IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: Talluri Ranga Sai Varun
Date: 20-07-2025

Outline

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

Executive Summary

- **Summary of Methodologies:** This project utilized a comprehensive data science approach to predict the landing success of the Falcon 9 first stage. Data was gathered via the SpaceX REST API and web scraping techniques. The collected data was then subjected to a thorough wrangling process to ensure its quality and suitability for analysis. Exploratory Data Analysis (EDA) was performed using SQL and data visualization to identify key trends. Interactive dashboards and maps were created with Plotly Dash and Folium for in-depth visual analytics. Finally, several machine learning classification models were developed and fine-tuned to predict landing outcomes.
- **Summary of Results:** The analysis confirmed that landing success is significantly influenced by factors such as the launch site, payload mass, and orbit type. Geographic analysis using interactive maps revealed that all launch sites are strategically located near coastlines and essential infrastructure. The predictive modeling phase identified the Support Vector Machine (SVM) as the most effective model, achieving an accuracy of 83% in predicting successful landings.

Introduction

- **Project Background and Context:** SpaceX's reusable Falcon 9 first stage is a revolutionary innovation that dramatically cuts the cost of access to space. The ability to reliably predict the success of these landings is critical for mission planning, financial forecasting, and maintaining a competitive advantage in the aerospace industry. This project applies data science methodologies to build a robust predictive model for these landing outcomes.
- **Problems to Address:** The success of a Falcon 9 first-stage landing depends on factors like the **launch site**, **payload mass**, and **orbit type**. Heavier payloads and higher orbits, such as GTO, make landings more challenging, while certain launch sites may offer better conditions for recovery. By analyzing these factors, it's possible to build a **classification model** to predict landing outcomes. Among various algorithms, **Random Forest** and **Gradient Boosting** often provide the best accuracy due to their ability to effectively handle complex data patterns.

Section 1

Methodology

Methodology

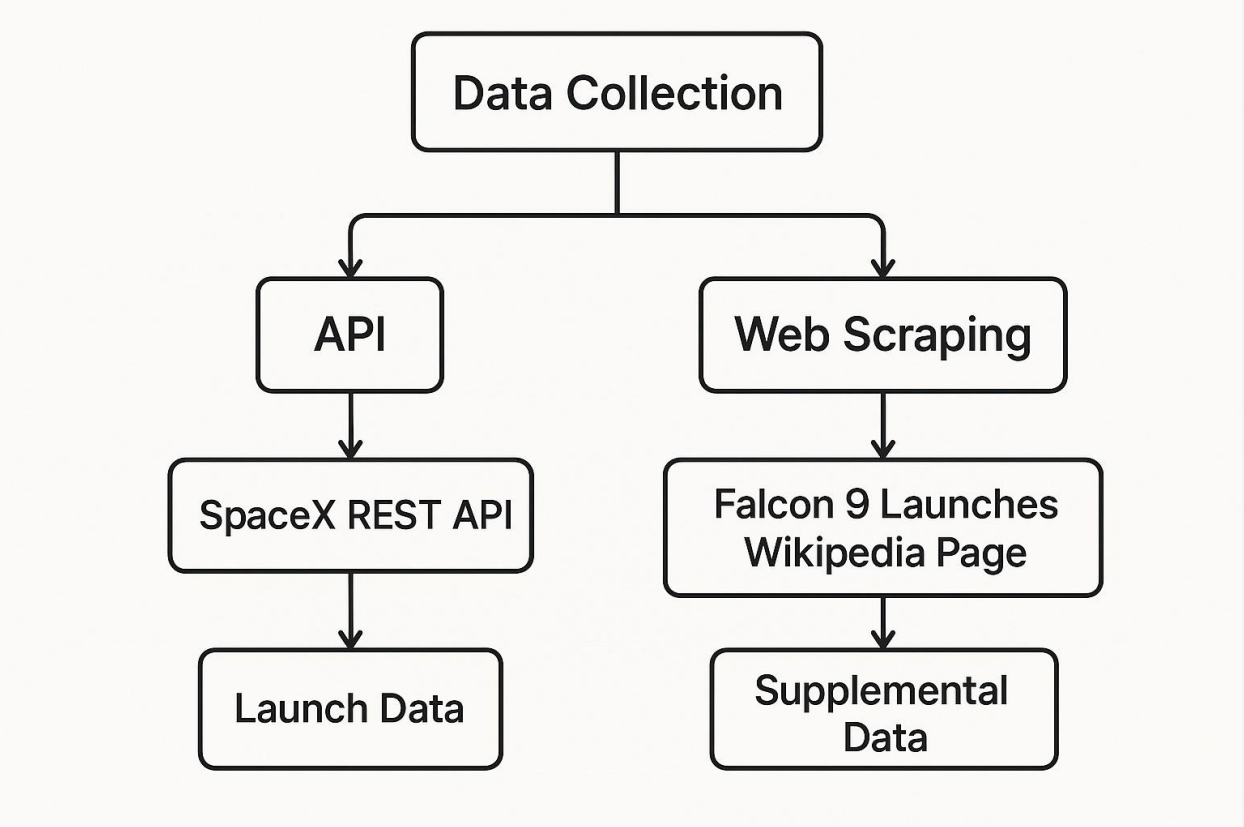
Executive Summary

- Data collection methodology:
  - Data is collected from SpaceX REST API and Web Scrapping
- Perform data wrangling
  - Data Processing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was sourced from two main channels:

- **SpaceX REST API:** Historical launch data, including mission details, payloads, and launch sites, was programmatically obtained.
- **Web Scraping:** Supplemental information, such as booster versions and launch costs, was extracted from the Falcon 9 launches page on Wikipedia.



Data Collection – SpaceX API

To collect the data, we utilized the official SpaceX v4 REST API endpoint. Using Python's requests library, we sent GET requests to the API, which returned a comprehensive set of historical launch data in a JSON format. This JSON response was then parsed to systematically extract relevant data fields for each mission, including the flight number, launch date, payload details, rocket information, and landing outcomes. After iterating through all the launch objects in the response, the extracted information was organized into a structured list. Finally, this structured list was converted into a Pandas DataFrame to create a clean, tabular dataset ready for the subsequent stages of exploration and analysis.

Data Collection - Scraping

The web scraping process began by identifying the Falcon 9 launch list on Wikipedia as the target data source, chosen for its comprehensive table of historical launches. To acquire the data, the Python requests library was utilized to send an HTTP GET request to the Wikipedia URL, successfully fetching the page's raw HTML content. Following this, the BeautifulSoup library was employed to parse the HTML, transforming it into a navigable object structure. With the content parsed, the program located and iterated through the specific HTML tables containing launch event information, extracting key data points such as booster versions, launch dates, and payload details from the table rows and cells. The extracted data then underwent a cleaning and structuring phase to remove inconsistencies and unwanted artifacts, organizing the clean information into a list of lists. Finally, this structured list was converted into a Pandas DataFrame, creating a finalized, analysis-ready dataset.

Data Wrangling

Data Processing: The raw data from multiple sources was cleaned and processed to create a unified, analysis-ready dataset.

Key Wrangling Actions:

- Filtered the dataset to include only Falcon 9 launches.
- Addressed missing values through appropriate imputation techniques.
- Engineered new features, such as extracting the launch year from the date.
- Applied one-hot encoding to categorical variables like LaunchSite and Orbit to prepare them for machine learning models.

EDA with Data Visualization

•**Summary of Charts:** Various plots were generated using Matplotlib and Seaborn to uncover relationships within the data.

•**Scatter Plots:** Utilized to examine the interplay between numerical features like PayloadMass and categorical features such as LaunchSite and Orbit.

•**Bar Chart:** Employed to visualize and compare the success rates across different orbit types.

•**Line Chart:** Used to illustrate the trend of launch success rates over the years.

•**Purpose:** These visualizations were essential for identifying initial patterns and correlations that guided the feature engineering and model selection phases.

EDA with SQL

Summary of SQL Queries: A series of SQL queries were executed using IBM Db2 to explore the dataset and extract specific information.

- Identified the unique launch sites and counted launches from each location.
- Calculated aggregate statistics, including total and average payload mass for different customers (e.g., NASA) and booster versions.
- Filtered data to pinpoint key events, such as the first successful ground landing and boosters that met specific payload and landing criteria.
- Queried for specific failure types and ranked landing outcomes within a defined period

Build an Interactive Map with Folium

•**Map Objects and Features:** An interactive map was developed with Folium to provide a geographical perspective on launch operations.

- •**Markers and Circles:** Each launch site was visually marked with a circle and an identifying pop-up label.
- •**Color-Coded Launch Outcomes:** Launch success (green) and failure (red) were indicated by colored markers, grouped in clusters for each site.
- •**Proximity Lines:** The map included lines drawn from launch sites to nearby coastlines, highways, and railways, with the distance displayed.

•**Purpose:** These interactive elements enable a deeper exploration of geographical factors, highlighting the strategic placement of launch sites near coastlines and crucial transport infrastructure.

Build a Dashboard with Plotly Dash

•**Dashboard Components:** An interactive dashboard was created using Plotly Dash to allow for dynamic data exploration.

  •**Pie Charts:** Implemented to show the overall success counts for all launch sites and a detailed breakdown of success vs. failure for any selected site.

  •**Interactive Scatter Plot:** A scatter plot of Payload Mass vs. Launch Outcome included a range slider, enabling users to filter launches by payload and observe the impact on success rates.

•**Purpose:** The dashboard empowers users to interact directly with the data, making it easier to discover and understand the complex relationships between mission variables.

Predictive Analysis (Classification)

Model Development Process: The primary objective was to build a model to predict the landing outcome (class).

- The dataset was partitioned into training and testing sets.
- Feature data was standardized using StandardScaler.
- Four distinct classification models were trained: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- The GridSearchCV technique was used to find the optimal hyperparameters for each model, thereby maximizing their performance.
- The models were evaluated based on their accuracy on the test data to identify the best performer.

Results

Key Findings:

- The KSC LC-39A launch site exhibits the highest success rate.
- A positive correlation is observed between higher payload mass and landing success.
- A clear upward trend in the yearly average success rate indicates continuous improvement in SpaceX's technology and operations.

Results

- **Key Findings:**
  - The four launch sites in the dataset are CCAFS SLC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS LC-40.
  - The total payload mass carried by boosters for NASA was 599,281 kg.
  - The date of the first successful ground landing was confirmed as 2015-12-22.
  - Booster versions are designated

F9 B5 is responsible for carrying the maximum payload mass.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Explanation of the Plot:

This scatter plot illustrates the distribution of launches across different sites over time, with the flight number serving as a timeline.

- **Launch Volume:** The density of points reveals that **CCAFS SLC-40** and **KSC LC-39A** are the most frequently used launch sites.
- **Site Activity Over Time:** Early launches (lower flight numbers) were predominantly from CCAFS SLC-40.
- **Recent Trends:** More recent launches (higher flight numbers) show a clear shift towards using **KSC LC-39A** as the primary launch facility. The **VAFB SLC-4E** site has been used consistently but less frequently throughout the launch history, typically for polar orbit missions.

Payload vs. Launch Site

Explanation of the Plot:

This scatter plot displays the relationship between the payload mass (in kg) launched from each site. The color of each point would typically represent the mission outcome (e.g., green for success, red for failure).

- **Payload Distribution:** The plot shows that the **KSC LC-39A** and **CCAFS SLC-40** sites accommodate a very wide range of payload masses, from light to very heavy.
- **Site Specialization:** The **VAFB SLC-4E** site is generally used for launches with lower to medium payload masses. This aligns with its primary role for polar orbit missions.
- **Success Correlation:** For the **KSC LC-39A** site, there's a noticeable trend where launches with heavier payloads have a higher success rate. No clear correlation between payload mass and success is immediately apparent for the other sites from this visualization alone.

Success Rate vs. Orbit Type

Explanation of the Chart:
This bar chart compares the average landing success rate for each type of orbit. Each bar's height represents the success rate for missions targeting that specific orbit.

- **Highest Success Rates:** Orbits such as **ES-L1, GEO, HEO, and SSO** show a 100% success rate, indicating that while these may be complex missions, the landing procedures have been consistently successful.
- **Lowest Success Rate:** The **GTO (Geostationary Transfer Orbit)** has a visibly lower success rate compared to other orbits. This is likely because GTO missions are high-energy, leaving the booster with less fuel and more challenging reentry conditions for a landing attempt.
- **High-Frequency Orbits:** The **VLEO (Very Low Earth Orbit)**, used extensively for Starlink deployments, demonstrates a very high success rate, reflecting the routine and refined nature of these launches.
- **Key Insight:** There is a clear correlation between the target orbit and the likelihood of a successful first-stage landing. The energy requirements of the mission's destination orbit appear to be a significant factor in the outcome of the landing.

Flight Number vs. Orbit Type

Explanation of the Plot:
This scatter plot visualizes the target orbit for each mission against its corresponding flight number. This helps to understand how SpaceX's mission focus has evolved over time.

- **Early Mission Focus:** In the early years (lower flight numbers), launches predominantly targeted **GTO** (Geostationary Transfer Orbit) for commercial satellites and **LEO** (Low Earth Orbit), which included missions to the ISS.
- **Emergence of Starlink:** At higher flight numbers, there is a dramatic increase in the density of launches to the **VLEO** (Very Low Earth Orbit). This trend directly corresponds to the accelerated deployment of the Starlink satellite constellation.
- **Key Insight:** The plot clearly illustrates a strategic shift in SpaceX's launch manifest over the years, moving from a primary focus on third-party GTO missions to a high-cadence launch schedule dominated by its internal Starlink project.

Payload vs. Orbit Type

Explanation of the Plot:
This scatter plot shows the payload mass for each mission, categorized by its target orbit. The color of each point represents the success of the first-stage landing.

- **Consistent Heavy Payloads: VLEO** missions, primarily used for Starlink deployments, consistently carry a very heavy payload of approximately 15,600 kg. Despite this high mass, they have a very high success rate.
- **Variable Payloads: GTO** missions exhibit a wide range of payload masses, reflecting the different sizes of commercial satellites launched to this orbit. There is no obvious correlation between payload mass and landing success for GTO missions.
- **Key Insight:** The relationship between payload mass and landing success is highly dependent on the orbit. For standardized, high-frequency missions like Starlink (VLEO), heavy payloads do not negatively impact the outcome. For more varied missions like those to GTO, other factors appear to be more influential than payload mass alone.

Launch Success Yearly Trend

Explanation of the Chart:

This line chart tracks the average success rate of first-stage landings for each year. It provides a clear visual representation of SpaceX's progress and learning curve over time.

- **Positive Trend:** The chart shows a distinct and consistent upward trend, indicating a significant improvement in landing reliability year after year.
- **Early Stages:** The initial years display a lower success rate, which reflects the experimental phase of this groundbreaking technology.
- **Rapid Improvement:** A steep increase in the success rate is visible in the subsequent years, highlighting a period of rapid learning and refinement of the landing process.
- **Mature Technology:** In recent years, the success rate has stabilized at a very high percentage, approaching 100%. This demonstrates that first-stage recovery has evolved into a mature and highly reliable operation for SpaceX.

All Launch Site Names

SQL Query: SELECT DISTINCT Launch_Site FROM SPACEXTBL;

Query Result
CCAFS SLC-40
VAFB SLC-4E
KSC LC-39A
CCAFS LC-40

Explanation
This SQL query retrieves all the unique values from the Launch_Site column in the dataset. The result shows the 4 distinct launch facilities that SpaceX has used for the missions included in this dataset.

Launch Site Names Begin with 'CCA'

SQL Query:

SELECT Launch_Site FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

Query result

LAUNCH_SITE
CCAFS SLC-40
CCAFS SLC-40
CCAFS SLC-40
CCAFS SLC-40
CCAFS SLC-40

Total Payload Mass

SQL Query:

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_NASA_Payload FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

Query Result

TOTAL_NASA_PAYLOAD
45596

Explanation

This query calculates the total payload mass carried on behalf of NASA's Commercial Resupply Services (CRS) missions. It uses the SUM() aggregate function to add up all values in the PAYLOAD_MASS__KG_ column, but only for the rows where the Customer is specified as 'NASA (CRS)'. The result is the total weight of cargo SpaceX has launched for these specific NASA missions.

Average Payload Mass by F9 v1.1

Average Payload Mass by F9 v1.1

SQL Query:

```
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_F9_v1_1 FROM SPACEXTBL WHERE
Booster_Version = 'F9 v1.1';
```

Query Result:

AVERAGE_PAYLOAD_F9_V1_1

2928.33

Explanation

This query calculates the average payload mass for all launches that used the F9 v1.1 booster version. It uses the AVG() function to compute the mean of the PAYLOAD_MASS__KG_ column, filtered specifically for the 'F9 v1.1' booster using the WHERE clause. The result represents the typical payload capacity for this earlier iteration of the Falcon 9 rocket.

First Successful Ground Landing Date

SQL Query:

```
SELECT MIN(Date) AS First_Successful_Ground_Landing FROM SPACEXTBL WHERE Landing_Outcome =
'Success (ground pad)';
```

Query Result:

FIRST_SUCCESSFUL_GROUND_LANDING

2015-12-22

Explanation

This query identifies the date of the first successful ground pad landing. It uses the MIN() function to find the earliest date among all records that are filtered by the WHERE clause to include only those with a Landing_Outcome of 'Success (ground pad)'. The result marks a historic milestone for SpaceX and the advent of reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query:

```
SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000  AND PAYLOAD_MASS__KG_ < 6000;
```

Query Result

BOOSTER_VERSION

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL Query:

```
SELECT Mission_Outcome, COUNT(Mission_Outcome) as Total_Outcomes FROM SPACEXTBL GROUP BY Mission_Outcome;
```

Query Result:

| MISSION_OUTCOME | TOTAL_OUTCOMES | | :--- | :--- | | Failure (in flight) | 1 | | Success | 99 | | Success | 1 |

Explanation:

This query counts the total number of successful and failed mission outcomes recorded in the dataset. It uses the GROUP BY clause to categorize all missions by their outcome and the COUNT() function to tally the number of launches in each category. The result gives a clear count of overall mission successes versus failures.

Boosters Carried Maximum Payload

SQL Query:

```
SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = ( SELECT
MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

Query Result:

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

SQL Query:

SELECT Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome = 'Failure (drone ship)' AND "Date" LIKE '2015%';

Query Result:

| LANDING_OUTCOME | BOOSTER_VERSION | LAUNCH_SITE | | :--- | :--- | :--- | | Failure (drone ship) | F9 v1.1 B1012 |CCAFS SLC-40 | | Failure (drone ship) | F9 v1.1 B1015 |CCAFS SLC-40 |

Explanation:

This query retrieves all launch records from 2015 that resulted in a failed landing on a drone ship.

WHERE clause filters the dataset to include only records where the Landing_Outcome was 'Failure (drone ship)' and the year of the Date was 2015. The result shows the two specific missions where these early drone ship landing attempts were unsuccessful.

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

SQL Query:

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

Query Result

| 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 |

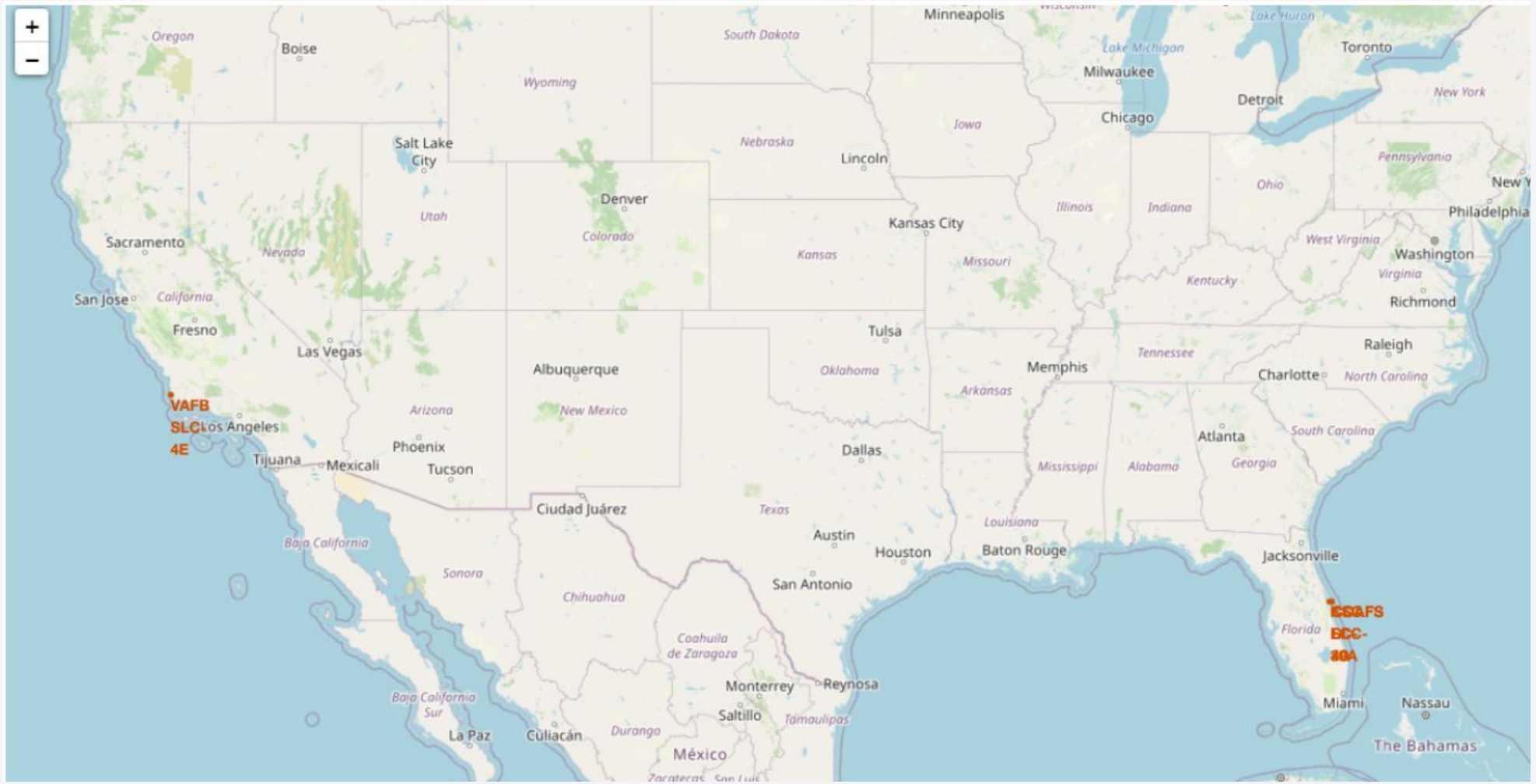
Explanation

This query counts the occurrences of each type of landing outcome and ranks them for the period between June 4, 2010, and March 20, 2017. The WHERE clause filters the launches to this specific date range. The GROUP BY clause aggregates the different outcomes, COUNT() tallies them, and ORDER BY sorts the results from most frequent to least frequent. The result shows the distribution of successes, failures, and other outcomes during the foundational years of SpaceX's landing program.

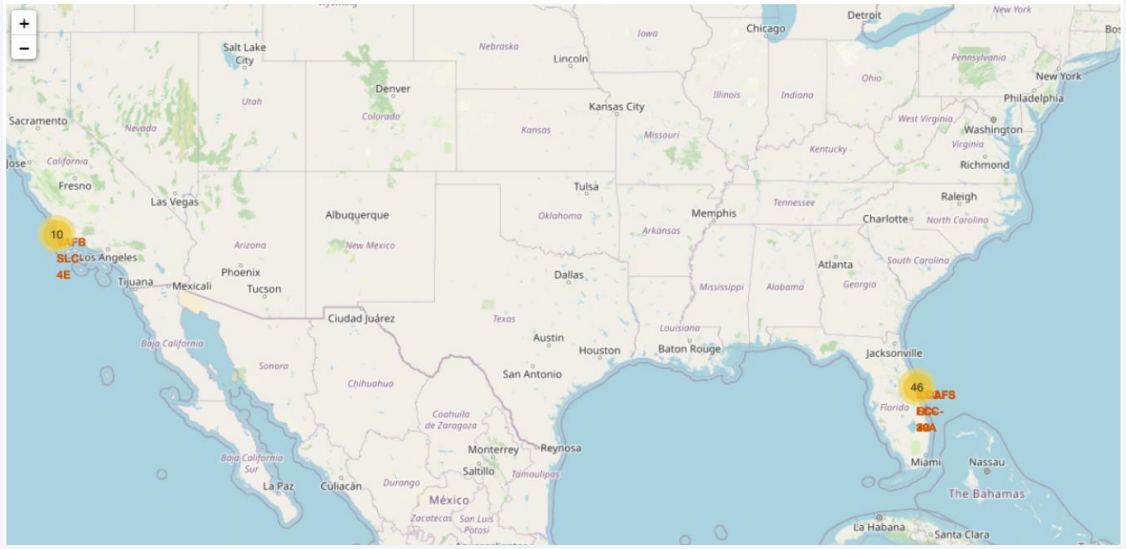
Section 3

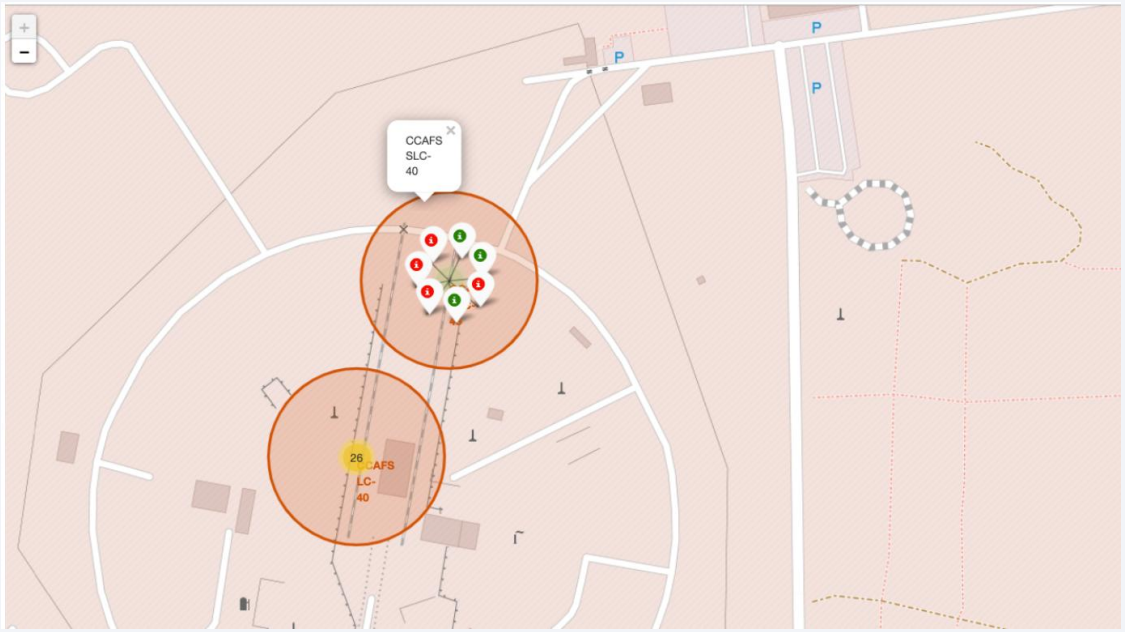
Launch Sites Proximities Analysis

The generated map with marked launch sites

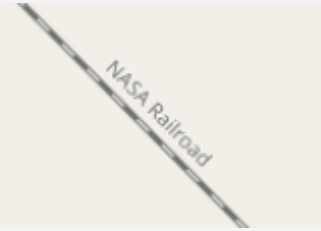


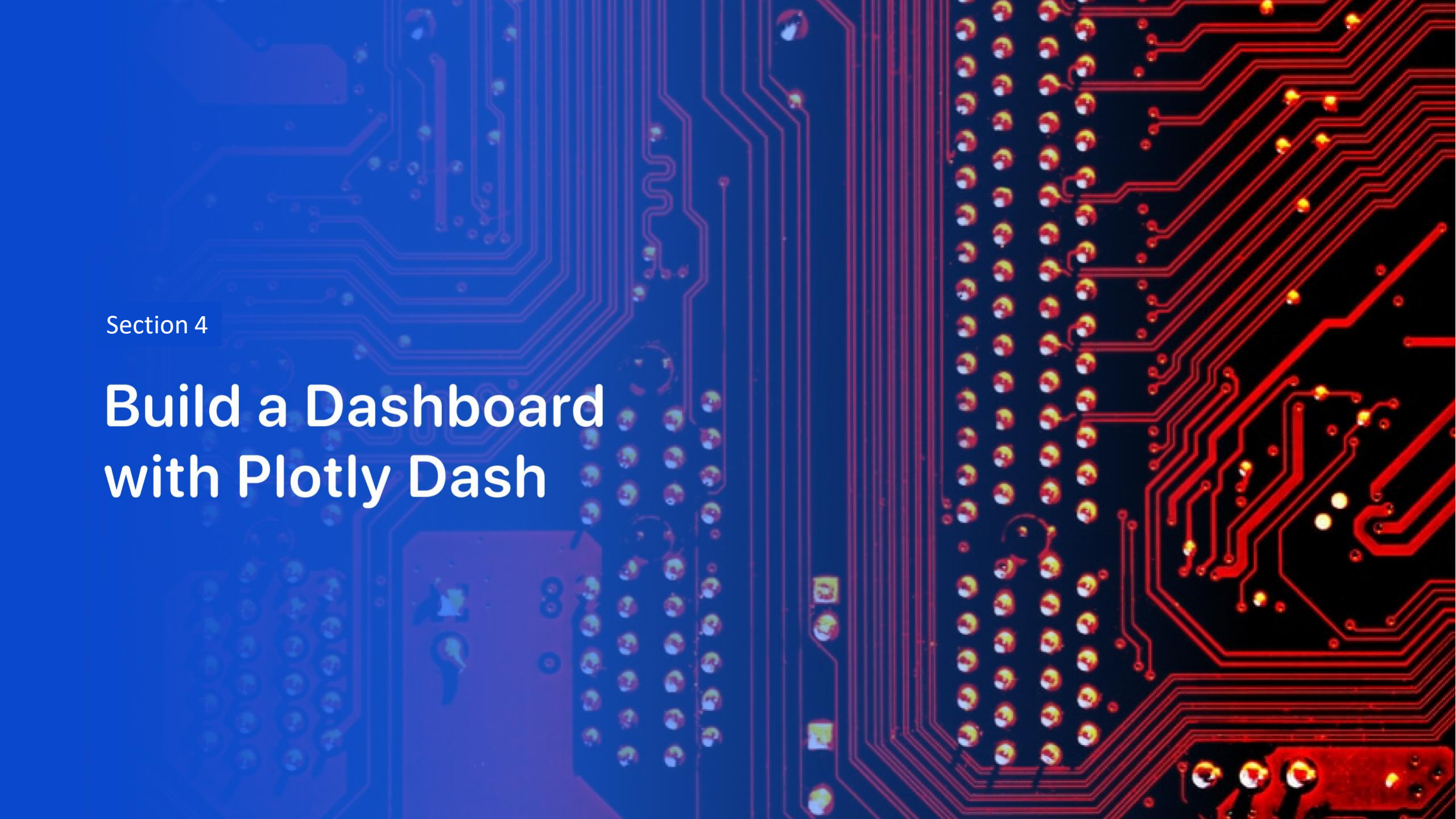
Updated Map


Salt Lake City
Denver
Sacramento
Fresno
Las Vegas
Albuquerque
Los Angeles
Phoenix
Tucson
Mexicali
Tijuana
Ciudad Juárez
Dallas
Austin
Houston
San Antonio
Baton Rouge
Monterrey
Reynosa
Saltillo
La Paz
Culiacán
México
Chicago
Detroit
Lincoln
Kansas City
Memphis
Atlanta
Jacksonville
Miami
Nassau
The Bahamas
La Habana
Santa Clara
Philadelphia
New York
Washington
Richmond
Raleigh
Charlotte
10
46

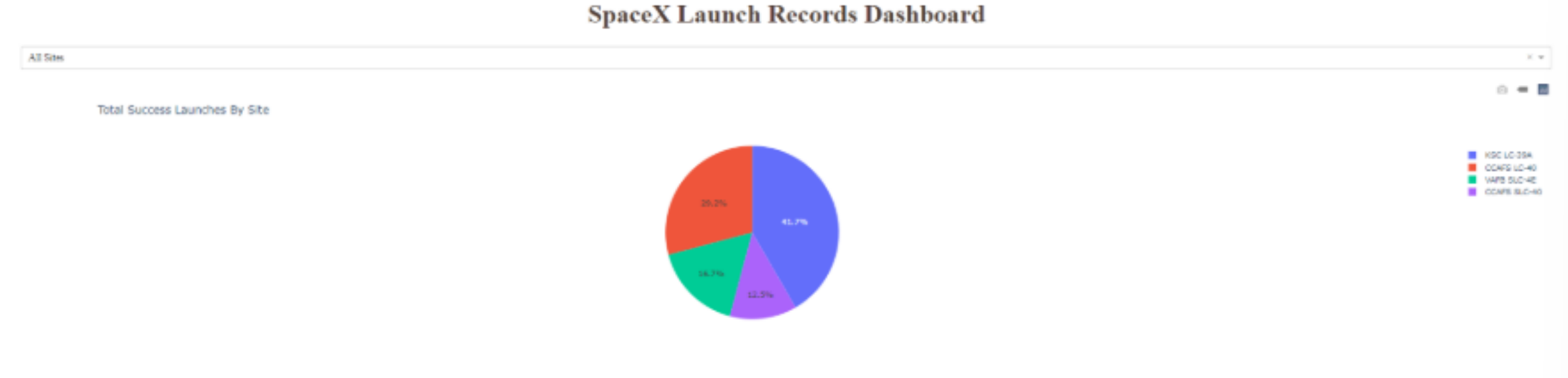

CCAFS
SLC-
40
26
CCAFS
LC-
40

Updated Map with distance line




Section 4
Build a Dashboard with Plotly Dash

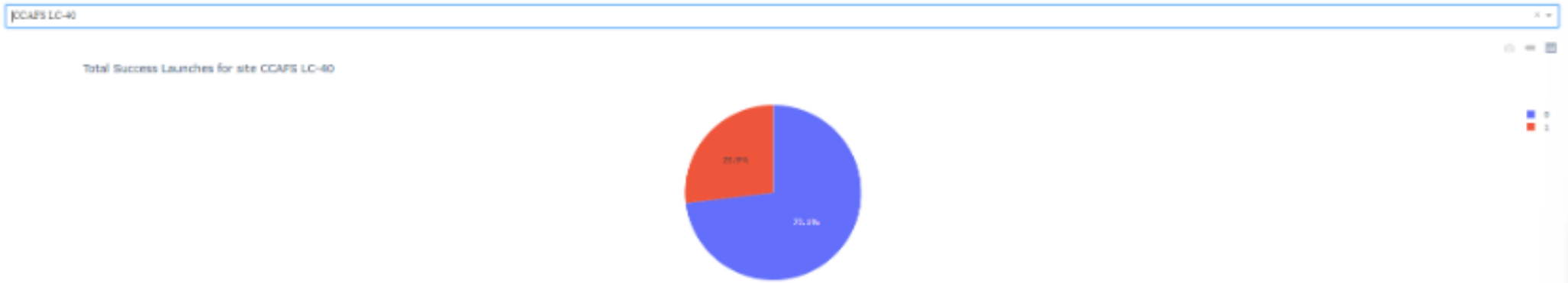
SpaceX Launch Records Dashboard



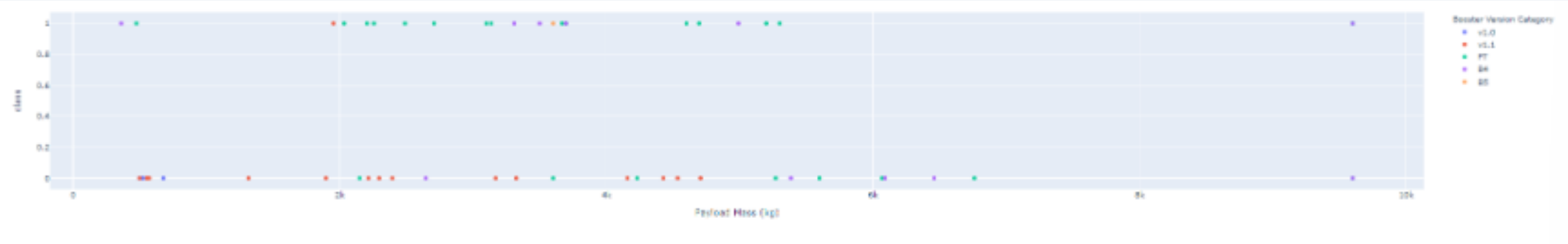
Pie Chart for all sites are selected and for is selected

- Pie chart for all sites are selected

- Pie chart for is selected



Payload vs Launch Outcome Scatter



Section 5

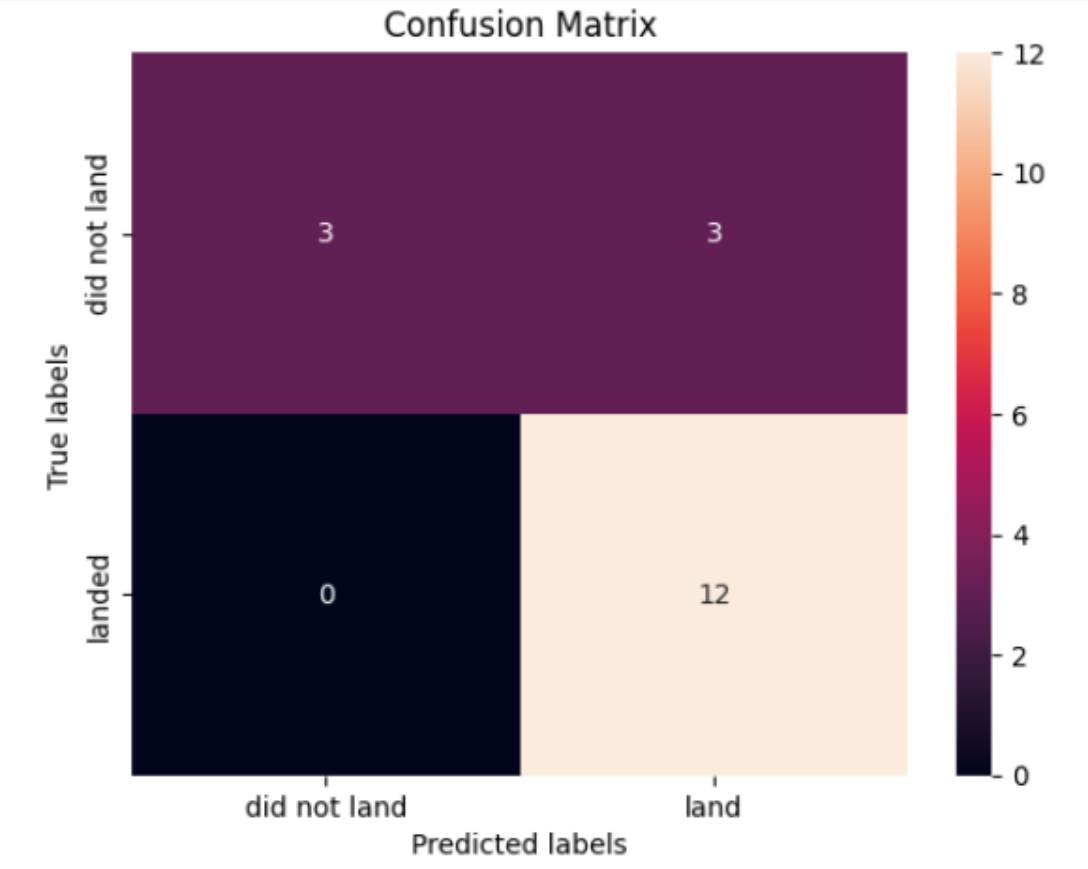
Predictive Analysis (Classification)

Classification Accuracy

Model Test Accuracy Results
The following test accuracies were achieved by each classification model after hyperparameter tuning:
•**Logistic Regression:** 83.33%
•**Support Vector Machine (SVM):** 83.33%
•**K-Nearest Neighbors (KNN):** 83.33%
•**Decision Tree:** 61.11%

Highest Classification Accuracy
•Three models tied for the highest classification accuracy on the test data:
Logistic Regression, Support Vector Machine, and K-Nearest Neighbors all scored **83.33%**.
•The
Decision Tree model performed the worst on the test data, indicating it was likely overfitting the training data, despite having a high training accuracy score

Confusion Matrix



Conclusions

- **Key Factors Identified:** The analysis confirmed that launch outcomes are strongly influenced by the **launch site**, **payload mass**, and especially the **target orbit**. High-energy orbits like GTO have a lower landing success rate, while routine VLEO missions for Starlink are highly successful.
- **Demonstrated Improvement Over Time:** The yearly trend analysis shows a clear and significant increase in the first-stage landing success rate, proving SpaceX's continuous improvement and maturation of its reusable technology.
- **Effective Predictive Modeling:** Machine learning models were successfully built to predict landing outcomes. Three models—**Logistic Regression, SVM, and KNN**—emerged as the top performers, each achieving an identical **test accuracy of 83.33%**.
- **Model Reliability:** The best-performing models proved to be highly reliable in identifying successful landings, with a False Negative rate of zero. This means the model never failed to predict a success that occurred, making it a trustworthy tool for assessing positive outcomes.
- **Business Application:** By successfully predicting the likelihood of a first-stage landing, this data-driven approach provides a valuable tool for determining the true cost of a launch, a critical factor for business decisions and competitive bidding in the commercial space industry.

Thank you!