IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Madina Benhamouda
23 March 2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data collected through SpaceX API and web-scraping
  - Data prepared and transformed into a useable subset
  - Model selected trough comparison
- Summary of all results
  - EDA identified the best features to predict successful landings;
  - Machine Learning model predicted if the Falcon 9 first stage will land successfully (0.833).

Introduction

- Project background and context
  - SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.
- Problems you want to find answers
  - We need a model able to predict the success of a landing to calculate the cost of a launch.
  - We also need to identify the best site launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data Dealing with missing values Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuning and evaluation of classification models to ensure the best results

Data Collection

- We collected data sets from SpaceX API and using web scraping techniques, the Falcon9 Wikipedia page information.

Data Collection – SpaceX API

- Data collection with SpaceX REST calls

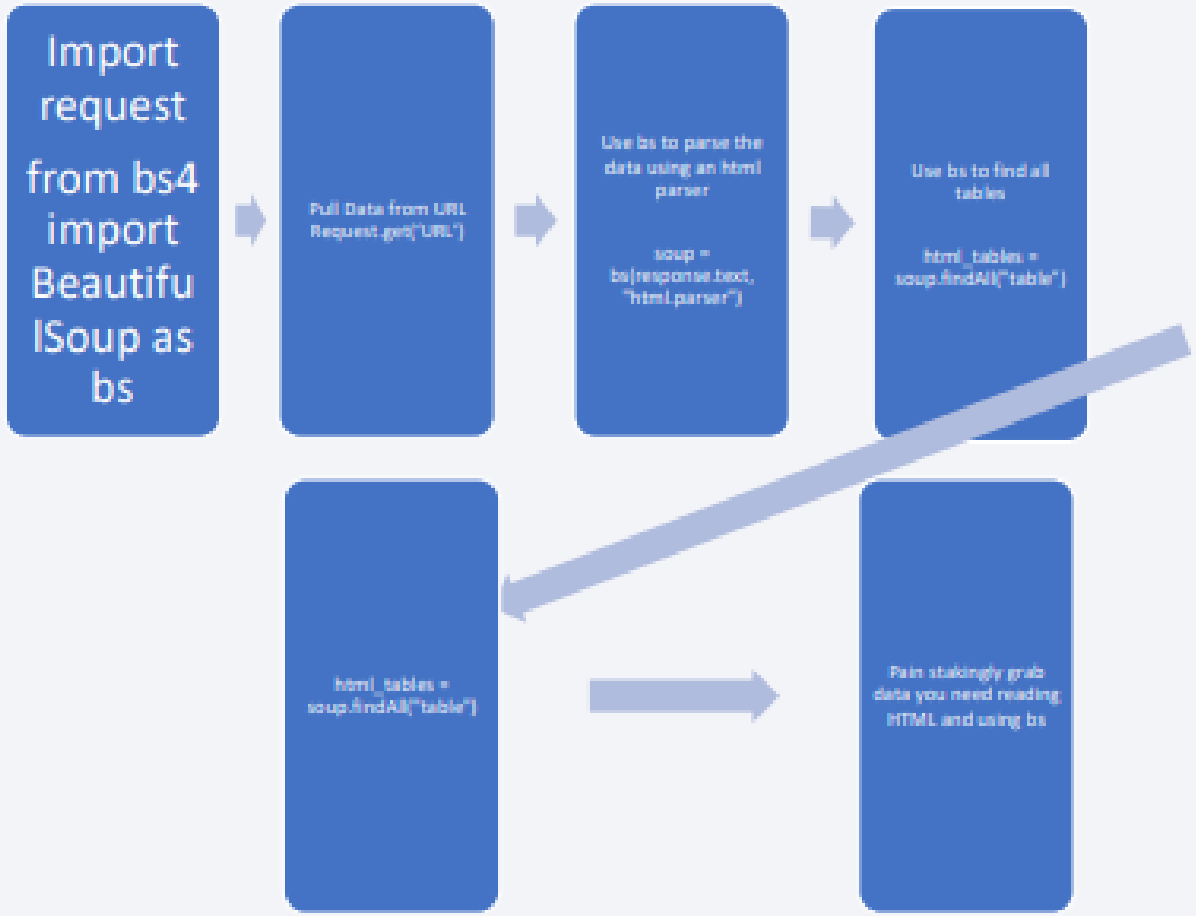
- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1/jupyter-labs-spacex-data-collection-api-v2%20(1).ipynb

Data Collection - Scraping

- Web scraping process

- During this process we scrapped Falcon 9 launch records with BeautifulSoup:
  - Extracting a Falcon 9 launch records HTML table from Wikipedia
  - Parsing the table and convert it into a Pandas data frame

- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1



Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find patterns in the data and determine what would be the label for training supervised models.

- The main contribution was converting landing outcomes into training labels with ‘1’ meaning the booster successfully landed and ‘0’ representing the landing was unsuccessful.

- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1/labs-jupyter-spacex-Data%20wrangling-v2%20(1).ipynb

EDA with Data Visualization

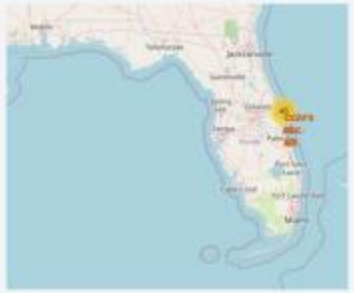
- We rendered scatter plots and bar plots to visualize the relationship between pair of features:
  - Payload Mass vs Flight Number
  - Launch Site vs Flight Number
  - Launch Site vs Payload Mass
  - Orbit vs Flight Number
  - Payload vs Orbit

EDA with SQL

- We used SQL queries to:
  - 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 successful landing outcome in ground pad was achieved.
  - 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.

Build an Interactive Map with Folium

- We used markers, circles, lines and marker clusters on Folium Maps to highlight:
  - The location of launch sites
  - Points of interest, such as the like NASA Johnson Space Center
  - Distance between launch sites and geographical features, like a coastline
  - Cluster of events with close coordinates



- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Added a drop down list to enable launch site selection
- Added a pie chart to show the total successful launches count for all sites
- Added a callback function for `site-dropdown` as input, `success-pie-chart` as output
- Added a callback function for `site-dropdown` and `payload-slider` as inputs, `successpayload-scatter-chart` as output

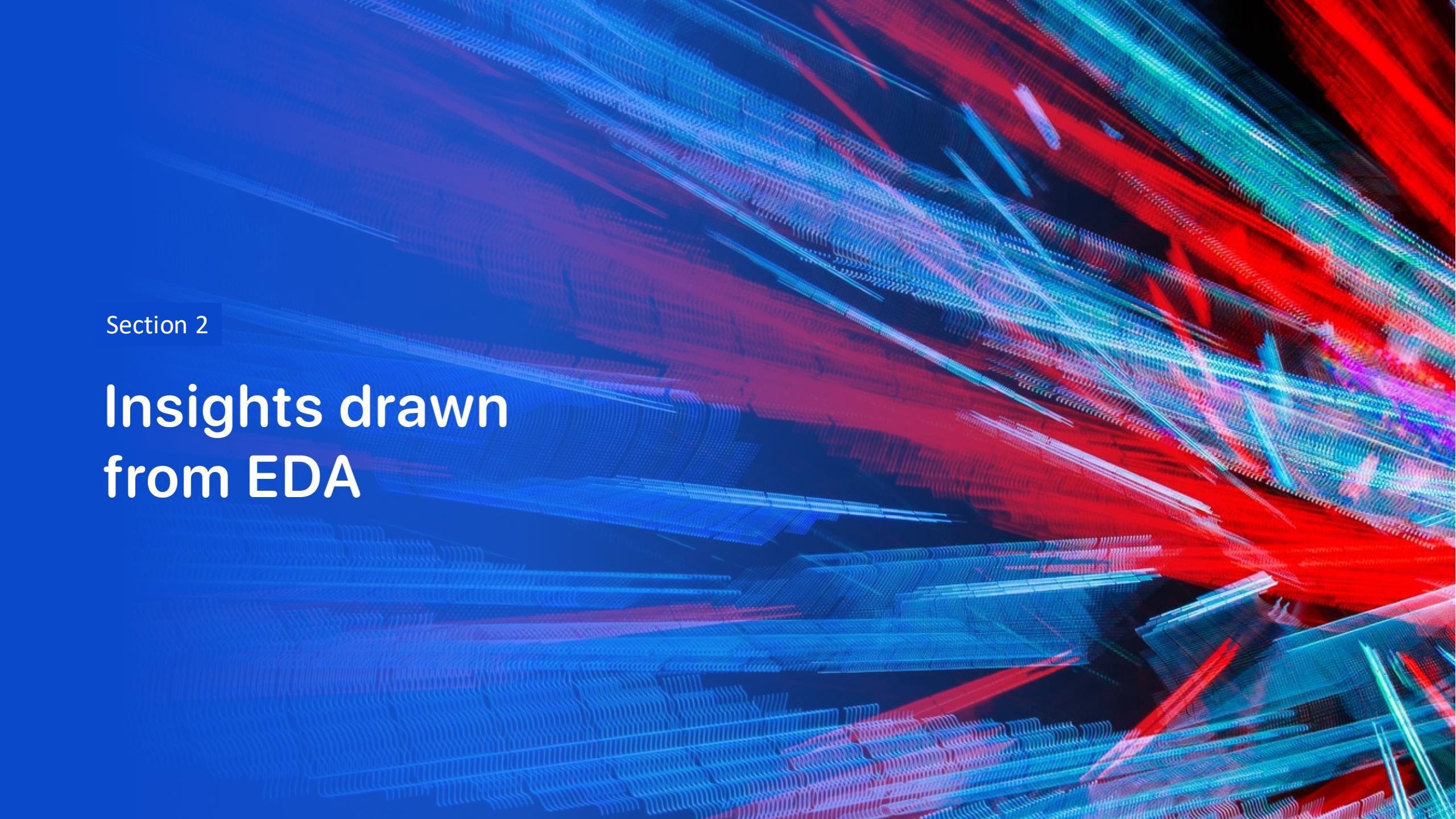
- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1/spacex-dash-app.py.ipynb

Predictive Analysis (Classification)

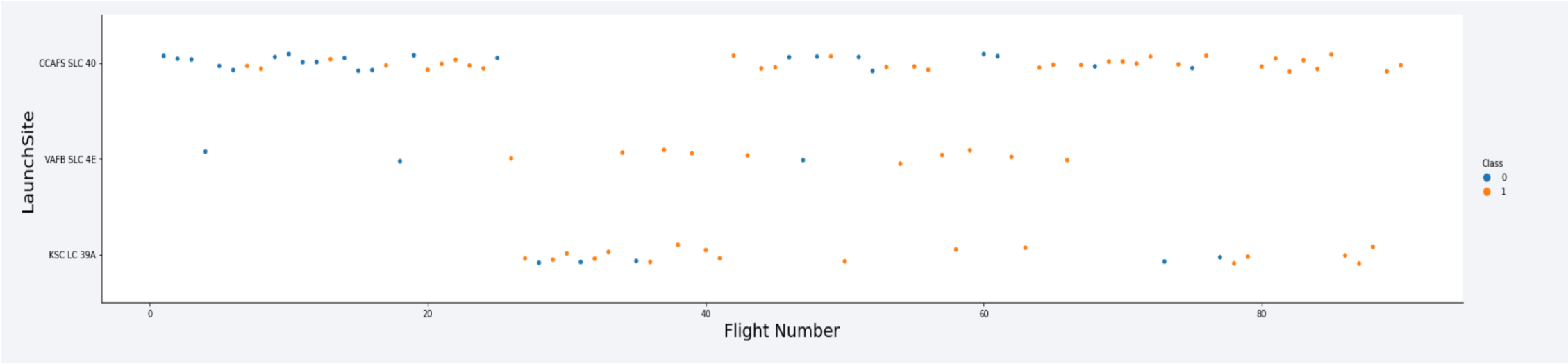
- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- https://github.com/Mad2025/XSpace/blob/ae103c90e40faef4582917be274859b9c8a2a2c1/SpaceX_Machine_Learning_Prediction.ipynb

Results

- Exploratory data analysis results
  - We see that as the flight number increases, the first stage is more
  - likely to land successfully. The payload mass is also important; it
  - seems the more massive the payload, the less likely the first
  - stage will return.
- Predictive analysis result:
  - Our decision tree model performed the best with 0.8333

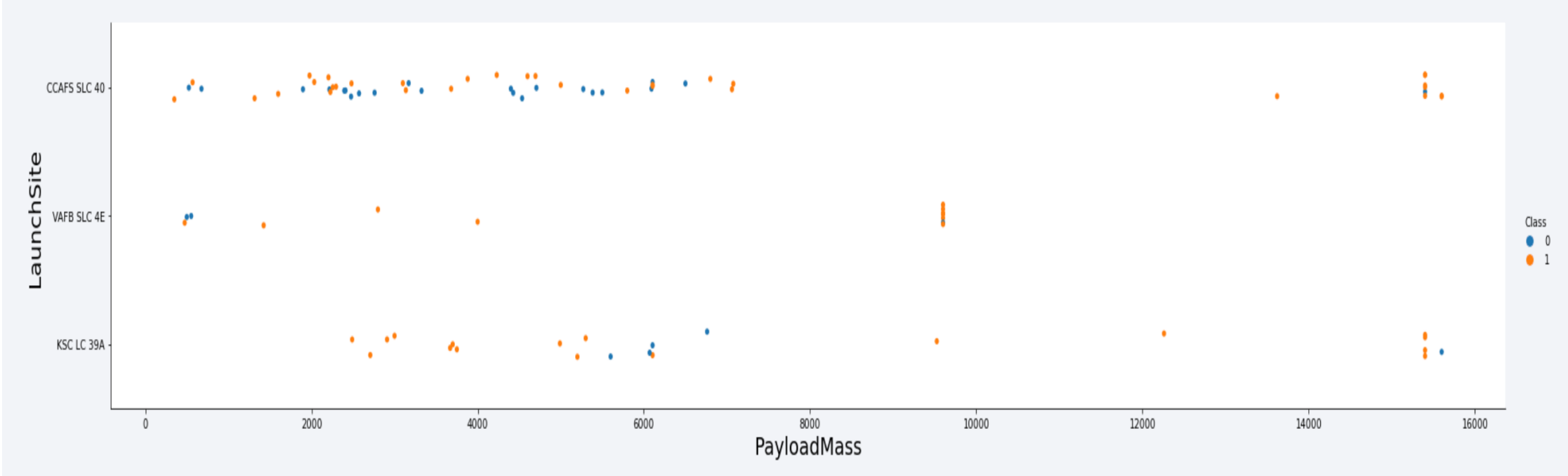

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- According to the plot above, the most successful launch site is CCAF5 SLC 40
- In second place VAFB SLC 4E and third place KSC LC 39A;
- Also, the overall success rate is improving over time.

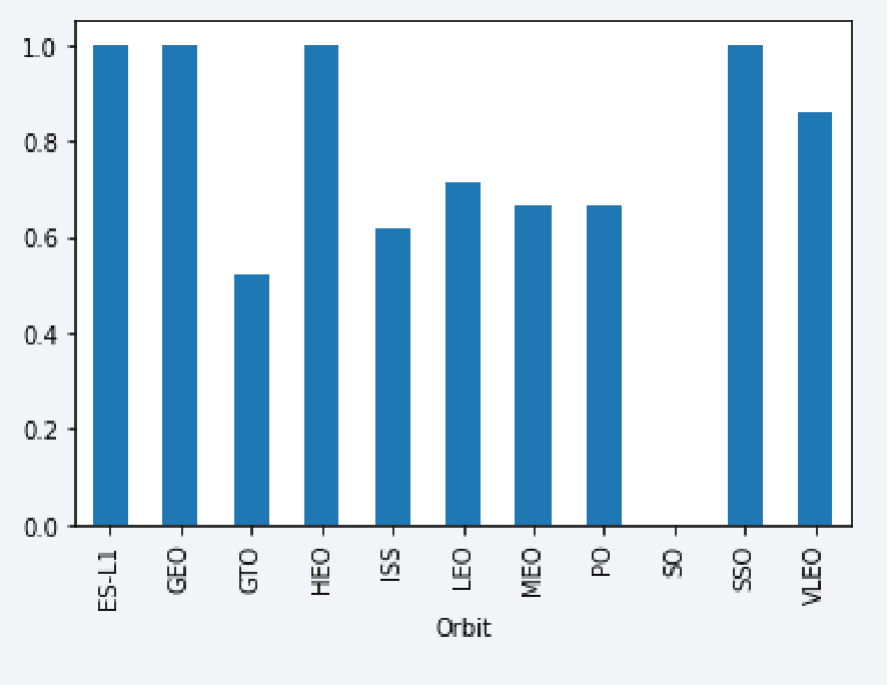
Payload vs. Launch Site



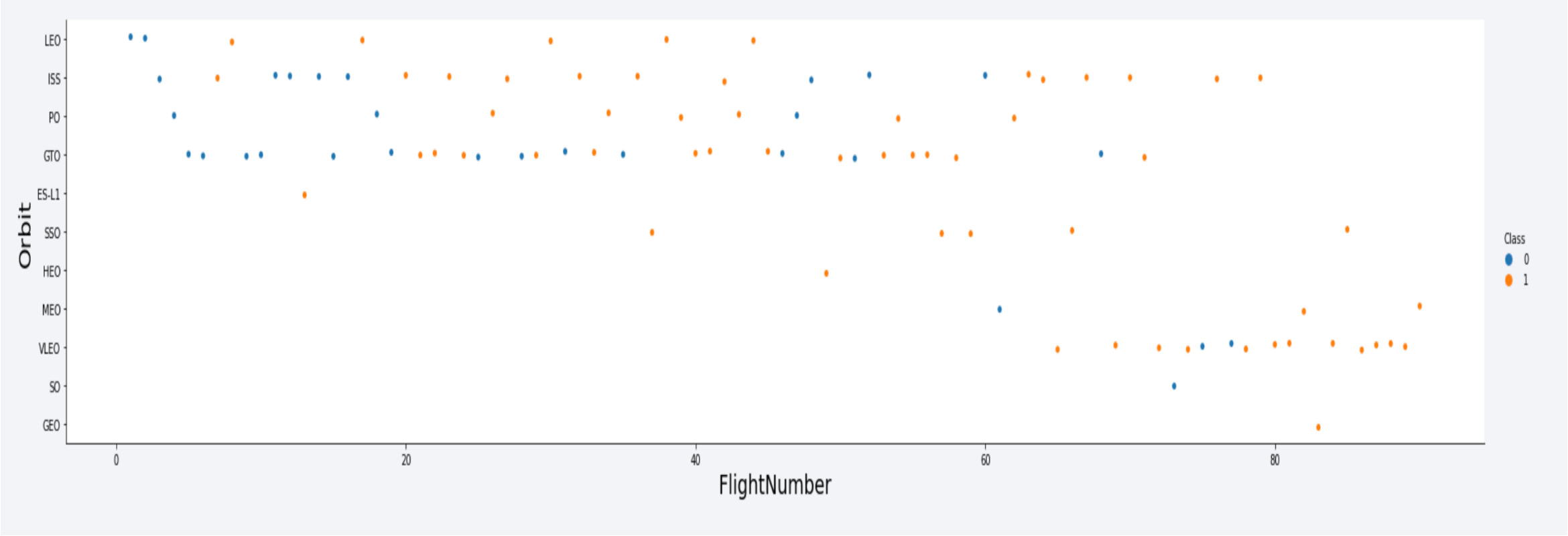
- Higher payloads (over 12,000kg) have a better success rate at CCAFS SLC 40 and KSC LC 39A launch sites

Success Rate vs. Orbit Type

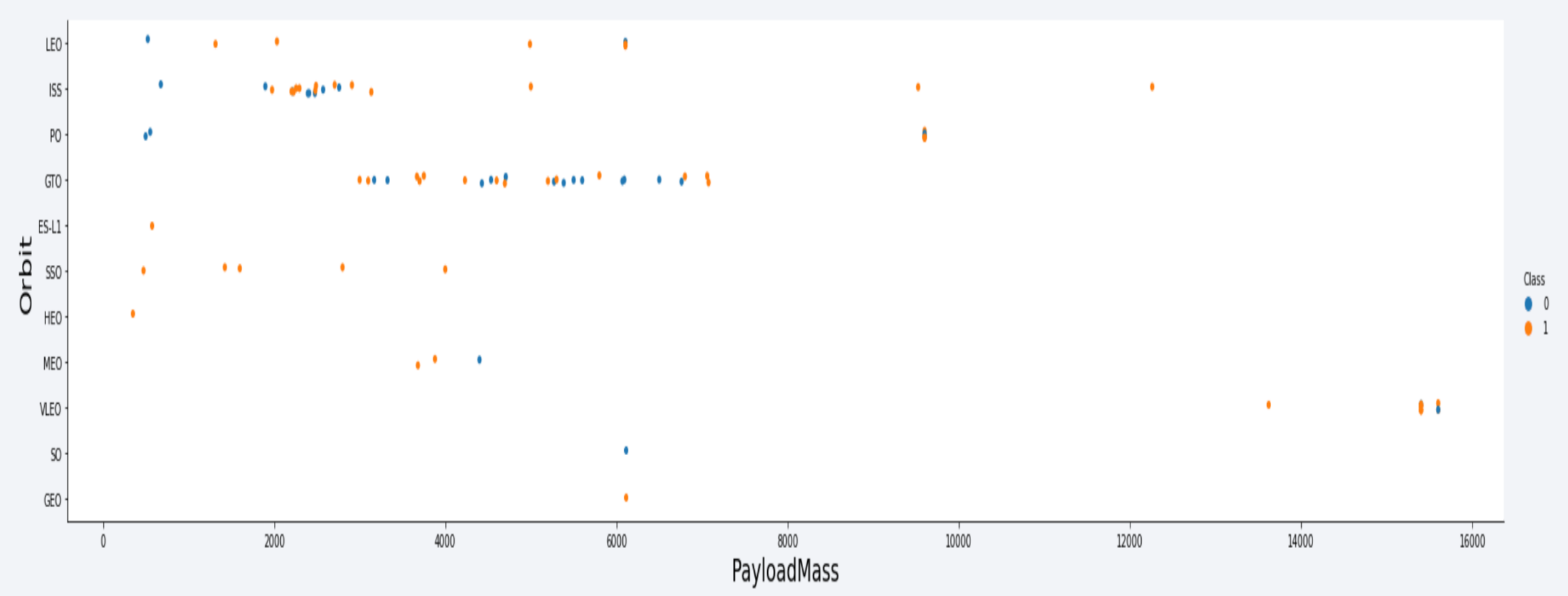
- The following orbits have the highest success rates:
  - ES-L1
  - GEO
  - HEO
  - SSO



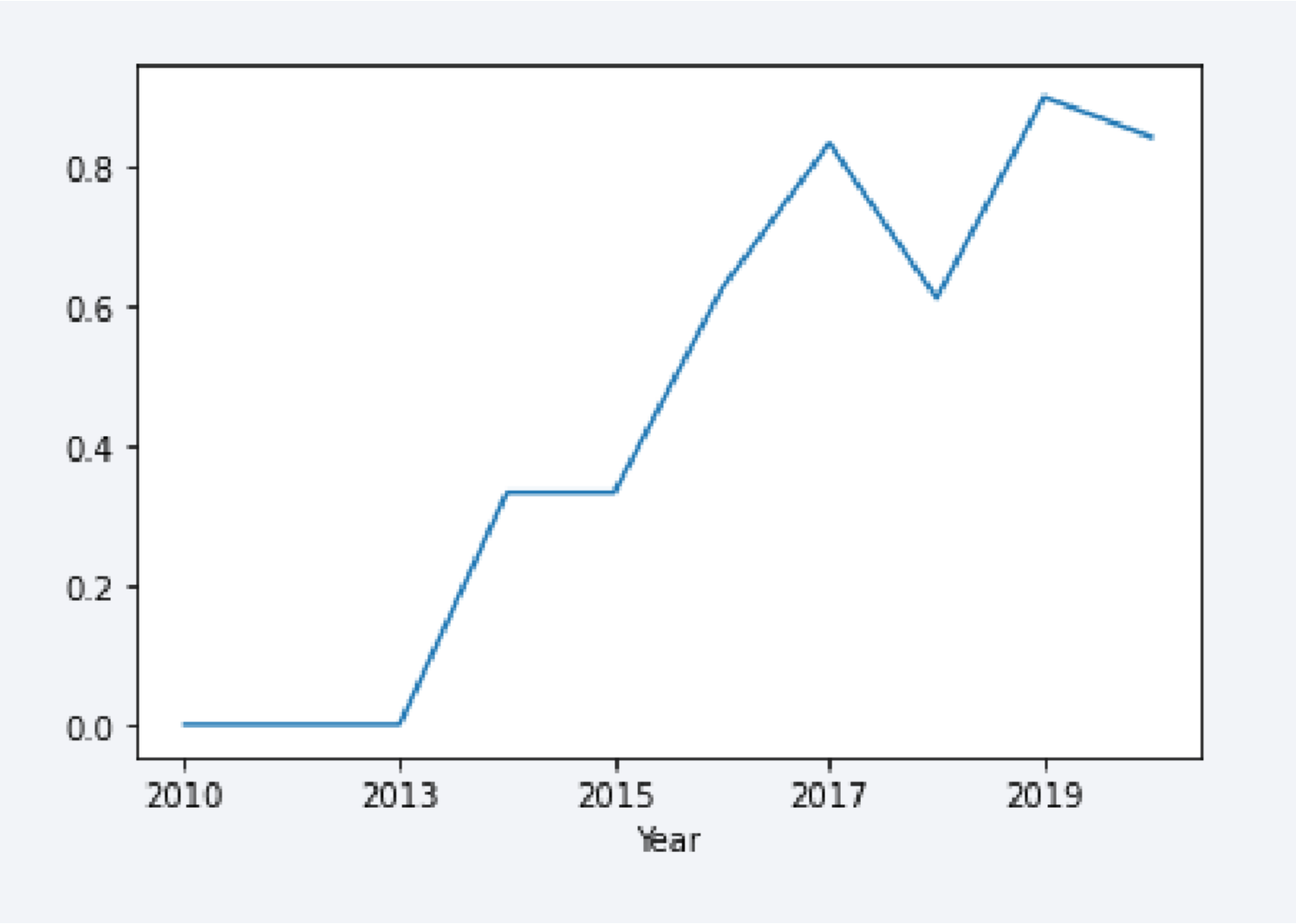
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- According to data, there are four launch sites:

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- They are obtained by selecting unique occurrences of “launch_site” values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin 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	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Here we can see five samples of Cape Canaveral launches.

Total Payload Mass

- Total payload carried by boosters from NASA:

Total Payload (kg)
111.268

- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

Avg Payload (kg)
2.928

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

Min Date
2017-01-05

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1031.2
F9 FT B1022

- Selecting distinct booster versions according to the filters above, these 2 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	44
Success (payload status unclear)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- Boosters which have carried the maximum payload mass

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered 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

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40

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

- Ranking of all landing outcomes between the date 2010-06-04 and 2017-03-20:

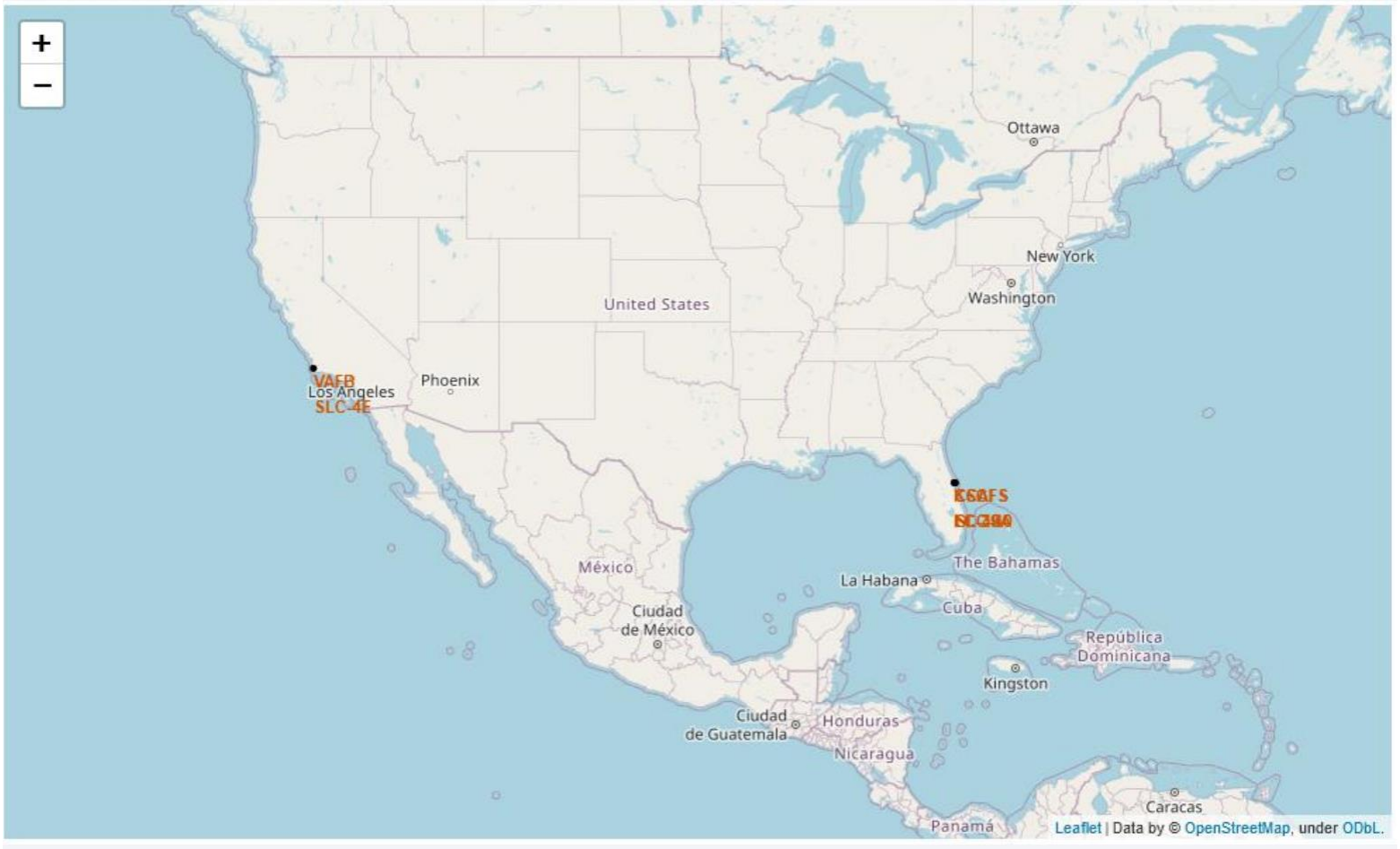
Landing Outcome	Occurrences
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

- This view of data alerts us that “No attempt” must be taken into account.

Section 3

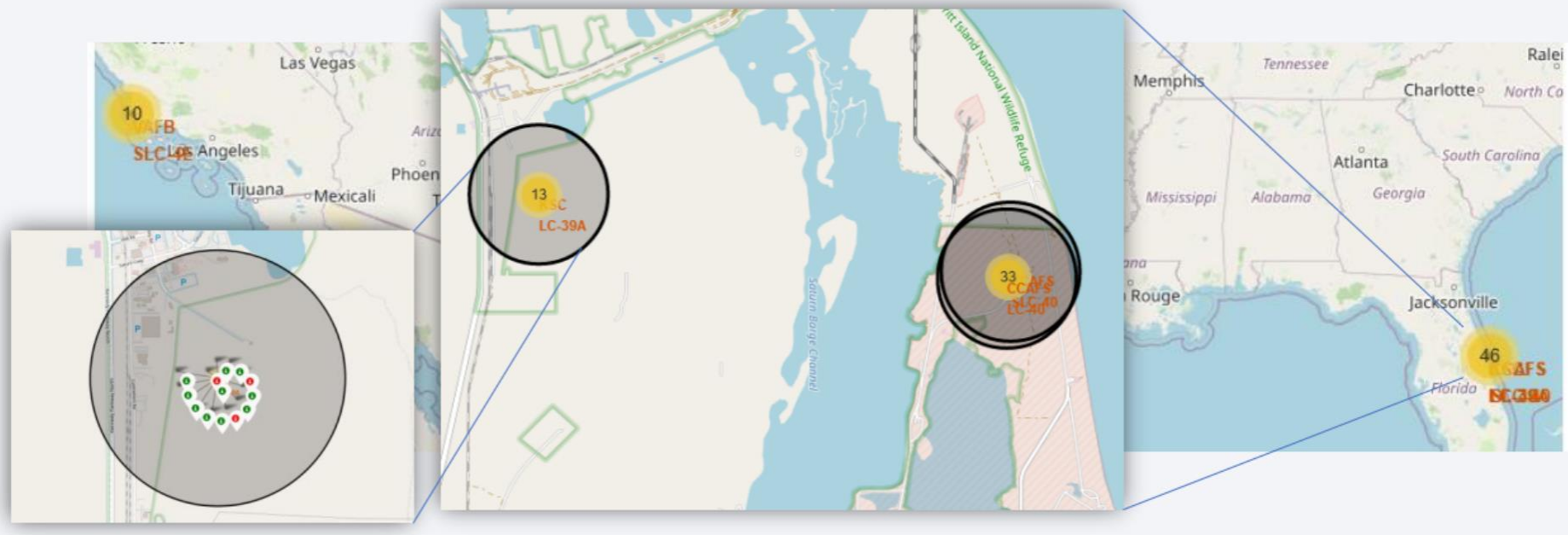
Launch Sites Proximities Analysis

All launch sites

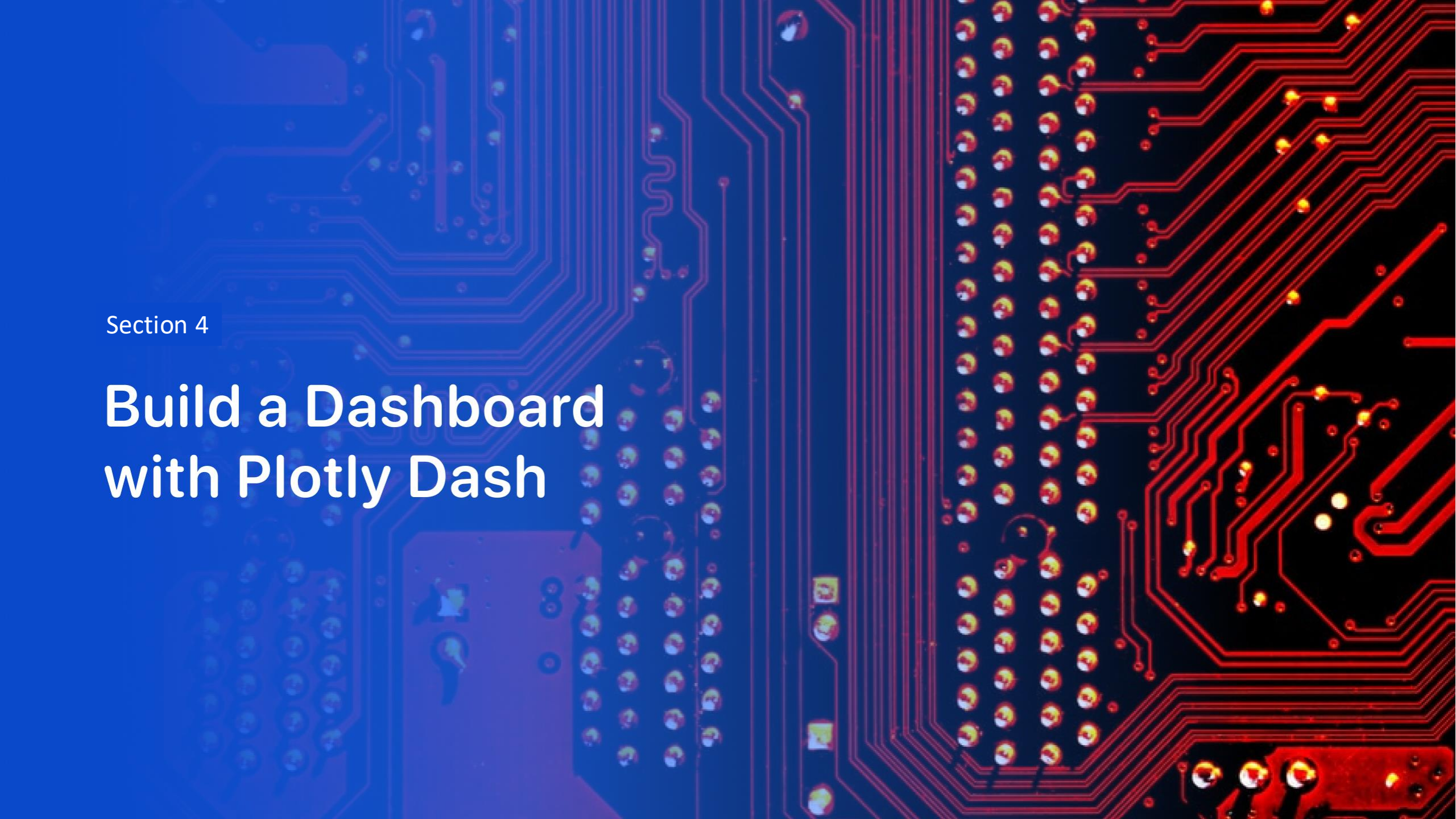


Launch outcome by launch site

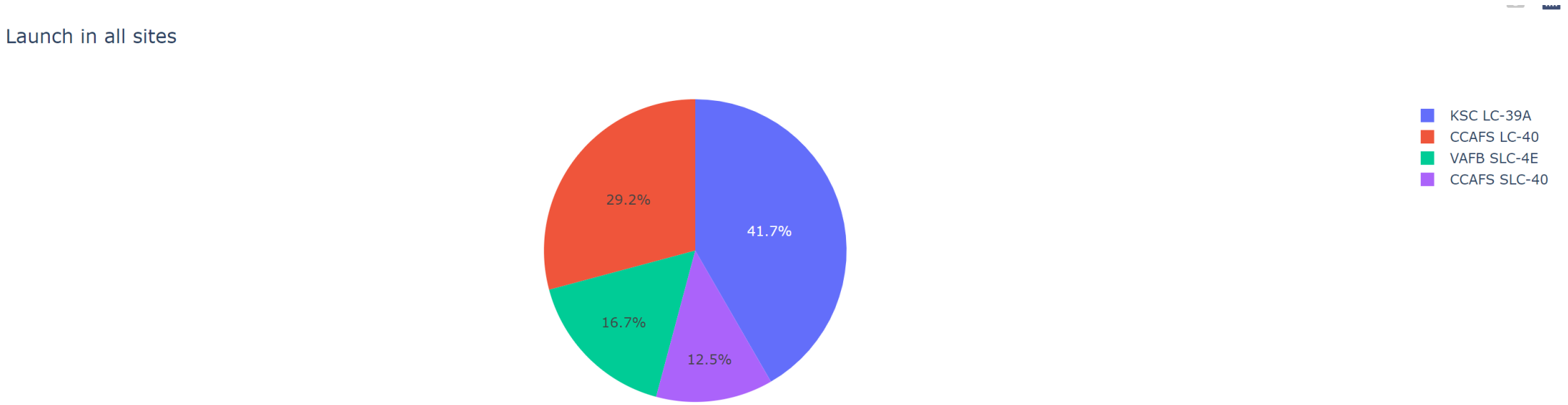
- Example of KSC LC-39A launch site launch outcomes



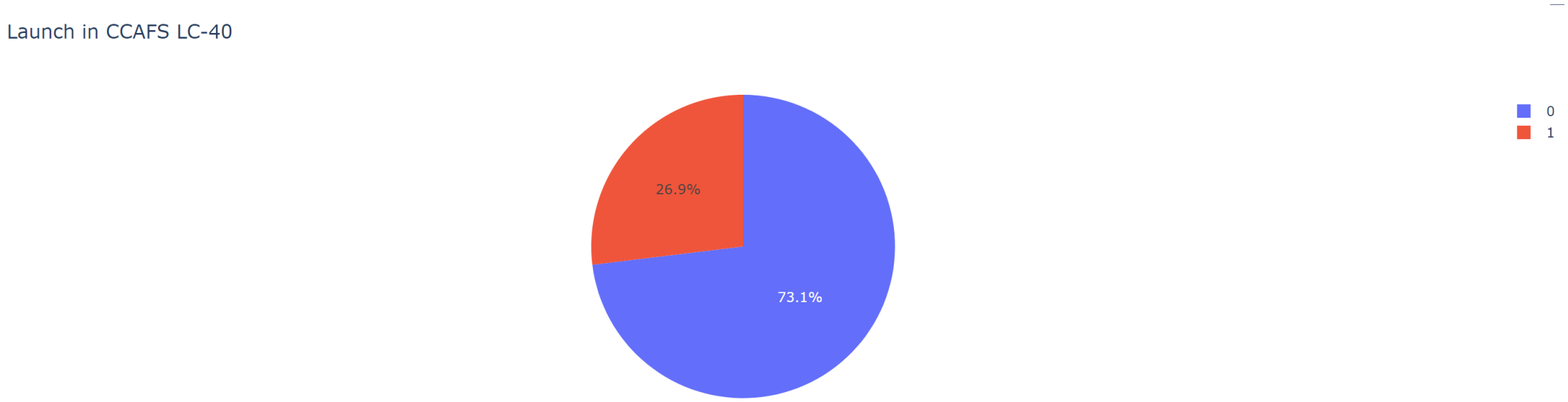
- Green markers indicate successful and red ones indicate failure.


Section 4
Build a Dashboard with Plotly Dash

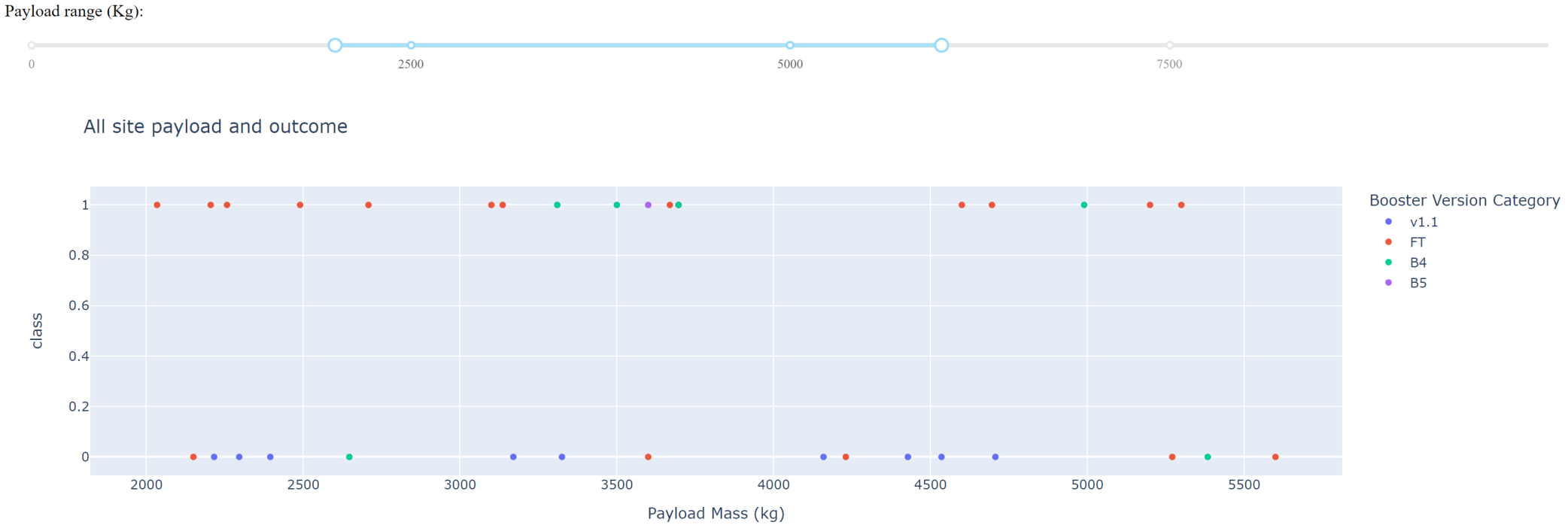
Successful Launches by Site



Launch Success Ratio for KSC LC-39A



Payload vs Launch Outcome

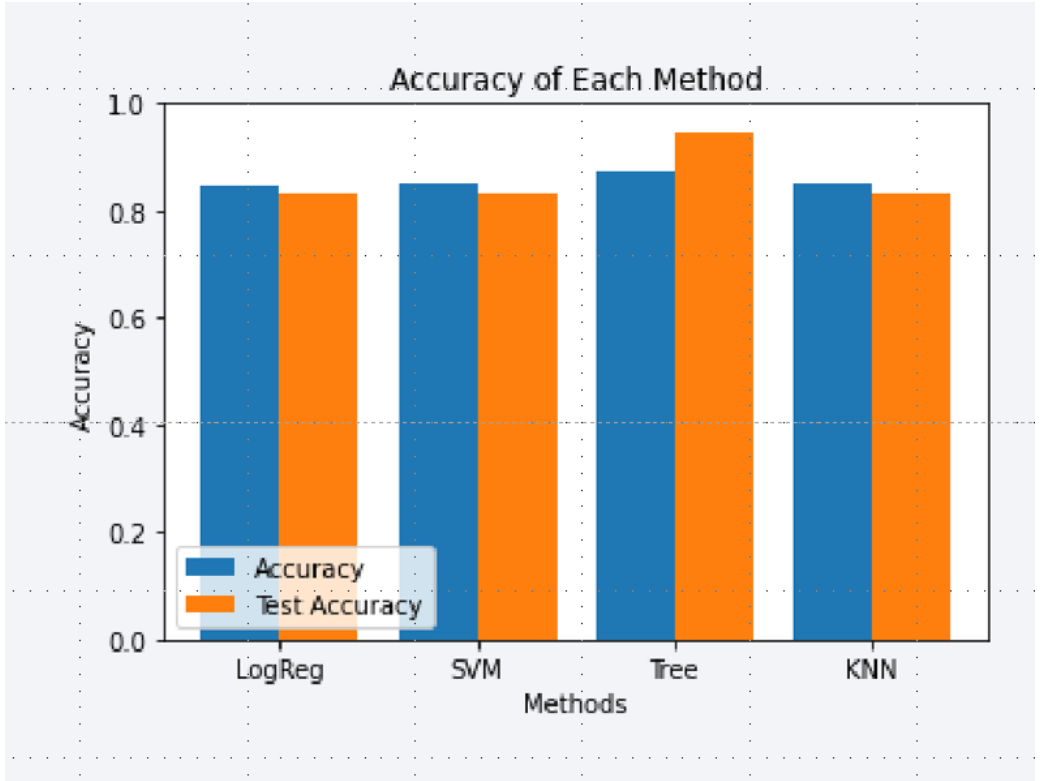


Section 5

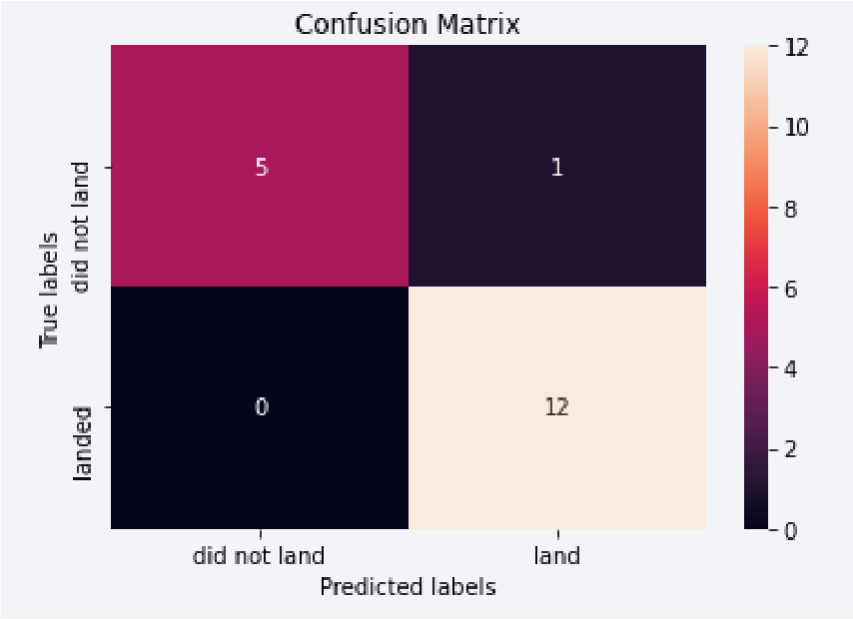
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside
- The classifier model with the highest accuracy is the Decision Tree



Confusion Matrix



The Decision Tree confusion matrix shows the largest number of true positive and true negative compared to other models.

Conclusions

- KSC LC-39A is the launch site with the highest success rate
- Higher payload launches are more successful
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- SpaceX API https://api.spacexdata.com

Thank you!