Winning Space Race with Data Science

Elaheh Sarshar
04 Oct 2023

Outline

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

Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA with Visualization
- EDA with SQL
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis

Summary of all results

- EDA Results
- Interactive Analytics
- Predictive Analysis

Introduction

Predicting SpaceX Falcon 9 Launch Costs

Context:

- SpaceX's Competitive Edge is that Falcon 9 launches are advertised at $62 million, a stark contrast to competitors' >$165 million.

Significance:

- A major factor in this cost difference is SpaceX's ability to reuse the Falcon 9's first stage, a feature not common with other providers.

Objective of Our Study:

- To predict the landing success of Falcon 9's first stage using historical launch data.

Approach:

- We aim to harness historical launch data and employ a machine learning model to make these predictions.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Utilizing SpaceX's official REST API
- Extracting data via Wikipedia web scraping

Perform data wrangling:

- Cleaning tasks include filling missing payload masses with average values and omitting non-essential columns
- Establishing a label for landing outcomes
- Adjusting data: Applying One Hot Encoding for categorical data and standardizing numerical attributes

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models:

- Implemented models: Linear Regression, K Nearest Neighbors, Support Vector Machine, and Decision Tree
- Evaluation to identify the optimal classifier

Data Collection

Key Phrases:

Data Sources Identification: Determine the origins of data collection.

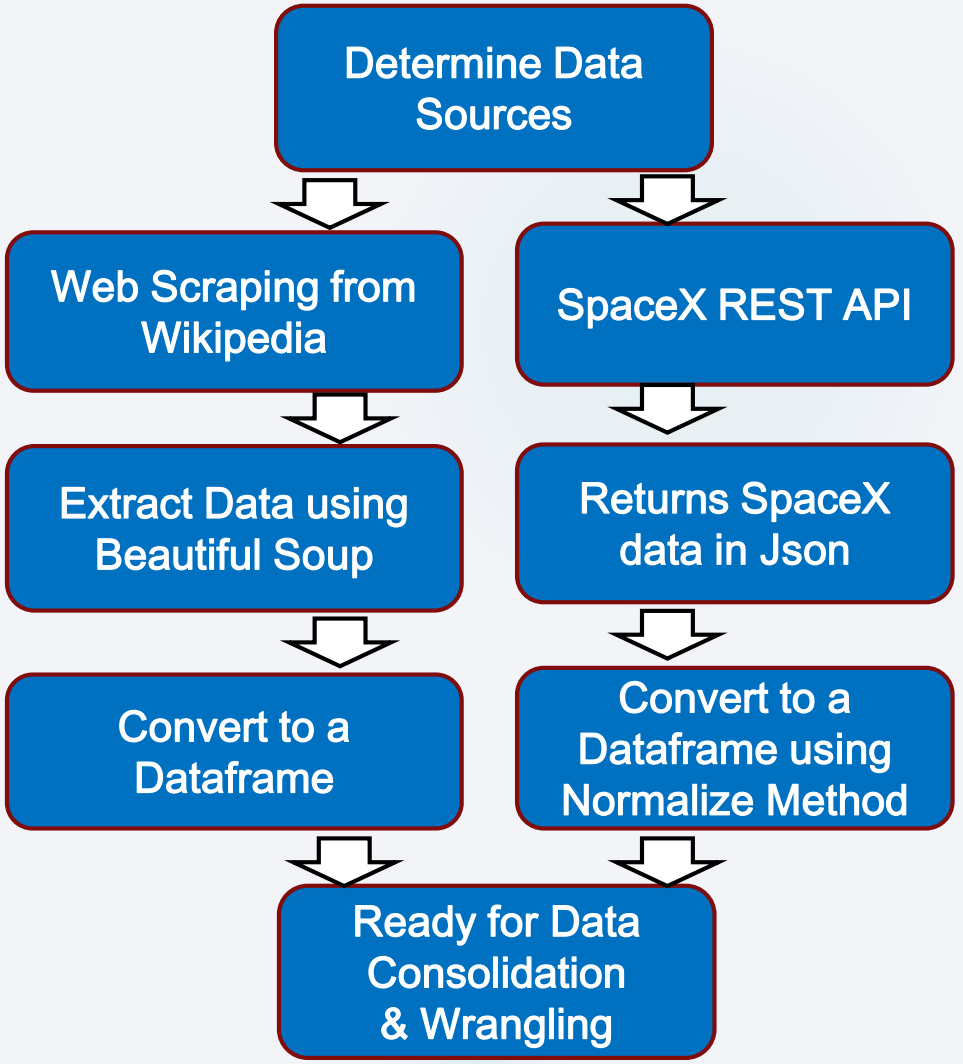
SpaceX REST API Access: Retrieve launch details from “https://api.spacexdata.com/v4/launches/past”.

JSON to DataFrame: Process the JSON data from the REST API and convert it into a pandas dataframe using json_normalize.

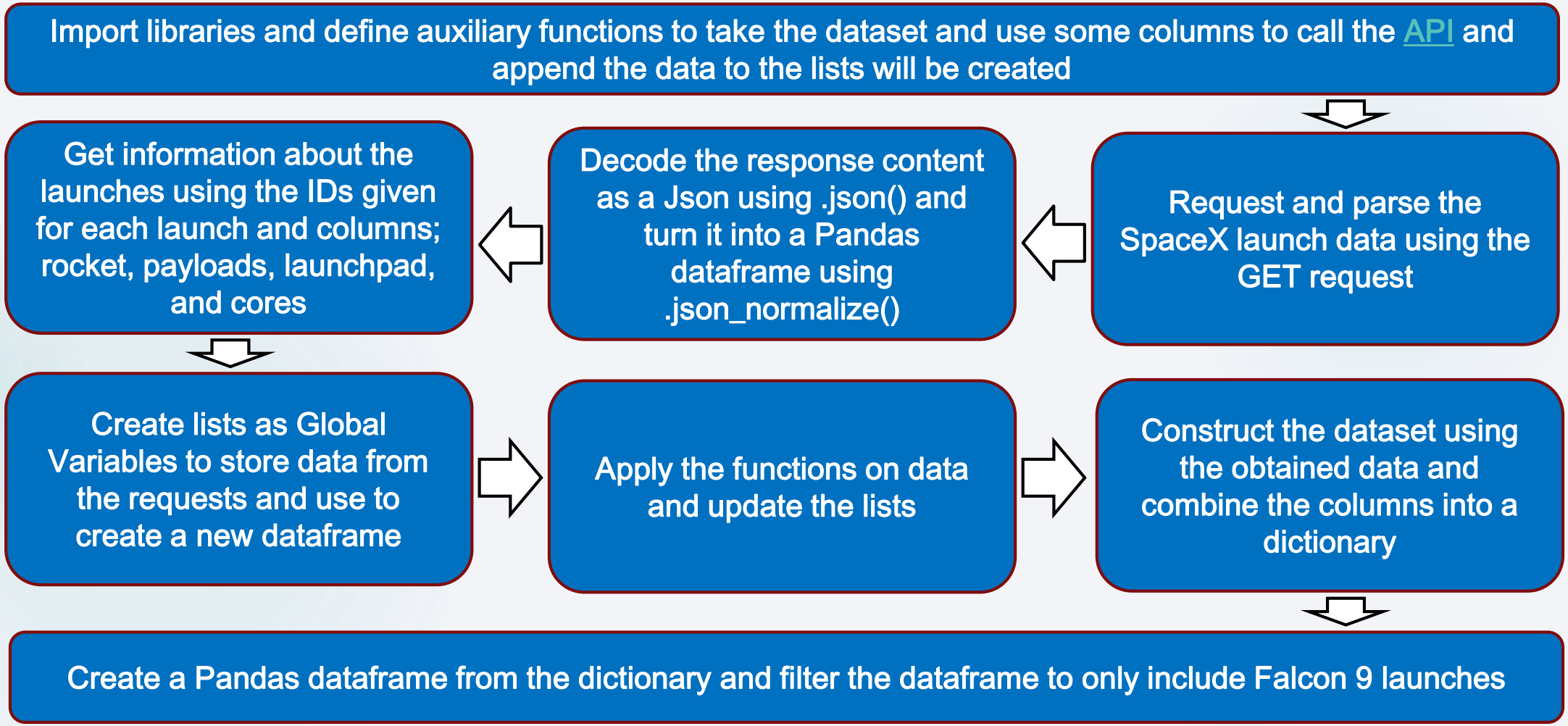
:**Wikipedia Webscraping:** Scrape SpaceX launch data from the designated Wikipedia page.

HTML Parsing with BeautifulSoup: Extract launch records from the Wikipedia table using BeautifulSoup.

Data Preparation: Each data source (REST API and Wikipedia) is individually processed and transformed into a structured dataframe format, ready for the data wrangling phase.

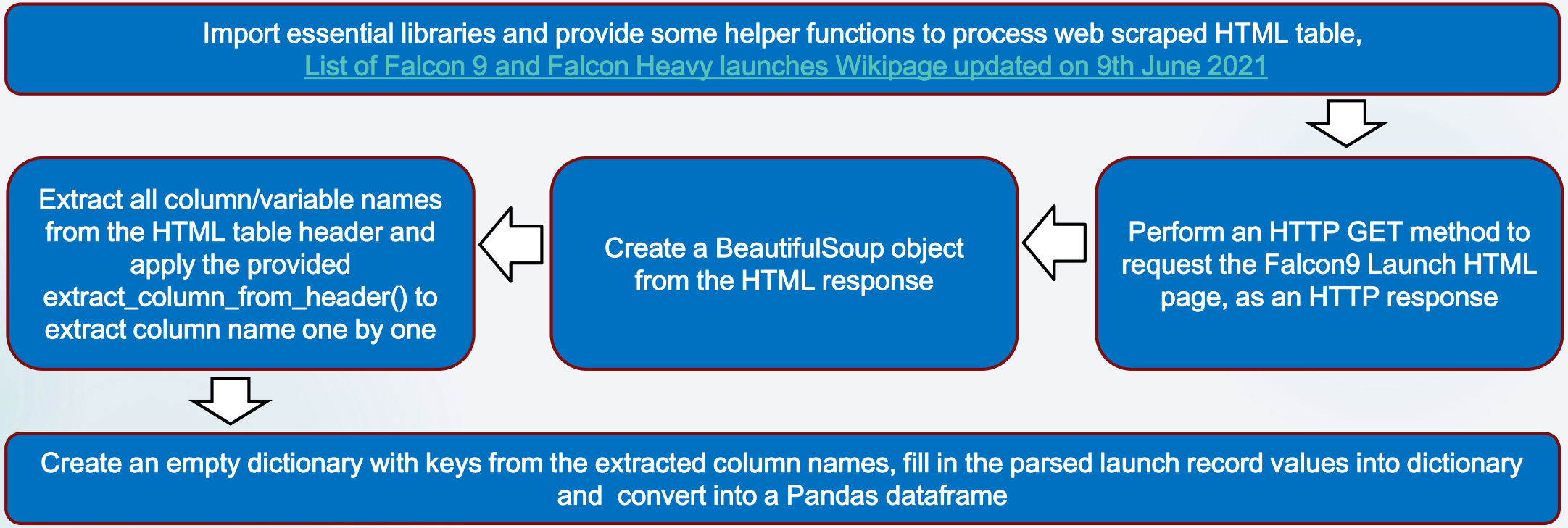


Data Collection – SpaceX API



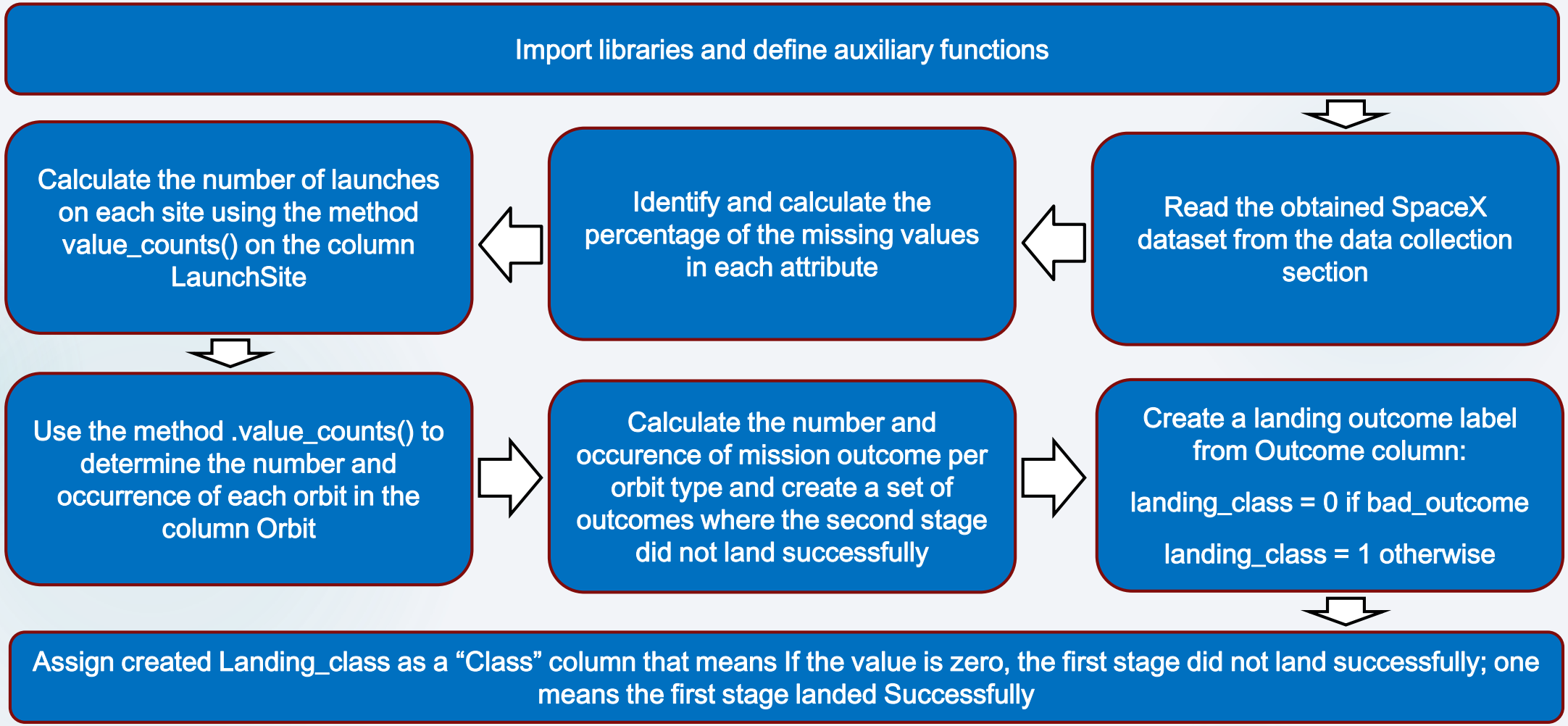
For a detailed review, refer to the GitHub notebook

Data Collection - Scraping



For a detailed review, refer to the GitHub notebook

Data Wrangling



For a detailed review, refer to the GitHub notebook

EDA with Data Visualization

To explore data, scatterplots, bar plots, were used to visualize the relationship between pair of features.

Scatter plots of FlightNumber vs PayloadMass, Flight number vs launch site, PayloadMass vs Launch site, Flight number vs Orbit type and Payload vs Orbit type, help to find out how much the options like increasing flight numbers, payload heaviness, launch sites and orbit types would affect the successful landing.

Bar chart of “Success Rate by Orbit Type” visualize if there are any relationship between success rate and orbit type.

Line charts are used to show and track changes in data over a specified period of time. The year trend of the average launch success rate was shown through a line chart

For a detailed review, refer to the GitHub notebook

EDA with SQL

SQL Queries Executed:

- Display the unique launch sites used in space missions.
- Show 5 records where launch sites start with the string 'KSC'.
- Calculate the total payload mass for boosters launched by NASA (CRS).
- Determine the average payload mass for the booster version F9 v1.1.
- Identify the date when a successful landing on a drone ship occurred.
- List boosters that successfully landed on a ground pad with a payload mass between 4000 and 6000.
- Count the total successful and failed mission outcomes.
- Identify booster versions that have transported the heaviest payloads.
- Display records of month names, successful ground pad landings, and launch sites for 2017.
- Rank successful landings between 2010-06-04 and 2017-03-20 in descending order.

For a detailed review, refer to the GitHub notebook

Build an Interactive Map with Folium

Map Objects in Folium: A Summary

Markers:

What? Placed markers on each launch site.
Why? To distinctly identify and locate each launch site on the map, providing a quick reference point for viewers.

Circles (or Circle Markers):

What? Used circles to represent launch outcomes, with different colors indicating success or failure.
Why? Circles provide a visual representation of performance, allowing for an immediate understanding of how each site has fared over time.

Lines (or Polylines):

What? Drew lines connecting launch sites to nearby infrastructures like railways, highways, and coastlines. Additionally, used Polyline to trace paths or routes on the map.
Why? Lines and polylines help in understanding the geographical context, proximity, and potential routes, highlighting logistical advantages or challenges for each launch site.

MousePosition:

What? Integrated the MousePosition plugin to display the latitude and longitude of the mouse pointer.
Why? To provide real-time geographic coordinates as users navigate the map, enhancing interactivity and aiding in precise location identification.

For a detailed review, refer to the GitHub notebook

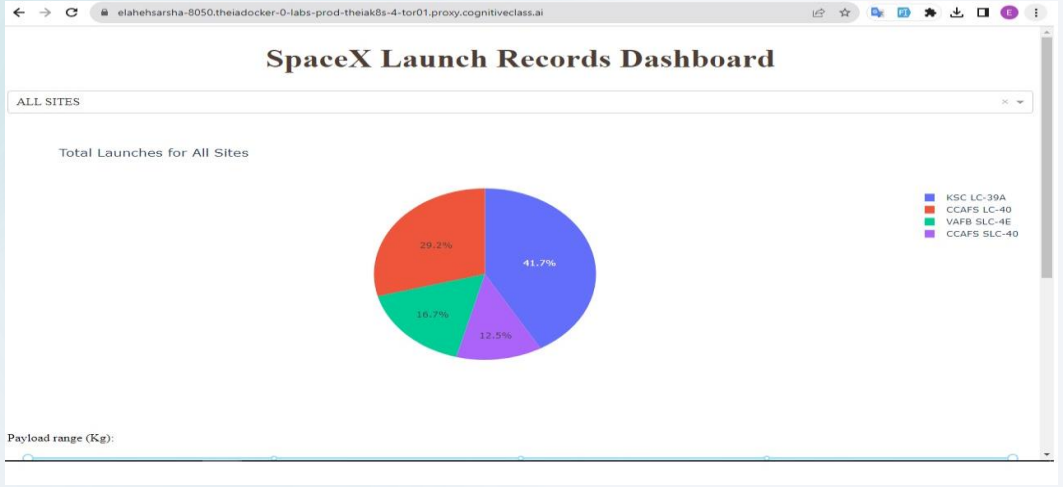
Build a Dashboard with Plotly Dash

Real-time Analytics with Plotly Dash

The Plotly Dash applications enable real-time interactive visual analytics on SpaceX launch data. Our dashboard has been developed with the following features:

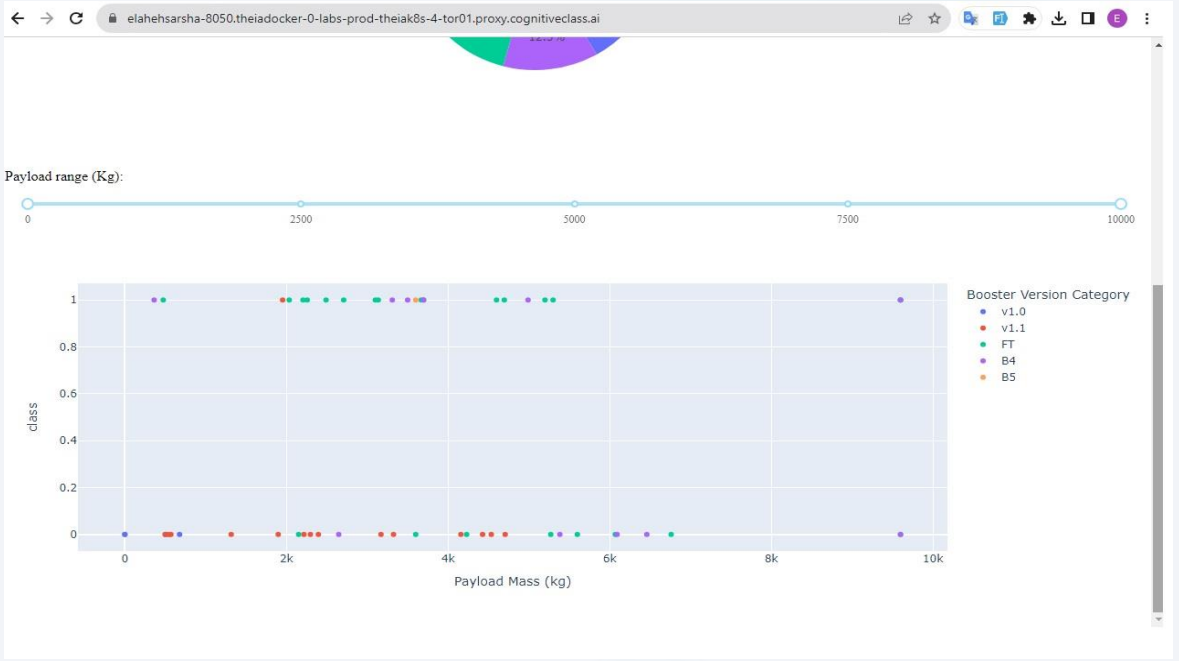
Pie Chart:

Provides a visual representation of launch successes and failures, specifically based on the selected launch site. It offers an intuitive overview of performance metrics for each site.



Scatter Plot:

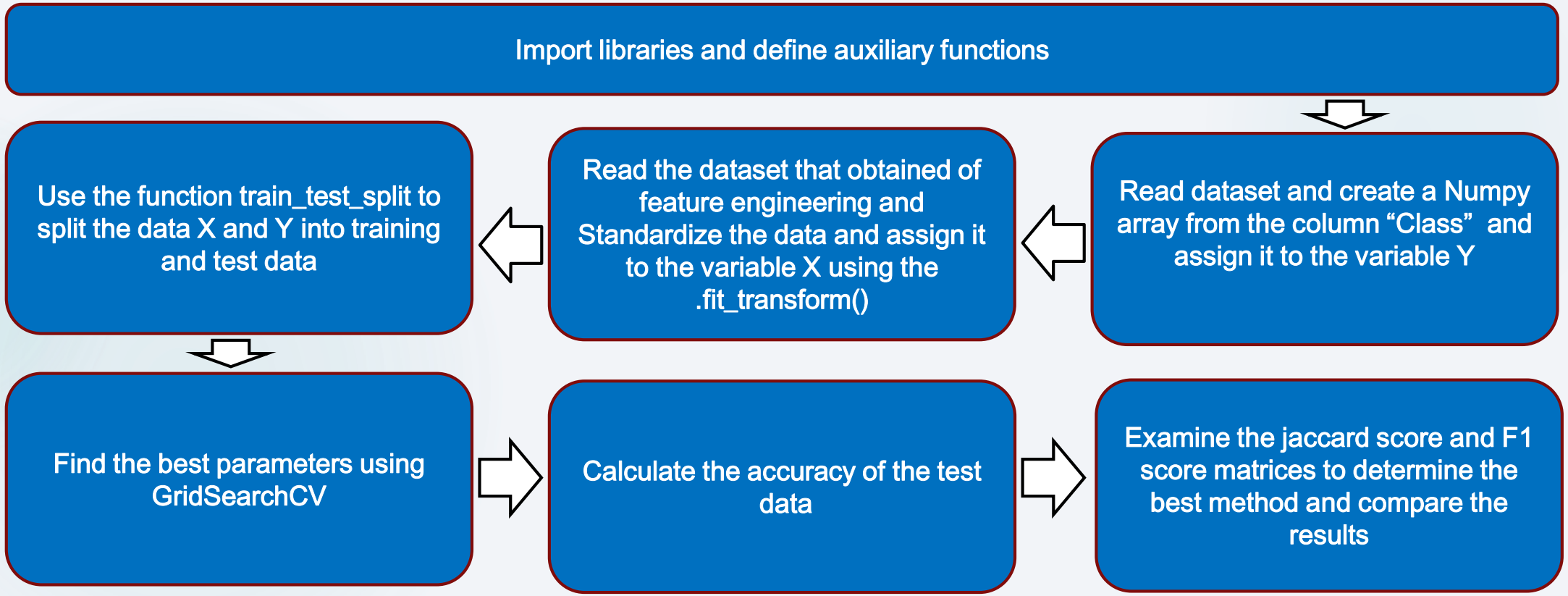
Showcases the correlation between launch success and payload mass.
Includes an integrated Range Slider that allows users to adjust the payload range, providing insights into specific payload segments and their impact on launch outcomes.



For a detailed review, refer to the GitHub notebook

Predictive Analysis (Classification)

Four classification models including Linear Regression, K Nearest Neighbors, Decision Tree models, and Support Vector Machine, were built and compared.

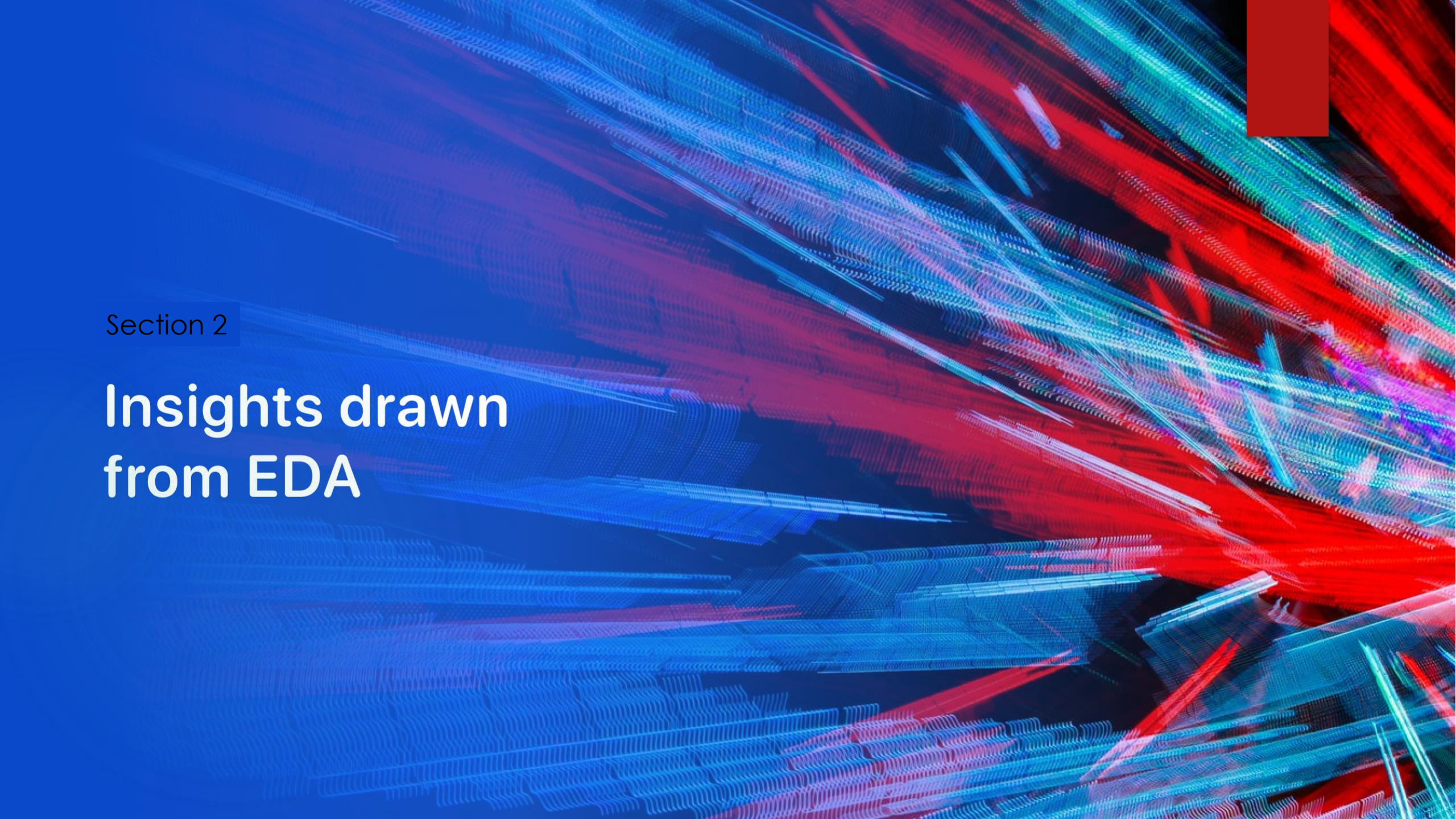


For a detailed review, refer to the GitHub notebook

Results

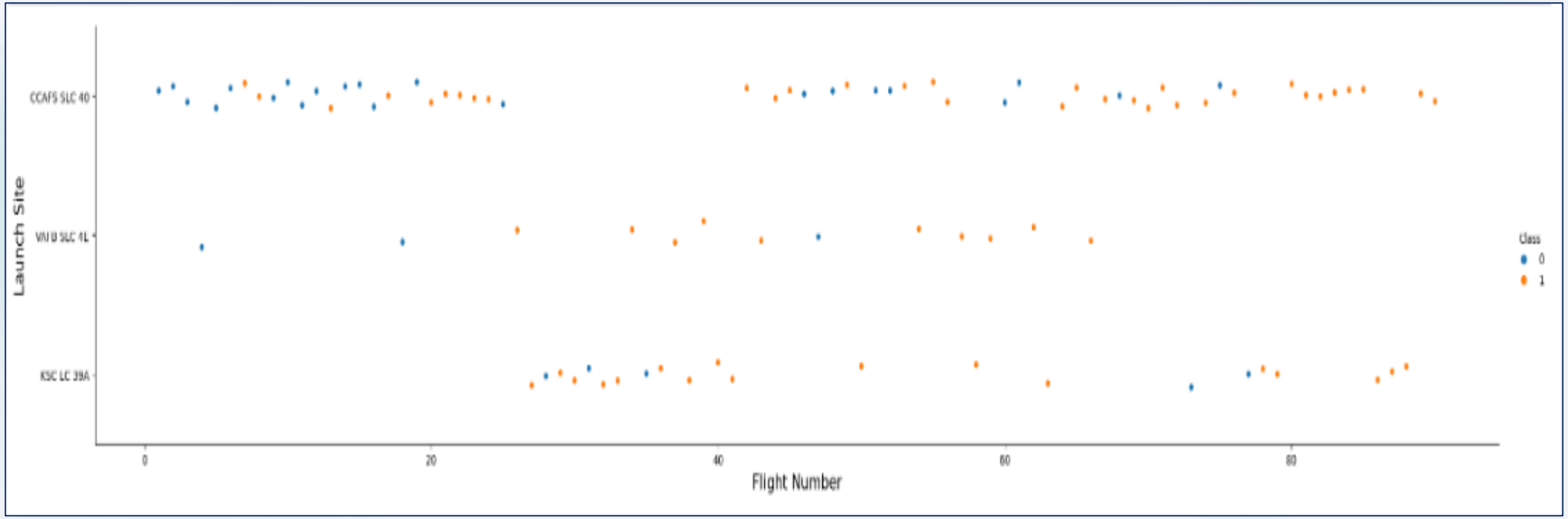
Exploratory Data Analysis & Predictive Analysis Insights

1. **Ground Pad Landing Milestone:** The inaugural successful ground pad landing was achieved on December 22, 2015.
2. **Trend in Success Rates:** From 2013 to 2020, SpaceX launches exhibited a consistent upward trajectory in success rates.
3. **Significance of Launch Location:** The choice of launch location has emerged as a crucial determinant in the success of missions.
4. **Top Performing Launch Site:** KSC LC-39A has distinguished itself by recording the highest number of successful launches among all sites.
5. **Payload Weight vs. Success:** Lighter payloads have demonstrated a superior success rate in comparison to heavier ones.
6. **Orbits with High Success Rates:** ES-L1, GEO, HEO, and SSO orbits have consistently achieved commendable success rates.
7. **Heavy Payloads & Orbit Preference:** For bulkier payloads, the success rates are notably higher in VLO and ISS orbits.
8. **Predictive Analysis Outcome:** The Decision Tree model has proven to be the most accurate in predicting launch outcomes for the given dataset.


Section 2
Insights drawn from EDA

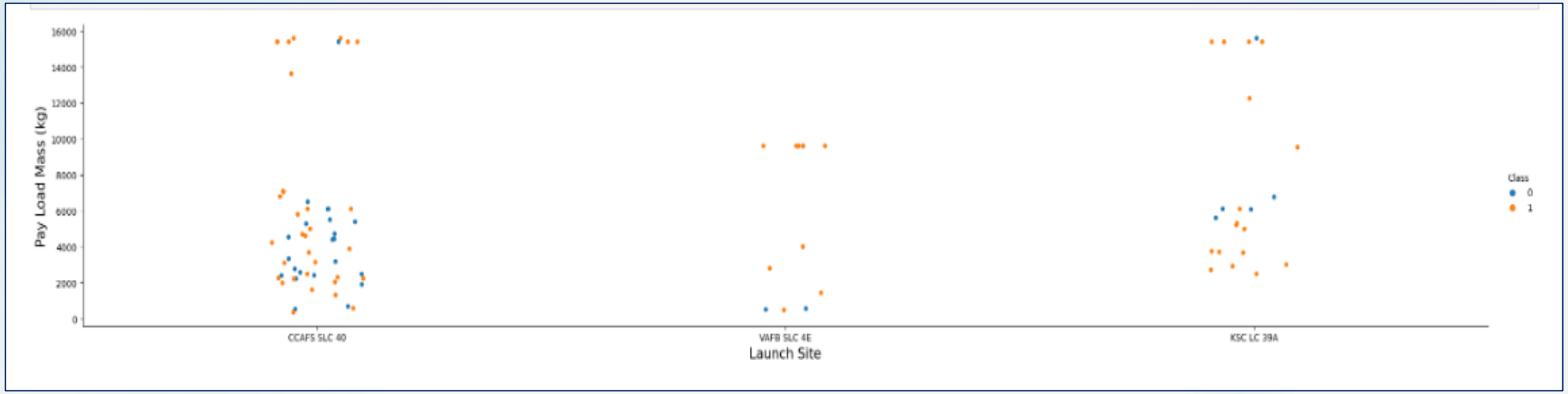
Flight Number vs. Launch Site

The majority of launches originated from the CCAFS SLC 40 launch site.
Initial space missions predominantly utilized the CCAFS SLC 40 launch site.
Recent launches from CCAFS SLC 40 have seen a significant uptick in success.
Over time, there has been a consistent improvement in success rates across all launch sites.



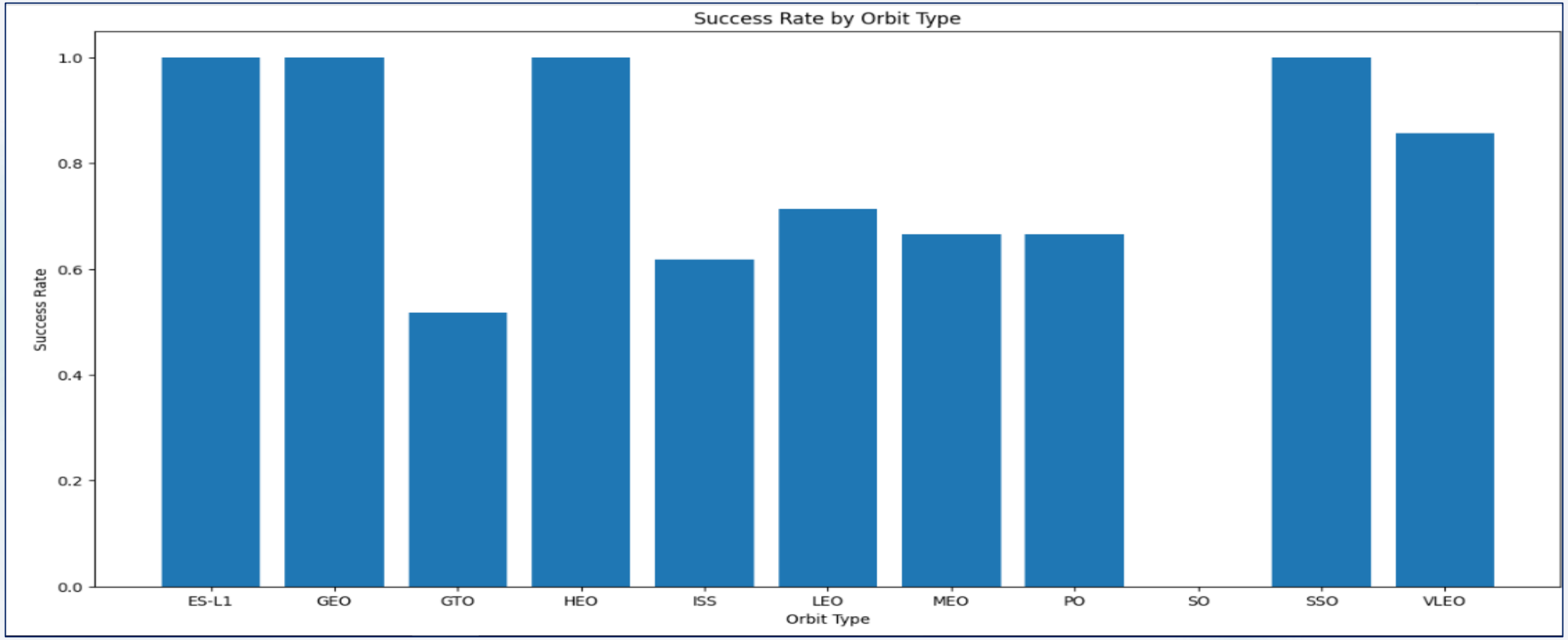
Payload vs. Launch Site

- The VAFB SLC 4E launch site predominantly manages launches with lighter payloads, notably with no recorded launches exceeding 10,000 kg.
- CCAFS SLC 40 stands out for its adaptability, hosting a diverse range of launches that encompass both heavier and lighter payload weights.
- Launches with payloads tipping the scales at over 9,000 kg have predominantly seen successful outcomes.



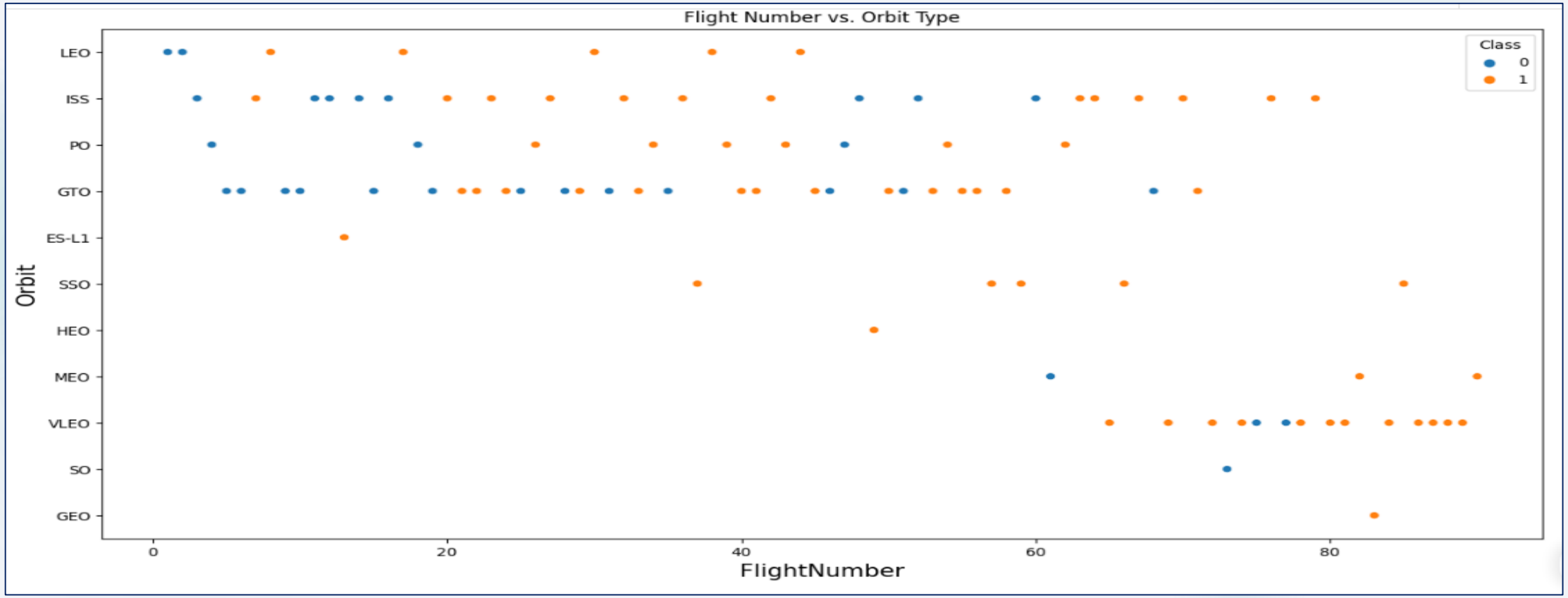
Success Rate vs. Orbit Type

Orbits such as ES-L1, GEO, HEO, and SSO have achieved a flawless success rate of 100%. Following closely, VLEO orbits have a commendable success rate of over 80%, and LEO orbits maintain a success rate of more than 70%.



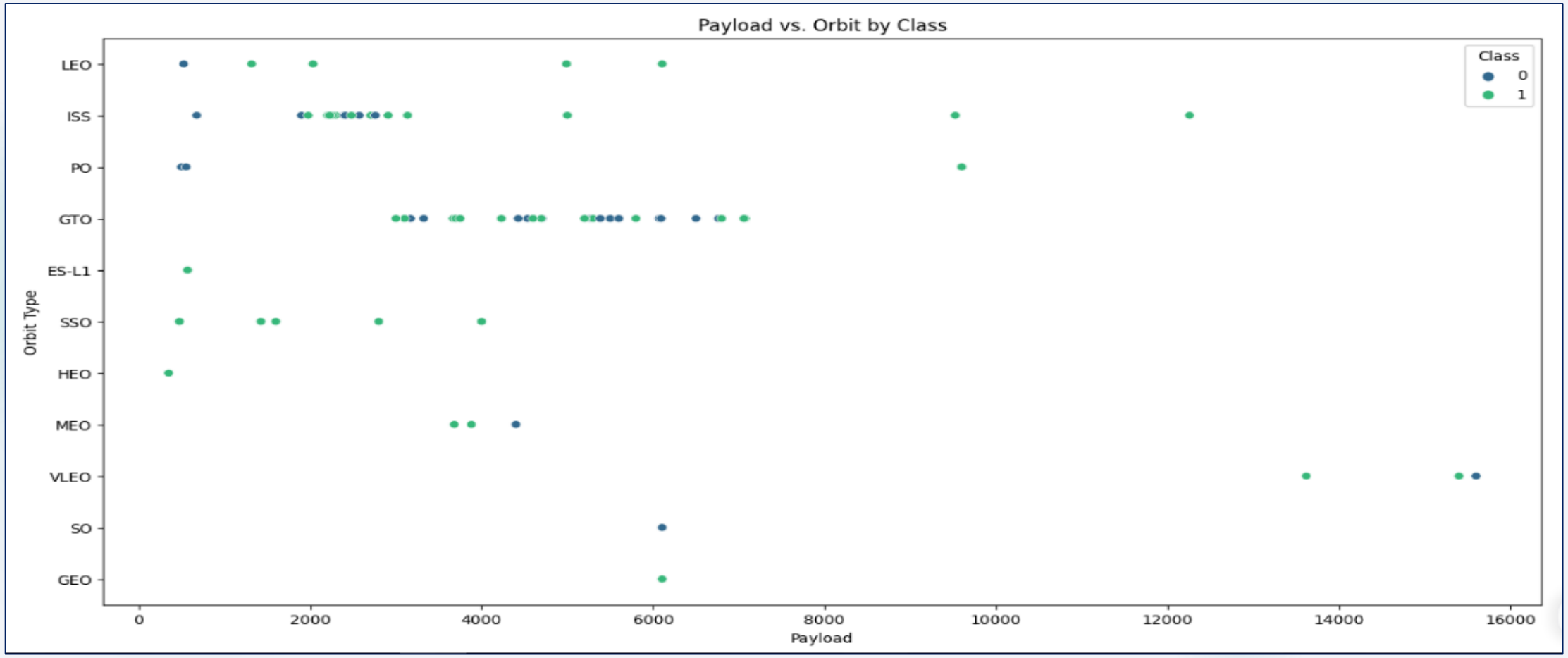
Flight Number vs. Orbit Type

In recent years, there's been a marked shift towards missions targeting Very Low Earth Orbits (VLEO), which have consistently achieved high success rates. Conversely, while the GTO orbit registers a lower success rate, it's evident that the flight number doesn't directly influence the success rate within this orbit.



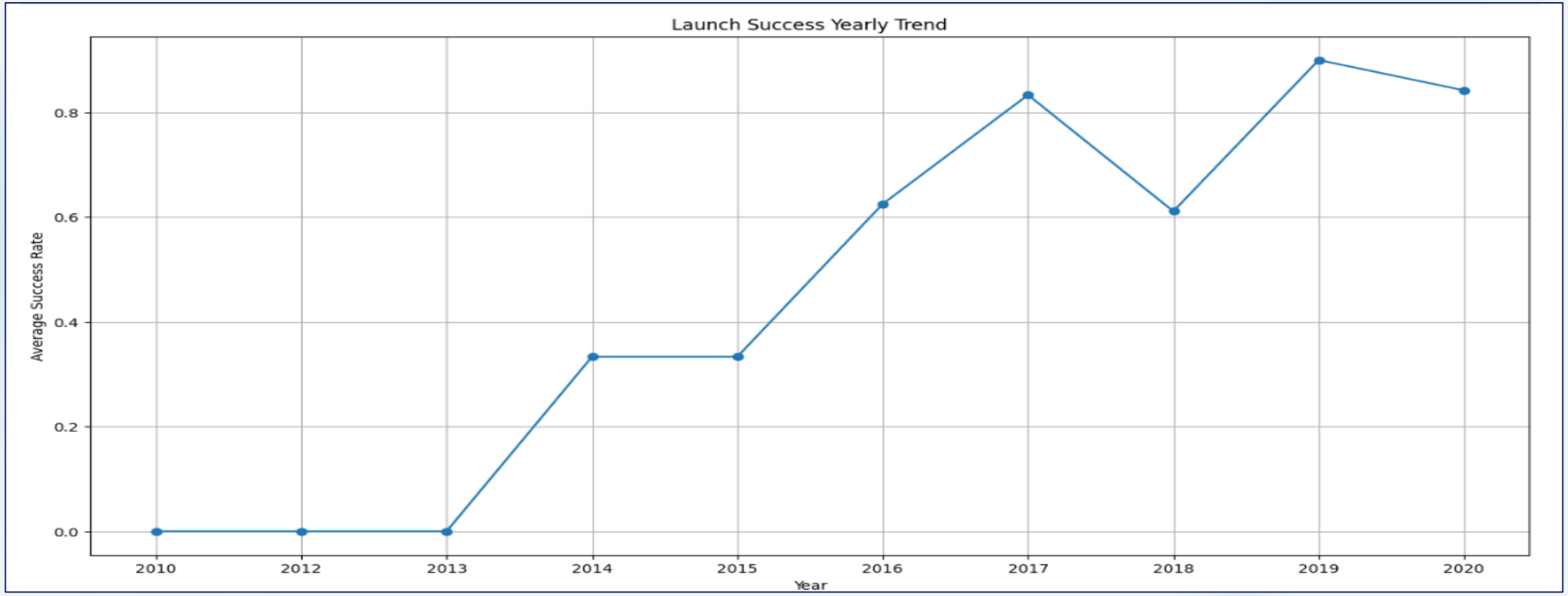
Payload vs. Orbit Type

For heavy payloads, the rates of success are higher for VLO and ISS orbits. In the case of GTO, there seems no apparent relationship between payload and the rate of success.



Launch Success Yearly Trend

- The rate of success has seen a notable rise since 2013 and continued to rise until 2020, possibly attributed to technological advancements.
- The first three years (2010-2013) seem to have been a phase focused on fine-tuning and technological enhancement.



All Launch Site Names

The “distinct” function allows to select and list only the unique sites without repetition.

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

```
In [8]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

The “limit” enables to display only 5 names.

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

```
In [9]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Out[9]:

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

Total Payload Mass

The “sum” function sums up the selected data. The total payload carried by boosters from NASA (CRS) is 45596 (kg).

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

```
In [10]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_payload_mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

Out[10]:

Total_payload_mass
45596

Average Payload Mass by F9 v1.1

The “avg” function calculates the average of the selected data. The average payload mass carried by booster version F9 v1.1 is 2928.4 (kg).

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

```
In [11]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

Out[11]:

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

The dates will be ordered hence the need for the “min” function. This function will select the lowest (earliest) date. The first successful ground pad landing took place on December 22, 2015.

```
In [12]: %sql SELECT MIN(DATE) AS First_successful_landing_gound FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';
 * sqlite:///my_data1.db
Done.
Out[12]: First_successful_landing_gound
---------------------------------------
                              2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

There are 2 conditions specified after the “where” clause. The first is for mass greater than 4000, the second is for mass less than 6000. The use of the “and” operator returns data from when both conditions are met or true.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [13]:
```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000;
```

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

Out[13]:

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

Total Number of Successful and Failure Mission Outcomes

The “count” function allowed for counting of each instance of a mission outcome.

List the total number of successful and failure mission outcomes

In [14]:
```
%%sql SELECT Mission_Outcome, COUNT(*) AS Total_count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

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

Out[14]:

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

Boosters Carried Maximum Payload

The “MAX” function list the boosters’ names that have carried maximum payload mass.

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

In [15]:
```
%%sql SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

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

Out[15]:

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

The unsuccessful landing outcomes on drone ships, along with their corresponding booster versions and launch site names, during the year 2015, are provided below:

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [16]:
```
%%sql SELECT
 CASE
        WHEN Date like '%-01-%' THEN 'January'
        WHEN Date like '%-02-%' THEN 'February'
        WHEN Date like '%-03-%' THEN 'March'
        WHEN Date like '%-04-%' THEN 'April'
        WHEN Date like '%-05-%' THEN 'May'
        WHEN Date like '%-06-%' THEN 'June'
        WHEN Date like '%-07-%' THEN 'July'
        WHEN Date like '%-08-%' THEN 'August'
        WHEN Date like '%-09-%' THEN 'September'
        WHEN Date like '%-10-%' THEN 'October'
        WHEN Date like '%-11-%' THEN 'November'
        WHEN Date like '%-12-%' THEN 'December'
    END AS Month,
    Landing_Outcome AS Failure_Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Date like'%2015%' AND Landing_Outcome LIKE 'Failure (drone ship)';
```

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

Out[16]:

Month	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

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

Landing outcomes listed in descending order by making use of the “Order by” function.
As shown, the highest count among the landing outcomes is attributed to "No attempt", totaling 10.
This is followed by "Success (ground pad) ", "Success (drone ship) ", and "Failure (drone ship) " each occurring 5 times, while "Controlled (ocean)" is observed 3 times.

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.

In [17]:
```
%%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

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

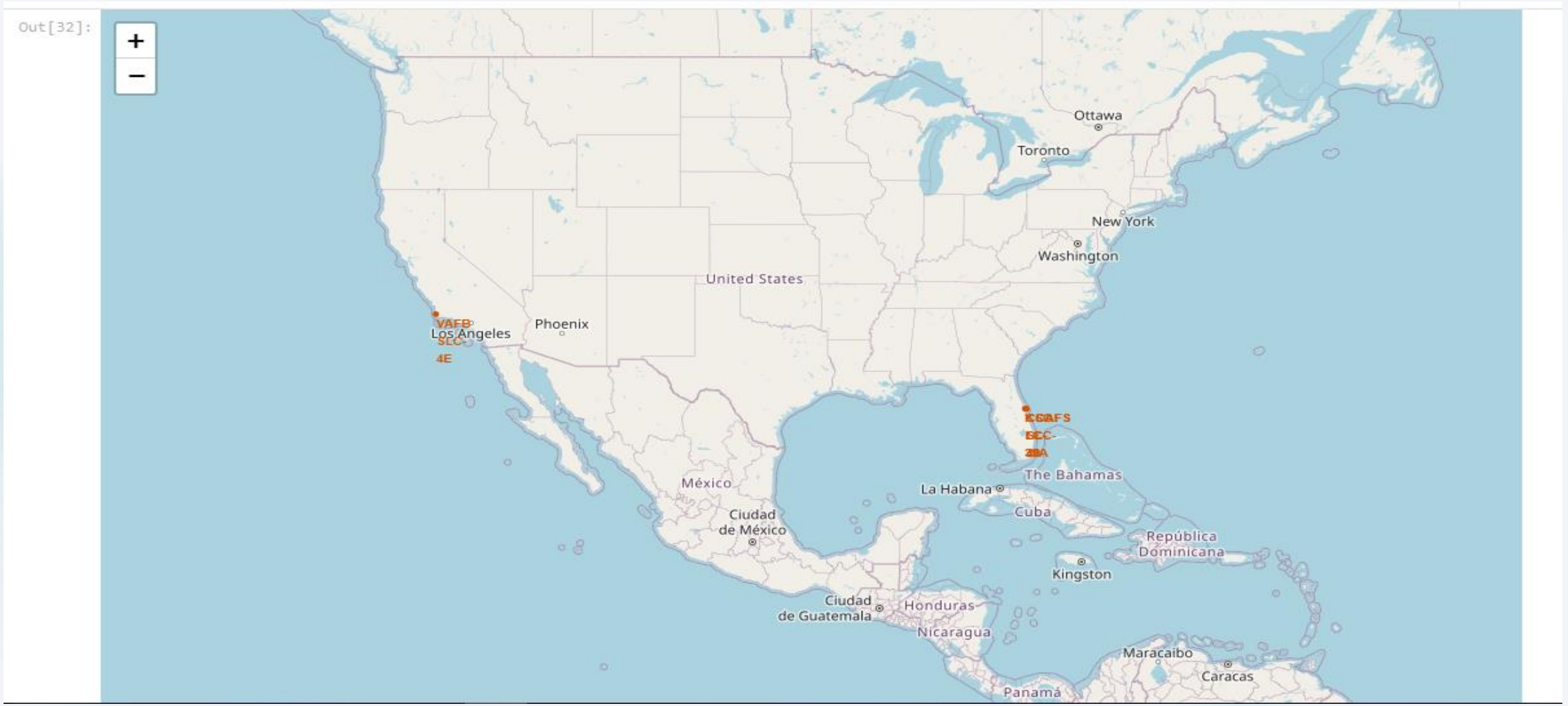
Out[17]:

Landing_Outcome	Outcome_Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

Launch Site Locations

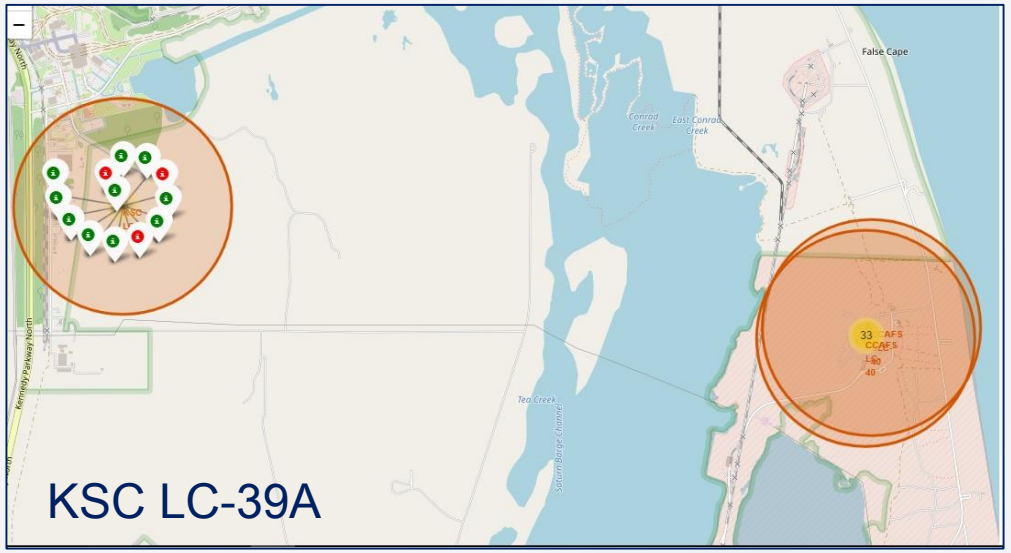
As shown in the map, All launch site locations are near a coastline and close to the sea.

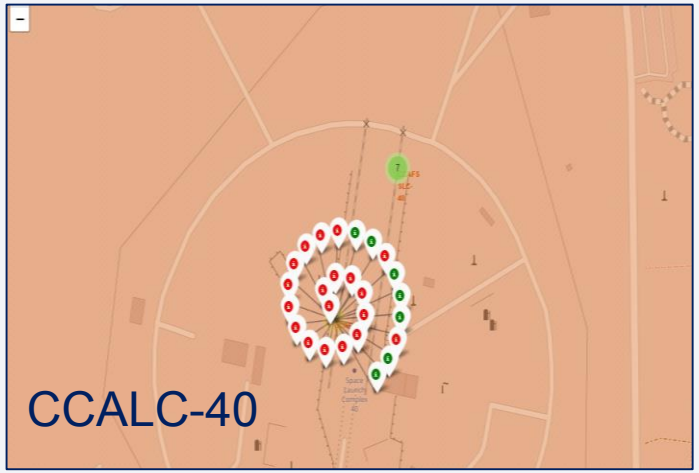


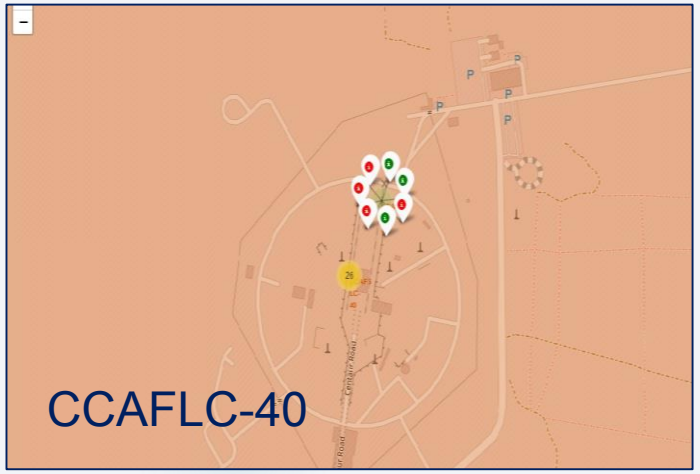
Success/Failed Launches for Launch Sites

Upon zooming in, we can see the success (green) and failure (red) marks for each site.

Out of 13 launches, KSC LC-39A has achieved the highest success rate with 10 successful missions (10/13=76.9%).

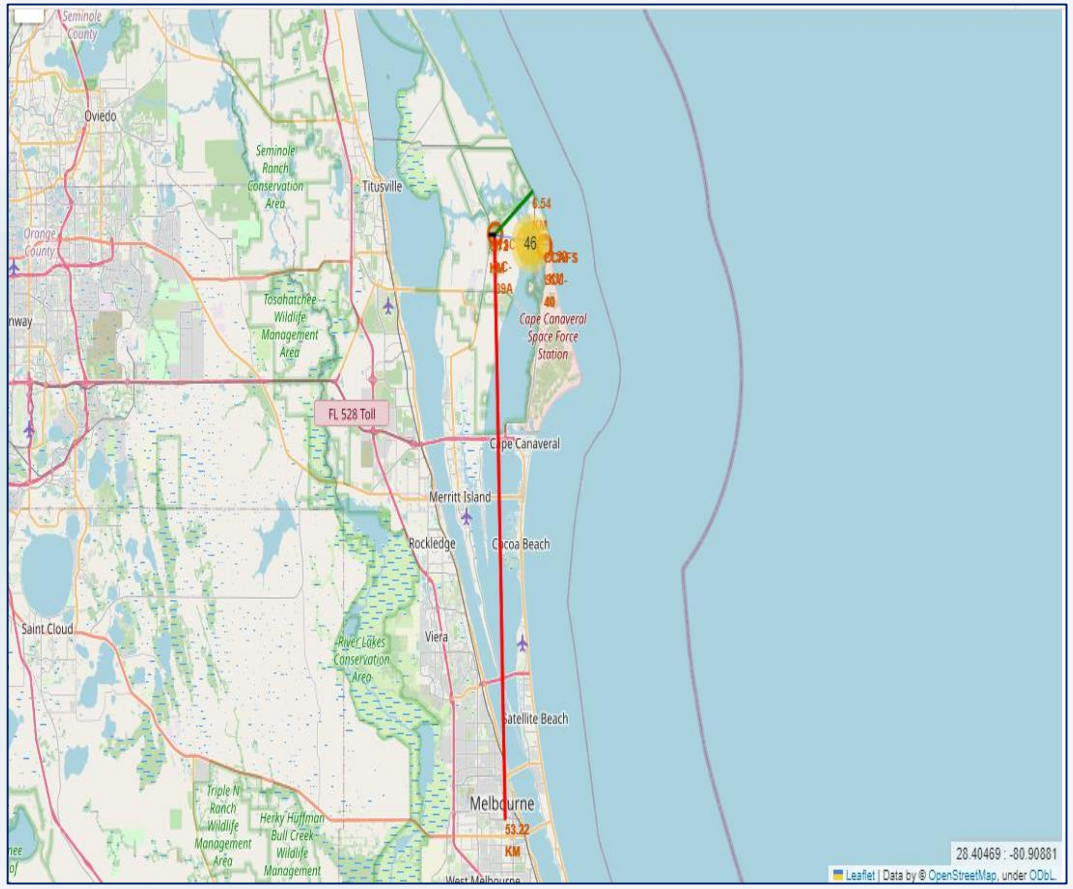


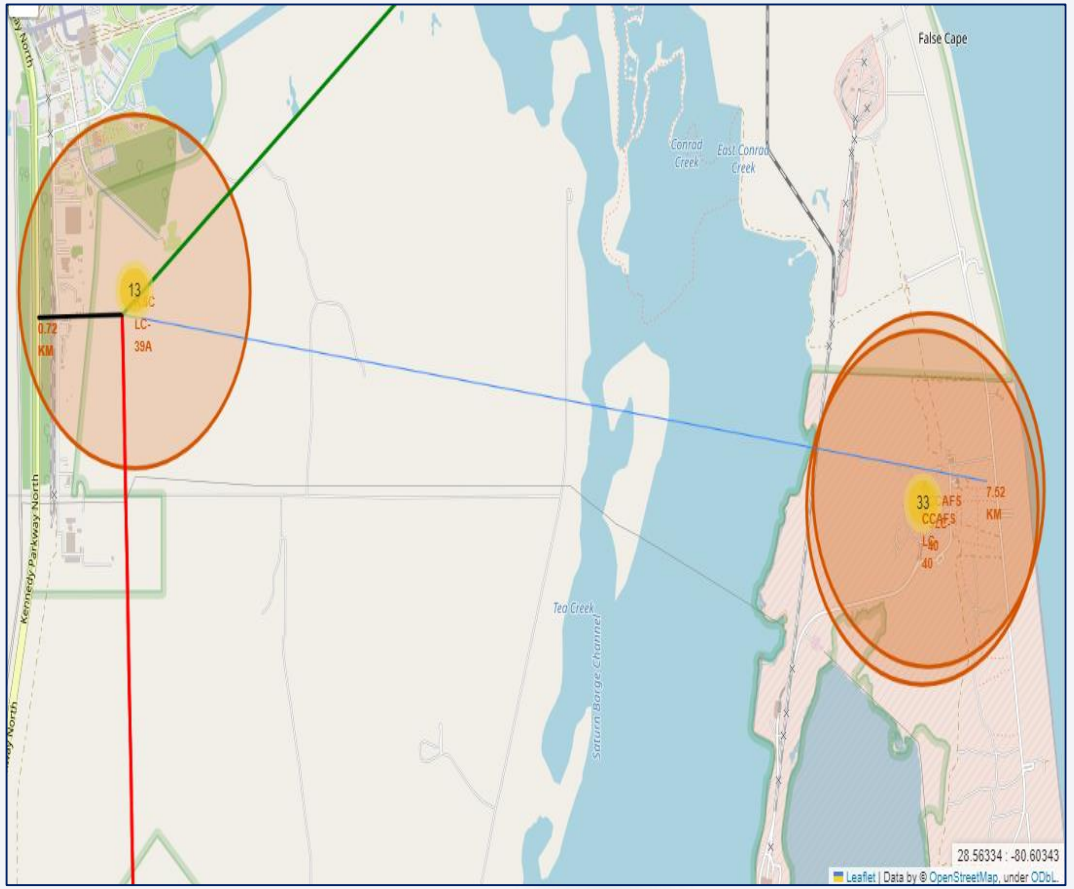


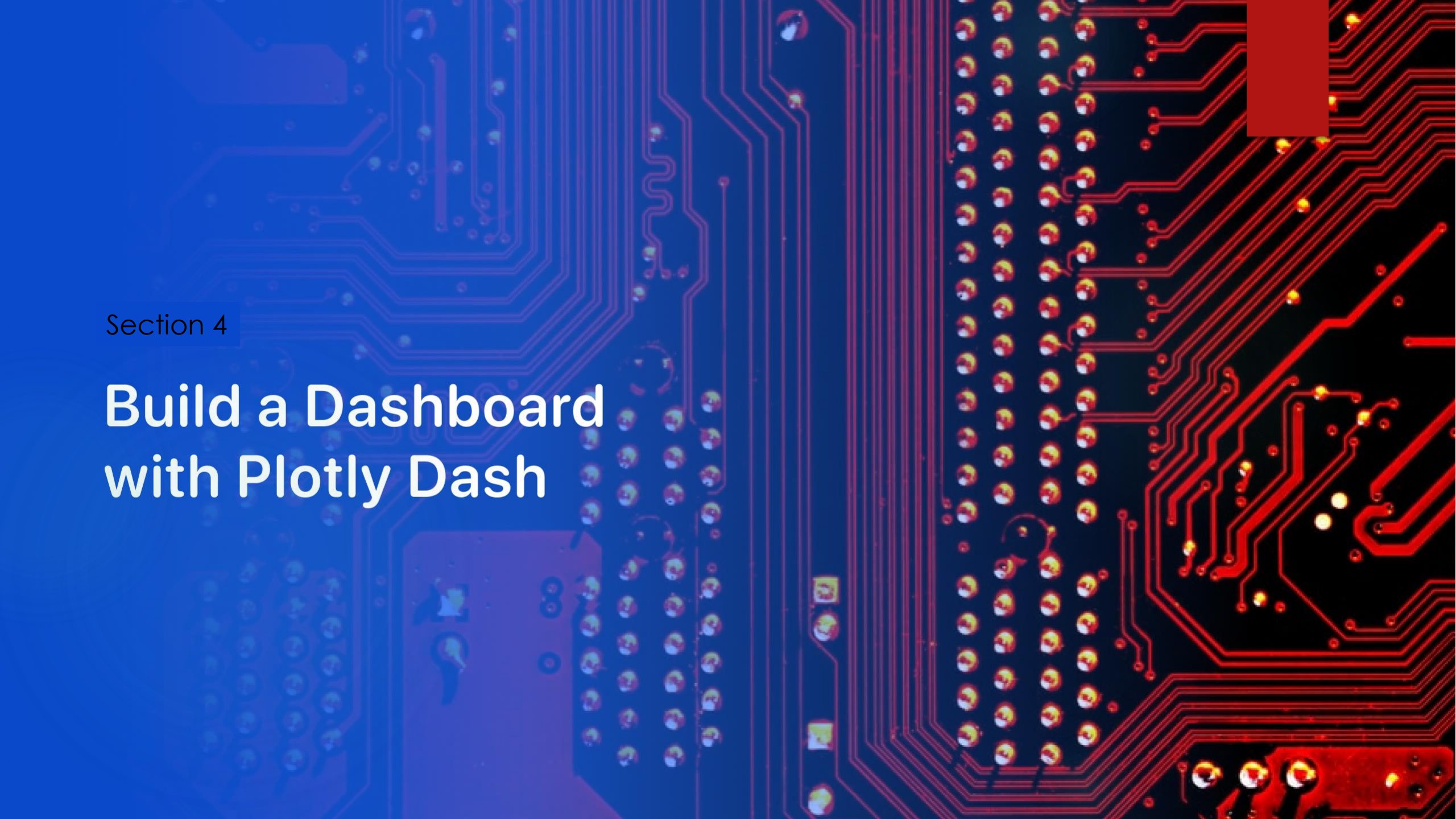


Distance from Launch Sites to Proximities

These figures below, show a PolyLine between site KSC LC - 39 A to the selected coastline point such as railway, highway and City of Melbourne.

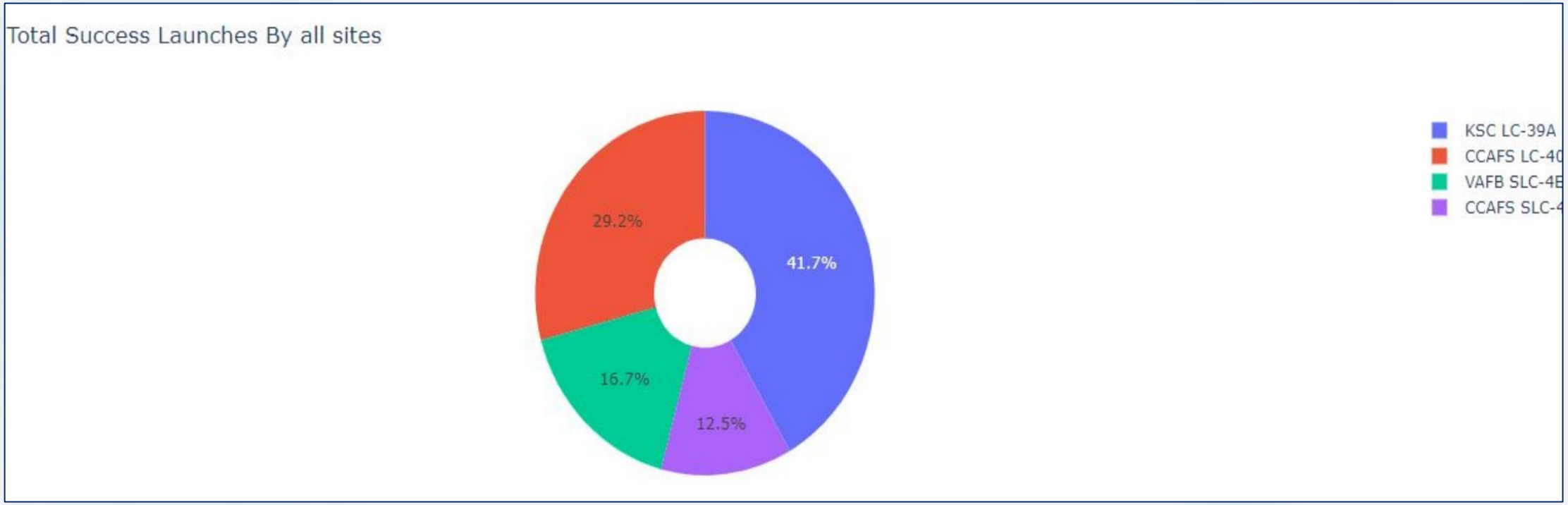





Section 4
Build a Dashboard with Plotly Dash

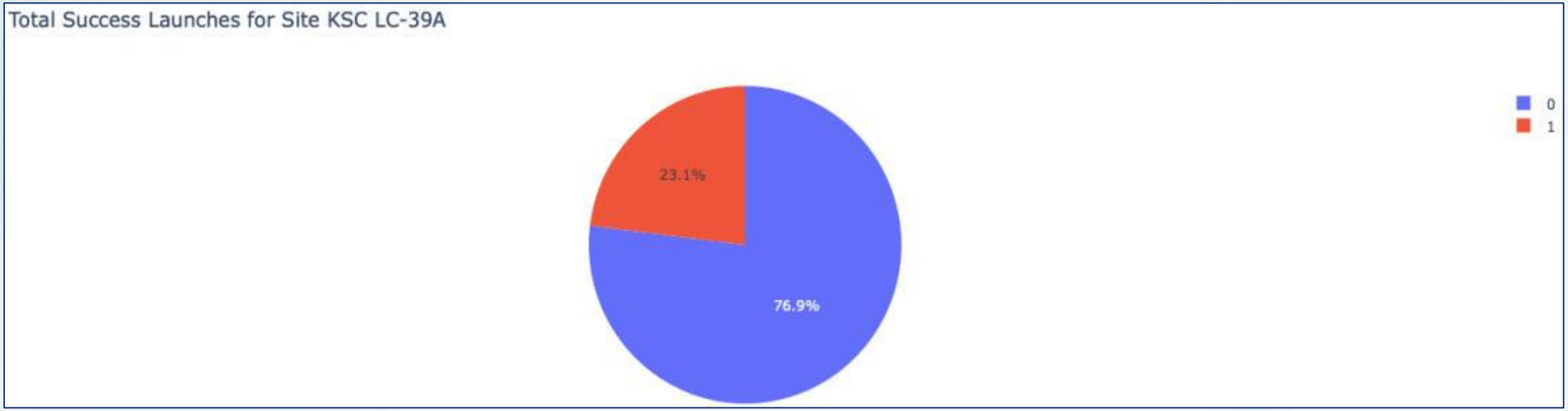
Total Success Launches By All Sites

The location of launch appears to be a significant contributing factor to the success of missions.
KSC LC-39A has the most successful launches compared to other sites



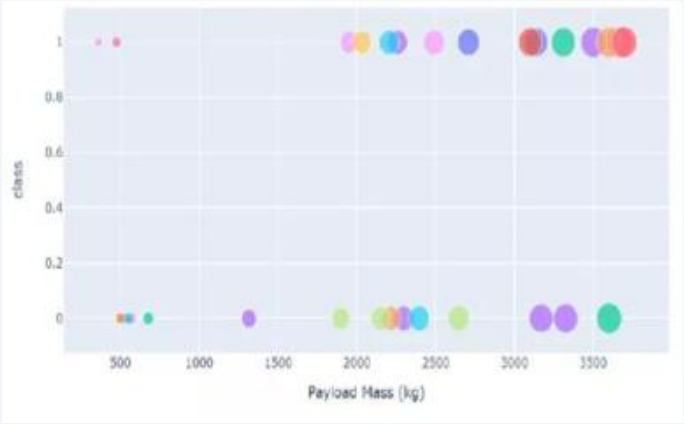
Launch site with highest launch success ratio

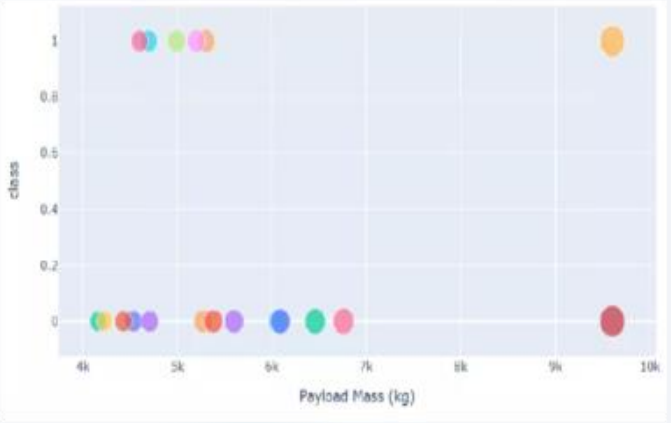
Using the dropdown menu on the dashboard allows for viewing single site launches.
At KSC LC-39A, 76.9% of the launches resulted in success, while 23.1% experienced failure.



Payload vs. Launch Outcomes

Through the Range Slider, we gain insights into the results of launches—both successful and unsuccessful—across various booster versions, emphasizing the payloads they carried. It's evident that launches with lighter payloads consistently achieve higher success rates compared to those with heavier payloads.



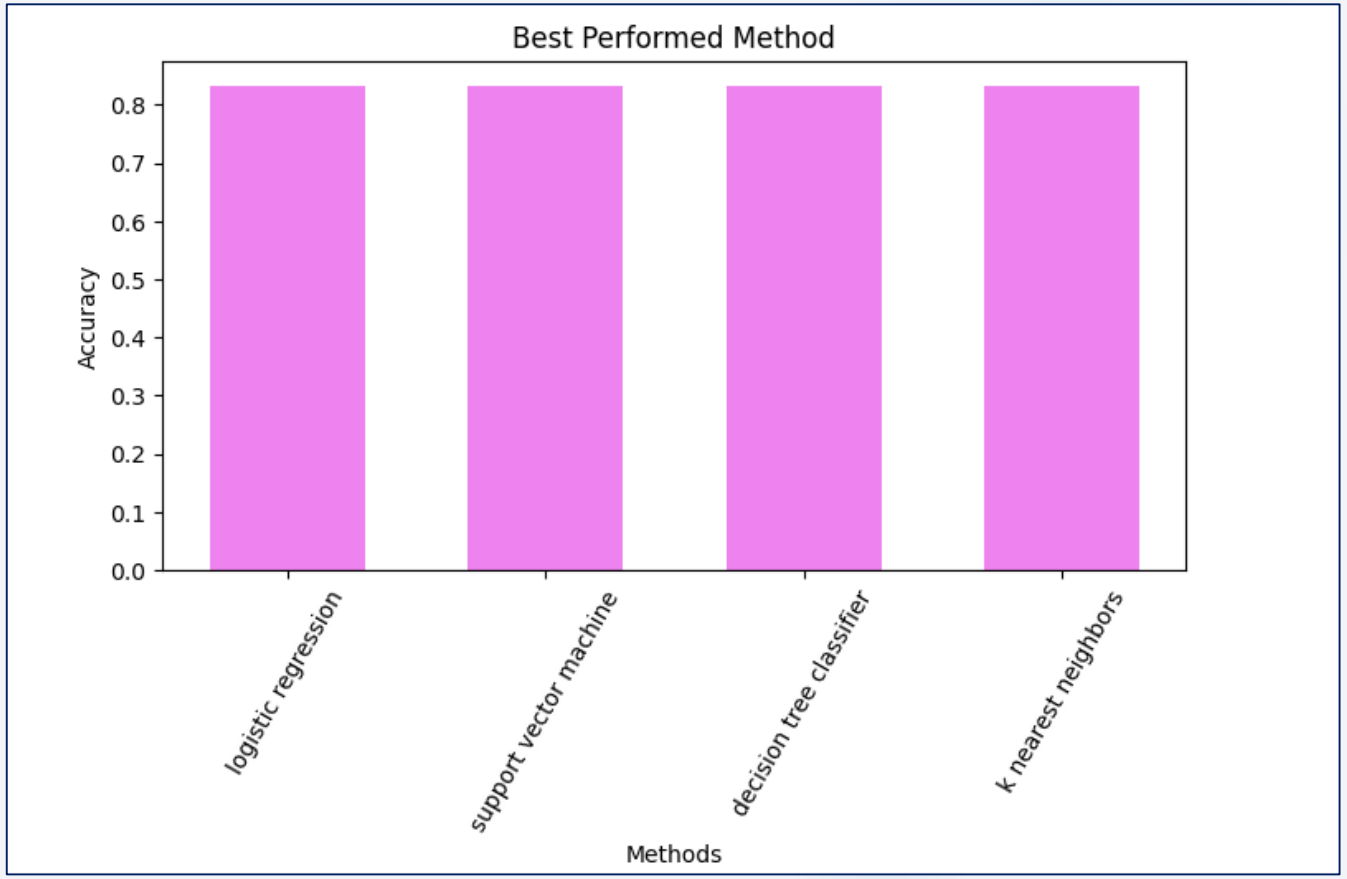


Section 5

Predictive Analysis (Classification)

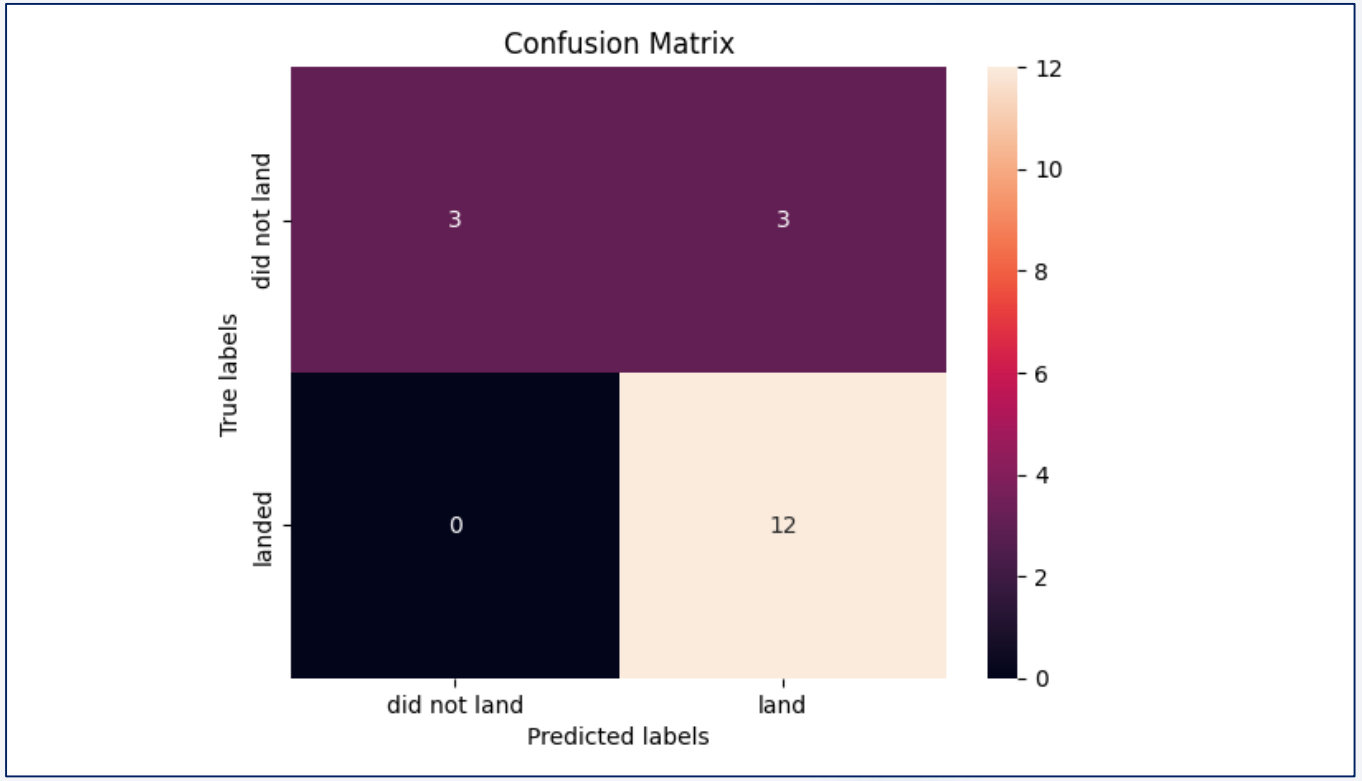
Classification Accuracy

All models have the same classification accuracy.



Confusion Matrix

All four models can distinguish between the different classes.
However, the major problem for all is False Positives (n=3).



Conclusions

- The success rates of SpaceX launches have shown improvement from 2013 to 2020.
- Among launch sites, KSC LC-39A boasts the highest number of successful launches.
- Lighter payloads tend to have a higher success rate than their heavier counterparts. The orbits of ES-L1, GEO, HEO, and SSO have recorded the highest success rates.
- For payloads on the heavier side, VLO and ISS orbits demonstrate superior success rates.
- For predictive accuracy with this dataset, Decision Tree models stand out as the most effective.

By leveraging data and thorough analysis, rocket companies can identify optimal strategies to reduce launch costs. This not only mitigates the risk of client attrition but also ensures a competitive edge in the market.

Appendix

1.Data Sources:

1. SpaceX REST API
2. Wikipedia: List of Falcon 9 and Falcon Heavy launches
3. Python Notebooks: Github Repository

2.Tools & Technologies Used:

1. Plotly Dash for interactive visual analytics
2. Folium for map visualizations
3. SQL for data querying
4. Python libraries: Pandas, NumPy, Scikit-learn, Matplotlib

3.Key Definitions:

1. **VLEO:** Very Low Earth Orbit
2. **GTO:** Geostationary Transfer Orbit
3. **CCAFS:** Cape Canaveral Air Force Station
4. **KSC:** Kennedy Space Center

4.Acknowledgments:

1. SpaceX for providing open access to launch data
2. Open-source communities for tools and libraries

5.Contact Information:

1. Elaheh Sarshar
2. Email: elahehsarshar.es@gmail.com
3. LinkedIn: elahehsarshar
4. Github: elahehsarshar

Thank you!