IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Caroline Savickiene
25/06/2025

Outline

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

Executive Summary

This project used the CRISP-DM methodology to guide the data science workflow from problem understanding to model deployment. The process involved data cleaning, wrangling, feature engineering, and scaling. Data was collected through web scraping. Supervised machine learning models, including logistic regression and decision trees, were built and evaluated using cross-validation, confusion matrix, and key performance metrics. Hyperparameter tuning was performed with Grid Search. The final model predicted SpaceX landing success with 83% accuracy, with launch site, booster version, and payload mass as key factors. This approach provided reliable forecasting to support mission planning.

Introduction

- **Project Background:**
  SpaceX's reusable rocket program aims to cut launch costs, making successful landings a key performance metric.
- **Objective:**
  Analyze launch data to identify key factors influencing landing success.
- **Key Questions:**
- What features impact landing outcomes?
- Can we predict landing success before launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected by using web scraping techniques to automatically extract relevant information from publicly available online sources.
- Perform data wrangling
  - Data wrangling involved cleaning, handling missing values, removing duplicates, and restructuring the data. Features were engineered and numerical values scaled to prepare for modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build classification models by training on selected features, tune with Grid Search, and evaluate using accuracy, precision, recall, F1 score, confusion matrix, and cross-validation.

Data Collection

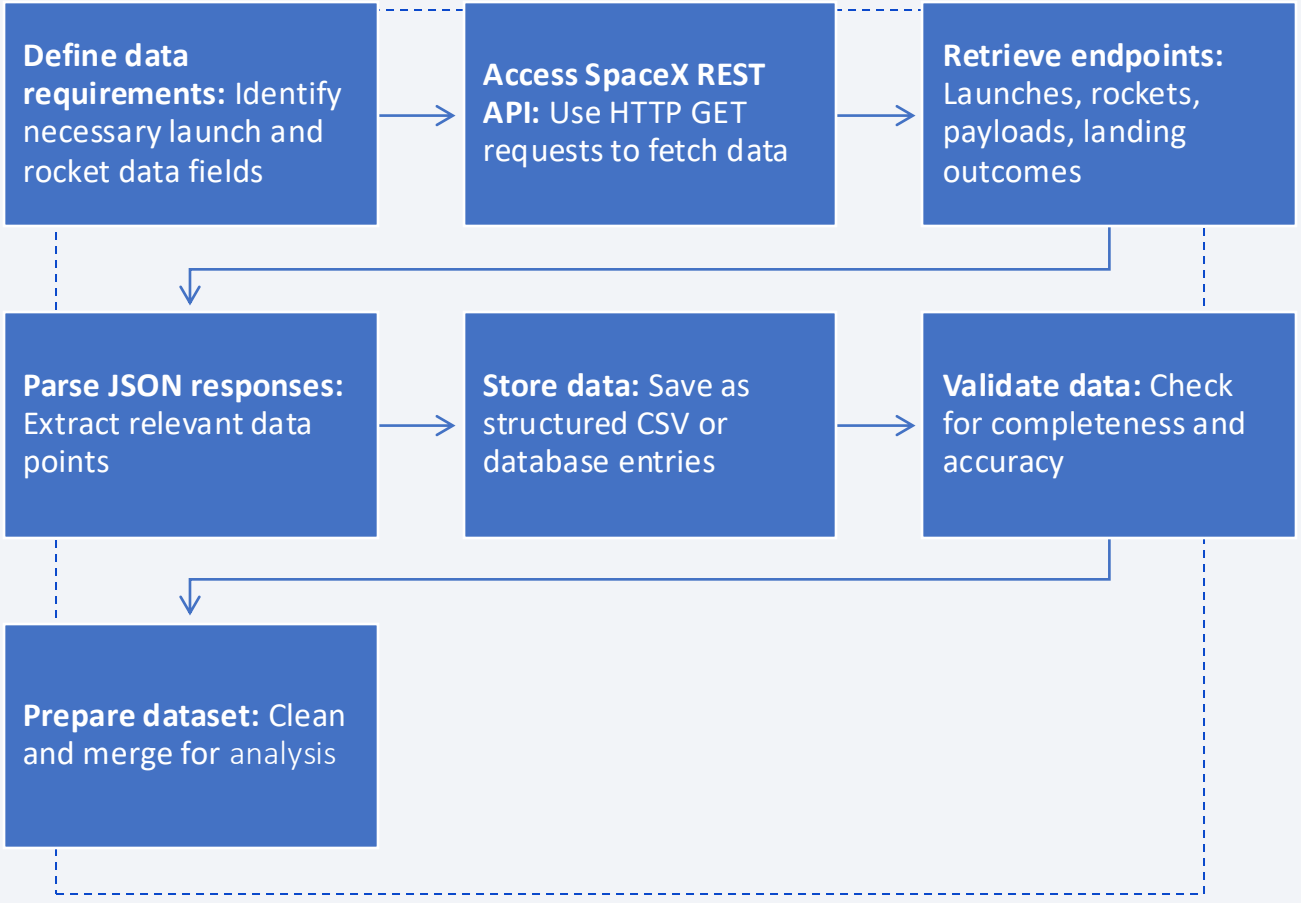
The data collection process began by identifying key sources such as official SpaceX launch records, public APIs, and online databases. Automated web scraping scripts were then used to extract structured data from relevant websites. The raw data was cleaned and stored in CSV and JSON formats. Next, the dataset was verified for completeness by checking for missing or inconsistent entries. Finally, the cleaned data was consolidated into a unified dataset, ready for analysis.

- Define data requirements
- Select sources (websites, APIs)
- Run web scraping scripts
- Extract and save raw data
- Clean and validate data
- Merge into final dataset

Data Collection – SpaceX API

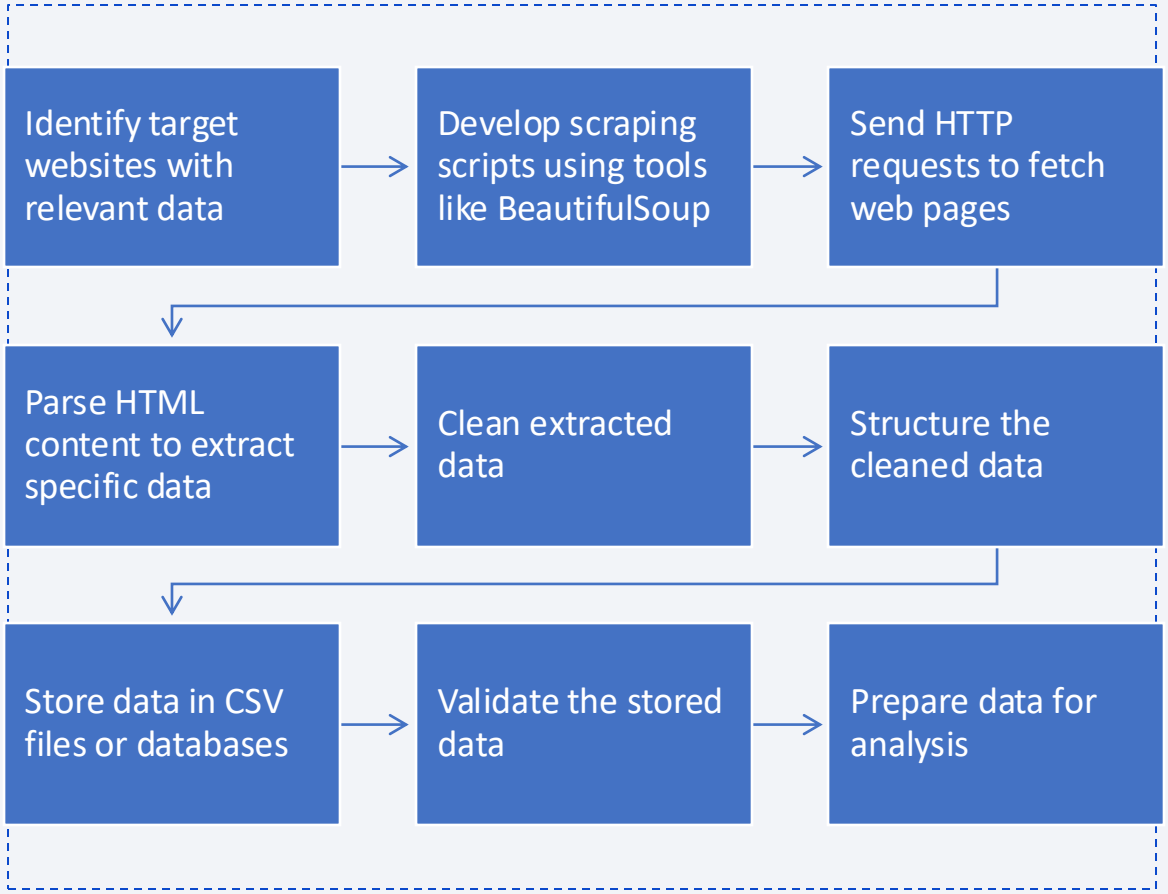
- Data was collected by sending GET requests to SpaceX REST API endpoints like `/launches` and `/rockets`. JSON responses were parsed, relevant data extracted, stored in structured formats, validated for accuracy, and prepared for analysis.

- https://github.com/CarolineEmilieS/Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

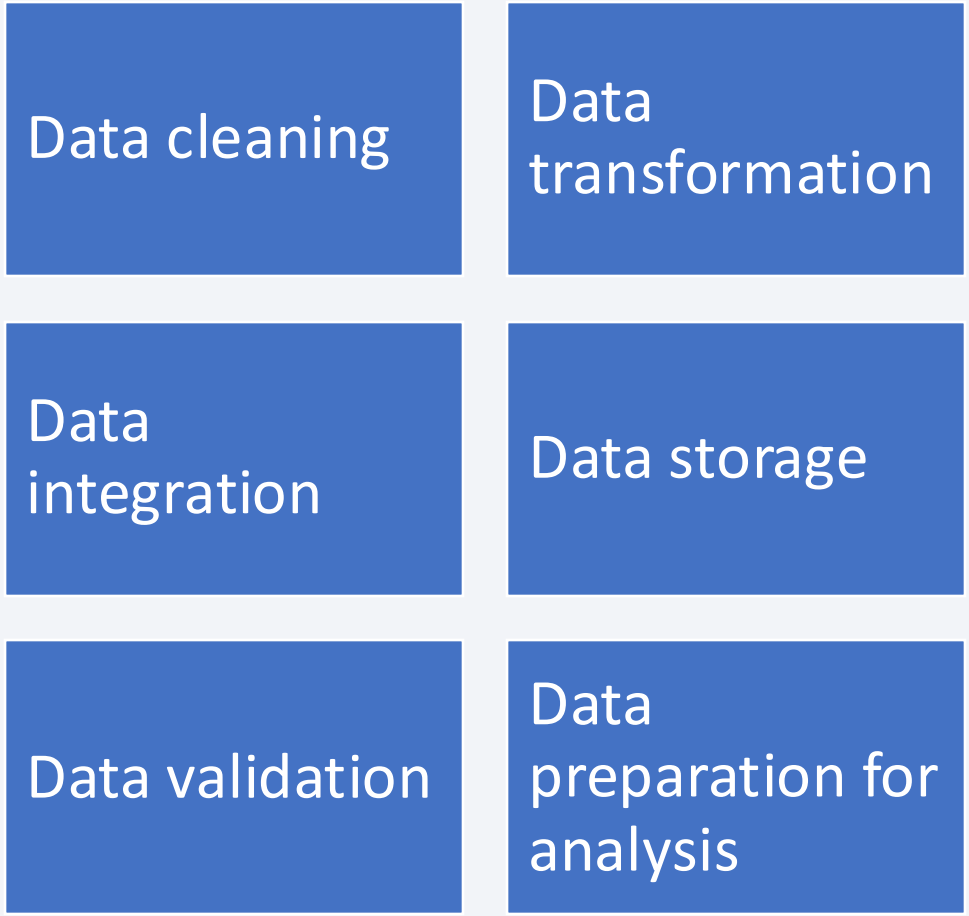
- The web scraping process began by identifying target websites with relevant data. Scraping scripts were developed using tools like BeautifulSoup to send HTTP requests and fetch web pages. Extracted data was cleaned, structured, stored in CSV or databases, then validated and prepared for analysis.
- https://github.com/CarolineEmilieS/Capstone/blob/main/jupyter-labs-webscraping.ipynb



DATA WRANGLING

- Data extraction involved sending HTTP requests and parsing HTML to gather information. The data was cleaned, transformed into a structured format, and stored in CSV files or databases. Finally, it was validated and prepared for analysis.

- https://github.com/CarolineEmilieS/Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- Bar charts were plotted to compare the number of launches by year and by rocket type, helping to highlight SpaceX's growth and rocket usage trends. Line charts illustrated the timeline of launch success rates, showing improvements over time. Scatter plots were used to examine relationships between payload weight and launch success or failure. Pie charts displayed the distribution of launch outcomes and customer types. These charts were chosen to clearly visualize key aspects of SpaceX's launch performance and mission profiles, aiding in deeper analysis and understanding.
- https://github.com/CarolineEmilieS/Capstone/blob/main/edadataviz.ipynb

EDA with SQL

- Selected specific columns from launch data to analyze key attributes like launch date, rocket type, and outcome
- Filtered launches by year to study performance trends over time
- Counted total launches grouped by rocket type to identify the most used rockets
- Calculated success rates by grouping launches based on outcome status
- Joined launch data with payload information to explore payload-related patterns
- Aggregated launch counts by customer to determine major clients
- Ordered launches chronologically to observe temporal launch patterns

- https://github.com/CarolineEmilieS/Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

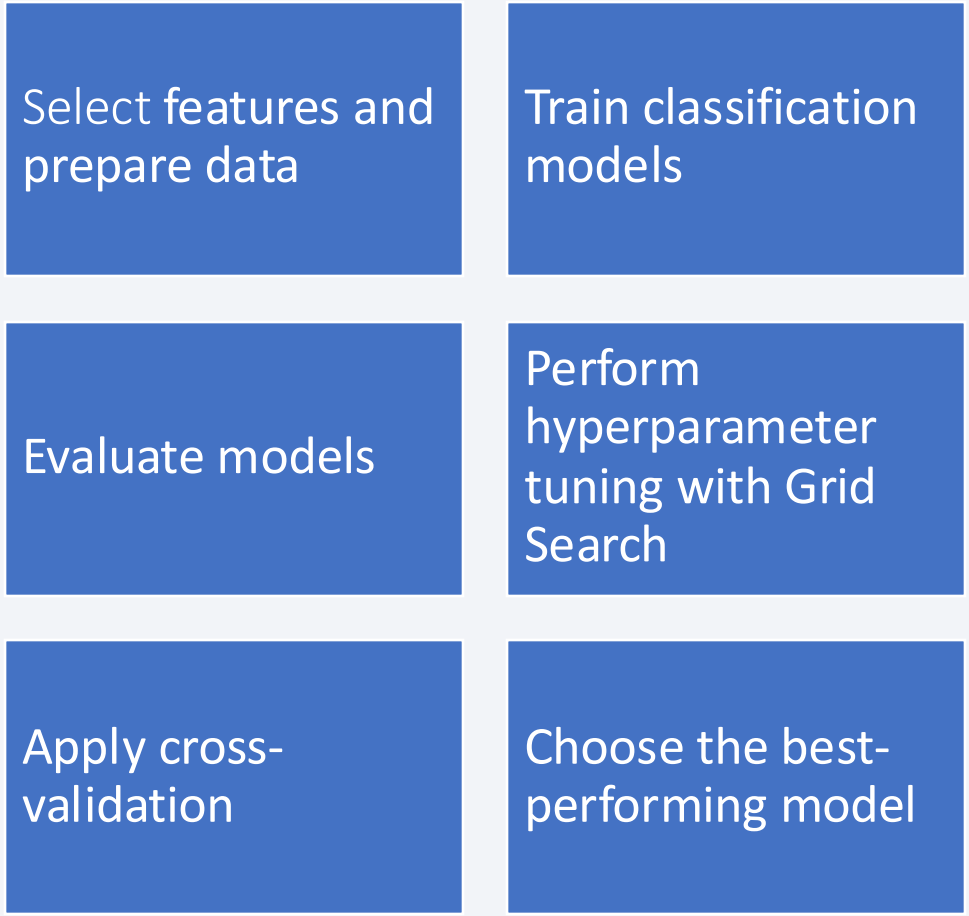
- Markers were added to pinpoint specific launch sites, providing clear visual references for each location. Circles highlighted launch areas with radius size indicating the density or significance of launches nearby. These objects were included to make the map interactive and informative, allowing easy identification of key locations and understanding of spatial patterns related to SpaceX launches.
- https://github.com/CarolineEmilieS/Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The dashboard includes key plots such as a bar chart showing launch success rates by site, a scatter plot comparing payload mass against landing outcomes, and a line chart tracking success trends over time. Interactive filters allow users to select specific launch sites, booster versions, and date ranges to explore patterns dynamically. These visualizations and interactions were added to help users easily identify factors affecting SpaceX landing success and enable flexible, insightful data exploration for better decision-making.
- https://github.com/CarolineEmilieS/Capstone/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Built classification models, evaluated them with key metrics, tuned hyperparameters using Grid Search, and applied cross-validation. The best model was chosen for its high and consistent accuracy.
- https://github.com/CarolineEmilieS/Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

- The exploratory data analysis revealed key insights into the dataset. Distributions of variables like payload mass and launch site showed significant variation affecting landing success. Correlation analysis highlighted strong relationships between booster version and landing outcomes. Visualizations identified patterns and outliers, guiding feature selection and data preprocessing steps.
- The predictive models achieved strong performance in forecasting SpaceX landing success, with the best model reaching 83% accuracy. Key features like launch site, booster version, and payload mass were significant predictors.

SpaceX Launch Records Dashboard

All Sites

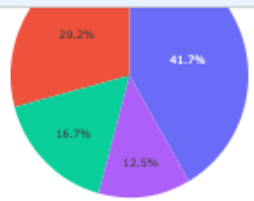
All Sites

CCAFS LC-40

CCAFS SLC-40

KSC LC-39A

VAFB SLC-4E

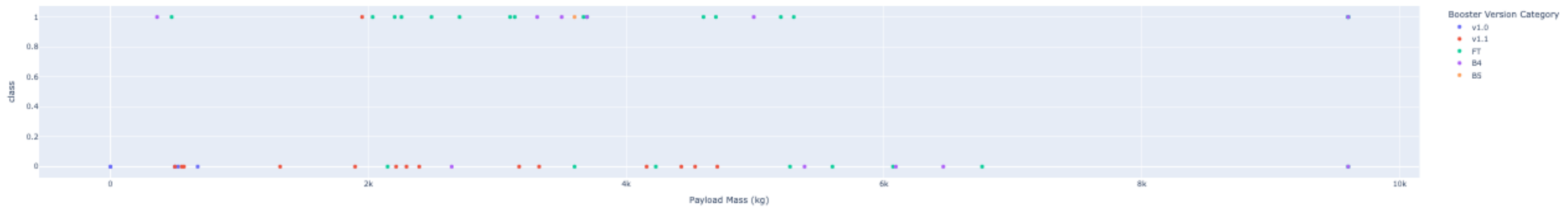


CCAFS SLC-40

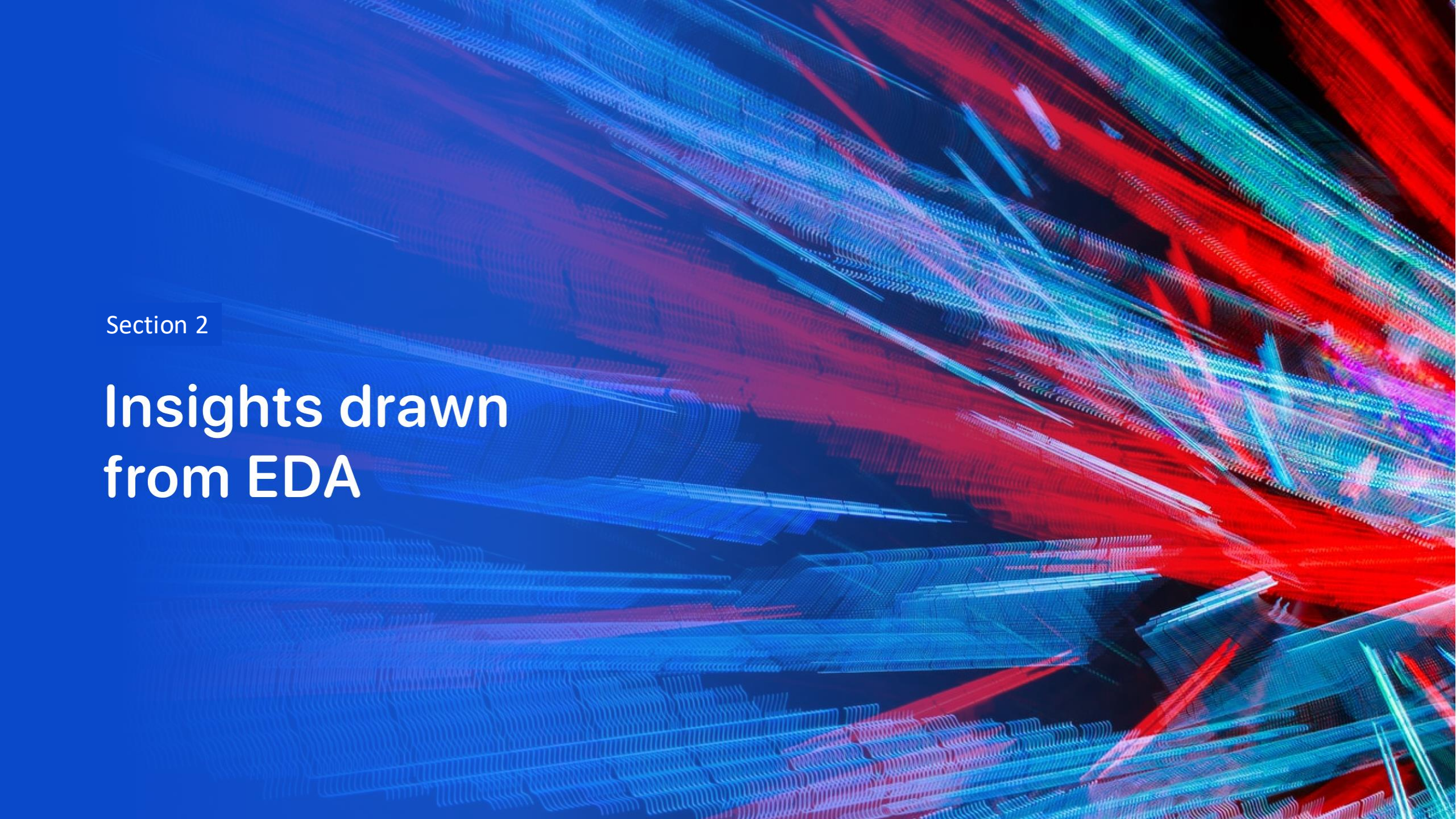
Payload range (Kg):

0 100

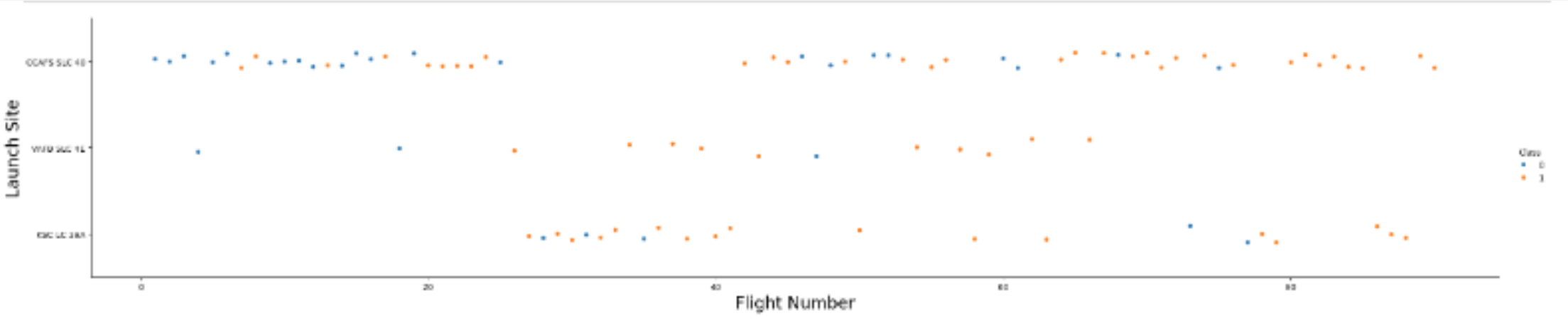
Payload vs. Outcome for All Sites



Errors Callbacks | v3.0.4 | Server

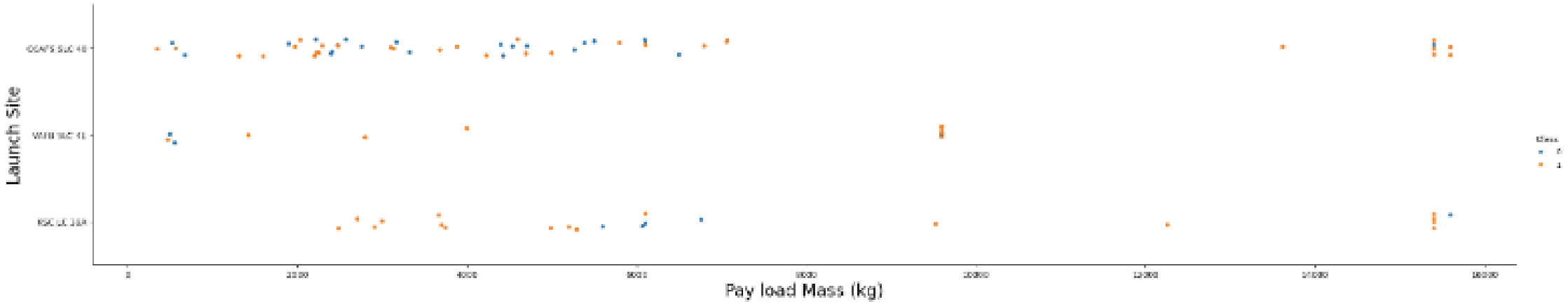

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



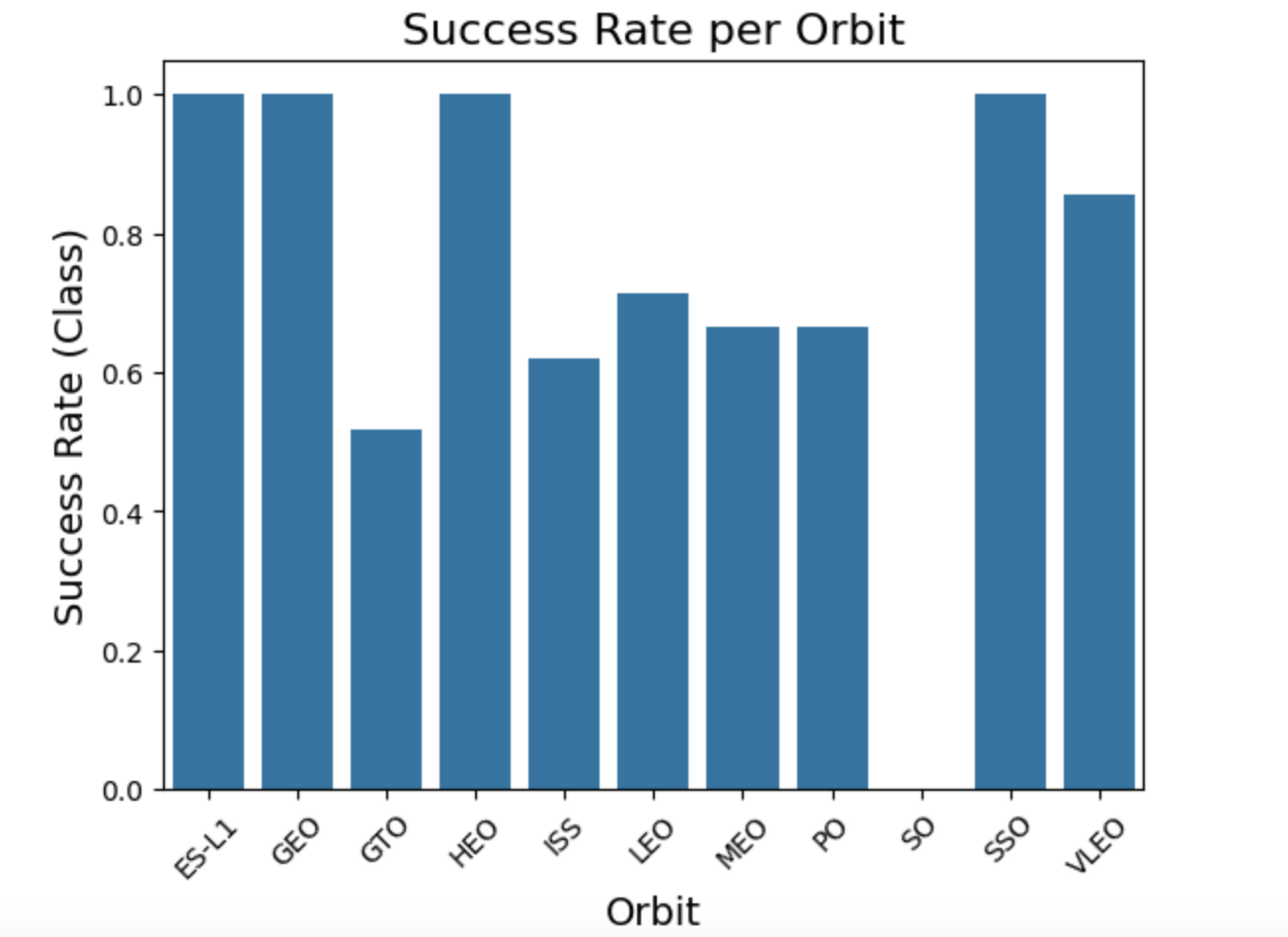
On the scatterplot it is seen that there is not much relationship between the launch site and the flight number.

Payload vs. Launch Site



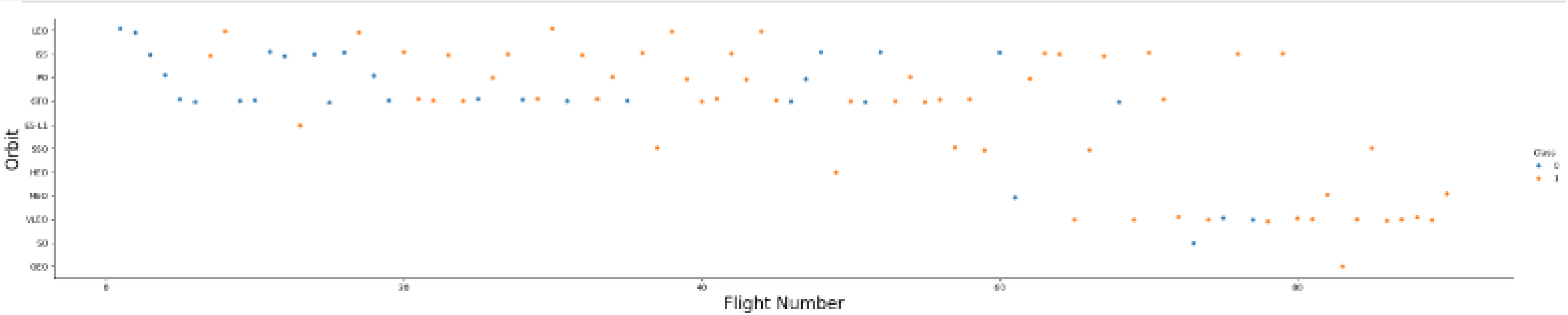
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



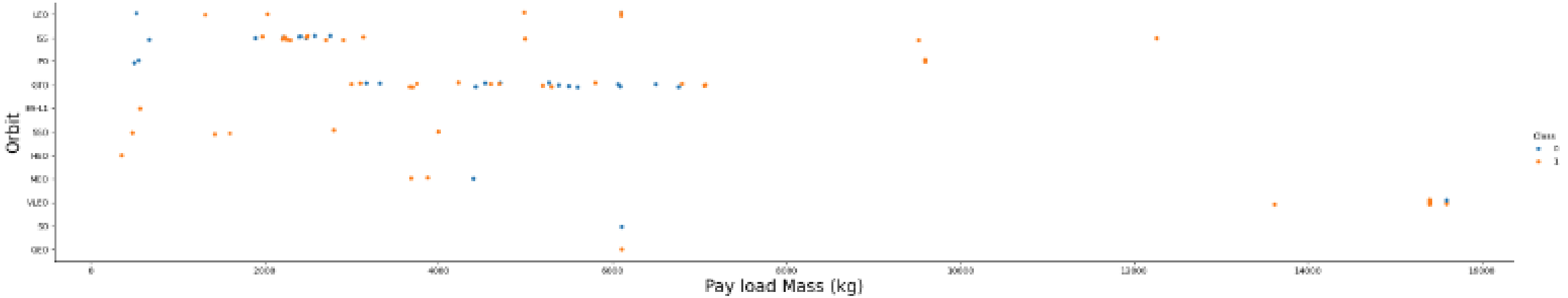
On the bar chart it is seen that the orbits ES-L1, GEO, HEO, SSO have the highest success rates and VLEO is also quite high.

Flight Number vs. Orbit Type



You can observe that 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.

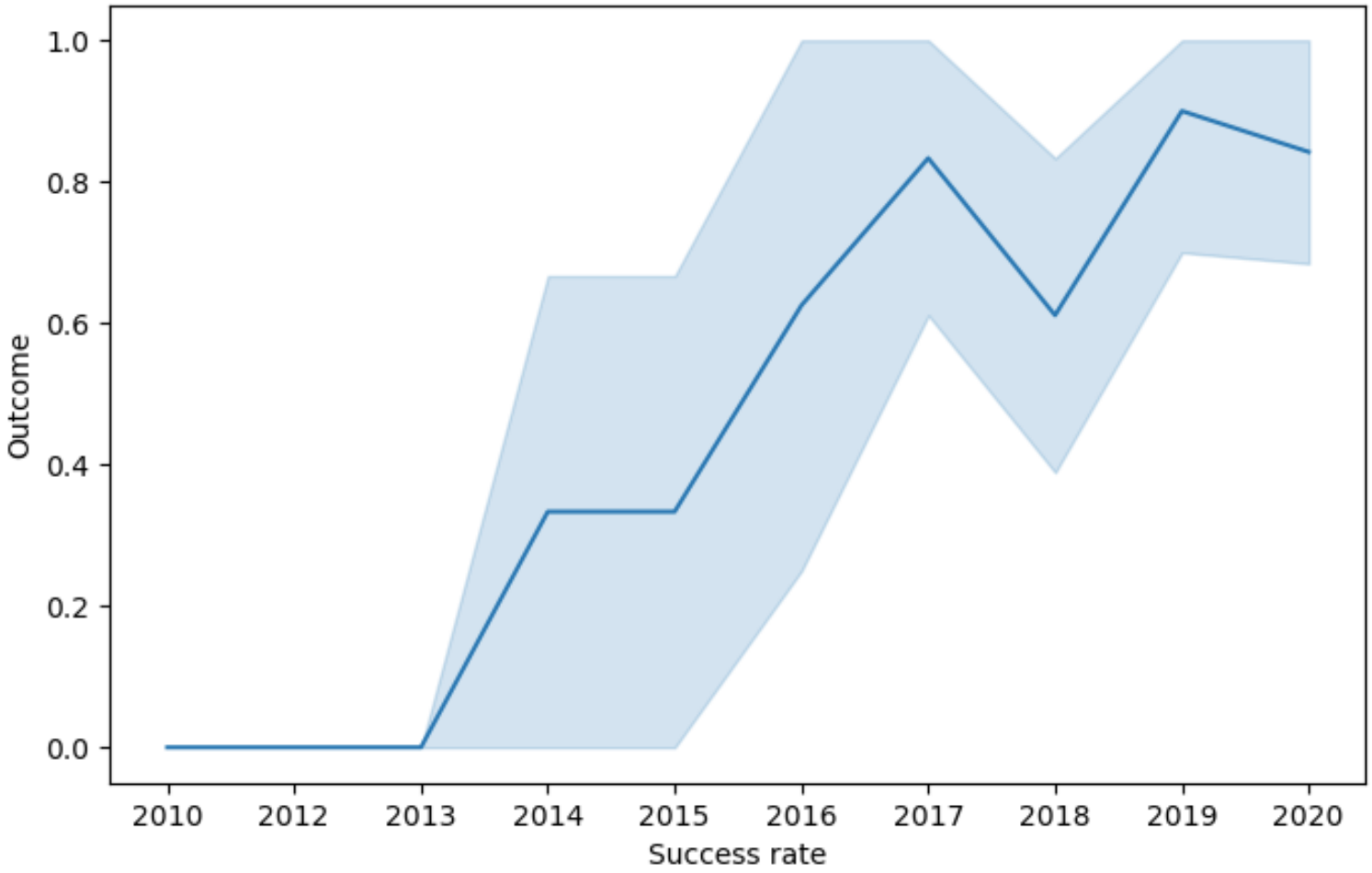
Payload vs. Orbit Type



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

Launch_Sites

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

The results show that there are 4 different launch sites.

Launch Site Names Begin with 'CCA'

6]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (par
2010-12-08	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 (par
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No a
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No a
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No a

Here are 5 instances where a launch site was used that started with CCA

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

The total payload carried by boosters from NASA is 45596

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

The average payload mass carried by booster version F9 v1.1 is 2928.4kg

First Successful Ground Landing Date

The first successful ground landing was December 22nd 2015.

```
SELECT DATE FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
ORDER BY DATE LIMIT 1;
```

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

Out[23]:

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

t[25]:

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

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

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

Boosters Carried Maximum Payload

List of the names of the boosters that have carried the maximum payload mass

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

MonthName	Failure_Landing_Outcome	Booster_Version	Launch_Site
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
June	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

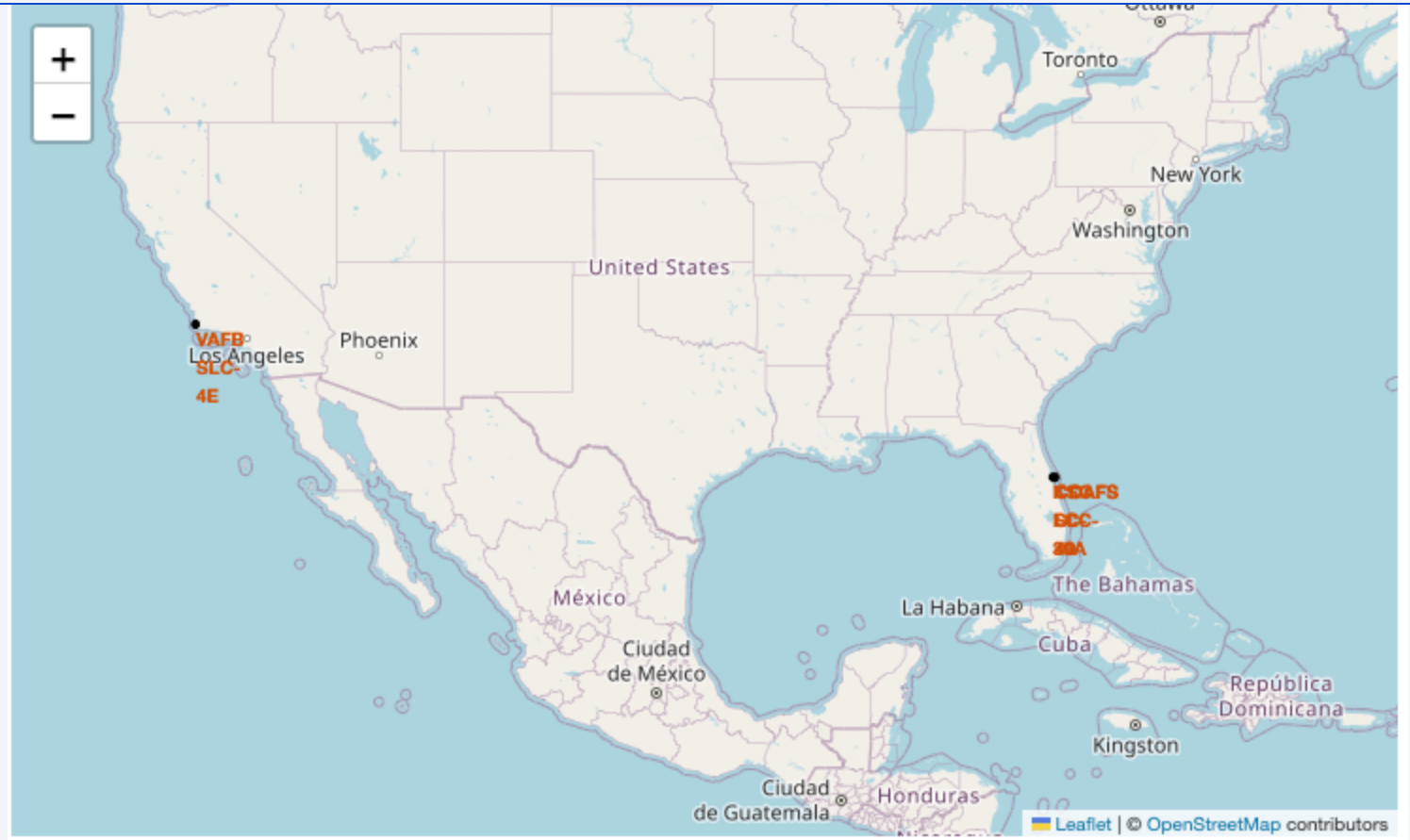
The count of landing outcomes ranked in descending order between the date 2010-06-04 and 2017-03-20

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

Section 3

Launch Sites Proximities Analysis

Map of the Launch Sites



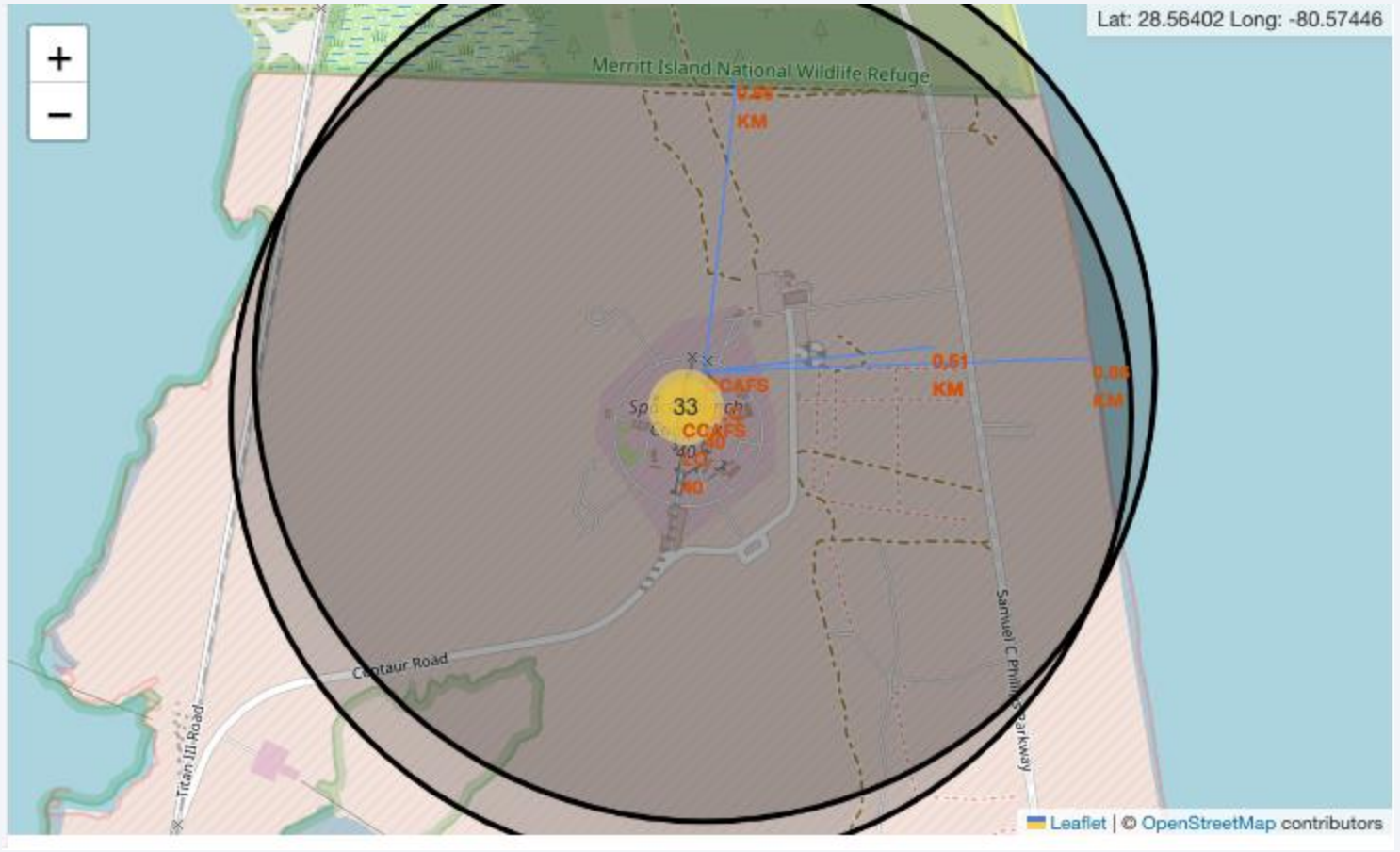
The map shows the launch sites and we can see on the map that they are all in close proximity to an ocean.

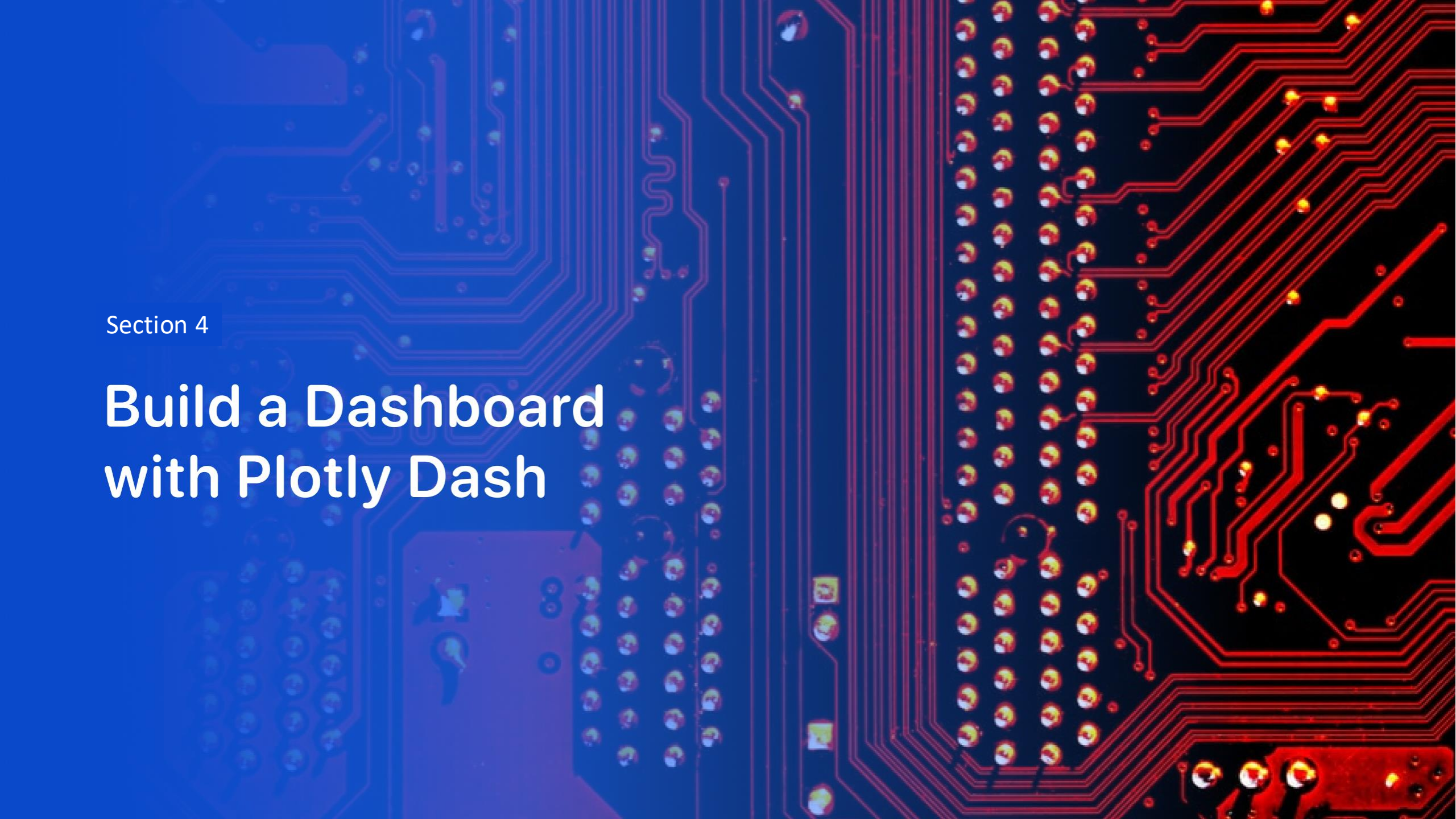
Map of Successful and Failed launches

On the map it is possible to see the outcomes for each launch; red for failed launch an green for successful launch.

<Folium Map Screenshot 3>

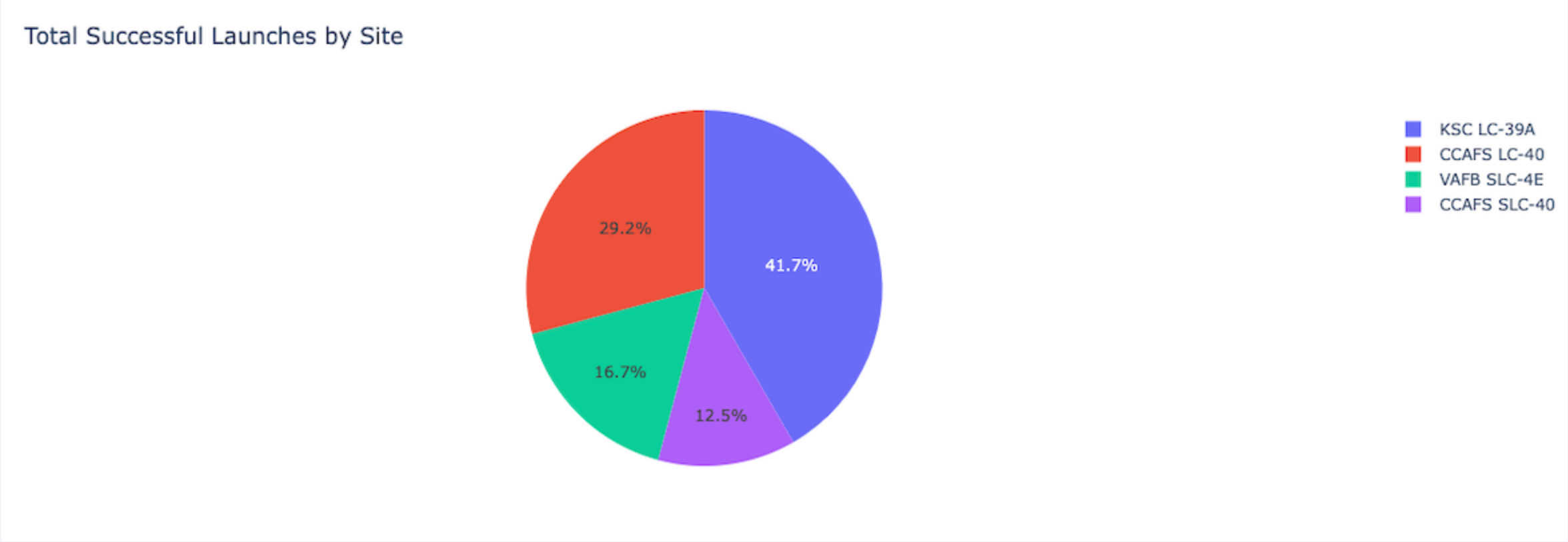
The map shows the distance in km from the launch site CCAF SLC-40 to the nearest coastline, road and forest.




Section 4
Build a Dashboard with Plotly Dash

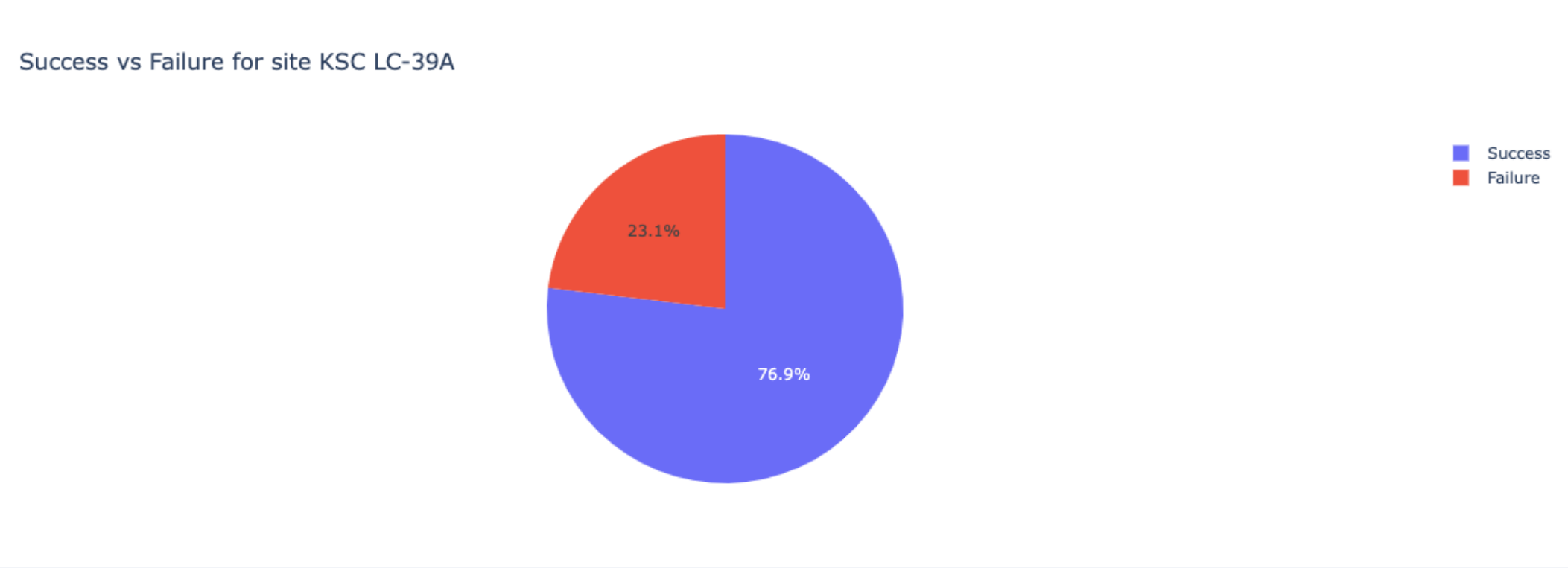
Total Successful Launches by Site

The pie chart shows the total successful launches for each of the sites and the corresponding percentage.



Successful and Failed launches for the site KSC LC-39A

The piechart shows the launch site with highest launch success ratio. Its shows the percentage of successful and failed launches for the site KSC LC-39A.



Payload vs. Launch Outcomes

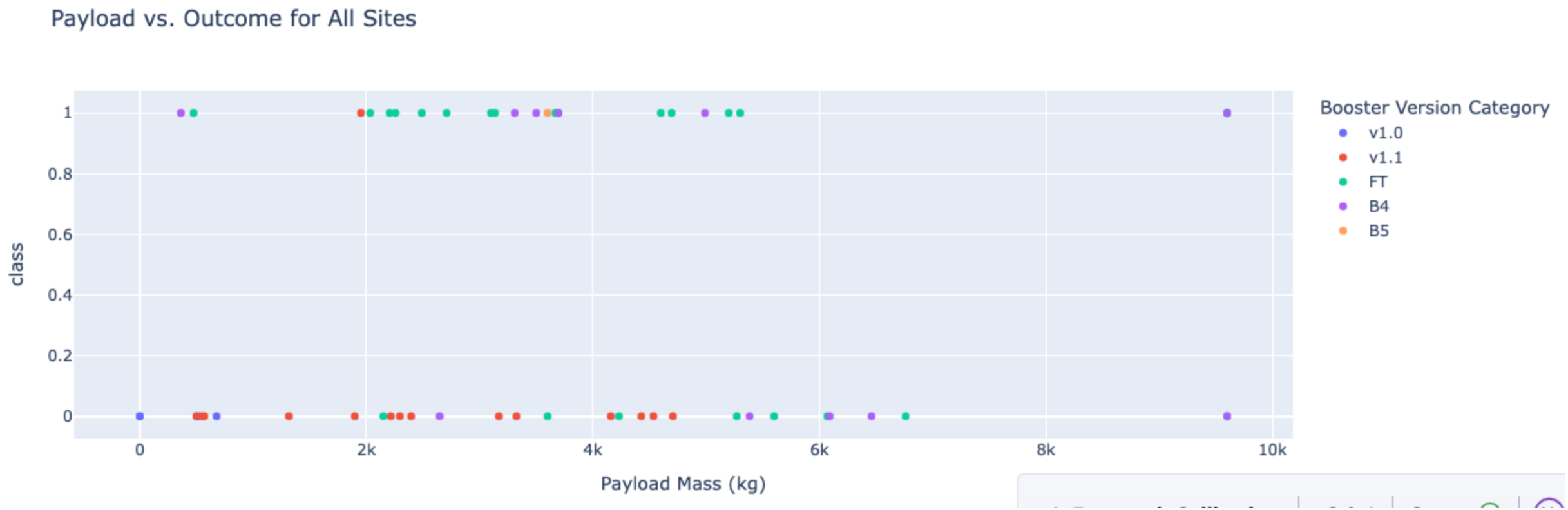
The scatter plots show Payload vs. Launch Outcome for all sites, with different payload selected in the range slider.

Based on the scatter plots we can see that the Booster Version B4 has the highest success rate when the Payload is very high.

Similarly we can see that with a lower Payload the Booster Version FT has a high success rate, while the model v1.1 has a low success rate.

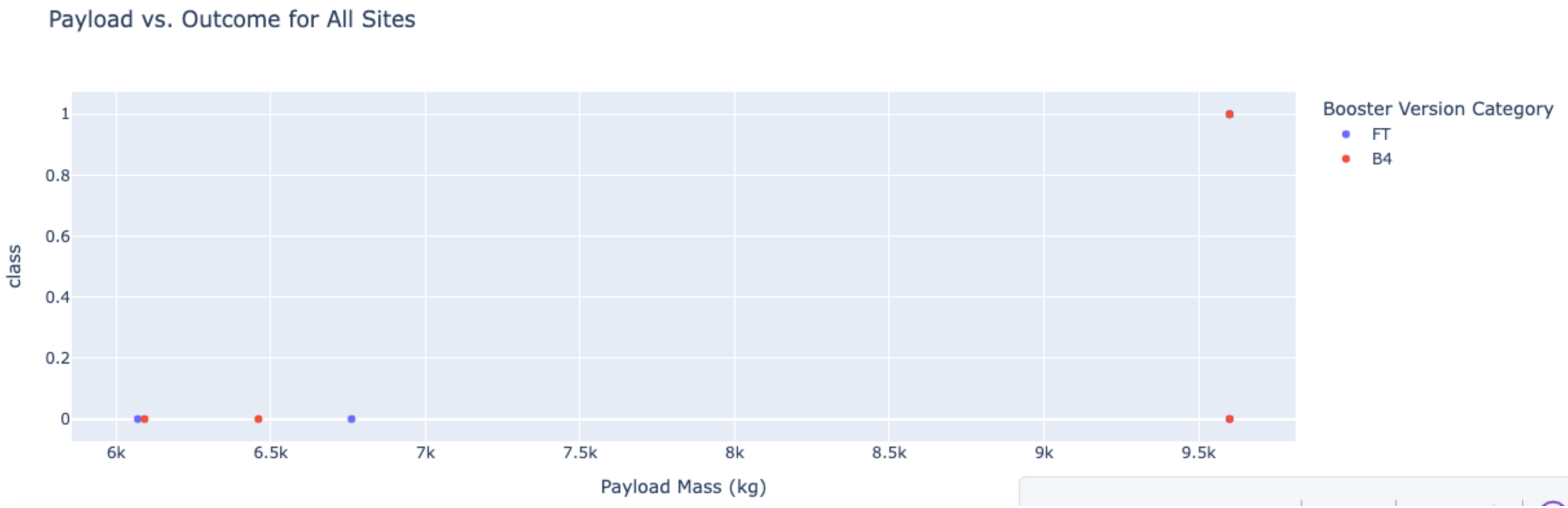
Payload range (Kg):

0100



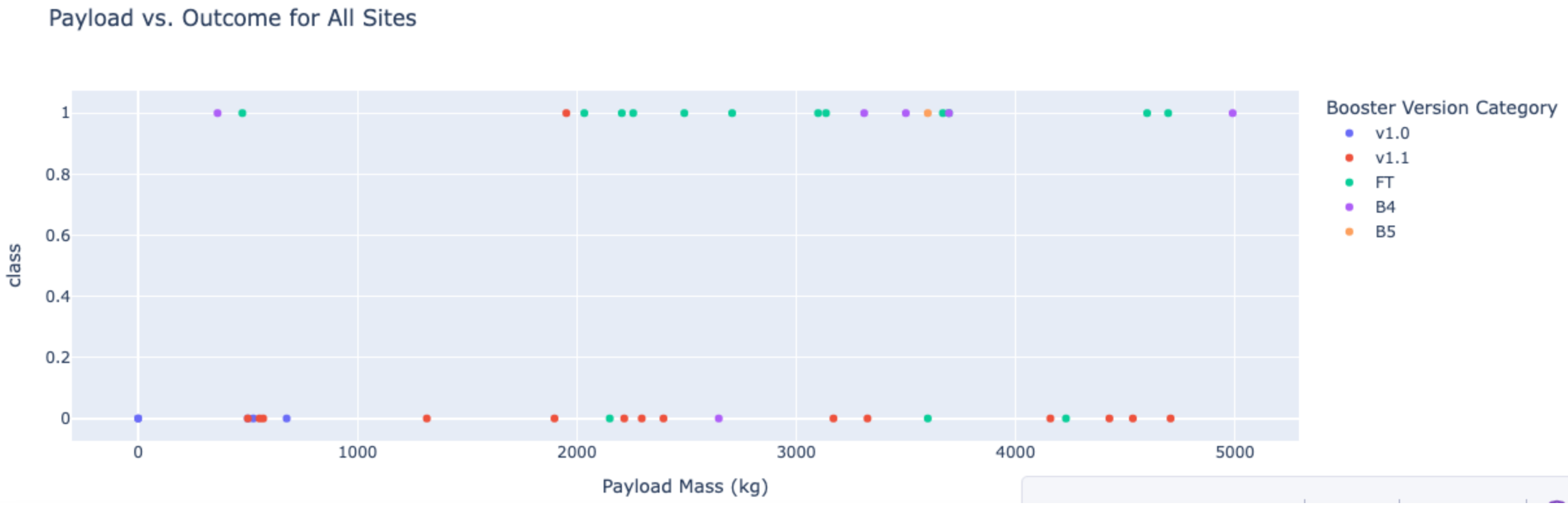
Payload range (Kg):

0100



Payload range (Kg):

0100

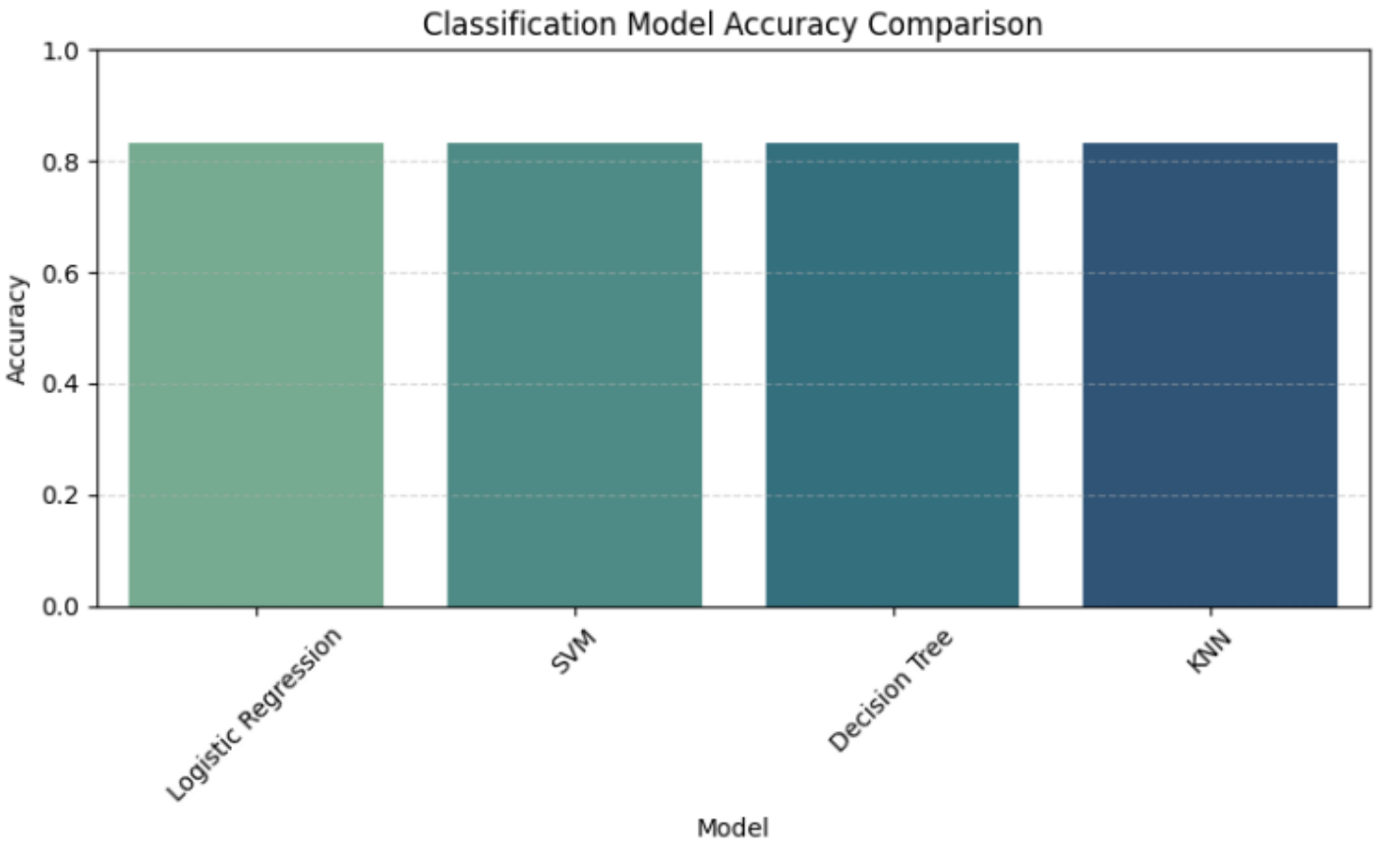


Section 5

Predictive Analysis (Classification)

Classification Accuracy

On the bar chart we can see that all the models have very similar classification accuracy. This can be because the test data is very small or there could have been data leakage.

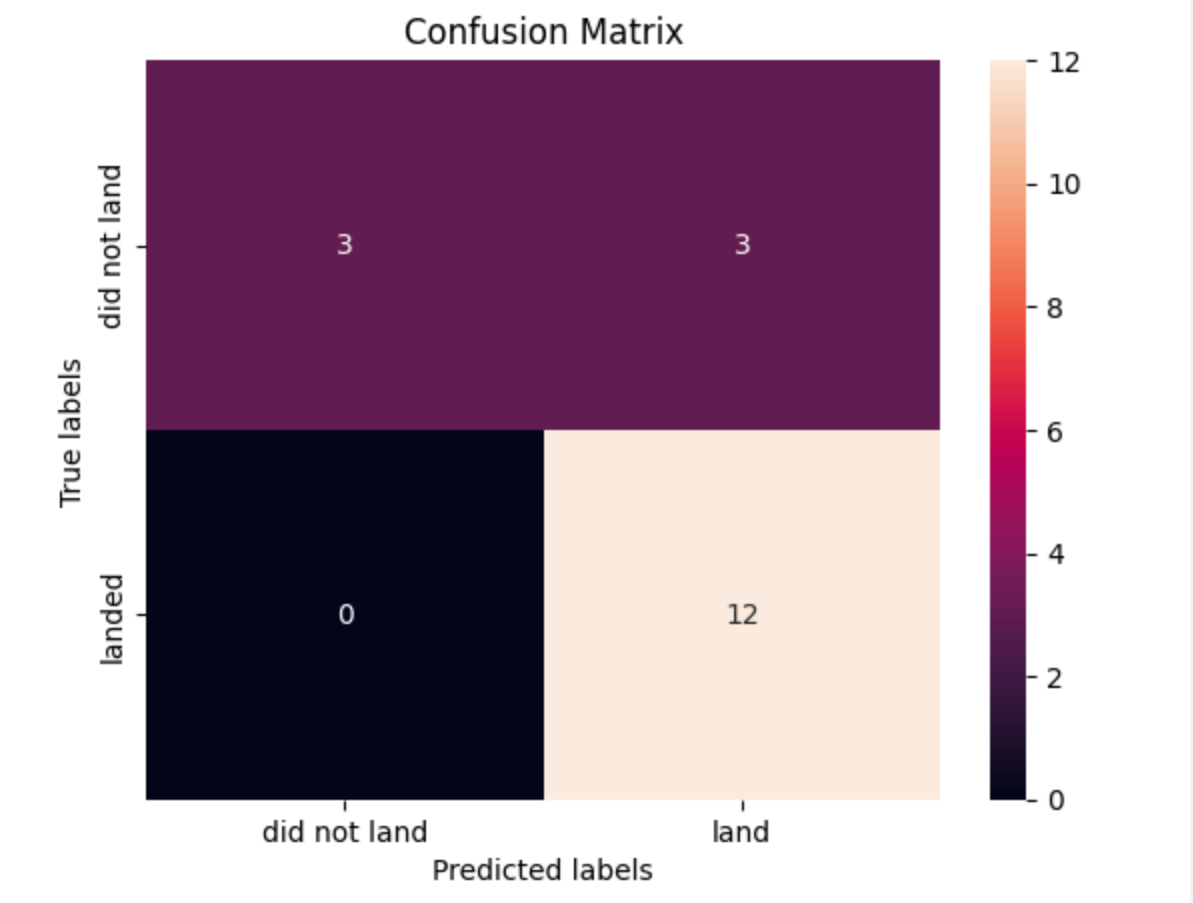


Confusion Matrix

The best performing model is the logistic regression model with an accuracy of 83.3%.

On the confusion matrix it is shown that:

- True Postive - 12 (True label is landed, Predicted label is also landed)
- False Postive - 3 (True label is not landed, Predicted label is landed)



Conclusions

- The project followed the CRISP-DM methodology, which gave a structured approach from problem understanding to model evaluation and deployment.
- Data collection was performed using both the SpaceX REST API and web scraping techniques, allowing for the extraction of relevant launch data from multiple sources.
- The extensive cleaning and wrangling of the data included handling missing values, transforming formats, validating accuracy, and structuring the dataset for analysis.
- Exploratory Data Analysis (EDA) showed important patterns, such as features like payload mass, launch site, and booster version. These were found to significantly influence landing success rates.
- Several classification models were built and compared, including logistic regression, decision trees, SVM, and KNN. These models were tuned using Grid Search and evaluated with cross-validation, accuracy and confusion matrix.
- Logistic regression model achieved the highest performance with an accuracy of 83.3%, showing consistent and reliable predictions of landing success.
- The visual analytics dashboard provided valuable tools for predicting launch outcomes.
- Interactive visualizations and maps helped present complex data a format that is easier to understand and this improves insights and exploratory analysis.

Thank you!