IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

M. Panov
April 2024

Outline

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

Executive Summary

- Summary of methodologies
  1. Data Collection:
     - Utilize SpaceX REST API
     - Employ Web Scraping Techniques
  2. Exploratory Data Analysis (EDA):
     - Data Wrangling for Structured Data
     - Data Visualization for Insights
     - Interactive Visualization for Accessibility
  3. Predictive Analysis:
     - Employ Machine Learning Algorithms
     - Forecast Trends and Outcomes
- Summary of all results
  1. Key Insights for EDA:
     - Identified significant factors influencing launch success
     - Visualized trends and patterns in launch data
     - Established correlations between variables impacting mission outcomes
  2. Predictive Analysis results:
     - Identified the best-performing predictive model
     - Evaluated model accuracy and performance metrics
     - Provided insights for strategic decision-making and risk assessment

Introduction

- Project background and context
  - In the rapidly evolving landscape of commercial space travel, affordability and reliability are paramount. SpaceX has emerged as a frontrunner, offering cost-effective launches through innovations like reusable first stages. To compete in this space, Space Y aims to leverage data analytics and machine learning to predict launch success and determine pricing, enabling informed decision-making and strategic positioning in the market.

- Problems you want to find answers
  - The key challenges revolve around understanding the factors influencing the success of the first stage landing and its impact on launch cost determination. Specifically, the project seeks answers to:
    1. What are the primary factors influencing the successful landing of the first stage?
    2. How do these factors contribute to determining the cost of a launch?
    3. Can machine learning models effectively predict the success of the first stage landing based on historical data?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API and web scraping techniques.
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was collected through a combination of accessing SpaceX REST API and employing web scraping techniques to gather comprehensive historical and real-time launch data.

Data Collection – SpaceX API

Requesting rocket launch data from SpaceX API with an URL and GET method.

Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize().

Use the API again to get information about the launches using the IDs given for each launch.

Filter the dataframe to only include Falcon 9 launches.

Dealing with Missing Values using the .mean() and .replace() function to replace np.nan values.

- Jupyter Notebook: GitHub Link

Data Collection – Scraping

Send a request to the Falcon9 Launch Wiki page from its URL with HTTP GET method.

Create a BeautifulSoup object from the HTML response text content.

Collect all relevant column names from the HTML table header with find_all function.

Iterate through <th> elements to extract column names.

Create a data frame by parsing the launch HTML tables.

- Jupyter Notebook: GitHub Link

Data Wrangling

Identify and calculate the percentage of the missing values in each attribute. Identify which columns are numerical and categorical.

Calculate the number of launches on each site using .value_counts() function.

Determine the number and occurrence of each orbit in the column using .value_counts() function.

Calculate the number and occurrence of mission outcome of the orbits using .value_counts() function.

Create a landing outcome label from Outcome column.

- Jupyter Notebook: GitHub Link

EDA with Data Visualization

Scatter Plots

- Understand how Payload variable would affect the launch outcome.
- Visualize the relationship between Flight Number and Launch Site.
- Visualize the relationship between Payload and Launch Site
- Visualize the relationship between Flight Number and Orbit Type.
- Visualize the relationship between Payload and Orbit Type.

Bar Plot of Success Rate and Orbit Type:

- Visually check if there are any relationship between success rate and orbit type.

Line Plot of Launch Success by Year:

- Visualize the launch success yearly trend.

- Jupyter Notebook: GitHub Link

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

- Jupyter Notebook: GitHub Link

EDA with SQL

- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- 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.

- Jupyter Notebook: GitHub Link

Build an Interactive Map with Folium

- Exploratory Data Analysis (EDA) was conducted using Folium, a Python library for visualizing geospatial data, to enhance the understanding of launch sites and outcomes. This included:
  - Adding Circle Markers to depict launch sites with pop-up labels and text labels.
  - Utilizing Colored Markers to represent launch outcomes, grouped in marker clusters.
  - Incorporating Lines to illustrate distances between launch sites and points of interest.
- Various Folium functions such as Folium.marker(), Folium.circle(), and Folium.polyline() were employed to create visual elements on the map. Markerclusters() helped simplify maps with multiple markers sharing identical coordinates, while MousePosition facilitated obtaining coordinates upon hovering over map points.
- Jupyter Notebook: GitHub Link

Build a Dashboard with Plotly Dash

- An interactive Dashboard was created using Plotly Dash, showcasing the following functionalities:
  - Dropdown list for selecting launch sites.
  - Pie charts presenting success rates for individual or all sites.
  - Slider to choose payload mass range.
  - Scatterplots depicting the correlation between payload mass and success rate.

- Jupyter Notebook: GitHub Link

Predictive Analysis (Classification)

Create a NumPy array from the column Class in data, by applying the method to_numpy().

Standardize the data with preprocessing.StandardScaler() function and fit X data as type float.

Split the data into training and testing data using the function train_test_split.

Models are trained and hyperparameters are selected using the function GridSearchCV.

Find which method performs best using test data.

- Jupyter Notebook: GitHub Link

Results

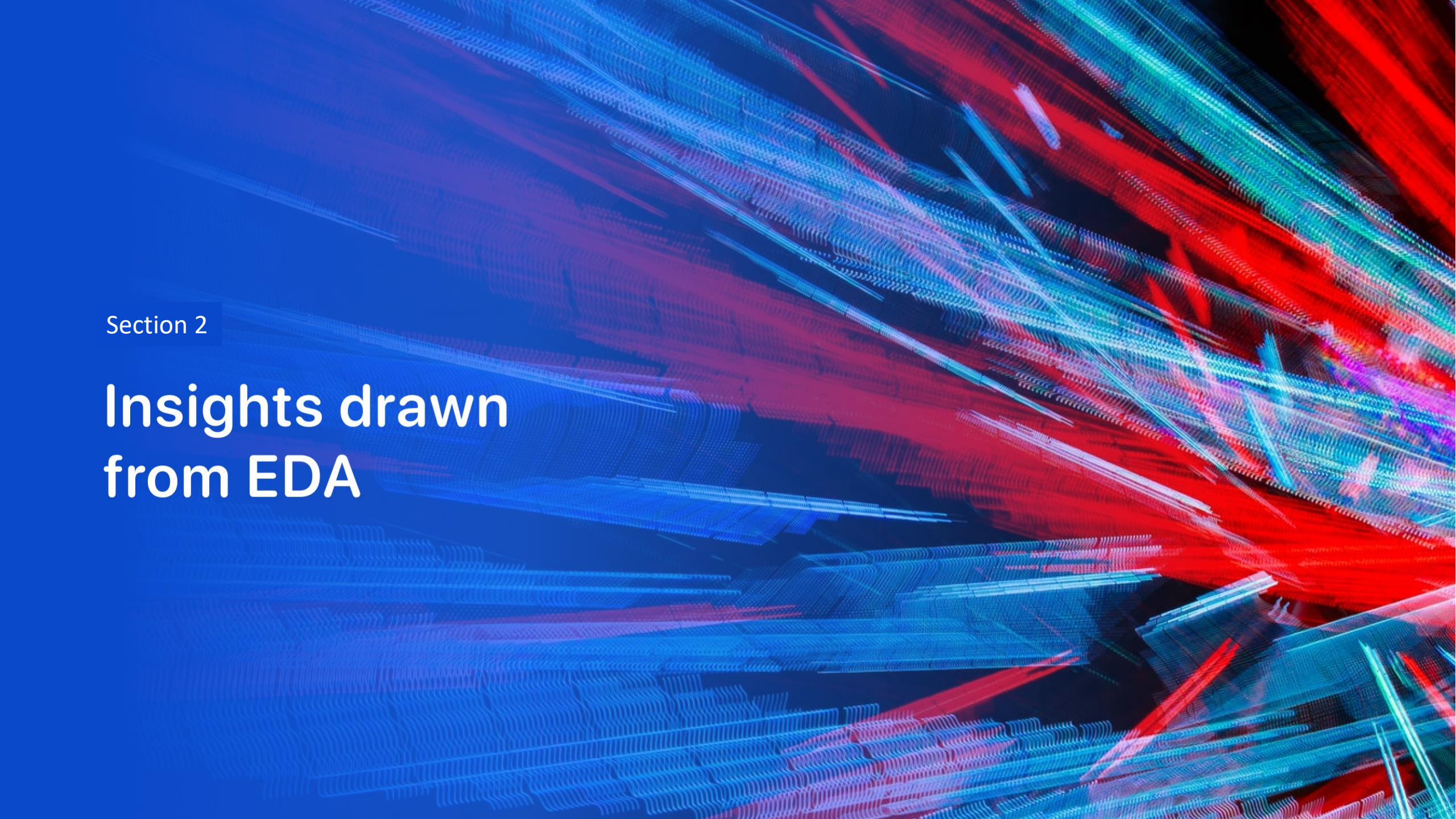
Exploratory data analysis results

- By Visualization
- By SQL
- By Folium Map
- By Interactive Dashboard

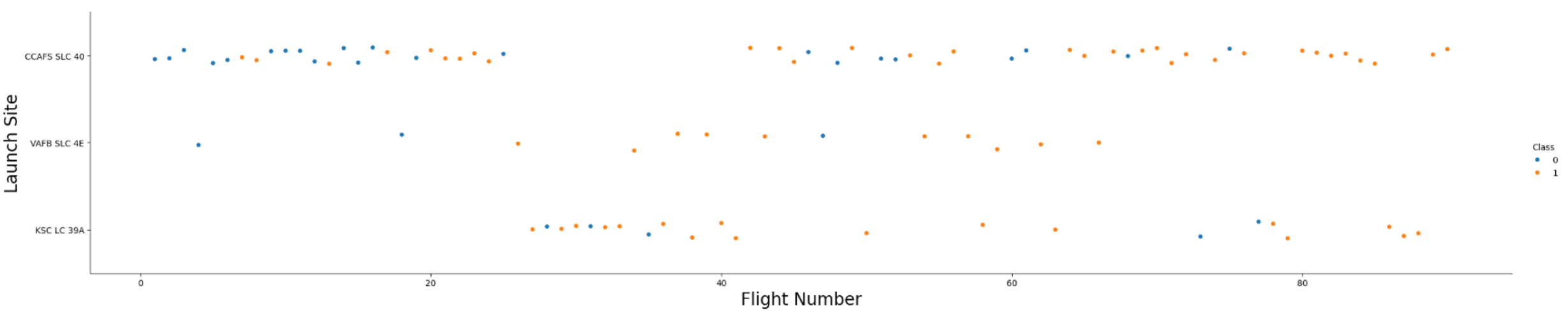
Interactive analytics demo in screenshots

Predictive Analysis results

- By Classification

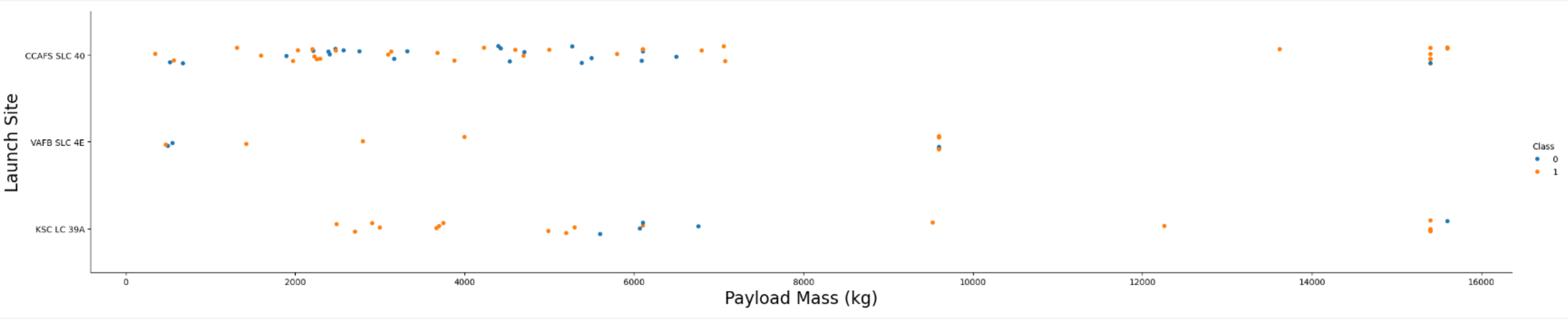

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



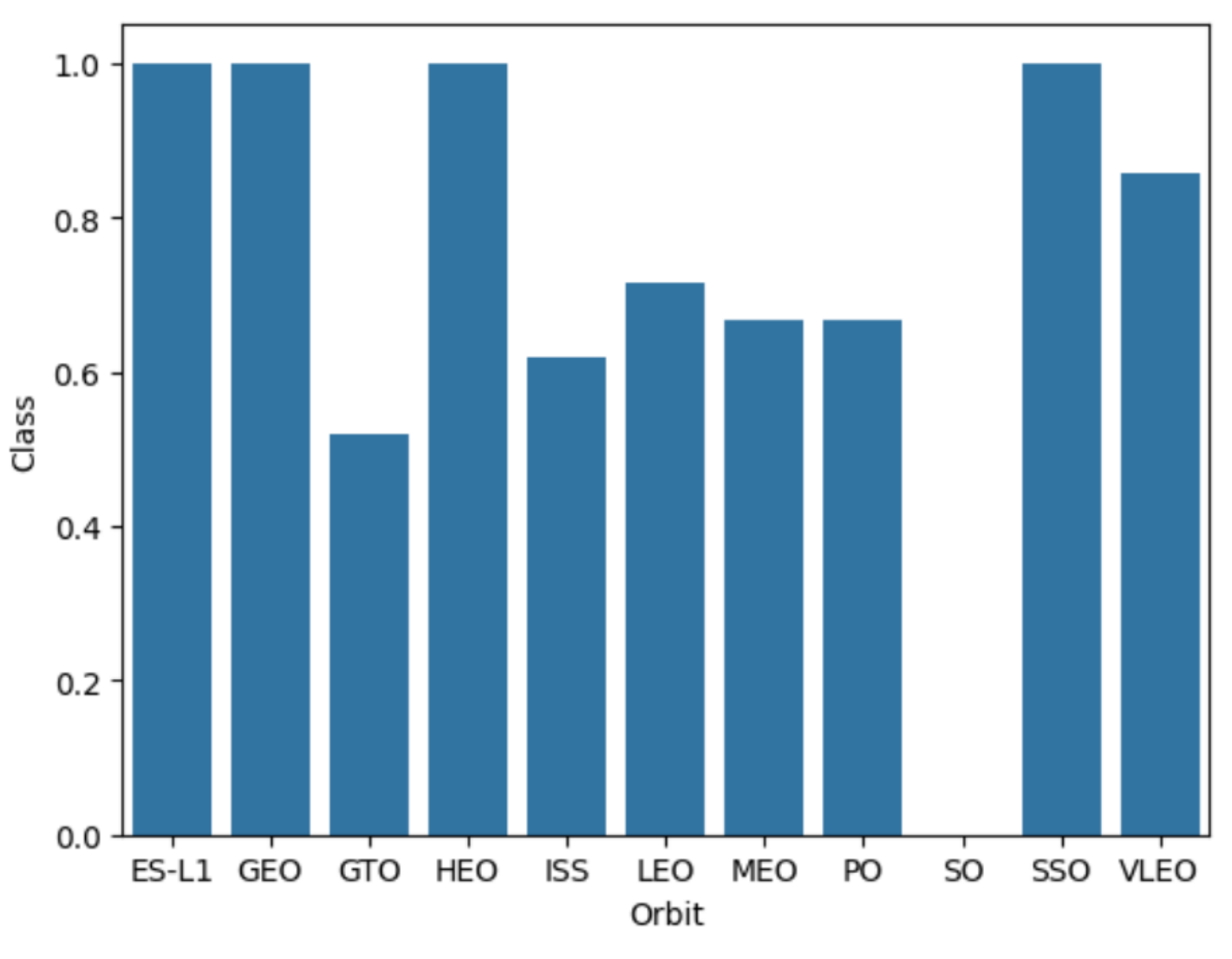
- The scatter plot represents the correlation between flight numbers and launch sites, with data points differentiated by two classes: unsuccessful (0) and successful (1) launches.
  - Certain launch sites seem to have more successful launches (VAFB SLC 4E success rate of 77% and KSC LC 39A of 77%) than others (CCAFS SLC40 of 60%).
  - As the flight numbers increase, the launch sites being used appear to change. This could be due to a variety of factors such as improvements in technology, changes in mission requirements, or lessons learned from previous launches.
  - Most launches are from CCAFS SLC40 (Total of 55).

Payload vs. Launch Site



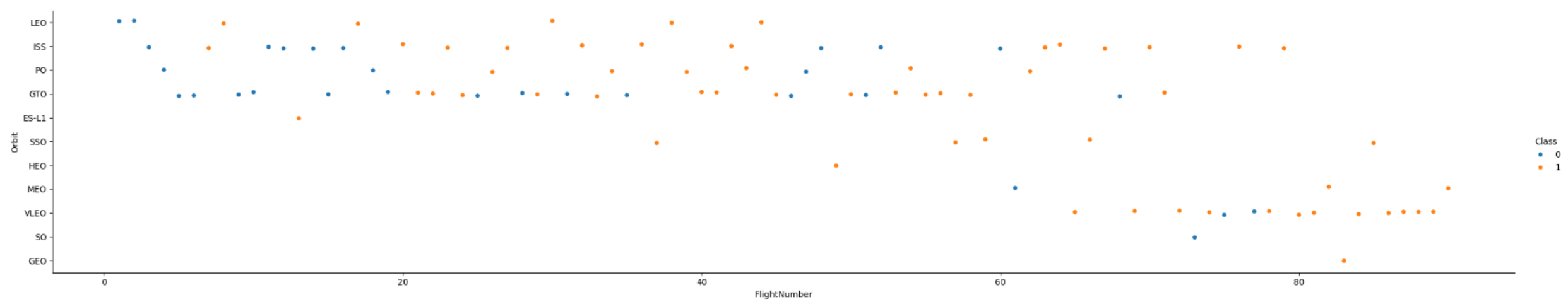
- The scatter plot represents the correlation between payload mass and launch sites, with data points differentiated by two classes: unsuccessful (0) and successful (1) launches.
  - The plot shows a correlation between different launch sites and payload mass, which could indicate that certain launch sites are specialized or more frequently used for heavier payloads.
  - For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000 kg).
  - Success rate for heavy payload mass launches is higher.
  - KSC LC 39A has a 100% success rate for payload mass under 5700 kg.

Success Rate vs. Orbit Type



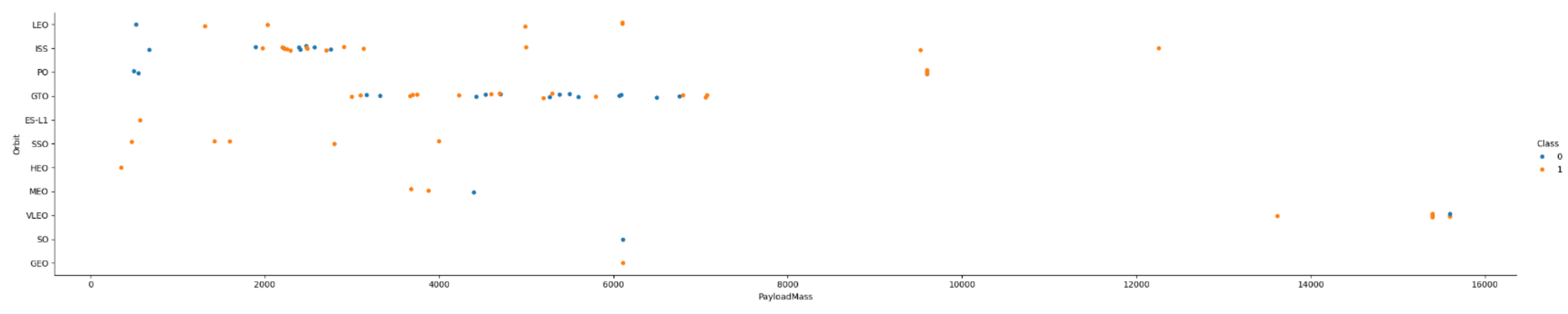
- The bar plot represents the relationship between success rate and orbit type.
  - The plot shows a correlation between different launch sites and payload mass, which could indicate that certain launch sites are specialized or more frequently used for heavier payloads.
  - The success rate of ES-L1, GEO, HEO, and SSO is 100%.
  - The success rate of SO is 0%.
  - Other orbits with low success rate are GTO and ISS (~50%-60%).

Flight Number vs. Orbit Type



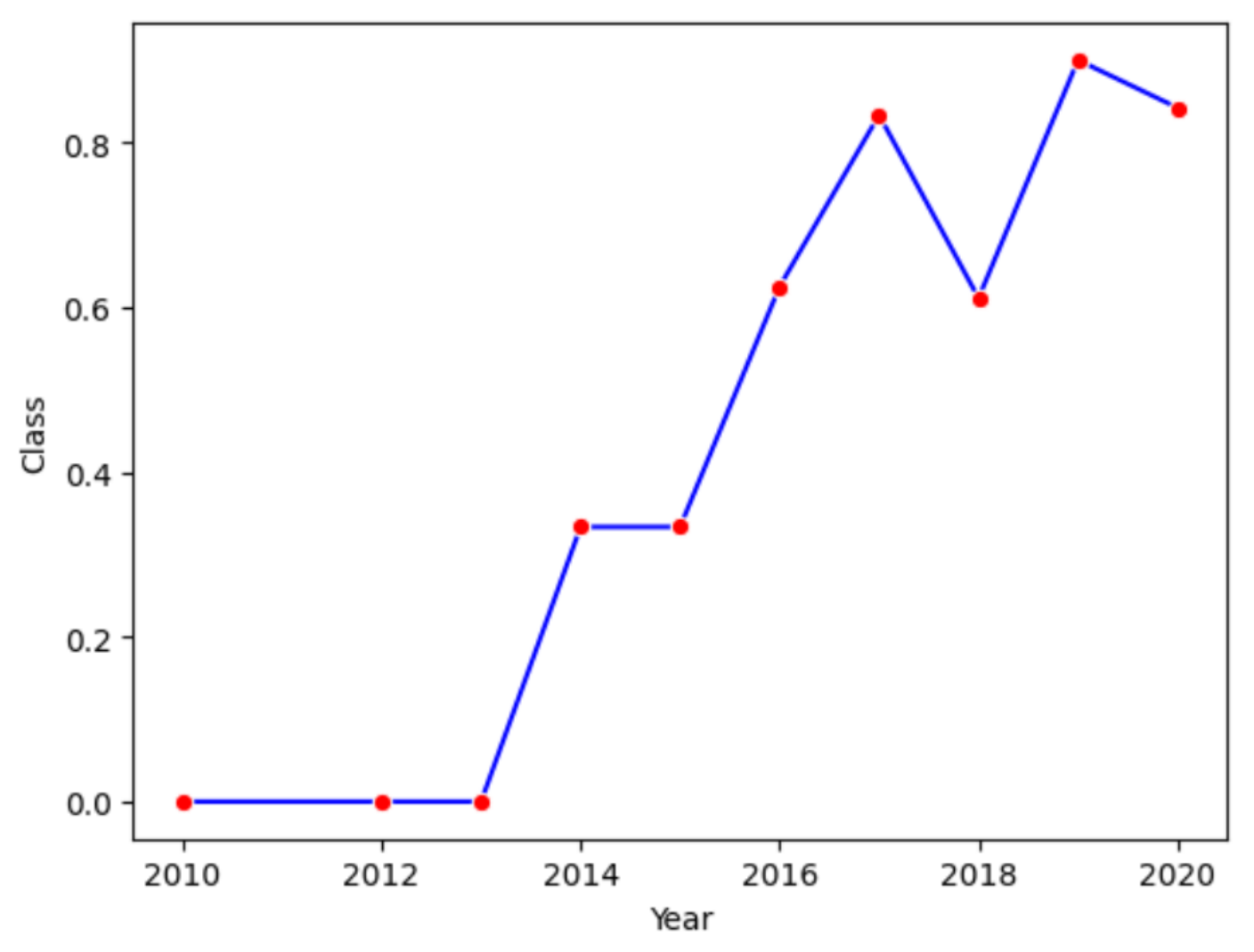
- The scatter plot represents the correlation flight number and orbit type, with data points differentiated by two classes: unsuccessful (0) and successful (1) launches.
  - SSO success rate is 100%.
  - Orbits used for later launches are SSO, HEO, MEO, VLEO, SO, and GEO.

Payload vs. Orbit Type



- The scatter plot represents the correlation payload mass and orbit type, with data points differentiated by two classes: unsuccessful (0) and successful (1) launches.
  - With heavy payloads the successful landing are more for PO, LEO and ISS.
  - Heavier payloads have better success rates.
  - GTO is used for missions with payload under 8000 kg.
  - VLEO is used for mission with most heavy payload above 13000 kg.

Launch Success Yearly Trend



- The line plot represents the relationship between success rate and year.
  - 4 years were needed for the first successful launch (in 2014).
  - A significant increase is observed starting in 2016, indicating a notable event or change that year.
  - In 2016 the success rate reached above 50%.
  - The recent 4 years the mission have success rare of above 80% with exception in 2018.

All Launch Site Names

```
%sql SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE
```

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

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

- Used to retrieve unique values from the "Launch_Site" column in the SPACEXTABLE, ensuring that each launch site appears only once in the result set.
- 4 launch sites are recognized.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE '%CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- Retrieves up to 5 records from the SPACEXTABLE where the Launch_Site column contains the substring "CCA“.

Total Payload Mass

```
%sql SELECT Customer, SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

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

Customer	SUM(PAYLOAD_MASS__KG_)
NASA (CRS)	45596

- Retrieves the total payload mass (in kilograms) for launches carried out for NASA under the CRS program from the SPACEXTABLE.
- Total payload mass is 45,596 kg.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version LIKE '%F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- Calculates the average payload mass (in kilograms) for launches involving booster versions containing the substring "F9 v1.1" in the SPACEXTABLE.
- Average payload mass is 2,534.67 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

- Retrieves the earliest date of successful ground pad landings from the SPACEXTABLE.
- It is on 22nd of December 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version, Landing_Outcome, PAYLOAD_MASS__KG_ FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ >= 4000 AND PAYLOAD_MASS__KG_ <= 6000
```

```
 * 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

- Retrieves data from the SPACEXTABLE, including the booster version, landing outcome (specifically 'Success (drone ship)'), and payload mass (in kilograms) for launches where the payload mass falls within the range of 4000 kg to 6000 kg and the landing outcome is successful on a drone ship.
- 4 Booster Versions are found to fulfill the requirement.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

- Calculates the count of occurrences for each unique mission outcome in the SPACEXTABLE and presents it alongside the respective mission outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
ORDER BY Booster_Version
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

- Retrieves the booster version and payload mass (in kilograms) from the SPACEXTABLE where the payload mass is equal to the maximum payload mass recorded in the table. The results are ordered by booster version.

2015 Launch Records

```
%sql SELECT SUBSTR(Date, 6, 2) AS Month, 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.
```

Month	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

- Extracts the month from the Date column, labels it as "Month", and selects Landing_Outcome, Booster_Version, and Launch_Site from the SPACEXTABLE. It filters the results to include only failures on drone ships and launches that occurred in 2015.
- Two months with failure outcomes on Drone ship are found - January and April 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Count_Landing_Outcome FROM SPACEXTABLE WHERE Date >= '2010-06-04' AND Date <= '2017-03-20'
AND Landing_Outcome LIKE '%Success%' GROUP BY Landing_Outcome ORDER BY Count_Landing_Outcome DESC
```

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

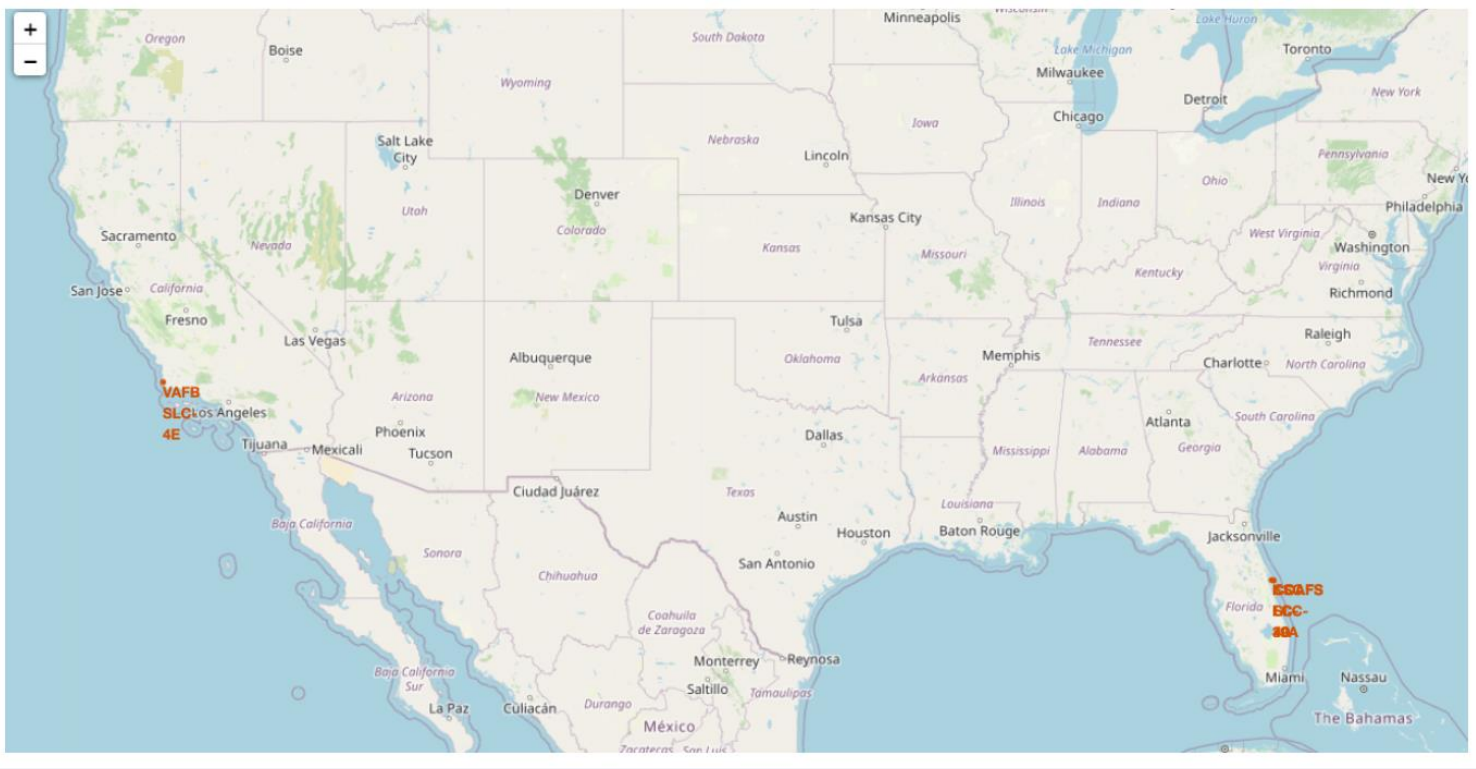
Landing_Outcome	Count_Landing_Outcome
Success (drone ship)	5
Success (ground pad)	3

- Retrieves the count of successful landing outcomes within the specified date range (from June 4, 2010, to March 20, 2017) from the SPACEXTABLE. It groups the results by landing outcome and orders them in descending order based on the count of each outcome.
- A total of 8 successful landings are conducted in this period – 5 at Drone ship and 3 at Ground pad.

Section 3

Launch Sites Proximities Analysis

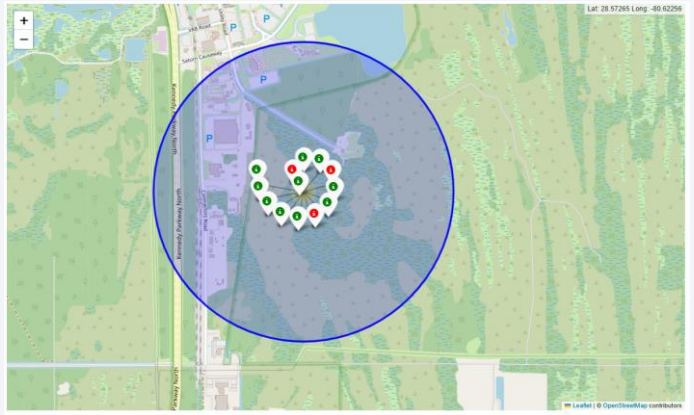
Launch site's locations



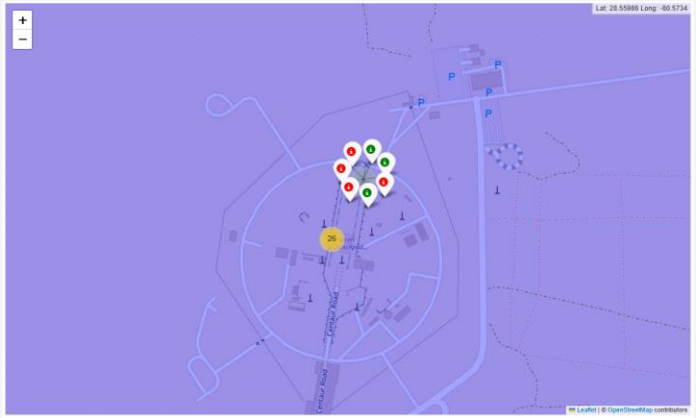
- Launching from near the equator provides a significant velocity boost due to the Earth's rotation. This equatorial advantage reduces the amount of fuel needed for spacecraft to achieve desired orbits, making launches more cost-effective.
- Launching over the ocean minimizes risks to populated areas in case of launch failures or malfunctions. Launching from coastal areas allows rockets to fly over uninhabited regions and provides a clear flight path, reducing potential hazards to people and property.
- Coastal launch sites offer easy access to water for transportation of rocket components, equipment, and fuel. Water transport is often more cost-effective than overland transportation, especially for large and heavy payloads.

Map with Launch outcome

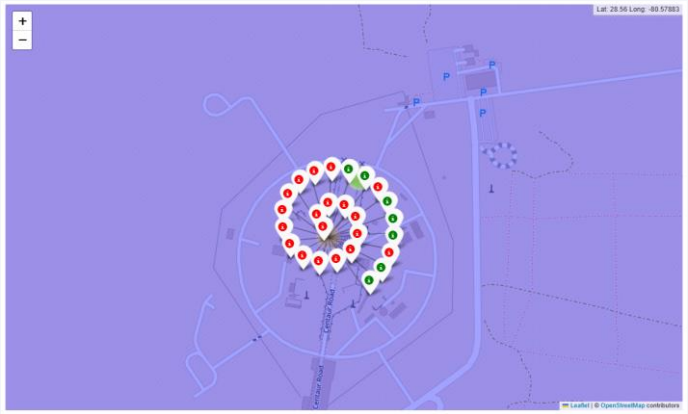
- VAFB SLC-40



- KSC LC-39A



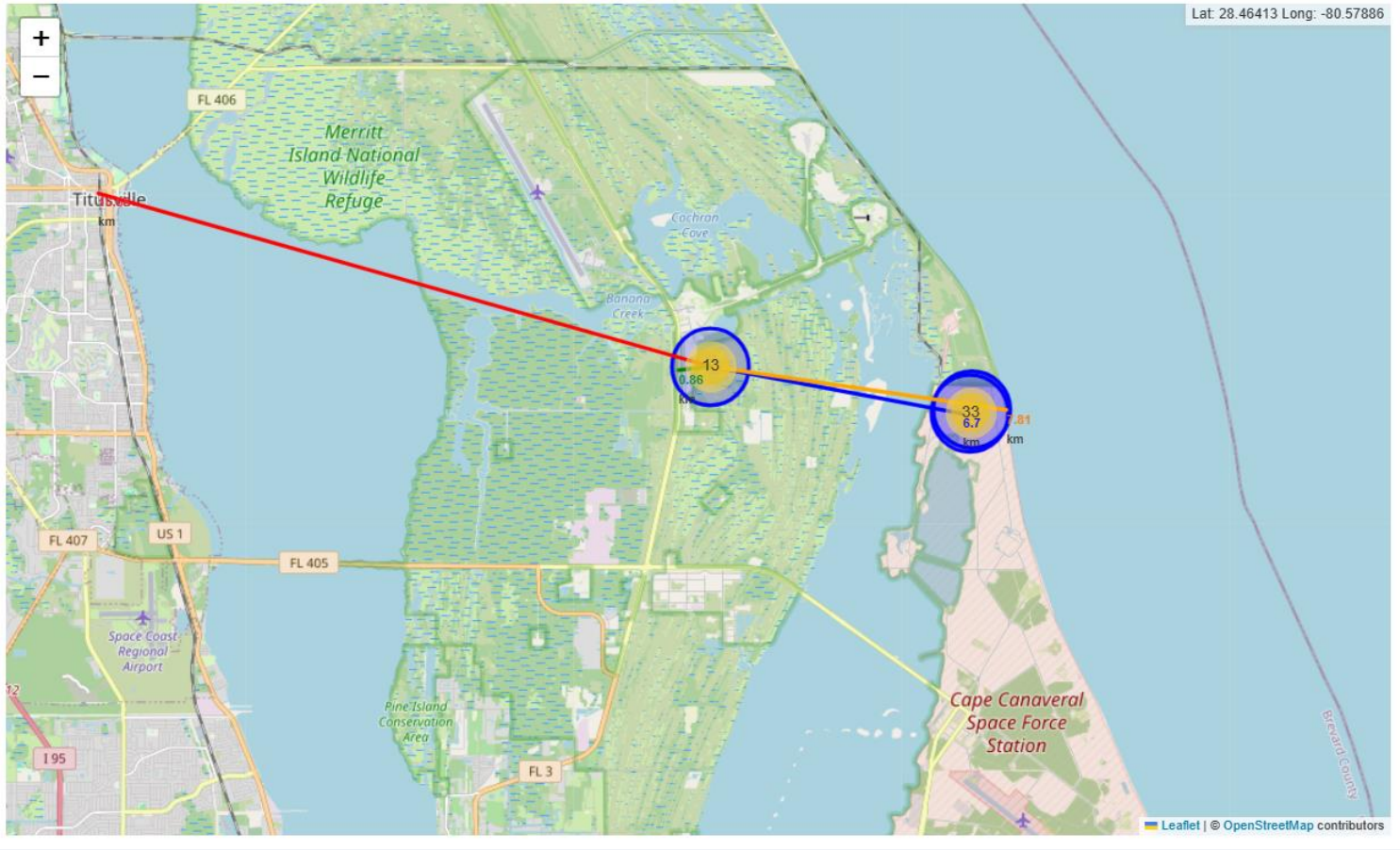
- CCAFS SLC-40



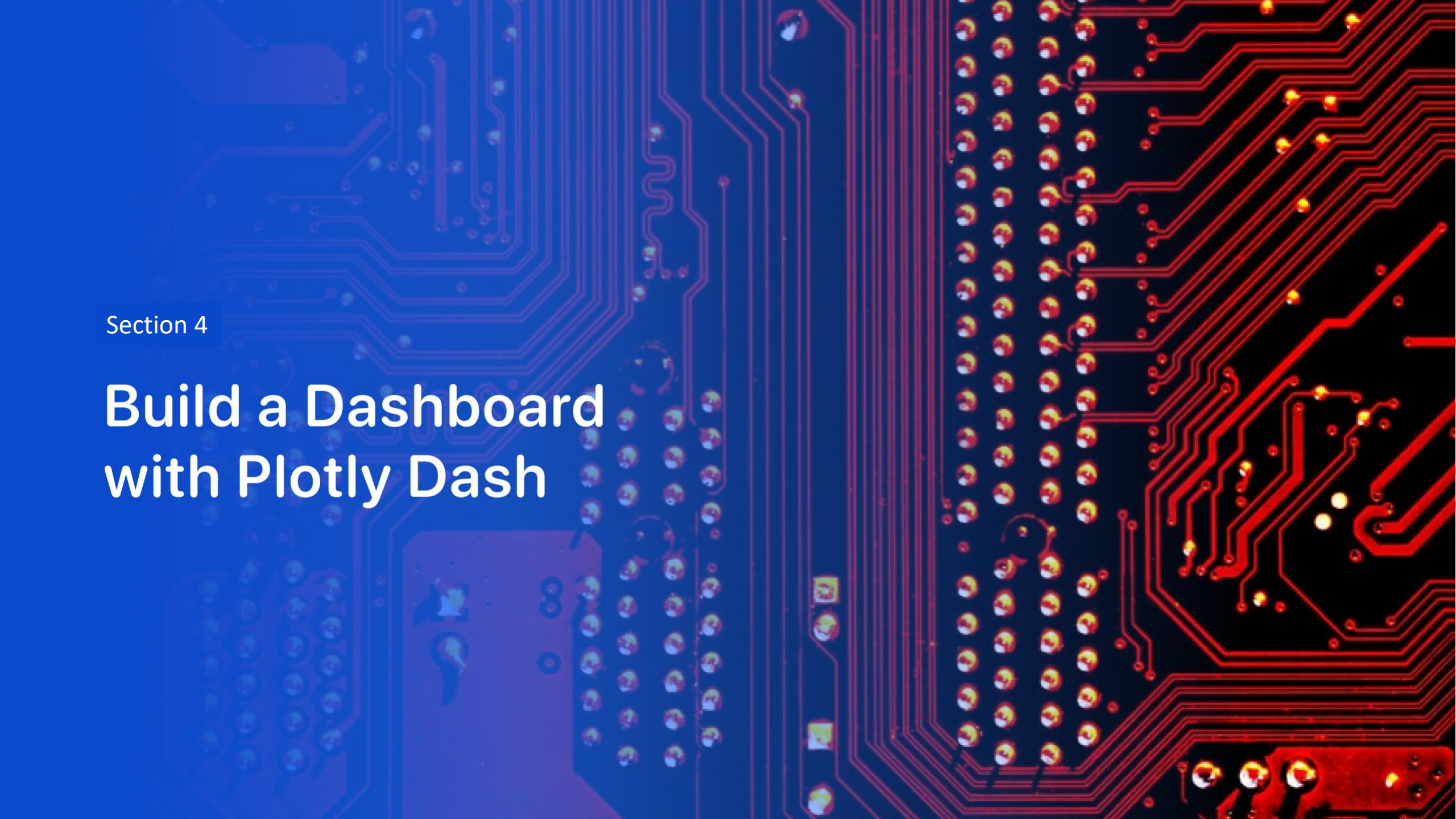
- CCAFS LC-40

- Easy visual identification of the success rate of each launch site.
- Red = Failure launch
- Green = Successful launch

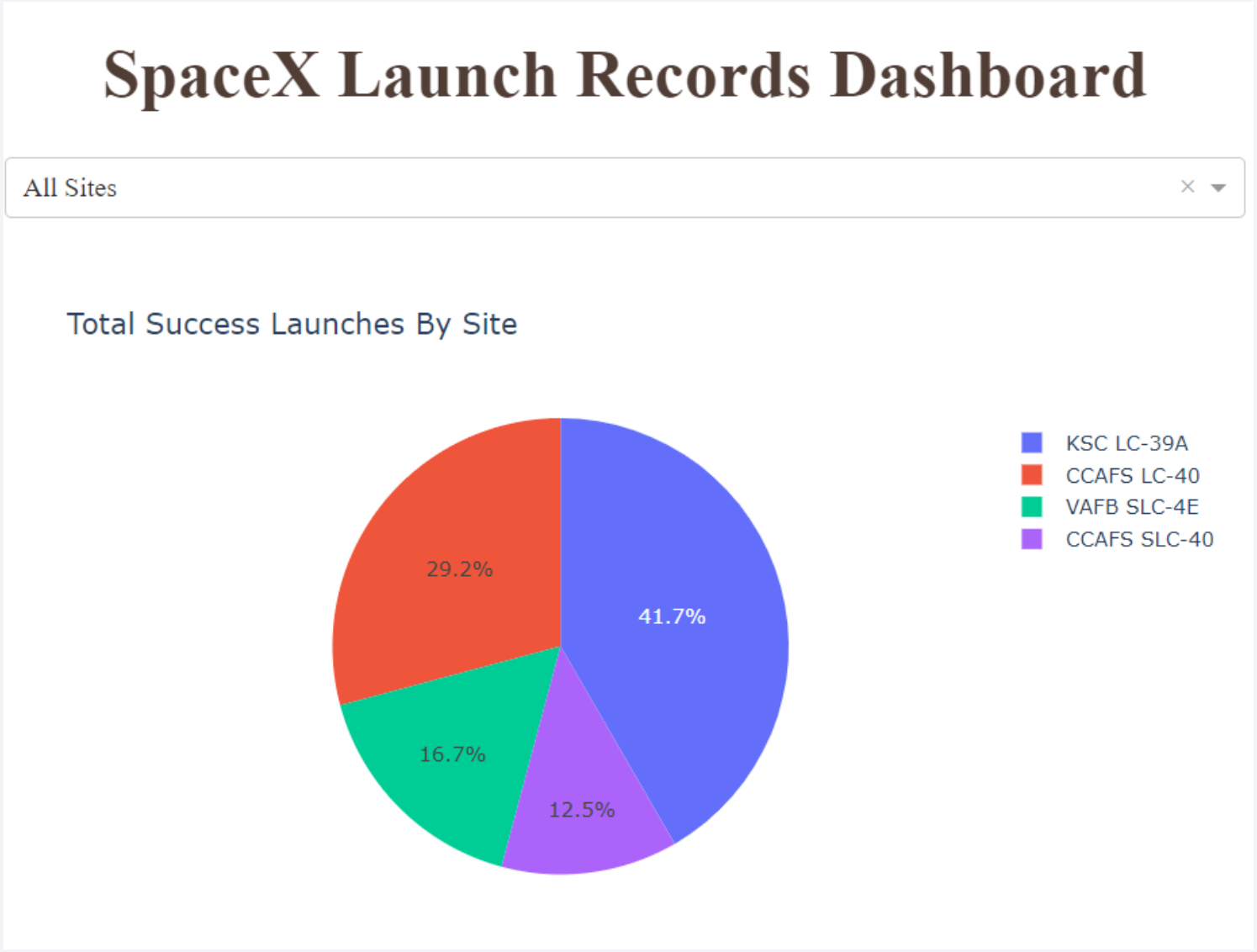
Launch map proximity



- This map calculates and visually illustrates the distance between the launch pad and various key features such as cities, highways, coastline, and more.

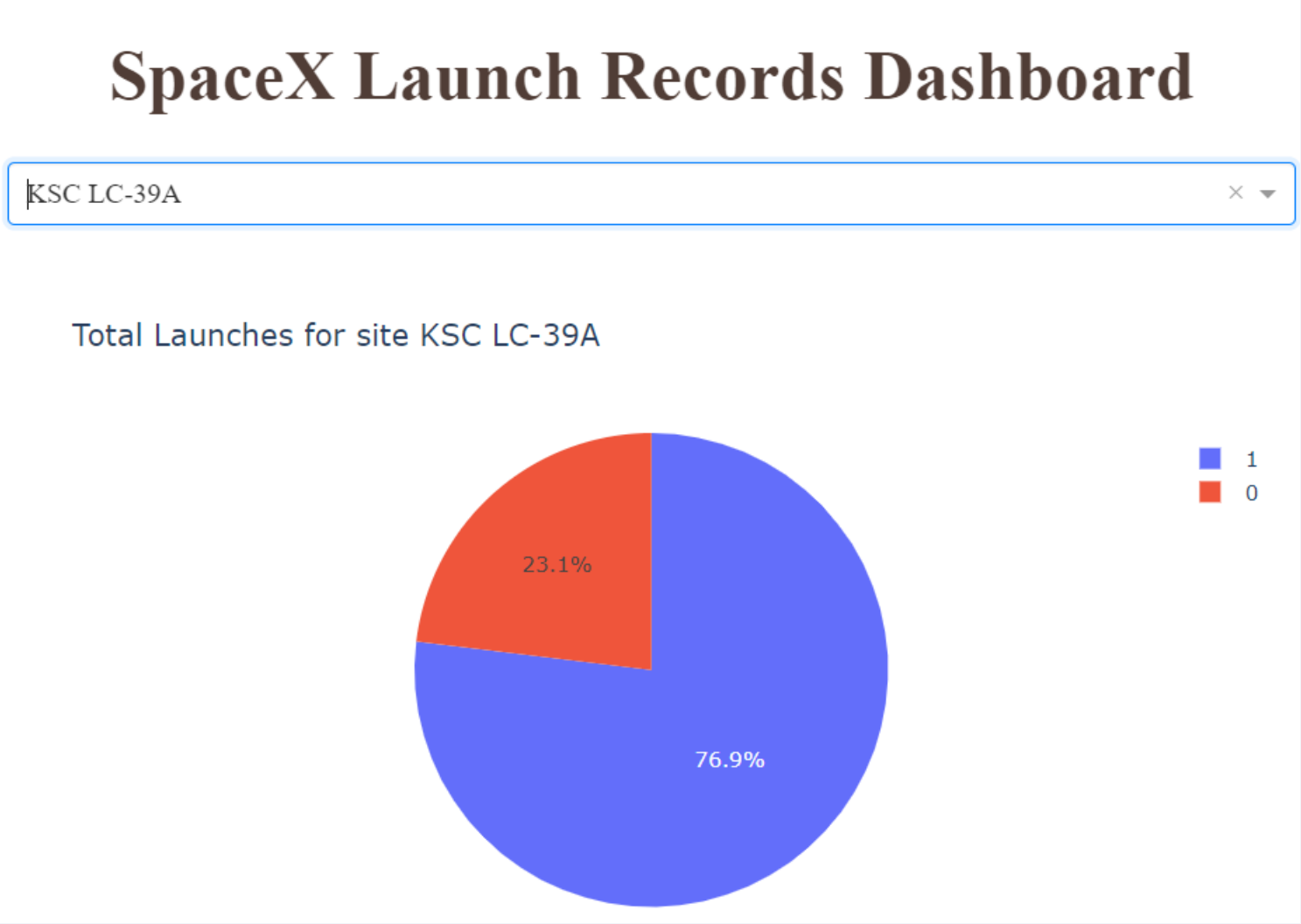

Section 4
Build a Dashboard with Plotly Dash

Interactive dashboard – Total Success Launches



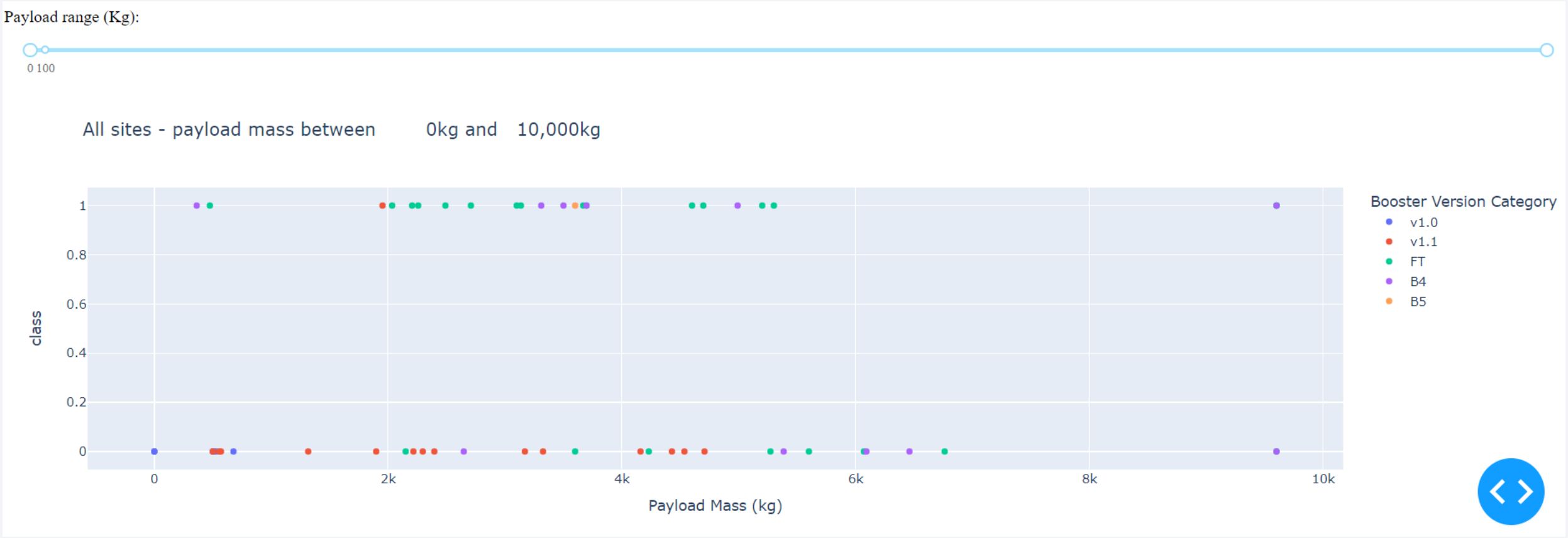
- Pie chart represent the total successful launches by launch site.
- KSC LC-39A has most successful launches.

Interactive dashboard – KSC LC-39A success rate



- Pie chart represent the success rate oft the launch site with the most successful launches which is KSC LC-39A.
- Success rate is 76.9%.

Interactive dashboard – Payload mass and outcome



- Visualize the effect of payload mass on launch outcome across various booster category versions.

Section 5

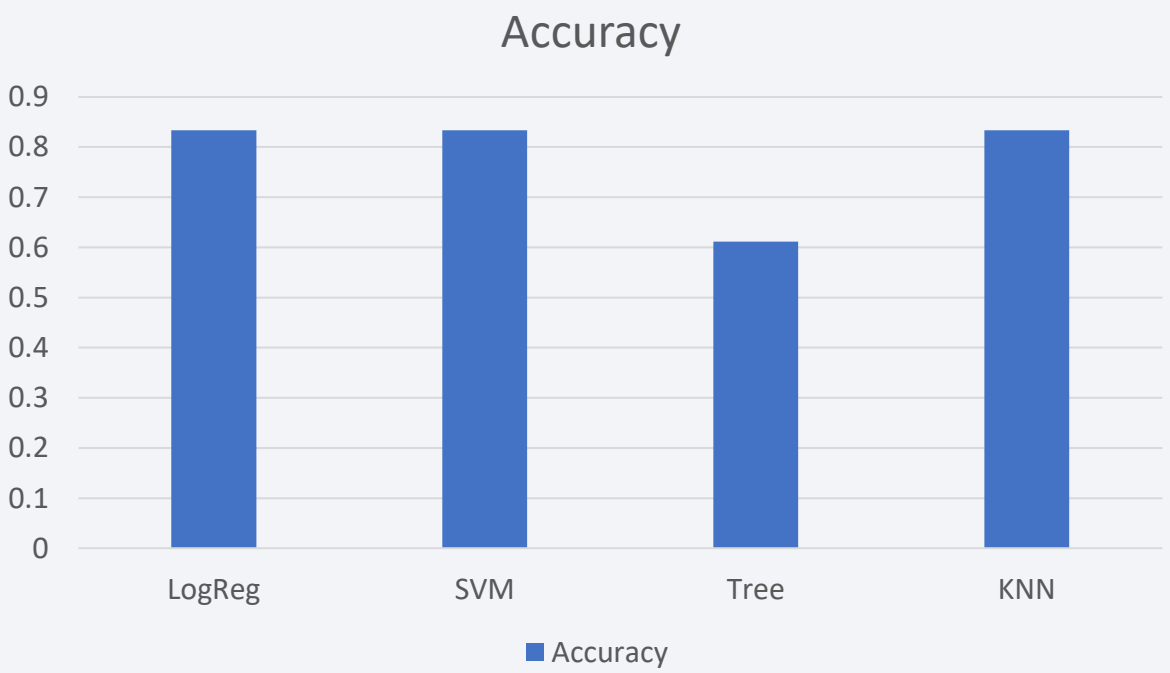
Predictive Analysis (Classification)

Classification Accuracy

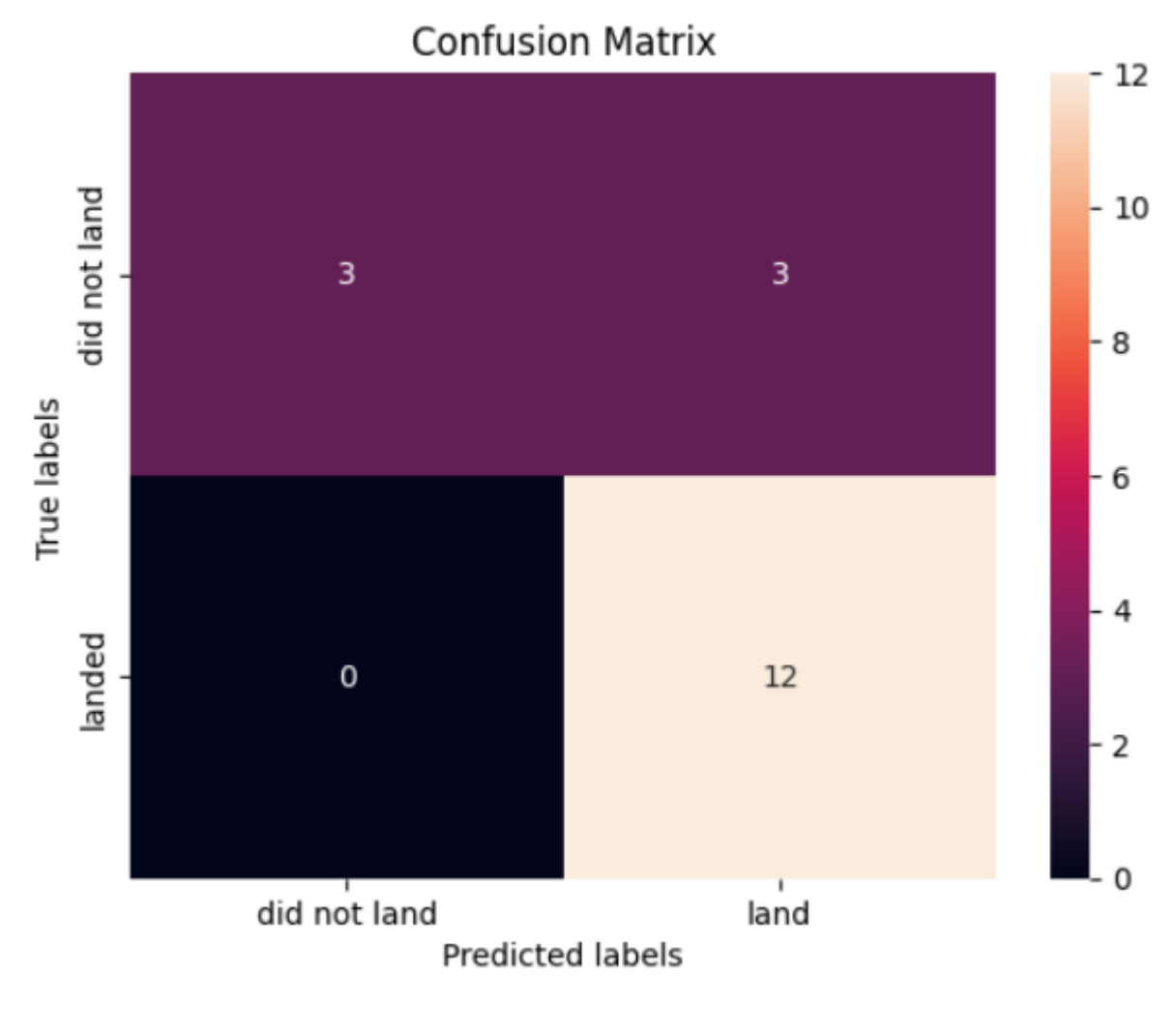
```
print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.6111111111111112
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.533333	0.800000
F1_Score	0.888889	0.888889	0.695652	0.888889
Accuracy	0.833333	0.833333	0.611111	0.833333



Confusion Matrix



- The confusion matrix shows a high rate of true negatives.

Conclusions

- Correlation exists between launch sites and payload mass, suggesting site specialization.
- KSC LC 39A achieves a 100% success rate for payloads under 5700 kg.
- Some orbits consistently achieve 100% success rates, while others, like SO, have a 0% success rate.
- Notable increase in successful launches is observed starting in 2016, with success rates exceeding 50%.
- Recent years show consistent success rates above 80%, with exceptions in 2018.
- KSC LC-39A stands out with the highest success rate of 76.9%.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!