IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Albert Henry Ntarmah
July 27, 2024

Outline

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

Executive Summary

- Methodologies
  - In this project, I empolyed two main approaches to data collection: utilized the SpaceX API and web scraping Wikipedia for Falcon 9 launch records, followed by comprehensive data wrangling and exploratory data analysis (EDA) using SQL and interactive visual analytics with Folium and Plotly Dash.
  - For predictive analysis, I standardized the data, split it into train and test sets, optimized hyperparameters, and evaluated model accuracy.
- Results
  - The EDA results revealed that SpaceX's success rate increased steadily from 2013 to 2020. Notable milestones include the first successful ground pad landing on December 22, 2015, and the total payload carried by NASA boosters being 48,213 kg. The average payload for the F9 v1.1 booster is 2928.4 kg. Between June 2010 and March 2017, there were 5 failed drone ship landings compared to 3 successful ground pad landings. In 2015, two Falcon 9 missions from CCAFS LC-40 failed to land on a drone ship.
  - According to the plotly dash results, the KSC LC-39A launch site has the highest success rate at 76.9%, while CCAFS SLC-40 has the lowest at 12.5%. Among booster versions, F9 Booster version B4 stands out with the highest success rate and the ability to carry the largest payload, up to 9600 kg.
  - The predictive analysis results revealed that the models (Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors) show consistent accuracy (83%) on the test data and are effective at distinguishing between different classes (success or failure).

Introduction

- SpaceX has gained worldwide attention for a series of historic milestones. It is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars wheras other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- This project aims to predict the successful landing of the Falcon 9 first stage, a critical factor in SpaceX's cost-effective launch strategy. Specifically, this project uses data science metholodogies and models to answer key questions:
  - How do payload mass, launch site, number of flights, and orbits affect the success of first-stage landings?
  - What is the rate of successful landings over time?
  - Which predictive model is most effective for forecasting successful landings?
- By addressing these questions, the project aims to enhance the understanding of SpaceX's launch dynamics and can make more informed bids against SpaceX for a rocket launch. to contribute to more efficient and reliable space missions.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - I imported the necessary libraries in Python, defined auxiliary functions, performed a GET request on the SpaceX API, and converted the response into a DataFrame.
  - I web scraped Falcon 9 launch records using BeautifulSoup by extracting the HTML table from Wikipedia, parsing it, and converting it into a Pandas DataFrame.
- Perform data wrangling
  - Handled missing values in the dataset, replacing null PayloadMass entries with the mean value.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - I standardized the data, split it into train and test sets, optimized hyperparameters for each model, and calculated test data accuracy.

Data Collection

- In this section, I will describe how datasets were collected, focusing on the SpaceX API and web scraping approaches.

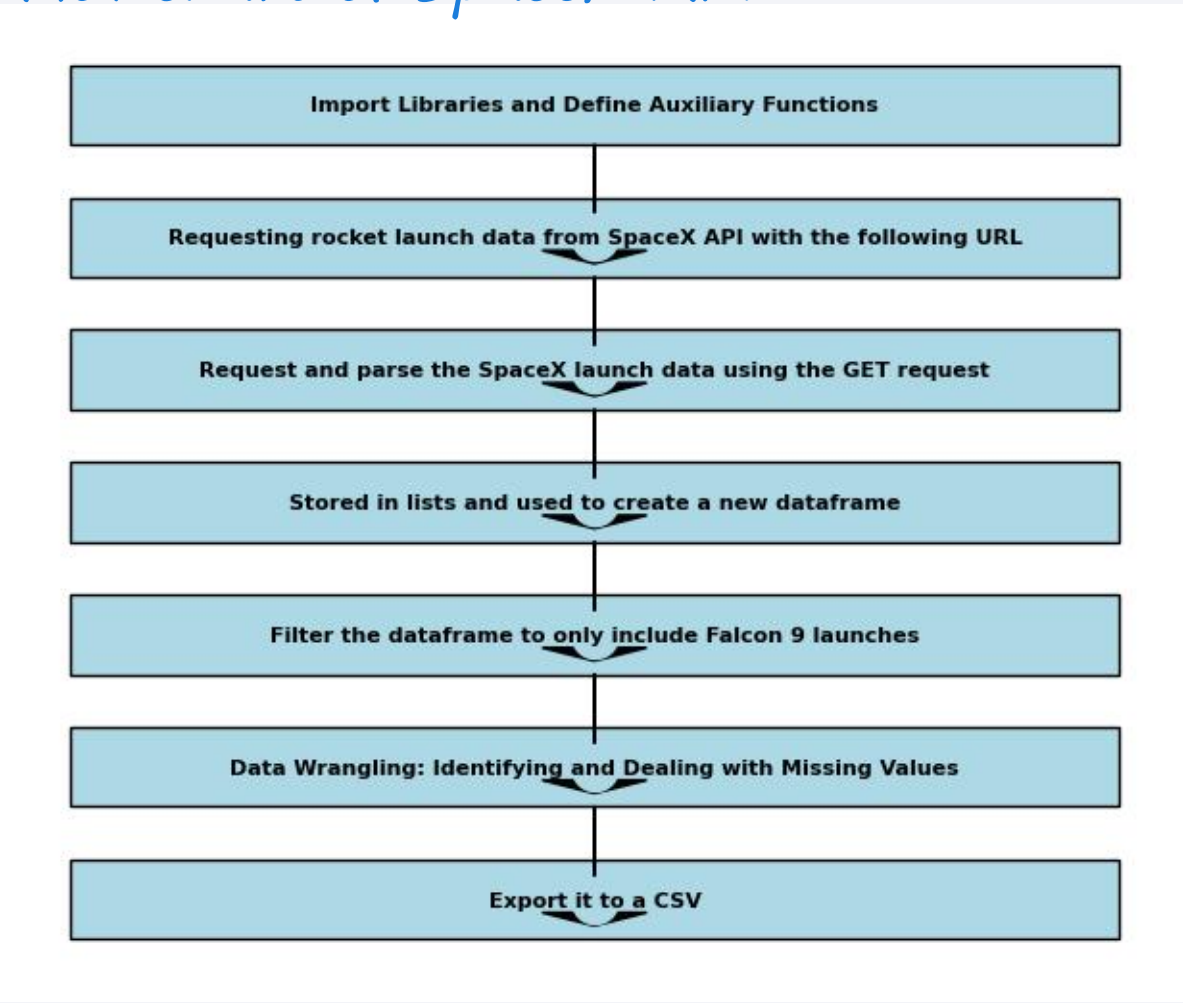
- For each approach, I will present the data collection process using key phrases and flowcharts.

Data Collection – SpaceX API

- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/jupyter-labs-spacex-data-collection-api%20(1).ipynb

- See the GitHub URL for the codes

Flowchart of SpaceX API

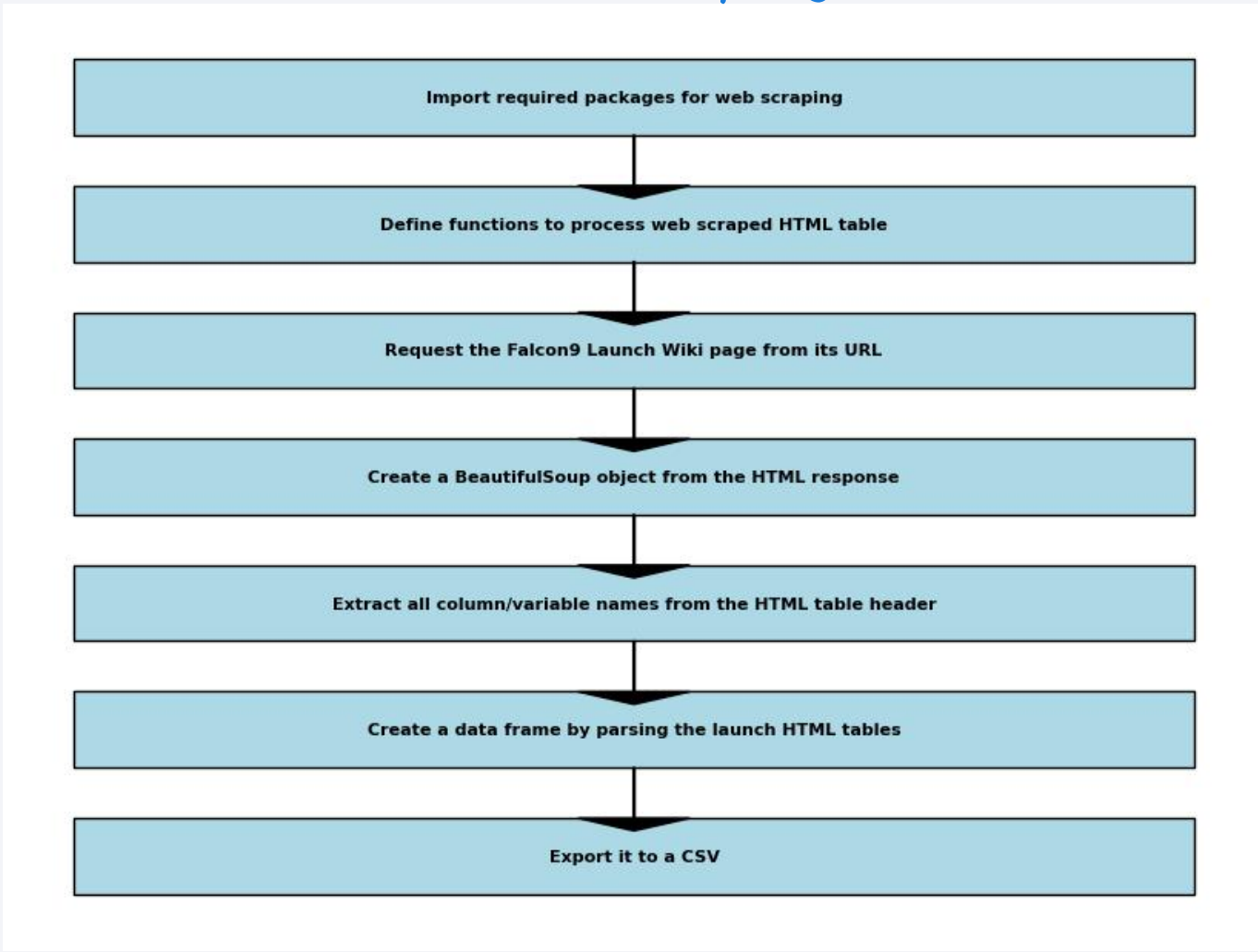


Data Collection - Scraping

- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/jupyter-labs-webscraping.ipynb

- See the GitHub URL for the codes

Flowchart of web scraping

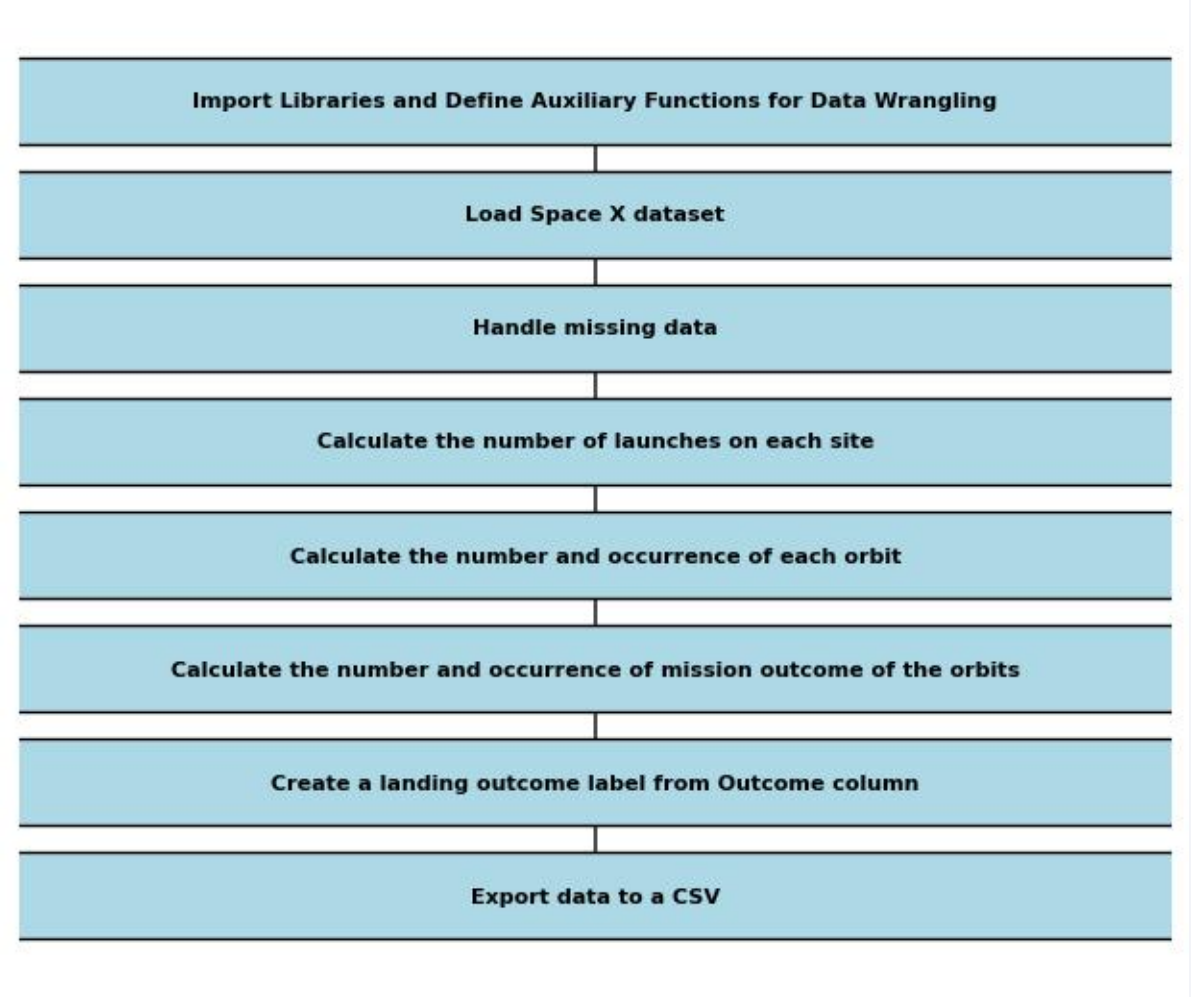


Data Wrangling

GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

See the GitHub URL for the codes

Data Wrangling

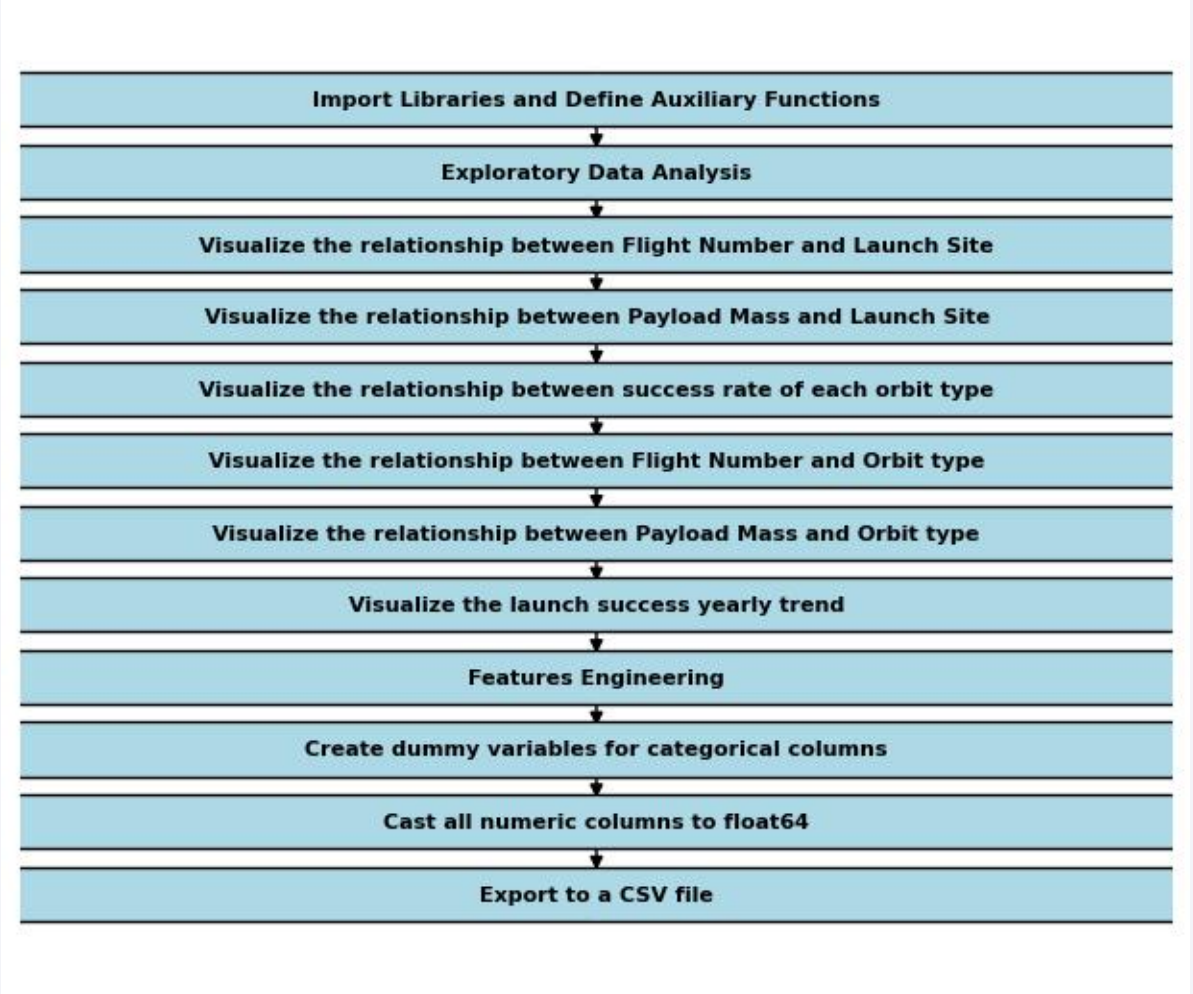


EDA with Data Visualization

- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/edadataviz.ipynb

- See the GitHub URL for the codes

Flow Chart of EDA with Data Visualization



EDA with SQL

The SQL queries performed

- Download the datasets
- Connect to the database
- Remove blank rows from table
- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- The folium map includes circles to highlight the locations of each SpaceX launch site and markers to label them, providing clear visual and textual identification. Additionally, colored markers indicate launch outcomes, marker clusters group overlapping markers for better readability, and poly lines illustrate distances to nearby geographical features, enhancing the map's analytical capabilities.
- These map objects collectively create an interactive and informative visualization that allows for geographical analysis of SpaceX launch sites, their performance, and their proximity to various points of interest.
- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/lab_jupyter_launch_site_location%20(6).ipynb
- Note: See the last end of the folium lab link for the links to all the maps generated.

Build a Dashboard with Plotly Dash

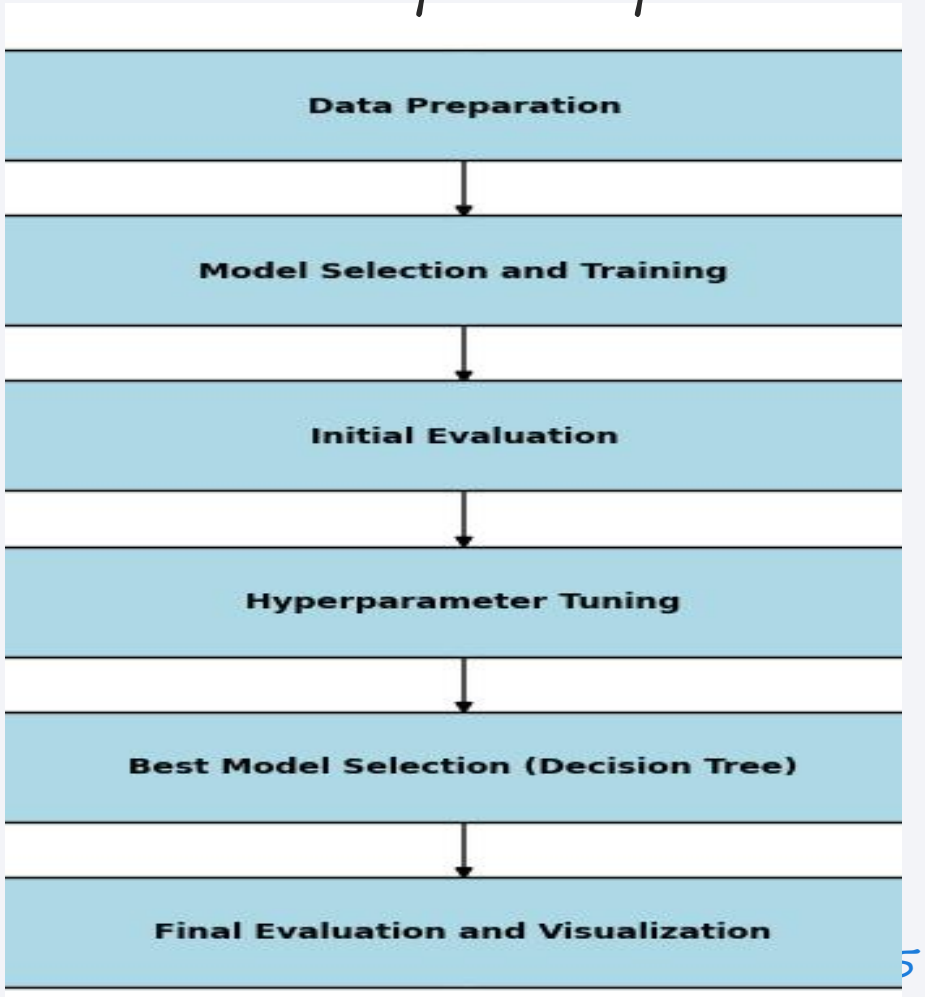
- The dashboard features a pie chart (success pie chart) that visualizes launch success rates and a scatter plot (success-payload scatter plot) that explores the relationship between payload mass and launch outcomes, both of which are interactive and respond to user selections from a dropdown and a range slider.
- These plots were added to provide users with insights into launch performance across different sites and payloads, facilitating the analysis of success rates and potential correlations with payload mass.

- Add the GitHub URL: Codes:
  https://github.com/AlbertNtarmah/testrepo/blob/main/spacex_dash_app%20(1).py
- URL for plots:
  https://github.com/AlbertNtarmah/testrepo/blob/main/PLOTLY%20DASH%20CHARTS%20AND%20PLOTS.pdf

Predictive Analysis (Classification)

- The model development process involved data preparation, model selection and training, initial evaluation, hyperparameter tuning, selection of the best model (Decision Tree), and final evaluation with visualization.
- I implemented four classification algorithms: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN)
- GitHub URL: https://github.com/AlbertNtarmah/testrepo/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Model development process



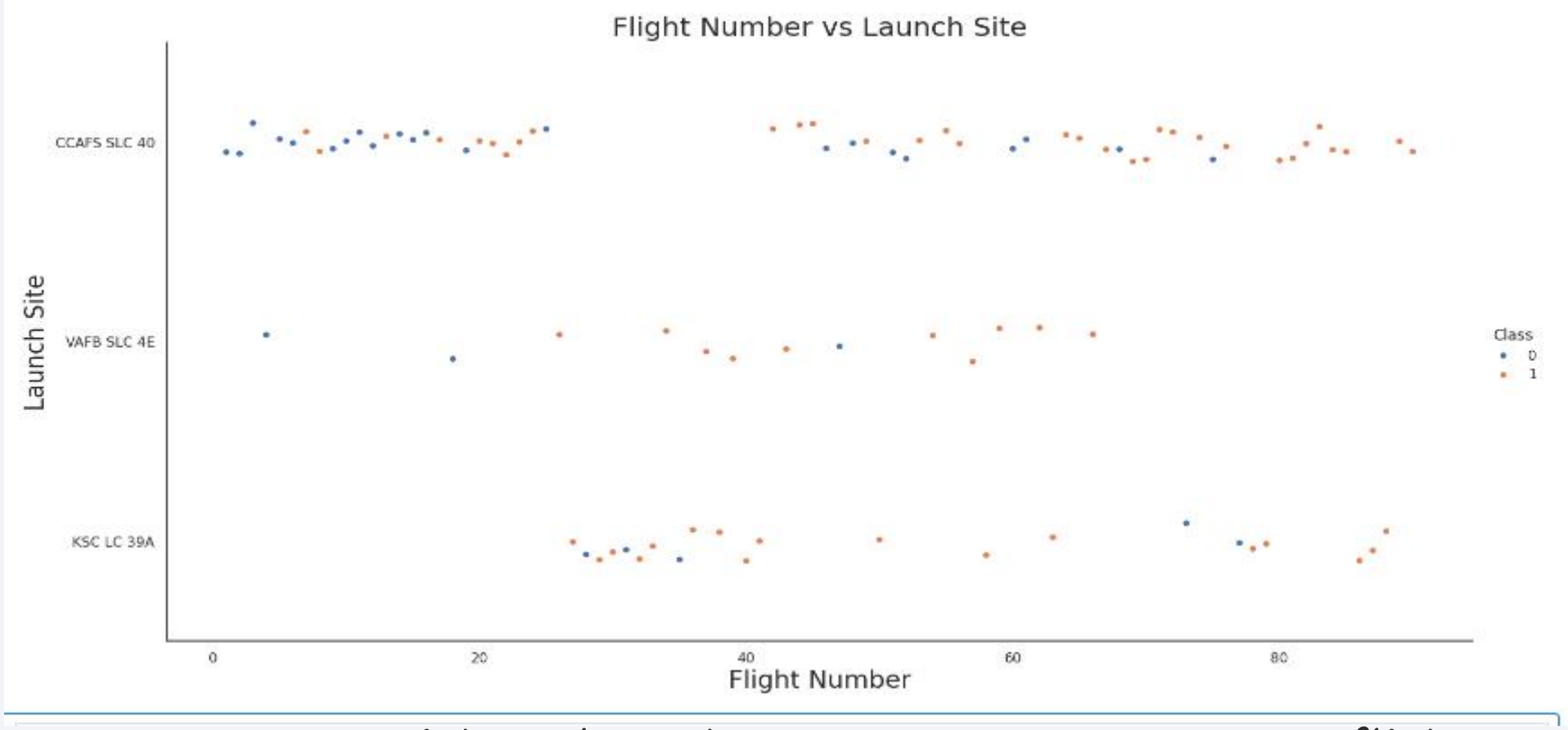
Results

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

Section 2

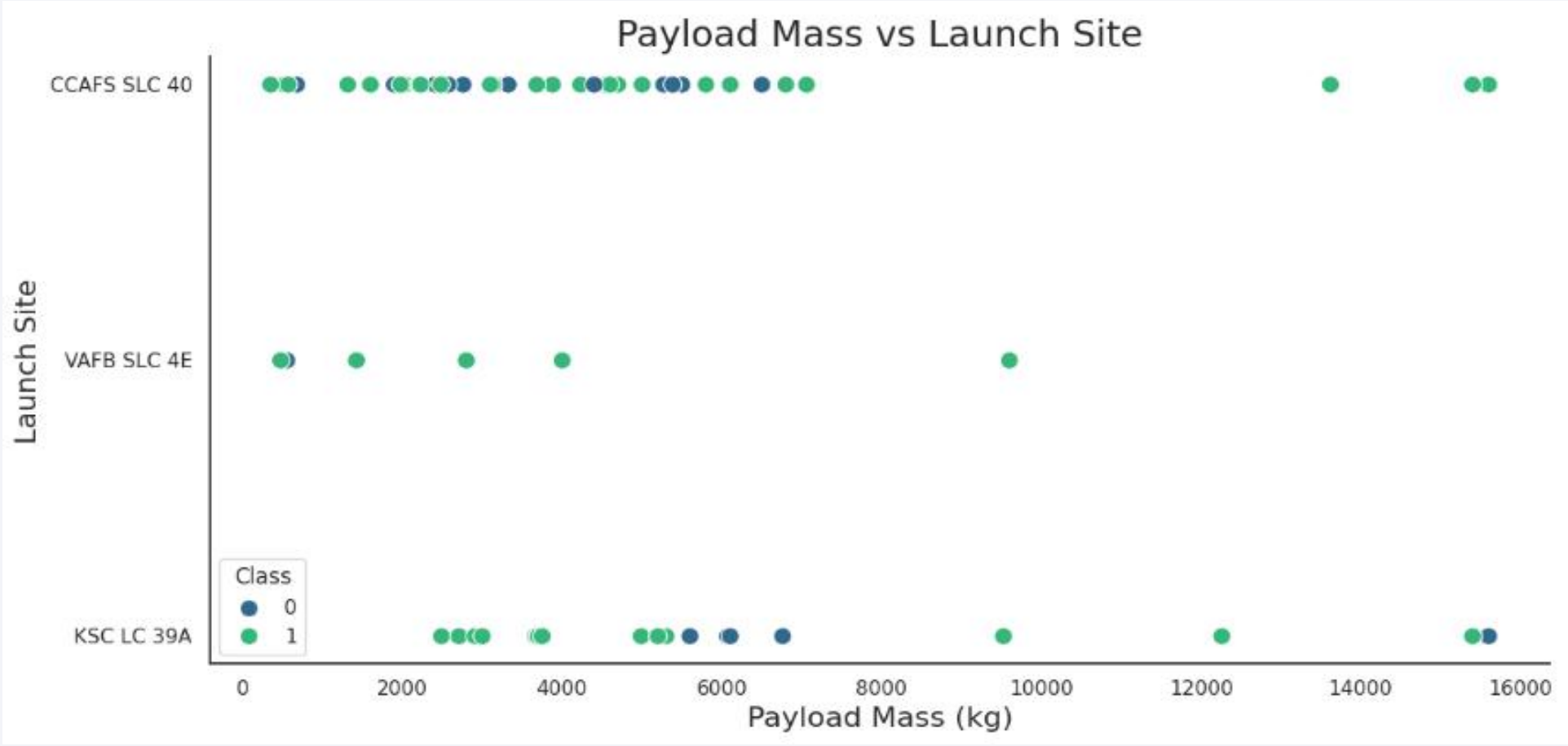
Insights drawn from EDA

Flight Number vs. Launch Site



- CCAFS SLC-40: Highest launches; more success rate 70 flights.
- VAFB SLC-4E: No launches beyond 70 flights
- KSC LC-39A: Launches start above 20 flights.

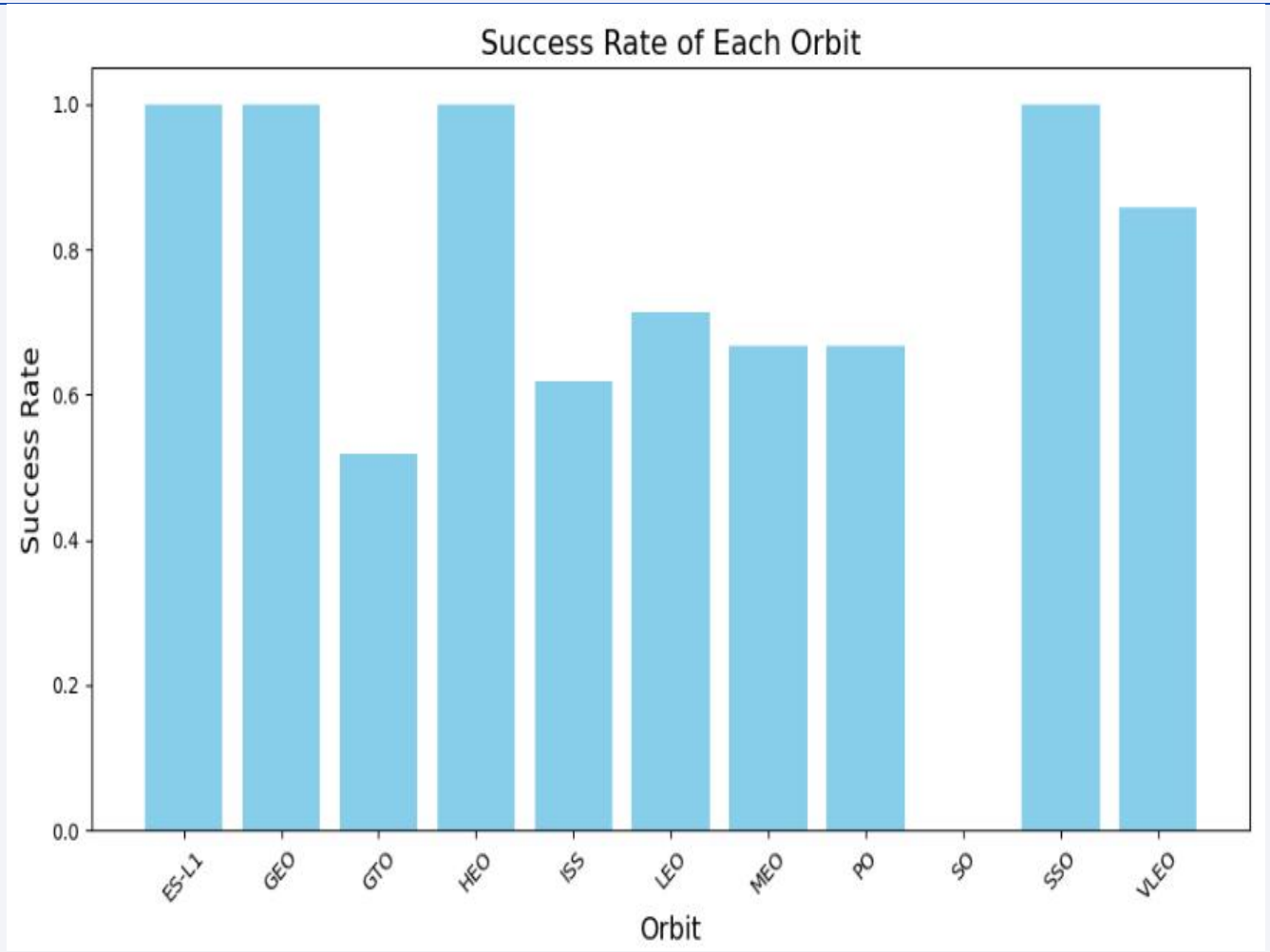
Payload vs. Launch Site



- VAFB-SLC launchsite had no rockets launched for heavypayload mass(>10000).

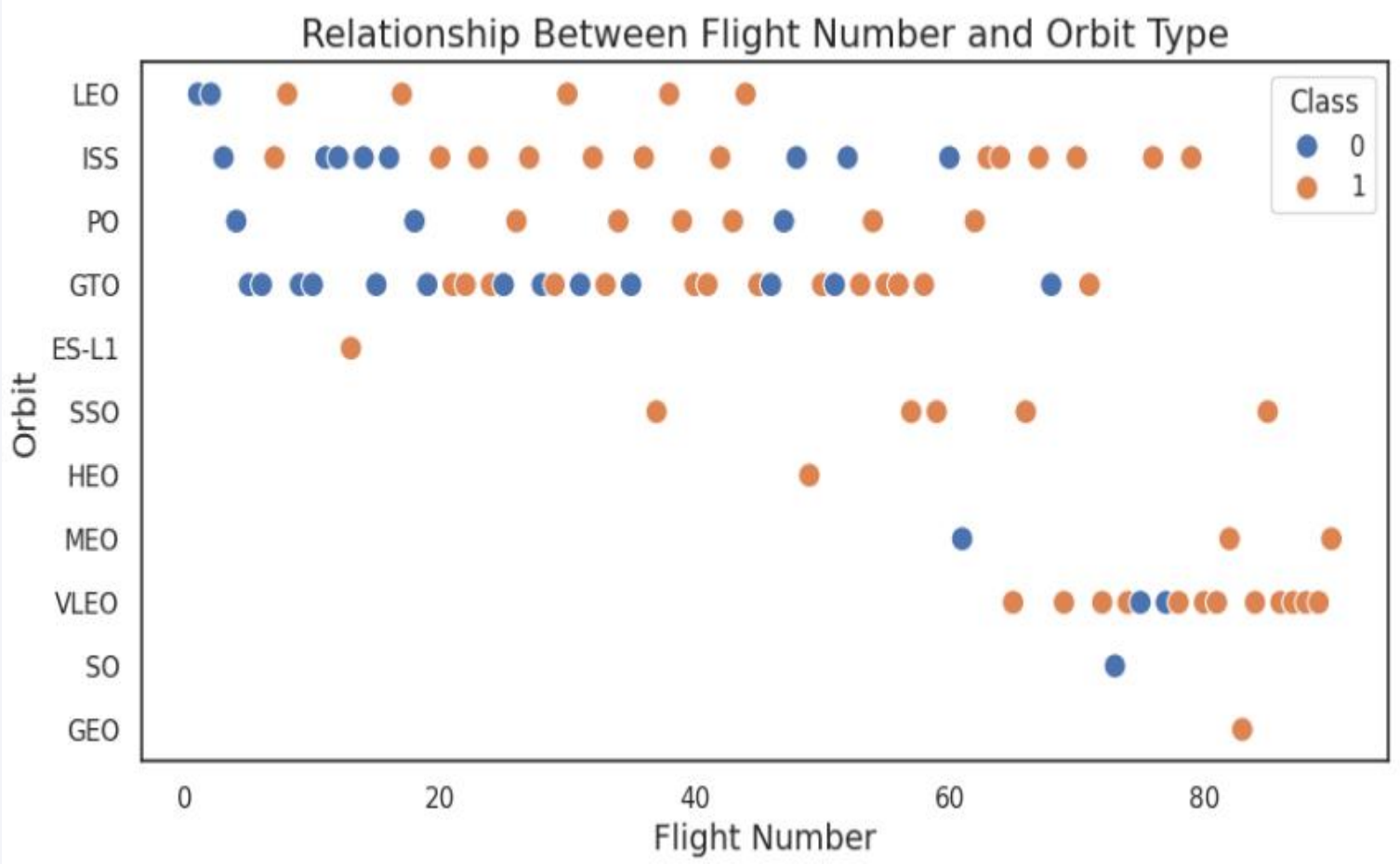
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO: Highest success rates
- GTO: Lowest success rate
- SO: No successful launches recorded



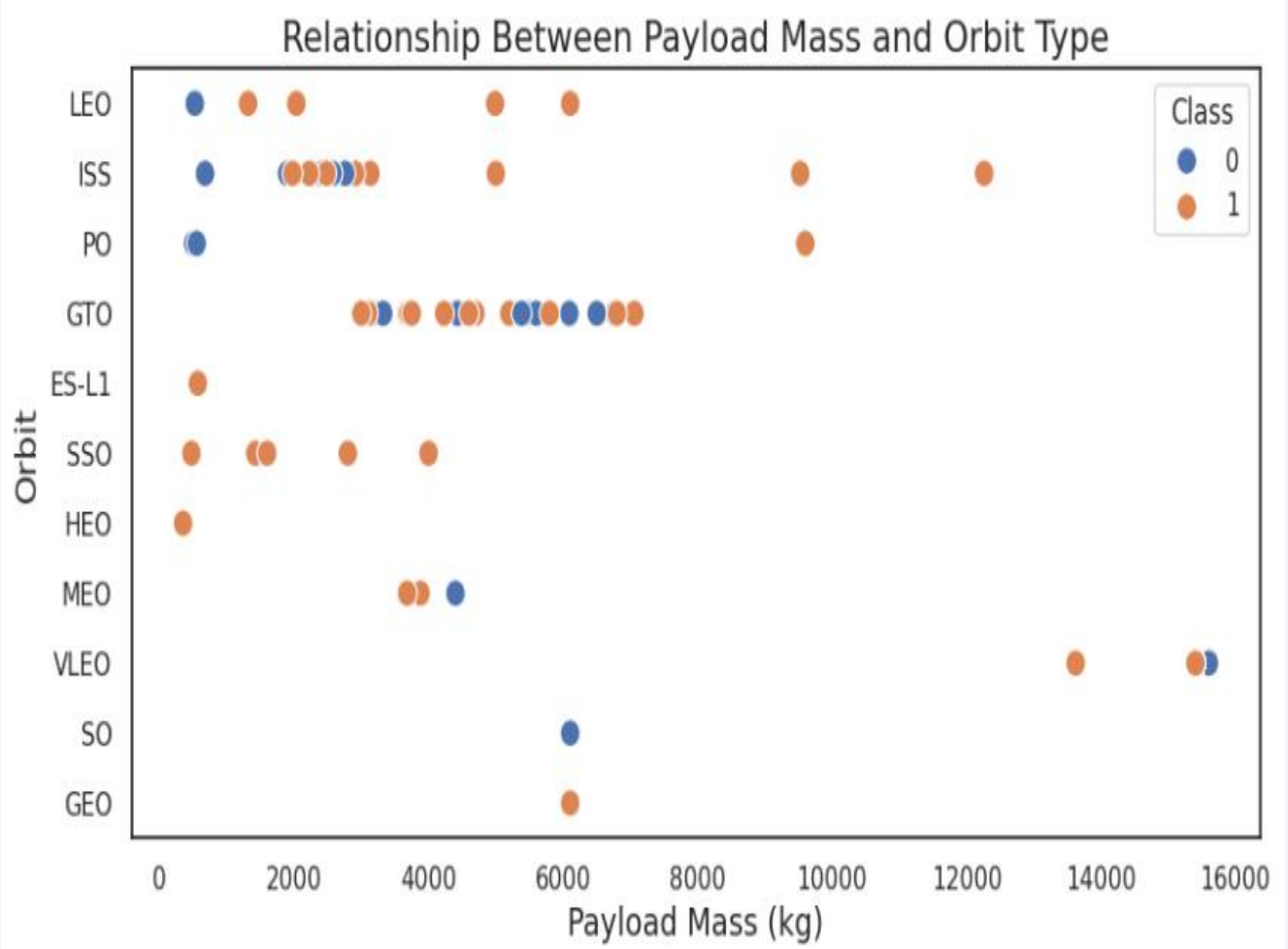
Flight Number vs. Orbit Type

- In the LEO orbit, success seems to be related to the number of flights.
- Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.



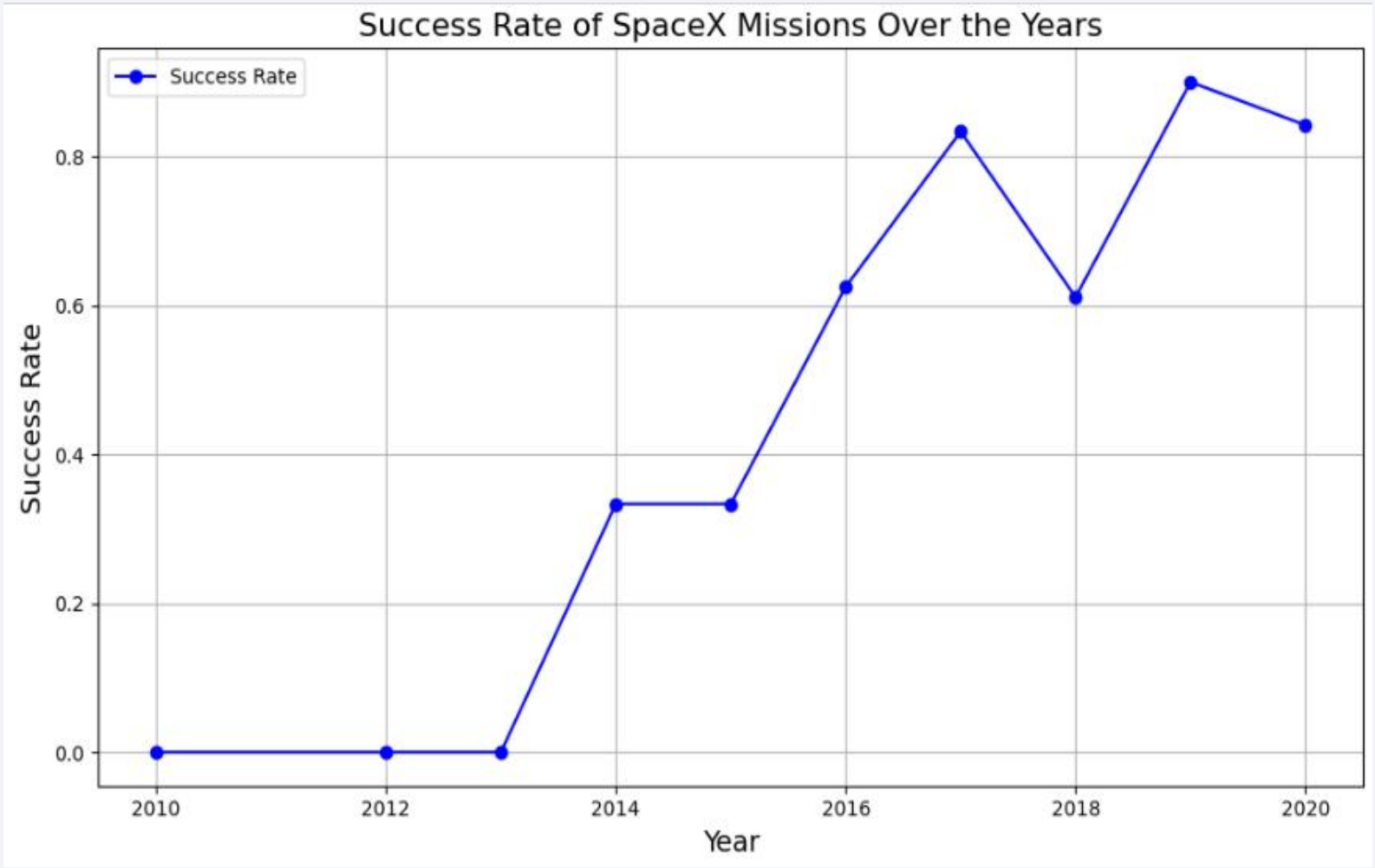
Payload vs. Orbit Type

- LEO, ISS, and PO Orbits had positive relationship with payload mass
- ES-L1, SSO, and HEO Orbits seem to have no failure record.
- GTO orbit's payload mass outcomes were between 3000 and 8000 kg.
- VLEO is the only Orbit with with payload mass records beyond 14000 kg



Launch Success Yearly Trend

- Sucess rate since 2013 kept increasing till 2020



All Launch Site Names

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

SpaceX has four unique launch sites. They are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parach
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 (parach
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

- Five records of launch sites begining with `CCA` are all CCAFS LC-40 Launch Site
- All of them are also F9 v1.0 Booster_Versions

Total Payload Mass

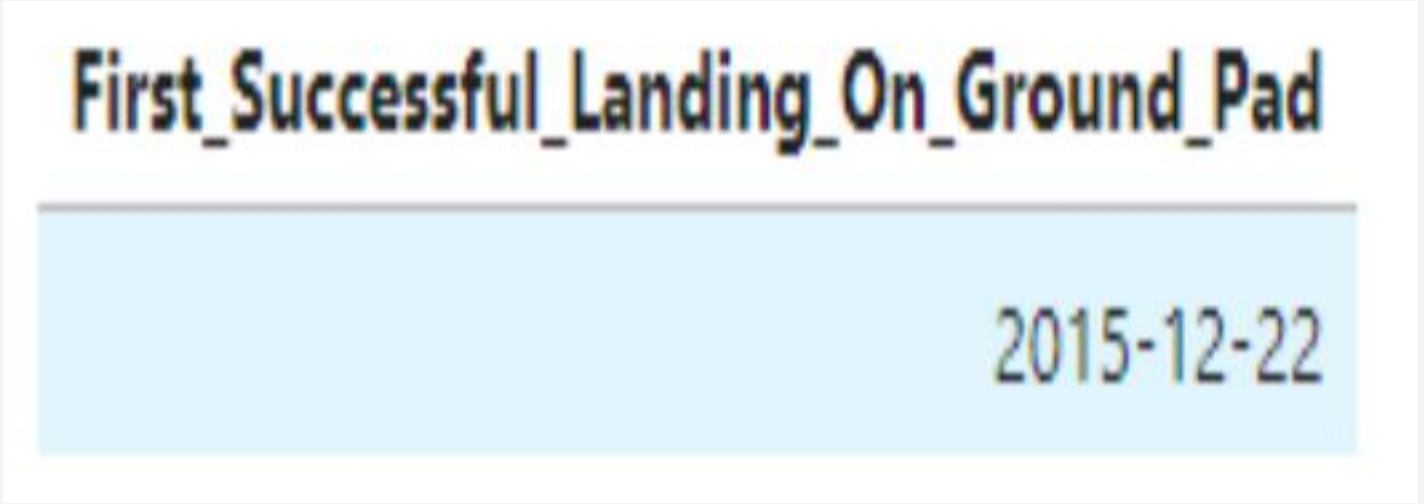
The total payload mass is 48,213 kg

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg

First Successful Ground Landing Date



The date the of the first successful landing outcome on ground pad was 22nd December 2015.

This date is important as it mark the initial successful operations on ground pad, which is crucial for planning and further development of space missions.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Names of Boosters Successfully Landed on Drone Ship with Payload Mass Between 4000 and 6000 kg are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2

These boosters successfully landed on a drone ship and had payload masses that fall within the specified range of 4000 kg to 6000 kg. This demonstrates the Falcon 9 rocket's capability to handle significant payloads while achieving successful drone ship landings.

Total Number of Successful and Failure Mission Outcomes

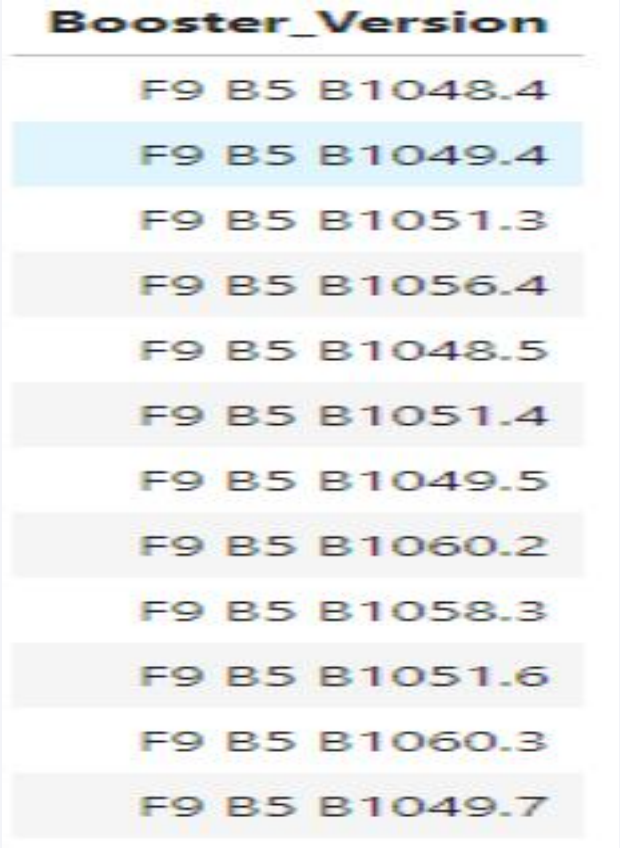
- There were 98 successful missions in total.
- It appears there was not failure

Mission_Outcome	Total_Outcomes
Success	98

Boosters Carried Maximum Payload

- The list of boosters in the image are those that carried maximum payload mass.
- All of them are F9 B5 booster versions
- This indicates that these specific missions achieved the highest payload capacities with the Falcon 9 B5 booster versions.
- The ability of these boosters to carry such large payloads underscores the impressive payload capacity of the Falcon 9, making them particularly significant for missions requiring heavy payloads.

Image of List of Boosters Carried Maximum Payload



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

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

- In 2015, two Falcon 9 missions attempting to land on a drone ship ended in failure.
  - January 2015: F9 v1.1 B1012 attempted a drone ship landing but did not succeed.
  - April 2015: F9 v1.1 B1015 also failed in its attempt to land on a drone ship.
- Both of these attempts were launched from CCAFS LC-40

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

Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3

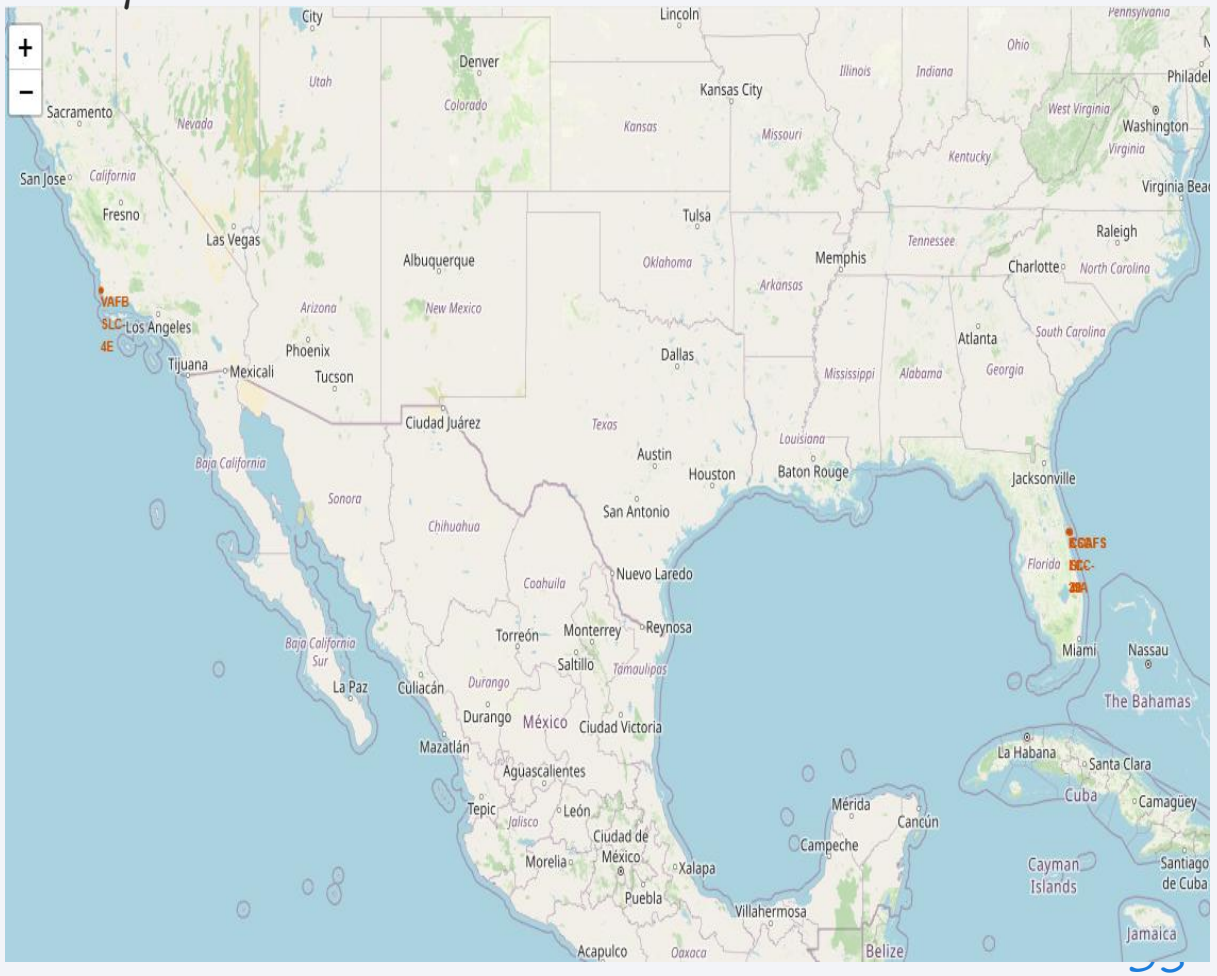
Between June 4, 2010, and March 20, 2017, there were 5 failed attempts to land on a drone ship compared to 3 successful landings on a ground pad. This reflects the challenges of perfecting drone ship landings while demonstrating some early successes in ground pad landings.


Section 3
Launch Sites
Proximities Analysis

Map of all launch sites' location markers on a global map

- The right side image shows all the four launch sites' locations.
- However, on the global map, it appears only two sites are on the map.
- VAFB SLC-4E is located at the western coast of the United States.
- The remaining three sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A are located at the eastern coast of the United States(See the zoom map below for these launch sites).

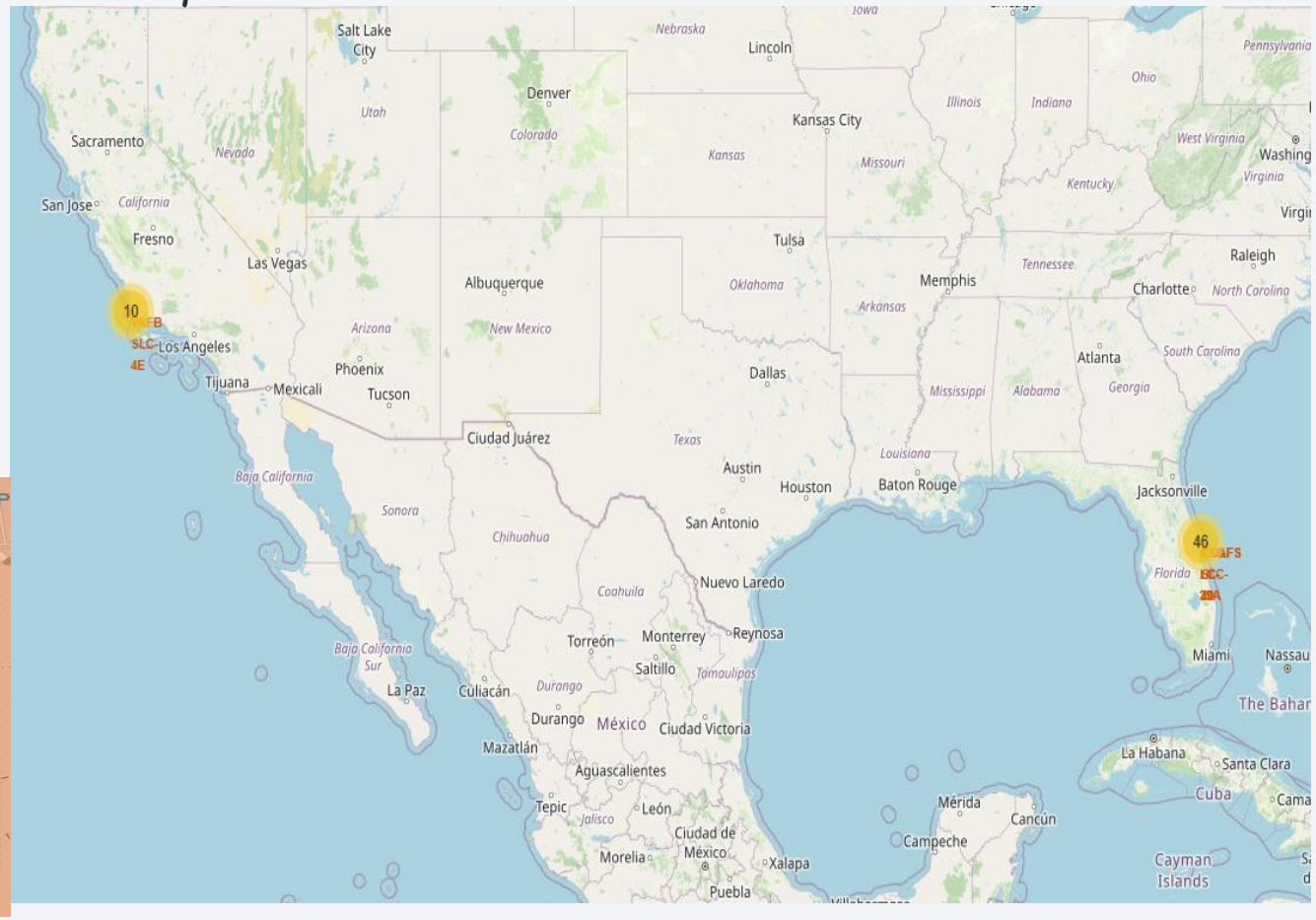
Map of all launch sites' location markers



Color-labeled launch outcomes

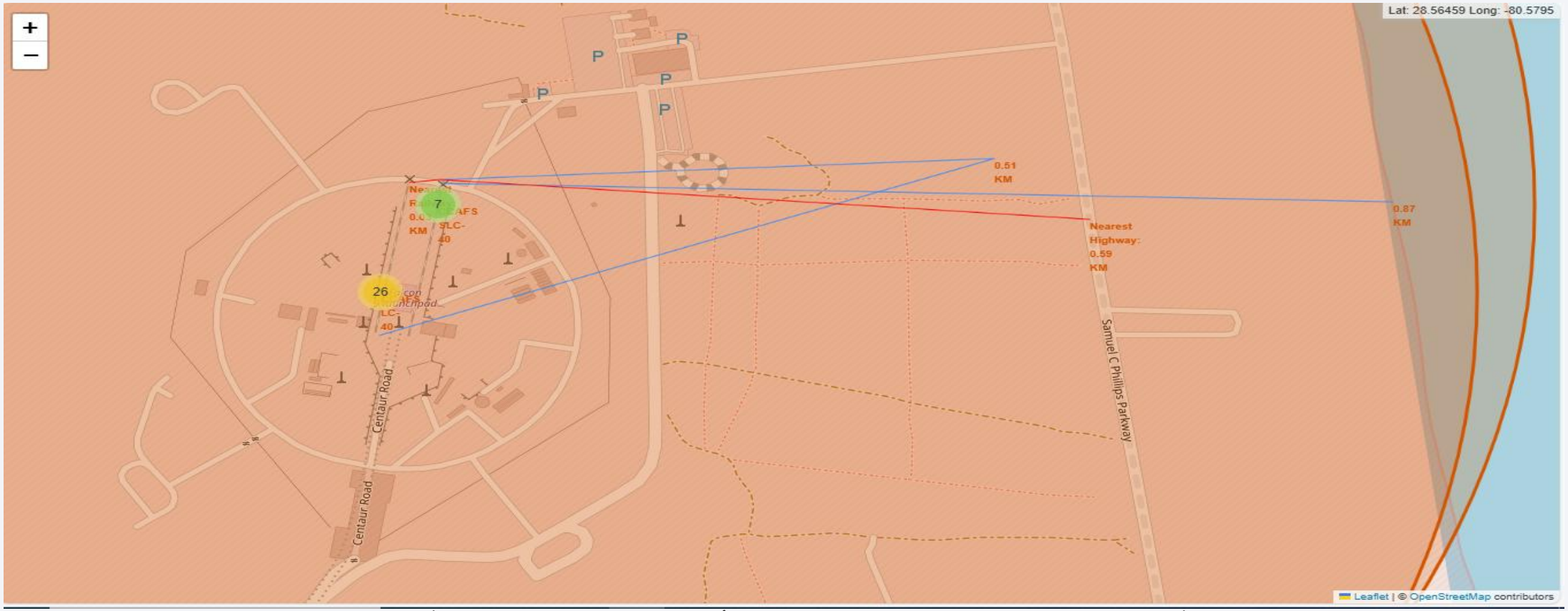
- VAFB SLC-4E had 10 launch success.
- CCAFS LC-40 (26), CCAFS SLC-40 (7), and KSC LC-39A (10) had a total of 46 launch success (see the zoomed images below).

Map of color-labeled launch outcomes

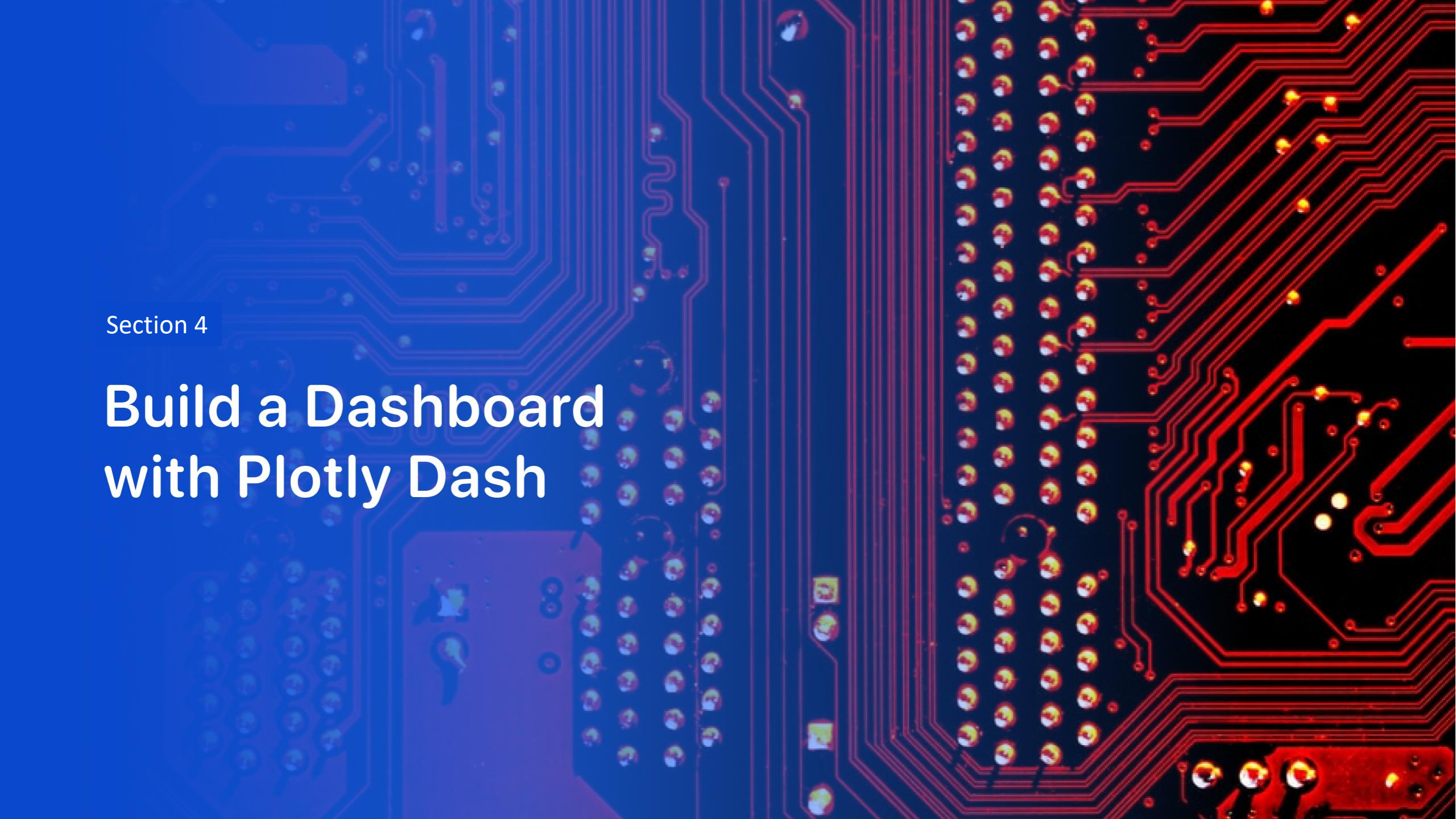


Launch site to its proximities

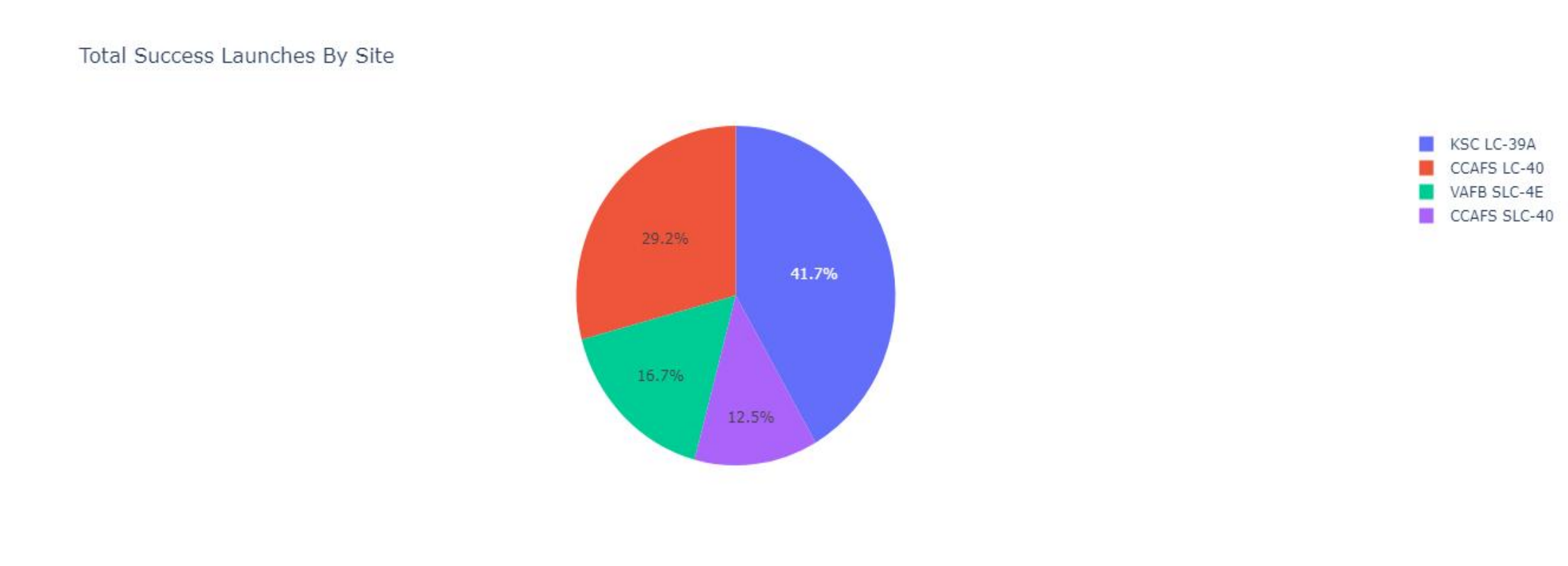
Map of a selected launch site to its proximities such as railway, highway, and coastline



- From CCAFS SLC-40 launch site, the distance to the nearest: railway station is 0.03KM; highway is 0.59 KM; and coastline is 0.87KM.

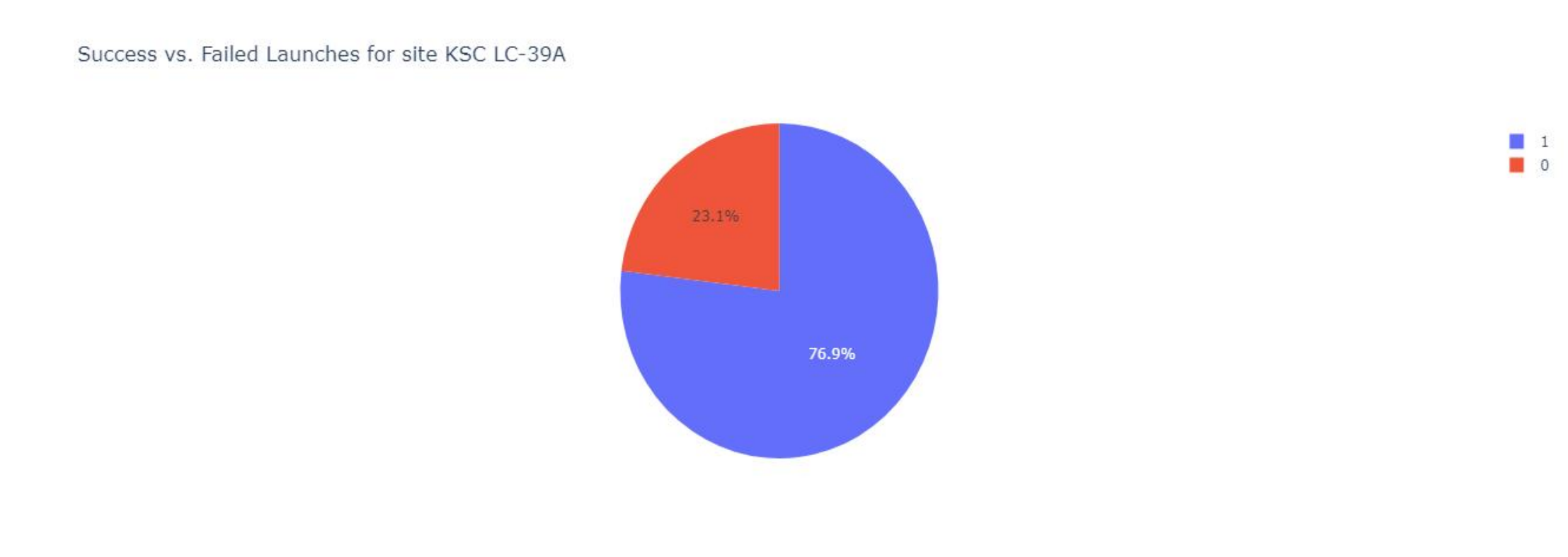

Section 4
Build a Dashboard with Plotly Dash

Luach Success for All Sites



KSC LC-39A is the sitewith the largest successful launches (76.9%) while CCAFS SLC-40 has the smallest launches (12.5%)

Launch Site with Highest Launch Success Ratio



- KSC LC-39A site has the highest launch success rate at 76.9%.

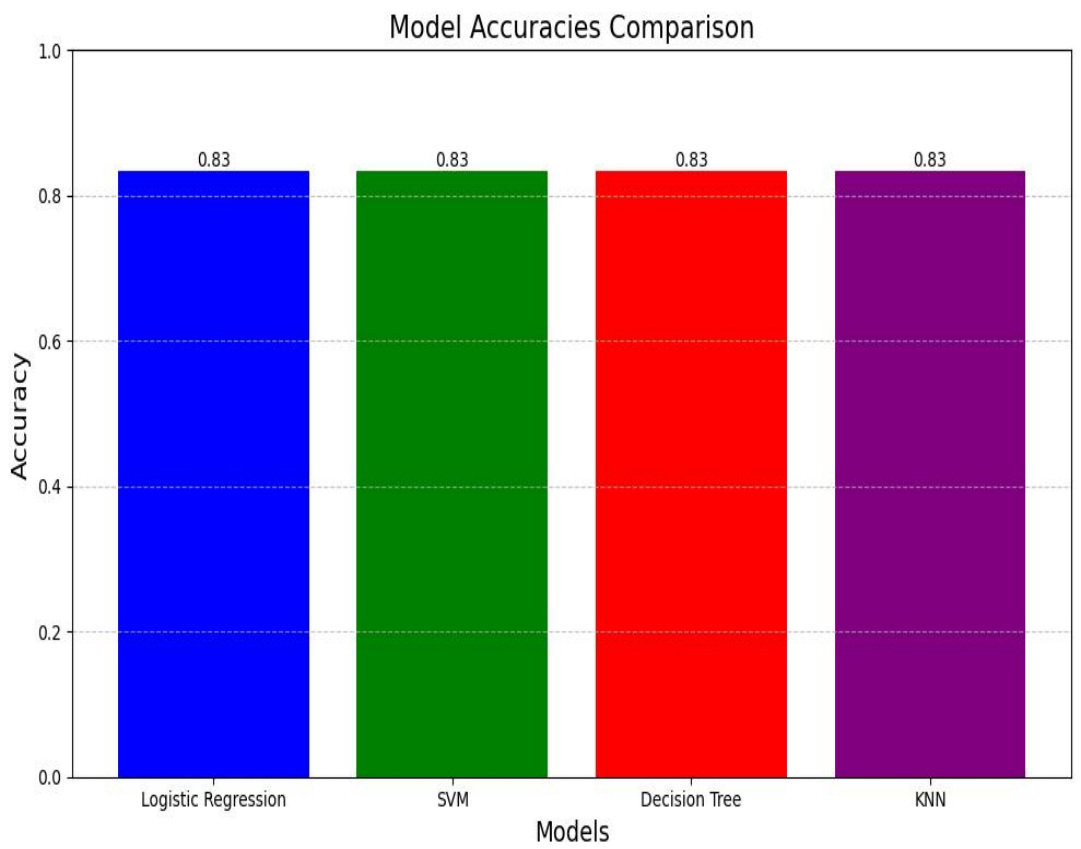
Payload vs. Launch Outcome scatter plot

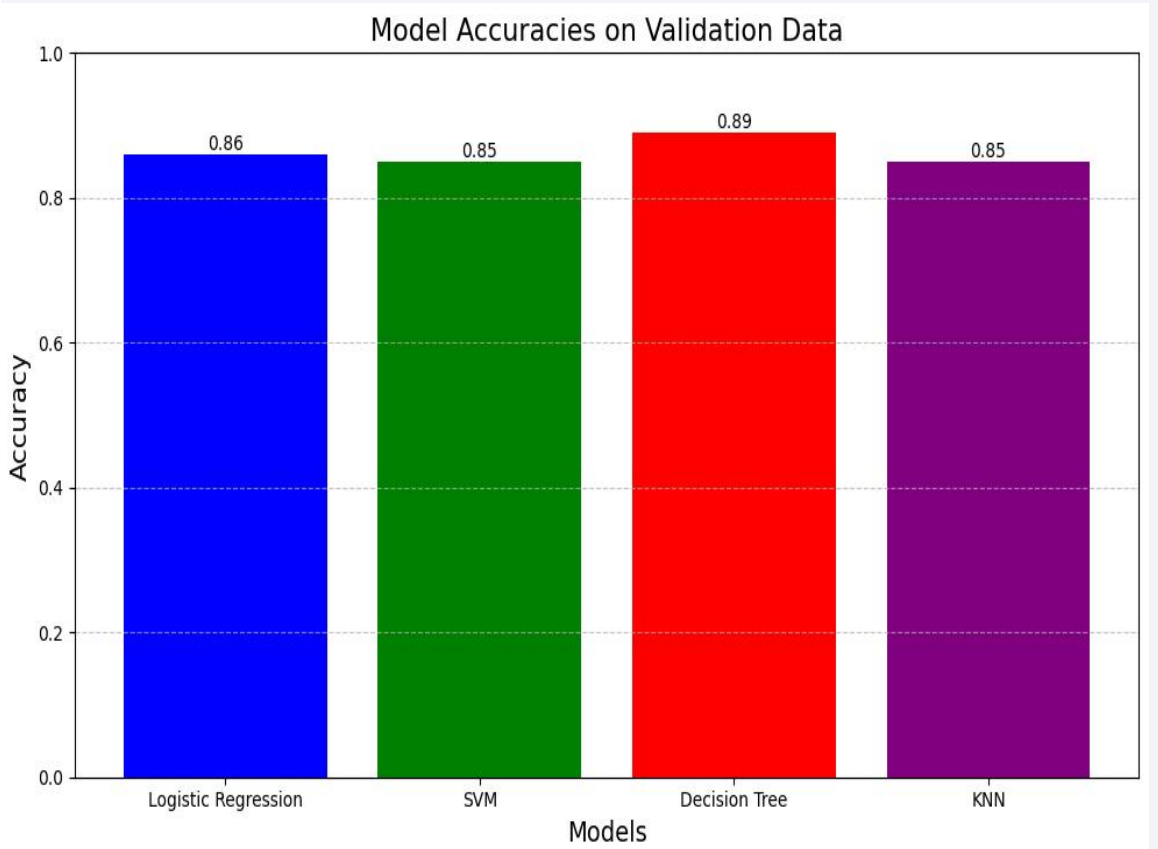
- F9 Booster version B4 has the highest launch success rate/payload ranges (9600 KG)

Section 5

Predictive Analysis (Classification)

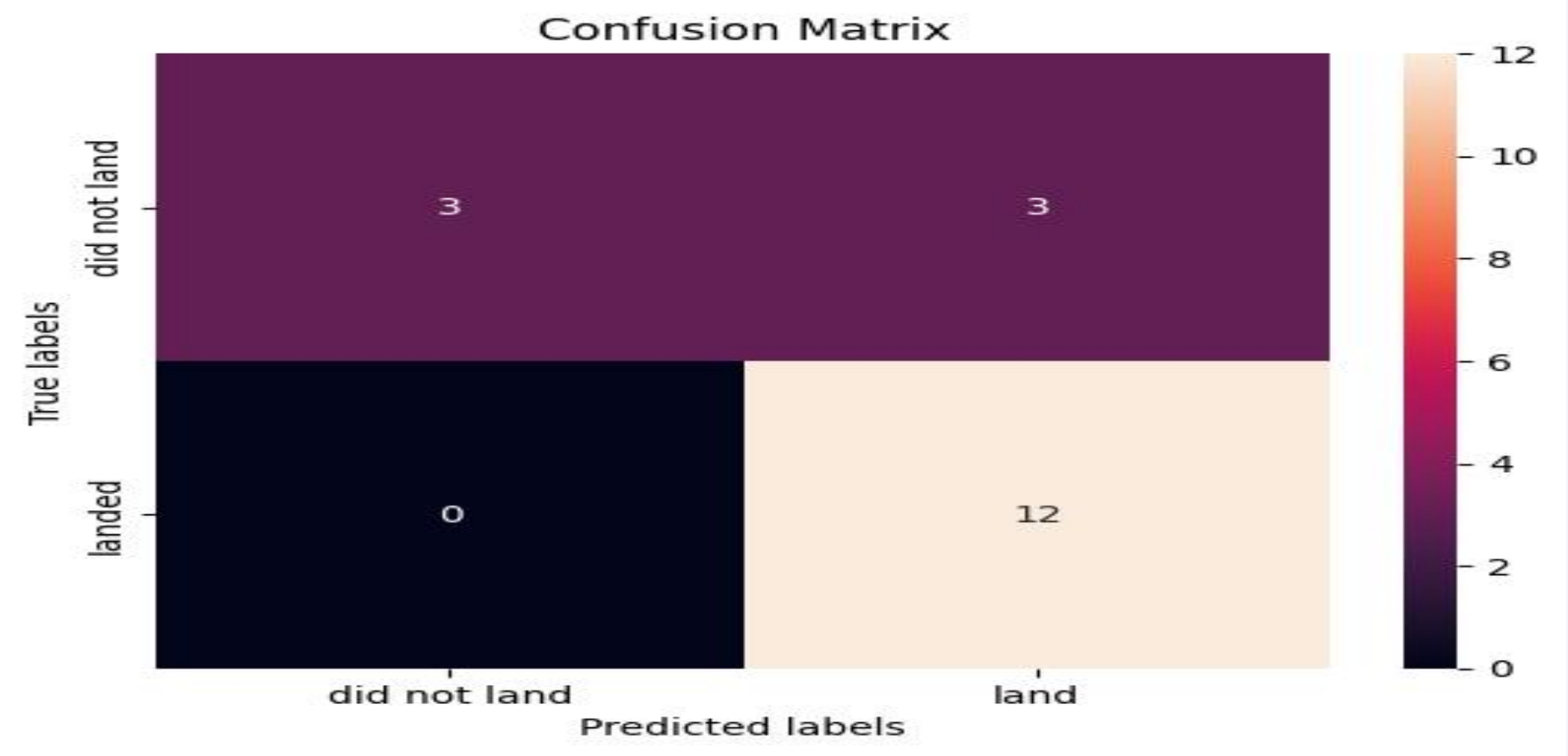
Classification Accuracy





- The models' accuracies on the test data are consistent for all models.
- In terms of accuracy on validation data, Decision tree offers high validation accuracy.

Confusion Matrix

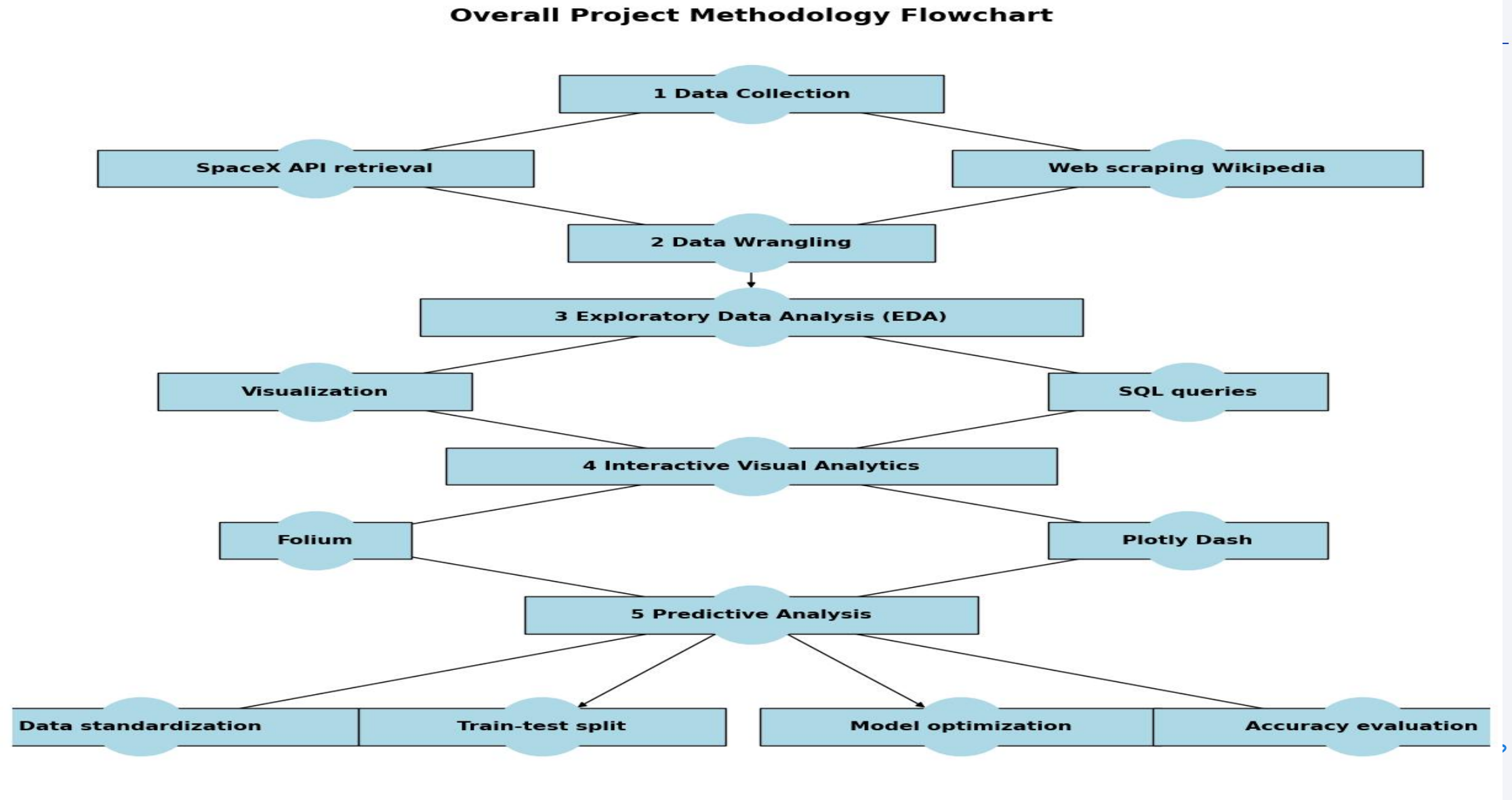


- Examining the confusion matrix, we see that the models can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- The project aimed to predict Falcon 9 landing success, crucial for SpaceX's cost efficiency.
- Data was collected using the SpaceX API and Wikipedia, then analyzed with SQL, Folium, and Plotly Dash and predictive models.
- EDA showed a rising success rate from 2013 to 2020, with key milestones including the first successful ground pad landing in 2015 and varied payload capacities.
- KSC LC-39A had the highest success rate at 76.9%, while CCAFS SLC-40 had the lowest at 12.5%, and F9 Booster version B4 had the highest payload capacity and success rate.
- Predictive models achieved 83% accuracy and effectively distinguishing between classes (success or failure).

Appendix


Overall Project Methodology Flowchart
1 Data Collection
SpaceX API retrieval
Web scraping Wikipedia
2 Data Wrangling
3 Exploratory Data Analysis (EDA)
Visualization
SQL queries
4 Interactive Visual Analytics
Folium
Plotly Dash
5 Predictive Analysis
Data standardization
Train-test split
Model optimization
Accuracy evaluation

Thank you!