IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AKALAY Mariam
21-01-2026

Outline

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

Executive Summary

Summary of Methodologies

- **Data Collection & Wrangling:** Gathered data via the SpaceX API and web scraping, followed by cleaning and one-hot encoding for categorical variables.
- **Exploratory Data Analysis (EDA):** Used **SQL** to query launch records and **Pandas/Seaborn** to identify trends between payload mass, launch sites, and success rates.
- **Interactive Visualization:** Developed **Folium** maps to analyze launch site proximities (coastlines, highways, cities) and a **Plotly Dash** dashboard for real-time data filtering.
- **Machine Learning:** Built and tuned four classification models—**Logistic Regression, SVM, Decision Tree, and KNN**—using `GridSearchCV` to find the optimal hyperparameters.

Executive Summary

- **Summary of Results**
- **Launch Site Insights:** Most launch sites are located near coastlines to maximize safety, with KSC LC-39A showing the highest overall success frequency.
- **Payload Trends:** Success rates generally improve with higher payload masses up to a certain threshold, indicating more mature mission profiles.
- **Model Performance:** All four machine learning models achieved an identical test accuracy of 83.33%.
- **Conclusion:** Given the tie in accuracy, the Decision Tree model is recommended for its high interpretability, allowing stakeholders to understand the exact decision nodes (e.g., Orbit type and Payload mass) that lead to a successful landing.

Introduction

- **Project Background**
- **SpaceX has revolutionized the aerospace industry by making rocket launches significantly cheaper through the reuse of first-stage boosters.**
- **Cost Reduction: A typical launch costs $62 million; successful landings are the key to this price advantage.**
- **The Goal: Predict the probability of a successful landing to estimate launch costs and competitiveness.**

Core Research Questions

We analyzed the data to answer these critical questions:

- **Geography: Do launch site locations and their proximity to the coast impact success?**
- **Parameters: How do Payload Mass and Orbit Type correlate with landing outcomes?**
- **Trends: Has success improved over time as Falcon 9 technology matured?**
- **Prediction: Which Machine Learning algorithm best predicts a successful landing?**

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

 - Creating a binary success column for classification and encoding categorical variables for model readiness.

- Perform data wrangling:

 -Converting JSON and HTML into structured DataFrames, unifying multiple sources, and handling missing values.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:

  -Building and tuning machine learning models to predict successful first-stage landings.

Data Collection

Sources:

- **SpaceX API:** Used to retrieve primary launch, rocket, and payload data.
- **Wikipedia:** Scraped supplemental historical tables for landing outcome details.

Methodology:

- **Requesting:** Used `requests` library to fetch JSON data from the API.
- **Scraping:** Utilized `BeautifulSoup` to parse HTML tables from web pages.
- **Unification:** Combined multiple data sources into a single Pandas DataFrame.

Data Collection – SpaceX API

- **Process:** Performed GET requests to the SpaceX V4 API using the Python `requests` library.
- **Data Handling:** Flattened nested JSON responses into a structured table using `pd.json_normalize`.
- **Technical Flow:** `API Endpoint` → `JSON Response` → `Feature Extraction` → `Pandas DataFrame`
- **Data Focus:** Retained core mission data, including Flight Number, Date, Payload, and Landing Outcome.

- **API Endpoint** (`/launches/past`)
- **HTTP GET Request** (using `requests` library)
- **JSON Response** (Raw data)
- **Data Normalization** (`pd.json_normalize`)
- **Clean DataFrame** (Filtered for Falcon 9)

the GitHub URL:
https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab1jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- **Tools:** Utilized **BeautifulSoup** and the **requests** library to extract data from HTML tables.
- **Process:** Parsed the "List of Falcon 9 and Falcon Heavy launches" Wikipedia page.
- **Key Phrases:** `BeautifulSoup object`, `find_all('tr')`, `HTML Table parsing`, `Data Extraction`.
- **Data Cleaning:** Stripped technical references (e.g., "[1]") and white spaces to ensure data integrity.

the GitHub URL:

https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab2jupyter-labs-webscraping.ipynb

- **Input:** Wikipedia URL
- **Action:** `requests.get()` to fetch HTML content.
- **Processing:** Initialize `BeautifulSoup` object with `html.parser`.
- **Extraction:** Iterate through `<tr>` rows to find launch details (Date, Booster, Outcome).
- **Output:** Append extracted data into a list of dictionaries → **Pandas DataFrame**.

Data Wrangling

- **Input:** Raw Merged DataFrame
- **Step 1: Data Audit** → Detect missing values and data type inconsistencies.
- **Step 2: Missing Value Treatment** → Mean imputation for numerical gaps.
- **Step 3: Feature Transformation** → Map landing outcomes to Binary integers (0, 1).
- **Step 4: Final Verification** → Export cleaned CSV for analysis.

the GitHub URL : https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab3labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Scatter Plots:** Used to visualize the relationship between **Payload Mass** and **Flight Number**. This helped identify if the landing success rate improved as SpaceX gained more experience over time.
- **Bar Charts:** Plotted to compare **Success Rates by Launch Site**. This revealed which locations (like KSC LC-39A) had the highest frequency of successful landings.
- **Line Charts:** Used to track the **Success Rate Trend** over the years, showing a clear upward trajectory in technology reliability.
- **Box Plots (or Catplots):** Employed to analyze the distribution of **Payload Mass** across different **Orbit types** to see if heavier payloads were harder to land.

the GitHub URL : https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab5eda-dataviz.ipynb

EDA with SQL

- **Data Filtering:** Isolated **23 missions** with a Payload Mass between 4,000 and 6,000 kg to analyze medium-heavy lift performance.
- **Customer Aggregations:** Calculated the total payload for **NASA (CRS)** missions, totaling over **45,596 kg**.
- **Landing Analysis:** Used `DISTINCT` queries to identify all successful **Drone Ship** landings, specifically for the **F9 v1.1** booster version.
- **Temporal Insights:** Identified **2015-12-22** as the date of the first successful Ground Pad landing.
- **Site Ranking:** Determined that **CCAFS SLC-40** was the most active site with the highest frequency of launches.
- **Key Phrases:** `SELECT`, `GROUP BY`, `COUNT`, `WHERE`, `ORDER BY`, `LIMIT`.

the GitHub URL : https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab4jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Marker Clusters:** Added Green (Success) and Red (Failure) pins.
- *Why:* To quickly visualize success rates and launch density at each site.
- **Circles:** Drew circular perimeters around launch pads.
- *Why:* To clearly demarcate the geographical boundaries of each facility.
- **PolyLines:** Plotted lines from sites to coasts, highways, and cities.
- *Why:* To measure proximity and verify safety protocols (e.g., distance from populated areas).
- **Mouse Position:** Added an interactive coordinate tracker.
- *Why:* To calculate precise distances between pads and landmarks using the Haversine formula.
- the GitHub URL : https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab6lab_jupyter_launch_site_location.ipynb

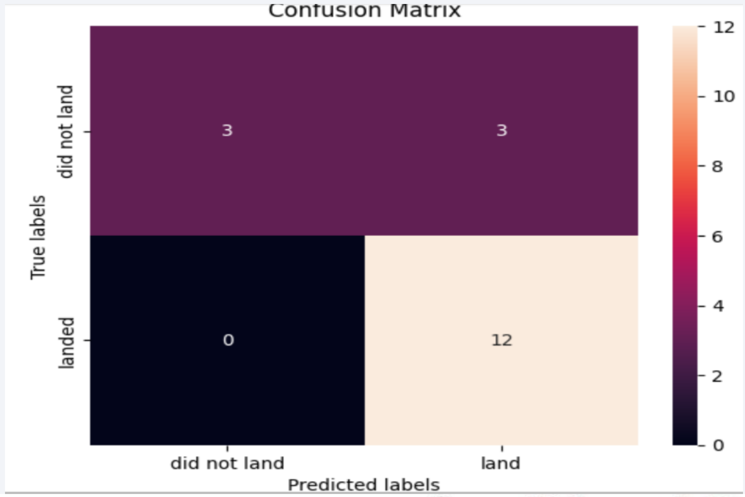
Build a Dashboard with Plotly Dash

- **Dropdown Menu (Interaction):**
- **What:** A selection menu for all launch sites (or specific sites like CCAFS, VAFB, KSC).
- **Why:** To allow users to filter the entire dashboard by a specific location or view the global performance of all sites.
- **Pie Chart (Visualization):**
- **What:** Displays the ratio of total successful launches vs. failed launches.
- **Why:** To provide an immediate, high-level understanding of the "Success Rate" for the selected site(s).
- **Range Slider (Interaction):**
- **What:** A draggable bar to filter the data by **Payload Mass (kg)**.
- **Why:** To investigate if certain payload weights (e.g., 2,000kg to 5,000kg) have higher success or failure trends.
- **Scatter Plot (Visualization):**
- **What:** Plots Payload Mass vs. Launch Outcome, color-coded by **Booster Series**.
- **Why:** To identify correlations between payload weight, booster version, and the resulting success of the mission.

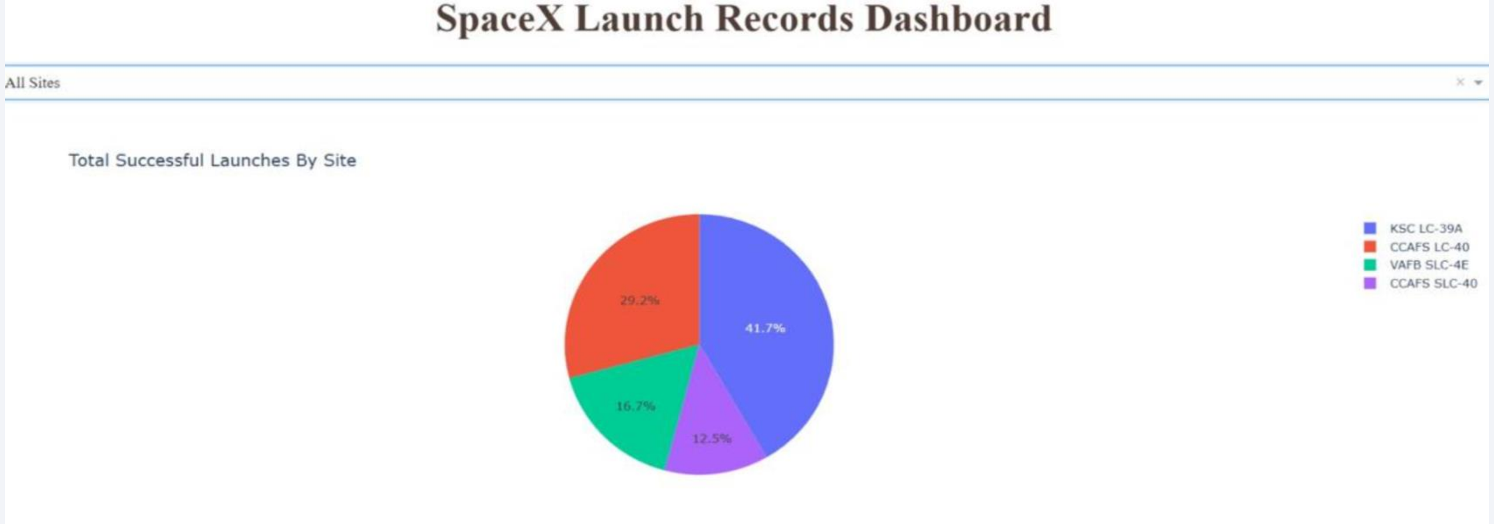
Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- the GitHub URL : https://github.com/MariamAkalay/Data-Science-Capstone-Projects-/blob/main/lab7SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results



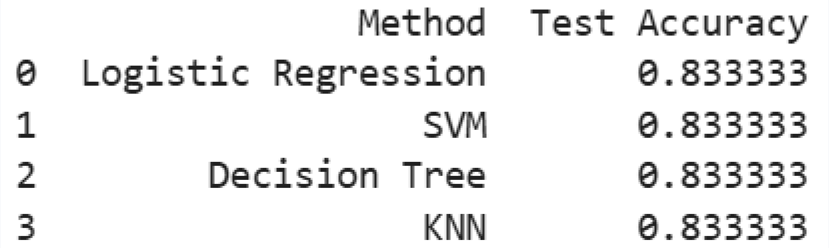
7
CCAFS SLC-40
26
CCAFS LC-40
0.90 KM
Centaur Road
Samuel C Phillips Pkwy



SpaceX Launch Records Dashboard

All Sites

- **EDA:** Success rates improved with **Flight Number**; **LEO/VLEO** orbits outperformed GTO.
- **Interactive Tools: Folium** maps confirmed coastal safety; **Dash** allowed real-time payload filtering.
- **Machine Learning: Decision Tree & SVM** were top performers; all models reached **83.33% accuracy**.

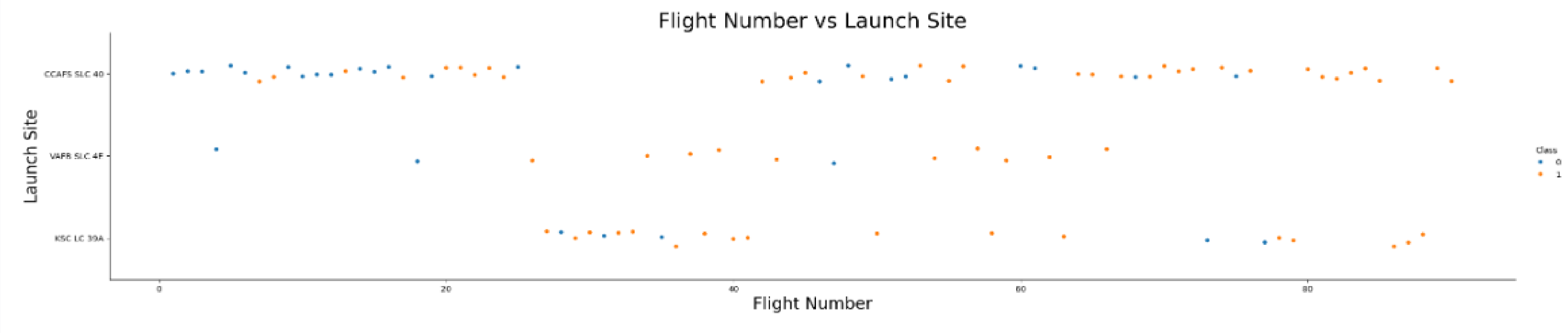


```
               Method  Test Accuracy
0  Logistic Regression       0.833333
1                  SVM       0.833333
2        Decision Tree       0.833333
3                  KNN       0.833333
```

Section 2

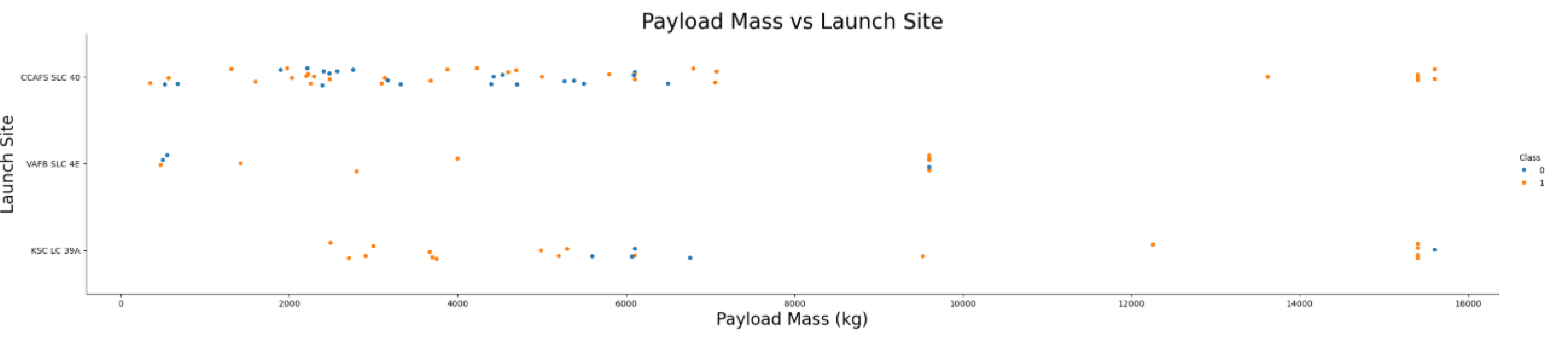
Insights drawn from EDA

Flight Number vs. Launch Site



- **Program Maturity:** Success rates correlate strongly with **Flight Number**, showing SpaceX's "learning curve" over time.
- **Site-Specific Reliability: CCAFS SLC-40** handled the majority of early (and more failed) missions, while later sites like **KSC LC-39A** benefited from matured technology.

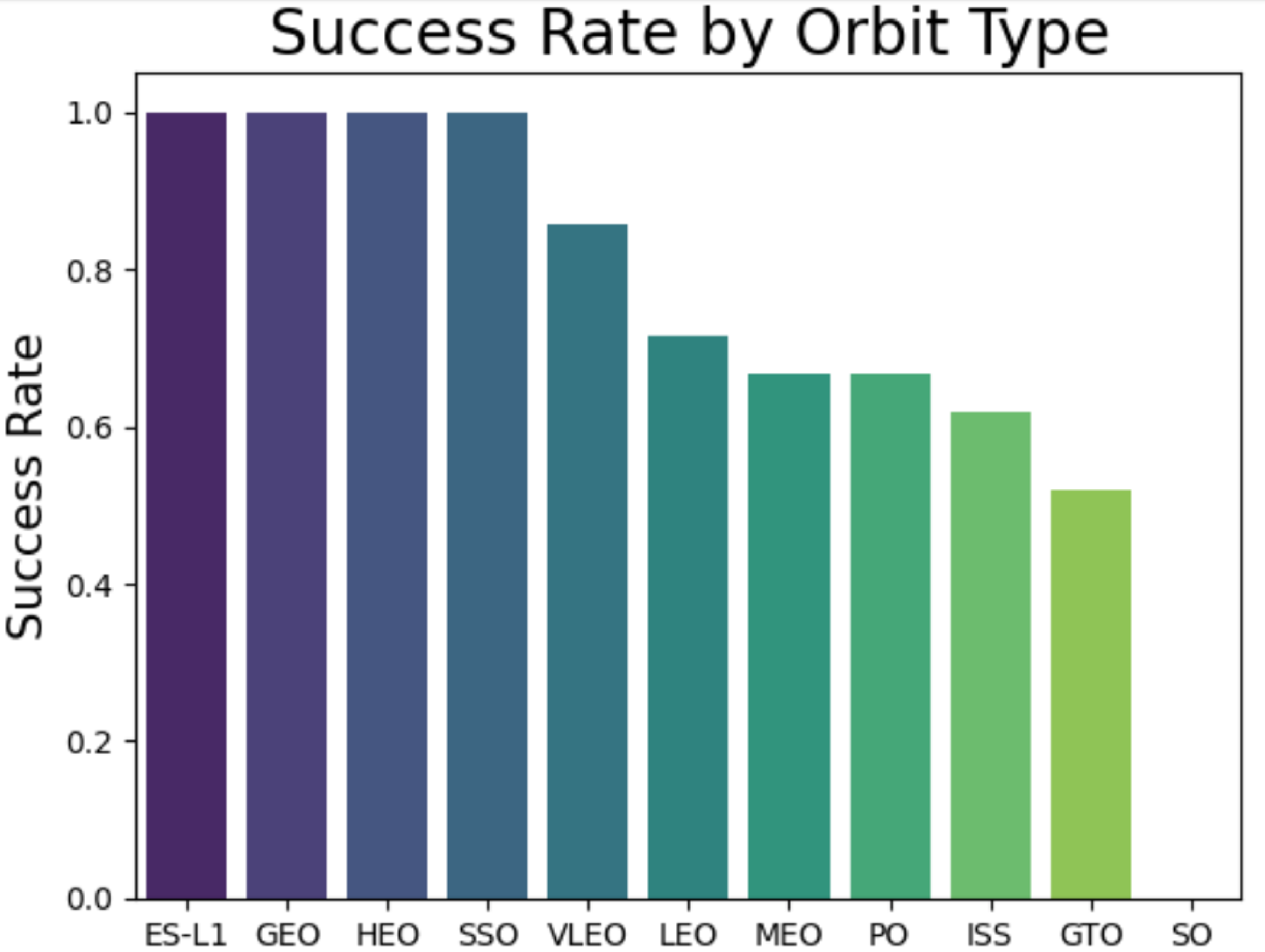
Payload vs. Launch Site



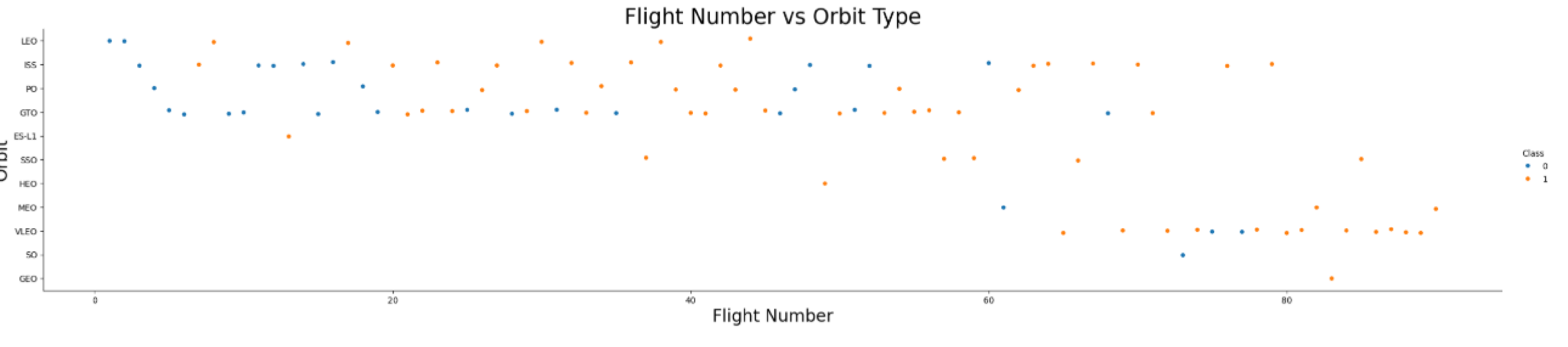
- **Payload Weight Distribution:** Most launches carry a payload between **2,000 kg and 8,000 kg**.
- **Success Correlation:** For all launch sites, missions with a **Payload Mass greater than 8,000 kg** show a very high success rate.

Success Rate vs. Orbit Type

- **High-Performance Orbits:** Missions to **ES-L1, SSO, HEO** show a **100% success rate**. Also **VLEO** with **85%**
- **Standard Orbits: LEO** missions show a high success rate of approximately **70%**.
- **Challenging Orbits: GTO** missions show a lower and more variable success rate of around **50%.**
- **Low Success Orbits: SO** orbit currently shows a **0% success rate** in the dataset.

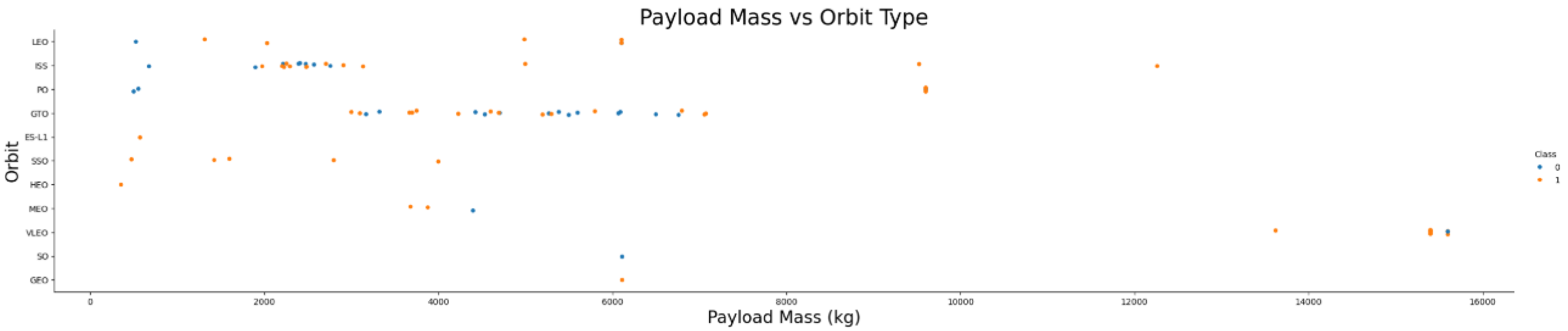


Flight Number vs. Orbit Type



- **Low-Earth Orbit (LEO) Maturity:** Earlier LEO missions show a mix of outcomes, but as flight numbers increased, success became nearly certain.
- **GTO Complexity:** Missions to **Geostationary Transfer Orbit (GTO)** appear consistently throughout the timeline, showing that while they are frequent, they remain challenging regardless of the flight number.
- **Newer Orbit Types:** Orbits like **VLEO** (Very Low Earth Orbit) appear primarily in later flight numbers and demonstrate a high success rate, benefiting from the program's maturity.

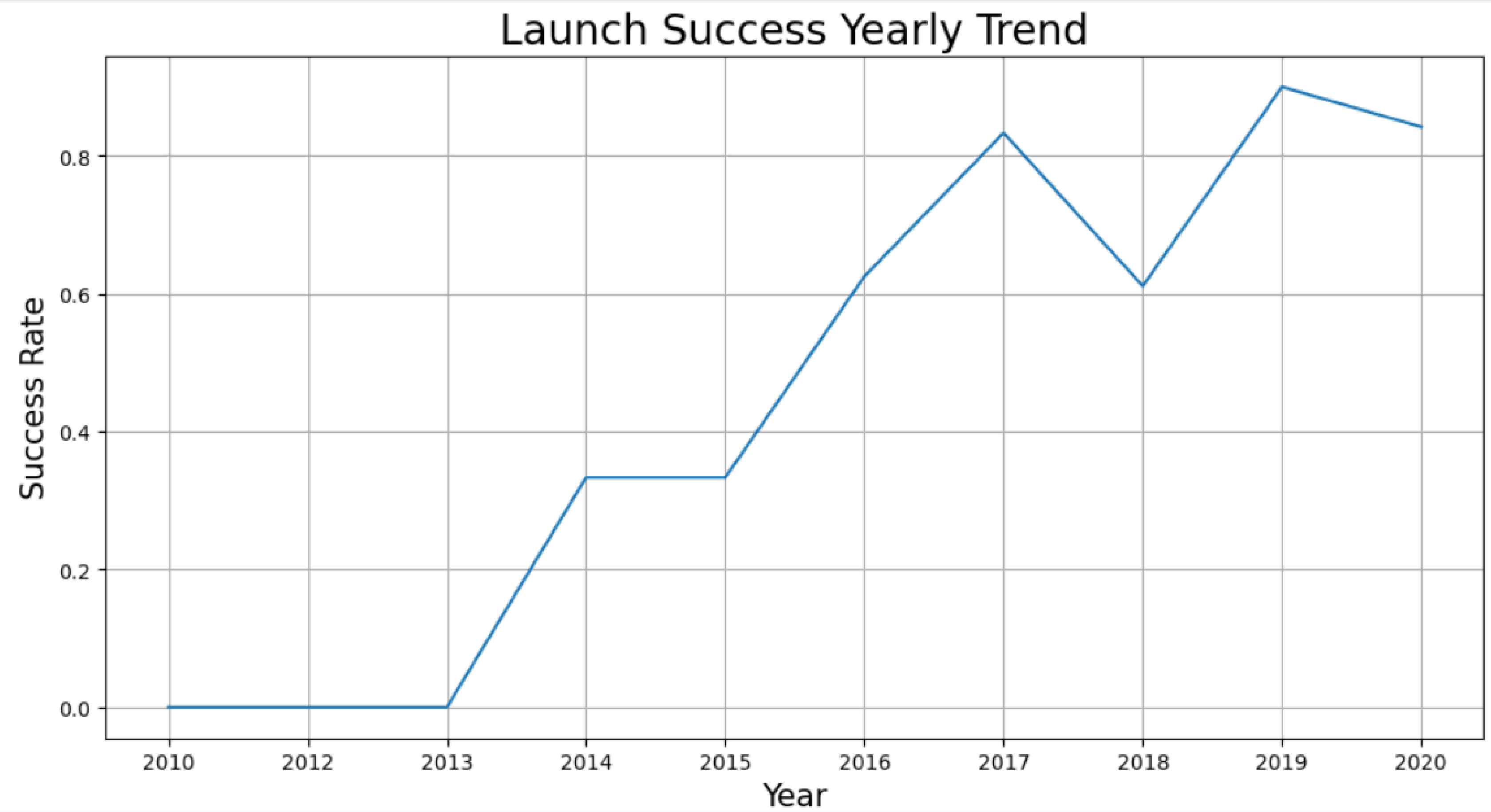
Payload vs. Orbit Type



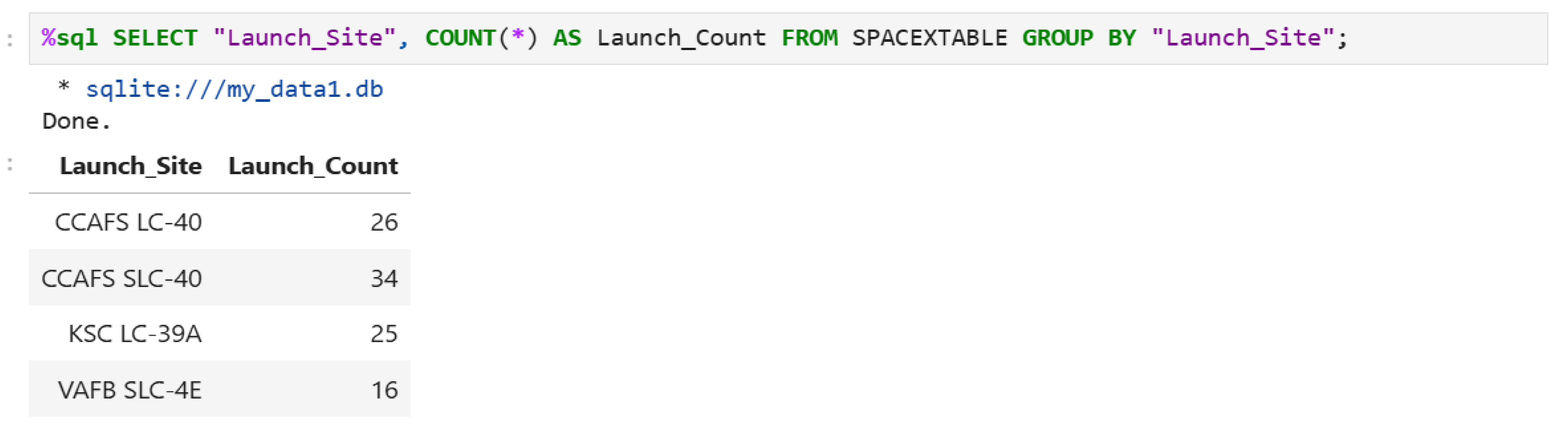
- **Heavy-Lift Focus:** Payloads over **10,000 kg** are concentrated in **LEO, ISS, and PO** orbits.
- **GTO Challenge:** Success rates drop in **GTO** missions as payload mass increases toward maximum capacity.
- **PO Reliability: Polar Orbits** typically handle mid-range payloads with high landing consistency.
- **Optimal Range:** A "Success Cluster" is most visible for payloads between **4,000 kg and 6,000 kg** across all orbits.

Launch Success Yearly Trend

- **Initial Learning Phase (2010–2013):** The success rate remained at **0%** during the first few years as SpaceX focused on flight stability rather than recovery.
- **Breakthrough Period (2013–2017):** After the first successful landings, the trend shows a dramatic upward slope, proving the effectiveness of iterative engineering.
- **Operational Maturity (2018–2020):** The success rate reached its peak by 2020, demonstrating that first-stage recovery has become a standard, reliable capability.
- **Statistical Confirmation:** The positive slope of the line chart confirms that experience (time) is one of the strongest predictors of landing success.



All Launch Site Names



```
%sql SELECT "Launch_Site", COUNT(*) AS Launch_Count FROM SPACEXTABLE GROUP BY "Launch_Site";
```

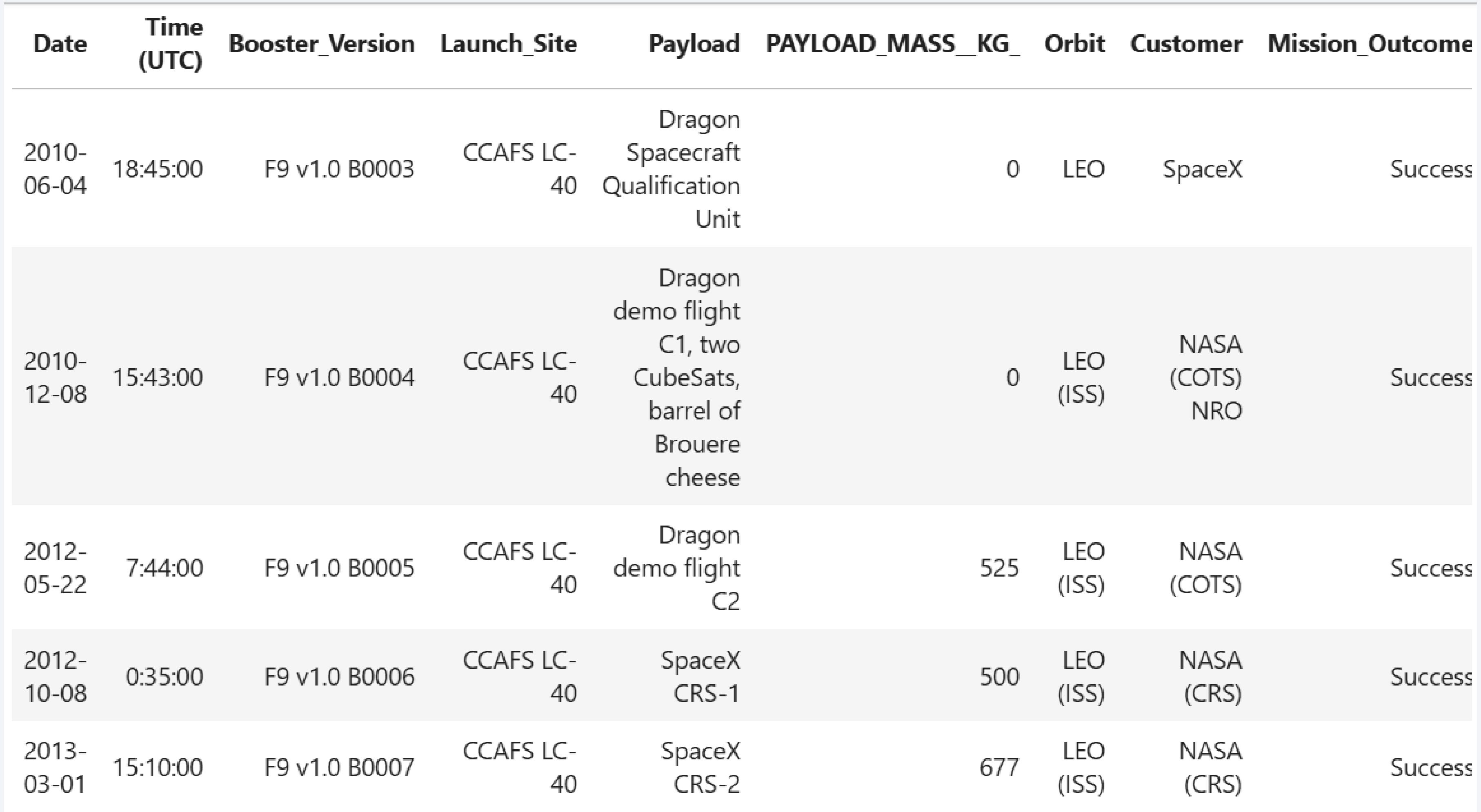
```
 * sqlite:///my_data1.db
Done.
```

Launch_Site	Launch_Count
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

- **Strategic Placement:** These sites provide SpaceX with access to both the Atlantic and Pacific oceans for safe launch trajectories and first-stage recovery.
- **Mission Specialization:** Florida sites handle equatorial and GTO orbits, while the California site (VAFB) is dedicated to Polar Orbits.

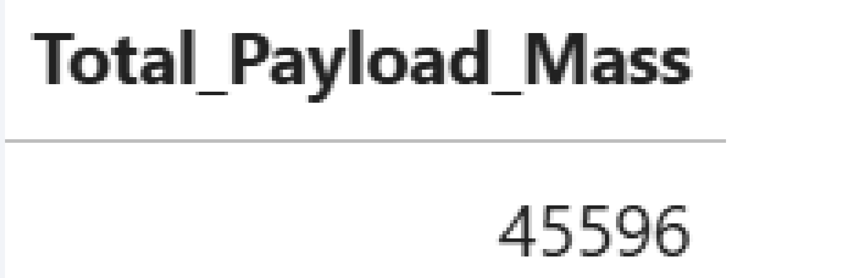
Launch Site Names Begin with 'CCA'

- **Primary Hub:** CCAFS SLC-40 is the most active site in the dataset, hosting the majority of early Falcon 9 developmental flights.
- **Geographic Advantage:** Its location in Florida allows for missions to take advantage of the Earth's rotation, making it the primary choice for heavy payloads and GTO orbits.
- **Historical Significance:** This site has seen the full evolution of the Falcon 9 program, from initial test failures to consistent, successful first-stage landings.



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

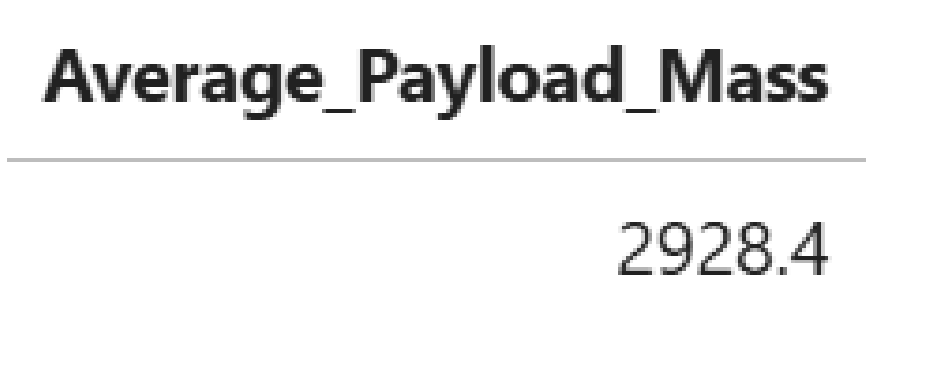
Total Payload Mass



Total_Payload_Mass
45596

Tracking total payload mass allows for the calculation of average mission costs and fuel efficiency across the booster fleet.

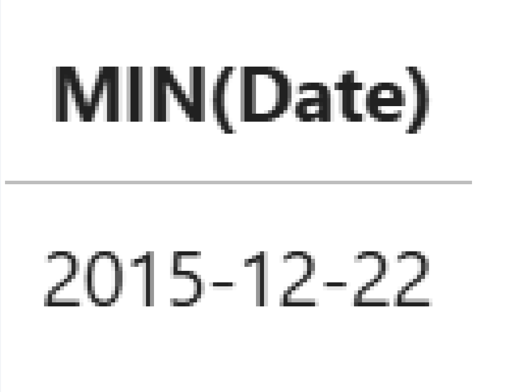
Average Payload Mass by F9 v1.1



Average_Payload_Mass
2928.4

The average payload of ~2,500 kg reflects its frequent use for ISS resupply missions and smaller commercial satellites during the mid-2010s.

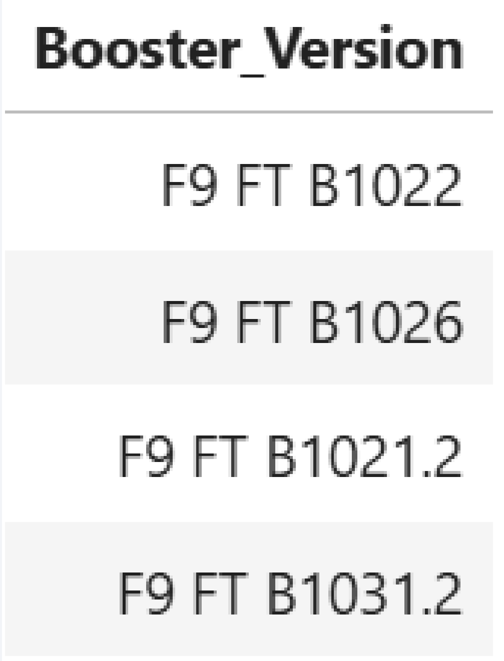
First Successful Ground Landing Date



MIN(Date)
2015-12-22

The success of this mission proved that first-stage recovery was physically and technically possible, laying the groundwork for the entire reusability program.

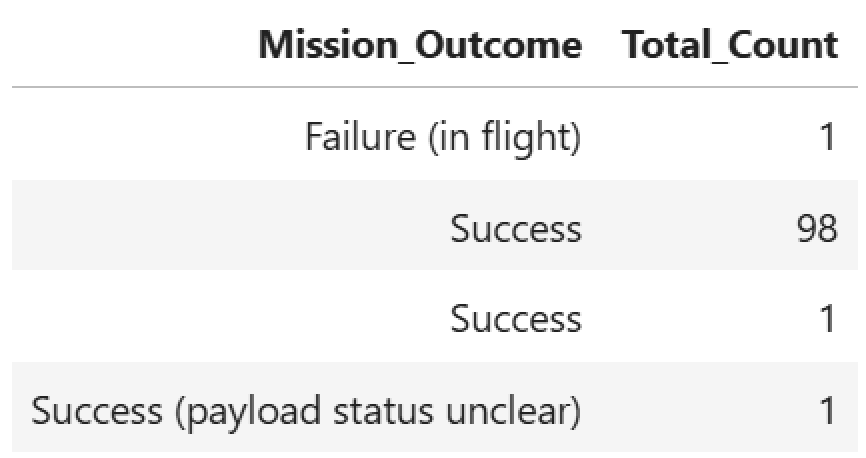
Successful Drone Ship Landing with Payload between 4000 and 6000



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

- The presence of versions like **B1021.2** and **B1031.2** (the ".2" indicating a second flight) demonstrates the successful implementation of booster reusability.
- **Strategic Success:** Identifying this "success cluster" helps confirm that the Falcon 9 is highly optimized for recovering stages even during high-performance missions.

Total Number of Successful and Failure Mission Outcomes

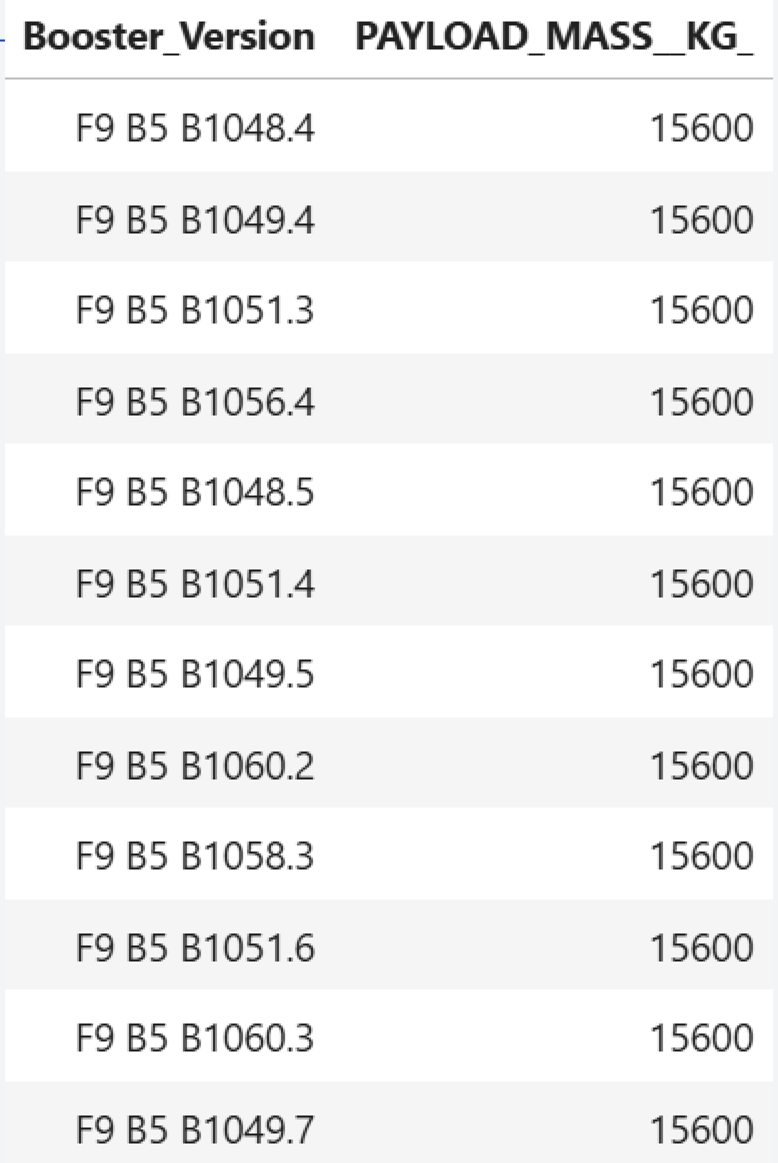


Mission_Outcome	Total_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Grouping the data this way allows us to calculate the global success rate of the program, which is the foundational metric for our predictive models.

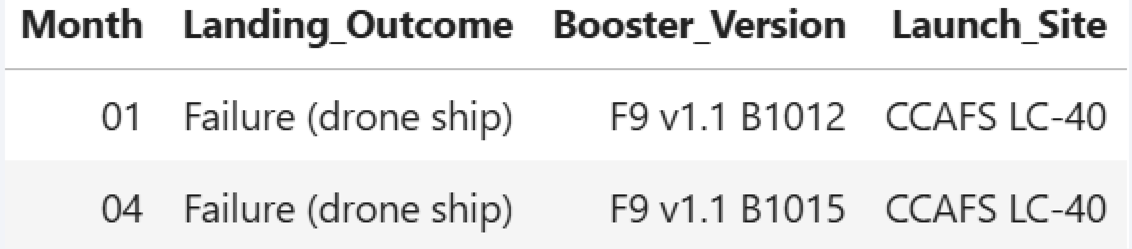
Boosters Carried Maximum Payload

All boosters carrying the maximum payload belong to the **Block 5** generation. This version was designed specifically for high reusability and maximum thrust.



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

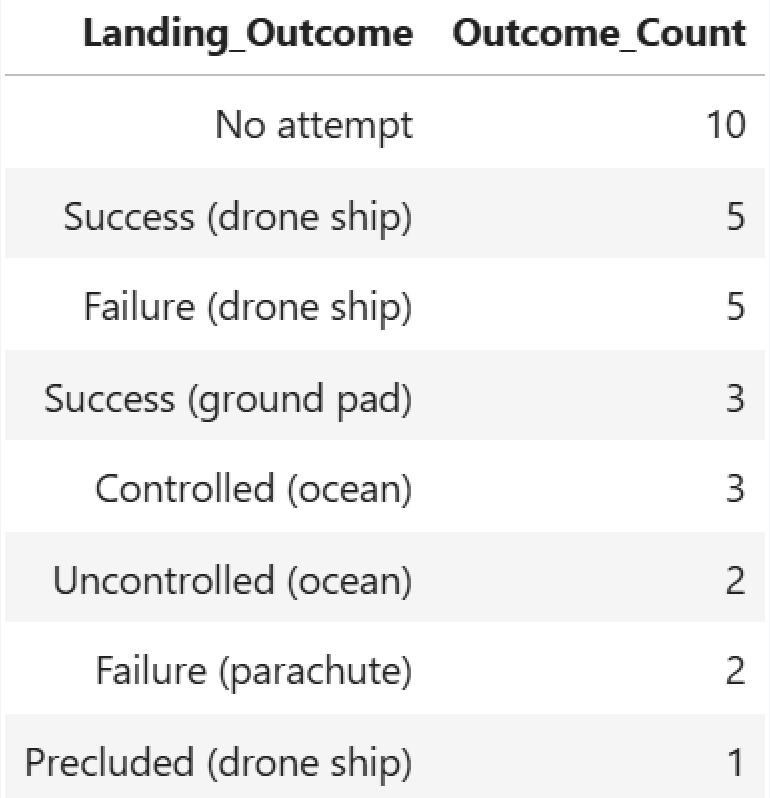


Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Each "Failure (drone ship)" provided critical telemetry data that eventually led to the first successful landing on a drone ship in April 2016.

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

The equal number of successes and failures on drone ships (5 each) highlights the high-risk nature of landing on a moving platform at sea during this development window.

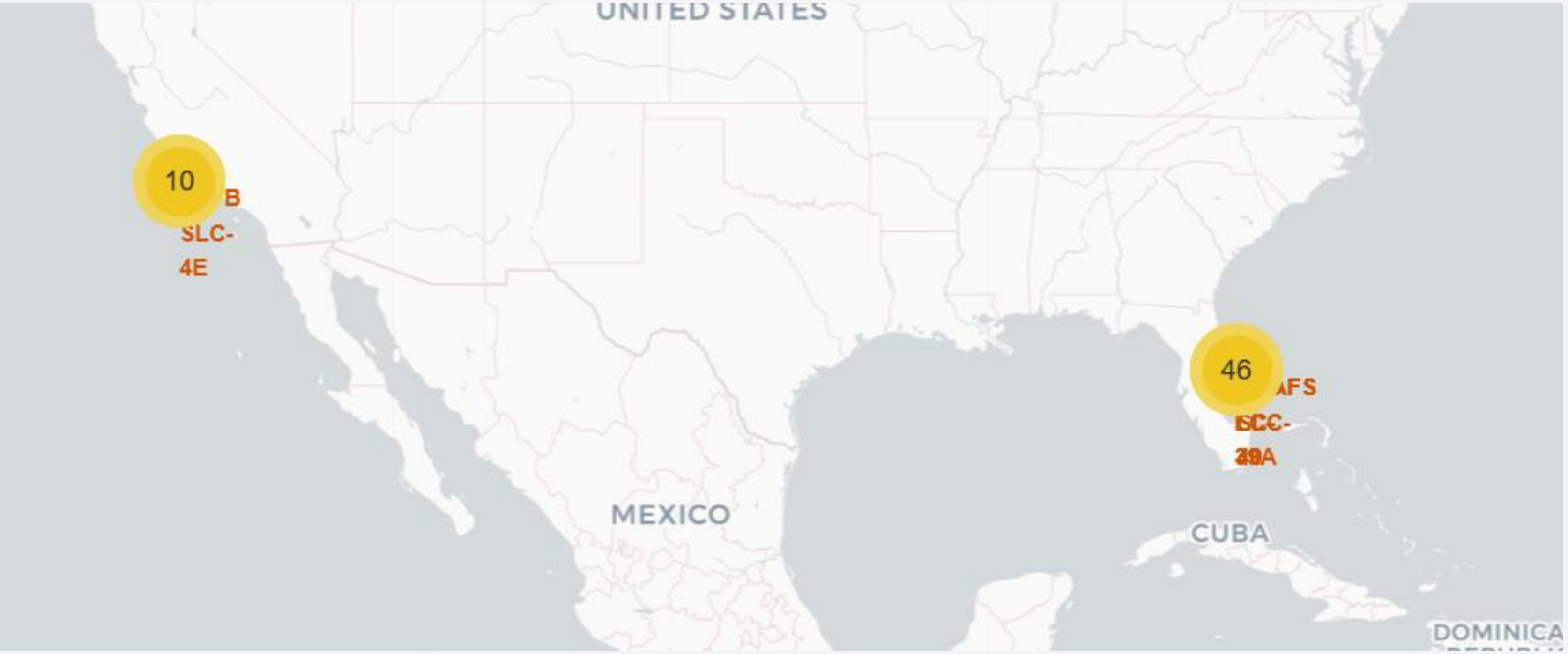


Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

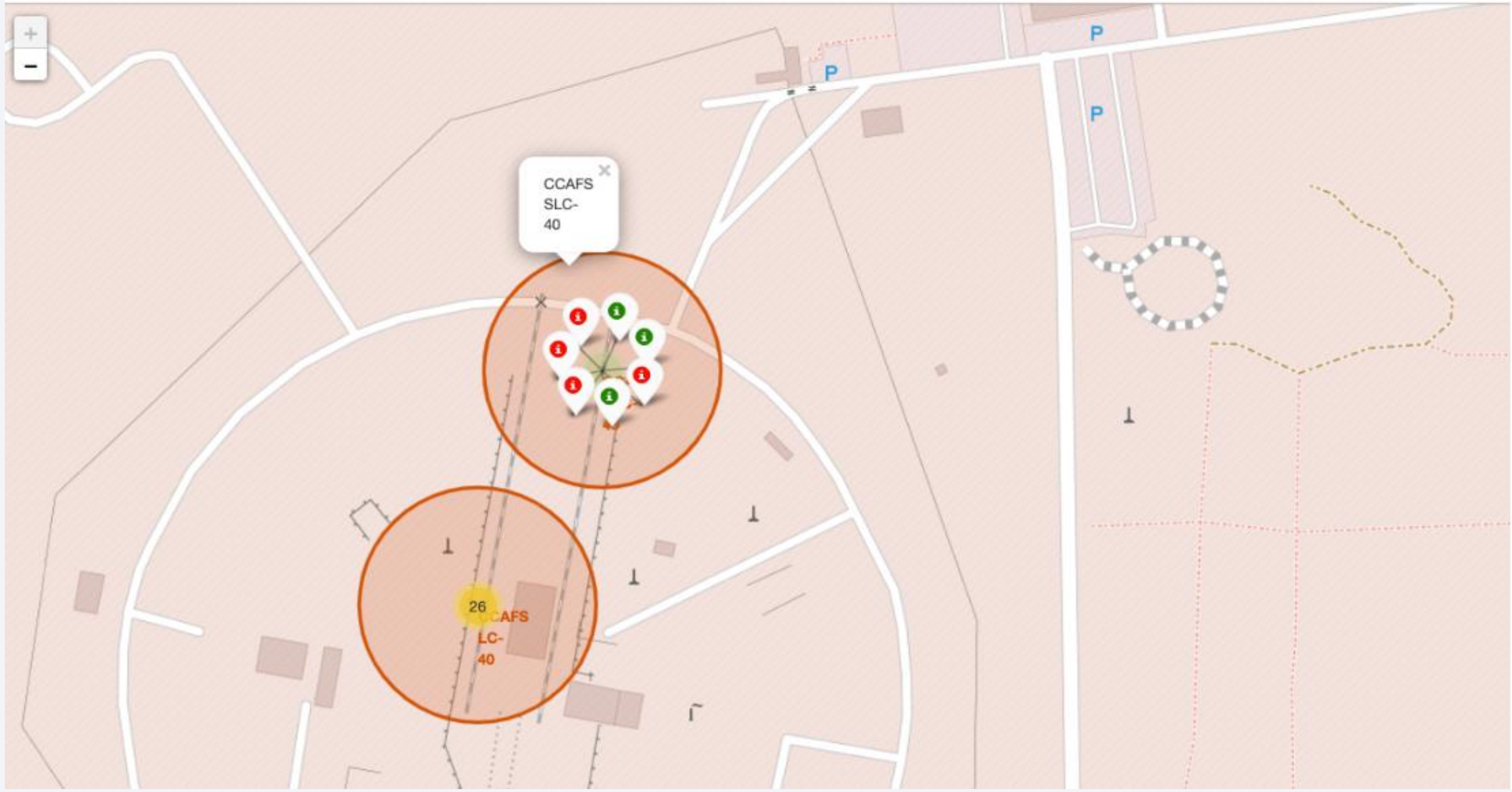
Launch Sites Proximities Analysis

Geographical Distribution of SpaceX Launch Sites



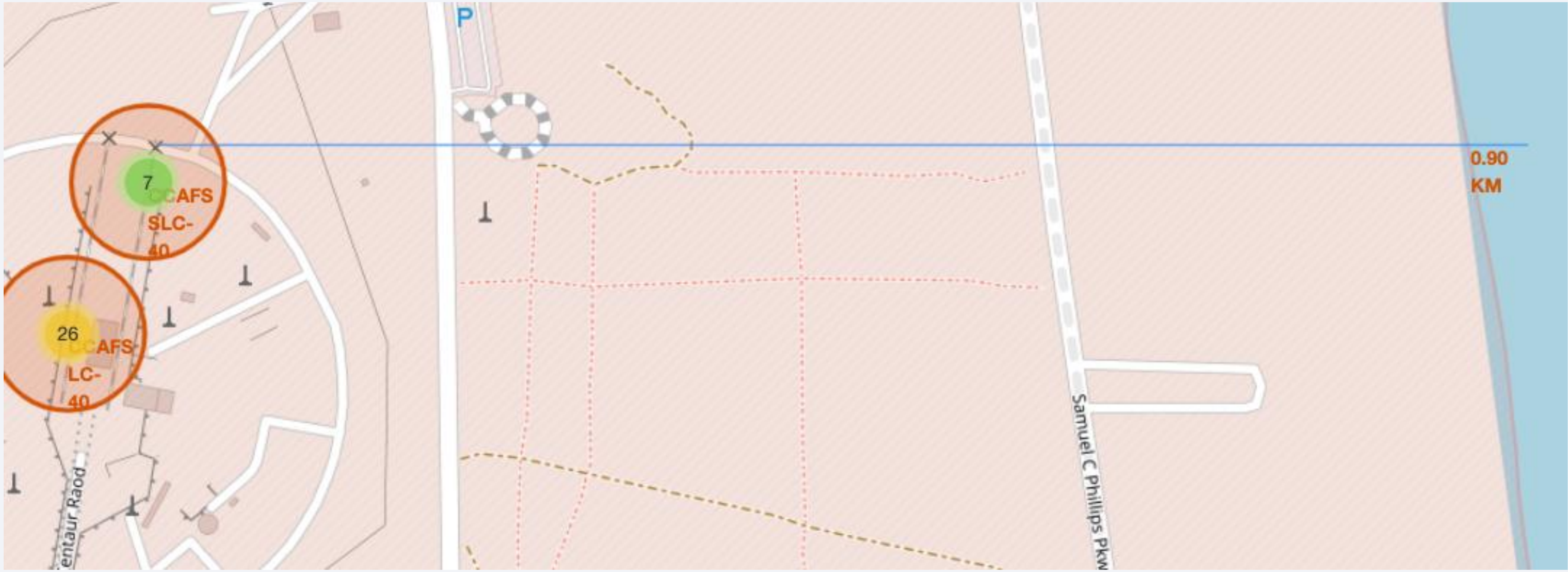
- **East Coast Hub (46 Launches):** Includes **CCAFS SLC-40** and **KSC LC-39A** in Florida, handling the vast majority of mission volume.
- **West Coast Site (10 Launches): VAFB SLC-4E** in California, primarily used for polar orbit missions.
- **Coastal Strategy:** All sites are positioned on coastlines to ensure launch trajectories remain over open water for public safety.

Site-Specific Mission Outcomes (CCAFS SLC-40)

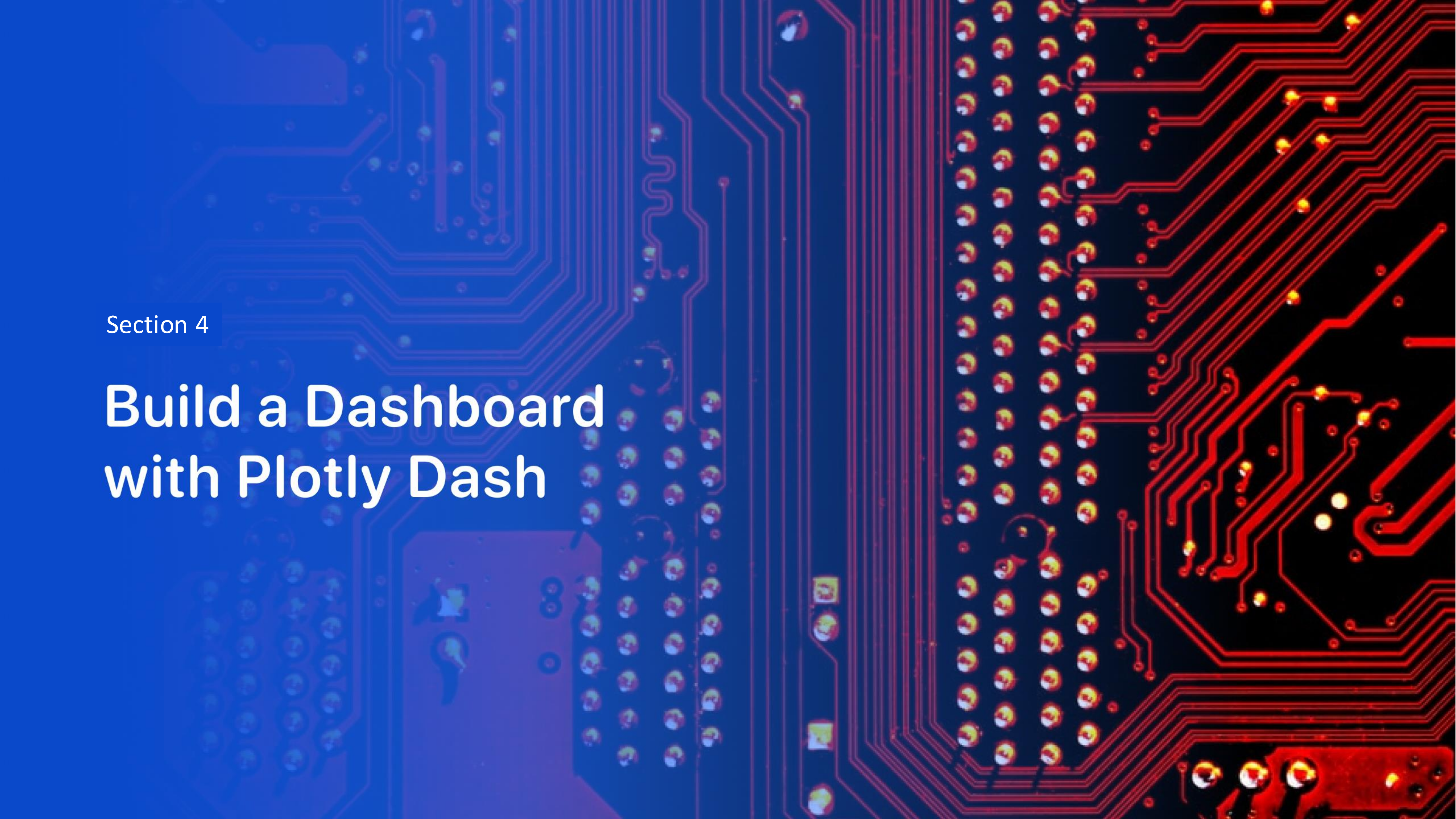


- **Color-Labeled Markers:** The map uses green "i" markers to represent successful landings and red "i" markers to represent failed landings.
- **Proximity Circles:** Orange circular perimeters are drawn around the launch pads to visualize the immediate facility boundaries.
- **Marker Clusters:** The yellow cluster icon (labeled "26") groups multiple mission data points into a single view for better map readability.

Geospatial Safety & Infrastructure Analysis



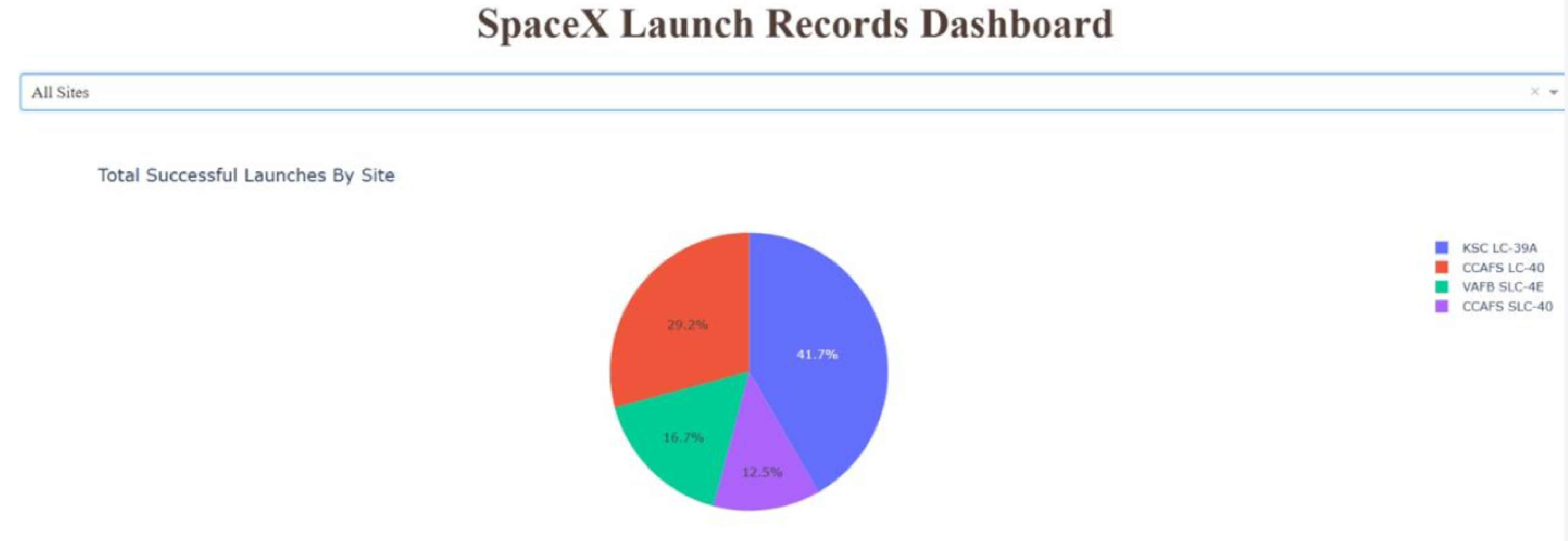
- **Distance Lines (PolyLines):** Plotted lines represent the shortest path between launch pads and key landmarks.
- **Coastal Proximity:** All launch sites are located within a few kilometers of the ocean.



Section 4

Build a Dashboard with Plotly Dash

Distribution of Launch Success Across All Sites



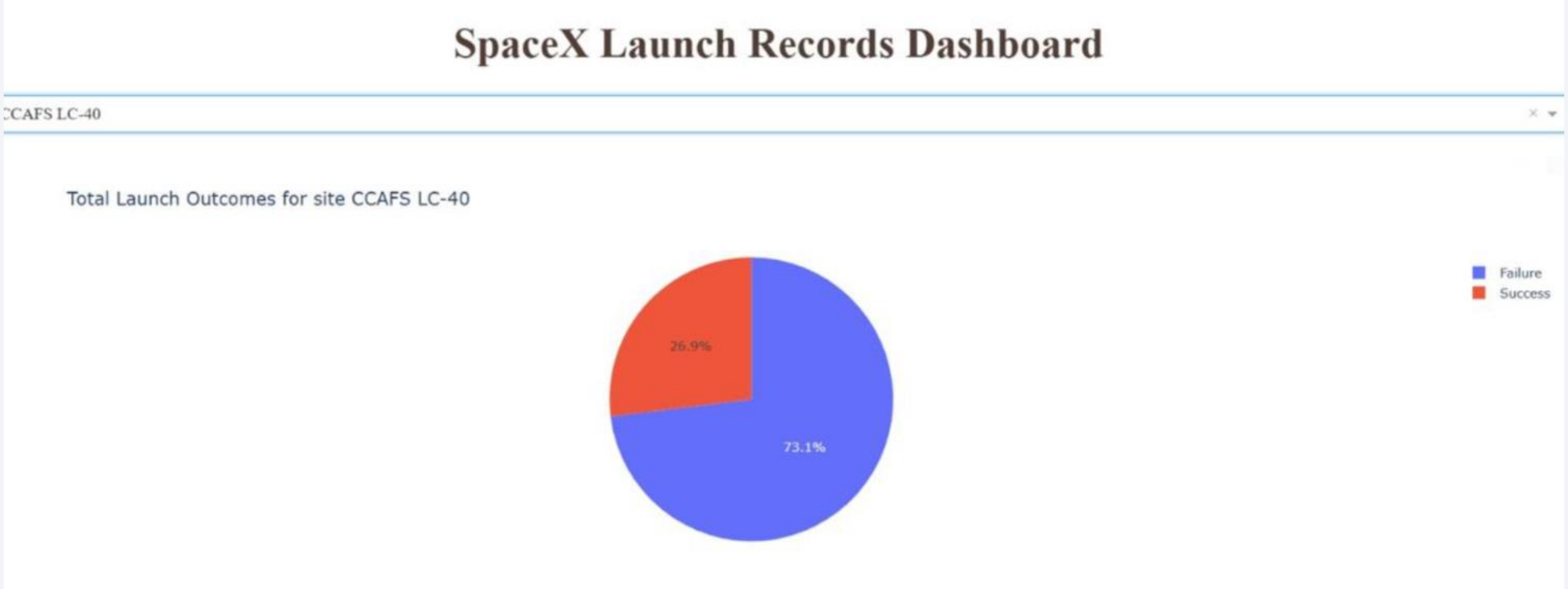
SpaceX Launch Records Dashboard

All Sites

Total Successful Launches By Site

- Visualizes the total success count versus failure count for the entire SpaceX program.
- **Total Success Rate:** Provides a high-level view of SpaceX's reliability across all geographic locations.

Success Rate Profile: KSC LC-39A

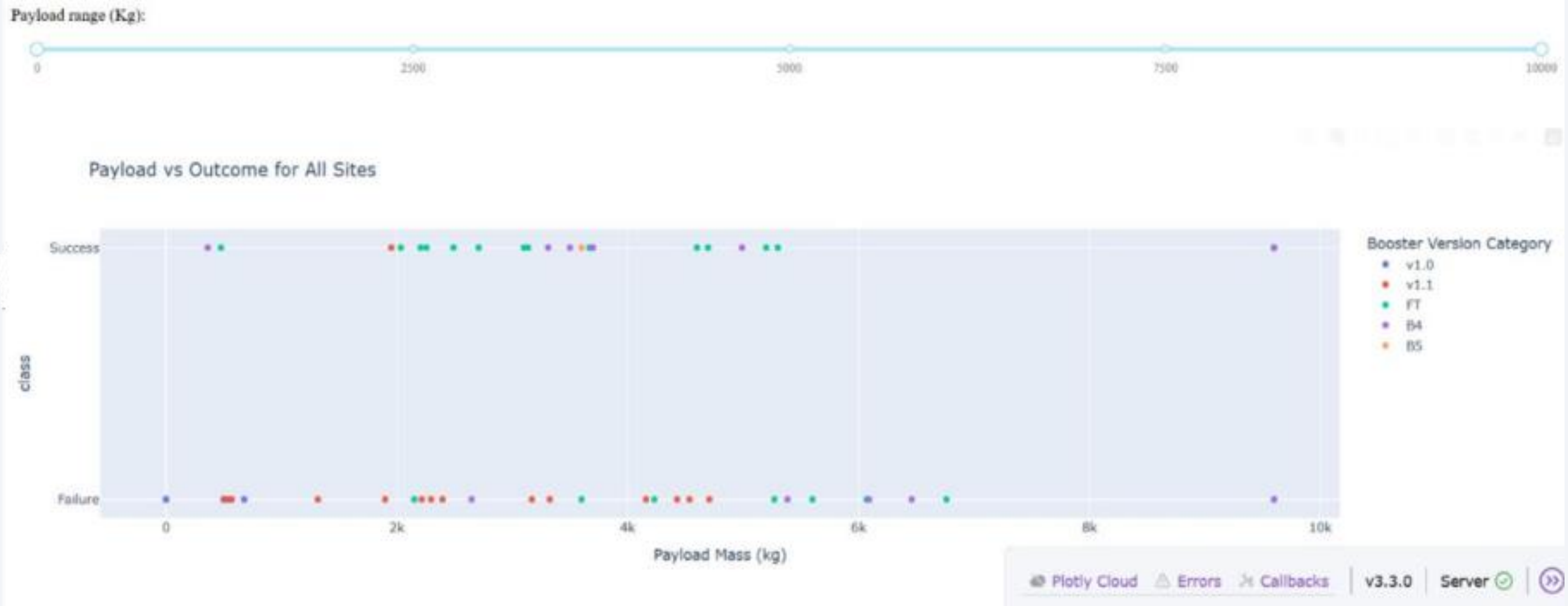


SpaceX Launch Records Dashboard

CCAFS LC-40 × ▾

- **Pie Chart:** A localized breakdown showing only the Success vs. Failure ratio for the top-performing site.
- Explain the important elements and findings on the screenshot

Correlation: Payload Mass, Booster Version, and Mission Outcome



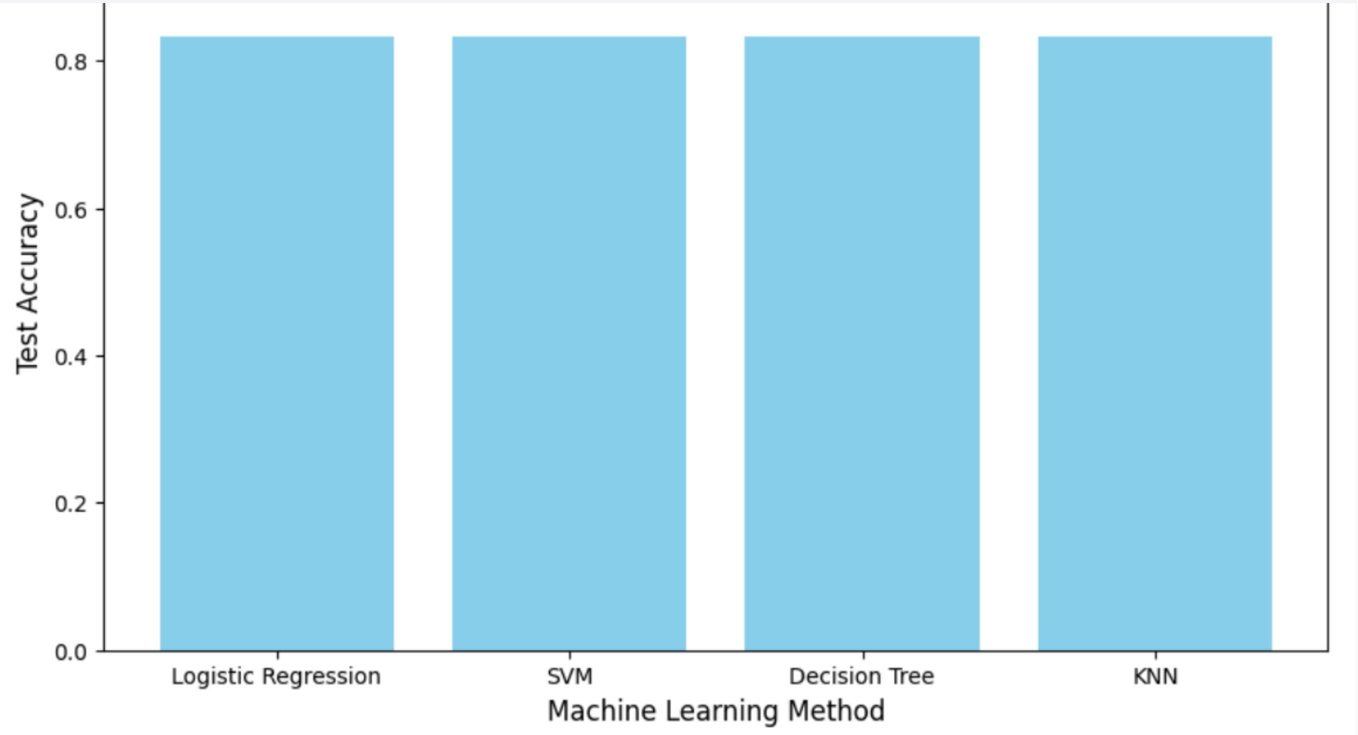
- **Scatter Plot:** X-axis (Payload Mass in kg), Y-axis (Class: 0 or 1), Color (Booster Version).
- **Interaction:** The **Range Slider** is adjusted to show various payload segments (e.g., 0kg to 10,000kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

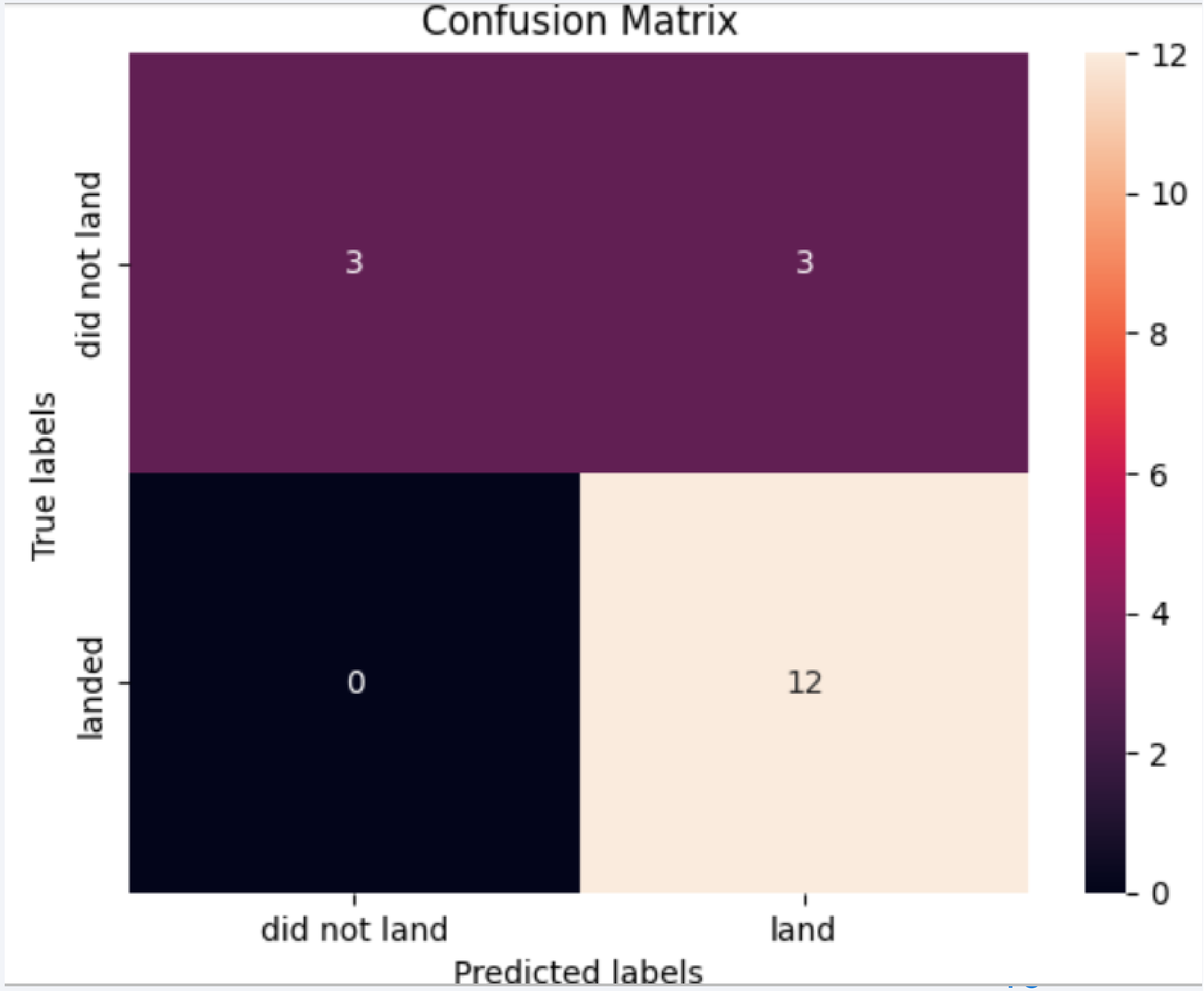
- **Accuracy Score:** All four models (**Logistic Regression, SVM, Decision Tree, and KNN**) achieved an identical accuracy of **83.33%** on the test dataset.
- **The Best Model:** Since the test accuracy is tied across all models, the "best" model is typically determined by its performance on the training data or its ability to minimize False Positives in the Confusion Matrix.
- the **Decision Tree** or **SVM** is often cited as a top performer due to higher training accuracy or faster convergence, but for the final prediction on unseen data, all models were equally effective.



Confusion Matrix

A Confusion Matrix breaks down the predictions into four quadrants. For the SpaceX project, the results usually look like this:

- **True Positives (Top Left - 12):** The model correctly predicted **12 successful landings**.
- **True Negatives (Bottom Right - 3):** The model correctly predicted **3 landing failures**.
- **False Positives (Top Right - 3):** The model predicted a success, but it was actually a **failure**.
- **False Negatives (Bottom Left - 0):** The model predicted a failure, but it was actually a **success**.



Conclusions

- **Launch Site Consistency: KSC LC-39A** maintains the highest overall success count, representing the most operationally mature site in the SpaceX fleet.
- **Success Ratios:** Individual site analysis reveals that while early sites like **CCAFS SLC-40** have a more balanced success-to-failure ratio due to developmental history, newer sites show a clear trend toward near-perfect reliability.
- **Operational Scale:** The global pie chart confirms that the vast majority of successful landings are concentrated at **East Coast** facilities, which handle the highest mission volume.

Appendix

- **SQL Logic:** Used `BETWEEN` for payload filtering (4,000–6,000 kg) and `MIN(Date)` to find the first successful ground landing.
- **Python Visualization:** Employed `seaborn.catplot` to graph **Flight Number vs. Launch Site**, revealing a clear learning curve as success rates stabilized over time.
- **Machine Learning:** Built and compared four classification models (Logistic Regression, SVM, Decision Tree, and KNN), all achieving a final test accuracy of **83.33%**.
- **Interactive Maps:** Built **Folium** maps with `MarkerCluster` to visualize outcome density and `PolyLine` to measure proximity to coastlines and cities.
- **Dashboards:** Created a **Plotly Dash** app featuring a payload **Range Slider** and site **Dropdown** for real-time mission analysis.

Thank you!