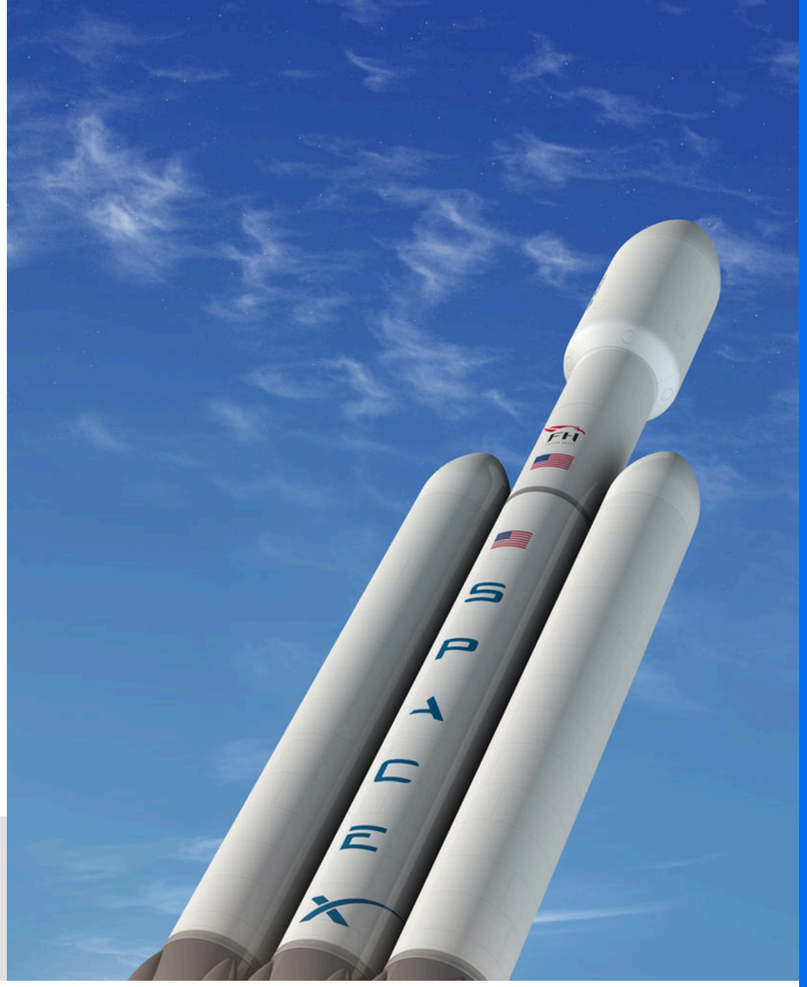
DATA SCIENCE CAPSTONE PROJECT



By Truc Ho

15.5.2025

REPORT OUTLINE

TOPIC HIGHLIGHTS

1. EXECUTIVE SUMMARY

SUMMARY OF METHODDOLOGIES

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

SUMMARY OF RESULTS

- Exploratory Data Analysis results
- Interactive analytics
- Predictive analysis results

2. INTRODUCTION

🔎 BACKGROUND AND PROBLEM STATEMENT

The rise of commercial space companies has made space travel more accessible. **SpaceX** leads this effort by significantly cutting launch costs, thanks to its ability to **reuse the first stage** of the Falcon 9 rocket. While a Falcon 9 launch costs about **62 million USD**, competitors may charge over **165 million USD** due to non-reusability.

Predicting whether the **first stage will land successfully** is key to estimating launch costs. In this project, acting as data scientists for a new company, **SpaceY**, we will use **public data and machine learning** to forecast first stage recovery outcomes.

2. INTRODUCTION

? KEY QUESTIONS TO EXPLORE

- How do factors like **payload, launch site, orbit, and flight number** impact landing success?
- Has the **success rate** improved over time?
- What is the **best algorithm** for binary classification in this scenario?

3. METHODOLOGY

01 Data collection: using SpaceX Rest API and Web Scrapping from Wikipedia

02 Data Wrangling: filtering data, handle missing values, using One Hot Encoding

03 Exploratory Data Analysis: using visualization and SQL

04 Interactive visual analytics: using Folium, Plotly Dash

05 Built and optimized classification models for predictive analysis.

3. METHODOLOGY

DATA COLLECTION

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from table in SpaceX's Wikipedia entry.

SPACEX REST API

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

WIKIPEDIA WEB SCRAPING

Data Columns are obtained:
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

3. METHODOLOGY

SPACEX REST API

DATA COLLECTION

Send API Request
Retrieve rocket launch data from SpaceX API

Convert JSON to DataFrame
Use .json() and json_normalize().

Extract Required Information
Apply custom functions to get specific data.

Build a Dictionary
Organize the extracted data into a dictionary.

Export to CSV
Save the cleaned data to a CSV file.

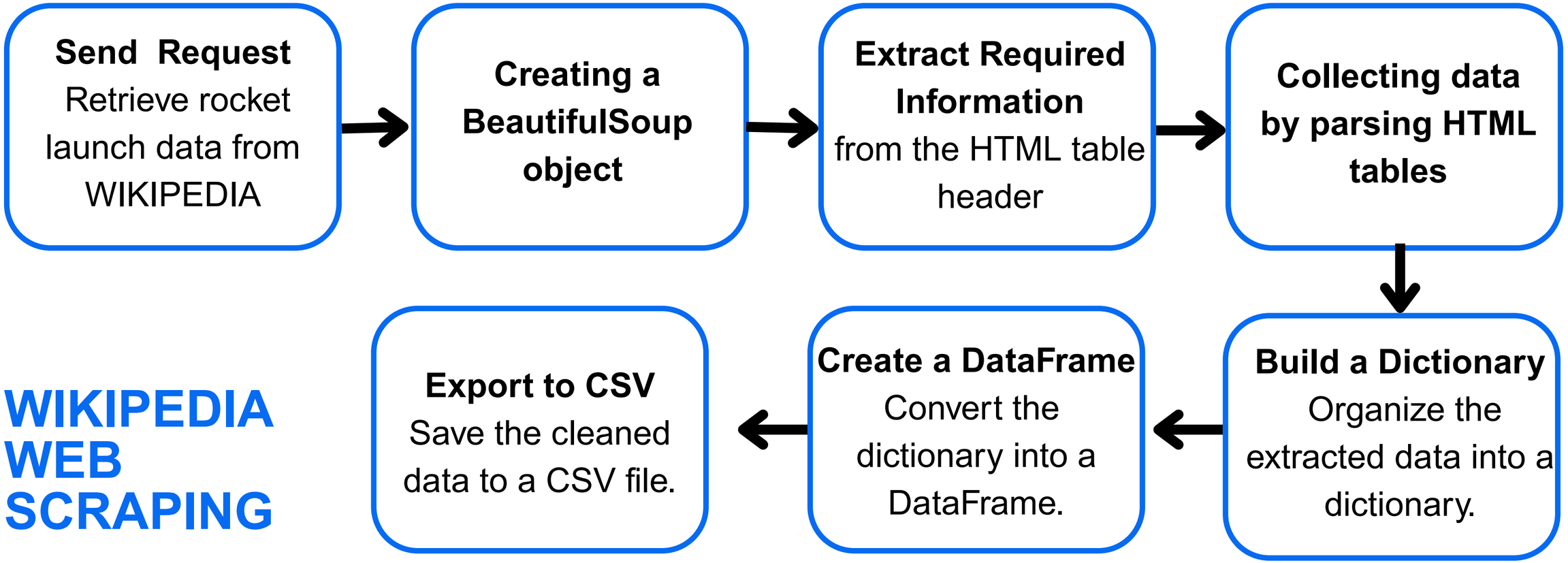
Handle Missing Values
Replace missing values with the column mean

Filter for Falcon 9 Launches
Keep only Falcon 9 launch records.

Create a DataFrame
Convert the dictionary into a DataFrame.

3. METHODOLOGY

DATA COLLECTION



WIKIPEDIA WEB SCRAPING

3. METHODOLOGY

DATA WRANGLING

Understand Landing Outcomes

Outcomes like True/False Ocean, True/False RTLS, and True/False ASDS indicate whether the booster landed successfully or failed (in ocean, on land, or on a drone ship).

Label the Data

Convert outcomes into training labels:

- "1" = successful landing
- "0" = unsuccessful landing

Data Processing Steps

- Perform EDA and define labels
- Count launches per site
- Count and analyze orbits
- Analyze outcomes by orbit type
- Create outcome labels from the original data
- Export the cleaned dataset to CSV

3. METHODOLOGY

EDA WITH DATA VISUALIZATION

Charts created:

- Flight Number vs. Payload Mass / Launch Site / Orbit Type
- Payload Mass vs. Launch Site / Orbit Type
- Orbit Type vs. Success Rate
- Success Rate Trend by Year

Chart types & purposes:

- **Scatter Plots** → Show relationships between numerical variables (used for model input analysis)
- **Bar Charts** → Compare categories (e.g., launch sites or orbit types)
- **Line Charts** → Show trends over time (e.g., yearly success rate)

3. METHODOLOGY

Key SQL Queries Performed:

Launch Sites:

- List unique launch sites
- Filter launch sites starting with 'CCA'

Payload Analysis:

- Total payload mass by NASA launches
- Average payload mass for booster version F9 v1.1
- Boosters on drone ships with payload between 4000–6000
- Booster versions with maximum payload mass

Landing Outcomes:

- First successful ground pad landing date
- Count of successful vs. failed missions
- Failed drone ship landings in 2015 with booster & site details
- Ranked outcomes (Success/Failure) between 2010-06-04 and 2017-03-20

3. METHODOLOGY

BUILD AN INTERACTIVE MAP WITH FOLIUM

1. Launch Site Markers:

Placed circular markers with popup and text labels using latitude/longitude for:

- NASA Johnson Space Center (start location)
- All launch sites (to visualize geographic spread and proximity to coasts/equator)

2. Launch Outcome Markers:

- Used green markers for successful and red markers for failed launches
- Applied Marker Cluster to group markers and identify high-success launch sites

3. Distance Visualization:

Drew colored lines to show distances from a launch site (e.g., KSC LC-39A) to nearby:

- Railways, Highways, Coastlines, and Closest Cities

3. METHODOLOGY

BUILD A DASHBOARD WITH PLOTLY DASH

1. Launch Sites Dropdown List:
- Enables users to filter data by selecting specific launch sites.

2. Pie Chart (Launch Success Visualization):
- Shows total success vs. failure launches for: All launch sites, a selected launch site (if chosen from dropdown)

3. Payload Mass Range Slider:
- Allows selection of a payload mass range to filter data dynamically.

4. Scatter Chart (Payload vs. Success Rate by Booster Version):
- Visualizes correlation between payload mass and launch success, broken down by booster version.

3. METHODOLOGY

1. Prepare the Data

- Extract the Class column into a NumPy array (used as the target variable).

2. Preprocess the Features

- Standardize the features using StandardScaler to normalize the data.

3. Split the Dataset

- Use train_test_split to divide the data into training and testing sets.

4. Model Selection & Hyperparameter Tuning

- Create a GridSearchCV object with cross-validation (cv=10) to find the best model parameters.

3. METHODOLOGY

5. Model Training

Apply GridSearchCV on:

- Logistic Regression (LogReg)
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

6. Model Evaluation

Evaluate model performance by:

- Computing accuracy on test data using .score()
- Analyzing the confusion matrix
- Comparing Jaccard Score and F1 Score

4. RESULTS

Understand Landing Outcomes

Outcomes like True/False Ocean, True/False RTLS, and True/False ASDS indicate whether the booster landed successfully or failed (in ocean, on land, or on a drone ship).

Label the Data

Convert outcomes into training labels:

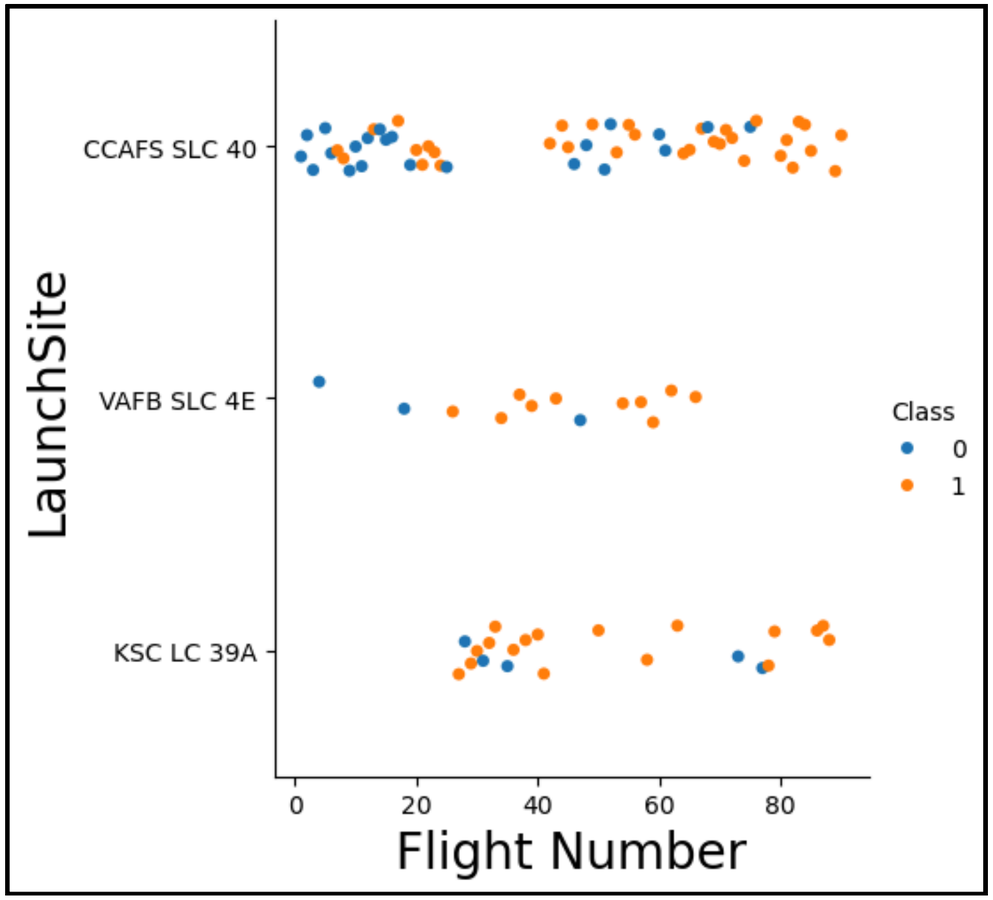
- "1" = successful landing
- "0" = unsuccessful landing

Data Processing Steps

- Perform EDA and define labels
- Count launches per site
- Count and analyze orbits
- Analyze outcomes by orbit type
- Create outcome labels from the original data
- Export the cleaned dataset to CSV

4. RESULTS

EDA WITH VISUALIZE



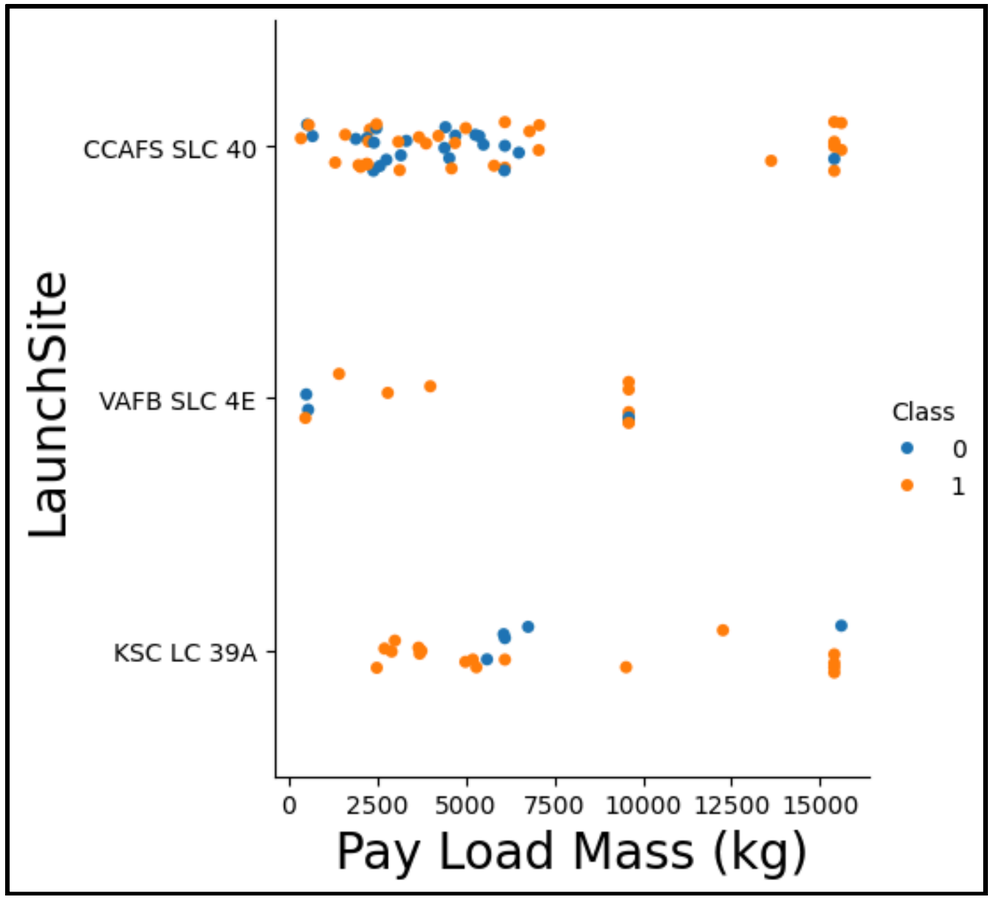
FLIGHT NUMBER AND LAUNCH SITE

From the scatter plot, we observe that:

- Most launches from CCAFS SLC 40 and KSC LC 39A resulted in successful missions (Class 1), especially as the flight number increases, indicating improved reliability over time.
- Launches from VAFB SLC 4E show a mix of successes and failures, with fewer overall flights compared to the other sites.
- There is a noticeable trend where higher flight numbers tend to be associated with more successes, suggesting learning or technological improvements over time.

4. RESULTS

EDA WITH VISUALIZE

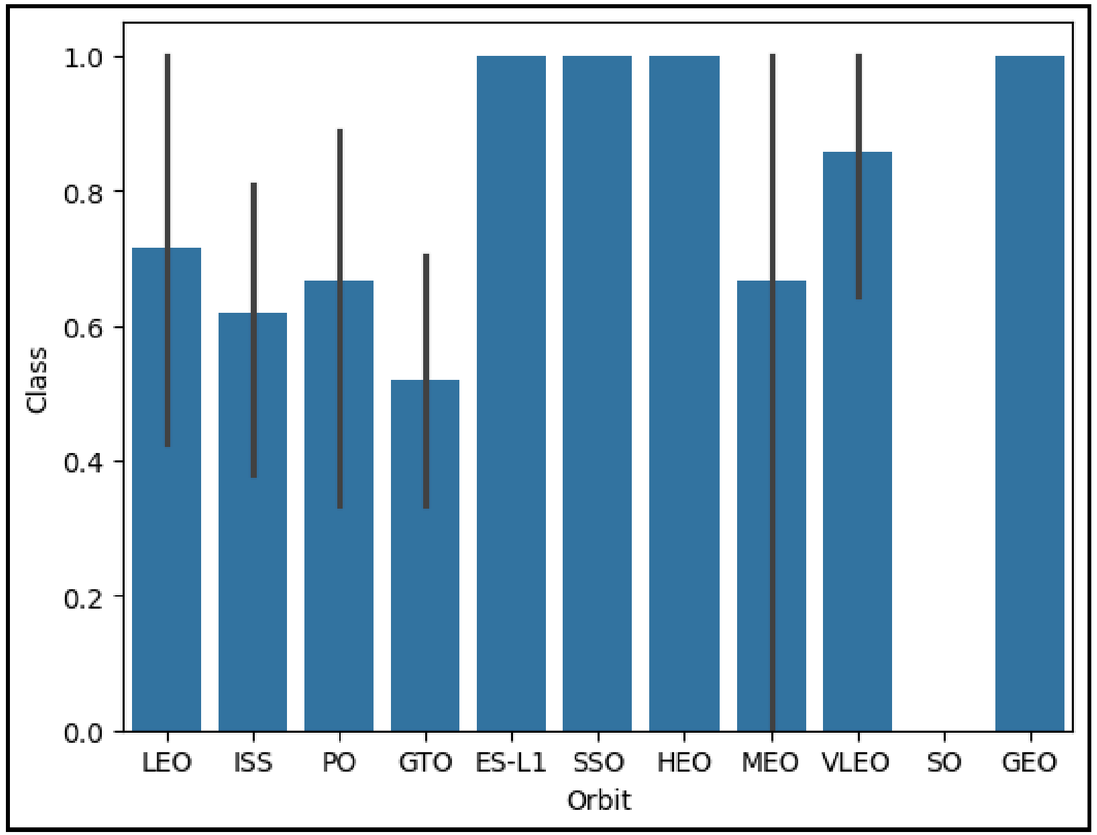


FLIGHT NUMBER AND PAY LOAD MAX

The plot shows that CCAFS SLC 40 had the most launches with high success rates. KSC LC 39A handled the heaviest payloads, mostly with successful outcomes. VAFB SLC 4E had fewer launches with mixed results. Heavier payloads were generally launched successfully.

4. RESULTS

EDA WITH VISUALIZE

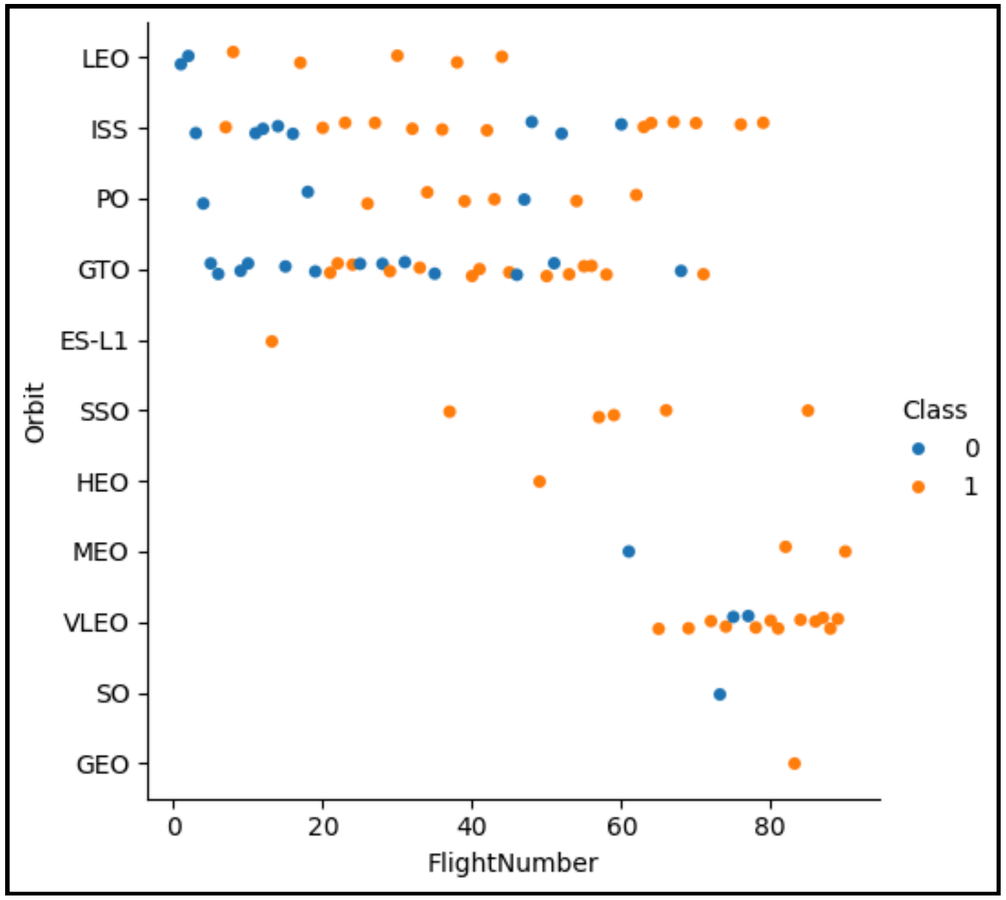


SUCCESS RATE OF EACH ORBIT TYPE

The chart shows that SSO, ES-L1, HEO, and GEO orbits had the highest success rates (100%), while GTO had the lowest. Most orbit types had high success, but several had large error bars, indicating variability in outcomes.

4. RESULTS

EDA WITH VISUALIZE

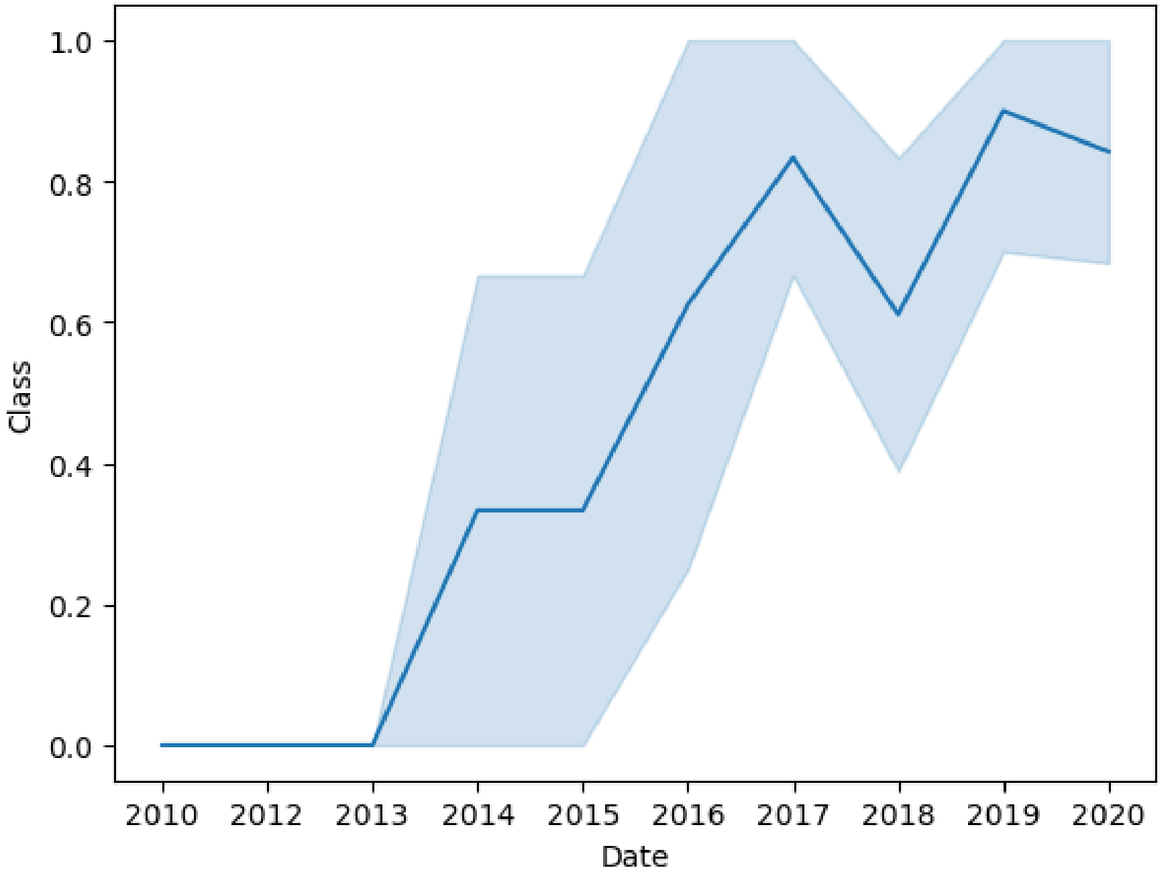


FLIGHTNUMBER AND ORBIT TYPE

Most orbits were used throughout the flight history. High success (orange) is observed for SSO, ES-L1, and GEO. Failures (blue) are more common in early flights and for orbits like LEO and GTO.

4. RESULTS

EDA WITH VISUALIZE



SUCCESS YEARLY TREND

The graph shows a rising success trend from 2010 to 2020, with a sharp increase starting in 2013 and peaking around 2019. Fluctuations indicate variability, but the overall trajectory is upward, suggesting significant progress over the years

4. RESULTS


```
%sql select DISTINCT(Landing_Outcome) as unique_launch from SPACEXTBL
```

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

unique_launch
Failure (parachute)
No attempt
Uncontrolled (ocean)
Controlled (ocean)
Failure (drone ship)
Precluded (drone ship)
Success (ground pad)
Success (drone ship)
Success
Failure
No attempt

ALL LAUNCH SITE NAMES

4. RESULTS

EDA WITH SQL

```
%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	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

DISPLAY 5 RECORDS WHERE LAUNCH SITES BEGIN WITH THE STRING 'CCA'

4. RESULTS

EDA WITH SQL

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

DISPLAY THE TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)

4. RESULTS

EDA WITH SQL

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

DISPLAY THE TOTAL PAYLOAD MASS CARRIED BY BOOSTERS LAUNCHED BY NASA (CRS)

4. RESULTS

EDA WITH SQL

```
%sql select AVG(PAYLOAD_MASS__KG_) as  `average payload mass` from SPACEXTBL where Booster_Version LIKE 'F9 v1.1%'
```

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

average payload mass
2534.6666666666665

DISPLAY AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9 V1.1

4. RESULTS

EDA WITH SQL

```
%sql select MIN(Date) as `first succesful landing outcome in ground pad was acheived` from SPACEXTBL where Landing_Outcome == 'Success (ground pad)'
```

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

first succesful landing outcome in ground pad was acheived
2015-12-22

THE FIRST SUCCESFUL LANDING OUTCOME IN GROUND PAD WAS ACHEIVED.

4. RESULTS

EDA WITH SQL

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome = 'Precluded (drone ship)' and 4000<PAYLOAD_MASS__KG_<6000
```

Python

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

Booster_Version
F9 v1.1 B1018

THE BOOSTERS WHICH HAVE SUCCESS IN DRONE SHIP AND HAVE PAYLOAD MASS GREATER THAN 4000 BUT LESS THAN 6000

4. RESULTS

EDA WITH SQL

```
%sql select DISTINCT(Mission_Outcome), COUNT(*) from SPACEXTBL GROUP BY Mission_Outcome
```

✓ 0.0s Python

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

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

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

4. RESULTS


```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

✓ 0.0s

```
 * 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

ALL THE BOOSTER_VERSIONS THAT HAVE CARRIED THE MAXIMUM PAYLOAD MASS. USE A SUBQUERY.

4. RESULTS

EDA WITH SQL

```
%sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTBL where substr(Date,0,5)='2015' and Landing_Outcome  = 'Failure (drone ship)'
```

✓ 0.0s

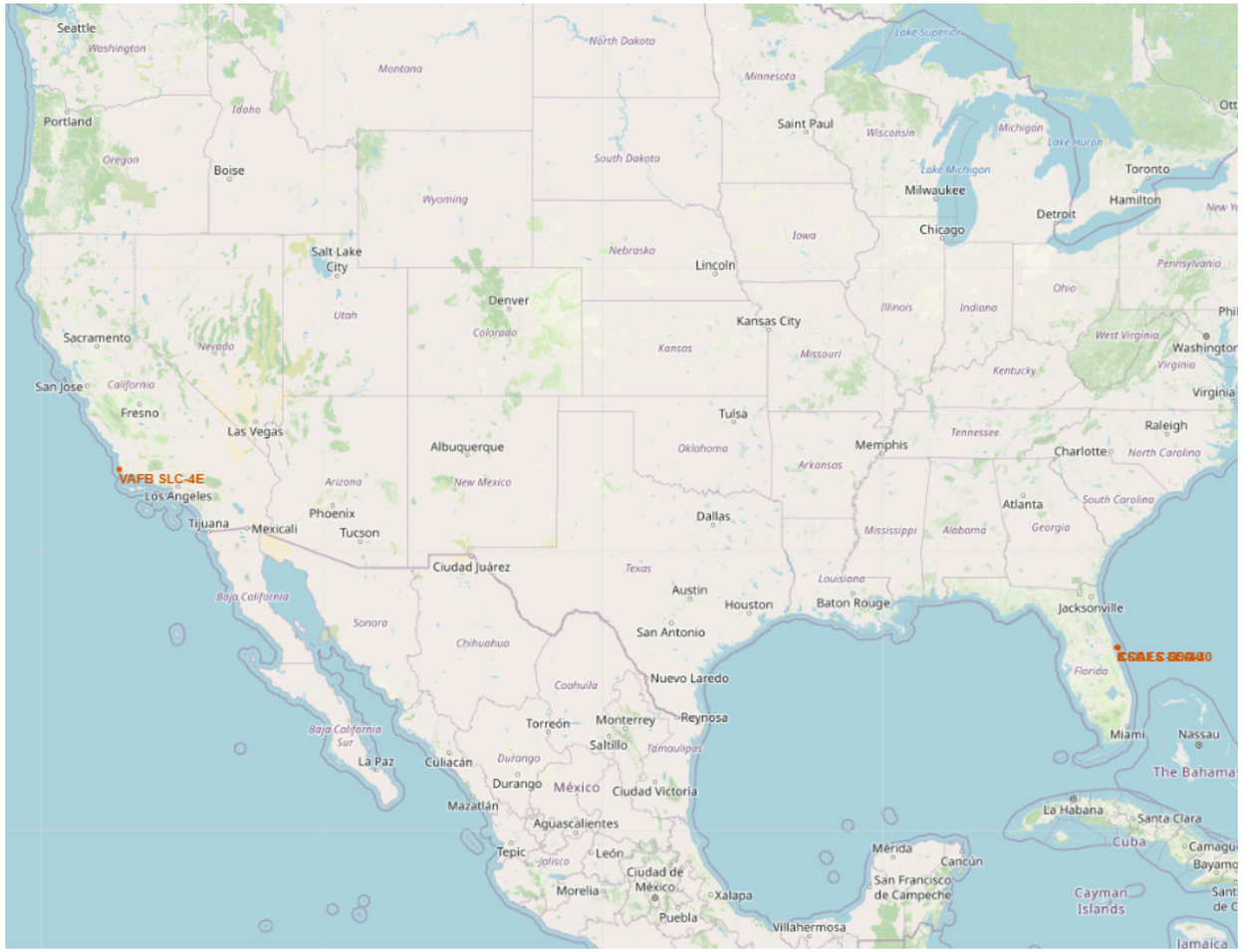
```
 * sqlite:///my_data1.db
Done.
```

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

THE RECORDS WHICH WILL DISPLAY THE MONTH NAMES, FAILURE LANDING_OUTCOMES IN DRONE SHIP ,BOOSTER VERSIONS, LAUNCH_SITE FOR THE MONTHS IN YEAR 2015.

4. RESULTS

INTERACTIVE MAP WITH FOLIUM



Most launch sites are near the equator because the Earth's rotation gives rockets an extra speed boost of 1670 km/h due to inertia, helping them achieve orbit more efficiently. Additionally, launch sites are usually close to the coast to reduce the risk of debris falling in populated areas.

4. RESULTS

INTERACTIVE MAP WITH FOLIUM

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

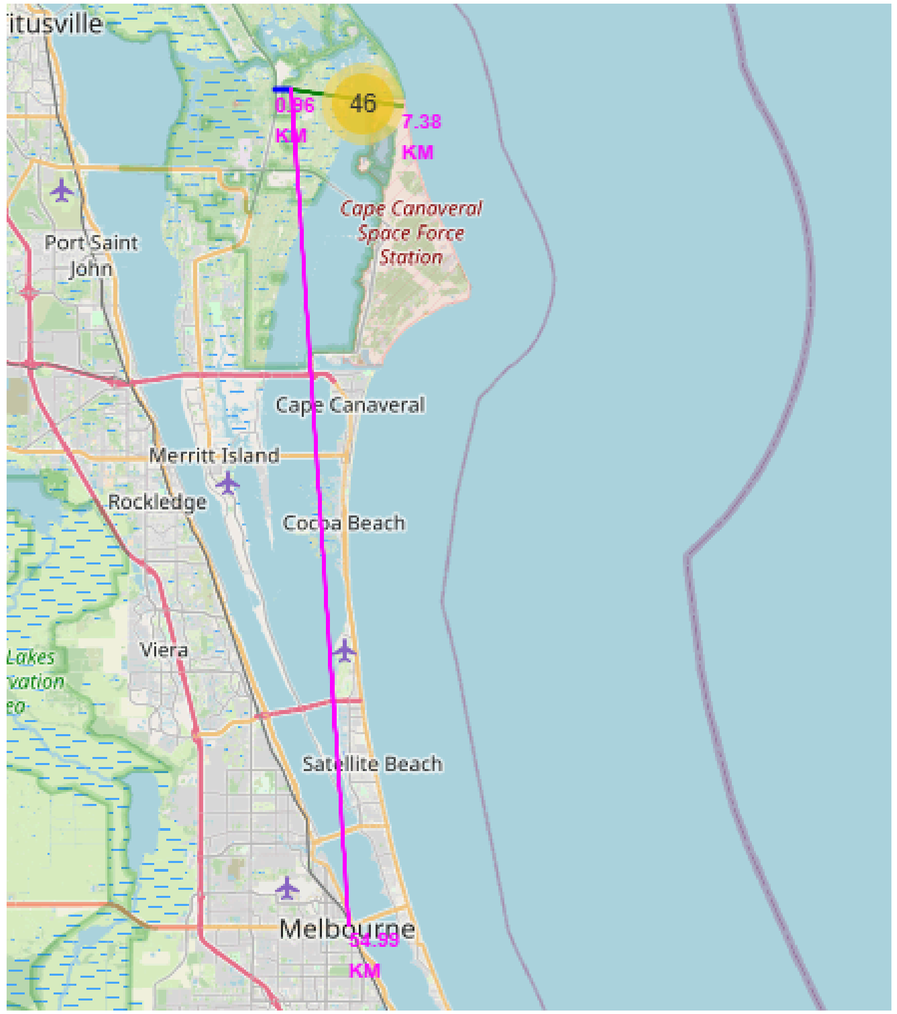
- **Green Marker** = Successful Launch

- **Red Marker** = Failed Launch

• Launch Site KSC LC-39A has a very high Success Rate.

4. RESULTS

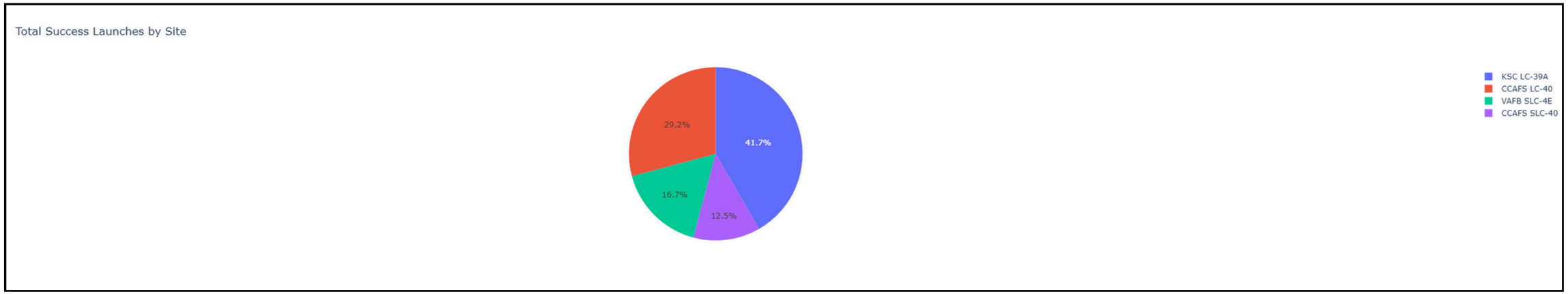
INTERACTIVE MAP WITH FOLIUM



The launch site, marked in orange on the map, is located at Cape Canaveral Space Force Station. It is positioned northeast of Melbourne, approximately 46 km away. The site is also near the coastline, with a distance of about 7.38 km offshore indicated on the map. This strategic location allows launches to occur over the ocean, minimizing risk to populated areas.

4. RESULTS

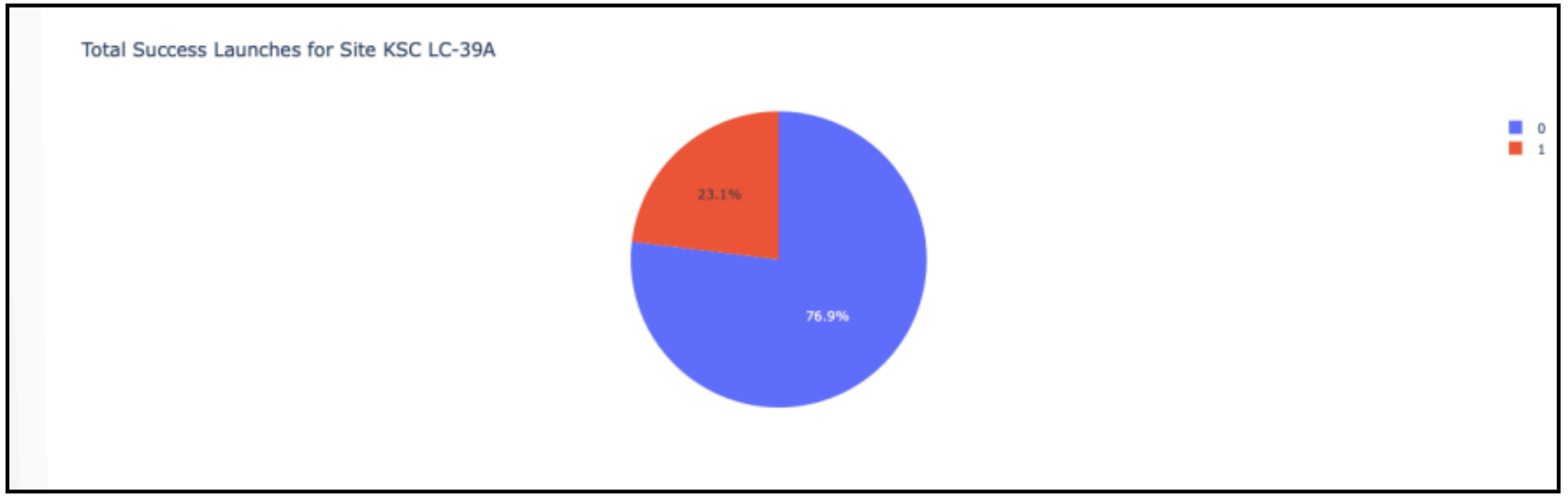
DASHBOARD WITH PLOTY DASH



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

4. RESULTS

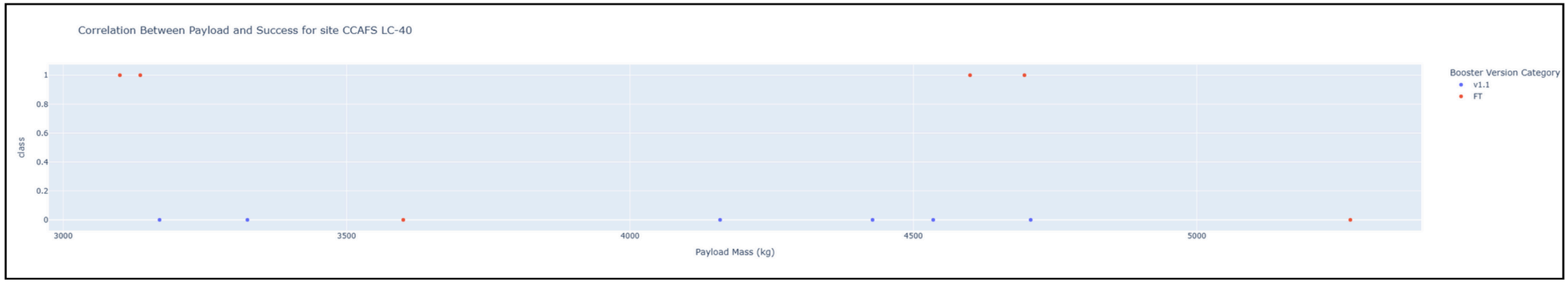
DASHBOARD WITH PLOTY DASH

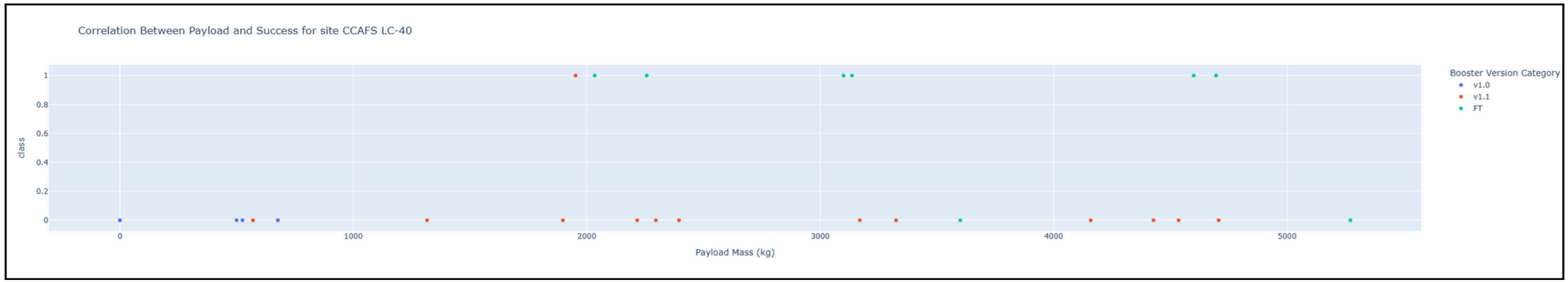


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings

4. RESULTS

DASHBOARD WITH PLOTY DASH





4. RESULTS

PREDICTIVE ANALYSIS

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

SCORES AND ACCURACY OF THE TEST SET

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

SCORES OF PREDICTION ON WHOLE DATASET

- The tables compare four machine learning models—Logistic Regression, SVM, Decision Tree, and KNN—based on Jaccard Score, F1 Score, and Accuracy.
- In test results, all models achieve 83.33% accuracy. However, Decision Tree performs best on the full dataset with 91.11% accuracy, followed by SVM (87.78%), Logistic Regression (86.67%), and KNN (85.56%).
- Decision Tree shows the highest overall performance, but consistency across different datasets should be considered before selecting the best model.

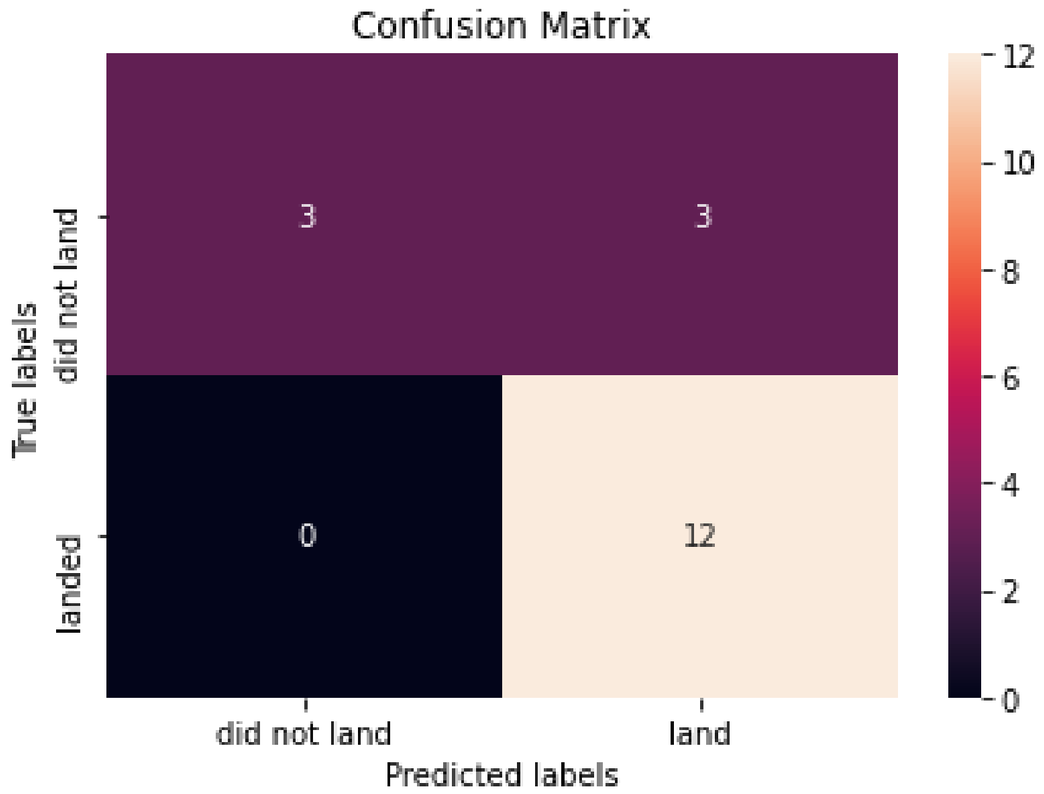
4. RESULTS

CONFUTION MATRIX

Confusion Matrix

			Ground Truth Label	
		Total Observations (n)	*has disease*	*no disease*
			Condition Positive (CP)	Condition Negative (CN)
Predicted Label	*test positive*	Test Outcome Positive (TOP)	True Positive (TP)	False Positive (FP)
	test negative	Test Outcome Negative (TON)	False Negative (FN)	True Negative (TN)

Figure 1: Basic colour coded confusion matrix with marginal sums



Confution matrix of decision tree

5. CONCLUSION

- The Decision Tree model demonstrates the highest performance for this dataset.
- Launches with lighter payloads tend to yield better results compared to those with heavier payloads. Most launch sites are situated near the Equator and are positioned very close to the coastline, likely optimizing launch conditions.
- Over time, the overall success rate of launches has improved. Among all sites, KSC LC-39A has the highest success rate. Additionally, certain orbital destinations—ES-L1, GEO, HEO, and SSO—achieve a perfect 100% success rate.