IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joachim Lelle
April 2024

Outline

Executive Summary

- METHODOLOGIES: Implemented a multi-stage process encompassing data collection via API calls and web scraping, data cleansing and transformation, exploratory data analysis with advanced visualizations, and predictive modeling using machine learning algorithms.

- RESULTS: The predictive analysis yielded a reliable model with high accuracy in forecasting successful landings, which could significantly enhance launch cost predictions and competitive bidding strategies.

Introduction

- **Project Background and Context:** In an era of commercial space travel, efficient launch operations are critical. This project delves into SpaceX's launch data to understand patterns and outcomes, focusing on the pivotal role of the Falcon 9 rocket's first-stage landing success, which is key for reducing costs and increasing launch frequency.

- **Problems to Find Answers To:** The core challenge is to predict the success of Falcon 9 first-stage landings. By doing so, we aim to ascertain the financial viability of reuse in rocket launches and provide actionable insights for potential competitors in the space launch market.

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

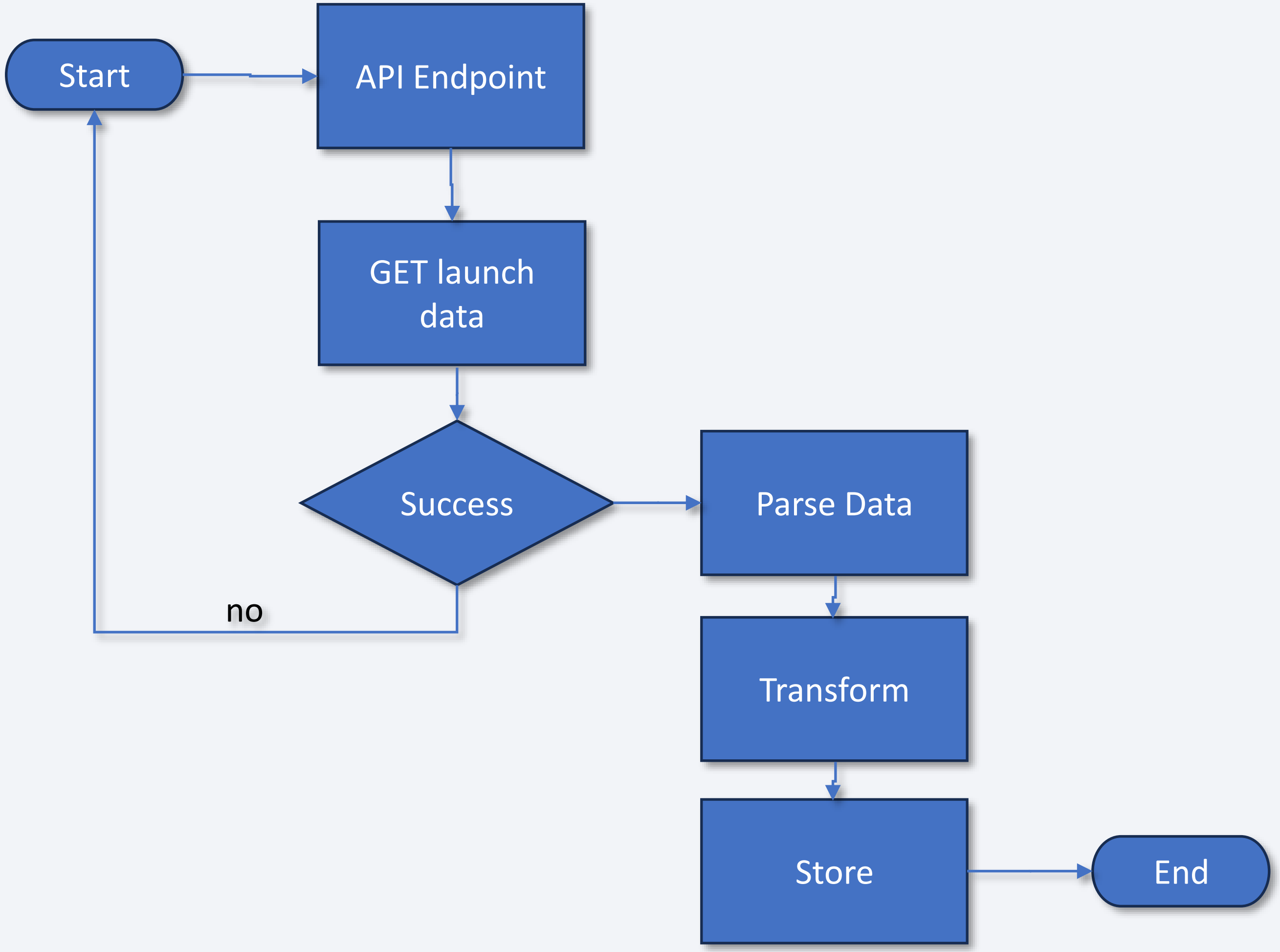
Data Collection

- How data sets were collected:
  - Data sets were meticulously compiled via SpaceX API, which involved making GET requests to retrieve launch data.
  - Essential details like booster version, payload mass, launch site, and landing outcomes were extracted and formatted into a structured data frame.
  - Auxiliary functions were employed to parse the data further, ensuring meaningful insights could be derived from the specific characteristics of each launch.
- The data collection process was iterative and thorough, leveraging the power of API calls to shape raw data into an analytical foundation .

- Identify Data Source
- API Calls / Scraping
- Authentication
- Data Retrieval
- Data Parsing
- Data Cleaning
- Data Storage
- Wrangling & Transform.
- Data Analysis

Data Collection – SpaceX API

- Automated Data Access
- Real-Time Updates
- Consistency and Standardization



GitHub URL of the completed SpaceX API calls notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab01-1.spacex-data-collection-api_(completed).ipynb

Data Collection - Scraping

- Extraction of Launch Records
- Efficiency and Accuracy
- Structured Data Retrieval

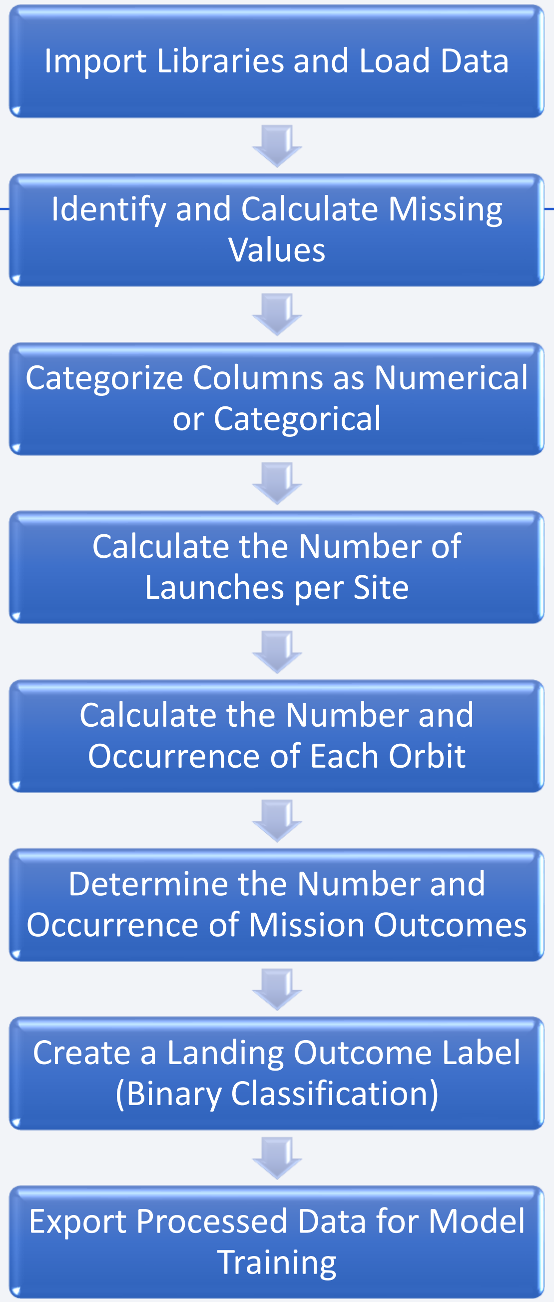
1. Load the necessary libraries (BeautifulSoup for parsing HTML and requests for making HTTP requests).
2. Request the webpage content using the requests library.
3. Parse the webpage content using BeautifulSoup to create a soup object.
4. Use BeautifulSoup methods like find, find_all to extract relevant data (tags, attributes, text) from the soup object.
5. Store the extracted data in a structured format (like a pandas DataFrame).
6. Clean and preprocess the data as needed (handling missing values, type conversion, etc.).
7. Export the cleaned and structured data for further analysis or use.

GitHub URL of the completed web scraping notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab01-2.webscraping_(completed).ipynb

Data Wrangling

- Prrocessed and transformed SpaceX Falcon 9 launch data to prepare it for a binary classification model predicting the success of first stage landings.

GitHub URL of the completed data wrangling related notebooks: https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab02-01.Spacex-Data%20wrangling_(completed).ipynb



EDA with Data Visualization

- KEY METHODOLOGIES:
- Utilizing Python libraries such as Pandas for data manipulation, and Matplotlib and Seaborn for data visualization.
- Exploring relationships between flight number, payload mass, launch site, orbit type, and their effects on launch success.
- Implementing visualizations like scatter plots, bar charts, and line charts to identify trends and patterns in the data.

GitHub URL of the completed EDA with Data Visualization notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab04-01.eda-dataviz_(completed).ipynb

- FINDINGS AND CHARTS PLOTTED:
- A positive correlation between flight number and success rate, suggesting improvements in landing success over time.
- The effect of payload mass on landing success, with heavier payloads showing different success rates across launch sites.
- The success rate varying by orbit type, with some orbits showing higher success rates than others.
- Yearly trends indicating an overall increase in launch success rates over time.

EDA with SQL

- **Key methodologies:**
- Utilized SQLite database to store and query SpaceX launch data.
- Employed SQL queries to extract specific information related to launch sites, payload masses, orbit types, and landing outcomes.
- Analyzed data distributions, correlations, and aggregations to understand the factors influencing launch success.

GitHub URL of the completed EDA with SQL notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab03-01.eda-sql-coursera_sqllite_(completed).ipynb

- **Findings and Charts Plotted:**
- Launch Sites Analysis: number of launches from sites to understand their activity levels.
- Success Rate by Orbit Type: for different orbit types, revealing how orbit designations influence mission success.
- Payload Mass Impact: relationship between payload mass and launch success, suggesting the potential impact of payload mass on mission outcomes.
- Yearly Launch Trends: counts and success rates, identifying growth and improvement in launch operations over time.

Build an Interactive Map with Folium

- OBJECTIVE: The lab aimed to showcase the use of Folium for interactive map visualization, focusing on visualizing the geographical distribution of SpaceX launch sites and analyzing their success rates.

GitHub URL of the completed Map with Folium notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab05-01.launch_site_location_(completed).ipynb

These elements were added to provide an intuitive and interactive way to explore SpaceX launch sites' geographical contexts.

- Markers give a clear point of reference for each site representing a location of a SpaceX launch site.
- Circles emphasize the area around the launch sites.
- Lines measure distances crucial for logistical considerations and environmental impact assessments

Build a Dashboard with Plotly Dash

Plots/Graphs and Interactions:

- Dropdown list
- Range slider
- Pie chart
- Scatter plot

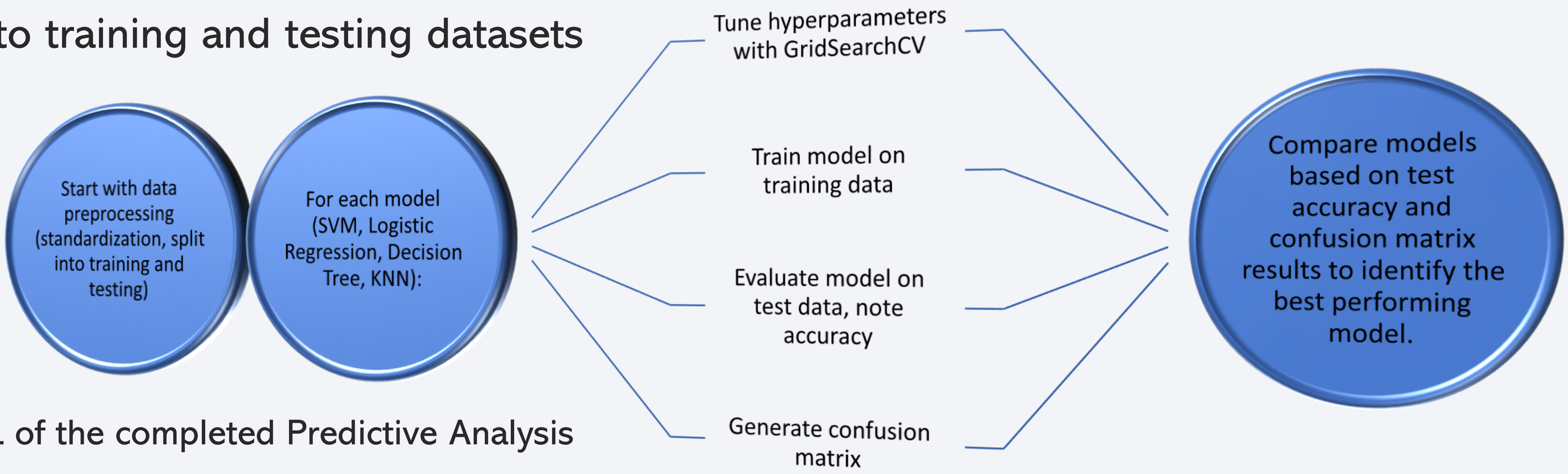
GitHub URL of the completed Dashboard with Plotly and Dash notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/spacex_dash_app.py

Reasons for Adding:

- Dropdown list allows users to filter launch data by specific launch sites to analyze site-specific launch success rates.
- Range slider lets users explore the impact of payload mass on launch success, enabling a detailed analysis within selected payload ranges.
- Pie chart provides a visual summary of launch success rates, offering a quick overview of successful versus failed launches for a chosen site or across all sites.
- Scatter plot reveals patterns or correlations between payload mass and launch outcomes, with booster version differentiation to further explore the influence of technology on success rates.

Predictive Analysis (Classification)

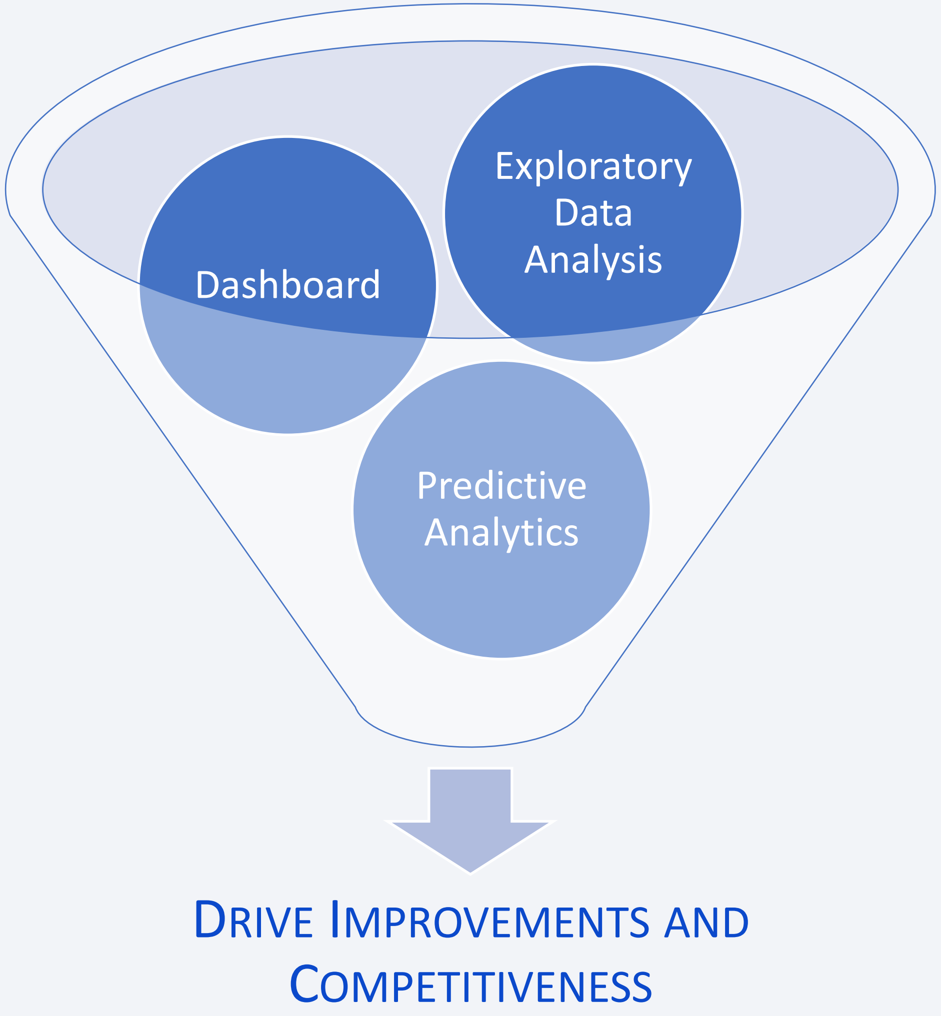
- Exploratory Data Analysis and Determining Training Labels
- Data Standardization
- Split into training and testing datasets
- Hyperparameter Optimization
- Model Evaluation



GitHub URL of the completed Predictive Analysis notebook:
https://github.com/Joe-L1964/ds24-rep/blob/main/DSCP.Lab06-01.SpaceX_Machine_Learning_Prediction_(completed).ipynb

Results

- The launch site has a significant impact on the success rate of landings. Certain sites demonstrate higher success rates, potentially due to geographical advantages or better infrastructure.
- Orbit Types and Success Rate: Different orbits have varied success rates for the first stage landing.
- A relationship exists between the payload mass and the success rate of landings. Lighter payloads have shown a higher success rate, suggesting the limitations of the Falcon 9 first stage under heavier load conditions.



Qualitative Conclusions and Suggestions

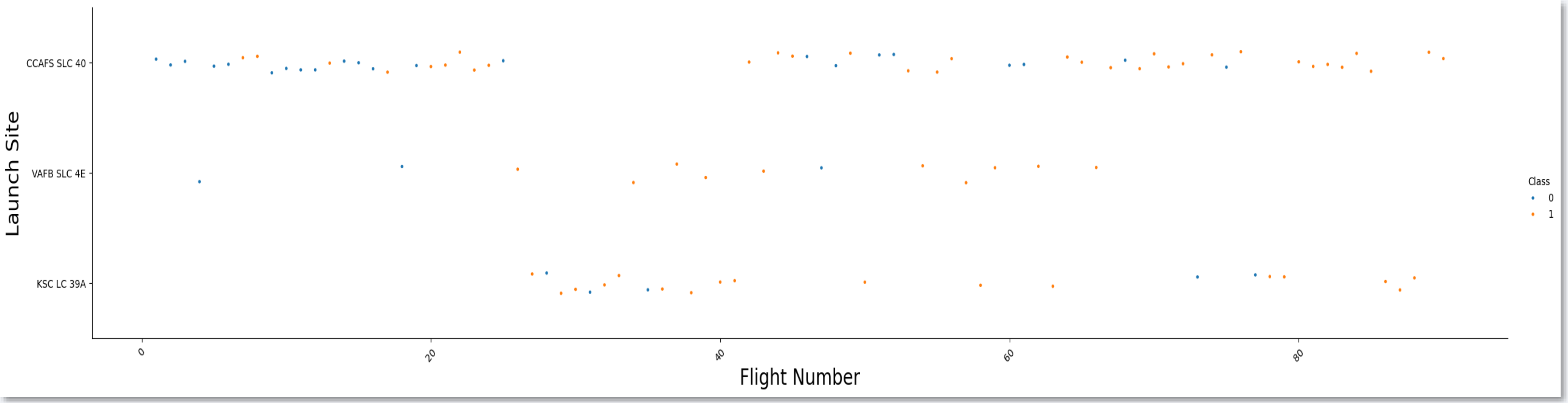
- The EDA and interactive analytics provided DEEP INSIGHTS into factors affecting the Falcon 9 first stage landing success, valuable for operational improvements and strategic planning.
- The predictive analysis identified the DECISION TREE MODEL AS A PROMISING TOOL for forecasting landing outcomes, although further refinement and validation are necessary.
- The dashboard equips the company with ONGOING ACTIONABLE INSIGHTS into the complexities of reusable rocket landings, supporting strategic decisions in future missions and technological developments.
- Future work should explore additional features, more complex models, and a larger dataset for IMPROVED PREDICTIVE ACCURACY. Additionally, incorporating weather conditions and technical details of each flight could offer deeper insights into landing success factors.

Section 2

Insights drawn from EDA

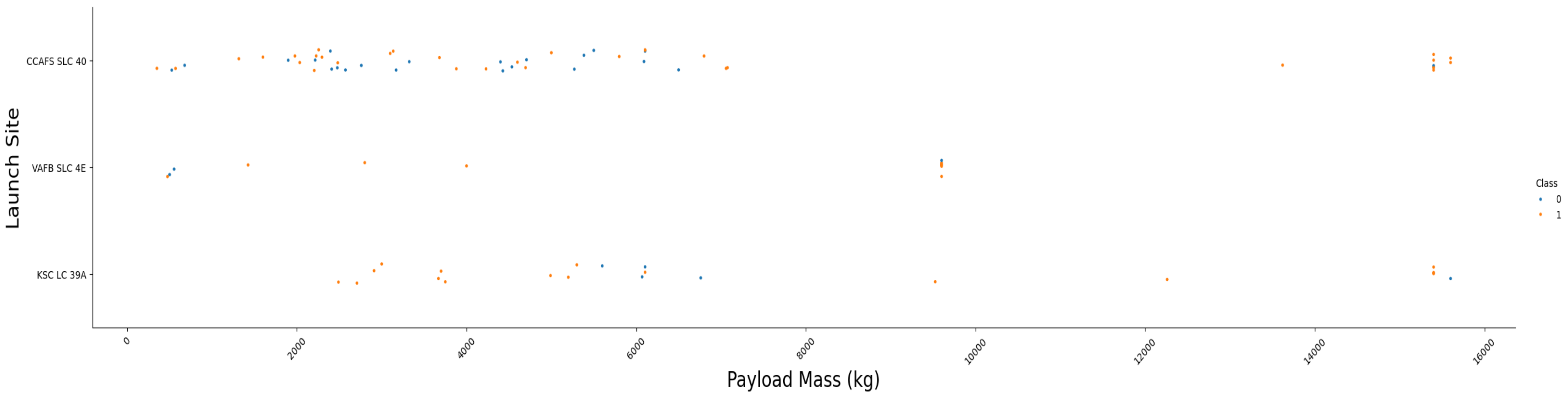
Flight Number vs. Launch Site

- Showing the relationship between SpaceX flight numbers and their respective launch sites, with an additional classification denoted by color.
  - The success rate (assuming class '1' indicates success) appears relatively high across all launch sites.
  - The plot may suggest an improving trend of success rates over time, given that higher flight numbers tend to have a higher density of successful class markers.



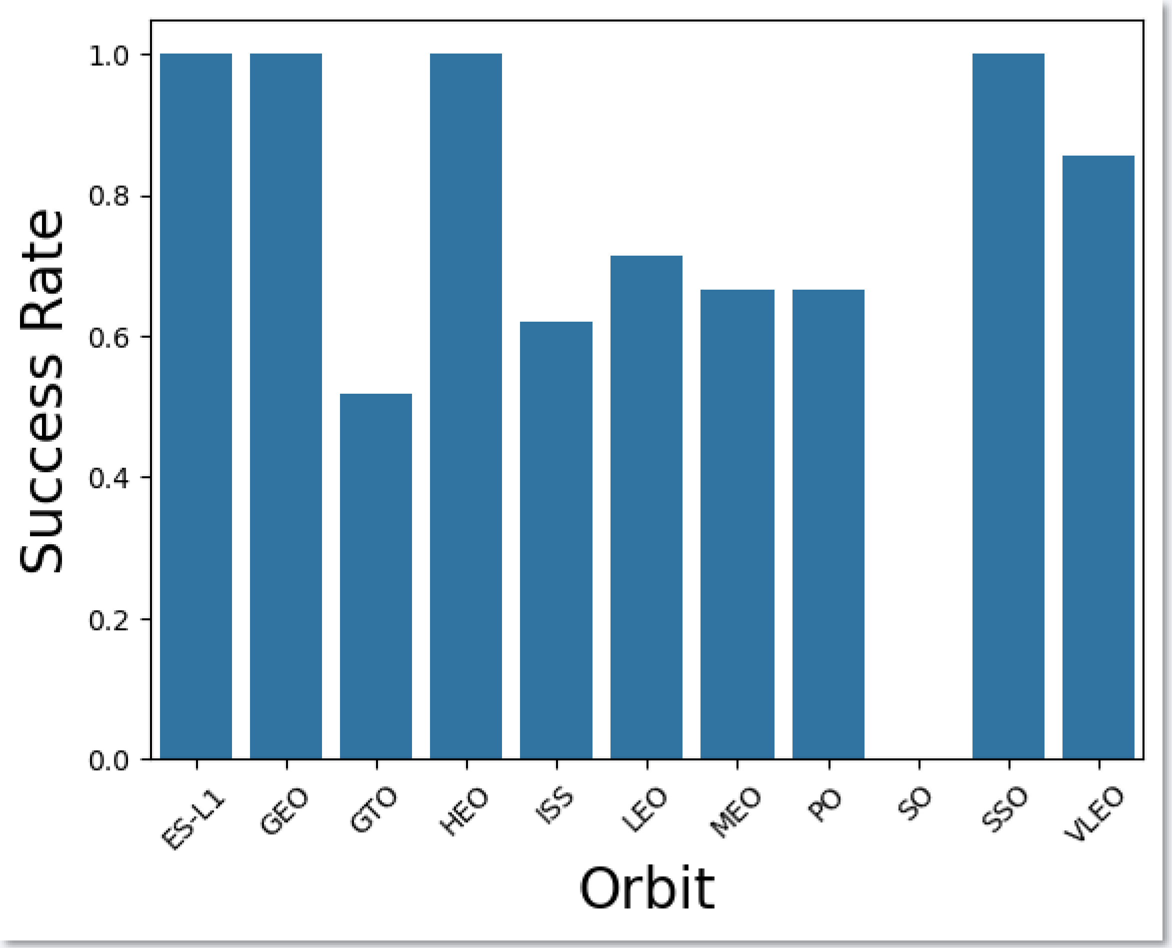
Payload vs. Launch Site

- Displaying the relationship between payload mass (in kilograms) and the corresponding launch sites, with an additional dimension indicating the class - which likely represents the success ('1') or failure ('0') of the booster landing.
  - We could infer that SpaceX's launch success is not adversely affected by higher payload masses, which is significant for the viability of their launch operations.



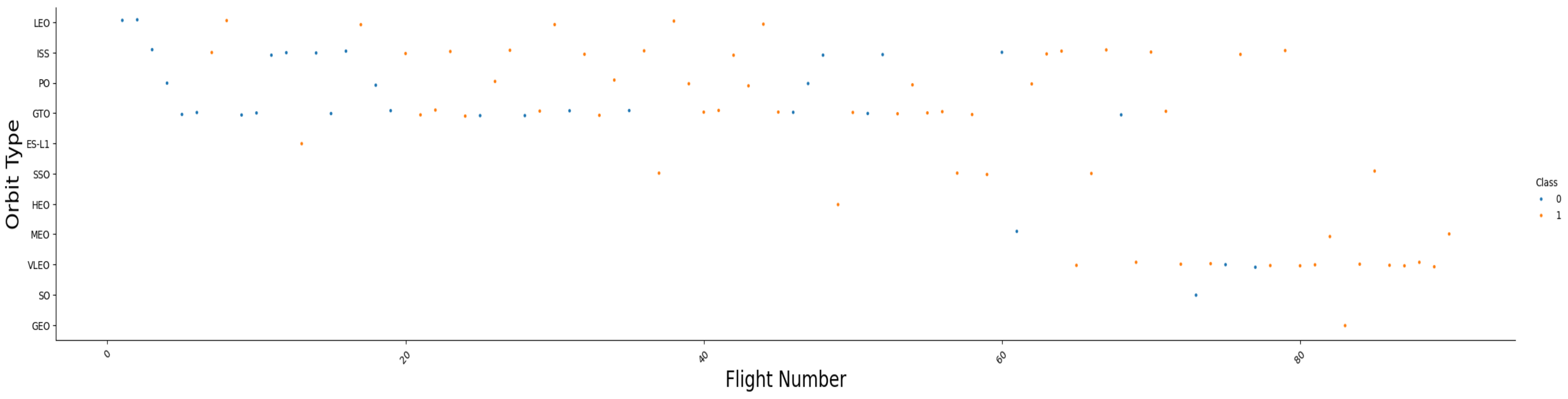
Success Rate vs. Orbit Type

- Stakeholders may infer where future mission planning and resource allocation might be directed to improve success rates in the more challenging orbits.
  - Certain orbits, such as ES-L1, GEO, and VLEO, have higher success rates close to 1.
  - HEO and MEO, have notably lower success rates.
  - The LEO (Low Earth Orbit) also has a high success rate, which is significant as it is a common destination for many satellites.



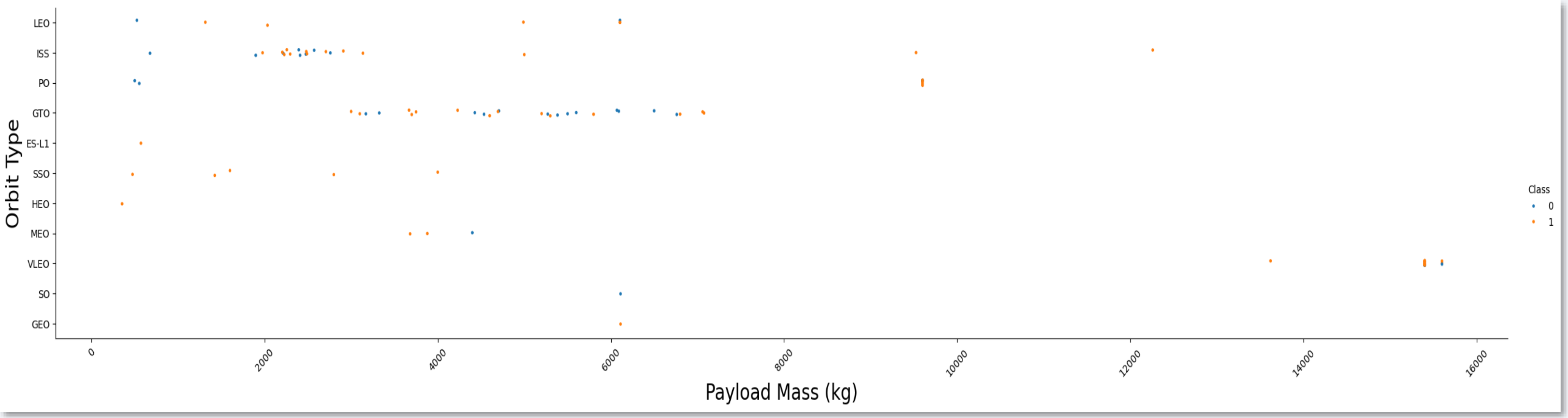
Flight Number vs. Orbit Type

- Relationship between the flight number of a rocket launch and the type of orbit it was destined for, with an additional classification indicated by color.
  - Early flights mostly reached LEO, with increasing variation to other orbit types in later flights.
  - Recent flights seem to target other orbits, with a focus on VLEO.
  - May suggest that success rates do not significantly vary across different orbit types.



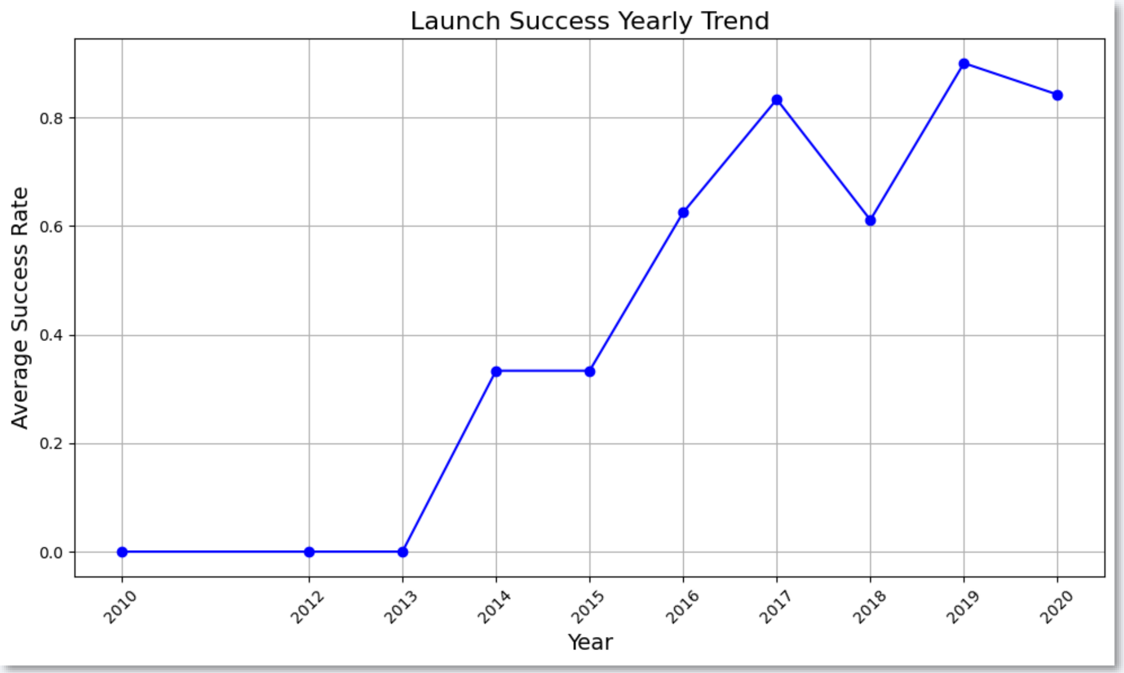
Payload vs. Orbit Type

- Relationship between the payload mass (presumably in kilograms) of rocket launches and the types of orbits achieved, with color coding to indicate the success (1) or failure (0) of each mission.
    - Indicating the flexible capabilities of the rocket in terms of mission requirements.
    - Might suggest that failure rates do not significantly correlate with the increase in payload mass.
    - Indicating either fewer launches with heavy payloads or a limit to the rocket's payload capacity.



Launch Success Yearly Trend

- Average launch success rate of rockets by year.
- Significant increase in launch success rate from 2013 to 2014.
- The peak success rate occurs in 2017 and 2019.
- Noticeable dip in success rate in 2018
- Trend of increasing success rate over the years.



All Launch Site Names

- CCAFS LC-40: Cape Canaveral Air Force Station Launch Complex 40, located in Florida, has been used for a wide array of satellite launches.
- VAFB SLC-4E: Vandenberg Air Force Base Space Launch Complex 4E, situated in California, is employed for polar orbits due to its favorable location.
- KSC LC-39A: Kennedy Space Center Launch Complex 39A, also in Florida, is a historic site that has been adapted for launches of the Falcon 9 and Falcon Heavy rockets.
- CCAFS SLC-40: This is another pad at Cape Canaveral Air Force Station and serves as a primary site for Falcon 9 launches.

Launch Site Names Begin with 'CCA'

The following is a limited list of launches, restricted to only 5 records and only records where the Lauch_Site name starts with “CCA”

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

Total_Payload_Mass
45596

According to the Lab, this is the total payload in KG carried for the SpaceX customer: NASA (CRS)

Average Payload Mass by F9 v1.1

Avg_Payload_Mass
2928.4

According to the Lab this is the average all time payload carried by boosters of version ‘F9 v1.1’

First Successful Ground Landing Date

These are the dates of the first successful landing outcome on ground pad.

The first successful ground landing was in the year 2015 on the 22nd of December.

Successful_Landing
2015-12-22
2016-07-18
2017-02-19
2017-05-01
2017-06-03

Successful Drone Ship Landing with Payload between 4000 and 6000

Here are the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 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

The result seems to show multiple entries for "Success“:

- This can happen if the Mission_Outcome entries have trailing spaces or are not consistently cased, or if there are other invisible characters.
- It's also possible that the outcomes are categorized differently (for example, "Success" and "Success (payload status unclear)" are considered different outcomes).

A cleanup is suggested, to get a more straightforward count of success and failure outcomes.

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

Boosters Carried Maximum Payload

These are the names of the booster which have carried the maximum payload mass.

These boosters were selected after first identifying the maximum payload mass and subsequently only finding the booster that had carried that 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

Here are the failed landing outcomes in drone ship, their booster versions, and launch site names for the year 2015, organized by month.

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 is a ranking of landing outcomes by outcome (such as Failure (drone ship) or Success (ground pad))

- between the date 2010-06-04 and 2017-03-20,
- in descending order

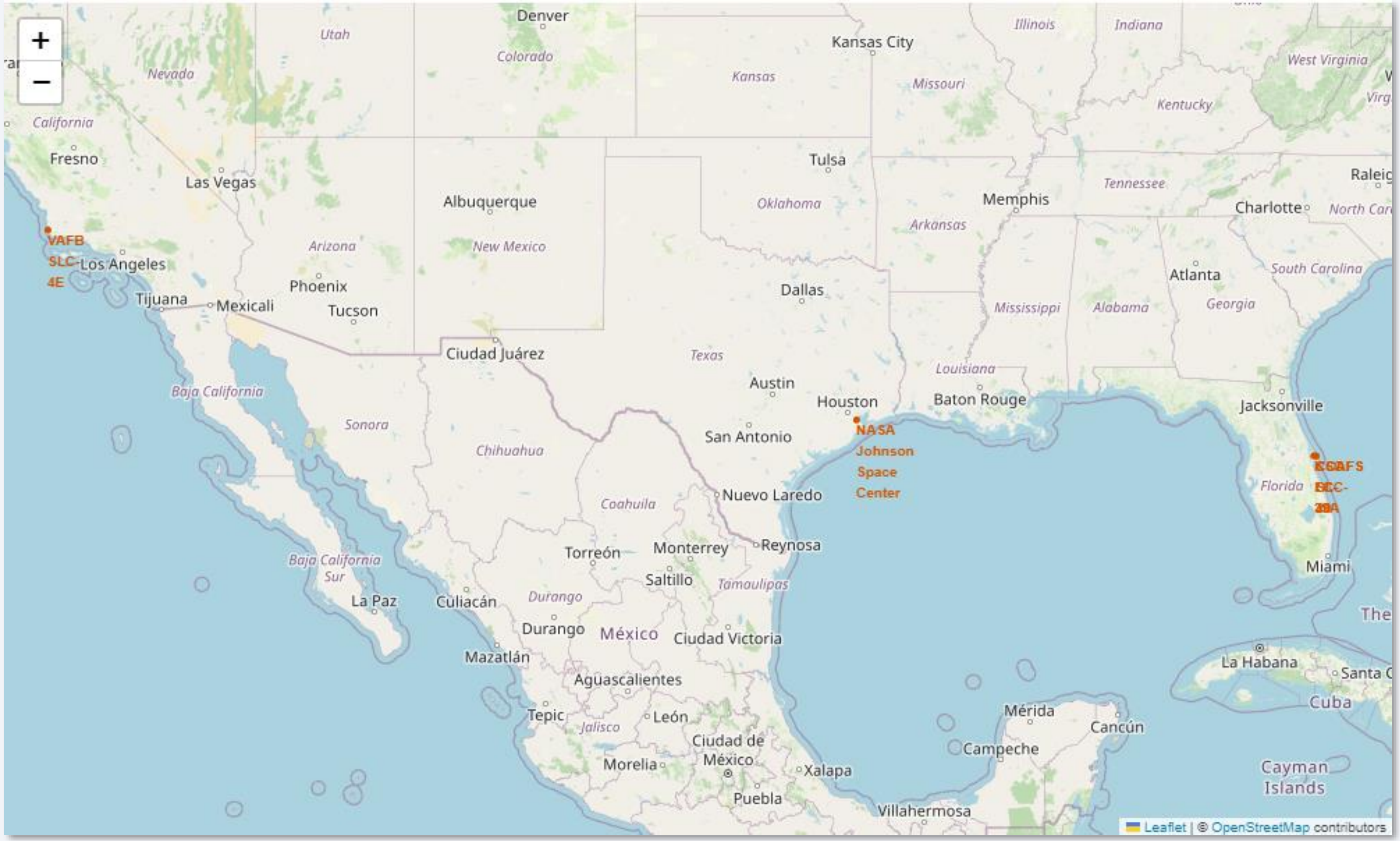
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

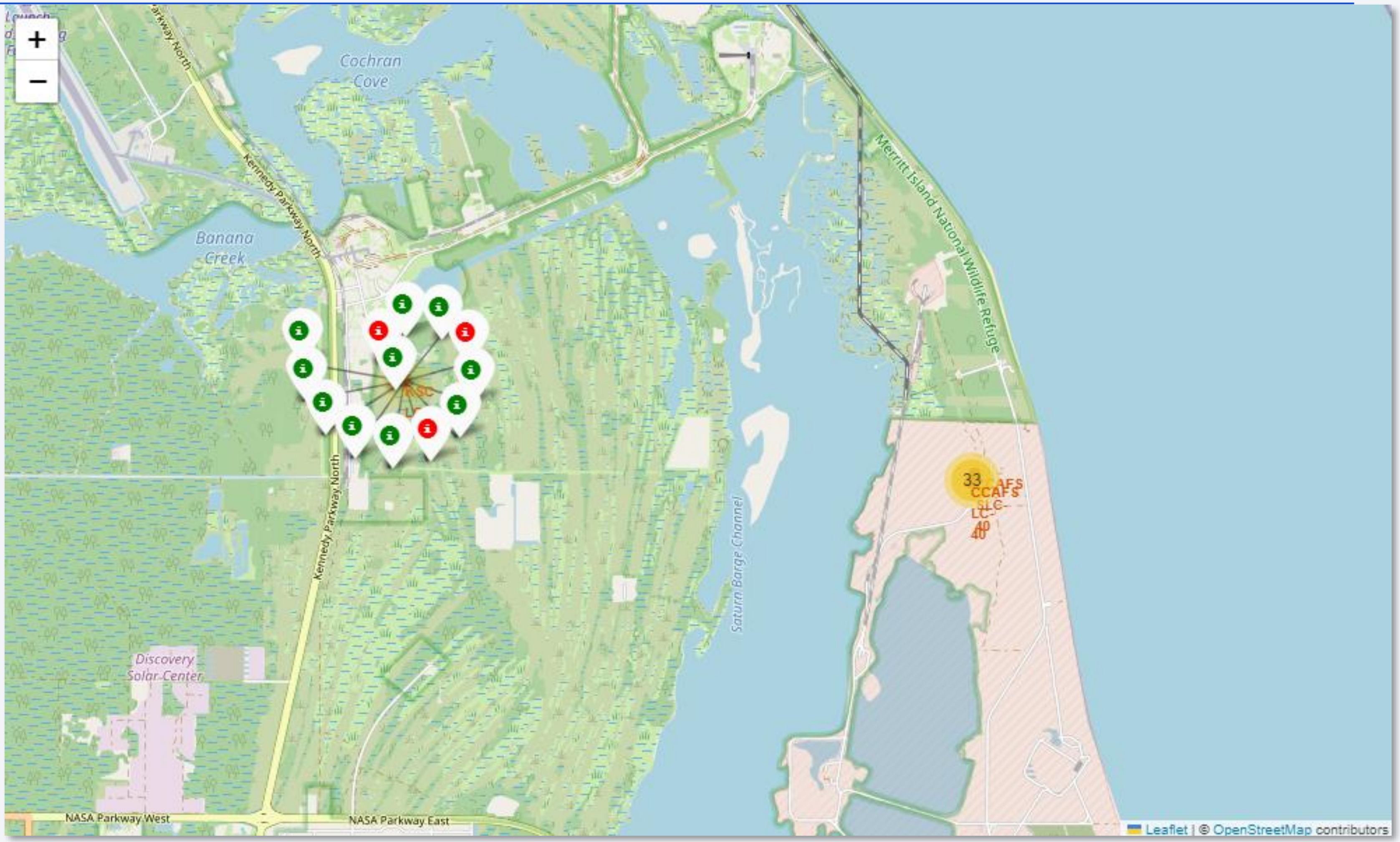
All launch sites' location

- Launch Sites are on the West Coast in Florida and in California, north of Los Angeles.
- The map also shows NASA´s Johnson Space Center near Houston, TX.



Color-coded Labels showing Outcomes per Site

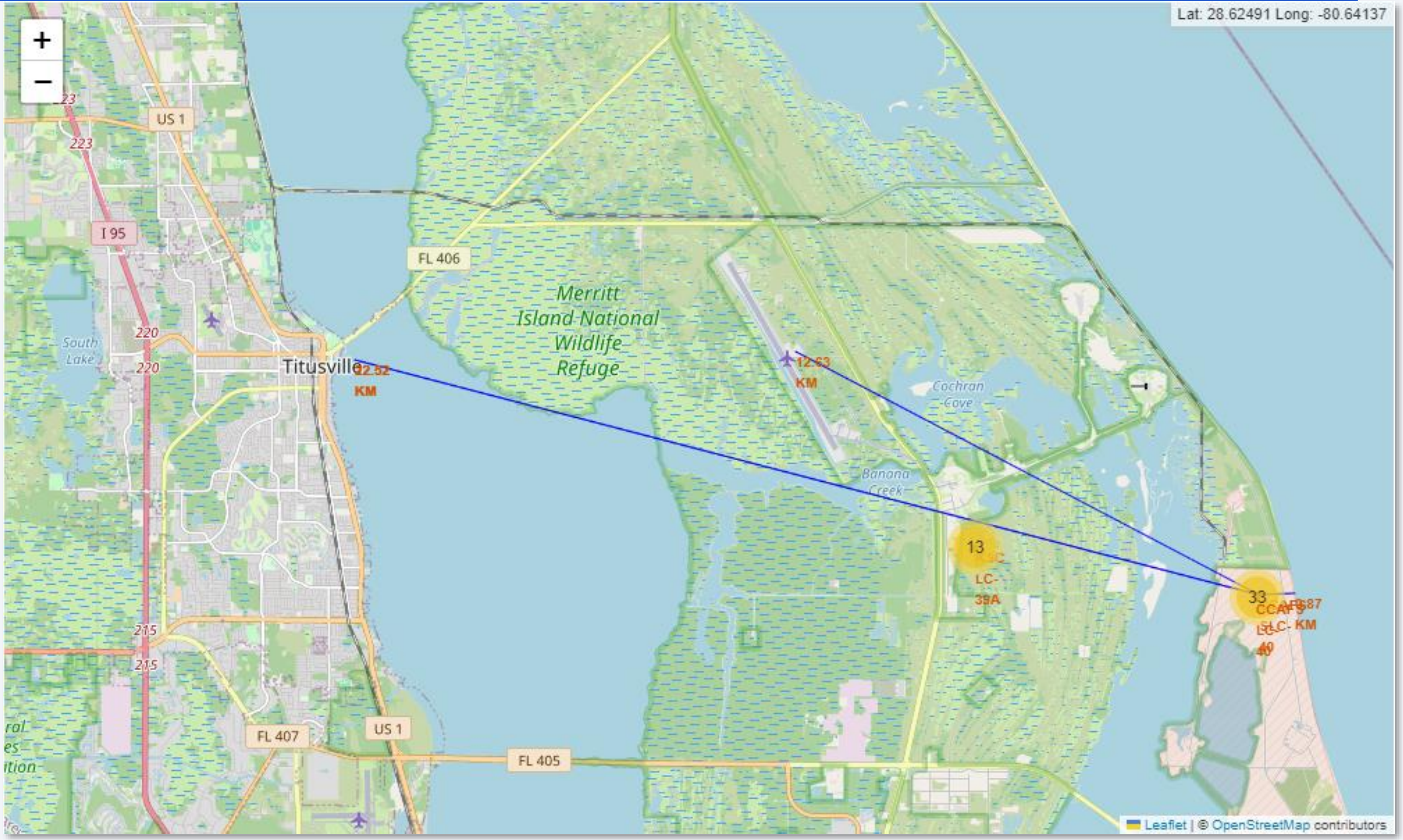
- Launch Sites outcomes are organized in cluster displays.
- When clicked, the outcomes for the selected launch site will show:
    - Green = success
    - Red = failure



Launch Site Proximities

- Launch Site can be analysed in respect to its proximities such as railway, highway, coastline, with distance calculated and displayed.

- Proximities play an important role in considering aspects of guaranteeing a successful outcome

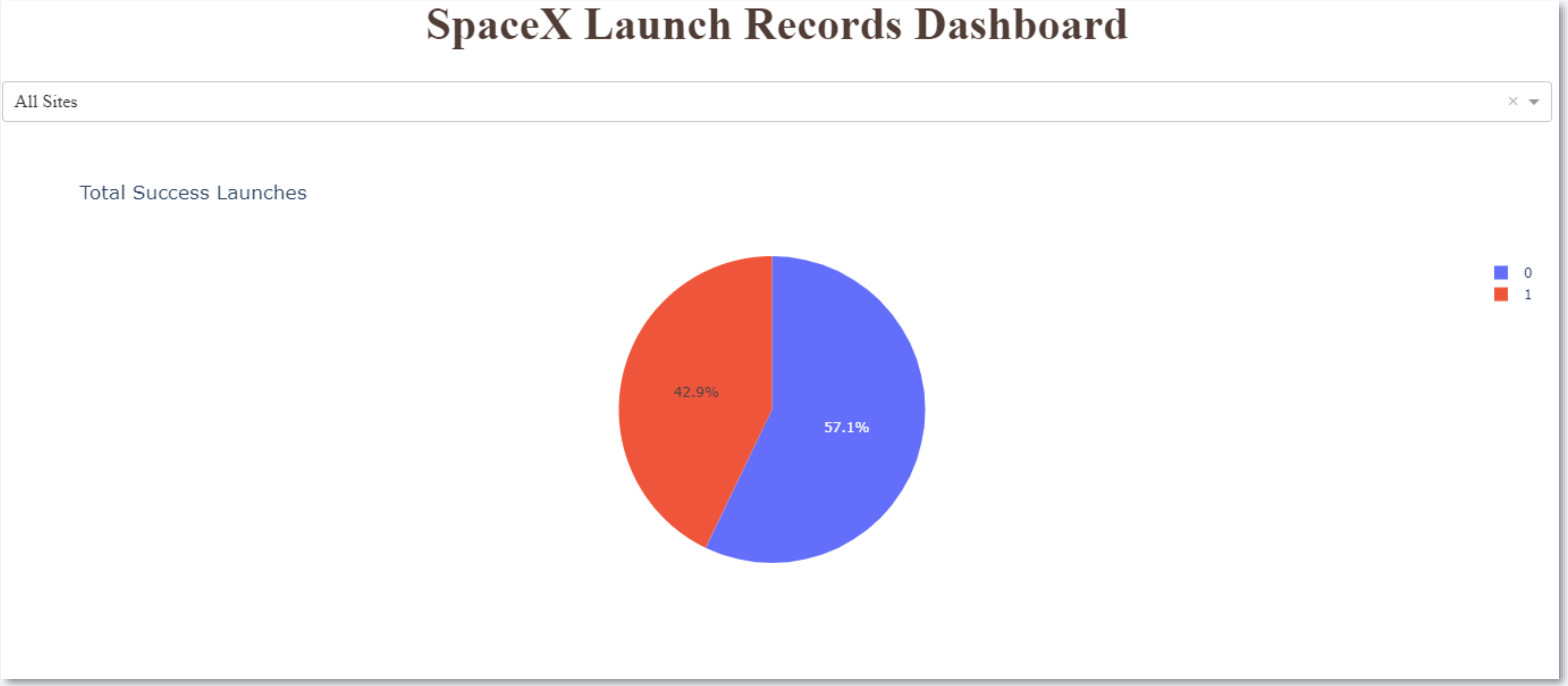


Section 4

Build a Dashboard with Plotly Dash

Launch success count for all sites

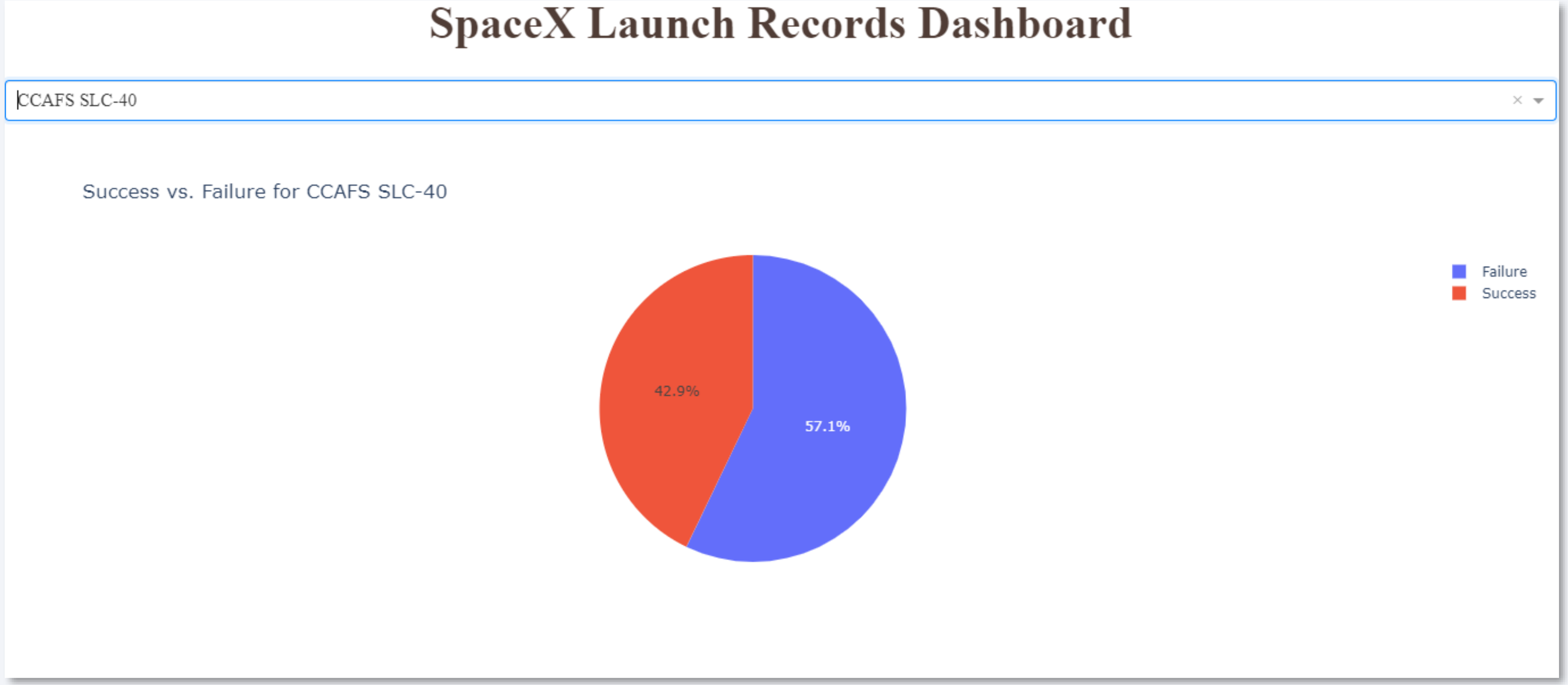
- Dashboard allows to select all or a particular site and shows the success by color-coded pie chart.



Highest launch success ratio

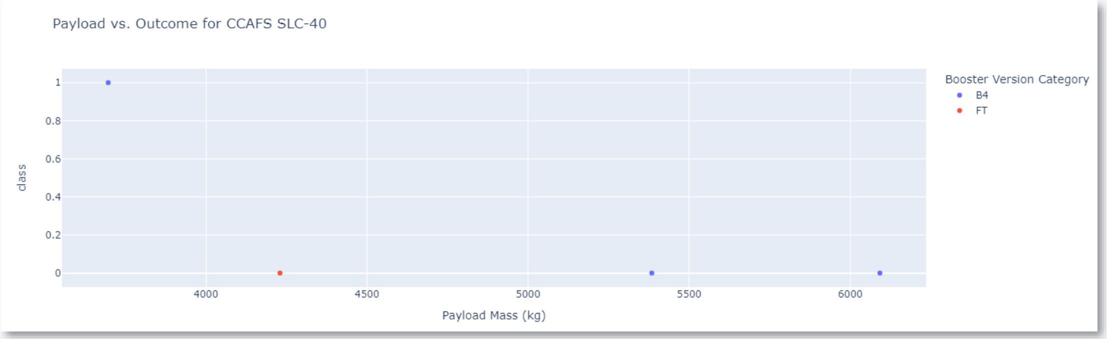
- The dashboard allows to check and verify the launch outcomes for each site.

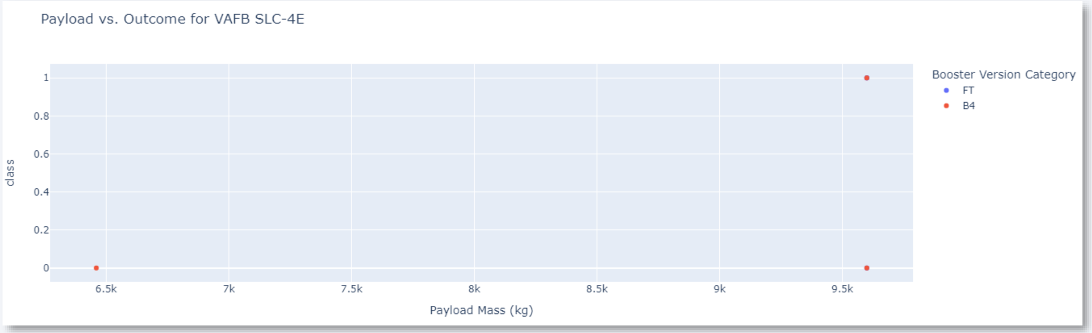
- This is the display of CCAFS SLC-40 which has the highest ratio for launch successes.

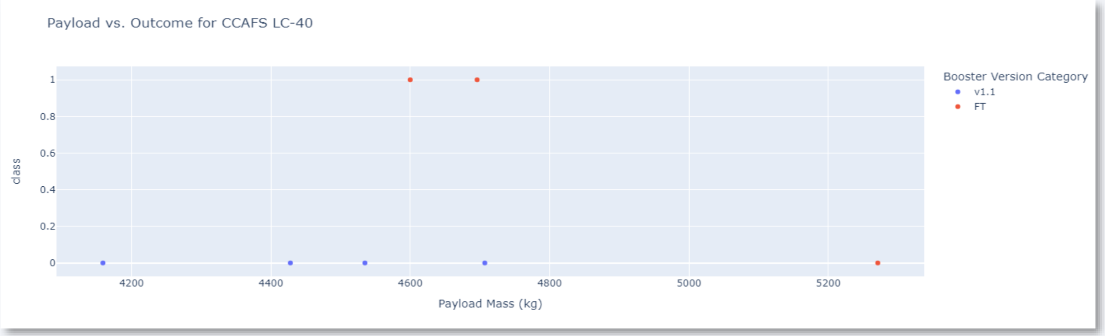


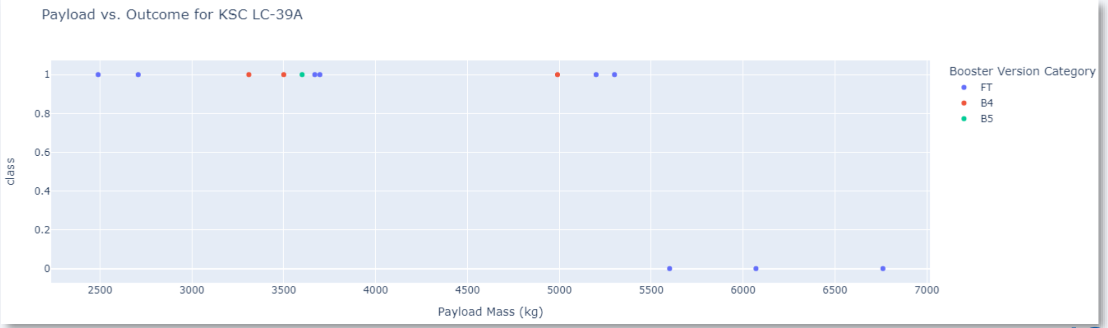
Payload vs. Launch Outcome

- Using the combination of selecting site vs payload a detailed relationship can be explored. The color-coding for the booster version allows even more insight.







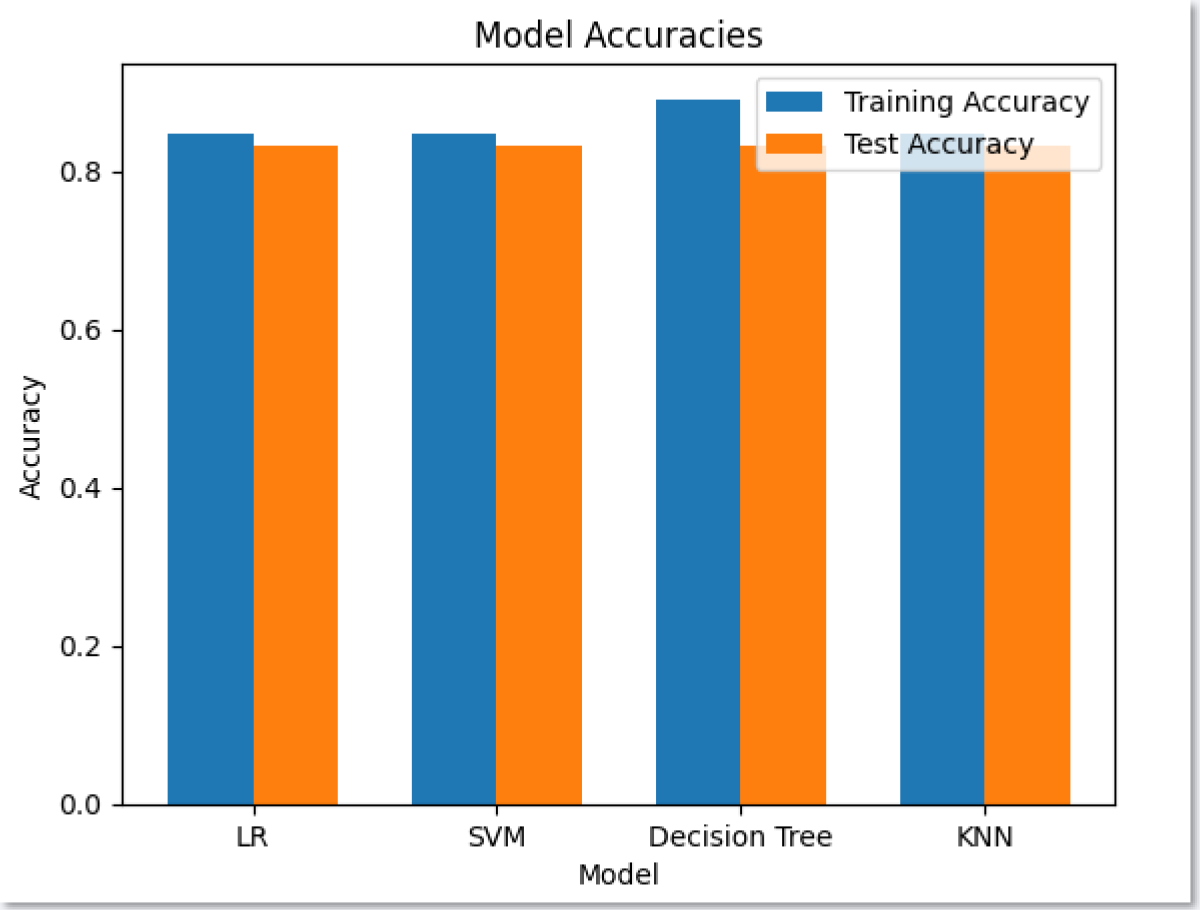


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualizing the built model accuracy for all built classification models.
- The Decision Tree appears to be the model that performs best, because it has the highest classification accuracy.

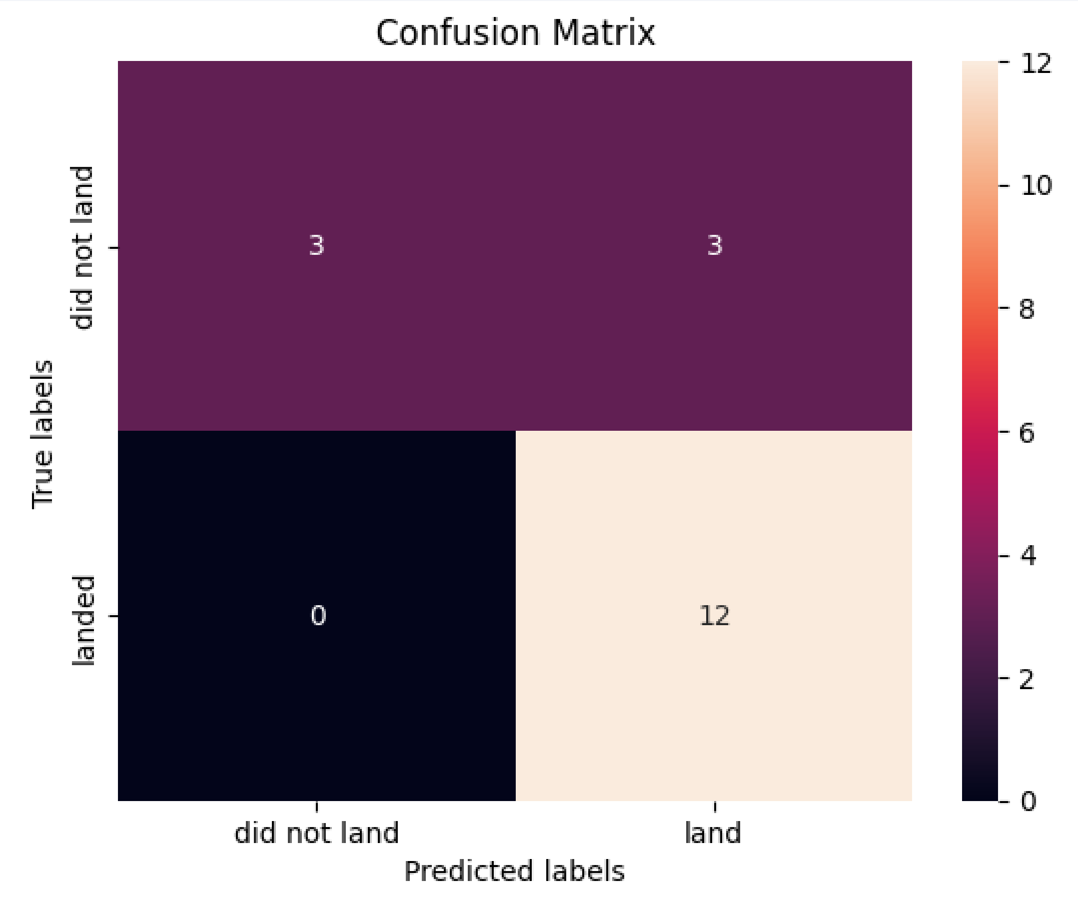


Confusion Matrix

- Here is the confusion matrix of the best performing model: decision tree.

- We see that a decision tree can distinguish between the different classes.

- We also see that the major problem is false positives.



Conclusions

- Predictive Precision: Successfully established a predictive framework that forecasts Falcon 9 landing outcomes with a noteworthy degree of accuracy.
- Cost Efficiency: The model underscores the financial viability of reusable launch systems, offering SpaceX a strategic advantage in competitive pricing.
- Launch Site Performance: Identified the most efficient launch sites, equipping SpaceX with data to optimize launch scheduling and resource allocation.
- Payload Correlations: Unveiled trends between payload mass and launch success, providing insights for payload planning and risk assessment.
- Orbital Success Patterns: Revealed the orbit types with higher success rates, guiding mission selection and customer engagement strategies.
- Temporal Trends: Illustrated a positive trajectory in launch success rates over time, reflecting SpaceX's technological advancements and operational refinement.
- Strategic Recommendations: Based on the analyses, recommend focusing on the most successful launch sites and orbits, and further investigation into the outliers in payload mass that affect success rates.
- Enhancement Roadmap: Suggest further model enhancements with expanded data sets and advanced algorithms to elevate predictive accuracy and reliability.

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!