Predictions of Rocket Launches' Cost Based on First Stage Reusability

Amanda G. Braga

04-Dez-23

Outline

Executive Summary

- A succinct summary encapsulating the primary findings and significant insights derived from the analysis.

Introduction

- Providing an overview of the project's background, objectives, and the dataset used for analysis.

Methodology

- Detailing the step-by-step processes involved in data collection, wrangling, exploratory data analysis, interactive visual analytics, predictive analysis, and classification methodologies.

Results

- Presenting the key findings and observations obtained from the analysis, encompassing various aspects such as launch success, payload impact, site-specific trends, and orbital success rates.

Conclusion

- Summarizing the essential insights and concluding remarks derived from the analysis, highlighting optimal models, payload impact, site placements, success rate trends, and specific successful orbits.

Executive Summary

Methodologies

- **Data Collection**
- **Data Wrangling**
- **Exploratory Data Analysis (EDA)**
  - EDA with SQL
  - EDA with Data Visualization
- **Interactive Visual Analytics**
  - Interactive Map with Folium
  - Dashboard with Plotly Dash
- **Predictive Analysis - Classification Models**

Results

- **Exploratory Data Analysis**
  - Identified trends in payload mass
  - Launch success rates
  - Site-specific success ratios.
- **Interactive Analytics Demo**
  - Showcased interactive screenshots
  - Demonstrated insights via Folium and Plotly Dash.
- **Predictive Analysis Results:**
  - Evaluated model performances
  - Highlighting Decision Tree as the best-performing algorithm.

Introduction

In the dynamic realm of commercial space travel, SpaceX is pioneering accessible and affordable space exploration. SpaceX's success is underpinned by its groundbreaking Falcon 9 rocket's reusability, driving launch costs down to $62 million, in contrast to competitors charging over $165 million per launch.

The project aims to predict the success of the first stage's landing using publicly available data and machine learning models. The background and context are crucial:

SpaceX's cost-saving reusability.

Our startup's goal to compete with SpaceX.

Three key questions:

1. Factors influencing first-stage landing success.
2. Temporal trends in landing success.
3. Optimal binary classification algorithm.

This project empowers us to compete with SpaceX effectively and navigate the evolving commercial space landscape.

Methodology

Data Collection:

- Utilized SpaceX Rest API and Web Scraping from Wikipedia for comprehensive data acquisition.

Data Wrangling:

- Processed the gathered data, including filtering and addressing missing values.
- Employed One Hot Encoding to prepare the dataset for binary classification tasks.

Exploratory Data Analysis (EDA):

- Conducted EDA using visualization techniques to unveil patterns and insights.
- Leveraged SQL queries to extract and analyze specific information from the dataset.

Interactive Visual Analytics:

- Engaged in interactive visual analytics using Folium for geographical insights.
- Developed dynamic visualizations using Plotly Dash for comprehensive data exploration.

Predictive Analysis - Classification Models:

- Executed predictive analysis by constructing, fine-tuning, and evaluating classification models to attain optimal results.

Data Collection and Data Wrangling Methodology

The project's data collection involved a two-fold approach utilizing REST API and web scraping from Wikipedia. These methods provided comprehensive data essential for our detailed analysis.

GitHub URL:

REST API

Web Scraping

REST API Data Collection

Initiated data collection through SpaceX's REST API via GET requests.

Decoded API responses as JSON and converted them into a Pandas DataFrame using *json_normalize()*.

Extracted relevant launch information using custom functions.

Filtered the DataFrame to isolate Falcon 9 launches.

Transformed the dictionary into a DataFrame.

Aggregated the obtained data into a dictionary.

Handled missing values in the Payload Mass column by imputing the mean.

Exported the cleaned data to CSV for further analysis.

Web Scraping Data Collection

1. Retrieved Falcon 9 launch data from SpaceX's Wikipedia entry.
2. Created a *BeautifulSoup* object from the HTML response.
3. Extracted column names from the HTML table header.
4. Collected data by parsing HTML tables.
5. Constructed the collected data into a dictionary.
6. Transformed the dictionary into a DataFrame.
7. Exported the data to CSV, complementing the REST API dataset.

Data Wrangling Methodology

GitHub URL:
Data Wrangling

In the data wrangling phase, the focus was on refining and preparing the dataset for analysis.

- **Exploratory Data Analysis (EDA) and Training Labels:**
  - Conducted EDA to gain insights into the dataset.
  - Determined Training Labels by converting booster landing outcomes into a binary format: "1" for successful landings and "0" for unsuccessful ones.
- **Launch Site Analysis:**
  - Calculated the number of launches at each site to understand distribution and frequency.
- **Orbit Analysis:**
  - Computed the number and occurrence of each orbit to identify patterns and trends.
- **Mission Outcome per Orbit Type:**
  - Analyzed the relationship between mission outcomes and orbit types.
  - Calculated the number and occurrence of mission outcomes for each orbit type.
- **Landing Outcome Label Creation:**
  - Derived a landing outcome label from the "Outcome" column.
  - Introduced binary labels: "1" for successful landings and "0" for unsuccessful ones.
- **Data Export:**
  - Completed the wrangling process by exporting the refined data to a CSV file for subsequent analysis.

EDA and Interactive Visual Analytics Methodology

In the pursuit of comprehensive insights, the exploratory data analysis (EDA) journey seamlessly blended traditional visualizations and advanced interactive tools. Through diverse charts, including Flight Number vs. Payload Mass and Orbit Type vs. Success Rate, intricate relationships were dissected within the dataset. Notably, scatter plots unveiled nuanced variable connections, laying the groundwork for potential machine learning integrations. Supplementing this visual approach, the SQL queries unearthed unique launch site names and diverse statistical summaries. Transitioning from static analyses, the power of Folium was harnessed to construct an interactive map, showcasing launch sites, their outcomes, and spatial proximities. Elevating the analytics, an intuitive Plotly Dash dashboard was crafted, uniting diverse visualizations and enabling dynamic exploration of launch data, payload specifics, and success rates. This methodology facilitated a holistic understanding, seamlessly marrying traditional EDA with cutting-edge visual analytics tools.

EDA with SQL

GitHub URL:
EDA with SQL

Executed SQL queries for diverse insights, including unique launch site names, records with launch sites starting with 'CCA,' total payload mass by NASA (CRS), average payload mass for booster version F9 v1.1, and more.

Extracted information like the date of the first successful ground pad landing, boosters with success in drone ship and specific payload mass ranges, and the ranking of landing outcomes between specified dates.

Name of all Launch Sites

```
%sql select distinct LAUNCH_SITE as "Launch_Sites" from SPACEXTBL;
```

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

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

Insights

Space Mission Launch Sites

- This display highlights the unique names of all launch sites involved in the space mission. It provides a comprehensive overview of the diverse locations utilized for launching missions within the space program. Space Mission Launch Sites: This display highlights the unique names of all launch sites involved in the space mission. It provides a comprehensive overview of the diverse locations utilized for launching missions within the space program.

Launch Sites starting with the String 'CCA'

```
%sql select LAUNCH_SITE from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Insights

Filtered Launch Site Records

- Displaying five records where launch sites commence with 'CCA' provides a targeted view of specific launch sites within the space mission. This focused display offers insights into a subset of sites sharing a common naming convention.

Total Payload Mass

```
%sql select sum (PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)';

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

sum (PAYLOAD_MASS__KG_)
45596

Insights

Aggregate Payload Mass by NASA (CRS)

- This display reveals the total payload mass collectively carried by boosters specifically launched by NASA under the CRS program. It provides a cumulative measure of payloads associated with NASA's Commercial Resupply Services missions.

Total Payload Mass for F9 v1.1

```
%sql select sum (PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1';
```

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

sum (PAYLOAD_MASS__KG_)
14642

Insights

Average Payload Mass - F9 v1.1

- This display presents the average payload mass carried specifically by booster version F9 v1.1. It offers a representative measure of the typical payload weight associated with this booster version across missions.

First Successful Landing

```
%sql select min (DATE) as "First Succesful Landing" from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)';
```

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

First Succesful Landing
2015-12-22

Insights

First Successful Ground Pad Landing

- This display reveals the significant milestone of the first successful landing outcome achieved on a ground pad. It marks a pivotal event in the space program, signifying the inception of successful ground pad landings within the missions.

Drone Ship Success (Payload 4000kg-6000kg)

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pad)' and PAYLOAD_MASS__KG_ > 4000 and P
```

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

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Insights

Successful Drone Ship Landings - Payload Range

- This display presents the names of boosters that achieved successful landings on a drone ship with a payload mass greater than 4000kg but less than 6000kg. It showcases specific boosters that effectively completed missions within this payload range via drone ship landings.

Mission Outcomes

```
%sql select sum(case when MISSION_OUTCOME like '%Success%' then 1 else 0 end) as "Successful Mission", sum(case when MISSIO
```

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

Successful Mission	Failure Mission
100	1

Insights

Mission Outcome Summary

- This display provides a summary of the total number of successful and failed mission outcomes. It offers a comprehensive count of missions categorized as successful or failed, providing an overview of the mission success rate.

Booster Versions which carried the Max Payload Mass

```
%sql select distinct BOOSTER_VERSION as "Booster Versions which carried the Max Payload Mass" from SPACEXTBL where PAYLOAD
```

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

Booster Versions which carried the Max Payload Mass
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

Insights

Maximum Payload Carriers - Booster Versions

- This display utilizes a subquery to showcase the names of booster versions that carried the maximum payload mass. It highlights specific booster versions that achieved the highest payload capacities within the dataset.

Rank Landing Outcomes (2010/06/04 - 2017/03/20)

```
%sql select LANDING_OUTCOME as "Landing Outcome", count(LANDING_OUTCOME) as "Total Count" from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' \
group by LANDING_OUTCOME \
order by count(LANDING_OUTCOME) desc;
```

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

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

Insights

Ranked Landing Outcomes - Date Range

- This display ranks the count of landing outcomes, including categories like Failure (drone ship) or Success (ground pad), between June 4, 2010, and March 20, 2017, in descending order. It offers a ranked overview of the frequency of these landing outcomes during the specified time frame.

EDA with Data Visualization

GitHub URL:
EDA with Data Visualization

Utilized various charts, including Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs. Orbit Type, and Success Rate Yearly Trend.

Employed scatter plots to examine variable relationships, providing insights for potential integration into machine learning models.

Utilized bar charts for discrete category comparisons, illustrating relationships between specific categories and measured values.

Employed line charts to depict trends over time, especially in the context of a time series.

Flight Number vs. Launch Site

Insights

Temporal Success

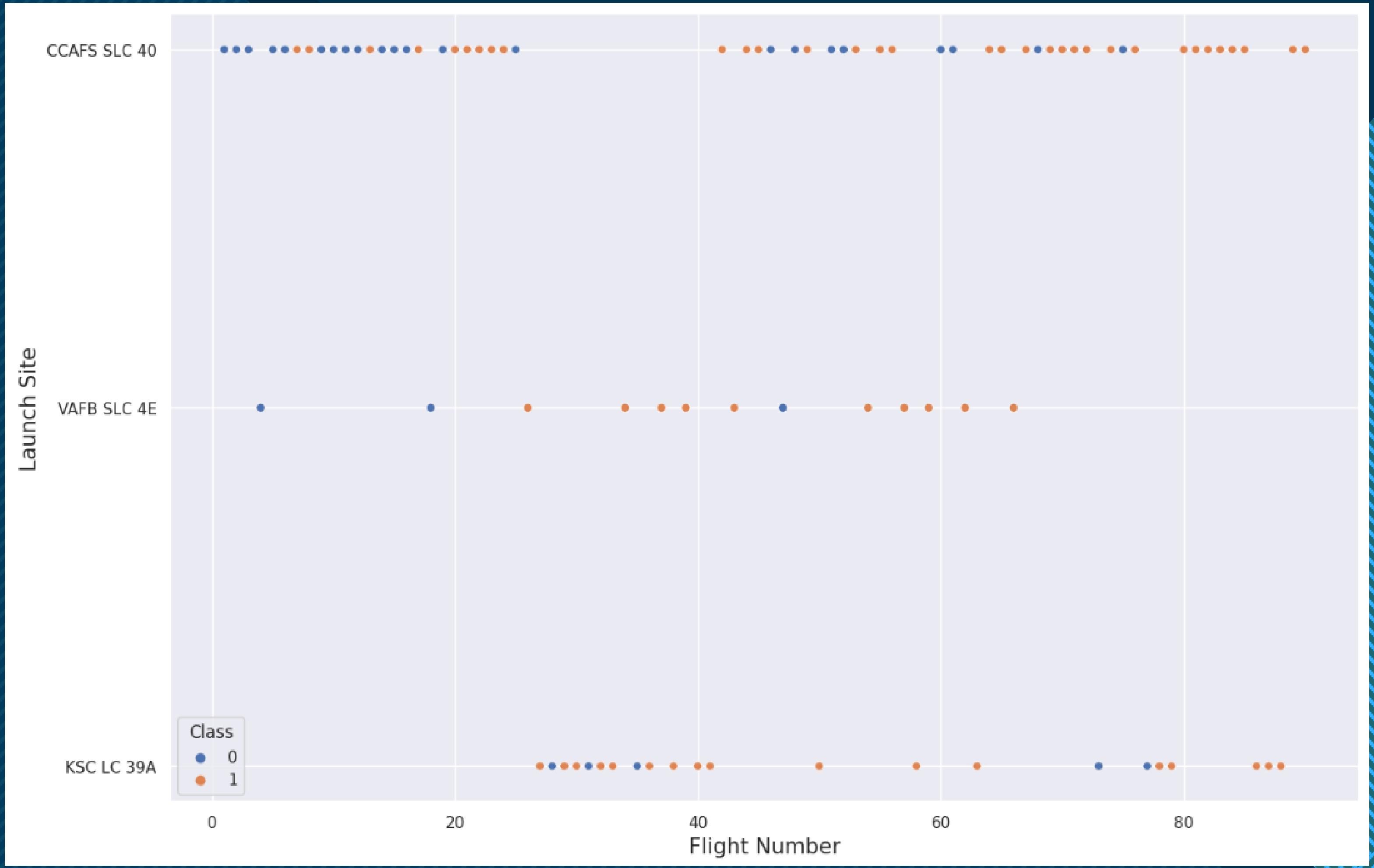
- Early flights exhibit higher failure rates, while recent flights trend towards consistent success.

Launch Site Impact

- CCAFS SLC 40 hosts a significant portion of launches, while VAFB SLC 4E and KSC LC 39A show higher success rates.

Progressive Success

- A noticeable trend indicates improving success rates with newer launches.



Payload vs. Launch Site

Insights

Payload Impact

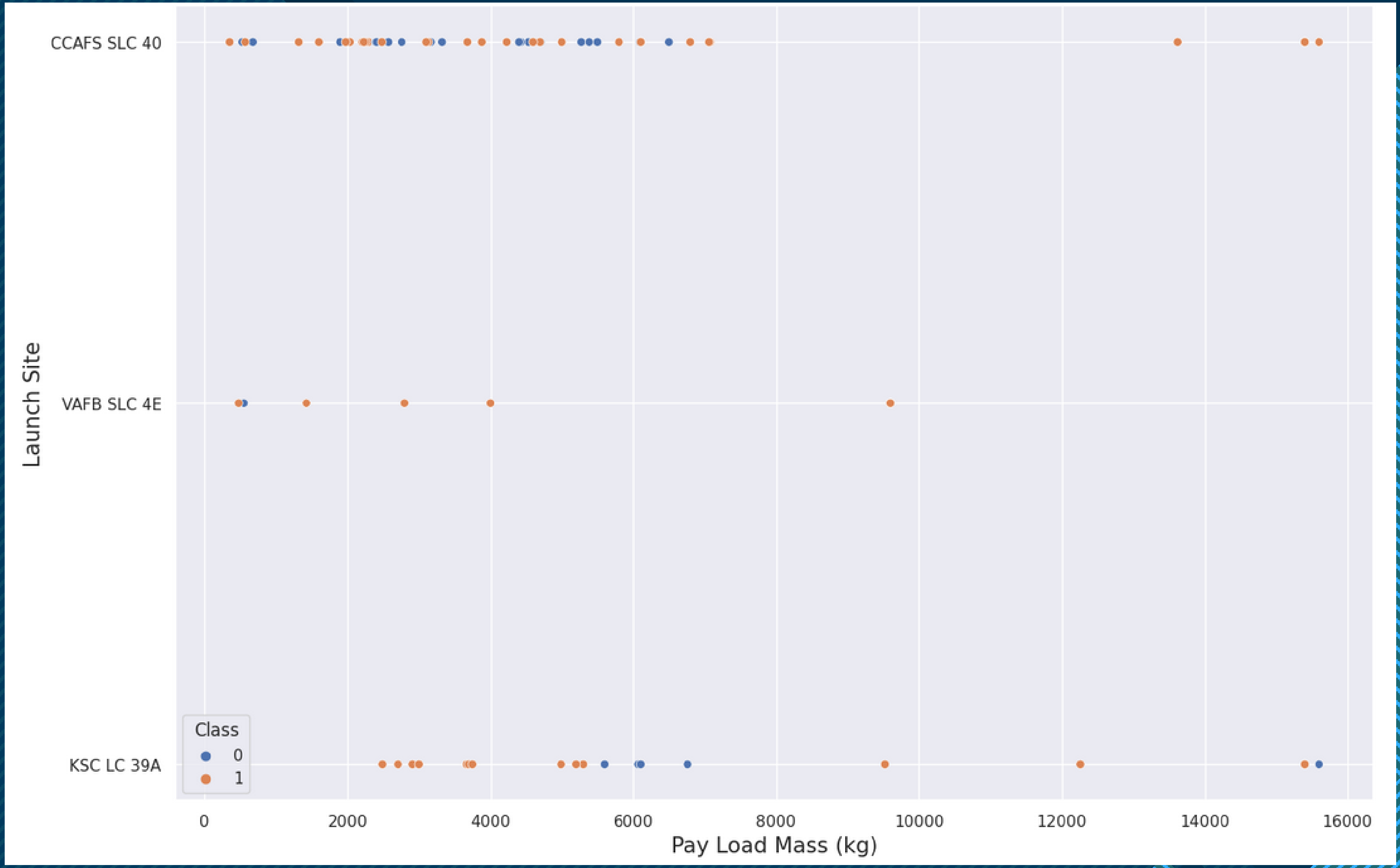
- Across all launch sites, higher payload masses correlate with increased success rates, indicating a positive relationship between payload mass and launch success.

Threshold for Success

- Launches with a payload mass exceeding 8000 kg predominantly resulted in successful missions, suggesting a potential threshold for optimal mission success.

Site-Specific Success

- KSC LC 39A boasts a 100% success rate for lower payload masses.



Success Rate vs. Orbit type

Insights

High Success Orbits

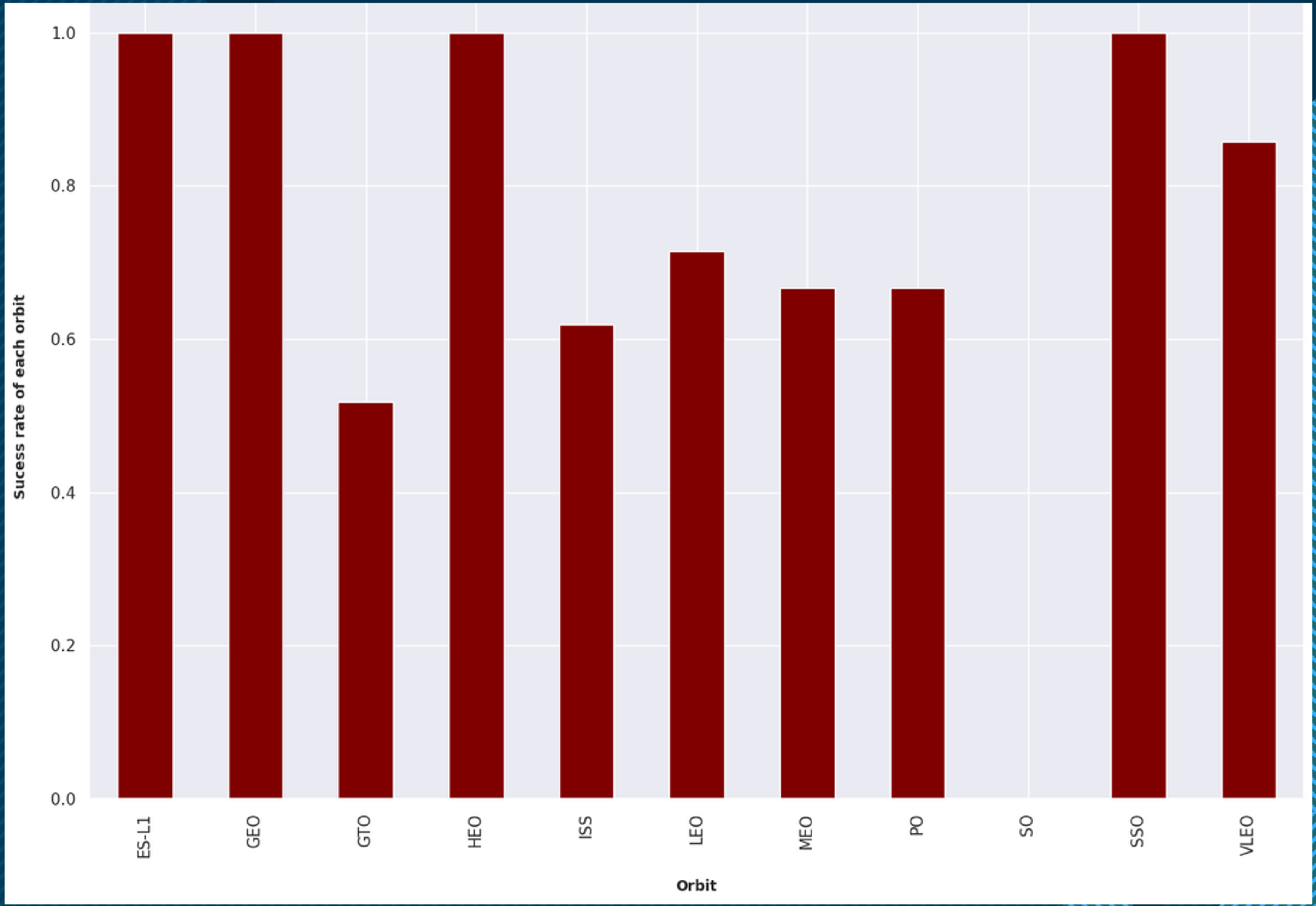
- Certain orbits—ES-L1, GEO, HEO, and SSO—consistently exhibit a 100% success rate, suggesting reliability and efficacy in mission outcomes.

Zero Success Orbits

- Notably, the SO orbit presents a 0% success rate, indicating significant challenges or complexities associated with missions in this orbit.

Moderate Success Orbits

- Orbits like GTO, ISS, LEO, MEO, and PO demonstrate success rates ranging between 50% and 85%, indicating a moderate level of mission success variability within these orbital parameters.



Flight Number vs. Orbit type

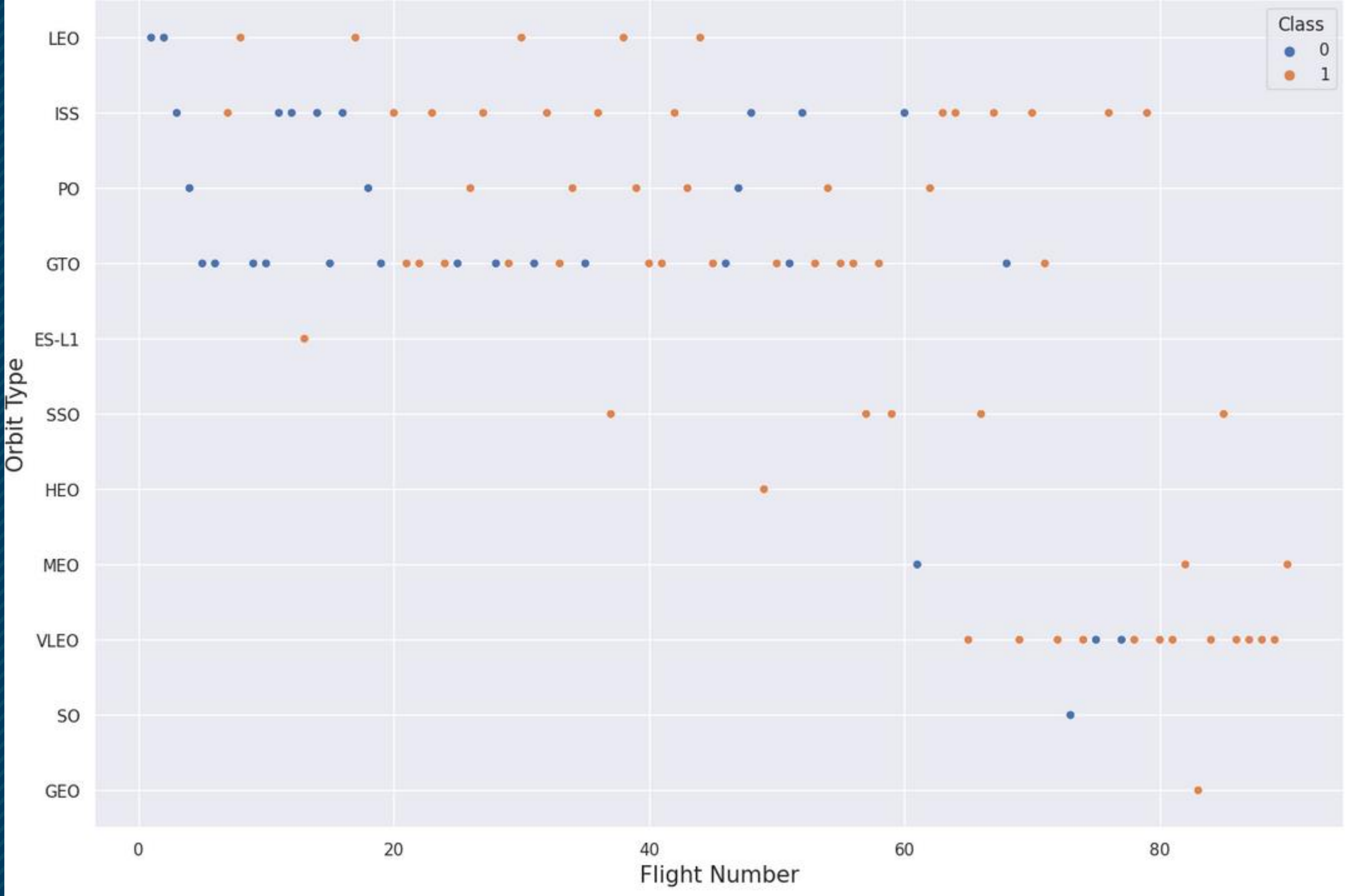
Insights

LEO Orbit Dynamics

- Within the LEO orbit, a visible relationship emerges between the number of flights and success rates. Higher flight numbers correspond to increased success, implying a potential learning curve or operational improvements over successive missions.

GTO Orbit Contrast

- Conversely, in the GTO orbit, no apparent correlation exists between flight number and success rates. This suggests a distinct operational landscape where flight frequency may not significantly impact mission success within this orbit.



Payload Mass vs. Orbit type

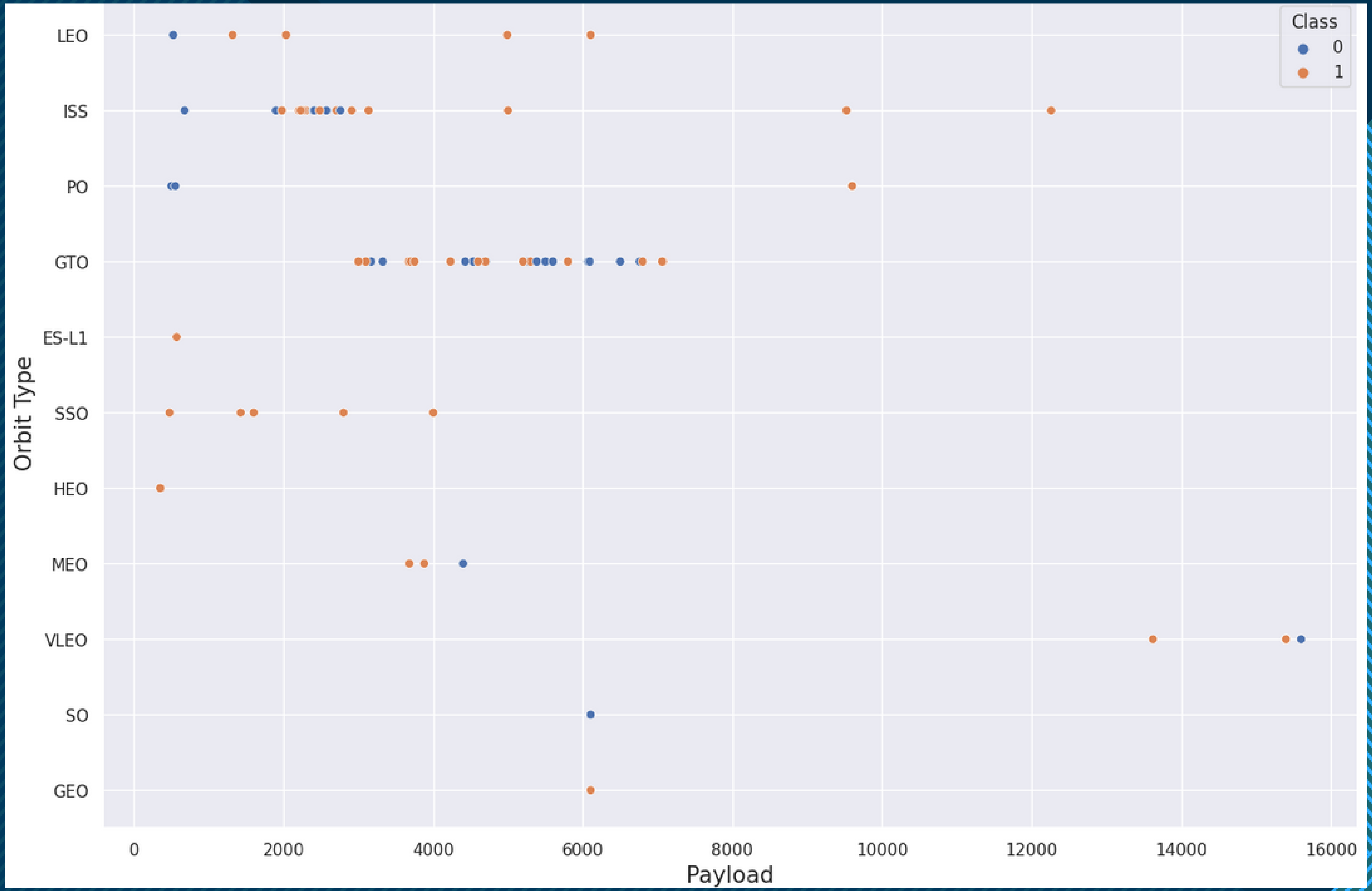
Insights

Payload Influence on Success

- Heavy payloads demonstrate higher success rates in Polar, LEO, and ISS orbits. The correlation suggests that these orbits exhibit more favourable outcomes for missions with heavier payloads.

GTO Complexity

- In the GTO orbit, distinguishing success rates based on payload mass becomes challenging. Both positive (successful) and negative (unsuccessful) landing rates are observable, indicating complexity and variability in mission outcomes within this orbit, regardless of payload mass.



Launch Success vs. Yearly Trend

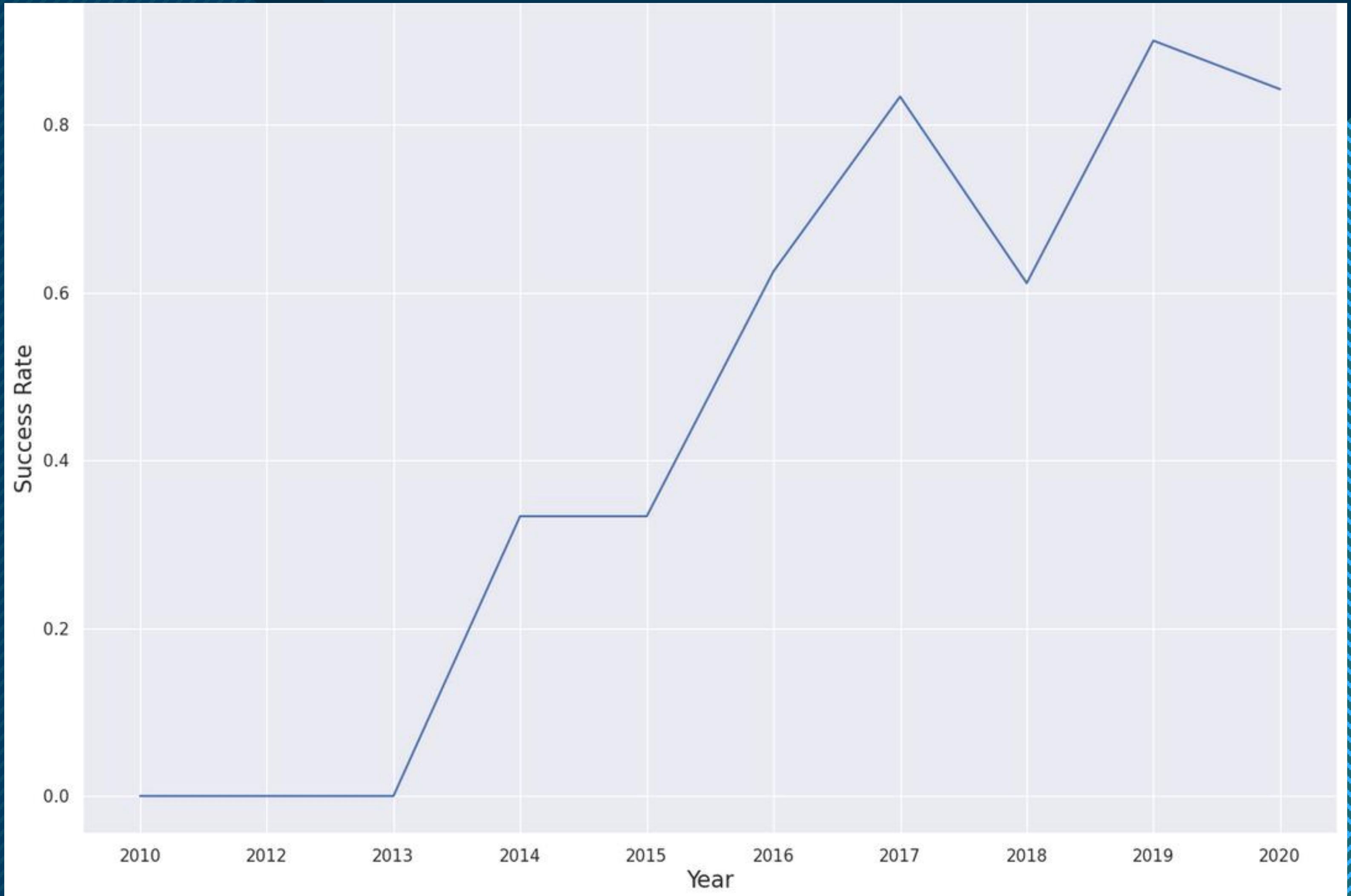
Insights

Progressive Success

- The trend showcases a consistent increase in success rates from 2013 to 2020, indicating an overall positive trajectory in mission success over the years.

Notable Dip in 2018

- However, a discernible dip in success rates is observed in 2018, interrupting the otherwise upward trend, suggesting potential anomalies or challenges faced during missions in that particular year.



Interactive Map with Folium

GitHub URL:
Interactive Map with Folium

Created markers for all launch sites, each displaying relevant information.

Utilized colored markers to signify success (green) and failure (red) launches, aiding in identifying sites with higher success rates.

Implemented colored lines to display distances between launch sites and proximities, enhancing geographical understanding.

All Launch Sites on a Map

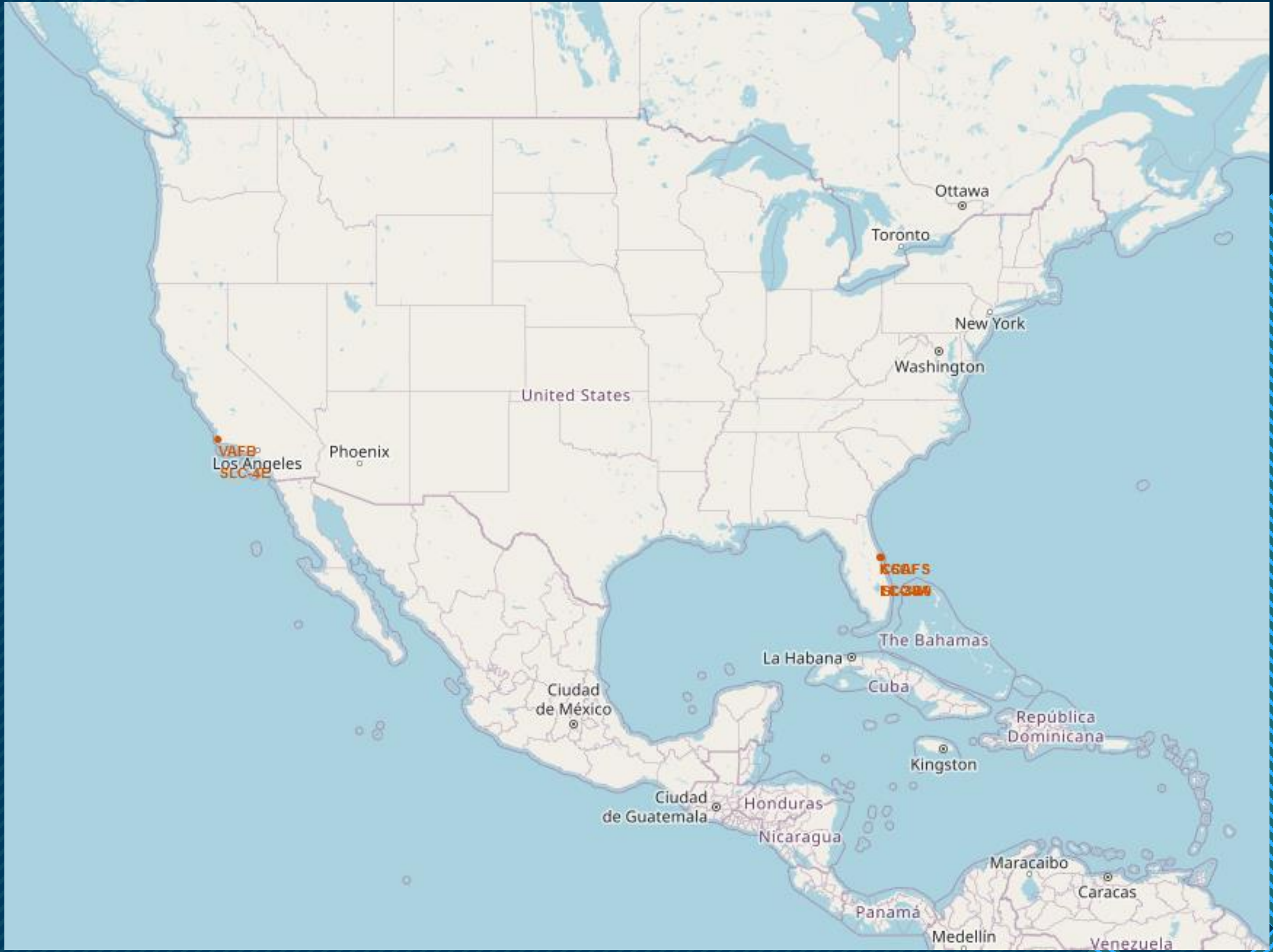
Insights

Equatorial Proximity

- Observing most launch sites in proximity to the Equator highlights the strategic location choice. The Equator's faster rotational speed (1670 km/hour) contributes to initial spacecraft velocity, aiding in achieving and maintaining orbits through inertia.

Coastal Placement

- Notably, all launch sites are situated very close to coastlines. Launching rockets towards the ocean significantly mitigates risks associated with potential debris or explosions, reducing the likelihood of hazardous occurrences near populated areas.



Launch Records by Colour on the map

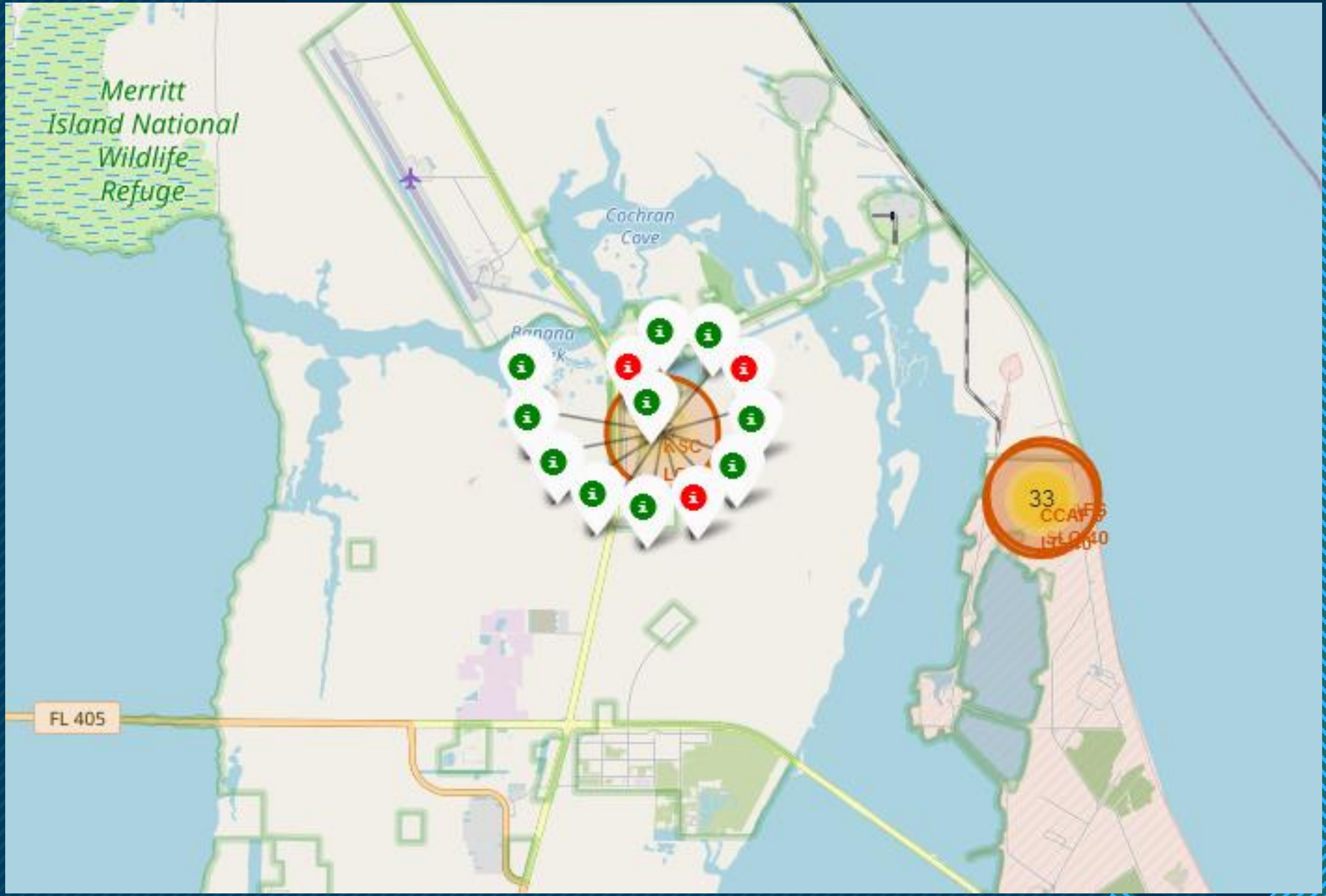
Insights

Colour-Coded Success Indicators

- The colour-labelled markers offer a straightforward visual identification of launch outcomes. Green markers denote successful launches, while red markers signify failed launches, enabling quick identification of sites with relatively high success rates.

High Success Rate Highlight

- The specific observation of Launch Site KSC LC-39A having a notably high success rate stands out amidst the markers, indicating its consistent track record of successful launches.

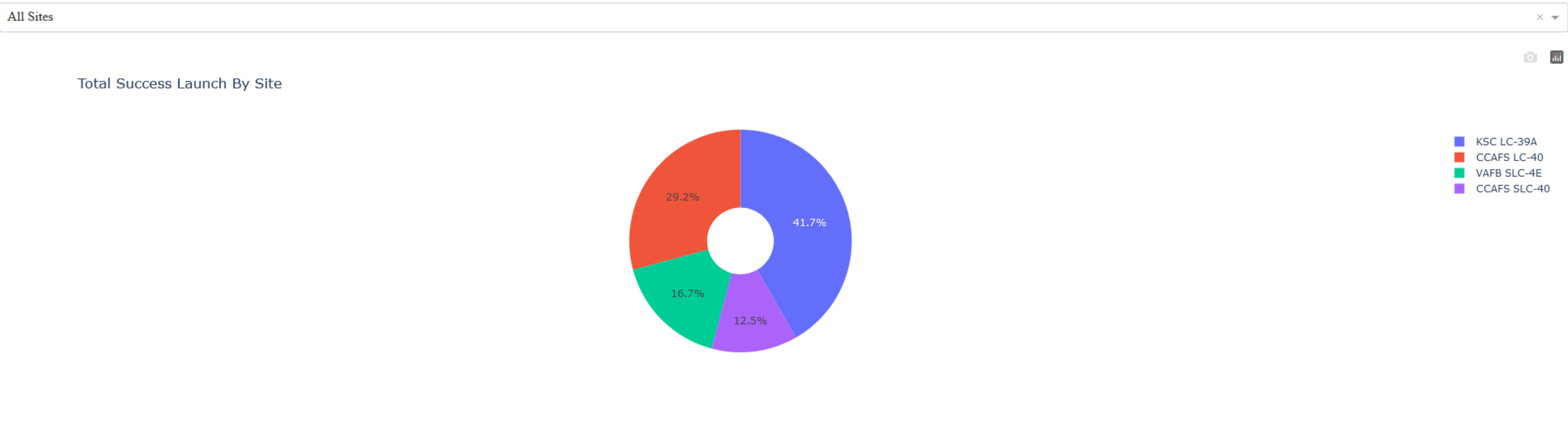


Dashboard with Plotly Dash

GitHub URL:
Dashboard with Plotly Dash

- Developed a user-friendly dashboard with features such as a Launch Sites Dropdown List for site selection.
- Integrated a pie chart illustrating total successful launches and site-specific success vs. failure counts.
- Implemented a slider for Payload Mass Range selection.
- Incorporated a scatter chart illustrating the correlation between Payload and Launch Success for various Booster Versions.

Total Success Launch by Site

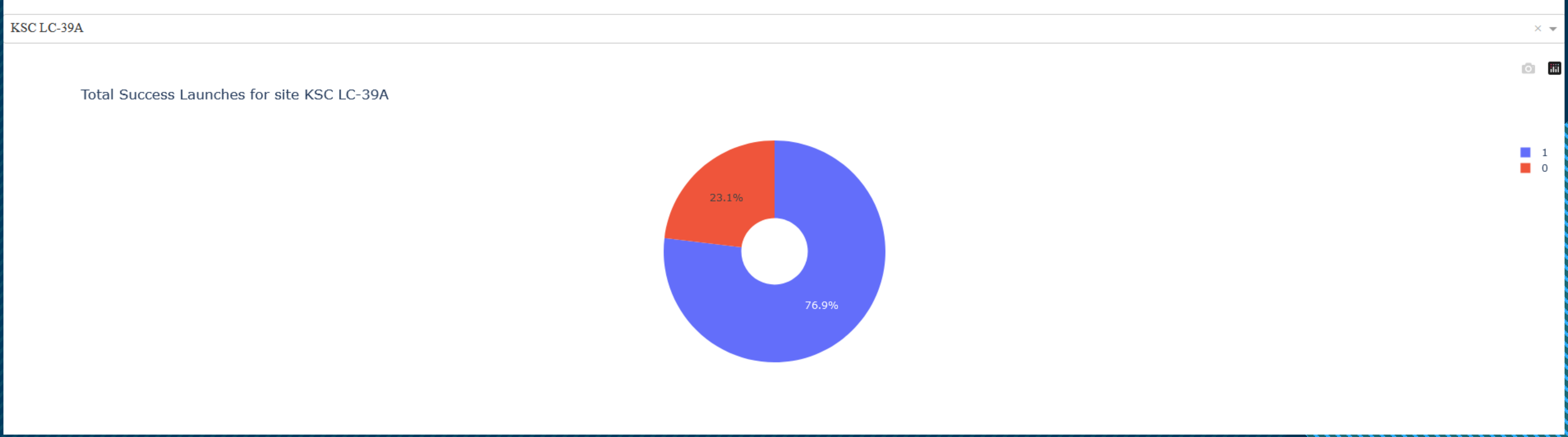


Insights

Site-Specific Success Comparison

- The chart prominently illustrates that among all sites, KSC LC-39A stands out with the highest count of successful launches. Its success count nearly doubles that of the second most successful site, demonstrating its significantly higher success rate in comparison to other launch sites.

Total Success Launch for Site KSC LC-39A



Insights

Exceptional Success Rate - KSC LC-39A

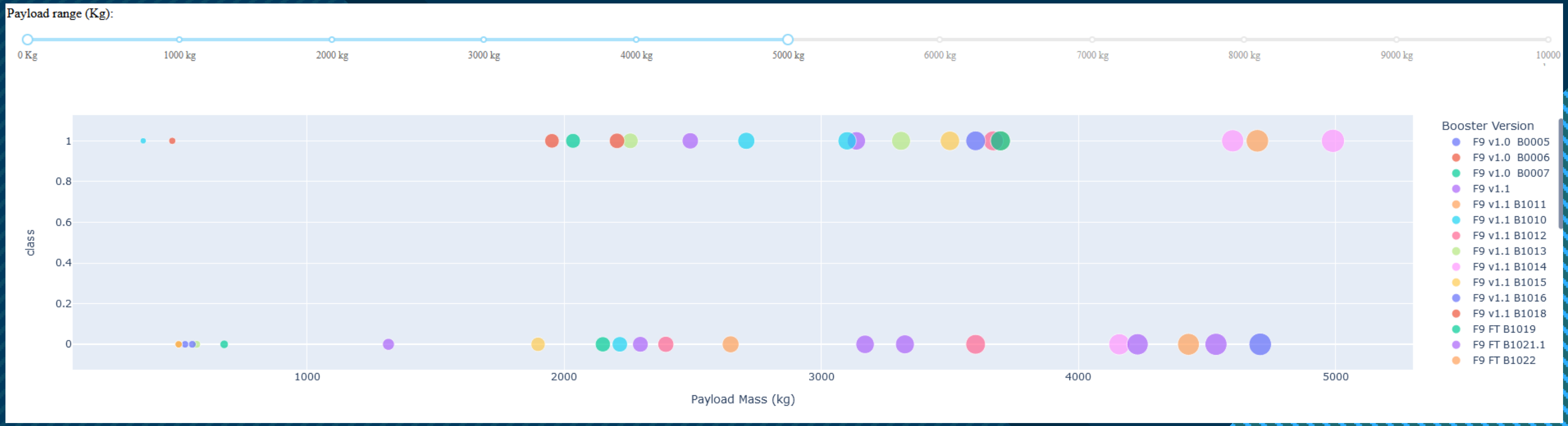
- The data reveals that KSC LC-39A boasts the highest launch success rate among the sites, standing at 76.9%. This site achieved 10 successful landings, significantly outnumbering the mere 3 failed landings, showcasing its robust success rate.

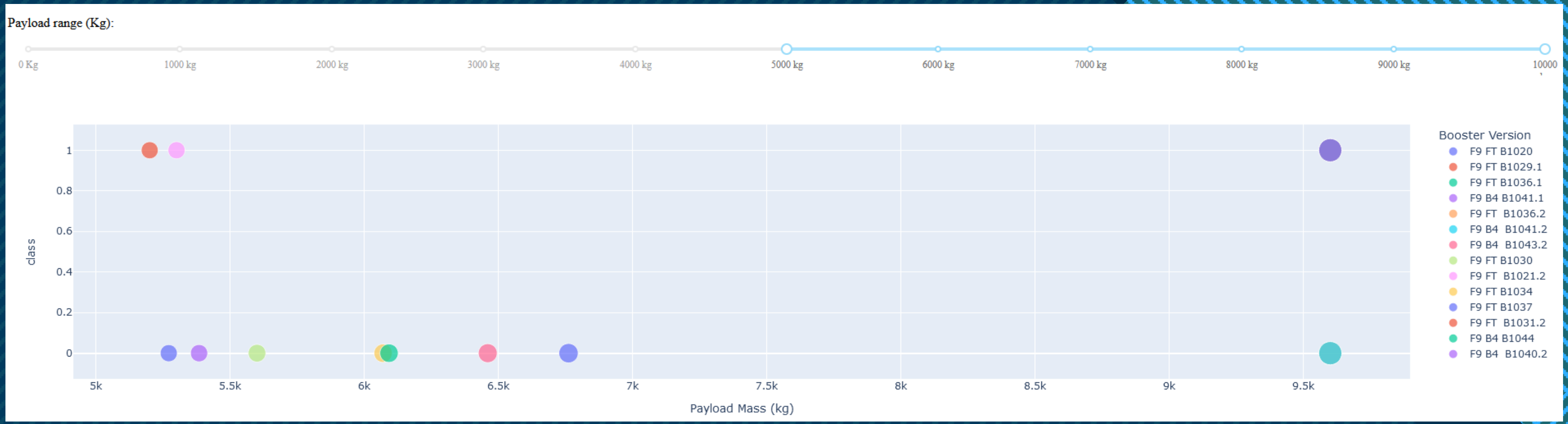
Payload Mass vs. Launch Outcome (All Sites)

Insights

Optimal Payload Success Range

- The scatter plot charts showcase that payloads ranging between 2000kg and 5500kg exhibit the highest success rates. This range demonstrates a consistently higher success rate compared to other payload masses, suggesting an optimal range for successful launches across all sites.





Predictive Analysis Methodology

In the pursuit of insights, the methodology employed classification techniques to predict categorical outcomes.

GitHub URL:

REST API

Web Scraping

Predictive Analysis Methodology:

GitHub URL:
Classification

Data Preparation:	Created a NumPy array from the "Class" column, serving as the target variable for classification.
Standardization and Transformation:	Standardized the dataset using StandardScaler to ensure uniform scaling, fitting, and transforming it for optimal model performance.
Data Splitting:	Segmented the dataset into training and testing sets using the train_test_split function, vital for model evaluation and validation.
Hyperparameter Tuning:	Established a GridSearchCV object with a cross-validation of 10 folds, seeking the most optimal parameters for enhanced model performance.
Model Evaluation:	Applied GridSearchCV across LogReg, SVM, Decision Tree, and KNN models, exploring their performance on the dataset.
Accuracy Assessment:	Computed the accuracy scores using the .score() method to evaluate model performance on the test data
Confusion Matrix Examination:	Analyzed the confusion matrices for all models, providing insights into their predictive capabilities and error patterns.
Performance Metrics Analysis:	Leveraged Jaccard score and F1 score metrics to identify the superior-performing model based on evaluation criteria.

Scores Accuracy

Insights

Test Set Ambiguity

- The scores from the Test Set alone do not distinctly indicate the best-performing method, potentially due to limitations from the small sample size (18 samples). To resolve this ambiguity, evaluation across the entire dataset was conducted.

Decision Tree Model Superiority

- Assessment across the entire dataset validates the Decision Tree Model as the best-performing model. This model not only yields higher scores but also showcases the highest accuracy among the tested methods.

Score from Test Sets

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.529412	0.800000
F1_Score	0.888889	0.888889	0.692308	0.888889
Accuracy	0.833333	0.833333	0.555556	0.833333

Score from Whole Data Set

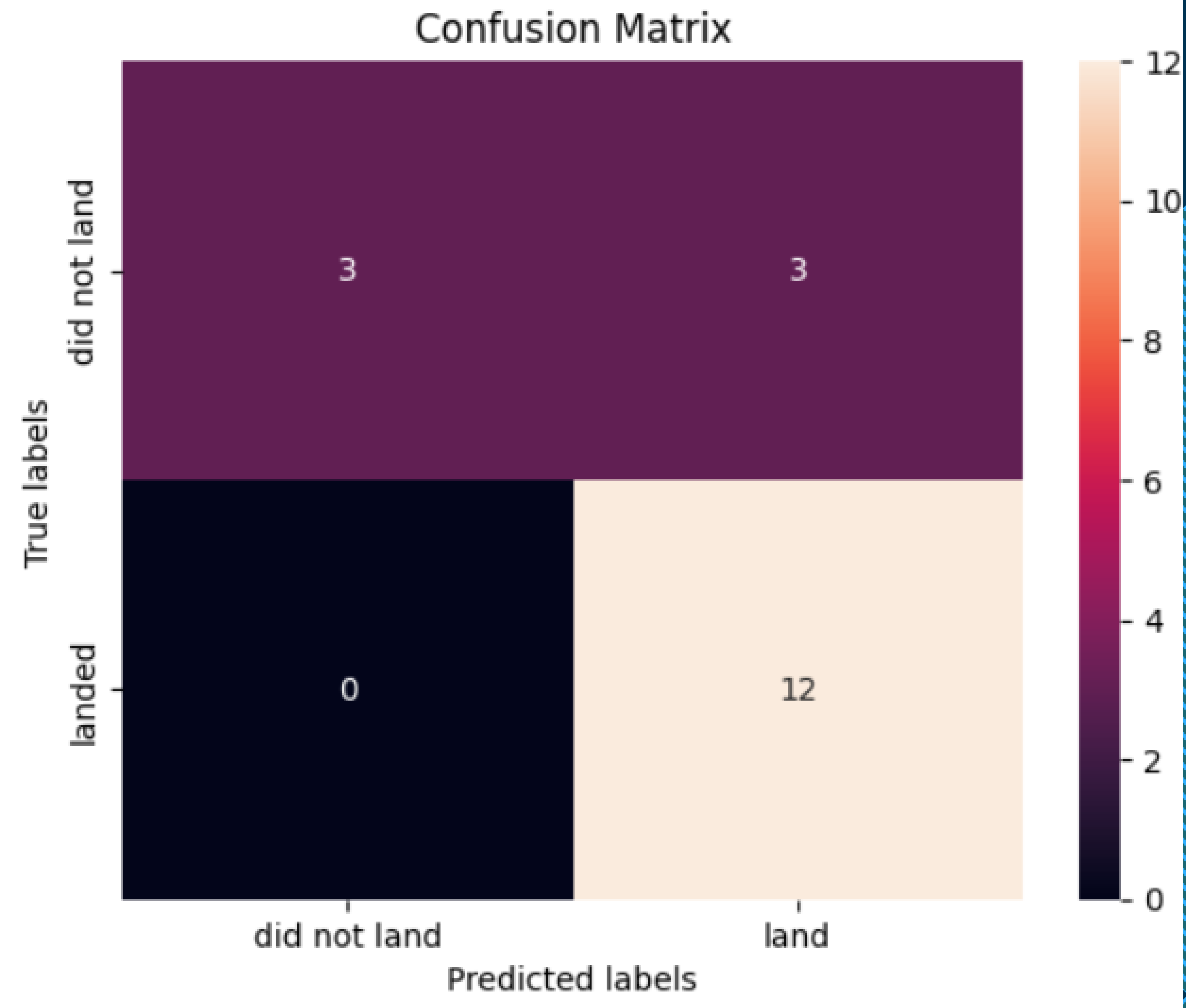
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.750000	0.819444
F1_Score	0.909091	0.916031	0.857143	0.900763
Accuracy	0.866667	0.877778	0.777778	0.855556

Confusion Matrix

Insights

Logistic Regression Class Differentiation

- The examination of the confusion matrix indicates that logistic regression effectively distinguishes between different classes. However, a notable issue arises with false positives, signifying instances where the model incorrectly predicts positive outcomes when they are negative.



Results

Data Collection:

- Successfully gathered comprehensive data from SpaceX Rest API and Wikipedia through web scraping, ensuring a rich dataset for analysis.

Data Wrangling:

- Efficiently processed and cleaned the dataset, addressing missing values and preparing it for binary classification using One Hot Encoding.

Exploratory Data Analysis (EDA):

- Unveiled crucial insights and patterns through visualizations, offering a comprehensive understanding of the dataset.
- Extracted targeted information using SQL queries, providing nuanced insights.

Interactive Visual Analytics:

- Constructed an interactive map using Folium, offering geographical insights into launch sites and outcomes.
- Developed an intuitive dashboard with Plotly Dash, enabling dynamic exploration of launch data and success rates.

Predictive Analysis - Classification Models:

- Successfully built, fine-tuned, and evaluated classification models, identifying optimal parameters for predictive accuracy.

Conclusion

Decision Tree Model Superiority

- The analysis concludes that the Decision Tree Model stands out as the optimal algorithm for this dataset, showcasing superior performance among the tested models.

Payload Mass Impact

- Notably, launches with lower payload masses exhibit higher success rates compared to those with larger payload masses, signifying a correlation between payload mass and launch success.

Strategic Site Placement

- The analysis reveals that most launch sites are strategically located near the Equator, leveraging its higher rotational speed for launch velocity. Additionally, all sites are positioned in close proximity to coastlines, minimizing risks associated with debris near populated areas.

Increasing Success Rates

- Over the years, there is a discernible trend of increasing success rates in launches, indicating advancements or improved efficiencies within the space program.

KSC LC-39A Success Dominance

- Among all sites, KSC LC-39A notably boasts the highest success rate, highlighting its consistent track record of successful launches.

Highly Successful Orbits

- Orbits ES-L1, GEO, HEO, and SSO maintain a perfect 100% success rate, underscoring the reliability and efficacy of missions within these specific orbital parameters.

Thank You