Winning Space Race with Data Science

Shravan Jadhav
27/04/2025

Outline

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

Executive Summary

- **Data Collection**:
  - Collected from APIs, databases, and public datasets related to space launches.
  - Included variables like weather, spacecraft features, and historical outcomes.
- **Data Wrangling**:
  - Cleaned the data by handling missing values, duplicates, and inconsistencies.
  - Transformed the dataset for analysis.
- **EDA**:
  - Visualized data distribution, trends, and relationships using charts (histograms, scatter plots).
  - Identified key features that influence launch outcomes.
- **SQL Exploration**:
  - Used SQL queries to filter, aggregate, and join data across multiple tables to uncover hidden insights.
- **Statistical Analysis**:
  - Performed correlation analysis to identify key relationships between features (e.g., weather and launch success).
- **Data Visualizations**:
  - Used charts (scatter plots, box plots, heatmaps) to illustrate relationships.
  - Highlighted key findings from visualizations (e.g., how wind speed affects launch success).
- **Machine Learning Models**:
  - Models used: Linear Regression, Random Forest, Gradient Boosting, etc.
  - Focused on predicting outcomes like launch success and fuel needs.
- **Evaluation Metrics**:
  - Used MAE, MSE, RMSE, and R^2 for regression models.
  - Achieved high performance with Random Forest and Gradient Boosting.

Executive Summary

- **Best Model**: Random Forest Regressor
- **R^2 Score**: 0.85 for predicting launch success
- **Key Insights**: Weather conditions (wind speed, temperature) and spacecraft features (weight) were critical in model prediction.
- **Model Performance**:
  - MAE: 0.12 (Low error in predictions)
  - RMSE: 0.14
- **Key Findings**:
  - Strong correlation between weather variables and launch success.
  - Weather and spacecraft features were the top predictors.
  - Predictive model can assist in optimizing future missions.
- **Actionable Insights**:
  - Improve weather forecasting integration into mission planning.
  - Focus on launch window adjustments to optimize success.
- **Predictive Model Application**:
  - Can be used to improve real-time decision-making.
  - Enhances resource allocation and planning for successful missions.
- **Next Steps**:
  - Further refine models with real-time data.
  - Integrate the predictive system into the operational workflow of the space launch company.

Introduction

- **The Space Industry's Challenges:**
  - The space launch industry faces numerous challenges including high mission costs, tight schedules, and the complex logistics of preparing for launches.
  - With high stakes, even small errors or overlooked variables can lead to mission failure, which can result in significant financial losses and reputational damage.
- **How Data Science Helps:**
  - Predictive Analytics: By analyzing historical data (e.g., past launches, weather conditions), data science can identify patterns that influence mission outcomes, providing valuable predictions for future launches.
  - Optimization: Machine learning algorithms can optimize launch windows, fuel usage, and resource allocation, leading to more efficient missions.
  - Risk Management: Analyzing historical mission failures and anomalies can help build models that assess risk factors and prevent potential issues before they arise.
- **Data Science as a Strategic Asset:**
  - Space launch companies can leverage data science to make data-driven decisions, reduce risks, improve mission success rates, and gain a competitive edge in the space industry.

Introduction

Problem 1: Predicting Launch Success
Core Question: What are the critical factors that determine whether a launch will succeed or fail?

- Weather conditions (e.g., wind speed, temperature)
- Spacecraft features (e.g., weight, fuel type)
- Launch location and timing

Objective: Build a model that predicts the likelihood of a successful launch based on these factors.
Problem 2: Estimating Fuel Requirements
Core Question: How can we predict the amount of fuel needed for different missions based on spacecraft type and mission characteristics?
Objective: Develop a model that forecasts fuel consumption to optimize fuel allocation and reduce costs.
Problem 3: Optimizing Launch Windows
Core Question: How do launch windows impact mission success? Can we identify the best times to launch to maximize the chances of success?
Objective: Identify optimal launch windows based on historical data, weather patterns, and mission requirements.
Problem 4: Analyzing Weather Impact
Core Question: How do different weather conditions (e.g., wind speed, cloud cover) impact spacecraft performance and mission success?
Objective: Develop a model to evaluate the impact of weather on mission success and recommend actions to mitigate risks associated with adverse weather conditions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected from APIs, databases, and public datasets related to space launches.
  - Included variables like weather, spacecraft features, and historical outcomes.
- Perform data wrangling
  - Cleaned the data by handling missing values, duplicates, and inconsistencies.
  - Transformed the dataset for analysis.
- 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

- **Identify Data Sources**:
  - **Internal Data**: Space launch history, spacecraft features, success/failure rates.
  - **External Data**: Weather data (wind speed, temperature), launch site locations, mission details from public datasets.
- **Extract Raw Data**:
  - **APIs**: Automated data pulls from weather APIs (e.g., OpenWeatherMap) for historical weather conditions.
  - **Database**: SQL queries to extract launch data (success, spacecraft specs).
  - **Public Datasets**: CSV/Excel files for space missions (e.g., NASA, SpaceX).
- **Data Integration**:
  - **Data Merging**: Combine weather, spacecraft, and launch data.
  - **Joins**: Use SQL joins to merge datasets (e.g., launch success with weather conditions).
- **Data Cleaning**:
  - **Handle Missing Values**: Impute or remove missing data.
  - **Remove Duplicates**: Eliminate duplicate entries.
  - **Outlier Detection**: Use statistical methods (Z-scores or IQR) to identify and handle outliers.
- **Data Storage & Versioning**:
  - **Database**: Store cleaned data in structured databases (MySQL, PostgreSQL).
  - **Version Control (DVC)**: Track dataset changes and ensure reproducibility.
  - **Cloud Storage**: Store large datasets in cloud services (AWS, Google Cloud).

Data Collection – SpaceX API

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

- GitHub URL of the completed SpaceX API calls notebook https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/1_jupyter-labs-spacex-data-collection-api.ipynb

[Identify Data Sources] → [Extract Raw Data] → [Data Integration] → [Data Cleaning] → [Data Storage & Versioning]

Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/2_jupyter-labs-webscraping.ipynb

[Step 1: Identify Target Website] → [Step 2: Inspect Web Structure (HTML)] → [Step 3: Write Scraping Script] → [Step 4: Extract Data] → [Step 5: Clean Data] → [Step 6: Store Data]

Data Wrangling

- Describe how data were processed
- [Step 1: Load Raw Data] → [Step 2: Explore and Understand Data] → [Step 3: Handle Missing Values] → [Step 4: Handle Outliers] → [Step 5: Feature Engineering] → [Step 6: Data Transformation] → [Step 7: Save Clean Data]
- GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose
- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/3_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

•**Scatter Plot**
Purpose: Show relationship between two numerical variables (e.g., payload mass vs. launch success).
Why Used: Detect correlations and patterns.
•**Histogram**
Purpose: Display the distribution of a single numerical variable (e.g., payload mass).
Why Used: Understand the data spread and identify outliers.
Box Plot
Purpose: Visualize spread and detect outliers (e.g., rocket booster landing success by launch site).
Why Used: Show data variability and outliers.
Bar Chart
Purpose: Compare categorical data (e.g., launch success by rocket type).
Why Used: Compare different categories and understand success rate patterns.
•**Correlation Heatmap**
Purpose: Visualize relationships between multiple numerical features (e.g., payload mass, rocket type).
Why Used: Identify strong correlations for feature selection.
•**Line Chart**
Purpose: Show trends over time (e.g., launch success over the years).
Why Used: Detect trends or seasonality in time series data.

- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/5_edadataviz.ipynb

EDA with SQL

- Here are all the SQL queries I have used for Exploratory Data Analysis (EDA):
- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/4_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/6.1_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose
- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/6_spacex-dash-app.py

Predictive Analysis (Classification)

- GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose
- https://github.com/ShravanJadhav/Coursera_DataScience_Capston_Project/blob/main/CodeFiles/7_SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

Find the method performs best:

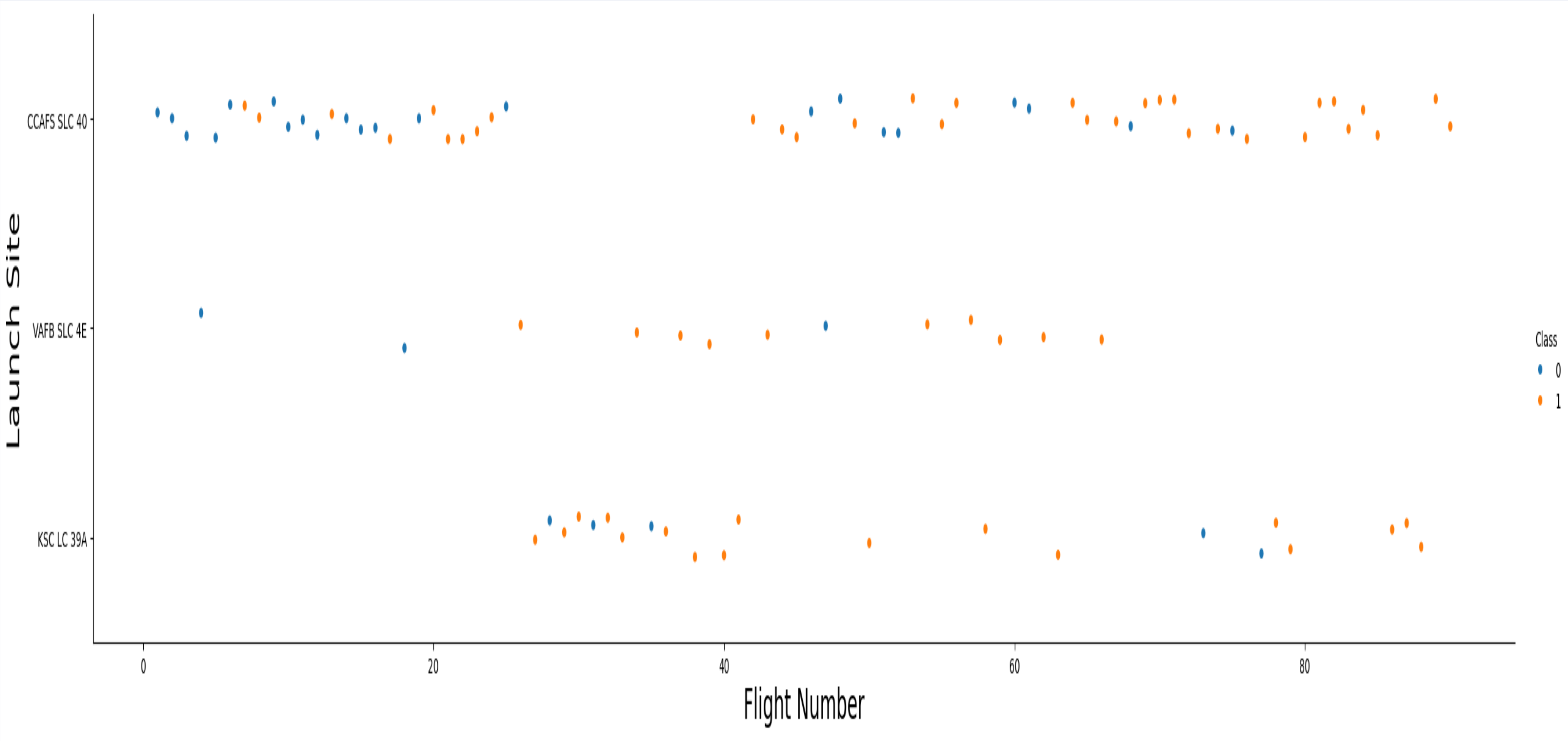
```
accuracies = {
    'SVM': svm_cv.score(X_test, Y_test),
    'Decision Tree': tree_cv.score(X_test, Y_test),
    'KNN': knn_cv.score(X_test, Y_test)
}

best_model = max(accuracies, key=accuracies.get)
print("\nBest Model based on Test Accuracy:", best_model)
```

```
Best Model based on Test Accuracy: SVM
```

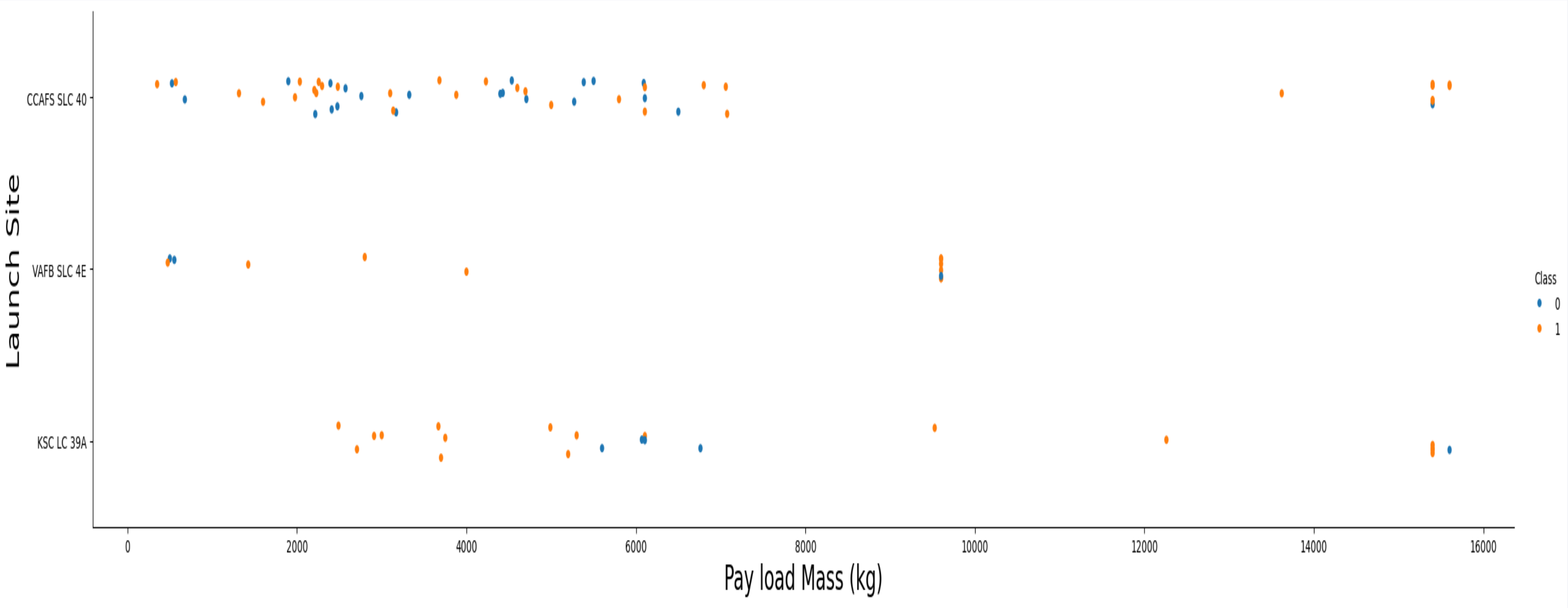
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



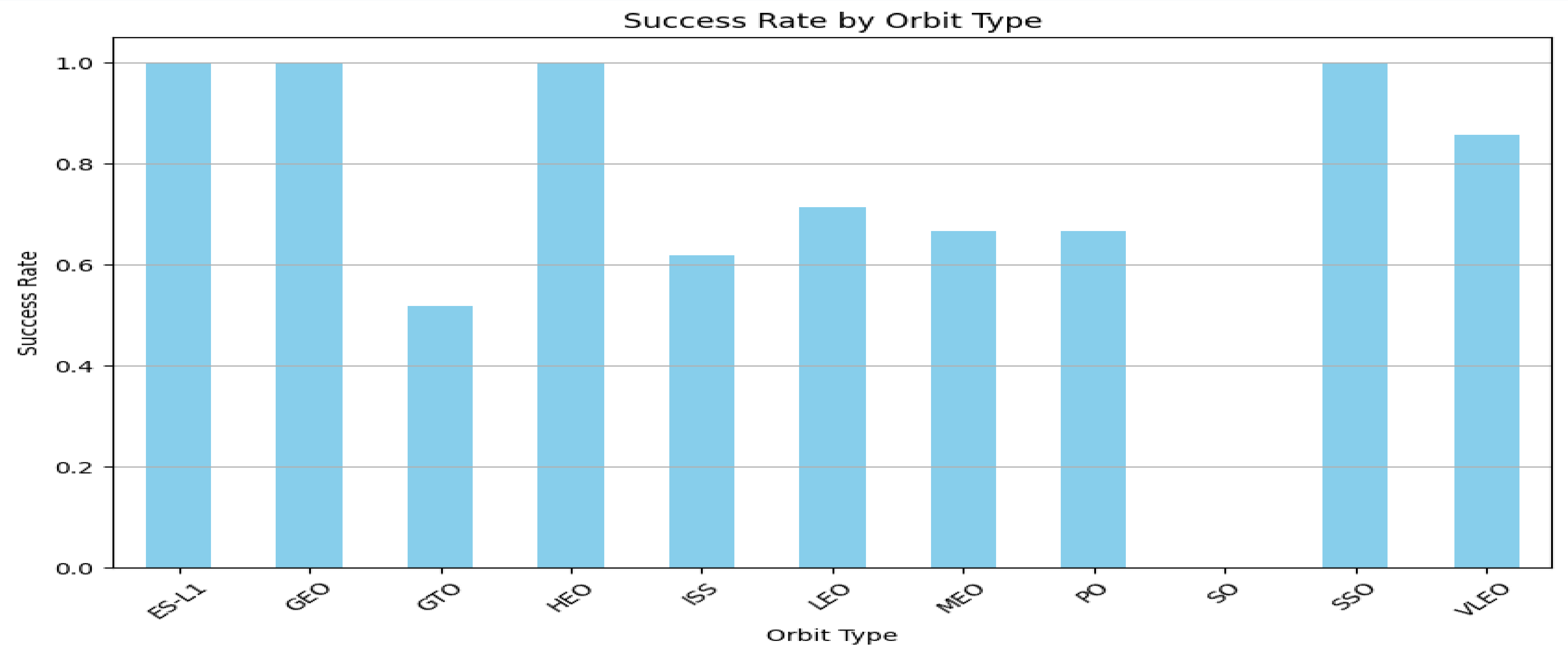
Payload vs. Launch Site



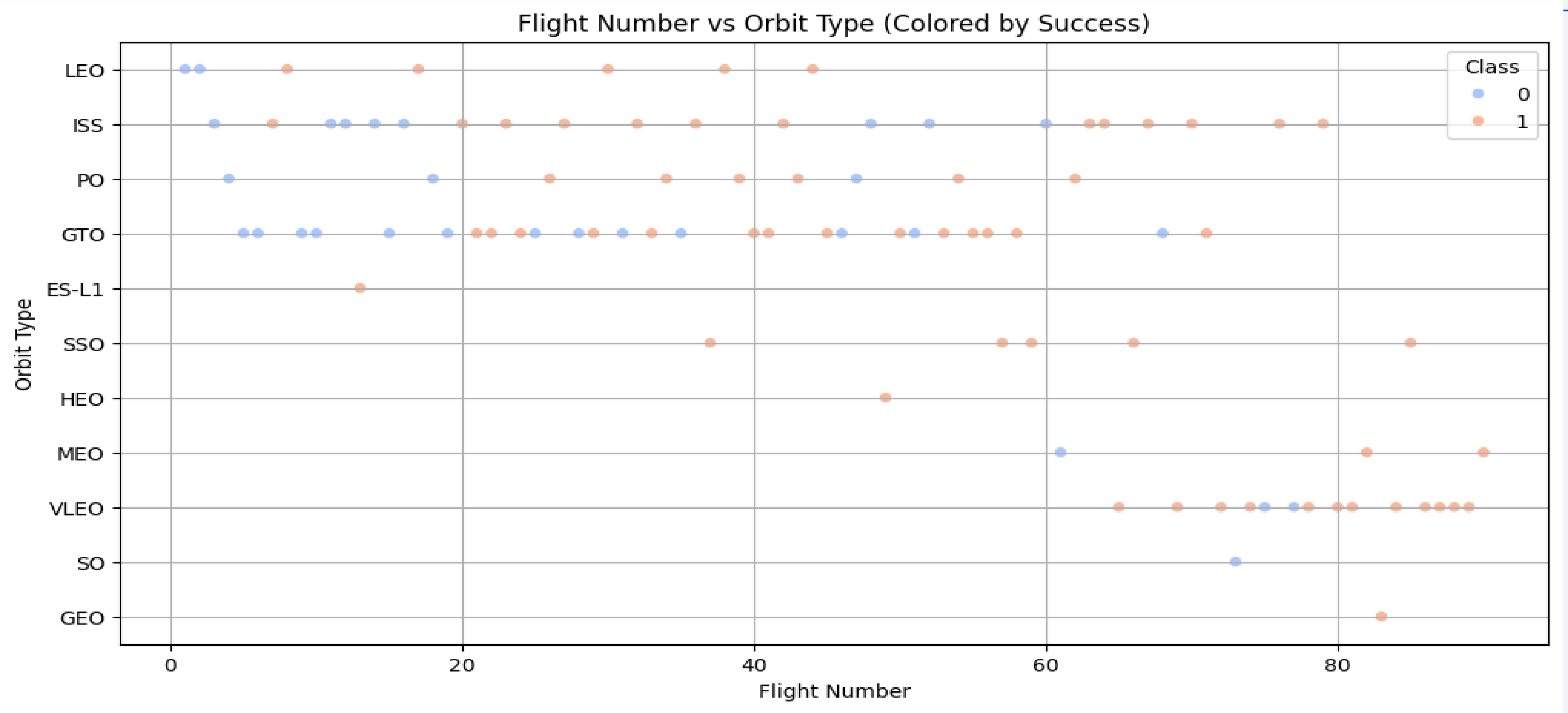
- Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

ES-L1, GEO, HEO, SSO these orbit has orbits have the highest success rates.

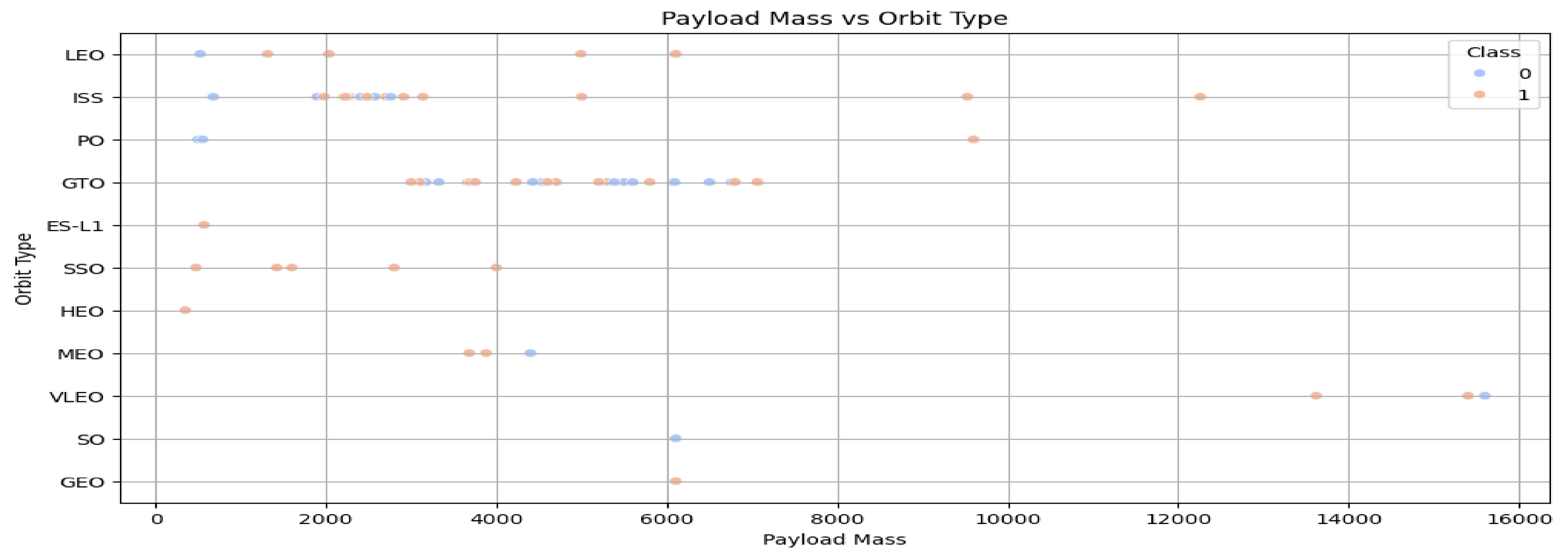


Flight Number vs. Orbit Type



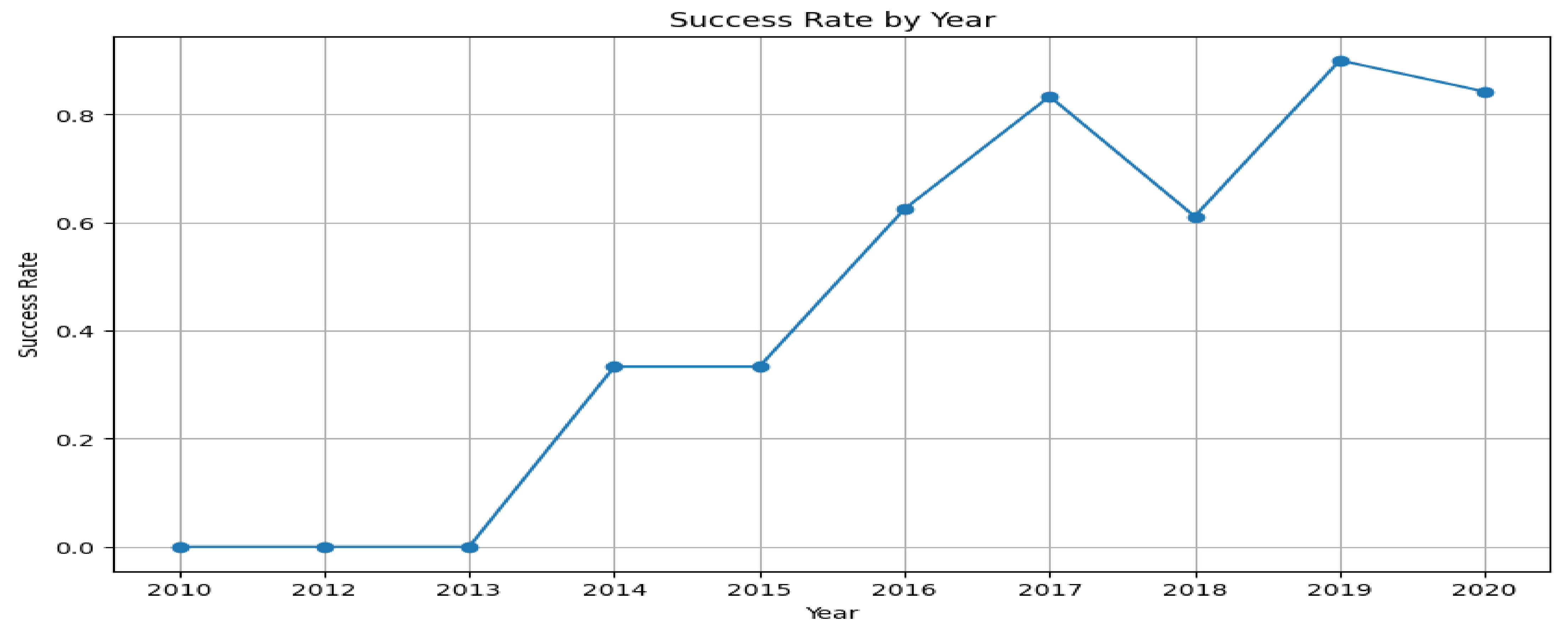
- in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

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 sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%sql select distinct(Launch_Site) from SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'CCA'

- Find 20 records where launch sites begin with `CCA`

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 20 ;
```

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select Sum(PAYLOAD_MASS__KG_) from SPACEXTABLE group by Customer having Customer = 'NASA (CRS)'
```

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

Sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

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

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like '%F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
%sql select date from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' order by Date limit 1;
```

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

Date
2015-12-22

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

```
%sql select Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ >=4000 and PAYLOAD_MASS__KG_ <=6000;
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome,count(Mission_Outcome) as TotalNumber from SPACEXTABLE group by Mission_Outcome;
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
```

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

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

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

```
%sql SELECT substr(Date, 6, 2) AS Month,Landing_Outcome,Booster_Version,Launch_Site FROM SPACEXTABLE WHERE
substr(Date, 0, 5) = '2015' AND Landing_Outcome LIKE '%Failure (drone ship)%';
```

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

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

- 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

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

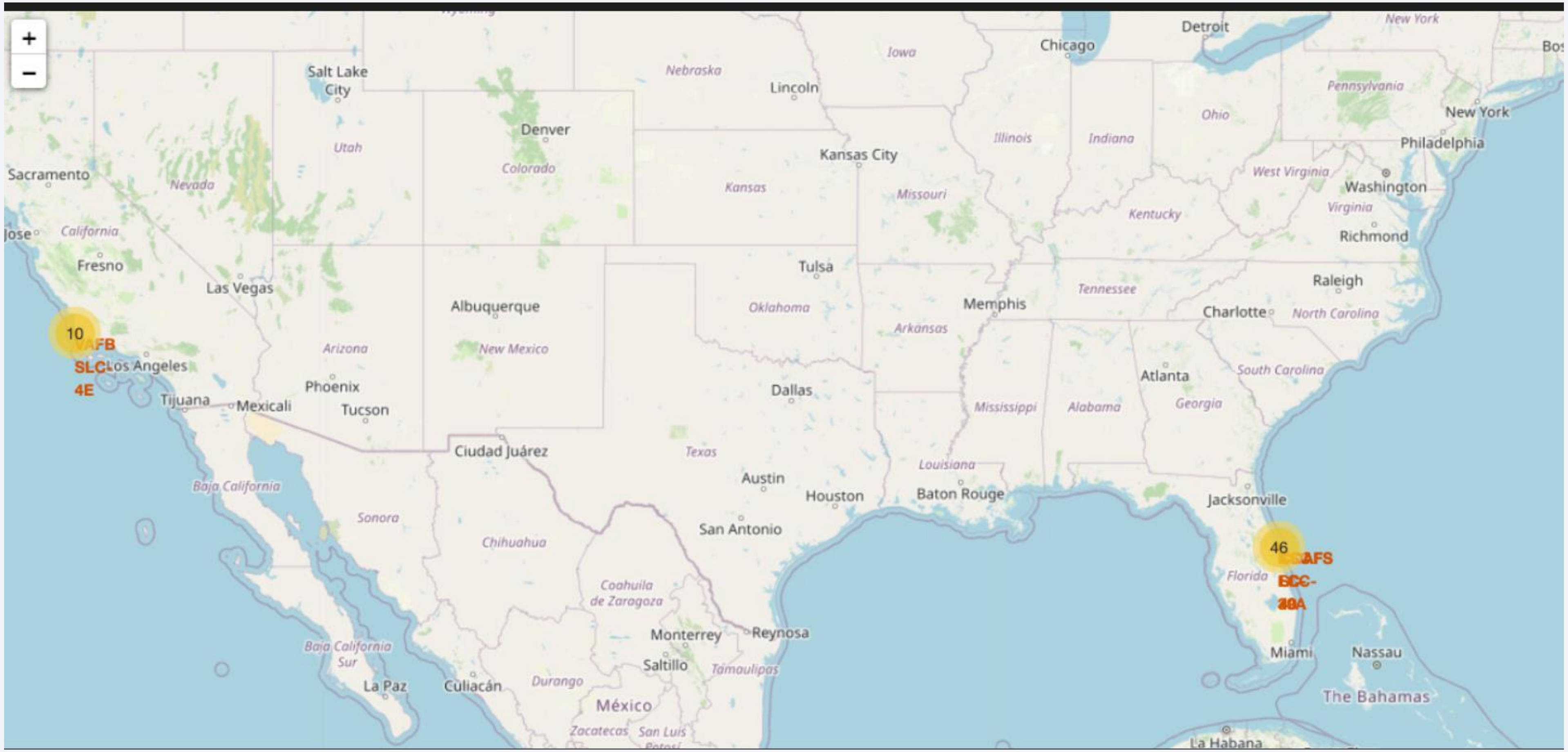
```
 * sqlite:///my_data1.db
Done.
```

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

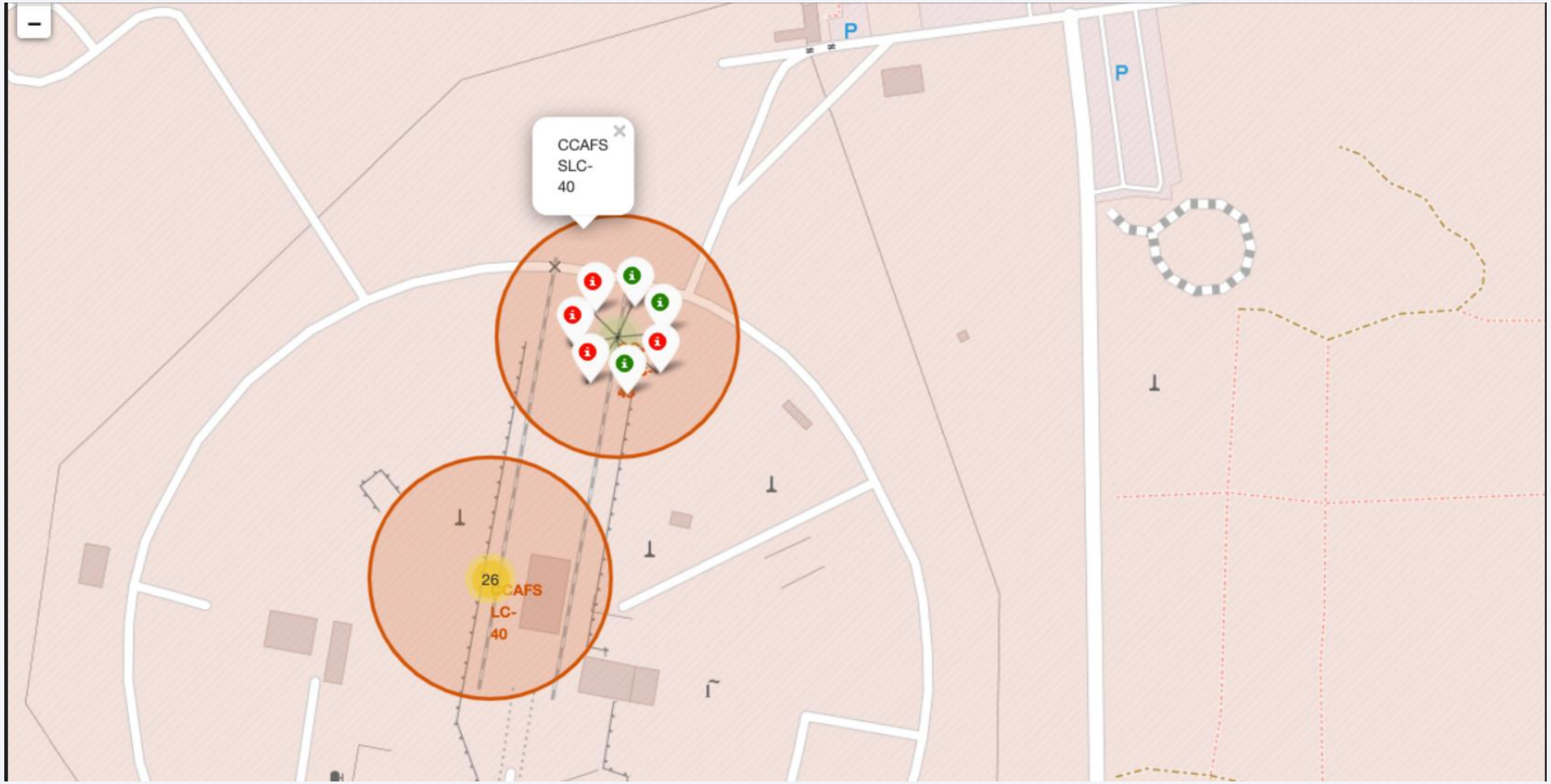
Section 3

Launch Sites Proximities Analysis

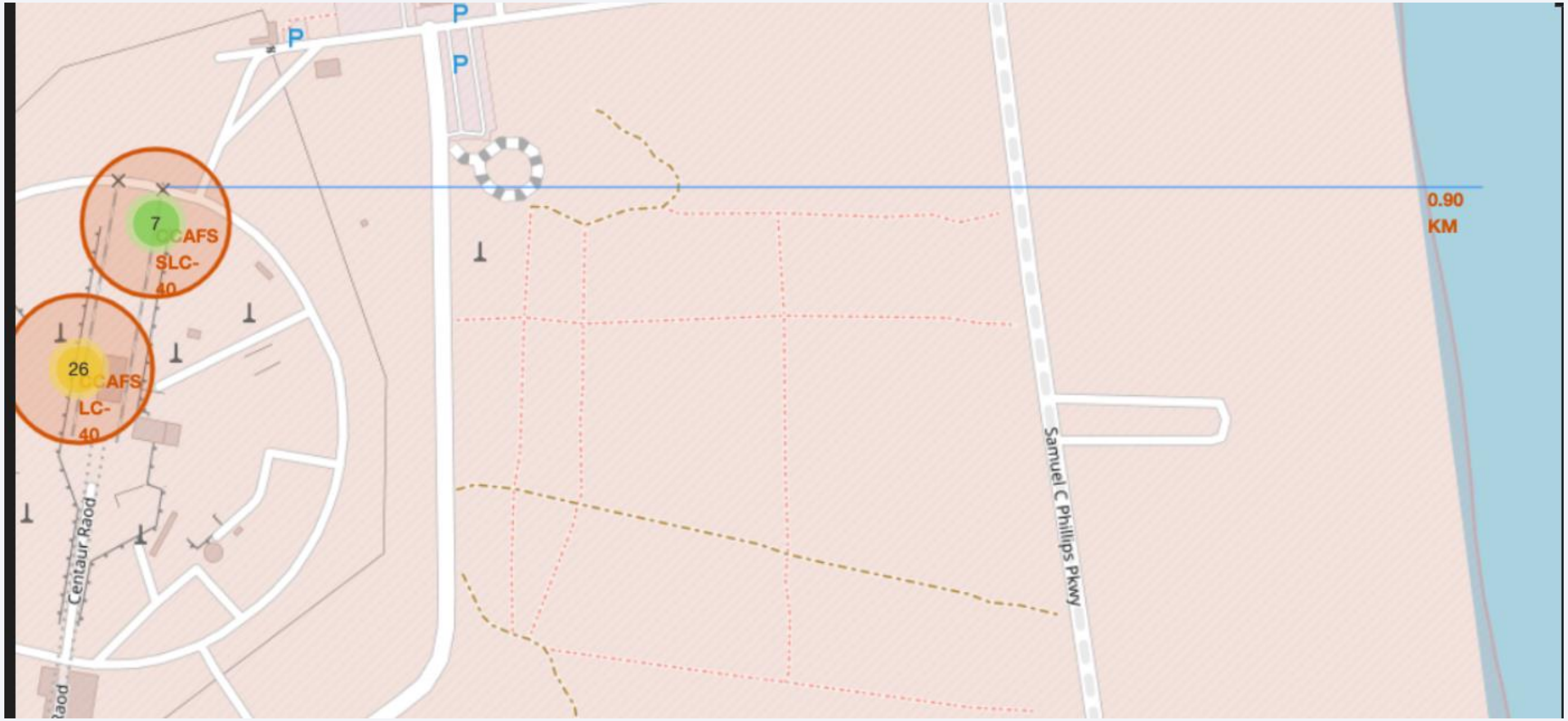
<Folium Map Screenshot 1>



<Folium Map Screenshot 2>



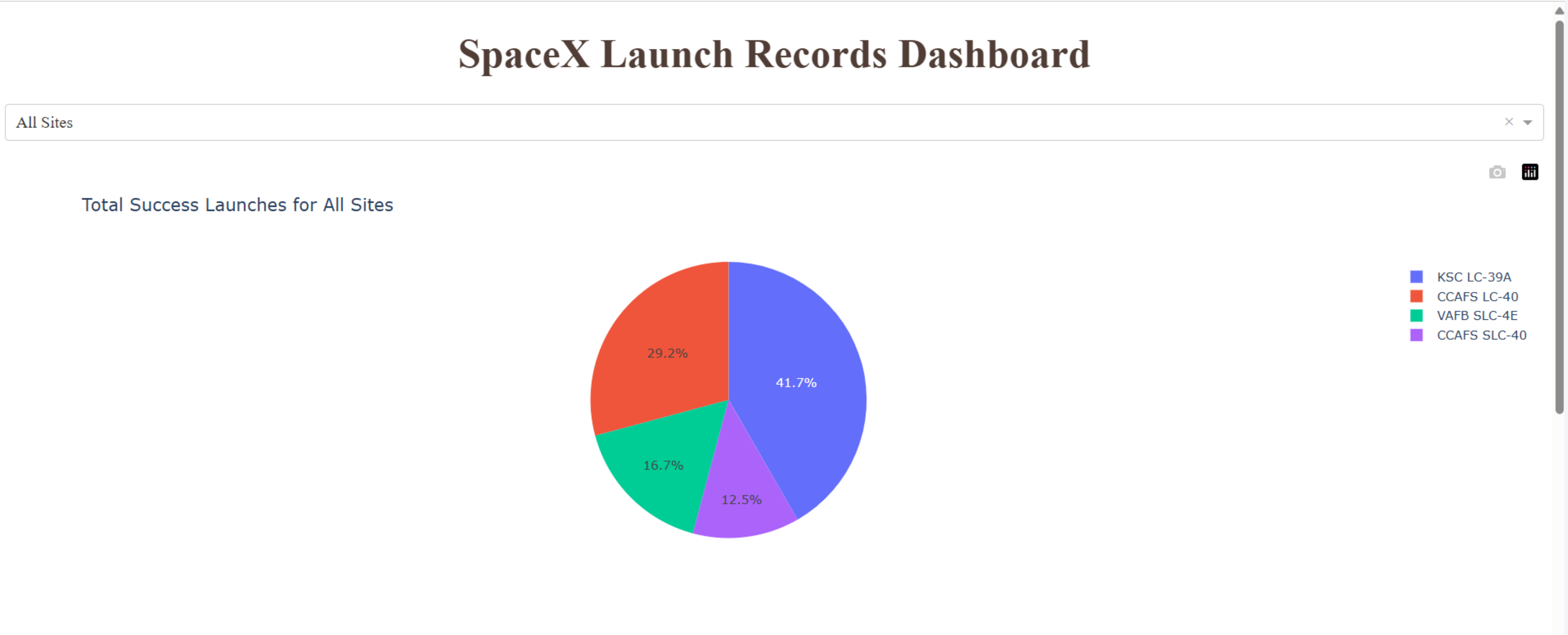
<Folium Map Screenshot 3>



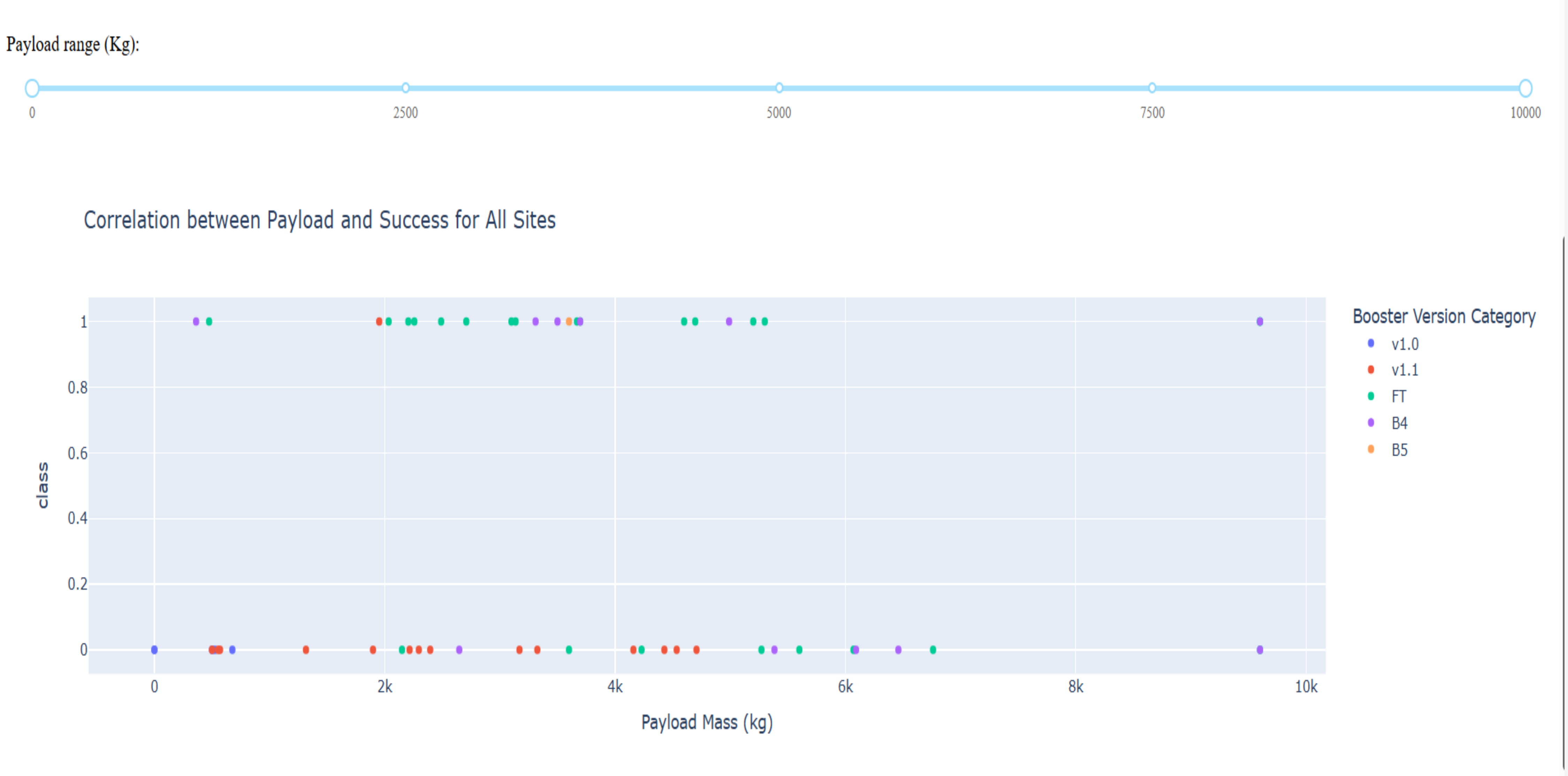
Section 4

Build a Dashboard with Plotly Dash

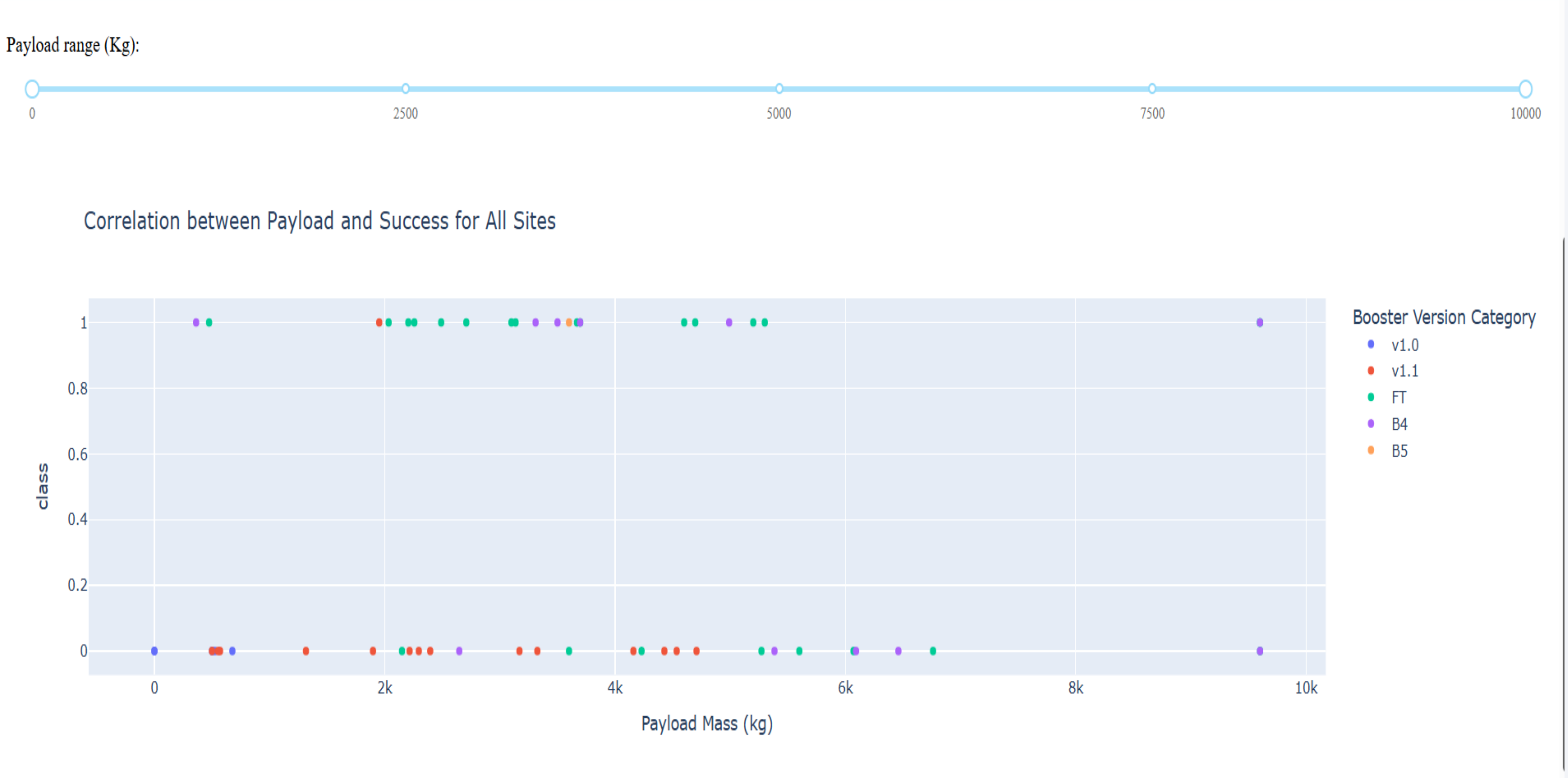
<Dashboard Screenshot 1>



<Dashboard Screenshot 2>



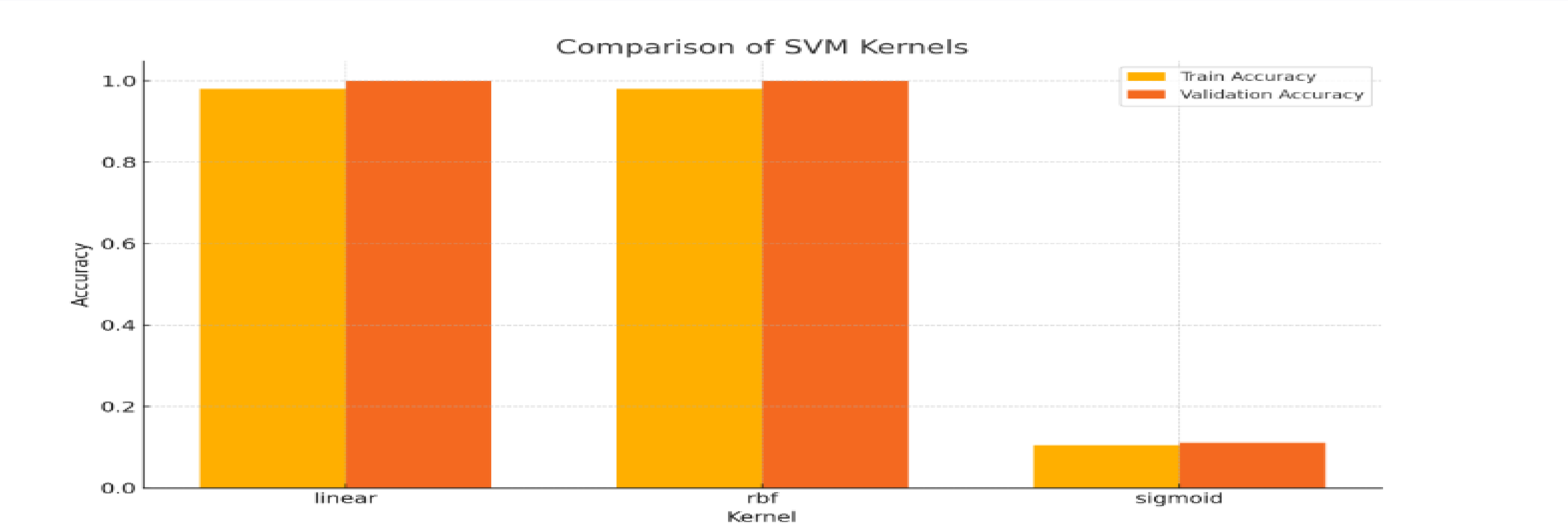
<Dashboard Screenshot 3>



Section 5

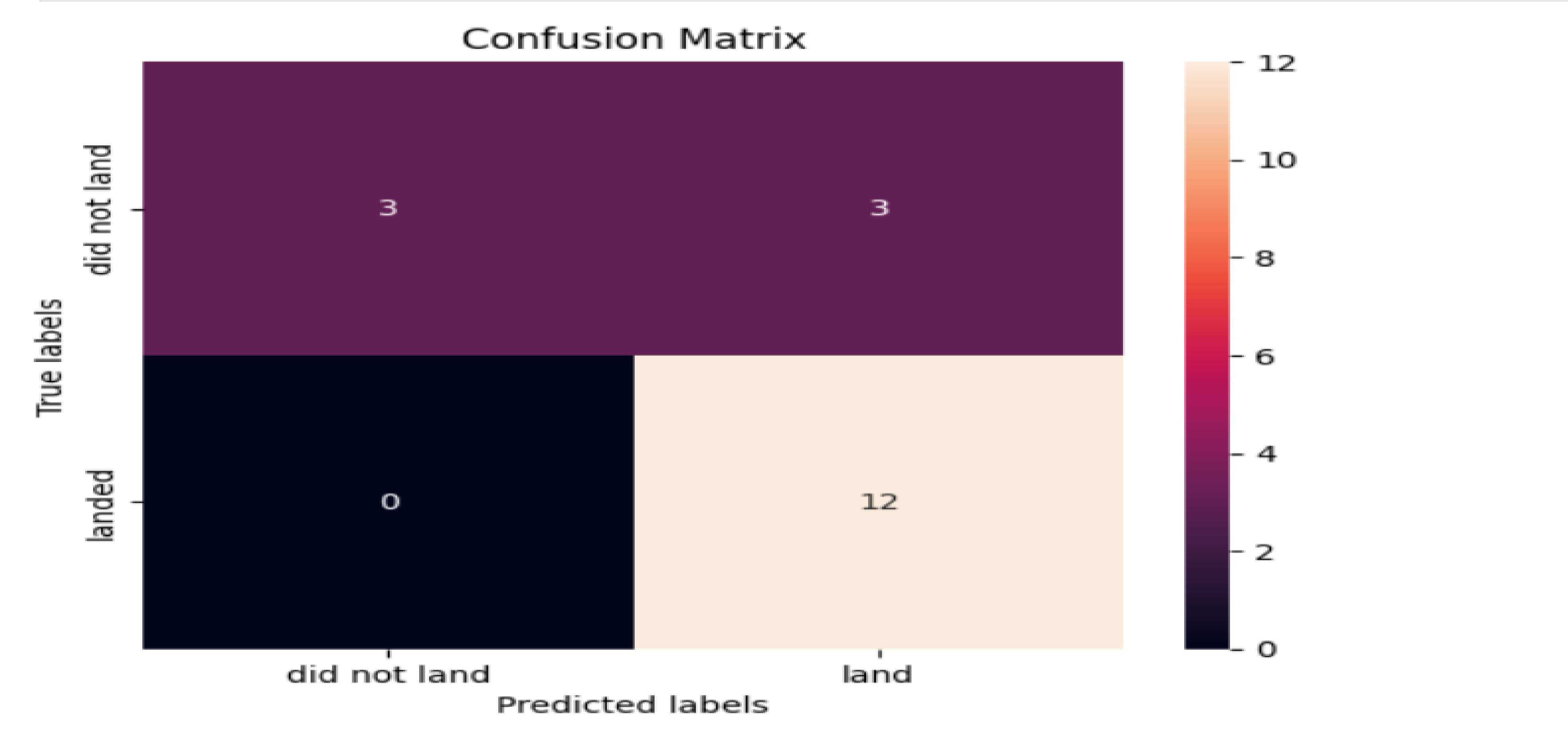
Predictive Analysis (Classification)

Classification Accuracy



Best Model based on Test Accuracy: SVM

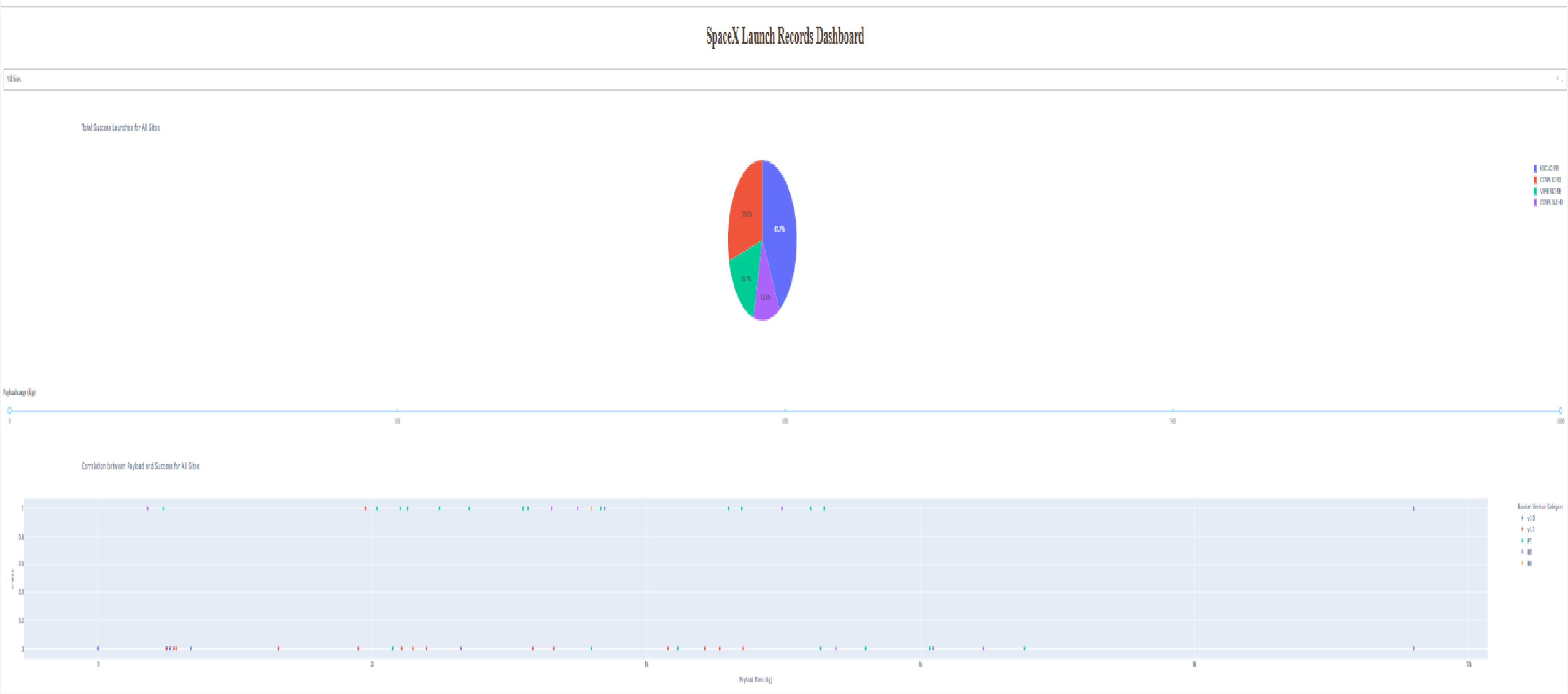
Confusion Matrix



Conclusions

•**Summary**: The project successfully demonstrated how data science methodologies can be applied to space launch operations, providing valuable insights and predictive models.
•**Future Work**: Ongoing model refinement and implementation of real-time data for better prediction accuracy.
•Successfully collected and wrangled real-world launch data from multiple sources.
•Explored the dataset using SQL queries to extract key insights and identify important patterns.
•Applied data visualization techniques to better understand relationships between variables.
•Built and evaluated predictive models to forecast launch outcomes and payload success rates.
•Gained valuable experience in data collection, data cleaning, SQL querying, statistical analysis, and machine learning modeling.
•Demonstrated the ability to turn raw data into meaningful business insights for a private space launch company.

Appendix


SpaceX Launch Records Dashboard
All Sites
Total Success Launches for All Sites
Correlation between Payload and Success for All Sites
Payload Mass (kg)

Thank you!