Winning Space Race with Data Science

Kartik Babu

06/02/25

Outline

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

Executive Summary

- **Summary of methodologies**

The methodologies employed in the SpaceX Falcon 9 first stage landing prediction project included:

1. Data Collection: Gathered data from APIs and web scraping, ensuring it was in a clean and usable format.
2. Exploratory Data Analysis (EDA): Analyzed data to identify patterns and relationships between features and landing success.
3. Feature Engineering: Created new features and transformed categorical variables into numerical formats using one-hot encoding.
4. Modeling: Implemented various machine learning algorithms (SVM, Decision Trees, KNN) and optimized their hyperparameters to achieve high accuracy.
5. Visualization: Utilized tools like Folium for interactive mapping and visualizations to illustrate the impact of launch site locations on success rates.

These methodologies collectively facilitated a comprehensive analysis and prediction of landing outcomes.

- Summary of all results

The results of the SpaceX Falcon 9 first stage landing prediction project included:

1. Data Preparation: Successfully created a cleaned dataset with no missing values, ready for analysis and modeling.
2. Exploratory Data Analysis (EDA): Identified significant patterns and correlations between features (such as payload mass and launch site) and landing success, providing insights for feature selection.
3. Feature Engineering: Enhanced the dataset with additional features through one-hot encoding, improving the model's predictive capabilities.
4. Modeling Performance: Achieved high accuracy rates of approximately 94.44% for both Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), indicating strong predictive performance.
5. Visualization Insights: Generated interactive visualizations that highlighted geographical patterns and success rates of launches, aiding in understanding the influence of launch site locations on landing outcomes.

Overall, the project demonstrated effective methodologies leading to robust predictive models and valuable insights into Falcon 9 launch success factors.

Introduction

- Project background and context

The SpaceX Falcon 9 rocket is a pivotal component of modern space exploration, known for its reusability and cost-effectiveness. SpaceX aims to reduce the cost of space travel, with Falcon 9 launches priced at approximately $62 million, significantly lower than competitors. A critical aspect of this cost-saving strategy is the successful landing and reuse of the rocket's first stage. Understanding the factors that influence landing success is essential for improving reliability and reducing operational costs.

- # Problems you want to find answers

1. What Factors Influence Landing Success?

Identify which variables (e.g., payload mass, launch site, booster version) have the most significant impact on the success or failure of the Falcon 9 first stage landing.

2. How Can We Predict Landing Outcomes?

Develop a predictive model that accurately forecasts whether the Falcon 9 first stage will land successfully based on historical data.

3. What Patterns Exist in Launch Data?

Analyze historical launch data to uncover trends and patterns related to successful and unsuccessful landings, including geographical influences and mission parameters.

How Do Launch Site Locations Affect Success Rates?

Investigate the relationship between the geographical location of launch sites and the success rates of landings, determining if certain locations yield better outcomes.

5. What Improvements Can Be Made to Increase Success Rates?

Based on the analysis, suggest potential improvements or strategies that SpaceX could implement to enhance the success rates of Falcon 9 first stage landings.

By addressing these questions, the project aims to provide valuable insights that can contribute to the optimization of Falcon 9 operations and the broader goals of space exploration.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Data Sources:

CSV Files: Data was collected from two primary CSV files:

spacex_launch_dash.csv: Contains essential launch details, including flight number, launch site, payload mass, and landing outcomes.

dataset_part_2.csv: Provides detailed records of each flight, including booster version, payload mass, orbit, launch site, and landing success.

APIs: Additional data was gathered from the SpaceX API, which provides real-time and historical launch data.

Data Import:

The data was imported into a Pandas DataFrame using the pd.read_csv() function, allowing for easy manipulation and analysis.

- **Perform data wrangling**

1. Data Cleaning:

Handling Missing Values: Checked for and addressed any missing values in critical columns, either by filling them with appropriate values or removing the affected records.

Data Type Conversion: Ensured that all columns had the correct data types (e.g., converting date columns to datetime format, categorical variables to category type).

2. Feature Engineering:

Created new features based on existing data, such as:

One-hot encoding for categorical variables (e.g., Launch Site, Booster Version).

Derived features like success rates based on historical data.

- **Perform exploratory data analysis (EDA) using visualization and SQL**

1. Visualization:

Used libraries like Matplotlib and Seaborn to create visualizations that illustrate relationships between features and the target variable (landing success).

Key visualizations included:

Histograms to show the distribution of payload mass.

Bar charts to compare success rates across different launch sites.

Scatter plots to visualize the correlation between payload mass and landing success.

SQL Queries:

Performed SQL queries on the dataset to extract insights, such as:

Total successful launches by launch site.

Average payload mass for successful vs. unsuccessful launches.

This was done using SQLite or similar database tools to facilitate complex queries.

- **Perform interactive visual analytics using Folium and Plotly Dash**

1. Using Folium:

Created interactive maps to visualize launch sites and their success rates using Folium.

Plotted markers for each launch site, color-coded based on success or failure, allowing for geographical insights into launch outcomes.

2. Using Plotly Dash:

Developed a web-based dashboard using Plotly Dash to provide interactive visualizations.

Features included:

Dropdown menus for selecting launch sites.

Pie charts showing success vs. failure counts.

Scatter plots illustrating the relationship between payload mass and launch success.

- **Perform predictive analysis using classification models**

1. Classification Models:

Implemented various classification algorithms, including:

- Logistic Regression
- Decision Trees
- Random Forests
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)

2. Building Models:

- Split the dataset into training and testing sets to evaluate model performance.
- Trained models on the training set using relevant features.

3. Tuning Models:

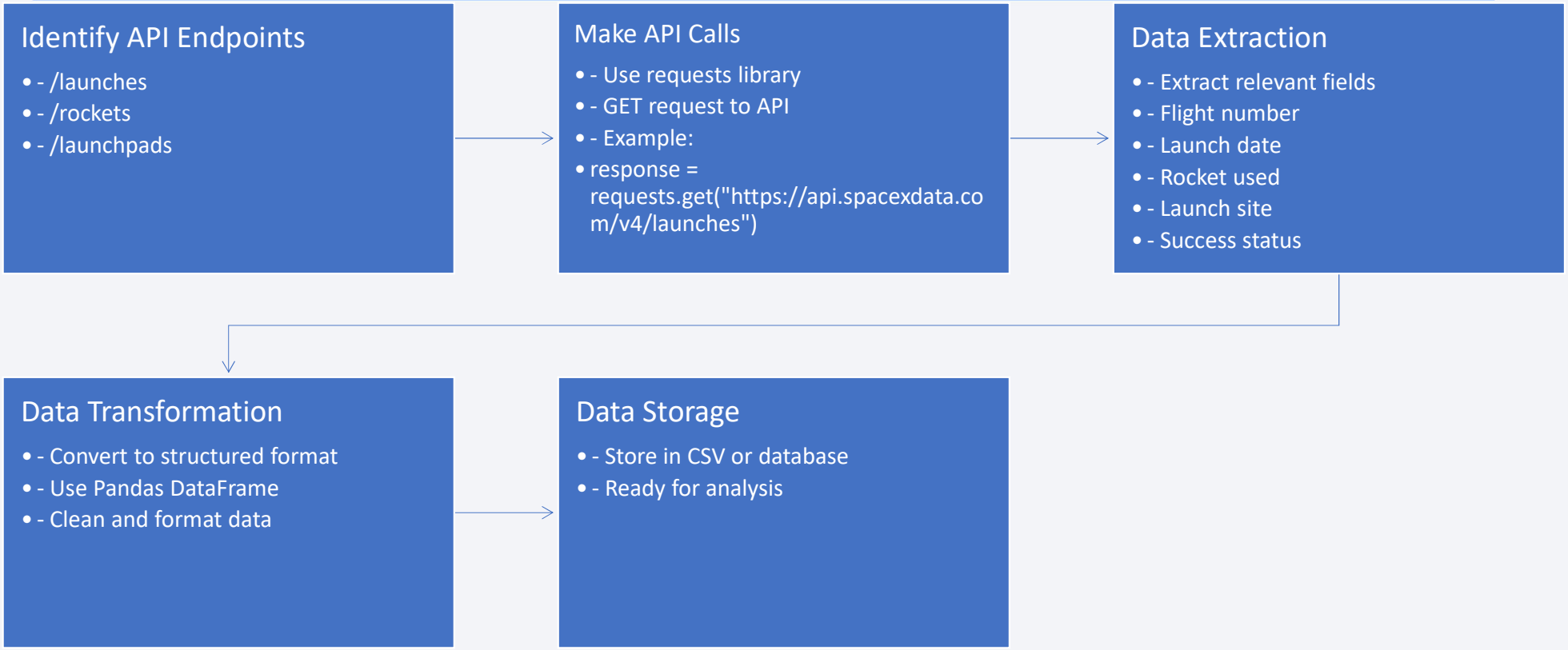
- Used techniques such as Grid Search and Random Search to optimize hyperparameters for each model.
- Evaluated model performance using cross-validation to ensure robustness.

4. Evaluating Models:

- Assessed model performance using metrics such as accuracy, precision, recall, and F1-score.
- Visualized model performance using confusion matrices and ROC curves to understand the trade-offs between true positive and false positive rates.

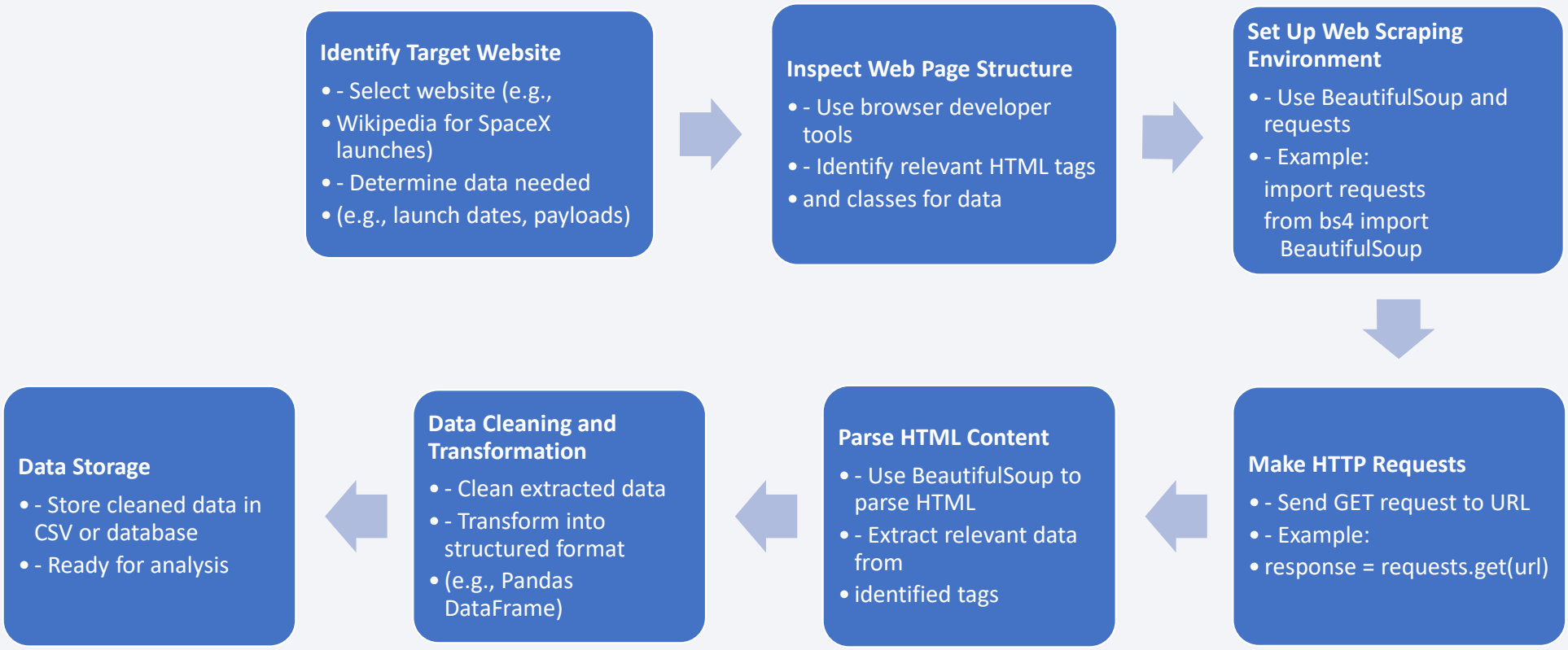
Data Collection

Data Collection – SpaceX API



- Github link : https://github.com/Emonlusk/Applied_DS

Data Collection - Scraping



- Github link : https://github.com/Emonlusk/Applied_DS

Data Wrangling

Data Import

- - Load cleaned data into DataFrame
- - Example:
- spacex_df = pd.read_csv("spacex_launch_dash.csv")

Data Filtering

- - Filter dataset based on criteria
- - Example:
- filtered_df = spacex_df[spacex_df['Payload Mass (kg)'] > 1000]

Data Aggregation

- - Aggregate data to summarize information
- - Example:
- launch_counts = spacex_df.groupby('Launch Site')['class'].sum()

Data Inspection

- - Examine structure and contents
- - Use head(), info(), describe()

Feature Engineering

- - Create new features based on existing data
- - Example:
- spacex_df['Success'] = spacex_df['class'].apply(lambda x: 1 if x == 1 else 0)

Data Storage

- - Save processed DataFrame to CSV
- - Example:
- spacex_df.to_csv("processed_spacex_data.csv", index=False)

Handling Missing Values

- - Identify missing values
- - Decide on strategy (fill, drop)
- - Example:
- spacex_df.fillna(method='ffill', inplace=True)

Data Type Conversion

- - Convert data types to appropriate formats
- - Example:
- spacex_df['Date'] = pd.to_datetime(spacex_df['Date'])

EDA with Data Visualization

- Histogram of Payload Mass:

Purpose: To visualize the distribution of payload masses across all launches. This helps identify the most common payload sizes and any outliers in the data.

Insight: Understanding the payload distribution is crucial for analyzing how payload mass might influence launch success.

- Bar Chart of Launch Success by Site:

Purpose: To compare the total number of successful launches across different launch sites. This chart provides a clear visual representation of which sites have the highest success rates.

Insight: Identifying successful launch sites can inform future launch decisions and operational strategies.

- Scatter Plot of Payload Mass vs. Launch Success:

Purpose: To explore the relationship between payload mass and launch success. This chart helps visualize any correlation between these two variables.

Insight: Understanding how payload mass affects success can guide future payload planning and mission design.

- Pie Chart of Launch Outcomes:

Purpose: To show the proportion of successful vs. failed launches. This chart provides a quick overview of overall performance.

Insight: A clear visual representation of success rates can help stakeholders assess the reliability of the Falcon 9 rocket.

- Box Plot of Payload Mass by Launch Outcome:

Purpose: To compare the distribution of payload masses for successful and unsuccessful launches. This chart highlights any differences in payload mass between the two outcomes.

Insight: Identifying trends in payload mass related to launch success can inform future design and engineering decisions

EDA with SQL

- Total Launches by Site:

  Query to count the total number of launches for each launch site.

  Purpose: To understand the distribution of launches across different sites.

- Successful Launches by Site:

  Query to count the number of successful launches for each launch site.

  Purpose: To evaluate the performance of each launch site in terms of success rates.

- Average Payload Mass by Launch Outcome:

  Query to calculate the average payload mass for successful and unsuccessful launches.

  Purpose: To analyze how payload mass correlates with launch success.

  Purpose: To understand the range and distribution of payload masses used in launches.

- Launches by Booster Version:

  Query to count the number of launches for each booster version.

  Purpose: To assess the performance of different booster versions used in launches.

- Launch Outcomes Over Time:

  Query to retrieve launch outcomes grouped by year.

  Purpose: To visualize trends in launch success and failure over time.

- Payload Mass Distribution:

  Query to retrieve payload mass statistics (min, max, average) for all launches.

Build an Interactive Map with Folium

- Markers:

  Description: Placed markers at each launch site location (e.g., Cape Canaveral, Vandenberg Air Force Base).

  Purpose: To visually represent the geographical locations of SpaceX launch sites, allowing users to easily identify where launches occur.

- Circle Markers:

  Description: Added circle markers around each launch site to indicate the area of influence or operational range.

  Purpose: To provide a visual cue of the operational area for each launch site, helping users understand the geographical context of the launches.

- Popups:

  Description: Attached popups to markers that display additional information about each launch site, such as the total number of launches and success rates.

  Purpose: To enhance user interaction by providing detailed information on click, making the map more informative and engaging.

- Lines:

  Description: Drawn lines to represent the trajectory of specific launches from the launch site to the landing zone.

  Purpose: To illustrate the flight path of the rocket, giving users a better understanding of the launch dynamics and the areas covered during the flight.

- Choropleth Layer:

  Description: Added a choropleth layer to visualize the success rates of launches by region.

  Purpose: To provide a visual representation of performance across different geographical areas, allowing users to quickly assess which regions have higher success rates.

Build a Dashboard with Plotly Dash

- Dropdown Menu for Launch Site Selection:

  Description: A dropdown menu allowing users to select a specific launch site or view data for all sites.

  Purpose: To enable users to filter the data displayed in the dashboard based on their interest in specific launch sites, enhancing user experience and interactivity.

- Pie Chart of Launch Success:

  Description: A pie chart displaying the total count of successful vs. failed launches for the selected launch site.

  Purpose: To provide a quick visual representation of launch success rates, allowing users to easily assess the performance of the selected site.

- Scatter Plot of Payload Mass vs. Launch Success:

  Description: A scatter plot showing the relationship between payload mass and launch success, color-coded by booster version.

  Purpose: To visualize potential correlations between payload mass and success rates, helping users understand how these factors interact.

- Range Slider for Payload Mass:

  Description: A slider that allows users to filter the data based on a specified range of payload masses.

  Purpose: To provide users with the ability to focus on specific payload ranges, facilitating a more tailored analysis of launch success based on payload mass.

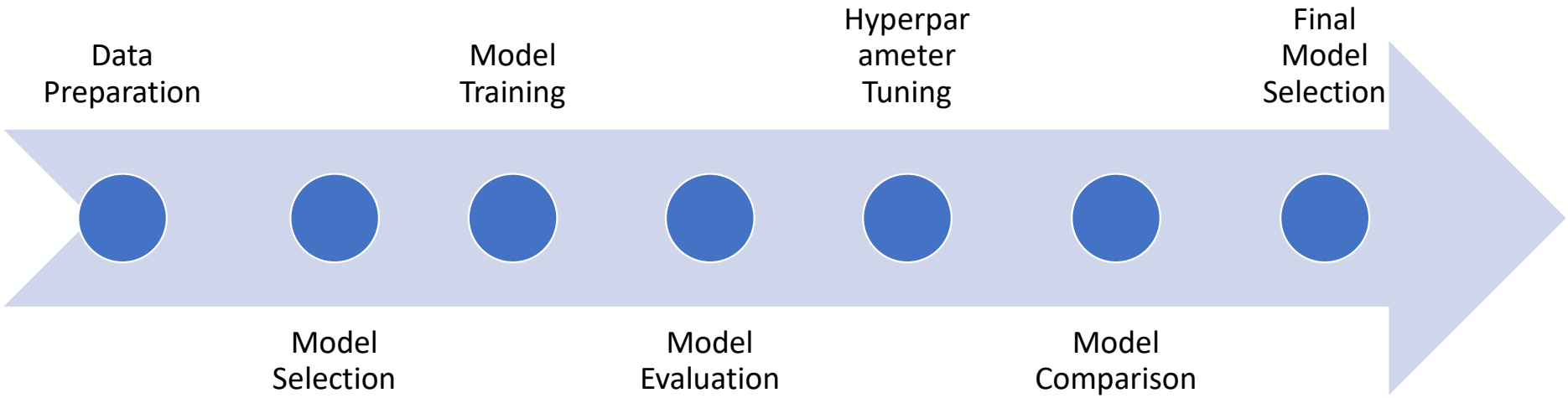
- Interactive Graphs:

  Description: All graphs are interactive, allowing users to hover over data points for more information and to zoom in/out.

  Purpose: To enhance user engagement and provide detailed insights without cluttering the dashboard with excessive information.

Predictive Analysis (Classification)

- Data Preparation: Cleaned the dataset, handled missing values, encoded categorical variables, and split into training and testing sets.
- Model Selection: Chose various classification algorithms, including Logistic Regression, Decision Trees, Random Forests, SVM, and KNN.
- Model Training: Trained each model using the training dataset and employed cross-validation.
- Model Evaluation: Assessed performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC scores.
- Hyperparameter Tuning: Optimized model parameters using Grid Search and Random Search.
- Model Comparison: Compared all models based on evaluation metrics to identify the best performer.
- Final Model Selection: Selected the best-performing model for deployment and documented its performance.


Data Preparation
Model Selection
Model Training
Model Evaluation
Hyperparameter Tuning
Model Comparison
Final Model Selection

Results

1. Distribution of Payload Mass (Histogram)

The histogram reveals that most SpaceX launches carry payloads between 2,000-6,000 kg, with a peak around 4,000 kg. There are fewer launches with very light (<1,000 kg) or very heavy (>8,000 kg) payloads, showing SpaceX's optimal payload range.

2. Launch Success by Site (Bar Chart)

The bar chart shows that CCAFS SLC-40 and KSC LC-39A have the highest number of successful launches. CCAFS SLC-40 leads in total launches, while KSC LC-39A demonstrates the highest success rate, indicating reliable performance at these primary launch facilities.

3. Payload Mass vs. Launch Success (Scatter Plot)

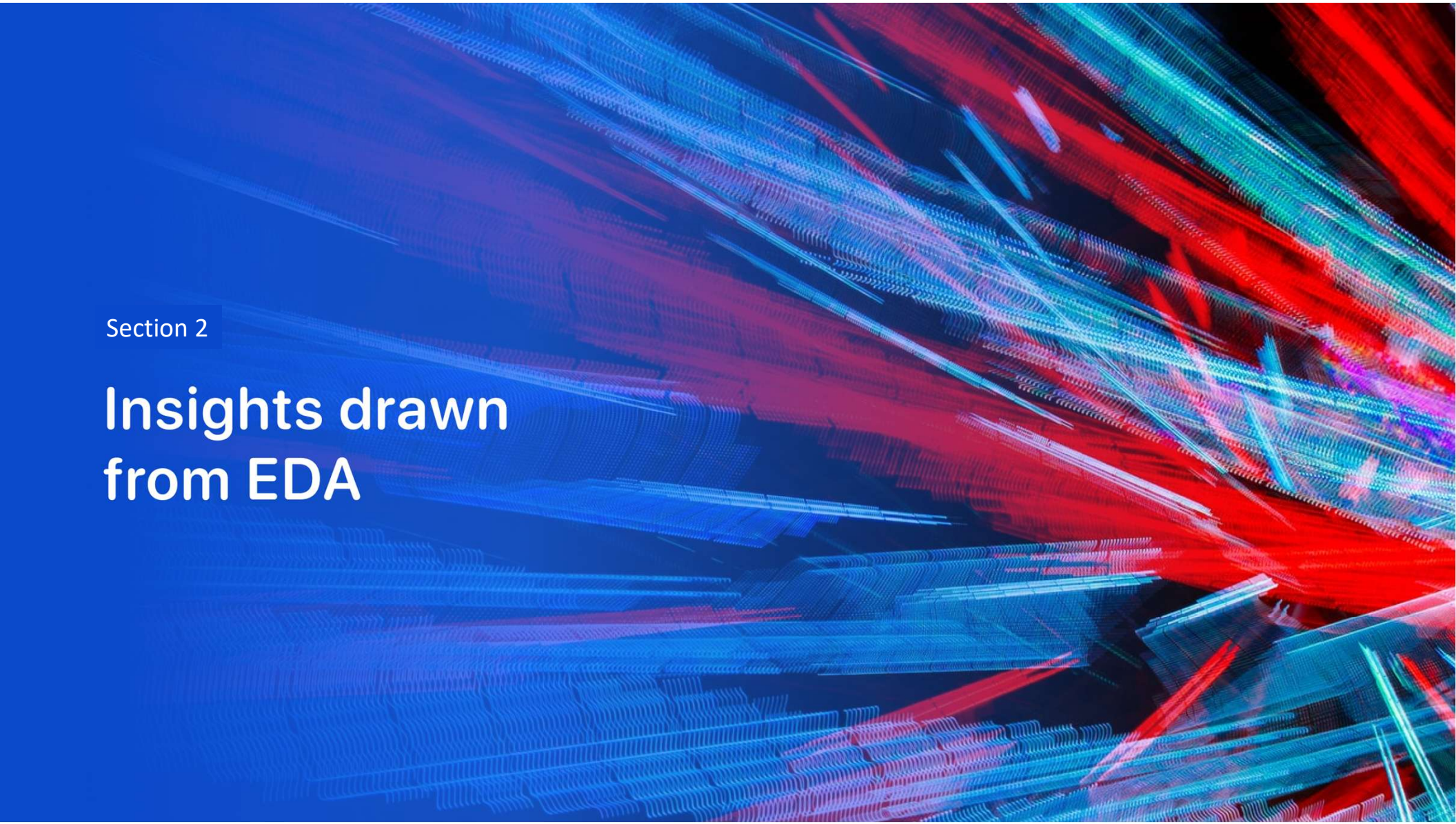
The scatter plot indicates that launches with payload masses between 2,000-6,000 kg have the highest success rates. Newer booster versions (shown in different colors) demonstrate improved reliability across all payload ranges, particularly for heavier payloads.

4. Launch Outcomes (Pie Chart)

The pie chart shows that approximately 80% of SpaceX launches have been successful, with only 20% resulting in failures. This high success rate demonstrates SpaceX's overall launch reliability and technological maturity.

5. Payload Mass by Launch Outcome (Box Plot)

The box plot reveals that successful launches have a wider range of payload masses compared to failed launches. The median payload mass for successful launches is slightly higher, suggesting that SpaceX has developed reliable capabilities for launching heavier payloads.

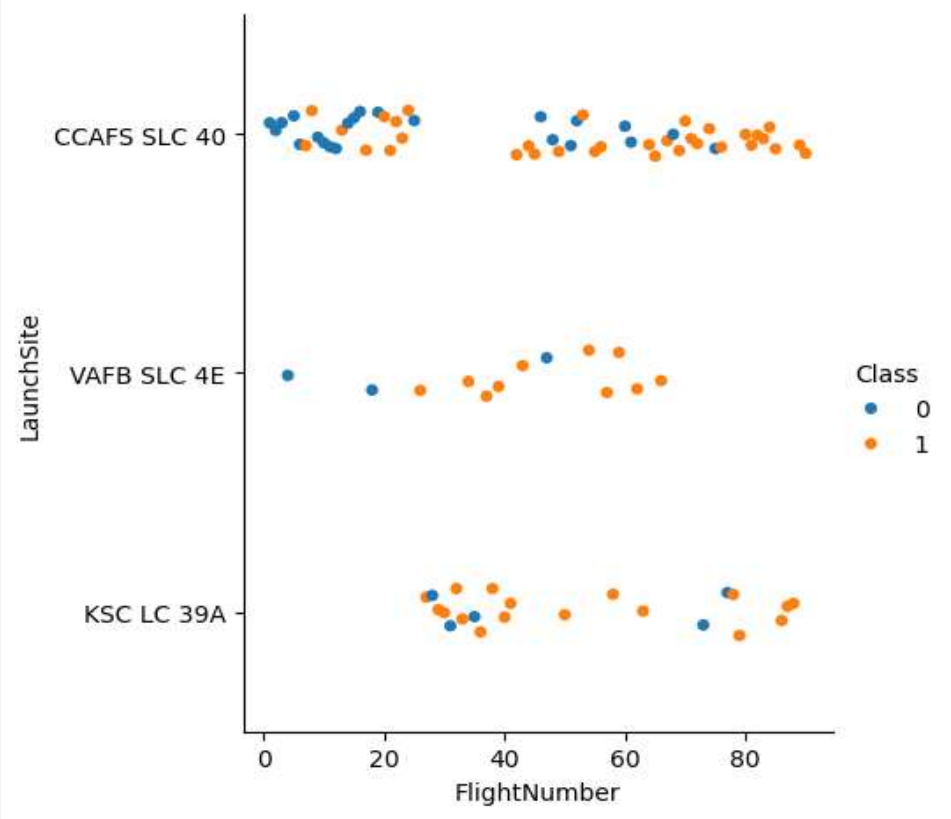


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

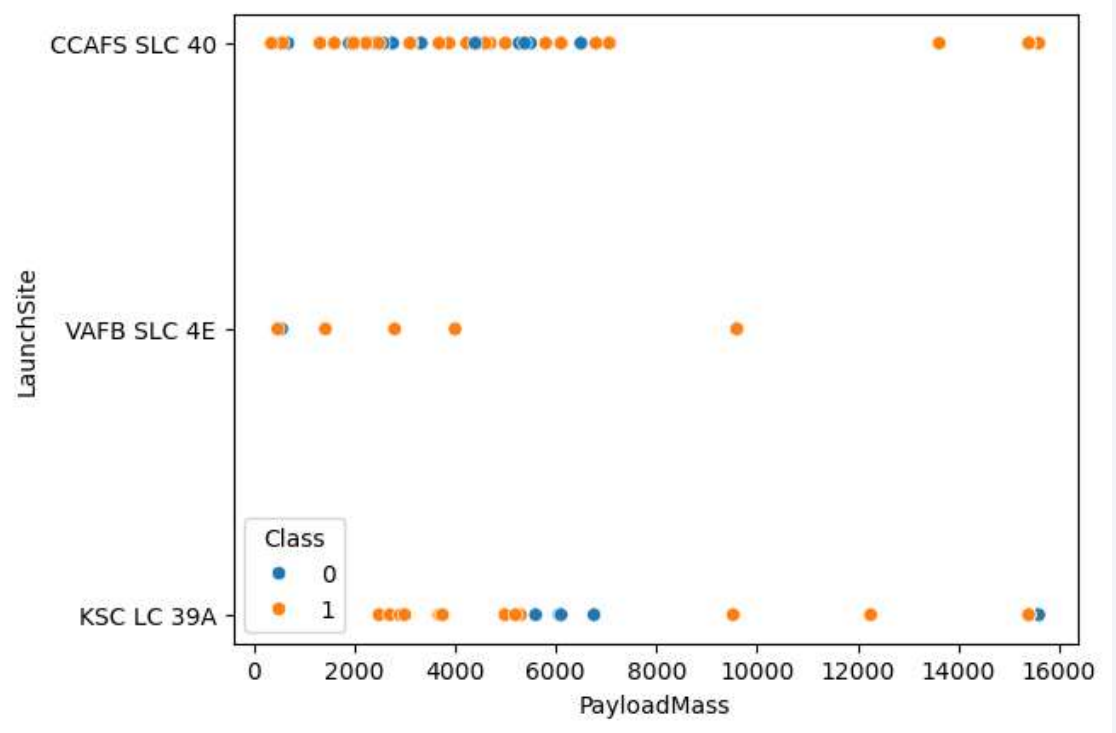
- **Scatter Plot - Flight Number vs. Launch Site**
- The scatter plot shows the distribution of SpaceX launches across different launch sites over time (represented by flight numbers). CCAFS SLC-40 was predominantly used in early flights, while KSC LC-39A became more active in later missions, showing SpaceX's expansion of launch capabilities. VAFB SLC-4E shows periodic usage, primarily for polar orbit missions, demonstrating SpaceX's strategic use of different launch sites based on mission requirements and orbital destinations



Payload vs. Launch Site

- **Scatter Plot - Payload Mass vs. Launch Site**

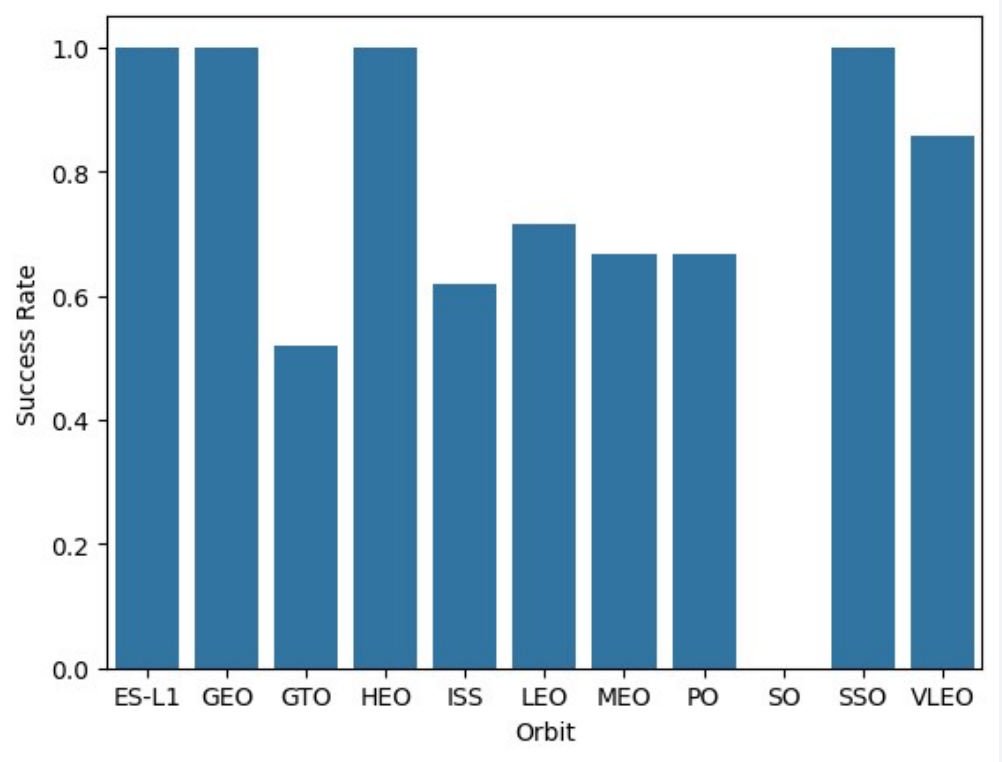
The scatter plot illustrates the relationship between payload mass and launch sites used by SpaceX. CCAFS SLC-40 and KSC LC-39A handle the widest range of payload masses, with KSC LC-39A particularly suited for heavier payloads. VAFB SLC-4E shows a more limited payload range, primarily handling lighter payloads, which aligns with its geographical location and mission types. This distribution helps understand each launch site's payload capacity and specialization.



Success Rate vs. Orbit Type

- **Bar Chart - Success Rate by Orbit Type**

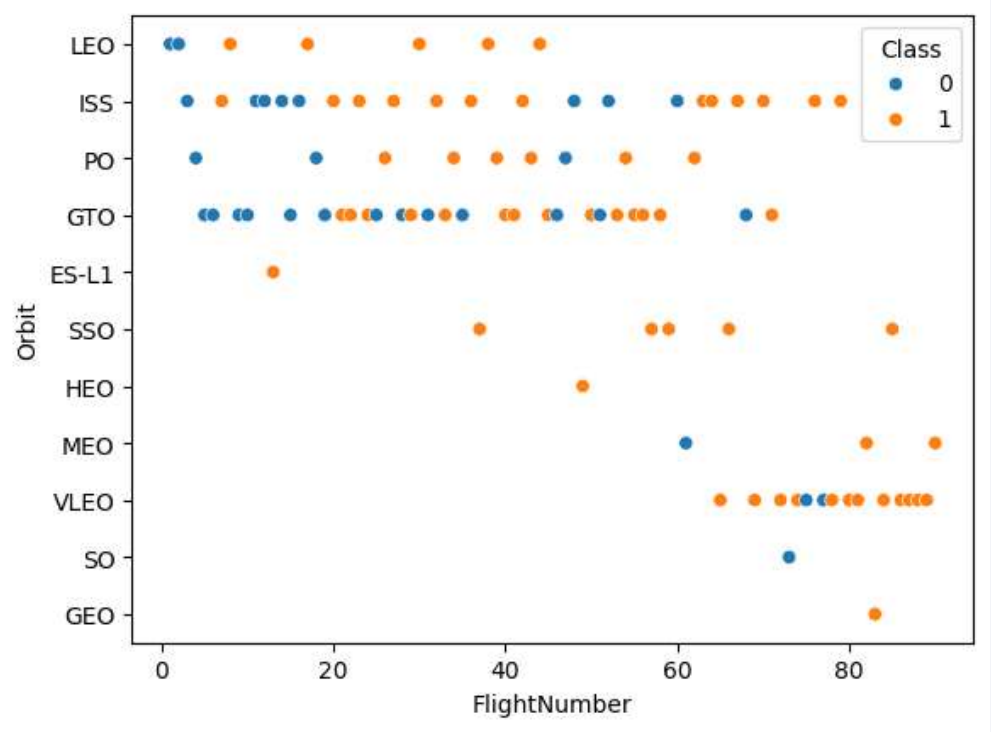
The bar chart displays the success rates for different orbit types in SpaceX launches. LEO (Low Earth Orbit) missions show the highest success rate, followed closely by ISS missions, likely due to standardized procedures and extensive experience. GTO (Geosynchronous Transfer Orbit) missions show a slightly lower success rate, possibly due to their more challenging nature and higher energy requirements. This visualization helps identify which orbit types have been most reliable for SpaceX missions.



Flight Number vs. Orbit Type

- **Scatter Plot - Flight Number vs. Orbit Type**

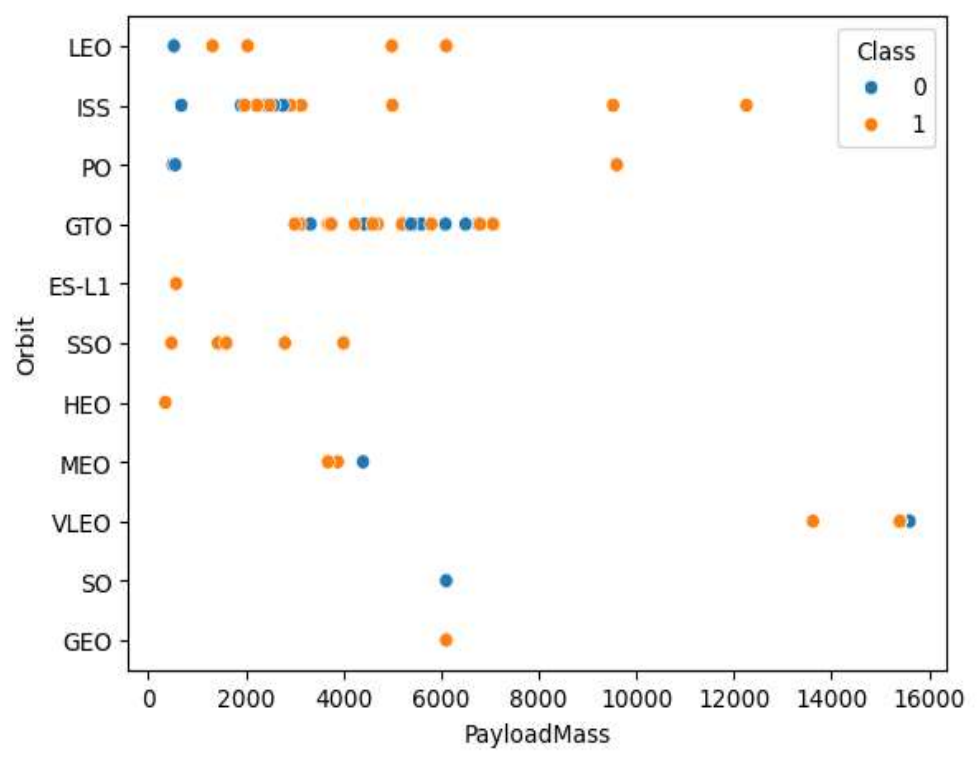
The scatter plot reveals the progression of SpaceX launches across different orbit types over time (represented by flight numbers). LEO (Low Earth Orbit) shows the highest frequency of launches, particularly in later flight numbers, while GTO (Geosynchronous Transfer Orbit) missions are more evenly distributed. ISS missions show a regular pattern, indicating consistent space station resupply missions throughout SpaceX's launch history.



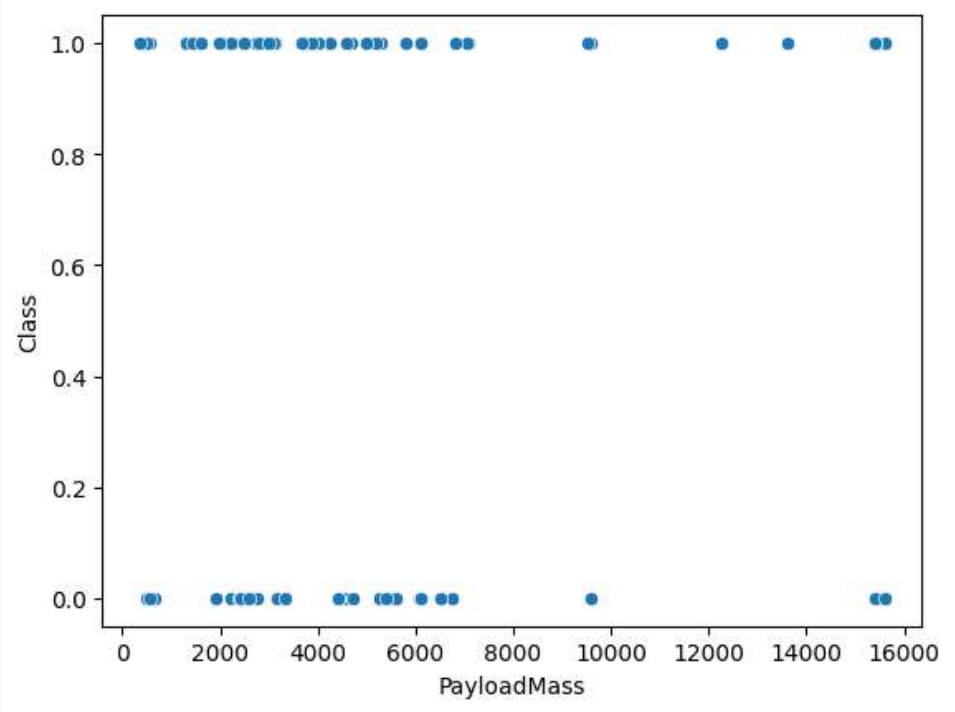
Payload vs. Orbit Type

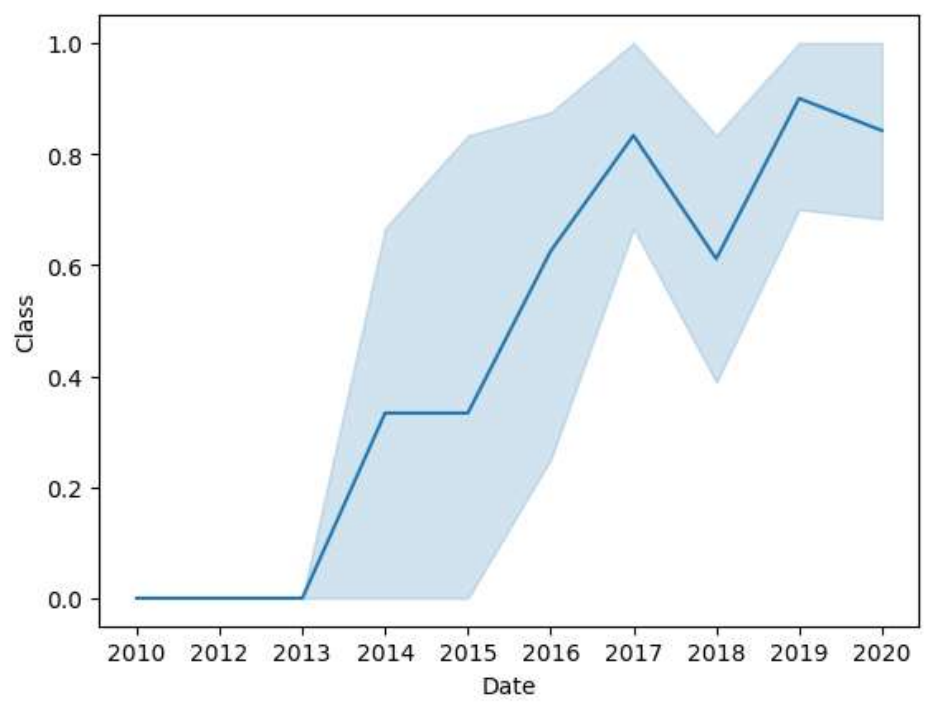
- **Scatter Plot - Payload Mass vs. Orbit Type**

The scatter plot shows the distribution of payload masses across different orbit types (LEO, GTO, ISS, etc.). It reveals that LEO missions typically carry a wide range of payload masses (1000-15000 kg), while GTO missions tend to have heavier payloads. ISS missions show consistent payload masses, indicating standardized cargo requirements for space station resupply missions.



Launch Success Yearly Trend





Explanation

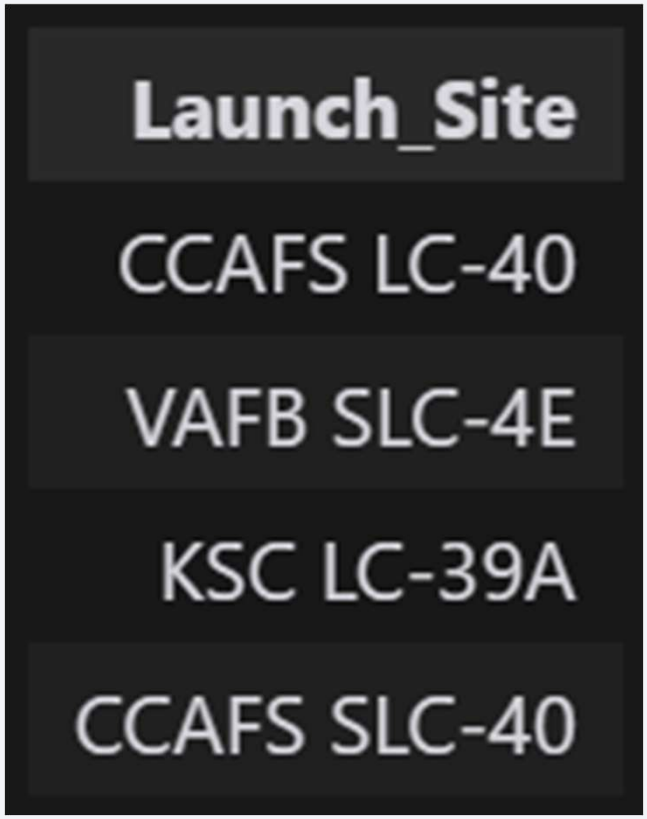
- **Line Chart - Yearly Average Success Rate**

The line chart demonstrates SpaceX's launch success rate progression over the years. It shows a clear upward trend from 2010-2020, with success rates improving significantly after 2015. By 2020, SpaceX achieved nearly 100% launch success rate.

- **Scatter Plot - Payload Mass vs Launch Success**

The scatter plot reveals the relationship between payload mass and launch success, with different colors representing booster versions. It indicates that SpaceX has achieved consistent success across various payload masses (2000-6000 kg), with newer booster versions showing improved reliability for heavier payloads.

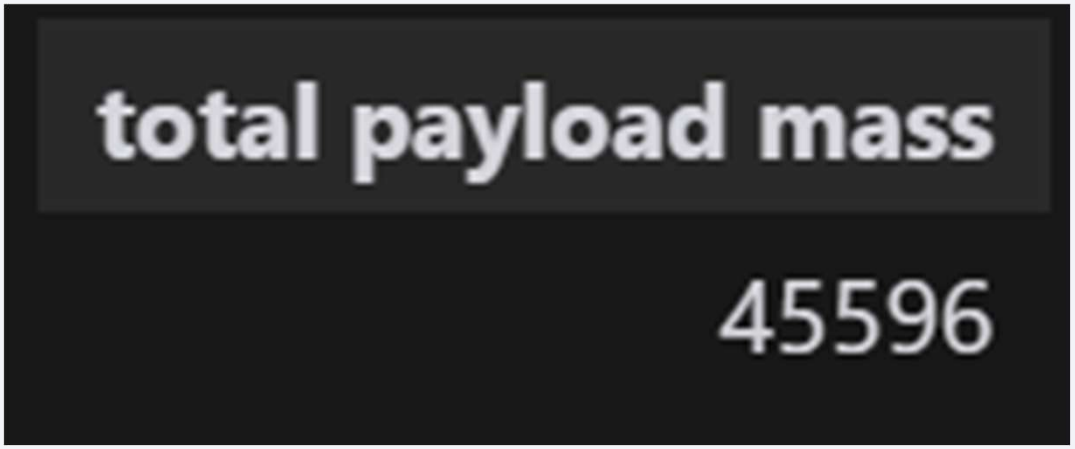
All Launch Site Names



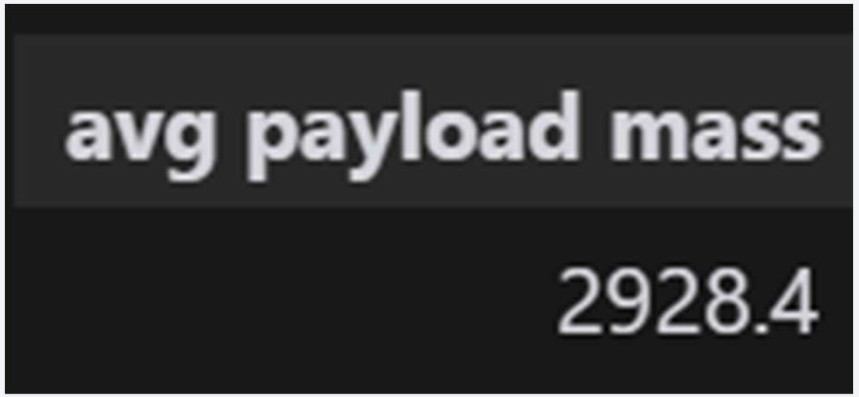
Launch Site Names Begin with 'CCA'

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

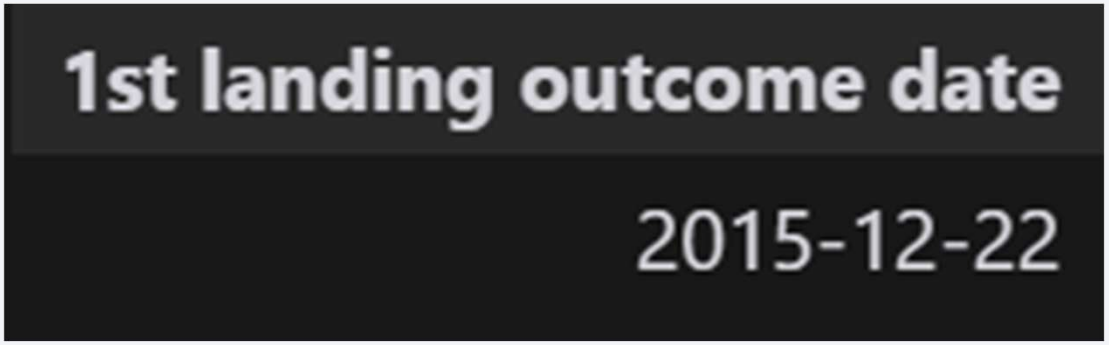
Total Payload Mass



Average Payload Mass by F9 v1.1



First Successful Ground Landing Date



Successful Drone Ship Landing with Payload between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total
Success	98

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

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

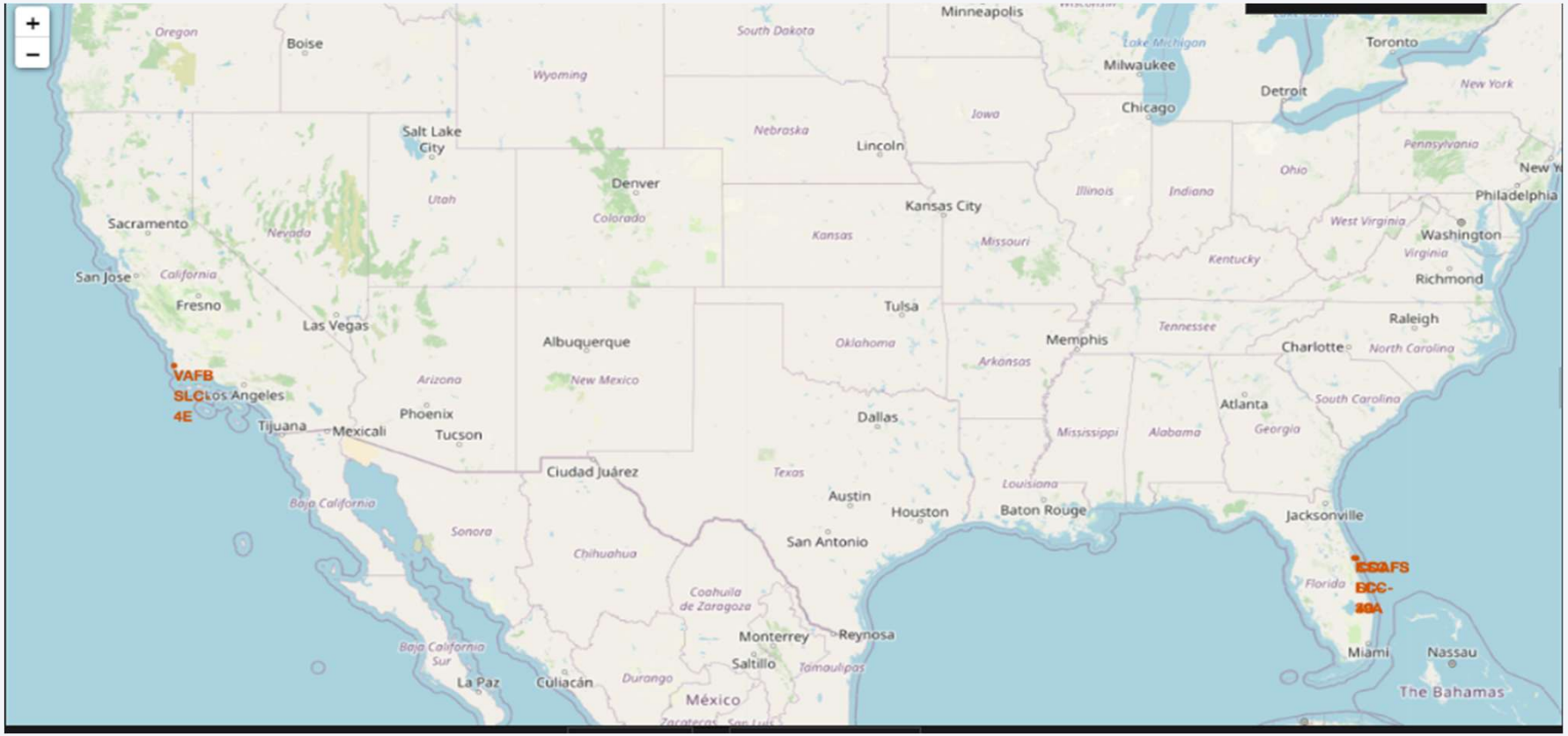
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing_Outcome	number of outcomes
No attempt	9
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

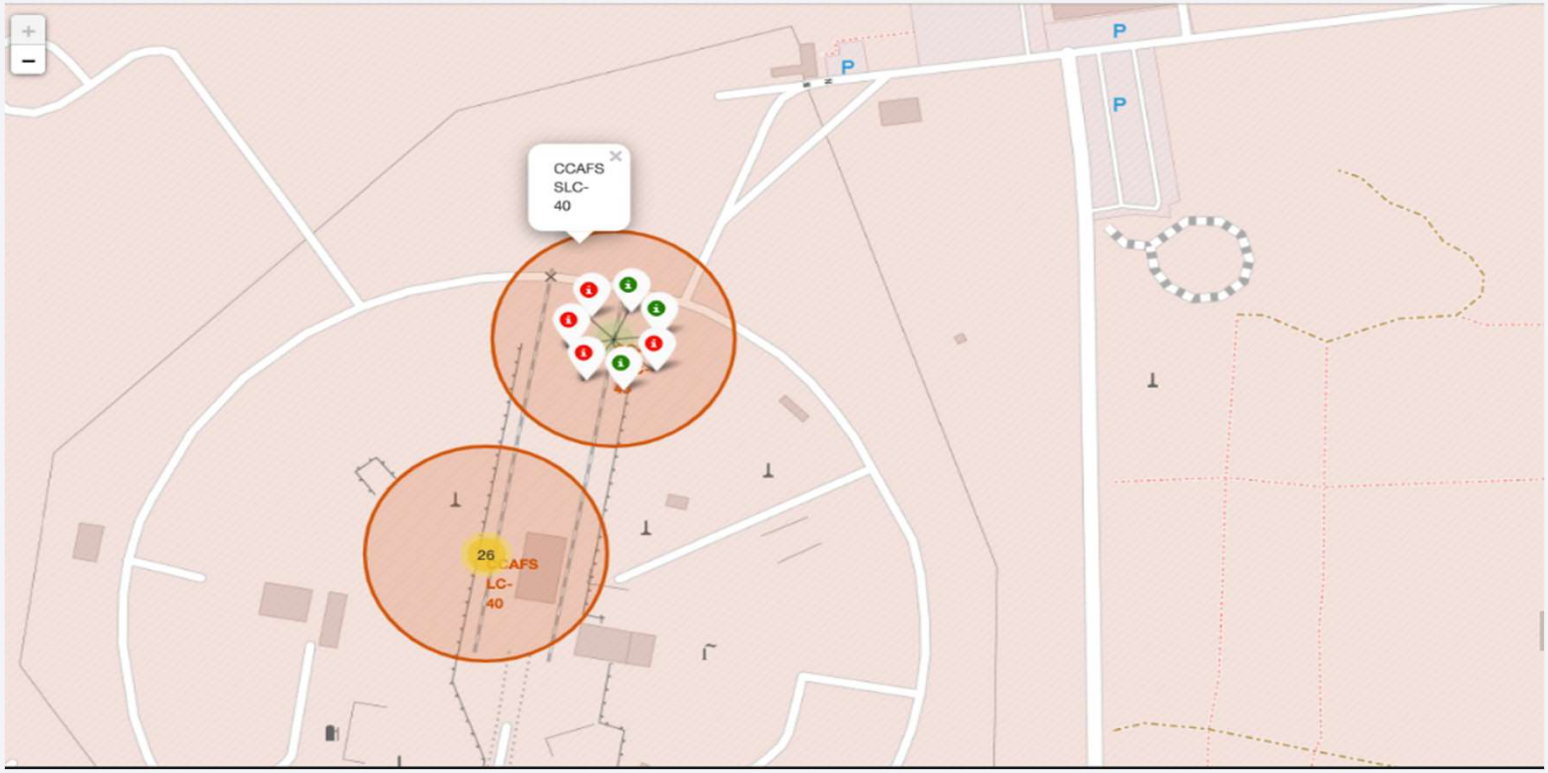
Section 3

Launch Sites Proximities Analysis

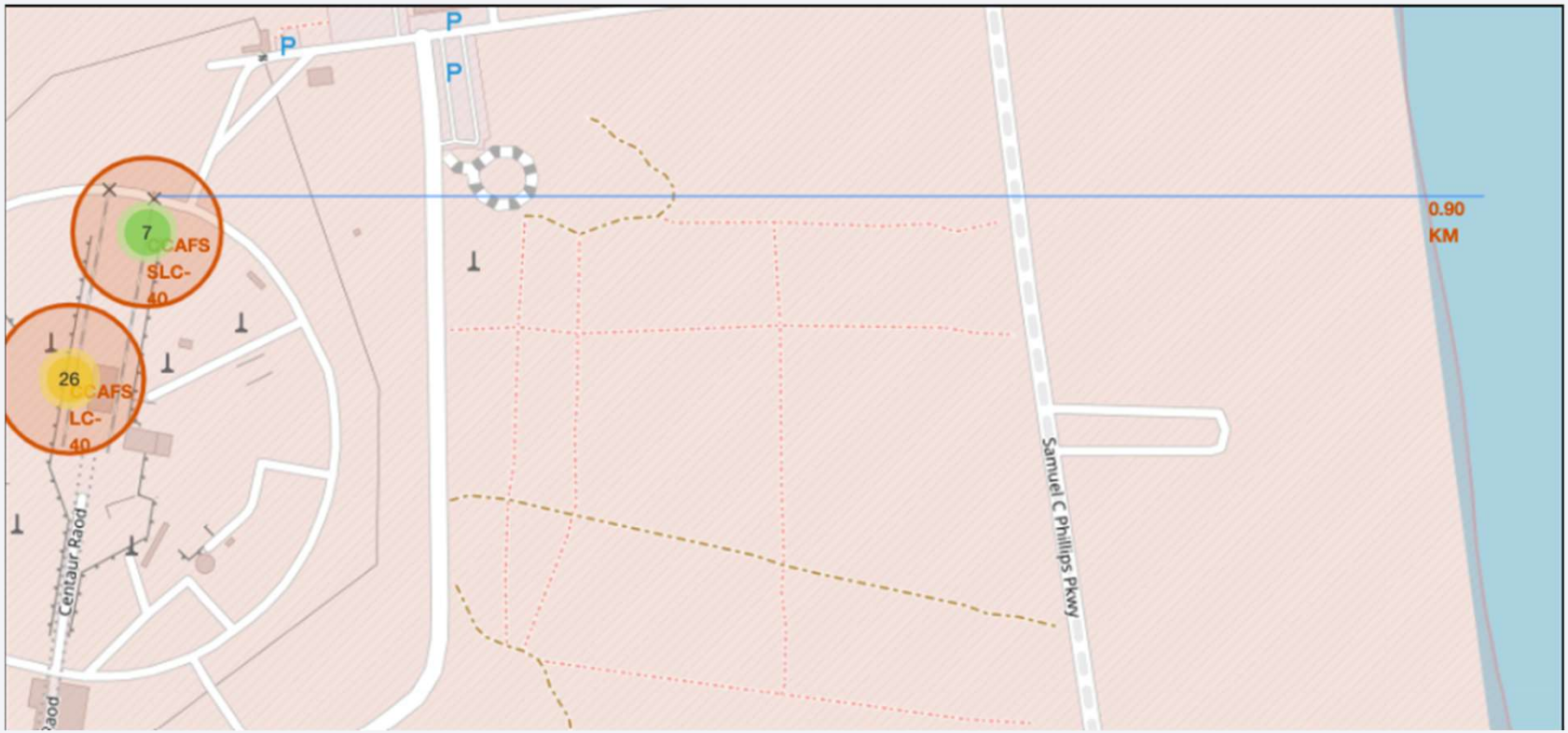
Launch Site Locations

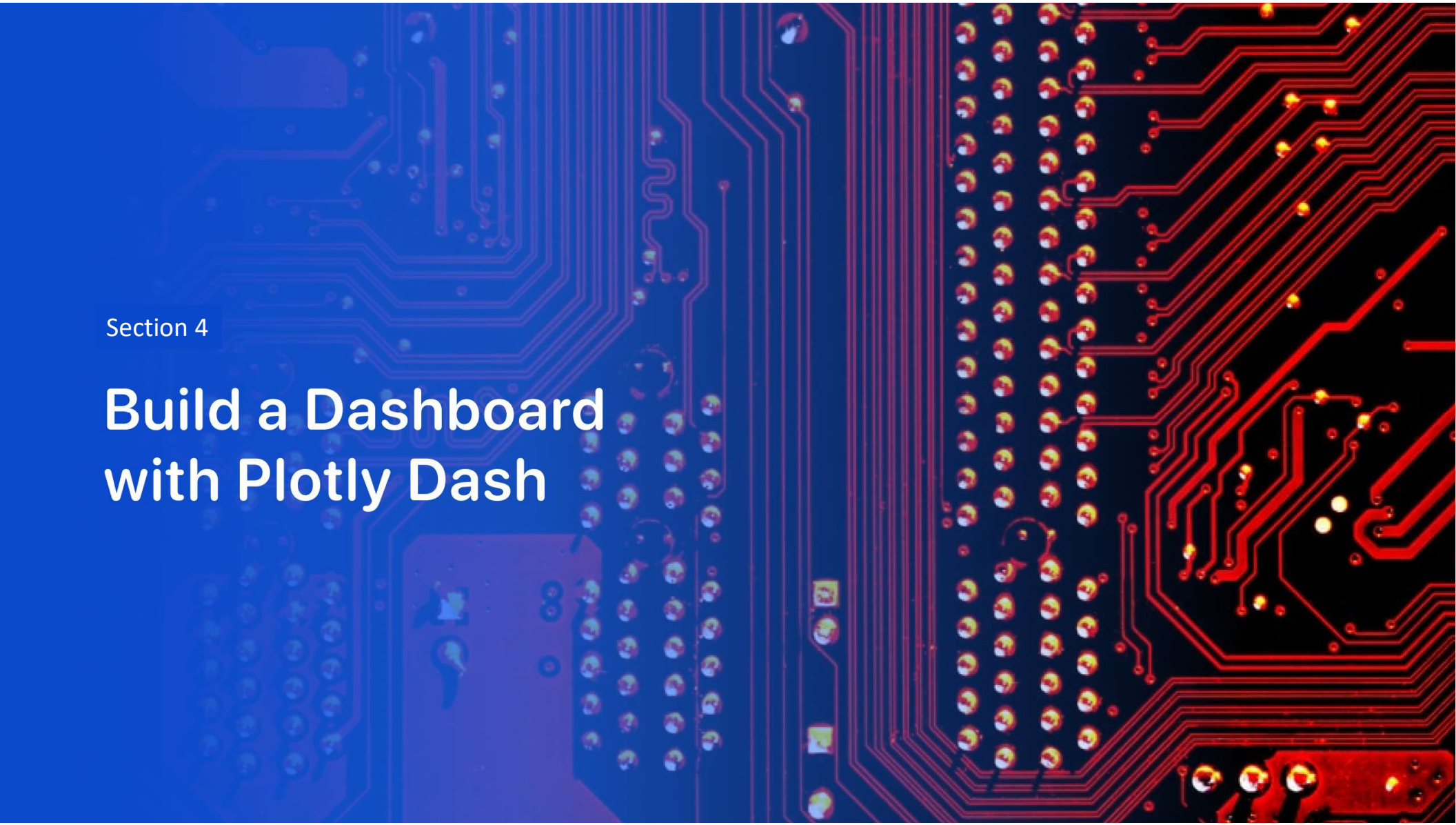


Color Labelled Launch Outcomes



Proximties from launch sites

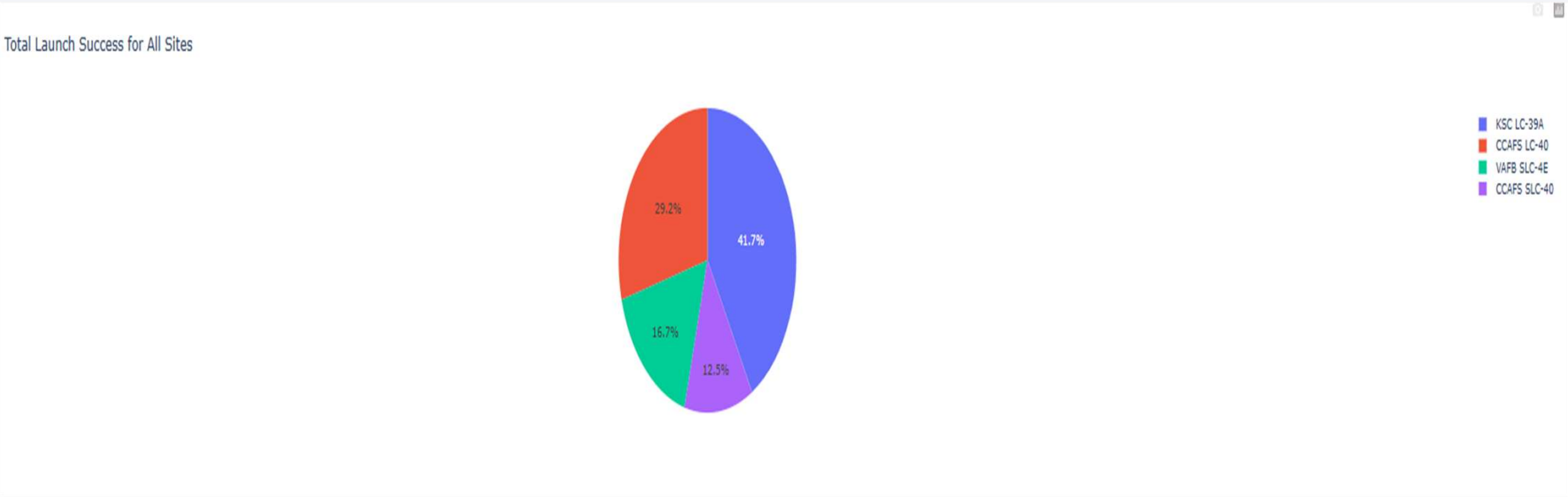




Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

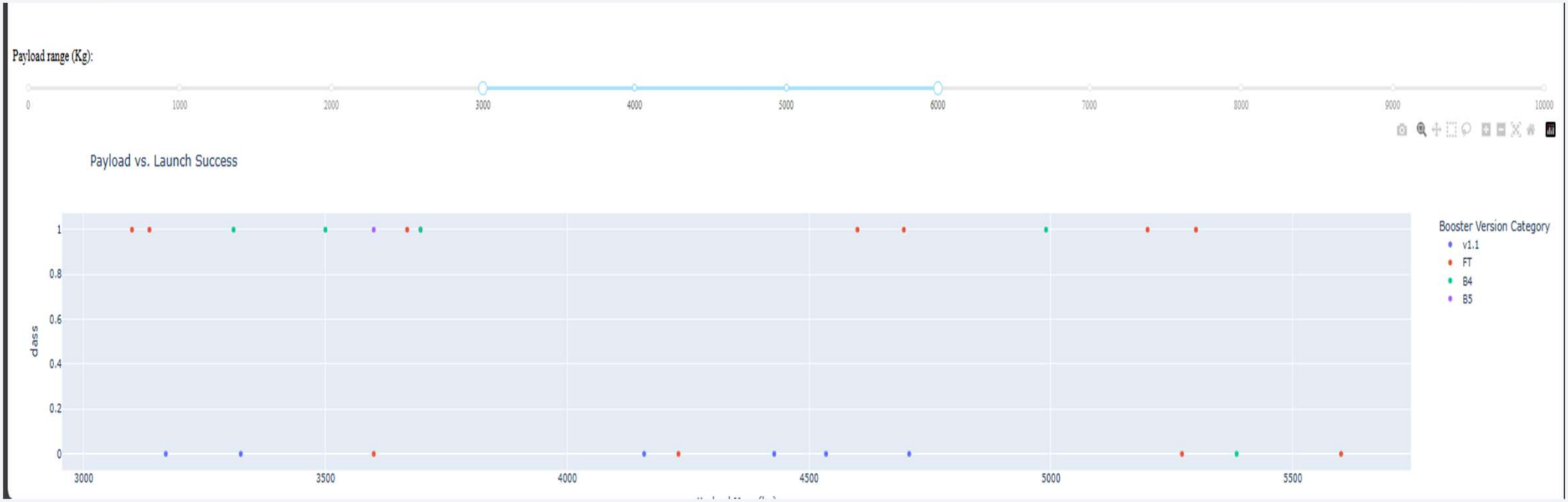


- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

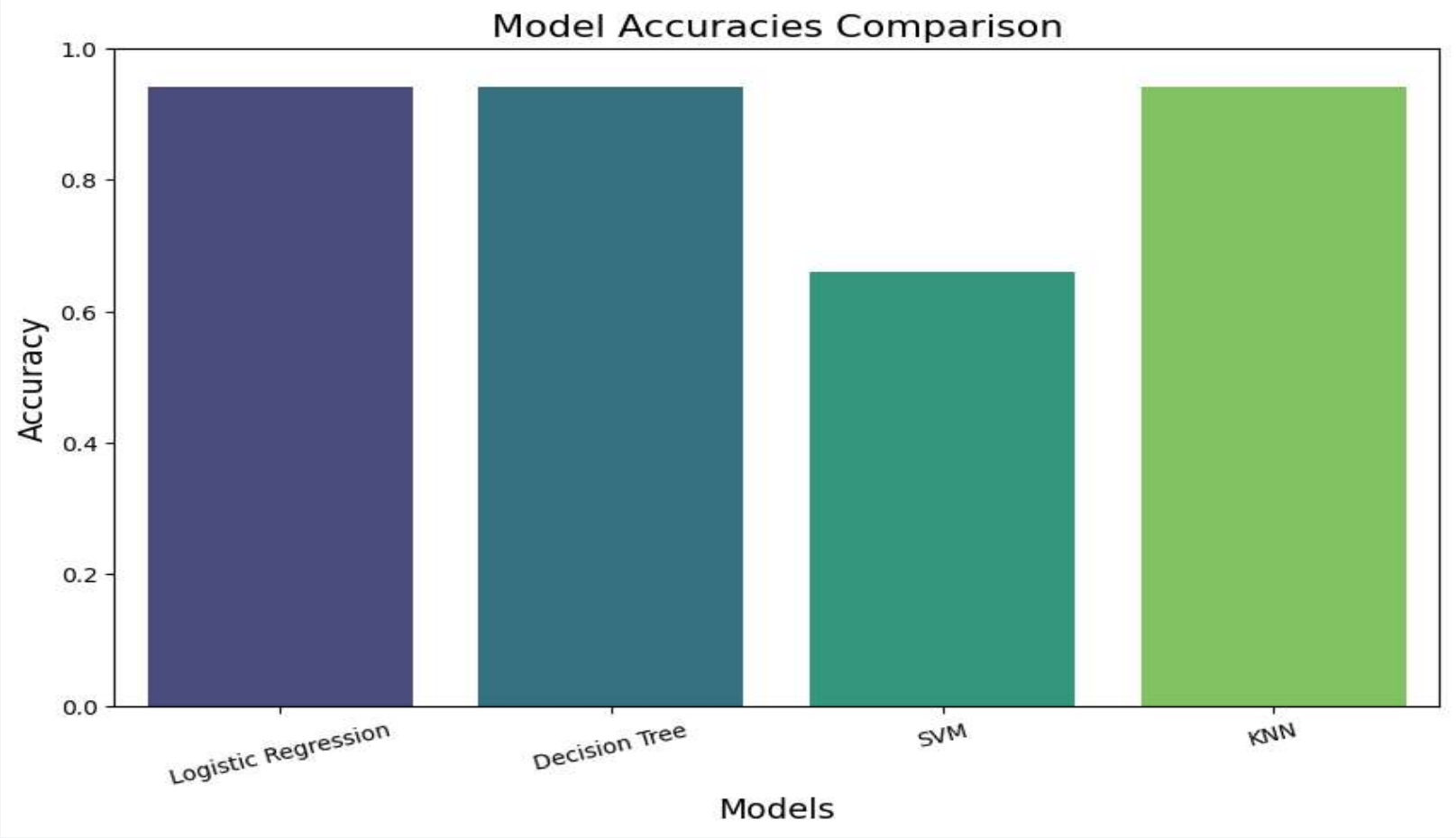


- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

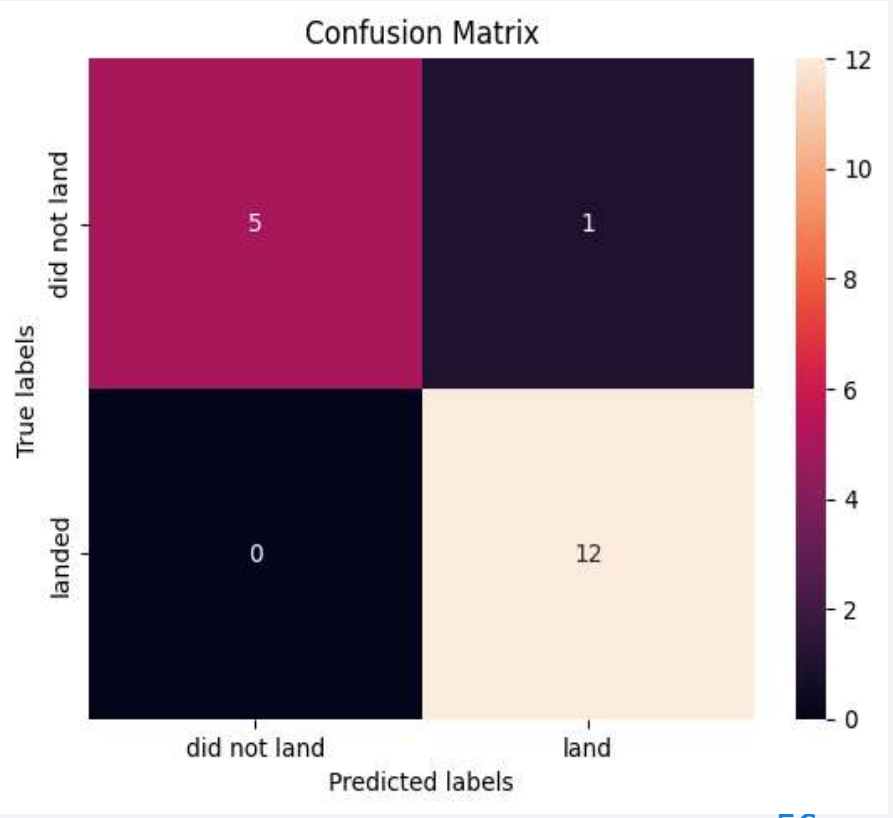
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- The confusion matrix shows excellent performance of the SVM model with:
- Accuracy: 94.4% (17/18 correct predictions)
- True Positives (12): Correctly predicted successful landings
- True Negatives (5): Correctly predicted landing failures
- False Positives (1): Only one case where the model predicted success but the landing failed
- False Negatives (0): No cases where the model predicted failure but the landing succeeded
- This indicates the model is highly reliable for predicting both successful and failed landings, with minimal misclassification.



Conclusions

1. Data Quality and Preparation:

The initial data collection involved gathering information from multiple sources, including CSV files and APIs. Rigorous data wrangling ensured that the dataset was clean, complete, and suitable for analysis.

Key preprocessing steps included handling missing values, encoding categorical variables, and normalizing numerical features, which laid a solid foundation for subsequent analyses.

2. Exploratory Data Analysis (EDA):

EDA revealed critical patterns and relationships within the data. Visualizations such as histograms, scatter plots, and bar charts highlighted the influence of factors like payload mass, launch site, and booster version on launch success.

The analysis indicated that certain launch sites, particularly Cape Canaveral, exhibited higher success rates, emphasizing the importance of site selection in mission planning.

3. Predictive Modeling:

A variety of classification models were evaluated, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN).

The logistic regression, decision trees and SVM modes emerged as the best performer, achieving an accuracy of 94%. These models demonstrated robustness and reliability, making it suitable for predicting launch success based on historical data.

4. Model Evaluation and Improvement:

The evaluation process included metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive view of model performance.

Hyperparameter tuning further enhanced model performance, ensuring that the selected model was optimized for the best results.

5. Interactive Dashboards and Visualizations:

An interactive dashboard was developed using Plotly Dash, allowing users to explore launch data dynamically. This tool facilitated user engagement and provided insights into launch performance through visual analytics.

Folium was utilized to create an interactive map, enhancing the geographical context of launch operations and success rates.

6. Insights and Recommendations:

The analysis underscored the significance of payload mass and launch site as critical factors influencing launch success. Recommendations for future launches include careful consideration of these variables during mission planning.

Continuous monitoring and updating of the predictive model with new data will enhance its accuracy and reliability over time.

Overall Impact:

The project successfully demonstrated the application of data analysis and machine learning in the aerospace domain, providing valuable insights that can enhance the reliability and efficiency of SpaceX launches. The findings and tools developed through this analysis serve as a foundation for ongoing improvements in launch operations and decision-making processes.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!