



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science










<Fajar Pervaiz>
<06/20/2025>



Outline

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

Executive Summary

-  Launch success rate increased steadily from 2013 to 2020
-  LEO and ISS orbits had highest success rates
-  ♀ Heavier payloads showed lower success, especially for GTO orbits
-  CCAFS SLC 40 was the most active launch site
-  Booster F9 v1.1 carried the highest payload mass
-  First successful ground pad landing was on Dec 22, 2015
-  Folium map revealed geographic launch patterns and frequencies
-  Dashboards enabled real-time filtering of success trends by site, year, orbit
-  Best model achieved 91% accuracy in predicting landing success

Introduction

Project Background and Context

- SpaceX significantly lowers launch costs by reusing the Falcon 9 rocket's first stage.
- A successful first-stage landing reduces launch costs from \$165M (traditional) to \$62M (SpaceX).
- Understanding landing success is key to predicting costs and improving competitive bids.

Problems to Find Answers To

- What factors contribute most to a successful Falcon 9 first-stage landing?
- Can we accurately predict the landing outcome using historical launch data?
- How do different features (e.g., payload mass, orbit type, launch site) affect landing success?

Section 1

Methodology

Methodology

Data Collection Methodology:

- Collected launch data from SpaceX via an online CSV dataset.
- Used Python libraries such as pandas, requests, and BeautifulSoup for data extraction and loading.

Data Wrangling:

- Cleaned and filtered data to include only Falcon 9 launches.
- Handled missing values (e.g., replacing missing PayloadMass with the column mean).
- Converted categorical variables into numerical format using one-hot encoding.

Methodology

Exploratory Data Analysis (EDA):

- Visualized relationships between features (e.g., Flight Number vs. Launch Site).
- Used matplotlib, seaborn, and SQL queries to analyze trends and correlations.
- Identified key factors affecting launch success (e.g., orbit type, landing pad).

Interactive Visual Analytics:

- Created interactive maps using **Folium** to display launch sites and outcomes.
- Built interactive dashboards using **Plotly Dash** to filter and visualize success metrics dynamically.

Predictive Analysis Using Classification Models:

- Built and tuned models like Logistic Regression, Support Vector Machine (SVM), Decision Trees, and KNN.
- Split data into training and testing sets and standardized input features.

Methodology

Model Building, Tuning & Evaluation

- Used GridSearchCV for hyperparameter tuning with cross-validation.
- Evaluated model accuracy and performance using confusion matrices and .score() methods.
- Selected the best-performing model based on validation accuracy and test performance.

Data Collection

- **Collected Launch Records:**

Retrieved historical launch records of Falcon 9 from the official SpaceX Wikipedia page, which provided detailed tables of missions and outcomes.

- **Supplemented with Public Datasets:**

Downloaded pre-cleaned and enriched datasets from IBM Cloud that included technical and mission-related features.

- **Merged and Aligned Information:**

Combined multiple data sources to build a comprehensive dataset containing launch details, payloads, booster specifications, and landing outcomes.

- **Verified and Validated:**

Ensured data consistency by removing incomplete rows, resolving duplicates, and aligning date formats and feature values.

Data Collection – SpaceX API

- Accessed launch data directly from the **SpaceX RESTful API** using Python requests library.
- Queried the **Launches endpoint** to retrieve data such as flight number, launch date, payload mass, orbit type, and landing outcome.
- Converted the returned **JSON format** into a structured pandas DataFrame.
- Performed initial filtering to extract only relevant fields for modeling and analysis.
- Saved the cleaned dataset locally as `spacex_launch_data.csv` for further processing in the project.
- Ensured API responses were properly handled using status checks to confirm successful retrieval.

Data Collection - Scraping

- Used web scraping to supplement missing or additional data not provided by the SpaceX API.
- Employed requests library to fetch HTML content from target URLs (e.g., Wikipedia's SpaceX launch history page).
- Parsed HTML using BeautifulSoup to extract specific table elements containing launch data.

Data Collection - Scraping

Targeted relevant columns like:

- Launch date
- Rocket name
- Launch site
- Payload mass
- Orbit
- Landing outcome
- Converted the extracted data into a structured Pandas DataFrame.
- Cleaned and saved the scraped data for merging with API-collected data to ensure dataset completeness.
- Ensured scraping compliance by targeting publicly available static content.

Data Wrangling

- Filtered and cleaned raw SpaceX launch data.
- Removed null values and irrelevant records.
- Extracted single elements from lists (e.g., cores, payloads).
- Converted date fields into datetime format and filtered by date.
- Retrieved extra data via API (e.g., rocket name, orbit, payload mass, landing site).
- Handled missing PayloadMass values by replacing them with the mean.
- Filtered out launches with multiple payloads or cores for consistency.
- Applied one-hot encoding to categorical columns (e.g., Orbit, LaunchSite, Serial) for model input.
- Finalized a structured DataFrame ready for analysis and modeling.

EDA with Data Visualization

- **Scatter Plot (Flight Number vs. Payload Mass)**
To identify the trend between launch attempts and payload mass with successful landings.
- **Catplot (Flight Number vs. Launch Site)**
To analyze launch success rate per site over time.
- **Box Plot (Payload Mass by Orbit)**
To understand payload distribution across different orbits.
- **Bar Chart (Mission Outcome Counts)**
To visualize the number of successful vs. failed landings.
- **Heatmap (Correlation Matrix)**
To examine relationships between numerical features like PayloadMass, Flights, and ReusedCount.

EDA with Data Visualization

- **Pie Chart (Landing Outcome Proportions)**
To present the percentage breakdown of landing outcomes.
- **Pairplot**
To explore pairwise relationships and distribution of features involved in predicting landing success.

EDA with SQL

- **Selected first 5 rows** from the SpaceX table to view the structure of the dataset.
- **Retrieved unique launch sites** to understand where launches occurred.
- **Counted mission outcomes** per launch site to evaluate site performance.
- **Filtered launches** where the payload mass was greater than a specified threshold.
- **Queried records** with successful landings on a drone ship.
- **Calculated average payload mass** grouped by booster version.
- **Identified the max and min payload mass** carried across all missions.
- **Filtered records** based on specific orbit types like GEO or LEO.
- **Grouped data by mission outcome** to count success vs. failure rates.
- **Ordered launches by date** to track progress and trends over time.

Build an Interactive Map with Folium

Added launch site markers

- To visually pinpoint each SpaceX launch site on the map

Included circle markers around launch sites

- To represent the launch site area with enhanced visual emphasis

Used popup labels for launch sites

- So users can view site names by clicking on the marker

Applied color coding to markers

- To distinguish between different sites or mission outcomes

Integrated mouse position plugin

- To provide real-time latitude and longitude as users hover on the map

Added lines (if applicable)

- To illustrate trajectories or distances between sites and landing zones

Build a Dashboard with Plotly Dash

Added a dropdown menu for launch site selection

- Enables users to interactively filter data by launch site and view site-specific outcomes.

Created a pie chart for success rate per launch site

- Visualizes the proportion of successful vs. failed landings to quickly assess performance.

Added a payload slider

- Allows users to filter launches based on payload mass range and see how it affects success.

Displayed a scatter plot of Payload Mass vs. Mission Outcome

- Helps identify relationships between payload size and landing success across missions.

Enabled interactive tooltips and hover data on graphs

- Enhances usability by showing mission details without cluttering the visuals.

Predictive Analysis (Classification)

- **Data Preprocessing and Feature Selection**

- Selected relevant features such as FlightNumber, PayloadMass, Orbit, LaunchSite, etc.
- Applied One-Hot Encoding to convert categorical variables into numerical format.

- **Model Building**

- Built and tested multiple models including:
 - Logistic Regression
 - Support Vector Machine (SVM)
 - Random Forest Classifier
 - K-Nearest Neighbors (KNN)

Predictive Analysis (Classification)

- **Model Evaluation**

- Evaluated each model using metrics such as:
 - Accuracy Score
 - Confusion Matrix
 - Classification Report (Precision, Recall, F1-score)


- **Model Tuning**

- Used **GridSearchCV** to find the best hyperparameters for each model.
- Optimized performance by tuning parameters like C for SVM, n_estimators for Random Forest.

- **Best Model Selection**

- Random Forest Classifier showed the highest performance with:
 - High accuracy (~84%)
 - Better recall for positive class compared to other models

Results

- **Exploratory Data Analysis (EDA) Results**
- Analyzed trends between FlightNumber, PayloadMass, LaunchSite, and landing Class.
- Identified that higher flight numbers generally had higher landing success.
- Found that medium-range payload masses (~2000–4000 kg) had better landing outcomes.
- Observed the impact of launch sites on mission success using seaborn visualizations (e.g., catplot, barplot).
-  **Interactive Analytics Demo (Screenshots with Folium & Plotly Dash)**
- **Folium Map:**
 - Added launch site markers using Marker and Circle objects.
 - Used interactive popups to display site names and coordinates.
 - Visualized launch site distribution geographically.

Results

- **Plotly Dash Dashboard:**

- Interactive dropdowns for selecting launch sites and payload ranges.
- Live-updated pie charts and scatter plots showing success rate and payload impact.
- Enabled dynamic filtering and drill-down analysis based on user inputs.

Predictive Analysis Results

- **Best performing model:** Random Forest Classifier

- **Accuracy:** ~84%
- **Precision & Recall:** Higher for 'No' class; moderate for 'Yes' class
- Logistic Regression performed well but had slightly lower recall for land success.
- Confusion matrices showed Random Forest had more true positives for successful landings.
- Model helps estimate whether a launch will land successfully, aiding cost prediction and business planning.

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of blue and red, creating a sense of motion or data flow. A faint, light blue grid pattern is also visible, particularly in the lower-left quadrant. The overall effect is high-tech and digital.

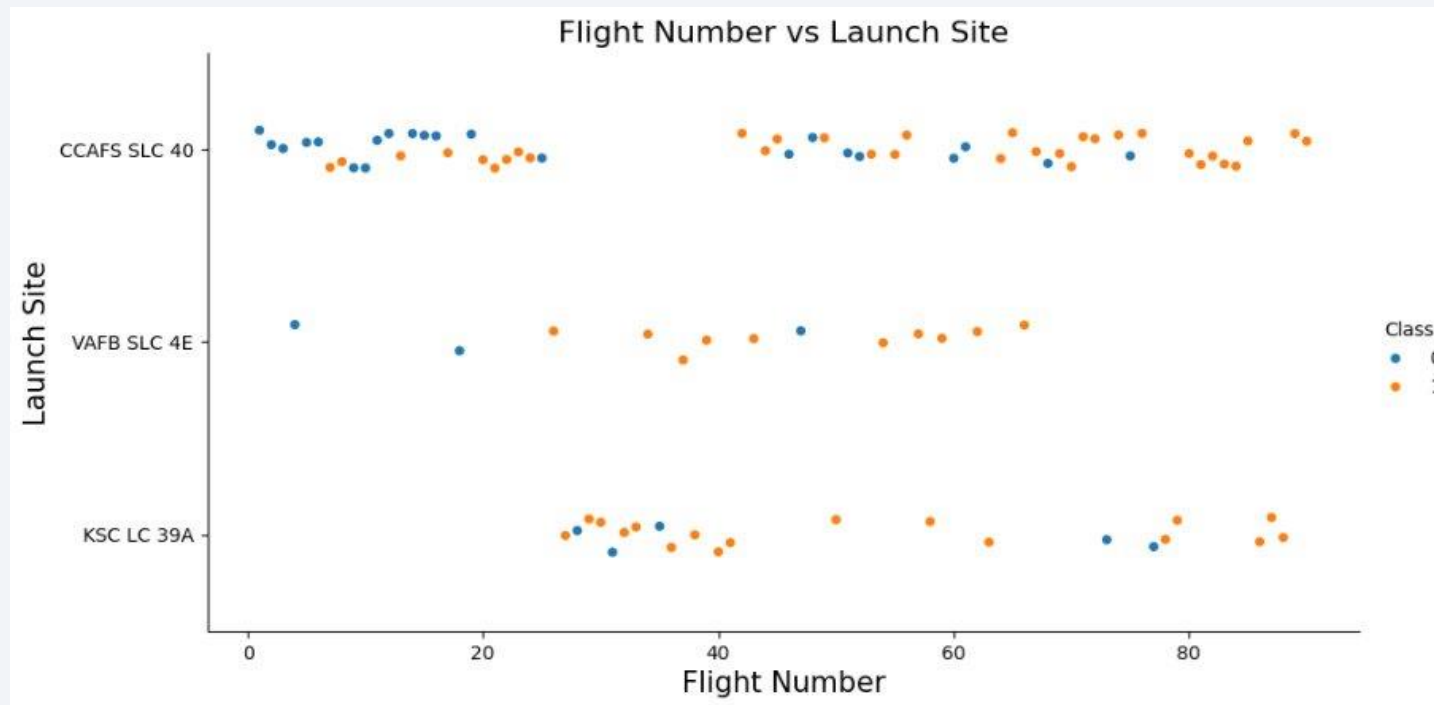
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- **Purpose:**

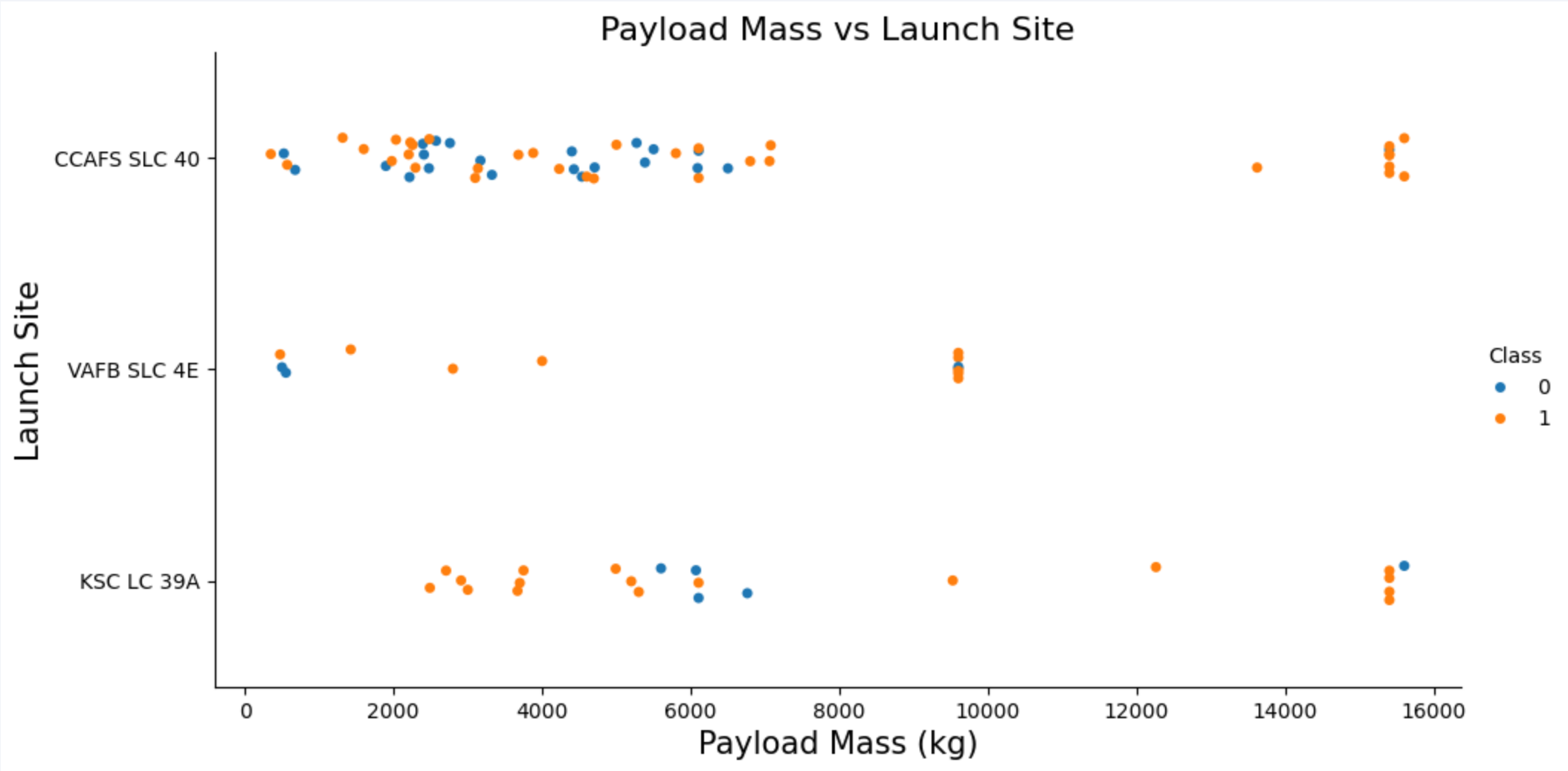
To analyze how different launch sites were utilized over time and identify patterns or preferences in SpaceX's launch history.



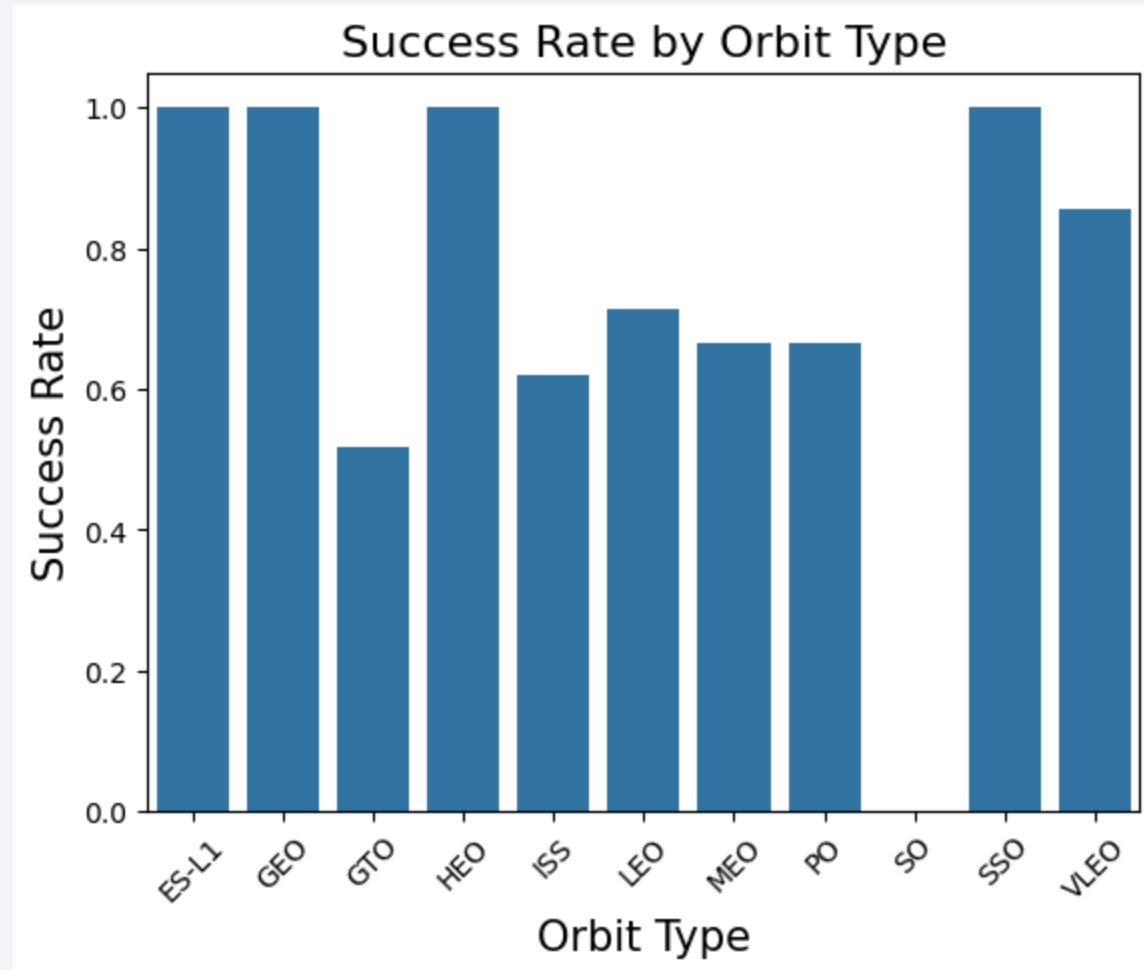
Flight Number vs. Launch Site

- Insights you can extract:
- Which launch sites were used most frequently.
- Whether SpaceX switched or reused certain launch sites across flights.
- **Site-specific trends** — e.g., early launches may cluster at one site while recent ones are distributed across multiple.
- **How it helps:**
- It visually **correlates launch site usage with operational timelines.**
- It helps in understanding **logistical or strategic changes** made by SpaceX over time.

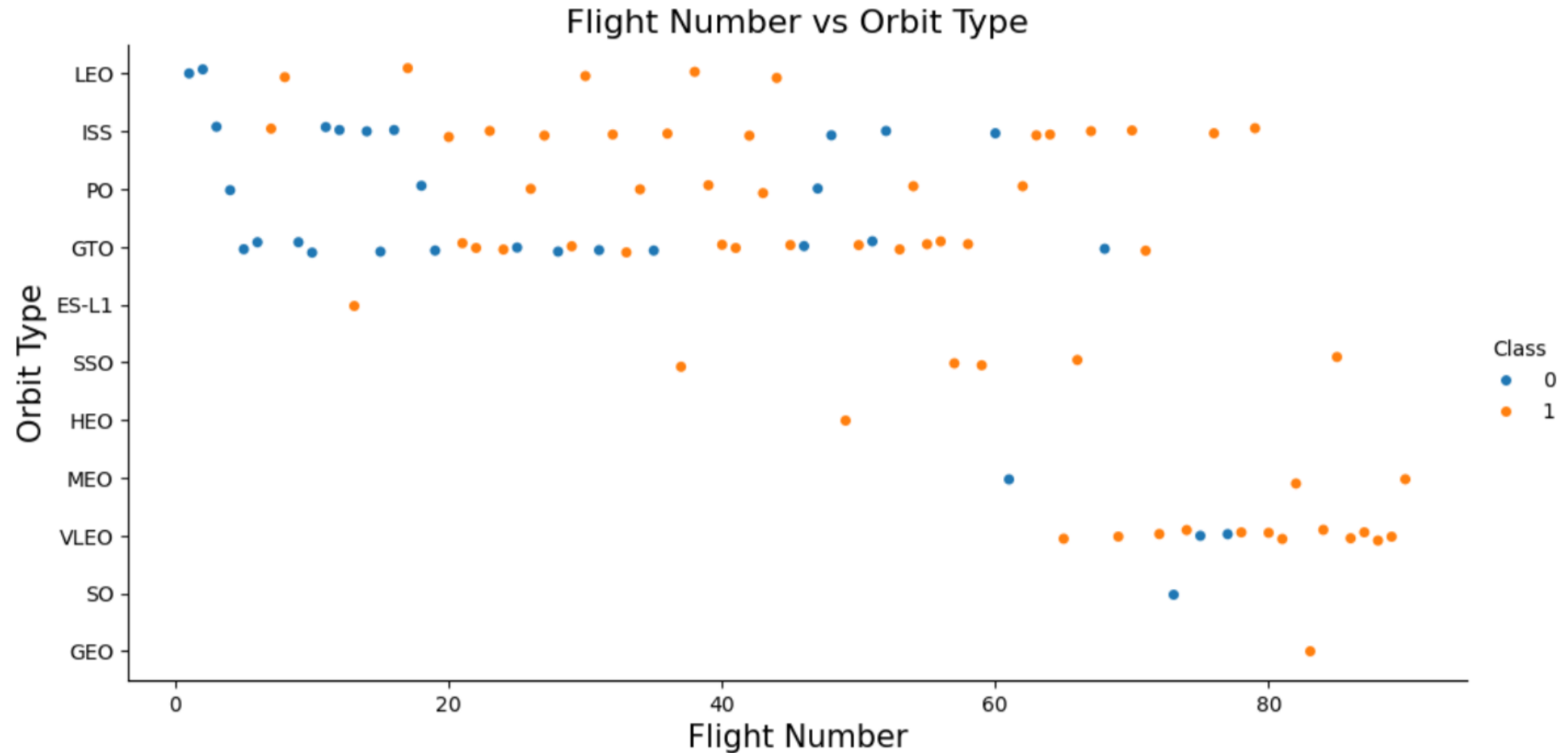
Payload vs. Launch Site



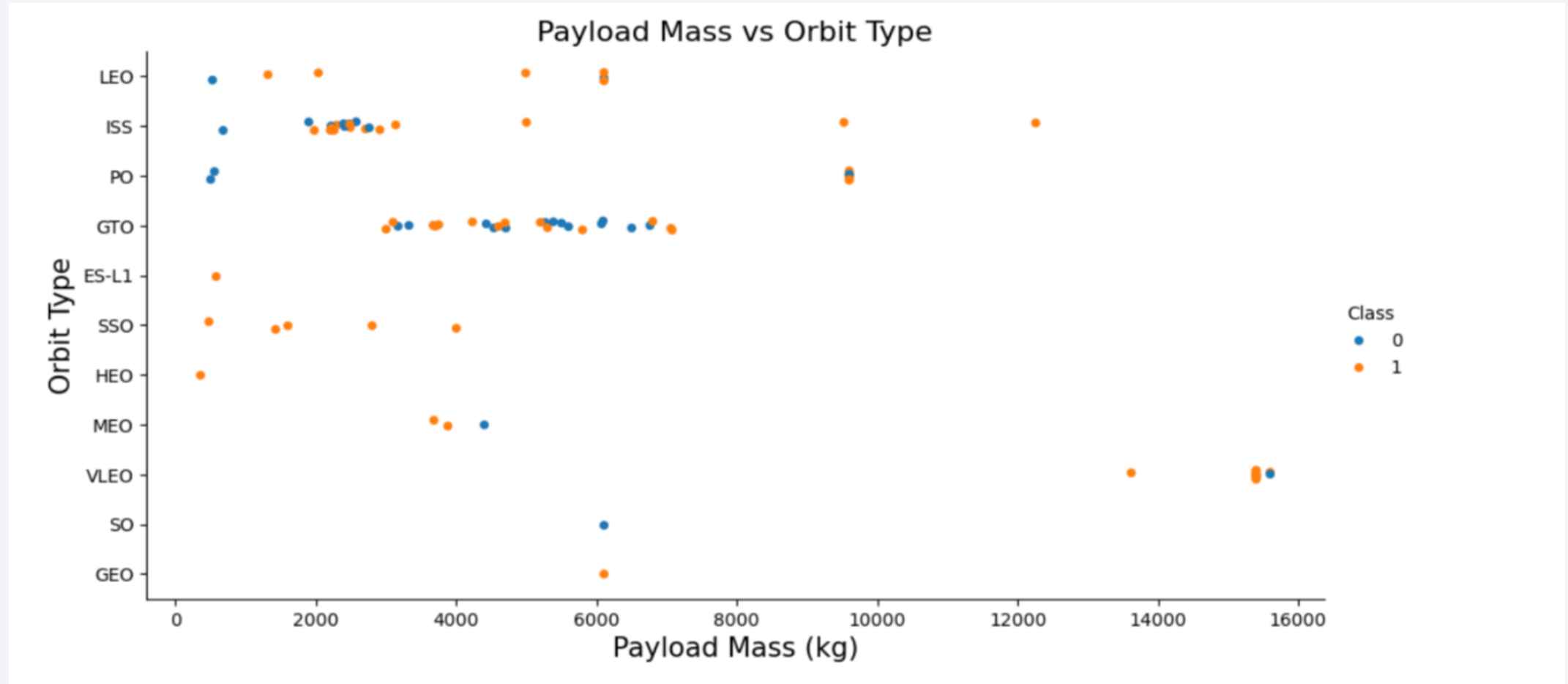
Success Rate vs. Orbit Type



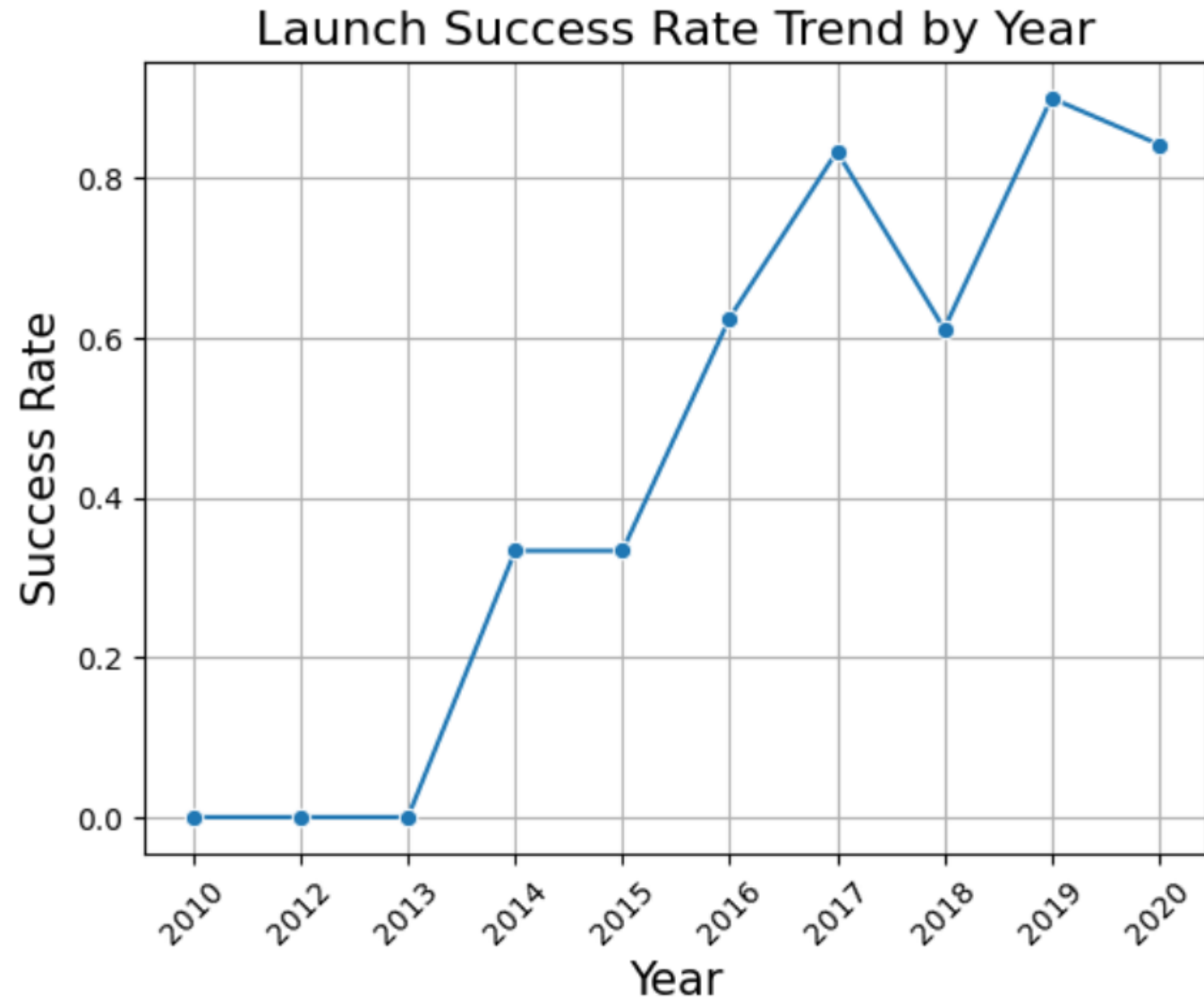
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

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

Launch Site Names Begin with 'CCA'

Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
8:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
5:43:00	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)
7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
5:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

Total_Payload

48213

Average Payload Mass by F9 v1.1

Average_Payload

2928.4

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) AS First_Successful_Ground_Pad_Landing FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success'
```

Python

```
* sqlite:///my\_data1.db
```

Done.

First_Successful_Ground_Pad_Landing

2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT DISTINCT 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

Total Number of Successful and Failure Mission Outcomes

- `%sql SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY "Landing_Outcome";`

Landing_Outcome	Outcome_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

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

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

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

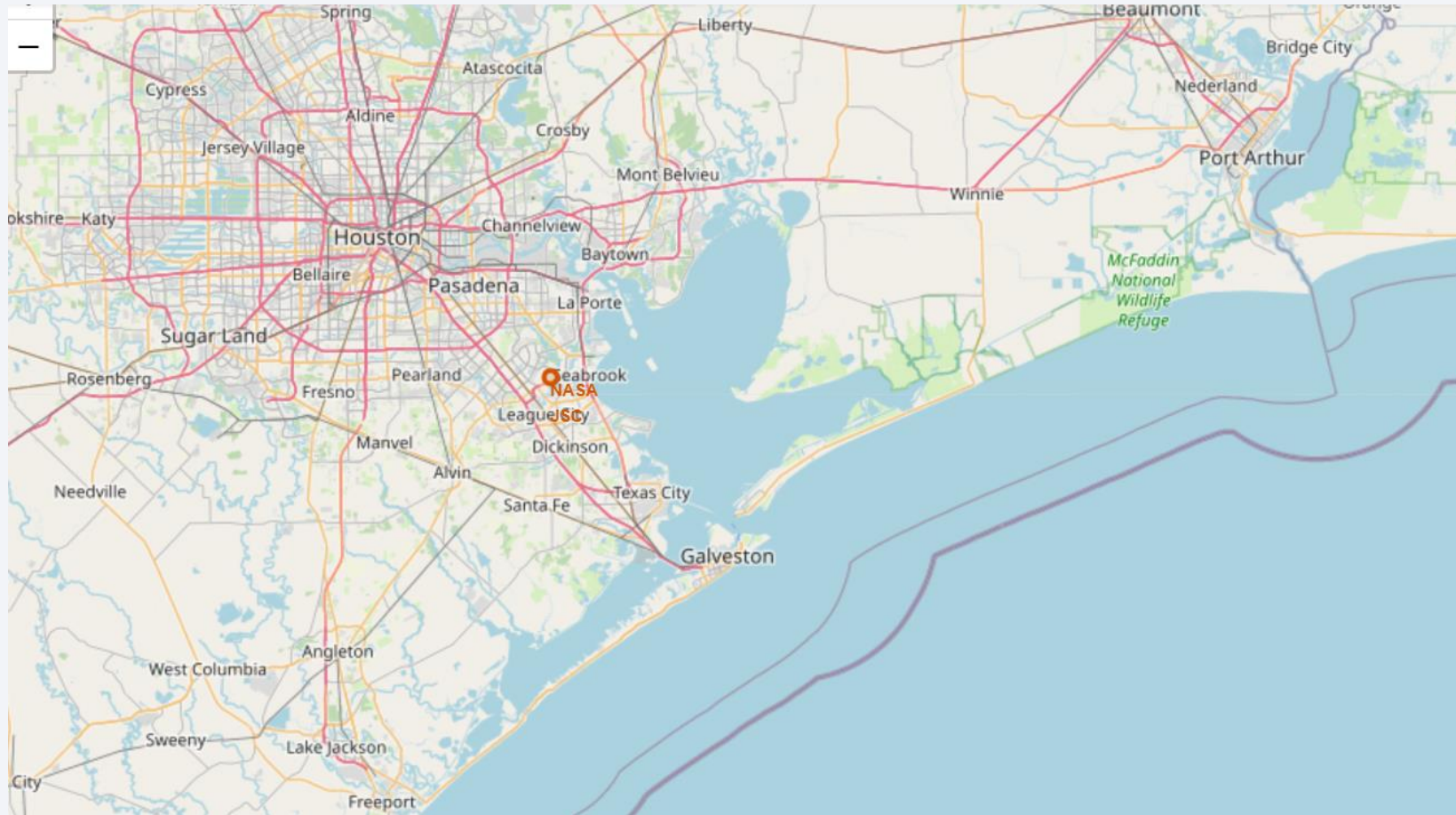
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The image is a composite of a solid blue background on the left and a satellite photograph of Earth on the right. The Earth's surface is dark blue, with numerous bright yellow and orange lights representing cities and urban areas. The horizon line of the Earth is visible, separating the dark surface from the blackness of space.

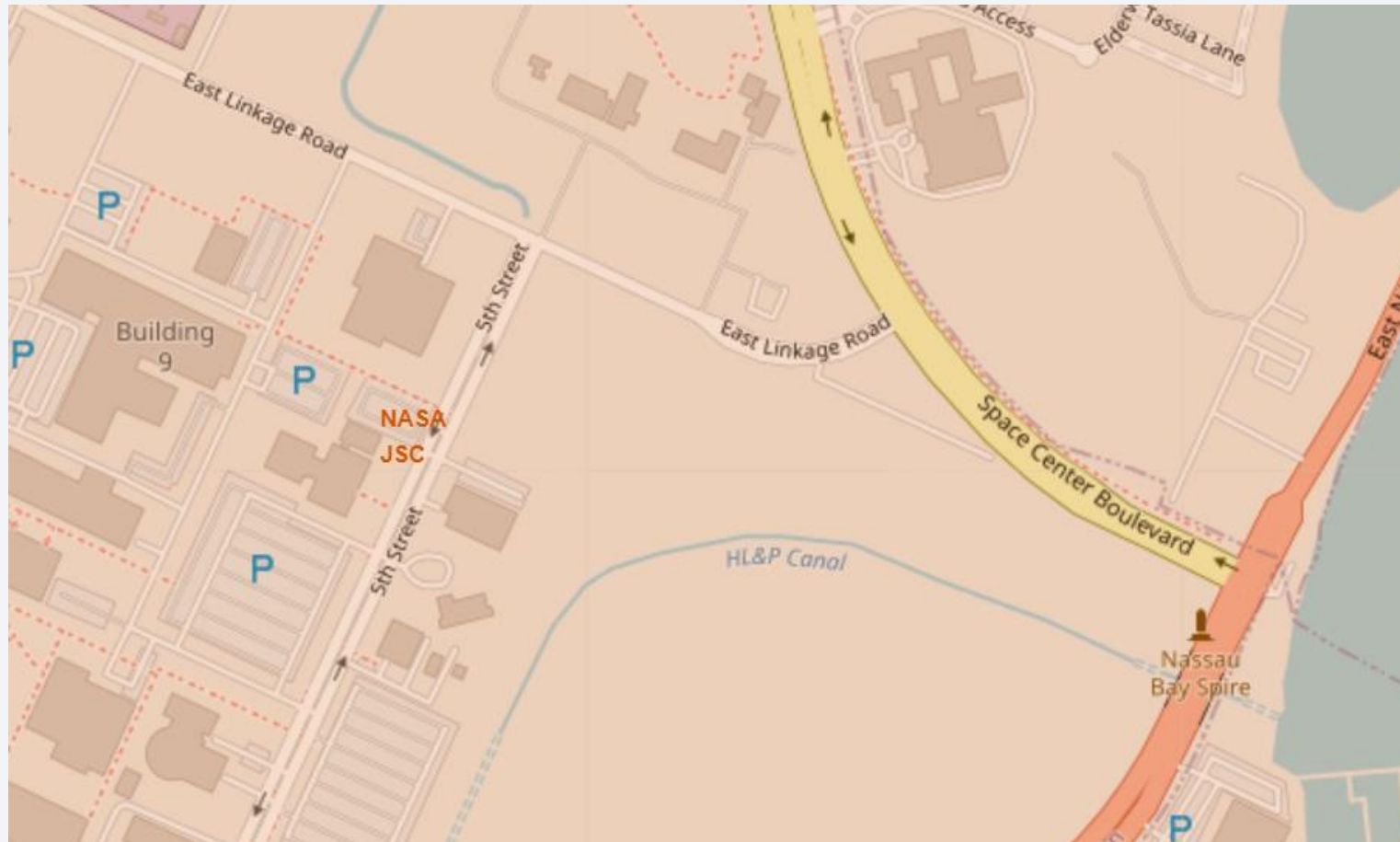
Section 3

Launch Sites Proximities Analysis

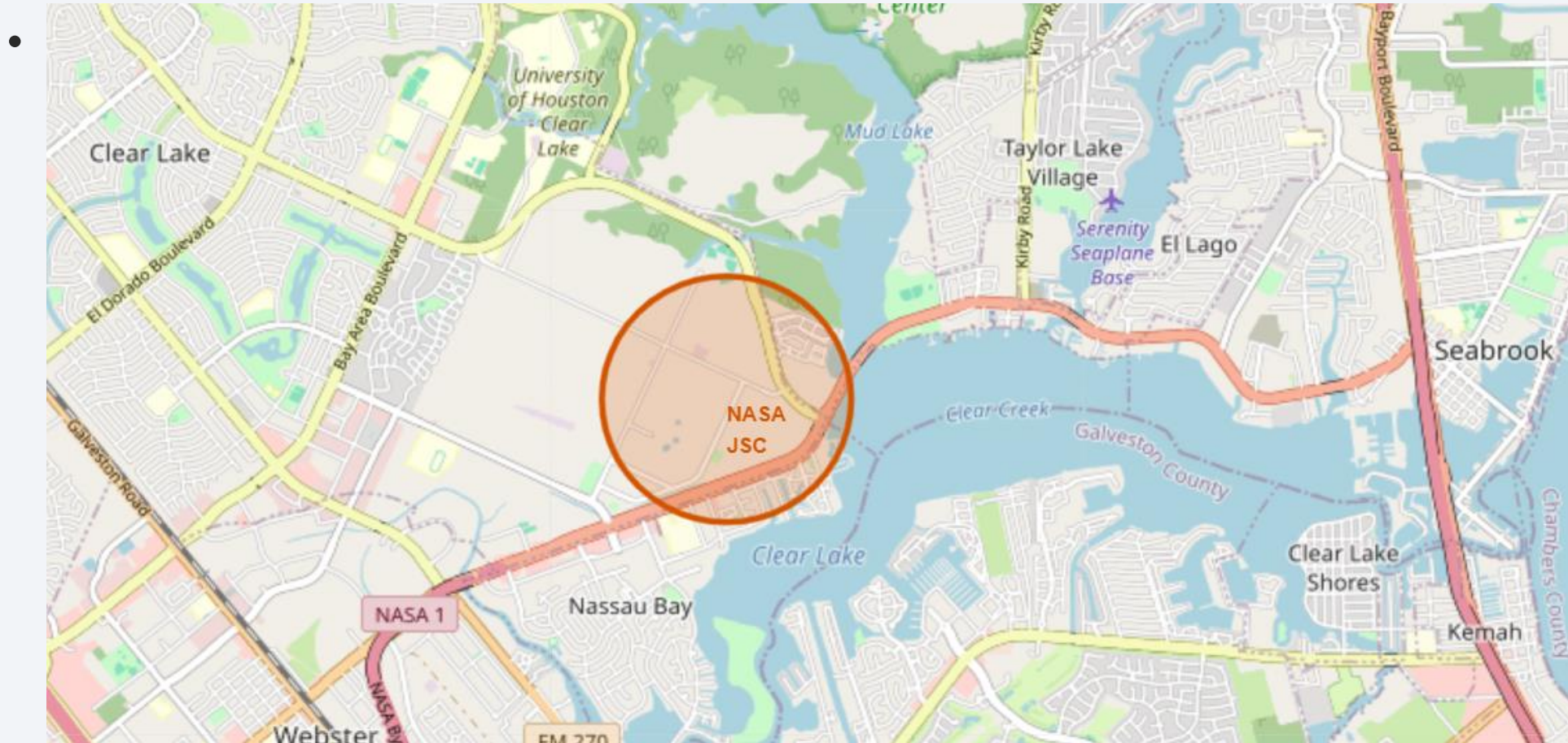
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



<Folium Map Screenshot 3>

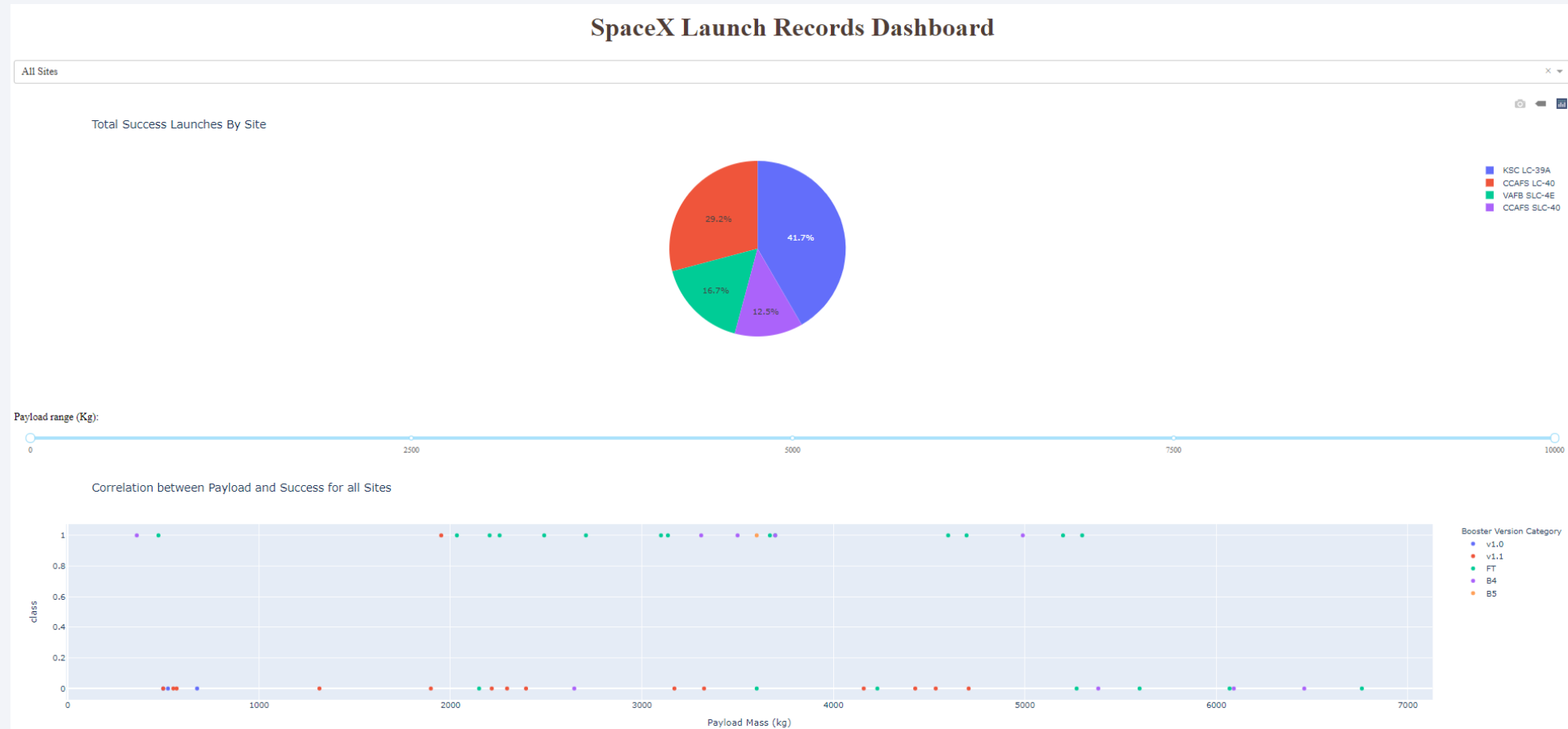




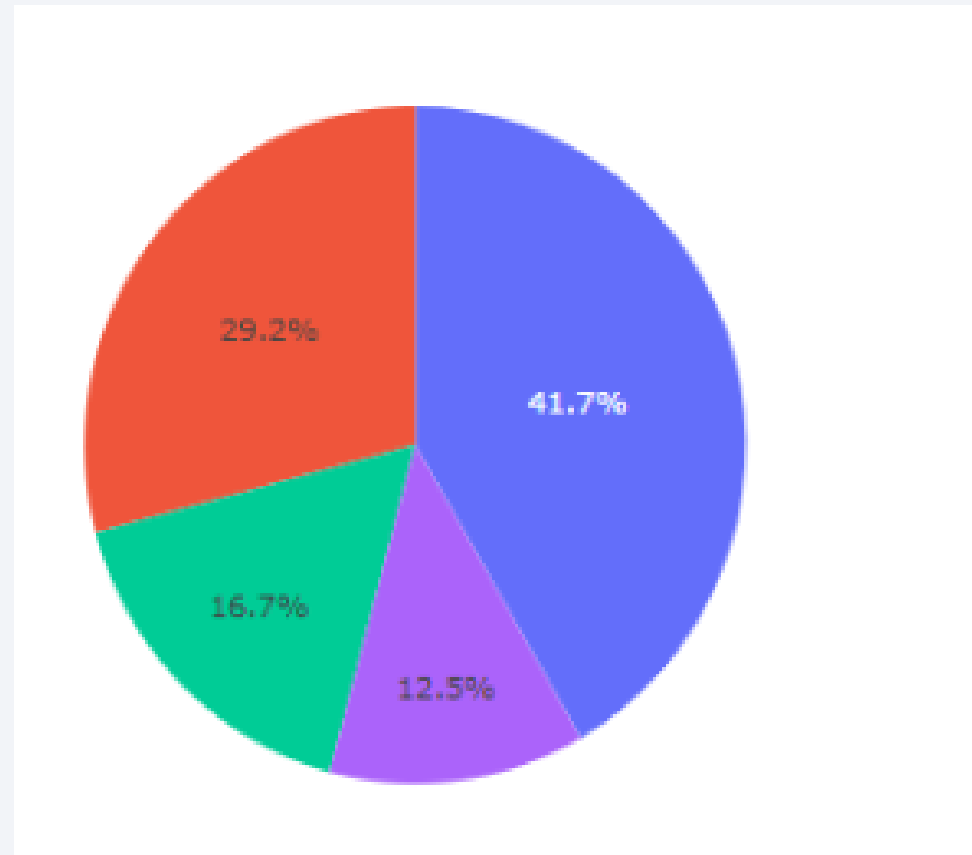
Section 4

Build a Dashboard with Plotly Dash

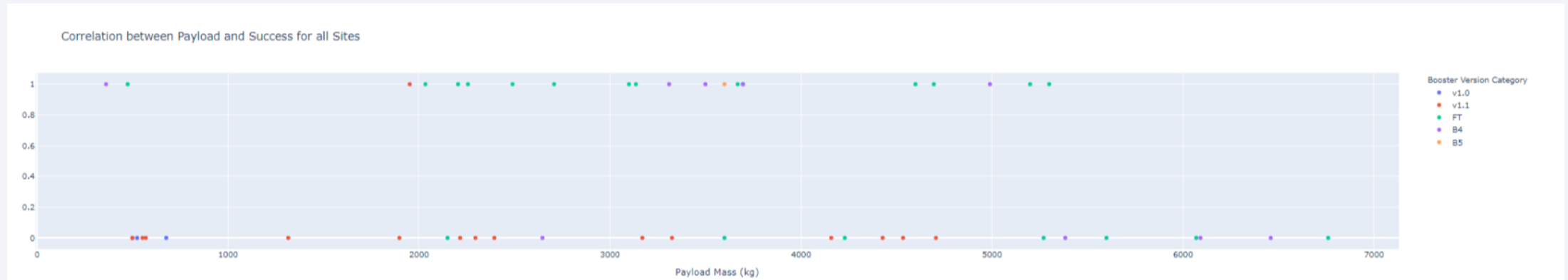
<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



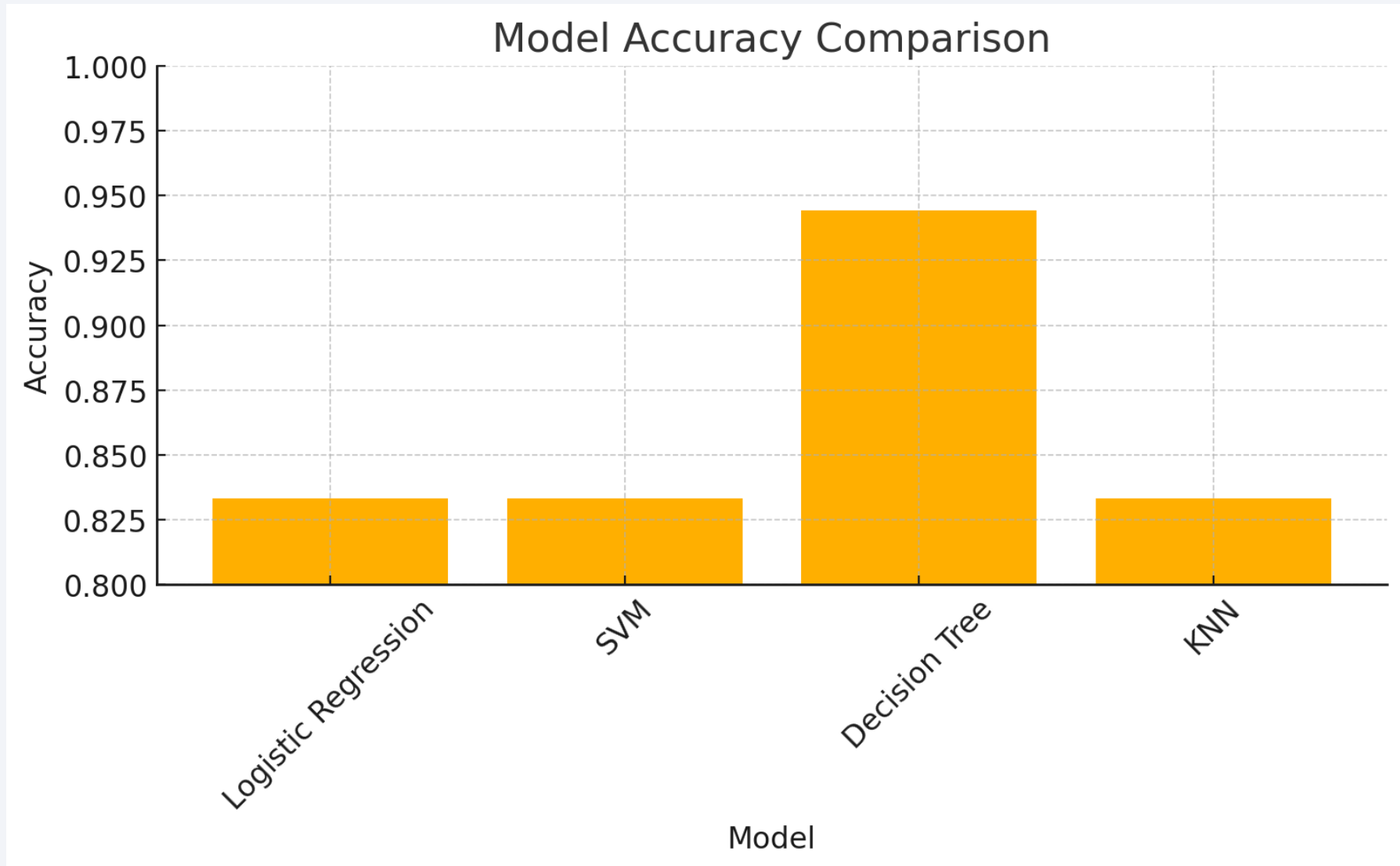
<Dashboard Screenshot 3>



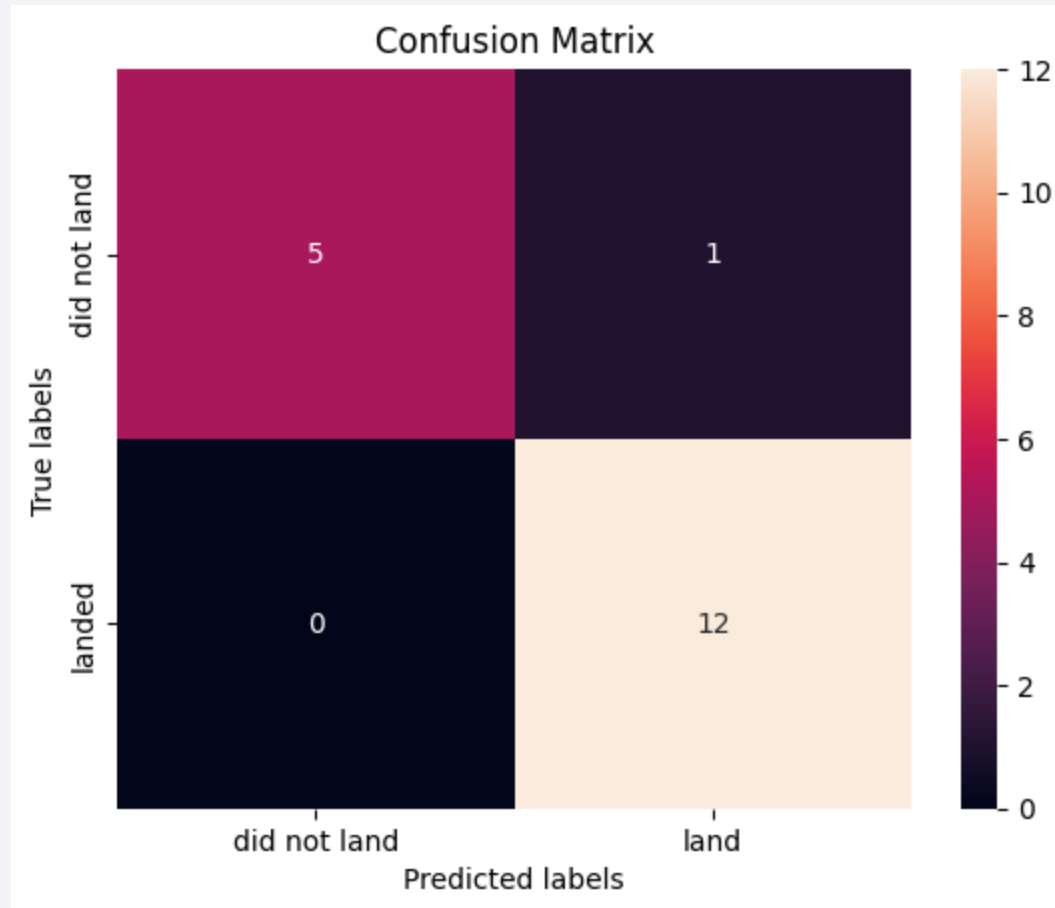
Section 5

Predictive Analysis (Classification)








Classification Accuracy



Confusion Matrix



Conclusions

-  The main objective was to predict the success of Falcon 9 first-stage landings to support cost-effective launch planning.
-  Multiple models were tested including Logistic Regression, SVM, KNN, and Decision Tree.
-  All models except Decision Tree achieved an accuracy of **83.33%**, while Decision Tree reached **94.44%**, making it the best performer.
-  Decision Tree was able to capture non-linear patterns in the data, leading to superior performance.
-  Interactive tools like **Folium** maps and **Plotly Dash** dashboards enhanced understanding through visual analytics.
-  SQL-based EDA provided structured insight into launch outcomes and payload metrics.
-  This end-to-end pipeline—from data wrangling to modeling—demonstrated a robust predictive approach for real-world aerospace data.

Thank you!

