IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

WILMAR CRISTANCHO TRIANA
OCTOBER 30 2024

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- The project focuses on collecting, cleaning, and analyzing data to extract insights, build predictive models, and present findings in a compelling story.
- Identify patterns, predict trends, and uncover insights

Section 1

Methodology

Methodology

Executive Summary

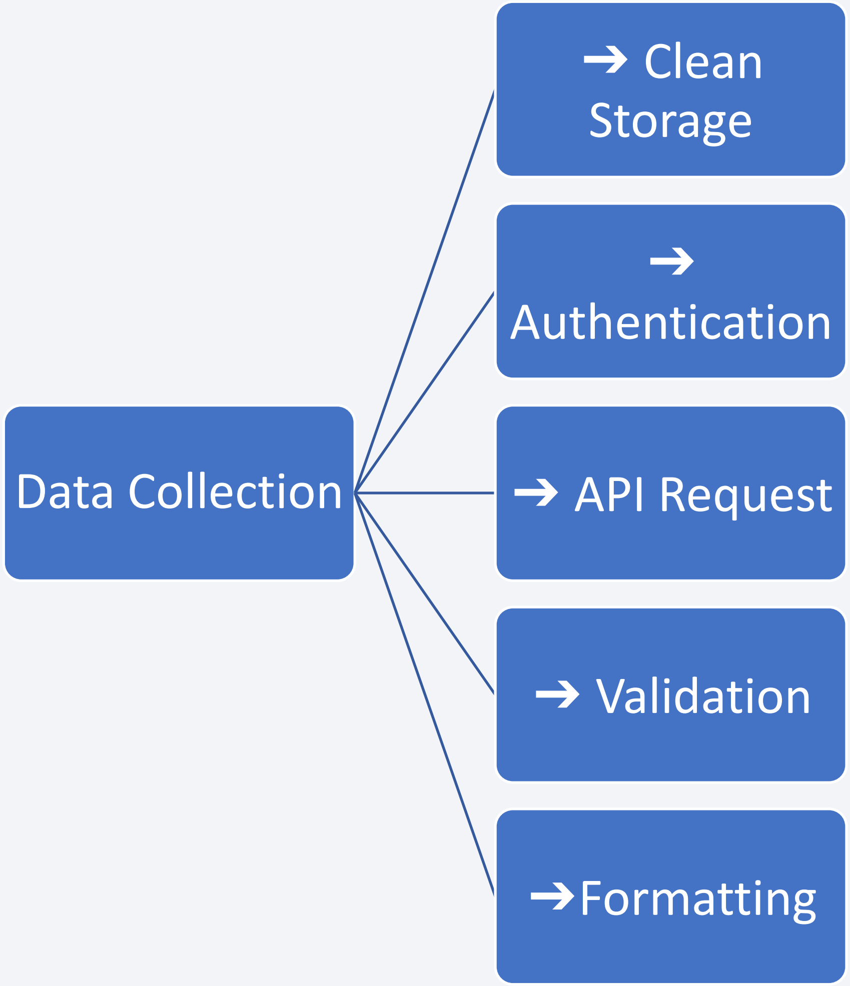
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using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

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

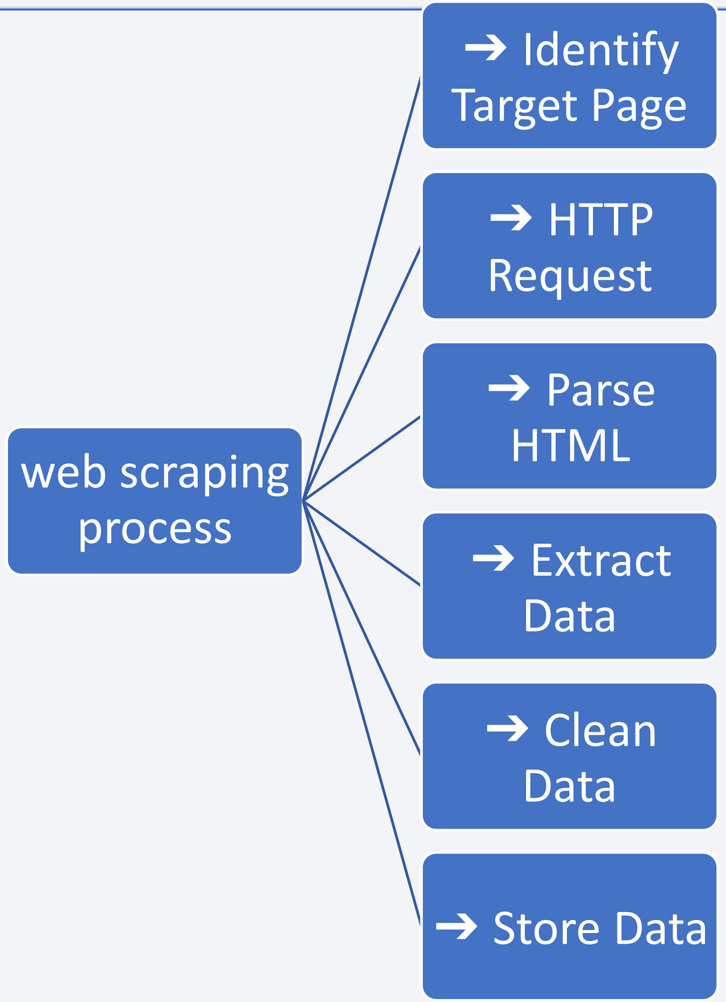
Data Collection – SpaceX API

- Start ➔ Identify SpaceX APIAuthentication (if needed) ➔
- Establish ConnectionMake API Call ➔ Request Specific Data (e.g., launch status, landings)
- Validate and Format Data ➔ Handle Missing Values and StructureStore in
- Clean Format (CSV, DB) ➔ End
- [GitHub link Data Collection API]()



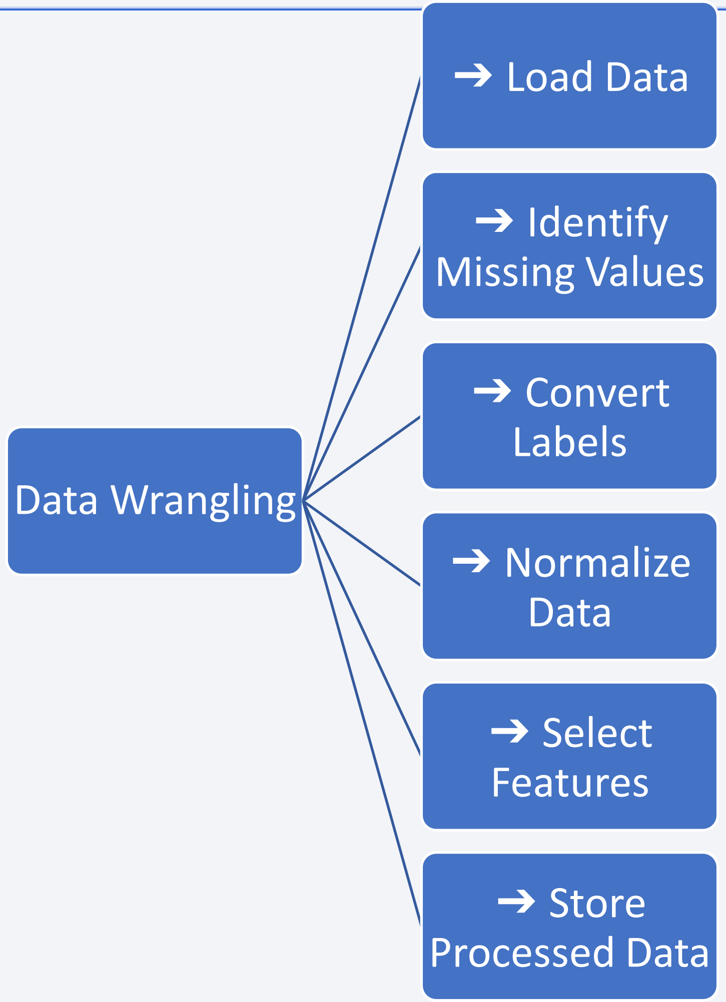
Data Collection - Scraping

- Identify Target Page: Wikipedia page for Falcon 9 & Falcon Heavy launches
- Send HTTP Request: Retrieve page content with requests library
- Parse HTML: Use BeautifulSoup to extract structured data
- Extract Launch Data: Locate tables and pull relevant launch records
- Data Cleaning: Remove extra formatting, handle missing values
- Store Data: Save cleaned data to CSV or database
- [Github link Data Collection Scraping]



Data Wrangling

- Load Data: Import dataset for analysis
- Identify Missing Values: Detect and handle any missing data
- Convert Labels: Map landing outcomes to binary values (1 = successful, 0 = unsuccessful)
- Normalize Data: Scale numeric values if needed for model readiness
- Feature Selection: Choose relevant columns for analysis
- Save Processed Data: Store the cleaned and labeled data
- You need to present your data wrangling process using key phrases and flowcharts
- [Link GitHub Data Wranging]



EDA with Data Visualization

- Bar Charts: Displayed frequencies of successful vs. unsuccessful landings, helping visualize overall landing success rates.
- Histograms: Used to analyze the distribution of payload masses across launches, revealing any patterns or ranges linked to successful landings.
- Scatter Plots: Explored relationships between payload mass and landing success, as well as orbit types, identifying trends in success likelihood.
- Box Plots: Compared payload mass and orbit type across success outcomes, highlighting any significant statistical differences.
- [Link GitHub EDA]

EDA with SQL

- Loaded Dataset: Imported SpaceX dataset into a Db2 database table for querying.
- Total Launches: Queried the total number of Falcon 9 launches to establish dataset scope.
- Landing Success Rate: Calculated the proportion of successful first-stage landings to evaluate reliability.
- Payload Analysis: Filtered launches by payload mass and orbit type to identify any success patterns.
- Launch Site Success: Queried launch sites individually to compare landing success rates by location.
- Cost Estimation: Used landing success data to predict the cost effectiveness of launches for potential bidders.
- Link GitHub EDA SQL

Build an Interactive Map with Folium

- Markers: Placed on each launch site to show exact locations.
- Circle Markers: Indicated success or failure of each launch.
- Lines: Measured distances from launch sites to nearby points for proximity analysis.

- Markers: Show launch site locations.
- Circle Markers: Indicate launch success/failure.
- Lines: Assess proximity impact on success rates.

- Link GitHub map Folium

Build a Dashboard with Plotly Dash

- Pie Chart: Shows launch success rates by site.
- Scatter Plot: Visualizes payload vs. launch success, color-coded by booster version.
- Dropdown & Slider: Select site and payload range for dynamic data filtering.

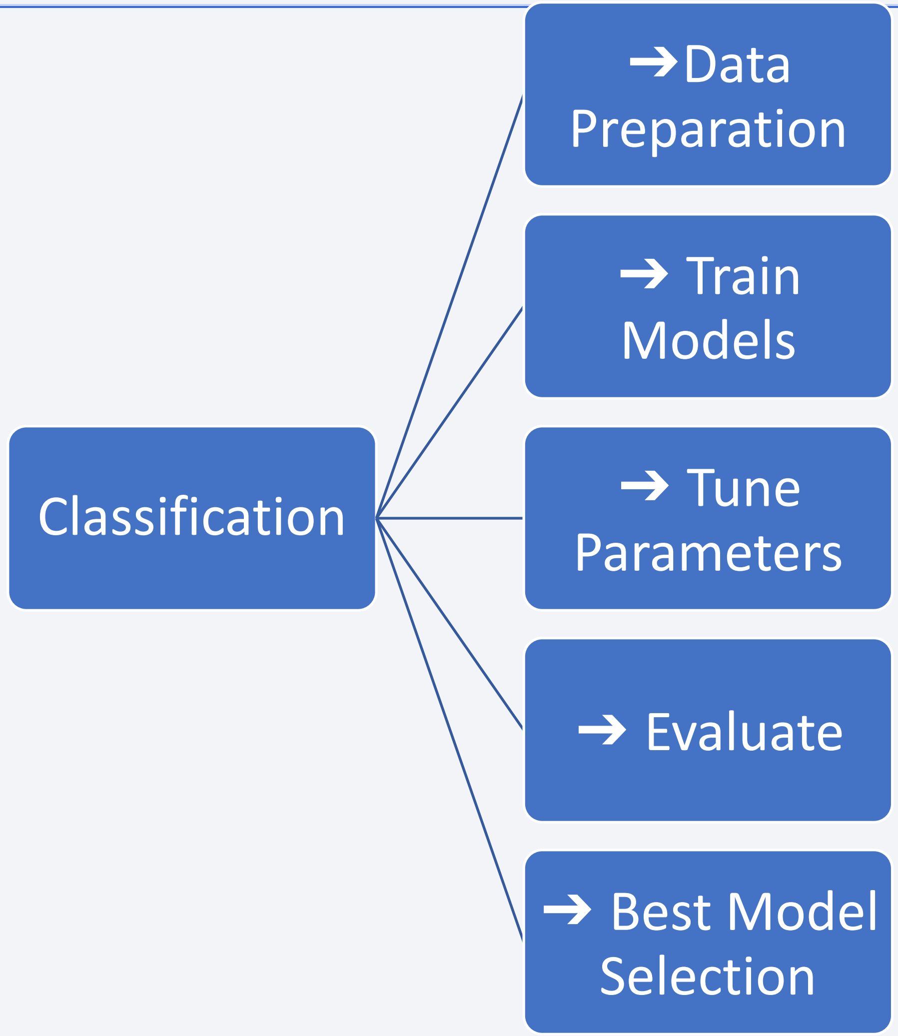
- Pie Chart: To compare launch success by site.
- Scatter Plot: To analyze payload impact on success.
- Dropdown & Slider: For flexible, targeted data exploration.

- Link GitHub

Predictive Analysis (Classification)

- Built Models: Trained SVM, Decision Trees, and Logistic Regression.
- Evaluated: Tested models on test data.
- Improved: Tuned hyperparameters with Grid Search.
- Best Model: Selected based on highest accuracy on test set.

- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose



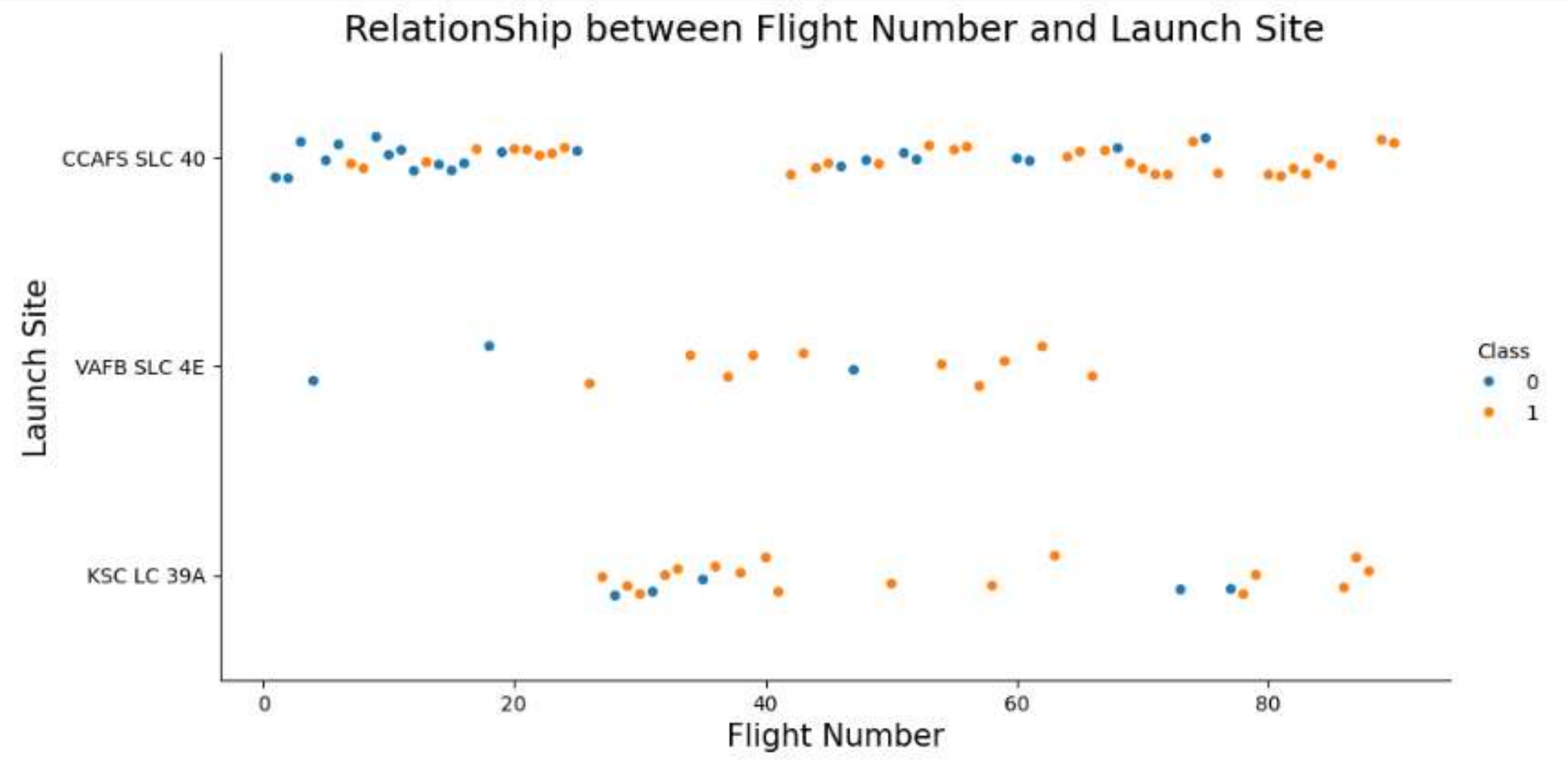
Results

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

Section 2

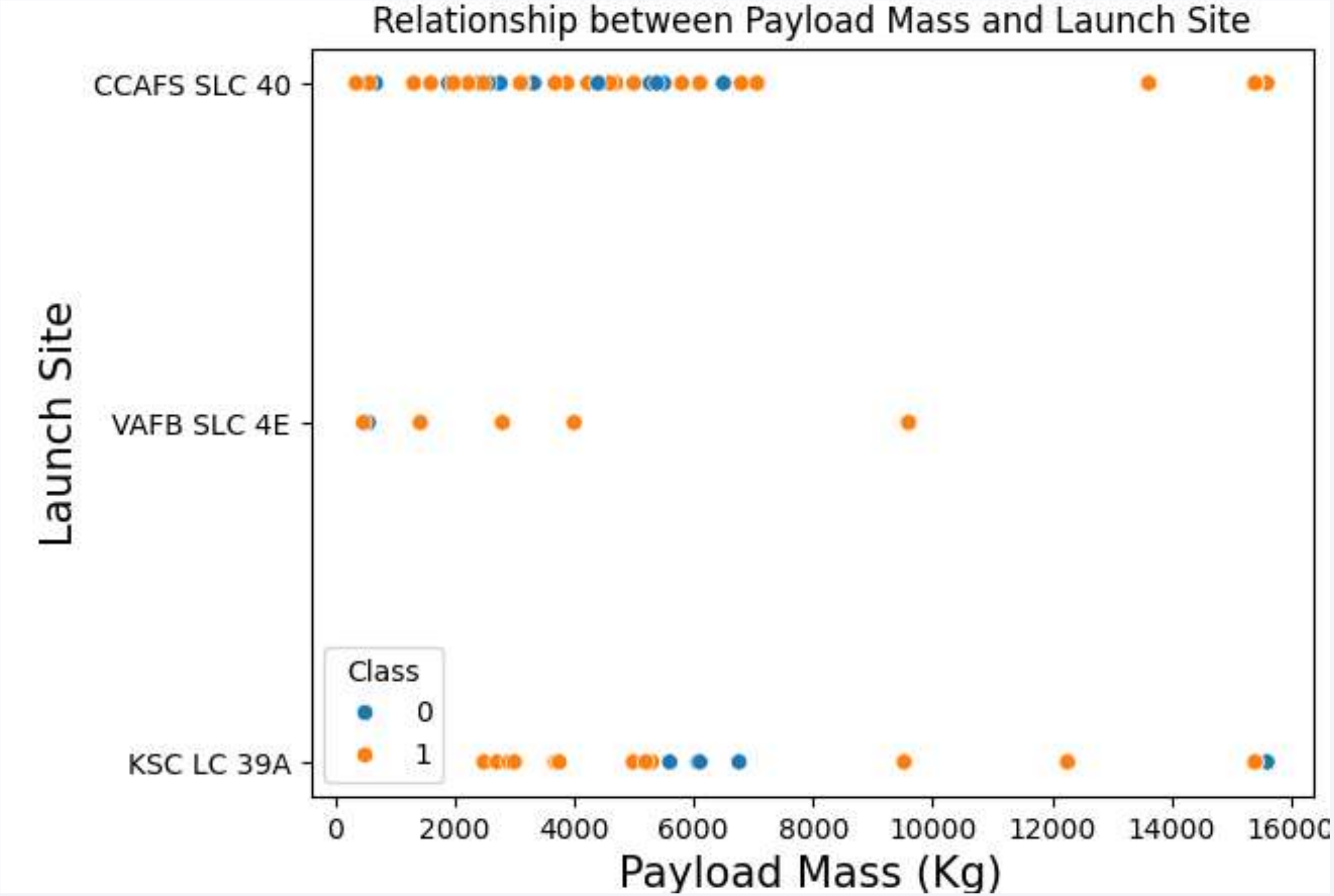
Insights drawn from EDA

Flight Number vs. Launch Site



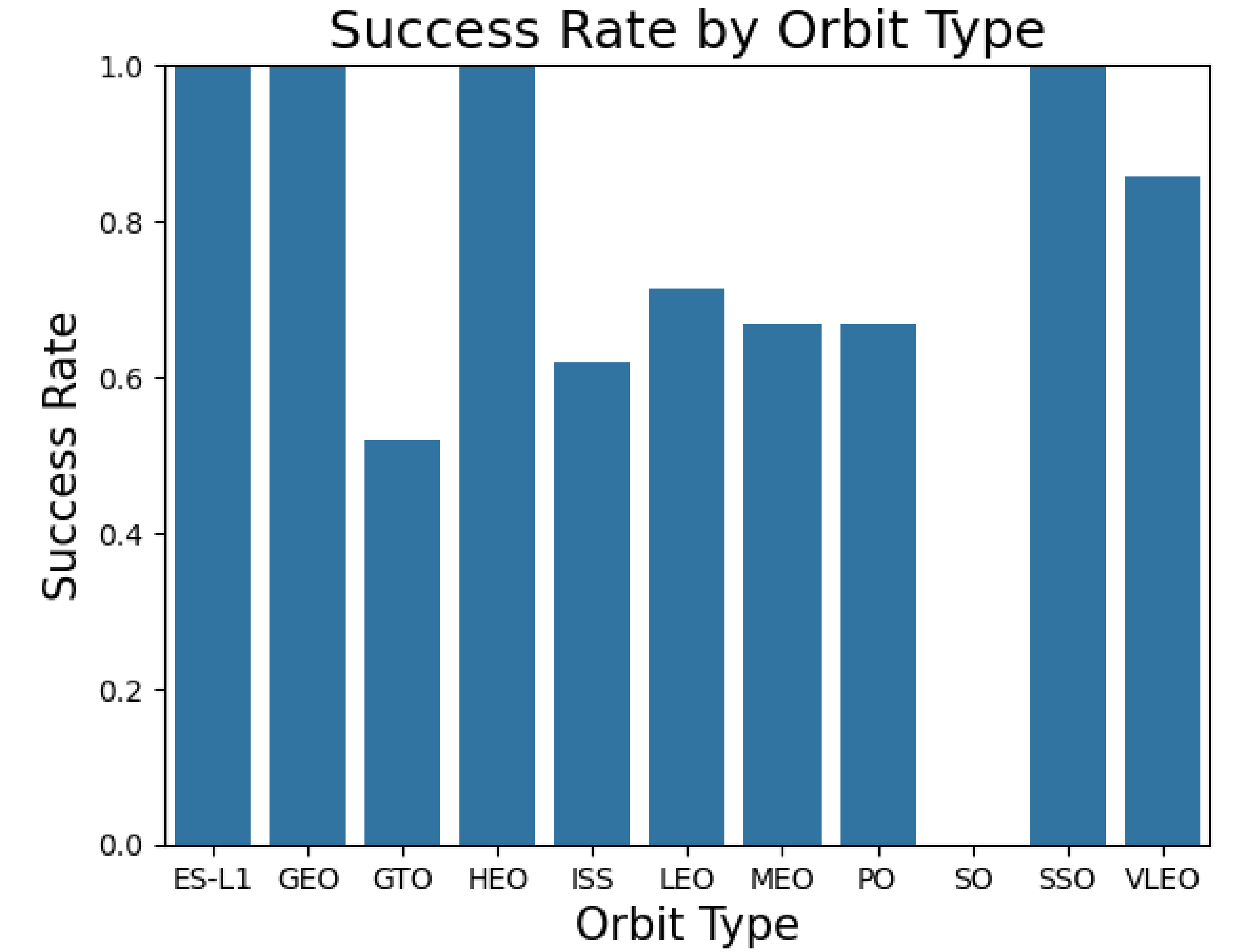
- The plot shows the relationship between flight number and launch site. Each point represents a flight, with color indicating success (orange) or failure (blue). Higher flight numbers generally show more successful landings, especially at CCAFS SLC 40 and KSC LC 39A.

Payload vs. Launch Site



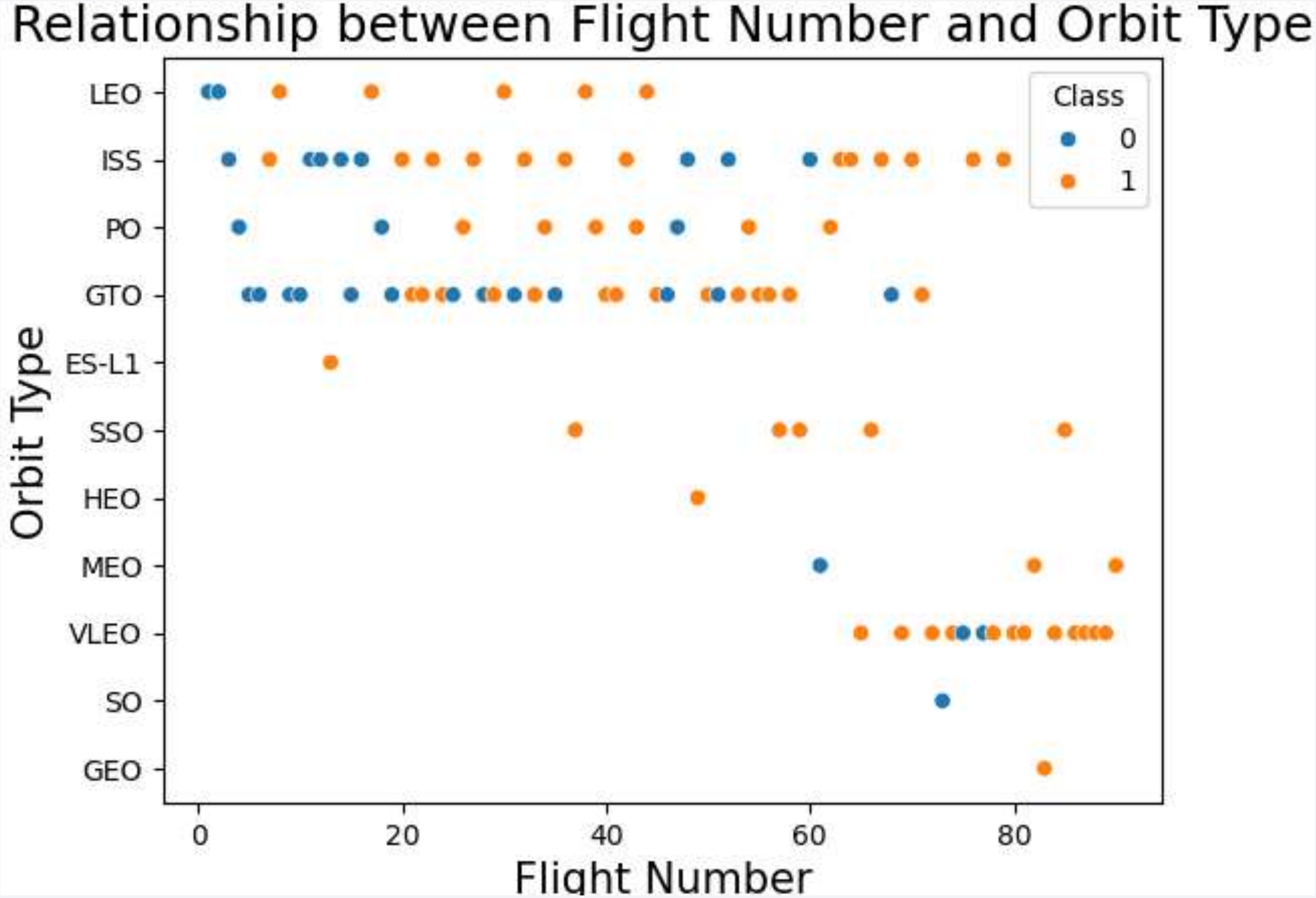
- The plot illustrates the relationship between payload mass and launch site success. Successful landings (orange) are more frequent at CCAFS SLC 40 across a wide range of payloads, while VAFB SLC 4E and KSC LC 39A have fewer successful landings, especially as payload mass increases.

Success Rate vs. Orbit Type



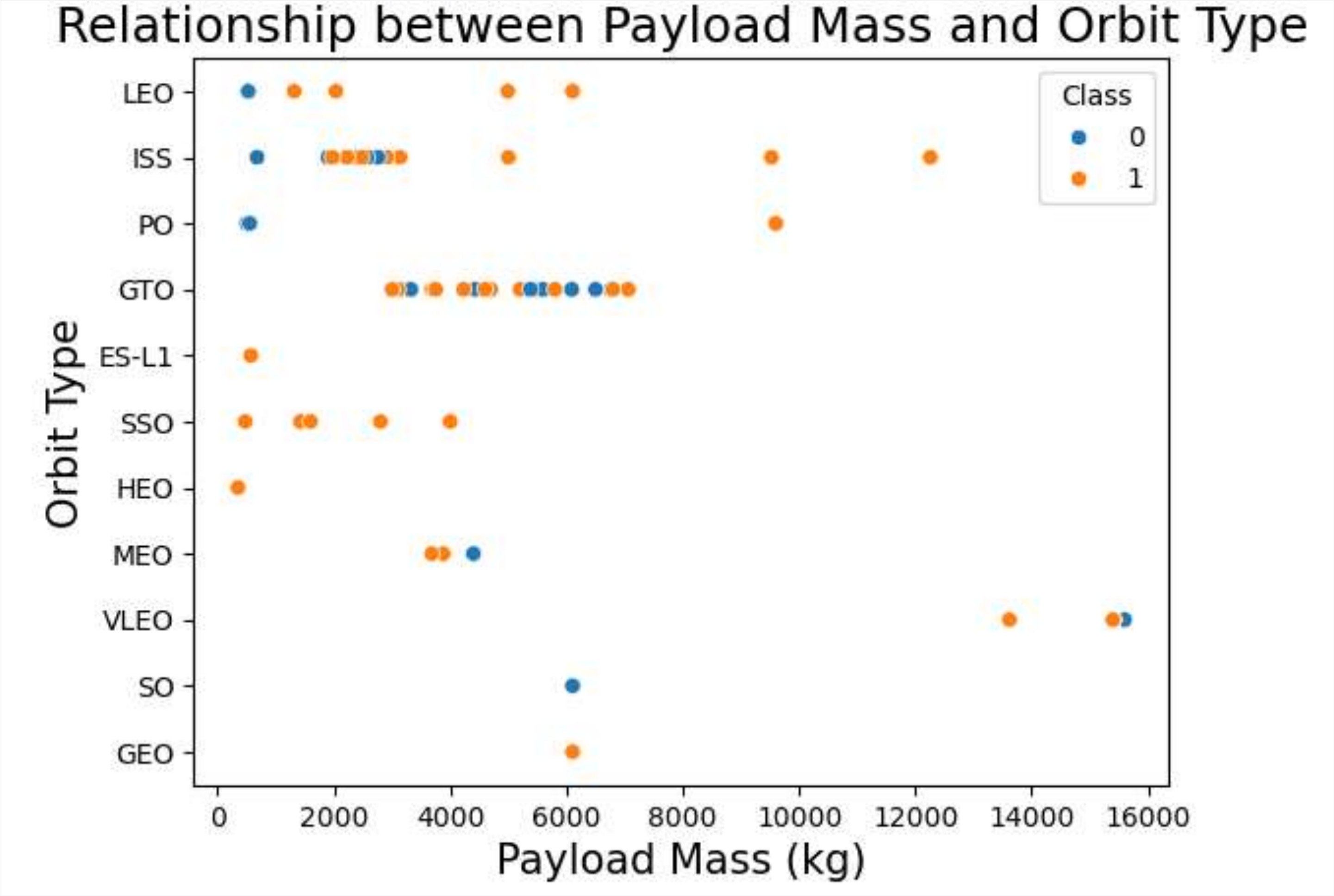
- The plot illustrates the success rate for different orbit types. ES-L1, GEO, SSO, and VLEO orbits show the highest success rates, indicating strong performance in these orbit categories. In contrast, GTO and HEO orbits have noticeably lower success rates, suggesting more challenges associated with these orbit types. Other orbits like LEO and MEO show moderate success rates.

Flight Number vs. Orbit Type



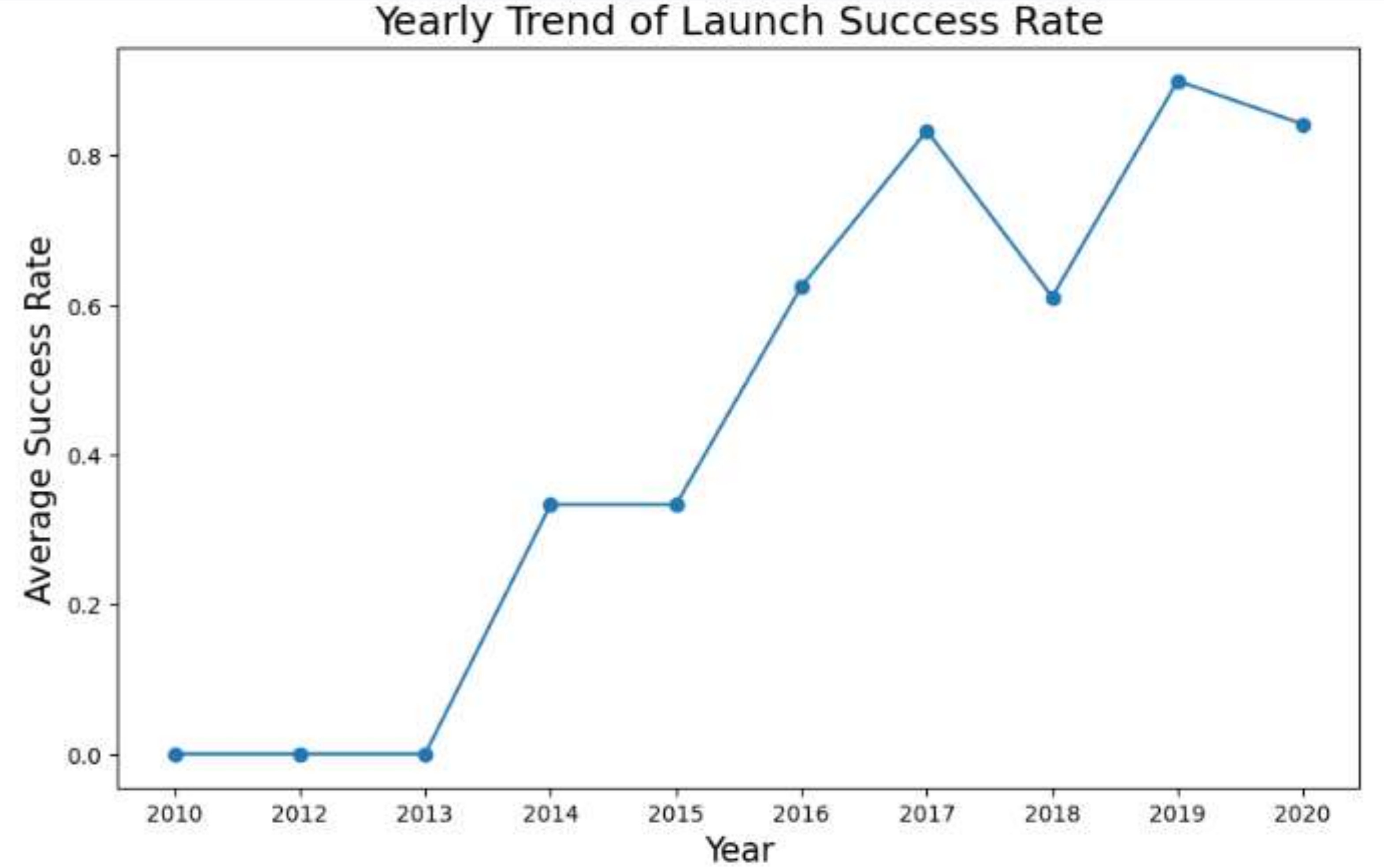
- The plot shows that higher flight numbers generally lead to more successful landings (orange) across most orbits, especially in LEO, ISS, and SSO. Challenging orbits like GTO and PO display mixed outcomes, indicating variability in landing success.

Payload vs. Orbit Type



- The plot shows the relationship between payload mass and orbit type, with success (orange) and failure (blue) outcomes. Successful landings are observed across various payloads for most orbits, especially in LEO and ISS. However, higher payloads, particularly in GTO and VLEO, show mixed success rates, indicating that increased payload mass may impact landing success in these orbits.

Launch Success Yearly Trend



- The plot shows a steady increase in launch success rates over the years, peaking around 2019, indicating improved reliability over time. This trend suggests advancements in technology and procedures at SpaceX.

All Launch Site Names

- The query result shows four unique SpaceX launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40. These represent the primary locations used for Falcon 9 launches, each playing a key role in mission deployments.

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

Launch Site Names Begin with 'CCA'

- This query selects the first 5 records from SPACEXTABLE where the Launch_Site starts with "CCA," filtering for launch sites located at Cape Canaveral.

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

Total Payload Mass

- This query calculates the total payload mass for launches where the customer is NASA (CRS) in the SPACEXTABLE. It provides the combined payload weight for all NASA Commercial Resupply Services (CRS) missions.

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass for launches with the booster version F9 v1.1 in the SPACEXTABLE, providing insight into the typical payload capacity for this specific booster version.

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- This query finds the earliest date for a successful landing on a ground pad in the SPACEXTABLE, identifying when the first successful ground landing occurred.

First_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query retrieves the Booster Version for launches that successfully landed on a drone ship and carried a payload mass between 4000 kg and 6000 kg. It helps identify booster versions capable of handling mid-range payloads with successful drone ship landings.

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

Total Number of Successful and Failure Mission Outcomes

- This query counts the number of occurrences for each Mission Outcome in the SPACEXTABLE, grouping results by outcome type. It provides a summary of how frequently each mission outcome (e.g., success, failure) has occurred.

Mission_Outcome	Outcome_Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- This query retrieves the Booster Version associated with the heaviest payload in the SPACEXTABLE. It identifies which booster was used for the mission carrying the maximum payload mass.

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

- This query selects the Month, Landing Outcome, Booster Version, and Launch Site for launches in 2015 that had a landing outcome of Failure (drone ship). The substr function is used to extract the month and year from the date, helping to analyze failed drone ship landings within that specific year.

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

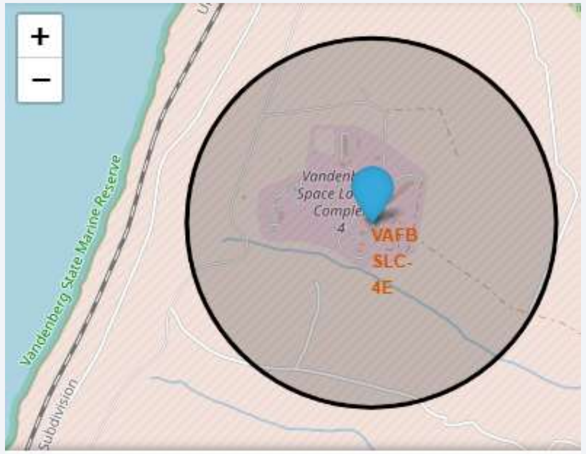
- This query counts the occurrences of each Landing Outcome in the SPACEXTABLE for launches between June 4, 2010 and March 20, 2017. It groups results by landing outcome, sorts them in descending order of frequency, and provides insight into the most common landing outcomes during this period.

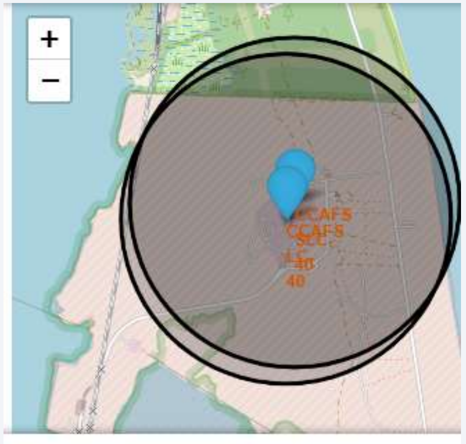
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

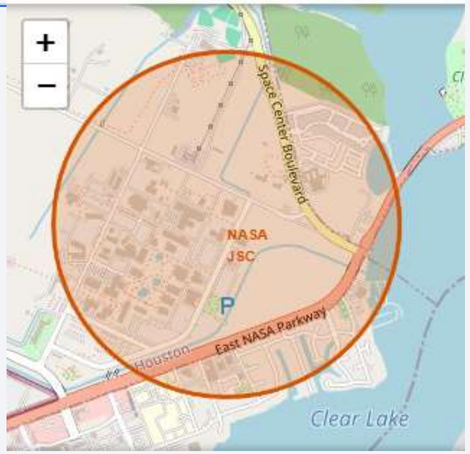
Section 3

Launch Sites Proximities Analysis

Location of Space Launch Sites







Launch Sites: Blue markers show U.S. rocket launch sites, mainly on the east and west coasts.

Labels: Orange labels indicate site names, including VAFB on the west coast, CCAFS in Florida, and NASA JSC in Texas.

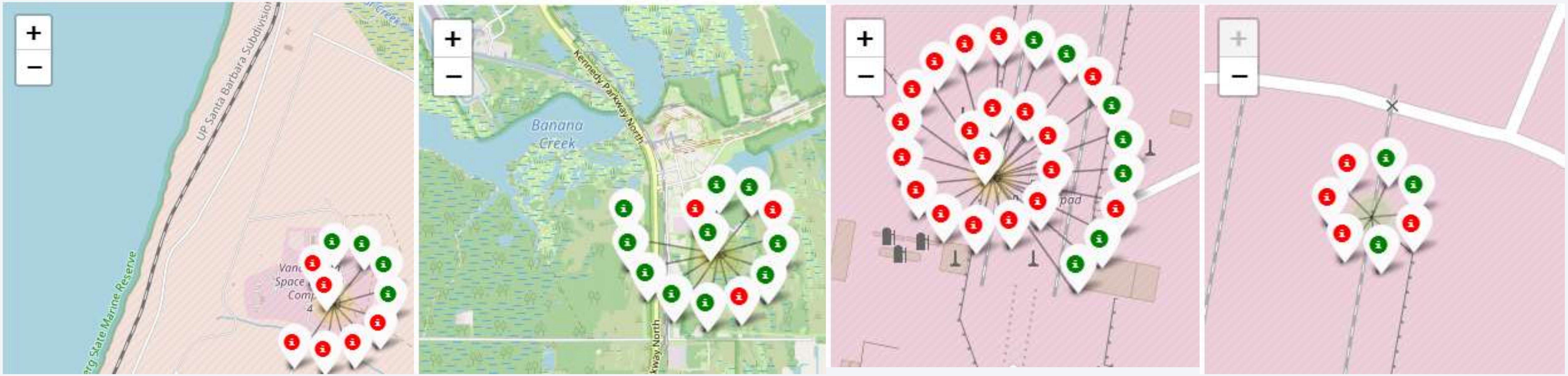
Strategic Locations: Coastal positioning allows safe launches over the ocean and efficient eastward trajectories.

Insights

Coastal Concentration: U.S. launch sites are mostly coastal, optimizing for different orbits.

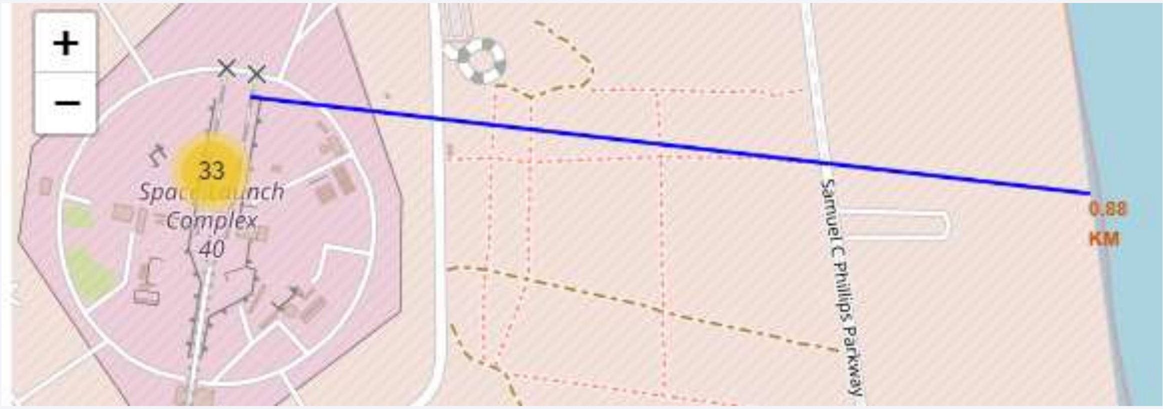
East Coast Hub: Cape Canaveral hosts multiple launch sites, highlighting its role in U.S. space missions.

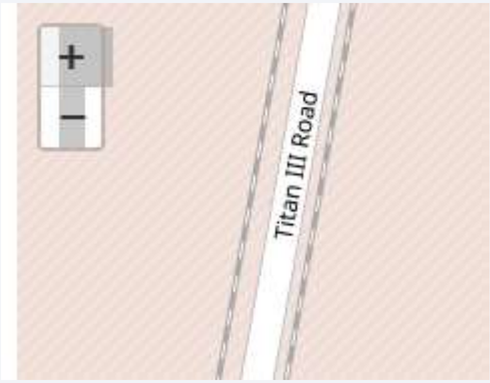
SpaceX Launch Success and Failure Location



- Clustered Markers: Groups markers at busy launch sites; click to expand.
- Color-Coded Outcomes: Green for success, red for failure, showing launch results at a glance.
- Key Sites: Major sites include CCAFS SLC-40 and LC-39A in Florida.

Nearest Launch Site to Coastline with Distance Marker



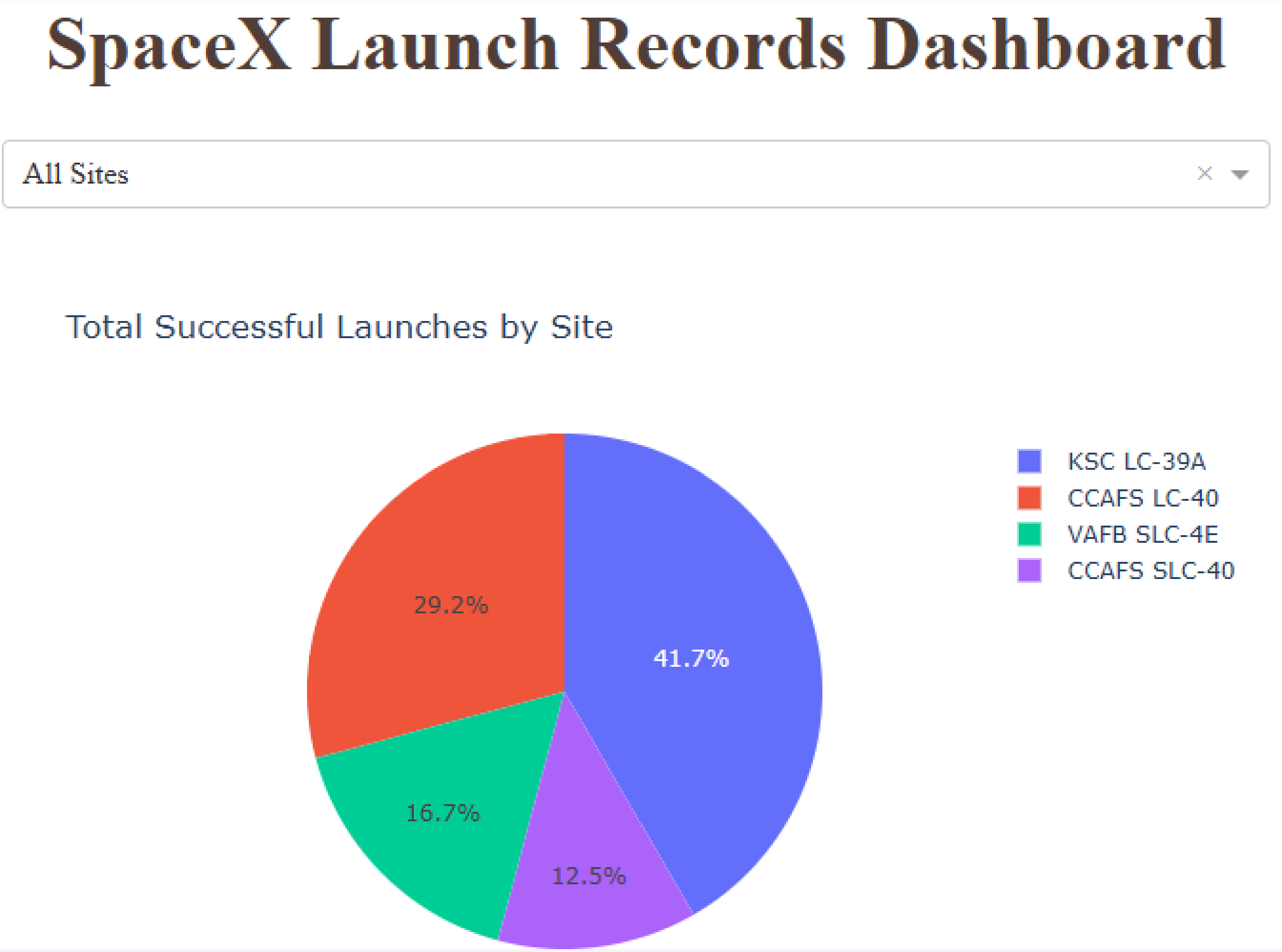


- Launch Site: "Space Launch Complex 40" is marked, indicating a SpaceX launch pad.
- Coastline Distance: A blue line shows the 0.88 km distance from the launch site to the nearest coastline point.

Section 4

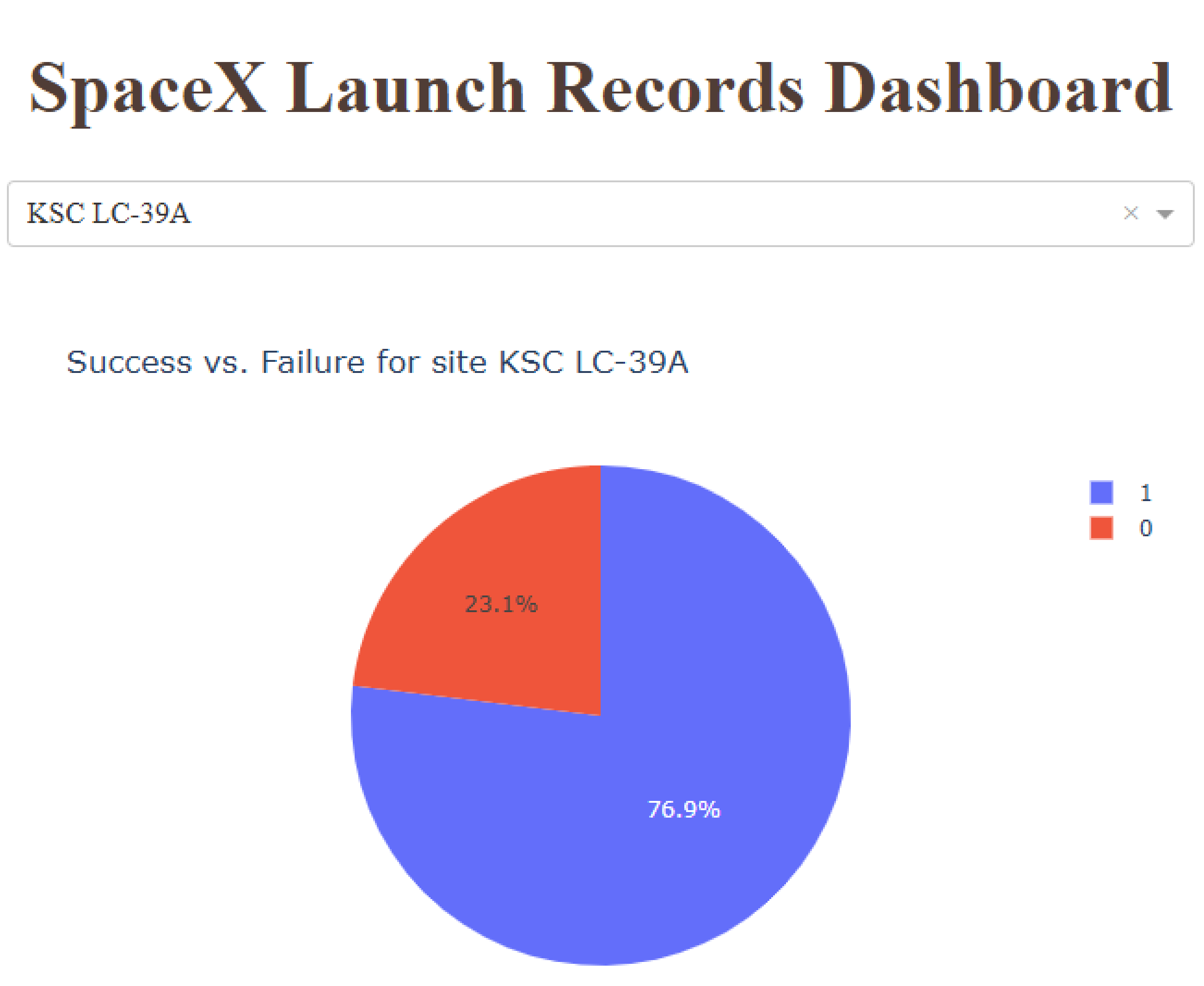
Build a Dashboard with Plotly Dash

Total Successful Launch



- Most Successful Site: KSC LC-39A has the highest success rate, with 41.7% of total successful launches.
- Diversity in Launch Sites: Successes are spread across multiple sites, but some contribute more than others, showing varying levels of activity.

High Success Rate



- High Success Rate: KSC LC-39A has a strong success rate, with nearly 77% of launches being successful.

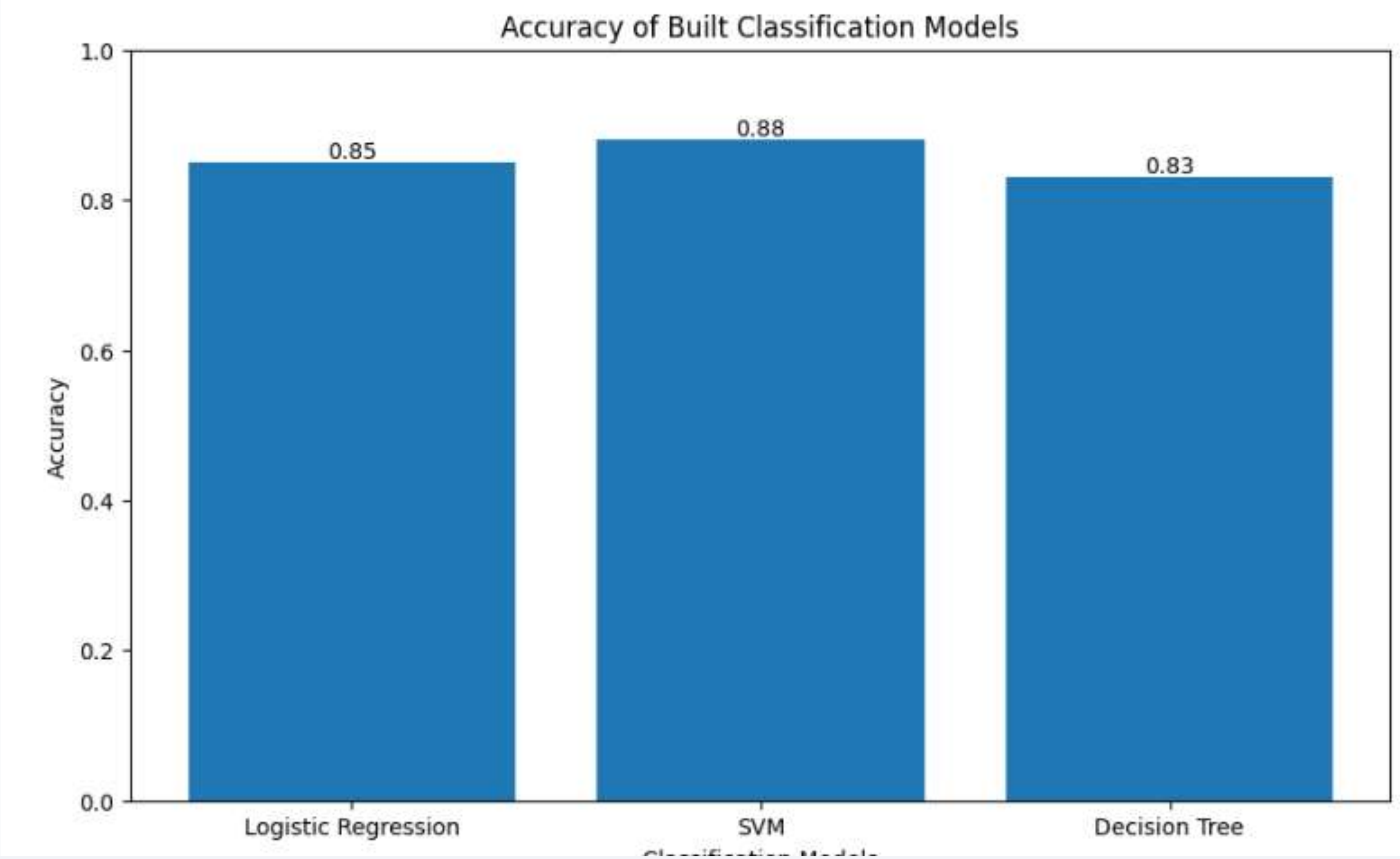
Impact of Payload Mass and Booster Version on Launch Success

- High Success Rates: Success is more common in the lower payload ranges (0–6,000 kg).
- Booster FT Dominance: The FT booster version appears frequently in successful launches, indicating high reliability.

Section 5

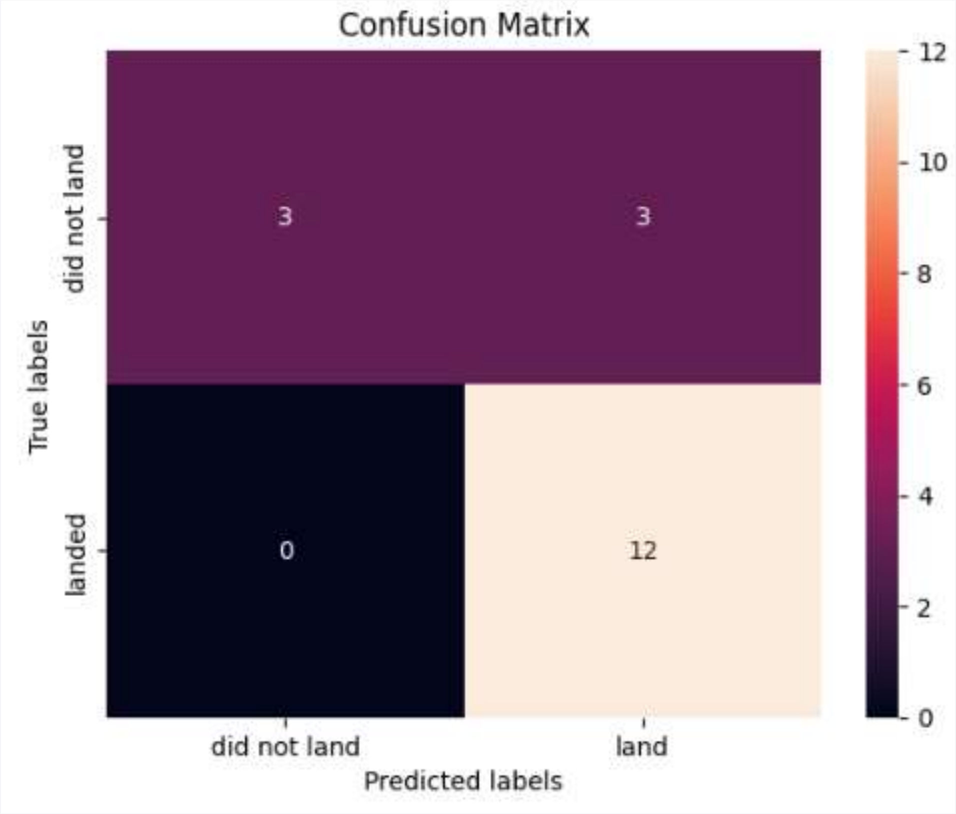
Predictive Analysis (Classification)

Classification Accuracy



- The highest accuracy, achieved by the SVM model at 88%, indicates it best captures patterns in the data. This is likely due to effective hyperparameter tuning and SVM's robustness in handling complex patterns without overfitting.

Confusion Matrix



- True Positives (12): Correctly predicted "landed."
- True Negatives (3): Correctly predicted "did not land."
- False Positives (0): No incorrect "landed" predictions.
- False Negatives (3): Incorrectly predicted "did not land."
- The model accurately identifies most landings, with minimal errors

Conclusions

- Point 1: The SVM model achieved the highest accuracy, making it the most reliable for this classification task.
- Point 2: The confusion matrix shows strong performance in predicting successful landings, with few misclassifications.
- Point 3: Hyperparameter tuning contributed to the model's effectiveness, optimizing its parameters for accuracy.
- Point 4: The model's high accuracy can support cost-effective decisions for future rocket launches by accurately predicting reusable landings.

Appendix

- Python Code: Key data processing and analysis snippets
- SQL Queries: Essential queries for data insights
- Charts: Visuals showing key findings
- Notebook Outputs: Significant results
- Datasets: Original and processed data for reference

Thank you!