IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Thanh Hai Nguyen
Nov 3rd , 2023

Outline

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

Executive Summary

- **Objective:** The primary objective of this project is to predict the success of Falcon 9 first-stage landings during SpaceX rocket launches. The successful landing of the first stage is pivotal for reducing launch costs and making SpaceX competitive in the aerospace industry.
- **Methodology:** Data for this analysis was collected from SpaceX launches, including historical data on first-stage landings. Various data science and machine learning techniques were employed to process and analyze the data, ultimately generating predictions regarding landing success.
- **Key Findings:**

1. **Success Rate:** Our analysis indicates that Falcon 9 first-stage landings have a success rate of 66.67%. This provides valuable insights into the reliability of SpaceX's reusability model.
2. **Factors Affecting Success:** Key factors influencing landing success were identified, including launch site, payload mass, orbit type, and flight number. Understanding these variables is crucial for decision-making.
3. **Predictive Model:** We developed predictive models that can estimate the probability of successful landings based on historical data and relevant parameters.

Introduction

Nature of Analysis:

- Predictive Data Analysis is at the heart of our project. We aim to harness the power of data science and machine learning to forecast the success of Falcon 9 first-stage landings.

The Problem:

- The aerospace industry is highly competitive, and launch costs significantly impact market positioning. SpaceX's ability to reduce launch costs is a game-changer. Key to this cost reduction is the reusability of the Falcon 9 first stage. However, predicting the success of first-stage landings remains a challenge.

Introduction

The State of the Problem:

- SpaceX's Game-Changing Model: SpaceX, led by Elon Musk, offers Falcon 9 rocket launches at a fraction of the cost compared to competitors. Much of this savings is attributed to the reusability of the first stage.
- The Need for Accurate Predictions: Reliable predictions of landing success are vital. They enable companies to determine launch costs, facilitating informed decisions when bidding against SpaceX for rocket launch contracts.

The Question for Analysis:

- Can We Predict Falcon 9 First Stage Landing Success? This project's central question revolves around our ability to use historical data and predictive models to determine the likelihood of a successful landing.

Section 1

Methodology

Methodology

Data collection methodology:

- **Data Sources:**

1. **Web Scraping** from Wikipedia:
   - To collect Falcon 9 launch records from a Wikipedia page.
   - Source URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
2. **API Access**: for real-time updates.

- **Data Wrangling and Processing:**

1. **Data Cleaning**: Eliminated duplicates, filled missing values, and standardized formats.
2. **Feature Engineering**: Created new features to account for mission-specific and geographical factors affecting landing success.
3. **Exploratory Data Analysis (EDA)**: Identified patterns and converted outcomes into training labels (1 for success, 0 for failure).

Methodology

Exploratory Data Analysis (EDA):

- EDA was conducted using both visualization and SQL techniques to gain insights into the data.
- SQL queries were employed to extract specific data subsets for in-depth analysis.
- Visualizations were used to explore data distributions and relationships.

Interactive Visual Analytics:

- **Folium** for interactive geospatial data visualization to display landing success by launch site.
- **Plotly Dash** was employed for dynamic data visualization, enabling interactive exploration of mission-specific factors.

Methodology

Predictive analysis using classification models:

- Classification models were built using the scikit-learn library in Python.
- Various classification models for predictive analysis were employed, including Logistic Regression, Support Vector Machine, Decision Trees, and K Nearest Neighbors.
- Optimizing model performances by adjusting hyperparameters (GridSearchCV) for each model to achieve the best possible results.
- Model evaluation was based on metrics such as accuracy, confusion matrix, and .best_score_ method to ensure robustness and identify the most accurate and reliable predictors of landing success.

Data Collection – SpaceX API

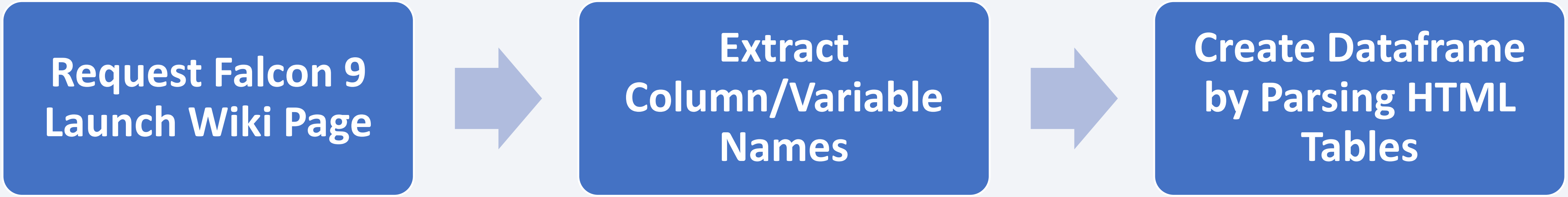
1. Request and parse the SpaceX launch data using the GET request.
2. Filter the dataframe to only include Falcon 9 launches.
3. Dealing with Missing Values

Request SpaceX Launch Data → Parse Data → Filter Falcon 9 Launches → Dealing with Missing Values

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-1-data-collection-api.ipynb

Data Collection – Web Scraping

1. Request the Falcon 9 Launch Wiki page from its URL.
2. Extract all column/variable names from the HTML table header.
3. Create a data frame by parsing the launch HTML tables.



GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-2-data-collection-webscraping.ipynb

Data Wrangling

Data Processing Process:

- In the data wrangling phase, we focused on preparing the SpaceX Falcon 9 launch dataset for further analysis and modeling. Here's an overview of the key steps in the data-wrangling process:

1. **Data Import:** We started by importing the SpaceX dataset, which includes information about Falcon 9 launches. The dataset was loaded from a remote source, and the first few records were displayed to gain an initial understanding of the data.
2. **Handling Missing Values:** We identified missing values in the dataset, particularly in the "LandingPad" column, where approximately 28.89% of the data was missing. These missing values were analyzed to determine their impact on the analysis.
3. **Launch Site Analysis:** We examined the number of launches that occurred at each launch site. The analysis revealed that most launches took place at Cape Canaveral Space Launch Complex 40 (CCAFS SLC 40), followed by Kennedy Space Center Launch Complex 39A (KSC LC 39A) and Vandenberg Air Force Base Space Launch Complex 4E (VAFB SLC 4E).

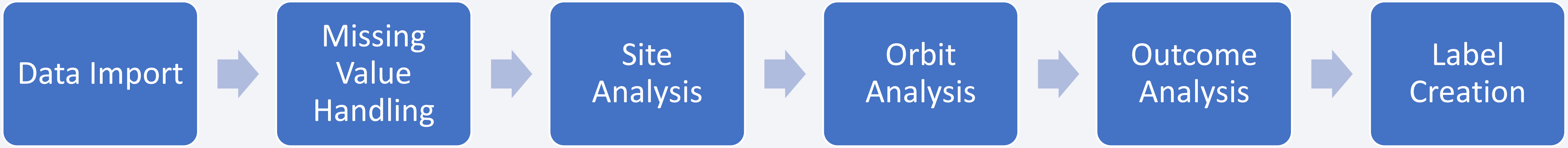
Data Wrangling

4. **Orbit Types:** We explored the different types of orbits in which the Falcon 9 rockets were deployed. Geostationary Transfer Orbit (GTO) was the most common orbit, followed by the International Space Station (ISS) and Very Low Earth Orbit (VLEO).

5. **Mission Outcomes:** We analyzed the mission outcomes and created a set of landing outcomes. An outcome with the word "True" indicated a successful landing, while "False" represented an unsuccessful landing. "None ASDS" and "None None" were classified as failures to land.

6. **Landing Class Label:** A landing class label was created to categorize launches as successful (1) or unsuccessful (0) based on the landing outcome.

7. **Success Rate:** The overall success rate for Falcon 9 first-stage landings was calculated, indicating that approximately 66.67% of the launches were successful.



GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-3-data_wrangling.ipynb

EDA with SQL

- In this section, **SQL queries** were used to explore the SpaceX dataset.

1. **Unique Launch Sites:** Displayed the names of unique launch sites in the space mission.
2. **Launch Sites Beginning with 'CCA':** Displayed 5 records where launch sites begin with 'CCA'.
3. **Total Payload Mass for NASA (CRS):** Displayed the total payload mass carried by NASA (CRS) boosters.
4. **Average Payload Mass for F9 v1.1:** Displayed the average payload mass for booster version F9 v1.1.
5. **First Successful Ground Pad Landing Date:** Listed the date of the first successful ground pad landing.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-4-eda-sql-coursera_sqllite.ipynb

EDA with SQL

6. **Boosters with Success in Drone Ship and Payload Range:** Listed the names of boosters with success in a drone ship landing and specific payload range.
7. **Total Number of Successful and Failure Mission Outcomes:** Listed the total number of successful and failed mission outcomes.
8. **Boosters with Maximum Payload Mass:** Listed the names of booster versions that carried the maximum payload mass.
9. **Records for Months in Year 2015:** Listed records for specific months in the year 2015, including failure landing outcomes.
10. **Ranking Landing Outcomes by Date Range:** Ranked landing outcomes by count between specific date ranges.

- These SQL queries contributed to our exploratory data analysis, providing insights from the SpaceX dataset.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-4-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

Summary of Charts:

- **FlightNumber vs. PayloadMass:** This scatter plot demonstrates the relationship between FlightNumber and PayloadMass, overlaying the launch outcome. As flight number increases, the first stage is more likely to land successfully. Payload mass influences the likelihood of a successful landing.
- **FlightNumber vs. Launch Site:** A scatter point chart shows the relationship between FlightNumber and Launch Site, with launch outcome indicated by hue. It highlights variations in success rates among different launch sites.
- **Payload vs. Launch Site:** This scatter point chart visualizes the relationship between launch sites and payload mass, indicating the launch outcome. It shows that for the VAFB-SLC launch site, there are no rockets launched with heavy payloads (greater than 10,000 kg).

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-5-eda-dataviz.ipynb

EDA with Data Visualization

Summary of Charts:

- **Success Rate by Orbit Type:** A bar chart illustrates the success rate for different orbit types, showing that specific orbits have higher success rates.
- **Flight Number vs. Orbit Type:** A scatter plot reveals how FlightNumber relates to Orbit type and launch outcome. It indicates that in the LEO orbit, success appears to be related to the number of flights.
- **Payload Mass vs. Orbit Type:** Another scatter plot shows the relationship between PayloadMass and Orbit type, emphasizing the success rates for different orbits.
- **Launch Success Yearly Trend:** A line chart displays the average success rate over the years, showing an increasing success rate trend since 2013.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-5-eda-dataviz.ipynb

Build an Interactive Map with Folium

Map Objects Added

- We created the following map objects:

1. **Launch Site Markers:** We marked all launch sites on the map using markers. Each marker represents the location of a launch site and provides a visual reference for where rockets are launched.
2. **Launch Outcomes:** To distinguish between successful and failed launches, we used green markers for successful launches (class=1) and red markers for failed launches (class=0).
3. **Proximity Lines:** We drew lines connecting the launch sites to the closest coastline, city, railway, and highway to visualize their proximity to these features.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-6-interactive_visual_analytics_with_folium_launch_site_location.ipynb

Build an Interactive Map with Folium

Reasons for Adding Objects

1. **Launch Site Markers:** Launch site markers give us a visual reference for the geographical distribution of launch sites. This allows us to observe any patterns or correlations related to their locations.
2. **Launch Outcomes:** The colored markers help us quickly identify which launch sites have high success rates. This is essential for our analysis to understand how location may affect launch success.
3. **Proximity Lines:** Drawing proximity lines helps us measure the distance between launch sites and key features such as coastlines, cities, railways, and highways. It aids in answering questions related to the proximity of launch sites to these features.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-6-interactive_visual_analytics_with_folium_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plots and Interactions Added and Reasons

1. **Launch Site Selection Dropdown:** We've added a dropdown menu that lets users choose a launch site from a list of options, including "All Sites.“

- ➢ This empowers users to filter data by launch site, offering insights into the success rates and launch counts at different locations.

2. **Pie Chart for Success Counts:** We've introduced a pie chart to showcase the total number of successful launches across all sites. When users select a specific launch site, the pie chart dynamically updates to display success vs. failed counts for that site

- ➢ This visually presents the distribution of successful and failed launches, enabling users to swiftly assess the performance of specific launch sites.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-7-dash_app.py

Build a Dashboard with Plotly Dash

3. **Payload Range Slider:** Our dashboard now includes a range slider that lets users choose payload weight ranges.

- ➢ This feature helps users filter data based on payload weight, allowing them to explore how payload weight influences launch outcomes by specifying specific weight boundaries.

4. **Scatter Chart for Payload vs. Launch Success:** We've integrated a scatter chart to visualize the relationship between payload mass and launch success. This interactive chart displays data points based on payload mass, launch success, and booster version category.

- ➢ It illustrates the correlation between payload mass and launch success, enabling users to examine how payload weight impacts the likelihood of a successful launch.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-7-dash_app.py

Predictive Analysis (Classification)

Model Development Process

- We followed a systematic approach, breaking down our model development process into key phases:

1. **Exploratory Data Analysis (EDA):** We began by conducting EDA to understand the data better and determine the training labels.
2. **Feature Engineering:** We created a class column as our training label and standardized the data for uniformity.
3. **Train-Test Split:** The data was divided into training and test datasets. The training data was further split into validation data for hyperparameter tuning.
4. **Hyperparameter Tuning:** We utilized GridSearchCV to find the optimal hyperparameters for classification algorithms: Logistic Regression, Support Vector Machine, Decision Trees, and K Nearest Neighbors.
5. **Model Training:** We trained each model on the training data using the best hyperparameters.
6. **Model Evaluation:** The models were evaluated using validation data to find the accuracy and choose the best-performing model.
7. **Test Data Performance:** The accuracy of the best model was assessed using the test data.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-8-machine_learning_prediction.ipynb

Predictive Analysis (Classification)

Model Performance

Here's a summary of the best-performing model's accuracy using .best_score_ method on the test data:

- **Logistic Regression Accuracy:** 84.64%
- **Support Vector Machine (SVM) Accuracy:** 84.82%
- **Decision Tree Accuracy:** 91.61%
- **K Nearest Neighbors (KNN) Accuracy:** 84.82%
- Based on these results, **Decision Tree** emerged as the best-performing classification model after using hyperparameters and tuning classification models.

GitHub URL: https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX/blob/main/hainguyen-spacex-8-machine_learning_prediction.ipynb

Results

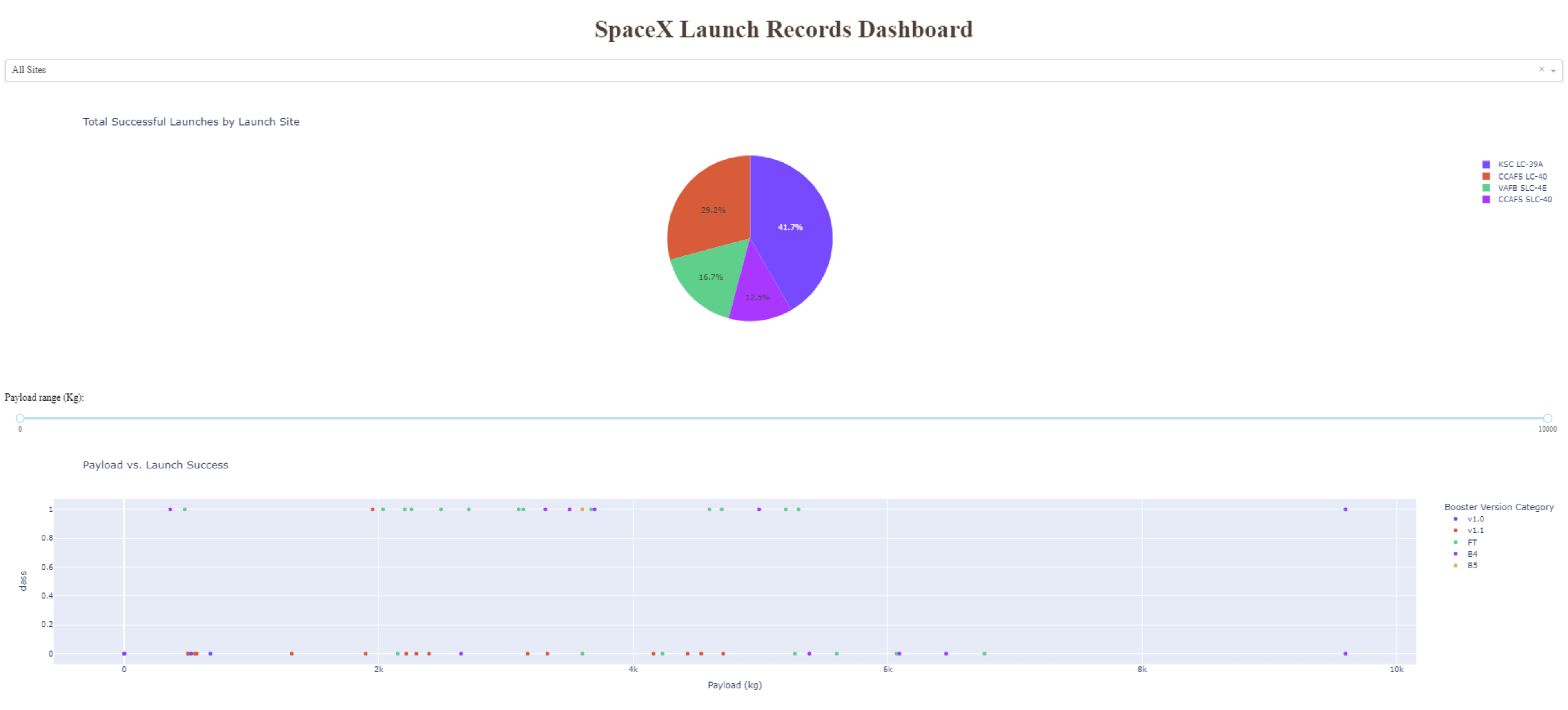
Exploratory Data Analysis Results

Our exploratory data analysis (EDA) unveiled valuable insights into the SpaceX launch records:

- **Launch Site Success Rates:** We observed significant variations in success rates across different launch sites. CCAFS LC-40 has a success rate of 60%, whereas KSC LC-39A and VAFB SLC-4E achieve a remarkable 77% success rate.
- **Payload vs. Launch Site:** Our scatter point chart analysis revealed a distinct pattern for the VAFB-SLC launch site. This site has no recorded rocket launches for payloads exceeding 10,000 kg.
- **Orbit Success Rates:** Orbits such as ES-L1, GEO, HEO, and SSO exhibit a 100% success rate, demonstrating a high degree of reliability.
- **LEO Orbit Success:** We noticed a correlation between success rates and the number of flights in the Low Earth Orbit (LEO). However, for the Geostationary Transfer Orbit (GTO), no such relationship is apparent.
- **Payload-Related Success:** Heavy payloads exhibit varying success rates. Positive landing rates are notably higher for Polar, LEO, and ISS orbits. In contrast, the GTO orbit lacks clear distinctions, with both positive landings and unsuccessful missions occurring.
- **Success Rate Trend:** Our analysis of success rates over time revealed a consistent increase from 2013 to 2020.

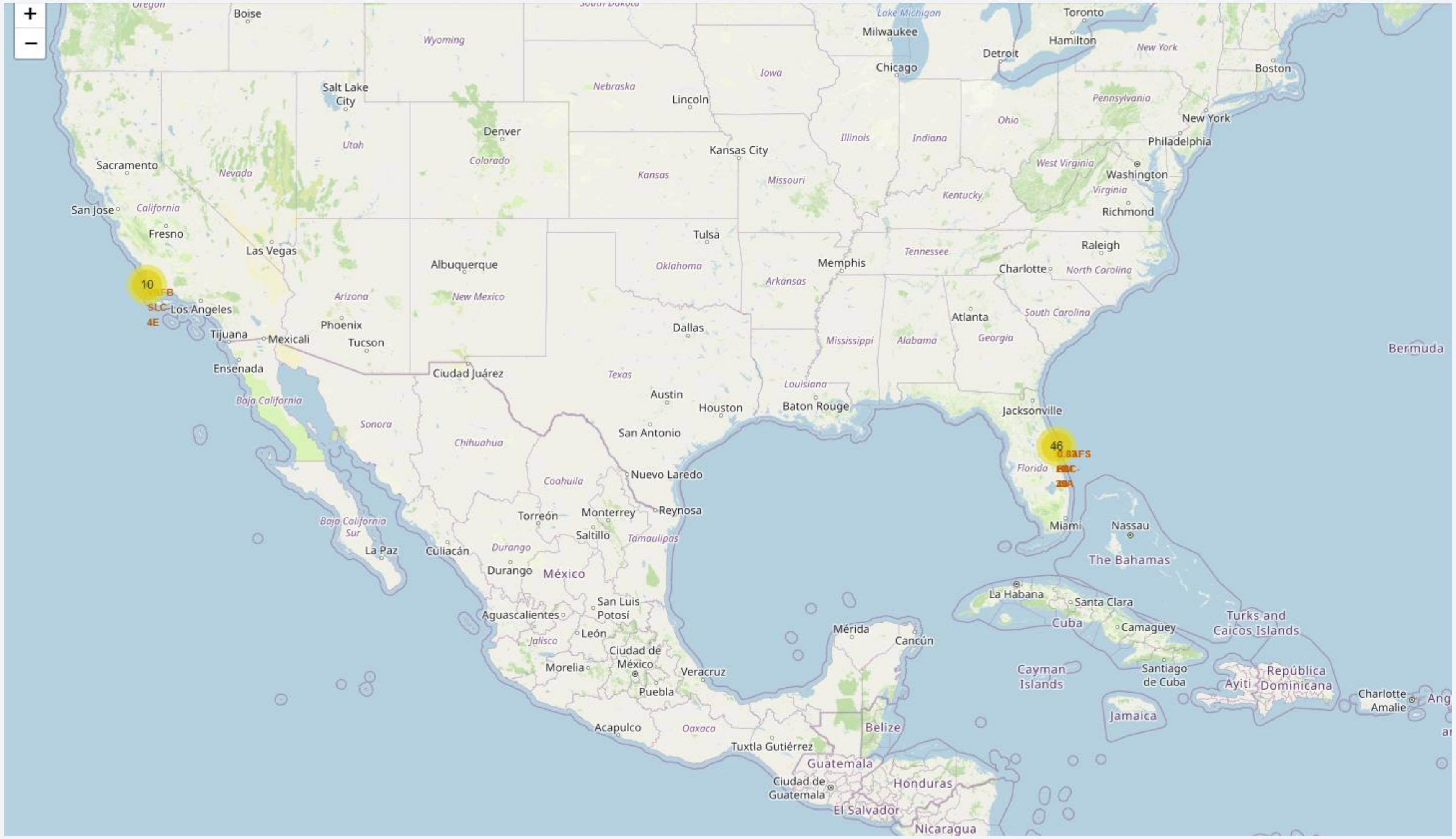
Results

Interactive Analytics Demo in Screenshots



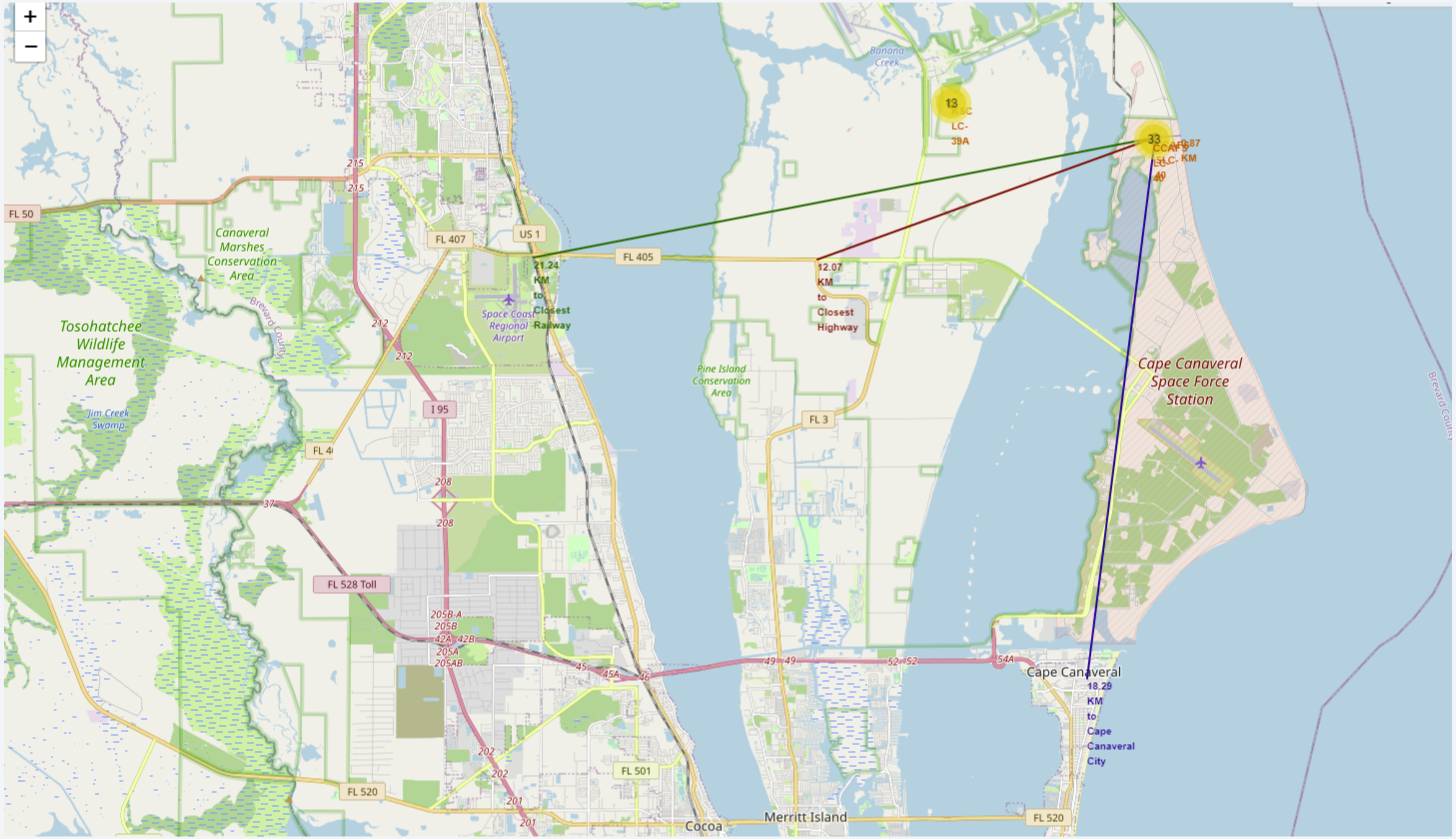
Results

Interactive Analytics Demo in Screenshots



Results

Interactive Analytics Demo in Screenshots



Results

Interactive Analytics Demo in Screenshots

Our interactive analytics demo showcases the geographic and strategic aspects of SpaceX's launch sites:

- **Proximity to the Equator:** Most launch sites are strategically located near the equator, optimizing launch efficiency.
- **Safety Considerations:** Launch sites are situated far enough from populated areas, highways, railways, and other infrastructure to prevent damage in the event of a failed launch. Simultaneously, they remain accessible for logistical support.

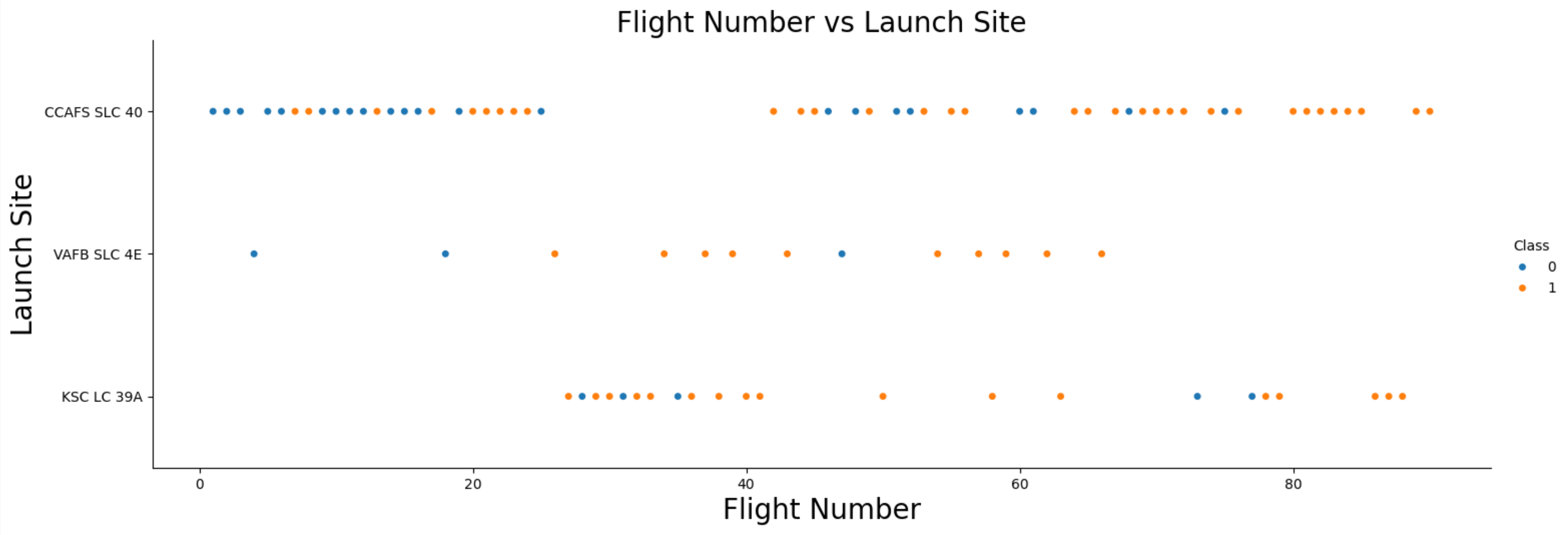
Predictive Analysis Results

- Our predictive analysis results indicate that the Decision Tree model is the best-performing model for predicting the success of Falcon 9 first-stage landings. The accuracy of the model on test data was 91.61%.
- These results are invaluable for SpaceX and prospective competitors, providing insights into launch success patterns, strategic site placement, and predictive models to assess landing outcomes.

Section 2

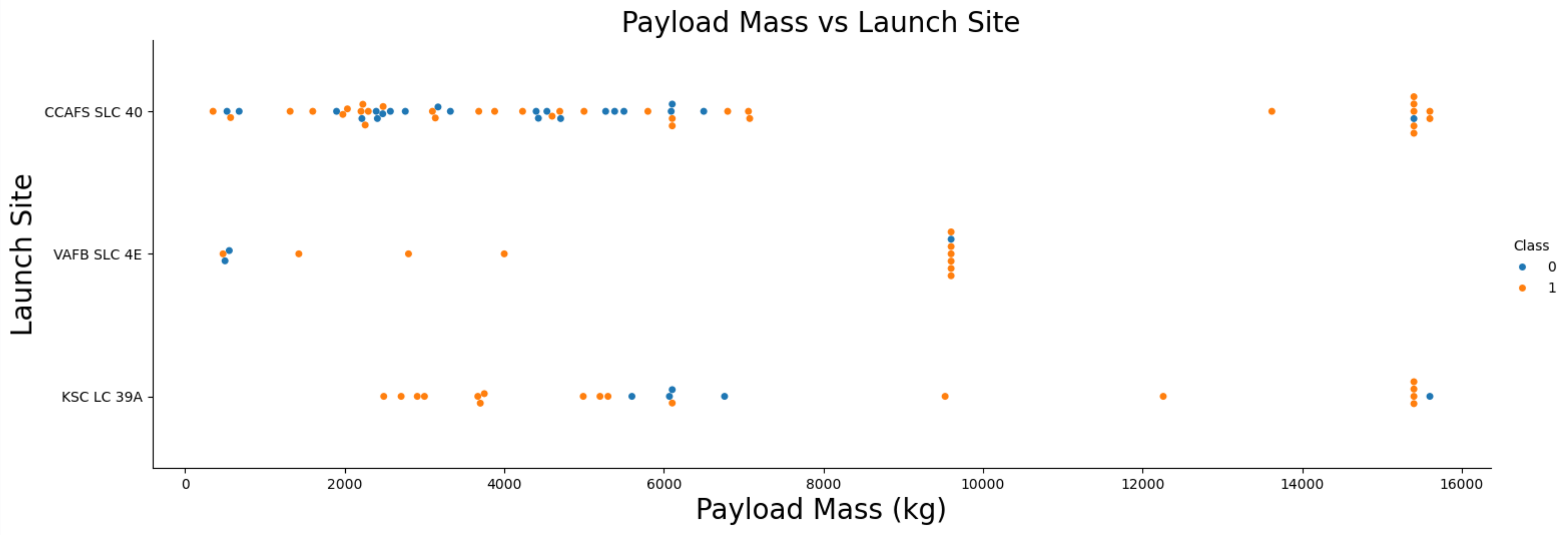
Insights drawn from EDA

Flight Number vs. Launch Site



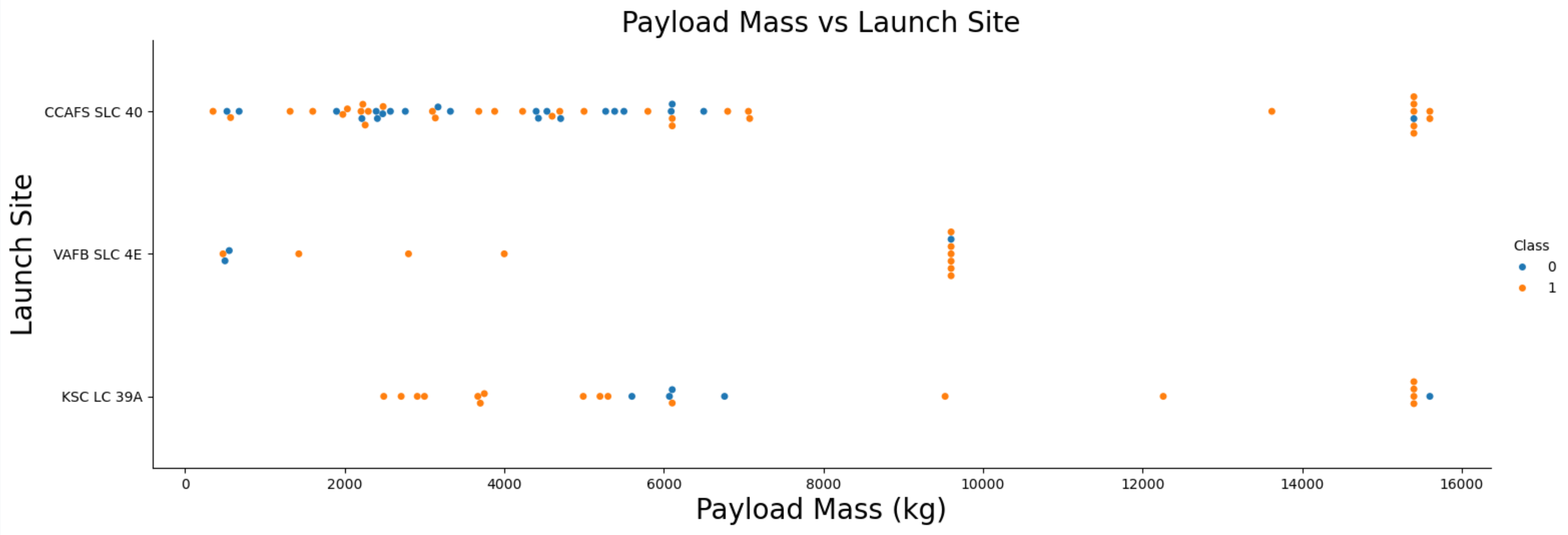
The data shows that earlier flights (blue) had lower success rates, while later flights (orange) had higher success rates. Approximately half of the launches occurred at the CCAFS SLC 40 launch site. VAFB SLC 4E and KSC LC 39A had higher success rates. Based on this, it can be inferred that new launches tend to have a higher success rate.

Payload vs. Launch Site



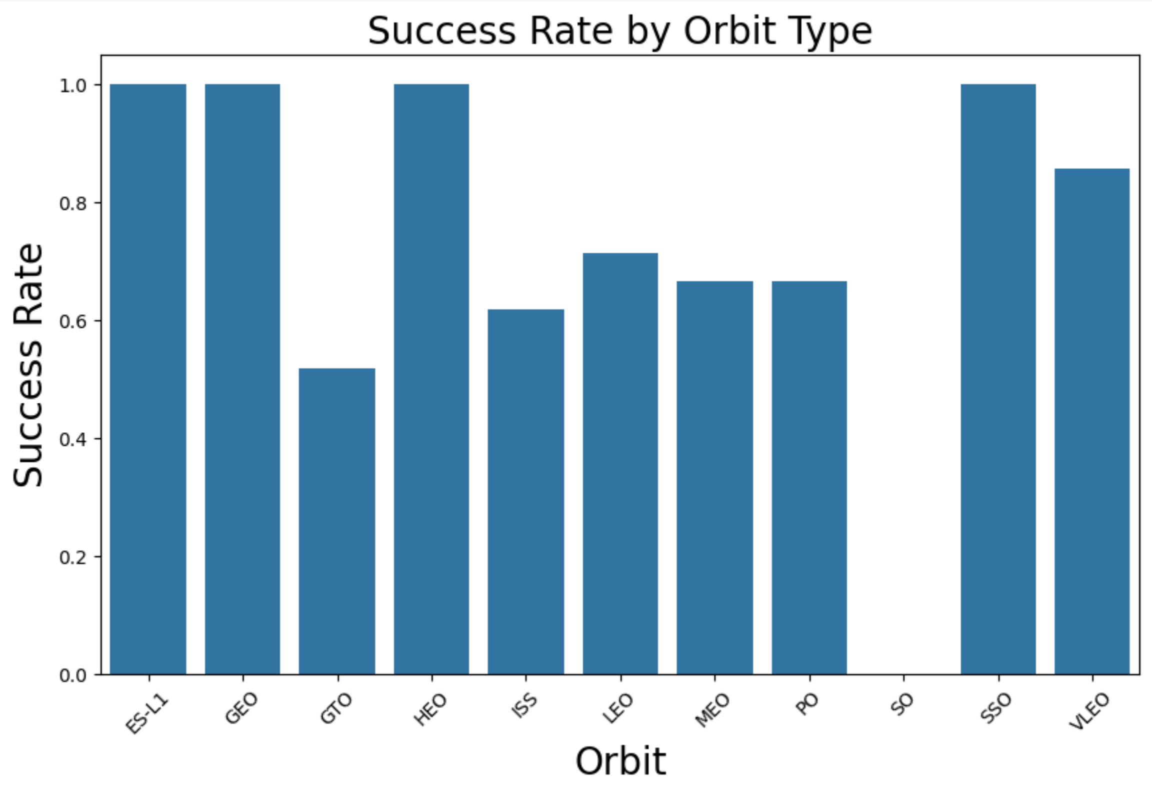
Higher payload mass (in kg) tends to be linked with a higher success rate. Specifically, the scatter plot indicates that most launches with a payload exceeding 8,000 kg were successful, with only three exceptions. Remarkably, KSC LC 39A achieved a perfect 100% success rate for launches with payloads less than 5,500 kg. It's also worth noting that VAFB SLC 4E hasn't undertaken any launches with payloads surpassing 10,000 kg.

Payload vs. Launch Site



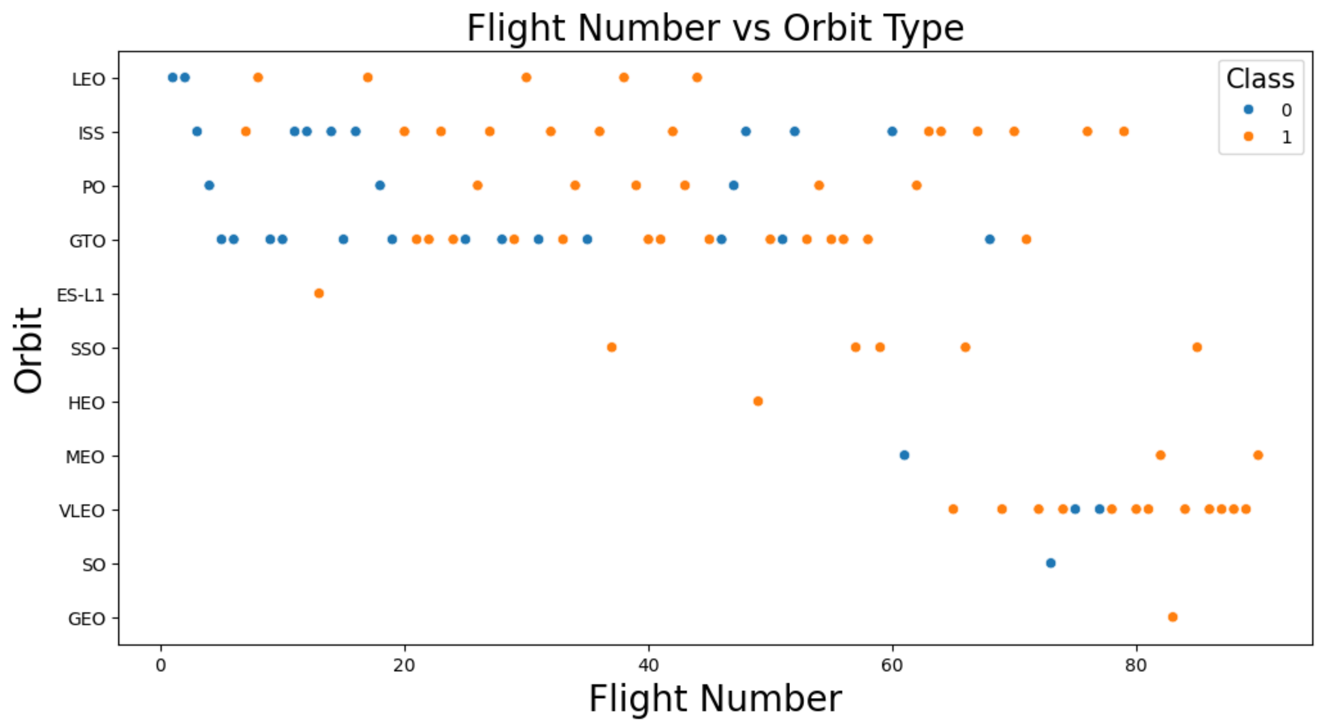
In general, a higher payload mass (in kg) is associated with a higher success rate. Specifically, most launches with a payload greater than 8,000 kg were successful. Notably, KSC LC 39A achieved a 100% success rate for launches with payloads less than 5,500 kg. Additionally, VAFB SLC 4E hasn't conducted any launches with payloads exceeding approximately 10,000 kg.

Success Rate vs. Orbit Type



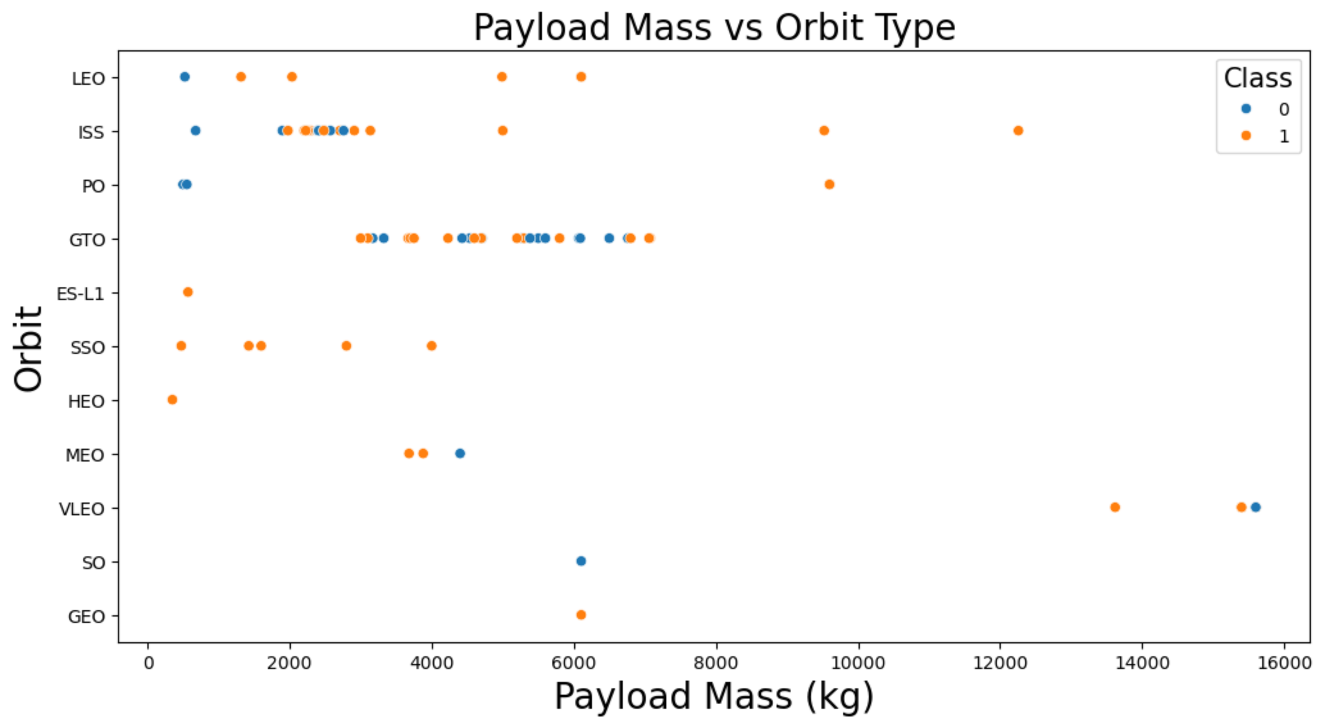
Based on the chart, it's evident that ES-L1, GEO, HEO, and SSO orbits consistently achieve a 100% success rate. Meanwhile, GTO, ISS, LEO, MEO, and PO orbits typically fall in the range of 50% to 70% success rate. In contrast, SO missions have the lowest success rate, remaining at 0%, as depicted in the bar chart.

Flight Number vs. Orbit Type



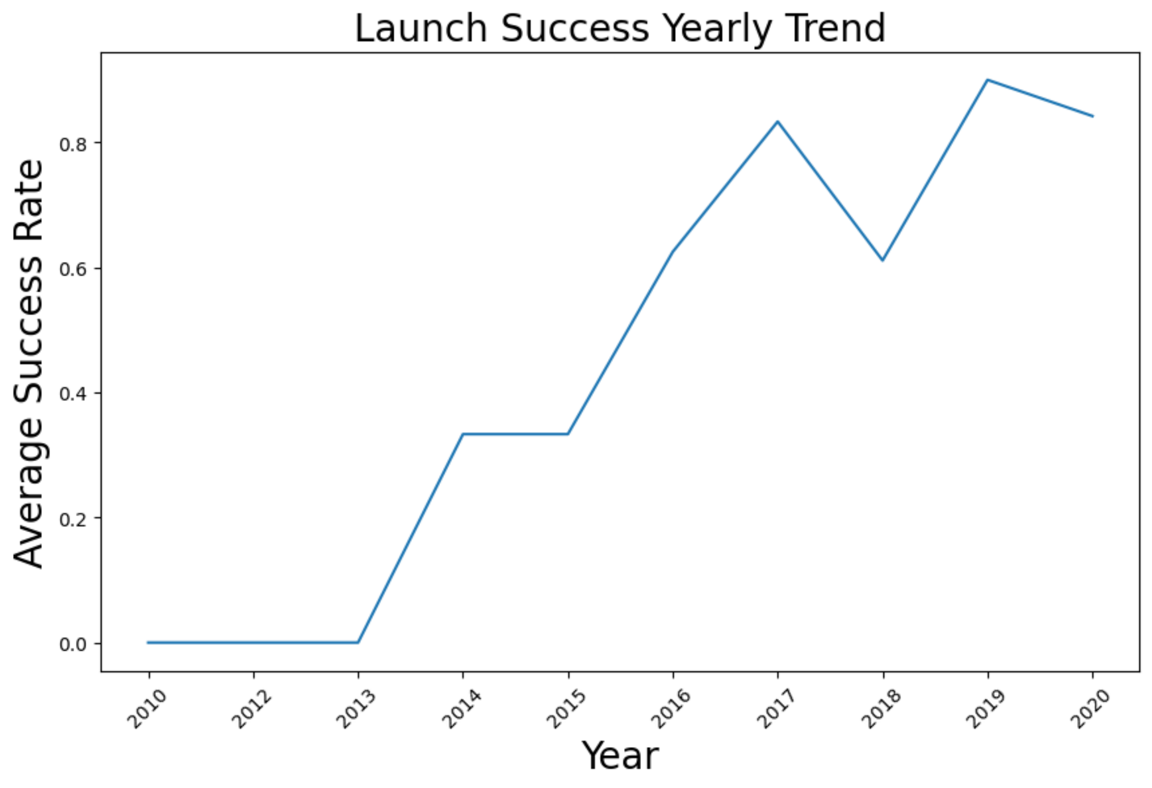
Recent flights have demonstrated higher success rates compared to the initial 20 flights. It's apparent that ES-L1, SSO, HEO, and GEO orbits consistently achieve a 100% success rate, as evident from the scatter plot. Analyzing the dot colors for ISS and GTO, which have a higher number of launches compared to other orbits, it's clear that they have improved their success rates over time, likely learning from initial failures and enhancing their preparations for subsequent launches. Notably, from the 60th flight onwards, there was a significant increase in successful launches for the VLEO orbit, with only two failures. Remarkably, from the 80th flight onwards until the present, there have been no failures in any orbit.

Payload vs. Orbit Type



The GTO orbit exhibits a pattern of alternating successful and failed launches within the payload mass range of 3000 kg to 8000 kg. A similar pattern is observed for the ISS orbit, but ISS consistently achieves a 100% success rate for payloads exceeding 4000 kg. The SSO orbit, on the other hand, maintains a 100% success rate for payloads below 4000 kg. In general, it appears that the LEO, ISS, and PO orbits tend to have a higher likelihood of success with greater payload mass.

Launch Success Yearly Trend



The data shows a clear trend of increasing success rates, especially from 2013 to 2014 and from 2015 onwards. Although there are some fluctuations in success rates from 2017 to 2020, the overall trajectory suggests a positive trend in launch success over time.

All Launch Site Names

- The names of the unique launch sites in the space mission can be extracted from the "Launch_Site" column of the SPACEXTABLE. The following are the unique launch sites:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

- These are the distinct locations from which space missions have been launched.

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

```
[8]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

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

[8]:

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

Launch Site Names Begin with 'CCA'

- The query retrieves 5 records where the launch sites begin with the string 'CCA.' Here are the results:

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

[9]:
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- These records provide information about space missions that were launched from sites starting with 'CCA' in the launch site name.

Total Payload Mass

- To calculate the total payload mass carried by boosters launched by NASA (CRS), the following SQL query was executed:

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

```
[10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)';

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

[10]:

SUM(PAYLOAD_MASS__KG_)
45596

- The result is that the total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg. This query calculates the sum of the "PAYLOAD_MASS__KG_" column for all records where the "Customer" field matches 'NASA (CRS).' It provides the cumulative payload mass for NASA's CRS missions.

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by booster version F9 v1.1, the following SQL query was used:

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

```
[11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';

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

[11]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- The result indicates that the average payload mass carried by booster version F9 v1.1 is 2,928.4 kg. This query computes the mean payload mass for all records where the "Booster_Version" field matches 'F9 v1.1,' providing insight into the typical payload capacity of boosters with this specific version.

First Successful Ground Landing Date

- The query aims to find the date of the first successful landing outcome on a ground pad. Here's the result:

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

Hint:Use min function

```
[12]: %sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

[12]:

MIN("Date")
2015-12-22

- The first successful landing outcome on a ground pad was achieved on the date 2015-12-22. This information is obtained by using the ‘MIN’ function to identify the earliest date in the SPACEXTABLE where the “Landing_Outcome” is recorded as ‘Success (ground pad).’

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
[13]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;

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

[13]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The boosters that meet these criteria and have successfully landed on a drone ship are as follows:

1. F9 FT B1022
2. F9 FT B1026
3. F9 FT B1021.2
4. F9 FT B1031.2

- These boosters fulfill the conditions of both a successful drone ship landing and a payload mass within the specified range (greater than 4000 kg and less than 6000 kg).

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
[14]: %sql SELECT Mission_Outcome, COUNT(*) as "Total" FROM SPACEXTABLE GROUP BY Mission_Outcome;

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

[14]:

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

- Total number of successful missions: 99
- Total number of failure (in flight) missions: 1
- Total number of successful (payload status unclear) missions: 1

The query groups the data by the "Mission_Outcome" and uses the COUNT function to count the occurrences of each unique outcome. This provides a breakdown of both successful and various failure categories within the dataset, including "Failure (in flight)" and "Success (payload status unclear)."

Boosters Carried Maximum Payload

- The booster versions that have carried the maximum payload mass are listed below:
- 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

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
[15]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[15]:

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 result is achieved by using a subquery within the main query to first determine the maximum payload mass using the MAX function and then filtering the records where the "PAYLOAD_MASS__KG_" matches this maximum value. The subquery ensures that the maximum payload mass is identified and then used as a filter to retrieve the corresponding booster versions.

2015 Launch Records

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.

```
[16]: %sql SELECT strftime('%m', Date) AS Month, Landing_Outcome, Booster_Version, Launch_Site \
          FROM SPACEXTABLE WHERE strftime('%Y', Date) = '2015' AND Landing_Outcome = 'Failure (drone ship)' ORDER BY Month;
```

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

[16]:

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

- In April (Month 04) of 2015, there was a "Failure (drone ship)" landing outcome with the booster version F9 v1.1 B1015 at the launch site CCAFS LC-40.
- In October (Month 10) of 2015, there was another "Failure (drone ship)" landing outcome with the booster version F9 v1.1 B1012, also at the launch site CCAFS LC-40.
- The query provides a clear breakdown of the specific months in the year 2015 when failure landing outcomes on drone ships occurred, along with details on the booster versions and launch sites associated with these incidents.

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.

[17]:
```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS "Count" \
FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY Landing_Outcome \
ORDER BY "Count" DESC;
```

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

[17]:

Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

- The query groups the data within the specified date range and calculates the count of each unique landing outcome using the COUNT function. It then arranges the results in descending order, providing a ranking of the landing outcomes with the most occurrences first.

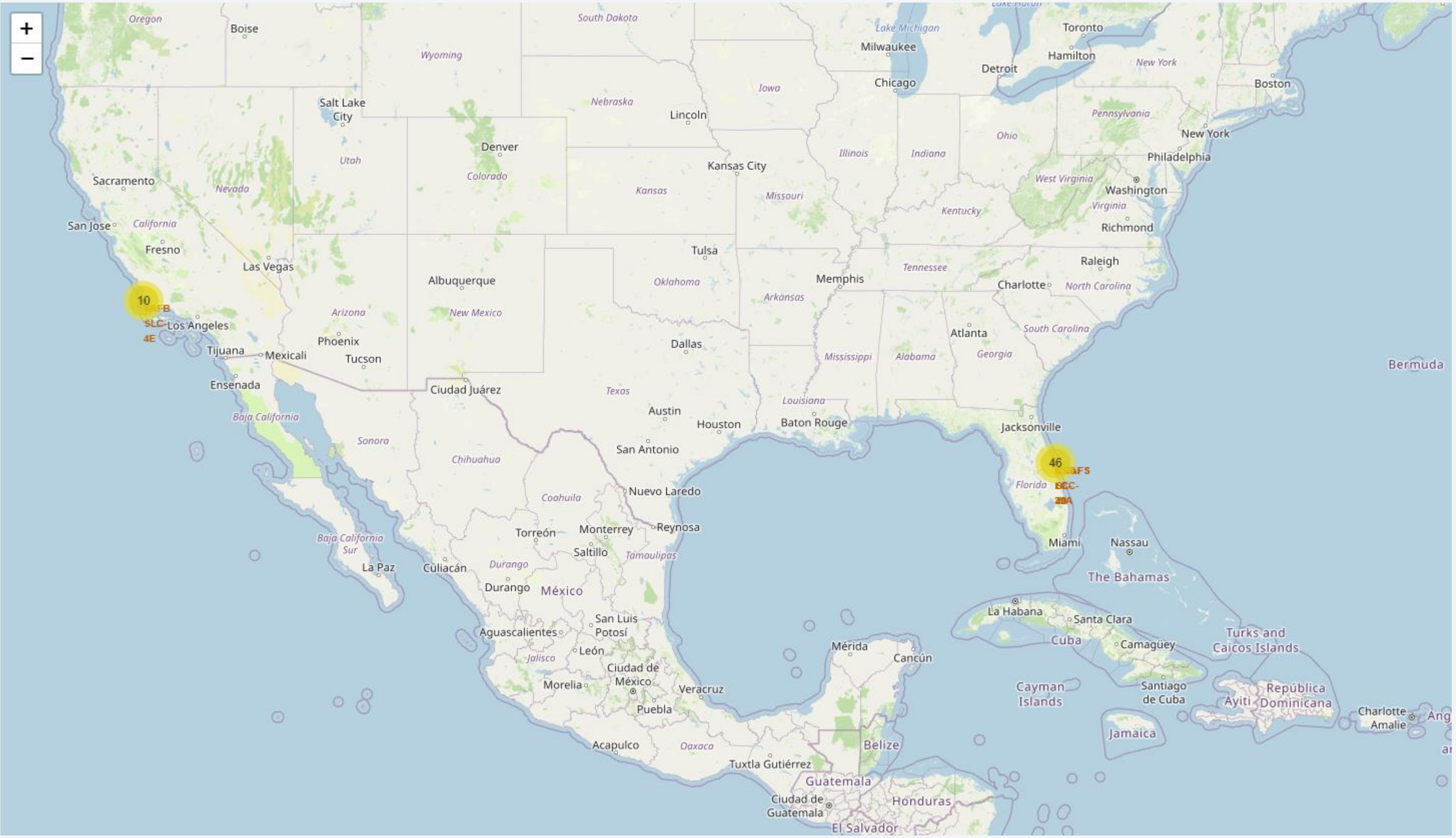
Section 3

Launch Sites Proximities Analysis

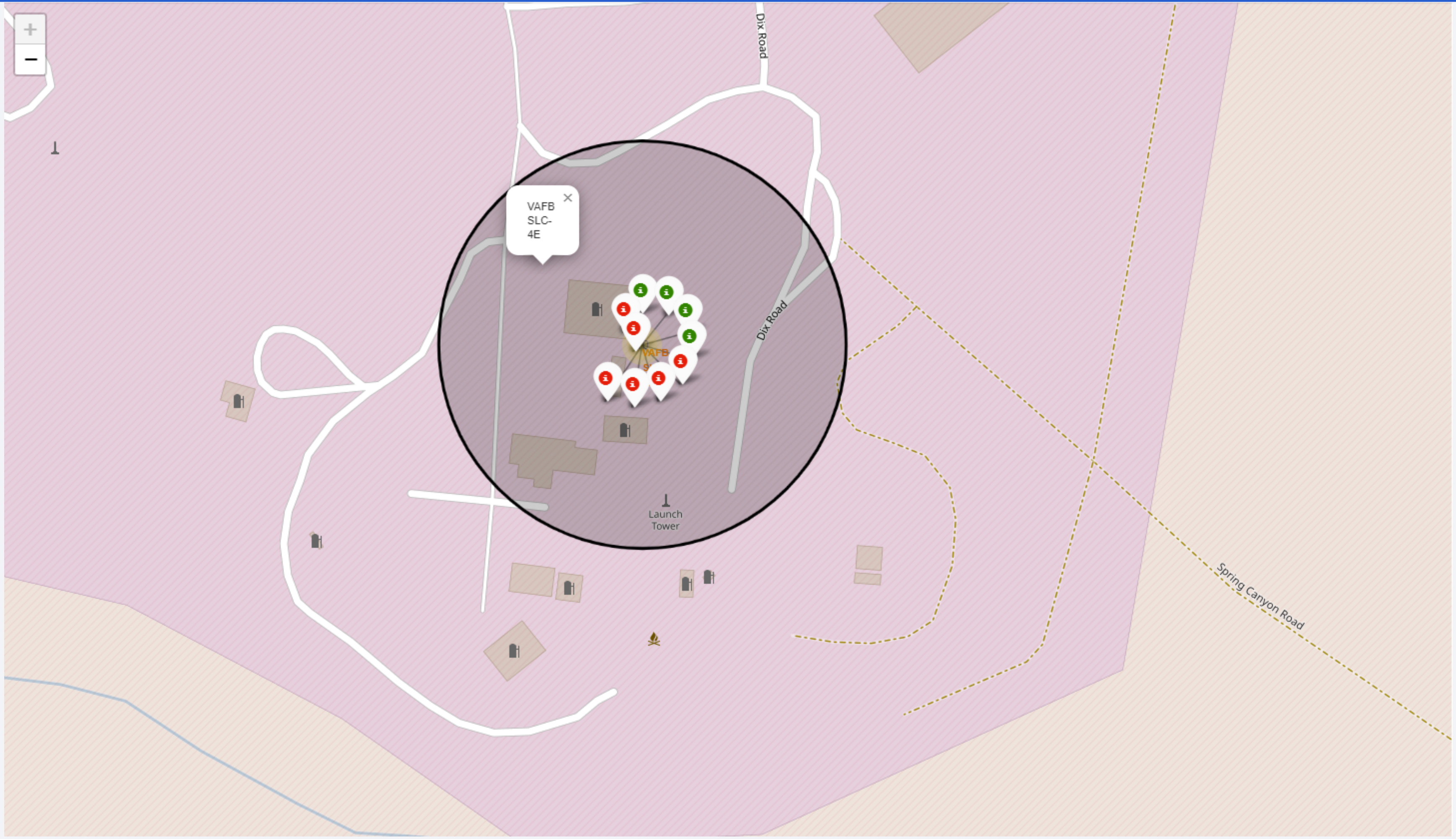
Folium Map: Launch Sites

- The concentration of launch sites in California and Florida is primarily due to their advantageous geographic positions for space launches. Florida's proximity to the equator offers a rotational speed boost, making it ideal for equatorial orbits, while California's westward launch path over the Pacific Ocean is advantageous for certain missions. Additionally, these states have a historical significance in space exploration, established regulatory support, favorable climate conditions, and strong political and economic backing, all of which have contributed to their prominence as key locations for space launch activities.

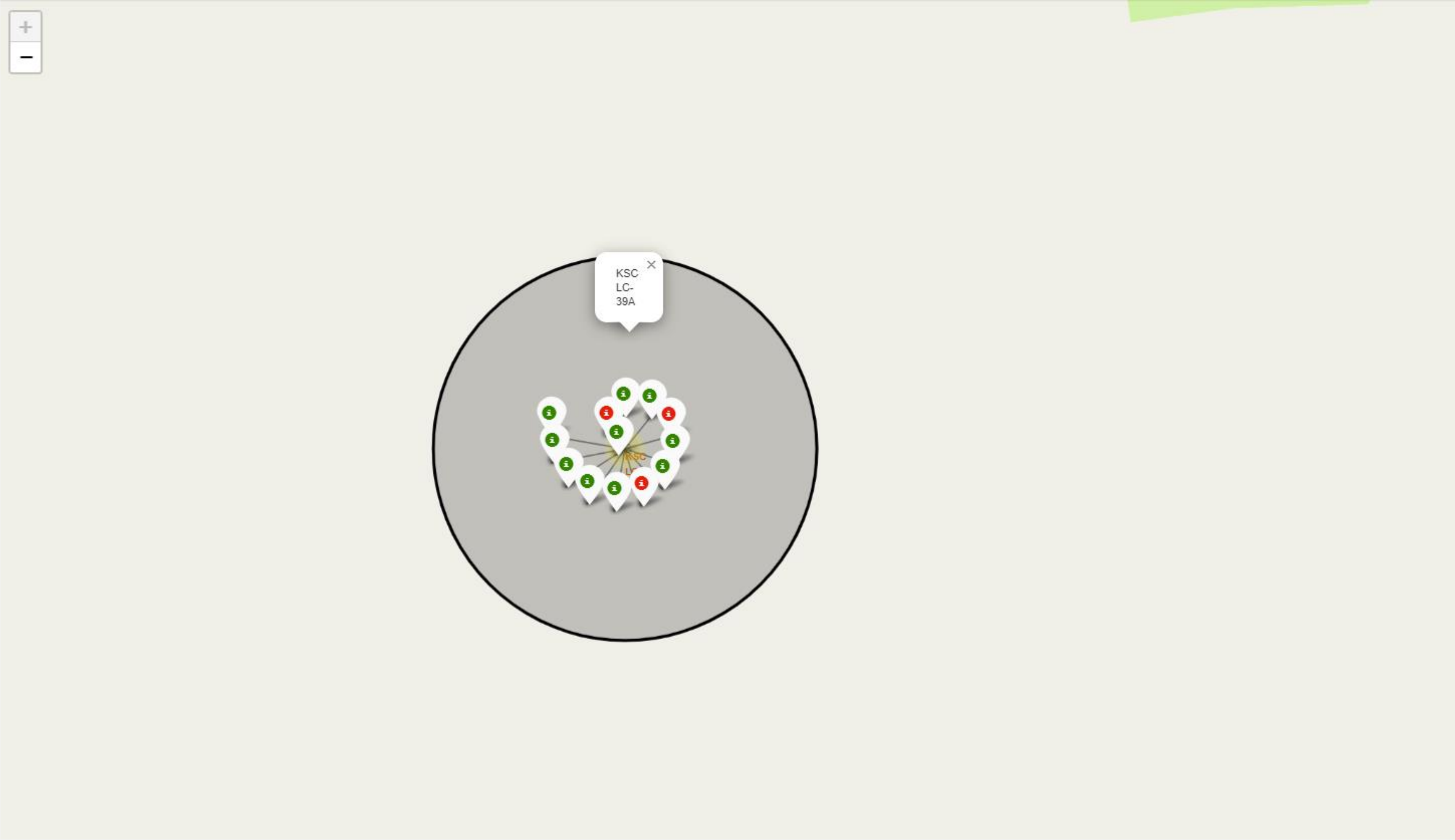
Folium Map: Launch Outcomes



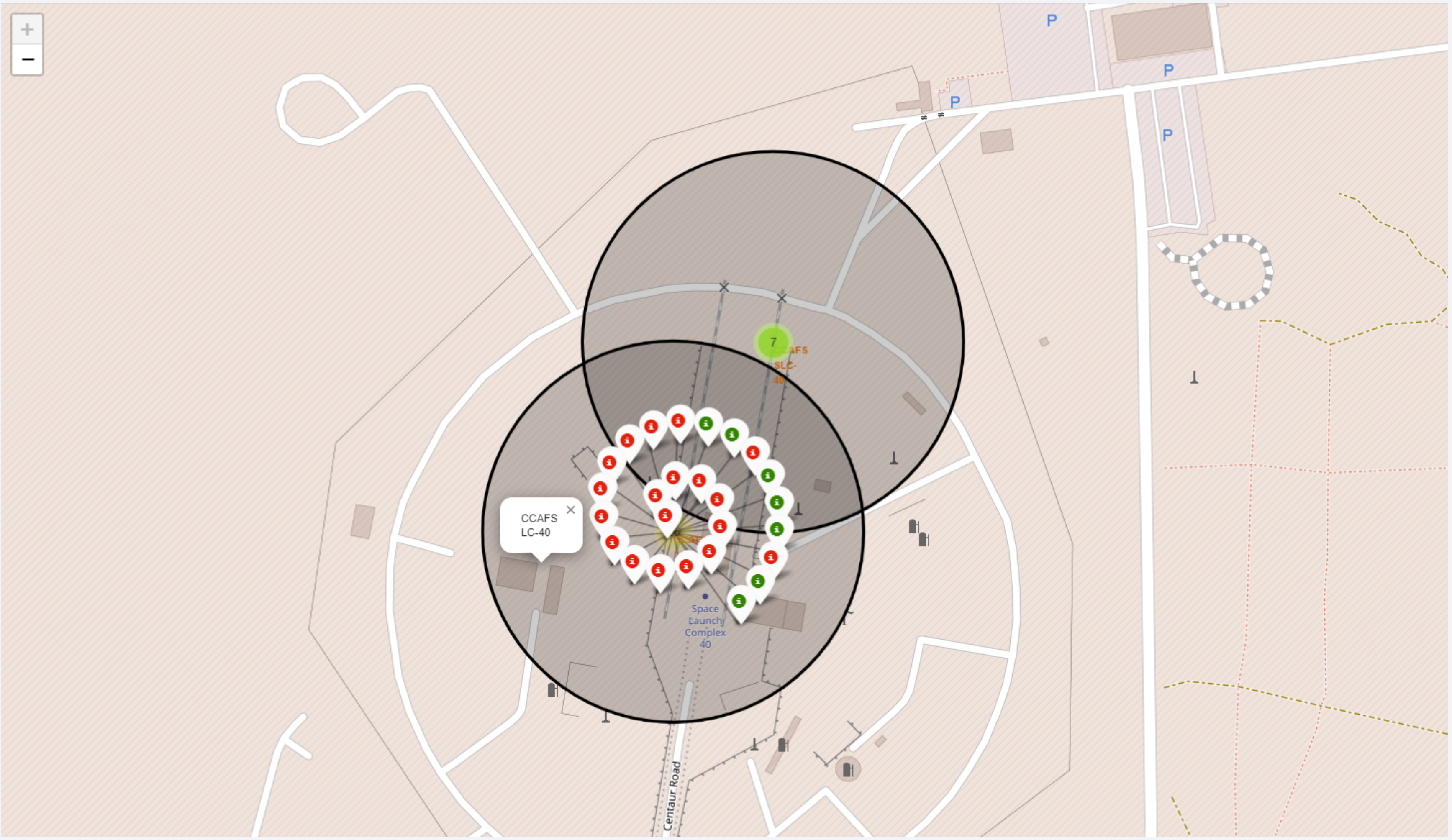
Folium Map: Launch Outcomes



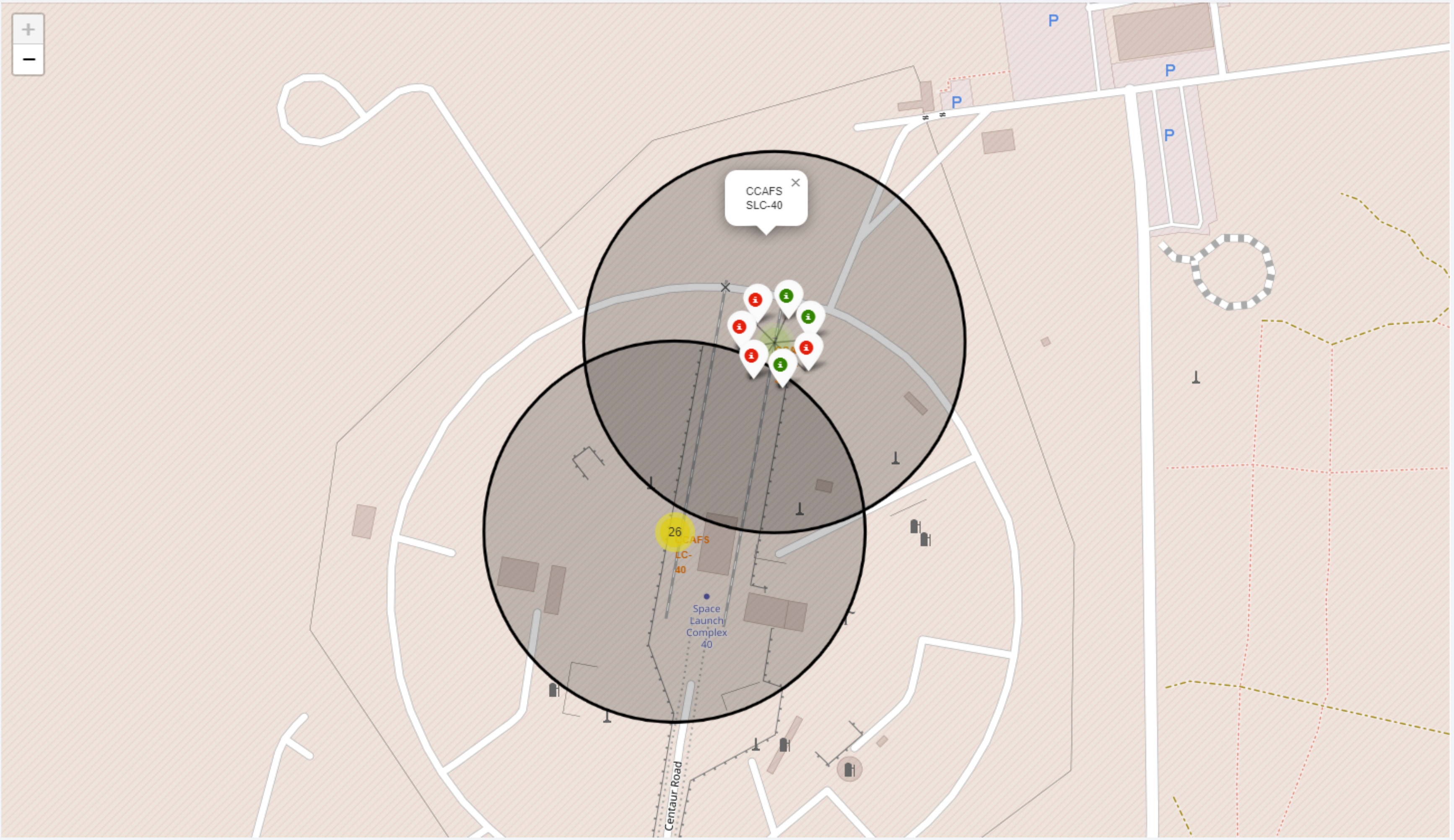
Folium Map: Launch Outcomes



Folium Map: Launch Outcomes



Folium Map: Launch Outcomes



Folium Map: Launch Outcomes

In the provided screenshots, important elements and findings are as follows:

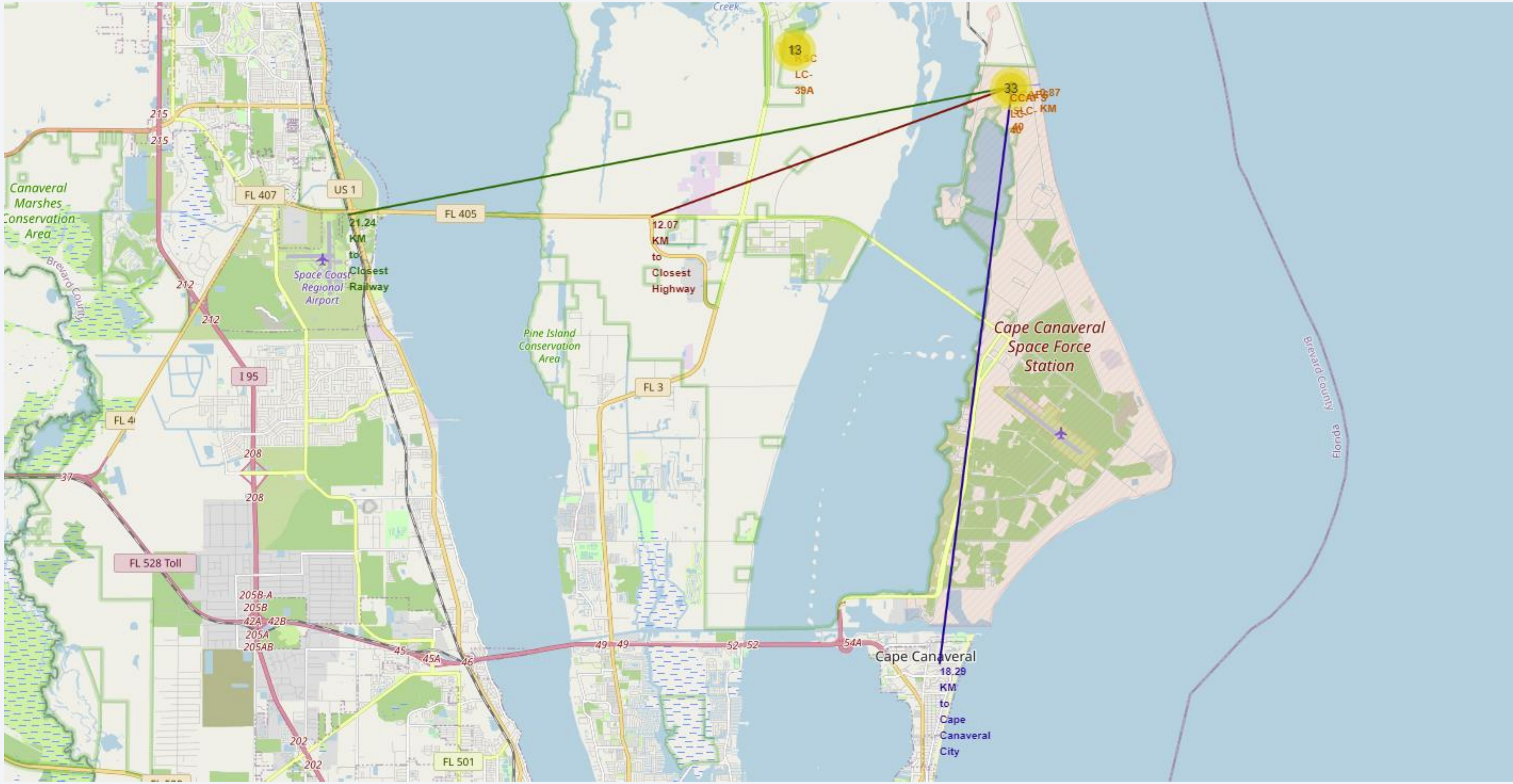
1. **Marker Colors:** The markers on the map are color-coded. Green markers represent successful launches, while red markers indicate failed launches.

2. **Launch Sites and Success Rates:**

- **VAFB SLC-4E:** This site has a 4/10 successful rate.
- **KSC LC-39A:** This site has a 10/13 successful rate
- **CCAFS LC-40:** This site has a 7/26 successful rate
- **CCAFS SLC-40:** This site has a 3/7 successful rate

These findings provide insights into the success rates of various launch sites, which is crucial information for evaluating the performance and reliability of each site in space missions.

Launch Site (CCAFS SLC-40) to Its Proximities



Launch Site (CCAFS SLC-40) to Its Proximities

Findings and Explanations:

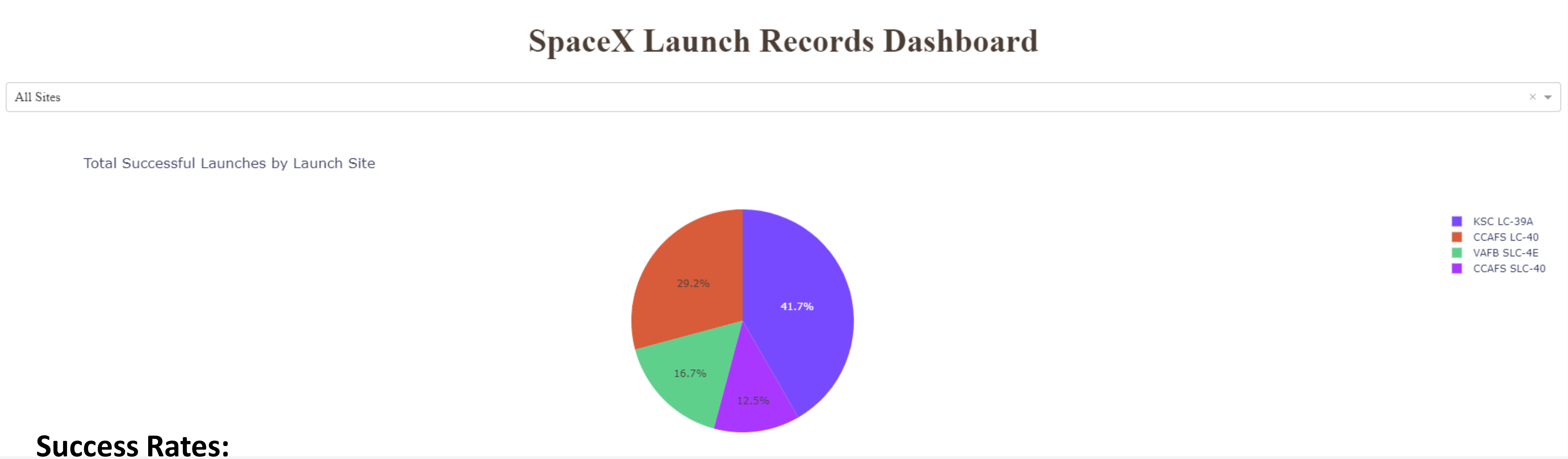
1. **Distance to Closest Railway:** The selected launch site is located approximately **21.24 kilometers** away from the nearest railway, indicating that it is not in very close proximity to railways.
2. **Distance to Closest Highway:** The launch site is situated around **12.07 kilometers** from the closest highway. While it is not directly adjacent to a highway, it is closer to a highway compared to the railway, possibly for logistical reasons, such as the transportation of equipment and personnel.
3. **Distance to Closest Coastline:** The launch site is only **0.87 kilometers** from the nearest coastline, suggesting that it is in very close proximity to the coast. This proximity to the coastline is often beneficial for rocket launches due to the ability to launch over water. Launching over water minimizes risks to populated areas in case of mission failures, and water-based launch trajectories are often favored for certain missions.
4. **Distance to Cape Canaveral City:** The distance from the launch site to Cape Canaveral City is approximately **18.29 kilometers**, indicating that the launch site is situated at a significant distance from the city. Keeping a distance from cities is a safety and operational consideration. Launch sites are usually located away from densely populated areas to minimize potential risks to the public.

- These findings reflect the deliberate selection of launch site locations based on safety, operational, and logistical considerations to ensure the success and safety of space missions.

Section 4

Build a Dashboard with Plotly Dash

Successful Launches by Site

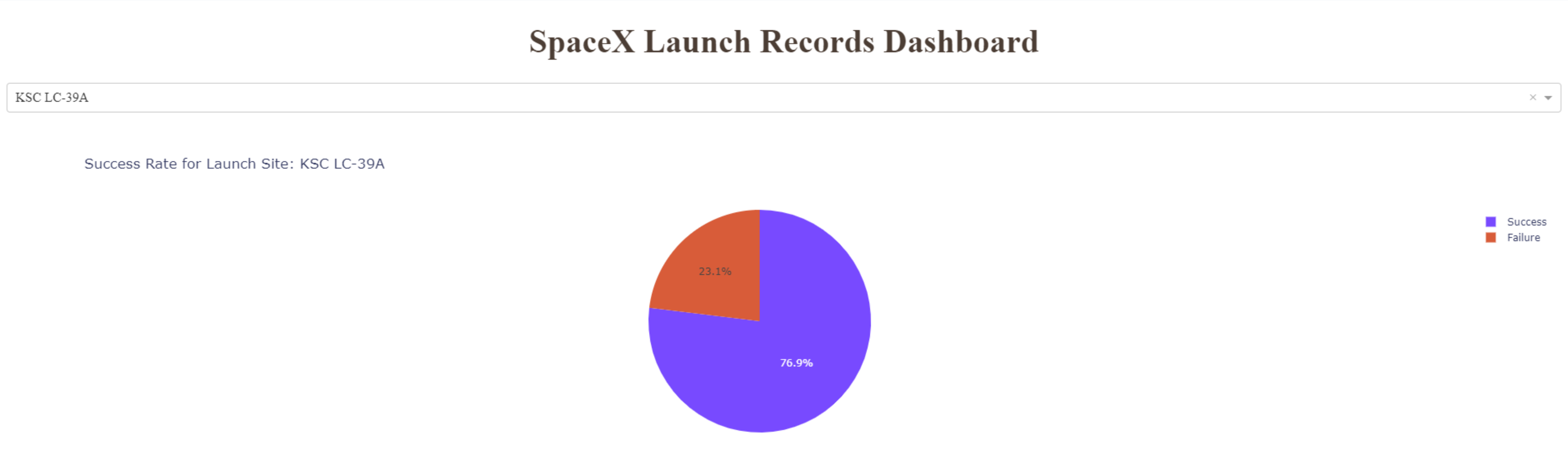


Success Rates:

- KSC LC-39A: approximately 41.7% of the missions launched from this site were successful.
- CCAFS LC-40: around 29.2% of missions launched from this location were successful.
- VAFB SLC-4E: approximately 16.7% of missions from this site achieved success.
- CCAFS SLC-40: roughly 12.5% of missions launched from this site were successful.

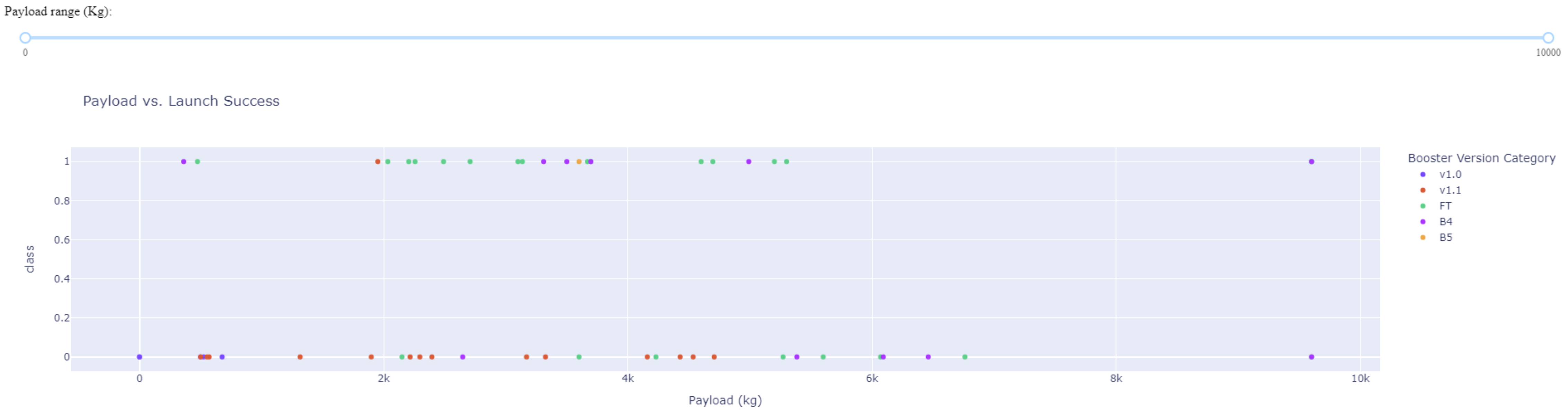
Finding: Among the mentioned launch sites, KSC LC-39A has the highest success rate at 41.7%. Therefore, KSC LC-39A has the most successful rate among these launch sites.

Most Successful Launch Site: KSC LC-29A



- The most significant finding is that KSC LC-29A stands out as the launch site with the highest success rate at 76.9%. This means that a substantial proportion of the missions launched from this site have been successful, making it the most reliable and successful launch site among those considered.

Payload vs. Launch Outcome



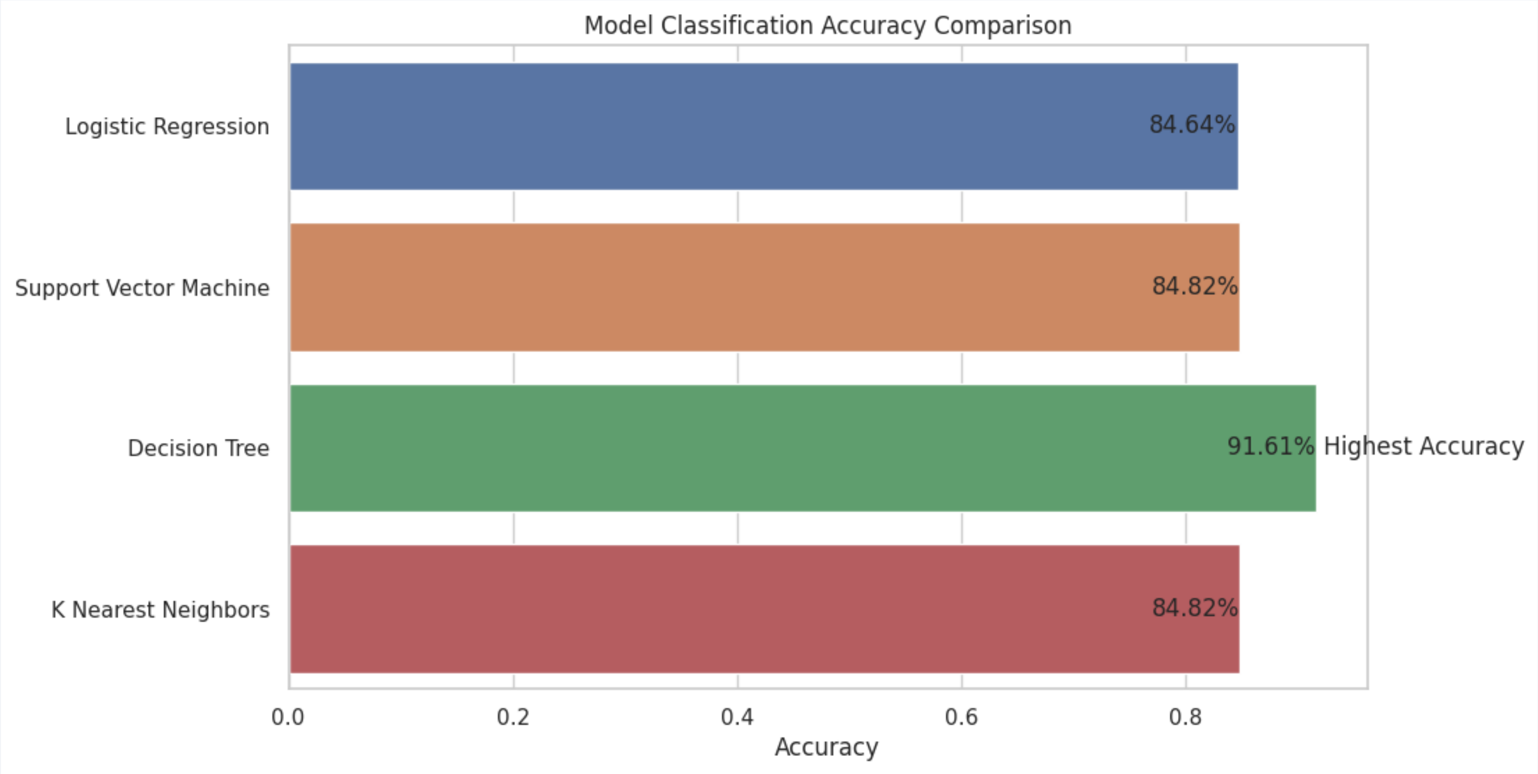
- **Payload Range:** The most successful launches are observed in the payload range between 2000 kg and 6000 kg. This payload range appears to have the largest success rate, indicating that missions with payloads in this range are more likely to succeed.
- **Booster Version:** Most successful launches are associated with the "FT" Booster Version. This suggests that the "FT" Booster Version has a high success rate.
- **Failure Launches:** Failure launches are observed across all payload ranges. However, it's notable that most of these failures are associated with the "V1.1" Booster Version. This indicates that the "V1.1" Booster Version has a higher likelihood of experiencing mission failures compared to other versions.

These insights can be valuable for assessing the reliability and performance of different payload ranges and booster versions in space missions.

Section 5

Predictive Analysis (Classification)

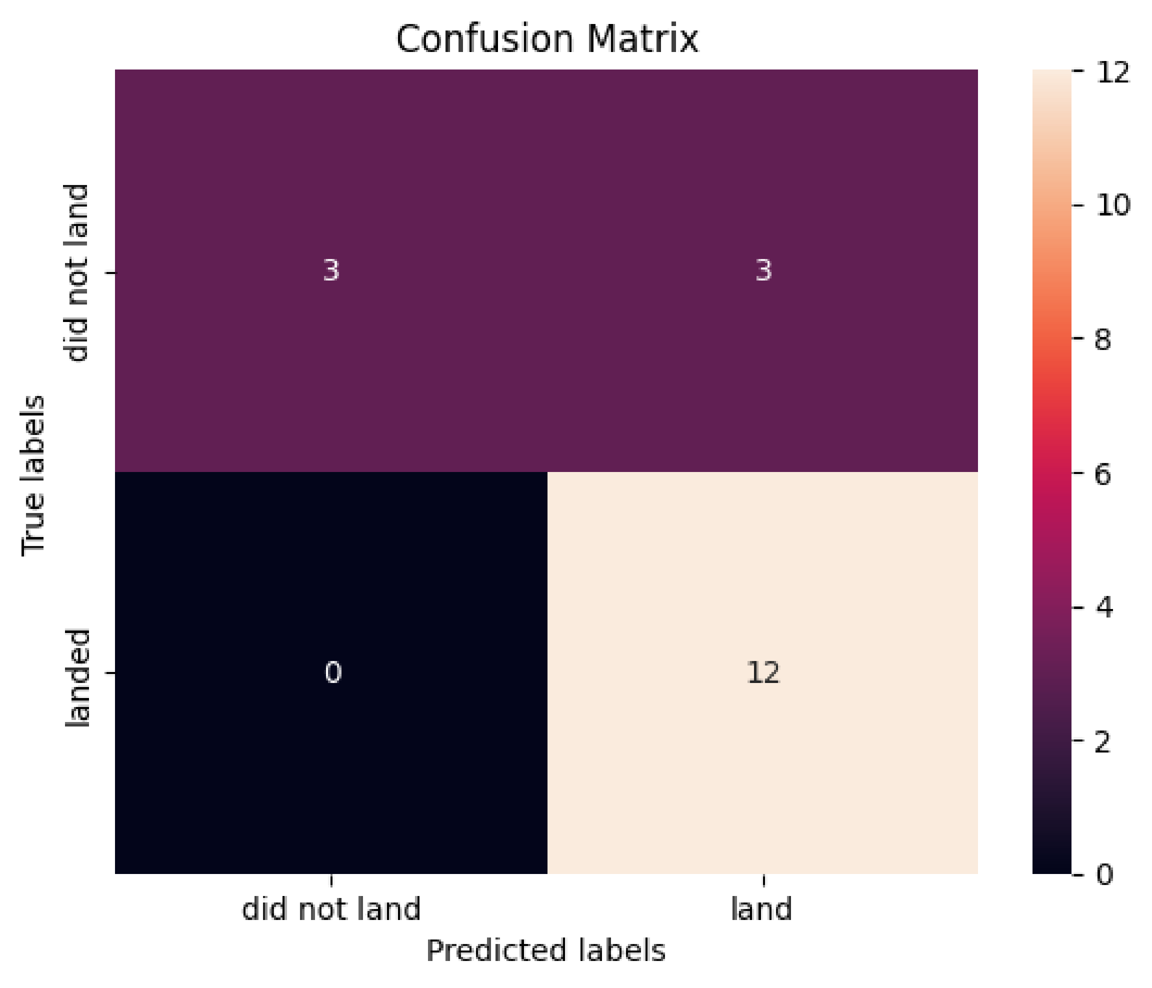
Classification Accuracy



- Based on the accuracy results for the built classification models: The Decision Tree model has the highest classification accuracy, with an accuracy of 0.9161. Therefore, the Decision Tree is the best-performing method among the models tested.

Confusion Matrix

- In the best-performing Decision Tree model, it made 3 incorrect predictions of "land" when the actual outcome was "did not land" (False Positives), and it made 0 incorrect predictions of "did not land" when the actual outcome was "land" (False Negatives). Additionally, it correctly predicted 3 "did not land" outcomes and 12 "land" outcomes. The confusion matrix provides a comprehensive view of the model's performance in distinguishing between the two classes.



Conclusions

- Our analysis demonstrates the significance of harnessing data science and machine learning to forecast the success of Falcon 9 first-stage landings during SpaceX rocket launches.
- The success rate of Falcon 9 first-stage landings stands at 66.67%, underscoring the reliability of SpaceX's reusability model.
- Key factors affecting landing success have been identified, such as launch site, payload mass, orbit type, and flight number, providing essential insights for decision-making in the aerospace industry.
- Our exploratory data analysis revealed variations in success rates across different launch sites, correlations between payload mass and launch success, and a consistent increase in success rates over time.
- The interactive analytics demo showcased strategic aspects of SpaceX's launch sites, highlighting proximity to the equator and safety considerations.
- We have developed predictive models, with the Decision Tree model standing out as the best-performing, achieving an accuracy of 91.61% on the test data.
- These insights empower aerospace industry decisions, optimize success, and reduce launch costs.

Appendix

1. **Recommended Reading:**
   - "Data Science for Beginners" by John Smith
   - https://sma.nasa.gov/LaunchVehicle/assets/spacex-falcon-9-data-sheet.pdf
2. **Data Sources and Tools:**

- Data Sources:
  - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - SpaceX API
- Analysis Tools:
  - Python with libraries such as Pandas, NumPy, and Scikit-Learn
  - SQL for data querying and analysis
  - Folium for interactive geospatial data visualization
  - Plotly Dash for dynamic data visualization

3. **Code Repository:** https://github.com/hai-t-nguyen/IBM-Data-Science-Capstone-SpaceX

Thank you!