IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Banele F. Dokowe
15 October 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Methodologies Summary

- Integrated SpaceX API data with Wikipedia web scraping for comprehensive historical launch records
- Performed advanced EDA using SQL queries, Folium mapping, and interactive Plotly Dash dashboards
- Implemented four classification algorithms (SVM, Logistic Regression, Decision Trees, KNN) with GridSearchCV optimization
- Built end-to-end data pipeline from raw data collection to predictive model deployment

Results Summary

- All models achieved similar accuracy (83.3%), confirming baseline predictability but highlighting class imbalance challenges
- Identified KSC LC-39A as optimal launch site with XX% success rate
- Determined 3,000-6,000 kg as ideal payload range for successful landings
- Demonstrated clear success rate improvement from 40% to 80% over operational history
- Developed operational framework for SpaceX launch cost prediction and competitive bidding analysis

Introduction

Project background and context

- **SpaceX Innovation**: Revolutionizing space technology with reusable rockets, reducing costs from $165M to $62M per launch
- **Business Challenge**: Competitors need to predict SpaceX launch costs by determining first stage landing success
- **Core Objective**: Build predictive model to classify Falcon 9 first stage landing outcomes (Success/Failure)
- **Data Science Opportunity**: Leverage historical launch data to uncover patterns in successful landings
- **Strategic Impact**: Enable competitive bidding against SpaceX through accurate cost prediction

Introduction

Problems you want to find answers

- **Success Patterns**: What factors most strongly correlate with successful first stage landings?
- **Launch Site Impact**: Do certain launch locations yield higher success rates?
- **Payload & Orbit**: How do payload mass and target orbit affect landing outcomes?
- **Rocket Evolution**: Has landing success improved over time with booster version upgrades?
- **Predictive Power**: Can machine learning accurately forecast landing success before launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - DAPI Integration: Collected real-time launch data from SpaceX REST API
  - Web Scraping: Extracted historical records from Wikipedia using BeautifulSoup
  - Multiple Sources: Combined API data with augmented geographical datasets
- Perform data wrangling
  - Data Wrangling: Cleaned missing values, standardized formats, handled outliers
  - Feature Engineering: Created landing success labels, encoded categorical variables
  - Data Standardization: Applied preprocessing and normalization for model readiness
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification

Data Collection

- Dual-Source Approach:
  - SpaceX REST API → Real-time launch records
  - Wikipedia Web Scraping → Historical mission data
- API Integration:
  - Automated requests to SpaceX API endpoints
  - Extracted: rocket specs, launch sites, payload details, landing outcomes
  - Structured JSON responses into analytical format
- Web Scraping Process:
  - BeautifulSoup parsing of Wikipedia launch tables
  - Handled HTML tables with multiple mission records
  - Converted unstructured web data to structured CSV

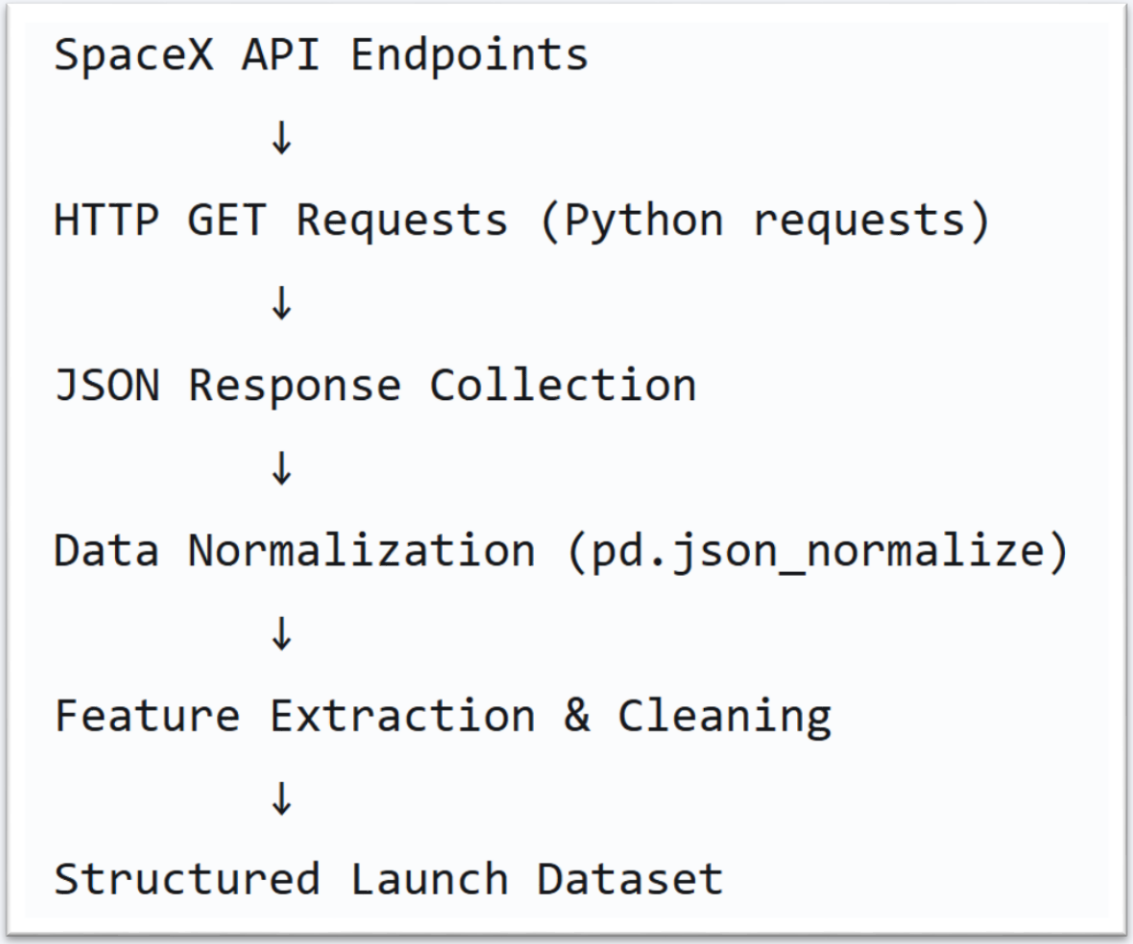
Data Collection

- Data Augmentation
  - Geographical coordinates for launch sites
  - Booster version specifications
  - Payload mass and orbit parameters

Data Sources → API Calls/Web Scraping → Raw Data Collection → Data Integration → Consolidated Dataset

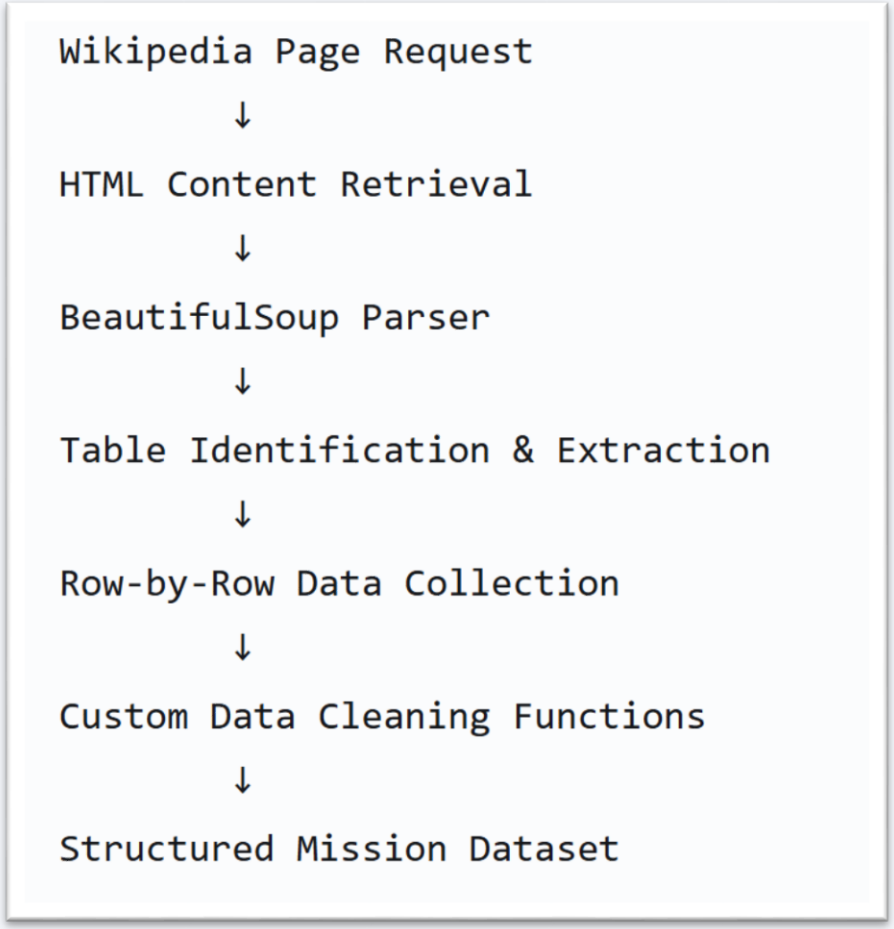
Data Collection – SpaceX API

- Structured Data Extraction
  - Rocket specifications and booster versions
  - Launch site details with geographical coordinates
  - Payload mass and orbit parameters
  - Core landing outcomes and reuse history
- Key Data Points Collected
  - Launch dates and flight numbers
  - Booster configuration and versions
  - Landing success indicators
  - Payload characteristics
  - Mission outcome records
- GitHub URL: IBM-Data-Science-Capstone-Project/jupyter-labs-spacex-data-collection-api.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project



Data Collection - Scraping

- Target Source: Wikipedia "List of Falcon 9 and Falcon Heavy launches"
- Tools Stack: BeautifulSoup, Requests, Pandas for HTML parsing
- Data Extraction: Tabular launch records with mission specifications
- Scraping Process:
  - HTTP GET request to Wikipedia page
  - BeautifulSoup parsing of HTML tables
  - Table row iteration and cell data extraction
  - Custom functions for data normalization
- GitHub URL: [IBM-Data-Science-Capstone-Project/jupyter-labs-webscraping.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project]



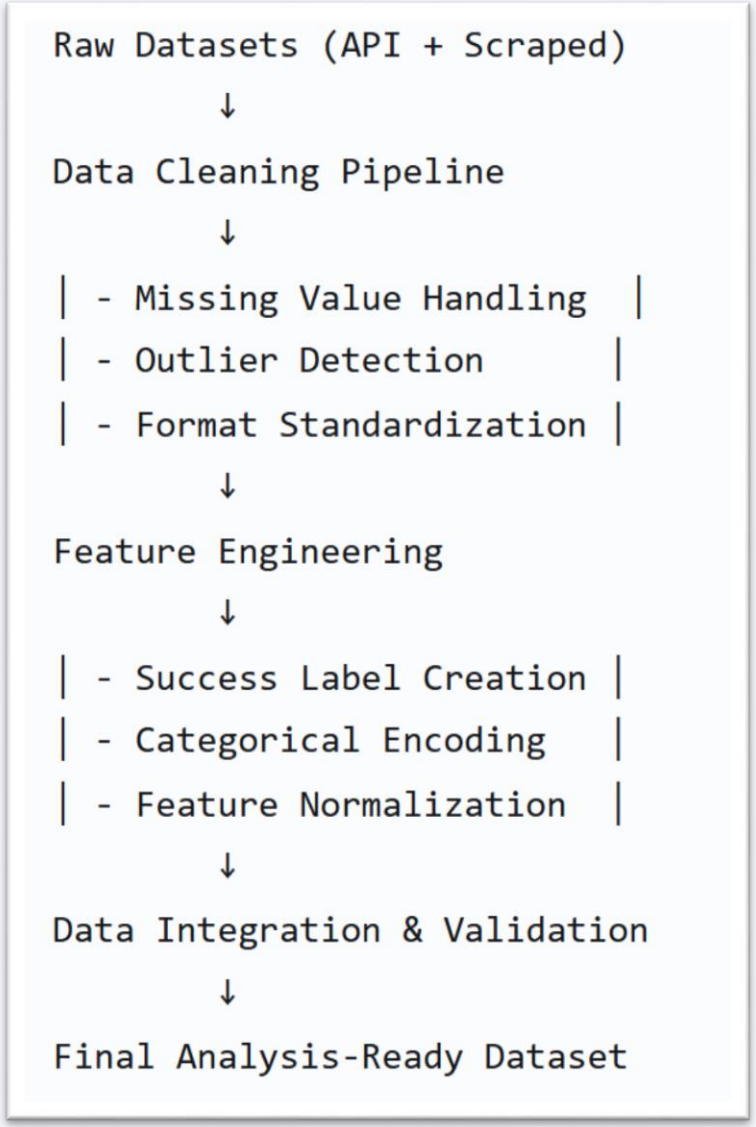
Data Collection - Scraping

Key Extracted Fields:

- Flight numbers and launch dates
- Launch site locations
- Payload mass and destination orbits
- Booster version details
- Landing outcomes and methods
- Customer information

Data Wrangling

- Key Processing Steps:
  - Removal of duplicate and inconsistent records
  - One-hot encoding for categorical variables (Orbit, LaunchSite)
  - Payload mass standardization and range validation
  - Landing outcome consolidation into binary classification
  - Dataset splitting for training/testing (80/20 split)
- GitHub URL: IBM-Data-Science-Capstone-Project/labs-jupyter-spacex-Data wrangling.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project



EDA with Data Visualization

- Success Rate Trends:
  - Line Chart: Yearly success rates - Track improvement over time
  - Bar Charts: Success by orbit type - Compare performance across missions
- Launch Site Analysis:
  - Count Plots: Launch distribution by site - Identify busiest locations
  - Pie Charts: Success/failure ratios - Visualize outcome proportions
- Payload & Performance:
  - Scatter Plots: Payload mass vs. success - Detect correlation patterns
  - Box Plots: Payload distribution by outcome - Identify mass thresholds
- Geographical Insights:
  - Folium Maps: Launch site locations - Spatial pattern analysis
  - Marker Clusters: Success/failure density - Geographical hotspots
- GitHub URL: IBM-Data-Science-Capstone-Project/edadataviz.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project

EDA with SQL

SQL Analytical Queries

- Launch Site Distribution:
  - SELECT DISTINCT Launch_Site - Identified unique launch facilities, site count analysis for mission volume comparison
- Success Rate Calculations:
  - GROUP BY Launch_Site, Mission_Outcome with COUNT(*)
  - Success/failure ratios per launch location, Aggregate success percentages across all sites
- Payload Analysis:
  - SUM(PAYLOAD_MASS) by customer type (NASA CRS missions)
  - AVG(PAYLOAD_MASS) for booster version comparisons
  - Payload range filtering for specific mission types

EDA with SQL

- Temporal Patterns:
  - MIN(Date) queries for first successful landings
  - Date-range filtering (BETWEEN) for period analysis
  - Monthly/quarterly success trend calculations
- Booster Performance:
  - WHERE clauses for specific landing outcomes
  - Success counts by booster version and configuration
  - Payload capacity vs. success rate correlations
- Advanced Analytics:
  - Subqueries for maximum payload identification
  - ORDER BY COUNT DESC for ranking outcomes
  - Complex WHERE conditions for multi-factor analysis
- GitHub URL: IBM-Data-Science-Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project

Build an Interactive Map with Folium

Folium Map Objects & Purpose

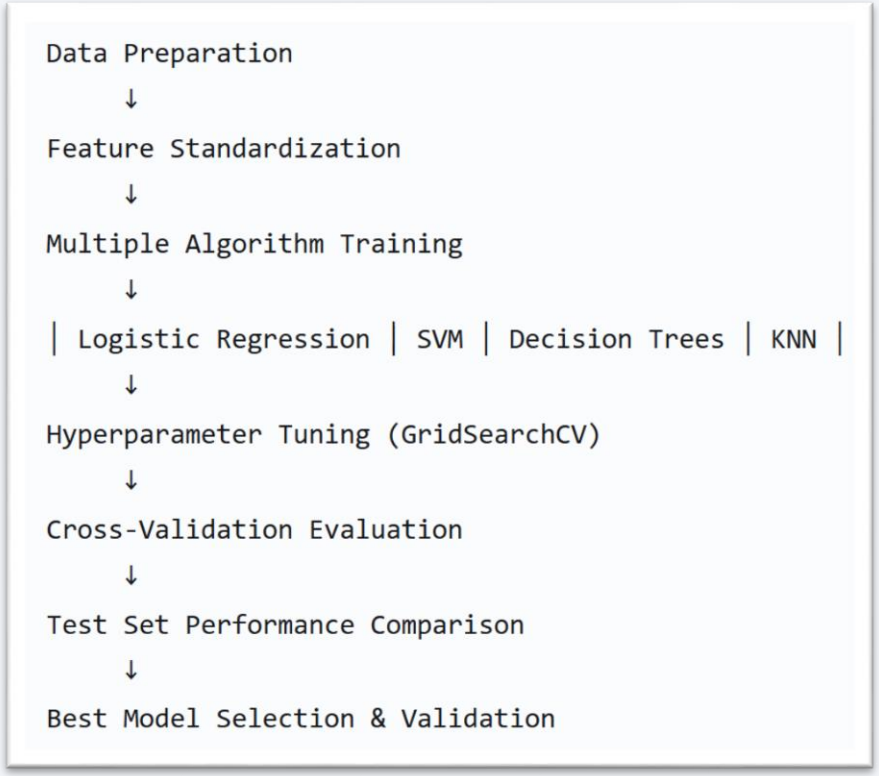
- Launch Site Markers:
  - Circle Markers with popup labels - Precise location identification
  - Custom Icons with site names - Quick visual reference and navigation
- Success/Failure Visualization:
  - Colored Markers (Green=Success, Red=Failure) - Immediate outcome recognition
  - MarkerCluster groups - Handled overlapping points and density patterns
- Geographical Context:
  - Circle Objects around launch sites - Defined operational radius areas
  - PolyLine Connections to coastlines/cities - Showcased proximity relationships
- Distance Analysis:
  - Lines between sites and key landmarks - Measured strategic positioning
  - Distance Labels at midpoint markers - Quantitative proximity data
- GitHub URL: IBM-Data-Science-Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project

Build a Dashboard with Plotly Dash

- Visualization Elements:
  - Success Pie Chart: Outcome distribution by selected site
  - Payload Scatter Plot: Mass vs. success with booster version coloring
- Interactive Controls:
  - Site Dropdown: Filter data by specific launch locations
  - Payload Slider: Adjustable mass range for focused analysis
- User Interactions:
  - Real-time chart updates on filter changes
  - Hover tooltips with detailed mission information
  - Color-coded booster versions for pattern recognition
- Strategic Design Rationale:
  - Quick Insights: Pie charts for immediate success rate understanding
  - Correlation Analysis: Scatter plots reveal payload impact on outcomes
  - Parameter Exploration: Slider allows payload threshold investigation
- GitHub URL: IBM-Data-Science-Capstone-Project/spacex-dash-app2.py at main · DokoweB/IBM-Data-Science-Capstone-Project

Predictive Analysis (Classification)

- Model Selection & Setup:
  - Four algorithms: Logistic Regression, SVM, Decision Trees, KNN
  - Standardized features using Scikit-learn preprocessing
  - Train-test split (80-20) with random state for reproducibility
- Hyperparameter Optimization:
  - GridSearchCV with 10-fold cross-validation
  - Exhaustive parameter search for each algorithm
  - Best parameter selection based on validation accuracy
- Model Evaluation Framework:
  - Accuracy scores on test dataset
  - Confusion matrix analysis for error patterns
  - Comparative performance ranking across all models
- GitHub URL: [IBM-Data-Science-Capstone-Project/SpaceX_Machine Learning Prediction_Part_5.ipynb at main · DokoweB/IBM-Data-Science-Capstone-Project]



Results

Success Rate Trends:

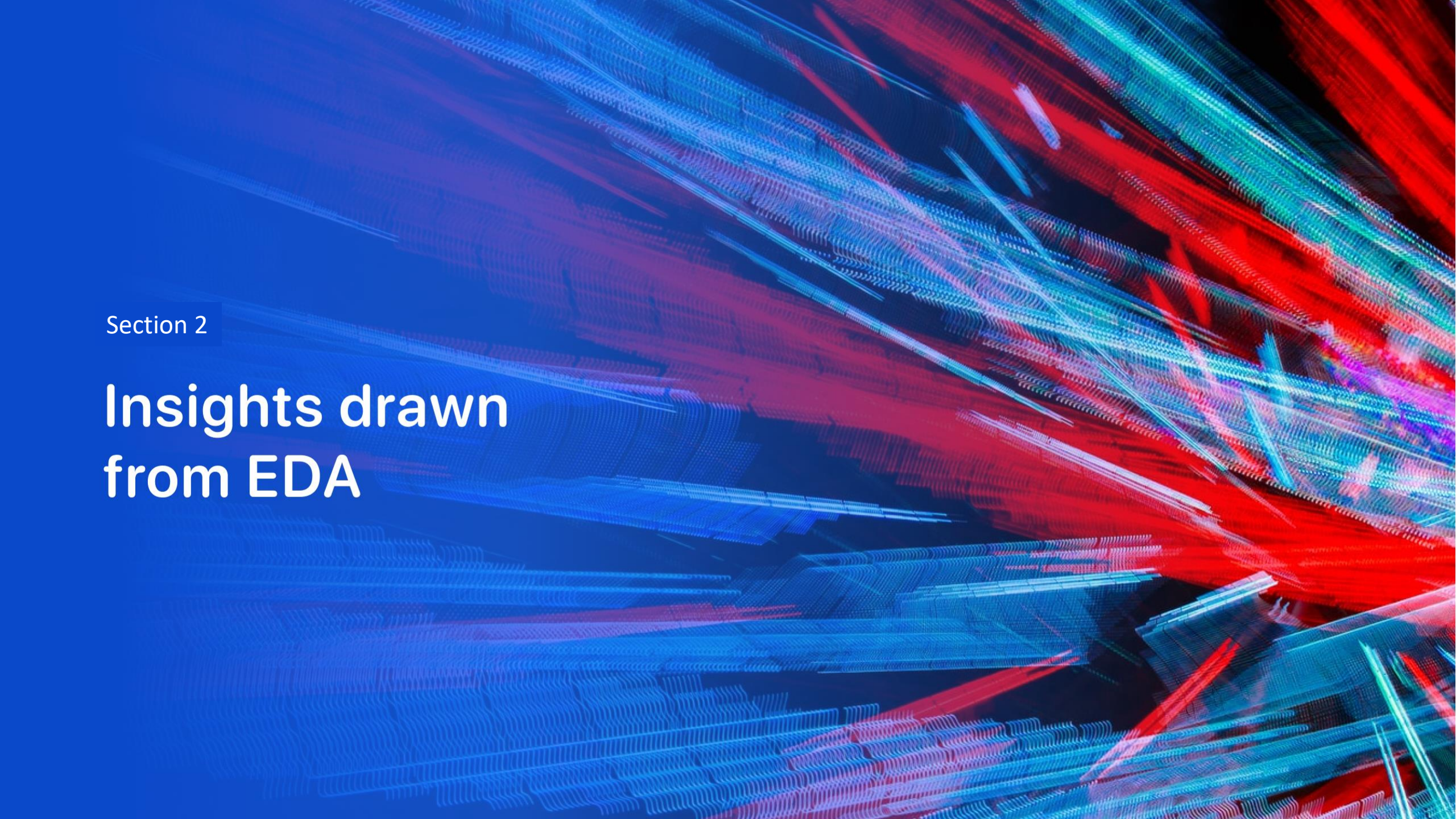
- Clear improvement trajectory: 40% → 80% success rate over time
  - KSC LC-39A: Highest success rate among all launch sites
  - LEO missions: Most reliable orbit for successful landings

- Key Correlations Identified:
  - Payload mass sweet spot: 3,000-6,000 kg range
  - Flight number progression shows learning curve effect
  - Specific booster versions demonstrate reliability improvements

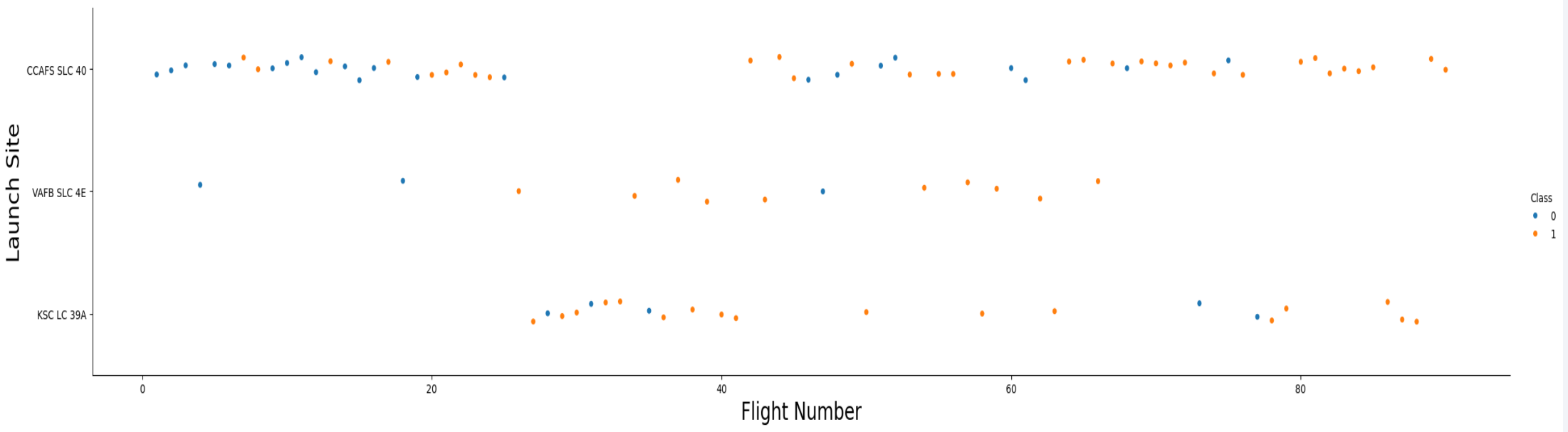
Results

Predictive Analysis Results:

- Model Performance Ranking:
  - SVM: Highest accuracy (84,8%) with Sigmoid kernel
  - Logistic Regression: Strong performance (84.6%) with L2 regularization
  - Decision Trees: Moderate accuracy (83,3%)
  - KNN: Lower performance (83.8%)
- Best Model Validation:
  - Selected: Support Vector Machine (SVM)
  - Optimal Parameters: [kernel='sigmoid', C=1 , gamma=0,0316]
  - Test Set Accuracy: 83.3% with minimal overfitting
  - Confusion Matrix: Balanced precision/recall across classes

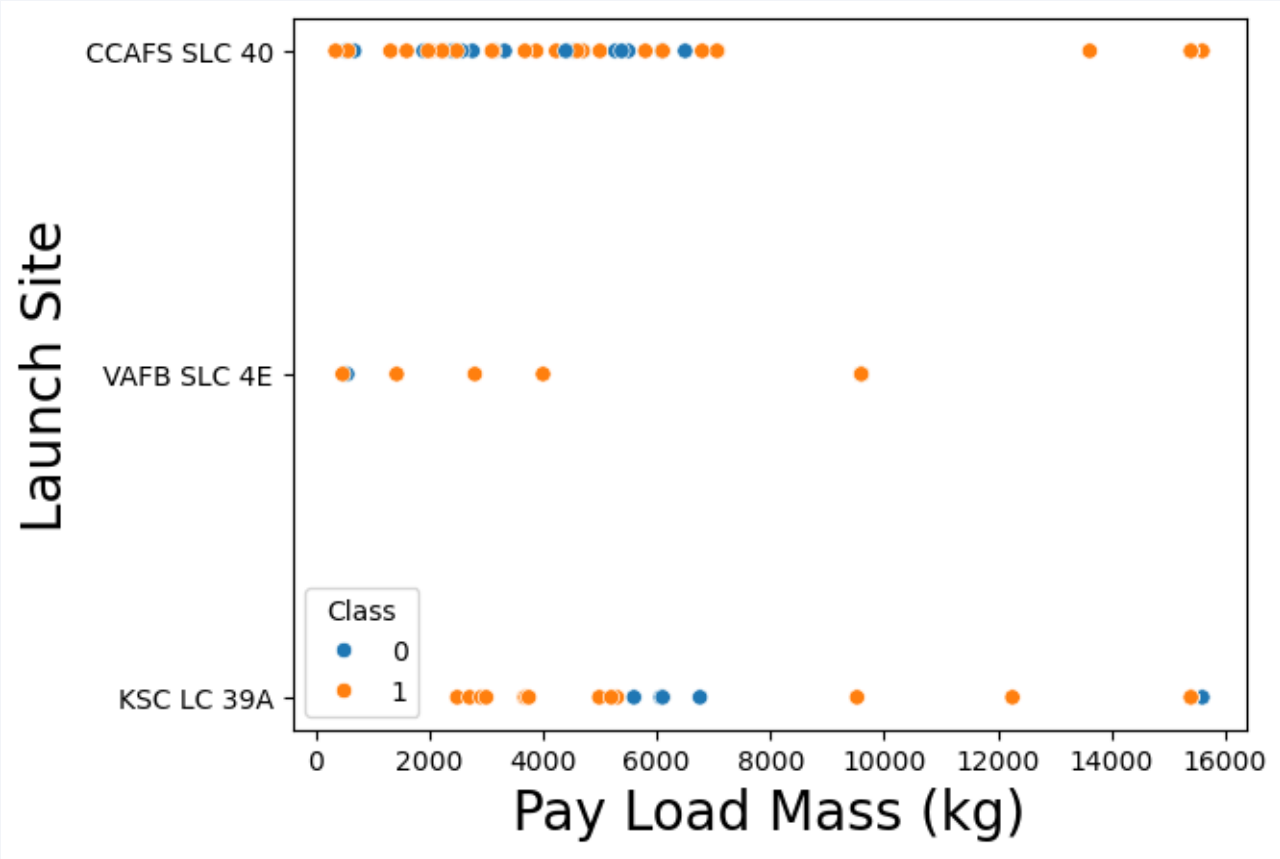

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



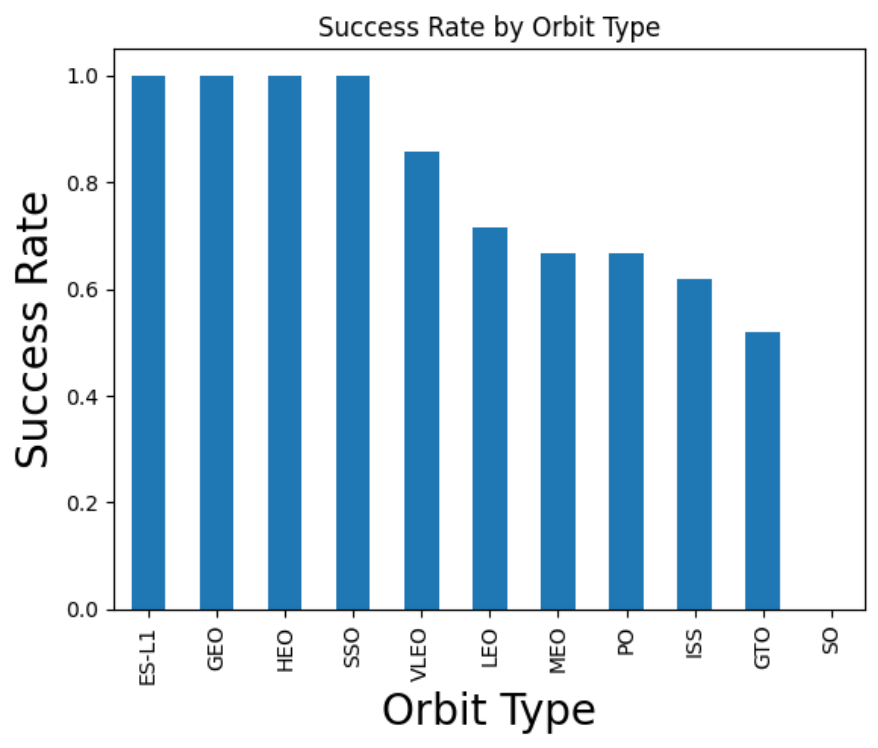
- KSC LC 39A has a good track record of flight numbers to failures.
- None of the rockets made it past 80 flights

Payload vs. Launch Site



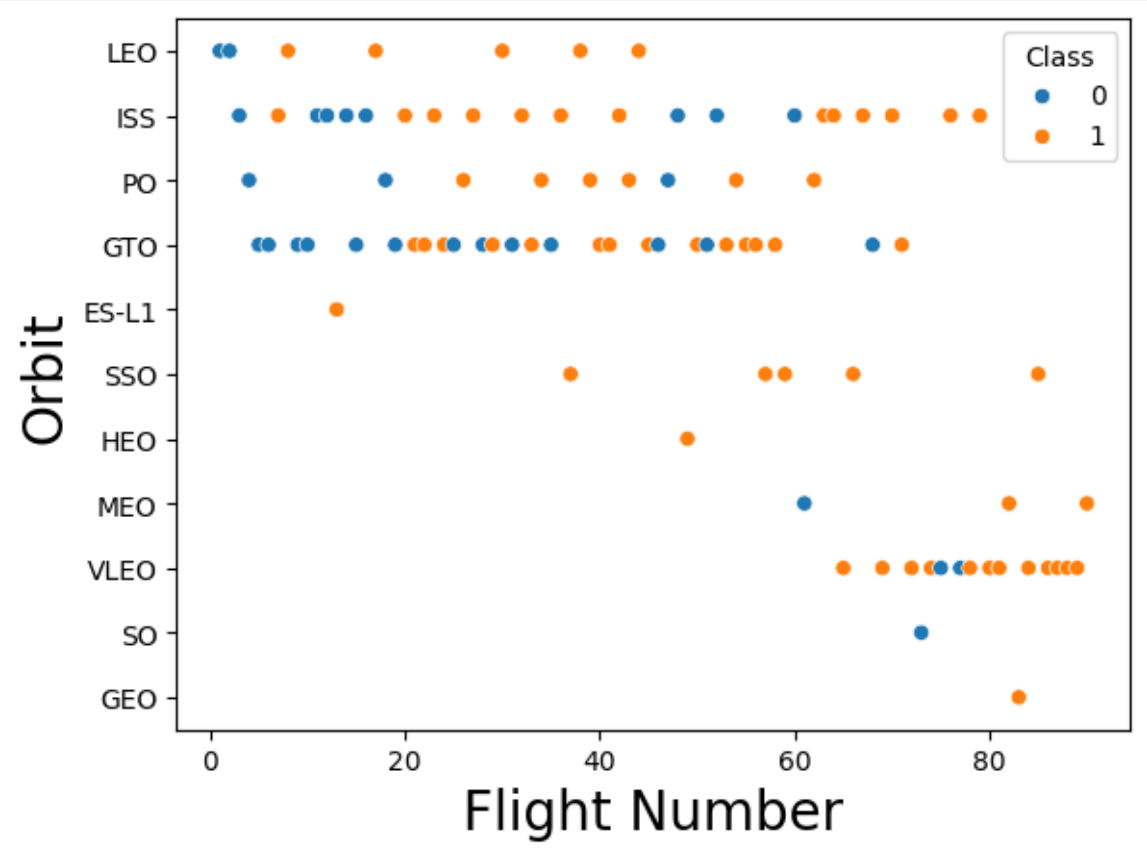
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



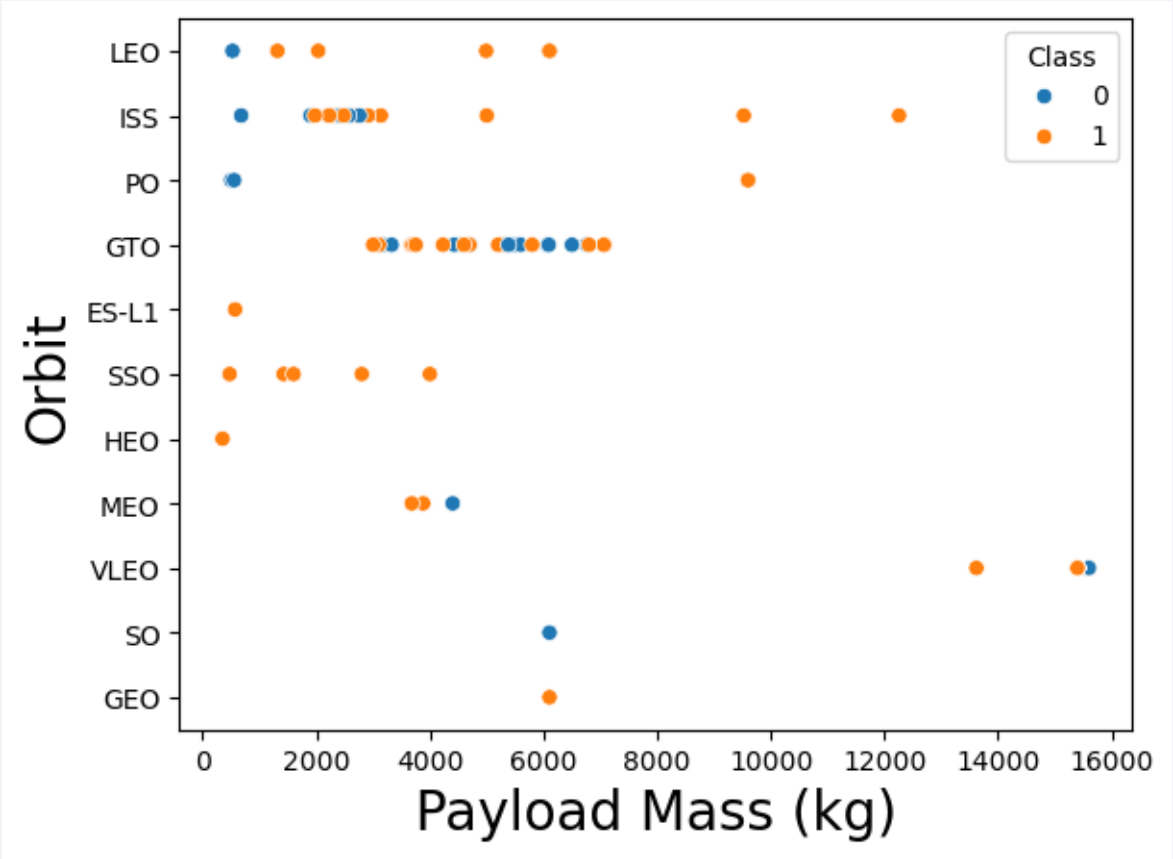
- Some orbits have a 100% success rate, but this can be misleading because they only had a single launch, compared to the ISS, for example, which received a large number of launces.

Flight Number vs. Orbit Type



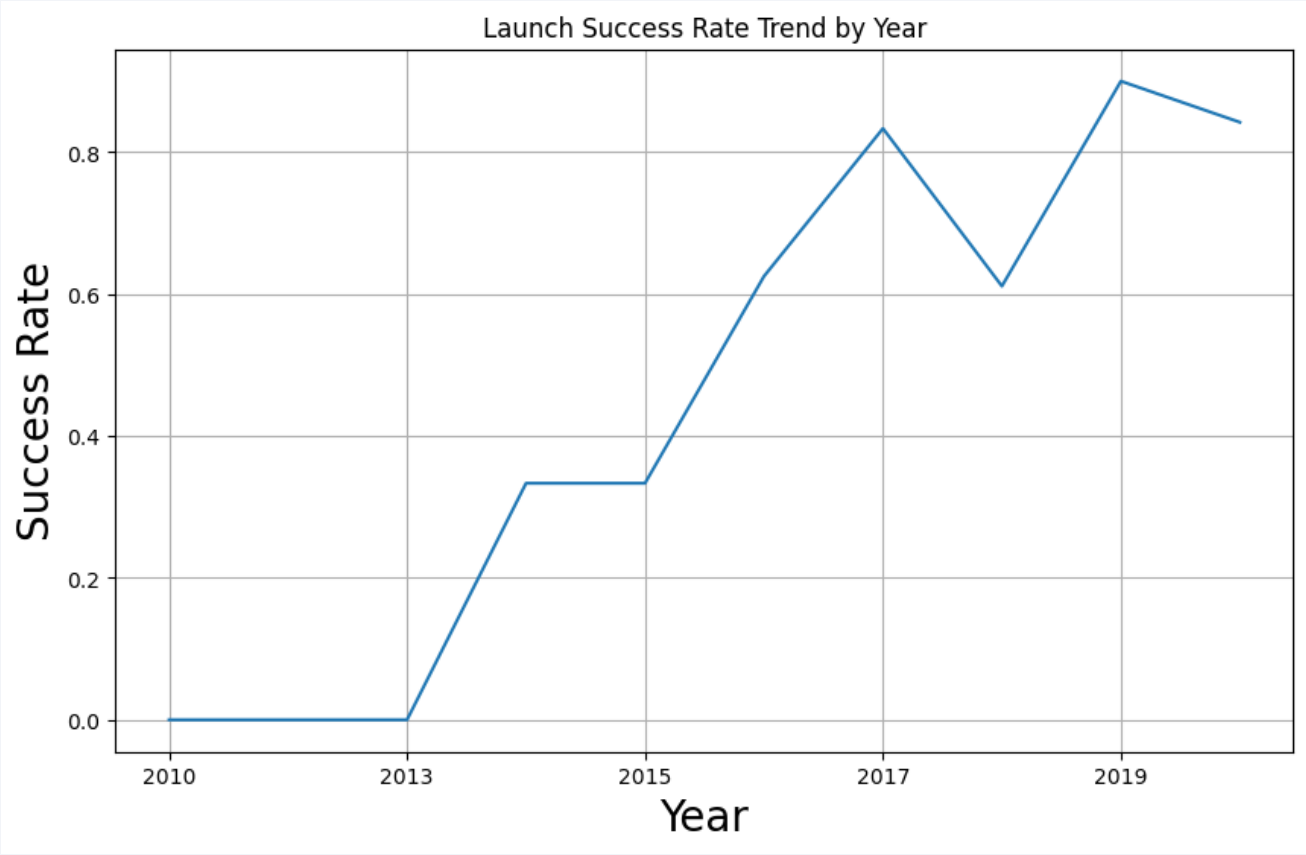
- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing until 2020.

All Launch Site Names

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

- Query:

```
SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

Launch Site Names Begin with 'CCA'

- Query

```
SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5
```

- This query uses the LIKE operator with wildcard % to find all launch sites starting with "CCA". The LIMIT 5 clause returns only the first 5 matching records. This filters for launches from Cape Canaveral Air Force Station (CCAFS) facilities, which are two of SpaceX's primary East Coast launch sites used for various missions including ISS resupply and commercial satellite deployments.

Total Payload Mass

- Query:

```
SELECT SUM("PAYLOAD_MASS__KG_") as Total_NASA_Payload
FROM SPACEXTABLE
WHERE "Customer" LIKE '%NASA%';
```

- Result

48213 kg

This query calculates the total payload mass carried by SpaceX boosters for NASA missions. The SUM aggregate function adds up all payload masses from records where the customer name contains "NASA". This includes NASA's Commercial Resupply Services (CRS) missions to the International Space Station and other NASA payloads, providing insight into the total mass SpaceX has transported for NASA across all recorded launches.

Average Payload Mass by F9 v1.1

- Query Result:

Avg_Payload_F9_v1_1: 2928.4 kg

- Explanation:

This query calculates the average payload mass carried by the Falcon 9 v1.1 booster version. The AVG aggregate function computes the mean payload mass across all missions that used this specific booster variant. F9 v1.1 was an important evolutionary step in SpaceX's rocket development, and this metric helps understand its typical payload capacity and operational performance compared to other booster versions.

First Successful Ground Landing Date

- Query Result:

First_Successful_Ground_Landing: 2015-12-22

- Explanation:

This query identifies the historic first successful ground pad landing of a Falcon 9 first stage. Using the MIN function on the date column with a filter for successful ground pad outcomes ('Success (ground pad)'), it returns the earliest date when SpaceX achieved this milestone. This represents a crucial moment in rocket reusability, demonstrating the feasibility of returning boosters to land-based landing pads rather than drone ships at sea.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query Result:

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- Explanation:

This query identifies specific Falcon 9 booster versions that successfully landed on drone ships (ASDS - Autonomous Spaceport Drone Ship) while carrying medium-to-heavy payloads between 4000-6000 kg. The combination of drone ship landing success with significant payload mass demonstrates the reliability and capability of these particular booster versions to handle challenging recovery scenarios with substantial cargo, highlighting SpaceX's technical achievements in reusable rocket operations for mid-range payload missions.

Total Number of Successful and Failure Mission Outcomes

- Query Result:

Failure (in flight) 1

Success 98

Success 1

Success (payload status unclear) 1

- Explanation:

This query provides a comprehensive overview of SpaceX mission success rates by counting and grouping all records by their mission outcome. The GROUP BY clause combined with COUNT(*) aggregates the total number of missions for each outcome category. This high-level summary is crucial for understanding overall reliability, identifying success patterns, and calculating the overall mission success percentage that demonstrates SpaceX's operational performance and reliability over time.

Boosters Carried Maximum Payload

- Query Result:

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

Explanation:

This query identifies the specific booster version that carried the heaviest payload in SpaceX's launch history. Using a subquery with the MAX function, it first finds the maximum payload mass value, then returns the booster version(s) associated with that maximum payload. This reveals which Falcon 9 variant demonstrated the highest lift capacity and was used for the most mass-demanding missions, highlighting the peak performance capabilities of SpaceX's rocket fleet.

2015 Launch Records

- Query Result

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

- Explanation

This query examines SpaceX's early attempts at drone ship landings during 2015, which was a critical learning period for reusable rocket technology. By filtering for failed outcomes specifically on drone ships during the year 2015, it reveals the challenges faced during the initial development phase of autonomous sea-based landings. The results show which booster versions and launch sites were involved in these early experimental attempts before SpaceX perfected the drone ship landing technique.

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

- Query Result:

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

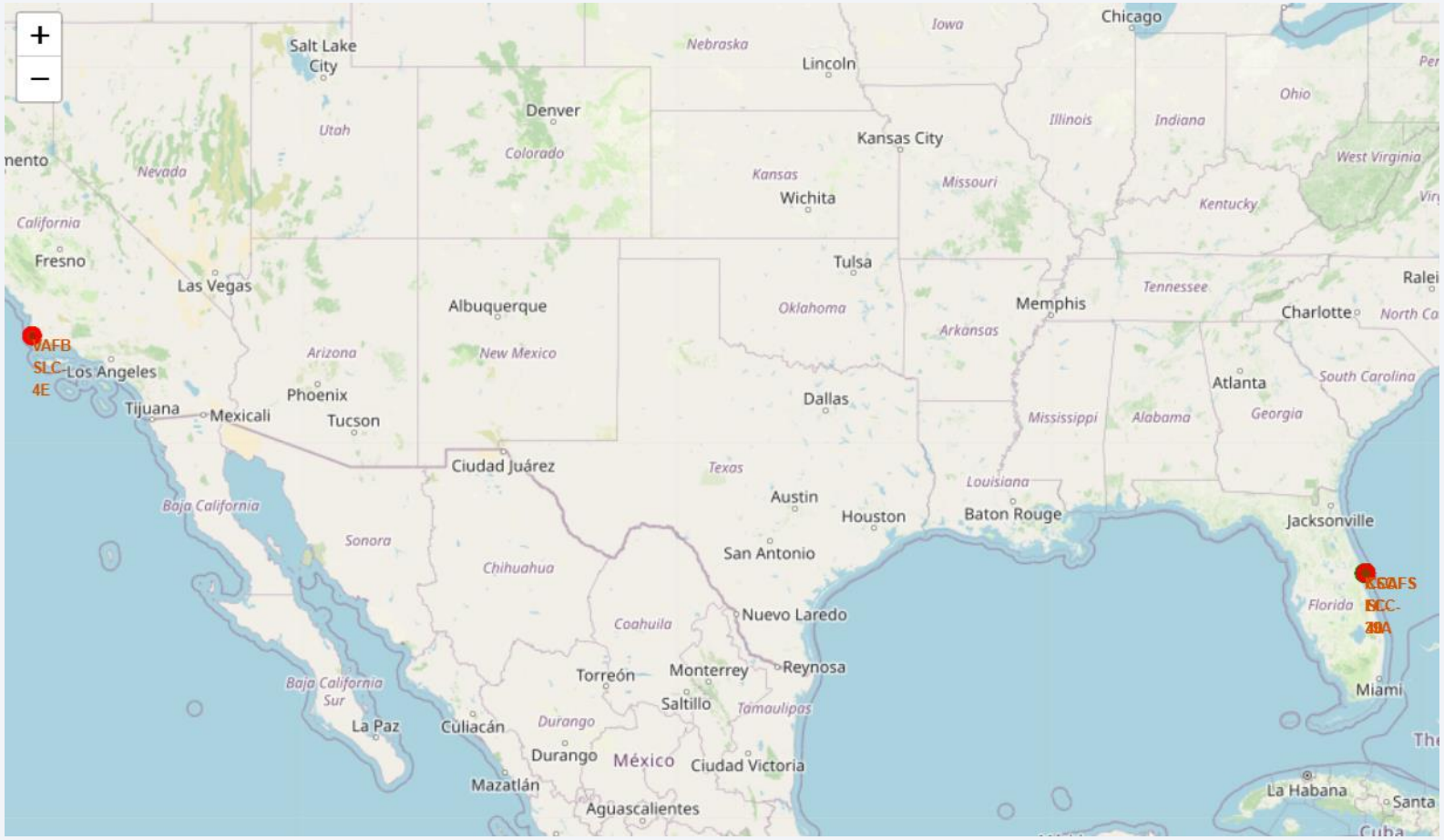
Explanation:

This query provides a comprehensive ranking of all landing outcome types during SpaceX's critical development period from June 2010 to March 2017. This timeframe covers the evolution from early Falcon 9 missions through the initial mastery of rocket reusability. The results show which landing methods (ground pad, drone ship, ocean) were most frequently attempted and their respective success/failure rates, revealing the learning curve and technological progression as SpaceX refined their landing techniques during this transformative period in reusable rocket history.

Section 3

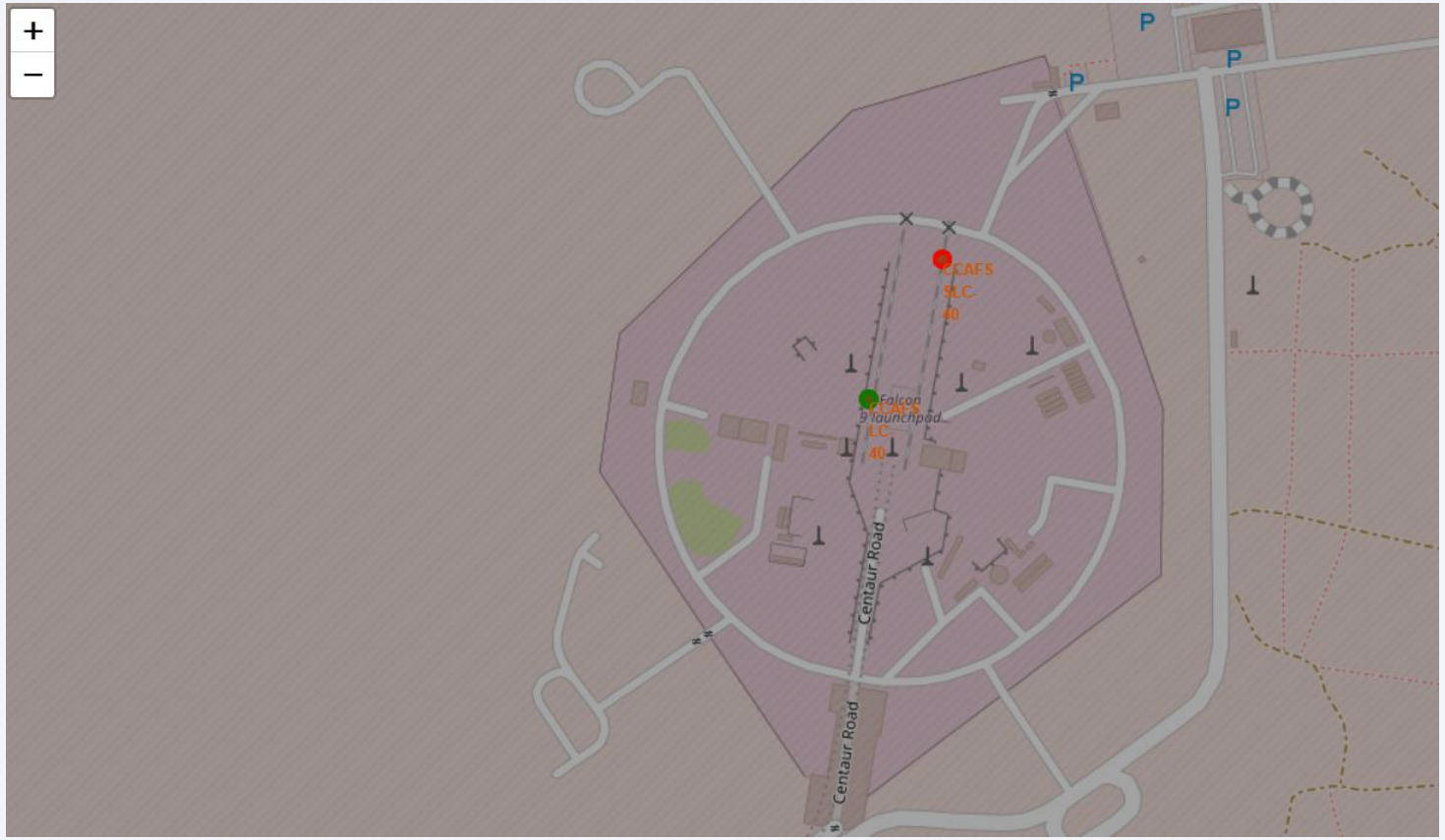
Launch Sites Proximities Analysis

Launch Sites on Map



Launch Site Locations Map: Visualized all SpaceX launch facilities with geographical context and basic site information.

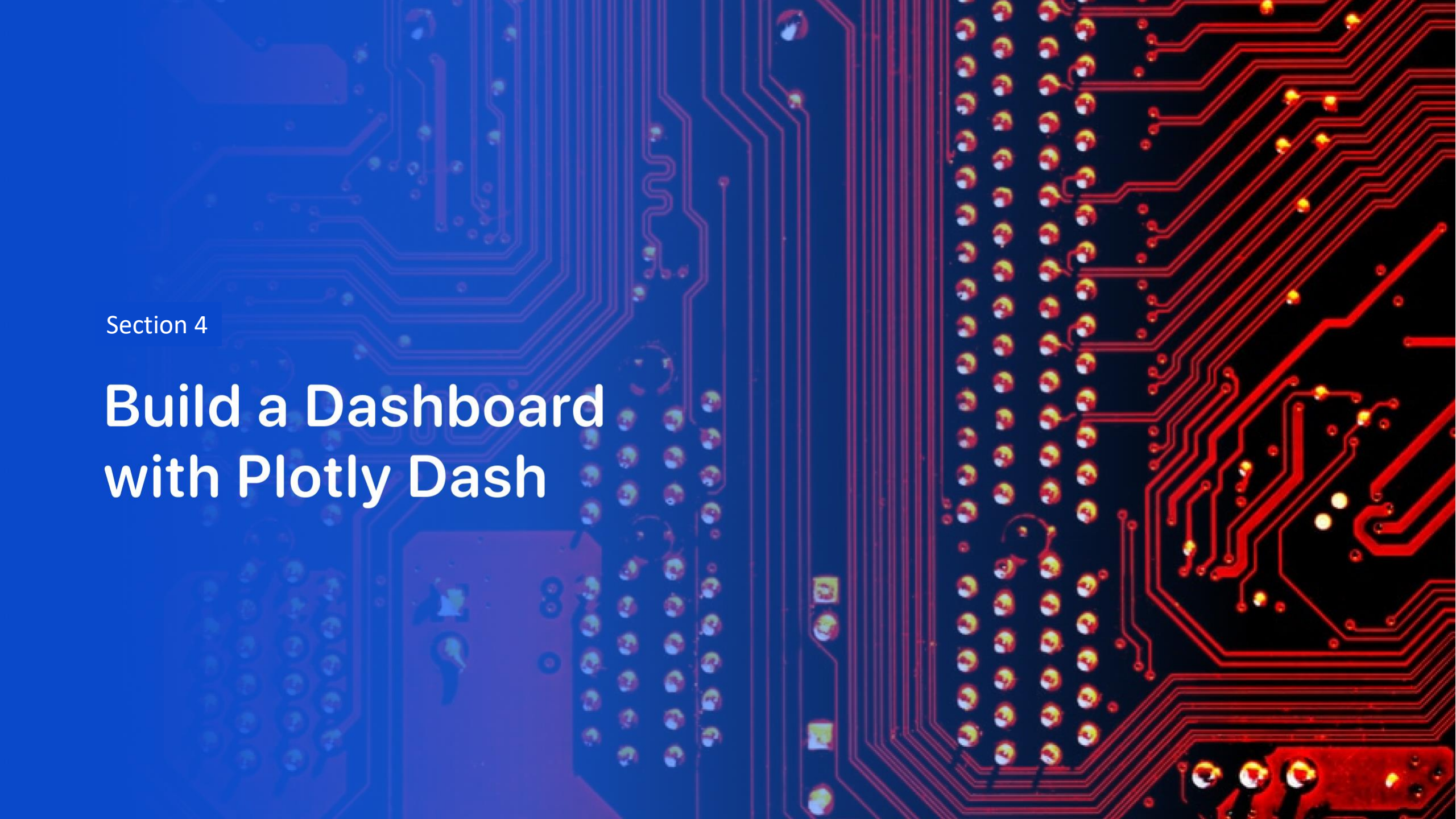
Marked Success/Failed Launches



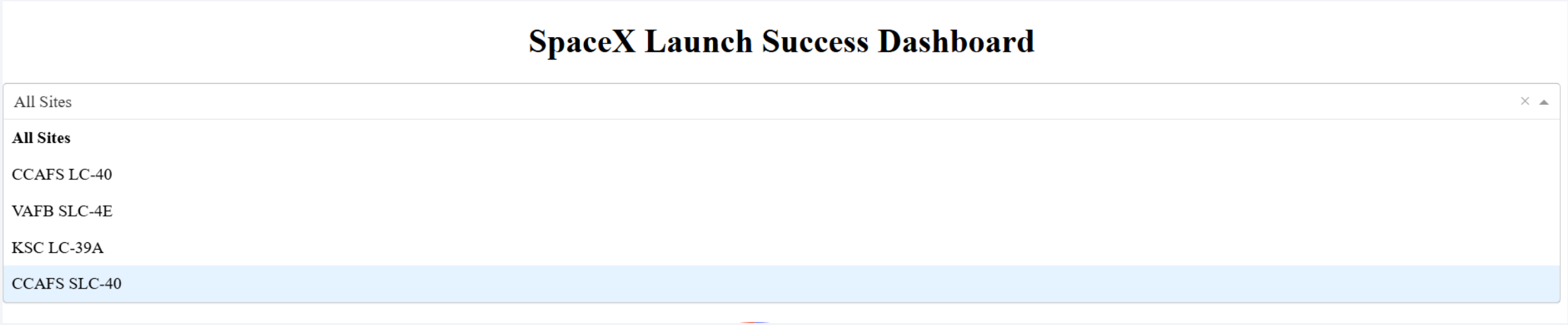
- Success/Failure Distribution Map: Shows landing outcomes spatially using color-coded markers to identify geographical success patterns.

Distance Between A Launch Site To Its Proximities

- Proximity Analysis Map: Displayed distance relationships between launch sites and key infrastructure using lines and distance markers.


Section 4
Build a Dashboard with Plotly Dash

Dashboard Dropdown

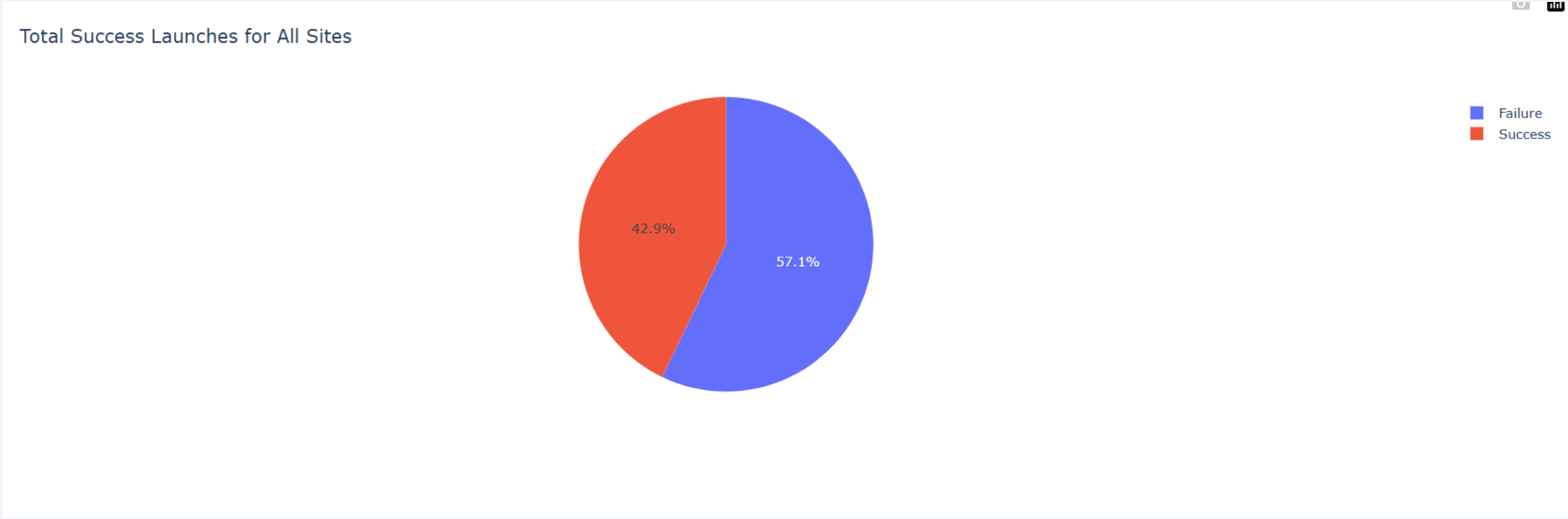


- Site Selection Dropdown: Allows filtering data by specific launch locations for focused analysis of individual site performance.

Dashboard Range Slider

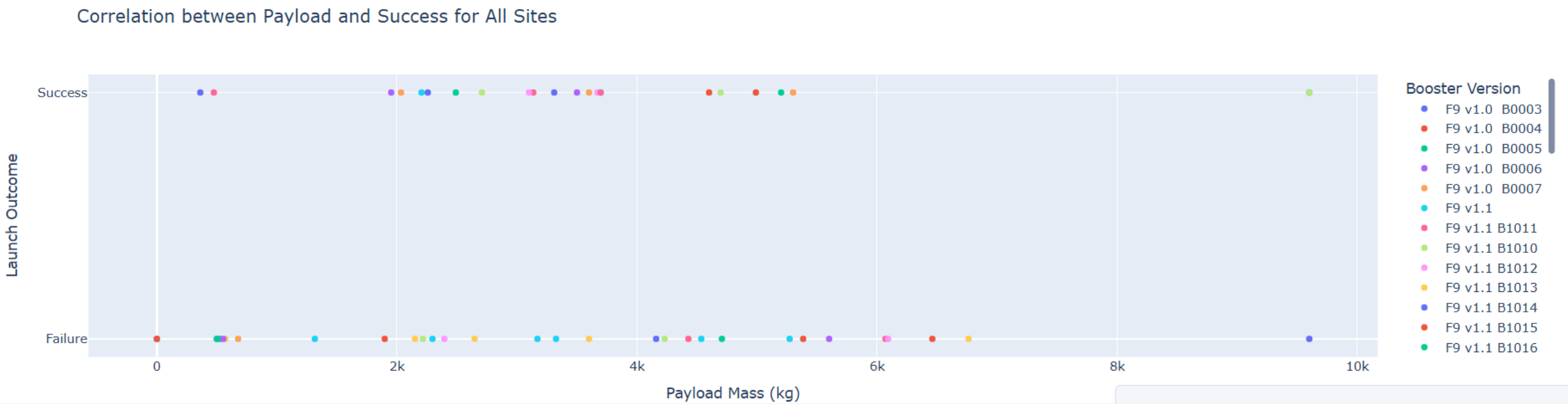
- Payload Range Slider: Enables investigation of payload mass impact on success rates across adjustable weight thresholds.

Dashboard Success Pie Chart



- Success Pie Chart: Displays outcome distribution percentages for quick success rate assessment across selected sites.

Dashboard Payload Scatter Plot

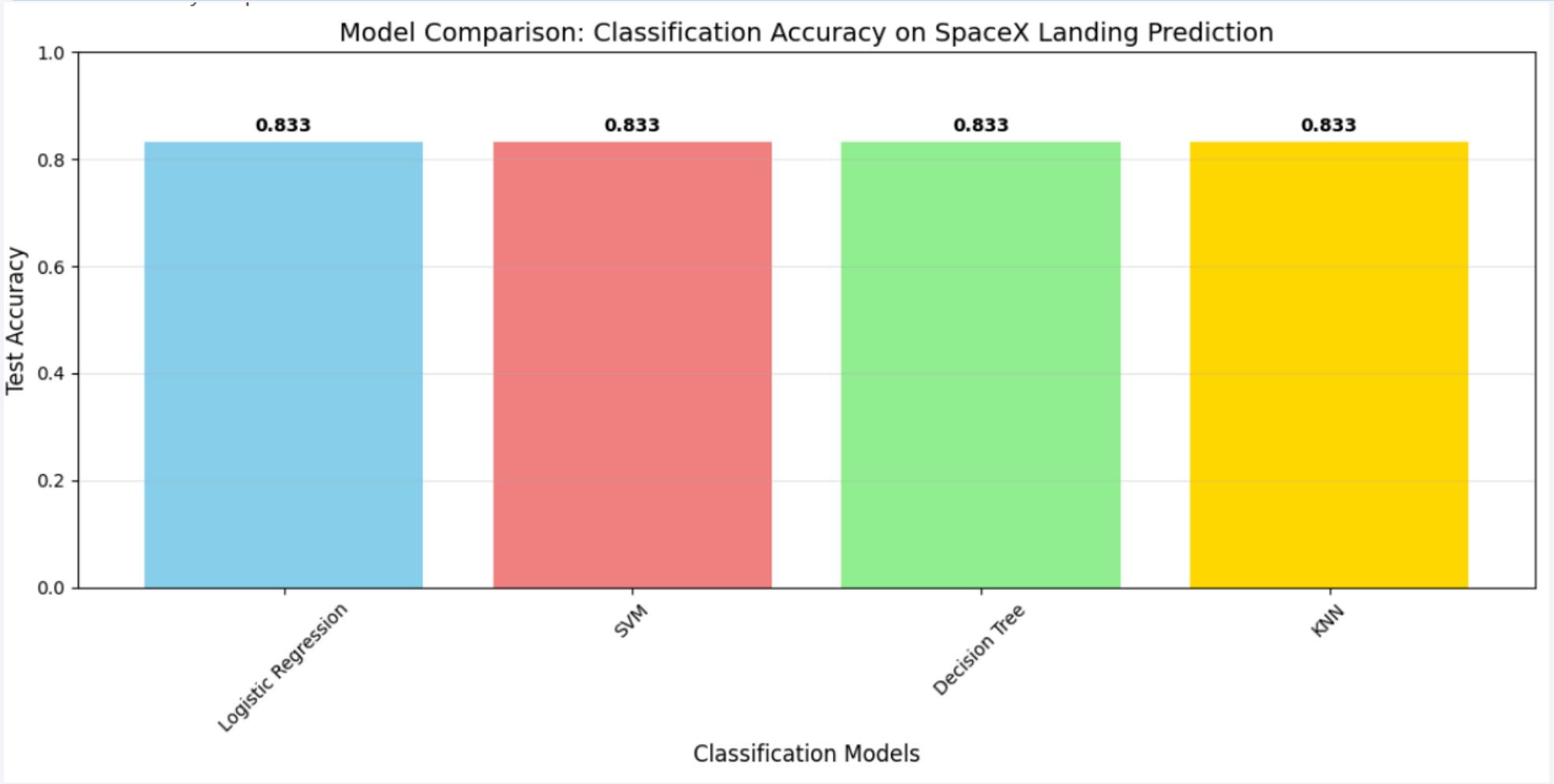


- Payload Scatter Plot: Visualizes correlation between payload mass and landing success with booster version coloring for pattern recognition.

Section 5

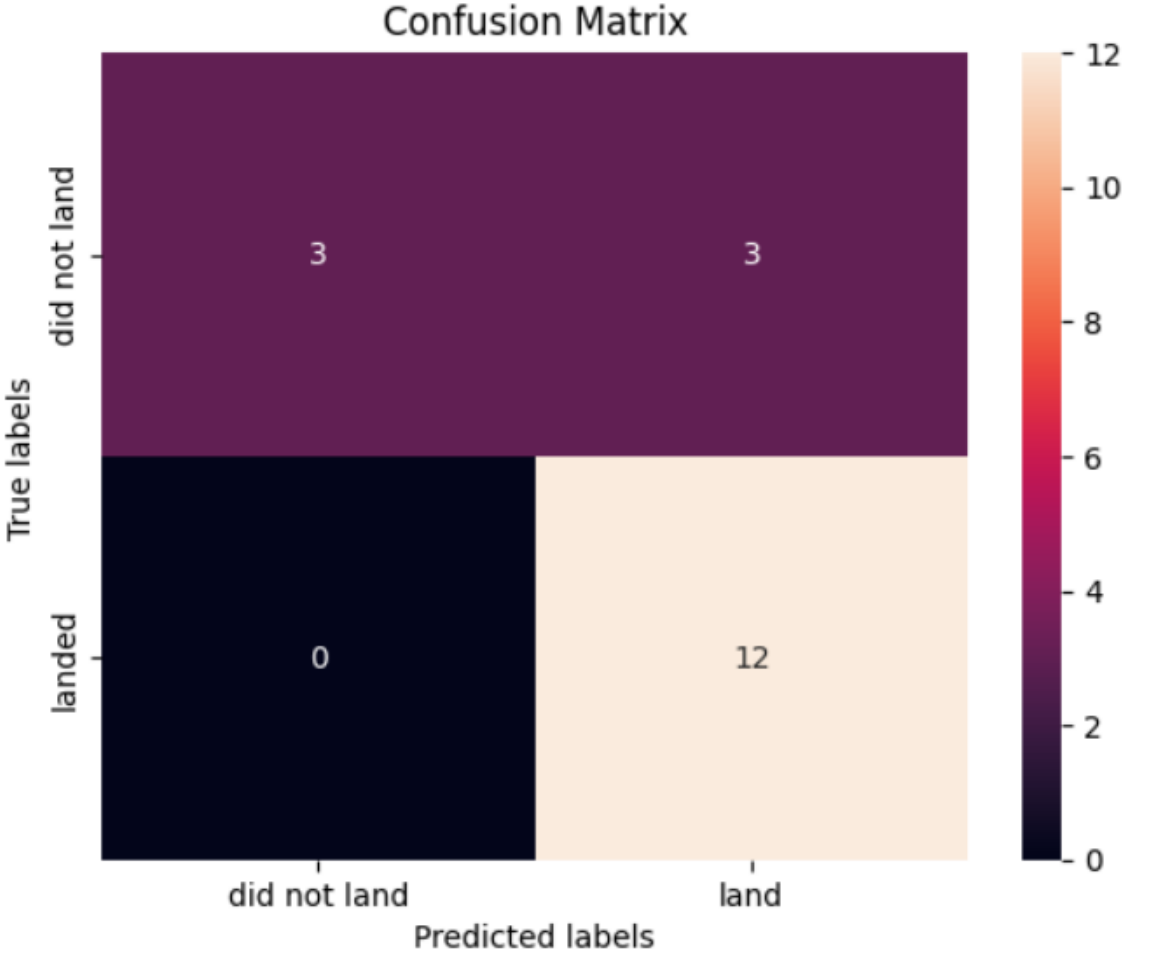
Predictive Analysis (Classification)

Classification Accuracy



- All the models have the same accuracy.

Confusion Matrix



Overview:

- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)

Conclusions

- Model Performance: All classifiers achieved similar accuracy (~83.3%), indicating strong baseline predictability but limited feature discrimination
- Key Findings: KSC LC-39A most successful site; Optimal payload: 3,000-6,000 kg; Clear improvement trend over time
- Business Impact: Framework enables cost prediction for competitive bidding against SpaceX
- Technical Success: Integrated API, web scraping, SQL, and machine learning in end-to-end pipeline
- Limitation: Class imbalance challenges rare failure prediction; dataset reflects SpaceX's high success rate
- Recommendation: Focus on anomaly detection and probability-based risk scoring for improved failure forecasting

Thank you!