The Roadmap to the Space

Presented by Adama Boubacar Oumarou

December, 2024

Agenda

01 Executive Summary

02 Introduction

03 Methodology

04 Results

05 Conclusion

06 Appendix

Executive Summary

The project used several methodologies to collect, analyze, and visualize data. Data was gathered through API integration and web scraping techniques, followed by data wrangling to clean and prepare the information for analysis. Exploratory Data Analysis (EDA) was performed using SQL to uncover patterns, and additional insights were gained through data visualization. Interactive visual analytics were created using Folium, enhancing the data's geographical context. Machine learning models were developed for predictive analytics, providing forecasts based on the data.

The results from each step, including EDA findings, interactive analytics, and predictive outcomes, were summarized to present a comprehensive view of the analysis. Screenshots of the interactive analytics were included to illustrate the visual insights.

Introduction

Background

SpaceX, dominant in the space industry, aims to make space exploration more accessible and affordable. The company has achieved notable milestones, from sending spacecraft to the International Space Station to deploying a global satellite network for internet access and launching crewed space missions. A key factor in SpaceX's ability to offer relatively low-cost launches ($62 million each) is its innovative reuse of the Falcon 9 rocket's first stage. In contrast, other providers, lacking this capability, charge over $165 million per launch. By predicting whether the first stage will land successfully, we can better estimate launch costs. Leveraging public data and machine learning, we can develop models to forecast the likelihood of first-stage reuse, benefiting both SpaceX and potential competitors in the space launch market.

Questions

How do payload mass, launch site, number of flights, and orbit type influence first-stage landing success?
What is the trend in successful landings over time?
Which predictive model is most effective for forecasting successful landings (binary classification)?

Methodology

Data Collection

Data was collected using the SpaceX API and some web scraping techniques.

Data Wrangling

Data was prepared for analysis and modeling through cleaning(addressing values, filtering) and encoding.

Data Exploration

Explanatory Data Analysis was performed using SQL and some data vizualization techniques.

Data Visualization

Data was vizualized using folium and an interactive dashboard.

Models Building

Different predictive models are built and evaluated to choose the best performing.

Data Collection
SpaceX API

- Fetch rocket launch data from the SpaceX API
- Decode the API response using .json() and normalize it into a DataFrame with json_normalize().
- Use custom functions to request and process launch information.
- Structure the launch data into a dictionary.
- Convert the dictionary into a DataFrame.
- Filter the DataFrame to include only Falcon 9 launches.
- Handle missing values in the Payload Mass column by replacing them with the calculated mean.
- Export the processed data to a CSV file.

Data Collection
Web Scraping

- Fetch Falcon 9 launch data from Wikipedia.
- Use BeautifulSoup to parse the HTML response.
- Extract column names from the table headers in the HTML.
- Parse the HTML table rows to collect data.
- Organize the extracted data into a dictionary.
- Convert the dictionary into a DataFrame.
- Save the DataFrame to a CSV file.

Data Wrangling

- Perform Exploratory Data Analysis (EDA) to identify patterns and define training labels.
- Calculate Launch Statistics:
  1. Count the number of launches for each launch site.
  2. Determine the frequency of each orbit type.
  3. Analyze mission outcomes by orbit type, calculating their counts and occurrences.
- Generate a new label from the Outcome column to indicate landing success or failure.
- Export Processed Data

EDA with Visualization

Charts were created to visualize the data:

- 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
- Success Rate Yearly Trend

- Scatter plots were used to explore variable relationships, with patterns or correlations helping to guide machine learning features.

- Bar charts compared discrete categories, showing their relationships with measured values.

- Line charts captured trends over time, providing insights into time series patterns.

EDA with SQL

- Display the unique launch site names.
- Retrieve 5 records where launch site names start with 'CCA'.
- Calculate the total payload mass carried by NASA (CRS) boosters.
- Compute the average payload mass for booster version F9 v1.1 when launched from ground pads.
- Find the date of the first successful landing outcome.
- List booster names with successful drone ship landings and payloads between 4000 and 6000.
- Count the total successful and failed mission outcomes.
- Identify booster versions carrying the maximum payload mass.
- List failed drone ship landings in 2015, with booster versions and launch site names.
- Rank landing outcomes (e.g., Failure or Success) by count within a specific date range.

Data Visualization with Folium

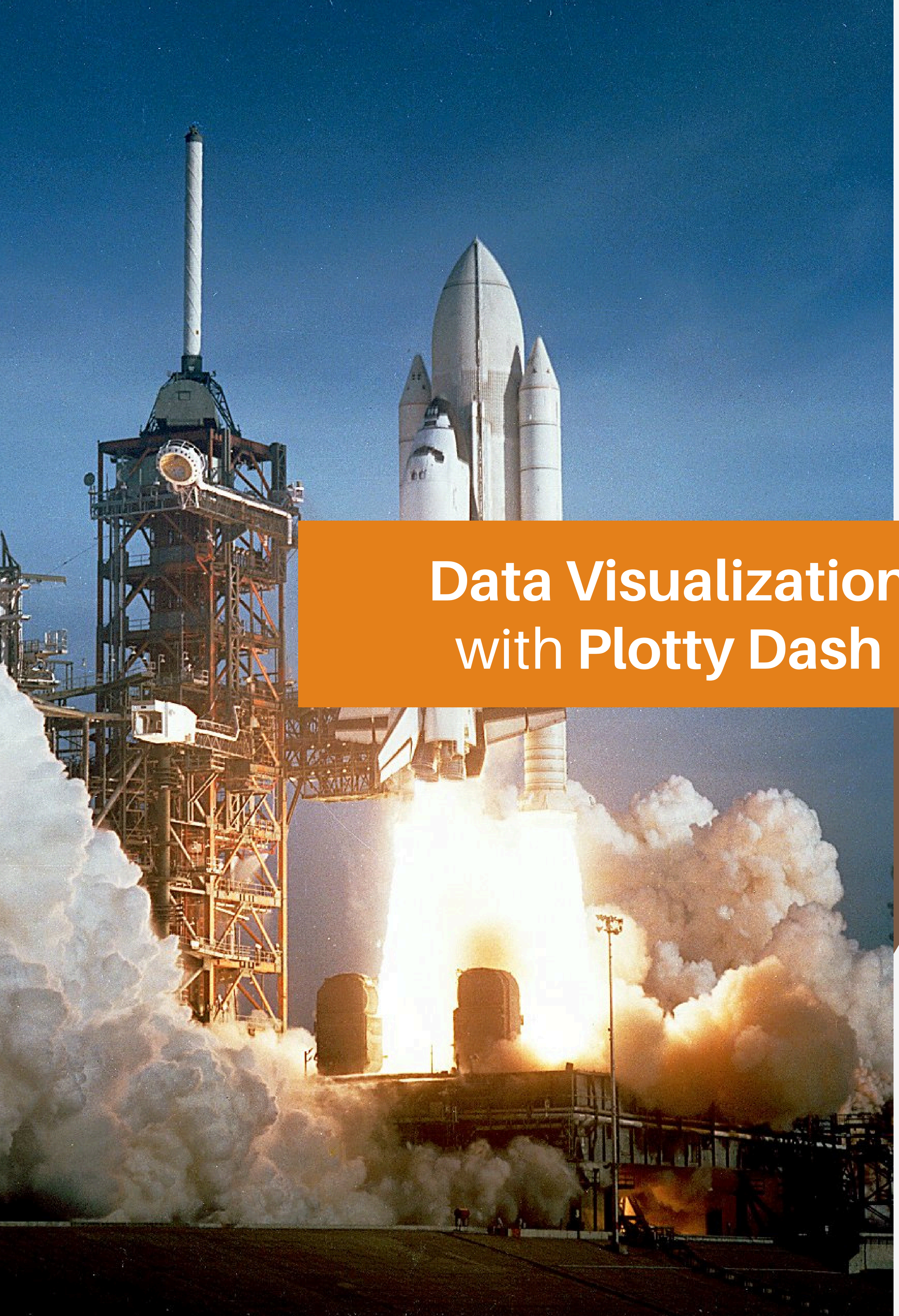
Charts were created to visualize the data:

- 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
- Success Rate Yearly Trend

- Scatter plots were used to explore variable relationships, with patterns or correlations helping to guide machine learning features.

- Bar charts compared discrete categories, showing their relationships with measured values.

- Line charts captured trends over time, providing insights into time series patterns.



Data Visualization
with **Plotty Dash**

- Convert the Class column into a NumPy array.
- Standardize the data using StandardScaler, applying fit() and transform().
- Split the data into training and testing sets using train_test_split.
- Create a GridSearchCV object with 10-fold cross-validation for parameter optimization.
- Apply GridSearchCV to various algorithms:
    1. Logistic Regression (LogisticRegression())
    2. Support Vector Machine (SVC())
    3. Decision Tree Classifier (DecisionTreeClassifier())
    4. K-Nearest Neighbor (KNeighborsClassifier()
- Calculate the accuracy of each model on the test data using score().
- Evaluate the confusion matrix for each model.
- Determine the best model based on Jaccard Score, F1 Score, and Accuracy.

Launch Site Markers:

- A blue circle marks the NASA Johnson Space Center, labeled with its name using latitude and longitude.
- Red circles mark all other launch sites, each labeled with its name using their coordinates.

Launch Outcome Markers:

- Green markers represent successful launches, and red markers indicate unsuccessful ones, visualizing success rates by site.

Proximity Distances:

- Colored lines show the distances from CCAFS SLC-40 to the nearest coastline, railway, highway, and city.

Model Building

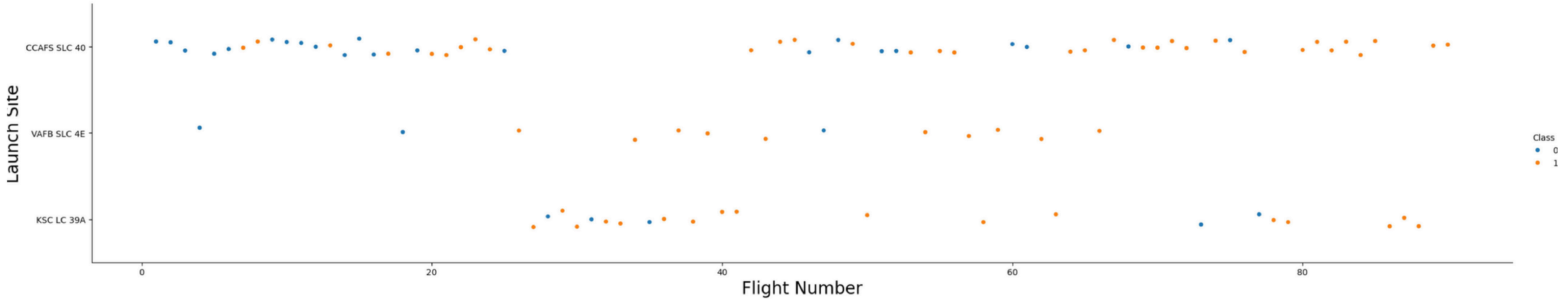
Results and Discusions

Flight Number Vs. Lunch Site

- Earlier flights showed a lower success rate, while later flights demonstrated a higher success rate.
- Approximately half of the launches originated from the CCAFS SLC-40 launch site.
- The VAFB SLC-4E and KSC LC-39A launch sites have achieved higher success rates.

This suggests that newer launches tend to have improved success rates.

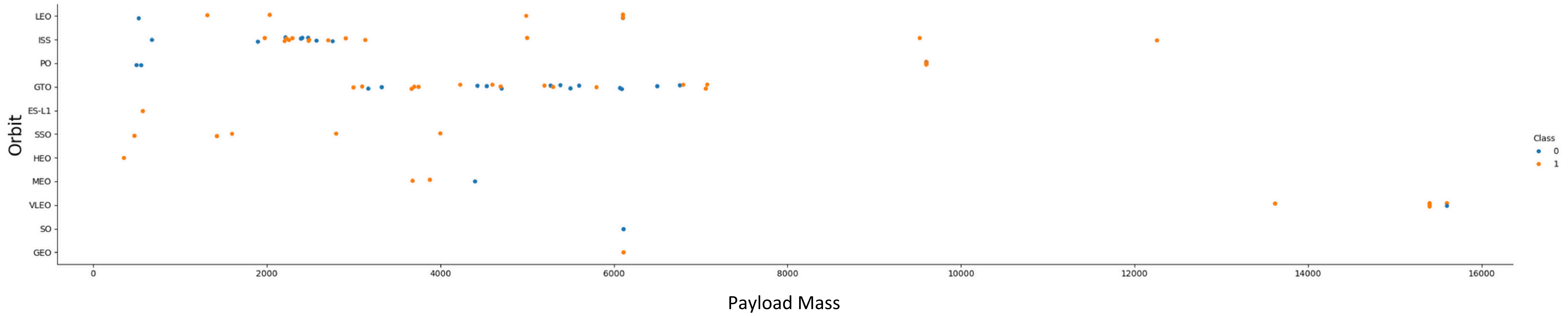
Blue = Fail
Orange = Success



Payload MassVs. Orbit

- Heavy payloads decrease success rates in GTO orbits, indicating challenges for such missions.
- Heavy payloads are more successful in VLEO, especially for Polar orbits (PO) and ISS missions.
- Success rates with heavy payloads in GTO orbits are inconsistent, depending on mission specifics.

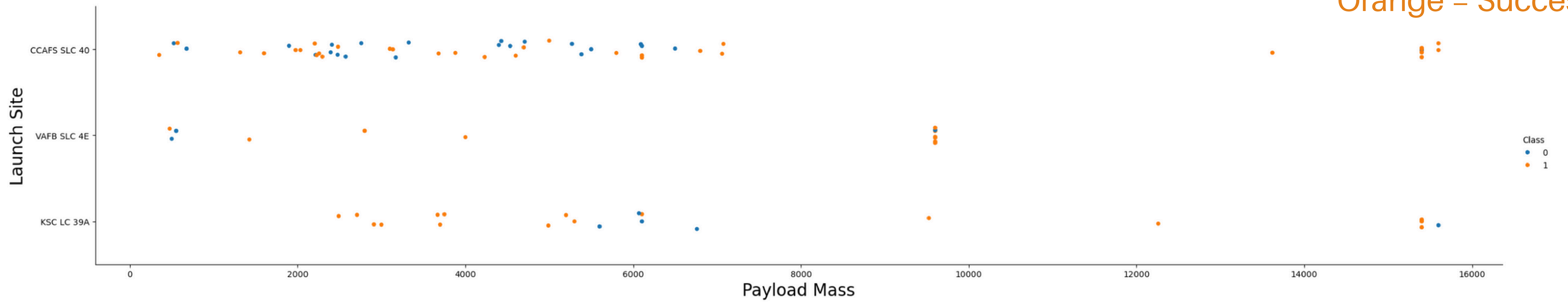
Blue = Fail
Orange = Success



Payload Mass Vs. Lunch Site

- Higher payload masses generally result in higher success rates, with most payloads over 7000 kg being successful.
- KSC LC-39A has a 100% success rate for payloads under 5500 kg and strong performance overall.
- 3VAFB SLC-4E demonstrates high success rates, particularly for payloads above 6000 kg.
- CCAFS SLC-40 shows more variability, with some failures occurring at lower payload masses.

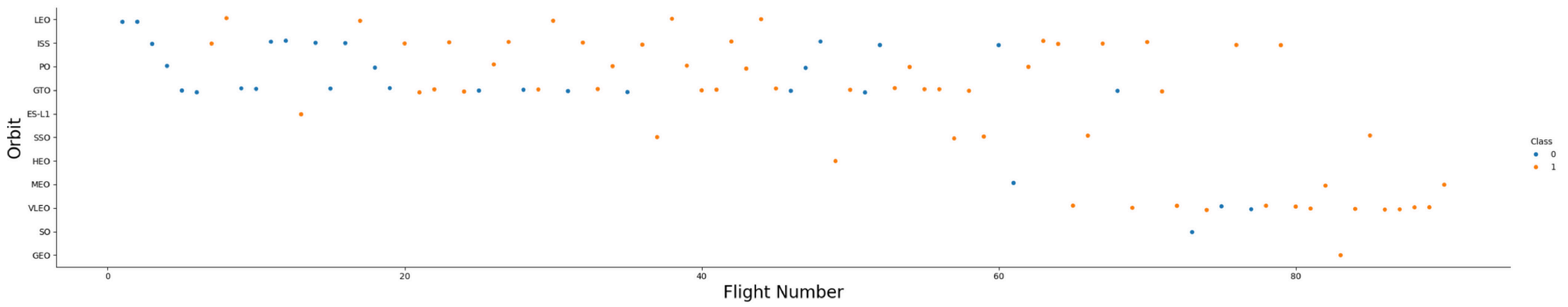
Blue = Fail
Orange = Success



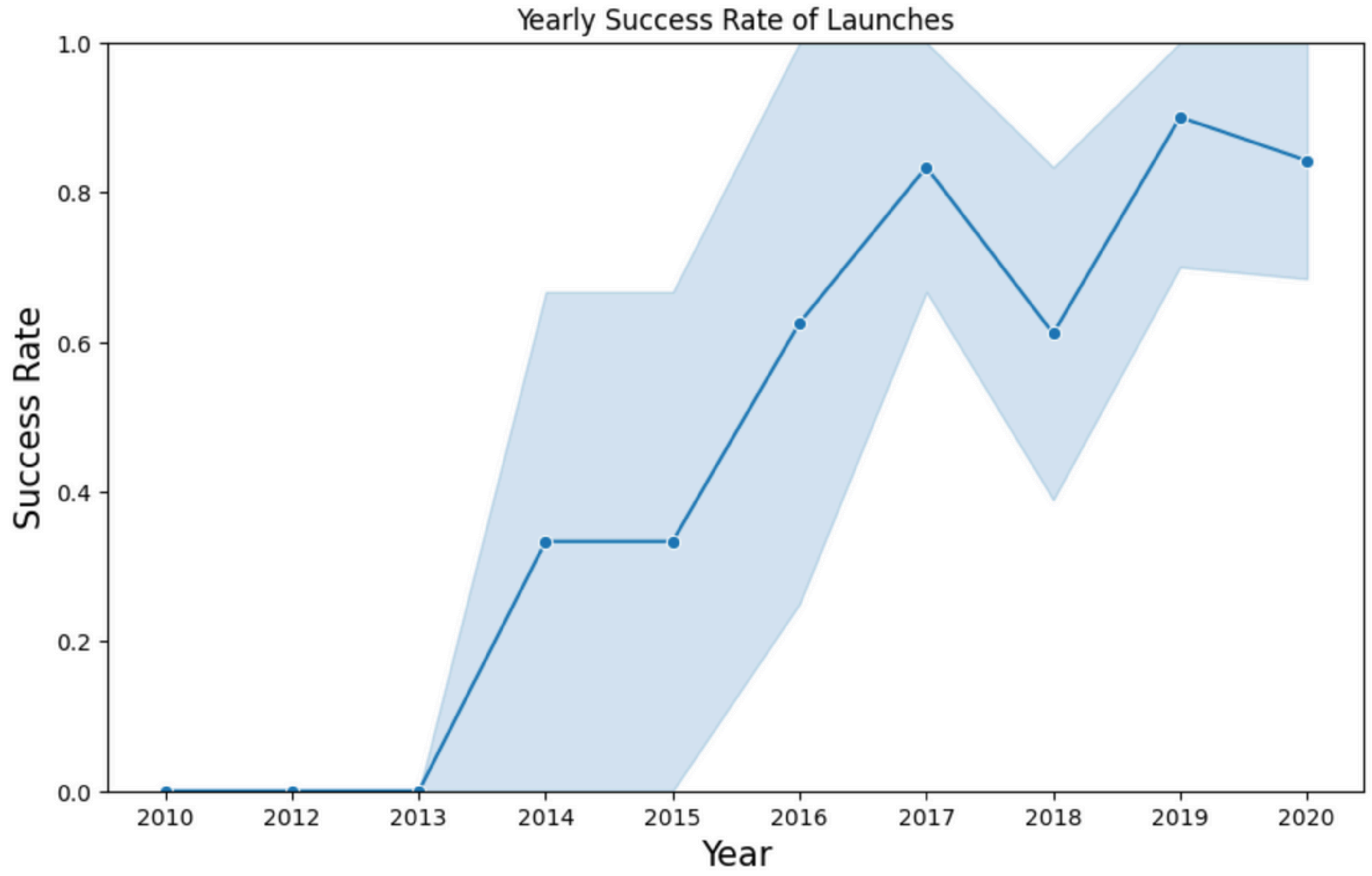
Orbit Vs. Flight Number

- Success rates generally improve with the number of flights for most orbits, especially in LEO.
- The LEO orbit shows a clear correlation between a higher number of flights and increased success rates.
- Unlike LEO, GTO orbit success rates do not appear to be influenced by the number of flights.

Blue = Fail
Orange = Success

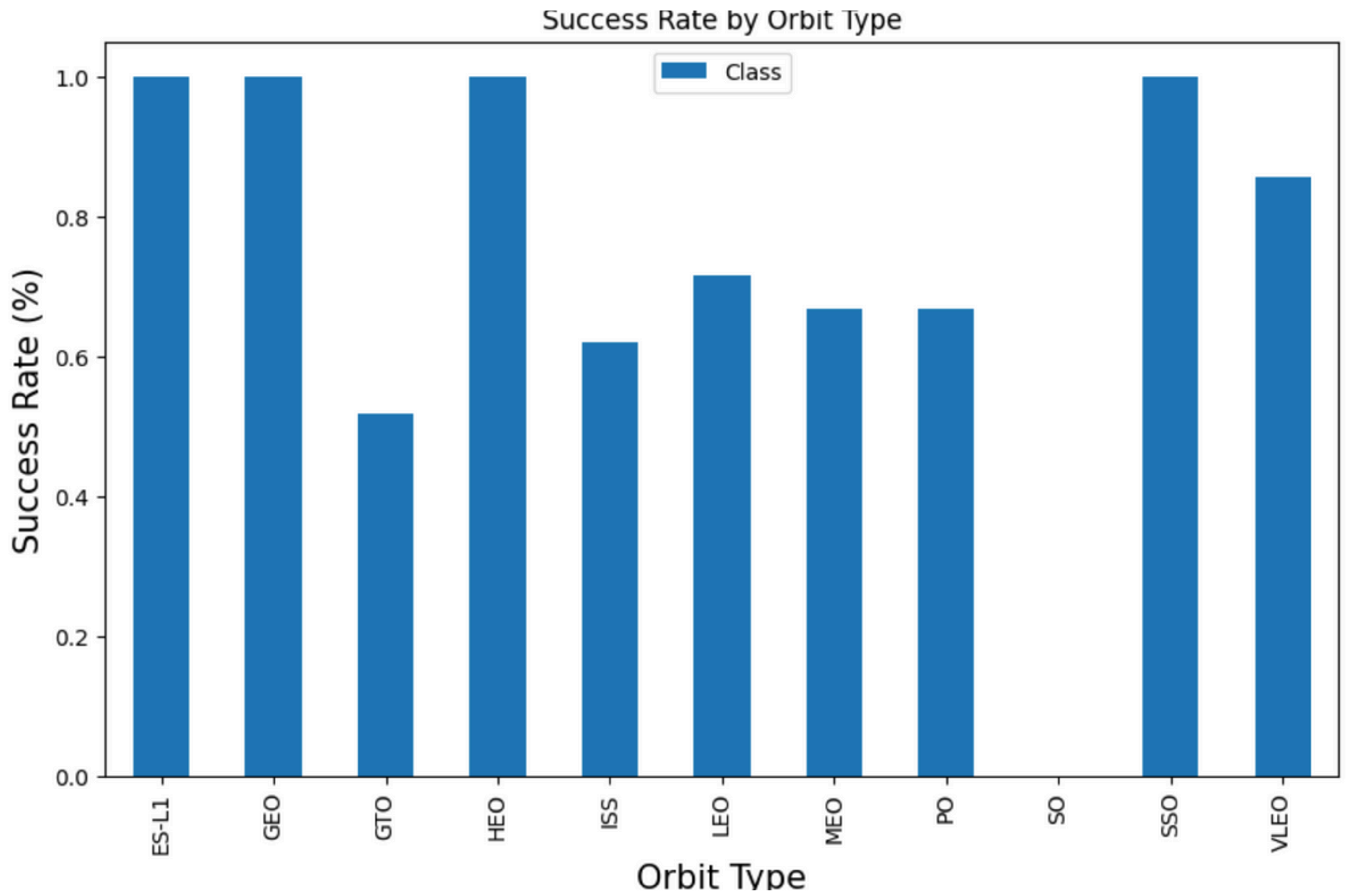


Launch Success Over Time



- There is a clear trend of increasing success rates over the years.
- Starting from 2013, success rates have significantly improved, peaking in recent years.

Rate Success by Orbit



- Orbits like ES-L1, GEO, and SSO have a 100% success rate.
- GTO has a lower success rate compared to other orbit types.
- so has 0% of success.

Launch Site

Out [4]:

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

Out [5]:

time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
15:43:00	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
07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attemp
00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Launch Site Names

Displaying 5 records of the Launch Site

Payload Mass

Out[6]:

total_payload_mass
45596

Total payload carried by boosters from NASA

Out[7]:

average_payload_mass
2534

Average payload mass carried by booster version F9 v1.1

Landing, Mission and Boosters Details

Out[15]:

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

- booster that have carried the maximum payload

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

- dates of the first successful landing outcome on ground pad

Out[9]:

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

- Successful Drone Ship Landing with Payload between 4000 and 6000

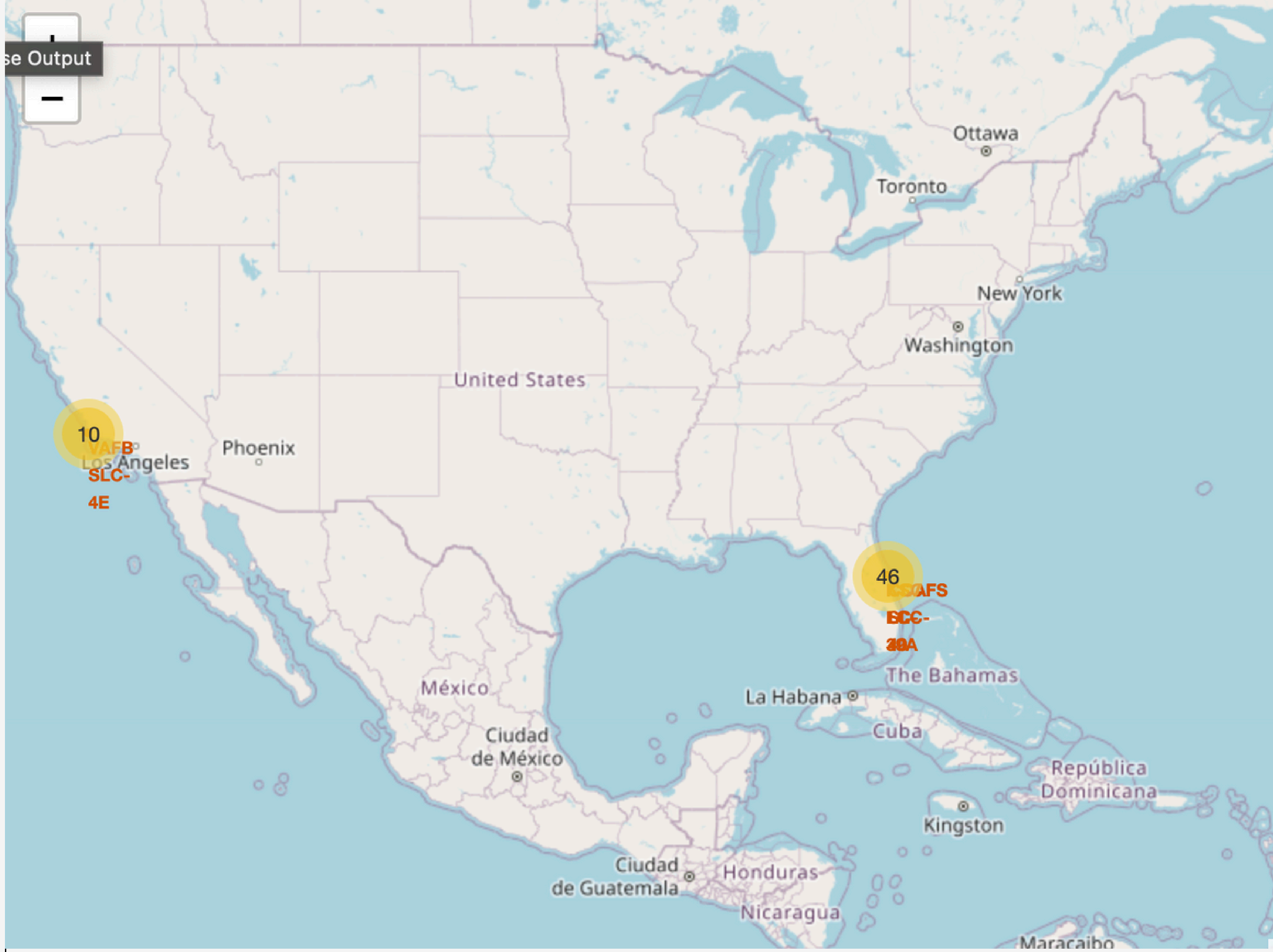
Landing Success Outcomes

Out[37]:

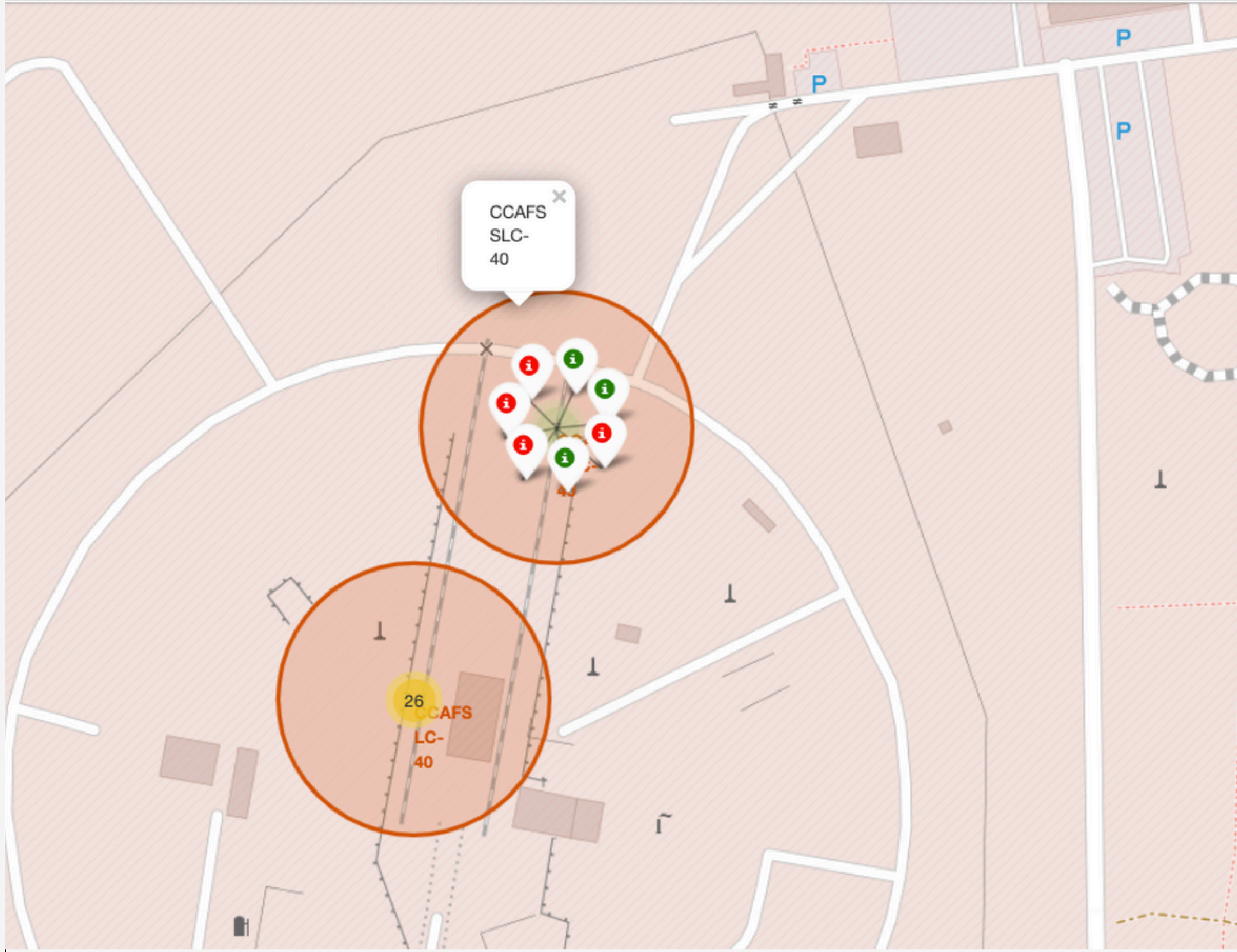
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Details of Landing Outcomes Between 2010-06-04 and 2017-03-20

Map of the Launch Sites



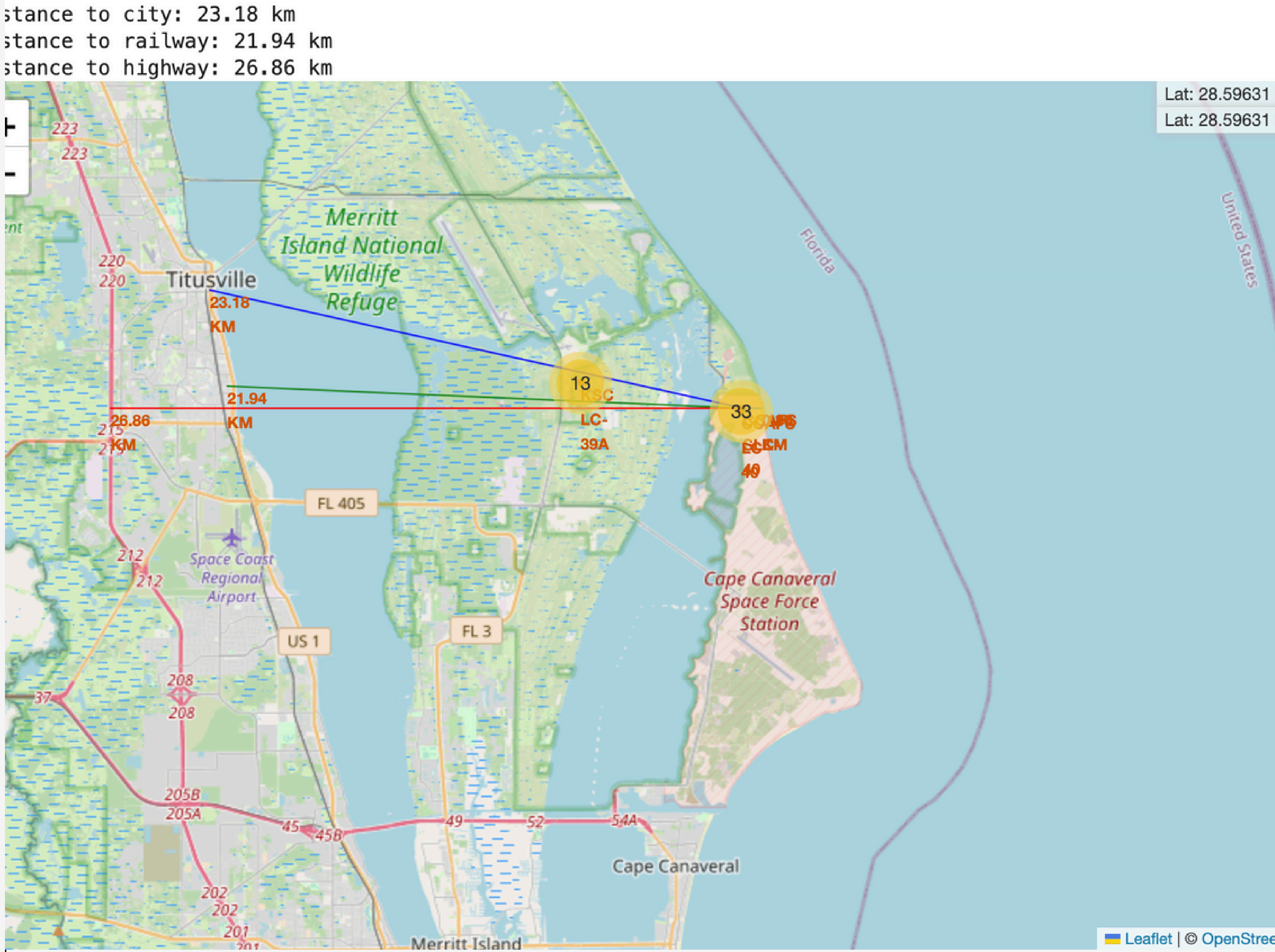
Launch Outcomes



- **Green markers** for successful launches
- **Red markers** for unsuccessful launches

Looking at the map launch site CCAFS SLC-40 has 3 out of 7 successes

Launch Outcomes

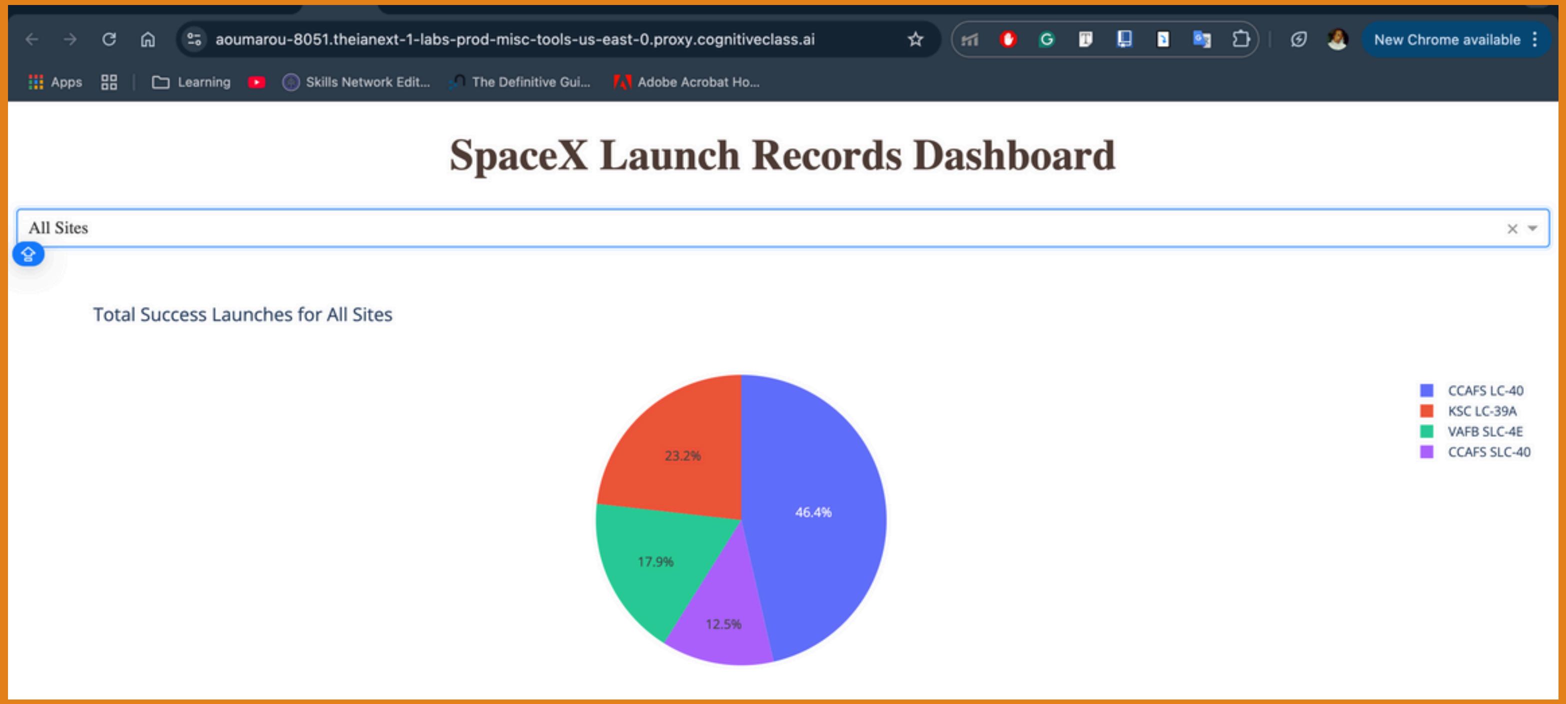


- The launch sites are 21.94 km away from railways, indicating they are not in close proximity to railway infrastructure.
- The launch sites are 26.86 km from highways, suggesting they are moderately distant from major roadways.

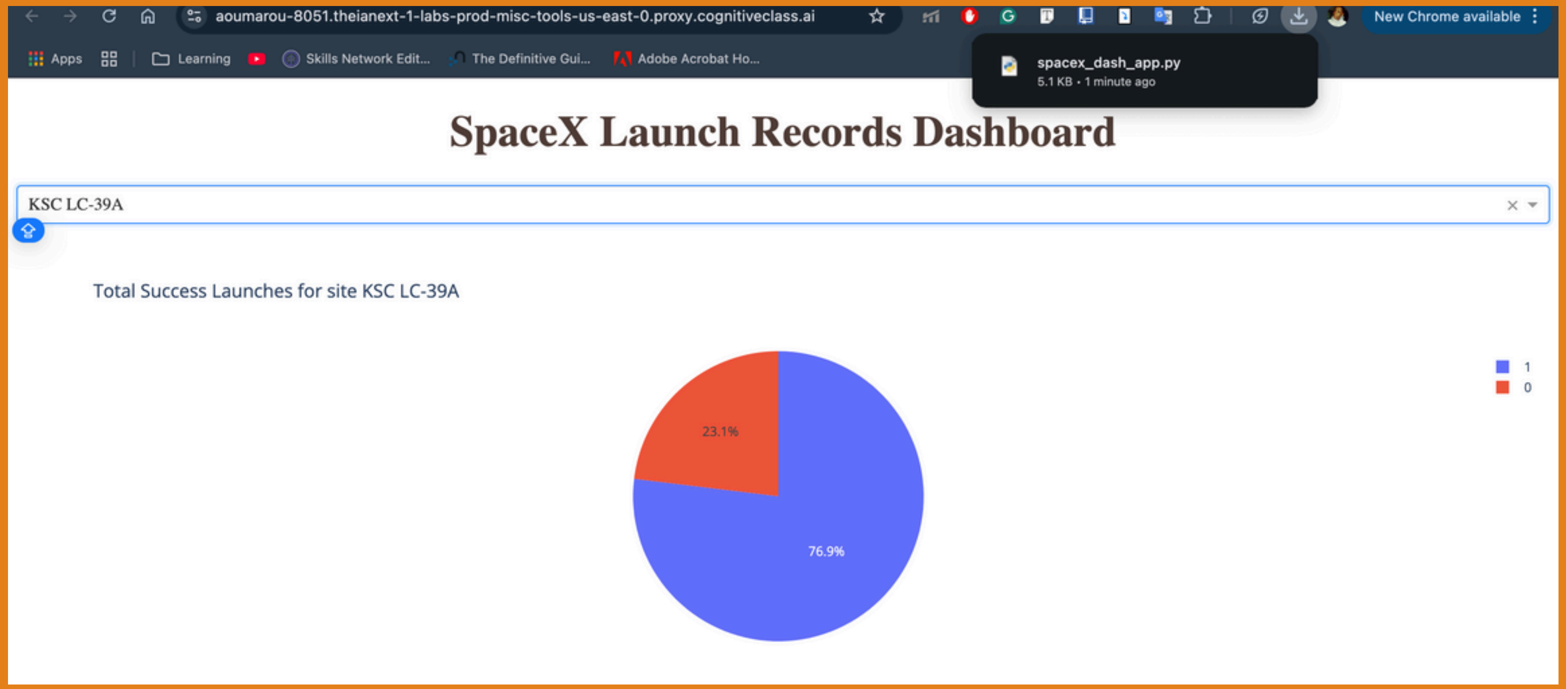
- The launch sites are only 0.86 km from the coastline, making them very close to maritime access.

- The launch sites are 23.18 km away from cities, ensuring a safe distance from urban areas.

Site Launch Success



With a success rate of 41.2%, KSC LC-39A appears to be thhe most successful launches amongst launch sites.

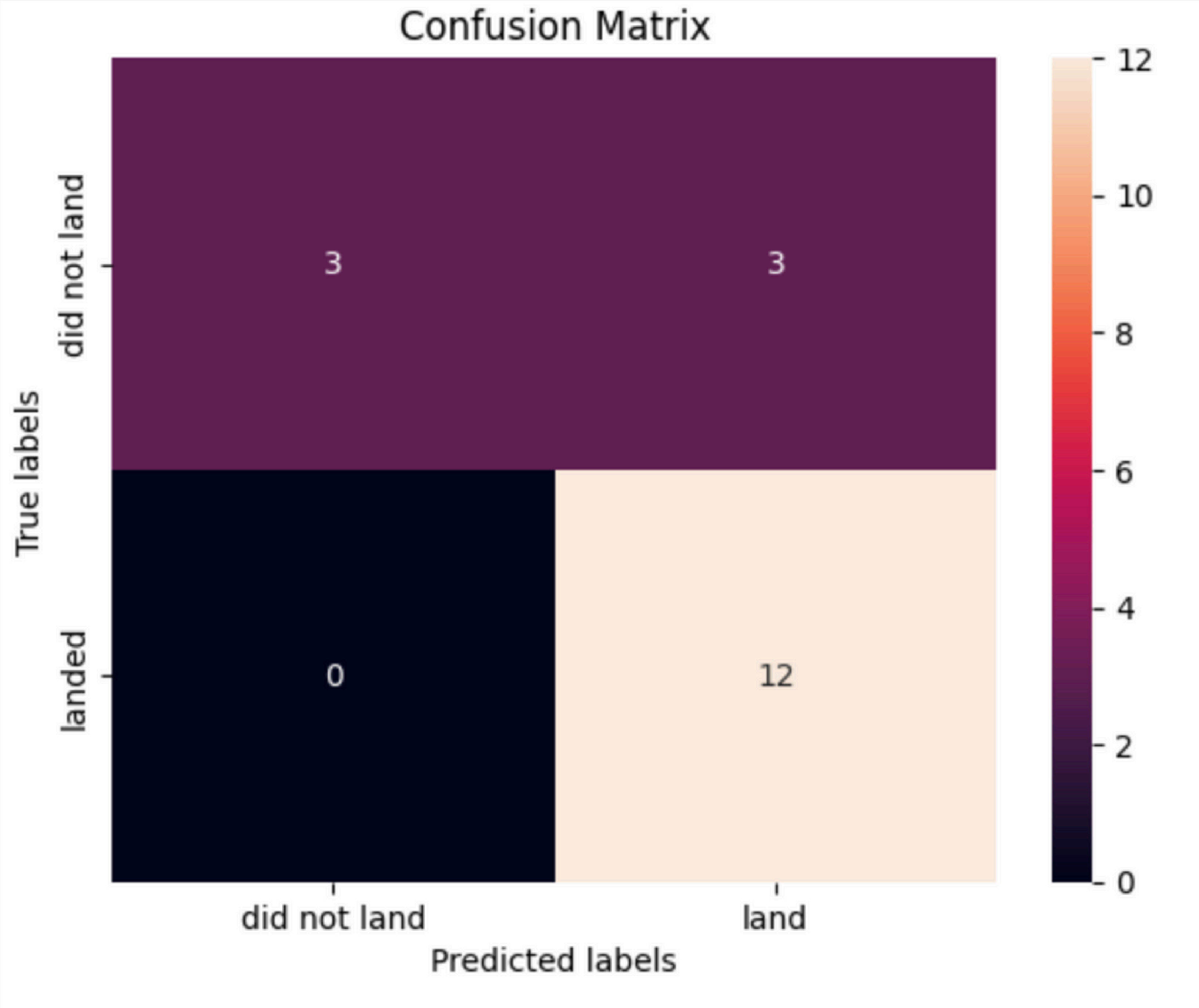


KSC LC-39A has the highest success rate among launch sites at 76.9%, with 10 successful launches out of 13 total.

Site Launch Success

Payloads between **2,000 kg** and **5,000 kg** exhibit the highest success rate. In the context of the classification, a "1" indicates a successful outcome, while a "0" signifies an unsuccessful outcome. This suggests that the decision tree classifier has a higher accuracy in predicting successful landings for payloads within this weight range.

Classification Accuracy



Performance Summary:

- A confusion matrix is used to summarize the performance of a classification algorithm.
- All confusion matrices were identical.
- The presence of false positives (Type 1 error) is undesirable.

Analysis:

The confusion matrix for the decision tree classifier indicates that the model can differentiate between classes, but it struggles with false positives, where unsuccessful landings are incorrectly labeled as successful. This issue of false positives is a key concern, as highlighted in the performance summary, which shows a precision of 0.80 and recall of 1.00, indicating a good recall but a moderate precision due to the false positives.

Conclusion

- Launch sites near the equator and close to the coast benefit from reduced fuel costs and logistical advantages.
- Launch success rates have increased over time, particularly from 2013 to 2020.
- KSC LC-39A has the highest success rate, especially for payloads under 5,500 kg.
- Orbits like ES-L1, GEO, HEO, and SSO have a 100% success rate.
- Higher payload masses are correlated with higher success rates across all launch sites.
- The decision tree classifier performed best among the models tested.
- Future improvements could come from a larger dataset, feature analysis (PCA), or the use of XGBoost.

Thank you!

December, 2024