Winning Space Race with Data Science

Matias Ferraro
22/01/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection via API.
  - Data Collection with Web Scraping.
  - Data Wrangling.
  - Exploratory Data Analysis with Data Visualization.
  - Exploratory Data Analysis with SQL.
  - Visual analytics with maps using Folium and Dashboard creation with Plotly Dash.
  - Predictive analysis using Machine Learning techniques.
- Summary of all results
  - EDA results.
  - Visual results including maps and dashboards.
  - Predictive Results.

Introduction

The aim of this project is to predict the success of the Falcon 9 first stage landing in order to determine the cost of a rocket launch by SpaceX. Falcon 9 rocket launch cost is 62 million dollars, and other providers cost upward of 165 million dollars each. This information can be used by other companies to bid against SpaceX for a rocket launch. The cost savings for SpaceX is largely due to the reuse of the first stage, making the prediction of a successful landing crucial for determining the cost of a launch. So, this information will be interesting for another company if it wants to compete against SpaceX.

Therefore, all of this leads us to the most important question, which is trying to understand if the Falcon 9 rocket will land successfully with certain features and what factor determine if the rocket will land successfully or not.

Section 1

Methodology

Methodology

Executive Summary

The data was collected in two different ways. First, the SpaceX API was used to obtain the first set of information, and then web scraping methods were used to obtain a dataframe formed from information that was stored on Wikipedia. Finally, this data was combined to work with all the data obtained during this processes.

Data wrangling methods including cleaning and transformation were performed next. This process involved techniques such as one-hot encoding for categorical features and dropping unnecessary columns, among other things. Once everything was cleaned it up, exploratory data analysis using visualization and SQL was performed.

Then, Folium and Plotly Dash were used to perform interactive visual analytics and, after that, work was done on classification models to perform predictive analysis. Specifically, four different machine learning classification models were used, with the goal of selecting the most accurate model

Data Collection

Datasets were collected from the SpaceX API and Wikipedia using web scraping techniques.

First, the SpaceX API was called using the Rest API URL and a JSON was returned, which was used to create a dataframe. This dataframe was then cleaned and exported.

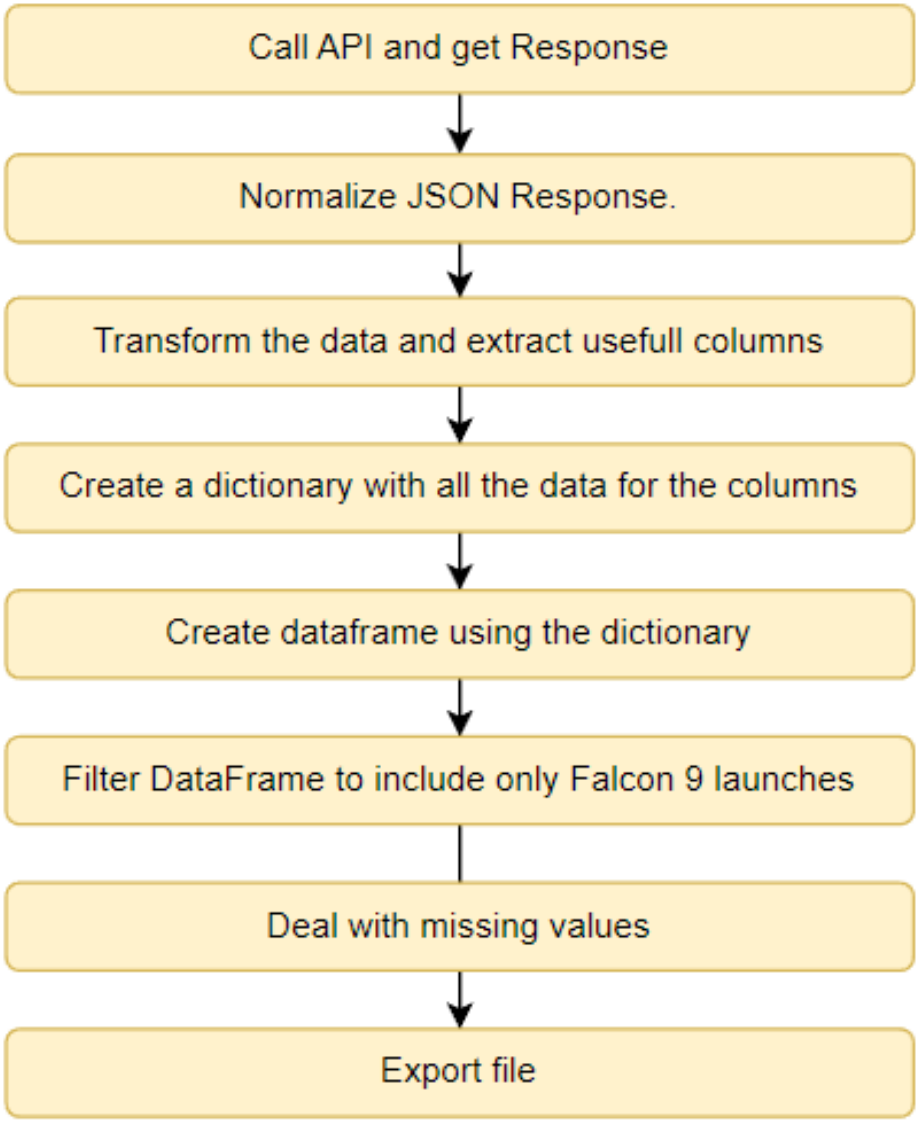
On the other hand, a Wikipedia URL containing data on all SpaceX rocket launches was used and data was extracted using BeautifulSoup. This data was then used to create a dataframe, which was also exported

Data Collection – SpaceX API

- Data was collected using the Rest API, then data was cleaned to get a DataFrame and export it to a file.

- GitHub URL of the completed SpaceX API calls notebook:

  https://github.com/MatNF/IBM_Capstone-Project/blob/master/1%20-%20Collecting%20the%20data%20with%20API.ipynb



Data Collection - Scraping

- Data obtained using Beautiful Soup for web scraping, finding tables and creating dictionary with column variables to finally create a DF to export as csv file.

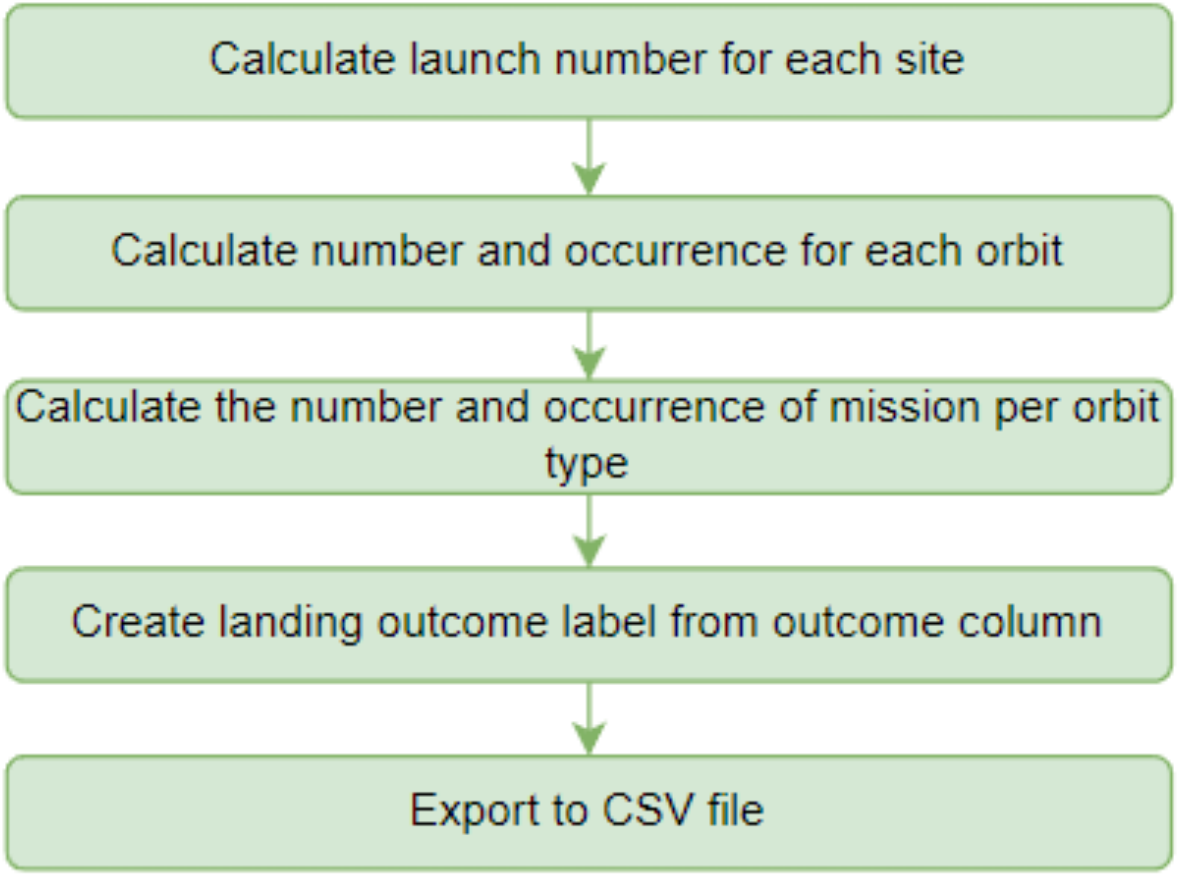
- GitHub URL of the completed web scraping notebook:

https://github.com/MatNF/IBM_Capstone-Project/blob/master/2%20-%20Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

- The number of launches for each site was calculated, the number of occurrences for each orbit was calculated, and the number of occurrences of mission outcome per orbit type. On the other hand, since there are several different cases in the dataset where the booster does not land correctly and given that these data are categorical, the categories were transformed using one-hot encoding to obtain a new column with values of 1 (indicating success) and 0 (indicating failure). Finally, the file was exported in CSV format
- GitHub URL:

  https://github.com/MatNF/IBM_Capstone-Project/blob/master/3%20-%20Data%20Wranling%20-%20EDA.ipynb



EDA with Data Visualization

Different charts were used to analyze the data correctly.

- **Charts**:
  - **Scatter**:
    - Flight Number Vs. Payload Mass. - Flight Number Vs. Launch Site - Payload Vs. Launch Site. - Orbit Vs. Flight Number. - Payload Vs. Orbit Type. - Orbit Vs. Payload Mass.
  - **Bar**:
    - Success Rate Vs. Orbit Types.
  - **Line**:
    - Success Rate Vs. Year.

GitHub:

https://github.com/MatNF/IBM_Capstone-Project/blob/master/5%20-%20EDA%20with%20visualization.ipynb

EDA with SQL

- SQL Queries performed were the following:
  - 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.
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass.
  - List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub: https://github.com/MatNF/IBM_Capstone-Project/blob/master/4%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

The Folium map was populated with different map objects such as markers, circles, and lines to mark the success or failure of launches for each site. Circles were used to mark NASA Johnson Space Center's coordinates, as well as to map each launch site's coordinate and show the launch site's name. Markers were used to show successful or unsuccessful landings, using green for successful and red for unsuccessful. Additionally, markers were used to show the distance between launch site key locations such as railways, highways, coasts, and cities by plotting lines between them.

With all these objects, we were able to better understand the problem and have a spatial view of the situation we were dealing with, managing to observe the distances and real sites where each launch was carried out.Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

GitHub: https://github.com/MatNF/IBM_Capstone-Project/blob/master/6%20-%20Visual%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

An interactive dashboard with Plotly Dash was created and two charts were added:

- A Pie chart that shows the total success and total failure for the launch site selected in a dropdown.
- Scatter plot graph that shows the relationship between Success and Payload Mass.

Also, two more components were added:

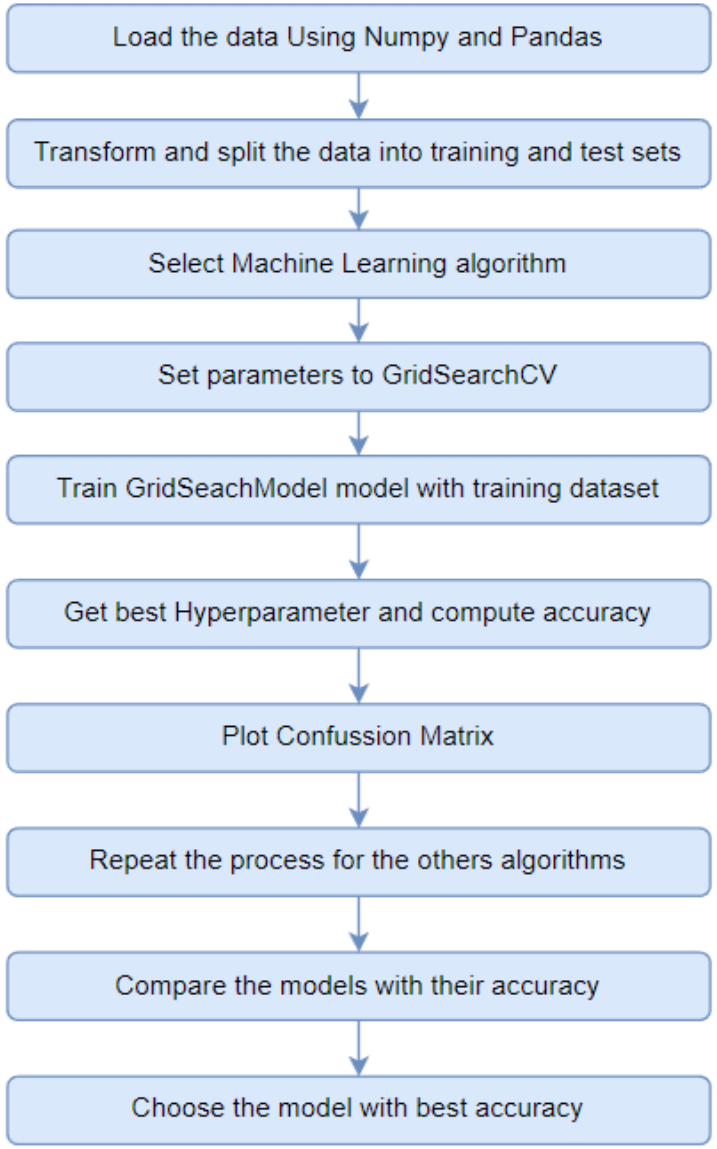
- A dropdown that allows the user to select a launch site.
- A rangeslider allowing the user to select a payload mass in a fixed range.

GitHub: https://github.com/MatNF/IBM_Capstone-Project/blob/master/7%20-%20spacex%20dash%20app.py

Predictive Analysis (Classification)

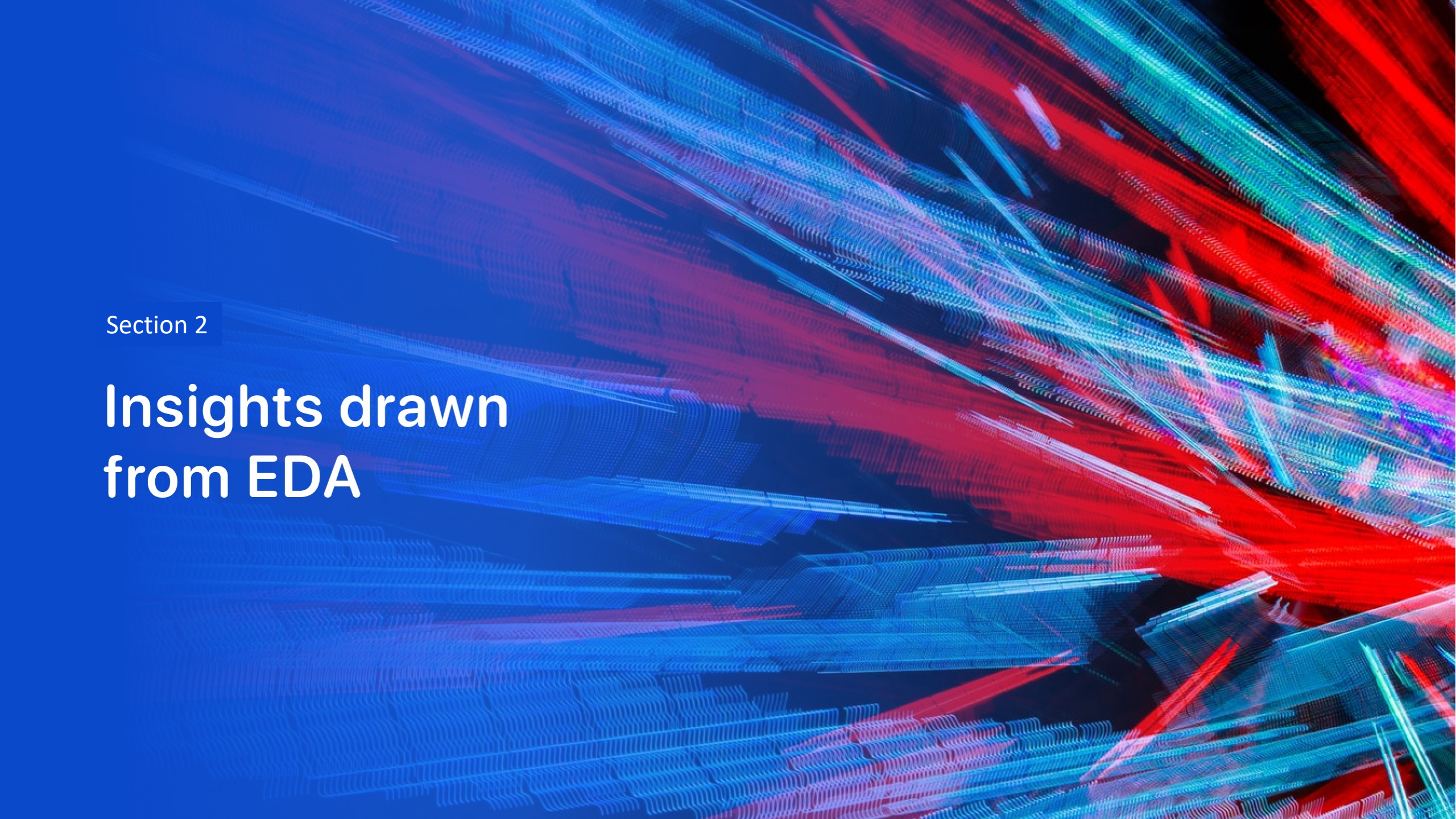
In summary, the data was loaded, a split was made to prepare the training and test set. Different machine learning models were used including logistic regression, support vector machines, k-nearest neighbors, and decision trees (all of this using GreadSearchCV). Finally, the accuracy was calculated for each model and the one with the best accuracy was chosen.

GitHub: https://github.com/MatNF/IBM_Capstone-Project/blob/master/8%20-%20Machine%20Learning%20Prediction.ipynb

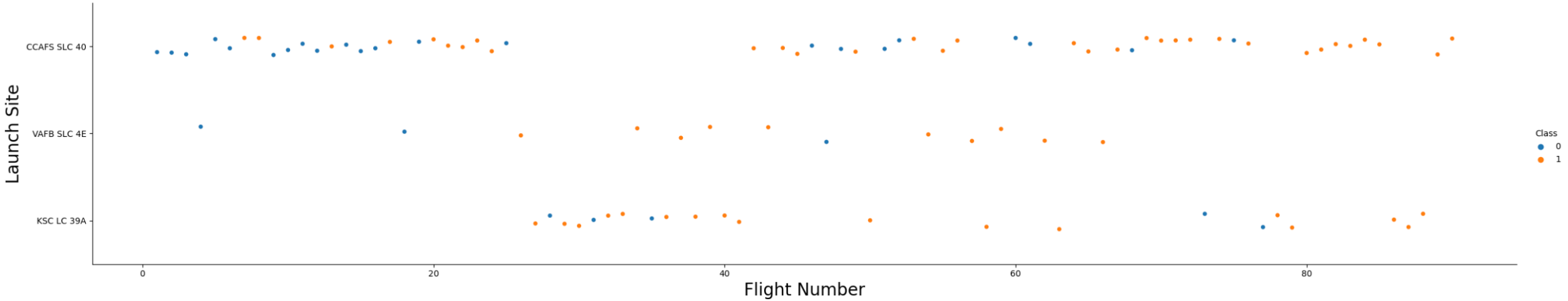


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

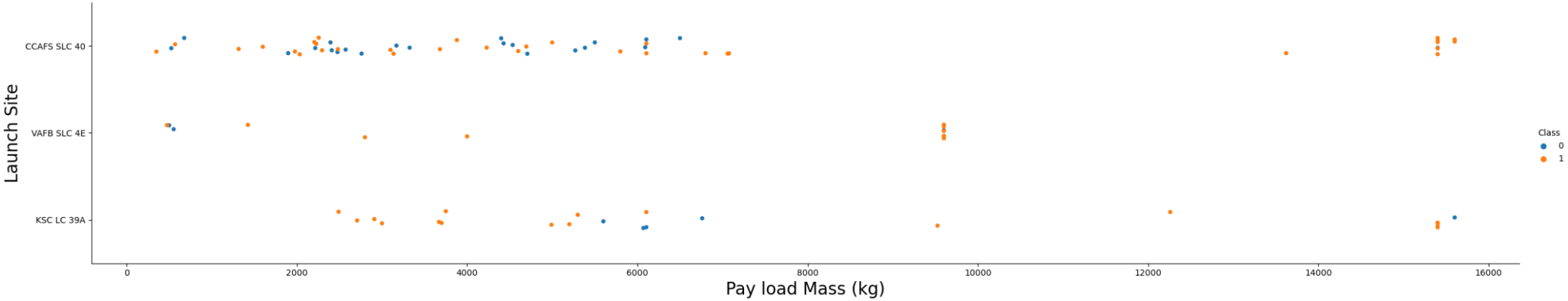

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



Early flights were unsuccessful while recent flights have been successful. The more flights launched from a site, the higher the success rate at that site.

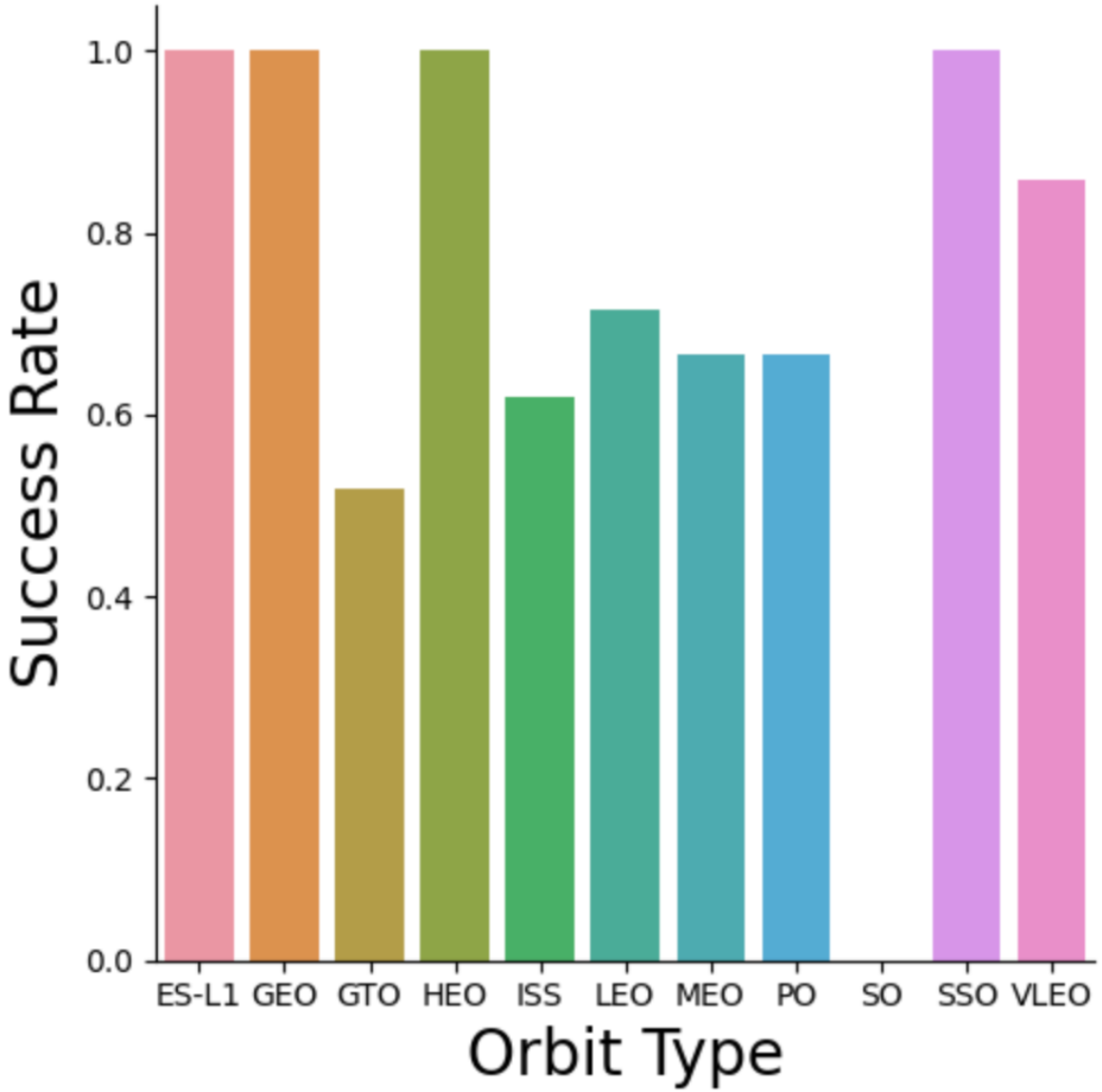
Payload vs. Launch Site



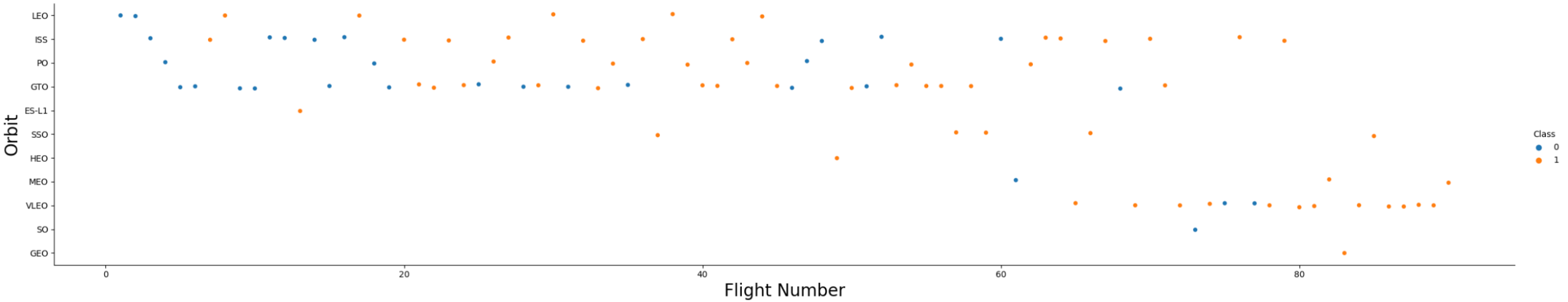
The CCAFS SLC 40 launch site has the highest capacity for heavy payloads. The weight of the payload can be a factor in determining the success of a landing, with heavier payloads potentially requiring a specific launch site. However, a payload that is too heavy can result in a failed landing.

Success Rate vs. Orbit Type

Definitely, ES-L1, GEO, HEO, and SSO have the highest success rate, while GTO has the lowest success rate.Those orbits are likely to be the preferred choices.

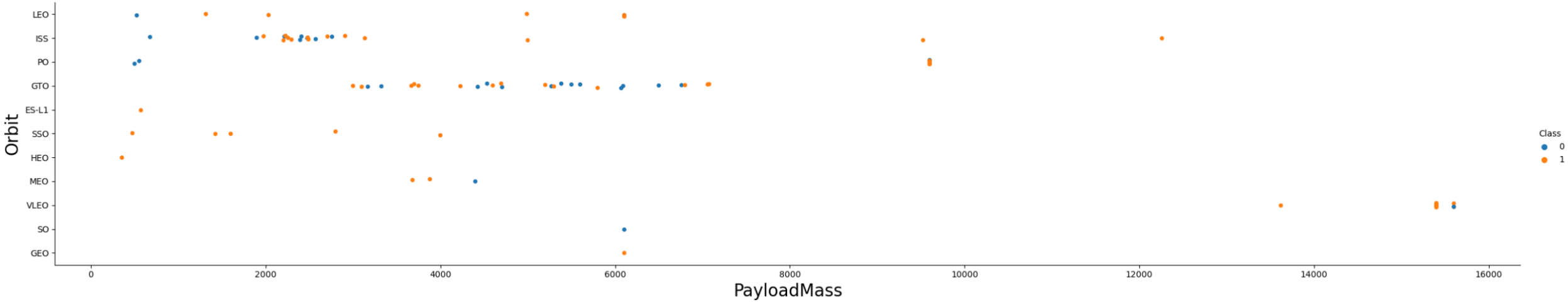


Flight Number vs. Orbit Type



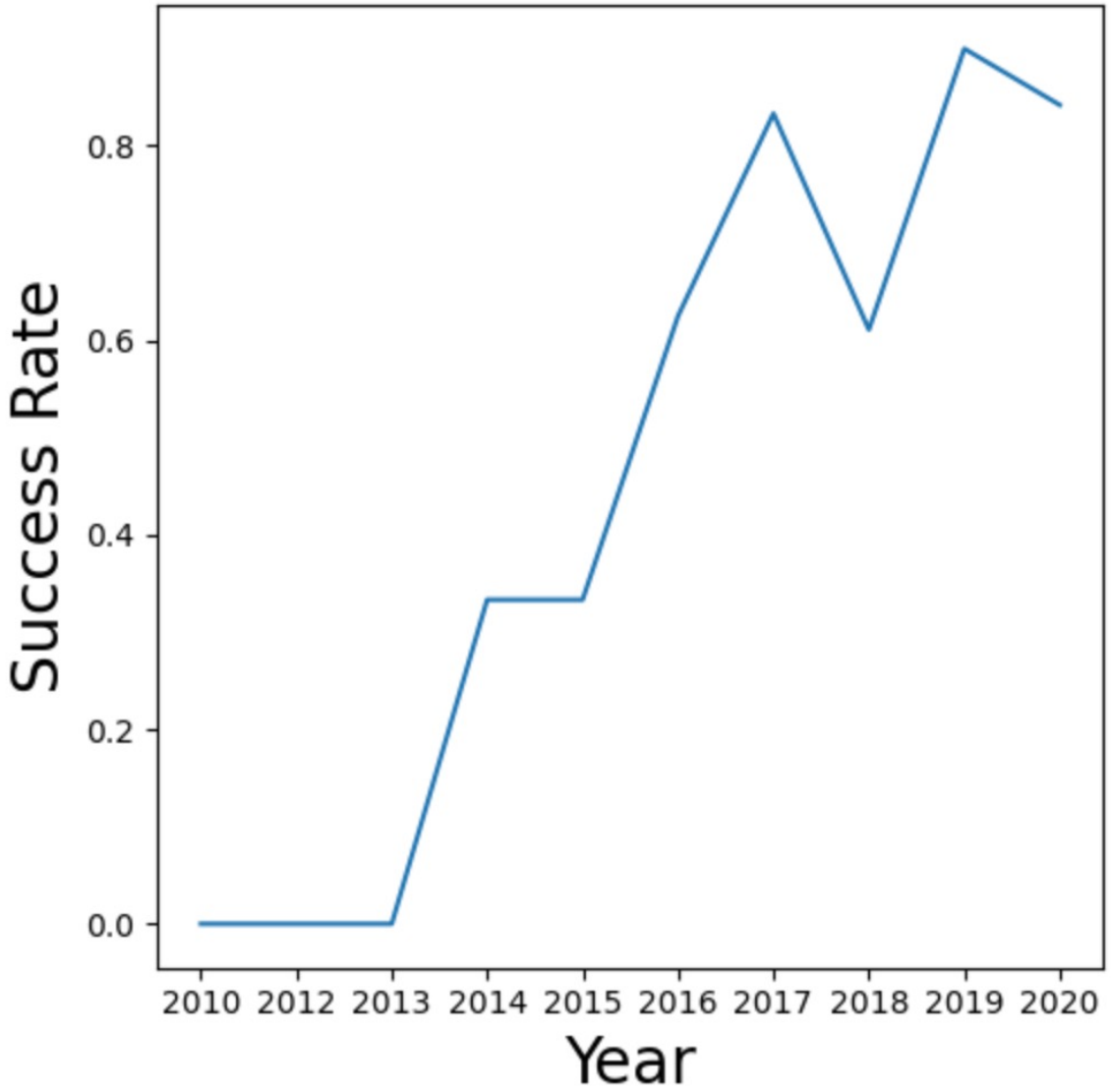
In the LEO orbit, success is correlated with the number of flights, whereas in the GTO orbit, there is no correlation between the number of flights and success.

Payload Vs. Orbit Type



Payload weight can greatly impact the success rate of a LEO orbit launch. Lowering the weight of payloads for GTO orbit launches increases the chances of a successful launch.

Launch Success Yearly Trend



Clearly, with the knowledge gained over the years, success rates for landings have consistently improved. However, in 2019, there was a decline in landing success rates, which recovered in 2020

All Launch Site Names

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

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

Explanation:
There are four unique launch sites. To find this four unique sites, DISTINCT clause was used.

Launch Site Names Begin with 'CCA'

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

Explanation:
Using the LIKE operator allow to find the launch sites beginning with ‘CCA’. The result was limited to show only 5 rows of data.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") as "Total Payload" from SPACEXTBL
```

Total Payload
619967

Explanation:
Using the SUM operator on the Payload Mass KG column, we were able to obtain the total sum of the payload. The final value is 619,967 Kg.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") as "Average Paylaod"
FROM SPACEXTBL
WHERE "Booster_Version" LIKE "F9 v1.1%"
```

Average Paylaod
2534.6666666666665

Explanation:
Using the AVG operator on the Payload Mass KG column, I was able to obtain the total AVG value of the payload. The final value is 2534,66 Kg.

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "Success%"
```

MIN(DATE)
01-05-2017

Explanation:
Using the MIN operator on the dataset and filtering the column landing outcome to find the value of a successful landing.
The first date was 01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT Booster_Version
FROM SPACEXTBL
WHERE
    "PAYLOAD_MASS__KG_" > 4000
    AND
    "PAYLOAD_MASS__KG_" < 6000
    AND
    "Landing _Outcome" = 'Success (drone ship)'
```

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

Explanation:
Values of column payload mass kg were filtered to find those values over 4000 and less than 6000.
Booster versions are as showed in this image.

Total Number of Successful and Failure Mission Outcomes

```
%%sql

SELECT Mission_Outcome, COUNT(*) as mission_outcomes_total
FROM SPACEXTBL
GROUP BY Mission_Outcome
```

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

Explanation:
Using the COUNT and grouping by the Mission Outcome column I find that 100 were success and 1 failure.

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT(Booster_Version)
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
ORDER BY BOOSTER_VERSION
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Explanation:
Distinc to find unique values and where clause with a subquery to find those boosters that match with the maximum payload mass kg value.

2015 Launch Records

```
%%sql
SELECT SUBSTR("Date",4,2) as Month, Booster_Version, Launch_Site, Mission_Outcome
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Failure (drone ship)'
    AND SUBSTR("Date",7,4) = '2015'
```

Month	Booster_Version	Launch_Site	Mission_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Success
04	F9 v1.1 B1015	CCAFS LC-40	Success

Explanation:

Using substrings to filter with selected date of 2015 and landing outcome like Failure (drone ship) to retrieve the launch records asked.

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

```
%%sql
SELECT "Landing _Outcome", COUNT(*) AS conteo
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE 'Success%'
    AND "DATE" BETWEEN '04-06-2010' AND '20-03-2017'
ORDER BY conteo DESC
```

Landing _Outcome	conteo
Success (drone ship)	34

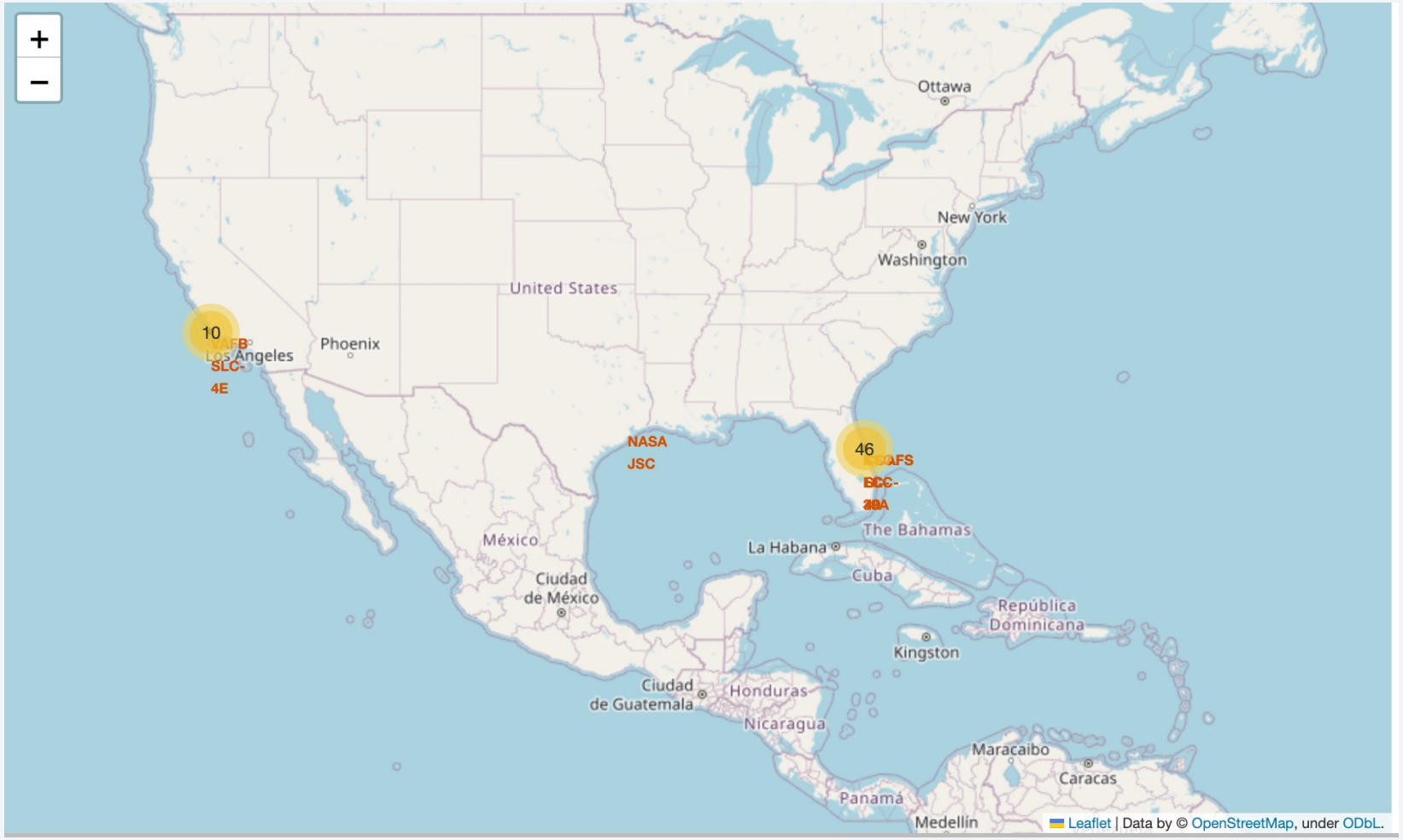
Explanation:

Using the COUNT and filtering landing outcome with the word ”success” and the date between 04-06-2010 and 20-03-2017.

This shows a total count value of 34 success (drone ship) rows.

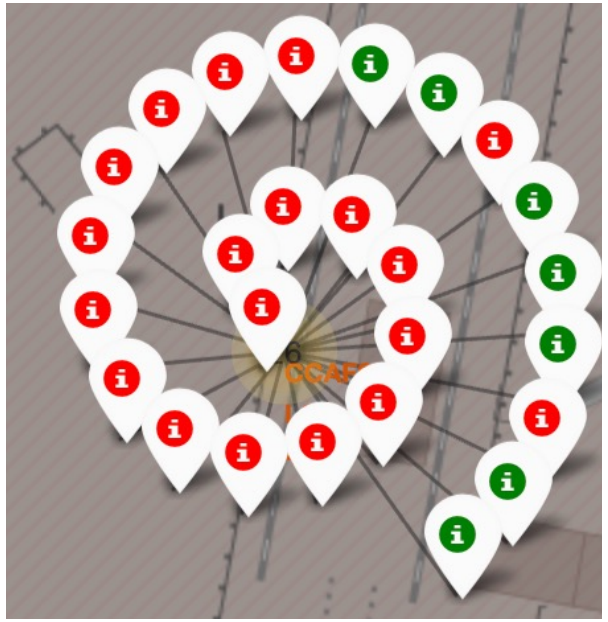

Section 3
Launch Sites
Proximities Analysis

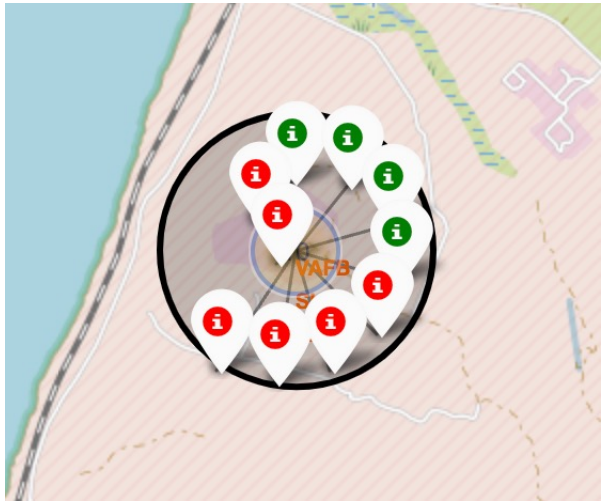
Folium Map – Launch Sites

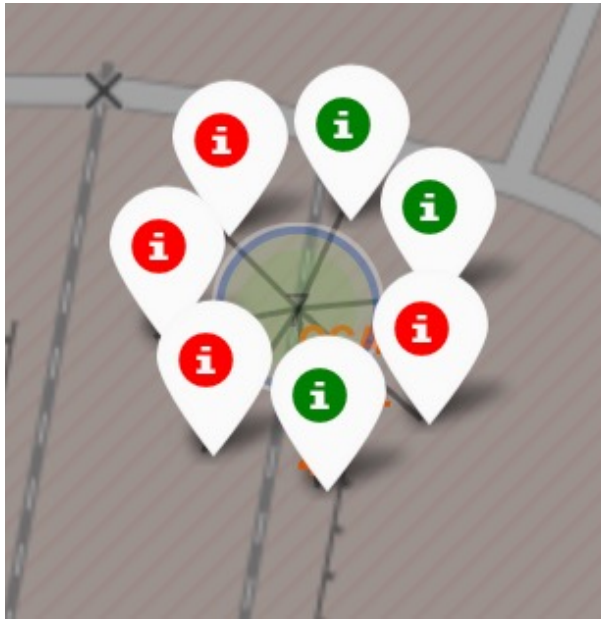


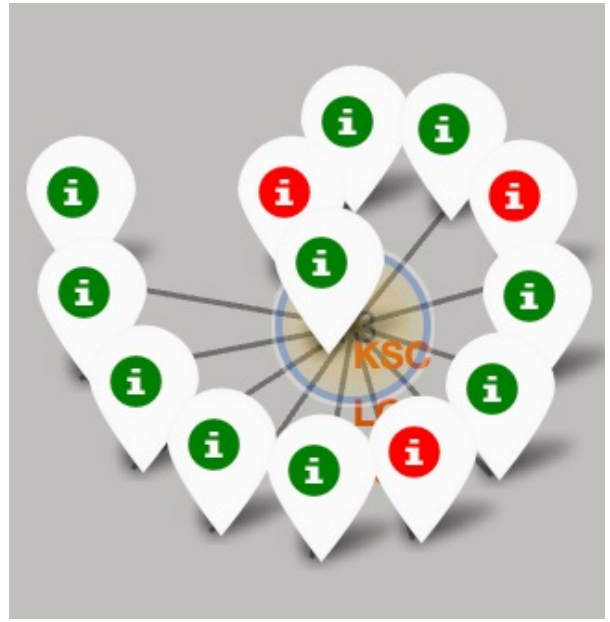
Launch sites are located near coasts of United States

Folium Map – Color labeled Launch Sites





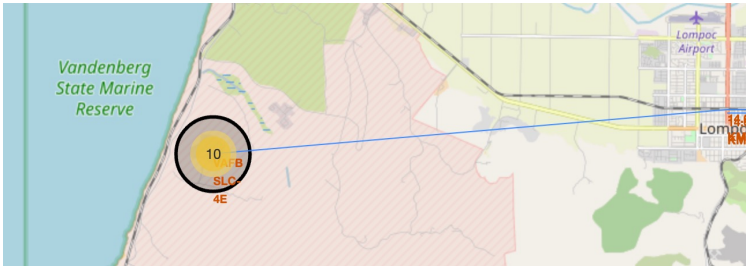


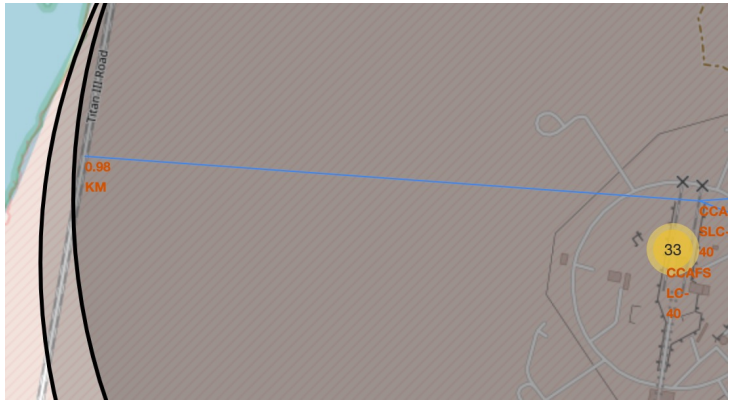


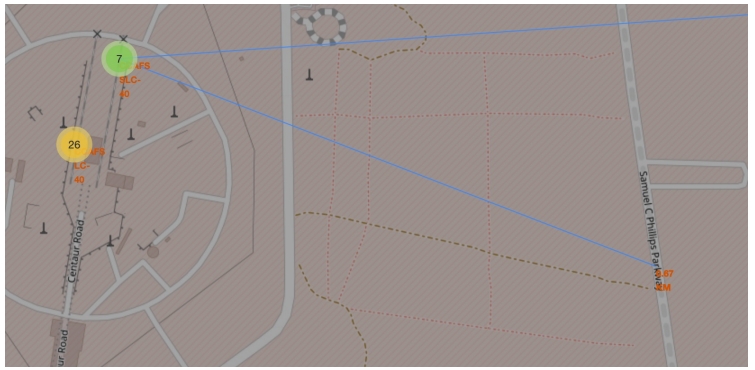
Green markers: successful launches Red markers: Failed launches

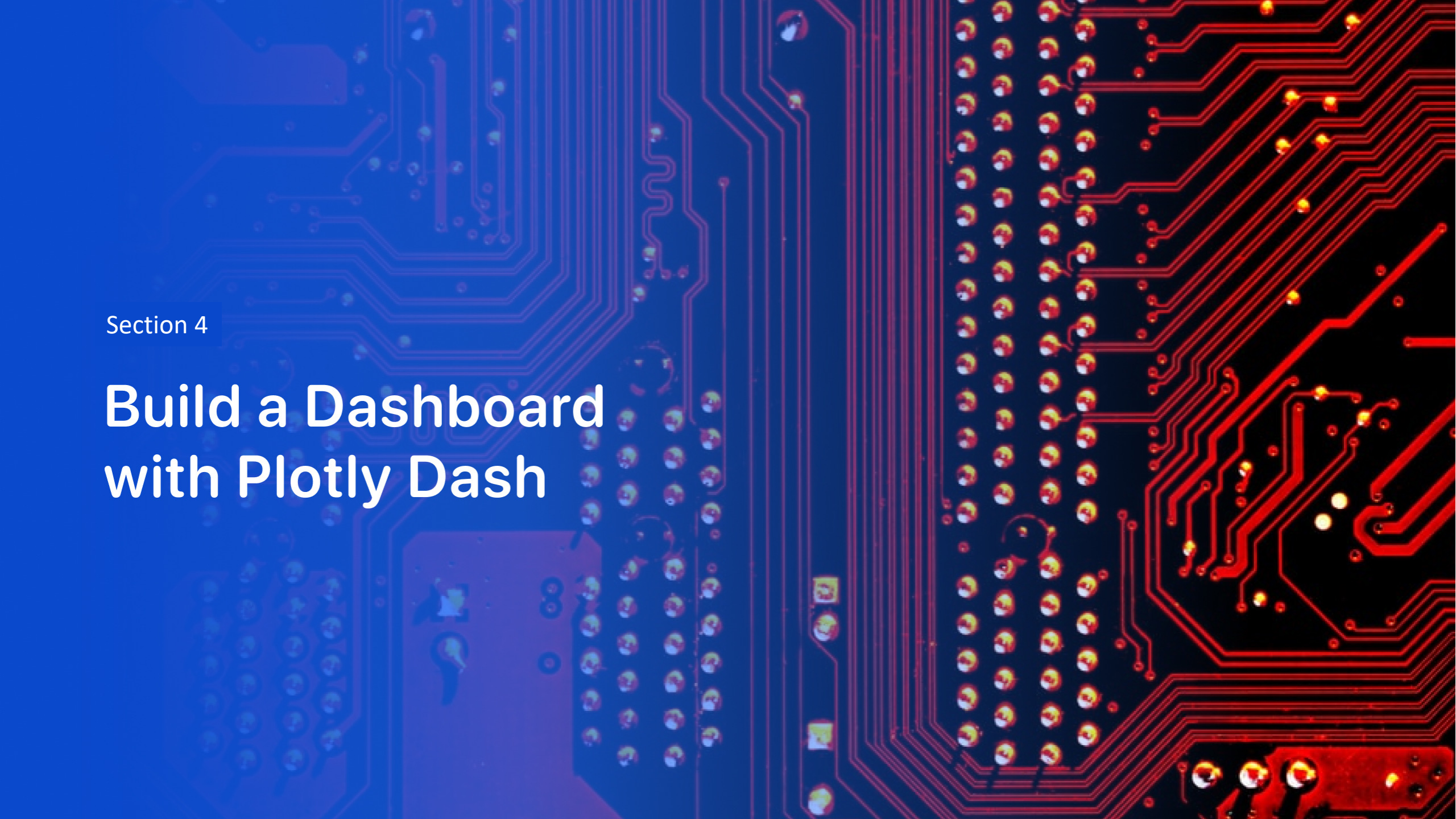
Folium Map - Proximity Sites

- Are launch sites in close proximity to railways? YES
- Are launch sites in close proximity to highways? YES
- Are launch sites in close proximity to coastline? YES
- Do launch sites keep certain distance away from cities? YES

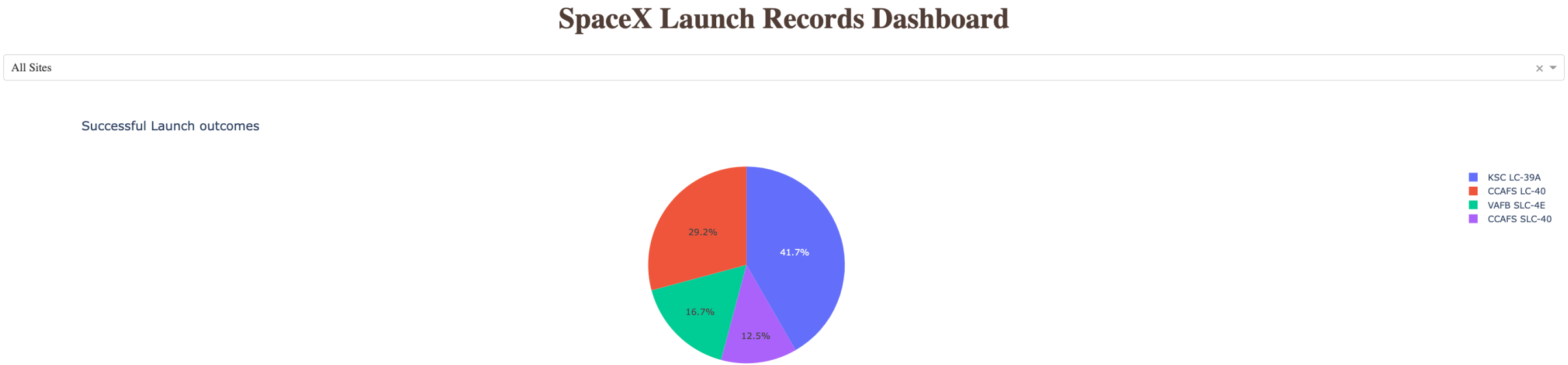






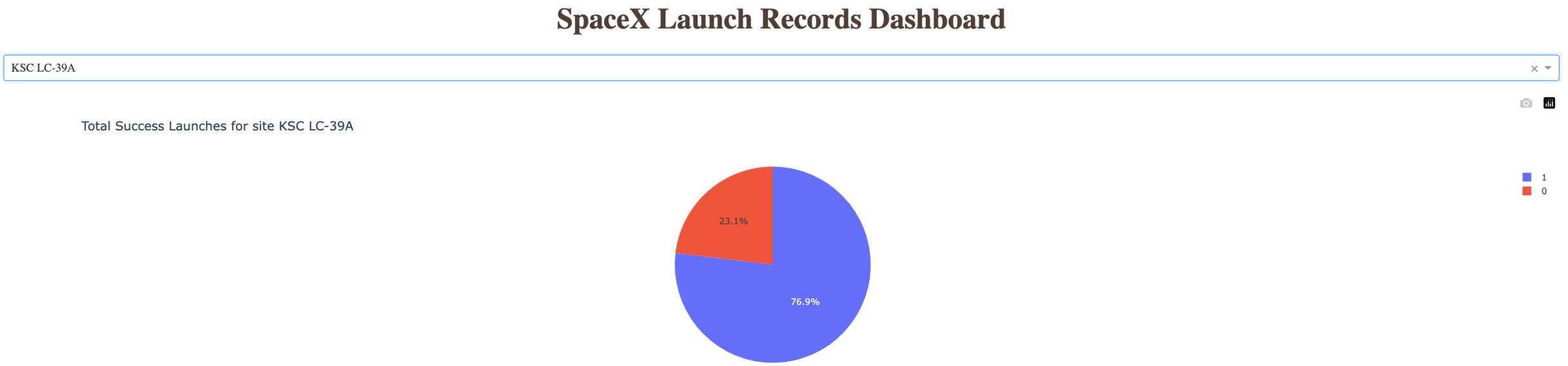

Section 4
Build a Dashboard with Plotly Dash

Pie Chart - All Sites



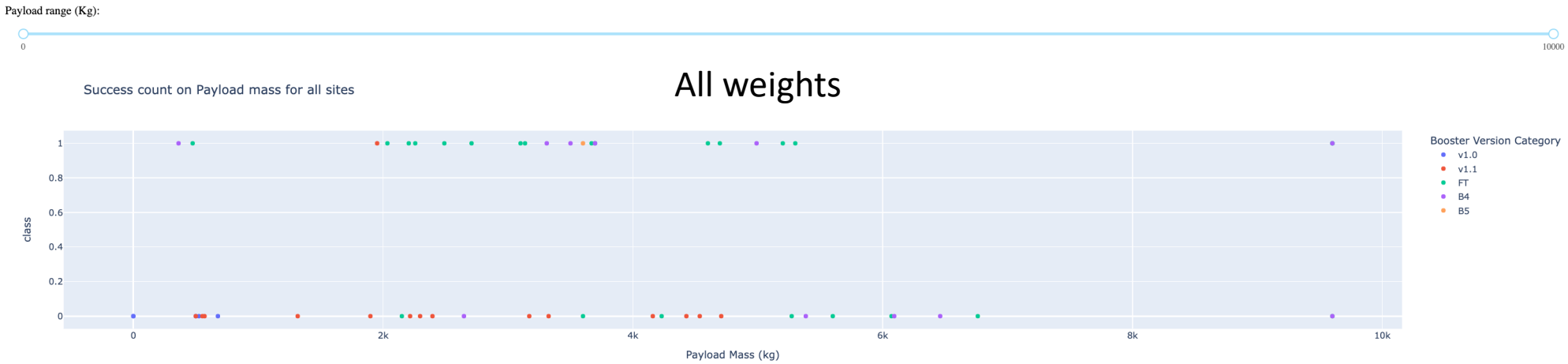
The KSC LC-39A launch site has the highest success rate for launches.

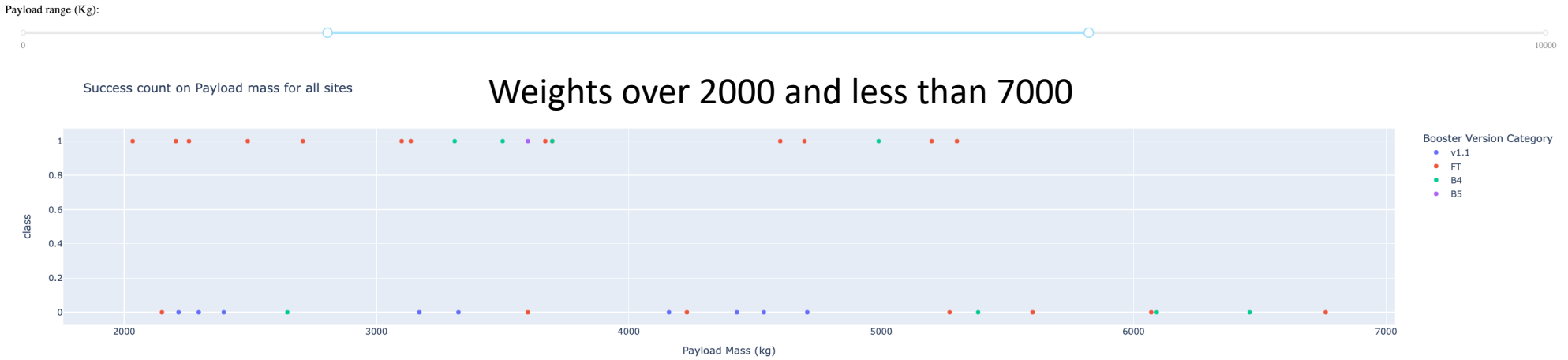
Pie Chart for the Highest Launch Success



"The KSC LC-39A launch site has a 76.9% success rate and a 23.1% failure rate."

Payload Mass Vs. Launch Outcome for all sites with different payload range selecteds

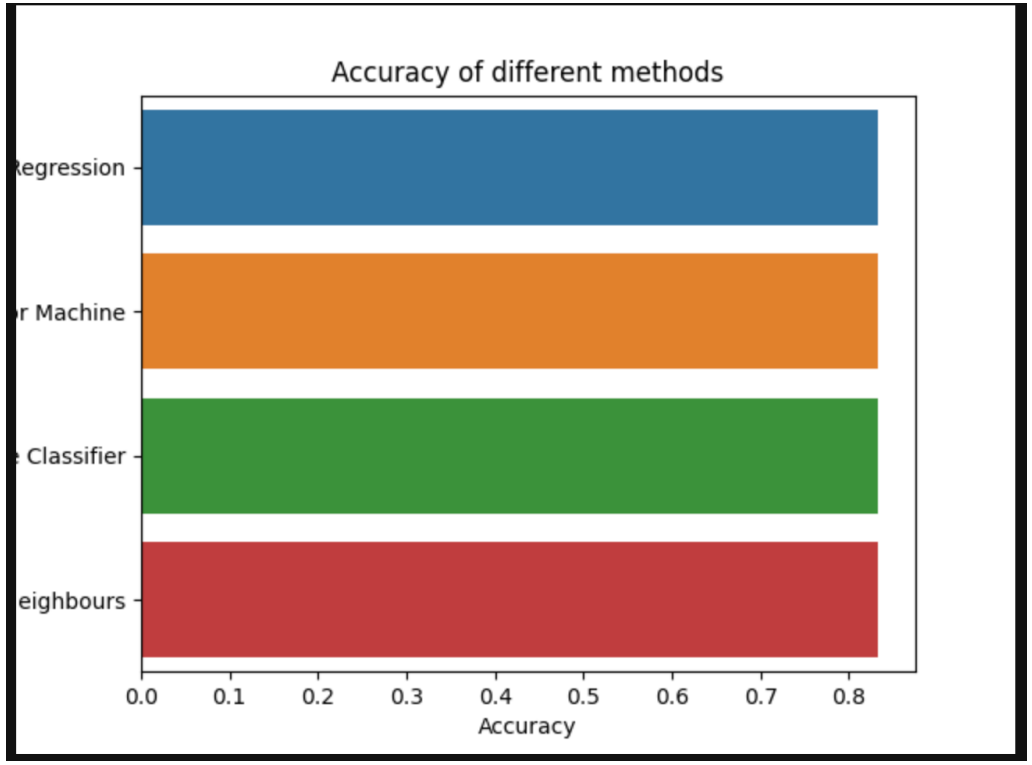




Section 5

Predictive Analysis (Classification)

Classification Accuracy

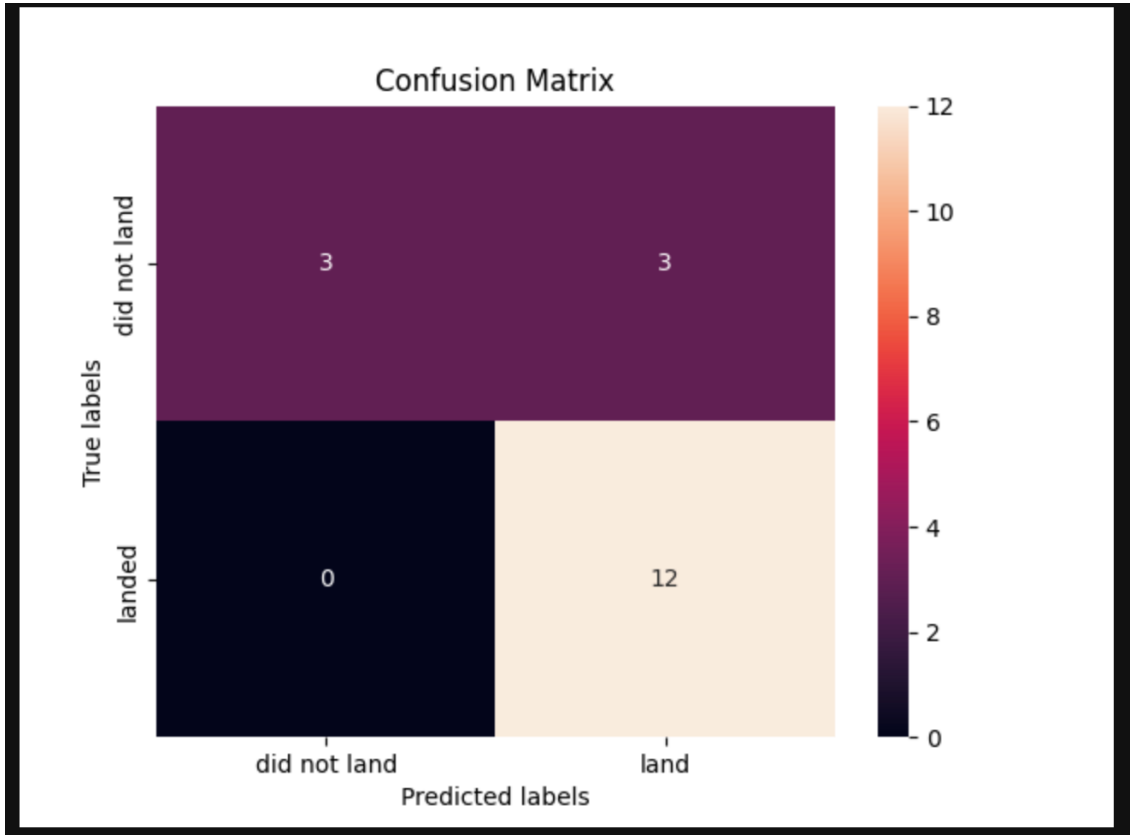


	methods	accuracy
0	Logistic Regression	0.833333
1	Support Vector Machine	0.833333
2	Decision Tree Classifier	0.833333
3	K Nearest Neighbours	0.833333

In the accuracy test, all methods performed similar.

Confusion Matrix

- Since all models had the same accuracy, the decision tree's confusion matrix was chosen as it appears to be the most optimal.



Decision Trees

Conclusions – SpaceX

- More flights at a launch site lead to a higher success rate.

The weight of the payload can be a crucial factor in determining the success of a mission depending on the orbit.

- Launches have been increasingly successful from 2013 till 2020.

- Orbits ES-L1, GEO, HEO, SSO, VLEO have the most favorable outcomes.

- KSC LC-39A launch site has the highest number of successful launches.

- The Decision tree classifier was selected as the best model for this task.

Thank you!