APPLIED DATA SCIENCE CAPSTONE PROJECT

ROBIN MASAWI

23.02.2024

OUTLINE

Executive Summary

Introduction

Methodology

Results

Conclusion

EXECUTIVE SUMMARY

- **Methodologies Utilized**
  - Data collection via the SpaceX API and Webscraping
  - Data wrangling in Python
  - Exploratory data analysis with SQL and Data Visualization
  - Building an interactive visual map with Folium
  - Designing a dashboard with Plotly Dash
  - Predictive analysis using Classification

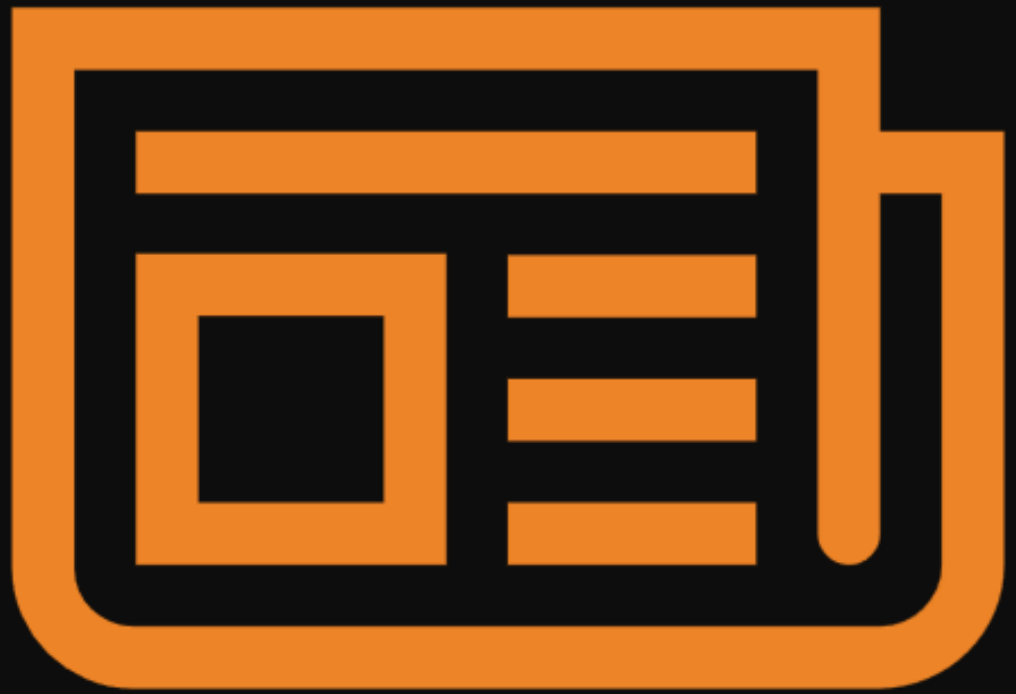
- **Results to be Showcased**
  - Exploratory data analysis
  - Interactive visual map and dashboard as screenshots
  - Classification modeling

INTRODUCTION

- **Project Background**

- The most prosperous business of the commercial space era, SpaceX has reduced the cost of space travel. On its website, the business promotes the launch of Falcon 9, a 62 million dollar rocket.
Due in large part to SpaceX's ability to reuse the first stage, other providers—whose costs can reach 165 million dollars each—can afford to charge less. Thus, we can calculate the launch cost if we can ascertain if the first stage will land. We will forecast if SpaceX will reuse the first stage based on available data and machine learning techniques.

- **Questions to be Addressed**
    - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
    - Do successful landings increase over the years?
    - What is the best algorithm to be utilized in this case?

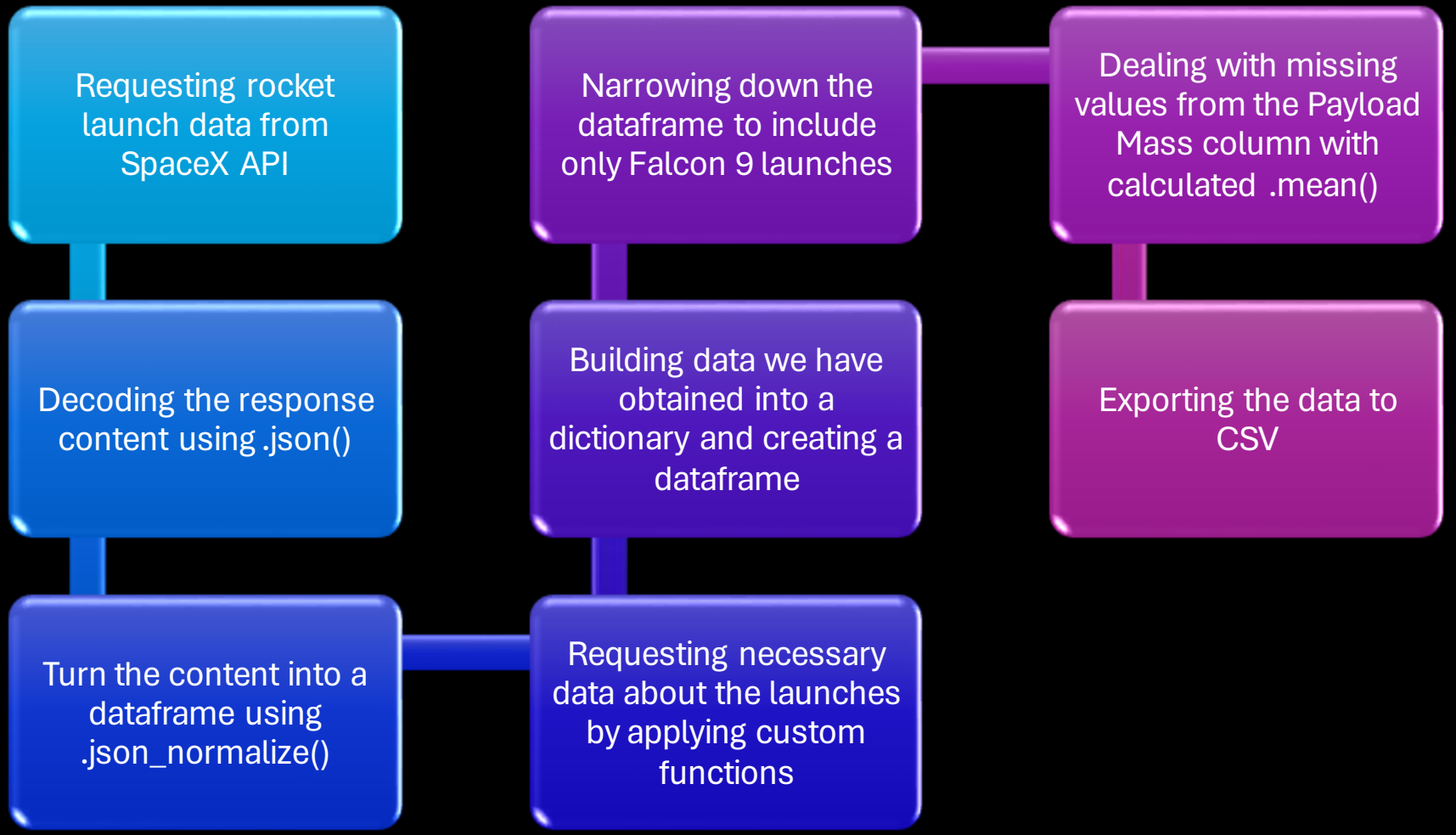


METHODOLOGY

- Data Collection: utilize SpaceX Rest API and Web Scrapping from Wikipedia.
- Data Wrangling: preparing the data to be ready for binary classification.
- Exploratory Data Analysis (EDA): made use of SQL and visualizations.
- Interactive Visual Analytics: applied Folium and Plotly Dash.
- Predictive Analysis: classification models came into play to ensure the best results.

Data Collection SPACEX API

- Data Columns Retrieved:
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude



Requesting Falcon 9 launch data from Wikipedia

Constructing a BeautifulSoup object from the HTML response

Taking the names of every column out of the HTML table header

Collecting the data by parsing HTML tables

Developing data obtained into a dictionary and creating a dataframe

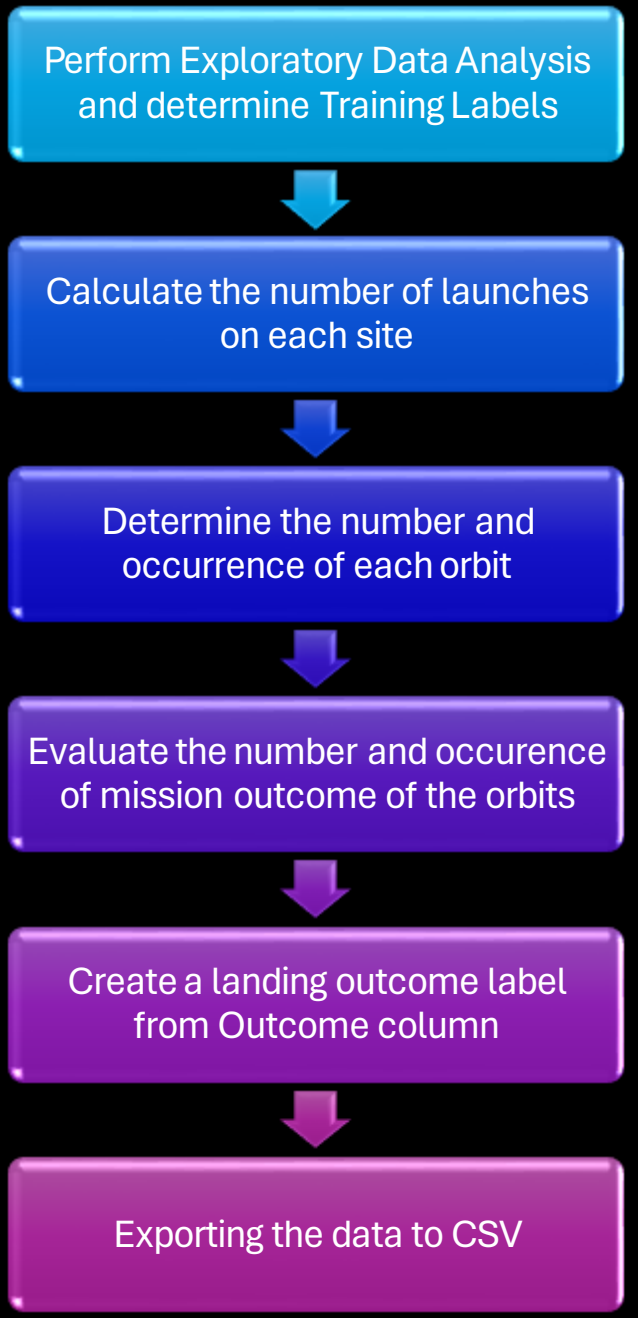
Exporting the data to CSV

Data Collection
Wikipedia Web Scraping

- Data Columns Retrieved:
  - Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Wrangling

- There are other instances in the data set where the booster failed to land properly.
- There are instances where a landing attempt was made but was unsuccessful due to an accident; for instance, True Ocean denotes a successful landing in a particular area of the ocean, whereas False Ocean denotes an unsuccessful landing in a particular area of the ocean.
- When a mission is declared successful and lands on a ground pad, it is said to have true RTLS. A mission outcome that was unsuccessfully landed on a ground pad is indicated by a false RTLS.
- A successful landing of the mission outcome on a drone ship is referred to as a true ASDS. False ASDS indicates that the mission's outcome—a drone ship—was not properly landed.
- Data wrangling was mainly about converting those outcomes into Training Labels with 1 as the booster successfully landed 0 as it was unsuccessful.



Visualizing Relationships

- Between Flight Number and Launch Site
- Between Payload and Launch Site
- Between success rate of each orbit type
- Between FlightNumber and Orbit type
- Between Payload and Orbit type
- Envision the launch success yearly trend

Feature Engineering

- Create dummy variables to categorical columns.
- Cast all numeric columns to float64

Exporting data to CSV

EDA with Data Visualization

- The main objectives are performing Exploratory Data Analysis and Feature Engineering using Pandas and Matplotlib.
- Charts utilized were:
  - Line charts – highlight data trends over time.
  - Scatter plots - exhibit the relationship between variables.
  - Bar charts - show comparisons among discrete categories.

EDA with SQL

- The main objective are to:
  - Understand the SpaceX DataSet.
  - Load the dataset into the corresponding table in a Db2 database.
  - Execute SQL queries to answer questions to be addressed.

SQL Queries Parsed

- Displaying the names of unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- 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

Interactive Visual Analytics Folium

- The primary goal is to identify certain geographic trends about launch sites by:
  - Marking all launch sites on a map.
  - Marking the success/failed launches for each site on the map.
  - Calculating the distances between a launch site to its proximities.

Mark all launch sites on a map

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts

Mark the success/failed launches for each site on the map

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates

Calculate the distances between a launch site to its proximities

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City

Interactive Visual Analytics Plotly Dash

- Building a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time.
- Components of the application include:
  - Launch Sites Dropdown List.
  - Pie Chart showing Success Launches (All Sites/Certain Site).
  - Slider of Payload Mass Range.
  - Scatter Chart of Payload Mass vs. Success Rate for different Booster Versions.

Launch sites dropdown list

- Added a dropdown list to enable Launch Site selection

Pie chart showing success launches (All Sites/Certain Site)

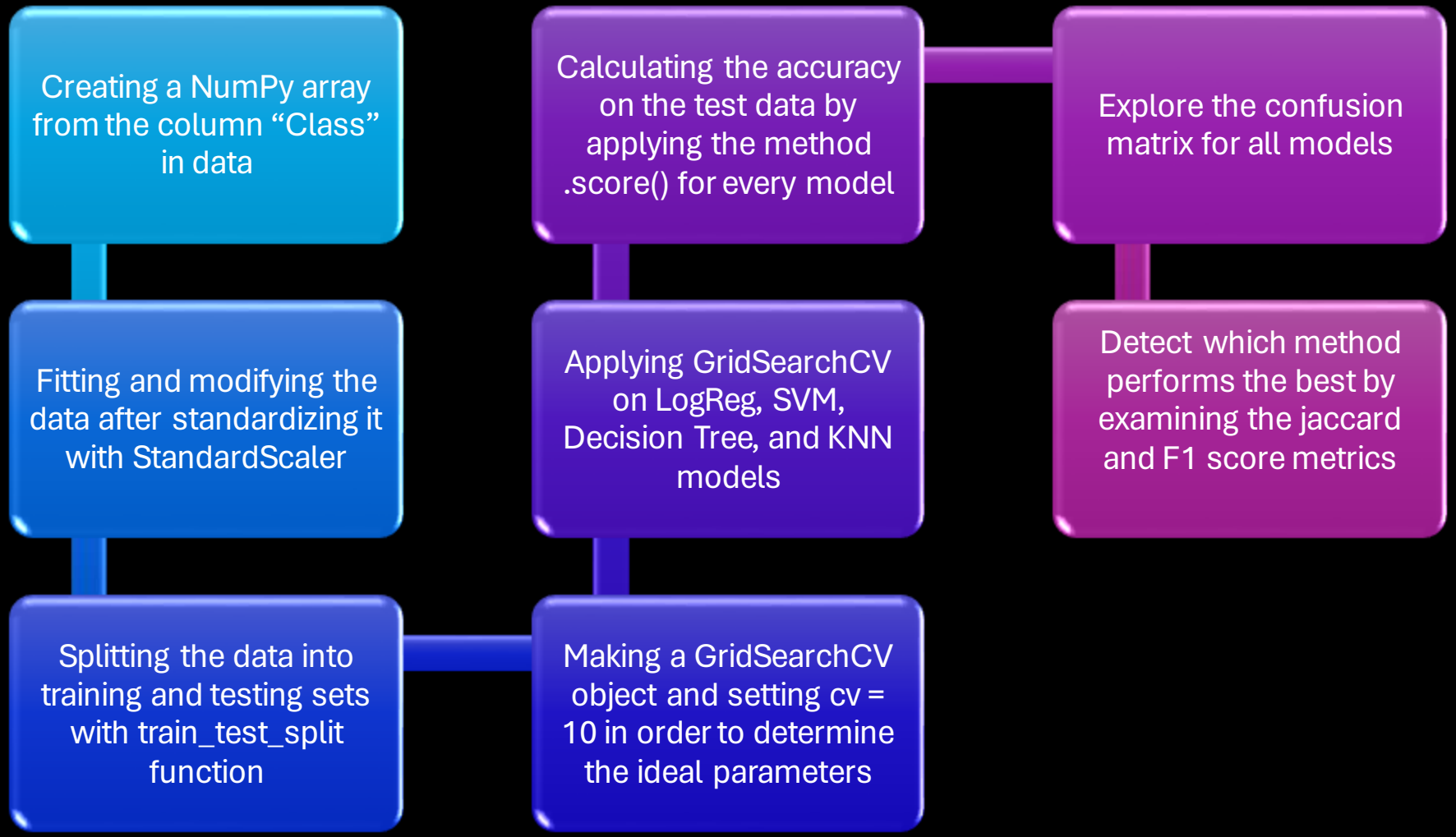
- Appended a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected

Slider for Payload Mass range

- Added a slider to select Payload range

Scatter chart of Payload Mass vs. Success Rate for different booster versions

- Attached a scatter chart to show the correlation between Payload and Launch Success



Predictive Analysis Classification

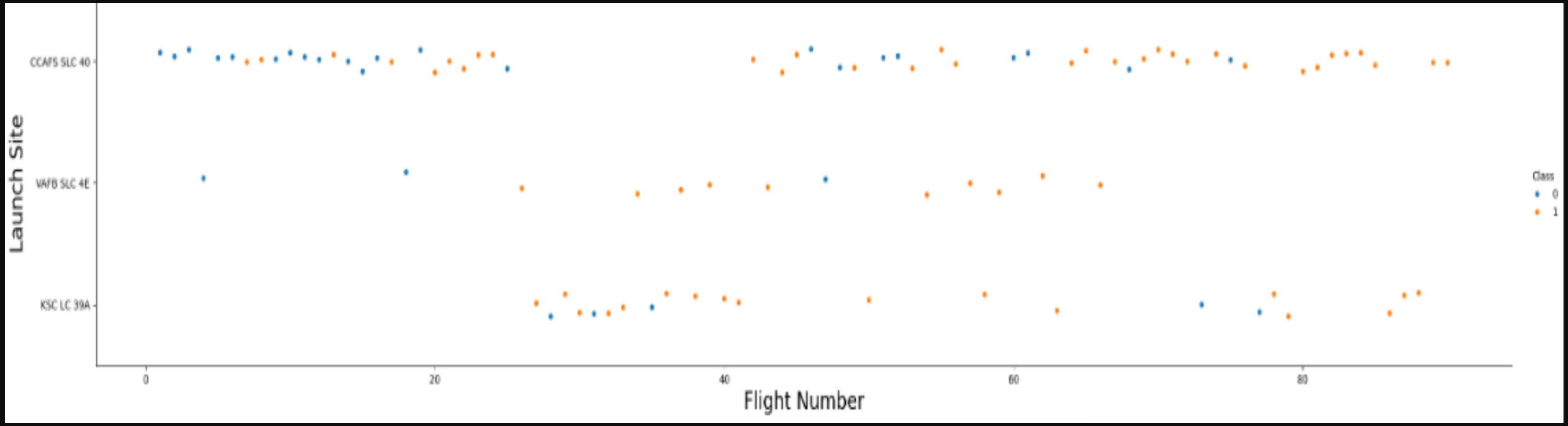
- The main objectives are:
  - Perform Exploratory Data Analysis and determine Training Labels.
    - Create a column for the class.
    - Standardize the data.
    - Split into training data and test data.
  - Detect the best hyperparameter for KNN, SVM, Decision Tree and Logistic Regression.
    - Find the method that performs best using test data.

RESULTS

- EDA with Data Visualization and SQL
- Interactive Map with Folium
- Plotly Dash dashboard
- Predictive Analysis (classification)

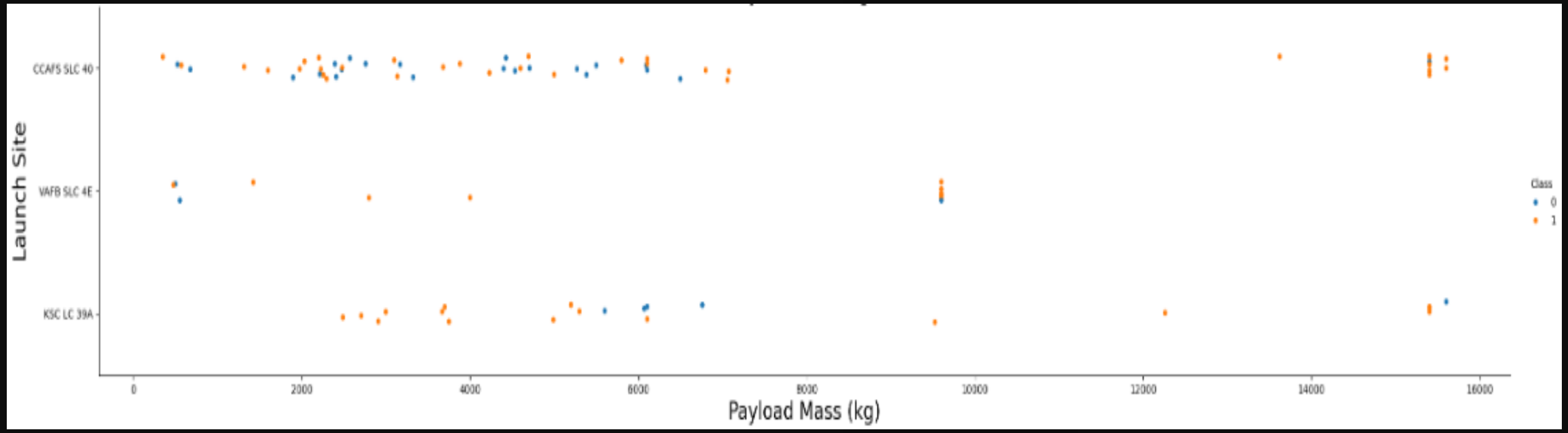
EDA with Visualization

Flight Number vs Launch Site



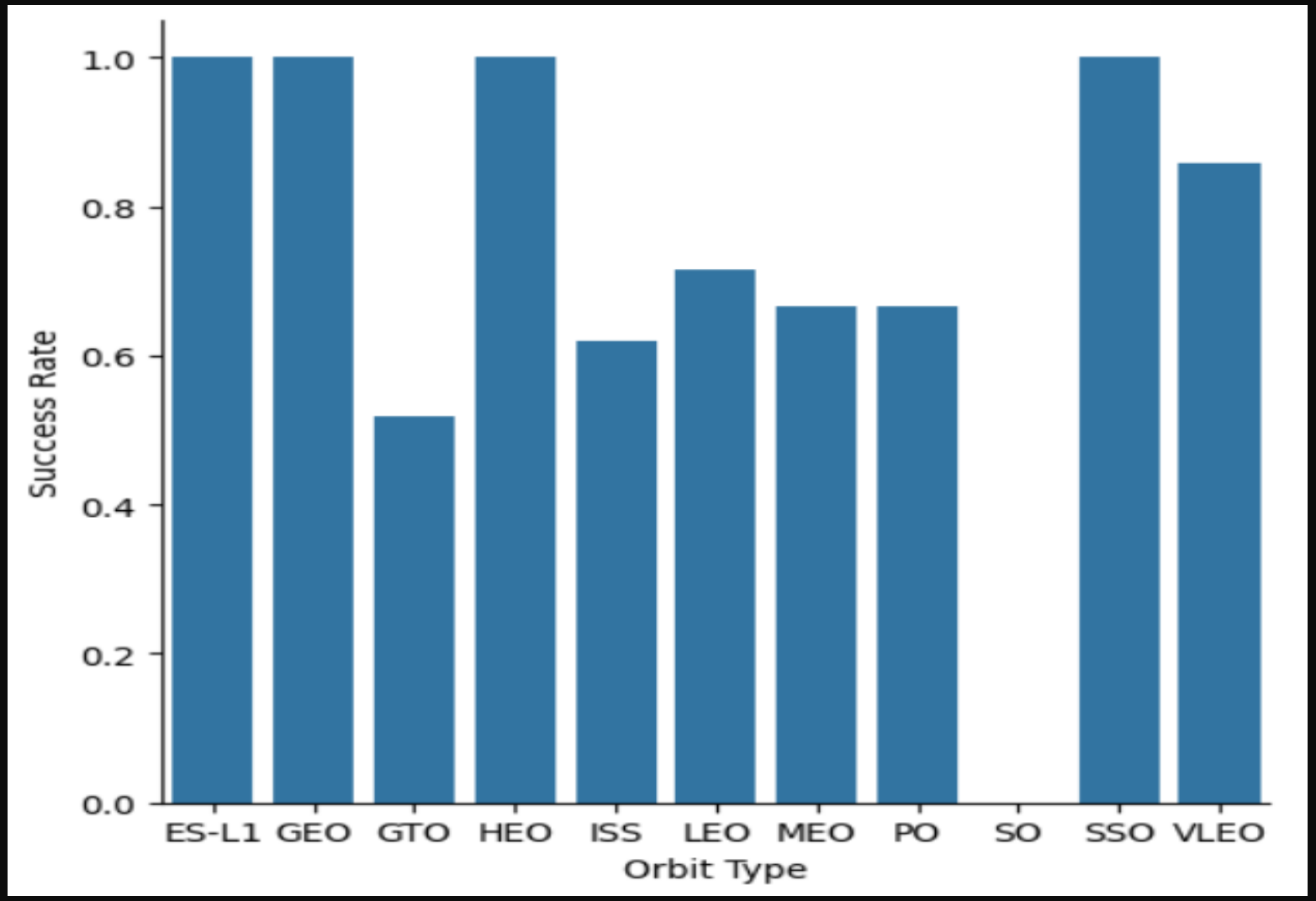
- Earliest flights all failed whereas the latest flights all succeeded.
- The assumption is that each new launch had a higher success rate.
- CCAFS SLC 40 launch site has the majority of the launches.
- VAFB SLC 4E and KSC LC 39A have much higher success rates.

Payload vs Launch Site



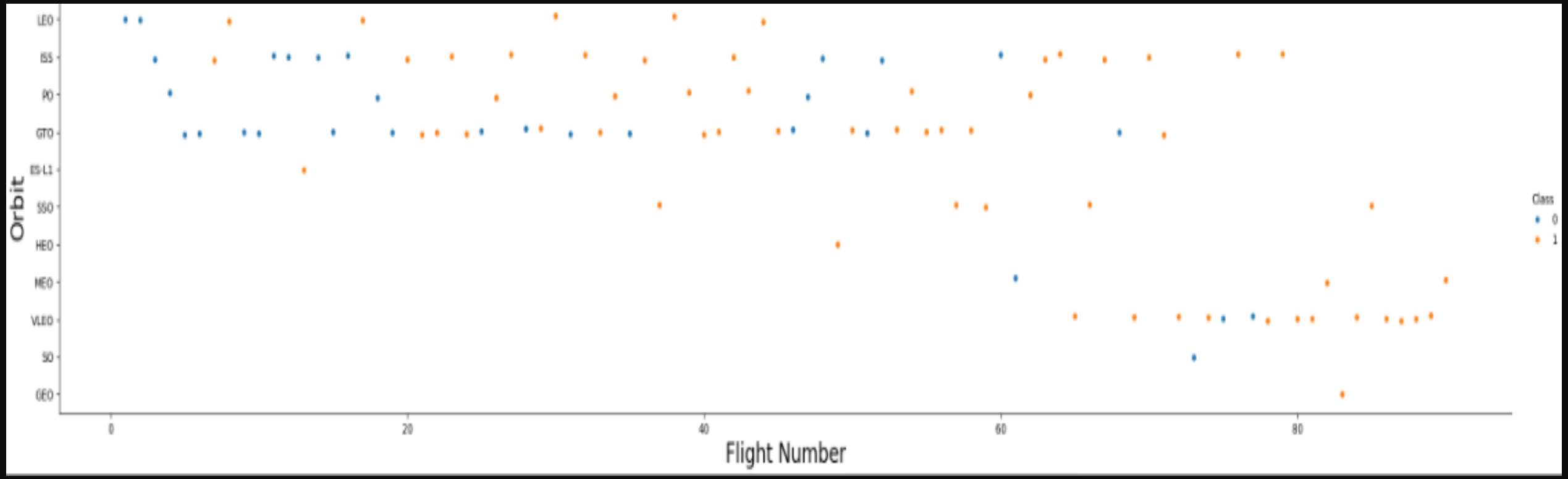
- CCAFS SLC 40 is the only launch site that launched rockets with a pay load mass greater than 10,000 kgs.
- For pay loads less than 5,500 kgs KSC LC 39A had a 100% success rate.
- The higher the pay load mass, the higher the success rate for most of these launch sites.

Success Rate vs Orbit Type



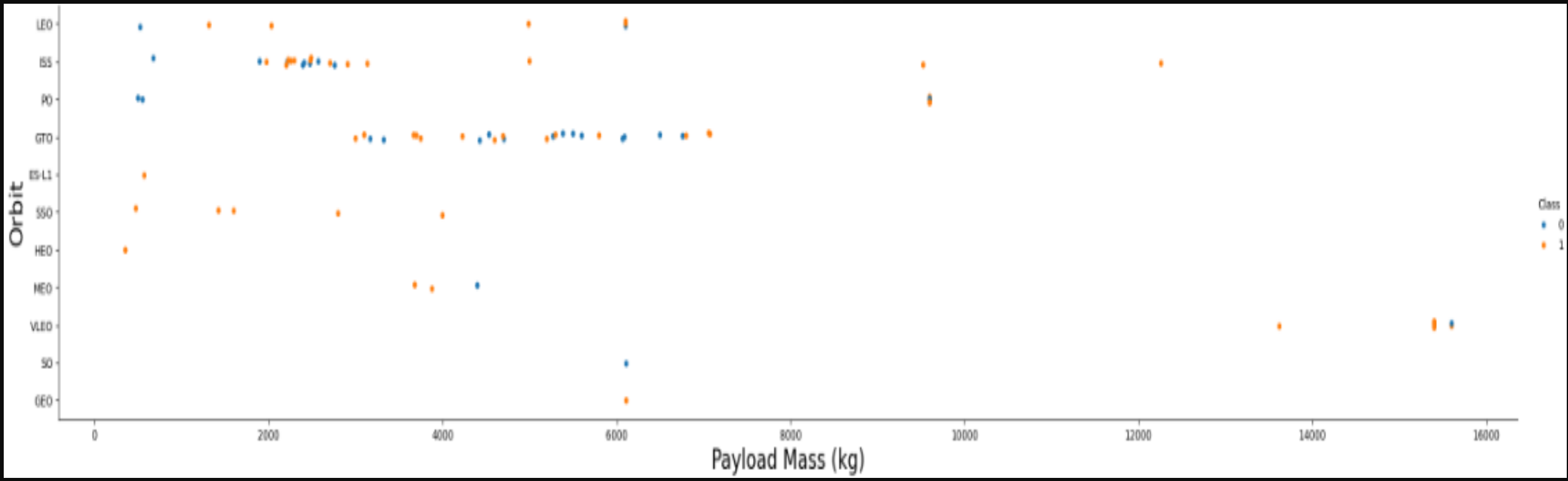
- ES-L1, GEO, HEO and SSO have a 100% success rate.
- GTO, ISS, LEO, MEO, PO and VLEO have a success rate between 50% and 90%.
- SO has 0% success rate.

Flight Number vs Orbit Type



- 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.

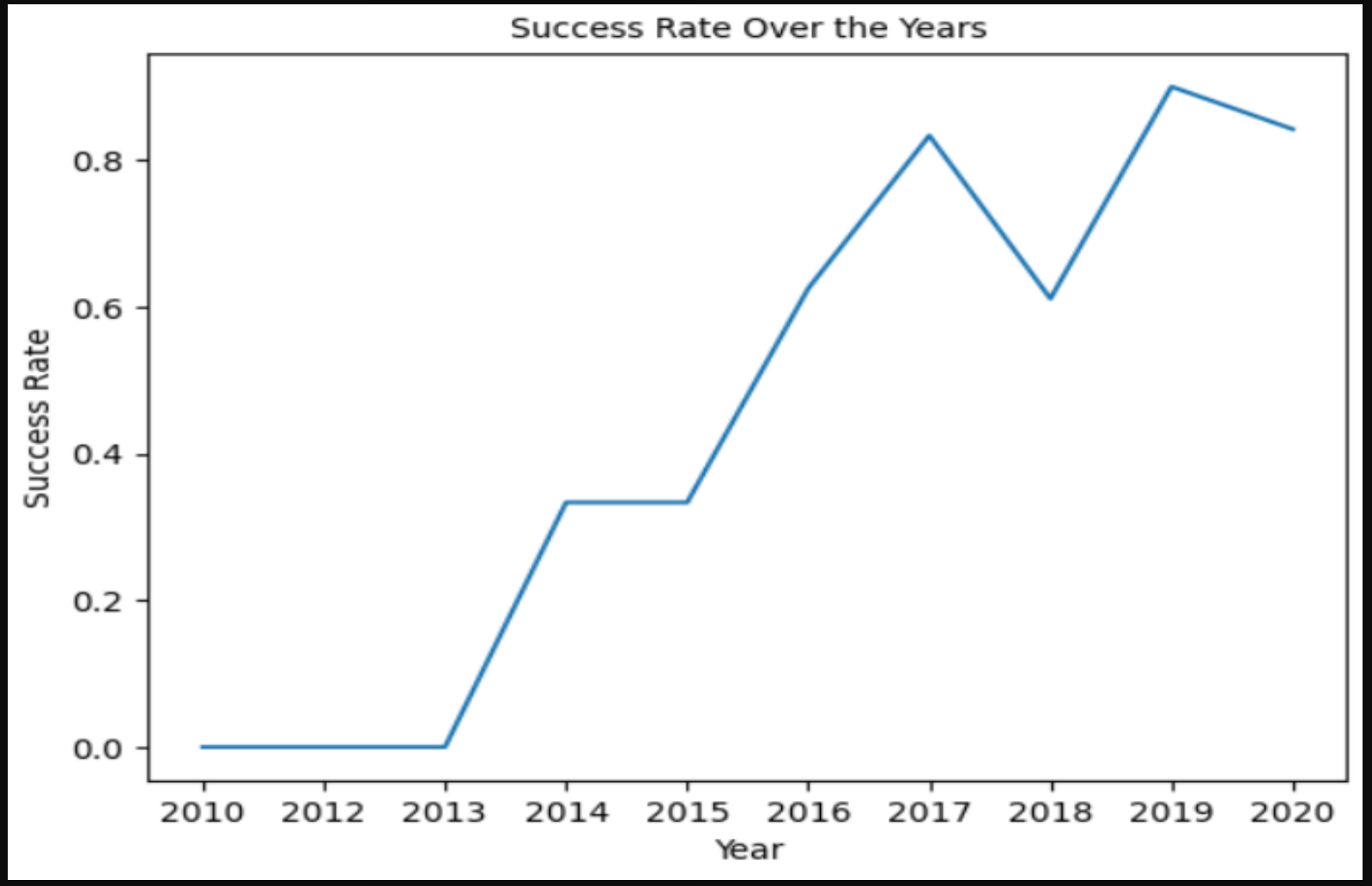
Payload vs Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Rate

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



EDA with SQL

Launch Site Names

```
%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

- Displaying the names of the unique launch sites in the space mission.
- There are four unique launch sites in the dataset.

Launch Site Names beginning with '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

- Showing 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass launched by NASA(CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- Exhibiting the total payload mass carried by boosters launched by NASA (CRS).
- The total payload mass is 45,596 kg.

Average Payload Mass carried by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = 'F9 v1.1'
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- Presenting the average payload mass carried by booster version F9 v1.1.
- The average payload mass is 2,928.4 kg

First Successful Ground Pad Landing Achieved

```
%sql select min(Date) from SPACEXTBL where Mission_Outcome = 'Success'
```

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

min(Date)
2010-06-04

- Listing the date when the first successful landing outcome in ground pad was achieved.
- 04-06-2010 was when it was accomplished.

Successful in Drone Ship and Payload Mass between 4,000 & 6,000 kgs

```
%sql select Booster_Version from SPACEXTBL where Landing_Outcome in ('Success (drone ship)') and PAYLOAD_MASS__KG_
     between 4000 and 6000
```

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

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

- Itemizing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- There are 4 boosters that meet those two criteria.

Mission Outcomes Count

```
%sql select Mission_Outcome, count(*) from SPACEXTBL group by Mission_Outcome
```

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

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

- Enumerating the total number of successful and failure mission outcomes.
- 96% of the mission outcomes are successful.

Boosters that carried maximum Payload Mass

```
%sql select Booster_Version from SPACEXTBL where "PAYLOAD_MASS__KG_" = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

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

- Registering the names of the booster versions which have carried the maximum payload mass utilizing a subquery.
- The result set has 12 booster versions.

2015 Failed Landing Outcomes in Drone Ship

```
%sql select substr(Date, 6, 2) AS month, Landing_Outcome, Booster_Version, Launch_Site
     from SPACEXTBL where substr(Date, 0, 5) = '2015' and Landing_Outcome = 'Failure (drone ship)' and Launch_Site IS NOT NULL;
```

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

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

- Listing the records which will display the month names, failure landing outcomes ,booster versions and launch site in drone ship for the months in year 2015.
- Only 2 records from the dataset meet the criteria specified.

Landing Outcomes between 04-06-2010 & 20-03-2017

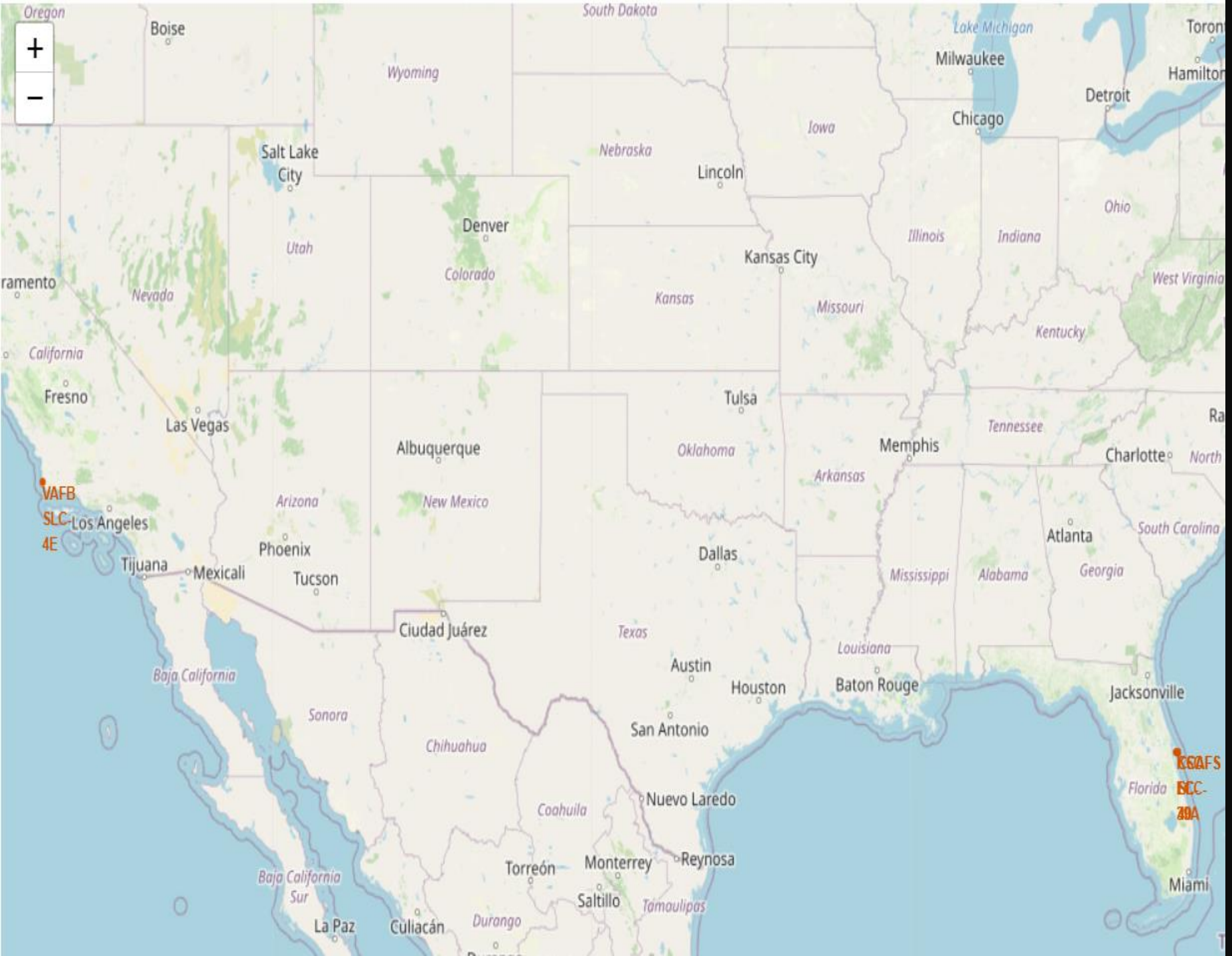
```
%sql select Landing_Outcome, count(*) from SPACEXTBL where Date between '2010-06-04' and '2017-03-20'
     group by Landing_Outcome order by count(*) desc
```

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

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

- 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.
- 32% of the landing outcomes in that time period are No Attempt.

Interactive Map with Folium



Mark Launch Sites on Map

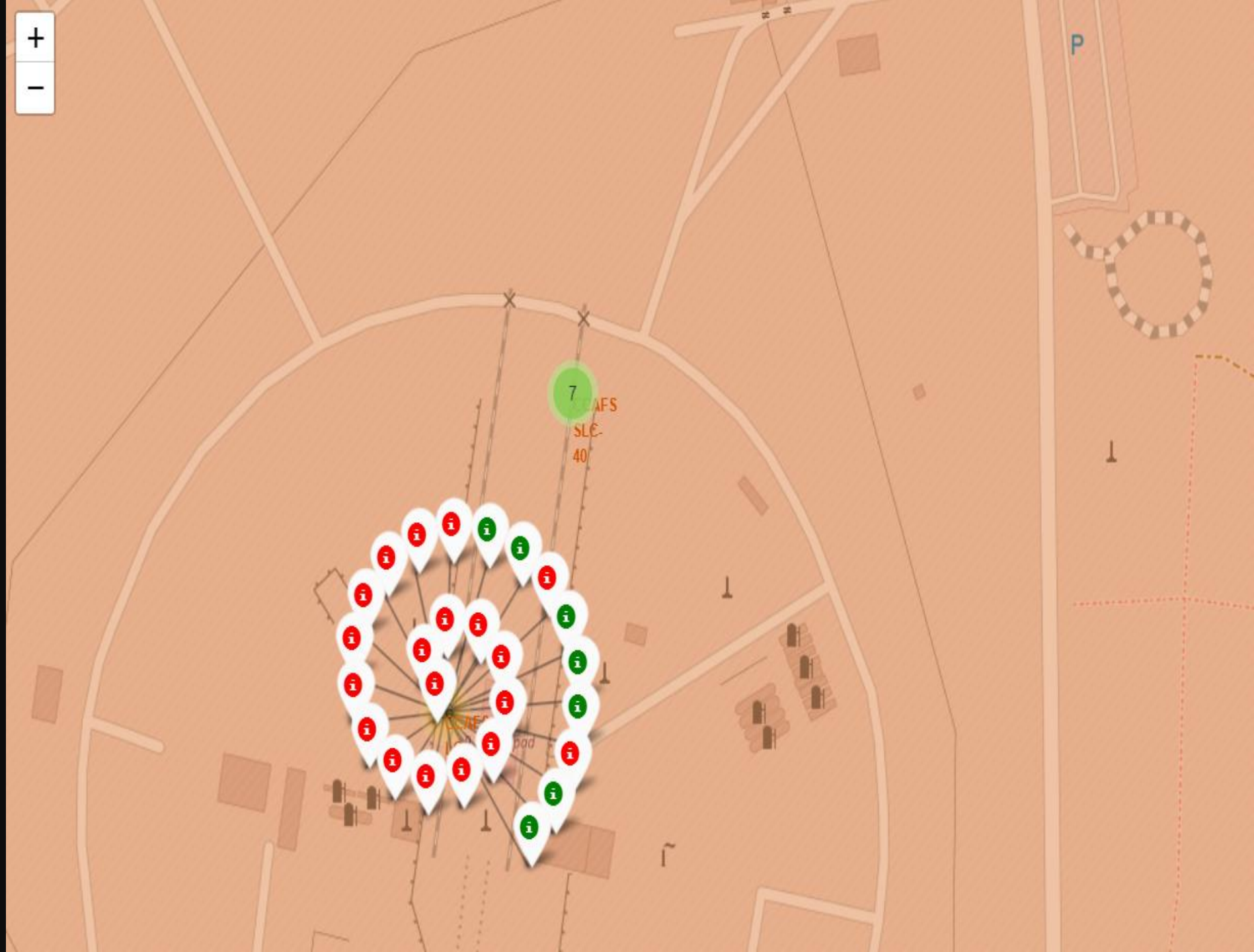
- The majority of the launch sites taken into consideration for this project are close to the Equator.
  - Due to the fact that everything on Earth's surface moves at its fastest potential speed (1670 km/h), launch sites are located as close to the equator as is practical.
- Launch locations taken into consideration for this project are quite near to the coast.
  - When launching rockets toward the ocean, it reduces the possibility of debris falling or exploding in close proximity to people.

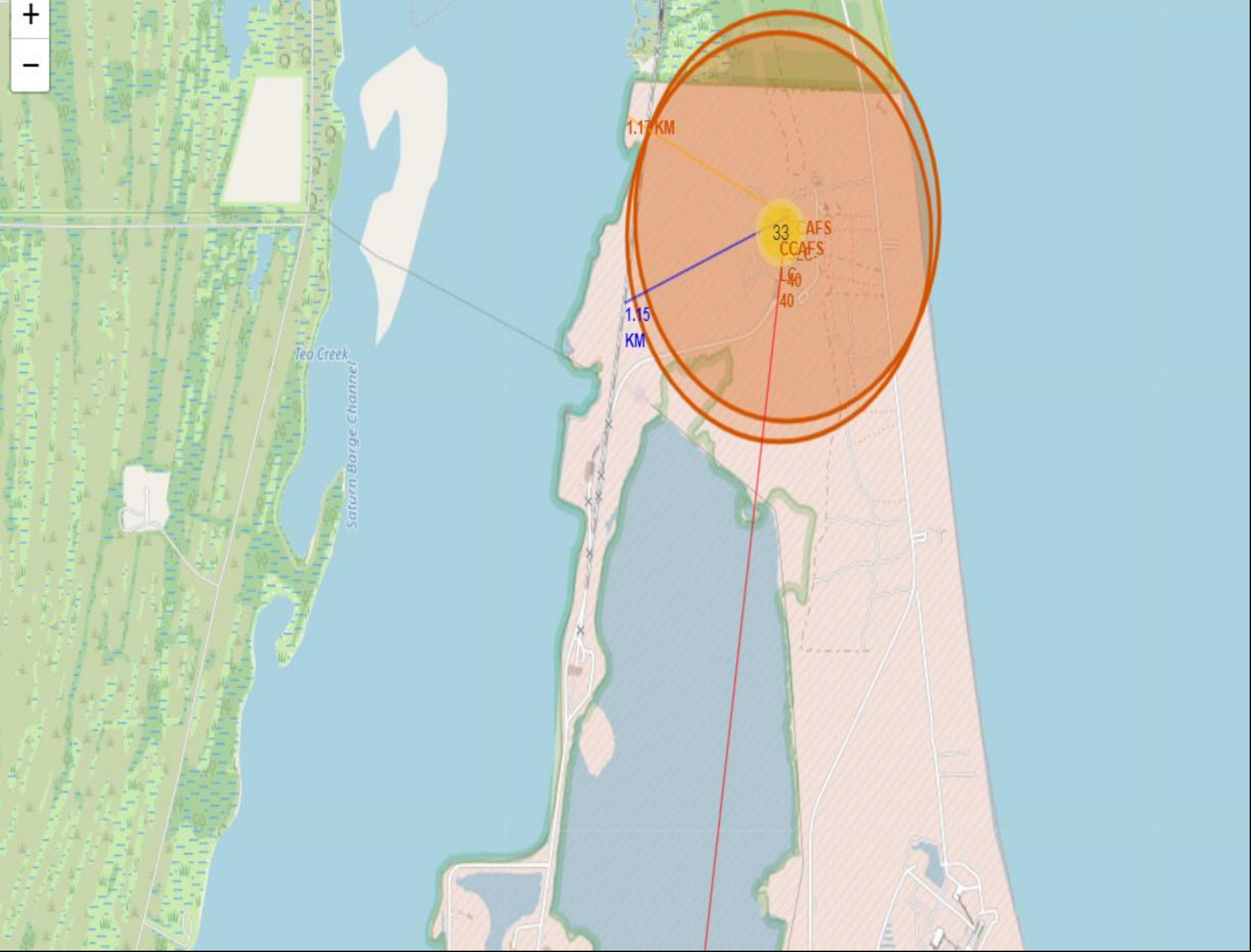
Mark Launch Outcomes for each Site on Map

- Utilizing color labeled markers it is easier to identify which launch sites have relatively high success rates.

  Green: Successful Launch
  Red: Failed Launch

- Launch site CCAFS LC-40 has a very low success rate as indicated on the map.





Distance between Launch Site CCAFS SLC-40 and its Proximities

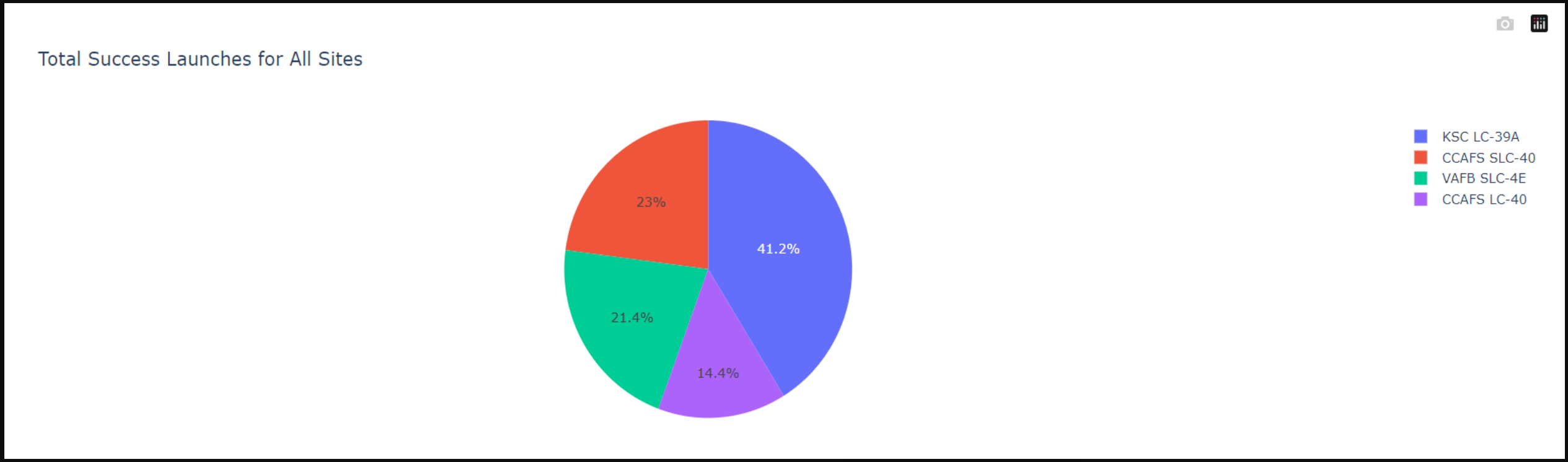
- The visual analysis of the launch site CCAFS SLC-40 shows that it is on some form of island, a distance from the mainland.

- On the island we can clearly see that:
  - distance to road, 1.15km
  - distance to coastline, 1.17km
  - distance to city, 18.21km

- A failed rocket can travel up to 15-20 km in a matter of seconds due to its high speed. It might pose a threat to densely inhabited areas.

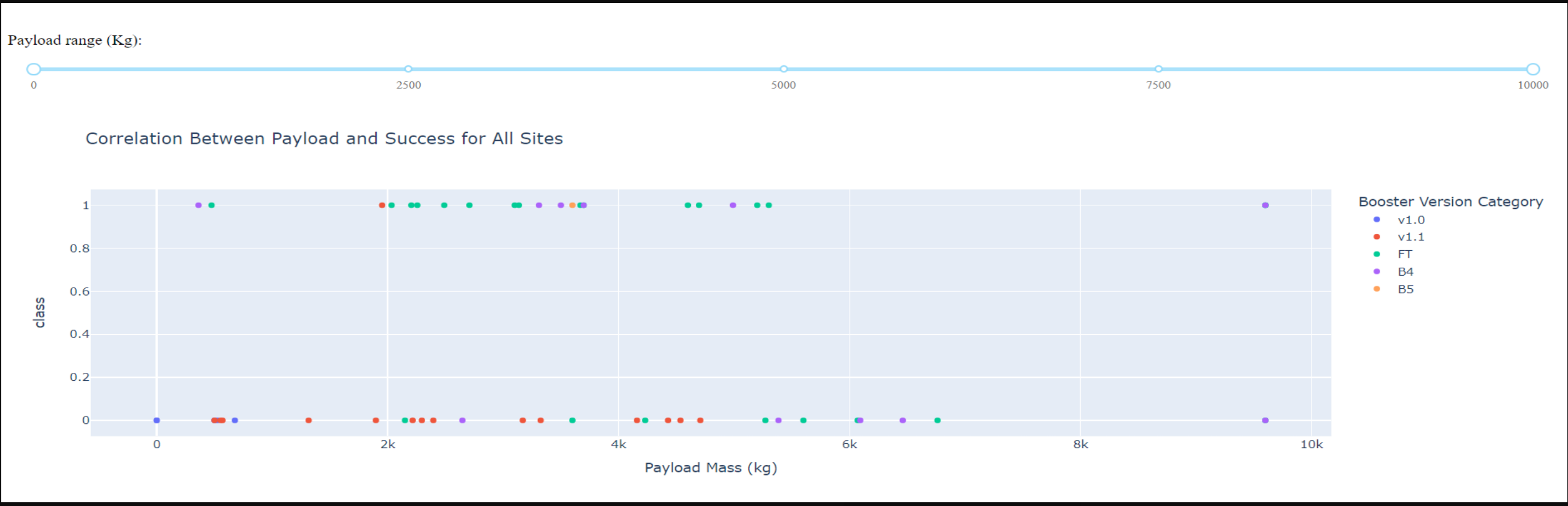
Plotly Dash Dashboard

Successful Launches by Site



- KSC LC-39A takes up a big chunk of the pie with 41.2% successful launches.
- CCAFS SLC-40, VAFB SLC-4E and CCAFS LC-40 follow in the same order.

Payload Mass vs Launch Outcome

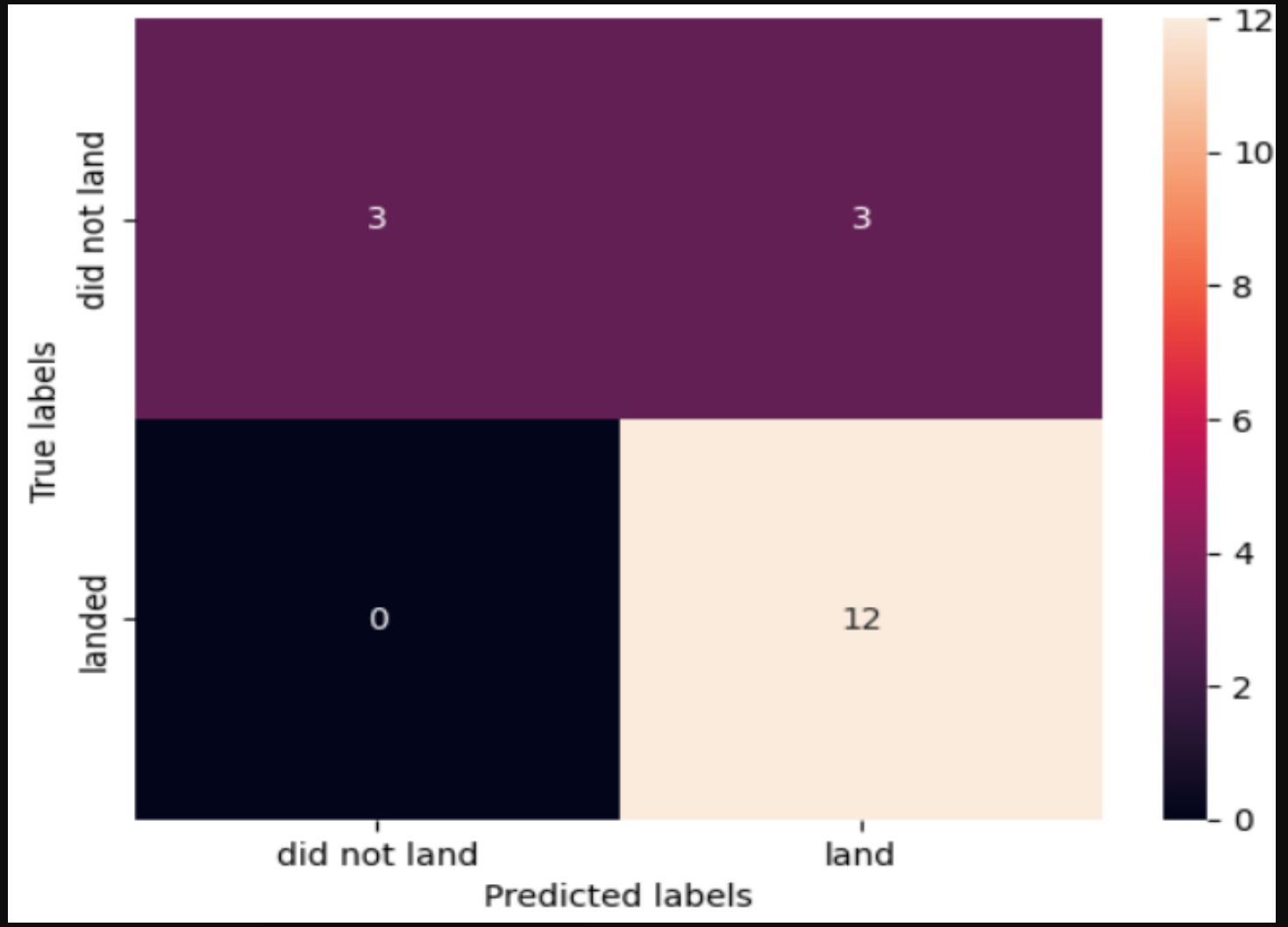


- Launches with a Payload Mass between 2,000 and 5,500 kg show the highest success rate.
- The Booster Version FT has the most successful launches compared to the rest.

Predictive Analysis
Classification

Confusion Matrix

- This is a confusion matrix for Support Vector Machine model.
- Clearly, it is able to distinguish between the different classes (the landing outcome).
- The major problem are the false positives.
- MM



Accuracy Metrics

Test Dataset

Model	Accuracy	Jaccard Index	F1 Score	Log Loss
Logistic Regression	0.833333	0.700000	0.814815	6.007276
Support Vector Machine	0.833333	0.700000	0.814815	0.000000
Decision Tree	0.666667	0.511111	0.666667	0.000000
K-Nearest Neighbors	0.833333	0.700000	0.814815	0.000000

Entire Dataset

Model	Accuracy	Jaccard Index	F1 Score	Log Loss
Logistic Regression	0.866667	0.755556	0.856061	4.80582
Support Vector Machine	0.877778	0.774491	0.869190	0.00000
Decision Tree	0.866667	0.767677	0.866667	0.00000
K-Nearest Neighbors	0.855556	0.739845	0.845407	0.00000

- From the test dataset all the models have the same scores except for the Decision Tree.
  This may be due to the small test sample size, hence why we tested the models on the entire dataset
- The Support Vector Machine model has the scores and accuracy, tested against the entire dataset, compared to the rest.

CONCLUSION

- Support Vector Machine is the best algorithm for this dataset.
- Launches with a Payload Mass between 2,000 and 5,500 kg show the highest success rate.
- Locations, launch sites, taken into consideration for this project are quite near to the coast.
- Launch sites VAFB SLC 4E and KSC LC 39A have much higher success rates.
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate.
- The success rate of launches has continued to increase over the years.