Winning Space Race with Data Science

Ashwaq Almalki
2/12/2024

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

Executive Summary

Data Collection: Utilized SpaceX REST API and web scraping techniques to gather relevant information.

Data Wrangling: Processed and cleaned the data to create a variable representing success or failure outcomes.

Exploratory Analysis: Leveraged data visualization to investigate various factors such as payload, launch site, flight number, and yearly trends.

Data Analysis: Used SQL to compute critical metrics including total payload, payload ranges for successful launches, and counts of successful vs. failed outcomes.

Launch Site Analysis: Examined the success rates of launch sites in relation to their geographic proximity to key markers.

Visualization: Highlighted the launch sites with the highest success rates and the payload ranges associated with successful outcomes.

Predictive Modeling: Developed models including logistic regression, support vector machines (SVM), decision trees, and k-nearest neighbor (KNN) to predict landing outcomes.

Introduction

- SpaceX is at the forefront of innovation in the space industry, making space exploration more affordable and accessible. One of its most remarkable advancements is the reuse of the first stage of its Falcon 9 rockets, drastically reducing launch costs to $62 million compared to over $165 million for traditional rockets. This project focuses on understanding the factors that influence the success of rocket landings and predicting their outcomes using data and machine learning models. By leveraging publicly available data, we aim to analyze launch trends, visualize success rates, and develop predictive tools to assess the likelihood of first-stage reusability, offering insights into the future of cost-effective space travel.

Section 1

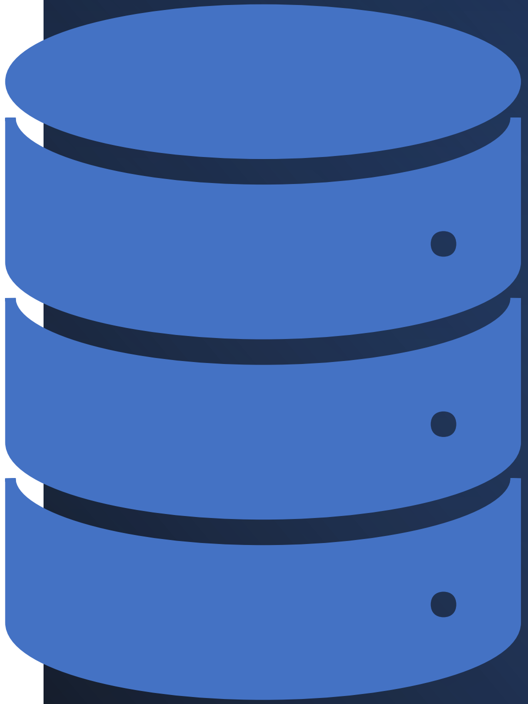
Methodology

Methodology

- Data collection
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

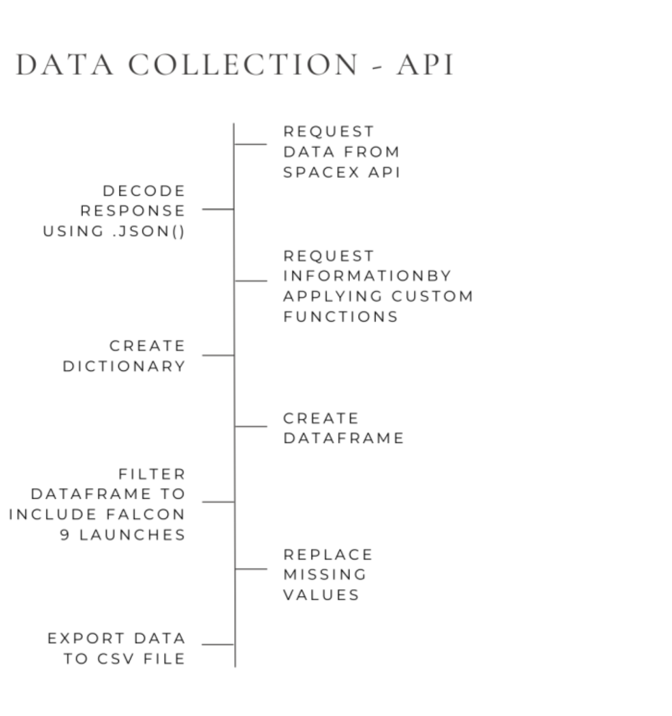
- The data collection process combined API requests from the SpaceX REST API with web scraping from a table on SpaceX's Wikipedia page. Utilizing both methods was essential to gather comprehensive information about the launches, enabling a more thorough and detailed analysis.



Data Collection – SpaceX API

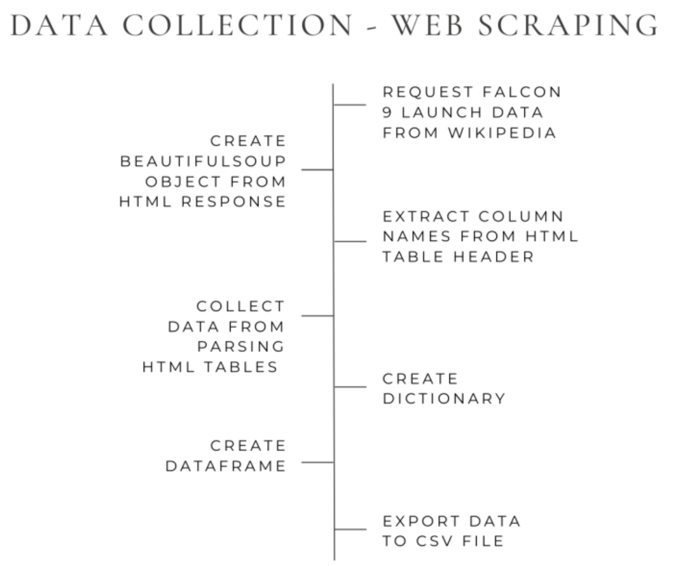
- Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.

- [Github URL : Data Collection - API]



Data Collection - Scraping

- Data Columns:Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time.
- [Github URL : Data Collection – Web Scraping]



Data Wrangling

Key steps in preparing the data included:

EDA and Metrics:

- Counted launches by site, orbit frequency, and mission outcomes per orbit.
- Created a binary landing outcome column:
  - (1 : Successful landing) True Ocean, True RTLS, True ASDS.
  - (0 : Unsuccessful landing) False Ocean, False RTLS, False ASDS.

Outcome Labels:

- Classified landing results as successful or unsuccessful based on location (ocean, ground pad, or drone ship).

Export:

- Processed data exported to a CSV for analysis and modeling.

GitHub: Data Wrangling

EDA with Data Visualization

- **Scatter plots :**
- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type
- **Bar charts :**
- Relationships among the categories
- **Line charts :**
- Trends in data over time
- GitHub URL : EDA with Visualization

EDA with SQL

- **Key SQL queries performed during the analysis:**
- **Display**:
  - Names of unique launch sites.
  - 5 records where launch sites start with 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass for booster version F9 v1.1.
- **List**:
  - Date of the first successful landing on a ground pad.
  - Names of boosters with successful drone ship landings and payload mass between 4,000 and 6,000 kg.
  - Total count of successful and failed missions.
  - Names of booster versions that carried the maximum payload mass.
  - Failed drone ship landings in 2015, including booster version and launch site.
  - Ranking of landing outcomes (e.g., failure on drone ship, success on ground pad) between 2010-06-04 and 2017-03-20 in descending order.
- [GitHub URL : EDA with SQL]()

Build an Interactive Map with Folium

- **Markers Indicating Launch Sites**:
- Placed a blue marker at NASA Johnson Space Center with a popup label using its latitude and longitude.
- Added red markers for all launch sites, showing their geographical coordinates and names.
- **Colored Markers for Launch Outcomes**:
- Used green markers for successful launches and red markers for unsuccessful ones.
- Implemented marker clustering to visualize success rates at each site.
- **Distances to Proximities**:
- Added colored lines to display distances from launch sites (e.g., KSC LC-39A or CCAFS SLC40) to nearby features such as coastlines, railways, highways, and cities.

- GitHub URL : Interactive Map with Folium

Build a Dashboard with Plotly Dash

- An interactive dashboard was developed using Plotly Dash, providing dynamic visualizations and user controls:
- **Launch Sites Dropdown**:
    - Added a dropdown menu to allow users to select all launch sites or a specific site for detailed analysis.
- **Payload Mass Slider**:
    - Included a slider to enable users to filter data by selecting a specific payload mass range.
- **Pie Chart of Launch Success**:
    - Visualized the proportion of successful vs. unsuccessful launches:
        - Displayed total success rates across all sites.
        - Showed success vs. failure rates for a selected launch site.
- **Scatter Plot of Payload vs. Success**:
    - Added a scatter chart to analyze the relationship between payload mass and launch success for various booster versions.
- GitHub URL : Dashboard with Plotly Dash

Predictive Analysis (Classification)

- **Steps:**
- **Create** NumPy array from the “Class” column.
- **Standardize** the data with StandardScaler and **Fit** and **transform** the data.
- **Split** the data using train_test_split.
- **Create a GridSearchCV** object with cv=10 for parameter optimization .
- **Apply GridSearchCV** on different algorithms: logistic regression, support vector machine, decision tree and, K-Nearest Neighbor.
- **Calculate accuracy** on the test data using .score() for all models .
- Examining the **confusion matrix** for all models .
- Identify the best model using **Jaccard_Score, F1_Score and Accuracy.**
- GitHub URL : Predictive Analysis

Results

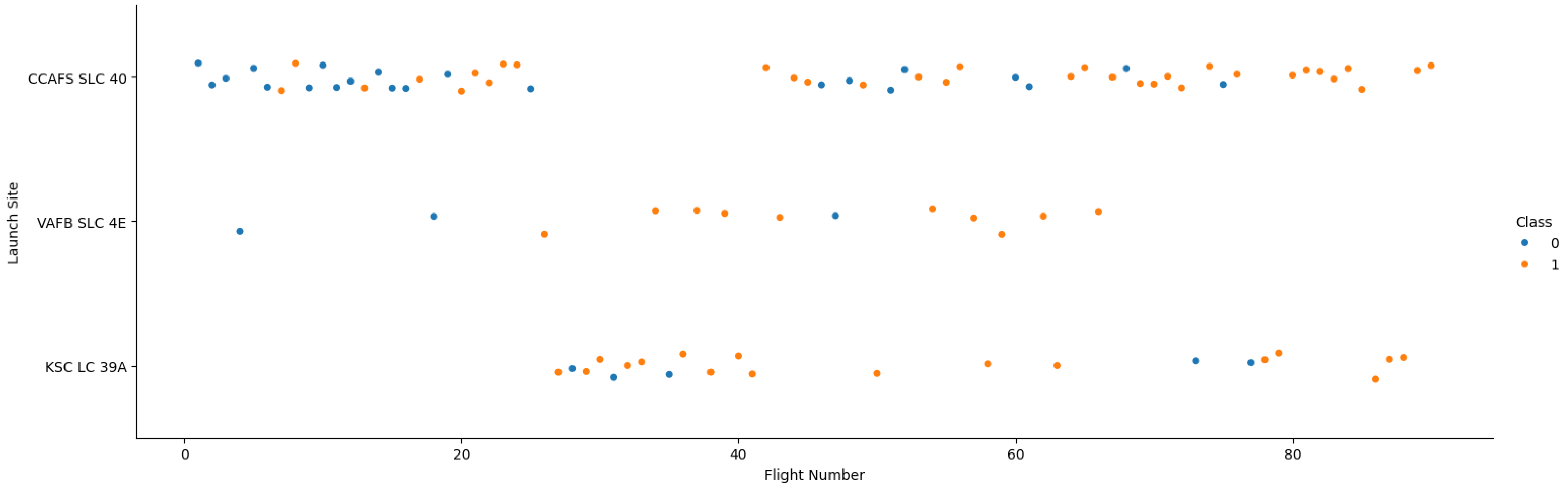
EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

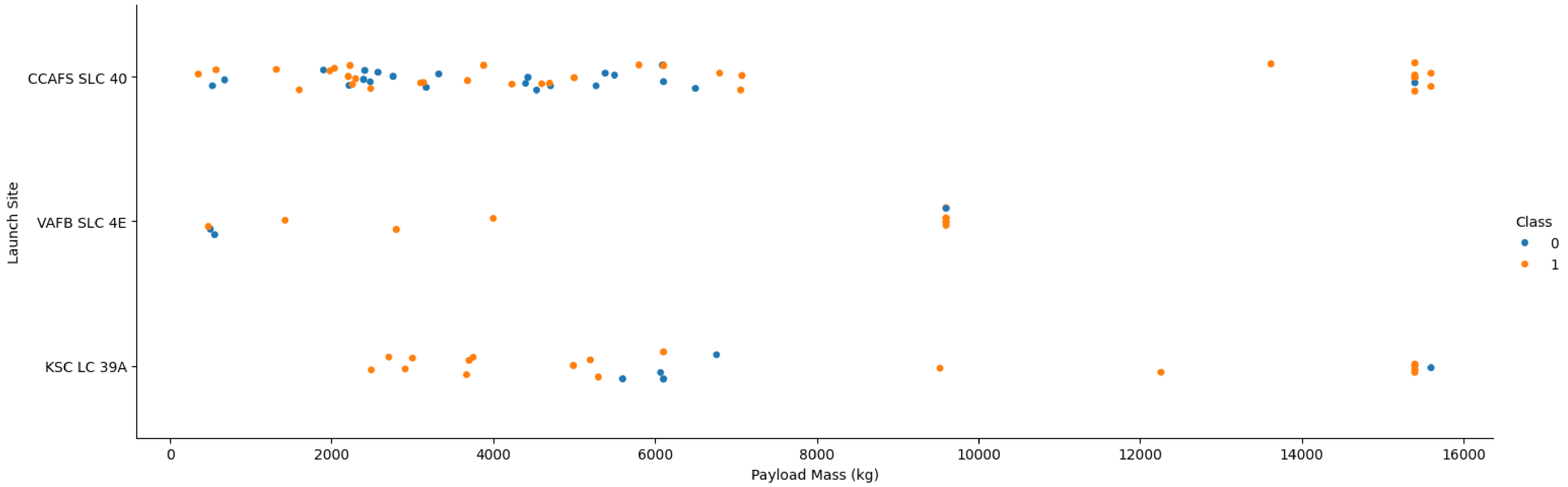
PREDICTIVE ANALYSIS RESULTS

Section 2

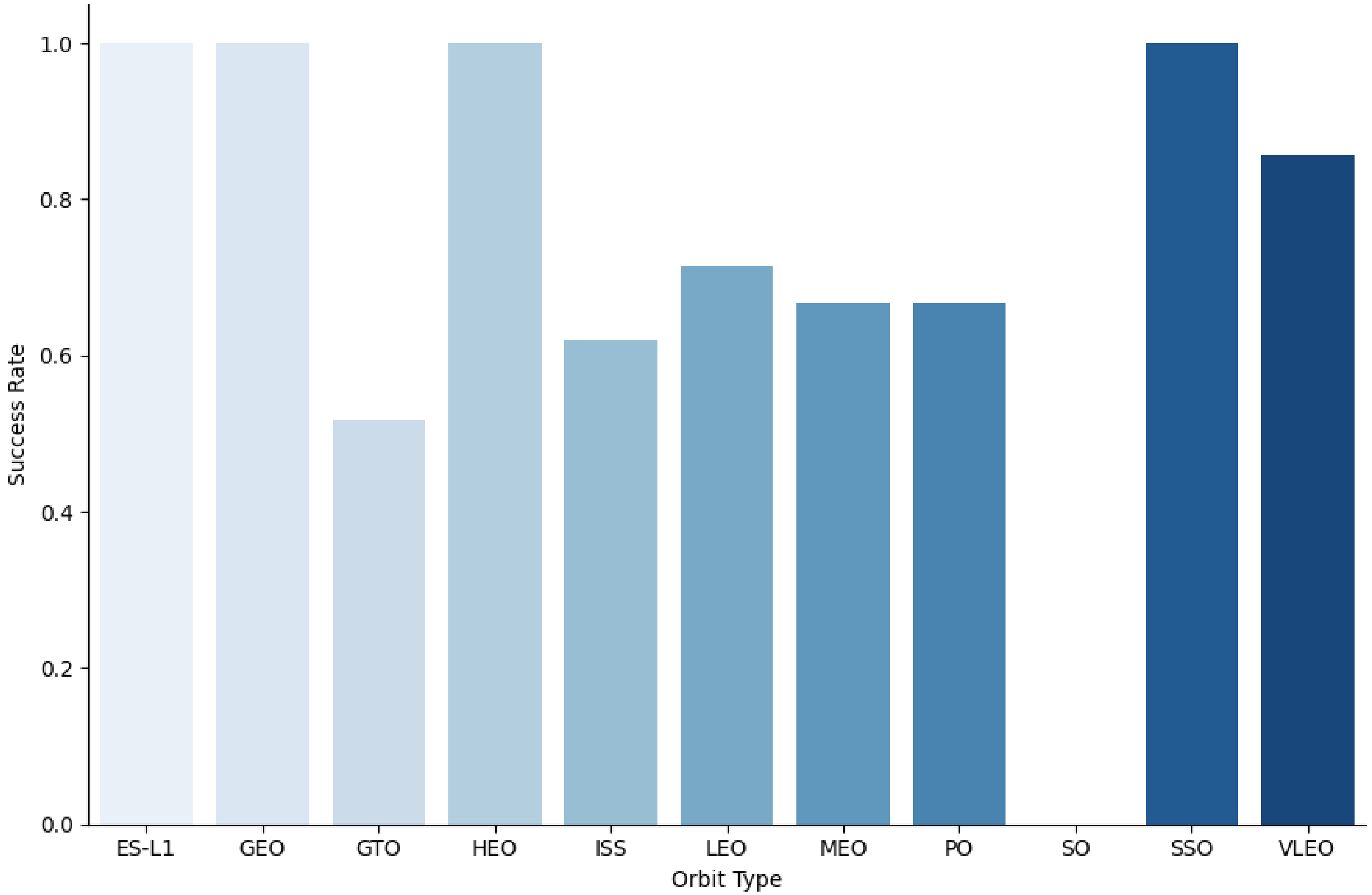
Insights drawn from EDA



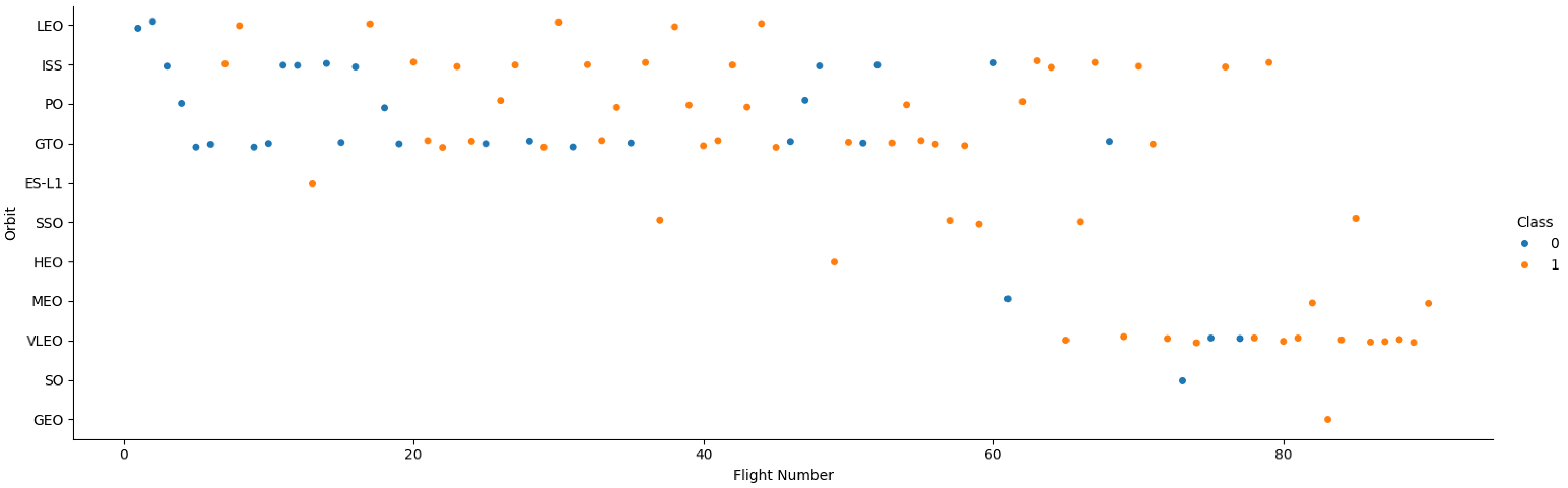
Flight Number vs. Launch Site



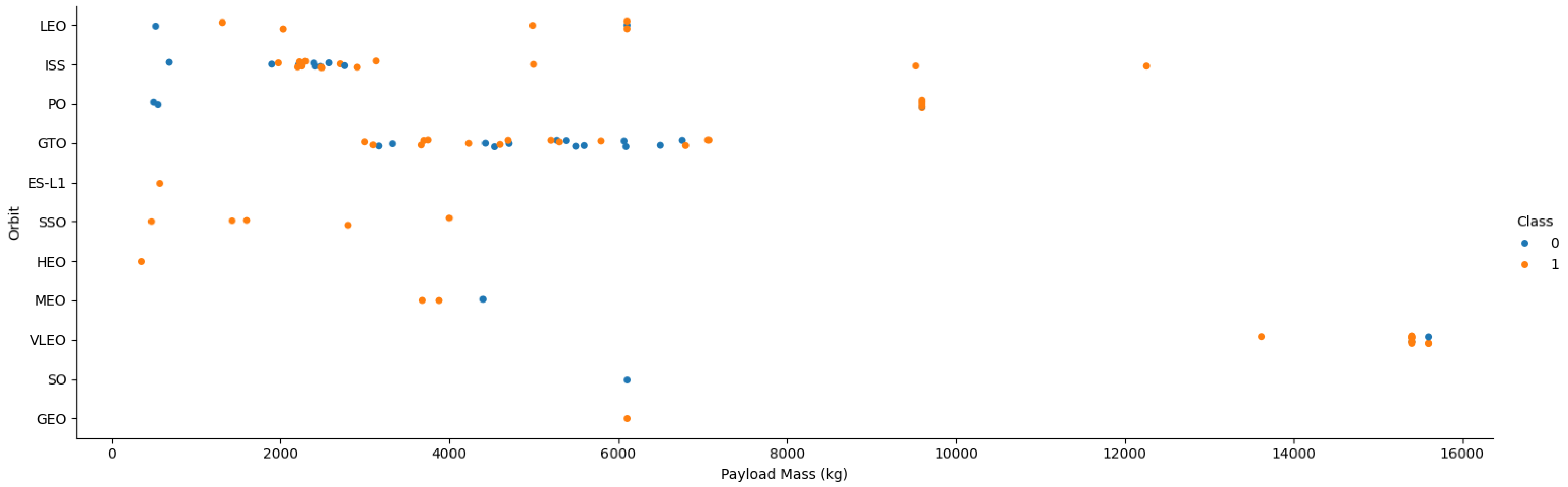
Payload vs. Launch Site



Success Rate vs. Orbit Type

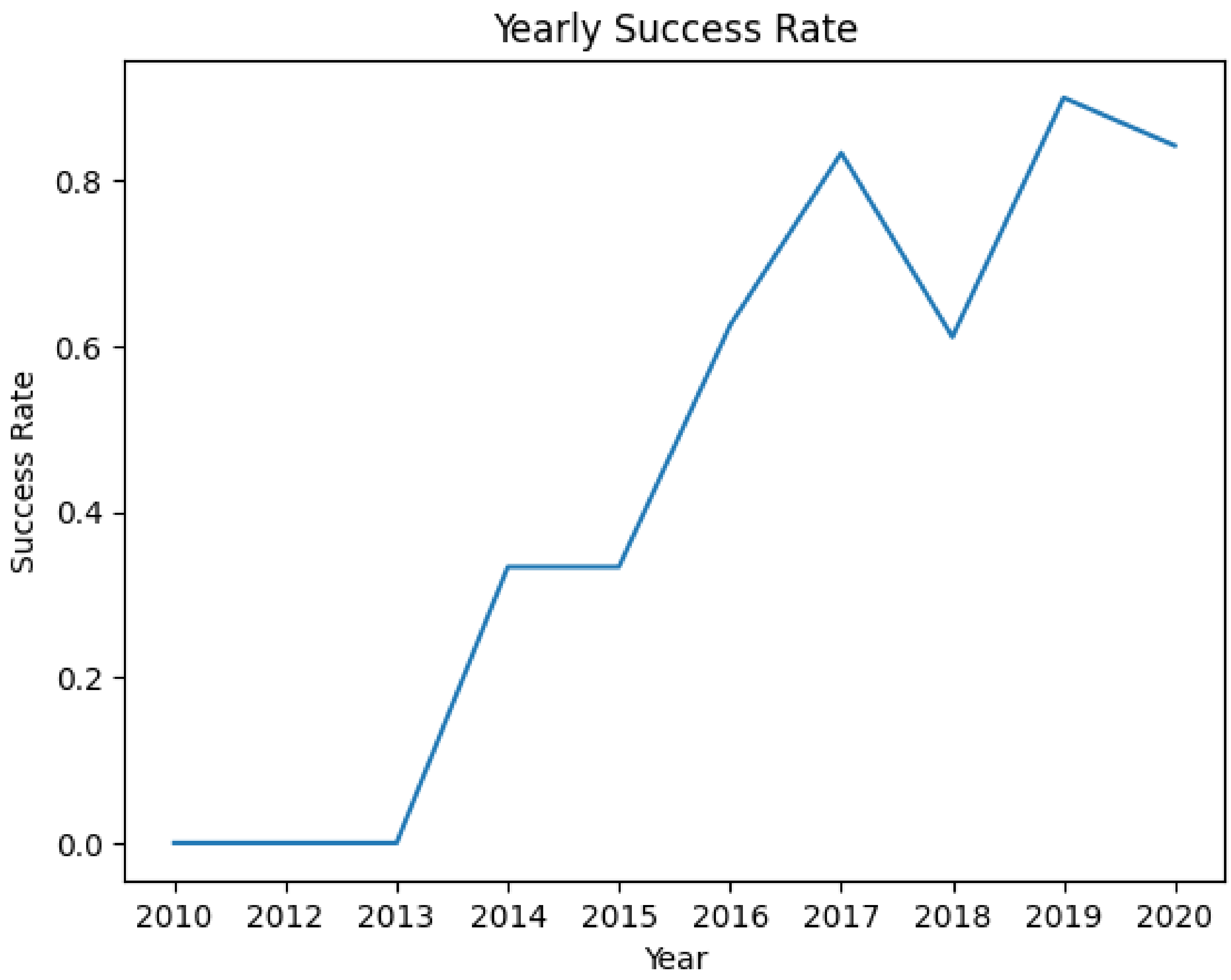


Flight Number vs. Orbit Type



Payload vs. Orbit Type

Launch Success Yearly Trend



All Launch Site Names

```
%%sql
select distinct Launch_Site from spacextbl
```

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

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

Launch Site Names Begin with 'CCA'

```
%%sql
select * from spacextbl 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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success

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

Average Payload Mass by F9 v1.1

```
%%sql
select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

```
%%sql
select min(Date) as min_date from spacextbl where Landing_Outcome = 'Success (ground pad)';
```

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

min_date
2015-12-22

```
%%sql

select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing_Outcome = 'Success (drone ship)');
```

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

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

```
%%sql
select Mission_Outcome, count(*) as counts from spacextbl group by Mission_Outcome;
```

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

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql
select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_)
```

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

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

Boosters Carried Maximum Payload

```
%%sql
select Landing_Outcome, Booster_Version, Launch_Site from spacextbl where Landing_Outcome = 'Failure (drone ship)'
```

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

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

2015 Launch Records

```
%%sql
select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

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

Landing_Outcome	LandingCounts
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

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

Section 3

Launch Sites Proximities Analysis

Launch Sites

- **Explanation**
- **Proximity to the Equator**:

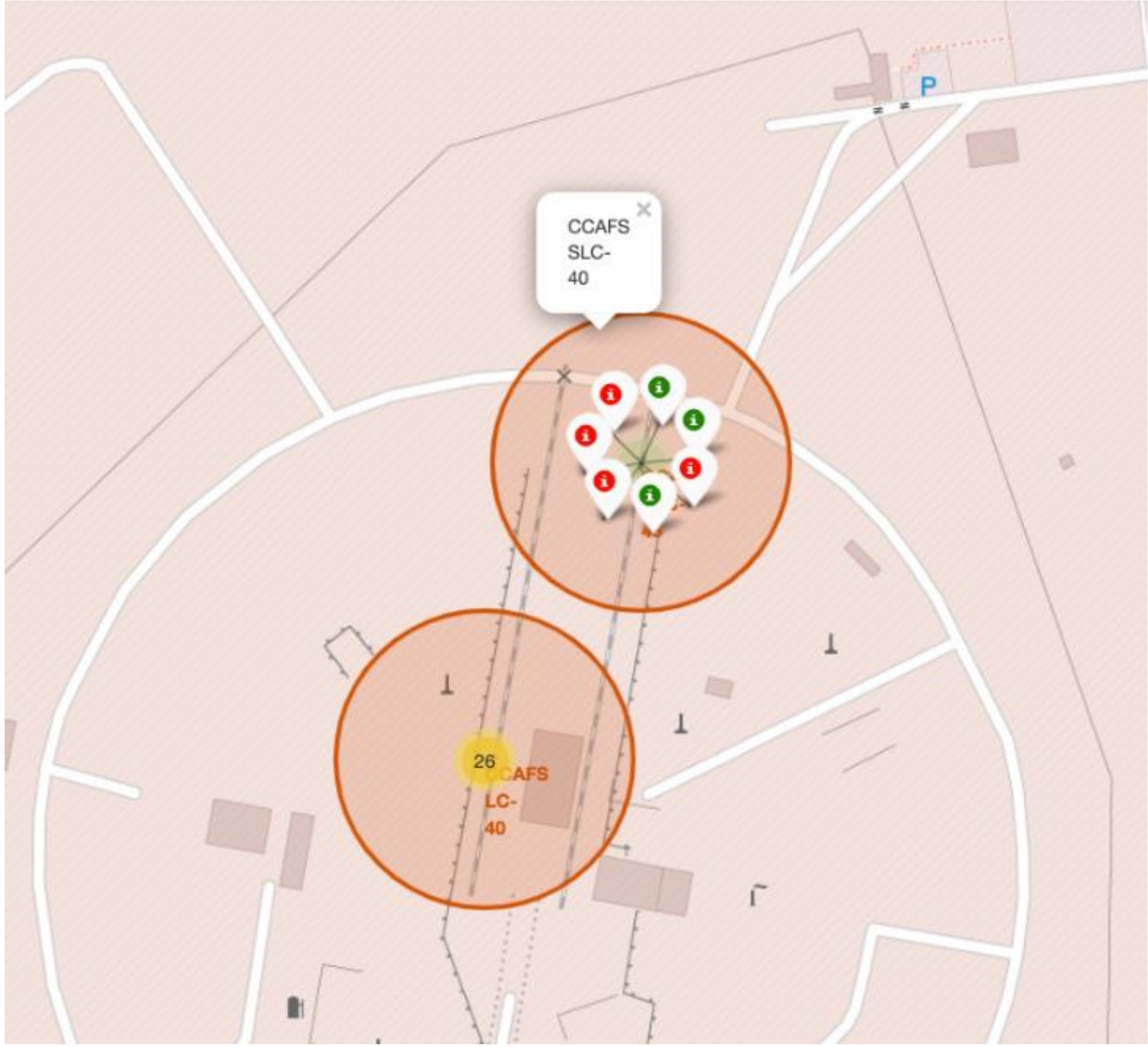
Launch sites near the equator benefit from Earth's rotational speed (1670 km/hour), providing an initial velocity boost to rockets. This helps spacecraft achieve and maintain orbital speed more efficiently.

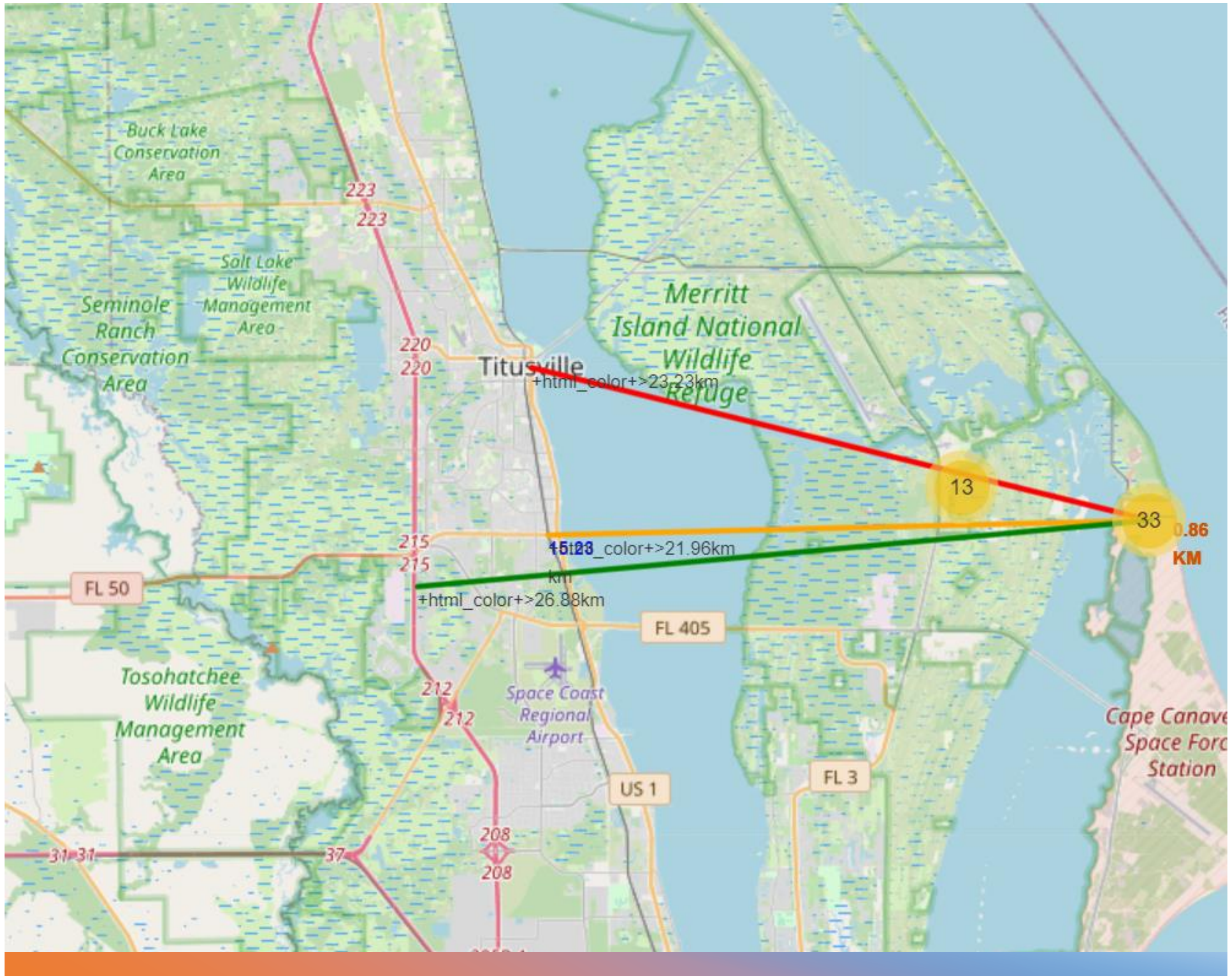
- **Coastal Locations**:

Launching rockets over the ocean reduces the risk of debris or explosions affecting populated areas, ensuring safer operations.

Launch Outcomes

- **Explanation**
- **Color-Coded Markers**:
  The markers make it easy to identify launch site success rates:
  - **Green**: Successful launches
  - **Red**: Failed launches
- **High Success Site**:
  Launch Site KSC LC-39A stands out with a notably high success rate.





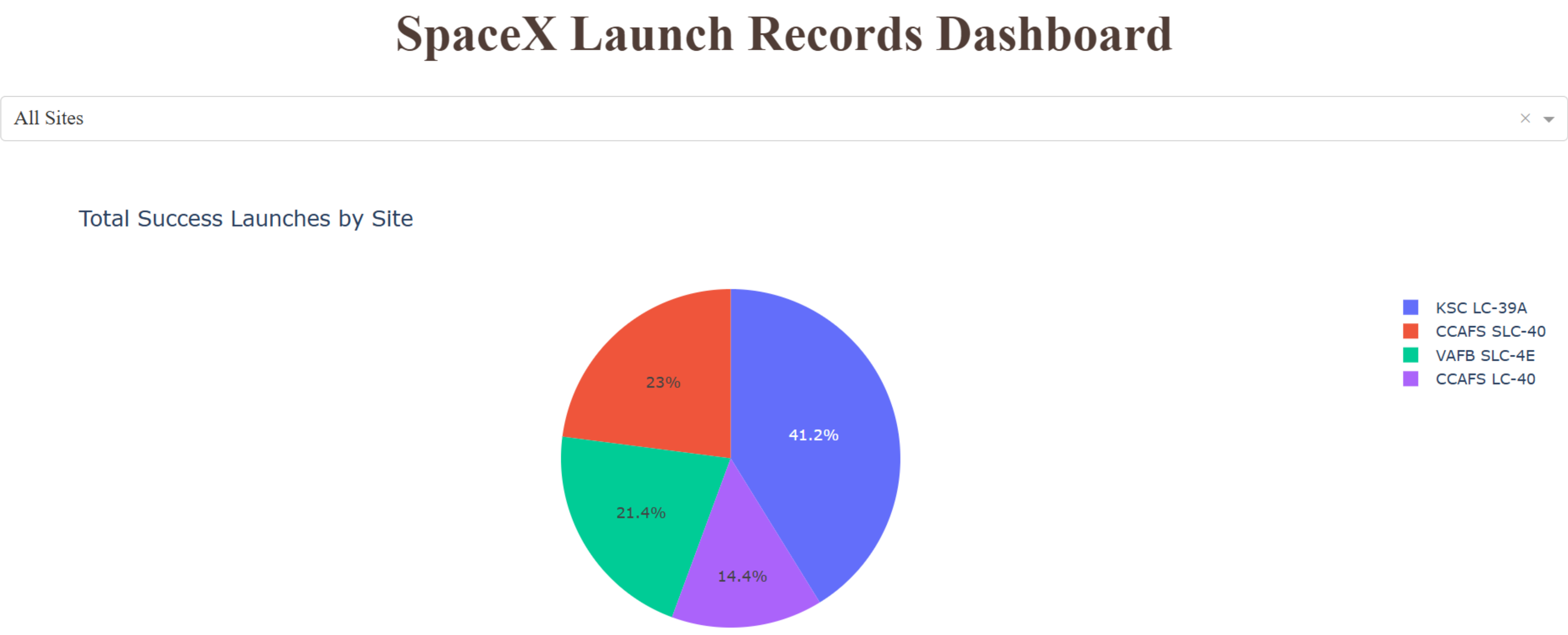
Strategic Placement of Launch Site

- **Coastal Proximity**: Launch sites, like KSC LC-39A (~13 km from the coast), are strategically located to minimize risks to populated areas by directing debris or failed launches over the ocean.
- **City Safety**: Positioned ~23 km from Titusville, ensuring a safe distance from urban areas while maintaining logistical convenience.
- **Infrastructure Access**: Close to transport facilities (~22 km from railways), enabling efficient movement of materials and personnel.

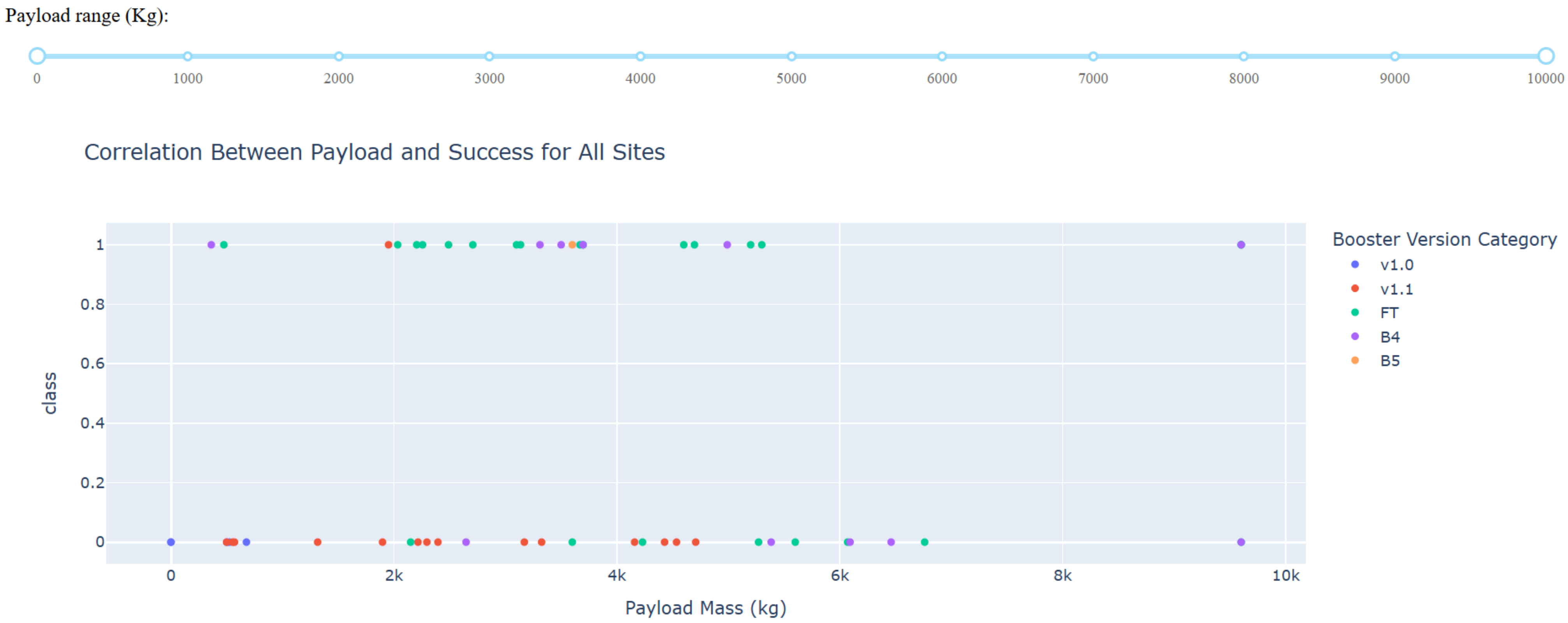
Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site



Payload Mass and Success



Section 5

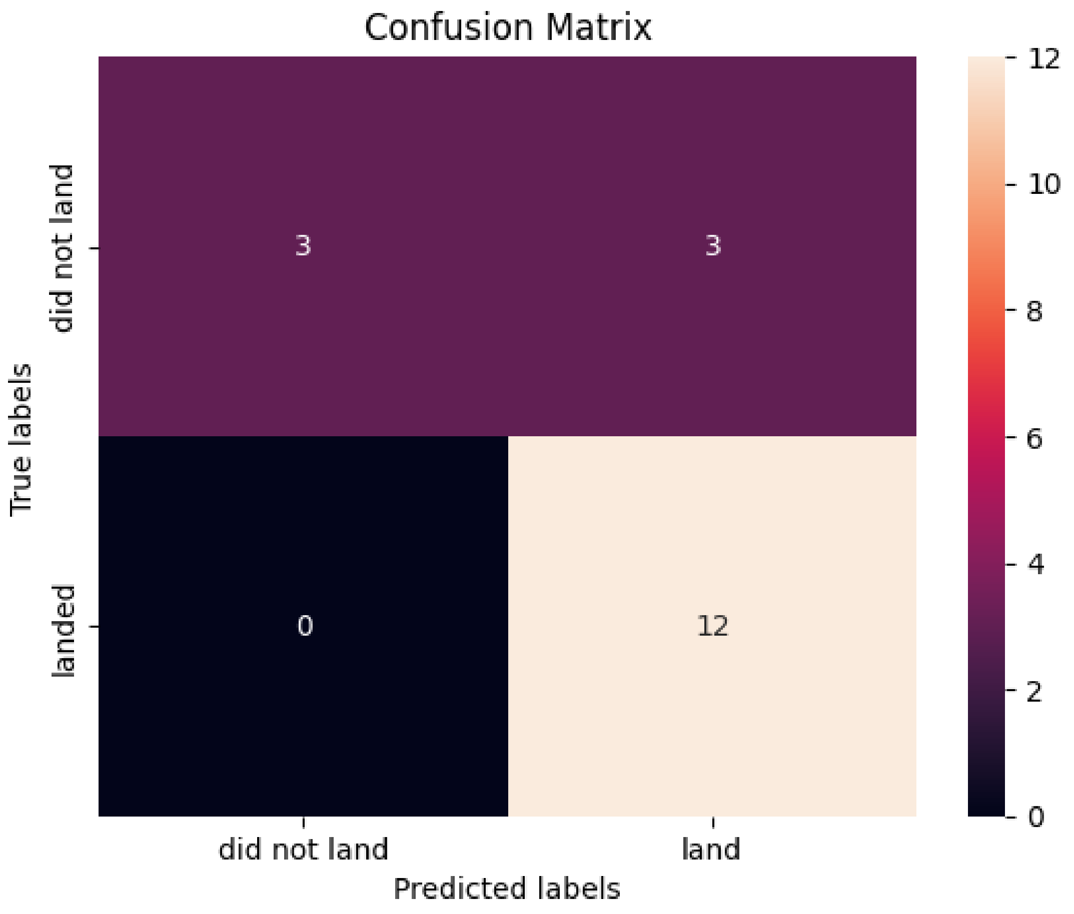
Predictive Analysis (Classification)

Classification Accuracy

- All models achieved similar performance and accuracy, likely due to the small dataset size.
- The Decision Tree model slightly outperformed the others based on its .best_score_ with accuracy of 0.889.
- The .best_score_ represents the average score across all cross-validation folds for a given parameter combination.

Confusion Matrix

Confusion Matrix: Decision Tree Model (Best Performer)



Conclusions

- **Model Performance**: All models performed similarly, with the Decision Tree model slightly outperforming the rest, making it the best algorithm for this dataset.
- **Launch Site Locations**: Most launch sites are strategically placed near the equator for a natural speed boost due to Earth's rotation, reducing fuel costs. All sites are also located close to the coast to ensure safety and minimize risks.
- **Launch Success Trends**:
    - Success rates have increased over time.
    - KSC LC-39A stands out as the most successful site, with a 100% success rate for payloads under 5,500 kg.
    - Orbits such as ES-L1, GEO, HEO, and SSO maintain a 100% success rate.
- **Payload Insights**:
    - Launches with lower payload masses tend to have higher success rates.
    - Across all sites, higher payload masses correlate with increased success rates.
- These findings highlight the importance of strategic site placement, payload considerations, and the increasing reliability of rocket launches over time.

Thank you!