IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juana Robacio
08/11/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Summary of methodologies

- Gathered public SpaceX launch data (payloads, launch sites, orbits, outcomes).
- Performed exploratory data analysis (EDA) to identify patterns affecting first-stage recovery.
- Built interactive dashboards to visualize launch statistics and cost-related factors.
- Trained several machine learning models to predict first-stage landing success.

Summary of all results

- Identified key variables influencing landing success (payload mass, orbit type, launch site).
- Developed a predictive model. SVM, Logistic Regression and KNN were the most accurate.

Introduction

Project background & context:

- The commercial space age is expanding and SpaceX dominates the market thanks to its reusable first-stage rockets, which significantly reduce launch costs. However, first-stage recovery is not guaranteed — landings can fail, or the stage may be intentionally expendable.
- A reliable way is needed to estimate launch prices and understand when reuse is possible.

Problems to solve & questions:

- Can we predict whether the Falcon 9 first stage will land using public data?
- How does the probability of successful landing relate to launch cost estimation?
- What factors (payload, orbit, launch site, etc.) influence reuse feasibility?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: The dataset was obtained using a combination of web scraping and publicly available SpaceX launch data sources. HTML content was retrieved using HTTP requests and parsed with BeautifulSoup to extract launch site information, outcomes, etc.
- Perform data wrangling: The dataset was cleaned, standardized, and merged. Missing values were handled, types were corrected, and new features were engineered.
- Perform exploratory data analysis (EDA) using visualization and SQL: Both visualization tools and SQL queries were used to explore launch frequencies, success rates, temporal trends, and relationships among variables.
- Perform interactive visual analytics using Folium and Plotly Dash: Folium maps were used to build interactive geographic visualizations, showing launch site locations, success/failure markers and distances to locations while Dash displayed launch statistics.
- Perform predictive analysis using classification models: The Machine learning workflw included: Splitting data into training and test sets, building baseline classifiers, tuning hyperparameters using GridSearchCV and evaluating performance with accuracy scores and confusion matrices.

Data Collection

The process of Data Collection in this Project was the following:

1. Identify data sources (SpaceX API, Webpages) → 2. Send HTTP requests (requests library) → 3. Parse HTML with *BeautifulSoup* → 4. Extract launch data (sites, dates, outcomes) → 5. Clean & normalize fields (strings, numbers, coordinates) → 6. Store in DataFrame (Pandas library) → 7. Merge with external data (CSV, API sources)

Data Collection – SpaceX API

Identify data source → Call REST API with get requests

↓

Receive JSON response → Parse JSON content

↓

Extract relevant fields (launch date, site, payload, outcome)

↓

Clean & normalize data → Store in Pandas DataFrame

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

Data Collection - Scraping

Identify web pages → Send HTTP request → Retrieve HTML

↓

Parse HTML with BeautifulSoup

↓

Extract relevant fields (e.g., booster version, mission type)

↓

Clean & normalize data → Store in DataFrame

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Load raw data
  ↓
- Handle missing values
  ↓
- Convert data types
  ↓
- Standardize column names
  ↓
- Filter and select relevant features
  ↓
- Create new derived features
  ↓
- Prepare final dataset for analysis

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts used:

- bar charts to compare successes and failures
- launch-site bar charts to identify the most successful locations
- scatter plots to examine how payload mass influenced landing outcomes
- trend plots to observe improvements over time
- count/pie charts to understand the distribution of categorical features such as booster versions and orbits.

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/edadataviz.ipynb

EDA with SQL

Summary of SQL queries:

- Filtered launches by success/failure to compute success rates.
- Selected launch sites, payloads, booster versions for analysis.
- Aggregated launches by year and site to identify trends.
- Calculated average payload mass per site, identifying unique launch sites.
- Joined multiple tables to combine launch info with booster and payload details.

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Objects created:

- Markers: Represented each launch site and individual launches.
- Marker Clusters: Grouped dense regions of launches to reduce map clutter.
- Circles: Indicated the area surrounding launch sites for visual emphasis.
- Lines (PolyLine): Connected launch sites to nearest coastlines to visualize distances.
- Popups /DivIcons: Displayed site names, distances, or success/failure info interactively.

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

Objects used:

- Bar Chart: Showed successes vs failures, launch site comparisons.
- Scatter Plots: Explored payload vs success, with success being class 1 and failure class 0. Using different colours for different booster versions.
- Dropdown menus: Allowed users to select specific sites, payload ranges, or years.

- https://github.com/juagr8-afk/Coursera-notebook/blob/main/spacex-dash-appp.py

Predictive Analysis (Classification)

Import scikit-learn and Split data into train/test

↓

Select features

↓

Train classification models

↓

Tune hyperparameters → Evaluate models (accuracy, confusion matrix) → Compare model performance

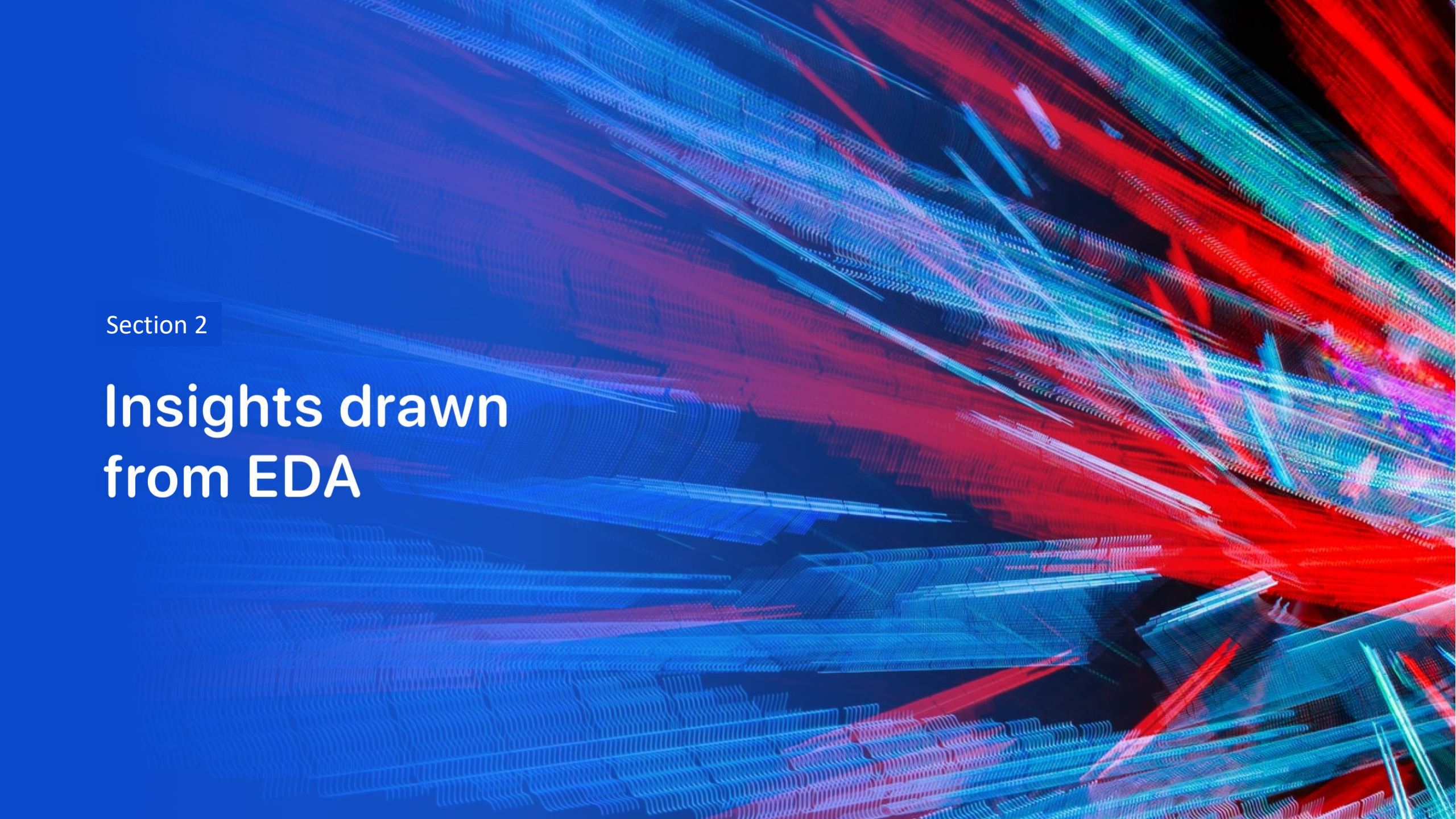
↑

Make predictions on new launches

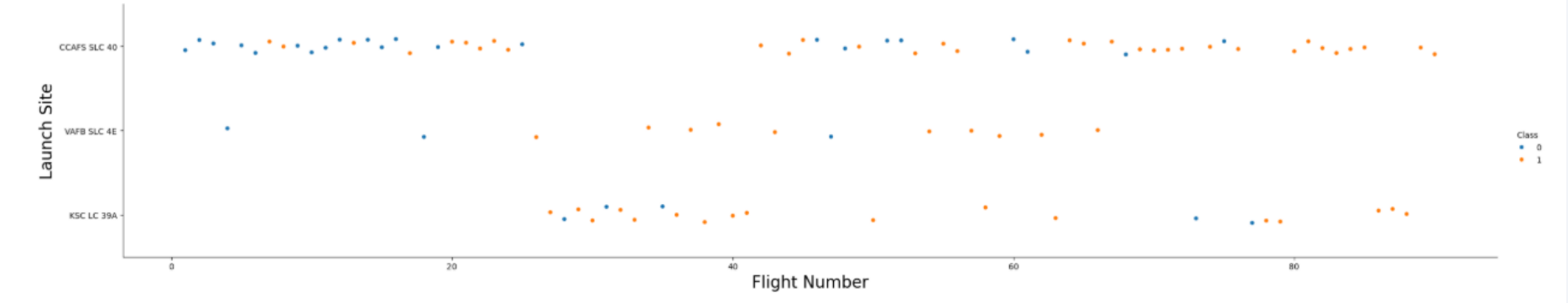
- https://github.com/juagr8-afk/Coursera-notebook/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis: EDA showed that KSC LC-39A was the most successful launch site, with the highest proportion of successful missions. The analysis also identified newer Booster versions and payload as the features most associated with launch success.
- Interactive analytics: Dashboards and Folium maps also revealed that the two closely located sites at Cape Canaveral (CCAFS SLC-40 and CCAFS LC-40) formed the densest cluster of launches, concentrating a large number of missions in a small geographic area near the coast and relatively far from cities. Color-coded markers showed that this region also had a moderate rate of successful launches.
- Predictive analysis:Logistic Regression, KNN, and SVM achieved the highest accuracy (~83%). Confusion matrices showed strong true-positive and true-negative performance, displaying that most launches will be successful.

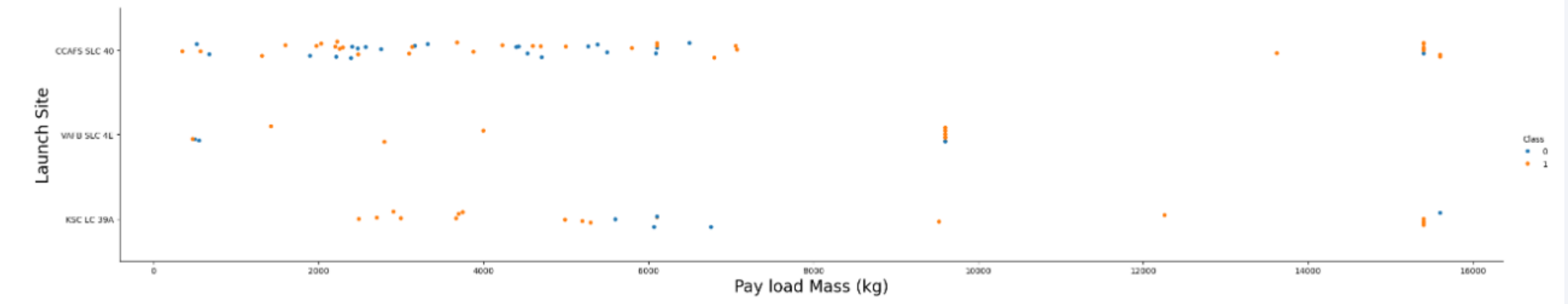

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- Looking at the scatter plot, in the first flight numbers there are more successes for CCAFS SLC 40 while from flight number 20 the successes decrease, being more spread out. The same happens for the other 2 locations.

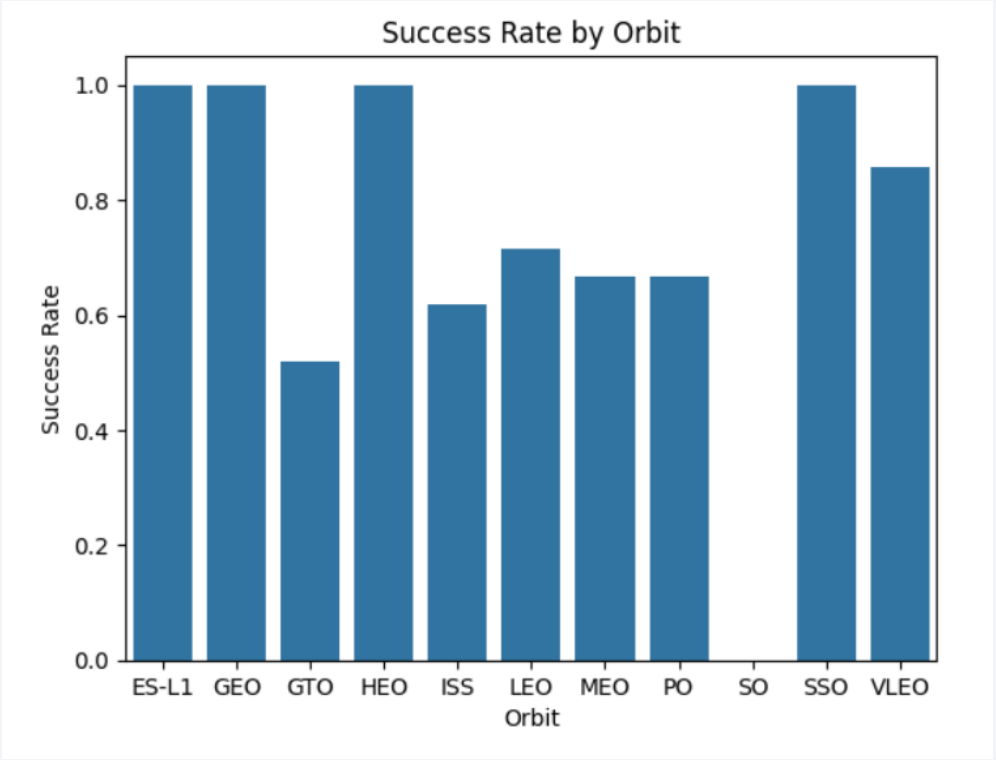
Payload vs. Launch Site



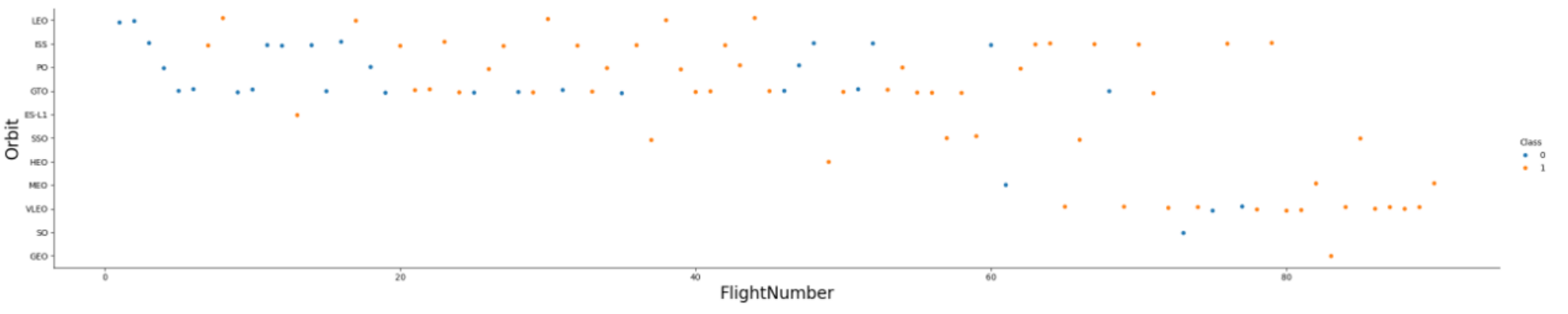
- It can be seen from the plot that CCAFS SLC 40 has the most successes for a payload mass from 2000 approximately to 6000.
- The other two sites don't have as many launches, but show successes for a very small payload range.

Success Rate vs. Orbit Type

- The most successful orbits are GEO, HEO, ES-L1 and SSO.
- The least successful one is GTO.

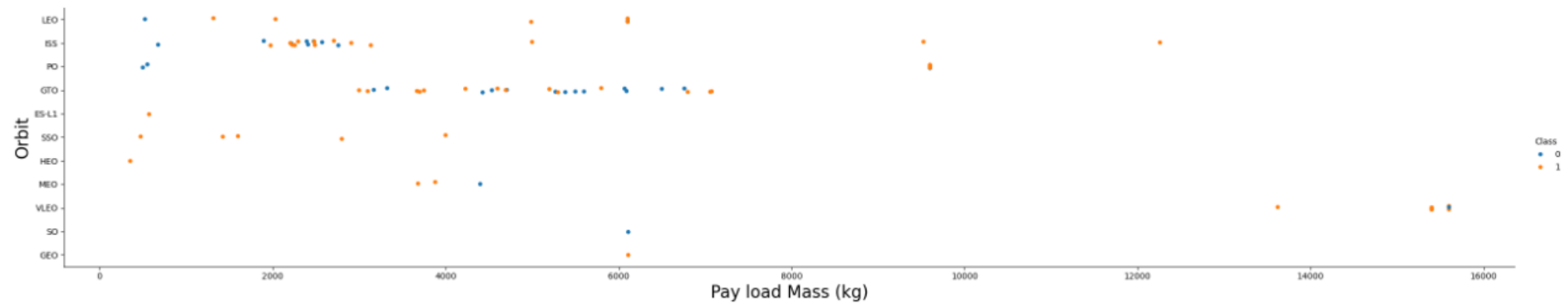


Flight Number vs. Orbit Type



- 1SS shows the most successes overall, while GTO shows a seemingly equal amount of successes and failures.
- VLEO seems to have the lowest failure ratio from flight number 60 on.
- SSO shows only a few successes.

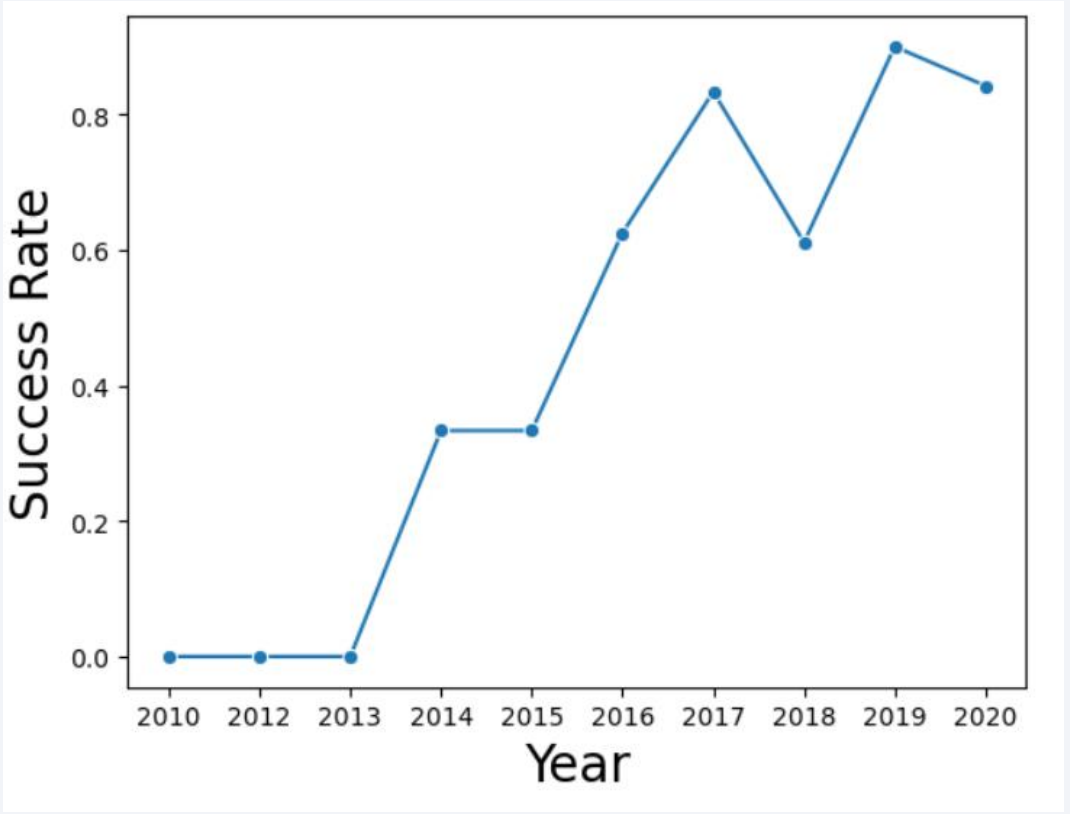
Payload vs. Orbit Type



- GTO shows the biggest number of failures for 3000 to 7000 kg payload mass, while 1SS shows some successes from 2000 kg to 3000 kg approximately.
- The rest show mostly succeses or no data.

Launch Success Yearly Trend

- Success rate increases rapidly since 2013.
- There's an upward trend in success rate overall, with a slight dip from 2017 to 2018.



All Launch Site Names

- Finding all unique names of launch sites yields the following results:

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

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

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

- Overall, the successes outnumber the failures by a great number.

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

Landing_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

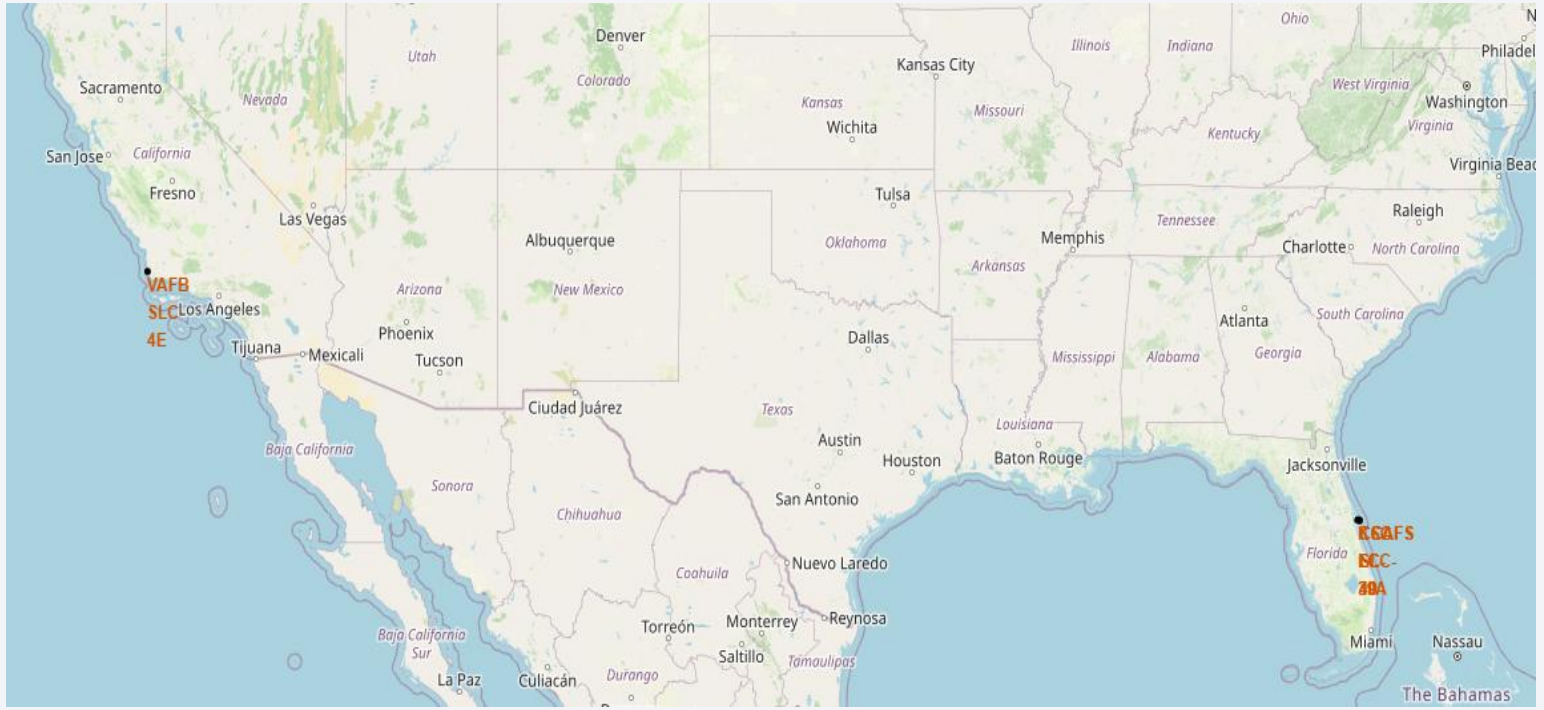
Section 3

Launch Sites Proximities Analysis

Launch Sites in Folium

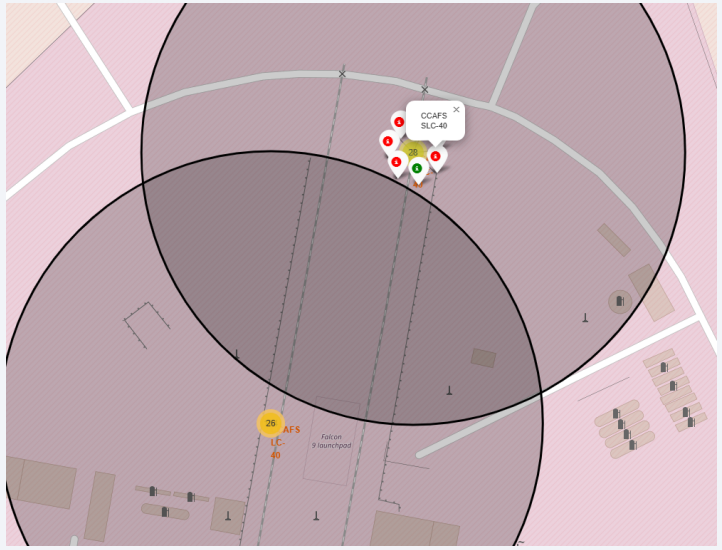
- The Florida cluster includes CCAFS SLC-40, CCAFS LC-40, and KSC LC-39A, which are geographically close and account for most of the launches.

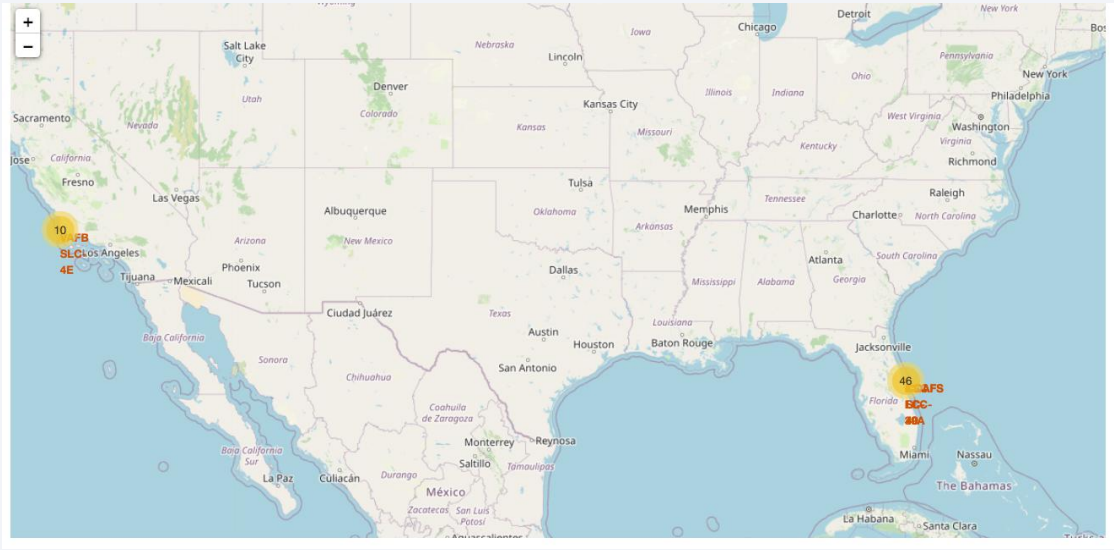
- The California cluster is more spread out and has fewer launches overall



Successes and Failures

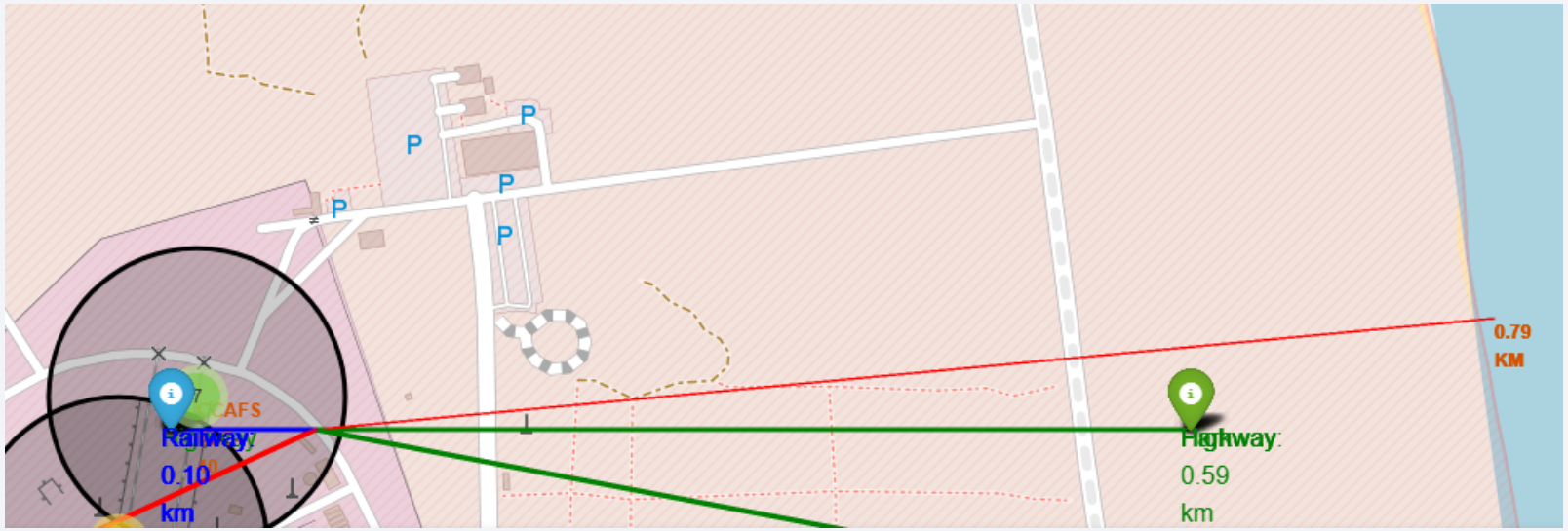
- On one side we have 10 missions carried out and on the other, 46, grouped by Launch Site. In green we can see successes and in red, failures.

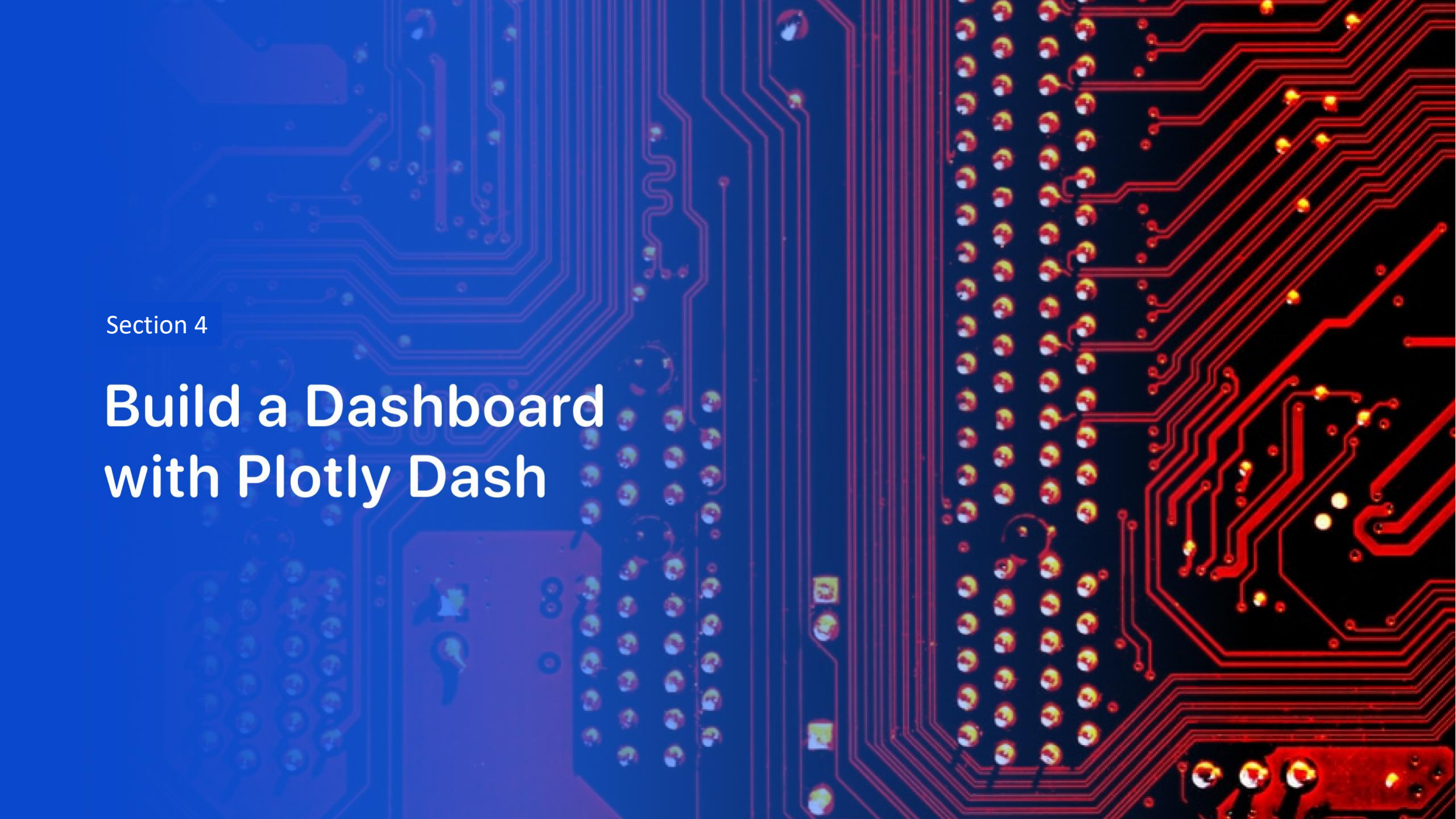




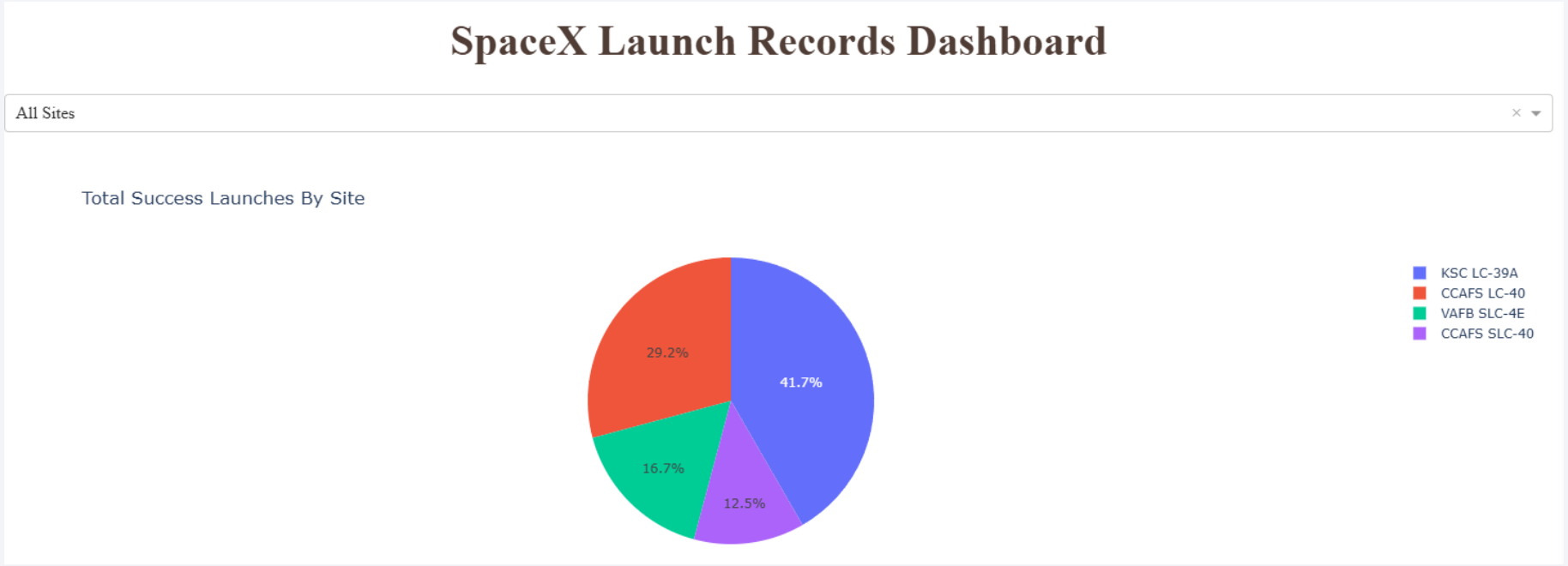
Nearby locations

- The launch sites in Florida (CCAFS SLC-40, CCAFS LC-40, KSC LC-39A) are located near highways, and railway lines, providing essential infrastructure for transport and logistics.
- Their proximity to the coastline is critical for safety, as launches over the ocean reduce risk to populated areas.



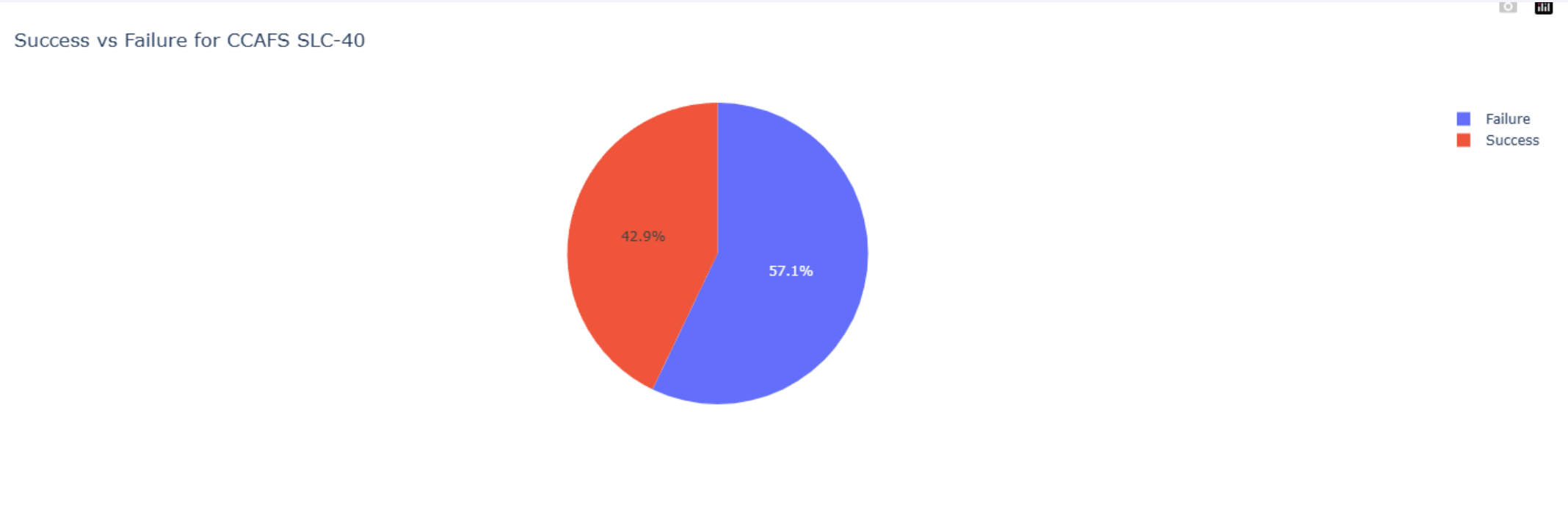

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches by Site



- As seen in the image, the Launch Site with most successful launches is KSC LC 39A, followed by CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40.

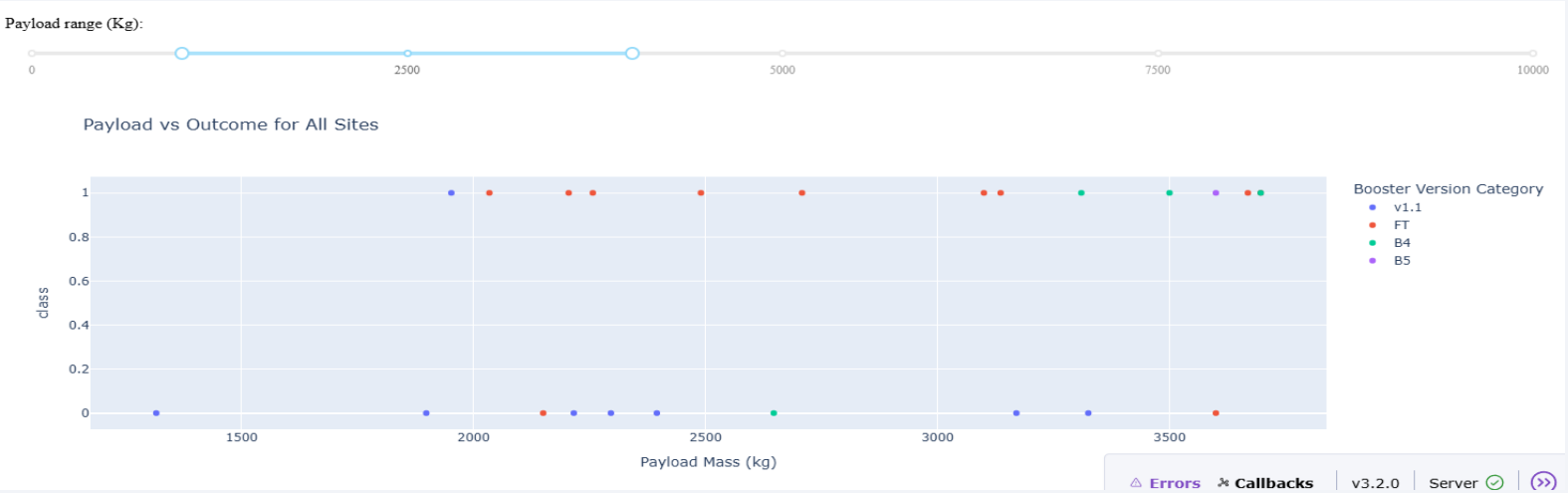
Successes and Failures per Launch Site



- Successes and Failures for the Launch Site with highest success-to-failure ratio, CCAFS SLC-40.

Payload vs Launch Outcome

- The payload range and booster version with the largest success rate is between 2k and 5k for the Falcon 9 FT.

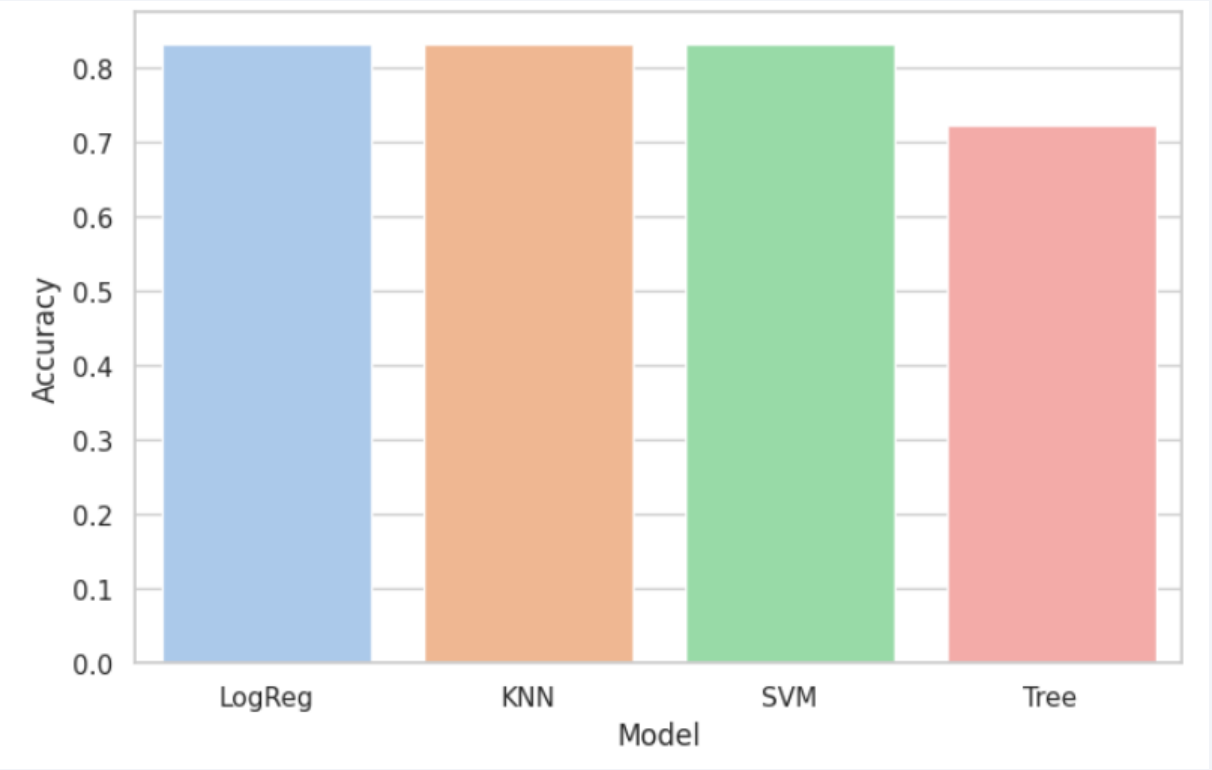


Section 5

Predictive Analysis (Classification)

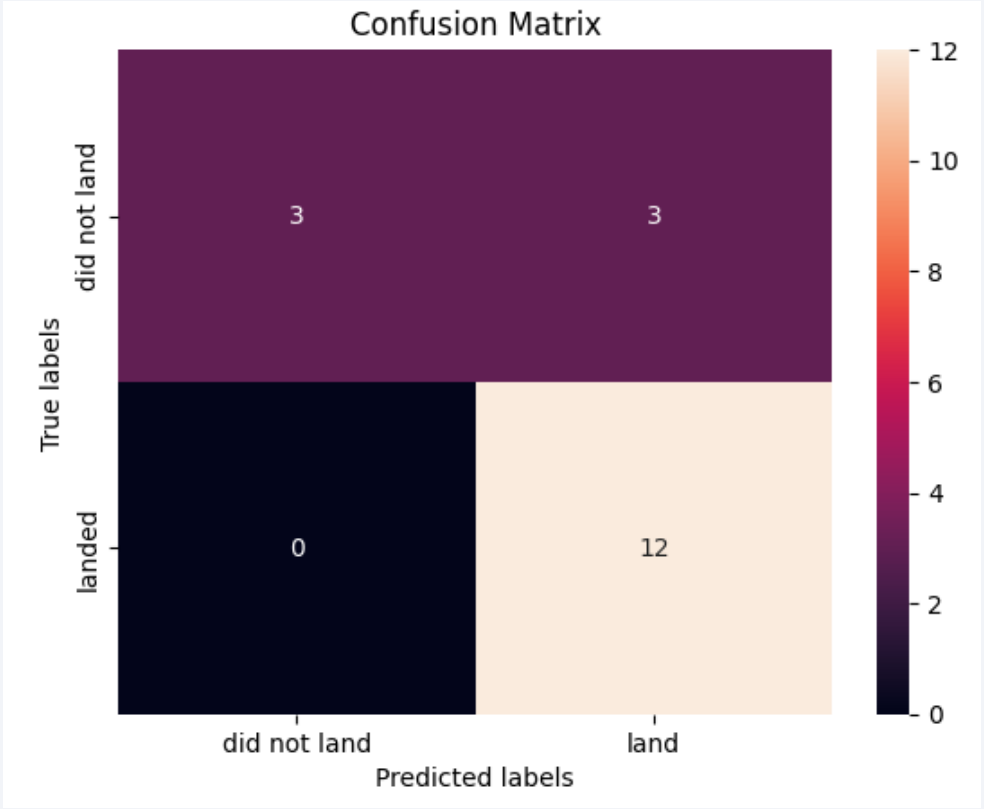
Classification Accuracy

- Decision Tree showed the lowest accuracy (~0.72), likely due to overfitting and high model variance because random state wasn't fixed, while Logistic Regression, Support Vector Machines and K-Nearest Neighbors all showed an accuracy of approximately 83%.



Confusion Matrix

- This is the confusion matrix of the Logistic Regression model.
- As shown, the confusion matrix shows higher True Positives and True Negatives, meaning the model correctly predicted both successful and failed landings most of the time. Misclassifications (False Positives/Negatives) were limited, indicating balanced predictive power.



Conclusions

- GEO, HEO, SSO and ES-L1 are the most successful orbits.
- CCAFS SLC-40 and KSC LC-39A had the largest number of successful launches, and most successful launches had payloads roughly between 2000 kg and 5000 kg, where success rates were consistently high, with Falcon 9 FT being the most successful Booster version.
- Confusion matrices confirmed the model's performance, showing high accuracy predicting mostly successful launches.

Thank you!