Winning Space Race with Data Science

King Ting Chan
24-06-2024

Outline

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

Executive Summary

In this project, we conducted an extensive analysis of SpaceX Falcon 9 launches employing various data science methodologies. Our approach included:

- **Data Collection:** Leveraging the SpaceX API to gather detailed launch data.
- **Data Wrangling**: Cleaning and preprocessing the data to ensure accuracy and consistency.
- **Exploratory Data Analysis (EDA) with Data Visualization**: Analyzing the data to uncover insights and trends, using tools like Matplotlib and Seaborn for visualization.
- **Exploratory Data Analysis with SQL**: Executing complex queries to extract specific insights and aggregate information from the dataset.
- **Building an Interactive Map with Folium**: Mapping the geographical distribution of launch sites.
- **Building a Dashboard with Plotly Dash**: Developing an interactive dashboard for dynamic data exploration and visualization.
- **Predictive Analysis (Classification)**: Utilizing machine learning techniques to predict the classification of the next landing, optimized with Grid Search.

Introduction

Project background and context

SpaceX has revolutionized the commercial space industry by making space travel more affordable, primarily through the innovative reuse of their rockets' first stages.

A Falcon 9 rocket launch costs $62 million, significantly less than the $165 million charged by competitors. This cost reduction is largely due to SpaceX's ability to re-land and reuse the first stage of their rockets.

Predicting the success of these landings is crucial for estimating launch costs more accurately. This project leverages public data and machine learning techniques to forecast the reusability of Falcon 9's first stage, which is vital for competitors aiming to price their services effectively against SpaceX.

As data scientists working to rival SpaceX, our objective is to develop a machine learning pipeline to predict first-stage landing outcomes, thus enabling more precise launch cost estimations and competitive bidding.

Questions to Be Answered

1. **Impact of Factors on Landing Success**: How do payload mass, launch site, number of flights, and orbits affect the success of a first-stage landing?

2. **Trends in Landing Success**: Has SpaceX's success rate for first-stage landings improved over the years?

3. **Predicting Landing Success**: Can we predict the success of a first-stage landing for new launches based on historical data?

4. **Optimal Classification Algorithm**: What is the most effective binary classification algorithm for predicting first-stage landing success?

5. **Enhancing Launch Success**: What factors ensure a successful launch, and how can this information be used to optimize future launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

Rocket launch data obtained via the SpaceX REST API and Falcon 9 data collected by web scraping Wikipedia.

- Perform data wrangling

Data processing involved converting the structured JSON API data to a normalized data frame, which was then saved as a CSV file. Categorical features were processed using one-hot encoding to prepare for binary classification. The data was filtered, missing values were handled, and missions were labeled according to orbit frequencies and landing outcomes (successful vs. failed).

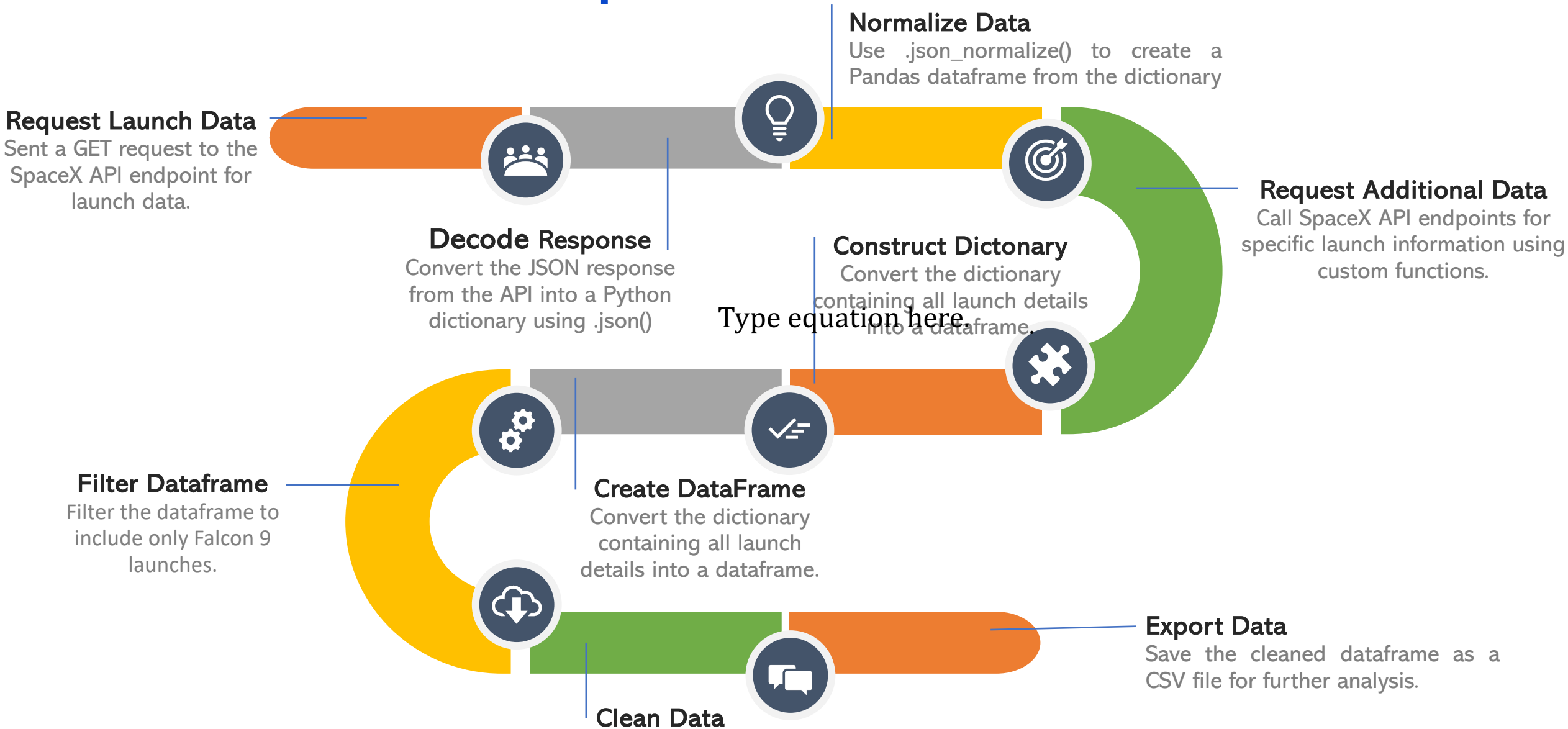
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Developed, tuned, and evaluated classification models to achieve optimal results, using GridSearchCV to determine the best-fitting model.

Data Collection

The data acquisition strategy involved using the SpaceX REST API and web scraping Falcon 9 historical launch information from Wikipedia to ensure access to the latest data. Data collection entailed gathering and measuring information on targeted variables to answer relevant questions and evaluate outcomes. For the REST API, we used a GET request, decoded the JSON response, and converted it into a pandas Data Frame with `json_normalize()`. We then cleaned the data and addressed any missing values. For web scraping, we employed BeautifulSoup to extract and parse HTML tables of launch records, converting them into a pandas Data Frame for further analysis.

Data Collection – SpaceX API



GitHub Link: [Data Collection API]

Data Collection – Scraping

Request Data
End a GET request to the Wikipedia page containing Falcon 9 launch data.

Parse Response
Use Beautiful Soup to create an object representing the HTML structure of the webpage.

Extract Headers
Identify and extract all column names (headers) from the HTML table containing launch data.

Parse Table Data
Iterate through the HTML table rows and columns, collecting launch information.

Construct Dictionary
Organize the collected launch data into a Python dictionary with headers as keys

Create Dataframe
Convert the dictionary containing all launch details into a dataframe.

Export Data
Save the cleaned dataframe as a CSV file for further analysis.

GitHub Link: Data Collection Scraping

Data Wrangling

Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA).

Import Libraries & Define Functions

- Import libraries like Pandas (data manipulation)
- Define functions for data cleaning/ transformation

Load SpaceX Dataset

Load the previously collected Falcon 9 launch data

Exploratory Data Analysis (EDA)

Through EDA, we dissect SpaceX launch data: site frequency, popular orbits, and cracking the "Outcome" code (True/ False for landing) to define our model's success metric.

Create Training Labels

Transform the "Outcome" column into a binary "Landing_Label." Successful = 1, unsuccessful = 0.This step creates a crucial target variable for model training.

GitHub Link: Data Wrangling

EDA with Data Visualization

Scatter Plot
Identify any patterns between the order of launches (Flight Number) and the location from where they took off (Launch Site)

Scatter Plot - Flight Number vs. Launch Site

Scatter Plot - Payload vs. Launch Site

Scatter Plot
Explore any potential relationships between the weight of the cargo carried (Payload) and the launch location (Launch Site)

Bar Chart
Visualize the success rate of missions for each specific orbit type (e.g., what percentage of launches achieved their intended orbit for each category)

Bar Chart - Success Rate by Orbit Type

Scatter Plot - Flight Number vs. Orbit Type

Scatter Plot
Uncover any patterns between the launch order (Flight Number) and the type of orbit the mission aimed to achieve (Orbit Type)

Scatter Plot
Explores any potential relationships between the cargo weight (Payload) and the type of orbit targeted by the launch (Orbit Type)

Scatter Plot - Payload vs. Orbit Type

Line Plot - Launch Success Yearly Trend

Line Plot
Identify trends in launch success rates over time (yearly). It reveals whether the success rate is improving, declining, or staying consistent across years.

GitHub Link: EDA with Data Visualization

EDA with SQL

Performed SQL queries:

Displaying the names of the 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

GitHub Link: EDA with SQL

Build an Interactive Map with Folium

Launch Sites:

Markers with circles indicate all launch site locations, with text labels providing additional information such as "NASA Johnson Space Center." These markers illustrate the geographical distribution of launch sites and their proximity to the equator and coastlines.

Launch Outcomes:

Green and red markers denote successful and failed launches, respectively. Marker clusters group these indicators, enabling viewers to quickly identify launch sites with high success rates.

Distances:

Lines connect specific launch sites (e.g., KSC LC-39A) to nearby features such as railways, highways, coastlines, and the nearest city, highlighting the proximity of launch sites to essential infrastructure and population centers.

GitHub Link: Interactive map with Folium

Build a Dashboard with Plotly Dash

Dropdown List:

This feature allows users to filter data by selecting a specific launch site. When a site is chosen, the dashboard focuses on that location's launch information.

Pie Chart:

This chart displays the total number of successful launches across all sites. When a specific site is selected from the dropdown list, the pie chart updates dynamically to show the success and failure rates for that location.

Slider:

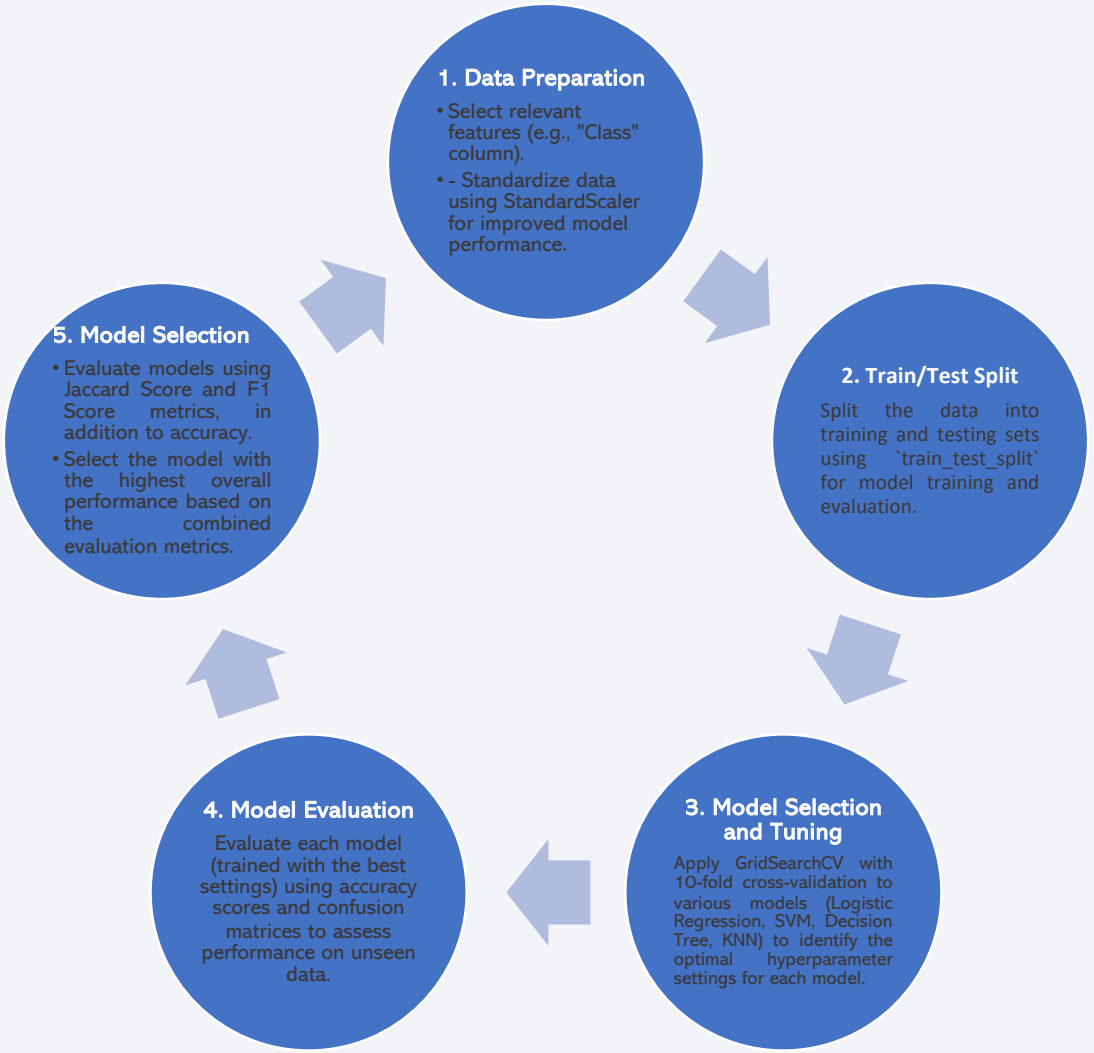
The slider enables users to select a specific range of payload mass, allowing for focused analysis of how launch success rates correlate with the weight of the cargo.

Scatter Chart:

This scatter plot visualizes the relationship between payload mass and launch success rates for different booster versions. By using the slider and dropdown filters, users can explore these correlations for specific launch sites or booster types.

GitHub Link: Dashboard with Plotly Dash

Predictive Analysis (Classification)



GitHub Link: [Predictive Analysis]

Results

- **Exploratory data analysis results**

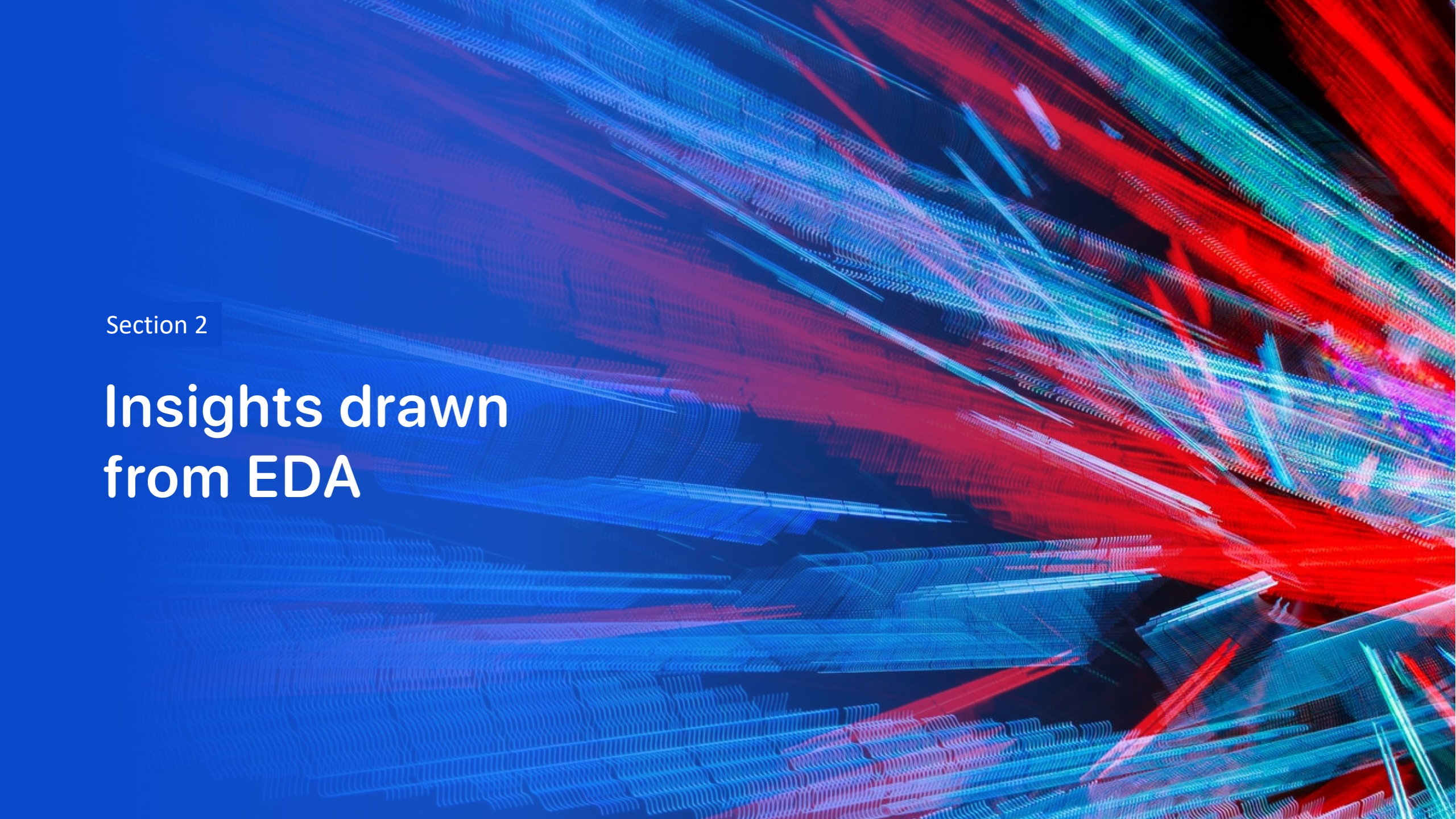
Different launch sites have varying success rates. CCAFS LC-40 has a success rate of 60%, while KSC LC-39A and VAFB SLC-4E each have a success rate of 77%.

- **Interactive analytics demo in screenshots**

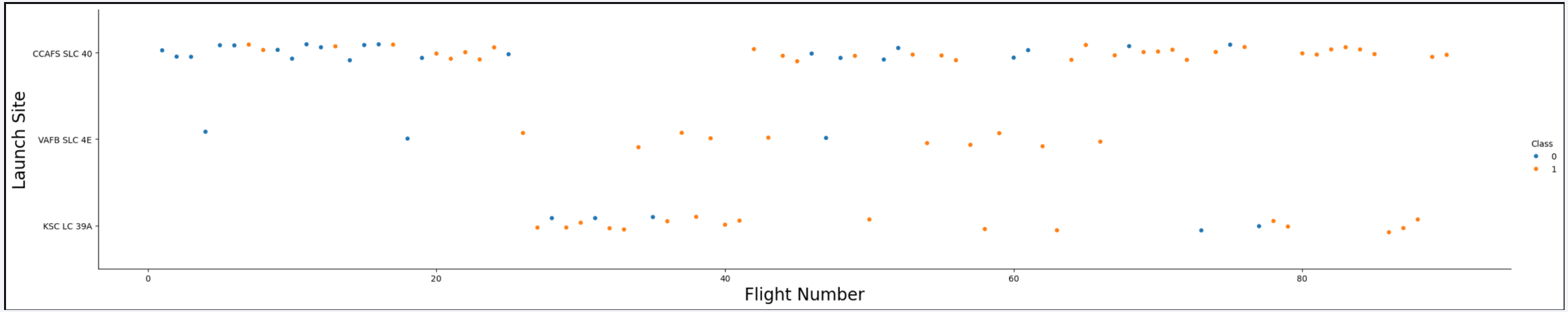
Refer to section 2

- **Predictive analysis results**

Although all models achieved an accuracy close to 85%, the Decision Tree performed slightly better.


Section 2
Insights drawn from EDA

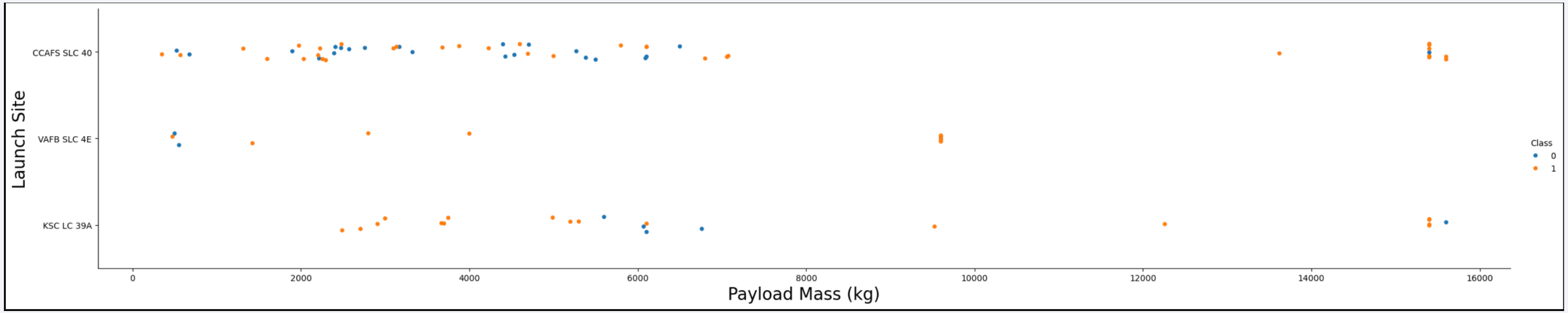
Flight Number vs. Launch Site



Early flights experienced failures, but recent flights have been successful. CCAFS SLC-40 hosts about half of all launches, with a success rate of 60%. In contrast, KSC LC-39A and VAFB SLC-4E have higher success rates of 77%. Overall, success rates have improved with each new launch.

GitHub Link: EDA with Data Visualization

Payload vs. Launch Site

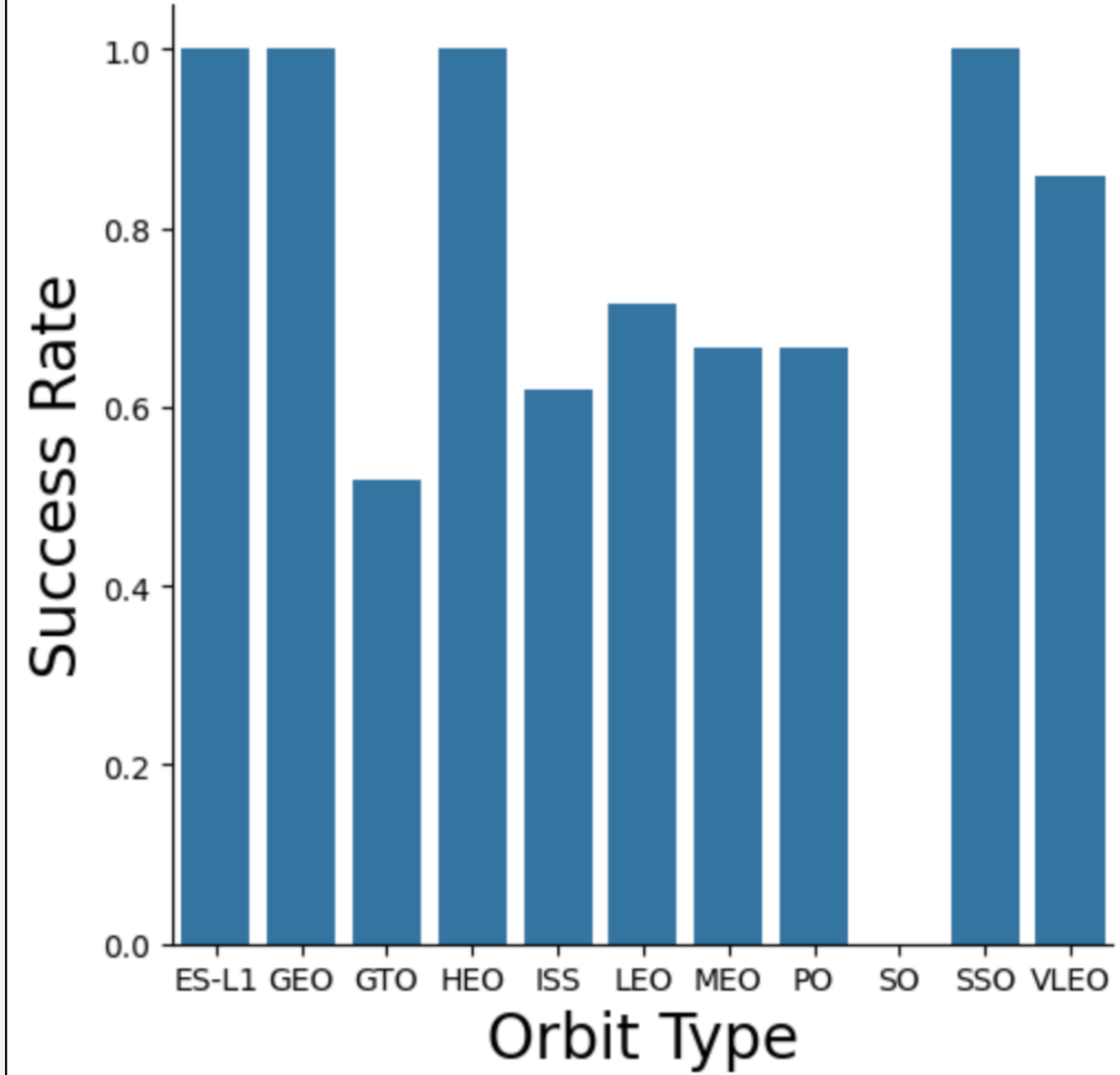


Higher payload mass generally leads to higher success rates, with most launches carrying over 7000 kg being successful. KSC LC-39A boasts a 100% success rate for payloads under 5500 kg. While early launches with lighter payloads struggled, SpaceX has improved significantly.

Heavier payloads see higher success across various sites, with KSC excelling in handling lighter ones. The scatter plot indicates that payloads above 7000 kg have a significantly increased probability of success, though there is no clear pattern linking launch site success rates to payload mass.

GitHub Link: EDA with Data Visualization

Success Rate vs. Orbit Type



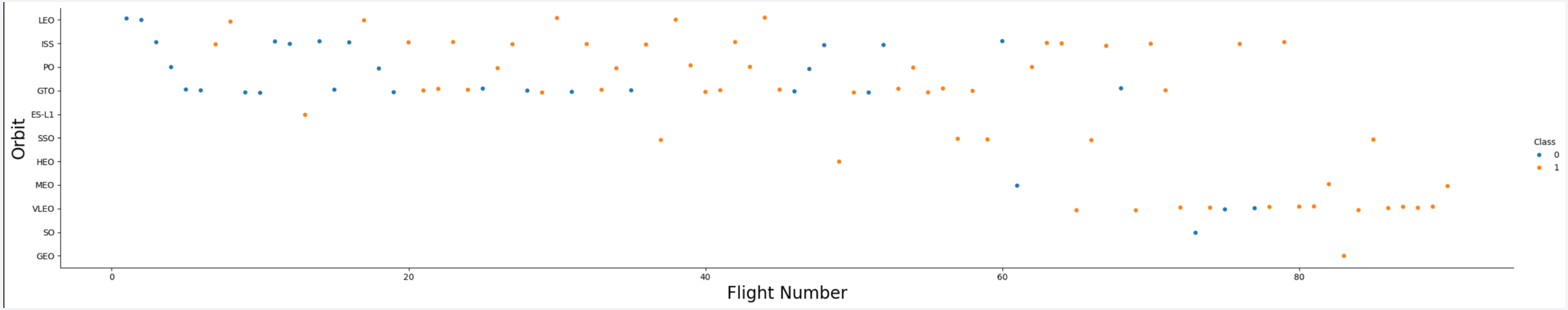
The bar chart shows that certain orbits, including ES-L1, GEO, HEO, and SSO, have a 100% success rate, while orbits such as GTO, ISS, LEO, MEO, PO, and VLEO have success rates between 50% and 80%. The SO orbit, however, has a 0% success rate.

Although orbits like ES-L1, GEO, HEO, and SSO show high success rates, it's important to note that some of these orbits, such as GEO, SO, HEO, and ES-L1, have only one occurrence in the dataset.

Therefore, more data is needed to identify any reliable patterns or trends before drawing definitive conclusions about the influence of orbits on landing outcomes.

GitHub Link: EDA with Data Visualization

Flight Number vs. Orbit Type

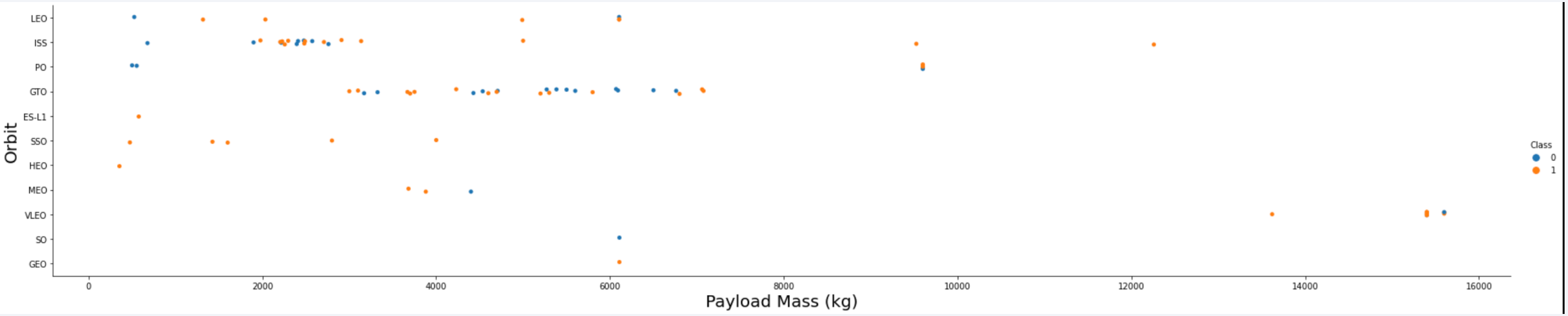


The scatter plot indicates that in the LEO orbit, success is related to the number of flights; the higher the flight number, the greater the success rate. In contrast, there is no apparent relationship between flight number and success in the GTO orbit.

Generally, increased flight numbers correlate with higher success rates across most orbits, especially LEO, except for GTO, which shows no such relationship. Orbits with only one occurrence should be excluded from this analysis as more data is needed for a conclusive pattern.

GitHub Link: EDA with Data Visualization

Payload vs. Orbit Type

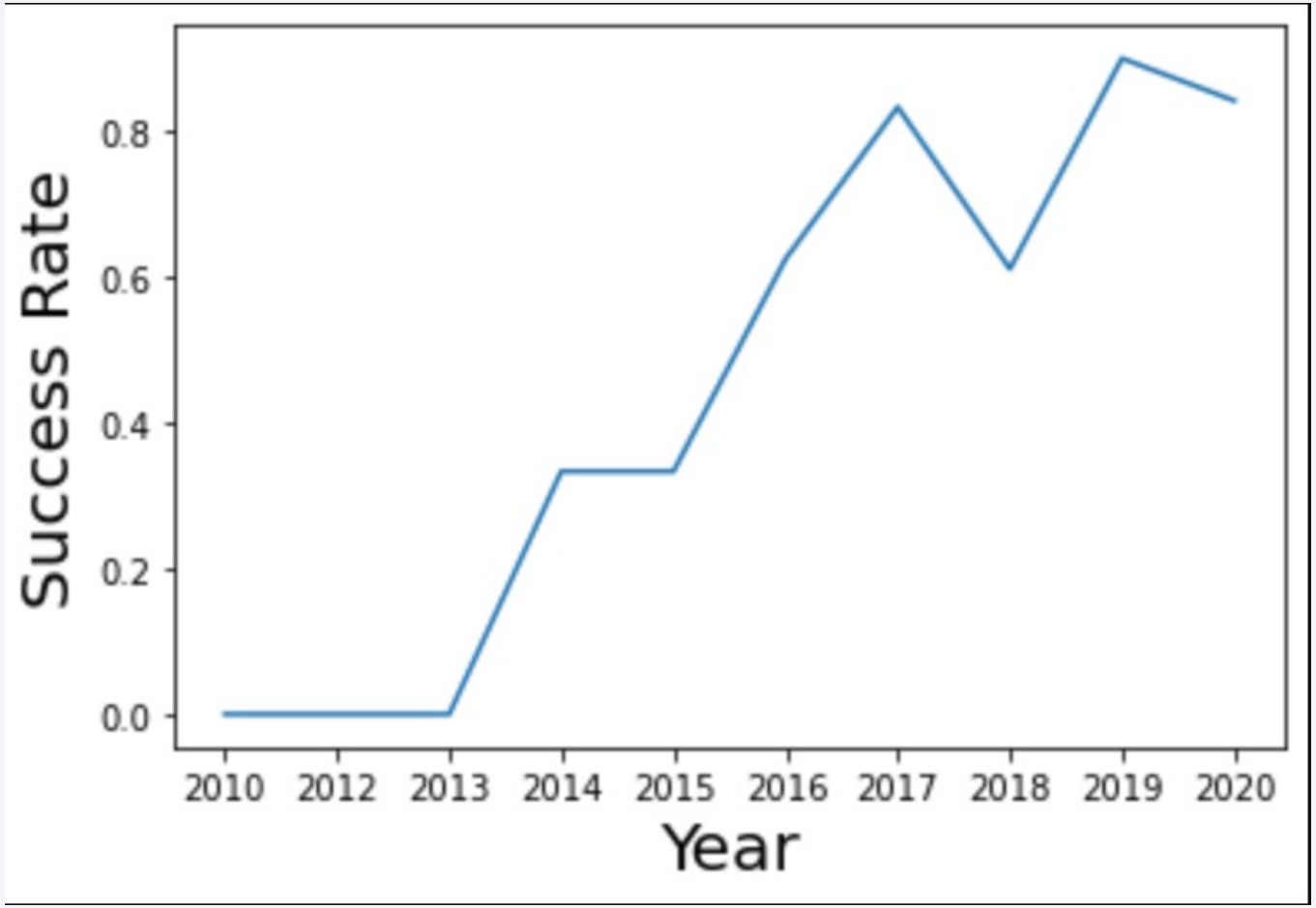


With heavy payloads, the success rate is higher for Polar, LEO, and ISS orbits. Heavier payloads positively impact LEO, ISS, and Polar orbits, while they have a negative impact on MEO and VLEO orbits.

For GTO, there is no clear relationship between payload weight and landing success, as both successful and unsuccessful landings occur. Additionally, orbits such as SO, GEO, and HEO require more data to determine any definitive patterns or trends.

GitHub Link: EDA with Data Visualization

Launch Success Yearly Trend



The line chart shows a clear increasing trend in success rates from 2013 to 2020, with the most significant growth occurring between 2013 and 2017.

If this trend continues, the success rate is expected to steadily rise, potentially reaching 100% in the coming years.

GitHub Link: EDA with Data Visualization

All Launch Site Names

```
%sql select distinct launch_site from SPACEXDATASET;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

SpaceX Launch Facilities:

- Cape Canaveral Space Launch Complex 40 (CCAFS SLC-40, CCAFS LC-40)
- Vandenberg Space Force Base Space Launch Complex 4E (VAFB SLC-4E)
- Kennedy Space Center Launch Complex 39A (KSC LC-39A)

GitHub Link: EDA with SQL

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

The following query displays records of launches conducted from all sites with names starting with 'CCA'.

GitHub Link: EDA with SQL

Total Payload Mass

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

```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

total_payload_mass
45596

The following query displays the total payload carried by the NASA (CRS) which is 45596 kg.

GitHub Link: EDA with SQL

Average Payload Mass by F9 v1.1

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

```
%sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

average_payload_mass
2534

The following query shows the average payload carried by the Booster Version F9 v1.1 which is 2928.4 .

GitHub Link: EDA with SQL

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

first_successful_landing
2015-12-22

The following query shows the first successful ground landing date which was 22nd December 2015.

GitHub Link: EDA with SQL

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

The following query indicates the Booster Versions which were successful in landing with payload between 4000 and 6000 kg.

GitHub Link: EDA with SQL

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

The following query groups and displays total number of mission outcomes be it success or failure.

GitHub Link: EDA with SQL

Boosters Carried Maximum Payload

```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
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

The following query gives a list of Booster Versions carrying maximum payload.

GitHub Link: EDA with SQL

2015 Launch Records

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

The following query displays a list of Booster Versions and launch site for a launches in 2015 where the landing outcome was failure on drone ship.

GitHub Link: EDA with SQL

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

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

```
%%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
      where date between '2010-06-04' and '2017-03-20'
      group by landing__outcome
      order by count_outcomes desc;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

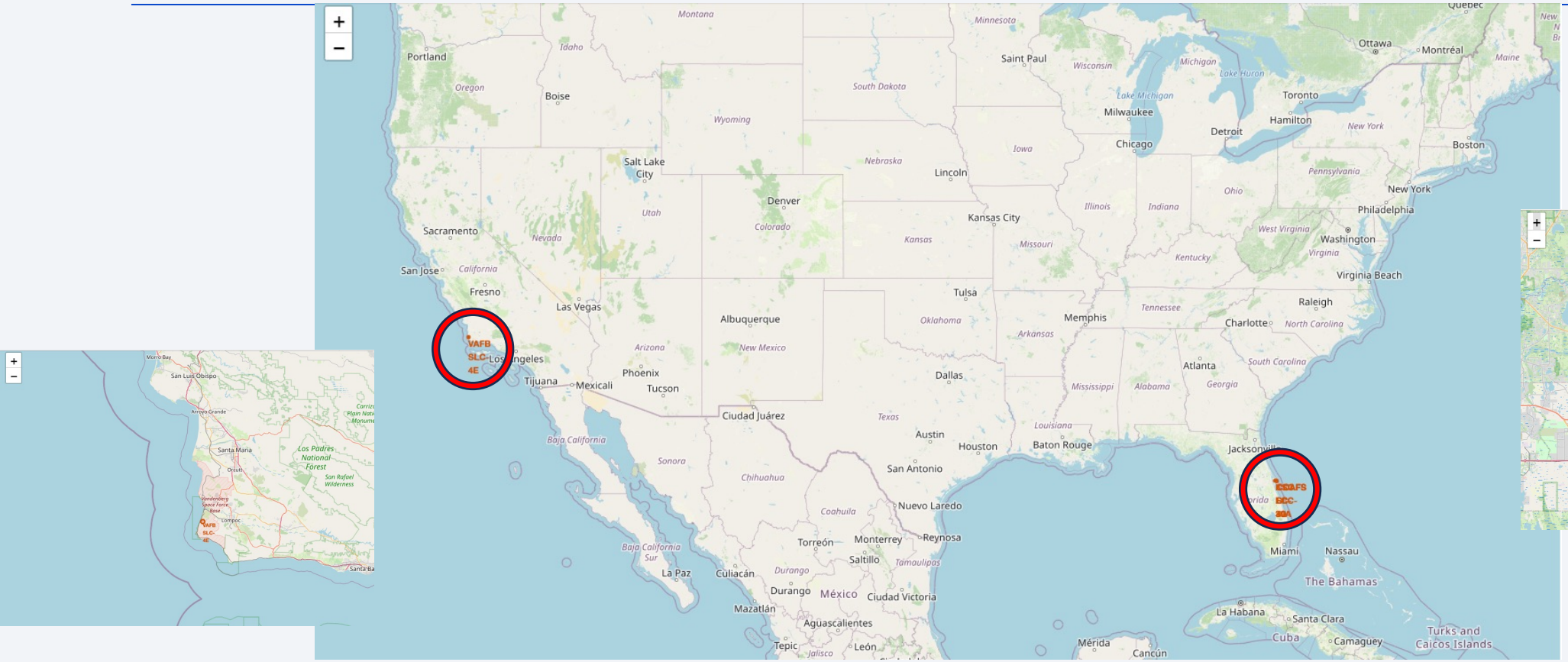
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

The following query ranks landing outcome between specific dates and groups them by landing outcome with their counts.

GitHub Link: EDA with SQL


Section 3
Launch Sites
Proximities Analysis

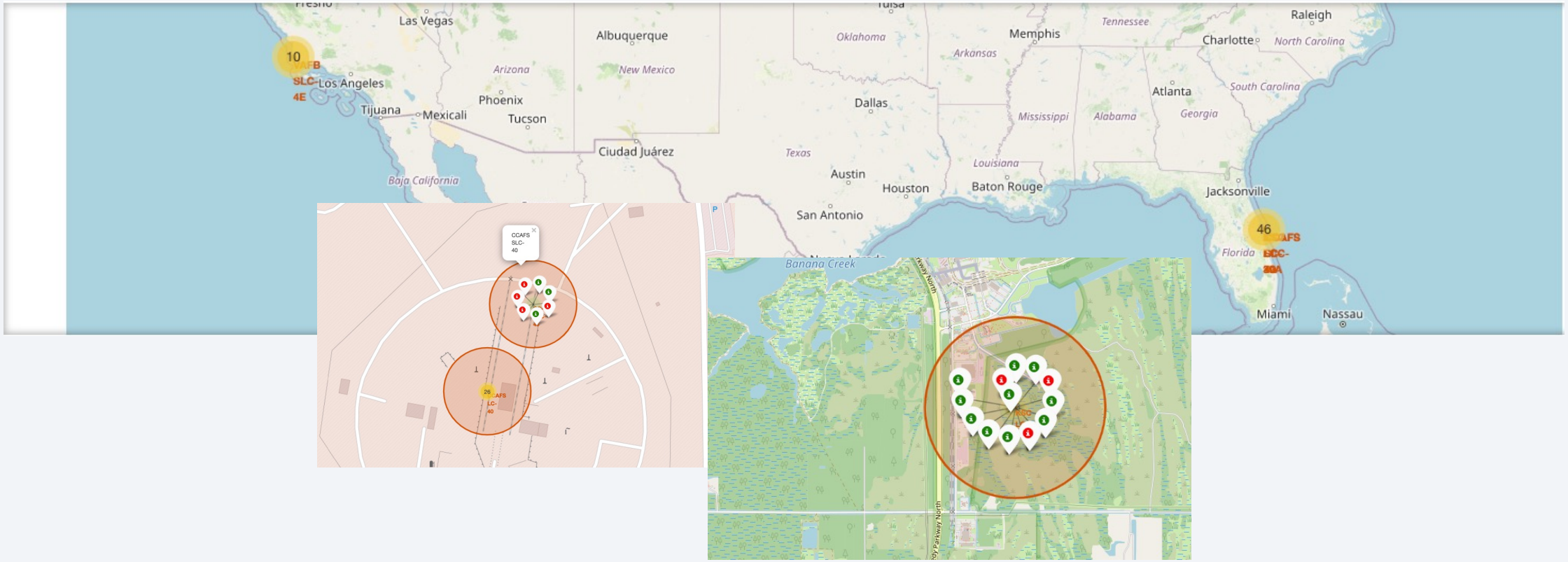
Location of SpaceX Launch Sites



SpaceX launch sites strategically placed near oceans aid in debris mitigation and are situated on both the East and West coasts of the southern United States, specifically Florida and California, taking advantage of their proximity to the equator for Earth's rotational boost.

GitHub Link: [Interactive map with Folium]

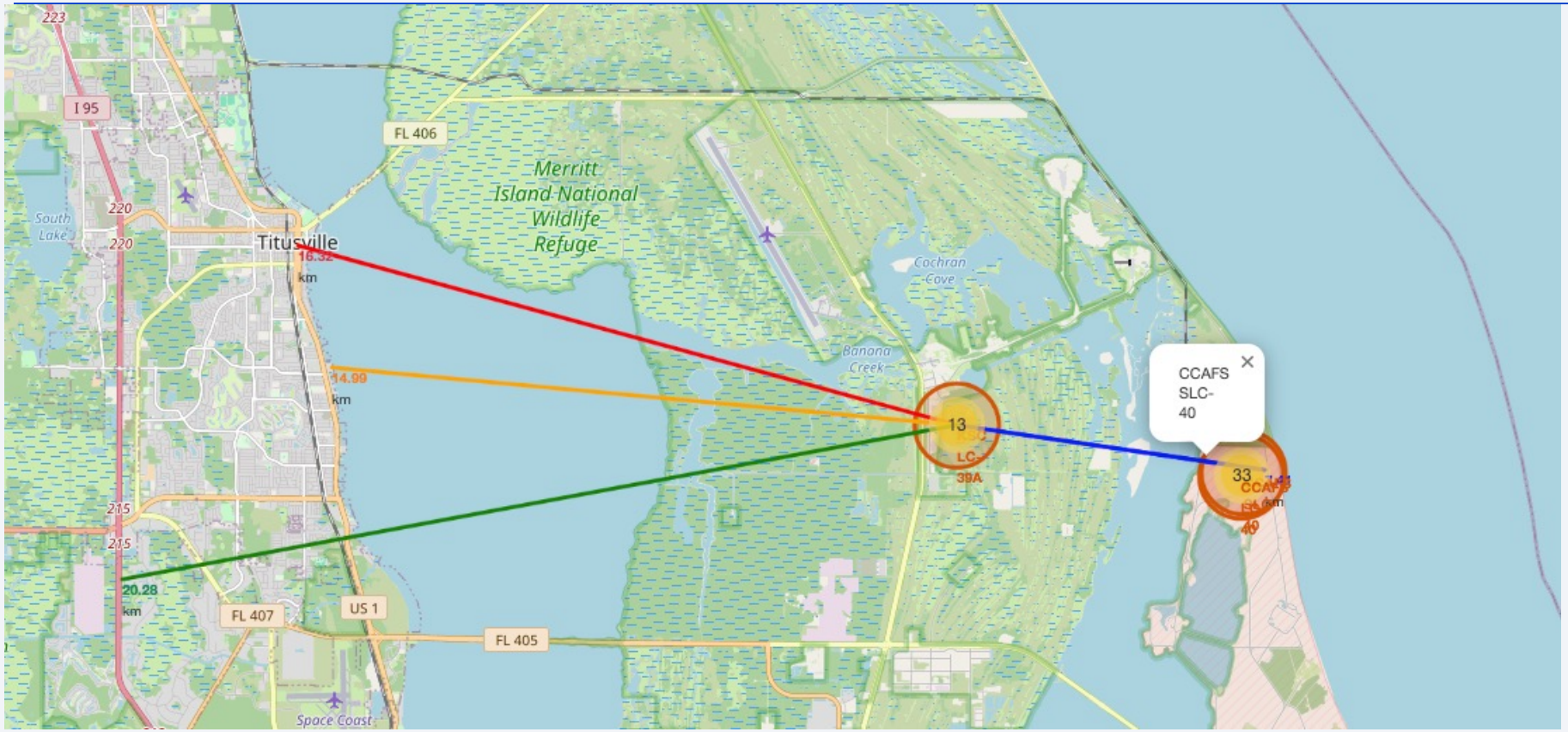
Launch Outcomes at Launch Sites



Launch outcome markers at launch sites indicate success with green markers and failure with red markers. By examining the color-labeled markers within clusters, we can readily discern which launch sites boast relatively high success rates.

GitHub Link: [Interactive map with Folium]

Proximities Distance From KSC LC-39A



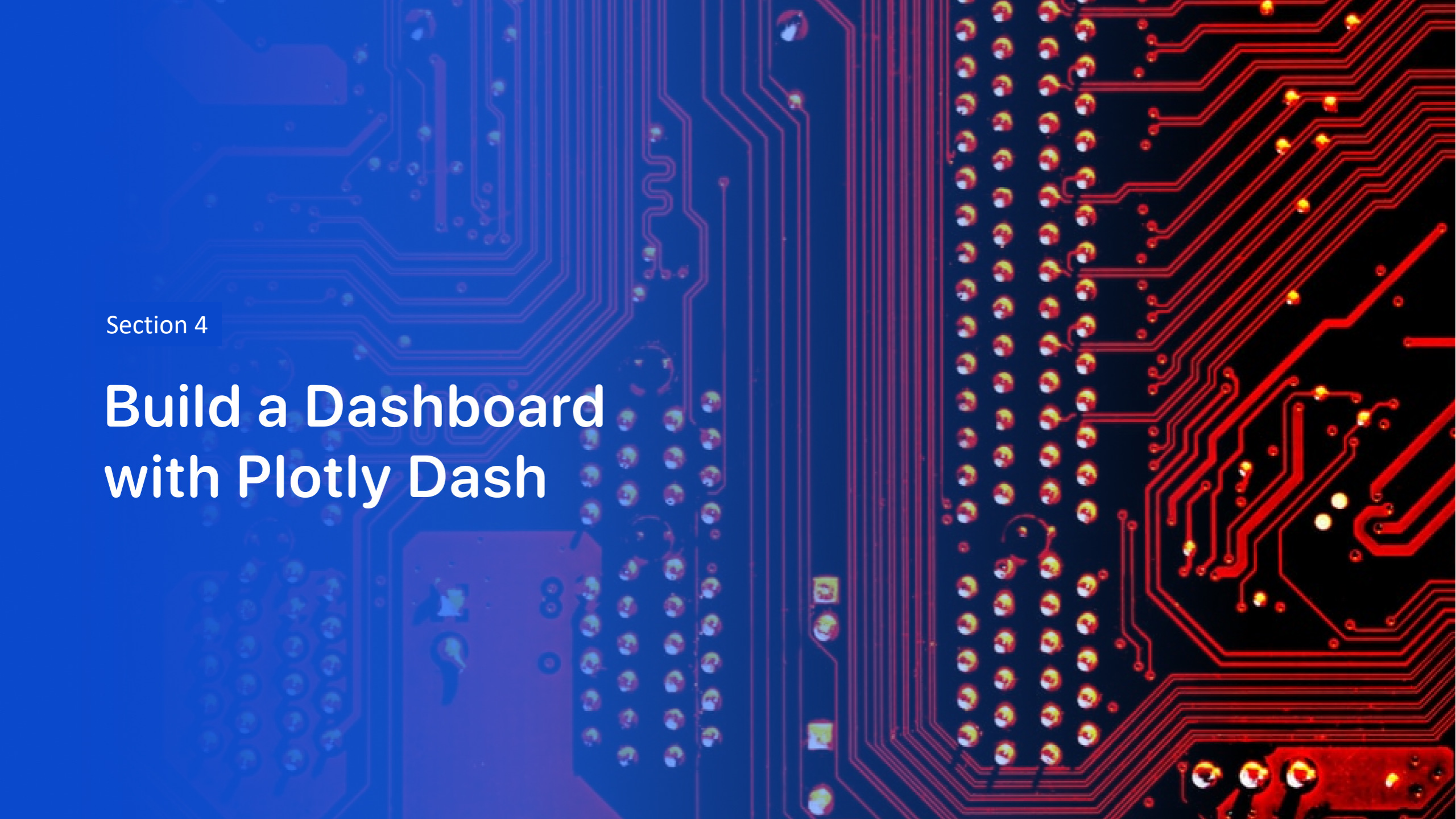
The closest railway to the launch site is approximately 15 km distant.

The nearest highway to the launch site is around 20 km away.

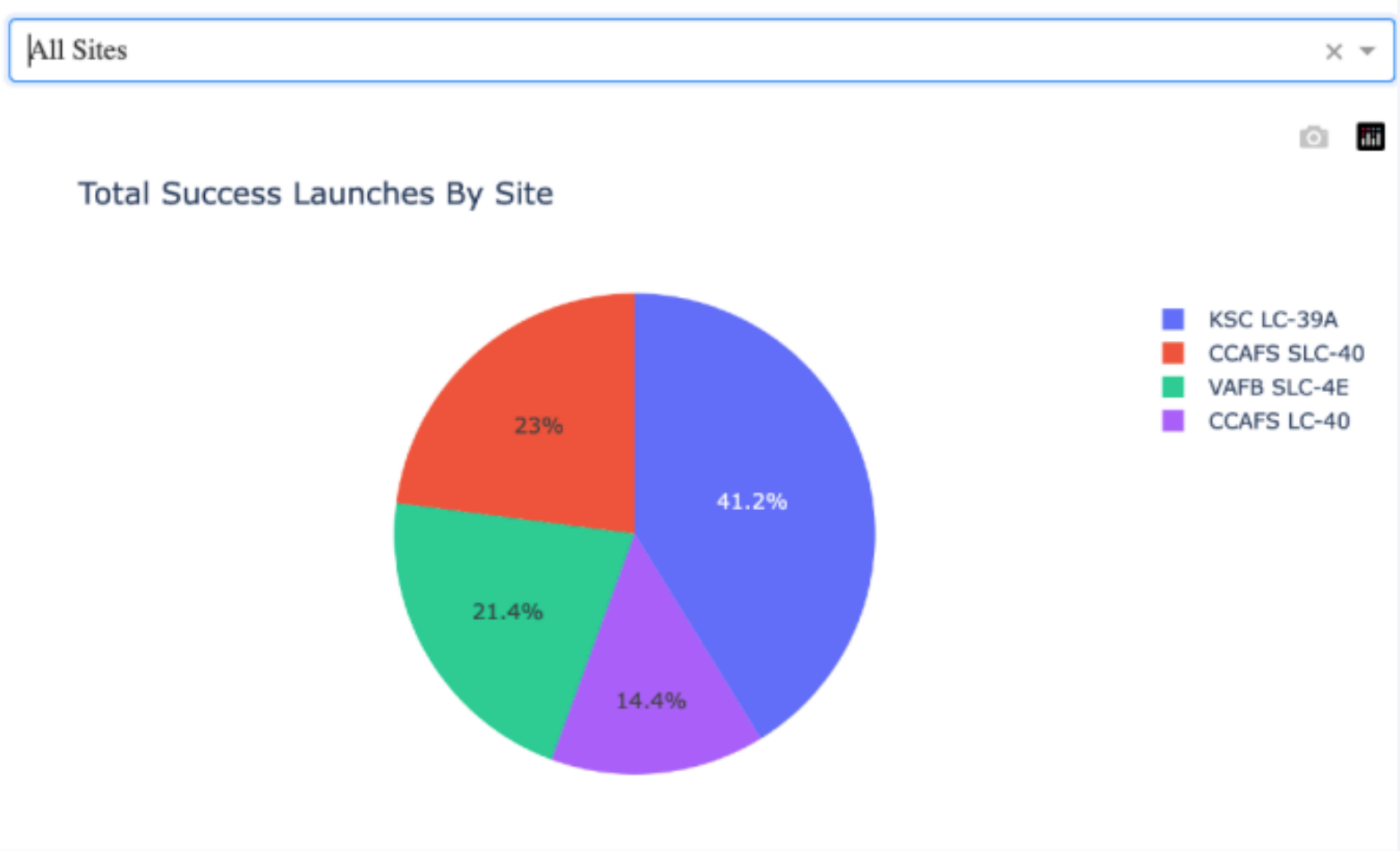
The closest coastline to the launch site is approximately 6 km away.

Are launch sites in close proximity to railways? NO
Are launch sites in close proximity to highways? NO
Are launch sites in close proximity to coastline? YES
Do launch sites keep certain distance away from cities? YES

GitHub Link: Interactive map with Folium


Section 4
Build a Dashboard with Plotly Dash

Total Success Launches By Site

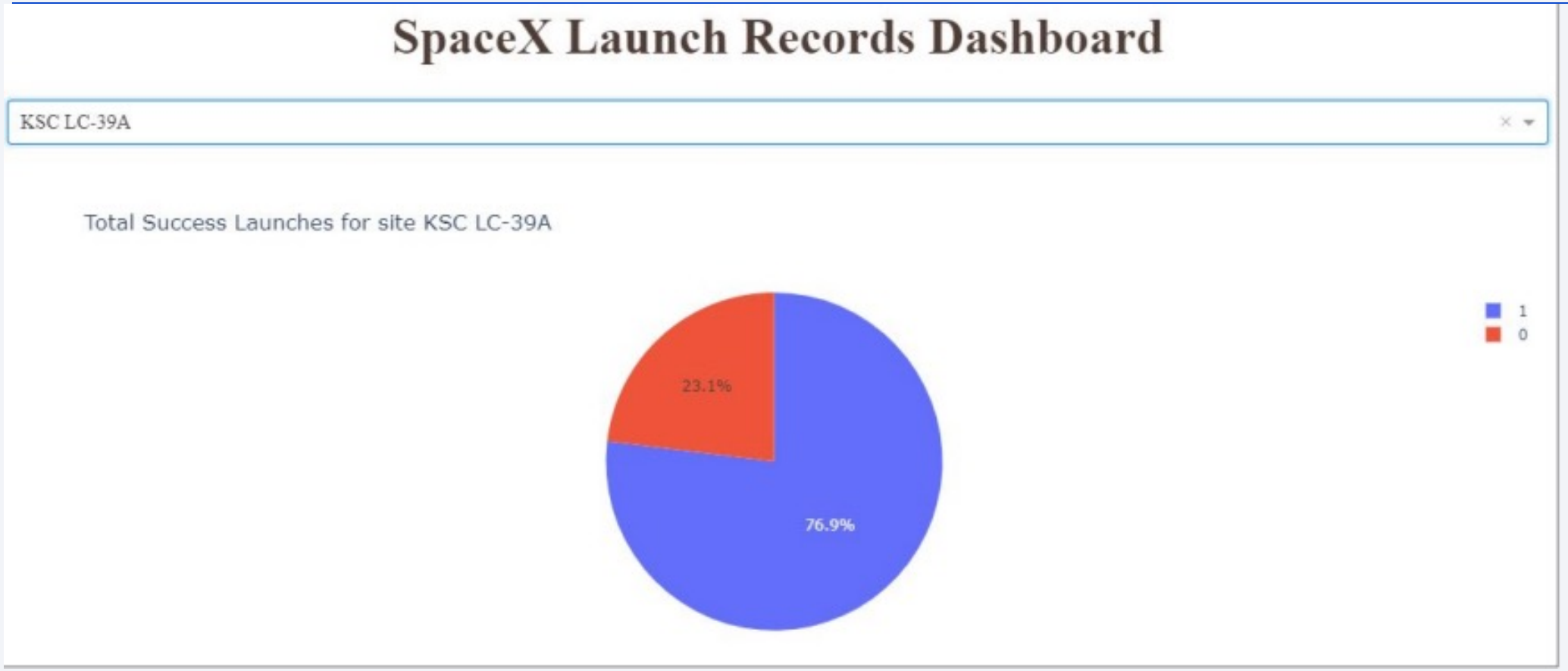


According to the pie chart, KSC LC-39A leads with the highest proportion of successful launches, accounting for approximately 41.2%, slightly ahead of CCAFS SLC-40 which contributes around 23%.

VAFB SLC-4E follows with 21.4% of successful launches, while CCAFS LC-40 has the lowest percentage at 14.4%.

GitHub Link: Dashboard with Plotly Dash

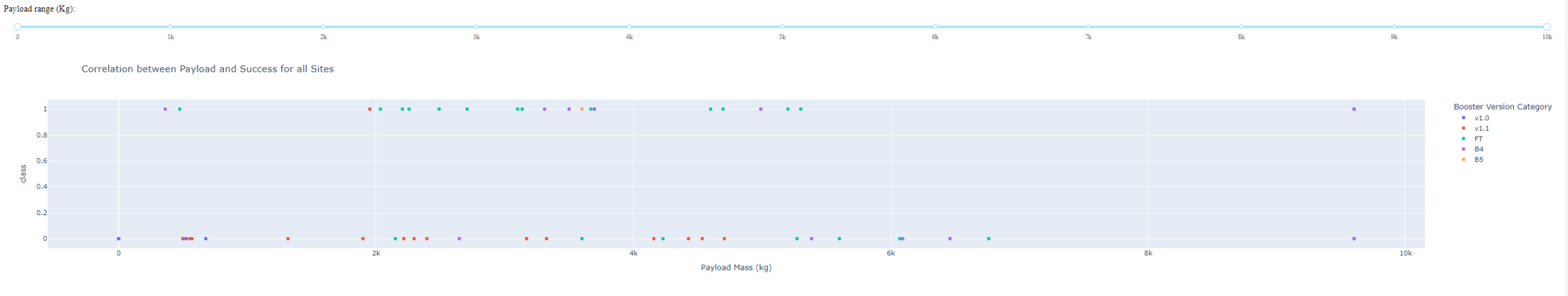
Launch Site With Highest Launch Success Ratio



KSC LC-39A boasts the highest success rate for launches at 76.9%, with 10 successful landings and 3 unsuccessful ones.

GitHub Link: Dashboard with Plotly Dash

Payload vs. Launch Outcome For Different Payloads



Payload Success Rates:
The highest success rate is seen for payloads weighing between 2000 and 5500 kg.

Booster Version FT:
It achieves the highest success rate overall and successfully launches the largest payload, weighing over 9.5 kg.

GitHub Link: [Dashboard with Plotly Dash]

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The test set scores are inconclusive in determining the best method due to the small sample size (18 samples). As a result, we evaluated all methods using the entire dataset.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

- Overall, SVM outperformed the others, achieving the highest Jaccard Score, F1 Score, and Accuracy.

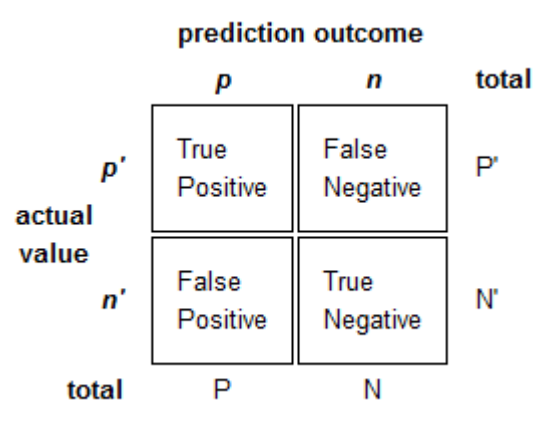
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.816901	0.819444
F1_Score	0.909091	0.916031	0.899225	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

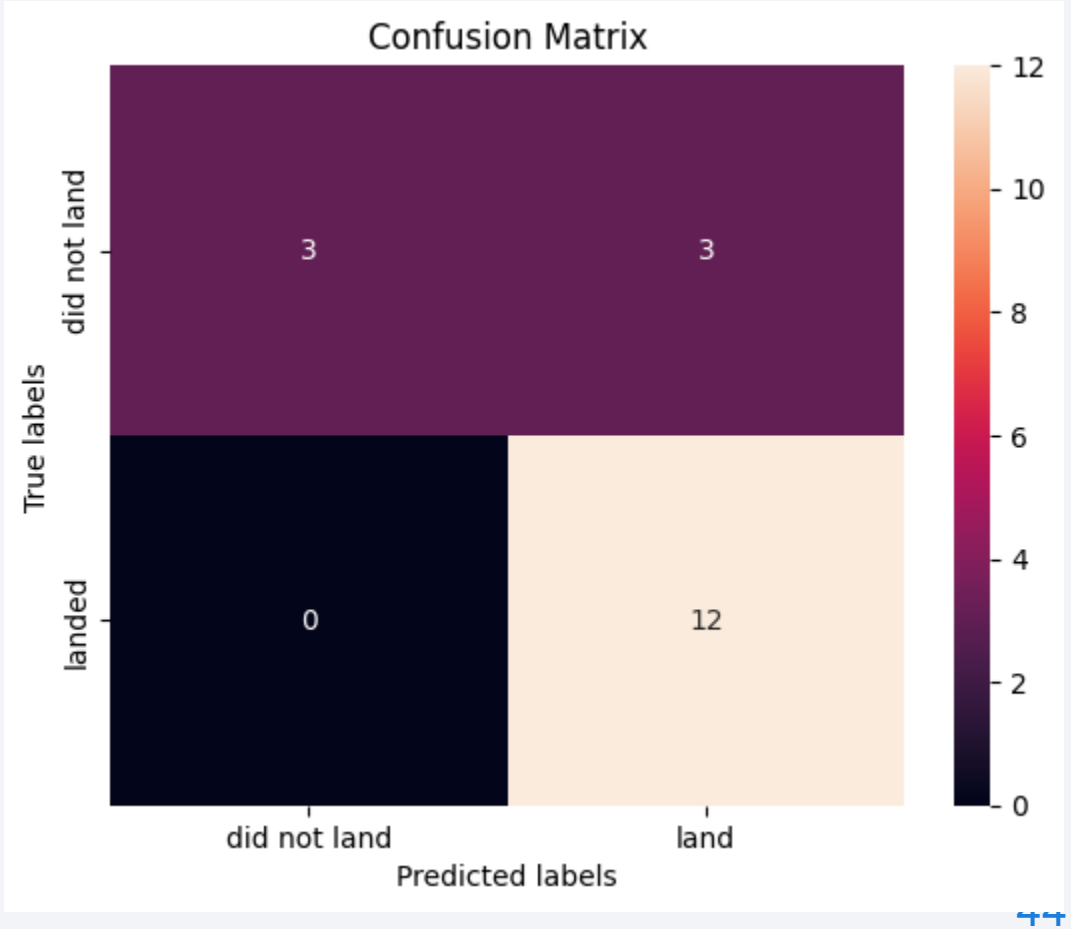
GitHub Link: Predictive Analysis

Confusion Matrix

SVM is capable of distinguishing between the two classes.

The primary issue, however, is the occurrence of False Positives.





GitHub Link: [Predictive Analysis]

Conclusions

Success Trends and Strategic Insights:

- Early Challenges and Recent Improvements: Initial launches encountered difficulties, but recent trends show significant improvements, indicating ongoing learning and refinement.
- Payload Mass and Success Rates:** Heavier payloads generally correlate with higher success rates across various launch sites.
- Orbit Success Rates: Certain orbits, such as ES-L1, GEO, HEO, and SSO, achieved a 100% success rate, while others like GTO, ISS, LEO, MEO, PO, and VLEO had success rates ranging from 50-80%.
- Increasing Success Rates: The overall launch success rate has been steadily increasing since 2013.

Strategic Sites and Performance:

- Ocean-Proximate Launch Sites: Launch sites near oceans prioritize safety through debris mitigation and leverage Earth's rotation (equator proximity) for an extra speed boost.
- Top Performing Sites: KSC LC-39A leads with 41.2% of successful launches, followed by CCAFS SLC-40 (23%), VAFB SLC-4E (21.4%), and CCAFS SLC-41 (14.4%).

Data Exploration and Modeling:

- Visualizing Launch Site Locations: Interactive maps were utilized to visualize the proximity of launch sites to essential infrastructure.
- Effective Predictive Model: The Decision Tree model was identified as the most effective for predicting launch success based on comprehensive dataset analysis.
- Best Algorithm: SVM emerged as the best algorithm for this dataset.

Thank you!