IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Stephen Mwangi Thuo
13th August 2023

Outline

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

Executive Summary

Our goal is to predict the successful landing of the SpaceX Falcon 9 first stage. This prediction is crucial as it directly impacts the cost of a launch.

Our approach involves utilizing various machine learning classification algorithms to achieve this prediction.

The methodology adopted encompasses key stages:
Data Collection, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization, and Machine Learning Prediction. Through this comprehensive process, we aim to identify patterns and insights within the dataset.

It is observed significant correlations between certain features of rocket launches and their success or failure outcomes.

Notably, our investigation revealed that multiple classification algorithms achieved the same accuracy on the test data, reaching an accuracy level of 83.33%.

However, it's noteworthy that the Decision Tree algorithm emerged as a strong contender for this specific problem.

Introduction

The project answers the question: Can we predict the successful landing of the Falcon 9 first stage?

SpaceX, has fundamentally altered the economics of rocket launches. By reusing the first stage, SpaceX drastically reduces costs.

This economic advantage hinges on the ability to determine whether the first stage will safely return to Earth.

Given the various variables associated with a Falcon 9 rocket launch, can we anticipate a successful landing of its first stage?

The analysis through machine learning techniques, not only unveils predictive potential but also underscores the transformational impact of such insights.

Section 1

Methodology

Methodology

Executive Summary

We harnessed the SpaceX API to tap into a trove of mission-related information. In addition web scraping was done to extract launch data from a Wikipedia page. Data was transformed and cleaned using Python's pandas library.

Visualization tools like Matplotlib and Seaborn were used to unravel trends, correlations, and patterns. SQL queries also served as a tool to slice through the dataset to extract valuable insights

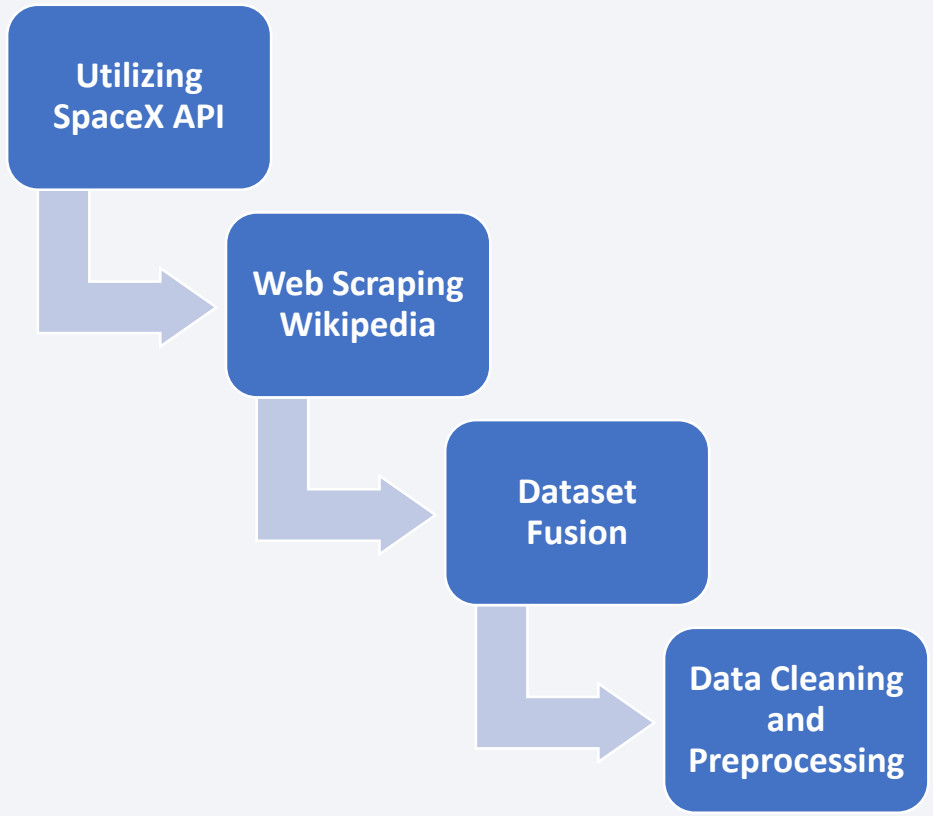
Folium enabled us to create dynamic maps, visualizing geographical aspects, while Plotly Dash was used to build interactive visuals.

Four models, namely logistic regression, support vector machines, k-nearest neighbors, and the decision tree classifier, were used and hyper parameter tuning done using Grid Search CV. For each model, we followed a structured approach. The foundation was laid with training, enabling models to learn from the data. Its important to note that the model were fed into the Grid Search pipeline to give the best parameters for each model.

The models were then tested against test data to determine their real-world performance. This iterative process revealed the strengths and weaknesses of each model, enabling us to select the optimal candidate for predictive analysis.

Data Collection

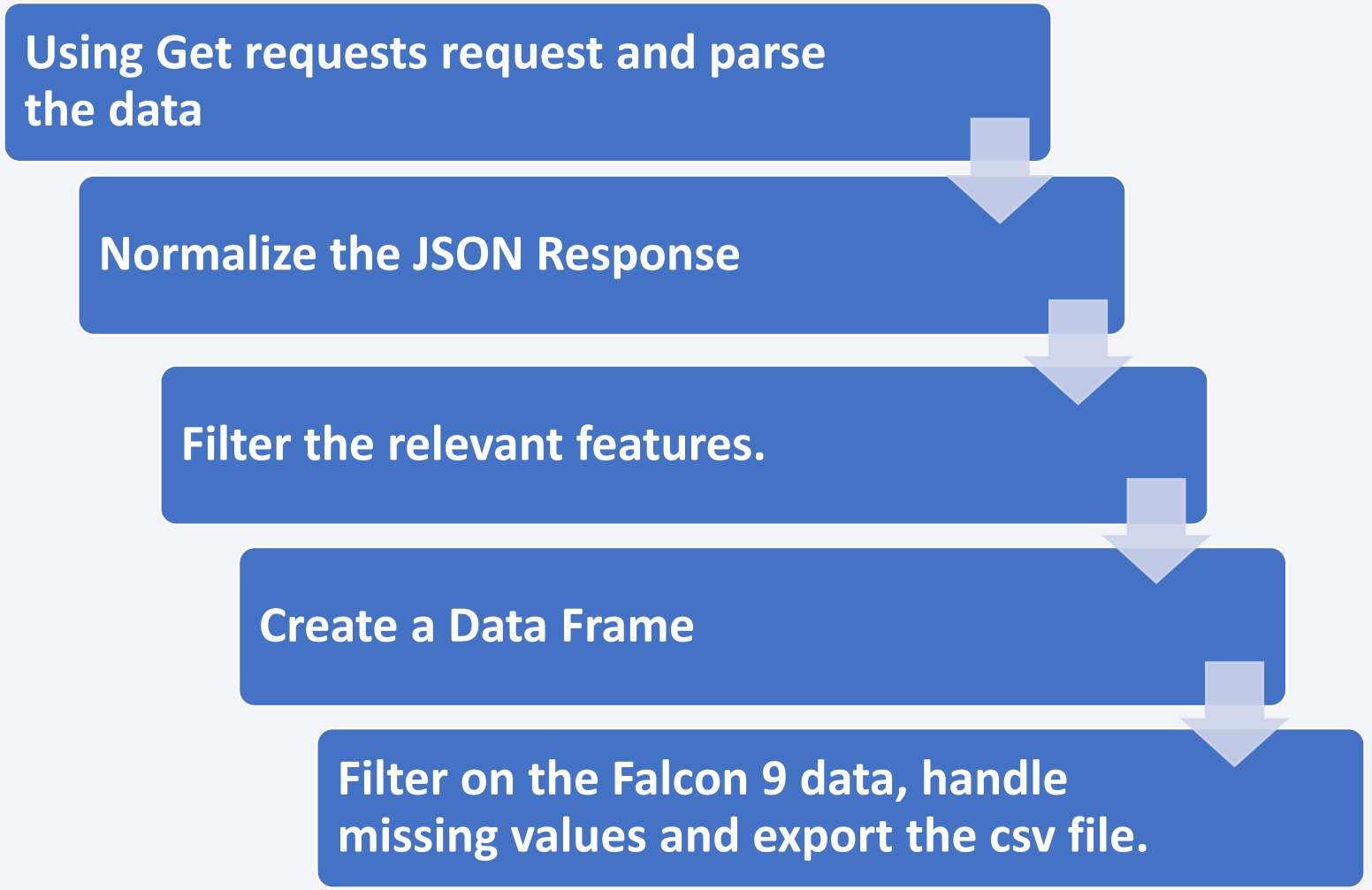
- The data collection process involved a combination of two methods: utilizing the SpaceX API and performing web scraping.



GitHub

Data Collection - SpaceX API

GitHub URL



Data Collection - Scraping

GitHub URL.

Send request from rocket launch data on Wikipedia.

Extracting column names from the HTML table.

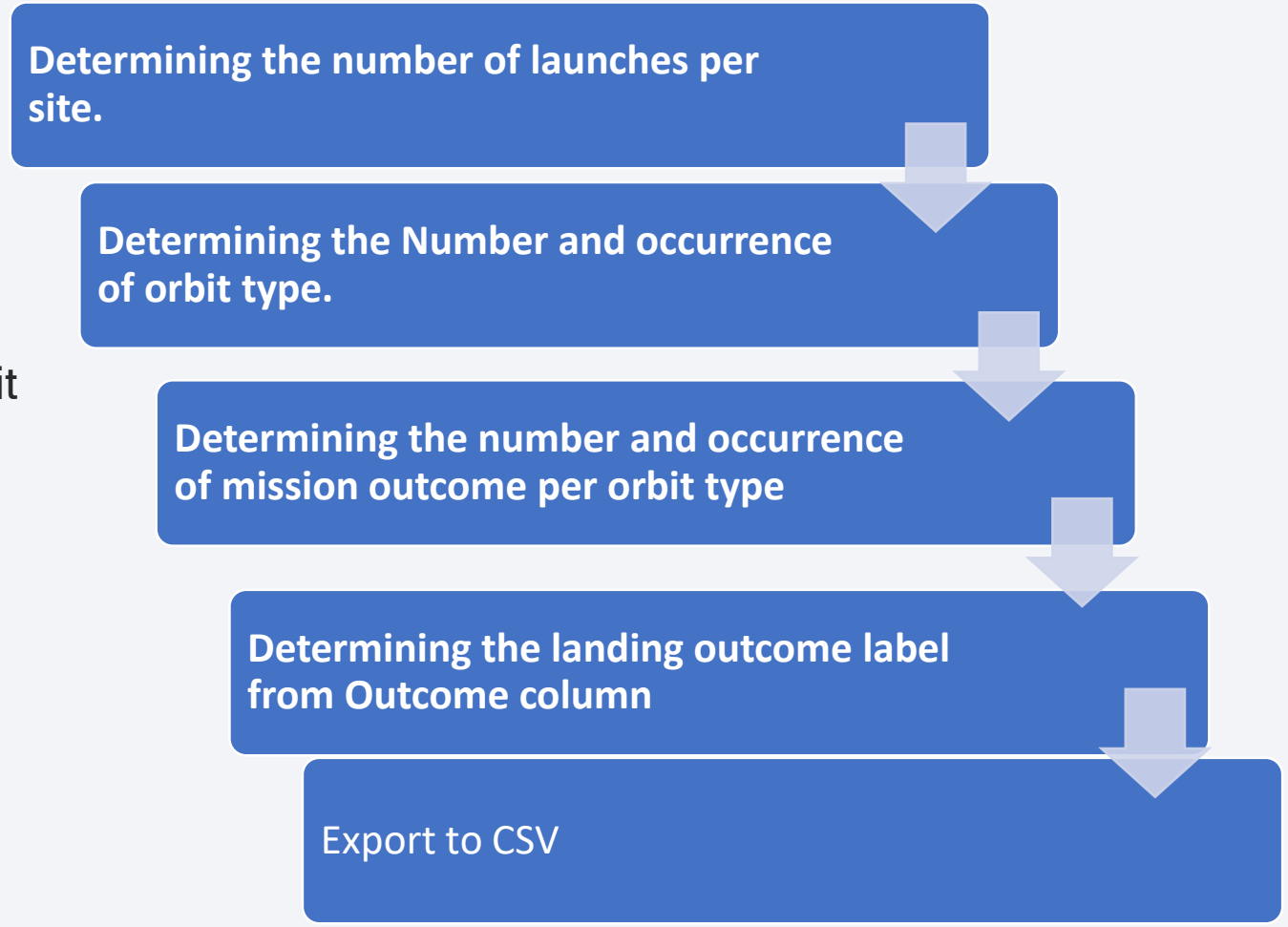
Adding the launch HTML tables data into a Data Frame.

Export to CSV.

Data Wrangling

- **Data Cleaning and Transformation**
  - Missing values were identified.
  - Data types were inspected.
  - Categorical variables were encoded using one-hot encoding to convert them into a numerical format.
  - Inspection of mission outcomes per orbit

GitHub URL.



EDA with Data Visualization

- Scatter Plots: Scatter plots were employed to visualize the relationships between pairs of variables. Various combinations of features were analyzed, including Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, and Payload vs. Orbit Type.

- Bar Chart: The utilization of bar charts facilitated quick comparisons of values across multiple categories. These charts featured a categorical x-axis and a discrete y-axis, allowing easy visualization of data distribution. They were applied to examine the Success Rate across different Orbit Types.

- Line Chart: Line charts were harnessed to display data trends over time. Specifically, a line chart was utilized to illustrate the changes in Success Rate over a specific span of years.

GitHub URL.

EDA with SQL

Below are SQL Queries made while exploring the data:

1. Presenting a compilation of distinct launch site names within the space mission.
2. Showcasing five records featuring launch sites starting with the prefix 'CCA.'
3. Illustrating the cumulative payload mass carried by boosters launched under NASA's (CRS) missions.
4. Demonstrating the average payload mass carried by booster version F9 v1.1.
5. Listing the date of the earliest successful landing on a ground pad.
6. Enumerating the names of boosters achieving success on a drone ship with a payload mass ranging from 4000 to 6000.
7. Enumerating the total count of both successful and failed mission outcomes.
8. Detailing the booster version names associated with the highest payload mass carried.
9. Displaying the names of booster versions, launch site names, and failed landing outcomes on drone ships in the year 2015.
10. Ranking the count of landing outcomes between June 4, 2010, and March 20, 2017, in descending order.

GitHub URL

Build an Interactive Map with Folium

- Incorporating various components onto a Folium map, the project involved creating and integrating objects.
- Specifically, marker objects were strategically employed to visually represent the locations of all launch sites on the map, while also providing insights into the success or failure of launches at each site.
- Additionally, line objects served the purpose of calculating and presenting the distances between launch sites and their respective neighboring points.
- This endeavor combined both interactive markers and informative lines to enhance the visual and analytical aspects of the map representation.

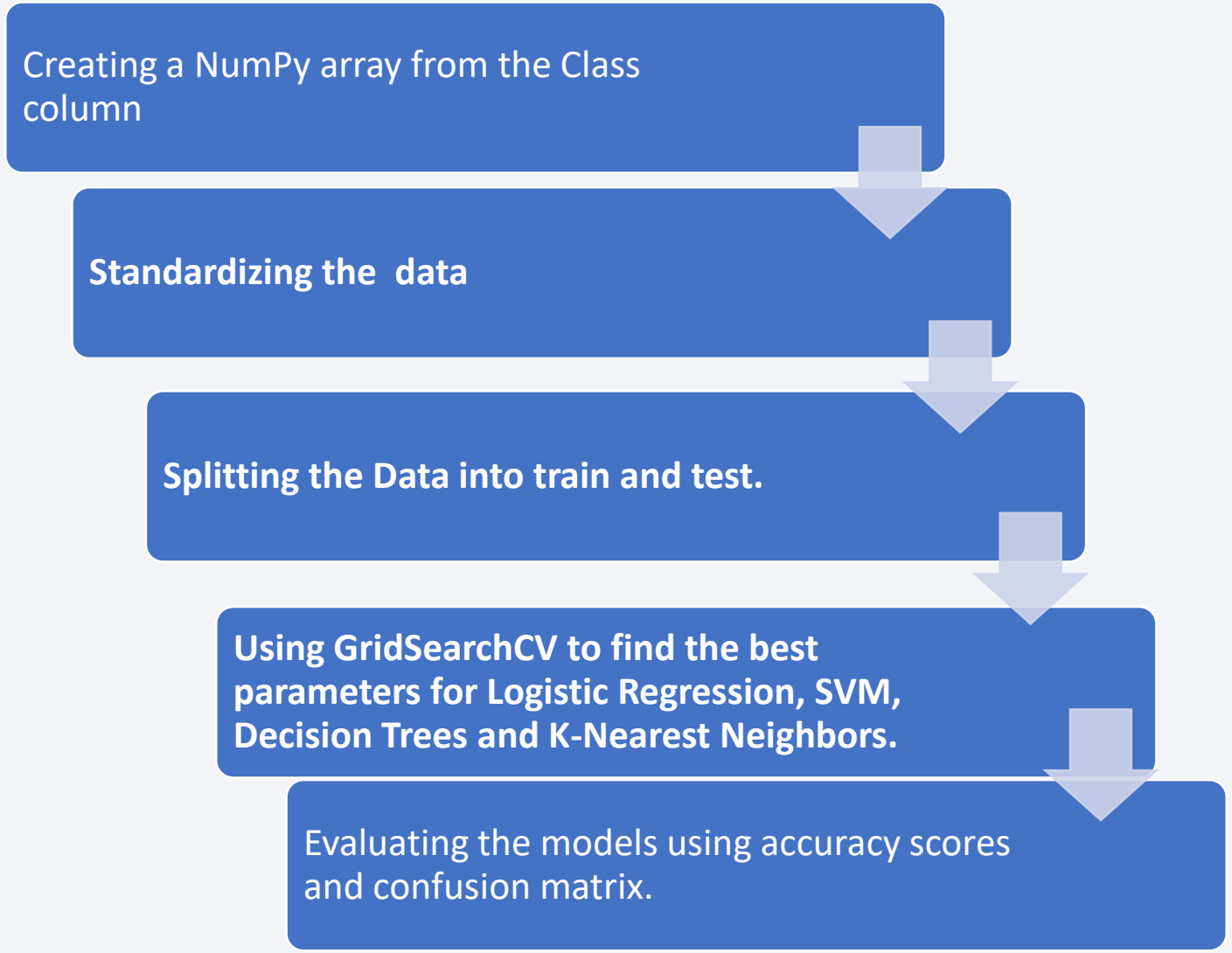
GitHub URL

Build a Dashboard with Plotly Dash

- Presenting a pie chart that comprehensively depicts the distribution of successful launches across each launch site..
- The dashboard incorporates a scatter chart that shows the connection between landing outcomes and the payload mass of distinct boosters.
- The dashboard's interactive features enable users to input specific launch site(s) and payload mass values, fostering an enriched understanding of how different variables intricately impact the landing outcomes of SpaceX missions.

GitHub URL

Predictive Analysis (Classification)

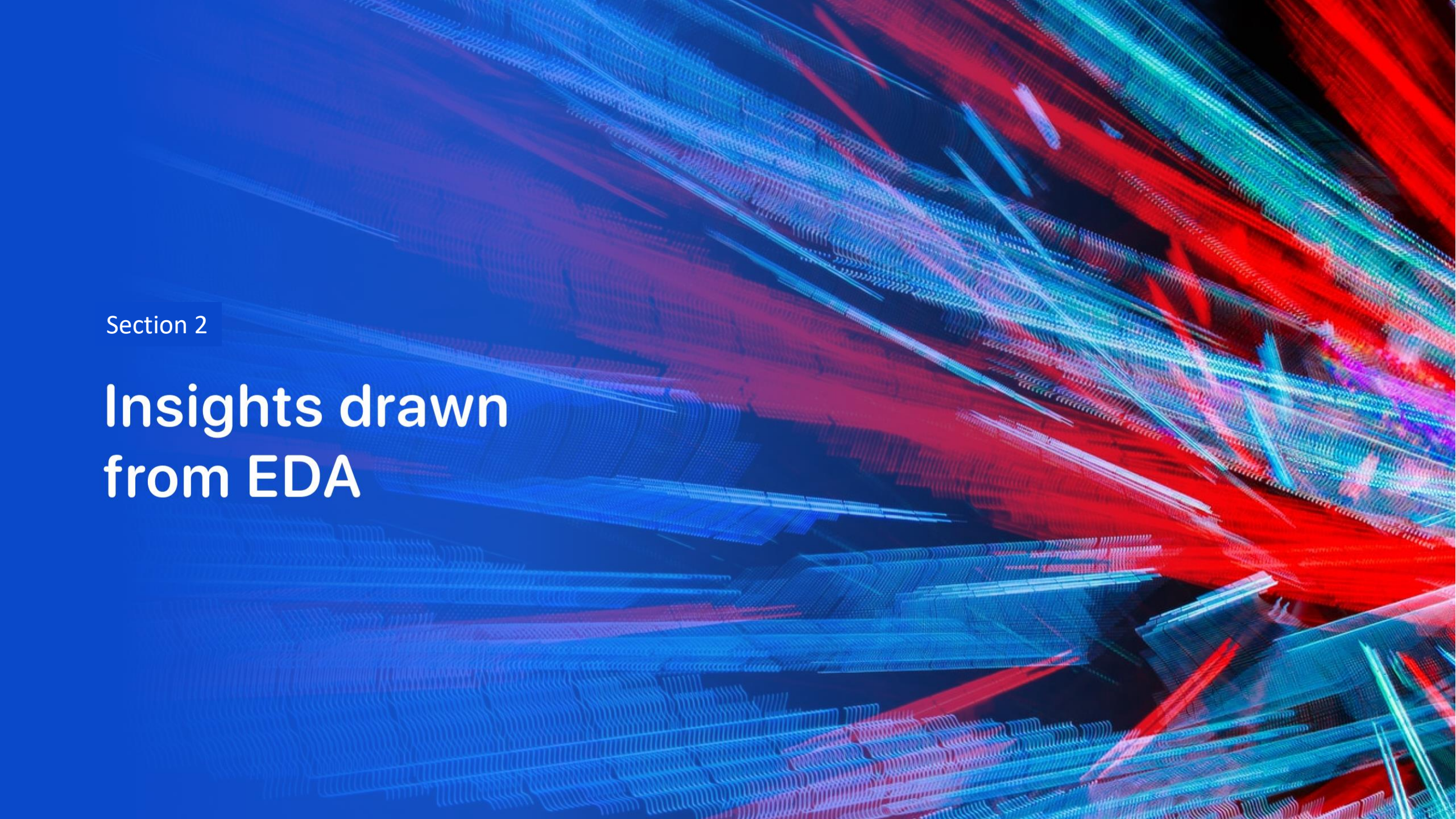


GitHub URL

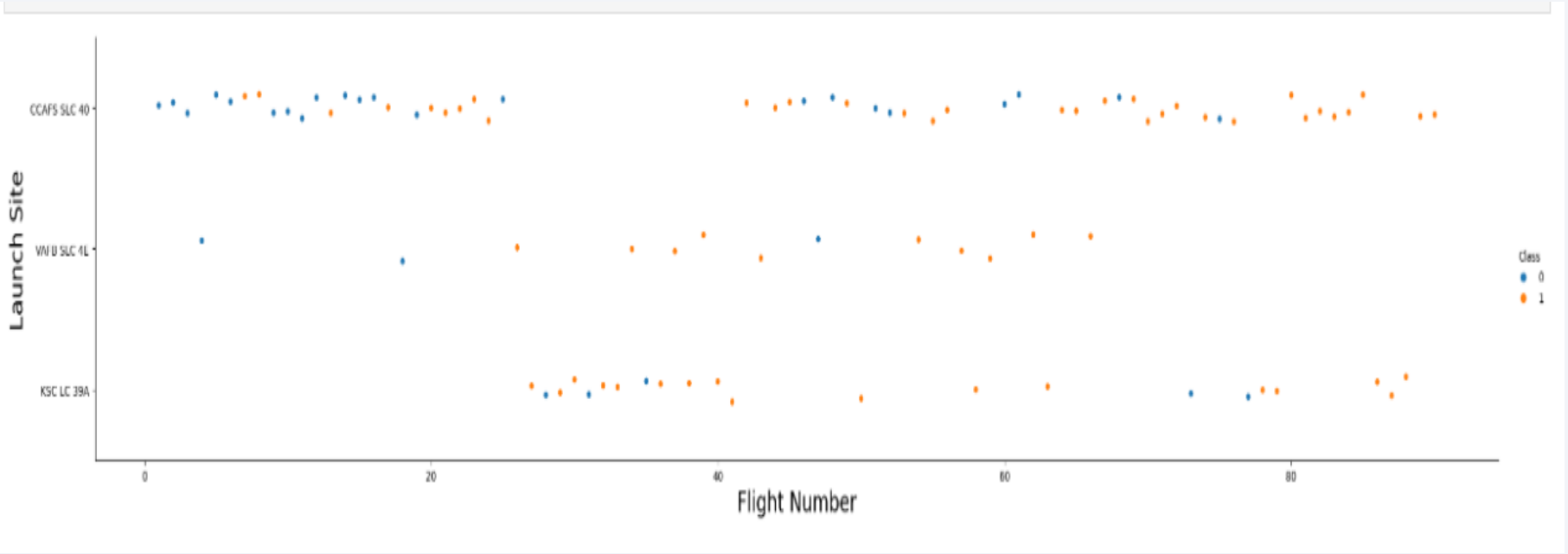
Results

The exploratory data analysis findings unveiled a Falcon 9 landing success rate of 66.66%. Our predictive analysis, on the other hand, yielded a significant outcome.

Among the classification algorithms tested, the Decision Tree algorithm stood out as the most effective, boasting an impressive accuracy of 94%. This robust accuracy underscores the reliability of the Decision Tree in predicting the success of Falcon 9 landings.

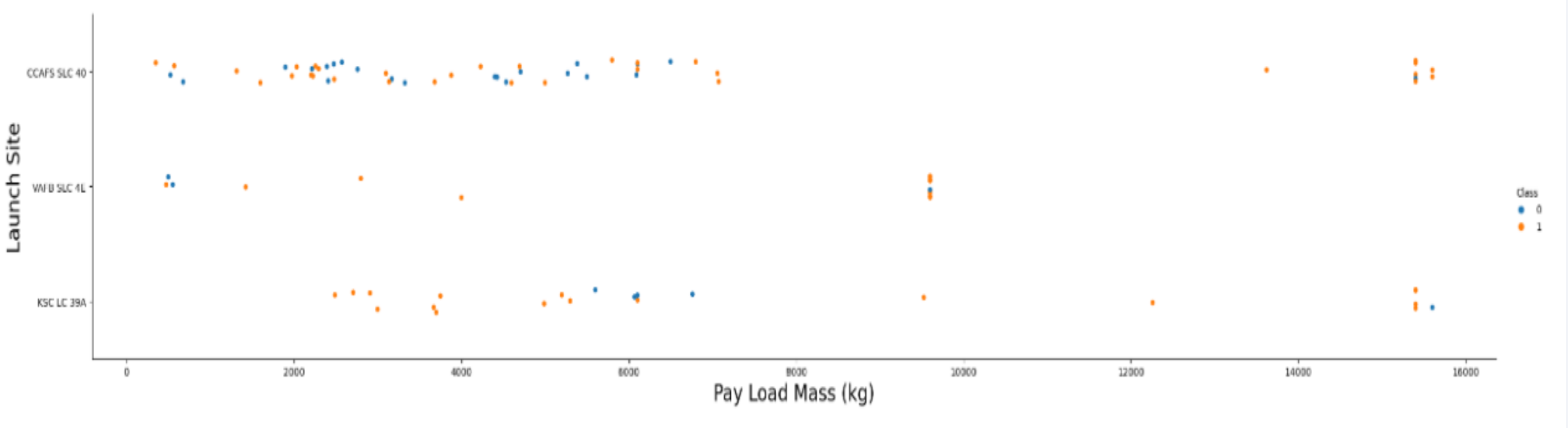

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



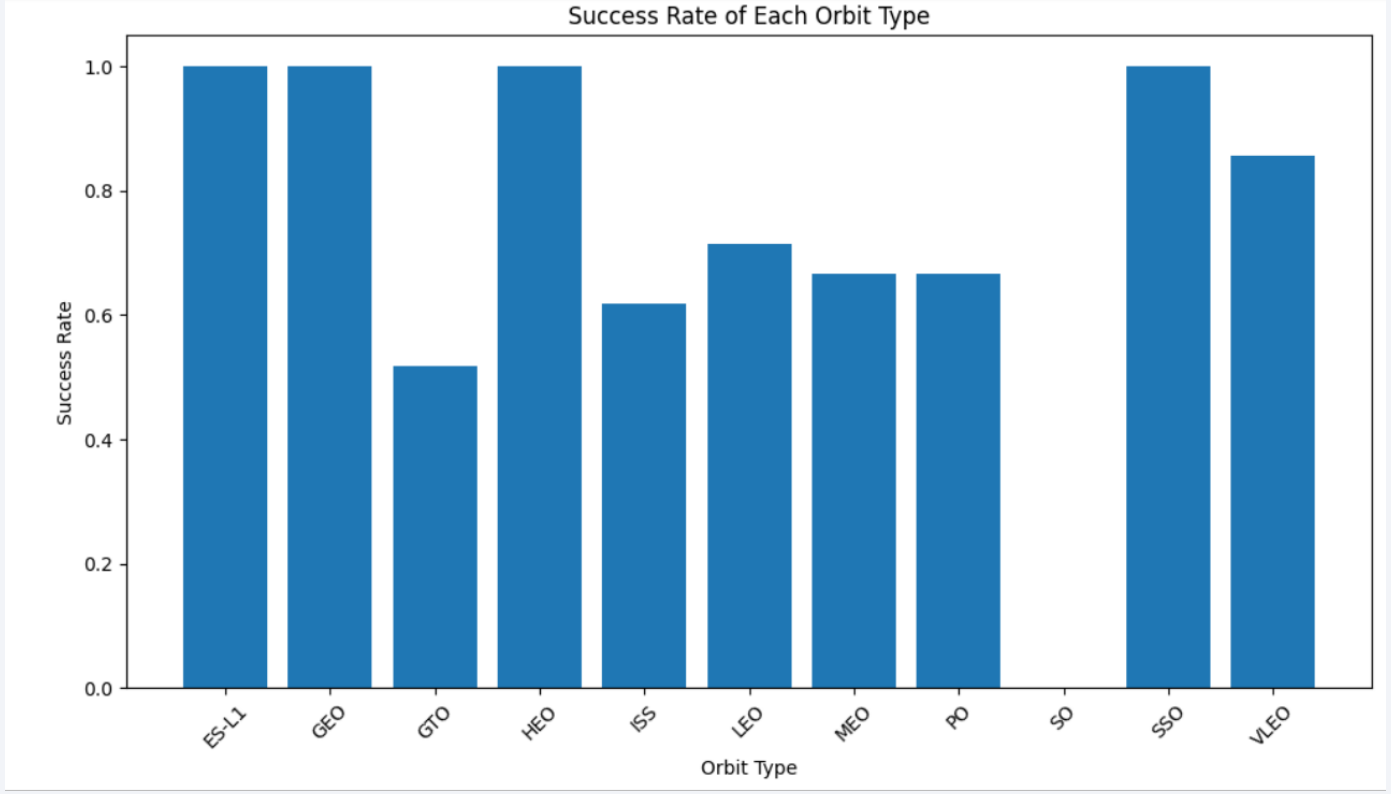
- The blue dots represent the successful launches while the red dot represent unsuccessful launches.
- There seems to be an increase in successful flights after the 40th launch.

Payload vs. Launch Site



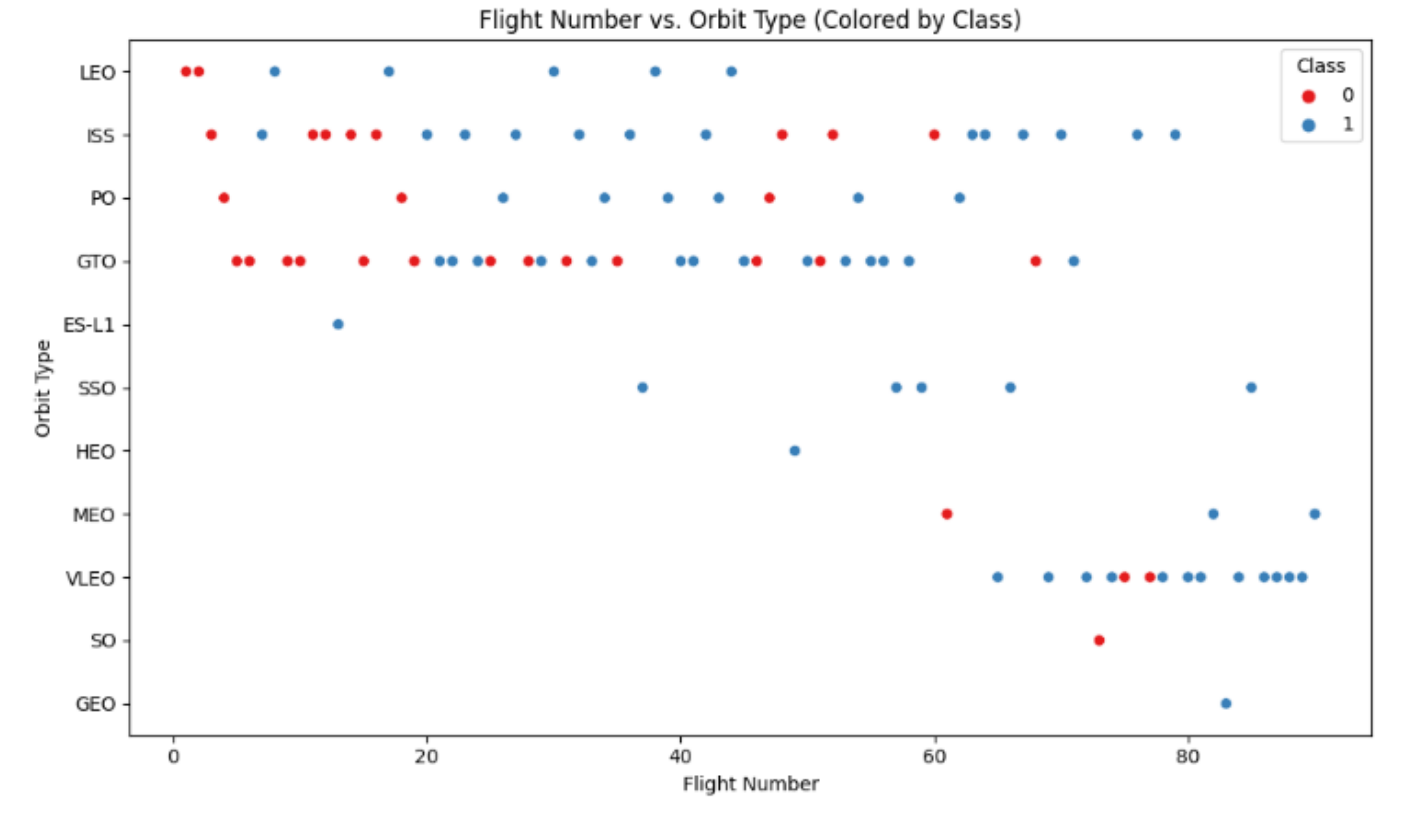
- There is a weak correlation between Payload and Launch Site.

Success Rate vs. Orbit Type



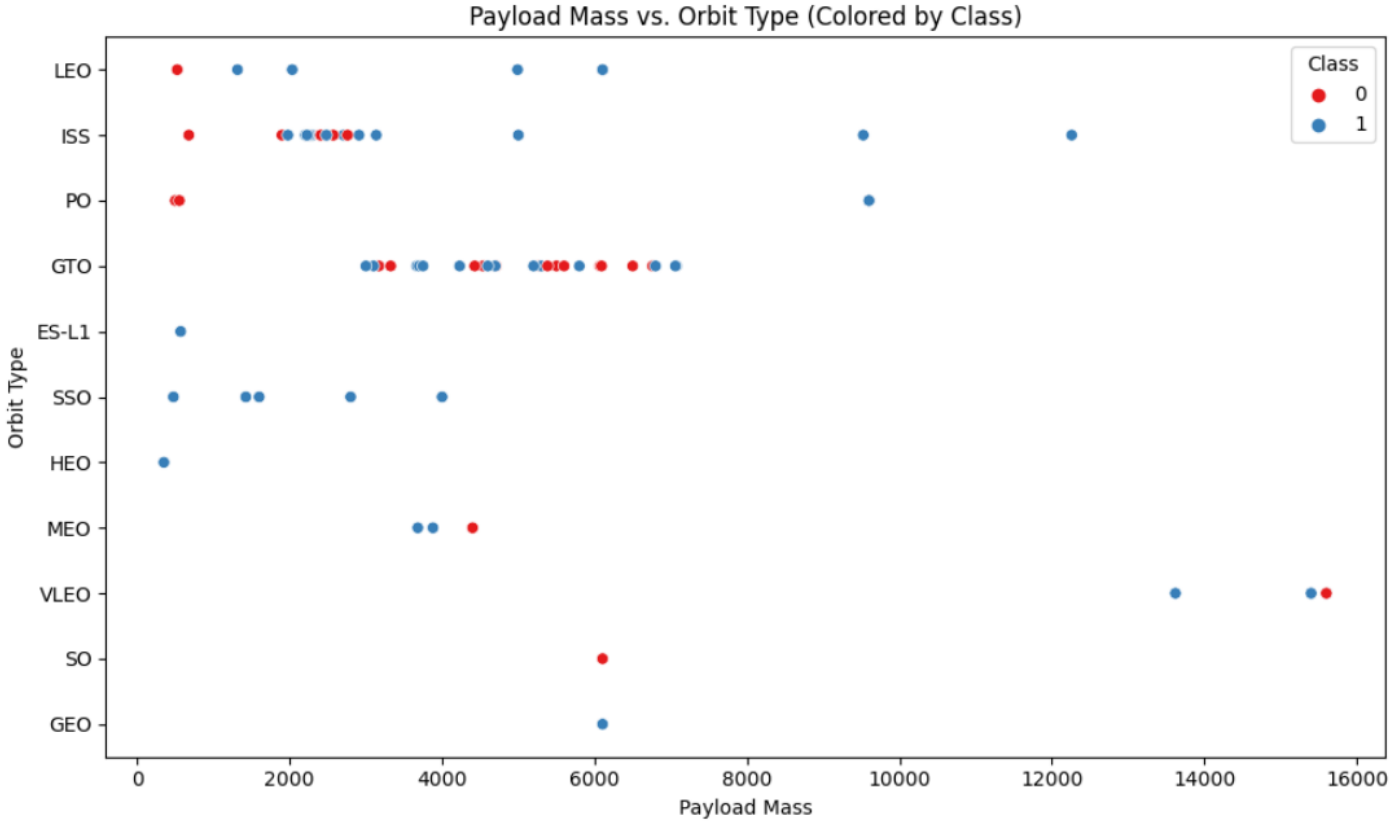
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.
- SO orbit did not have any successful launches with a 0% success rate.

Flight Number vs. Orbit Type



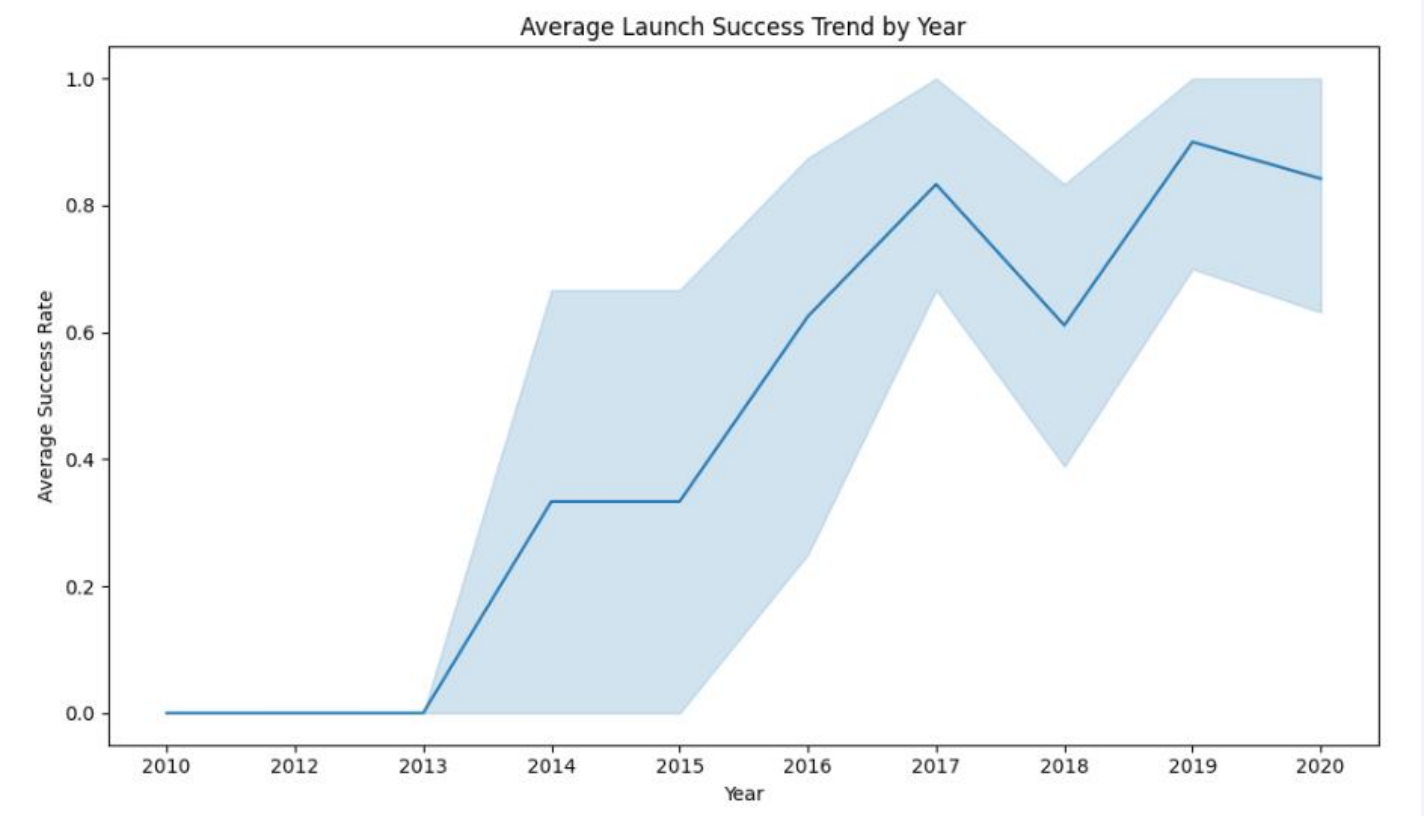
- In the LEO orbit, the success is positively correlated to the the number of flights.
- The SSO orbit has a 100% success rate however with fewer flights than the other orbits
- Flight numbers exceeding 40 exhibit a higher success rate compared to those falling within the range of 0 to 40..

Payload vs. Orbit Type



- With increasing payload mass, the success rate rises within the PO, SSO, LEO, and ISS orbits.
- The GTO orbit, there appears to be no clear correlation between orbit type and payload mass, as both successful and failed launches are evenly distributed.

Launch Success Yearly Trend



- The general trend of the chart shows an increase in landing success rate as the years pass.

All Launch Site Names

```
launch_sites_result = %sql SELECT DISTINCT "Launch_Site" AS launch_sites FROM SPACEXTBL;

# Convert the result to a pandas DataFrame
launch_sites_result
```

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

launch_sites
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

```
# Write the SQL query as a string
sample_CCA_records = %sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
sample_CCA_records
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No atte
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No atte
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No atte

- To display 5 records that we starts with CCA, LIMIT and LIKE clauses were used.

Total Payload Mass

```
# Execute the SQL query
total_CRS_payload_result = %sql SELECT SUM("PAYLOAD_MASS__KG_") AS "Total_NASA_CRS_Payload" FROM SPACEXTBL WHERE "Customer"

# Print the total payload mass carried by boosters launched by NASA (CRS)
total_CRS_payload_result
```

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

Total_NASA_CRS_Payload
45596.0

The SUM Function was used to get the total payload mass.

Average Payload Mass by F9 v1.1

```
# Execute the SQL query
average_payload_f9 = %sql SELECT AVG("PAYLOAD_MASS__KG_") AS "Average_F9_V1.1_Payload" FROM SPACEXTBL WHERE "Booster_Versior

# Print the total payload mass carried by boosters launched by NASA (CRS)
average_payload_f9
```

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

Average_F9_V1.1_Payload
2928.4

- The AVG() function was used to the calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on *booster_versions* only if they were named "F9 v1.1"

First Successful Ground Landing Date

```
first_success = %sql SELECT MIN("Date") AS 'First_Success' FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';

first_success
```

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

First_Success
01/08/2018

- The MIN(DATE) function was used to find the date of the first successful landing outcome on ground pad
- The WHERE clause ensured that the results were filtered to match only when the *'landing_*outcome' column is 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

```
success_boosters_with_payload_btn_4k_6k = %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (

success_boosters_with_payload_btn_4k_6k
```

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

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

- The BETWEEN clause was used to retrieve only those results of payload mass greater than 4000 but less than 6000. The WHERE clause filtered the results to include only boosters which successfully landed on drone ship

Total Number of Successful and Failure Mission Outcomes

```
total_success_failure = %sql SELECT "Mission_Outcome" AS 'Mission_Outcome', COUNT(*) AS 'Total_Outcome' FROM SPACEXTBL GROUP

total_success_failure
```

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

Mission_Outcome	Total_Outcome
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- The COUNT() function is used to count the number of occurences of different mission outcomes with the help of the GROUPBY clause applied to the '*mission_outcome*' column. A list of the total number of successful and failure mission outcomes os returned.
- There have been 99 successful mission outcomes out of 98 missions.

Boosters Carried Maximum Payload

```
booster_max_payload_result = %sql SELECT "Booster_Version" AS "Booster_Version_Max_Payload" FROM SPACEXTBL WHERE "PAYLOAD_MA

booster_max_payload_result
```

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

Booster_Version_Max_Payload
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

- The MAX() function was used in a subquery to retrieve a list of boosters which have carried the maximum payload mass

2015 Launch Records

```
records_2015_result = %sql SELECT CASE WHEN SUBSTR("Date", 4, 2) = '01' THEN 'January' WHEN SUBSTR("Date", 4, 2) = '02' THEN

# Get the result as a DataFrame
records_2015_result
```

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

Month_Name	Failure_Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- The SELECT statement was used to retrieve multiple columns from the table. The YEAR(DATE) function was used to retrieve only those rows with a 2015 launch date.

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

```
landing_outcome_rank_result = %sql SELECT "Landing_Outcome" AS "Landing_Outcome", COUNT(*) AS "Count_Landing_Outcome" FROM S
landing_outcome_rank_result
```

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

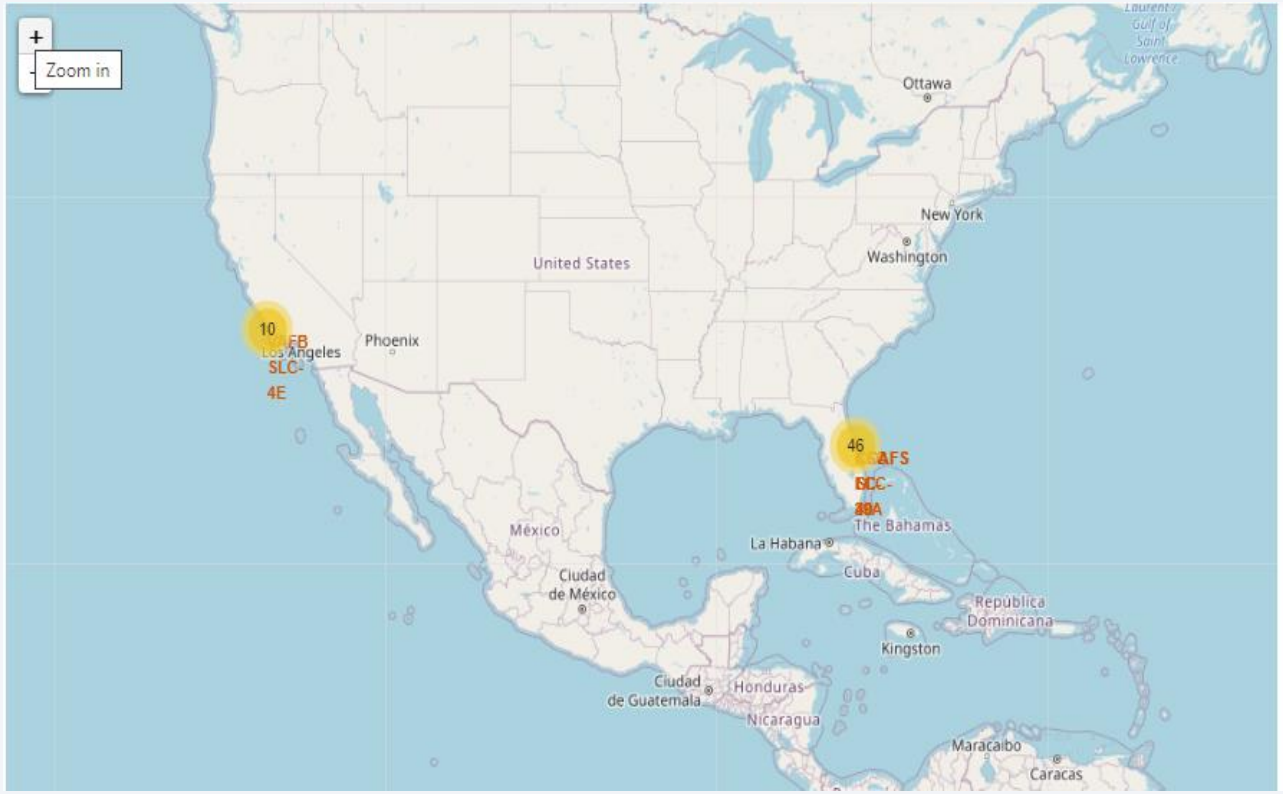
Landing_Outcome	Count_Landing_Outcome
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

- COUNT() function was used to count the different *landing outcomes.* The WHERE and BETWEEN clauses filtered the results to only include results between 2010-06-04 and 2017-03-20. The GROUPBY clause ensure that the counts were grouped by their outcome.

Section 3

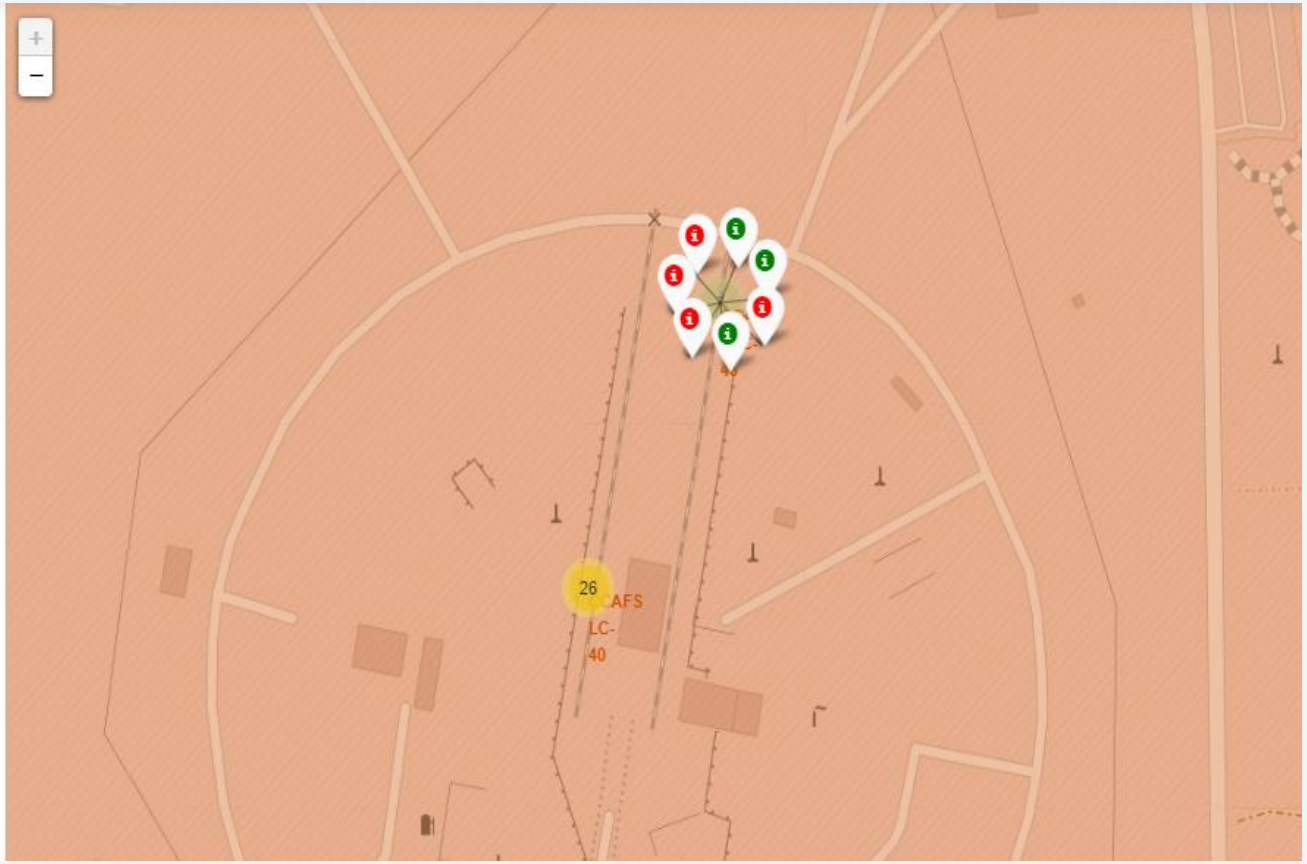
Launch Sites Proximities Analysis

SpaceX Launch Sites



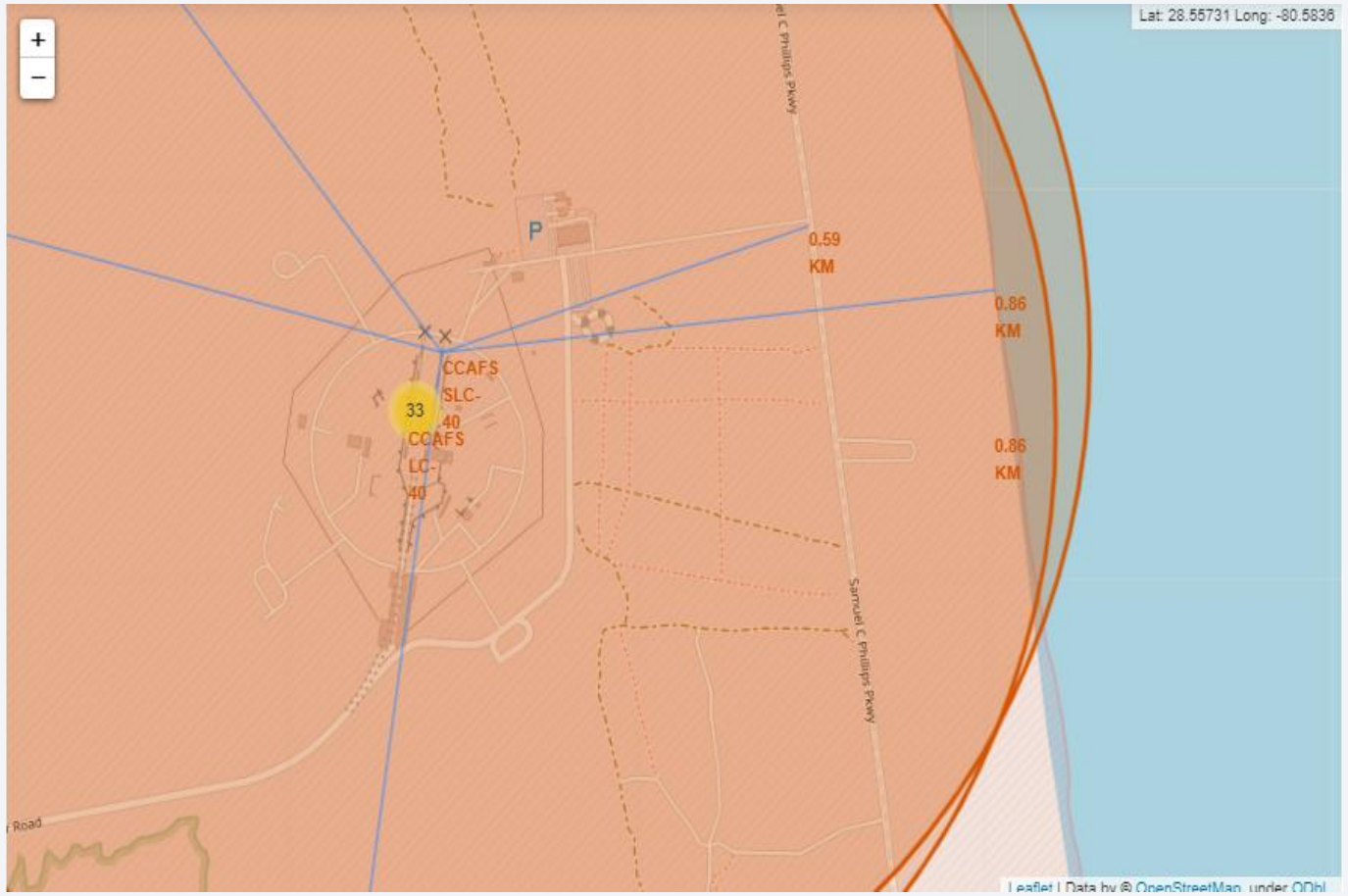
- The yellow markers are indicators of where the locations of all the SpaceX launch sites are situated in the US.
- The launch sites have been strategically placed near the coast

Label of Success or Failure

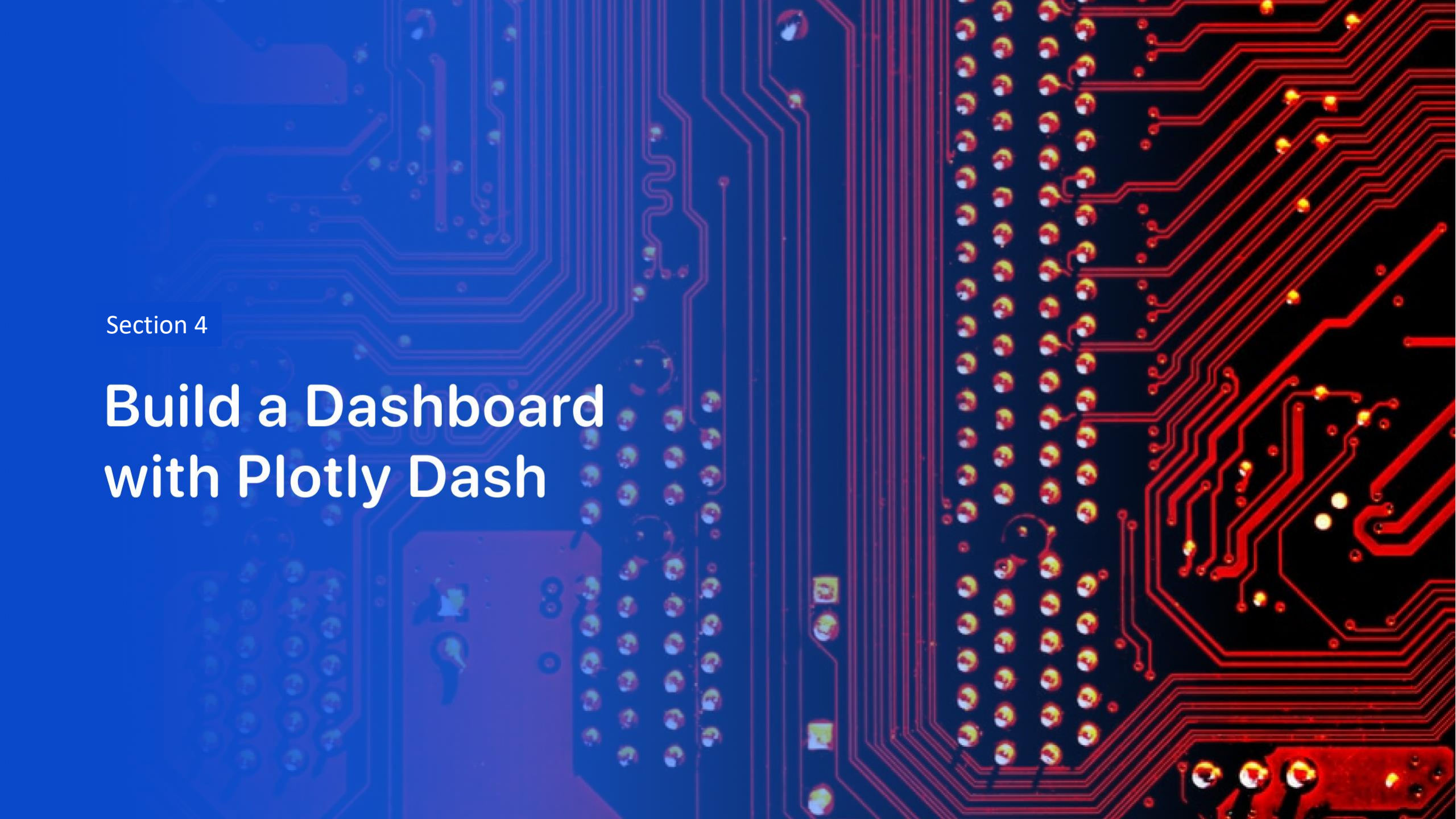


- Clicking on the launch site which will display marker clusters of successful landings (green) or failed landing (red).

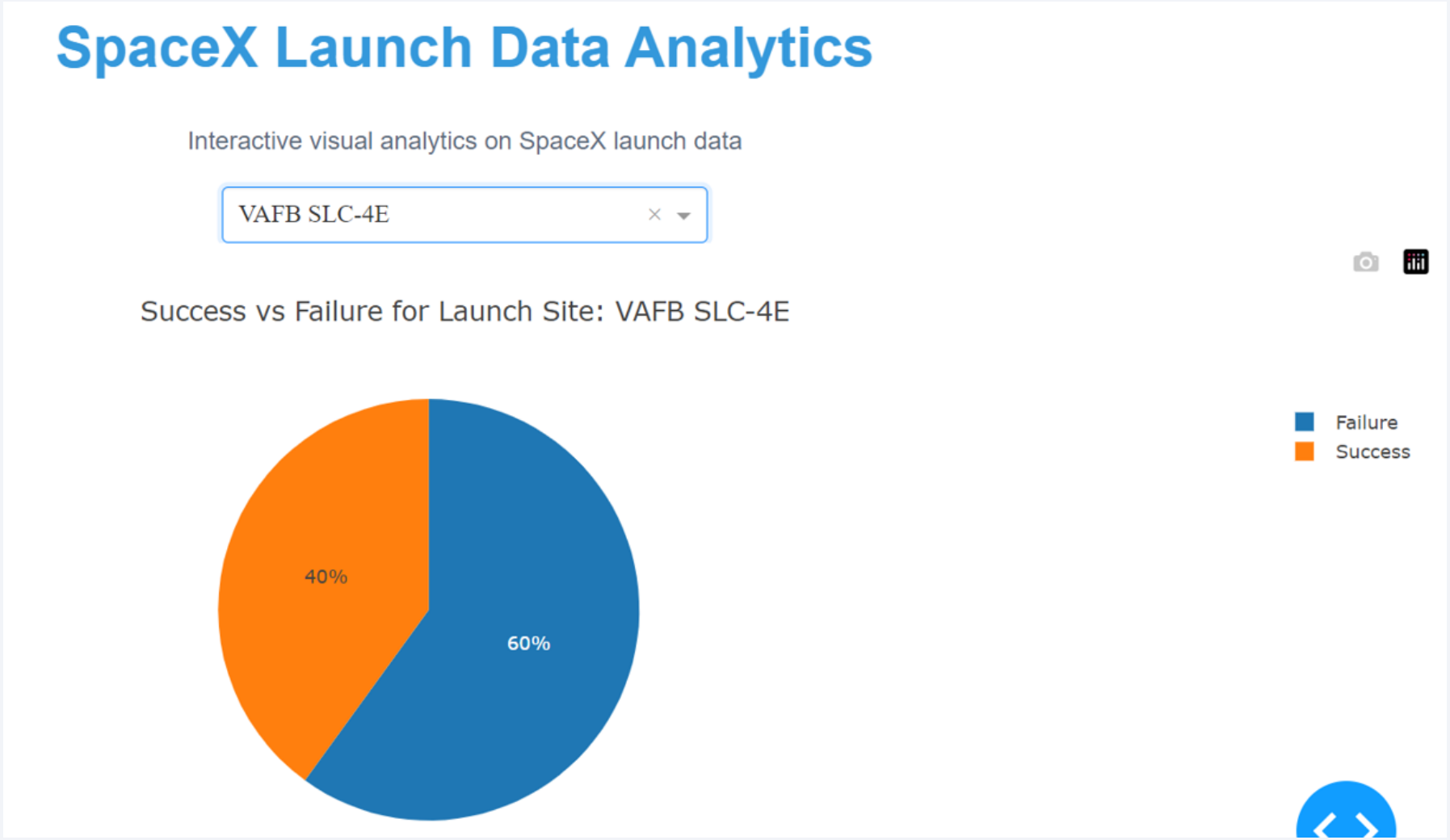
Launch Site Proximities



The generated map shows that the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.


Section 4
Build a Dashboard with Plotly Dash

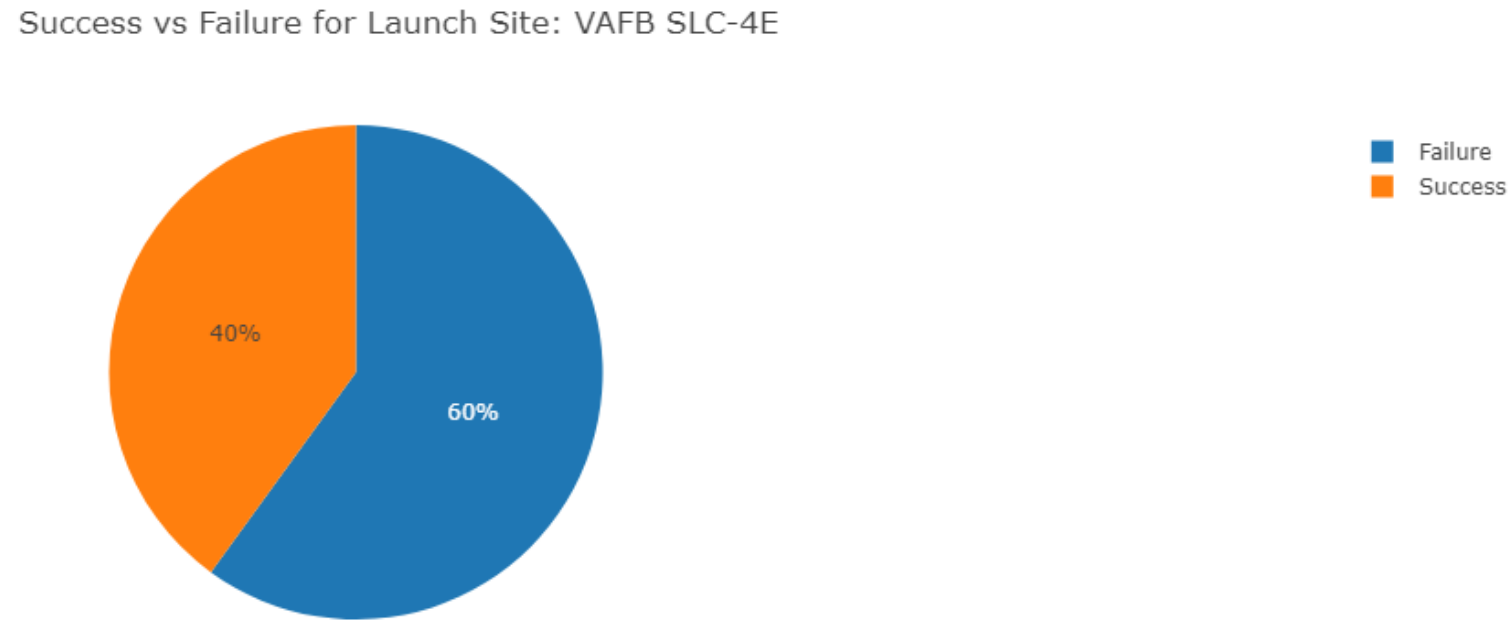
Total Successful Launch by Site



Payloads vs Launch Outcome

- The launch success rate for payloads 0-2500 kg is slightly lower than that of payloads 2500-5000 kg. There is in fact not much difference between the two.
- The booster version that has the largest success rate, in both weight ranges is the *v1.1*.

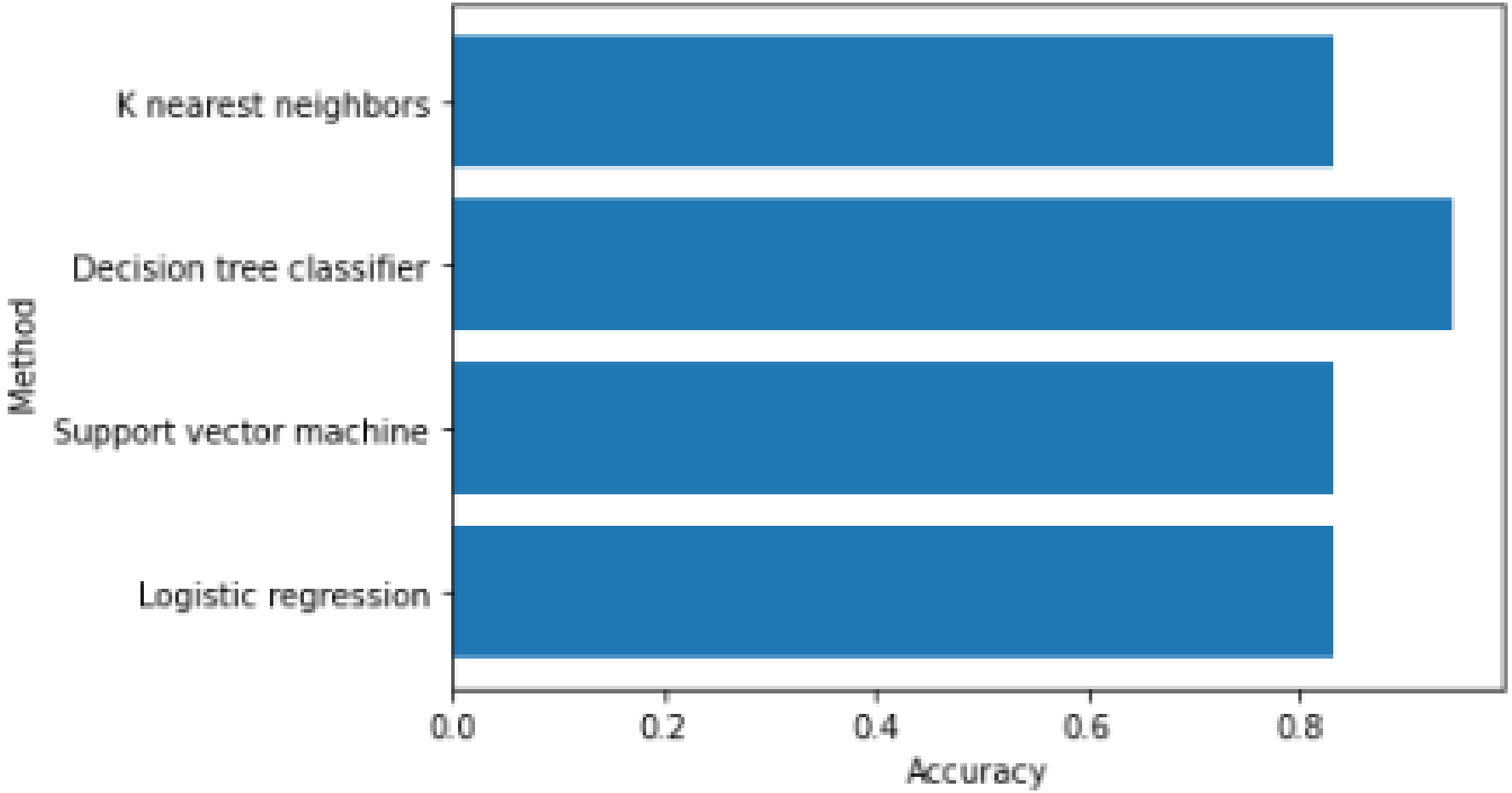
Dashboard


Success vs Failure for Launch Site: VAFB SLC-4E
40%
60%
Failure
Success

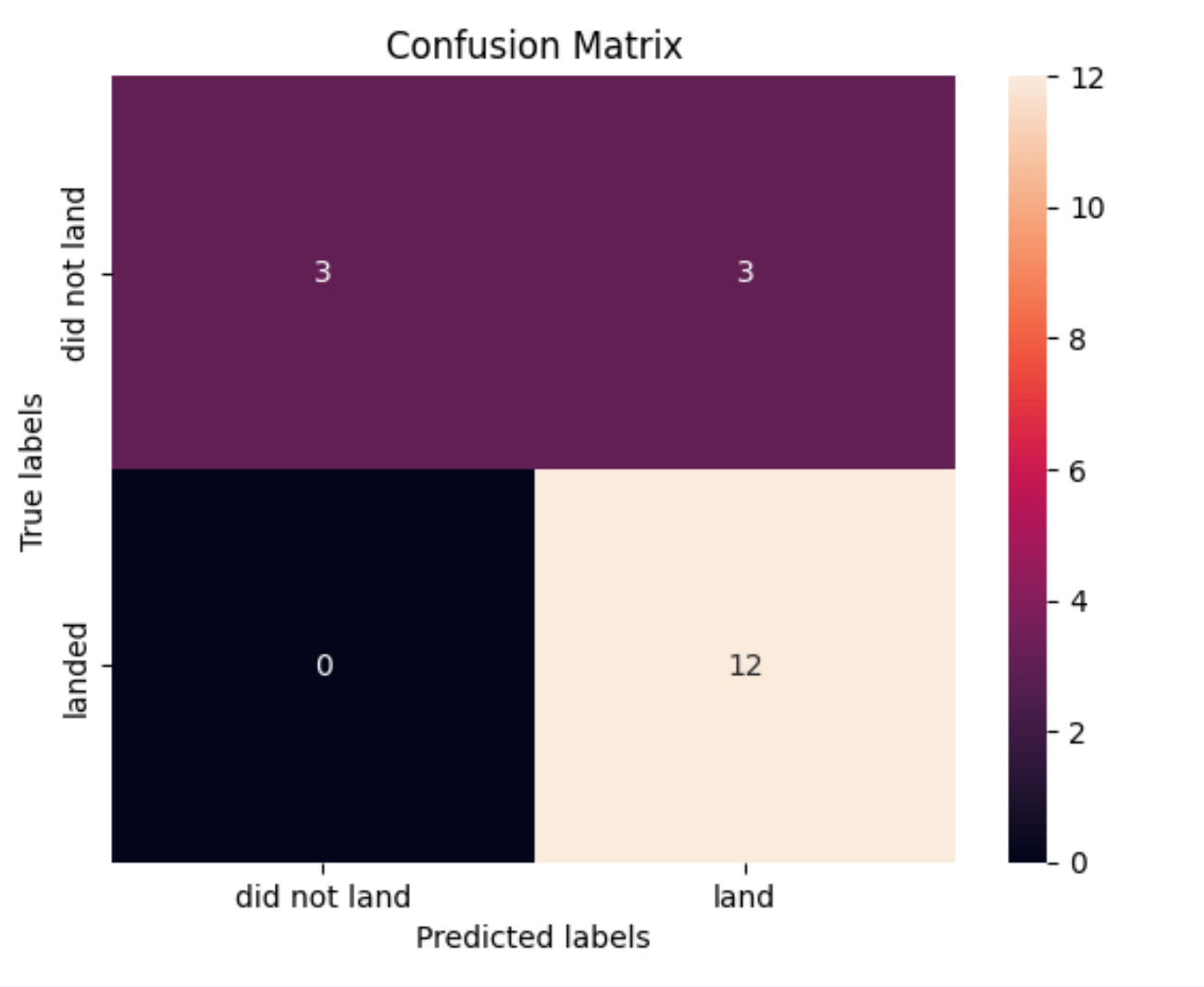
Section 5

Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix



- The model predicted 12 successful landings when the True label was successful (True Positive) and 3 unsuccessful landings when the True label was failure (True Negative).

Conclusions

- The analysis showed that there is a positive correlation between number of flights and success rate as the success rate has improved over the years.
- There are certain orbits like SSO, HEO, GEO, and ES-L1 where launches were the most successful.
- Success rate can be linked to payload mass as the lighter payloads generally proved to be more successful than the heavier payloads.
- The launch sites are strategically located near highways and railways for transportation of personnel and cargo, but also far away from cities for safety.
- The best predictive model to use for this dataset is the Decision Tree Classifier as it had the highest accuracy with 89%.

Appendix

GitHub Repository: The repository shows all the jupyter notebooks and the plotly dash app.

Thank you!