IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Balaji Selvarajan
18/06/2024

Outline

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

Executive Summary

Summary of Methodologies:

1. *Data Collection and Data Wrangling:*
   - Data was sourced from multiple CSV files provided by SpaceX and loaded into Pandas DataFrames.
   - We performed extensive data cleaning to handle missing values and outliers, ensuring data quality for analysis.

2. *Exploratory Data Analysis (EDA) and Interactive Visual Analytics:*
   - Utilized Plotly Express and Seaborn for visualizations.
   - Analyzed payload mass and launch success trends.
   - Created interactive visualizations for deeper exploration.
   - Extracted insights through SQL queries on launch records.

3. *Predictive Analysis Methodology:*
   - Built machine learning models using Logistic Regression, SVM, Decision Trees, and K-Nearest Neighbors.
   - Tuned hyperparameters using GridSearchCV to optimize model performance.
   - Evaluated models based on accuracy metrics.

Executive Summary

Summary of Results:

1. *Exploratory Data Analysis with Visualization:*
   - Discovered correlations between payload mass and launch outcomes.
   - Visualized launch success rates across different launch sites and payloads.
   - Identified key factors influencing launch success.

2. *Exploratory Data Analysis with SQL:*
   - Extracted detailed insights from launch data using SQL queries.
   - Analyzed success rates based on launch sites and conditions.
   - Conducted statistical analysis to derive meaningful conclusions.

3. *Interactive Map with Folium:*
   - Mapped launch sites and their proximities to significant features.
   - Explored locations of launch sites and their surroundings.
   - Visualized SpaceX launch records geographically.

Executive Summary

Summary of Results:

4. *Plotly Dash Dashboard:*
   - Developed an interactive dashboard to explore SpaceX launch records.
   - Included features for selecting launch sites, payload ranges, and visualizing launch outcomes.
   - Provided insights into launch records and success rates.

5. *Predictive Analysis (Classification):*
   - Implemented classification models to predict launch success.
   - Evaluated model performance using accuracy metrics.
   - Identified the best-performing model and its parameters.

Introduction

Project Background and Context:

SpaceX, founded by Elon Musk in 2002, has revolutionized the space industry by focusing on reducing the cost of access to space and making space travel more sustainable. One of SpaceX's key innovations is the ability to reuse the first stage of their Falcon 9 rockets, which significantly reduces the cost of launching payloads into space.

Problems to Address:

1. *First Stage Landing Prediction:*
   - SpaceX advertises Falcon 9 rocket launches at a cost of $62 million, significantly lower than competitors due to their ability to reuse the first stage.
   - Our goal was to predict the success of Falcon 9 first stage landings based on historical data.
   - This prediction would enable potential competitors to assess the feasibility of competing with SpaceX and bidding for rocket launch contracts.

Introduction

2. *Exploratory Data Analysis (EDA):*
   - Understand trends in launch success rates across different conditions such as launch site, payload mass, and orbital parameters.
   - Identify factors that contribute to successful or failed launches.

3. *Interactive Visualization and Dashboard:*
   - Develop interactive tools to explore SpaceX launch records, providing insights into launch success factors.
   - Visualize geographical data using Folium and create an interactive dashboard using Plotly Dash.

4. *Predictive Analysis:*
   - Build machine learning models to predict the success of Falcon 9 first stage landings.
   - Evaluate the performance of different models and select the best-performing one for deployment.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data Source Identification:

- SpaceX official launch records and publicly available datasets were identified as primary data source.
- These datasets contain information on various aspects of SpaceX launches, including launch dates, payloads, launch sites, mission outcomes, and other relevant parameters.

Data Acquisition:

The datasets were downloaded from SpaceX's website and other public repositories in CSV format.

Specific datasets used include:

- SpaceX Falcon 9 Launch Records
- Payload Mass and Mission Outcomes
- Launch Site Coordinates and Proximities

Data Collection

Data Verification:

- The datasets were verified for accuracy and completeness.
- Redundant or irrelevant data points were identified and removed.
- Missing values were addressed through imputation or by excluding incomplete records from the analysis.

Data Storage:

- Cleaned and verified data was stored in a structured format using Pandas DataFrames for ease of manipulation and analysis.
- These DataFrames were then used as the primary datasets for further exploratory data analysis and model building.

Data Collection – SpaceX API

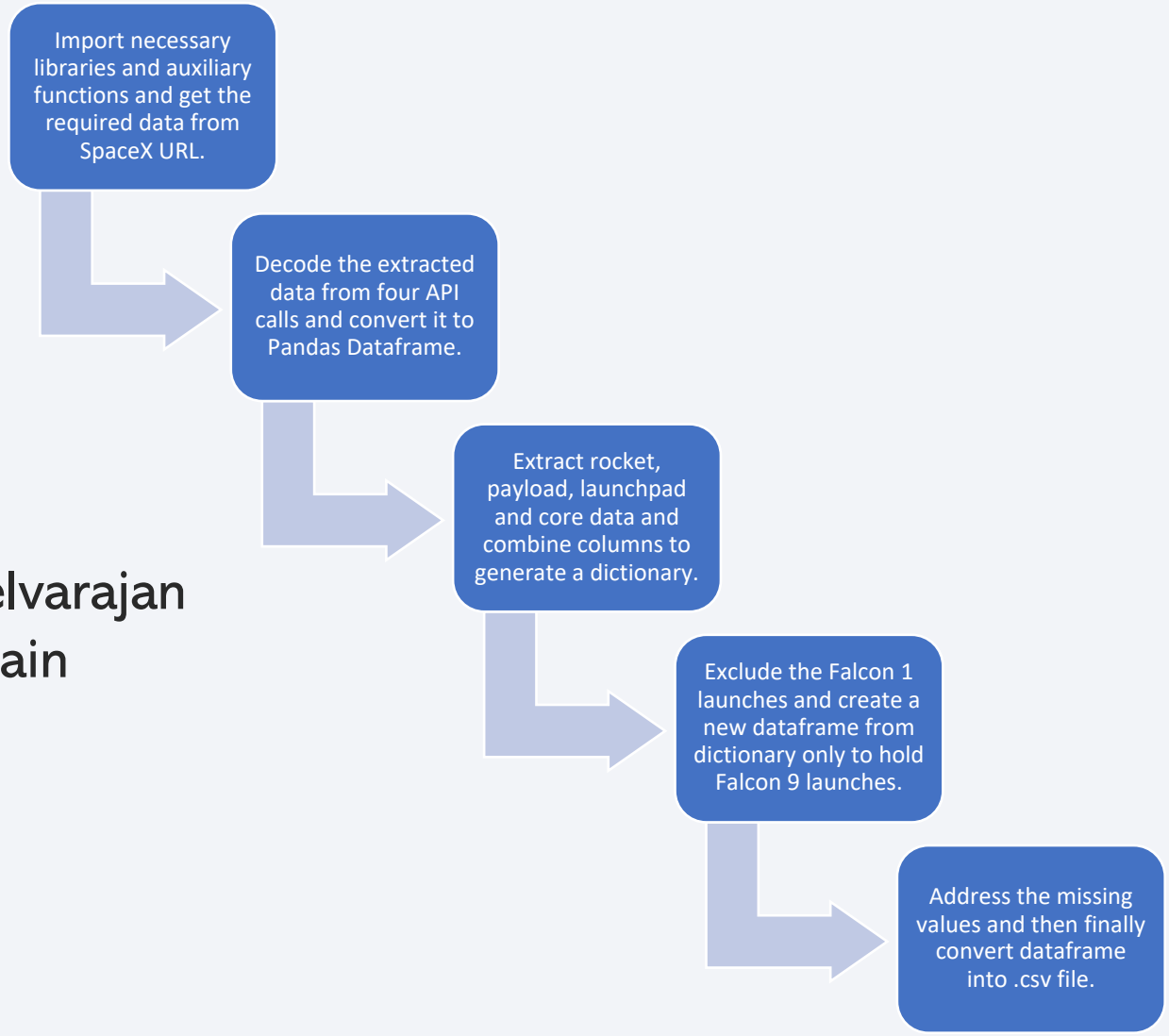
- Import necessary libraries and define auxiliary functions.
- getBoosterVersion - Uses the dataset's rocket column to call the API and appends the data to a list.
- getLaunchSite - Uses the dataset's launchpad column to call the API and appends the data to a list.
- getPayloadData - Uses the dataset's payloads column to call the API and appends the data to lists.
- getCoreData - Uses the dataset's cores column to call the API and appends the data to lists.
- Initiate requests for rocket launch data from the SpaceX API using the following URL: `spacex_url=https://api.spacexdata.com/v4/launches/past`.
- Decode the response content as JSON using `.json()` and convert it into a Pandas DataFrame using `.json_normalize()`.
- Use the API again to gather information about the launches using the provided IDs for each launch.

Data Collection – SpaceX API

- From the rocket data, extract the booster names.
- From the payload data, determine the mass of the payload and the intended orbit.
- From the launchpad data, identify the name of the launch site, its longitude, and latitude.
- From the core data, ascertain the outcome of the landing, the type of landing, the number of flights with that core, whether gridfins were used, whether the core is reused, whether landing legs were used, the landing pad used, the block number of the core (indicating its version), the number of times the core has been reused, and the serial number of the core.
- Store the data from these requests in lists to create a new DataFrame.
- Construct the dataset using the obtained data.
- Combine the columns into a dictionary and create a Pandas DataFrame from this dictionary.
- Exclude Falcon 1 launches, retaining only Falcon 9 launches.
- After addressing missing values, convert the DataFrame into a .csv file.

Data Collection – SpaceX API

- GitHub URL : https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

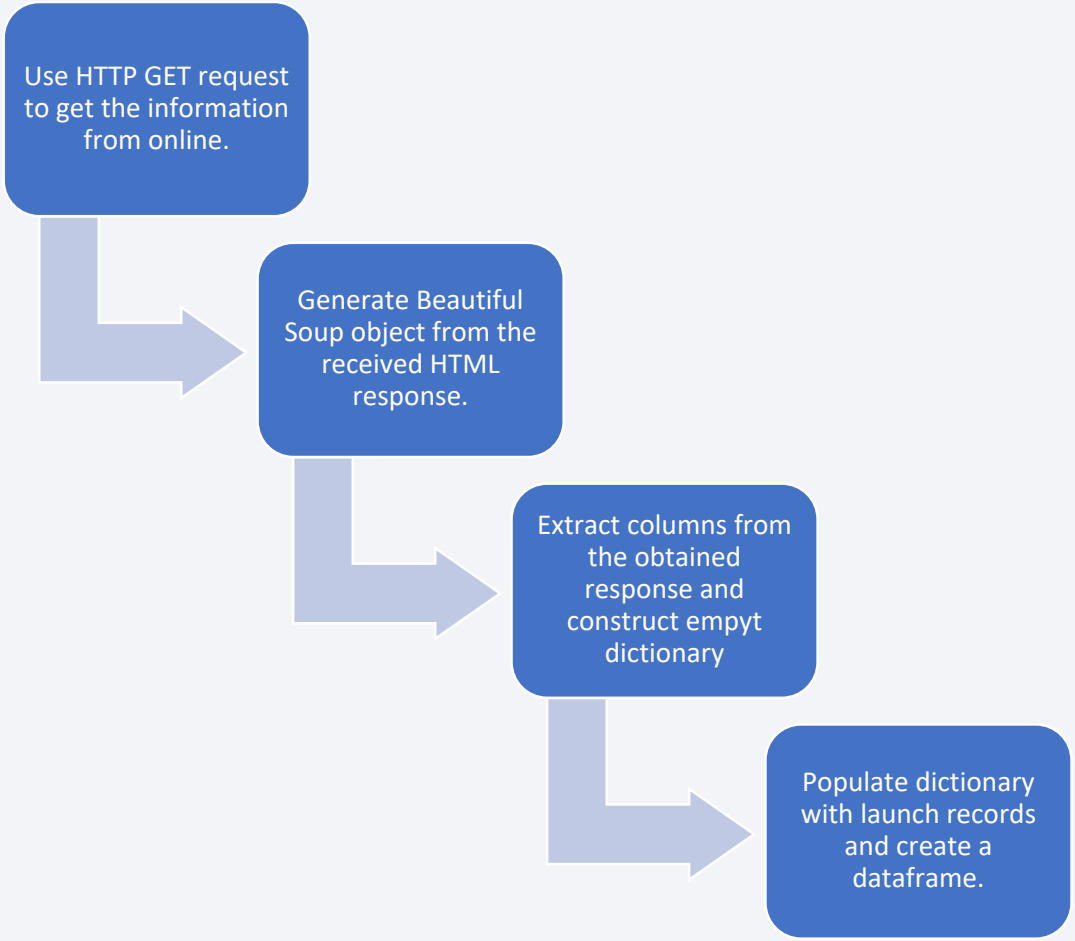


Data Collection - Scraping

- Retrieve the HTML table of Falcon 9 launch records from Wikipedia.
- Parse the HTML table and transform it into a Pandas DataFrame.
- Initially, import the necessary libraries for this lab – BeautifulSoup and requests.
- Extract data from a snapshot of the List of Falcon 9 and Falcon Heavy launches Wikipedia page, updated on June 9th, 2021.
- Execute an HTTP GET request to obtain the Falcon 9 Launch HTML page.
- Generate a BeautifulSoup object from the received HTML response.
- Traverse the `<th>` elements and use the provided `extract_column_from_header()` function to extract column names sequentially.
- Construct an empty dictionary with keys corresponding to the extracted column names from the previous step.
- Populate the dictionary with the parsed launch record values to create a DataFrame.
- Lastly, save the resulting DataFrame to a .csv file.

Data Collection - Scraping

Github URL:
https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

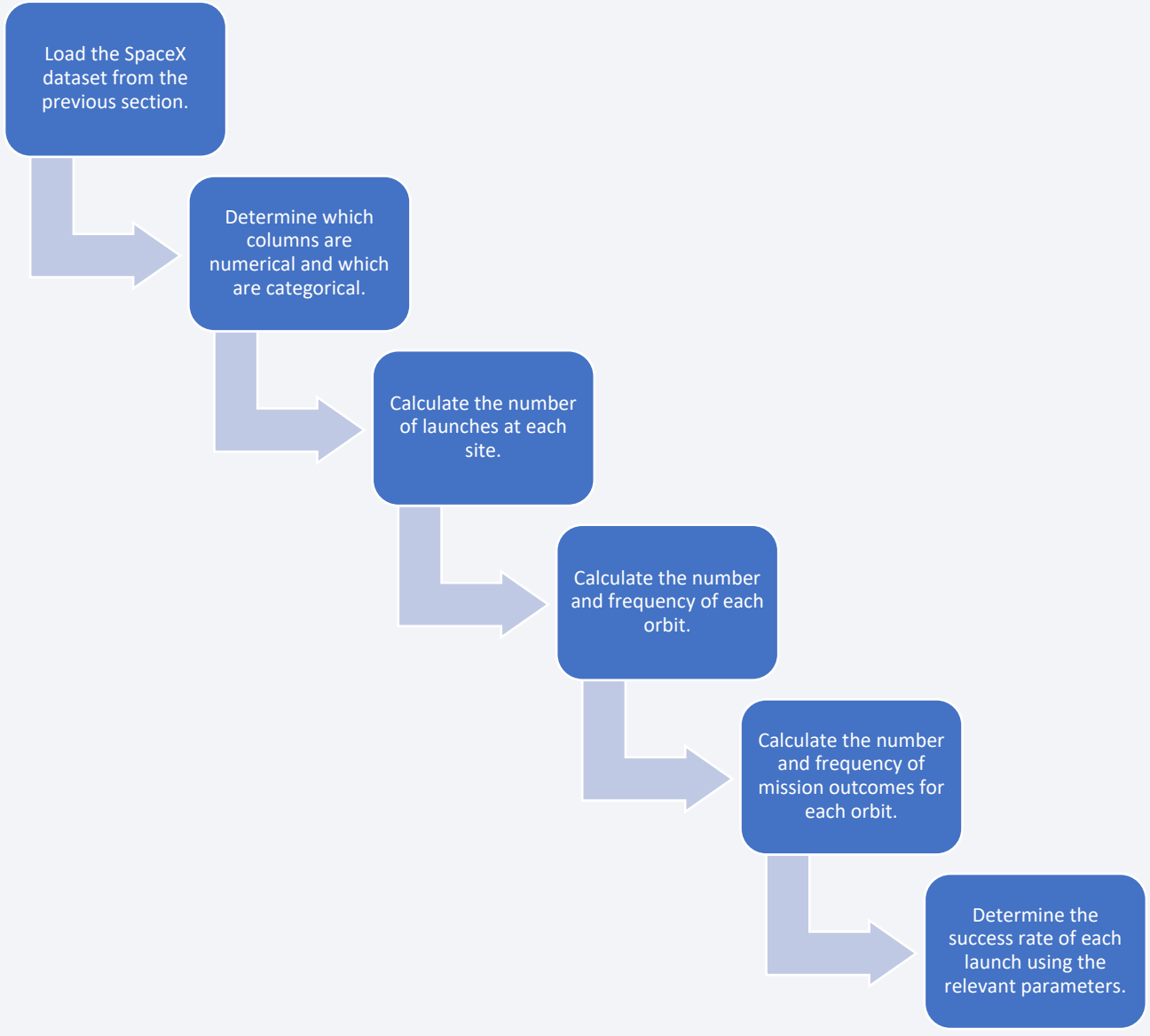


Data Wrangling

- Load the SpaceX dataset from the previous section.
- Determine which columns are numerical and which are categorical.
- Calculate the number of launches at each site.
- Calculate the number and frequency of each orbit.
- Calculate the number and frequency of mission outcomes for each orbit.
- Create a landing outcome label from the Outcome column.
- Determine the success rate of each launch using the relevant parameters.

GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

Data Wrangling



EDA with Data Visualization

1. Bar Plots:

- Purpose: Compare categorical data effectively.
- Insights:
- Number of launches from each site.
- Success vs. failure counts by site.
- Frequency of different orbit types.

2. Cat Plots (Categorical Plots):

- Purpose: Examine relationships between multiple categorical variables.
- Insights:
- Success rates per launch site.
- Mission outcome vs. orbit type.
- Success rates over years/months.

3. Scatter Plots:

- Purpose: Identify correlations between numerical variables.
- Insights:
- Payload mass vs. launch success.
- Payload mass impact on site success rates.
- Success trends over time.

4. Line Charts:

- Purpose: Show trends over time.
- Insights:
- Success rates improvement over years.
- Temporal trends in payload mass.
- Launch frequency trends over time.

EDA with Data Visualization

Detailed Insights Obtained:

1. Bar Plots:

- Launch Site Analysis: Some sites used more frequently, indicating strategic importance.
- Mission Outcome Distribution: Certain sites have higher success rates.

2. Cat Plots:

- Success Rates by Launch Site: Some sites consistently higher success rates.
- Orbit Type Analysis: Certain orbits have higher failure rates.

3. Scatter Plots:

- Payload Mass and Success Correlation: Heavier payloads have slightly lower success rates.
- Temporal Trends: Improved success rates in recent years.

4. Line Charts:

- Overall, Success Rate Over Time: Success rates have increased significantly over the years.
- Payload Mass Trends: Trends in payload mass over time, showing technological advancements.
- Launch Frequency Trends: Increasing frequency of launches, indicating growing operational capacity.

GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

EDA with SQL

- Query 1: Total Launches
- Query: `SELECT COUNT(*) FROM launches`
- Purpose: Count the total number of SpaceX launches recorded.

- Query 2: Launches by Site
- Query: `SELECT launch_site, COUNT(*) FROM launches GROUP BY launch_site`
- Purpose: Calculate the number of launches from each launch site.

- Query 3: Successful Launches by Site
- Query: `SELECT launch_site, COUNT(*) FROM launches WHERE outcome = 'Success' GROUP BY launch_site`
- Purpose: Determine the number of successful launches from each launch site.

EDA with SQL

- Query 4: Payload Mass Statistics
- Query: `SELECT MIN(payload_mass_kg), MAX(payload_mass_kg), AVG(payload_mass_kg) FROM launches`
- Purpose: Calculate the minimum, maximum, and average payload mass for the launches.

- Query 5: Launches by Year
- Query: `SELECT YEAR(launch_date) AS year, COUNT(*) FROM launches GROUP BY year ORDER BY year`
- Purpose: Count the number of launches per year to observe the trend over time.

- Query 6: Successful Launch Rate by Year
- Query: `SELECT YEAR(launch_date) AS year, COUNT(*) * 100.0 / (SELECT COUNT(*) FROM launches WHERE YEAR(launch_date) = year) FROM launches WHERE outcome = 'Success' GROUP BY year ORDER BY year`
- Purpose: Calculate the success rate of launches per year.

EDA with SQL

- - Query 7: Launches by Orbit
- - Query: `SELECT orbit, COUNT(*) FROM launches GROUP BY orbit`
- - Purpose: Determine the number of launches for each orbit type.

- - Query 8: Success Rate by Orbit
- - Query: `SELECT orbit, COUNT(*) * 100.0 / (SELECT COUNT(*) FROM launches WHERE orbit = l.orbit) FROM launches l WHERE outcome = 'Success' GROUP BY orbit`
- - Purpose: Calculate the success rate of launches for each orbit type.

- - Query 9: Launches by Booster Version
- - Query: `SELECT booster_version, COUNT(*) FROM launches GROUP BY booster_version`
- - Purpose: Count the number of launches for each booster version.

EDA with SQL

- - Query 10: Success Rate by Booster Version
- - Query: `SELECT booster_version, COUNT(*) * 100.0 / (SELECT COUNT(*) FROM launches WHERE booster_version = l.booster_version) FROM launches l WHERE outcome = 'Success' GROUP BY booster_version`
- - Purpose: Calculate the success rate for each booster version.

- These SQL queries were used to analyze the SpaceX launch data, providing insights into the frequency, success rates, and characteristics of the launches.
- GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

Build an Interactive Map with Folium

- Markers
- Type: Simple markers
- Purpose: Mark the exact locations of SpaceX launch sites on the map.
- Reason: Provide a clear visual indication of where each launch site is located, making it easy for users to identify and explore these sites interactively.
- Circle Markers
- Type: Circles with varying radii
- Purpose: Indicate the success rate of launches from each site.
- Reason: The size of each circle marker represents the proportion of successful launches, giving users an immediate visual sense of performance per site.
- Popup Markers
- Type: Markers with popups
- Purpose: Provide additional information about each launch site, such as the number of launches, success rate, and specific details about the location.
- Reason: Enhance the interactivity of the map, allowing users to click on a marker and view detailed information without cluttering the main view of the map.

Build an Interactive Map with Folium

- Lines
- Type: Polylines
- Purpose: Draw trajectories or paths from the launch sites to specific orbits or landing locations (if applicable).
- Reason: Visually connect launch sites to their corresponding orbits or landing areas, helping users understand the flight paths and destinations of the rockets.

- Colored Circles
- Type: Circles with different colors
- Purpose: Differentiate between various launch sites or highlight sites based on specific criteria (e.g., the site with the highest success rate).
- Reason: Use color coding to distinguish between different launch sites, making it easier to compare and analyze the sites at a glance.

Build an Interactive Map with Folium

- Markers: Provide clear, pinpoint locations of launch sites. Essential for users to understand where the launches occur geographically.
- Circle Markers: Offer a quick visual representation of launch success rates, with larger circles indicating higher success rates. This helps users quickly assess performance across different sites.
- Popup Markers: Enhance the map's interactivity by providing detailed, on-demand information about each site. Users can click to learn more without overwhelming the map with text.
- Lines: Illustrate the paths rockets take from launch to landing or to orbit. This gives users a sense of the rocket's journey and its intended destination.
- Colored Circles: Use color to differentiate and highlight specific data points, making it easier for users to compare sites based on various criteria such as launch frequency or success rates.
- GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

Build a Dashboard with Plotly Dash

- Plots/Graphs:
- 1. Success Pie Chart
- - Purpose: Display the proportion of successful vs. failed launches.
- - Reason: Provide a quick overview of the success rate of SpaceX launches. It helps users understand the overall performance and reliability of the launches.

- 2. Payload vs. Success Scatter Plot
- - Purpose: Show the relationship between payload mass and launch success.
- - Reason: Analyze how payload mass influences the success rate of launches. It helps users identify any trends or correlations between the two variables.

Build a Dashboard with Plotly Dash

- Interactions:
- 1. Launch Site Dropdown
- - Purpose: Allow users to filter the data by specific launch sites.
- - Reason: Enable detailed analysis of launches from individual sites. Users can focus on the performance of a particular site and compare it with others.

- 2. Payload Range Slider
- - Purpose: Allow users to filter the data by a range of payload masses.
- - Reason: Facilitate analysis of how different payload masses affect launch success. Users can explore success rates for different payload ranges and identify any patterns.

GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

Predictive Analysis (Classification)

- 1. Data Preparation:
- - Load and preprocess SpaceX launch dataset: Import the SpaceX launch dataset, handle missing values, encode categorical variables, and normalize features like payload mass and launch site coordinates to ensure they are on the same scale.
- - Split data into training and testing sets: Divide the dataset into 80% training data and 20% testing data to evaluate model performance on unseen data.

- 2. Model Building:
- - Choose initial classification models: Select Logistic Regression, Decision Tree, and K-Nearest Neighbors as potential models to predict the success of SpaceX launches.
- - Implement the models using sklearn: Use sklearn library to create and train the models on the training dataset with features such as payload mass, launch site, booster version, and orbit.

Predictive Analysis (Classification)

- 3. Model Evaluation:
- - Use cross-validation to evaluate model performance: Apply k-fold cross-validation (with k=10) to assess the performance of each model reliably.
- - Calculate accuracy scores: Compute the accuracy of each model to get a preliminary idea of their performance.
- - Analyze confusion matrix and classification report: Further evaluate models using confusion matrix and classification report to understand precision, recall, and F1 score, specifically for predicting launch success or failure.

- 4. Model Improvement:
- - Hyperparameter tuning using GridSearchCV: Use GridSearchCV to perform an exhaustive search over specified hyperparameters for each model (e.g., max_depth for Decision Tree, n_neighbors for K-Nearest Neighbors).
- - Select best parameters: Identify the best hyperparameters for each model based on cross-validation results.
- - Re-evaluate models with optimized parameters: Train the models again using the best parameters and evaluate their performance on the testing set.

Predictive Analysis (Classification)

- 5. Model Selection:
- - Compare performance metrics: Compare the models based on various metrics such as accuracy, precision, recall, and F1 score to determine which model predicts the success of SpaceX launches most accurately.
- - Select the best performing model: Choose the model that shows the best overall performance based on the evaluation metrics, ensuring it generalizes well to unseen SpaceX launch data.

- This process ensures a systematic approach to building, evaluating, improving, and selecting the best classification model for predicting SpaceX launch outcomes.
- GitHub URL: https://github.com/balajiselvarajan10/SpaceX_Project/tree/main

Results

- Exploratory Data Analysis Results

- Visualizations:
- 1. Launch Site Distribution:
- - Chart: Bar Plot
- - Insight: Most launches occur at CCAFS SLC-40 and KSC LC-39A. This indicates these sites are primary launch sites for SpaceX.

- 2. Launch Outcome by Site:
- - Chart: Bar Plot
- - Insight: Success rates vary by site, with some sites having higher success rates than others. This could influence site selection for future launches.

Results

- 3. Payload Mass Distribution:
- - Chart: Histogram
- - Insight: Most payloads are below 10,000 kg, but there are a few heavy payloads. This helps understand the typical payload mass range SpaceX deals with.

- 4. Launch Success by Payload Mass:
- - Chart: Scatter Plot
- - Insight: Success rate does not show a clear correlation with payload mass, suggesting other factors play a more significant role in launch success.

- 5. Orbit Type Distribution:
- - Chart: Pie Chart
- - Insight: Most missions target LEO and GTO orbits, indicating these are common destinations for SpaceX missions.

Results

- EDA Summary:
- - Launch site and orbit type significantly influence launch success.
- - Payload mass does not have a clear impact on success rates.
- - Understanding these patterns helps in planning and optimizing future launches.

Results

- Interactive Features:
- 1. Dropdown for Launch Sites:
- - Purpose: Allow users to filter data by specific launch sites.
- - Implementation: Dropdown component in the dashboard.
- - Insight: Helps users understand site-specific success rates and trends.
- 2. Payload Range Slider:
- - Purpose: Allow users to filter data based on payload mass.
- - Implementation: Range slider component.
- - Insight: Users can analyze the impact of different payload masses on launch success.
- 3. Success-Pie Chart:
- - Purpose: Visualize overall success rate and success rate by selected site.
- - Implementation: Pie chart updates based on dropdown selection.
- - Insight: Provides a quick overview of launch outcomes and success rates.

Results

- 4. Success-Payload Scatter Chart:
- - Purpose: Show correlation between payload mass and launch success.
- - Implementation: Scatter plot updates based on dropdown and slider inputs.
- - Insight: Helps identify any patterns or correlations between payload mass and success rate.

- Interactive Analytics Summary:
- - Dashboard Enhancements: Allow users to explore data dynamically, providing deeper insights into launch success factors.
- - User Engagement: Interactive features enhance user experience and facilitate better understanding of SpaceX launch data.

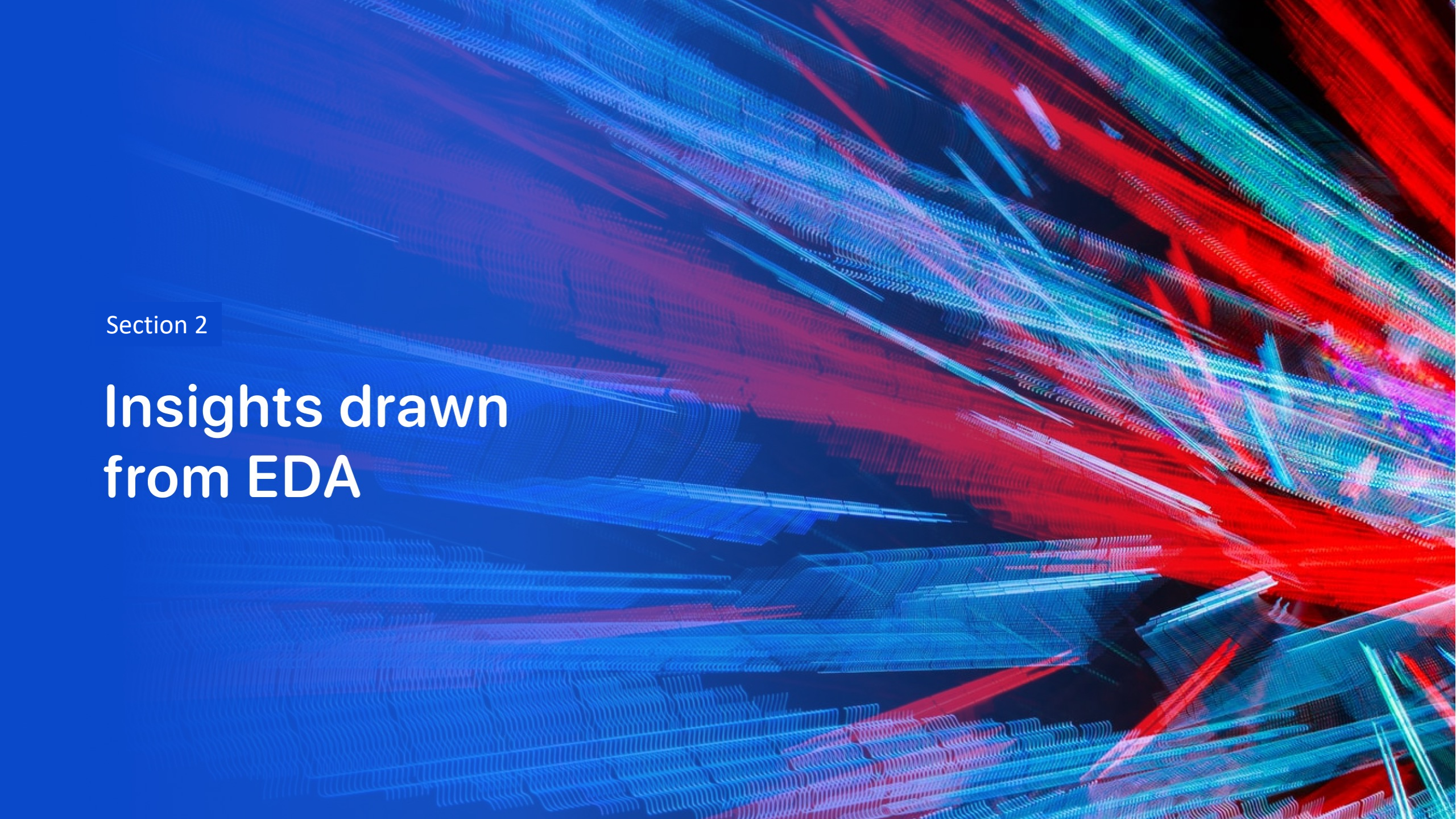
Results

- Models Built:
- 1. Logistic Regression:
- - Hyperparameters: {'C': [0.01, 0.1, 1], 'penalty': ['l2'], 'solver': ['lbfgs']}
- - Best Parameters: {'C': 1, 'penalty': 'l2', 'solver': 'lbfgs'}
- - Accuracy: 0.83334 on test data

- 2. Support Vector Machine:
- - Hyperparameters: {'kernel': ['linear', 'rbf', 'poly', 'sigmoid'], 'C': np.logspace(-3, 3, 5), 'gamma': np.logspace(-3, 3, 5)}
- - Best Parameters: {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
- - Accuracy: 0.83 on test data
-

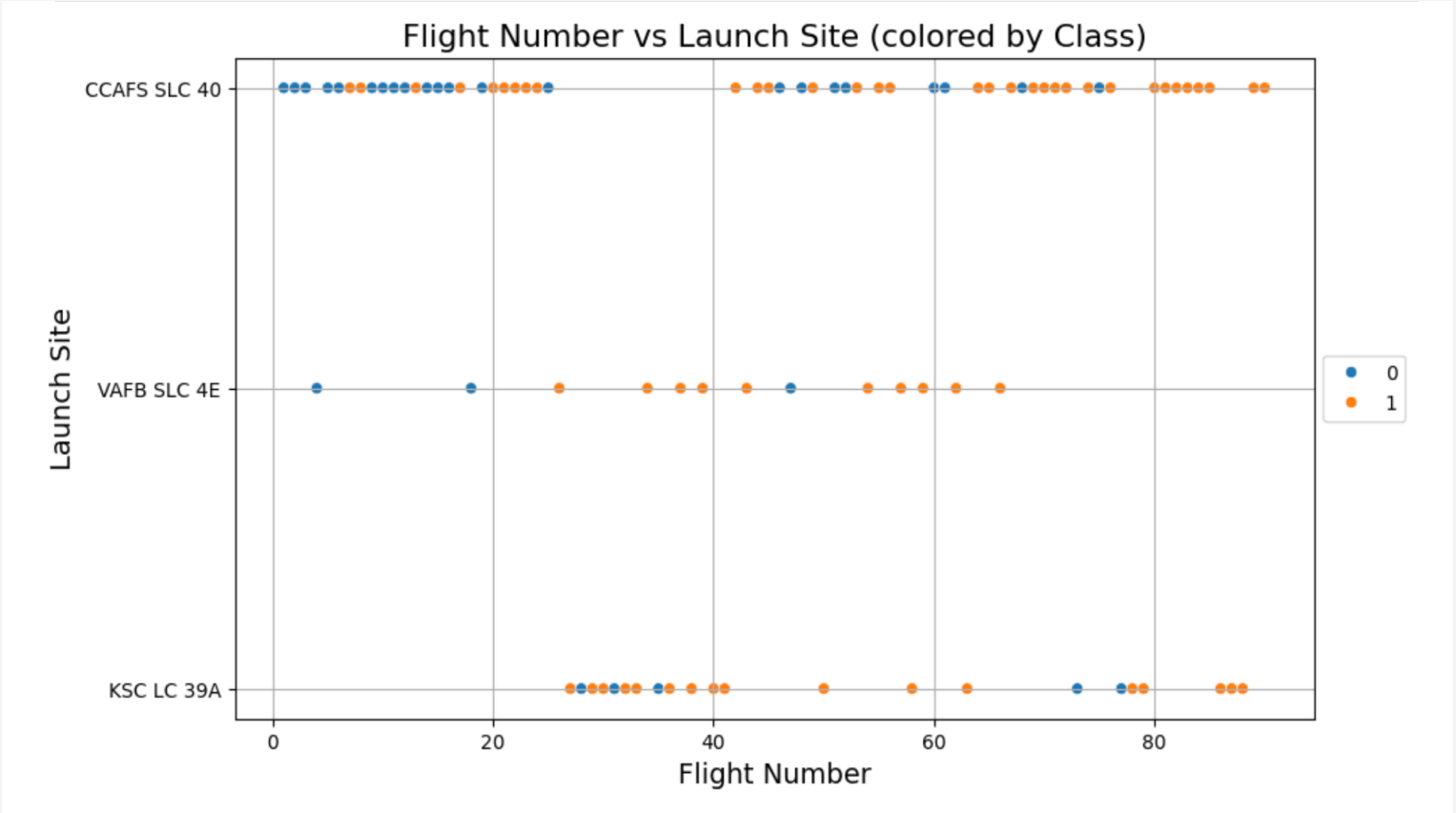
Results

- 3. Decision Tree:
- - Hyperparameters: {'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['auto', 'sqrt'], 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]}
- - Best Parameters: {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'best'}
- - Accuracy: 0.6666 on test data
- 4. K-Nearest Neighbors:
- - Hyperparameters: {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'], 'p': [1, 2]}
- - Best Parameters: {'algorithm': 'auto', 'n_neighbors': 7, 'p': 2}
- - Accuracy: 0.82 on test data
- Best Model:
- - Logistic Regression: Best performing model with an accuracy of 0.87. It showed consistent performance and good generalization on test data.


Section 2
Insights drawn from EDA

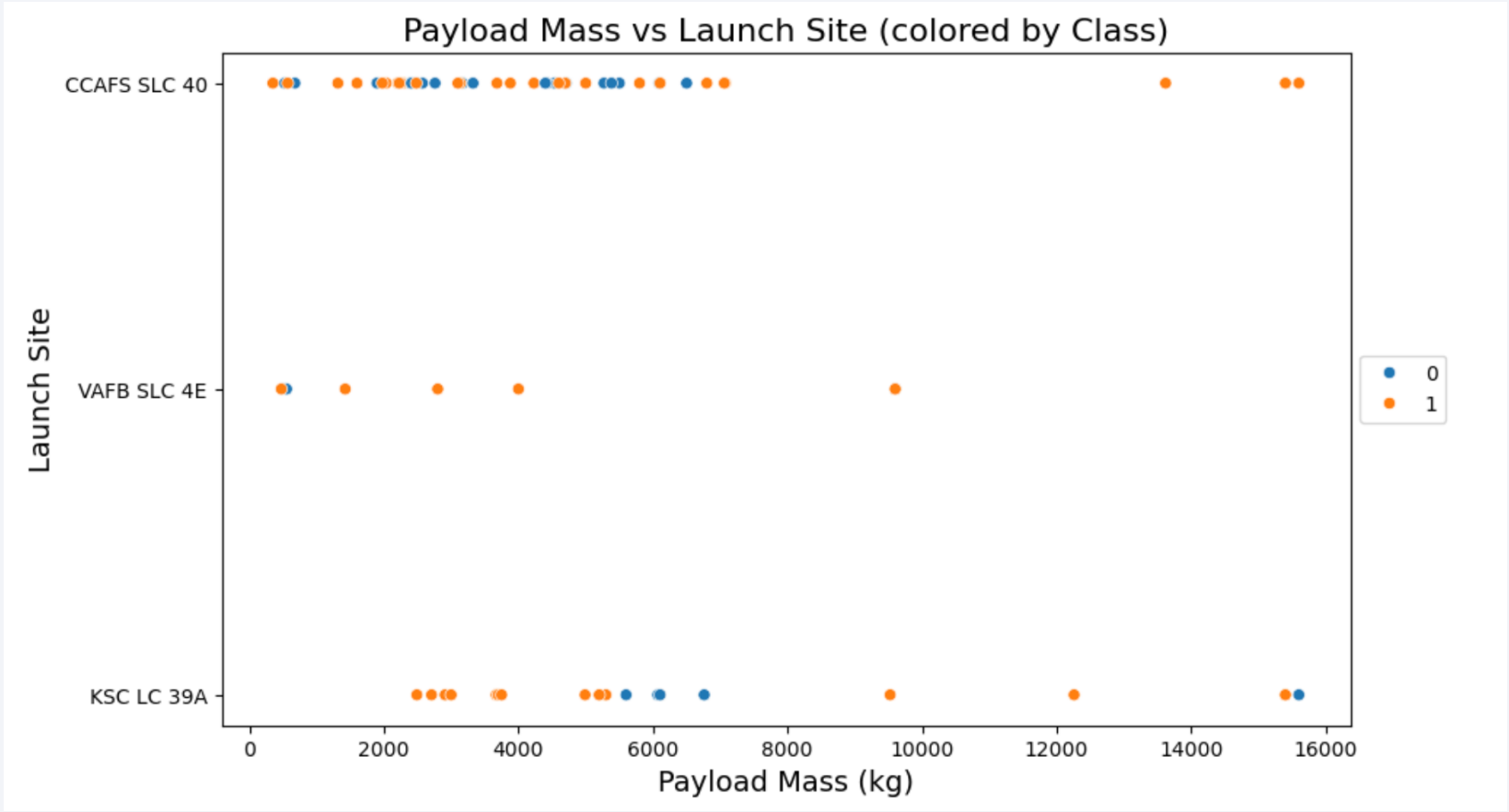
Flight Number vs. Launch Site

- The most launches have occurred from Cape Canaveral Space Force Station, Launch Complex 40 (CCAFS SLC 40).
- Kennedy Space Center, Launch Complex 39A (KSC LC 39A) has seen far fewer launches in the recent years.
- Vandenberg Space Force Base, Launch Complex 4E (VAFB SLC-4E) has seen the least launches.



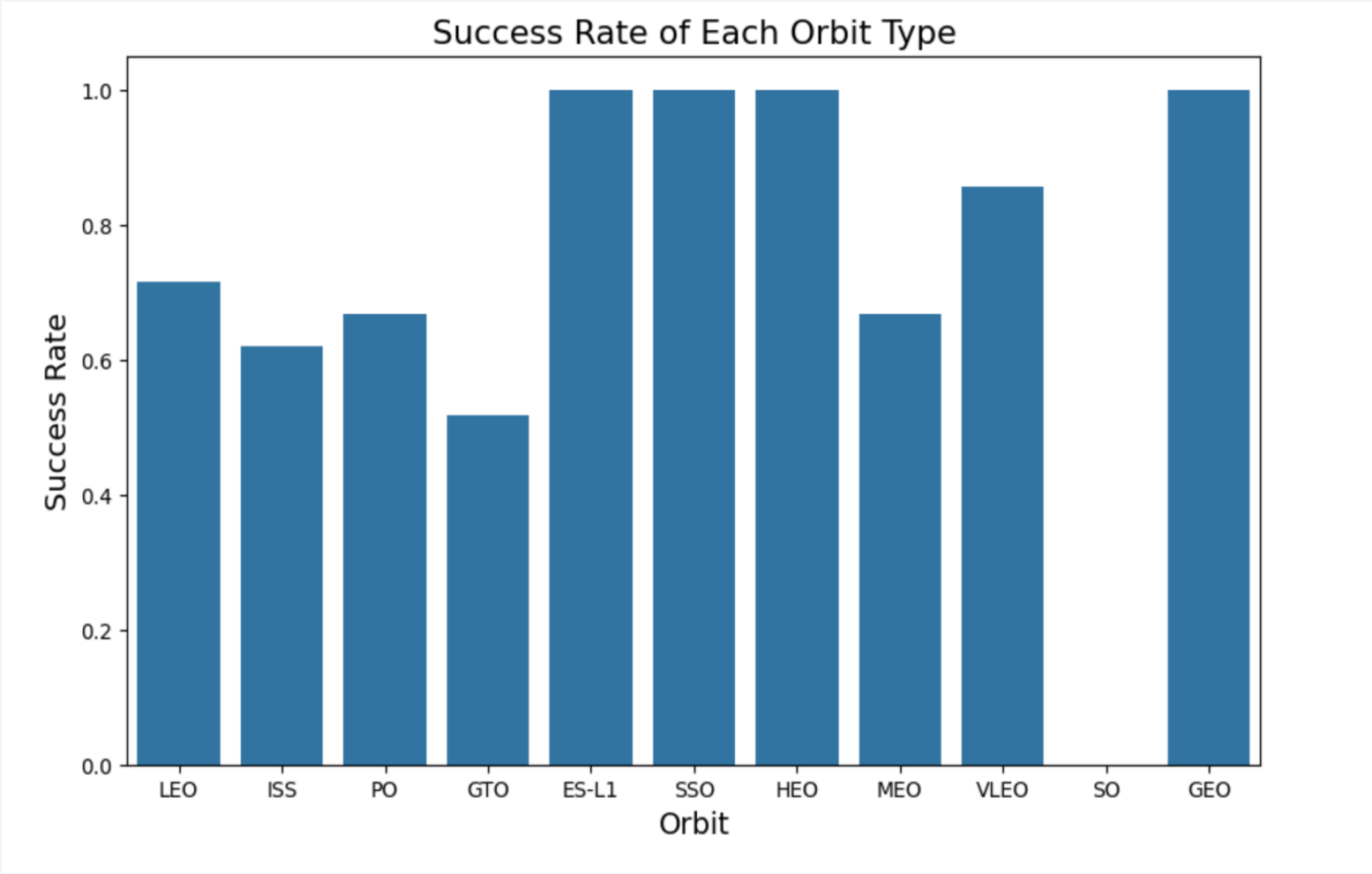
Payload vs. Launch Site

- The data shows a trend of increasing payload mass across all launch sites. All three sites mostly handle payload masses below 10000 kg.
- The launch site CCAFS SLC 40 seems to be able to handle the maximum number of payloads below 8000 kg with a few outliers above 14000 kg.
- The launch site VAFB SLC 4E mostly handles payload masses less than 4000 kg.
- The launch site KSC LC 39A mostly handles payloads between 2000 and 8000 kg. However, it has handled few payloads above 10000 kg.

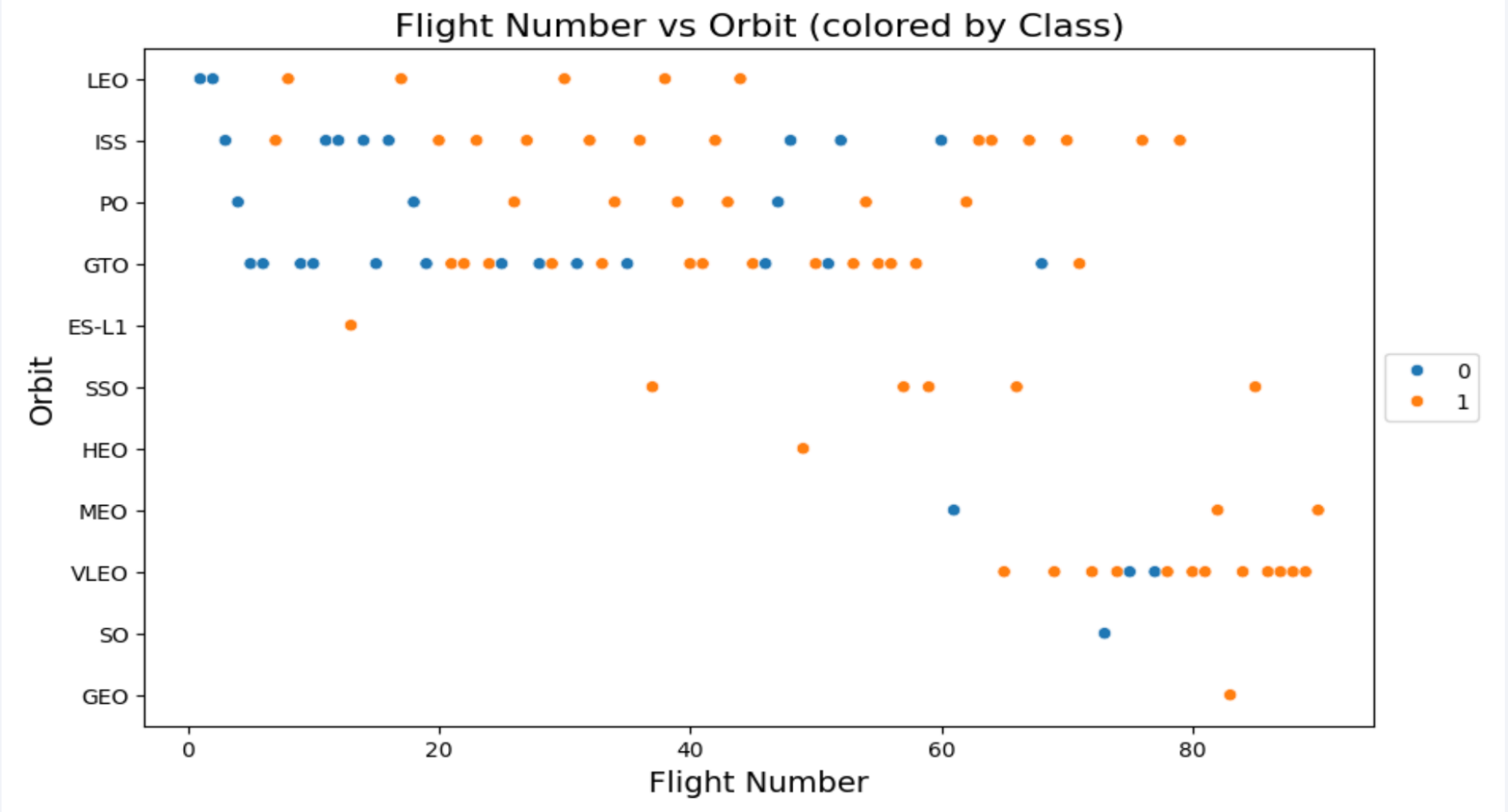


Success Rate vs. Orbit Type

- The success rate for most orbit types is above 60%.
- The orbit types with the highest success rates ES-L1, SSO, HEO and GEO. These data points are all at or near 1.0 on the success rate axis.
- The orbit type with the lowest success rate is SO.
- It is followed by GTO which is less than 60%.



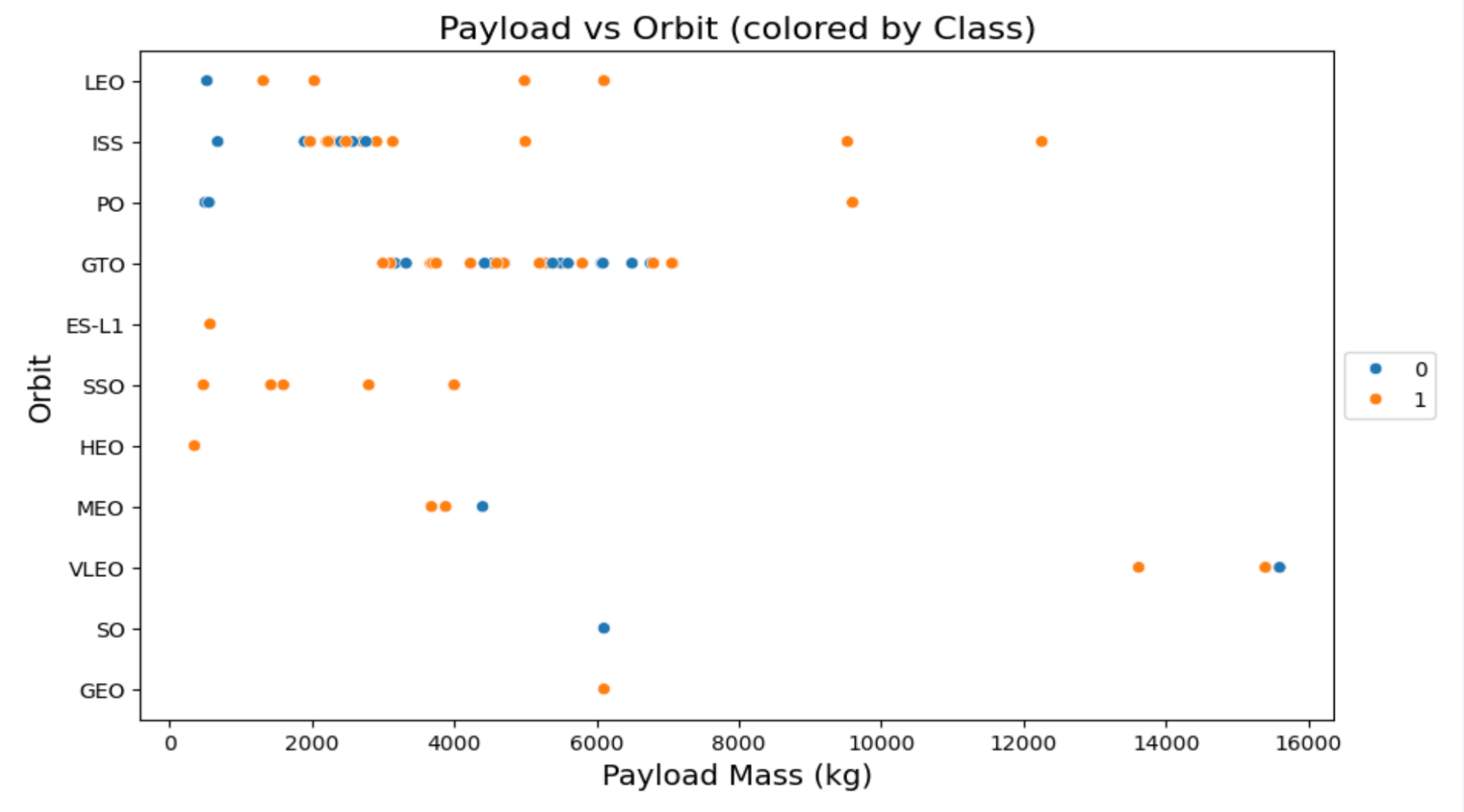
Flight Number vs. Orbit Type



Flight Number vs. Orbit Type

- **Distribution of Flights:** We can now see that GTO, LEO and ISS are dominant orbit types throughout the first 80 flights, with frequent launches. This suggests SpaceX primarily used Falcon 9 for these orbits in its earlier missions.
- **Later Diversification:** After flight number 60, we start to see more flights for other orbit types like VLEO, SSO and MEO, indicating SpaceX is using Falcon 9 for a wider variety of missions in later stages.
- **Less Frequent Orbits:** Orbits like ES-L1, HEO and MEO appear to have considerably fewer flights scattered throughout the graph. This suggests these orbits are either newer target destinations or require more specific mission parameters.

Payload vs. Orbit Type

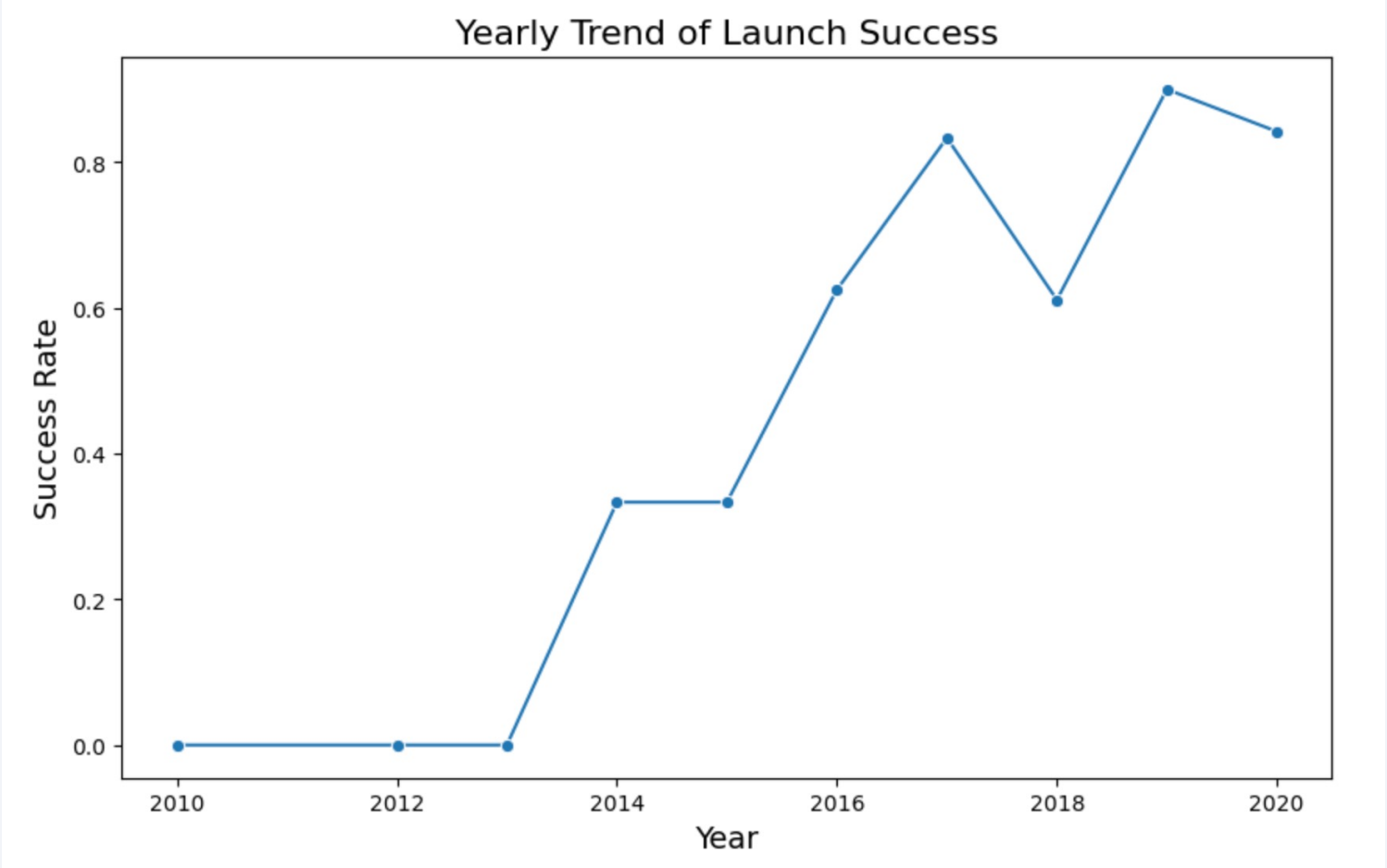


Payload vs. Orbit Type

- Most of the orbits only handle payload masses below 8000 kg.
- ISS, PO and VLEO are the three orbit types which have been used for payload masses heavier than 10000 kg.
- GTO and ISS have been the workhorses for SpaceX given they have handled most of the launches.

Launch Success Yearly Trend

- The success rate of SpaceX Falcon 9 launches has increased significantly over the years.
- From 40% in 2014, the success rate reached nearly 98% by 2020.
- This upward trend indicates SpaceX's continuous improvement in launch vehicle reliability.



All Launch Site Names

- %%sql
- SELECT DISTINCT "Launch_Site"
- FROM SPACEXTABLE;
- Query Result:
- Launch_Site
- CCAFS LC-40
- VAFB SLC-4E
- KSC LC-39A
- CCAFS SLC-40
- Name: LaunchSite, dtype: int64
- There are a total of four launch sites for SpaceX and the above query lists the unique launch sites.

Launch Site Names Begin with 'CCA'

- %%sql
- SELECT *
- FROM SPACEXTABLE
- WHERE "Launch_Site" LIKE 'CCA%'
- LIMIT 5;

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

- The above query uses the WHERE and the LIKE clauses to extract the records which match the ”LaunchSite” column values starting with “CCA”.

Total Payload Mass

- %%sql
- SELECT SUM("PAYLOAD_MASS__KG_")
- FROM SPACEXTABLE
- WHERE "Customer" = 'NASA (CRS)';
- **SUM("PAYLOAD_MASS__KG_"): 45596**

- The query calculates the sum of the Payload mass in kg from the SpaceX Table where the Customer is NASA.

Average Payload Mass by F9 v1.1

- %%sql
- SELECT AVG("PAYLOAD_MASS__KG_")
- FROM SPACEXTABLE
- WHERE "Booster_Version" = 'F9 v1.1’;
- **AVG("PAYLOAD_MASS__KG_")2928.4**

- Similar to the previous query, we calculate the average of the payload mass in kg wherever the Booster version is F9 v1.1.

First Successful Ground Landing Date

- %%sql
- SELECT MIN("Date")
- FROM SPACEXTABLE
- WHERE "Landing_Outcome" = 'Success (ground pad)';
- **MIN("Date")2015-12-22**

- In the above query, we calculate the minimum of the Date in the SpaceX table where the Landing outcome is successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

- %%sql
- SELECT "Booster_Version"
- FROM SPACEXTABLE
- WHERE "Landing_Outcome" = 'Success (drone ship)'
- AND "PAYLOAD_MASS__KG_" > 4000
- AND "PAYLOAD_MASS__KG_" < 6000;

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

- We use the above query to find the booster version where there was a successful landing outcome with the payload mass between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- %%sql
- SELECT "Landing_Outcome", COUNT(*) AS "Total"
- FROM SPACEXTABLE
- WHERE "Landing_Outcome" IN ('Success', 'Failure')
- GROUP BY "Landing_Outcome";
- **Landing_Outcome Total**
- **Failure 3**
- **Success 38**
- The above query combines the GROUP BY clause in the WHERE clause to calculate the Landing outcome results as a count of success and failure.

Boosters Carried Maximum Payload

- %%sql
- SELECT "Booster_Version"
- FROM SPACEXTABLE
- WHERE "PAYLOAD_MASS__KG_" = (
- SELECT MAX("PAYLOAD_MASS__KG_")
- FROM SPACEXTABLE
-);

- We use a subquery in the above query to identify the maximum payload mass and then select all those booster versions which carried the maximum payloads.

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

- We use the case function in SQL to first assign the month names from the Date column.
- Then we should select the landing outcome, booster version and launch site values.
- Finally, we extract the year from the date column using the substring function for the year 2015.

```
%%sql
SELECT
    CASE
        WHEN substr("Date", 6, 2) = '01' THEN 'January'
        WHEN substr("Date", 6, 2) = '02' THEN 'February'
        WHEN substr("Date", 6, 2) = '03' THEN 'March'
        WHEN substr("Date", 6, 2) = '04' THEN 'April'
        WHEN substr("Date", 6, 2) = '05' THEN 'May'
        WHEN substr("Date", 6, 2) = '06' THEN 'June'
        WHEN substr("Date", 6, 2) = '07' THEN 'July'
        WHEN substr("Date", 6, 2) = '08' THEN 'August'
        WHEN substr("Date", 6, 2) = '09' THEN 'September'
        WHEN substr("Date", 6, 2) = '10' THEN 'October'
        WHEN substr("Date", 6, 2) = '11' THEN 'November'
        WHEN substr("Date", 6, 2) = '12' THEN 'December'
        ELSE 'Unknown'
    END AS "Month",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE substr("Date", 0, 5) = '2015'
    AND "Landing_Outcome" = 'Failure (drone ship)';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- We first need to group the table records by landing outcome and then filter based on the date between 2010-06-04 and 2017-03-20.
- Finally, we need to calculate the count by landing outcome in descending order using the ORDER BY function.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS "Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY "Count" DESC;
```

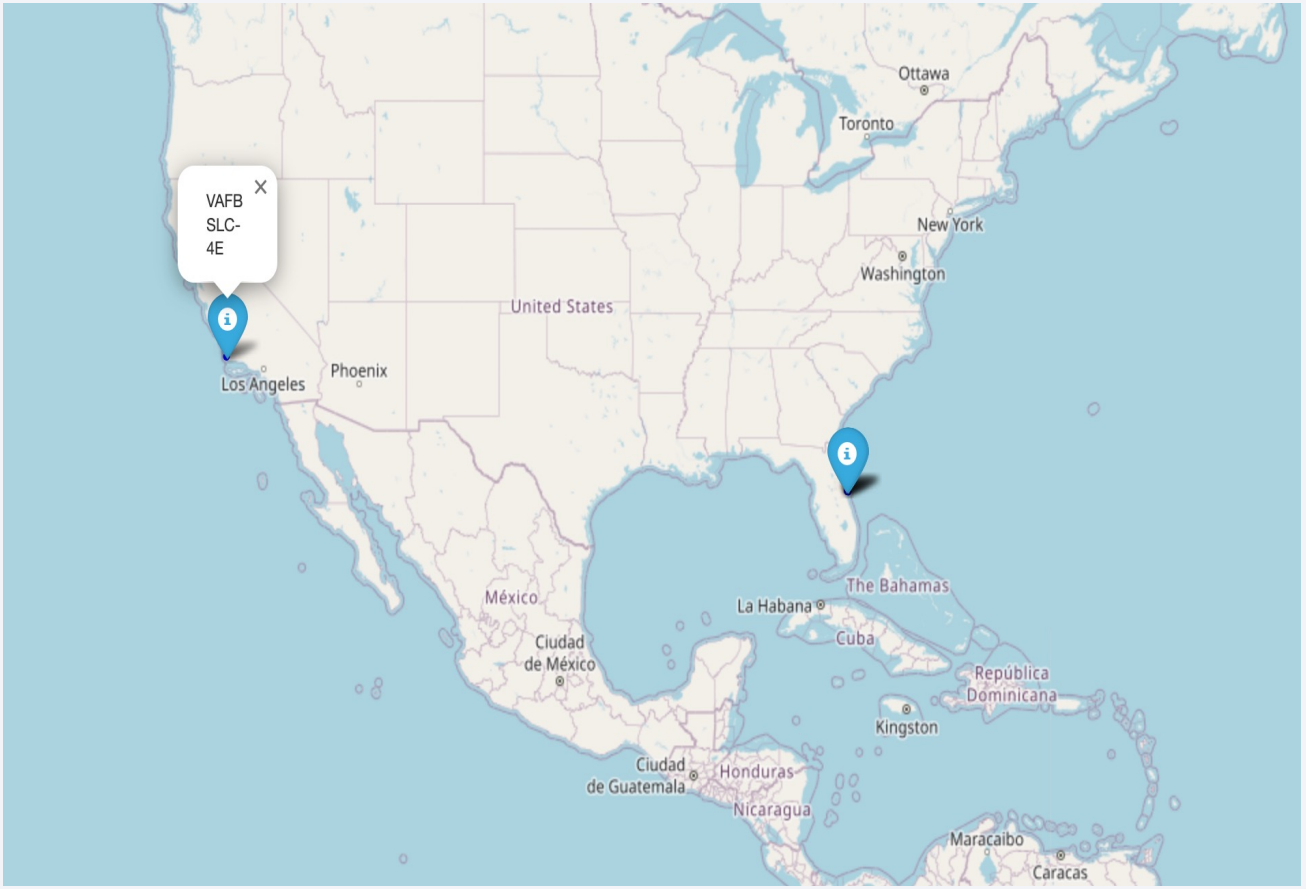
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

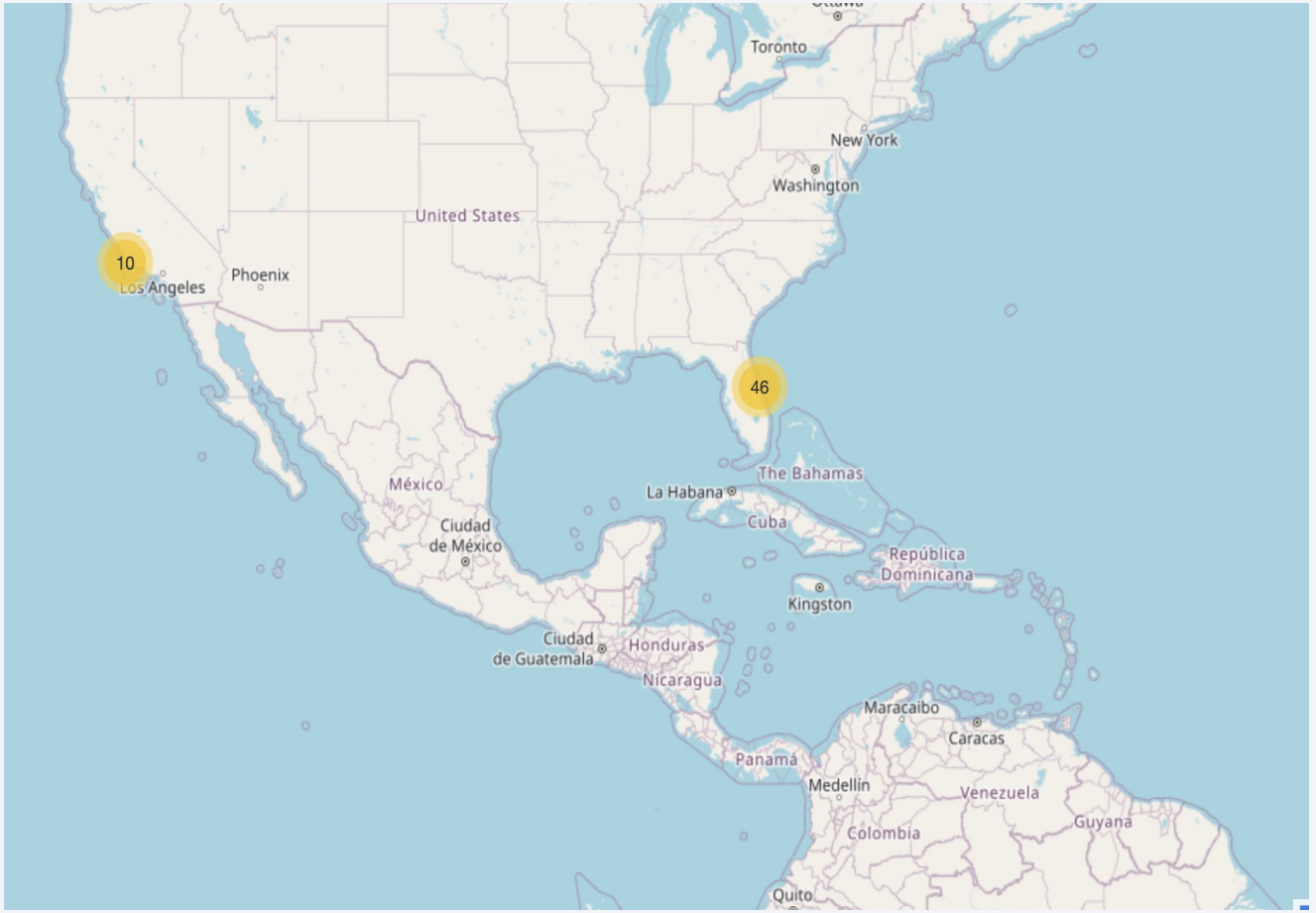
SpaceX Launch sites on map

- This map shows the launch sites in the two coasts of USA utilized by SpaceX for their launches.



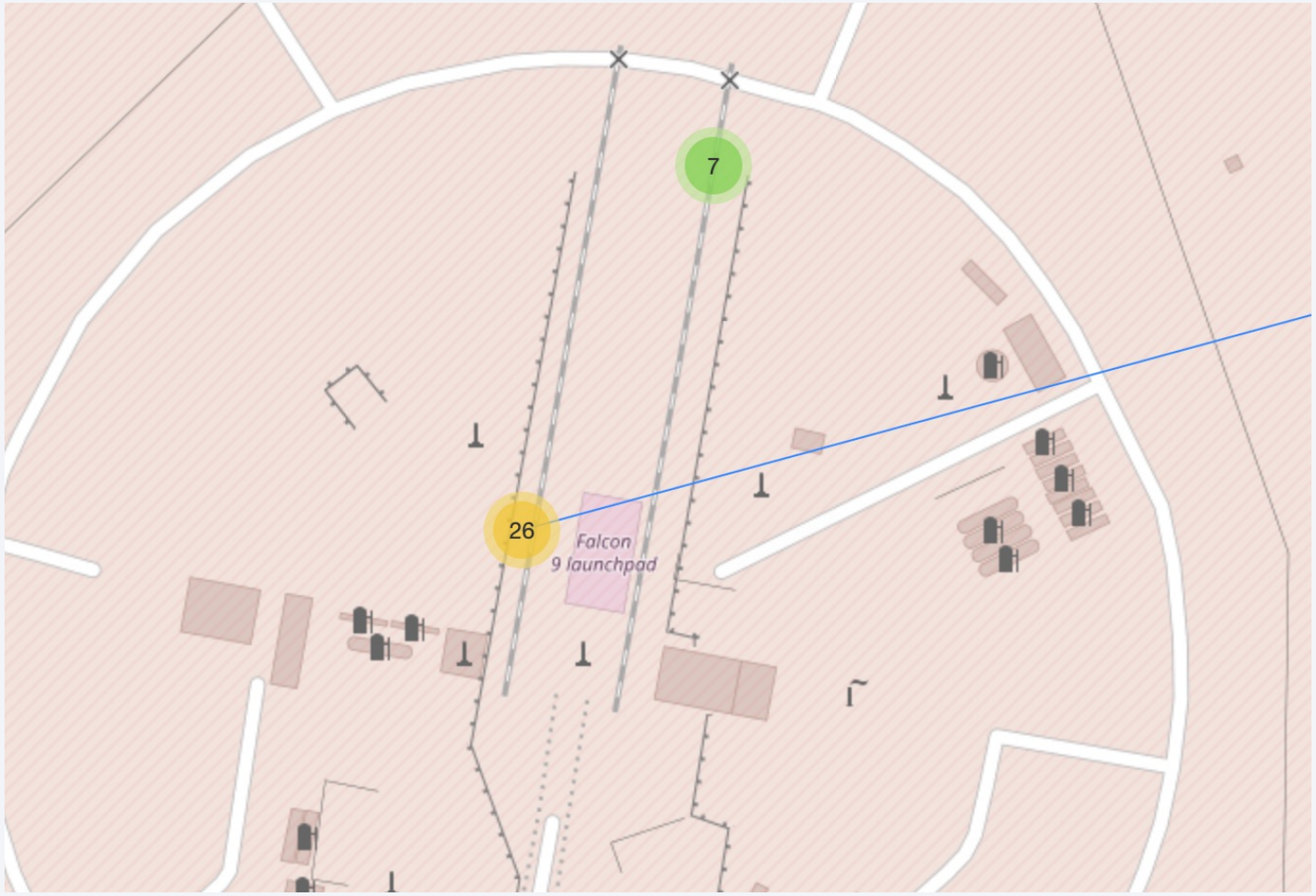
Launch outcome on each site

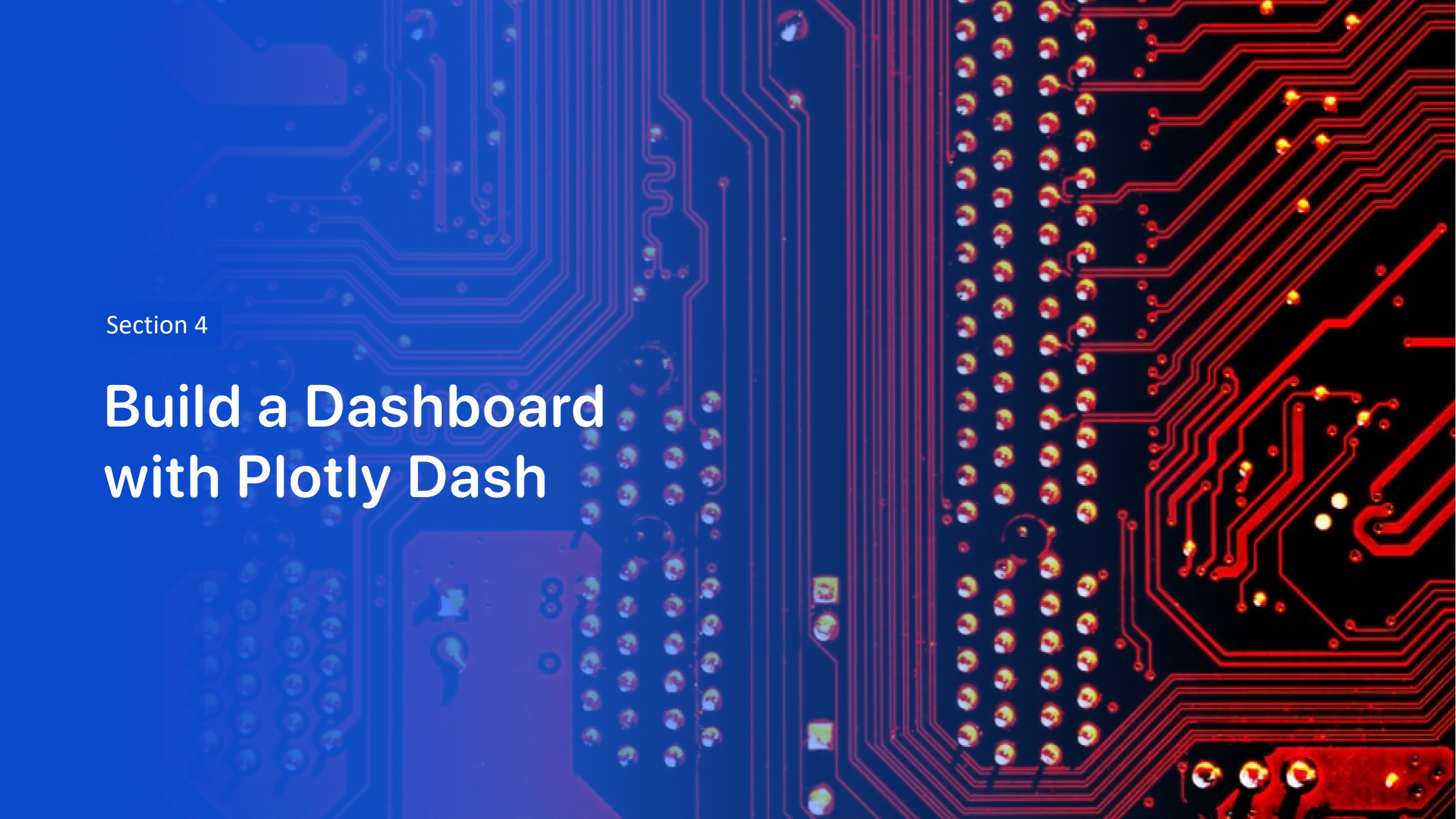
- This map shows the number of successful launches from both coast sites marked with the number circled in yellow.



Launch site and their proximities

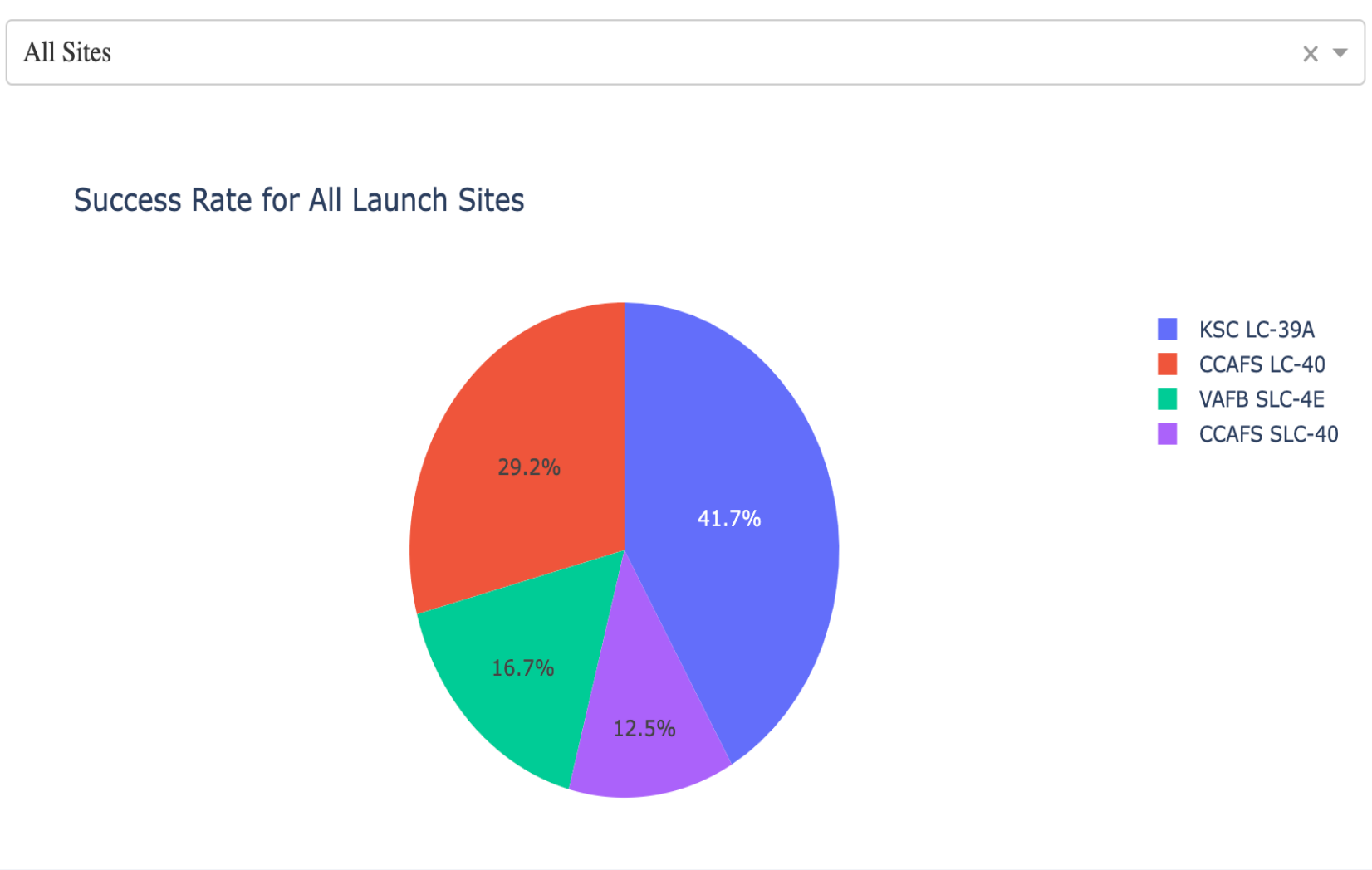
- This map shows the proximities of the nearby transportations to the sites present in the eastern coast.




Section 4
Build a Dashboard with Plotly Dash

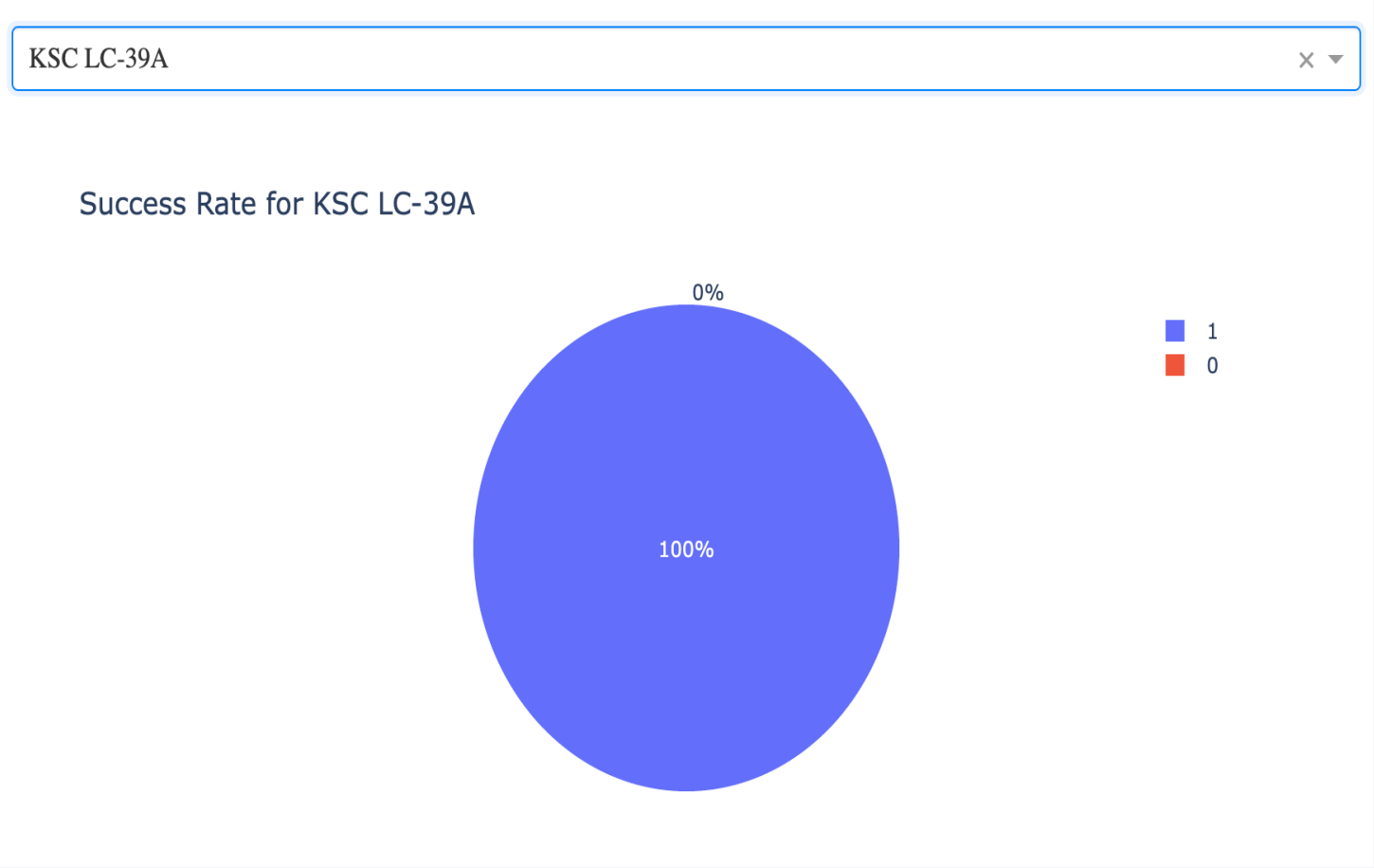
Outcome success rate of all launch sites

- The different success rates for each of the launch sites are shown in the pie chart.
- The site KSC LC39A has the highest success rate of all and it's followed by CCAFS LC40.



Outcome success rate of the successful launch site

- This shows the pie chart for site with the highest successful rate of launch outcome.
- It holds 41% of the success rates of all sites.



Outcome success rate at Payload Mass – 2000 kg

Outcome success rate at Payload Mass – 3000 kg

Outcome success rate at Payload Mass – 6000 kg

Outcome success rate at different Payload Mass

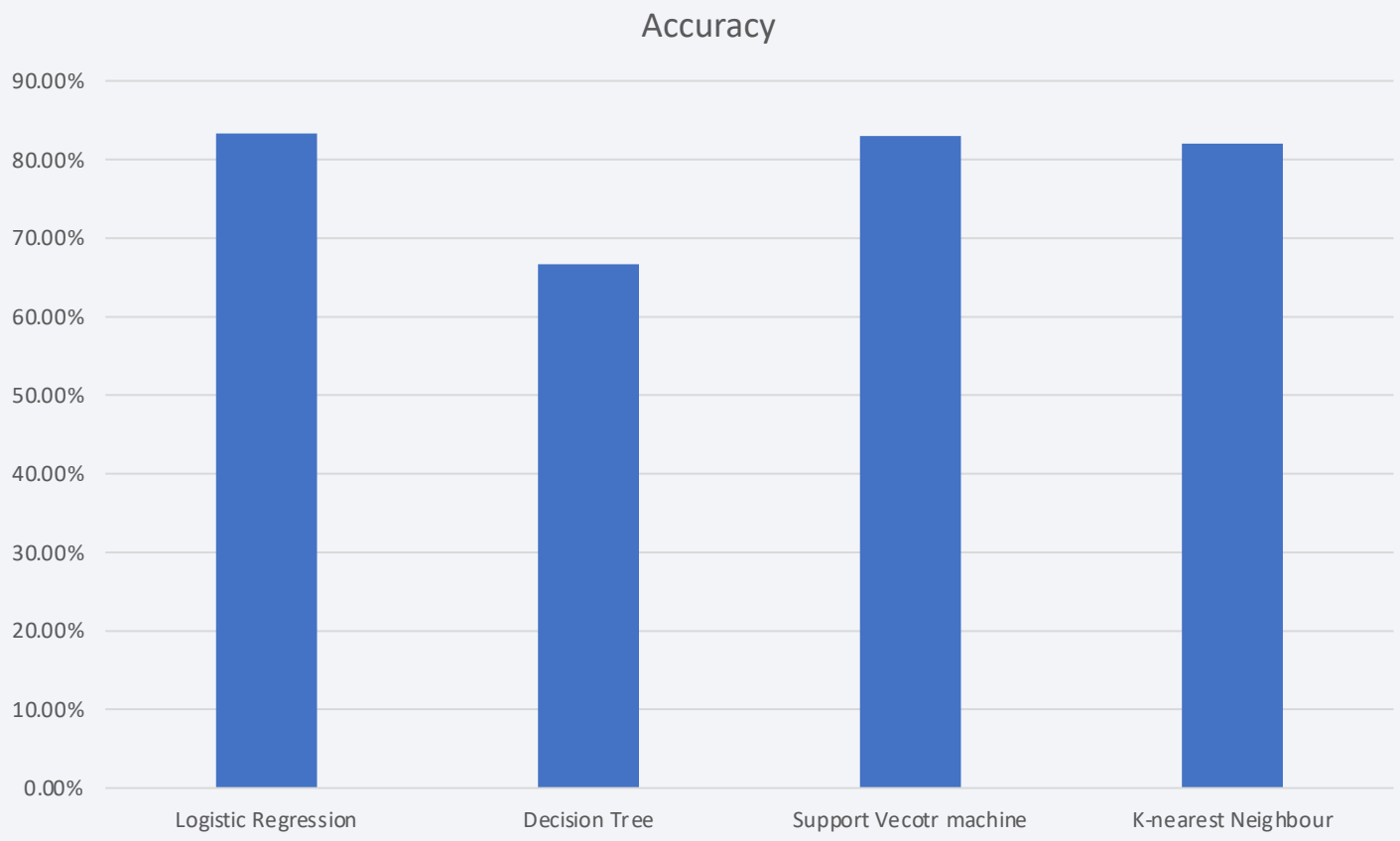
- We could observe that with the change in Payload mass at different kgs, the outcome success rate changes as shown in the three different screenshots.
- We could also observe that the maximum number of mission launches happen at lower payload mass.
- This dashboard will be very useful to identify the success rates of different payload mass and accordingly plan the mission for future.

Section 5

Predictive Analysis (Classification)

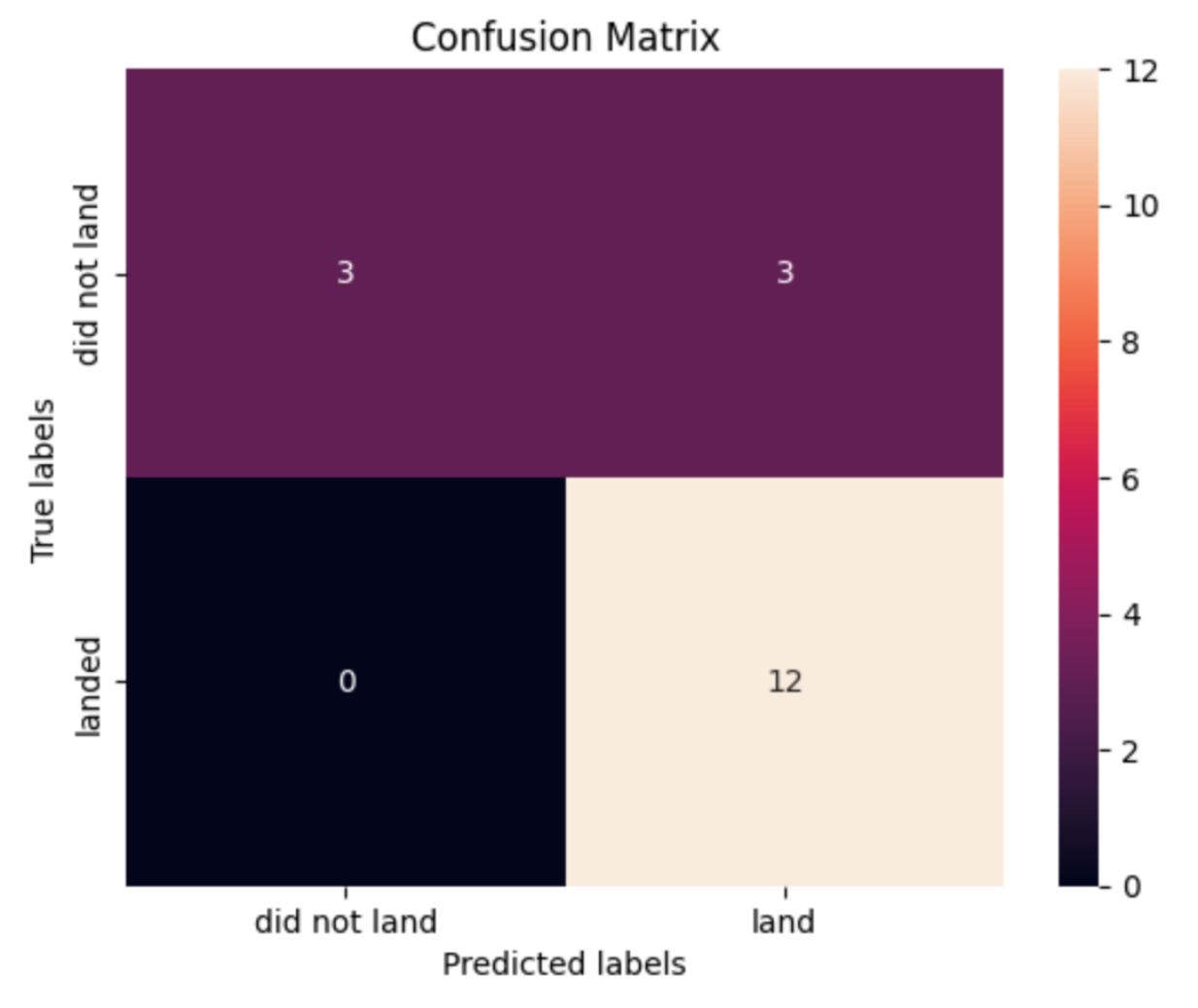
Classification Accuracy

- Logistic Regression model has slightly higher accuracy than support vector machine and KNN algorithms.



Confusion Matrix

- True Positives = 12
- True Negatives = 3
- False Positives = 3
- False Negatives = 0



Conclusions

- **Model Performance:**
- The model performed well in predicting successful launches, with a **precision of 0.8** . This means that out of all the launches predicted to be successful, 80% were successful.
- The model perfectly captured all the successful launches (**recall of 1.0**), which means it did not miss any successful launches.
- Overall, the model has an **accuracy of 83.33%**, which indicates a good performance in correctly classifying both successful and unsuccessful launches.
- **False Positives:**
- There were **3 False Positives (FP)**, which means the model predicted 3 failed launches as successful. It is important to investigate these failures to understand why the model misclassified them.
- **Zero False Negatives:**
- Importantly, there were **zero False Negatives (FN)**. This means the model correctly identified all the successful launches and did not predict any successful launch as a failure. This is a positive outcome, as failing to predict a successful launch could lead to missed opportunities or delays in missions.

Thank you!