IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Ballesteros
16/03/2025

Outline

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

Executive Summary

Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Introduction

Project Background and Context

SpaceX is a leader in the commercial space industry, making space travel more accessible through affordable launches. While Falcon 9 launches are advertised at $62 million, competitors can charge over $165 million, largely because SpaceX reuses the first stage. By predicting whether the first stage will land successfully, we can better estimate launch costs. Using publicly available data and machine learning, our goal is to predict the reusability of the first stage.

Questions to Be Answered

1.How do factors like payload mass, launch site, number of flights, and orbit types influence first stage landing success?
2.Has the success rate of first stage landings improved over time?
3.Which binary classification algorithm yields the best performance for predicting first stage landings?

Section 1

Methodology

Methodology

- **Data Collection Methodology:**
  - Obtained data via SpaceX REST API.
  - Scraped additional data from Wikipedia.
- **Data Wrangling:**
  - Filtered the dataset.
  - Handled missing values.
  - Applied One Hot Encoding for binary classification.
- **Exploratory Data Analysis (EDA):**
  - Visualized data using various plotting libraries.
  - Queried data using SQL.
- **Interactive Visual Analytics:**
  - Created interactive maps with Folium.
  - Developed dashboards with Plotly Dash.
- **Predictive Analysis:**
  - Built, tuned, and evaluated multiple classification models to optimize performance.

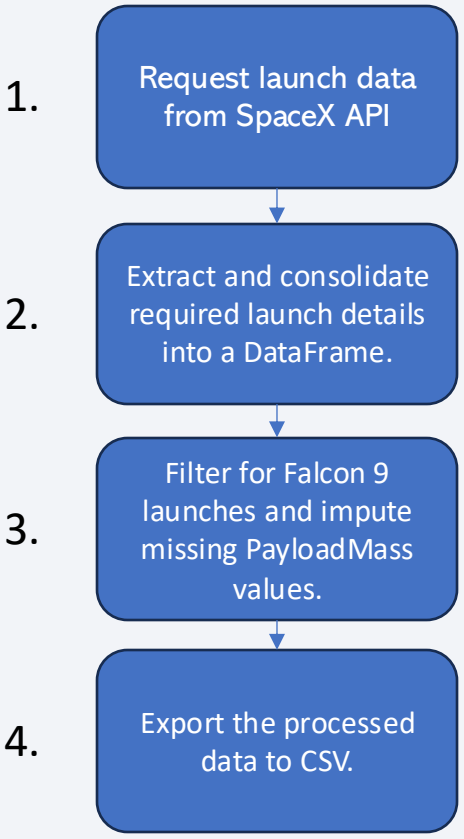
Data Collection

- **Combined Methods:**
  - Utilized SpaceX REST API for comprehensive launch details.
  - Performed web scraping on SpaceX's Wikipedia table for supplemental information.
- **SpaceX API Data:**
  - Collected key launch attributes:
    - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- **Wikipedia Data:**
  - Extracted additional columns for enriched analysis:
    - Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch Outcome, Booster Version, Booster Landing, Date, Time.
- **Purpose:**
  - Merged these datasets to achieve a more detailed and complete analysis of SpaceX launches.

Data Collection – SpaceX API

GITHUB - click here

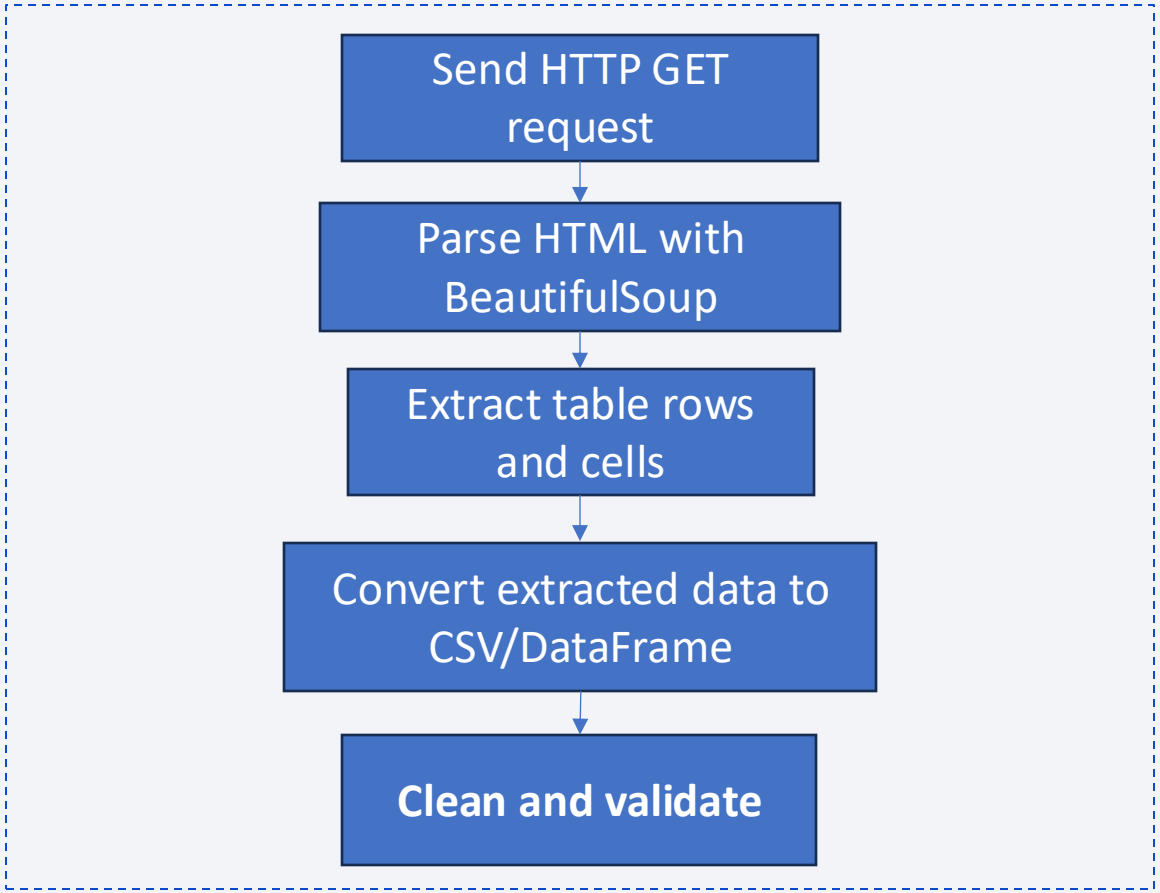
- **Approach Overview**
- **Data Retrieval:**
  - Employed the official SpaceX REST API to obtain detailed launch records.
- **Key Steps:**
  - **Initialize:** Define API endpoint URLs (e.g., https://api.spacexdata.com/v4/launches).
  - **Request:** Execute RESTful GET requests for each launch record.
  - **Parse:** Convert JSON responses into structured pandas DataFrames.
  - **Merge & Clean:** Combine multiple DataFrames and remove duplicates or incomplete records.
  - **Store:** Save the final consolidated dataset to CSV for further analysis.
- **Key Phrases:**
  - "RESTful GET requests"
  - "JSON to DataFrame"
  - "Merging & cleaning"
  - "CSV output"
- **Flowchart:**
  - Illustrates the process from API call → Data extraction → Parsing → Merging/Cleaning → Data storage.



Data Collection - Scraping

GITHUB - click here

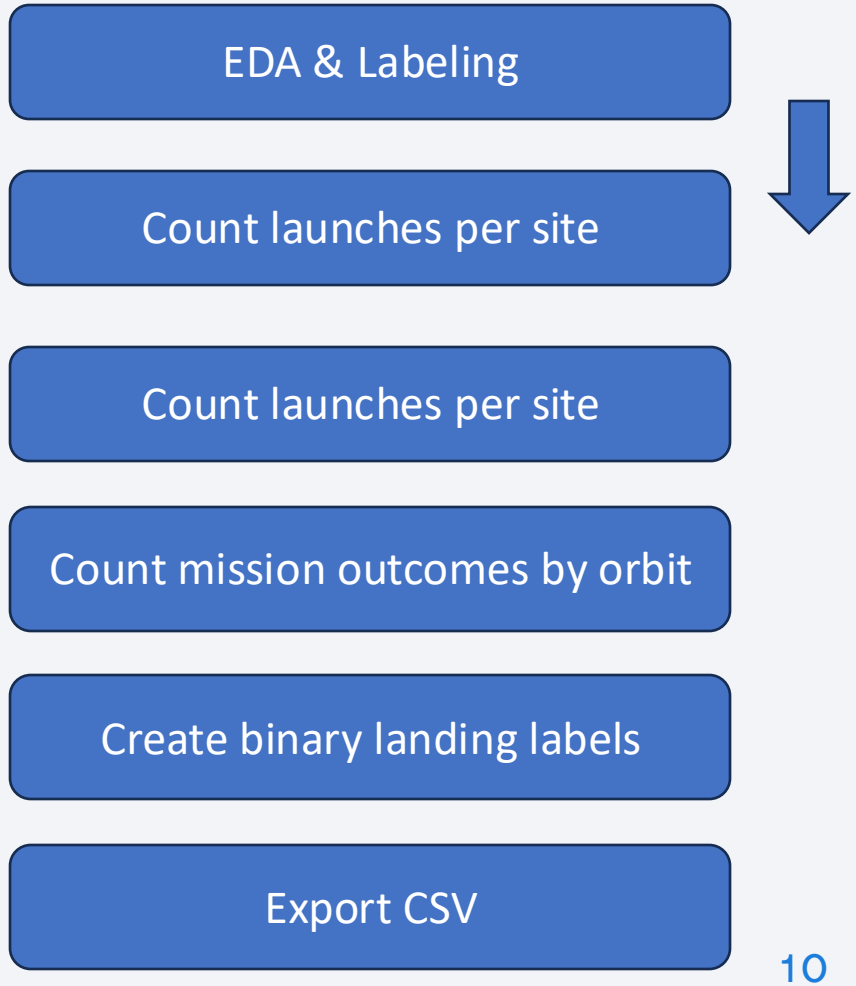
- **Web Scraping Overview**
- **Fetch HTML:**
  - Initiated an HTTP GET request to retrieve the HTML content from the SpaceX Wikipedia page containing Falcon 9 launch data.
- **Parse Content:**
  - Utilized BeautifulSoup to parse the HTML document.
  - Identified the specific HTML table that contained the launch details.
- **Extract Structure:**
  - Located and extracted the table header to capture column names.
  - Iterated through each table row and cell to extract the corresponding launch data.
- **Organize Data:**
  - Compiled the extracted information into a structured dictionary.
  - Converted the dictionary into a Pandas DataFrame to facilitate data manipulation and analysis.
- **Export Data:**
  - Saved the cleaned and structured DataFrame to a CSV file for further processing and validation.



Data Wrangling

GITHUB - click here

In the dataset, various scenarios indicate when the booster did not land successfully. There are instances where a landing was attempted but ended in failure, such as when "True Ocean" signifies a successful ocean landing, while "False Ocean" indicates an unsuccessful attempt at an ocean landing. Similarly, "True RTLS" denotes a successful landing on a ground pad, and "False RTLS" represents a failed ground pad landing. In the case of drone ship landings, "True ASDS" means the booster landed successfully on the ship, and "False ASDS" means it did not. We simplify these outcomes into training labels where a “1” represents a successful landing, and a “0” indicates a failure.



EDA with Data Visualization

GITHUB - click here

Charts plotted include:

•**Scatter plots:** Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Flight Number vs. Orbit Type, and Payload Mass vs. Orbit Type. These illustrate relationships between variables, which can inform machine learning models if strong correlations exist.

•**Bar charts:** Orbit Type vs. Success Rate, which compare discrete categories and their measured values.

•**Line charts:** Yearly Success Rate Trend, which display data trends over time.

EDA with SQL

GITHUB - click here

- Executed SQL queries to:
- Retrieve unique launch site names.
- Fetch 5 records with launch sites beginning with "CCA".
- Compute the total payload mass carried by NASA (CRS) boosters.
- Determine the average payload mass for booster version F9 v1.1.
- Identify the earliest date of a successful ground pad landing.
- List boosters that successfully landed on a drone ship with payload mass between 4000 and 6000.
- Sum the total numbers of successful and failed mission outcomes.
- Extract booster versions that carried the maximum payload mass.
- Select records for failed drone ship landings in 2015, including booster versions and launch site names.
- Rank landing outcomes between 2010-06-04 and 2017-03-20 in descending order based on their counts.

Build an Interactive Map with Folium

GITHUB - click here

Markers of All Launch Sites:

- Added a marker with a circle, popup, and text label for NASA Johnson Space Center using its latitude and longitude as the starting point.
- Added similar markers for all launch sites to display their geographic positions and proximity to the Equator and coastlines.

Colored Markers for Launch Outcomes:

- Implemented colored markers (green for success, red for failure) using Marker Cluster to highlight launch sites with higher success rates.

Distances to Nearby Features:

- Drew colored lines to illustrate distances from the launch site (e.g., KSC LC-39A) to nearby features like railways, highways, coastlines, and the closest city.

Build a Dashboard with Plotly Dash

GITHUB - click here

Launch Sites Dropdown List:

- Implemented a dropdown menu to allow selection of a specific launch site.

Pie Chart Showing Launch Success:

- Integrated a pie chart that displays the overall successful launch count across all sites, and when a site is selected, it shows the breakdown of successes versus failures.

Payload Mass Range Slider:

- Added a slider to adjust and select a specific payload mass range.

Scatter Chart of Payload Mass vs. Success Rate:

- Developed a scatter plot to visualize the correlation between payload mass and launch success across different booster versions.

Predictive Analysis (Classification)

GITHUB - click here

Model Development Process:

•**Data Preprocessing:** Cleaned the data, handled missing values, and standardized the features.

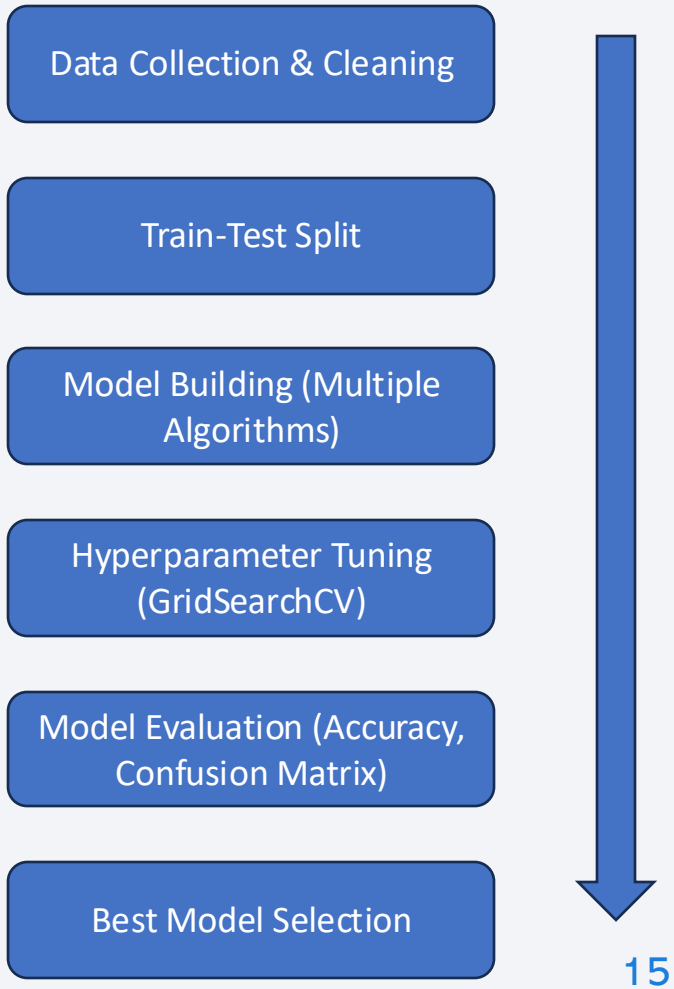
•**Data Splitting:** Divided the data into training (80%) and test (20%) sets.

•**Model Selection:** Evaluated multiple classifiers including Logistic Regression, SVM, Decision Tree, and KNN.

•**Hyperparameter Tuning:** Used GridSearchCV with cross-validation (cv = 10) to find the best hyperparameters.

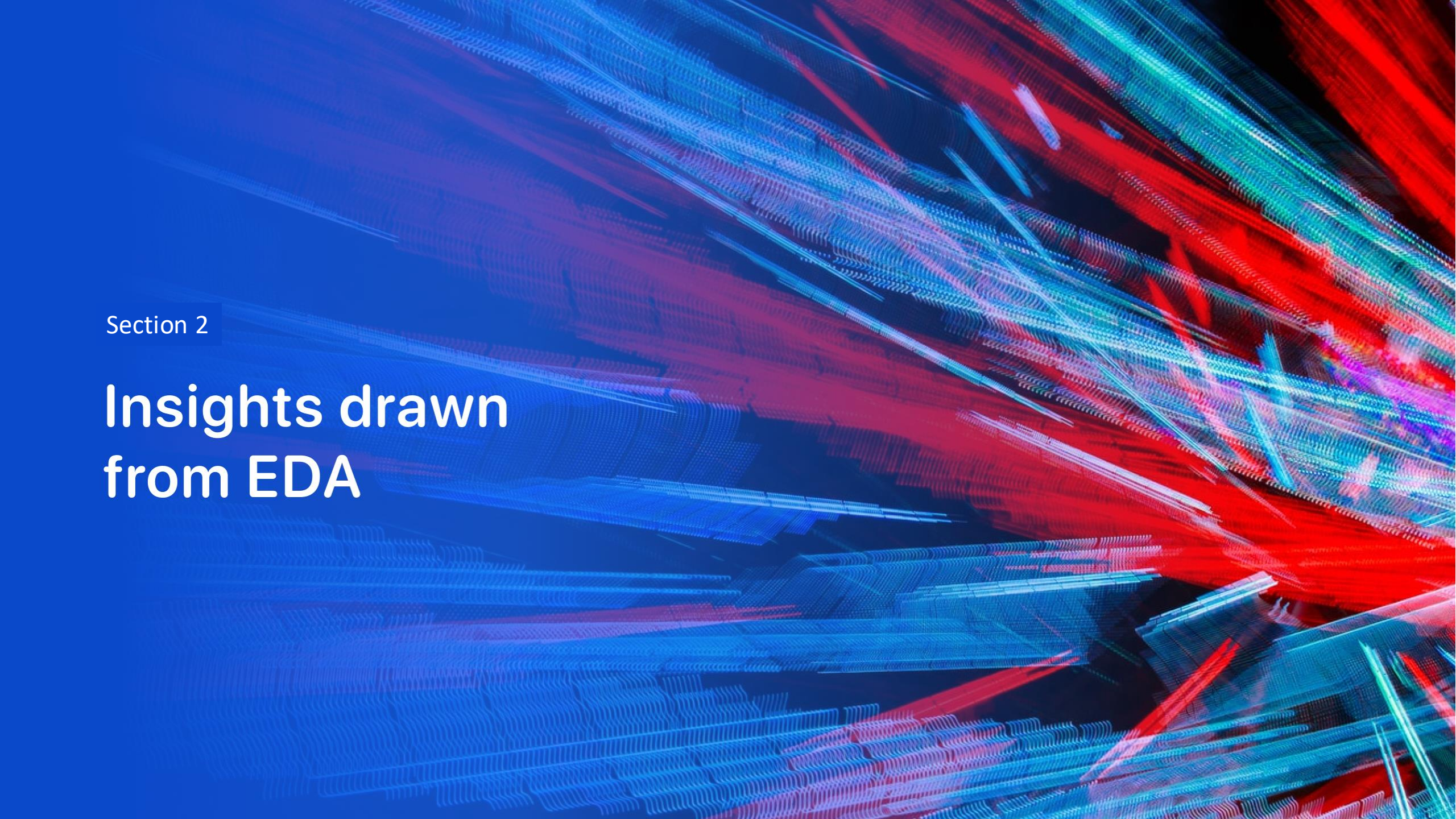
•**Model Evaluation:** Compared model performance using accuracy scores and confusion matrices.

•**Best Model Selection:** Determined the Decision Tree classifier performed best based on validation accuracy.

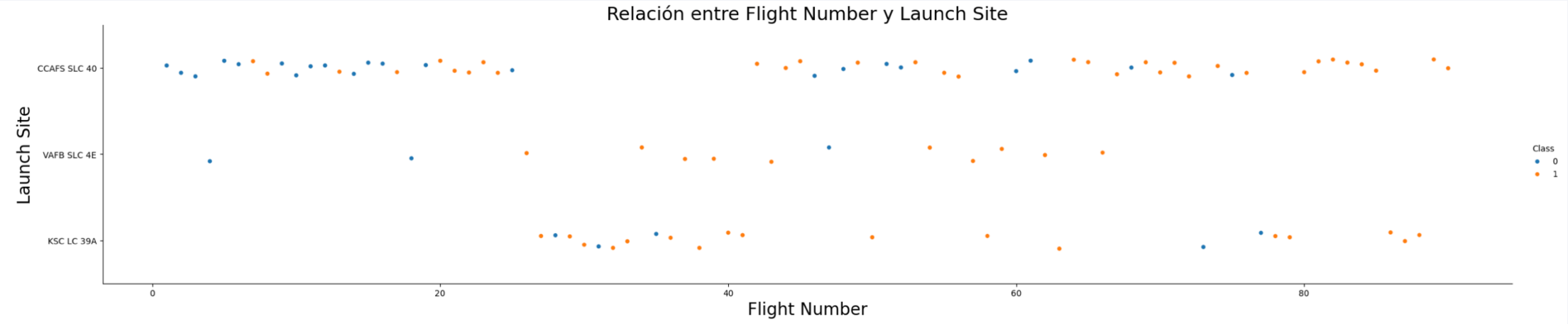


Results

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

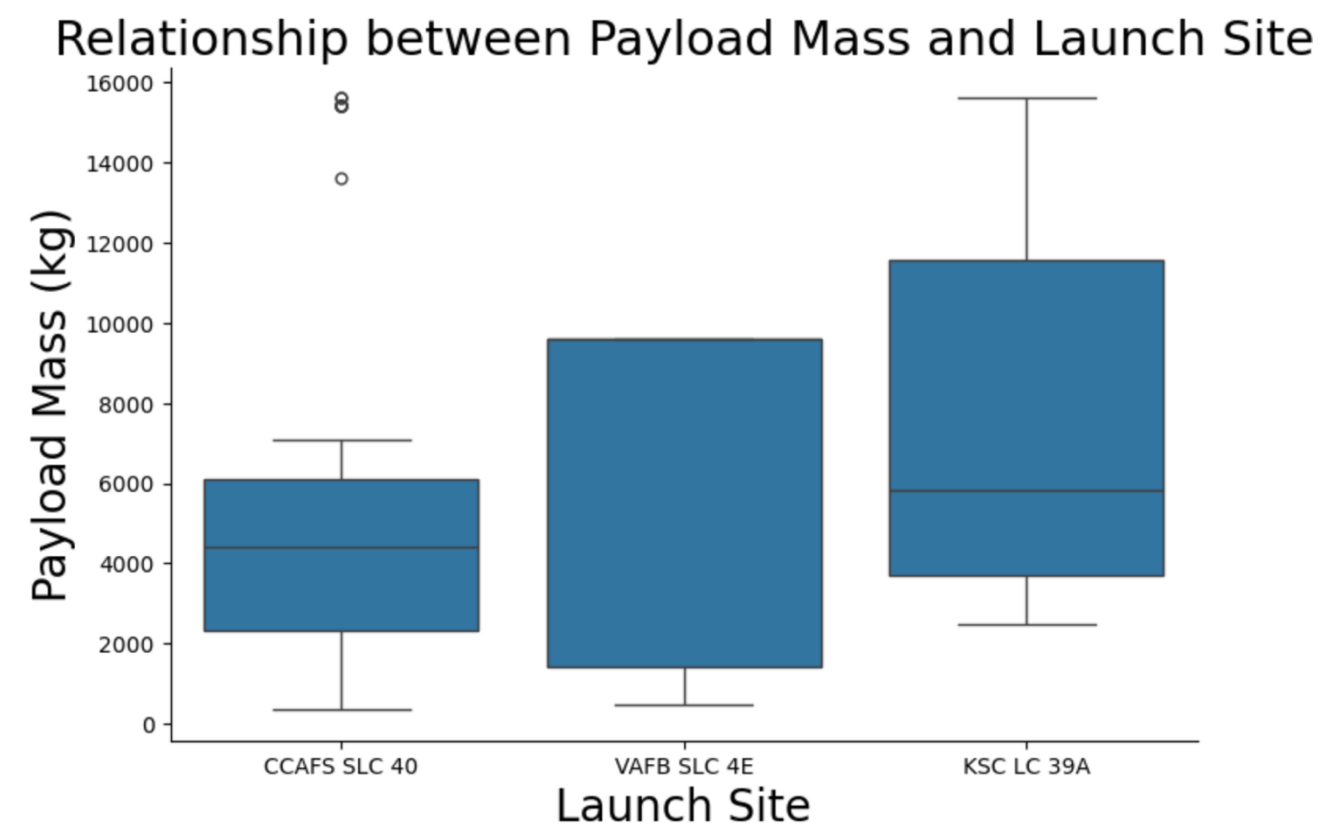

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



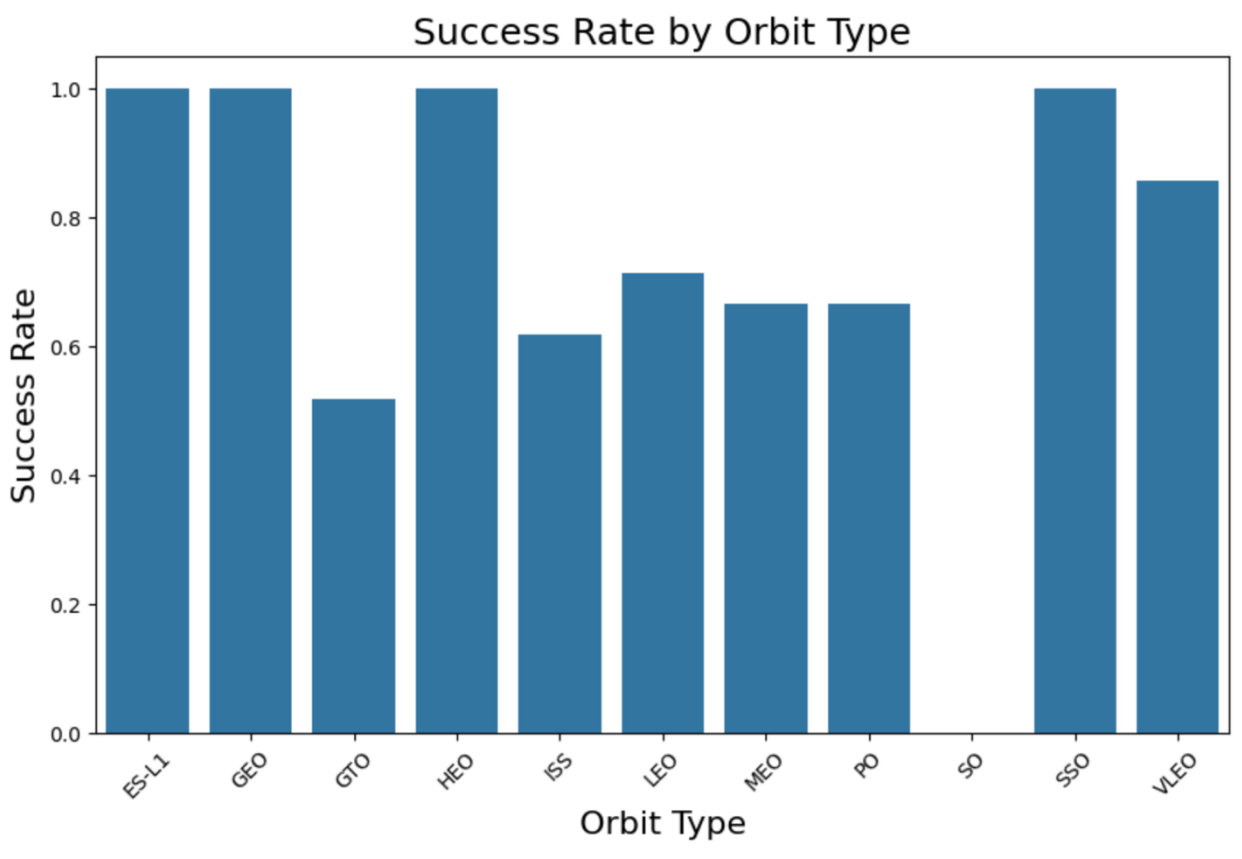
- The earliest flights resulted in failures, whereas the most recent flights have all been successful.
- Approximately 50% of all launches occurred at the CCAFS SLC 40 site.
- Both VAFB SLC 4E and KSC LC 39A demonstrate higher success rates.
- There is a clear trend showing that newer launches tend to have increased success.

Payload vs. Launch Site



- At every launch site, higher payload masses are associated with higher success rates.
- Most launches carrying more than 7000 kg were successful.
- Additionally, KSC LC 39A achieved a 100% success rate for payloads under 5500 kg.

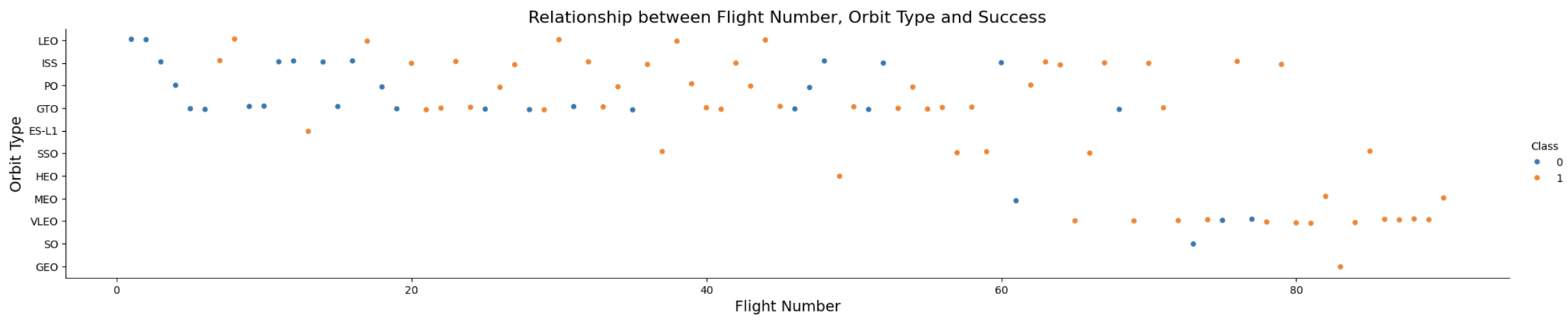
Success Rate vs. Orbit Type



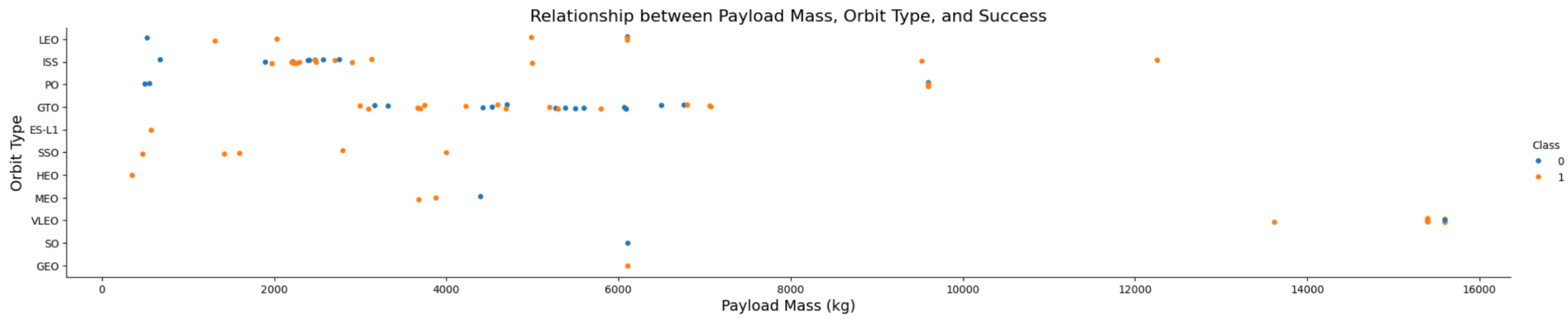
- **100% Success Rate:** ES-L1, GEO, HEO, and SSO
- **0% Success Rate:** SO
- **Moderate Success Rates (50%-85%):** GTO, ISS, LEO, MEO, and PO

Flight Number vs. Orbit Type

In LEO orbit, success correlates with the number of flights, while in GTO orbit, there is no apparent relationship with flight number.

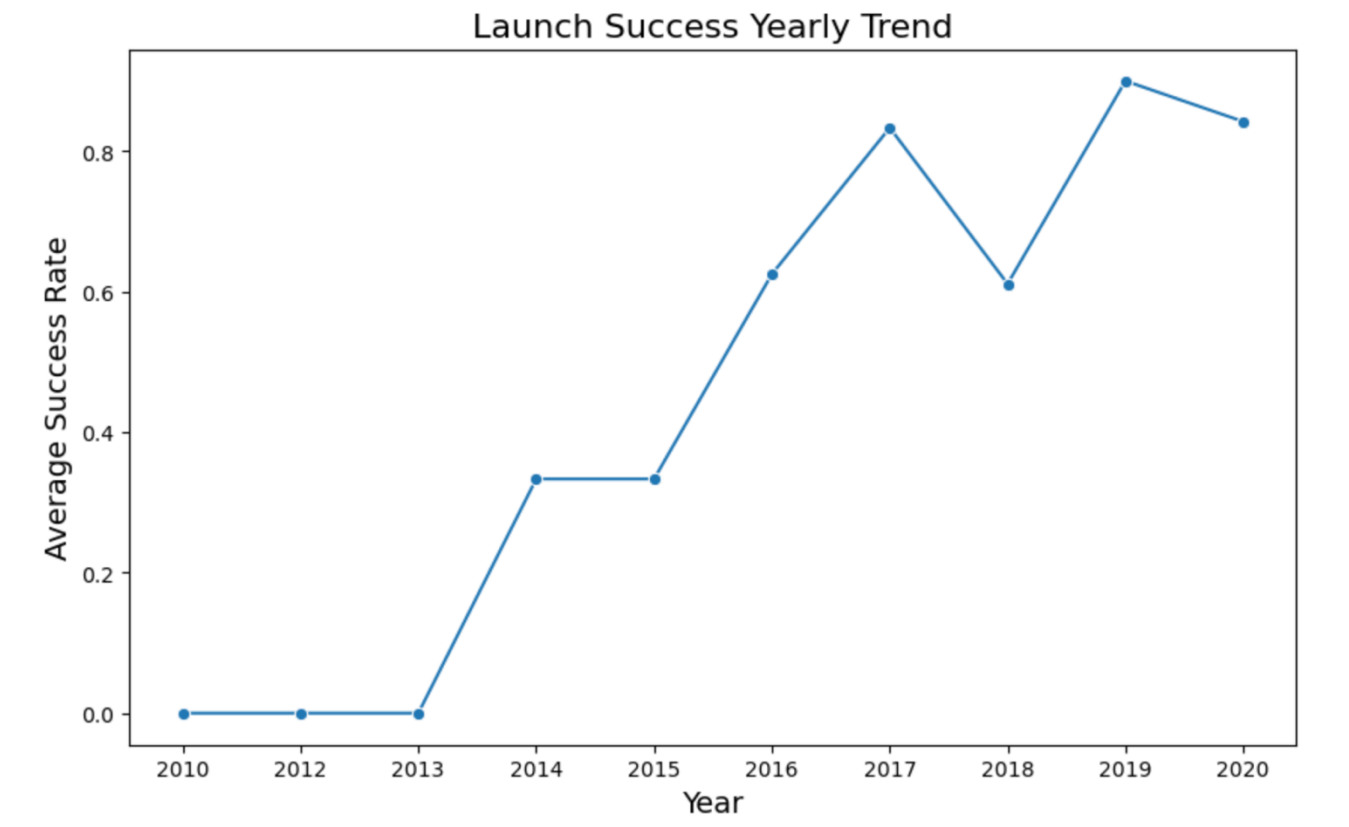


Payload vs. Orbit Type



Heavy payloads tend to lower success rates in GTO orbits, while they are positively correlated with success in Polar LEO (ISS) orbits.

Launch Success Yearly Trend



The success rate steadily increased from 2013 until 2020.

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

Uses a SELECT DISTINCT query on the "Launch_Site" column in SPACEXTABLE to retrieve unique launch site names. This quickly identifies which launch sites are present in the dataset without listing duplicates.

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

Uses a WHERE clause with the LIKE 'CCA%' pattern to filter rows where the "Launch_Site" begins with "CCA". The LIMIT 5 clause then displays only the first five matching records. This quickly inspects which entries match a specific prefix in the launch site names.

Total Payload Mass

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") AS TotalPayloadMass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

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

TotalPayloadMass
45596

- Calculates the total payload mass (in kilograms) for all boosters launched by “NASA (CRS)” by summing the "PAYLOAD_MASS__KG_" column for rows where the "Customer" equals "NASA (CRS)". The result is **45596** kg.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") AS AvgPayloadMass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

AvgPayloadMass
2928.4

Determines the **average payload mass** for all Falcon 9 v1.1 launches by calculating the mean of the "PAYLOAD_MASS__KG_" column for rows where the "Booster_Version" is 'F9 v1.1'. The result is approximately **2928.4 kg**.

First Successful Ground Landing Date

```
%%sql
SELECT MIN("Date") AS FirstGroundPadLanding
FROM SPACEXTABLE
WHERE "Outcome" = 'True RTLS';
```

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

FirstGroundPadLanding
None

This query attempts to find the earliest date (MIN(Date)) where the landing outcome was "True RTLS". However, the result is None, indicating there are no records in the table with a successful Return-To-Launch-Site ("True RTLS") landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT "Serial"
FROM SPACEXTABLE
WHERE "Outcome" = 'True ASDS'
  AND "PAYLOAD_MASS__KG_" > 4000
  AND "PAYLOAD_MASS__KG_" < 6000;
```

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

"Serial"

This query looks for boosters ("Serial") that successfully landed on a drone ship ("True ASDS") with a payload mass between 4000 and 6000 kg. If the query returns rows, it indicates which booster serial numbers met these criteria. If no rows are returned, then there were no such flights matching all conditions.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT
    CASE
      WHEN "Outcome" LIKE 'True%' THEN 'Success'
      ELSE 'Failure'
    END AS LandingOutcome,
    COUNT(*) AS Total
FROM SPACEXTABLE
GROUP BY LandingOutcome;
```

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

LandingOutcome	Total
Failure	101

This query classifies each launch as either “Success” (if "Outcome" starts with 'True') or “Failure” (otherwise), then counts the number of launches in each category. The result shows how many total successes versus failures occurred based on the "Outcome" field.

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT "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

This query identifies which booster versions carried the heaviest payload by comparing each launch's "PAYLOAD_MASS__KG_" to the maximum payload mass found in the table. The result shows all booster versions that reached that maximum payload.

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

```
%%sql
SELECT
    CASE
      WHEN "Outcome" LIKE 'True%' THEN 'Success'
      ELSE 'Failure'
    END AS LandingOutcome,
    COUNT(*) AS OutcomeCount
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LandingOutcome
ORDER BY OutcomeCount DESC;
```

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

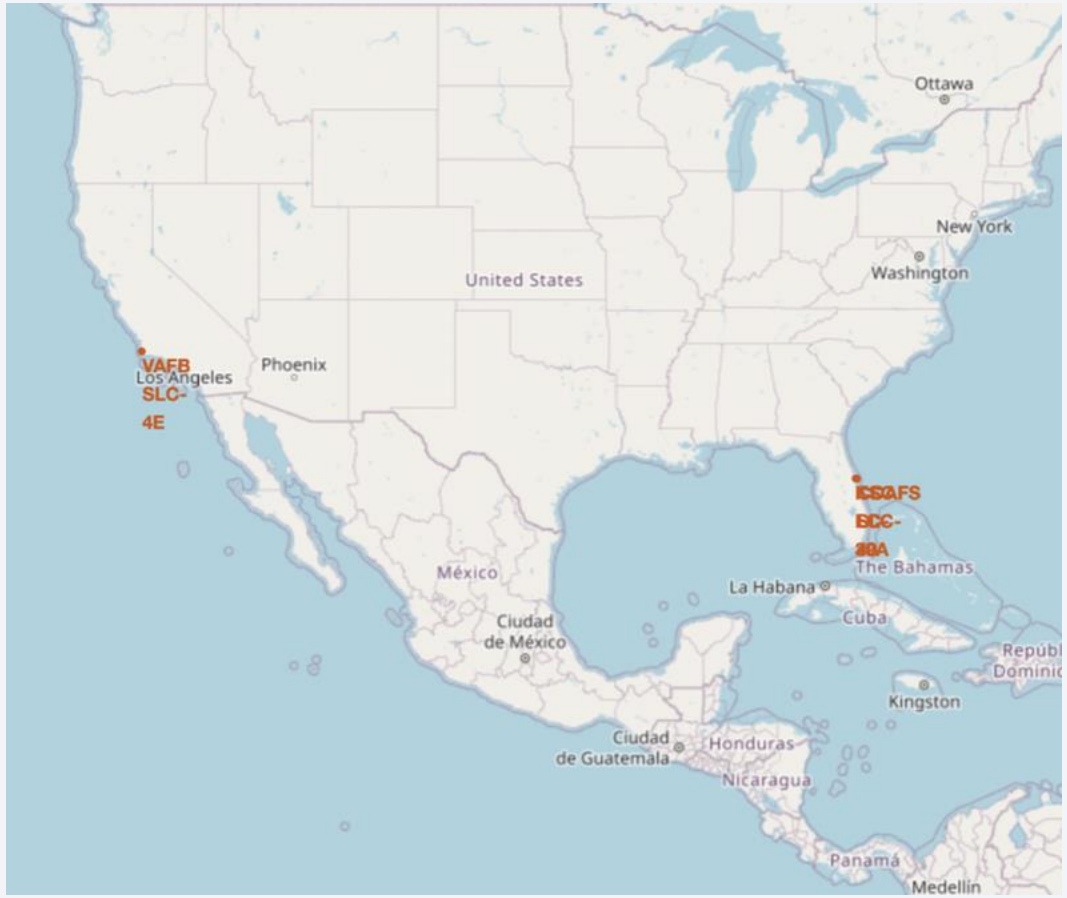
LandingOutcome	OutcomeCount
Failure	31

This query counts how many landings were “Success” or “Failure” between June 4, 2010, and March 20, 2017, then orders them in descending order of occurrence. Here, “Failure” had 31 total, indicating that within that date range, more launches ended unsuccessfully than successfully.

Section 3

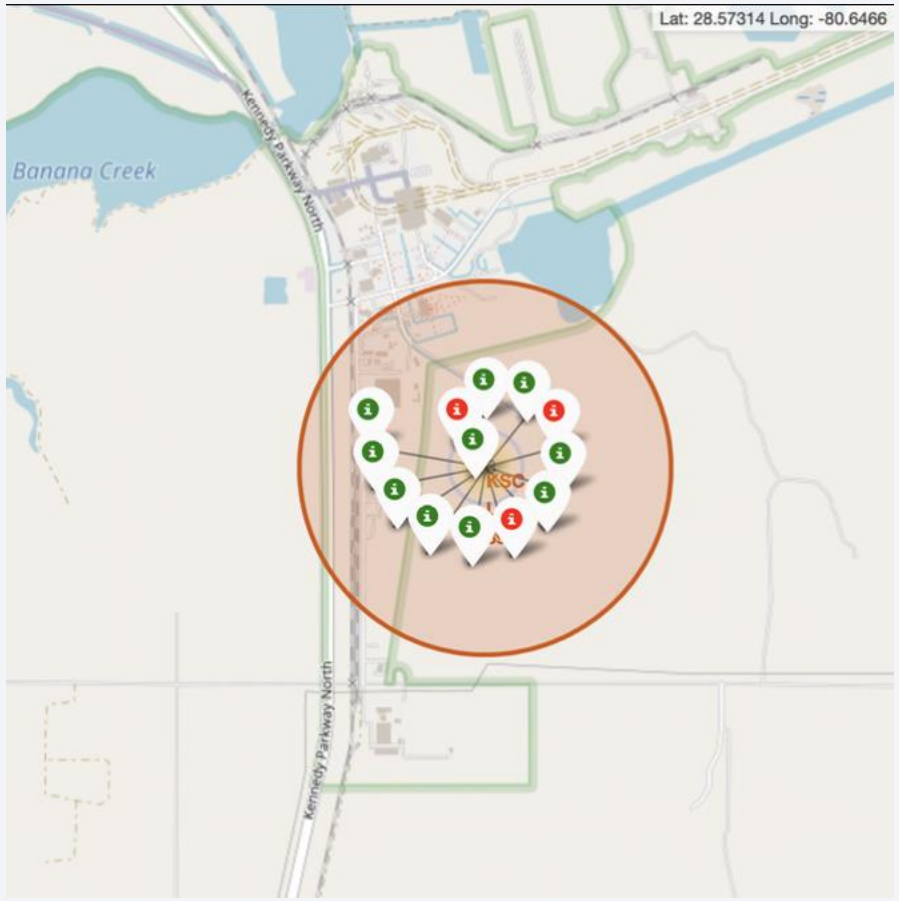
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



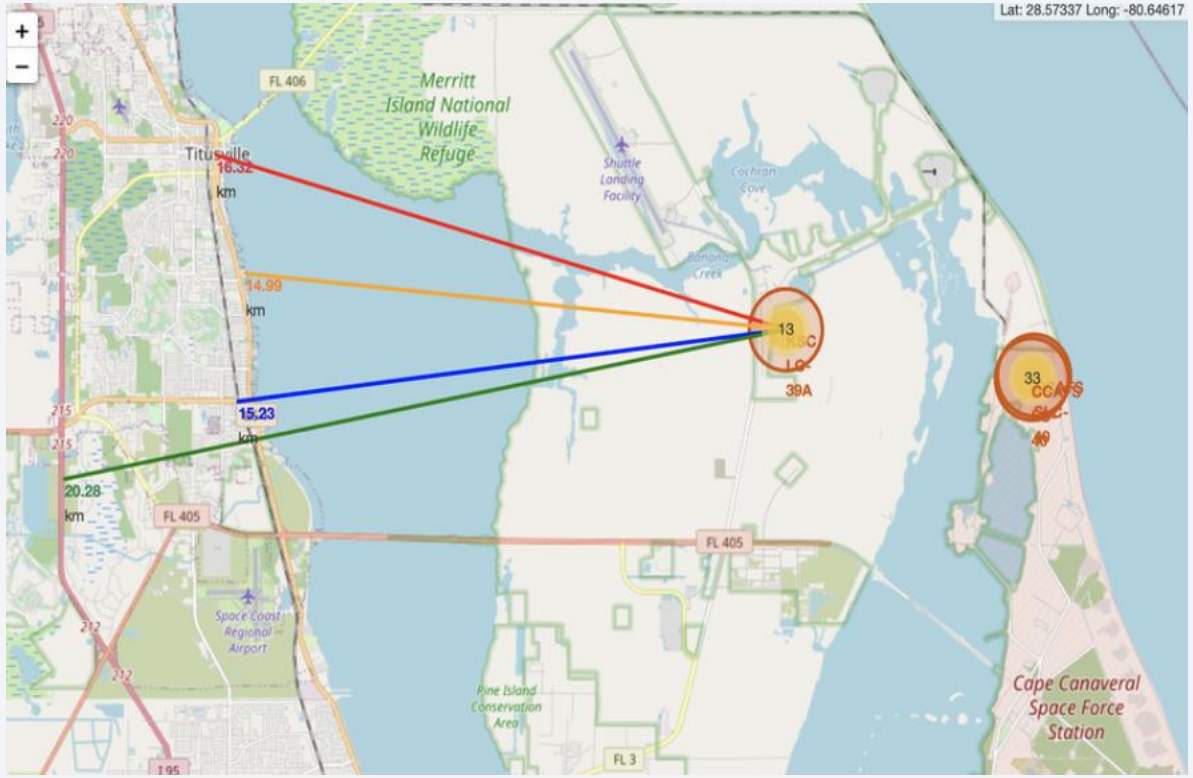
Many launch sites are located near the equator, where Earth's rotation provides an initial velocity boost of about 1670 km/h. This added speed, carried by inertia, helps rockets more easily achieve orbital velocity. Furthermore, positioning these sites along the coast allows launches to occur over open water, reducing the risk to populated areas if debris or a failure event occurs.

<Folium Map Screenshot 2>



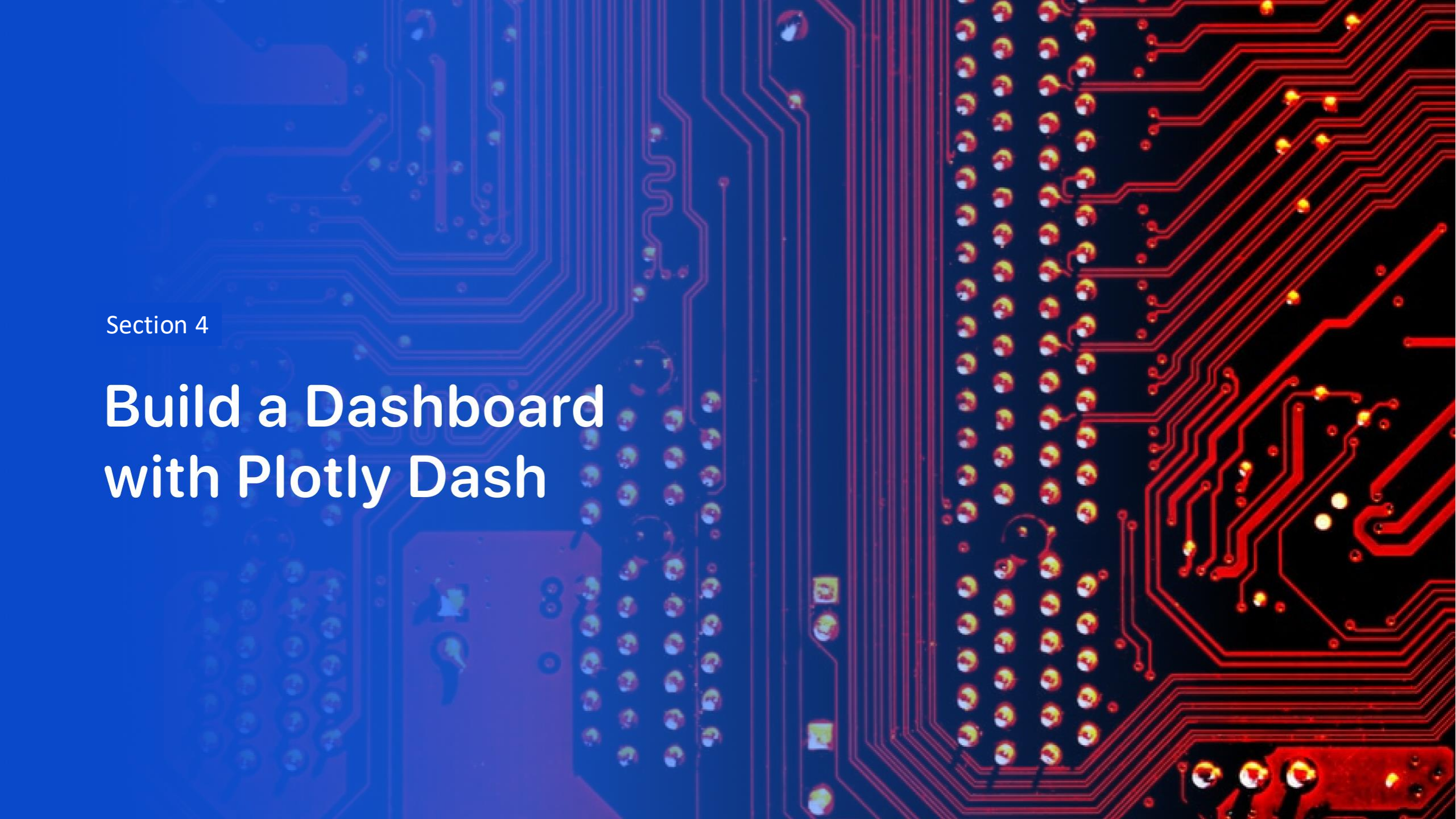
From the color-coded markers, it is straightforward to see which launch sites have higher success rates: green markers represent successful launches, while red markers represent failures. Notably, KSC LC-39A shows a particularly high success rate.

<Folium Map Screenshot 3>

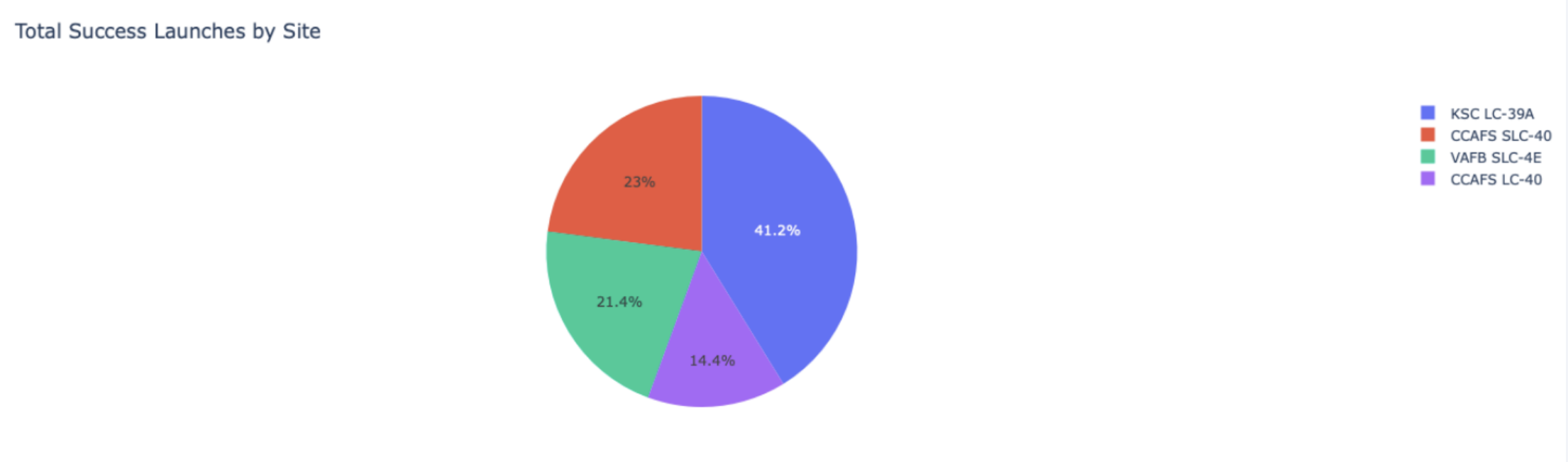


Analysis of KSC LC-39A shows that:

- It is located approximately 15.23 km from a railway, 20.28 km from a highway, and 14.99 km from the coastline.
- It is also about 16.32 km from the nearest city, Titusville.
- Given that a failed rocket can travel 15–20 km in seconds, its proximity to populated areas poses potential risks.

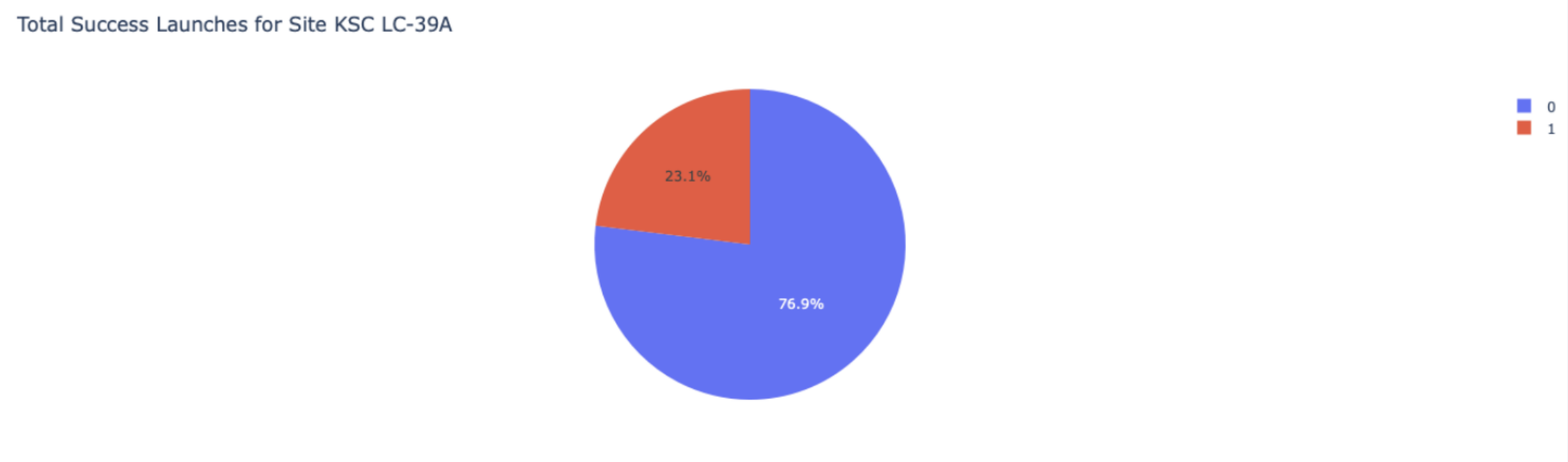

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

<Dashboard Screenshot 2>



- KSC LC-39A achieved a 76.9% success rate, with 10 successful landings compared to only 3 failures.

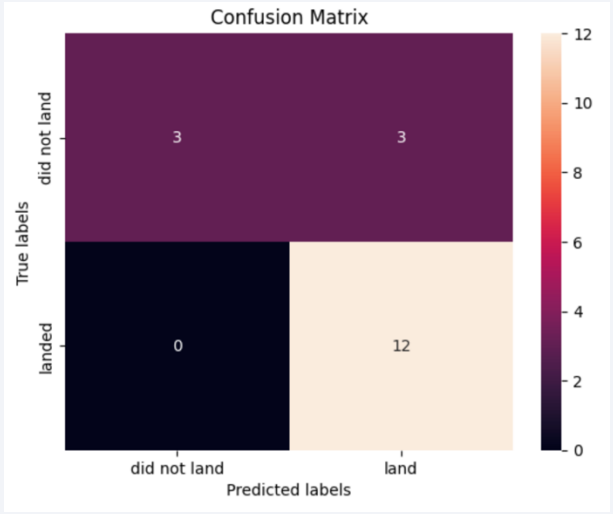
<Dashboard Screenshot 3>

- The charts indicate that payloads in the 2000–5500 kg range achieve the highest success rate.

Section 5

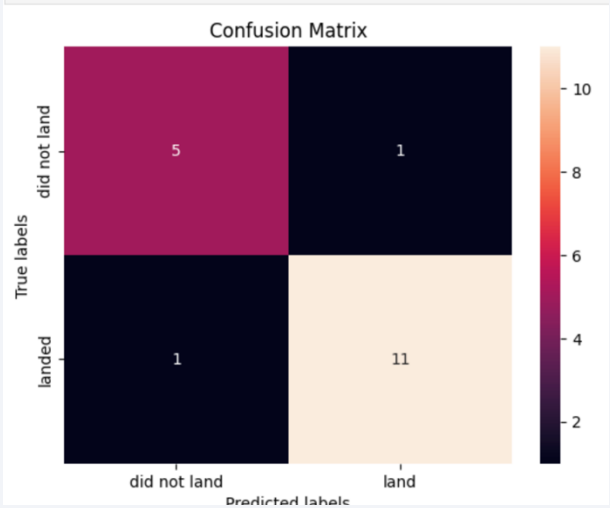
Predictive Analysis (Classification)

Classification Accuracy



The model correctly predicts “did not land” 3 times and “land” 12 times, but incorrectly labels 3 “did not land” flights as “land.” It never mistakes “land” flights as “did not land.”

The model correctly predicts “did not land” 5 times and “land” 11 times, with only 2 total mistakes (1 false positive, 1 false negative). Overall, it has fewer errors than Matrix 1.



Confusion Matrix

```
]: # TASK 11: Calculate the accuracy of knn_cv on the test data using the score method
   test_accuracy_knn = knn_cv.score(X_test, Y_test)
   print("Test Accuracy:", test_accuracy_knn)
```

```
Test Accuracy: 0.8333333333333334
```

We can plot the confusion matrix

```
]: yhat = knn_cv.predict(X_test)
   plot_confusion_matrix(Y_test,yhat)
```

Among Logistic Regression, SVM, KNN, and Decision Tree, the Decision Tree model showed the best overall accuracy, outperforming others on the test set. This suggests that Decision Tree is most effective for predicting booster landing outcomes under the given data and conditions.

Conclusions

- **Best Model:** The Decision Tree model outperformed others for predicting booster landing outcomes.
- **Payload Influence:** Launches with lower payload mass tend to achieve higher success rates.
- **Launch Site Location:** Most sites are near the Equator and coastlines, which aids in achieving orbit and minimizing risks.
- **Trend Over Time:** Launch success rates have steadily increased over the years.
- **Top Site:** KSC LC-39A shows the highest success rate among all launch sites.
- **Orbital Performance:** Orbits ES-L1, GEO, HEO, and SSO demonstrate a 100% success rate.

Thank you!

Daniel Ballesteros
16/03/2025