IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Labeabo Putra Asmar
01 March 2024

Outline

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

Executive Summary

- ## Summary of methodologies

This project focused on analyzing historical records of SpaceX Falcon 9 launches to predict the outcome of the first stage landing. The methodology employed various techniques, including data collection through the SpaceX API, data wrangling, exploratory data analysis (EDA) using visualization and SQL, interactive visual analytics using Folium and Plotly Dash, and predictive analysis using classification models.

- ## Summary of all results

The exploratory data analysis provided valuable insights into SpaceX's launch operations, revealing high overall success rates, diverse payload masses, and the geographic distribution of launch sites. The visualization of correlations between variables enhanced understanding of mission outcomes and identified potential dependencies. Furthermore, **the machine learning models successfully predicted Falcon 9** first stage landing outcomes, with the **SVM classifier** achieving the **highest accuracy of 83.33%** on the test data, serving as a reliable tool for future launch success predictions.

Introduction

- **Project background and context**

In recent years, the space industry has witnessed remarkable advancements, particularly with the emergence of private companies like SpaceX, founded by visionary entrepreneur Elon Musk. SpaceX's Falcon 9 rocket stands as a symbol of innovation and cost efficiency in space exploration. With its ability to successfully land and reuse the first stage, Falcon 9 has revolutionized the economics of space travel. This project delves into the predictive analysis of Falcon 9 first stage landings, aiming to leverage machine learning techniques to forecast the outcome of these critical maneuvers.

- **Problems we want to find answers**

The primary challenge addressed in this project revolves around determining the likelihood of a successful landing for the Falcon 9 first stage. Given the substantial cost savings associated with reusing this component, accurately predicting its landing outcome holds significant implications for cost estimation in rocket launches. By analyzing historical data and employing machine learning algorithms, we seek to answer fundamental questions such as:

1. Can we reliably predict whether the Falcon 9 first stage will land successfully?
2. How can this predictive capability aid in cost estimation and decision-making processes for space missions?

Through rigorous analysis and model development, we aim to provide actionable insights for stakeholders in the space industry, enabling informed decision-making and enhancing the efficiency of space exploration endeavors.

Section 1

Methodology

Methodology

Executive Summary

The methodology employed in this project encompasses several key steps, including data collection, preprocessing, exploratory data analysis (EDA), interactive visual analytics, and predictive analysis using classification models. Each stage is designed to systematically extract insights from the dataset and develop accurate predictions regarding the successful landing of SpaceX Falcon 9 first stage.

- **Data collection methodology**
  Data collection for this project primarily involved accessing historical records of SpaceX Falcon 9 launches through the SpaceX API. The API provided comprehensive information regarding launch details, including mission outcomes and landing statuses.

- **Perform data wrangling**
  Upon collecting the raw data, a data wrangling process was initiated to ensure its suitability for analysis. This involved cleaning the dataset by handling missing values, removing duplicates, and standardizing formats to facilitate further processing.

- **Perform exploratory data analysis (EDA) using visualization and SQL**
  Exploratory Data Analysis (EDA) serves as a crucial step in understanding the dataset's characteristics and uncovering insights that can inform subsequent analysis and modeling decisions. In this phase of the project, we leverage visualization techniques and SQL queries to gain valuable insights into the SpaceX launch data.

- **Perform interactive visual analytics using Folium and Plotly Dash**
  Interactive visual analytics represent a powerful tool for stakeholders to explore and manipulate data in real-time, enabling faster and more effective identification of patterns and insights. In this module, we leverage Folium and Plotly Dash to create interactive maps and dashboards, facilitating the exploration of SpaceX launch site data and enhancing the understanding of key variables.

- **Perform predictive analysis using classification models**
  Predictive analysis using classification models involves building a machine learning pipeline to predict the outcome of Falcon 9 first stage landings. This process encompasses several key steps, including preprocessing, model training, hyperparameter tuning, and evaluation.

Data Collection

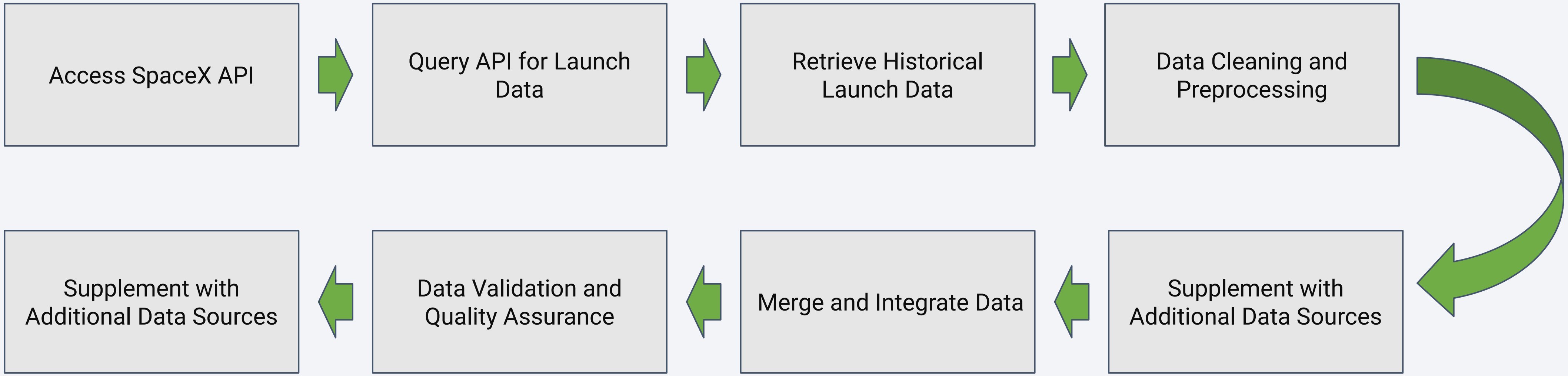
How data sets were collected

The datasets for this project were primarily collected through the SpaceX API, which provides access to comprehensive historical records of SpaceX Falcon 9 launches. The process of data collection involved the following steps:

1. Accessing the SpaceX API: The project team accessed the SpaceX API, which serves as a centralized repository for information regarding Falcon 9 launches. The API allows users to retrieve detailed data about each launch, including mission outcomes, landing statuses, launch dates, payloads, and other relevant information.
2. Retrieving Historical Launch Data: Using the endpoints provided by the SpaceX API, the team retrieved historical records of Falcon 9 launches. This involved querying the API for specific parameters, such as launch dates, launch sites, and mission outcomes, to gather relevant data for analysis.
3. Data Cleaning and Preprocessing: Upon retrieving the raw data from the SpaceX API, a data cleaning and preprocessing phase was initiated. This involved handling missing values, removing duplicates, and standardizing formats to ensure the dataset's integrity and suitability for analysis.
4. Supplementing Data Sources: In addition to the SpaceX API, supplementary data sources such as web scraping of relevant Wiki pages may have been utilized to gather additional information about Falcon 9 launches. This could include details about launch sites, mission objectives, payload information, and other relevant variables.

Data Collection

Data collection process flowcharts



Data Collection – SpaceX API

The capstone project, the main objective was to collect data from the SpaceX API and perform basic data wrangling and formatting. The context of the project revolves around predicting the success of Falcon 9 first stage landings, which is crucial for determining launch costs and competing against SpaceX in the rocket launch market.

Key Activities:

Request to the SpaceX API:
- GET request to the SpaceX API to retrieve data on Falcon 9 launches.
- The data obtained from the API included information about successful and unsuccessful landings, launch dates, mission outcomes, and other relevant details.

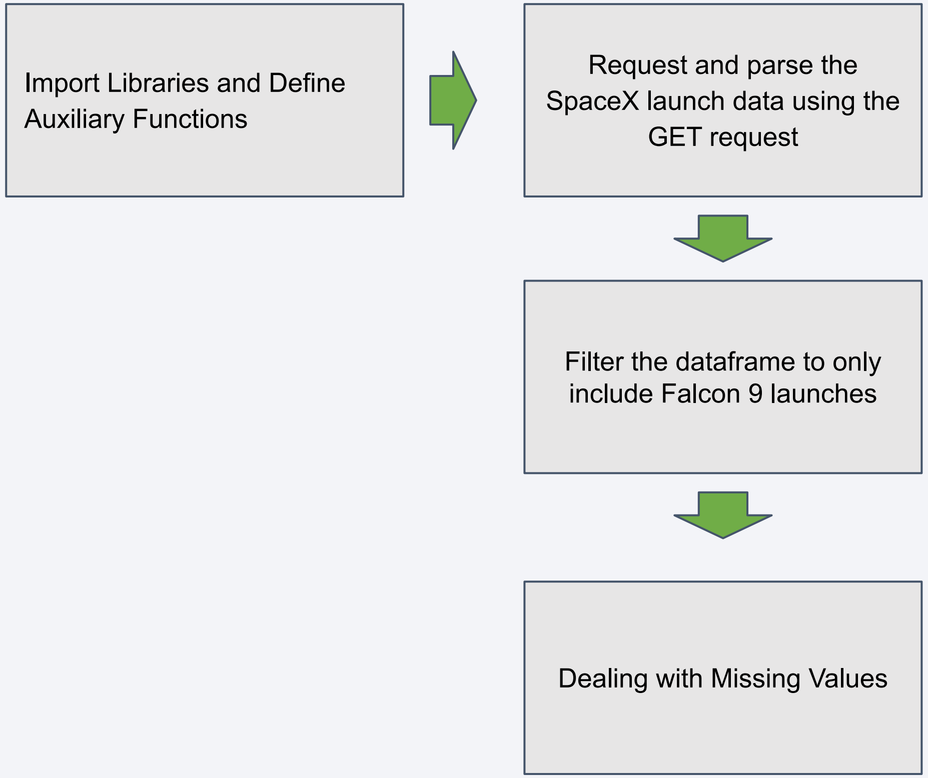
Cleaning the Requested Data:
- Basic data wrangling techniques were applied to clean the retrieved data.
- This involved handling missing values, removing duplicates, and ensuring that the data is in the correct format for further analysis.

Insights:
- The examples provided illustrated both successful and unsuccessful landings of Falcon 9 first stages, emphasizing the importance of predicting landing outcomes accurately.

- Understanding the context of planned controlled landings in the oceans by SpaceX further highlighted the complexities of analyzing landing data.

Github Url :
https://github.com/putra-asmarjoe/CapsW5/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

flowchart



Data Collection – Scraping

The objective was to perform web scraping to collect historical launch records of Falcon 9 rockets from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches." This data extraction process aimed to gather essential information for further analysis and prediction of Falcon 9 first stage landings.

Key Activities:

Web Scraping with BeautifulSoup:

- Students utilized the BeautifulSoup library to extract data from the Wikipedia page.
- The target was to extract the HTML table containing Falcon 9 launch records.

Parsing and Data Conversion:

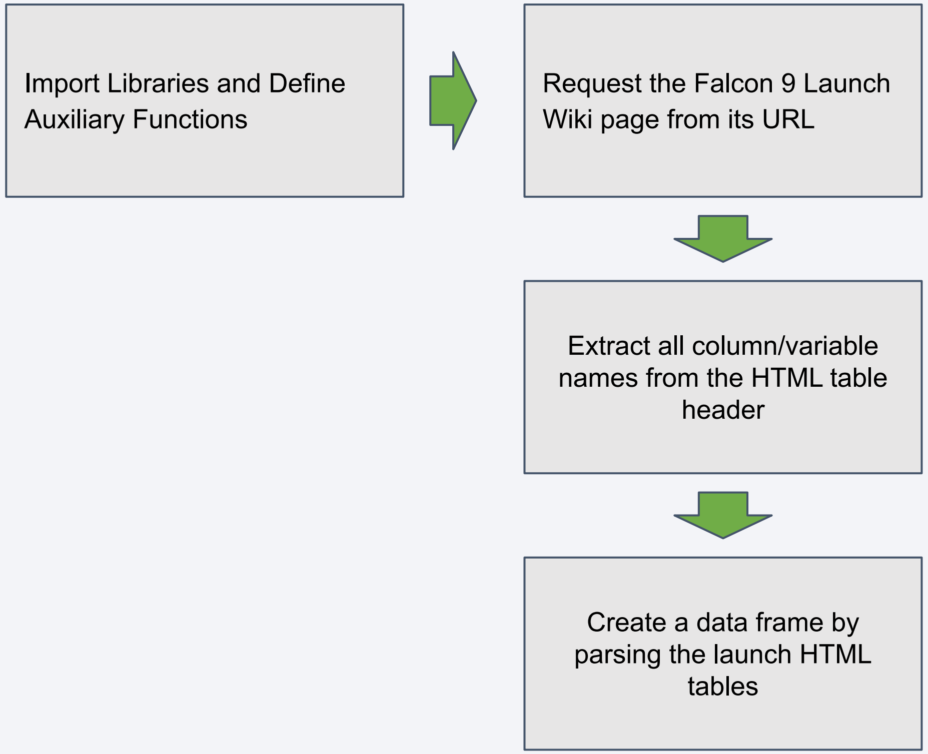
- Extracted HTML table data was parsed and converted into a structured format using Pandas DataFrame.
- This facilitated data manipulation and analysis for subsequent stages of the project.

Insights:

- The Wikipedia page provided a comprehensive list of Falcon 9 and Falcon Heavy launches, offering valuable data for analysis.
- Web scraping techniques allowed for efficient extraction of structured data from unstructured HTML content, enabling further exploration and analysis.

Github Url :
https://github.com/putra-asmarjoe/CapsW5/blob/main/jupyter-labs-webscraping.ipynb

flowchart



Data Wrangling

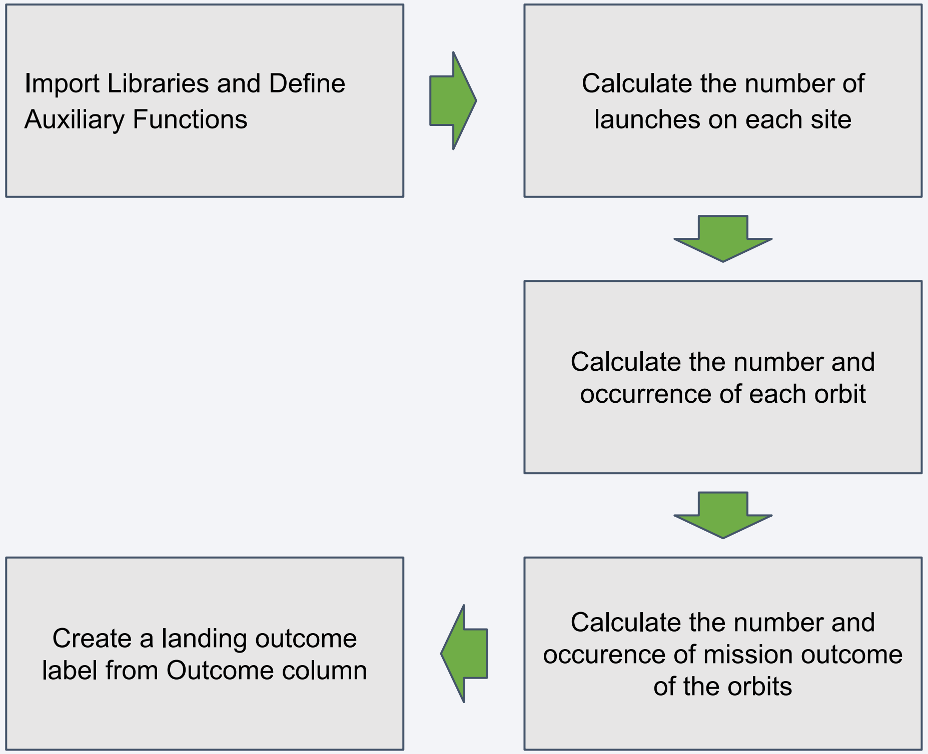
We focus was on data wrangling and exploratory data analysis (EDA) to identify patterns in the dataset and determine suitable training labels for supervised machine learning models. This step is crucial for preparing the data and defining the target variable for predictive modeling of Falcon 9 first stage landings.

Key Activities:

- Exploratory Data Analysis (EDA):
  a. Conducted exploratory data analysis to gain insights into the dataset's characteristics and distribution.
  b. Analyzed various variables and attributes to understand their significance and potential impact on the outcome of Falcon 9 landings.
- Determine Training Labels:
  a. Defined training labels based on the outcomes of Falcon 9 landings.
  b. Converted different landing outcomes (e.g., True Ocean, False RTLS) into binary labels, where 1 represents successful landings and 0 represents unsuccessful landings.
- Data Wrangling:
  a. Described the data processing steps undertaken to clean and prepare the dataset for analysis.
  b. This included handling missing values, standardizing formats, and transforming variables as needed.

Github Url :
https://github.com/putra-asmarjoe/CapsW5/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

flowchart



EDA with Data Visualization

Summary of the charts plotted during the exploratory data analysis (EDA) and the reasons why they were used:

1. **FlightNumber vs. PayloadMass Scatter Plot**: This chart was used to visualize the relationship between the flight number and the payload mass. It helps to identify any patterns or trends in payload mass across different flight numbers.
2. **FlightNumber vs. LaunchSite Scatter Plot**: This chart was used to visualize the relationship between the flight number and the launch site. It helps to identify any patterns or trends in launch site selection over time.
3. **Payload vs. LaunchSite Scatter Plot**: This chart was used to visualize the relationship between the payload mass and the launch site. It helps to identify any patterns or trends in payload mass across different launch sites.
4. **Bar Chart of Success Rate by Orbit**: This chart was used to visualize the success rate of launches for different orbit types. It helps to identify which orbits have higher success rates compared to others.
5. **FlightNumber vs. Orbit Scatter Plot**: This chart was used to visualize the relationship between the flight number and the orbit type. It helps to identify any patterns or trends in orbit selection over time.
6. **Payload vs. Orbit Scatter Plot**: This chart was used to visualize the relationship between the payload mass and the orbit type. It helps to identify any patterns or trends in payload mass across different orbit types.
7. **Line Chart of Launch Success Yearly Trend**: This chart was used to visualize the trend in launch success rates over the years. It helps to identify whether there has been an overall improvement or decline in launch success rates over time.

These charts were selected to explore various aspects of the dataset, including the relationship between different variables and the overall trend in launch success rates. They provide valuable insights into the factors that may influence the success of space launches.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Summary of the SQL queries performed during the EDA with SQL:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. List the date when the first successful landing outcome in ground pad was achieved
5. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
6. List the total number of successful and failure mission outcomes
7. List the names of the booster versions which have carried the maximum payload mass, using a subquery
8. List the records displaying the month names, failure landing outcomes in drone ship, booster versions, and launch sites for the months in the year 2015
9. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order
10. Display the total payload mass carried by boosters launched by NASA (CRS)

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

In this analysis, the following map objects were created and added to a Folium map:

Markers: Markers were added to represent the launch sites and the closest coastline points. Each marker displayed information such as the launch site name or the distance to the coastline in a popup.

Circles: Circles were added to represent the launch sites, with each circle centered at the corresponding latitude and longitude coordinates. These circles provided a visual indication of the area covered by each launch site.

Polylines: Polylines were drawn between the launch sites and the closest coastline points to illustrate the distances between them. These polylines provided a clear visual representation of the proximity between the launch sites and the coastline.

Overall, these map objects were used to visualize the locations of launch sites, their distances to the coastline, and the connections between them, allowing for a better understanding of their spatial relationships.

By adding these objects to the map, we can visually explore the spatial distribution of launch sites, their relationships with the coastline, and gain insights into factors that may influence launch site selection and operations.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Map Preview Url : https://nbviewer.org/github/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

In the Plotly Dash dashboard, I added the following plots and interactions:

1. Dropdown for Launch Site Selection: Users can select a specific launch site or view data for all sites. This interaction allows users to focus on specific launch locations or analyze data across all sites.
2. Pie Chart for Success Counts: Displays the total counts of successful launches across all sites or for a specific site selected using the dropdown. This plot provides a quick overview of launch success rates and helps identify trends among different launch sites.
3. Range Slider for Payload Selection: Allows users to select a payload range of interest, enabling the exploration of how payload mass correlates with launch outcomes. This interaction helps users identify any patterns or relationships between payload mass and launch success.
4. Scatter Plot for Payload vs. Outcome: Shows the correlation between payload mass and launch outcomes (success or failure). Users can observe how payload mass affects the likelihood of a successful launch, with the ability to filter data by launch site using the dropdown. This plot provides deeper insights into the relationship between payload characteristics and mission outcomes.

I added these plots and interactions to provide users with a comprehensive and interactive visualization of SpaceX launch data. The dropdown menu allows users to customize their analysis by selecting specific launch sites, while the range slider enables exploration of payload characteristics. The pie chart and scatter plot offer visual insights into launch success rates and payload-outcome relationships, empowering users to make data-driven decisions and gain a better understanding of SpaceX's launch operations.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/spacex_dash_app.py

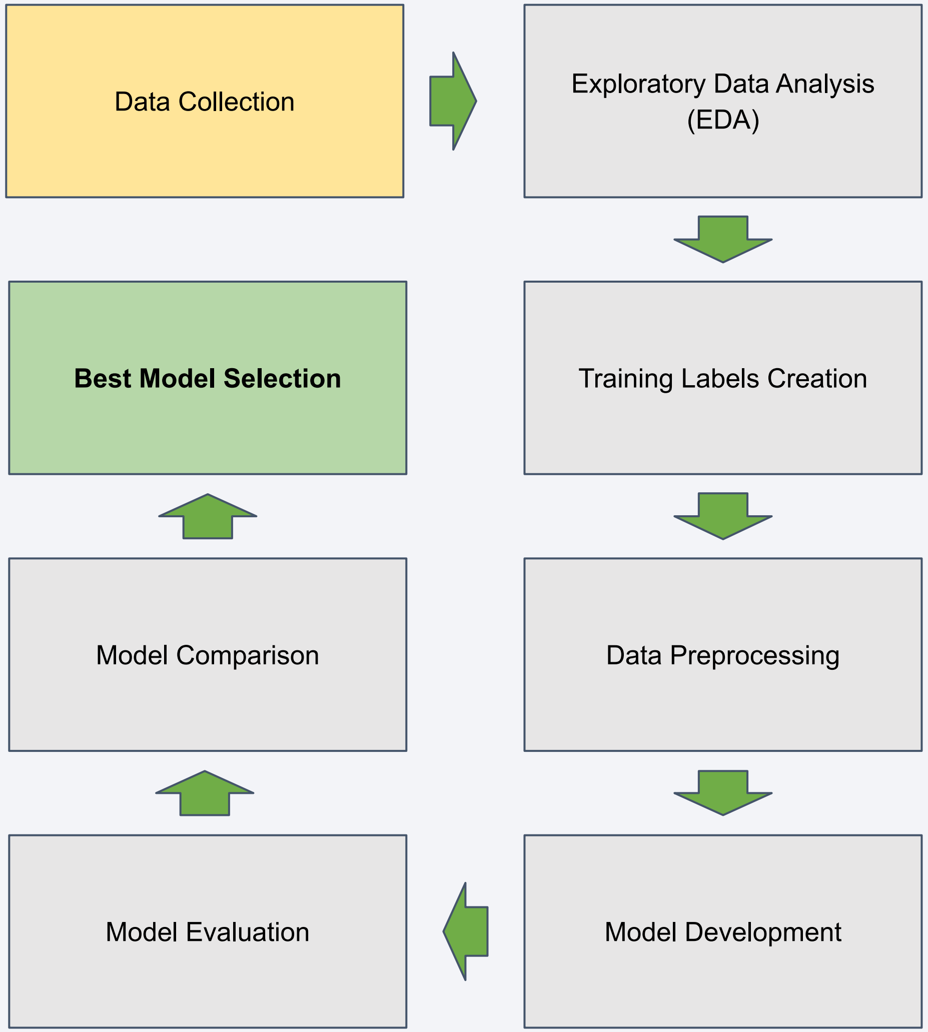
Predictive Analysis (Classification)

By exploring historical data on launch outcomes and associated variables, we seek to identify patterns and insights that can inform decision-making in future space missions. summary of the tasks performed in the SpaceX Falcon 9 First Stage Landing Prediction assignment:

1. Perform exploratory Data Analysis and determine Training Labels: Created a column for the class based on landing outcomes.
2. Standardize the data: Standardized the input features.
3. Split into training data and test data:Split the data into training and test sets.
4. Find best Hyperparameter for SVM, Classification Trees, and Logistic Regression: Utilized GridSearchCV to find the best hyperparameters for SVM, Decision Trees, Logistic Regression, and K-Nearest Neighbors classifiers.
5. Evaluate model performance:
    - Calculated the accuracy on the test data for each model.
    - Visualized the confusion matrices for each model.
6. Summary of Model Performance:
    - Logistic Regression: Accuracy on test data: 83.33%
    - Support Vector Machine (SVM): Accuracy on test data: 83.33%
    - Decision Tree Classifier: Accuracy on test data: 72.22%
    - K-Nearest Neighbors (KNN): Accuracy on test data: 83.33%

Github Url :
https://github.com/putra-asmarjoe/CapsW5/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

flowchart



Results

Overall, the exploratory data analysis provides valuable insights into SpaceX's launch operations, including the factors influencing launch success, payload characteristics, and geographic considerations. These insights serve as a foundation for further analysis and model development in predicting Falcon 9 first stage landing outcomes

Launch Success Rate:
- We analyzed the overall launch success rate and found that a majority of Falcon 9 rocket launches were successful.
- The success rate provides valuable insights into the reliability and performance of SpaceX's launch operations.

Payload Mass Distribution:
- We examined the distribution of payload masses carried by Falcon 9 rockets.
- The payload mass distribution helps understand the range and variability of payloads launched by SpaceX.

Launch Sites Analysis:
- We explored the distribution of launch sites used by SpaceX, including Cape Canaveral Air Force Station (CCAFS) and Vandenberg Air Force Base (VAFB).
- Understanding the distribution of launch sites provides insights into geographic factors influencing launch operations.

Orbit Types:
- We investigated the distribution of different orbit types targeted by SpaceX missions, such as Low Earth Orbit (LEO) and Geostationary Transfer Orbit (GTO).
- Analyzing orbit types helps understand the diversity of mission objectives and destinations.

Mission Outcomes:
- We analyzed the outcomes of SpaceX missions, including successful and unsuccessful launches.
- Understanding mission outcomes is crucial for evaluating the overall performance and reliability of SpaceX's launch operations.

Correlation Analysis:
- We explored correlations between variables such as payload mass, launch site, and mission outcomes.
- Identifying correlations helps uncover potential relationships and dependencies within the dataset.

Visualization:
- We utilized various visualization techniques, including histograms, scatter plots, and bar charts, to illustrate key findings and trends in the data.
- Visualization enhances the interpretation of data and facilitates communication of insights to stakeholders.

Conclusion:
The best-performing model is the **SVM classifier with a sigmoid kernel**, achieving an a**ccuracy of 83.33%** on the test data.
Overall, the assignment involved preprocessing the data, building classification models, tuning hyperparameters, and evaluating model performance to predict the success of Falcon 9 first stage landing.

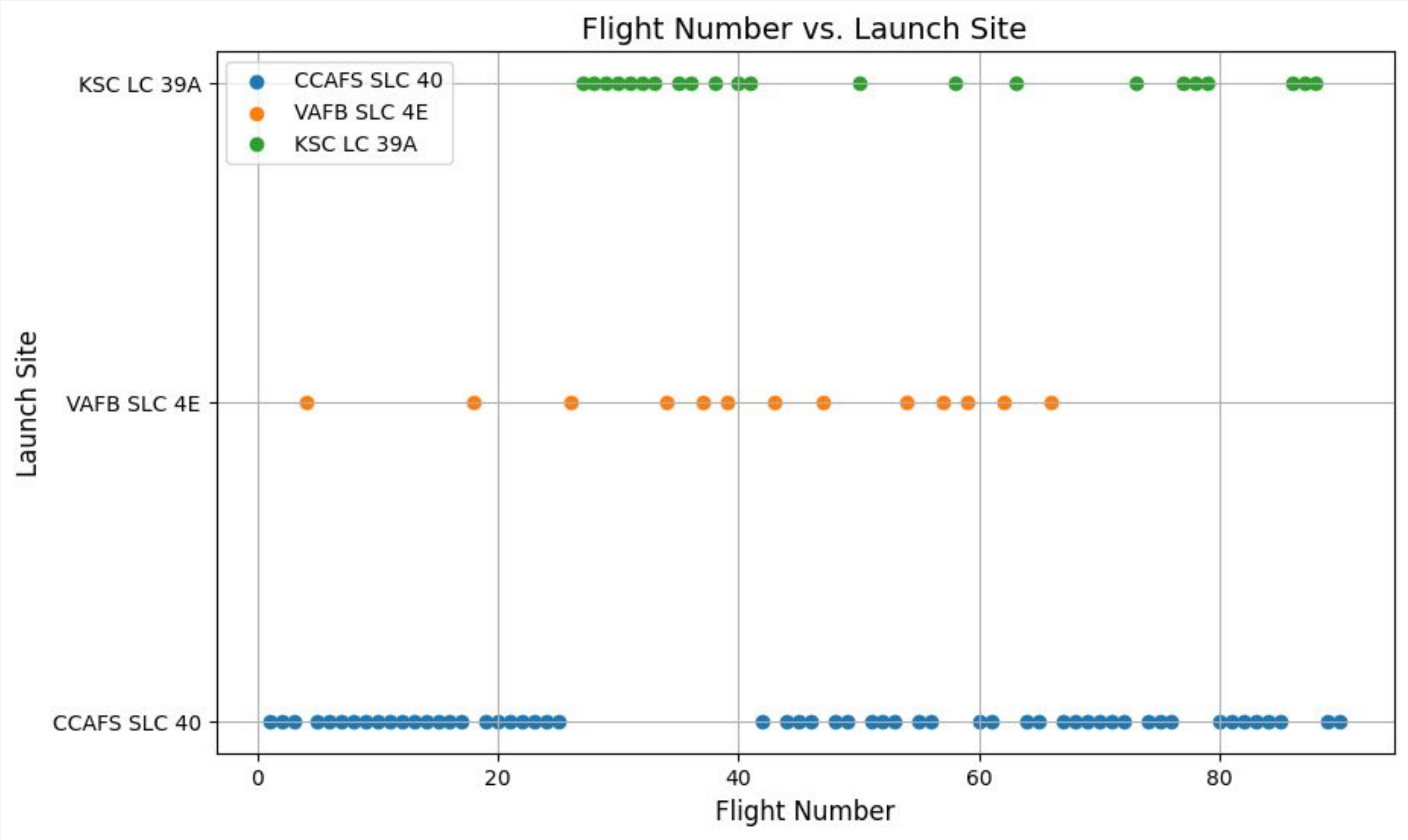
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

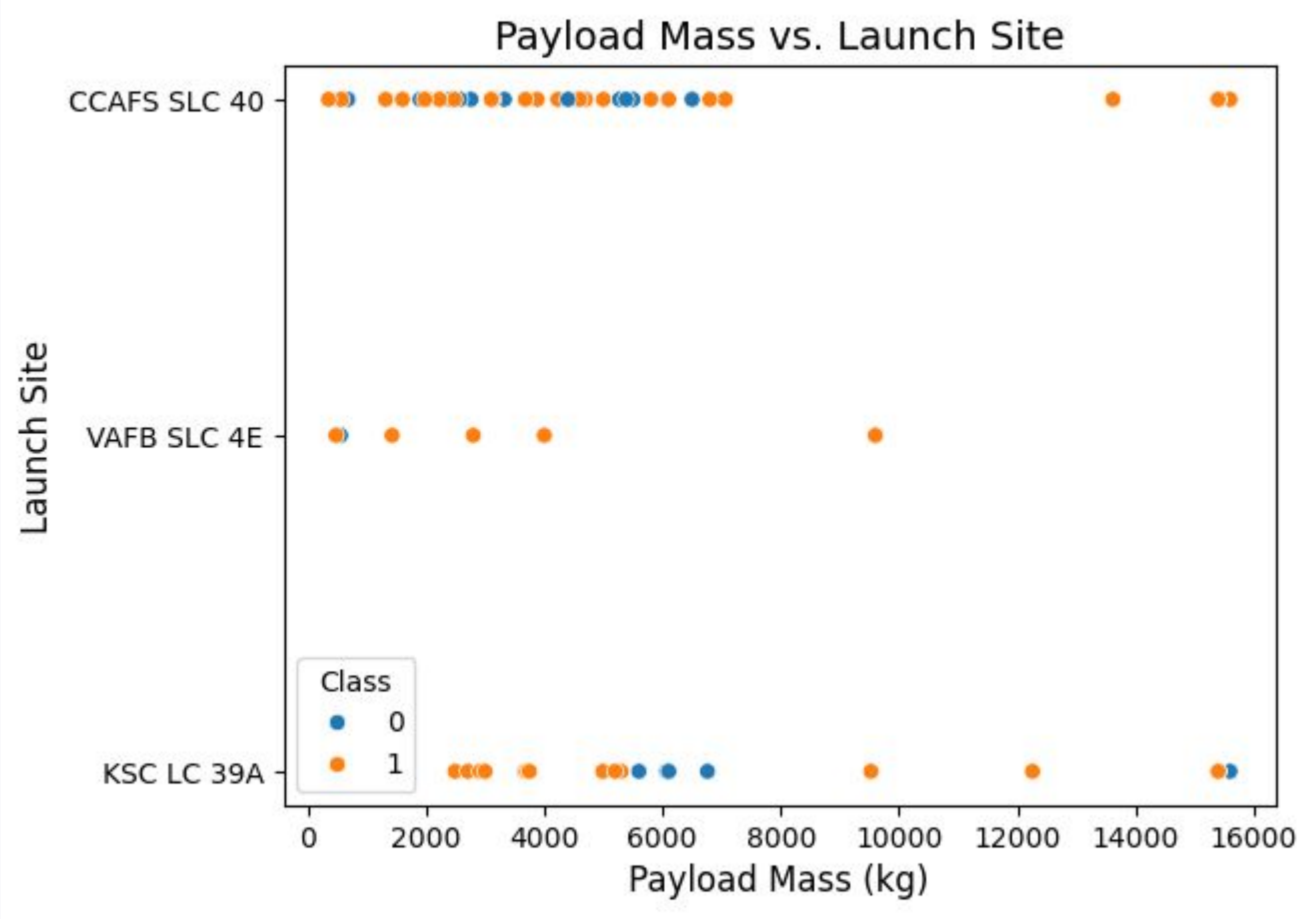
Each point on the plot represents a specific launch, with the x-axis representing the flight number and the y-axis representing the launch site. Each launch site is distinguished by a different color, with a legend indicating the mapping between launch sites and colors.



Payload vs. Launch Site

Observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

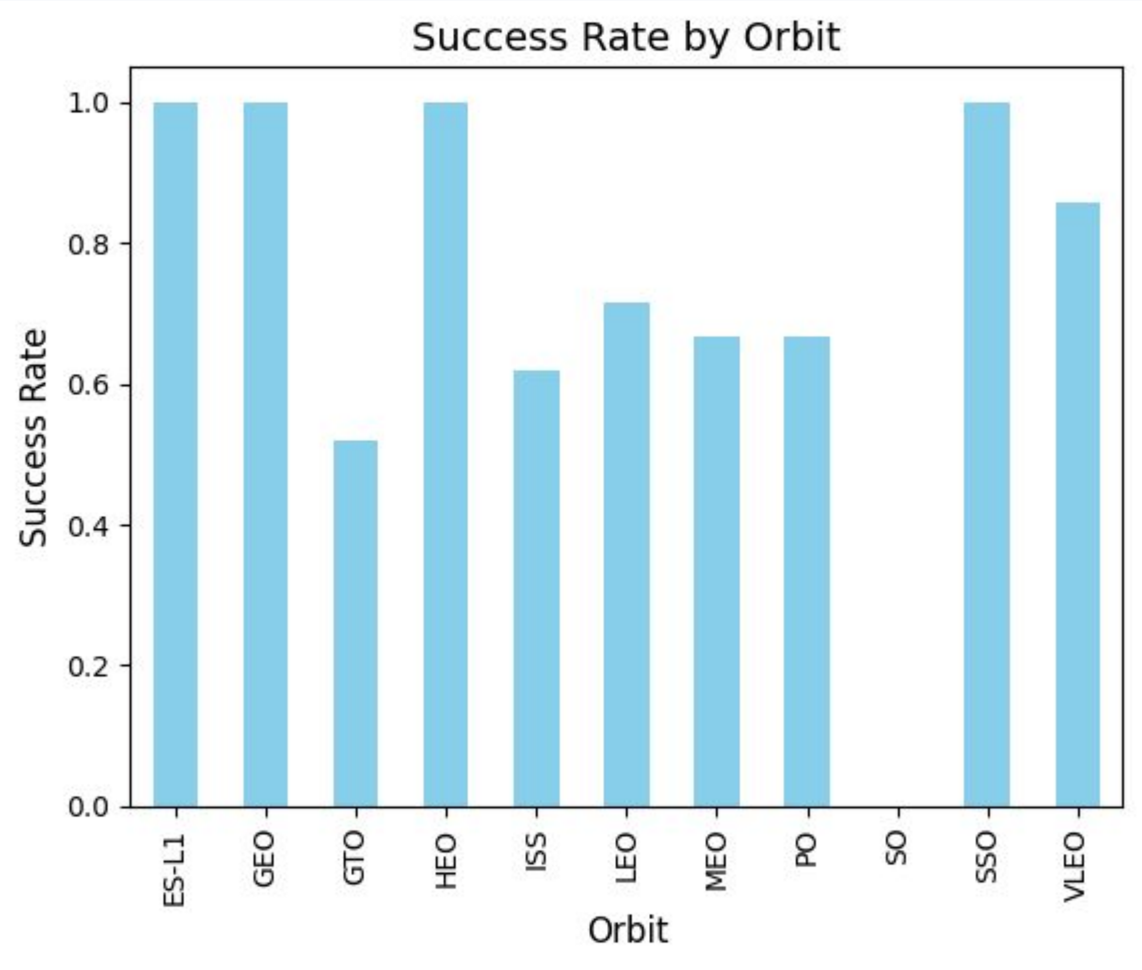
a scatter plot that visualizes the relationship between the payload mass and launch site, with points color-coded based on the launch outcome (success or failure). However, the task description mentions observing the relationship between the success rate of each orbit type, which suggests that the 'Class' variable may not be suitable for this analysis.



Success Rate vs. Orbit Type

The provided code calculates the mean success rate for each orbit type using the groupby method on the 'Orbit' column and then visualizes this data using a bar chart. This bar chart helps identify which orbits have a high success rate.

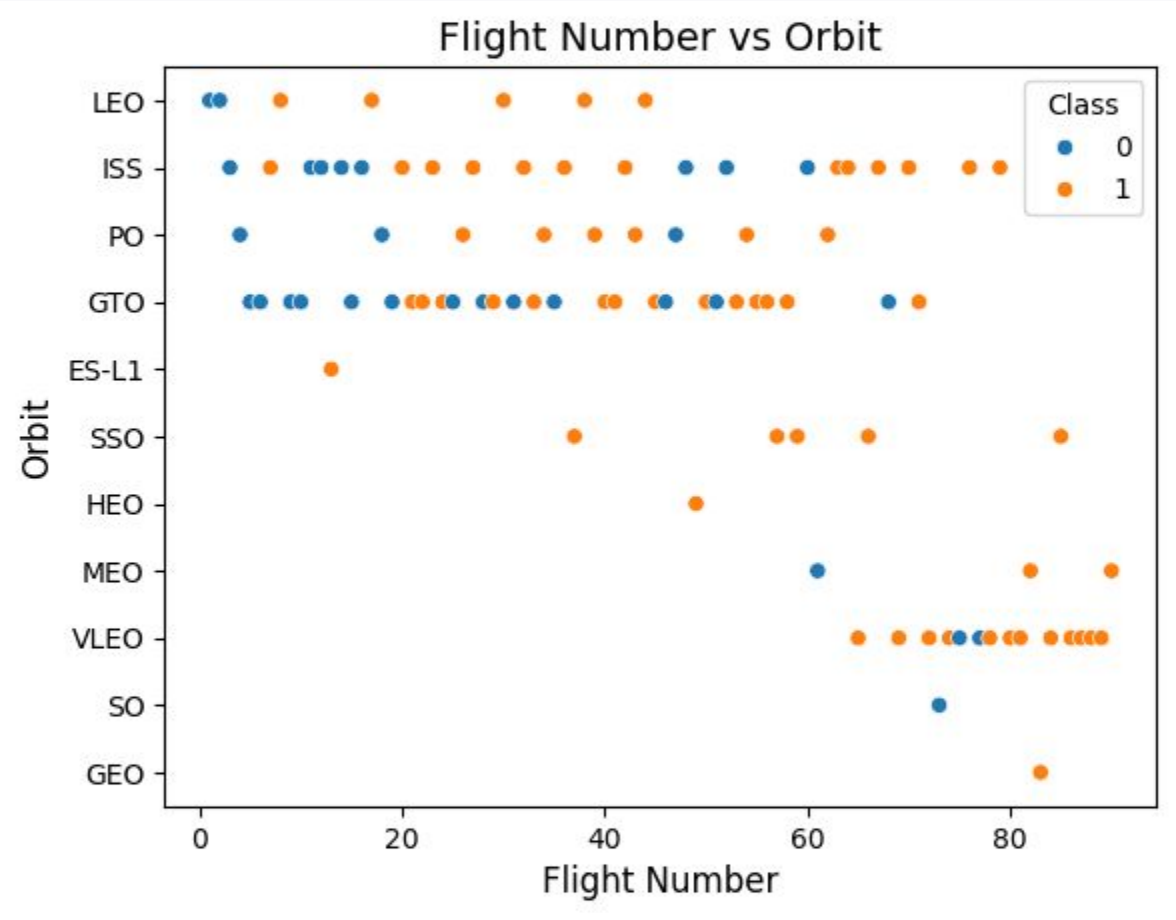
However, the analysis could be enhanced by including additional information, such as confidence intervals or error bars, to indicate the uncertainty in the success rate estimates. Additionally, it would be beneficial to compare the success rates across different orbits visually.



Flight Number vs. Orbit Type

The provided code creates a scatter plot to visualize the relationship between FlightNumber and Orbit type. The x-axis represents the FlightNumber, the y-axis represents the Orbit, and the hue represents the Class value (success or failure).

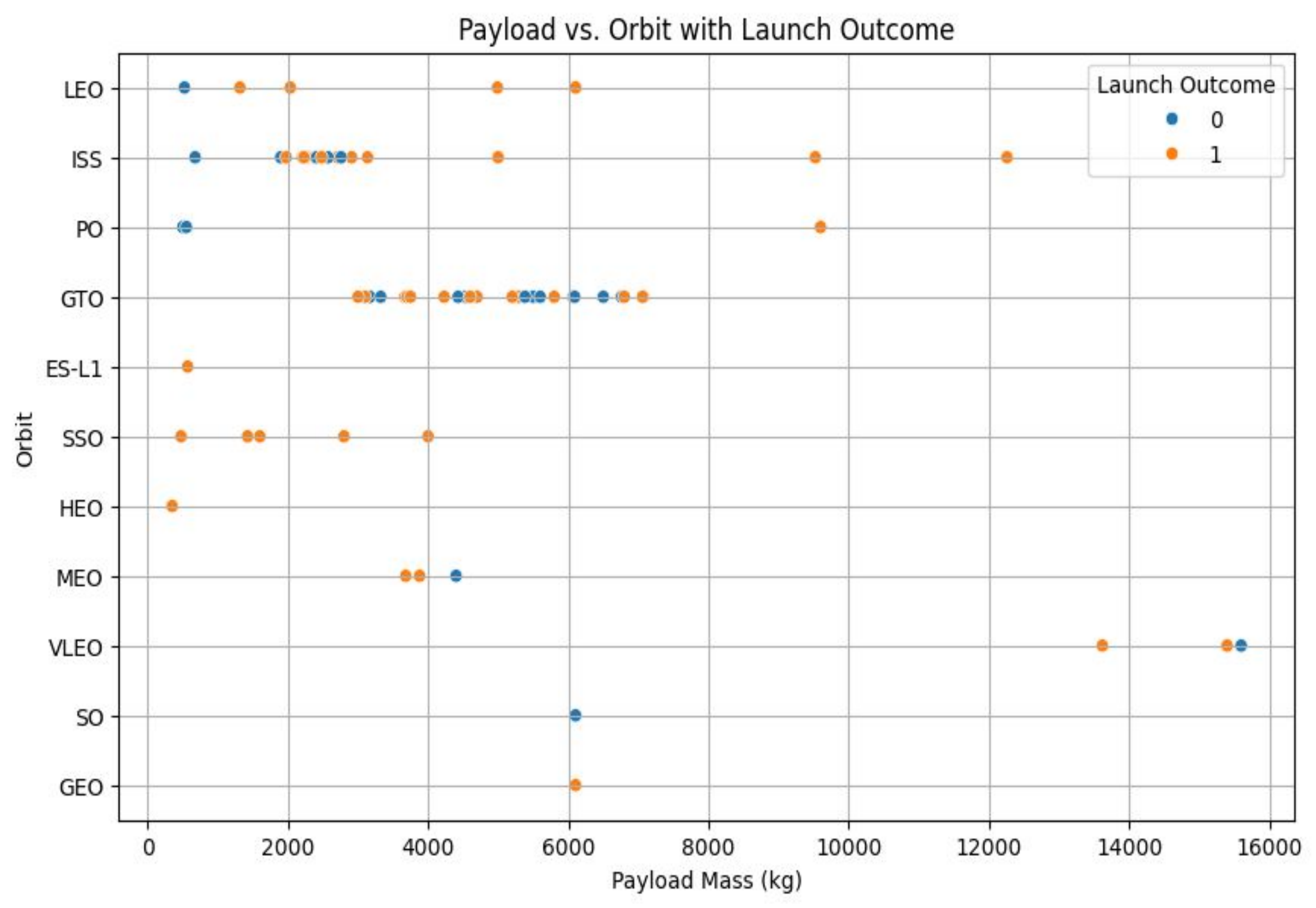
This scatter plot helps visualize the relationship between FlightNumber and Orbit type. It shows whether there is any apparent pattern or correlation between the two variables across different orbits. For example, in the LEO (Low Earth Orbit), the success appears related to the number of flights, while in the GTO (Geostationary Transfer Orbit), there seems to be no relationship between flight number and orbit success.



Payload vs. Orbit Type

scatter plot to visualize the relationship between Payload and Orbit type. The x-axis represents the Payload Mass, the y-axis represents the Orbit, and the hue represents the Class value (success or failure)

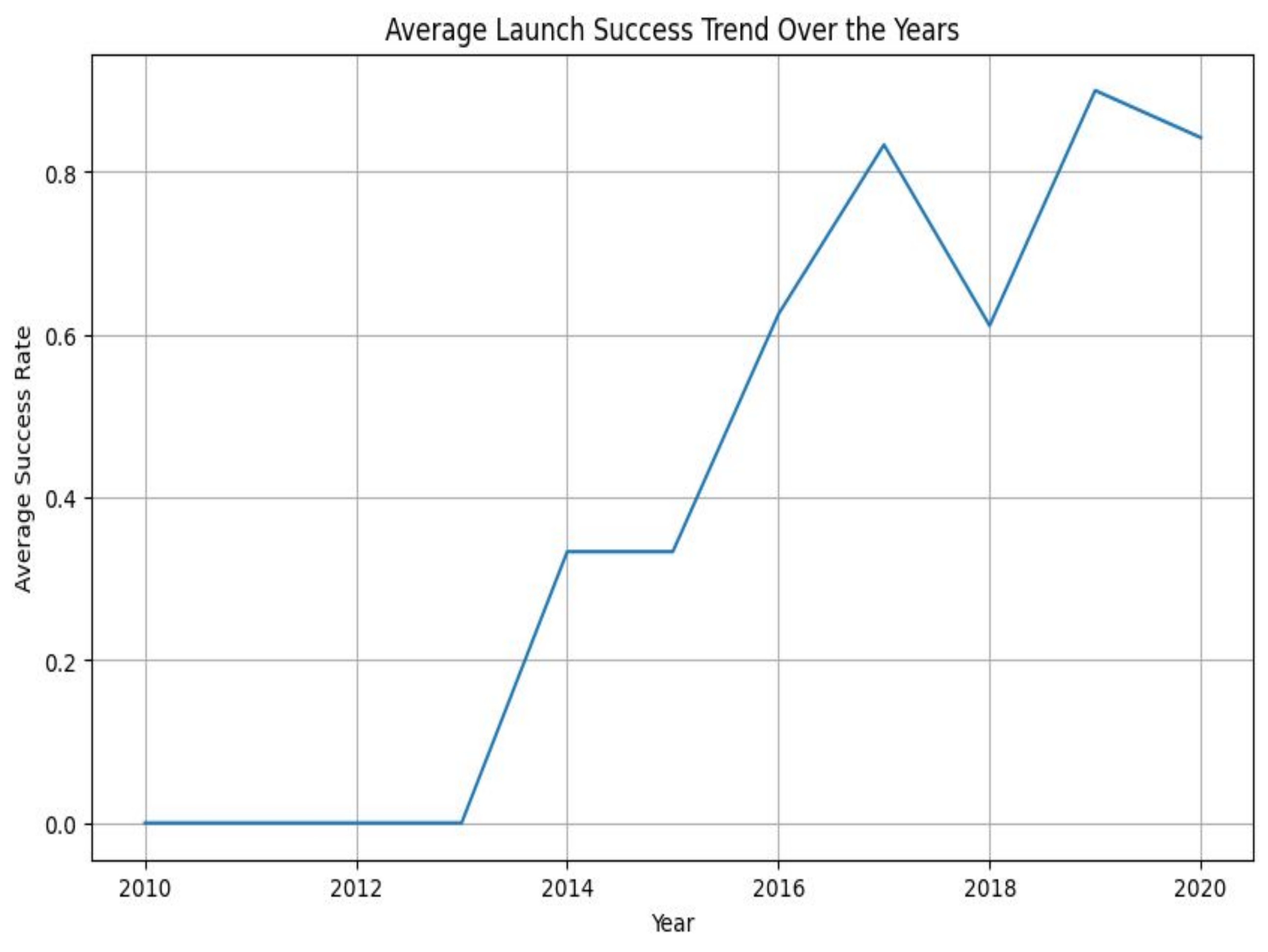
This scatter plot helps visualize the relationship between Payload and Orbit type, while also indicating the launch outcome (success or failure) with different hues. It provides insights into how different orbits are affected by payload mass and the corresponding launch outcomes.



Launch Success Yearly Trend

Performs feature engineering by extracting the year from the Date column and then calculates the success rate for each year. Finally, it plots a line chart to visualize the average launch success trend over the years.

This line chart helps visualize the average launch success trend over the years. From the plot, it's evident that the success rate has been increasing steadily since 2013 until around 2020, indicating an overall improvement in SpaceX's launch success rate over the years.



All Launch Site Names

*"**SELECT** **DISTINCT(Launch_Site) FROM SPACEXTABLE**"*

The SQL query retrieves unique values from the column Launch_Site in the table SPACEXTABLE. Here's what each part of the query means:

- SELECT DISTINCT(Launch_Site): This part of the query specifies that we want to select unique values from the column Launch_Site.
- FROM SPACEXTABLE: This specifies the table from which we are selecting the data, which in this case is SPACEXTABLE.

So, when you execute this query, it will return a list of unique launch site names from the Launch_Site column in the SPACEXTABLE table. If there are multiple records with the same launch site name, only one instance of that name will be returned in the result set.

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

Launch Site Names Begin with 'CCA'

“SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5”

The SQL query retrieves 5 records from the SPACEXTABLE where the Launch_Site column starts with the string 'CCA'.

Overall, the query returns a subset of data containing 5 records where the Launch_Site column starts with the string 'CCA'. This allows for quick retrieval and analysis of specific data related to launch sites with names beginning with 'CCA'.

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

"SELECT SUM(PAYLOAD_MASS__KG_) as total FROM SPACEXTABLE WHERE Customer LIKE 'NASA%'"

The SQL query retrieves the total payload mass carried by boosters launched by NASA (CRS) from the SPACEXTABLE.

Overall, the query returns the total payload mass carried by boosters launched by NASA (CRS) by summing up the payload mass values from relevant records in the SPACEXTABLE. **Total Payload Mass is 99.980.**

```
[16]: %sql select SUM(PAYLOAD_MASS__KG_) as total FROM SPACEXTABLE WHERE Customer LIKE 'NASA%'

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

[16]:

total
99980

Average Payload Mass by F9 v1.1

“SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1'“

The SQL query retrieves the average payload mass carried by boosters with the version 'F9 v1.1' from the SPACEXTBL.

Overall, the query returns the average payload mass carried by boosters with the version 'F9 v1.1' from the SPACEXTBL.

Average Payload Mas is 2.928,4

```
[19]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
```

[19]:

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

"SELECT MIN(Date) AS First_Successful_Landing_Date FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%'"

The SQL query retrieves the date when the first successful landing outcome on a ground pad was achieved from the SPACEXTBL table.

This means that the first successful landing outcome on a ground pad was **achieved on December 22, 2015.**

```
[22]: %sql SELECT MIN(Date) AS First_Successful_Landing_Date FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%';

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

[22]:

First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

“SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%' AND Payload_Mass__KG_ > 4000 AND Payload_Mass__KG_ < 6000 AND Landing_Outcome LIKE '%drone%';”

The SQL query retrieves the names of the boosters that have successfully landed on a drone ship and have a payload mass greater than 4000 kg but less than 6000 kg from the SPACEXTBL table.

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

Total Number of Successful and Failure Mission Outcomes

“SELECT CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 'Failure' ELSE 'Success' END AS Outcome_Category, COUNT(*) AS Count FROM SPACEXTBL GROUP BY Outcome_Category”

The SQL query calculates the total number of successful and failed mission outcomes from the SPACEXTBL table.

The result of the query is: This indicates that there is 1 failure and 100 successful mission outcomes in the dataset.

Outcome_Category	**Count**
Failure	1
Success	100

Boosters Carried Maximum Payload

"SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);"

The SQL query retrieves the names of booster versions that have carried the maximum payload mass.

The result of the query is a list of booster versions that have carried the maximum payload mass.

Booster_Version	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

“SELECT SUBSTR(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE SUBSTR(Date, 0, 5) = '2015' AND Landing_Outcome LIKE '%Failure%' AND Landing_Outcome LIKE '%drone ship%';”

The records displaying the month names, failure landing outcomes in the drone ship, booster versions, and launch sites for the months in the year 2015 are as follows:

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

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

“SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC;”

The landing outcomes between the dates June 4, 2010, and March 20, 2017, are ranked based on their counts in descending order as follows:

Landing_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

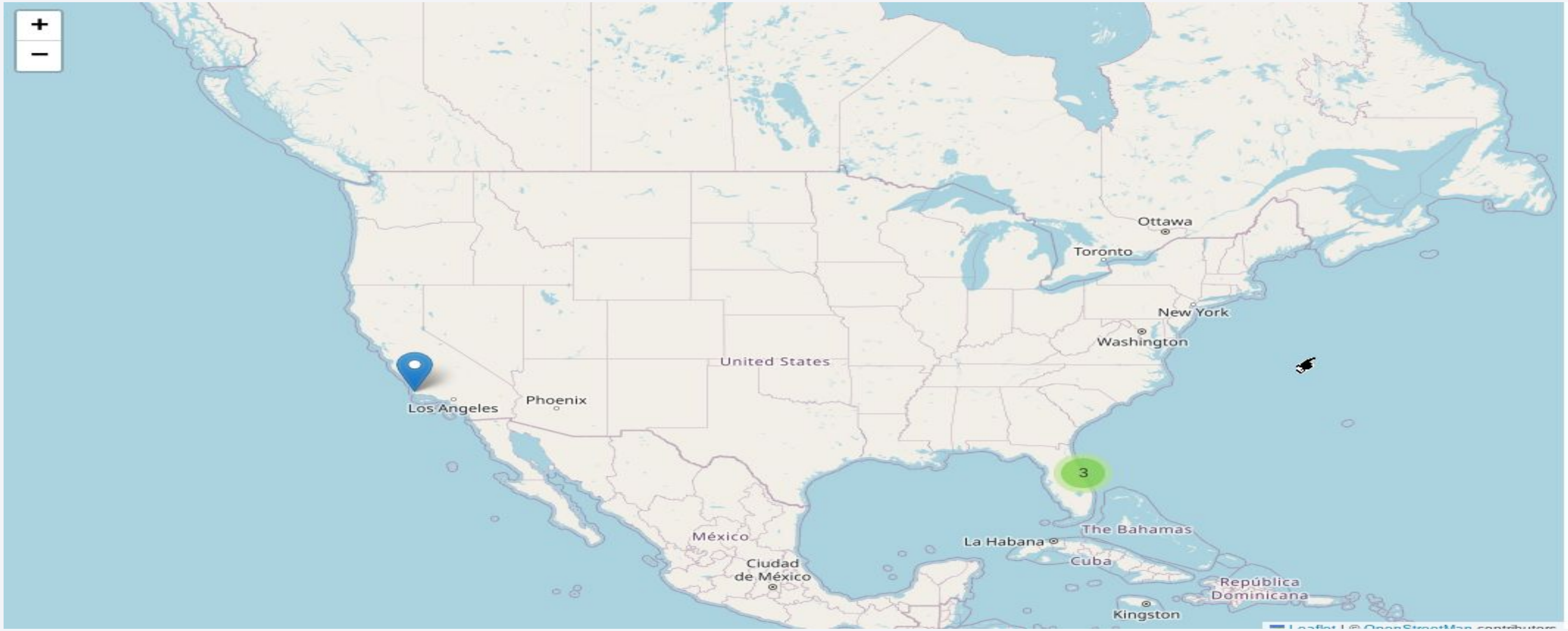
Section 3

Launch Sites Proximities Analysis

Mark all launch sites on a map

The map visualization displays four launch sites: **CCAFS LC-40, CCAFS SLC-40 and KSC LC-39A in Florida**, USA; and **VAFB SLC-4E in Los Angeles**. Each marker on the map represents a launch site, providing a geographical overview of SpaceX's launch infrastructure.

Map PreviewUrl : https://nbviewer.org/github/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb



Spatial Distribution of SpaceX Launch Outcomes

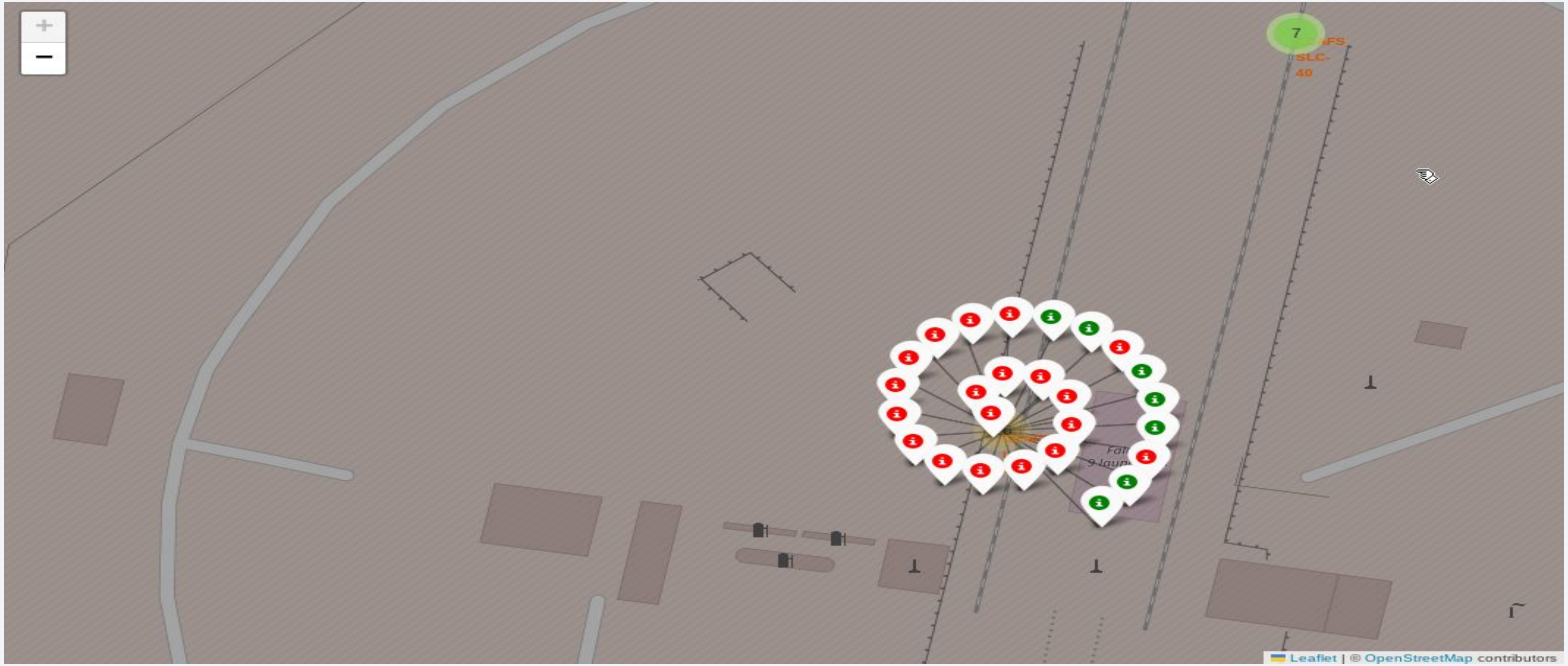
The map visualization illustrates the spatial distribution of SpaceX launch outcomes, depicting a total of 56 launches. Among these, 10 launches occurred in the Los Angeles area, while the remaining 46 launches took place in Florida. This geographical representation provides insights into the success and failure patterns across different launch sites.

Map Preview Url : https://nbviewer.org/github/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Launch Outcomes at CCAFS LC-40 Launch Pad

The subsequent map reveals the launch outcomes at the CCAFS LC-40 launch pad, depicting a total of 26 launches. Among these, 7 launches were successful, while 19 ended in failure. This spatial representation provides detailed insights into the success and failure rates at the specific launch site.

Map Preview Url : https://nbviewer.org/github/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

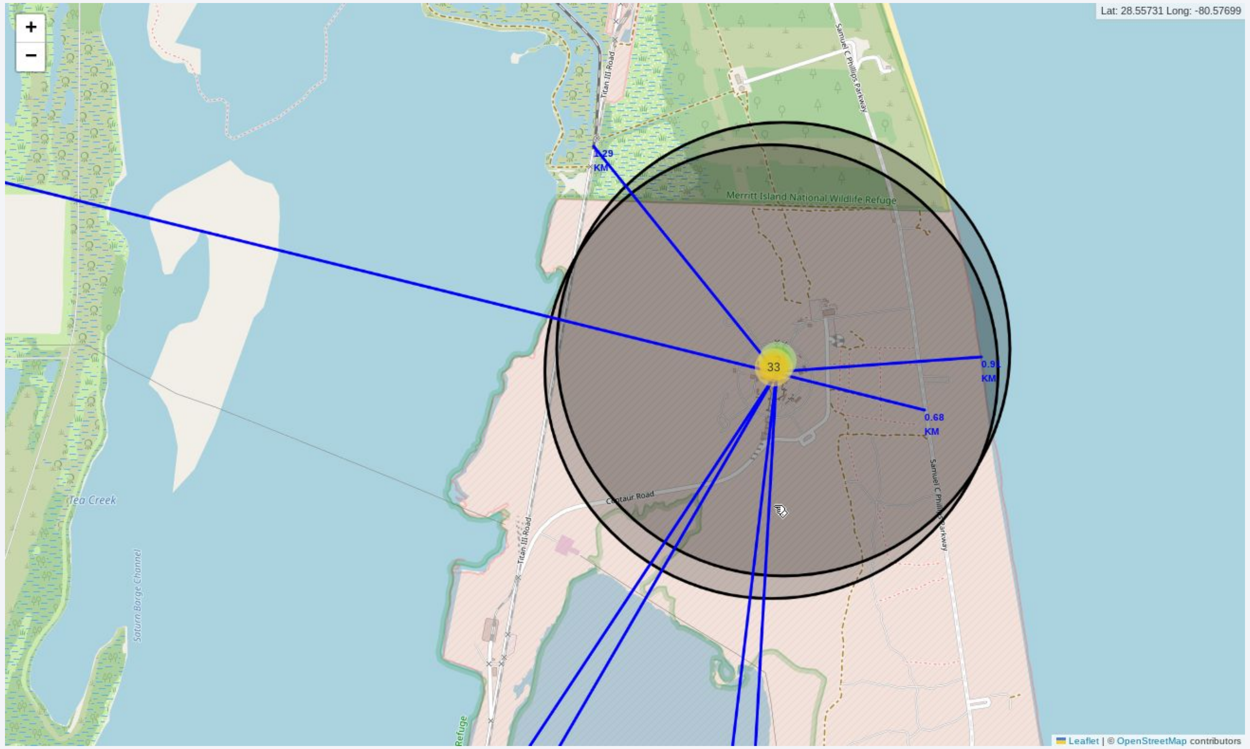


Proximity Analysis of Falcon 9 Launch Pad

In the subsequent map analysis for the Falcon 9 launch pad, we calculated the distances to the nearest coastline point, railway, highway, and five nearest cities. The results indicate the following distances:

- ➔ Nearest coastline point: 0.91 kilometers
- ➔ Nearest highway (Samuel C Philips Parkway): 0.68 kilometers
- ➔ Nearest railway (NASA Railroad): 1.29 kilometers
- ➔ Distance to nearest cities:
    a. Titusville City: 23.23 kilometers
    b. Cocoa City: 27.59 kilometers
    c. Rockledge City: 31.15 kilometers
    d. Melbourne City: 53.91 kilometers

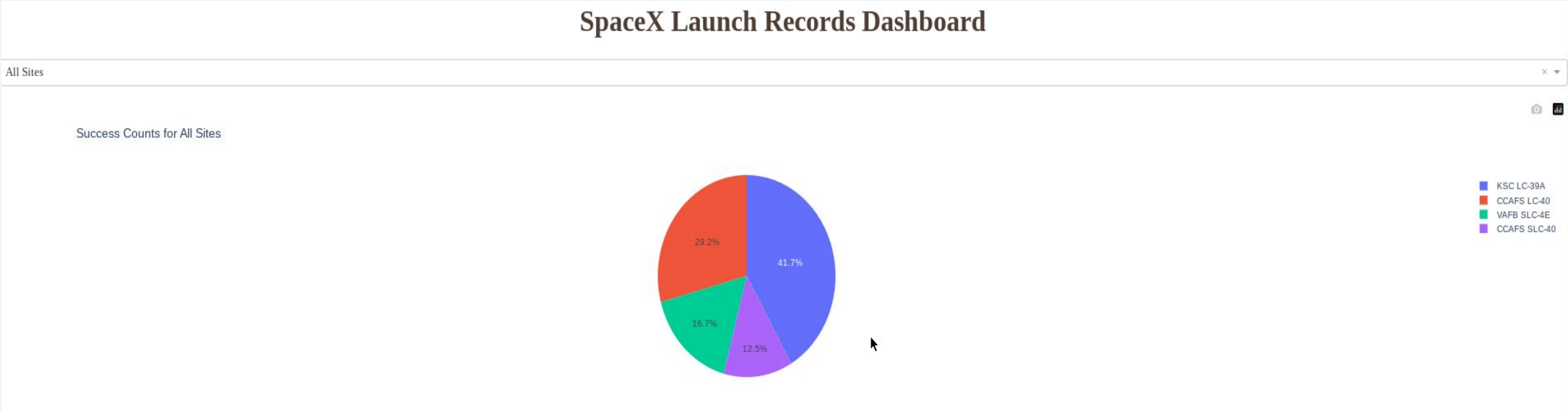
Map Preview Url :
https://nbviewer.org/github/putra-asmarjoe/CapsW5/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb



Section 4

Build a Dashboard with Plotly Dash

Success Rate Analysis of SpaceX Launch Sites

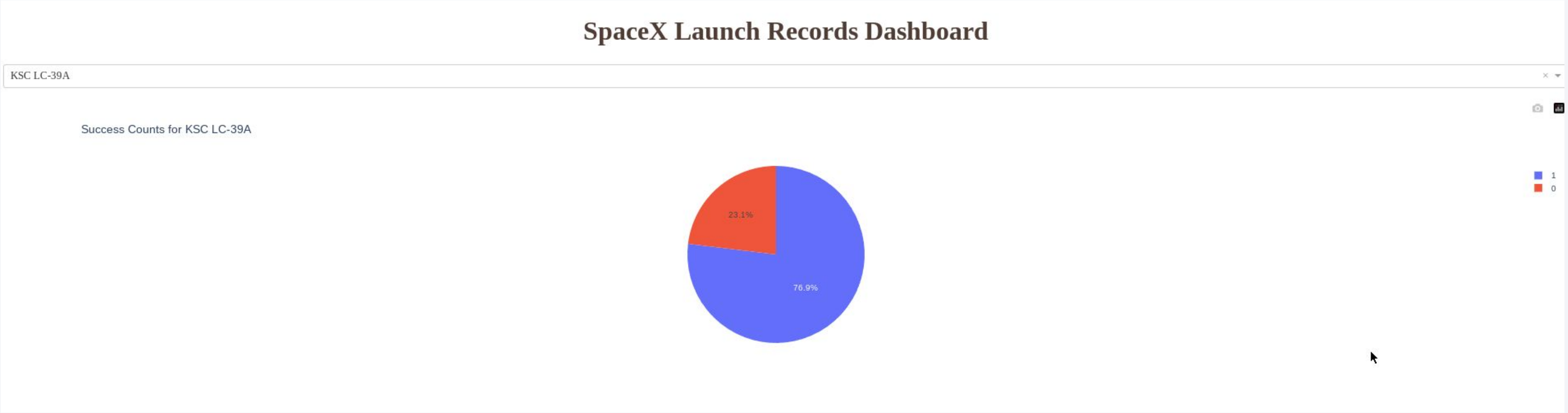


Upon examining the success rates across all four launch sites in the dashboard, we observe the following proportions of successful launches out of a total of 24. These insights provide a comparative analysis of success rates among the different launch sites, with KSC LC-39A exhibiting the highest success rate among them.

- CCAFS LC-40: 7 successes, accounting for 29.2% of the total
- CCAFS SLC-40: 3 successes, representing 12.5% of the total
- KSC LC-39A: 10 successes, constituting 41.7% of the total
- VAFB SLC-4E: 4 successes, comprising 16.7% of the total.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/spacex_dash_app.py

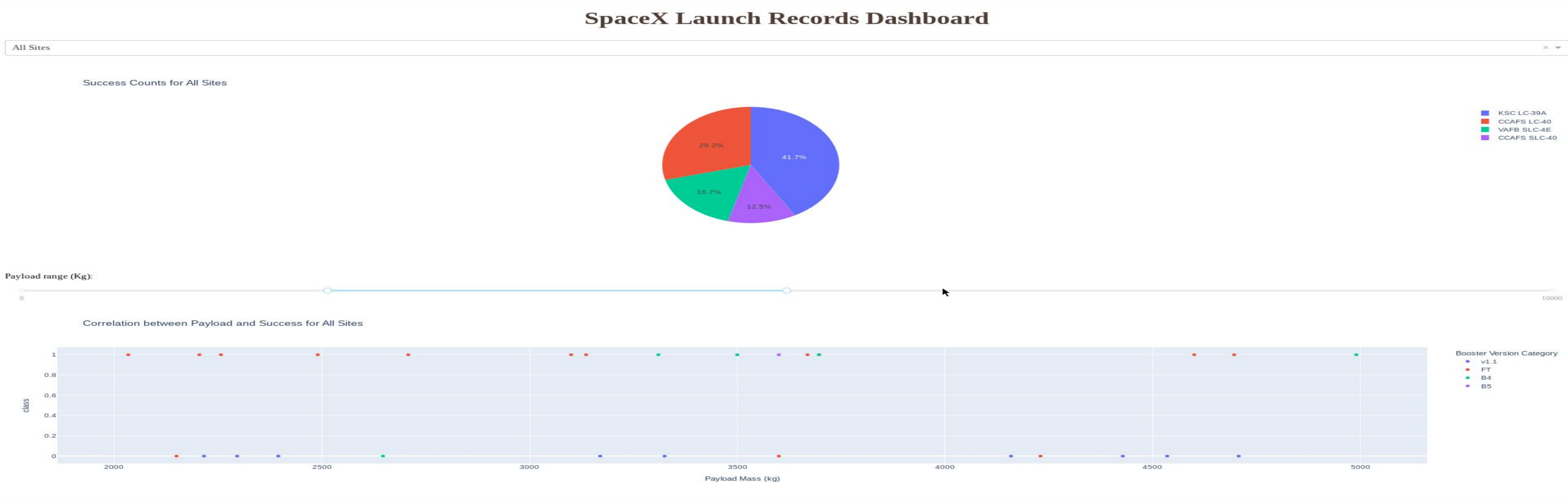
Dashboard View of Launch Success Ratio for Top Launch Site



Upon reviewing the pie chart visualization, it's evident that KSC LC-39A stands out with the highest launch success ratio. Out of 13 launches, 10 were **successful (76.9%)**, while **3 were unsuccessful (23.1%)**.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/spacex_dash_app.py

Dashboard Visualization: Payload vs. Launch Outcome Scatter Plot for All Sites



Using the "All Site" option to display all four launch sites and selecting a payload range of 2000 - 500 kg on the range slider, we observed the following results:

- 15 successful launches with the most frequent booster version being **FT**, contributing to 10 successful launches.
- 13 failed launches with the most frequent failed booster version being **v1.1**, with 9 launch failures.

Github Url : https://github.com/putra-asmarjoe/CapsW5/blob/main/spacex_dash_app.py

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Here is the list of classification model results based on the provided data:

Logistic Regression:
- Cross-validation accuracy: 0.8464
- Test data accuracy: 0.8333

SVM:
- Cross-validation accuracy: 0.8482
- Test data accuracy: 0.8333

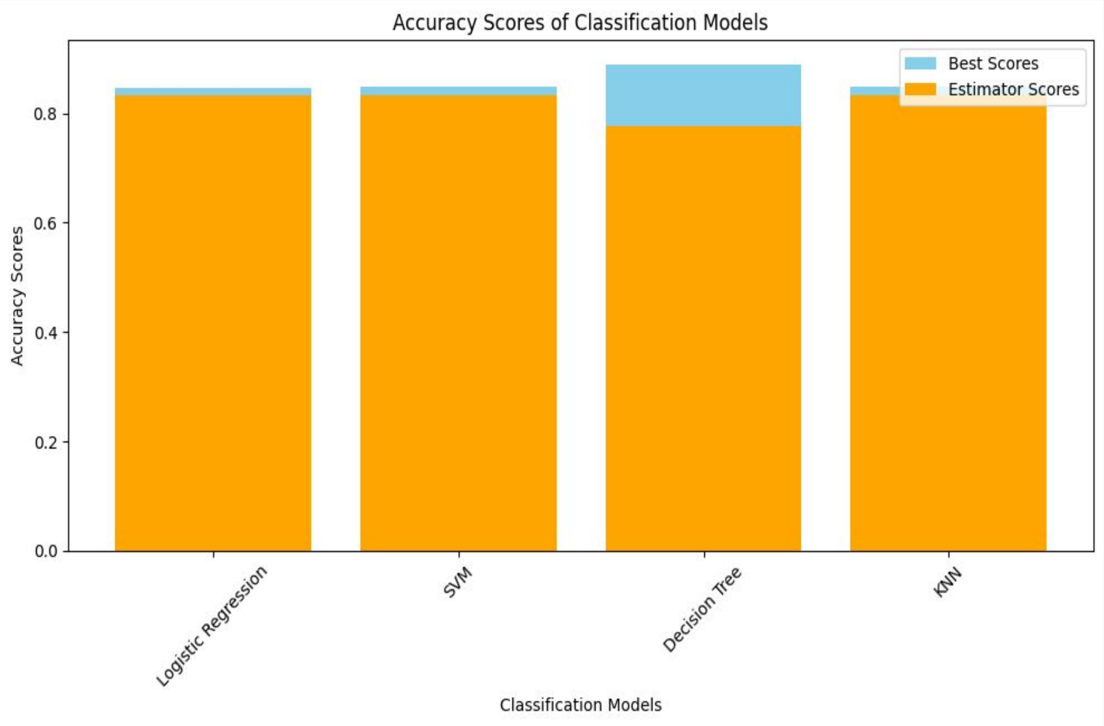
Decision Tree:
- Cross-validation accuracy: 0.8893
- Test data accuracy: 0.7778

KNN:
- Cross-validation accuracy: 0.8482
- Test data accuracy: 0.8333

Based on the provided accuracy scores, the Decision Tree model achieved the highest best score of 0.889 and the Logistic Regression, SVM, and KNN models achieved best scores of around 0.846. However, when considering the accuracy scores on the test data (estimator scores), Logistic Regression, SVM, and KNN models all achieved the same accuracy of 0.833, while the Decision Tree model had a lower accuracy of 0.778.

High Accuracy: The SVM model with the sigmoid kernel achieved one of the **highest accuracies** during cross-validation, indicating its effectiveness in predicting the success of Falcon 9 first stage landings.

Choosing the SVM model with the best kernel (sigmoid) based on the best score is a reasonable decision. The sigmoid kernel performed well in terms of accuracy during cross-validation and is likely to generalize well to unseen data.



Confusion Matrix

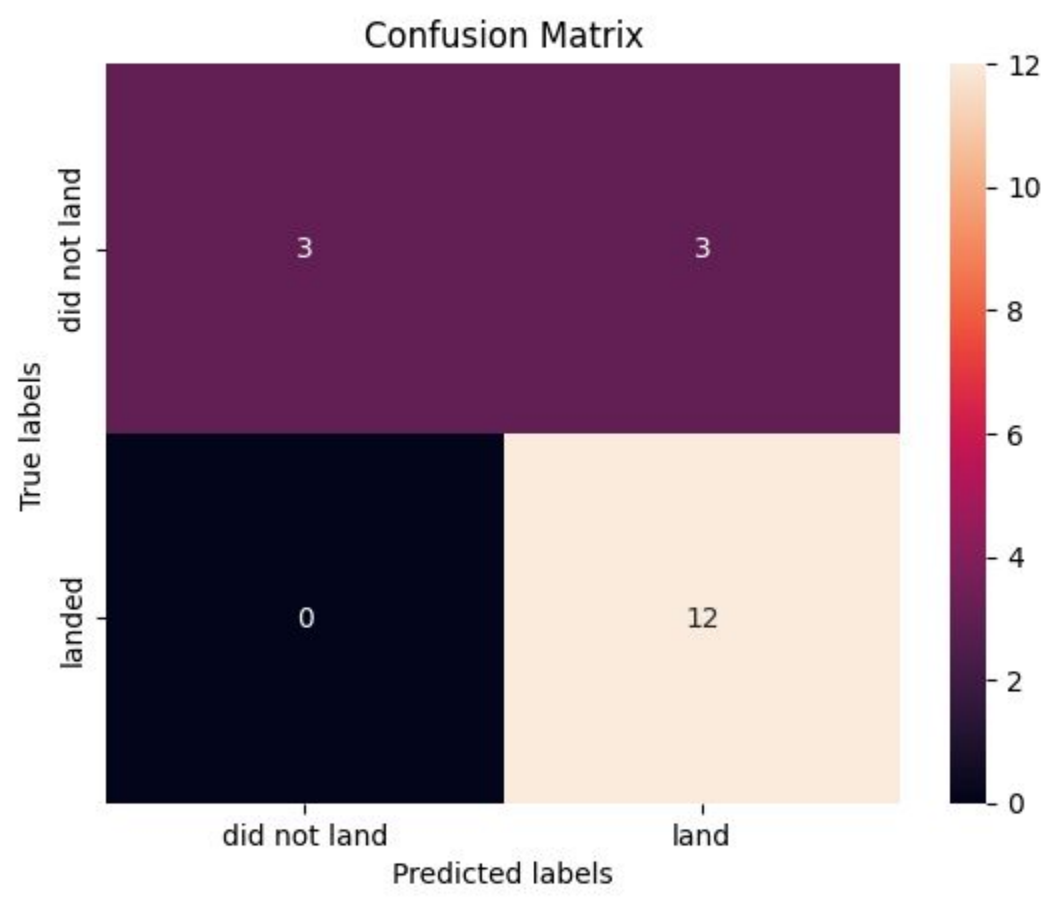
The confusion matrix for the SVM model with a kernel of sigmoid reveals the following:

- True Positives (TP): 3
- False Positives (FP): 0
- False Negatives (FN): 0
- True Negatives (TN): 12

Explanation:

- True Positives (TP): The model correctly predicted 3 instances where the first stage landing was successful.
- False Positives (FP): There were no instances where the model incorrectly predicted a successful landing when it was actually unsuccessful.
- False Negatives (FN): The model incorrectly predicted 0 instances as unsuccessful when they were actually successful.
- True Negatives (TN): The model correctly predicted 12 instances where the first stage landing was unsuccessful.

Overall, the SVM model with a sigmoid kernel demonstrates strong predictive performance, as indicated by the high number of true positives and true negatives in the confusion matrix. Additionally, the absence of false positives and false negatives suggests that the model effectively distinguishes between successful and unsuccessful first stage landings.



Conclusions

1. Data Collection: Obtained SpaceX launch data containing information about launch sites, payload mass, booster version, and launch outcomes.
2. Data Wrangling: Cleaned the dataset by handling missing values, converting data types, and filtering out irrelevant columns.
3. Exploratory Data Analysis (EDA) with SQL: Utilized SQL queries to explore patterns and insights in the SpaceX launch data, such as success rates across different launch sites and booster versions.
4. Exploratory Data Analysis (EDA) with Visualization:
   a. Visualized the distribution of payload mass and launch success outcomes using histograms and pie charts.
   b. Investigated the relationship between payload mass and launch outcomes using scatter plots.
   c. Explored the success rates across different launch sites using pie charts and bar charts.
5. Visual Analytics with Folium: Created an interactive map using Folium to visualize SpaceX launch sites and their locations on Earth.
6. Visual Analytics with Plotly Dash:
   a. Built a dashboard application using Plotly Dash to allow users to interactively explore SpaceX launch records.
   b. Implemented dropdowns, sliders, and charts to visualize launch success counts, payload correlations, and booster versions.
7. Machine Learning Prediction:
   a. Trained several machine learning models, including Logistic Regression, Support Vector Machine (SVM), Decision Trees, and K-Nearest Neighbors (KNN), to predict the success of SpaceX first-stage landings.
   b. Tuned hyperparameters using cross-validation to improve model performance.
   c. Evaluated model accuracy and selected the best performing model based on accuracy scores.
8. Overall Conclusions:
   a. The dataset provides valuable insights into the factors influencing the success of SpaceX launches, such as launch site locations, payload mass, and booster versions.
   b. The interactive visualizations and dashboard applications enhance the accessibility and usability of the data for stakeholders and analysts.
   c. Machine learning models can effectively predict the success of SpaceX first-stage landings, with Decision Trees achieving the highest accuracy among the models tested.
   d. The success of SpaceX launches is influenced by various factors, and further analysis could delve deeper into understanding these factors and optimizing launch strategies.

Appendix

Python Code Snippets:

- Data collection script - jupyter-labs-spacex-data-collection-api.ipynb
- Data wrangling and preprocessing code - labs-jupyter-spacex-Data wrangling.ipynb
- Exploratory data analysis (EDA) code using SQL queries and Python libraries - jupyter-labs-eda-sql-coursera_sqllite.ipynb
- Visualization code using Matplotlib, Seaborn and Plotly - jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb
- Visualization code using Folium Maps - lab_jupyter_launch_site_location.jupyterlite.ipynb - MAP Preview
- Visualization code using Plotly Dash - spacex_dash_app.py
- Machine learning model building and evaluation code - SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

SQL Queries:

- Queries used for data extraction and analysis from databases - SS/SQL

Charts:

- Visualizations generated during EDA, including histograms, scatter plots, bar charts, pie charts - SS/EDA
- Visualizations generated during EDA using Folium Maps - SS/MAP

Data Sets:

- Raw data sets used for analysis and modeling, including CSV files, Excel spreadsheets, or database exports - DATASET

Model Evaluation Metrics:

- Evaluation metrics for machine learning models - SS/MATRICS

Dashboard:

- Visualizations interactive dashboards , showcasing interactive visual analytics using Plotly Dash - SS/DASH

Thank you!