Winning Space Race with Data Science

Rutuparna S Dharan
25th, January 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - EDA with Visualization and SQL
  - Interactive Maps
  - Dashboard
  - Predictive Analytics
- Conclusion
- Appendix

Executive Summary

Summary of methodologies:

Our data-driven approach involved meticulous data collection through the SpaceX REST API and web scraping. Through thorough wrangling, we established a crucial success/failure variable.

Exploratory Data Analysis uncovered trends, with launch success steadily improving over time. KSC LC-39A emerged as the top-performing landing site, and certain orbits boasted a perfect 100% success rate.

Visualization highlighted the strategic locations of launch sites near the equator and coastal areas.

In Predictive Analytics, various models were employed, with the decision tree model showing a slight performance edge on the test set.

Summary of all results:

KSC LC-39A leads with the highest success rate.

Certain orbits consistently achieve a 100% success rate.

Strategic launch site locations near the equator and coast.

Comparable predictive performance among models, with Decision Tree slightly outperforming.

Models accurately predict 83.33% of the success class instances in the imbalanced dataset.

Introduction

Project background and context

As a data scientist at SpaceY, my role centers on utilizing public information to assess the probability of successful first-stage landings for SpaceX's reusable Falcon 9 rockets. The primary goal is to inform strategic bidding against SpaceX by understanding the determinants of first-stage landing success.

SpaceX's innovative approach to space exploration, marked by cost-effective launches due to the reuse of Falcon 9's first stage, has revolutionized the industry. Determining the predictability of first-stage landings is critical for SpaceY's competitive strategy against providers with higher launch costs.

Research Focus:

- How does payload mass, launch site, number of flights, and orbits influence first-stage landing success?
- What is the rate of successful landings over time?
- Which model proves to be the best for predicting successful landings in binary classification?

Section 1

Methodology

Methodology

Executive Summary

Data collection: Data collection involved making requests to SpaceX's API using HTTP and BeautifulSoup. The extracted data was then organized into a structured pandas table for further analysis.

Data Wrangling: Data wrangling was executed using pandas, ensuring that the collected information was properly formatted, cleaned, and structured for analysis.

Exploratory data analysis (EDA): The processed data underwent EDA using a combination of visualization techniques and SQL queries to derive meaningful insights.

Interactive Visual Analytics: Folium and Plotly Dash were utilized for interactive visual analytics, offering dynamic insights into the relationships within the data.

Predictive Analysis: Classification models, including Logistic Regression, SVM, Decision Tree, and KNN, were employed for predictive analysis. The process involved building, tuning, and evaluating these models to achieve optimal predictive performance.

Data Collection

Data Collection - API:

API Request: Requested data from SpaceX API, specifically rocket launch data.

Decode Response: Decoded API response using json().

DataFrame Conversion: Converted decoded data to a dataframe using .json_normalize().

Custom Function Request: Requested launch information from SpaceX API using custom functions.

Dictionary Creation: Created a dictionary from the obtained data.

DataFrame Creation: Created a dataframe from the dictionary.

Filtering Falcon 9 Launches: Filtered the dataframe to include only Falcon 9 launches.

Missing Values Handling: Replaced missing values of Payload Mass with calculated mean().

Export to CSV: Exported the cleaned data to a CSV file.

Flowchart·

Data Collection

Data Collection - Web Scraping:

Request Data: Requested Falcon 9 launch data from Wikipedia.

Create BeautifulSoup Object: Created a BeautifulSoup object from the HTML response.

Extract Column Names: Extracted column names from the HTML table header.

Collect Data from Parsing: Collected data by parsing HTML tables.

Dictionary Creation: Created a dictionary from the collected data.

DataFrame Creation: Created a dataframe from the dictionary.

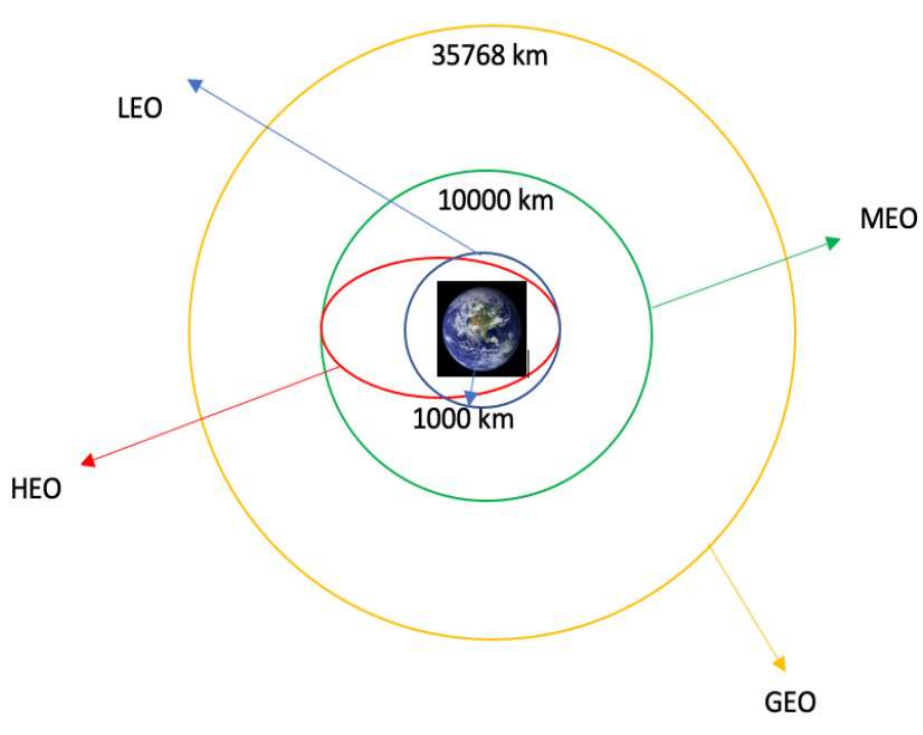
Export to CSV: Exported the data to a CSV file.

Flowchart:

Data Wrangling

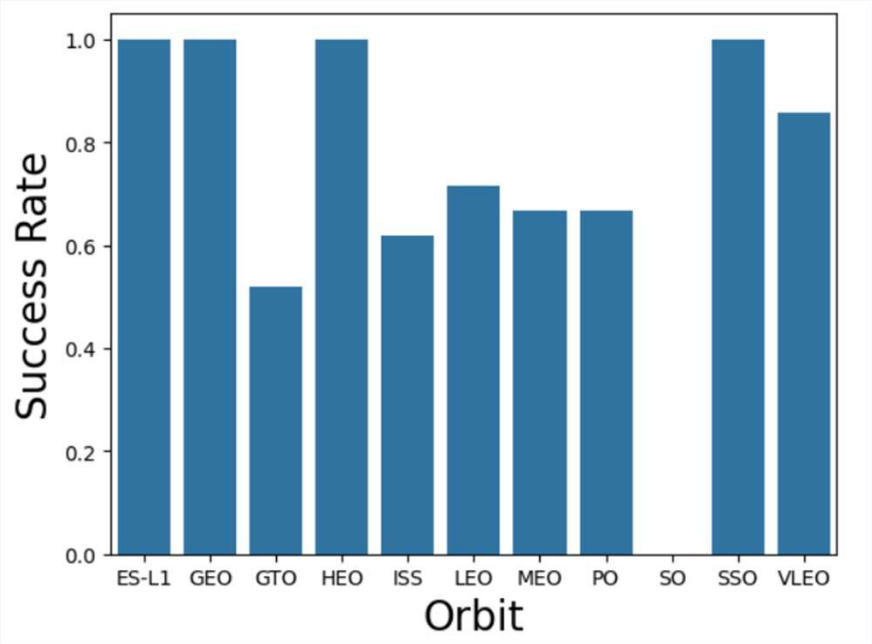
During the exploratory data analysis phase, we meticulously assessed the dataset and derived the training labels. Our analysis included computing the count of launches at each site and documenting the occurrence of different orbits. To enhance the dataset, we engineered a landing outcome label from the existing outcome column, providing a more informative feature for our predictive models.

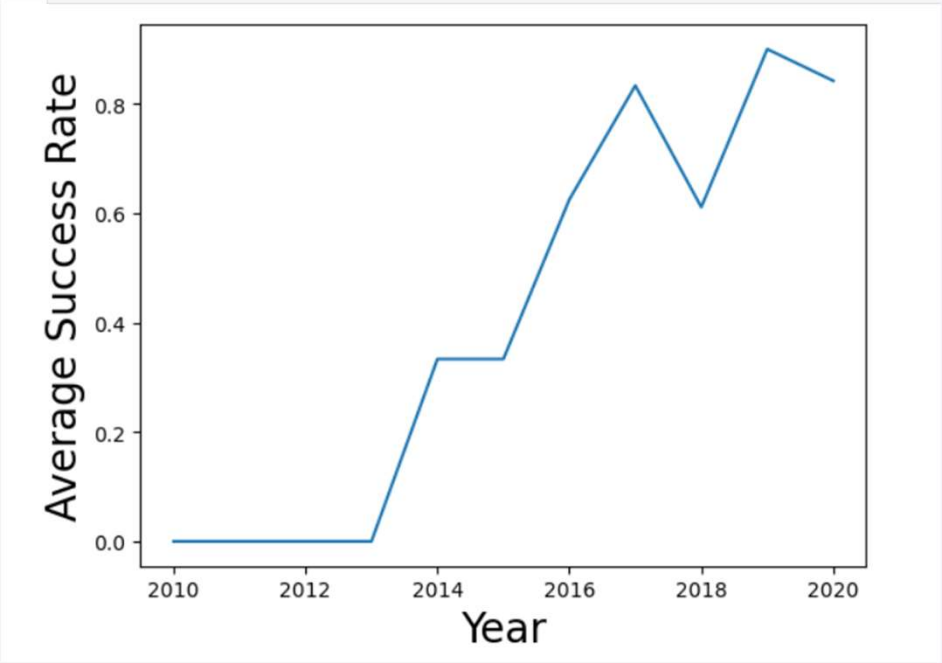
The comprehensive results of our exploration and data manipulation have been exported to a CSV file for further utilization. For a detailed walkthrough of our data wrangling process, you can access the Jupyter notebook on our GitHub repository: Data Wrangling Notebook. This notebook captures the intricacies of our data analysis, showcasing the steps taken to refine and prepare the dataset for subsequent stages in our analysis pipeline.



EDA with Data Visualization

In our data exploration, we visualized key relationships such as flight number with launch site, payload with launch site, success rates for each orbit type, flight number with orbit type, and the yearly trend in launch success. To view these insights, refer to our analysis in the notebook: [Data Exploration Notebook]





EDA with SQL

- Calculated the number of launches at each site using SQL.
- Determined the number and occurrence of each orbit type in the dataset.
- Created a landing outcome label from the outcome column in the SQL query.
- Exported the results, including the calculated values, to a CSV file.

- For a detailed overview of the SQL queries and exploratory data analysis (EDA) process, you can refer to the notebook on our GitHub repository: EDA with SQL Notebook

Build an Interactive Map with Folium

We incorporated various map objects into our Folium map during the interactive visualization process:

Markers: Marked all launch sites on the map to provide a visual representation of their locations.

Circles: Added circles to represent the success or failure of launches for each site, with color-coded markers for outcomes (0 for failure, 1 for success).

Lines: Incorporated lines to highlight connections or specific features related to launch sites.

The addition of these objects was pivotal for conveying meaningful information about launch outcomes, site locations, and their spatial relationships. The markers and circles, in particular, facilitated a clear distinction between successful and unsuccessful launches, aiding in identifying patterns or clusters.

Furthermore, we used the color-labeled marker clusters to discern launch sites with relatively high success rates, offering a quick visual summary of success patterns.

Additionally, we conducted spatial analyses by calculating distances between launch sites and their proximities. This allowed us to answer questions such as whether launch sites are near railways, highways, coastlines, or if they maintain a certain distance from cities.

GitHub repository: Interactive Map with Folium

Build a Dashboard with Plotly Dash

In developing the Plotly Dash interactive dashboard, we implemented several plots and interactions to facilitate a comprehensive exploration of SpaceX launch data:

Pie Charts: Integrated pie charts to visually represent the distribution of total launches across different launch sites. This offers a quick and intuitive overview of each site's contribution to the overall launch count.

Scatter Graph: Utilized a scatter graph to illustrate the relationship between launch outcomes and payload mass (Kg) for various booster versions. This dynamic visualization aids in identifying correlations or trends between these critical factors.

The incorporation of these specific plots and interactions aimed to enhance user engagement and provide a streamlined interface for users to explore key insights within the SpaceX launch dataset.

GitHub repository: Plotly Dash Lab

Predictive Analysis (Classification)

In constructing and refining our classification model, the following key steps and techniques were employed:

Data Loading and Transformation: Loaded the data using NumPy and pandas. Transformed the data to prepare it for model development.

Data Splitting: Segmented the dataset into training and testing sets, ensuring an appropriate division for model training and evaluation.

Model Building: Developed various machine learning models to assess their performance.

Metric Selection: Utilized accuracy as the primary metric to evaluate model performance.

Identification of Best Performing Model: Evaluated multiple models and identified the one demonstrating superior performance.

GitHub repository: Predictive Analysis Lab

Results

Exploratory Data Analysis Results:

Improved success trend observed in Falcon 9 first-stage landings.

KSC LC-39A leads with the highest success rate.

Certain orbits consistently achieve a 100% success rate.

Strategic launch site locations near the equator and coast.

Comparable predictive performance among models, with Decision Tree slightly outperforming.

Interactive Analytics Demo in Screenshots:

Visual representation of launch success rates at various sites.

Exploration of launch sites with the most success and successful payload ranges.

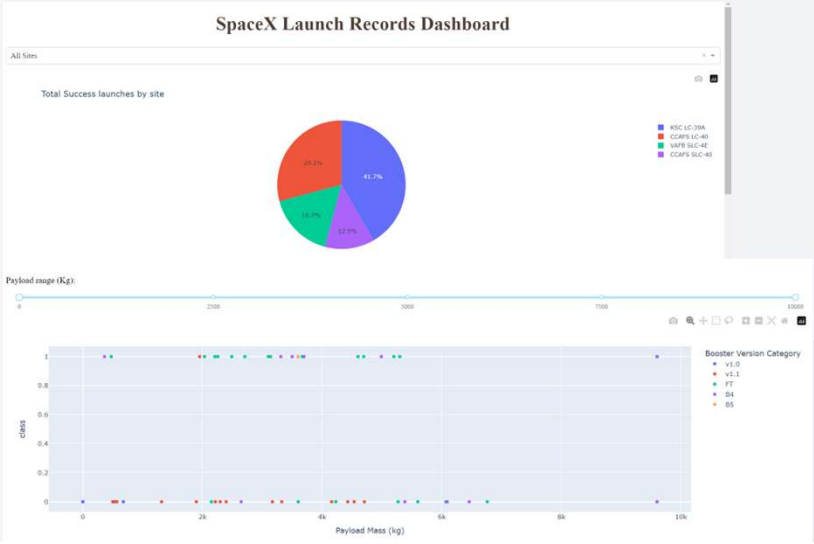
A dynamic view of the relationship between payload mass, orbit types, and launch success.

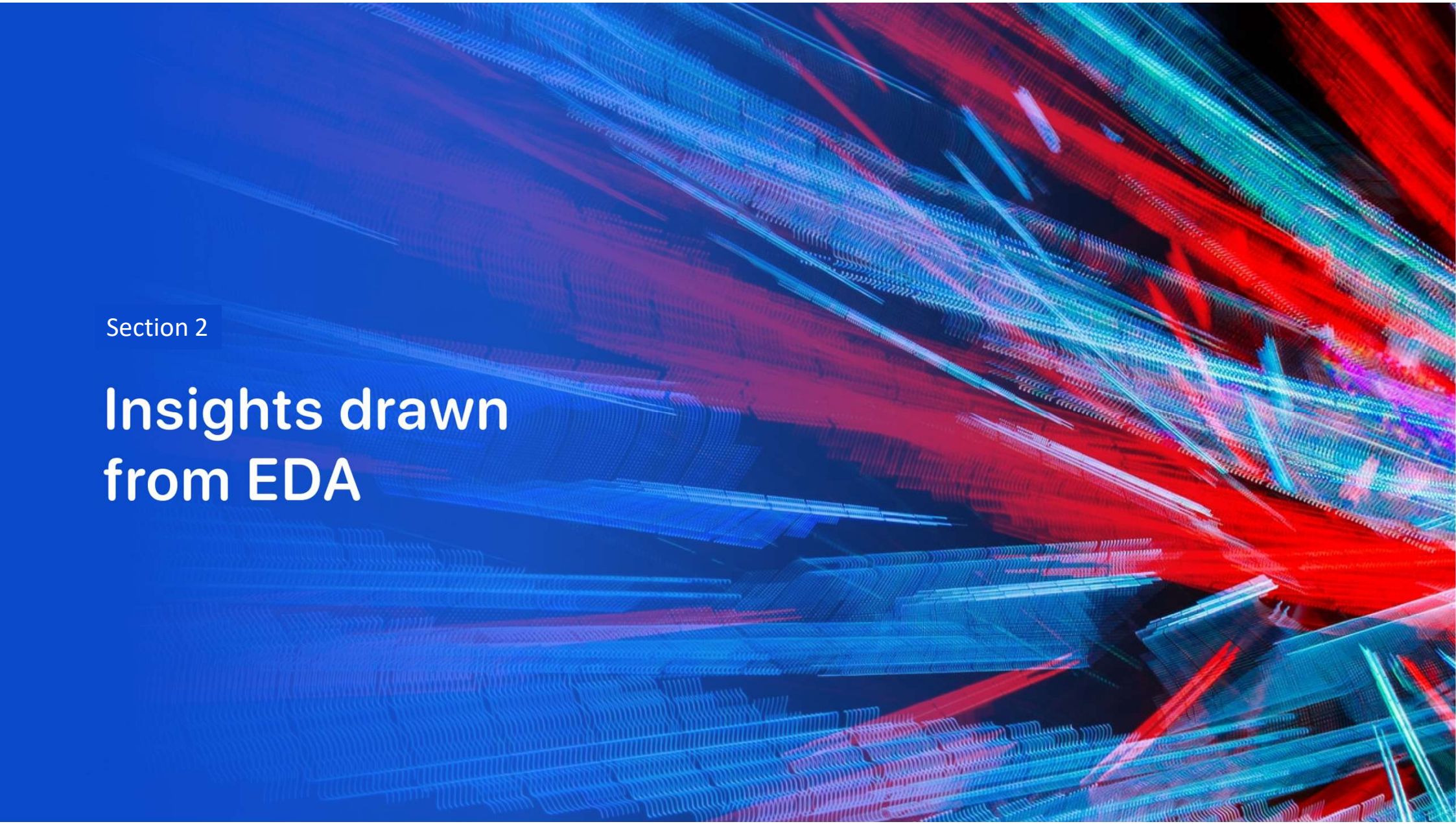
Predictive Analysis Results:

Building models using Logistic Regression, SVM, Decision Tree, and KNN.

Hyperparameter tuning with GridSearchCV to enhance model performance.

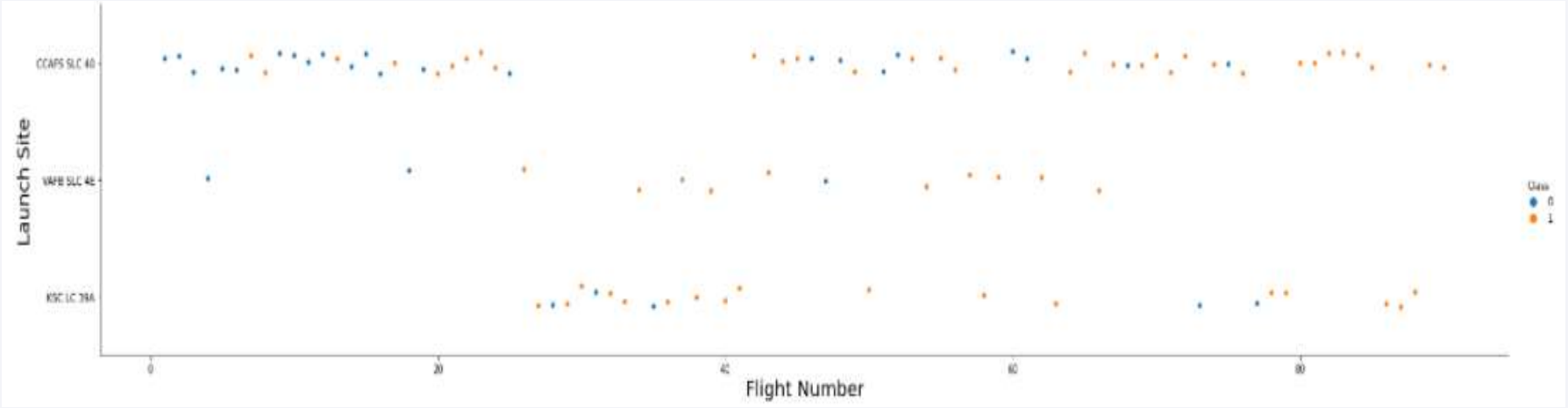
Models accurately predict 83.33% of the success class instances in the imbalanced dataset.




Section 2
Insights drawn
from EDA

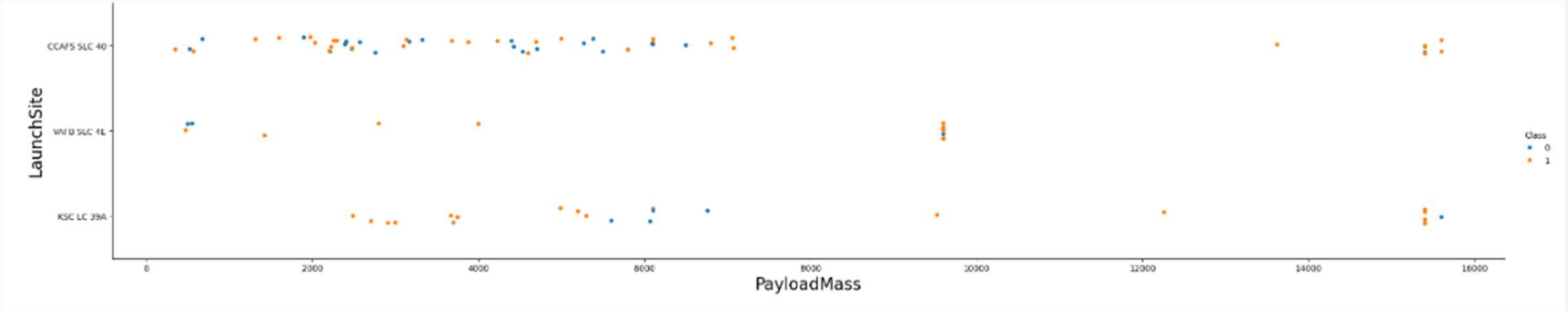
Flight Number vs. Launch Site

The plot suggests that as the flight count increases, the success rate tends to be higher. This finding implies a potential relationship between the launch site's experience or operational history, as reflected by the number of flights, and the likelihood of successful outcomes. The larger the flight amount, the greater the success rate.



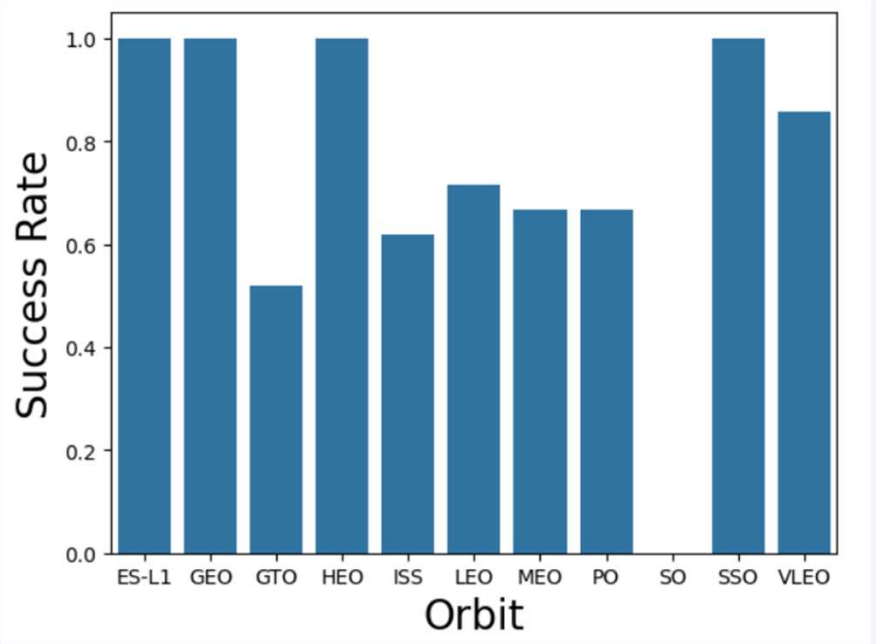
Payload vs. Launch Site

Specifically, for the launch site VAFB-SLC, there are no recorded rocket launches for heavy payload masses (greater than 10,000 kg). This observation suggests a potential limitation or operational constraint at the VAFB-SLC launch site regarding handling rockets with heavier payload masses.



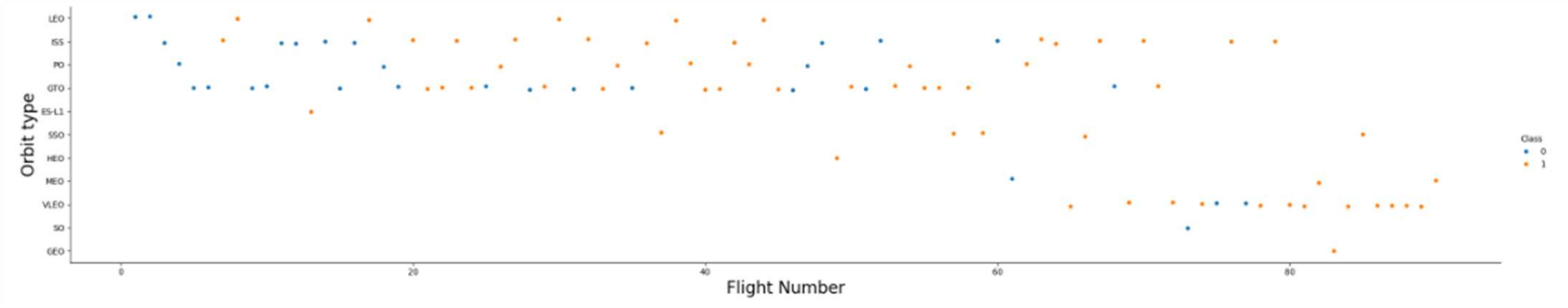
Success Rate vs. Orbit Type

The analysis of the plot highlights that the orbit types ES-L1, GEO, HEO, SSO, and VLEO exhibit the highest success rates. This observation suggests that rockets targeting these specific orbital trajectories have been more consistently successful in their first-stage landings.



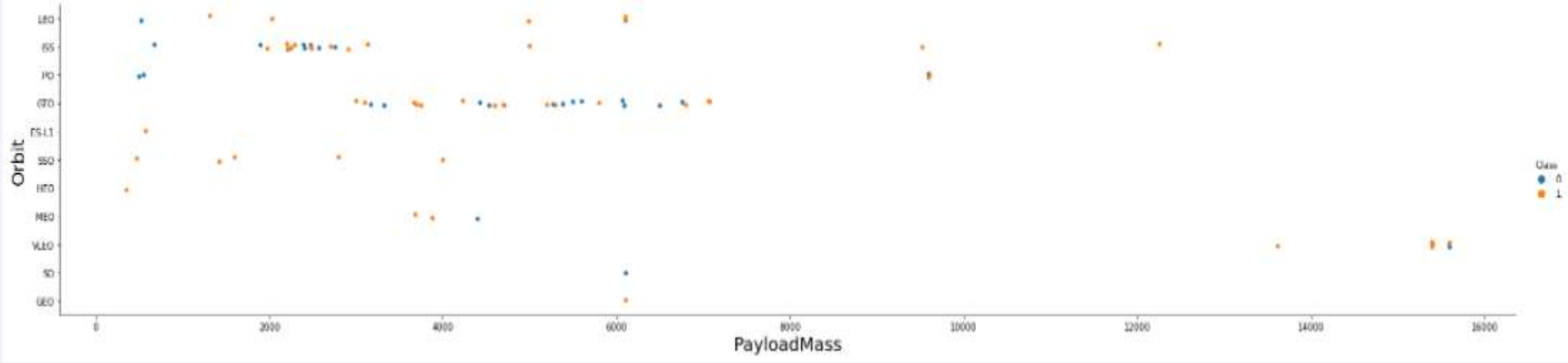
Flight Number vs. Orbit Type

In the Flight Number vs. Orbit Type plot, success in Low Earth Orbit (LEO) is correlated with increasing flight numbers. However, in the Geostationary Transfer Orbit (GTO), there is no discernible relationship between flight number and success.



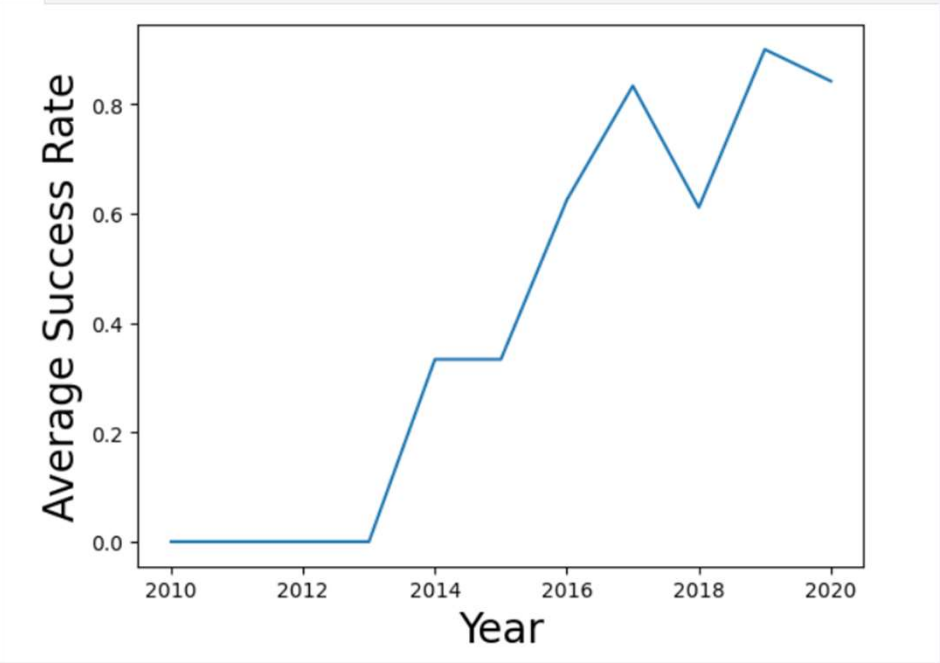
Payload vs. Orbit Type

The observation indicates that heavy payloads are associated with a higher likelihood of successful landings, particularly for Polar Orbit (PO), Low Earth Orbit (LEO), and International Space Station (ISS) orbits.



Launch Success Yearly Trend

We can observe a consistent increase in the success rate from 2013 to 2020. This trend suggests an improvement in the overall success of rocket launches over the specified time period.



All Launch Site Names

DISTINCT in the SQL query enabled the display of only unique launch sites from the SpaceX data. This approach ensures that each launch site is represented only once in the query result, providing a clear and concise list of distinct launch locations.

```
%sql select distinct LAUNCH_SITE from SPACEXTBL;
```

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

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

Launch Site Names Begin with 'CCA'

This query was employed to retrieve and display five records where launch sites begin with the prefix `CCA`. This specific use of the SQL query presents information related to launch sites that share a common starting sequence.

```
11]: %sql select * from SPACEXTBL where LAUNCH_SITE like "CCA%" limit 5;
 * sqlite:///my_data1.db
Done.
```

11]:

ate	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
10--04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
10--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)
12--22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
12--08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
13--01	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 query used successfully calculated the total payload carried by boosters from NASA, resulting in a total of 45,596. This SQL operation efficiently aggregated and summed the payload values associated with NASA boosters.

```
In [14]: %sql select Customer, sum(PAYLOAD_MASS__KG_) as Total_NASA_CRS_mass from SPACEXTBL where Customer = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Customer	Total_NASA_CRS_mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

The mentioned query effectively calculated the average payload mass carried by the booster version F9 v1.1, yielding a value of 2,928.4. This SQL operation aggregated the payload masses associated with F9 v1.1 boosters and computed the average.

```
]: %sql select Booster_Version     , avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version = "F9 v1.1";
 * sqlite:///my_data1.db
Done.
```

]:

Booster_Version	avg(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4

First Successful Ground Landing Date

The date of the first successful landing outcome on a ground pad was December 22, 2015. This specific date marks a significant milestone in SpaceX's achievements, capturing the initial success of landing a rocket on solid ground

```
16]: %sql select  Mission_Outcome, min(Date) as Date_First_Land from SPACEXTBL where Landing_Outcome ='Success (ground pad)';
 * sqlite:///my_data1.db
Done.
```

16]:

Mission_Outcome	Date_First_Land
Success	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The use of the WHERE clause, combined with the AND condition, allowed for the filtering of boosters that have successfully landed on a drone ship with a payload mass greater than 4000 and less than 6000.

```
%sql select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTBL where (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_M
```

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

Booster Version	Landing Outcome	PAYLOAD MASS KG
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

We filter records where the Mission Outcome was either a success or a failure.

```
In [18]: %sql select Mission_Outcome, count(Mission_Outcome) as "Total (Success or failure)" from SPACEXTBL GROUP BY MISSION_OUTCOME
 * sqlite:///my_data1.db
Done.
```

Out[18]:

Mission_Outcome	Total (Success or failure)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The use of a subquery in the WHERE clause, combined with the MAX() function, enabled the determination of the booster that has carried the maximum payload. This query efficiently identified the specific booster associated with the highest payload value in the dataset.

```
In [19]: %sql select Booster_Version,Landing_Outcome, PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOA
 * sqlite:///my_data1.db
Done.
```

Out[19]:

Booster Version	Landing Outcome	PAYLOAD MASS KG
F9 B5 B1048.4	Success	15600
F9 B5 B1049.4	Success	15600
F9 B5 B1051.3	Success	15600
F9 B5 B1056.4	Failure	15600
F9 B5 B1048.5	Failure	15600
F9 B5 B1051.4	Success	15600
F9 B5 B1049.5	Success	15600
F9 B5 B1060.2	Success	15600
F9 B5 B1058.3	Success	15600
F9 B5 B1051.6	Success	15600
F9 B5 B1060.3	Success	15600
F9 B5 B1049.7	Success	15600

2015 Launch Records

The combination of the WHERE clause, LIKE, AND, and BETWEEN conditions was employed to filter for failed landing outcomes on drone ships, including information about their booster versions and launch site names for the year 2015.

```
%sql select  Date, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTBL where Landing_Outcome= 'Failure (drone ship
```

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

Date	Booster_Version	Launch_Site	Landing_Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

The query effectively selected landing outcomes and the count of landing outcomes from the data, applying the WHERE clause to filter for landing outcomes between June 4, 2010, and March 20, 2010. Additionally, the query utilized the GROUP BY clause to group the landing outcomes and the ORDER BY clause to arrange the grouped outcomes in descending order.

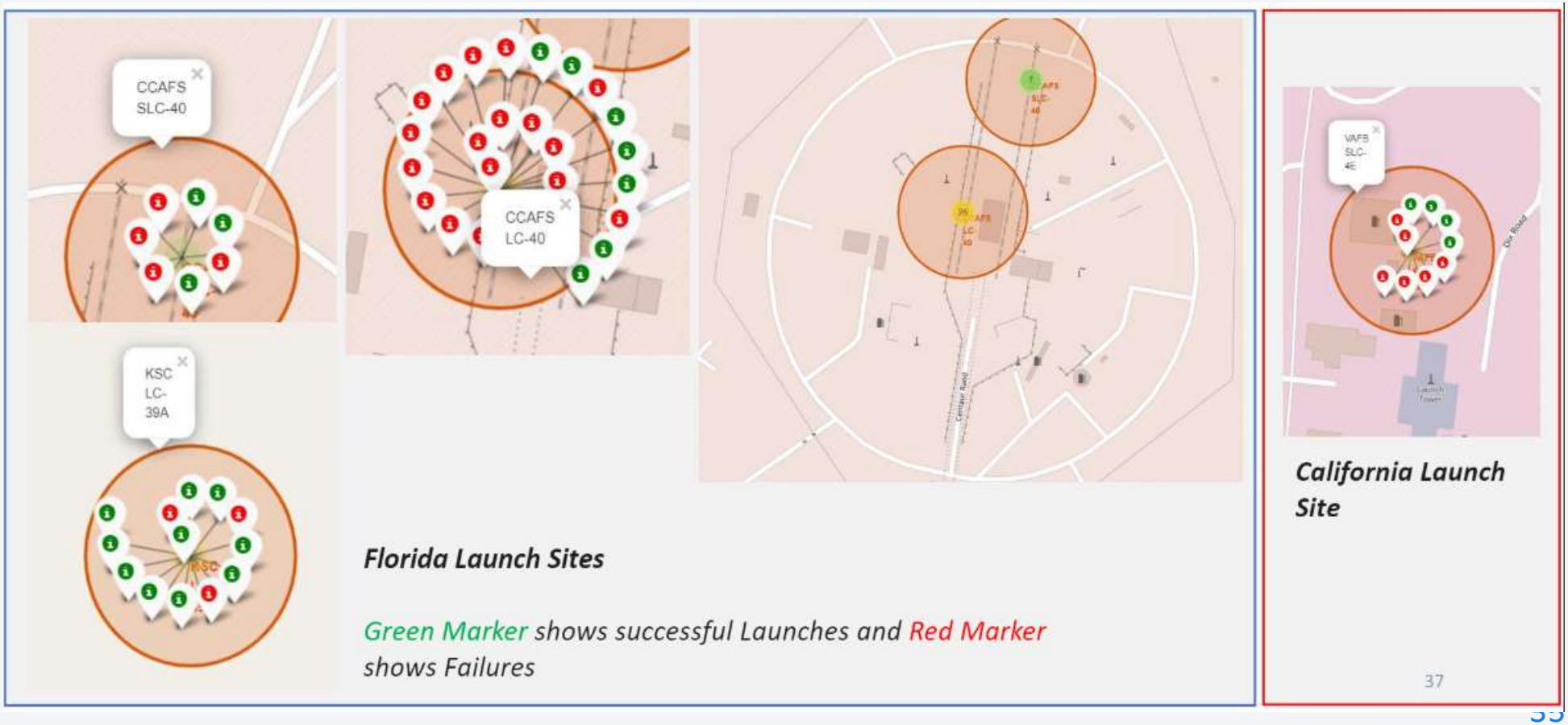
```
%sql select Landing_Outcome, count(Landing_Outcome) as "Total Count" from SPACEXTBL where Landing_Outcome = "Failure (drone
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

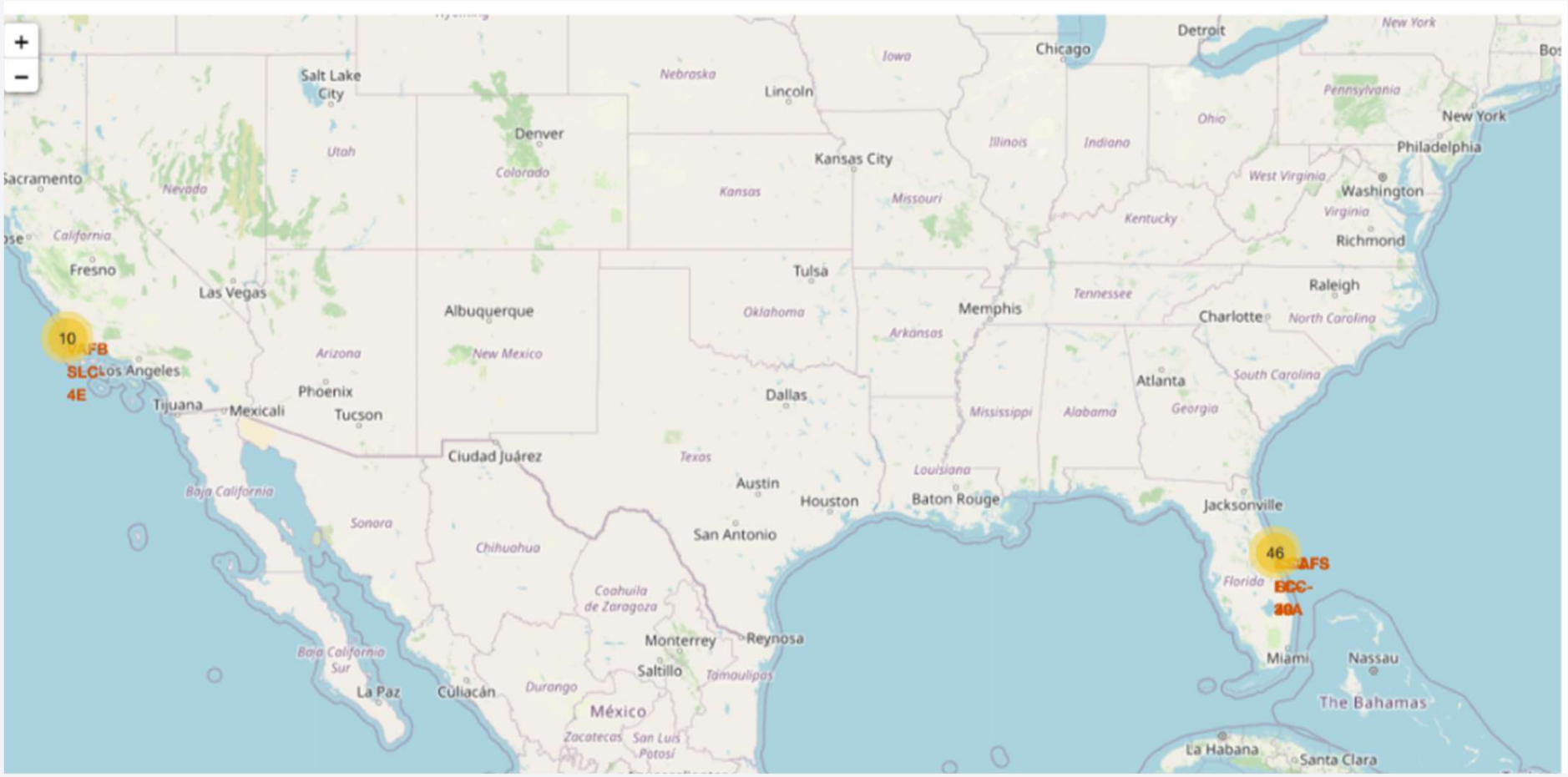

Section 3
Launch Sites
Proximities Analysis

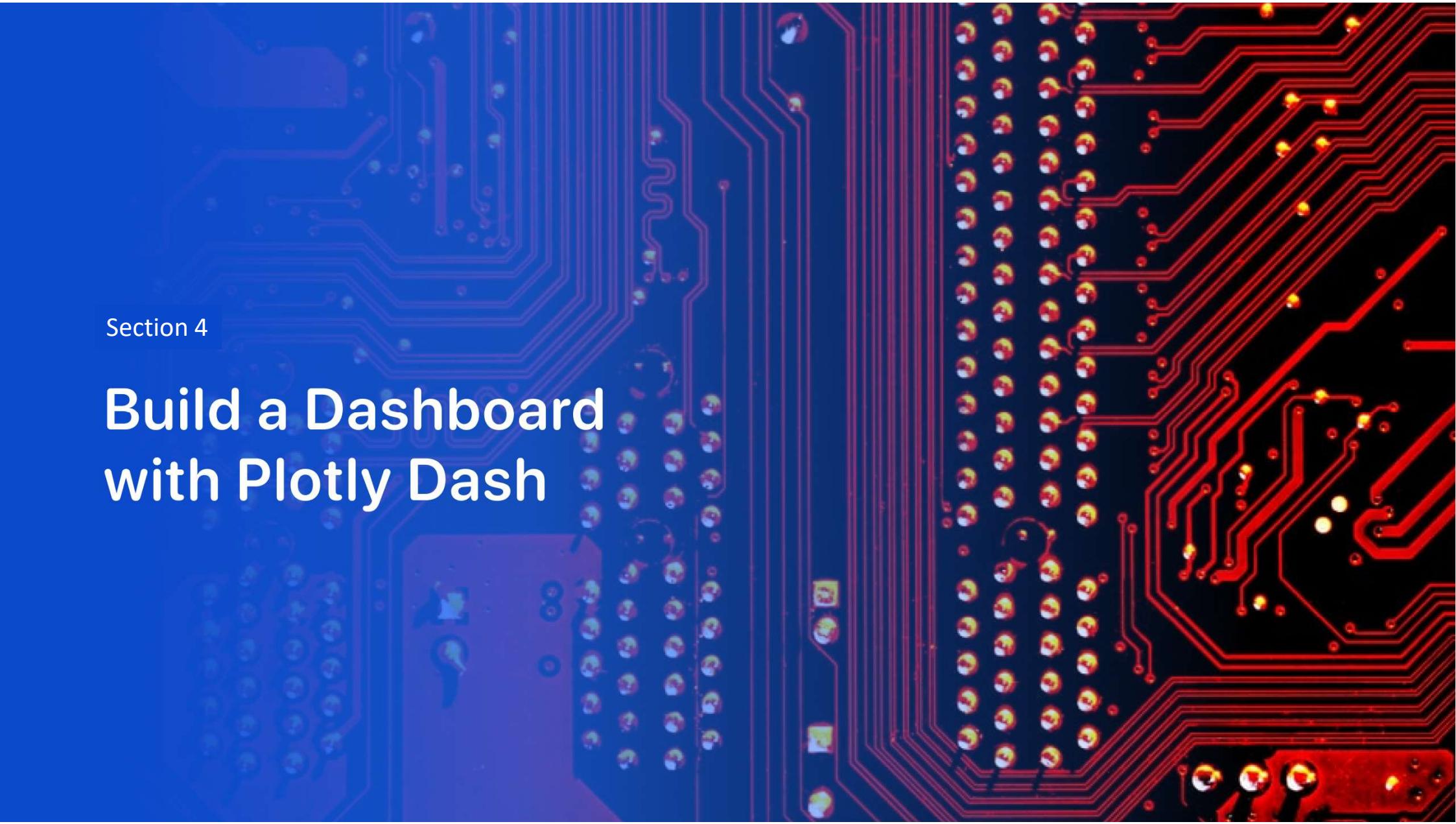
All launch sites global map markers

Markers showing launch sites with color labels



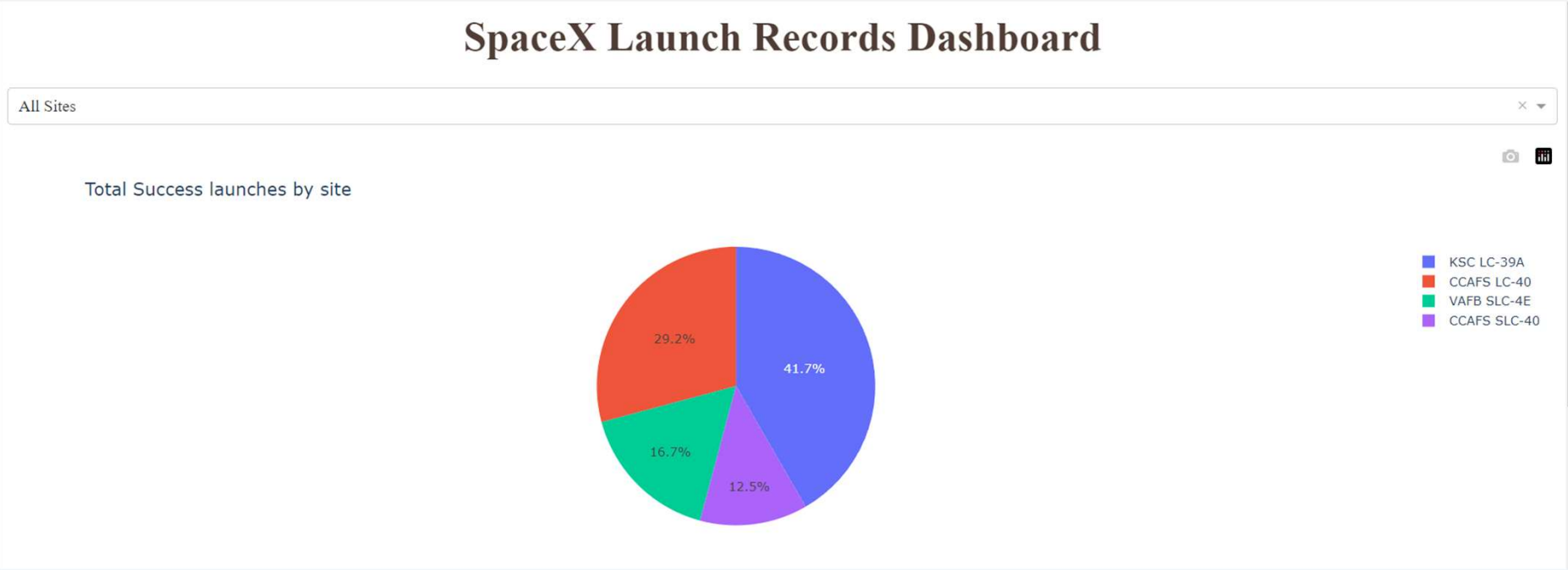
Markers Highlighting the Launch Site




Section 4
Build a Dashboard with Plotly Dash

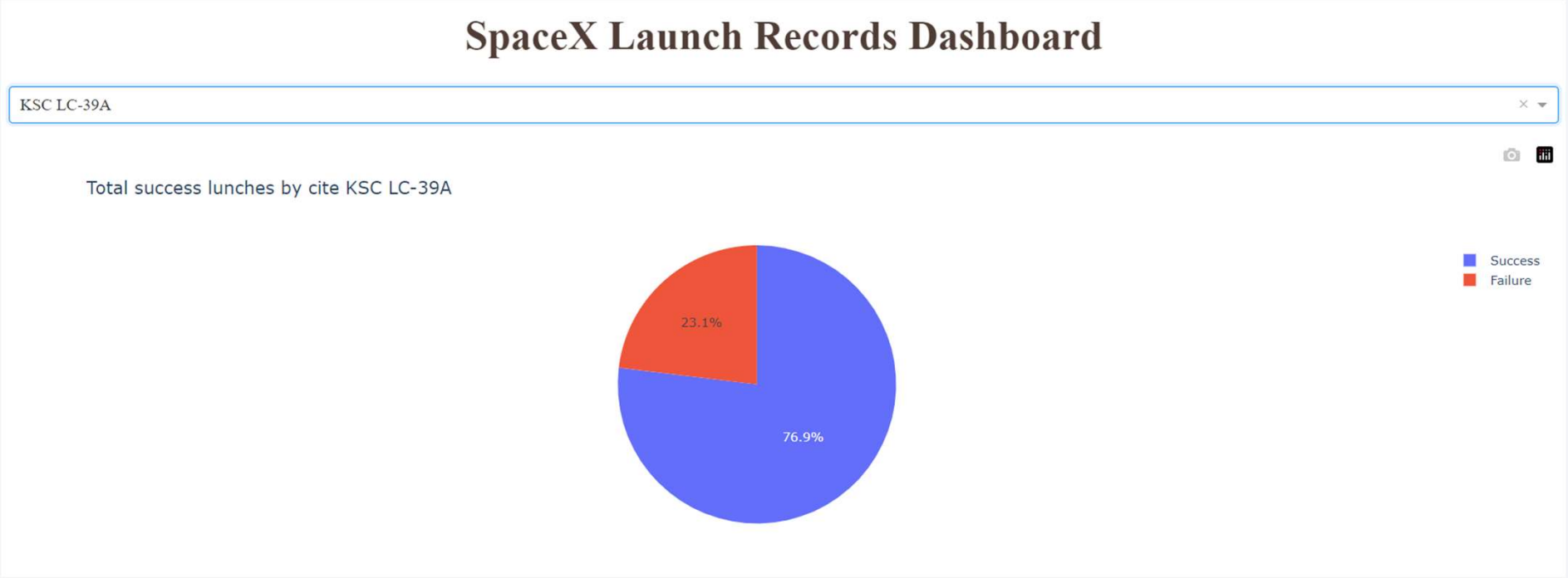
Pie chart showing the success percentage achieved by each launch site

The observation suggests that KSC LC-39A stands out with the most successful launches among all the launch sites. This insight emphasizes the historical success of launches conducted from KSC LC-39A and may indicate the site's efficiency and reliability in terms of achieving successful outcomes.



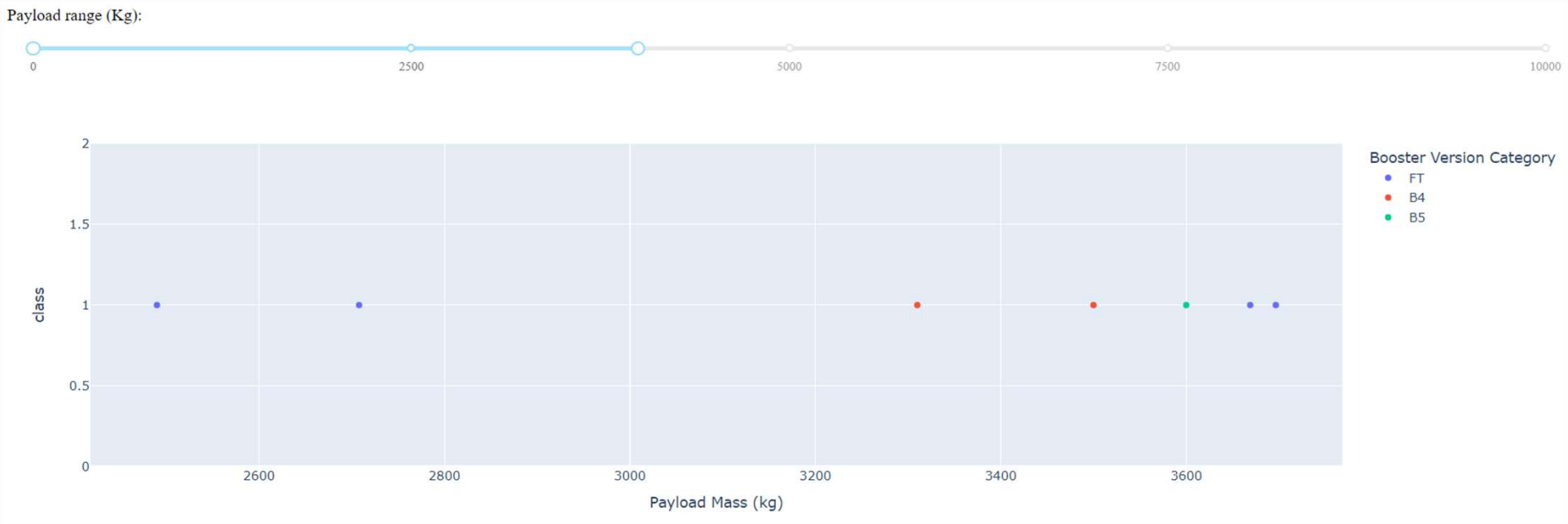
Pie chart showing the Launch site with the highest launch success ratio

The reported statistics indicate that KSC LC-39A achieved a success rate of 76.9% and a failure rate of 23.1%. These success and failure rates provide a quantitative measure of the site's performance in terms of successful rocket launches.



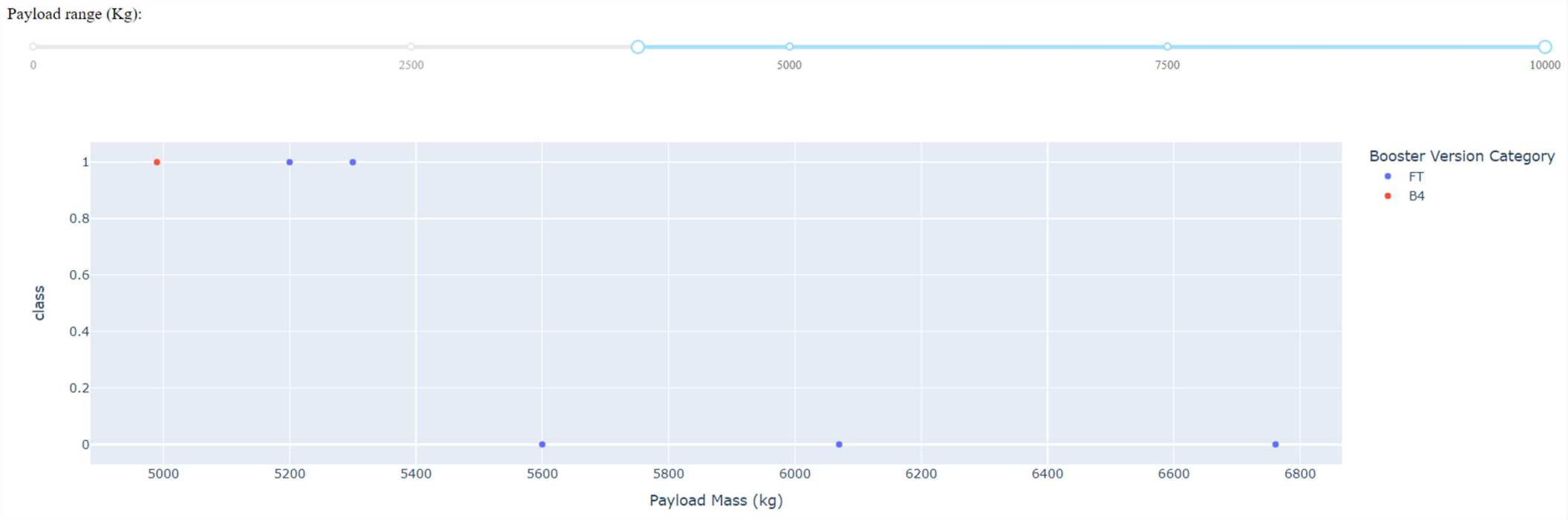
Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

Low Weighted Payload 0kg - 4000kg



Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider

Heavy Weighted Payload 4000kg - 10000kg

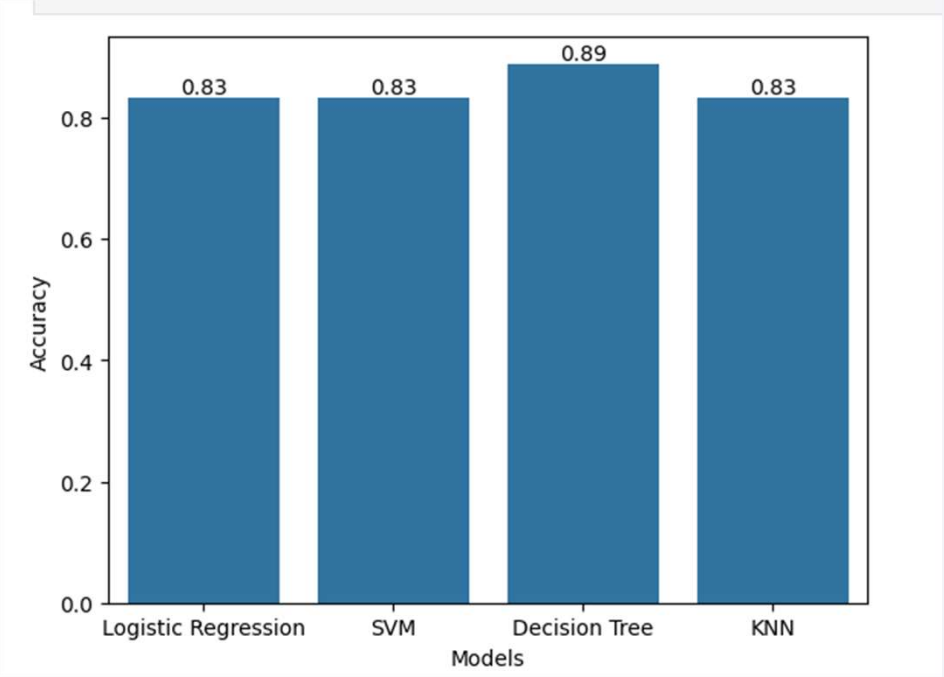


Section 5

Predictive Analysis (Classification)

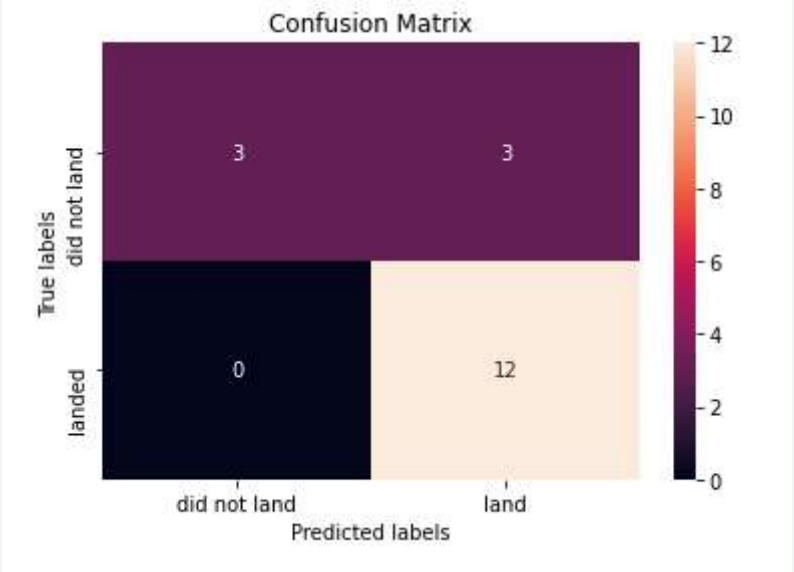
Classification Accuracy

Among the machine learning models evaluated, the Decision Tree classifier achieved the highest classification accuracy. This highlights the effectiveness of the Decision Tree model in accurately predicting and classifying outcomes in the context of the specific task or dataset under consideration.



Confusion Matrix

The analysis of the confusion matrix for the Decision Tree classifier reveals that the classifier is capable of distinguishing between different classes. However, a notable issue lies in false positives, where unsuccessful landings are incorrectly classified as successful landings. This observation suggests a specific area for improvement in the classifier's performance.



Conclusions

The conclusions drawn from the analysis are:

- A positive correlation exists between the number of flights at a launch site and the success rate at that site.
- Launch success rates have exhibited an increasing trend from 2013 to 2020.
- Orbits ES-L1, GEO, HEO, SSO, and VLEO have shown the highest success rates.
- KSC LC-39A stands out with the most successful launches among all sites.
- The Decision Tree classifier is identified as the most effective machine learning algorithm for the given task.

Thank you!