IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aliya Abdugulova
04/04/2025

Outline

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

Executive Summary

- **Summary of methodologies**

  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Ploty Dash
  - Machine Learning Prediction

- **Summary of all results**

  - Exploratory Data Analysis Result: Data trends, distributions, and key correlations were uncovered using SQL queries and visualizations.
  - Interactive Analytics in Screenshots: Dynamic maps and dashboards enabled real-time exploration and intuitive data interaction.
  - Predictive Analytics Result: Machine learning models provided accurate predictions, highlighting critical factors influencing outcomes.

Introduction

Project background and context

SpaceX is an aerospace company revolutionizing space travel with reusable rockets, and satellite deployments.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

This project aims to predict the success of Falcon 9 first-stage landings based on public information and machine learning models.

Problems you want to find answers:

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Used SpaceX Rest API and Web Scraping data from Wikipedia
- Perform data wrangling
  - Calculated the number of launches at each site, the frequency of each orbit type, and the mission outcomes per orbit: unsuccessful first-stage landing and a successful landing.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Trained different classification models – Support Vector Machine (SVM), Decision Trees, and Logistic Regression.
  - Compared predictive classification models to identify the best-performing method.

Data Collection

- The data collection process combined API requests from the SpaceX REST API with web scraping from a table in SpaceX’s Wikipedia entry. Both methods were necessary to obtain comprehensive launch information for a more detailed analysis.

- SpaceX API URL: "https://api.spacexdata.com/v4/launches/past"

- Wikipedia page URL: "https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches"

Data Collection – SpaceX API

1. Request data
Request rocket launch data from SpaceX API

2. Parse JSON response
Decode the response content as a Json and turn it into a Pandas dataframe

3. Extract key information
Request necessary launch information from the SpaceX API using custom functions.

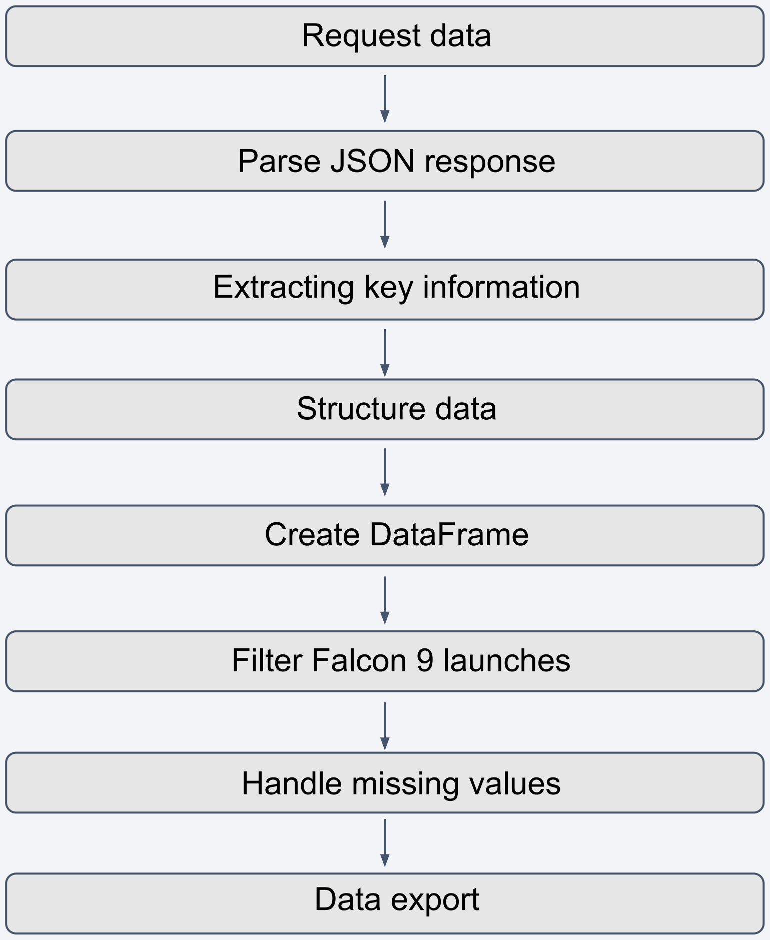
4. Structure data
Combine the columns into a dictionary using the obtained data

5. Create DataFrame
Create a dataframe from the dictionary

6. Filter data
Filter the DataFrame to only include Falcon 9 launches

7. Handle missing values
Replace missing values of Payload Mass with the mean value

8. Data export
Save processed DataFrame to CSV



The GitHub URL: [Data Collection - SpaceX API]

Data Collection – Scraping

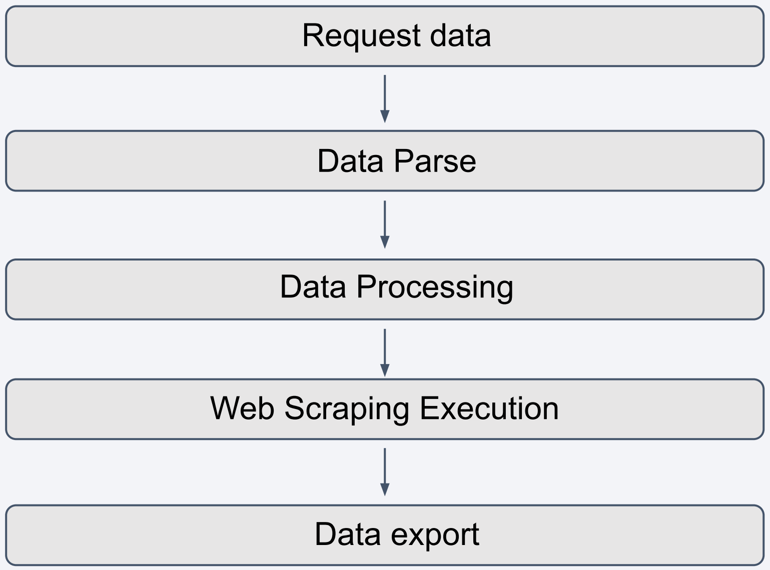
1. Request data
Extract a Falcon 9 launch records HTML table from Wikipedia

2. Data Parse
Process the retrieved HTML content and extract the relevant table.

3. Data Processing
Use helper functions to clean and format the extracted table data.

4. Web Scraping Execution
Perform an HTTP GET request, parse the HTML response with BeautifulSoup, extract table headers, and convert the table into a Pandas DataFrame

5. Data export
Saving processed DataFrame to CSV



The GitHub URL: [Data Collection - Scraping]

Data Wrangling

- The Data Wrangling process involves cleaning and transforming raw data to make it suitable for analysis.
- The dataset includes multiple cases where the booster did not land successfully, with some attempts failing due to accidents; landings are categorized based on whether they were successful or unsuccessful on ocean, ground pads, or drone ships.
- These outcomes are converted into training labels, where 1 indicates a successful landing and 0 represents an unsuccessful attempt.

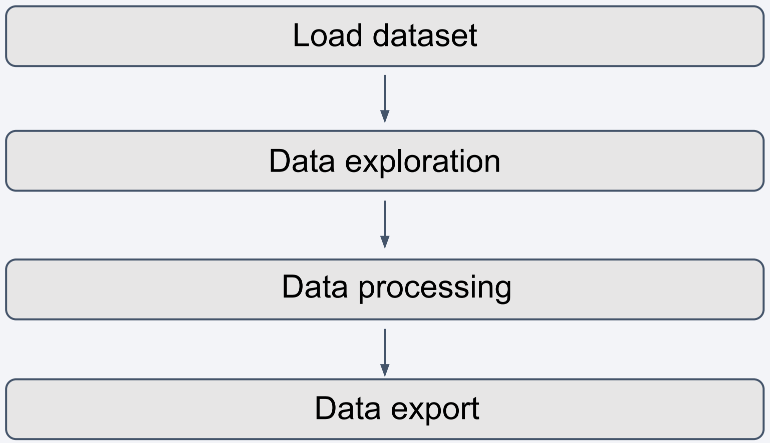
Data Wrangling

1. Load dataset
   Import the SpaceX dataset for analysis.

2. Data exploration
   - Compute the number of launches per site.
   - Analyze the frequency of each orbit type.
   - Determine the distribution of mission outcomes across different orbits.

3. Data processing
   - Generate a landing outcome label based on the Outcome column.
   - Calculate the success..

4. Data export
   Save the processed data to a CSV file for further analysis.



The GitHub URL: [Data Wrangling]

EDA with Data Visualization

Exploratory Data Analysis (EDA) involves visually exploring and summarizing the main characteristics of a dataset. The goal is to understand the data's distribution, identify patterns, and uncover relationships between variables.

Charts were plotted to visualize the relationship between variables:

- 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 reveal how variables relate to each other, making them useful for identifying patterns that could inform machine learning models.

Bar charts highlight differences between distinct categories, aiming to illustrate how specific groups correspond to a measured value.

Line charts depict data trends over time, effectively visualizing changes in time series data.

The GitHub URL: [EDA with Data Visualization]

EDA with SQL

SQL Queries Performed:

- Show each unique launch site
- Show 5 records where launch site names begin with 'CCA'
- Display the total payload mass carried by boosters launched by 'NASA (CRS)'
- Display the average payload mass carried by the v1.1 Falcon 9 booster
- List the date of the first successful ground landing outcome
- List the booster versions with successful outcomes landing on the drone ship with payloads between 4000kg and 6000kg.
- List the number of successful and failed mission outcomes
- List all of the booster versions that carried the max payload mass
- List the month name, outcome, booster version, and launch site for missions with failure outcomes landing on a drone ship in 2015.
- Show the distribution of outcomes between June 4th, 2010 and March 20th, 2017

The GitHub URL: [EDA with SQL]

Build an Interactive Map with Folium

To identify geographical patterns in the data, the following elements were marked on a map of launch sites:

- **All Launch Sites**: This provides an overview of where launches occur, helping to analyze spatial distribution and identify potential geographic influences on launch success.

- **Successful and Failed Launches**: By distinguishing between successful and failed launches, we can examine whether location plays a role in mission outcomes, such as weather conditions, terrain, or proximity to infrastructure.

- **Distances Between a Launch Site and Nearby Landmarks**: Measuring distances to features like coastlines, cities, highways, and railways helps assess how location affects logistics, safety, and accessibility. For example, proximity to the coastline minimizes risk in case of launch failure, while closeness to transportation networks facilitates efficient operations.

The GitHub URL: [An Interactive Map with Folium]

Build a Dashboard with Plotly Dash

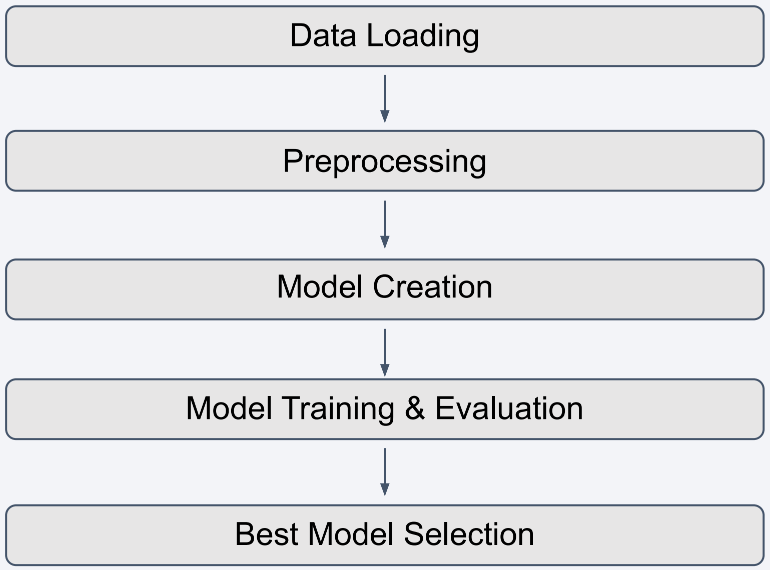
To facilitate interactive data exploration, a Plotly Dash dashboard was created with the following features:

- Dropdown Selector for Launch Sites: Allows users to choose a launch site, dynamically updating the visualizations:
    - Pie Chart:
        - When all sites are selected: Displays the distribution of successful outcomes across all launch sites.
        - When a specific site is selected: Shows the breakdown of successful vs. failed launches for that site.
    - Scatter Plot:
        - When all sites are selected: Visualizes launch outcomes based on payload mass and booster version across all sites.
        - When a specific site is selected: Displays launch outcomes by payload mass and booster version for that site.
- Payload Mass Range Selector: Filters data points on the scatter plot based on the selected payload mass range.

The GitHub URL: [Dashboard with Plotly Dash]

Predictive Analysis (Classification)

1. Data Loading
   - Loaded dataset using
2. Preprocessing
   - Split data into training & testing sets
3. Model Creation
   - Implemented multiple classification models
4. Model Training & Evaluation
   - Trained each model on the training set
   - Predicted outcomes on the test set
   - Evaluated each model using: Accuracy score and Confusion matrix
5. Best Model Selection
   - Compared models based on accuracy scores
   - Identified the best-performing classification model



The GitHub URL: [Machine Learning prediction]

Results

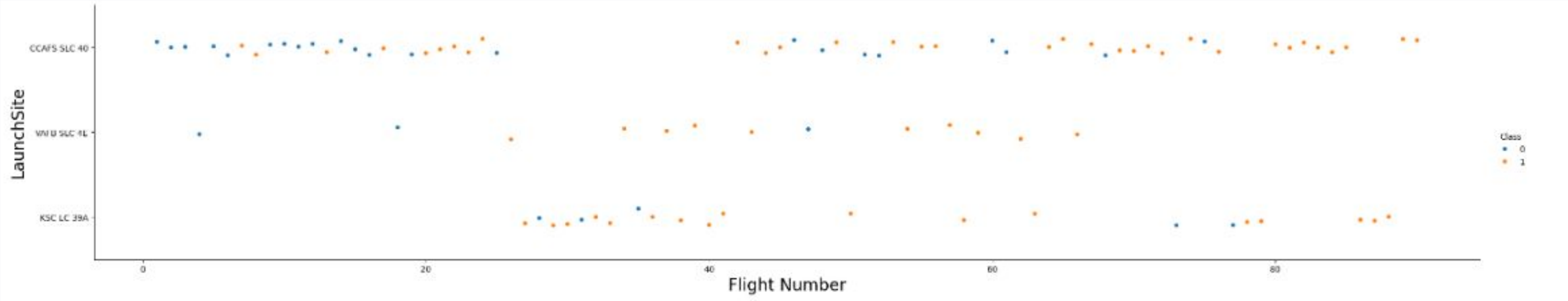
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

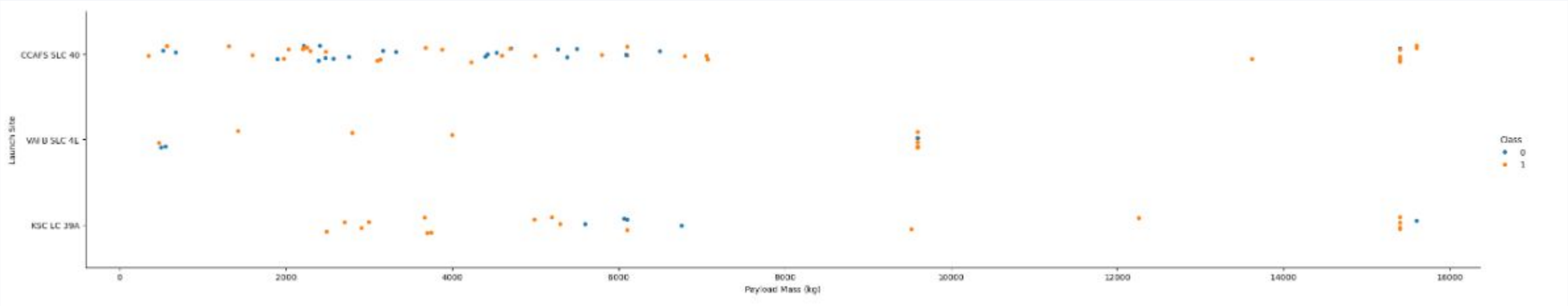
Flight Number vs. Launch Site

This scatter plot shows that as the number of flights increases at a launch site, so does the success rate.



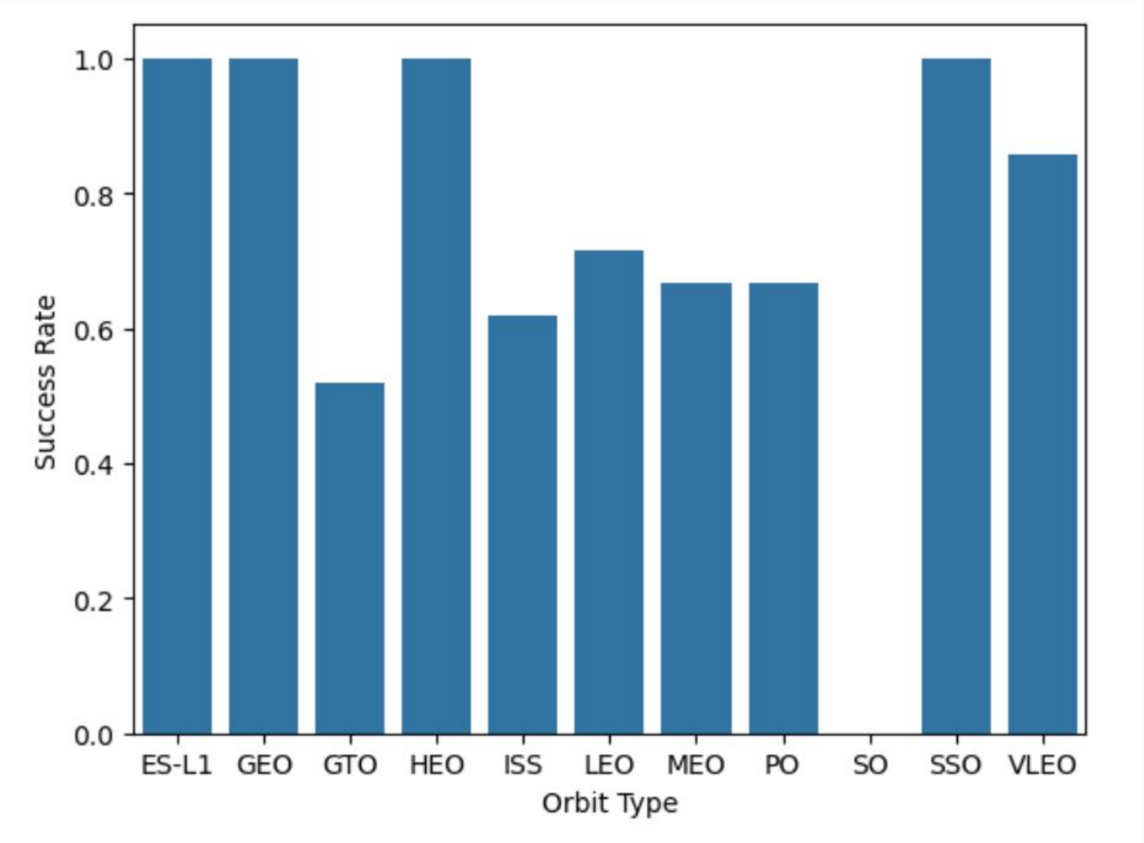
Payload vs. Launch Site

The scatter plot displays how payload and launch site are related, revealing that the success rate increases with larger payloads.



Success Rate vs. Orbit Type

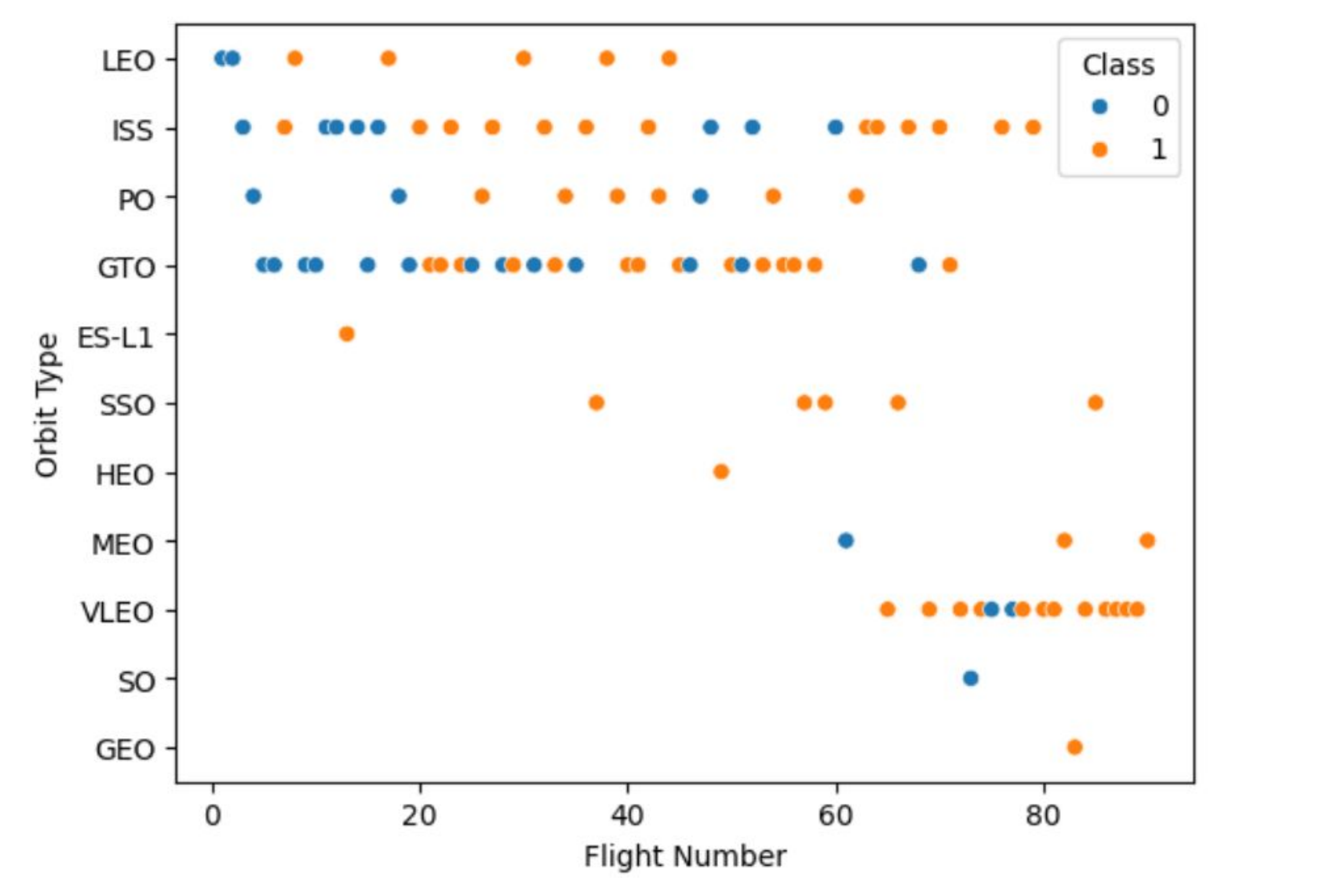
- Some orbits, such as ES-L1, SSO, HEO, and GEO consistently show high success rates.
- Others such as GTO show more mixed outcomes, suggesting some orbit types may introduce operational or technological challenges.
- With only one launch, there is not enough data for the SO orbit type to provide an accurate analysis.



Flight Number vs. Orbit Type

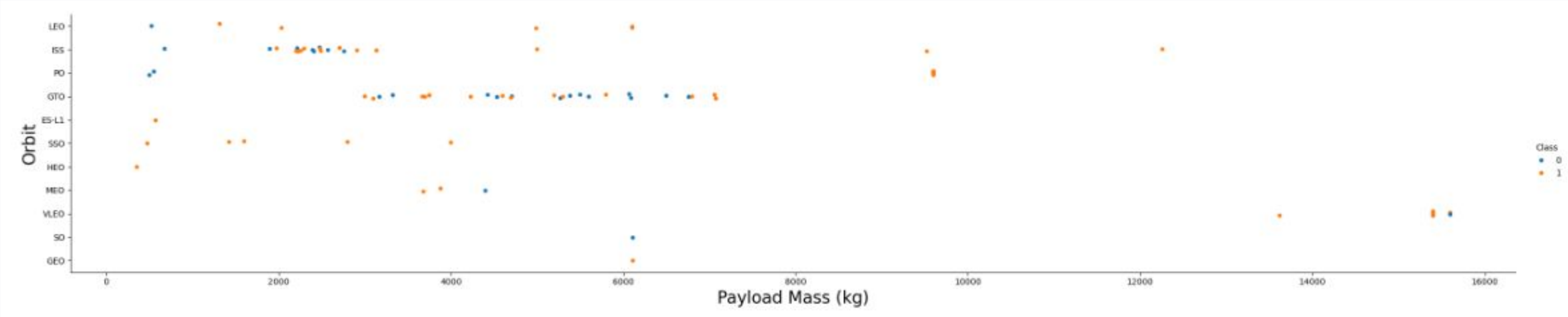
- Numerous orbits are covered across the flight number range, while some are only attempted in later missions.

- Landing success improves significantly with higher flight numbers, reflecting gained experience and continuous enhancements.



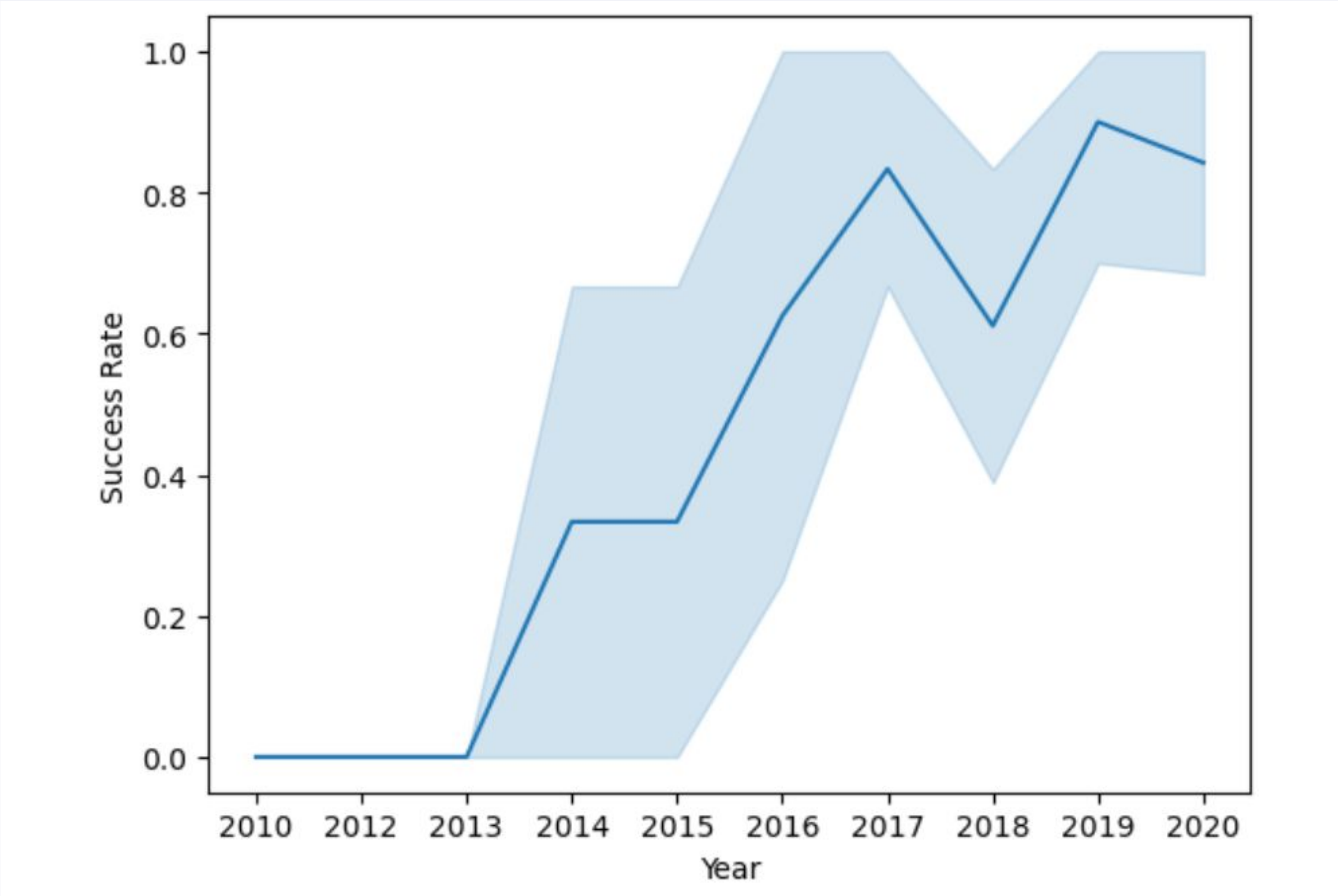
Payload vs. Orbit Type

- Many orbits are represented across a broad spectrum of payload masses, whereas others, such as SSO, MEO, HEO, and GEO, generally fall within a lower range.
- Orbits with a narrower payload range tend to exhibit higher landing success rates.
- While payload mass does not directly dictate mission success, its relationship with orbit suggests a notable correlation.



Launch Success Yearly Trend

According to the line chart the success rate kept increasing since 2013 to 2020.



All Launch Site Names

The following query displays the names of all launch sites used which are CCSFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Display the names of the unique launch sites in the space mission

In [24]:
```
%sql select distinct launch_site from SPACEXTABLE;
```

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

Out[24]:

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

Launch Site Names Begin with 'CCA'

The table below displays 5 records where launch sites begin with the string 'CCA'

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

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD_MASS
111268

Average Payload Mass by F9 v1.1

- The following query displays the average payload carried by the Booster Version F9 v1.1 which is 2928.4 .

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTABLE WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

The first successful landing outcome on ground pad was December 12, 2015

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
```

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

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

See below the list of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 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

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

Mission Status	**Count**
Failure	1
Success	100

List the total number of successful and failure mission outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

Listing the names of the booster versions which have carried the maximum payload mass.

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

2015 Launch Records

List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE DATE LIKE '2015-%' AND \
LANDING_OUTCOME = 'Failure (drone ship)';
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

%sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;

Landing_Outcome	QTY
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 Site Locations

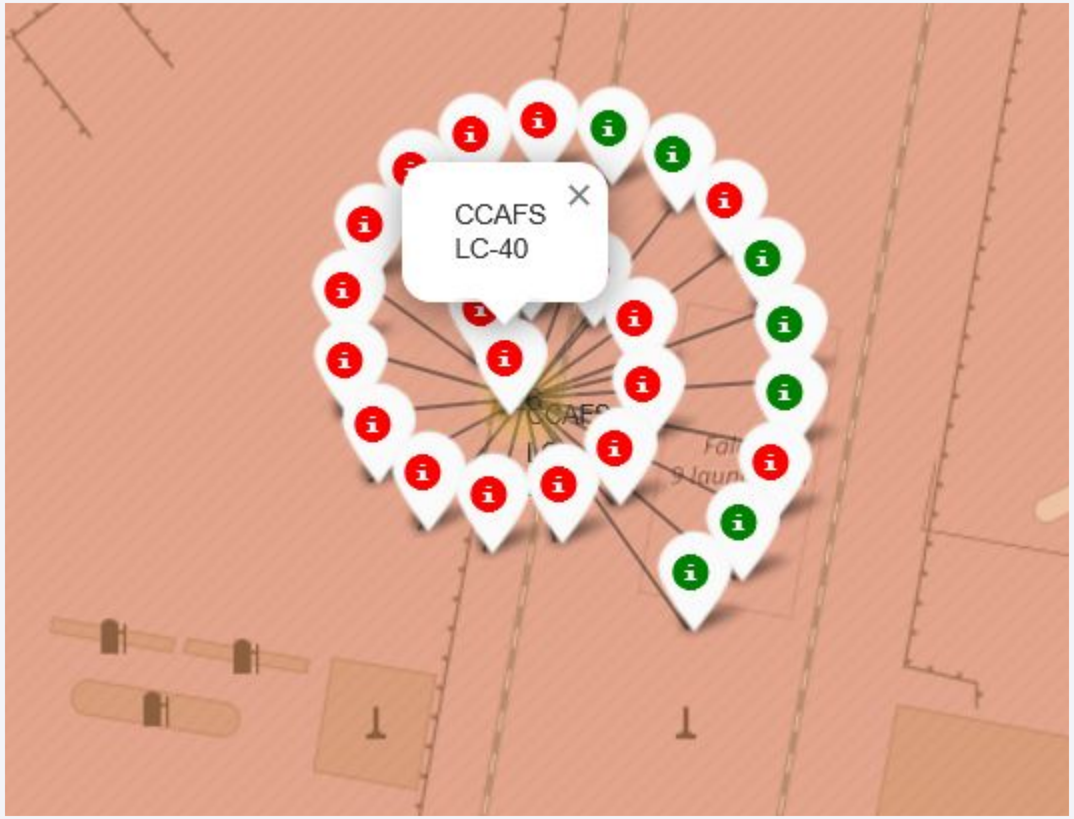
All launch sites are in very close proximity to the coast, while launching rockets towards the ocean it minimises the risk of having any debris dropping or exploding near people.

Launch Outcomes

Note that each launch occurs at one of the four launch sites, meaning many launch records share the same coordinates. Marker clusters help simplify a map with numerous markers at identical locations.

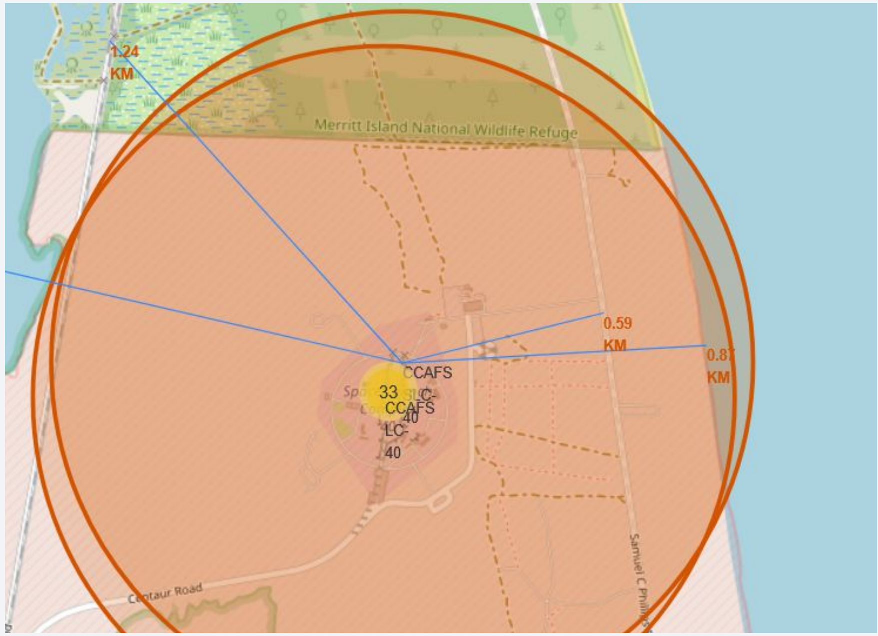
The success/failed launches for each site on the map

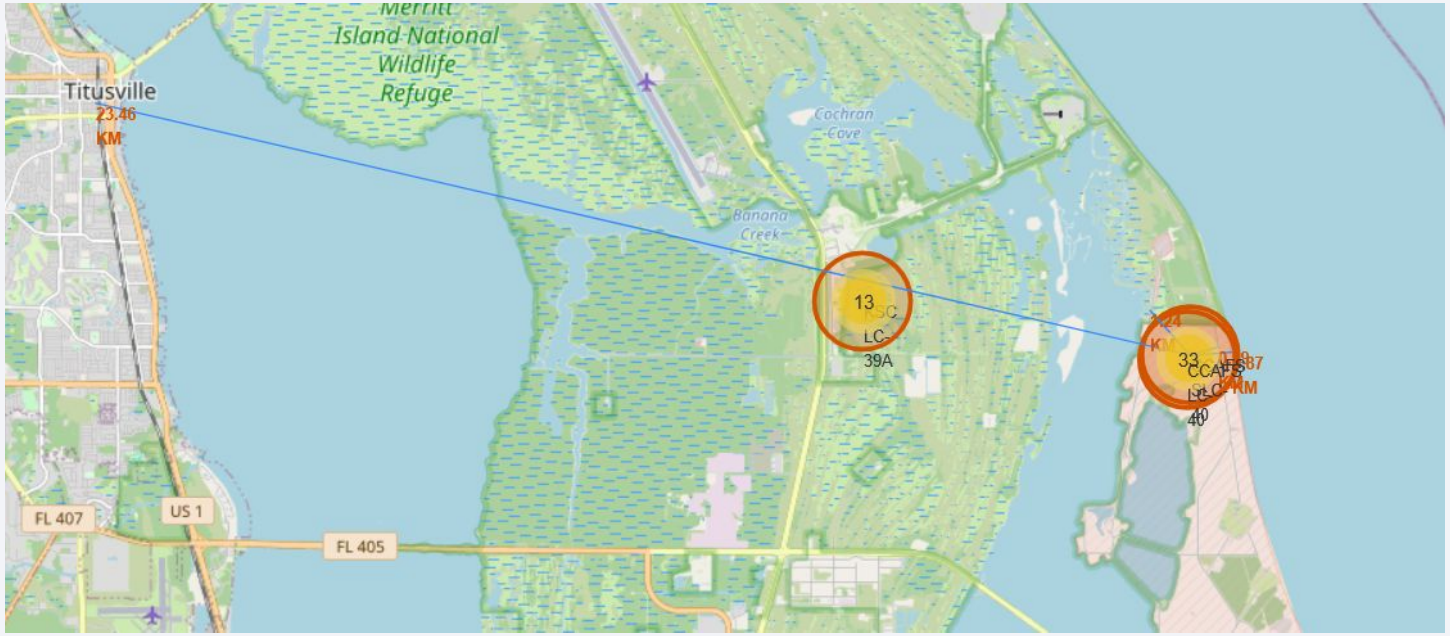
There are color-labeled markers in marker clusters that easily identify which launch sites have relatively high success rates.



Notable Proximate Locations

The screenshots of a launch site show its proximity to features such as railways, highways, and coastlines, with distances calculated and displayed.





Choose the launch site **CCAFS SLC-40**.

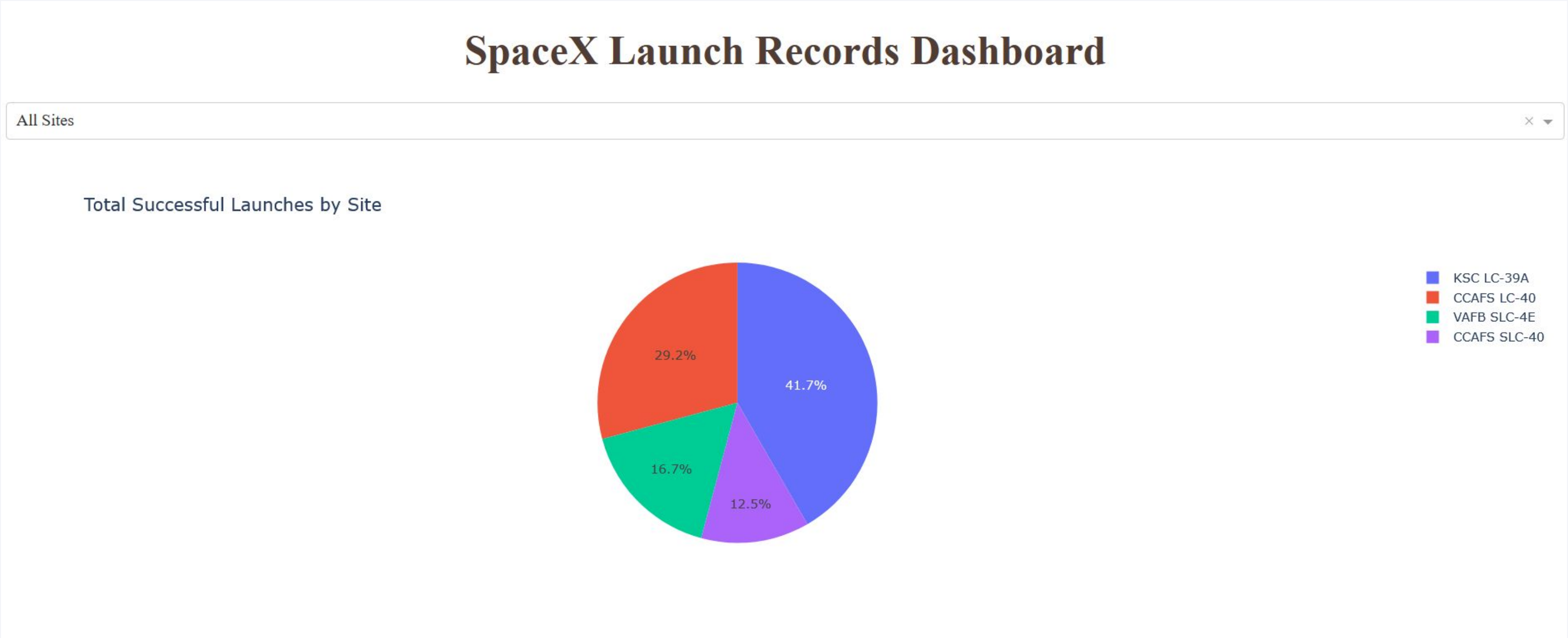
- The distance between the coastline point and the launch site is **0.87 km**.
- The distance between the city of **Titusville** and the launch site is **23.46 km**.
- The distance between the closest railway, **NASA Railroad**, and the launch site is **1.24 km**.
- The distance between **Samuel C. Phillips Parkway** and the launch site is **0.59 km**.

Section 4

Build a Dashboard with Plotly Dash

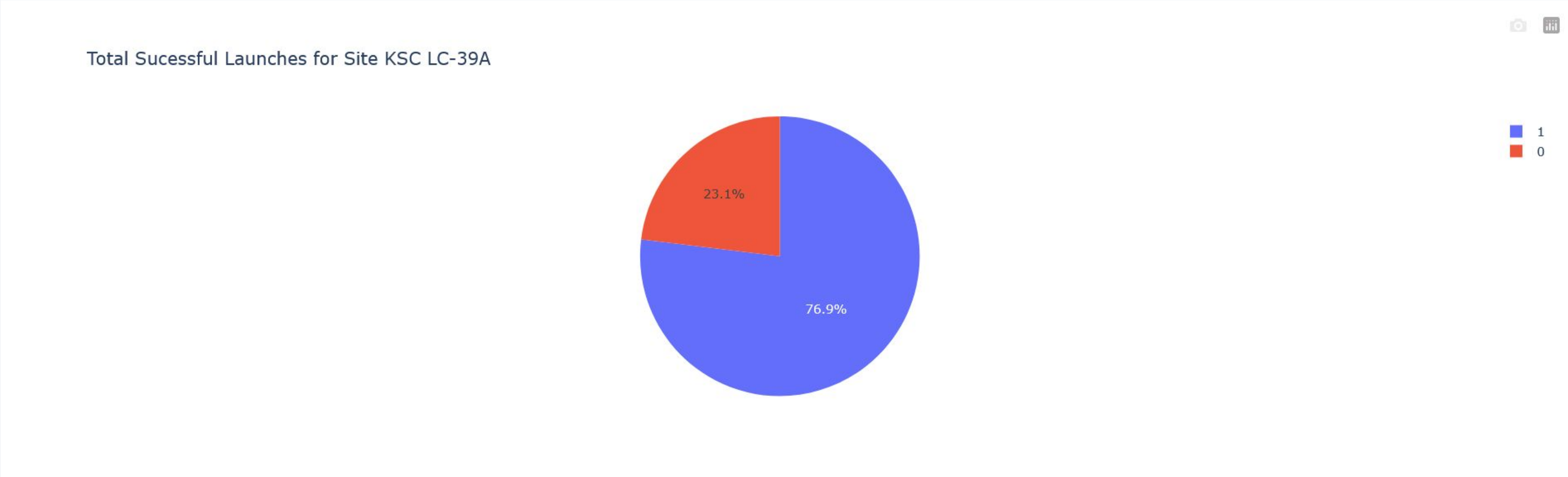
All Launch Sites: Successful Landings

KSC LC-39A experienced the highest proportion of successful landings, followed by CCAFS LC-40. VAFB SLC-4E and CCAFS SLC-40 the lowest.



Launch site with highest launch success ratio

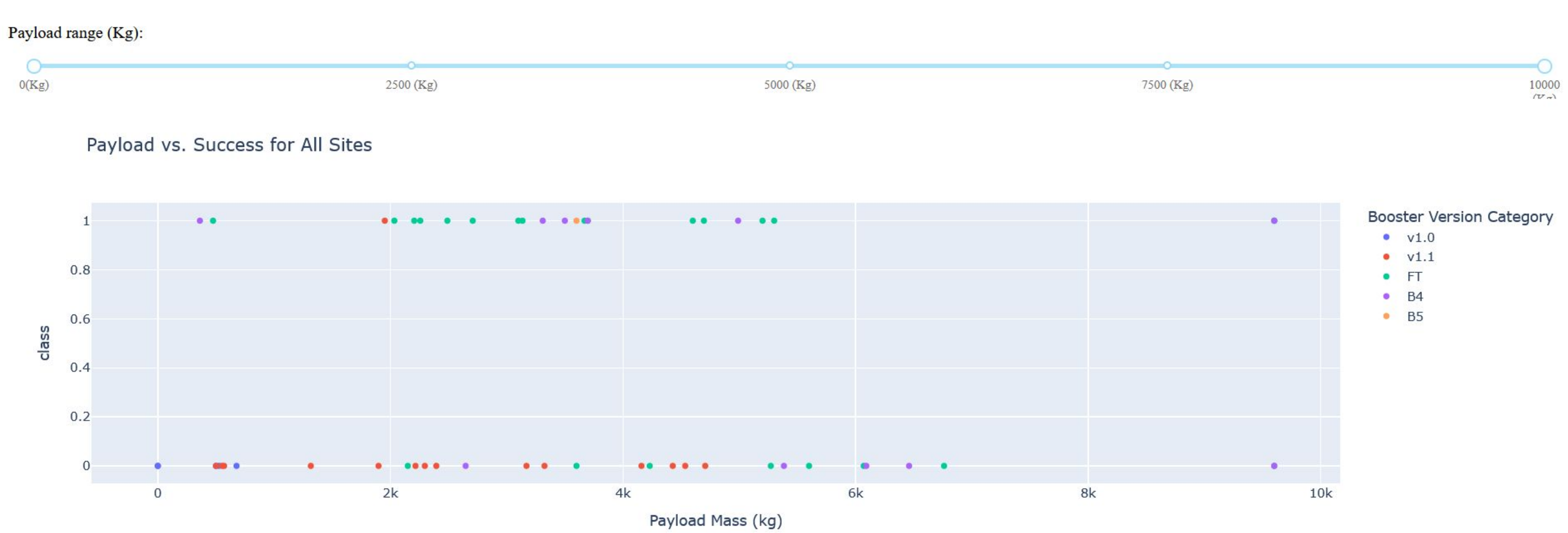
KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.



Payload Mass vs. Launch Outcome for all sites

The scatter plot reveals that payloads between **2000 and 5500 kg** achieve the highest success rate, with a noticeable concentration of successful launches in this range.

Additionally, **newer booster versions** demonstrate improved reliability, as indicated by their higher proportion of successful outcomes compared to older versions.

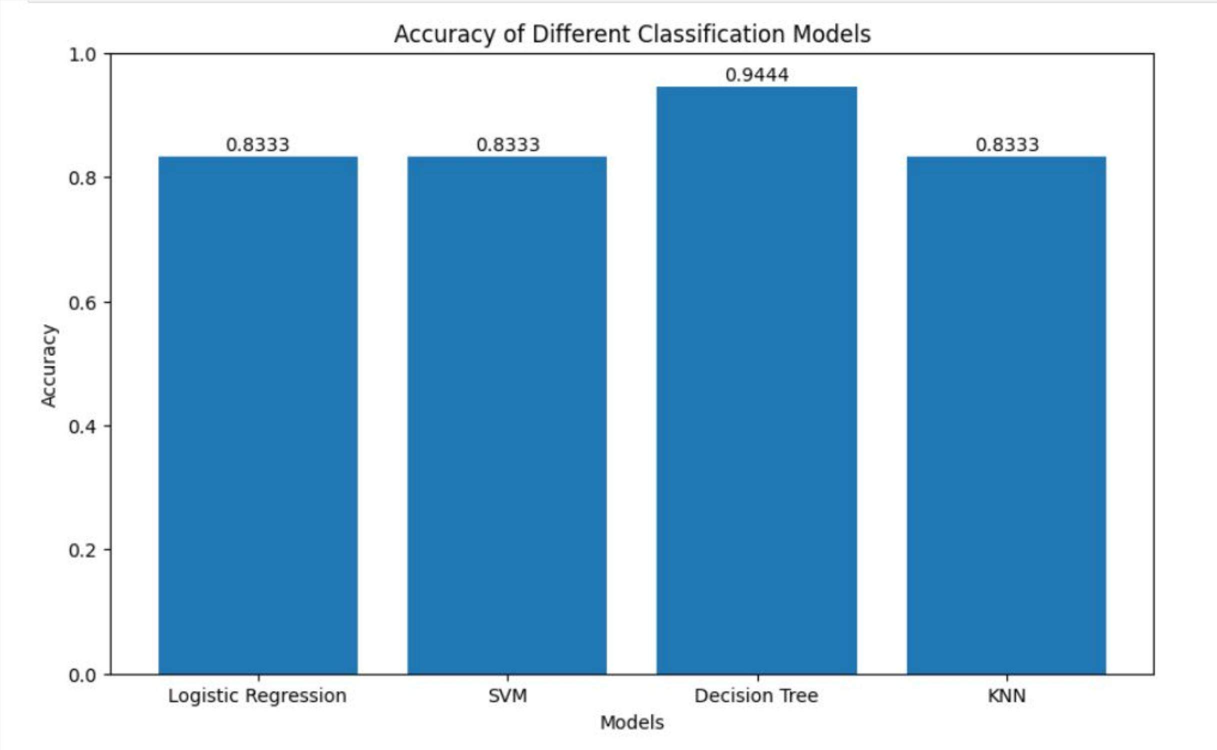


Section 5

Predictive Analysis (Classification)

Classification Accuracy

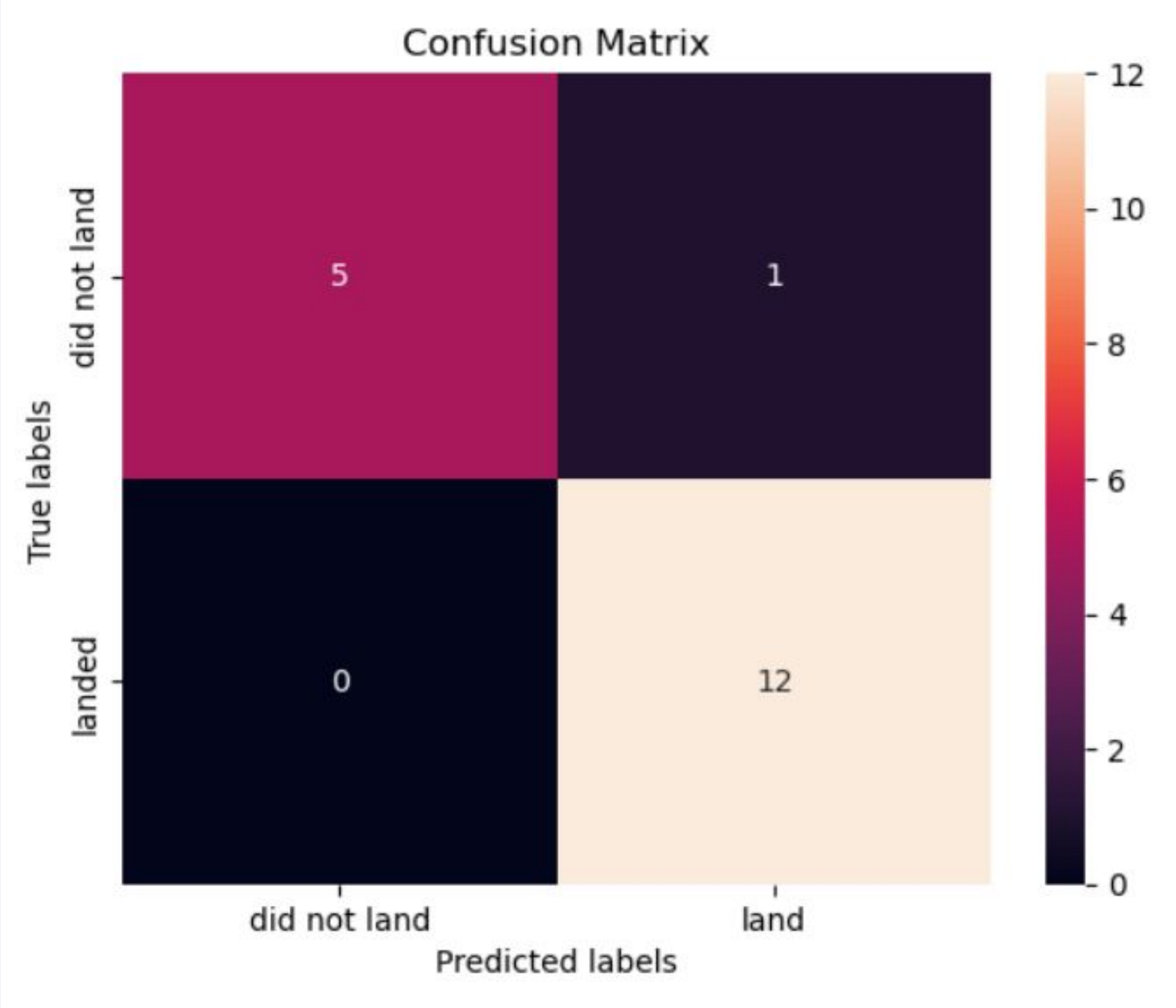
The Decision Tree Classifier has the highest classification accuracy



Confusion Matrix

This confusion matrix evaluates the performance of a classification model that predicts whether a rocket booster successfully lands or not. Here's the breakdown:

- True Positives: The model correctly predicted 11 landings.
- True Negatives: The model correctly predicted 5 instances where the rocket did not land.
- False Positives: The model incorrectly predicted a landing when the rocket did not land.
- False Negatives: The model incorrectly predicted a failure to land when the rocket actually landed.



Conclusions

1. The Decision Tree Model is the most suitable algorithm for this dataset.
2. Launches with lower payload mass tend to have higher success rates compared to those with heavier payloads.
3. Most launch sites are located very close to the coastline.
4. The success rate of launches has shown an increasing trend over the years.
5. Among all sites, KSC LC-39A records the highest launch success rate.
6. Orbits ES-L1, GEO, HEO, and SSO have achieved a 100% success rate.

Thank you!