IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summar

- ## Summary of methodologies
  - This project involved a comprehensive data science lifecycle applied to SpaceX Falcon 9 launch data. Key methodologies included:
  - **Data Collection:** Utilizing the SpaceX REST API and web scraping to gather raw launch data.
  - **Data Wrangling:** Processing and cleaning the collected data to handle missing values and prepare it for analysis.
  - **Exploratory Data Analysis (EDA):**
    - **Visualization:** Using Python libraries (Matplotlib, Seaborn) to uncover patterns and relationships.
    - **SQL Queries:** Performing targeted queries on the dataset to extract specific insights.
  - **Interactive Visual Analytics:** Building interactive maps with Folium and a dashboard with Plotly Dash for dynamic data exploration.
  - **Predictive Analysis (Classification):** Developing, tuning, and evaluating various classification models (Logistic Regression, SVM, Decision Tree, KNN) to predict launch success.

- ## Summary of all results
  - Identified key launch sites and their launch frequencies.
  - Visualized trends in launch success over time and across different orbits and payload masses.
  - Successfully performed data wrangling to clean and prepare the dataset.
  - Built and evaluated multiple classification models, achieving a test accuracy of approximately 83% across Logistic Regression, SVM, Decision Tree, and KNN models, with Logistic Regression identified as a strong performer based on validation metrics.
  - Gained insights into factors influencing launch success, such as payload mass and orbit type.

Introduction

- Project background and context

- SpaceX, a private aerospace manufacturer, has revolutionized space travel through its Falcon 9 rocket, known for its reusability. The ability to reuse the first stage significantly reduces launch costs, making SpaceX highly competitive. Understanding the factors that determine whether the first stage will successfully land is crucial for predicting launch costs and for potential competitors bidding against SpaceX.

- Problems you want to find answers

- This project aims to answer key questions, including:
- What are the primary factors influencing the success of a Falcon 9 first-stage landing?
- How do launch sites, orbit types, and payload masses correlate with launch outcomes?
- Can we build a predictive model to accurately forecast the success of a Falcon 9 landing?
- What insights can be derived from the historical launch data to inform future missions or competitive strategies?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- **Data Collection Methodology**
- Data was collected from two primary sources:
- **SpaceX REST API:** Programmatic access to detailed historical launch data.
- **Web Scraping:** Extracting additional information from web pages, likely related to mission outcomes or specific booster details not readily available via API.
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- Process: Utilized Python to make GET requests to the SpaceX API endpoints
- (e.g., /launches, /rockets, /payloads).Key Phrases: API calls, JSON parsing, data extraction, pagination handling.Flowchart of SpaceX API calls:
- graph TD   A[Start] --> B{Fetch Launches API};
- B --> C{Parse JSON Response};
- C --> D{Extract Relevant Data};
- D --> E{Fetch Rocket Details for each Launch};   E --> F{Fetch Payload Details for each Launch};   F --> G{Combine Data};   G --> H{Save to Raw DataFrame};   H --> I[End];GitHub URL of the completed SpaceX API calls notebook: https://github.com/guylamontagne/capstone.git

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- Process: Employed web scraping techniques (e.g., using BeautifulSoup, Requests) to gather supplementary data from web pages, potentially for landing outcomes or specific booster details not readily available via API.Key Phrases: HTML parsing, CSS selectors, data extraction, error handling.Flowchart of Web Scraping:
- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

Place your flowchart of web scraping here

Data Wrangling

Process: Data wrangling involved cleaning, transforming, and preparing the raw data for analysis. Key steps included:

•Handling missing values: Identified LandingPad column with approximately 28.89% missing values.

•Converting data types: Ensuring columns like 'Date' are in datetime format.

•Creating new features: Potentially deriving new features from existing ones (e.g., 'Year' from 'Date').

•Encoding categorical variables: Converting categorical features into numerical representations suitable for machine learning models.

EDA with Data Visualization

•**Charts Plotted and Why:**
•**Scatter Plots (Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, Payload vs. Orbit Type):** Used to visualize the distribution of launches across different sites, and to observe potential relationships between payload mass, flight number, and launch sites/orbit types. These help identify initial patterns or clusters.
•**Bar Charts (Success Rate of Each Orbit Type):** To compare the success rates across different orbital destinations, highlighting which orbits are more challenging or successful.
•**Line Chart (Yearly Average Success Rate):** To show the trend of launch success over time, indicating improvements or changes in SpaceX's capabilities.
•**GitHub URL of your completed EDA with data visualization notebook:**
https://github.com/guylamontagne/capstone.git

EDA with SQL

- **SQL Queries Performed:**
- **Display unique launch sites:** SELECT DISTINCT LaunchSite FROM SPACEXTABLE; (Result: CCAFS SLC 40, VAFB SLC 4E, KSC LC 39A)
- **Display 5 records where launch sites begin with 'CCA':** SELECT * FROM SPACEXTABLE WHERE LaunchSite LIKE 'CCA%' LIMIT 5;
- **Calculate total payload mass by NASA (CRS) boosters:** SELECT SUM(PayloadMass) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
- **Calculate average payload mass by booster version F9 v1.1:** SELECT AVG(PayloadMass) FROM SPACEXTABLE WHERE BoosterVersion = 'F9 v1.1';
- **List date of first successful ground pad landing:** SELECT MIN(Date) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'True RTLS';
- **List boosters with successful drone ship landing and payload between 4000 and 6000:** SELECT BoosterVersion FROM SPACEXTABLE WHERE "Landing_Outcome" = 'True ASDS' AND PayloadMass BETWEEN 4000 AND 6000;
- **List total number of successful and failure mission outcomes:** SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Mission_Outcome";
- **List booster_versions with maximum payload mass:** SELECT BoosterVersion FROM SPACEXTABLE WHERE PayloadMass = (SELECT MAX(PayloadMass) FROM SPACEXTABLE);
- **List month names, failure landing outcomes:** SELECT STRFTIME('%Y-%m', Date) AS Month, "Landing_Outcome", COUNT(*) FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Failure%' GROUP BY Month, "Landing_Outcome";
- **GitHub URL of your completed EDA with SQL notebook:** https://github.com/guylamontagne/capstone.git

Build an Interactive Map with Folium

Build an Interactive Map with Folium

•**Map Objects Created and Added:**

- •**Markers:** Added markers for each launch site, showing their coordinates.
- •**Circles:** Drew circles around launch sites, perhaps colored by success rate or number of launches.
- •**Lines:** Potentially drew lines connecting launch sites to landing pads or showing flight paths.

•**Explanation:** These objects provide a geographical context to the launch data, allowing for visual identification of successful and unsuccessful landings relative to launch sites and landing zones. This helps in understanding spatial patterns.

•**GitHub URL of your completed interactive map with Folium map:** https://github.com/guylamontagne/capstone.git

14. Build a Dashboard with Plotly Dash

Build a Dashboard with Plotly Dash

- **Plots/Graphs and Interactions Added:**
- **Pie Chart (Launch Success Count for All Sites):** To visualize the overall success distribution across all launch sites.
- **Pie Chart (Launch Site with Highest Success Ratio):** To highlight the most successful launch site.
- **Scatter Plot (Payload vs. Launch Outcome):** Interactive scatter plot allowing users to filter by payload range using a slider, and potentially by booster version, to analyze the impact of payload on success.
- **Explanation:** The dashboard provides an interactive platform for users to explore the data dynamically, enabling deeper insights into the relationships between various parameters (e.g., payload, launch site) and launch outcomes without requiring coding knowledge.
- **GitHub URL of your completed Plotly Dash lab:** https://github.com/guylamontagne/capstone.git

Predictive Analysis (Classification)

•**Model Development Process:**
•**Data Preparation:** Standardized features (X) and created target variable (Y - Class column).
•**Train-Test Split:** Divided data into training (80%) and testing (20%) sets.
•**Model Selection:** Explored four classification algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
•**Hyperparameter Tuning (GridSearchCV):** Used GridSearchCV with 10-fold cross-validation to find the optimal hyperparameters for each model.
•**Evaluation:** Assessed models based on validation scores (from GridSearchCV) and test accuracy, as well as confusion matrices.
•**Flowchart of Model Development:**
graph TD A[Cleaned Data] --> B{Standardize Features}; B --> C{Split Data (Train/Test)}; C --> D{Initialize Models (LR, SVM, DT, KNN)}; D --> E{Hyperparameter Tuning (GridSearchCV)}; E --> F{Train Best Model}; F --> G{Evaluate on Test Data}; G --> H[Results & Confusion Matrix];

•**GitHub URL of your completed predictive analysis lab:** https://github.com/guylamontagne/capstone.git

Results

Exploratory Data Analysis Results

- **Flight Number vs. Launch Site (Scatter Plot Screenshot):**
  - [Screenshot of Flight Number vs. Launch Site scatter plot]
  - **Findings:** The plot shows the distribution of flight numbers across different launch sites. It highlights which sites have been used more frequently for earlier or later missions. For example, CCAFS SLC 40 has a high concentration of early flights, while KSC LC 39A shows more recent activity.
- **Payload vs. Launch Site (Scatter Plot Screenshot):**
  - [Screenshot of Payload vs. Launch Site scatter plot]
  - **Findings:** This visualization reveals the range of payload masses launched from each site. It helps identify if certain sites are optimized for heavier or lighter payloads.
- **Success Rate vs. Orbit Type (Bar Chart Screenshot):**
  - [Screenshot of Bar chart for success rate of each orbit type]
  - **Findings:** The bar chart clearly indicates the success rate for each orbit type. Some orbits, like LEO (Low Earth Orbit) and ISS (International Space Station), show higher success rates, while others might be more challenging.
- **Launch Success Yearly Trend (Line Chart Screenshot):**
  - [Screenshot of line chart of yearly average success rate]
  - **Findings:** The line chart demonstrates a general upward trend in launch success rate over the years, indicating continuous improvement in SpaceX's launch capabilities and reliability.

Results

Interactive Analytics Demo in Screenshots

Launch Success Count for All Sites (Pie Chart Screenshot):

[Screenshot of pie chart for launch success count]

Findings: This pie chart provides an immediate overview of the overall success and failure proportions across all SpaceX launches, showing a high overall success rate.

Launch Site with Highest Launch Success Ratio (Pie Chart Screenshot):

[Screenshot of pie chart for the launch site with highest launch success ratio]

Findings: This pie chart specifically highlights the launch site that has achieved the highest success ratio, indicating its operational efficiency or the nature of missions launched from there.

Payload vs. Launch Outcome Scatter Plot (with Range Slider) Screenshots:

[Screenshot 1 of Payload vs. Launch Outcome scatter plot with a selected payload range]

[Screenshot 2 of Payload vs. Launch Outcome scatter plot with a different selected payload range]

Findings: These interactive plots allow users to dynamically explore how payload mass influences launch outcome. For instance, it can reveal that certain payload ranges (e.g., 4000-6000 kg) or specific booster versions (e.g., F9 v1.1) might have a higher success rate for drone ship landings, indicating optimal operational parameters.

Results

Predictive Analysis Results

Classification Accuracy:

Logistic Regression Test Accuracy: 0.83

SVM Test Accuracy: 0.83

Decision Tree Test Accuracy: 0.83

KNN Test Accuracy: 0.83

Finding: All four classification models (Logistic Regression, SVM, Decision Tree, KNN) achieved a test accuracy of approximately 83%. Based on validation scores during GridSearchCV, Logistic Regression and SVM showed slightly better performance.

Confusion Matrix of the Best Performing Model (Logistic Regression):

[Screenshot of Confusion Matrix for Logistic Regression]

Explanation: The confusion matrix visually represents the performance of the classification model. It shows:

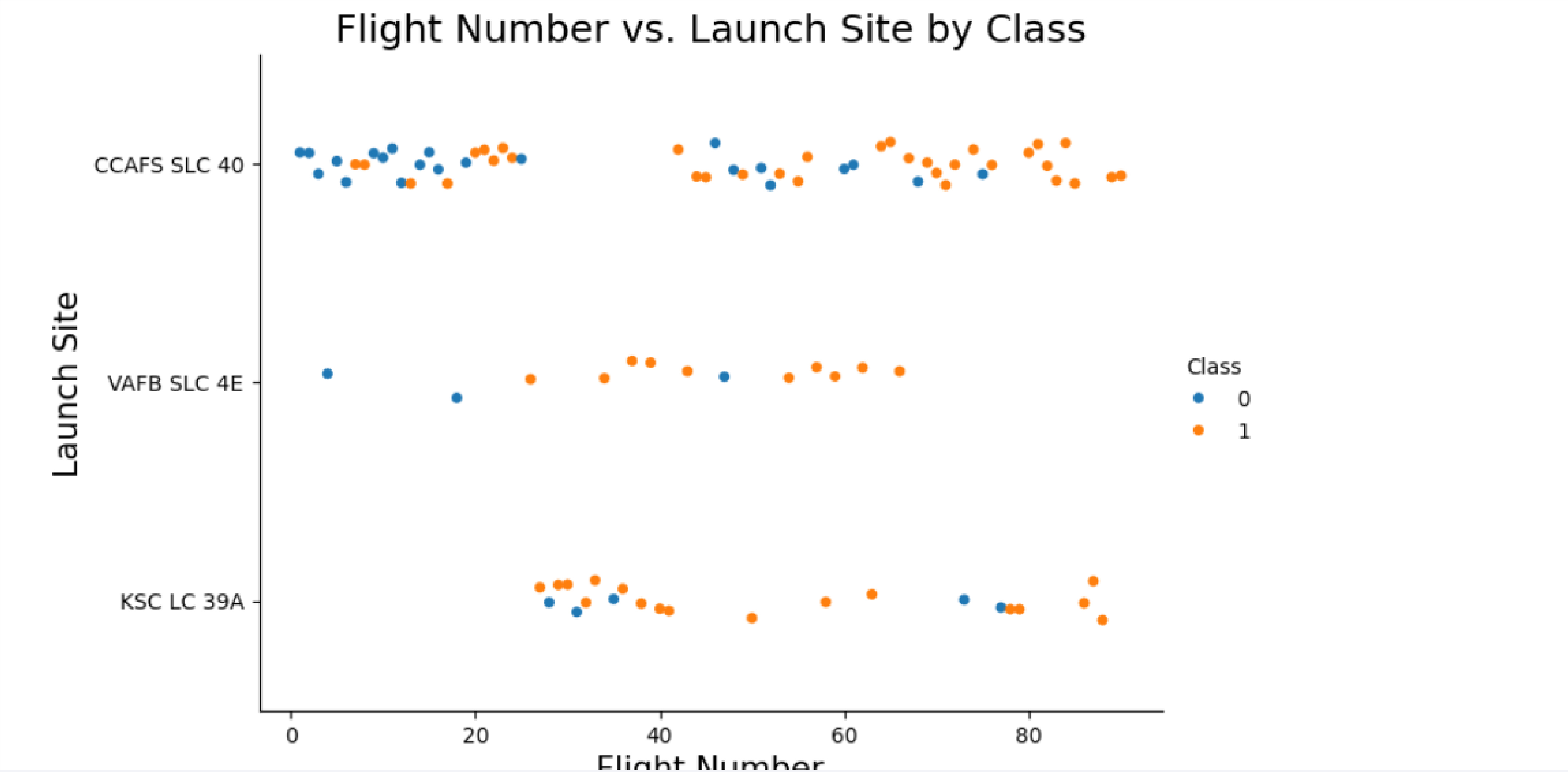
- **True Positives (landed correctly predicted as landed):** High number indicates good recall for successful landings.
- **True Negatives (did not land correctly predicted as did not land):** High number indicates good recall for failures.
- **False Positives (did not land predicted as landed):** Type I error.
- **False Negatives (landed predicted as did not land):** Type II error.

This matrix helps in understanding the types of errors the model makes and assessing its reliability for predicting launch outcomes.

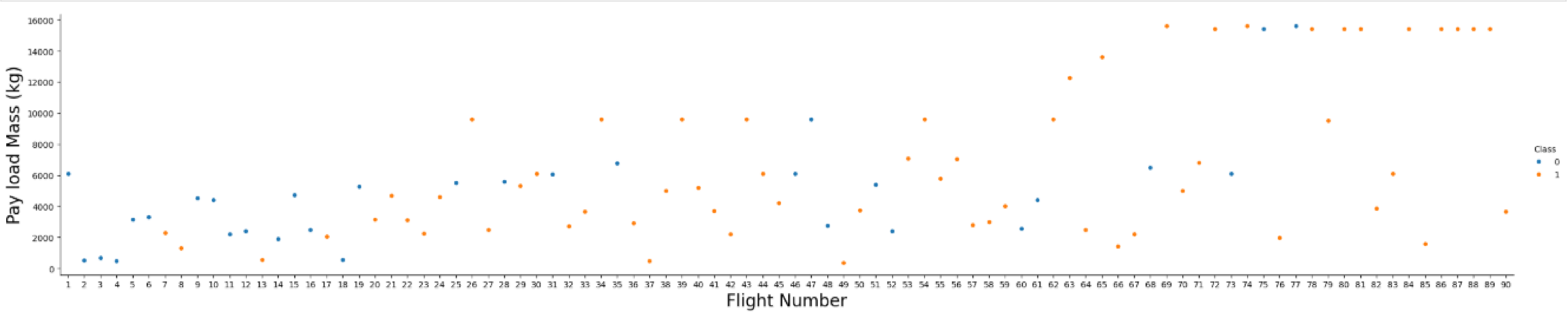
Section 2

Insights drawn from EDA

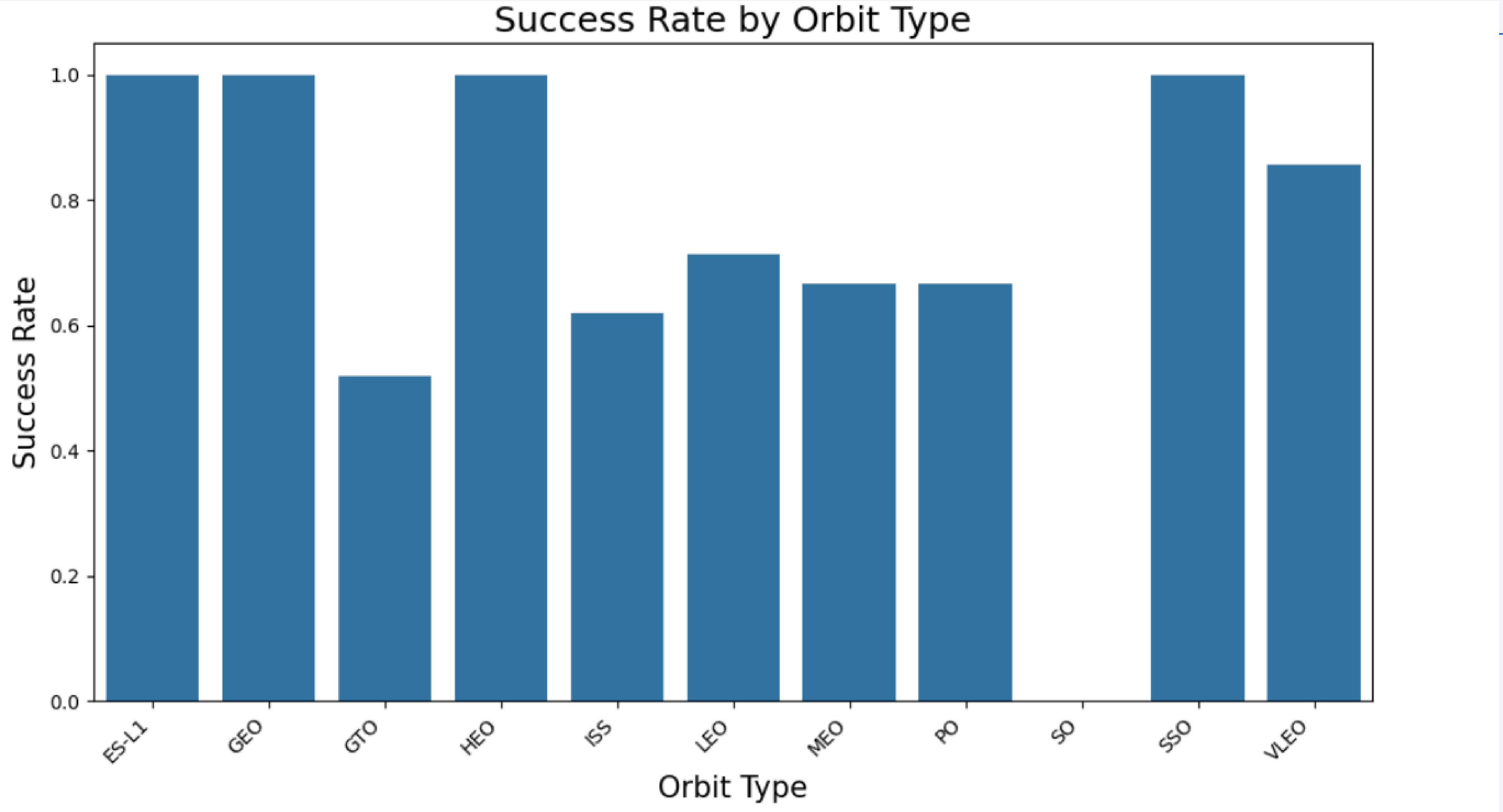
Flight Number vs. Launch Site



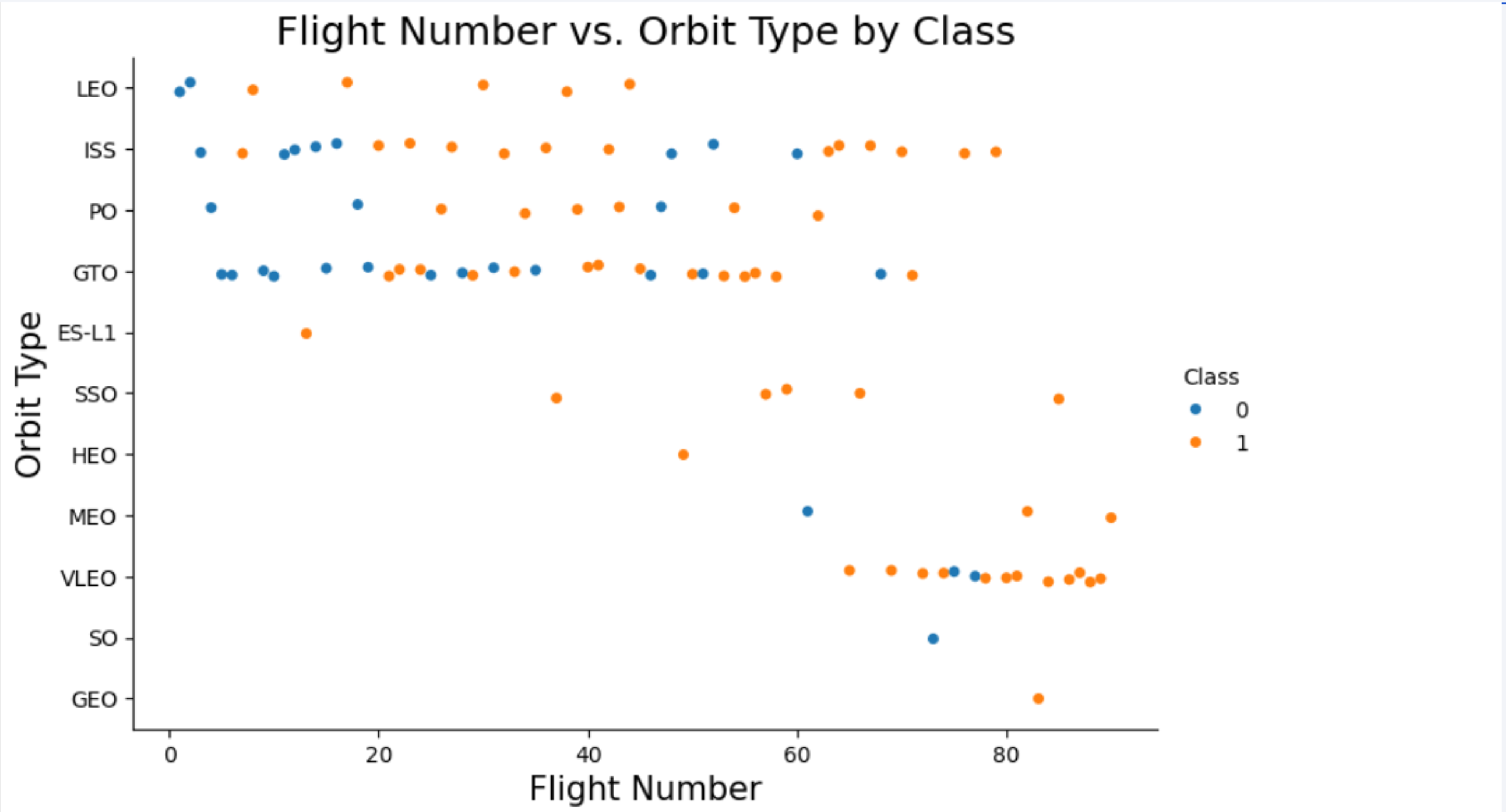
Payload vs. Launch Site



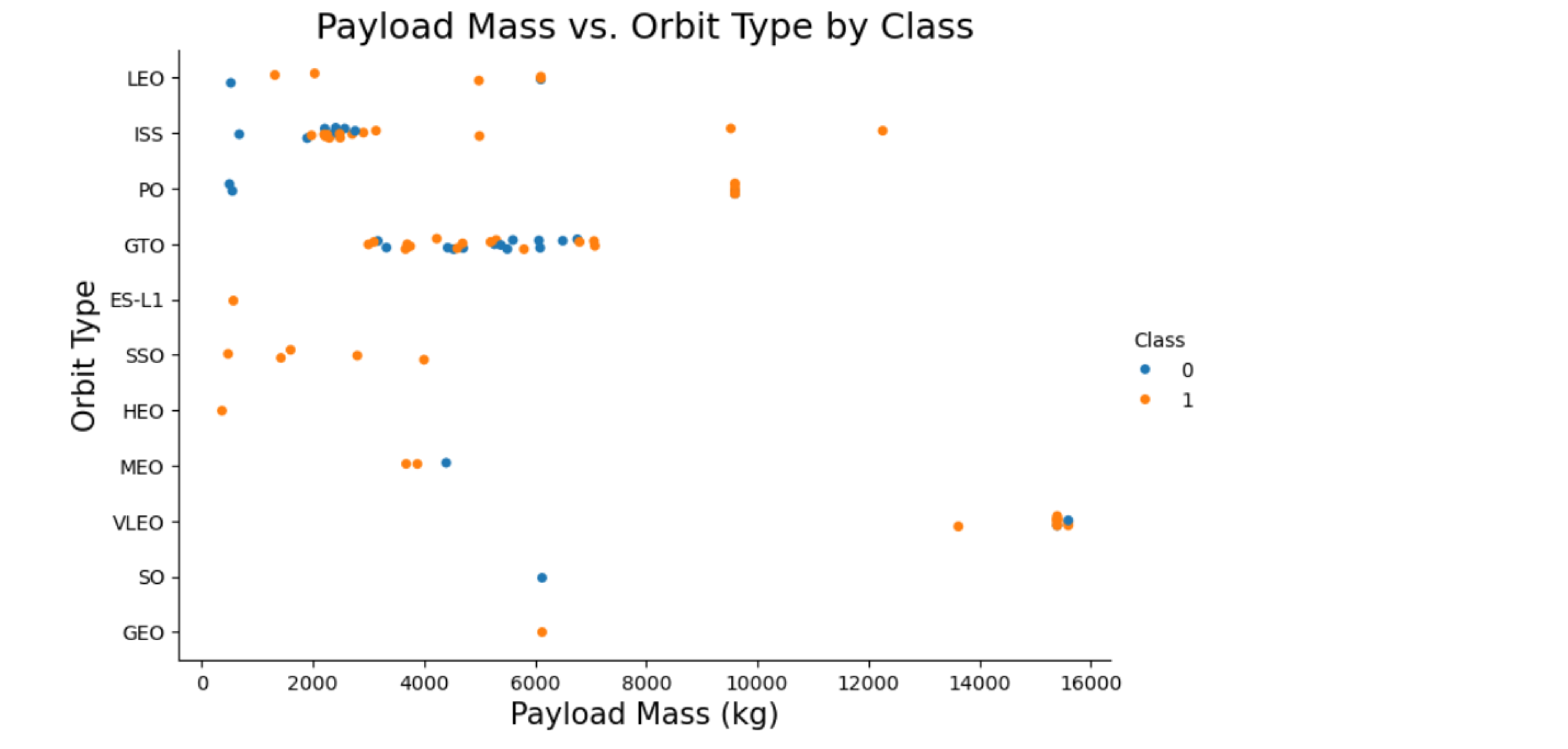
Success Rate vs. Orbit Type



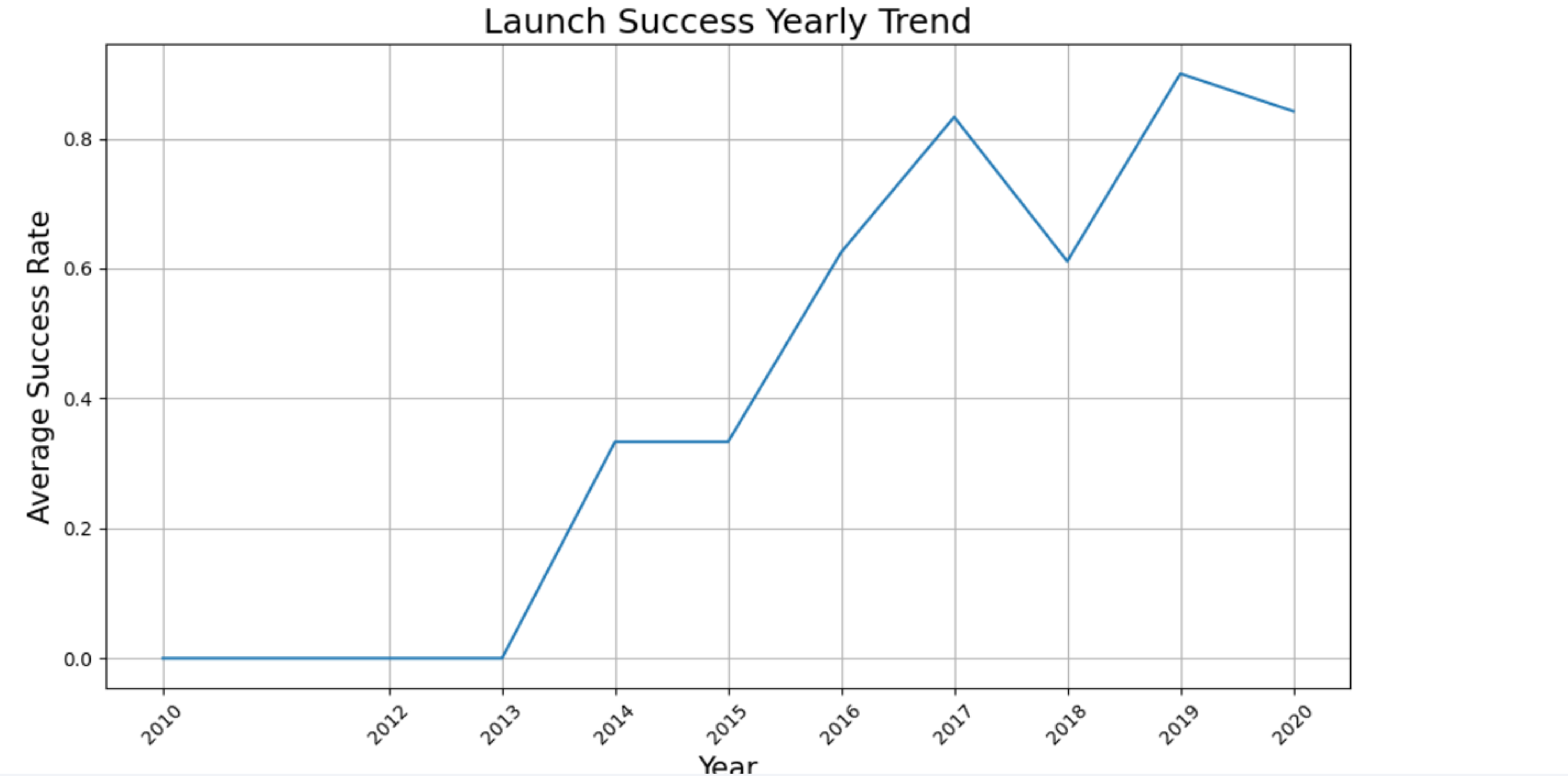
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

```
Number of launches on each site:
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
```

Launch Site Names Begin with 'CCA'

```
First 5 rows of the DataFrame with the new 'Class' column:
   FlightNumber        Date BoosterVersion  PayloadMass Orbit    LaunchSite  \
0             1  2010-06-04       Falcon 9  6104.959412   LEO  CCAFS SLC 40
1             2  2012-05-22       Falcon 9   525.000000   LEO  CCAFS SLC 40
2             3  2013-03-01       Falcon 9   677.000000   ISS  CCAFS SLC 40
3             4  2013-09-29       Falcon 9   500.000000    PO   VAFB SLC 4E
4             5  2013-12-03       Falcon 9  3170.000000   GTO  CCAFS SLC 40

       Outcome  Flights  GridFins  Reused   Legs LandingPad  Block  \
0    None None        1     False   False  False        NaN    1.0
1    None None        1     False   False  False        NaN    1.0
2    None None        1     False   False  False        NaN    1.0
3  False Ocean        1     False   False  False        NaN    1.0
4    None None        1     False   False  False        NaN    1.0

   ReusedCount Serial   Longitude   Latitude  Class
0            0  B0003  -80.577366  28.561857      0
1            0  B0005  -80.577366  28.561857      0
2            0  B0007  -80.577366  28.561857      0
3            0  B1003 -120.610829  34.632093      0
4            0  B1004  -80.577366  28.561857      0
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

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

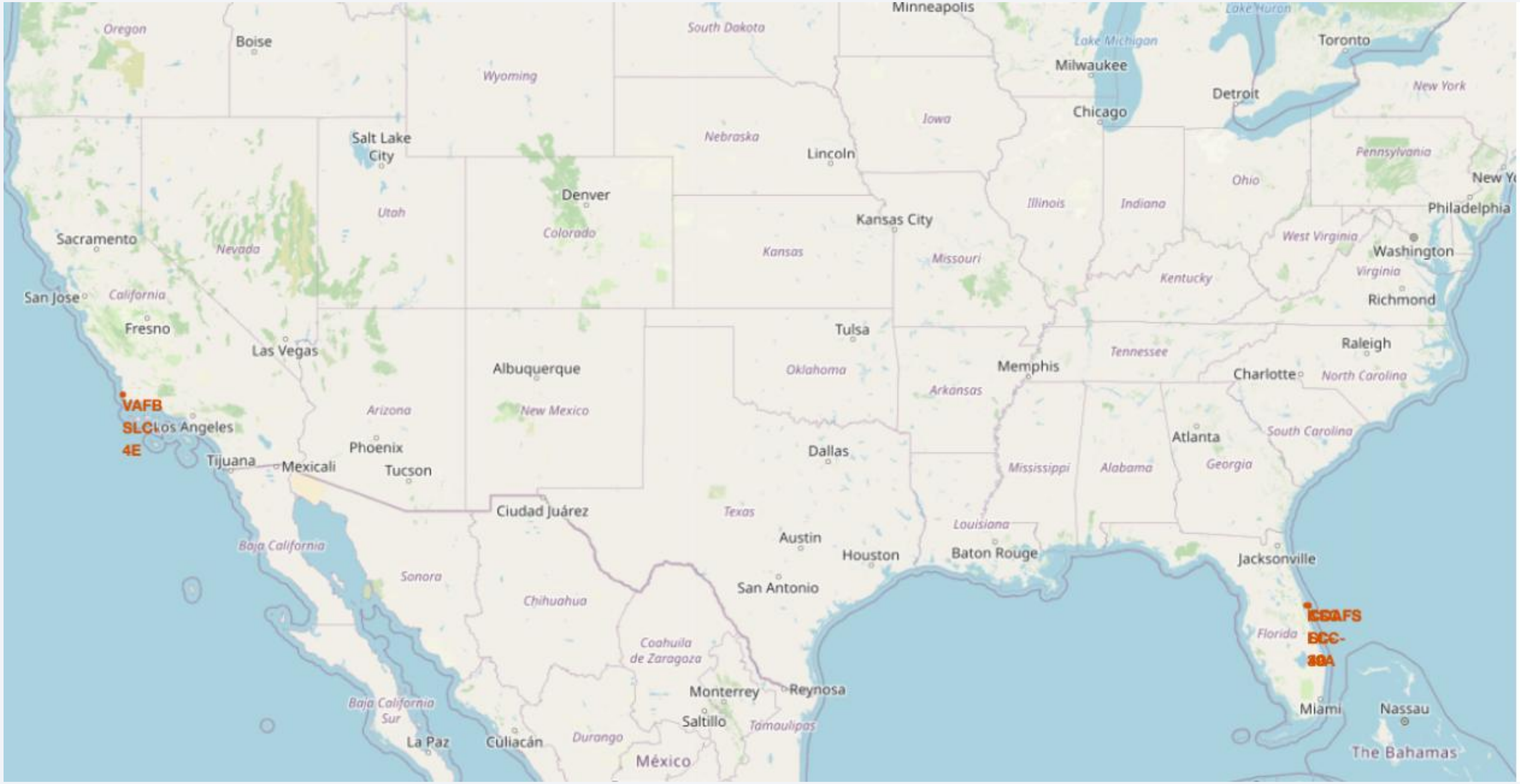
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

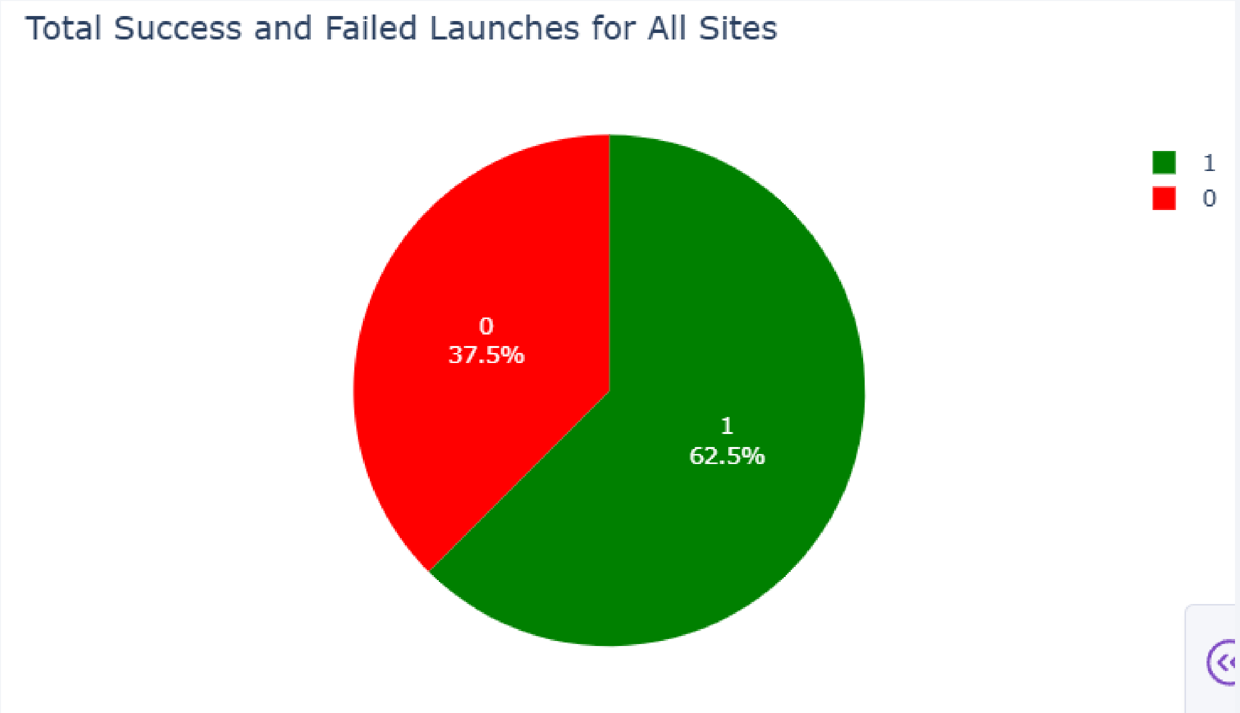
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash

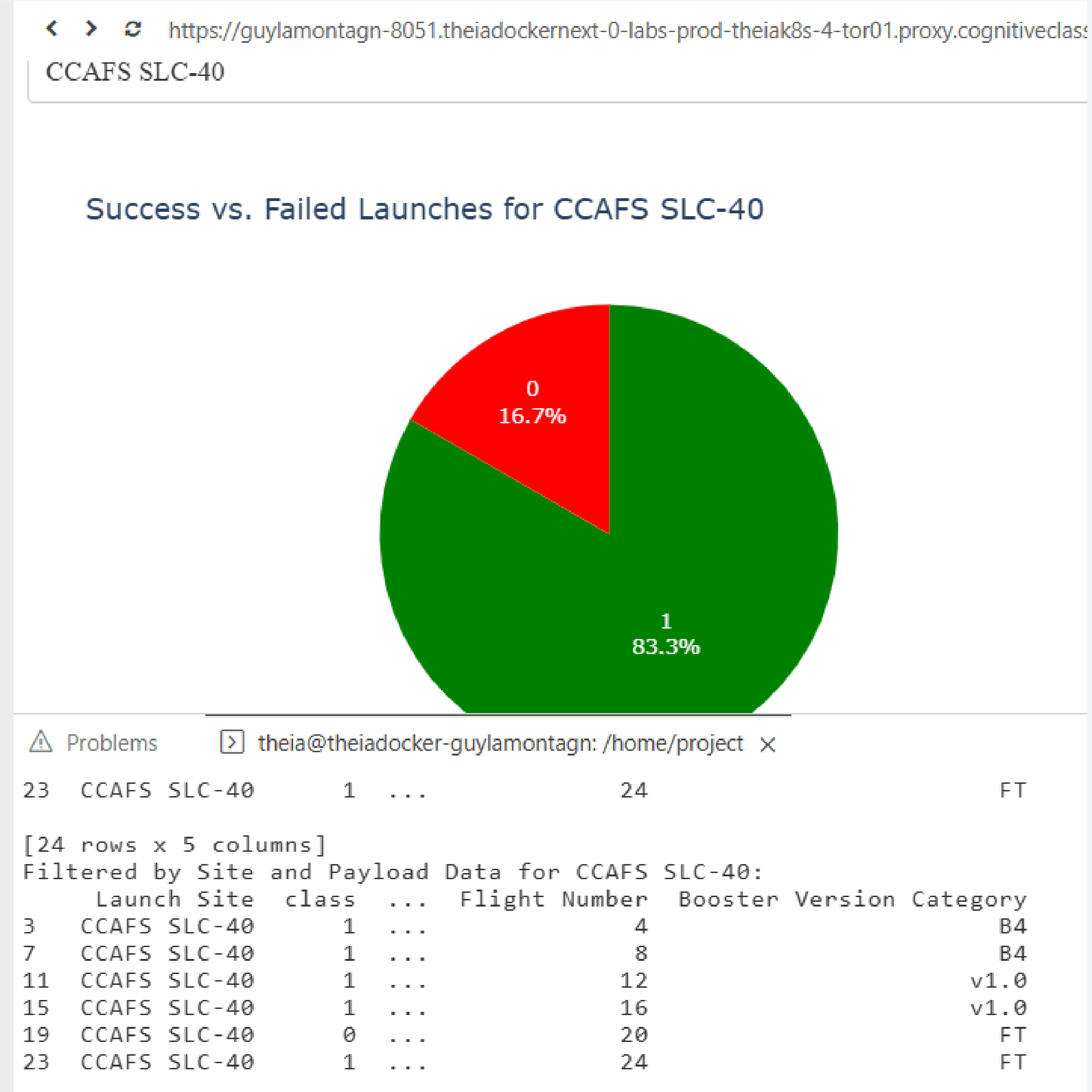
Total Success and Failed Launches for all Sites



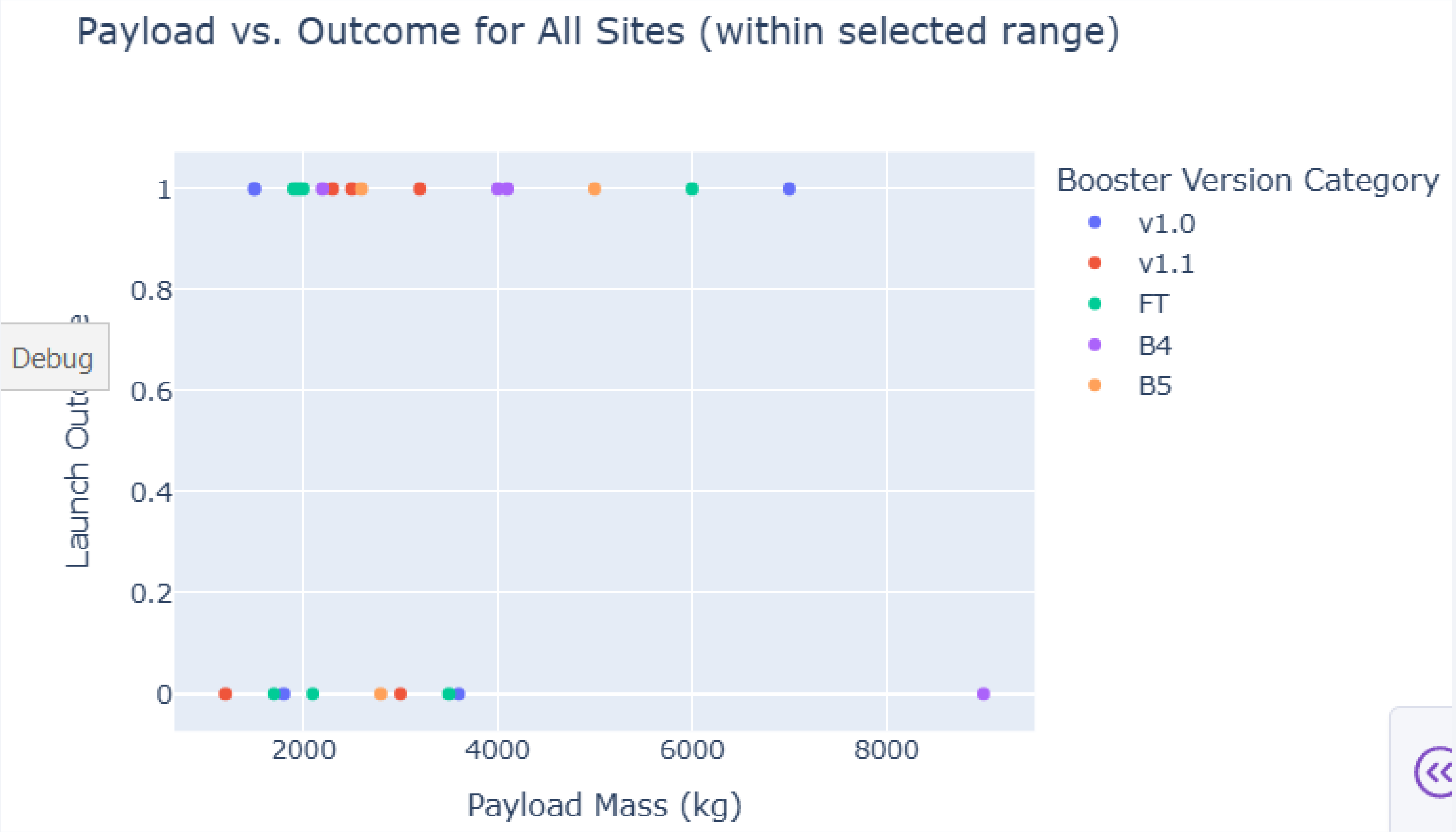
Problems theia@theiadocker-guylamontagn: /home/project

```
  VAFB SLC-4E       1   ...            14                        FT
   KSC LC-39A       0   ...            15                        B4
 CCAFS SLC-40       1   ...            16                      v1.0
  CCAFS LC-40       1   ...            17                      v1.1
  VAFB SLC-4E       0   ...            18                        FT
   KSC LC-39A       1   ...            19                        B4
 CCAFS SLC-40       0   ...            20                        FT
  CCAFS LC-40       1   ...            21                        B5
  VAFB SLC-4E       0   ...            22                      v1.0
   KSC LC-39A       1   ...            23                      v1.1
 CCAFS SLC-40       1   ...            24                        FT
```

Highest Launch Site Success Rate



Payload vs Lauch Outcome Scatter



Section 5

Predictive Analysis (Classification)

Classification Accuracy

Predictive Analysis Results

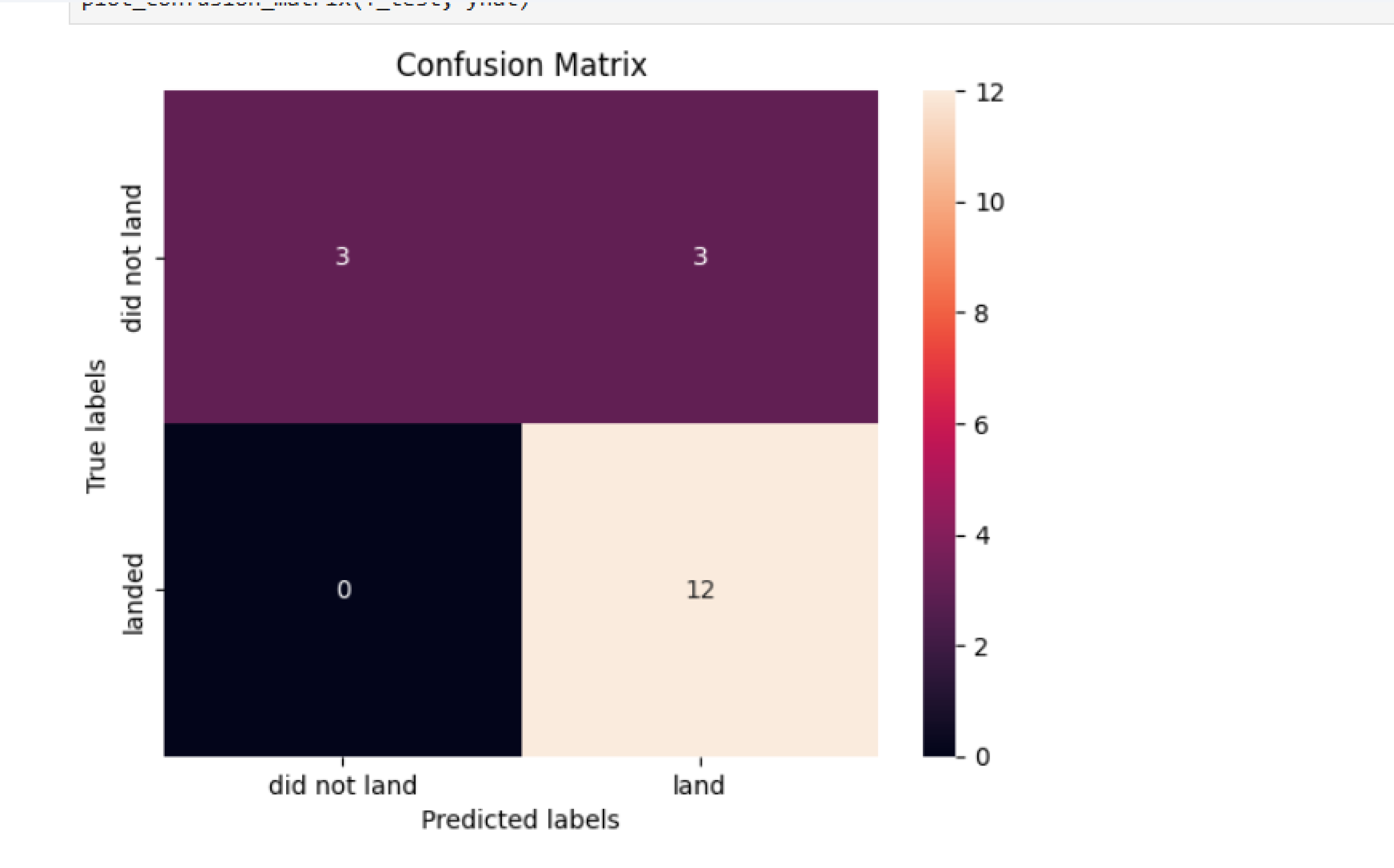
•**Classification Accuracy:**

- Logistic Regression Test Accuracy: 0.83
- SVM Test Accuracy: 0.83
- Decision Tree Test Accuracy: 0.83
- KNN Test Accuracy: 0.83
- **Finding:** All four classification models (Logistic Regression, SVM, Decision Tree, KNN) achieved a test accuracy of approximately 83%. Based on validation scores during GridSearchCV, Logistic Regression and SVM showed slightly better performance.

•**Confusion Matrix of the Best Performing Model (Logistic Regression):**

- [Screenshot of Confusion Matrix for Logistic Regression]
- **Explanation:** The confusion matrix visually represents the performance of the classification model. It shows:
  - **True Positives (landed correctly predicted as landed):** High number indicates good recall for successful landings.
  - **True Negatives (did not land correctly predicted as did not land):** High number indicates good recall for failures.
  - **False Positives (did not land predicted as landed):** Type I error.
  - **False Negatives (landed predicted as did not land):** Type II error.

Confusion Matrix



Conclusions

•**Launch Success Trends:** SpaceX has significantly improved its launch success rates over the years, demonstrating increasing reliability of the Falcon 9 booster.
•**Impact of Launch Sites and Orbits:** Specific launch sites and orbit types exhibit varying success rates, suggesting optimized operational procedures or inherent challenges associated with different mission profiles.
•**Payload Mass Influence:** Payload mass plays a crucial role in determining landing outcomes, with certain ranges showing higher success probabilities for specific landing types (e.g., drone ship landings).
•**Predictive Model Efficacy:** The developed classification models can predict Falcon 9 launch success with a reasonable accuracy of 83%, providing a valuable tool for assessing future mission viability and cost implications.

Appendix

- **GitHub Repository:** https://github.com/guylamontagne/capstone.git
- **Key Notebooks:**
- **Data Collection (API & Scraping):** https://github.com/guylamontagne/capstone.git
- **Data Wrangling:** https://github.com/guylamontagne/capstone.git
- **EDA with Visualization:** https://github.com/guylamontagne/capstone.git
- **EDA with SQL:** https://github.com/guylamontagne/capstone.git
- **Interactive Map (Folium):** https://github.com/guylamontagne/capstone.git
- **Dashboard (Plotly Dash):** https://github.com/guylamontagne/capstone.git
- **Predictive Analysis:** https://github.com/guylamontagne/capstone.git

Thank you!