IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ala Marar
05/12/2024

Outline

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

Executive Summary

- **Objective**: Predict the success of Falcon 9 first stage landings.

- **Process**
  - **Data Collection**: Webscraped Falcon 9/Heavy launch data from Wikipedia/SpaceX.
  - **Data Wrangling**: Cleaned and formatted data for analysis.
  - **Exploratory Data Analysis** (EDA): Identified patterns and defined labels for supervised learning.
  - **Database Integration**: Loaded dataset into a SQL database for advanced queries.
  - **Feature Engineering**: Developed features to enhance predictive accuracy.
  - **Interactive Visualizations**: Utilized Folium for geospatial insights and created a dashboard with Dash.
  - **Machine Learning Pipeline**: Built and evaluated models for landing predictions.

- **Insights Gained:** Our analysis indicates correlations between launch features and outcomes, with Decision Tree emerging as the most effective algorithm for predicting Falcon 9 first-stage landing success.

Github Repo: Link

Introduction

SpaceX and Falcon 9

- SpaceX revolutionized space exploration with reusable rockets.
- Falcon 9 first stage landings are crucial for cost efficiency and sustainability.
- SpaceX pricing is around $62 million vs $165 million normally

Problem Statement

- Determining the likelihood of a successful first-stage landing remains complex.
- Accurate predictions can inform design, planning, and mission strategy.

Project Objective

- Build a predictive model to determine if the Falcon 9 first stage will successfully land, using historical launch data. This information can be used if an alternate company (Space Y) wants to bid against SpaceX for a rocket launch.

Main Question: For a set of features related to a Falcon 9 rocket launch (such as payload mass, flight number, orbit type and more), will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology – Data Collection

Data Source: Space X API and Wikipedia Falcon 9 and Falcon Heavy launch records.

Tools Used:

- **Python Libraries:** Helper functions, Beautiful Soup.
- **HTTP Requests:** Retrieved data using a GET request.

Data Extraction:

- Extracted HTML tables containing launch data.
- Parsed table headers to identify variable names.

Filtering:

- Retained only Falcon 9 launches for analysis.

Data Cleaning:

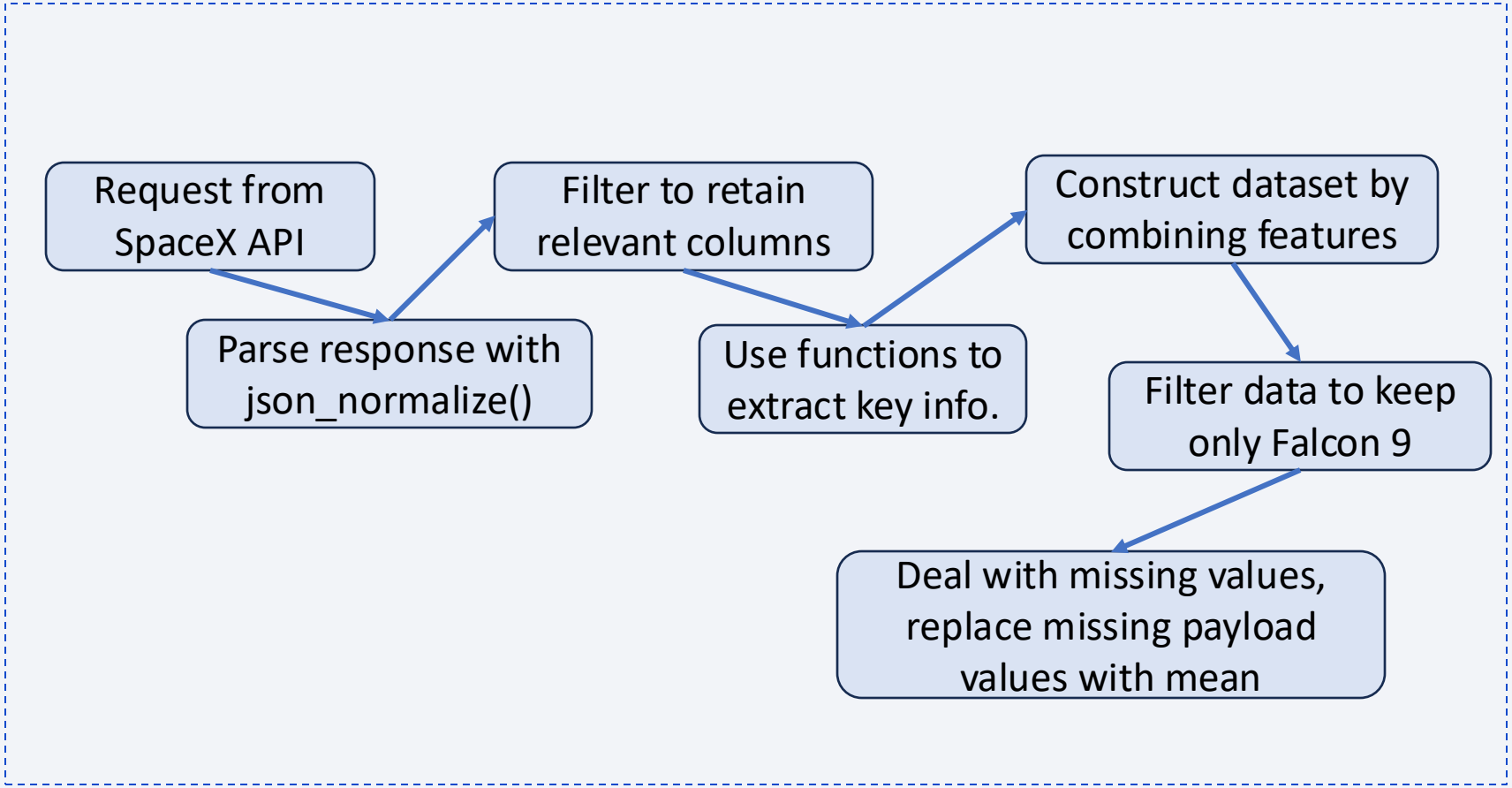
- Ensured correct data types for all variables.
- Identified and addressed missing data.
- Replaced missing values in **Payload Mass** with the mean.

Data Collection – SpaceX API

API Data

- The API used for SpaceX data is https://api.spacexdata.com/v4
- From the API we extract the Booster Version, LaunchSite, PayloadData and Core Data
- Gives us 90 rows and 17 features

Github URL: Link

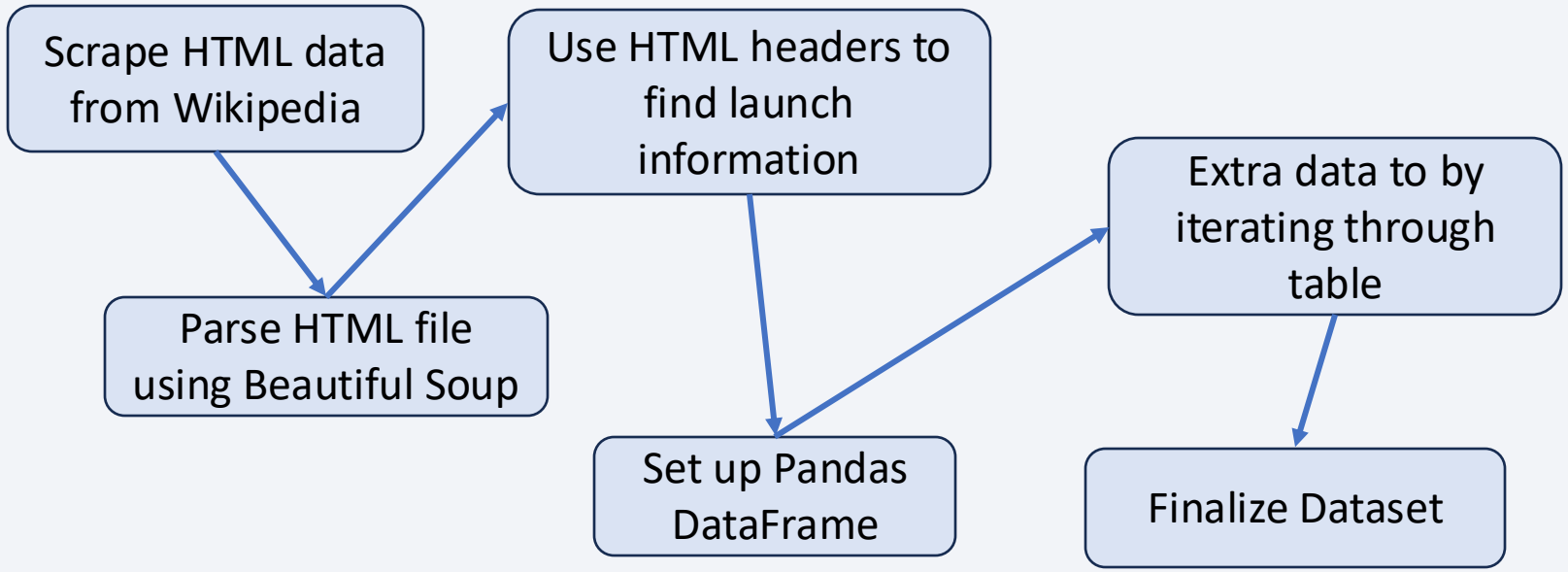


Data Collection – Web Scraping

Wikipedia Data

- We web scrape data from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Gives us 121 rows and 11 features

Github URL: Link



Methodology - Data Wrangling

- We create a landing outcome class label where a landing was successful (1) and where it was a failure (0)

- This is done based on the outcome column. The outcome column would have both the outcome component and the landing location component.

- We use this to determine a column label a a training label

Github URL: Link

Methodology – EDA with Visualization

Data Exploration:

- Counted launches by site.
- Analysed frequency and types of orbits.
- Examined mission outcomes for each orbit.

Visualizing Patterns:

- Flight Number vs. Launch Site: Explored for successful and failed outcomes.
- Payload Mass vs. Launch Site: Analyzed for success and failure patterns.
- Used Scatter plots, bar plots, and line plots to help assess the relationship between variables

Orbit Analysis:

- Examined success rates by orbit type.
- Investigated the relationship between flight number, orbit type, and success likelihood.

Temporal Trends:

- Analysed changes in success rates and patterns over time (years).

Github URL: Link

EDA with SQL

Loaded data into IBM DB2 Database

- Queried data from database using SQL Lite
- Used SQL to understand the data further before predictive analysis

SQL Analysis:

- Identified boosters with the highest successful landings.
- Calculated failures and successes by month.
- Counted occurrences of different outcomes.

Github URL: Link

Methodology – Interactive Visualization with Folium

Interactive Mapping:

- Plotted **launch sites** on an interactive map using **Folium**.
- Marked launch sites based on **success** or **failure** outcomes.

Proximity Analysis:

- Calculated **distances** between launch sites and nearby features (e.g., railways, highways, coastlines, cities).
- Visualised the possible the impact of proximity on launch success or failure.

Objects added:

- Markers for all launch sites on the map, succeeded/failed launches, line for distances to major points, and a longitude and latitude marker for your mouse (cursor)

Github URL: Link

Build a Dashboard with Plotly Dash

Dashboard Creation:

- Created a dashboard with dropdown functionality to help visualise the data. This was done with **Dash.**

Visualisations created:

- We display the total successful launches from the different launch sites. This is done with a **pie chart.**
- We display the correlation between mission outcome and the payload mass to help us identify if there is a relationship between mass and outcome. This is done with a **scatter plot**.

Customisability

- The dashboard allows users to choose a specific launch site to manipulate the pie chart and see its success rate.

GitHub URL: Link

Methodology – Predictive Analysis Preparation

Feature Engineering:

- Created dummy categorical variables for non-numeric features.
- Corrected variable types, e.g., converting numeric columns to floats.

Data Standardization:

- Standardized the data to ensure consistent scale for model training.

Data Splitting:

- Divided data into training and testing sets for model evaluation.

Ready for Predictions:

- Data is now prepared for machine learning model training and testing.

Methodology – Predictive Analysis

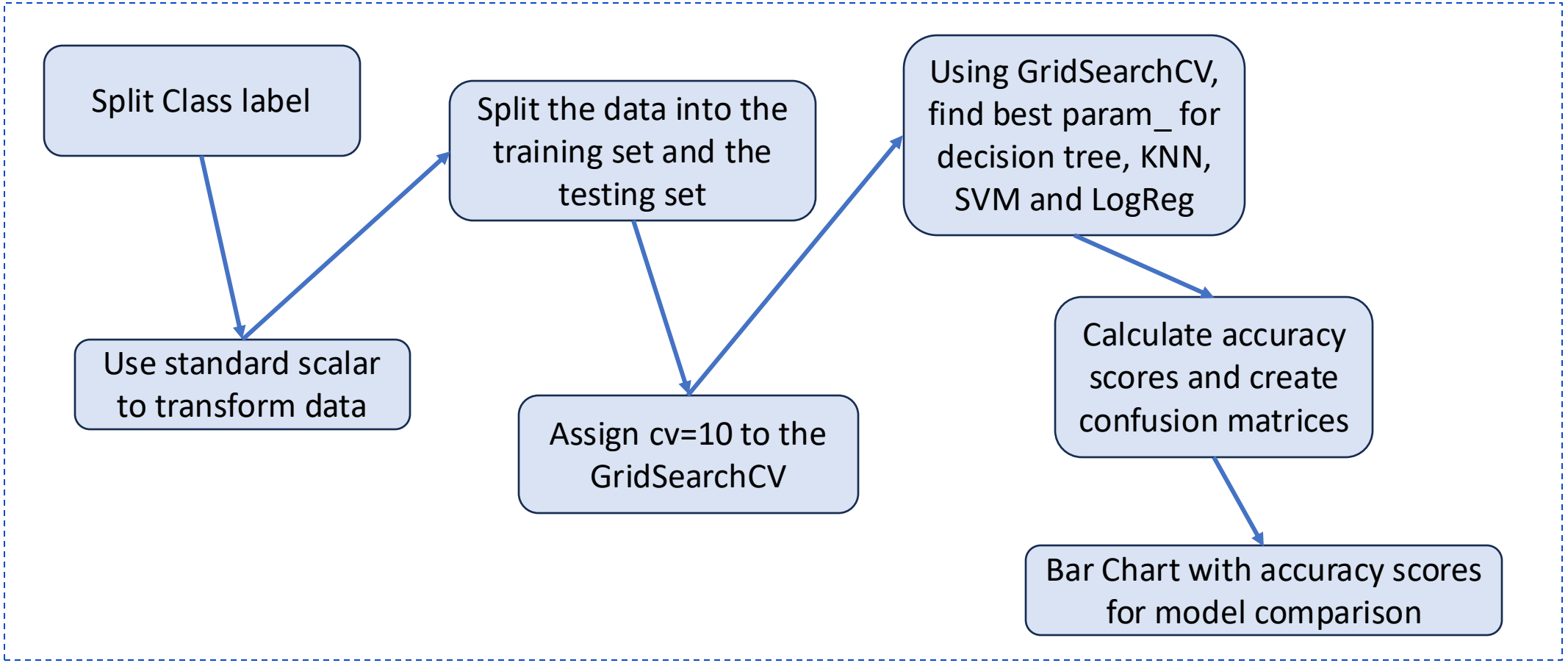
Model Creation & Hyperparameter Tuning:

- **Logistic Regression**: Used GridSearchCV to find the best parameters.
- **Support Vector Machine (SVM):** Tuned hyperparameters using GridSearchCV.
- **Decision Tree Classifier:** Optimized with GridSearchCV for best parameters.
- **K-Nearest Neighbours (KNN):** Tuned using GridSearchCV for optimal performance.

Model Evaluation:

- Tested each model on the test data.
- Visualized accuracy scores to determine the best-fitting model.

Methodology – Predictive Analysis

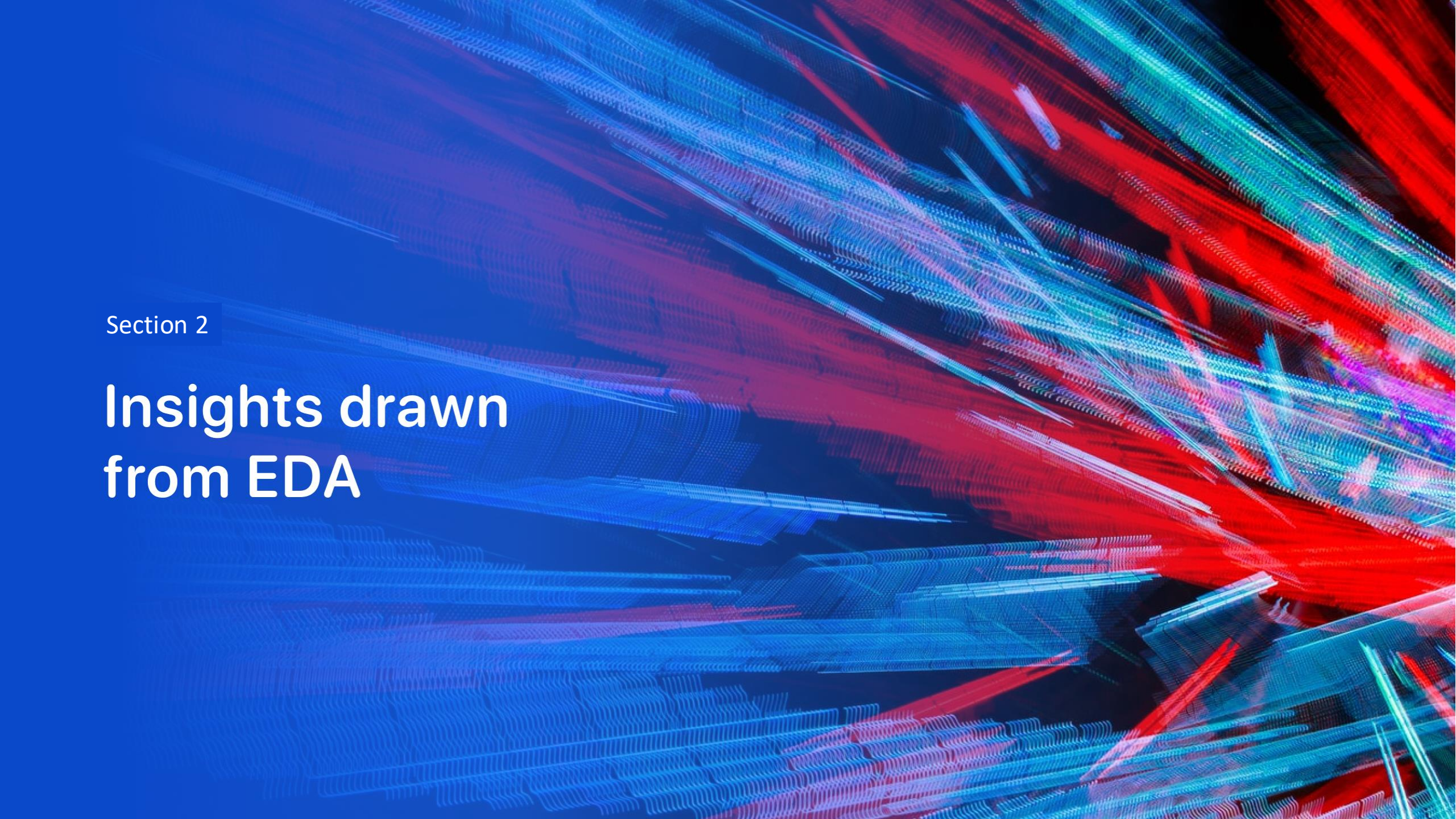


GitHub URL: Link

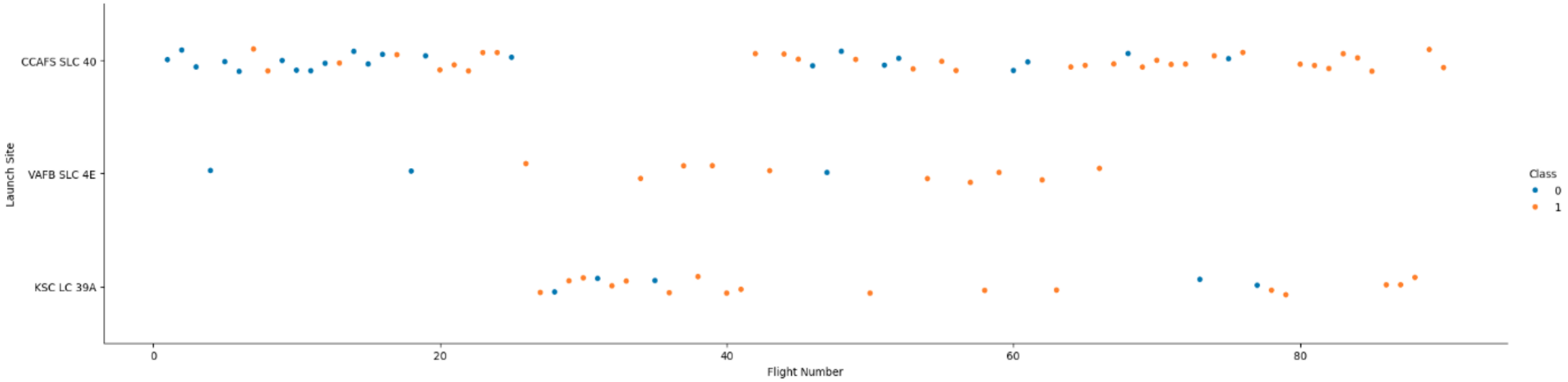
Results

The results are now going to be split in several sections:

- **Exploratory data analysis results**
- **Interactive analytics demo in screenshots (Plotly and Dash)**
- **Predictive analysis results**


Section 2
Insights drawn from EDA

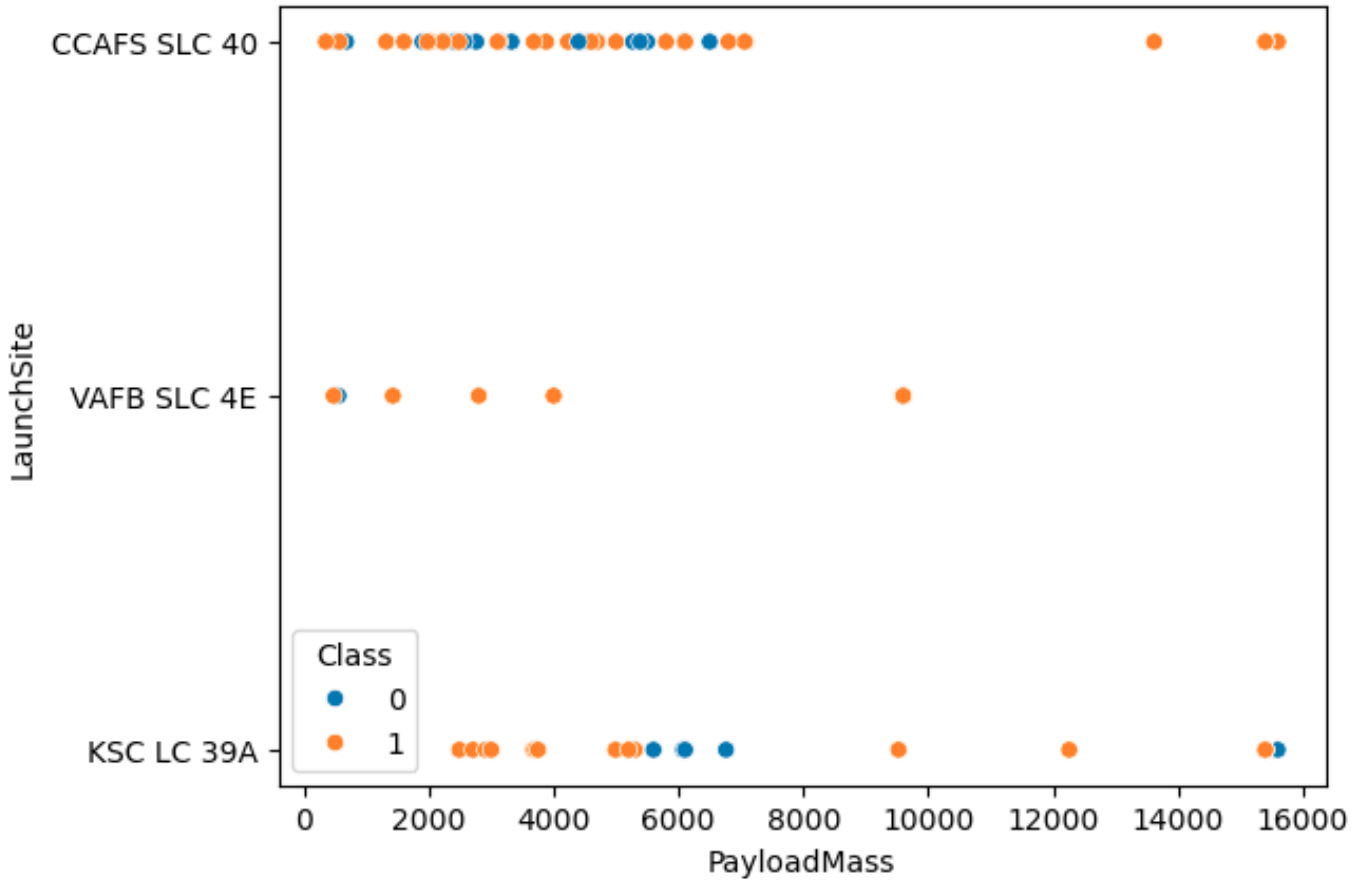
Flight Number vs. Launch Site



- This is a scatter plot of Flight Number vs. Launch Site where the class indicates success (1 - orange) or failure (0 - blue)
- We can see that at the highest flight numbers (above 80) the success rate is 100%
- We can see that below 10 the failure rate is quite high for the CCAFS SLC 49

Payload vs. Launch Site

- This is a scatter plot of Payload vs. Launch Site. Where class indicates success / failure
- We can see that for CCAFS SLC 40 launchsite, they have a higher success rate at above 6000KG payload
- We can see that VAFB SLC 4E has the highest success rate.



Success Rate vs. Orbit Type

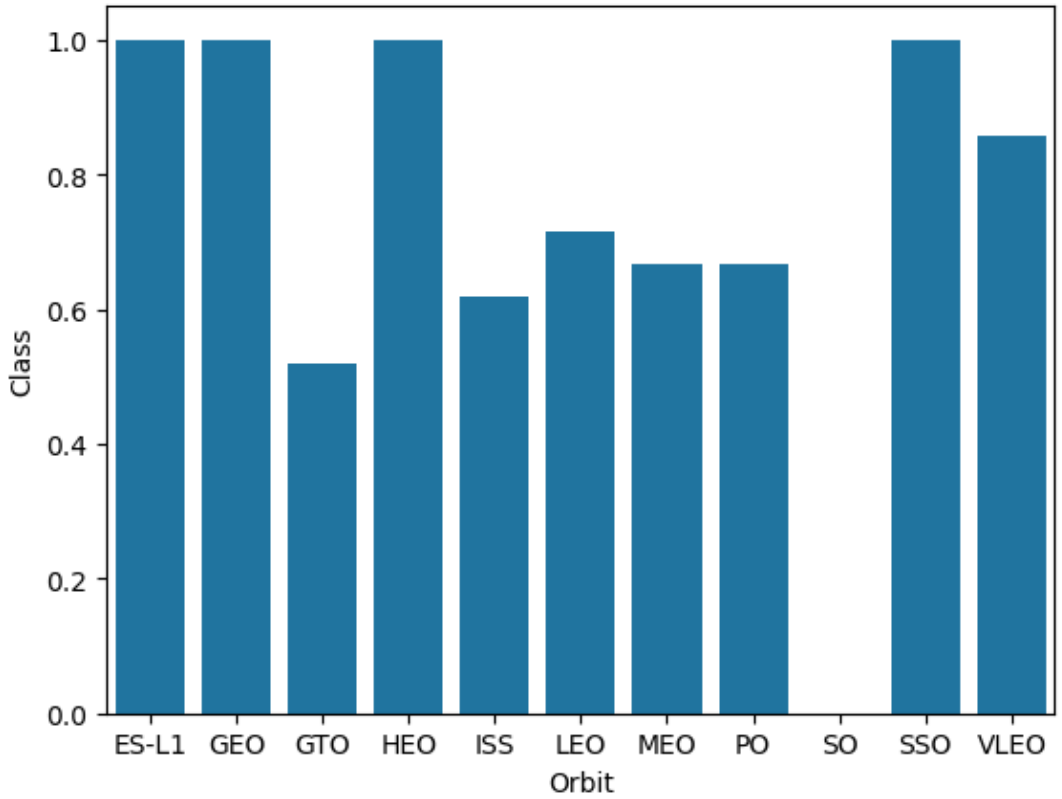
- This is a bar chart for the success rate of each orbit type

- We can see that SSO, HEO, GEO and ES-L1 have the highest success rate
- We can see that SO has a 100% failure rate.

```
# HINT use groupby method on Orbit column and get the mean of Class column
success_rate = df.groupby("Orbit")["Class"].mean().reset_index()

sns.barplot(x="Orbit",y="Class",data = success_rate)
```

```
<AxesSubplot:xlabel='Orbit', ylabel='Class'>
```

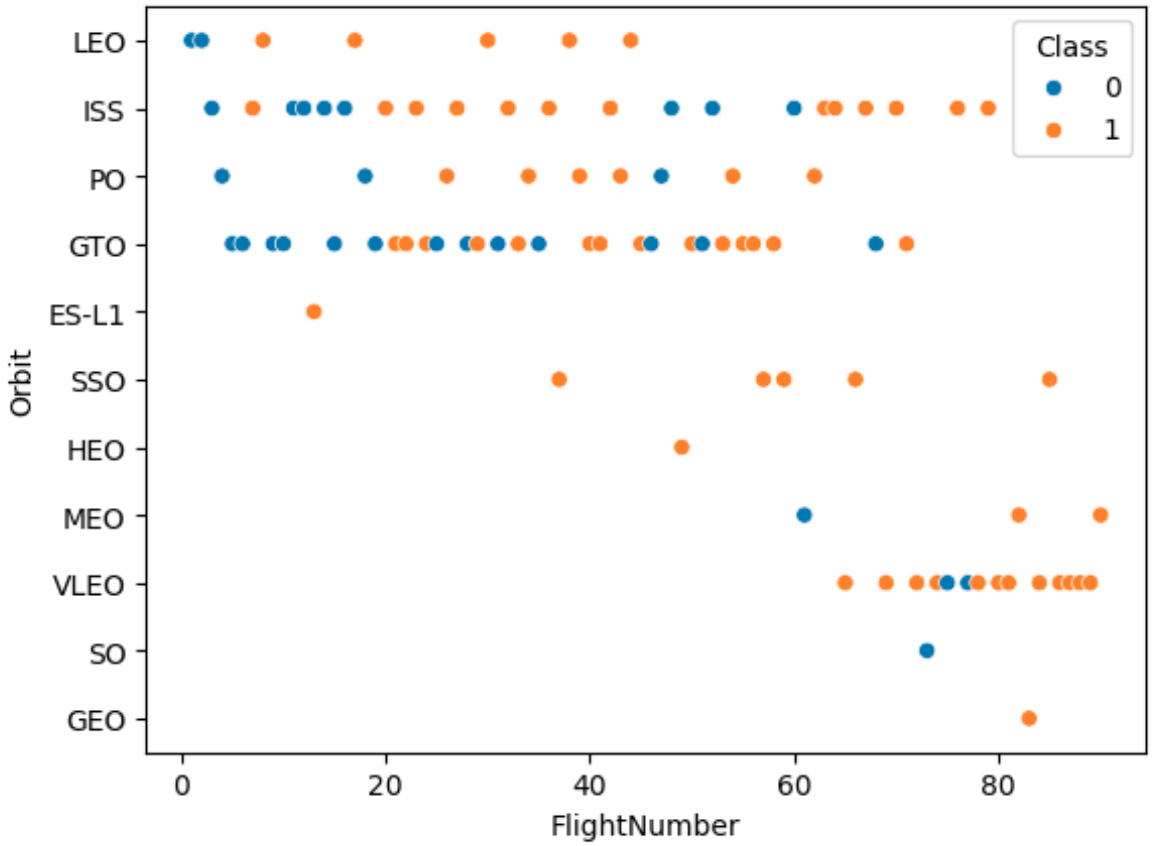


Analyze the plotted bar chart to identify which orbits have the highest success rates.

Flight Number vs. Orbit Type

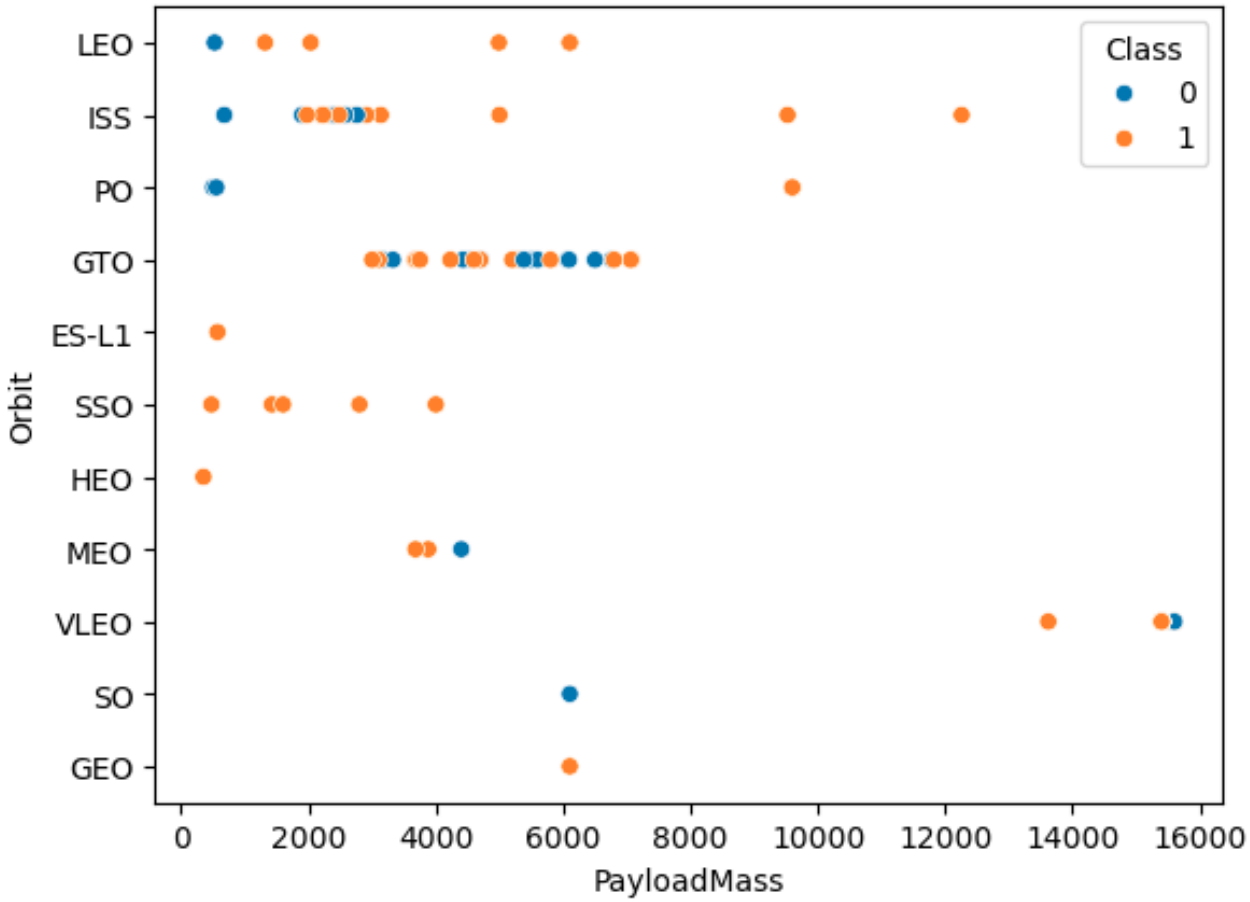
- This is a scatter plot of Flight number vs. Orbit type, where class indicates success/failure

- We can see that for LEO and MEO, the higher the flight number the greater the success rate.
- We can see that for some orbits there is no correlation between flight number and success rate



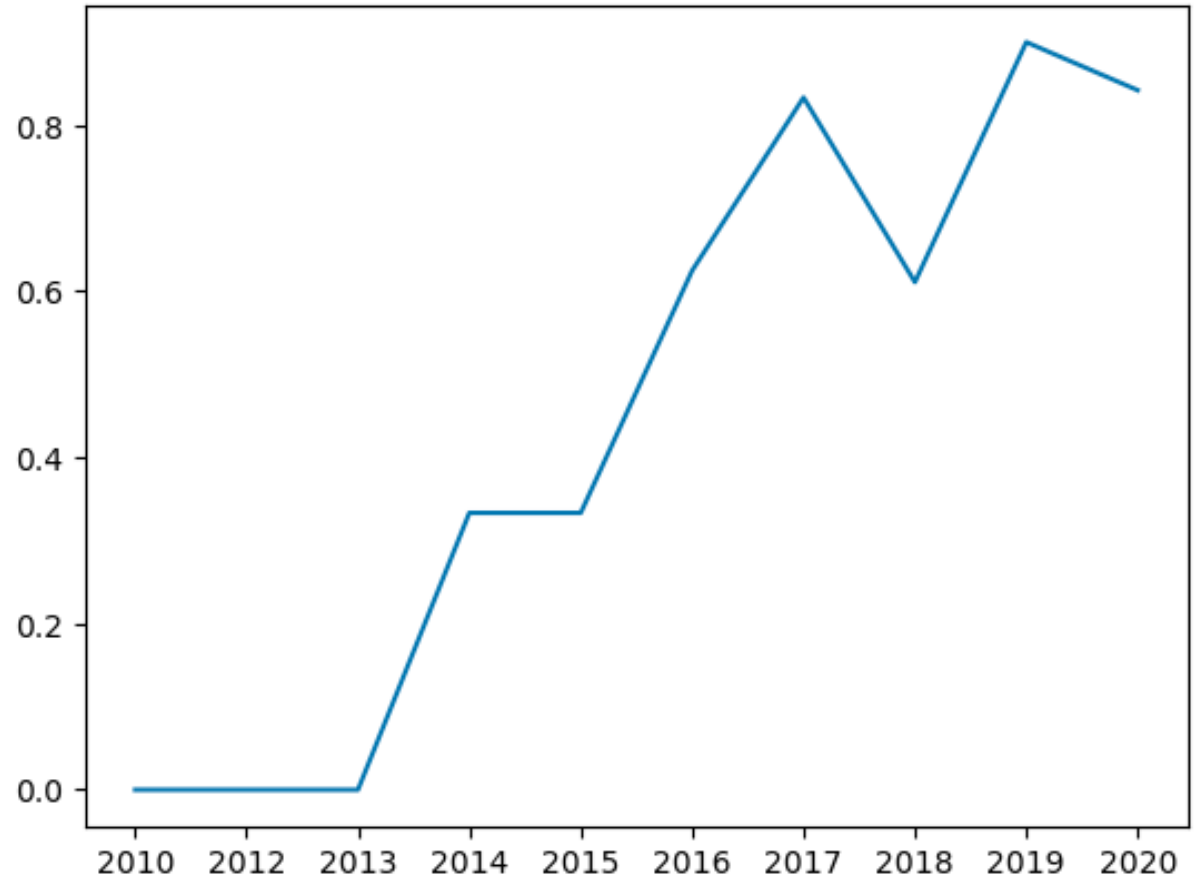
Payload vs. Orbit Type

- This is a scatter plot of payload vs. orbit type. Where class 1 indicates a success.
- We can see that for LEO, ISS and PO, the higher the payload mass the higher the success rate.
- We can see that ES-L1, SSO, HEO, GEO have the highest success rate.



Launch Success Yearly Trend

- Show a line chart of yearly average success rate
- Show the screenshot of the scatter plot with explanations



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

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

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

We extract 5 records where the name of the launch site starts with ”CCA”, we do this using the ”LIKE” method with a % after CCA. We limit the response to 5 so we only get 5 records.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE "NASA%"
```

```
 * sqlite:///my_data1.db
Done.
```

SUM(PAYLOAD_MASS__KG_)
99980

We calculate the total payload mass by boosters that NASA launched. We do this by filtering ”Customer” to NASA, and using the SUM function in SQL on the payload mass variable. The total payload mass is **99,980KG**

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload FROM SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1%"
```

```
 * sqlite:///my_data1.db
Done.
```

average_payload
2534.6666666666665

We calculate the average payload mass for booster versions F9 v1.1. The average payload is approx. 2534.67KG

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT Date as first_successful_landing FROM SPACEXTBL WHERE Mission_Outcome = "Success" ORDER BY Date ASC Limit 1
```

```
 * sqlite:///my_data1.db
Done.
```

first_successful_landing
2010-06-04

After ordering all successful landings by date (in ascending order) we take the first record using the limit 1 method. We find the first successful landing to be on **04-06-2010**.

Successful Drone Ship Landing with Payload between 4000 and 6000

We list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE Mission_Outcome = "Success" AND Landing_Outcome LIKE "%drone%" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Total Number of Successful and Failure Mission Outcomes

successful	failure
100	1

We calculate the number of successful and failure mission outcomes. We do this by using the CASE WHEN method to filter missions for successes and failures separately. We find that there were 100 success and 1 failure.

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT DISTINCT(Booster_Version) FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

We list the names of the booster which have carried the maximum payload mass

We find around 12 booster versions which 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

2015 Launch Records

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

We list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

We do this by converting strings into month format and year format. We then filter for failures and the year 2015.

We have two failures from 2015, in January and in April.

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

We 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

We find the following:

- launches where the landing was not attempted were the most frequent
- When attempted, most landings are attempted on drone ship.
- Landings on drone ship have a 50% success rate.

Landing_Outcome	landing_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

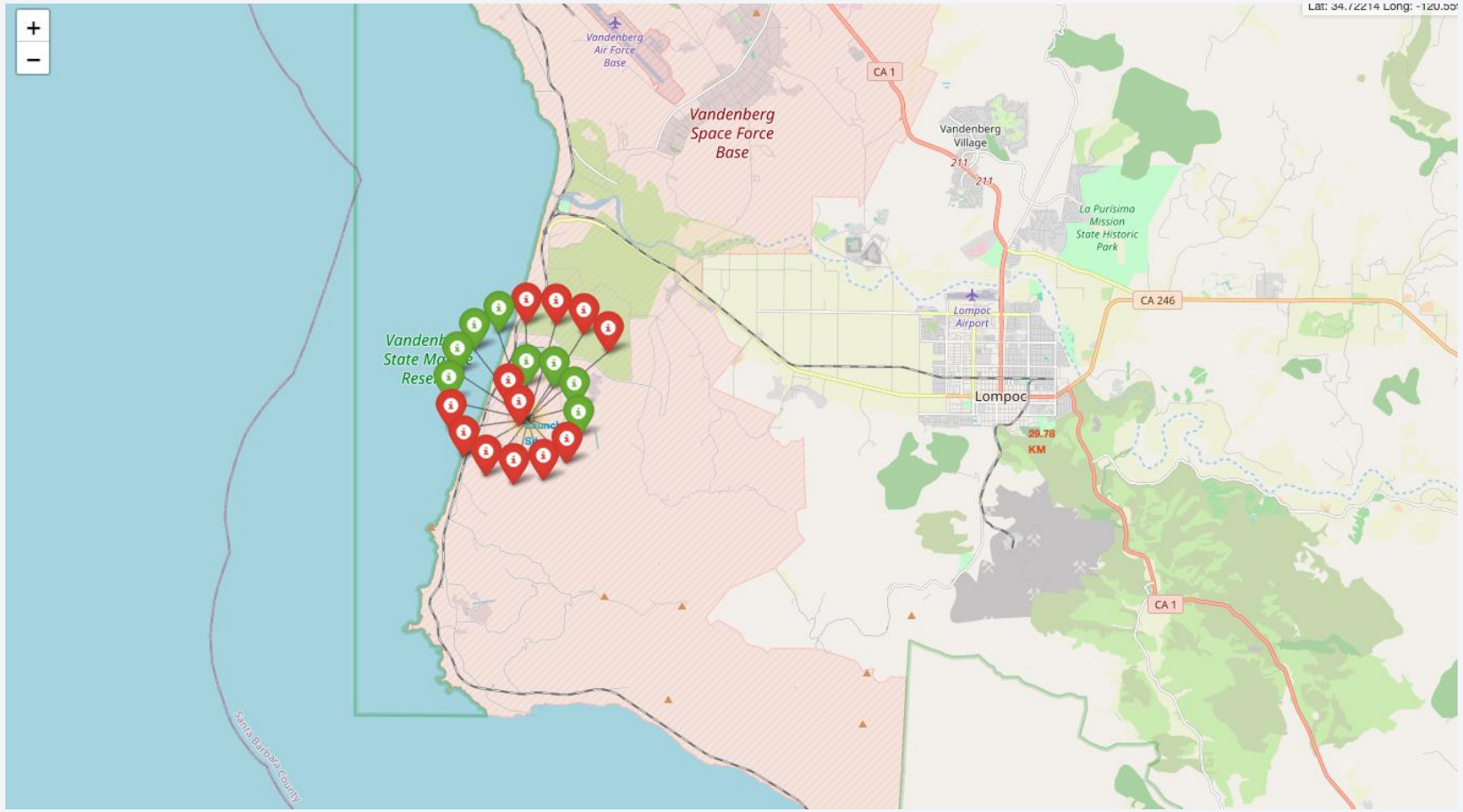
Section 3

Launch Sites Proximities Analysis

Initialize Folium Map

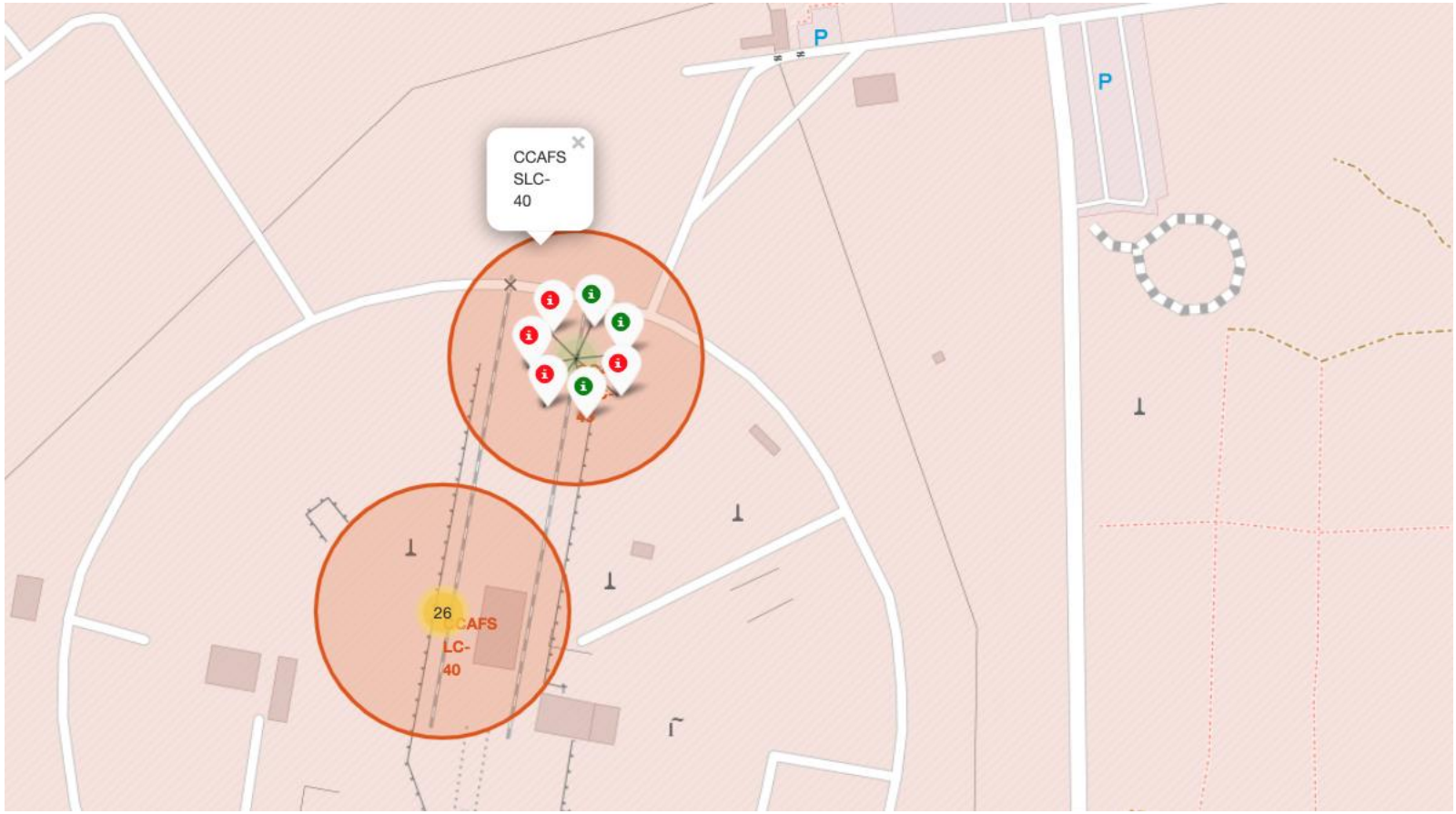
We first initialize the folium map and add the specific launch sites.

Folium Map with Success Indicator



We improve label visibility by color coding the successful vs failure launch sites.

Overlaying labels and the Mouse Position

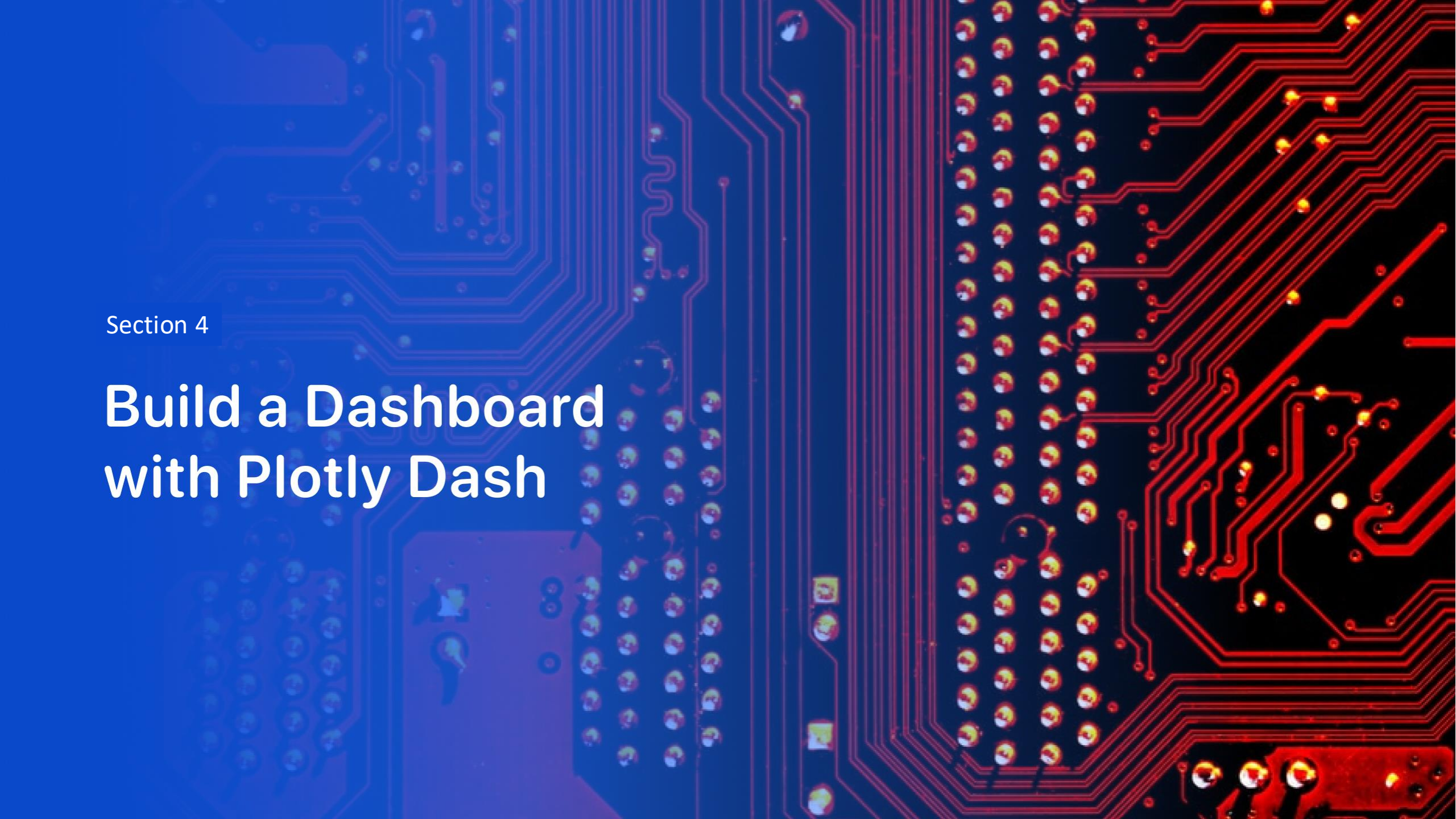


We introduce a label for the different launch site to help navigation on the map. We also introduce the mouse position so we can see the Longitude and Latitude of our mouse.

Folium Map with Proximities.

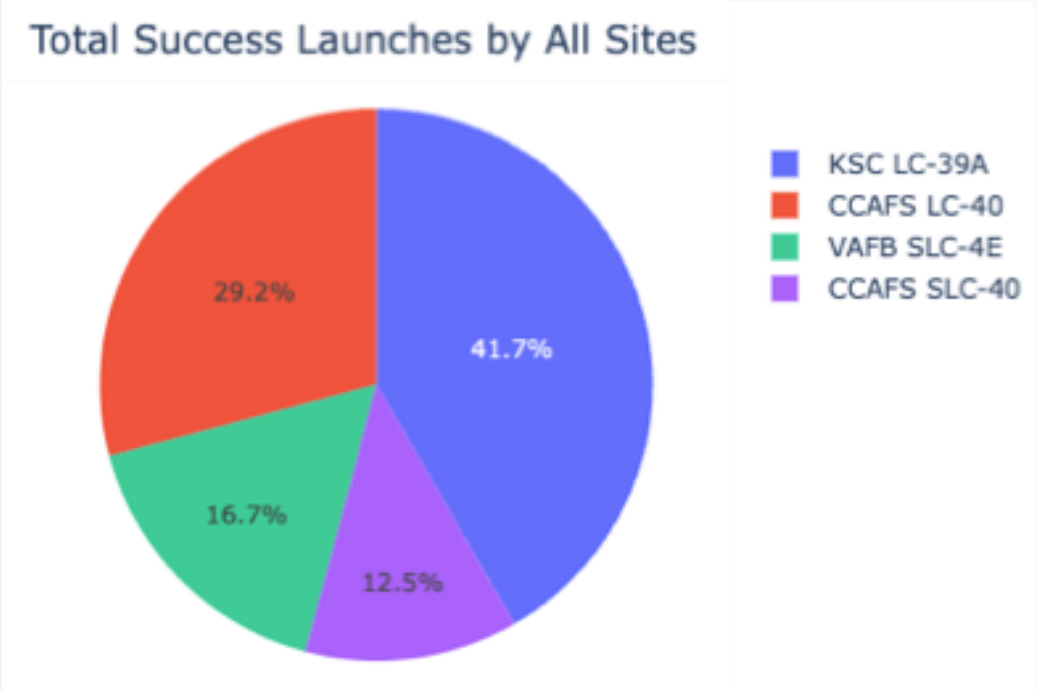
We choose a specific launch site and begin to calculate the approximate distance between it and various points, such as highways, railways and more.

In this photo, we look at the proximity from the coastline.


Section 4
Build a Dashboard with Plotly Dash

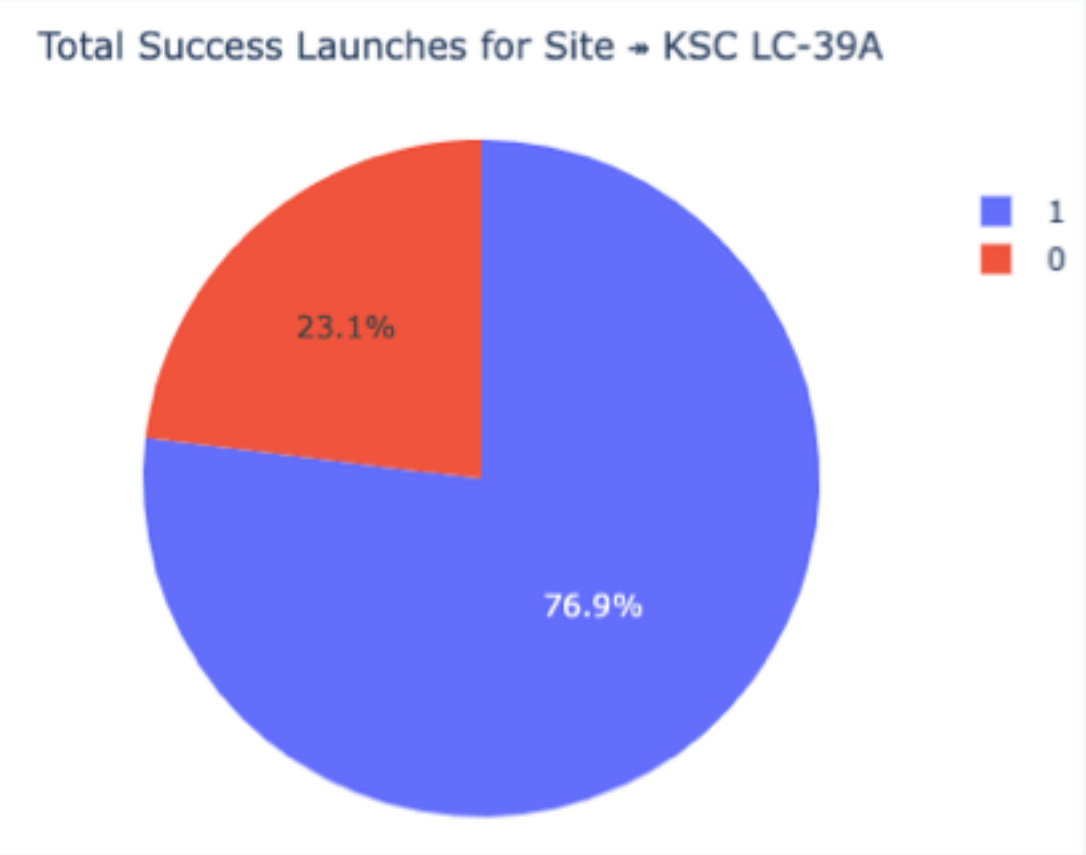
Total Success Launches by All Sites

- This is a pie chart indicating the percentage of successful launches by launch site.
- The legend indicates the color of the respective launch site.
- This is currently showing All Sites. The Launch Site KSC LC 39A is the most successful launch.



Success Rate for KSC LC -39A

- This is a pie chart showing the proportion of successful launches for the Site KSC LC -39A
- Shaded area in blue indicates the successful launches.
- For the KSC LC -39A, the success rate is 76.9%



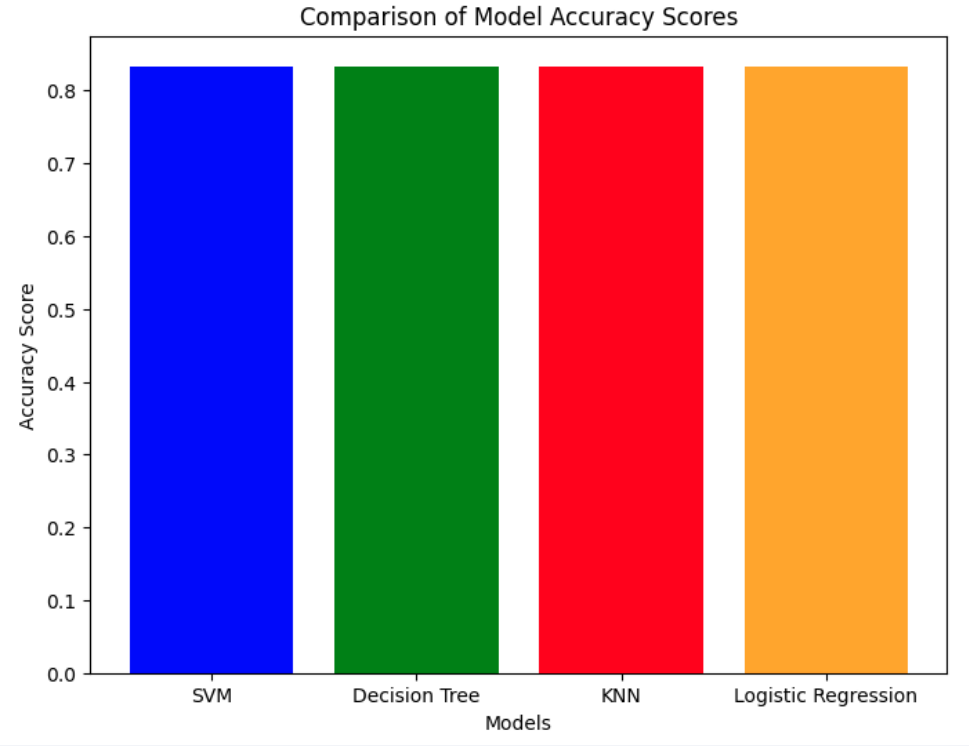
Payload Mass vs Success Rate

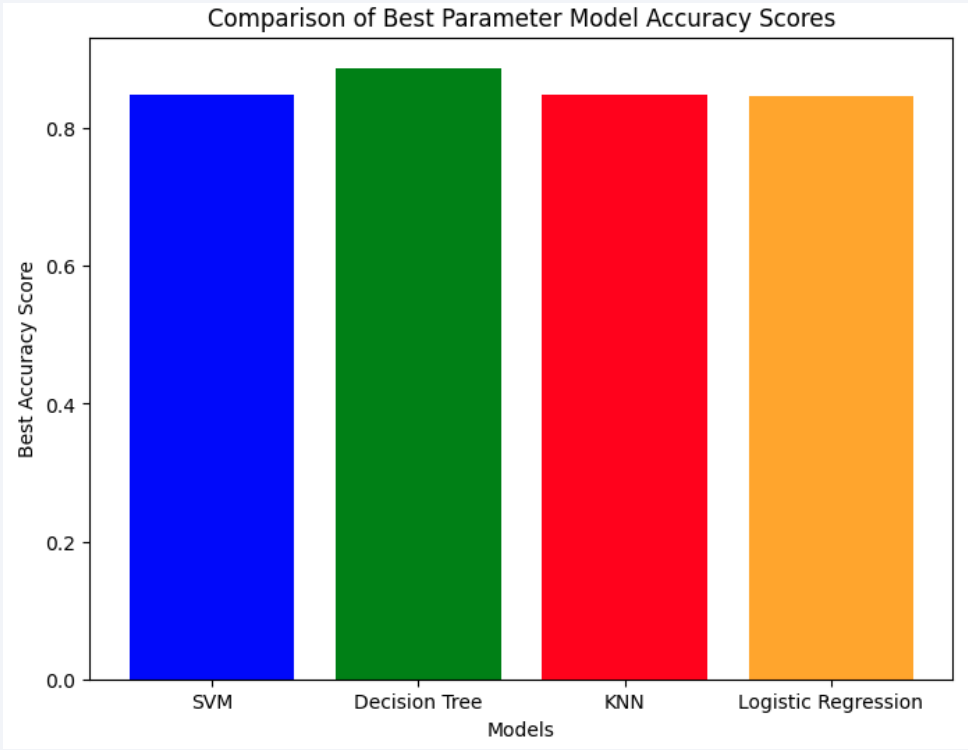
- This shows a Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- The legend indicates the booster version category
- The slider at the top allows us to constrict the payload range to manipulate the chart

Section 5

Predictive Analysis (Classification)

Classification Accuracy





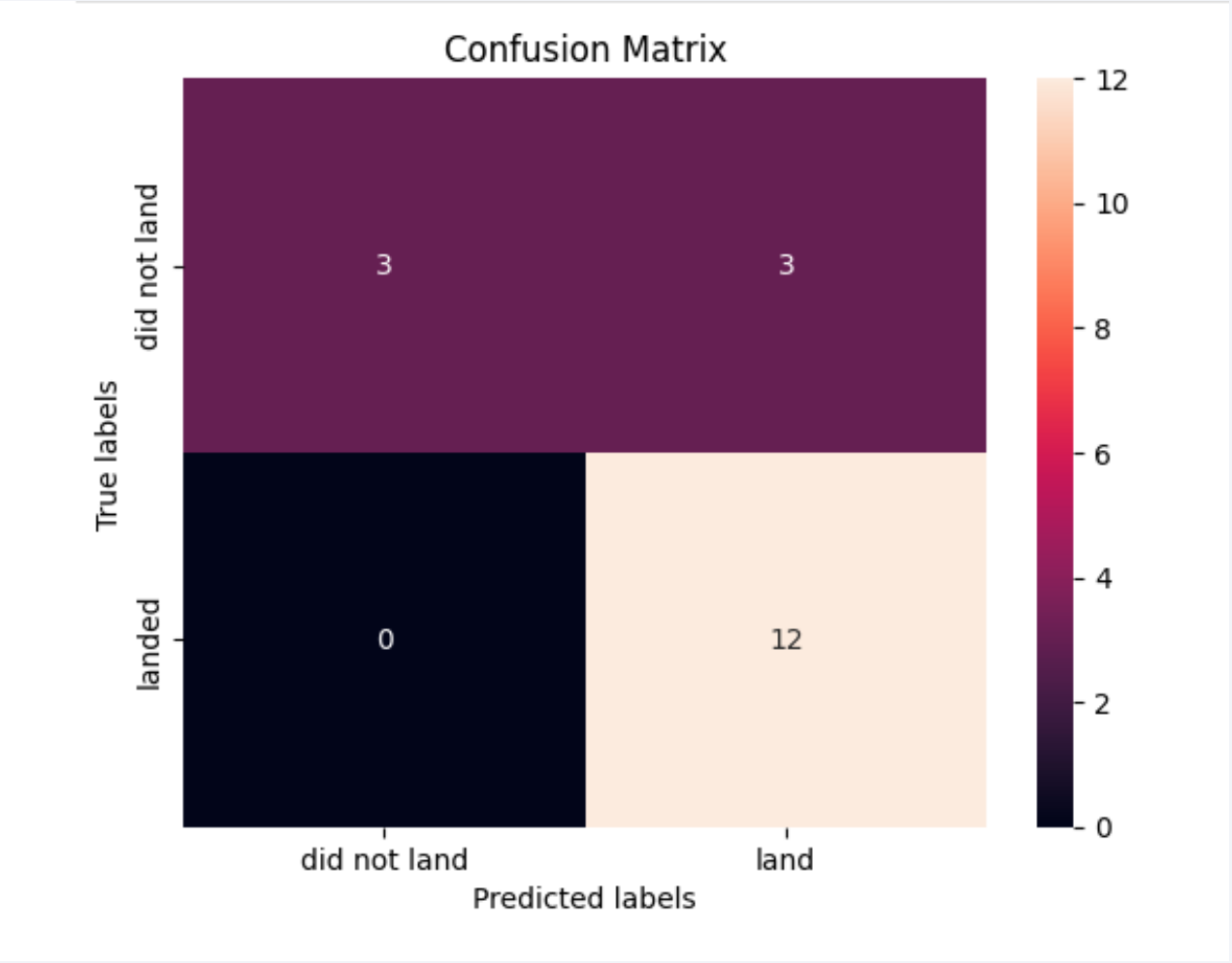
The Decision Tree model has the best accuracy. Even though the accuracy score on the testing data appears to be the same for all models. Under the best parameters, the accuracy score on the training set is highest for the decision tree model, with a score of: **0.8875**

Confusion Matrix

This is the confusion matrix for the decision tree model when tested on the test data.

It shows that of the launches that landed successfully, all 12 were correctly classified by the model.

However, of the 6 that did not land, only half were correctly classified as having landed.



Conclusions

Summary of Findings:

- Visualizations indicate possible correlations between features and mission outcomes.
  - **Payload Mass & Orbit Type:** Higher success rates observed for heavy payloads in **Polar**, **LEO**, and **ISS** orbits.
  - **GTO Orbit:** Success patterns are less distinct due to mixed outcomes.

Discussion:

- Each feature impacts the mission outcome, though the exact relationships are complex.
- Machine learning models help uncover patterns in historical data, offering predictive insights for future missions.

Project Achievements:

- Built predictive models to assess Falcon 9 first-stage landing success.
- **Decision Tree Model:** Achieved the highest accuracy among the four algorithms tested.

Conclusion:

- Predictive insights contribute to estimating launch costs and optimizing mission planning, aiding SpaceX's goal of cost-effective and sustainable space exploration.

Thank you!