IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Belal Alfutayh
14/08/2024

Outline

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

Executive Summary

- **Data Collection**: Successfully gathered data from the SpaceX API and web scraping.
- **Data Wrangling**: Cleaned data for missing values and added a Class column to indicate launch success (1) or failure (0).
- **Exploratory Data Analysis (EDA)**: Executed all necessary SQL queries and visualized data effectively.
- **Interactive Visualization**: Created interactive visualizations using Folium maps and Plotly Dash apps.
- **Predictive Analysis**: Found that the Decision Tree model outperformed others with the highest accuracy.

Introduction

- Project Background and Context
  - The goal of this project is to collect and analyze relevant data to gain insights into the SpaceX Falcon 9 space launches. The process begins with gathering data from multiple sources, followed by data wrangling to clean and prepare the data for analysis.
  - Next, I'll explore the data using SQL and visualize relationships between variables. By segmenting the data and applying statistical analysis, I aim to uncover deeper patterns. Finally, I'll build and refine predictive models to better understand what drives successful launches, and present my findings in a clear and compelling way.
- Problems to Address
  - How can I efficiently collect and clean data on space launches?
  - What insights can be gained through SQL queries and data visualization?
  - Which predictive model best forecasts launch success?
  - How can I effectively present my analysis?

Section 1

Methodology

Methodology

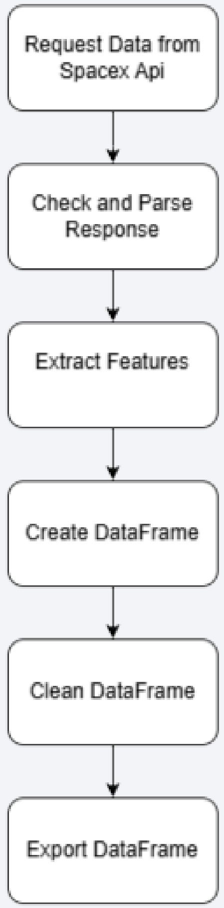
- **Data collection methodology:**
  - I gathered historical data on Falcon 9 and Falcon Heavy launches by using the SpaceX API for detailed mission information and web scraping a Wikipedia page for additional details.
- **Perform data wrangling:**
  - Process involved importing the SpaceX dataset, inspecting its structure, and checking for missing values and correct data types. I analyzed launch sites, orbit types, and mission outcomes, then labeled each launch as successful or unsuccessful.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - I built and tuned classification models using GridSearchCV to find the best parameters. I evaluated each model's performance by checking accuracy and analyzing the confusion matrix to assess how well the models classified data.

Data Collection

- To gather historical data on Falcon 9 and Falcon Heavy launches, I used two sources:
  1. **SpaceX API**: Retrieved detailed launch data directly from the SpaceX API (https://api.spacexdata.com), which provides information on each mission, including dates, rocket types, and outcomes.
  2. **Wikipedia Web Scraping**: Complemented the API data by web scraping a specific Wikipedia page to collect additional details, ensuring a comprehensive dataset.
- These two sources combined provided a thorough historical record of Falcon 9 and Falcon Heavy launches for analysis.

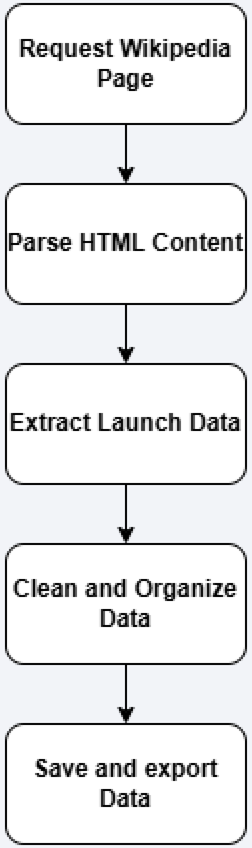
Data Collection – SpaceX API

- Request Data from API
  - Send a GET request to the SpaceX API to retrieve historical launch data.
- Check and Parse Response
  - `200 OK`: Proceed with data parsing.
  - Normalize the JSON response into a Pandas DataFrame for structured data analysis.
- Extract Features
  - Extract relevant features like `BoosterVersion`, `LaunchSite`, `PayloadData`, and `CoreData`.
- Create DataFrame
  - Create DataFrame from the Extracted Features and only include the FALCON 9 .
- Clean DataFrame
  - Clean the Dataframe from any null, missing values, etc... .
- Export DataFrame
  - Export DataFrame into any format ready for analysis.



Data Collection - Scraping

- **Request Wikipedia Page**
  - Send a request to retrieve the Wikipedia page content.
- **Parse HTML Content**
  - Use BeautifulSoup parser to extract relevant data (tables, sections) from the HTML.
- **Extract Launch Data**
  - Identify and extract the specific Falcon 9 launch data from the parsed content.
- **Clean and Organize Data**
  - Process and structure the extracted data into a usable format (e.g., CSV, JSON).
- **Save Data**
  - Store the cleaned data in a file or database.
- Check out the Webscraping Notebook in Github



Data Wrangling

- Data Wrangling Process Description
  - **Load Data**: Import the SpaceX dataset into a Pandas DataFrame.
  - **Inspect Data**: View the first few rows to understand the data structure.
  - **Check Missing Values**: Identify missing values and their percentages in each column.
  - **Check Data Types**: Verify the data types of each column to ensure correctness.
  - **Analyze Features**:
    - **Launch Sites**: Count the number of launches per site.
    - **Orbit Types**: Count occurrences of each orbit type.
  - **Analyze Outcomes**: Examine mission outcomes to classify successful and unsuccessful landings.
  - **Create Labels**: Add a new column `Class` to label each launch as successful (`1`) or unsuccessful (`0`).

- Check the Data Wrangling Notebook at [Github]

EDA with Data Visualization

Analysis	Chart Type	Key Insight
Flight Number vs. Payload Mass	Scatter Plot	Higher flight numbers correlate with better success rates.
Flight Number vs. Launch Site	Scatter Plot	Success rates vary significantly across different launch sites.
Payload Mass vs. Launch Site	Scatter Plot	Certain sites are less effective with heavier payloads.
Success Rate by Orbit Type	Bar Chart	Some orbit types consistently achieve higher success rates.
Yearly Success Rate Trend	Line Chart	Success rates have steadily increased since 2013.
Flight Number vs. Orbit Type	Scatter Plot	LEO orbits show a strong positive correlation with flight numbers; GTO orbits do not.

Check out the EDA DATA VISUALIZATION Notebook in [Github]

EDA with SQL

Task	Description	SQL keywords
1	Retrieve unique launch site names.	SELECT DISTINCT
2	Get 5 records where launch sites start with 'CCA'.	WHERE ... LIKE, LIMIT
3	Calculate total payload mass for NASA (CRS).	SUM()
4	Find average payload mass for F9 v1.1.	AVG()
5	List the date of the first successful ground pad landing.	WHERE ... = 'Success (ground pad)'

Check the SQL Queries Notebook in [Github]

Task	Description	SQL keywords
6	Identify boosters with payloads between 4000-6000 kg.	WHERE ... = 'Success (drone ship)', AND ... BETWEEN
7	Count the number of successful and failed outcomes.	COUNT(), WHERE ... LIKE
8	List booster versions carrying the maximum payload mass.	SELECT MAX(), Subquery
9	Display records for drone ship landing failures in 2015.	AVG()
10	Rank successful landings from June 2010 to March 2017.	ORDER BY DATE DESC

Build an Interactive Map with Folium

- Map Objects Added to Folium Map :
  - **Markers and Circles for Launch Sites:**
    - **What:** Added markers and circles to pinpoint SpaceX launch sites (e.g., CCAFS LC-40, KSC LC-39A).
    - **Why:** To visualize the exact locations of each launch site on the map.
  - **Success/Failure Launch Markers:**
    - **What:** Green and red markers were added to indicate successful (green) and failed (red) launches.
    - **Why:** To visually assess the performance and success rates of launches at different sites.
  - **Mouse Position Tool:**
    - **What:** A tool to display the latitude and longitude of the mouse cursor's position on the map.
    - **Why:** To easily determine the coordinates of points of interest (e.g., coastlines, highways).
  - **Distance Lines:**
    - **What:** Lines representing distances from launch sites to nearby features like coastlines.
    - **Why:** To analyze the proximity of launch sites to important geographical features, identifying potential influences on launch outcomes.
- Check the Interactive Map Notebook in Github

Build a Dashboard with Plotly Dash

- **Plots/Graphs:**
  - **Pie Chart**: Displays total successful launches by site.
    - **Interaction**: Select a launch site from the dropdown to filter by specific site; otherwise, shows data for all sites.
  - **Scatter Plot**: Illustrates the correlation between payload mass and launch success.
    - **Interaction**: Use the dropdown to filter by launch site and the slider to select a payload mass range
- **Why These Elements?**
  - **Pie Chart**: Provides a quick overview of success rates across different launch sites, enabling easy comparison.
  - **Scatter Plot**: Allows users to explore how payload mass impacts launch success, filtered by site and payload range for detailed insights.

- Check the Plotly Dashboard Python at Github

Predictive Analysis (Classification)

- Model Development Process
  - **Data Preparation**: Loaded and explored the dataset, created a 'Class' column as the target variable, and standardized the feature data.
  - **Data Splitting**: Split the data into training and test sets (80% training, 20% testing).
  - **Model Training**: Evaluated three classification models—Support Vector Machine (SVM), Decision Trees, and Logistic Regression, K Nearest Neighbor—using GridSearchCV to find the best hyperparameters.
  - **Model Evaluation**: Tested the models on the test set and compared their performance.
  - **Best Model Selection**: Chose the best-performing model based on accuracy and confusion matrix analysis.

- Check The Predictive Analysis Notebook in Github

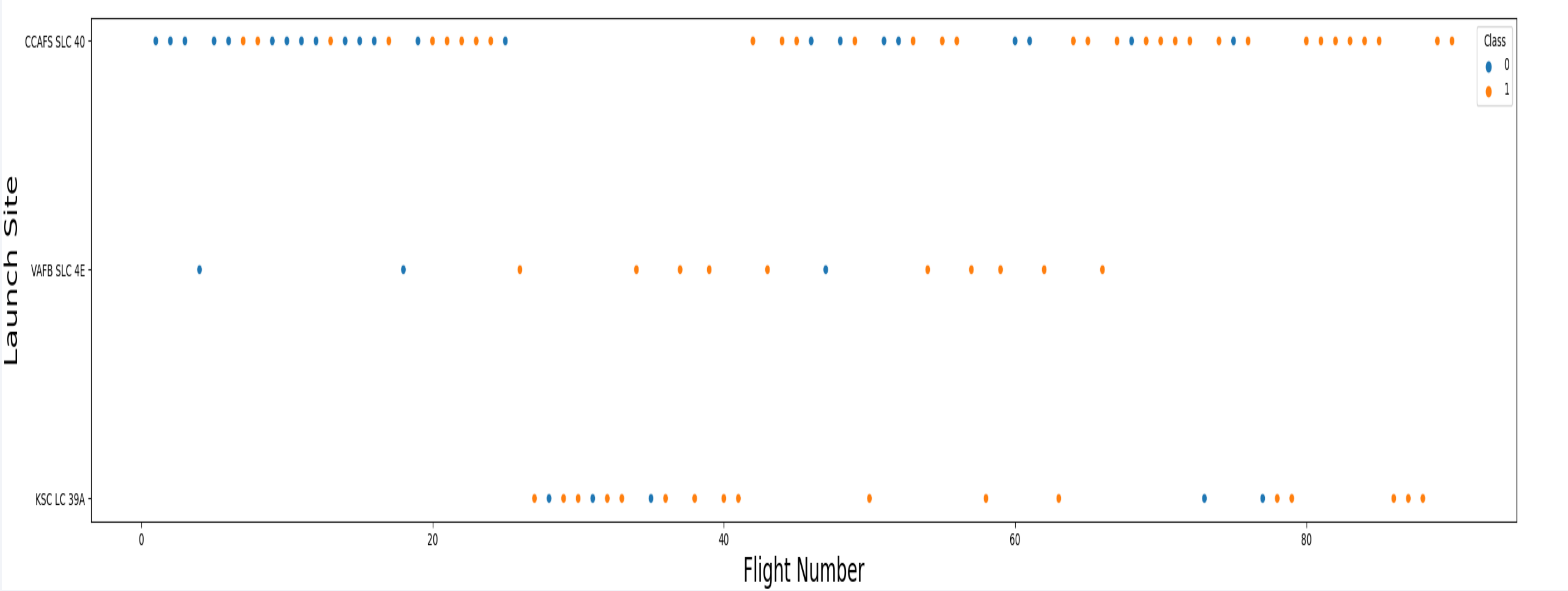
Results

- **Exploratory Data Analysis (EDA):**
  - **Launch Sites:** CCAFS LC-40 was the most active, showing both successes and failures.
  - **Payload Mass & Booster Performance:** NASA (CRS) missions had a total payload of 45,596 kg.
  - **Success Rates by Orbit & Flight:** Higher flight numbers correlated with increased landing success; high payloads often led to positive outcomes.
- **Predictive Analysis:**
  - **Mission Success Prediction:** Predicting launch success using the "Class" column.
  - **Model Performance:** The Decision Tree model outperformed other models with high accuracy in predicting successful launches.
- **Plotly Dash Visualizations:**
  - **Pie Chart:** KSC LC-39A has the highest launch success rate, with 41.7% of total successful launches.
  - **Scatter Plot:** When payloads range between 2000 and 7000 kg, the highest success rate is observed for a payload of 5300 kg with the Booster Version Category.

Section 2

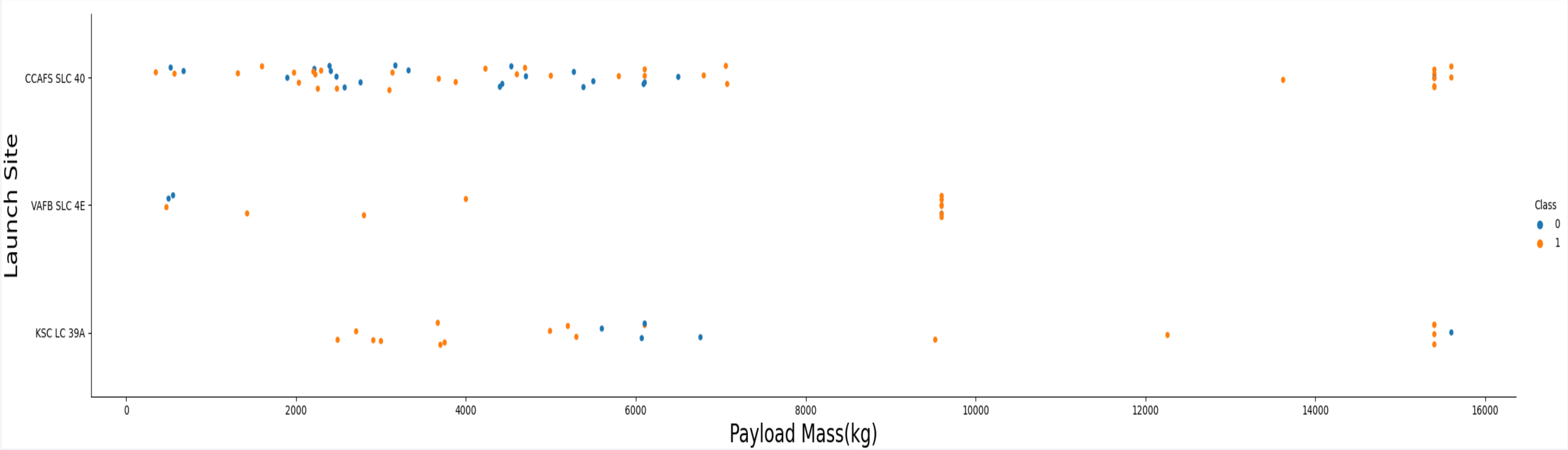
Insights drawn from EDA

Flight Number vs. Launch Site



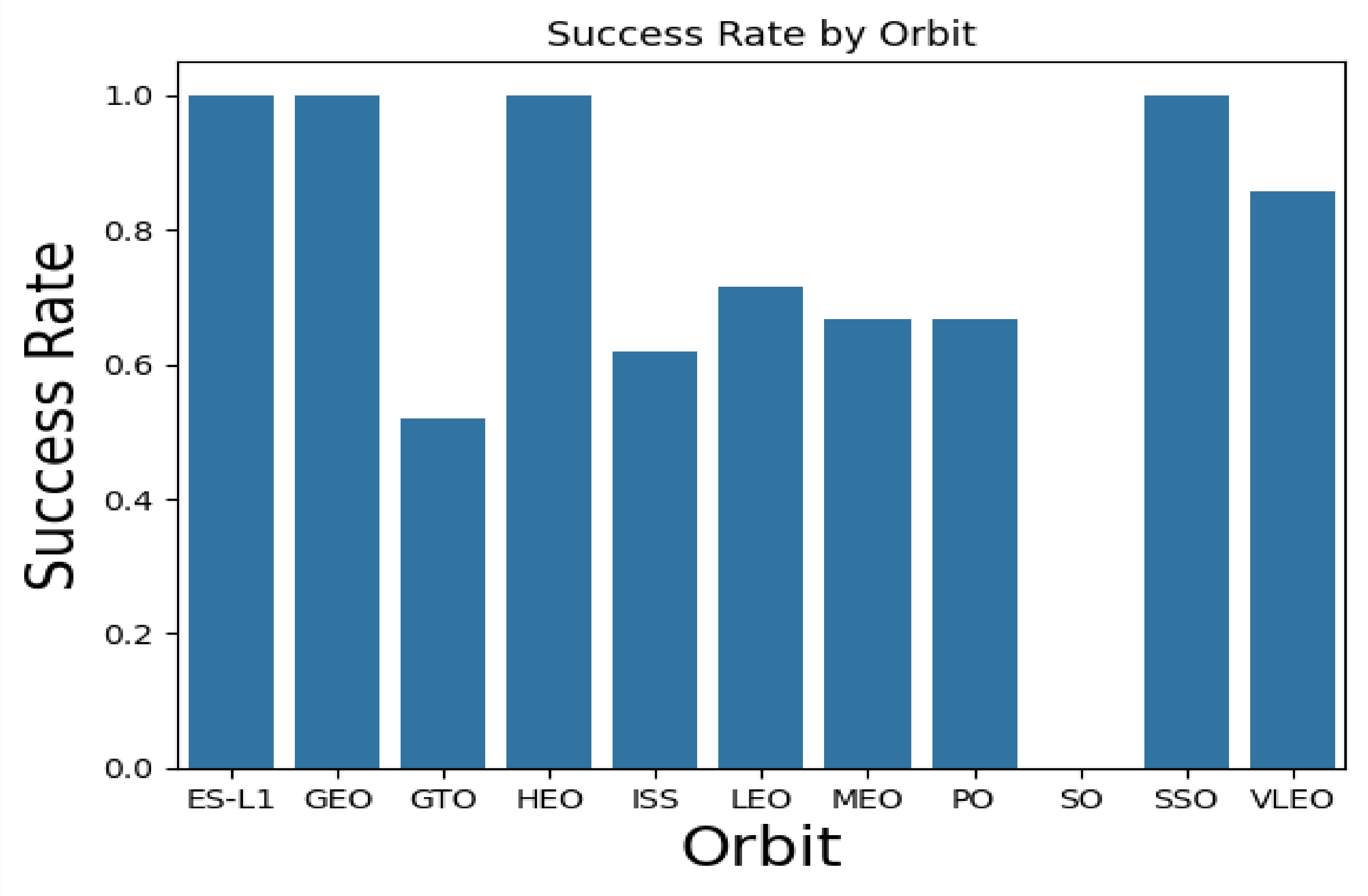
The scatter plot visualizes the relationship between Flight Number and Launch Site, with the hue representing the success or failure of each mission (Class). Each dot represents a flight, positioned according to its Flight Number and the associated Launch Site.

Payload vs. Launch Site



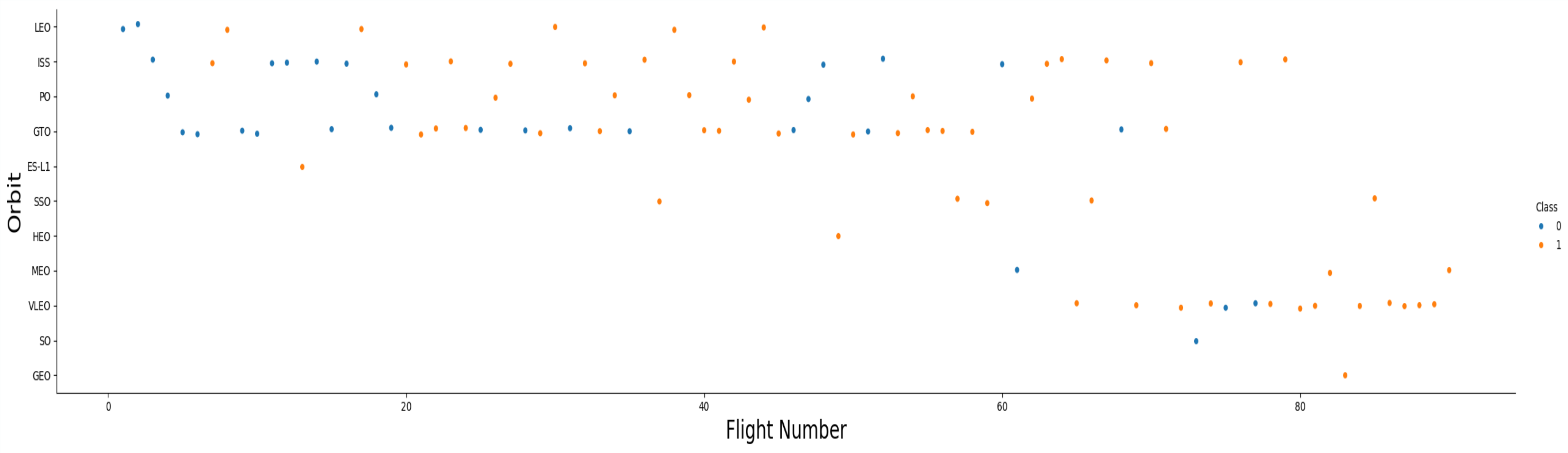
This scatter plot visualizes the relationship between the payload mass of a mission and the launch site from which it was launched. The color (hue) indicates whether the mission was successful (Class = 1) or not (Class = 0).

Success Rate vs. Orbit Type



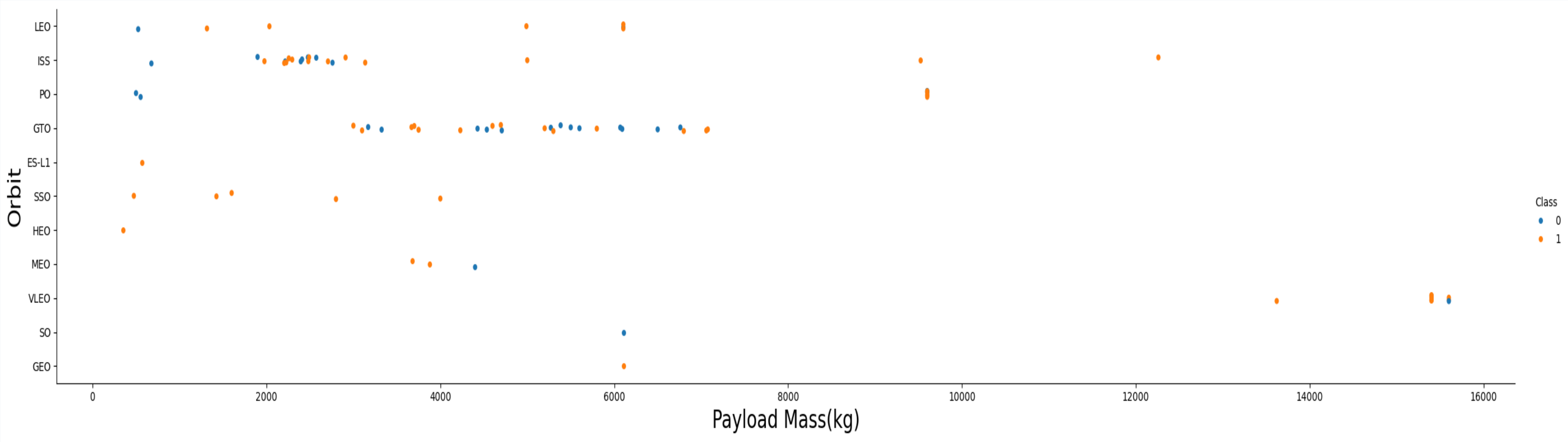
This bar chart shows the average success rate for each orbit type. It helps identify which orbits, such as LEO, GEO, and MEO, have higher mission success rates.

Flight Number vs. Orbit Type



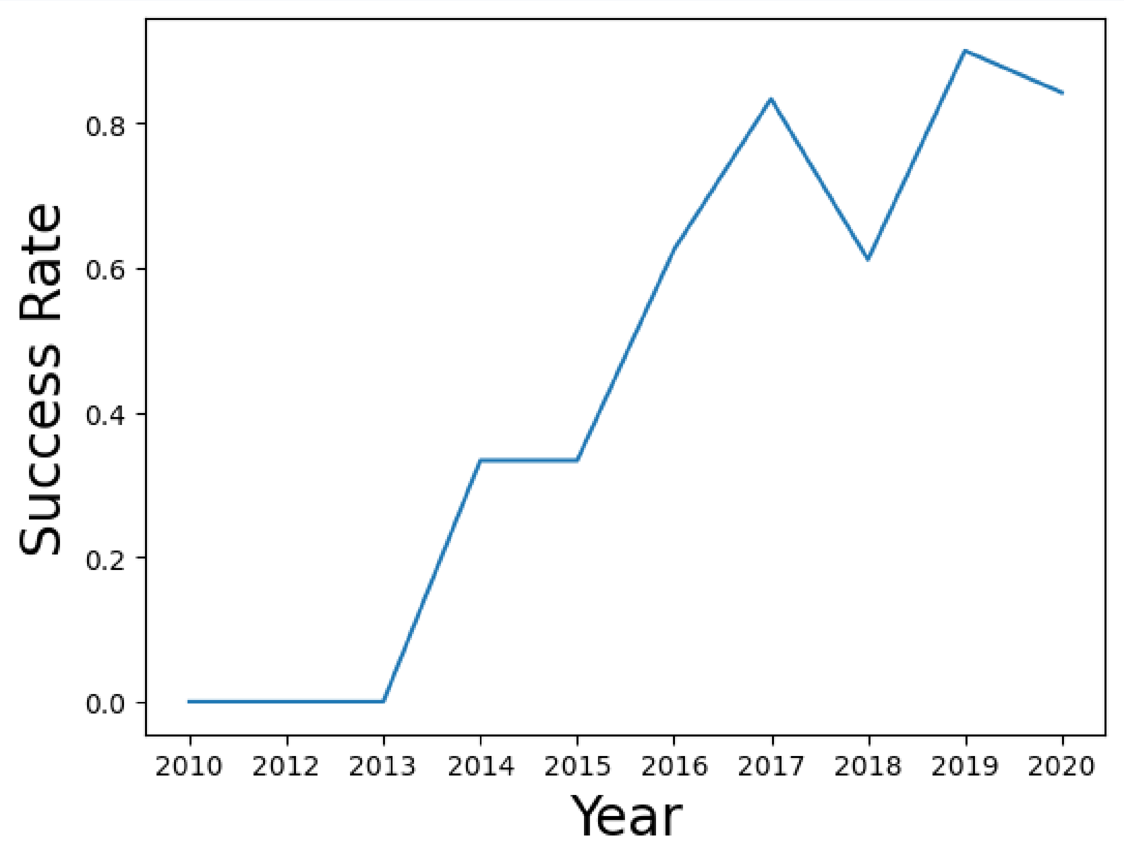
This scatter plot examines the relationship between flight numbers and orbit types, with success indicated by color. The purpose is to identify patterns in success rates across different orbits as the number of flights increases. The plot shows that for LEO, higher flight numbers tend to correlate with more success, while in GTO, no clear relationship is observed.

Payload vs. Orbit Type



This scatter plot shows the relationship between payload mass and orbit type, with success indicated by color. It reveals that heavier payloads are associated with higher success rates in orbits such as Polar, LEO, and ISS. Conversely, in GTO, there is no clear distinction between successful and unsuccessful landings based on payload mass.

Launch Success Yearly Trend



This line chart illustrates the trend of average launch success rates over the years. It shows that from 2013 to 2020, the success rate has generally increased, indicating an improving trend in launch success over this period.

All Launch Site Names

- **Task:**
  - Display the names of the unique launch sites used in space missions.
- **Result:**
  - **Unique Launch Sites:**
    - CCAFS LC-40
    - VAFB SLC-4E
    - KSC LC-39A
    - CCAFS SLC-40
- **Explanation:**
  - The data reveals the distinct locations from which space missions have been launched, highlighting key sites in the space program.

Launch Site Names Begin with 'CCA'

- **Task:**
  - Display 5 records where the launch sites begin with the string "CCA."
- **Result:**

Booster Version	Launch Site	Customer
F9 v1.0 B0003	CCAFS LC-40	SpaceX
F9 v1.0 B0004	CCAFS LC-40	NASA (COTS) NRO
F9 v1.0 B0005	CCAFS LC-40	NASA (COTS)
F9 v1.0 B0006	CCAFS LC-40	NASA (CRS)
F9 v1.0 B0007	CCAFS LC-40	NASA (CRS)

- **Explanation:**
  - These records focus on launches that took place at sites starting with "CCA," specifically identifying the Cape Canaveral Air Force Station (CCAFS LC-40).

Total Payload Mass

- **Task:** Display the total payload mass carried by boosters launched by NASA (CRS).
- **Result:**
  - **Total Payload Mass:** 45,596 kg
- **Explanation:**
  - This value represents the cumulative mass of all payloads launched by NASA under the Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

- **Task:** Display the average payload mass carried by booster version F9 v1.1.
- **Result:**
    - **Average Payload Mass:** 2,534.67 kg
- **Explanation:**
    - This figure represents the average mass of payloads transported by the Falcon 9 v1.1 booster version across its various missions.

First Successful Ground Landing Date

- **Task:**
  - List the date when the first successful landing outcome on a ground pad was achieved.
- **Result:**
  - **First Successful Landing on Ground Pad: December 22, 2015**
- **Explanation:**
  - This date marks the first time a landing outcome was successfully achieved on a ground pad, setting a significant milestone in space exploration.

Successful Drone Ship Landing with Payload between 4000 and 6000

- **Task:** List the names of the boosters that successfully landed on a drone ship and carried a payload mass greater than 4,000 kg but less than 6,000 kg.
- **Result:**
  - **Booster Versions:**
    - **F9 FT B1022**
    - **F9 FT B1026**
    - **F9 FT B1021.2**
    - **F9 FT B1031.2**
- **Explanation:**
  - These boosters not only achieved successful landings on a drone ship but also transported payloads within the specified mass range, highlighting their reliability and capability in handling mid-weight payloads.

Total Number of Successful and Failure Mission Outcomes

- **Task:** List the total number of successful and failed mission outcomes.
- **Result:**
  - Total Number of Mission Outcomes:
    - **101** (combined successful and failed missions)
- **Explanation:**
  - This count represents the sum of all missions that either succeeded or failed, providing a quick overview of the overall mission success rate within the dataset.

Boosters Carried Maximum Payload

- **Task:** List the names of the booster versions that carried the maximum payload mass using a subquery.
- **Result:**
  - Booster Versions:
    - **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**

- **Explanation:**
  - This count represents the sum of all missions that either succeeded or failed, providing a quick overview of the overall mission success rate within the dataset.

2015 Launch Records

- **Task:**
  - List records for the year 2015 showing:
    - Month names (extracted as numbers)
    - Failure landing outcomes on drone ships
    - Booster versions
    - Launch sites

- **Result:**

Month	Landing Outcome	Booster Version	Launch Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- **Explanation:**
  - In 2015, there were two instances where launches failed to land successfully on a drone ship. These failures occurred in January and April, both involving booster versions F9 v1.1 launched from CCAFS LC-40.

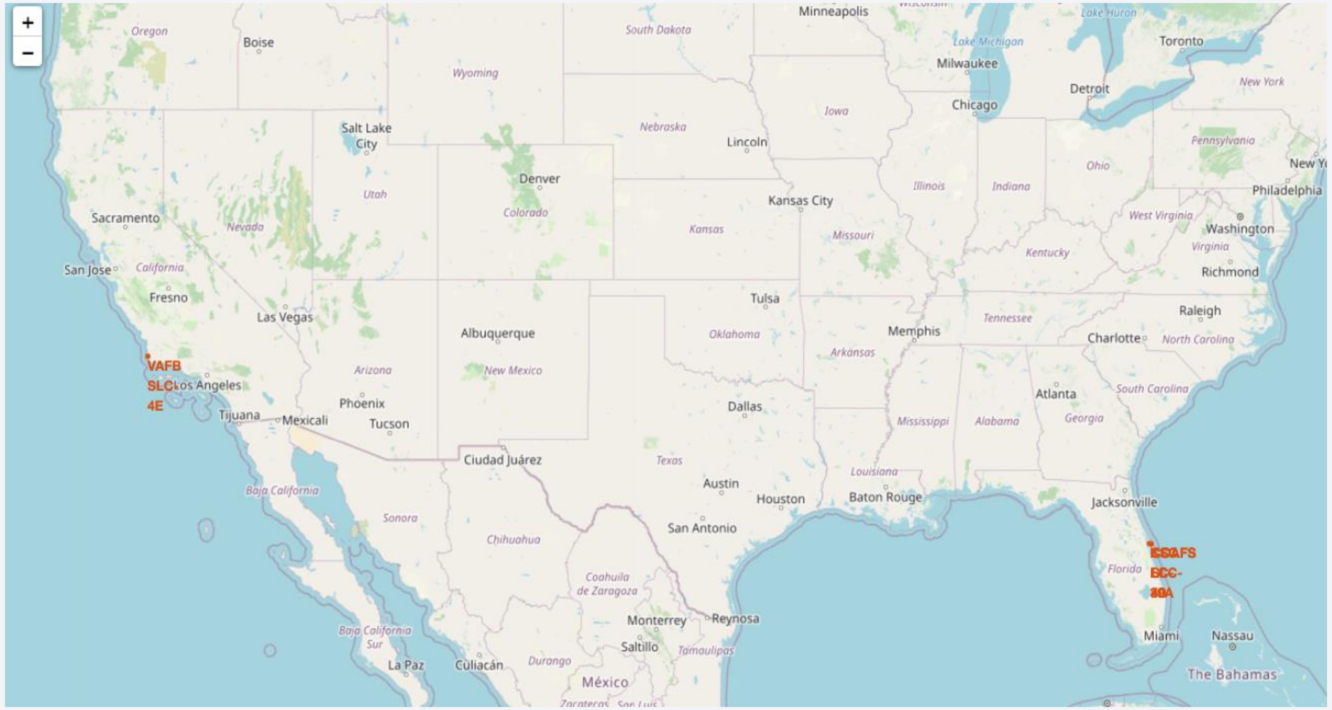
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- **Task:**
  - Rank the count of different landing outcomes (e.g., "Failure (drone ship)" or "Success (ground pad)") between the specified dates, in descending order.
- **Result:**
  - **No attempt** - **10** occurrences
  - **Success (drone ship)** - **5** occurrences
  - **Failure (drone ship)** - **5** occurrences
  - **Success (ground pad)** - **3** occurrences
  - **Controlled (ocean)** - **3** occurrences
  - **Uncontrolled (ocean)** - **2** occurrences
  - **Failure (parachute)** - **2** occurrences
  - **Precluded (drone ship)** - **1** occurrence
- **Explanation:**
  - The most frequent outcome during this period was "No attempt" with 10 occurrences.
  - Both "Success (drone ship)" and "Failure (drone ship)" each had 5 occurrences.
  - Other outcomes, such as "Success (ground pad)" and "Controlled (ocean)," were less common.

Section 3

Launch Sites Proximities Analysis

Display All Launch Sites



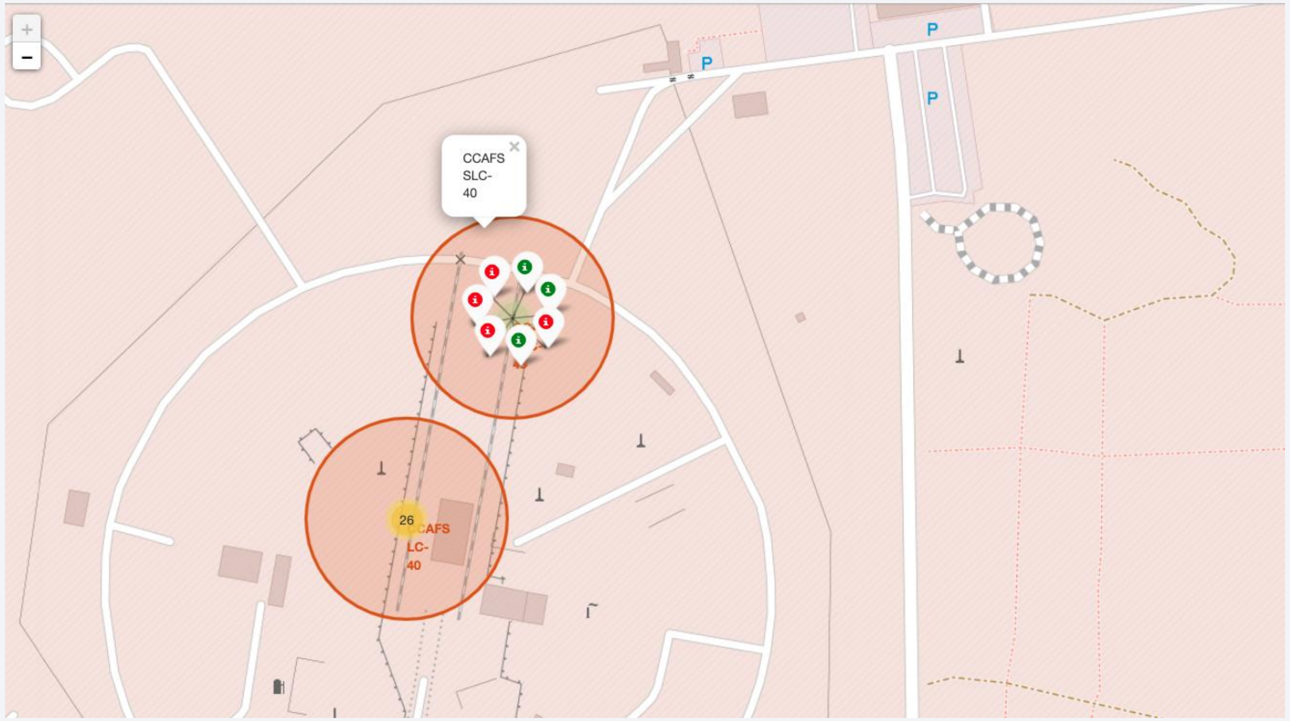
1. Map Overview:

- The map displays various launch sites marked with color-coded circles.
- Each circle represents a launch site and is labeled with the site's name.

2. Elements to Include:

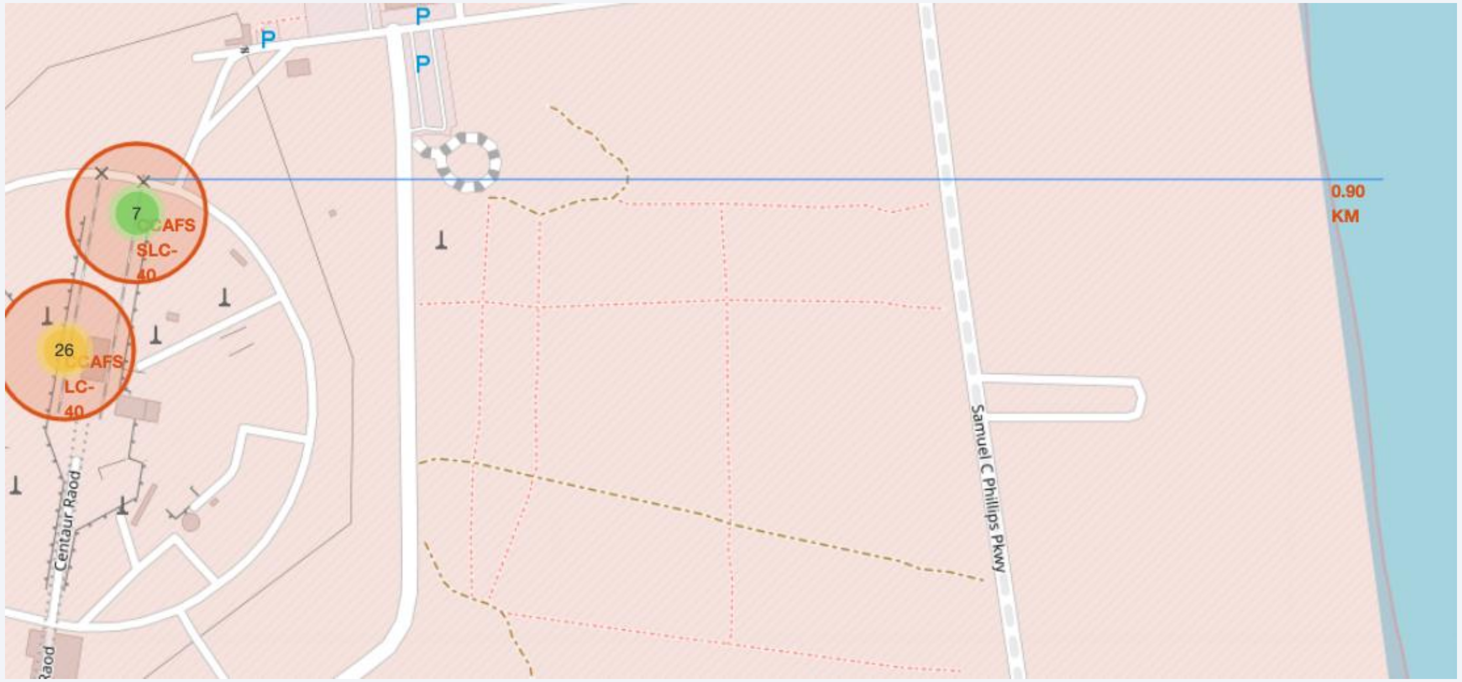
- **Launch Sites:** Show the locations of different launch sites with colored circles.
- **Popup Labels:** Indicate launch site names when hovering over the circles.
- **Circle Color:** Use a distinct color (e.g., orange) for clarity and visibility.

Display Success/Failed Launches for each site



1. **Map Overview:**
   - The map displays clustered markers indicating the success or failure of launches.
   - Markers are color-coded: green for successful launches and red for failed launches.
2. **Elements to Include:**
   - **Marker Cluster:** Shows the concentration of launch outcomes.
   - **Color Coding:** Indicates the outcome of each launch (green/red).

Calculate Distances between Launch Sites to its proximities

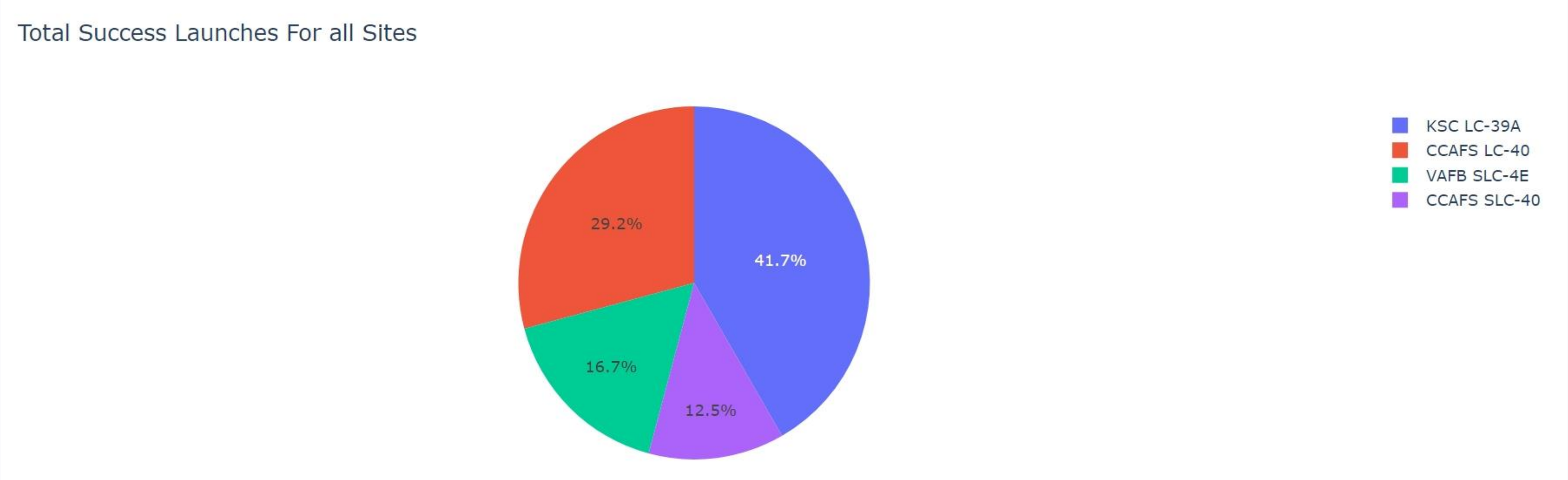


- **Map Overview:**
  - The map displays a launch site with markers and lines showing distances to a coastline and a city.
  - Each line is color-coded, and markers include distance annotations.
- **Elements to Include:**
  - **Markers:** Show locations of coastline and city with distance annotations.
  - **Lines:** Indicate the distance from the launch site to these landmarks.
  - **Colors:** Use different colors for different lines to distinguish between distances.

Section 4

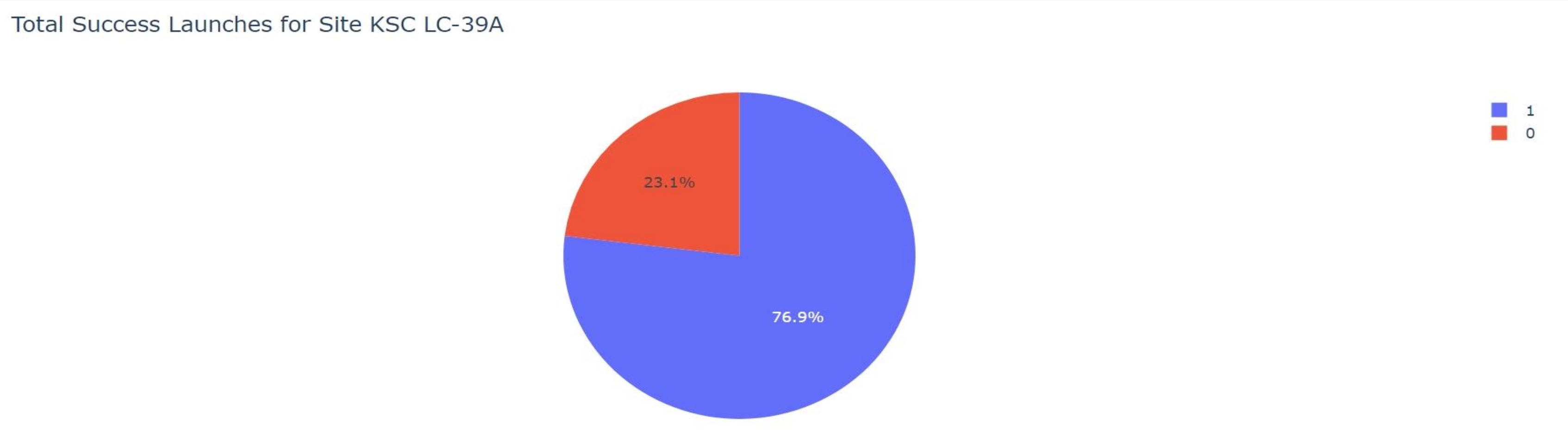
Build a Dashboard with Plotly Dash

Success Launches for sites Pie Chart



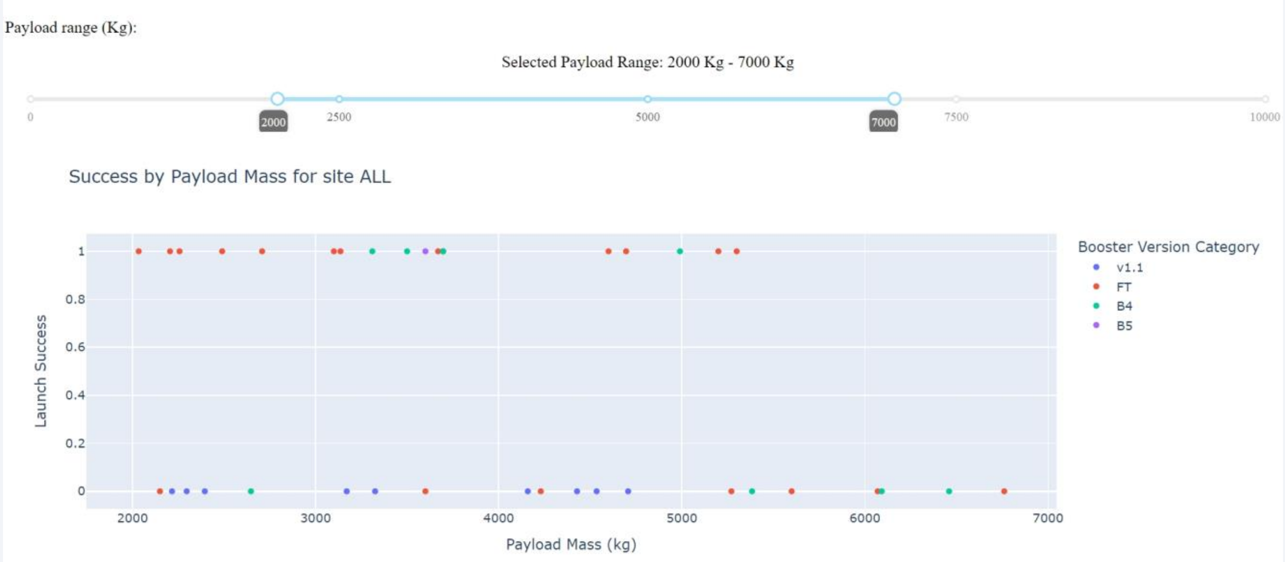
- **Findings:**
  - **Overall Success Rates:** Overview of launch success across sites.
  - **Site Performance:** Highlights top and low-performing sites.
  - **Visual Insight:** Easy comparison of success rates by site.
- **Key Elements:**
  - **Chart Purpose:** Displays success rates for SpaceX launch sites.
  - **Interactivity:** Select sites or view all.
  - **Color Coding:** Differentiates success and failure.
  - **Labels:** Show success percentages and site names.

Highest Success Rate Launch Site Pie Chart



- **Findings:**
  - **High Success Rate:** KSC LC-39A is the most successful launch site.
  - **Performance Insight:** Majority of launches from this site are successful.
  - **Visual Clarity:** Clear representation of the dominance of successful launches.
- **Key Elements:**
  - **Chart Focus:** Success vs. failure rates at KSC LC-39A.
  - **Success Ratio:** Displays the highest success rate among all sites.
  - **Color Coding:** Green for success, red for failure.
  - **Labels:** Percentage of successful and failed launches.

Launch Success vs. Payload Mass

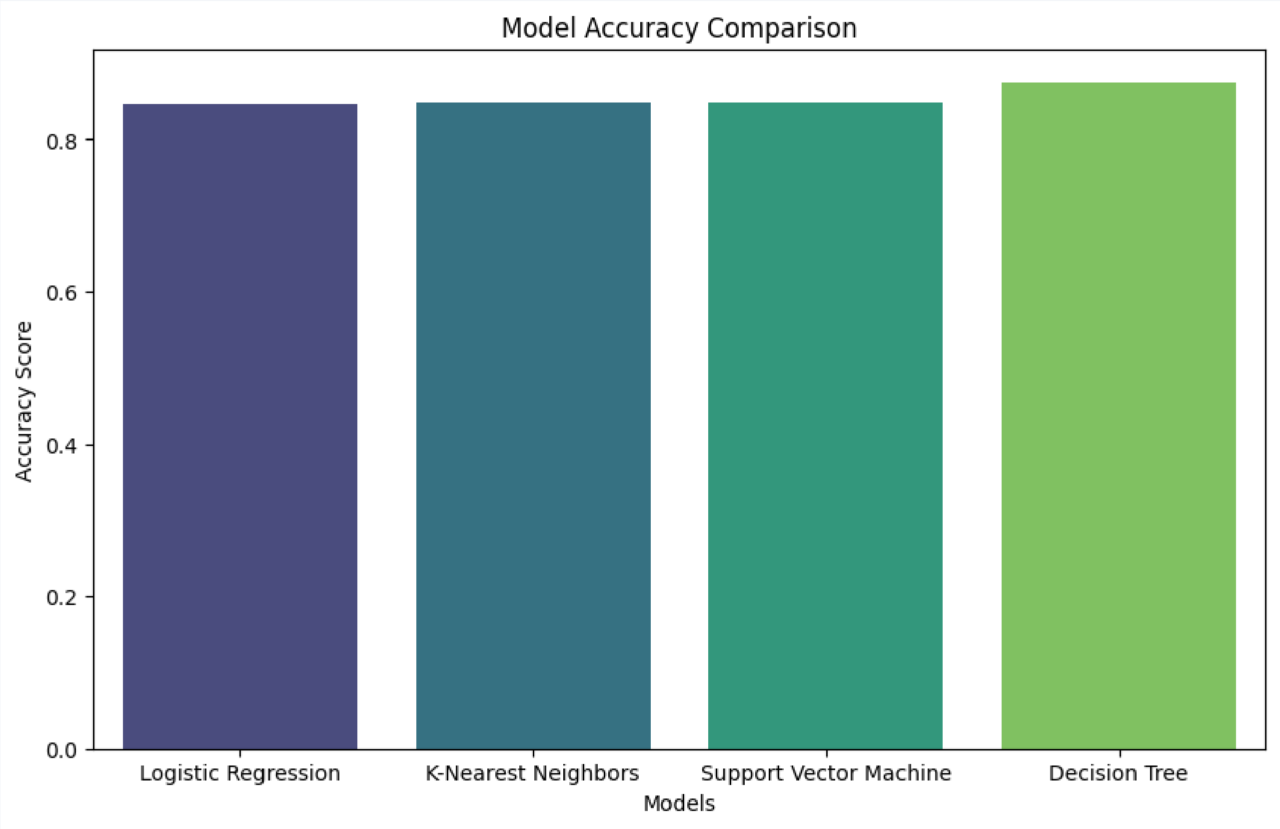


- **Findings:**
  - **FT Booster**: Highest success rate, especially with a payload around 5300 kg.
  - **2000-7000 kg Range**: High success rate for most launches.
  - **Visual Clarity:** Clear representation of the dominance of successful launches.
- **Key Elements:**
  - **X-axis**: Payload mass (kg).
  - **Y-axis**: Launch success (1 for success, 0 for failure).
  - **Data Points**: Represent individual launches, color-coded by Booster Version.
  - **Range Filter**: Adjustable payload range using a slider.

Section 5

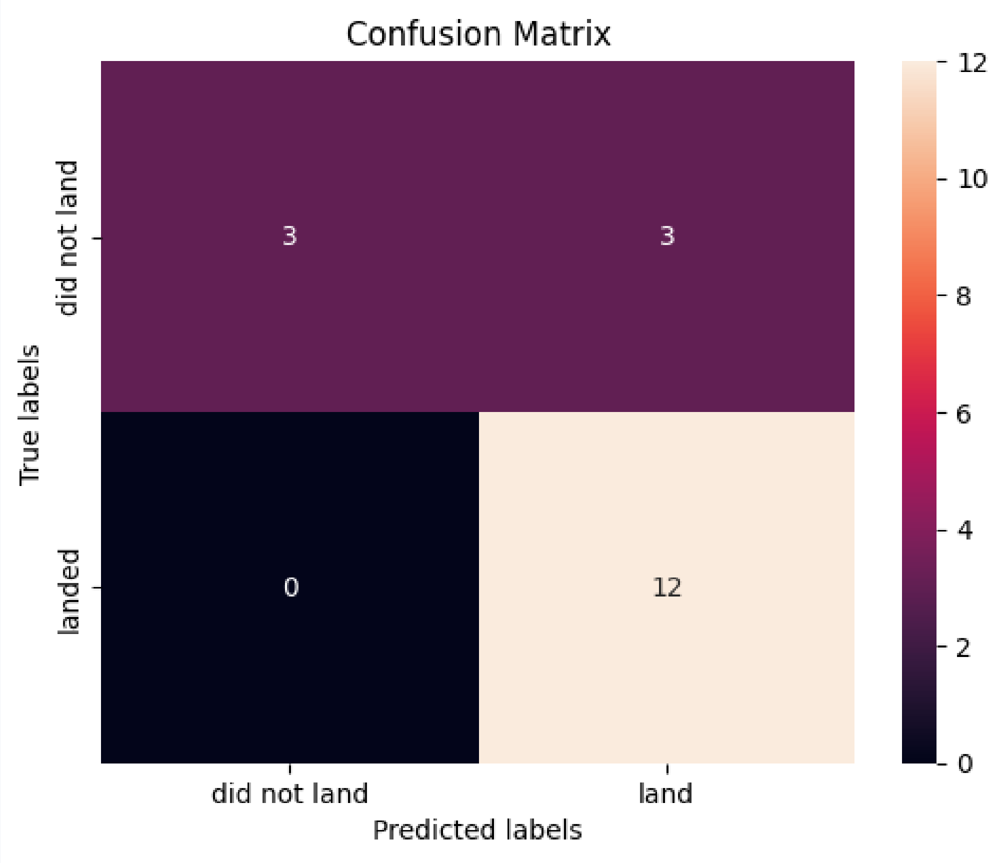
Predictive Analysis (Classification)

Classification Accuracy



- The bar chart compares the accuracy of different machine learning models on the given dataset.
  - **Best Model**: The Decision Tree model outperformed the others, achieving the highest accuracy of **87.5%**.
  - **Model Parameters**: The Decision Tree was optimized using parameters like the Gini criterion, a max depth of 4, and specific settings for splitting and leaf samples.

Confusion Matrix



- The confusion matrix shows how well the Decision Tree model performed in predicting whether a rocket landed successfully or not:
  - **True Positive (12 cases)**: The model correctly predicted that the rocket landed when it actually did.
  - **False Positive (3 cases)**: The model incorrectly predicted that the rocket landed when it didn't.
- This means the model is generally accurate but sometimes mistakenly predicts a successful landing when there wasn't one.

Conclusions

- In this project, I embarked on a comprehensive journey to analyze SpaceX Falcon 9 space launches, aiming to extract meaningful insights and enhance my understanding of launch success.
- **Data Collection:** I successfully gathered data from both the SpaceX API and through web scraping, ensuring a robust foundation for our analysis.
- **Data Wrangling:** The data was meticulously cleaned to address missing values and enhanced with a Class column to distinguish between successful (1) and failed (0) launches.
- **Exploratory Data Analysis (EDA):** I conducted thorough SQL queries and visualized the data effectively, revealing key relationships and patterns.
- **Interactive Visualization:** Leveraging Folium maps and Plotly Dash apps, I created interactive visualizations that provided intuitive insights into the data.
- **Predictive Analysis:** My efforts culminated in the evaluation of various predictive models, where the Decision Tree model emerged as the most accurate in forecasting launch success.
- By systematically addressing the challenges of data collection, cleaning, visualization, and prediction, I have gained a deeper understanding of the factors influencing SpaceX launches. This project not only highlights the power of data-driven insights but also demonstrates the effectiveness of my chosen analytical and predictive methodologies.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!