IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Karim Nasr
March 14th, 2024

Outline

- **Executive Summary**
  - Data-driven approach for Falcon 9 landing prediction
  - Methodologies: Data Collection, Wrangling, EDA, Visual Analytics, Predictive Analysis
  - Key Results: Model Performance, Insights from EDA, Interactive Visualizations
- **Introduction**
  - Background: SpaceX and Falcon 9's importance in commercial spaceflight
  - Project Goals: Predicting landing success, identifying influential factors, improving prediction accuracy
- **Methodology**
  - Data Collection: SpaceX API, Web Scraping
  - Data Wrangling: Cleaning, Labeling, Assessing Data Quality
  - EDA: Visualization, SQL Queries
  - Predictive Analysis: Classification Models, Hyperparameter Tuning

Outline

- **Results**
  - EDA Findings: Correlations, Patterns, Success Rates
  - Interactive Analytics: Geospatial Visualization, Dashboard Creation
  - Predictive Analysis: Model Development, Evaluation, Best Model Selection
- **Conclusions**
  - Model Comparisons: Logistic Regression, SVM, Decision Tree
  - Prediction Accuracy: Around 83.33% on test data
  - Future Work: Model Refinement, Additional Data Sources
- **Appendix**
  - GitHub Repository: Full Project, Notebooks, Plots, and Charts

Executive Summary

- **Methodologies Employed:** The project was anchored on a data-driven approach, which was meticulously designed to predict the success of Falcon 9 first stage landings.The methodologies employed were as follows:
  - **Data Collection:** The project began with data collection, which was primarily done through the SpaceX API and web scraping. This ensured that the most recent and relevant data was used for the analysis.
  - **Data Wrangling:** The collected data underwent a rigorous data wrangling process to ensure its quality and reliability. This involved cleaning the data, handling missing values, and transforming the data into a format suitable for analysis.
  - **Exploratory Data Analysis (EDA):** EDA was conducted using visualization and SQL to understand the underlying patterns and trends in the data. This step was crucial in identifying the key factors that influence the success of the landings.
  - **Interactive Visual Analytics:** The project also incorporated interactive visual analytics using tools like Folium and Plotly Dash. These tools provided dynamic data representation, which greatly enhanced the strategic decision-making process.
  - **Predictive Analysis:** The final step involved predictive analysis using machine learning classification models. The models were trained and tested on the processed data, with the aim of predicting the success of Falcon 9 first stage landings.

Executive Summary

- **Key Results:** The project yielded several key results that are of significant value to the field of aerospace data science:
  - **Predictive Model Performance:** The Decision Tree Classifier model, which was used for the predictive analysis, achieved an accuracy of 83.33% on the test data. This high accuracy rate indicates a strong correlation between the factors considered and the success of the landings.
  - **Insights from EDA:** The EDA revealed several interesting patterns. For instance, it was found that the payload mass and the launch site have a significant influence on the success rates. These insights can be instrumental in planning future missions.
  - **Interactive Visualizations:** The interactive visualizations developed using Folium and Plotly Dash were highly effective in representing the data dynamically6. These visualizations can be used to make strategic decisions in the commercial spaceflight market.
- **In conclusion,** this project successfully developed a predictive tool for SpaceX's Falcon 9 landings. The detailed analysis, coupled with the interactive visualizations, underscore the project's contribution to advancing the field of aerospace data science. The methodologies and results encapsulated in this summary highlight the project's success and its potential impact on future space missions.

Introduction

- **Project Background and Context**
  - SpaceX, a private American aerospace manufacturer and space transportation company, has revolutionized the economics of space travel with its Falcon 9 rocket. The Falcon 9's first stage is reusable, which significantly reduces the cost of each launch. This cost-effectiveness has given SpaceX a competitive edge in the commercial spaceflight market.
  - However, the successful landing of the Falcon 9's first stage is not guaranteed. The ability to predict the success of these landings is of great interest to SpaceX and other stakeholders. Accurate predictions could help determine the cost of a launch and could be particularly useful for alternate companies considering bidding against SpaceX for a rocket launch.

Introduction

- **Problems We Want to Find Answers For**
  - In this capstone project, we aim to answer the following questions:
  - 1. Can we predict if the Falcon 9 first stage will land successfully?
  - 2. What factors most influence the success or failure of these landings?
  - 3. How can we use machine learning to improve the accuracy of our predictions?

To answer these questions, we will follow a comprehensive data science workflow. This workflow includes data collection and cleaning, exploratory data analysis, interactive visual analytics, predictive analysis using machine learning, and finally, presenting our data-driven insights. Our goal is to develop a robust predictive model that can accurately determine the success of Falcon 9 first stage landings. This model could potentially assist in strategic decision-making in the commercial spaceflight market.

Let's embark on this exciting journey of data exploration, analysis, and prediction!

Section 1

Methodology

Methodology

- **Data Collection:**
  - Utilized SpaceX API and web scraping to gather comprehensive data on rocket, launchpad, payload, and core details. Ensured thorough data extraction and organization for analysis.
- **Data Wrangling:**
  - Conducted data cleaning and formatting, converting mission outcomes into binary labels for supervised learning, and assessed data quality by analyzing missing values and data types.
- **EDA with Visualization and SQL:**
  - Implemented exploratory data analysis to identify patterns and relationships using visual tools and SQL queries, extracting insights from launch data.

Methodology

- **Interactive Visual Analytics:**
  - Employed Folium and Plotly Dash for dynamic data representation, enabling interactive exploration of launch sites, payload masses, and success rates.
- **Predictive Analysis with Classification Models:**
  - Developed and refined classification models through exploratory data analysis, preprocessing, model selection, hyperparameter tuning, and evaluation to predict Falcon 9 first stage landing outcomes.

Data Collection

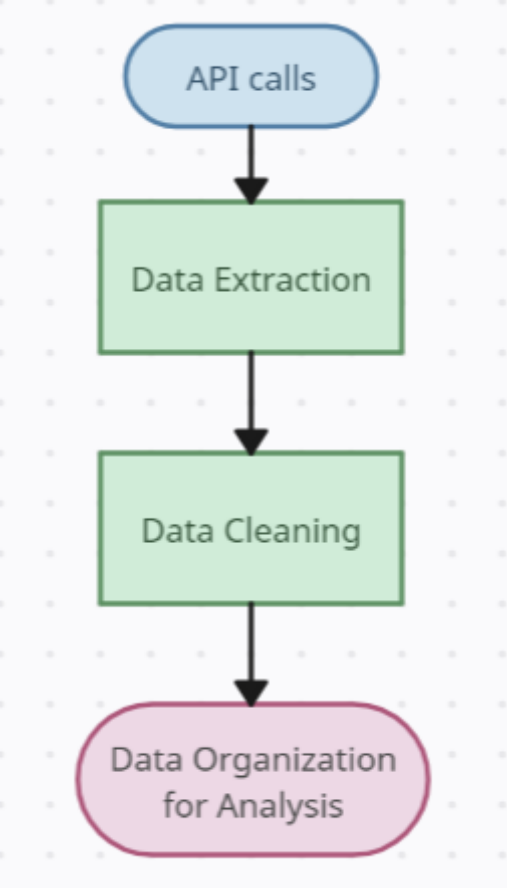
- The data sets were collected through two distinct methods: utilizing the SpaceX API and web scraping.
- **1. SpaceX API Data Collection:**
  - API Requests: Data was gathered using SpaceX API to collect information on rocket, launchpad, payload, and core data.
  - Data Wrangling: This step involved cleaning and formatting the collected data to make it suitable for analysis.
  - Helper Functions: These are defined to assist in extracting relevant information from the API.
  - Analysis Preparation: Ensuring that the data is ready for landing prediction analysis.
  - The flowchart shows a process starting with “API calls” leading to “Data Extraction,” followed by “Data Cleaning,” and ending at “Data Organization for Analysis.”

Data Collection

- **2. Web Scraping Data Collection:**
  - Request Page: An HTTP GET request was made to retrieve the HTML content of a specific Wikipedia page.
  - Parse HTML: BeautifulSoup was used to parse the HTML content and extract relevant information.
  - Extract Data: Specific data points such as flight number, date, payload, and outcome were extracted from the HTML table.
  - Create DataFrame: The extracted data was then organized into a Pandas DataFrame for analysis.
  - The flowchart illustrates a process beginning with “Request Page,” moving on to “Parse HTML,” then “Extract Data,” and concluding at “Create DataFrame.”

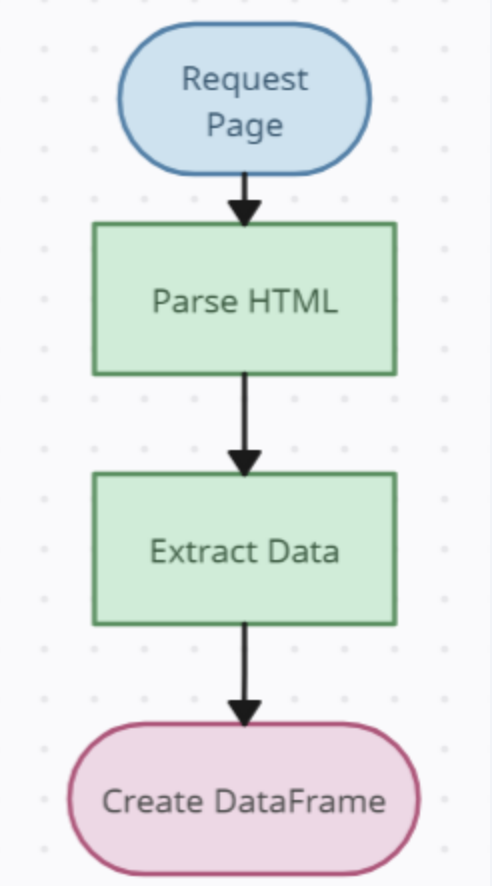
Data Collection – SpaceX API

- **Process**
  - API Requests: Using SpaceX API to gather rocket, launchpad, payload, and core data.
  - Data Wrangling: Cleaning and formatting the collected data.
  - Helper Functions: Defined to assist in data extraction from the API.
  - Analysis Preparation: Ensuring data is ready for landing prediction analysis.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/9f4b6f75ab9bc2f1ae619d5e8d50011c852715ea/jupyter-labs-spacex-data-collection-api.ipynb



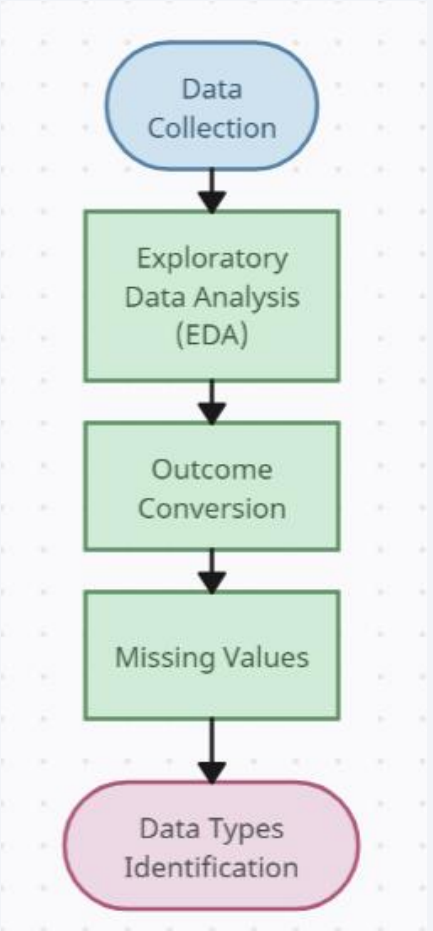
Data Collection – Scraping

- **Process**
  - Request Page: An HTTP GET request was made to retrieve the HTML content of the Wikipedia page.
  - Parse HTML: BeautifulSoup was used to parse the HTML content and extract relevant information.
  - Extract Data: Specific data points such as flight number, date, payload, and outcome were extracted from the HTML table.
  - Create DataFrame: The extracted data was then organized into a Pandas DataFrame for analysis.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/1894e2054eec49460cbeab9febe2346411c47168/2-jupyter-labs-webscraping.ipynb



Data Wrangling

- **Process**
  - Exploratory Data Analysis (EDA): Performed to identify patterns and determine labels for supervised models.
  - Outcome Conversion: Mission outcomes were converted into binary training labels (1 for successful landing, 0 for unsuccessful)1.
  - Missing Values: Calculated the percentage of missing values in each attribute to assess data quality.
  - Data Types Identification: Determined which columns are numerical and categorical for further analysis.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/300bd76246e2777852e5c7f6a366778f5bbaf4e2/3-labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- **Summary of the charts plotted and their purpose**
  - **Flight Number vs. Payload Mass:** This scatter plot visualizes the relationship between each flight's number and its payload mass, highlighting how these factors may influence the success of the Falcon 9 first stage landing.
  - **Launch Site Success Rates:** A bar chart showing the success rates of different launch sites (CCAFS LC-40, KSC LC-39A, VAFB SLC 4E) to analyze how location impacts landing outcomes.
  - **Flight Number vs. Launch Site:** Another scatter plot to explore how the number of flights from each launch site correlates with successful landings.
  - **Payload vs. Launch Site:** This chart investigates if the mass of payloads affects the success rate differently across various launch sites.
  - **Orbit Type Success Rates:** A bar chart to determine which orbit types have higher success rates, indicating the effectiveness of the Falcon 9 in different orbital missions.
  - **Flight Number vs. Orbit Type:** A scatter plot to see if there's a pattern between the flight number and the success rate within specific orbit types.
  - **Payload vs. Orbit Type:** This chart is used to examine how the payload mass correlates with the success rate for different orbit types.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/a0000afd14b1afde53cc426a96ef6287351426c6/5-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- **Summary of the SQL queries performed**
  - **Unique Launch Sites:** Retrieved distinct names of launch sites used in SpaceX missions.
  - **Launch Sites with 'CCA':** Selected records where launch sites start with 'CCA'.
  - **NASA Payload Mass:** Calculated the total payload mass carried by boosters for NASA (CRS).
  - **Average Payload Mass:** Determined the average payload mass carried by the booster version F9 v1.11.
  - **First Successful Landing:** Found the date of the first successful landing on a ground pad.
  - **Boosters with Specific Criteria:** Listed boosters with successful drone ship landings and payload mass between 4000 and 6000 kg2.
  - **Mission Outcomes:** Counted the total number of successful and failed mission outcomes.
  - **Maximum Payload Boosters:** Identified booster versions that carried the maximum payload mass.
  - **2015 Drone Ship Failures:** Displayed records of failed drone ship landings in 2015.
  - **Landing Outcome Rankings:** Ranked the count of different landing outcomes within a specified date range.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/c4e0648b2e7074ce0a1ddae0cd0b55765c1b901c/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

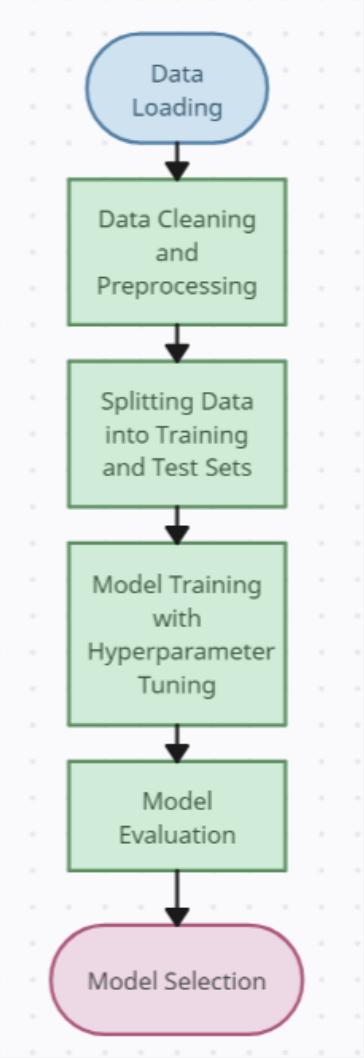
- **Summary of the map objects created and added to a Folium map, along with the reasons for their addition:**
- **Markers & Circles:**
  - To visually represent the locations of launch sites and their success or failure outcomes. Green markers indicate successful launches, while red markers denote failed ones1.
- **Marker Clusters:**
  - To manage and simplify the display of multiple markers at the same launch site coordinates, enhancing map readability.
- **Lines & PolyLines:**
  - To depict distances from launch sites to various proximities like coastlines, cities, railways, and highways, providing geographical context and analysis2.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/f54eb4869f8e98db07f18eb4d6f31f7a9e9ecfb8/6-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Summary of the plots/graphs and interactions added to the SpaceX Launch Records Dashboard:**
  - **Launch Site Selection:** A dropdown menu allows users to filter data by launch site, with an option to view all sites by default. This interaction enables users to focus on specific sites or compare across all sites.
  - **Success Rate Visualization:** A pie chart displays the success rate of launches, either for all sites collectively or for a selected site. This visual representation helps users quickly grasp the success rates and compare them easily.
  - **Payload Mass Range Selection:** A slider lets users select a range for payload mass, which can be used to filter the displayed data. This interaction provides a way to explore how payload mass affects launch success.
  - **Scatter Plot:** A scatter chart shows the correlation between payload mass and launch success, colored by booster version category. This plot is crucial for identifying trends and patterns in the data, such as the impact of payload mass on the likelihood of a successful launch.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/3dcd19c00971b6b318421dd00a7ecb3c8a46423f/7-spacex_dash_app.ipynb

Predictive Analysis (Classification)

- **Summary of the model development process:**
  - **Exploratory Data Analysis:** Identified features and labels, explored data distributions.
  - **Preprocessing:** Standardized data, split into training and test sets.
  - **Model Selection:** Evaluated Logistic Regression, SVM, Decision Tree, and KNN classifiers.
  - **Hyperparameter Tuning:** Used GridSearchCV to find optimal parameters for each model.
  - **Model Evaluation:** Assessed models using accuracy score and confusion matrix.
  - **Best Model:** Determined the best performing model based on test data accuracy.
- **GitHub Notebook URL**
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr/blob/fa21d6176cd8aefdc57735d2fd938f18d962a9b6/8-SpaceX_Machine_Learning_Prediction_Part_5.ipynb



Results

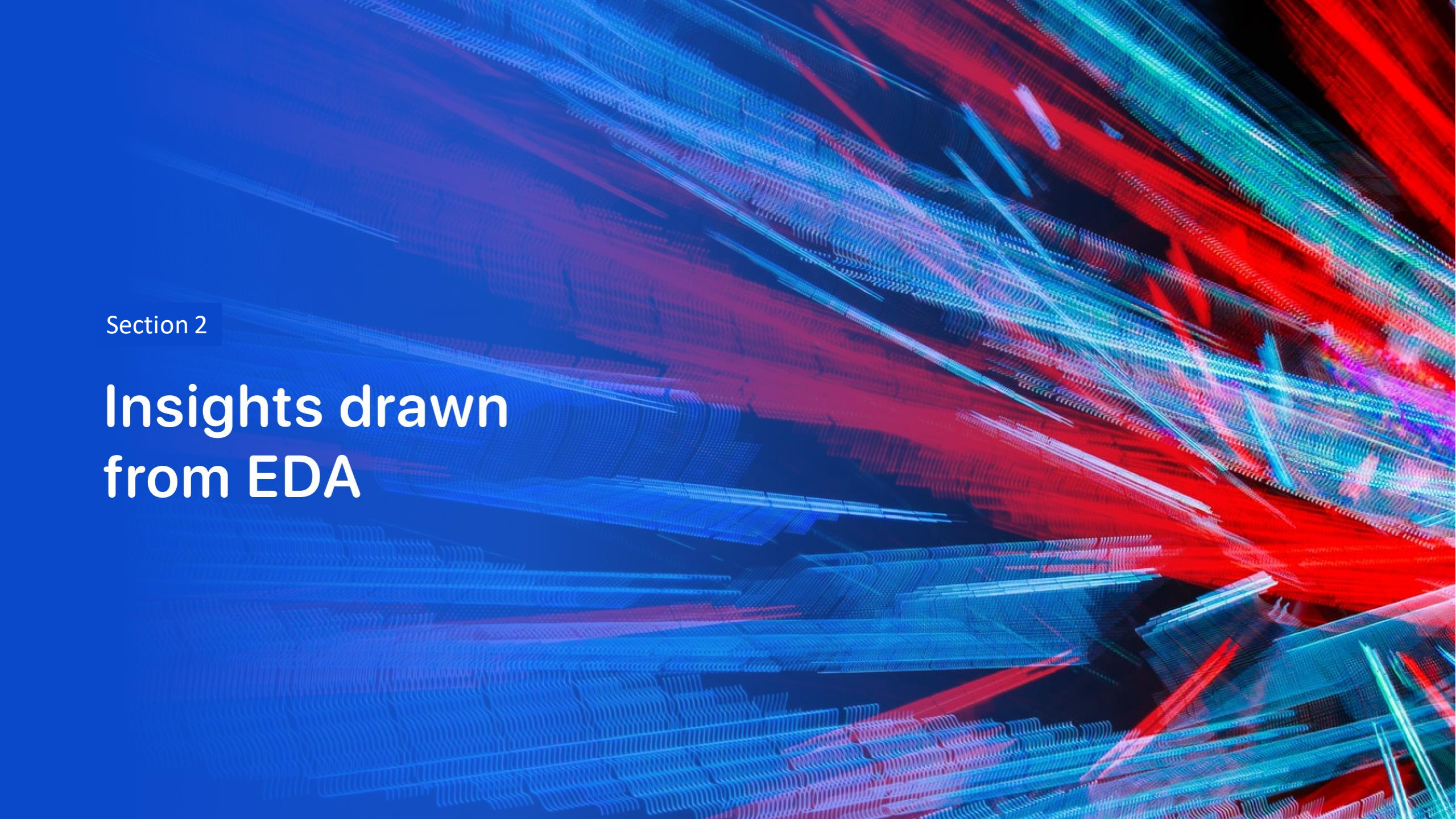
- **Exploratory Data Analysis (EDA):**
  - The EDA was conducted using Python's Pandas and Seaborn libraries. It revealed significant correlations and patterns within the dataset. For instance, a positive correlation was observed between the flight number and payload mass, indicating that as SpaceX's technology advanced, they were able to launch heavier payloads. Furthermore, the success rates varied across different launch sites and orbit types, suggesting that these factors significantly influence the outcome of a launch.
- **Interactive Analytics:**
  - Interactive visual analytics were implemented using Folium for geospatial visualization and Plotly Dash for creating interactive dashboards. This allowed for a dynamic representation of data, enabling users to interactively explore the geographical distribution of launch sites, the variation in payload masses, and the success rates of different orbit types.

Results

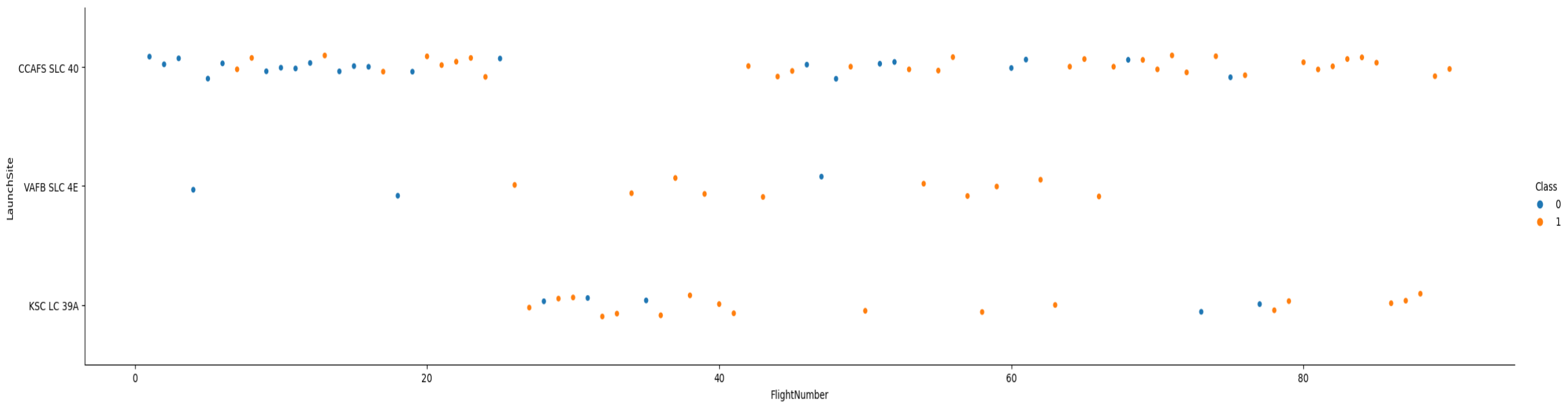
- **Predictive Analysis:**
  - A series of classification models were developed using Scikit-learn to predict the outcome of Falcon 9 first stage landings.
  - The dataset was preprocessed using techniques such as one-hot encoding for categorical variables and normalization for numerical variables.
  - Several models were trained and evaluated, including Logistic Regression, Decision Trees, and Random Forests. Hyperparameter tuning was performed using GridSearchCV to optimize model performance.
  - The Decision Tree Classifier emerged as the best model, achieving an accuracy of 87.5% on the training set and 83.33% on the test set.
  - The confusion matrix revealed that the major issue with the Logistic Regression model was a high number of false positives, which was addressed during model refinement.
- These results underscore the effectiveness of the applied data science workflow in extracting insights and predicting the success of Falcon 9 first stage landings. The interactive tools and predictive models developed in this project provide valuable resources for strategic decision-making in the commercial spaceflight market.

Section 2

Insights drawn from EDA

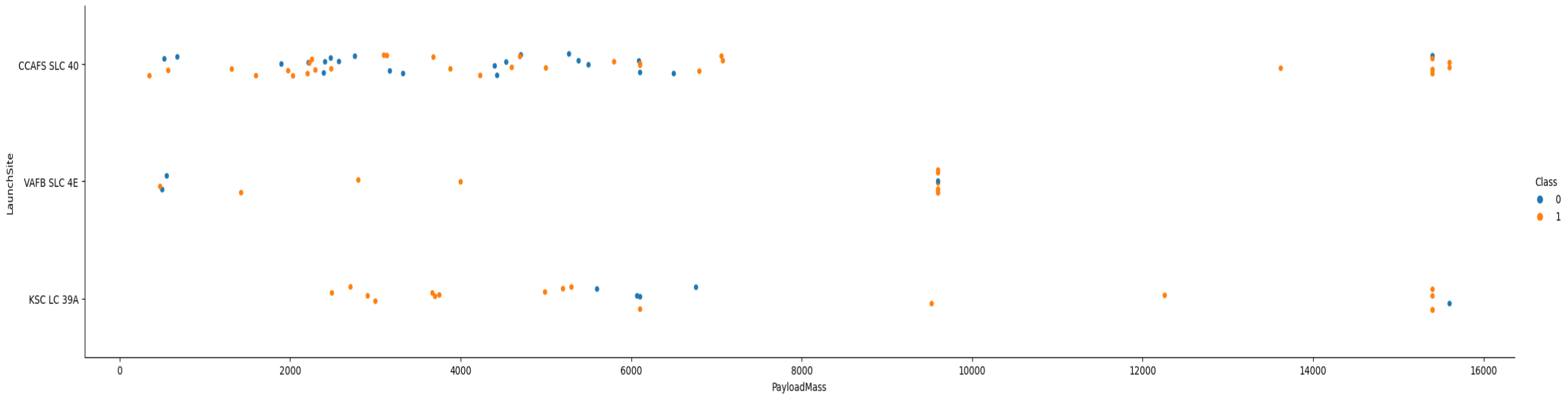


Flight Number vs. Launch Site



The scatter plot visualizes the relationship between Flight Number and Launch Site, with different classes represented by distinct colors. It appears that earlier flights, primarily launched from CCAFS SLC 40, have a mix of both class 0 and 1. As flight numbers increase, there is a noticeable shift towards class 1 outcomes, indicating improvements in launch success rates over time. KSC LC 39A and VAFB SLC 4E also become more common launch sites for higher numbered flights. This suggests a correlation between flight number, launch site, and launch success.

Payload vs. Launch Site

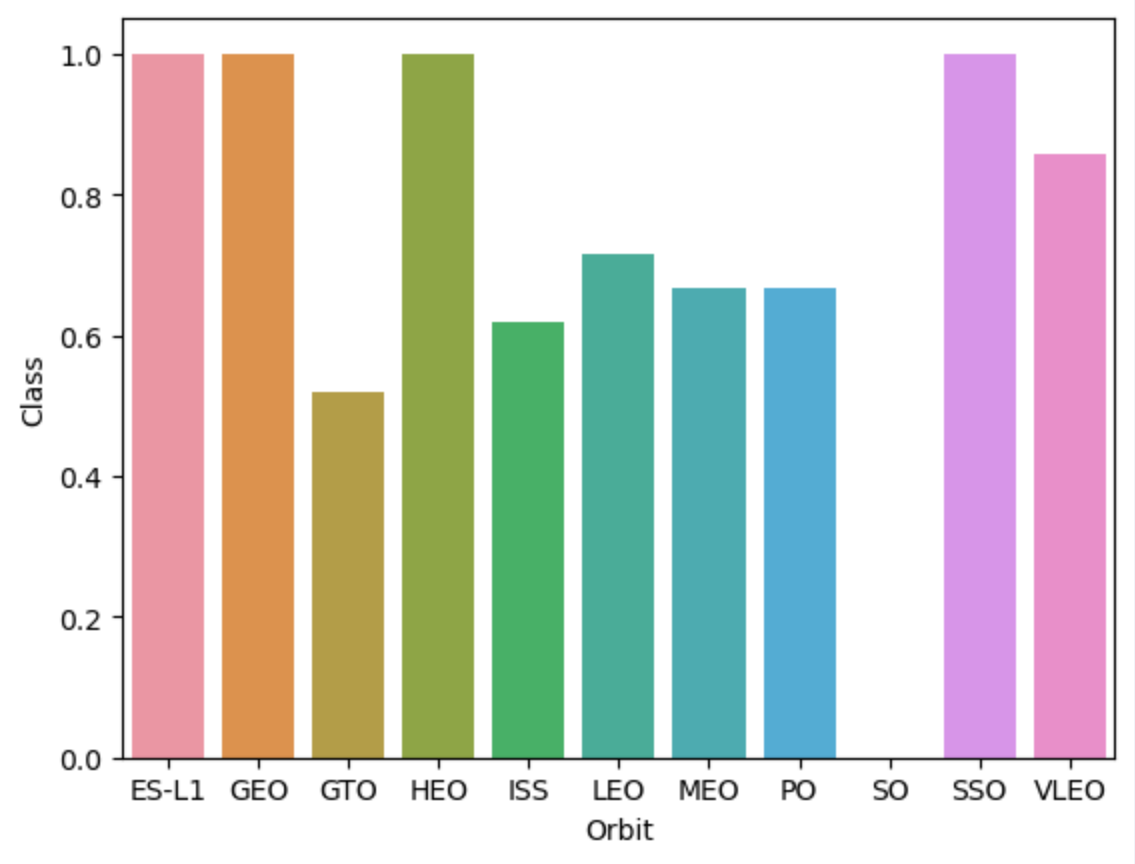


The scatter plot shows the relationship between rocket payload mass and launch sites. The X-axis represents payload mass, and the Y-axis represents launch sites. Points are colored by class. Notably, the VAFB-SLC launch site has no launches with a payload mass greater than 10000 kg, suggesting it may not be equipped for heavy payload launches. In contrast, CCAFS SLC 40 and KSC LC 39A sites show a broader range of payload masses, indicating a wider capability.

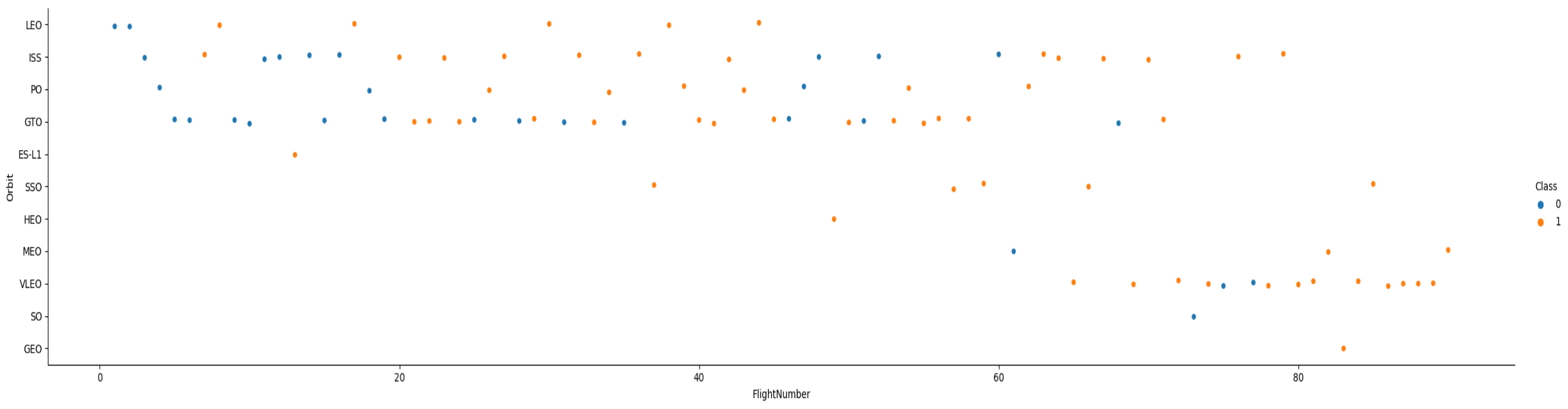
Success Rate vs. Orbit Type

The bar chart represents the success rates of various orbit types. High success rates are observed for ES-L1, GEO, and VLEO orbits, indicating most missions to these orbits are successful. Moderate success rates are seen for GTO and HEO, while ISS, LEO, MEO, PO, SO, and SSO orbits show varied success rates below 0.8.

This chart helps understand the relative reliability of missions to different orbit types. The success rate is calculated as the mean of the 'Class' column for each 'Orbit' type. The higher the bar, the higher the success rate.

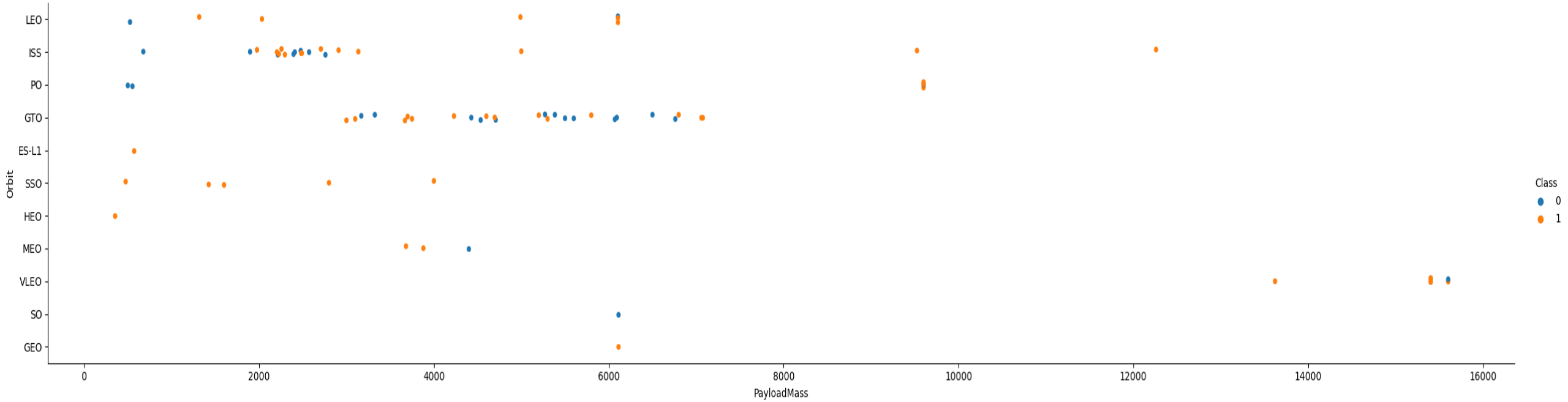


Flight Number vs. Orbit Type



The scatter plot visualizes the relationship between Flight Number and Orbit type. Different colors indicate the class value. In the LEO orbit, there is a noticeable pattern where success is associated with higher flight numbers, suggesting an improvement over time. However, for GTO orbit, the data points are scattered without a discernible pattern, indicating no clear relationship between flight number and success rate. This analysis helps in understanding the success rate of flights in different orbits.

Payload vs. Orbit Type

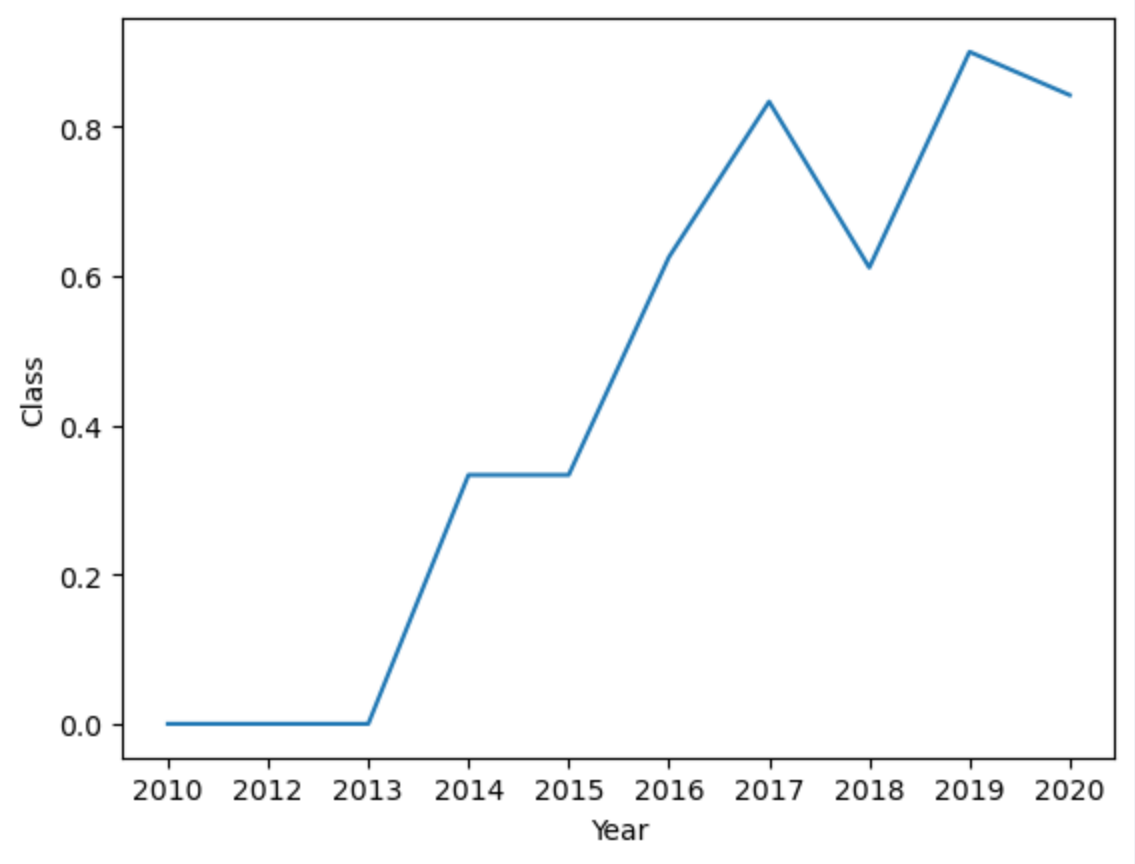


The scatter plot visualizes the relationship between payload mass and orbit type, with different colors indicating mission success or failure. It shows that successful landings are more common with heavy payloads for Polar, LEO, and ISS orbits. However, for GTO orbits, there is no clear distinction as both successful and unsuccessful missions are present. This visualization aids in understanding how payload mass correlates with mission outcomes across various orbit types.

Launch Success Yearly Trend

The plot visualizes the average launch success rate from 2010 to 2020. The x-axis represents the year, and the y-axis represents the success rate. The line chart shows a noticeable increase in success rate starting from 2013, with a plateau in 2014 and another significant increase after 2015.

This suggests an overall improvement in launch successes over these years. The code provided extracts the year from the date and plots the success rate against it. The success rate is calculated as the mean of the 'Class' column for each year.



All Launch Site Names

```
In [8]:  1 %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

The query is designed to select distinct launch sites from a database table named “SPACEXTABLE”. The database being queried is located at “sqlite:///my_data1.db”. The result of the query is displayed below the code cell, showing four distinct launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

This indicates that there are four unique launch sites in the “SPACEXTABLE” database table. One of the launch site names, KSC LC-39A, is highlighted, possibly indicating it was selected or of particular interest.

Launch Site Names Begin with 'CCA'

```
In [11]:  1 %sql select "Launch_Site" from "SPACEXTABLE" where "Launch_Site" like 'CCA%' limit 5
 * sqlite:///my_data1.db
Done.
```

Out[11]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

The query is designed to select the “Launch_Site” from a table named “SPACEXTABLE” where the “Launch_Site” starts with ‘CCA’. The output is limited to 5 results. The result of this query displays that all five entries for “Launch_Site” are “CCAFS LC-40”.

This indicates that the first five records in the “SPACEXTABLE” database table where “Launch_Site” starts with ‘CCA’ are all “CCAFS LC-40”.

Total Payload Mass

```
In [13]:  1 %sql select SUM(PAYLOAD_MASS__KG_) from "SPACEXTABLE" where "Customer" = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
```

Out[13]:

SUM(PAYLOAD_MASS__KG_)
45596

The query is designed to select the sum of the payload mass in kilograms from a table named "SPACEXTABLE" where the customer is 'NASA (CRS)'. The result of this query, shown below it, indicates that the total payload mass for NASA (CRS) is 45596 kilograms.

This suggests that the total mass of payloads sent by NASA (CRS) according to the records in the "SPACEXTABLE" database table is 45596 kg. This information could be useful for understanding the scale of NASA's (CRS) space missions.

Average Payload Mass by F9 v1.1

```
In [14]:  1  %sql select AVG(PAYLOAD_MASS__KG_) from "SPACEXTABLE" where "Booster_Version" = 'F9 v1.1'
 * sqlite:///my_data1.db
Done.
```

Out[14]:

AVG(PAYLOAD_MASS__KG_)
2928.4

The query is designed to select the average payload mass in kilograms from a table named "SPACETABLE" where the "Booster_Version" is 'F9 v1.1'. The result of this query, shown below it, indicates that the average payload mass for this specific booster version is 2928.4 kg.

This suggests that the average mass of payloads sent using the 'F9 v1.1' booster, according to the records in the "SPACETABLE" database table, is 2928.4 kg. This information could be useful for understanding the typical payload capacity of the 'F9 v1.1' booster.

First Successful Ground Landing Date

```
In [16]:  1 %sql select MIN(Date) from "SPACEXTABLE" where "Landing_Outcome" = 'Success (ground pad)'
```

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

Out[16]:

MIN(Date)
2015-12-22

The query is designed to select the minimum date from a table named “SPACEXTABLE” where the “Landing_Outcome” is ‘Success (ground pad)’. The result of this query, shown below it, indicates that the earliest successful landing on a ground pad occurred on December 22, 2015.

This suggests that the first successful landing on a ground pad, according to the records in the “SPACEXTABLE” database table, occurred on December 22, 2015. This information could be useful for understanding the timeline of successful landings on ground pads.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [19]:

```
%%sql
select "Booster_Version" from "SPACEXTABLE" where
"PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000 and
"Landing_Outcome" = 'Success (drone ship)'
```

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

Out[19]:

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

The query is designed to select the “Booster_Version” from a table named “SPACEXTABLE” where the payload mass in kilograms is greater than 4000 and less than 6000, and where the landing outcome is marked as ‘Success (drone ship)’. The result of this query, shown below it, lists four booster versions that meet these criteria: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

This suggests that these four booster versions have successfully landed on a drone ship with a payload mass between 4000 and 6000 kg, according to the records in the “SPACEXTABLE” database table. This information could be useful for understanding the performance and capabilities of different booster versions.

Total Number of Successful and Failure Mission Outcomes

In [25]:
```
%%sql
SELECT "Mission_Outcome", COUNT(*) FROM "SPACEXTABLE" GROUP BY "Mission_Outcome"
```

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

Out[25]:

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

The query is designed to select the “Mission_Outcome” from a table named “SPACETABLE” and count each occurrence, grouping by the outcome. The result of this query, shown below it, indicates that there has been 1 “Failure (in flight)”, 98 “Success”, 1 additional “Success” with an unclear payload status, and another where the outcome is not specified.

This suggests that according to the records in the “SPACETABLE” database table, there have been 98 successful missions, 1 failed mission in flight, 1 successful mission with an unclear payload status, and 1 mission with an unspecified outcome. This information could be useful for understanding the distribution of mission outcomes.

Boosters Carried Maximum Payload

In [27]:
```
%%sql
select "Booster_Version" from "SPACEXTABLE"
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from "SPACEXTABLE")
```

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

Out[27]:

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 query is designed to select the “Booster_Version” from a table named “SPACEXTABLE” where the payload mass in kilograms is equal to the maximum payload mass in the same table. The result of this query, shown below it, lists several booster versions.

This suggests that these booster versions have been used for missions with the maximum payload mass, according to the records in the “SPACEXTABLE” database table. This information could be useful for understanding which booster versions are capable of carrying the heaviest payloads.

2015 Launch Records

In [21]:
```
%%sql
select substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site from "SPACEXTABLE"
where Landing_Outcome= 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

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

Out[21]:

substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The query is designed to select records from the “SPACEXTABLE” where the landing outcome was a failure on a drone ship, and the date substring matches “2015”. The result of this query is displayed below it, showing two records that meet these criteria. Each record includes data on the date substring, landing outcome, booster version, and launch site.

This suggests that in 2015, there were two instances where the landing outcome was a failure on a drone ship, with the booster versions and launch sites specified in the results.

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

In [22]:

```
%%sql
select Landing_Outcome, COUNT(*) from "SPACEXTABLE"
where Date >= '2010-06-04' and
Date <= '2017-03-20'
ORDER BY COUNT(*) DESC
```

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

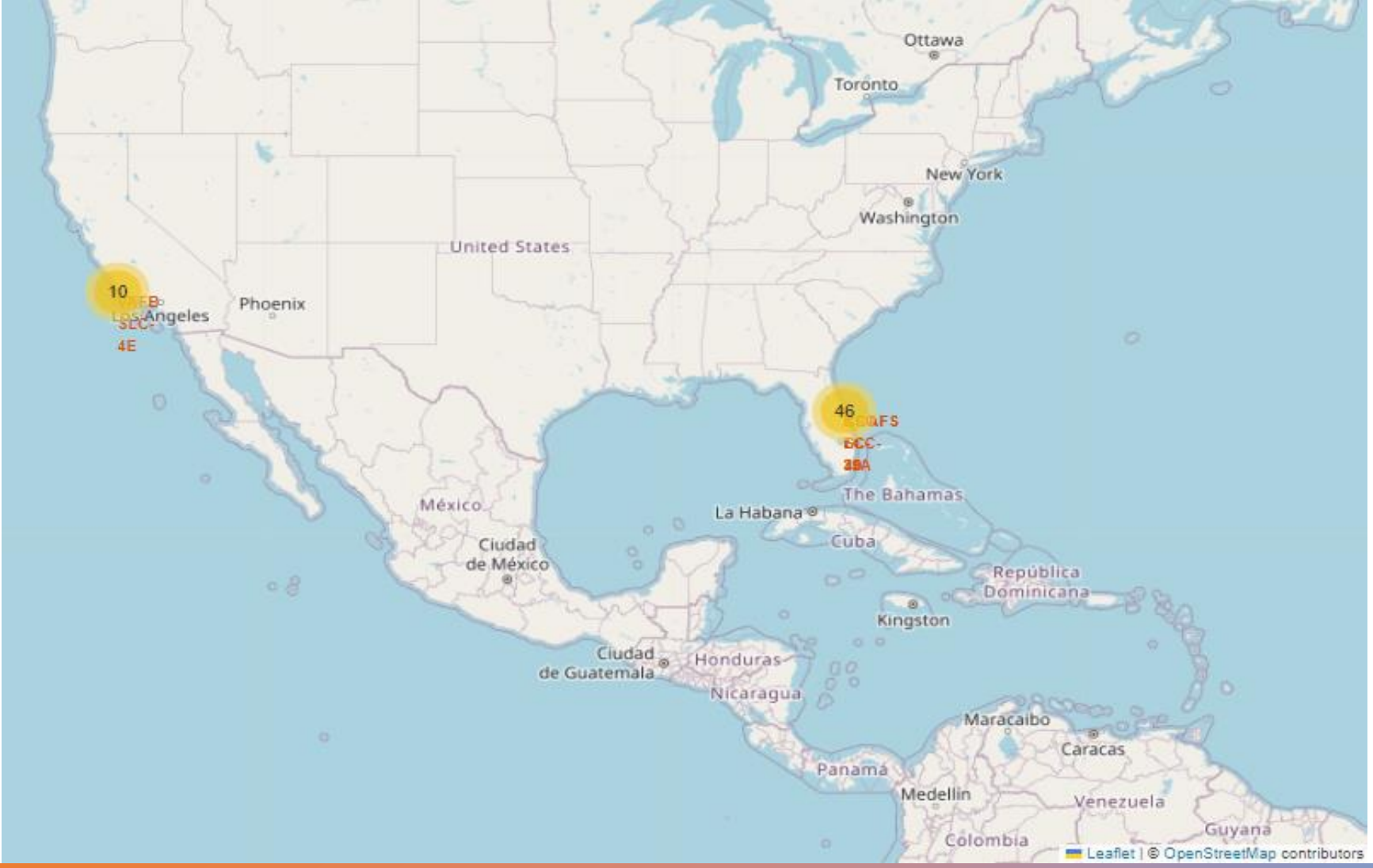
Out[22]:

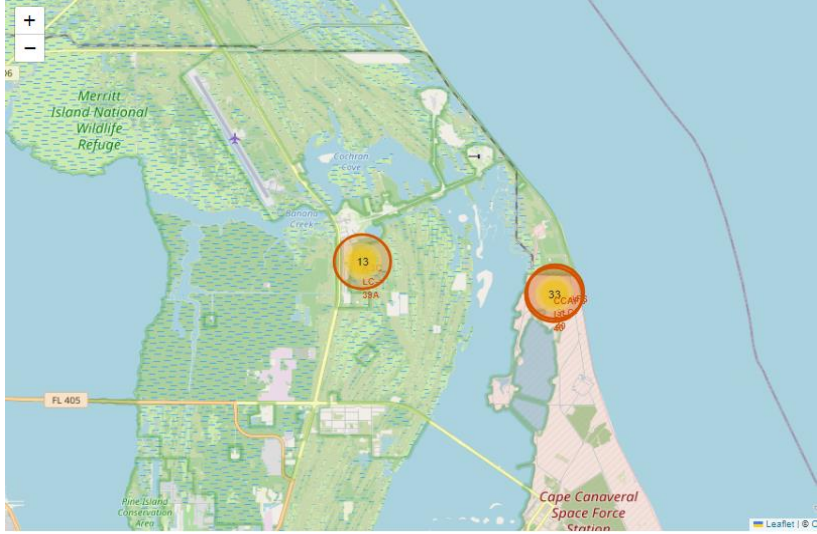
Landing_Outcome	COUNT(*)
Failure (parachute)	31

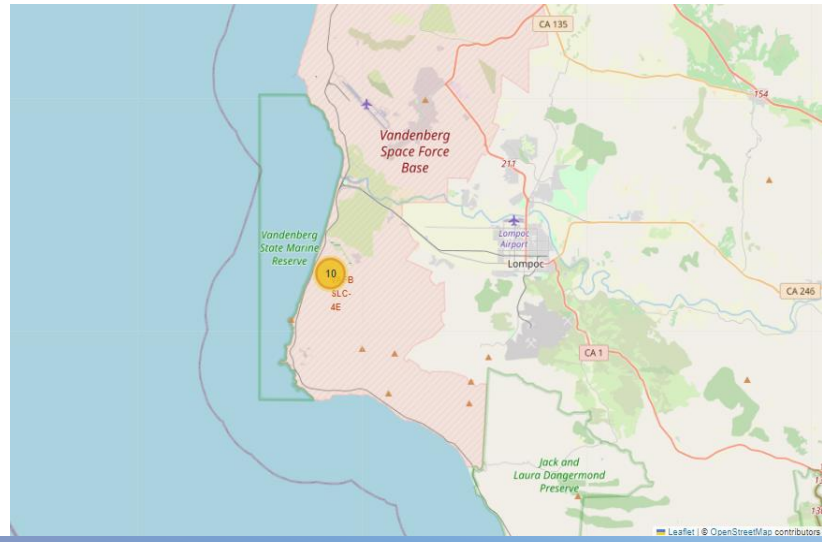
The query is designed to select the "Landing_Outcome" and count of occurrences from a table named "SPACEXTABLE" where the date is between '2010-06-04' and '2017-03-20'. It orders the results by the count in descending order. The result of this query, shown below it, indicates there were 31 instances of "Failure (parachute)" landing outcomes within the specified date range.

This suggests that between June 4, 2010, and March 20, 2017, there were 31 instances where the landing outcome was a failure on a parachute, according to the records in the "SPACEXTABLE" database table. This information could be useful for understanding the frequency of this specific type of landing failure during this period.


Section 3
Launch Sites
Proximities Analysis

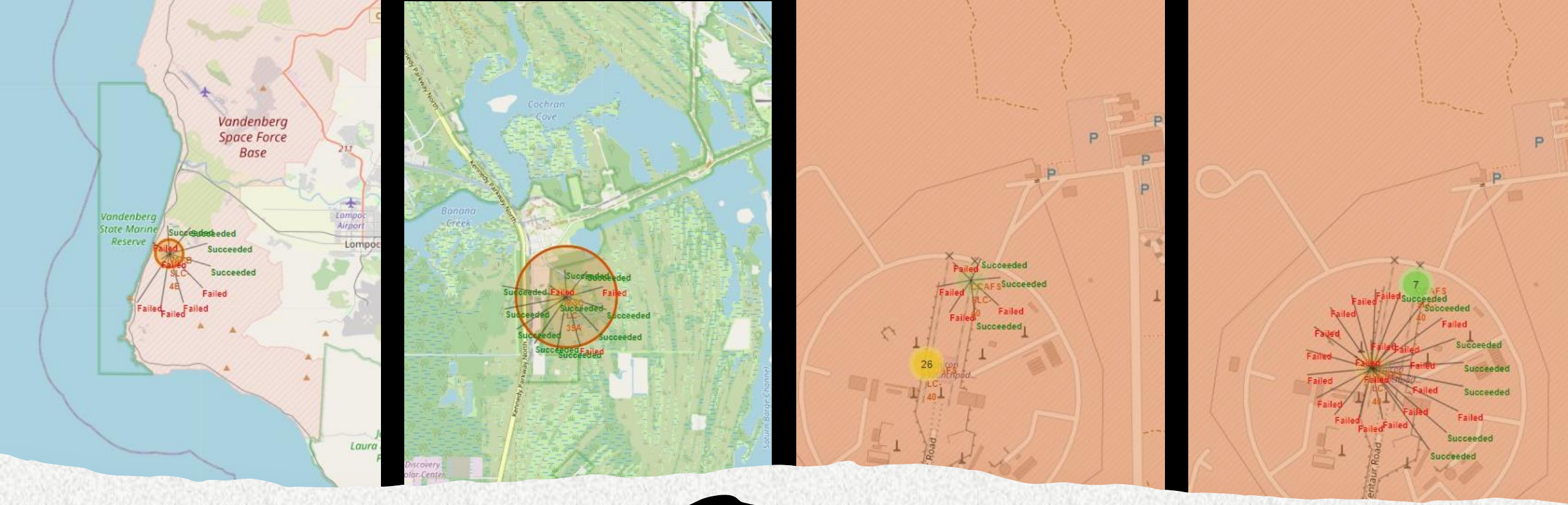






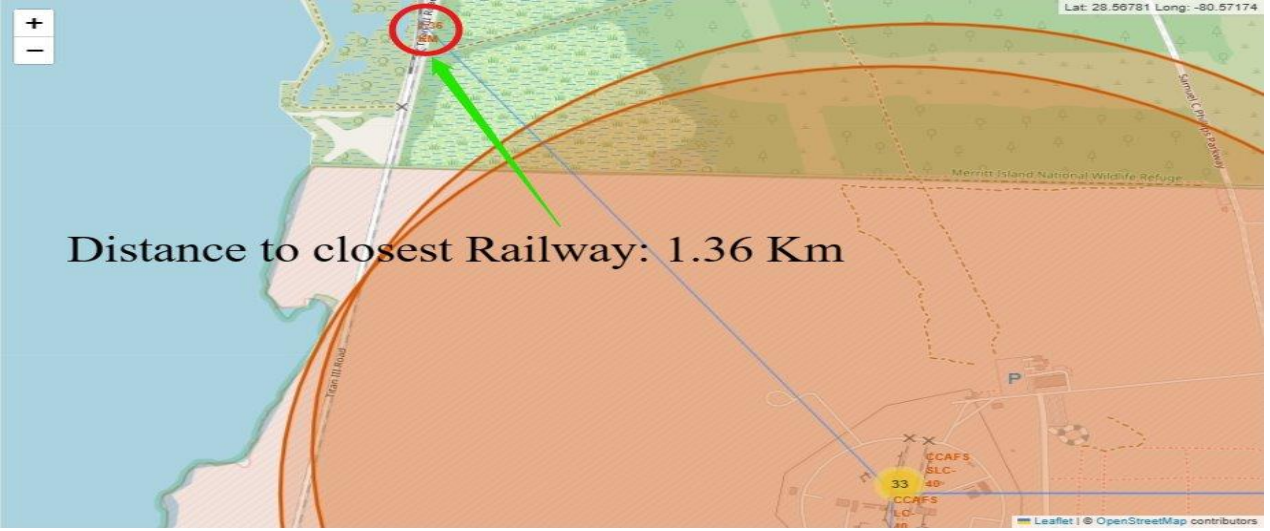
Launch Sites Locations

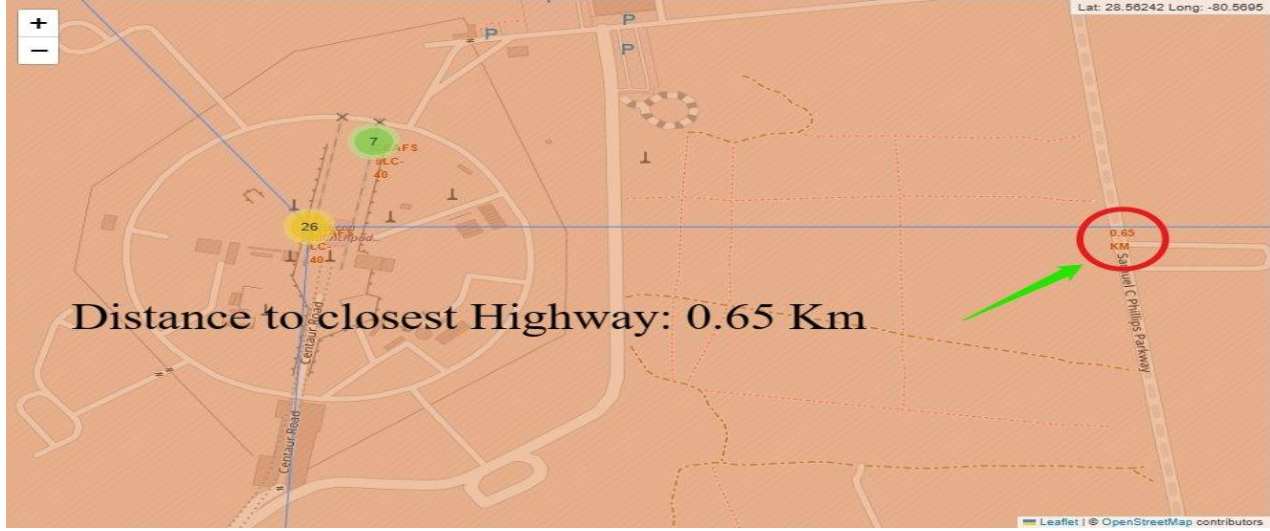
There are four launch sites, one in the west of US, in Los Angeles and three in the east of US, in Florida.

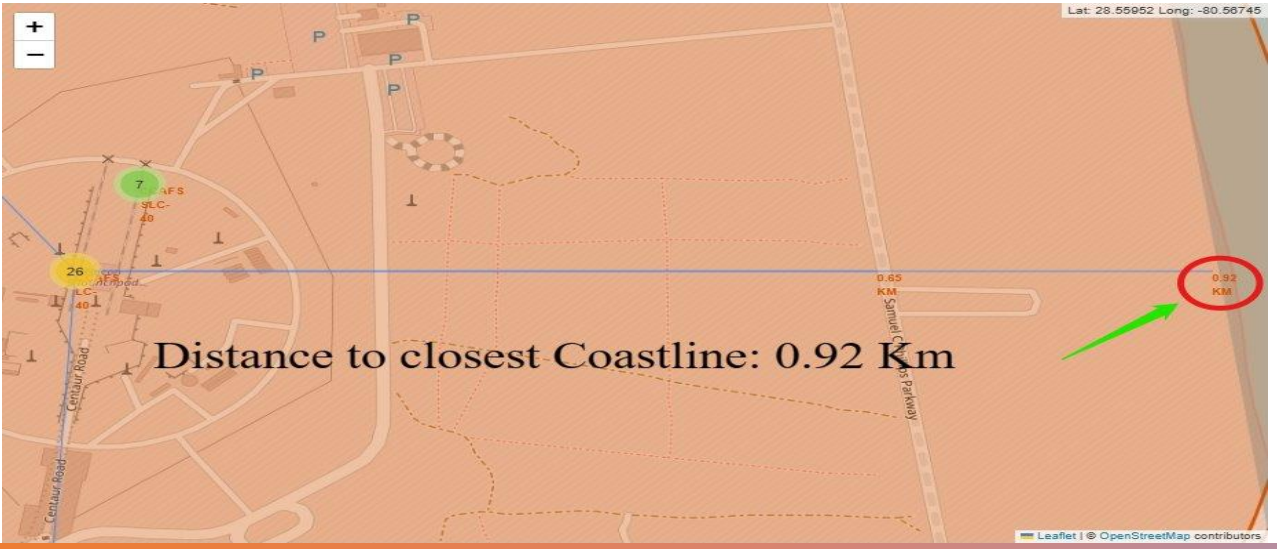


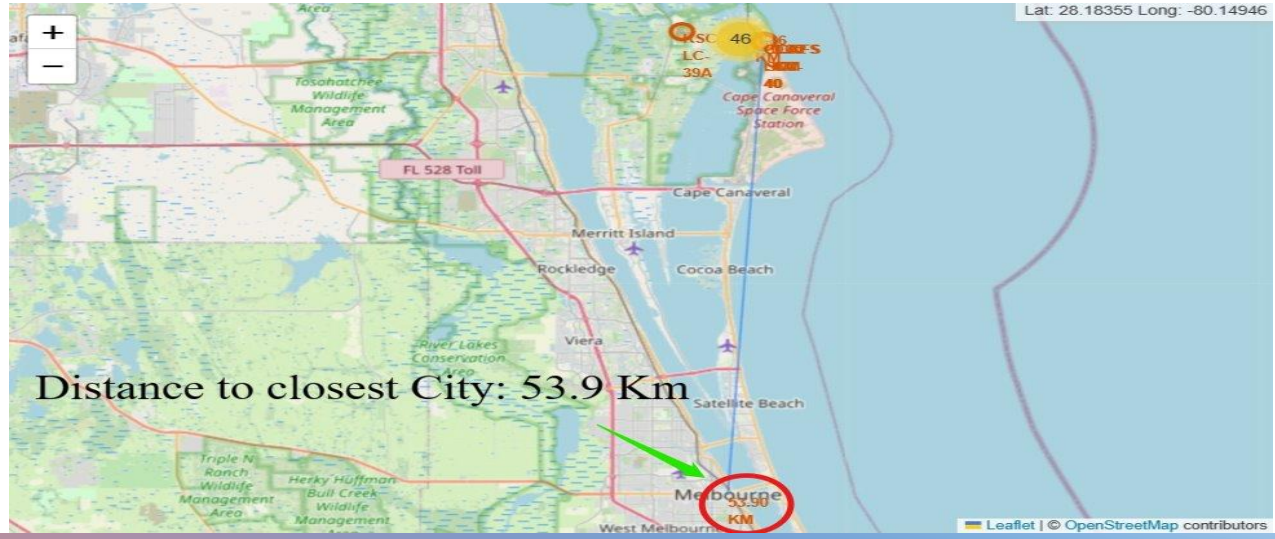
Launch Sites Success Rates

For each launch site we can see the success rate based on the markers, green markers for successful launches and red ones for a failed one.



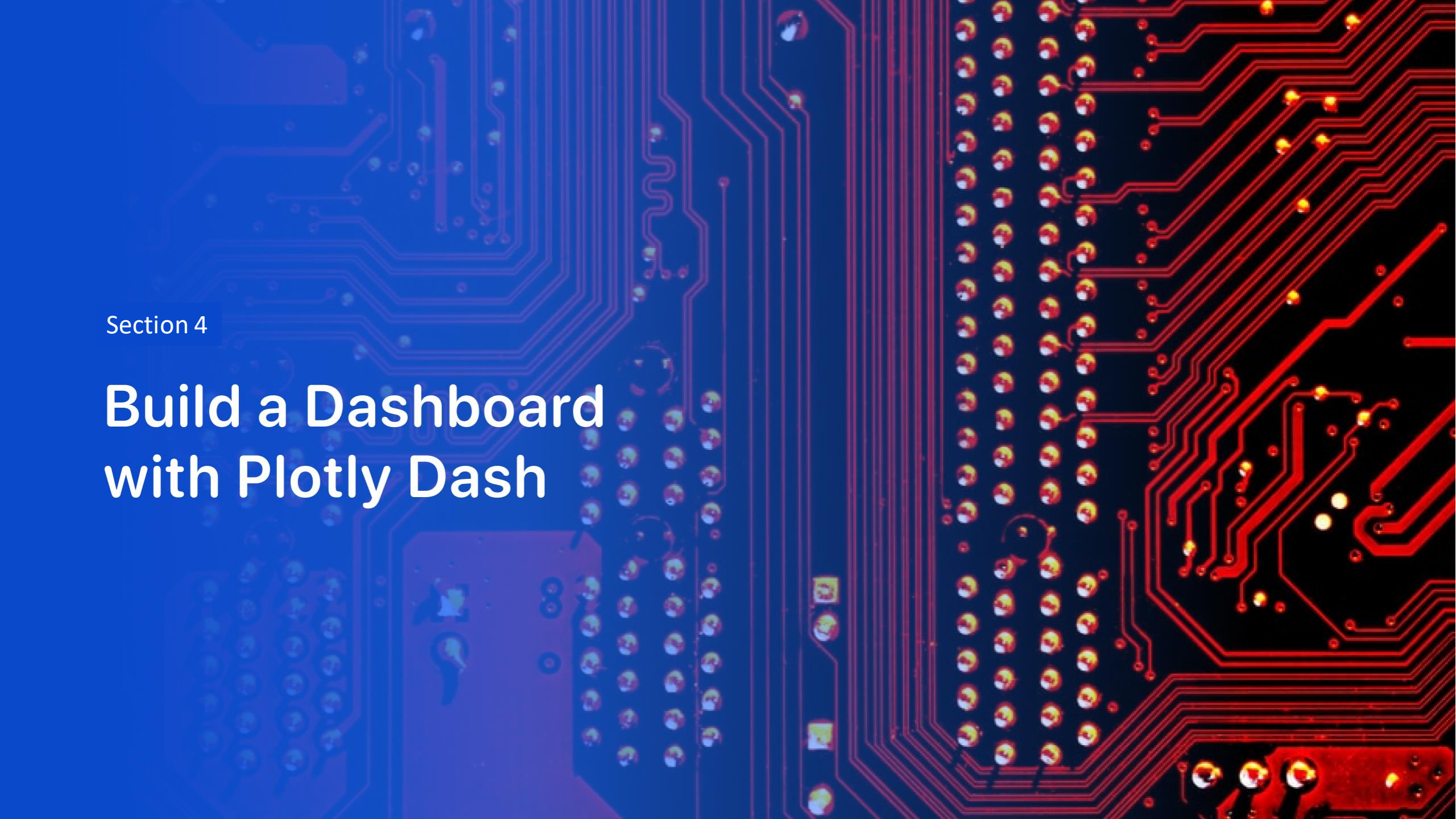


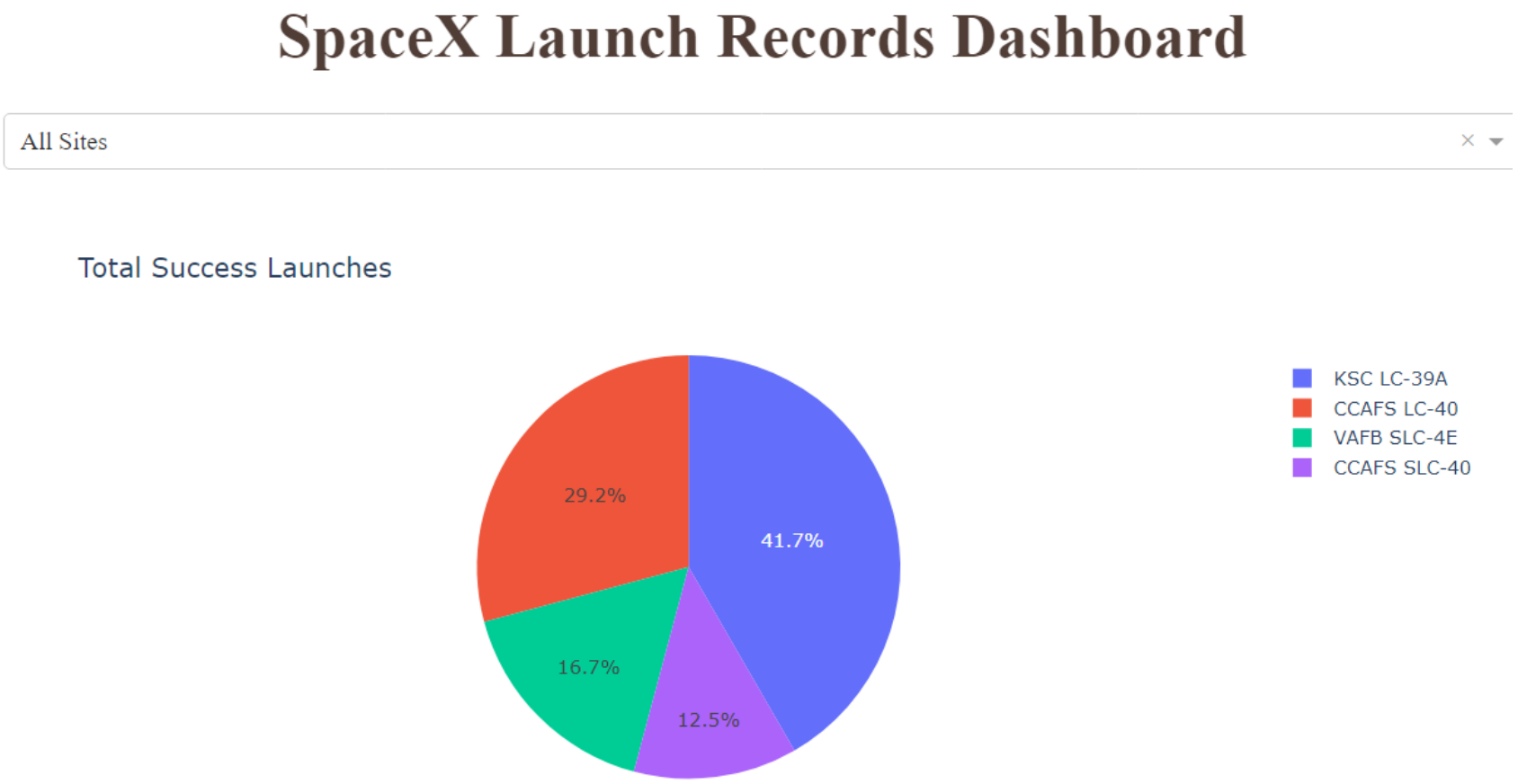




Launch Sites Proximities

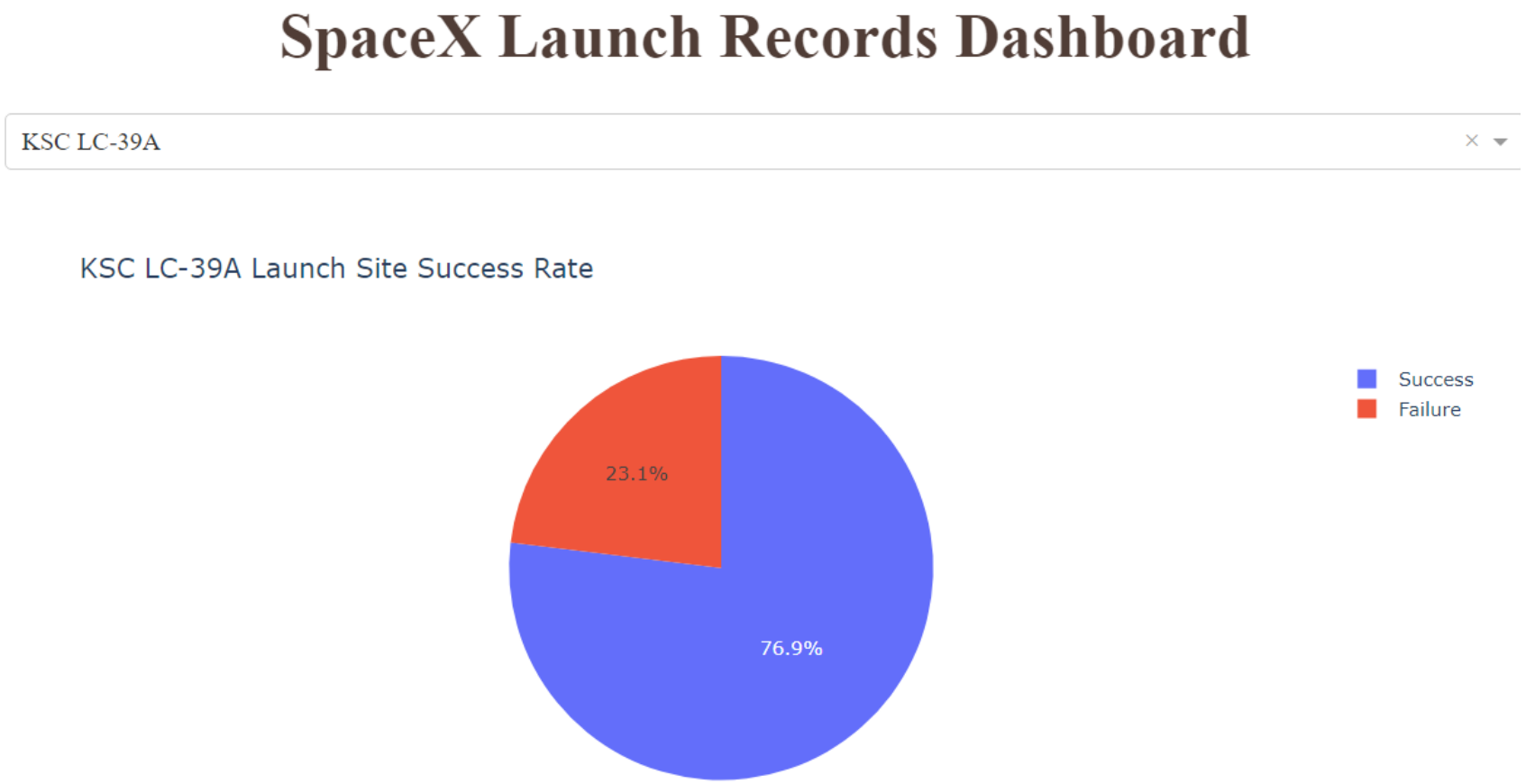
Proximities of one of the launch sites; the closest railway, highway, coastline and city.


Section 4
Build a Dashboard with Plotly Dash



Launch Success Count for All Sites

- **Screenshot key elements and findings**
  - **Launch Sites:** Four different colors represent four different SpaceX launch sites on the pie chart:
    - **KSC LC-39A** (in blue) accounts for 41.7% of total successful launches.
    - **CCAFS SLC-40** (in red) accounts for 29.2%.
    - **VAFB SLC-4E** (in green) has a 16.7% share.
    - **CCAFS LC-40** (in purple) contributes to 12.5% of



launch Site with Highest Launch Success Ratio

- The pie chart shows data for the **“KSC LC-39A”** Launch Site, which is the one with the highest launch success rate.
- **Screenshot key elements and findings**
  - Success (Blue): The blue section represents successful launches, covering 76.5% of the chart.
  - Failure (Red): The red section represents failed launches, covering 23.5% of the chart.

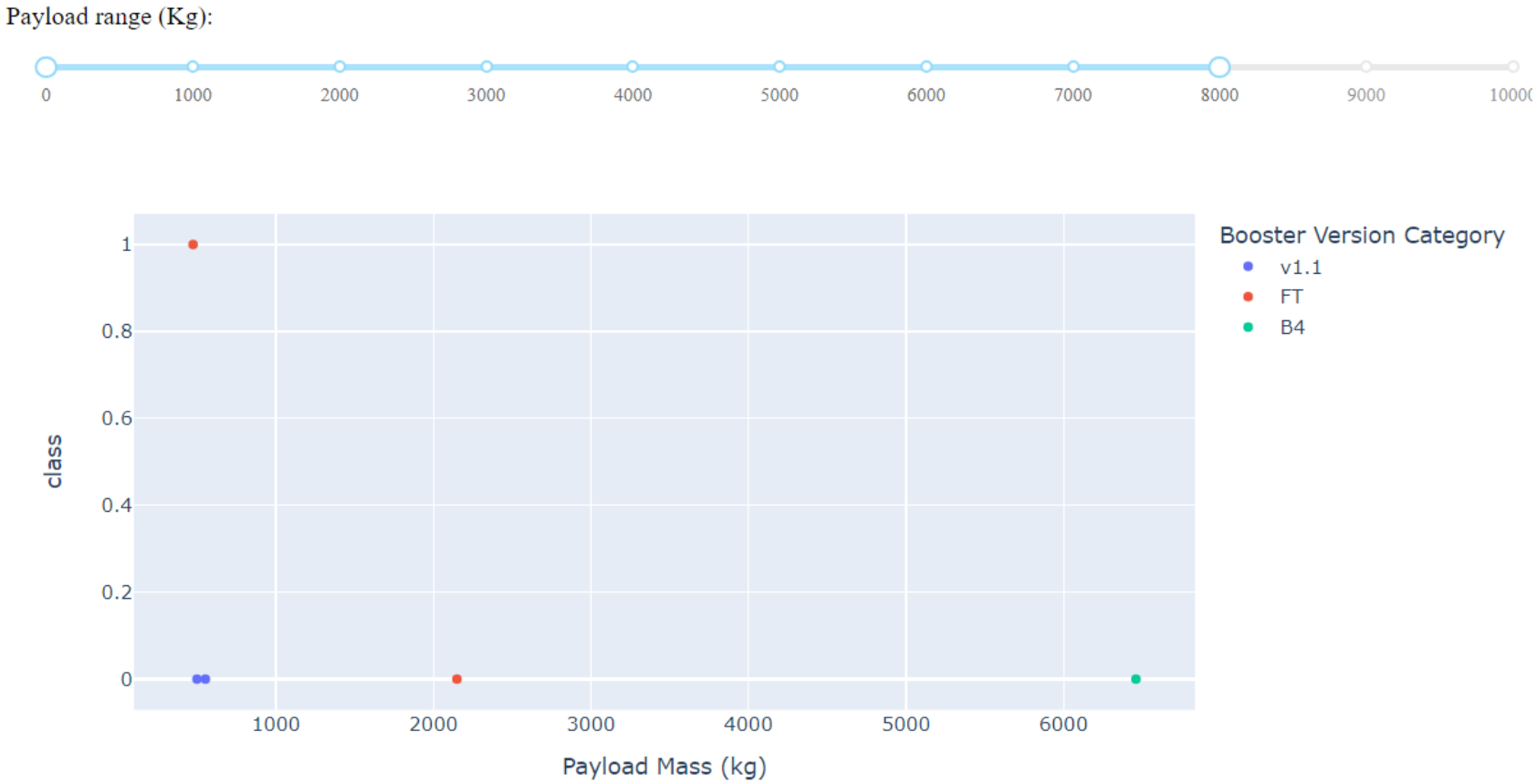
Payload vs. Launch Outcome scatter plot for **CCAFS LC-40** Launch Site

The scatter plot shows Payload vs. Launch Outcome for CCAFS LC-40 Launch Site. The x-axis represents payload mass, and the y-axis represents launch outcome. Three booster versions are shown: v1.0 (blue), v1.1 (red), and FT (green).

These are some observations bas ed on the provided plot.

- The v1.0 version shows that all launches failed.
- The v1.1 version shows that almost all launches failing, except for one launch at 2000 kg
- The FT version has the highest success rate, with payloads ranging from 2000 to 5000 kg.

Payload vs. Launch Outcome scatter plot for **VAFB SLC-4E** Launch Site

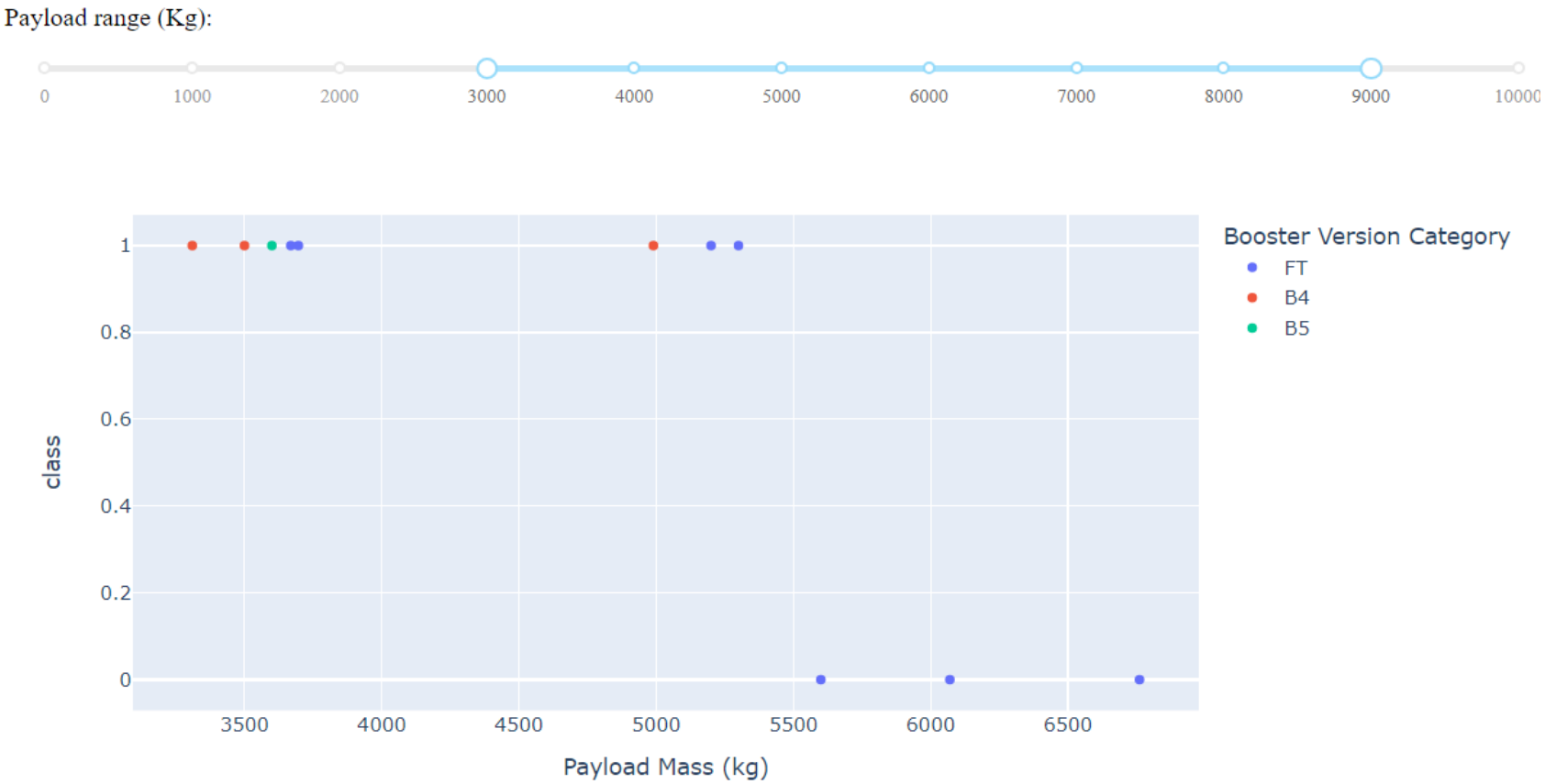


The scatter plot illustrates Payload vs. Launch Outcome for VAFB SLC-4E Launch Site. Three booster versions are depicted: v1.1 (blue), FT (red), and B4 (green).

These are some observations bas ed on the provided plot.

- The v1.1 version experienced failure at a lower payload mass, indicating its unreliability.
- The FT version has one successful launch at around 500 kg and one failed launch at 2000 kg
- The B4 show only one failed launch at around 6500 kg

Payload vs. Launch Outcome scatter plot for **KSC LC-39A** Launch Site

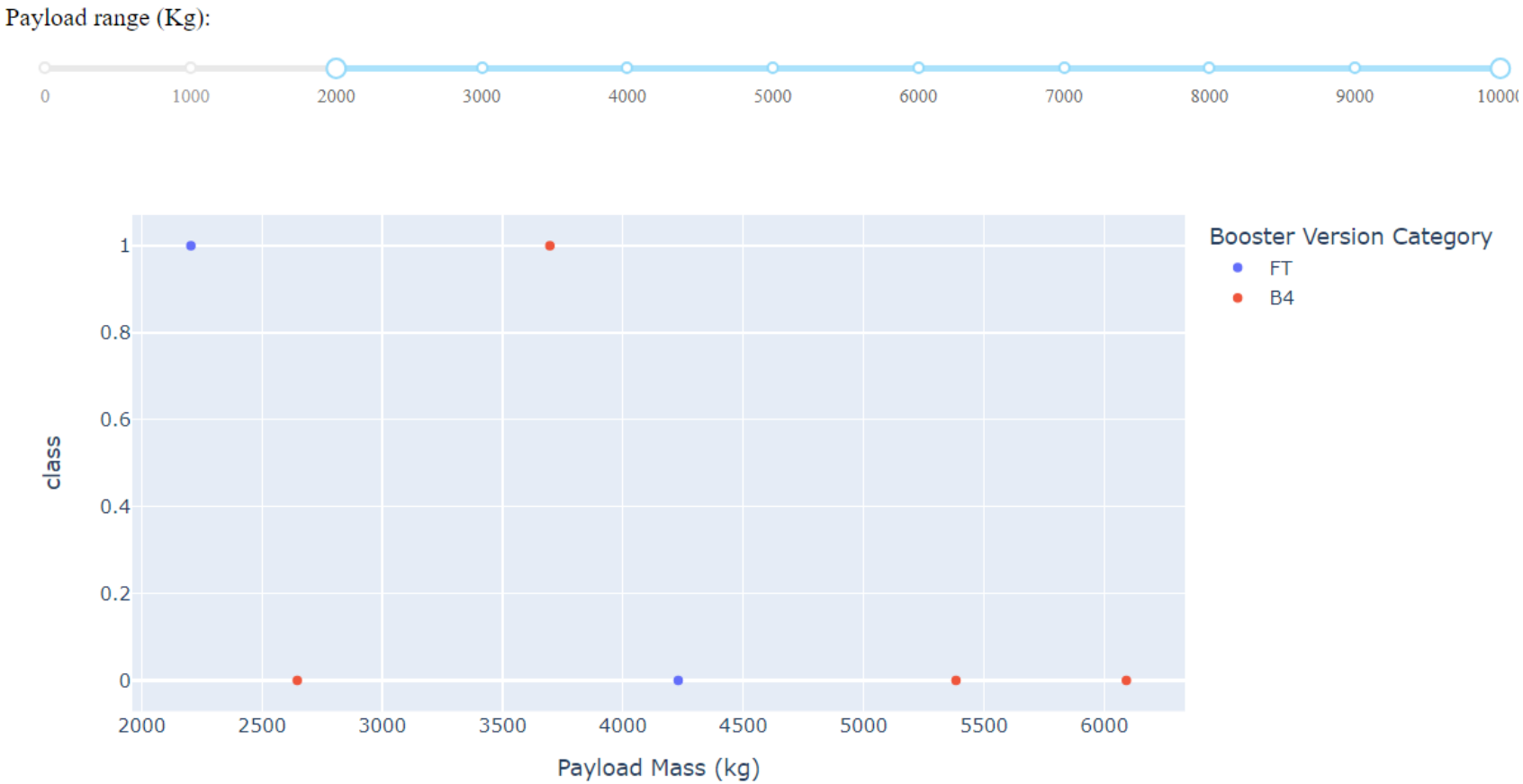


The scatter plot shows Payload vs. Launch Outcome for KSC LC-39A Launch Site. Three booster versions are shown: FT (blue), B4 (red), and B5 (green).

These are some observations bas ed on the provided plot.

- The FT version has a high success rate, especially with payloads lower than 5500 kg, but fails for higher payloads.
- The B4 version shows that all launches were successful for payloads lower than 5000 kg.
- The B5 version, with only one data point at around 3500 kg payload, indicates a successful launch.

Payload vs. Launch Outcome scatter plot for **CCAFS SLC-40** Launch Site



The scatter plot shows Payload vs. Launch Outcome for CCAFS SLC-40 Launch Site. Two booster versions are shown: FT (blue) and B4 (red).

These are some observations bas ed on the provided plot.

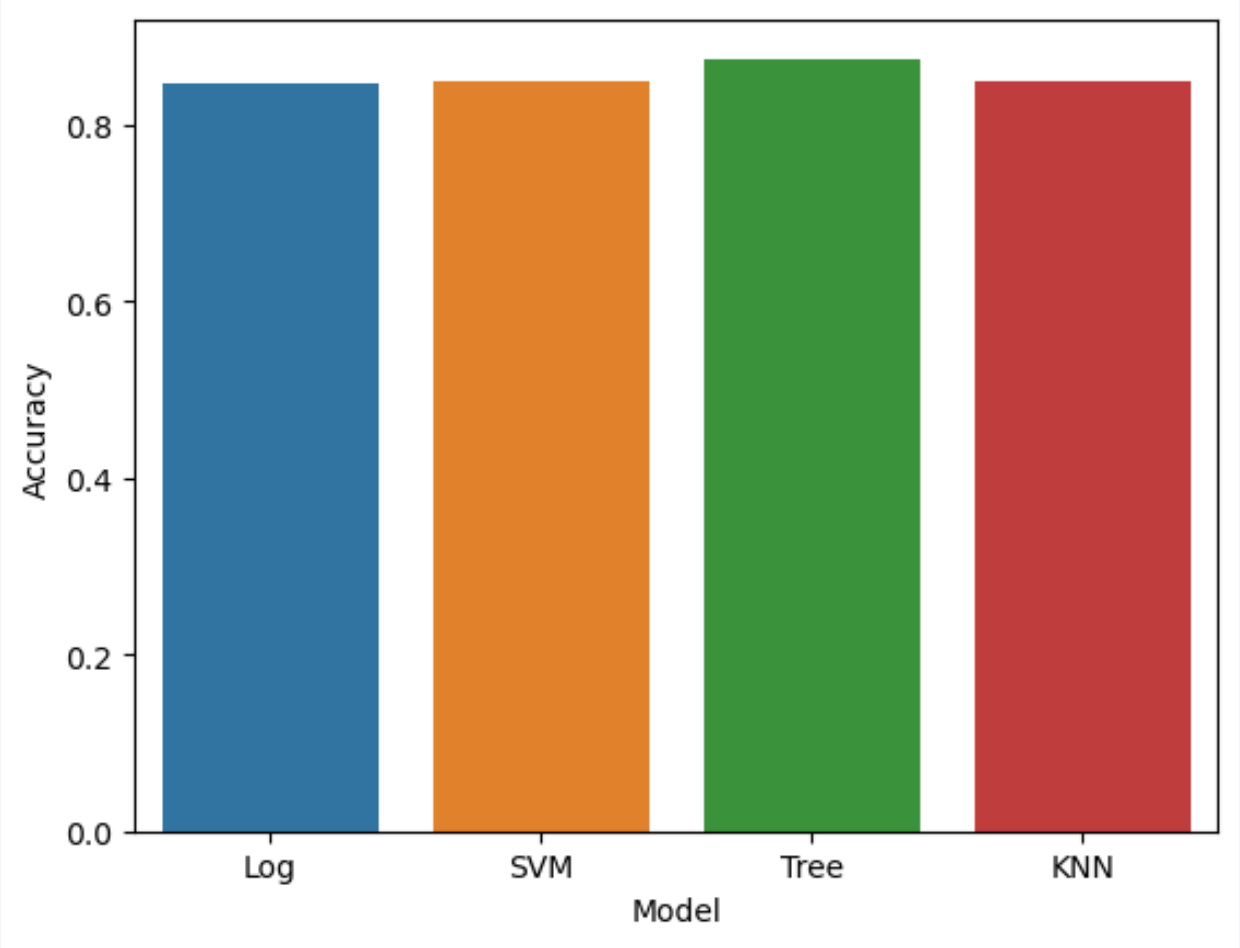
- The FT version has a 50% success rate, with one successful launch at around 2000 kg payload and one failure at approximately 4500 kg payload.
- The B4 version has three launches failing across varying payloads from about 2500 kg to over 6500 kg and one successful launch at around 3750 kg.

Section 5

Predictive Analysis (Classification)

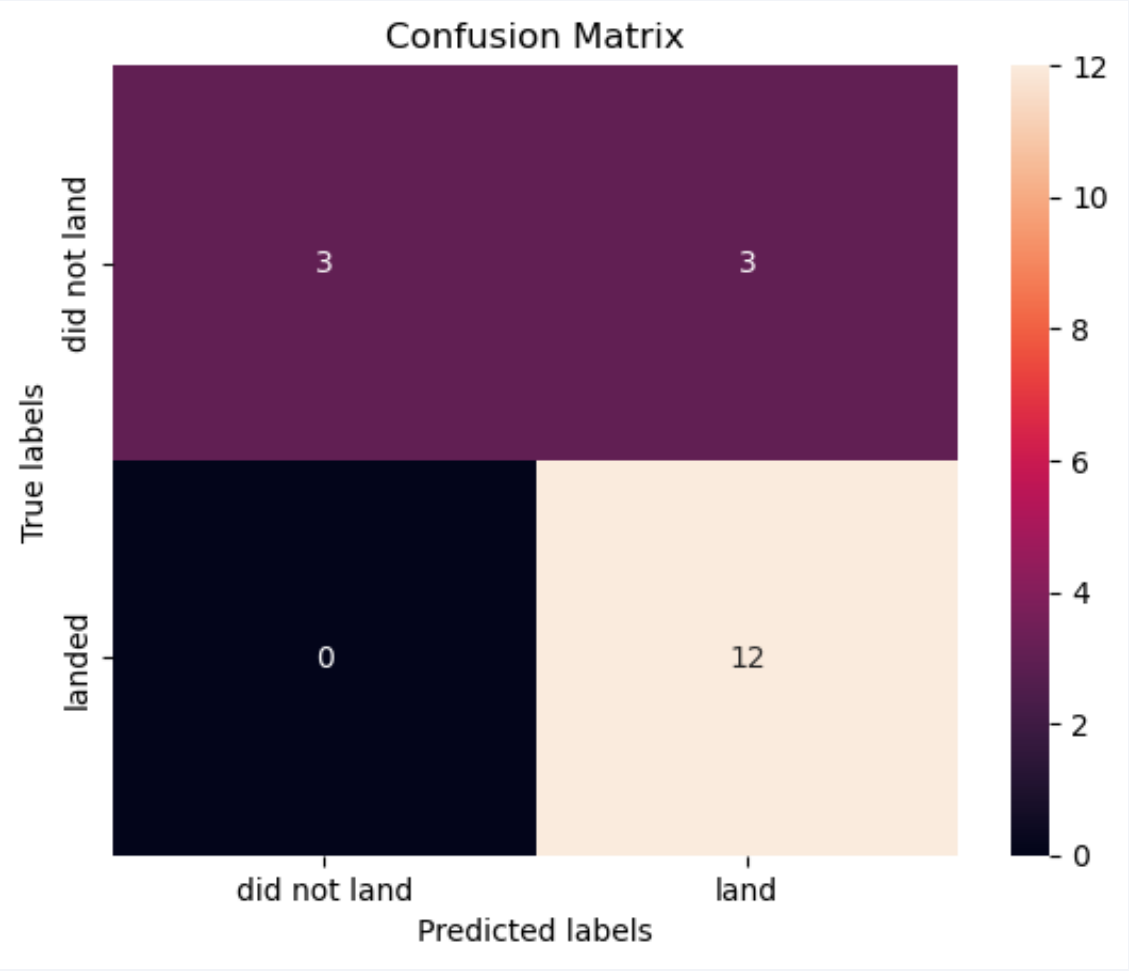
Classification Accuracy

- The Decision Tree Classifier has the highest classification accuracy.



Confusion Matrix

- The confusion matrix for the Decision Tree Classifier shows:
  - **True Positive (TP):** These are cases where the model predicted “landed”, and they did indeed land. In the matrix, there are 12 such instances.
  - **True Negative (TN):** The model predicted “did not land”, and they did not land. The matrix shows 3 such instances.
  - **False Positive (FP):** The model predicted “landed”, but they did not actually land. According to the matrix, there are 3 such instances.
  - **False Negative (FN):** The model predicted “did not land”, but they did land. In this matrix, there are no such instances.



Conclusions

- **Logistic Regression:**
  - Achieved an accuracy of 83.33% on the test data using a GridSearchCV to find the best hyperparameters.
  - Identified false positives as the major issue based on the confusion matrix analysis.
- **Support Vector Machine (SVM):**
  - Achieved an accuracy of 83.33% on the test data after hyperparameter tuning using GridSearchCV.
- **Decision Tree Classifier:**
  - Achieved the highest accuracy (87.5%) among the three models on the training data.
  - However, its performance on the test data remained at 83.33%.

Conclusions

- **Overall:**
  - All three models (Logistic Regression, SVM, Decision Tree) achieved similar accuracy (around 83.33%) in predicting the landing outcome on the test data.
  - Further investigation and model tuning might be necessary to improve the model's ability to generalize and achieve higher accuracy.

Conclusions

- **Final Conclusion:**
  - We've made significant progress in developing a tool to predict the outcome of SpaceX rocket launches.
- **Here's what we achieved:**
  - We built a system that analyzes various data points about the launch, like the weight of the payload and the launch location.
  - Based on this analysis, the system can predict with around 83% accuracy whether the first stage of the rocket will land successfully.
- **What this means for the future:**
  - This is a promising initial step. With further refinement and potentially using more data, we might be able to improve the prediction accuracy.

Conclusions

- **A more precise prediction tool could be valuable for:**
  - Cost estimation: Knowing the landing outcome beforehand allows for a more accurate cost assessment for each launch.
  - Decision-making: Stakeholders can make informed decisions based on the predicted landing outcome, potentially impacting launch strategies or resource allocation.
- **Next steps:**
  - We'll continue to explore ways to improve the model's accuracy. This might involve incorporating additional data sources or trying different machine learning techniques.
  - We can also investigate how to present the predictions in a way that is most helpful for decision-making.
  - Overall, this project demonstrates the potential of using machine learning to gain valuable insights from launch data and contribute to a more efficient and cost-effective space program.

Appendix

- Full Project link on GitHub (including all Notebooks):
  - https://github.com/KarimAboelfath/IBM-Applied-Data-Science-Capstone---KarimNasr
  - All plots, figures and charts are included in the notebooks in the GitHub repository.

Thank you!