Winning Space Race with Data Science

Amin Aghakasiri
September 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This project utilizes SpaceX's Data API and Falcon 9's Wikipedia page to collect data on successful and failed first stage launches and the variables that played a part in that outcome. Using tools such as SQL, Data Visualization and Machine Learning the following was learned

- The variable with the highest impact on successful landing was Payload Mass, with most successful landing happening with a payload mass between 2000 to 6000 kg. The most successful landing site was KSC LC-39A with a 76.9% success rate. Lastly our machine learning algorithms outputted an R2 score of 0.83 and F1 score of about 0.89 displaying their successful ability to predict future landing outcomes.

Introduction

- SpaceX has a strong grip on space exploration and travel due to its sharp decrease in costs compared to its competitors. This decrease is a result of their ability to reuse the First Stage.

- This project assumes a new client wishes to replicate SpaceX's success. In order to do so it predicts the success rate of the first stage landing in an attempt to help determine the cost of a launch.

Section 1

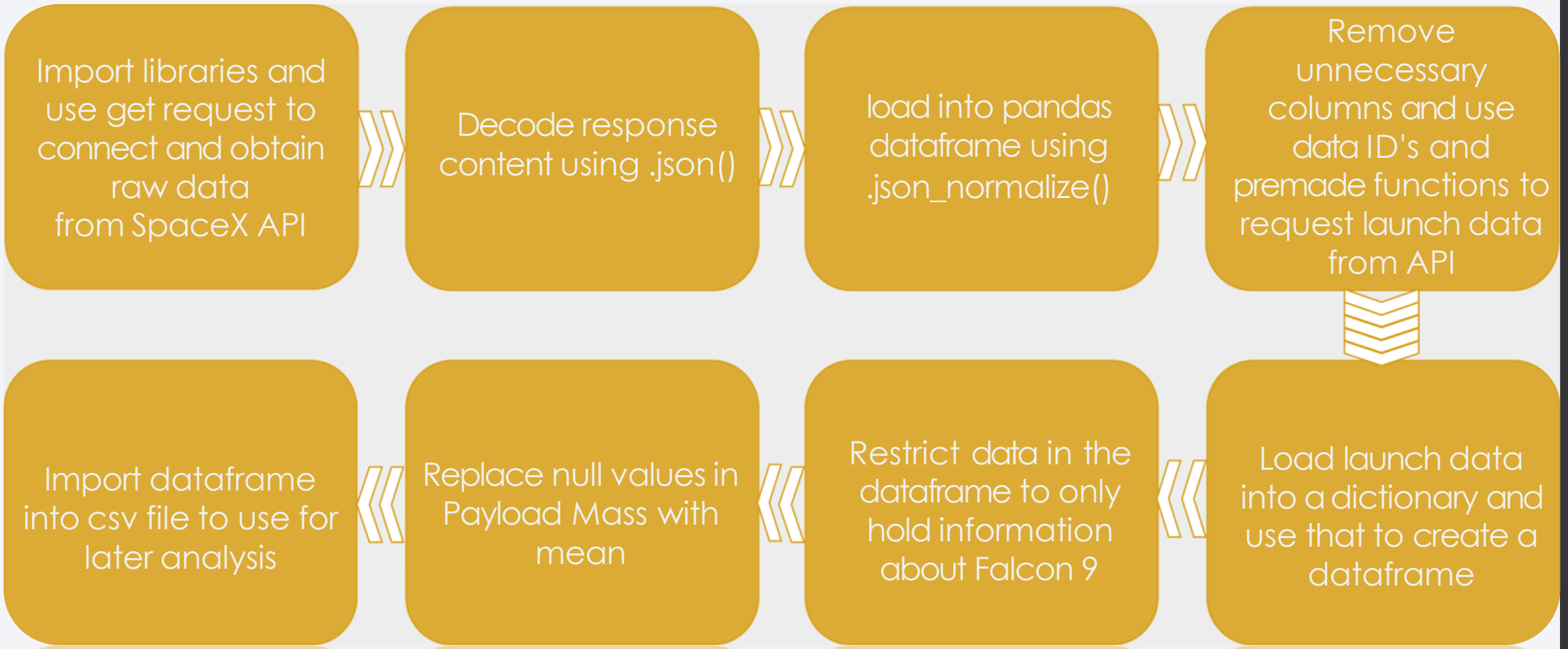
Methodology

Methodology

Executive Summary

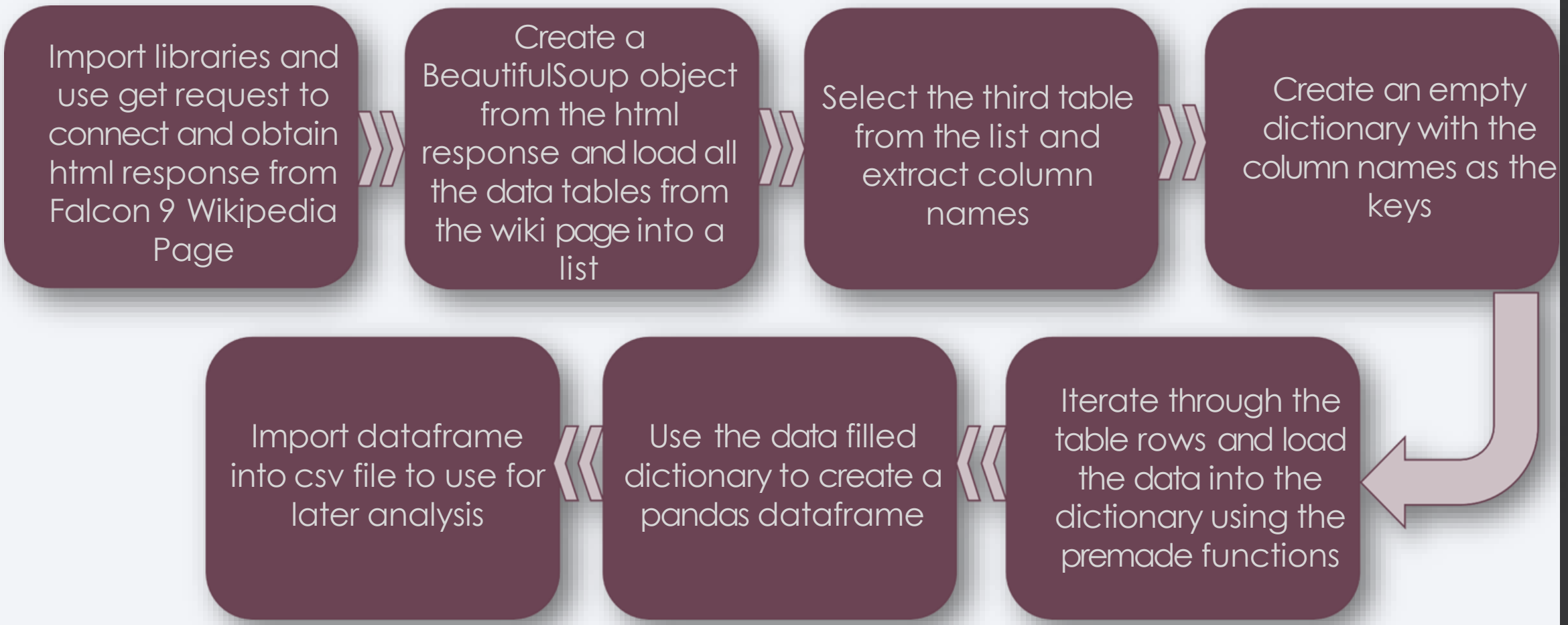
- Data collection methodology:
  - Data was collected through Falcon 9's Wikipedia page and SpaceX's data API
  - Perform data wrangling
  - Data was processed to create a new variable based on landing outcomes that depicted a 1 in the case the landing outcome was successful and 0 when it failed
  - Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The following models were used: Logistic Regression, SVM, Decision Tree and KNN. Each Model had gone through a hyperparameter optimization process to select its most efficient parameters

Data Collection-SpaceX API



GitHub Link: Data Collection API

Data Collection - Scraping



GitHub Link: Data Collection Web Scraping

Data Wrangling

Data Formatting

Look for null values, and incorrect data types within each column of the data set

Using .value_counts() calculate the number and occurrence of each Launch Site, Orbit, and Landing Outcome

Create a new column called landing class within the data set that outputs a 1 if the landing outcome was successful and a 0 if it failed

Calculate the mean of this new column to find the success rate of all the missions

After ensuring the data is ready further analysis save it so a csv file

In order for further analysis to result in accurate insights, data must be formatted in a way that does not skew or misinterpret information.

In this scenario the column "landing class" was formatted into a binary class that would later allow Machine Learning Algorithms to understand it.

GitHub Link: Data Wrangling

EDA with Data Visualization

1. Bar plots: Great at comparing categorical values with a numerical value.

Bar Plots:
- Orbit Type vs Success Rate

2. Scatter plots: Excel at comparing two independent variables with each other and can show possible correlation

Scatter Plots:
- Payload Mass vs Orbit Type

3. Line plots: Help show trends in data a period of time

Line Plots:
- Date vs Success Rate

GitHub Link: Data Visualization

EDA with SQL

SQL Queries:

- Load data set into pandas data frame then used dataframe.to_sql() to create SQL Table
- Displayed Distinct Launch Sites
- Displayed 5 Launch Sites whose name begin with CCA
- Displayed Total Payload Mass carried by boosters launched by NASA(CRS)
- Displayed Average Payload Mass carried by booster version F9 v1.1
- Displayed the date in which the first successful ground pad landing occurred
- Displayed the names of the boosters which successful land in drone ships and have a payload mass greater than 4000 kg but less than 6000
- Displayed the total number of Successful and Failure mission Outcomes
- Using a Subquery displayed the names of all booster versions which have carried the maximum payload mass
- Displayed the month of failed landing outcomes in drone ships, booster versions, and launch site in the year 2015
- Displayed the different types of landing outcomes between 2010-06-04 and 2017-03- 20 and the number of times they occurred in descending order

GitHub Link: EDA with SQL

Build an Interactive Map with Folium

Main Map

- Centered at Nasa Johnson Space Center (JSC)

Circles

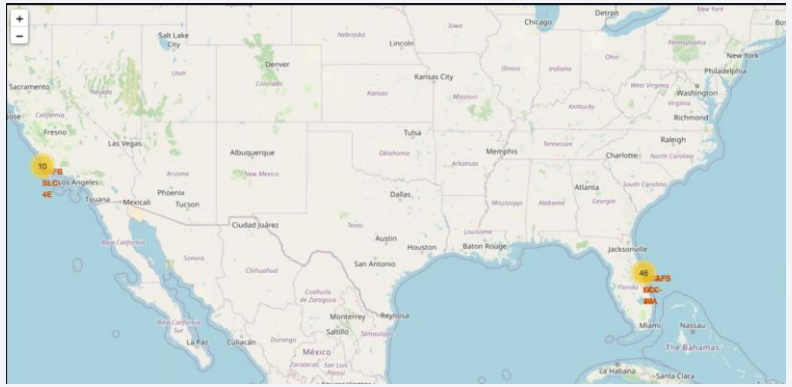
- Red Circle positioned on NASA headquarters that shows displays a popup with its name when clicked
- 4 different colored circles positioned on each rocket launch site that display their names as a popup when clicked

Clusters

- Marker clusters positioned on each rocked launch site that display the total number of landings
- When zoomed in and clicked on a spiral of clusters is displayed with a green marker for each successful landing and a red marker for each failed landing

Mouse Position

- An add-on that displays the current latitude and longitude of your cursor as it hovers over the map



GitHub Link: SpaceX Folium Map

Build a Dashboard with Plotly Dash

Dashboard Outline

Title: SpaceX Launch Records Dashboard

- Title is bolded and centered

Equipped with a Dropdown menu

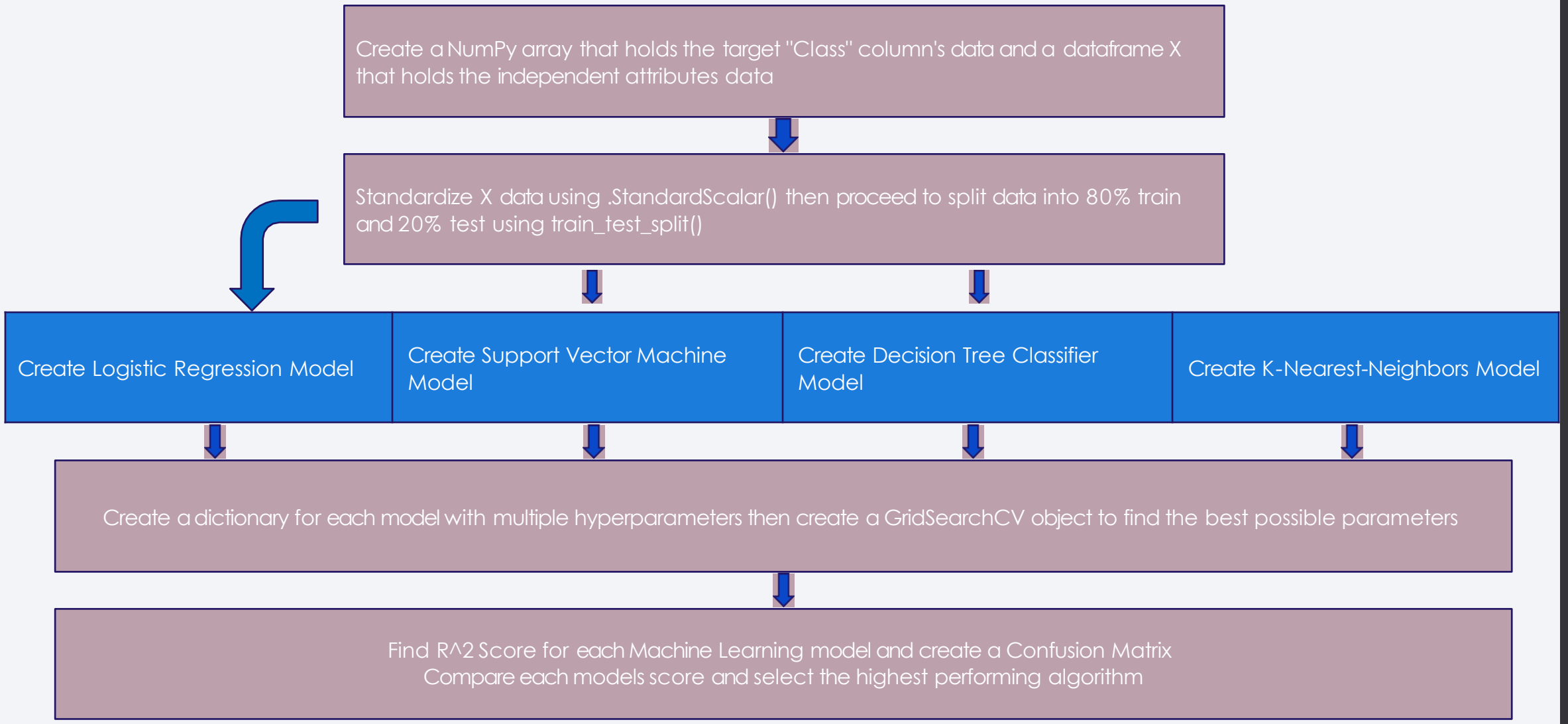
- Allows you to choose All Launch Site statistics or an individual Launch Site's statistics

Dashboard Visualizations

- Pie Charts
  - The All Sites dropdown shows a pie chart that compares the number of successful landings between all the launch sites
  - Each individual site dropdown shows a pie chart that
    - compares is landing success rate
- Scatter Plot and Range Slider
  - A scatter plot is shown that compares the Payload Mass (x-axis) and whether or not the launch successfully landed (y-axis)
    - The points are colored based on Booster Version
      - Category
  - The slider allows you to control the Payload Mass range

GitHub Link: Interactive Dashboard Analysis

Predictive Analysis (Classification)

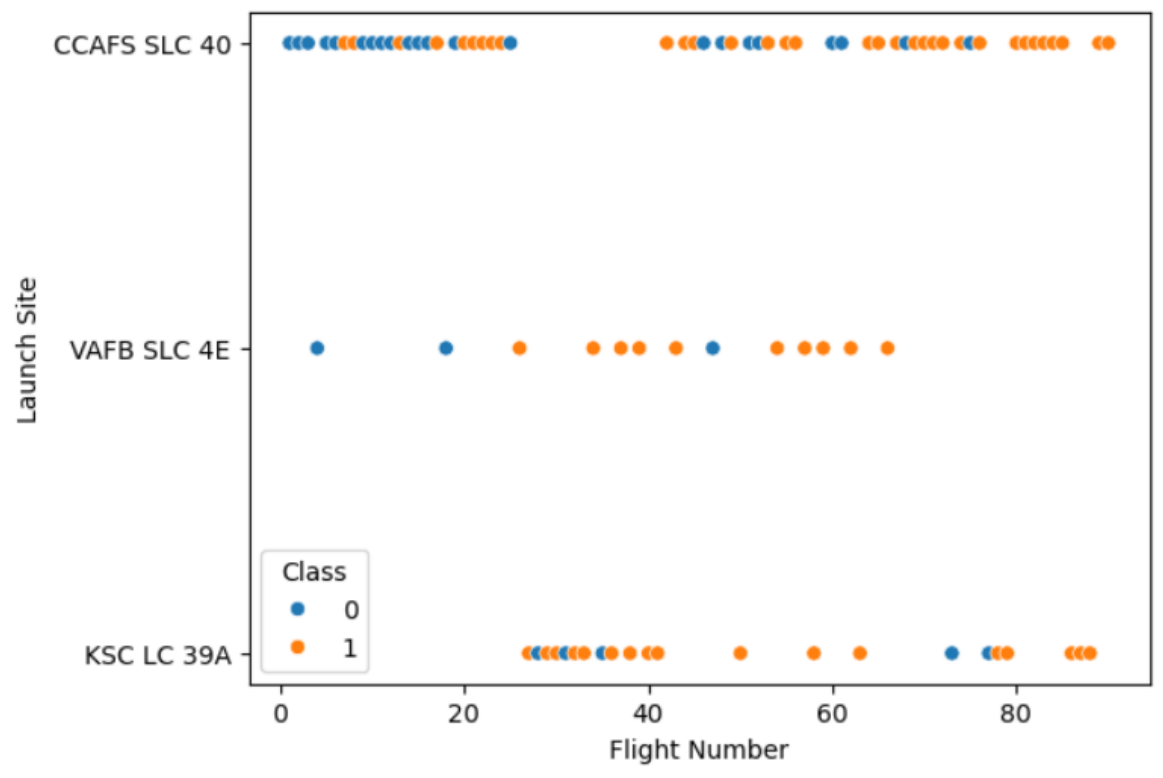


GitHub Link: Machine Learning Prediction Algorithms

Section 2

Insights drawn from EDA

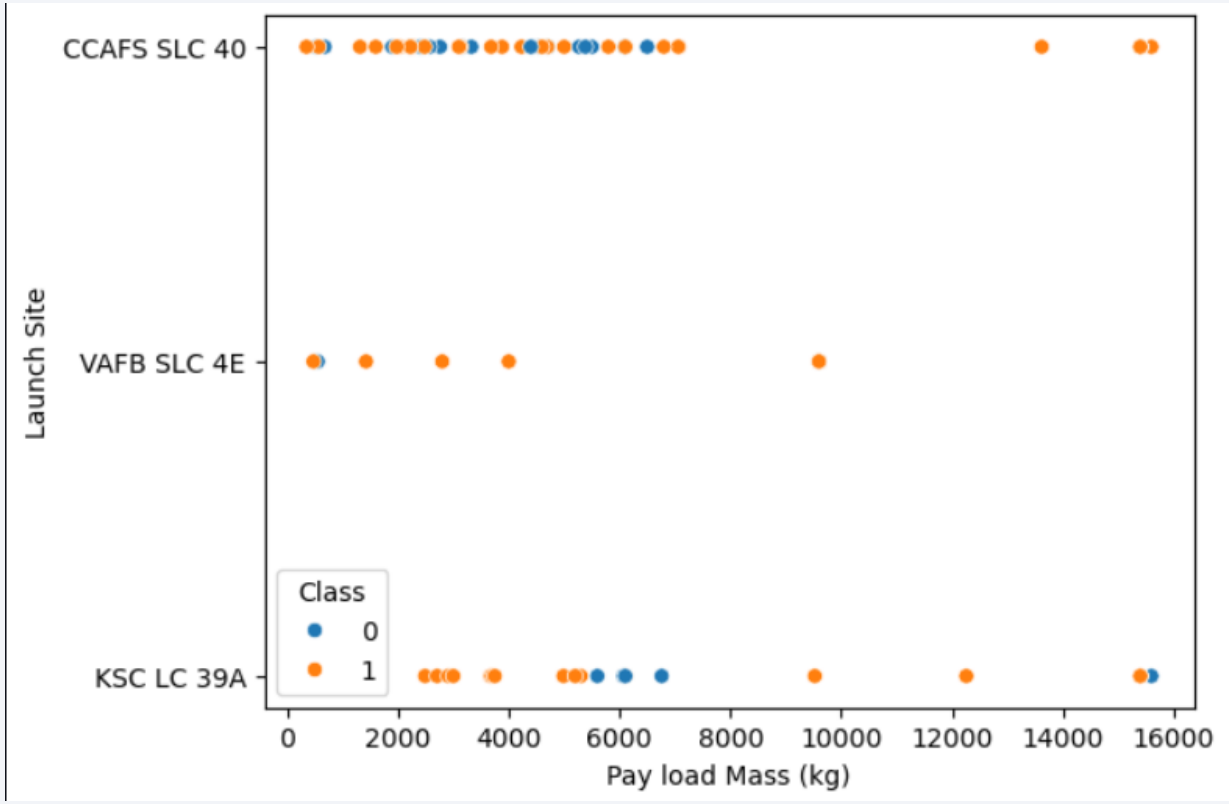
Flight Number vs. Launch Site



Insight:

- CCAFS SLC 40 has the highest amount of launches
- Success rate increases as flight number increases

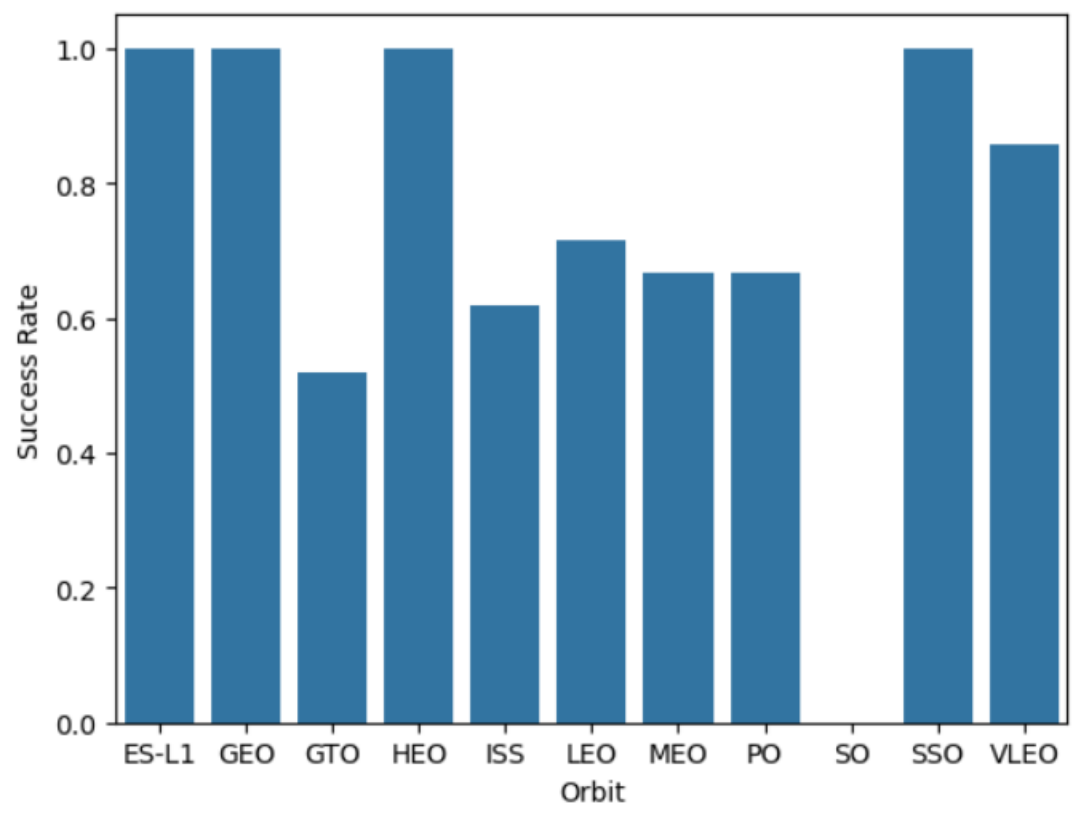
Payload vs. Launch Site



Insight:

- VAFB SLC 4E has no launches for rockets with a payload mass greater than 10000
- Majority of launches with a high payload mass had a higher success rate than those with low payload mass

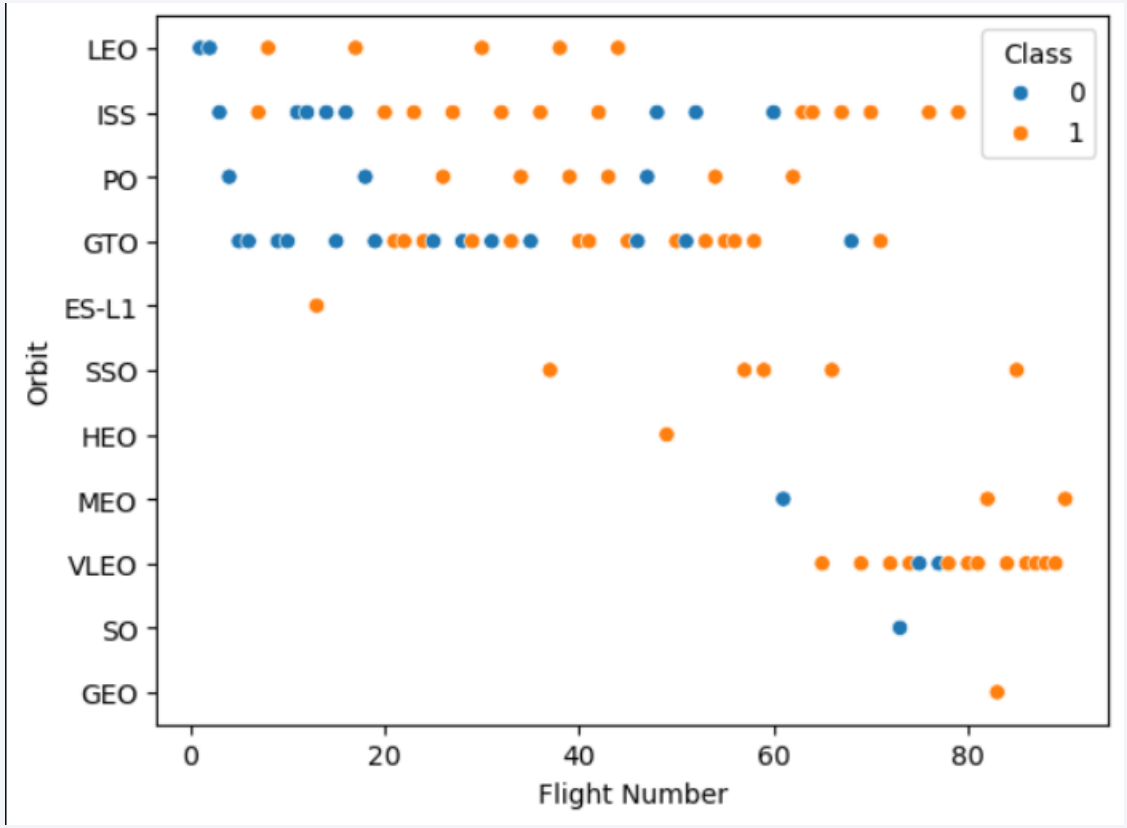
Success Rate vs. Orbit Type



Insights:

- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Orbit VLEO has the second highest success rate
- Orbit SO has a 0% success rate
- Orbit GTO has the second lowest success rate

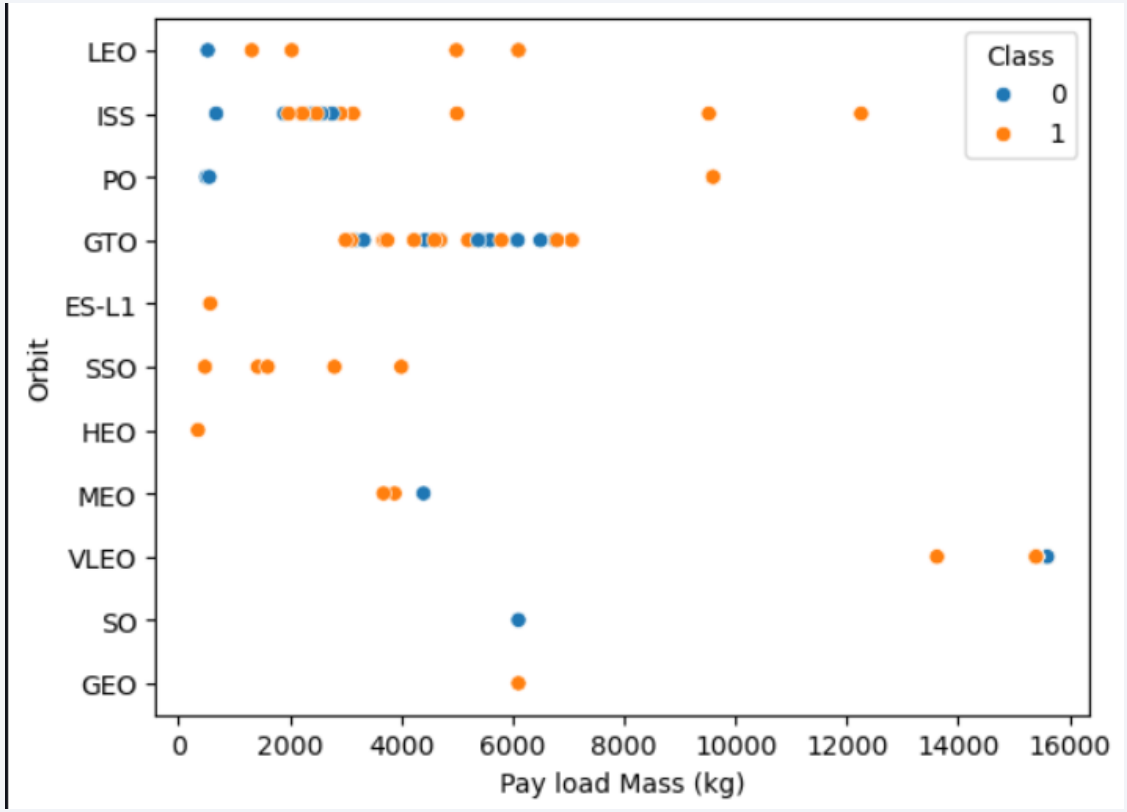
Flight Number vs. Orbit Type



Insights:

- Orbit VLEO only has launches with a flight number above 60
- Orbits ES-L1, GEO, and HEO only have a single launch
  - All of which are successful
- Orbit LEO has a 100% success rate after its first two failures
- Orbits ISS, GTO, and VLEO have the most launches
- Orbit GTO's success rate has no relationship with flight number

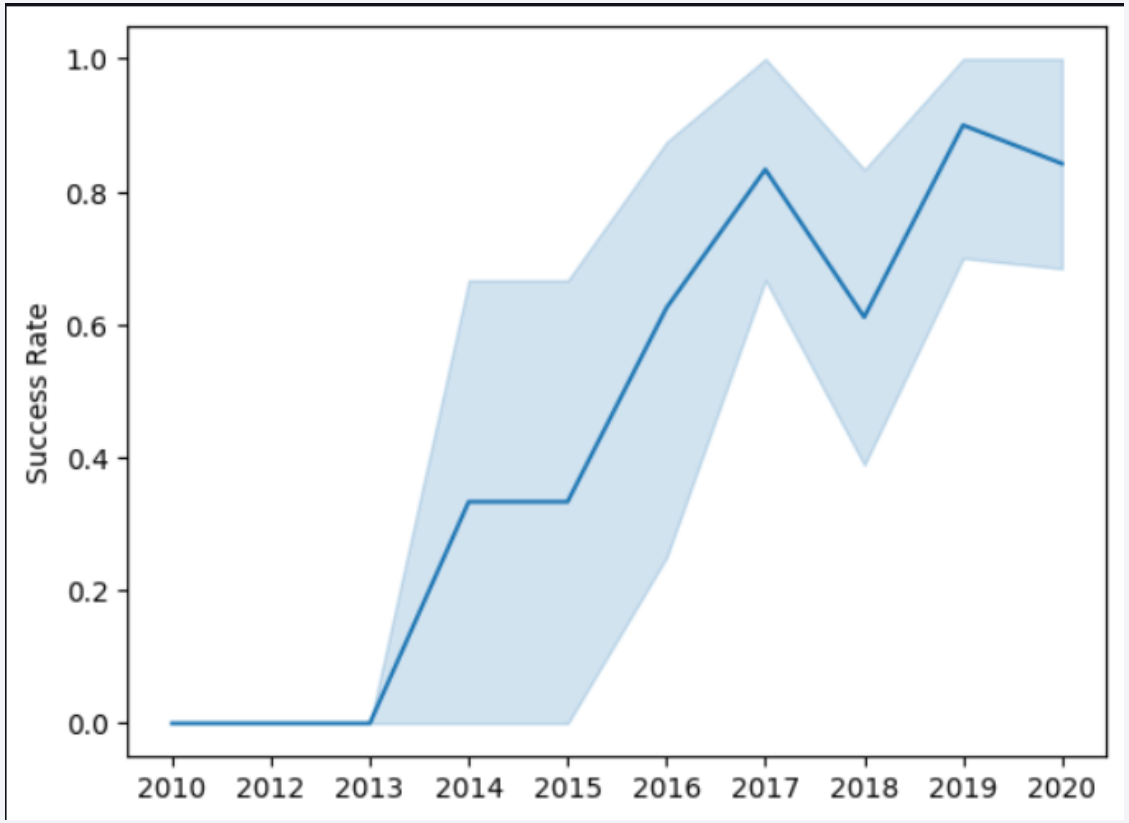
Payload vs. Orbit Type



Insight:

- Orbit GTO only has launches with a payload mass between 2000 kg and 8000 kg
- Orbit BLEO only has launches with payload mass above 13000 kg
- Orbits PO, LEO, and ISS have a strong positive relationship with payload mass

Launch Success Yearly Trend



Insights:

- Success rate had periods of sharp increases between 2013-2014, 2015–2017, and 2018-2019
- Success rate had periods of stagnation between 2010–2013 and 2014-2015
- Success rate had periods of decline between 2017–2018, and 2019–2020
- Overall success rate has increased drastically since 2013

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTABLE where 'Time (UTC)' = (select min('Time (UTC)') from SPACEXTABLE)
```

Python

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

min(date)
2010-06-04

Total Number of Successful and Failure Mission Outcomes

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

Python

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

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

Boosters Carried Maximum Payload

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

Python

```
 * 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

2015 Launch Records

```
%sql select distinct strftime('%m',Date) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where strftime('%Y',Date) = '2015' a
```

Python

```
 * 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

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

```
%sql select count(Landing_Outcome) from SPACEXTABLE where date between '2010-06-04' and '2017-03-20'
```

Python

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

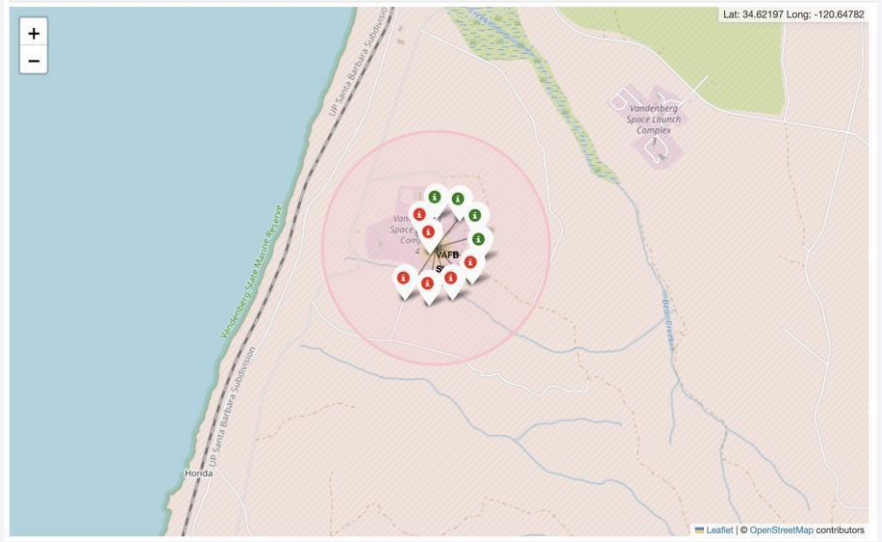
count(Landing_Outcome)
31

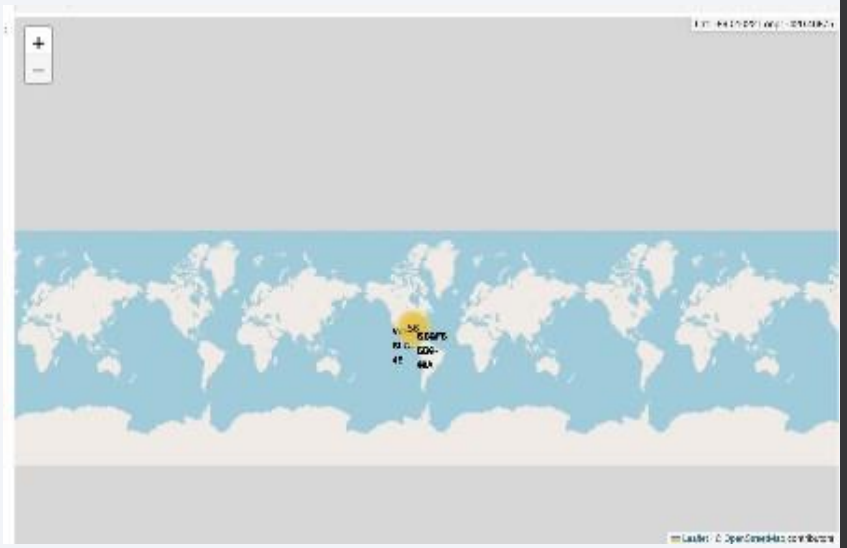
Section 3

Launch Sites Proximities Analysis

Folium Map - Launch Sites

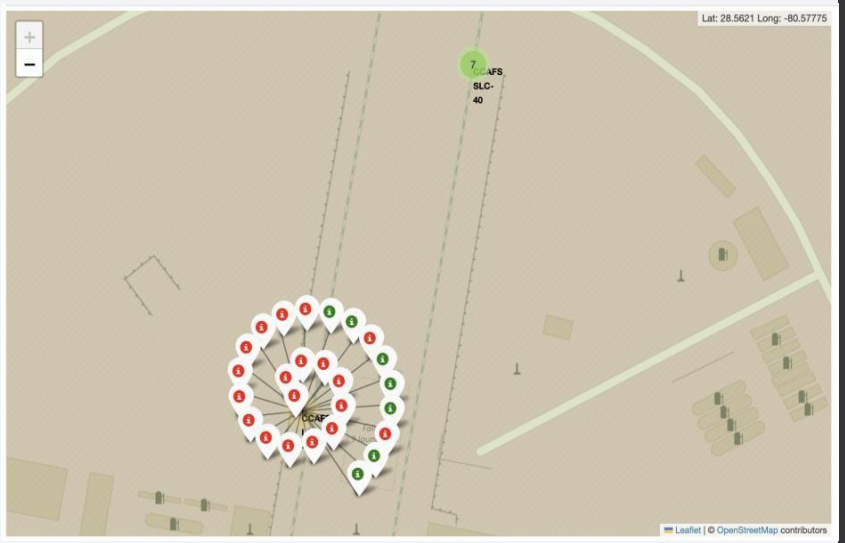
Folium Map – Launch Outcome

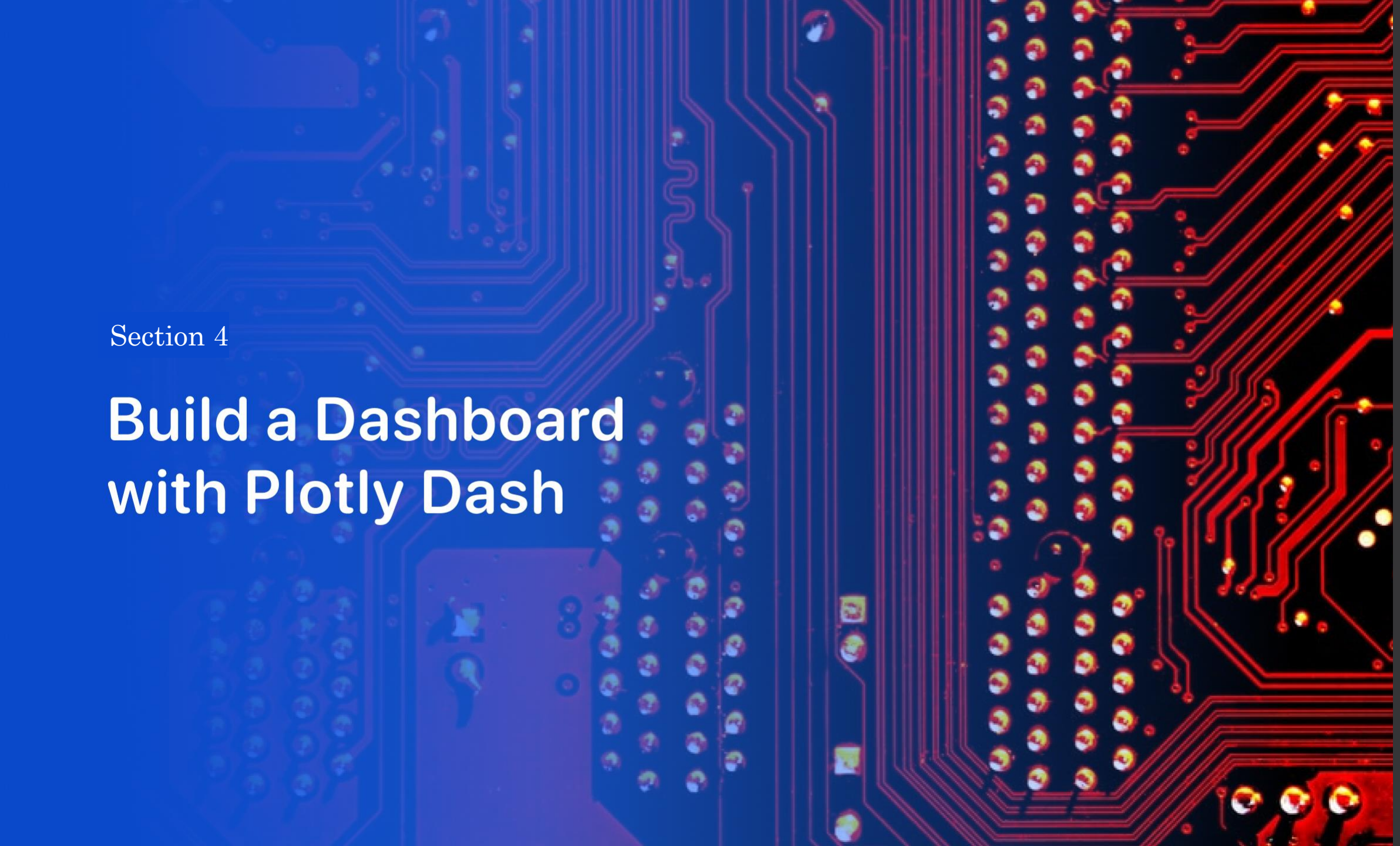




An interactive map displaying the total number of successful and failed launch outcomes.

Successful launches are green while failed launches are red

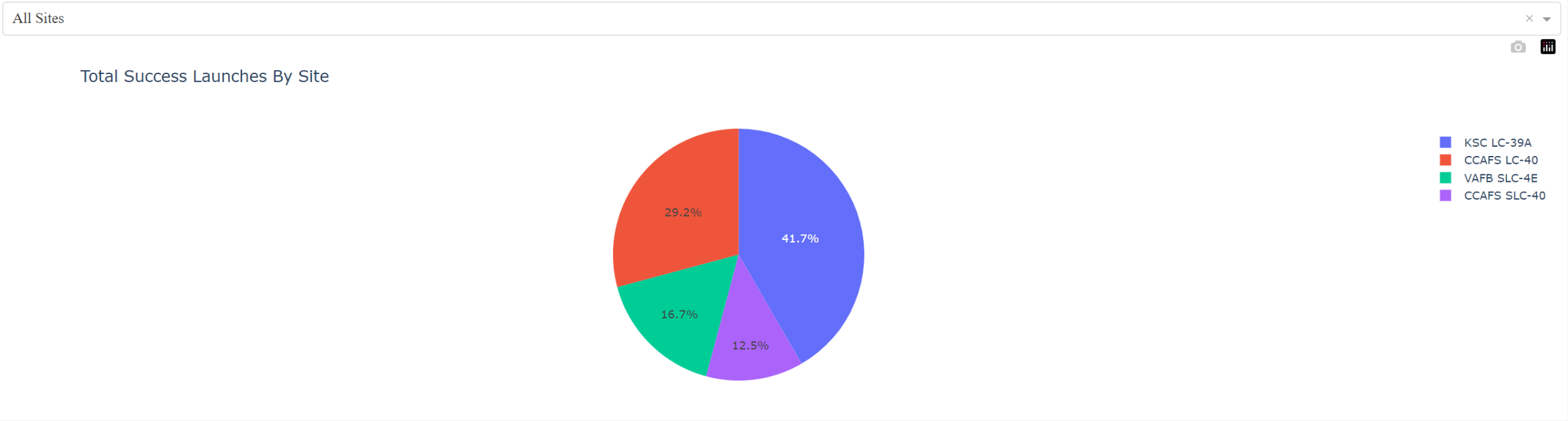




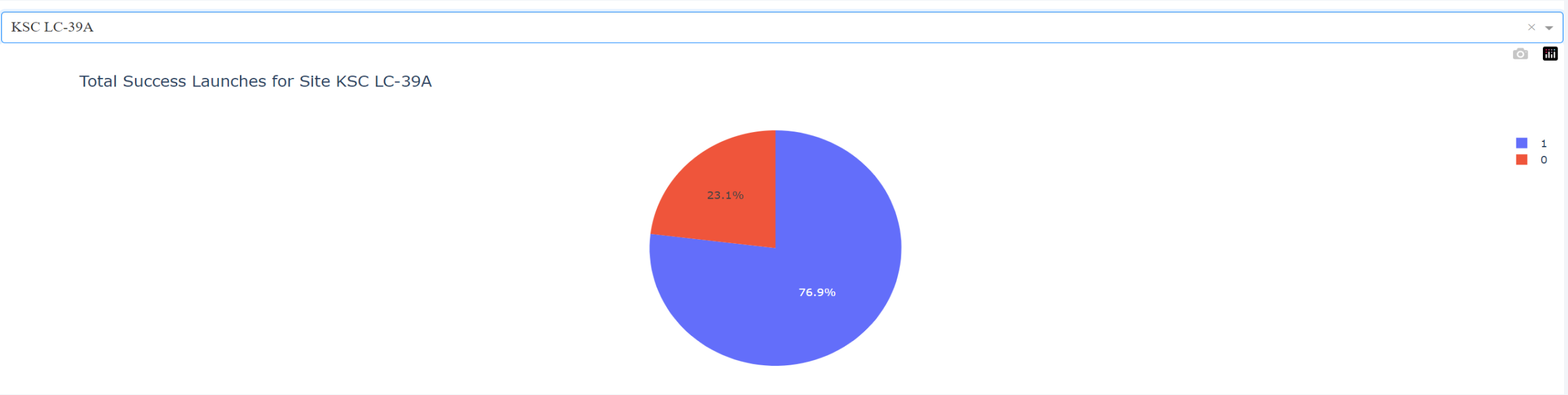
Section 4

Build a Dashboard with Plotly Dash

Dash – Launch Site Success Rates



Dash – Highest Launch Site Success Rate



Dash – Payload Mass vs Success Rate Scatter Plot

A Scatter Plot that helps show the correlation between payload mass and success rate

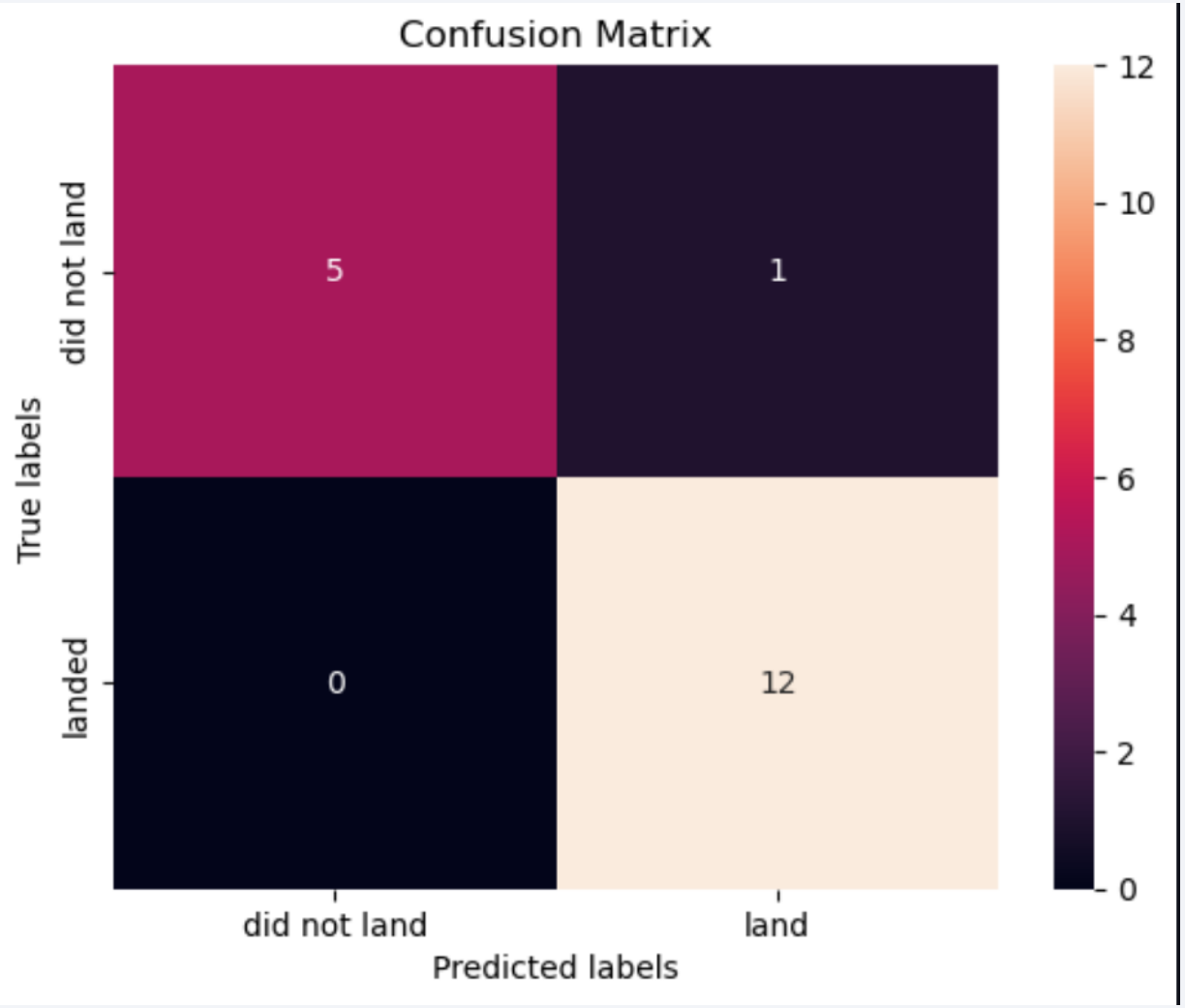
- Booster Version Category types are separated by color
- The range slider above the graph allows you to customize the range of the x-axis

Most successful launches seem to be with a payload mass between 2000 to 6000 kg

Section 5

Predictive Analysis (Classification)

Confusion Matrix



A confusion matrix displays 4 different quadrants of information

- Quadrant 1: False Positive
  - Incorrectly predicts rocket successfully landed
- Quadrant 2: True Negative
  - Correctly predicts rocket failed to land
- Quadrant 3: False Negative
  - Incorrectly predicts rocket failed to land
- Quadrant 4: True Positive
  - Correctly predicts rocket successfully landed

Due to the small sample size, all ml algorithms have almost the same accuracy around 84 percent.

```
0.8333333333333334
```

Conclusions

- The machine learning algorithms have an F1 score of 0.89 which shows it's skill in accurately predicting both successful and failed launches
- Launches with a payload mass between 2000 and 6000 kg show the highest rate of success
- Launch Site KSC LC 39-A has the highest success rate among the other launch sites
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
  - Orbits ISS, GTO, and VLEO have the most launches
- Launch Sites are located in close proximity to coastlines

Thank you!