IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matias Herrera Muñoz
20-06-2025

Outline

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

Executive Summary

- **Goal**: Predict Falcon 9 first-stage landings to support launch cost estimation for Space Y, using public SpaceX data.
- **Methodology**: Cleaned and explored data; trained ML models including Decision Tree (best accuracy: 88.9%); built interactive dashboards.
- **Key Insights**: Landing success improves with flight number; KSC LC-39A has the highest success rate; SSO orbits had 100% success over multiple launches

Introduction

- The rise of commercial spaceflight has made launch costs a key competitive factor. SpaceX leads the industry by reusing the Falcon 9's first stage, significantly lowering launch prices compared to other providers.
  However, not all missions recover the first stage. This project simulates the role of a data scientist at "Space Y", a competitor aiming to estimate launch costs and predict first-stage landings using public SpaceX data.

- Key Question:
  - Can we predict if the first stage will land successfully?
  - What features most influence landing success?
  - How do launch sites and orbits affect outcomes?
  - Can we build interactive dashboards to support decisions?

Allon Mask
My boss

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API of the SpaceX data site.
  - WebScrapping from Wikipedia.
- Perform data wrangling
  - Data Cleaning and Filtered
- 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 Standardization trough a One-Hot Coding.

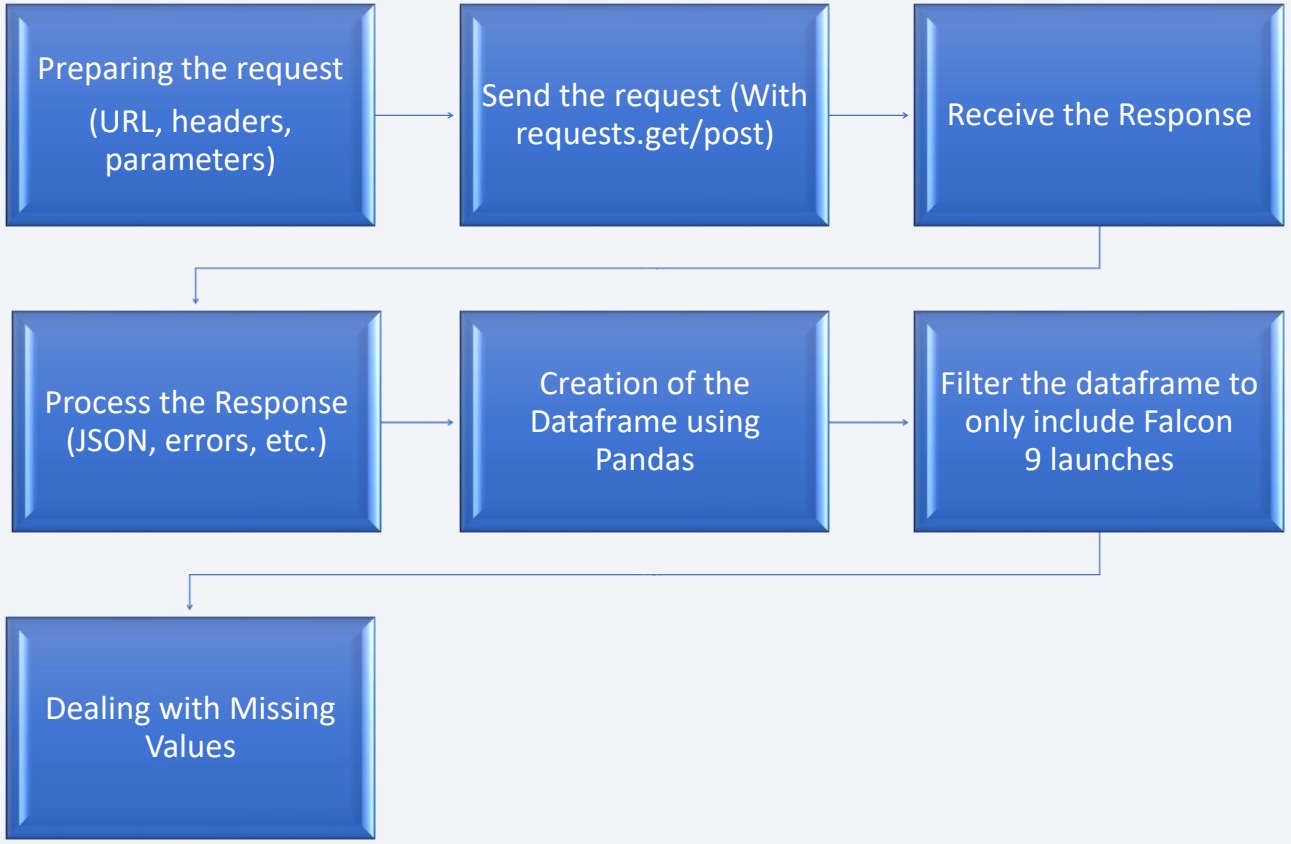
Data Collection

- The data was collected with two different methods:
  - API: The data was collected from the API https://api.spacexdata.com/v4/launches/past

- You need to present your data collection process use key phrases and flowcharts

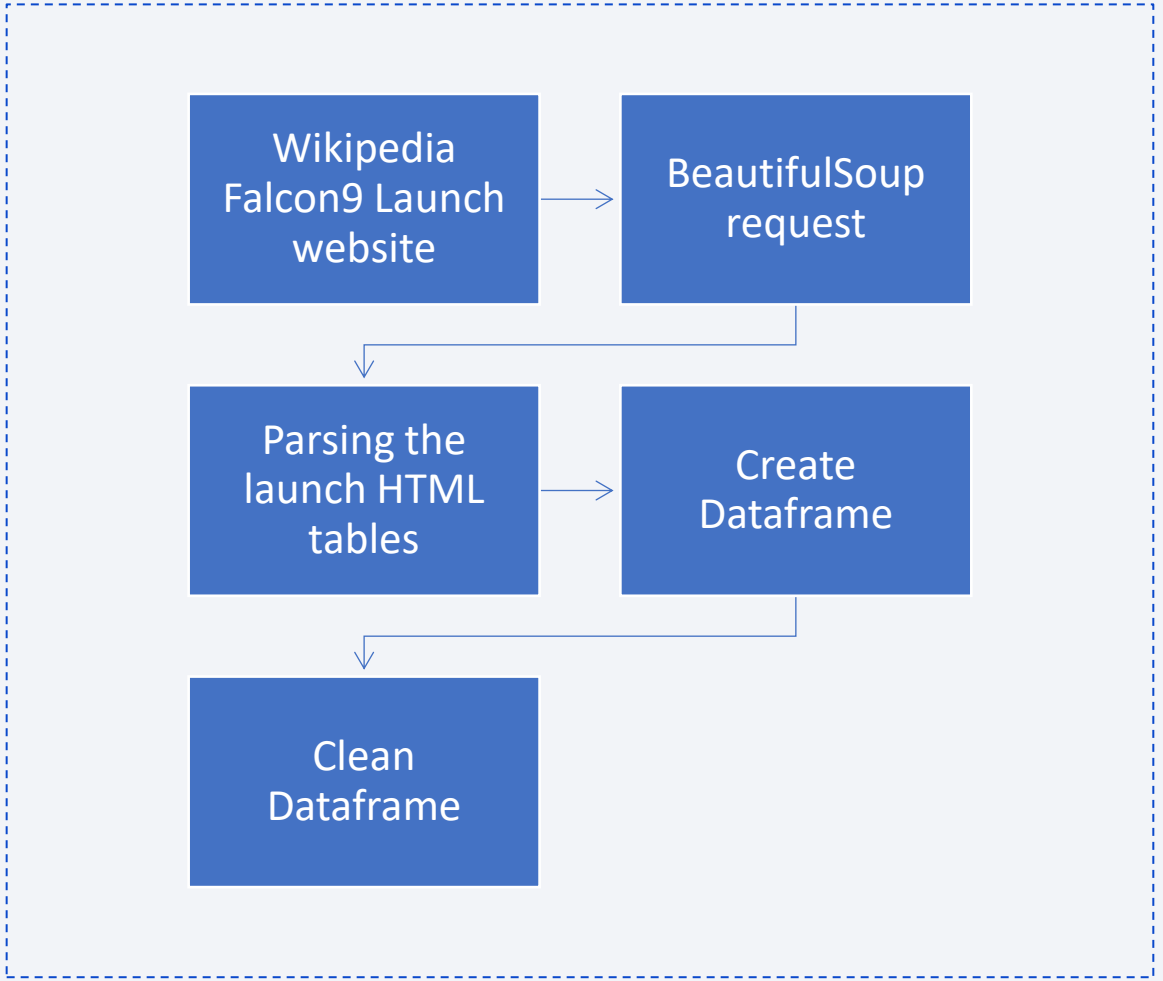
Data Collection – SpaceX API

- The Data was called using the request function, then with pandas I built the DataFrame.
- The DF is composed by the columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- All the collection work can be found in Github



Data Collection - Scraping

- The data was collected from Falcon9's Wikipedia page.
- With BeautifulSoup we extract all the page.
- Next identify the tables with the necessary information.
- Iterate trough the tables to organize the information into a DataFrame
- The process can be found on GitHub.



Data Wrangling

- With all the data in a DataFrame, we handled missing values by replacing NaNs with the mean for numerical columns and the most frequent value for categorical columns. However, in the specific case of the 'LandingPad' column, the NaN values were not replaced.
- Then the column ‘Landing_class’ was added, in this column the value is zero if the landing failed, and the value is 1 if the landing success.
- This task can be found in GitHub.

EDA with Data Visualization

- We try to see the relevant data that help us to identify if the first stage will land, to achieve this we will use:
  - FlightNumber vs Payload Mass scatter plot
  - LaunchSite vs Flight Number scatter plot
  - LaunchSite vs Payload Mass scatter plot
  - Orbit type ratio of success bar chart.
  - Orbit vs Flight Number Scatter plot
  - Orbut vs PayloadMass Scatter Plot
  - Success Rate vs Date Line graph
- All this plot can be found on Github

EDA with SQL

- Some queries were implemented using SQL on the database. The following are key pieces of information that were gathered:

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

Figure 1. List of Mission Outcome

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

Figure 2. Rank of Landing Outcome

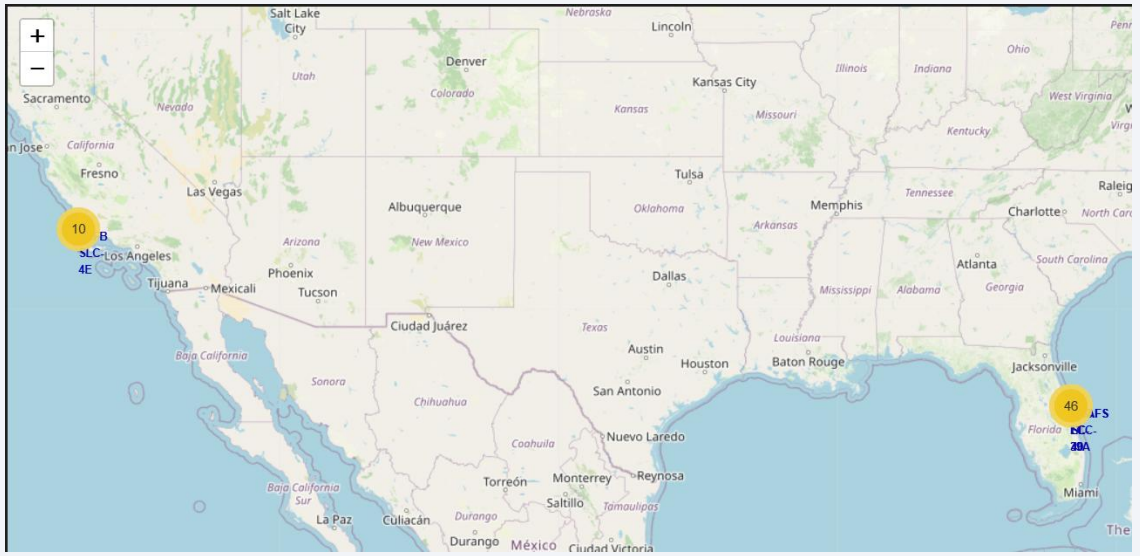
Total Payload of Nasa (CRS) [KG]
45596

Figure 3. Total Payload Mass transported by Nasa

- All the queries can be found on GitHub

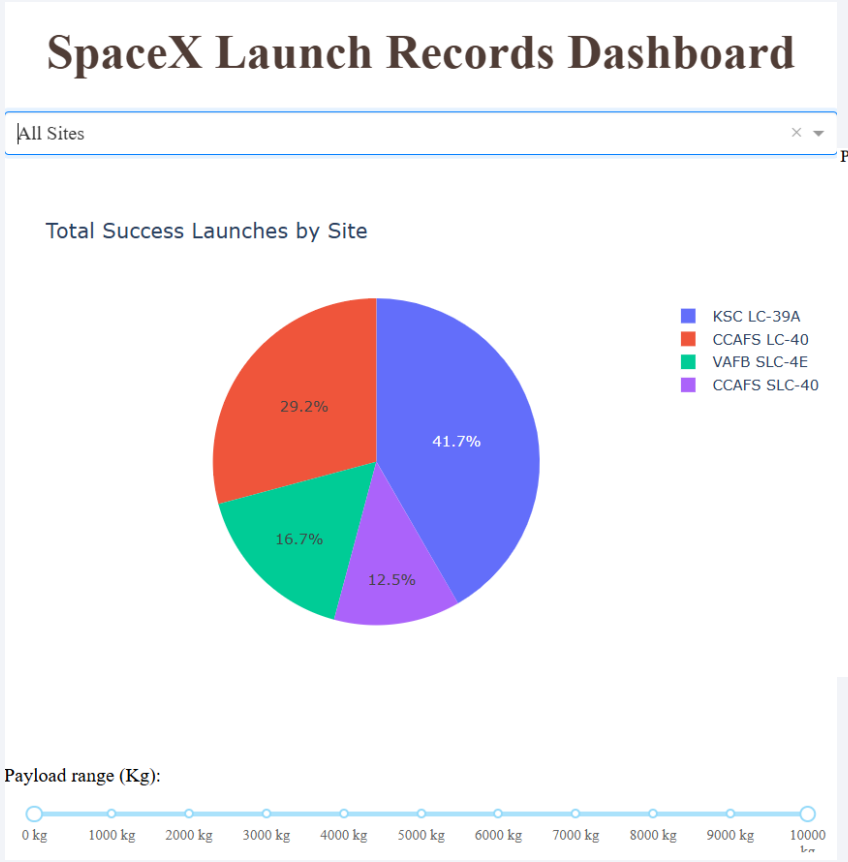
Build an Interactive Map with Folium

- I create interactive Maps with folium to recognize the launch sites
- To achieve this, I use markers to mark the position of the launch sites.
- The markers had different color for the landing outcome, green for success, and red for failed landings.
- The code can be found on GitHub



Build a Dashboard with Plotly Dash

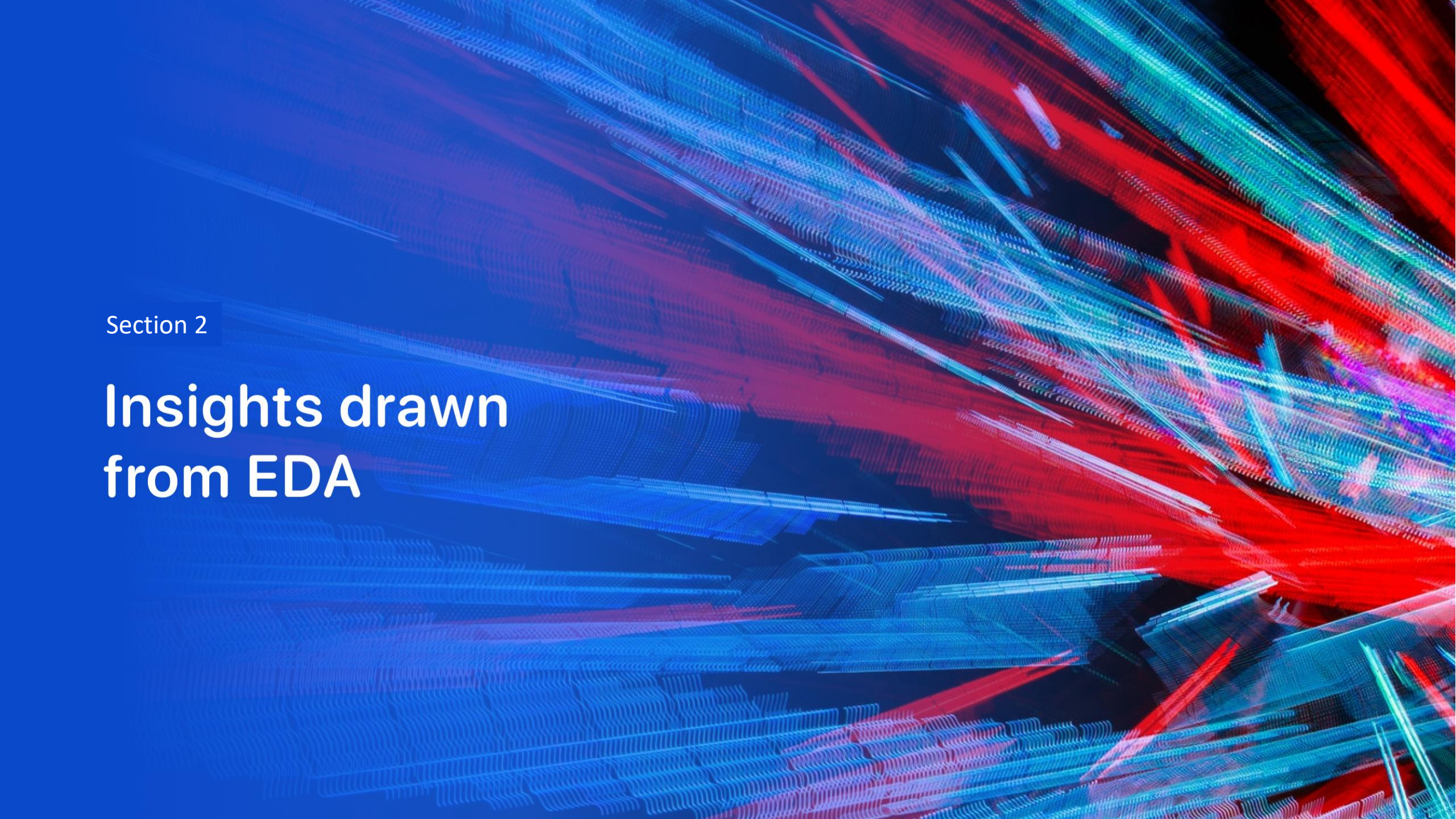
- It was created interactive visualization with Plotly Dash
- The visualization consist in:
  - Total success launches by site
  - Payload Mass vs Launch success for Launch Sites



Predictive Analysis (Classification)

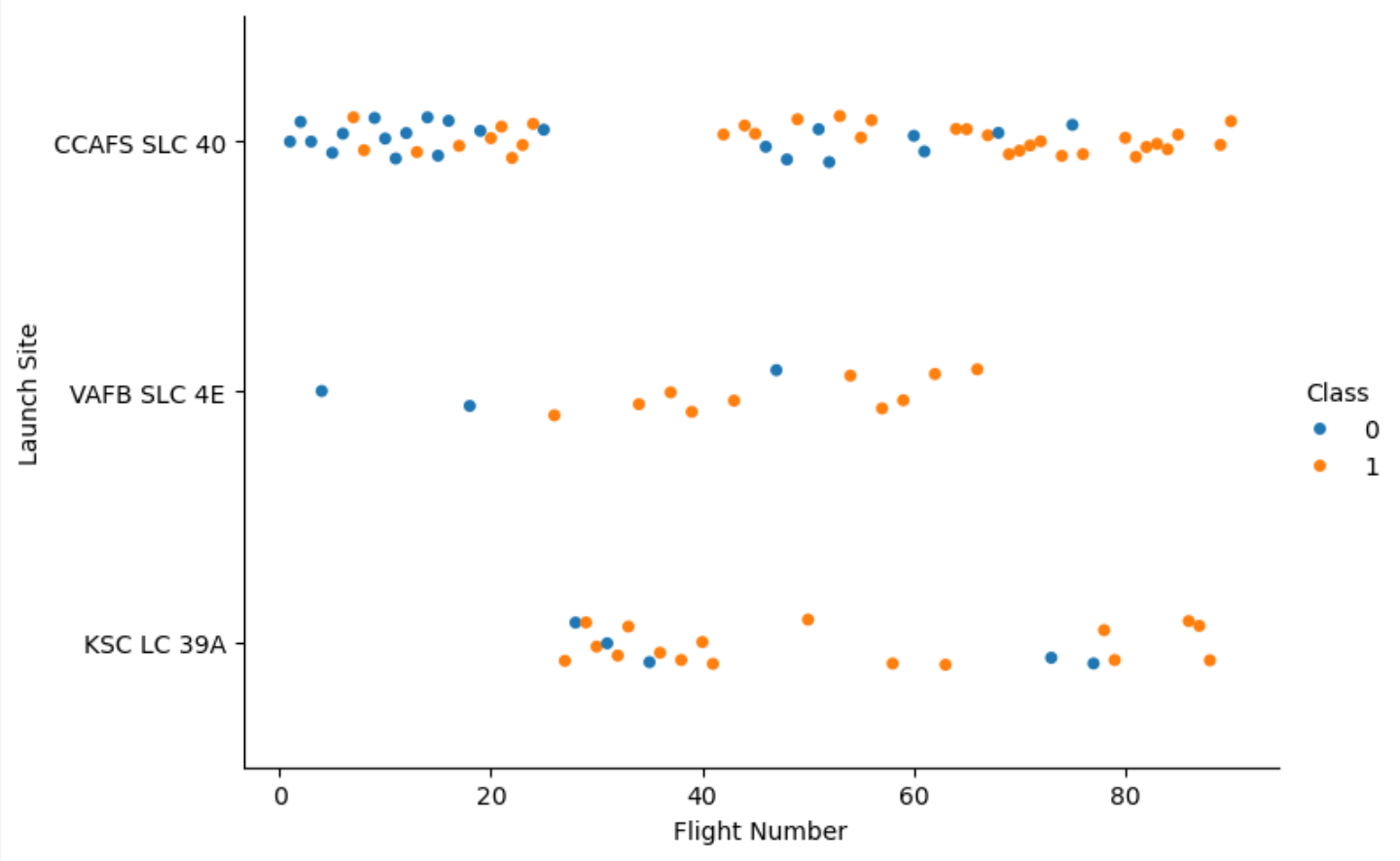
- Four models were used:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN

- To compare the performance, we use the same train test split.
- All the procedure can be found on GitHub.


Section 2
Insights drawn from EDA

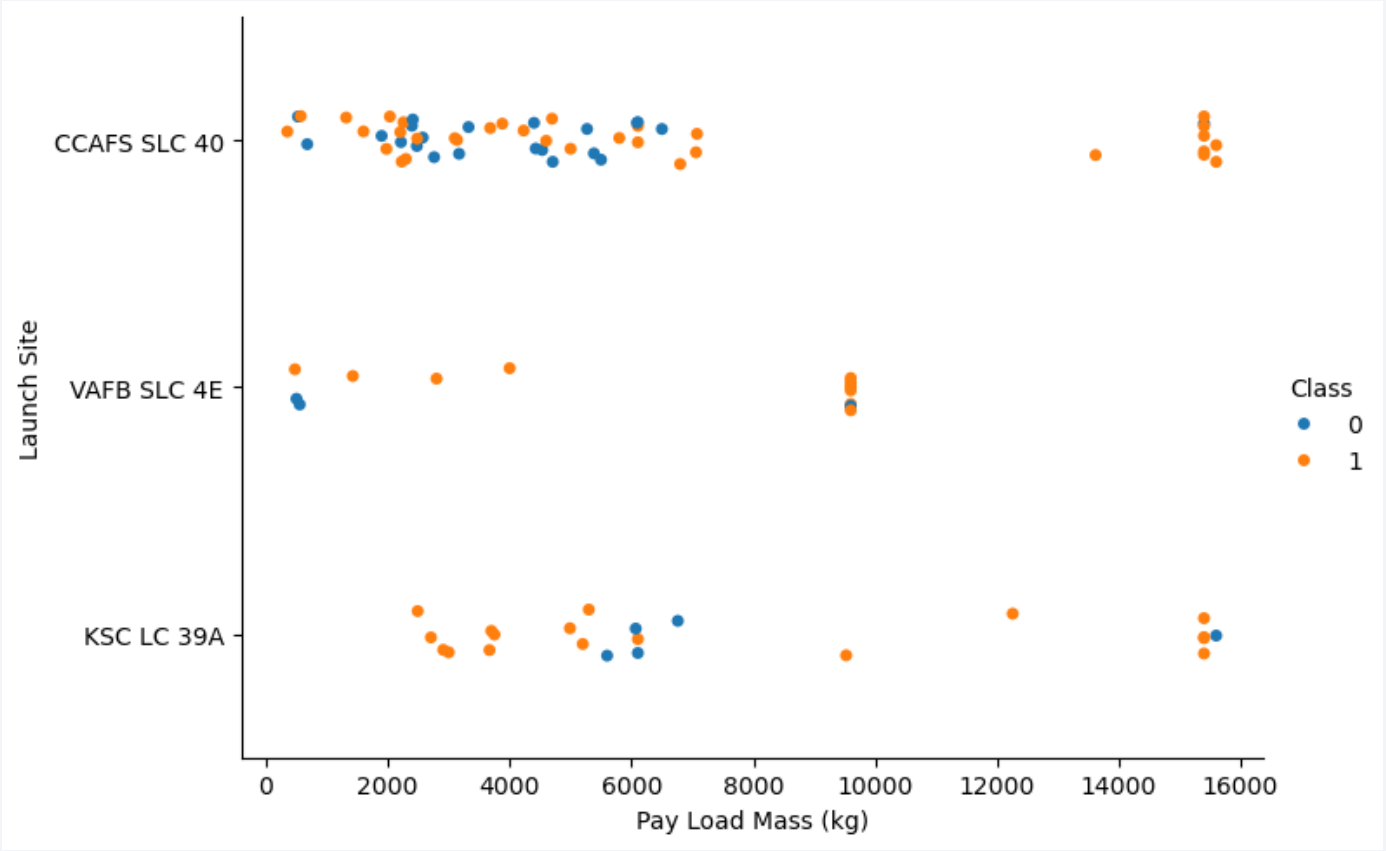
Flight Number vs. Launch Site

- The launch site VAFB SLC-4E shows the lowest number of recorded launches in the dataset, suggesting limited operational use, possibly as a test site that was later decommissioned.
- In contrast, CCAFS SLC-40 is the most utilized launch site, especially in the early missions. However, between flights 30 and 40, there is a noticeable decline in its usage, likely due to maintenance or upgrades. During this period, the KSC LC-39A site shows a significant increase in activity and stands out with the highest recorded success rate among all launch sites.



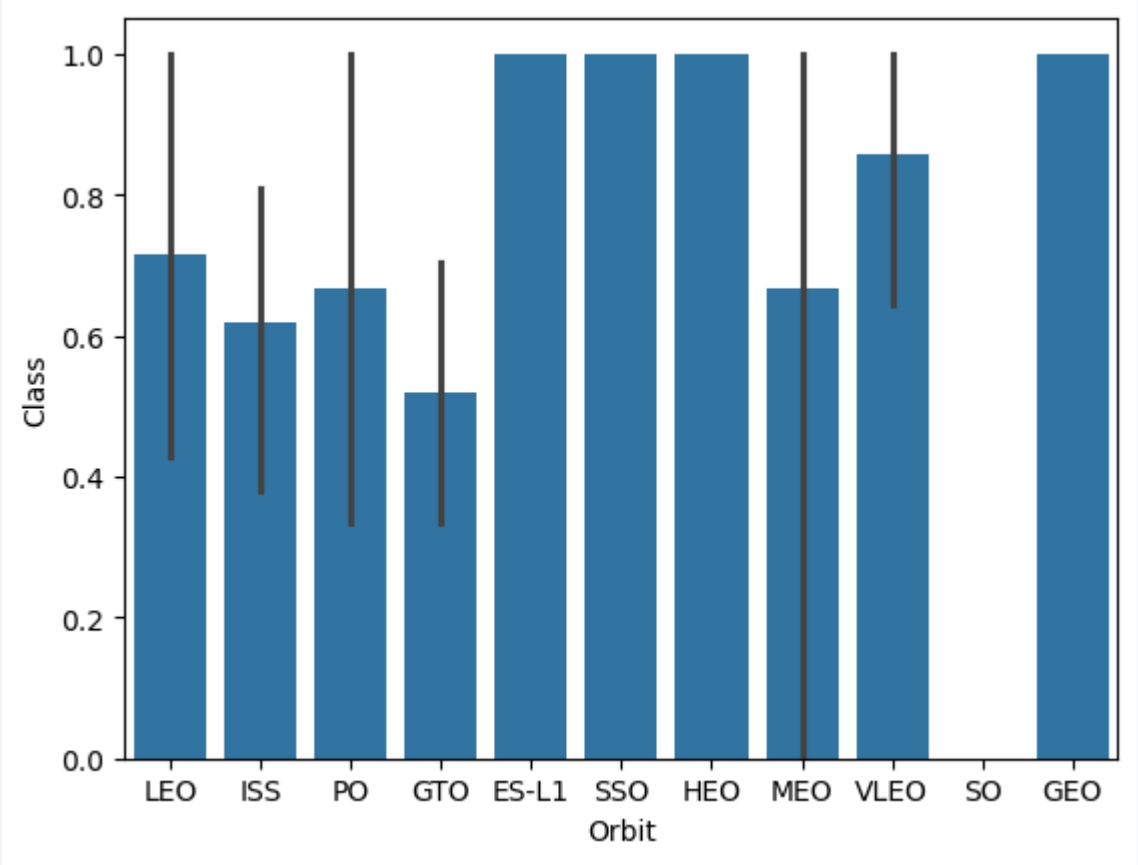
Payload vs. Launch Site

- CCAFD and KSC are the only launch sites with the maximum payload mass
- It can be seen some sort of trouble for the KSC launch site around the 6000kg of Payload Mass



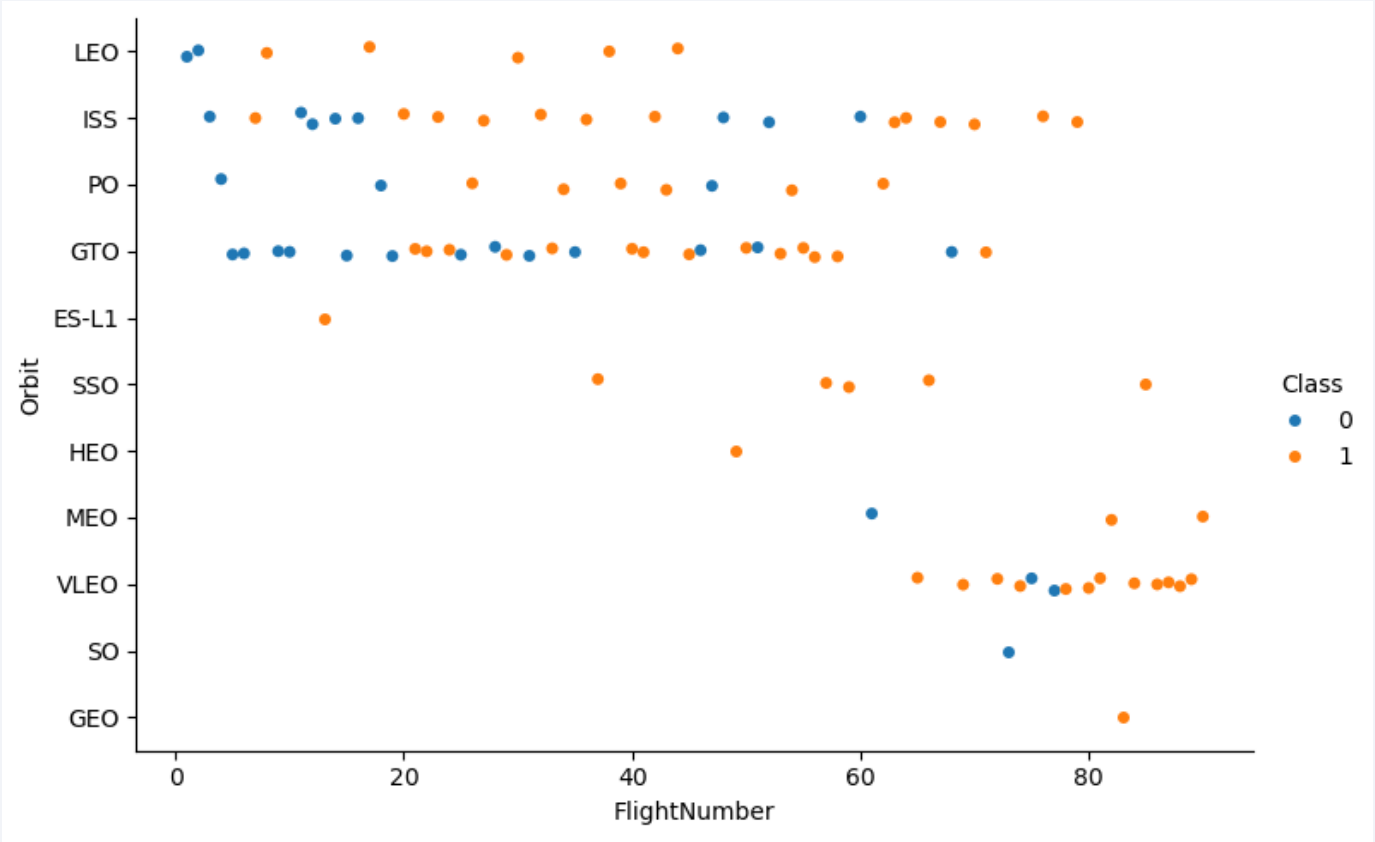
Success Rate vs. Orbit Type

- The data indicates that four different orbit types achieved a 100% success rate. Among them, the Sun-Synchronous Orbit (SSO) is the only one with more than a single recorded launch, totaling five successful missions conducted between 2017 and 2020.



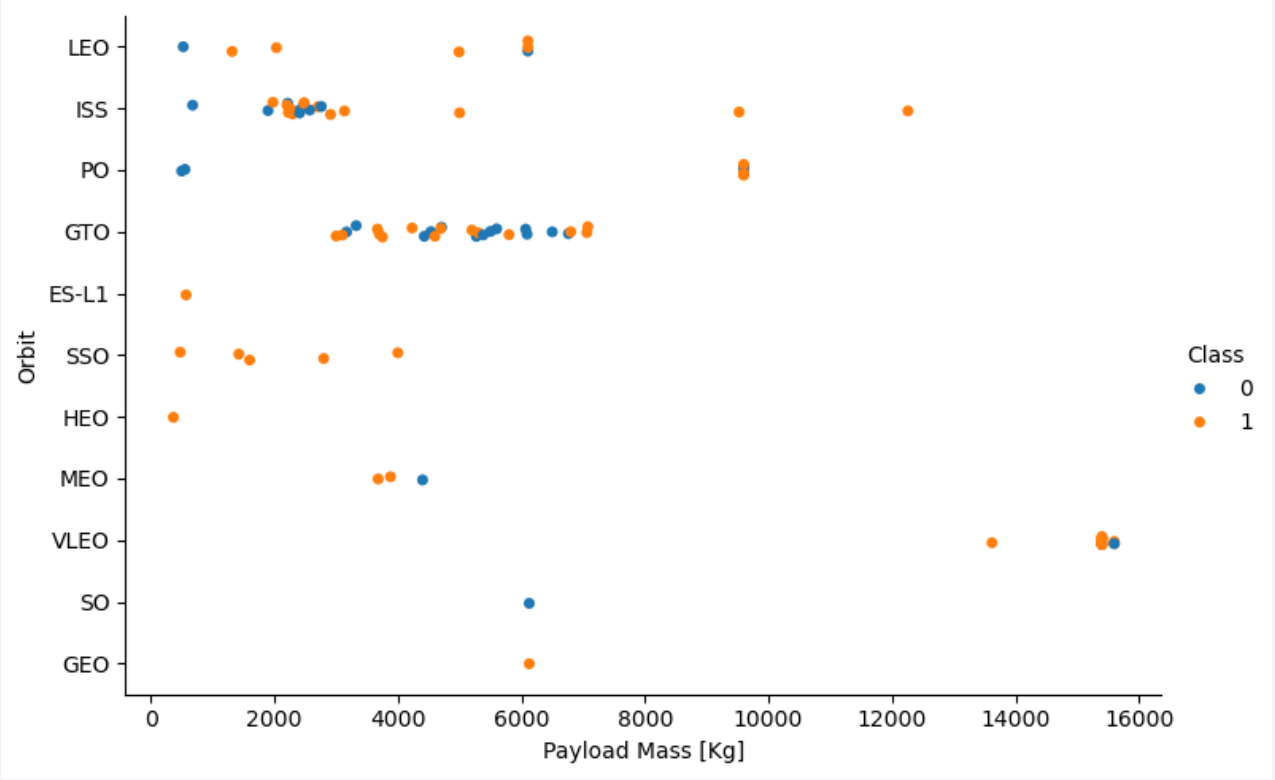
Flight Number vs. Orbit Type

- The LEO orbit is the only that show a relation between the Flight Number.



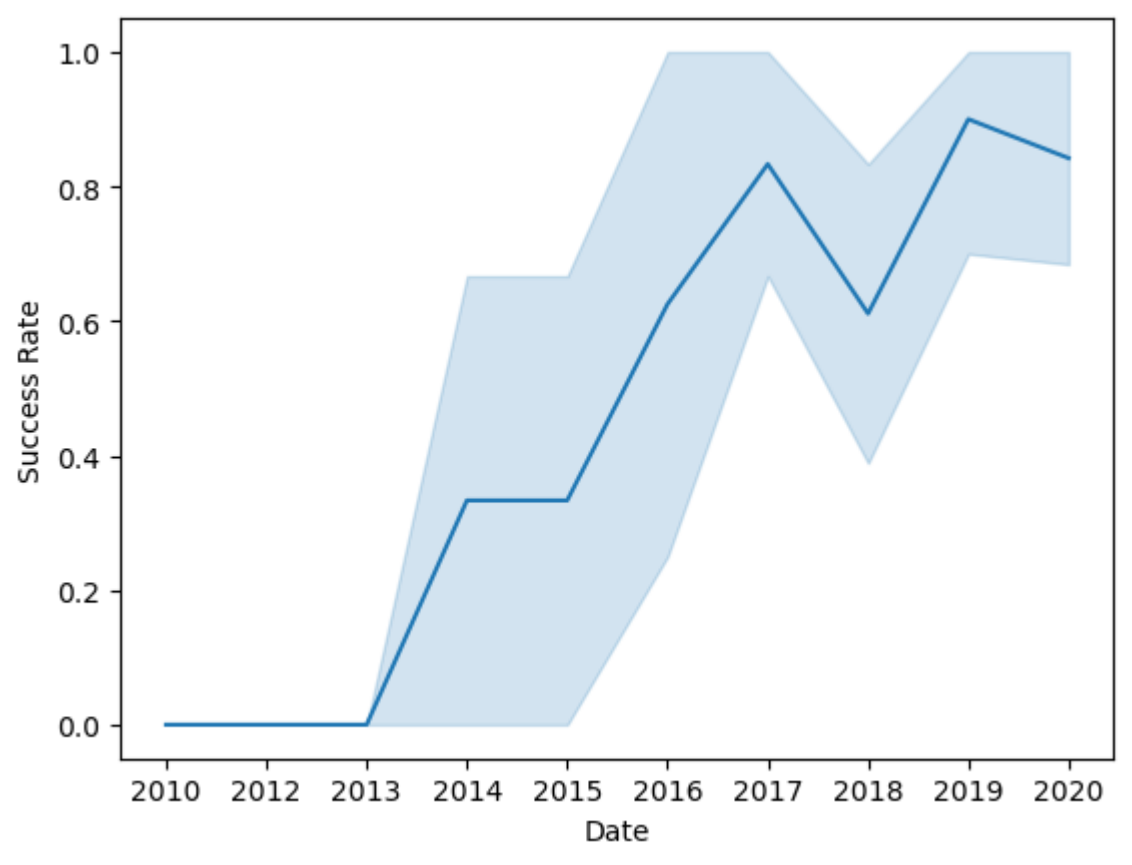
Payload vs. Orbit Type

- 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 see that the success rate increase since 2013
- There is a minimal decrease in the year 2018



All Launch Site Names

- Using SQL queries, I find the names of all the launch Sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

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

Launch Site Names Begin with 'CCA'

- To see the format of the Dataset, I query the information of the launche sites that begin with ‘CCA’
- In the previous query we see that are only two site that begin with CCA: CCAFS LC-40 and CCAFS SLC-40

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

Other minor queries

- The total payload carried by boosters from NASA

Average Payload mass carried by booster VF9
340.4

- Date of the first successful landing in ground pad

First Succesful landing in ground pad
2015-12-22

- Booster which have success landing and have a payload between 4000 and 6000kg

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

- All the booster_versions that have carried the maximum 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

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

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

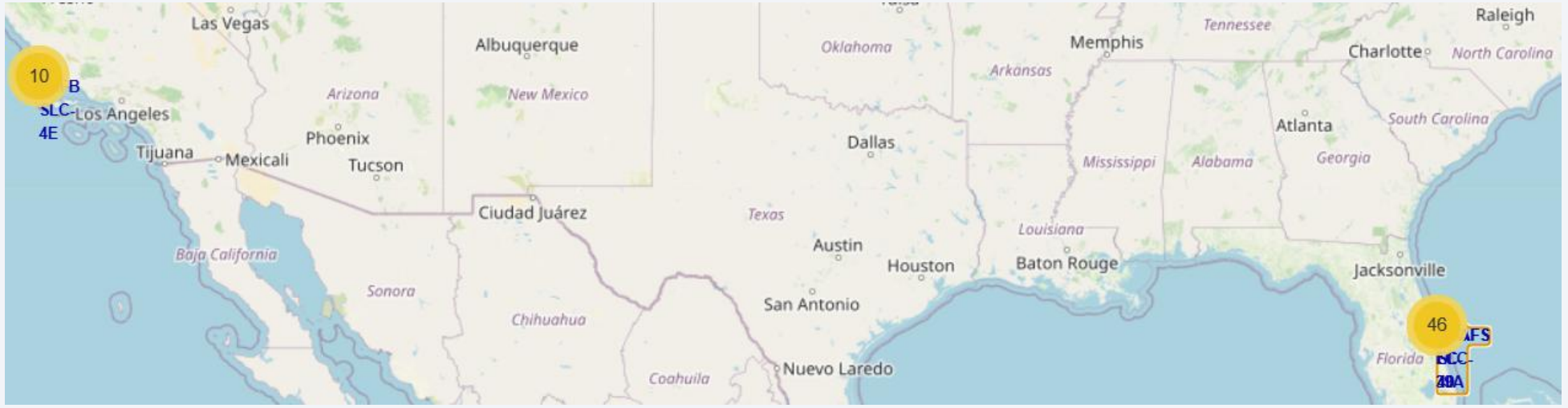
- List the total number of successful and failure mission outcomes

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

Section 3

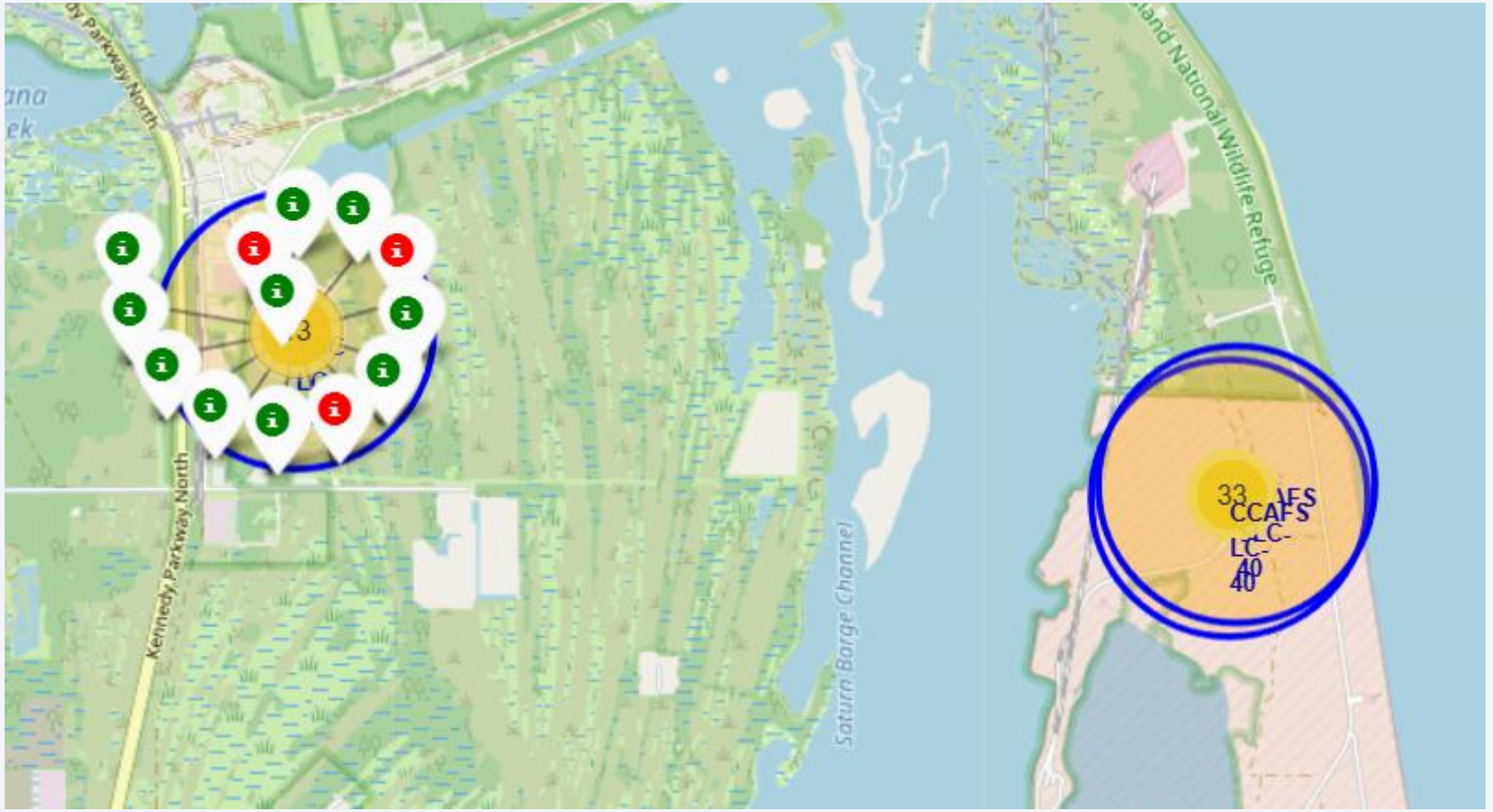
Launch Sites Proximities Analysis

Position of all the launch sites



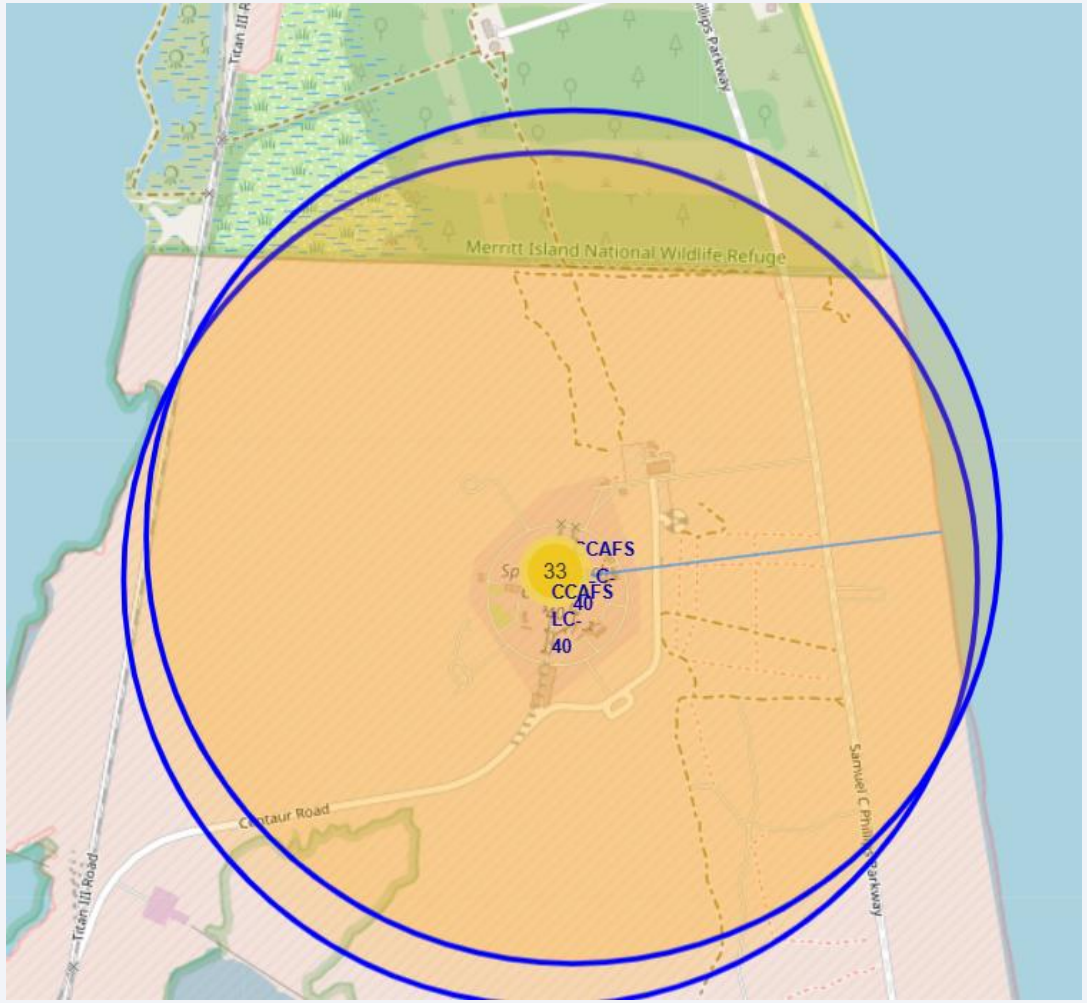
We can see how the launch sites are divided into two groups, one in california and the other in Florida

Marker Styles

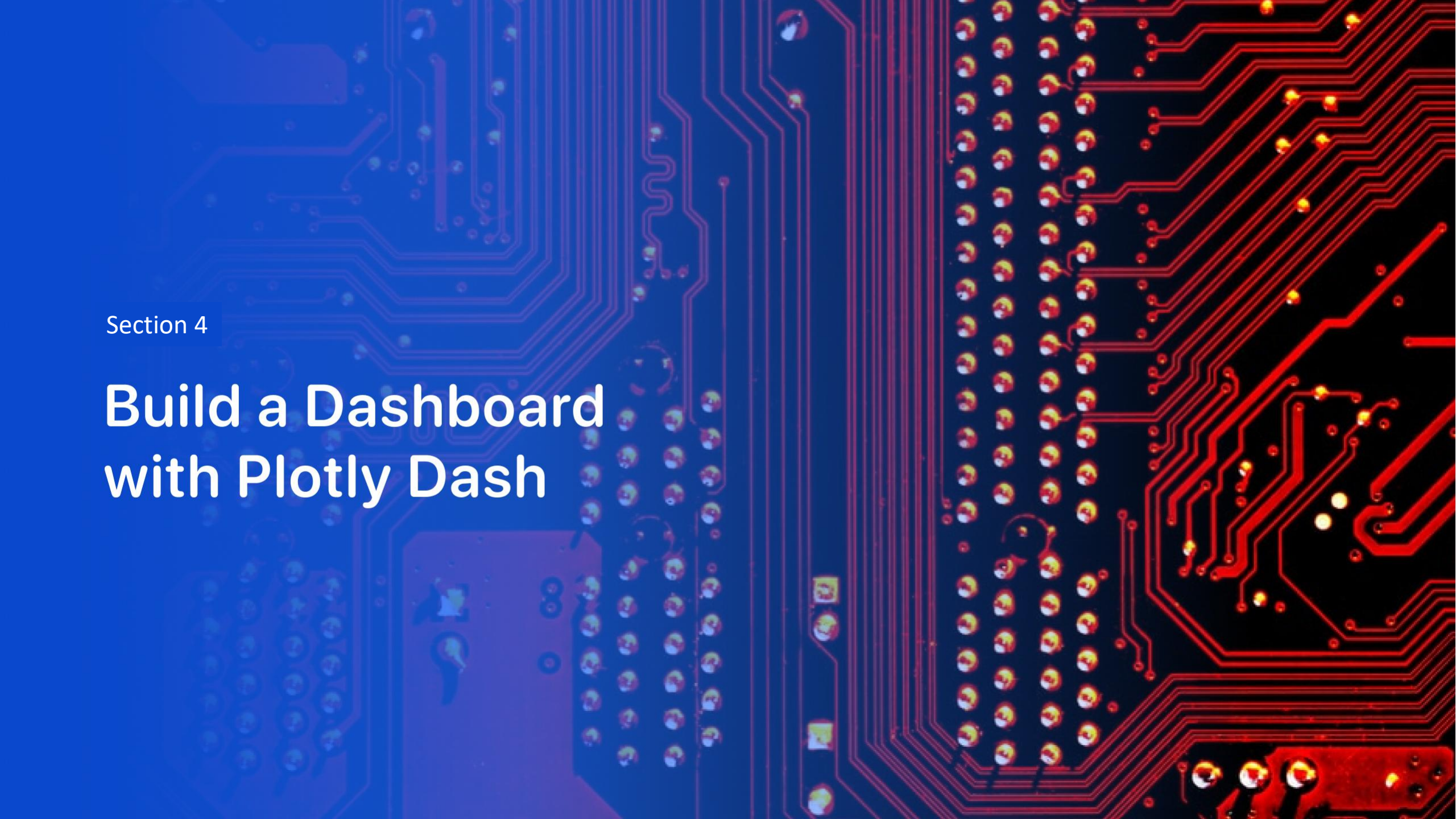


The marker are green for the Success landings, and red for the failed landing

Coastline distance from the launch site



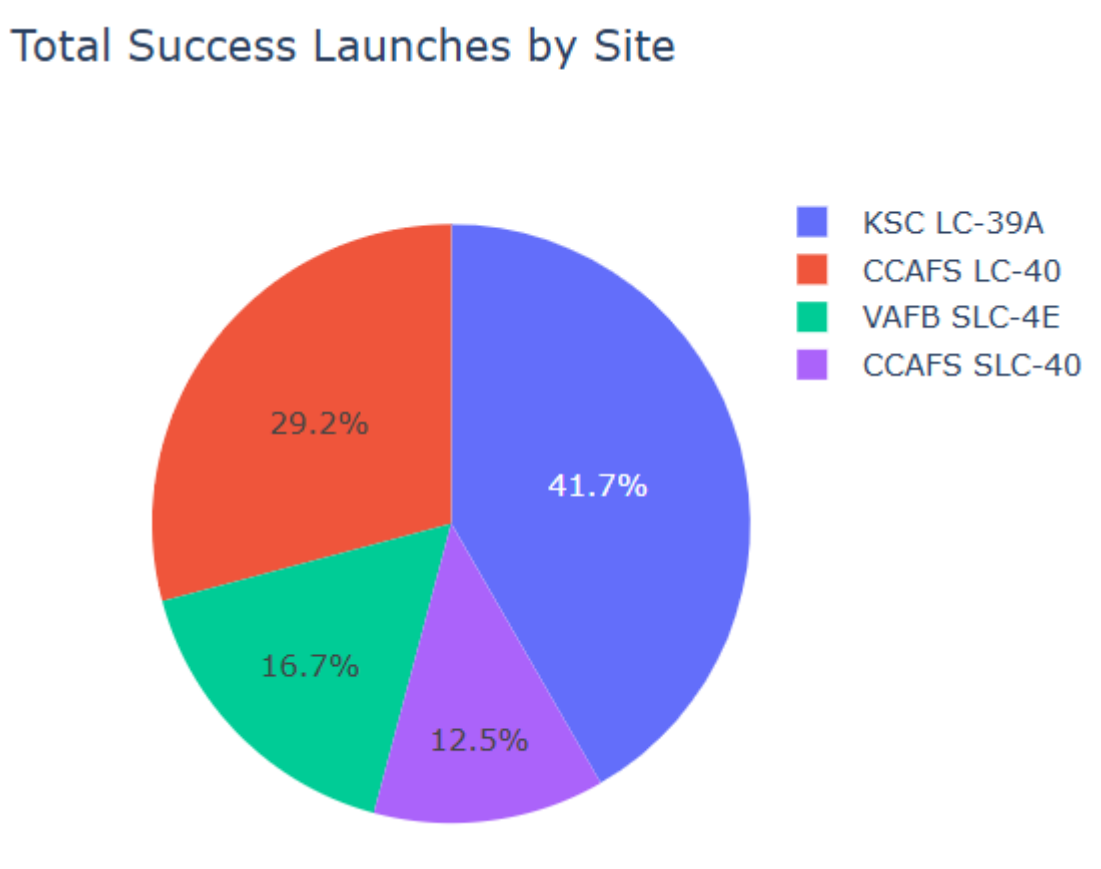
In the Figure we can see how close is the coastline from the launch site


Section 4
Build a Dashboard with Plotly Dash

Success Launches by Site

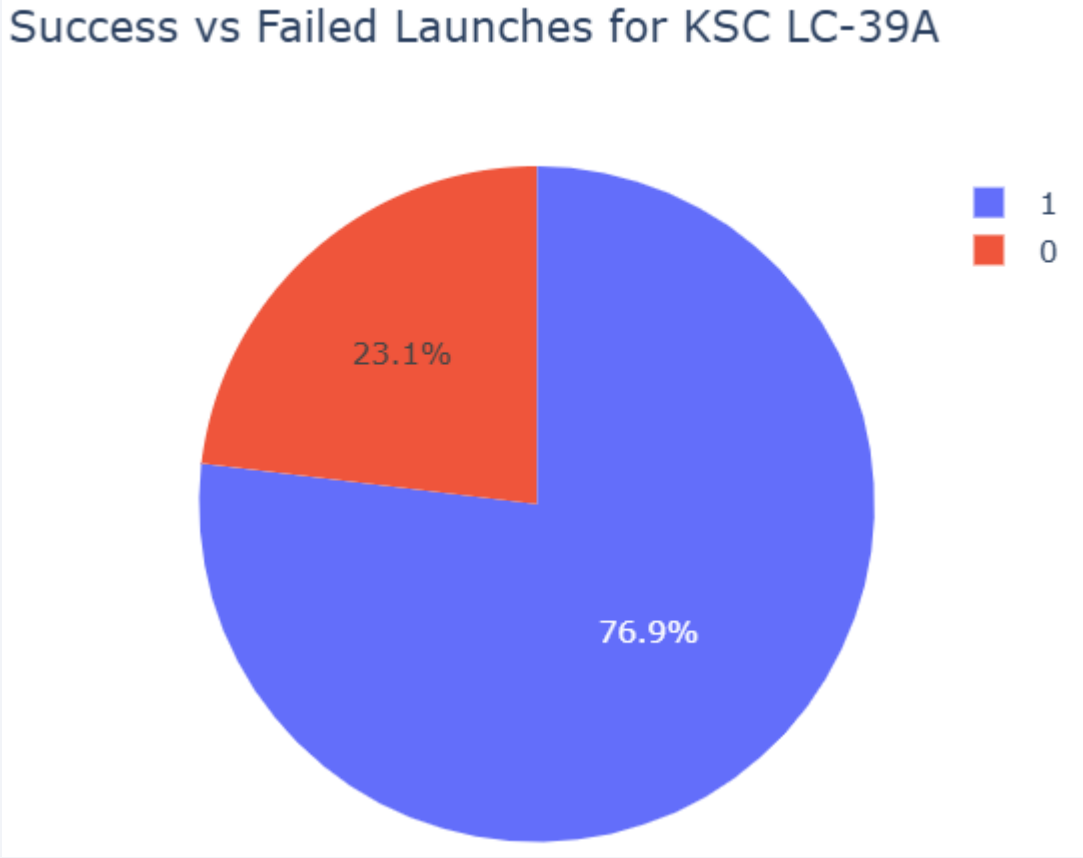
- In the Pie Chart we can see the percentage of success launches of all the launch site.

- The site with mire success launches is the KSC LC-39A with 10 launches.



Launch Site with more success

- The KSC LC-39A has the highest success rate, with
  - 10 success launches
  - 3 failed launches



Correlation of payload and Launch Succes

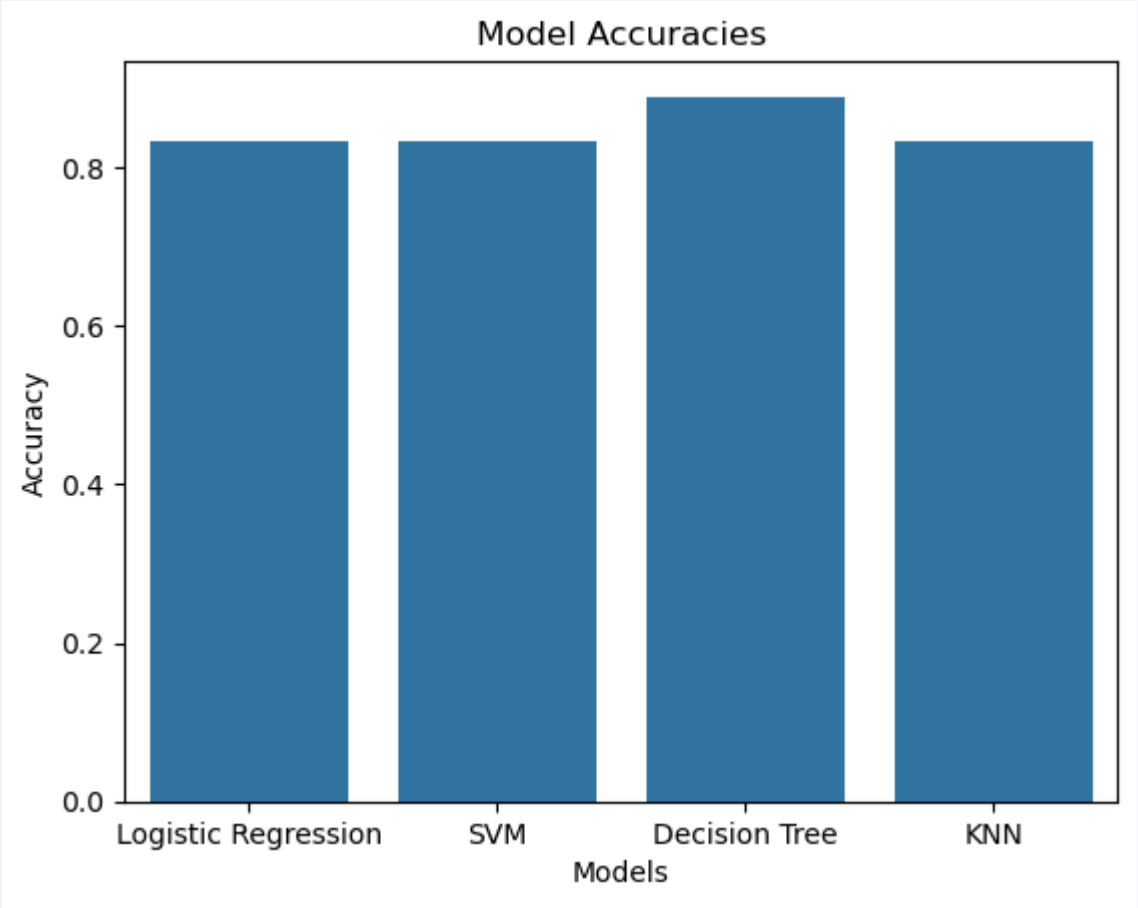
- We can see the predominance of KSC LC-39A Launch site on the success class.

Section 5

Predictive Analysis (Classification)

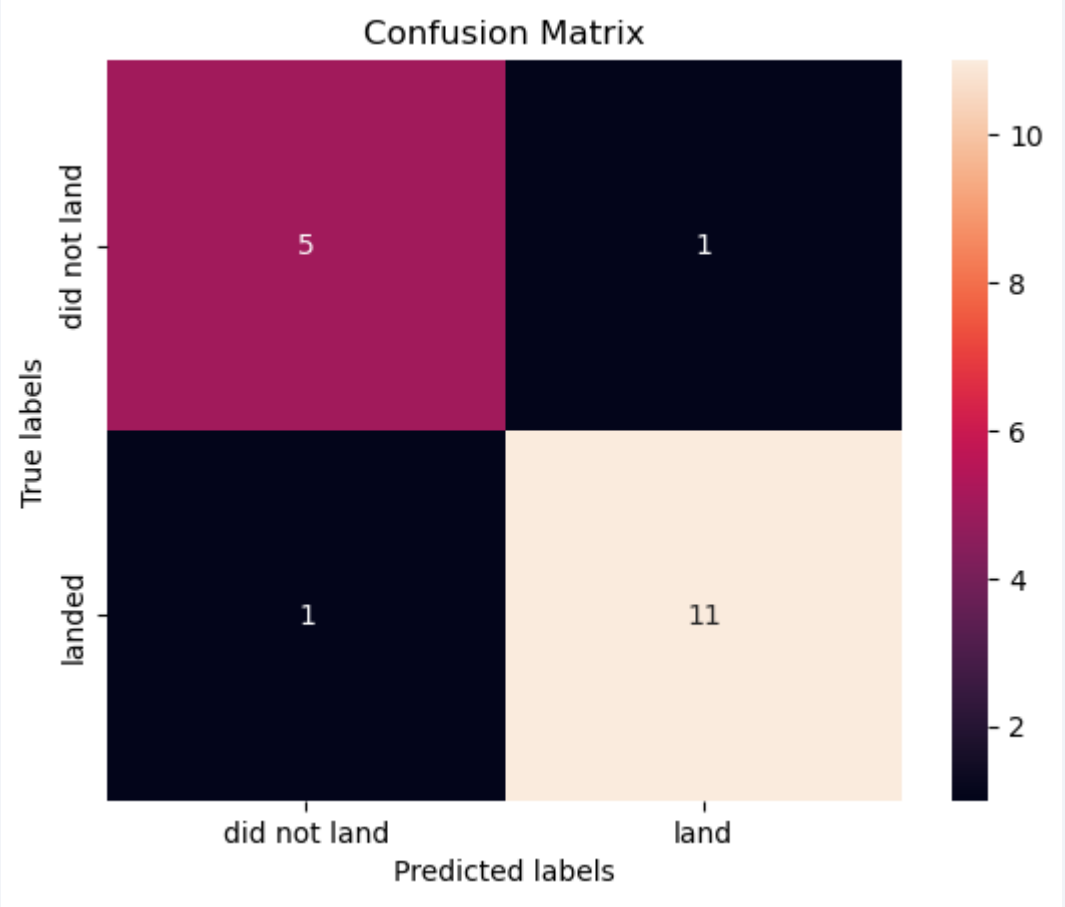
Classification Accuracy

- In the bar chart we can see the accuracies for all the models.
- The Decision Tree has the higher accuracy with 88,88%



Confusion Matrix for the decision tree

- This is de Confusion Matrix for the decision tree
- The confusion matrix shows that the decision tree classifier performs well, with a high number of correct predictions along the diagonal and relatively few misclassifications. This indicates that the model is effectively capturing the underlying patterns in the data.



Conclusions

- The Decision Tree model achieved the highest accuracy among the tested models, with a score of 88.9%.
- The EDA suggests that technological advancement over time is a key factor contributing to mission success. This is supported by the Flight Number, which shows the strongest correlation with other variables.
- The KSC LC-39A launch site appears promising due to its high success rate; however, its significance is limited by the relatively small sample size of only 13 recorded landings.

Thank you!