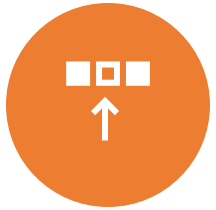
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bamwesigye Calvin Kiiza

24/November/2023

Outline



EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Methodology

Data Collection and Preparation: Acquired comprehensive datasets related to Falcon 9 launches, encompassing key parameters such as flight number, payload mass, launch site, success outcomes.

Exploratory Data Analysis (EDA): Conducted extensive data exploration and preprocessing to gain insights and ensure data quality, handling missing values, encoding categorical variables, and feature engineering.

MODEL DEVELOPMENT: EMPLOYED VARIOUS MACHINE LEARNING ALGORITHMS INCLUDING LOGISTIC REGRESSION, SUPPORT VECTOR MACHINES (SVM), DECISION TREES, AND K-NEAREST NEIGHBORS (KNN) TO TRAIN PREDICTIVE MODELS.

HYPERPARAMETER TUNING AND MODEL EVALUATION: UTILIZED GRIDSEARCHCV TO OPTIMIZE MODEL PARAMETERS AND ASSESSED MODEL PERFORMANCE USING RELEVANT EVALUATION METRICS LIKE ACCURACY, PRECISION, AND RECALL.

Results

Model Performance: The predictive models exhibited varying levels of accuracy in determining the first stage landing outcome. SVM emerged as the most accurate model, achieving an accuracy rate of 83% on the test dataset.

Insights and Implications: The developed predictive model provides valuable insights into estimating the success probability of Falcon 9 first stage landings. This information can guide cost estimation and strategic decision-making for organizations exploring alternatives in the rocket launch market.

Introduction

- SpaceX, a pioneering private aerospace manufacturer, has disrupted the space industry with innovations aiming to reduce the cost of space travel. One of its revolutionary creations, the Falcon 9 rocket, presents a significant advancement in reusable rocket technology. The key to cost-effectiveness lies in the ability to land and reuse the first stage of the Falcon 9 rocket.
- This project centers on predicting the successful landing of the Falcon 9 first stage. The predictive model aims to leverage historical data from previous Falcon 9 launches to estimate the probability of a successful landing. By doing so, it aspires to provide a crucial tool for estimating mission reliability and associated costs for stakeholders and potential bidders in the rocket launch industry.
-

Problem statement

The fundamental inquiry driving this project revolves around the anticipation of Falcon 9 first stage landings. The primary aim is to ascertain the probability of a successful first stage landing post-launch, thereby facilitating a comprehensive assessment of mission feasibility and cost projections.

•

Section 1

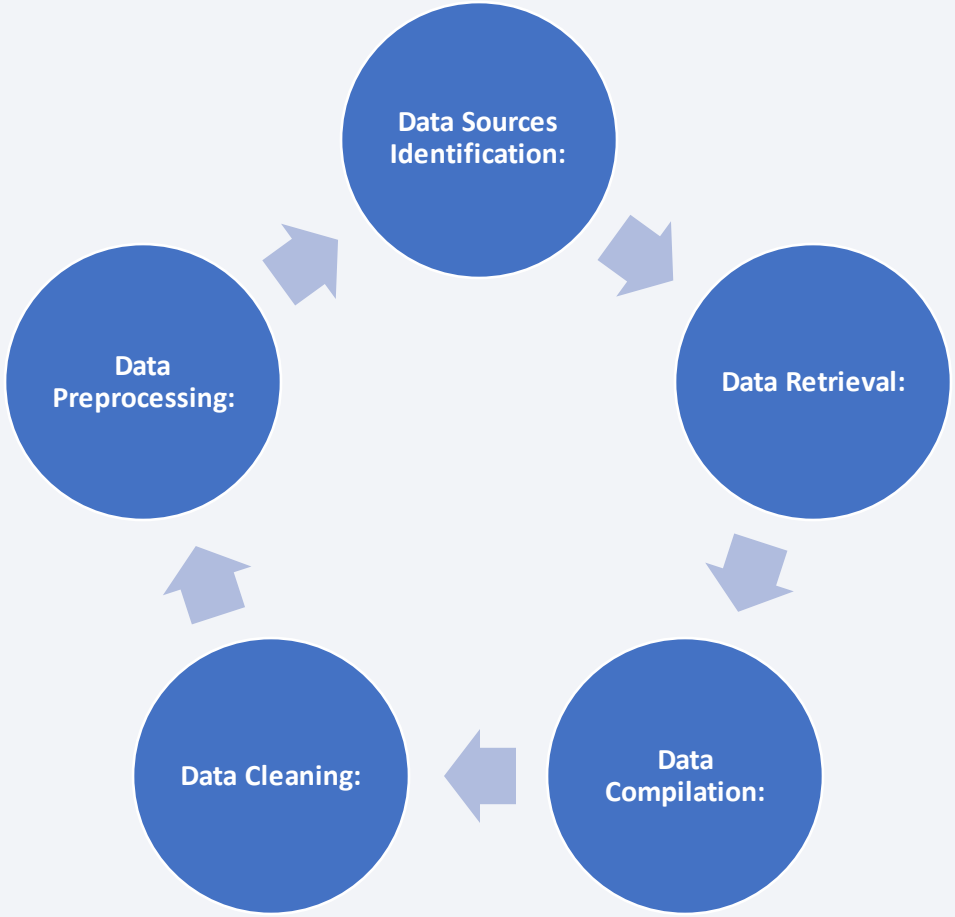
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describes how data was collected
- Perform data wrangling
  - Describes 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 Sources Identification:

- Identified and selected primary sources: SpaceX historical records, space databases, and official documentation.
- Secondary sources may include space agencies' data repositories or open-source space exploration databases.

Data Retrieval:

- Extracted relevant data from SpaceX databases and APIs or other publicly available data sources.
- Used web scraping techniques or accessed APIs to retrieve Falcon 9 launch details.

Data Compilation:

- Collated Falcon 9 launch records from different sources into a unified dataset.
- Ensured consistency and uniformity in data format across different sources.

Data Cleaning:

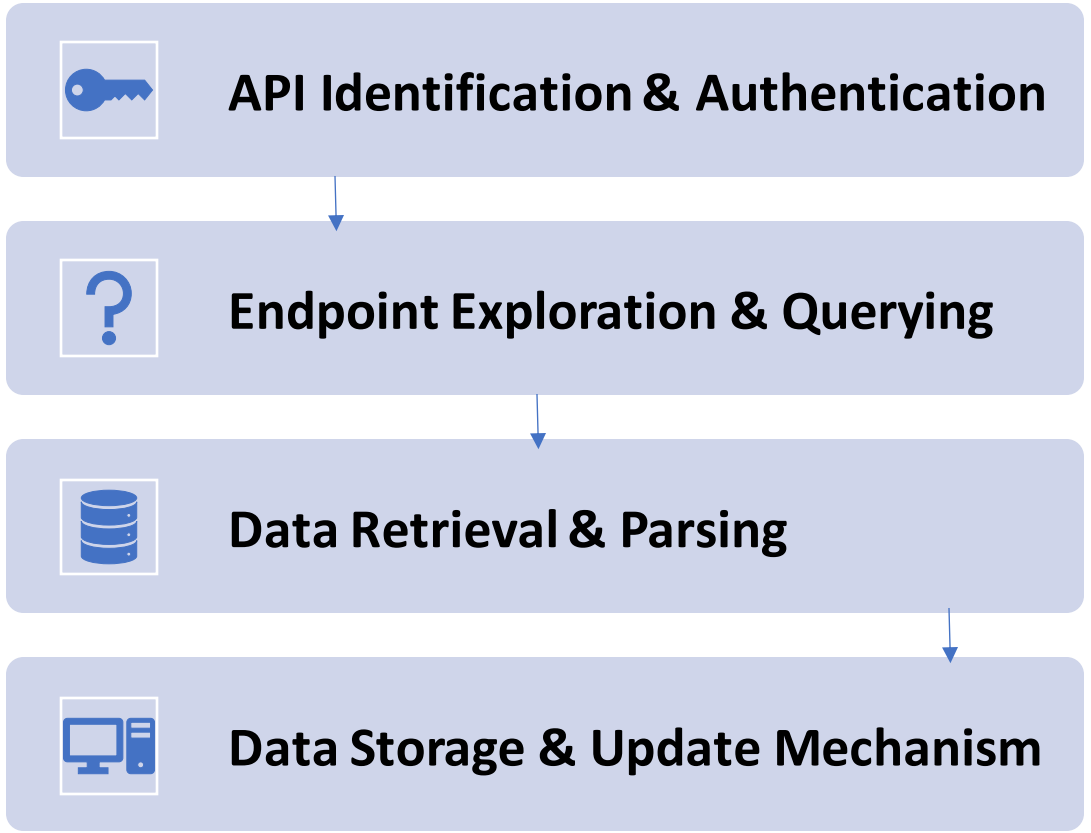
- Detected and rectified inconsistencies, missing values, or outliers within the collected dataset.
- Performed data validation and quality checks to ensure accuracy.

Data Preprocessing:

- Preprocessed the collected data by formatting and structuring it for analysis.
- Normalized numerical data and encoded categorical variables for machine learning compatibility.

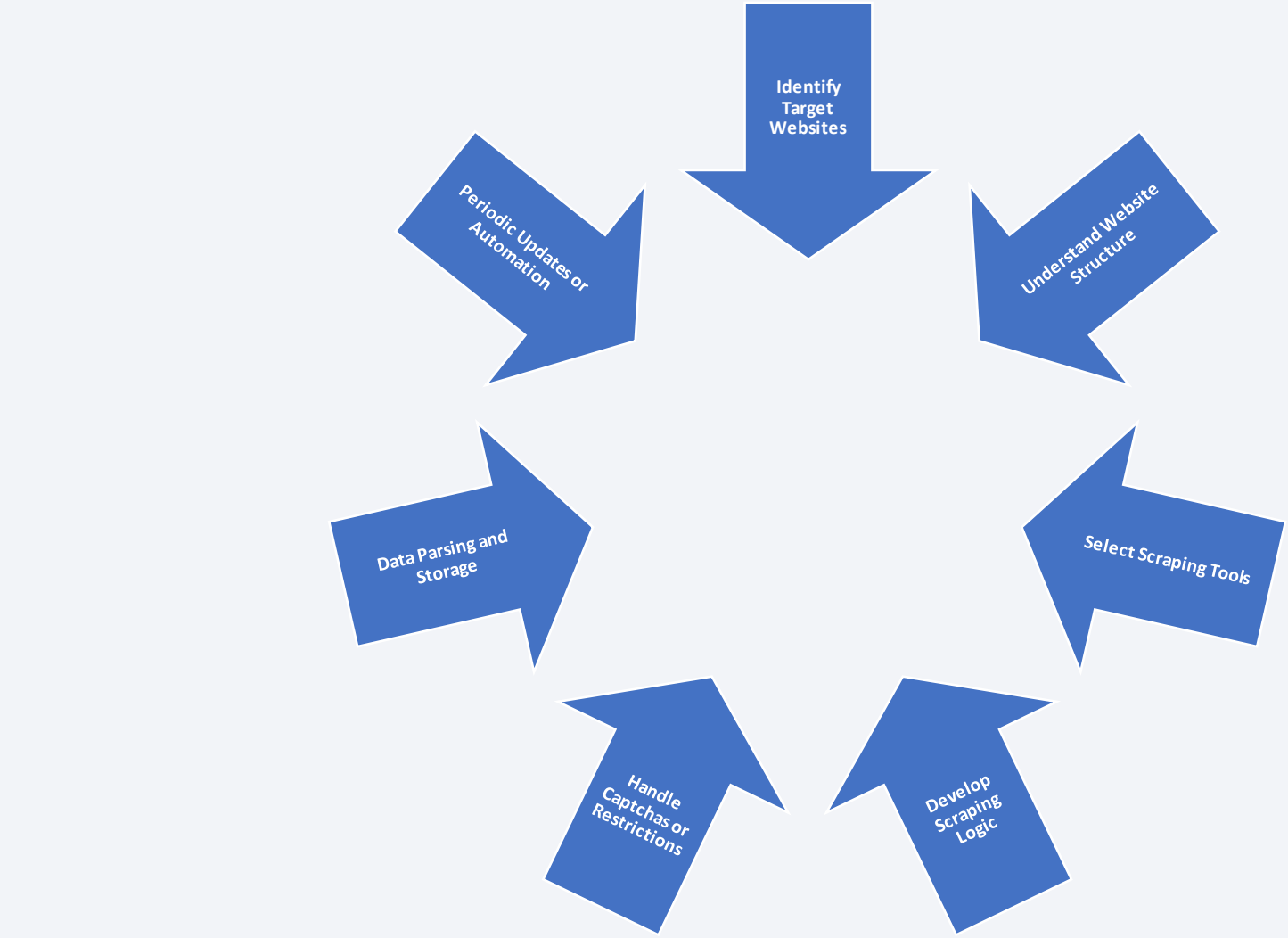
Data Collection – SpaceX API

API calls and results



Data Collection - Scraping

- webscrapping notebook



Data Wrangling

Data Cleaning and Preprocessing:

- **Handling Missing Values:** Imputation, deletion, or estimation of missing data points.
- **Dealing with Duplicates:** Removing or deduplicating identical records.
- **Standardizing Data Formats:** Ensuring consistency in date formats, units, or other data formats.
- **Error Correction:** Identifying and rectifying inaccuracies or inconsistencies in the dataset.

Data Transformation and Feature Engineering:

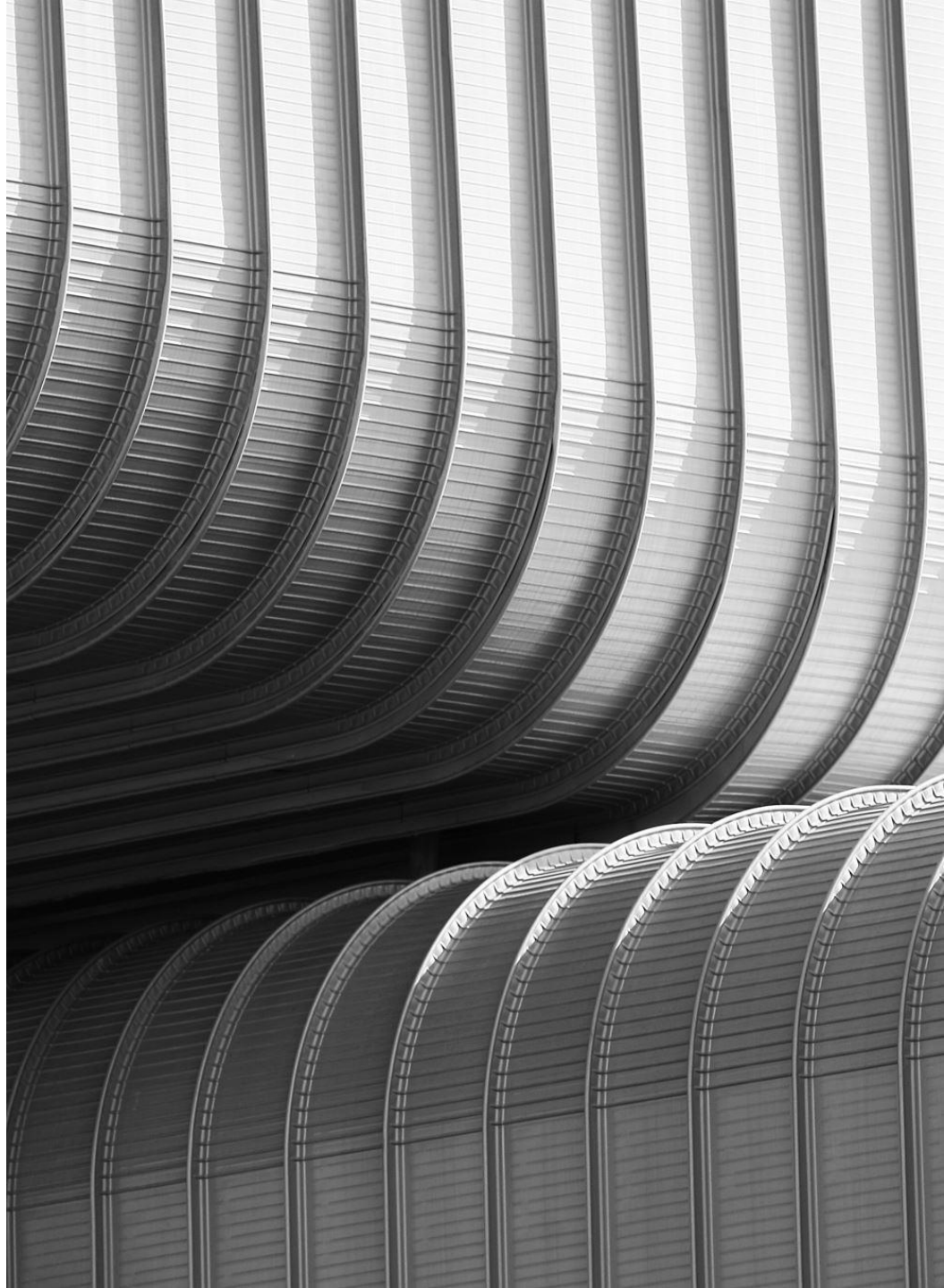
- **Creating New Features:** Deriving additional features from existing ones to enhance predictive power.
- **Scaling and Normalization:** Rescaling features to comparable ranges, preventing dominance by certain features.
- **Encoding Categorical Variables:** Converting categorical variables into numerical representations for modeling.
- **Text Preprocessing:** Tokenization, stemming, or lemmatization for natural language processing tasks.

Data Integration and Aggregation:

- **Merging Datasets:** Combining multiple datasets based on common keys or variables.
- **Aggregating Data:** Summarizing or grouping data to a desired level for analysis.

Validation and Quality Assurance:

- **Data Quality Checks:** Verifying data accuracy, consistency, and adherence to predefined rules.
- **Outlier Detection and Handling:** Identifying and managing outliers that might impact model performance



EDA with Data Visualization

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

  This plot indicates how the payload mass varies concerning different launch sites. It helps in understanding if certain launch sites tend to accommodate heavier or lighter payloads.

- **Orbit and Mean Variations Bar Chart:**

  This chart illustrates the mean variations within different orbits. It provides an overview of the average variability observed among various orbit types.

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

  This scatter plot visualizes the distribution of flight numbers across different orbits. It helps in analyzing the frequency and distribution of flights among different types of orbits.

- **Success Rate Over Years:**

  Depicts the success rate of SpaceX launches across different years. This chart enables the observation of trends in launch success over time, allowing for insights into the historical performance of SpaceX missions.

- Completed EDA

EDA with SQL

- Displayed the names of unique launch sites in the space mission.
- Retrieved 5 records where launch sites begin with 'CCA'.
- Calculated the total payload carried by boosters from NASA.
- Calculated the average payload mass carried by booster version F9 v1.1.
- Retrieved the dates of the first successful landing outcome on a ground pad.
- Listed the names of boosters that successfully landed on a drone ship with a payload mass between 4000 and 6000.
- Calculated the total number of successful and failed mission outcomes.
- Listed the names of boosters with the maximum payload mass carried.
- Retrieved failed landing outcomes in drone ships, along with their booster versions and launch site names for the year 2015.
- Ranked the count of landing outcomes between specific dates in descending order.

- EDA SQL

Build an Interactive Map with Folium

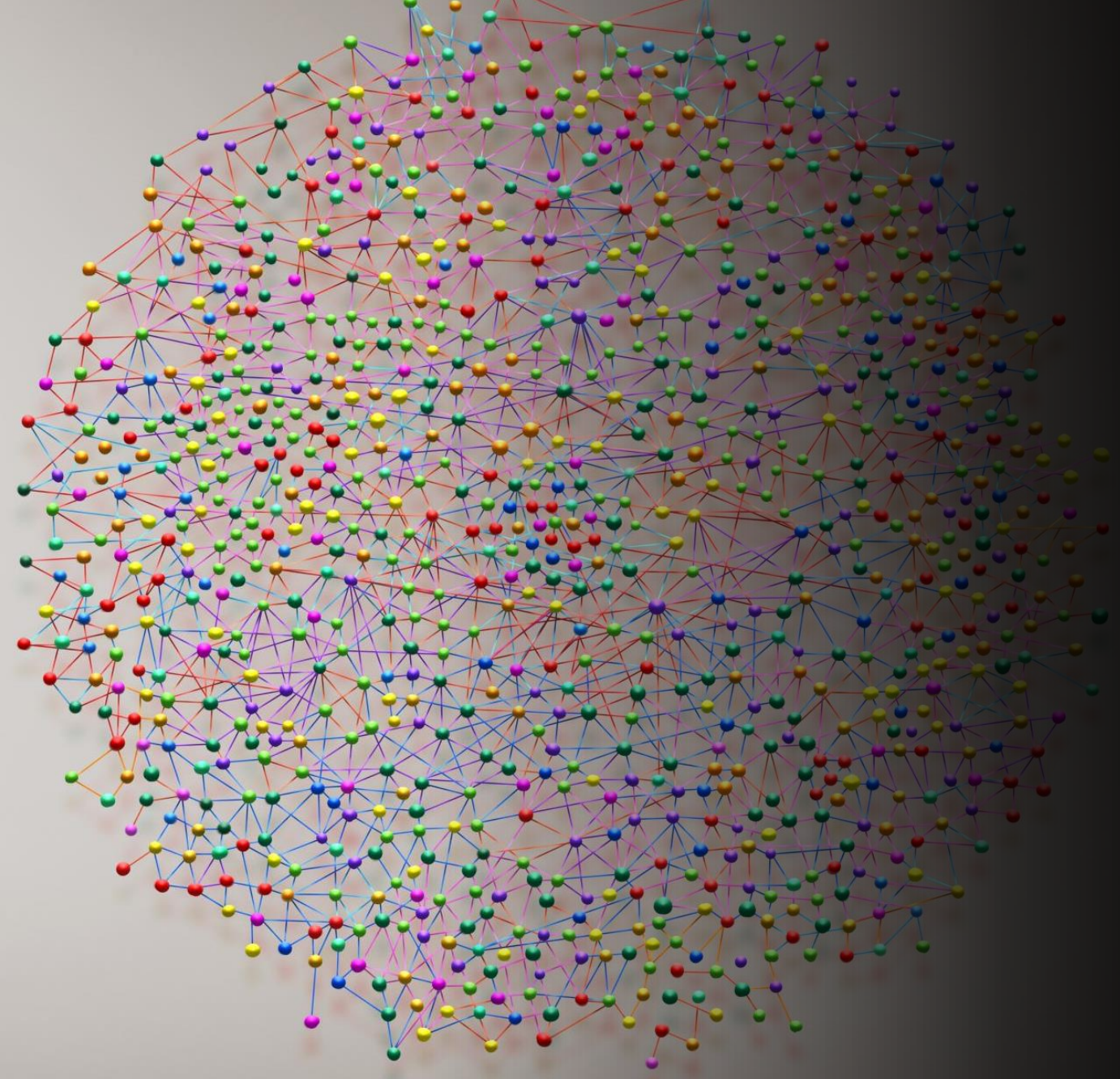
Markers:

Markers were placed to represent the specific launch sites used by SpaceX. Each marker denotes a unique launch site, making it easier to visualize the geographical distribution of launch locations.

Circles:

Circles were incorporated to highlight specific areas around each launch site. These circles indicate certain radii, aiding in visualizing the approximate reach or impact areas surrounding the launch sites. They might represent safety zones, restricted areas, or other important spatial boundaries.

MAPS PLOTTED



- **Geospatial Visualization:** Markers help to geographically pinpoint each launch site on the map.
- **Area Representation:** Circles provide an immediate understanding of the radius around each launch site, showing spatial constraints or safety zones.
- **Enhanced Understanding:** Together, these objects enhance the visual comprehension of the geographical distribution of launch sites and their surrounding areas.

Build a Dashboard with Plotly Dash

- **Success Pie Chart:**

  This chart displays the distribution of successful and failed launches. It aids in understanding the success rates across different launch sites or for all sites combined, providing a quick overview of mission outcomes.

- **Payload Range Slider:**

  This slider enables users to select a specific payload range, allowing them to filter launches based on payload mass. It helps in understanding the relationship between payload mass and launch success.

- **Success Payload Scatter Chart:**

  This scatter chart depicts the correlation between payload mass and launch success. It allows users to observe how varying payload masses affect the success rate of launches.

  - DASH BOARD

Predictive Analysis (Classification)

Model Building:

Employed various classification algorithms like Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors.

Developed initial models using default parameters.

Model Evaluation:

Assessed model performance metrics such as accuracy, precision, recall, and F1-score.

Used cross-validation techniques to ensure model generalization.

Hyperparameter Tuning:

Utilized GridSearchCV or similar techniques to fine-tune model hyperparameters.

Tuned parameters to optimize model performance.

Model Selection:

Selected the best-performing model based on evaluation metrics.

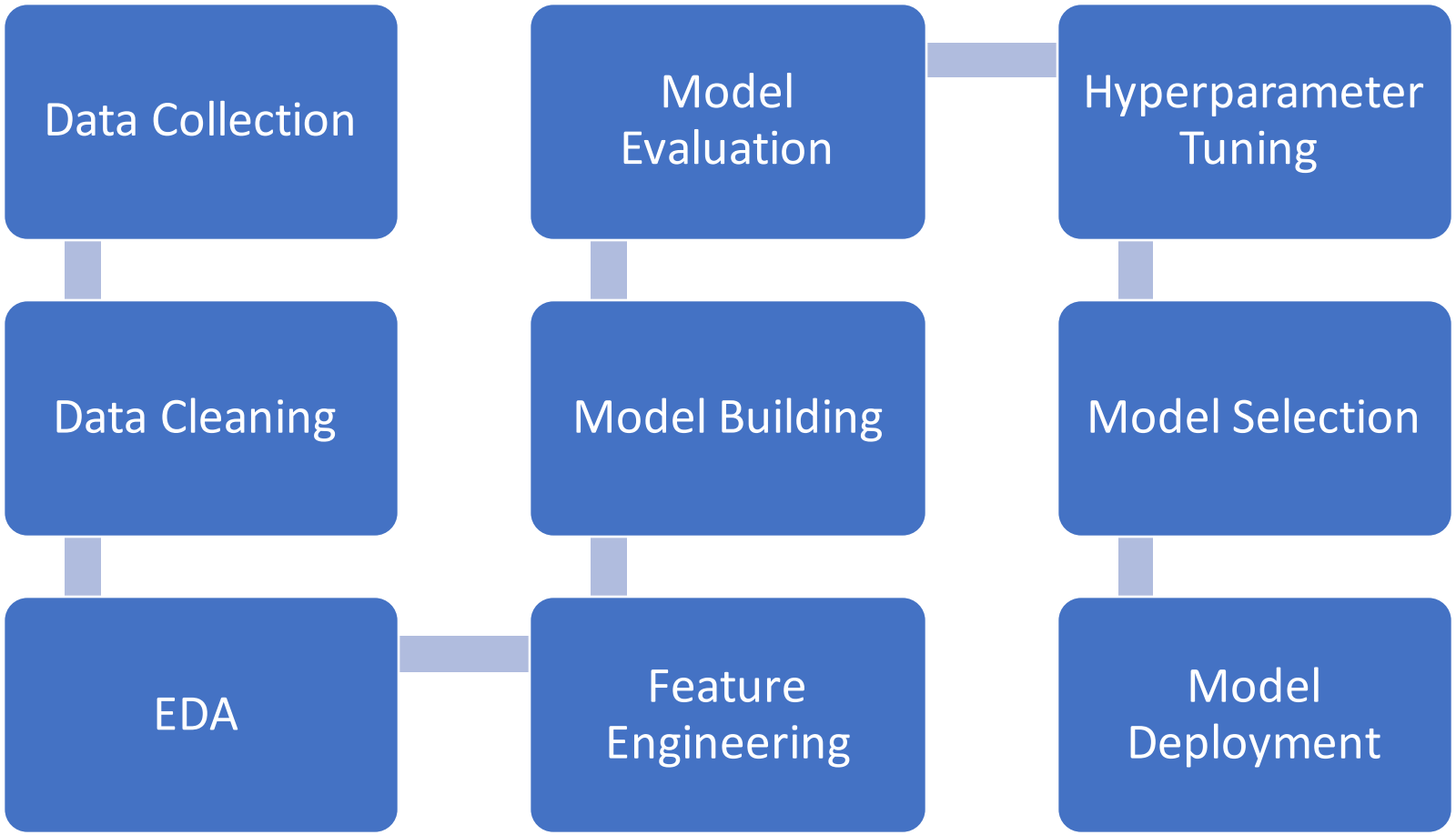
Improvement Iterations:

Iteratively refined models by adjusting features, algorithms, and hyperparameters.

Final Model Deployment:

Deployed the best-performing model for prediction and validation on new or test data.

Flowchart

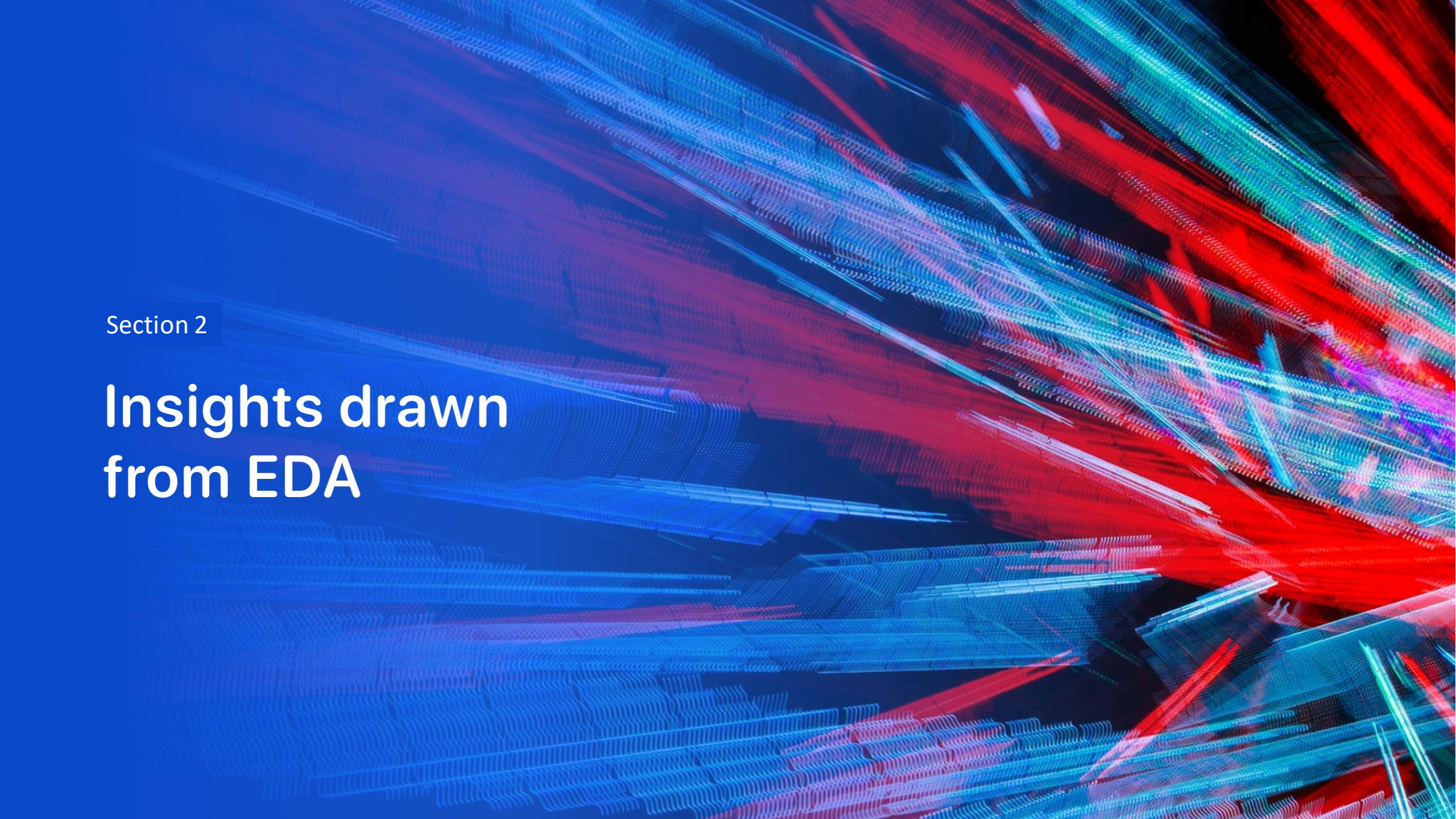

Data Collection
Data Cleaning
EDA
Feature Engineering
Model Building
Model Evaluation
Hyperparameter Tuning
Model Selection
Model Deployment

Results

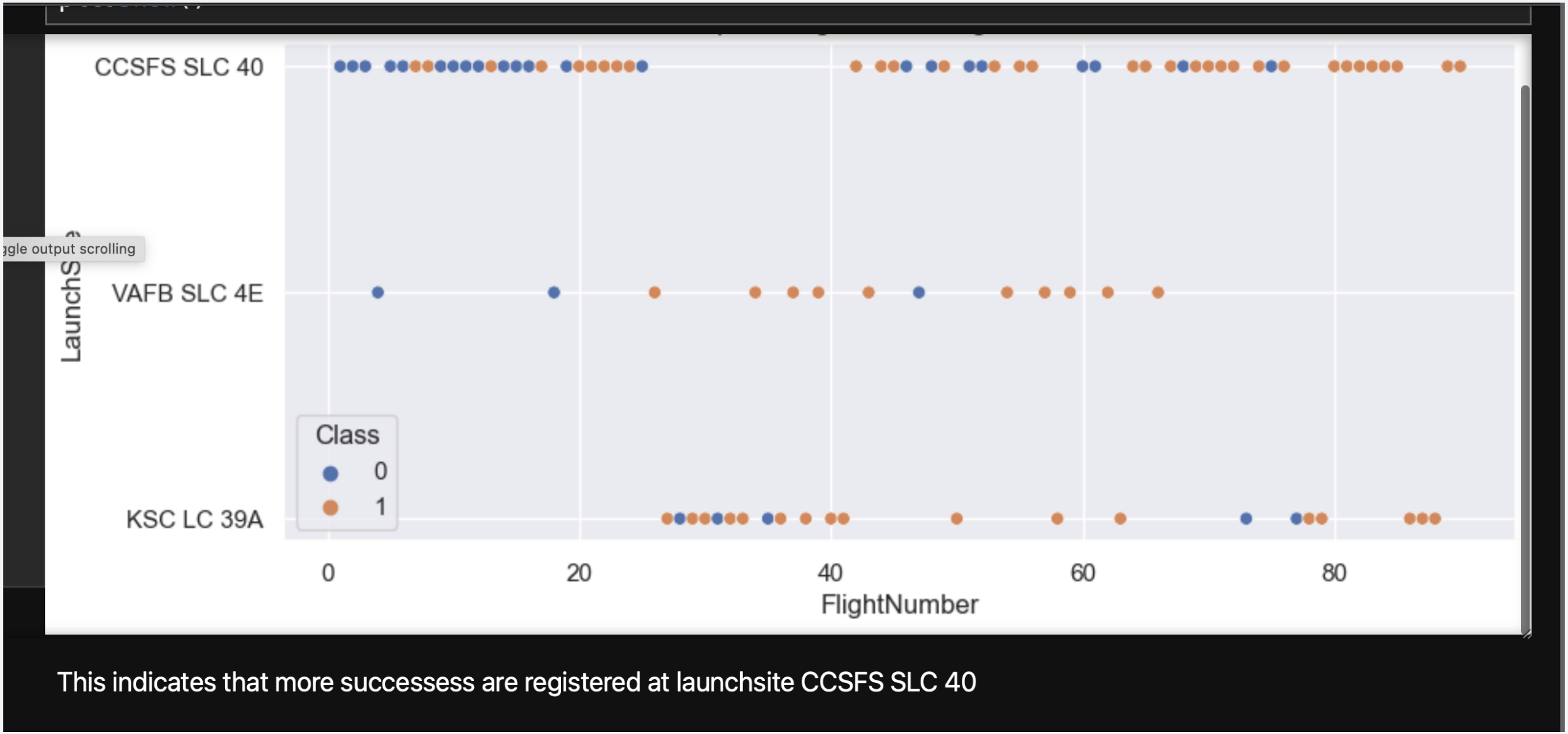
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA



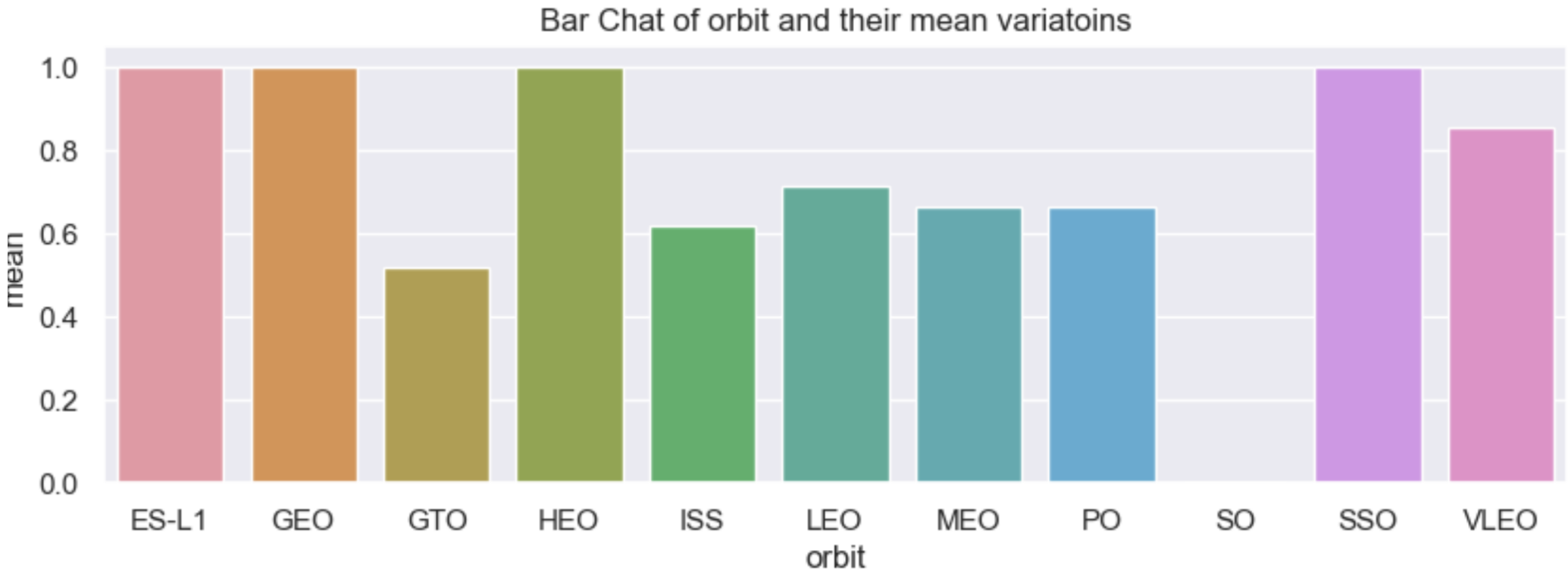
Flight Number vs. Launch Site



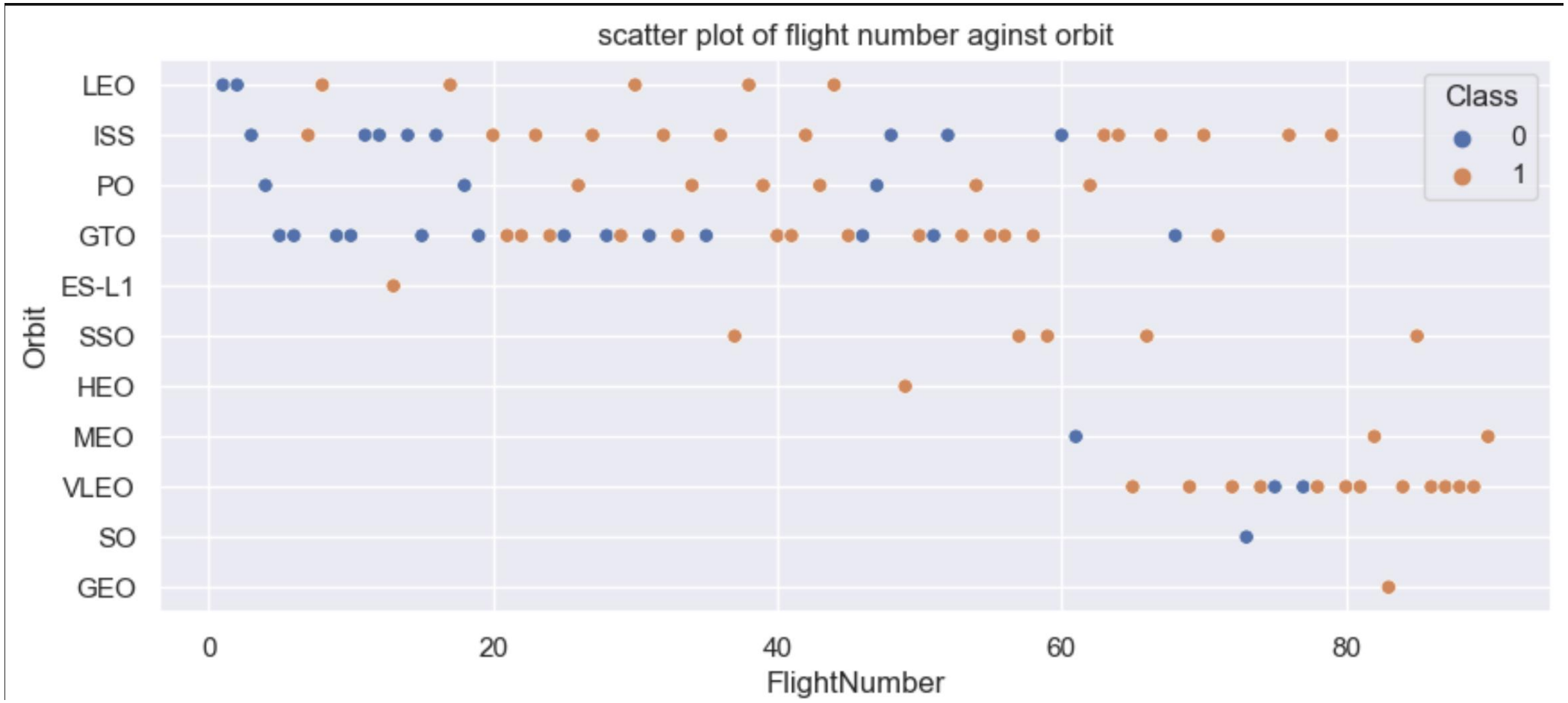
This indicates that more successess are registered at launchsite CCSFS SLC 40

Payload vs. Launch Site

Success Rate vs. Orbit Type

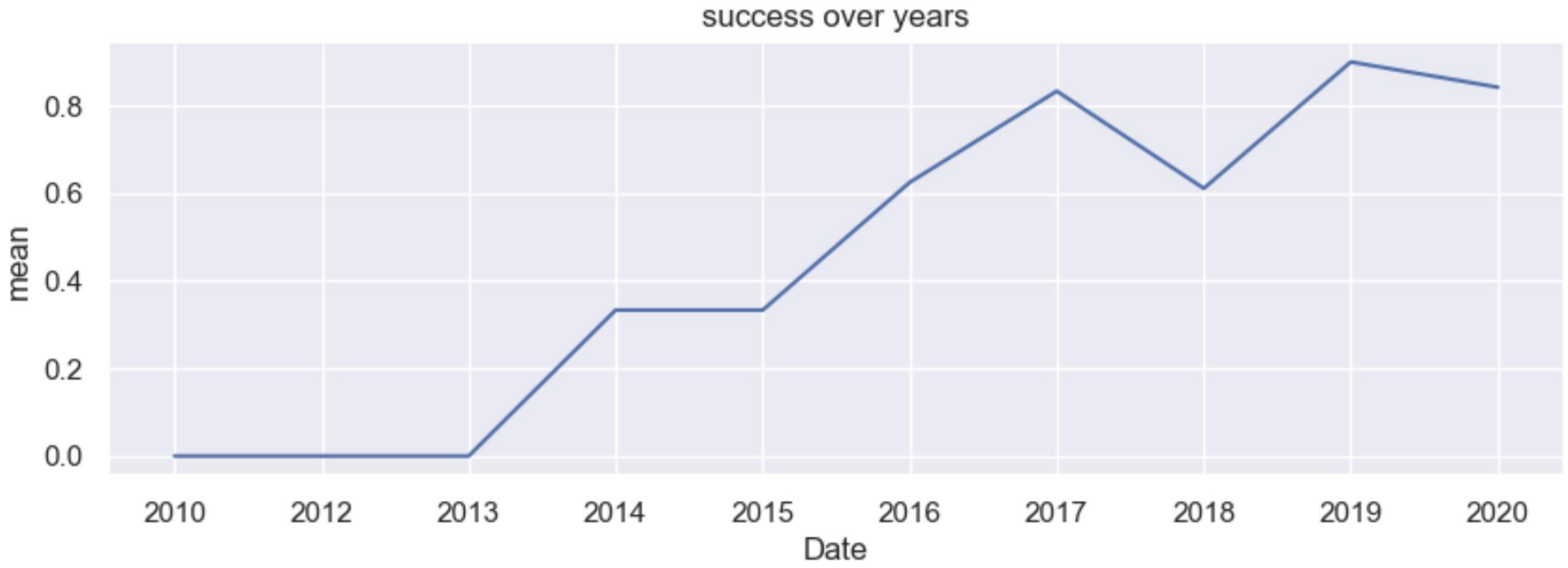


Flight Number vs. Orbit Type



Launch Success Yearly Trend

- This shows how the launch successes have been increasing over the years



All Launch Site Names

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

- "%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;"
- The sql statement fetches unique values with the help of DISTINCT key word from launch site column from a table named spacextbl

Launch Site Names Begin with 'CCA'

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

- **SELECT ***: Retrieves all columns from the table.
- **FROM SPACEXTBL**: Specifies the table named 'SPACEXTBL' from which the data will be fetched.
- **WHERE "LaunchSite" LIKE 'CCA%'**: Filters the records based on the condition that the "LaunchSite" column values must start with 'CCA'. The **LIKE** operator is used to match patterns, and 'CCA%' signifies that the "LaunchSite" values should begin with 'CCA', followed by any characters.
- **LIMIT 5**: Limits the output to only 5 records.

Total Payload Mass

- SELECT SUM("PayloadMass") FROM SPACEXTBL WHERE "CustomerName"='NASA';

- The SQL query selects the sum of the "PayloadMass" column from the 'SPACEXTBL' table where the "CustomerName" is 'NASA'. It computes the total payload carried by boosters associated with NASA missions

Average Payload Mass by F9 v1.1

- %sql SELECT AVG("PayloadMass") AS "AveragePayloadMass_F9_v1.1" FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1";
- This SQL query uses the **SELECT** statement to calculate the average payload mass (**PayloadMass**) carried by booster version 'F9 v1.1' from the 'SPACEXTBL' table. The **AVG()** function calculates the average value of the payload mass column. The **WHERE** clause filters the data to only include records where the booster version matches 'F9 v1.1'. The result is presented as 'AveragePayloadMass_F9_v1.1'.

First Successful Ground Landing Date

- %sql SELECT MIN(Date) AS First_Successful_Landing_On_Ground_Pad FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';

- This SQL query utilizes the **MIN()** function to find the earliest date of a successful landing outcome specifically on the ground pad. It searches the **SPACEXTBL** table for entries where the **Landing_Outcome** column matches **'Success (ground pad)'**. The **MIN(Date)** function then identifies and retrieves the earliest date associated with this particular successful landing outcome on the ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- **SELECT "Booster_Version"**: Specifies the column to be retrieved, which is the Booster_Version.
- **FROM SPACEXTBL**: Specifies the table name from which to retrieve data, in this case, the SPACEXTBL table.
- **WHERE "Landing_Outcome" = 'Success (drone ship)'**: Filters the rows where the Landing_Outcome column value is 'Success (drone ship)'.
- **AND "Payload_Mass__kg_" > 4000**: Adds a condition to filter rows with a Payload_Mass__kg_ value greater than 4000 kg.
- **AND "Payload_Mass__kg_" < 6000**: Adds another condition to filter rows with a Payload_Mass__kg_ value less than 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT "Mission_Outcome", COUNT(*) AS "Total_Count" FROM SPACEXTBL WHERE "Mission_Outcome" IN ('Success', 'Failure') GROUP BY "Mission_Outcome";

- **SELECT "Mission_Outcome", COUNT(*) AS "Total_Count"**: Selects the Mission_Outcome column and calculates the count of each unique value in that column. It uses the COUNT(*) function to count the occurrences.
- **FROM SPACEXTBL**: Specifies the table from which the data will be retrieved.
- **GROUP BY "Mission_Outcome"**: Groups the result set by the Mission_Outcome column, so the count is calculated for each unique mission outcome.

Boosters Carried Maximum Payload

- %sql SELECT "Booster_Version", MAX("Payload_Mass_kg") AS "Max_Payload" FROM SPACEXTBL GROUP BY "Booster_Version" ORDER BY "Max_Payload" DESC LIMIT 1;

- **SELECT "Booster_Version", MAX("Payload_Mass_kg") AS "Max_Payload"**: Selects the Booster_Version column and calculates the maximum Payload_Mass_kg. The MAX() function retrieves the maximum payload mass, and it is aliased as "Max_Payload".
- **FROM SPACEXTBL**: Specifies the table from which the data will be retrieved.
- **GROUP BY "Booster_Version"**: Groups the data by Booster_Version to find the maximum payload for each booster.
- **ORDER BY "Max_Payload" DESC**: Orders the result set by Max_Payload in descending order so that the highest payload appears first.
- **LIMIT 1**: Limits the output to only one row, which represents the booster(s) with the maximum payload mass.
-

2015 Launch Records

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

- **SELECT "Landing_Outcome", "Booster_Version", "Launch_Site"**: Specifies the columns to be retrieved in the query, namely Landing_Outcome, Booster_Version, and Launch_Site.
- **FROM SPACEXTBL**: Indicates the table from which the data will be retrieved.
- **WHERE EXTRACT(YEAR FROM "Date") = 2015**: Filters the data to include only records where the year extracted from the Date column is equal to 2015.
- **AND "Landing_Outcome" LIKE '%Failure (drone ship)%'**: Further filters the results to only include records where the Landing_Outcome column contains the phrase "Failure (drone ship)", indicating failed landing outcomes on the drone ship.
-

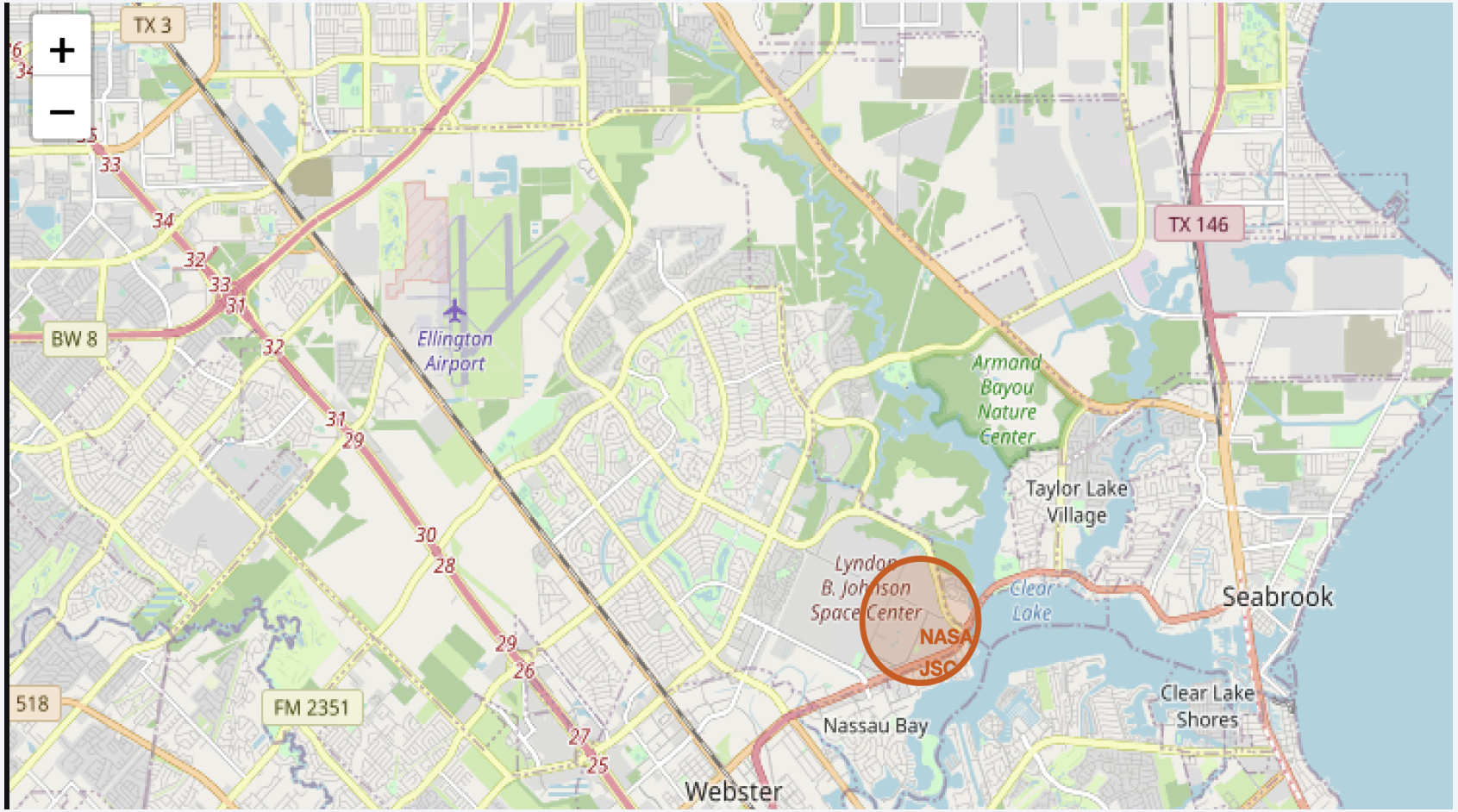
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

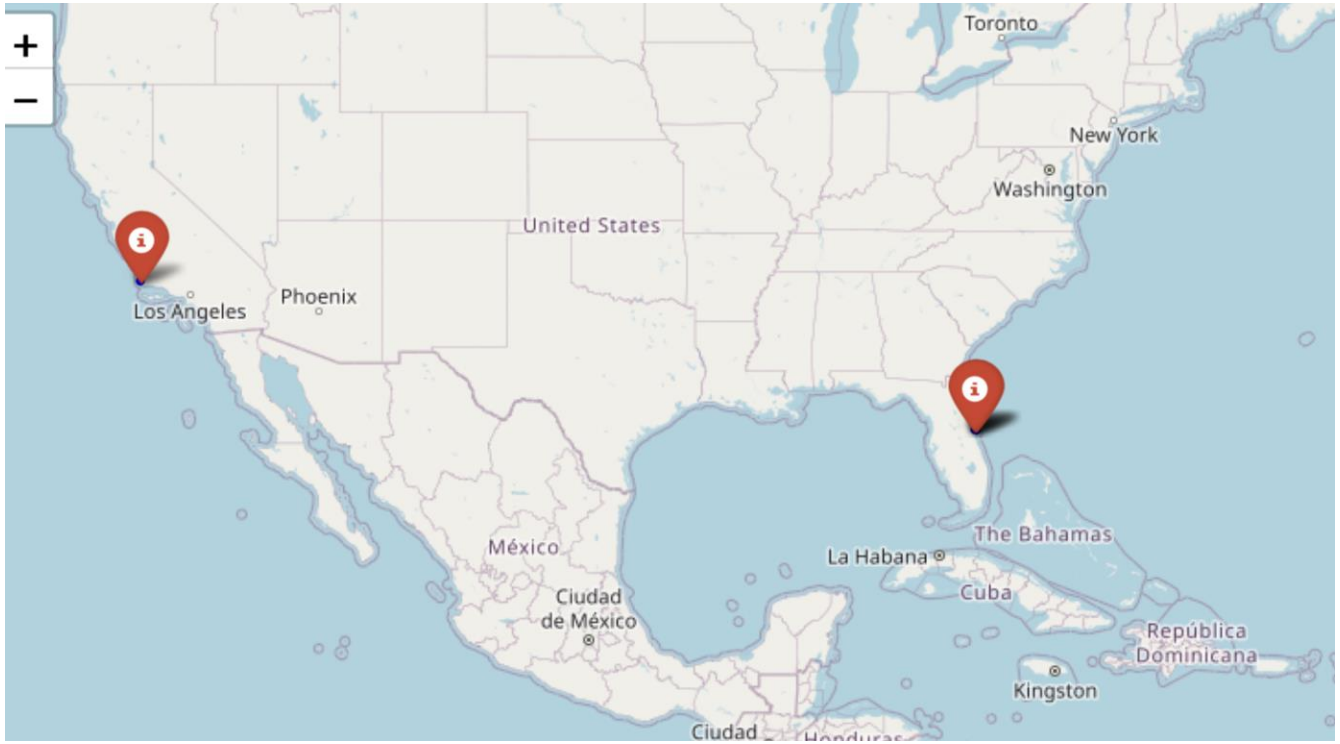
- **SELECT "Landing_Outcome", COUNT(*) AS "Count"**: Specifies the columns to be retrieved in the query, including Landing_Outcome and the count of occurrences (as "Count").
- **FROM SPACEXTBL**: Indicates the table from which the data will be retrieved.
- **WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'**: Filters the data to include only records where the Date falls within the specified date range.
- **GROUP BY "Landing_Outcome"**: Groups the data based on the landing outcomes.
- **ORDER BY "Count" DESC**: Orders the result set in descending order based on the count of landing outcomes.


Section 3
Launch Sites
Proximities Analysis

Nasa location

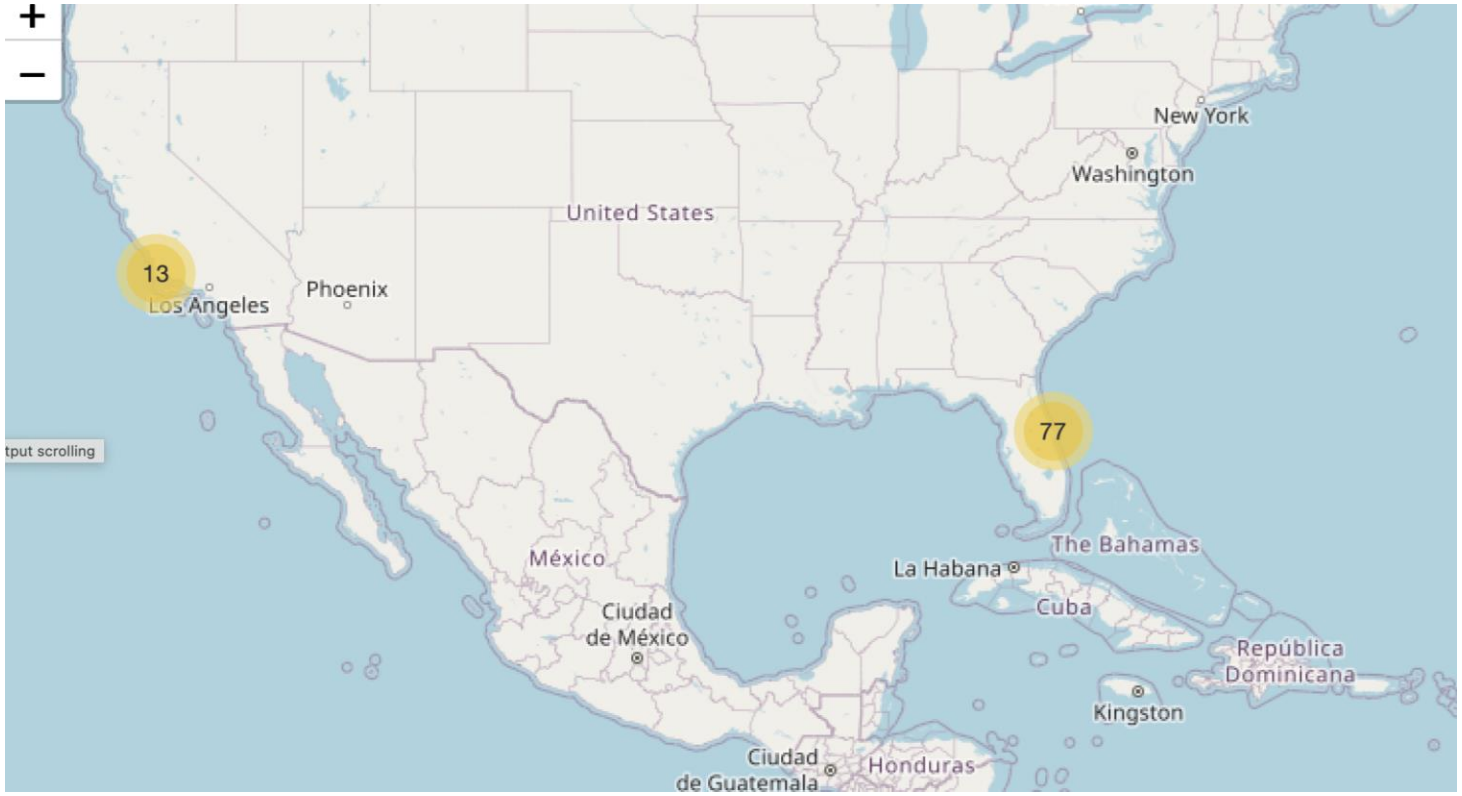


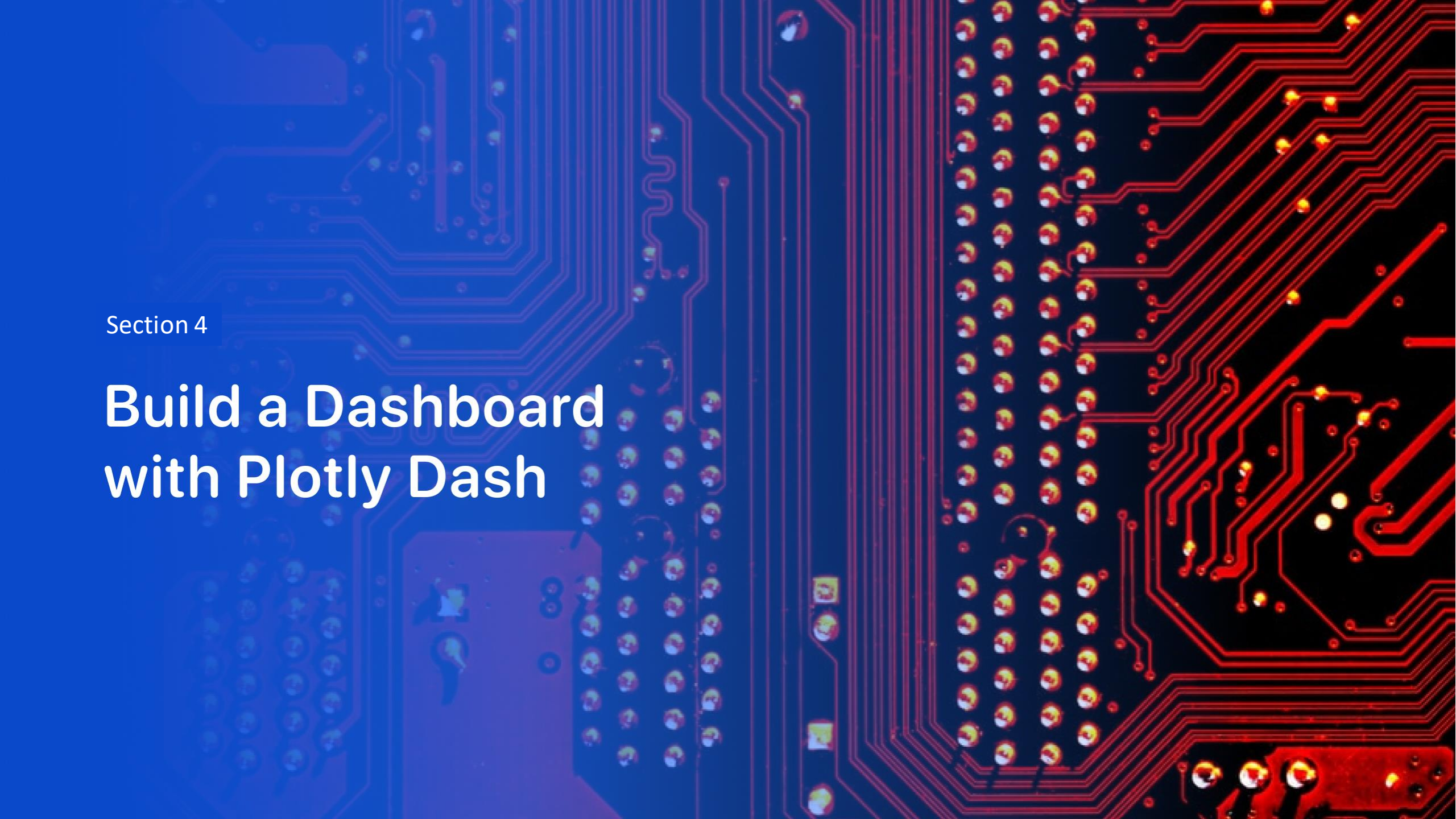
Launch site locations



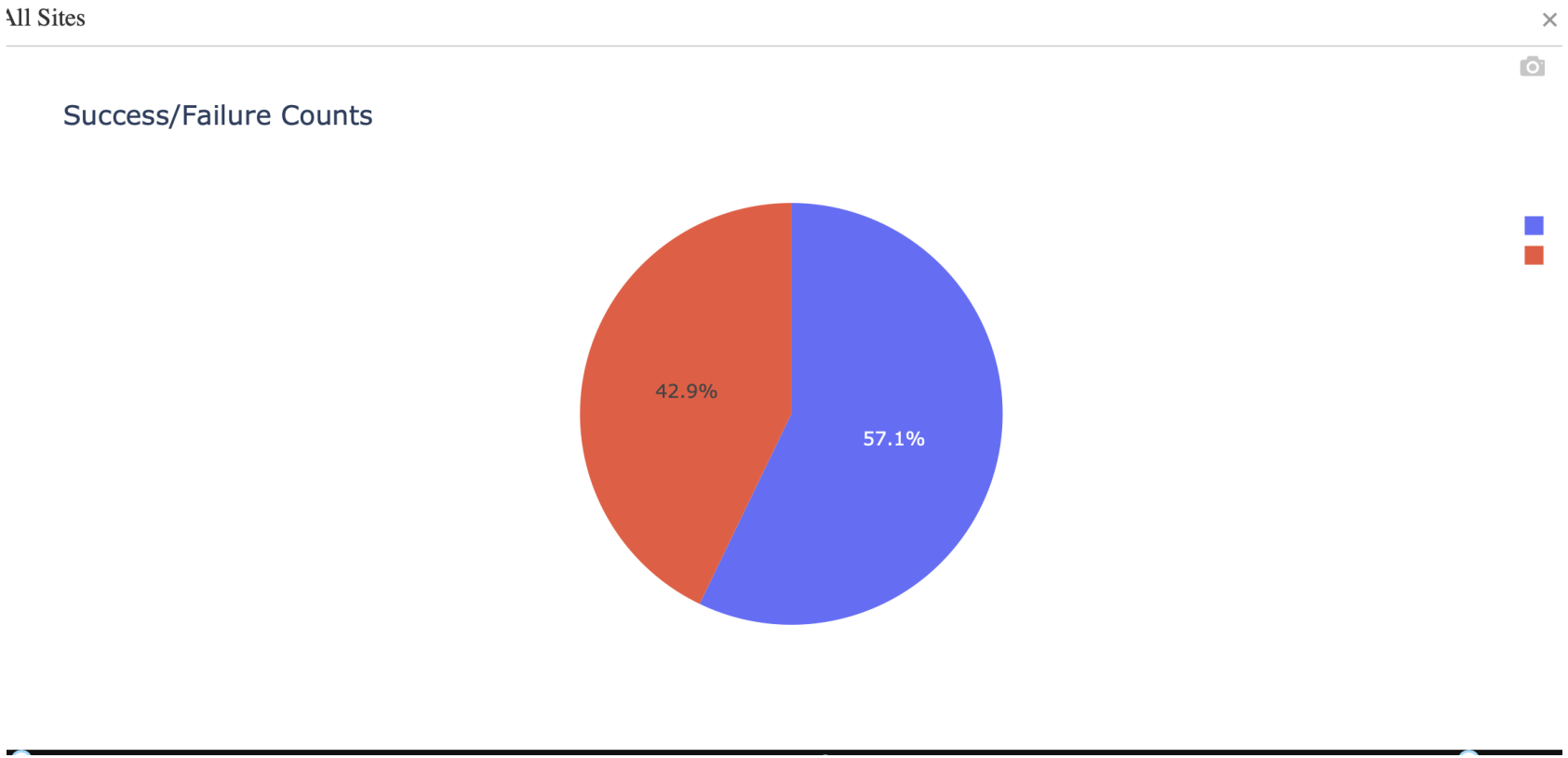
This showcases the distribution of the launchsites near the banks of the oceans

Successful Launches per location

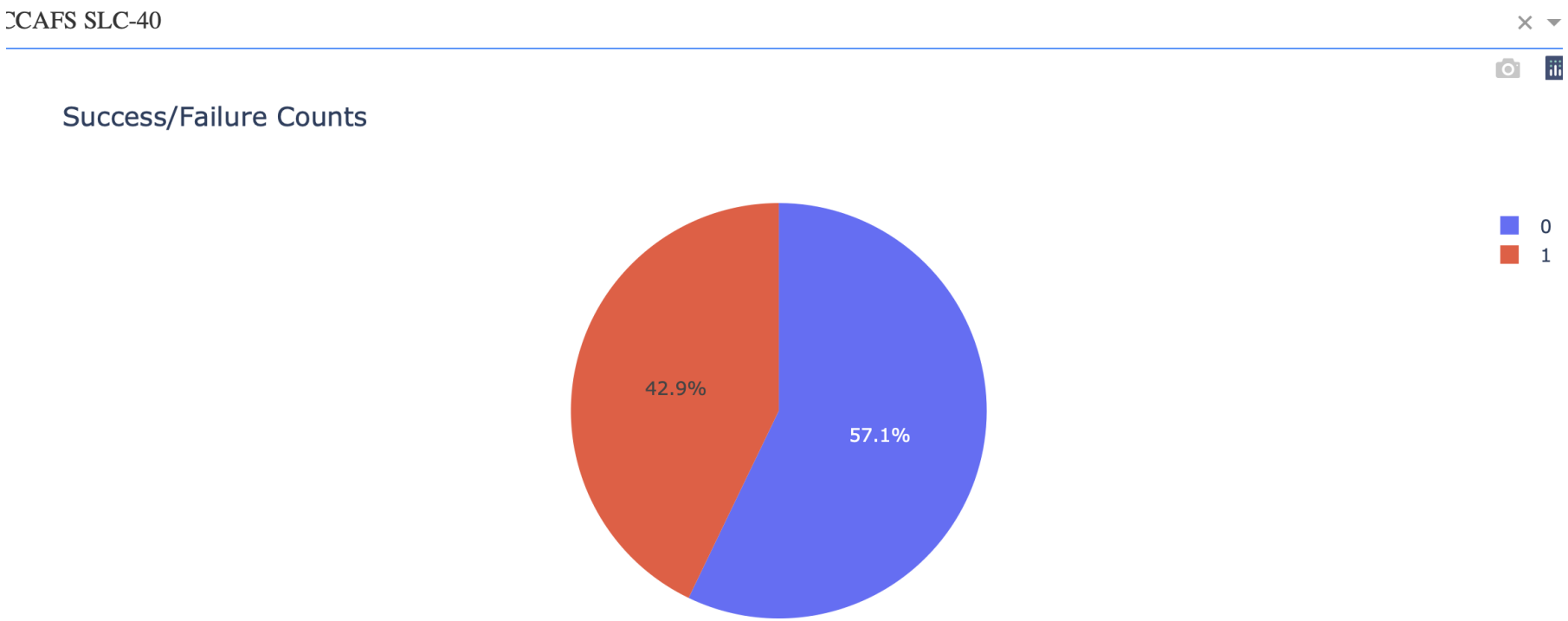



Section 4
Build a Dashboard with Plotly Dash

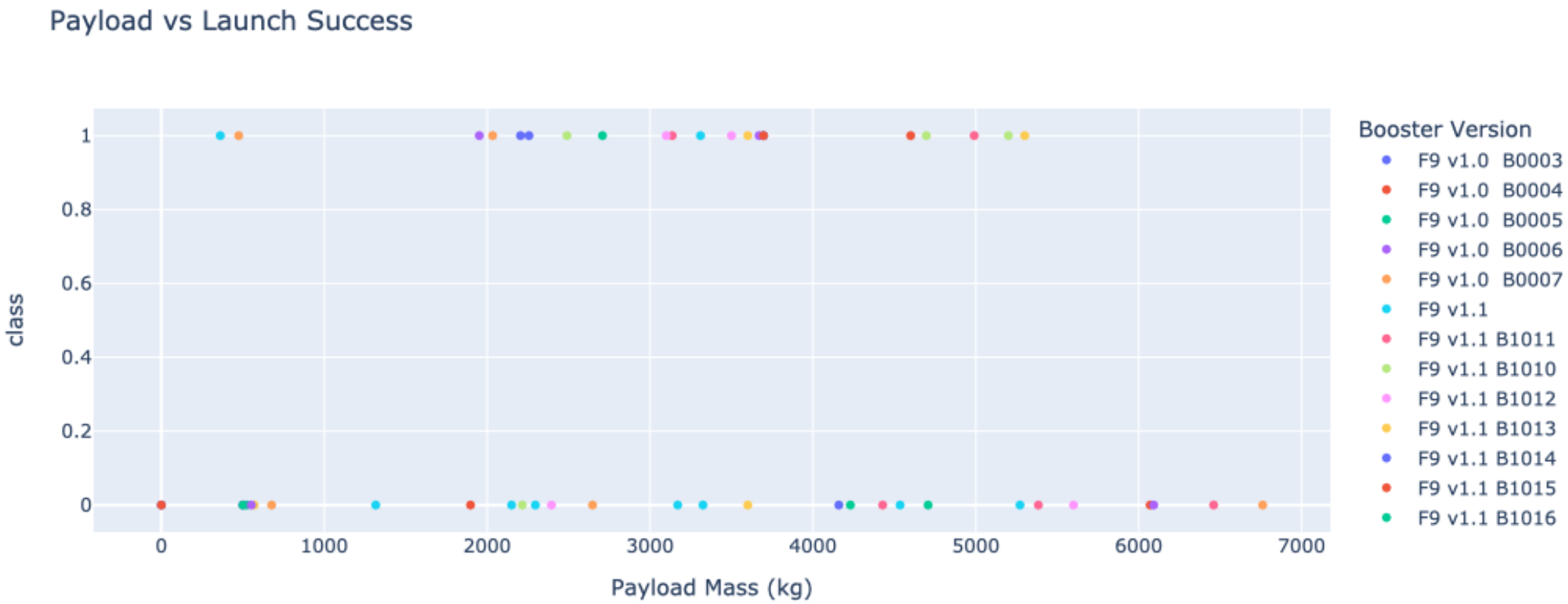
Successful counts in all sites



Most successful launches



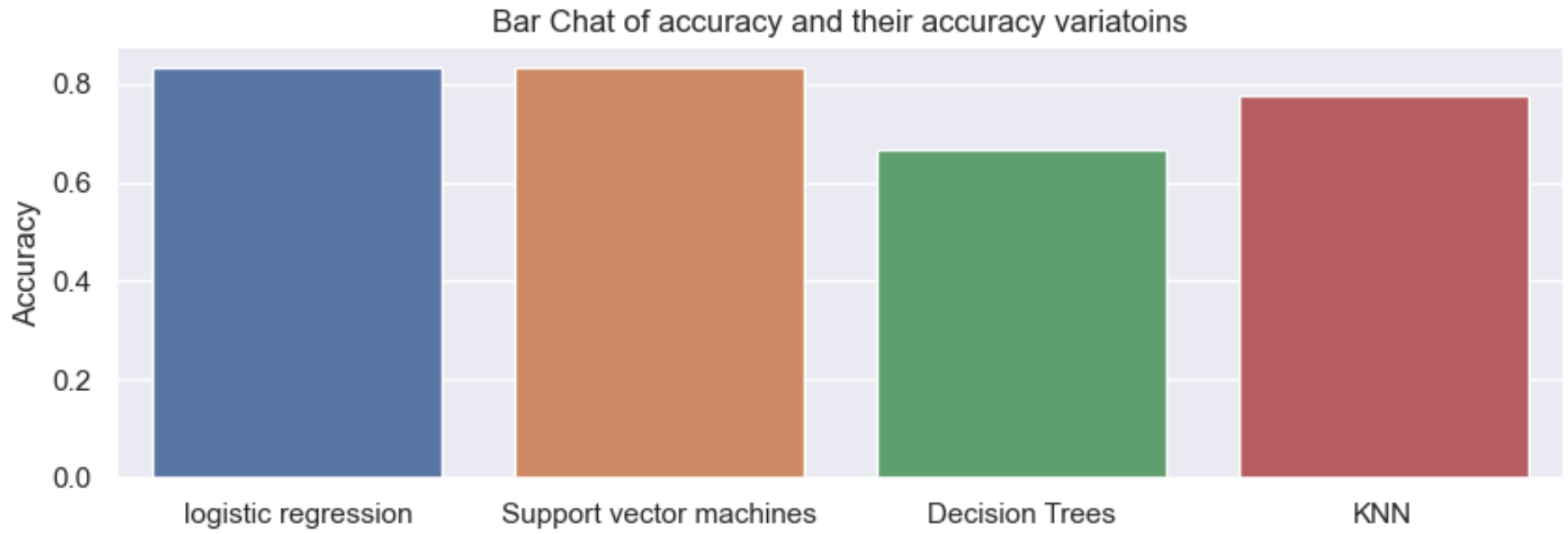
Payload vs. Launch Outcome scatter plot for all sites



Section 5

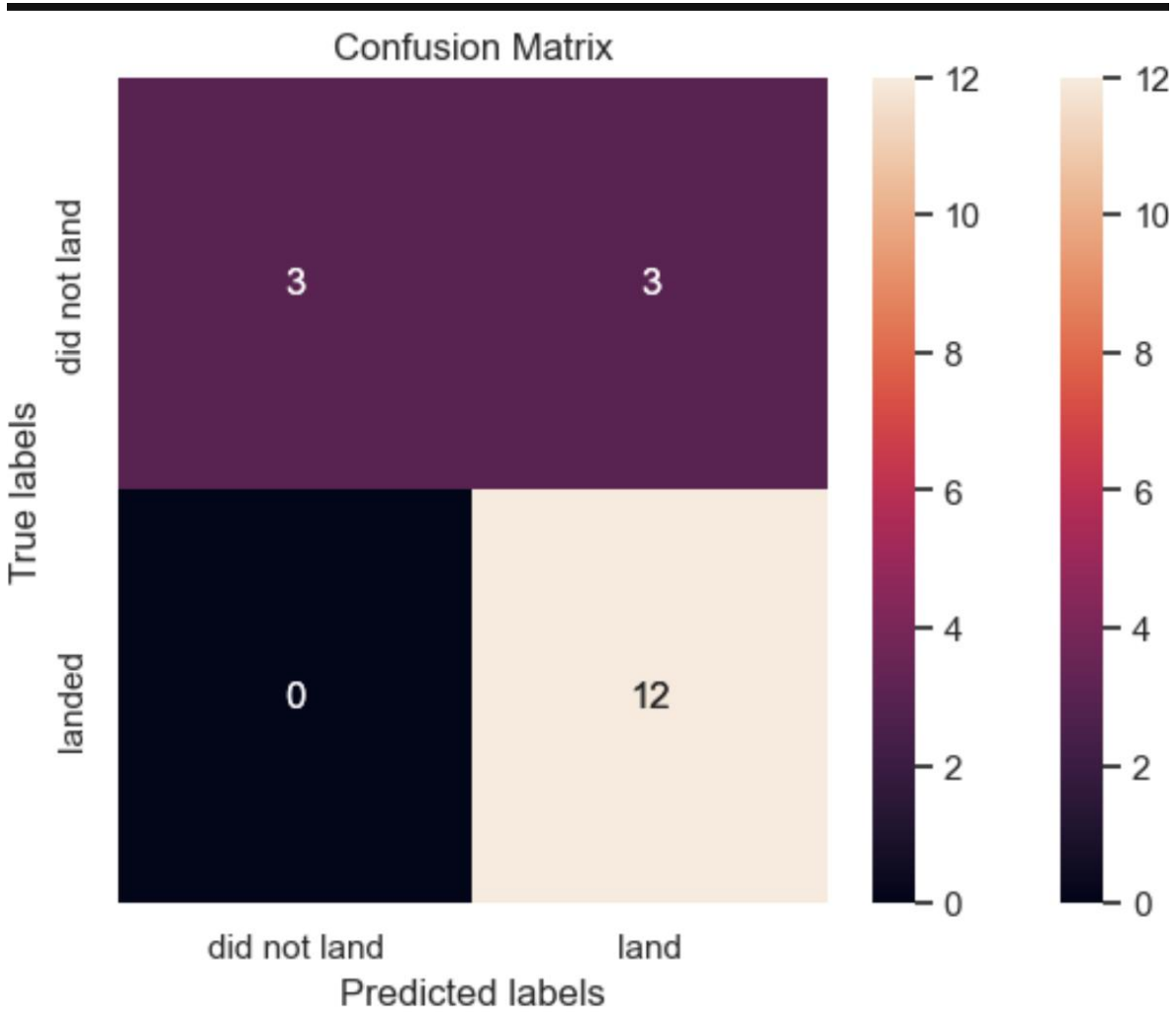
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix

- From the plot, SVM predicted 12 successful landings out of 12 that actually landed and out of the ones that dint land, 3 were wrongly predicted and the other 3 were correctly predicted

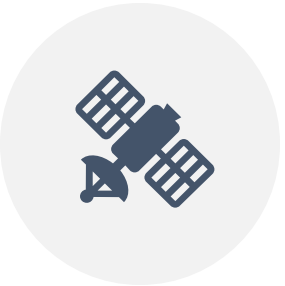


Conclusions

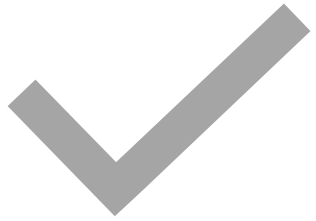
Prediction Objective: The project aimed to predict the successful landing of Falcon 9 rocket first stages.

Model Performance: Logistic Regression and Support Vector Machines achieved the highest accuracy at 83.33%, followed by K-Nearest Neighbors at 77.78%. Decision Trees resulted in an accuracy of 66.67%. These models were developed using hyperparameter tuning for optimal performance.

Cost Efficiency and Industry Relevance: Accurate predictions of Falcon 9 landing success are crucial for estimating mission reliability and associated costs.Potential implications for alternate companies bidding against SpaceX in the rocket launch industry.



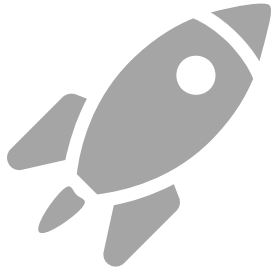
Insights from Model Analysis: Exploratory Data Analysis revealed correlations between flight numbers, payload mass, orbits, and launch success.Features like payload mass, launch site, and orbit were found to be significant predictors of landing success.



Model Limitations and Future Considerations:

The models achieved good accuracy but may not encompass all influential factors.

Future improvements could involve collecting additional data on weather conditions, technical specifics, or environmental factors during launches.



Value of Predictive Models:

Demonstrated the potential of machine learning in estimating rocket landing success, aiding decision-making in the aerospace industry.

Potential contributions to cost estimation and decision-making processes in commercial space launches.

Thank you!