IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abdullah Aloyaydi
07-10-2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- **Project Background and Context**

SpaceX offers Falcon 9 rocket launches at a cost of $62 million, significantly lower than other providers, which charge over $165 million. This cost advantage comes from the ability to reuse the first stage of the rocket. By predicting whether the first stage will land successfully, we can better understand launch costs and help other companies compete with SpaceX.

- **Problems to Address**

- What factors influence the success of rocket landings?
- How do different features interact to affect landing success rates?
- What conditions are necessary for a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia
- Perform data wrangling
  - The code was applied with the identification of the label for the supervised model training and determination of the augmented success in landing.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data Collection Methods:

- The data was collected through a **GET request** using the SpaceX API ("https://api.spacexdata.com/v4/"). Specific values such as "rocket," "payloads," etc., were requested for each API call. After retrieval, the data was cleaned and processed.
- **Additionally, web scraping** was used to collect data from Wikipedia's "List of Falcon 9 and Falcon Heavy launches" page using **BeautifulSoup4**.
- The table headers were scraped to extract the column names.
- The launch data was scraped from the table rows and converted into a **pandas DataFrame**.
- The final DataFrame was exported as spacex_web_scraped.csv for further analysis.

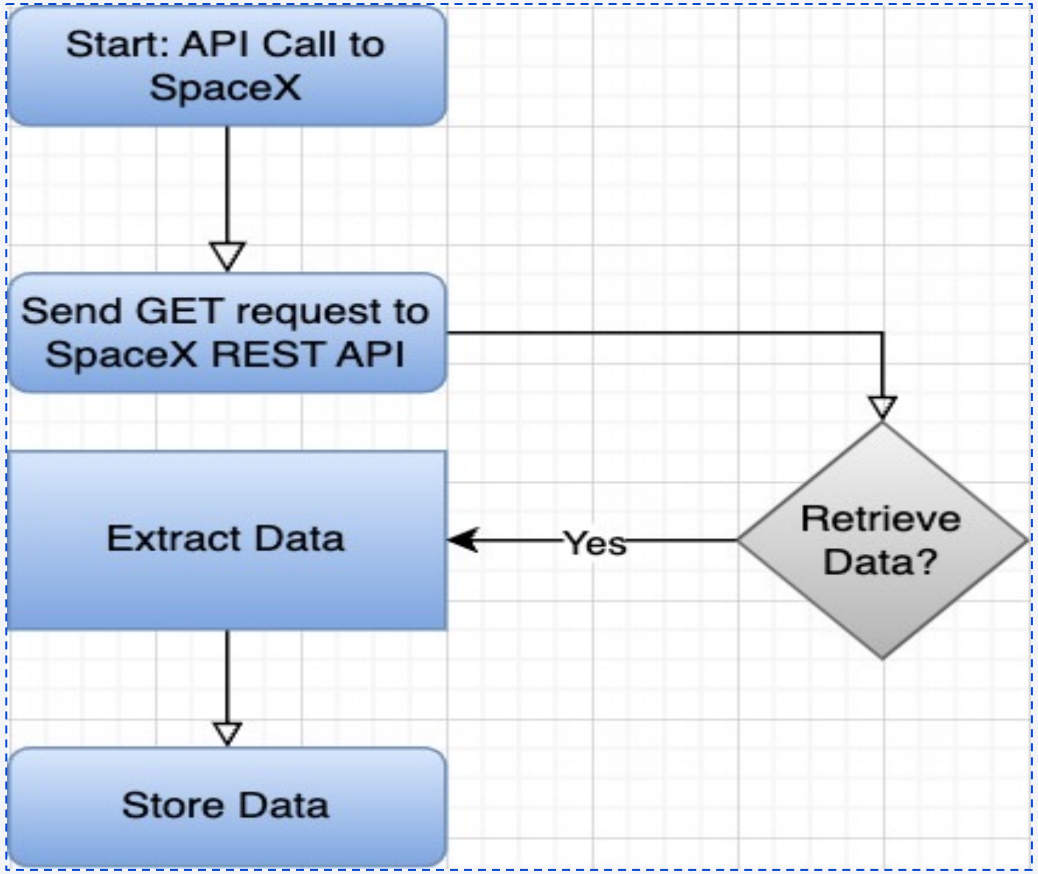
Data Collection – SpaceX API

- **Data Collection Overview:**

The data was collected through SpaceX REST API calls. Key data points such as "rocket," "payloads," and other values were retrieved using GET requests.

- **GitHub URL :**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/DataCollectionAndWebscarbaing/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

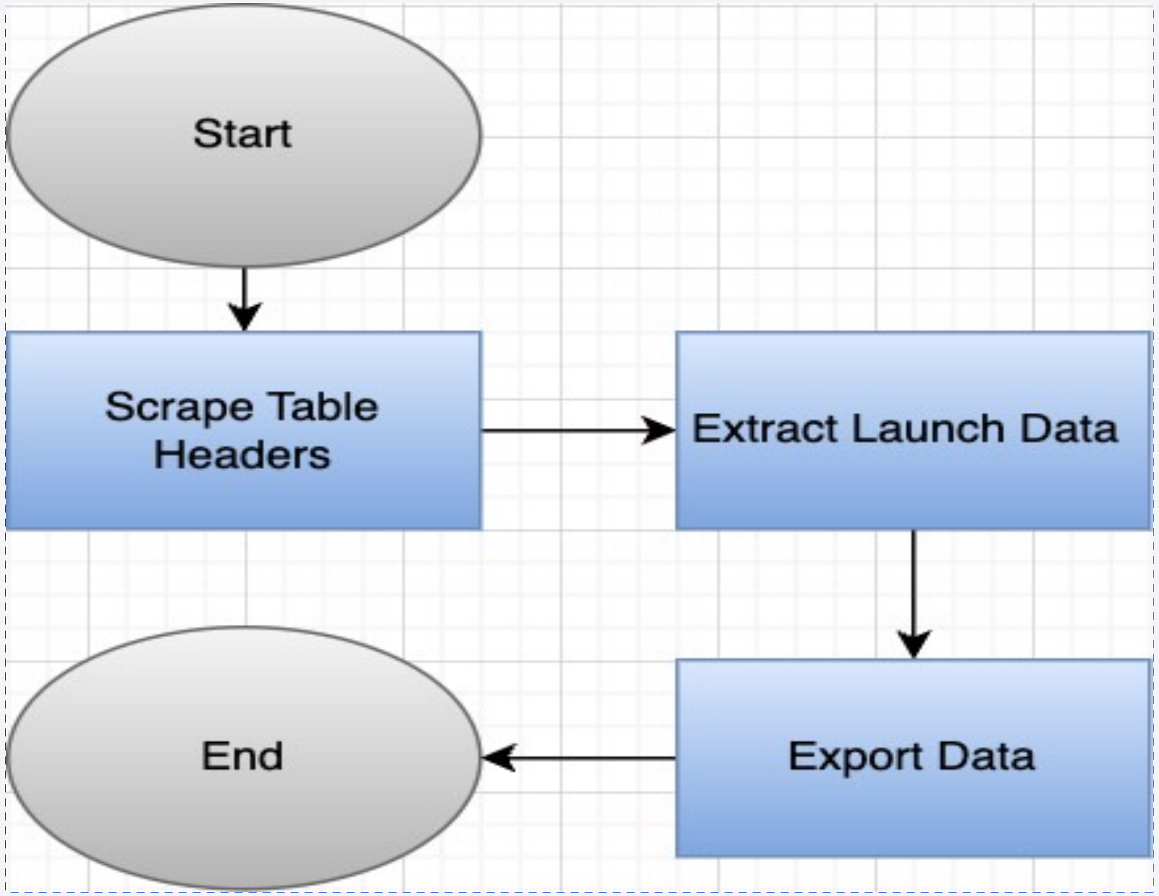
- **Web Scraping Overview:**

Data was collected from Wikipedia’s "List of Falcon 9 and Falcon Heavy launches" page using BeautifulSoup4. The key steps included:

- **Scraping table headers for column names.**
- **Extracting launch data into a pandas DataFrame.**
- **Saving the data as spacex_web_scraped.csv for analysis.**

- **GitHub URL :**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/DataCollectionAndWebscarbaing/jupyter-labs-webscraping.ipynb



Data Wrangling

- **Data Processing Overview:**
  - Data was processed to prepare it for analysis and model training.
- **Key Steps in Data Wrangling:**
  - **Launch Counts:** Counted the number of launches from each site.
  - **Orbit Analysis:** Calculated the occurrence of each orbit.
  - **Mission Outcomes:** Analyzed the mission outcomes for different orbits.
  - **Landing Labels:** Created labels for landing outcomes:
    - 1 = Successful landing
    - 0 = Unsuccessful landing
- **GitHub URL :**

  https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/DataWrangling/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Overview:**
  This section highlights the key visualizations created during the Exploratory Data Analysis (EDA) phase, which are essential for understanding patterns in the SpaceX launch data.

- **Visualizations:**

  Flight Number vs. Launch Site: Shows how flight numbers vary by launch site.

  Payload Mass vs. Launch Site: Compares payload mass across different launch sites.

  Success Rate by Orbit Type: Analyzes success rates for various orbit types.

  Flight Number vs. Orbit Type: Displays the distribution of flight numbers by orbit type.

  Payload Mass vs. Orbit Type: Examines the relationship between payload mass and orbit type.

  Yearly Launch Success Trend: Tracks trends in successful launches over the years.

  Dummy Variables: Converts categorical data into numerical form for analysis.

  Casting Numeric Columns: Ensures numeric columns are properly formatted for analysis.

These visualizations provide key insights that inform the overall analysis and predictive modeling.

- **GitHub URL:**

  https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/ExploringDataAnalysis/edadataviz.ipynb

EDA with SQL

- **Overview:**
  - Summary of SQL queries used for data analysis.
- **SQL Queries Performed:**
  - **Unique Launch Sites:**Retrieved names of distinct launch sites in space missions.
  - **Launch Sites Starting with 'CCA':**Displayed 5 records of launch sites that begin with 'CCA'.
  - **Total Payload Mass for NASA (CRS):**Calculated total payload mass for boosters launched by NASA.
  - **Average Payload Mass for F9 v1.1:**Computed average payload mass for Falcon 9 version 1.1 boosters.
  - **First Successful Ground Pad Landing:**Listed the date of the first successful landing on a ground pad.
  - **Boosters Successful on Drone Ship:**Identified boosters that successfully landed on drone ships with payload mass between 4000 and 6000.
  - **Total Successful and Failed Missions:**Summarized total successful and failed mission outcomes.
  - **Max Payload Mass Booster Versions:**Listed booster versions that carried the maximum payload mass using a subquery.
  - **Failure Landing Outcomes in 2015:**Retrieved records of failed landings on drone ships in 2015, including month names and booster versions.
  - **Ranking Landing Outcomes:**Ranked landing outcomes from 2010-06-04 to 2017-03-20 in descending order.
- **GitHub URL :**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/ExploringDataAnalysis/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Map Elements Added:**

  - **Launch Site Markers:** Plotted all launch sites on the map.
  - **Success/Failure Markers:** Marked successful and failed launches at each site.
  - **Distance Circles:** Displayed distances between launch sites and nearby locations.

Why These Elements?

  - **Markers:** Help identify launch sites and outcomes visually.
  - **Distance Circles:** Show the proximity of important locations around launch sites.

- **GitHub URL:**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/InteractiveVisualanalytics/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Key Features We Built:**
  - **-Launch Site Dropdown:** Allows users to filter launches by site or view all launches.
  - **-Success Pie Chart:** Displays total successful launches or success vs.failure for a selected site.
  - **-Payload Range Slider:** Enables users to adjust data based on payload weight.
  - **-Scatter Plot (Payload vs. Success):** Shows how payload impacts launch success rates.

- **Why We Built These Features:**
  - -Interactive controls let users explore the data easily.
  - -Charts provide quick insights into launch outcomes and payload effects.

- **GitHub URL:**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/InteractiveVisualanalytics/spacex_dash_app.py

Predictive Analysis (Classification)

- In this section, we prepared the data by creating a target variable and standardizing it. We then split the data into training and testing sets. Several classification models were built, including Logistic Regression, SVM, Decision Tree, and KNN. Each model's parameters were optimized using GridSearchCV to identify the best configurations. Finally, we evaluated the accuracy of each model on the test data to select the best performer.

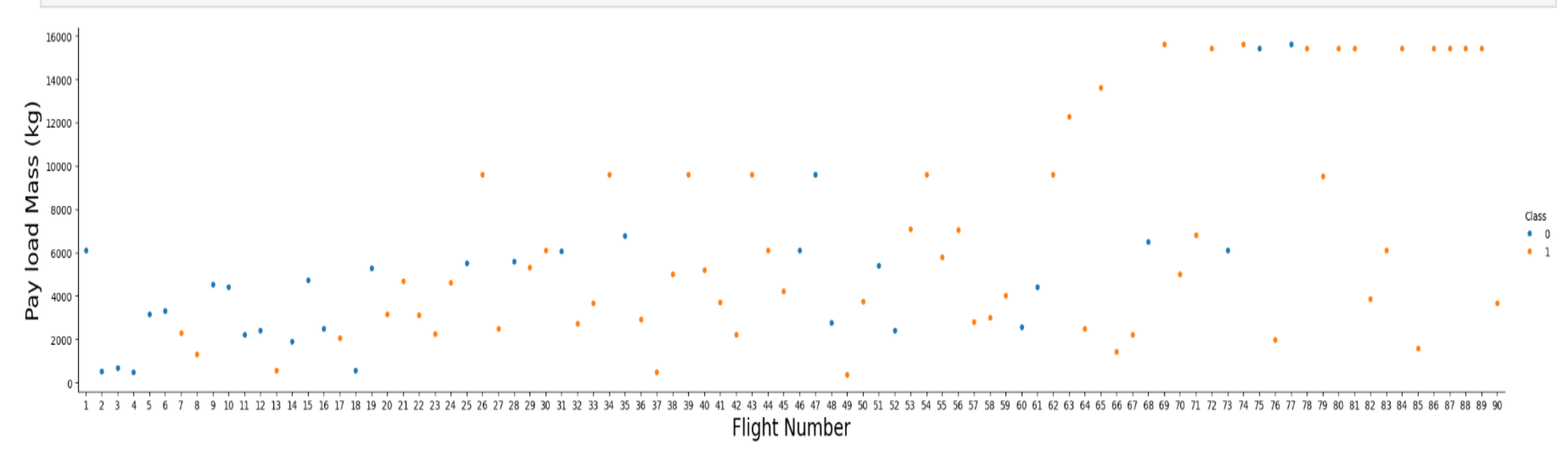
- **GitHub URL:**

https://github.com/AbdullahAloyaydi/LAbCapstone/blob/main/PredictiveAnalysis/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- **Exploratory Data Analysis Results:**

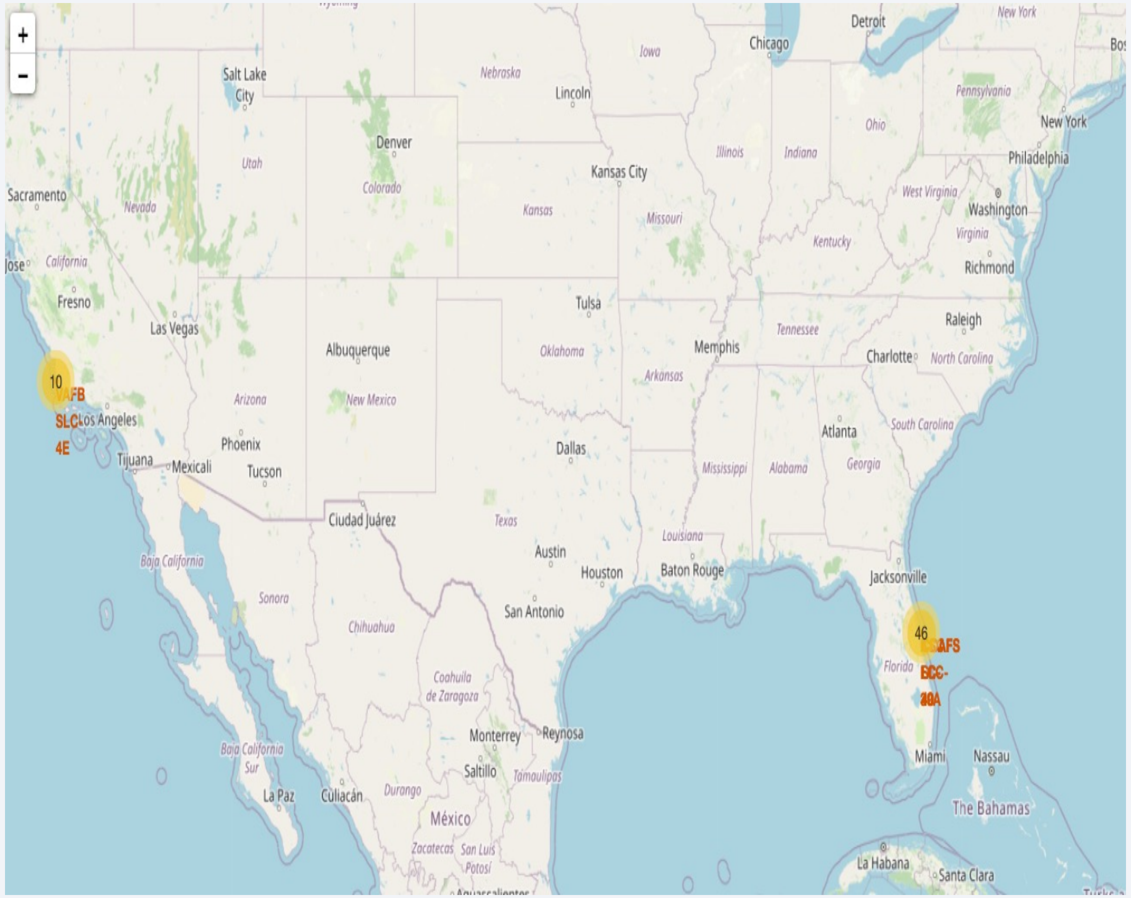
As the Flight Number increases, the likelihood of successful landings improves, while heavier payloads do not significantly hinder the success of first-stage landings. This relationship is illustrated in the accompanying screenshot of the plot showing Flight Number vs. Payload Mass overlaid with launch outcomes.



Results

- Interactive Analytics Demo (Screenshots):

The launch success rate is influenced by factors like payload mass, orbit type, and launch site location. In this analysis, we used Folium to create interactive maps. We marked all launch sites, highlighted successful and failed launches, and calculated distances to nearby areas. The screenshots showcase these visualizations and reveal geographical patterns that help us understand the performance of each launch site.



- Predictive Analysis Results:

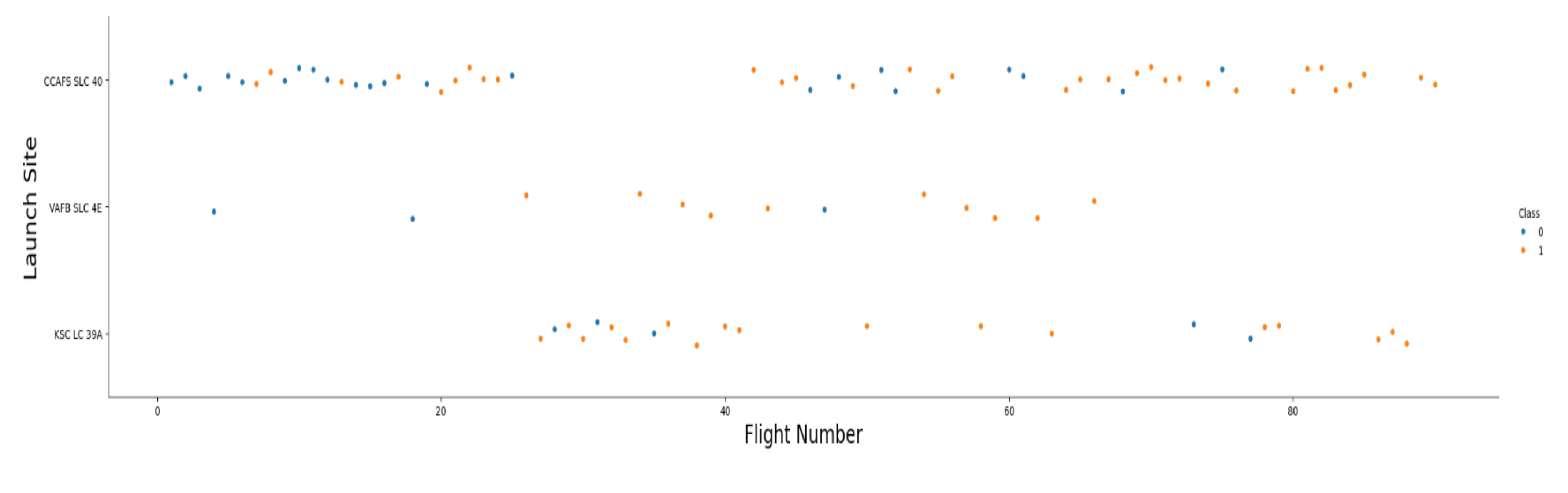
We built a machine learning model to predict whether the Falcon 9 first stage would land successfully, which helps determine launch costs. After analyzing the data and creating a target class for success, we standardized the information and split it into training and testing sets. We then compared different models, including Support Vector Machines, Classification Trees, and Logistic Regression.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

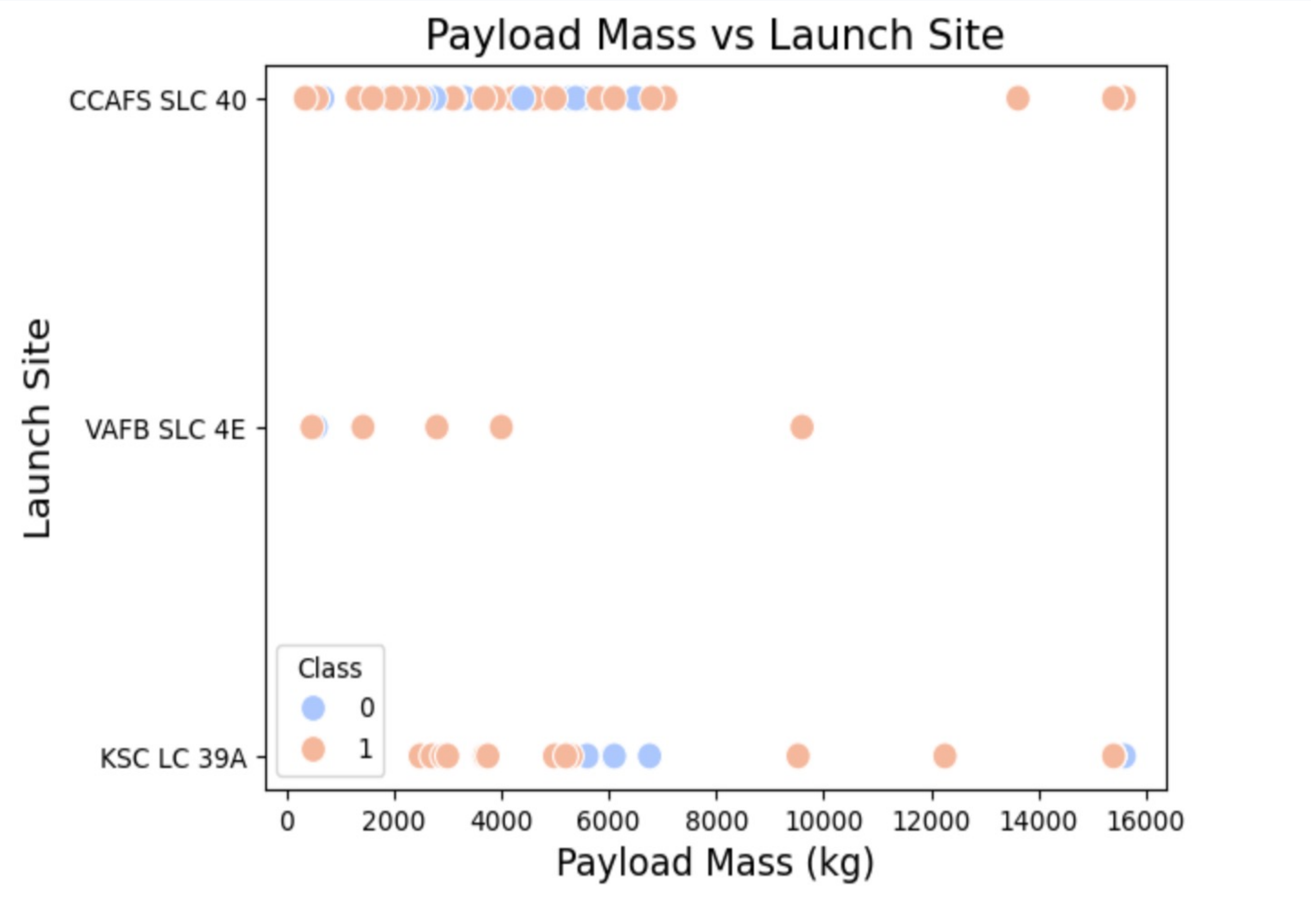
- This plot shows the relationship between the number of flights and each launch site. It reveals that as the number of flights increases at a launch site, the success rate tends to improve.



Payload vs. Launch Site

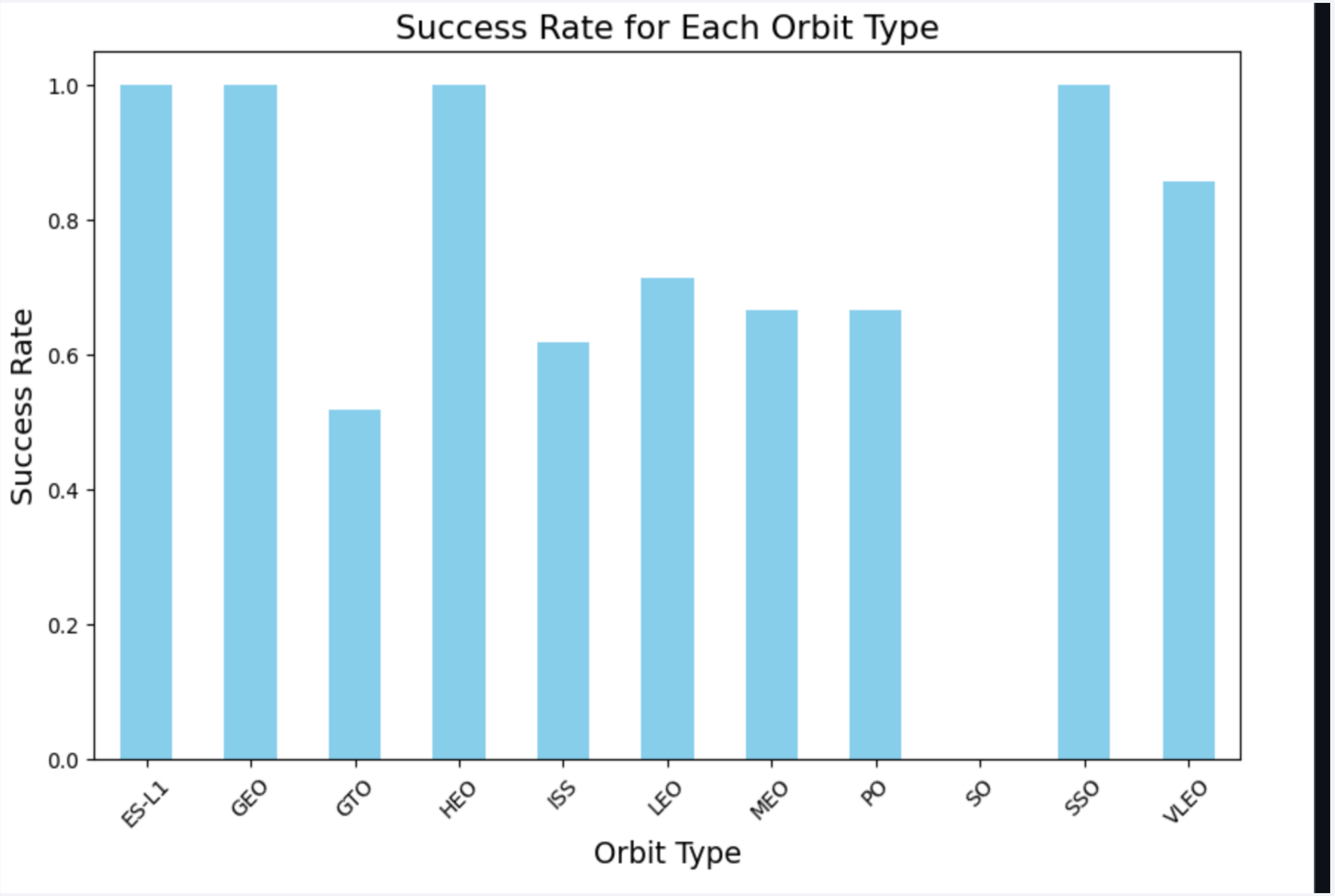
- **Scatter Plot:**

This scatter plot shows the payload mass for each launch at different SpaceX launch sites. Each point represents a launch, helping us see if there's any pattern between payload size and launch location.



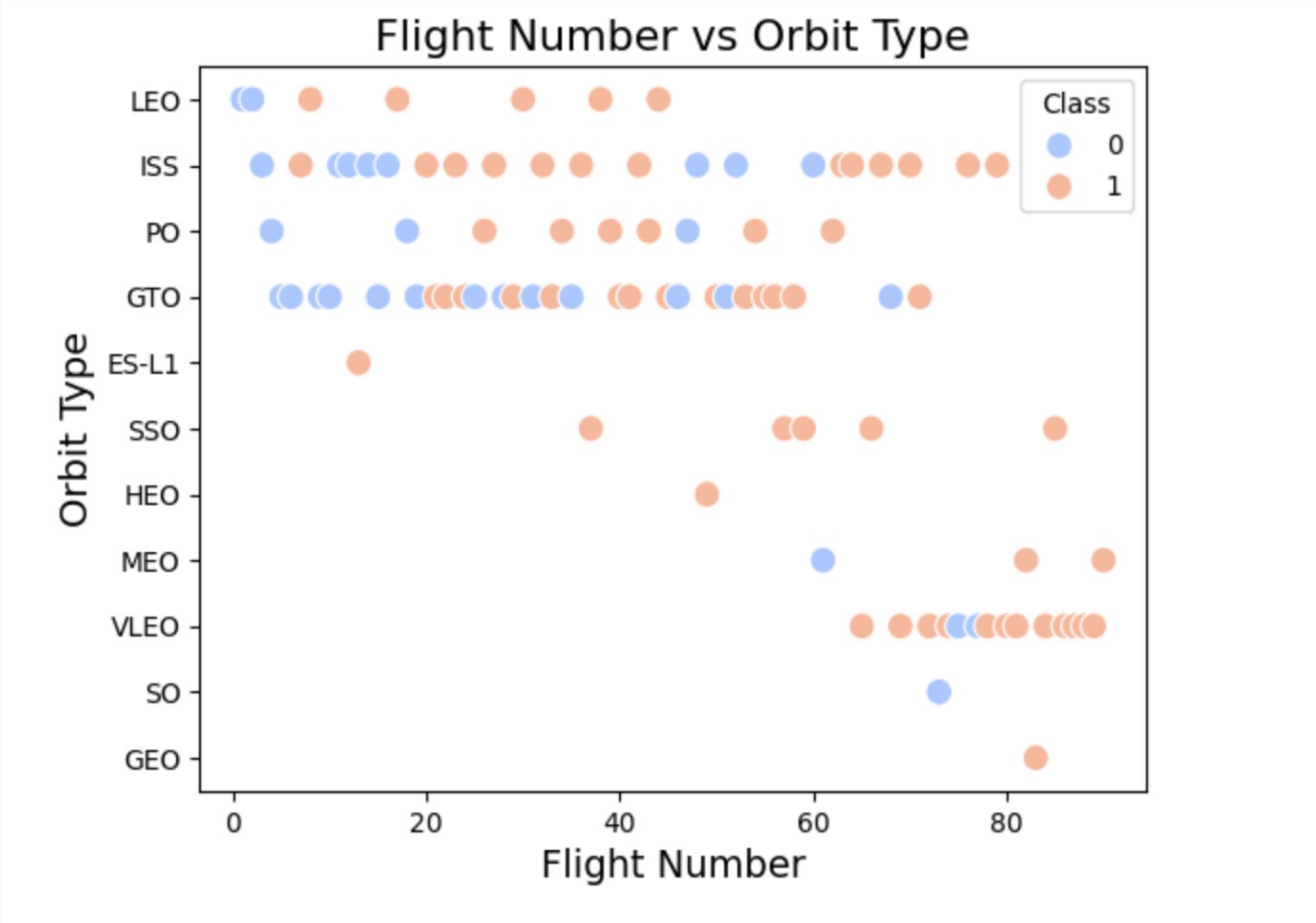
Success Rate vs. Orbit Type

- This bar chart illustrates the success rates for different orbit types. It helps identify which orbit types have higher success rates, offering insights into how orbit selection might influence mission outcomes.



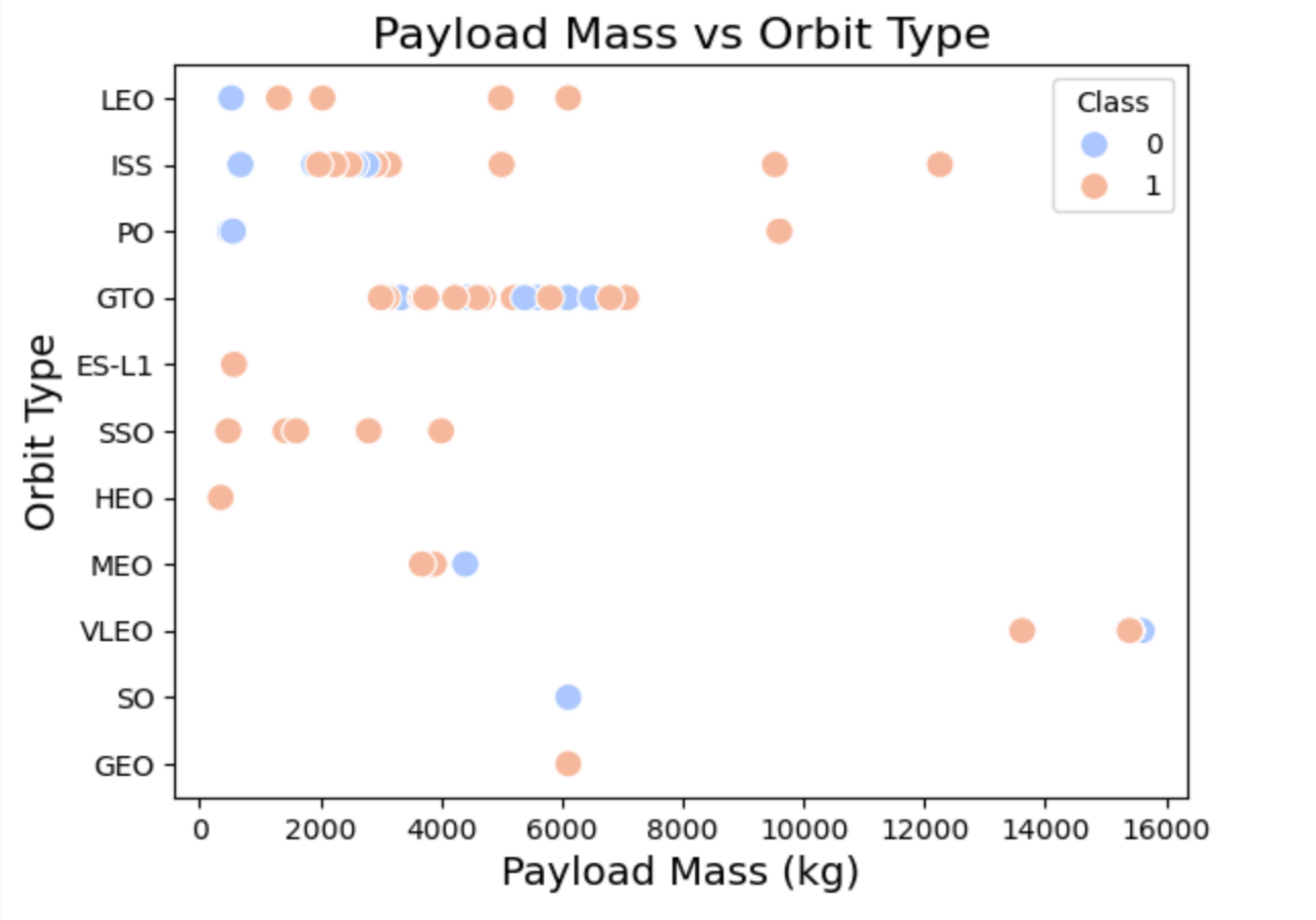
Flight Number vs. Orbit Type

- This scatter plot shows the relationship between flight numbers and orbit types. It highlights how different orbit types are distributed across varying flight numbers, giving insight into how orbit type might impact the frequency or success of launches.



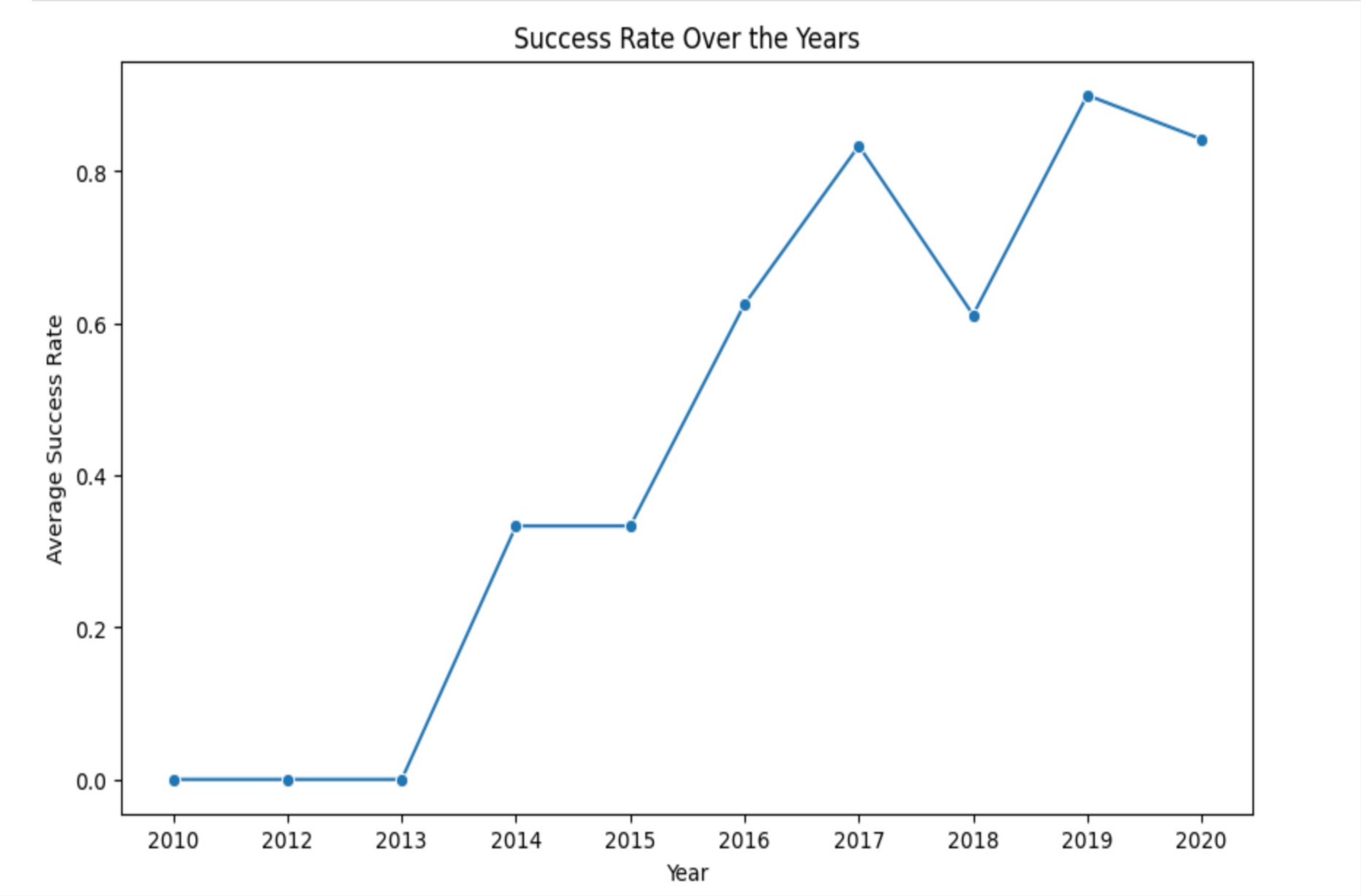
Payload vs. Orbit Type

- This scatter plot displays the relationship between payload mass and orbit type. It helps to identify how payload mass is distributed across different orbit types, which can reveal trends in the types of payloads launched to specific orbits.



Launch Success Yearly Trend

- This line chart shows the yearly average success rate of launches over time. It illustrates how SpaceX's launch success rate has improved year by year, reflecting the company's advancements in rocket technology and landing accuracy.



All Launch Site Names

- We retrieved the unique names of all launch sites in the SpaceX dataset. This provides an overview of the various locations used for Falcon 9 launches, helping us understand the geographical spread of launch activities.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
# unique_launch_sites = pd.read_sql(query, con)

# # Display the result
# print(unique_launch_sites)
```

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

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

Launch Site Names Begin with 'CCA'

- This query retrieves the first five records of launch sites whose names begin with "CCA." This helps identify specific launch locations associated with the "CCA" prefix, providing insight into regional launch activities.

```
%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

Total Payload Mass

- We calculated the total payload mass carried by boosters launched by NASA under the Commercial Resupply Services (CRS) contracts. This gives insights into the overall contribution of NASA's missions in terms of payload mass over time.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';
```

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

total_payload_mass
48213

Average Payload Mass by F9 v1.1

- calculates the average payload mass carried by the Falcon 9 booster version F9 v1.1. By filtering the records to only include this specific booster version, the result provides valuable insights into its typical payload capacity, which can inform future mission planning and capacity assessments.

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

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

average_payload_mass
2928.4

First Successful Ground Landing Date

- This query uses the MIN function on the Date column to find the earliest successful landing date for ground pad landings. The result provides valuable insight into the history of SpaceX's landing achievements, showcasing the advancements in their landing technology over time.

```
%sql SELECT MIN(Date) AS first_successful_landing_date FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE '%Success%Ground%';
```

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

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This analysis identifies boosters that successfully landed on a drone ship with a payload mass between 4000 kg and 6000 kg. The SQL query executed retrieves unique booster versions that meet these criteria, helping us understand which boosters have the capability to handle significant payloads while achieving successful landings at sea.

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Successful Landing (Drone Ship)'
```

```
* sqlite:///my_data1.db
(sqlite3.OperationalError) no such column: PAYLOAD_MASS_KG
[SQL: SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Successful Landing (Drone Ship)'
AND PAYLOAD_MASS_KG > 4000 AND PAYLOAD_MASS_KG < 6000;]
```

Total Number of Successful and Failure Mission Outcomes

- This analysis calculates the total number of successful and failed mission outcomes based on the landing outcomes recorded in the dataset. The SQL query groups the landing outcomes and counts the occurrences of each, providing insights into the overall performance of missions. This information is crucial for understanding the reliability and effectiveness of the launch operations.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS total_outcomes FROM SPACEXTABLE GROUP BY "Landing_Outcome";
```

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

Landing_Outcome	total_outcomes
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

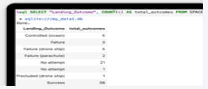
Boosters Carried Maximum Payload

- This analysis identifies the names of the boosters that have carried the maximum payload mass recorded in the dataset. By using a subquery to find the maximum payload mass, the SQL query retrieves the booster versions that achieved this feat. This information highlights the capabilities of different boosters and can inform future decisions regarding payload transportation.

```
%sql SELECT "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



2015 Launch Records

- This analysis focuses on the launch records from the year 2015, specifically identifying the failed landing outcomes on drone ships. The SQL query retrieves the relevant details, including the landing outcome, booster versions, and launch site names for each failed attempt during this period. Understanding these failures helps assess the challenges faced by SpaceX in their launch operations and provides insights for future improvements.

```
%sql SELECT Landing_Outcome,Booster_Version,Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Failure (Drone Ship)' AND D
```

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

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

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

- This analysis ranks the landing outcomes for SpaceX launches within the specified date range, from June 4, 2010, to March 20, 2017. The SQL query retrieves the count of each landing outcome type, including successes and failures, and organizes the results in descending order based on the count. This ranking provides insights into the overall success rates and challenges faced by SpaceX during this period, allowing for a better understanding of their operational performance.

```
%sql SELECT "Landing_Outcome", COUNT(*) AS outcome_count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP
```

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

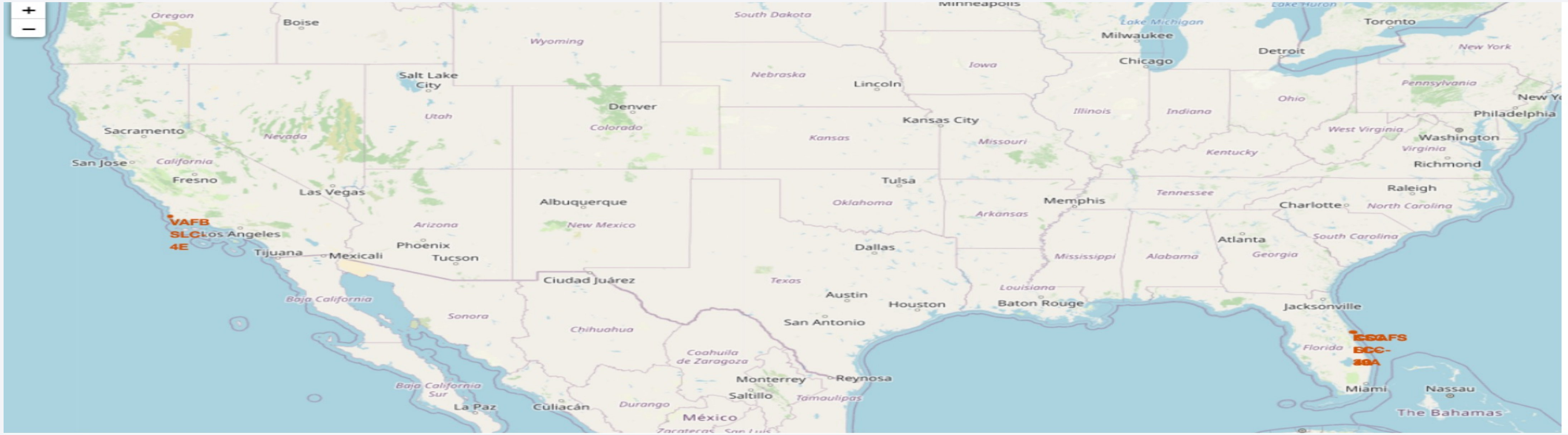
Landing_Outcome	outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

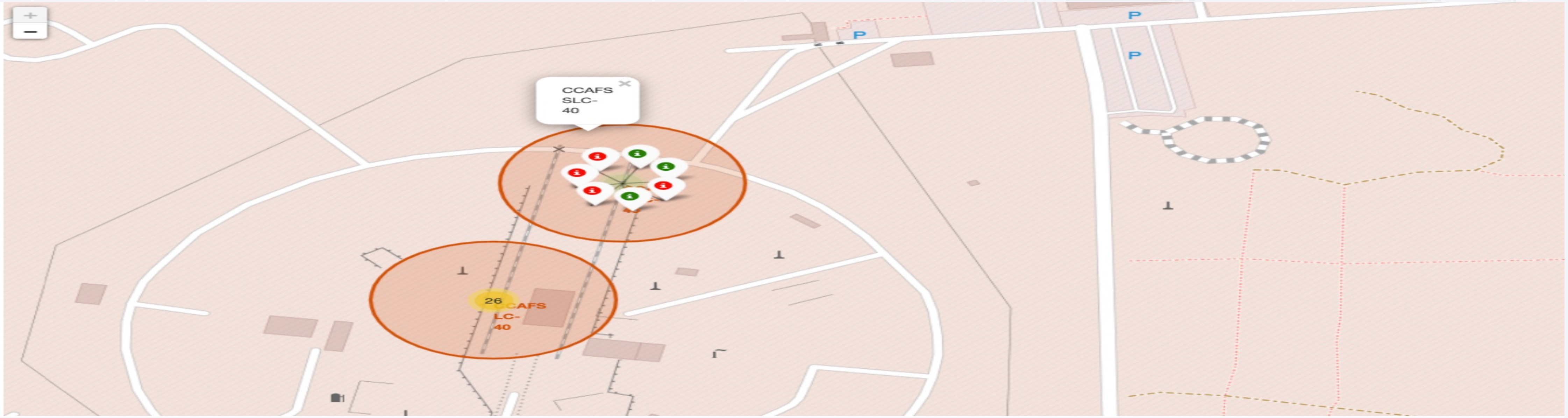
Interactive Folium Map of SpaceX Launch Sites

- This screenshot displays an interactive Folium map showing all the SpaceX launch sites worldwide, marked with distinct location markers. Each marker helps identify where launches occur and highlights the strategic placement of these sites to optimize rocket trajectories and ensure safety. The map's interactive features allow users to zoom in, zoom out, and click on markers for additional information, making it easier to understand the geographical distribution of launch sites.



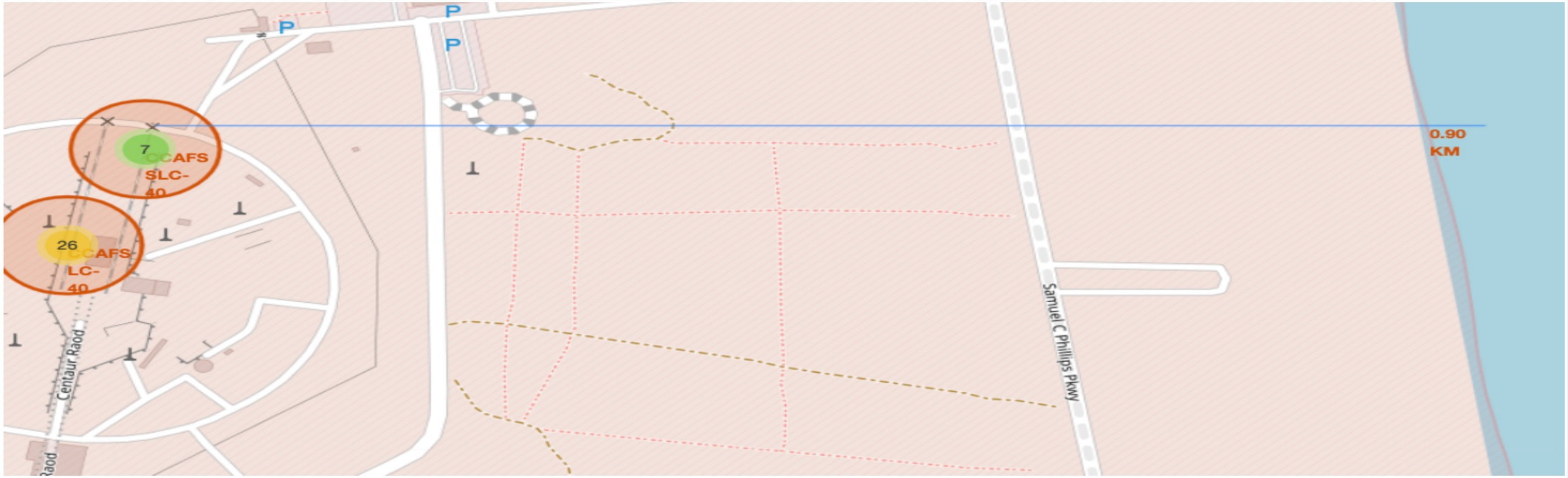
Launch Outcomes Visualized with Folium Map

- This screenshot displays a Folium map highlighting the outcomes of rocket launches across various sites. Successful launches are marked in green, while failed attempts are shown in red. Each marker represents a specific launch site, allowing for quick assessment of performance trends based on location. The map reveals geographical patterns, indicating that some sites have higher success rates than others. This visualization is valuable for understanding how location influences launch outcomes and can guide future decisions regarding launch site selection.



Folium Map Showing Launch Site Proximity to Key Infrastructure

- This screenshot displays a Folium map of a selected launch site, highlighting its proximity to key features such as railways, highways, and coastlines. The distances to these infrastructures are shown, indicating accessibility for operations. The close proximity to transportation routes suggests improved logistical support, enhancing the efficiency of launch operations.

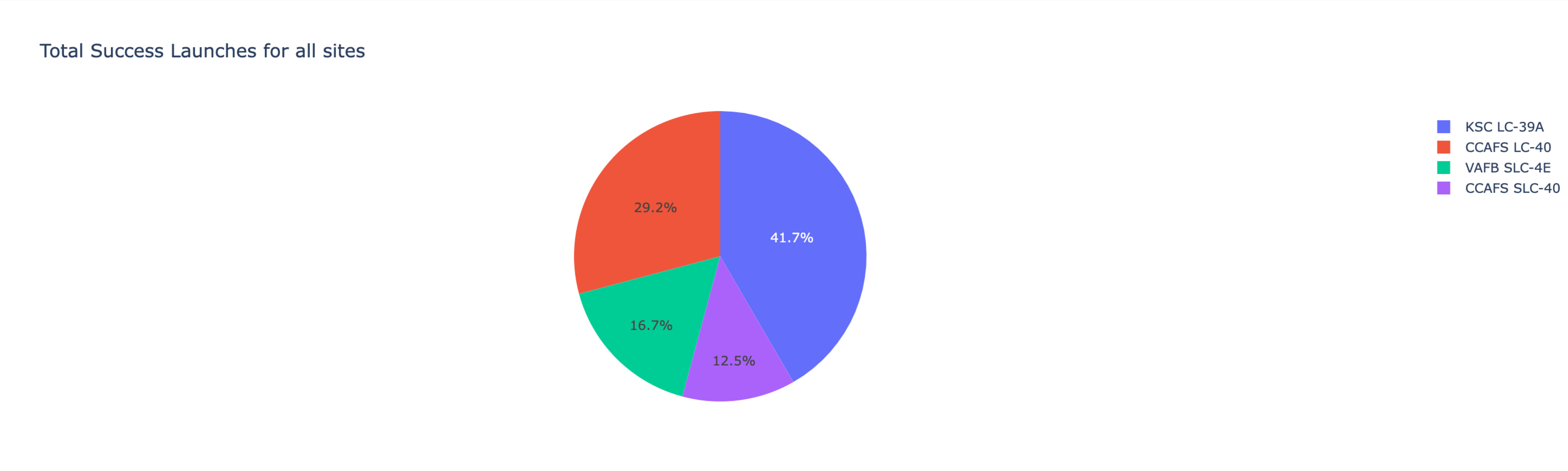


Section 4

Build a Dashboard with Plotly Dash

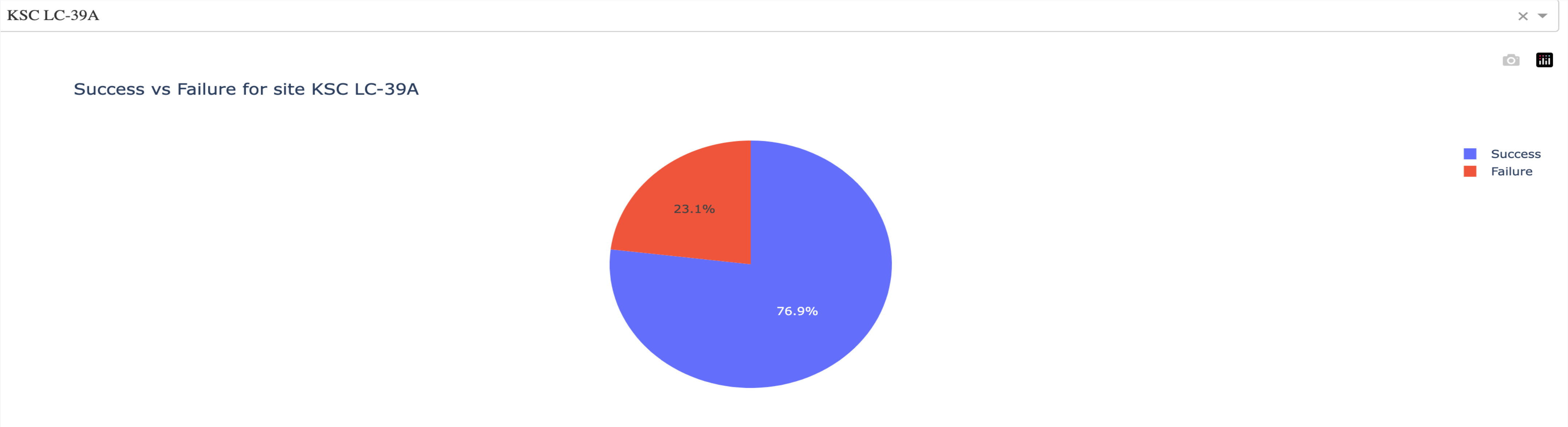
Successful SpaceX Launches by Site

- This pie chart displays the distribution of successful SpaceX launches by site. Each segment represents the number of successful launches from different locations, allowing for easy comparison. The chart highlights which sites have the highest success rates and offers insights into launch performance, helping to identify the most effective locations for future missions.



Launch Success Ratio for Top Performing Site

- This pie chart illustrates the launch success ratio for the launch site with the highest success rate. The segments represent the proportion of successful launches compared to failed ones, providing a clear visual of the site's performance. From the chart, it's evident that the majority of launches from this site were successful, highlighting its effectiveness and reliability. This information is crucial for stakeholders when evaluating launch site performance and making strategic decisions about future missions.



Scatter Plot of Payload vs. Launch Outcome

- This dashboard screenshot displays the scatter plot of Payload versus Launch Outcome for all sites, incorporating different payload ranges selected using the range slider. The plot highlights key findings, indicating that as the payload mass increases, there tends to be a higher success rate for launches. Notably, payloads within the moderate range (around 2000 to 4000 kg) show the largest success rates, while heavier payloads often result in varying outcomes. This analysis helps identify optimal payload ranges for future launches to maximize success rates.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

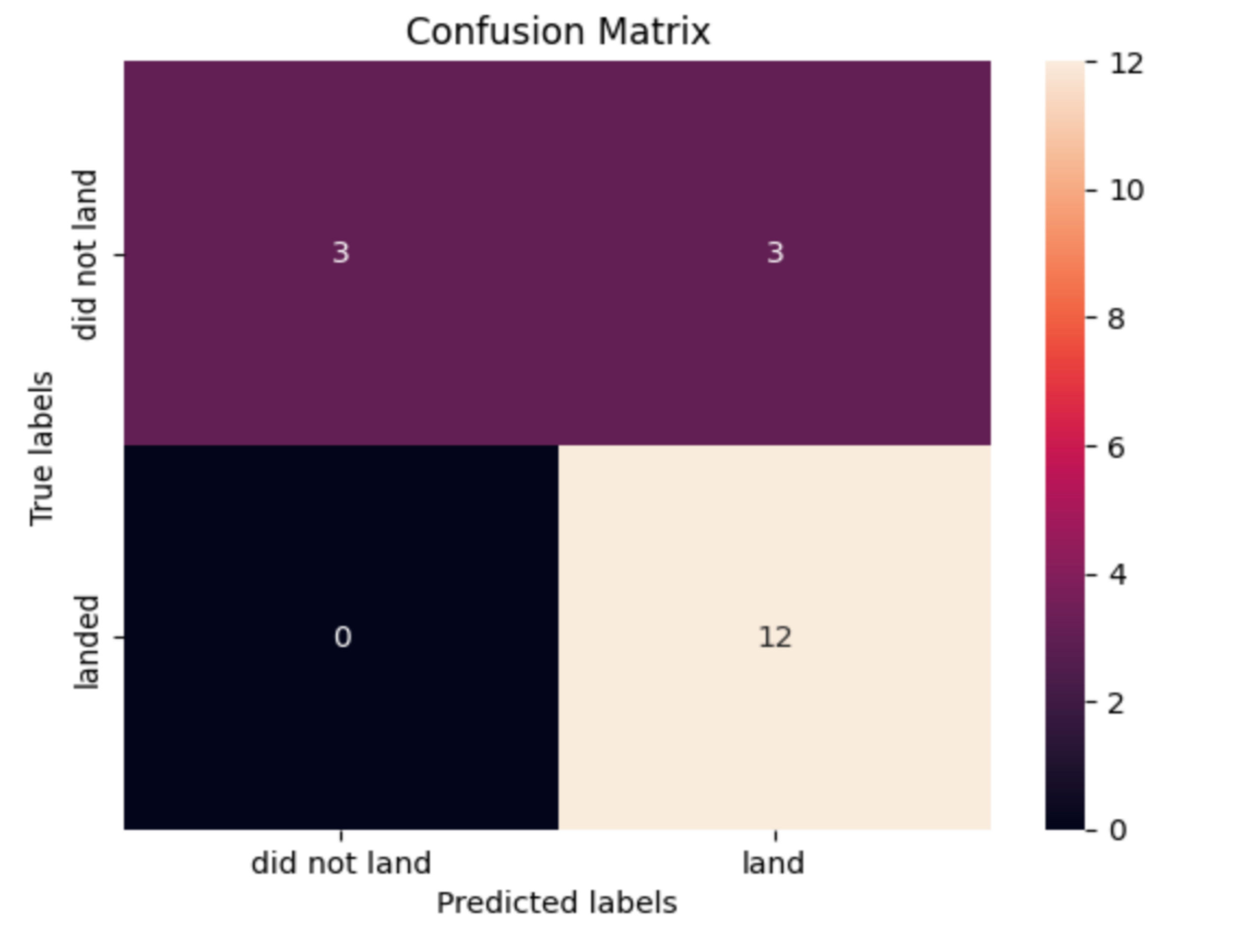
- This slide visualizes the accuracy of various classification models built during the analysis. The bar chart presents the accuracy scores of each model, allowing for an easy comparison. The model with the highest accuracy indicates the best performance in predicting successful landings, which is crucial for improving launch decision-making processes.

```
# Step 1: Calculate accuracy for Logistic Regression
logreg_accuracy = logreg_cv.score(X_test, Y_test)
print("Logistic Regression Accuracy: ", logreg_accuracy)

# Step 2: Calculate accuracy for SVM
svm_accuracy = svm_cv.score(X_test, Y_test)
print("SVM Accuracy: ", svm_accuracy)

# Step 3: Calculate accuracy for Decision Tree
tree_accuracy = tree_cv.score(X_test, Y_test)
print("Decision Tree Accuracy: ", tree_accuracy)

# Step 4: Calculate accuracy for KNN
knn_accuracy = knn_cv.score(X_test, Y_test)
print("KNN Accuracy: ", knn_accuracy)

# Step 5: Identify the best performing model
accuracies = {
    'Logistic Regression': logreg_accuracy,
    'SVM': svm_accuracy,
    'Decision Tree': tree_accuracy,
    'KNN': knn_accuracy
}

best_model = max(accuracies, key=accuracies.get)
print("Best performing model:", best_model)
```

```
Logistic Regression Accuracy:  0.8333333333333334
SVM Accuracy:  0.8333333333333334
Decision Tree Accuracy:  0.5555555555555556
KNN Accuracy:  0.8333333333333334
Best performing model: Logistic Regression
```

Confusion Matrix

- The confusion matrix illustrates the performance of the best classification model predicting Falcon 9 first-stage landings. It includes True Positives (TP) for successful landings correctly predicted, True Negatives (TN) for unsuccessful landings correctly predicted, False Positives (FP) for successful landings incorrectly predicted, and False Negatives (FN) for unsuccessful landings incorrectly predicted. This matrix is essential for assessing the model's accuracy, with a higher number of true positives and true negatives, and fewer false positives and negatives indicating better performance.



Conclusions

- **Flight Amount vs. Success Rate**: A larger number of flights at a launch site correlates with a higher success rate.
- **Trend in Launch Success**: The launch success rate has steadily increased from 2013 to 2020.
- **Orbit Success Rates**: The orbits ES-L1, GEO, HEO, SSO, and VLEO demonstrated the highest success rates.
- **Top Launch Site**: KSC LC-39A recorded the most successful launches among all sites.
- **Best Machine Learning Model**: The Decision Tree classifier proved to be the most effective algorithm for predicting landing outcomes.

Appendix

This section includes important materials from the project:

- **Python Code Snippets:** Key code used in the analysis.
- **SQL Queries:** Queries to extract data.
- **Charts:** Visuals showing data trends.
- **Notebook Outputs:** Results from the Jupyter Notebook.
- **Data Sets:** Datasets used in the project.

These materials provide additional context and support for the project findings.

- GitHub Repository:

  https://github.com/AbdullahAloyaydi/LAbCapstone

Thank you!