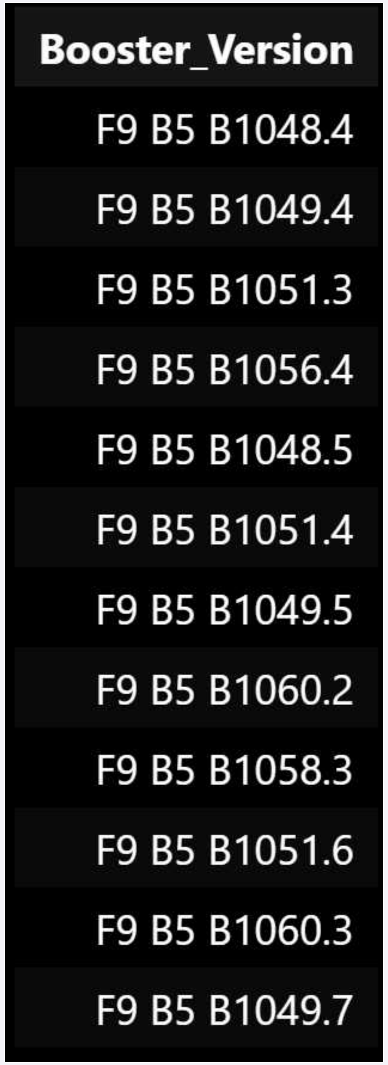
Mission success rate of approximately 99%, demonstrating exceptional mission reliability.



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

Boosters Carried Maximum Payload

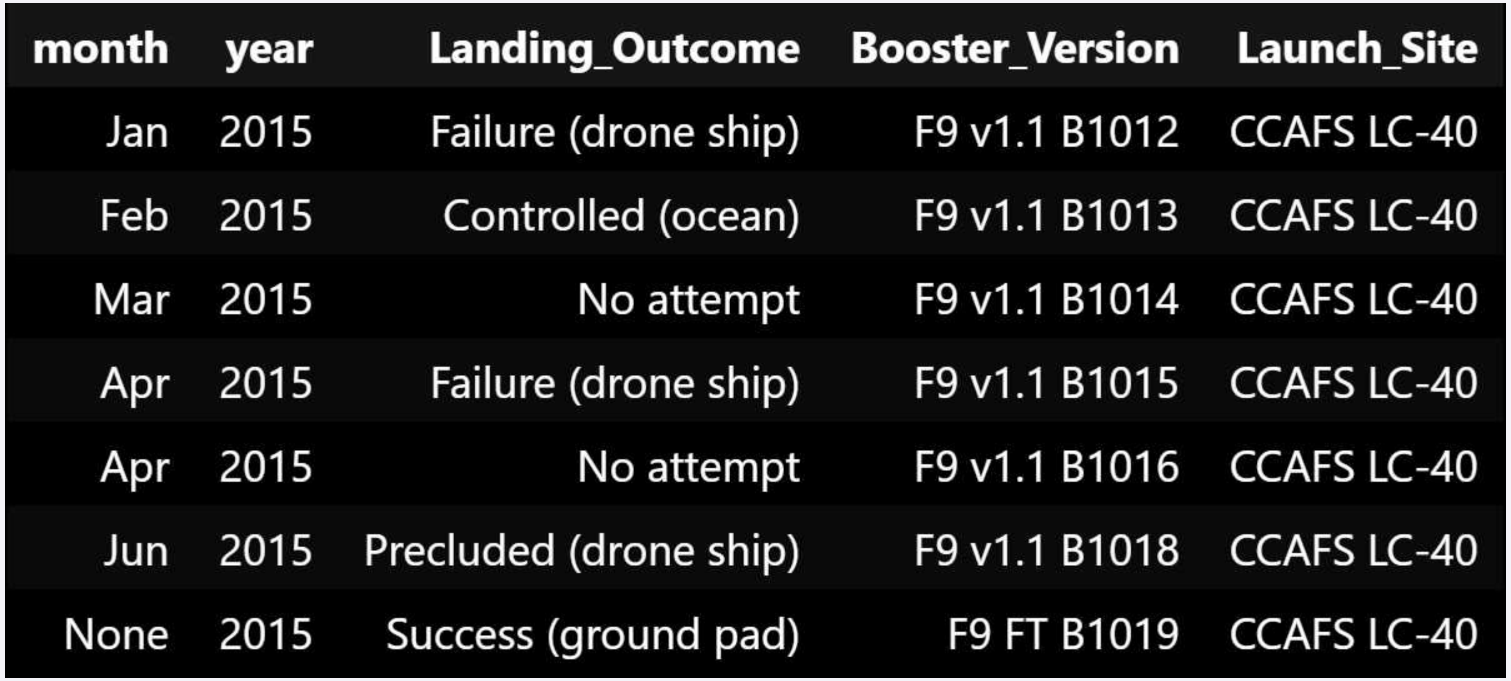
Multiple Falcon 9 booster versions have carried the maximum payload value.



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

2015 marked an important year with both breakthrough successes and significant failures, including first successful landings and the CRS-7 mission failure.

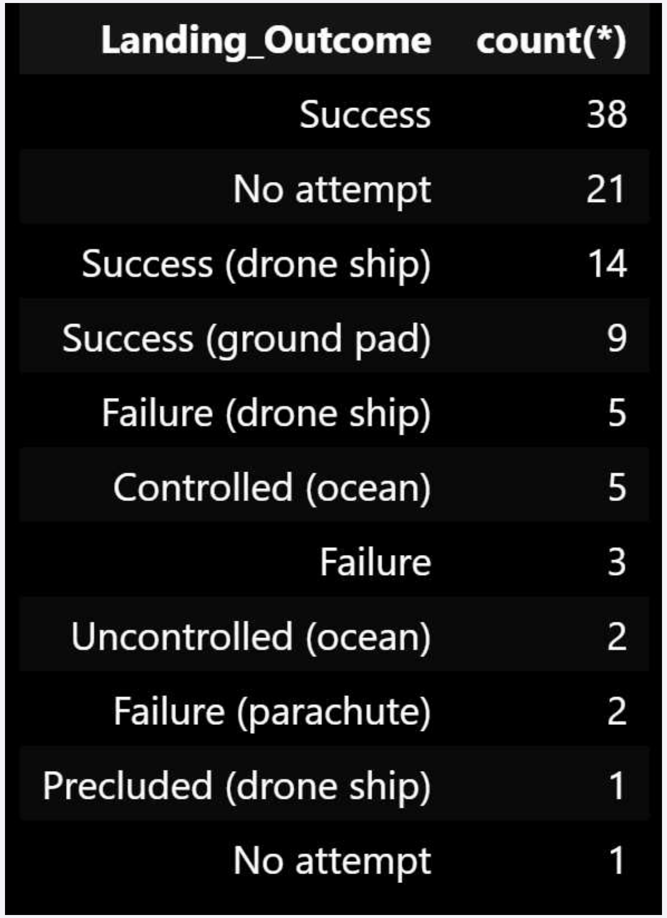


month	year	Landing_Outcome	Booster_Version	Launch_Site
Jan	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Feb	2015	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
Mar	2015	No attempt	F9 v1.1 B1014	CCAFS LC-40
Apr	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
Apr	2015	No attempt	F9 v1.1 B1016	CCAFS LC-40
Jun	2015	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
None	2015	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

During this 7-year period: 38 general successes, 21 no attempts, 14 drone ship successes, 9 ground pad successes, with various failure modes totaling significantly fewer occurrences.

The data shows clear progression from "no attempt" landings in early years to successful recovery methods by 2017. This timeline reveals the development maturity curve, indicating when landing technology became reliable enough.



Landing_Outcome	count(*)
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Section 3

Launch Sites Proximities Analysis

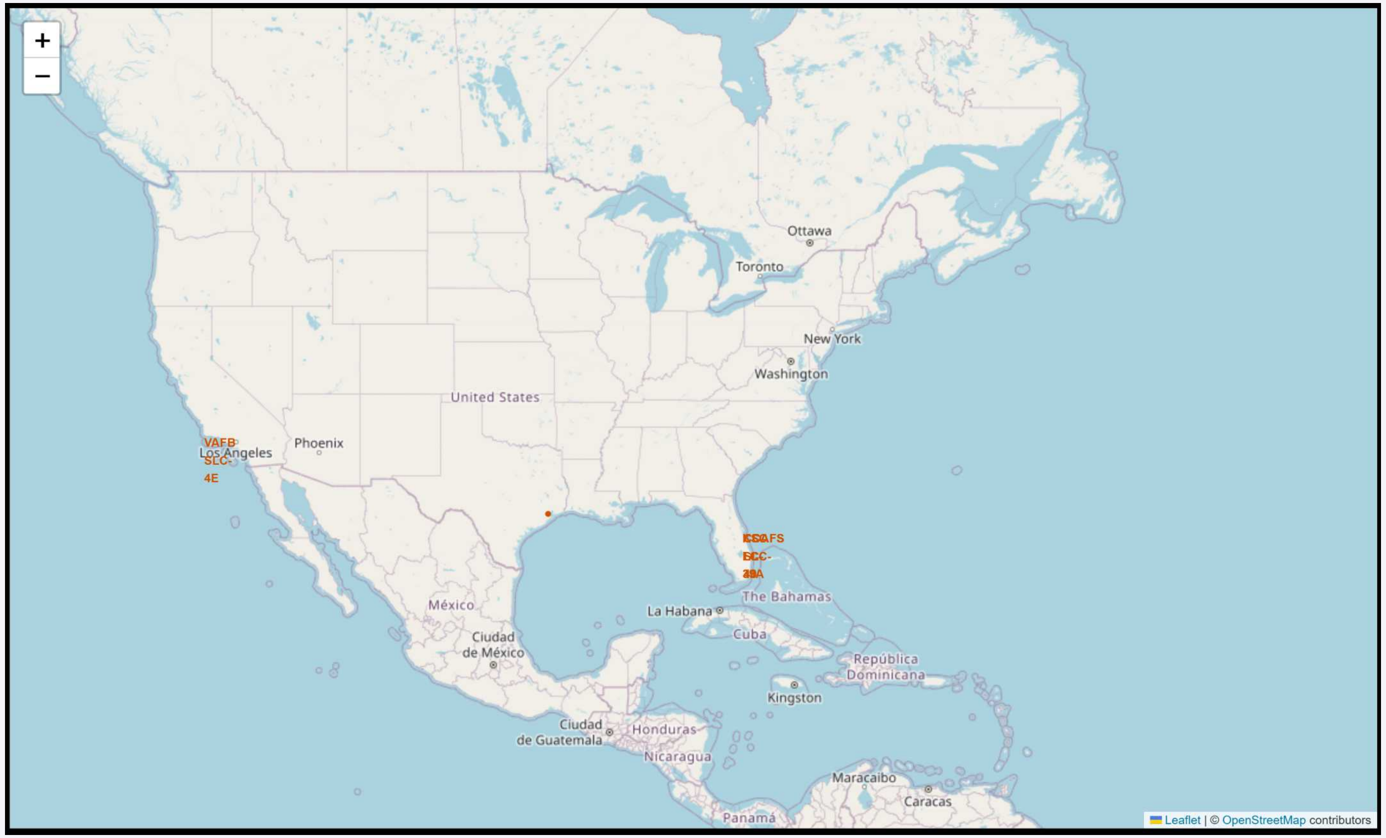
SpaceX Launch Sites - Strategic Geographic Distribution

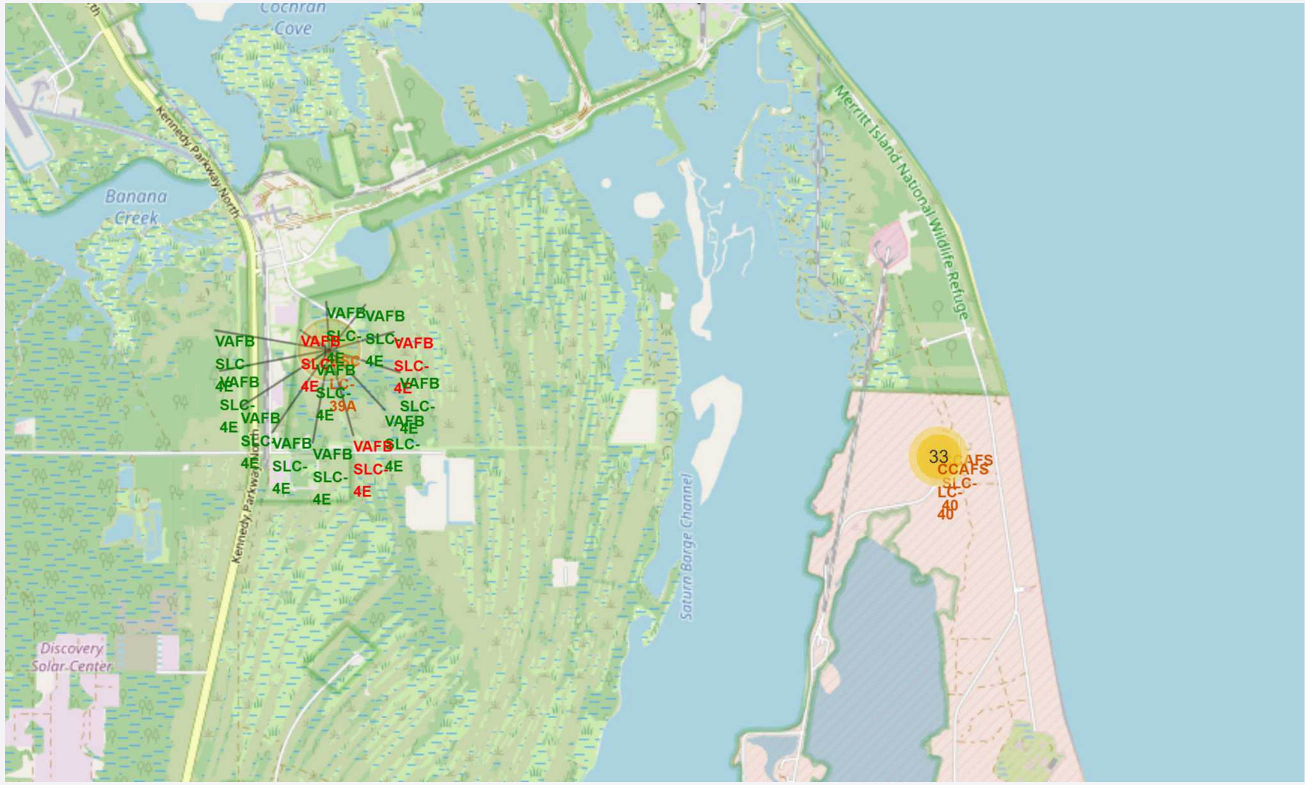
This global map displays all four SpaceX launch sites with clear geographic positioning:

- **VAFB SLC-4E** (California, West coast)
- **CCAFS LC-40** and **KSC LC-39A** (Florida, East Coast)
- **CCAFS SLC-40** (Florida, East Coast)

Key Geographic Advantages:

- All sites positioned near coastlines for safety
- Florida sites leverage Earth's rotation for efficient orbits
- VAFB enables polar/sun-synchronous missions - Multiple Florida sites provide operational flexibility




Cochran Cove
Banana Creek
Kennedy Parkway North
Merritt Island National Wildlife Refuge
Saturn Barge Channel
Discovery Solar Center
VAFB SLC-4E
LC-39A
33
CCAFS SLC-40

<Folium Map Screenshot 3>

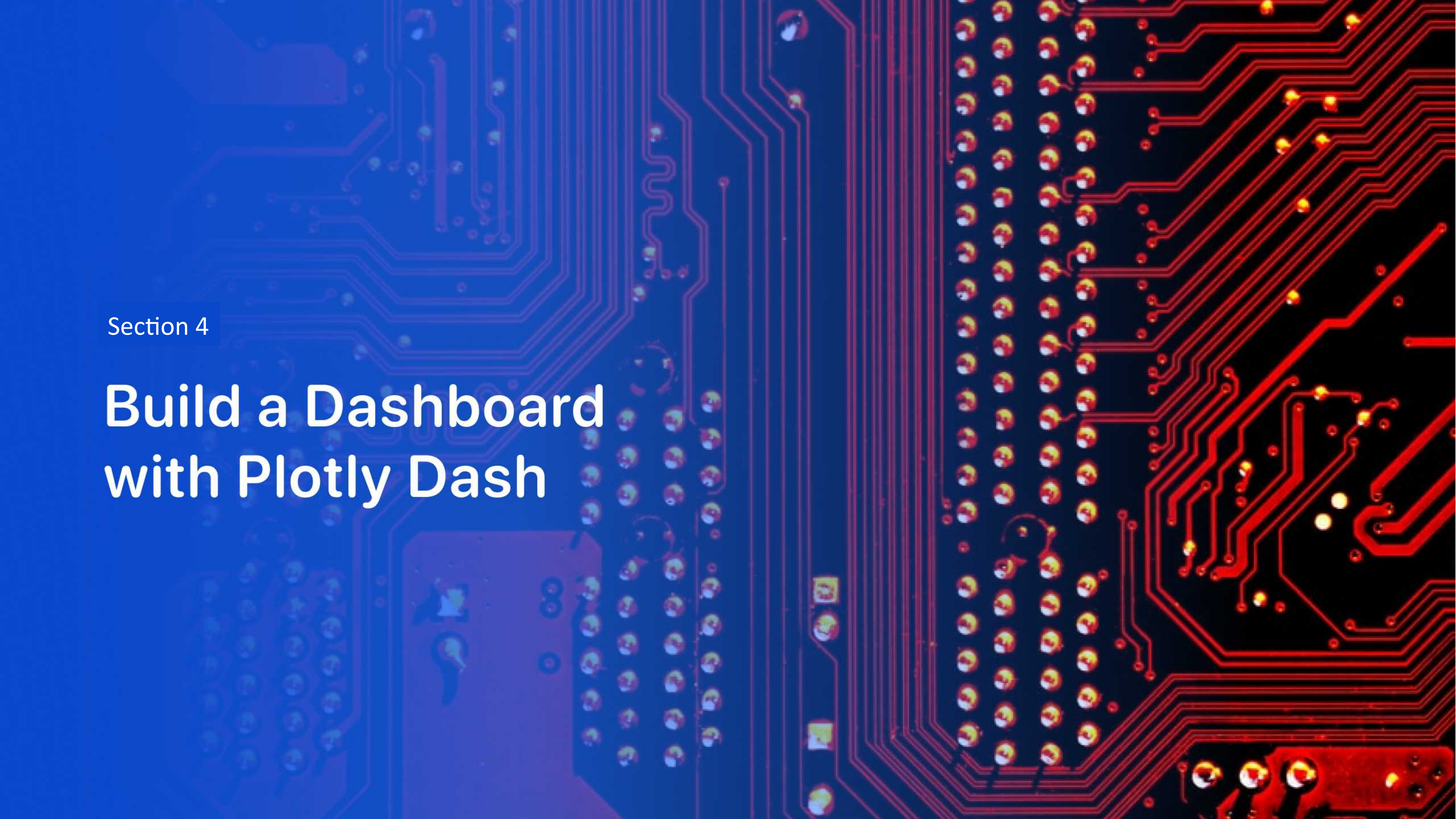
Replace <Folium map screenshot 3> title with an appropriate title

Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

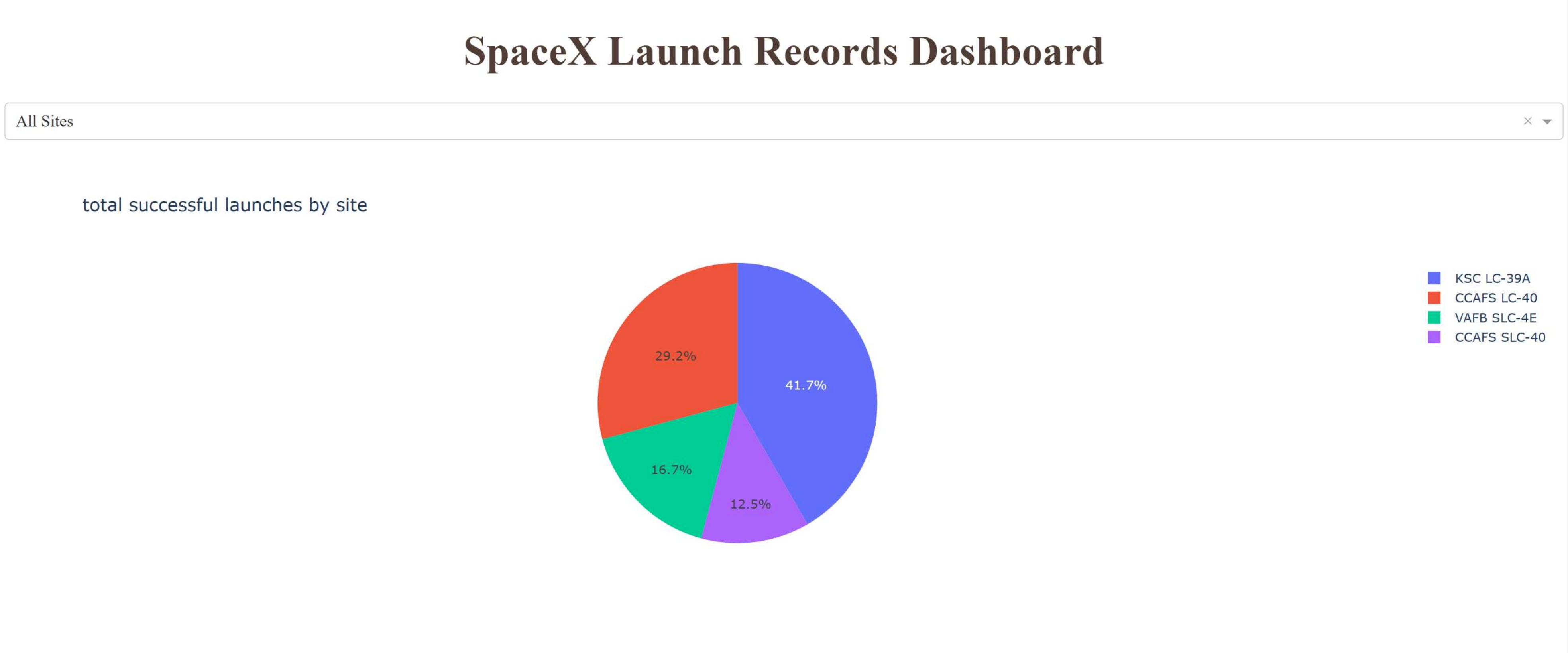
Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash



Launch Success Distribution Across SpaceX Launch Sites

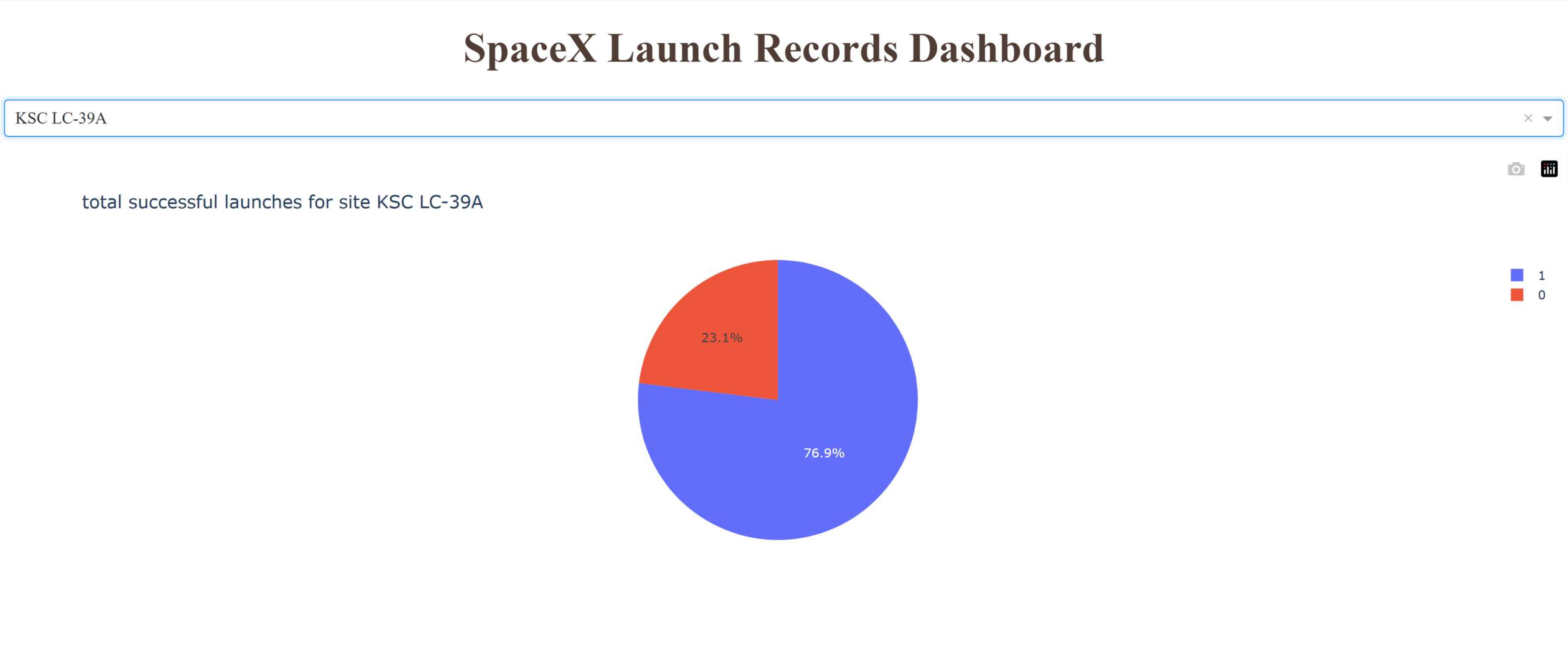


SpaceX Launch Records Dashboard

All Sites

Interactive pie chart displaying the distribution of total successful launches across all four SpaceX launch sites. KSC LC-39A demonstrates the highest success concentration at 41.7%, followed by CCAFS LC-40 at 29.2%. The visualization enables stakeholders to quickly identify which launch sites contribute most significantly to SpaceX's overall mission success rate, supporting site-specific performance analysis for predictive modeling.

KSC LC-39A Launch Success Rate Analysis



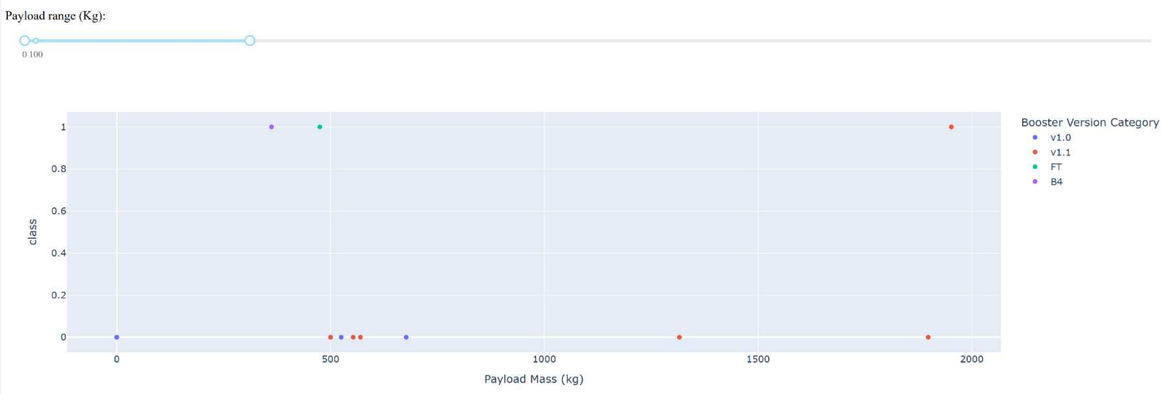
SpaceX Launch Records Dashboard

KSC LC-39A

total successful launches for site KSC LC-39A

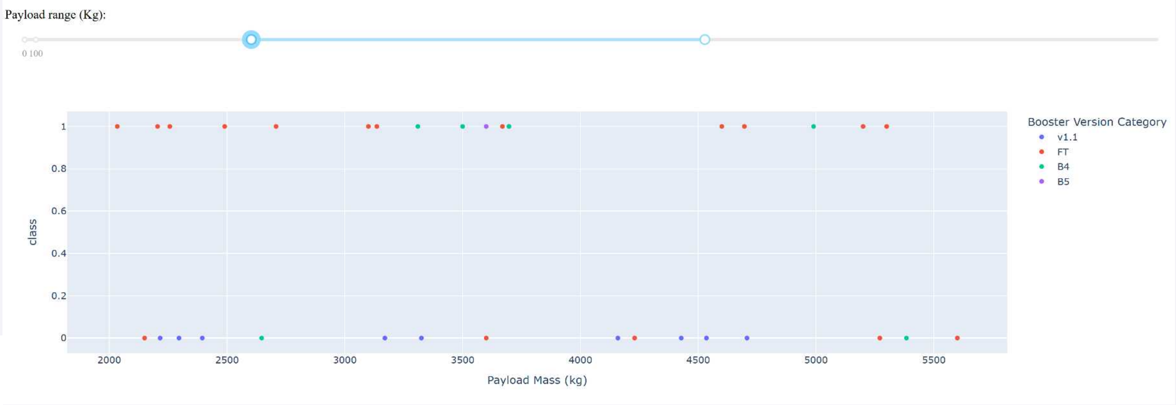
Performance analysis for Kennedy Space Center Launch Complex 39A shows a 76.9% mission success rate, significantly exceeding the SpaceX fleet average of 66%. With only 23.1% mission failures, KSC LC-39A emerges as SpaceX's most reliable launch facility.

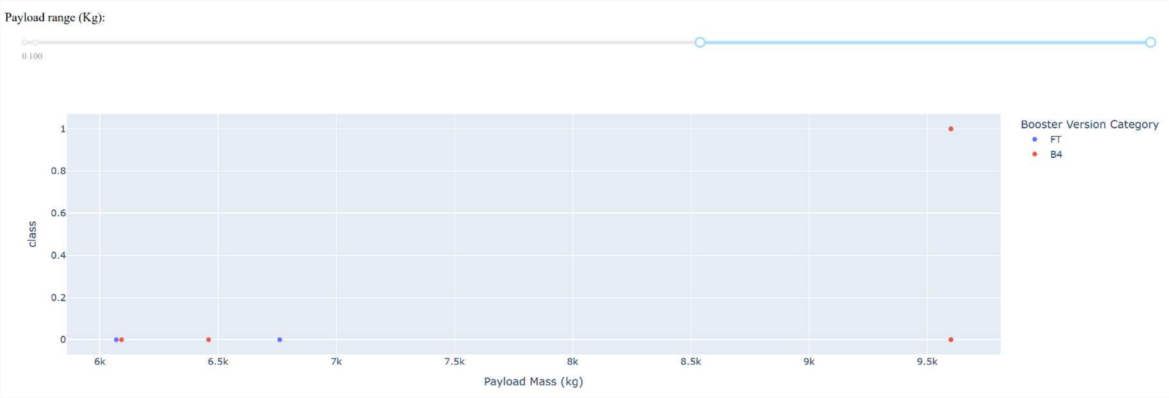
Payload Mass Impact on Launch Success Across Mission Profiles



0 - 2,000kg Range: Low payload missions demonstrate SpaceX's learning curve with early v1.0 and v1.1 boosters showing high failure rates (class=0). Success emerges with FT booster introduction, marking the transition from experimental to operational capabilities.

2,000 - 7,000kg Range: Mid-to-high payload missions showcase SpaceX's operational maturity with consistently high success rates (class=1) across FT, B4, and B5 boosters. The 3,000-5,000kg corridor represents peak performance reliability for commercial operations.





6,000 - 9,500kg Range: High-mass missions reveal performance limits with more failures than successes. Limited sample size means these represent SpaceX's maximum capacity missions, where payload constraints significantly impact landing success probability. This sparse data suggests heavy payloads remain a challenging frontier.

Section 5

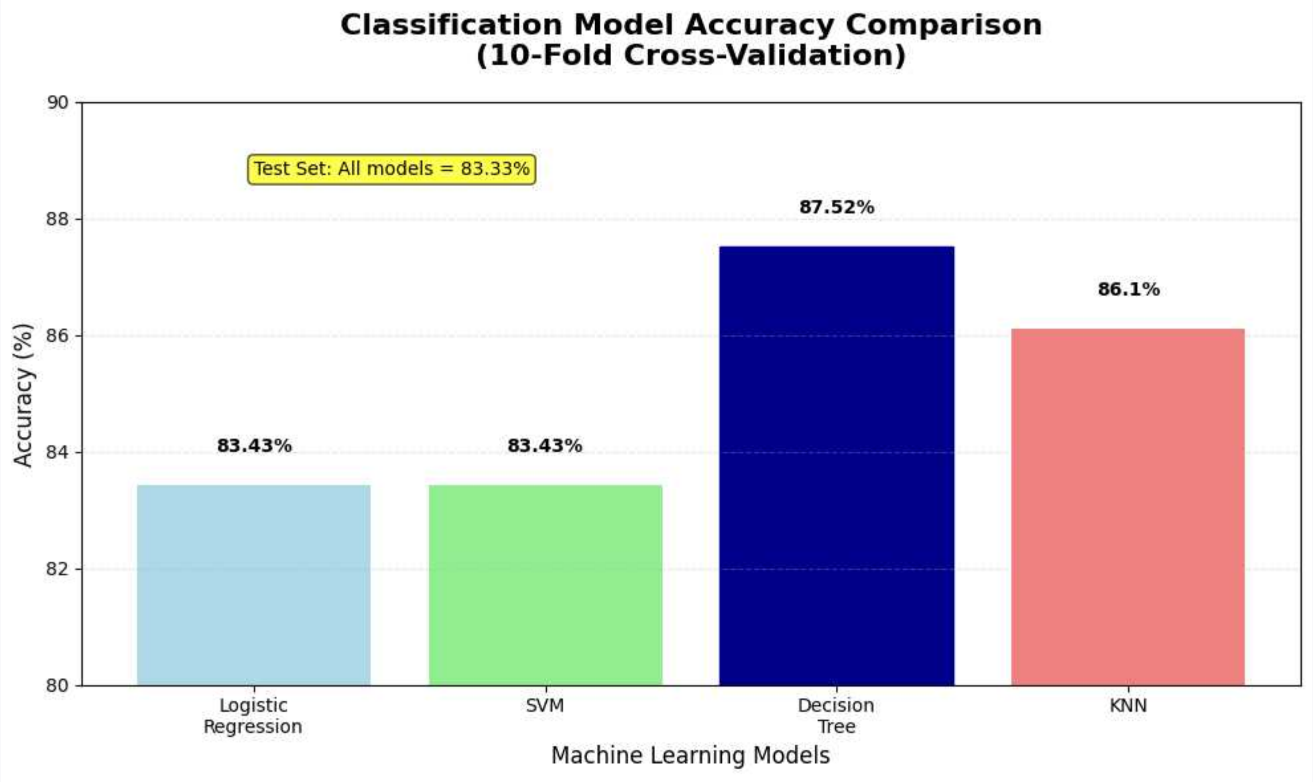
Predictive Analysis (Classification)

Classification Accuracy

Model Performance Comparison:

- Cross-Validation Results:
  - Decision Tree: 87.52% (highest)
  - KNN: 86.10%
  - Logistic Regression: 83.43%
  - SVM: 83.43%

- Test Set Results: All models achieved identical 83.33% accuracy
- Limitation: Small test set (18 samples) prevented meaningful model differentiation
- Best Model: **Decision Tree** based on cross-validation performance



Confusion Matrix

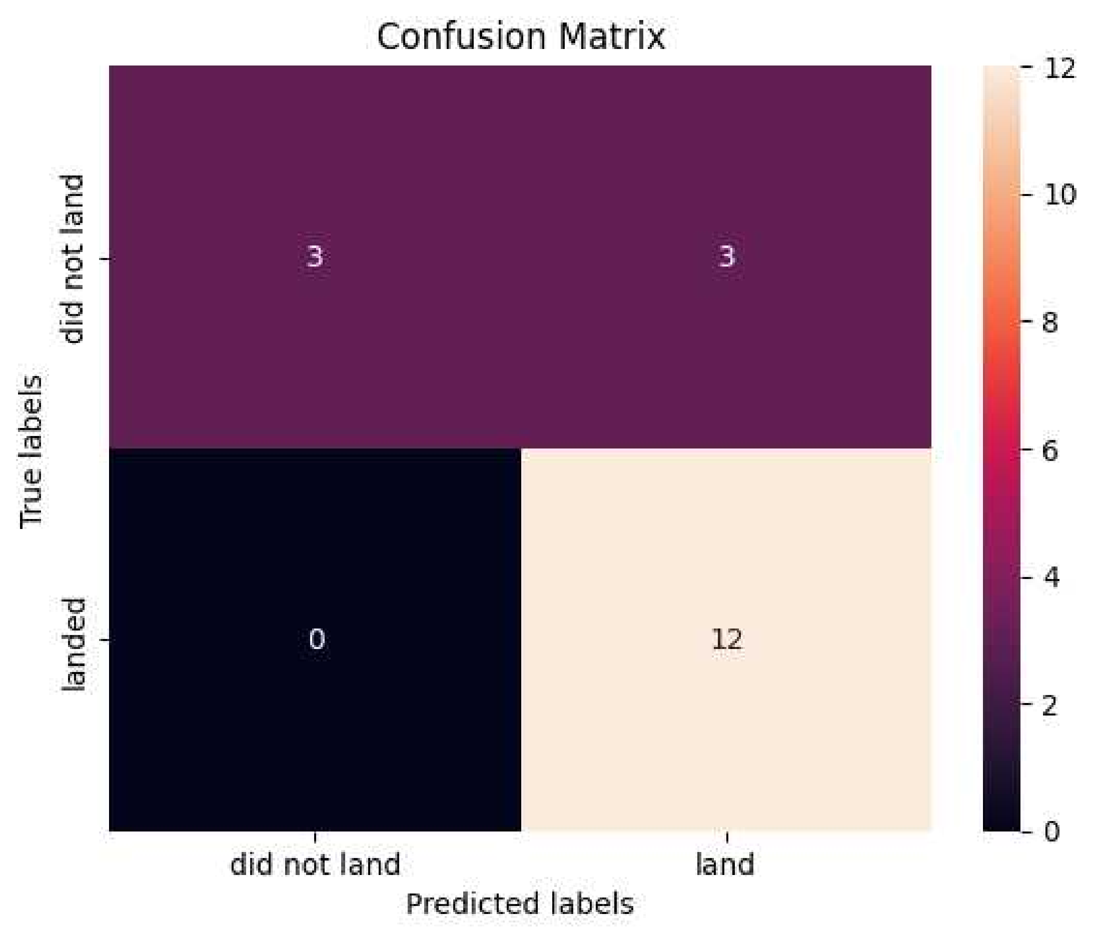
Results:

- True Positives: 12 (successful landings correctly predicted)
- True Negatives: 3 (failed landings correctly predicted)
- False Positives: 3 (failed landings incorrectly predicted as successful)
- False Negatives: 0 (no successful landings missed)

Key Insights:

- Perfect Recall: 100% of actual successful landings were correctly identified
- Precision Challenge: 50% precision for failed landings (3 FP out of 6 actual failures)
- Overall Accuracy: 83.33% (15 correct out of 18 predictions)

Business Impact: The model never misses a successful landing (critical for cost estimation) but occasionally overestimates landing success, leading to conservative cost predictions.



Conclusions

- **Point 1:** Decision Tree classifier achieved the highest cross-validation accuracy (87.52%), but all models converged to identical 83.33% performance on the small test dataset.
- **Point 2:** Machine learning models can reliably predict Falcon 9 landing success with >83% accuracy, enabling competitor cost estimation and strategic bidding against SpaceX.
- **Point 3:** Perfect recall for successful landings (100%) ensures no missed cost-saving opportunities, while conservative false positive predictions provide risk mitigation.
- **Point 4:** Dataset limitations (18 test samples, 90 total observations) restrict model differentiation and generalizability, highlighting the need for larger datasets in future work.
- **Point 5:** Launch site location, payload mass, and booster specifications provide sufficient predictive power for commercial space launch cost modeling.

Appendix

GitHub Repository Link: https://github.com/danielmh111/imb-ds-cert-capstone

Thank you!