Winning Space Race with Data Science

Nguyen Thien Thuan
11/11/2025

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

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

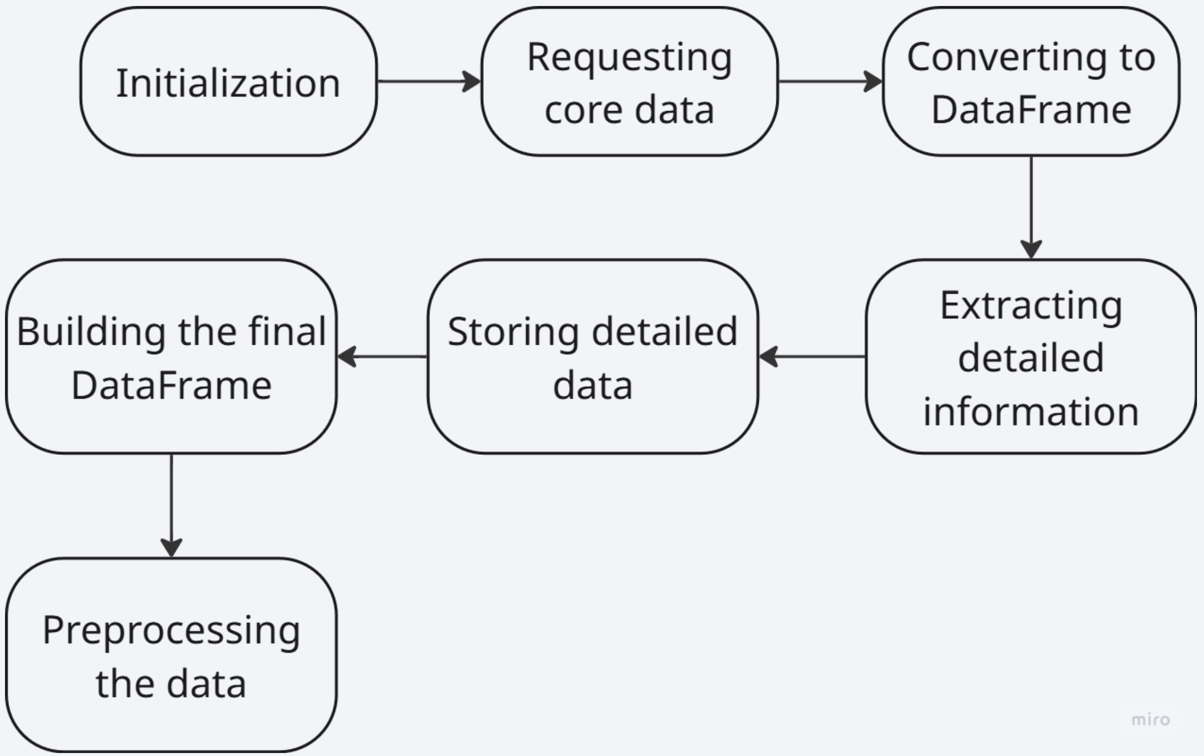
- Sending a GET request to the API
- Parsing the JSON response
- Extracting detailed information
- Storing the detailed data
- Creating a new DataFrame
- Filtering the data
- Handing missing values

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- This is the GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module1_introduction/jupyter-labs-spacex-data-collection-api.ipynb

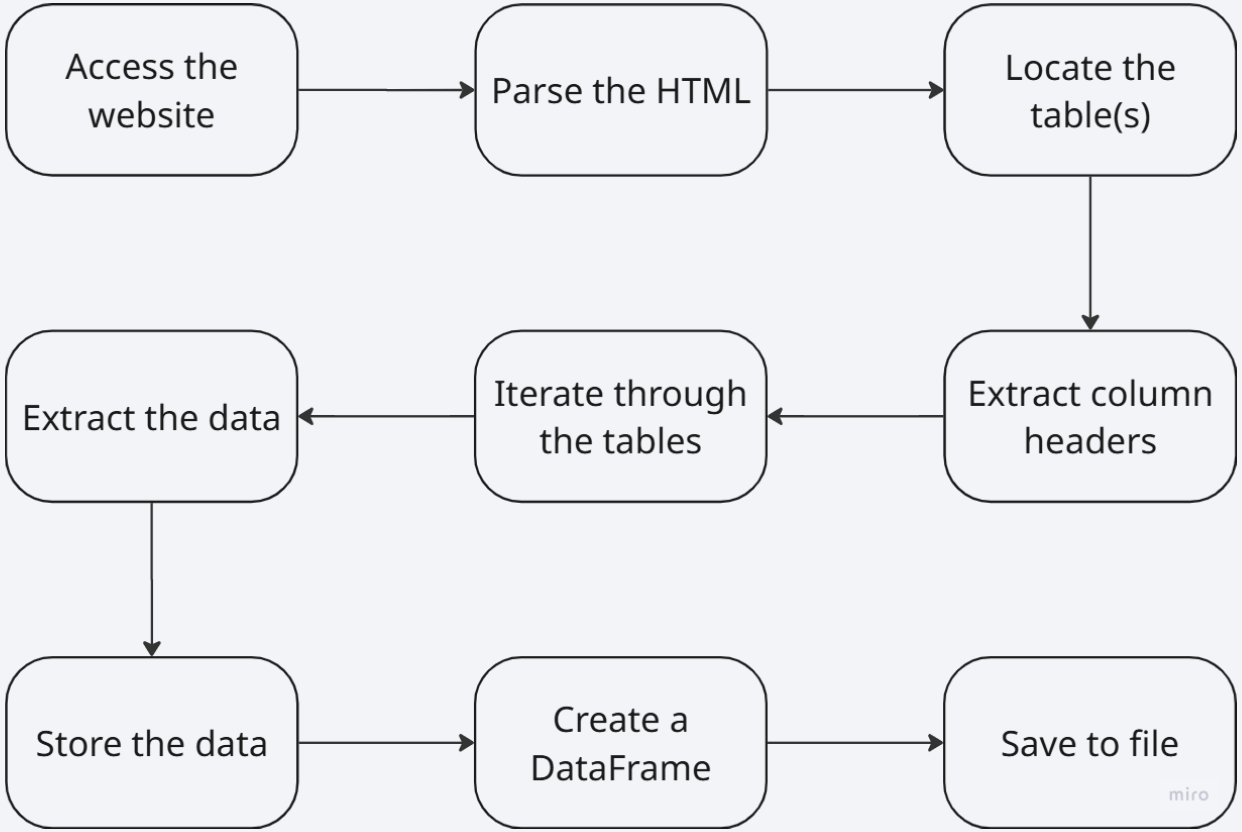


Data Collection - Scraping

- Web Scraping process

- This is the GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module1_introduction/jupyter-labs-webscraping.ipynb

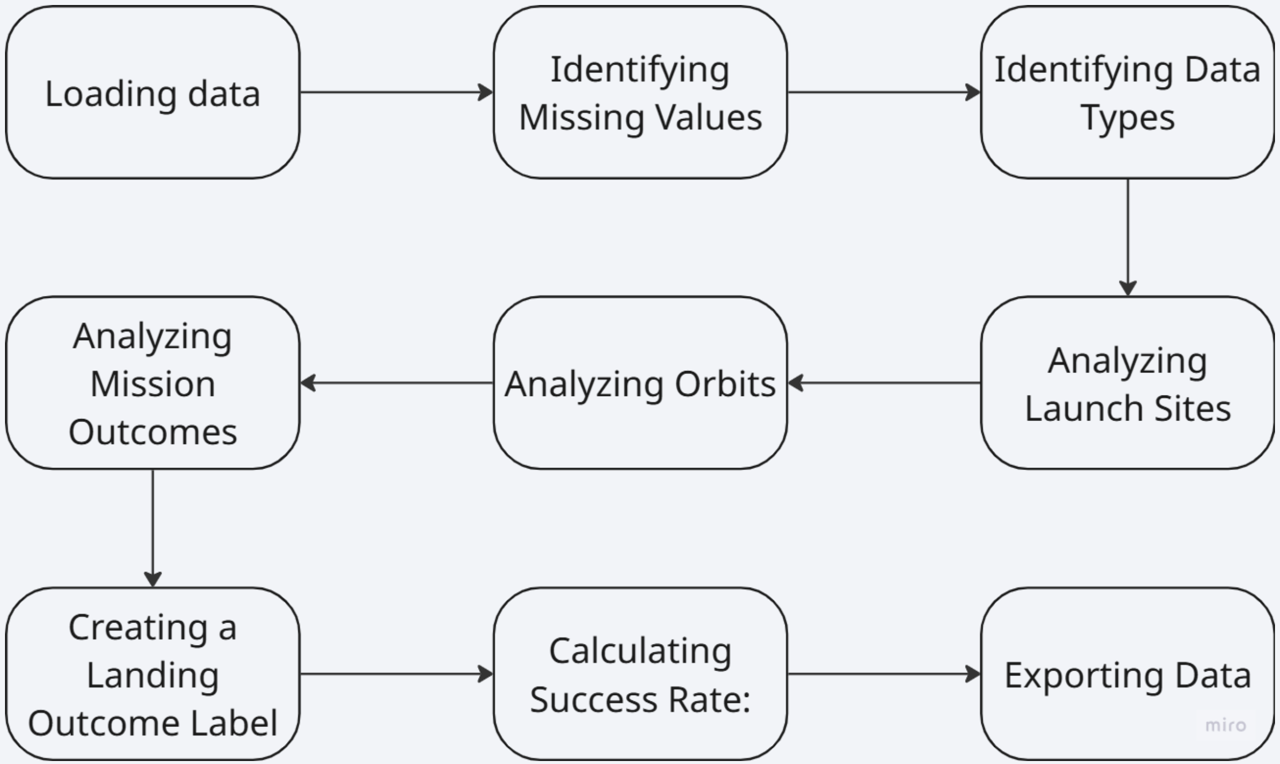


Data Wrangling

- Data processing workflow:

- This is the GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module1_introduction/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

Charts that were plotted:

1. Scatter point char: Flight Number vs. Payload Mass (with Outcome)

Reason for use:
To explore the relationship between two quantitive variables (FlightNumber, PayloadMass) and one categorical variable (Class). The scatter plot allows us to easily observe individual data points and identify any trends or clusters based on landing outcomes. Using *hue='Class'* provides a visual distinction between successful and failed launches.

EDA with Data Visualization

Charts that were plotted:

2. Scatter point chart: Flight Number vs. Launch Site (with Class)

Reason to use: To examine the relationship between a quantitive variable (FlightNumber) and a categorical variable (LaunchSite), categorized by outcome (Class). This help us understand the distribution of launches at each site over time and determine whether any location shows a better success rate over time

EDA with Data Visualization

Charts that were plotted:

3. Scatter point chart: PayLoadMass vs. Launch Site (with Class)

Reason to use: Similar to the previous chart, but replacing the quantitive variable with PayLoadMass. The purpose is to examine whether payload mass has different impacts on landing outcomes across various launch sites. It helps identify whether a specific launch site is more suitable for heavier payloads

EDA with Data Visualization

Charts that were plotted:

4. Bar chart of success rate by orbit:

Reason to use: To compare the average success rate across different types of orbits. A bar chart is the best choice for presenting the comparison of quantitative variable (success rate) across different froups (orbit types). This helps quickly identify which orbit type has the highest likelihood of successful landings

EDA with Data Visualization

Charts that were plotted:

5. Scatter point chart: FlightNumber vs. Orbit (with class):

Reason to use: To analyze the relationship between flight number and orbit type, categorized by outcome. This chart helps us examine whether the tred of success over time (represented by flight number) varis across difference orbit types

EDA with Data Visualization

Charts that were plotted:

6. Scatter point chart: PayLoadMass vs. Orbit (with class):

Reason to use: To examine the relationship between payload mass and orbit type, categorized by outcome. The goal is to determine whether landing outcomes vary based on payload mass within each specific orbit type

EDA with Data Visualization

Charts that were plotted:

7. Line chart of annual launch success trends:

This is GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module2_exploratory_data_analysis/edadataviz.ipynb

EDA with SQL

SQL Queries Performed:

- Display the names of unique launch sites.
- Show 5 records where the launch site names start with the string ‘CCA’
- Display the total payload mass carried by rockets launches by NASA (CRS)
- Show the average payload mass carried by the F9 v1.1 booster version
- List the date of the first successful landing outcome on a landing pad
- List the names of boosters that successful landed on a drone ship and carried payloads greater than 4000 but less than 6000
- List the total number of successful and failed mission outcomes

EDA with SQL

SQL Queries Performed:

- List all booster versions that carried the maximum payload mass, using a subquery with an appropriate aggregate function
- List records showing the month name, failed landing outcome on a dront ship, booster version, and launch site during the months of 2015
- Rank the number of landing outcomes (e.g., Failure (drone ship) or Success (landing pad)) between 2010-06-04 and 2017-03-20, in descending order

This is the GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module2_exploratory_data_analysis/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Objects added to the map:

1. folium.Map object:

This is the fundamental object used to create an interactive map. It provides a canvas for adding map layers and other objects. In this notebook, the map was initialized with the NASA Johnson Space Center as the initial center location.

2. folium.Circle object:

This object is used to draw circles on the map at specific coordinates. Initially, a blue circle was added at the location of NASA Johnson Space Center to demonstrate its usage. Later, smaller circles were added at each launch site to mark their positions on the map. A small radius (100 meters) helps accurately pinpoint the exact location of each launch site.

Build an Interactive Map with Folium

Objects added to the map:
3. folium.Marker object:

Markers are used to highlight specific points on the map with an icon or text “NASA JSC” to demonstrate how to use DivIcon for creating custom text labels. Later, markers were added for each launch site, displaying the site name as a text label. This is helps users easily identify the name of each launch site when viewing the map.

Markers were also used on Task 2 to mark individual launches. The color of these markers (green for success, red for failure) visually represents the launch outcoems at each site. Markers with distance labels were added to points of interest near the launch sites (e.g., coastline, city, railway) to visualize the distance from the launch site to those objects.

Build an Interactive Map with Folium

Objects added to the map:
4. folium.plugins.MarkerCluster object:

When there are many markers located close to each other – especially at the same coordinates (such as multiple launches from the same site) – MarkerCluster helps group them into a single cluster icon. When zoomed in, the cluster expands to reveal individual markes. This simplifies the map and improves performance when handling a large number of data points.

Build an Interactive Map with Folium

Objects added to the map:

5. folium.plugins.MousePosition object:

This plugin displays the coordinates of the mouse pointer on the map. It is very useful retrieving the coordinates of points of interest on the map for later distance calculations.

6. folium.PolyLine object:

PolyLine is used to draw straight or polyline paths on the map, connecting two or more points. PolyLine was used to draw lines from the launch site to nearby points of interest to visually represent the calculated distances.

Build an Interactive Map with Folium

This is GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module3_interactive_visual_analytics_and_dashboard/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Plot/Graphs added to the dashboard:

- Pie chart: Displays the total number of successful launches at each launch site.
- Scatter plot: Shows the correlation between payload mass and launch outcome

This is GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module3_interactive_visual_analytics_and_dashboard/spacex-dash-app.py

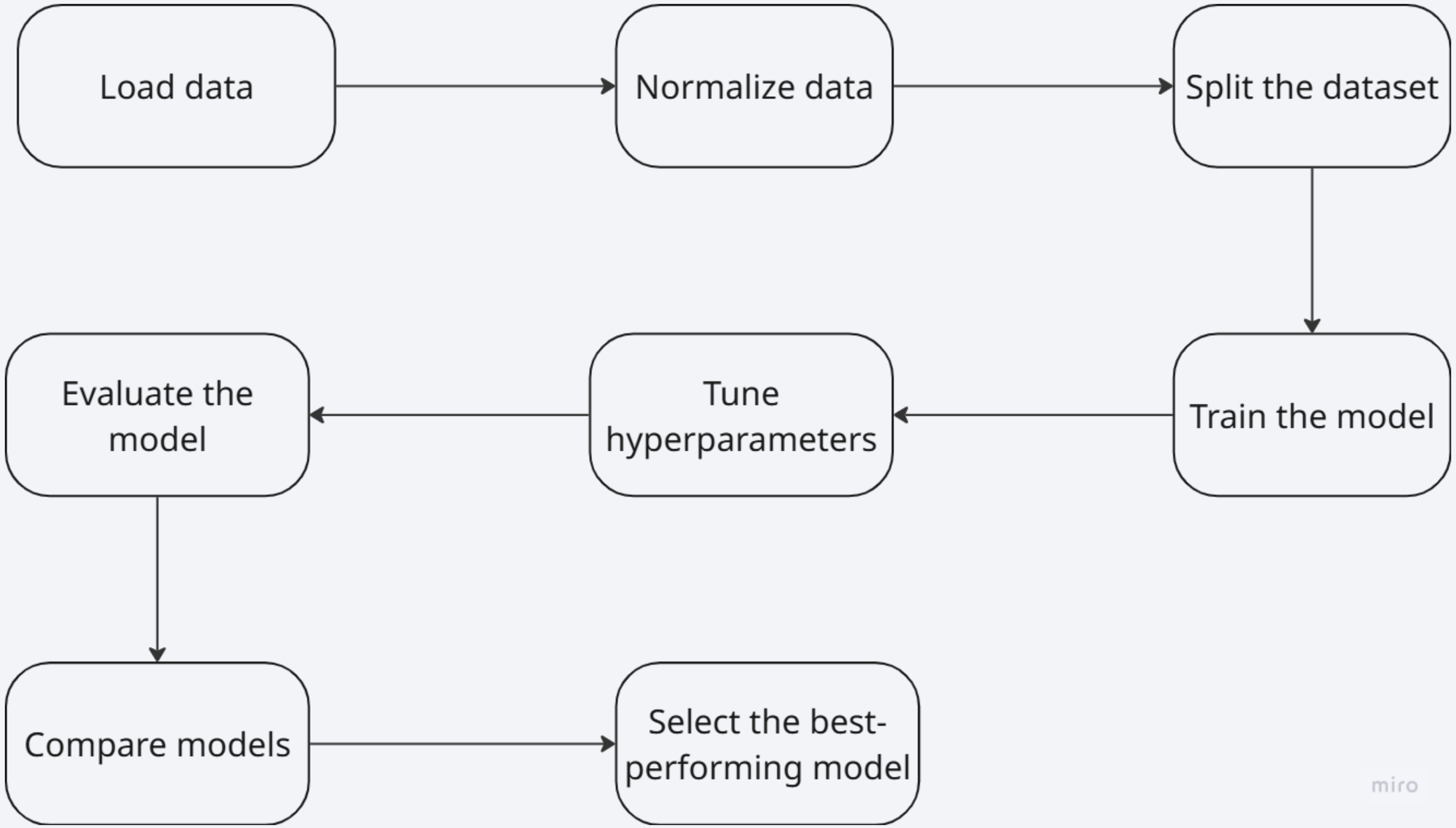
Predictive Analysis (Classification)

Summarize of how the best-performing classification model was built, evaluated improved, and selected:

- Data preparation
- Splitting data
- Model training and hyperparameter tuning
- Model evaluation
- Model comparison and selection

Predictive Analysis (Classification)

Model Development Process:



Predictive Analysis (Classification)

This is GitHub URL of the completed notebook:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course/blob/main/module4_predictive_analysis_classification/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

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

Results

Exploratory data analysis results

- Relationship between Flight Number and Payload Mass: As the flight number increases, the likelihood of a successful first-stage landing also increases. Payload mass also appears to be a factor influencing landing success.
- Relationship between Flight Number and Launch Site: Scatter plot analysis shows that different launch sites exhibit varying success trends over time (as represented by flight number).
- Relationship between Payload Mass and Launch Site: At the VAFB-SLC launch site, no rockets were launched with large payloads (greater than 10,000 kg). For other sites, the success rate with heavy payloads varies.

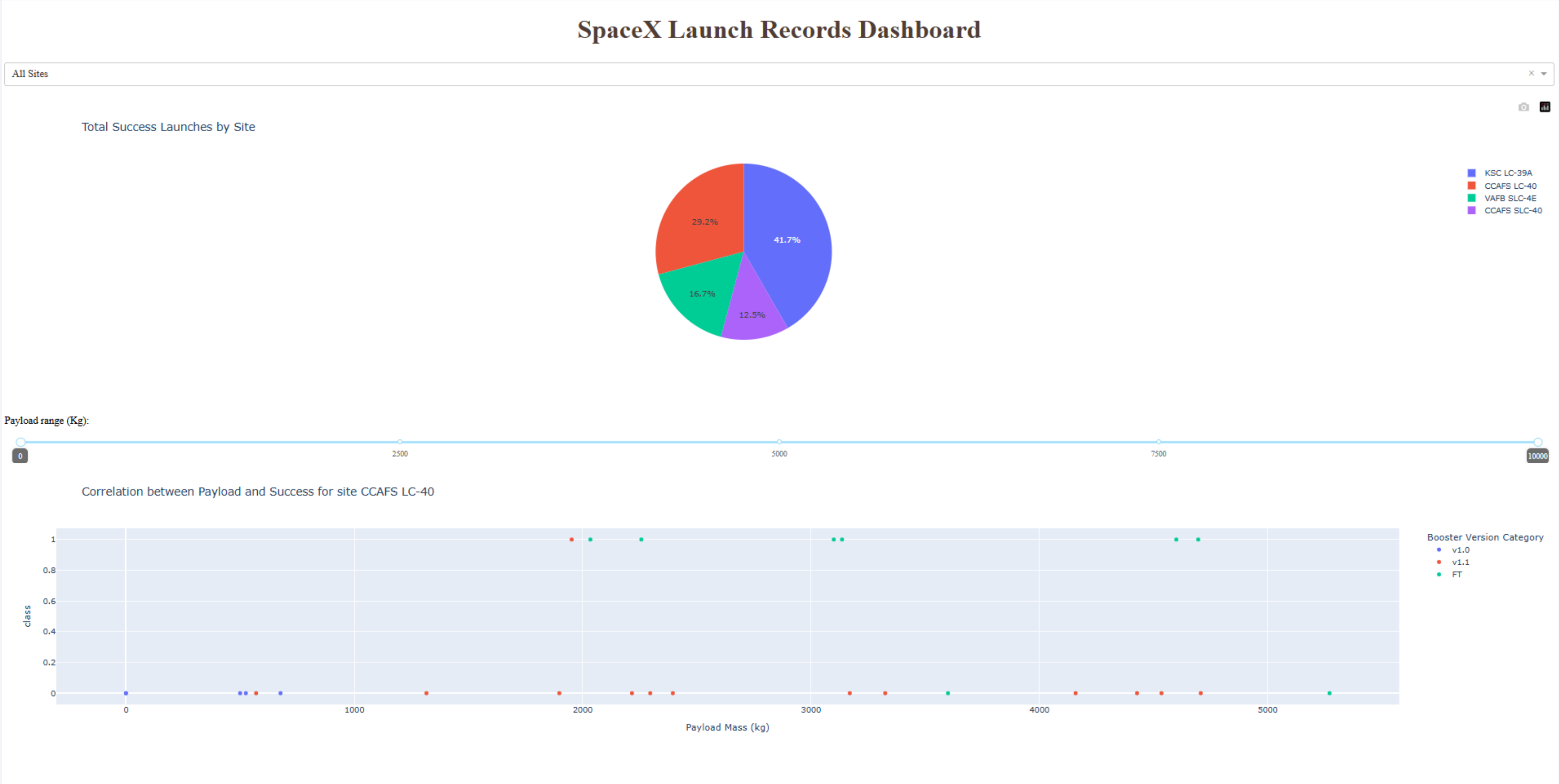
Results

Exploratory data analysis results

- Success Rate by Orbit Type: Orbit such as ES-L1, GEO, HEO, and SSO have the highest success rates (nealy 100%). Other orbits like GTO show lower success rates.
- Relationship between Flight Number and Orbit Type: In LEO orbit, success seems to be related to the number of launches. In contrast, for GTO orbit, there appears to be no clear relationship between number and success.
- Relationship between Payload Mass and Orbit Type: With heavy payloads, the success between successful and unsuccessful landings.
- Annual Launch Success Trend: The launch success rate has steadily increased from 2013 to 2020

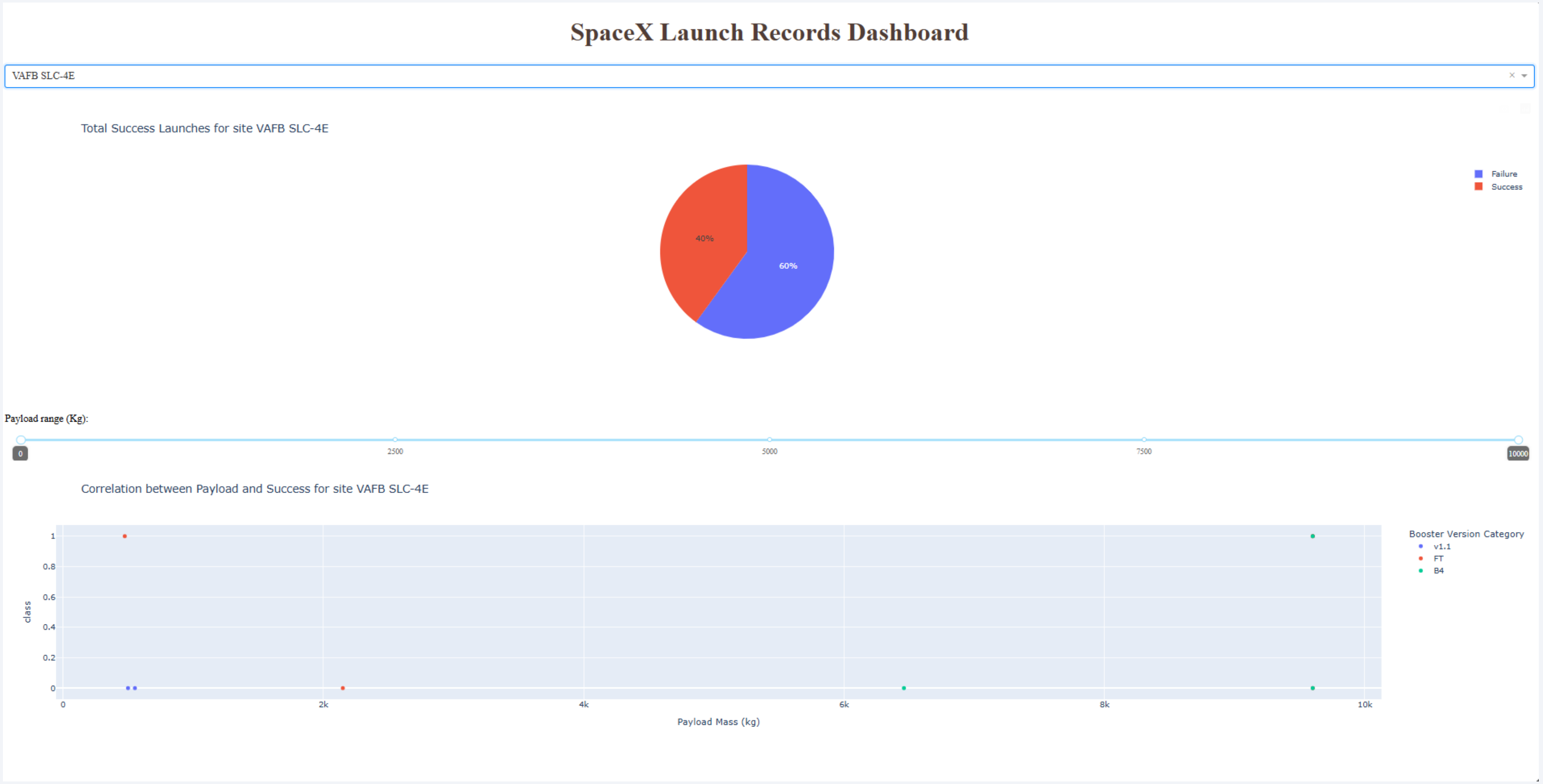
Results

Interactive analytics demo dashboard



Results

Interactive analytics demo dashboard



Results

Predictive analysis results

- Logistic Regression:
  - Best params: {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs’}
  - Validation accuracy ~ 0.8464
  - Test accuracy ~ 0.8333
  - Confusion matrix: 12 TP, 3 FP

Results

Predictive analysis results

- Support Vector Machine (SVM)
  - Best params: {'C': np.float64(1.0), 'gamma': np.float64(0.03162277660168379), 'kernel': 'sigmoid’}
  - Validation accuracy ~ 0.8482
  - Test accuracy ~ 0.8333

Results

Predictive analysis results

- Decision Tree
    - Best params: {'criterion': 'gini', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best’}
    - Validation accuracy ~ 0.875
    - Test accuracy ~ 0.333

Results

Predictive analysis results

- K-Nearest Neighbors (KNN)
  - Best params: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
  - Validation accuracy ~ 0.8482
  - Test accuracy ~ 0.8333

Results

Predictive analysis results

All models yielded similar accuracy on the test set (~83.33%) due to the small test size. The Decision Tree achieved the best validation accuracy (~87.5%), suggesting better learning during cross-validation.

The main error type was false positive (predicting a successful landing when it actually failed).

Additional metrics such as Precision, Recall, and F1-score were not yet evaluated- these should be included to better understand the balance between different types of errors.

Results

Predictive analysis results

Best model (validation): Decision Tree

Overall accuracy: ~83%

Main issue: False positive

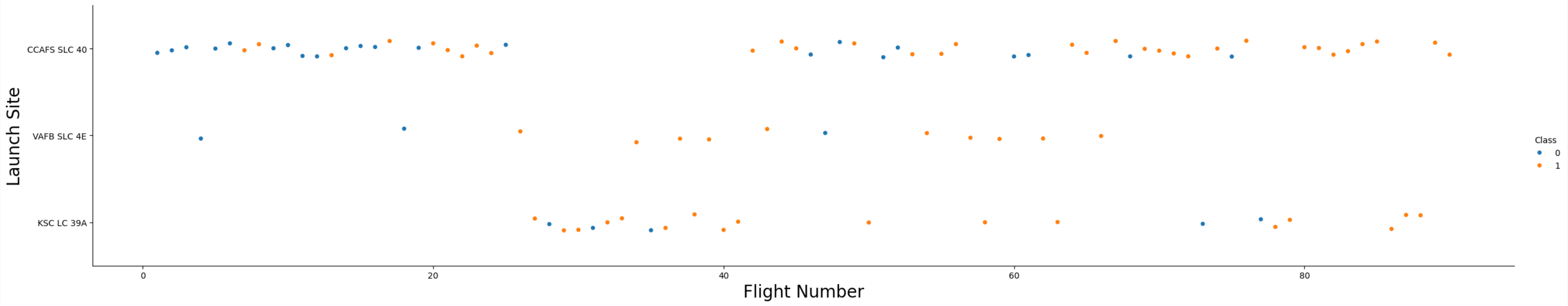
Recommendation: Further analysis of feature important and additional metrics should be included.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

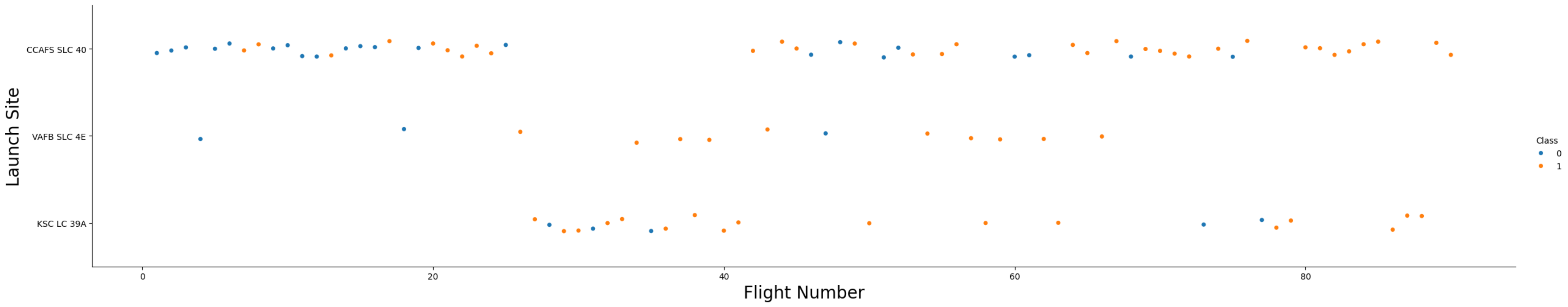
- Scatter plot of Flight Number vs. Launch Site



The relationship between flight number and launch site, categorized by landing outcome (success/failure)

Flight Number vs. Launch Site

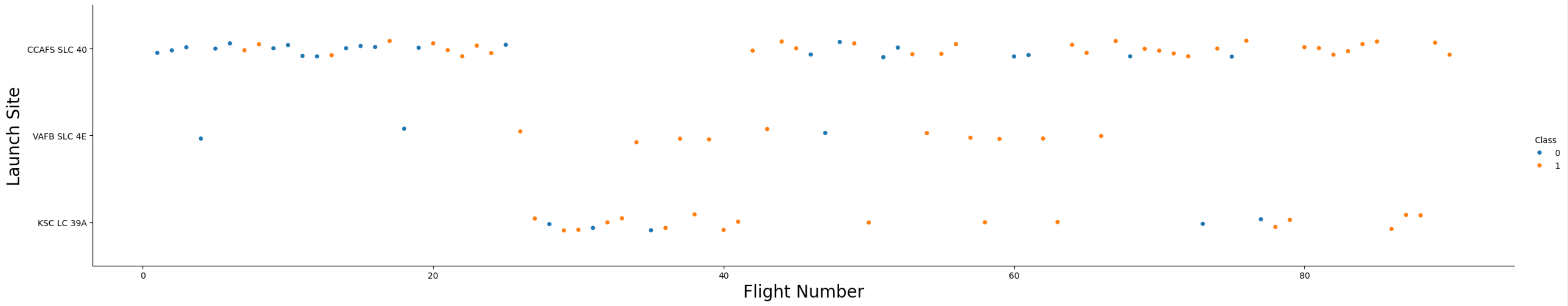
- Scatter plot of Flight Number vs. Launch Site



The success rate of Falcon 9 landings tends to increase significantly over time (as flight number increases).

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site

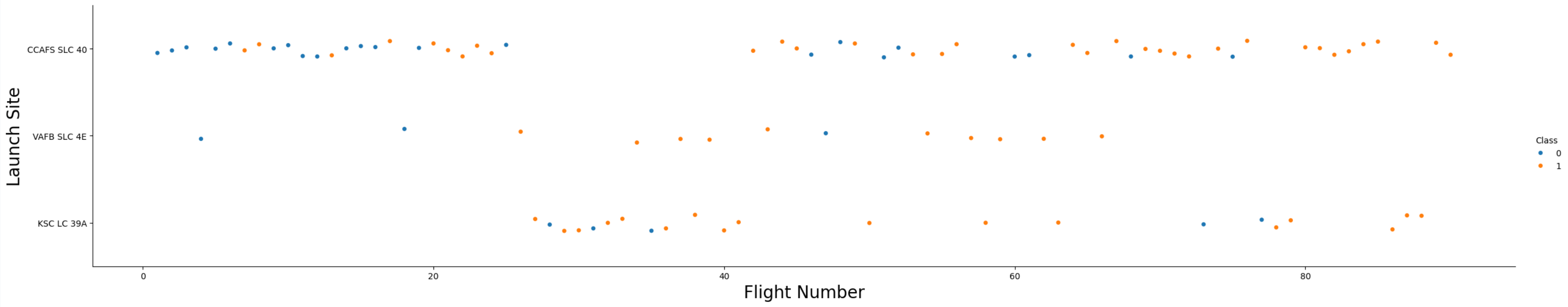


Success rates vary across launch sites:

- CCAFS SLC 40 and KSC LC 39A show clear improvements in landing success over time

Flight Number vs. Launch Site

- Scatter plot of Flight Numbervs. Launch Site

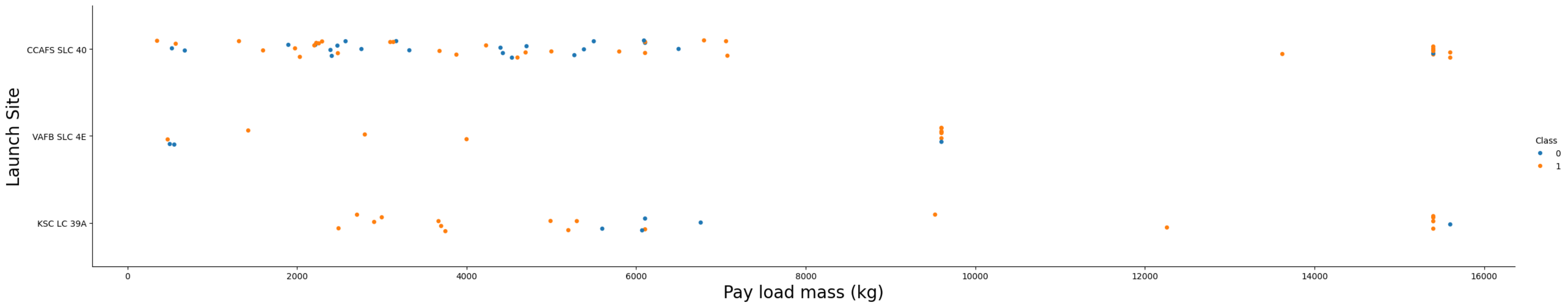


Success rates vary across launch sites:

- VAFB SLC 4E maintains a relatively stable success rate.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

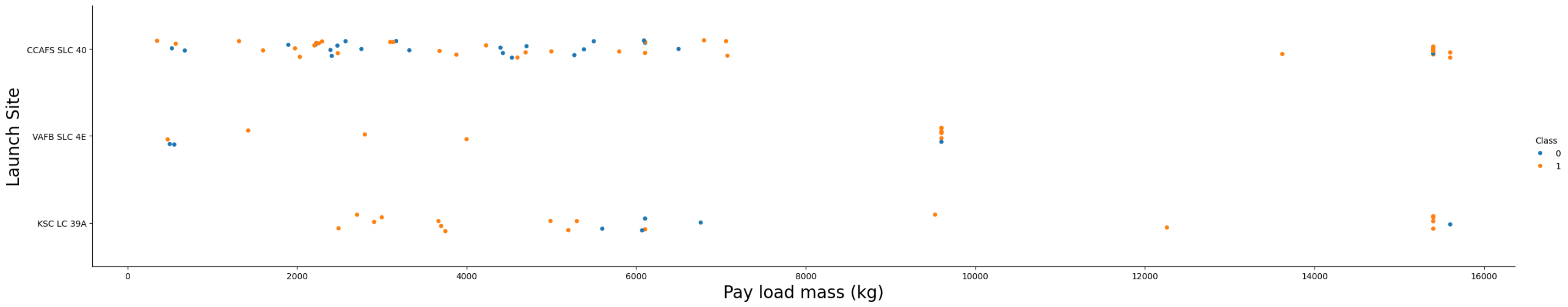


The relationship between payload mass and launch site, categorized by landing outcome (success/failure)

Launch sites tend to handle different types of payloads. VAFB-SLC 4E focuses on lighter payloads (less than 10,000kg)

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site



For heavier payloads, higher landing success rates are observed in Polar, LEO, and ISS orbits (although the chart does not derectly display orbit types, landing outcomes are related to them)

For GTO (an orbit typically associated with heavier payloads), distinguishing between successful and failed landings based on payload mass is more challenging.

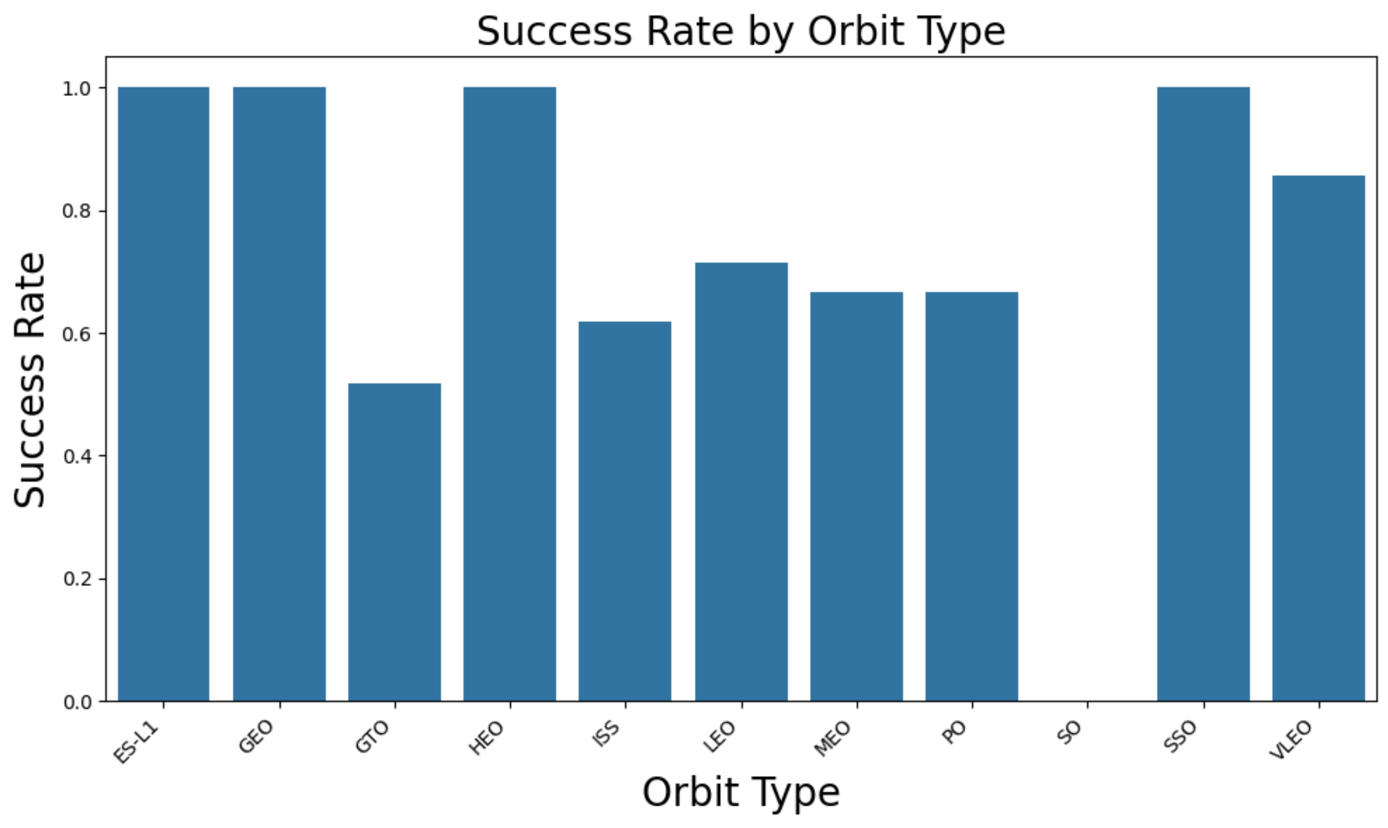
Success Rate vs. Orbit Type

- Show a bar chart for the success rate of each orbit type

The average landing success rate for each orbit type.

Orbits such as ES-L1, GEO, HEO, and SSO have very high success rate (close to 100%)

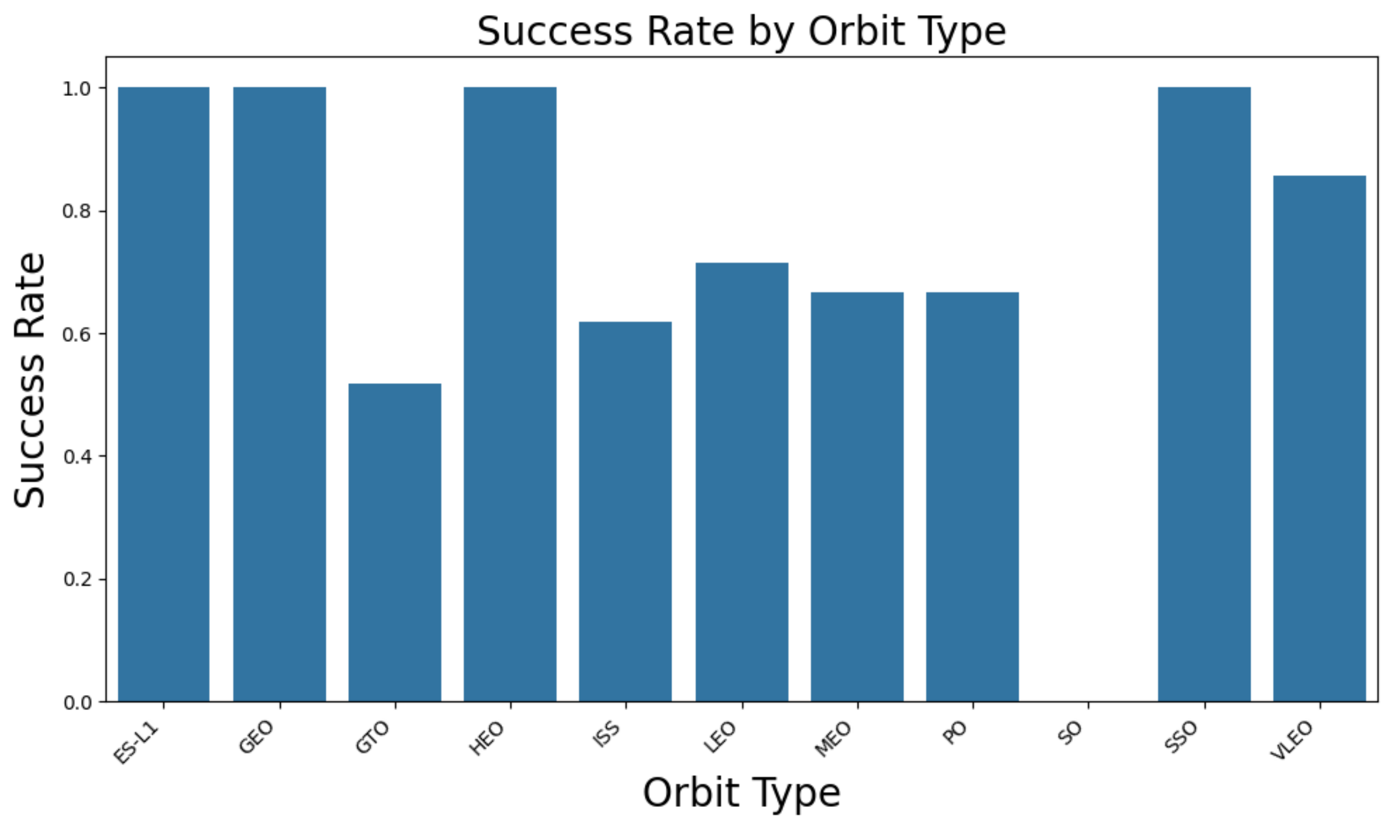
Other orbits like GTO, LEO, and ISS show lower success rates



Success Rate vs. Orbit Type

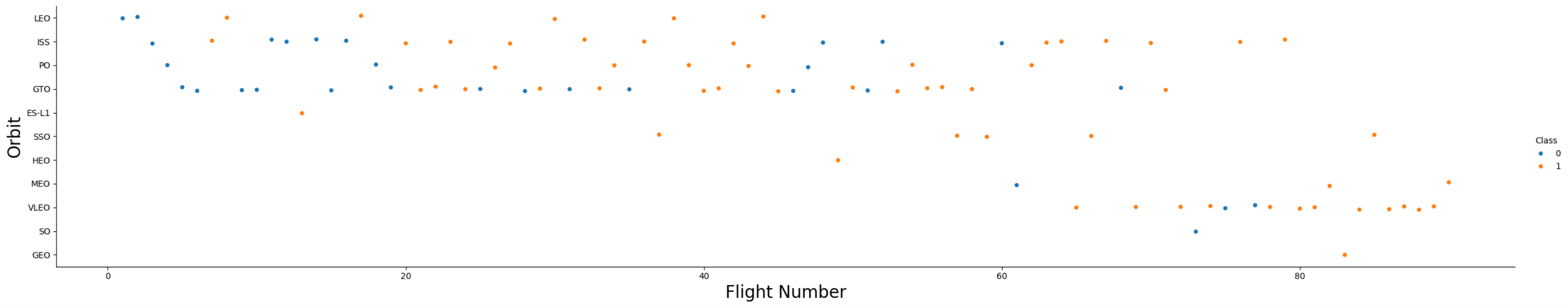
- Show a bar chart for the success rate of each orbit type

The type of orbit is a significant factor influencing the first-stage landing success.



Flight Number vs. Orbit Type

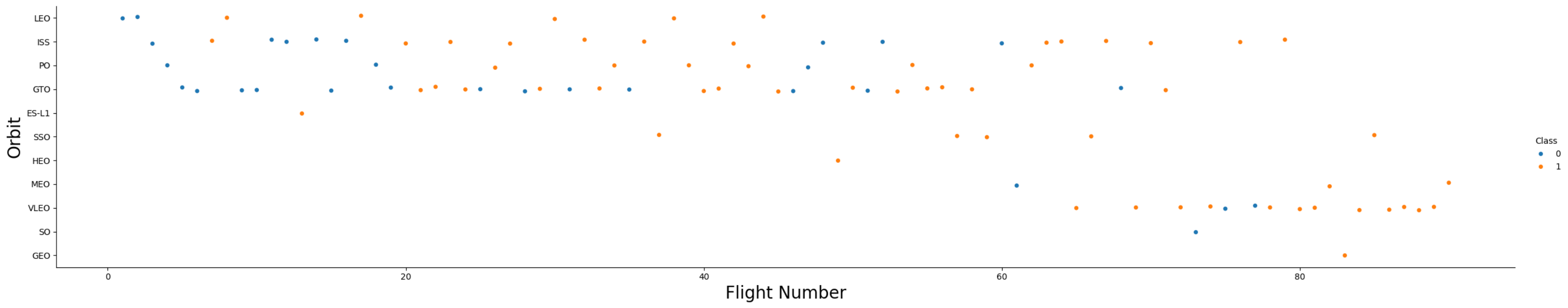
- Show a scatter point of Flight number vs. Orbit type



For LEO and ISS orbits, the success rate tends to increase as the flight number ries, indicating improvement over time.

Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



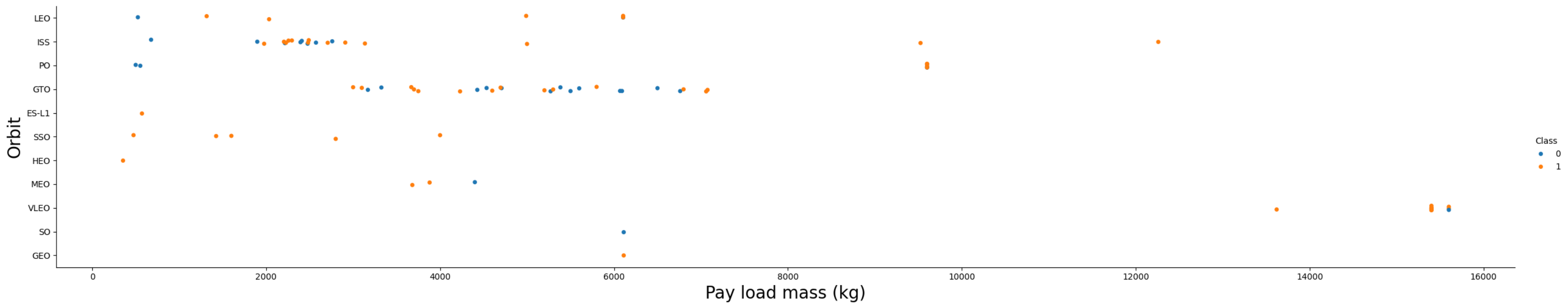
The flight number of each launch by orbit type, categorized by landing outcome.

For GTO orbit, there is no clear relationship between flight number and success rate

Technological advancements and accumulated experience have varying impacts on success rates depending on the orbit type

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



The average landing success rate of Falcon 9 over the years (2010-2020).

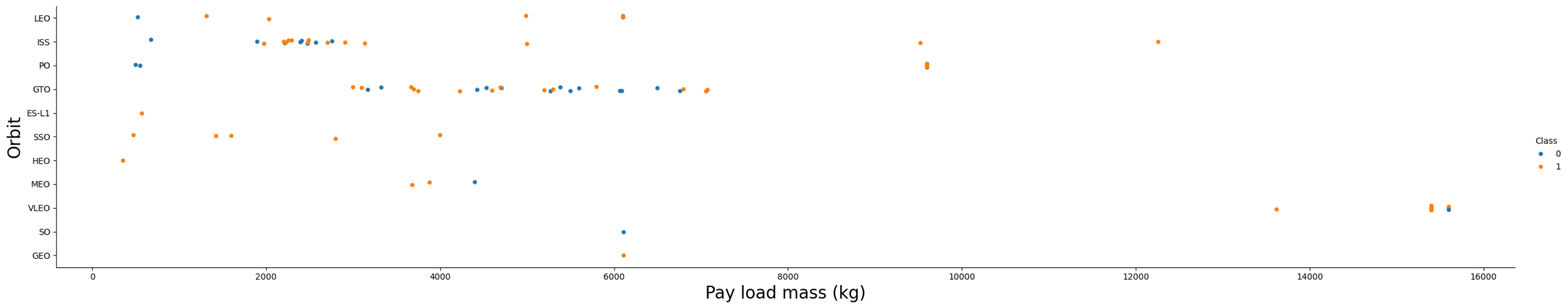
The initial success rate was very low.

A strong and consistent upward trned began around 2013.

A very high success rate was achieved by 2020.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type



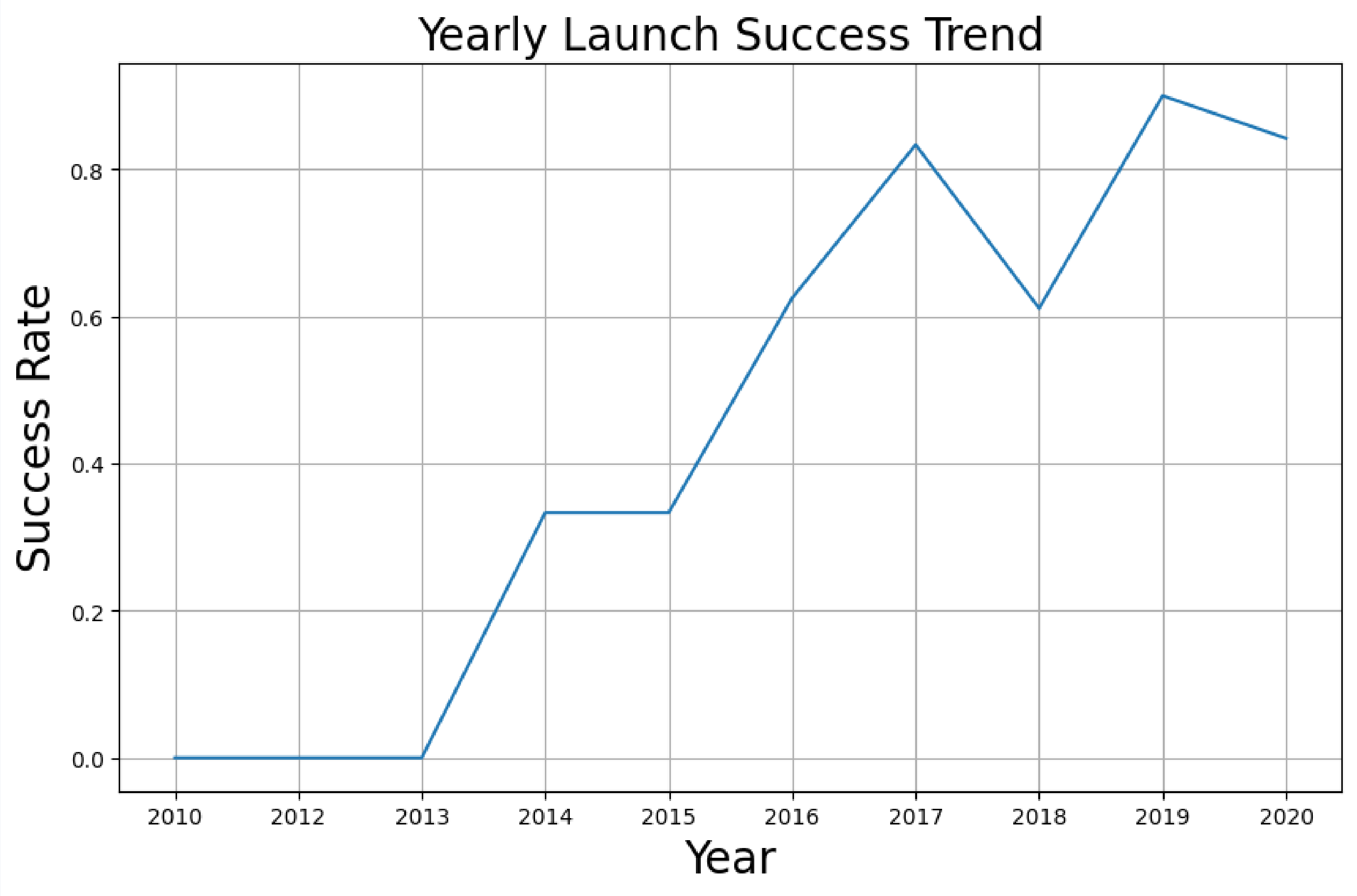
The chart demonstrates SpaceX's significant progress in rocket reusability, with a sharp increase in success over time thanks to technological improvements and accumulated experience.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate

The average landing success rate of Falcon 9 over years (2010-2020)

The initial success rate was very low.

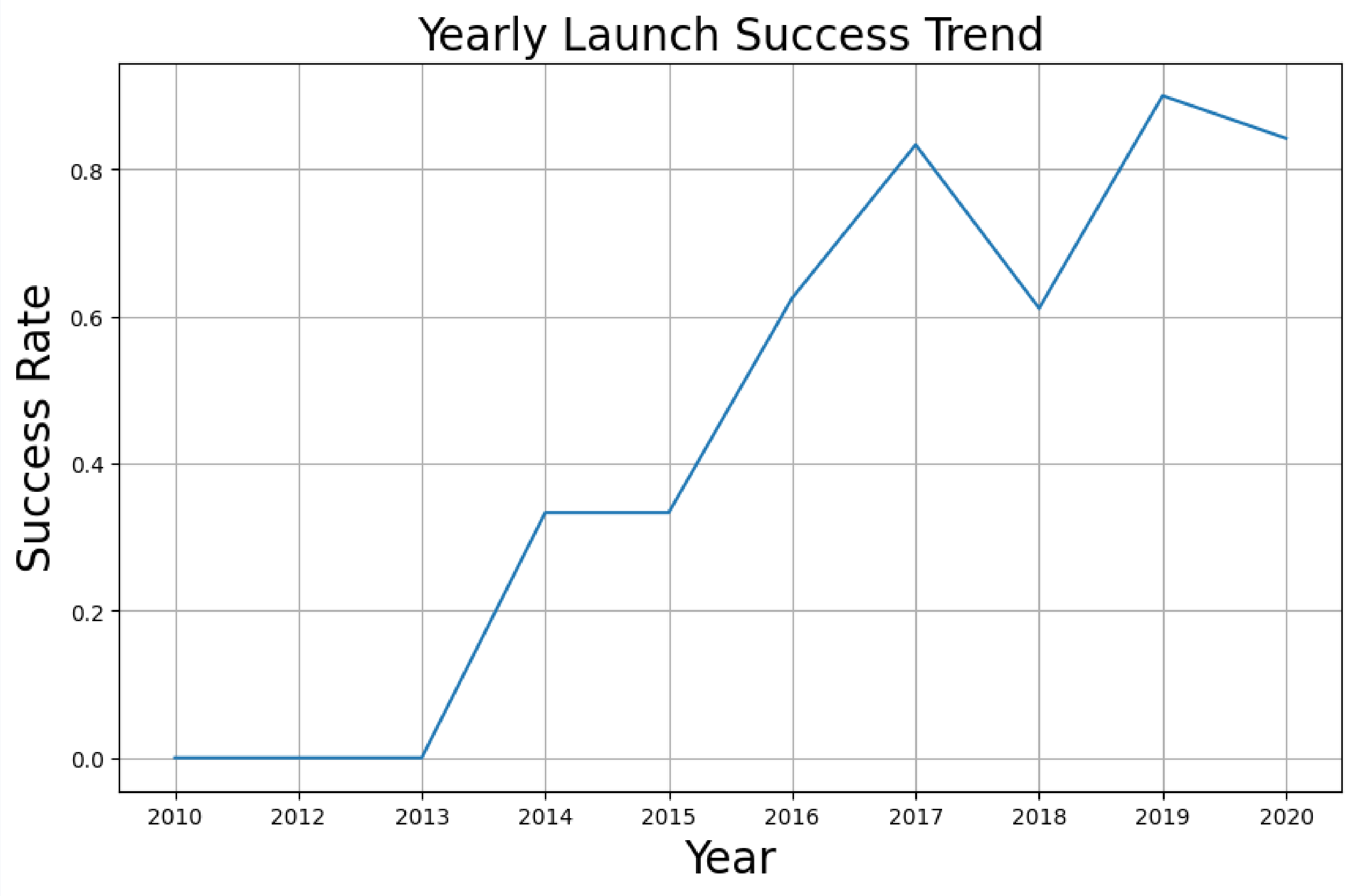


Launch Success Yearly Trend

- Show a line chart of yearly average success rate

A strong and continuus upward trend began around 2013.

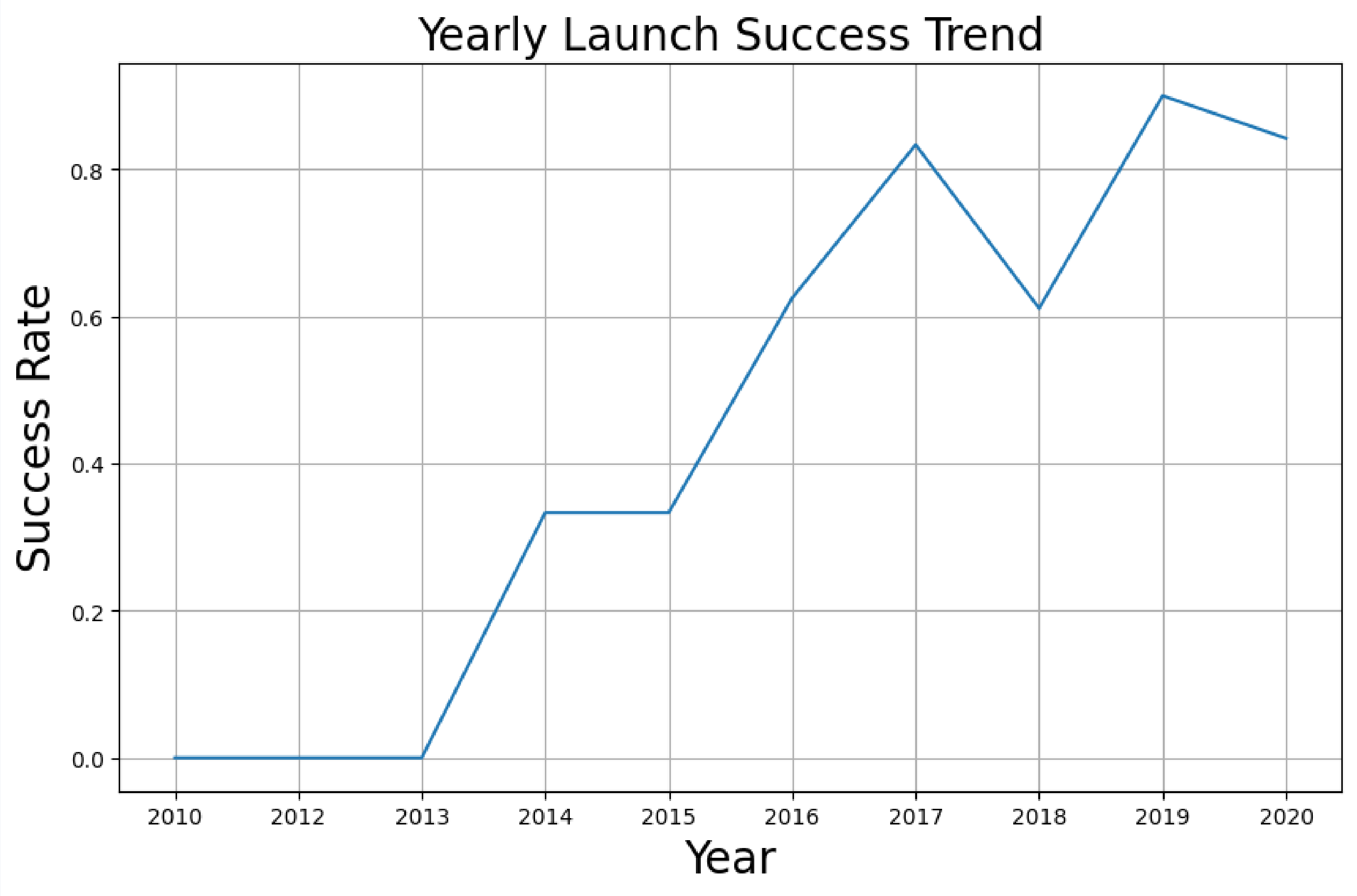
A very high success rate was achieved by 2020.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate

The chart demonstrates SpaceX's significant progress in rocket reusability, with a sharp increase in success rate over time to technological improvement and accumulated experience.



All Launch Site Names

- Unique Launch Sites:

Result:

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Result:

Date	Time (UTC)	Booster_Version	Launch_Site
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

Total Payload Mass

- The total payload carried by boosters from NASA

SUM("PAYLOAD_MASS__KG_")

45596

This number represents the total payload mass, measured in kilograms, carried by all launch vehicles launched by the customer ‘NASA (CRS)’ in the dataset.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1

Result:

AVG("PAYLOAD_MASS__KG_")

2928.4

This number represents the average payload mass (in kilograms) carried by launch vehicles of the ‘F9 v.1.1’ version.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad

Result:

MIN("Date")

2015-12-22

This indicates that the earliest date in the dataset is December 22, 2025.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Result:

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

The result of the query is a list of the following launch vehicle version. There are the rocket versions that successfully landed on a drone ship ("Success (drone ship)") and carried payloads ranging from 4000 kg to 6000 kg.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes

Result:

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

This result indicates that the majority of SpaceX missions in the dataset were successful.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

Result:

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

There are all the launch vehicle versions that carried the highest recorded payload mass in the dataset.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Result:

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

The result of the query shows the unsuccessful drone ship landings (“Failure (drone ship)”) in 2015. The data displays the month, landing outcome, rocket version, and launch site for each failure. Specifically, two failures were recorded: one in January with the rocket version F9 v1.1 B1012 at the launch site CCAFS LC-40, and another in April with the rocket version F9 v1.1 B1015, also at CCAFS LC-40.

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

The result of the query shows the number of occurrences for each landing outcome between June 4, 2010, and March 20, 2017, sorted in descending order

- No attempt: 10 times
- Success (drone ship): 5 times
- Failure (drone ship): 5 times
- Success (ground pad): 3 times

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

The result of the query shows the number of occurrences for each landing outcome between June 4, 2010, and March 20, 2017, sorted in descending order

- Controlled (ocean): 3 times
- Uncontrolled (ocean): 2 times
- Failure (parachute): 2 times
- Precluded (drone ship): 1 time

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

This result indicates that during this period, “No attempt” was the most common landing outcome, followed by “Success (drone ship)” and “Failure (drone ship)”, which occurred an equal number of times.

Section 3

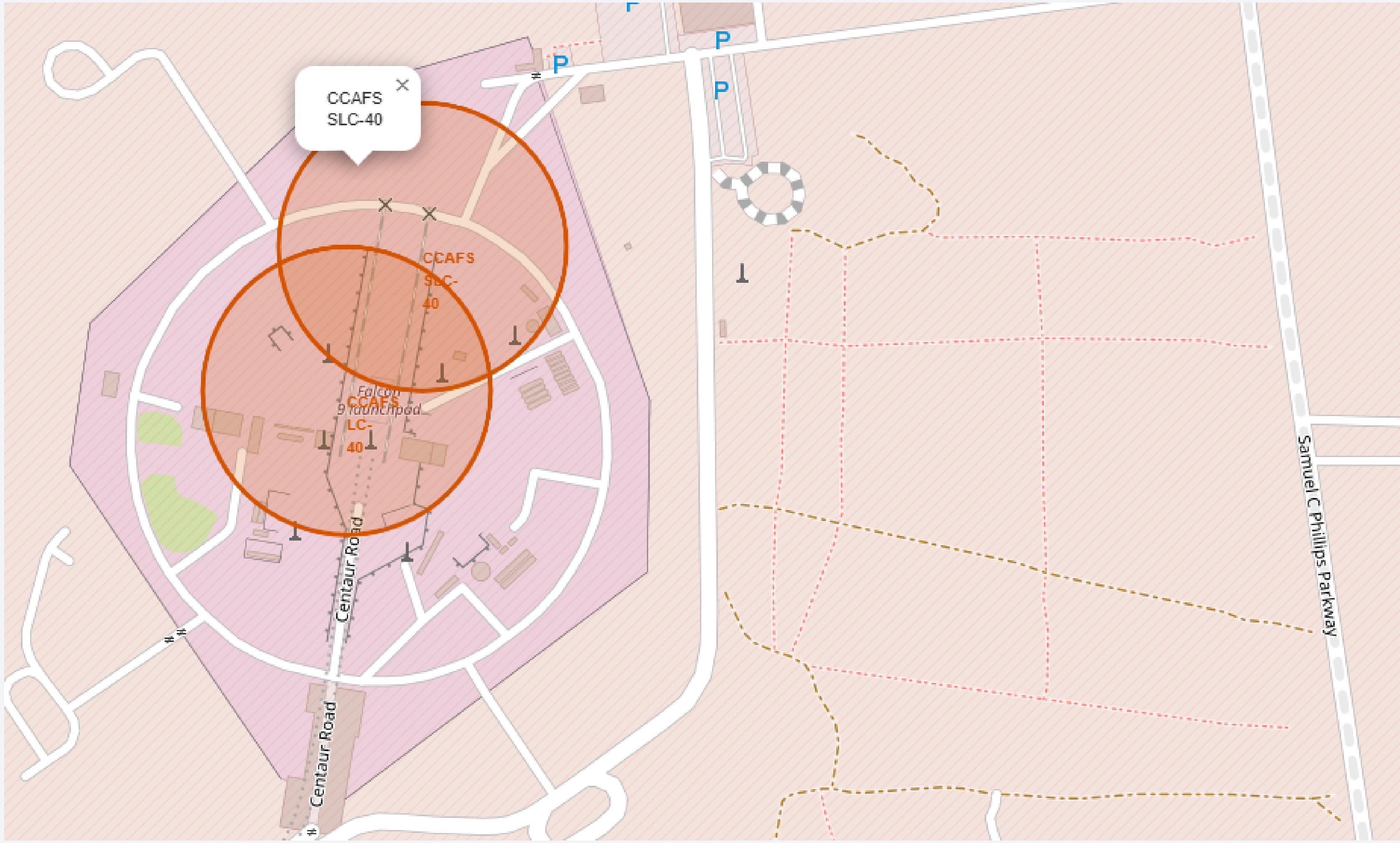
Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites

SpaceX landing outcomes are marked on the map

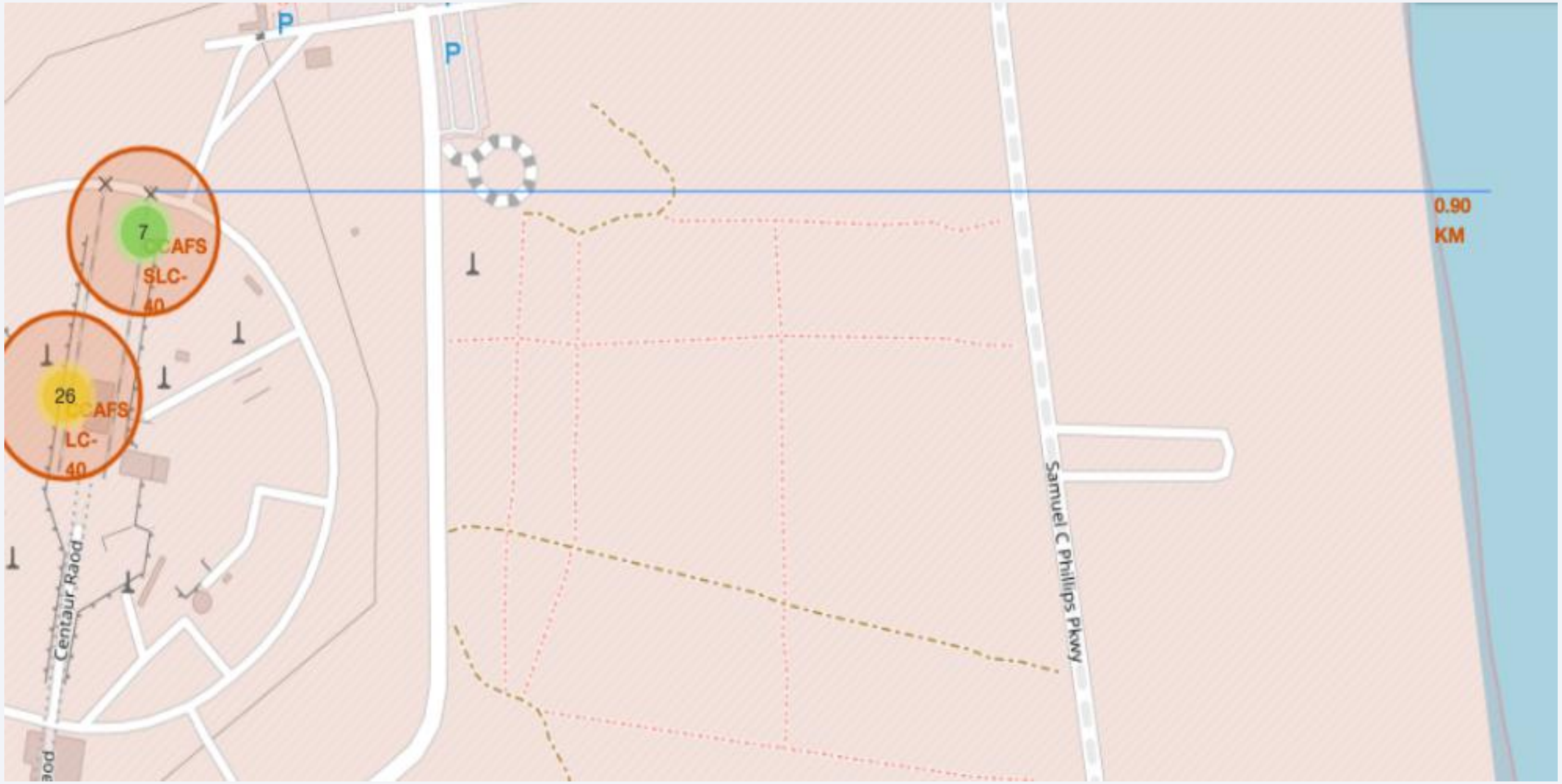
SpaceX landing outcomes are marked on the map

SpaceX landing outcomes are marked on the map



SpaceX landing outcomes are marked on the map

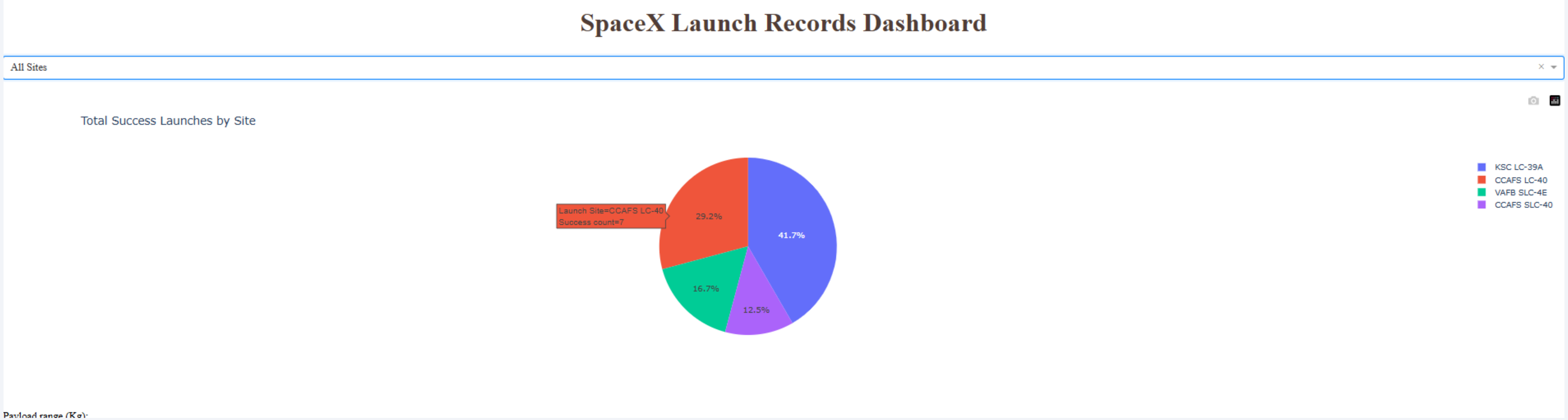
Distance from the launch site to nearby locations



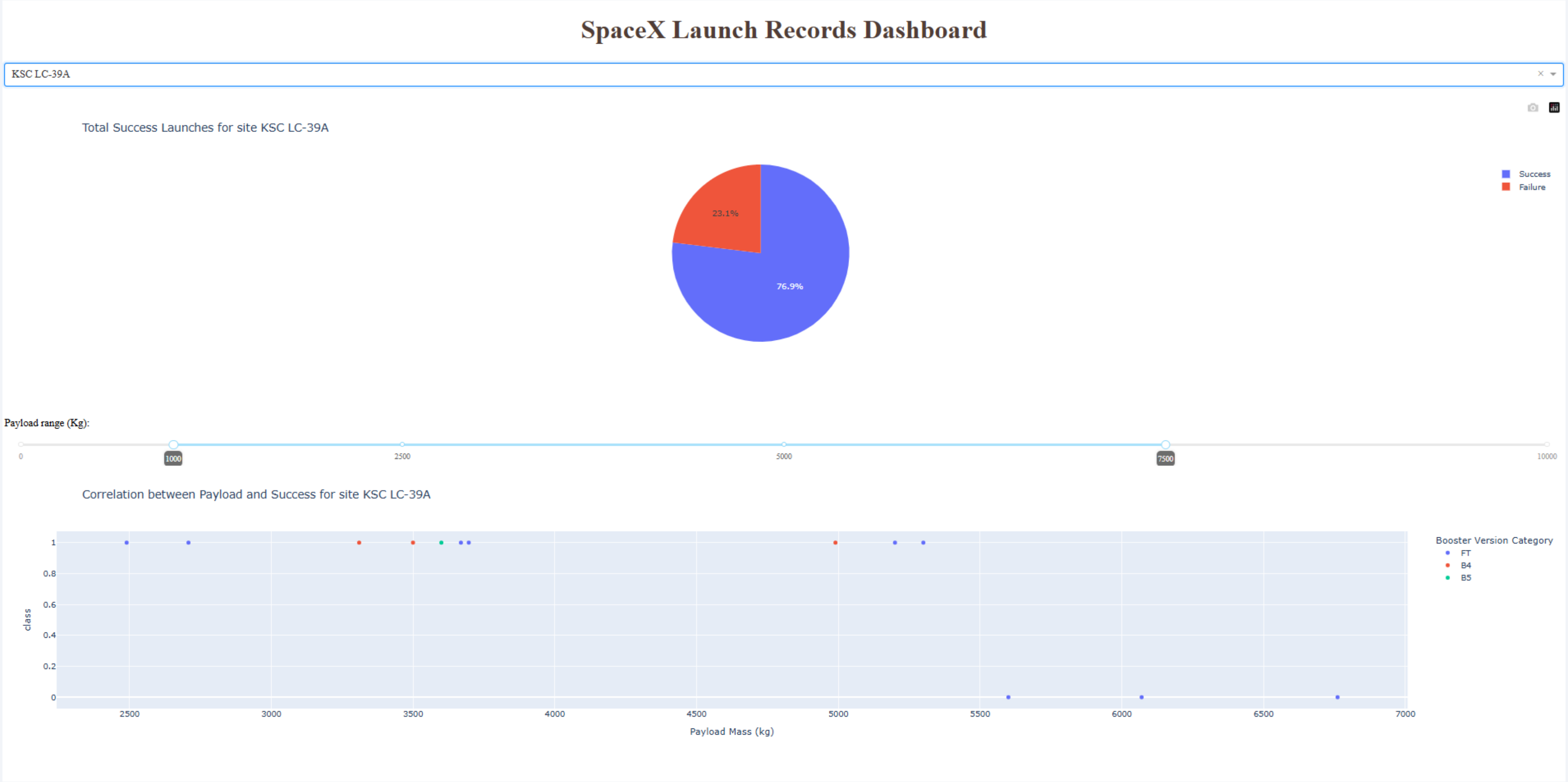
Section 4

Build a Dashboard with Plotly Dash

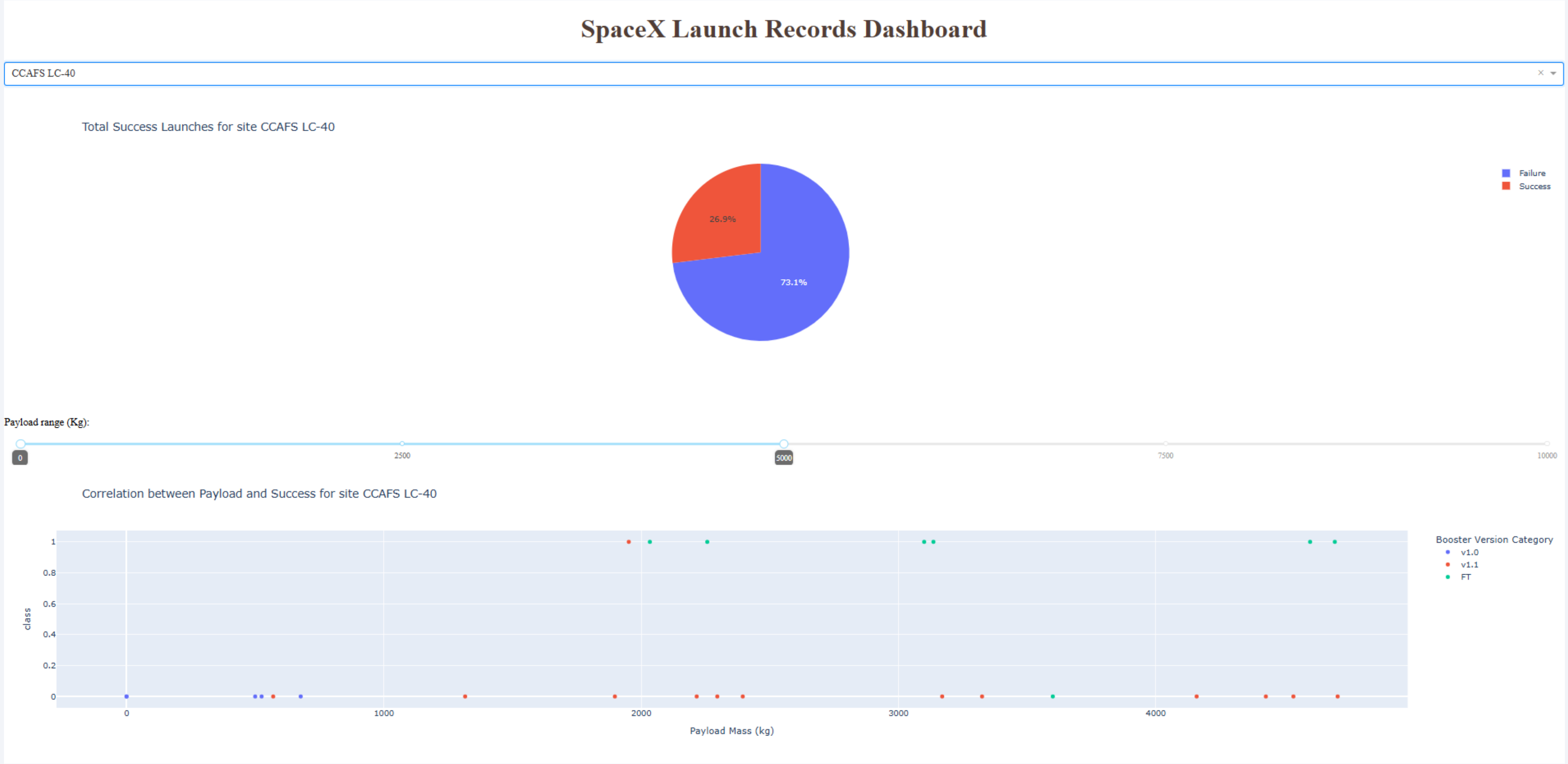
Pie Chart of Launch Success Counts by Site



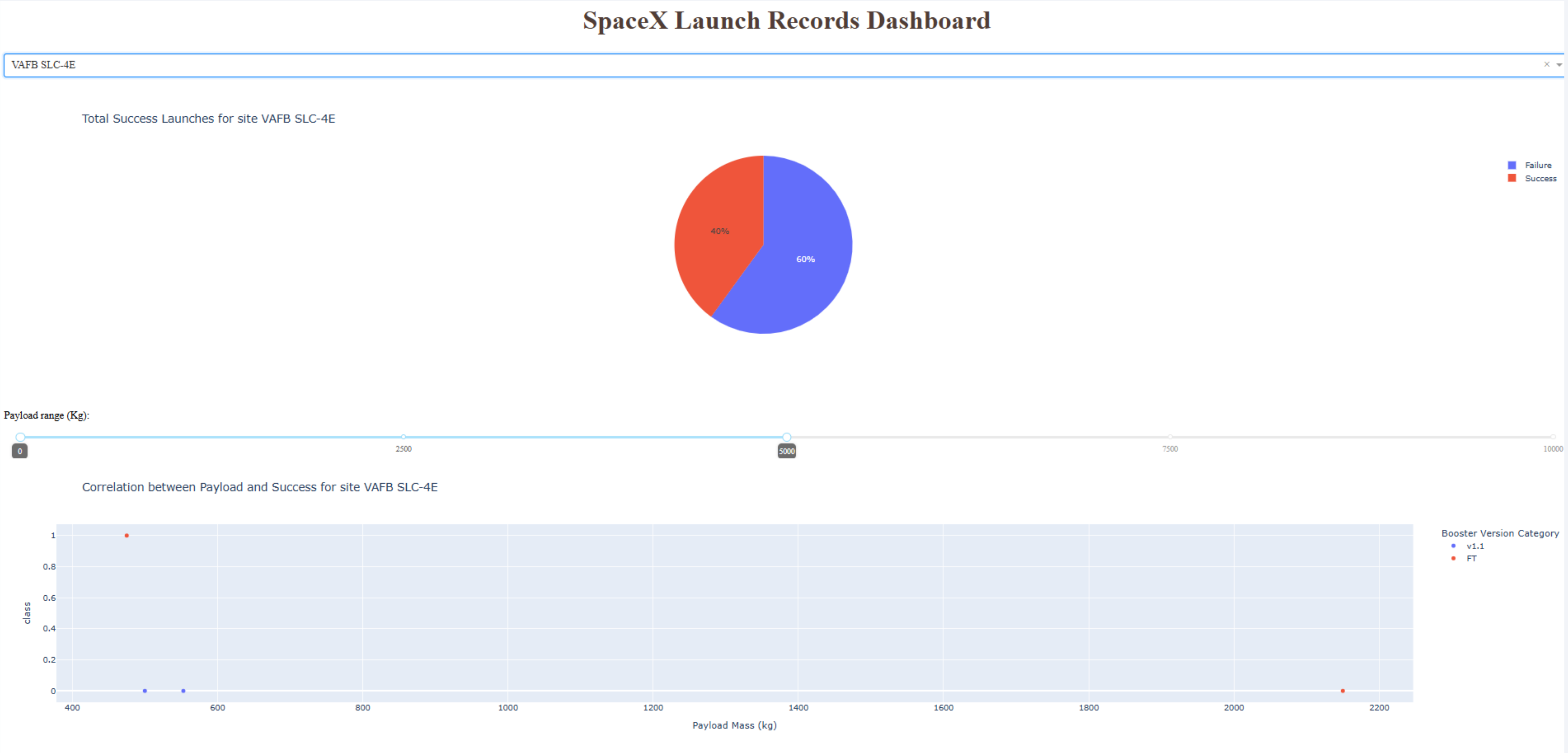
Top-performing Launch Site – Pie chart



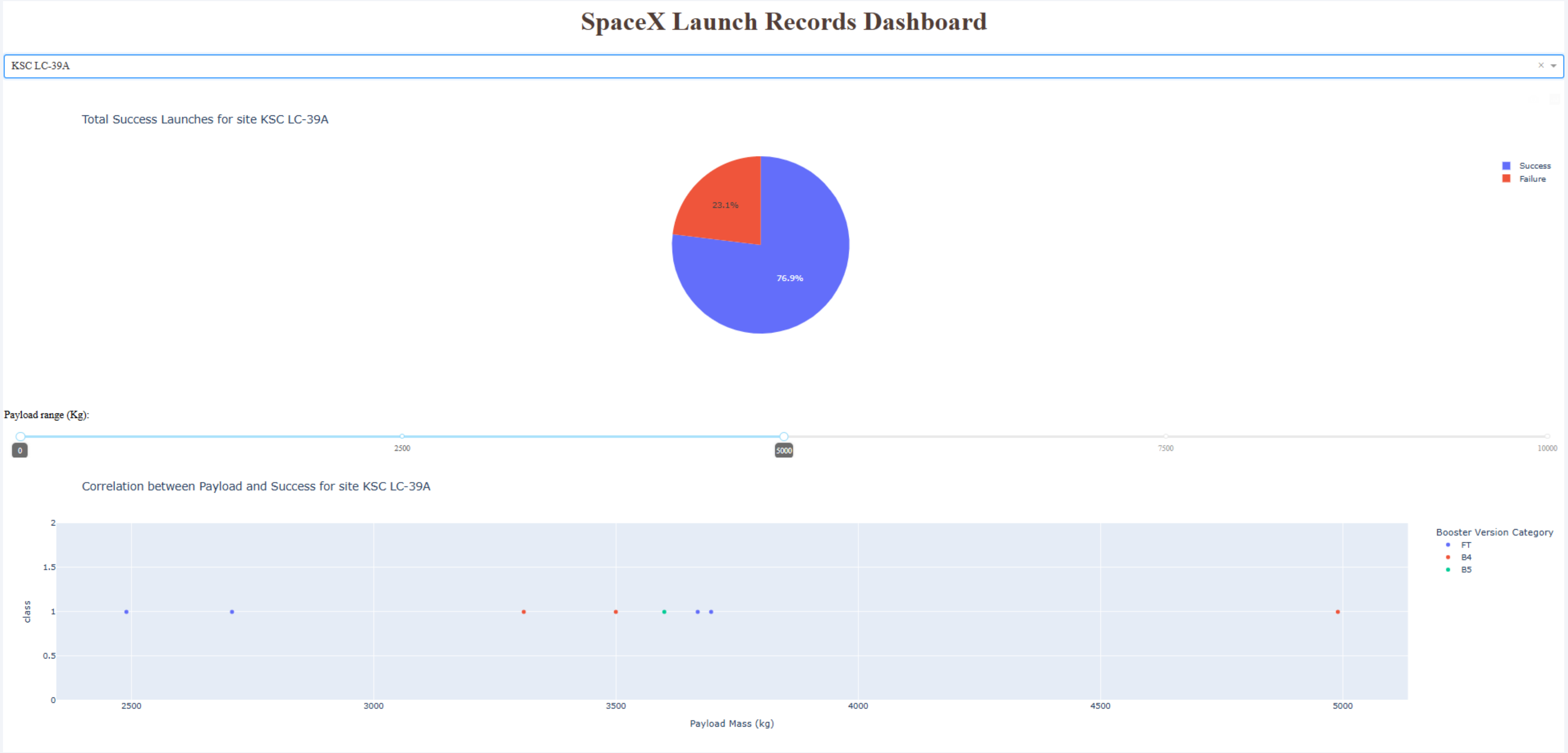
Scatter plot of payload versus launch outcome for CCSFS LC-40



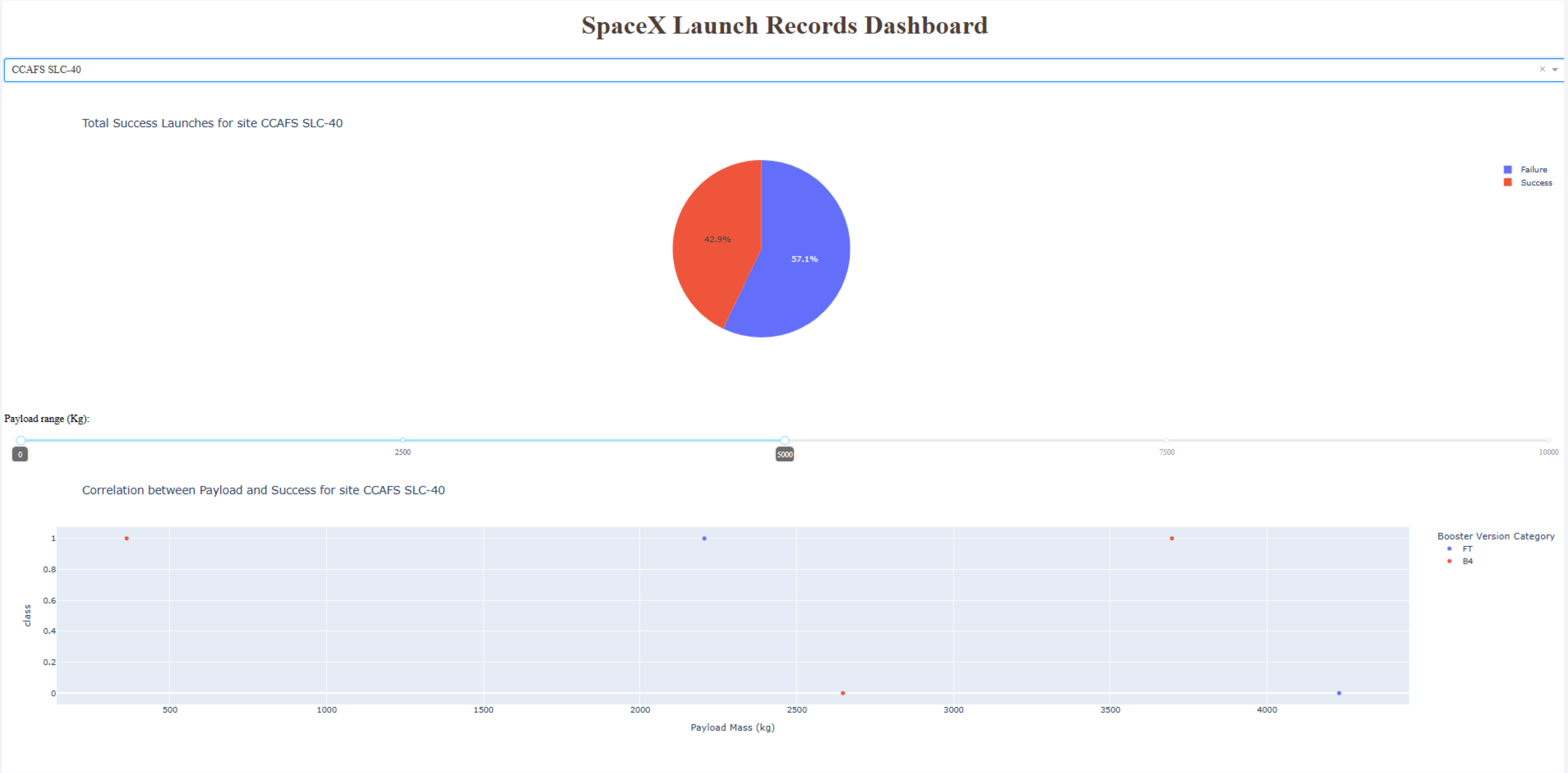
Scatter plot of payload versus launch outcome for VAFB SLC-4E



Scatter plot of payload versus launch outcome for KSCLC-39A



Scatter plot of payload versus launch outcome for CCAFS SLC-40

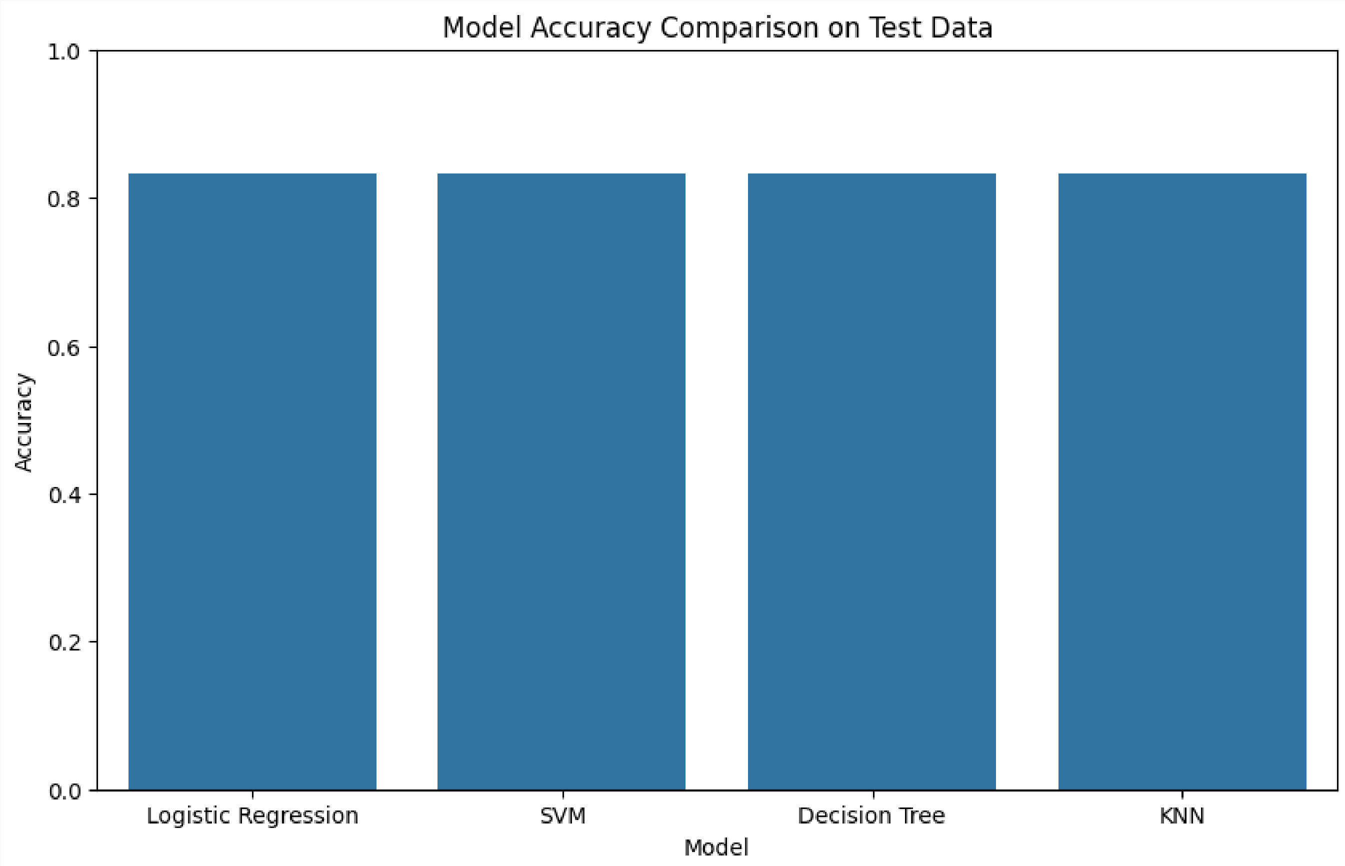


Section 5

Predictive Analysis (Classification)

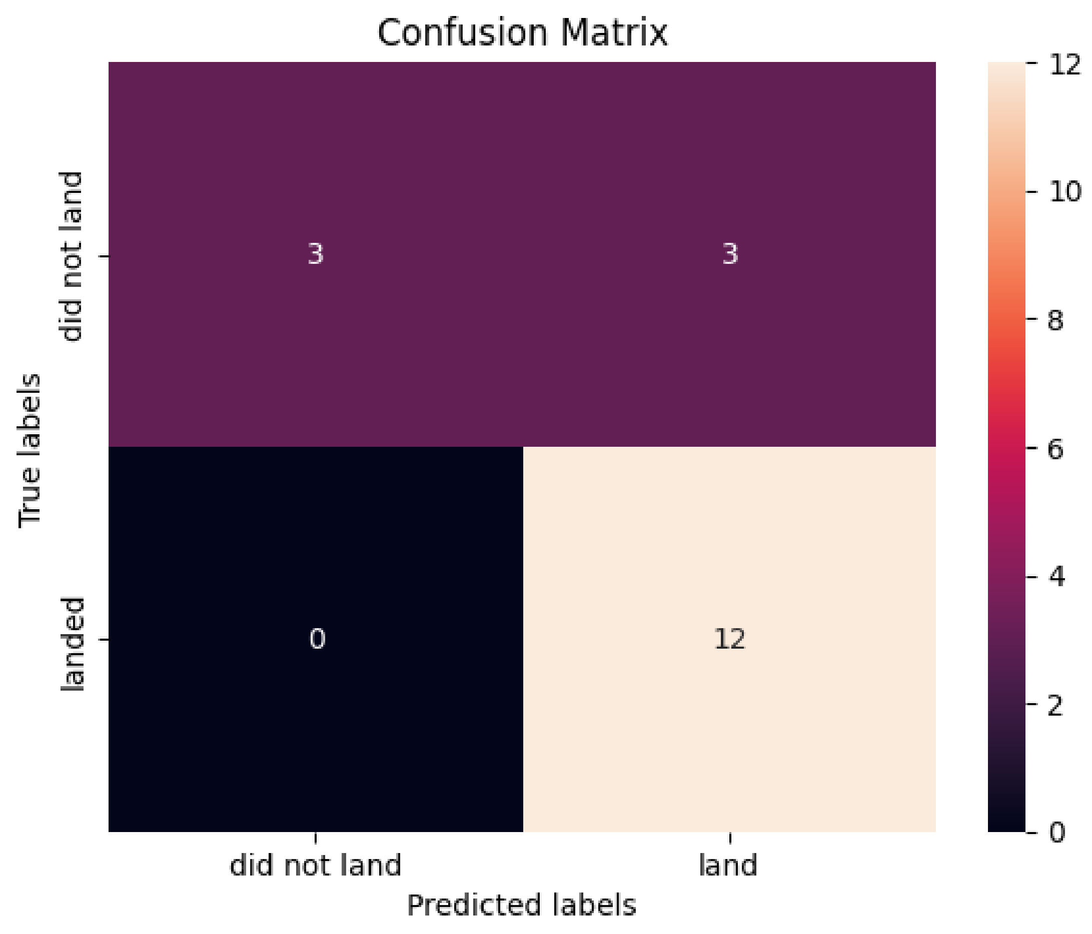
Classification Accuracy

The model have equal accuracy (~83.33%), however, the Decision Tree model is more intuitive and easier to interpret.



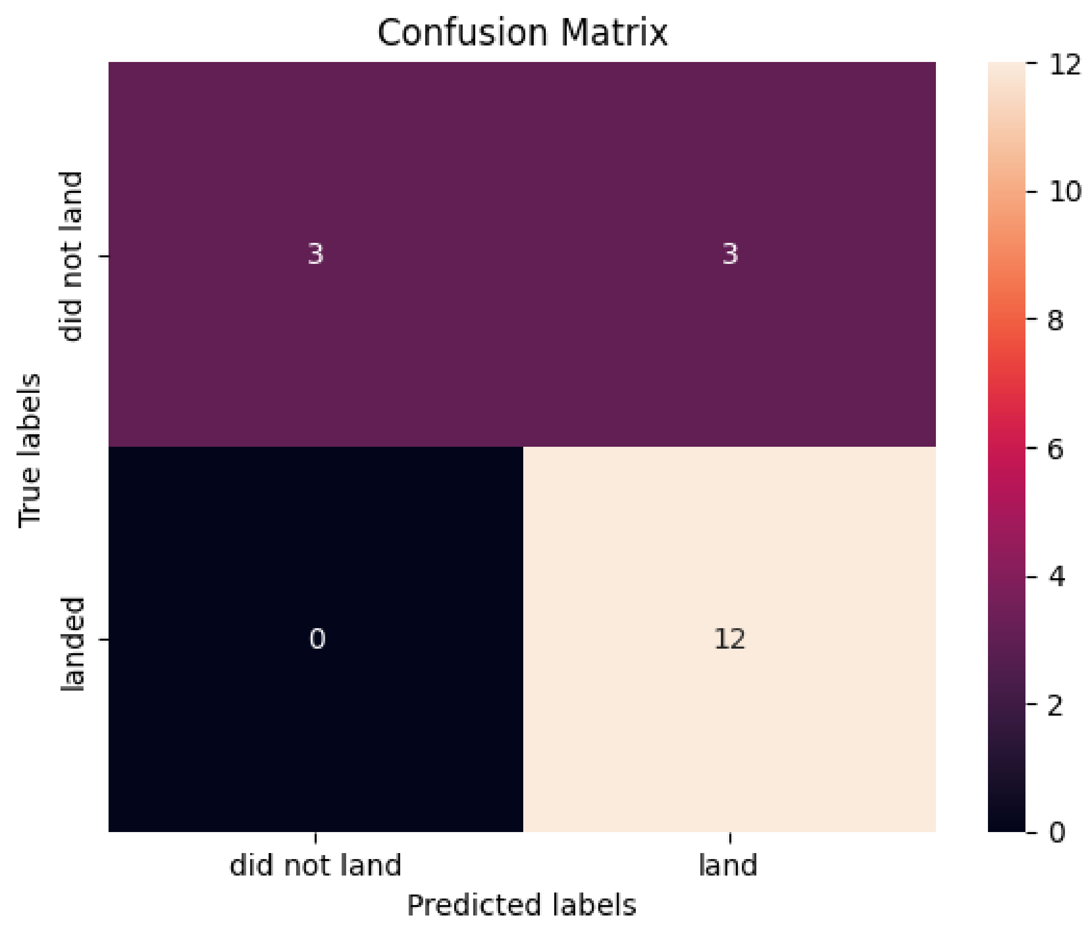
Confusion Matrix

- True Positive (TP=3): The model correctly predicted 3 times that the rocket did not land.
- False Positive (FP=3): The model incorrectly predicted 3 times that the rocket landed, while in reality it did not.



Confusion Matrix

- False Negative (FN=0): There were no instances where the model incorrectly predicted that the rocket did not land when it actually did.
- True Negative (TN=12): The model correctly predicted 12 times that the rocket landed.



Conclusions

- Launch success rates vary across launch sites
- Payload influences the probability of success
- Booster type is an important factor
- The prediction model achieves promising accuracy
- The model's primary error is false positives

Conclusions

- Optimizing launch planning based on effective sites and boosters: Prioritize launch sites and booster versions with a history of high success rates to reduce the risk of recovery costs or damage.
- Payload management: Keeping the payload within the "safe" range can increase the probability of a successful landing.
- A model with ~83% accuracy is sufficiently reliable as a decision-support tool, helping to flag missions with a high probability of failure.

Appendix

- The following link contains the complete source code of this project:

https://github.com/thienthuan25/Applies_Data_Science_Capstone_Course

Thank you!