IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

ALOR FAVOUR AMARACHI
03/10/2024

Outline

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

Executive Summary

- **Objective**: Analyze SpaceX's Falcon 9 launch data to evaluate success factors, predict first-stage landing outcomes, and identify key performance trends.
- **Methodologies:**

  Data collection cleaning and processing

  Exploratory Data Analysis (EDA) using visualization and SQL

  Interactive dashboards with Folium and Plotly Dash

  Predictive analysis using logistic regression, SVM, decision trees, and KNN
- **Results:** Identified significant predictors of launch success

  Achieved over 94% accuracy with the best classification model

  Insights to enhance future mission planning and safety

Introduction

Overview of SpaceX: The commercial space race has advanced with companies like SpaceX, which reduces launch costs through the reuse of rocket components, particularly the Falcon 9's first stage recovery. Founded in 2002 by Elon Musk, SpaceX aims to revolutionize space travel by significantly lowering transportation costs and enabling future Mars colonization.

Problems to Find Answers For:

Which factors influence the landing

What are the best conditions for the best result

What patterns in SpaceX's launch data can provide insights into optimizing future launch operations

Can we predict if the Falcon 9's first stage will successfully land after launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling Data was collected using the SpaceX API and web scraping to gather information about past Falcon 9 launches. JSON data was transformed into a Pandas DataFrame for analysis
- Cleaned data by handling missing values, filtering columns, and standardizing formats for analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- using regression and tree techniques

Data Collection

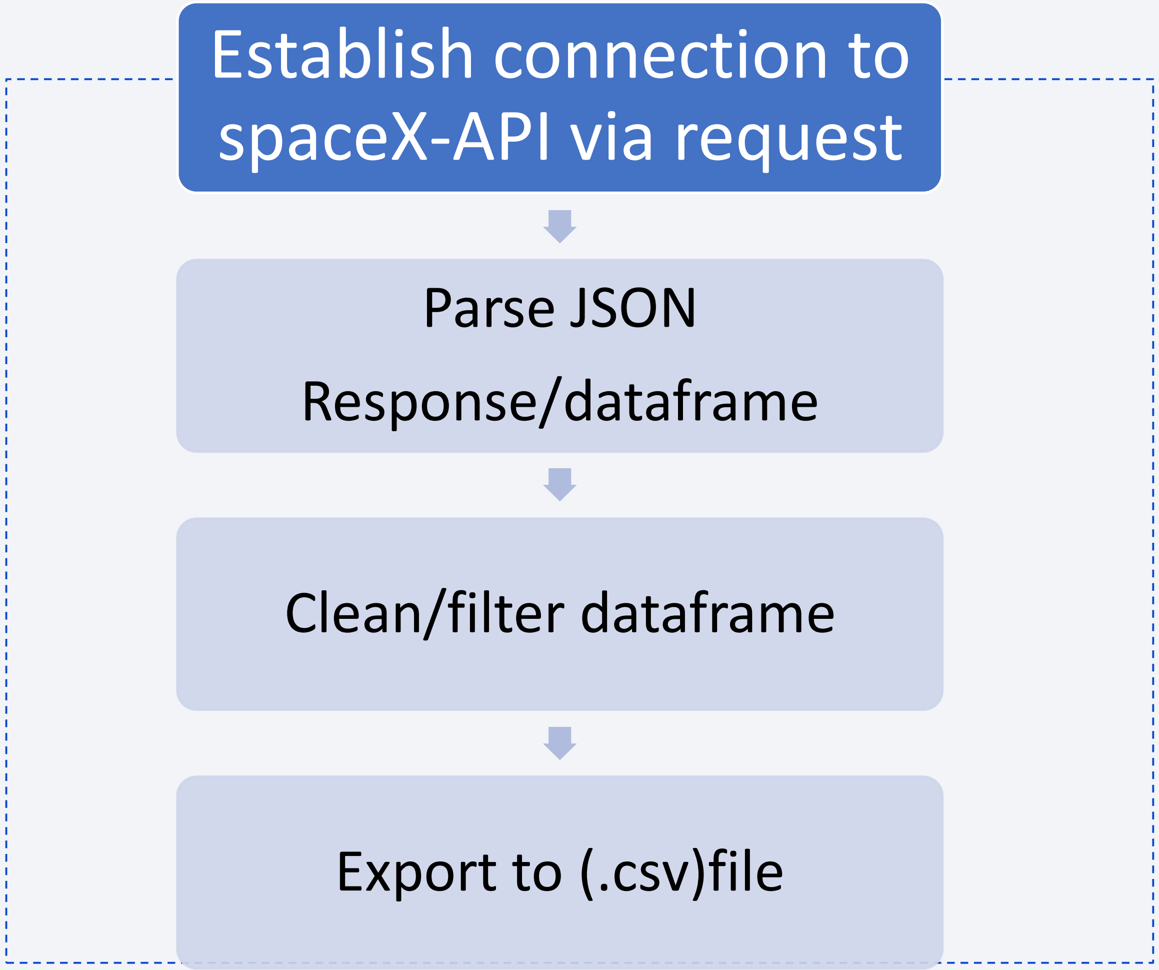
- Describe how data sets were collected.
- Data was sourced from official SpaceX launch records and public databases via API access and web scraping. This included details on launch dates, payloads, sites, and outcomes, which were subsequently cleaned and exported in CSV format.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- **Data Source:**
- SpaceX Launch Data from SpaceX REST API
- **Collection Process:**

1. **API Request**: Fetched launch data from the API.
2. **Data Conversion**: Transformed JSON responses into Pandas DataFrames.
3. **Data Preparation**: Cleaned and structured the data for analysis.
4. **Step 4:** Exported the cleaned data to a CSV file for further analysis.

https://github.com/AmasAura/My-data-science-class/blob/main/Data_Collection_API.ipynb

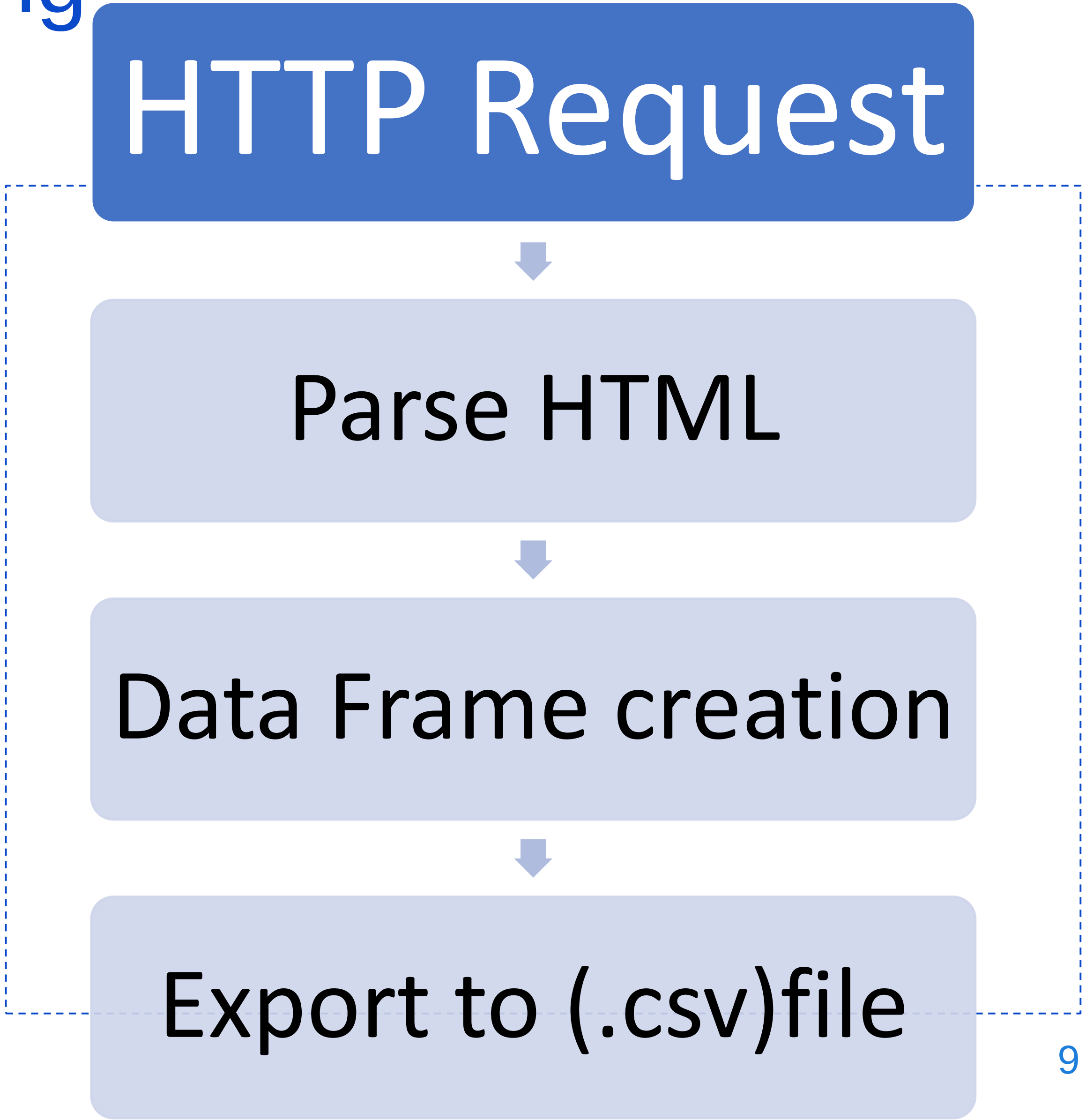


Data Collection - Scraping

- **Data Source:**
  Scraped SpaceX historical data from the official website using BeautifulSoup and requests.
- **Scraping Process:**

1. **Step 1:** Performed HTTP requests to access web pages.
2. **Step 2:** Parsed the HTML content with BeautifulSoup.
3. **Step 3:** Extracted key information such as mission names, launch dates, and rocket details into a pandas DataFrame.
4. **Step:** Exported the cleaned data to a CSV file for further analysis.

https://github.com/AmasAura/My-data-science-class/blob/main/Data_Collection_WebScraping.ipynb



Data Wrangling

Data process overview:

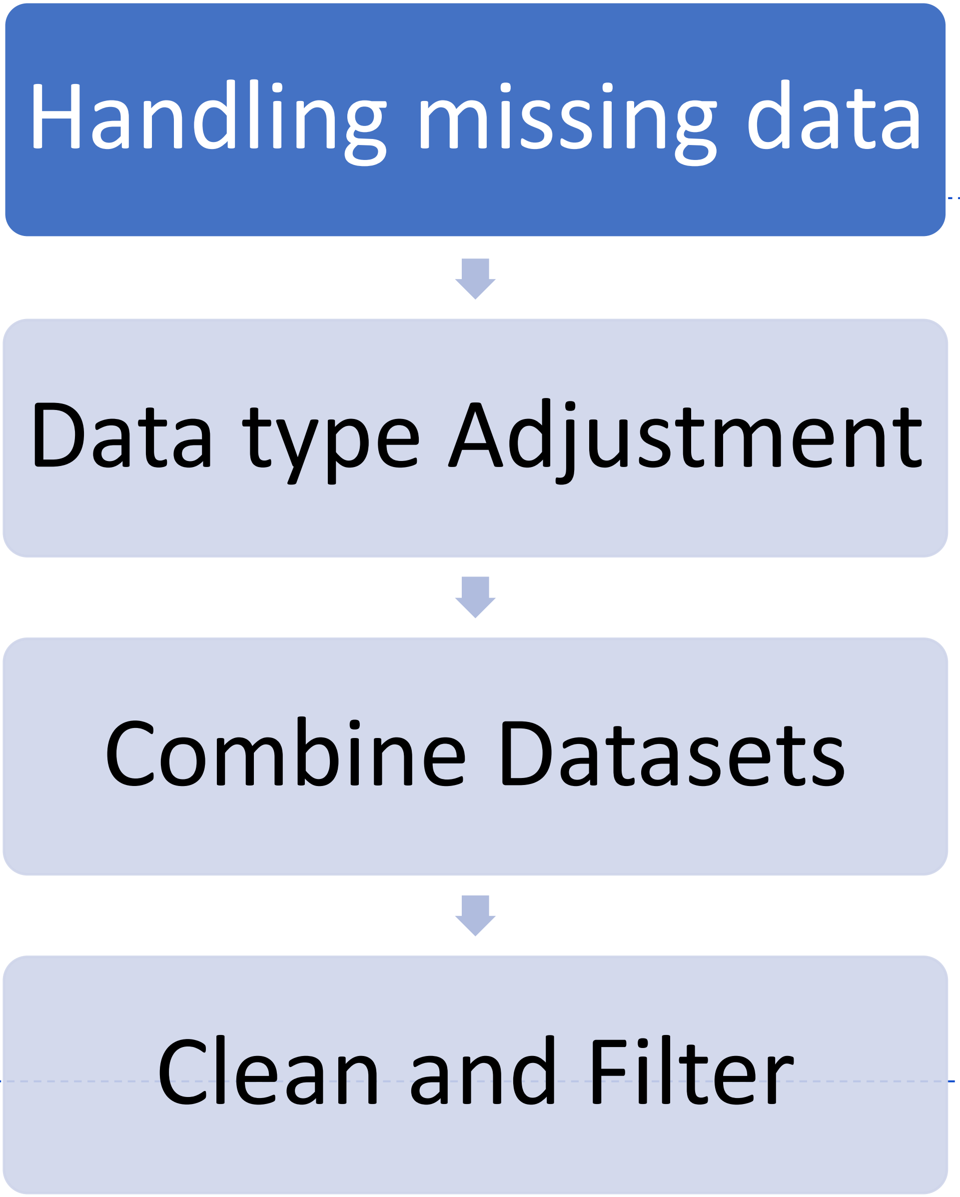
Step 1: Detected and addressed missing data using pandas method Filled or removed missing values based on relevance and impact on analysis

Step 2: Converted data types for numerical and datetime fields to ensure proper analysis.

Step3: Amalgamated several datasets to establish a comprehensive dataset for analysis.

Step 4: Eliminated inconsistencies and selected key data points for analysis.

https://github.com/AmasAura/My-data-science-class/blob/main/Data_Wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- **Charts Created:**
- **Bar Charts:** Visualized rocket launch success rates by year.
- **Pie Charts:** Represented the distribution of launch outcomes (success vs. failure).
- **Scatter Plots:** Analyzed the correlation between payload mass and launch success.
- **Line Plots:** Illustrated trends in launch frequency over time.
- **Chart Purposes:**
- **Bar & Pie Charts:** Emphasized categorical data for success analysis.
- **Scatter & Line Plots:** Explored trends and relationships over time.
- https://github.com/AmasAura/My-data-science-class/blob/main/EDA_Visualizations.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

➢ General overview over available landsides, in particular whose five entries who start with 'CCA'

➢ Number of successful/failed mission outcome

➢ List for failed ones in 2015

➢ List first successful landing outcome in drone ship

Specify, count and rank outcomes between ~2010 and ~2017

➢ Average Payload per booster version 'F9v1.1'/Total for boosters carried by NASA (CRS)

➢ Booster version with maximal payload

➢ Names of boosters with successful ground pad and certain payload

https://github.com/AmasAura/My-data-science-class/blob/main/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

➢ I added map icons, circle markers and lines

➢ Markers: Indicated launch sites for better visibility of locations.

➢ Circles: Represented launch zones to visualize operational areas.

➢ Polylines: Illustrated rocket trajectories to demonstrate flight paths.

- Explain why you added those objects

➢ This was done to draw insight as to weather the landing was a success or a failure

- https://github.com/AmasAura/My-data-science-class/blob/main/Folium_Visualization.ipynb

Build a Dashboard with Plotly Dash

Map Objects Summary:
I added map icons, circle markers, and lines to the folium map. These elements were included to provide insights into whether the landings were successful or failed.
Dashboard Components:
•**Pie Chart:** Displays the percentage of launches by site.
•**Scatter Plot:** Visualizes the relationship between payload and launch outcomes.

Rationale for Components:
The Pie Chart offers quick insights into the distribution of launches across sites, while the Scatter Plot analyzes the correlation between payload sizes and launch success rates.

https://github.com/AmasAura/My-data-science-class/blob/main/dashboard.py

Predictive Analysis (Classification)

Building the Classification Model

1. Data Preparation :Clean the dataset, handle missing values, and encode categorical variables Split the dataset into training and testing sets
2. Feature Selection• Identify key features that will be used for the classification task• Use techniques like correlation analysis or feature importance rankings
3. Model Selection• Select classification algorithms such as Logistic Regression, Random Forest, or Decision Trees base Evaluation of classification Model
4. Training the selected models using the training dataset. Apply techniques like cross-validation to ensure generalization
5. Evaluation Metrics• Use evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC to assess model Confusion matrix: Visualize correct vs. Incorrect classifications Model Comparison Compare different classification models based on the above metrics Use cross vallidation to measure how well the model performs on unseen data.

https://github.com/AmasAura/My-data-science-class/blob/main/Machine_Learning_Prediction_Analysis.ipynb

Results

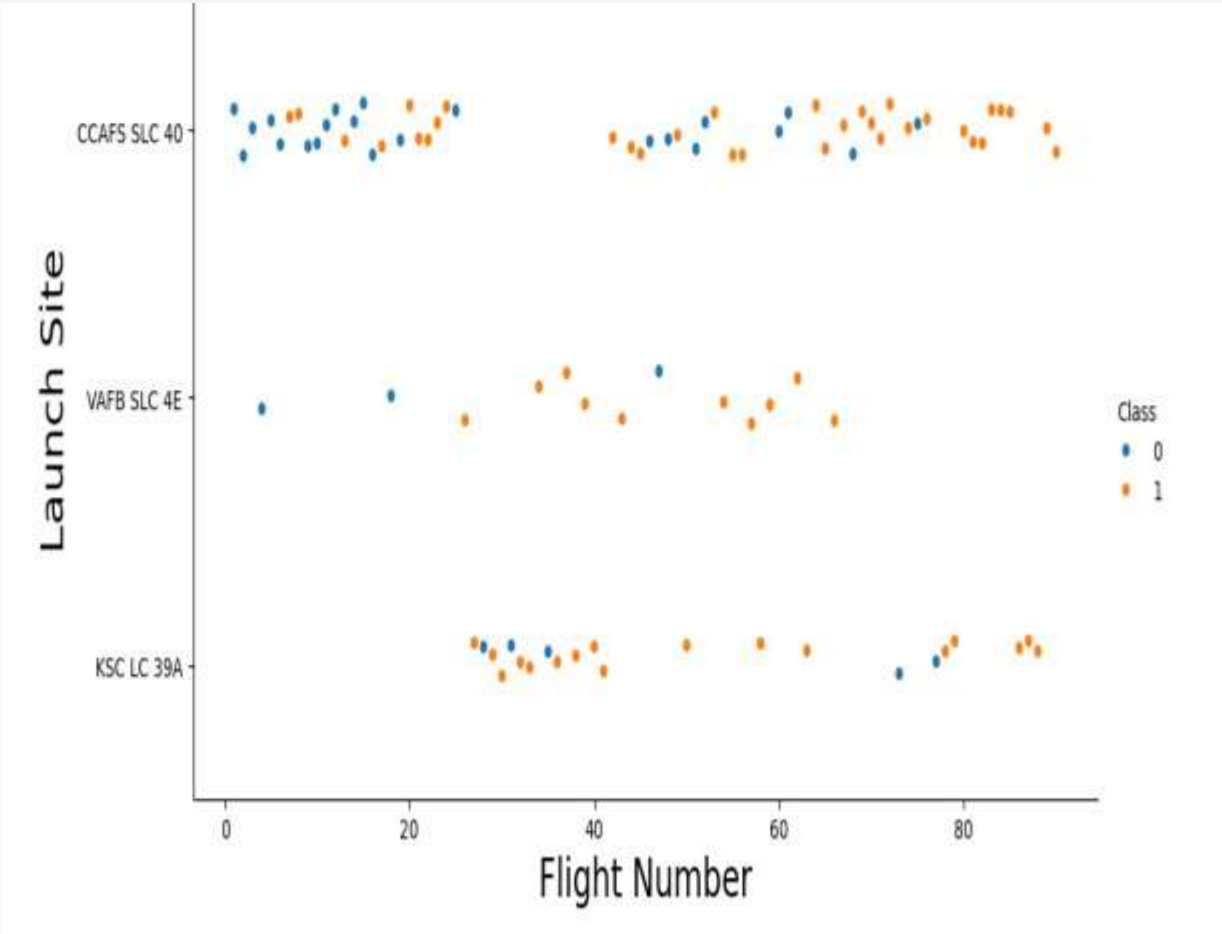
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

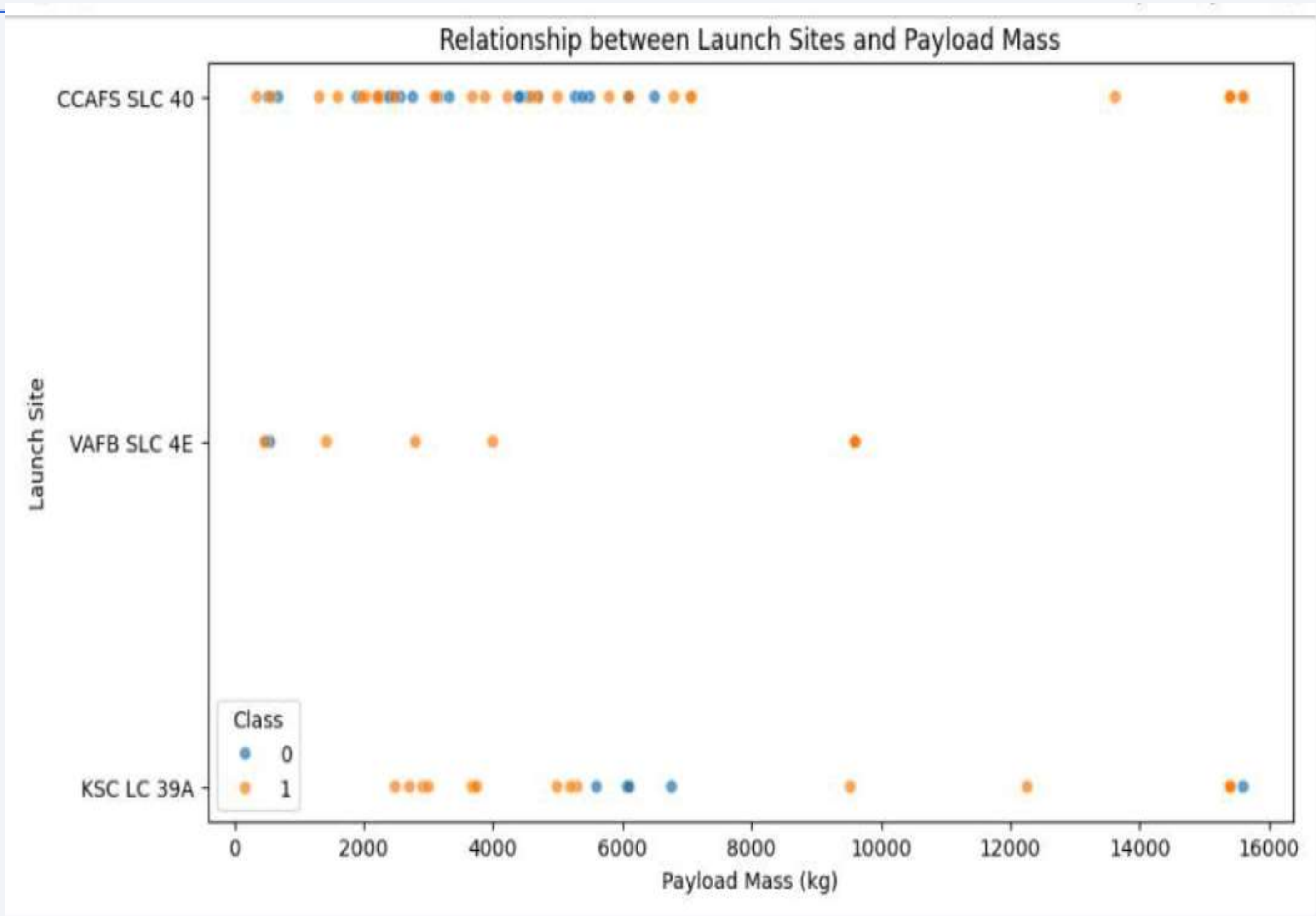
Flight Number vs. Launch Site

- **Flight Number (X-Axis):**
- Represents the sequential order of SpaceX launches, with each flight assigned a unique, incrementing number. This provides insight into the launch timeline and history.
- **Launch Site (Y-Axis):**
- Plots various SpaceX launch sites, including:
  - **CCAFS LC-40:** Cape Canaveral Air Force Station, Launch Complex 40
  - **KSC LC-39A:** Kennedy Space Center, Launch Complex 39A
  - **VAFB SLC-4E:** Vandenberg Air Force Base



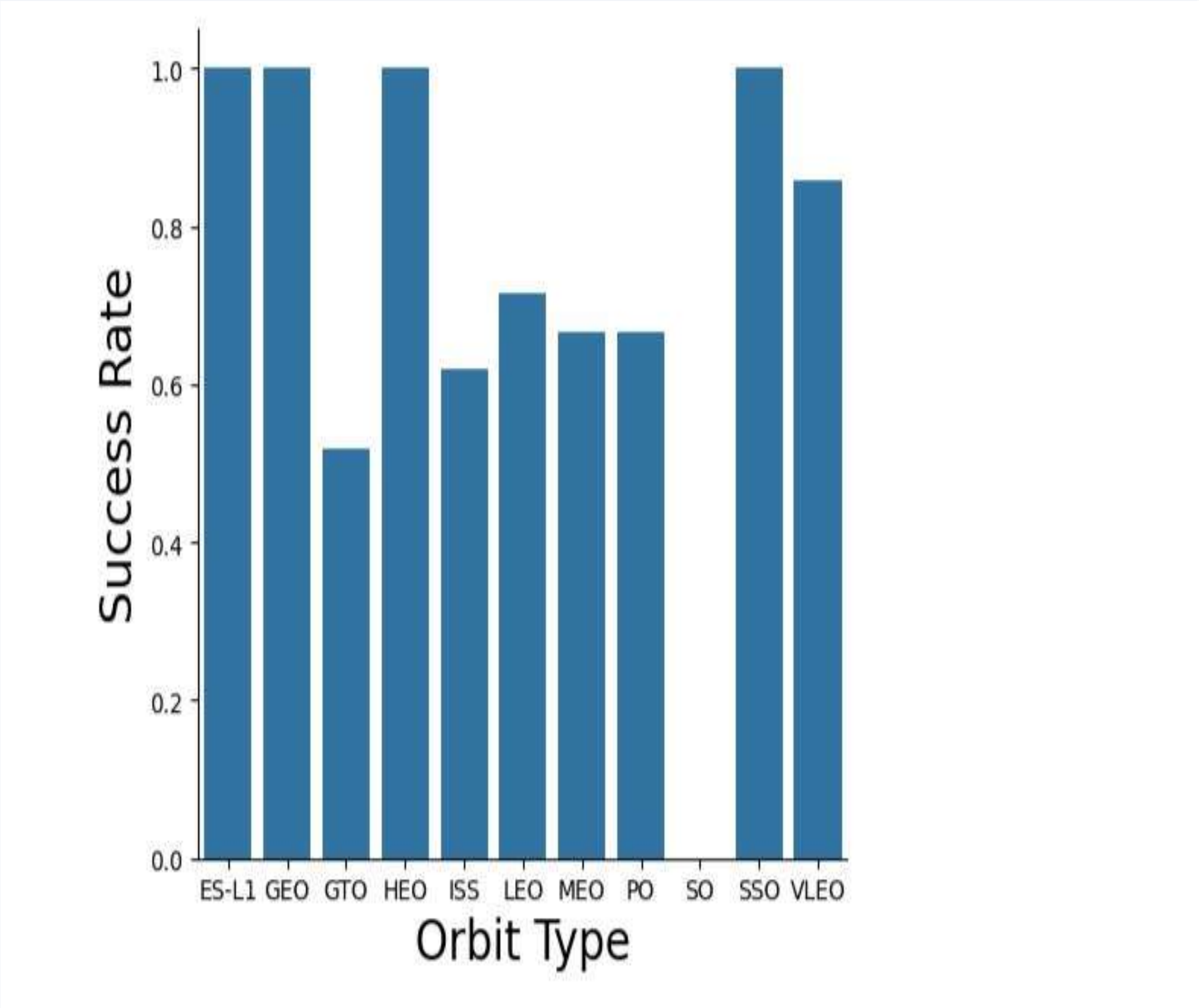
Payload vs. Launch Site

- By examining the scatter plot, you can identify which sites manage larger payloads. For instance, if heavier payloads cluster around **KSC LC-39A**, it indicates those locations are preferred for more powerful missions, like Falcon Heavy.
- **Variation in Payloads:** The spread along the x-axis shows the range of payloads launched from each site. Some locations accommodate diverse payloads, while others focus on smaller missions.
- **Trends and Outliers:** Outliers with unusually heavy or light payloads may represent significant missions or experimental launches.



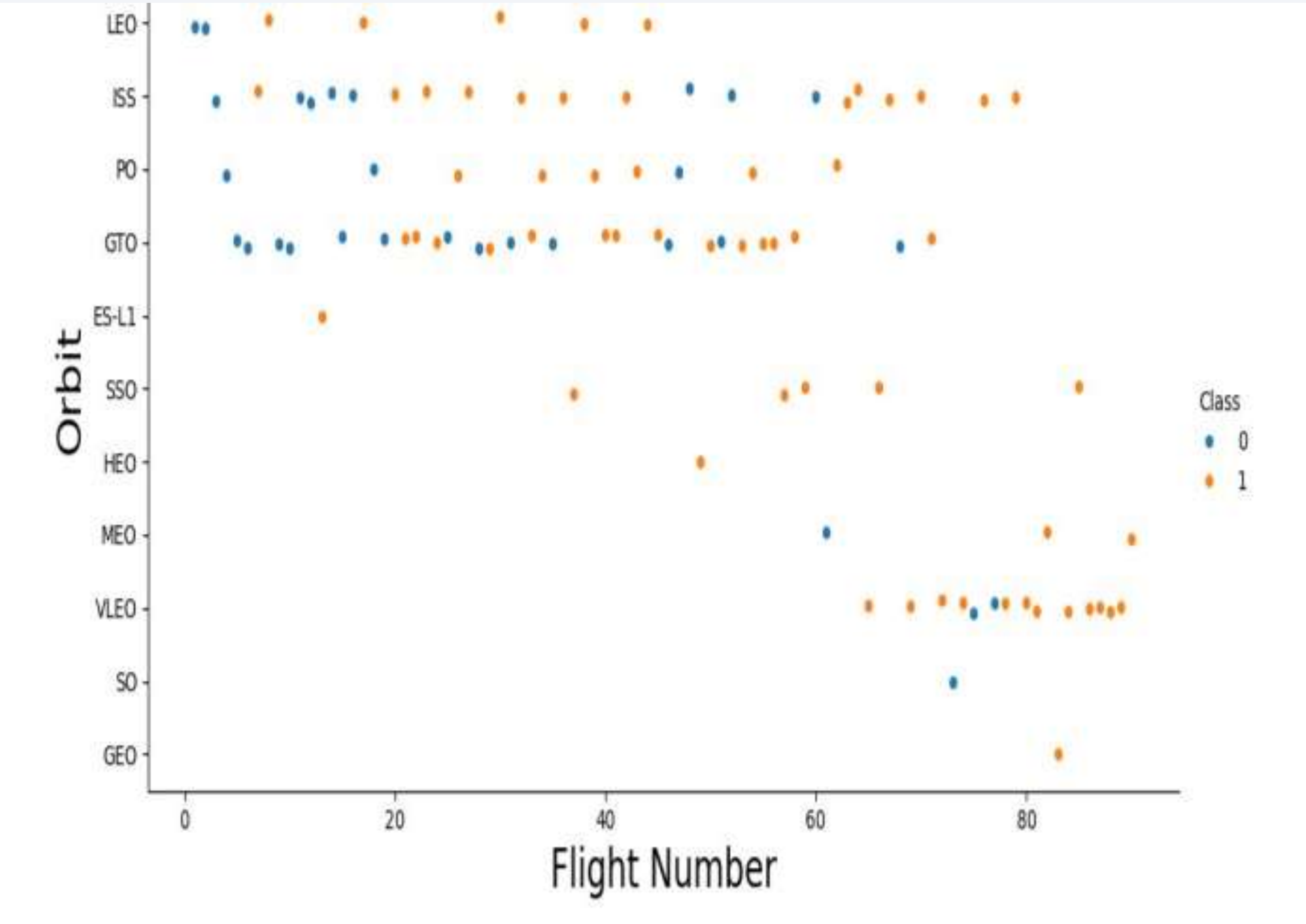
Success Rate vs. Orbit Type

•**Comparing Success Rates:** The chart reveals which orbits have higher or lower success rates. For example, a higher success rate in **LEO** compared to **GTO** indicates greater success in missions to lower orbits.
•**Operational Complexity:** Orbits like **GTO** require more energy, potentially affecting success rates.
•**Trends in Launch Outcomes:** Patterns in launch outcomes related to target orbits provide insights into SpaceX's performance across different mission profiles.



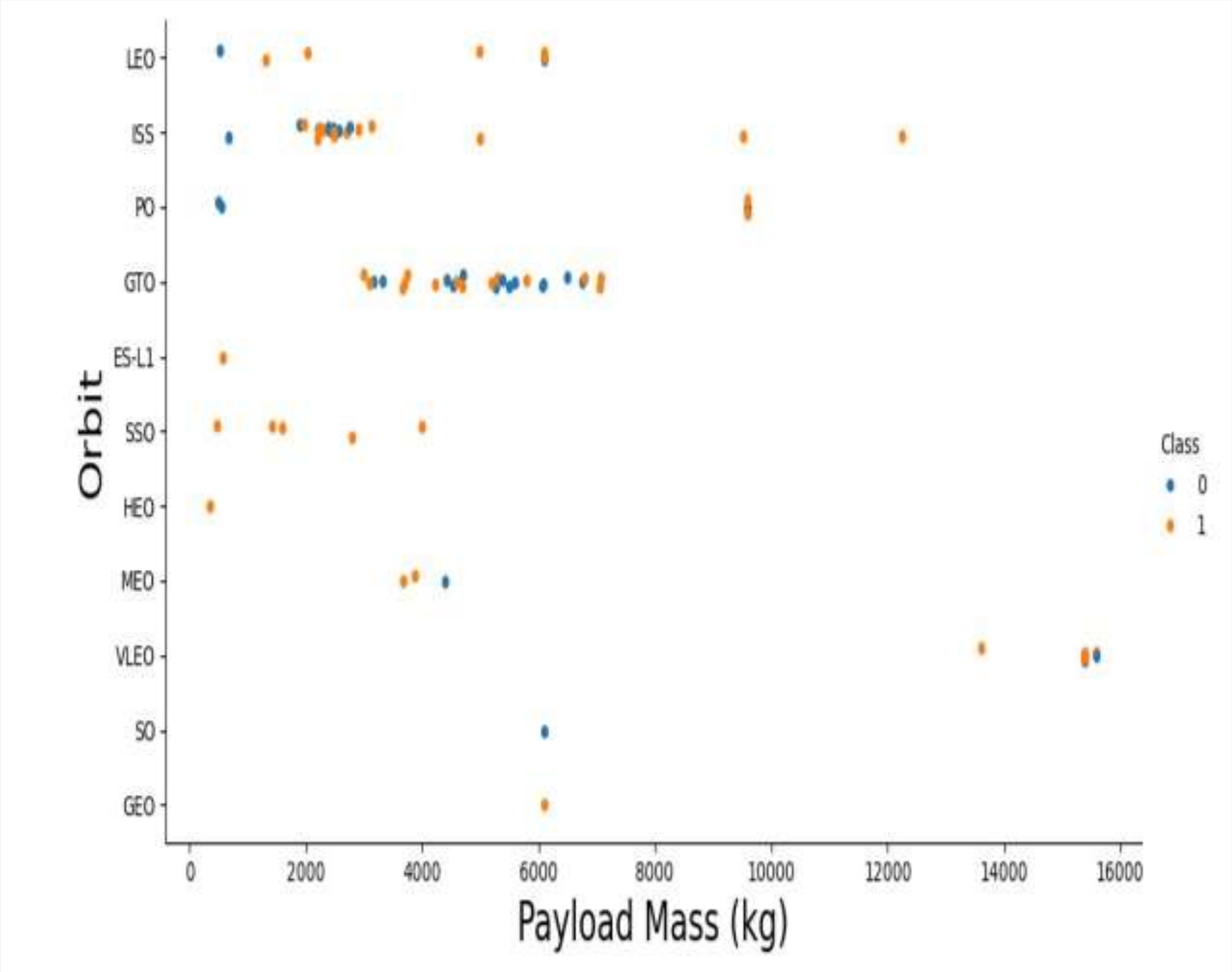
Flight Number vs. Orbit Type

- The scatter plot displays the distribution of SpaceX flight missions across different orbit types, with flight numbers on the x-axis and orbit types on the y-axis. The data points are color-coded by class:
- **Blue Dots (Class 0)**: Represent one category of missions, while
- **Orange Dots (Class 1)**: Represent another category.



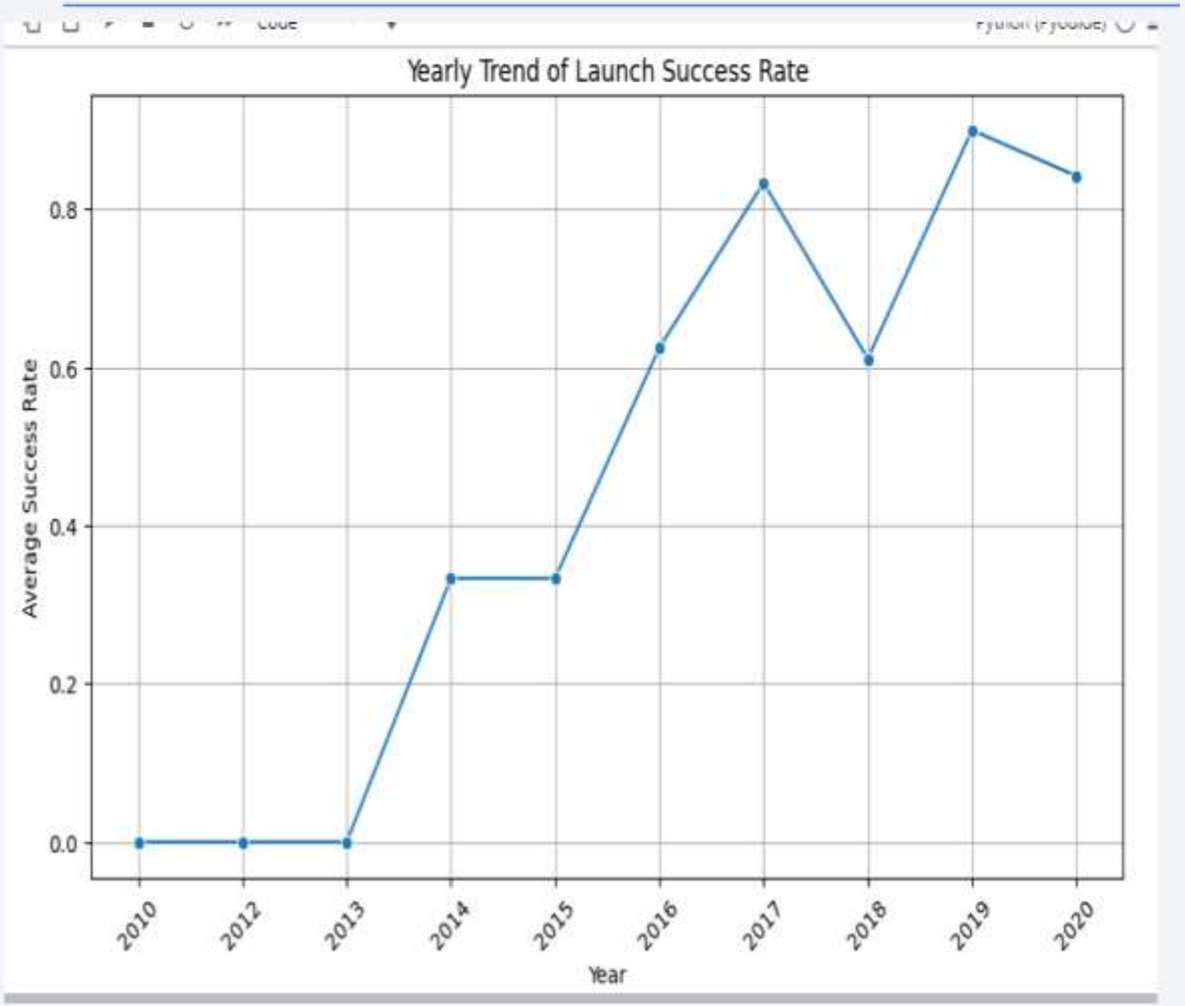
Payload vs. Orbit Type

The scatter plot illustrates that LEO is the most versatile and frequently successful orbit type for a wide range of payload masses, while GTO is also quite successful for heavier payloads.• The data suggests that missions targeting LEO have a higher success rate due to their relative simplicity, while GTO missions have successfully managed heavier, more complex payloads



Launch Success Yearly Trend

- This report summarizes the trends in SpaceX's launch success rates from its inception to the present.
- **Early Years (2002-2010)**
- SpaceX experienced a low success rate during its initial years, facing technical issues with the Falcon 1 and early Falcon 9 rockets.
- **Mid-Years (2011-2016)**
- After 2010, a clear upward trend in successful launches began, driven by the Falcon 9's deployment and reusable rocket technology, which improved reliability and reduced failures.
- **Recent Years (2017-Present)**
- From 2017 onwards, SpaceX saw a significant increase in successful launches, largely due to the Falcon 9 Block 5, NASA contracts, and Starlink missions.
- The reliability of reusable rockets has been pivotal in maintaining a high success rate



All Launch Site Names

SQL queries in Jupyter notebooks.

- **SELECT DISTINCT**: Retrieves unique values from the specified column, eliminating duplicates.
- **Launch_Site**: The column containing the names of launch locations.
- **FROM SPACEXTABLE**: The table containing data about SpaceX launches.
- **Purpose:** This query extracts a list of unique launch sites from the database, ensuring each site is listed only once, which helps identify all used launch locations without repetition.
- **Output:** A list of distinct launch sites from the SPACEXTABLE.

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

```
[14]: %sql select distinct(LAUNCH_SITE) from SPACEXTBL
```

 * sqlite:///my_data1.db
Done.

[14]:

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

Launch Site Names Begin with 'CCA'

- The query returns 5 launches from **CCAFS LC-40**, with details such as the date, time, booster version, payload, orbit, mission outcome, and landing outcome.
- All these missions were successful launches, although some had failed landings. The payloads include various missions such as Dragon demo flights and ISS resupply missions.
- The missions involved a mix of ISS-related payloads with varying outcomes, particularly showing

Task 2

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

```
[15]: %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5
```

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

[15]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload mass for NASA CRS missions is **45596kg**.
This result indicates that SpaceX has successfully delivered **45596 kilograms** of payload into space for NASA's CRS missions, which were crucial for resupplying the International Space Station (ISS). The payload includes various types of cargo, scientific equipment, and supplies for astronauts

Task 3

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

```
[16]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'

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

[16]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- this query calculates the average payload mass carried by boosters of the **F9 v1.1** version from the **SPACEXTABLE**.
- **Results:**
- **Average Payload Mass**: The average payload mass for the **F9 v1.1** booster is **2,928.4 kg**.
- This indicates that the Falcon 9 v1.1 booster typically carried about **2,928.4 kilograms** during its launches, providing insight into its carrying capacity compared to other Falcon 9 versions.

Task 4

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

```
[17]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'

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

[17]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- this query retrieves the date of the first successful landing on a ground pad using the **MIN** function.
- **Results:**
- **First Successful Landing Date**: The first successful landing on a ground pad occurred on **December 22, 2015**.
- This milestone marks the first successful landing of a Falcon 9 rocket on a ground pad, demonstrating the reusability of the booster. This achievement was crucial for SpaceX's efforts to reduce launch costs through booster recovery.

Task 5

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

Hint:Use min function

```
[24]: %sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)'
```

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

[24]:

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the booster versions that:
  - Successfully landed on a drone ship (Landing_Outcome = 'Success (drone ship)').
  - Carried a payload mass between **4,000 kg and 6,000 kg**.

Results:

- The query filters the dataset to include only successful drone ship landings with medium-sized payloads (between 4,000 kg and 6,000 kg).
- These successful landings demonstrate the advanced recovery techniques developed by SpaceX for its Falcon 9 boosters, contributing to the company's efforts in recovering and reusing rockets to lower the cost of space missions.

Task 6

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

```
%sql select distinct (Booster_Version) from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MAS
```

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

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

Total Number of Successful and Failure Mission Outcomes

- This query counts the total number of mission outcomes by grouping the data based on the **Mission Outcome** field, listing distinct outcomes and their occurrence counts.
- **Results:**
- The analysis reveals the following mission outcomes:
  - **98** missions were successful.
  - **1** in-flight failure.
  - **1** mission marked as successful but with unclear payload status, indicating potential ambiguity in the mission's objectives or payload performance.
  - An additional entry labeled as "Success" with **1** occurrence suggests a data anomaly or inconsistency that may require further investigation.

Task 7

List the total number of successful and failure mission outcomes

[27]:
```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

[27]:

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

Boosters Carried Maximum Payload

•This query identifies the booster versions that carried the maximum payload mass by:

1. Using a subquery (SELECT MAX(PAYLOAD_MASS_KG) FROM SPACEXTABLE) to find the maximum payload mass across all records.
2. The main query retrieves all **Booster_Version** entries where the payload mass matches this maximum value.

Results:

•The boosters listed are from the **F9 B5 (Block 5)** series, consistently associated with carrying the maximum payload mass in SpaceX's launches.

•The presence of various boosters within this series indicates that the Block 5 version of Falcon 9 is highly capable and frequently used for missions involving heavy payloads.

Task 8

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

[28]:
```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

[28]:

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

•his query retrieves records from the **SPACEXTABLE** with the following criteria:

- •It extracts the month from the **Date** column using the substr(Date, 6, 2) function to get the two-digit month.
- •The **WHERE** clause filters for records from the year **2015** where the **Landing Outcome** is **"Failure (drone ship)"**.

Results:

•The query returns records of failed landings on a drone ship during 2015.

•Notably, in April, the **F9 v1.1 B1015** booster failed its landing attempt at the **CCAFS LC-40** launch site.

•Both failures occurred at the **CCAFS LC-40** site, indicating a potential need for further investigation into booster performance during that period.

Task 9

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

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

```
%sql select CASE substr(Date, 6, 2) WHEN '01' THEN 'January' WHEN '02' THEN 'February' WHEN '03' THEN 'March'WHEN '04
```

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

Month_Name	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

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

- This query counts the occurrences of each unique **Landing_ Outcome** within the date range from **June 4, 2010, to March 20, 2017**.
- The results are grouped by **Landing Outcome** and ordered in descending order based on the count of each outcome.
- **Results:**
- The most common landing outcome during this period was **"No attempt,"** with **10 occurrences**, indicating many missions where landing attempts were not made.
- **Successful landings** on a drone ship and **failures** on a drone ship both had **5 occurrences**, suggesting a balanced success rate for drone ship landings.
- Additionally, there were **3 successful landings** on ground pads and **3 controlled landings** in the ocean, reflecting a variety of landing strategies employed.
- Overall, this data highlights the operational challenges and successes of SpaceX's missions during this timeframe, providing insight into the effectiveness of their landing strategies.

```
%sql select Landing_Outcome, COUNT(*) AS Outcome_Count from SPACEXTABLE where Date BETWEEN '2010-06-04' AND '2017-03-2
```

```
* 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

SpaceX Launch Sites' Location Markers on Global Map

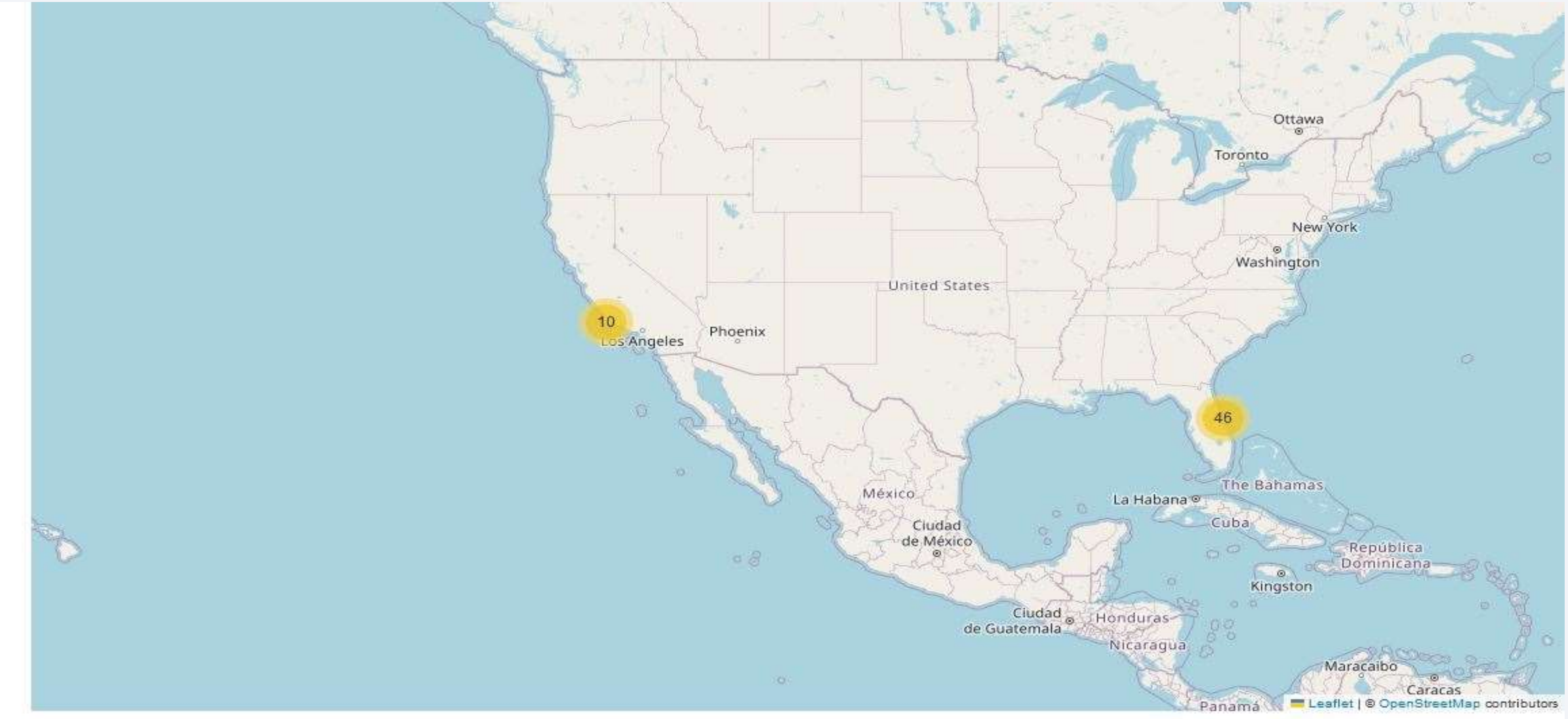
[6]:

	Launch Site
0	CCAFS LC-40
1	CCAFS SLC-40
2	KSC LC-39A
3	VAFB SLC-4E

Launch site are located on the east and west coast of the US, precisely in Califonia and Florida

The other significant cluster is located in Florida, specifically at Cape Space Canaveral Force Station (CCAFS) and Kennedy Space Center (LC-39A).These regions are major hubs for SpaceX launches, with most launches occurring from the Florida site

Launch Outcomes Depiction on Global Map



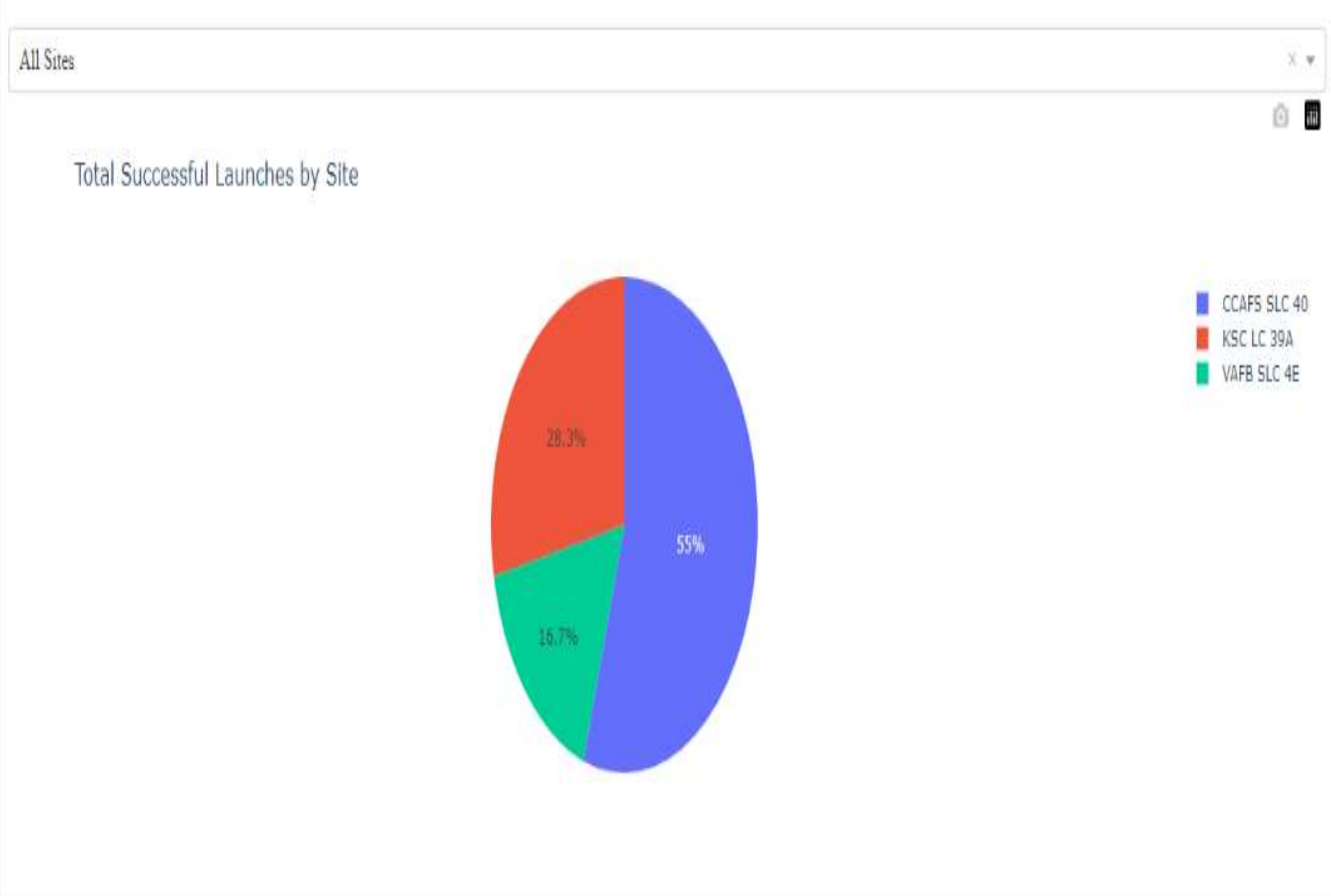
Proximity Analysis of Selected Launch Site on Global Map

Section 4

Build a Dashboard with Plotly Dash

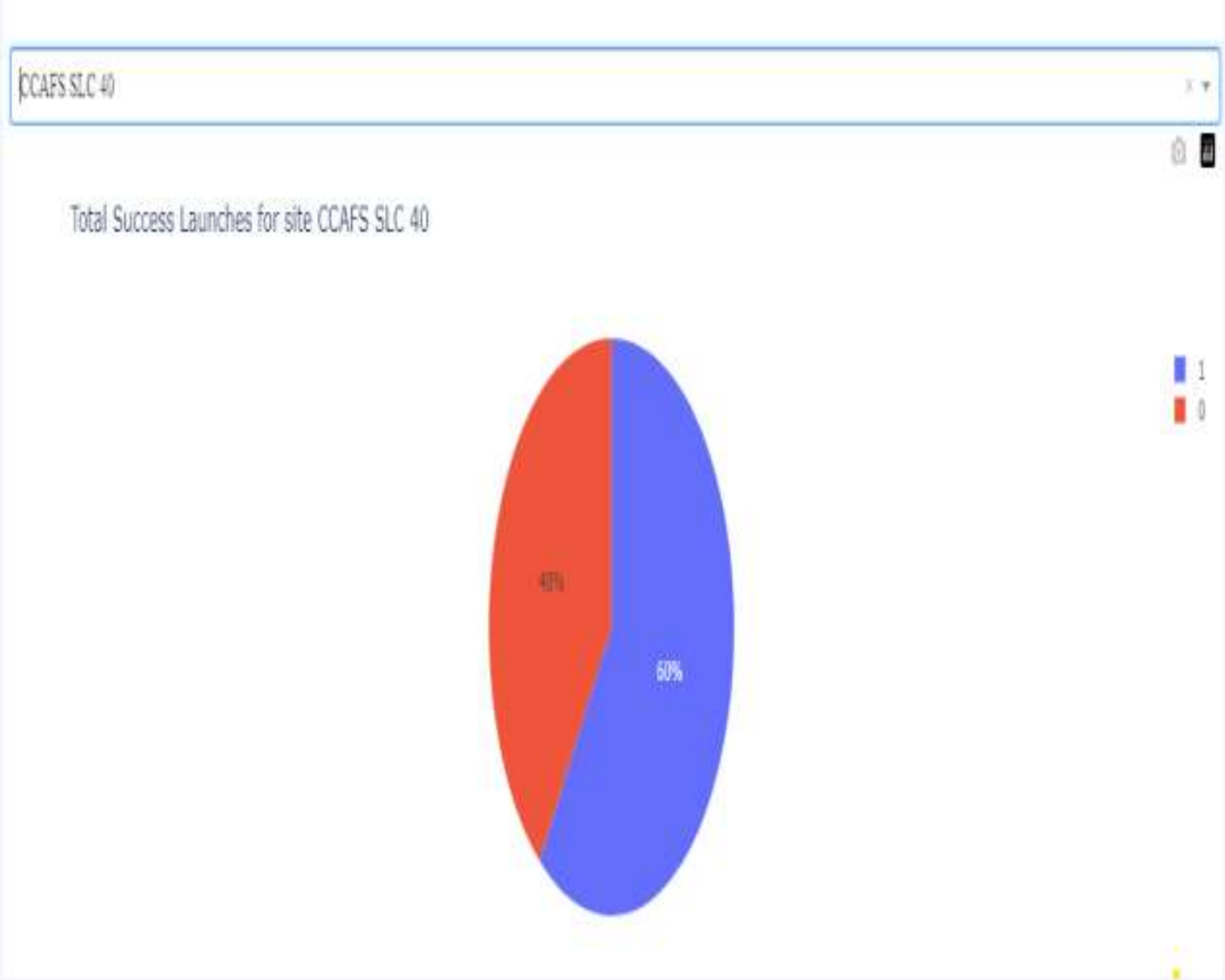
Total Successful Launches by Site

•**Key Elements:**
•**Colors:** Different colors for each launch site.
•**Proportional Size:** Segment size correlates with the number of successful launches.
•**Labels:** Each segment shows the launch site and percentage of successful launches.
•**Key Findings:**
•**Dominant Sites: KSC LC-39A** and **CCAFS LC-40** have larger segments, indicating higher success rates.
•**Smaller Sites:** Sites with smaller segments contributed fewer successful launches.



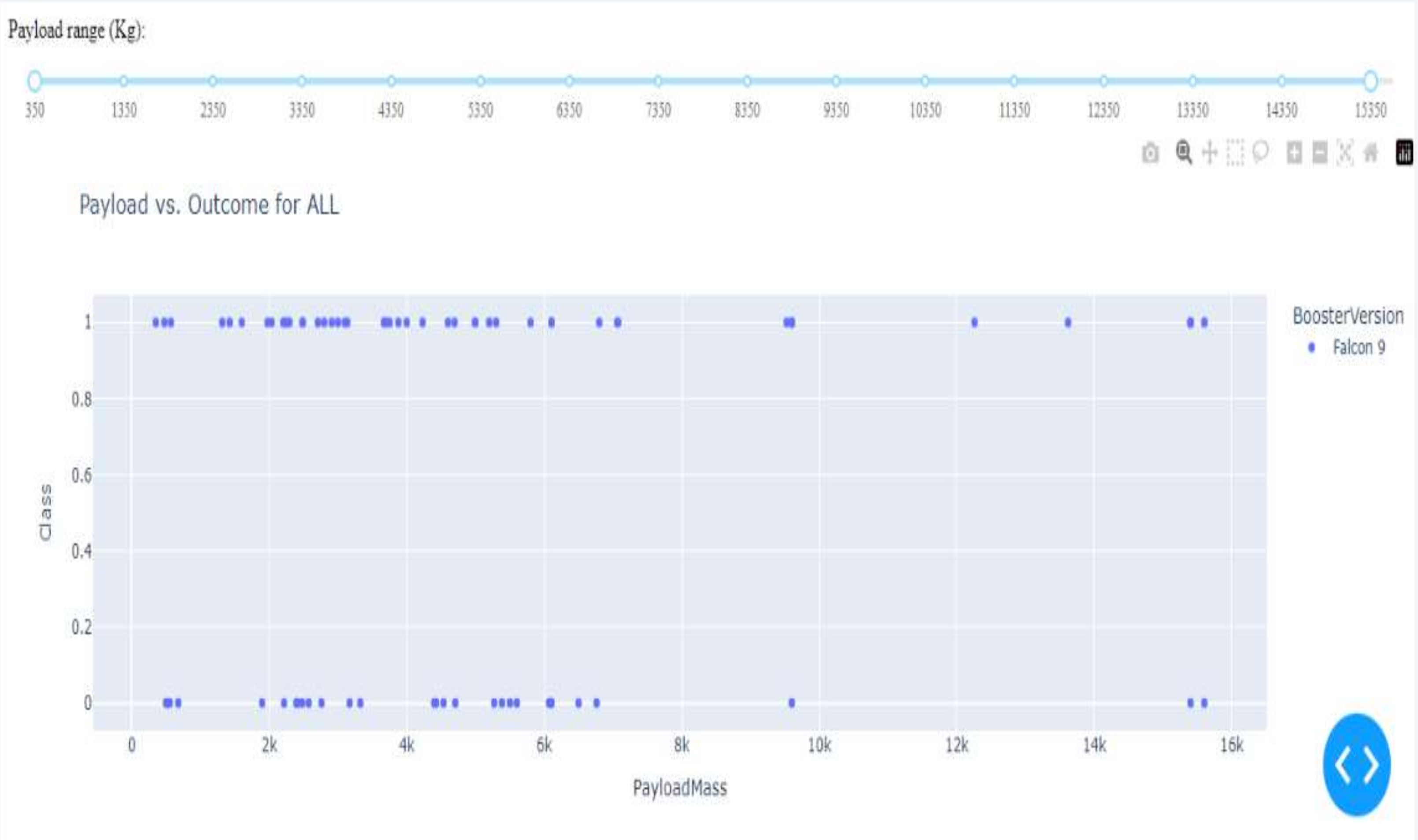
Total Success Launches for Site CCAFS SLC 40

- **Categories:**
  - **Blue Segment (76.9%):** Representssuccessful launches.
  - **Red Segment (23.1%):** Represents failed launches.
- **Labeling:** Success rate (76.9%) and failure rate (23.1%) are clearly labeled.
- **Color Legend:** Blue indicates success; red indicates failure.
- **Key Findings:**
- **High Success Rate:** The KSC LC-39A site has a success rate of 76.9%, highlighting its reliability.
- **Notable Failure Rate:** A failure rate of 23.1% indicates areas for improvement.



PayloadMass VS Outome for All Sites

- There is a noticeable trend indicating that lighter payloads tend to achieve higher success rates. This suggests that as payload mass decreases, the likelihood of mission success increases, potentially due to reduced stress on the launch vehicle

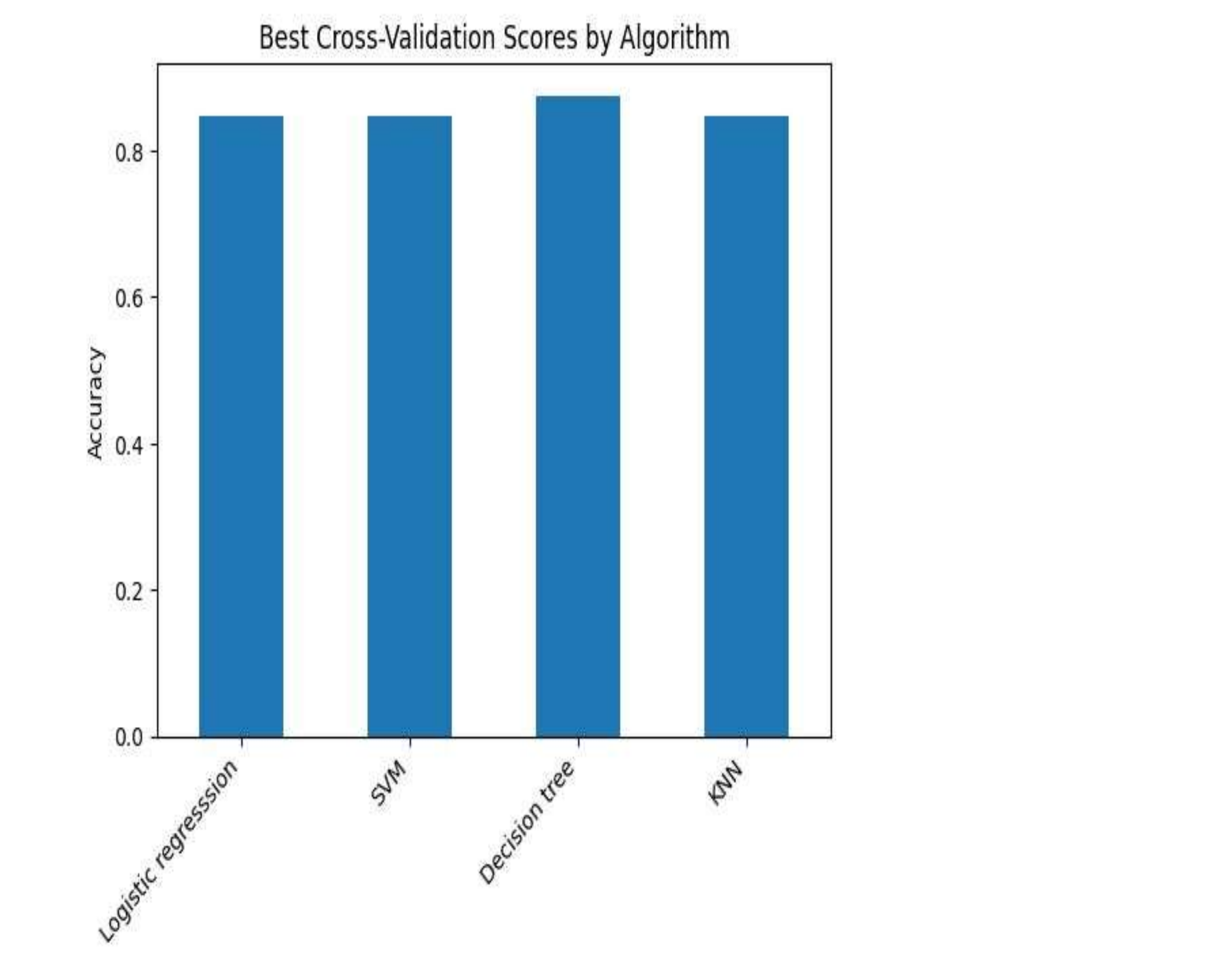


Section 5

Predictive Analysis (Classification)

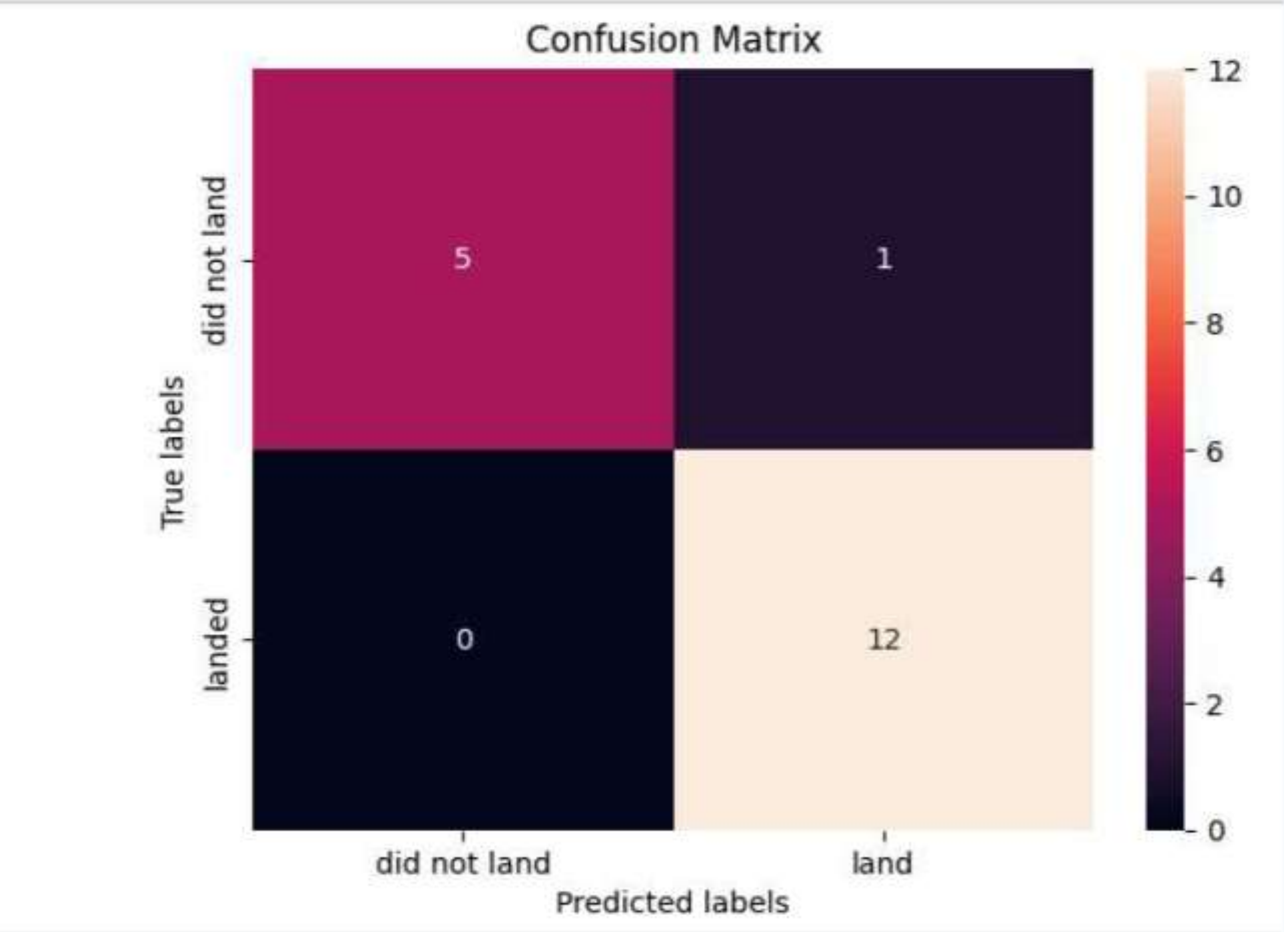
Classification Accuracy

•The Decision Tree model exhibits a high test set accuracy of 94.44%, demonstrating its ability to generalize effectively to unseen data. This level of accuracy indicates that the model can reliably predict landing outcomes based on the features used, making it a robust choice for this classification task.



Confusion Matrix

- **True Positives (TP)**: 12 successful landings accurately predicted, reflecting high precision and recall for success classifications.
- **True Negatives (TN)**: 5 unsuccessful landings correctly classified, indicating effective identification of failures.
- **False Positives (FP)**: Only 1 instance of incorrectly predicting success, showcasing a low false positive rate.
- **False Negatives (FN)**: 0 instances of missed successful landings, underscoring the model's reliability.
- With an overall accuracy of 94.44% and minimal errors (1 FP, 0 FN), the Decision Tree model effectively distinguishes between successful and unsuccessful landings. Its reliability makes it highly suitable for predicting future landing outcomes.



Conclusions

- This analysis concludes that the Decision Tree model is highly reliable, exhibiting minimal classification errors and strong predictive power for landing outcomes. Its ability to accurately identify both successful and unsuccessful landings makes it a valuable tool for future space missions, providing high-confidence predictions for mission planning and execution.
- SpaceX primarily uses a few key launch sites, which are optimized for frequent and successful launches.
- Across all launch sites, the higher the payload mass (kg), the higher the success rate
- Analysis shows a significant increase in the number of launches over time, indicating SpaceX's operational scalability and growth.

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!