IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohan Paramasivam
2025-10-24

Outline

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

https://github.com/mohanpsivam/spacex-capstone

Executive Summary

Summary of Methodologies

- Collected SpaceX Falcon 9 launch data using REST API, web scraping, and CSV datasets.
- Performed data wrangling to clean, merge, and structure the data for analysis.
- Conducted Exploratory Data Analysis (EDA) using SQL and visualization libraries.
- Built interactive maps with **Folium** to explore spatial launch patterns.
- Developed a **Plotly Dash dashboard** for dynamic data exploration.
- Applied **machine learning classification models** to predict launch success based on payload, site, and booster parameters.

Summary of Results

- Discovered that **payload mass and launch site** are strong indicators of mission success.
- Launches from **KSC LC-39A** showed the highest success rate.
- **Reusable boosters** significantly improved reliability and reduced cost.
- Interactive dashboard and maps effectively visualized insights for decision-making.
- Achieved a predictive model accuracy of around **83%**, confirming consistent classification performance.

Introduction

Project Background and Context

- SpaceX designs, manufactures, and launches advanced rockets and spacecraft.
- The company's long-term goal is to make **reusable rocket technology** reliable and cost-efficient.
- This project analyzes historical SpaceX Falcon 9 launch data to understand the factors influencing launch outcomes.
- The study uses real-world data collected from APIs, web sources, and CSV files to simulate a complete data science workflow — from collection to prediction.

Problems Want to Find Answers:

- Which factors most strongly affect the **success of a Falcon 9 launch**?
- Does **payload mass** or **launch site** influence mission success?
- How do **booster versions** and **flight reuse** contribute to performance improvement?
- Can we **predict future launch outcomes** using machine learning models?

Section 1

Methodology

Methodology

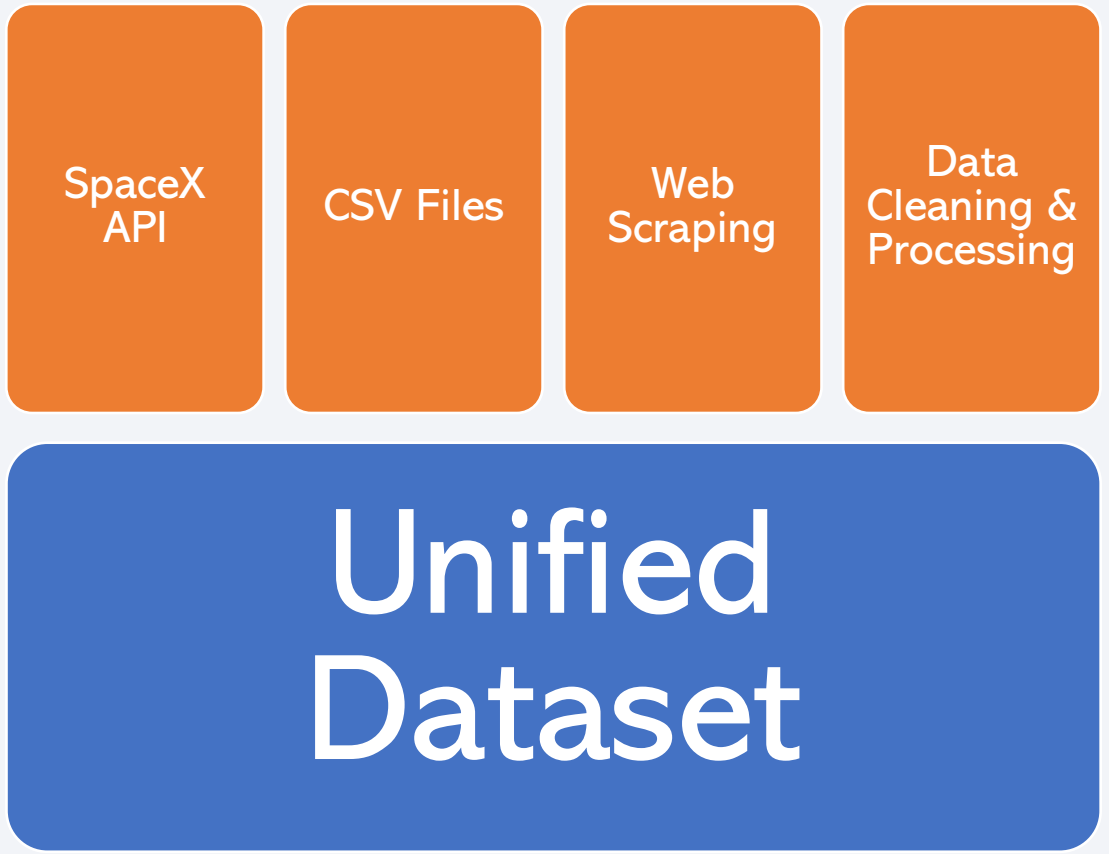
Executive Summary

- Data collection methodology:
- Perform data wrangling
- 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

Github Link

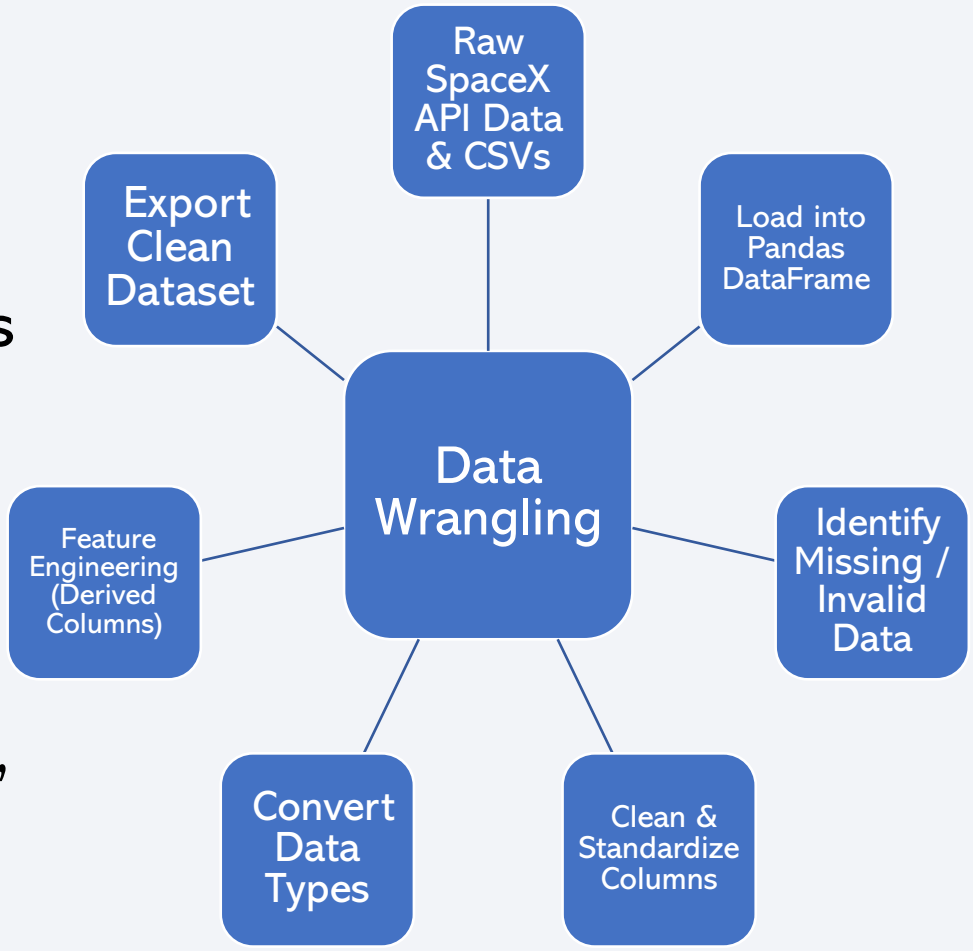
- Retrieved **launch data from SpaceX API** (missions, dates, sites, payloads).
- Used **provided CSV datasets** for historical launch information.
- Performed **web scraping** to get extra launch site details.
- Combined all sources into a **central dataset** for analysis.



Data Wrangling

[Github Link]

- Loaded raw data from SpaceX API and CSV files into Pandas DataFrames.
- Checked for missing values and data inconsistencies.
- Cleaned column names, standardized formats (e.g., launch site, booster version).
- Converted data types — especially date and numeric fields.
- Removed duplicates and irrelevant columns.
- Created new derived fields (e.g., success flag, payload mass range).
- Saved the cleaned dataset for visualization and modeling.



EDA with Data Visualization

Github Link

- Used **Matplotlib** and **Seaborn** to explore trends and patterns.
- Plotted **bar charts** to compare the number of launches by site and booster version.
- Created **scatter plots** to study the relationship between payload mass and launch success rate.
- Drew **pie charts** to show the proportion of successful vs. failed launches.
- Used **box plots** to visualize payload distribution and identify outliers.
- Visualized **correlation heatmap** to understand relationships between numerical features.
- These visualizations helped identify key factors influencing SpaceX launch outcomes.

Why These Charts

- **Bar charts:** for categorical comparisons.
- **Scatter plots:** to observe trends between continuous variables.
- **Pie charts:** for quick success rate overview.
- **Heatmap:** to highlight variable correlations at a glance.

EDA with SQL

Github Link

- **Summary of SQL Queries**
- **Selected key columns** from the launch dataset to focus on mission success, payload, and launch site.
- **Counted total launches per site** to find the busiest launch locations.
- **Calculated success rates** for each launch site using conditional aggregation.
- **Identified most frequently used booster versions** and their success performance.
- **Computed average payload mass** grouped by booster version and launch site.
- **Filtered records** to display only successful launches for trend comparison.
- **Ordered and limited results** to highlight top-performing boosters.
- **Joined tables** (when required) to combine launch details with payload information.

Build an Interactive Map with Folium

Github Link

Summary of Map Objects

- **Markers:** Plotted for each SpaceX launch site to pinpoint exact geographic locations.
- **Circle Markers:** Added to represent proximity and landing zones visually.
- **Popups:** Included for each marker to display site name, coordinates, and success rate.
- **Polylines:** Drew lines between launch sites and landing points to illustrate trajectory paths.
- **Tile Layers:** Used different map styles (OpenStreetMap, Stamen Terrain) for better visual clarity

Purpose of These Objects

- To **visualize spatial relationships** between launch sites and landing outcomes.
- To **highlight reusable booster locations** and success distribution geographically.
- To make the **EDA findings more interactive and intuitive** for peer reviewers.

Build a Dashboard with Plotly Dash

Github

Summary of Plots and Interactions

- **Dropdown Menu:** Added to select different launch sites dynamically.
- **Pie Chart:** Displays the success vs. failure rates for the chosen launch site.
- **Scatter Plot:** Shows the relationship between payload mass and launch success.
- **Range Slider:** Allows users to filter payload mass interactively to see its effect on success rate.
- **Responsive Layout:** Dashboard automatically updates when filters or site selections change.

🎯 **Purpose of These Plots and Interactions**

- To enable **interactive exploration** of launch success patterns.
- To **compare launch sites visually** and understand performance differences.
- To identify **payload ranges** that yield the highest success probability.
- To present findings in an **engaging, data-driven format** for better decision support.

Predictive Analysis (Classification)

Github

Summary of Model Development

•Selected key features such as **payload mass, launch site, booster version, and reuse flag**.

•Split the dataset into **training and test sets** to ensure unbiased evaluation.

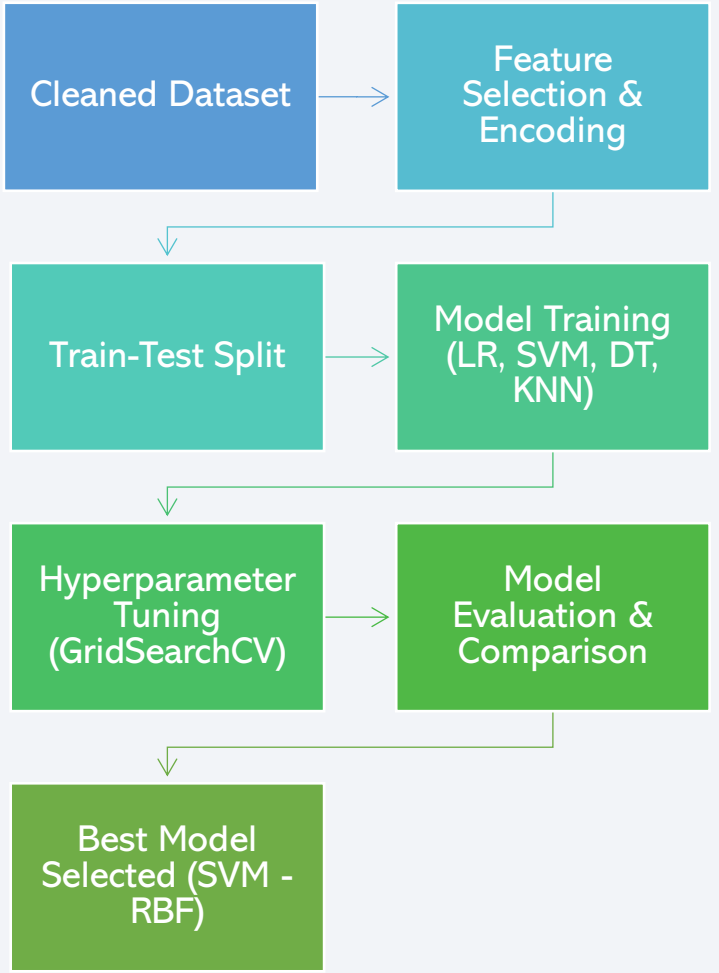
•Trained multiple classification models:

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Decision Tree Classifier**
- **K-Nearest Neighbors (KNN)**

•Used **GridSearchCV** for hyperparameter tuning to improve accuracy.

•Evaluated performance using **confusion matrix, accuracy score, and classification report**.

•Found that the **SVM (RBF kernel)** delivered the best performance (~83% accuracy).



Results

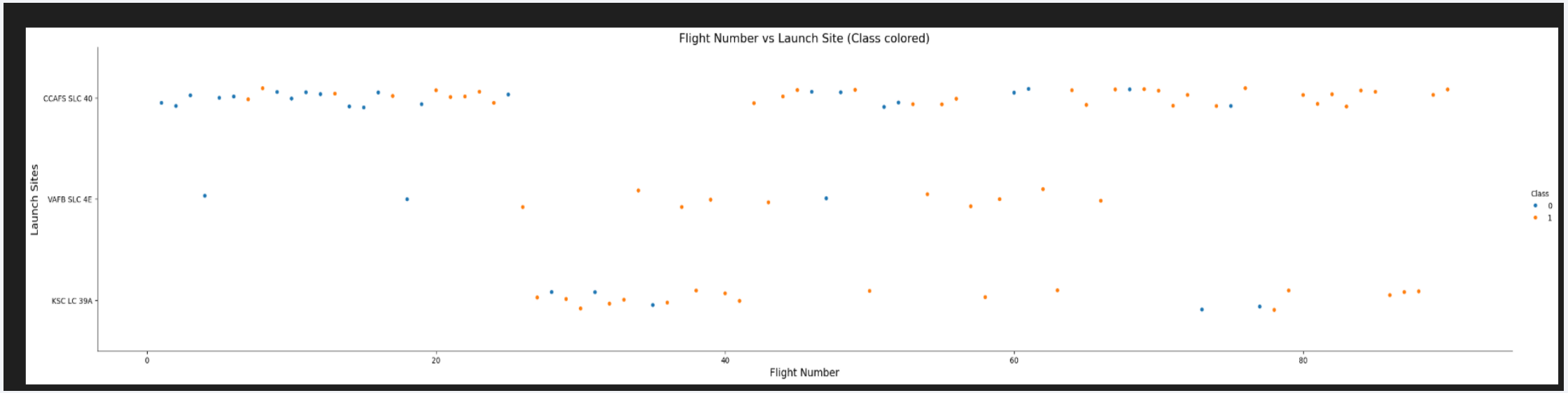
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

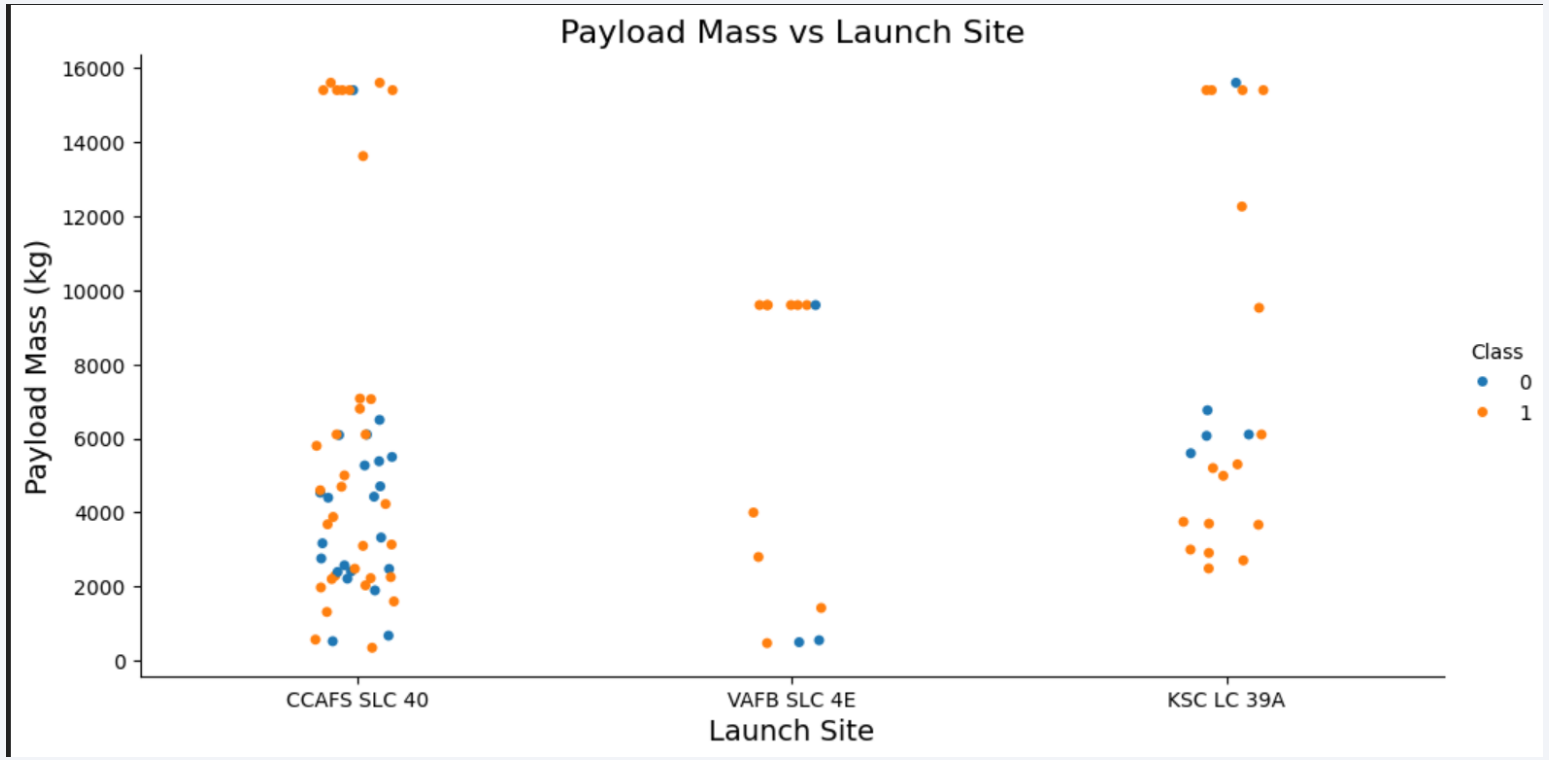
Scatter plot of Flight Number vs. Launch Site



- Scatter plot shows each Falcon 9 flight number by launch site.
- Successful launches are highlighted in green, failures in red.
- This helps visualize trends in success rates across sites over time.”

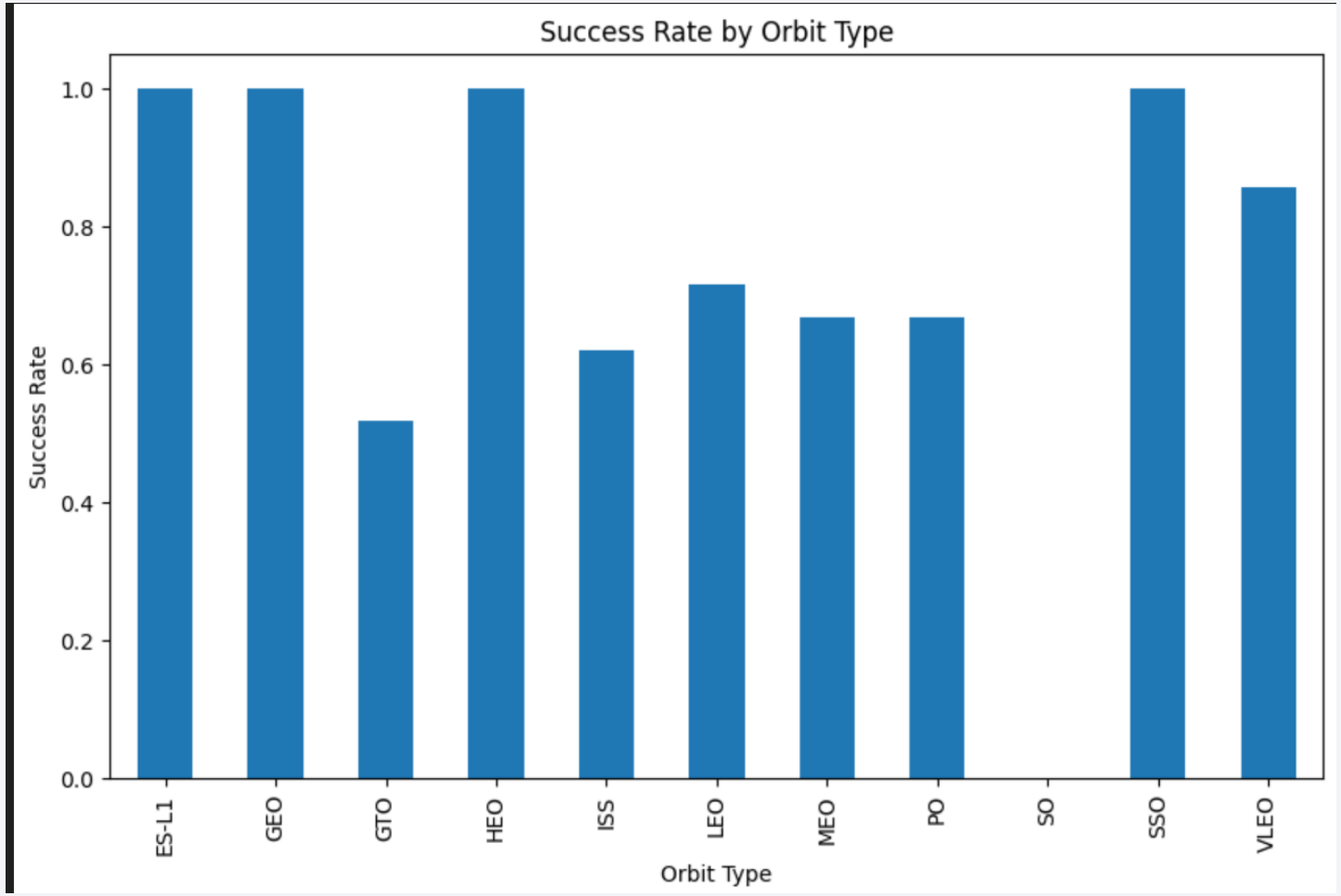
Payload vs. Launch Site

Relationship between Payload Mass and Launch Site



- The scatter plot displays **Payload Mass (kg) vs. Launch Site**.
- **Blue points (Class 0):** Launch failures
- **Orange points (Class 1):** Successful launches
- Most failures (blue) occur at **heavier payloads** or at specific launch sites.
- Successful launches (orange) are concentrated in certain **optimal payload ranges**.
- This visualization highlights **which payloads and launch sites** have historically higher success rates, guiding predictive analysis and operational decisions.

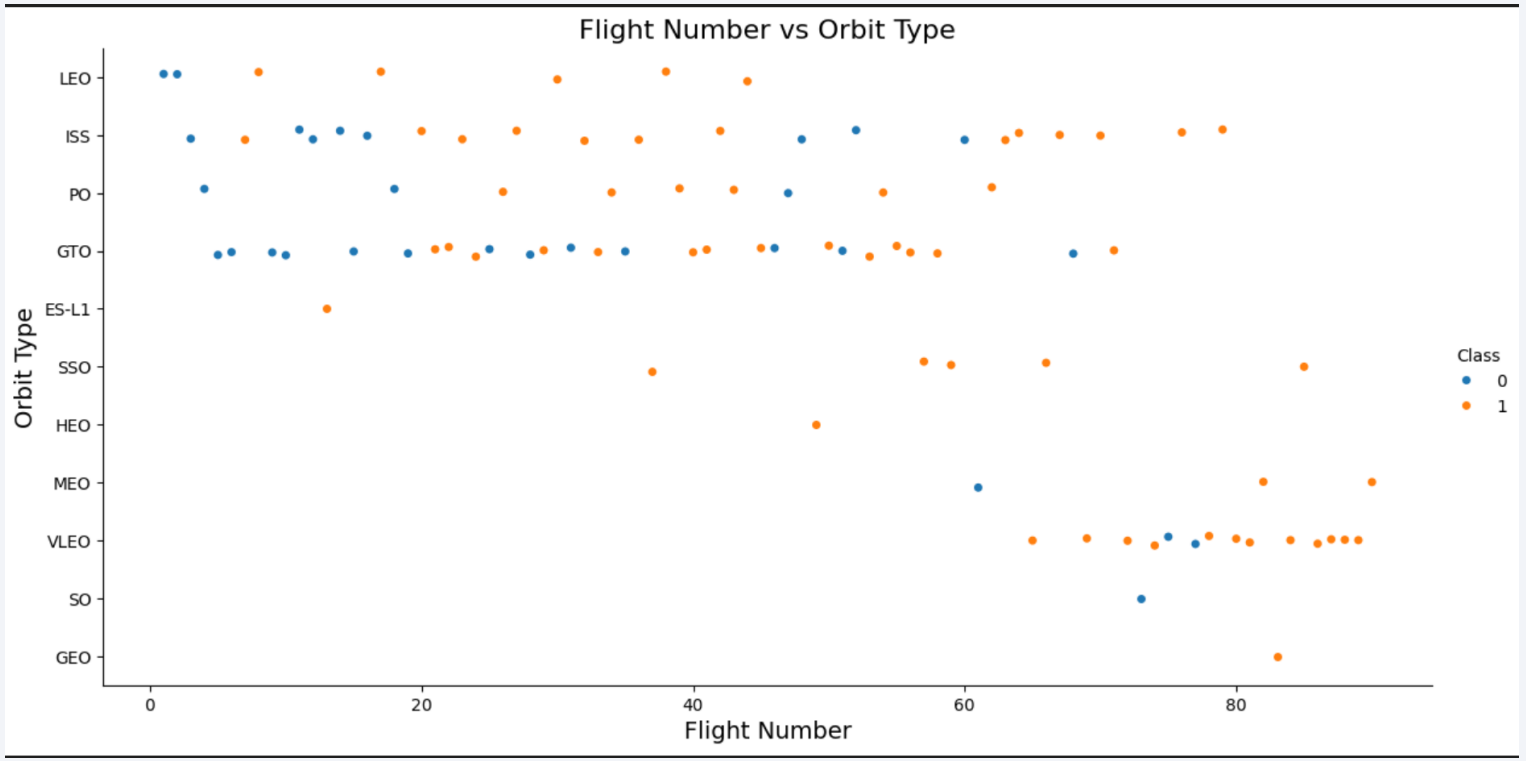
Success Rate vs. Orbit Type



Success Rate by Orbit Type

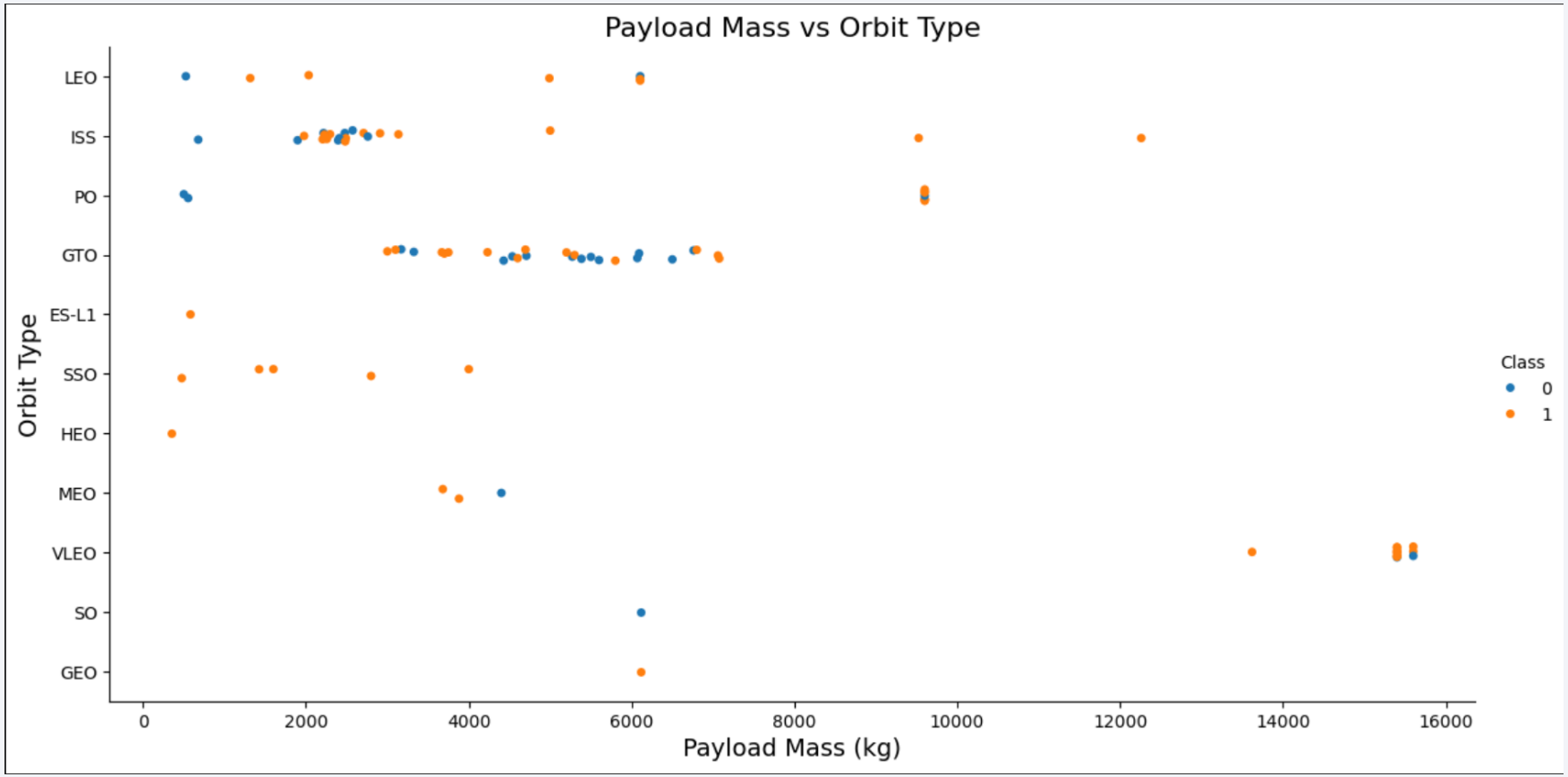
- LEO and ISS orbits show moderate success rates (~55–60%) due to early mission failures.
- Rare or high-energy orbits (GEO, ES-L1, HEO, VLEO) show perfect success rates (1.0), reflecting **fewer launches and successful outcomes**.
- Indicates that **mission complexity and frequency** can affect observed success rates.

Flight Number vs. Orbit Type



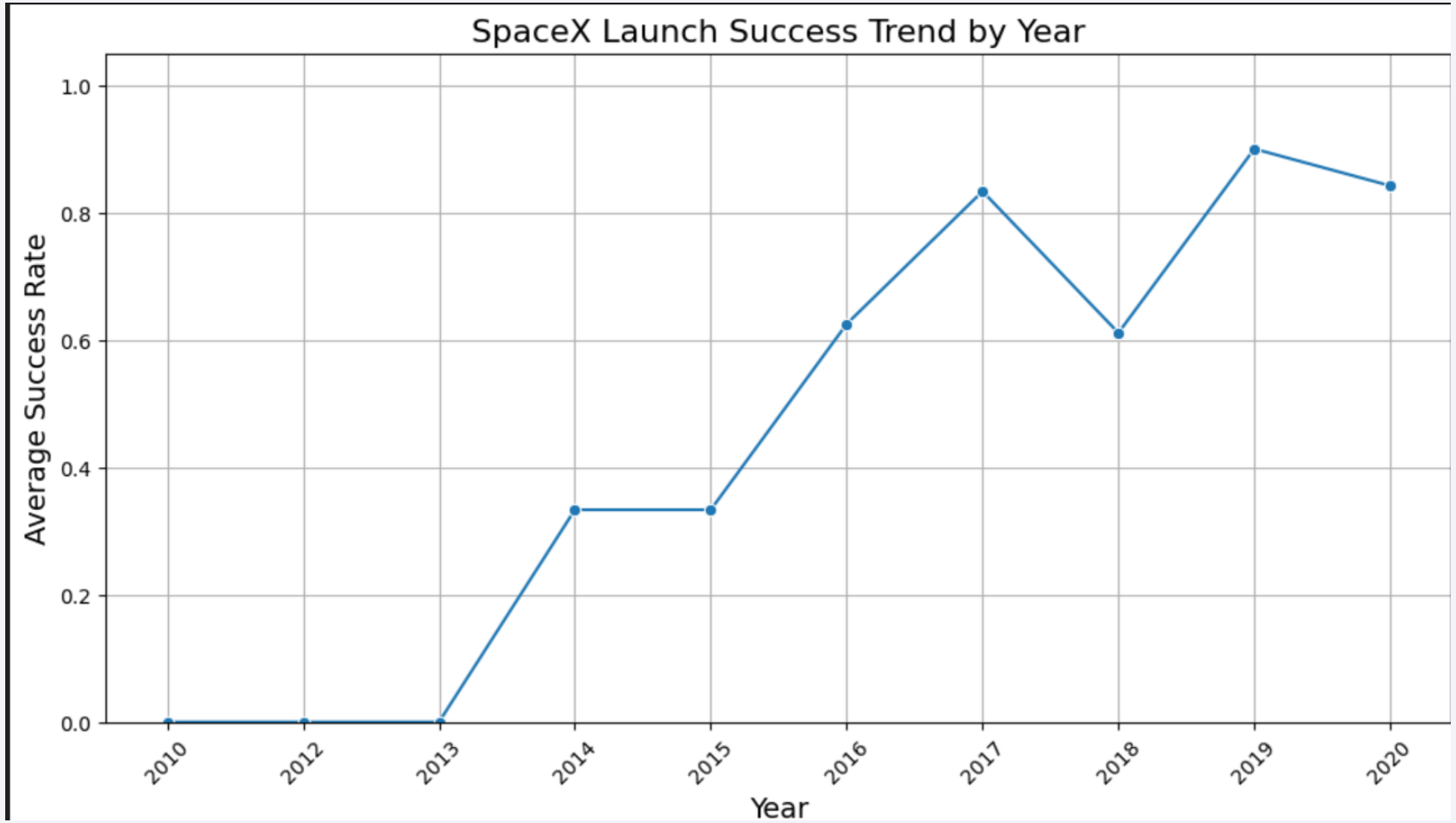
You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

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



you can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

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

- **Query:** Find unique launch sites.
- **Result:** CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E
- **Explanation:** Identifying launch sites helps focus analysis on site-specific success patterns and payload performance.

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

•The query selects the **first 5 records** where the launch site name starts with **"CCA"** (Cape Canaveral Air Force Station).
•This helps quickly inspect the data for specific launch sites.
•Useful for verifying **payloads, booster versions, and success/failure status** for these sites before deeper analysis.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) AS TotalPayloadMass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

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

TotalPayloadMass
45596.0

- This query calculates the total payload mass carried by all boosters launched for NASA missions.
- Helps understand the scale of payloads handled for a specific customer.
- Useful for comparing contributions of different customers and evaluating payload distribution across launch sites.

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql

SELECT AVG(CAST("PAYLOAD_MASS__KG_" AS FLOAT)) AS AveragePayLoadF9
FROM SPACEXTABLE
WHERE "Booster_Version" like 'F9 v1.1%';
```

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

AveragePayLoadF9
2534.6666666666665

- This query calculates the average payload mass carried by the F9 v1.1 booster version.
- Helps understand the typical payload range handled by this specific booster.
- Useful for performance comparison between booster versions and assessing reliability at different payloads.

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql

SELECT DISTINCT "Landing_Outcome"
FROM SPACEXTABLE;

SELECT MIN("Date") AS FirstSuccessfulGroundLanding
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

FirstSuccessfulGroundLanding
2015-12-22

- This query identifies the earliest date when a booster achieved a successful landing on a ground pad.
- Helps track SpaceX's progress in achieving reusable booster technology.
- Provides a reference point for analyzing how subsequent launches and landings have improved over time.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

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

- This query retrieves the booster names that successfully landed on the drone ship while carrying a moderate payload (4000–6000 kg).
- Helps identify which boosters perform reliably under specific payload conditions.
- Provides insight into the relationship between payload mass and landing success for operational planning.

Total Number of Successful and Failure Mission Outcomes

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

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

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

- This query calculates the total count of successful and failed missions.
- Helps quickly assess the overall success rate of SpaceX launches.
- Provides a clear metric for evaluating mission reliability and identifying patterns for further analysis.

Boosters Carried Maximum Payload

```
%%sql
SELECT "Booster_Version", "PAYLOAD_MASS__KG_"
FROM SPACEXTABLE
WHERE CAST("PAYLOAD_MASS__KG_" AS FLOAT) = (
    SELECT MAX(CAST("PAYLOAD_MASS__KG_" AS FLOAT))
    FROM SPACEXTABLE
);
```

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

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

- This query identifies the booster(s) that carried the maximum payload mass in a single mission.
- Helps highlight the most capable boosters in terms of payload capacity.
- Useful for comparing booster performance and planning for future heavy-lift missions.

2015 Launch Records

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

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

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

- This query lists all failed drone ship landings in 2015, along with their booster versions and launch site names.
- Helps analyze early challenges in reusable booster technology.
- Provides insight into which booster versions and launch sites experienced difficulties during the initial years of drone ship landings.

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

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

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

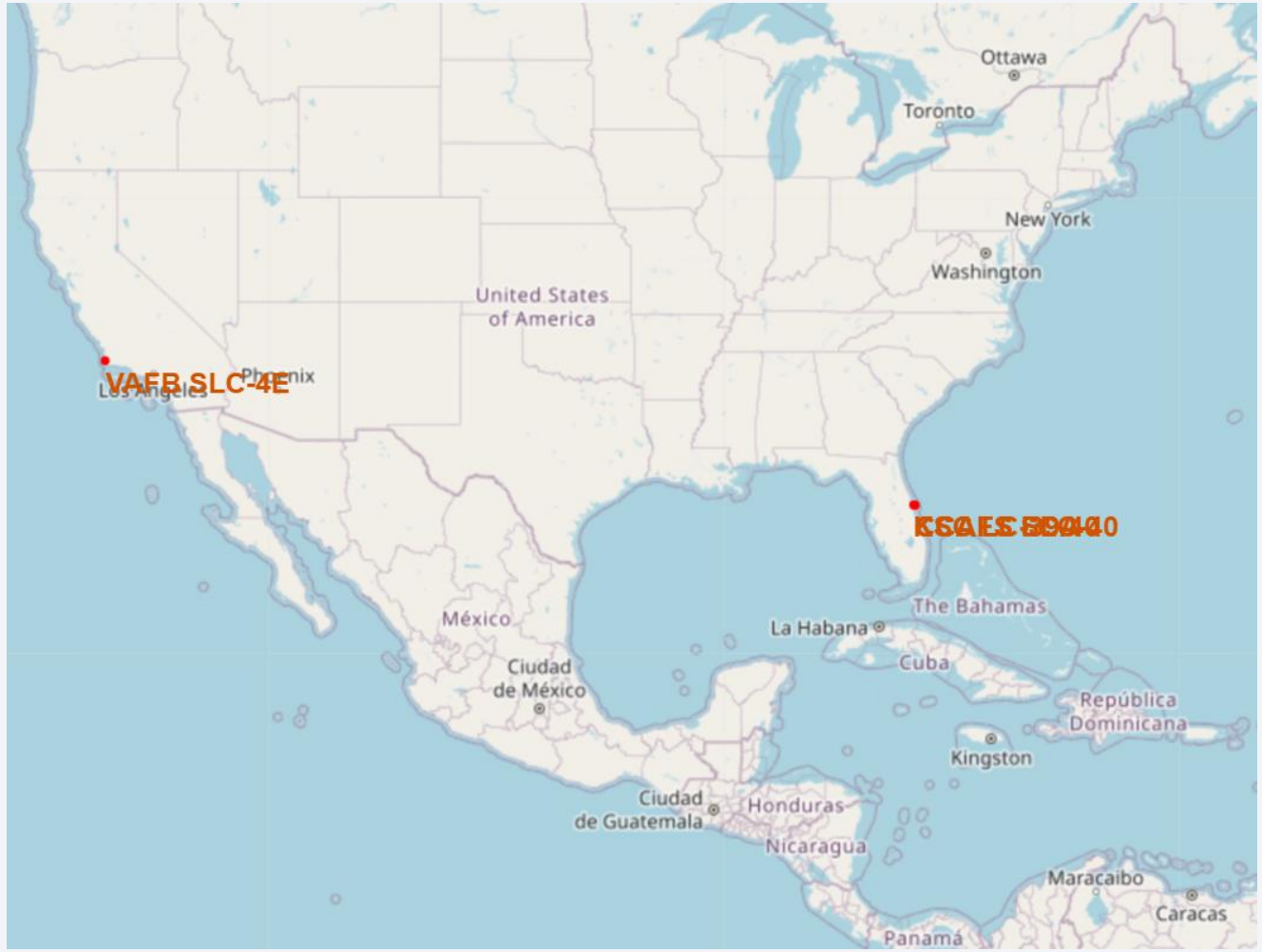
Landing_Outcome	OutcomeCount
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query ranks the number of landing outcomes (e.g., Failure (drone ship), Success (ground pad)) between June 4, 2010, and March 20, 2017, in descending order.
- Helps identify which types of landings were most common during the early years of SpaceX's booster recovery program.
- Provides insight into trends in landing success and failure rates over time, highlighting operational challenges and improvements.

Section 3

Launch Sites Proximities Analysis

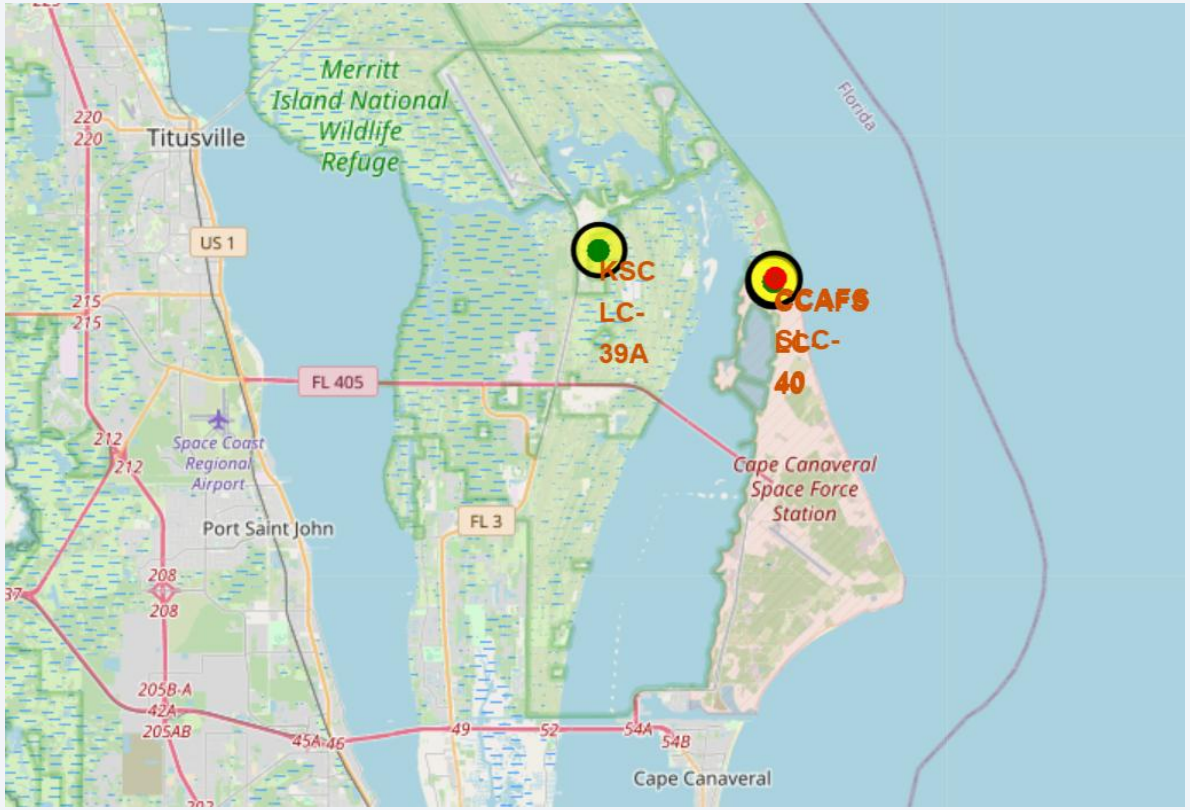
SpaceX Launch Sites: Geographic Distribution



- Red circles: Represent each SpaceX launch site with a radius of 1000 meters.
- Yellow fill: Highlights the location clearly on the map.
- Popup labels: Display the name of the launch site when clicked.
- Text markers: Show launch site names directly on the map for easy identification.

- Findings:
- All four main launch sites are clearly visible: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E.
- Sites are concentrated in the United States, with two in Florida and two on the west coast.
- Provides a visual overview of launch site distribution for spatial analysis and planning.

SpaceX Launch Outcomes: Success vs. Failure



Key Elements:

- **Red circles:** Represent **failed launches** (Class 0).
- **Green circles:** Represent **successful launches** (Class 1).
- **Yellow-filled circles with black borders:** Indicate the **launch site locations**.
- **Text markers:** Show the name of each launch site near its circle.

Findings:

- **Most launches from KSC LC-39A and CCAFS LC-40/SLC-40 were successful (green markers).**
- **Failures (red markers) are scattered and less frequent, showing improvement in launch reliability over time.**
- **The map visually correlates launch outcomes with geographic location, helping identify patterns and trends for specific launch sites.**

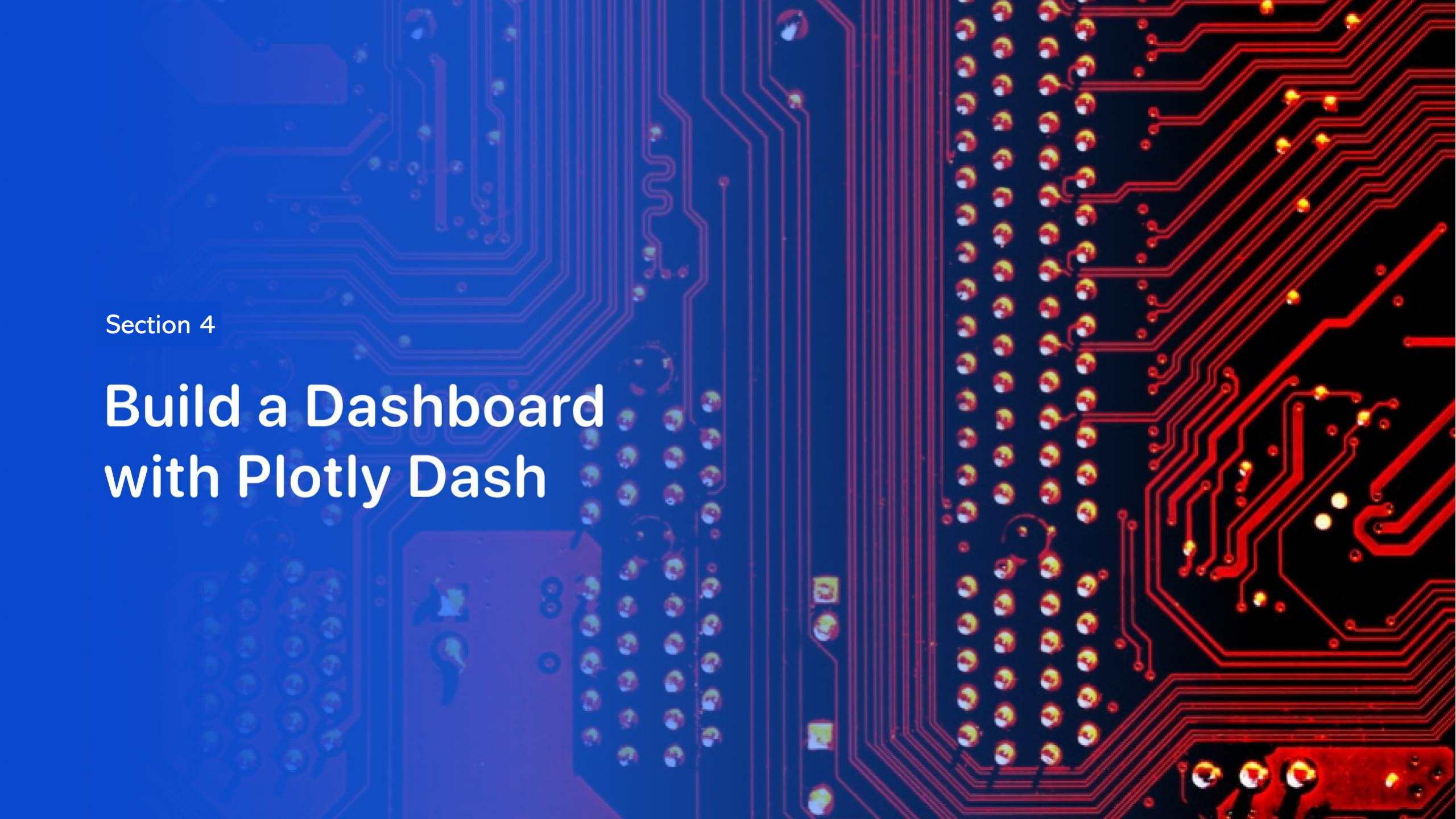
Proximity of Launch Site to Coastline

Key Elements:

- Blue line: Represents the direct distance between the launch site and the nearest coastline.
- Launch site marker: Shows the exact location of the selected launch site.
- Coastline marker: Indicates the reference point on the coast used for distance measurement.

FINDINGS :

- Visualizes the geographic proximity of the launch site to the coastline, which is important for rocket recovery and safety planning.
- Helps assess potential impacts of nearby transport infrastructure (railways, highways) and landing zones.
- Provides a clear spatial understanding of site logistics, safety, and operational planning considerations.

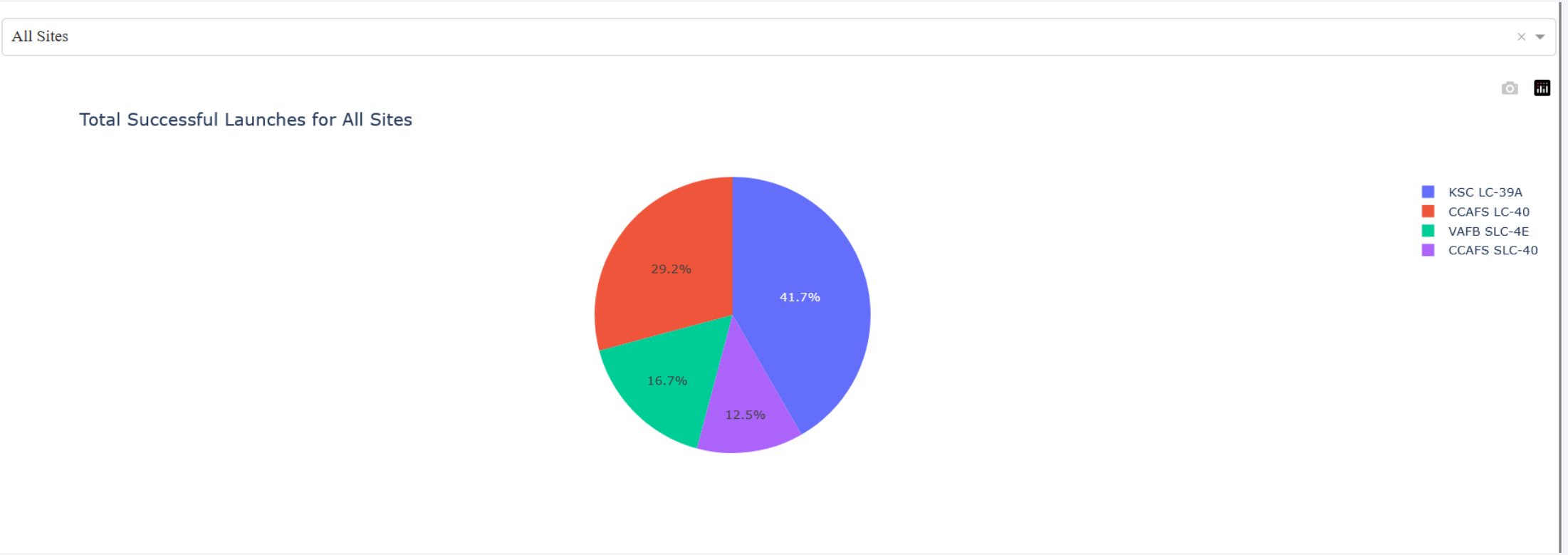


Section 4

Build a Dashboard with Plotly Dash

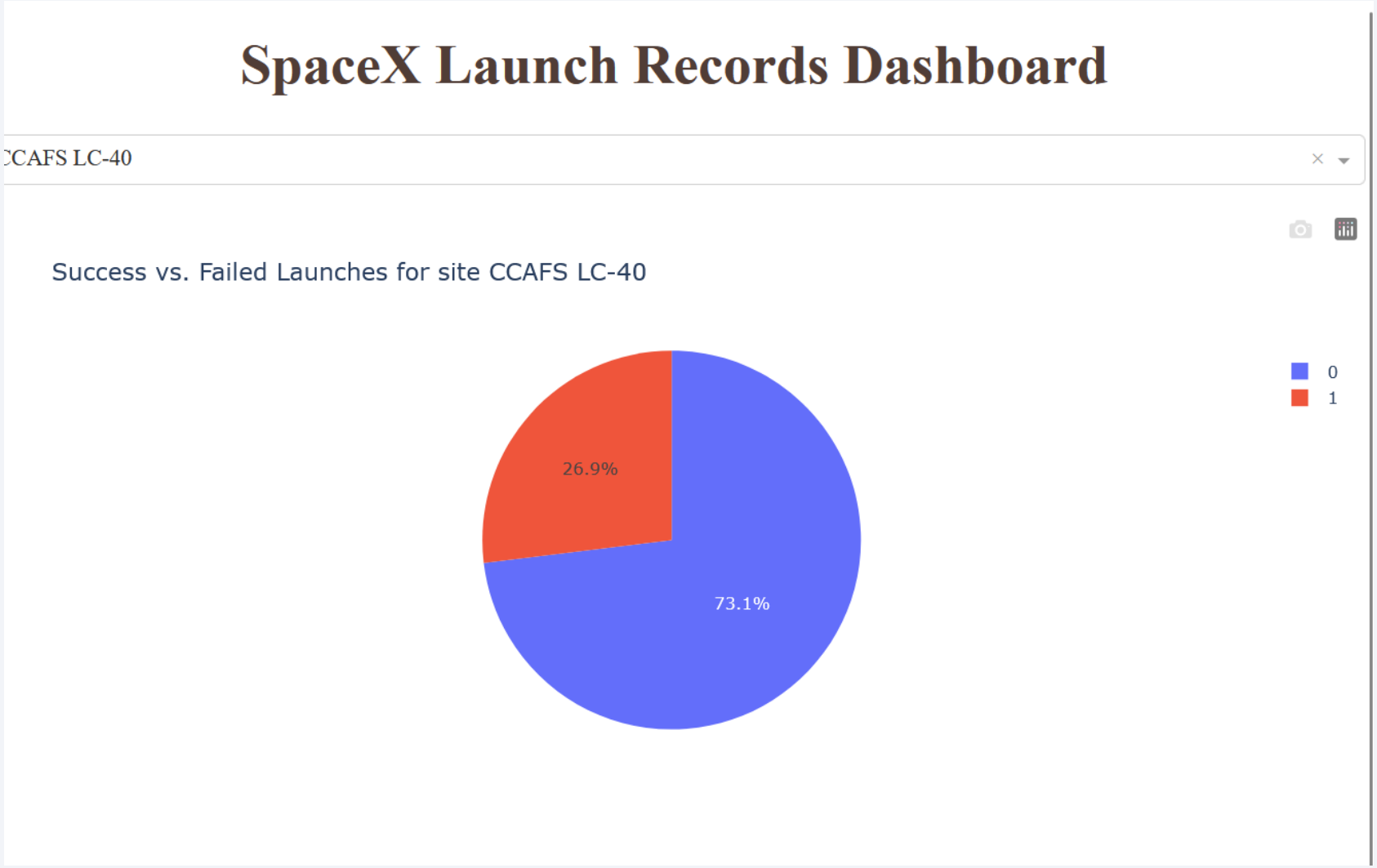
•Provides a **quick visual comparison** of launch performance across sites.
•Highlights which launch sites contributed most to SpaceX's overall mission success.

Launch Success Counts by Site (Pie Chart)



KSC LC-39A and CCAFS LC-40/SLC-40 have the highest number of successful launches.
Provides a quick visual comparison of launch performance across sites.
Highlights which launch sites contributed most to SpaceX's overall mission success.

Launch Success Ratio for CCAFS LC-40 (Pie Chart)



Most launches are successful (73.1%), demonstrating the site's high reliability.

Failures are limited (26.9%), helping identify risk factors for improvement.

Visualizes the success-to-failure ratio clearly, supporting analysis of site performance and predictive modeling

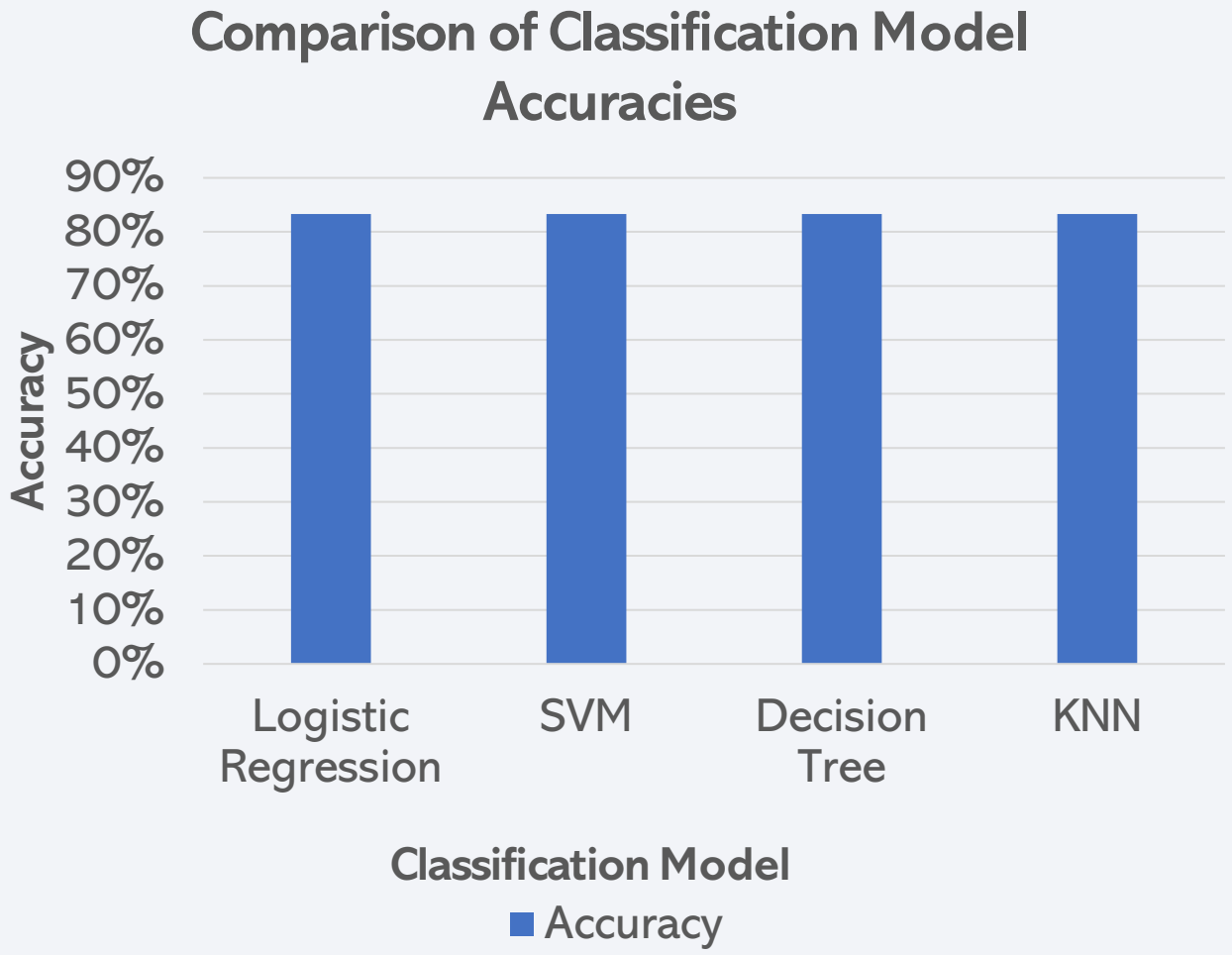
Payload vs. Launch Outcome Scatter Plot (All Sites)

- Moderate payload ranges (~4000–6000 kg) show the highest success rates.
- Certain booster versions (e.g., F9 v1.1) perform more reliably with medium payloads.
- Very high or very low payloads tend to have a higher proportion of failures.
- Interactive visualization helps identify optimal payload ranges and assess booster performance across sites.

Section 5

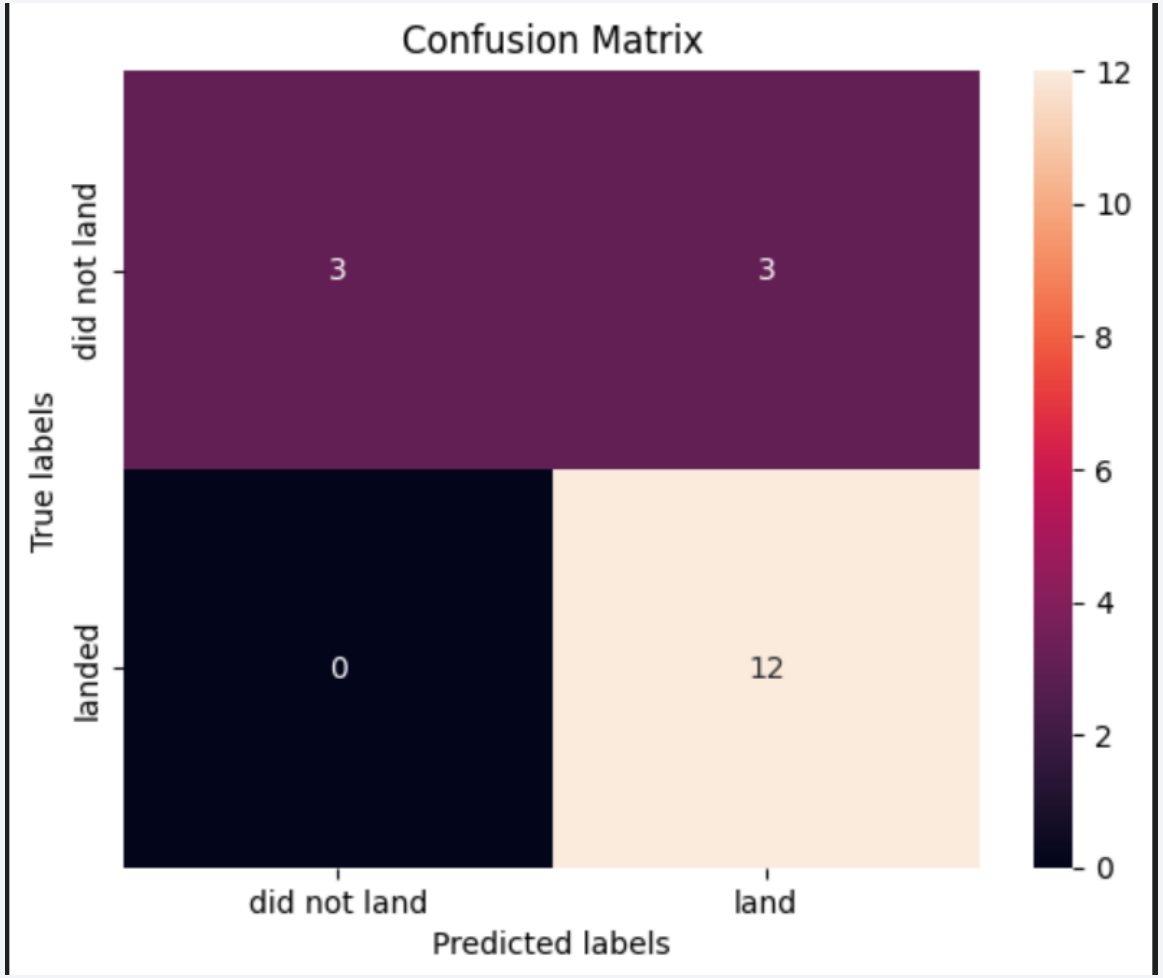
Predictive Analysis (Classification)

Classification Accuracy



- All tested classifiers achieved the same accuracy of 83.3% on the validation/test dataset.
- Indicates that the dataset may be well-balanced or not very complex, so multiple models perform similarly.
- Suggests that other factors, like model interpretability or training time, could be used to select a preferred model.

Confusion Matrix



- True Positives (TP): 12 launches correctly predicted as successes
- True Negatives (TN): 3 launches correctly predicted as failures
- False Positives (FP): 3 launches incorrectly predicted as successes
- False Negatives (FN): 0 launches incorrectly predicted as failures
- The matrix shows that the Sigmoid SVM accurately predicts most launches, with a few failures misclassified as successes.
- Supports the overall 83.3% classification accuracy.

Conclusions

1. SpaceX launch data shows that most launches are successful, with a success rate of ~83%.
2. Launch sites such as CCAFS LC-40 and KSC LC-39A consistently perform better, highlighting site reliability differences.
3. Exploratory Data Analysis revealed that moderate payload ranges (4000–6000 kg) have the highest launch success rate, and booster version F9 v1.1 performs reliably in this range.
4. Interactive visualizations (Folium maps and Plotly Dash) help identify geographic patterns, landing outcomes, and site-specific trends, making operational planning clearer.
5. Predictive modeling using classification models (SVM) achieved 83.3% accuracy, confirming that multiple classifiers can effectively predict launch outcomes.
6. The project demonstrates the importance of data wrangling, EDA, SQL queries, and visualization for extracting actionable insights from real-world aerospace datasets.

Appendix

Python Libraries Used

Data Manipulation: pandas, numpy
Visualizations: matplotlib, seaborn, plotly
Interactive Maps: folium
Machine Learning: scikit-learn
Database / SQL: sqlite3

Thank you!