IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Carlos Giovanni Garfias Renjel
Date: December 14 2025

Outline

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

Executive Summary

Project Methodologies:

- Data Collection: Acquired launch data from the public SpaceX API and supplemented it through web scraping from Wikipedia.
- Analysis & Visualization: Conducted exploratory data analysis (EDA) using SQL and statistical plots to identify trends in payload mass, orbit, and launch sites. Create interactive maps with *Folium* for spatial insights.
- Interactive Dashboard: Built a Plotly dashboard with real-time filtering and interactive charts for user exploration.
- Predictive Modeling: Developed, tuned, and validated multiple machine learning classifiers using *Scikit-learn* to predict landing outcomes**.**

Summary of Results:

- A clear learning curve was observed: landing success rates steadily increased with higher flight numbers, reflecting SpaceX's iterative improvements over time.
- Launch Site and Orbit Type emerged as the strongest predictors of landing success — particularly, LEO (Low Earth Orbit) missions achieved the highest success rates.
- A predictive machine learning model was successfully trained, with the Decision Tree Classifier achieving 87% accuracy on unseen test data.
- The final deliverable includes an interactive Plotly Dash dashboard, enabling dynamic, user-driven exploration of launch data, landing outcomes, and key influencing factors.

Github Url: https://github.com/GiovanniGarfias/IBM-Applied-Data-Science-Capstone/tree/main

Introduction

- **Project Background & Context**

The commercial space launch industry is undergoing a major transformation driven by rocket reusability — a concept pioneered by SpaceX through its Falcon 9 program.By successfully landing and reusing the rocket's first stage, SpaceX has dramatically reduced launch costs and increased the frequency of space missions.

This groundbreaking innovation has produced a rich dataset of launch records.Analyzing this data offers valuable insights into the key factors influencing landing success or failure, allowing us to apply machine learning to predict and better understand these complex outcomes.

- **Key Questions to Address**

This project focused on answering two main questions using SpaceX launch data:

1. What factors influence a successful first-stage landing?
   Is it mainly the launch site, the payload mass, the mission's orbit, or just experience gained over time?
2. Can we predict future landings?
   Based on past launches, can we train a machine learning model that accurately predicts whether a Falcon 9 will land successfully or not?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Retrieve information from Spacex Rest API
  - Web-scraping data from wikipedia
- Perform data wrangling

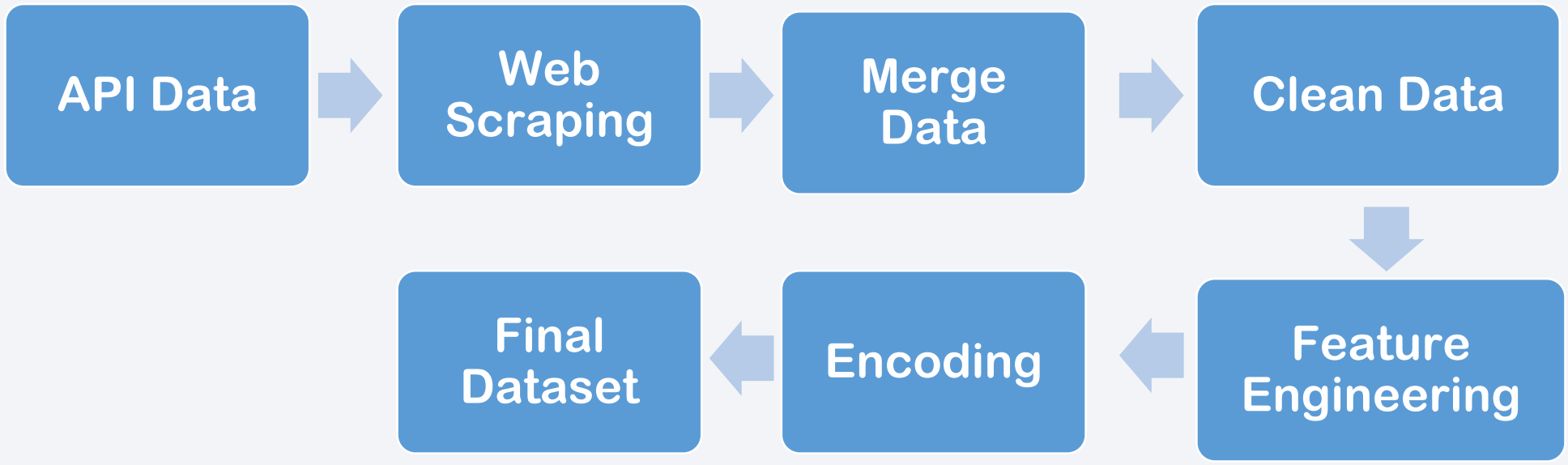
  We processed the data by imputing the missing payload values, filtering for Falcon 9 launches, and creating the binary class target variable (1=success, 0=failure)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

The launch data used in this project was collected from multiple public sources.

- First, I accessed the SpaceX REST API to retrieve structured information about Falcon 9 launches.
- Then, I used web scraping (BeautifulSoup) to gather additional details from Wikipedia, including booster versions and landing outcomes not available directly through the API.

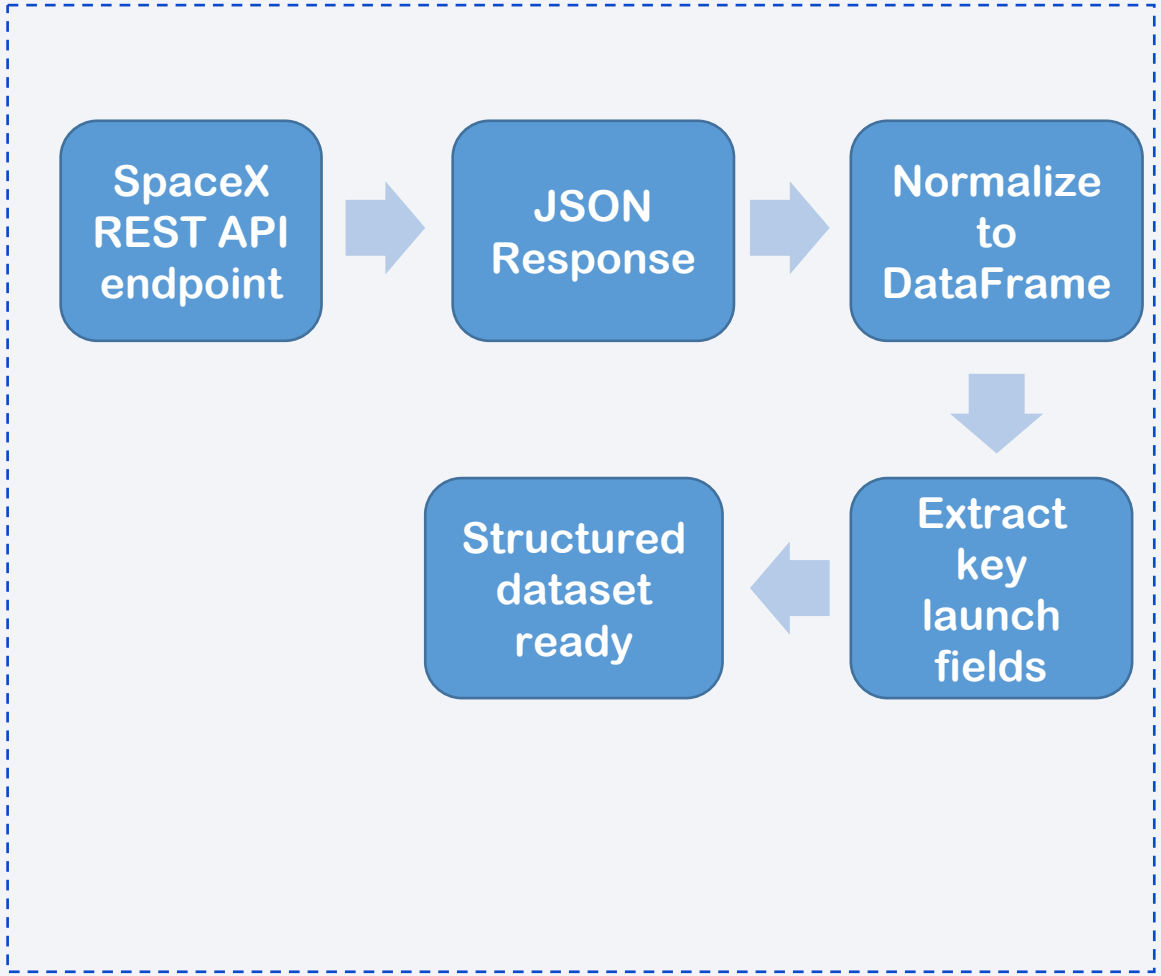
All datasets were combined and cleaned to create a single unified dataframe for analysis, visualization, and modeling.



Data Collection – SpaceX API

- Slide Text (short & clean)
- Launch data was collected directly from the SpaceX REST API, which provides JSON records of all Falcon 9 missions.
- API responses were parsed and normalized into a Pandas DataFrame using requests.get() and json_normalize().
- The API provided key information such as launch dates, payload mass, orbit type, booster version, and landing outcomes.
- These structured records formed the foundation of the full dataset, later enriched with additional scraped data.

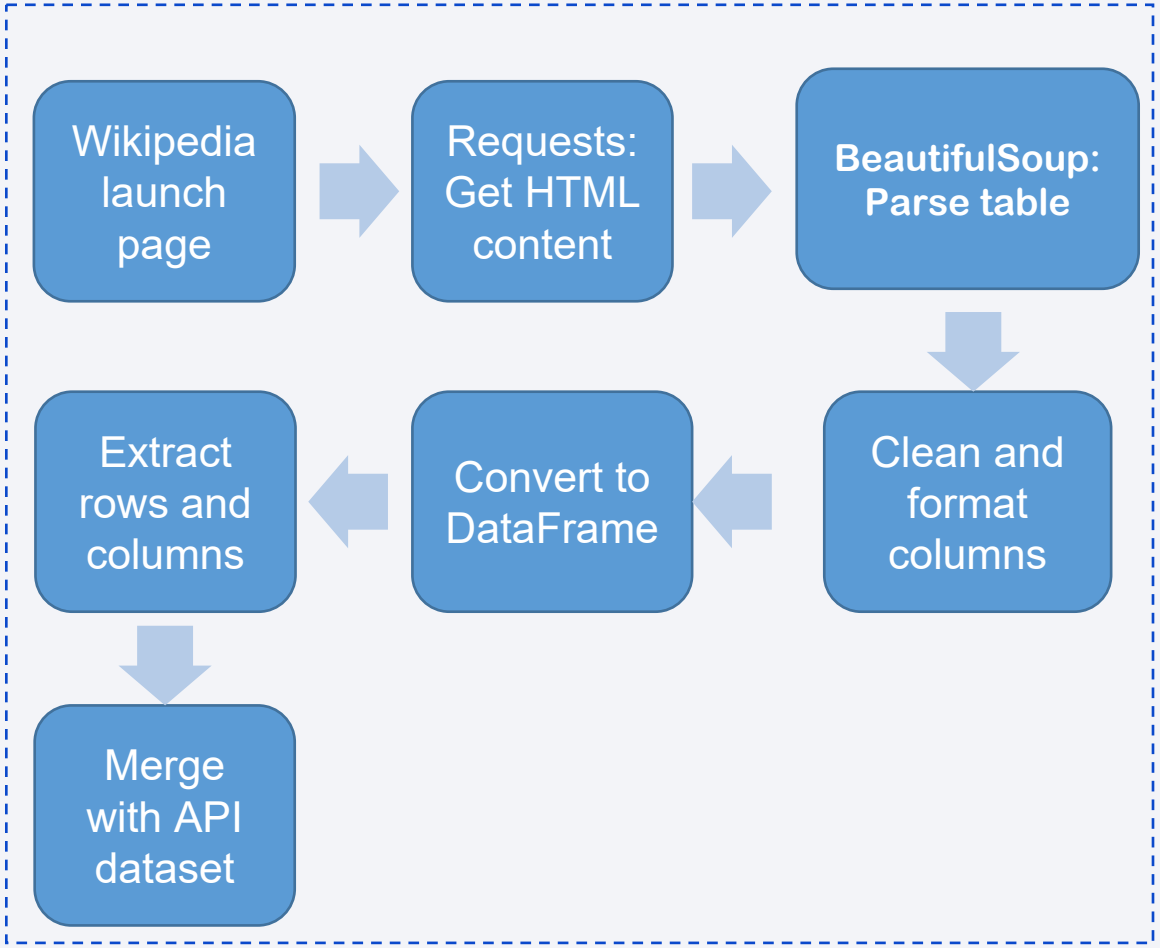
See the full notebook here.



Data Collection - Scraping

- To complement the SpaceX API data, I performed **web scraping** on the Falcon 9 & Falcon Heavy launch page from **Wikipedia**.
- Using **Requests** and **BeautifulSoup**, I extracted the launch table containing **booster versions, landing outcomes, payload details, and mission metadata**.
- The HTML table was parsed and converted into a **Pandas DataFrame**, followed by column cleanup and formatting.
- This scraped dataset provided additional fields not included in the API and was later **merged with the API dataset** to build a complete launch history.

See the complete notebook here.



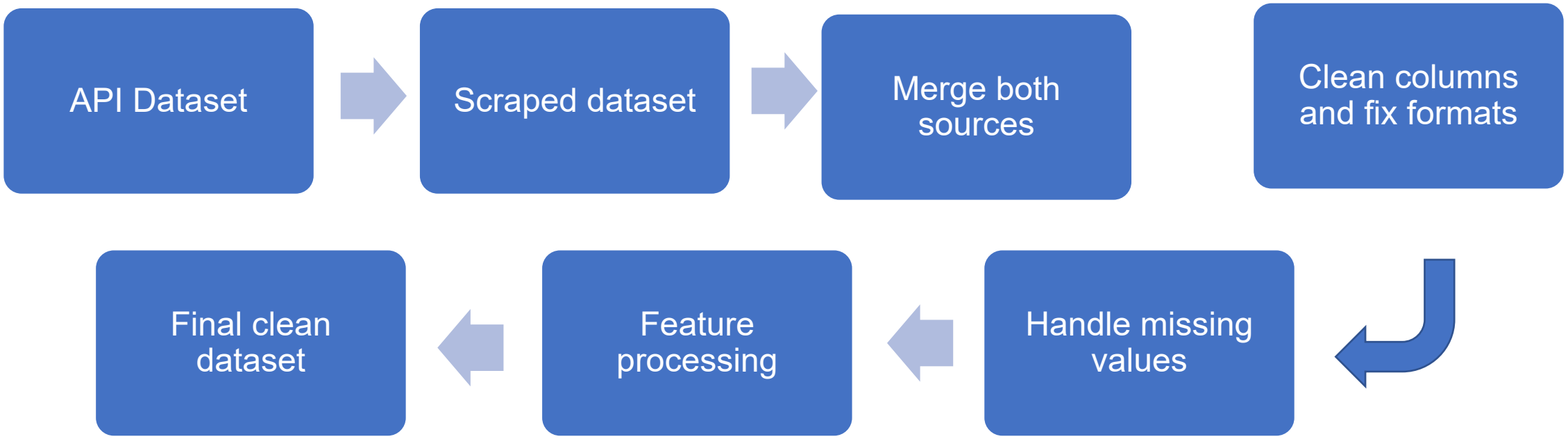
Data Wrangling

- After collecting the API and Wikipedia data, I combined both sources into a single unified dataset.
- Performed initial cleaning, including dropping irrelevant columns, standardizing text fields, and fixing inconsistent formats.
- Handled missing values by removing incomplete rows and aligning columns between the API and scraped data.
- Applied feature processing, such as converting datatypes, extracting useful fields, and preparing numerical and categorical attributes.
- Finalized a clean, structured DataFrame ready for EDA and machine learning.

See the complete notebook here.

Data Wrangling

- **Data Wrangling flowchart**



EDA with Data Visualization

- **Scatter Plot (Payload Mass vs Landing Outcome):**
  Used to check whether heavier or lighter payloads affect landing success.
- **Bar Charts (Launch Site vs Success Rate):**
  Helped compare performance across different launch sites to see which locations showed higher landing reliability.
- **Boxplots (Orbit Type vs Payload Mass):**
  Used to understand how payload mass varies by orbit, and whether orbit type influences landing outcomes.
- **Heatmap (Correlation Matrix):**
  Visualized relationships between numerical variables to identify possible predictors for the model.
- **Trend Plot (Flight Number vs Success Rate):**
  Helped show the learning curve over time, indicating improvement in landing outcomes as SpaceX gained experience.

See the complete notebook here.

EDA with SQL

- Connected to the SQLite database
- Removed any old table version using: **DROP TABLE IF EXISTS**.
- Created a **clean working table** (SPACEXTABLE) selecting only rows with a valid date.

- Retrieved all **distinct launch sites** to understand the categorical space.
- Previewed **sample records** using filters (LIKE 'CCA%') to verify consistency.
- Performed a full-table check to confirm column availability and data ranges.

- Calculated the **total payload mass** for NASA missions (SUM).
- Computed the **average payload** for a specific booster version (AVG).
- Identified the **earliest successful ground pad landing** (MIN on Date).
- See the complete notebook Github URL here:

EDA with SQL

- Counted the number of launches per **mission outcome** (GROUP BY).
- Counted **landing outcomes** within a specific timeframe (date filtering).
- Compared the number of launches **per launch site**.

- Retrieved the **booster version** associated with the **heaviest payload** (subquery using MAX).
- Extracted **monthly landing outcomes** for a specific year using substr() to manipulate dates.
- Performed targeted filters to explore “False ASDS” landings in a given year.

See the complete notebook Github URL here:

Build an Interactive Map with Folium

- **Map Objects Added**
- **Launch Site Circles:**
  Visual markers highlighting the exact geographic location of each launch site.
- **Text Labels (DivIcon):**
  Small labels added next to each site to clearly identify its name without clicking.
- **Outcome Markers (Success/Fail):**
  Individual launch points plotted on the map using:
- **Green markers** → successful landings (class = 1)
- **Red markers** → failed landings (class = 0)
  Grouped using a **MarkerCluster** to prevent overlap.
- **Distance Lines (PolyLines):**
  Straight lines drawn between each launch site and nearby features such as:
  coastline, highways, and railway tracks.
- **Distance Markers:**
  Custom text markers using **DivIcon** displaying exact distances (e.g., “5.23 KM”).

See the complete notebook Github URL here.

Build a Dashboard with Plotly Dash

What interactive plots and components were added

- Launch Site Dropdown Menu:
  Allows selecting *All Sites* or any individual launch site to dynamically filter the charts.
- Success Pie Chart:
  Automatically updates based on the dropdown selection, showing:
  • Total successes per site (when *All Sites* is selected)
  • Success vs. failure distribution for a specific site
- Payload Range Slider:
  Lets the user interactively filter launches based on payload mass (0–10,000 kg).
- Payload vs. Outcome Scatter Plot:
  Displays the relationship between payload size and landing outcome.
  Points are color-labeled by Booster Version Category to reveal additional patterns.

Why these components were added

- The dropdown enables quick comparison of launch success across sites and deeper inspection of each site's performance.
- The pie chart provides an intuitive, high-level view of success rates.
- The range slider helps explore how payload mass influences mission outcomes.
- The scatter plot reveals potential correlations between payload, success rate, and booster type, supporting deeper visual analysis.

See the Dashboard Github URL here.

Predictive Analysis (Classification)

How the classification model was built and selected

- **Feature Preparation:**
  Selected engineered numerical and one-hot encoded features related to flight history, payload, orbit, launch site, and landing conditions.
- **Data Standardization:**
  Applied *StandardScaler* to normalize feature scales and improve model convergence.
- **Train–Test Split:**
  Split the dataset into training (80%) and test (20%) sets to evaluate generalization on unseen data.
- **Model Training:**
  Trained four classification models:
- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

Hyperparameter Optimization:
Used *GridSearchCV (cv = 10)* to systematically tune hyperparameters for each model.

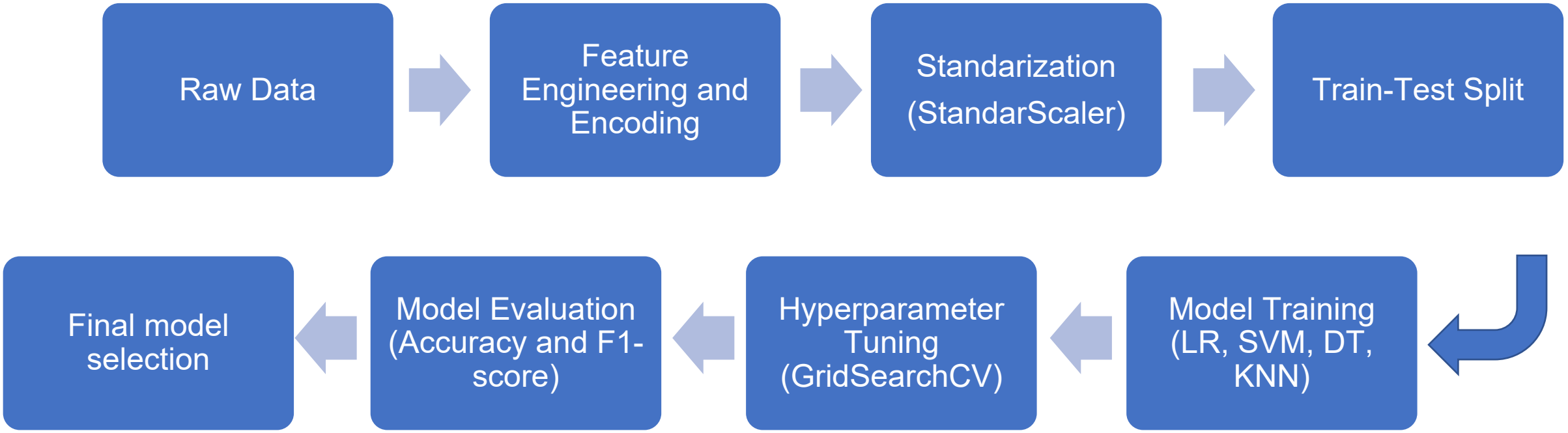
Model Evaluation:
Evaluated all models on the test set using **Accuracy** and **F1-score** to ensure balanced performance.

Model Selection:
All models achieved identical test performance. Logistic Regression was selected as the final model due to its simplicity, interpretability, and computational efficiency.

See the complete notebook Github URL here.

Predictive Analysis (Classification)

Predictive Analysis flowchart



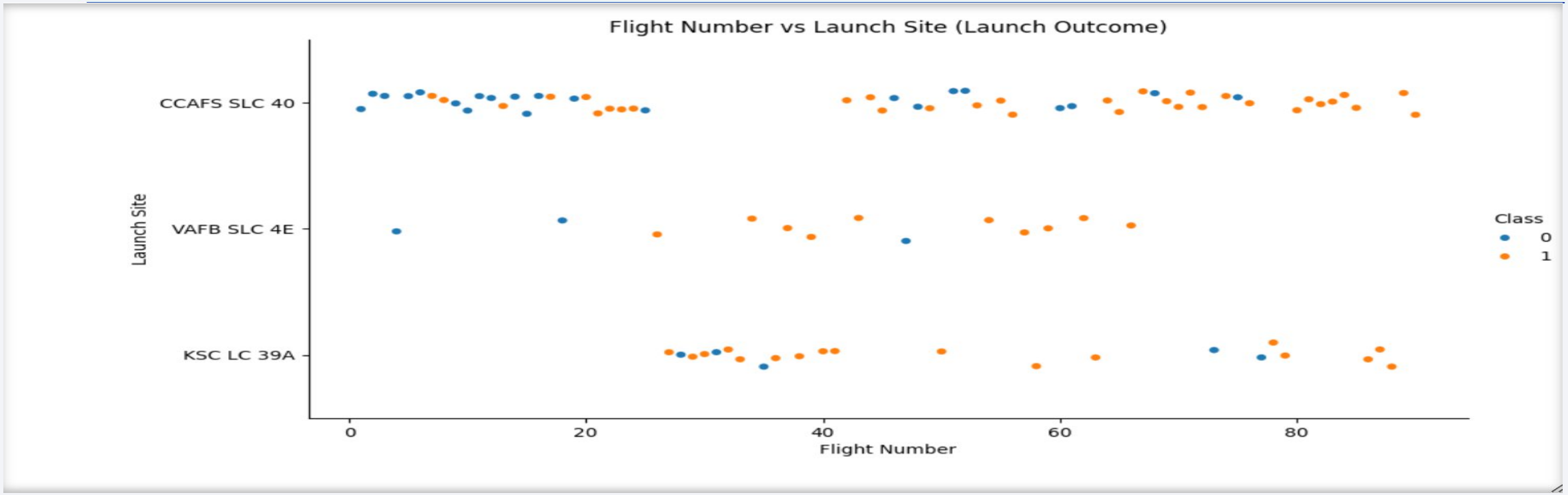
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

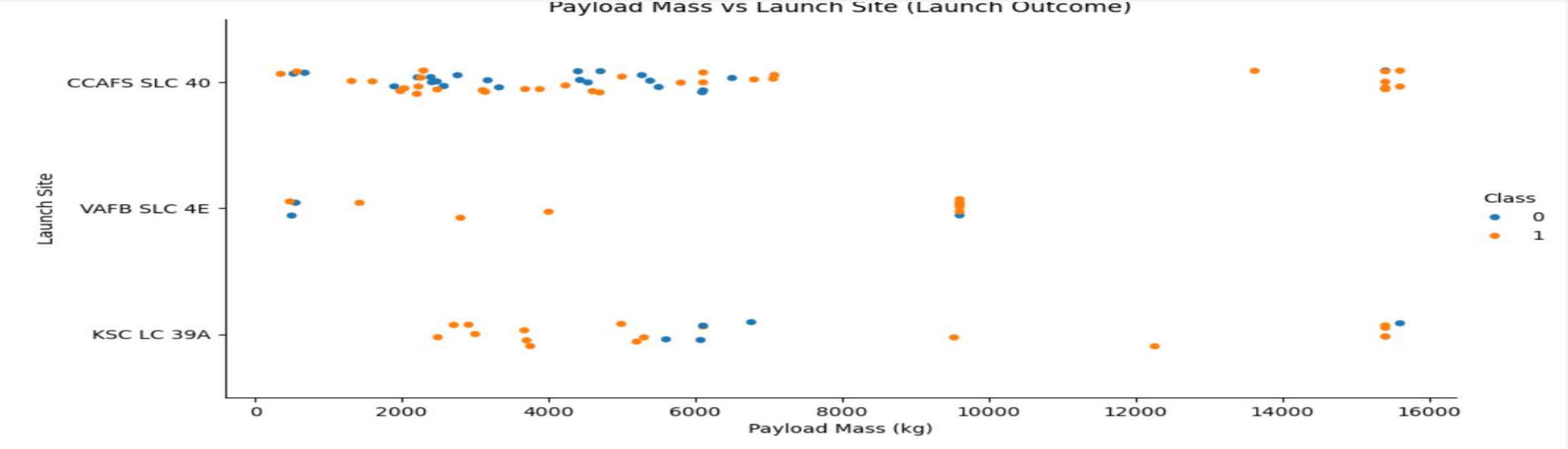
Insights drawn from EDA

Flight Number vs. Launch Site



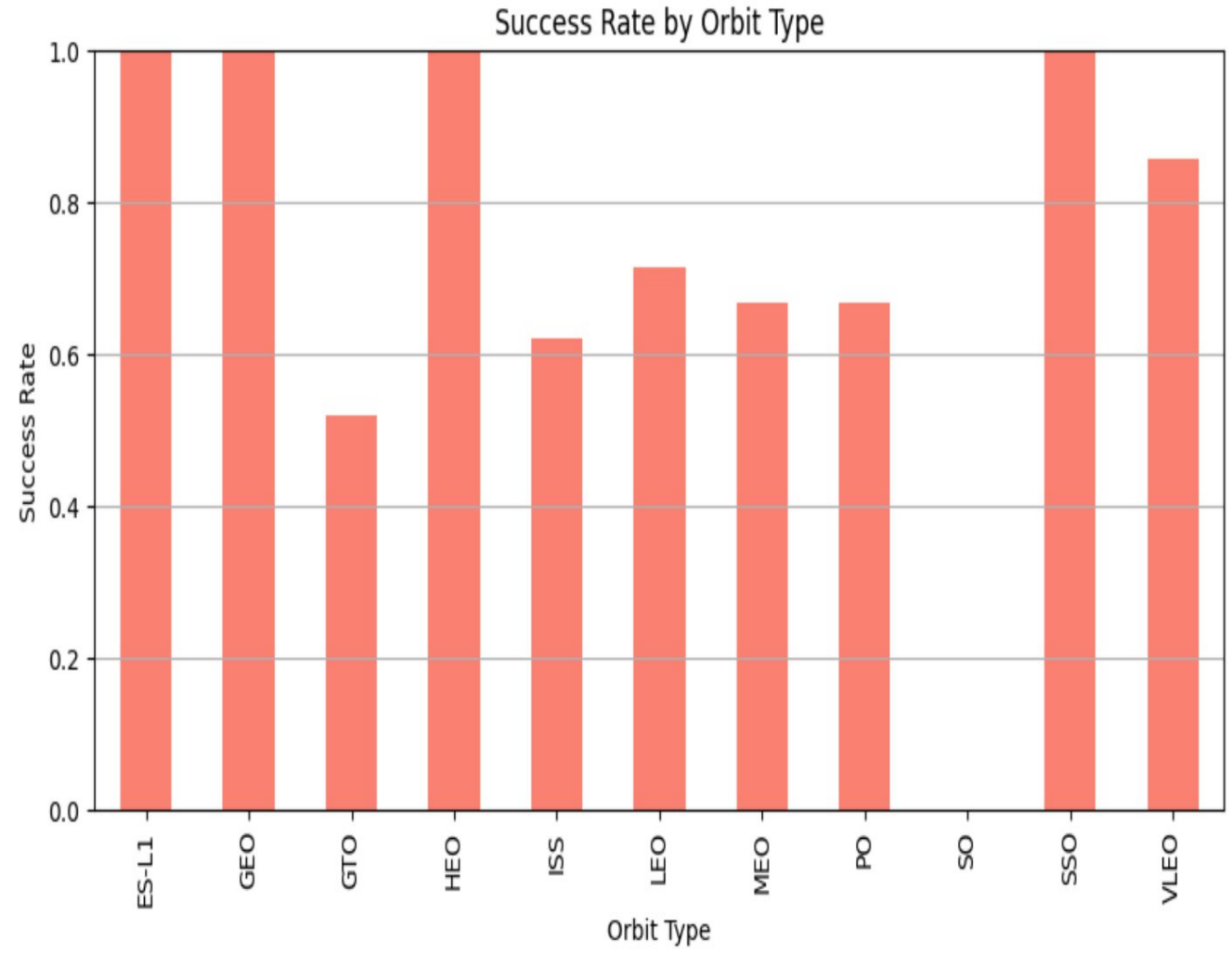
This scatter plot shows launch outcomes by flight number and launch site. CCAFS SLC-40 has the highest number of launches and demonstrates a strong improvement in landing success over time. VAFB SLC-4E has fewer launches but still shows a positive success trend, while KSC LC-39A displays a moderate number of attempts with increasing successful landings. Overall, all sites exhibit a clear learning effect as flight experience increases.

Payload vs. Launch Site



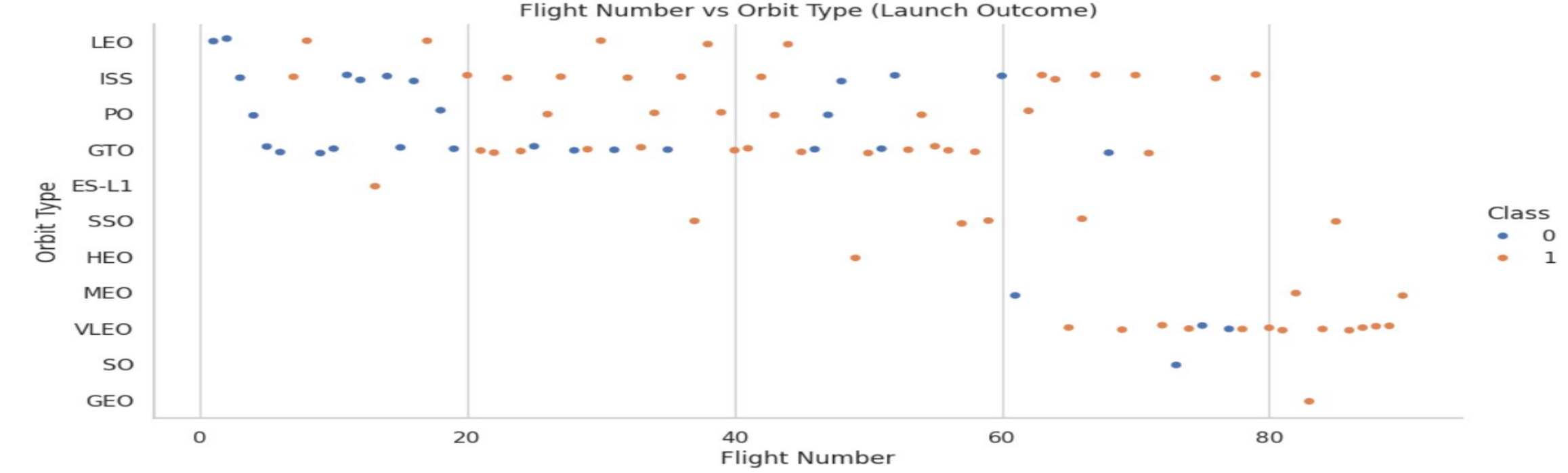
This scatter plot shows the relationship between payload mass and launch site, with launch outcomes highlighted by color. KSC LC-39A is associated with higher payload missions and demonstrates a higher proportion of successful landings. CCAFS SLC-40 covers a wide payload range and shows improving success rates, while VAFB SLC-4E has fewer launches with more constrained payload values. These patterns suggest that both payload mass and launch site are relevant factors influencing landing success.

Success Rate vs. Orbit Type



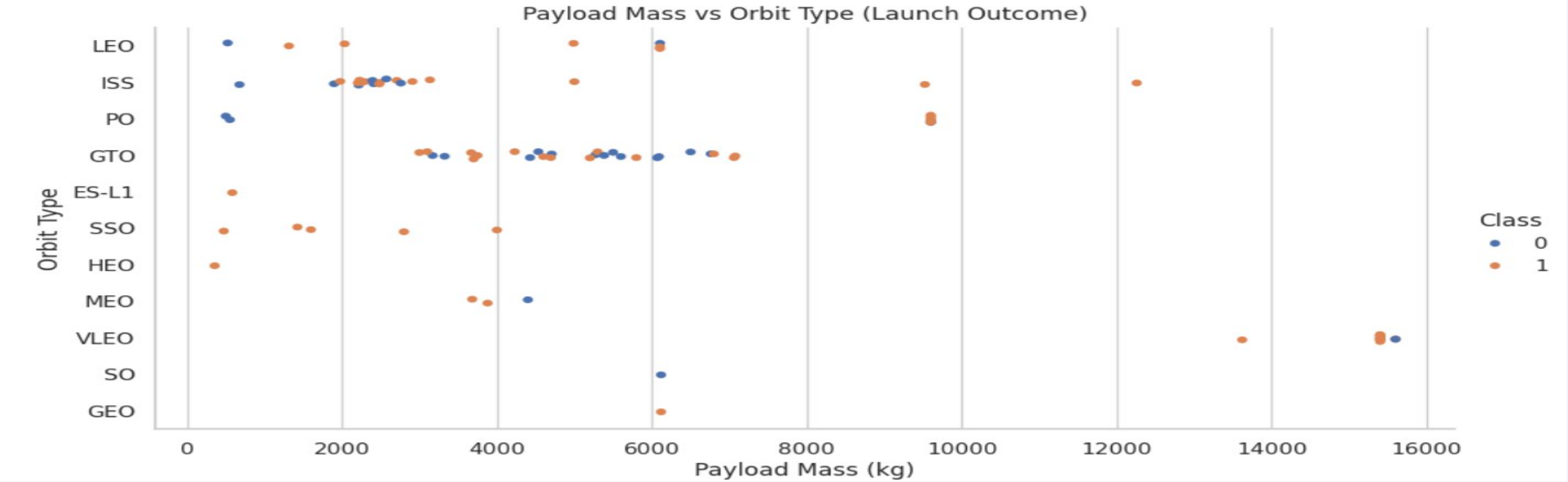
This bar chart shows the landing success rate for each orbit type. Low Earth Orbit (LEO) and ISS missions achieve noticeably higher success rates, while more demanding orbits such as GTO and HEO show lower success rates. This suggests that orbit type is a significant factor influencing landing performance and should be considered an important feature in predictive modeling.

Flight Number vs. Orbit Type



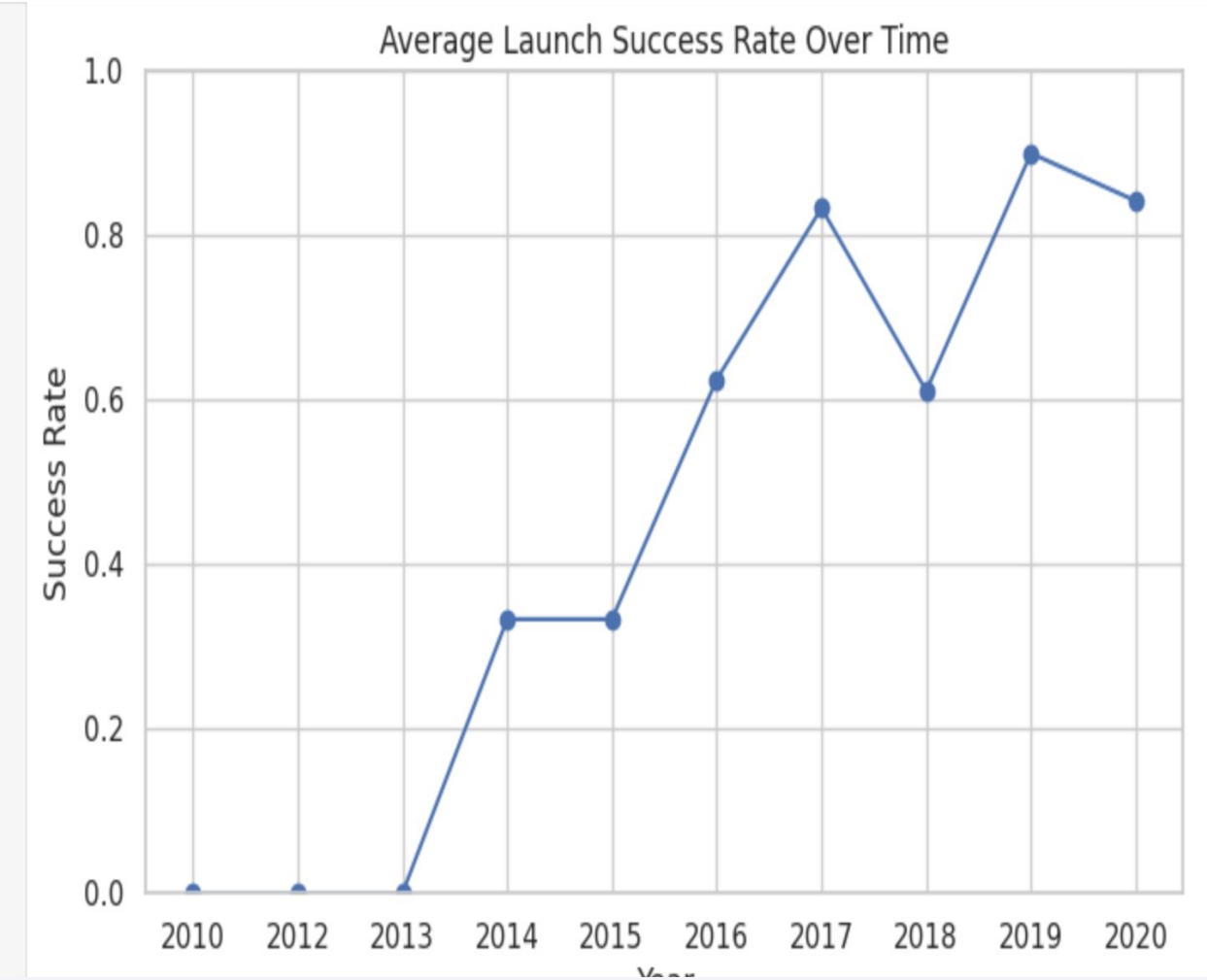
This scatter plot illustrates the relationship between flight experience and orbit type, with landing outcomes shown by color. Early flights across most orbit types exhibit higher failure rates, while later flights show a clear increase in successful landings. Simpler missions such as LEO and ISS reach higher success levels earlier, whereas more complex orbits improve more gradually. This highlights the combined effect of operational experience and orbit complexity on landing success.

Payload vs. Orbit Type



This scatter plot shows the relationship between payload mass and orbit type, with landing outcomes highlighted by color. Missions to simpler orbits such as LEO and ISS achieve high success rates across a broad payload range, while more demanding orbits like GTO exhibit lower success, particularly at higher payload masses. These patterns indicate that both payload mass and orbit type jointly influence landing performance.

Launch Success Yearly Trend



This line chart shows the average annual launch success rate. Early years exhibit lower success rates, while later years demonstrate a clear and consistent improvement. This trend highlights SpaceX's operational learning curve and increasing reusability reliability over time.

All Launch Site Names

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

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

This query retrieves the distinct launch sites from the SpaceX missions dataset. The result confirms that SpaceX operates launches from four different launch locations, which are later analyzed to evaluate their impact on mission success rates.

Launch Site Names Begin with 'CCA'

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

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

SUM(PAYLOAD_MASS__KG_)
45596

The query computes the total payload mass delivered by SpaceX during NASA CRS missions, highlighting the scale of cargo transported for NASA.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

AVG(PAYLOAD_MASS__KG_)
2928.4

The query calculates the average payload mass carried by the Falcon 9 v1.1 booster, summarizing its typical launch capacity.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

MIN(Date)
2015-12-22

This query identifies the earliest date on which SpaceX achieved a successful ground pad landing. By filtering records with a successful ground pad outcome and applying the MIN function to the launch date, the result highlights a key milestone in SpaceX's reusability program.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_Version
    FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (drone ship)'
    AND PAYLOAD_MASS__KG_ > 4000
    AND PAYLOAD_MASS__KG_ < 6000;
```

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

This query lists the booster versions that achieved a successful landing on a drone ship while carrying a payload mass between 4000 kg and 6000 kg. By combining landing outcome and payload constraints, the query identifies booster configurations that performed successfully under medium-to-heavy payload conditions.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome,  COUNT(*) as "Total" FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

This query groups all SpaceX missions by their mission outcome and counts the total number of missions in each category. The result provides an overall comparison between successful and failed missions, offering a high-level view of SpaceX mission reliability.

Boosters Carried Maximum Payload

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

```
%%sql SELECT Booster_Version
    FROM SPACEXTABLE
    WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE );
```

This query identifies the booster versions that carried the maximum payload mass in SpaceX missions. A subquery is used to first determine the highest payload mass in the dataset, and the outer query then retrieves all booster versions associated with that payload value.

2015 Launch Records

```
%%sql
SELECT
    substr(Date, 6, 2) AS month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 1, 4) = '2015'
  AND Landing_Outcome = 'Failure (drone ship)';
```

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

This query filters the dataset to select missions from the year 2015 with failed drone ship landings. By extracting the year and month from the launch date and filtering by a specific landing outcome, the result shows the month, booster version, and launch site for unsuccessful drone ship landings during 2015.

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

Landing_Outcome	Total
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

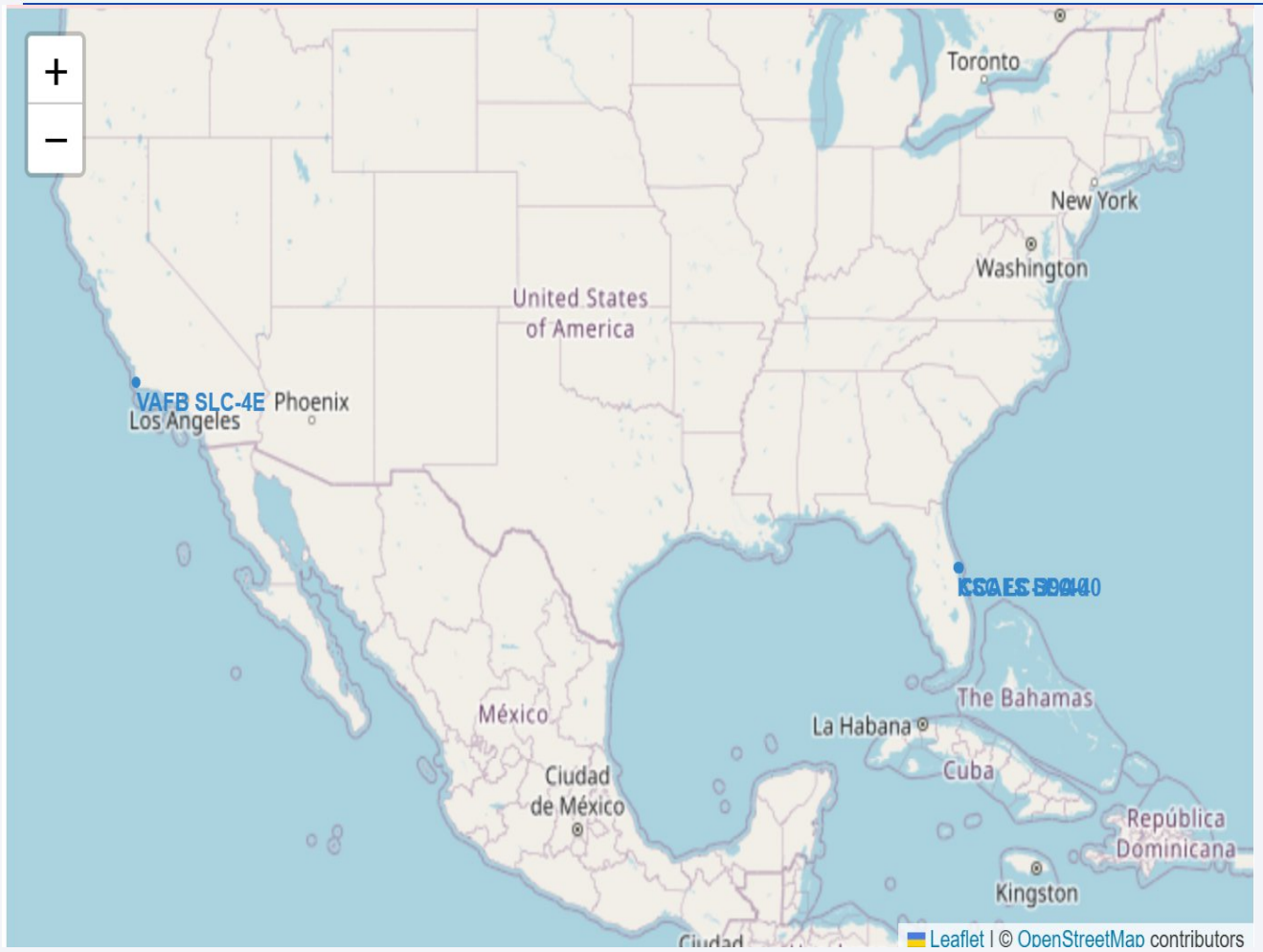
```
%%sql SELECT "Landing_Outcome",
    COUNT(*) as "Total"
    FROM SPACEXTABLE
    WHERE Date
    BETWEEN '2010-06-04'
    AND '2017-03-20'
    GROUP BY "Landing_Outcome"
    ORDER BY "Total" DESC;
```

This query ranks landing outcomes by their total count for SpaceX launches occurring between June 4, 2010 and March 20, 2017. By grouping missions by landing outcome and ordering them in descending order, the result highlights the most frequent landing results during the early reusability era.

Section 3

Launch Sites Proximities Analysis

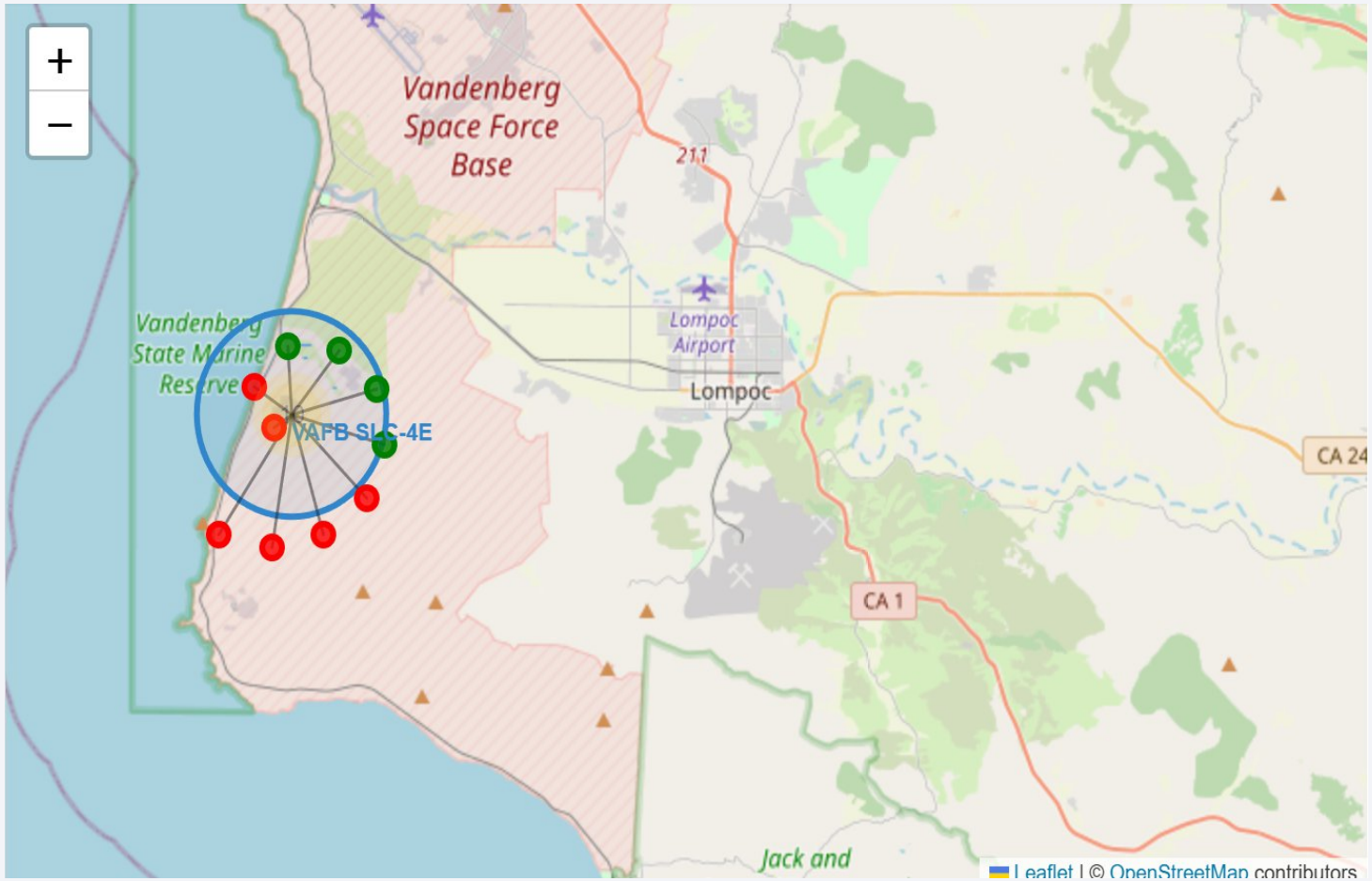
SpaceX Launch Sites – Geographic Distribution



This map shows the geographic locations of SpaceX launch sites across the United States. Each marker represents a distinct launch site, allowing a clear comparison of their spatial distribution.

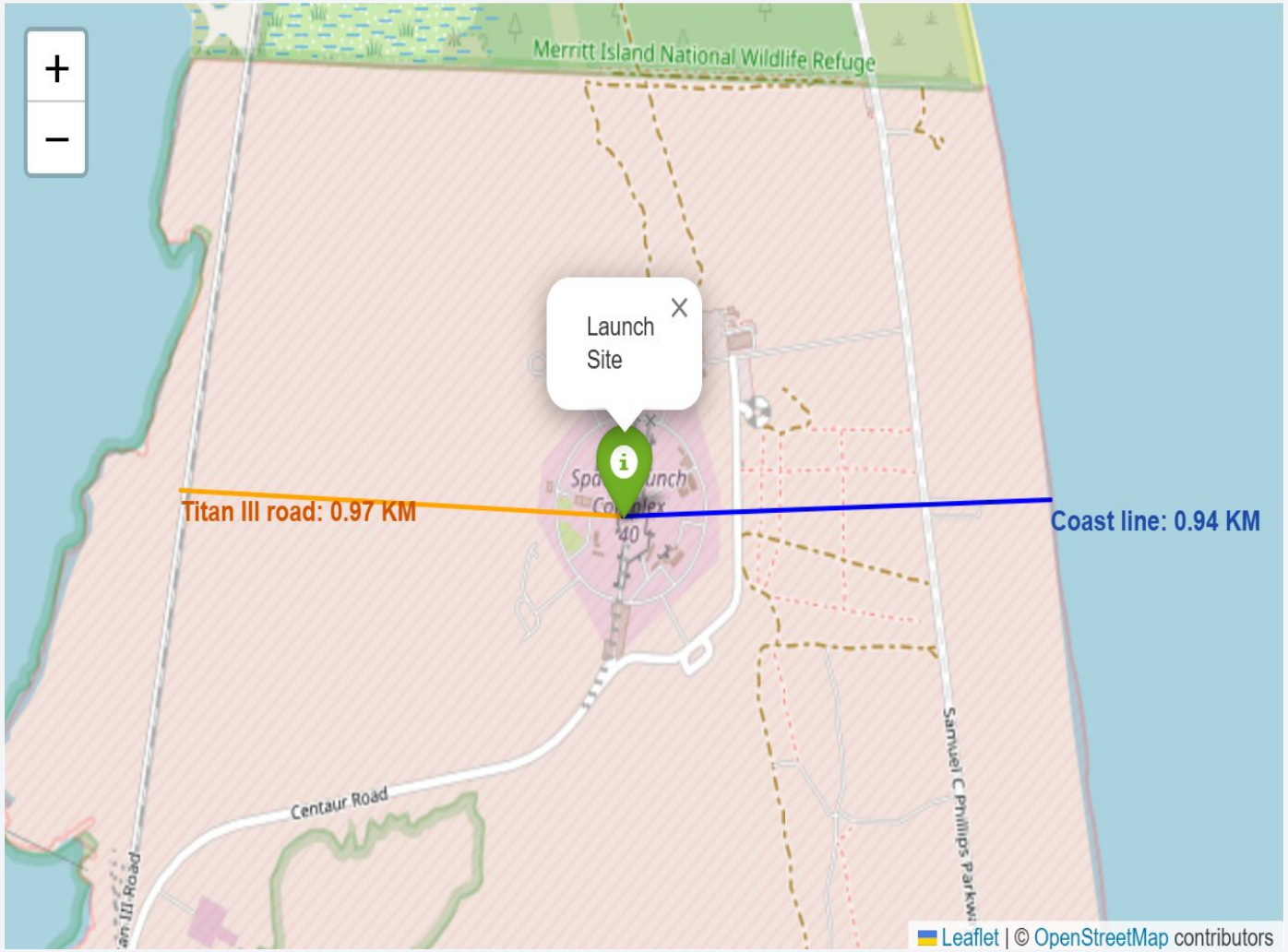
The visualization highlights that all launch sites are located near coastal areas, which is strategically important for safety and launch logistics. In addition, the sites are spread across different latitudes, supporting a variety of mission profiles and orbital inclinations.*

Color-Coded Launch Outcomes by Site

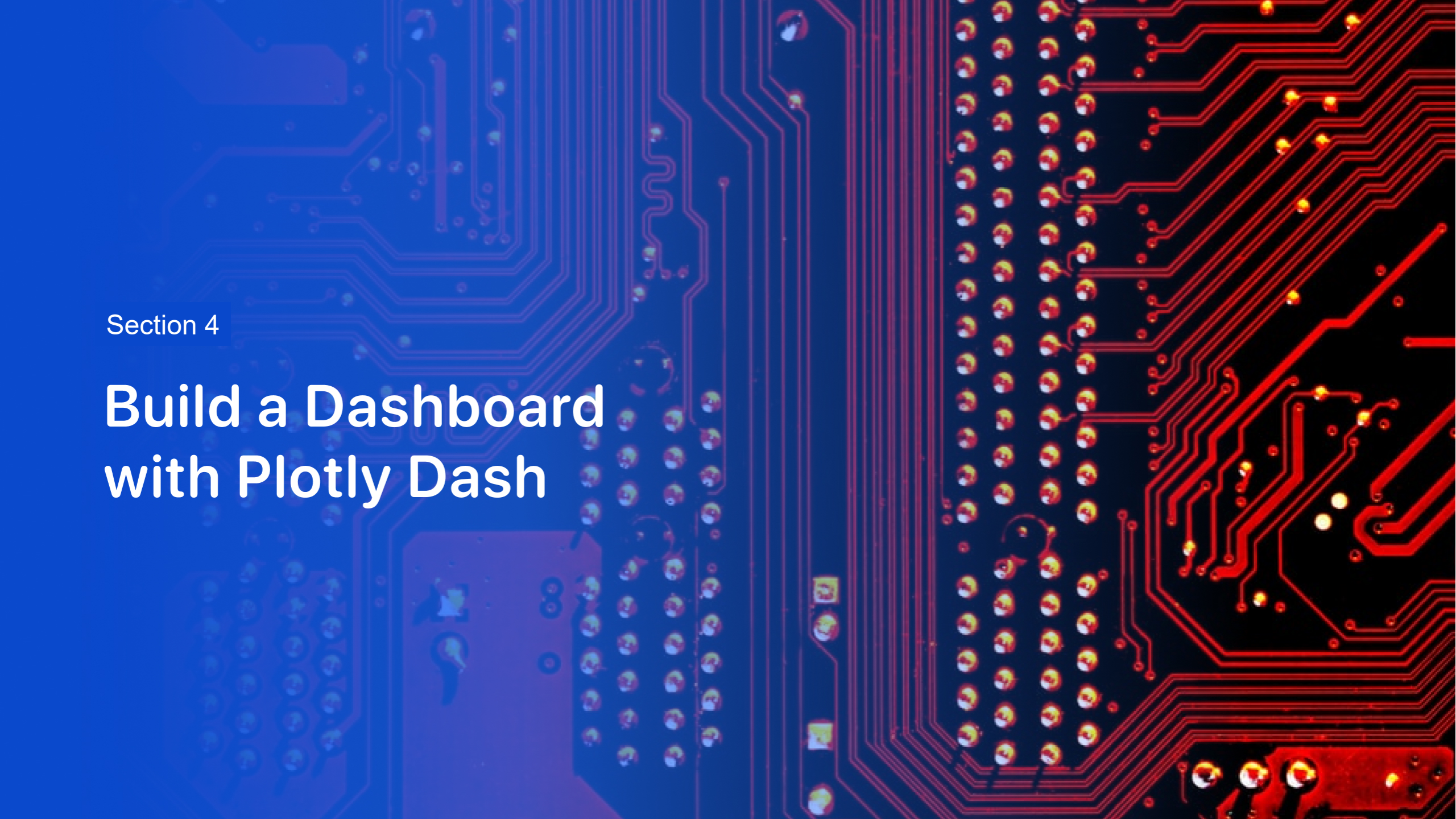


This interactive map displays SpaceX launch outcomes across all launch sites using color-coded markers. Each marker represents an individual launch attempt, where green markers indicate successful landings and red markers indicate failed landings. Marker clustering is used to group launches occurring at the same site, making it easier to visualize the overall proportion of successes and failures at each location. The map allows direct visual comparison of launch performance across different sites.

Launch Site Proximity to Coastline and Infrastructure



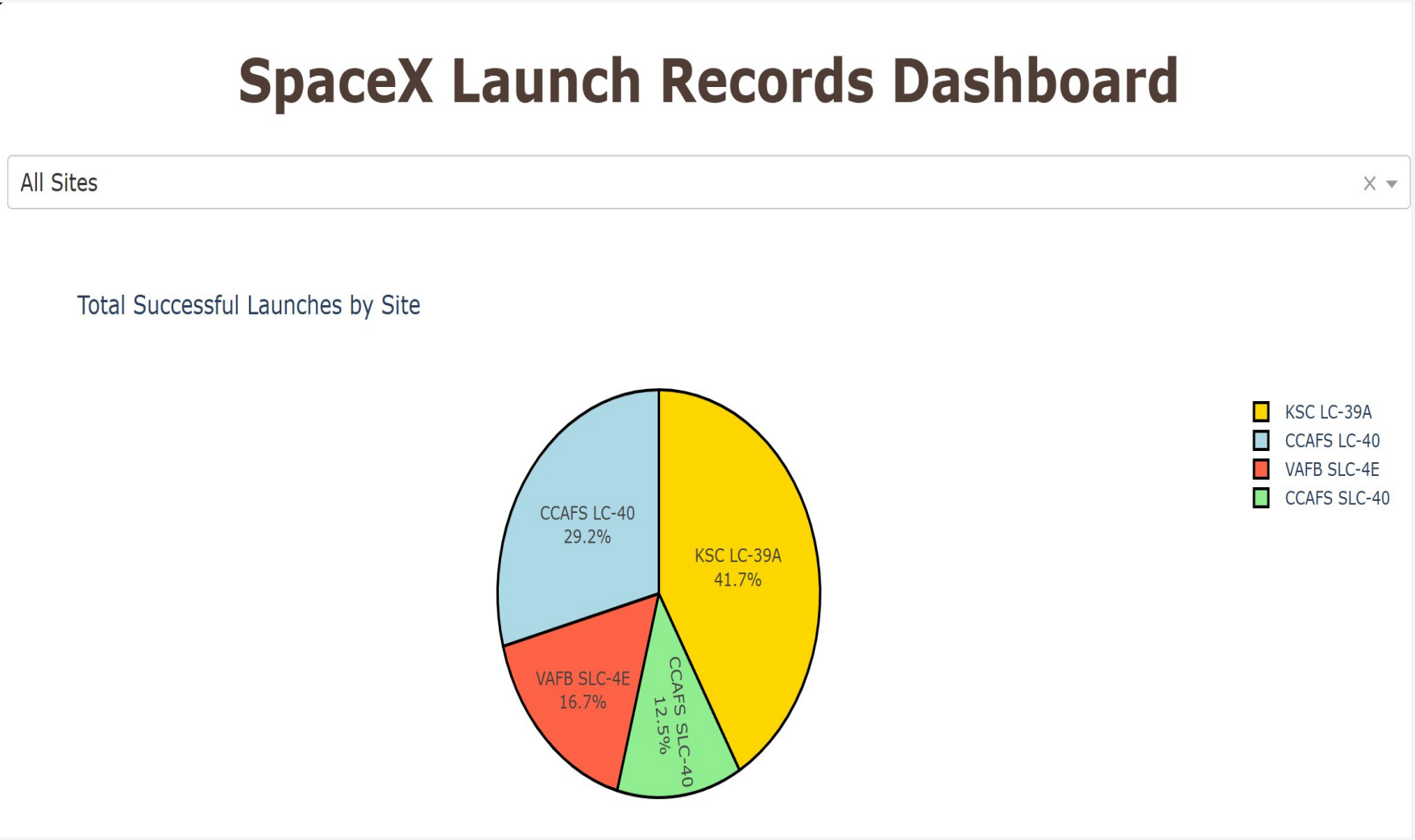
This map shows the distance from the launch site to two key reference points: the coastline and Titan III Road. Orange and blue polylines connect the launch pad to each location, while labeled markers display the distances in kilometers. This highlights how close the launch infrastructure is to both coastal safety zones and nearby ground transportation routes.



Section 4

Build a Dashboard with Plotly Dash

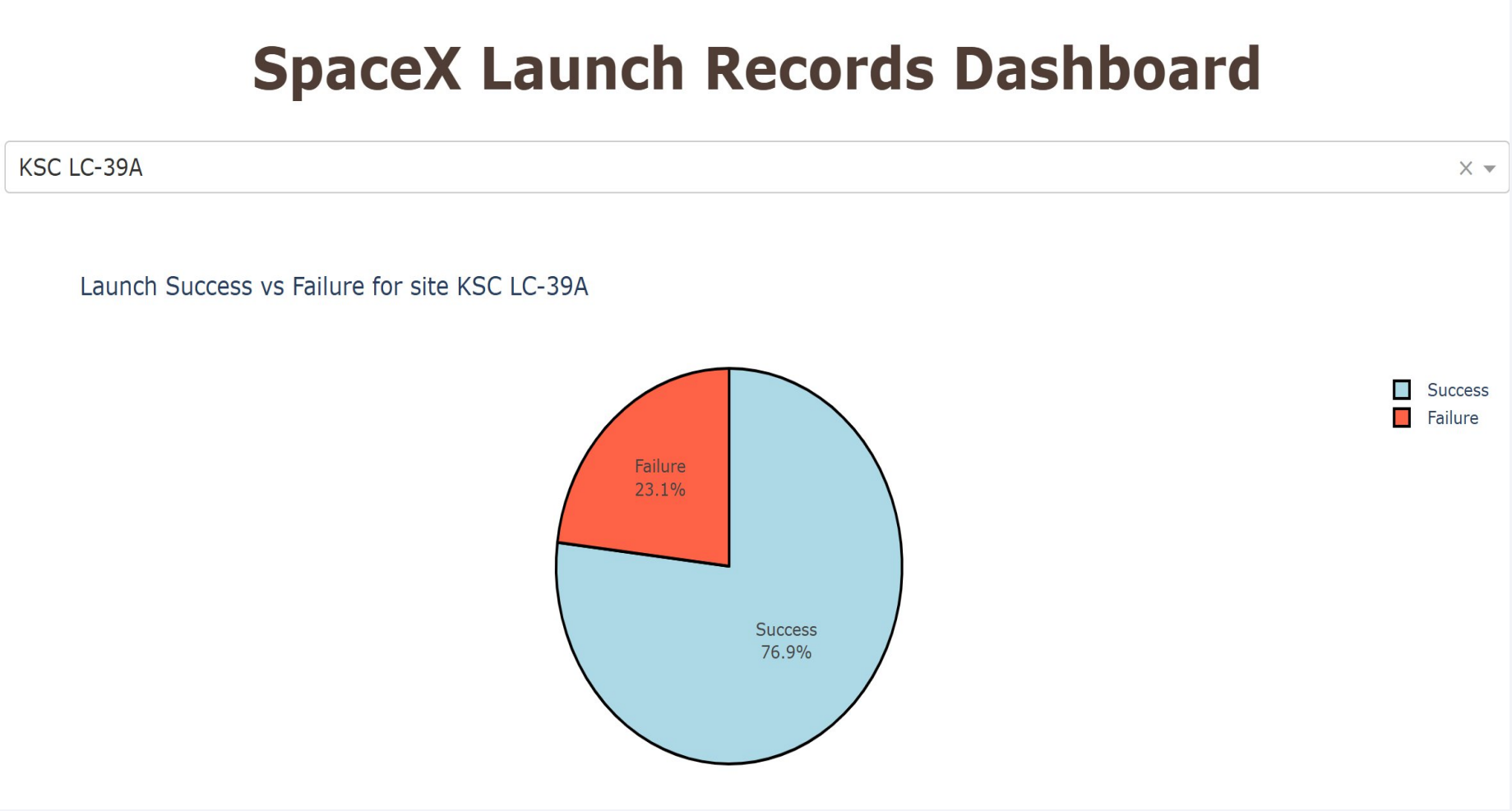
Successful Launch Share Across SpaceX Sites



This dashboard shows the distribution of successful launches across all SpaceX launch sites. The dropdown at the top is set to "All Sites", and the pie chart displays the proportion of successful missions contributed by each location.

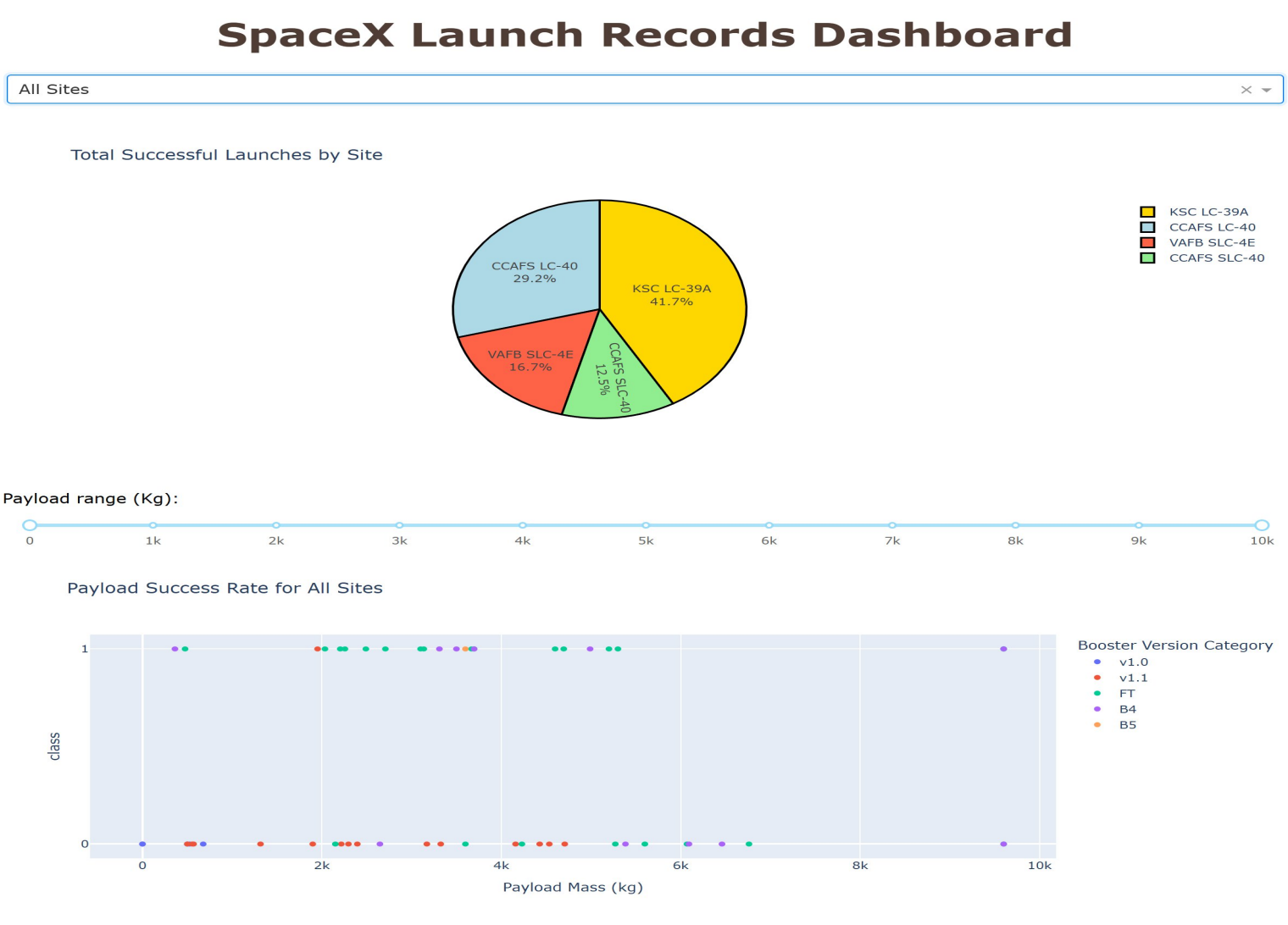
KSC LC-39A accounts for the largest share of successful launches (about 42%), followed by CCAFS LC-40 (~29%), VAFB SLC-4E (~17%), and CCAFS SLC-40 (~12%). This visualization highlights that launch activity and success are not evenly distributed across sites, with KSC LC-39A playing a particularly important role in successful missions

Launch site with the highest success ratio



This pie chart shows the success rate for launches at KSC LC-39A, the site with the highest total number of successful launches overall. When focusing only on this site, approximately 77% of launches were successful, while 23% resulted in failure, indicating a high reliability level for KSC LC-39A.

Key insights from the SpaceX launch data dashboard



This dashboard summarizes key insights obtained from the interactive SpaceX launch data dashboard. By combining launch site selection, payload range filtering, and visual comparisons, the dashboard reveals clear patterns in launch success behavior.

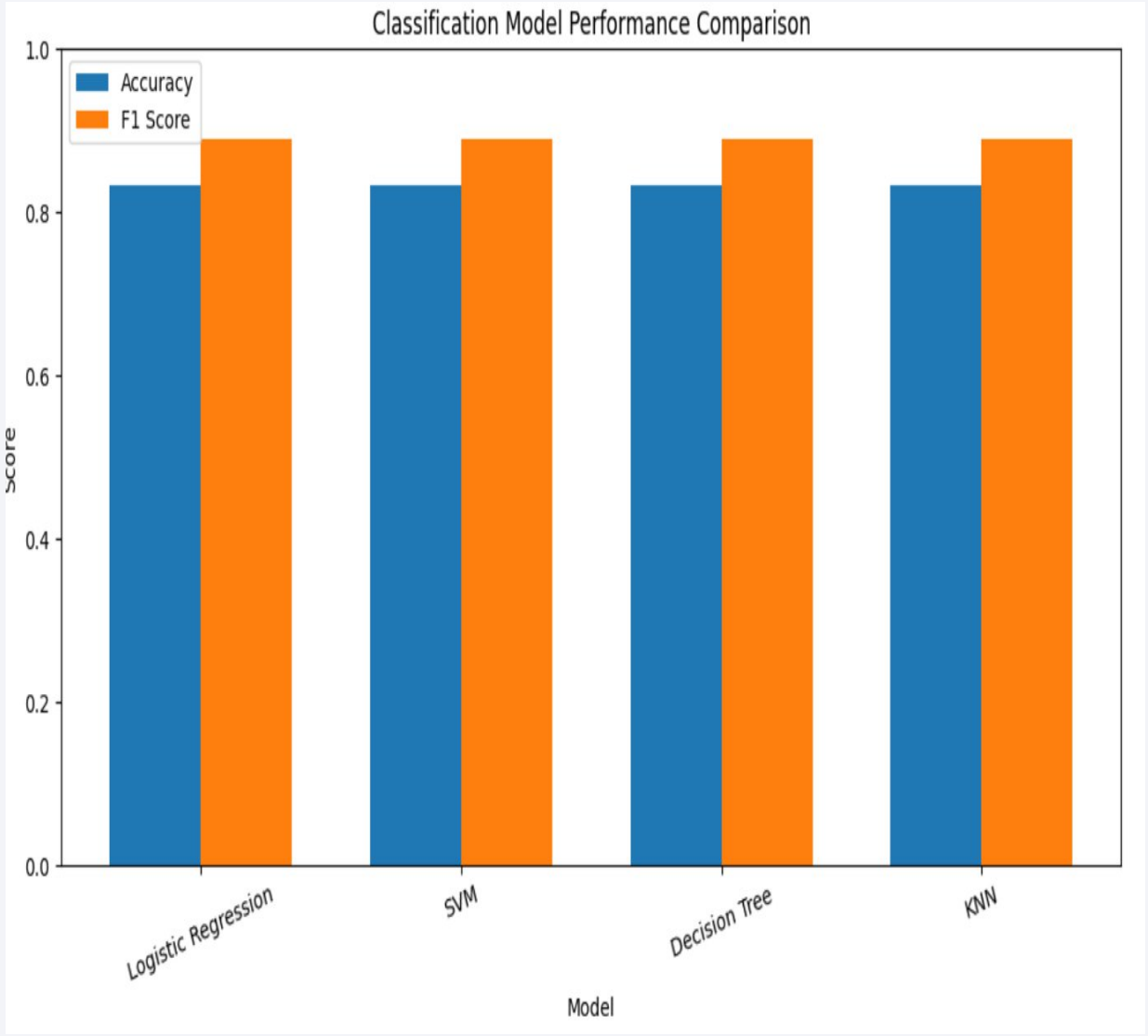
The analysis shows that launch success varies across sites, with KSC LC-39A contributing the highest number of successful launches. Additionally, payload mass plays an important role in mission outcomes, with higher success rates observed in specific payload ranges. The visualization also highlights improved reliability when using newer booster versions, such as FT, B4, and B5.

Overall, the dashboard enables interactive exploration of how launch site, payload mass, and booster technology jointly influence launch success

Section 5

Predictive Analysis (Classification)

Classification Accuracy

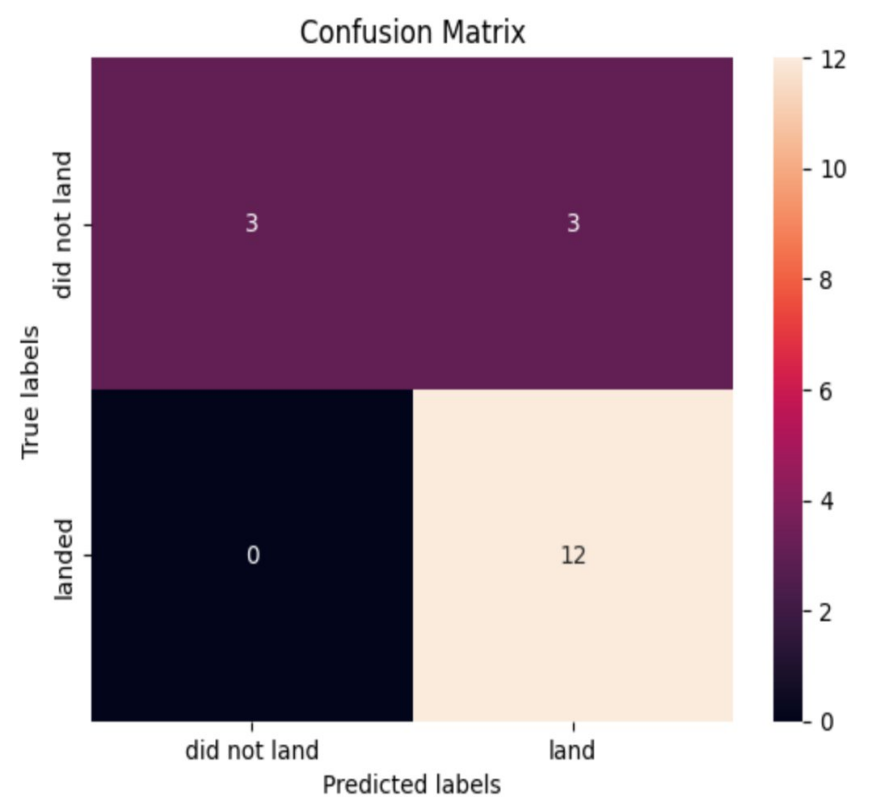


This chart compares the performance of four classification models—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN)—using both accuracy and F1-score on the test dataset.

All models achieved a similar performance, with an accuracy of approximately 83% and an F1-score close to 0.89. This indicates that the engineered features provide strong predictive signals and that the landing success prediction task is not highly sensitive to the choice of classification algorithm.

Since no single model clearly outperformed the others, the final model selection was based on factors such as simplicity, interpretability, and robustness rather than marginal differences in performance.

Confusion Matrix



The confusion matrix illustrates the performance of the Logistic Regression model on the test dataset by comparing predicted outcomes with the true landing results.
The model correctly predicted 12 successful landings (true positives) and 3 failed landings (true negatives). There were 3 false positives, where the model predicted a successful landing that did not occur, and no false negatives, meaning the model never misclassified a successful landing as a failure.
This result shows that the model is particularly effective at identifying successful landings, which is reflected in the high accuracy and F1-score. The confusion matrix confirms that the Logistic Regression model provides reliable and balanced predictions for this classification task

Logistic Regression :
Accuracy : 0.8333
F1 Score : 0.8889

Conclusions

- ***Operational Experience and Technology Improvements***
  Launch success rates increased over time as SpaceX gained operational experience and introduced improved Falcon 9 booster versions, particularly FT, B4, and B5.
- ***Impact of Launch Site and Payload Mass***
  Mission outcomes are influenced by both launch site and payload mass. KSC LC-39A accounted for the highest number of successful launches, and specific payload ranges consistently showed higher success rates, supporting their relevance as predictive features.
- ***Insights from Interactive and Geospatial Analysis***
  Interactive dashboards and geospatial visualizations enabled deeper exploration of how launch site location, payload mass, and booster configuration interact, revealing important operational patterns.
- ***Robust Machine Learning Results***
  All evaluated classification models achieved stable and consistent performance. The similarity in accuracy and F1-scores highlights the effectiveness of the feature engineering process. Logistic Regression was selected as the final model due to its simplicity, interpretability, and reliable predictive behavior.

Ackownledgements

I would like to thank Coursera and IBM for providing the tools, learning materials, and guidance that made this project possible. Special thanks to the instructors and content creators of the IBM Data Science Professional Certificate for their clear explanations and practical approach to data science and machine learning.

Giovanni Garfias
IBM Applied Data Science Capstone

Thank you!