IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Federico Ferrero
April 2, 2025

Outline

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

Executive Summary

- **Summary of methodologies**
  - **Data Collection & Wrangling:** Retrieved launch data via SpaceX API, converted JSON to Pandas DataFrame, filtered Falcon 9 launches, handled missing values, and ensured data integrity.
  - **Exploratory Data Analysis (EDA) & Visualization:** Analyzed launch counts, orbit occurrences, mission outcomes, and created a binary landing_class label; utilized SQL, Folium, and Plotly Dash for interactive visualizations.
  - **Predictive Modeling:** Built, tuned, and evaluated classification models (Logistic Regression, SVM, Decision Tree, KNN) to predict Falcon 9 first-stage landing success.
- **Summary of all results**
  - **Insights drawn from EDA**
  - **Launch Sites Proximities Analysis**
  - **Dashboard with Plotly Dash**
  - **Predictive Analysis (Classification)**

Introduction

- ## Project Background and Context
  - The cost of rocket launches varies significantly across providers, with SpaceX offering Falcon 9 launches at $62 million, while competitors charge upwards of $165 million.
  - A key factor in SpaceX's cost advantage is the reusability of the first stage.
  - Accurately predicting whether the first stage will successfully land is crucial for estimating launch costs. This insight can be valuable for companies looking to compete with SpaceX in the commercial space launch market.

- ## Research Problem
  - The goal is to predict the successful landing of the Falcon 9 first stage using data from past rocket launches advertised on SpaceX's website.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected launch data via SpaceX API, converted JSON to a Pandas DataFrame, filtered for Falcon 9 launches, handled missing values, and ensured data integrity for analysis.
- Perform data wrangling
  - Performed EDA by analyzing launch counts, orbit occurrences, and mission outcomes, as well as created a binary landing_class label for supervised learning.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - Built, tuned, and evaluated classification models (Logistic Regression, SVM, Decision Tree, and KNN) to better predict the successful landing of the Falcon 9 first stage.

Data Collection

- **API Data Retrieval**
  - Used requests library to make a GET request to the SpaceX API.
  - Retrieved historical launch data in JSON format.
- **Data Processing & Wrangling**
  - Converted JSON response into a Pandas DataFrame using pd.json_normalize().
  - Filtered the dataset to include only Falcon 9 launches.
  - Handled missing values to ensure data consistency.
- **Final Data Preparation**
  - Cleaned and formatted the dataset for analysis.
  - Verified data integrity before modeling.

Data Collection – SpaceX API

- Access the completed SpaceX API calls notebook on GitHub at this link:

https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%201%20-%20Collecting%20Data%20Ferrero.ipynb

Start

↓

Import Libraries (requests, pandas, numpy, datetime)

↓

Send GET request to SpaceX API

↓

Parse JSON response
(json() method)

↓

Convert to Pandas DataFrame
(pd.json_normalize())

↓

Filter Data (Falcon 9)

↓

Handle Missing Data

↓

Fina and Control Dataset

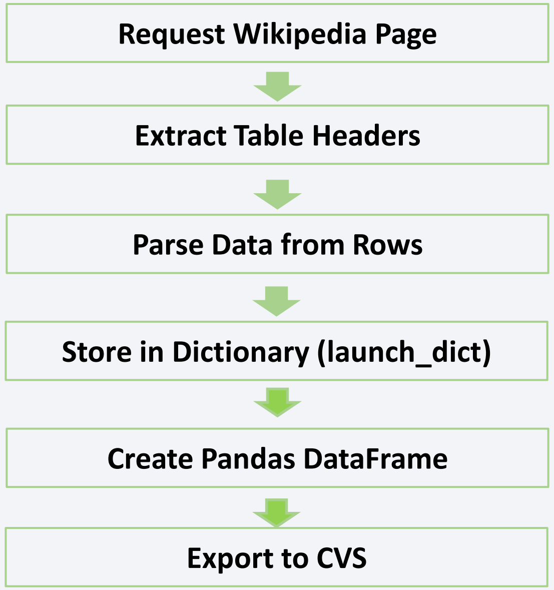
↓

Ready for Analysis

Data Collection - Scraping

Objective: Collect Falcon 9 historical launch records from the Wikipedia page List of Falcon 9 and Falcon Heavy launches using BeautifulSoup.

1. Request Webpage:
   - Use the URL to fetch the Wikipedia page (Snapshot: June 9, 2021).
   - Use requests.get() to send the request.
2. Extract Table Headers:
   - Parse the HTML table header to get column/variable names.
   - Use BeautifulSoup to identify the <th> tags in the table.
3. Parse the Table Data:
   - Extract data from each row of the launch records table.
   - Identify and extract the values for each launch record.
4. Store Data in a Dictionary:
   - Store parsed values in a dictionary (launch_dict).
   - Map each column/variable name to the corresponding data.
5. Create Pandas DataFrame and Export to CSV



- Access the GitHub URL of the completed web scraping notebook here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%202%20-%20Webscraping%20Ferrero.ipynb

Data Wrangling

- **Objective:** Perform EDA to find patterns and create labels for supervised model training.

  1. **Import Libraries & Define Functions**
     - Import pandas, numpy.
  2. **Launch Counts per Site**
     - Use value_counts() to get the number of launches per site.
  3. **Orbit Occurrence**
     - Use value_counts() to calculate the frequency of each orbit type.
  4. **Mission Outcome Occurrence**
     - Use value_counts() to analyze mission outcomes.
  5. **Landing Outcome Label**
     - Create a binary landing_class column based on mission outcome (0 for failure, 1 for success).
  6. **Export the data to CSV**

- Access the GitHub URL of the completed data wrangling notebook here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%203%20-%20Data%20wrangling%20Ferrero.ipynb

EDA with Data Visualization

- **Scatterplot: Flight Number vs. Launch Site**
  - Shows how launch frequency varies by site.
- **Scatterplot: Payload Mass vs. Launch Site**
  - Identifies how different sites handle varying payload weights.
- **Bar Chart: Success Rate by Orbit Type**
  - Highlights which orbit types have the highest launch success rates.
- **Scatterplot: Flight Number vs. Orbit Type**
  - Examines how orbit types change with increasing flight numbers.
- **Scatterplot: Payload Mass vs. Orbit Type**
  - Explores whether different orbit types accommodate varying payloads.
- **Line Chart: Yearly Launch Success Trend**
  - Tracks how launch success rates have evolved over time.

- Access the GitHub URL of the completed EDA with data visualization notebook here:

https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%205%20-%20EDA%20with%20Visualizations%20Ferrero.ipynb

EDA with SQL

Summary of Performed Queries:

1. Unique Launch Sites.
2. Filtered Launch Records (start with 'CCA').
3. Total Payload by NASA (CRS).
4. Average Payload Mass for F9 v1.1.
5. First Successful Ground Pad Landing.
6. Boosters with Successful Drone Ship Landings with payload mass between 4000-6000.
7. Mission Outcomes Count.
8. List of Boosters with Maximum Payload.
9. Failure Landings on Drone Ships in 2015.
10. Ranked Landing Outcomes between June 4, 2010, and March 20, 2017.

- Access the GitHub URL of the completed EDA with SQL notebook here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%204%20-%20%20EDA%20using%20SQL%20Ferrero.ipynb

Build an Interactive Map with Folium

- **Marked All Launch Sites**
  - Used folium.Marker to visualize the exact locations of all launch sites for better geographical context.
- **Highlighted Launch Outcomes**
  - Added MarkerCluster to display launch success and failure rates at each site, helping identify trends in performance.
- **Added Circle Markers**
  - Used folium.Circle to emphasize each launch site's location with labels, making them easily distinguishable on the map.
- **Mapped Distances to Key Locations**
  - Placed markers and drew folium.PolyLine to show distances from launch sites to the nearest city, coastline, and highway, illustrating infrastructure proximity and potential logistical challenges.

- Access the GitHub URL of the completed interactive map with Folium map here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%206%20-%20%20Interactive%20Visual%20Analytics%20with%20Folium%20Ferrero.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown for Launch Site Selection:** Allows users to filter data by specific launch sites or view all sites together.
- **Pie Chart for Launch Success Rates:** Displays the proportion of successful launches at each site. If a site is selected, it shows the success vs. failure rate for that site.
- **Payload Range Slider:** Enables users to filter launches based on payload mass, refining the displayed data.
- **Scatter Plot of Payload vs. Launch Success:** Illustrates the correlation between payload mass and launch success, color-coded by booster version category.

- Access the GitHub URL of the completed Plotly Dash lab here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%207%20-%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash%20Ferrero.py

Predictive Analysis (Classification)

- **Exploratory Data Analysis (EDA):**
  - Identified key patterns and defined training labels.

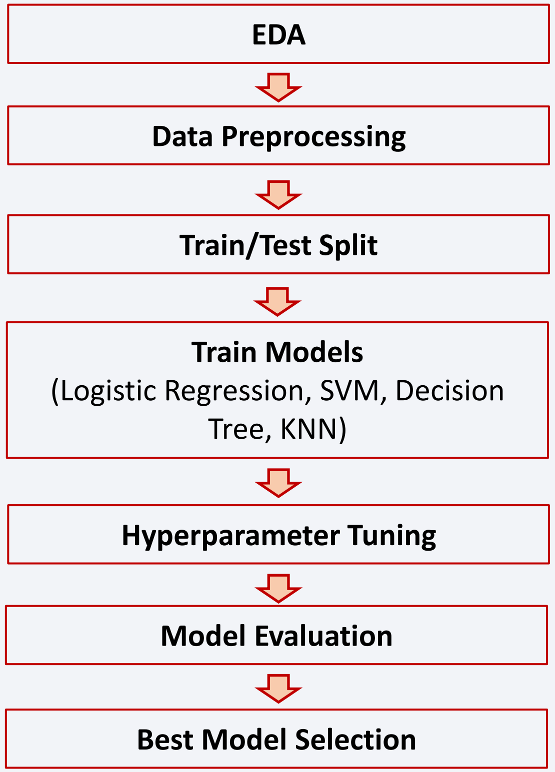
- **Data Preparation:**
  - Created a classification column, standardized features, and split data into training and test sets (80/20 split).

- **Model Training & Hyperparameter Tuning:**
  - Logistic Regression: Used GridSearchCV to find the best hyperparameters.
  - Support Vector Machine (SVM): Optimized using GridSearchCV with cross-validation.
  - Decision Tree Classifier: Tuned hyperparameters to improve accuracy.
  - K-Nearest Neighbors (KNN): Evaluated different values of k for best performance.

- **Model Evaluation:**
  - Assessed accuracy for each model using the score and confusion matrix.

- **Best Model Selection**
  - Compared accuracy scores to determine the most effective classification method.

- Access the GitHub URL of the completed predictive analysis lab here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/blob/main/Lab%208%20-%20Machine%20Learning%20Prediction%20Ferrero%20(1).ipynb



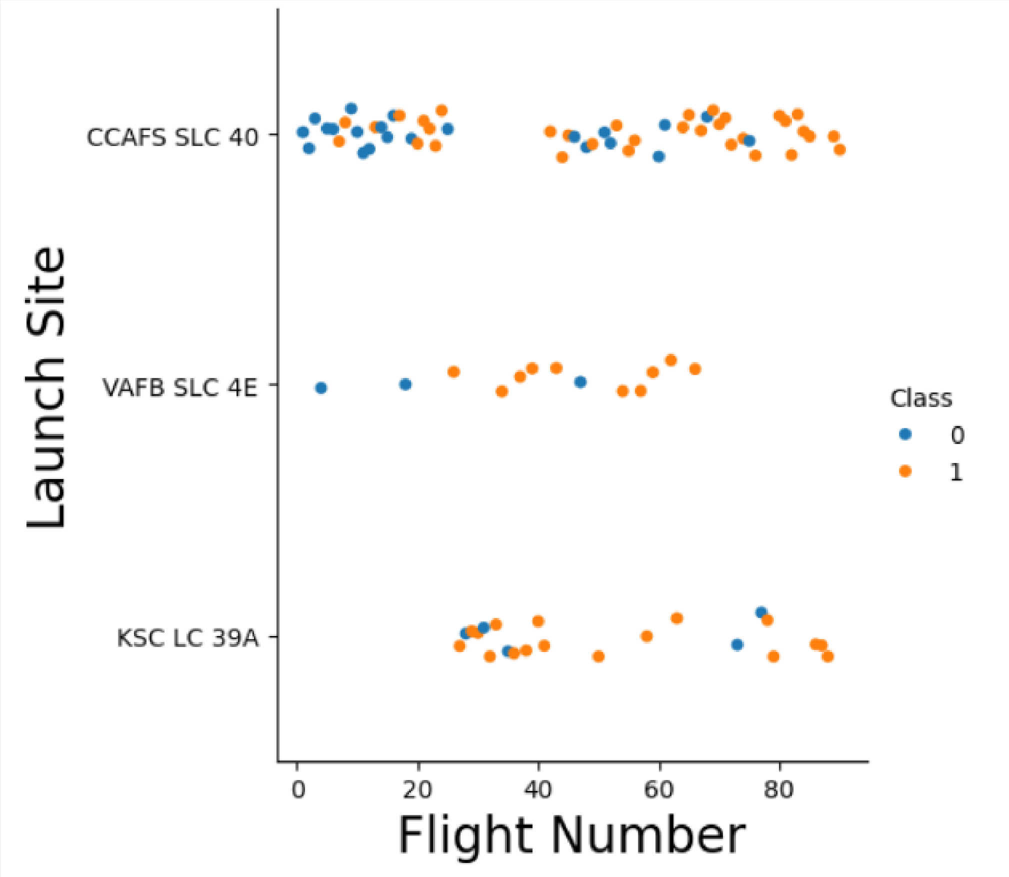
Results

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

Section 2

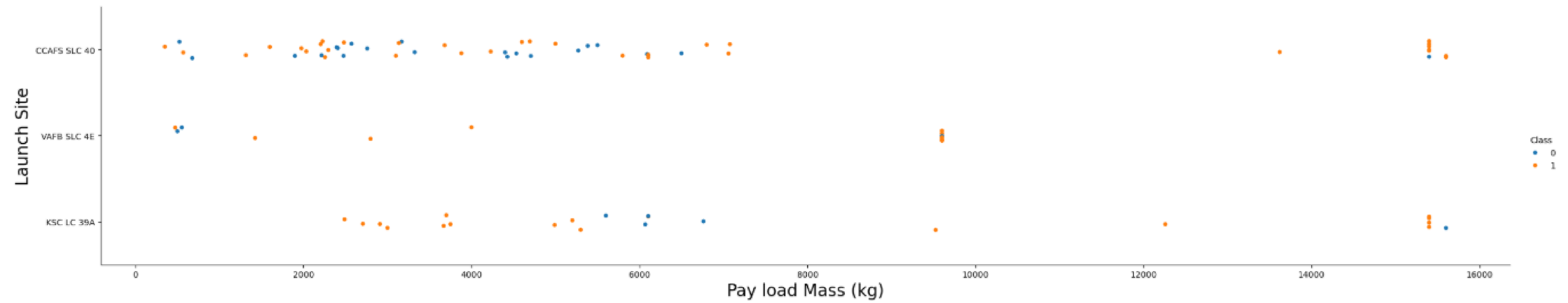
Insights drawn from EDA

Flight Number vs. Launch Site



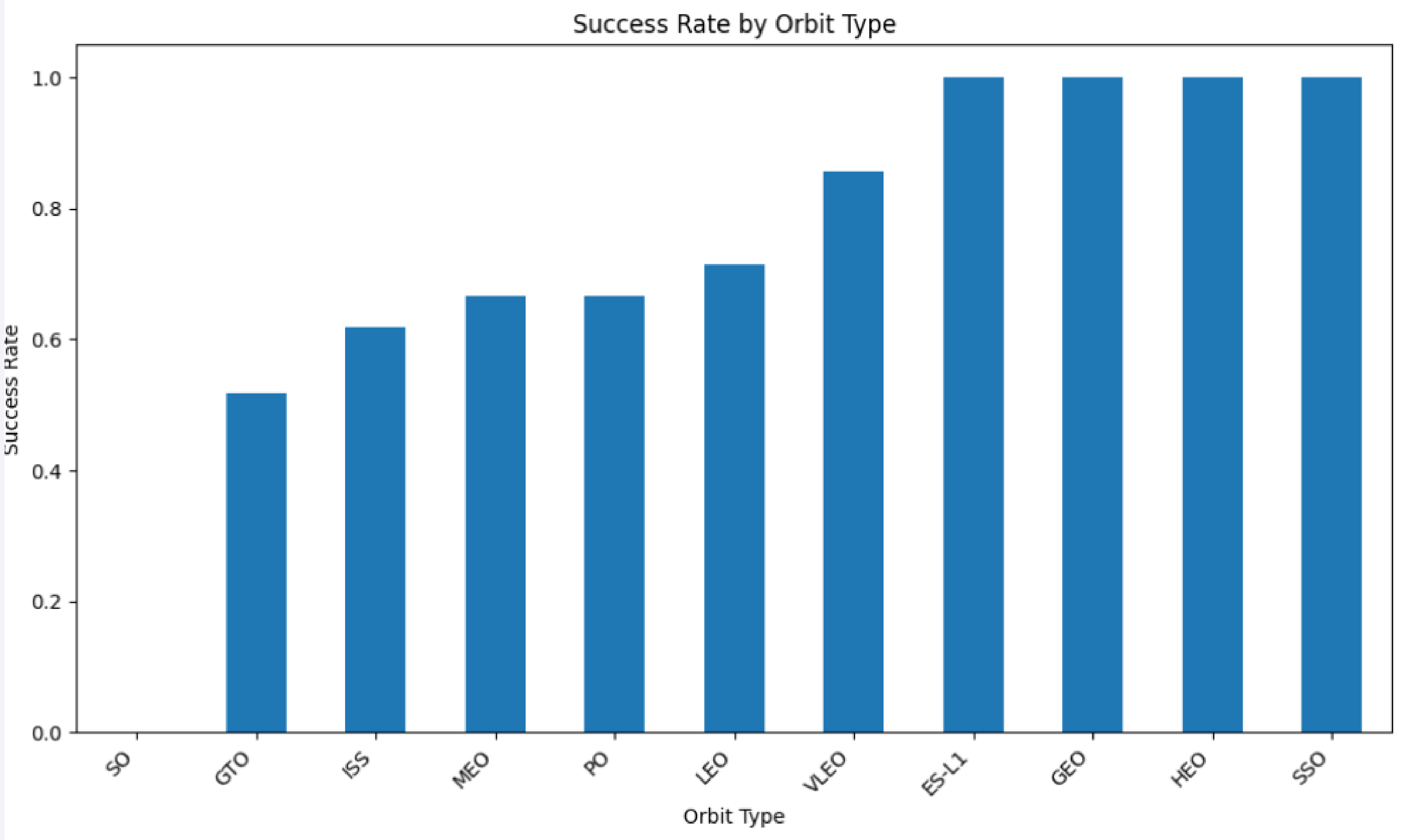
- As flight numbers increase, **CCAFS SLC-40** shows a higher likelihood of successful landings. **VAFB SLC-4E** and **KSC LC-39A** have recorded some failures. However, overall landing success rates have significantly improved.

Payload vs. Launch Site



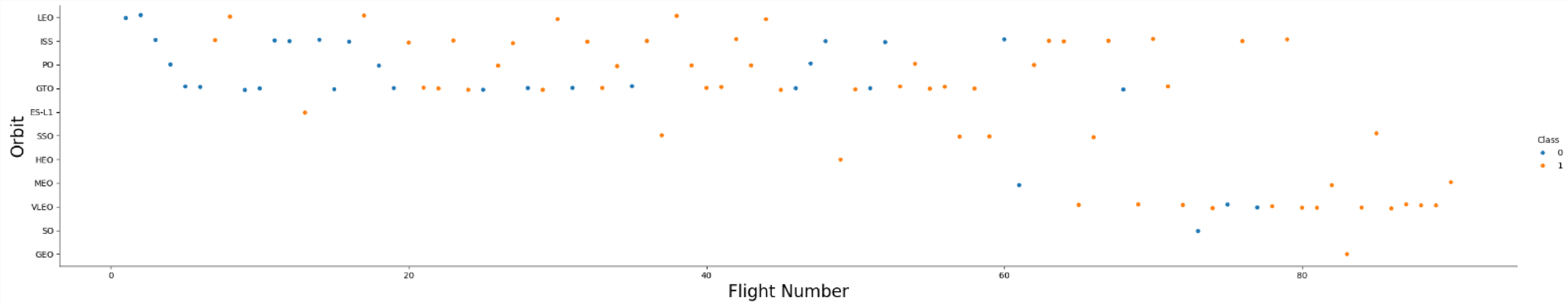
- **CCAFS SLC-40** and **KSC LC-39A** handle heavier payloads (>9,000 kg) with high landing success, while **VAFB SLC-4E** launches lighter payloads (<9,000 kg) with consistent performance. Increased payload mass does not significantly impact landing success.

Success Rate vs. Orbit Type



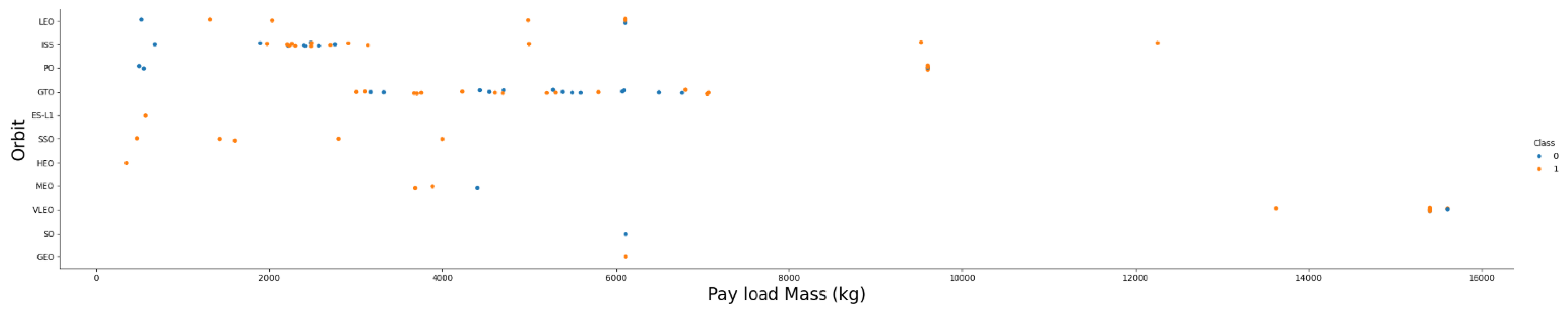
SSO, HEO, GEO, and **ES-L1** show 100% success. In contrast, **GTO** and **ISS** missions have lower success rates.

Flight Number vs. Orbit Type



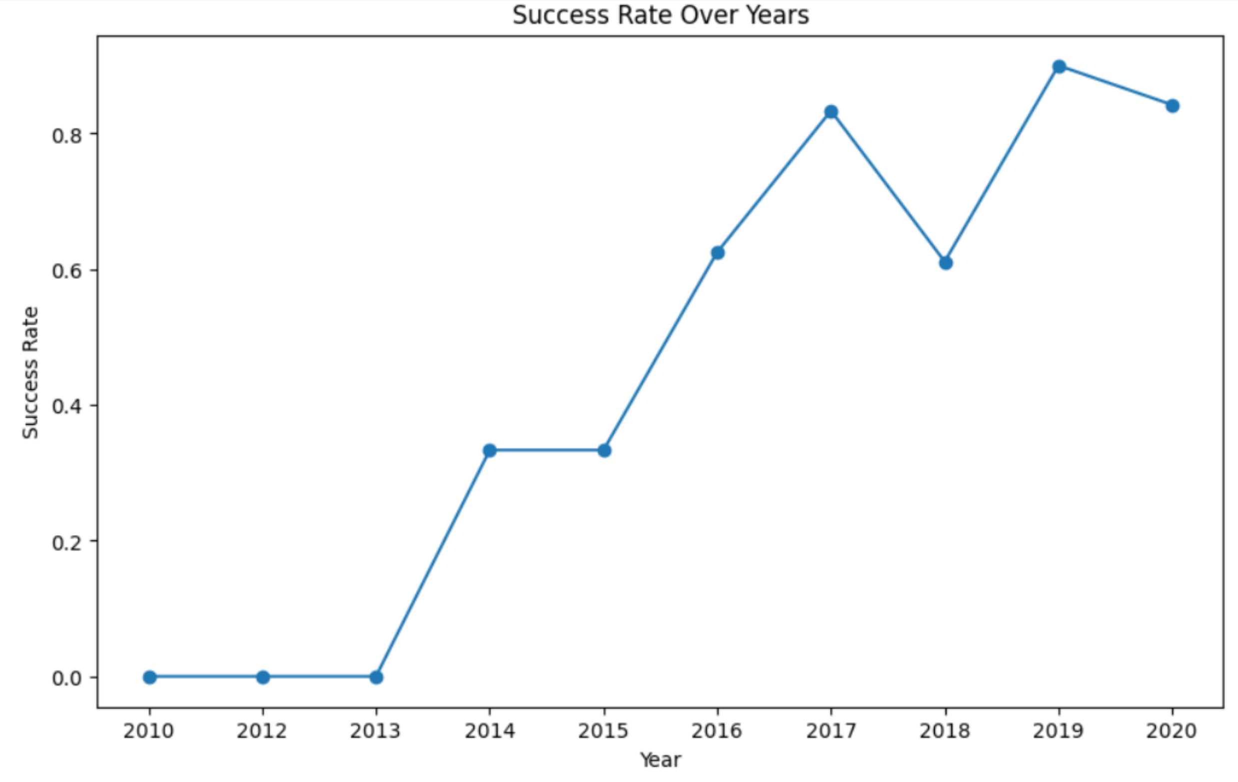
- In the **LEO** orbit, success appears to correlate with the number of flights, while in the **GTO** orbit, no such relationship between flight number and success is evident.

Payload vs. Orbit Type



- For heavy payloads, the successful landing rate is higher for **ISS** and **Polar** orbits. **VLEO** also shows some success with heavy payload landings. However, in **GTO**, distinguishing between successful and unsuccessful landings is challenging, as both outcomes occur.

Launch Success Yearly Trend



- Clearly, success rate since 2013 kept increasing until 2020.

All Launch Site Names

- Find the names of the unique launch sites: CCAFS 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 [83]:
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```

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

Out[83]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with \`CCA\`
- The 5 records where launch sites begin with “CCA” correspond to “CCAFS LC-40”.

Display 5 records where launch sites begin with the string 'CCA'

In [81]:
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Out[81]:

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

- The total payload mass carried by boosters launched by NASA is 45,596 Kg.

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

In [79]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass(Kgs)", Customer FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

Out[79]:

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

In [85]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as "Payload Mass Kgs", Booster_Version FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1
```

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

Out[85]:

Payload Mass Kgs	Booster_Version
2534.6666666666665	F9 v1.1 B1003

- The average payload mass carried by booster version F9 v1.1 is 2,534.66 Kg.

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad was achieved on December 22nd, 2015.

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [44]:
```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Out[44]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

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

In [88]:
```
%sql SELECT DISTINCT Booster_Version, Landing_Outcome, Payload, PAYLOAD_MASS__KG_ as "Payload Mass (Kg)" FROM SPACEXTBL WHEF
```

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

Out[88]:

Booster_Version	Landing_Outcome	Payload	Payload Mass (Kg)
F9 FT B1022	Success (drone ship)	JCSAT-14	4696
F9 FT B1026	Success (drone ship)	JCSAT-16	4600
F9 FT B1021.2	Success (drone ship)	SES-10	5300
F9 FT B1031.2	Success (drone ship)	SES-11 / EchoStar 105	5200

Total Number of Successful and Failure Mission Outcomes

- The total number of successful mission outcomes is 100, and the number of failed mission outcomes is only 1.

List the total number of successful and failure mission outcomes

In [52]:
```
%sql SELECT Mission_Outcome, COUNT(*) AS Total_Count FROM SPACEXTBL GROUP BY Mission_Outcome;
```

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

Out[52]:

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

Boosters Carried Maximum Payload

- The list of the boosters which have carried the maximum payload mass (15,600 Kg) is listed in the table on the right:

List all the booster_versions that have carried the maximum payload mass. Use a subquery.

In [77]:
```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM
```

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

Out[77]:

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

2015 Launch Records

- **F9 v1.1 B1012** and **F9 v1.1 B1015** are the booster versions that failed to land on the drone ship at the **CCAFS LC-40** launch site in 2015.

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [73]:

```
%sql SELECT substr(Date, 6, 2) AS Month, substr(Date,0,5) AS Year, Landing_Outcome, Booster_Version, Launch_Site FROM SPACE)
```

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

Out[73]:

Month	Year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Rank 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.

- From **2010-06-04 to 2017-03-20**, the most frequent landing outcome was **No attempt** (10), followed by **Success (drone ship)** and **Failure (drone ship)**, both with 5 occurrences.

Rank 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.

```
In [84]: %sql SELECT Landing_Outcome, COUNT(*) AS landing_count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUI
```

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

Out[84]:

Landing_Outcome	landing_count
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

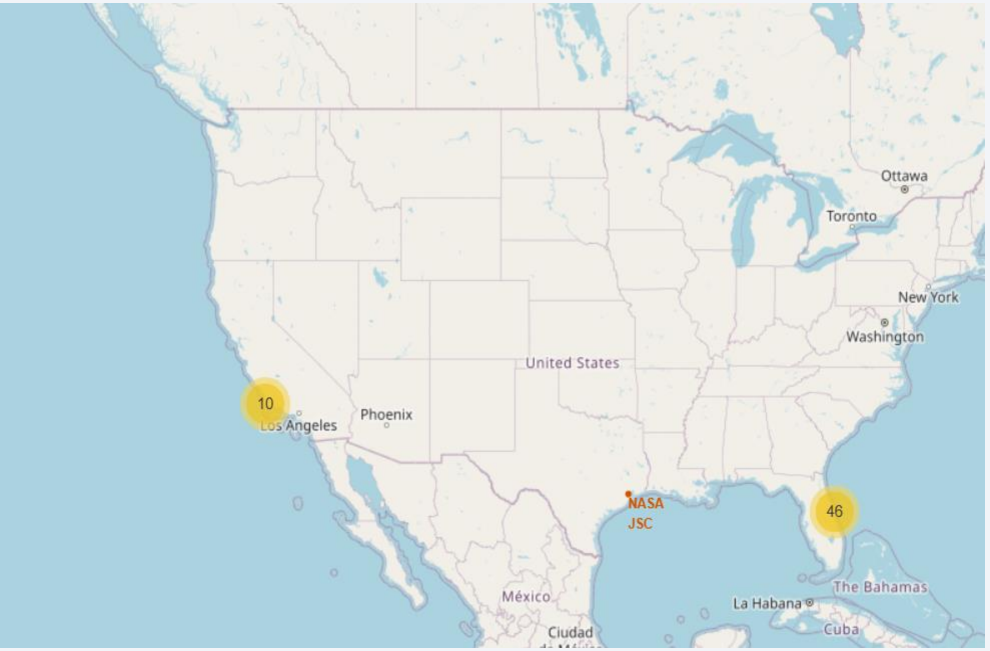
Section 3

Launch Sites Proximities Analysis

SpaceX Launch Sites' Locations on a Global Map

- All launch sites are situated near the Equator, positioned in the southern region of the U.S. map. Additionally, they are all located close to the coastline.

California: Success/Failed Launches for Each Site



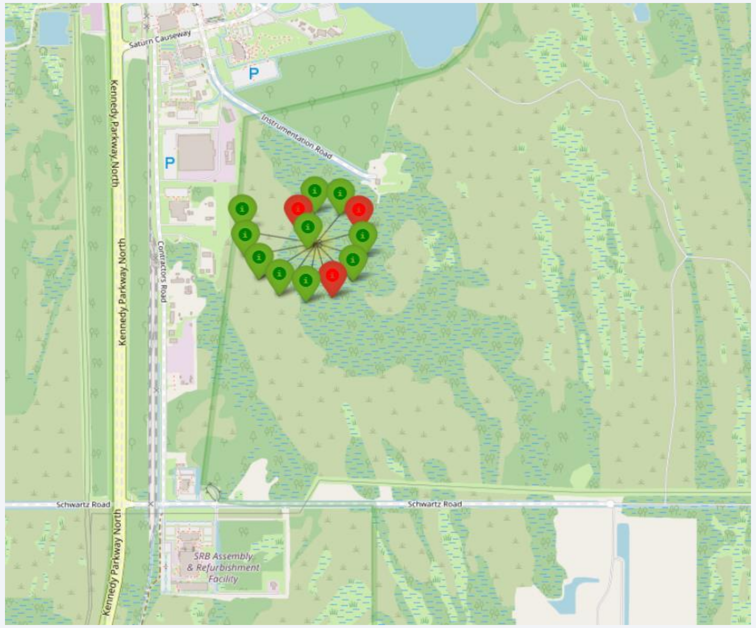
VAFB SLC-4E

- On the West Coast (California), the **VAFB SLC-4E** launch site successfully completed 4 out of 10 launches.

Florida: Success/Failed Launches for Each Site

CCAFS LC-40

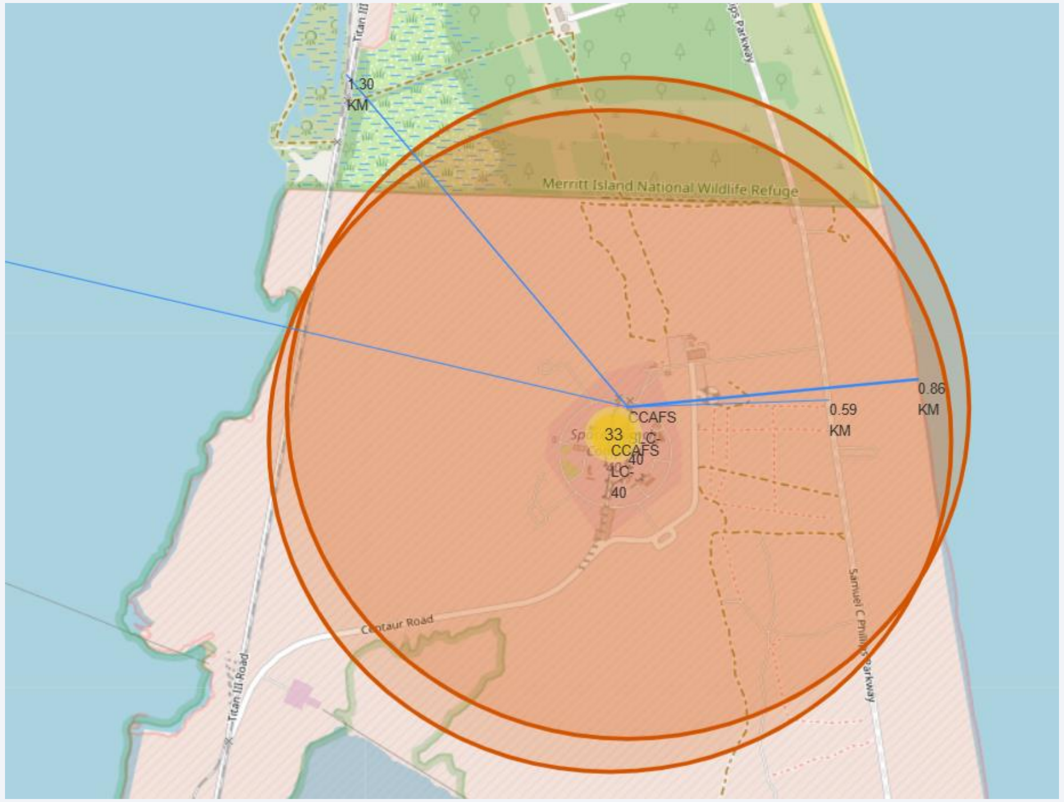
CCAFS SLC-40



KSC LC-39A

- On the East Coast (Florida), the **KSC LC-39A** launch site has a higher success rate compared to **CCAFS SLC-40** and **CCAFS LC-40**.

Distances between CCAFS SLC-40 to its proximities

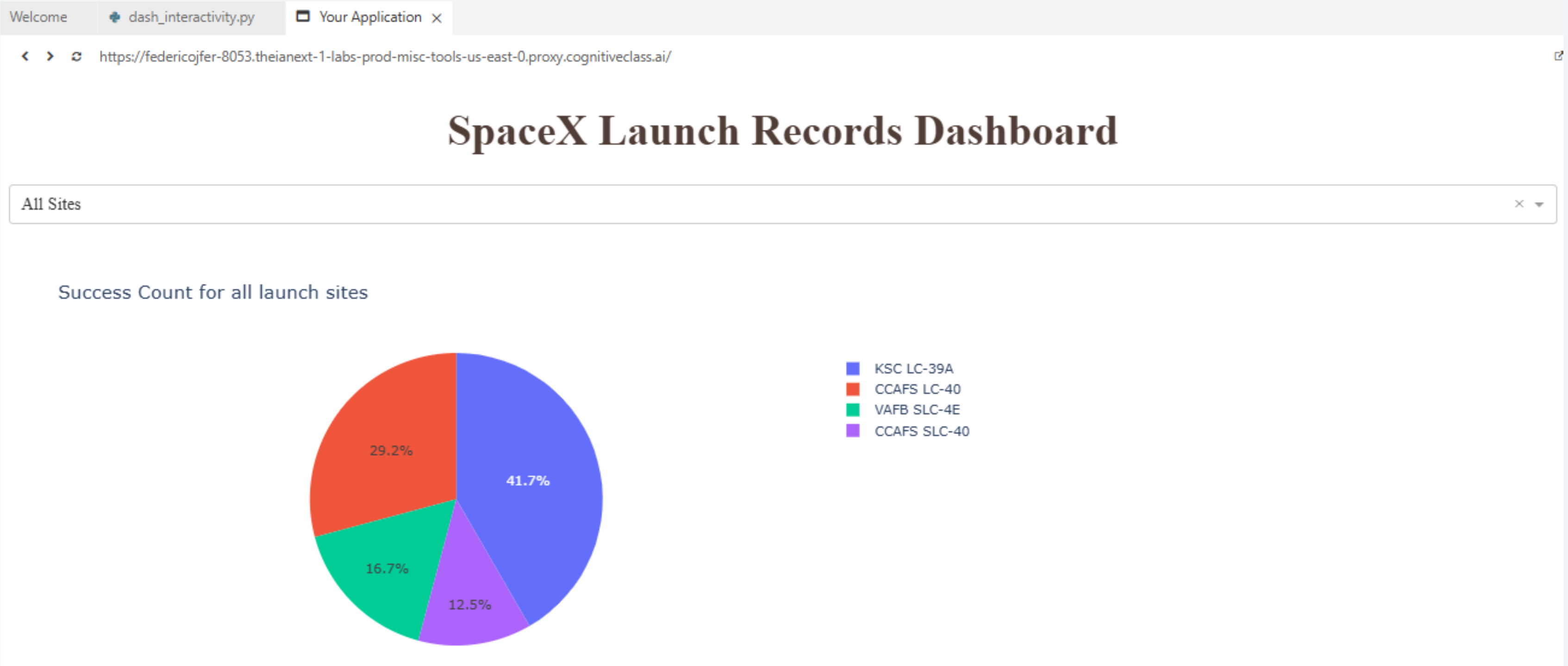


- The nearest city to **CCAFS SLC-40** is Titusville, located **23.19 km** away. The coastline is **0.86 km** from the site, and the nearest road is **0.59 km** away.

Section 4

Build a Dashboard with Plotly Dash

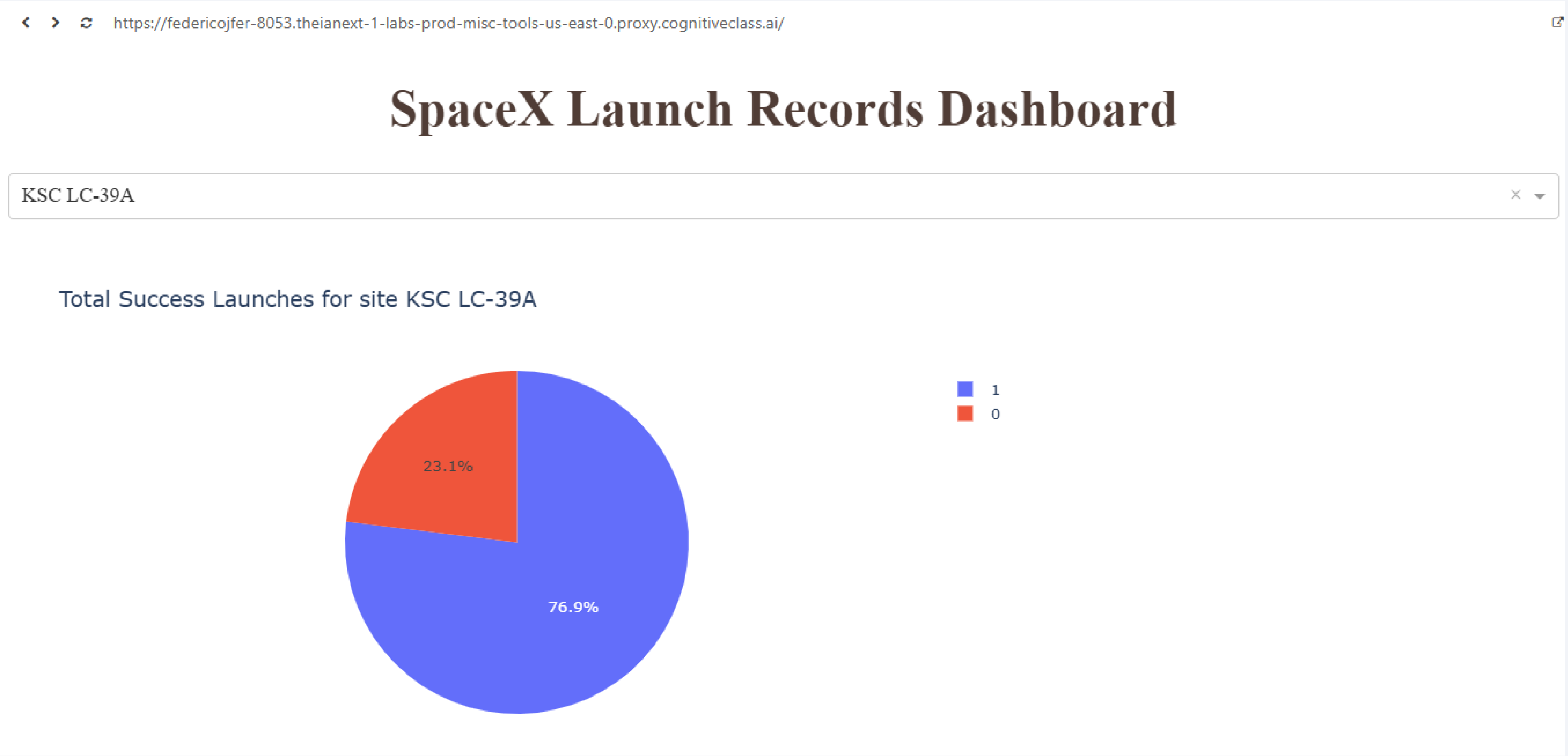
SpaceX Launch Success Count for All Sites



- The launch site **KSC LC-39A** has the highest success count at 42%, followed by **CCAFS LC-40** at 29%, **VAFB SLC-4E** at 17%, and **CCAFS SLC-40** with the lowest success count of 13%.

SpaceX Launch Site with the Highest Launch Success Ratio: KSC LC-39A

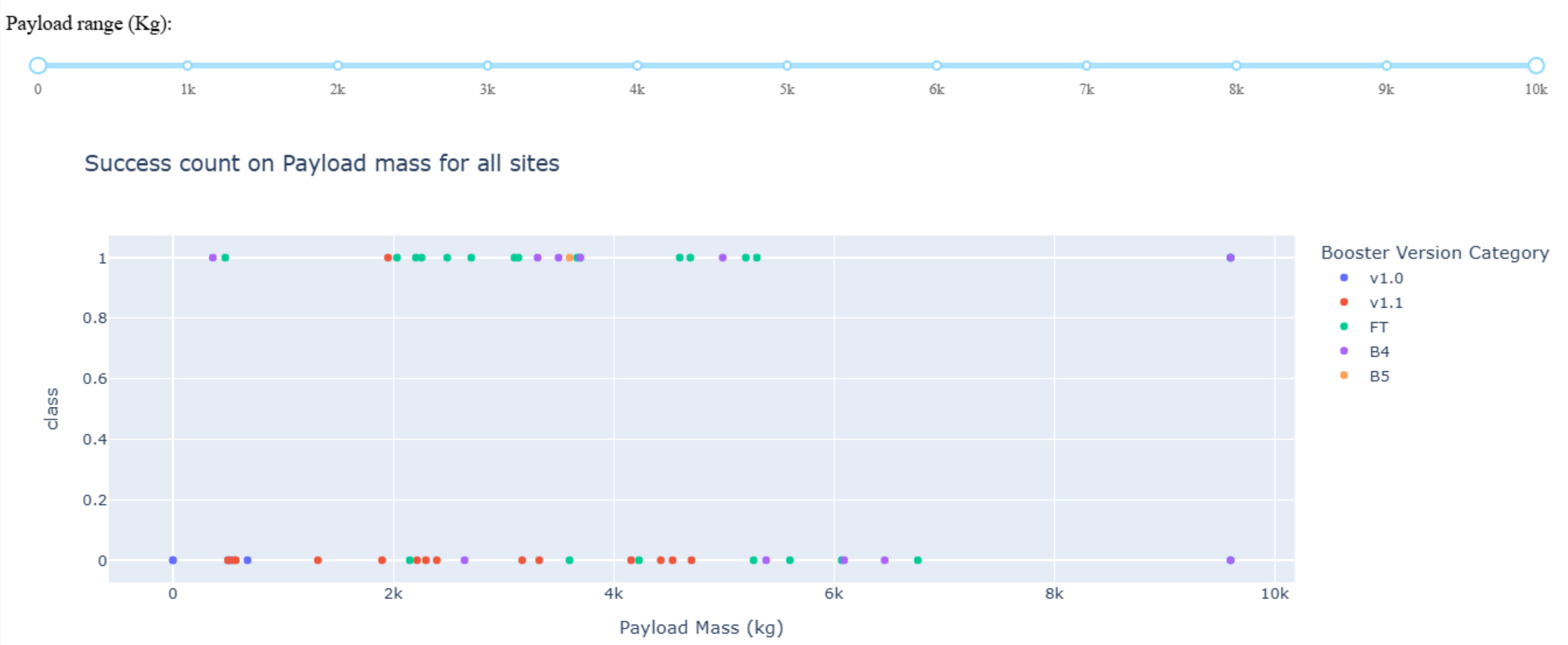
- When we click on the dropdown menu and select **KSC LC-39A**, the total success rate is 77%, which is the highest launch success ratio among the sites.



Payload vs. Launch Outcome for all sites

- When selecting all the sites with different payloads using the range slider, the plot suggests that payload size does not directly determine launch success. Other factors, such as site infrastructure and mission complexity, play a significant role in the outcomes.

- However, generally speaking, the **FT** booster version performs better (highest success count), typically for payloads between 2k and 5k mass, while **v1.1** has the lowest success count.

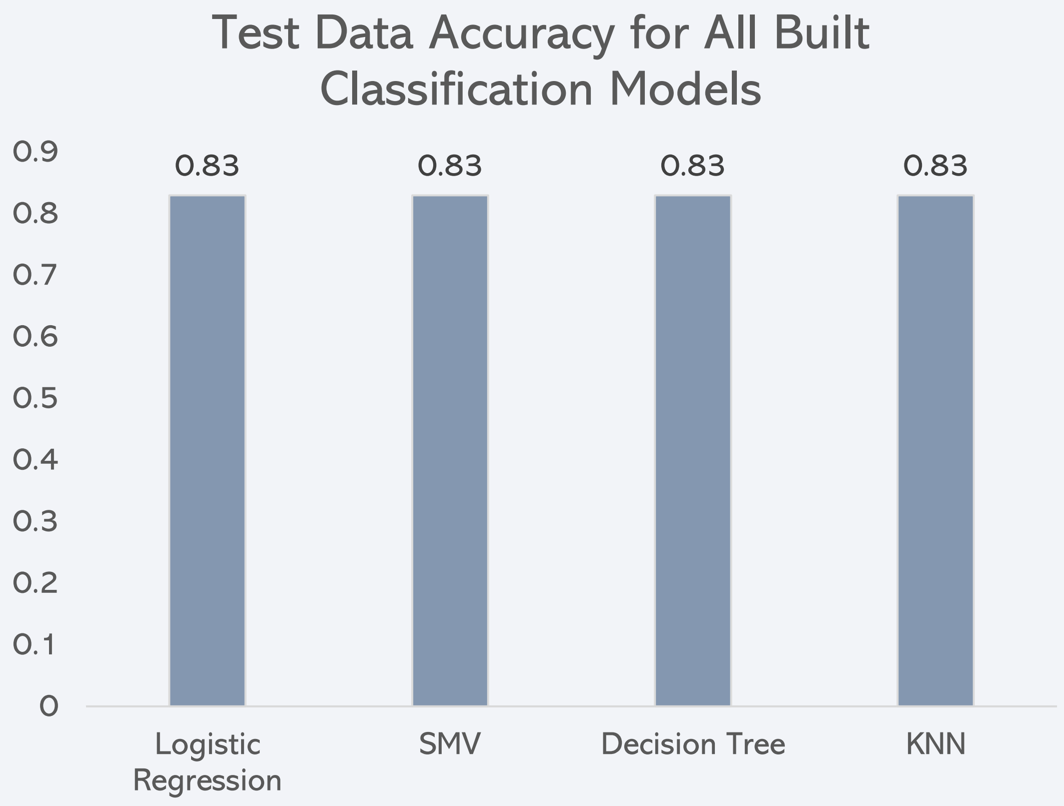


Section 5

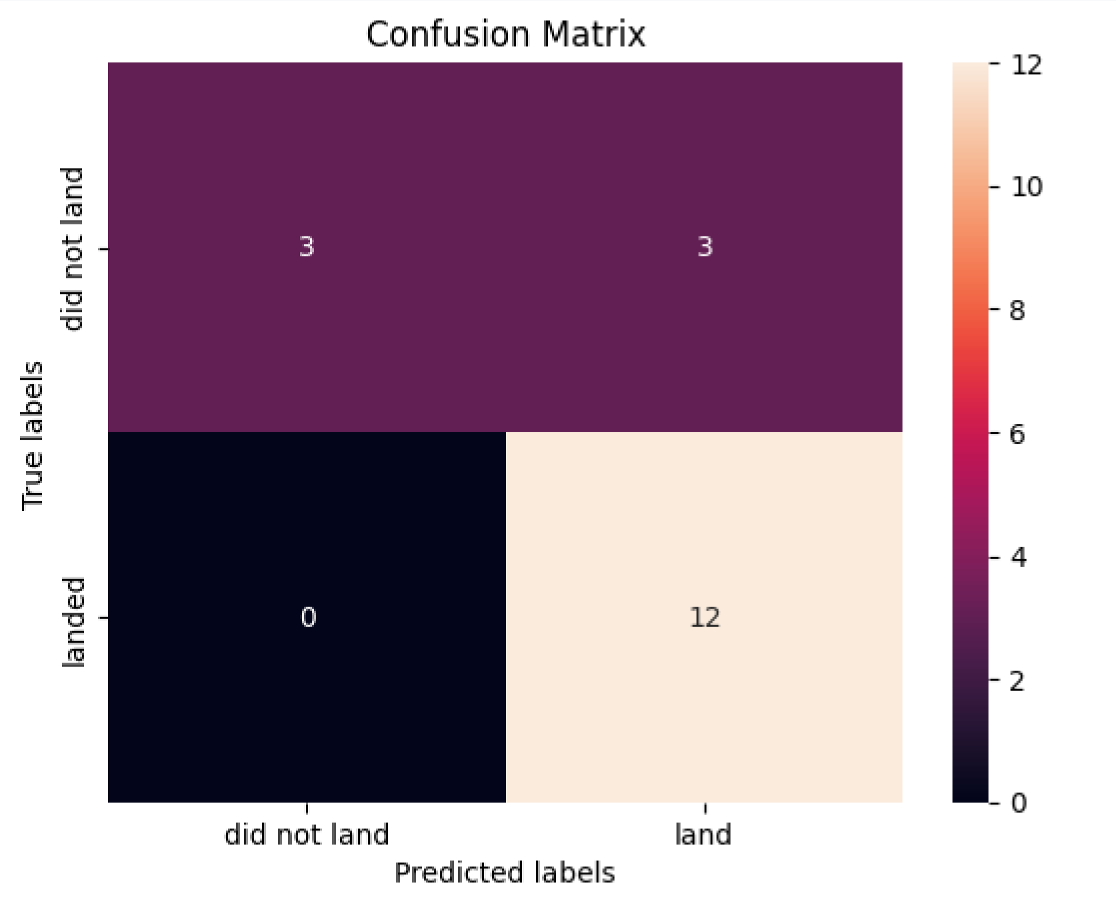
Predictive Analysis (Classification)

Classification Accuracy

- All methods yield identical performance on the test data, each achieving an accuracy of 0.833333.



Confusion Matrix



- Analyzing the confusion matrices, logistic regression, SMC, decision tree, and KNN successfully differentiate between the classes. However, the main issue lies in false positives.

- **True Positives:** 12 (Correctly predicted as landed)

- **False Positives:** 3 (Incorrectly predicted as landed)

- All methods perform the same. Almost all these algorithms produce the same outcome.

Conclusions

- **Landing Success by Launch Site:** As flight numbers increase, CCAFS SLC-40 shows a higher likelihood of successful landings, while VAFB SLC-4E and KSC LC-39A have recorded some failures. Overall, landing success rates have significantly improved over time.
- **Payload Size and Landing Success:** CCAFS SLC-40 and KSC LC-39A handle heavier payloads (>9,000 kg) with high landing success, while VAFB SLC-4E launches lighter payloads (<9,000 kg) with consistent performance. Payload size does not significantly impact landing success.
- **Orbit and Success Rates:** SSO, HEO, GEO, and ES-L1 orbits show 100% landing success, while GTO and ISS missions have lower success rates. In LEO, success seems to correlate with the number of flights, but no such relationship is observed in GTO orbit.
- **Launch Site Performance:** KSC LC-39A has the highest success rate at 77%, followed by CCAFS SLC-40. VAFB SLC-4E has the lowest success count, with payload size not directly determining launch success. Site infrastructure and mission complexity play a significant role.
- **Booster Version Impact:** The FT booster version performs better, typically for payloads between 2k and 5k mass, while the v1.1 version has the lowest success count.
- **When predicting successful landings**, all the models (Logistic Regression, SVM, Decision Tree, and KNN) perform similarly, with test accuracy values close to 0.833333 (or 83.3%) for each. There is no significant difference in performance among the models based on the test data.

Appendix

- Access the GitHub repository for the completed files and code here: https://github.com/federico-jf/SpaceX---Applied-Data-Science-Capstone/tree/main

Thank you!