Winning Space Race with Data Science

Jeff Anga
August 2nd 2022

Outline

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

Executive Summary

- **Summary of methodologies**

  -Data Understanding

  -Data Collection

  -Data Wrangling and Preparation

  -Exploratory Data Analysis using Data Visualizations and Interactive Visual Analytics

  -Predictive Analysis using Machine Learning (Classification)

- **Summary of all results**

Data on Rocket Launches were collected from the Space X API and public source like Wikipedia

Exploratory Data Analysis (EDA) gave more insight into the data and what feature we can use to predict successful launches and landings.

Predictive Analysis using Machine Learning predicted which

Introduction

- **Project Background and context**

SpaceX saves upward of 100 million dollars rocket launches compared to other providers, much of the savings is because SpaceX can reuse the first stage of a rocket launch. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information will be used if an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems you want to find answers**
  - Reliable estimators for total cost of launches and landings
  - What influences successful launches and landings
  - What conditions are necessary to ensure the highest successful landings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the Space X API and from Wikipedia using Web scraping
- Perform data wrangling
  - Data was processed and transformed using One Hot Encoding
- 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

- Data was collected from the Space X API
- https://api.spacexdata.com/v4/launches/past
- Data was also collected from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches"
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

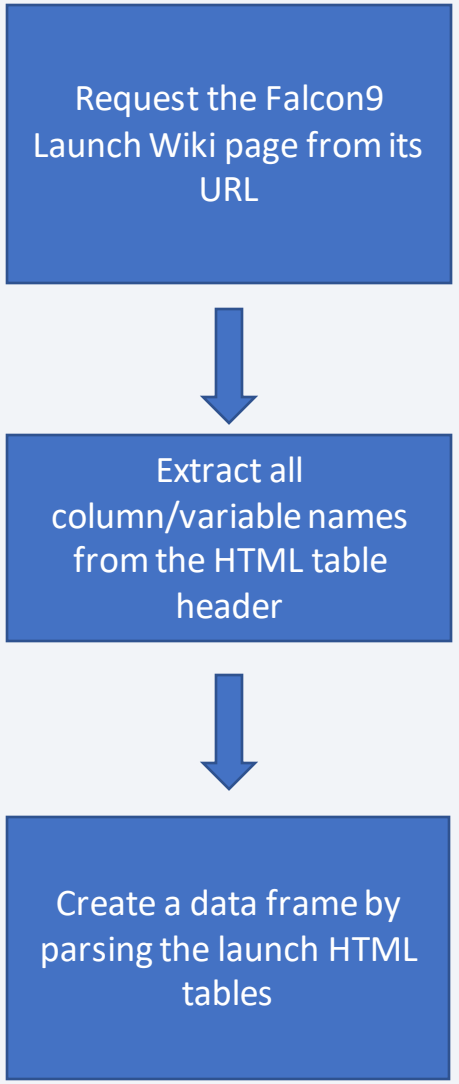
Data Collection – SpaceX API

- Data was obtained and collected from the SpaceX public API using request
- Data was then parse and stored in a .csv file
- Filtered the data and replaced missing values

- Data Collection API

Data Collection - Scraping

- Data was obtained and collected from the Wikipedia page on Falcon 9 launches using beautiful soup.
- Data was then parse and stored in a data frame then exported to a .csv file



- Data Collection- Web Scraping

Data Wrangling

- Performed Exploratory Data Analysis (EDA) to find some patterns in the data and then determined Training Labels.
- The number of launches on each site, the number and occurrence of each orbit and mission outcome per orbit type were calculated.
- Finally a landing outcome label from the outcome column was created.

- Data Wrangling

EDA with Data Visualization

- Performed Exploratory Data Analysis and Feature Engineering
- Several scatter plots were used to show how two variables would affect the launch outcome. The pairs used for the scatterplots are: Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload vs Launch Site, Flight Number vs Orbit type, Payload vs Orbit type
- A bar chart was used to visually check for a relationship between the success rate and Orbit type
- Finally a line chart was used to get the average launch success trend
- EDA with Data Visualizations

EDA with SQL

- Several SQL queries were performed, namely:
- 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
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass.
- The failed landing outcomes in drone ship, their booster versions, and launch site names for 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

- EDA with SQL

Build an Interactive Map with Folium

- Map objects such as markers, circles, lines were created and added to a folium map
- Markers were used to show the launch sites
- Circles show the highlighted areas around specific coordinates
- Clusters were used to show a point with many markers having the same coordinate
- Lines showed the distances between coordinates

- Interactive Map with Folium

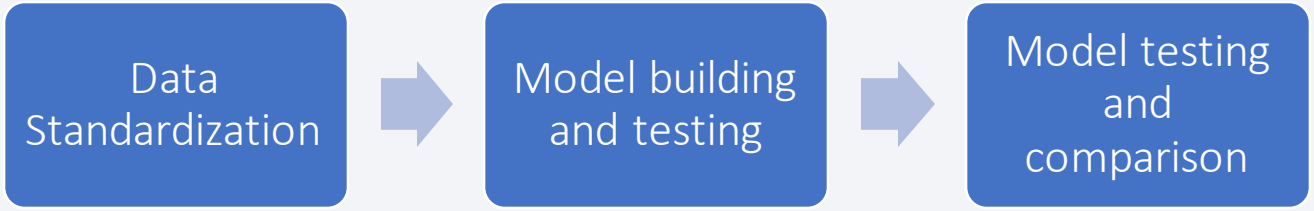
Build a Dashboard with Plotly Dash

- A pie chart was added to the dashboard to show the percentage of launches by a particular launch site or all the launch sites . The pie chart was used because it could effectively show the quantity and proportions of the launches per site visually.
- A scatter plot was added to the dashboard to show the relationship between payload mass and the outcome for different booster versions. This plot was used because of how easy it is to observe the relationship between the two variables.

- Plotly Dash

Predictive Analysis (Classification)

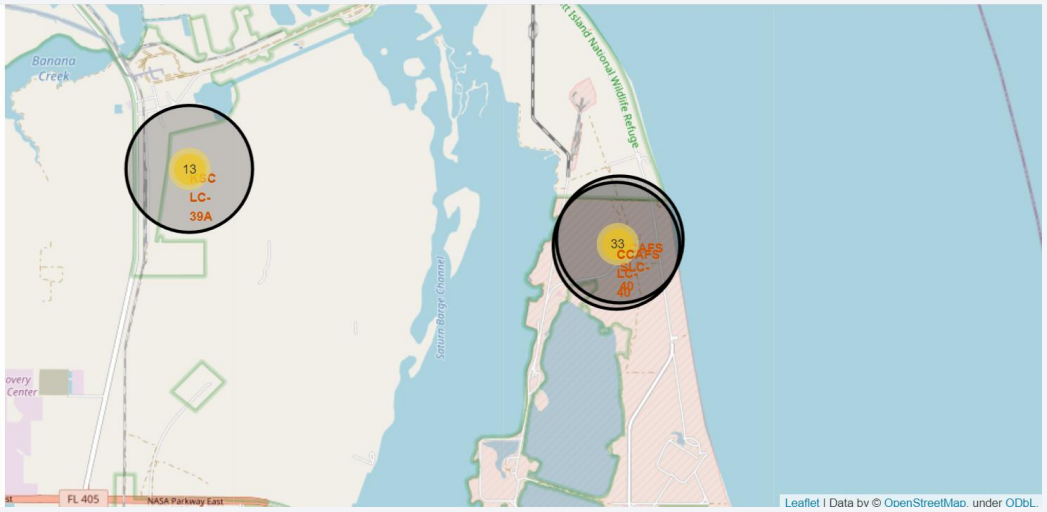
- Several classification models were considered: Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbours (KNN)
- Data was first prepared and then standardized and transformed to get it ready for the models. The data was then split into training and test sets.
- Each of the four classification models were then used on the data and used to train the dataset. The best hyperparameters for each model was gotten and the accuracy of the models were then calculated using the test data.
- A confusion matrix was plotted for each model to gain more insight on the accuracy of the models.
- Finally, the best classification model was now determined using the accuracy scores.



- Predictive Analysis

Results

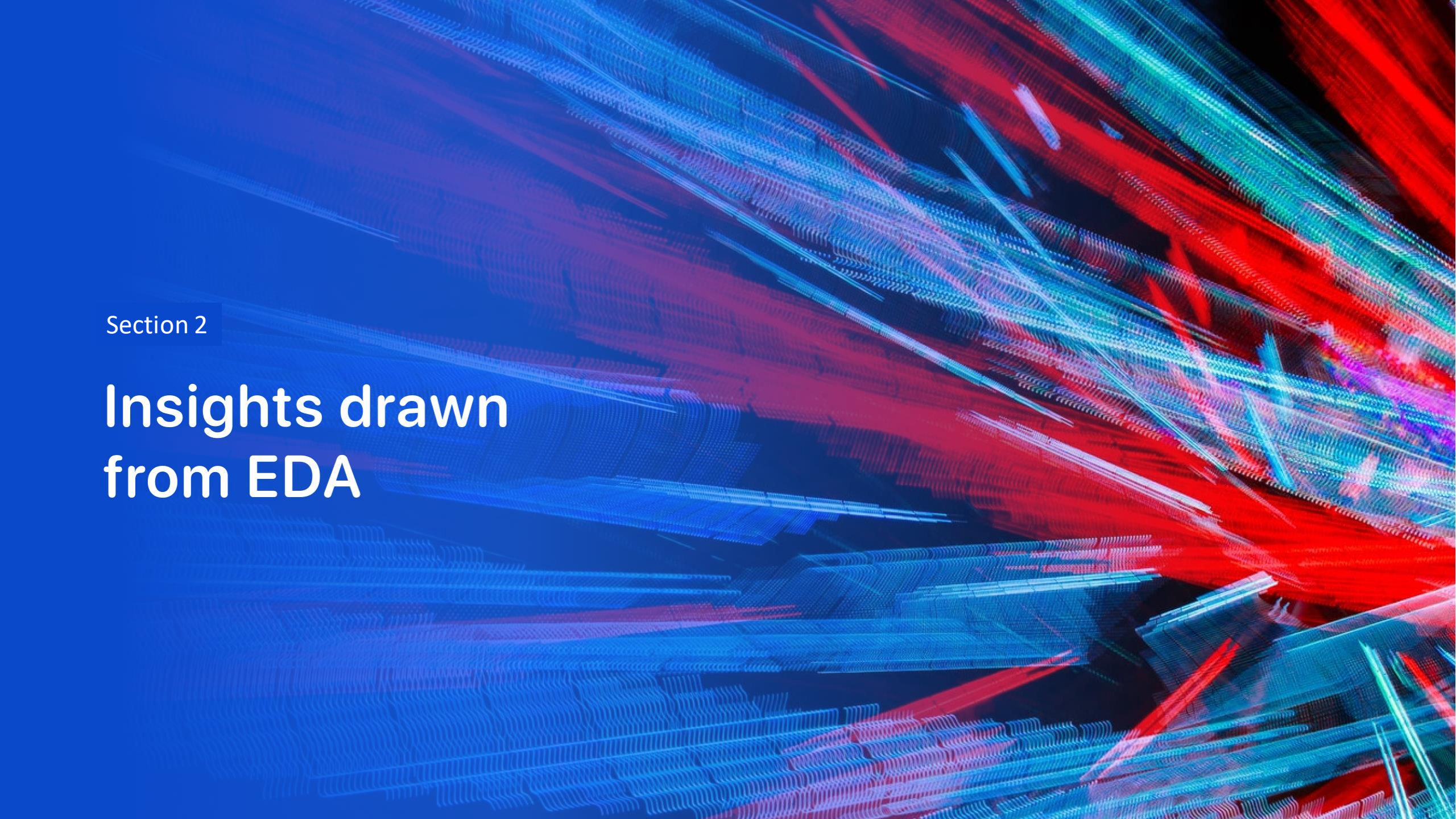
- Exploratory data analysis results showed that:
- Space X used 4 different launch sites, all by the sea.
- The number of successful landing outcomes improved over time
- Mission outcomes were nearly always successful.



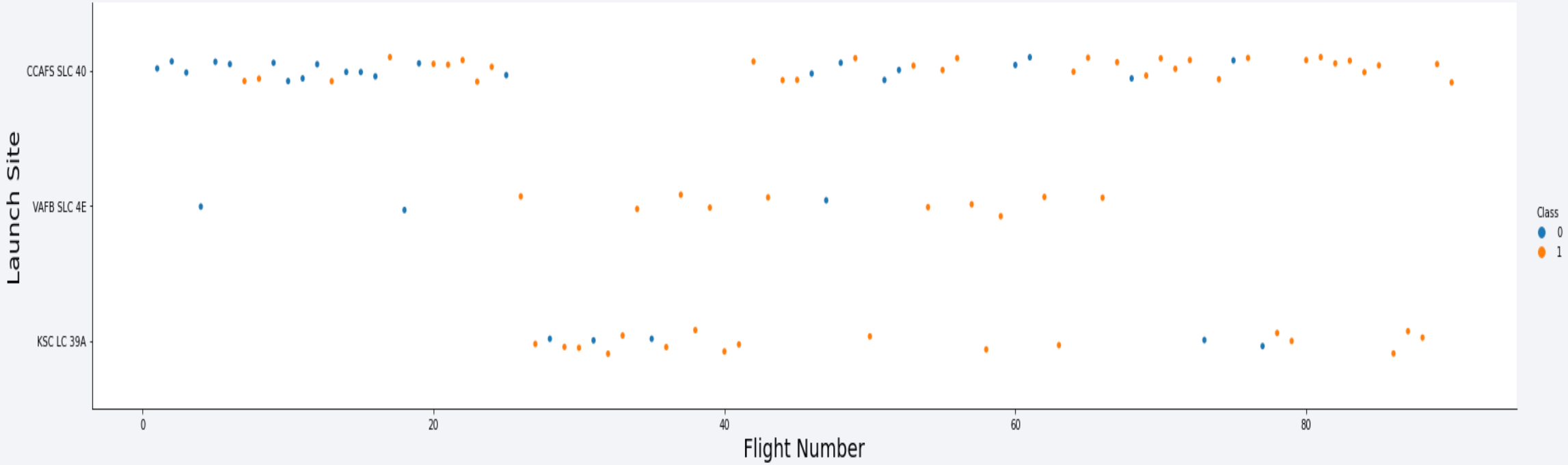
- Predictive analysis results showed that the Decision Tree is the best Classification model to predict successful landings with an accuracy of over 88%

Section 2

Insights drawn from EDA

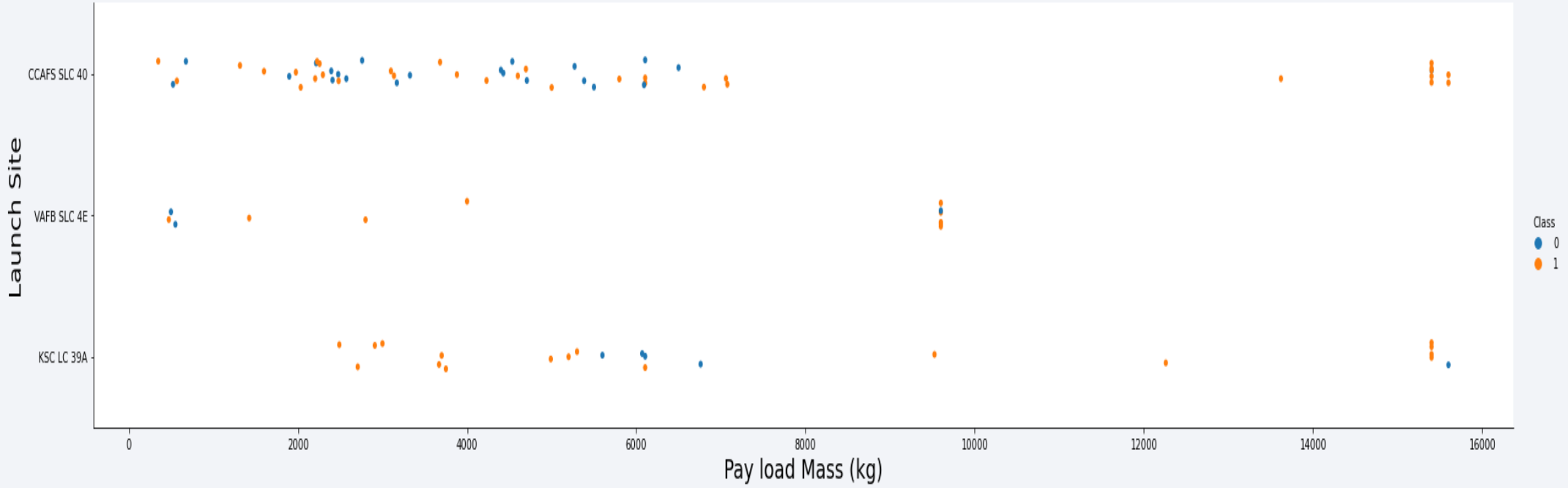


Flight Number vs. Launch Site



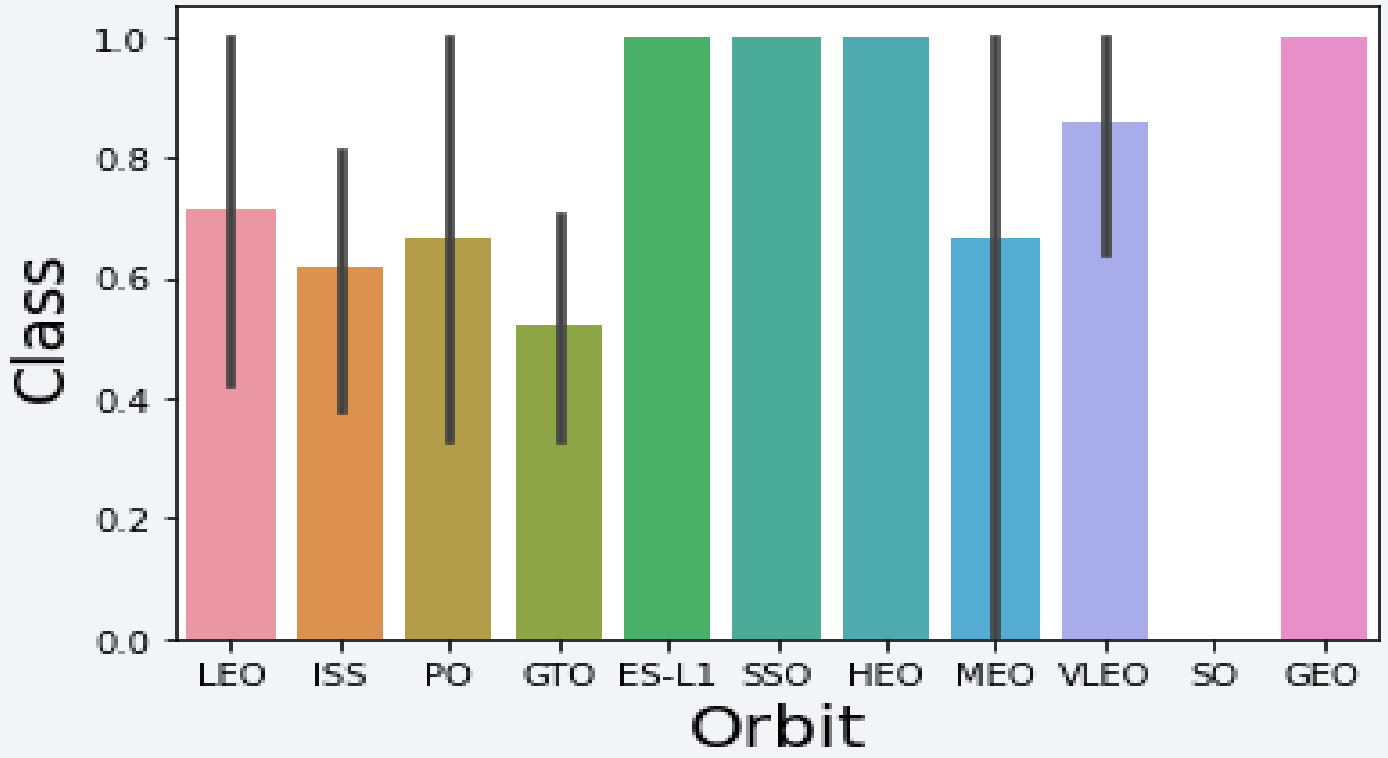
- We can see that CCAF5 SLC 40 has the most number of flight attempted. It also has the highest successful launches
- We can also see that VAFB SLC 4E has the least attempts but also has an over 80% success rate. The highest among the three sites

Payload vs. Launch Site



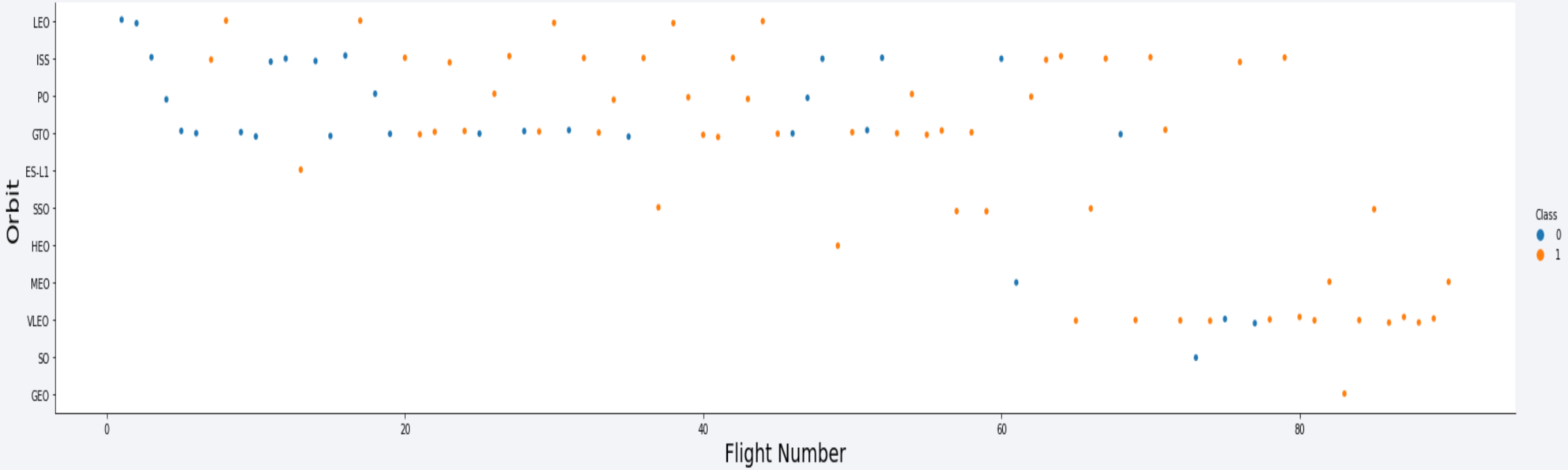
- Payloads over 9,000 kg have an near perfect success rate

Success Rate vs. Orbit Type



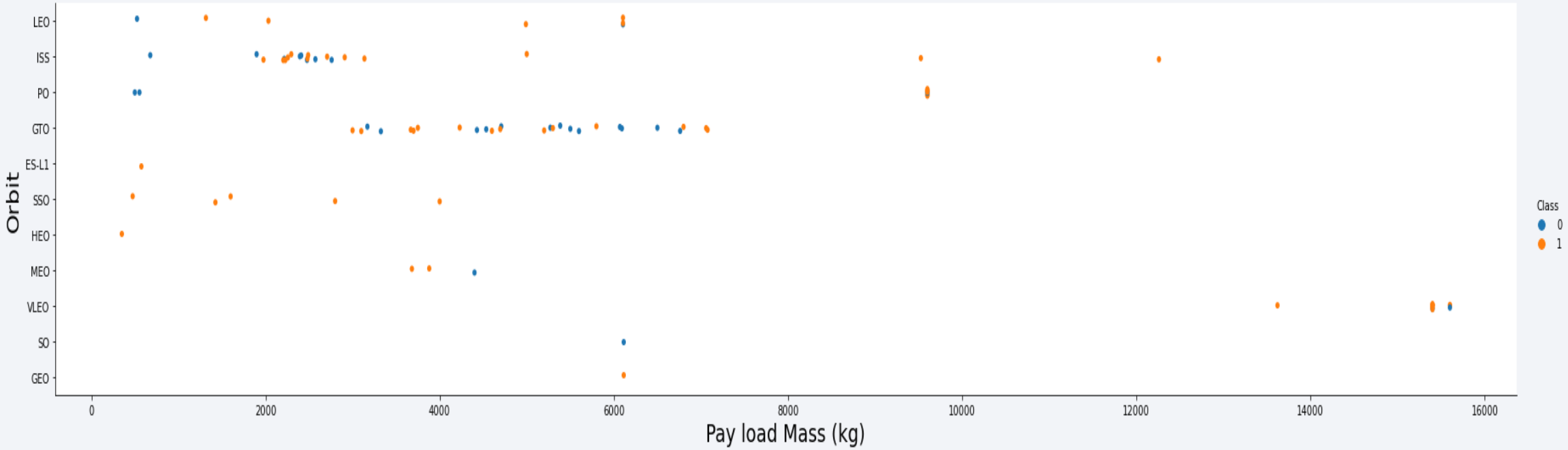
- The highest success rates happens with orbit type in ES-L1, SSO, HEO, GEO

Flight Number vs. Orbit Type



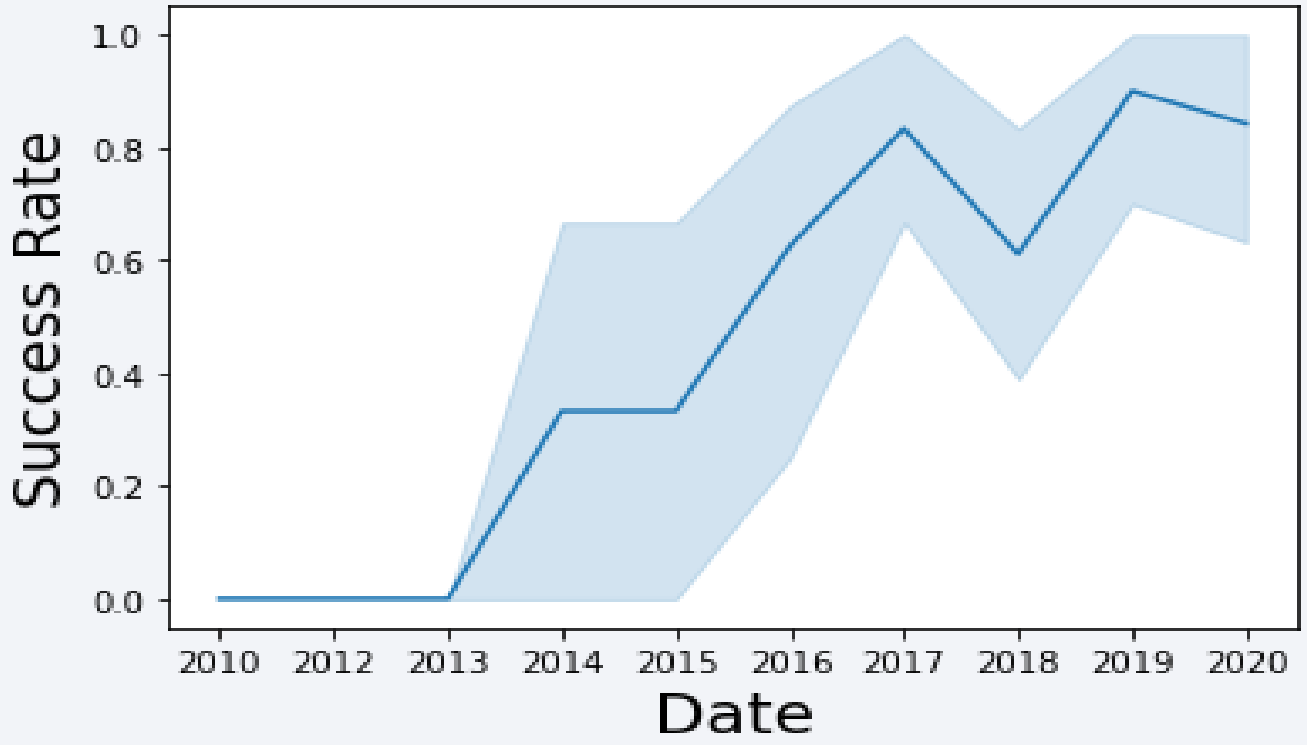
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend



- You can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- According to the data, these names of the unique launch sites:
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

- This selection was obtained by querying the distinct 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	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Displayed launch sites that has CCA in their name

Total Payload Mass

- Total payload carried by boosters from NASA:
  - 45596

- Payload was calculated by aggregating and summing all payload with the "NASA (CRS) in their names

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1:
- 2928

- Filtered the data and calculated the average payload mass of booster version F9 v1.1

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad:
- 2015-12-22

- Filtered the data by searching for the min date where landing outcome is success ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- Selected the boosters where landing outcome = success (ground pad) and payload mass is > 4000 and less < 6000

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes:
- Failure (in flight) 1
- Success 99
- Success (payload status unclear) 1

- got the count of the mission outcomes where the mission outcome is either a failure and success.

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass:
- Displayed the booster version that had the max payload 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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Displayed the landing outcome, booster version and launch sites where the date was 2015 and the landing outcome was failure

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

- Ranking 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:

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

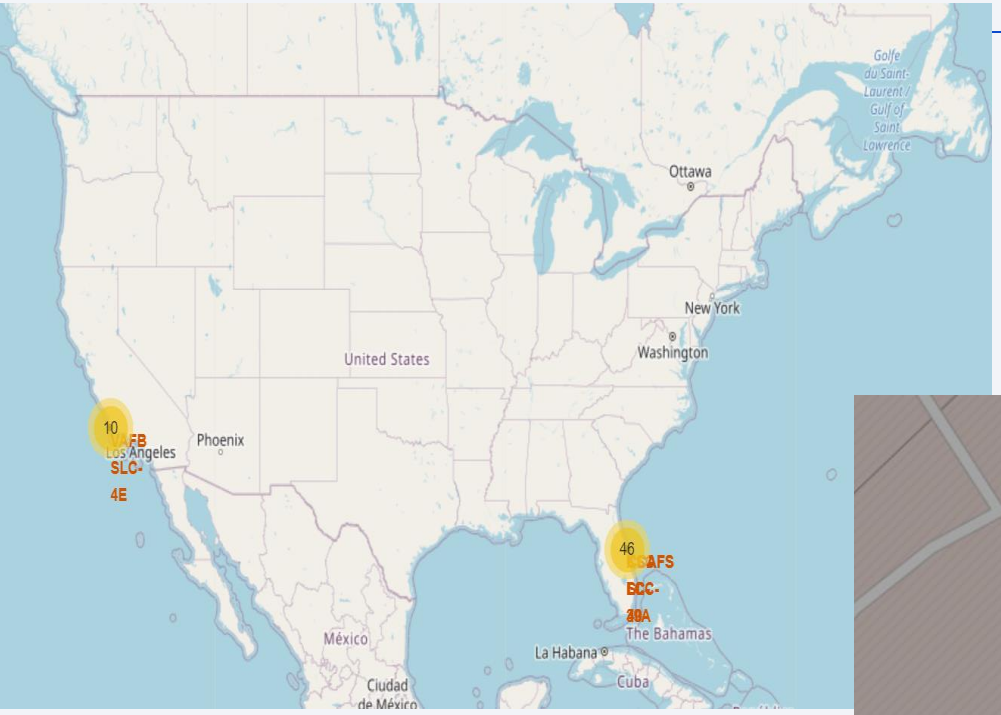
- Displayed the count of the landing outcome between the required dates.

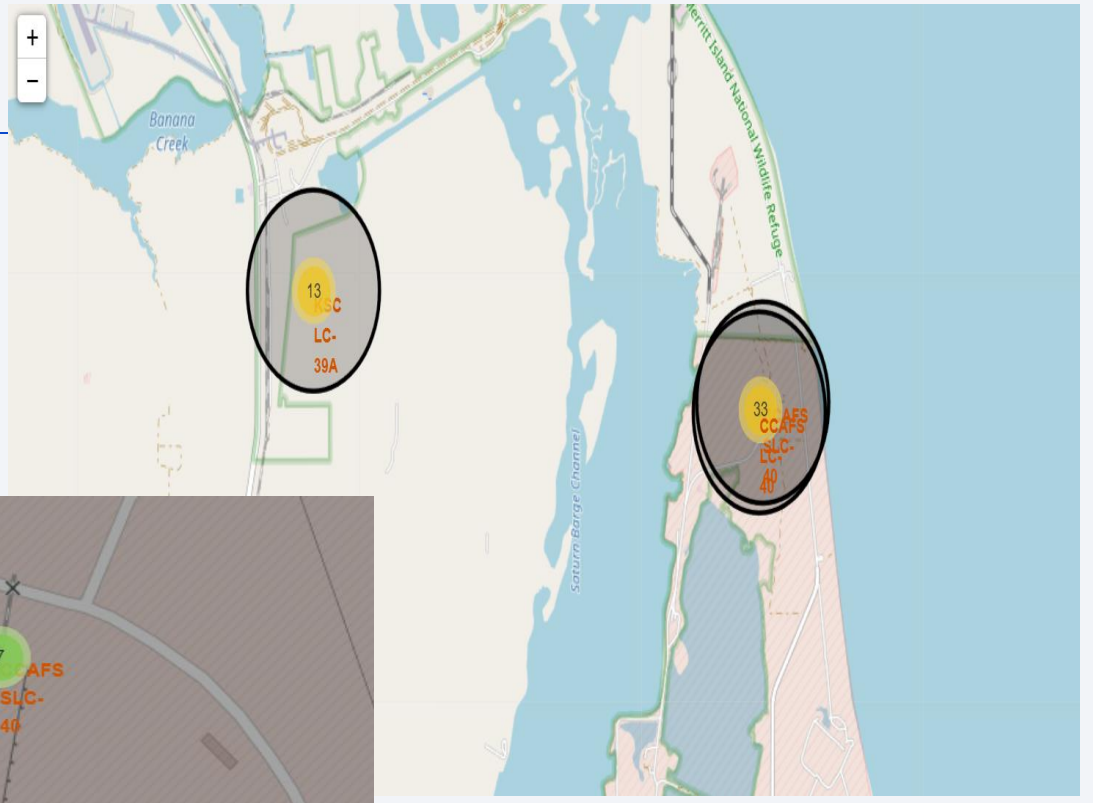

Section 3
Launch Sites
Proximities Analysis

All Launch Sites Markers

- As you can see, the launch sites are on the coast, by the sea in the United States

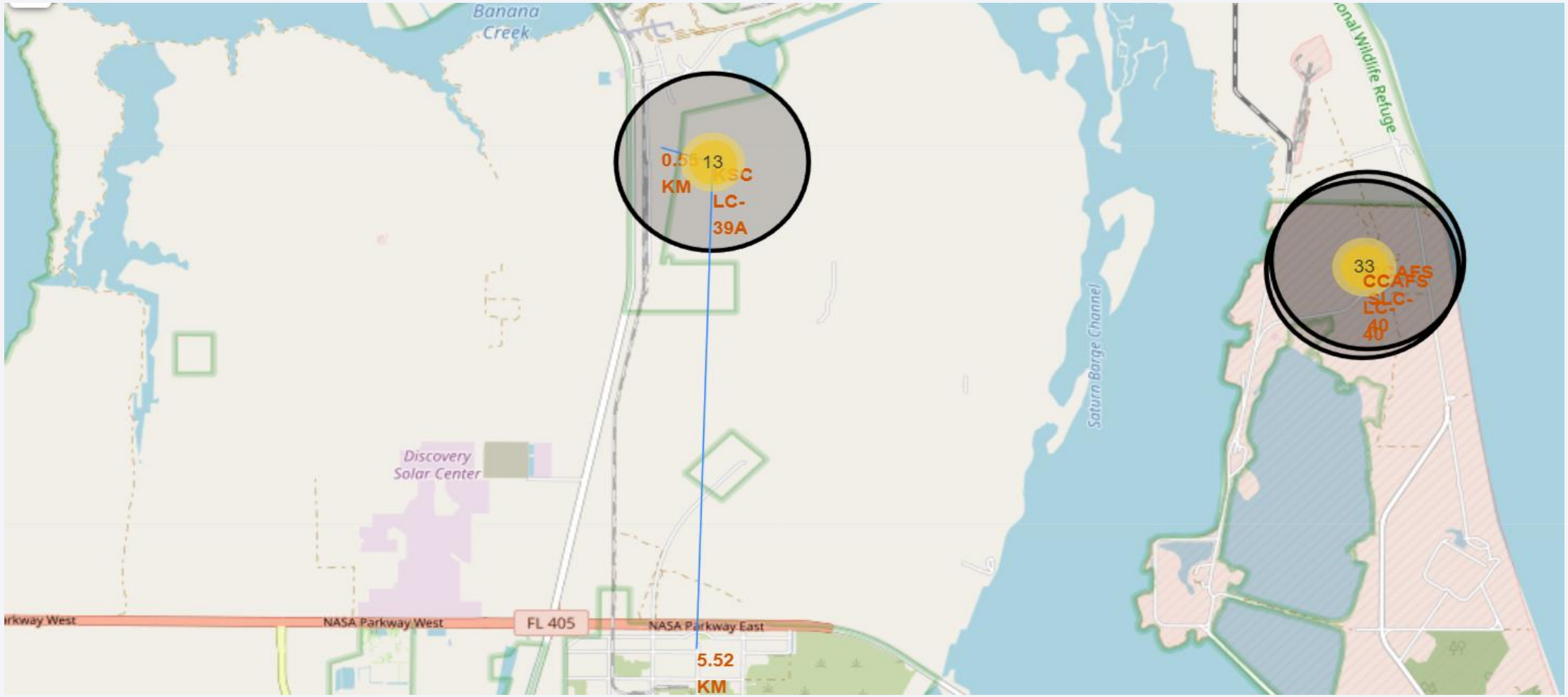
Launch Outcomes



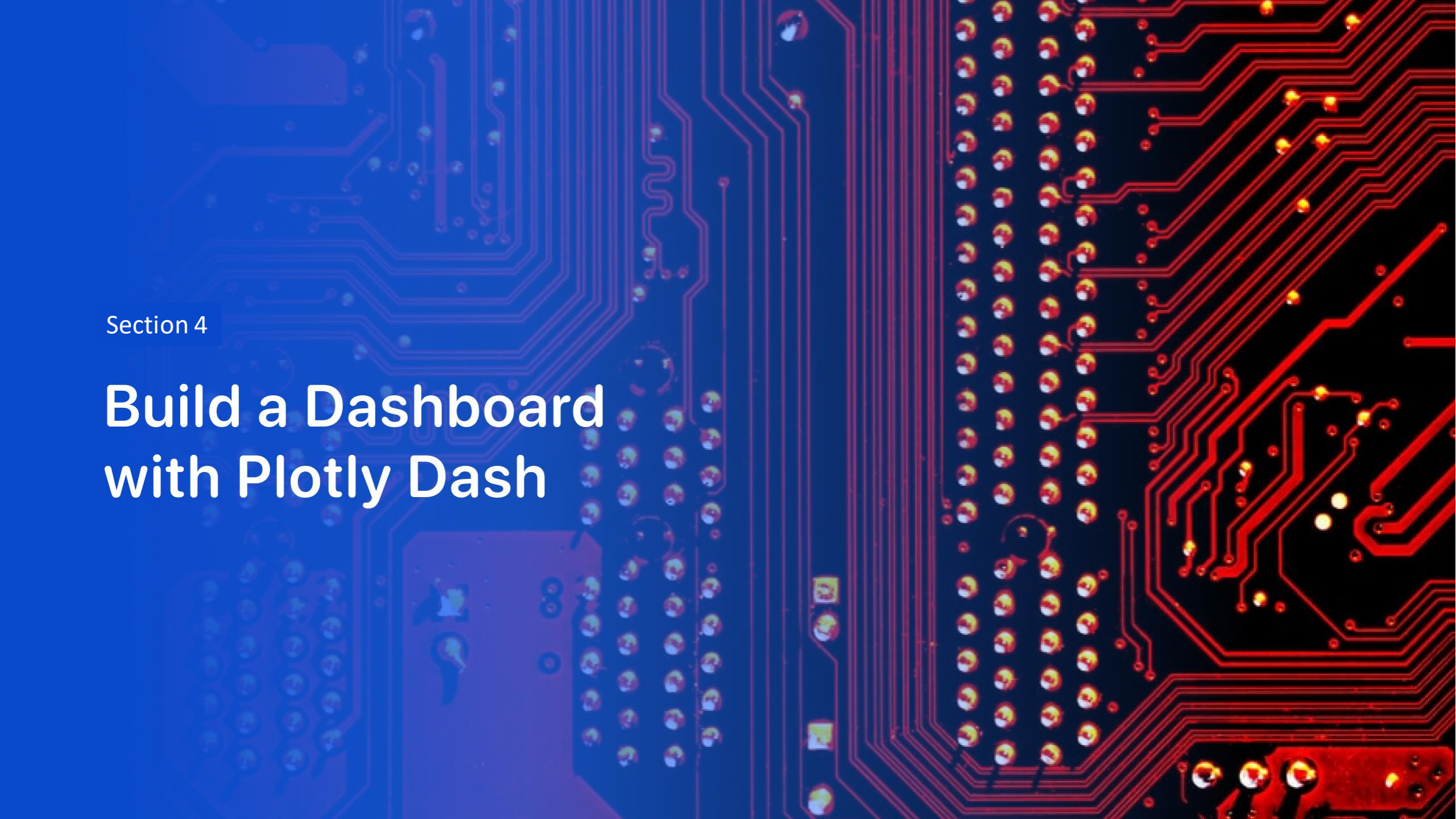


- Green Markers show successful launches and Red shows Failed outcomes

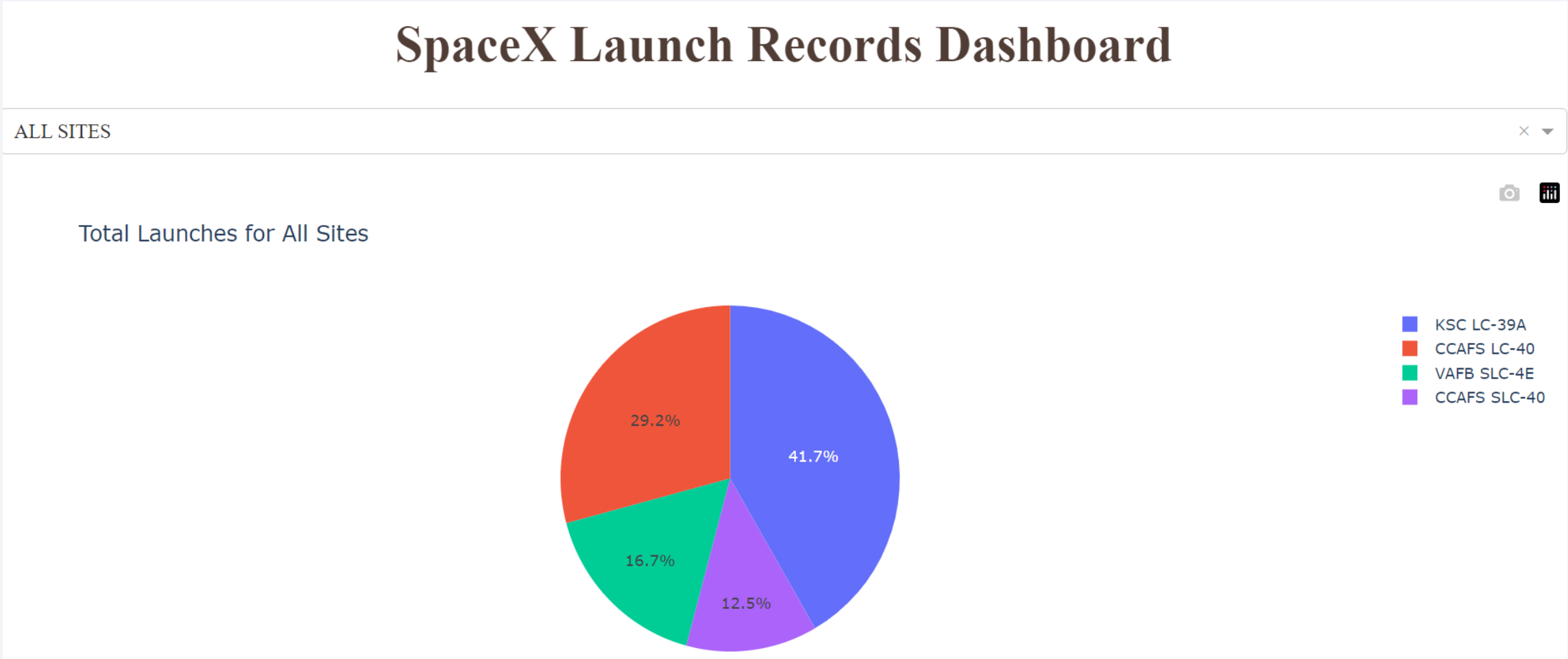
Launch Site Distances



- Launch sites aren't close to highways, railways or cities. They are close to the coastlines

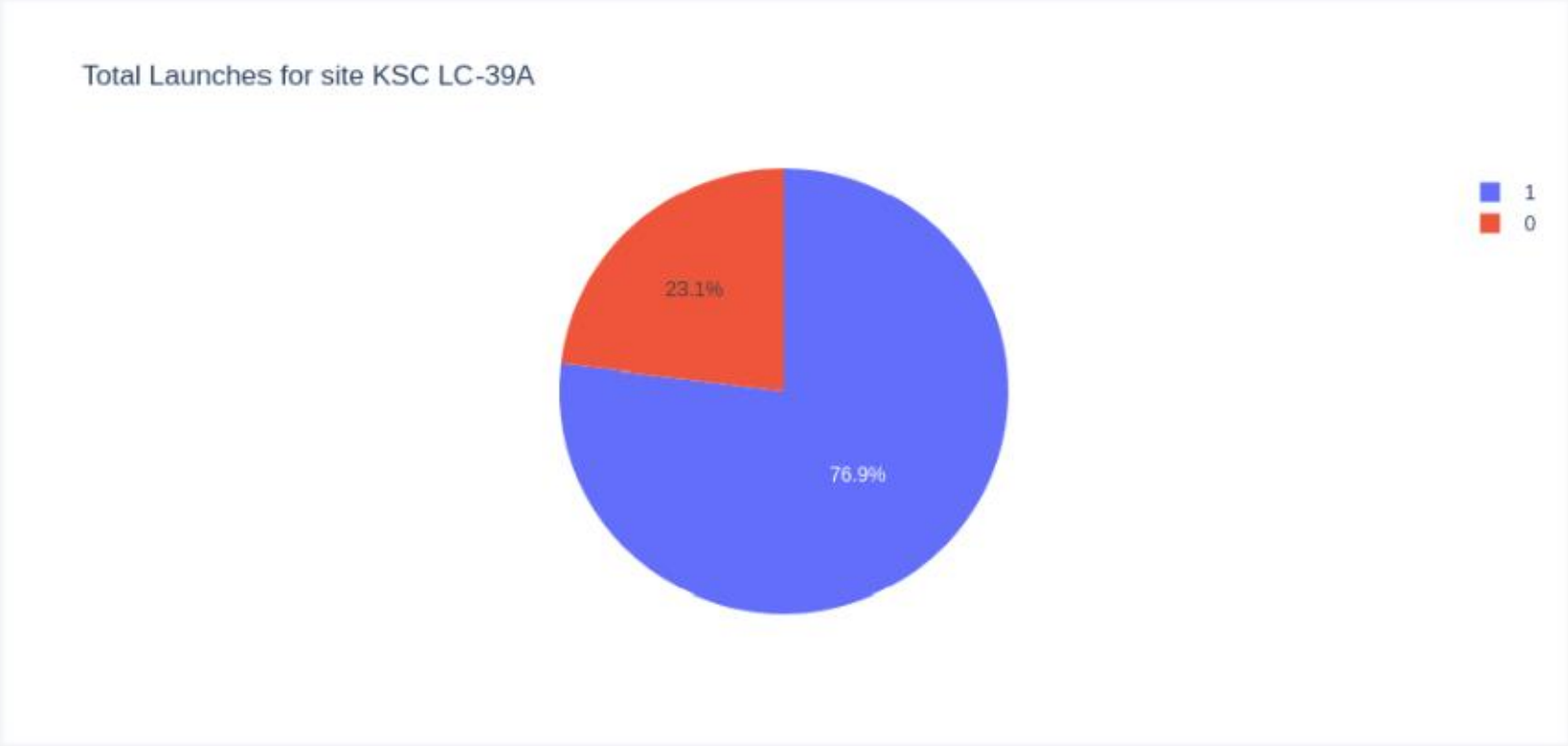

Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches By All Sites



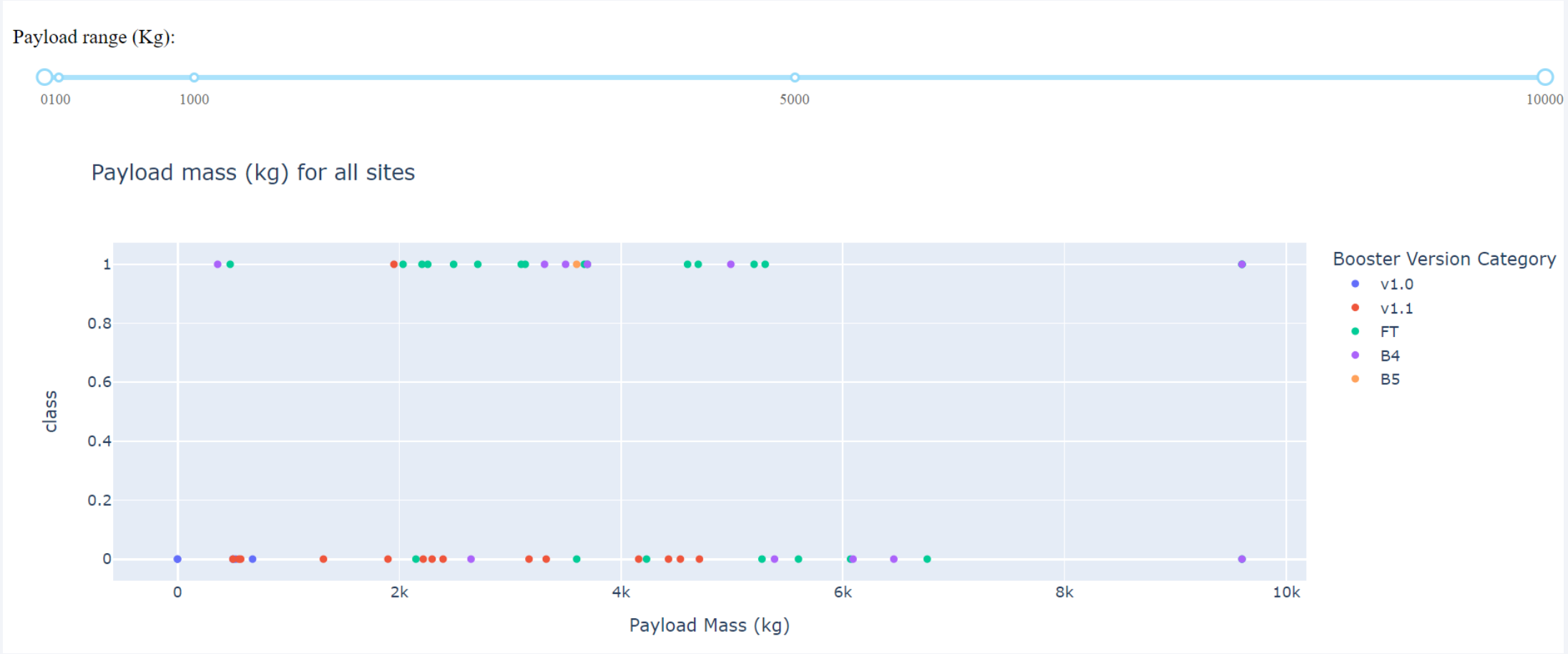
- Launch Sites seems to be an important factor. You can observe KSC LC-39A had the most successful launches of all sites

Site with the Highest Launch Success Ratio



- The KSC LC-39A has the highest success rate of 76.9%

Payload vs. Launch Outcome Scatter Plot

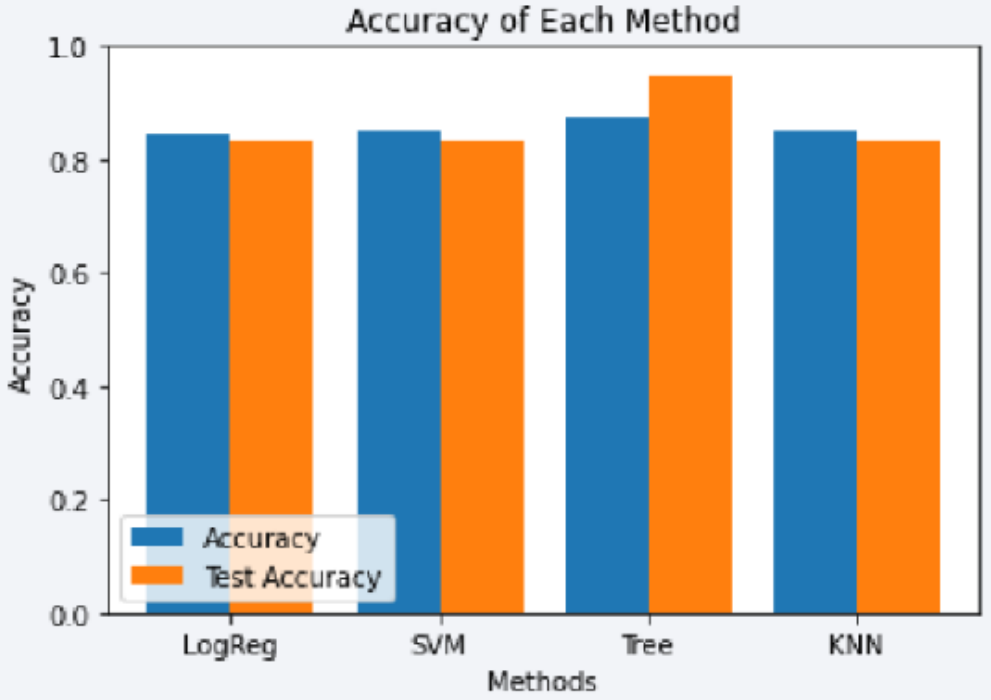


- The success rates for lower payloads are higher than the high payloads.

Section 5

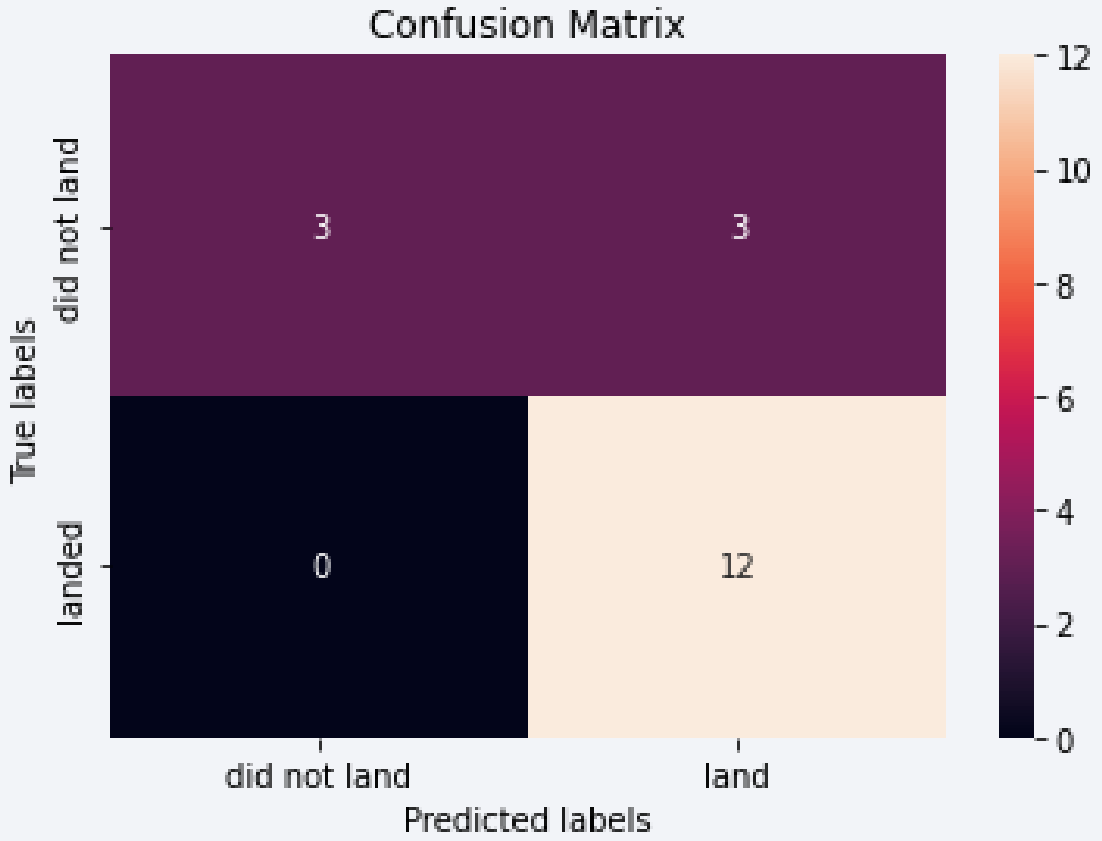
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree Classification Model has the highest accuracy of about 87%

Confusion Matrix



- You can see the Decision Tree Classification Model Confusion Matrix can distinguish between the different classes. The major problem is the false positives

Conclusions

- Mission Outcomes are generally successful
- Mission Outcomes success rate did improve over time, but that could be due to external factors like advancements in technology
- KSC LC-39A was the most successful launch site
- The Decision Tree Classification model is the most accurate model to be used to predict successful landings
- …

Appendix

- Every Code and Notebook is available on GitHub

Thank you!