IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Chekwube Ononuju
07/06/2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

The Applied Data Science Capstone Project aims to predict the landing success of SpaceX's Falcon 9 first stage. This initiative is crucial as SpaceX's ability to reuse the first stage of its rockets significantly reduces launch costs, positioning it competitively against other providers. By accurately predicting whether the first stage will land successfully, we can estimate the cost of a launch more precisely.

The main takeaway from the results is that the predictive model can accurately forecast the success of Falcon 9 first stage landing. Decision trees have emerged as the best model for predicting landing outcomes due to their high accuracy, interpretability, and ability to handle complex datasets effectively.

INTRODUCTION

- **Capstone Project: Predicting Falcon 9 First Stage Landing Success**
- **Objective:**
- To predict whether the Falcon 9 first stage will land successfully.
- **Context:**
- **SpaceX Launch Cost:** $62 million per launch.
- **Competitor Launch Cost:** Over $165 million per launch.
- **Cost Savings:** Due to the reuse of the first stage of the rocket.
- **Significance:**
- Accurate predictions of first stage landing success can:
    - Estimate the cost of a launch.
    - Aid alternate companies in competitive bidding against SpaceX.
    - Inform strategic planning in the space launch industry.

METHODOLOGIES SUMMARY

Data Collection and Web Scraping:

- •Gathered launch data using the SpaceX REST API
- •Web Scraping HTML tables from Wikipedia

Data Wrangling:

- •Created new features and Included attributes like Flight Number, Date, Booster version, Payload mass, Orbit, Launch Site, and Outcome for the Feature Engineering.
- •Collected detailed data using additional API endpoints and Filtered the data to include only Falcon 9 launches

Exploratory Data Analysis (EDA)

- •Analyzed success rates by launch site and payload mass.
- Used SQL for data selection, sorting, and analysis.

Interactive Visual Analytics and Dashboard

- Employed Plotly Dash **to d**evelop interactive dashboards.
- Built interactive maps for launch site proximity analysis with folium.

Predictive Analysis (Classification)

- Implemented K-Nearest Neighours, Decision Trees, Support Vector Machines and, Logistic Regression machine learning models.

Project Background and Context

SpaceX, a pioneering private aerospace manufacturer and space transportation company, has revolutionized the space industry with its Falcon 9 rocket. A significant factor contributing to SpaceX's competitive edge is its ability to reuse the first stage of the Falcon 9 rocket, significantly reducing launch costs. While other providers charge upwards of 165 million dollars for a rocket launch, SpaceX advertises Falcon 9 launches at 62 million dollars. The substantial cost savings primarily stem from the reusability of the first stage of the rocket.

Predicting the successful landing of the Falcon 9's first stage is crucial. The ability to reliably forecast this outcome not only impacts the cost of launches but also enhances the planning and operational efficiency of SpaceX. Furthermore, this predictive capability is valuable for alternate companies looking to compete against SpaceX in the rocket launch market. Accurate predictions can inform better bidding strategies and provide a clearer understanding of potential launch costs and risks.

PROBLEMS WE WANT TO FIND ANSWERS TO

WHAT ARE THE KEY FACTORS THAT INFLUENCE THE SUCCESSFUL LANDING OF THE FALCON 9 FIRST STAGE?

HOW ACCURATELY CAN WE PREDICT THE SUCCESSFUL LANDING OF THE FALCON 9 FIRST STAGE USING HISTORICAL DATA?

HOW DOES THE PROBABILITY OF A SUCCESSFUL LANDING AFFECT THE OVERALL COST SAVINGS FOR SPACEX?

WHAT IMPROVEMENTS CAN BE MADE TO INCREASE THE SUCCESS RATE OF FALCON 9 FIRST STAGE LANDINGS?

HOW CAN COMPETING COMPANIES LEVERAGE THIS INFORMATION TO BID MORE EFFECTIVELY AGAINST SPACEX?

Section 1

Methodology

METHODOLOGY

- Executive Summary
- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

DATA COLLECTION

DESCRIBE HOW DATA SETS WERE COLLECTED.

PRESENT DATA COLLECTION PROCESS WHILE USING KEY PHRASES AND FLOWCHARTS

Data Collection – SpaceX API

Data Collection with SpaceX REST API

To predict the successful landing of the Falcon 9 first stage, we need to gather relevant data from SpaceX's public REST API.

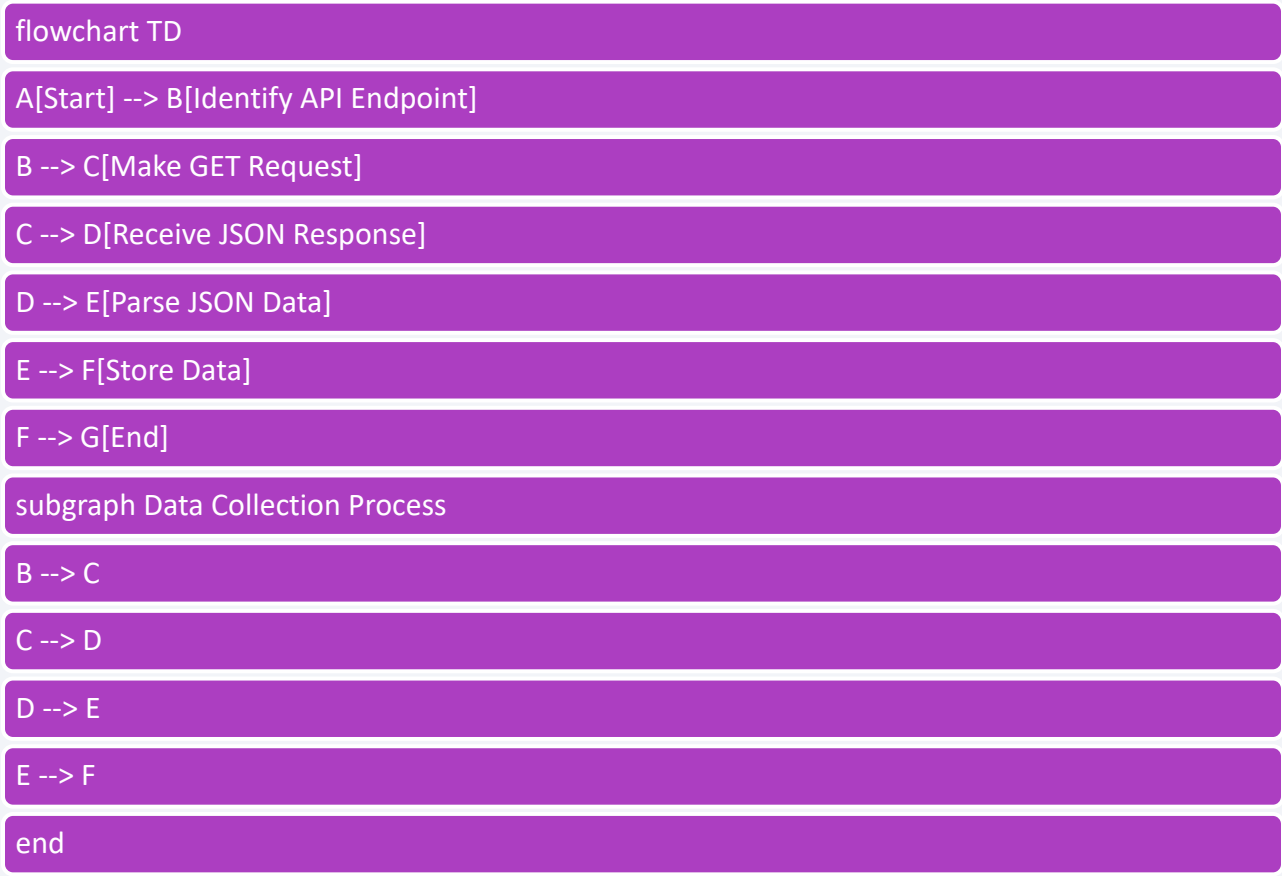
API Endpoint: The URL where the data is accessed.

GET Request: The HTTP method used to retrieve data from the API.

JSON Response: The format in which the data is received.

Data Parsing: Extracting relevant information from the JSON response.

Data Storage: Saving the extracted data for analysis.

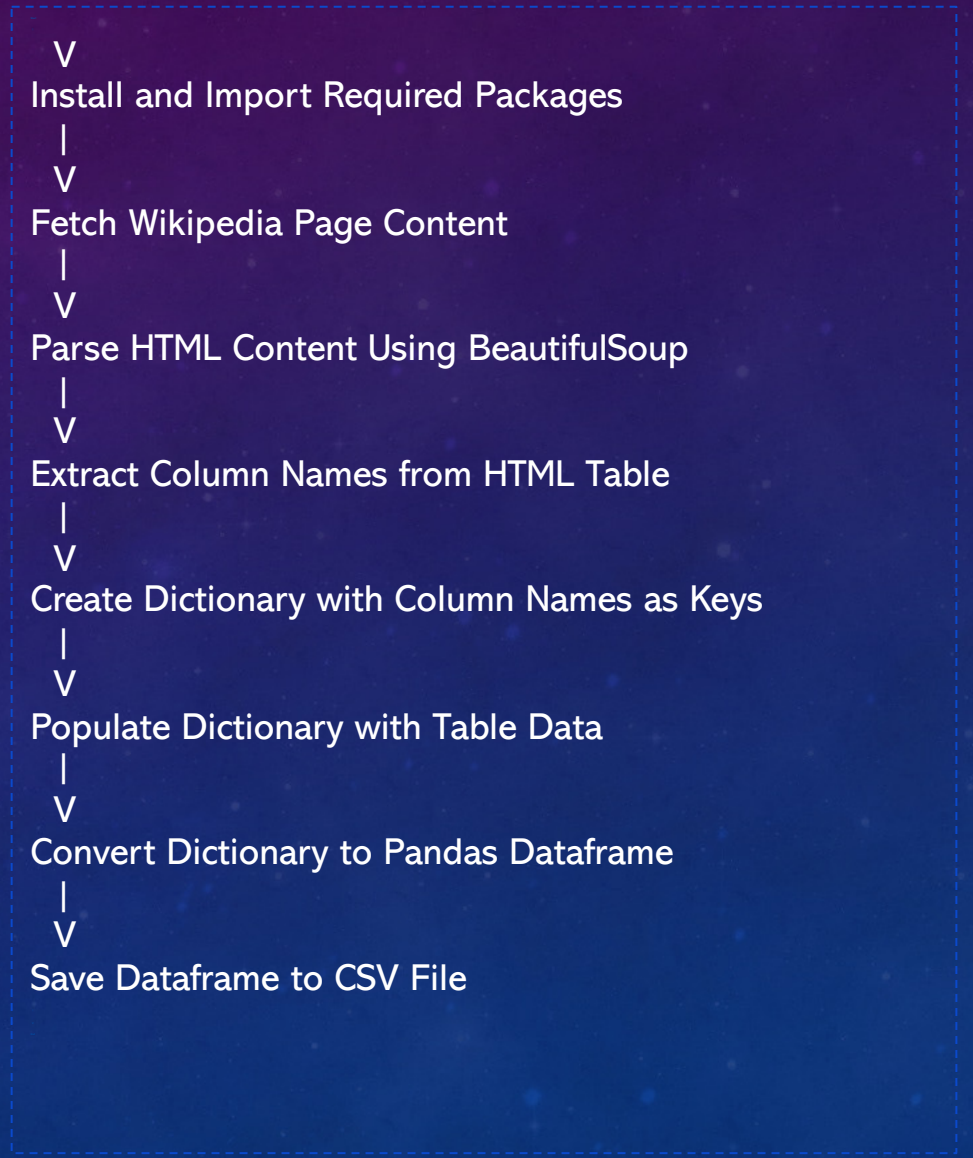


For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

Data Collection - Scraping

Web Scraping Process for Falcon 9 Launch Records:

Extract Falcon 9 launch records HTML table from Wikipedia.
Parse the table and convert it into a Pandas dataframe.
Install BeautifulSoup and requests using pip.
Import necessary packages (sys, requests, BeautifulSoup, pandas).
Perform an HTTP GET request to fetch the Wikipedia page containing the Falcon 9 launch records.
Identify and extract relevant column names from the HTML table header on the Wiki page.
Create a Dataframe by Parsing the Launch HTML Tables



For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

DATA WRANGLING

Data Wrangling Process:

1.Data Collection:

Source: Collected Falcon 9 launch records from SpaceX REST API and web scraping.

Methods: Used requests and BeautifulSoup for web scraping, and API calls to fetch JSON data.

2.Data Loading:

Pandas DataFrame: Loaded raw data into a pandas DataFrame for easier manipulation and analysis.

3.Data Cleaning:

Handling Missing Values: Identified and filled or removed missing values in critical columns.

Data Type Conversion: Converted data types to appropriate formats (e.g., dates to datetime objects, numerical values to floats/integers).

Duplicate Removal: Removed duplicate records to ensure data integrity.

4.Data Transformation:

Normalization: Standardized column names and values for consistency.

Feature Engineering: Created new features (e.g., 'Launch Year', 'Launch Success') based on existing data.

Filtering: Filtered out irrelevant data (e.g., launches outside the specified date range).

5.Data Integration:

Merging Data: Combined data from different sources into a single DataFrame.

Database Loading: Loaded the cleaned and processed DataFrame into an SQLite database for efficient querying.

6.Data Exploration:

Descriptive Statistics: Calculated summary statistics to understand data distribution and central tendencies.

Visualization: Created plots (e.g., histograms, scatter plots) to visualize data patterns and relationships.

Flowchart:

1. **Data Collection**:

   API Calls → JSON Responses → Load into DataFrame

   Web Scraping → HTML Parsing → Extract Tables → Load into DataFrame

2. **Data Loading**:

   Load CSV/JSON Files → Pandas DataFrame

3. **Data Cleaning**:

   Identify Missing Values → Fill/Remove Missing Values

   Convert Data Types → Remove Duplicates

4. **Data Transformation**:

   Normalize Data → Feature Engineering → Filter Data

5. **Data Integration**:

   Merge DataFrames → Load into Database

6. **Data Exploration**:

   Calculate Statistics → Generate Visualizations

For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

EDA WITH DATA VISUALIZATION

In the EDA various charts and graphs were plotted for various needs and purposes,

These graphs include:

•A scatter plot which shows the relationship between payload mass and success (1) or failure (0) of launches at the VAFB SLC-4E site.

•A scatter plot which depicts the relationship between payload mass and launch success for the CCAFS LC-40 site.

•A plot combines the data from all launch sites, showing the relationship between payload mass and success.

•A pie chart showing the proportion of successful launches from each site.

•A scatter plot showing the relationship between the flight number and the launch site.

•This plot shows the relationship between payload mass and the launch site.

•A bar chart showing the success rate for each type of orbit.

•A scatter plot shows the relationship between flight number and orbit type.

•A plot shows the relationship between payload mass and orbit type.

•A line chart showing the yearly trend of launch success rates.

•These visualizations were used to analyze the performance and trends of SpaceX launches, providing insights into which launch sites, payload masses, and orbit types have the highest success rates and how these factors have evolved over time.

For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

EDA WITH SQL

SQL Queries Summary:

Create Table with Non-Null Dates.

Load DataFrame into SQLite Database.

Select Distinct Launch Sites

Filter Launch Sites by Prefix

Calculate Total Payload Mass for NASA

Average Payload Mass for Specific Booster Version

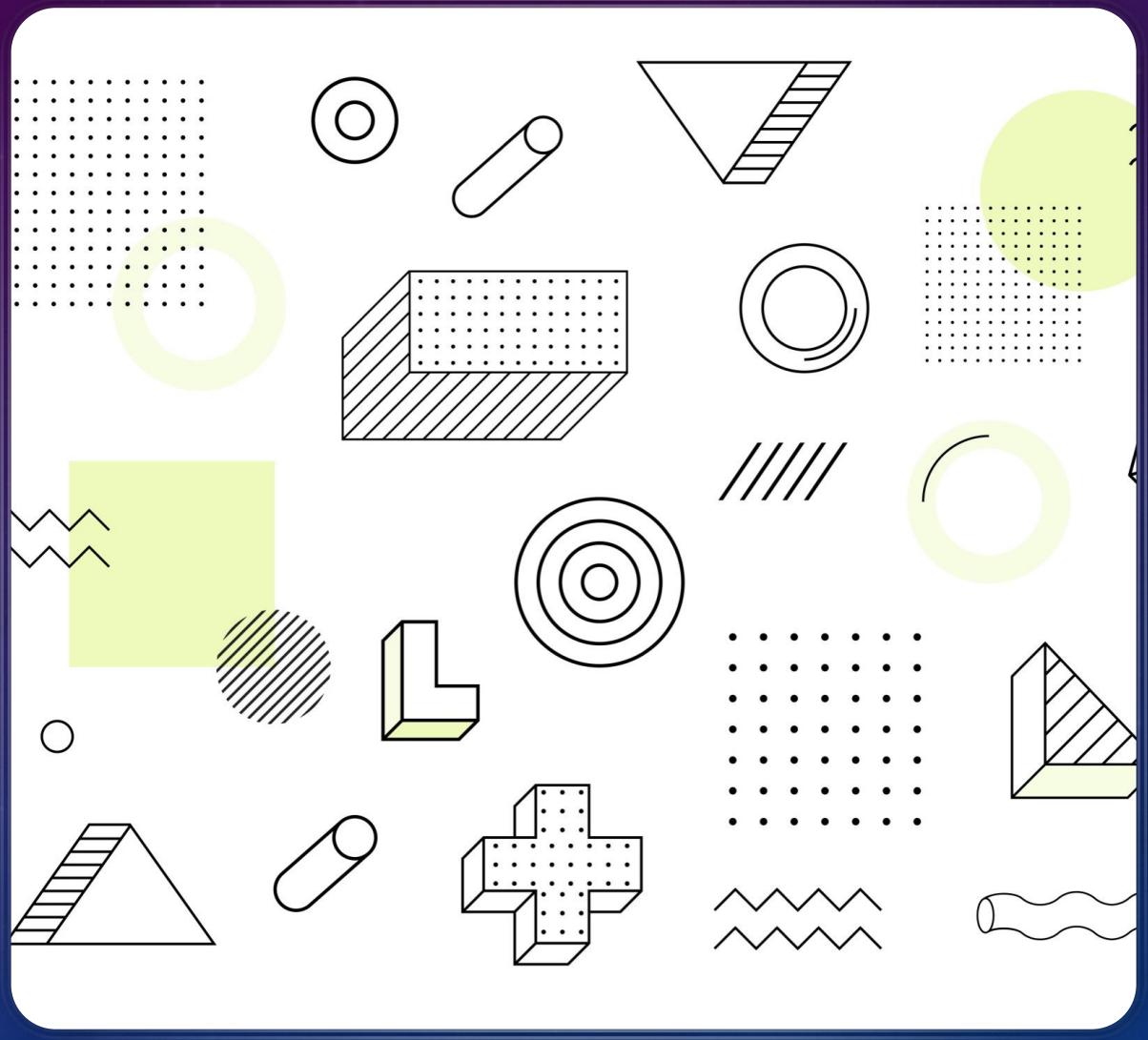
Find First Successful Landing Date on Ground Pad

Select Boosters with Payload Mass in Specific Range

Select Booster with Maximum Payload Mass

Identify Records with Specific Failure Conditions

Ranking Landing Outcomes by Frequency



For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook.

Interactive Map with Folium

- **Map Objects Added to Folium Map:**

  Launch Sites: Added markers at various SpaceX launch sites.

  Closest Highway: Added a marker indicating the closest highway to a specific point.

  Distance Line: Drew a line representing the distance from a specific point to the closest highway.

  Route Lines: Added lines connecting different launch sites to represent possible routes or connections.

  Launch Site Circles: Added circles around launch sites to highlight the areas and show their significance.

- **Explanation of Added Objects:**

1. **Markers**:
   1. **Launch Sites**: Placing markers at launch sites helps in visualizing their exact locations on the map. This is crucial for understanding the geographical distribution of SpaceX's launch infrastructure.
   2. **Closest Highway**: Adding a marker for the closest highway to a launch site provides context about accessibility and logistics. It helps in analyzing the ease of transportation and support infrastructure.
2. **Lines**:
   1. **Distance Line**: Drawing a line from a point to the closest highway helps in visualizing the proximity and understanding logistical considerations. It can be used to assess the travel distance for transporting equipment or personnel.
   2. **Route Lines**: Connecting different launch sites with lines helps in understanding possible logistical routes and connections between sites. This can be useful for planning transportation, resource allocation, and coordination between different launch sites.
3. **Circles**:
   1. **Launch Site Circles**: Adding circles around launch sites highlights their importance and gives a sense of the area they cover. It helps in understanding the spatial distribution and the potential impact area of each launch site.

For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

Dashboard with Plotly Dash

Summary of Plots/Graphs and Interactions Added to the Dashboard

Launch Success Yearly Trend: This plot shows the trend of launch success rates over the years.

Landing Outcomes: This chart displays the count of different landing outcomes (e.g., success on ground pad, success on drone ship, failure, etc.).

Mission Outcomes by Launch Site: This chart shows the success and failure counts for missions launched from different sites.

Proportion of Landing Outcomes: This chart visualizes the proportion of each type of landing outcome in a circular format.

Payload Mass Over Time: This chart shows the trend of payload mass over time, highlighting any changes in payload capabilities.

Launch Sites and Outcomes: A geographical heatmap showing launch sites and the density of successful and unsuccessful landings.

For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

Predictive Analysis (Classification)

Model Evaluation and Selection Process

1. Model Selection

•**Initial Models Chosen**:

- **Logistic Regression**
- **Decision Tree**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**

2. Model Training and Hyperparameter Tuning

•**Logistic Regression**:

- Hyperparameters tuned using GridSearchCV:
  - Regularization strength (C)
  - Solver (e.g., 'liblinear', 'saga')

•**Decision Tree**:

- Hyperparameters tuned using GridSearchCV:
  - Maximum depth (max_depth)
  - Minimum samples split (min_samples_split)
  - Criterion (e.g., 'gini', 'entropy')

•**Support Vector Machine (SVM)**:

- Hyperparameters tuned using GridSearchCV:
  - Kernel (e.g., 'linear', 'rbf')
  - Regularization parameter (C)
  - Gamma (for 'rbf' kernel)

•**K-Nearest Neighbors (KNN)**:

- Hyperparameters tuned using GridSearchCV:
  - Number of neighbors (n_neighbors)
  - Weights (e.g., 'uniform', 'distance')

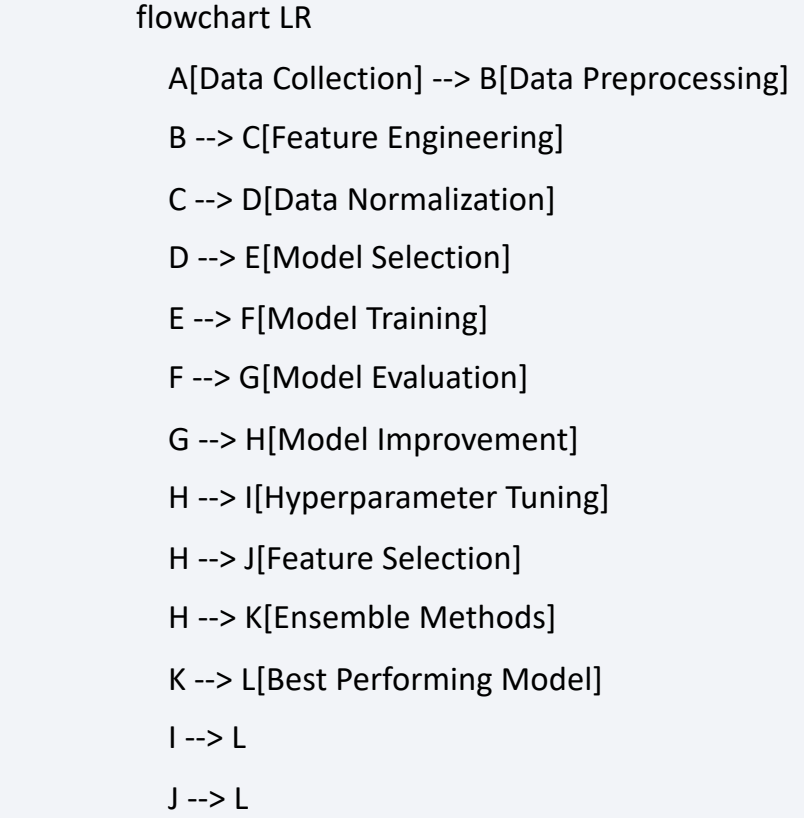
3. Model Evaluation

•**Evaluation Metrics Used**:

- **Accuracy**: Overall correctness of the model.
- **Grid Search CV Score**: Average performance across different folds during cross-validation.

4. Final Model Selection

Based on the evaluation metrics, the best performing model was selected for predicting the Falcon 9 first stage landing success



For more detailed exploratory data analysis and visualizations, refer to the EDA with data visualization notebook

RESULTS

The exploratory data analysis provided several valuable insights into the SpaceX

Falcon 9 launch dataset:

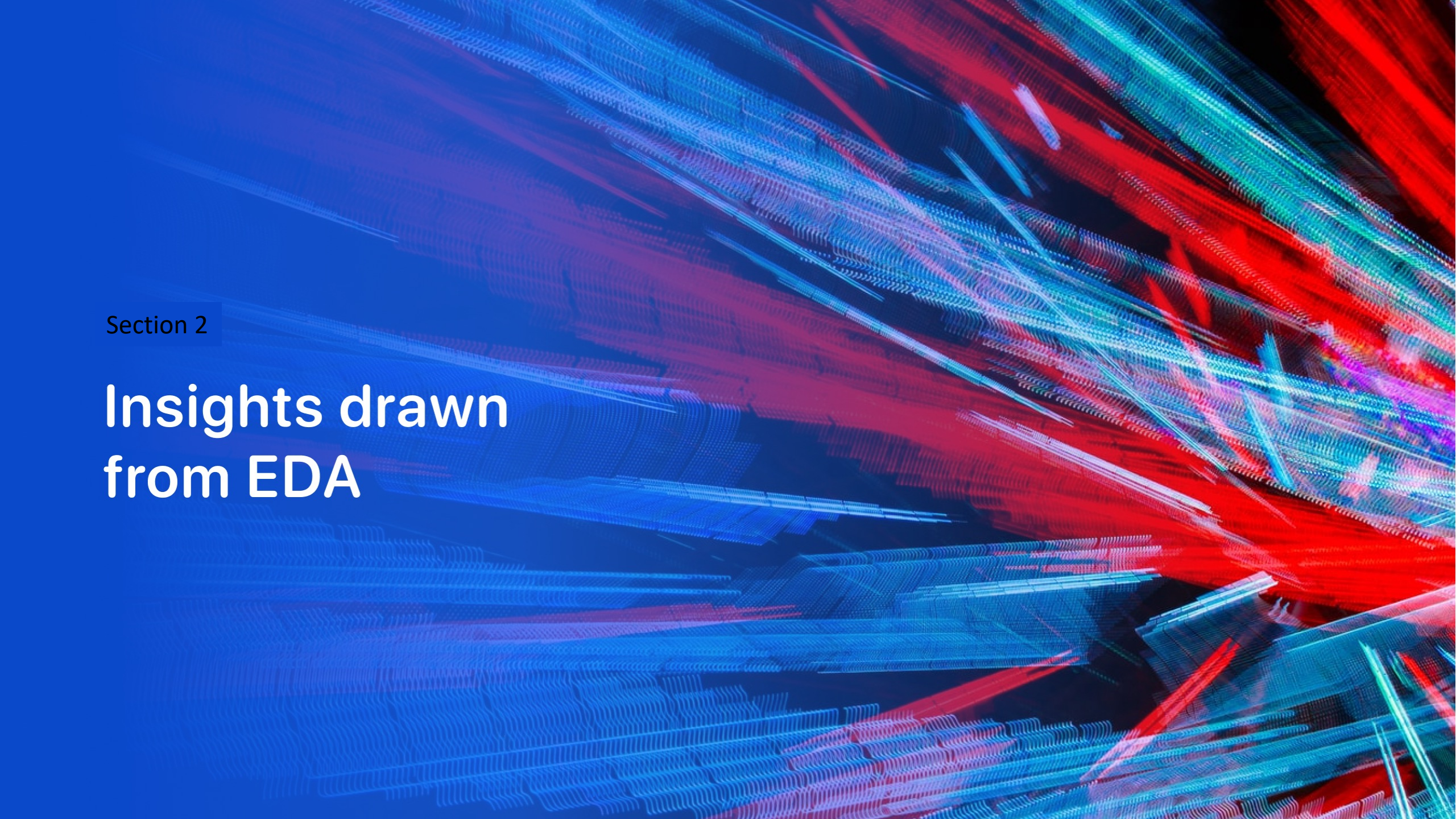
Trends Over Time: The success rate of launches has been improving over the years.

Launch Sites: Certain launch sites are more frequently used, indicating their importance or suitability.

Payload Distribution: Most payloads fall within a specific weight range, with some notable outliers.

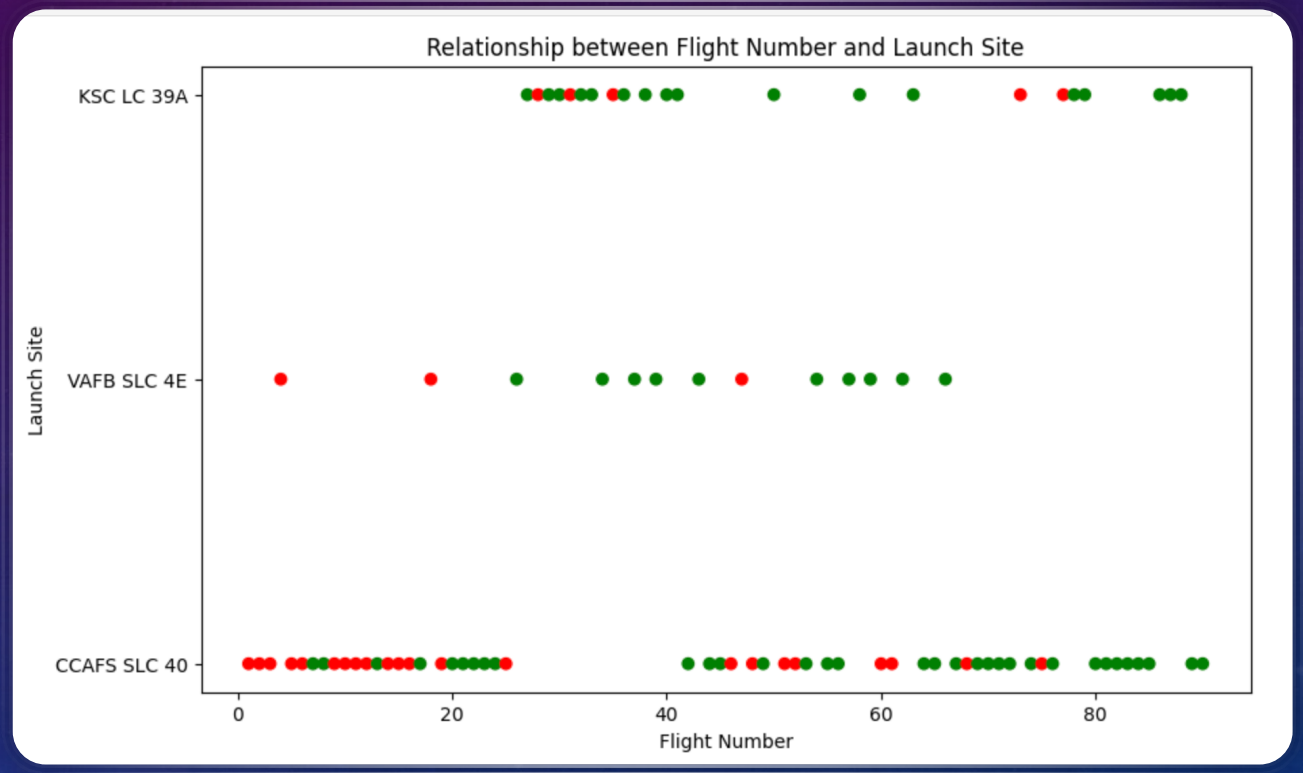
Mission and Landing Outcomes: The majority of missions and landings are successful, highlighting SpaceX's reliability.

Geospatial Distribution: Mapping launch sites provides a clear view of their geographical spread.


Section 2
Insights drawn from EDA

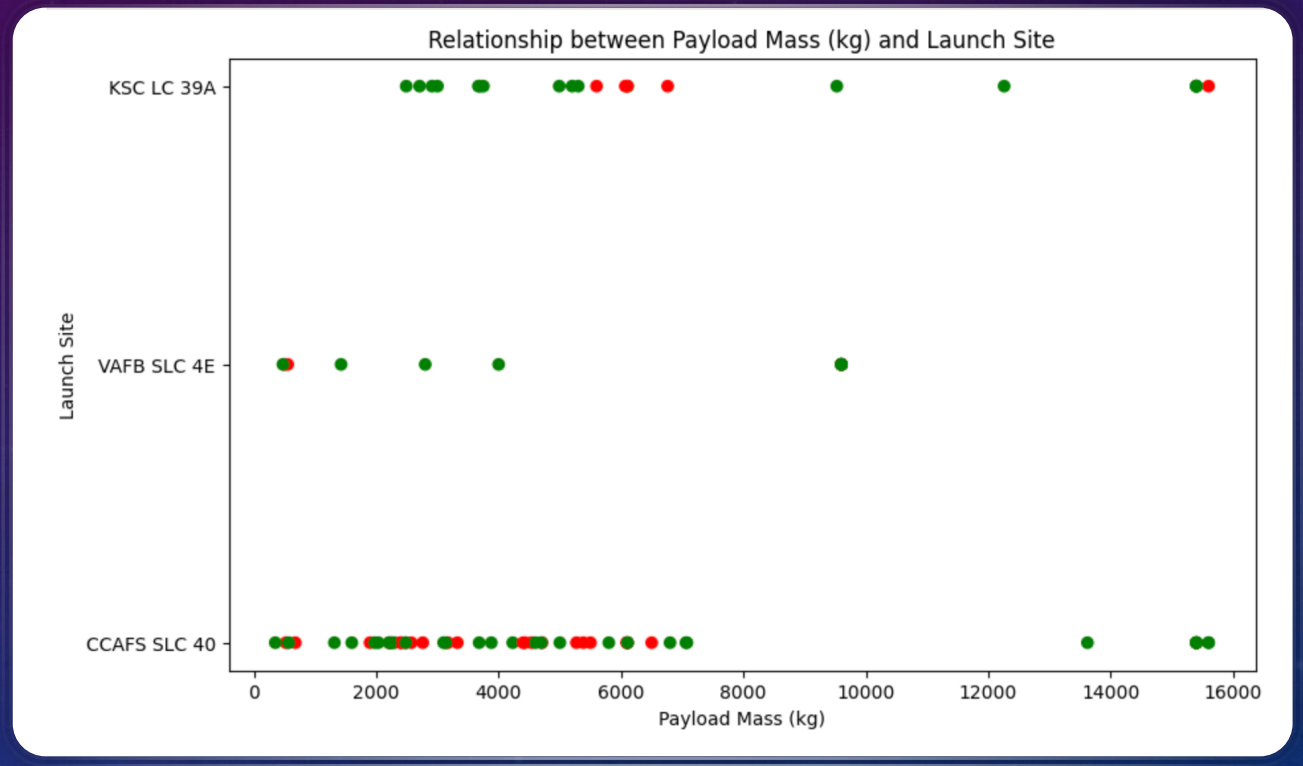
FLIGHT NUMBER VS. LAUNCH SITE

This scatter plot provides insights into the performance and reliability of different launch sites over time, as represented by flight numbers.



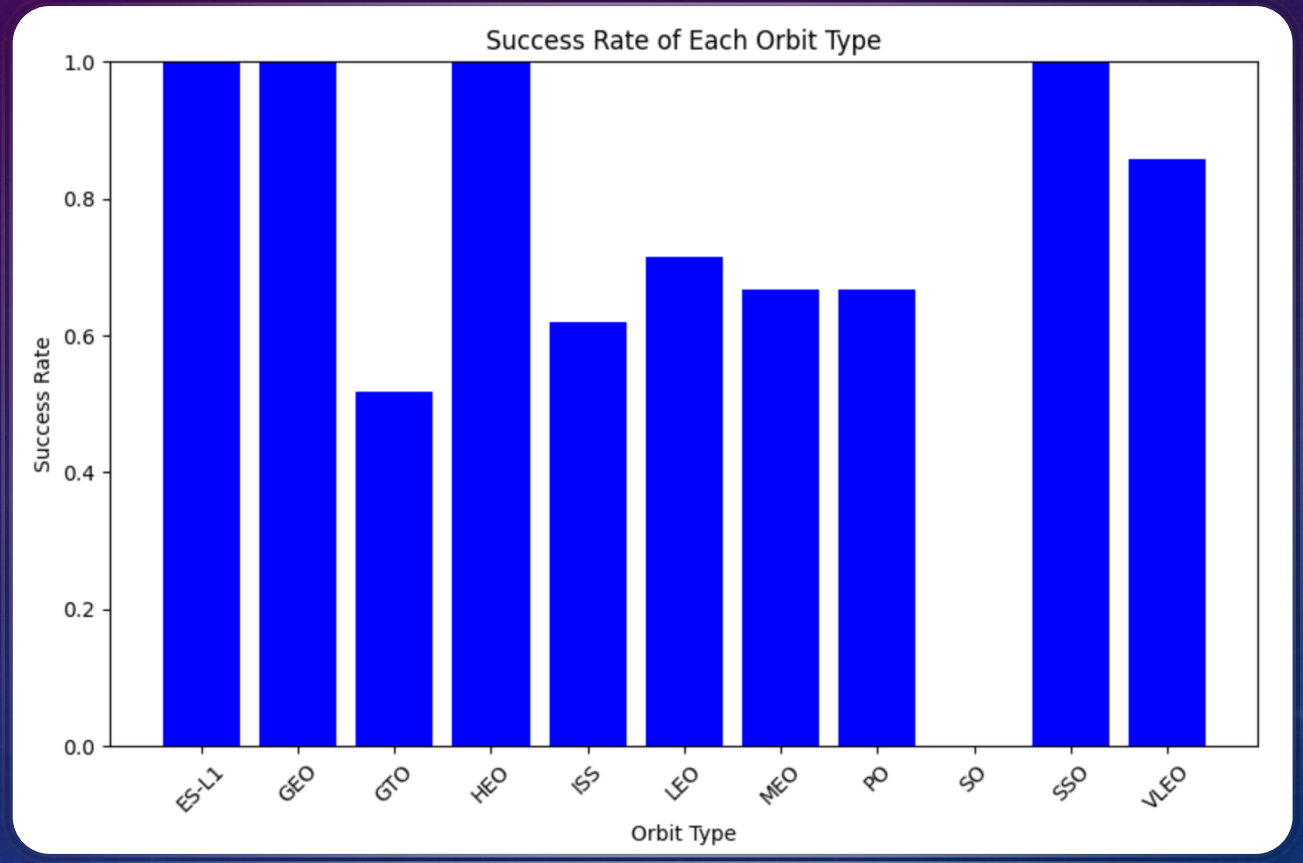
PAYLOAD VS. LAUNCH SITE

This scatter plot helps visualize the distribution of payload masses across different launch sites and can indicate the performance (success or failure) of launches with varying payloads at each site.



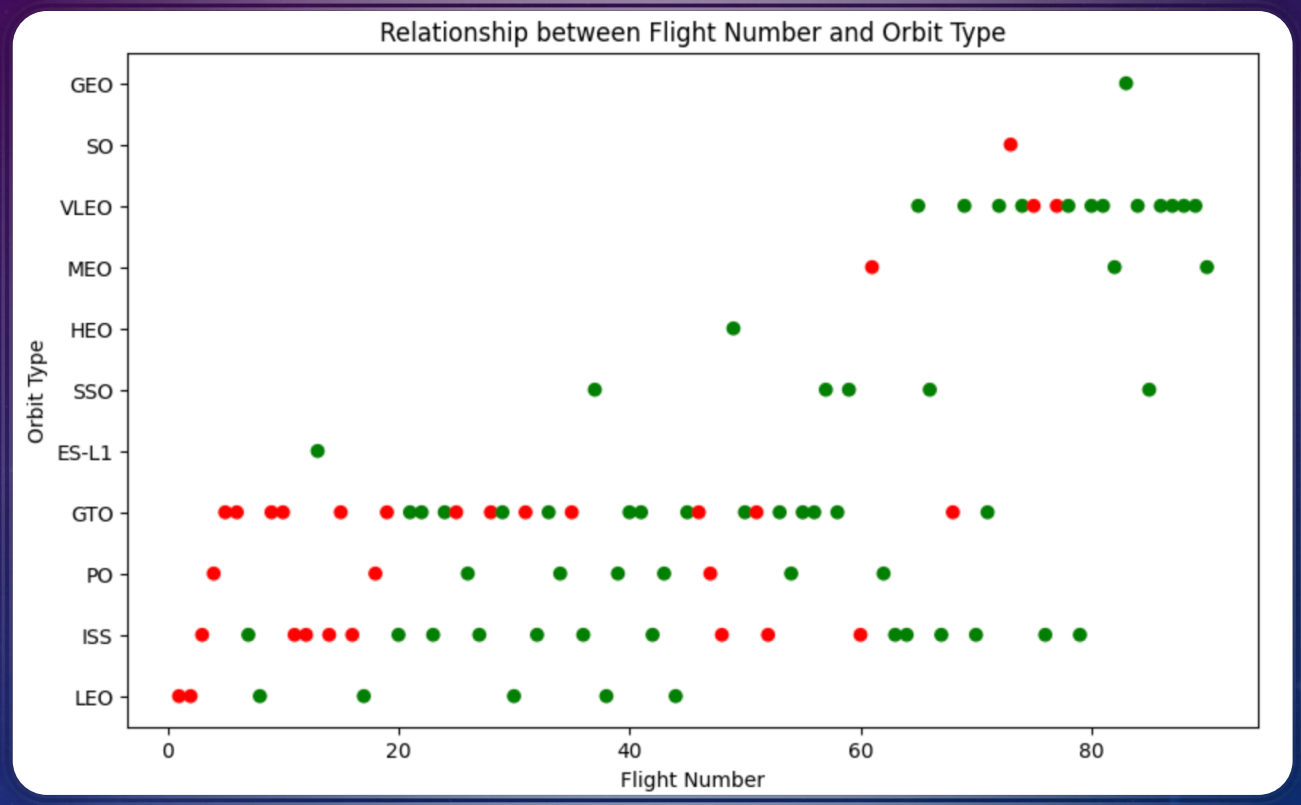
SUCCESS RATE VS. ORBIT TYPE

This bar chart effectively shows the success rates of Falcon 9 first stage landings for different orbit types.



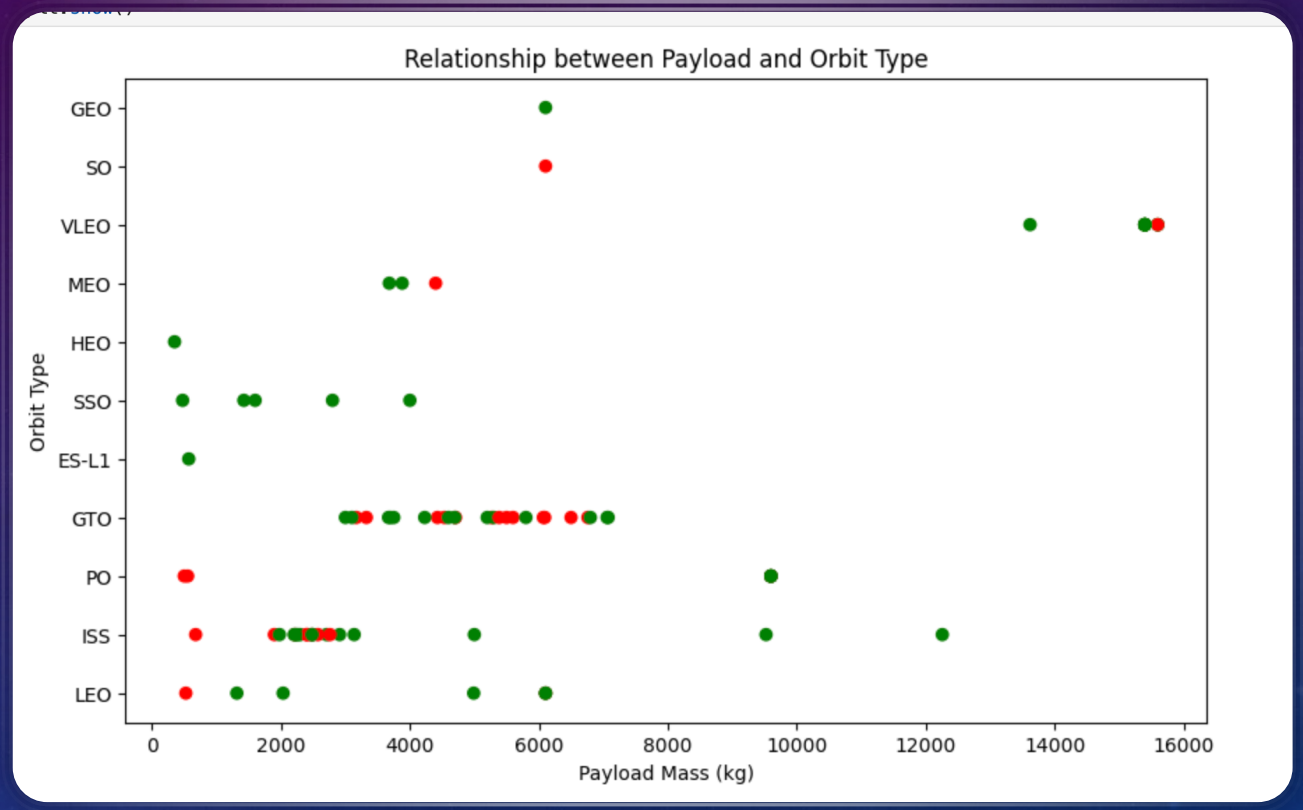
FLIGHT NUMBER VS. ORBIT TYPE

This plot helps visualize the distribution of successful and failed landings across different flight numbers and orbit types.



PAYLOAD VS. ORBIT TYPE

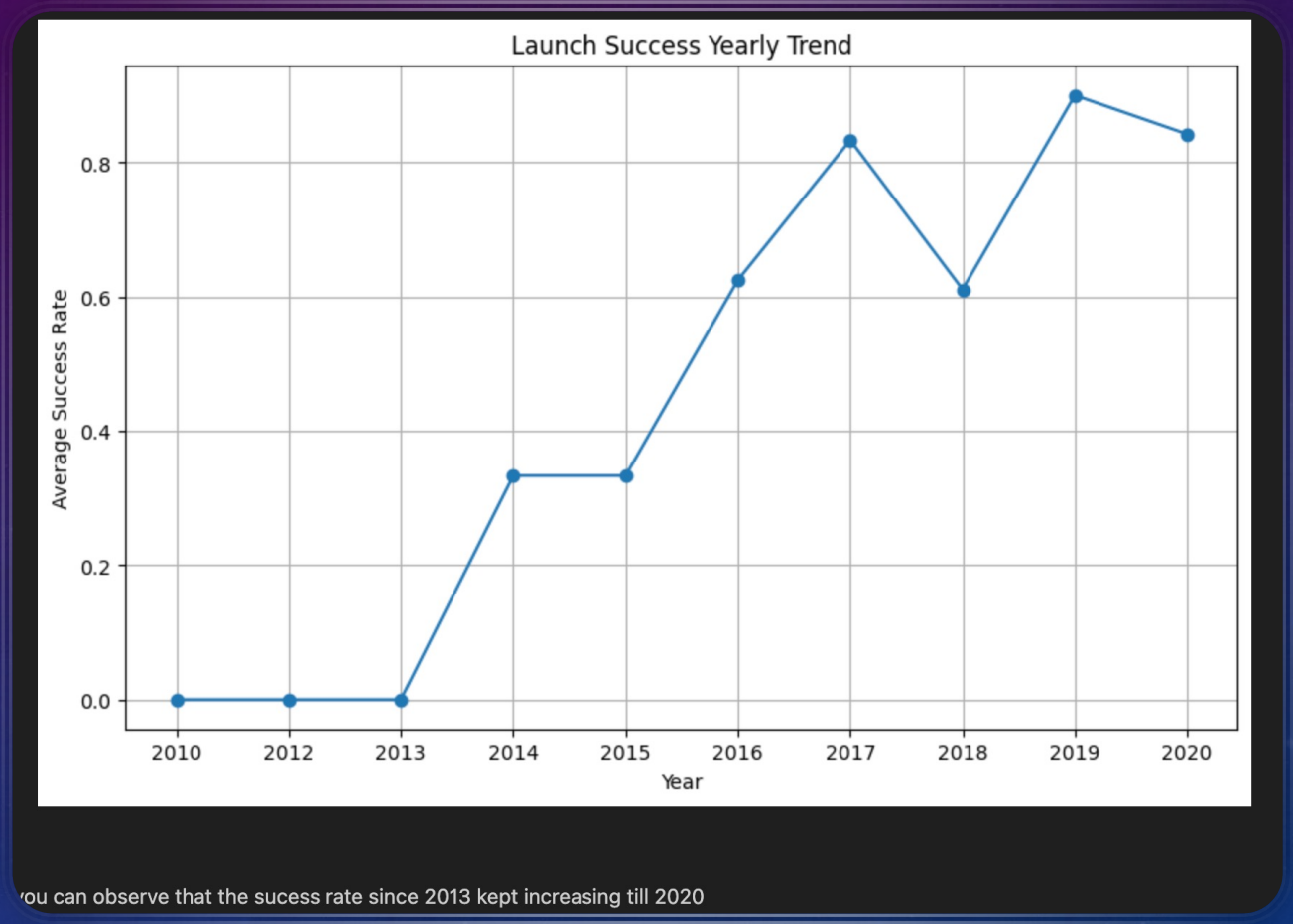
This plot helps visualize the distribution of payload masses across different orbit types and indicates the success rate of landings based on the color coding of data points.



LAUNCH SUCCESS YEARLY TREND

Explanation of the Scatter Plot:

The scatter plot illustrates a clear upward trend in the success rate of launches from 2013 to 2020, indicating improvements in launch technology and processes over the years. Despite slight fluctuations, the overall success rate has significantly increased, highlighting the advancements made by SpaceX in its launch capabilities.



The names of the unique launch sites are:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CAFS SLC-40

ALL LAUNCH SITE NAMES

Date	Time (UTC)	Booster Version	Launch Site
2010-06-04	18:45:00	F9 v1.0	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0	CCAFS LC-40
2012-05-22	07:44:00	F9 v1.0	CCAFS LC-40
2012-10-08	00:35:00	F9 v1.0	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0	CCAFS LC-40

Date: The date of each launch, ranging from 2010 to 2013.

Time (UTC): The specific time of the launch in Coordinated Universal Time.

Booster Version: The version of the Falcon 9 booster used for these launches.

Launch Site: All launches were conducted at the Cape Canaveral Air Force Station Space Launch Complex 40 (CCAFS LC-40).

TOTAL PAYLOAD MASS

•The total payload mass for the customer 'NASA (CRS)' is 45,596 kg. This result was obtained by summing up all the payload masses associated with launches conducted for NASA (Commercial Resupply Services) missions. This value represents the cumulative weight of the payloads launched to space for this specific customer.

AVERAGE PAYLOAD MASS BY F9 V1.1

- The average payload mass carried by the booster version 'F9 v1.1' is approximately 2,928.4 kg. This value was calculated by taking the mean of all the payload masses associated with launches conducted using the 'F9 v1.1' booster version. This average provides insight into the typical payload capacity handled by this particular booster version during its operational flights.

FIRST SUCCESSFUL GROUND LANDING DATE

First Successful Landing Date on Ground Pad:

- The SQL query result shows that the first successful landing date on a ground pad is December 22, 2015.
- This information is crucial as it marks the beginning of successful recoveries of the first stage, contributing to cost savings and reusability.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

Boosters with Specific Success Criteria:

- The boosters which have successfully landed on a drone ship and had payload masses between 4000 and 6000 kg include F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.
- These specific boosters are notable for meeting the criteria of successful drone ship landings while carrying substantial payload masses.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

The query result provides the total number of successful and failure mission outcomes for SpaceX launches:

- **Success**: 98
- **Success (payload status unclear)**: 1
- **Success**: 1
- **Failure (in flight)**: 1
- This indicates that out of the recorded launches, 100 missions were successful (including those with unclear payload status), and only 1 mission was a failure in flight.

Booster Versions that Carried the Maximum Payload Mas

- 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

BOOSTERS CARRIED MAXIMUM PAYLOAD

2015 LAUNCH RECORDS

2015 Drone Ship Landing Failures:

- **January:**
  - **Booster Version:** F9 v1.1 B1012
  - **Launch Site:** CCAFS LC-40
- **April:**
  - **Booster Version:** F9 v1.1 B1015
  - **Launch Site:** CCAFS LC-40

RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

Ranking of Landing Outcomes Between 2010-06-04 and 2017-03-20:

The table displays the various landing outcomes for SpaceX launches within the specified period, ranked by their frequency:

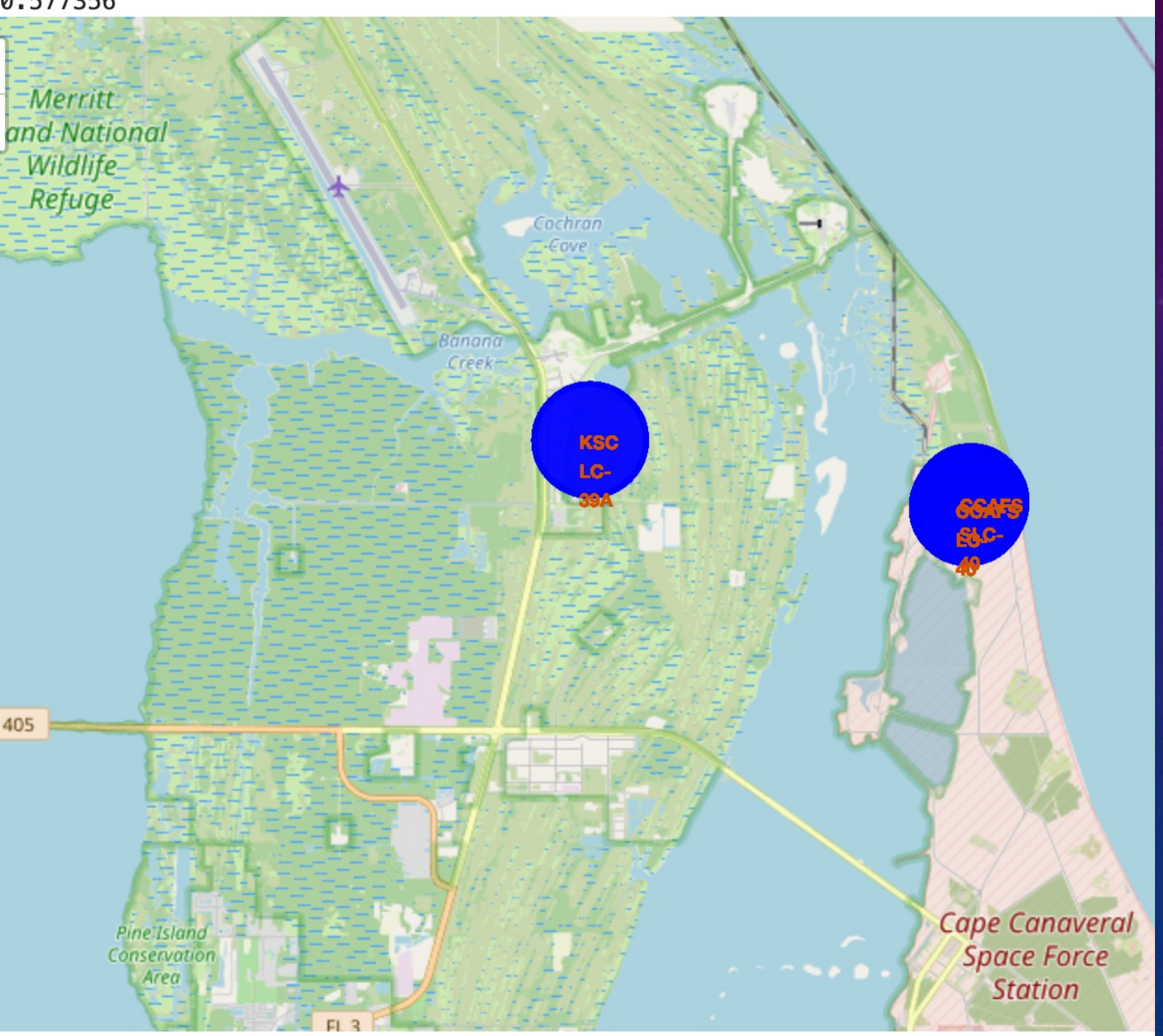
- **No attempt:** 10 instances where no landing attempt was made.
- **Success (drone ship):** 5 instances where the landing was successful on a drone ship.
- **Failure (drone ship):** 5 instances where the landing attempt on a drone ship failed.
- **Success (ground pad):** 3 instances where the landing was successful on a ground pad.
- **Controlled (ocean):** 3 instances where the landing was controlled in the ocean.
- **Uncontrolled (ocean):** 2 instances where the landing was uncontrolled in the ocean.
- **Failure (parachute):** 2 instances where the landing attempt using a parachute failed.
- **Precluded (drone ship):** 1 instance where the landing attempt on a drone ship was precluded.


Section 3
Launch Sites
Proximities Analysis

Launch Sites’ Location Markers on a Global Map

Map Overview:

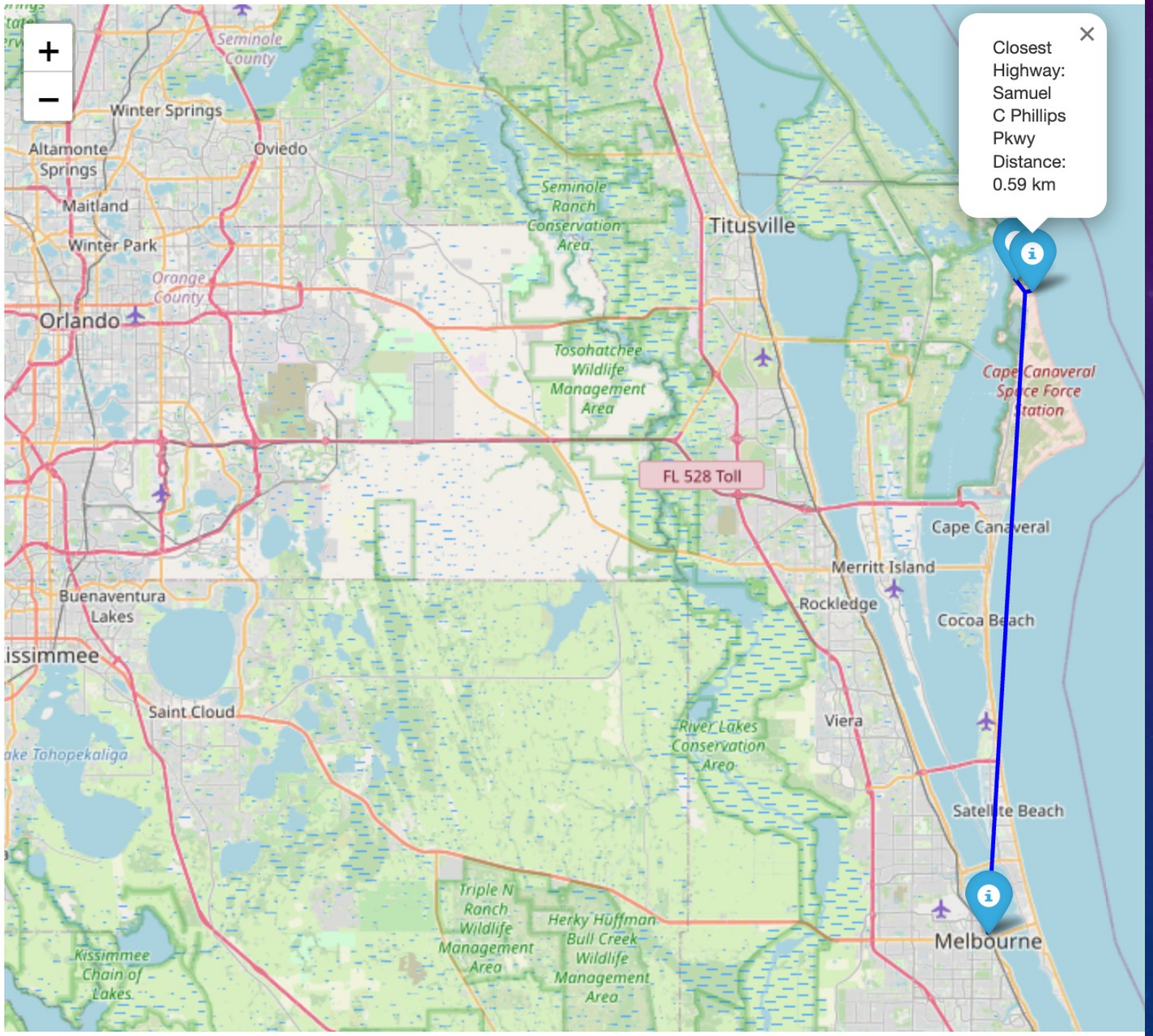
- The map displays the United States, focusing on the geographical distribution of key launch sites on both the East and West coasts.



Color-labeled Launch Outcomes on the Map

Purpose and Findings:

- The map provides a visual representation of the locations of major launch complexes at Cape Canaveral.
- Such a map is useful for understanding the spatial distribution of launch facilities and their surrounding infrastructure.
- The close proximity of these sites facilitates shared resources and logistics, benefiting space launch operations.

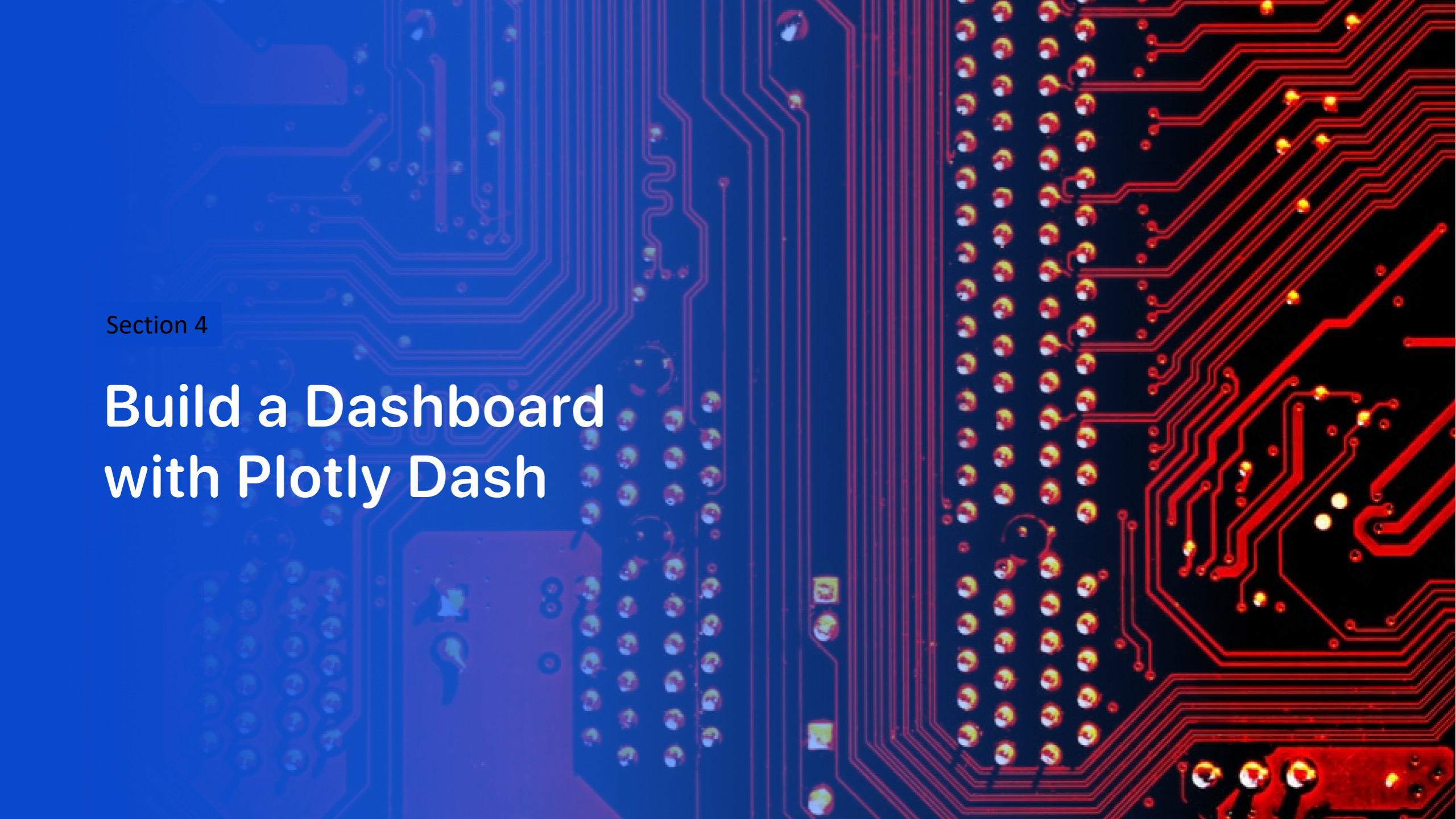


Selected Launch Site to its Proximities

The map represents the distance to the nearest highway which suggests ease of access or connectivity to the space launch facility.

Section 4

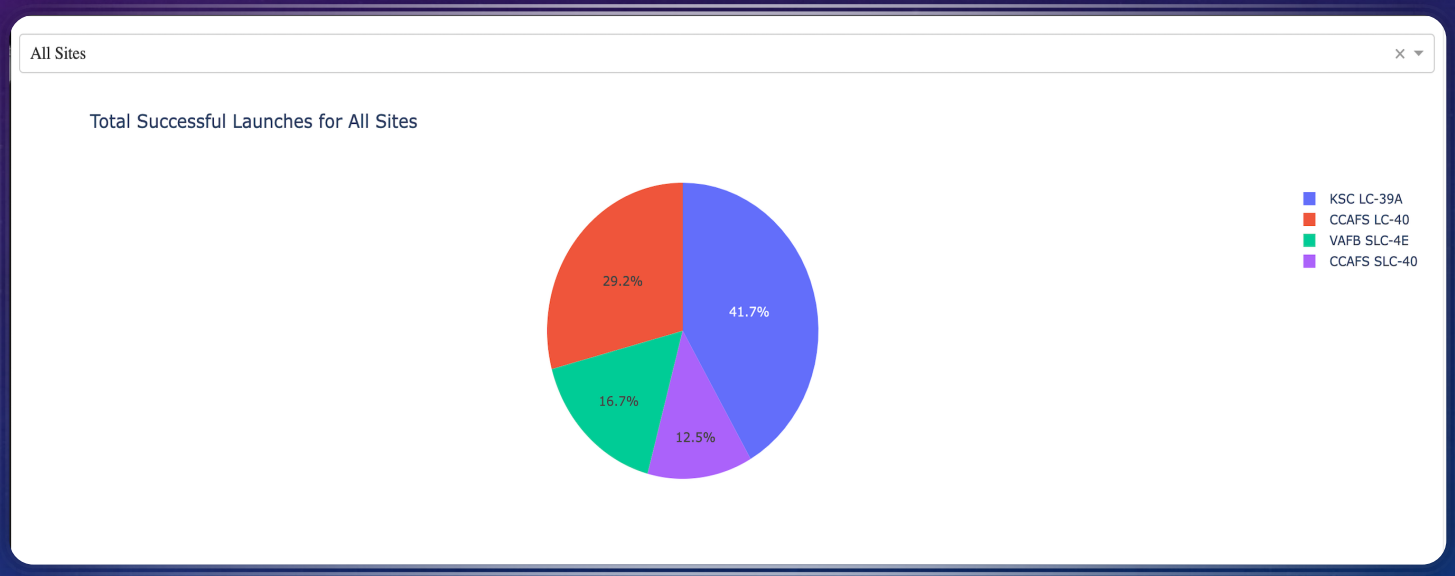
Build a Dashboard with Plotly Dash



Launch Success Count for all Sites

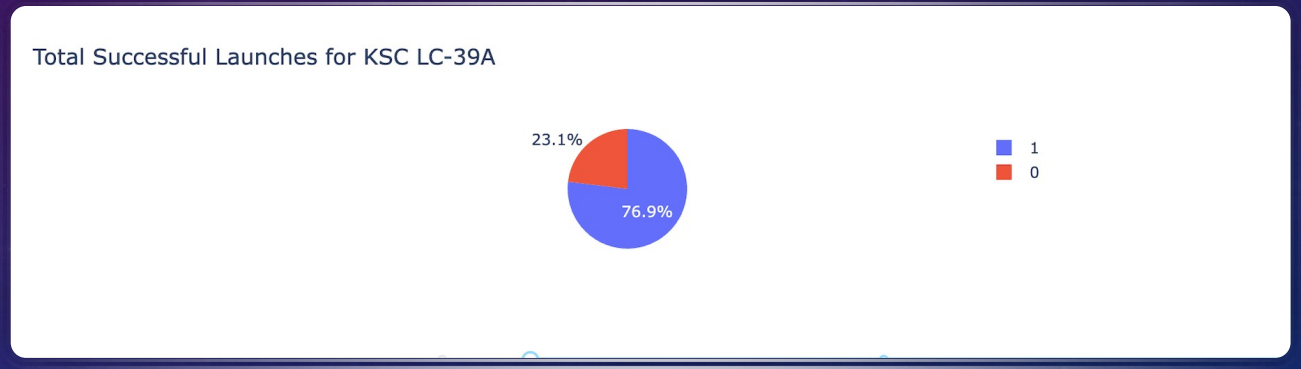
Findings:

- **KSC LC-39A**: This site has the highest number of successful launches, accounting for 41.7% of the total successful launches. This indicates that KSC LC-39A is the most frequently used and possibly the most reliable launch site.
- **CCAFS LC-40**: This site has the second-highest number of successful launches, with 29.2%. It is also a significant launch site for SpaceX.
- **VAFB SLC-4E**: This site accounts for 16.7% of the total successful launches. It is less used compared to KSC LC-39A and CCAFS LC-40 but still plays an important role.
- **CCAFS SLC-40**: This site has the least number of successful launches, with 12.5%. It is the least utilized launch site among the four.



Total Successful Launches for KSC LC-39A

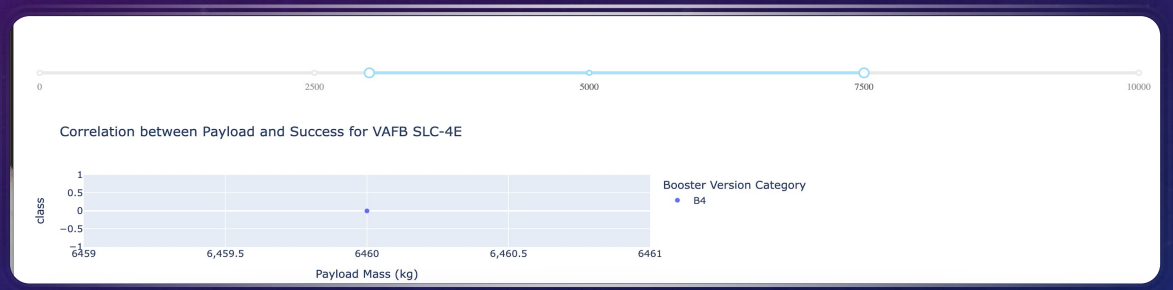
KSC LC-39A: Has the highest launch success rate of 76.9%, making it the most reliable launch site.

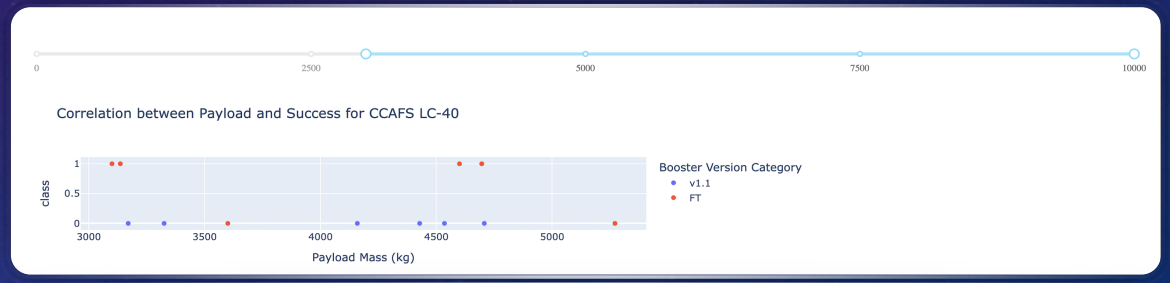


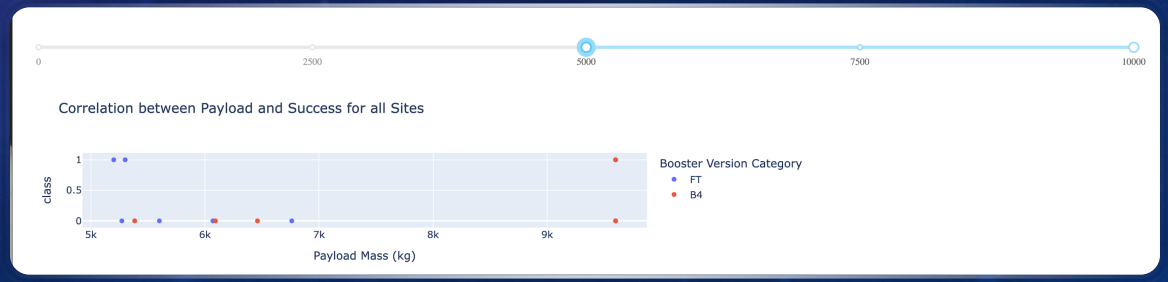
Payload vs. Launch Outcome

Key Findings:

- **Payload Mass**: There is no clear correlation between payload mass and landing success across the different sites. The success rate does not seem to be strongly influenced by the payload mass within the observed ranges.
- **Booster Version**: The FT booster version generally shows a higher success rate compared to the B4 version. This is evident from the higher number of successful landings in the data points for FT boosters.



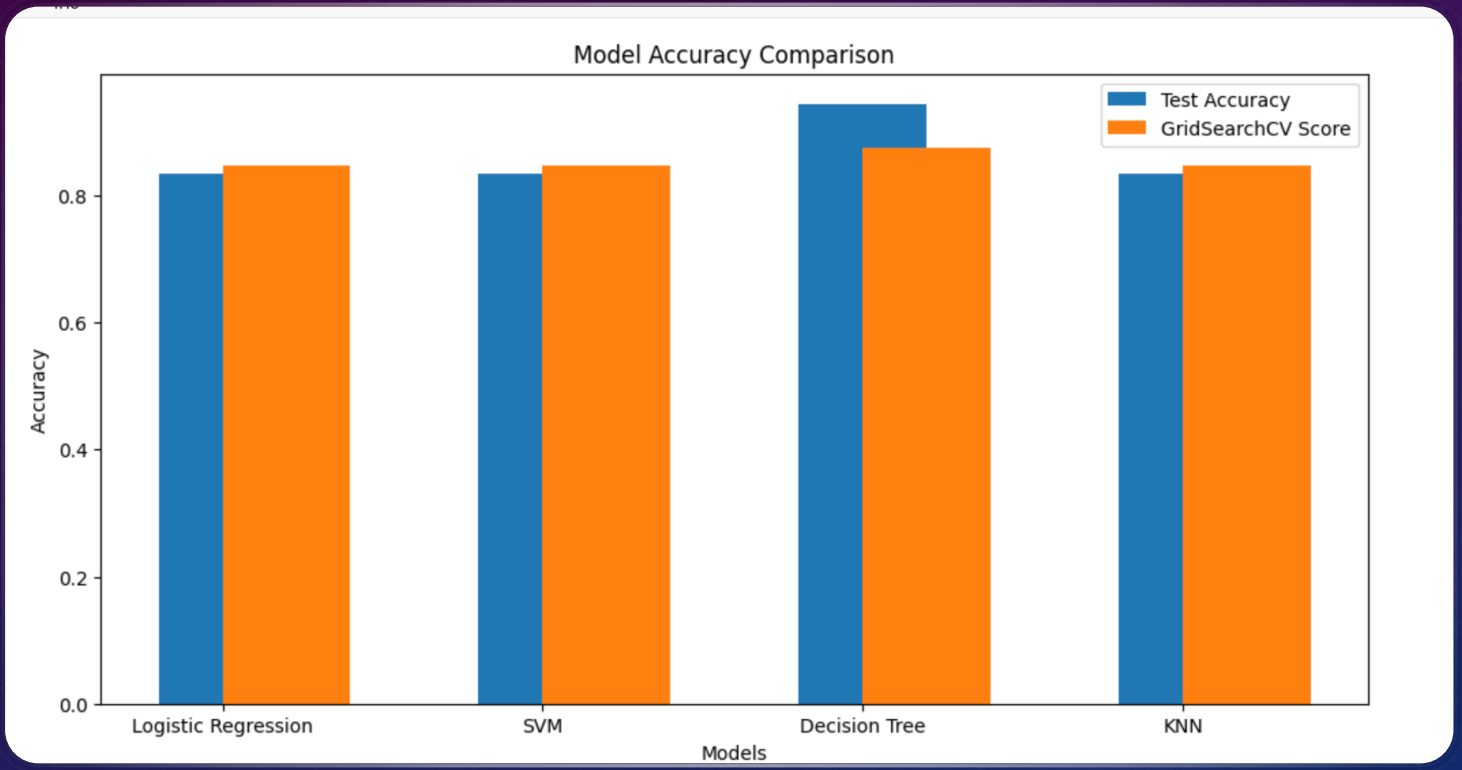




Section 5

Predictive Analysis (Classification)

CLASSIFICATION ACCURACY



Accuracy Test Scores:

• Logistic Regression Accuracy Test Score: 0.8333 Support Vector Machine Accuracy Test Score: 0.8333 Decision Tree Accuracy Test Score: 0.9444

• K-Nearest Neighbors Accuracy Test Score: 0.8333

GridSearchCV Scores:

•Logistic Regression GridSearchCV Score: 0.8464

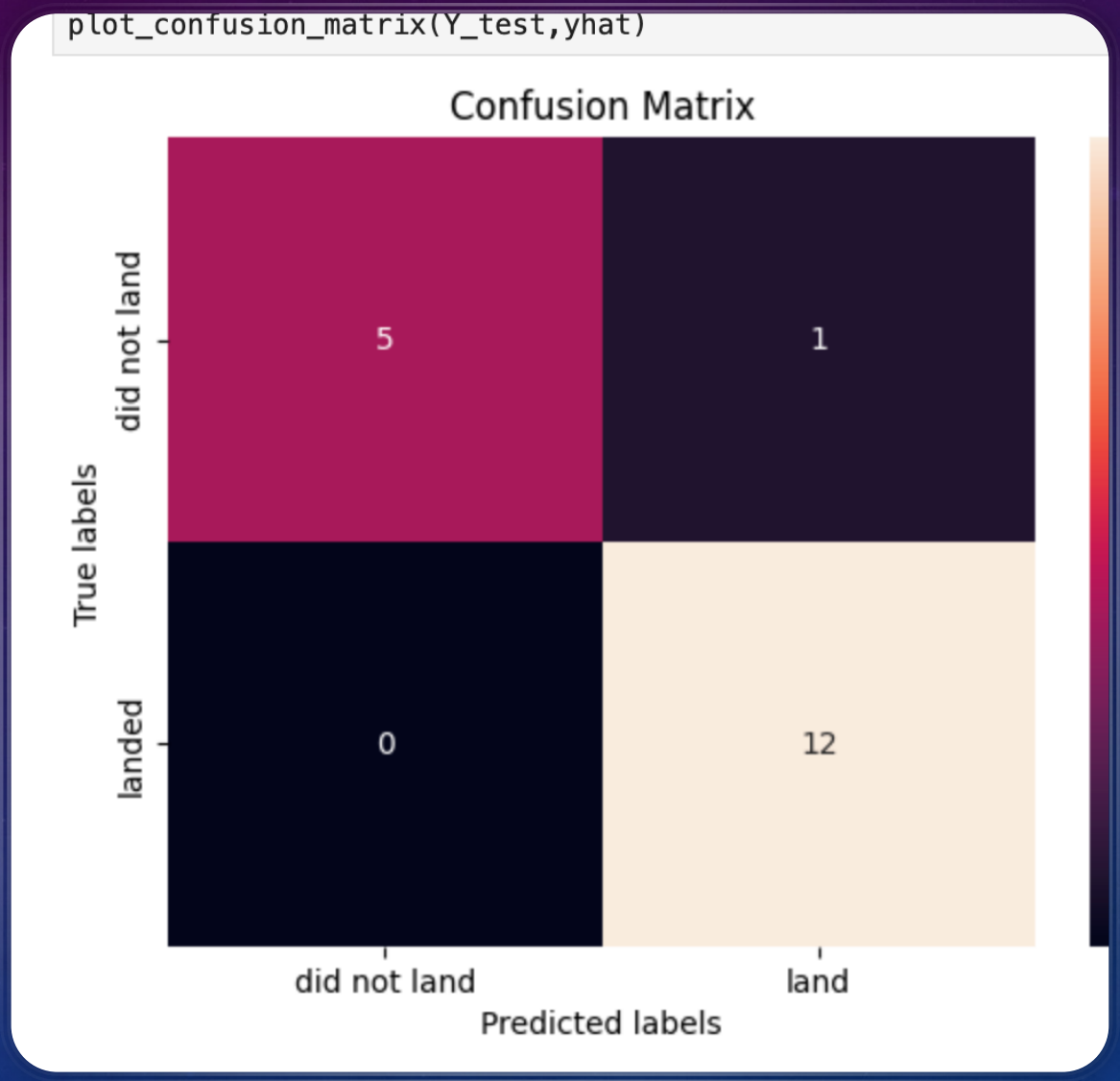
•Support Vector Machine GridSearchCV Score: 0.8482 Decision Tree GridSearchCV Score: 0.875

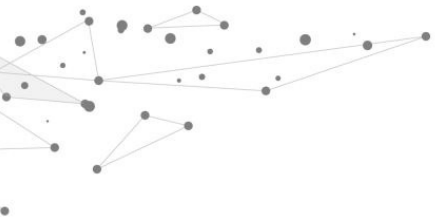
•K-Nearest Neighbors GridSearchCV Score: 0.8482

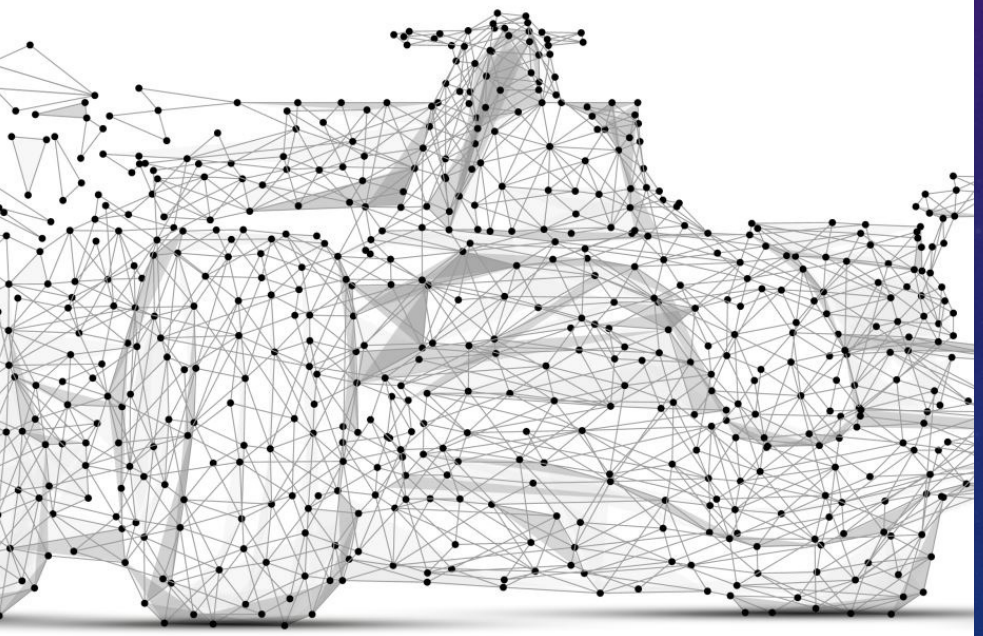
From these results: The Decision Tree model is the best-performing model based on both the accuracy test score and the GridSearchCV score.

CONFUSION MATRIX

Interpretation: High Accuracy: The high accuracy of 94.44% indicates that the Decision Tree model performs very well in predicting the landing outcome of the Falcon 9 first stage. High Recall for Landed Class: The perfect recall for the landed class means that the model is very effective at identifying successful landings. This is crucial for the project's goal as it reduces the risk of missing a successful landing. Few Errors: The few errors made by the model are primarily false positives, where the model predicted a landing when it did not occur. This is preferable over false negatives in this context, as false positives still indicate a potential for a successful landing.







CONCLUSIONS

Conclusion

The predictive model developed in this project successfully forecasts the success of Falcon 9 first stage landings.

The primary answer we gain from all the results is that **the predictive model can reliably forecast the success of Falcon 9 first stage landings**, which has significant implications for cost estimation, competitive bidding, and strategic planning in the space launch industry.

Thank you!