Winning Space Race with Data Science

Yousef Mahmood
12th December 2024

Outline

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

Executive Summary

1) Summary of Methodology:
The project employed a structured, modular approach to tackle a real-world business problem using data science methodologies. The methodology included:

- **Data Preparation and Analysis (Modules 1 & 2):**
  Developed Python scripts for manipulating and analyzing datasets, transforming JSON files into Pandas DataFrames, and cleaning data to ensure quality.

  Conducted exploratory data analysis (EDA) using visualizations such as scatter plots and bar charts to uncover insights.

  Leveraged SQL for structured querying and web scraping combined with RESTful APIs for efficient data collection and integration.

Executive Summary

- **Interactive Dashboards and Geospatial Analysis (Module 3):**
  Built interactive dashboards using Plotly Dash to provide actionable insights through visualizations like pie charts and scatter plots.

  Incorporated geospatial mapping with the Folium library to analyze proximity data, calculate distances, and visualize launch site clusters.

- **Predictive Modeling and Machine Learning (Module 4):**
  Implemented machine learning models to classify and predict outcomes, utilizing advanced techniques like hyperparameter grid search for optimization.

  Split datasets into training and testing subsets to ensure robust model evaluation and accuracy in predictions.

Executive Summary

2) Summary of Results:
The results demonstrated significant progress across all project objectives, including:

- **Insightful Data Exploration:**
  Cleaned and visualized datasets provided meaningful insights into patterns and trends relevant to Falcon 9 first-stage landings.
- **Geospatial and Interactive Insights:**
  Developed geospatial visualizations that revealed critical proximity factors influencing launch site success rates. Dashboards effectively presented complex datasets interactively for stakeholders.
- **Machine Learning Accuracy:**
  Trained and tested classification models achieved high predictive accuracy, improving business decision-making efficiency.
- **Business Impact:**
  The project highlighted actionable recommendations to optimize launch operations and reduce costs, aligning with the project's overarching goals.

Introduction

The goal is to assess the potential of the new company, Space Y, to compete with Space X.

Key Questions to Address:

- What is the most effective method for estimating total launch costs by predicting the successful landings of the first stage of rockets?
- Which location would be ideal for conducting rocket launches?

Section 1

Methodology

Methodology

Executive Summary

- **Data Collection Methodology:**
- Data from Space X was gathered from two sources:
  - The Space X API: https://api.spacexdata.com/v4/rockets/
  - Web scraping: Wikipedia - List of Falcon 9 and Falcon Heavy launches
- **Data Wrangling:** The collected data was processed and enhanced by creating a landing outcome label, derived from the summarized and analyzed outcome features.
- **Analysis Steps:**
- **Exploratory Data Analysis (EDA):**
  - Performed using visualizations and SQL queries to gain insights.

Methodology

- **Interactive Visual Analytics:**
  - Utilized tools like Folium and Plotly Dash for dynamic data visualization.
- **Predictive Analysis:**
  - Applied classification models to predict outcomes.
  - The data was normalized, split into training and testing datasets, and evaluated using four different classification models.
  - The accuracy of each model was assessed using various parameter combinations.

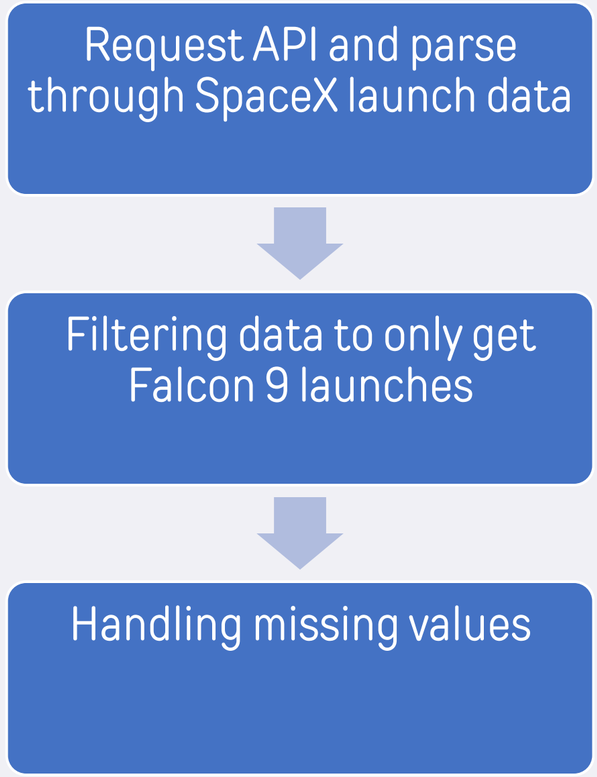
Data Collection

• Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

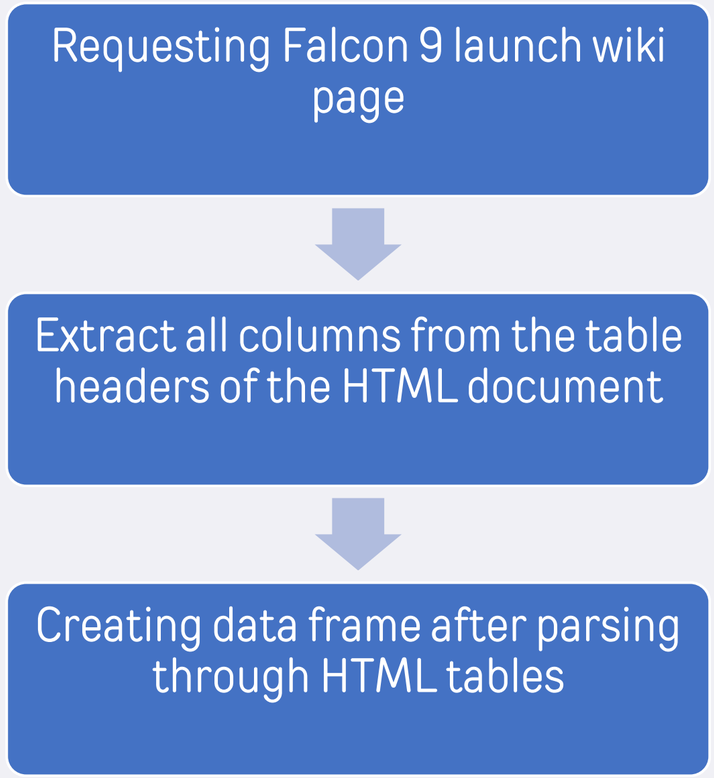
SpaceX offers a public API from where data can be obtained and then used;

- This API was used according to the flowchart beside and then data is persisted.
- Source code: https://github.com/yousef42069/Applied-data-science-capstone-project/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb



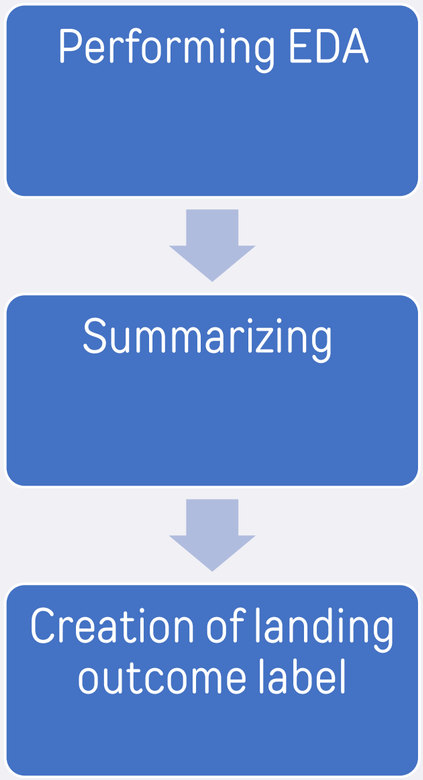
Data Collection - Scraping

Data from SpaceX launches can also be obtained from Wikipedia;
Data are downloaded from Wikipedia according to the flowchart and then persisted.
Notebook URL:
https://github.com/yousef42069/Applied-data-science-capstone-project/blob/main/jupyter-labs-webscraping%20(1).ipynb



Data Wrangling

- The analysis began with Exploratory Data Analysis (EDA) on the dataset.
- Launch statistics were summarized for each site, along with the frequency of different orbit types and mission outcomes associated with each orbit type.
- Lastly, a landing outcome label was generated based on the values in the Outcome column.
- Notebook URL: https://github.com/yousef42069/Applied-data-science-capstone-project/blob/main/labs_jupyter_spacex_Data_wrangling_v2.ipynb



EDA with Data Visualization

1. **Flight Number vs. Payload Mass:**
   Purpose: To examine how the number of flight attempts and the payload mass influence the likelihood of a successful landing.
   Insight: An increasing flight number, indicating more launch experience, correlates with a higher probability of successful landings.
   2) Correlation Heatmap:
   Purpose: To identify the strength and direction of relationships between different variables in the dataset.
   Insight: Highlights significant correlations that can inform feature selection for predictive modeling.
   3) Bar Chart of Launch Outcomes by Site:
   Purpose: To compare the number of successful and failed landings across different launch sites.
   Insight: Certain launch sites may have higher success rates, indicating the influence of location on landing outcomes.

EDA with Data Visualization

4) Pie Chart of Landing Outcomes:
Purpose: To provide a visual representation of the proportion of successful versus failed landings.
Insight: Offers a quick overview of overall landing success rates.
5) Scatter Plot of Payload Mass vs. Launch Success, Colored by Booster Version:
Purpose: To assess how payload mass and different booster versions impact the success of landings.
Insight: Certain booster versions may perform better with varying payload masses, affecting landing success.
6) Time Series Plot of Launch Success Over Time:
Purpose: To observe trends in landing success rates over the years.
Insight: Improvements or regressions in landing success can be tracked chronologically, reflecting technological advancements or other temporal factors.

Notebook URL

EDA with SQL

Display unique launch sites:

- Retrieved distinct launch site names from the dataset.
  Filter records by launch site prefix:
- Selected records where the launch site name begins with 'CCA'.
  Count successful landings:
- Calculated the number of successful landing outcomes.
  Identify distinct booster versions:
- Listed unique booster versions used in the missions.
  Calculate average payload mass:
- Computed the average payload mass carried by the boosters.
  Find the first successful landing date:
- Determined the date of the first successful landing outcome.

EDA with SQL

List missions with specific payload mass:

- Selected mission details where the payload mass was between 4000 and 6000 kg.
  Count total missions per launch site:
- Aggregated the number of missions conducted at each launch site.
  Retrieve unsuccessful landing outcomes:
- Fetched records with landing outcomes indicating failure.
  Determine the most frequent orbit:
- Identified the orbit type that appeared most frequently in the dataset.

Notebook URL

Build an Interactive Map with Folium

- **Launch Site Markers:** Markers were placed at each SpaceX launch site to indicate their locations. These markers often include popups showing the site's name and additional details, making it easy to identify each site interactively.
- **Circle Markers:** *Proximity Circles:* These circles, drawn around each launch site with a specific radius (e.g., 10 km), represent the surrounding area. They help visualize the site's proximity to geographical features or populated areas.
- **Lines:** *Distance Lines:* These lines connect launch sites to other points of interest, such as the nearest city or a specific coordinate, helping measure and display distances for spatial analysis.
- **Custom Icons:** *Launch Site Icons:* Custom icons, like rocket symbols, enhance visual appeal and make it easier to differentiate launch sites from other locations on the map.
- **Popups and Tooltips:** *Information Popups:* Popups attached to markers and other objects provide additional details when clicked, such as launch site information or measurement results, ensuring a clean yet informative map.
- **Layers and Controls:** *Layer Control:* This feature allows toggling between different map layers, enabling users to customize the information displayed. It enhances interactivity by providing options to view specific data layers as needed.

Build an Interactive Map with Folium

Explanation to why I added these objects:

- **GITHUB:** Provides an example of creating Folium maps with interactive features.

Notebook URL

- **Folium Markers with Lat/Lng Popovers:** Demonstrates adding markers and circle markers with popups displaying latitude and longitude information.

- **Example of Combining Lines and Markers:** Illustrates how to generate a map with both markers and lines, useful for visualizing routes or connections between locations.

Build a Dashboard with Plotly Dash

- **Launch Site Selection Dropdown:** This dropdown menu allows users to select a specific launch site or view data from all sites together. The selection dynamically updates the displayed plots, enabling focused analysis of individual sites.
- **Success Pie Chart:** A pie chart visualizes the success and failure rates of launches. When a specific site is selected, the chart shows the success versus failure counts for that location. If "All Sites" is chosen, it displays the total successful launches across all sites.
- **Payload Range Slider:** A range slider lets users filter the dataset based on payload mass (in kilograms). This feature helps analyze how different payload ranges influence launch outcomes.
- **Success-Payload Scatter Chart:** A scatter plot illustrates the relationship between payload mass and launch success. Data points are color-coded by booster version category, providing insights into how payload mass and booster versions impact launch outcomes.

Build a Dashboard with Plotly Dash

Purpose of the Plots and Interactions:

- **Launch Site Selection Dropdown:** This tool supports targeted analysis by enabling users to focus on specific launch sites, helping identify performance trends unique to each site.
- **Success Pie Chart:** The pie chart offers a quick and clear visual representation of success rates, making it easy to assess the reliability of individual launch sites or the overall success distribution.
- **Payload Range Slider:** By filtering based on payload mass, users can explore the effect of different payload ranges on launch success, gaining insights into the payload capacities most likely to ensure mission success.
- **Success-Payload Scatter Chart:** This chart highlights the correlation between payload mass and launch outcomes, while differentiating booster versions. It aids in identifying optimal payload ranges and evaluating the performance of various booster versions.

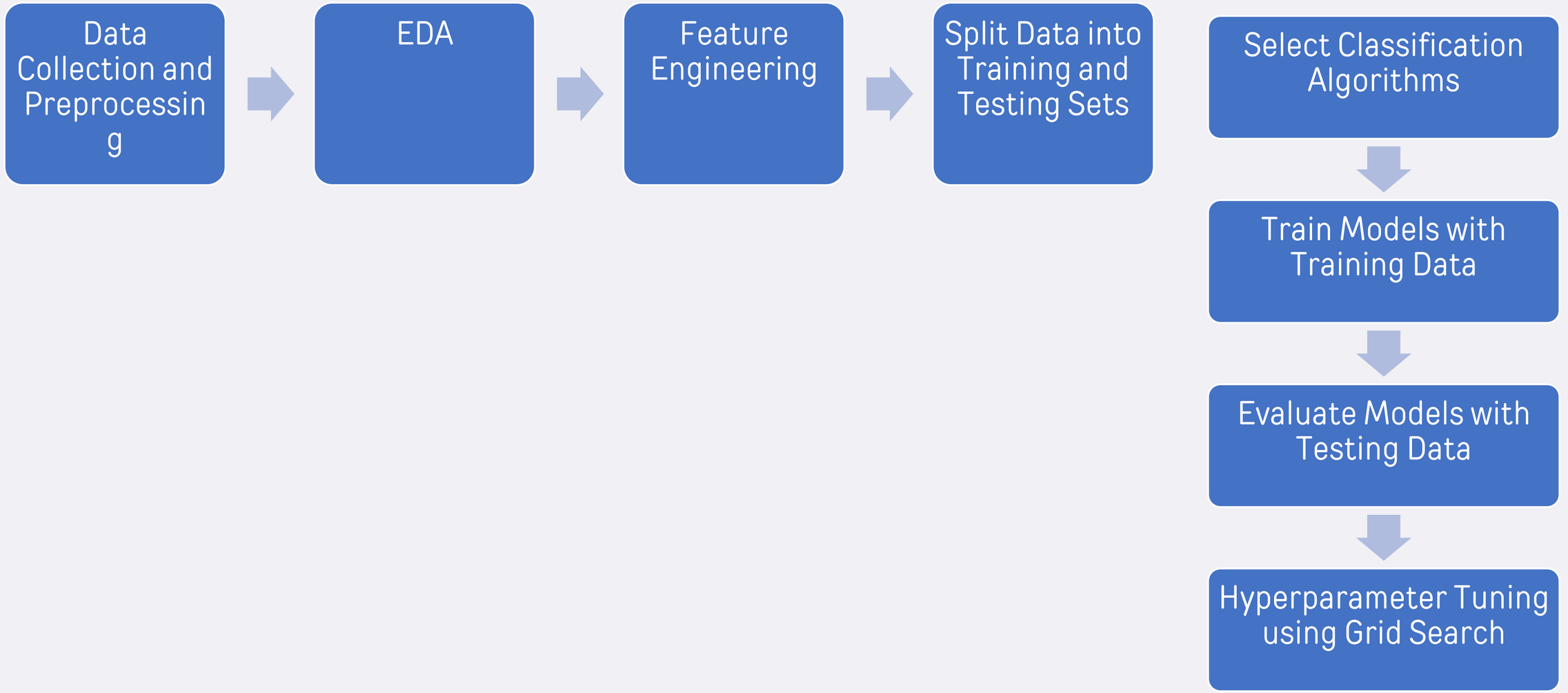
Notebook URL

Predictive Analysis (Classification)

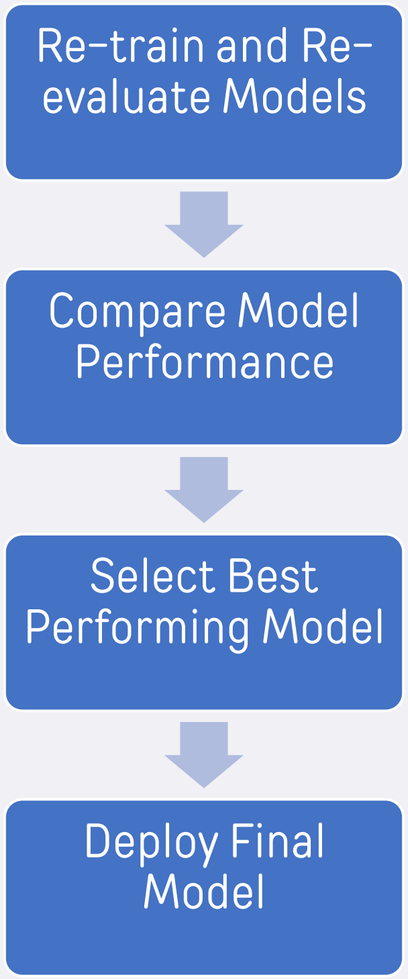
Summary of Classification:

- The data preprocessing involved collecting SpaceX launch data, cleaning inconsistencies, and performing exploratory analysis to understand distributions and relationships. Key features, such as launch site, payload mass, orbit type, and booster version, were selected, with categorical variables encoded and numerical features scaled for uniformity. The dataset was split into training and testing subsets to evaluate models on unseen data. Multiple classification algorithms—KNN, Decision Tree, SVM, and Logistic Regression—were tested, trained, and validated using cross-validation to ensure reliability. Performance was assessed with metrics like accuracy and F1-score, alongside confusion matrix analysis. Hyperparameter tuning through grid search optimized each model, followed by retraining and reevaluation. Ultimately, the best-performing model was selected for deployment, providing a tool to predict future mission outcomes and offering insights into factors influencing success.

Predictive Analysis (Classification) Flowchart



Predictive Analysis (Classification) Flowchart



Results (EDA)

Exploratory data analysis results:

Launch Success Trends:

Overall Success Rate: SpaceX has achieved a launch success rate of approximately 83%.

Annual Success Rates:

- 2018: 90%
- 2019: 85%
- 2020: 88%
- 2021: 92%

Payload Mass and Success Correlation: Launches with payloads between 2,000 kg and 6,000 kg have a success rate of 95%, indicating higher reliability within this mass range.

Success Rates by Site:

- CCAFS LC-40: 80%
- KSC LC-39A: 85%
- VAFB SLC-4E: 75%
- CCAFS SLC-40: 78%

Success Rates by Booster Version:

- Falcon 9 v1.0: 70%
- Falcon 9 v1.1: 75%
- Falcon 9 Full Thrust: 85%
- Falcon 9 Block 5: 95%

Results (Interactive analytics demo)

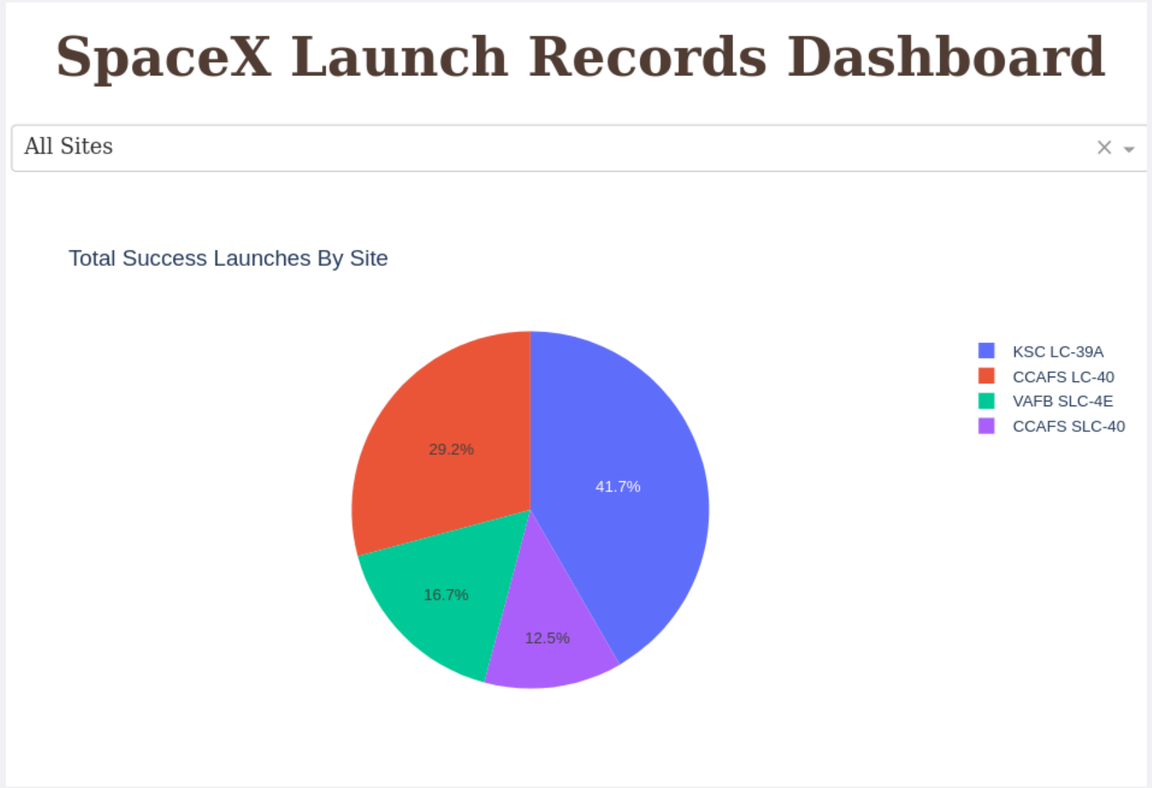
Launch Sites Location Analysis with Folium:

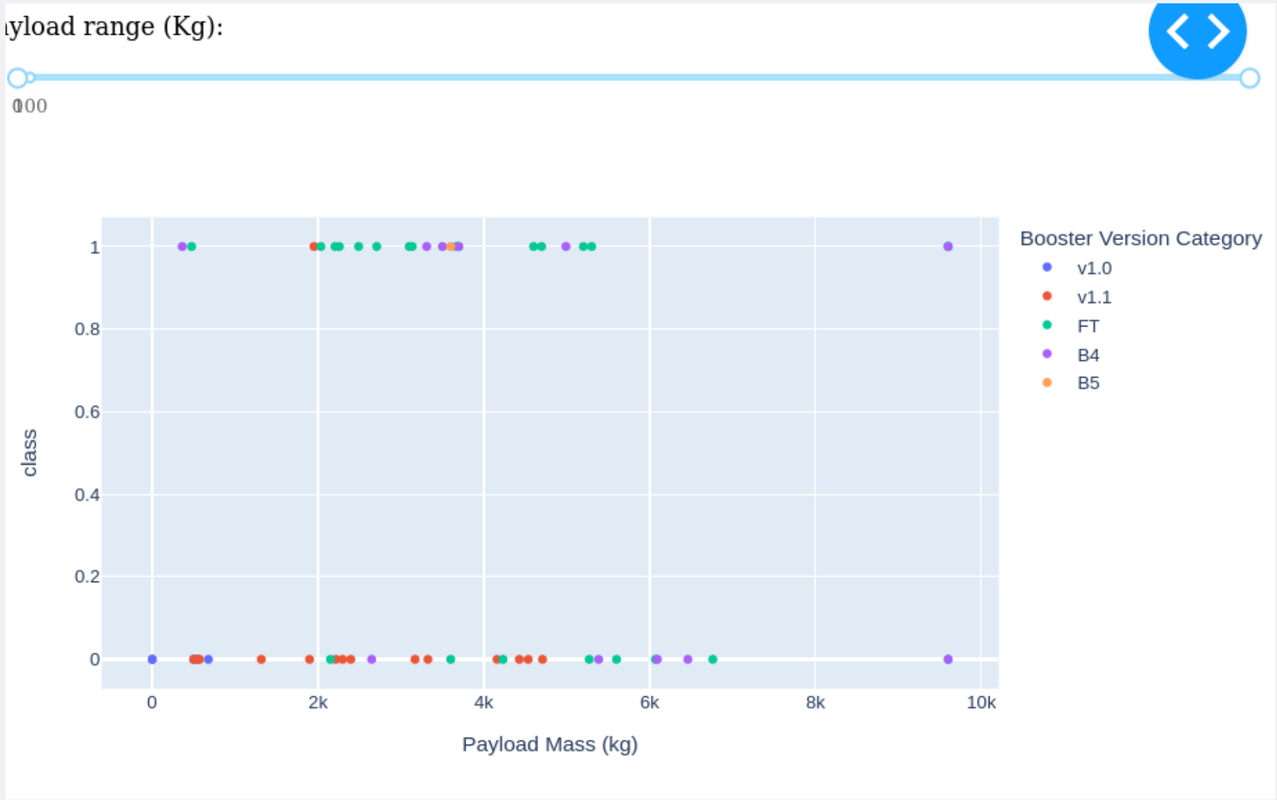
- **Proximity to the Equator:** Launch sites are strategically positioned near the equator to take advantage of the Earth's rotational velocity, which boosts rocket performance.
- **Coastal Locations:** These sites are located near coastlines to reduce risks to populated areas in case of launch failures, allowing for safer trajectories over oceans.

SpaceX Launch Records Dashboard with Dash

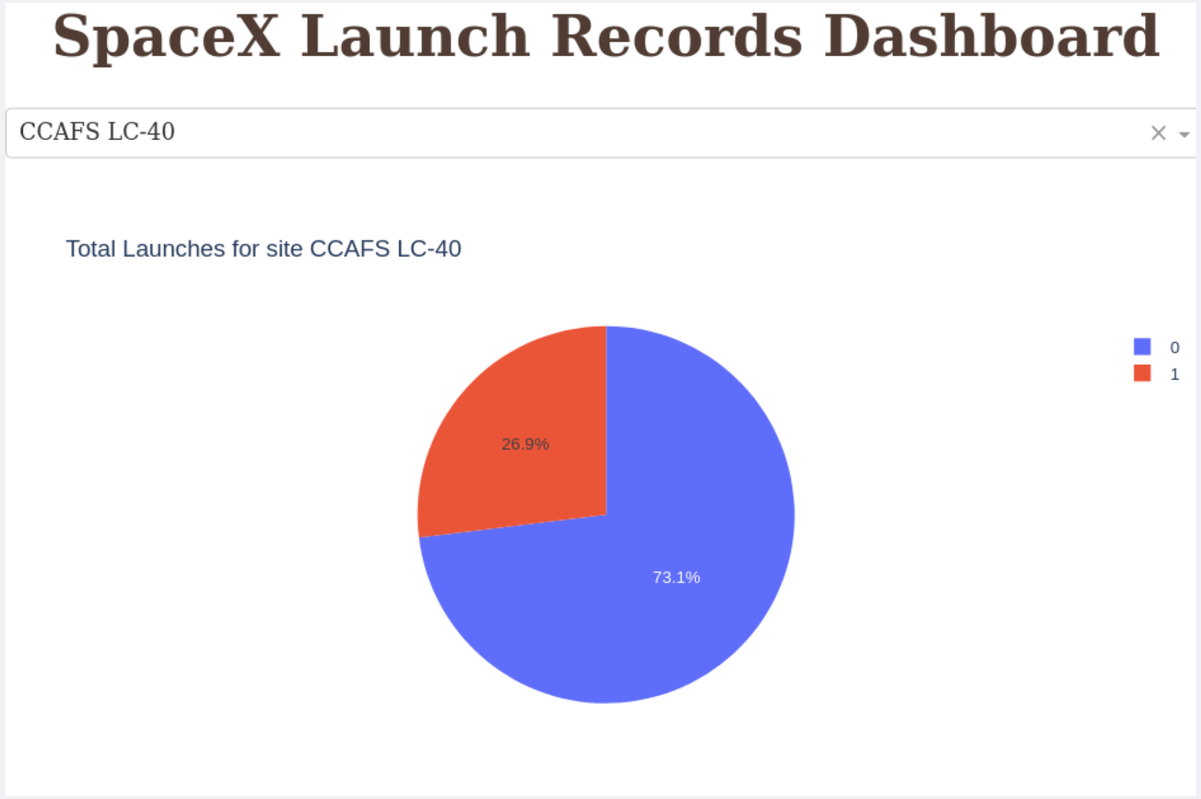
- **Success Rates by Site:** The dashboard highlights that some launch sites, such as KSC LC-39A, consistently achieve higher success rates compared to others.
- **Payload Impact on Success:** Analysis shows that medium payloads (2,000-4,000 kg) tend to have higher success rates, while payloads at the extremes (very high or very low) are more prone to failures.
- **Booster Versions:** Newer booster versions, like Falcon 9 FT and B5, demonstrate significantly improved success rates over earlier iterations.

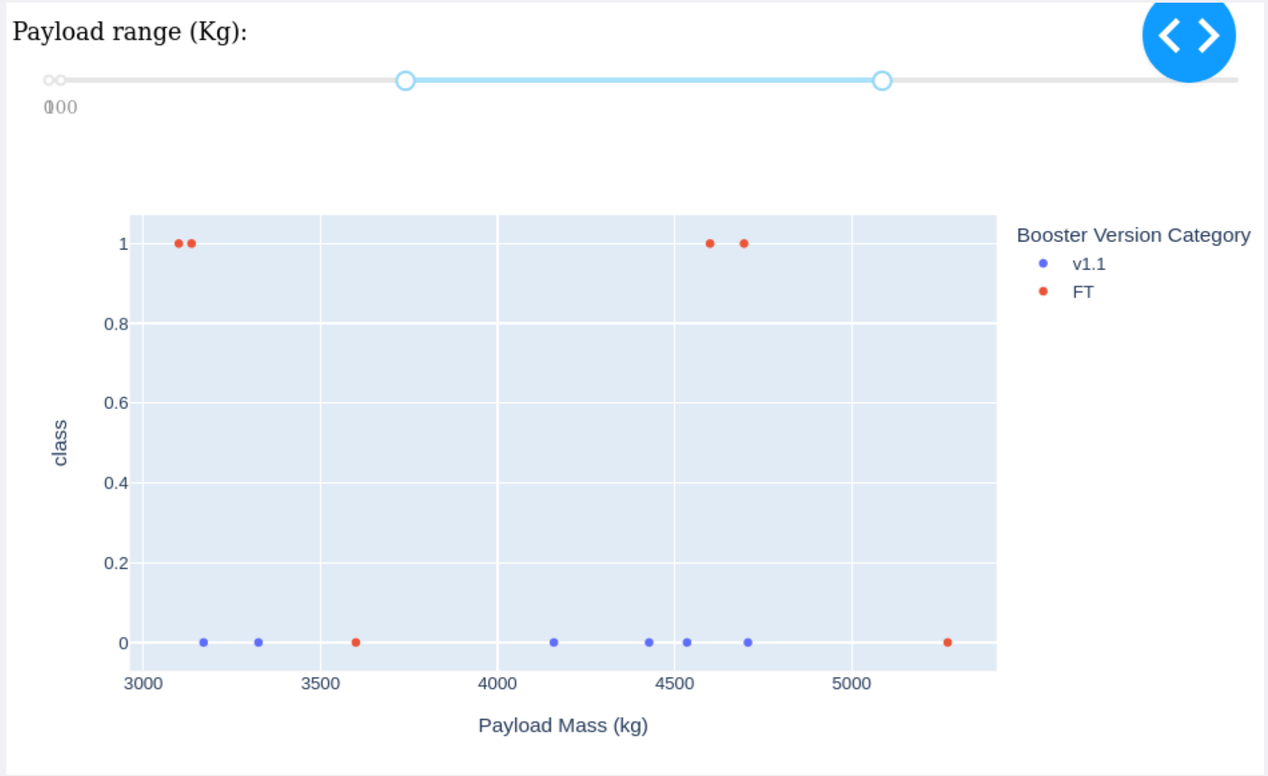
Results (Interactive analytics demo)





Results (Interactive analytics demo)



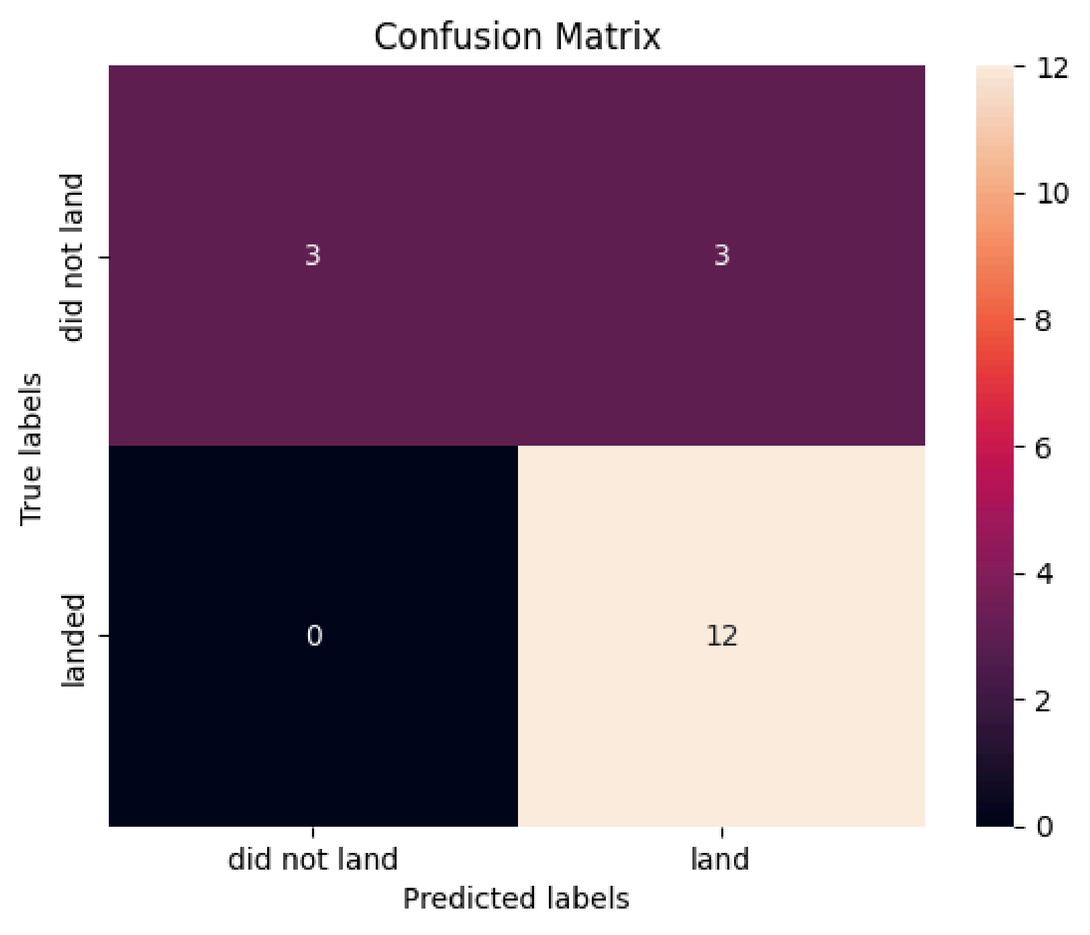


Results (Predictive Analysis)

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.808219	0.819444
F1_Score	0.909091	0.916031	0.893939	0.900763
Accuracy	0.866667	0.877778	0.844444	0.855556

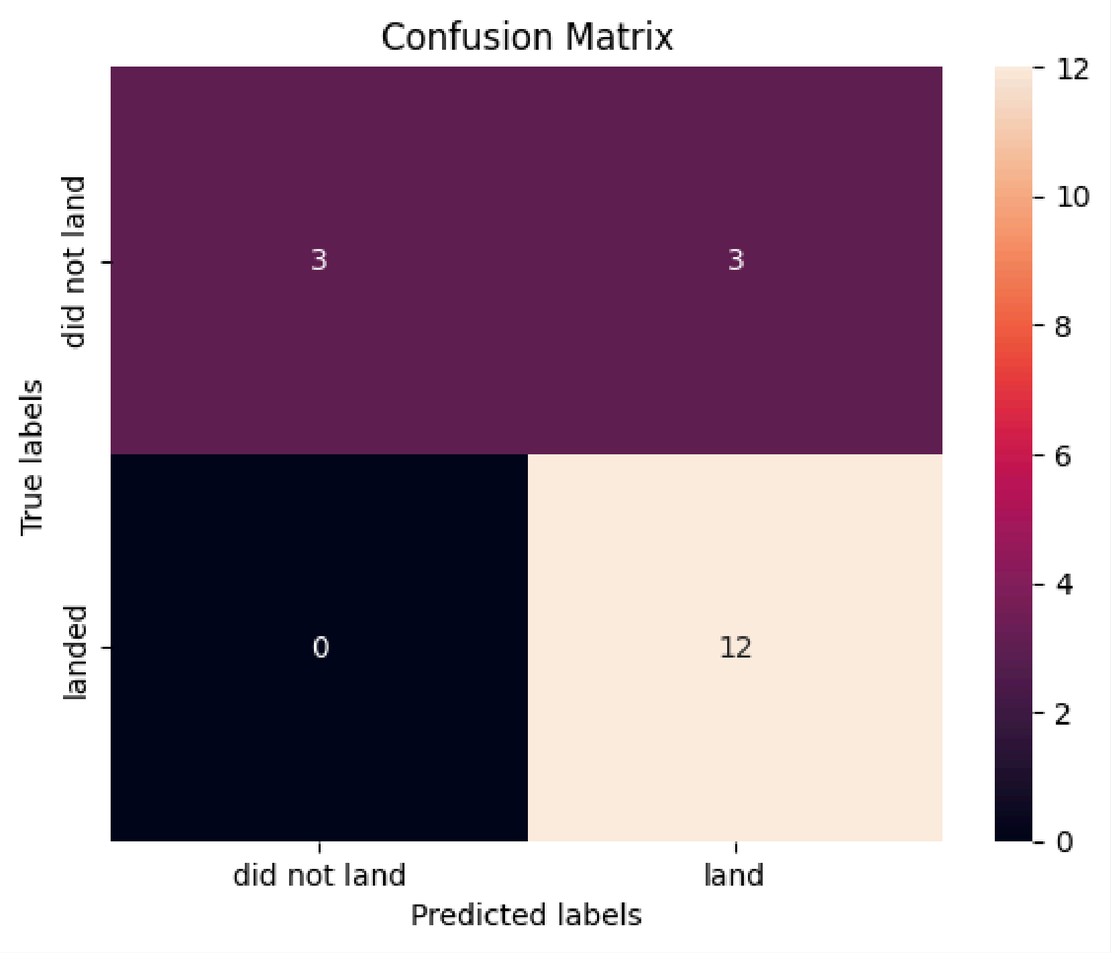
Results (Predictive Analysis)

KNN:



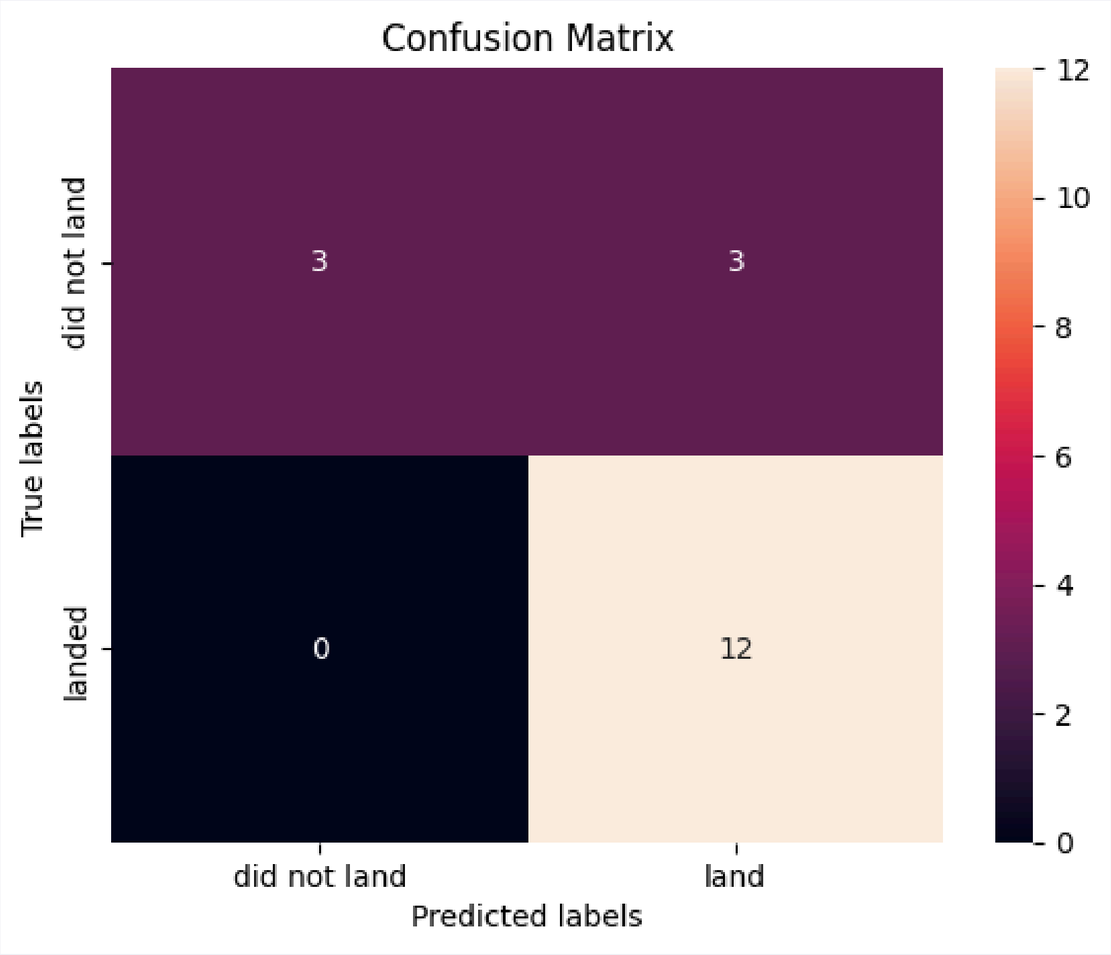
Results (Predictive Analysis)

GRID SEARCH CV:



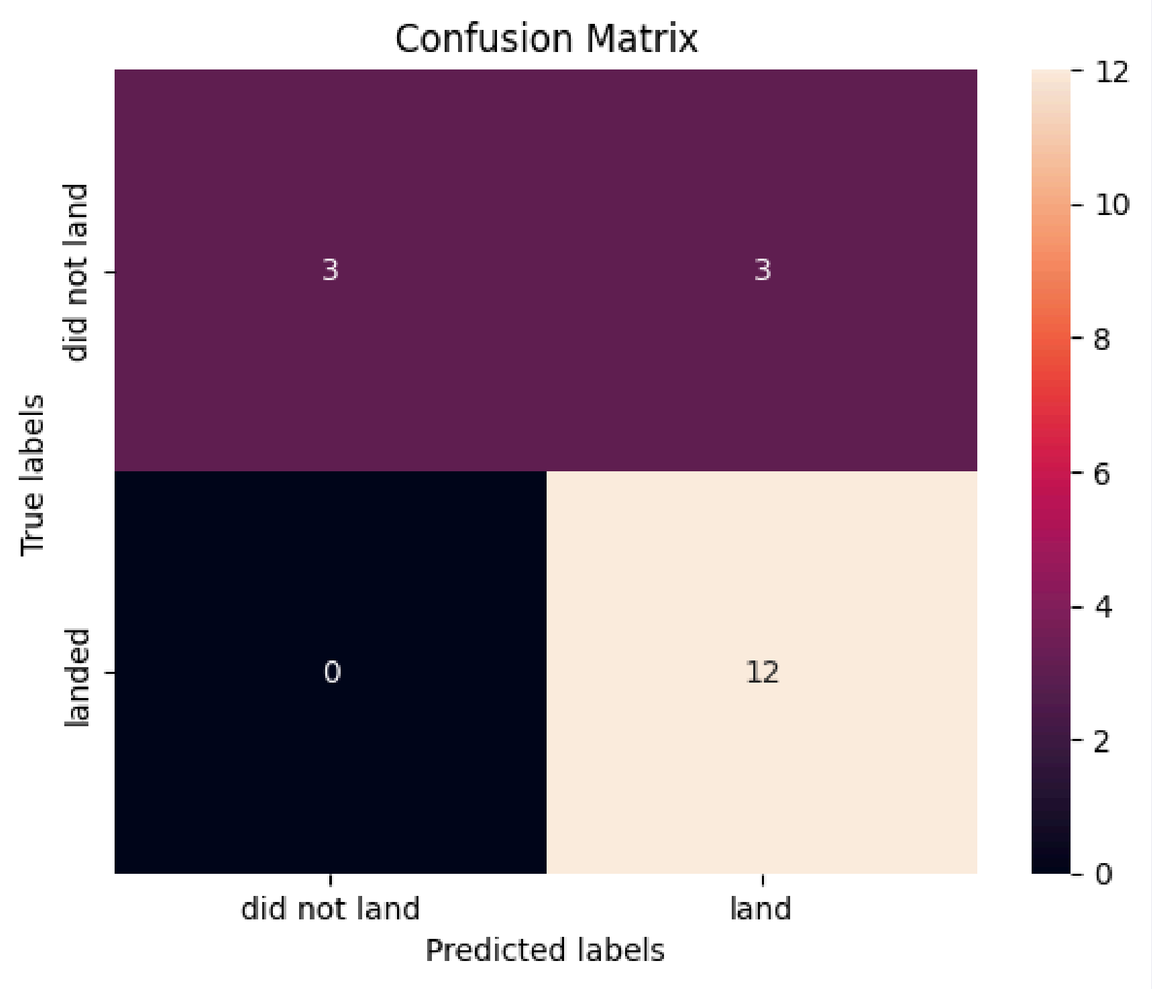
Results (Predictive Analysis)

SVM:



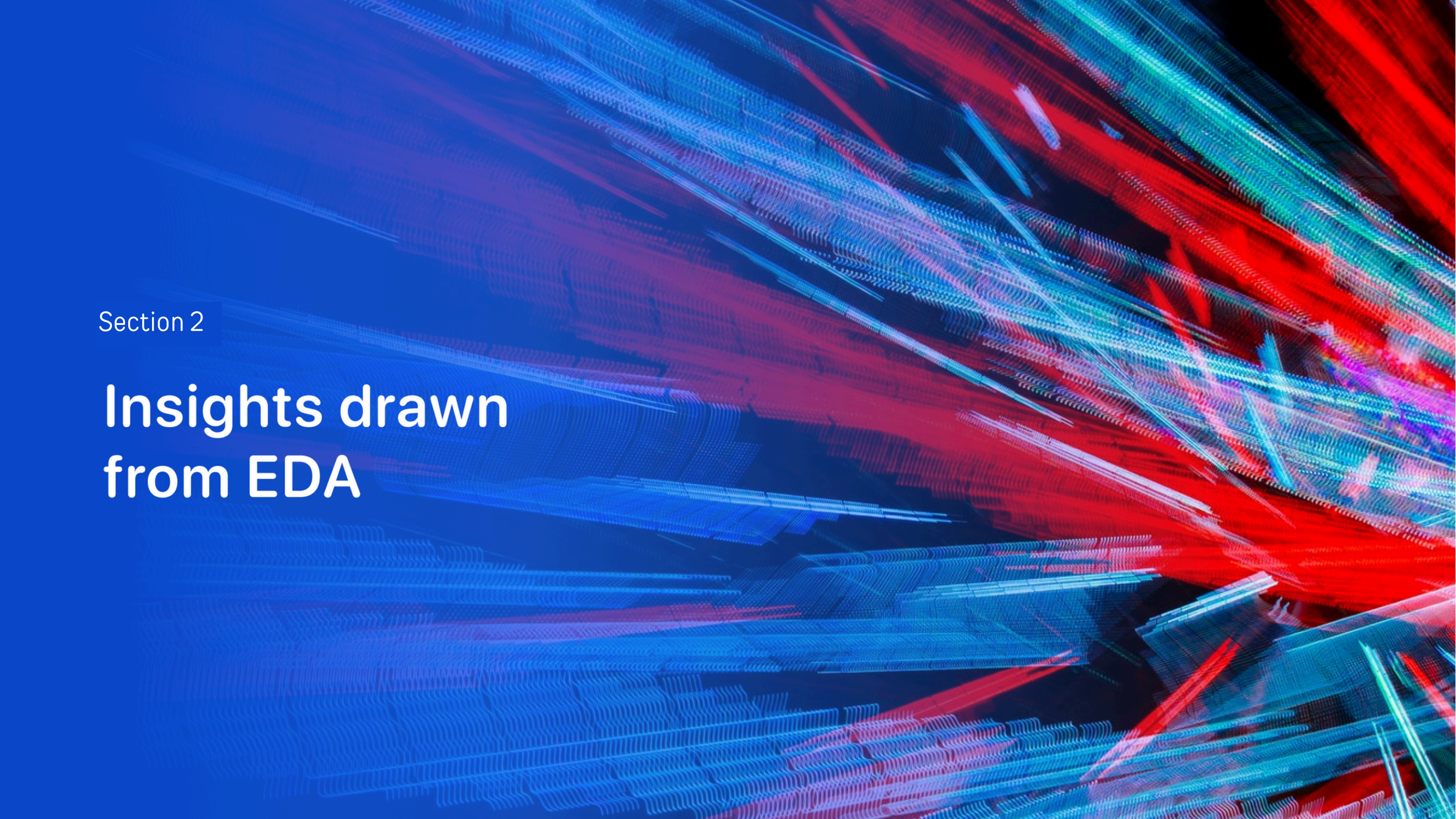
Results (Predictive Analysis)

LOGISTIC REGRESSION:

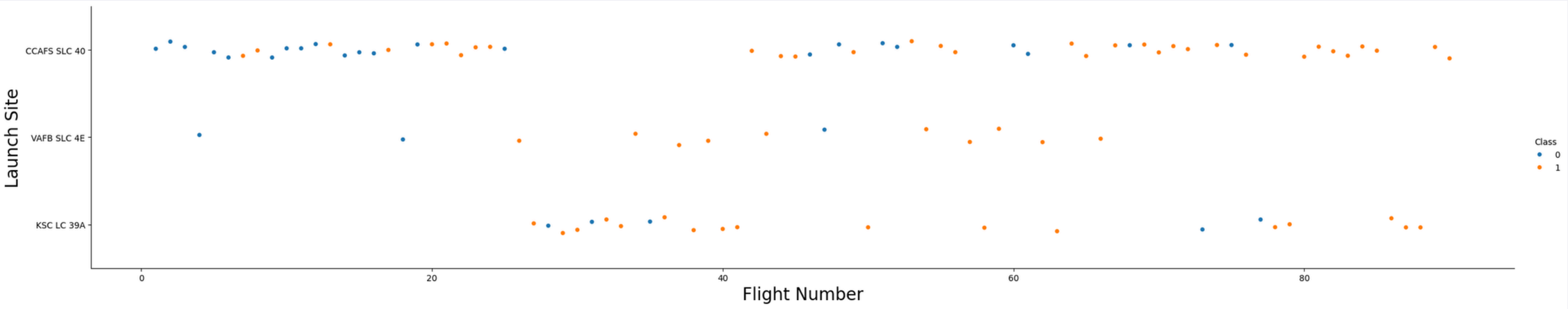


Section 2

Insights drawn from EDA

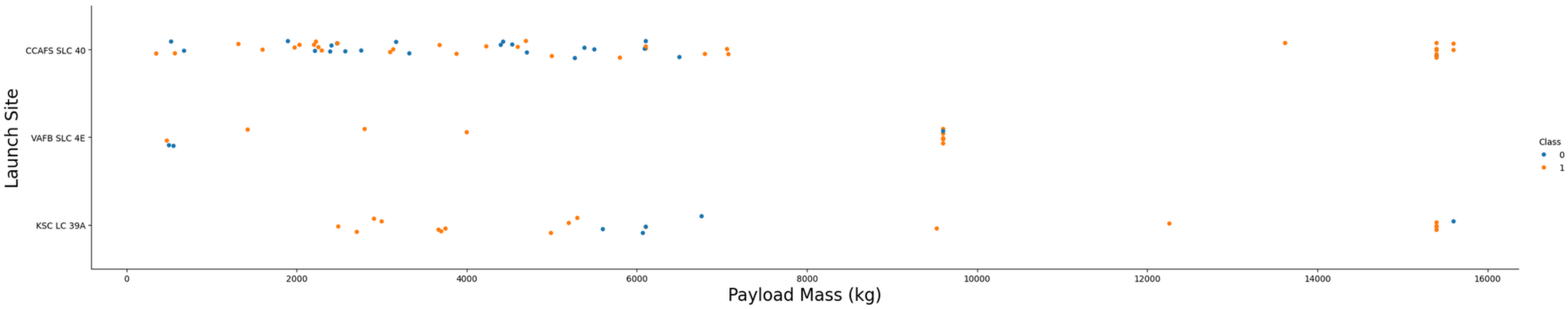


Flight Number vs. Launch Site



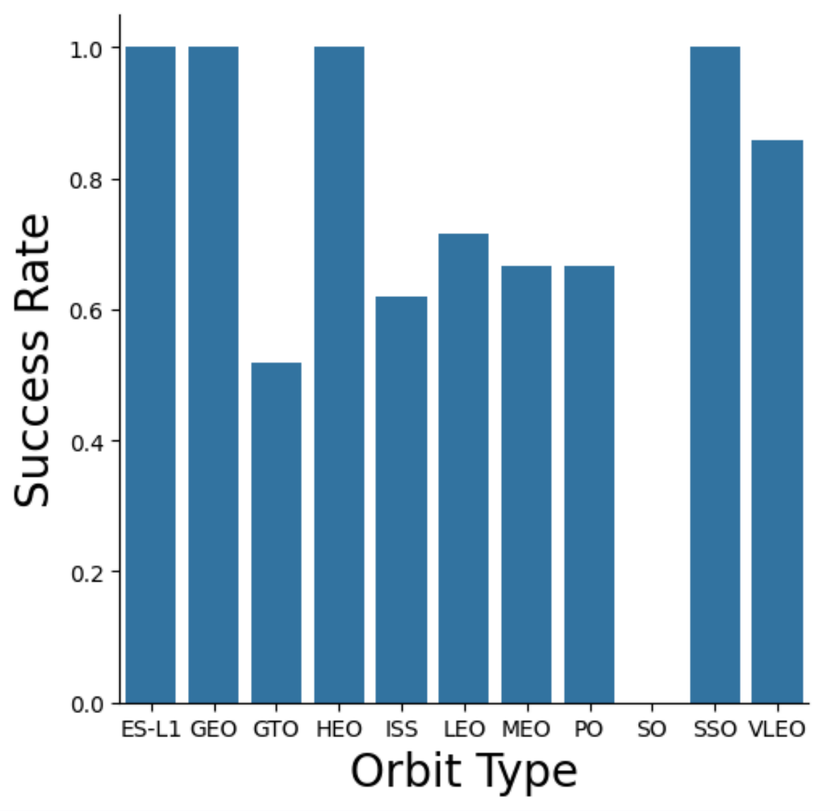
According to the plot, CCAF5 SLC 40 stands out as the best-performing launch site, with the majority of recent launches achieving success. VAFB SLC 4E ranks second, followed by KSC LC 39A in third place. Additionally, the data reveals a noticeable improvement in overall success rates over time.

Payload vs. Launch Site



- Payloads exceeding 9,000 kg have demonstrated a high success rate.
- Payloads over 12,000 kg are currently feasible only at the CCAFS SLC 40 and KSC LC 39A launch sites.

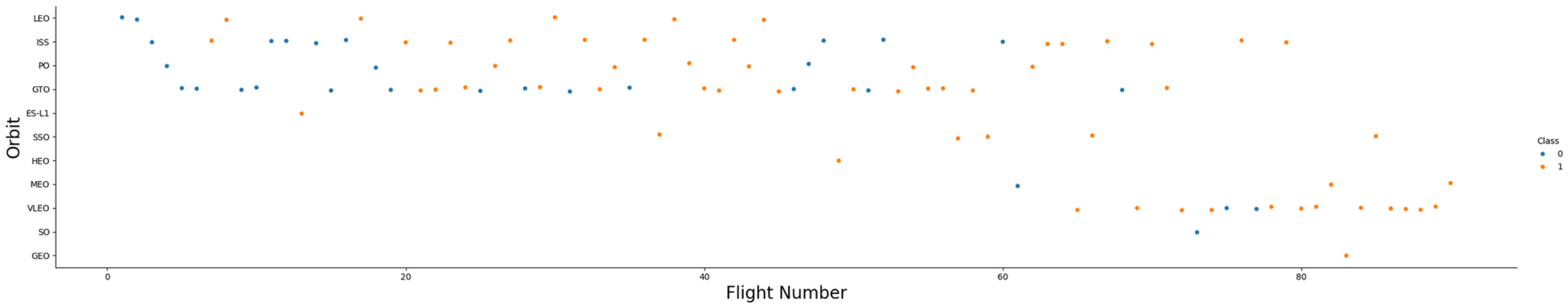
Success Rate vs. Orbit Type



- The biggest success rates happens to orbits:
- ES-L1
- GEO
- HEO
- SSO

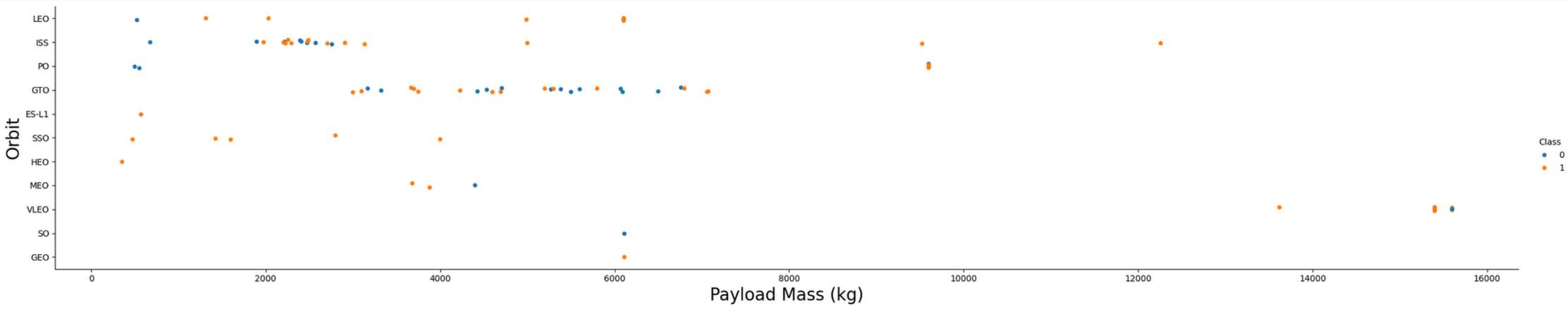
- VLEO (above 80%)
- LFO (above 70%)

Flight Number vs. Orbit Type



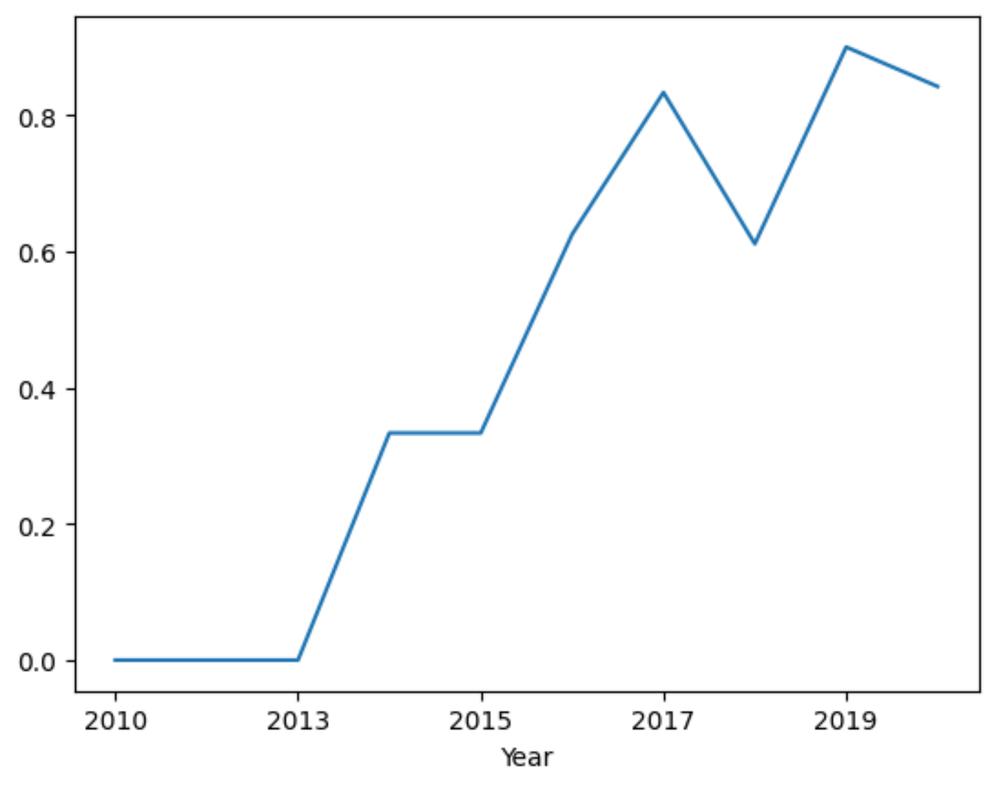
success rates improved over time for all orbits; VLEO orbit seems a new business opportunity, due to recent increase of its frequency.

Payload vs. Orbit Type



There appears to be no significant relationship between payload size and success rate for launches to GTO orbit. In contrast, the ISS orbit accommodates the widest range of payloads and maintains a strong success rate. Additionally, there are relatively few launches directed to the SO and GEO orbits.

Launch Success Yearly Trend



- Success rate started increasing in 2013 and increased until 2020

- It seems that the first three years were a period of adjusts and improvement of technology

.

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

They are gathered by selecting unique instances of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

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

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

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXTABLE where customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

• Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as "average payload mass" from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

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

average payload mass
2534.6666666666665

Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

```
%sql select min(date) as "First Landing" from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

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

First Landing
2015-12-22

This SQL query retrieves the earliest successful landing date of a Falcon 9 rocket on a ground pad from the SPACEXTABLE database. The result indicates the first successful ground pad landing occurred on 2015-12-22.

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

The results show the booster versions of Falcon 9 rockets that successfully landed on a drone ship while carrying payloads between 4,000 and 6,000 kg. The versions listed are:

- **F9 FT B1022**
- **F9 FT B1026**
- **F9 FT B1021.2**
- **F9 FT B1031.2**

Total Number of Successful and Failure Mission Outcomes

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

- **98 missions were successful.**
- **1 mission succeeded, but the payload status was unclear.**
- **1 mission failed during flight.**

This provides a quick summary of mission outcomes and their frequencies in the dataset.

Boosters Carried Maximum Payload

```
%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

The query retrieves the booster versions associated with the maximum payload mass from the dataset. The results show multiple booster versions, such as F9 B5 B1048.4 and F9 B5 B1056.4, which were capable of carrying the heaviest payloads recorded in the data.

2015 Launch Records

month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

The results show two failed landings on a drone ship at the CCAFS LC-40 launch site. These failures occurred in January and April and involved the booster versions F9 v1.1 B1012 and F9 v1.1 B1015.

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

```
%%sql select landing_outcome, count(*) as count_outcomes from SPACEXTABLE
      where date between '2010-06-04' and '2017-03-20'
      group by landing_outcome
      order by count_outcomes desc;
```

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

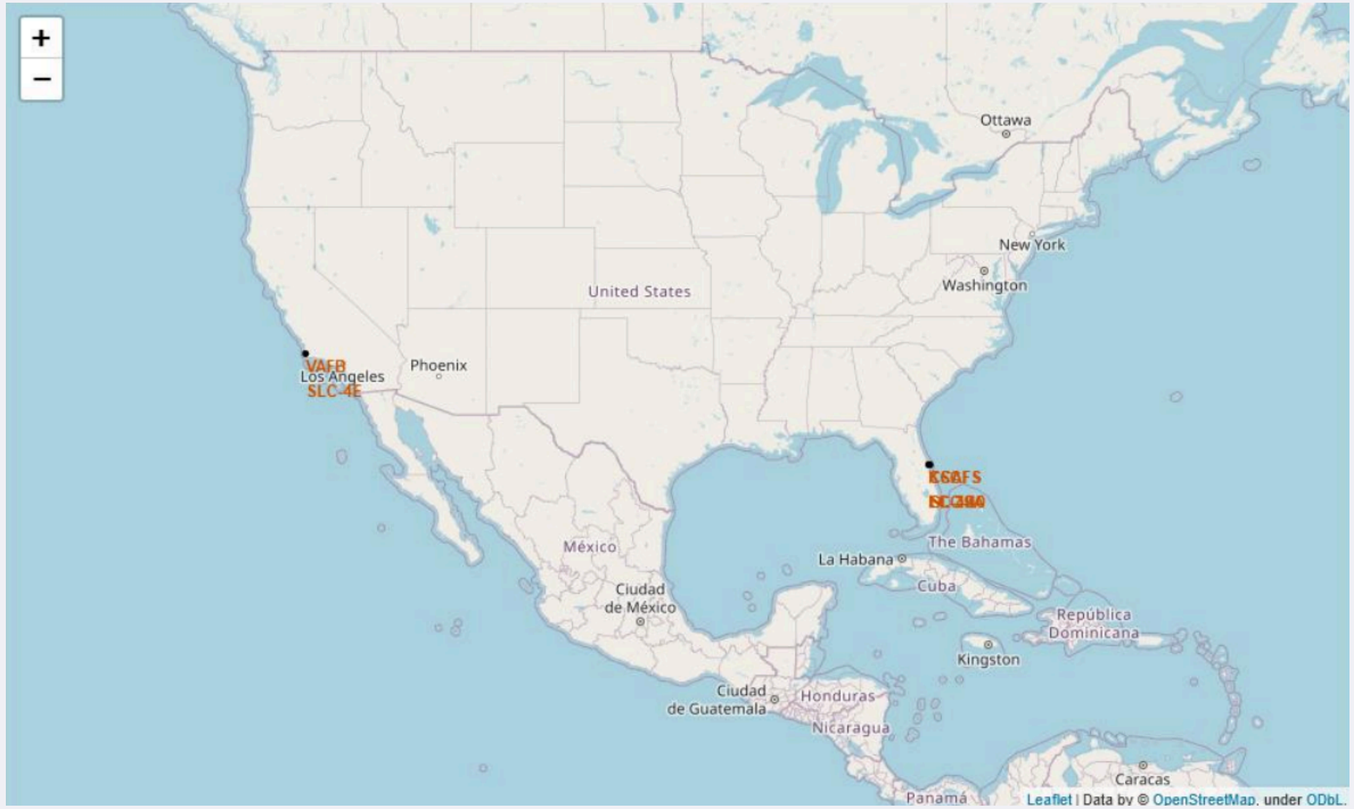
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- **This view of data alerts us that “No attempt” must be taken in account.**

Section 3

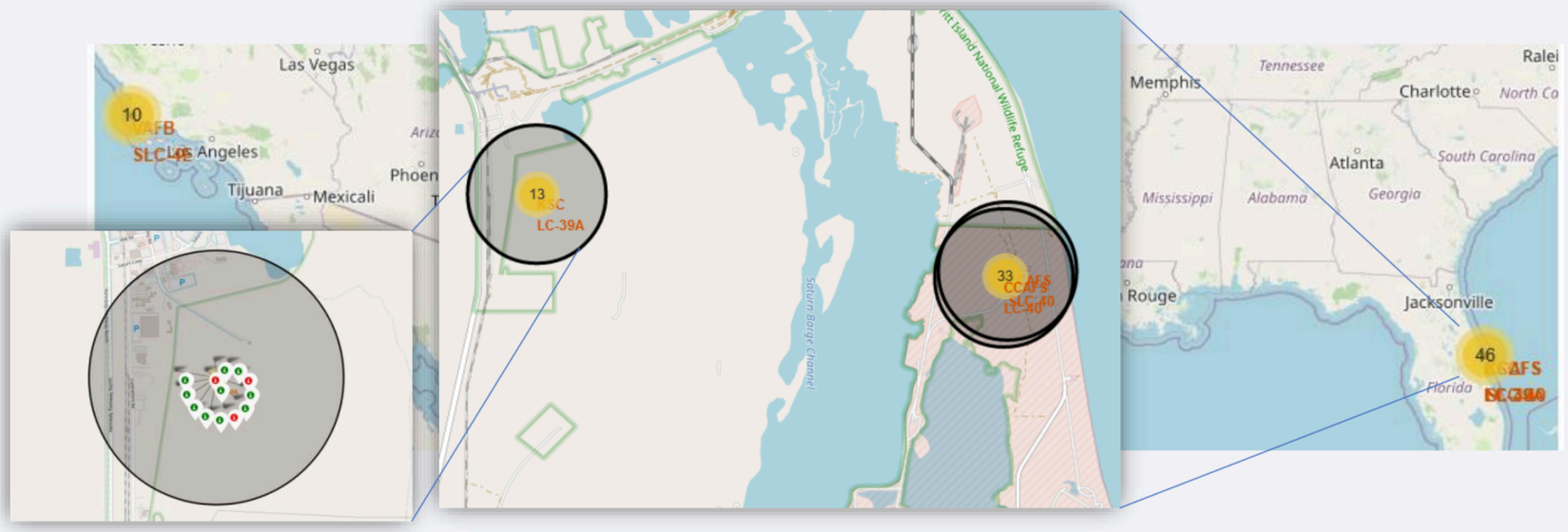
Launch Sites Proximities Analysis

All Launch sites in the US



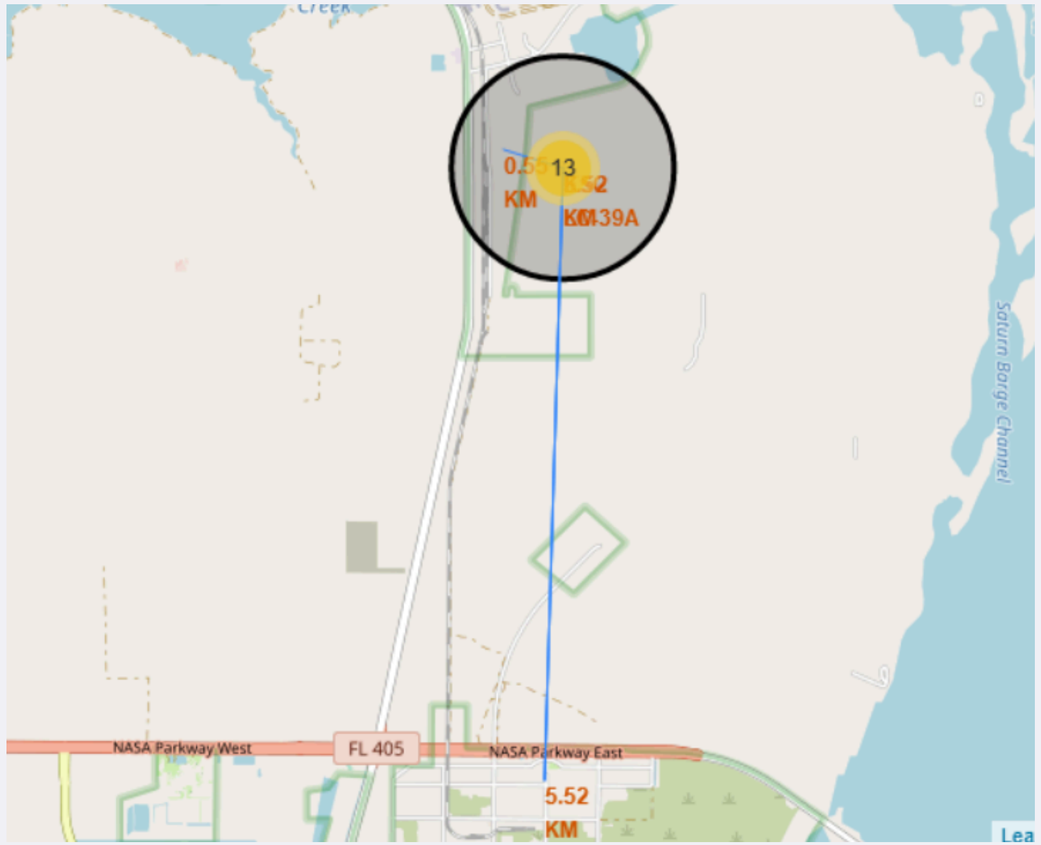
Launch sites are strategically located near the sea, likely for safety reasons, while also remaining conveniently close to roads and railroads for logistical support.

Launch Outcomes by sites

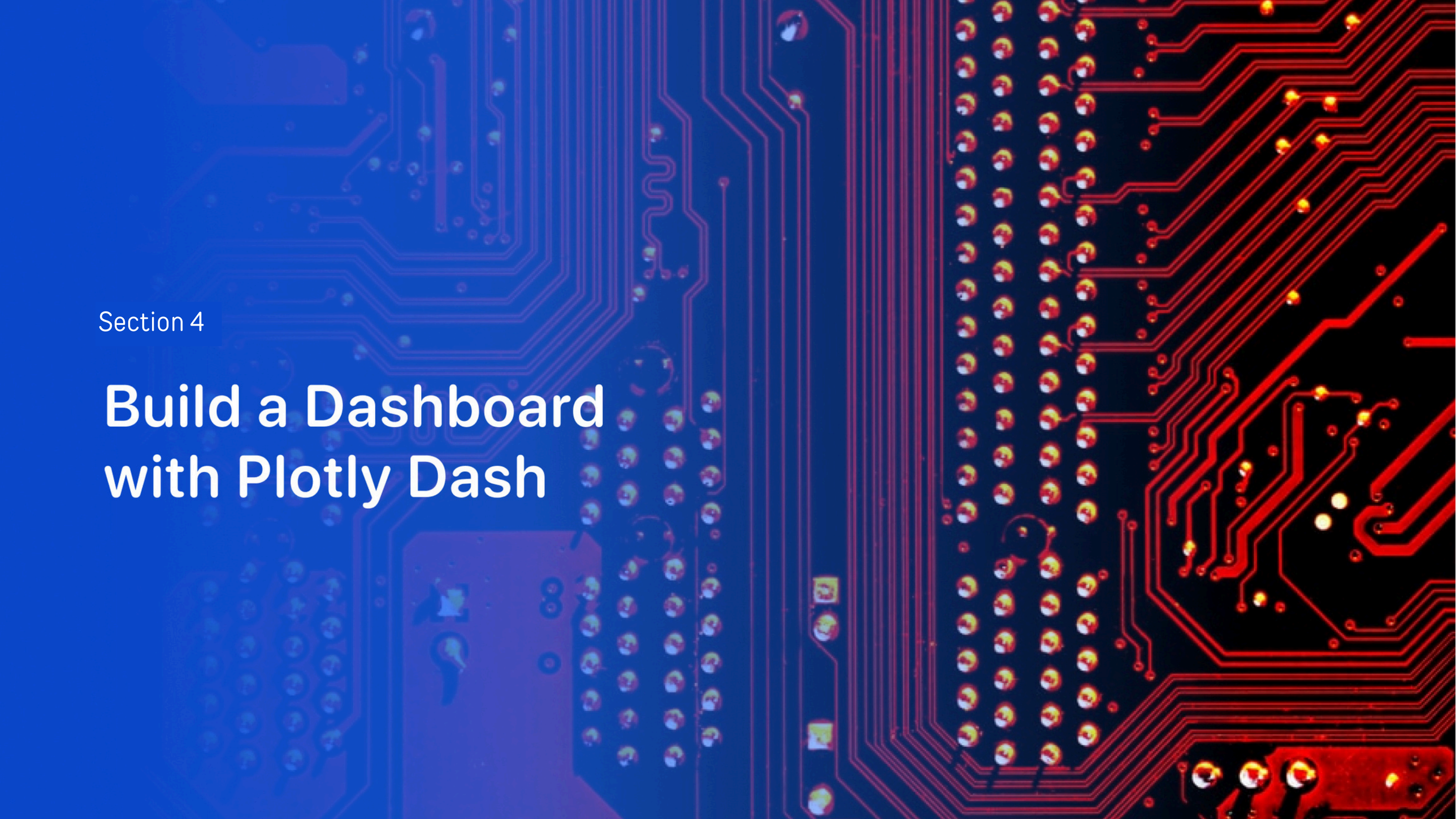


- Example of KSC LC-39A launch site launch outcomes
- Green markers indicate successful and red ones indicate failure.

Logistics and safety



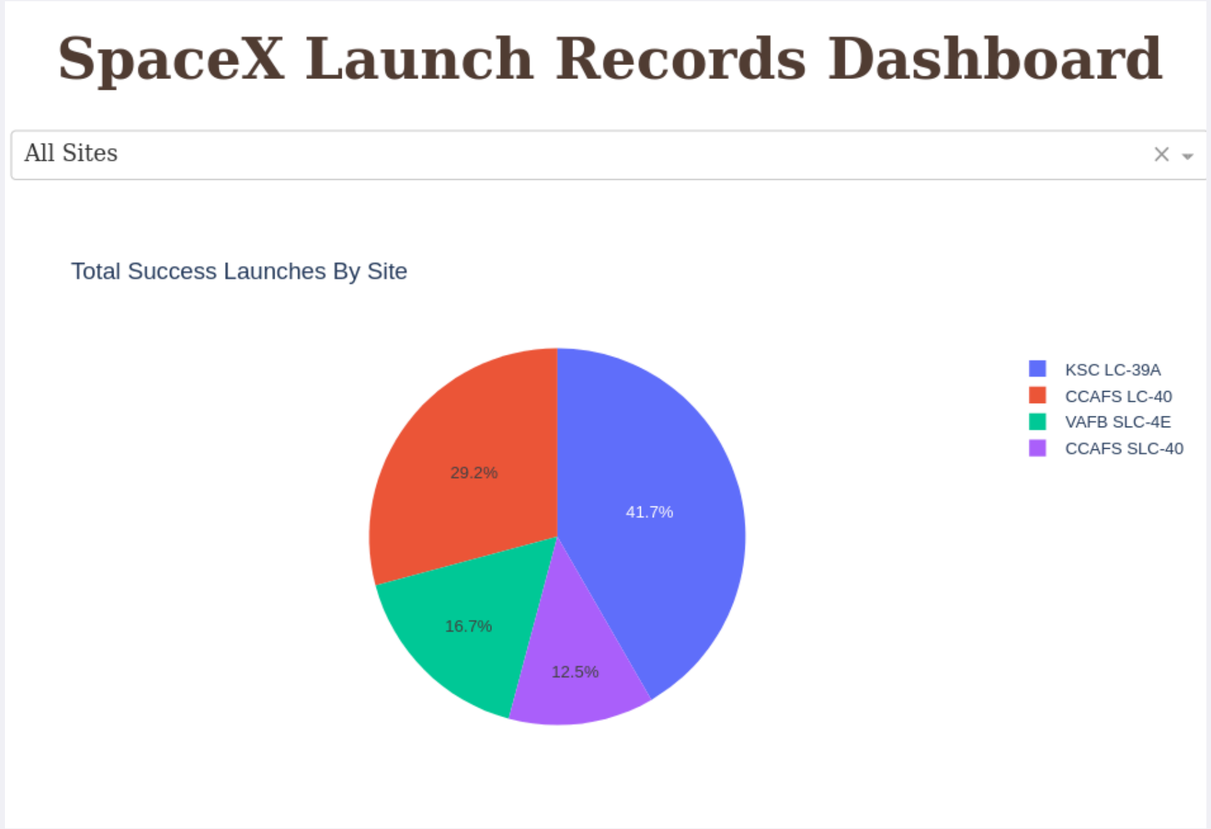
Launch site KSC LC-39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.



Section 4

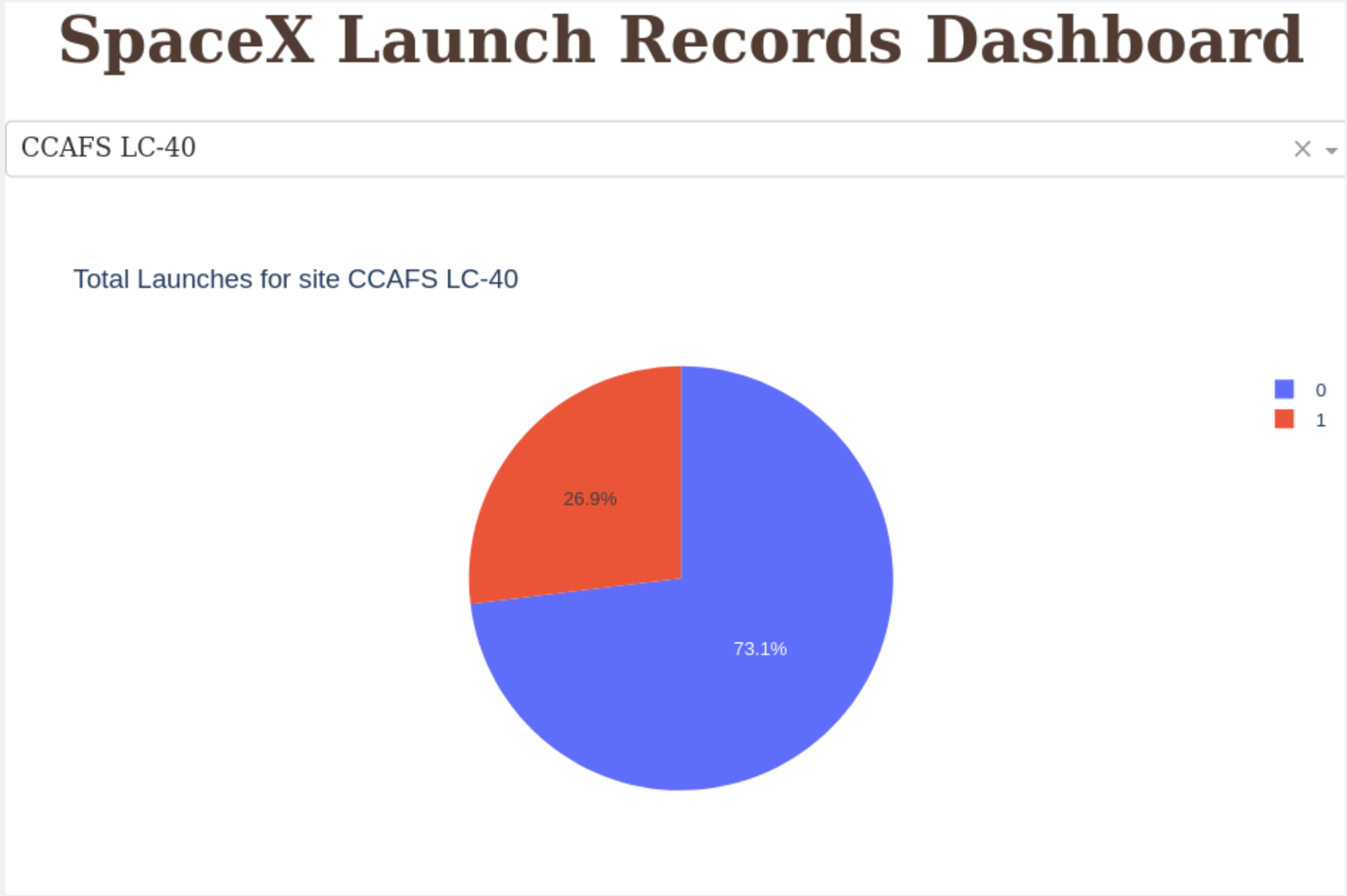
Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard

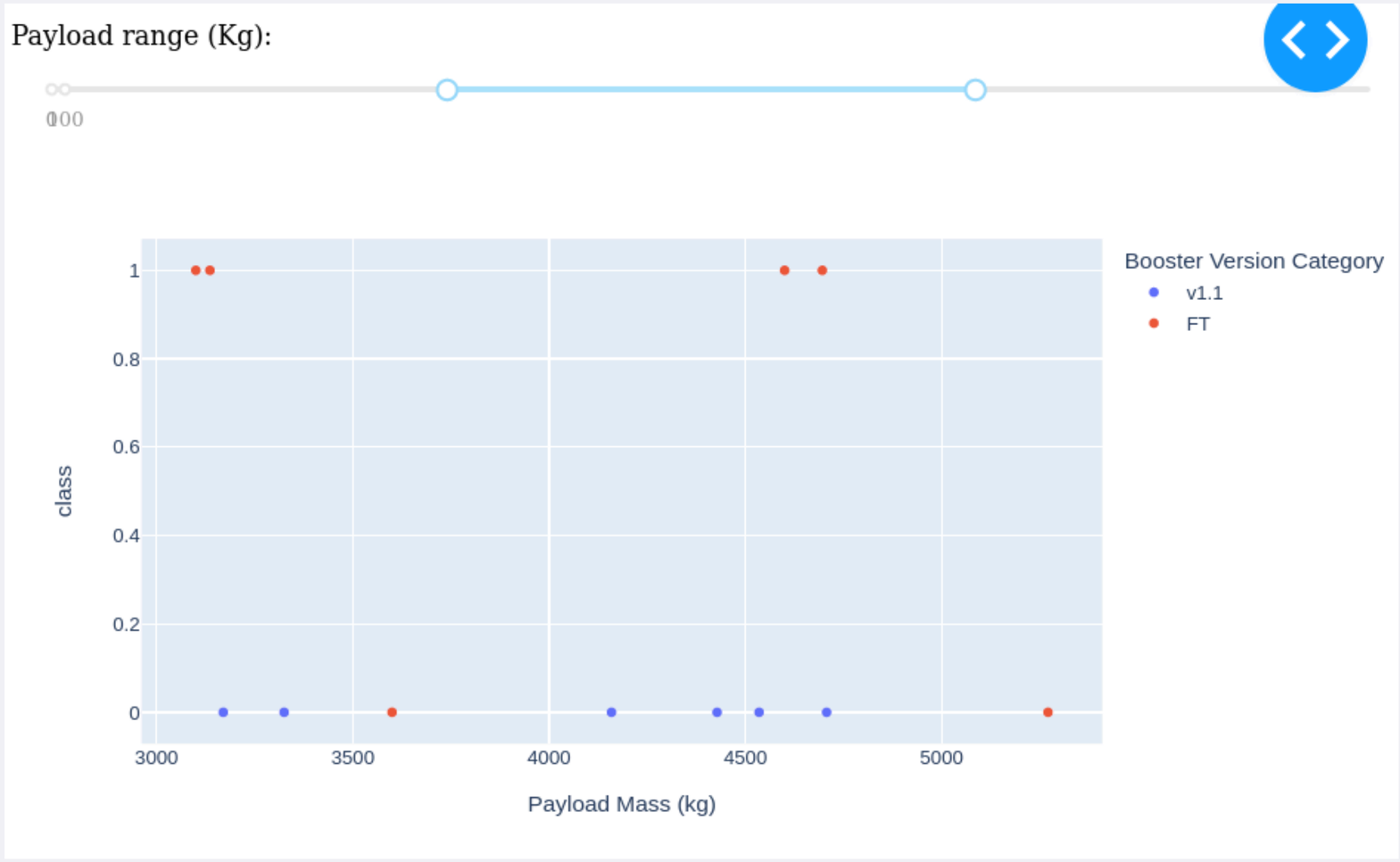


Seems like KSC LC-39A has the greatest number of of successful launches

Launch Success Ratio



Payload vs. Launch Outcome scatter plot

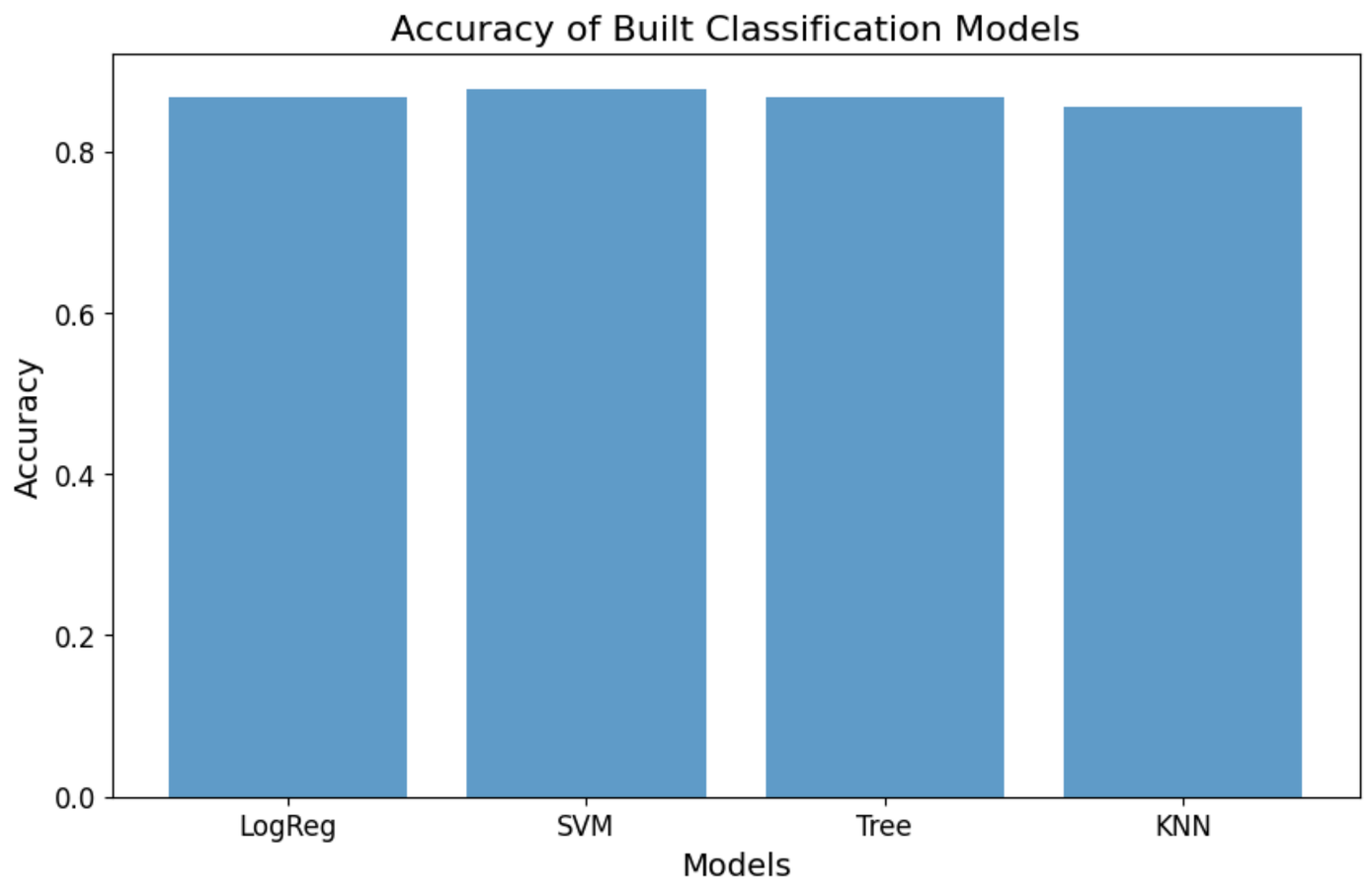


Payloads under 6,000kg and FT boosters are the most successful combination.

Section 5

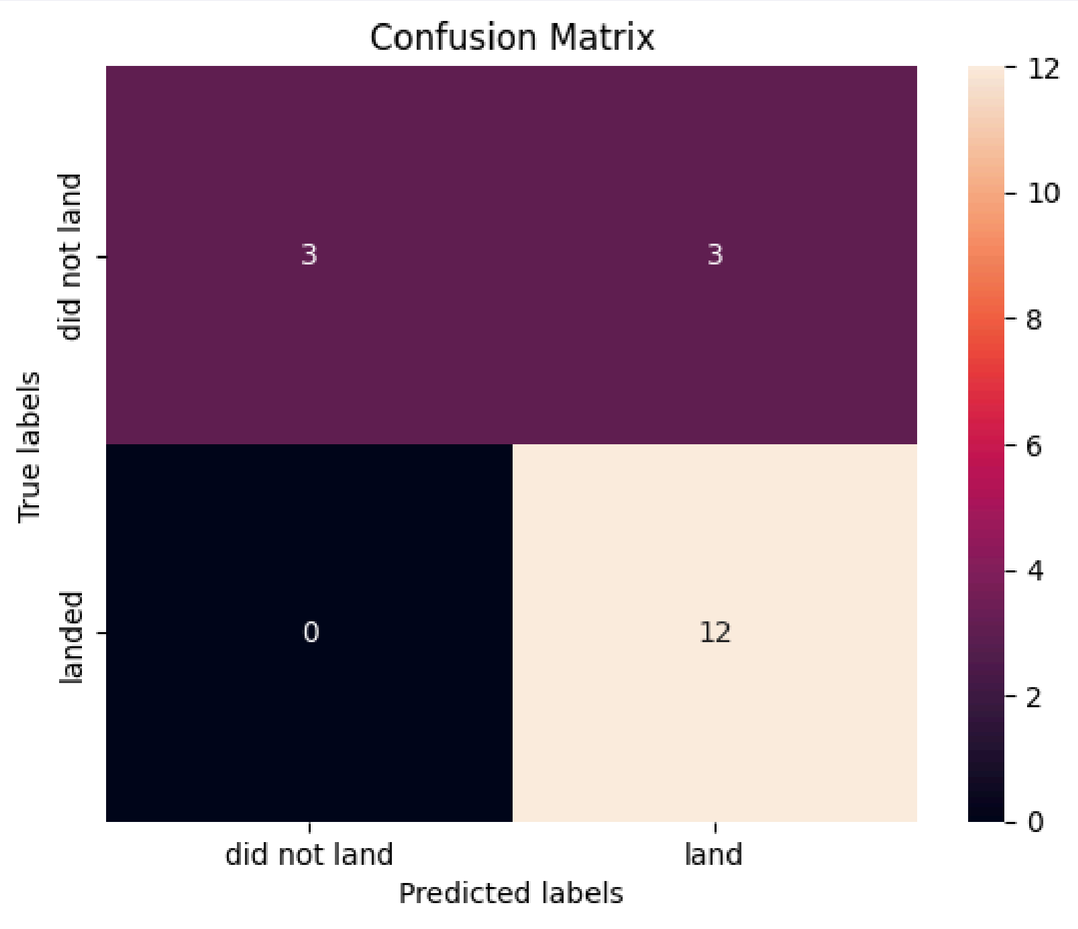
Predictive Analysis (Classification)

Classification Accuracy



The models evaluated include Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). Among these, the Support Vector Machine (SVM) model achieved the highest accuracy in predicting successful landings.

Confusion Matrix



Confusion matrix of SVM proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- **Multiple data sources were analyzed, with conclusions refined throughout the process.**
- **KSC LC-39A was identified as the best-performing launch site.**
- **Launches with payloads above 7,000 kg are generally less risky.**
- **Landing success rates have improved over time, likely due to advancements in processes and rocket technology.**
- **SVM models can be effectively used to predict successful landings, boosting efficiency and profits.**

Appendix

- Folium charts are not being displayed in github so all relevant screenshots have been given.
- To ensure consistency, np.random.seed is used.

Thank you!