Winning Space Race with Data Science

Leonardo Aldecocea
02/08/2024
https://github.com/Lalde004/Space_Y_Final_Project.git

Outline

Executive Summary

Methodology

1. Data preprocessing and feature engineering
2. Split data into training and test sets
3. Applied four machine learning models:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. K-Nearest Neighbors (KNN)

Results

Three models performed equally well (83.333% accuracy): linear regression, SVM, and KNN

Introduction

- Project Background and Context:

1. SpaceX's Falcon 9 rocket launches are revolutionizing the space industry with lower-cost space travel.
2. The ability to reuse the first stage of the Falcon 9 rocket is key to SpaceX's cost advantage.
3. Predicting the success of first stage landings is crucial for estimating launch costs.
4. Competing launch companies need to understand factors influencing landing success to remain competitive.

- Key Problems to Address:

1. Can we accurately predict the success of Falcon 9 first stage landings?
2. What are the most important factors influencing the success of these landings?
3. How do different launch parameters (e.g., payload mass, orbit type, launch site) affect landing success probability?
4. Can we develop a model that helps estimate the cost of a launch for competing companies?
5. How can we use historical launch data to inform future launch strategies and improve success rates?

Section 1

Methodology

Methodology

1. Data Collection:
    1. Gathered SpaceX launch data from public API
    2. Scraped additional information from SpaceX website
2. Data Wrangling:
    1. Cleaned and structured raw data
    2. Handled missing values and outliers
    3. Standardized data formats
3. Exploratory Data Analysis (EDA):
    1. Used SQL queries for initial data exploration
    2. Created visualizations to identify patterns and relationships
    3. Analyzed correlations between variables
4. Interactive Visual Analytics:
    1. Developed interactive maps using Folium to visualize launch sites
    2. Created dynamic dashboards with Plotly Dash for data exploration
5. Predictive Analysis:
    1. Prepared data for machine learning (feature selection, encoding)
    2. Split data into training and test sets
    3. Applied classification models: Logistic Regression, SVM, Decision Tree, KNN

Data Collection

Collection:

Primary Sources, API data retrieval, Web Scraping, Data Integration

- Flowchart:

- [SpaceX API] → [Data Extraction] → [JSON Storage] ↓ [SpaceX Website] → [Web Scraping] → [HTML Parsing] ↓ [Data Integration] → [Unified Dataset]

Data Collection – SpaceX API

- Process:

1. Identified SpaceX API endpoint for launch data
2. Configured request parameters (e.g., limit, offset)
3. Sent GET requests to retrieve launch records
4. Parsed JSON responses to extract launch details
5. Stored data in Pandas DataFrame for analysis
6. Cleaned and preprocessed the collected data

- Space_Y_Final_Project/jupyter-labs-spacex-data-collection-api.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

[Define API Endpoint] → [Set Request Parameters] ↓ [Send GET Request] → [Receive JSON Response] ↓ [Parse JSON Data] → [Extract Relevant Information] ↓ [Store in DataFrame] → [Clean and Preprocess Data]

Data Collection - Scraping

- Process:

1. Identified SpaceX website pages with relevant launch data
2. Sent HTTP requests to target URLs
3. Retrieved HTML content of the web pages
4. Used BeautifulSoup to parse HTML structure
5. Located specific HTML elements containing desired information
6. Extracted data from identified elements
7. Structured extracted data into a usable format
8. Stored scraped data in a Pandas DataFrame for further analysis

- [Space_Y_Final_Project/jupyter-labs-webscraping.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)]

[Identify Target URL] → [Send HTTP Request] ↓ [Receive HTML Content] → [Parse HTML with BeautifulSoup] ↓ [Locate Relevant Elements] → [Extract Data] ↓ [Structure Extracted Data] → [Store in DataFrame]

Data Wrangling

- Flowchart:

- [Raw Data] → [Remove Duplicates] ↓ [Handle Missing Values] → [Convert Data Types] ↓ [Feature Engineering] → [Standardize Formats] ↓ [Outlier Detection/Handling] → [Merge Datasets] ↓ [Final Cleaned Dataset]

- Process

1. Imported raw data from API and web scraping sources
2. Removed duplicate entries
3. Identified and handled missing values (imputation/removal)
4. Converted data types for consistency (e.g., dates, numerical values)
5. Standardized formats (e.g., units of measurement)
6. Detected and addressed outliers

- Space_Y_Final_Project/labs-jupyter-spacex-Data wrangling.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

EDA with Data Visualization

Summary of Charts and Their Purpose:

1. **Bar Charts**:
    1. Used to compare launch success rates across different sites
    2. Visualized the distribution of booster versions
2. **Pie Charts**:
    1. Displayed the proportion of successful vs. failed launches
    2. Showed the distribution of orbits for all launches
3. **Scatter Plots**:
    1. Explored relationship between payload mass and launch success
    2. Investigated correlation between flight number and launch outcome
4. **Line Graphs**:
    1. Tracked launch success rate over time
    2. Visualized trends in payload mass across years
5. **Box Plots**:
    1. Compared payload mass distributions for different orbits
    2. Identified outliers in flight durations
6. **Heatmaps**:
    1. Displayed correlation matrix of numerical features
    2. Visualized launch success rates by month and year

Space_Y_Final_Project/edadataviz.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

EDA with SQL

- Basic Data Exploration (SELECT, COUNT, DISTINCT)
- Launch Success Analysis (COUNT, GROUP BY, WHERE, HAVING)
- Payload Analysis (AVG, MAX, MIN, CASE)
- Temporal Analysis (EXTRACT, COUNT, GROUP BY, WHERE)
- Booster Analysis (JOIN, COUNT, GROUP BY, WHERE)
- Complex Queries (subqueries, CTEs, Window functions)
- Data Aggregation (SUM, AVG, ROLLUP)

Space_Y_Final_Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

Build an Interactive Map with Folium

1. **Markers**:
    1. Added for each launch site location
    2. Displayed site name and coordinates on click Purpose: To precisely indicate launch site locations
2. **Circles**:
    1. Created around each launch site
    2. Radius proportional to launch success rate Purpose: To visually represent the relative success of each site
3. **Polylines**:
    1. Connected launch sites to their respective landing zones
    2. Color-coded based on landing success rate Purpose: To illustrate the trajectory and success of first stage returns
4. **Heatmap Layer**:
    1. Showed density of launches across geographical areas Purpose: To highlight areas with high launch activity
5. **Choropleth Layer**:
    1. Colored states/countries based on number of launches Purpose: To display the distribution of launches across regions
6. **Custom Icons**:
    1. Different colors for successful vs. failed launches Purpose: To enhance visual distinction between launch outcomes
7. **Popup Windows**:
    1. Added to markers with detailed site information

- Space_Y_Final_Project/lab_jupyter_launch_site_location.ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

Build a Dashboard with Plotly Dash

- **Plots/Graphs**:

1. Pie Chart: Launch success rate by site
2. Scatter Plot: Payload mass vs. launch outcome
3. Bar Chart: Launch success count by year
4. Line Graph: Launch success rate over time

- **Interactions**:

1. Dropdown menu: Select launch site
2. Range Slider: Filter payload mass range
3. Radio Buttons: Choose orbit type
4. Date Picker: Select date range for analysis

Space_Y_Final_Project/spacex_records.py at main · Lalde004/Space_Y_Final_Project (github.com)

Predictive Analysis (Classification)

- Flowchart:

- [Preprocessed Data] → [Feature Selection] ↓ [Train-Test Split] → [Model Training] ↓ [Hyperparameter Tuning] → [Cross-Validation] ↓ [Model Evaluation] → [Model Comparison] ↓ [Select Best Model] → [Final Testing]

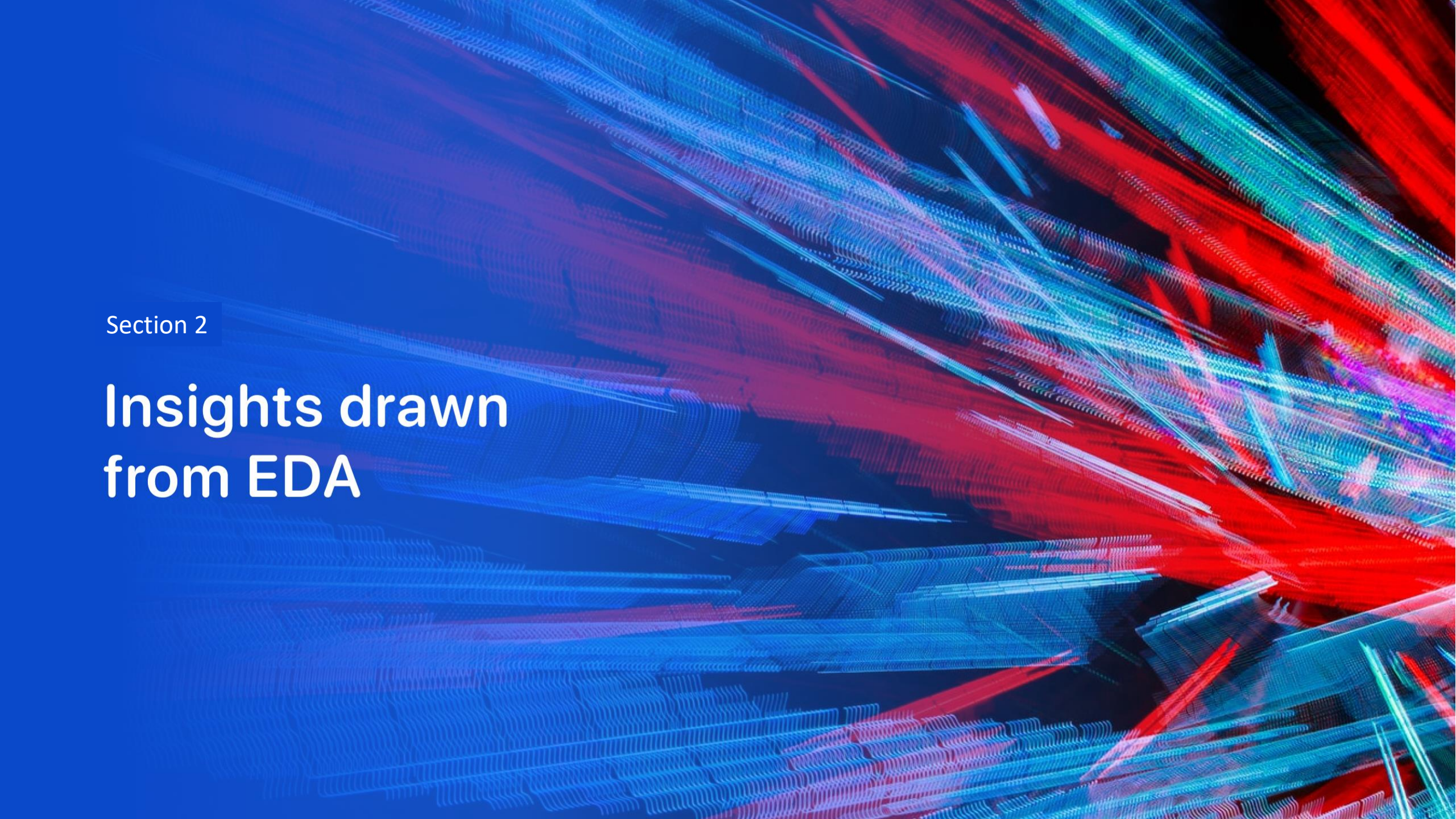
Process:

1. Selected relevant features based on EDA insights
2. Split data into training and testing sets (80-20 split)
3. Trained multiple models:
    1. Logistic Regression
    2. Support Vector Machine (SVM)
    3. Decision Tree
    4. K-Nearest Neighbors (KNN)
4. Performed hyperparameter tuning using GridSearchCV
5. Conducted 10-fold cross-validation for each model
6. Evaluated models using accuracy
7. Compared model performances

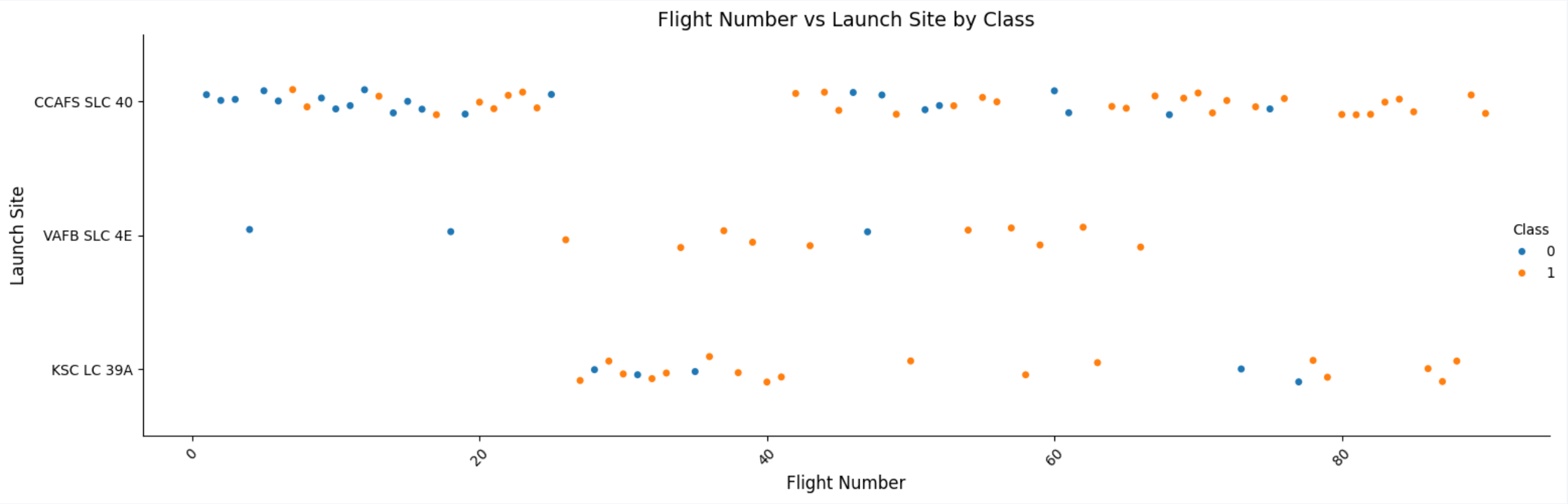
Space_Y_Final_Project/SpaceX_Machine Learning Prediction_Part_5 (1).ipynb at main · Lalde004/Space_Y_Final_Project (github.com)

Results

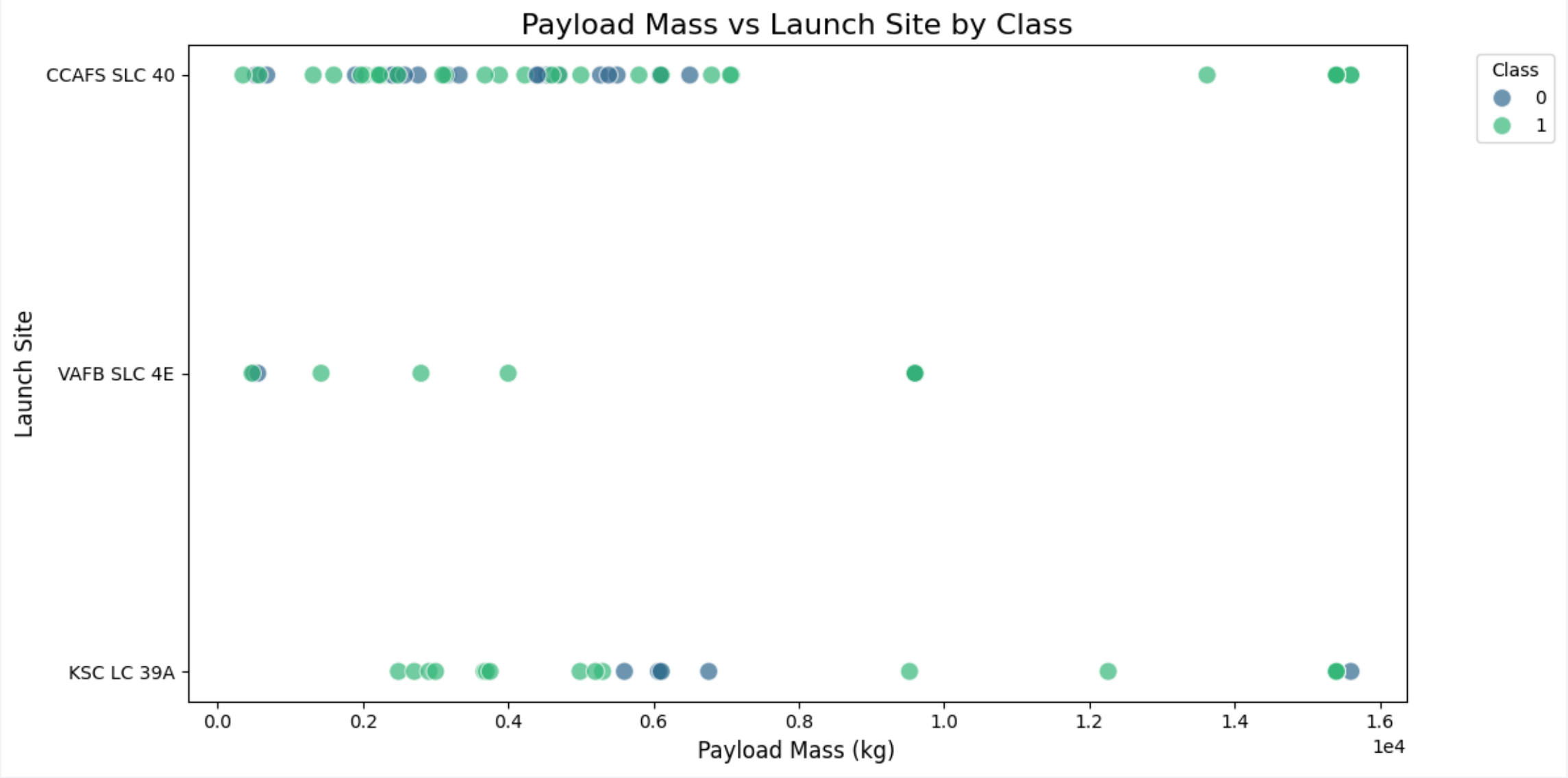
PREVIEW OF PLOTLY DASHBOARD


Section 2
Insights drawn from EDA

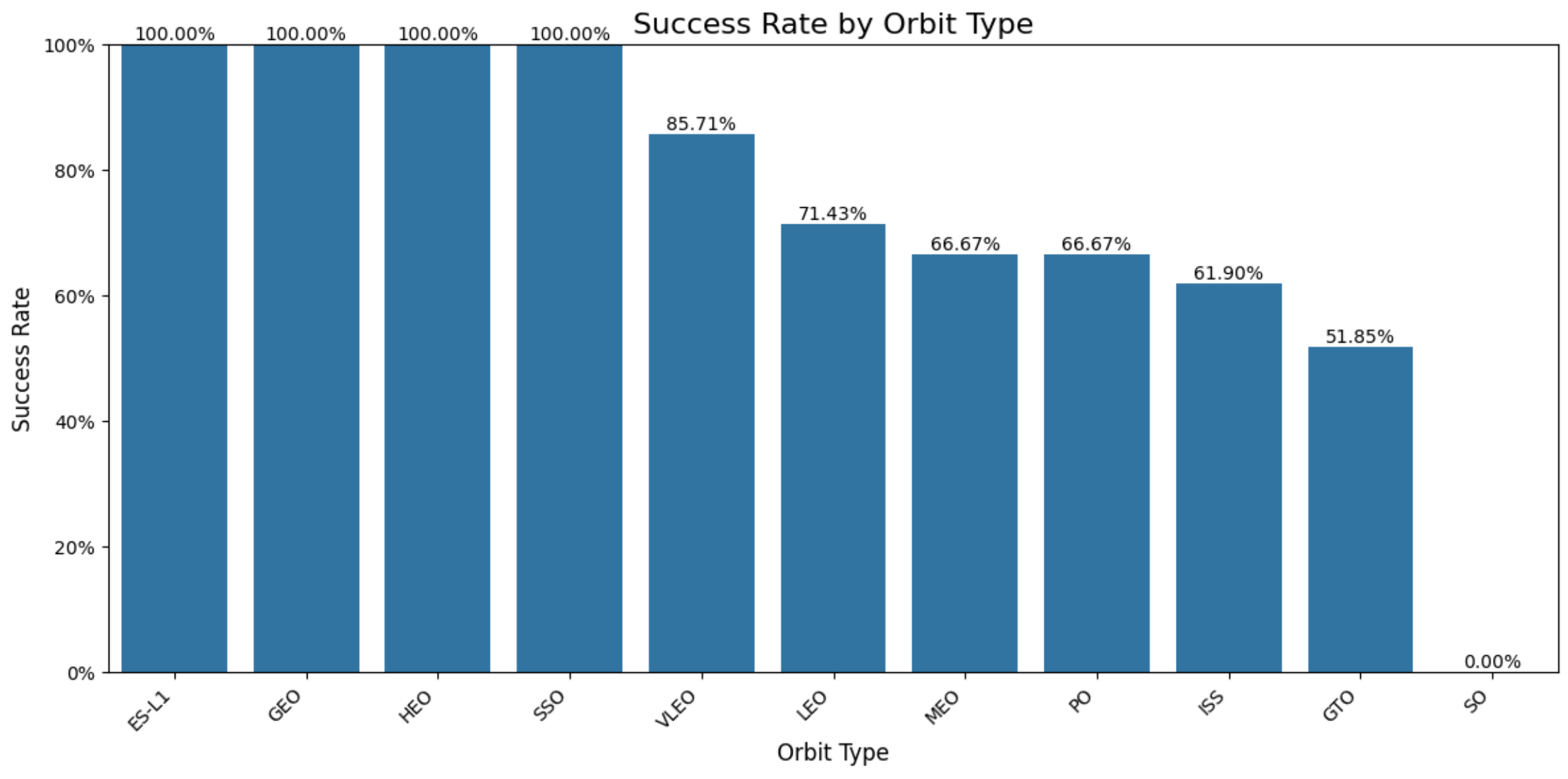
Flight Number vs. Launch Site



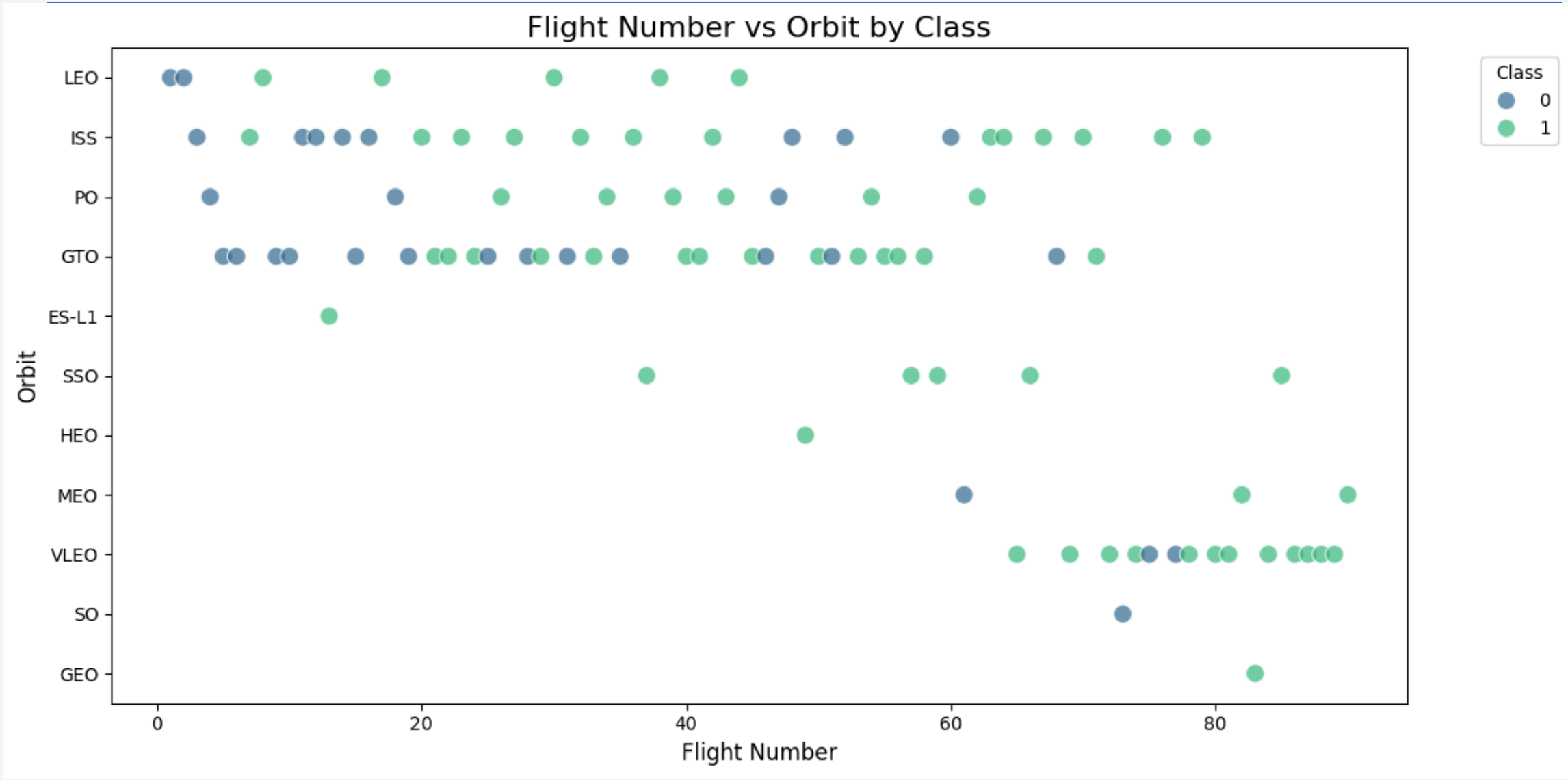
Payload vs. Launch Site



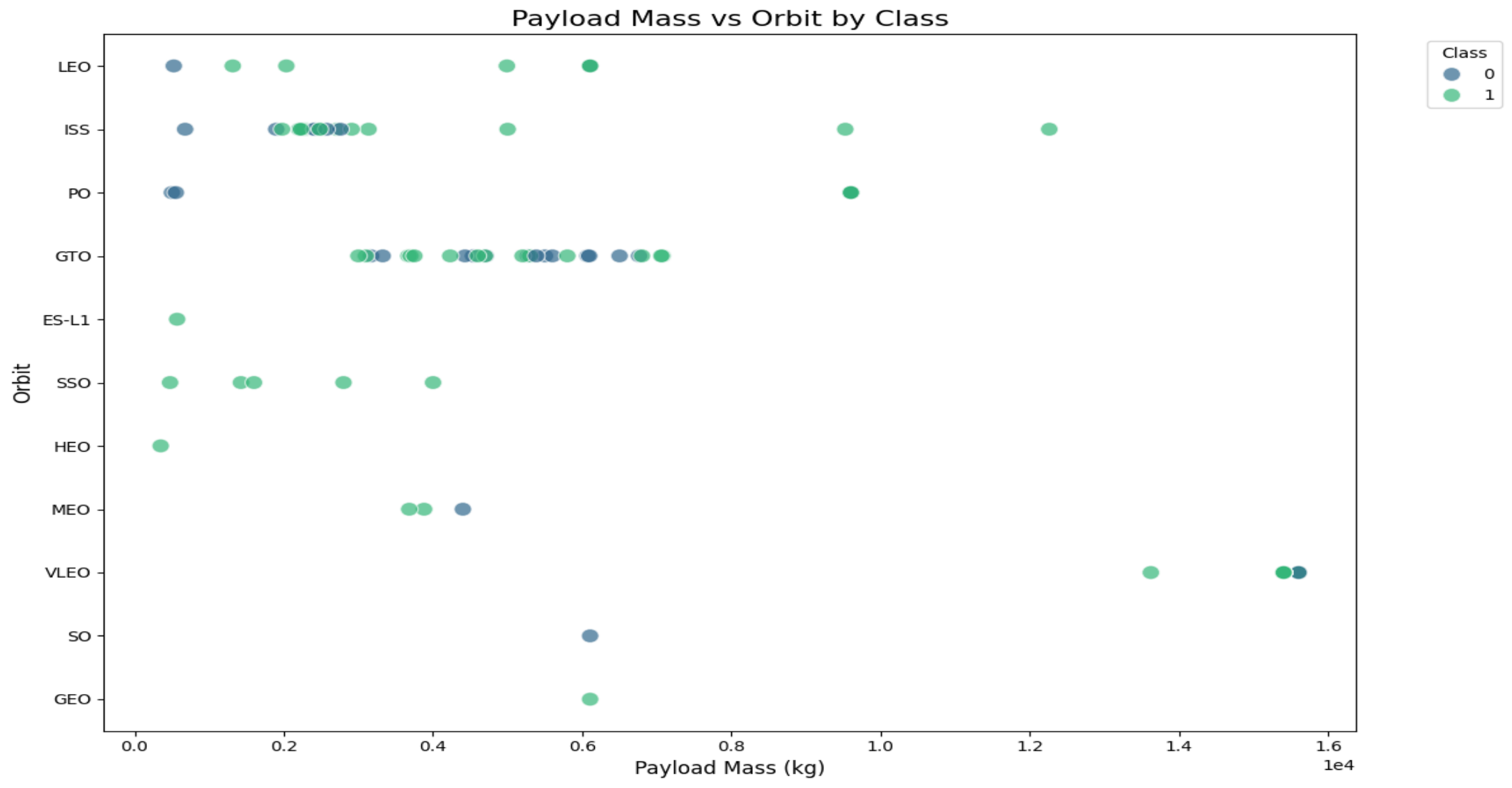
Success Rate vs. Orbit Type



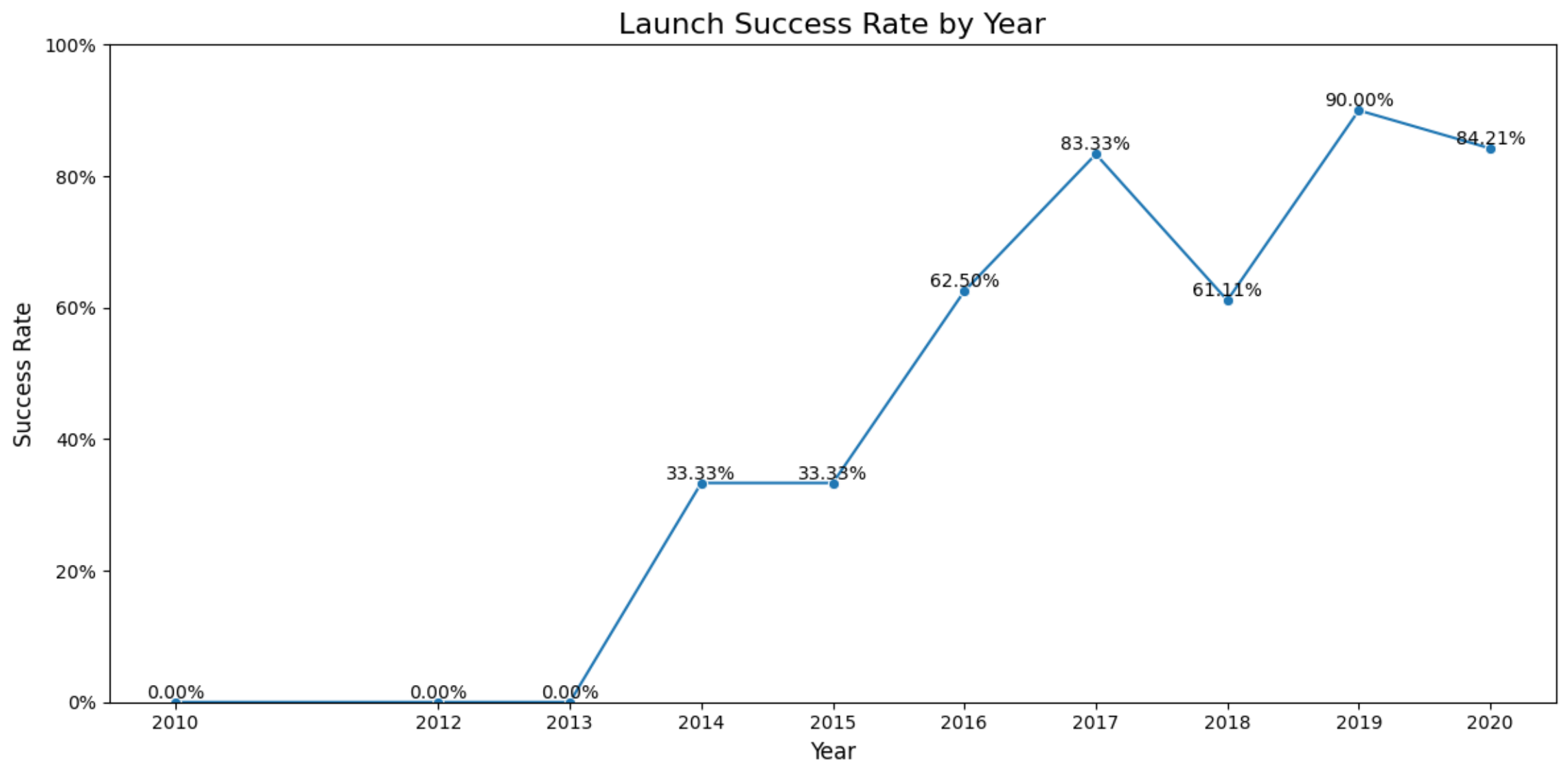
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- SELECT DISTINCT LaunchSite FROM SpaceXLaunchData ORDER BY LaunchSite;

- **LAUNCH SITES**

1.CCAFS LC-40: Cape Canaveral Air Force Station, Launch Complex 40

2.CCAFS SLC-40: Cape Canaveral Air Force Station, Space Launch Complex 40

3.KSC LC-39A: Kennedy Space Center, Launch Complex 39A

4.VAFB SLC-4E: Vandenberg Air Force Base, Space Launch Complex 4E

Launch Site Names Begin with 'CCA'

- `SELECT * FROM SpaceXLaunchData WHERE LaunchSite LIKE 'CCA%' LIMIT 5;`

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
1	2010-06-04	F9 v1.0	6104.959	LEO	CCAFS SLC-40	Success
2	2012-05-22	F9 v1.0	525	LEO	CCAFS SLC-40	Success
3	2013-03-01	F9 v1.0	677	ISS	CCAFS SLC-40	Success
5	2013-12-03	F9 v1.1	3170	GTO	CCAFS SLC-40	Success
6	2014-01-06	F9 v1.1	3325	GTO	CCAFS SLC-40	Success

Total Payload Mass

- SELECT SUM(PayloadMass) AS TotalNASAPayload FROM SpaceXLaunchData WHERE Customer = 'NASA' OR Customer LIKE '%NASA%';

- Total NASA Payload: 45,596.0 kg
- Key points about this result:

1. The query sums up the PayloadMass for all launches where the Customer is either exactly 'NASA' or contains 'NASA' in its name (to catch variations like 'NASA/NOAA').
2. The total of 45,596.0 kg represents the cumulative mass of all payloads SpaceX has launched for NASA missions.

Average Payload Mass by F9 v1.1

- SELECT AVG(PayloadMass) AS AveragePayloadMass FROM SpaceXLaunchData WHERE BoosterVersion = 'F9 v1.1';

- Average Payload Mass for F9 v1.1: 2,534.67 kg
- Key points about this result:

1.The Falcon 9 v1.1 was an upgraded version of the original Falcon 9, introduced in 2013.

2.The average payload of 2,534.67 kg (assuming the unit is kilograms) represents the typical mass of payloads this version of the Falcon 9 carried.

First Successful Ground Landing Date

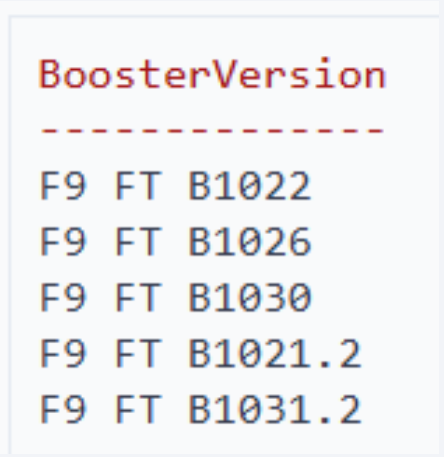
- SELECT Date, LaunchSite, LandingOutcome FROM SpaceXLaunchData WHERE LandingOutcome = 'Success (ground pad)' ORDER BY Date ASC LIMIT 1;

- **First Successful Ground Pad Landing:**
  - Date: December 22, 2015
- Launch Site: CCAFS SLC-40 (Cape Canaveral Air Force Station, Space Launch Complex 40)
- Landing Outcome: Success (ground pad)
  - Key points about this result:

1. This date marks a significant milestone in SpaceX's reusability efforts. It was the first time they successfully landed the first stage of a Falcon 9 rocket on solid ground after a mission.

Successful Drone Ship Landing with Payload between 4000 and 6000

- SELECT DISTINCT BoosterVersion FROM SpaceXLaunchData WHERE LandingOutcome = 'Success (drone ship)' AND PayloadMass > 4000 AND PayloadMass < 6000 ORDER BY BoosterVersion;



```
BoosterVersion
--------------
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1031.2
```

1. Five distinct booster versions met these criteria.
2. All boosters are variants of the Falcon 9 Full Thrust (FT) version.
3. The ".2" suffix on some boosters (e.g., B1021.2) indicates that these were reused boosters on their second flight.

Total Number of Successful and Failure Mission Outcomes

- SELECT SUM(CASE WHEN MissionOutcome = 'Success' THEN 1 ELSE 0 END) AS SuccessfulMissions, SUM(CASE WHEN MissionOutcome = 'Failure' THEN 1 ELSE 0 END) AS FailedMissions FROM SpaceXLaunchData;

- Successful Missions: 101
- Failed Missions: 1

Boosters Carried Maximum Payload

- SELECT DISTINCT BoosterVersion FROM SpaceXLaunchData WHERE PayloadMass = (SELECT MAX(PayloadMass) FROM SpaceXLaunchData);

BoosterVersion
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

1. Multiple boosters have carried the maximum payload mass.
2. All of these boosters are variants of the Falcon 9 Block 5 (F9 B5) version.

2015 Launch Records

- SELECT BoosterVersion, LaunchSite, LandingOutcome FROM SpaceXLaunchData WHERE YEAR(Date) = 2015 AND LandingOutcome LIKE 'Failure (drone ship)%';

BoosterVersion	LaunchSite	LandingOutcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

1. Two failed drone ship landing attempts in 2015:
2. a. First Attempt:
    1. Booster Version: F9 v1.1 B1012
    2. Launch Site: CCAFS LC-40 (Cape Canaveral Air Force Station, Launch Complex 40)
    3. Outcome: Failure (drone ship)
3. b. Second Attempt:
    1. Booster Version: F9 v1.1 B1015
    2. Launch Site: CCAFS LC-40
    3. Outcome: Failure (drone ship)

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

```
SELECT LandingOutcome, COUNT(*) as Count FROM SpaceXLaunchData WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LandingOutcome ORDER BY Count DESC;
```

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

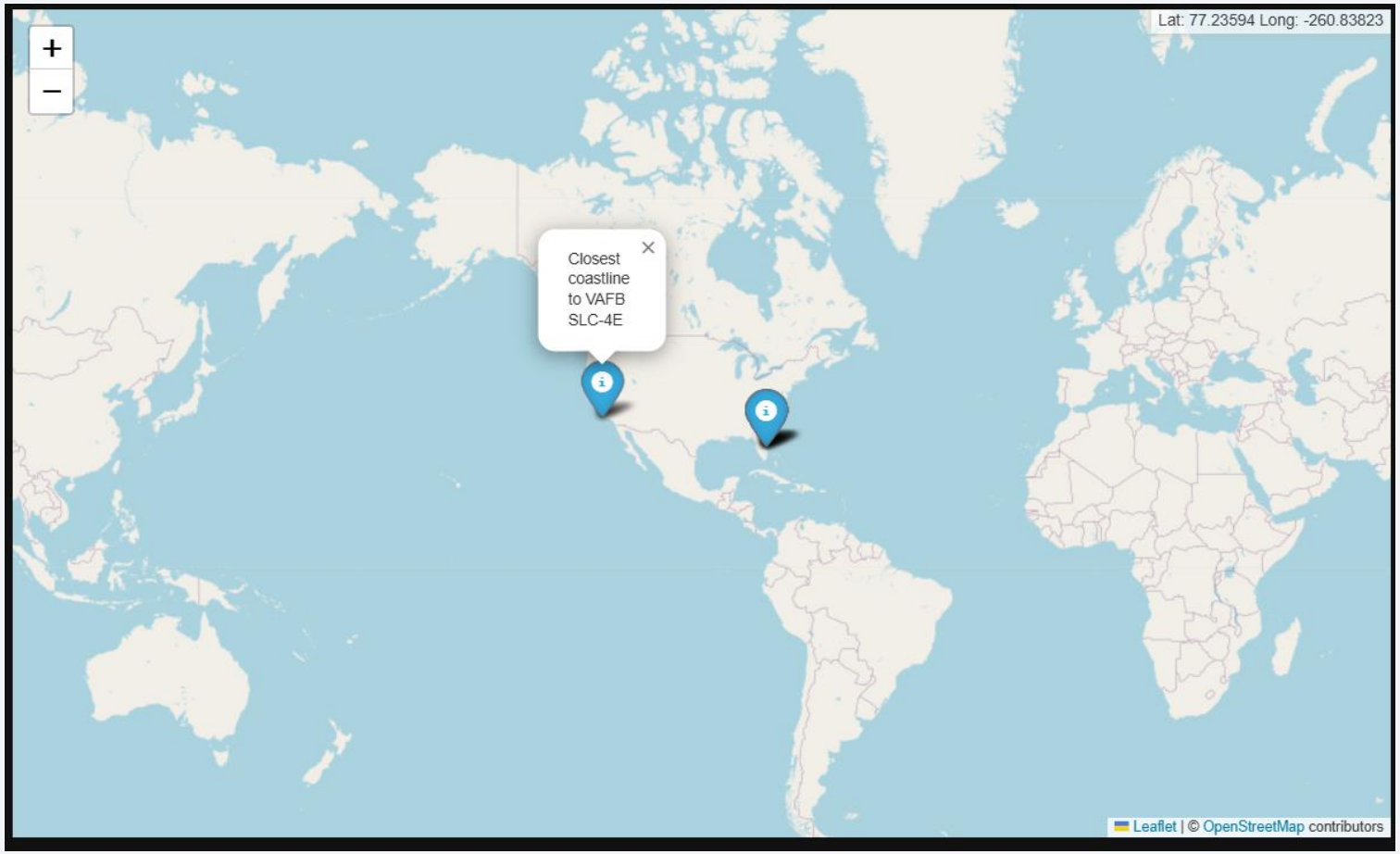
1.No attempt (10): The most common outcome, likely for earlier missions or those where landing wasn't planned.

2.Failure (drone ship) (5): Unsuccessful attempts to land on an autonomous spaceport drone ship.

Section 3

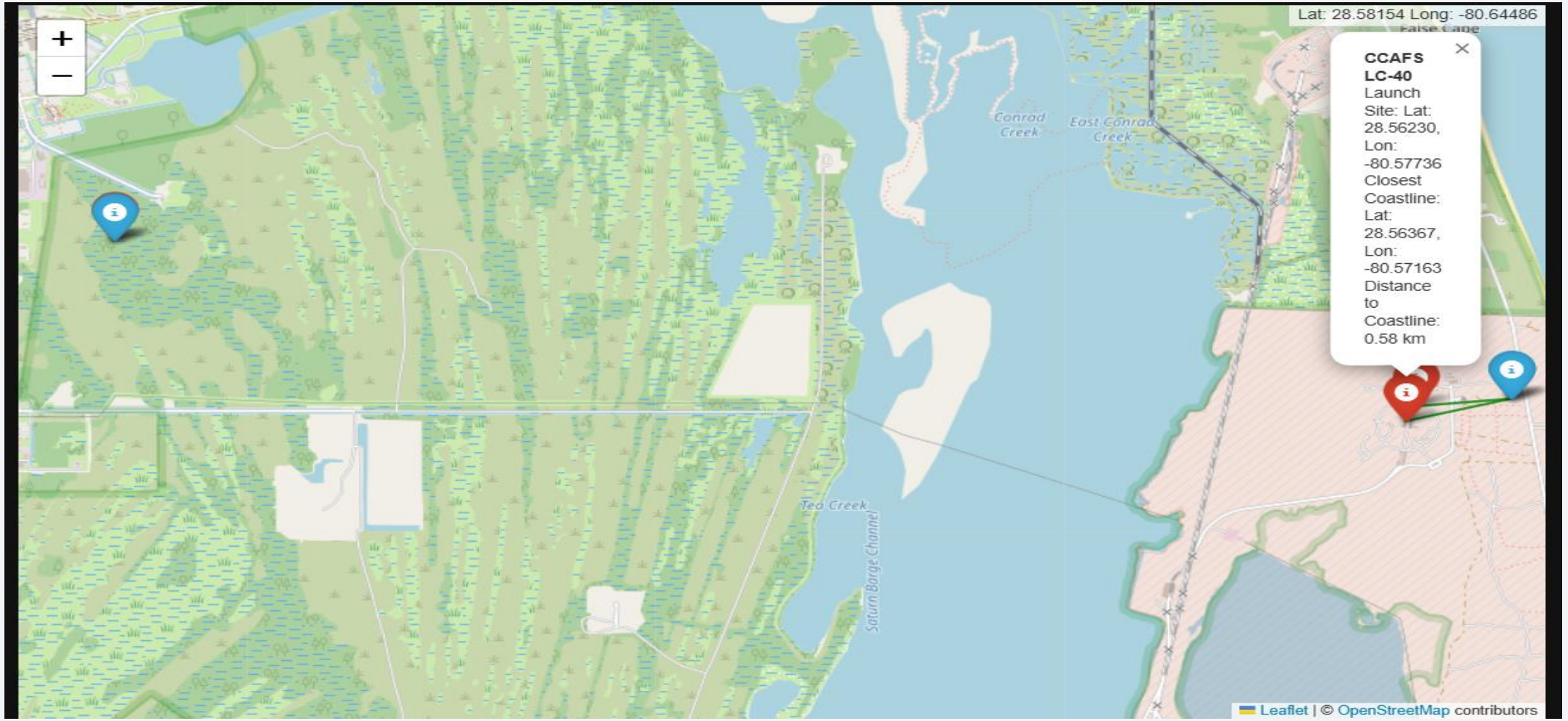
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



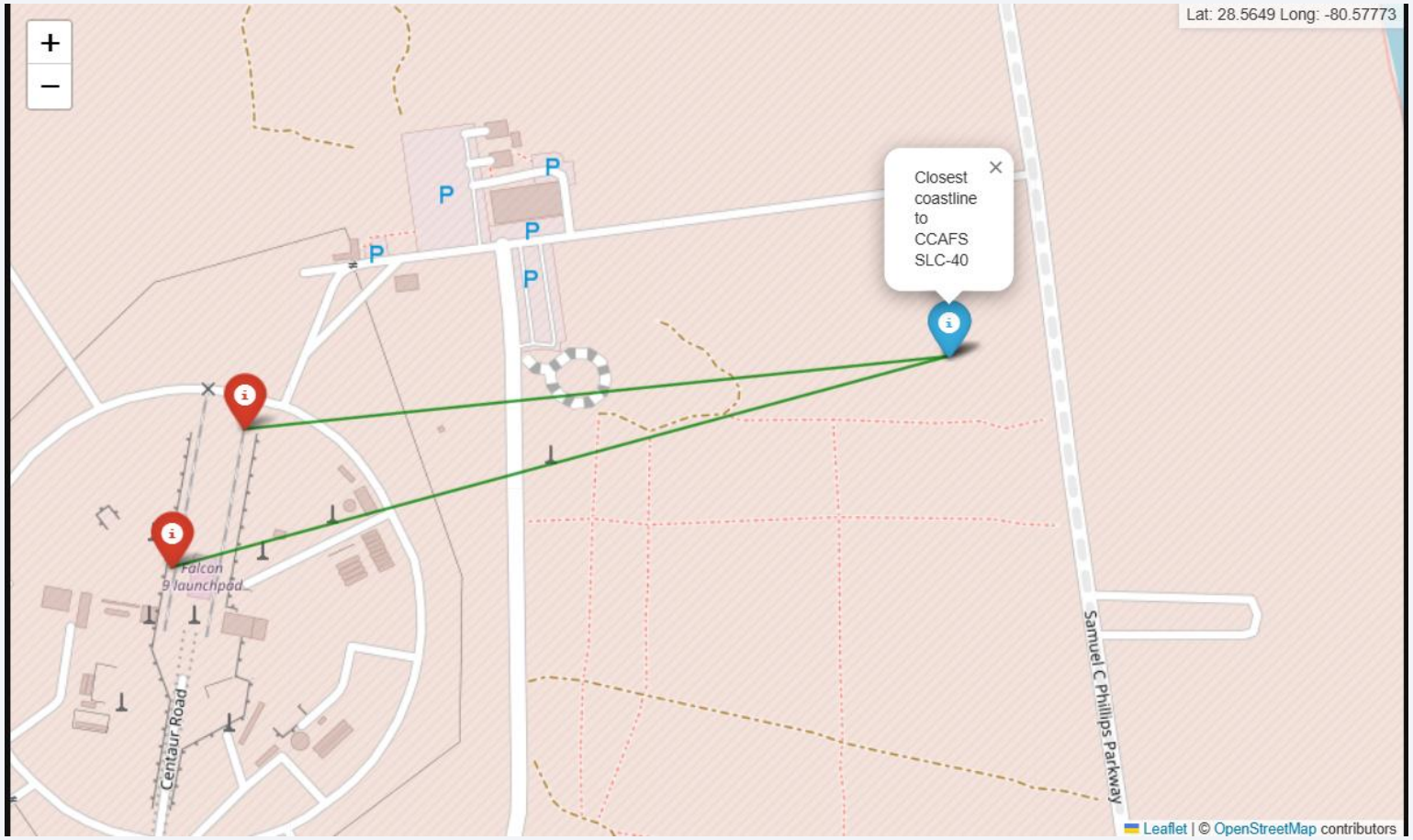
Global Map displaying SpaceX launching locations located in the United States

<Folium Map Screenshot 2>

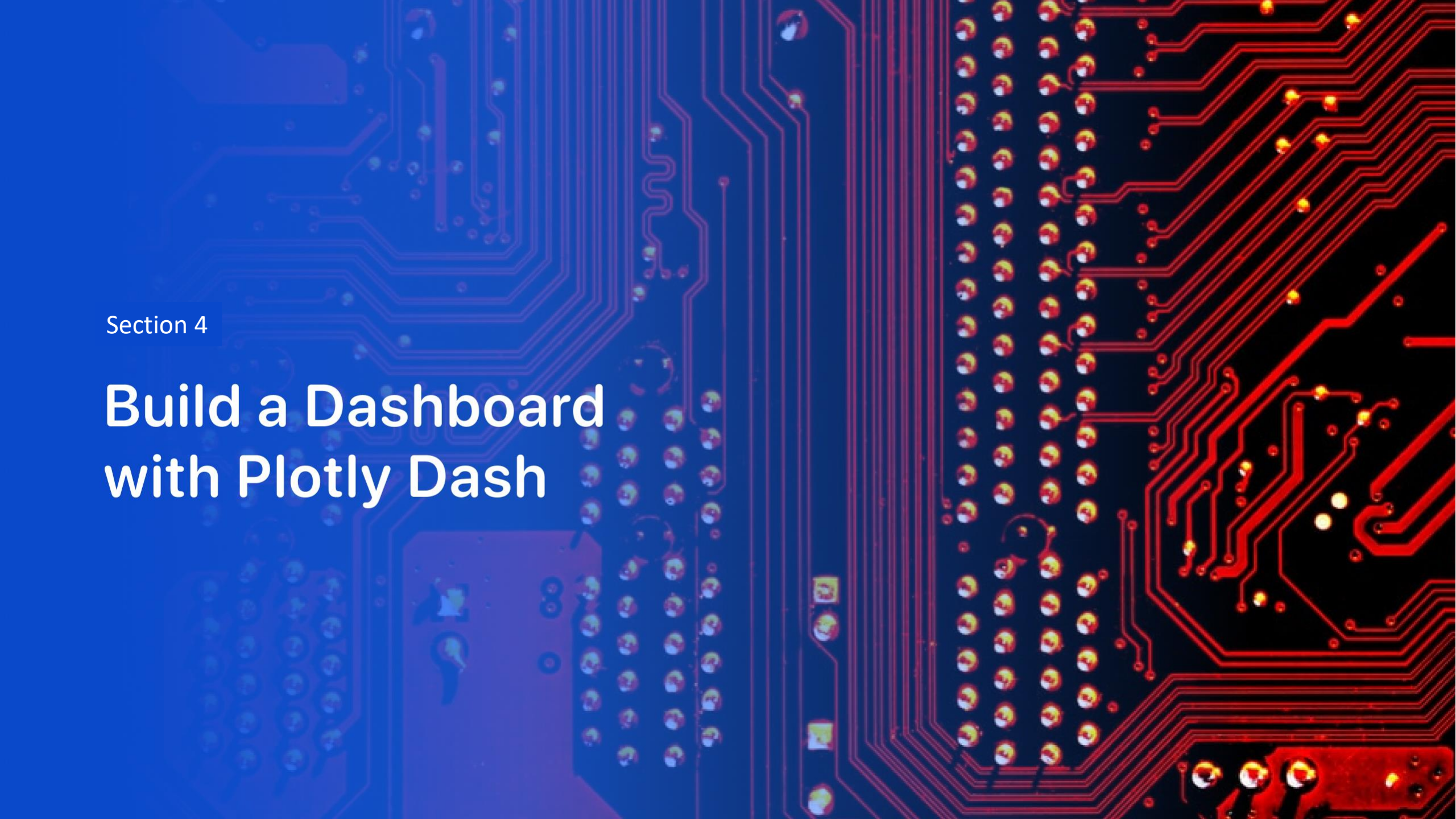


Map displaying red (failed) markers and blue (success) markers

<Folium Map Screenshot 3>



Detailed map displaying roads, railroads, etc. and nearest coastline


Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

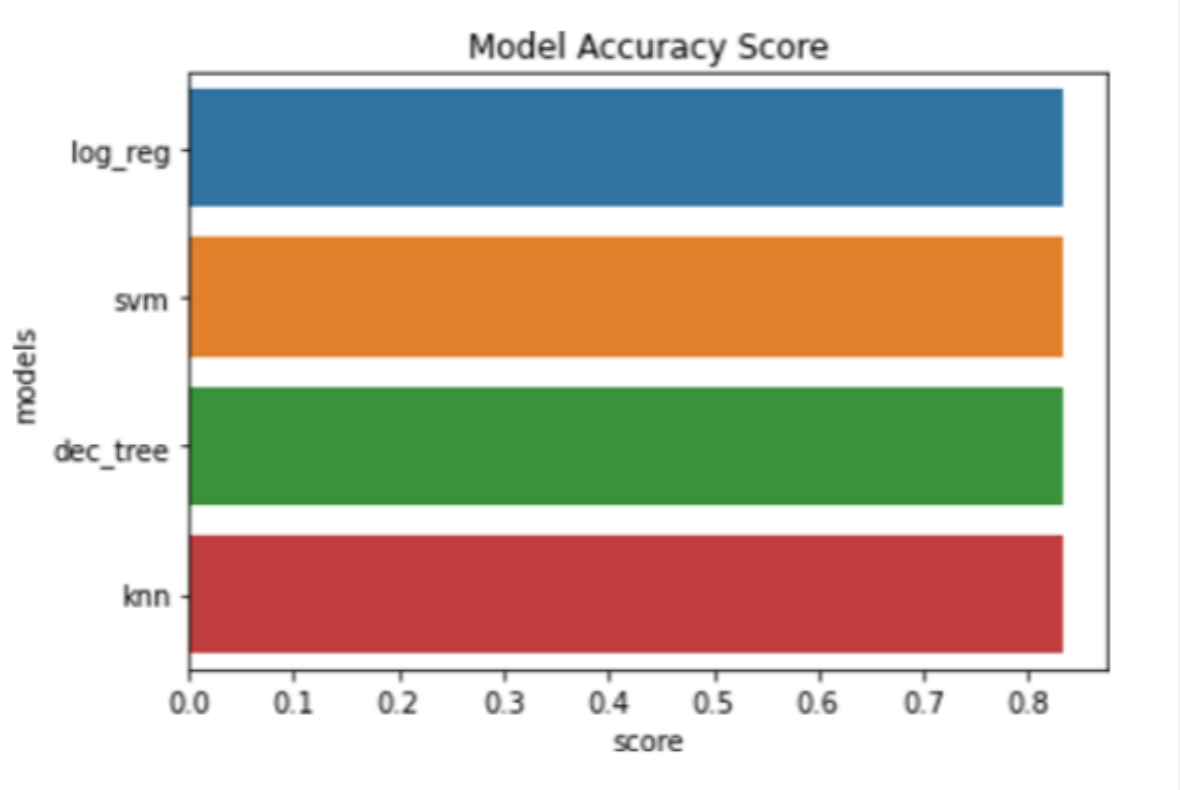
- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

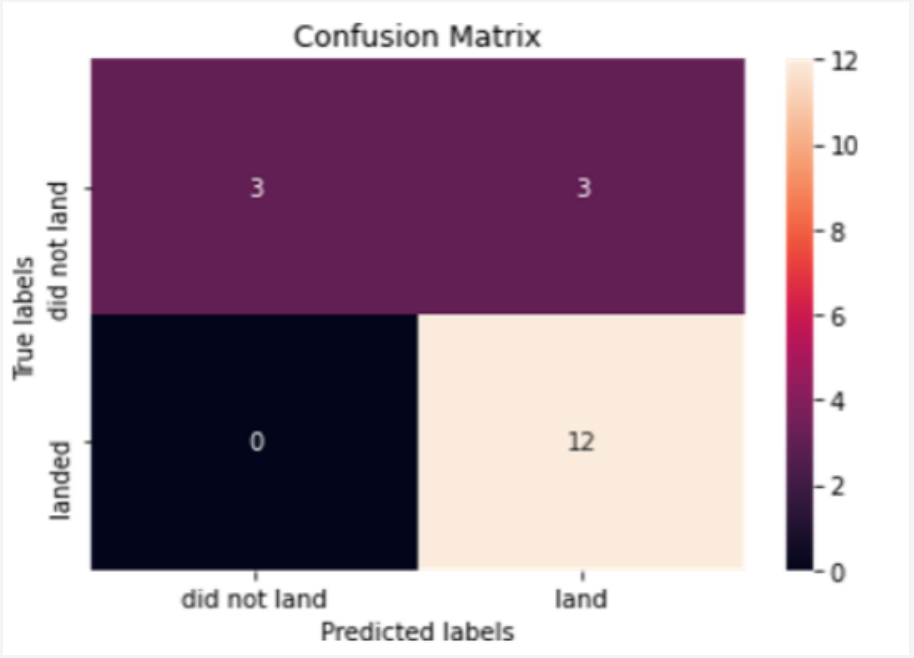
Classification Accuracy

- All four models displayed similar accuracy at 83.33% accuracy.
- A larger sample size is needed.



Confusion Matrix

- All models displayed similar accuracy.
- The small sample size (18) leads to over-prediction of success.



Conclusions

- Based on the analyses and queries we've performed, here are four key conclusions we can draw about SpaceX's launch operations:
- 1. High Success Rate and Rapid Evolution: SpaceX has demonstrated an impressive success rate in its missions, with 101 successful launches out of 102 total missions in our dataset.
- 2. Mastery of Reusability: The data shows a clear progression in SpaceX's reusability efforts. From early failures in drone ship landings to consistent successes with both drone ship and ground pad landings, SpaceX has mastered the art of recovering and reusing boosters. This is evidenced by boosters flying up to seven times (e.g., F9 B5 B1049.7)
- 3. Diverse Launch Capabilities: SpaceX has demonstrated the ability to handle a wide range of payload masses (from under 1000 kg to over 15,000 kg) and serve various orbits (LEO, GTO, ISS, etc.). The consistent performance of the Falcon 9 Block 5 across multiple launches and its ability to carry maximum payloads even after several reuses highlight SpaceX's versatility in meeting diverse customer needs.
- 4. Strategic Launch Site Utilization: The analysis of launch sites reveals SpaceX's strategic approach to space access. With primary sites on both the East Coast (Cape Canaveral, Kennedy Space Center) and West Coast (Vandenberg), SpaceX can efficiently reach a variety of orbits.

Thank you!