IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Shyam Sundar Vijayakumar
Date: 26.08.2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with folium
  - Building a dashboard with Plotly Dash
  - Predictive analysis
- Summary of all results
  - EDA results
  - Interactive dashboard
  - Classification evaluation results

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - Predict whether first stage of SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

Methodology

Executive Summary

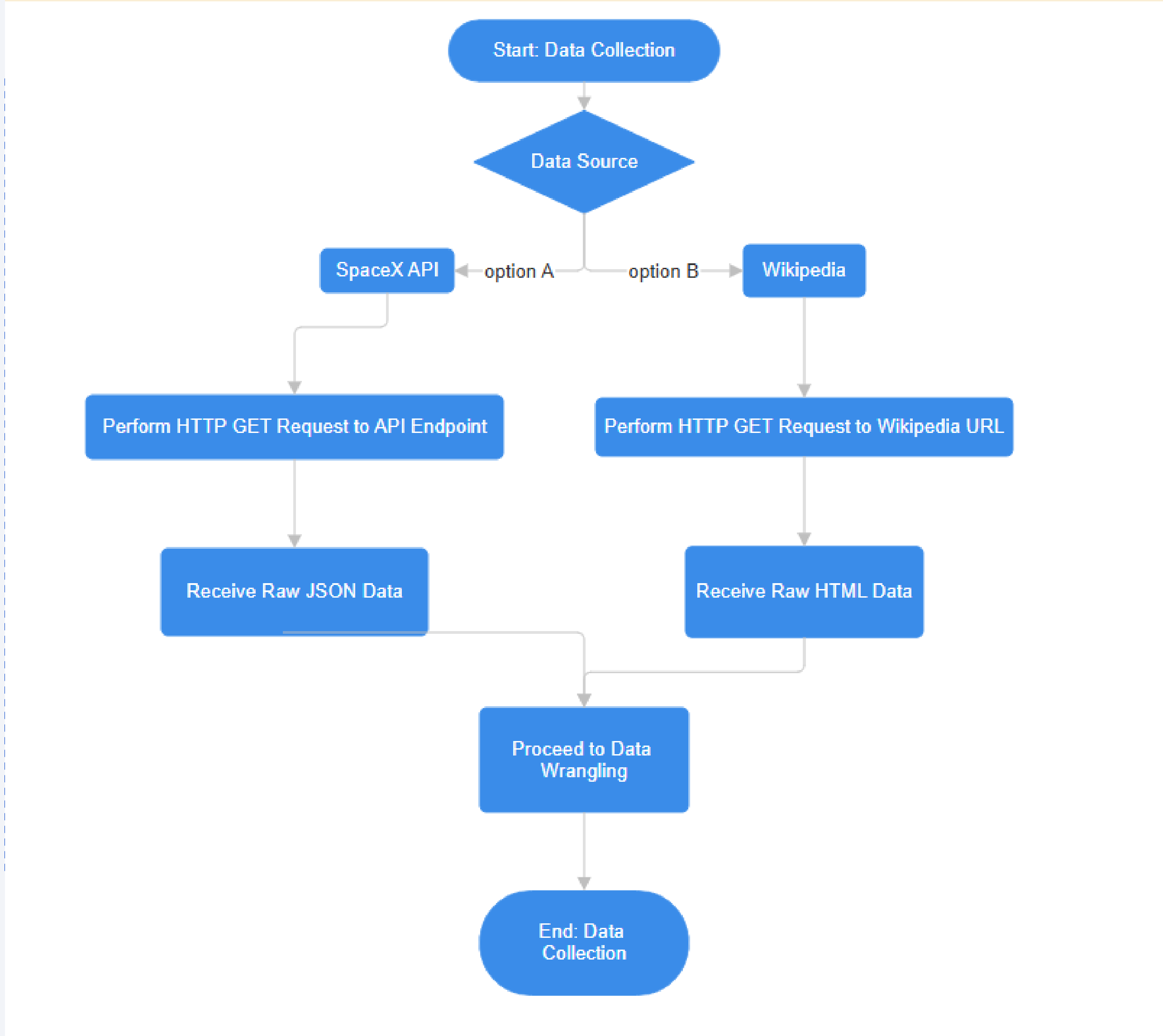
- Data collection methodology:
  - API Collection, Web Scraping
- Perform data wrangling
  - Implemented one hot encoding for categorical features
  - Cleaned data by dealing with missing values and eliminated unnecessary columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - K-Nearest Neighbors (KNN), Logistic Regression(LR), Decision trees (DT), Support Vectors Machine (SVM), grid search for hyperparameter tuning, confusion matrix for evaluation

Data Collection

- API Collection
- Source: SpaceX REST API
- Method: An HTTP GET request was made to the SpaceX API's /v4/launches/past endpoint to retrieve historical launch data.
- Web Scraping
- Source: Wikipedia page "List of Falcon 9 and Falcon Heavy launches"
- Method: The BeautifulSoup library was used to extract launch records from the HTML table on the Wikipedia page. An HTTP GET request was made to the static URL of the page.

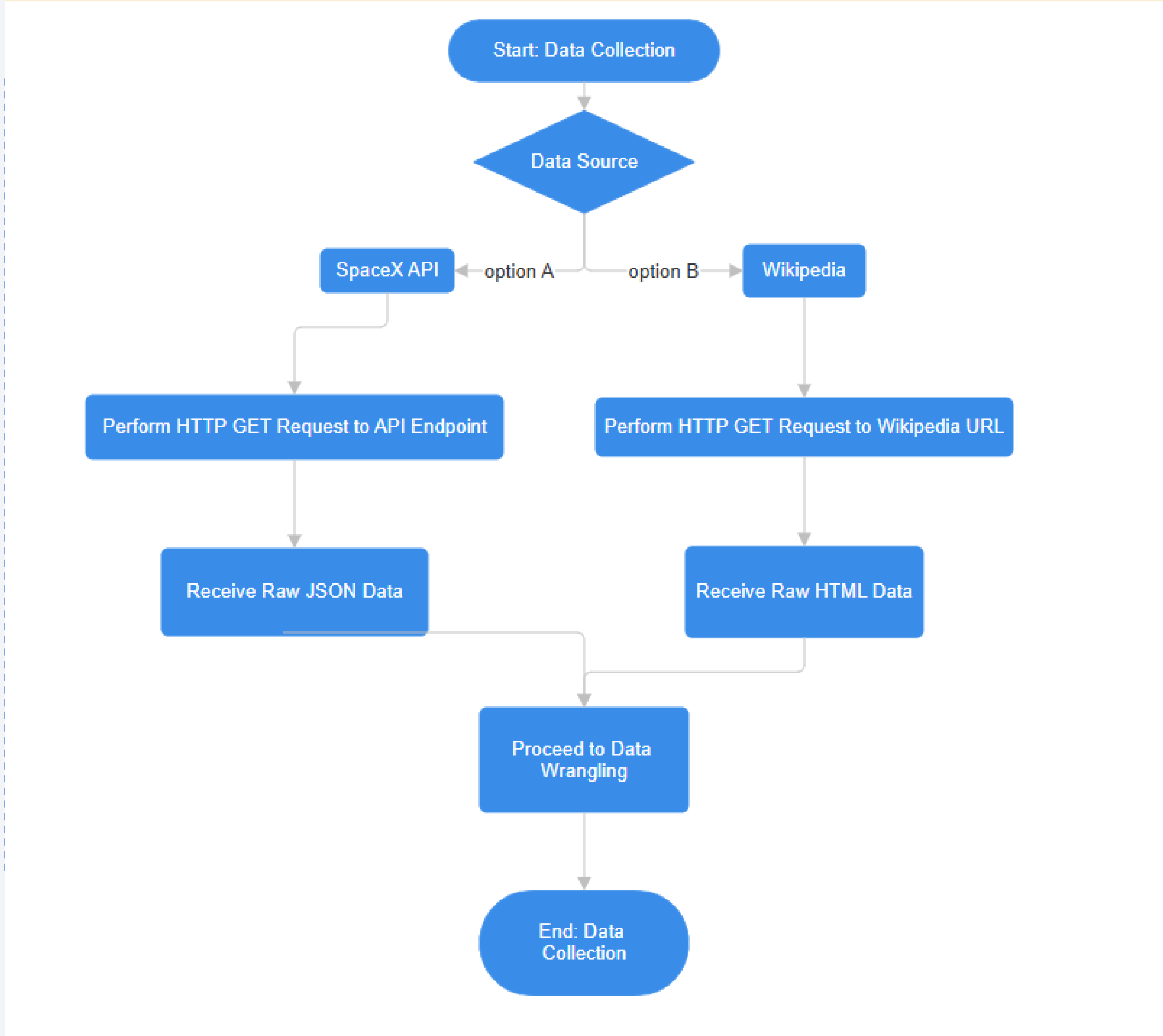
Data Collection – SpaceX API

- Key Phrase: "Request to the SpaceX API“, “parse the SpaceX launch data using the GET request”
- GitHub URL of the completed SpaceX API calls notebook Data_Collection.



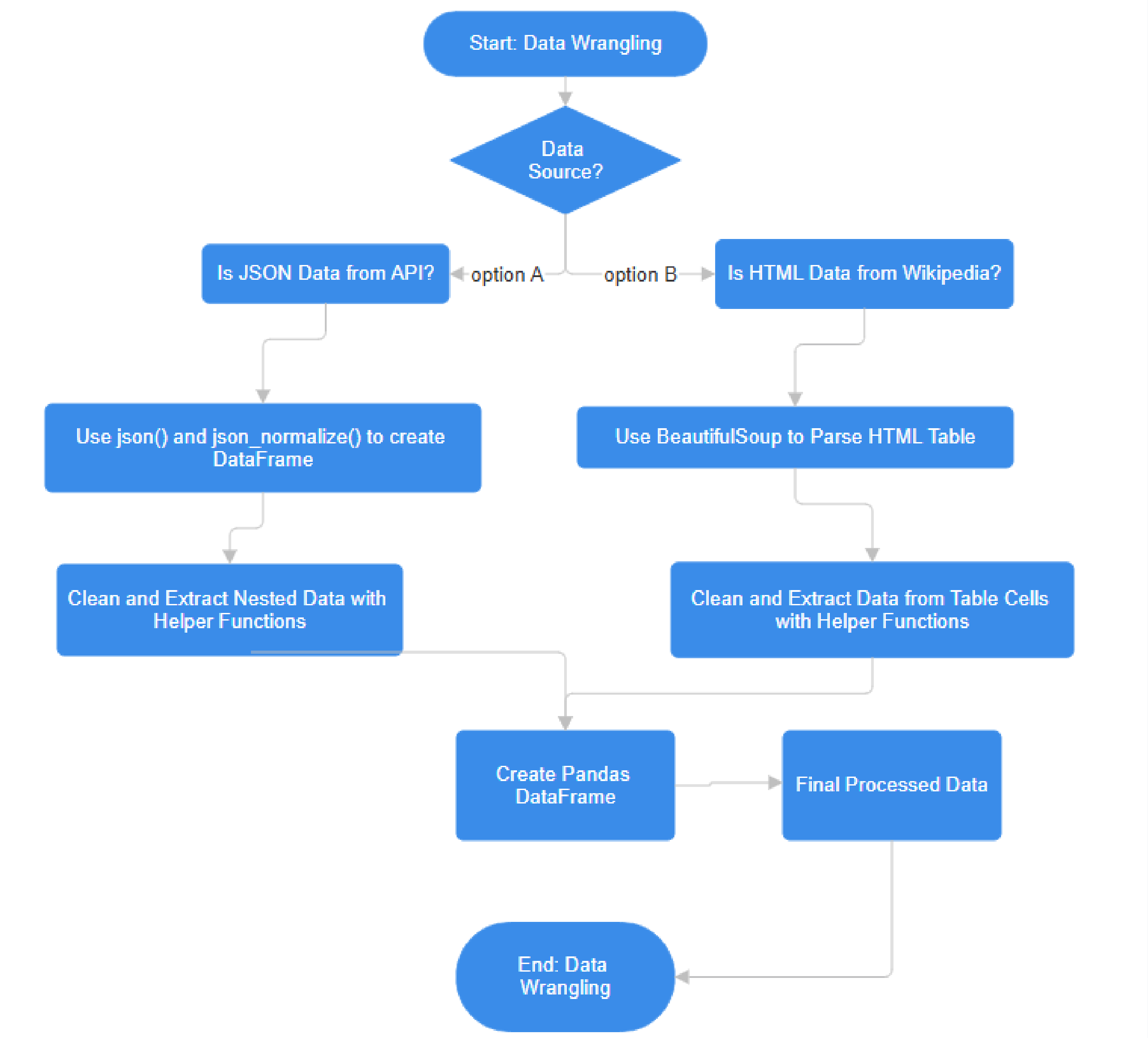
Data Collection - Scraping

- Key Phrase: "Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia"
- The GitHub URL of the completed [(Web Scraping)]() notebook.



Data Wrangling

- The data wrangling was performed to clean and format the raw data collected from both the SpaceX API and Wikipedia into a structured and usable format.
- Key Phrases:"Clean the requested data""basic data wrangling and formating""Parse the table and convert it into a Pandas data frame"
- The GitHub URL of completed [data wrangling]notebook.


Start: Data Wrangling
Data Source?
option A
option B
Is JSON Data from API?
Is HTML Data from Wikipedia?
Use json() and json_normalize() to create DataFrame
Use BeautifulSoup to Parse HTML Table
Clean and Extract Nested Data with Helper Functions
Clean and Extract Data from Table Cells with Helper Functions
Create Pandas DataFrame
Final Processed Data
End: Data Wrangling

EDA with Data Visualization

- Scatter Plots (Flight Number & Payload Mass): These charts helped us see trends over time. We could check if there was a relationship between the order of a flight (Flight Number) and the payload it carried, or if a site's success rate improved over time.
- Bar Chart (Success Rate by Launch Site): This chart was used for a direct, simple comparison. It quickly showed us which launch sites were the most and least successful overall.
- Line Plot (Success Rate by Payload Mass): This plot helped us analyze a key question: Does the weight of the payload affect a launch's success? It showed if certain mass ranges had a higher or lower success rate.
- Scatter Plot (Class vs. Orbit): This visualization helped us understand if the final destination of the rocket (Orbit) was a major factor in determining a successful landing.
- The GitHub URL of completed EDA with data visualization notebook.

EDA with SQL

- The SQL queries were used to analyze the SpaceX mission data to find key insights.
- Data exploration: We identified unique launch sites and filtered records to examine specific groups of launches.
- Performance analysis: We calculated total and average payload mass, and counted the number of successful vs. failed missions.
- Outcome investigation: We analyzed the data to find patterns related to mission outcomes, such as the date of the first successful landing or the performance of specific boosters.
- The GitHub URL of completed EDA with SQL notebook.

Build an Interactive Map with Folium

- Markers: We used markers to pinpoint the exact location of each of the four launch sites.
- Color Markers: We color-coded the markers (green for success, red for failure) to easily visualize the outcome of each launch directly on the map.
- Lines: Lines were drawn to show the distance from a launch site to a coastline, railways, highways, and cities. This was done to investigate if proximity to these features influenced the choice of a launch site.
- The GitHub URL of completed interactive map with Folium map.

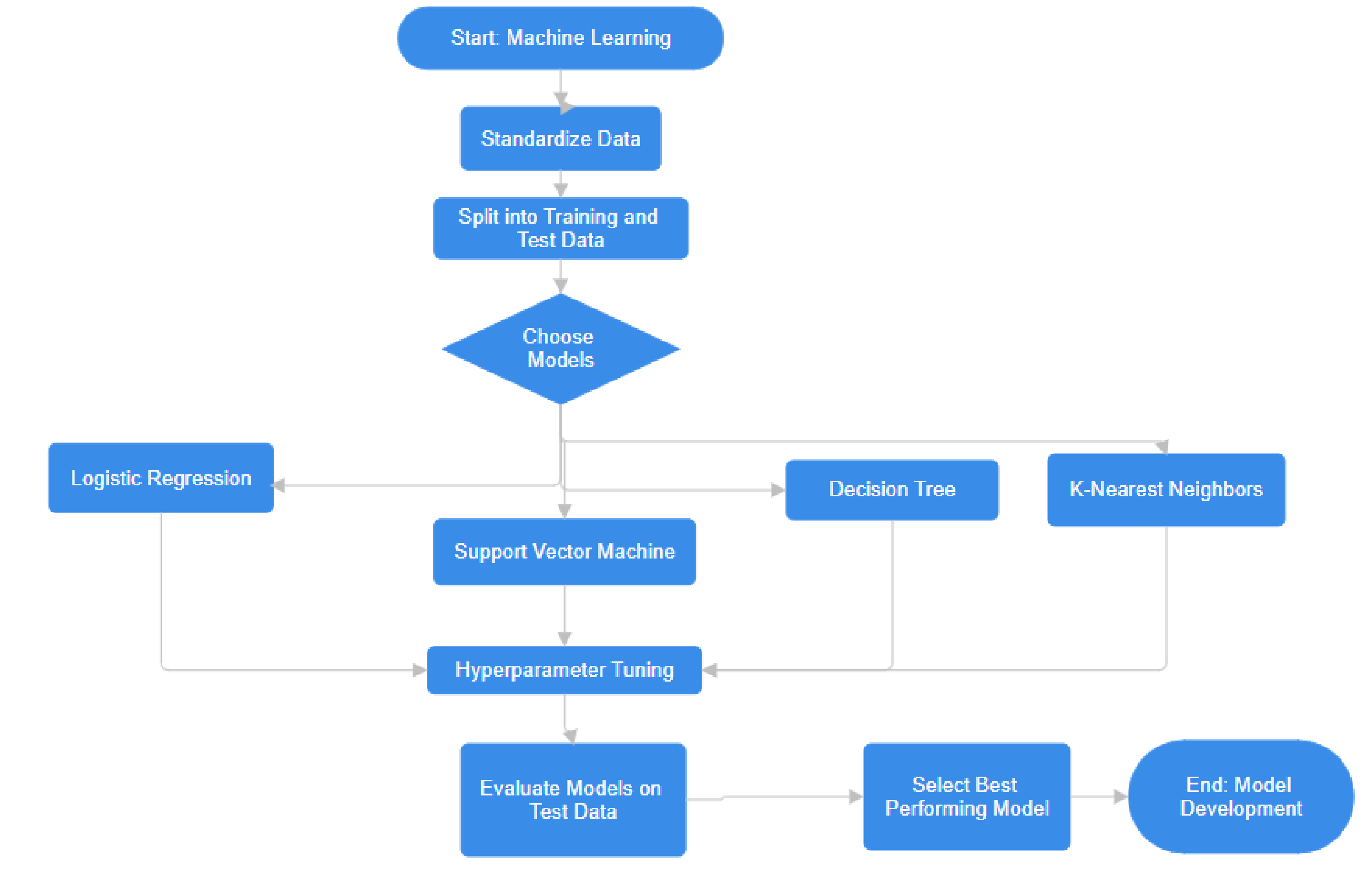
Build a Dashboard with Plotly Dash

The dashboard uses two interactive charts to analyze the SpaceX mission data:

- Launch Success Pie Chart: Shows the success vs. failure counts for a selected launch site, or the overall success count for all sites.
- Payload vs. Success Scatter Plot: Displays the correlation between a rocket's payload mass and its launch outcome. It can be filtered by a specific launch site and a chosen payload range.
- The GitHub URL of completed Plotly Dash lab.

Predictive Analysis (Classification)

- The development of the best performing classification model involved a sequence of key steps:
- Data Preparation: The data was first standardized to ensure all features had an equal impact on the model. This was followed by splitting the data into separate training and testing sets.
- Model Training & Tuning: Four different models—Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors—were trained and then fine-tuned using Hyperparameter tuning to optimize their performance.
- Best Model Selection: After evaluation on the test data, the Decision Tree was chosen as the best performing model. It achieved the highest training accuracy and performed as well as the other models on the test data.
- The GitHub URL of completed [predictive analysis lab]



Results

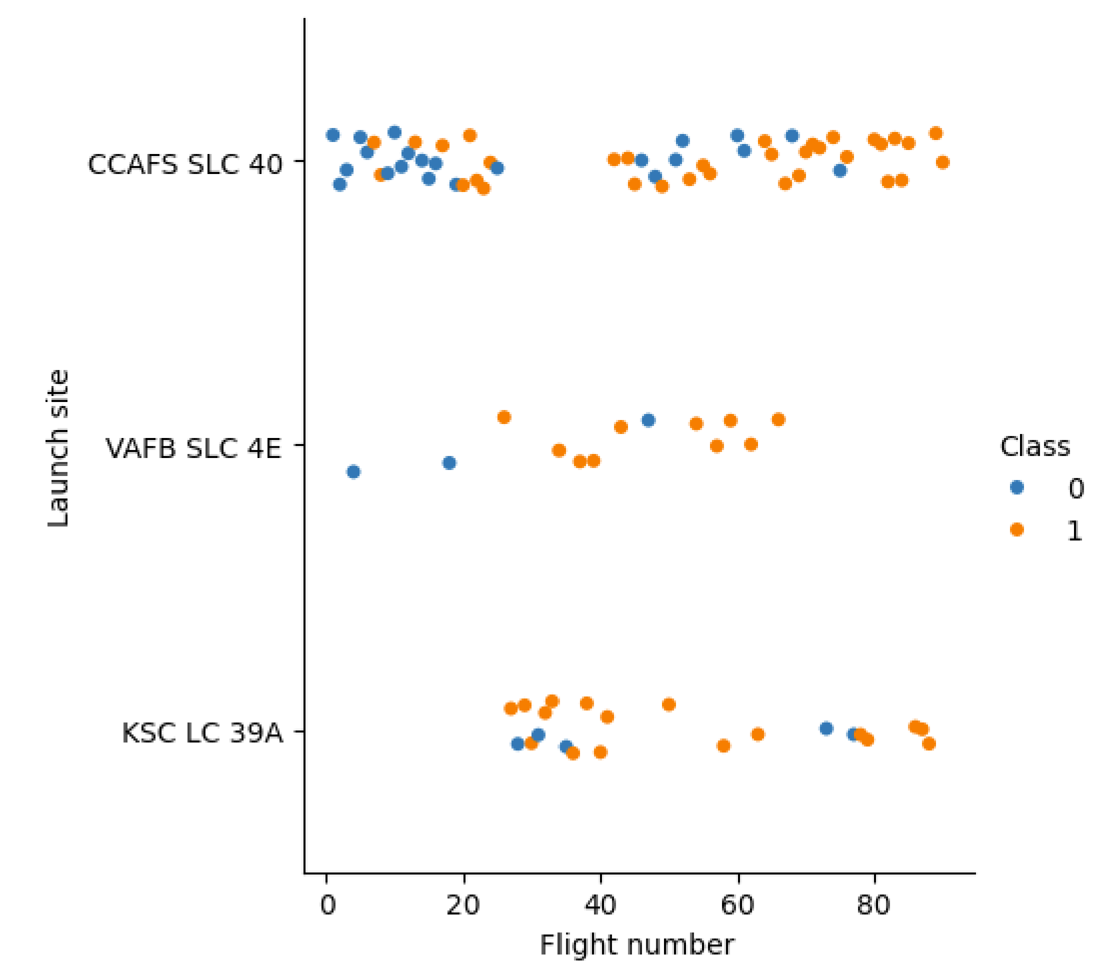
- SVM, KNN, LR are the best in terms of prediction Accuracy
- Low weighted payloads performed better than heavier payloads
- The success rate of successful launches is improving every year
- KSC LC 39A had the most successful launches out of all launch sites
- Orbit GEO, HEO, SEO, ES L1 has the best success rate

Section 2

Insights drawn from EDA

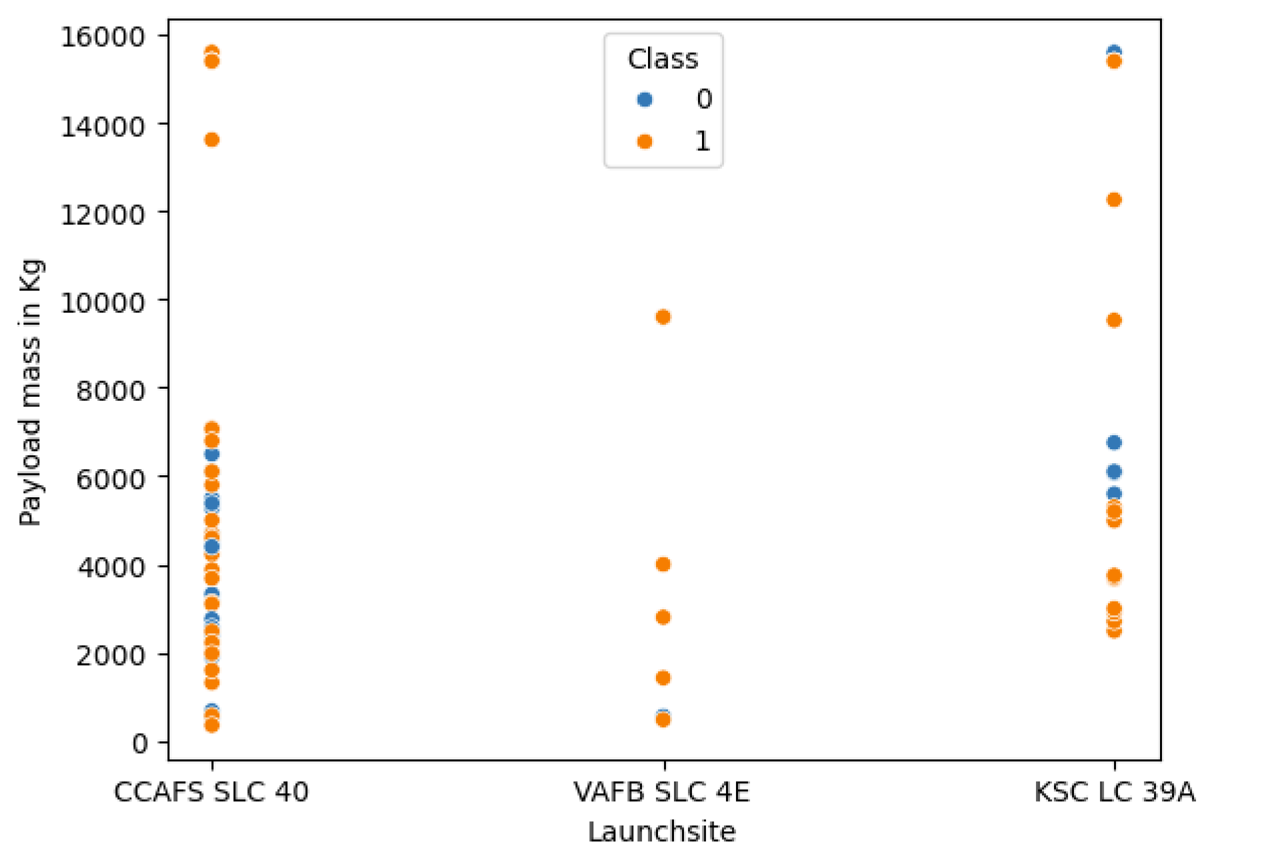
Flight Number vs. Launch Site

- The plot shows the relationship between a mission's Flight Number and its Launch Site
- The colors of the data points indicate the outcome of the launch: success or failure
- The purpose of this visualization is to see the launch success rate at each site



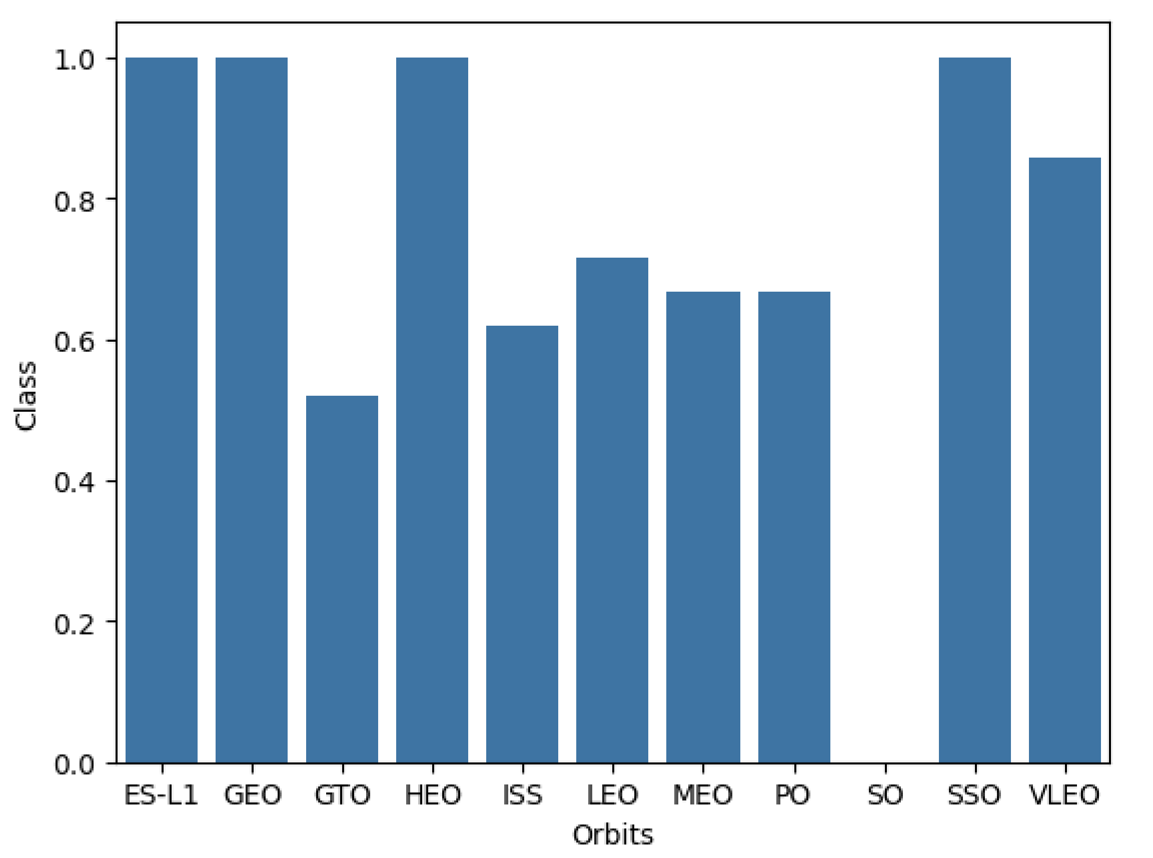
Payload vs. Launch Site

- This scatter plot visualizes the relationship between the payload mass of a rocket and the launchsite
- The colors of the data points would indicate the launch outcome (success or failure)
- The purpose of the plot is to analyze if specific launch sites were more successful with heavier or lighter payloads. It would help to see if any site had a tendency to have a higher success rate for a particular range of payload masses



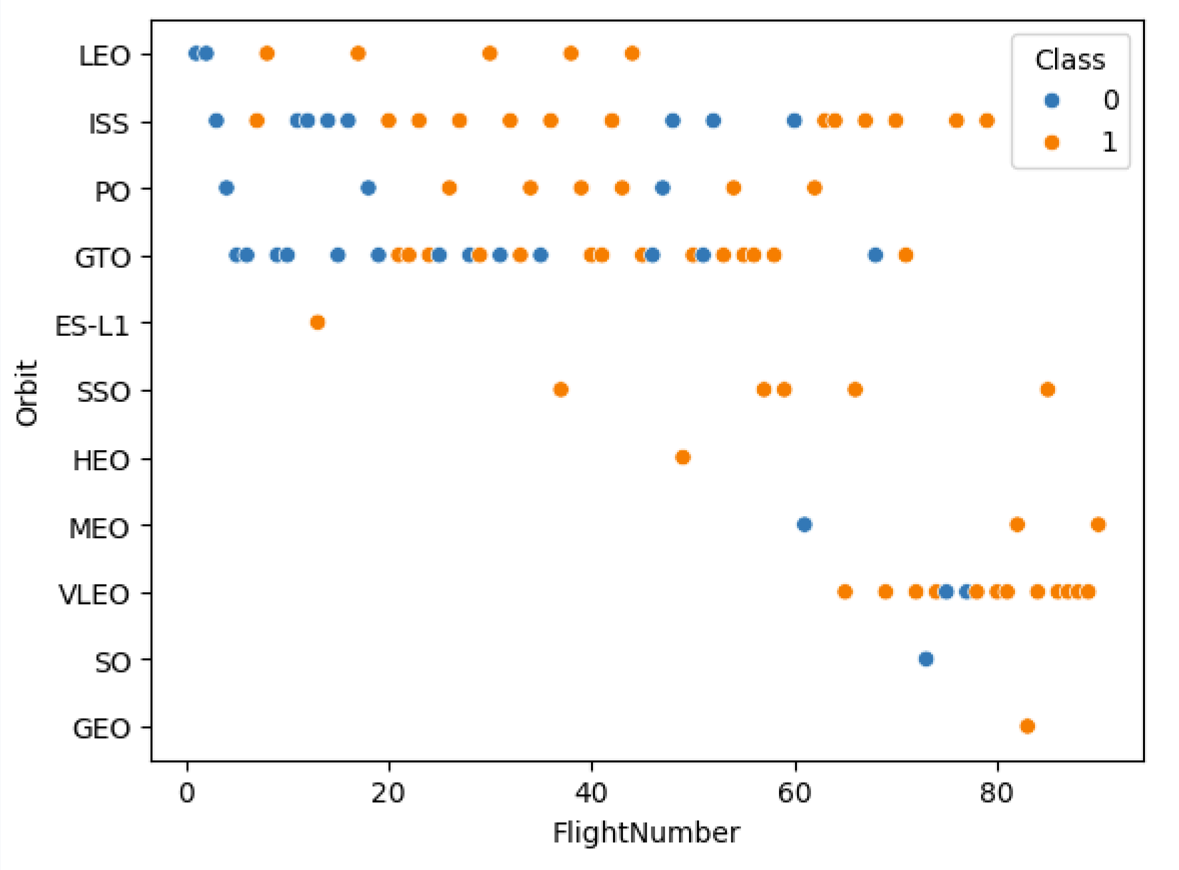
Success Rate vs. Orbit Type

- Bar Chart: Each bar represents a different Orbit type
- Success Rate: The height of each bar shows the overall success rate for that specific orbit, making it easy to compare the performance across all orbit types
- This chart helps to answer the question: "Does the type of orbit influence a launch's success?" By comparing the heights of the bars, you can quickly identify which orbits have a higher probability of a successful mission



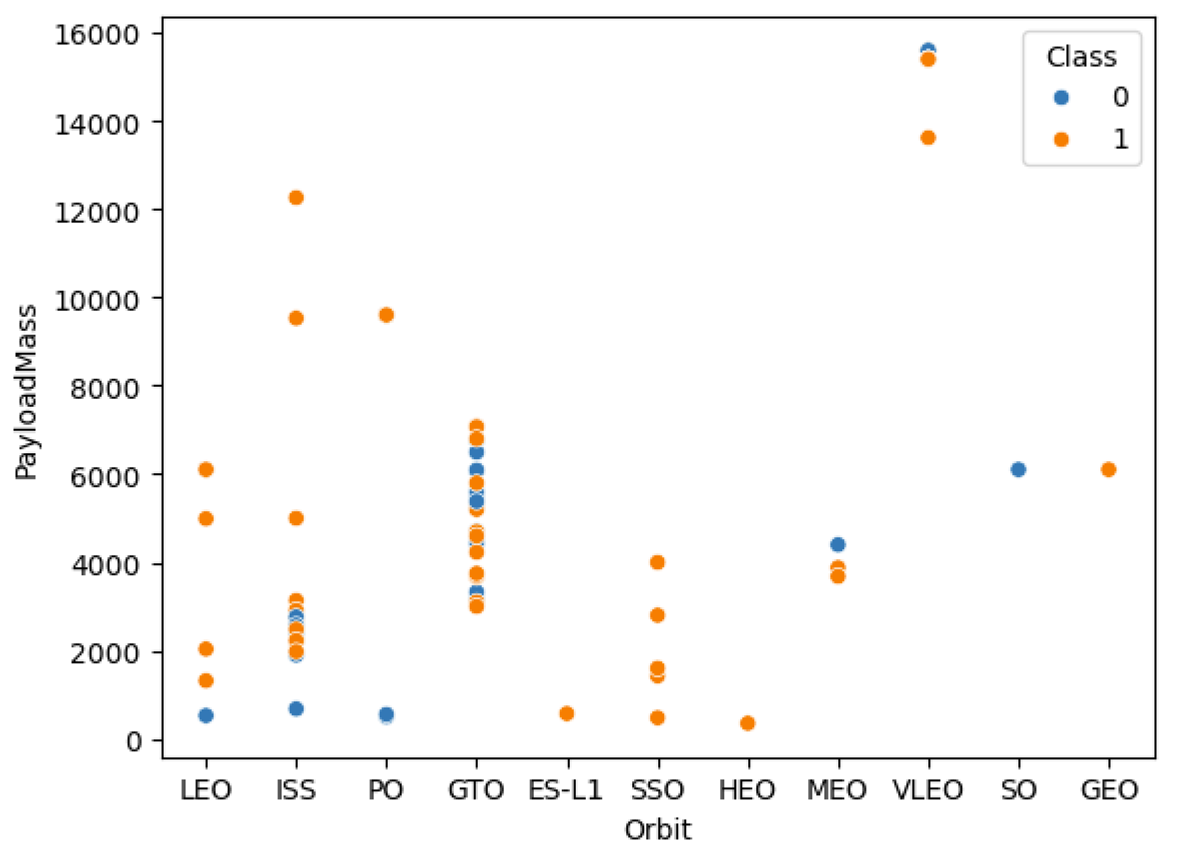
Flight Number vs. Orbit Type

- This scatter plot, visualizes the launch outcomes over time for each type of orbit
- The colors of the data points indicate the outcome of the launch: success or failure
- The purpose of this chart is to understand if certain orbits were successful compared to other orbits. You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success



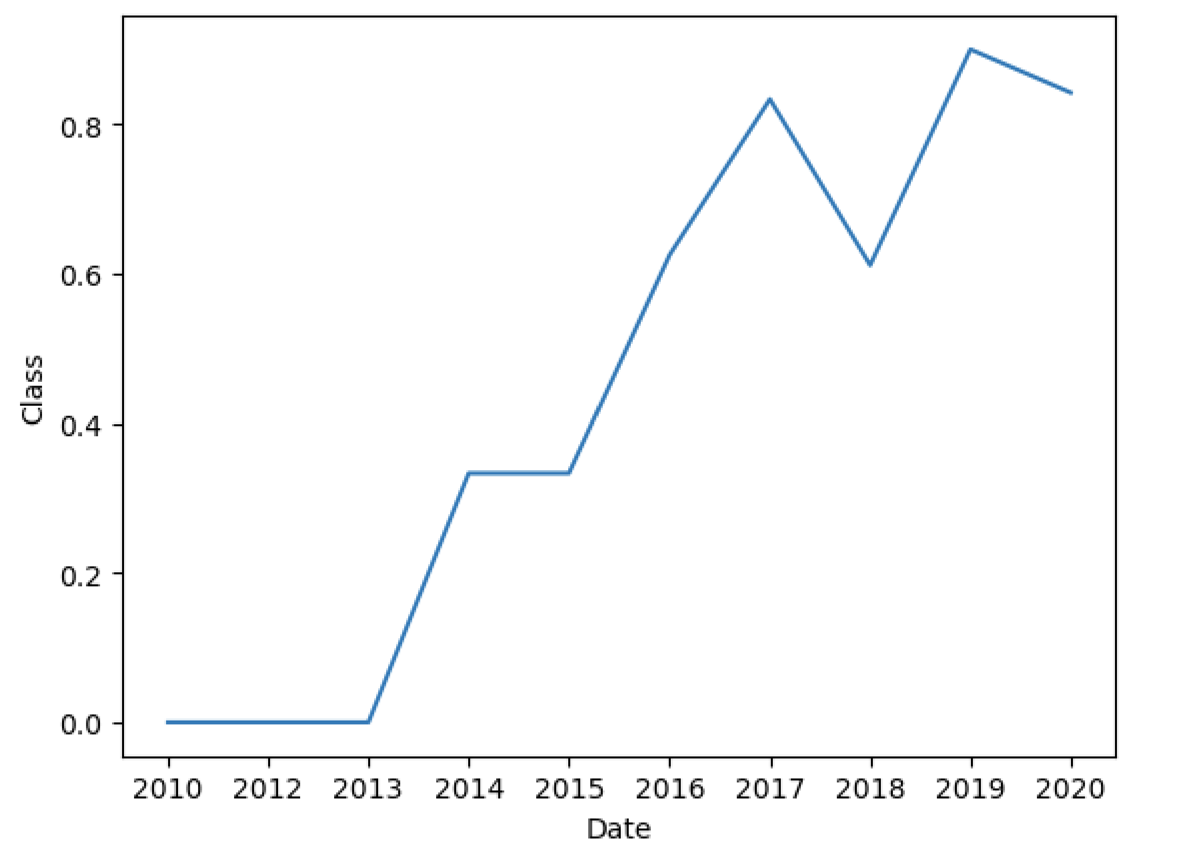
Payload vs. Orbit Type

- This chart helps to visualize the relationship between the payload mass and the type of orbit for each mission
- The colors of the data points indicate the launch outcome (success or failure)
- This plot would help determine if specific booster versions were used for particular orbits and how payload mass may have influenced the outcome for different mission types



Launch Success Yearly Trend

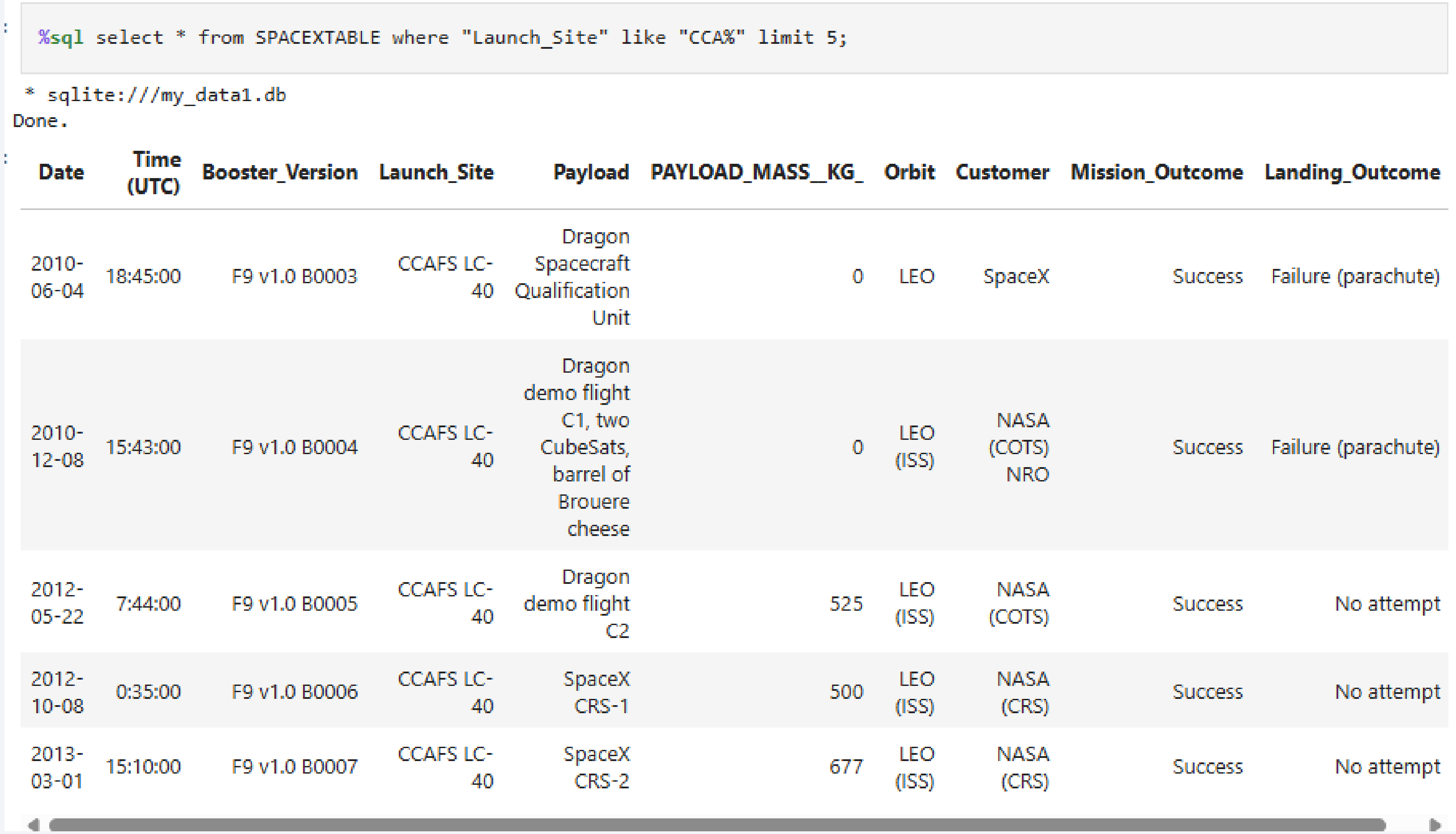
- This line chart shows the average launch success rate over the years, which is a key measure of the project's overall progress and reliability
- The line itself visualizes the trend, showing the success rate has been increasing consistently over time
- The purpose of this chart is to provide a high-level view of how SpaceX's performance has evolved, which is a crucial part of the exploratory data analysis



All Launch Site Names

- Here are the unique launch site names from the dataset.
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40

Launch Site Names Begin with 'CCA'



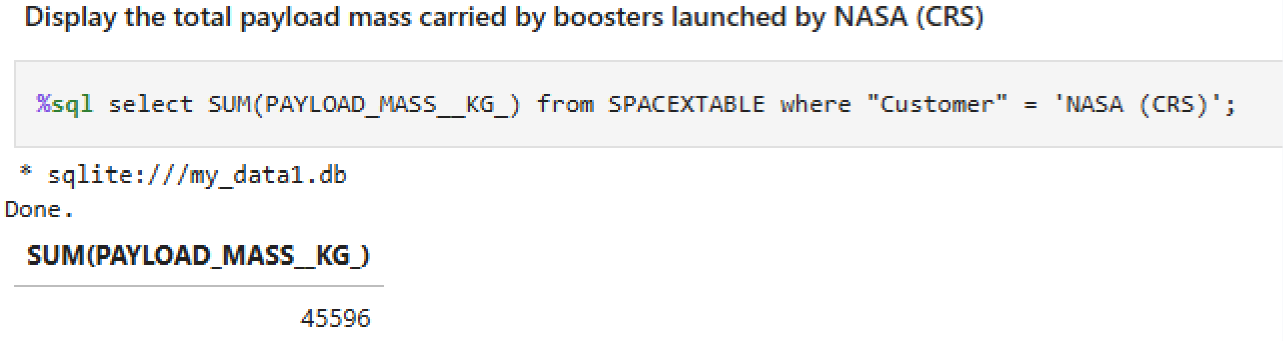
```
%sql select * from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5;
```

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

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

- This query selects all columns (*) from the SPACEXTABLE where the Launch_Site column starts with CCA (LIKE "CCA%"), and it limits the output to the first five results.

Total Payload Mass



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

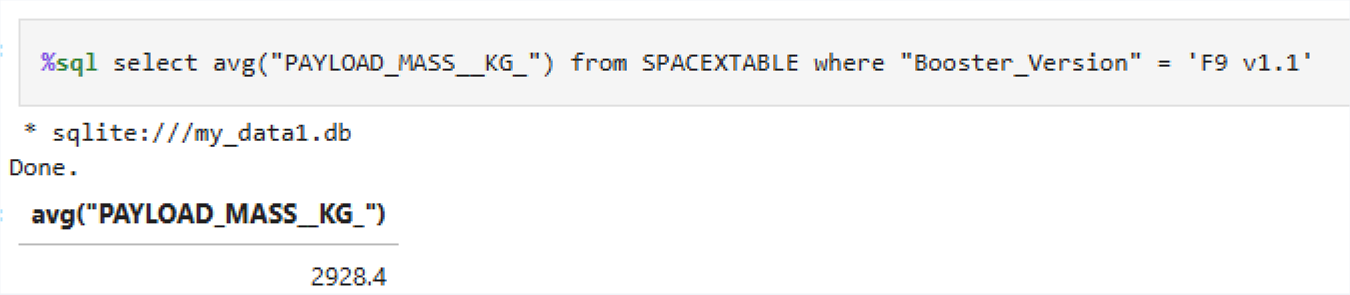
```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = 'NASA (CRS)';
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- This query uses the SUM function to calculate the total of all values in the PAYLOAD_MASS__KG_ column, but it only includes records where the Customer is NASA (CRS).

Average Payload Mass by F9 v1.1



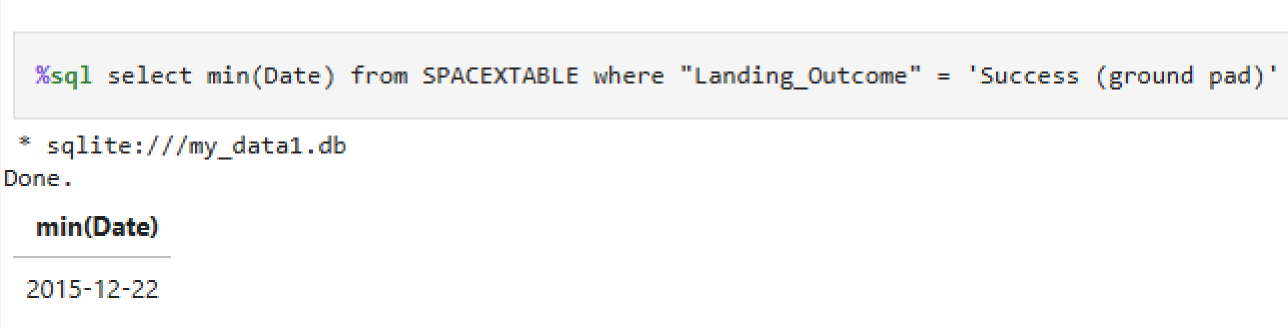
```
%sql select avg("PAYLOAD_MASS__KG_") from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'
```

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

avg("PAYLOAD_MASS__KG_")
2928.4

- This query uses the AVG function to compute the average of the PAYLOAD_MASS__KG_ column, specifically for records where the Booster_Version is F9 v1.1.

First Successful Ground Landing Date



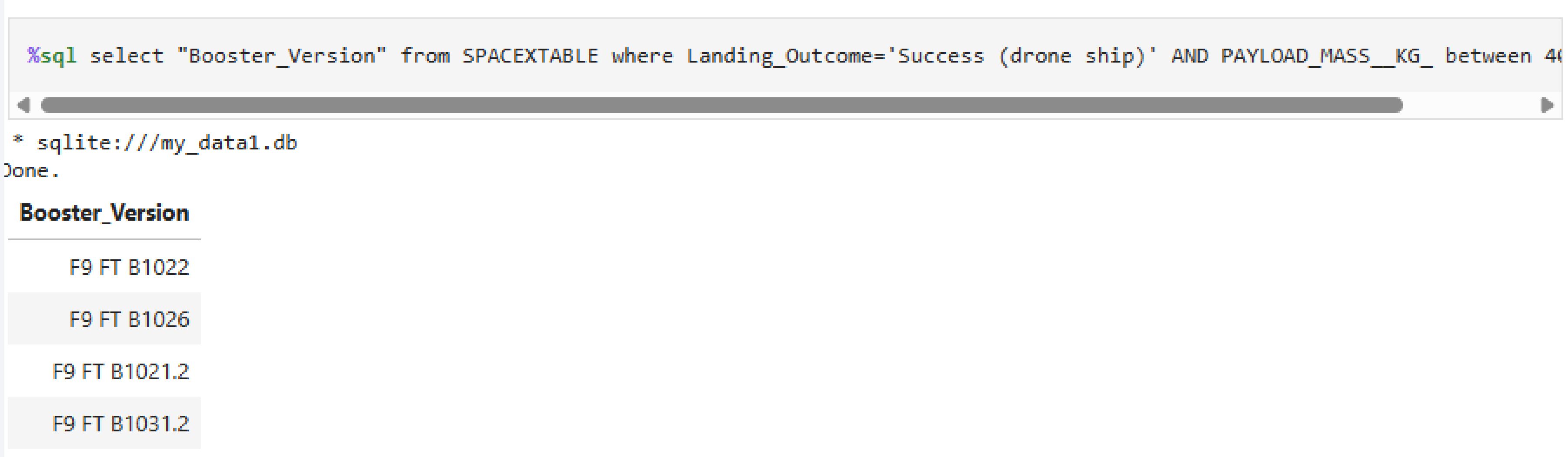
```
%sql select min(Date) from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

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

min(Date)
2015-12-22

- This query uses the MIN function to find the earliest date in the dataset where the Landing_Outcome was 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000



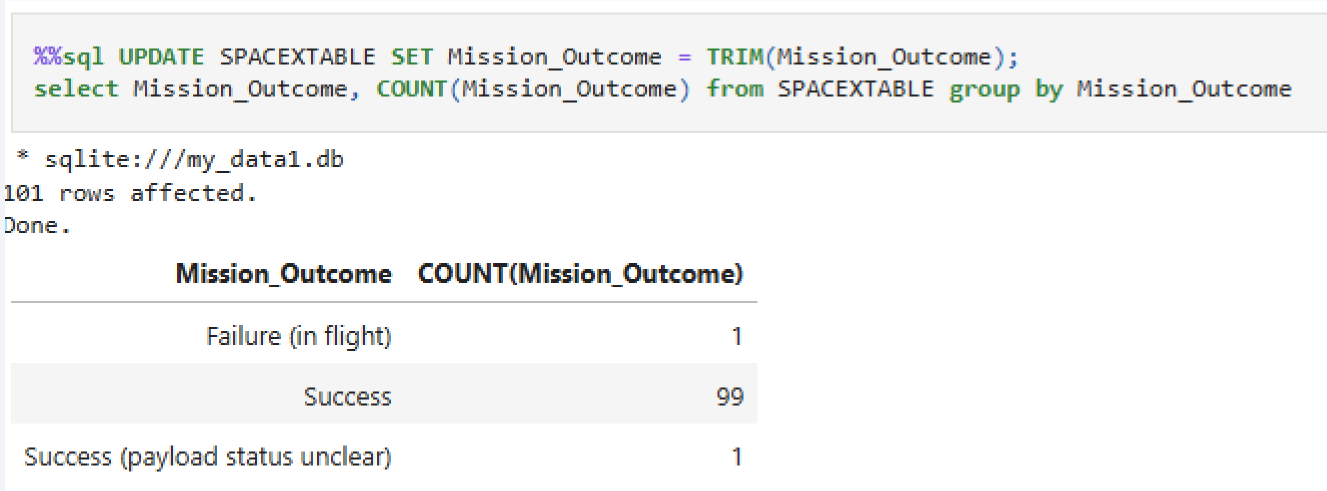
```
%sql select "Booster_Version" from SPACEXTABLE where Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ between 4
```

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

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

- select "Booster_Version" from SPACEXTABLE where Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ between 4000 AND 6000
- This query selects the Booster_Version from the table, filtering for records where the Landing_Outcome is exactly 'Success (drone ship)' AND the PAYLOAD_MASS__KG_ is between 4000 and 6000 (inclusive)

Total Number of Successful and Failure Mission Outcomes



```
%%sql UPDATE SPACEXTABLE SET Mission_Outcome = TRIM(Mission_Outcome);
select Mission_Outcome, COUNT(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

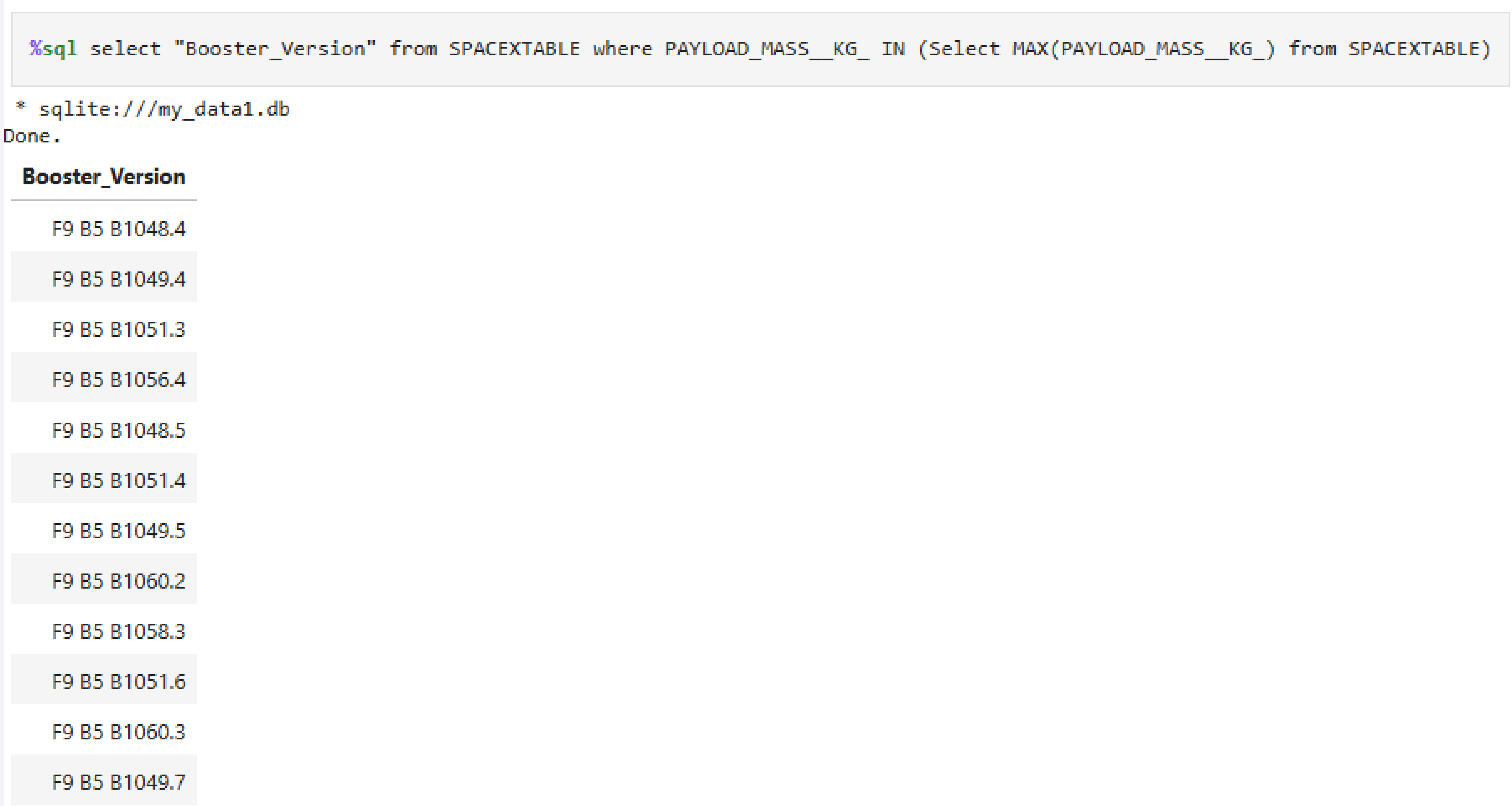
```
 * sqlite:///my_data1.db
101 rows affected.
Done.
```

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query groups the data by the Mission_Outcome column and then uses the COUNT function to tally the number of occurrences for each outcome. This provides a clear summary of how many missions were successful versus how many failed.

Boosters Carried Maximum Payload

This query uses a subquery to first find the single maximum value in the PAYLOAD_MASS__KG_ column. Then, the main query selects the Booster_Version for every record that matches this maximum payload mass



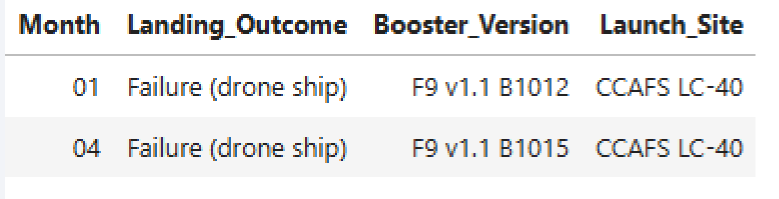
```
%sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_ IN (Select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

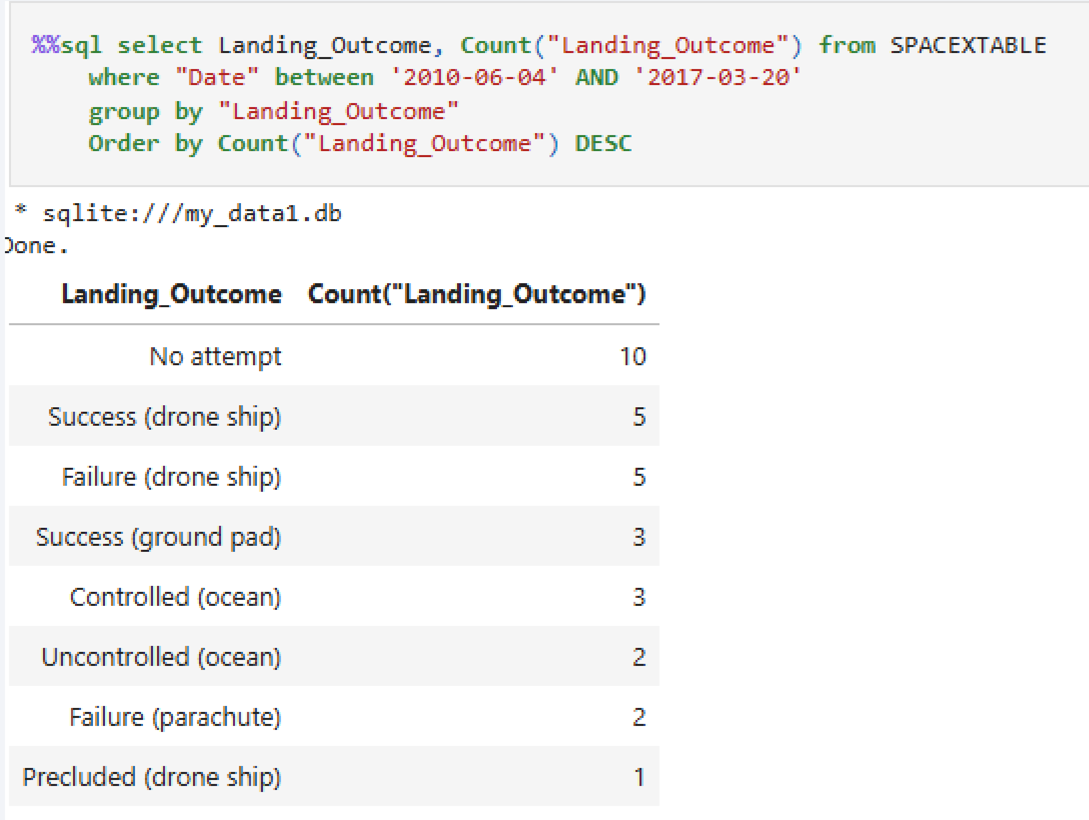
2015 Launch Records

- select substr(Date, 6,2) as Month,"Landing_Outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where substr(Date,0,5)='2015' AND "Landing_Outcome" ='Failure (drone ship)’
- This query selects the month (substr(Date, 6,2)), the Landing_Outcome, Booster_Version, and Launch_Site from the SPACEXTABLE. It filters the records to include only those from the year 2015 (substr(Date,0,5)='2015') and where the Landing_Outcome was 'Failure (drone ship)'



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



```
%%sql select Landing_Outcome, Count("Landing_Outcome") from SPACEXTABLE
    where "Date" between '2010-06-04' AND '2017-03-20'
    group by "Landing_Outcome"
    Order by Count("Landing_Outcome") DESC
```

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

Landing_Outcome	Count("Landing_Outcome")
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

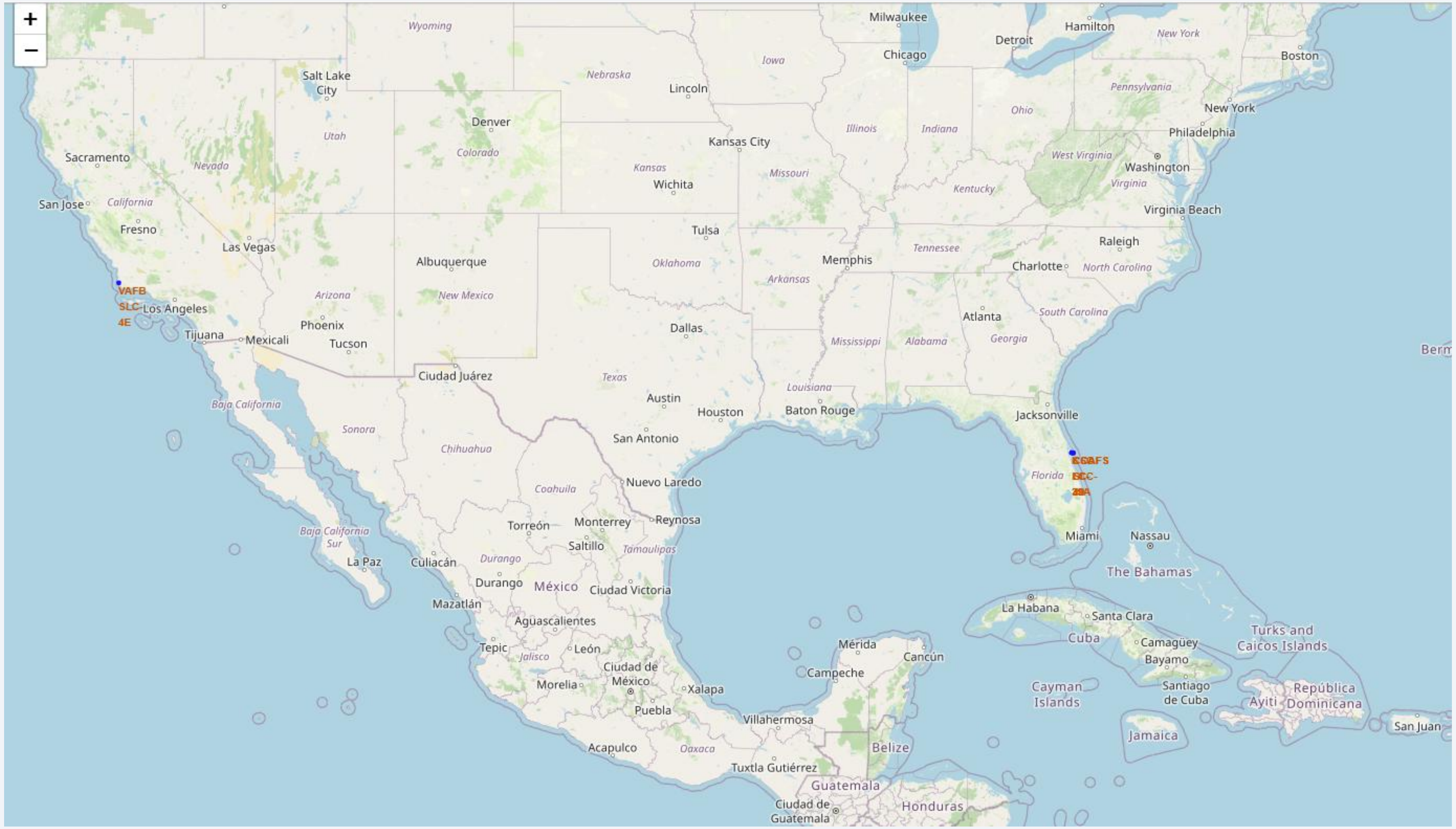
- This query filters the data for a specific date range, then groups the results by Landing_Outcome. It counts the number of times each outcome occurred and presents the results from most frequent to least frequent. This provides a clear ranking of the most common landing outcomes during that period

Section 3

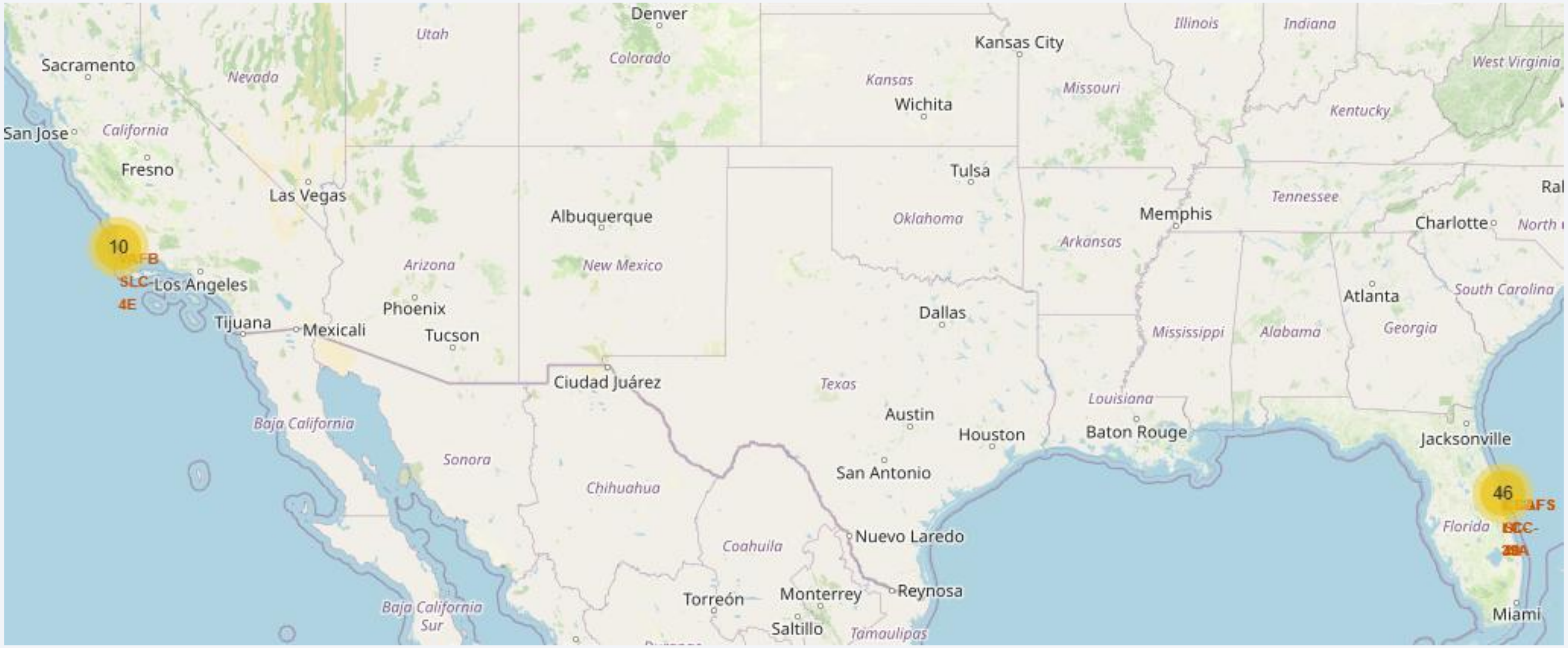
Launch Sites Proximities Analysis

Launch sites

Markers: Pinpoint the exact location of the four main launch sites
Lines: Show distances from each site to coastlines, railways, and cities, explaining why sites were chosen for safety and accessibility
Visual Confirmation: The map confirms that the sites are strategically located near the coast, away from populated areas, and with good access to transportation

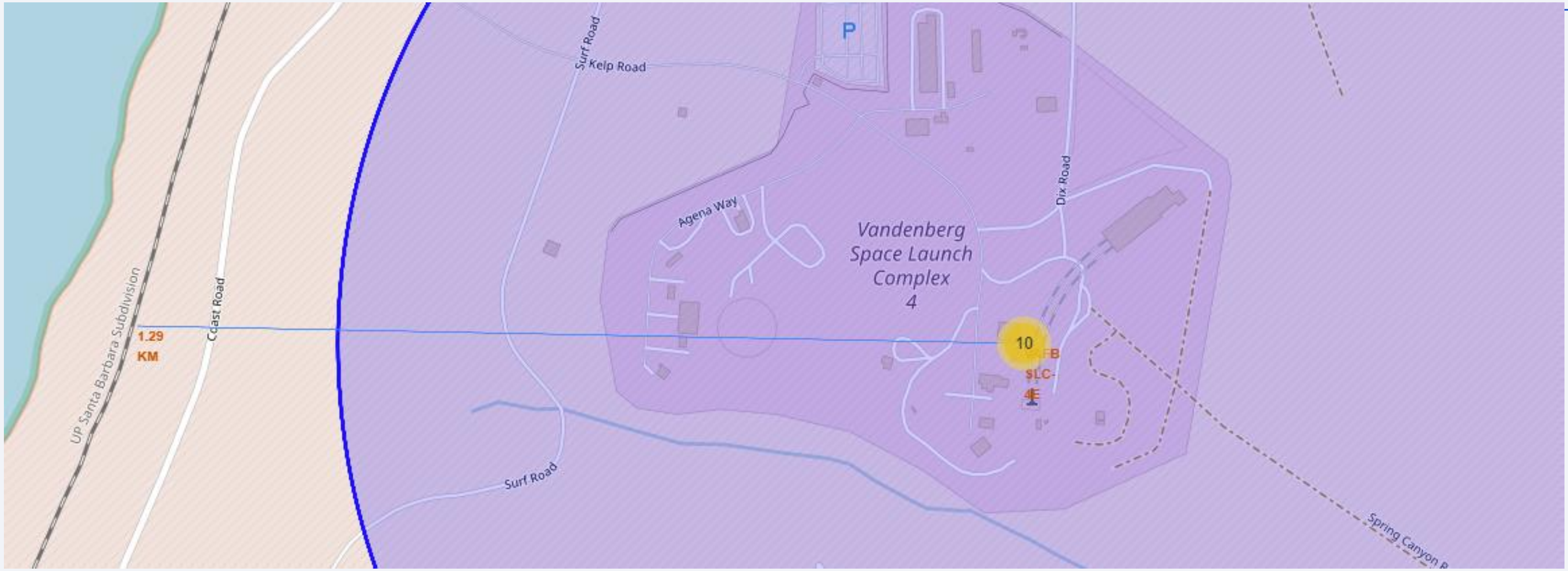


Launch Outcome Map



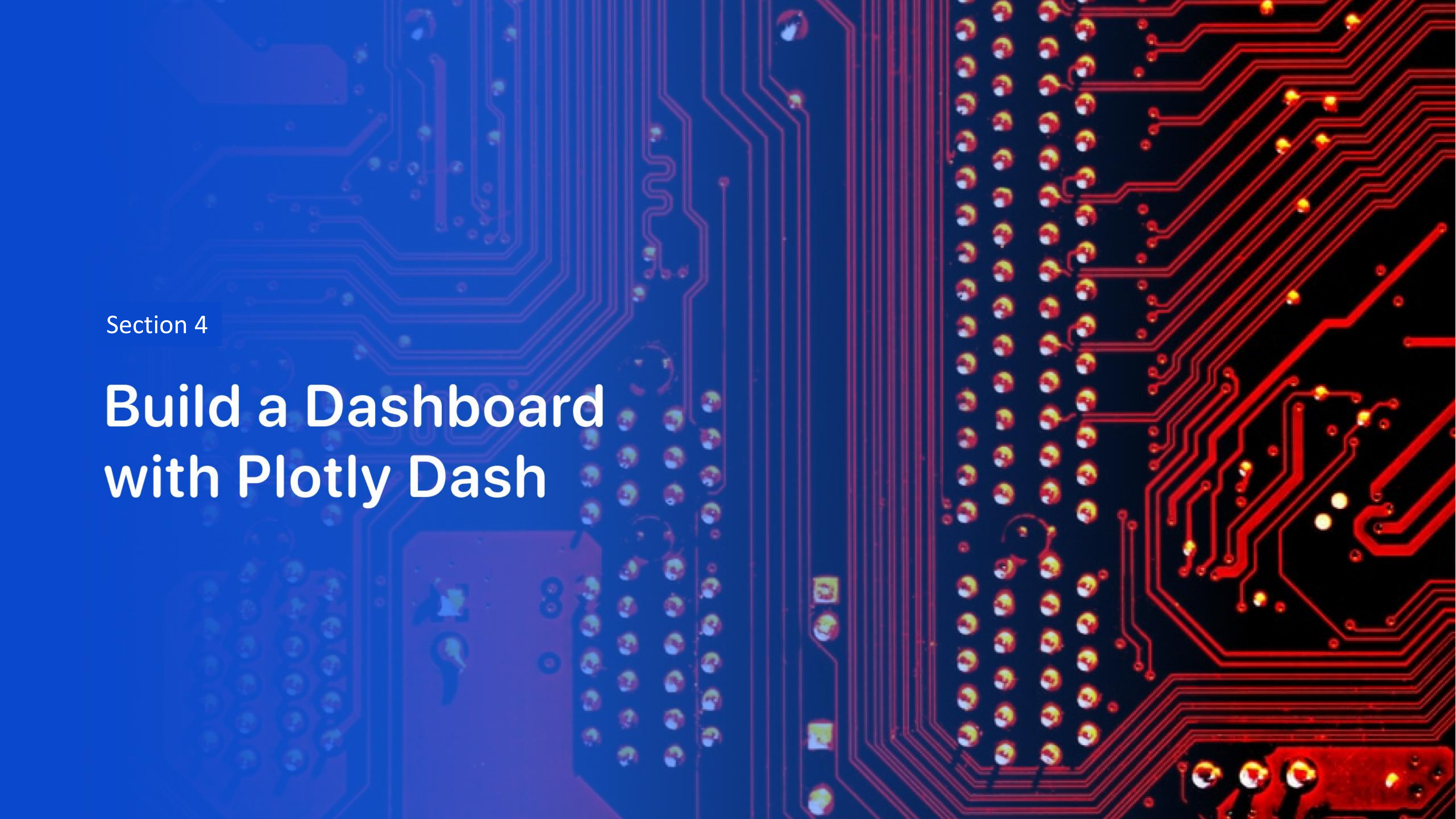
- This map uses color-coded markers to visualize the outcome of each launch directly on the map
- Markers: Each marker represents a single launch
- Colors: Green indicates a successful launch (class=1), Red indicates a failed launch (class=0)
- This visualization helps to quickly identify if a particular launch site has a high success or failure rate

Launch Site Proximity Map



This map visualizes the exact location of a selected launch site and calculates its distance to key geographic features

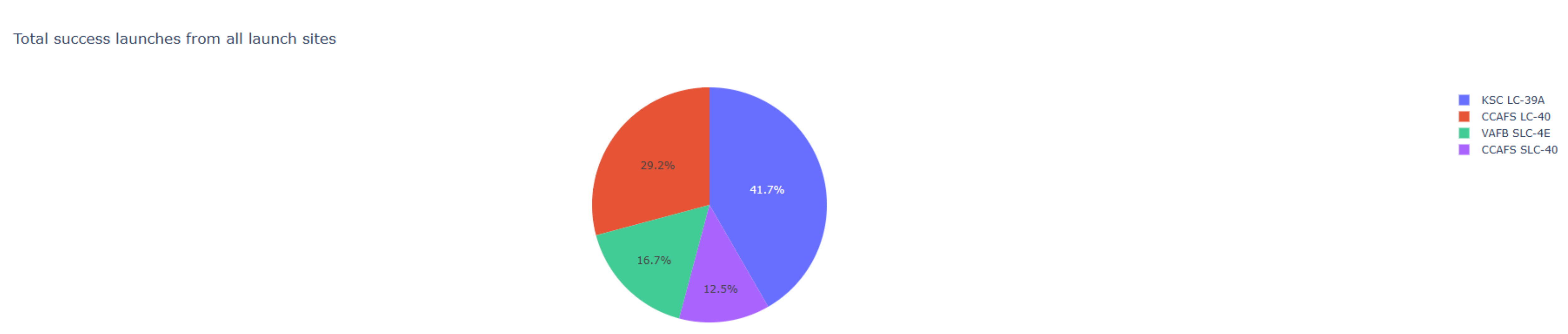
- **Distance to Key Features:** The map draws lines from the launch site to the nearest railway. The distance is calculated and displayed
- **Key Finding:** This confirms that launch sites are strategically located near coasts for safety, and have good access to transportation infrastructure like railways and highways



Section 4

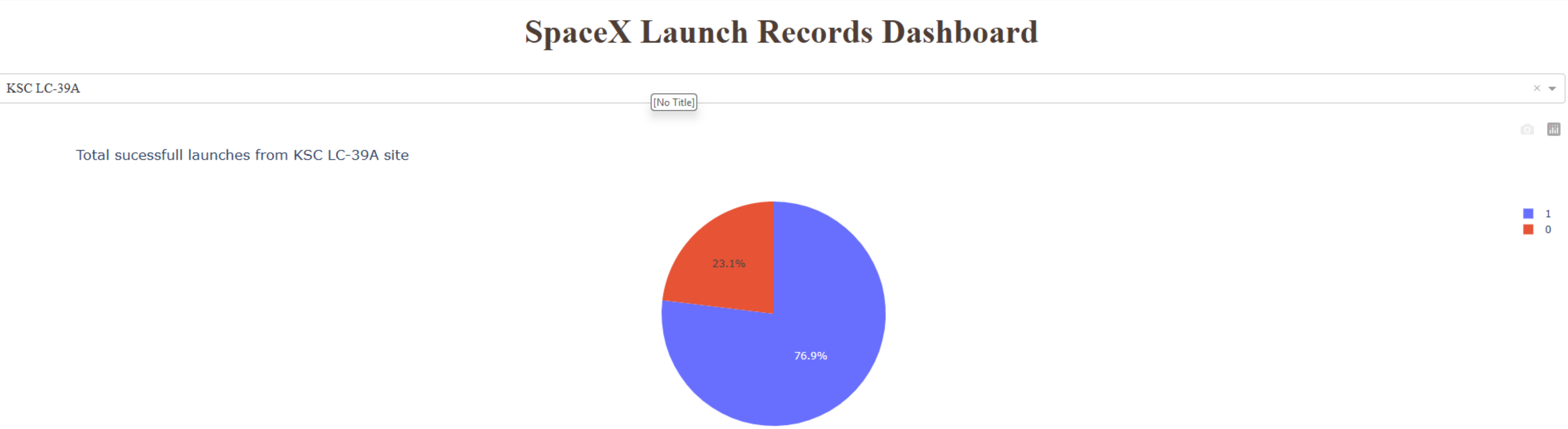
Build a Dashboard with Plotly Dash

Launch Success Across All Sites


Total success launches from all launch sites

- The pie chart demonstrates that overall, the SpaceX mission success rate
- It provides a simple and clear overview of the project's reliability
- From the pie chart it is clear that KSC LC 39A has high success rate out of these 4 sites

Launch Success at KSC LC-39A

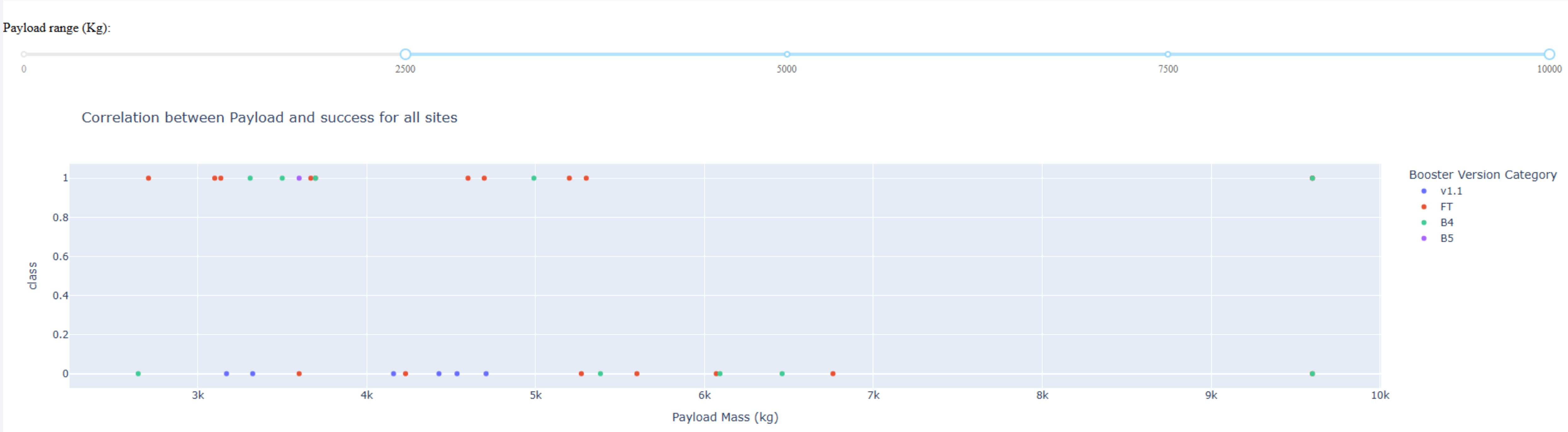


SpaceX Launch Records Dashboard

KSC LC-39A

- This chart visualizes the success in blue and failure count in red for the launch site with the highest success ratio, which is KSC LC-39A based on the overall data

Payload Mass and Launch Outcome



Payload range (Kg):

- This interactive scatter plot visualizes the correlation between a rocket's payload mass and the success of its launch

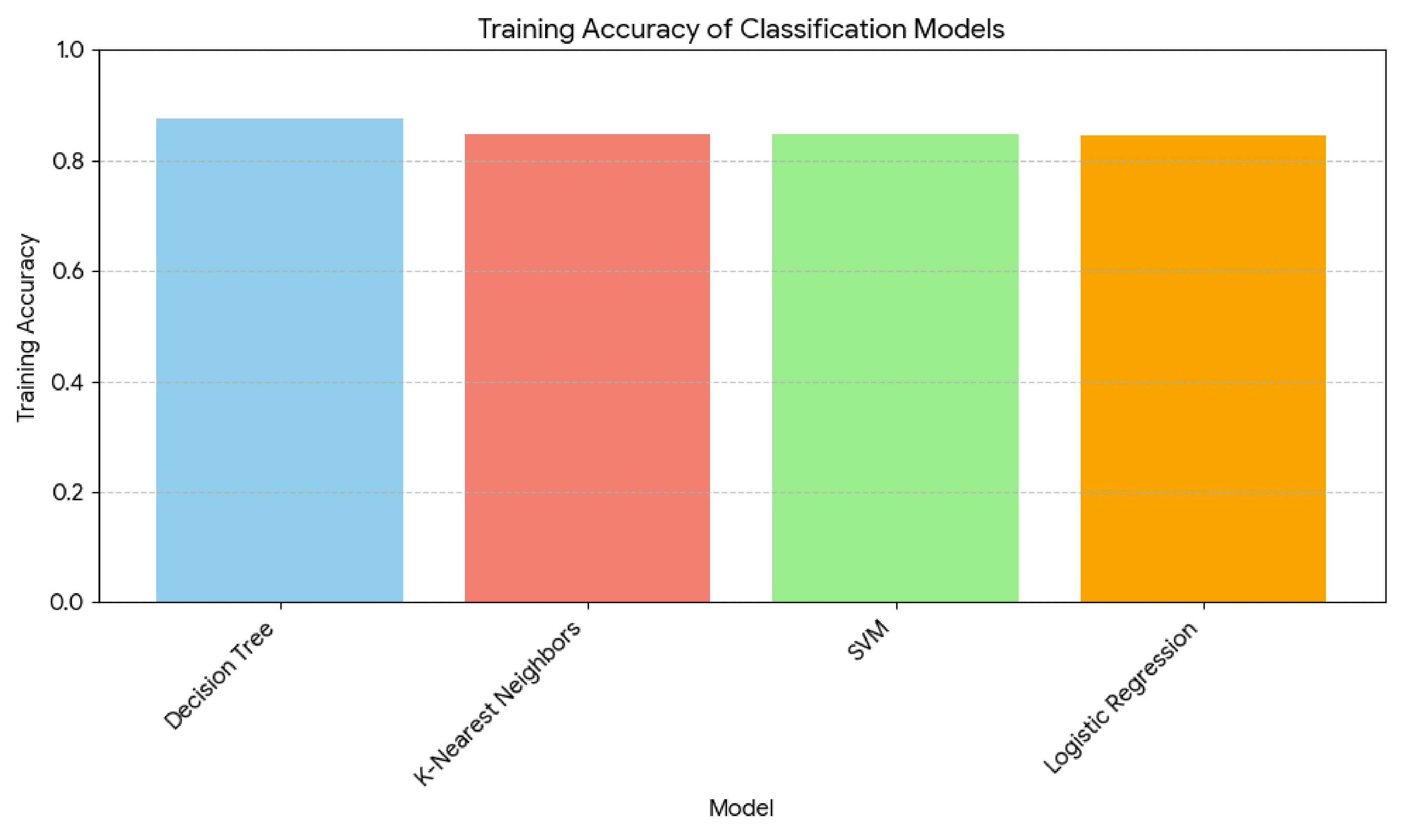
- From all sites low weighted payloads(kg) has relatively high success rates than heavy weighted payloads

Section 5

Predictive Analysis (Classification)

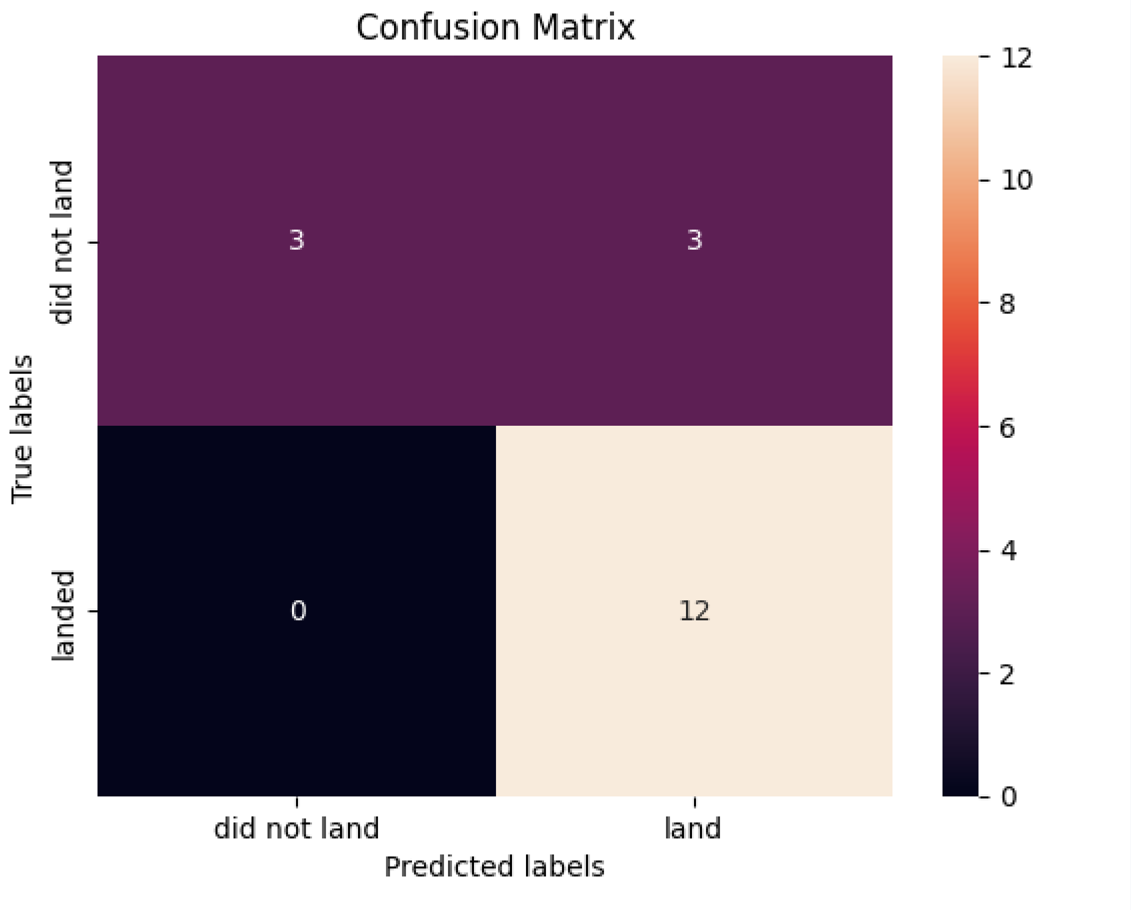
Classification Accuracy

- Based on the analysis, the Decision Tree model has the highest training accuracy at 0.8750



Confusion Matrix

- This matrix summarizes the model's performance on the 18 test outcomes
- **Correct Predictions (Total: 15):**
  - The model correctly predicted **3** instances of "did not land"
  - The model correctly predicted **12** instances of "land"
- **Incorrect Predictions (Total: 3):**
  - The model incorrectly predicted **3** "did not land" outcomes as "land."
  - The model did not make any incorrect predictions of "land" outcomes as "did not land"
- This shows that the model is very good at correctly identifying launches that "land" but has some difficulty with correctly identifying launches that "did not land"



Conclusions

- Low weighted payloads has successfully landed more often than heavy weighted payloads
- KSC LC 39A has most successful launches when compared to other launch sites
- Successful Predictive Modeling: We successfully developed a machine learning model to predict launch success based on various factors
- Decision Tree is the Best Performer: The Decision Tree model proved to be the most accurate in our analysis, achieving a training accuracy of 87.5%. This model is a strong candidate for predicting future launch outcomes
- Future Enhancements: The predictive model can be integrated into the dashboard to provide real-time launch success predictions, further enhancing its value for mission planning and analysis

Thank you!