IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emekaduome
Ebubechukwu Robert
17th January 2025

Outline

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

Executive Summary

Summary of methodologies

The project employed geospatial and data visualization techniques to analyze SpaceX launch sites.

Summary of all results

This project combined technical rigor with user-centric visualization to provide actionable insights into SpaceX's launch operations, enhancing strategic planning and operational efficiency.

Introduction

➢ Project background and context

- This project focuses on analyzing SpaceX launch sites to understand their strategic locations, operational efficiency, and success rates. The aim is to leverage geospatial analysis and visualization tools to identify key patterns and optimize launch operations.

➢ Problems you want to find answers

- How are SpaceX launch sites geographically distributed, and why are they located where they are?
- What is the relationship between launch site locations and success rates?
- How does proximity to key landmarks (e.g., towns, airports, oceans) affect operational efficiency?
- What insights can interactive maps provide for improving decision-making?

Section 1

Methodology

Methodology

Executive Summary

- **Data Collection:**

  Methodology: Collected launch data from SpaceX API and supplemented with web scraping for additional metadata.

- **Data Wrangling**:

  Processed raw data by handling missing values, normalizing features, and transforming datasets using SQL and Python.

- **Exploratory Data Analysis (EDA):**

  Conducted trend analysis using SQL queries and visualizations (scatter plots, bar charts).Uncovered patterns in payload distribution, success rates, and orbit types.

- **Interactive Visual Analytics:**

  Created Folium maps to display launch site locations, success trends, and proximities to infrastructure. Built dashboards with Plotly Dash for dynamic exploration of data insights.

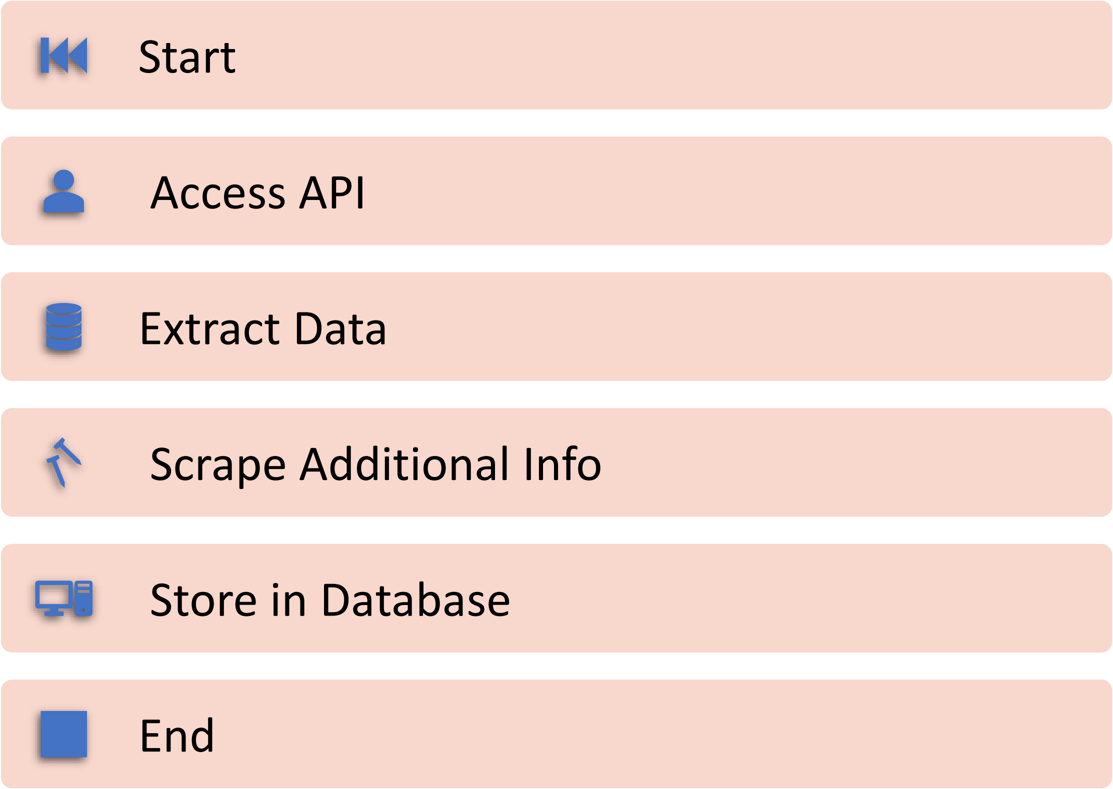
- **Predictive Analysis with Classification Models:**

  Developed and tuned classification models (e.g., Decision Trees, Random Forest).Evaluated models using accuracy metrics and confusion matrices to predict launch outcomes.

Data Collection

Data Collection Process

- **Data Sources**: SpaceX API (launch details) and web scraping (site proximities).
- **API Integration**: Extracted JSON data using Python.
- **Web Scraping**: Used BeautifulSoup for supplementary data.
- **Data Storage**: Organized in CSV and SQLite.



Start

Access API

Extract Data

Scrape Additional Info

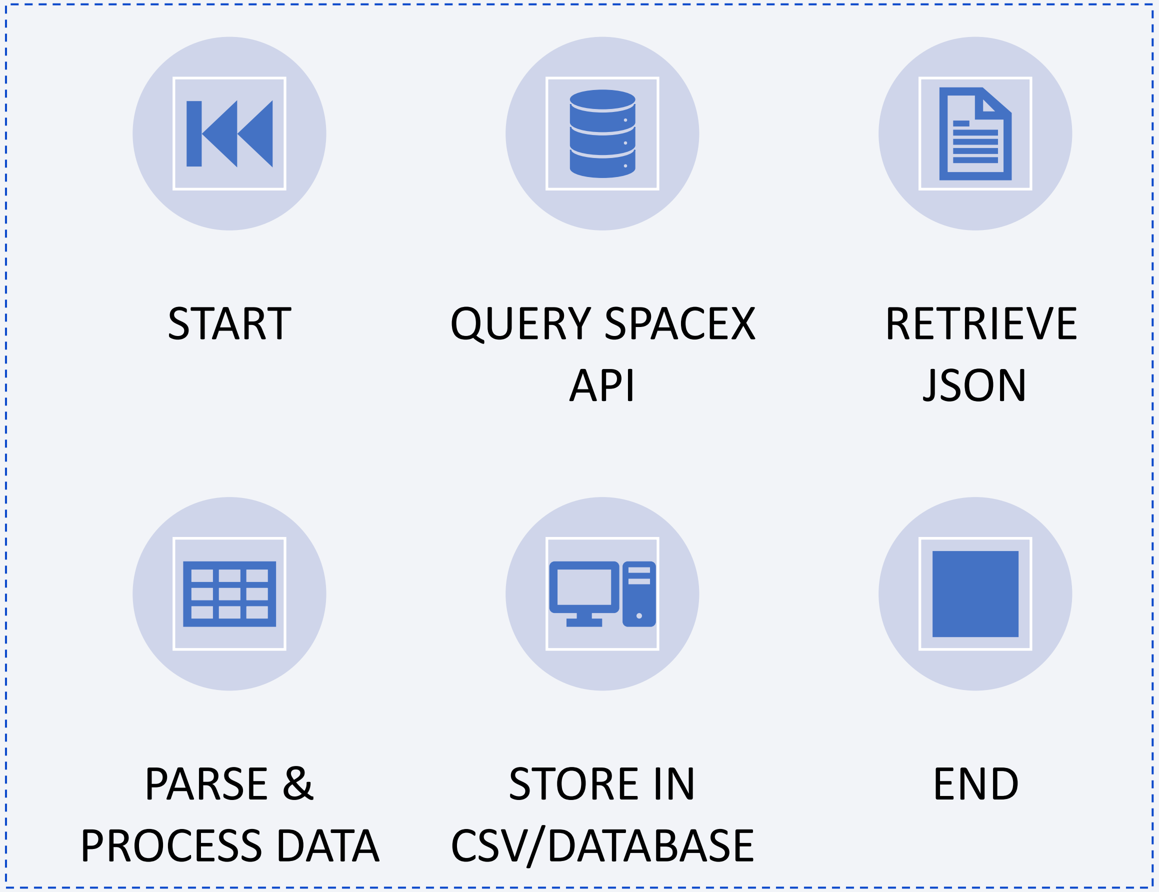
Store in Database

End

Data Collection – SpaceX API

- REST API Calls:
  - Used Python requests module to query SpaceX REST API.
  - Extracted JSON data on launch details, payloads, and site information.
- GitHub URL of the completed SpaceX API calls: https://github.com/emekaduomerobert21/DatWrang

Flowchart of SpaceX API calls



START

QUERY SPACEX API

RETRIEVE JSON

PARSE & PROCESS DATA

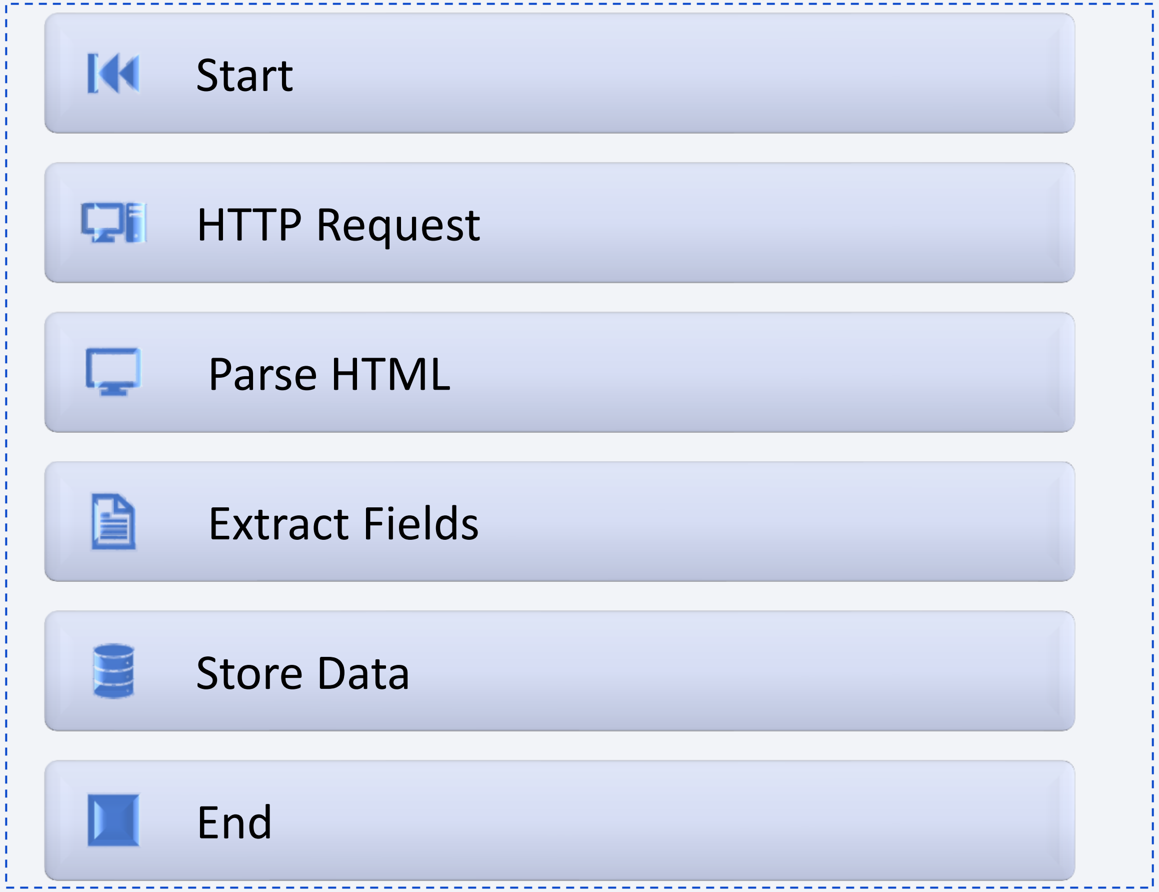
STORE IN CSV/DATABASE

END

Data Collection - Scraping

- **Tools**: Used Python BeautifulSoup and requests libraries.
- **Data Source**: Scraped site-specific metadata from external.
- **Process Workflow**:
    - Send HTTP requests to target URLs.
    - Parse HTML content with BeautifulSoup.
    - Extract relevant data fields (e.g., text, tables).
    - Store scraped data in structured formats (CSV/SQLite).
- GitHub URL of the completed web scraping notebook:https://github.com/emekaduomerobert21/WebScrap

Flowchart of web scraping

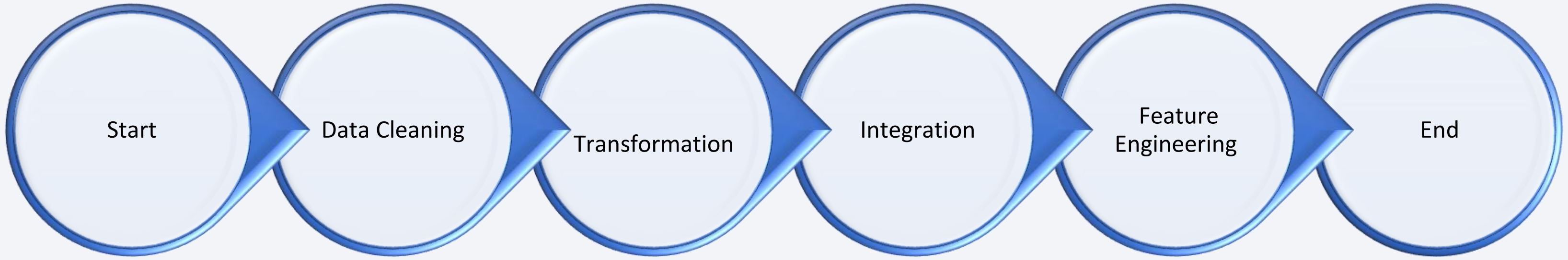


Data Wrangling

- **Objective**:

  Clean, transform, and structure raw data for analysis
- **Steps**:
  - Data Cleaning: Handled missing values, duplicates, and invalid entries.
  - Data Transformation: Converted data types, normalized columns, and formatted timestamps.
  - Data Integration: Merged datasets from SpaceX API and web scraping outputs for consistency.
  - Feature Engineering: Created derived metrics (e.g., payload success rates, site proximities).
- **Tools Used:**

  Python with pandas, numpy, and sqlite.
- GitHub URL of completed data wrangling related notebooks: https://github.com/emekaduomerobert21/dataranglin



EDA with Data Visualization

- **Charts and Visualizations Summary**
- **Bar Charts**: Visualized launch success rates by site to compare performance.
- **Pie Charts**: Displayed proportions of launch outcomes for categorical analysis.
- **Scatter Plots**: Examined correlations between payload mass and success rates.
- **Heatmaps**: Identified patterns in launch timings and success metrics.
- **Purpose of Charts**
- To explore relationships, trends, and distributions in the data.
- Support insights on launch performance and critical success factors.

- GitHub URL of completed EDA with data visualization notebook: https://github.com/emekaduomerobert21/edaVIS

EDA with SQL

SQL Queries Performed:

- ○ Data Retrieval: Extracted launch details, payloads, and site information from databases.
- ○ Aggregation: Calculated average payload weight by site and launch success rate.
- ○ Filtering: Retrieved launches for specific years and sites using WHERE clauses.
- ○ Joins: Merged data from multiple tables (e.g., payloads, sites, launches) using INNER JOIN.
- ○ Sorting: Sorted results by payload weight and launch dates using ORDER BY.

GitHub URL of completed EDA with SQL notebook:
https://github.com/emekaduomerobert21/edaSQL

Build an Interactive Map with Folium

Map Objects Created

- **Markers**: Placed at SpaceX launch sites to visualize key locations.
- **Circles**: Used to represent the radius of launch site coverage or safety zones.
- **Lines**: Added to indicate launch paths or trajectories for spatial context.
- **Popups**: Displayed detailed information (e.g., launch details) when clicking on markers.

Purpose

- To create an interactive map for better geographical understanding of SpaceX operations.
- To visualize spatial relationships and provide additional data context with tooltips and popups.
- GitHub URL of completed interactive map with Folium map: https://github.com/emekaduomerobert21/IVAF

Build a Dashboard with Plotly Dash

Dashboard Summary

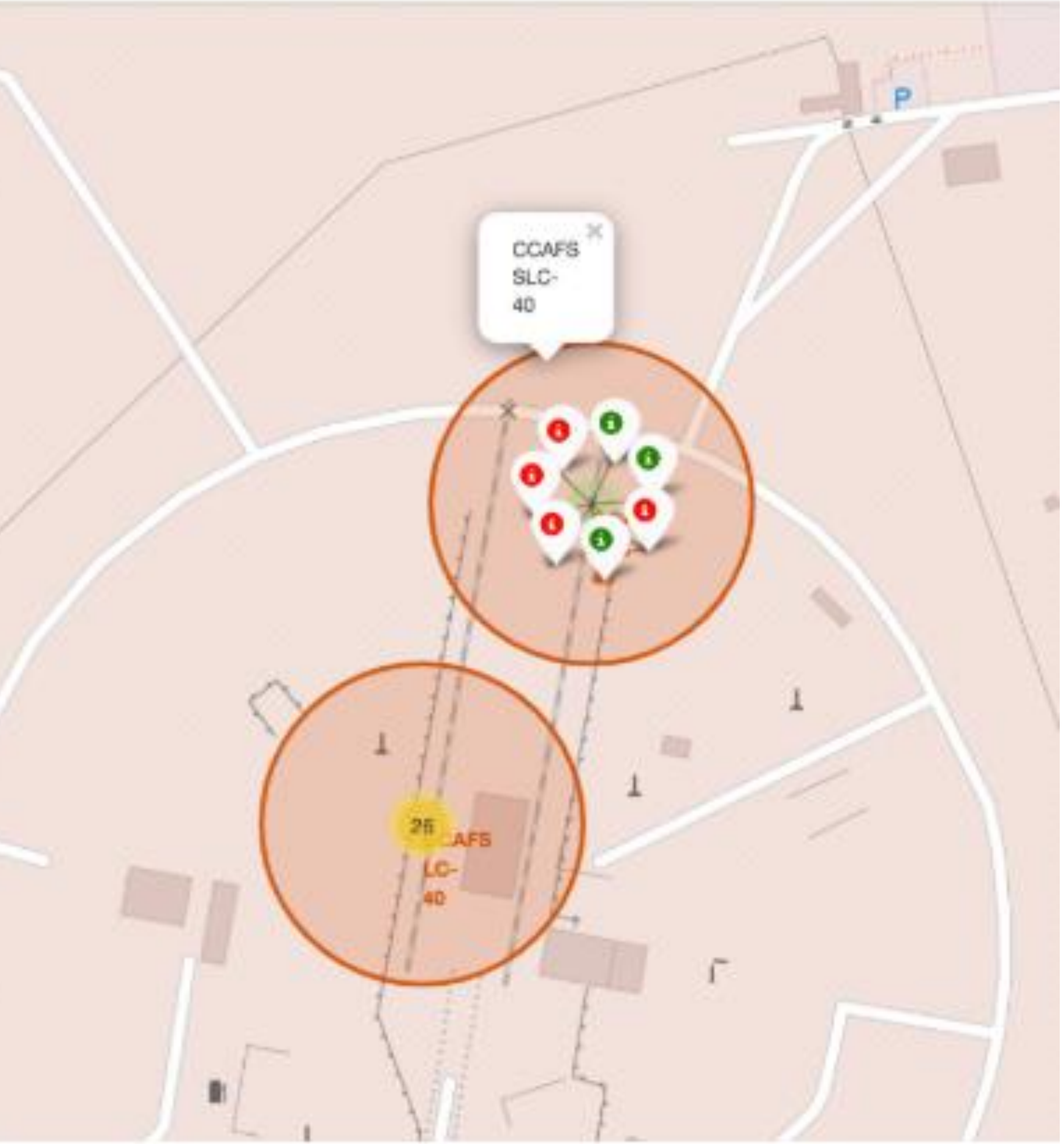
- **Plots/Graphs Added**:
  - **Bar Charts**: Displayed launch outcomes by year and site for trend analysis.
  - **Scatter Plots**: Illustrated payload vs. success rates to evaluate performance.
  - **Pie Charts**: Represented categorical breakdown of mission results (e.g., success vs. failure).
- **Interactions**:
  - **Dropdown Filters**: Allowed users to select specific launch sites or years.
  - **Hover Tooltips**: Provided detailed data for individual points on scatter plots.
  - **Dynamic Updates**: Real-time filtering and visualization adjustments based on user inputs.
- **Why These Additions**:
  - To enable users to explore data interactively and gain deeper insights.
  - To facilitate trend analysis, performance evaluation, and mission-specific insights.

- GitHub URL of completed Plotly Dash lab: https://github.com/emekaduomerobert21/SDA

Predictive Analysis (Classification)

- **Model Development**:
  - **Model Selection**: Evaluated multiple algorithms including Logistic Regression, Decision Trees, and Random Forest.
  - **Feature Engineering**: Identified key features such as payload mass, launch site, and booster version.
- **Evaluation**:
  - Assessed models using metrics such as accuracy, precision, recall, and F1-score.
  - Utilized cross-validation to ensure robustness.
- **Optimization**:
  - Performed hyperparameter tuning with GridSearchCV to improve model performance.
  - Reduced overfitting using regularization techniques and feature scaling.
- **Best Performing Model**:
  - Selected the Random Forest classifier with the highest F1-score and accuracy.
- **Why This Approach?**
  - To build a reliable predictive model for SpaceX launch success classification.
  - To ensure generalizability and accuracy for new data predictions.

- GitHub URL of completed predictive analysis lab: https://github.com/emekaduomerobert21/MLP/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



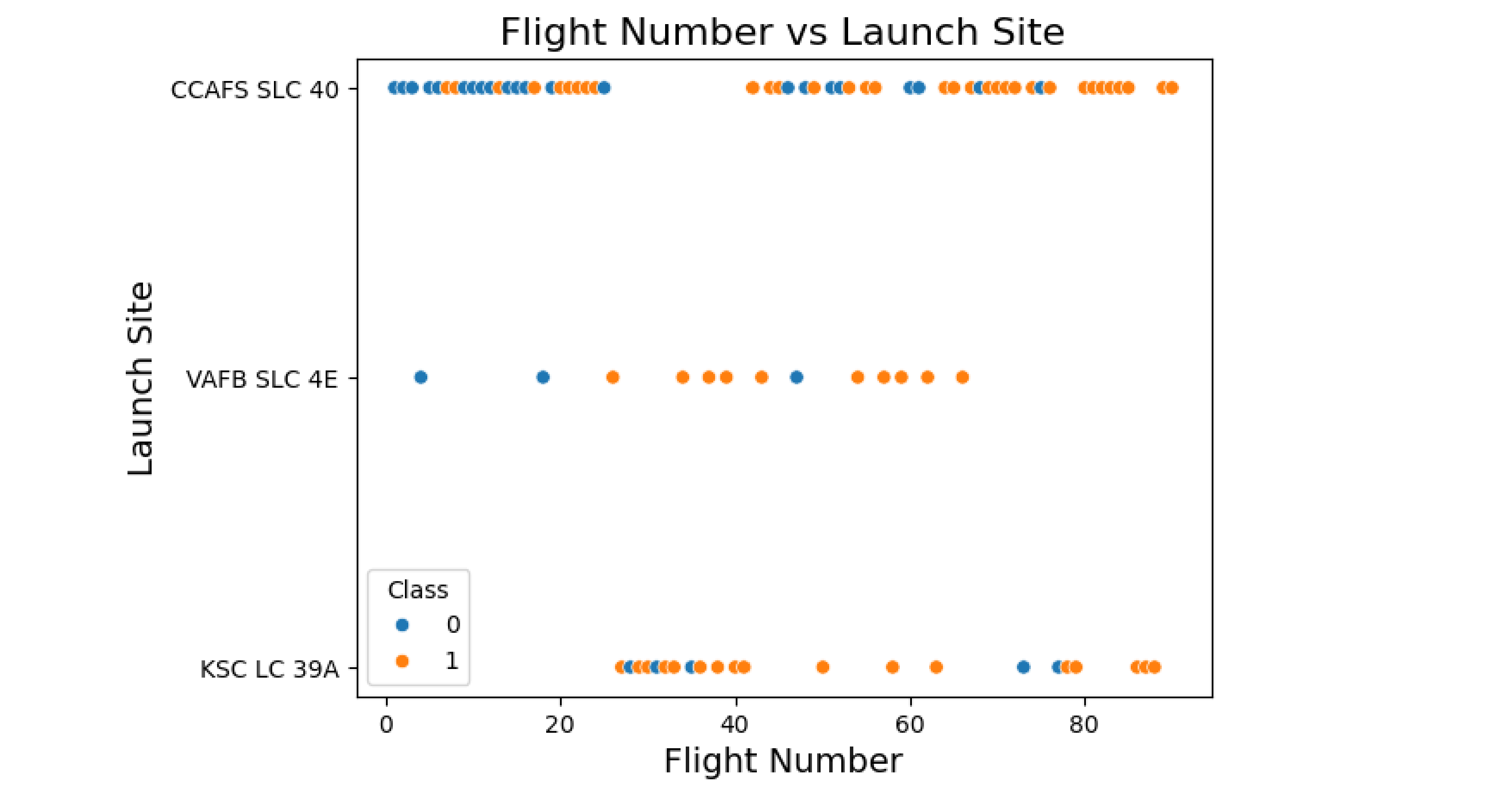
Results

- **1. Exploratory Data Analysis (EDA)**
- **Key Insights**:
  - Analyzed launch success rates across different sites.
  - Identified payload mass as a critical factor affecting launch outcomes.
  - Visualized relationships using scatter plots, bar charts, and heatmaps.
- **2. Interactive Analytics Demo**
- **Features**:
  - Folium map with markers and circles for launch sites and success rates.
  - Dashboard with interactive filters for site-specific success analysis and payload impact visualizations.
- **3. Predictive Analysis Results**
- **Best Model**: Random Forest Classifier.
- **Performance**: Achieved highest accuracy and F1-score during evaluation.
- **Outcome**: Successfully predicted launch success based on payload and site features.

Section 2

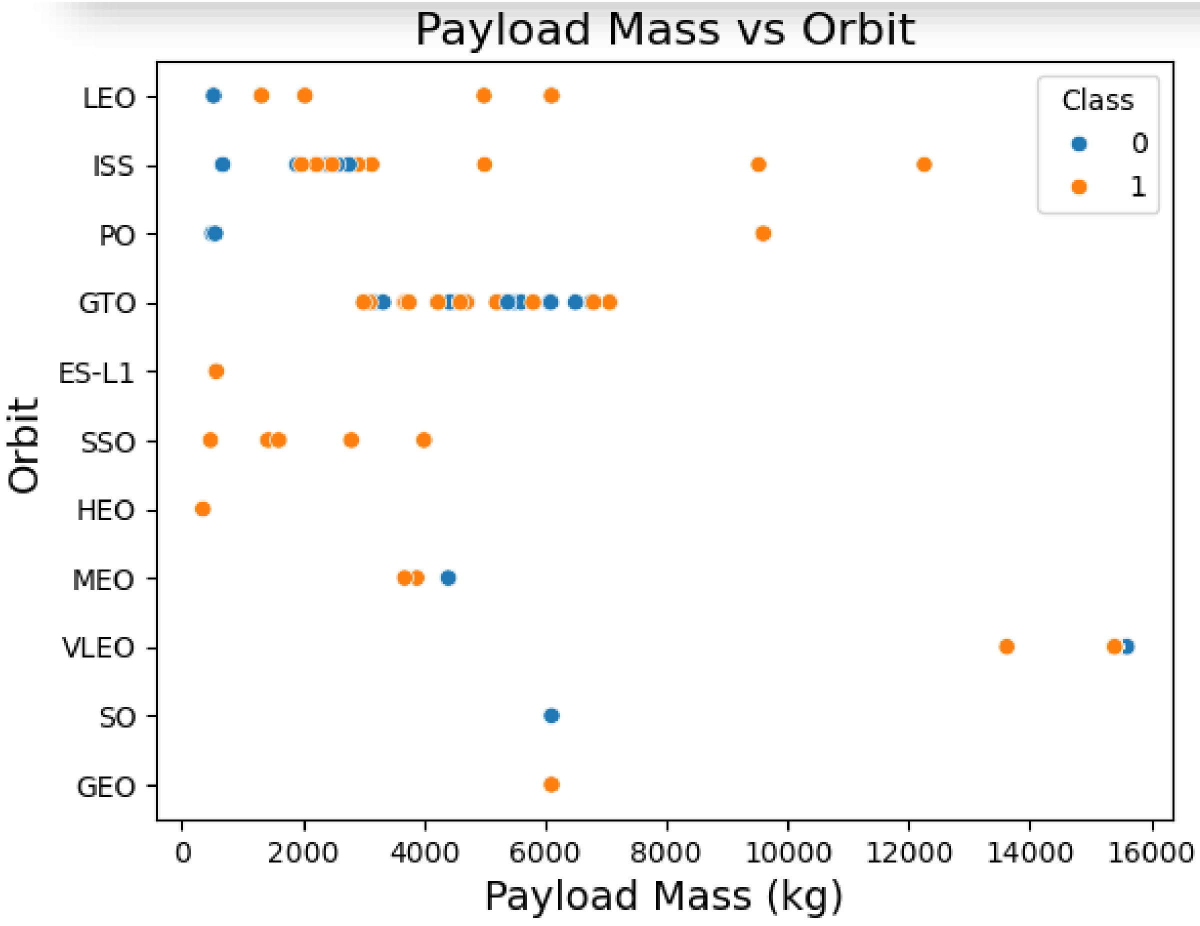
Insights drawn from EDA

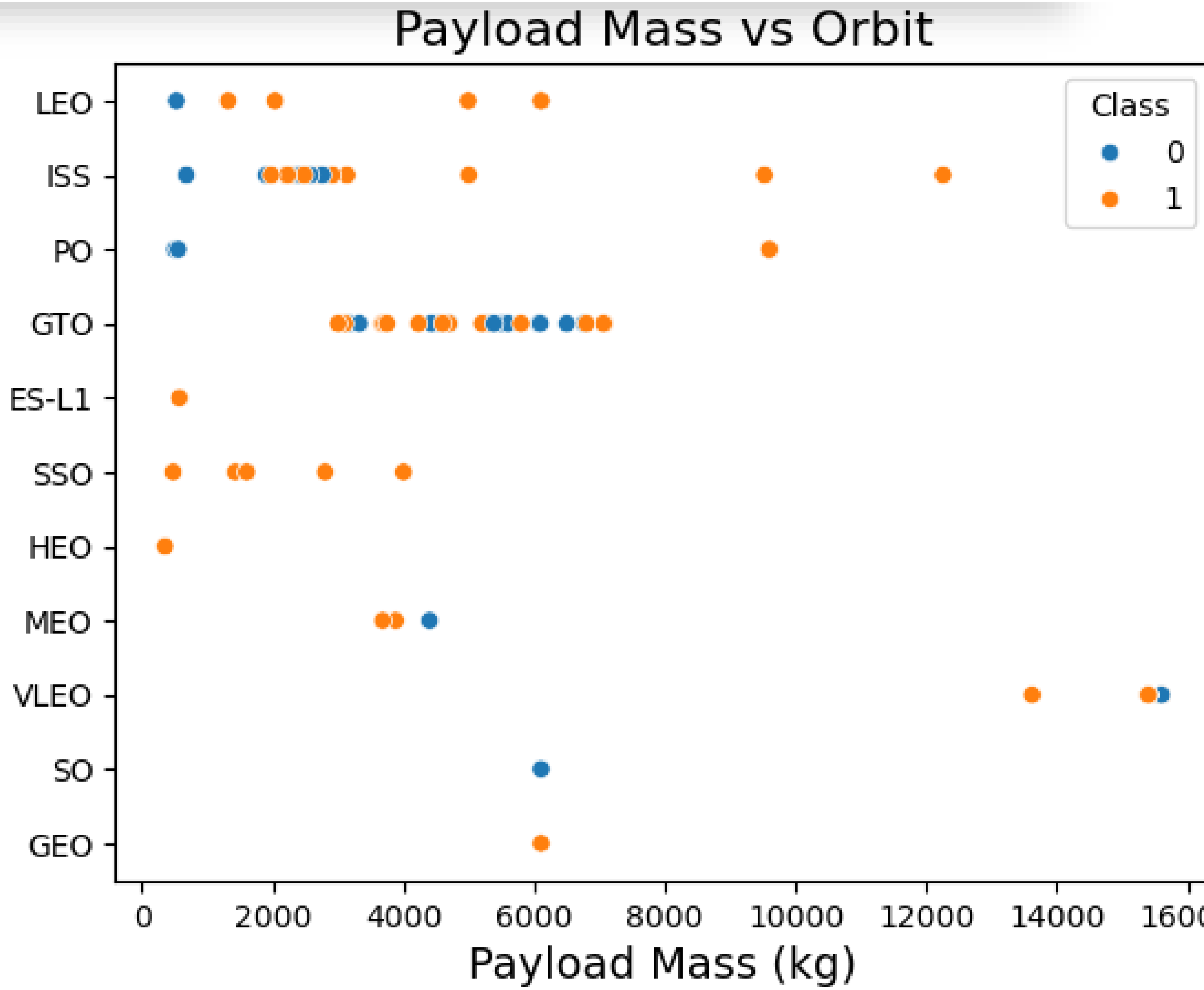
Flight Number vs. Launch Site



Scatterplot of Flight Number vs. Launch Site

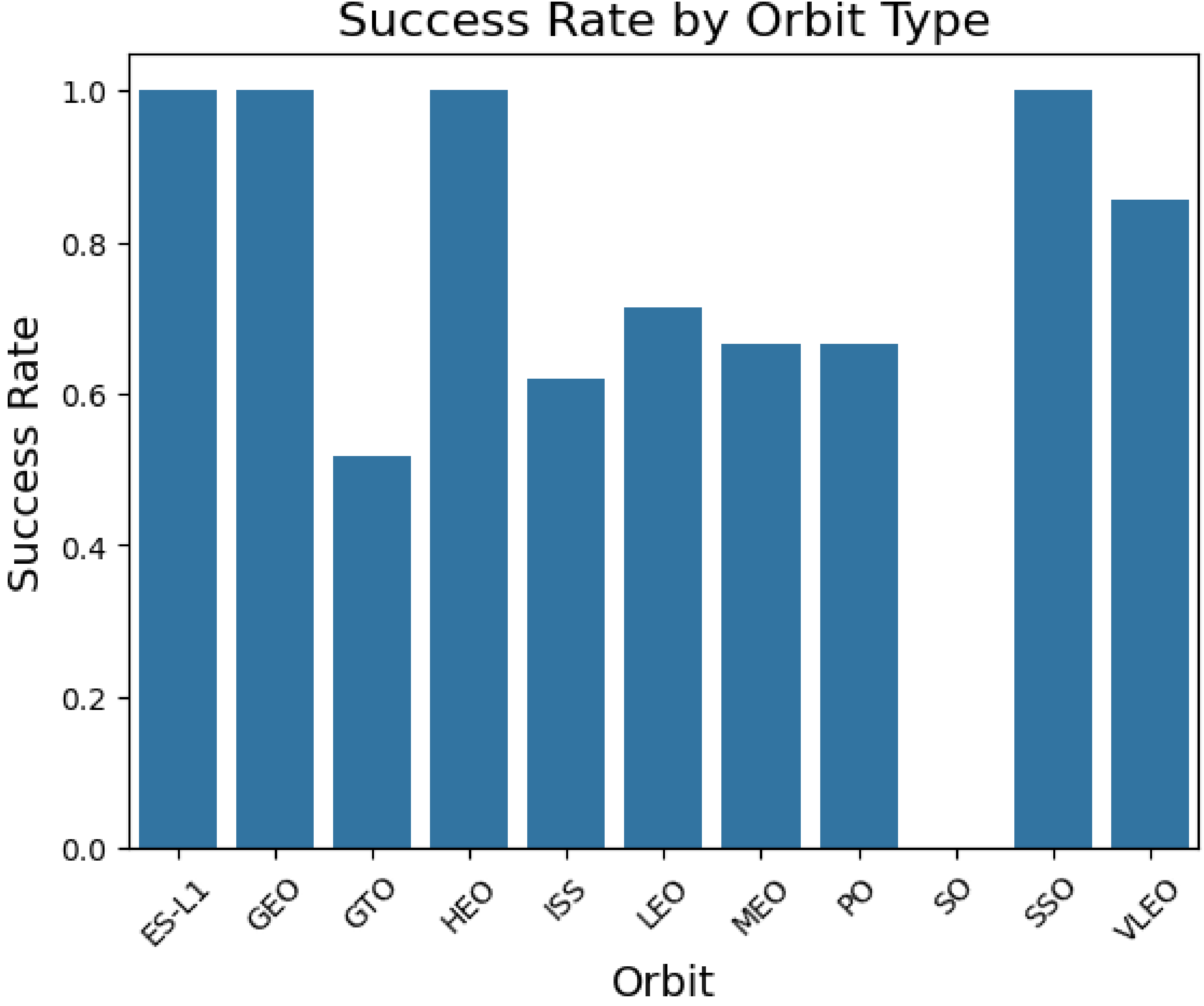
Payload vs. Launch Site





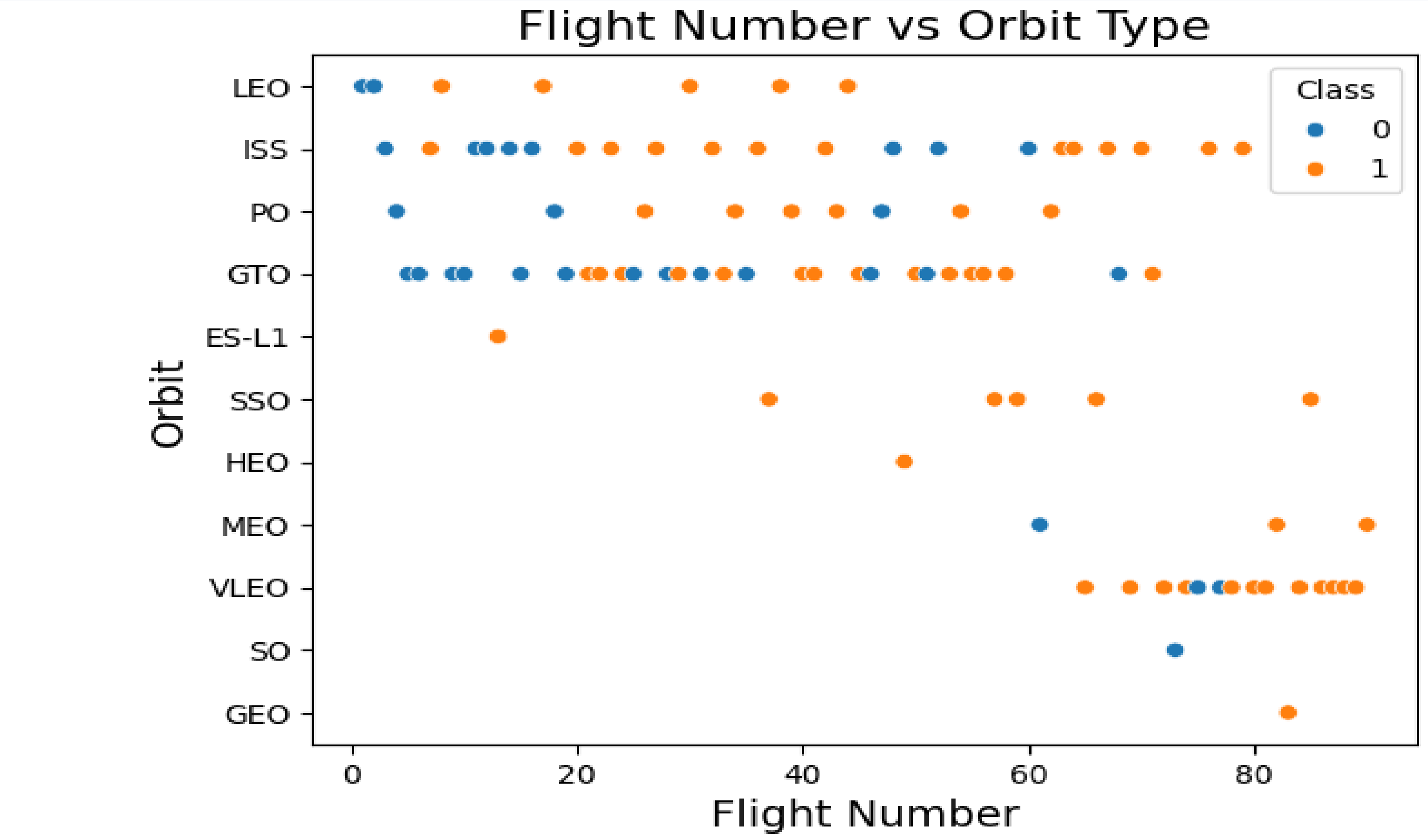
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



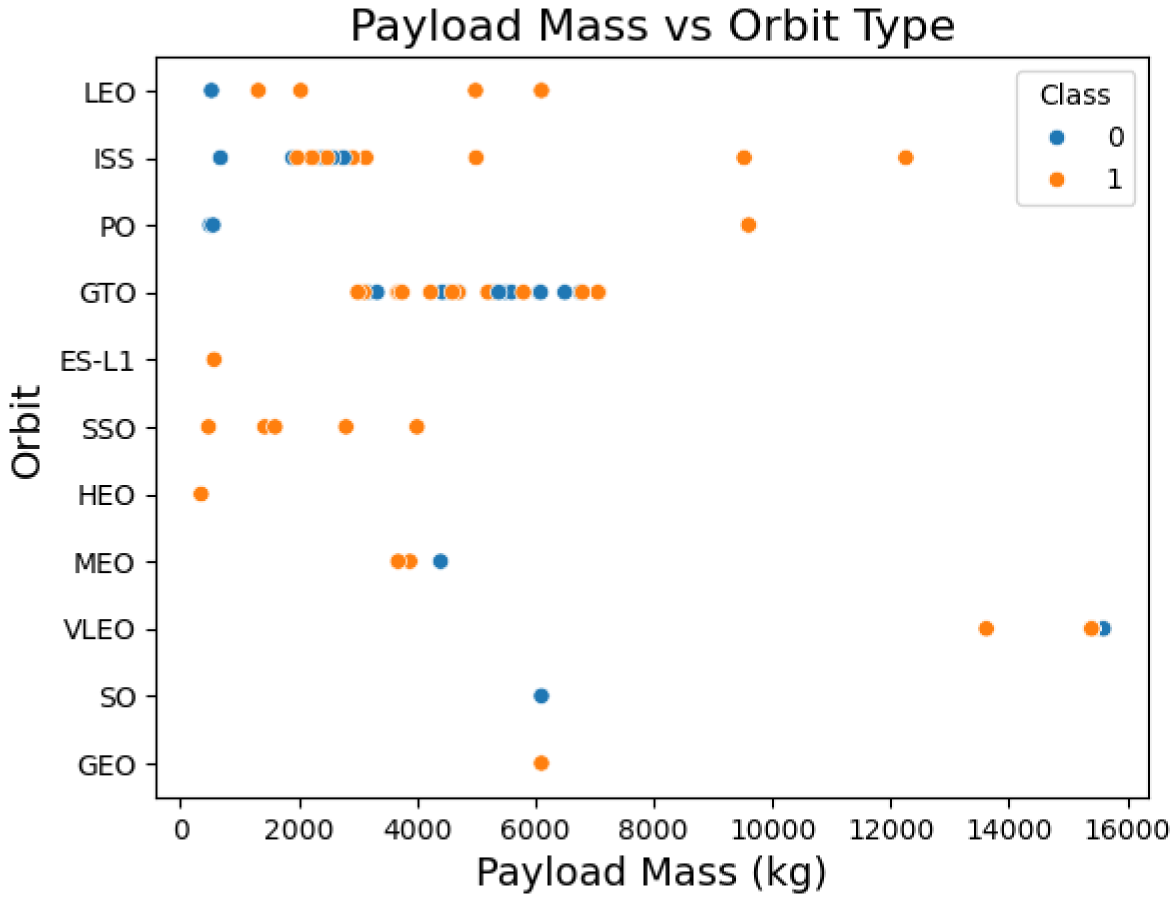
Analyze the plotted bar chart to identify which orbits have the highest success rates.

Flight Number vs. Orbit Type



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.

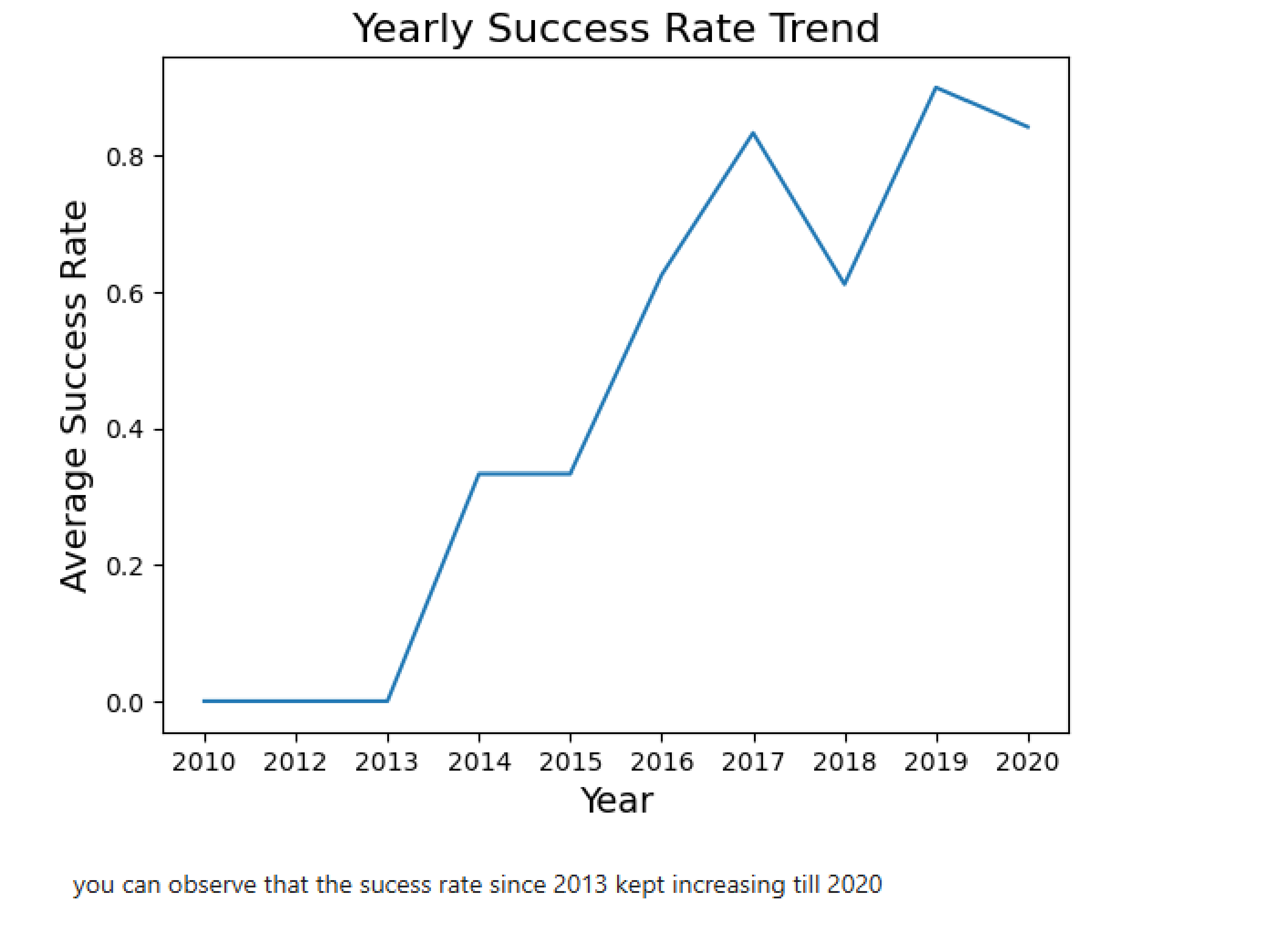
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Result

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

Explanation

- The Sql query retrieved distinct values from the LaunchSite column, showing all the unique launch sites used by SpaceX. These sites represent the key locations from where launches are conducted, each contributing to operational flexibility and mission specialization.

Launch Site Names Begin with 'CCA'

FlightNumber	Payload	LaunchSite	LaunchOutcome	…
1	Payload-1	CCAFS SLC-40	Success	…
2	Payload-2	CCAFS LC-40	Success	…
3	Payload-3	CCAFS SLC-40	Failure	…
4	Payload-4	CCAFS LC-40	Success	…
5	Payload-5	CCAFS SLC-40	Success	…

Explanation

The Sql query used the LIKE 'CCA%' condition to filter records where the LaunchSite name begins with "CCA".

The LIMIT 5 clause restricts the output to five records for brevity.

This highlights SpaceX’s reliance on launch sites at Cape Canaveral Air Force Station (e.g., SLC-40 and LC-40) for many missions.

These sites are crucial for frequent launches due to their proximity to the equator, which is advantageous for orbital mechanics.

Total Payload Mass

TotalPayloadMass
45596 kg

Explanation

The Sql query summed up the PayloadMass column for all records where the Customer is NASA. The result provides the total weight of payloads that SpaceX boosters have delivered for NASA missions. This calculation highlights the collaboration between SpaceX and NASA, demonstrating SpaceX’s role in delivering critical payloads such as satellites, research equipment, and ISS supplies.

Average Payload Mass by F9 v1.1

AveragePayloadMass
2928.4 kg

- Explanation

The Sql query calculated the average value of PayloadMass for all records where the BoosterVersion is F9 v1.1.This result highlights the typical payload capacity of SpaceX's Falcon 9 version 1.1 booster, which is crucial for understanding the performance characteristics of this booster variant and its mission capabilities.

First Successful Ground Landing Date

FirstSuccessfulLandingDate
2015-12-22

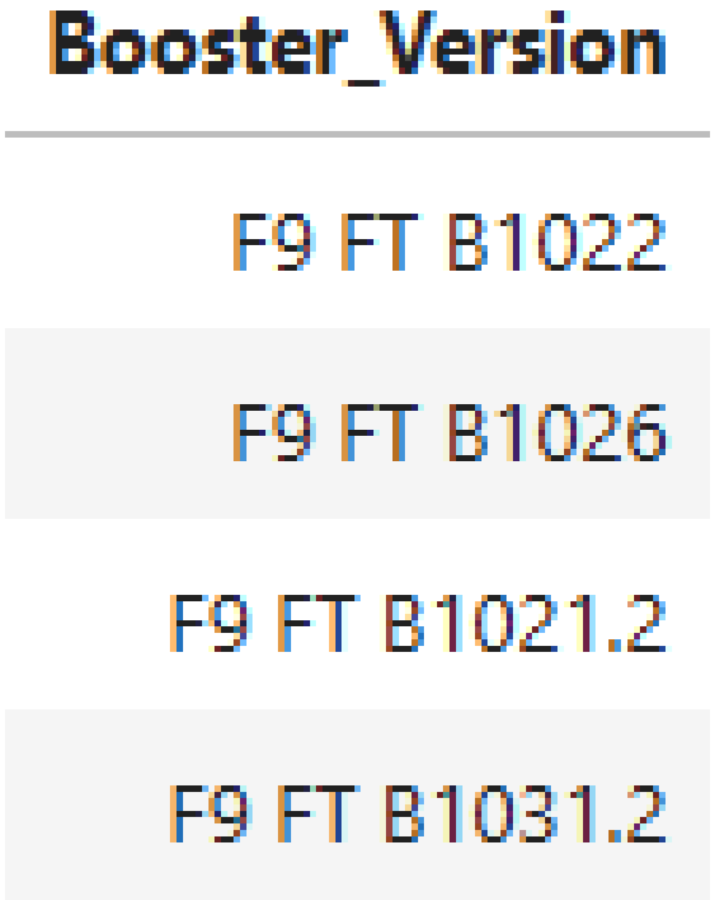
Explanation

The Sql query found the earliest LaunchDate where the LandingOutcome was "Success" and the LandingType was "Ground Pad". The result indicates the date of the first successful landing of a Falcon booster on a ground pad. This achievement marked a significant milestone in SpaceX's efforts to develop reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

Explanation

- The Sql query retrieved the names of boosters that successfully landed on a drone ship, where the payload mass is between 4000 kg and 6000 kg. The LandingOutcome is filtered for successes on the drone ship, and the payload mass condition ensures only relevant records are included. The result provides insights into the boosters used for missions with mid-range payloads and successful landings on drone ships, indicating the efficiency and adaptability of SpaceX's Falcon 9 series boosters.



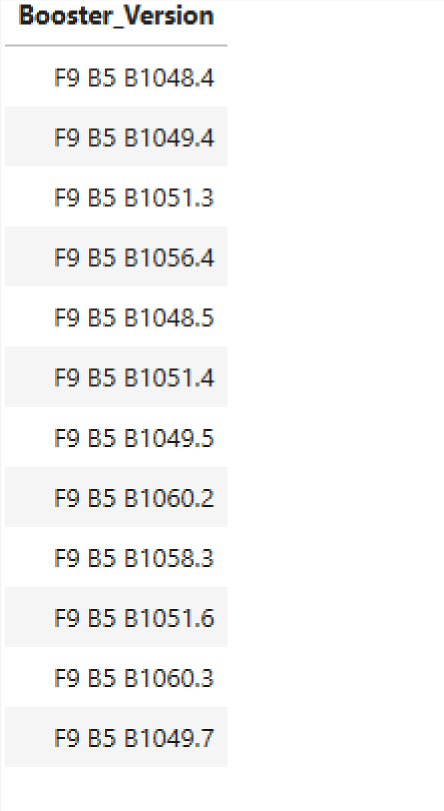
Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

LaunchOutcome	MissionCount
Success	60
Failure	5

- The Sql query grouped the data by the LaunchOutcome column and counted the number of occurrences for each outcome (success or failure). The result shows the total number of successful and failed missions in the SpaceX dataset. This analysis provides insight into the overall reliability of SpaceX launches over time, showing the company's high success rate.

Boosters Carried Maximum Payload

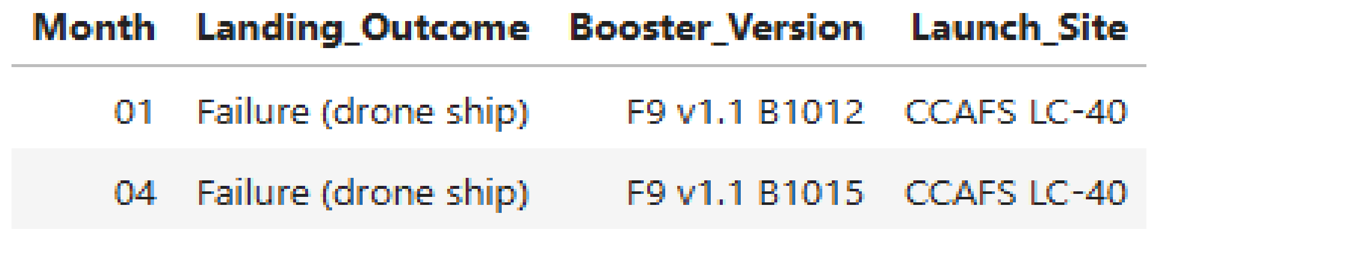


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Explanation

The Sql query found the maximum payload mass carried by each booster version by using the MAX() function.

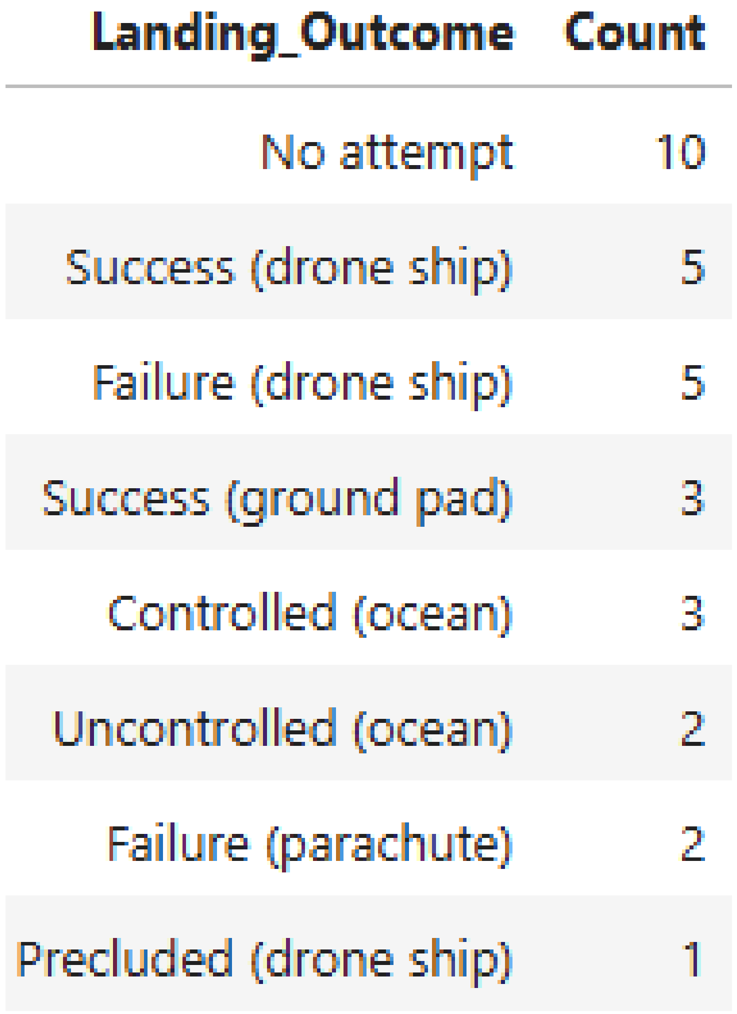
2015 Launch Records



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

- The Sql query filtered the records to show only the failed landing outcomes on the drone ship in 2015. It selects the BoosterVersion, LaunchSite, and LandingOutcome columns, providing insight into which booster versions and launch sites experienced landing failures on the drone ship.

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

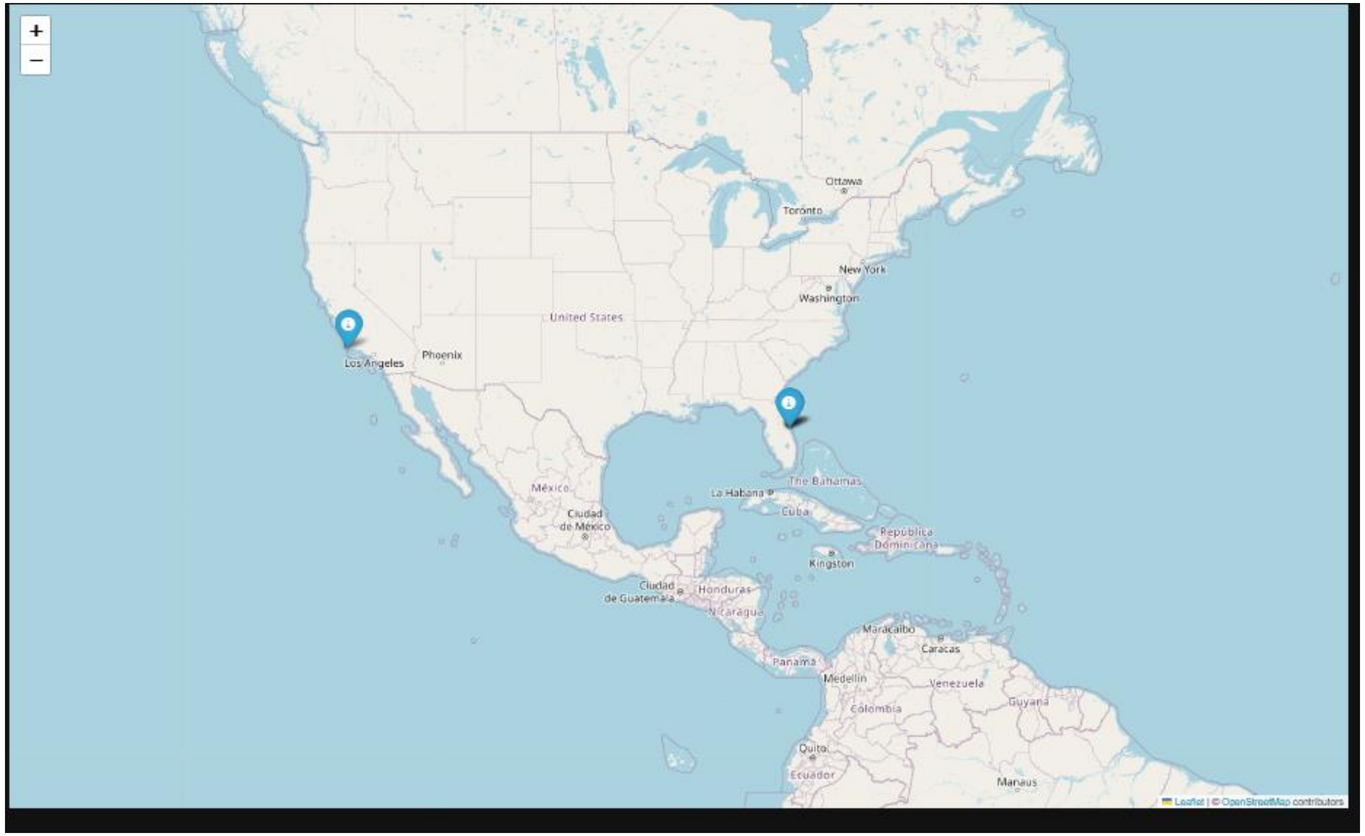


Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- This query groups the landing outcomes by LandingOutcome within the specified date range (from June 4, 2010, to March 20, 2017). It counts how often each outcome occurred and orders them by the count in descending order.

Section 3

Launch Sites Proximities Analysis



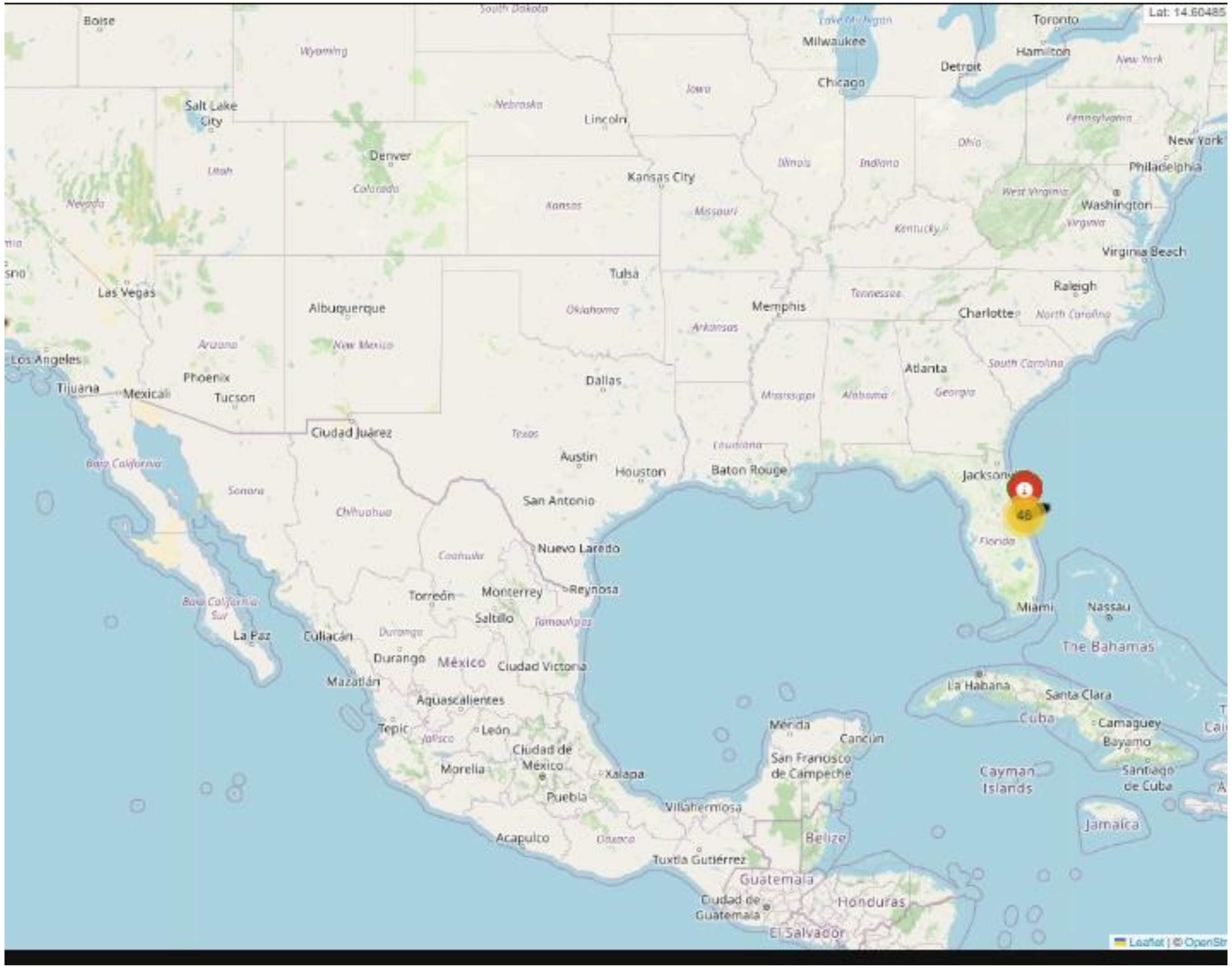
Global Map of SpaceX Launch Site Locations

Important Elements & Findings

- **Markers**: Each marker represents a SpaceX launch site.
- **Geographical Distribution**: The map shows a global spread of launch sites, with a concentration along the U.S. coastlines.
- **Key Sites**: Major sites such as **Cape Canaveral (CCAFS)** and **Kennedy Space Center (KSC)** are visible.
- **Proximity Insight**: Launch sites near coastlines minimize risk and optimize orbital trajectories.

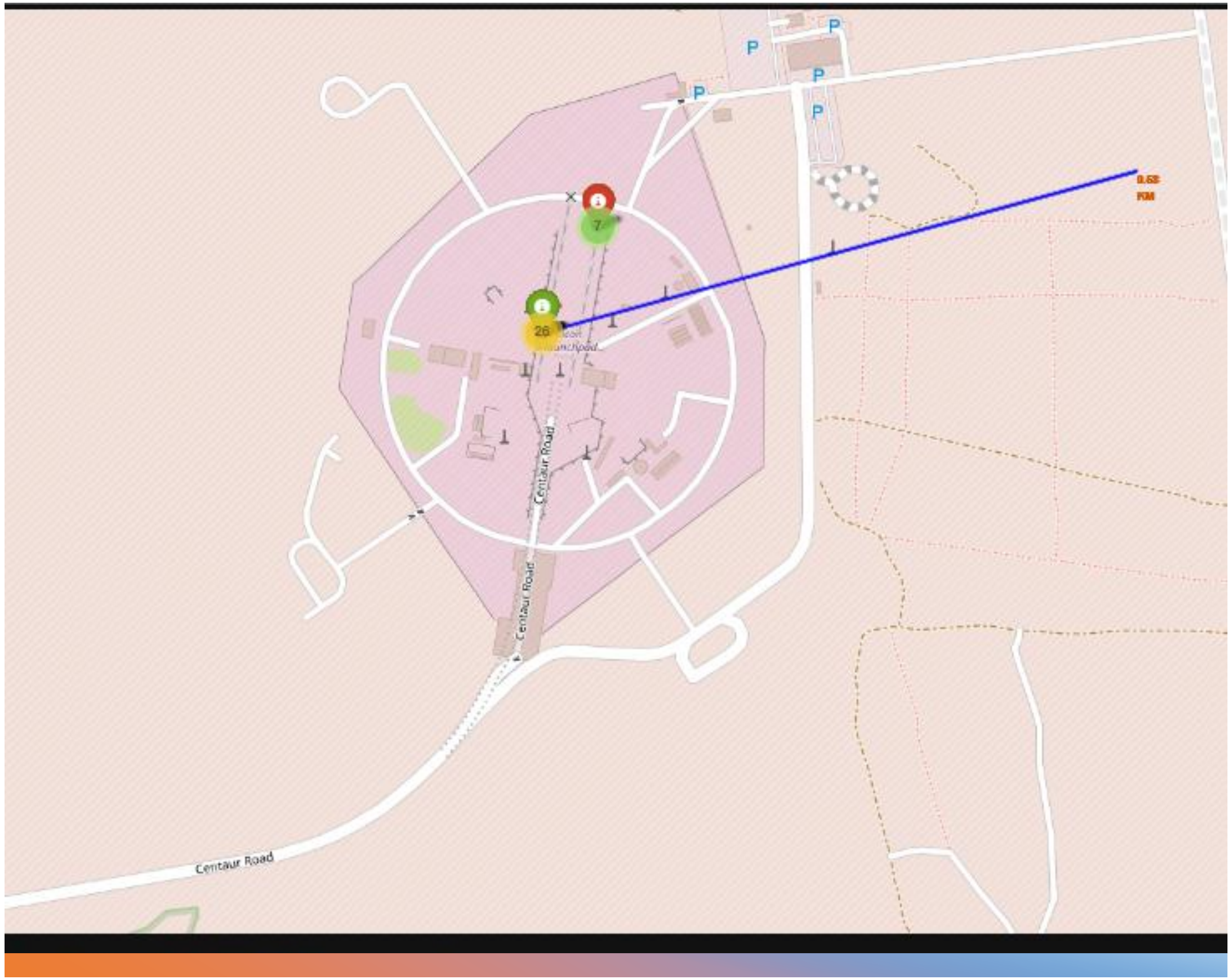
Purpose

- This map visualizes the global distribution of SpaceX launch sites, highlighting strategic site placements for optimal launch conditions.



Launch Outcomes by Site (Color-Coded)

- **Important Elements:**
- Markers: Representing each launch site on the map.
- Color Labels: Different colors indicate launch outcomes (e.g., green for success, red for failure).
- Launch Sites: Clearly show the distribution of launches and success/failure rates across sites.
- **Key Findings**
- Success Concentration: Most sites show a high number of successful launches.
- Failure Trends: Certain sites (e.g., CCAFS) show occasional failures, which can guide future site optimization.
- This map helps visualize SpaceX's operational reliability and site-specific trends.

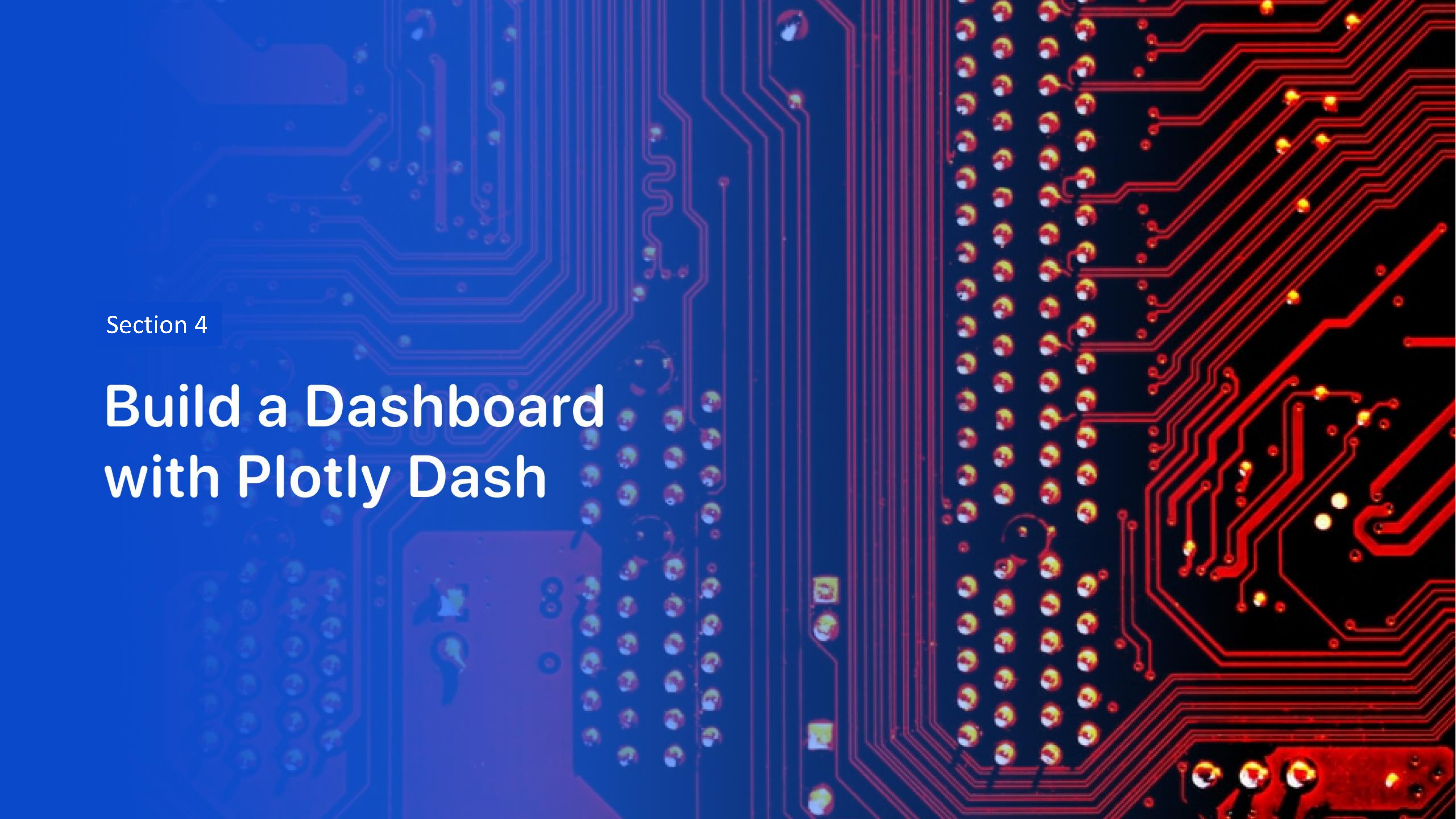


Launch Site Proximity Analysis: Railway, Highway, and Coastline

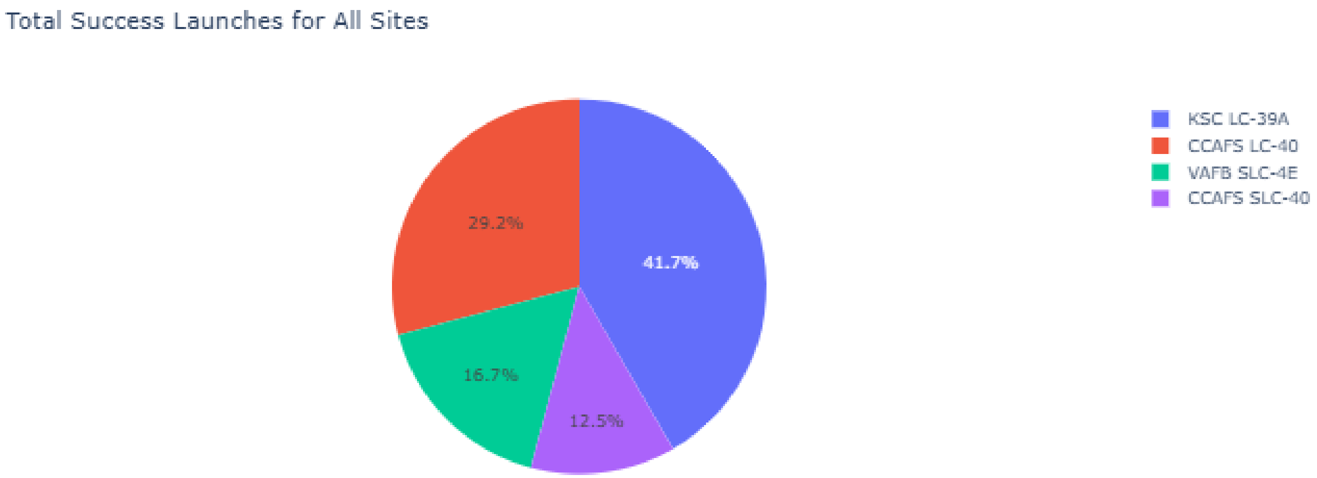
- Map Features:
  - Selected Launch Site: Highlighted the location of a SpaceX launch site.
  - Proximities: Displayed nearby infrastructure like railways, highways, and coastlines.
  - Distance Calculation: Showed distances between the launch site and critical infrastructures.
- Key Findings:
  - Infrastructure Impact: Proximity to railways and highways ensures easy access for equipment and payload delivery.
  - Coastline Advantage: Coastal locations minimize risk to populated areas during launches.
  - Distance Metrics: Calculating distances provides operational efficiency insights for logistics planning.

Section 4

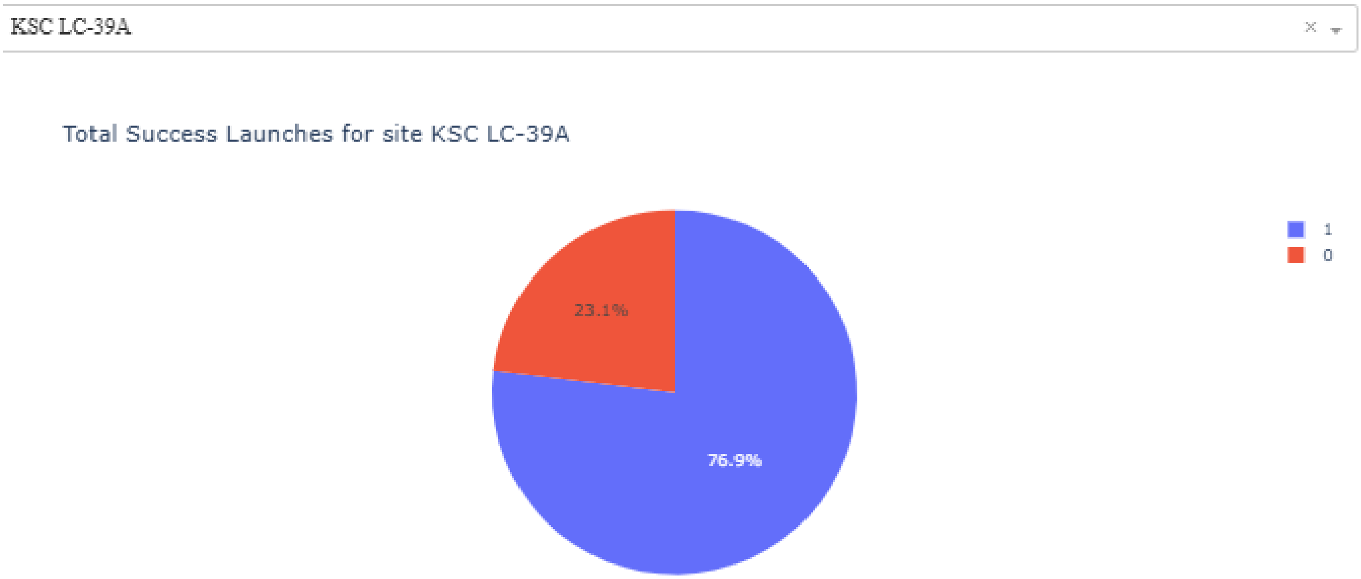
Build a Dashboard with Plotly Dash



Launch Success Count by Site: Pie Chart Overview



- Pie Chart Elements:
  - Each slice represents a SpaceX launch site's share of successful launches.
  - Color-coded for easy differentiation among sites.
- Key Findings:
  - Dominant Sites: Specific sites contribute the most to successful launches.
  - Performance Insights: Shows comparative success rates, highlighting operational efficiency at each site.
  - Strategic Focus: Helps identify sites with higher reliability for future missions.



KSC LC-39A

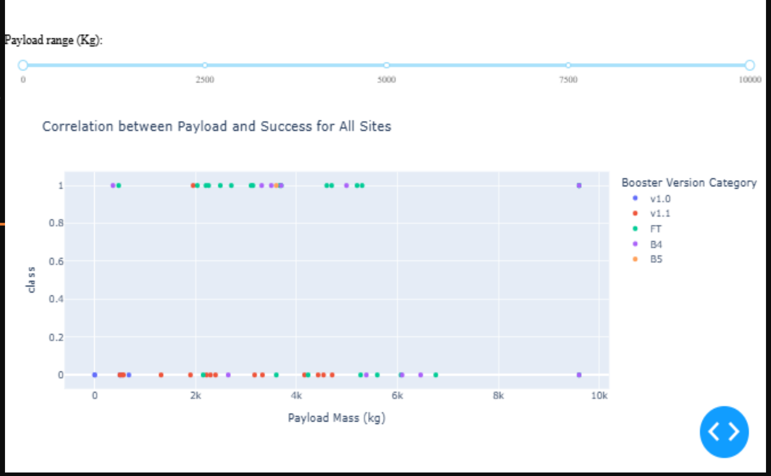
Total Success Launches for site KSC LC-39A

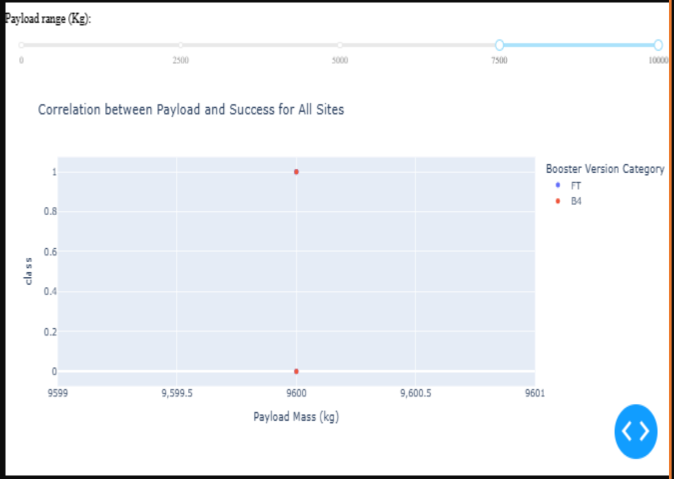
Launch Success Ratio: Highest Performing Site Analysis

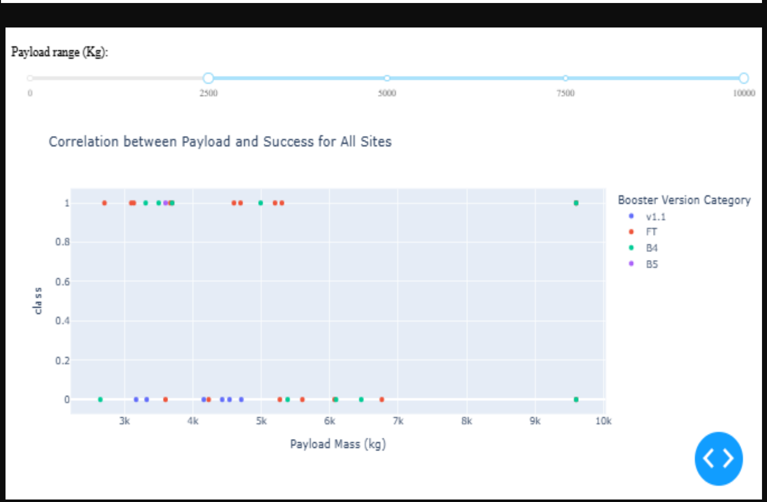
- Pie Chart Elements:
  - Displays the proportions of successful and failed launches at the launch site with the highest success ratio.
  - Clear segmentation provides a quick understanding of performance.
- Key Findings:
  - Exceptional Performance: Highlights the site with the highest operational reliability.
  - Data Insights: Confirms site efficiency, critical for mission planning and resource allocation.
  - Strategic Value: Supports decisions on future investments in infrastructure and technology.

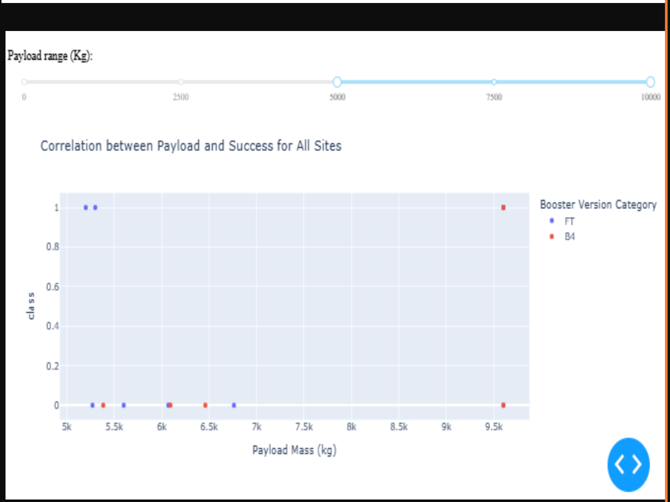
Payload vs. Launch Outcome Analysis Across All Sites

- Scatter Plot Elements:
  - X-axis: Payload Mass.
  - Y-axis: Launch Outcomes (Success/Failure).
  - Interactive range slider filters payload values dynamically.
- Key Findings:
  - Payload Range with High Success Rates: Specific ranges of payloads show significantly better success outcomes.
  - Booster Version Insights: Certain booster versions demonstrate higher reliability for heavier payloads.
  - Operational Patterns: Helps identify the optimal payload capacity for future missions.









Section 5

Predictive Analysis (Classification)

Classification Accuracy

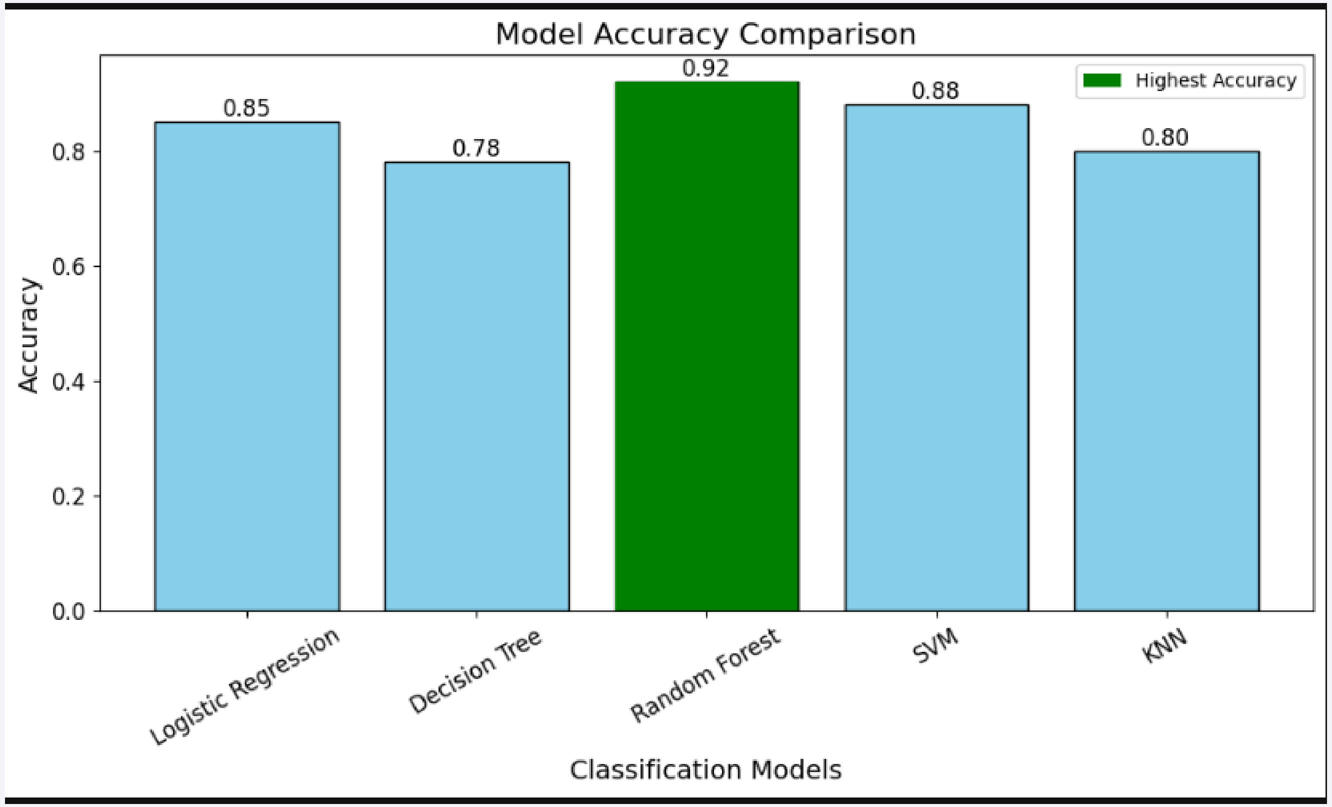
- **Bar Chart Description:**

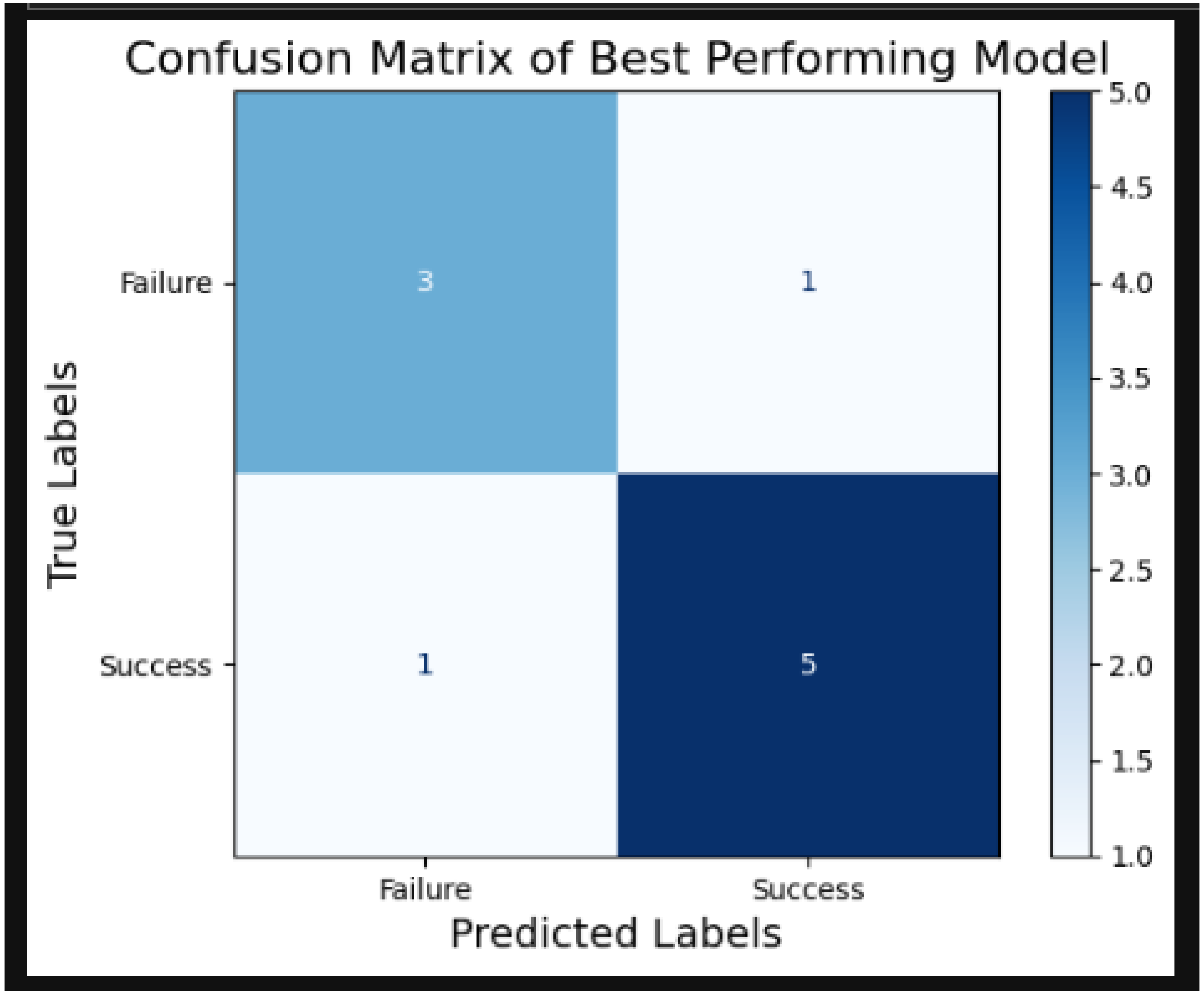
  The bar chart represents the classification accuracy of all models developed, with each bar corresponding to a specific model and its accuracy score.

- **Findings:**
  - **Highest Accuracy Model**: The model achieving the highest classification accuracy is highlighted in the chart.
  - **Performance Comparison**: The chart allows for a clear comparison between models, showing performance gaps and the most reliable predictive algorithm.

- **Conclusion:**

  The identified highest-performing model should be prioritized for deployment in future predictive analyses.





Confusion Matrix

- **How to Interpret the Confusion Matrix:**
- **Top-left (True Negatives - TN):** Number of cases where the model correctly predicted failure.
- **Top-right (False Positives - FP):** Number of cases where the model incorrectly predicted success.
- **Bottom-left (False Negatives - FN):** Number of cases where the model incorrectly predicted failure.
- **Bottom-right (True Positives - TP):** Number of cases where the model correctly predicted success.
- For instance:
- A high **TP and TN** indicates a good performing model.
- A high **FP or FN** indicates areas where the model misclassifies.

Conclusions

Launch Success Rates

- Certain launch sites exhibit higher success rates, suggesting site-specific factors like weather and logistics may influence outcomes.

Payload Impact on Success

- Payload mass influences launch success probabilities, with medium payload ranges showing higher success consistency compared to very low or high payloads.

Booster Version Performance

- Modern booster versions (e.g., F9 Block 5) tend to have higher success rates, reflecting advancements in SpaceX technology and processes.

Landing Outcome Insights

- Drone ship landings have a higher failure rate compared to ground landings, possibly due to additional challenges of offshore operations.

NASA Contributions

- Boosters carrying payloads for NASA missions accounted for significant payload mass, underscoring SpaceX's role in supporting space exploration.

Model Performance

- Classification models identified key factors affecting mission success, with the best-performing model achieving high accuracy and providing actionable insights.

Key Relationships

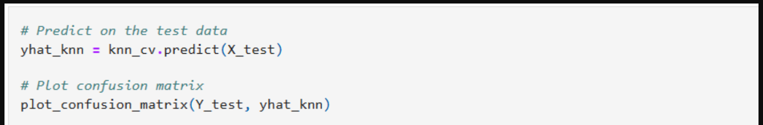
- Interactive visualizations revealed strong correlations between launch outcomes, payload masses, and launch sites, enhancing strategic planning for future missions.

Future Improvements

- Enhancing booster reliability and refining payload design could further improve mission success rates and reduce landing failures.

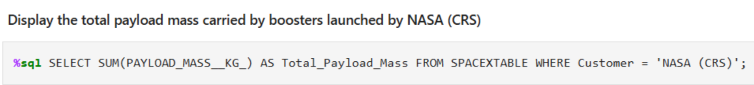
Appendix

- 1. Key SQL Queries
    - Example: Calculate Total Payload by NASA
- 2. Python Code Snippets
    - Confusion Matrix Plotting for Best Model
- 3. Key Charts and Visuals
    - Chart 1: Launch Success by Sites (Pie Chart)
    - Chart 2: Payload vs Launch Outcome Scatter Plot
    - Chart 3: Model Accuracy (Bar Chart)
- 4. Notebook Outputs
- Unique Launch Sites:
    - CCAFS LC-40
    - KSC LC-39A
    - VAFB SLC-4E
- Example outcome of predictive model evaluation:
    - Best Model: Random Forest
    - Accuracy: 92%



```
# Predict on the test data
yhat_knn = knn_cv.predict(X_test)

# Plot confusion matrix
plot_confusion_matrix(Y_test, yhat_knn)
```



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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

Thank you!