IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lampros Velentzas
7-Mar-2025

Outline

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

Executive Summary

Summary of methodologies

- **Data collection:** Scraped SpaceX launch data from SpaceX API and Wikipedia
- **Data Wrangling & Cleaning:** Processed missing values, engineered key features (launch site, payload mass, booster version etc.)
- **Exploratory Data Analysis (EDA):**
  - Visualized trends in launch success rates over time
  - Analyzed impact of launch site, payload mass and booster version
- **Machine Learning**
  - Built classification models (Logistic Regression, Decision Trees, SVM, KNN)
  - Selected best-performing model based on accuracy, precision, recall and F1-score

Key findindings & Results

- **Landing Success Rate:** Identified key factors influencing booster landings
- **Best Model Performance:** Decision Tree achieved 88.93% accuracy in predicting successful landings
- **Key Predictors:**
  - Payload mass has a non-linear impact on landing success
  - Certain launch sites have higher success rates
  - Newer booster versions show improved reliability
- **Business Impact:** Model can help SpaceX optimize launch decisions, reducing mission costs

Introduction

Project background and context

SpaceX has revolutionized the aerospace industry by successfully launching and landing reusable rockets, significantly reducing space exploration costs. The **Falcon 9** first-stage booster landing is a key milestone in ensuring the reusability of rockets, making space missions more economical and sustainable. However, predicting the success of the booster landing remains a complex task, as many factors affect the outcome, such as **payload mass, booster version, weather conditions, and launch site**.

As SpaceX aims to increase the frequency and reliability of their launches, a data-driven approach is crucial in predicting the success or failure of these landings. This project focuses on analyzing historical launch data to build a predictive model that can forecast the success of a Falcon 9 first-stage landing.

Problem Statement & Objectives

The primary **problem** addressed in this project is to **predict whether the Falcon 9 first-stage booster will successfully land** after launch. Specifically, we aim to:

- **Identify key factors** that influence the likelihood of a successful booster landing.
- **Build a predictive model** that can accurately classify the success or failure of the landing, based on historical launch data.
- **Evaluate the performance** of several machine learning models to determine the best approach for prediction.

By tackling these questions, the project aims to **inform future SpaceX missions** and improve decision-making regarding rocket reuse, ultimately contributing to **cost savings and efficiency**.

Section 1

Methodology

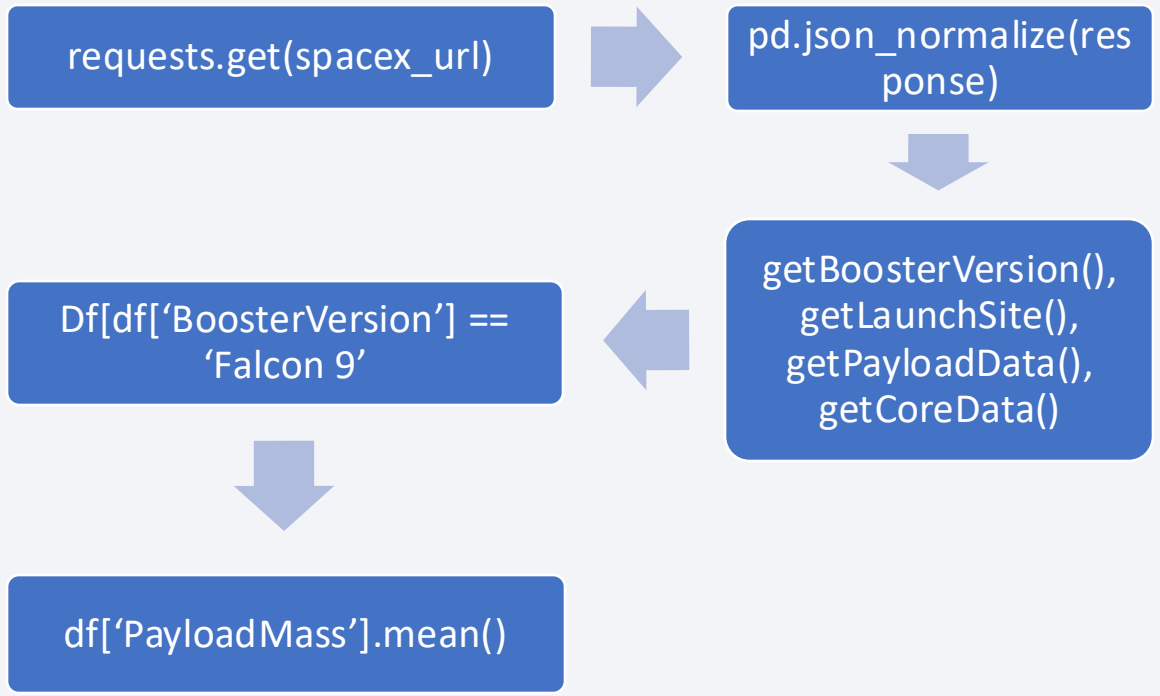
Methodology

Executive Summary

- Data collection methodology:
  - Requests to the SpaceX API
  - Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia
- Perform data wrangling
  - Using pandas performed standard data wrangling operations
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection – SpaceX API

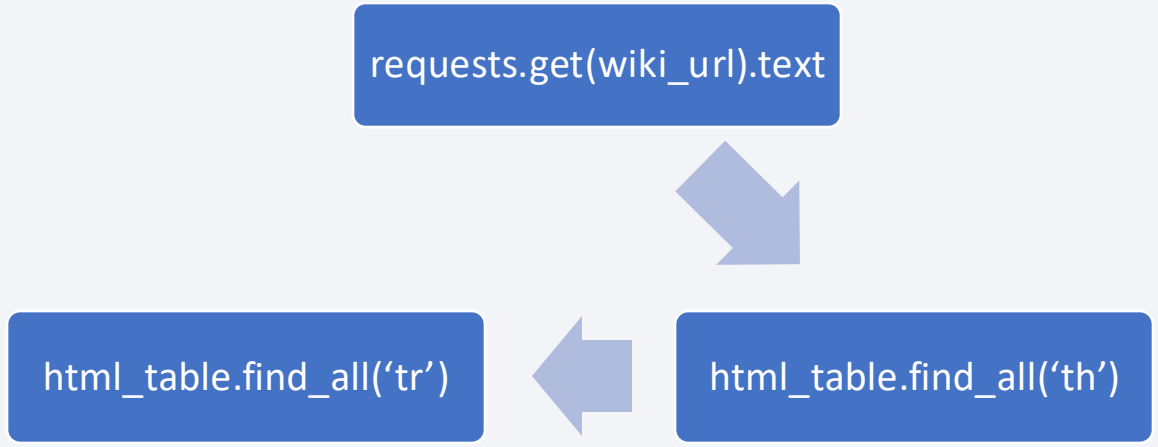
- Request the json payload via an API call from SpaceX API
- Decode the response and turn it to a pandas df
- Use helper functions to extract additional information
- Filter df to only include Falcon 9 launches
- Use mean() to replace 'PayloadMass' nulls



https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab1a-spacex-data-collection-api.ipynb

Data Collection - Scraping

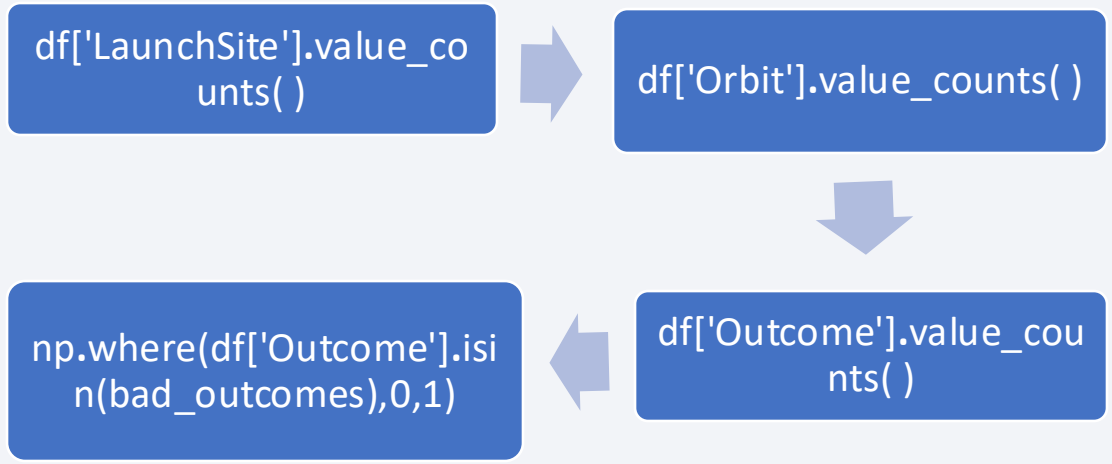
- Perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.
- Create a BeatifulSoup object from the HTML response
- Extract all column/variable names from the HTML table header
- Create a df by parsing the data from the table



https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab1b-webscraping.ipynb

Data Wrangling

- Calculate the number of launches on each site
- Calculate the number of occurrence of each orbit
- Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column



https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab1c-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Used scatterplot of ‘Flight Number’ vs ‘Launch Site’ to identify whether there are patterns that affect the outcome
- Used scatterplot of ‘Payload’ vs ‘Launch Site’ to identify whether there are patterns that affect the outcome
- Used barplot of ‘Success Rate’ vs ‘Orbit Type’ to identify whether there are orbits with higher success rate than others
- Used scatterplot of ‘Flight Number’ vs ‘Orbit Type’ to identify patterns
- Used scatterplot of ‘Payload’ vs ‘Orbit Type to identify patterns
- Used line plot of ‘Year’ vs ‘Success Rate’ to identify potential trend

https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab2b-edadataviz.ipynb

EDA with SQL

- Identified the names of the unique launch sites
- Displayed 5 records where launch sites begin with the string 'CCA'
- Identified the total payload mass carried by boosters launched by 'NASA (CRS)'
- Identified the average payload mass carried by booster version F9 v1.1
- Listed the date when the first successful landing outcome in ground pad was achieved
- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 kg but less than 6000 kg
- Listed the total number of successful and failure mission outcomes
- Listed the names of the boosters which have carried the maximum payload mass
- Listed the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Ranked the count of landing outcomes (such as Failure (drone ship) and Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab2a-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created a folium Map object, with initially centered location to be NASA Johnson Space Center at Houston, Texas.
- Added a folium.Circle (adds small yellow circle) and folium.Marker (to display name of the site) for each launch site on the map
- Used Marker object to color green for successful and red for failed launches
- Used MarkerCluster( ) to simplify the map which contains many markers with the same coordinates (for the same site)
- Used Marker to apply color label and easily identify which launch sites have relatively high success rates
- Used MousePosition() on the map to get coordinate for mouse over a point on the map
- Used icon property DivIcon() to display the distance between coastline point and launch site
- Used folium.PolyLine to update the map with the distance line

https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab3a-launch_site_location_Folium.ipynb

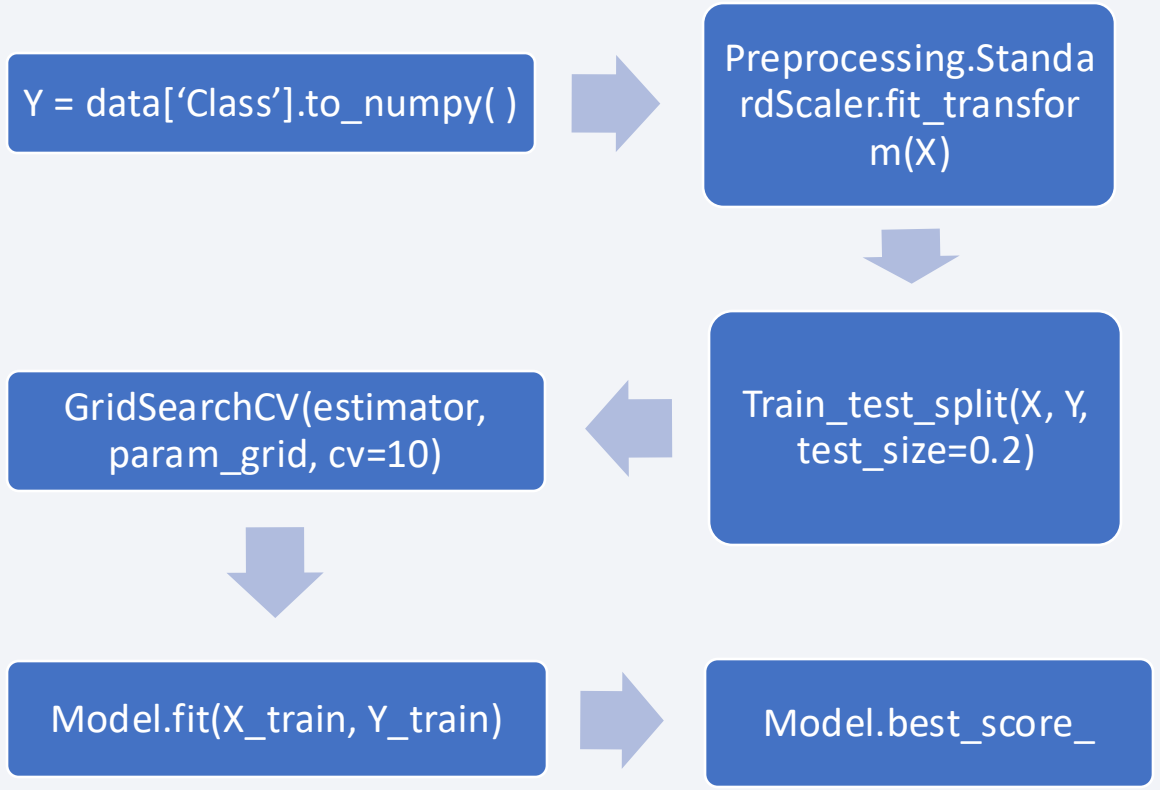
Build a Dashboard with Plotly Dash

- Used Pie-charts to visualize
  - Success launches by site over total
  - Success vs Failure rate per site
- Used colored scatter-plot to identify success over payload mass and booster version
- Used dropdown to add interactivity for both charts to select amongst Launch Sites or even all of them
- Used slider to add interactivity for scatter plot to filter across different payload ranges

https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab3b-Interactive_Dashboard_Plotly_Dash.txt

Predictive Analysis (Classification)

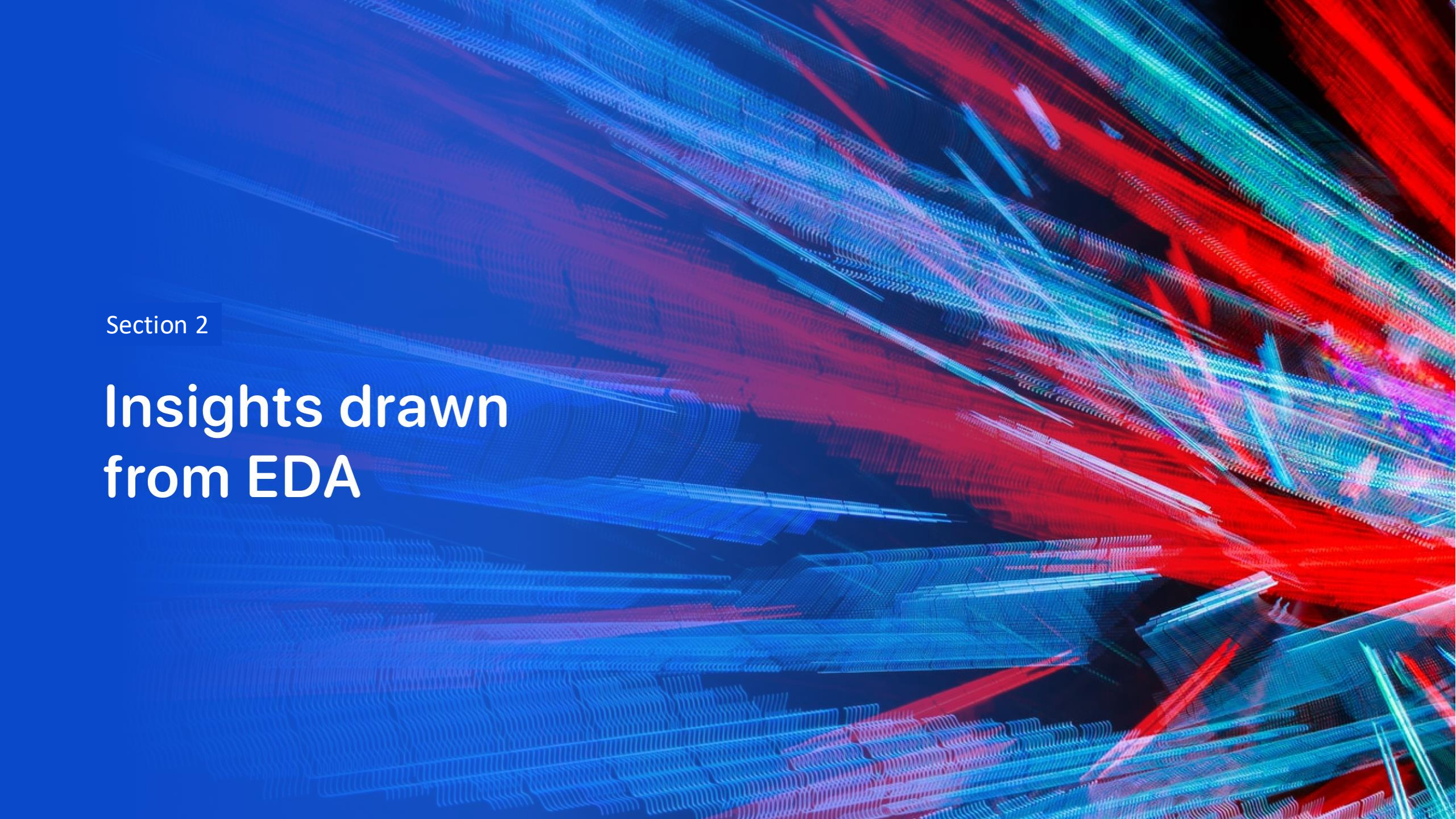
- Created the 'Class' array using to_numpy() method
- Standardised the training data using StandardScaler() and fit_transform()
- Split the data into train and test data using train_test_split()
- Built the algorithms (logistic regression, SVM, decision tree, KNN) with GridSearchCV object, parameters dictionary and 10-fold crossvalidation.
- Train the model with .fit() function
- Used attribute best_score() to calculate the accuracy of the model
- Compared accuracies across models for the test set



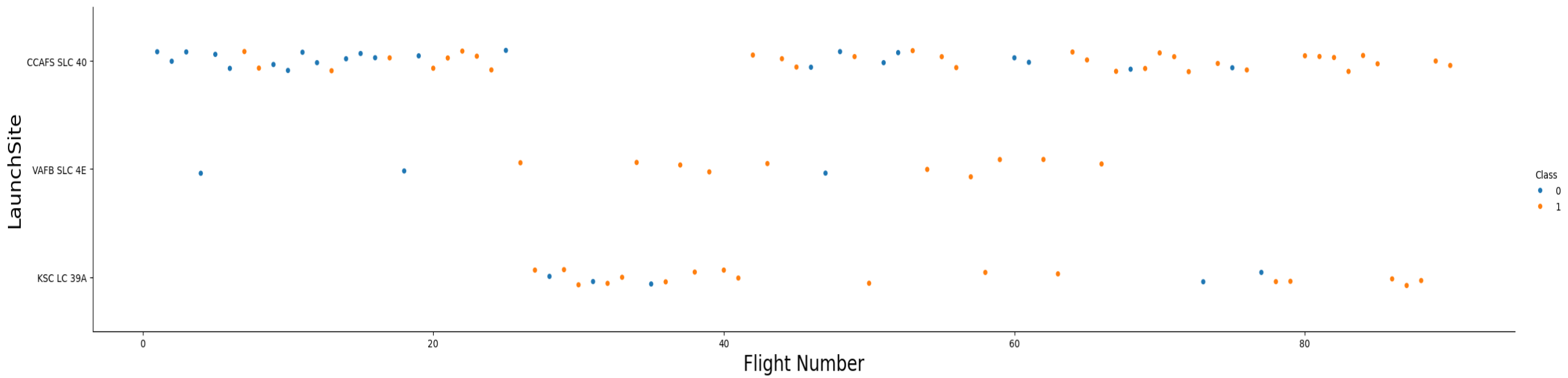
https://github.com/Labrosvel/Data-Scientist-Capstone/blob/main/Lab4a-SpaceX_Machine%20Learning%20Prediction.ipynb

Results

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

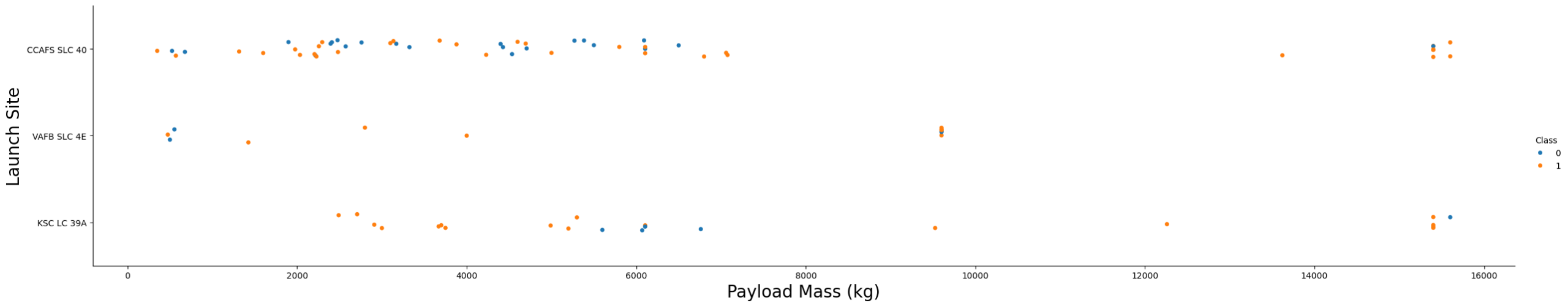

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- We can identify from the plot the following patterns:
  - 'VAFB SLC 4E' site hasn't had any launches since around 65th flight
  - 'CCAFS SLC 40' site has more launches than the others
  - Most recent flights had successful landings, especially after the 80th flight all landed successfully

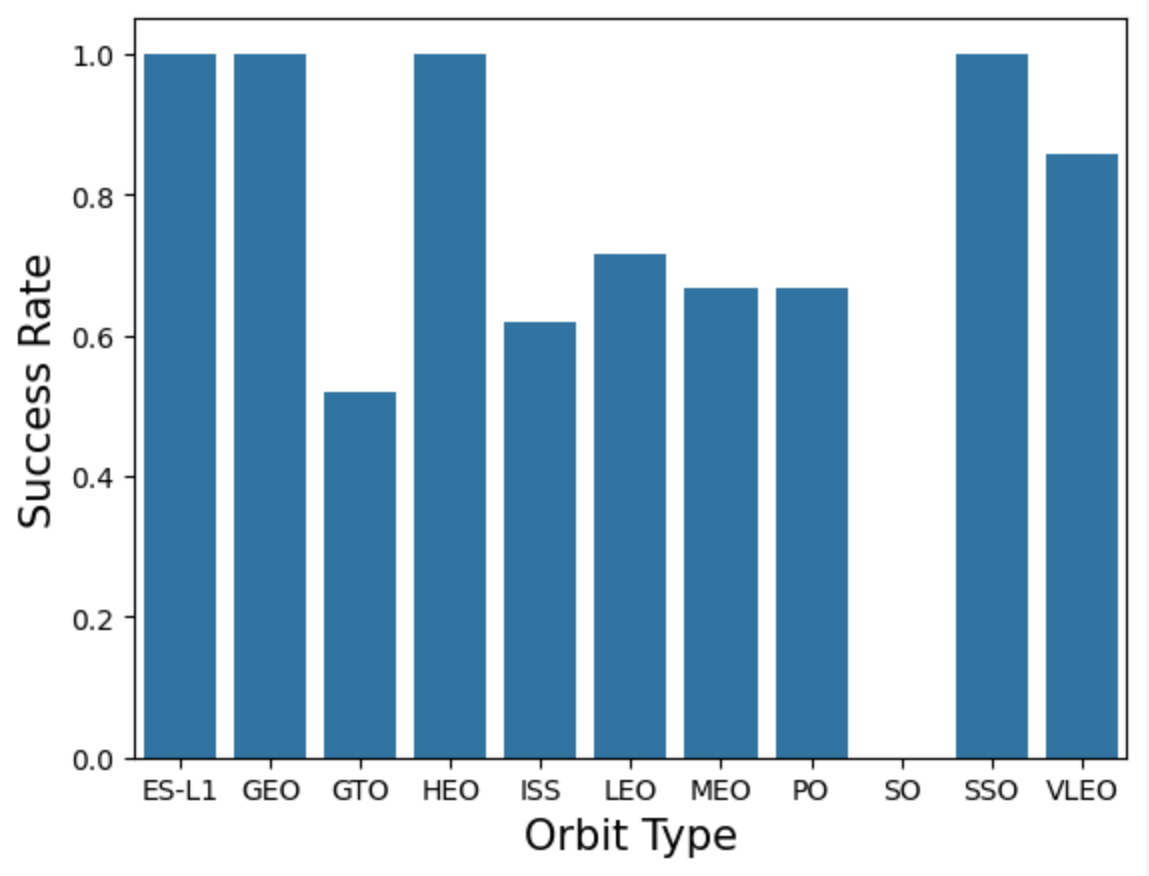
Payload vs. Launch Site



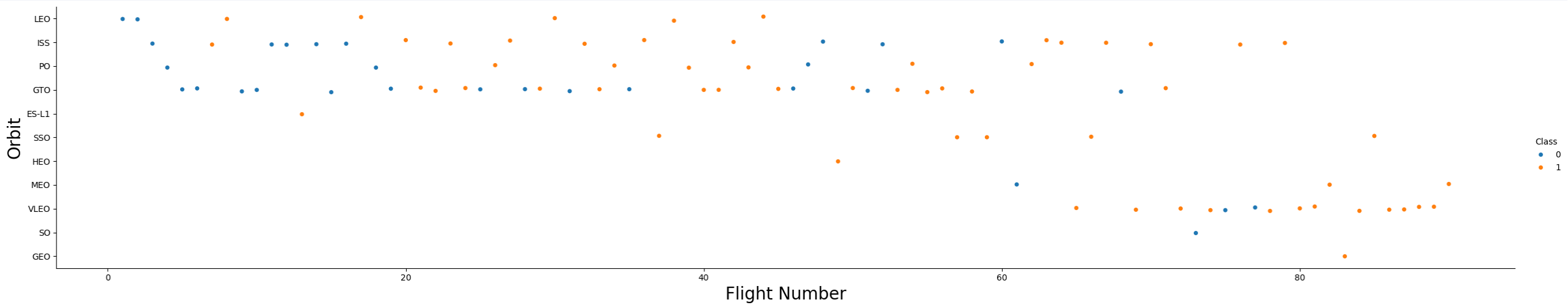
- We can identify from the plot that:
  - The VAFB-SLC launch site has no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type

- We can identify from the plot that:
  - Orbit types: ‘ES-L1’, ‘GEO’, ‘HEO’ and ‘SSO’ have 100% second stage recovery
  - All other orbits have success rate higher than 50%
  - ‘GTO’ orbit has the lowest success rate
  - There is no data related to ‘SO’ orbit

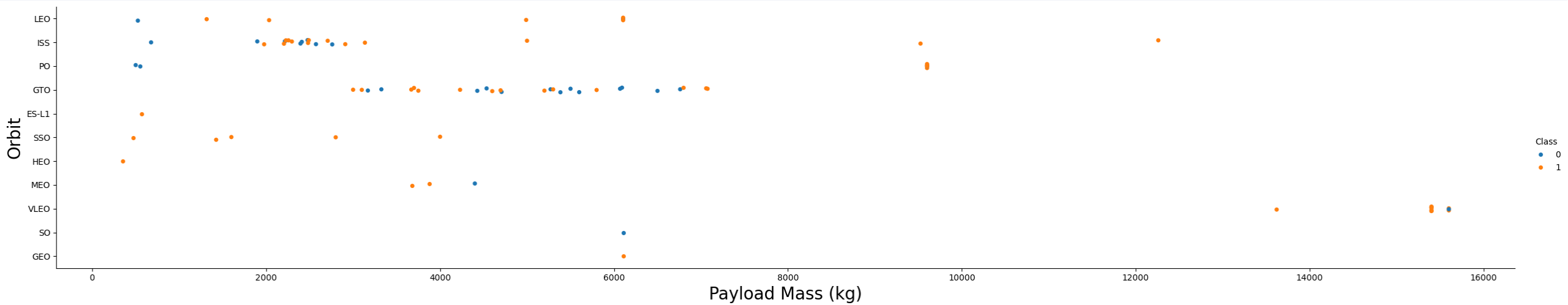


Flight Number vs. Orbit Type



- We can identify from the plot that:
  - In the LEO orbit, success seems to be related to the number of flights.
  - In the GTO orbit, there appears to be no relationship between flight number and success.

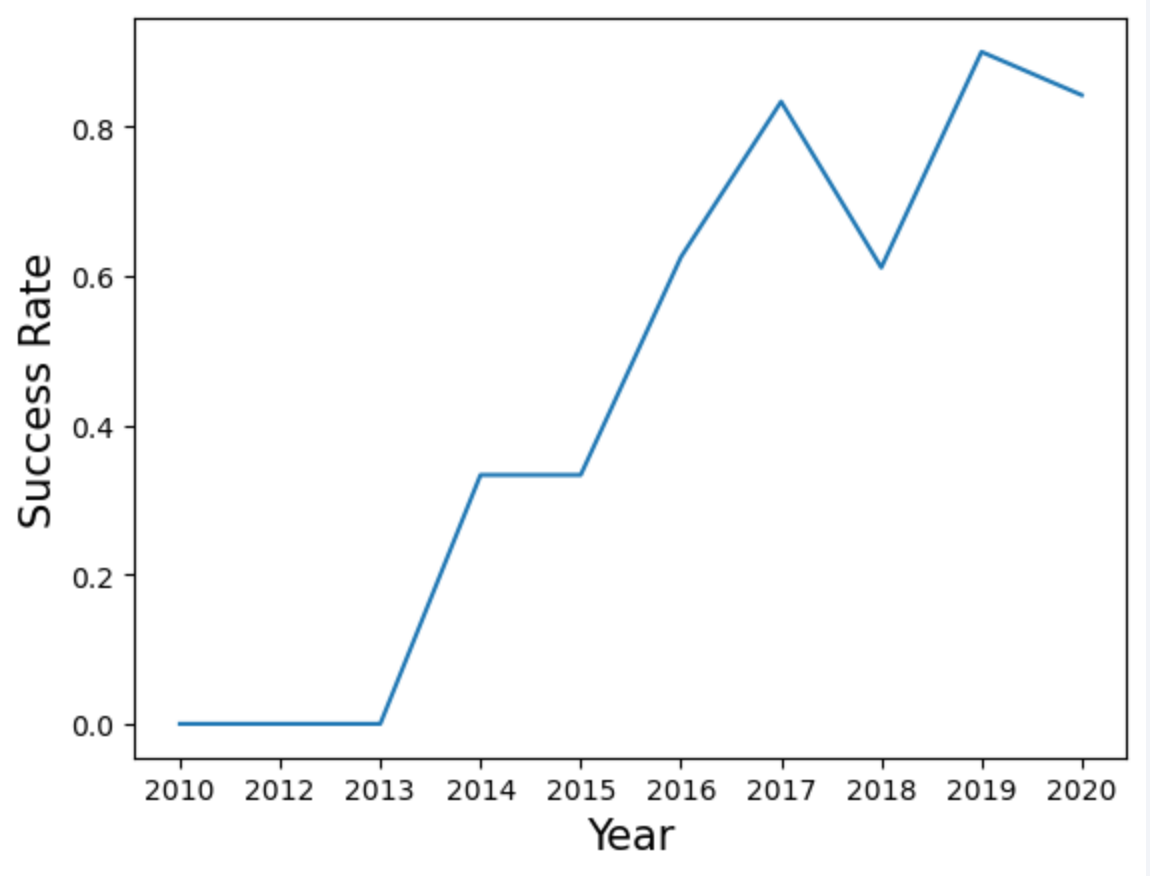
Payload vs. Orbit Type



- We can identify from the plot that:
  - With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
  - However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

- We can identify from the plot that:
  - The success rate since 2013 kept increasing till 2020



All Launch Site Names

- The names of the unique launch sites are:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

Launch Site Names Begin with 'CCA' (sample)

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

- The total payload carried by boosters from ‘NASA (CRS)’ was: 45,596kg

Average Payload Mass by F9 v1.1

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

First Successful Ground Landing Date

- The first successful landing outcome in ground pad was achieved on: 8th-Apr-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is presented in the following table:

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

Boosters Carried Maximum Payload

- The names of the boosters which have carried the maximum payload mass are presented in the following table:

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are presented in the following table:

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

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

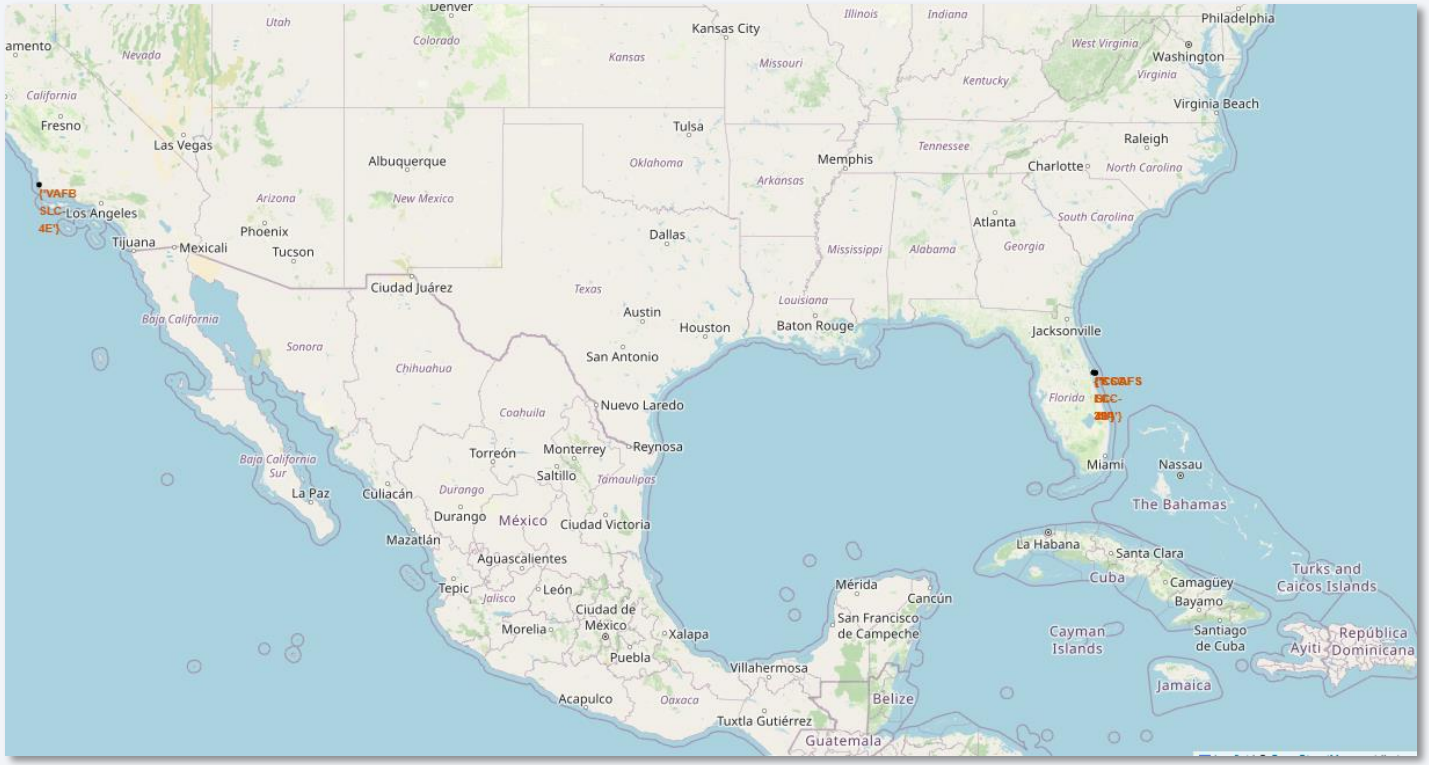
- The ranking of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are presented in the following table:

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

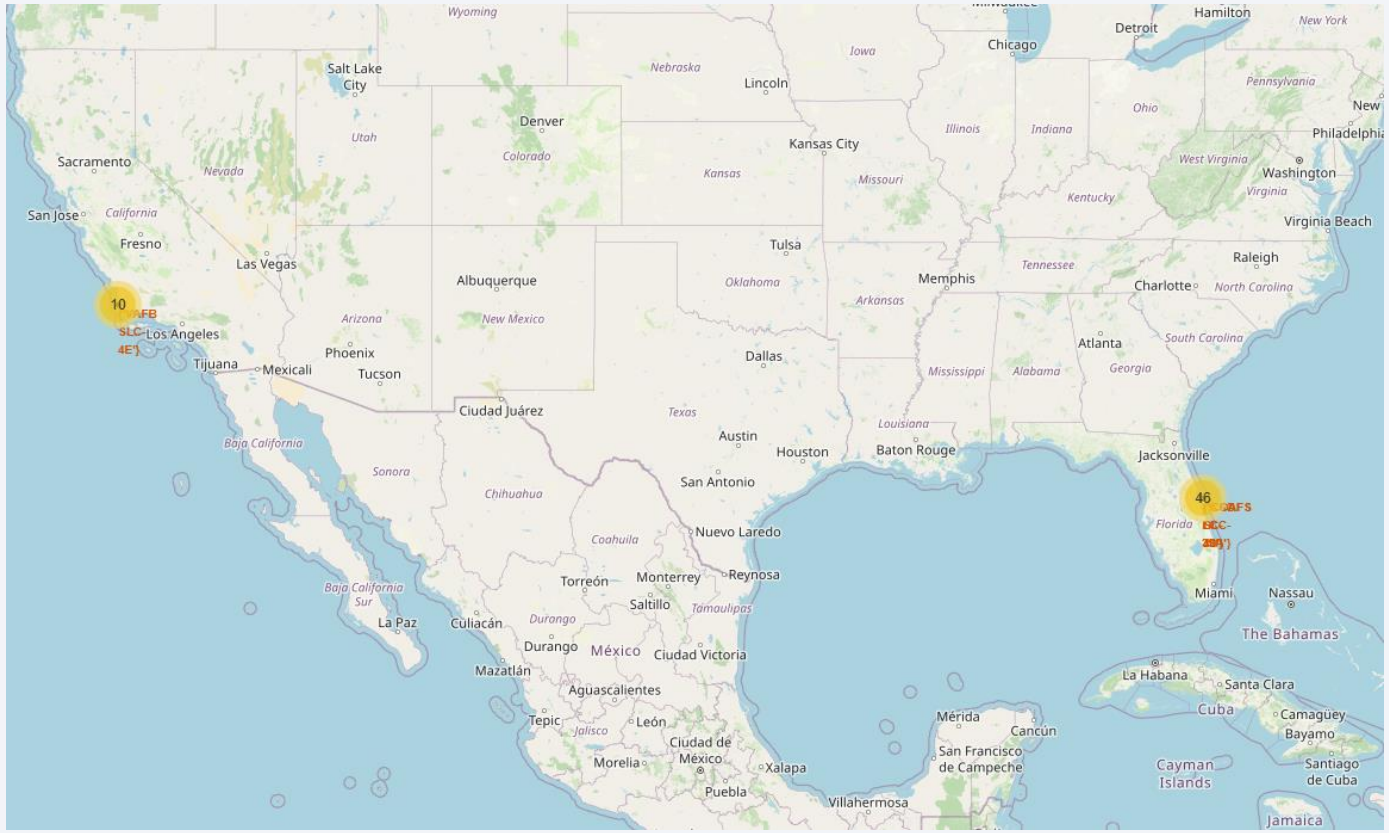
Launch Sites Proximities Analysis

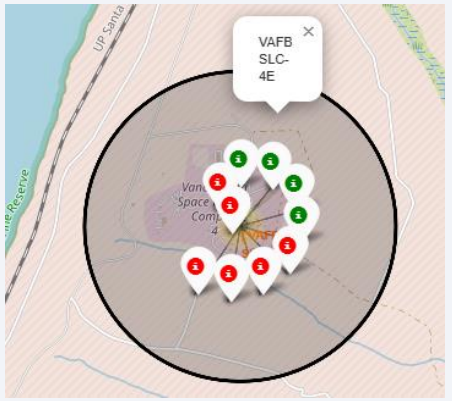
Launch sites on a map

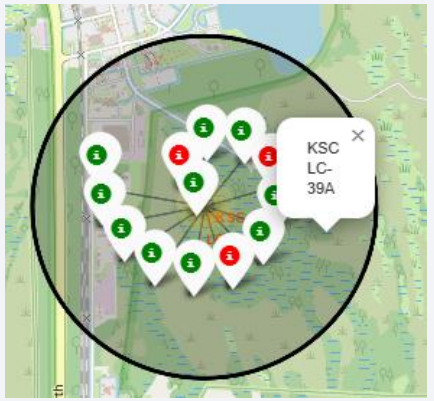


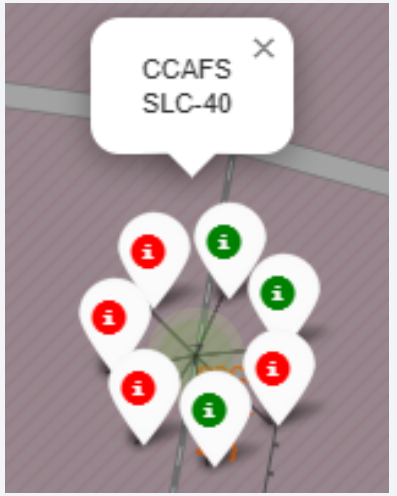
- Findings:
  - There are 4 launch sites.
  - All launch sites are close to the coast. One is close to west coast and the other three are close to the east coast and very close together
  - No particular proximity to the Equator line

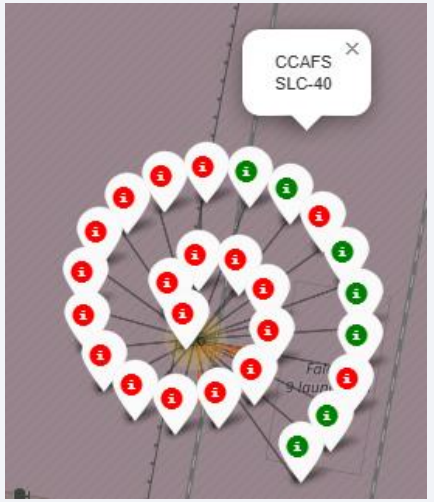
Colour labeled outcomes on a map







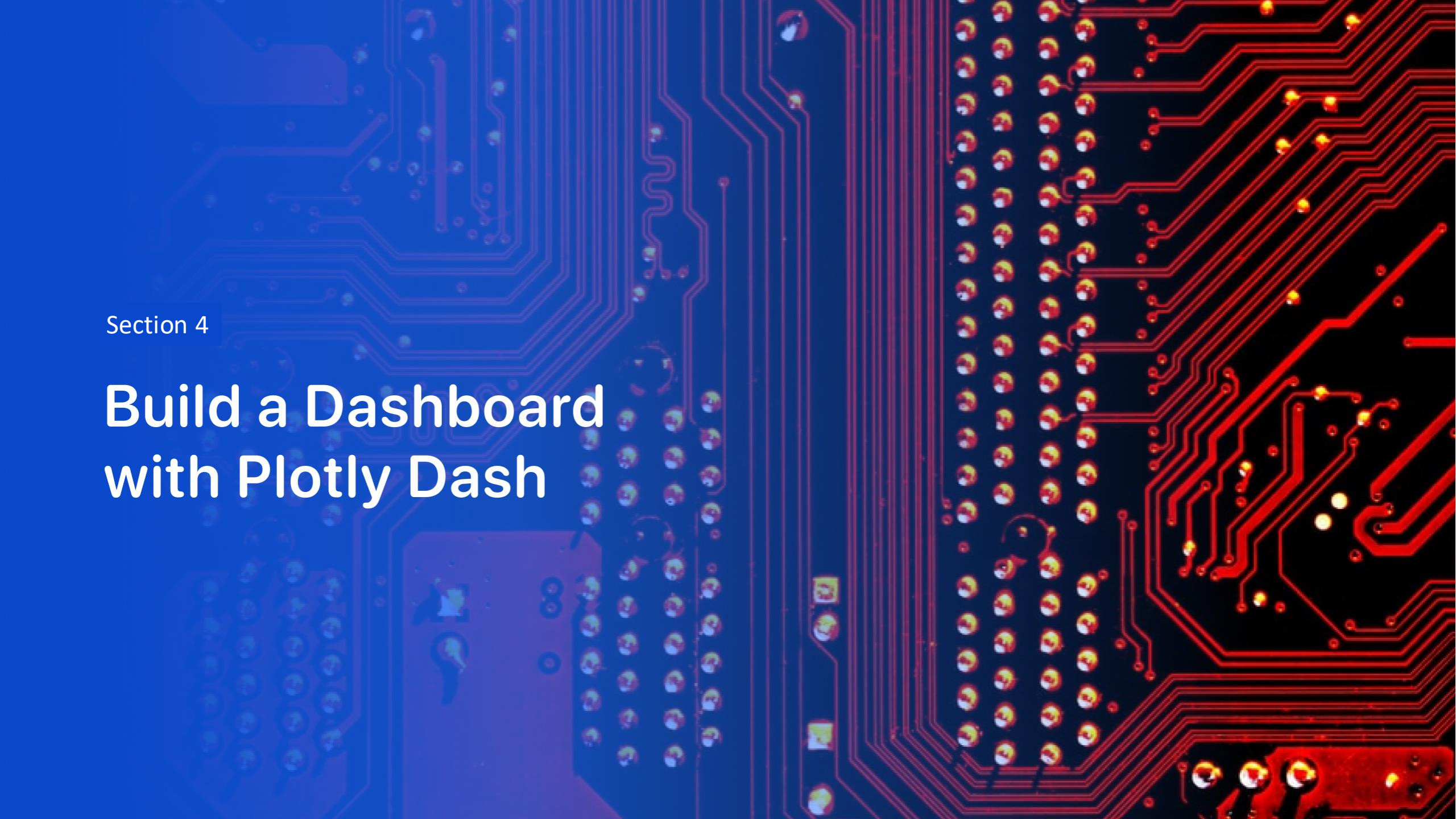




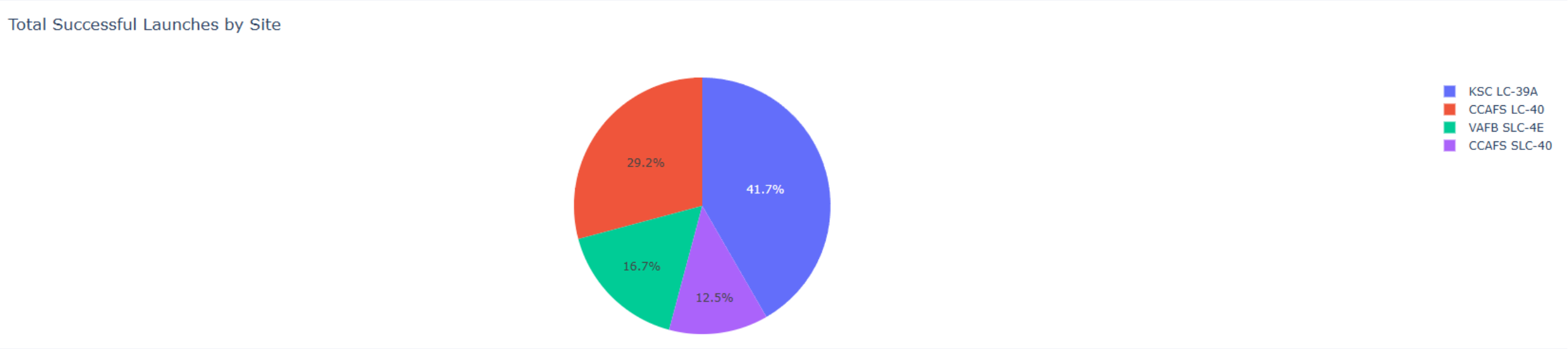
- VAFB SLC-4E launch site has 10 launches with not good success rate
- KSC LC-39A site has good success rate
- CCAFS SLC-40 site has very low success rate

Distance between launch site and its proximities

- CCAFS SLC-40 site distance from the coastline is approximately 0.90km

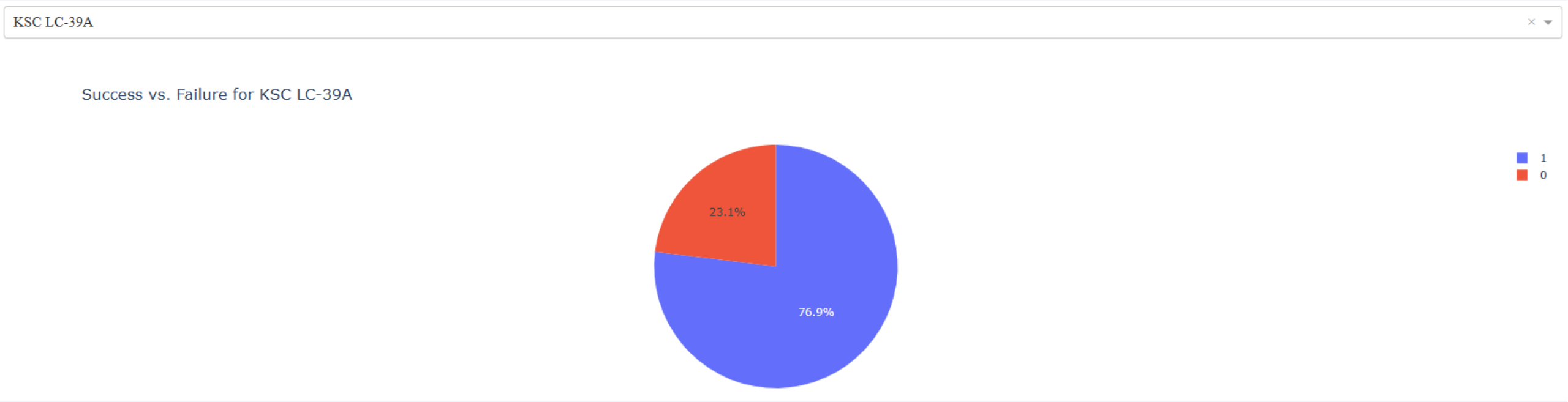

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



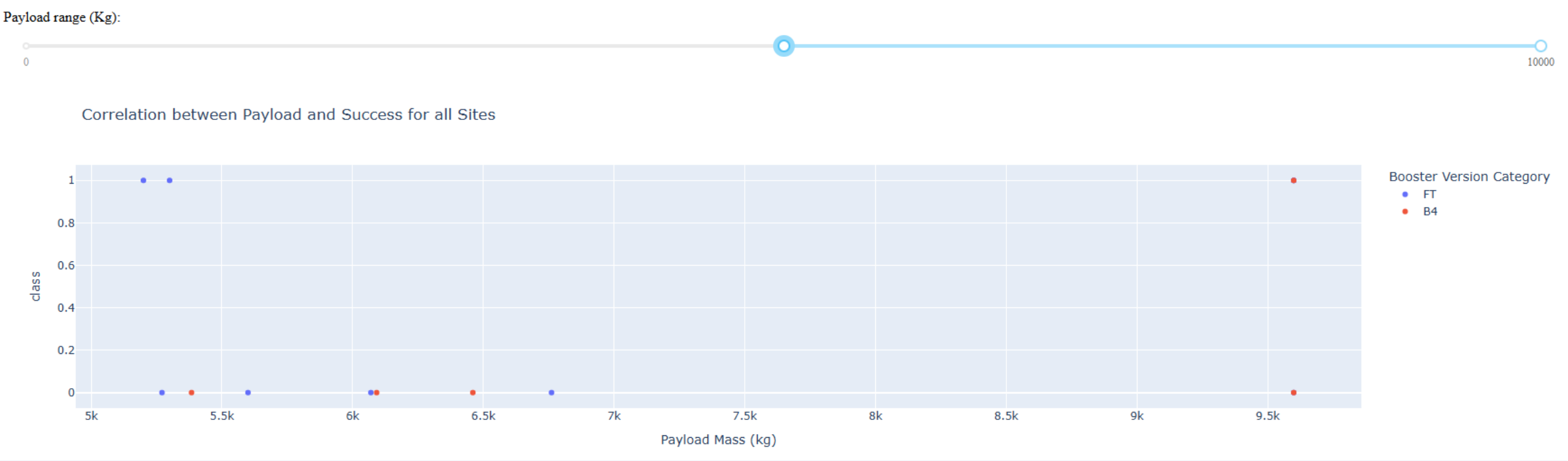
- Findings:
  - KSC LC-39A has the most successful launches, more than 40% of all
  - CCAFS SLC-40 has the least successful launches of all, around 12.5% of them

Highest Success Rate Launch Site



- Findings:
  - KSC LC-39A has also the highest success rate with 76.9% of launches from this site recovering the booster

Correlation between Payload (>6000kg) and Success



- Findings:
  - Only FT and B4 boosters weigh 5000kg or more
  - FT boosters have higher success rate than B4 ones. B4s are rarely recovered (20%)
  - Only two boosters were above 7000kg, both were B4 and only one recovered

Section 5

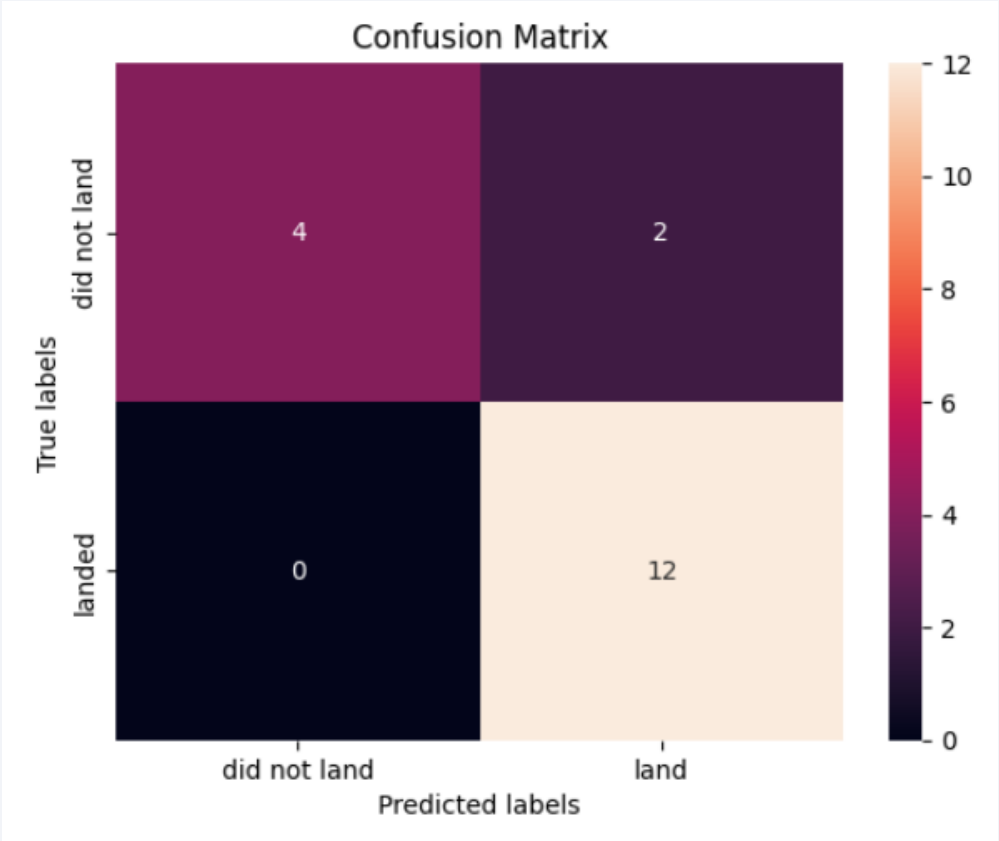
Predictive Analysis (Classification)

Classification Accuracy

- Accuracy results:
  - Logistic regression: 83.33%
  - SVM: 83.33%
  - Decision Tree: 88.89%
  - KNN: 83.33%

- Champion model
  - Decision tree: 88.89% accuracy

Confusion Matrix (Decision Tree)



- Findings:
  - Decision tree has TP rate of 100% identifying all successful landings
  - Decision tree identifies 4 out 6 failures. Decision Tree has better accuracy than the other algorithms that identify 3/6 failures.

Conclusions

- Key factors affecting landing success
  - Certain launch sites have significantly higher success rate (e.g. KSC LC-39A)
  - Booster versions play a crucial role – newer models tend to have higher landing success
  - Payload mass affects landing success, with heavier payloads reducing the probability of a successful landing
- Machine Learning Model Performance
  - Amongst the tested models (logistic regression, decision trees, SVM, KNN), the best-performing model achieved 88.89% accuracy
  - Feature importance analysis showed that launch site, booster version, and payload mass were the most influential predictors.
- Business Impact and Future Applications
  - The predictive model provides insights that can help SpaceX optimize launch conditions and reduce mission costs by improving landing success rates
  - The methodology can be extended to other aerospace companies looking to develop reusable rocket technology.

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!