IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SANA'A SALEM ALOUFI

15 / 10 / 2024

Outline

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

Executive Summary

My team and I are data scientists working for a new rocket company, Space Y. Space Y aims to compete with Space X by leveraging advanced data science techniques. Our project focuses on gathering and analyzing information about Space X launches to create insightful dashboards.

Key Objectives:

- Predict Reusability: Use machine learning models to predict if Space X will reuse the first stage of their rockets.
- Forecast Landing Success: Train models using public data to predict the likelihood of successful first stage landings

Introduction

- In this capstone project, we aim to predict whether the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches at a cost of $62 million, significantly lower than the $165 million charged by other providers. This cost efficiency is largely due to SpaceX's ability to reuse the first stage of the rocket. By predicting the success of the first stage landing, we can estimate the cost of a launch, which is valuable information for companies looking to compete with SpaceX.
- There are several scenarios where the booster did not land successfully. For instance, a landing attempt might fail due to an accident.

- The outcomes are categorized as follows:
- **True Ocean**: The mission successfully landed in a specific region of the ocean.
- **False Ocean**: The mission unsuccessfully landed in a specific region of the ocean.
- **True RTLS (Return to Launch Site)**: The mission successfully landed on a ground pad.
- **False RTLS**: The mission unsuccessfully landed on a ground pad.
- **True ASDS (Autonomous Spaceport Drone Ship)**: The mission successfully landed on a drone ship.
- **False ASDS**: The mission unsuccessfully landed on a drone ship.

Section 1

Methodology

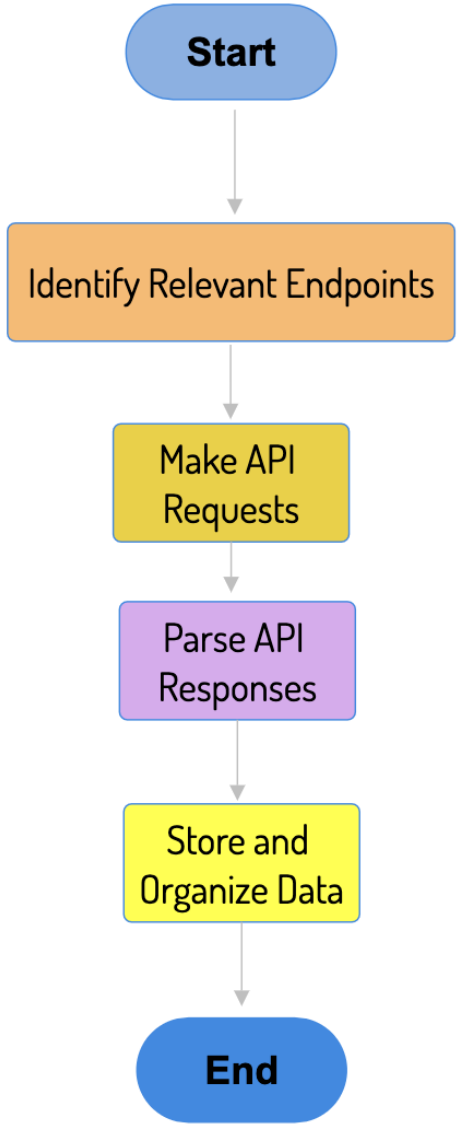
Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

- First, We will collect and ensure the data is in the correct format from an API, by making a GET request to the SpaceX API.
- Then, we will perform some basic data wrangling and formatting through the following steps:
  - Request data from the SpaceX API.
  - Clean the requested data.
- Finally, we Save the collected data and export it to a CSV file .



Data Collection – SpaceX API

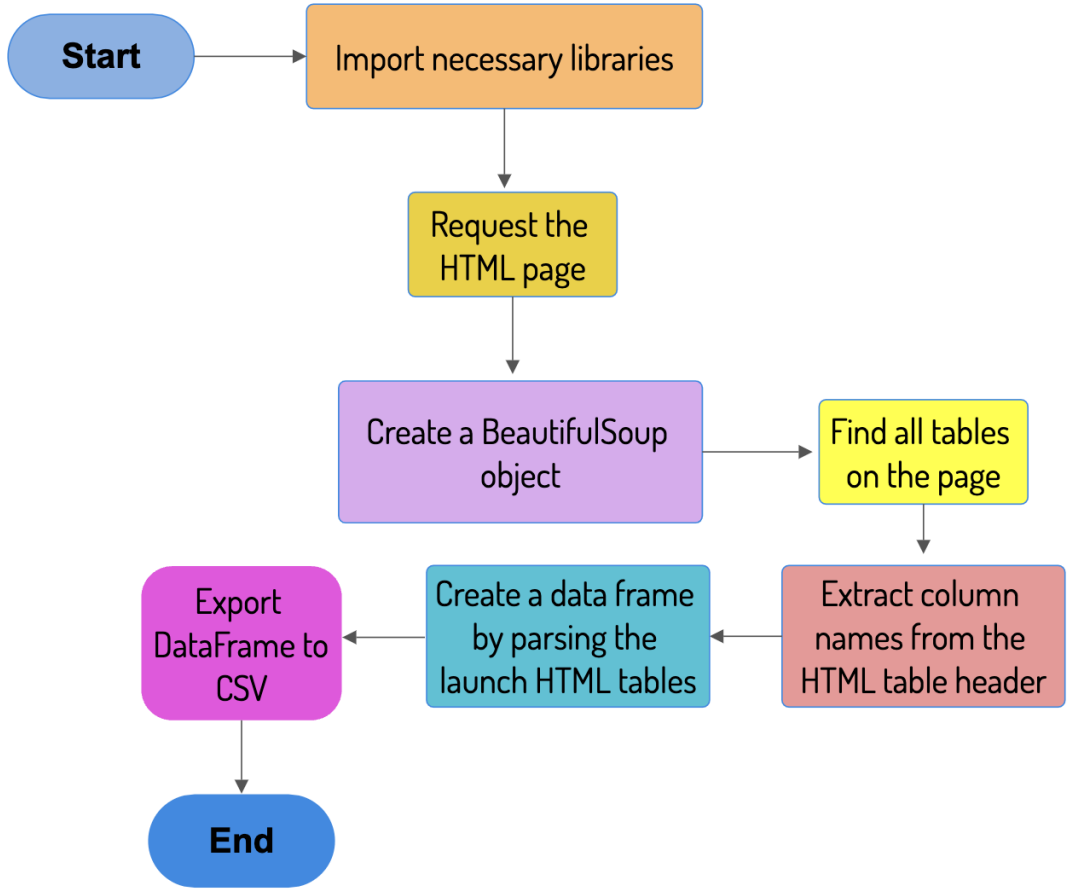
The goal of this task is to collect data from SpaceX using their REST API. This involves:

- **Identifying Relevant API Endpoints:** Determining the specific API endpoints that provide the desired data.
- **Making API Requests:** Sending HTTP requests to these endpoints using GET request .
- **Parsing API Responses:** Extracting and processing the SpaceX launch data.
- **Storing and Organizing Data:** Saving the collected data and export it to a CSV file .

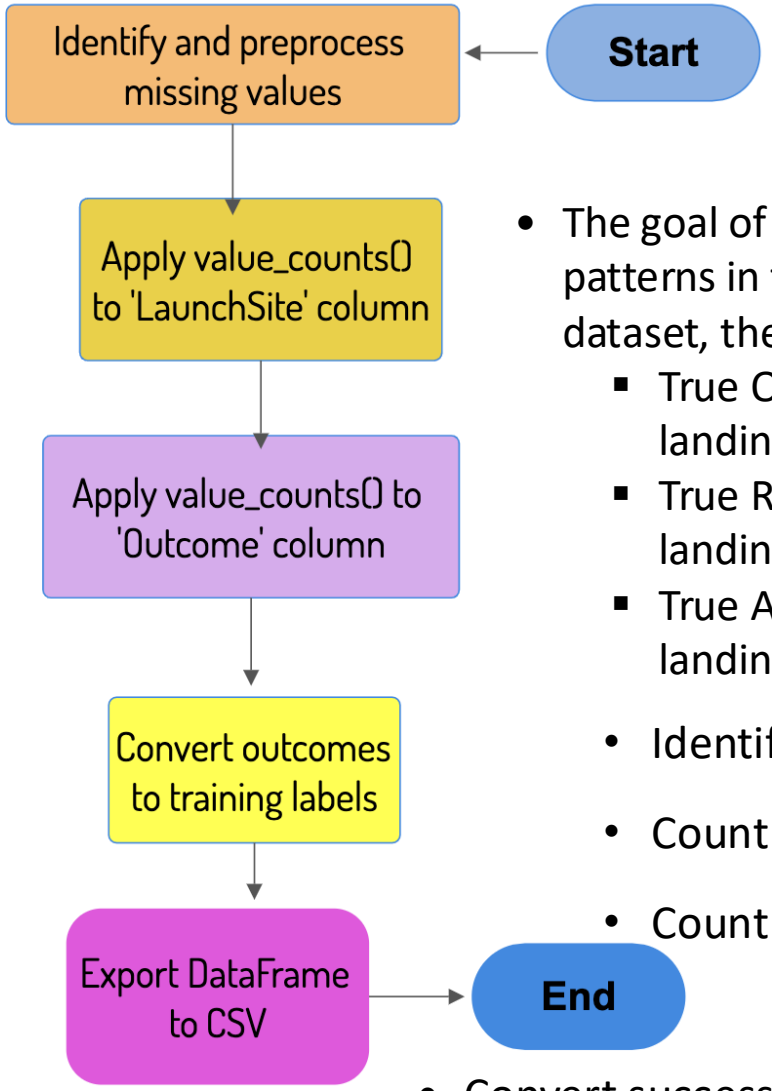
- GitHub URL of SpaceX API calls notebook **here**

Data Collection -Scraping

- The goal of this task is to web scraping to collect Falcon 9 launch records with BeautifulSoup through followeing steps:
  - Extract a Falcon 9 launch records HTML table from Wikipedia
  - Parse the table and convert it into a Pandas data frame
  - Saving the scraped data and export it to a CSV file
- GitHub URL of the completed web scraping notebook **here**



Data Wrangling



- The goal of this task is to perform some Exploratory Data Analysis (EDA) to identify patterns in the data and determine the labels for training supervised models. In the dataset, there are several different cases regarding where the booster landed:
  - True Ocean or False Ocean: Indicates whether the mission outcome was a successful landing in a specific region of the ocean.
  - True RTLS or False RTLS: Indicates whether the mission outcome was a successful landing on a ground pad.
  - True ASDS or False ASDS: Indicates whether the mission outcome was a successful landing on a drone ship.
- Identify and preprocess missing values:
- Count the number of launches at each launch site.
- Count the number of landing outcomes based on previous booster landings.
- Convert successful and unsuccessful landing outcomes into binary labels (1 and 0) and assign the training label values to a new column 'Class' in the dataset .
- Save the data wrangling results to a CSV file.
- GitHub URL of the completed web scraping notebook **here**

EDA with Data Visualization

- We are using scatter plots to visualize the relationships between several variables, including:
  - Flight Number and Launch Site
  - Payload Mass and Launch Site
  - Flight Number and Orbit Type
  - Payload Mass and Orbit Type
- To determine which orbits, have the highest success rates, we will employ a bar chart. Furthermore, we will plot a line chart with the Year on the x-axis and the average success rate on the y-axis to analyze the trend in average success rates over time.
- GitHub URL of a completed EDA with data visualization notebook is **here**

EDA with SQL

- Display the unique launch sites
- Displaying the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Display the unique launch sites.
- Some queries related to payload mass in the dataset include:
  - The total payload mass carried by boosters launched by NASA (CRS).
  - The average payload mass carried by booster version F9 v1.1.
  - The names of the booster versions that have carried the maximum payload mass.

EDA with SQL

- Display the total number of successful and failed mission outcomes.
- List the records that show the month names, failed drone ship landings, booster versions, and launch sites for the months in 2015 (using `substr(Date, 0, 5)` to extract the month).
- Rank the count of landing outcomes (e.g., Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.
- GitHub URL of completed EDA with SQL notebook **here** .

Build an Interactive Map with Folium

- Let's visualize launch site locations and success rates:
- Mapping Launch Sites: We can leverage the latitude and longitude coordinates of each launch site to create a map. We'll use Folium, a Python library, to add markers (folium.Marker) for each launch site on the map. This will allow us to see the geographical distribution of the launch sites.
- Visualizing Launch Success: We can further enhance the map by using colored markers to represent launch success. Green markers can indicate successful launches (class=1), while red markers can signify failed launches (class=0). Additionally, we can employ marker clusters (folium.MarkerCluster) to manage a large number of markers effectively, making the map easier to navigate. This will allow us to identify launch sites with high success rates at a glance.

Build an Interactive Map with Folium

- **Analyzing Launch Site Connectivity:** To understand the infrastructure surrounding launch sites, we can draw lines (folium.PolyLine) connecting each launch site to its closest city, railway, highway, and coastline points. This will help us answer questions about the proximity of launch sites to these key infrastructure elements:
  - Are launch sites located near major transportation routes like railways and highways?
  - Do launch sites benefit from access to coastlines?
  - Are launch sites deliberately located away from populated cities?
  - By creating this interactive map, we can gain valuable insights into the geographical context of launch sites and their success rates.
- GitHub URL of completed Launch Sites Locations Analysis with Folium notebook **here**

Build a Dashboard with Plotly Dash

- We've developed a Plotly Dash application that enables users to explore SpaceX launch data dynamically and interactively. This dashboard provides a user-friendly interface with input components like a dropdown list and a range slider, allowing users to customize the visualization.
- **Key Features:**
- **Launch Site Selection:** Users can choose a specific launch site from a dropdown menu or view data for all sites combined.
- **Success Rate Analysis:** A pie chart displays the success rate of launches for the selected launch site or overall.
- **Payload Exploration:** A range slider allows users to filter data based on payload mass.
- **Payload Success Scatter Plot:** A scatter plot shows the relationship between payload mass and launch success for the selected launch site or all sites.

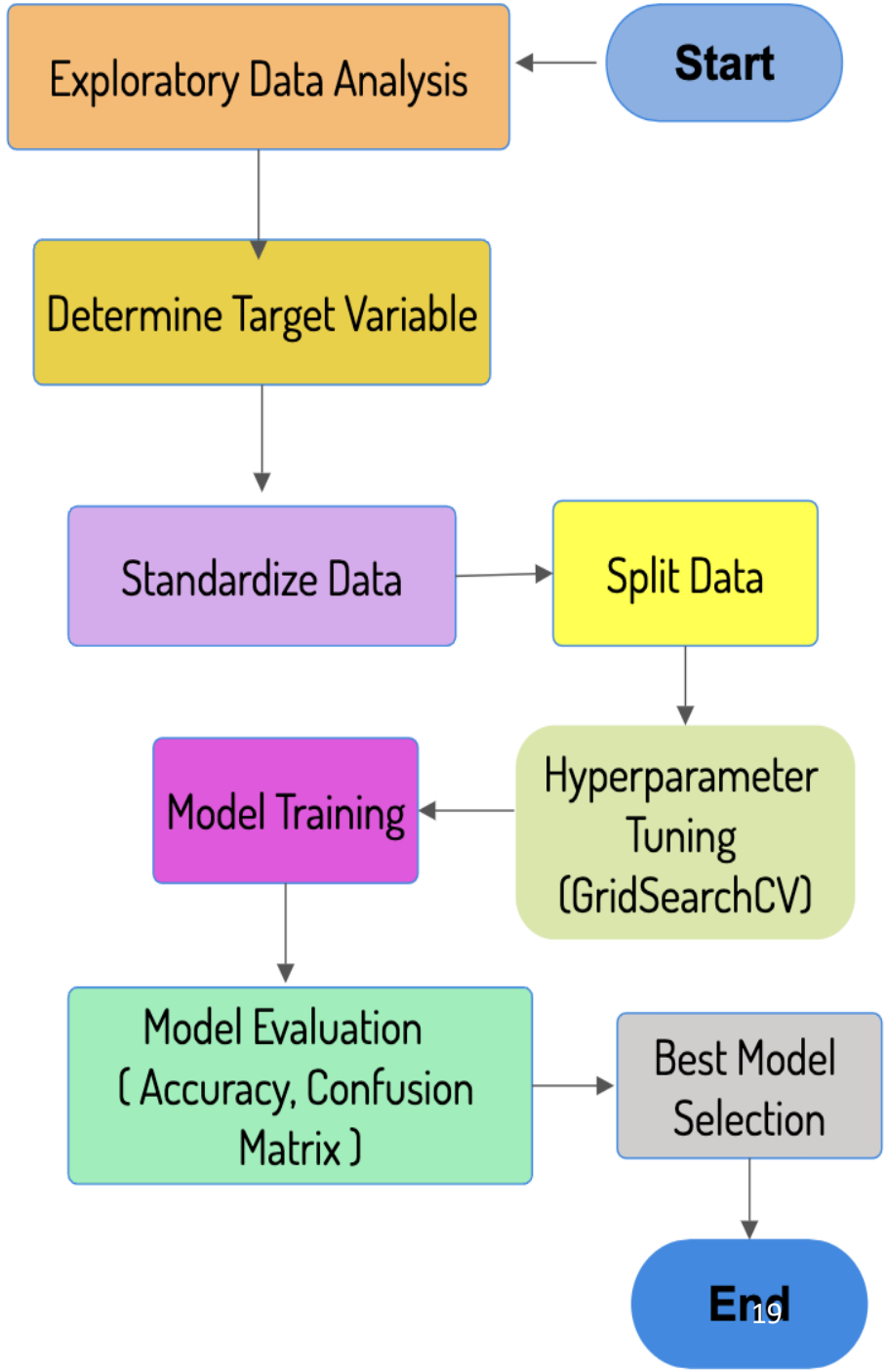
Build a Dashboard with Plotly Dash

- By using this interactive dashboard, users can uncover valuable insights, including:
- **Most Successful Launch Site:** Identifying the launch site with the highest number of successful launches.
- **Highest Launch Success Rate:** Determining which launch site boasts the best overall success rate.
- **Optimal Payload Ranges:** Pinpointing the payload mass ranges associated with the highest and lowest launch success rates.
- **F9 Booster Performance:** Assessing the launch success rates of different F9 Booster versions.

- GitHub URL of your completed Plotly Dash **here**

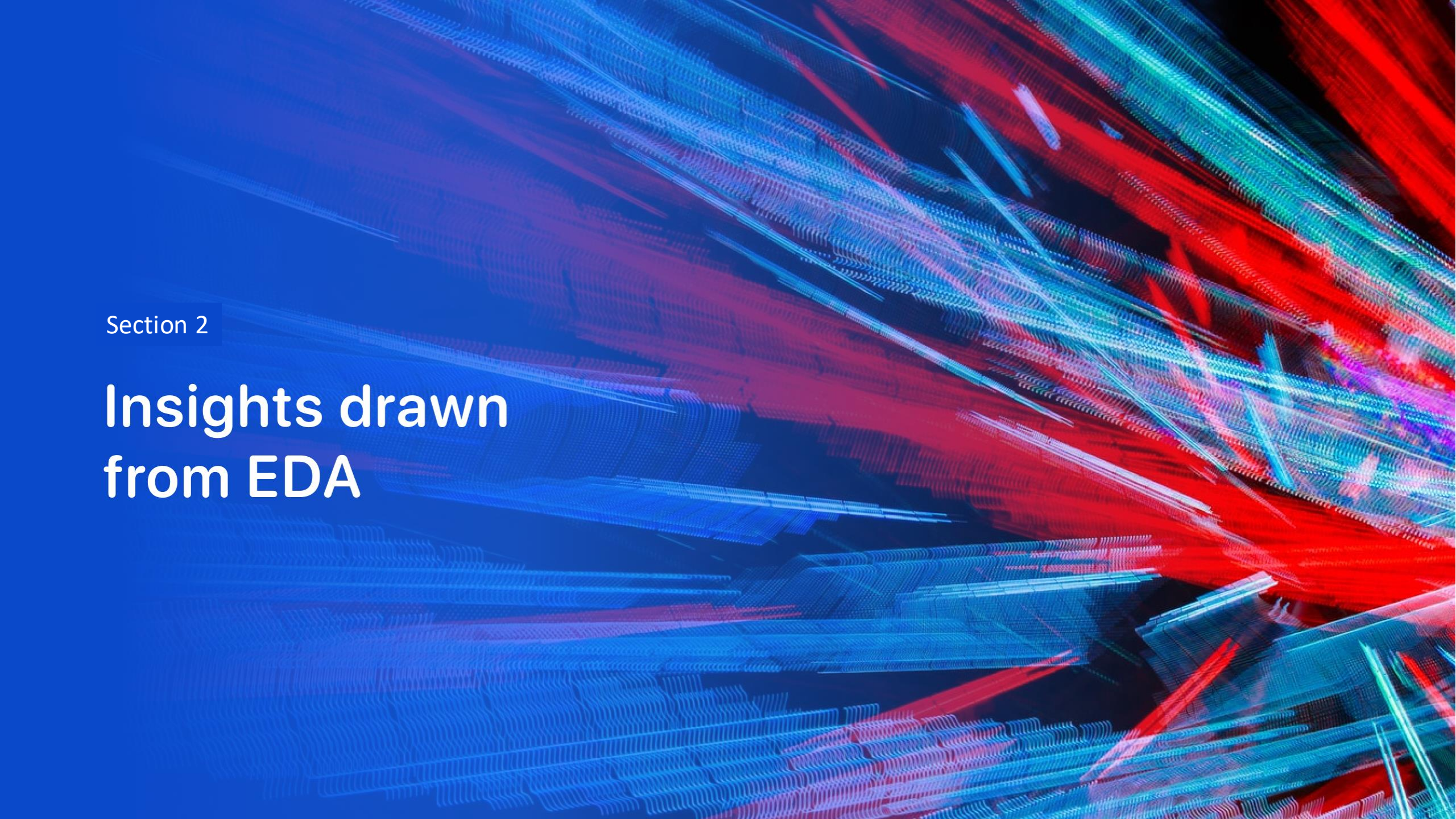
Predictive Analysis (Classification)

- In this task, we conducted exploratory data analysis on the dataset, focusing on the 'Class' column to determine the target variable.
- After standardizing the data to ensure consistency, we split the dataset into training and testing sets.
- To optimize model performance, we employed hyperparameter tuning using GridSearchCV.
- We then trained various machine learning models and evaluated their accuracy on the test data using the score function.
- Additionally, we visualized the confusion matrices of each model to understand their classification performance.
- By comparing the performance metrics, we identified the most effective model for our task.
- GitHub URL of completed predictive analysis **here**

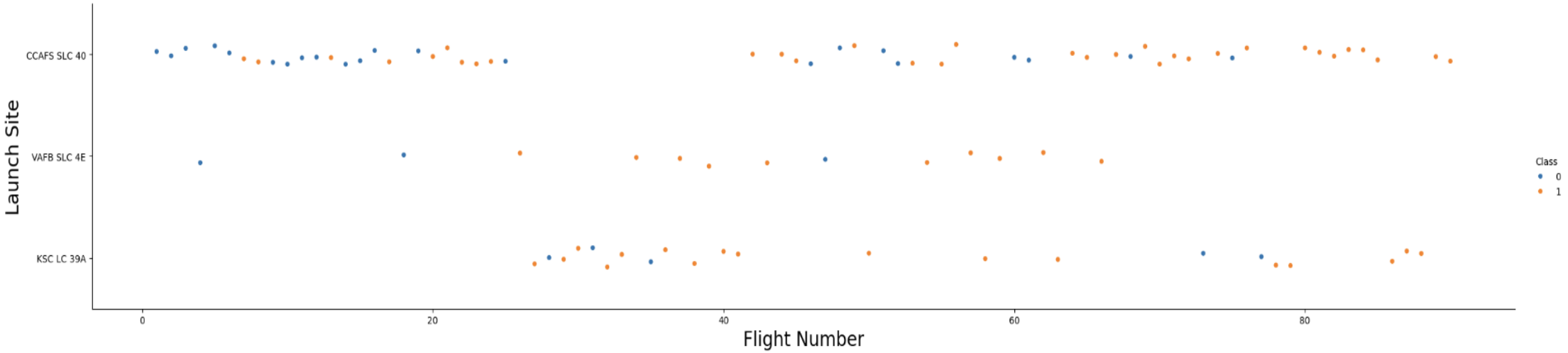


Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

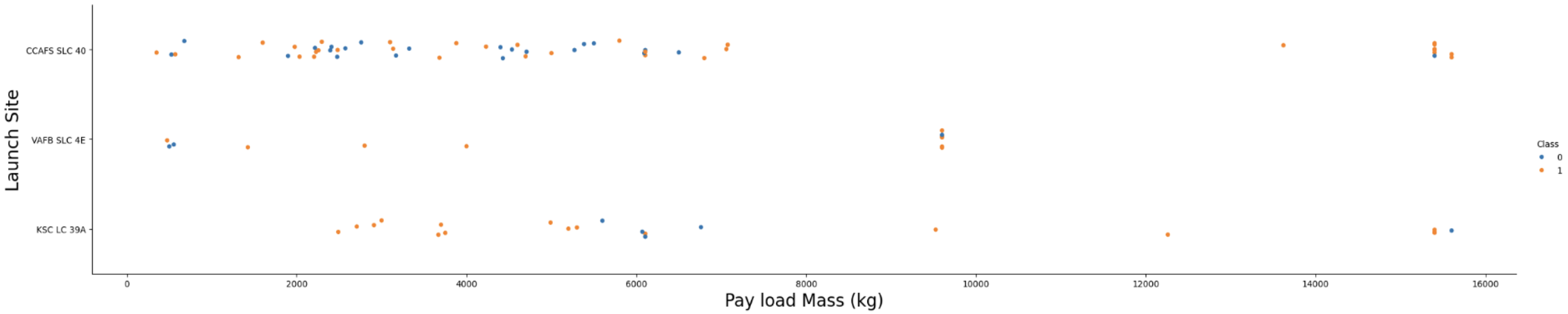
Flight Number vs. Launch Site



Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.

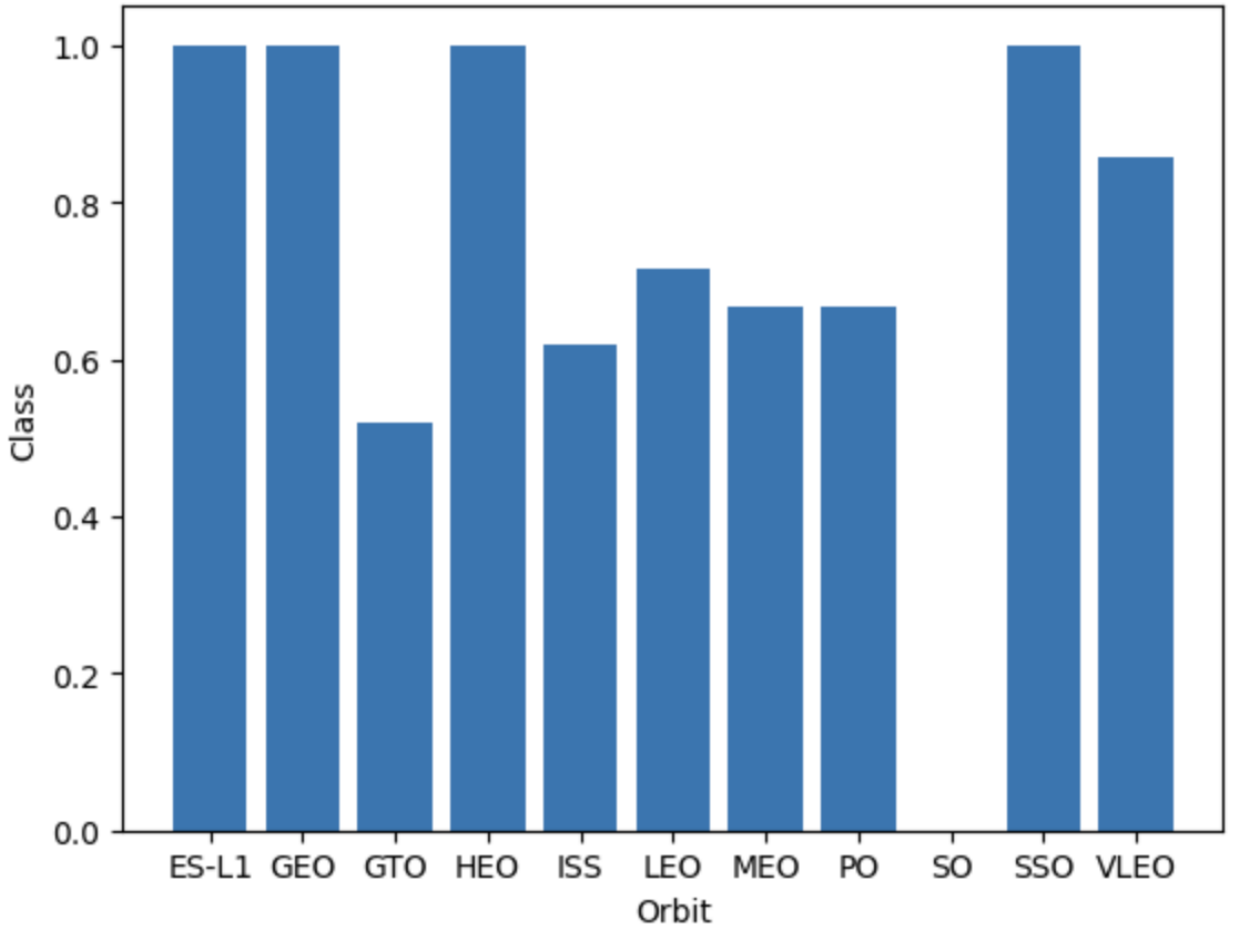
I see that as the flight number increases, the first stage is more likely to land successfully at each launch site .

Payload vs. Launch Site



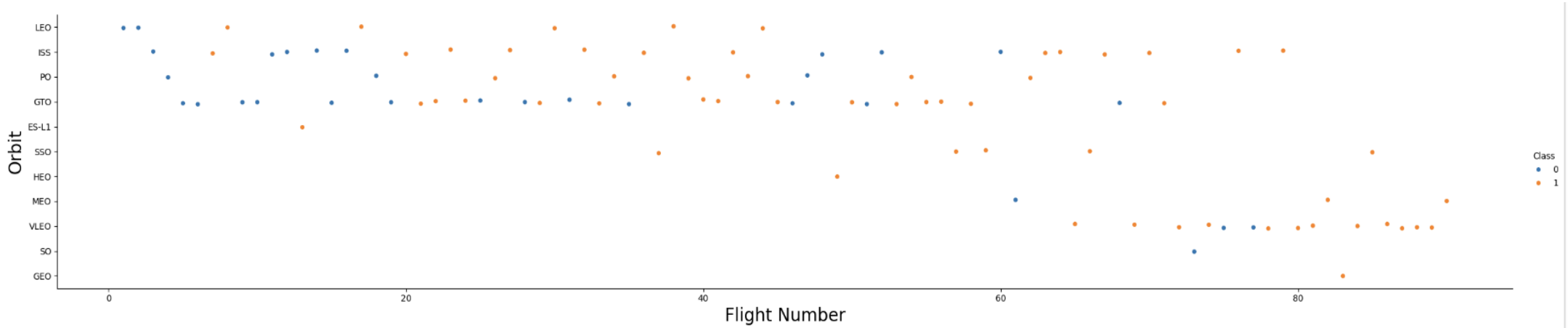
Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



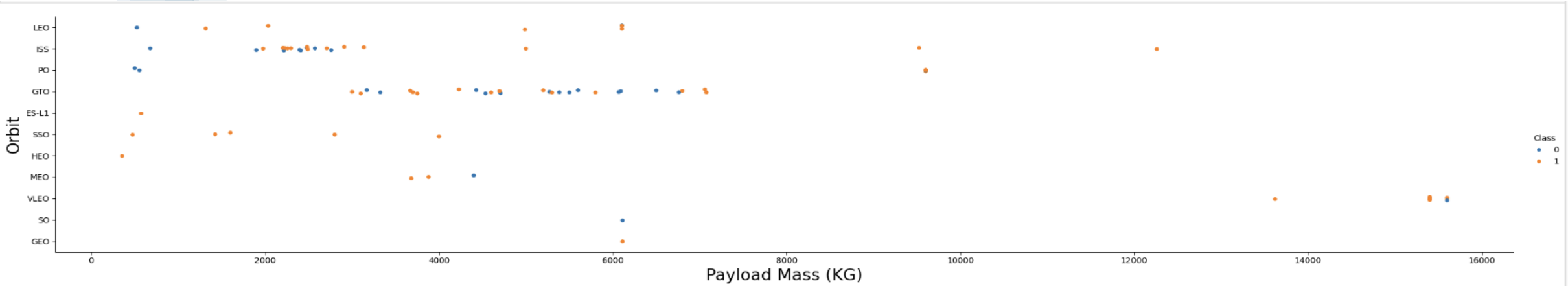
Analyze the plotted bar chart to identify which orbits have the highest success rates.

I see that the SSO, HEO, GEO and ES-L1 orbits have the highest success rates



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Flight Number vs. Orbit Type

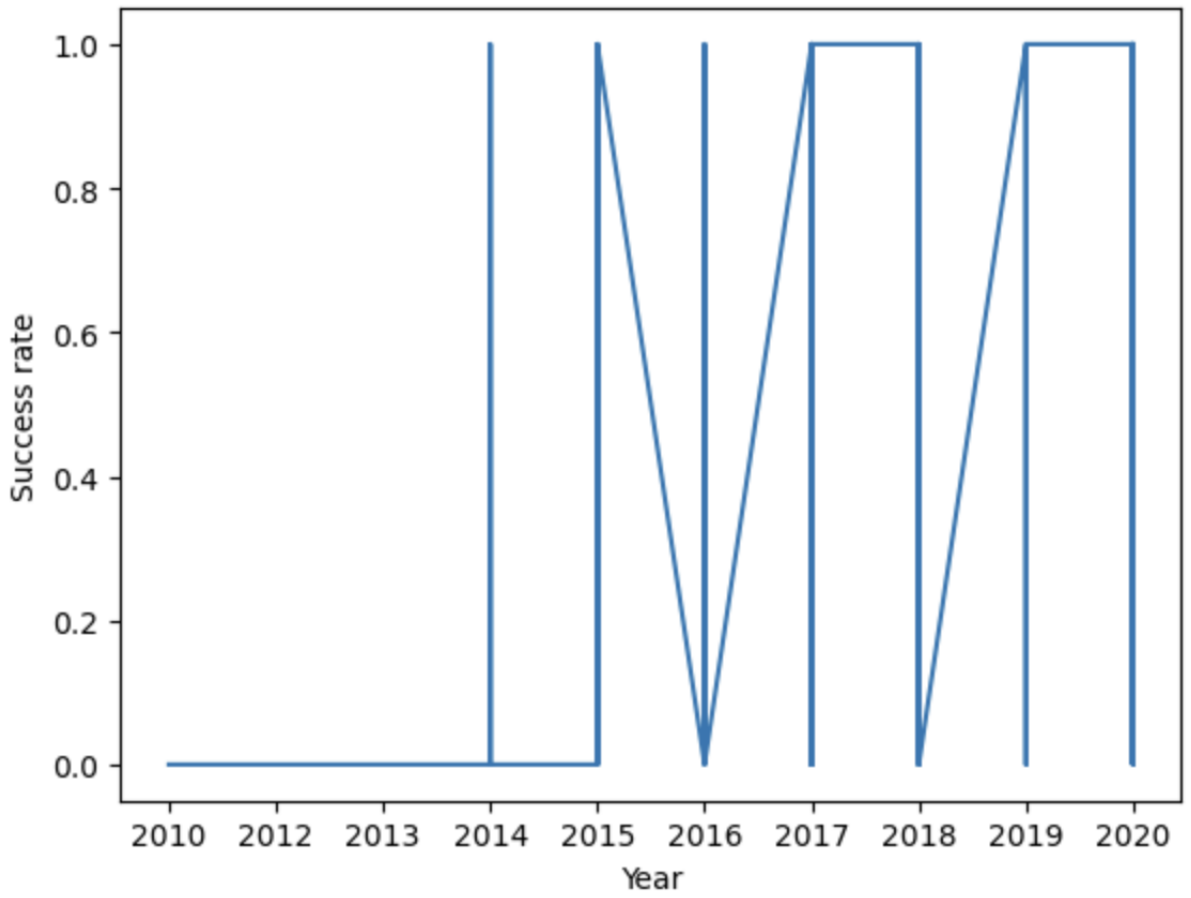


With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Payload vs. Orbit Type

Launch Success Yearly Trend



you can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

Task 1

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

10]:
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

10]:

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

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string '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_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Total Payload Mass

Task 3

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

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") as "Total Mass by NASA (CRS)" FROM SPACEXTABLE WHERE "Customer" = "NASA (CRS
```

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

Total Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

Task 4

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

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") as "AVERAGE_Mass by F9 v1.1" FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1";
```

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

AVERAGE_Mass by F9 v1.1
2928.4

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
]: %sql SELECT MIN("Date") AS "First day of a successful ground pad landing outcome" FROM SPACEXTABLE WHERE "Landing_Outco
```

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

]:

First day of a successful ground pad landing outcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND "PAYLOAD_MASS__KG_"
```

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
6]: %sql SELECT COUNT(CASE WHEN "Mission_Outcome" LIKE 'Success%' THEN 1 END) AS "Number of Successful Mission Outcomes",C
```

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

6]:

Number of Successful Mission Outcomes	Number of Failed Mission Outcomes
100	1

Boosters Carried Maximum Payload

Task 8

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

```
7]: %sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACE
```

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

7]:

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

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

3]:
```
%sql SELECT "Date", "Booster_Version", "Launch_Site", CASE substr("Date", 6, 2) WHEN '01' THEN 'January' WHEN '02' THE
```

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

3]:

Date	Booster_Version	Launch_Site	Month Name
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	January
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	April

2015 Launch Records

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

Task 10

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.

4]:
```
%sql SELECT "Booster_Version", "Landing_Outcome", "Date" FROM SPACEXTABLE WHERE ("Landing_Outcome" = "Failure (drone s
```

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

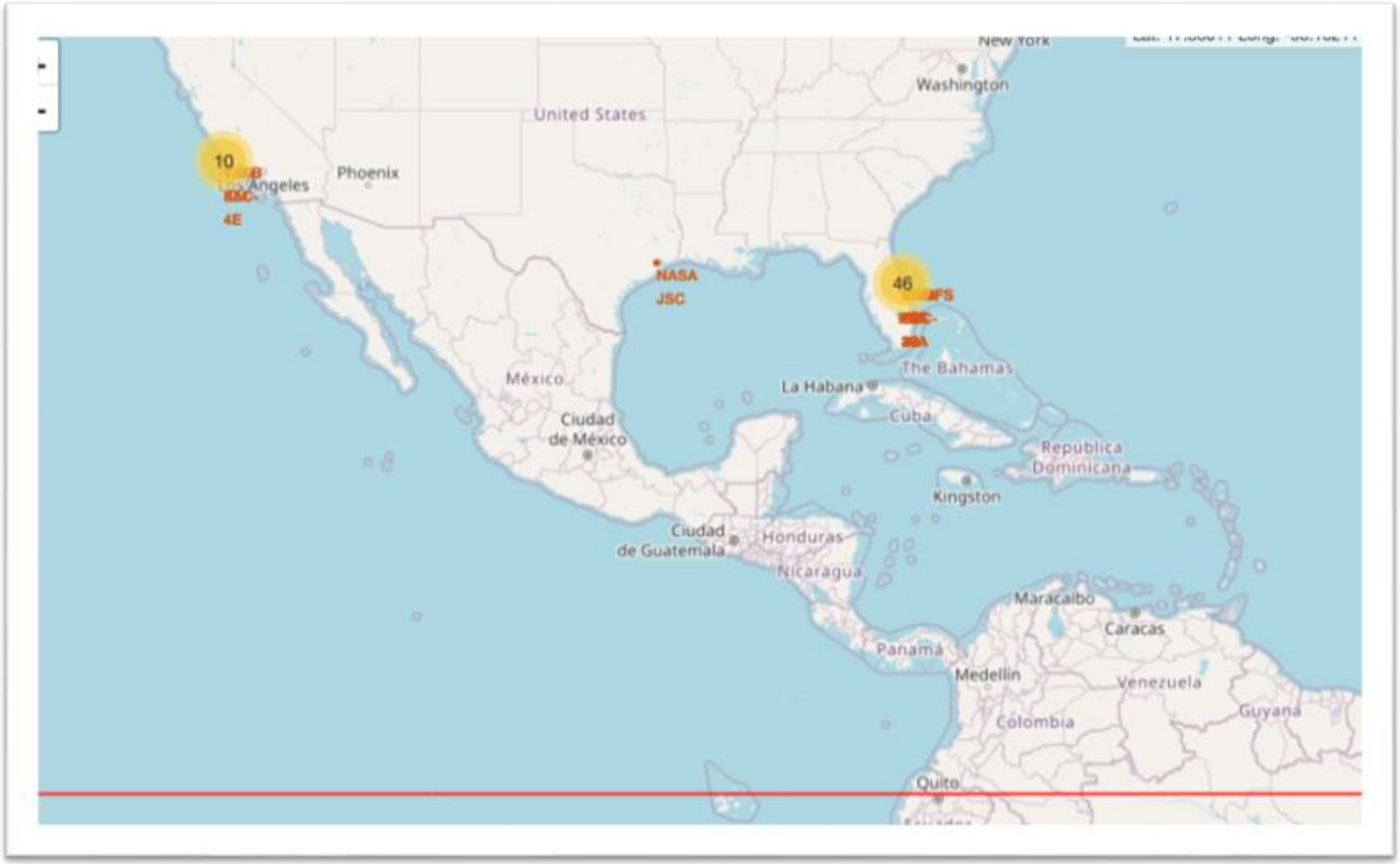
4]:

Booster_Version	Landing_Outcome	Date
F9 FT B1031.1	Success (ground pad)	2017-02-19
F9 FT B1025.1	Success (ground pad)	2016-07-18
F9 FT B1024	Failure (drone ship)	2016-06-15
F9 FT B1020	Failure (drone ship)	2016-03-04
F9 v1.1 B1017	Failure (drone ship)	2016-01-17
F9 FT B1019	Success (ground pad)	2015-12-22
F9 v1.1 B1015	Failure (drone ship)	2015-04-14
F9 v1.1 B1012	Failure (drone ship)	2015-01-10

Section 3

Launch Sites Proximities Analysis

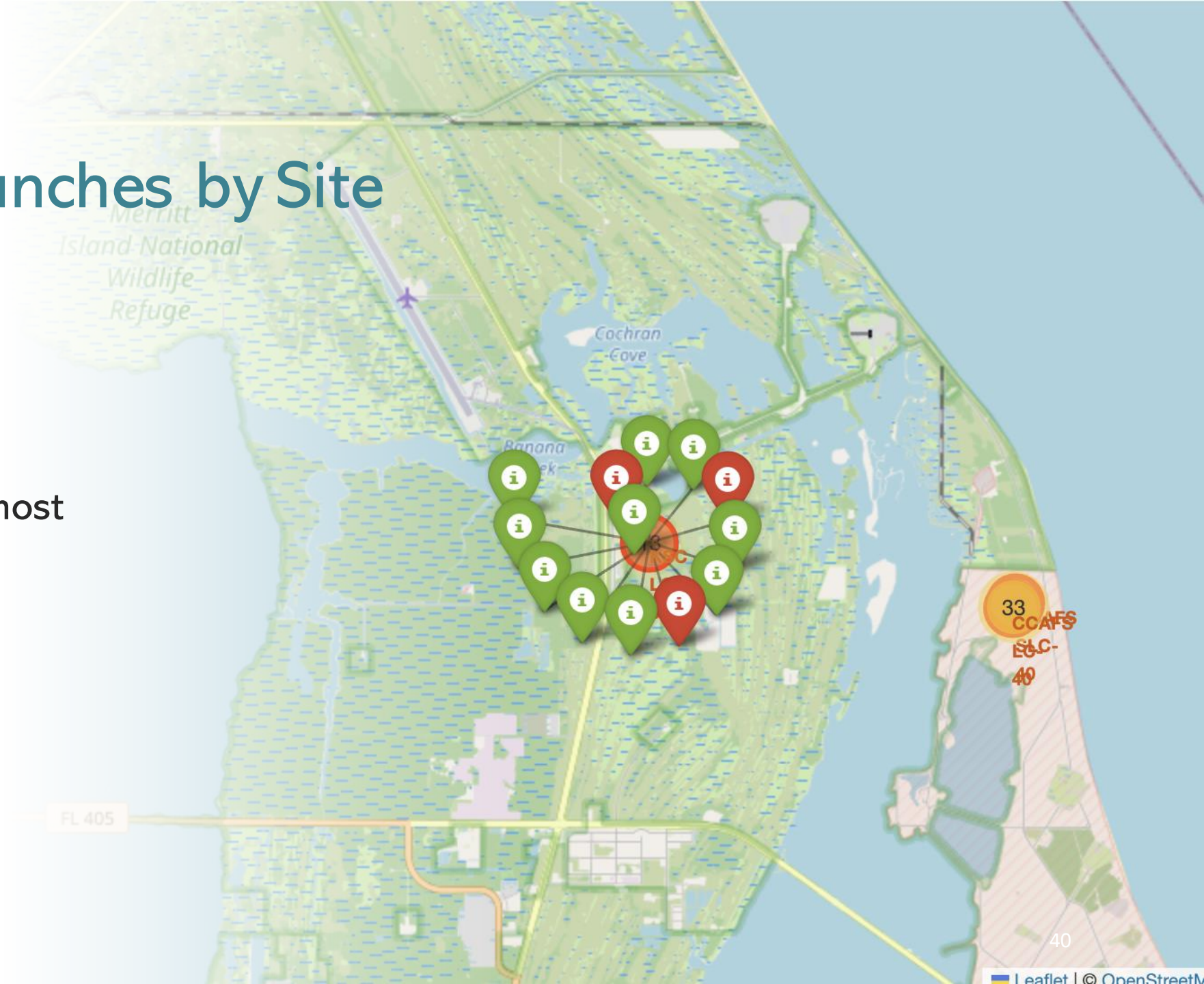
While launch sites are often located near the coast, they are generally far from the equator line.

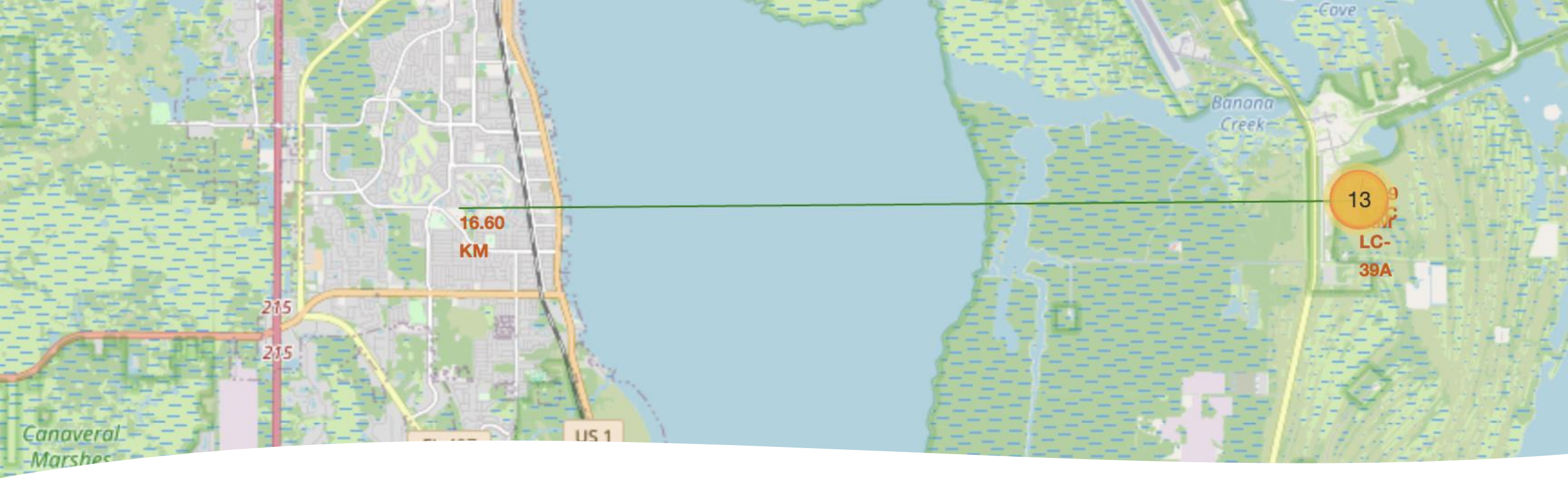


Global Launch Sites Mapped

Outcomes of Launches by Site

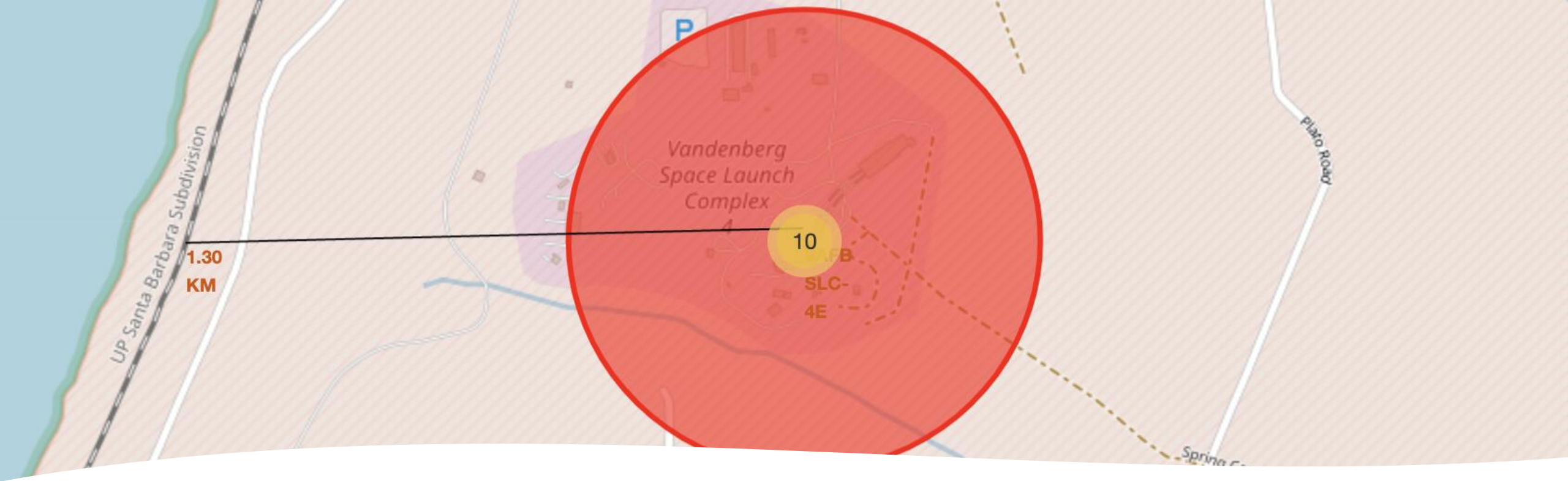
- KSC LC-39A has the most successful launches





Visualizing the Distance Between Launch Site and City

- Using the (PolyLine) function, we plot a line connecting a launch site to its nearest city and calculate the distance between them

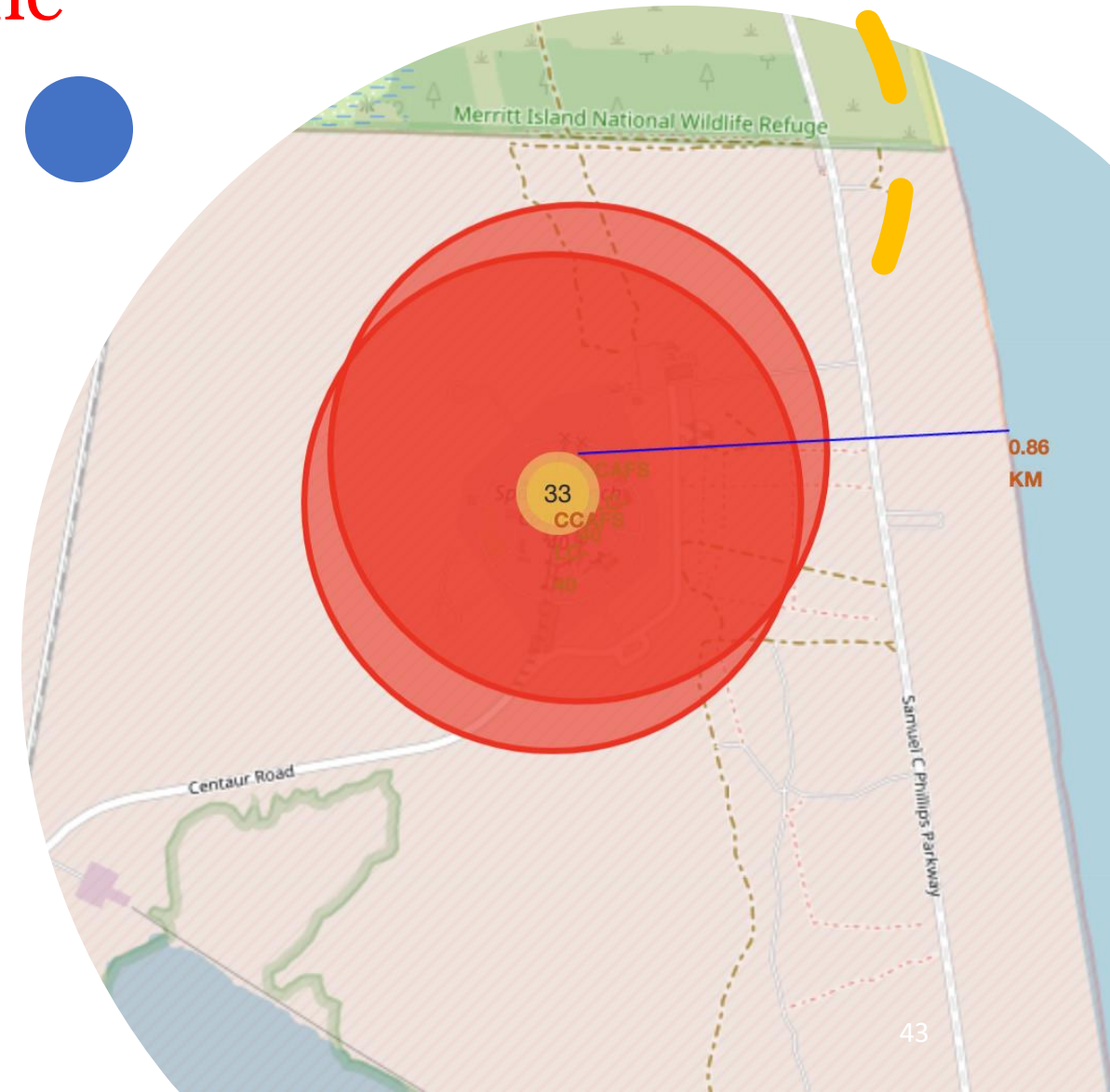


Visualizing the Distance Between Launch Site and railway

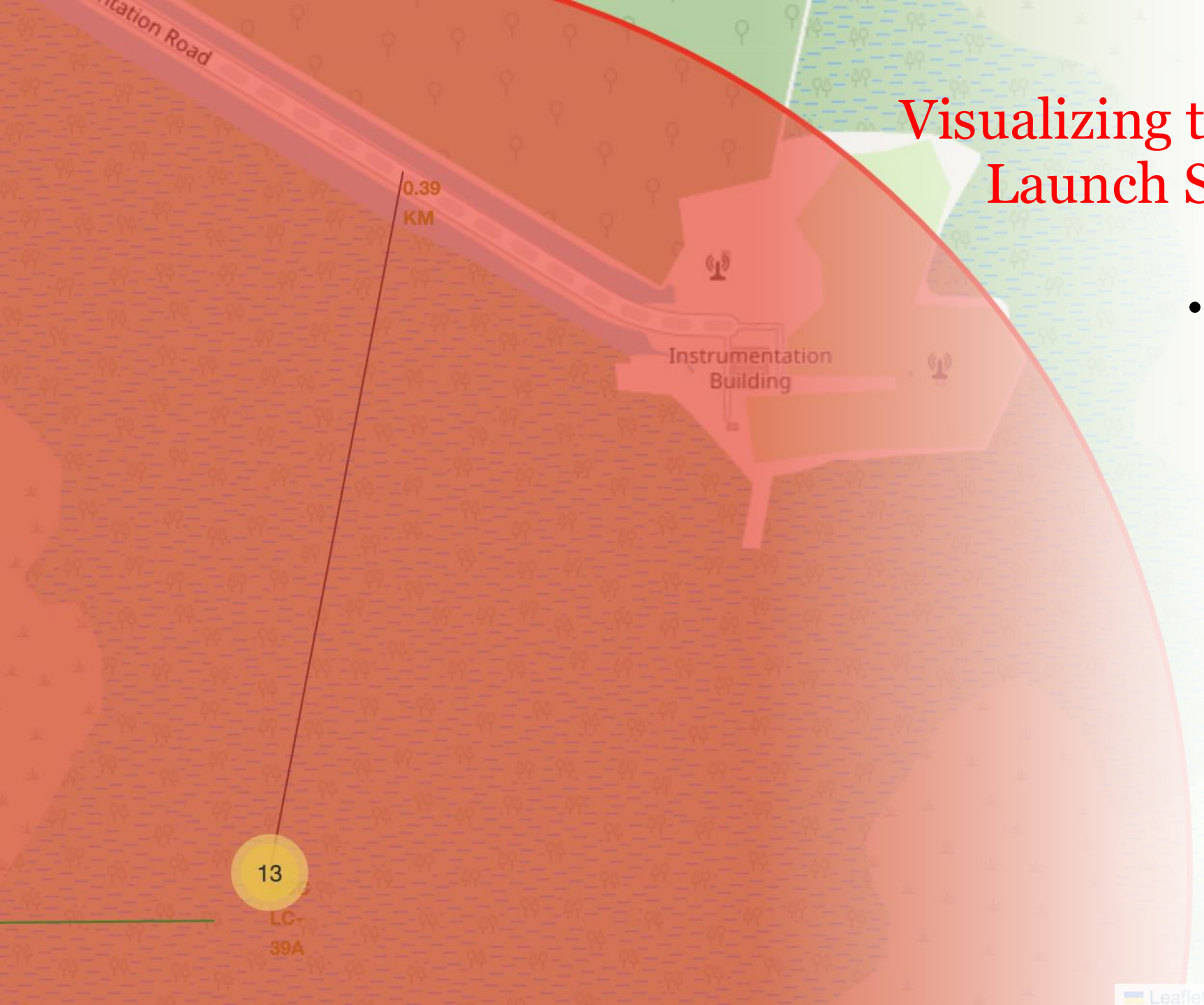
- Using the (PolyLine) function, we plot a line connecting a launch site to its nearest railway and calculate the distance between them

Visualizing the Distance Between Launch Site and coastline

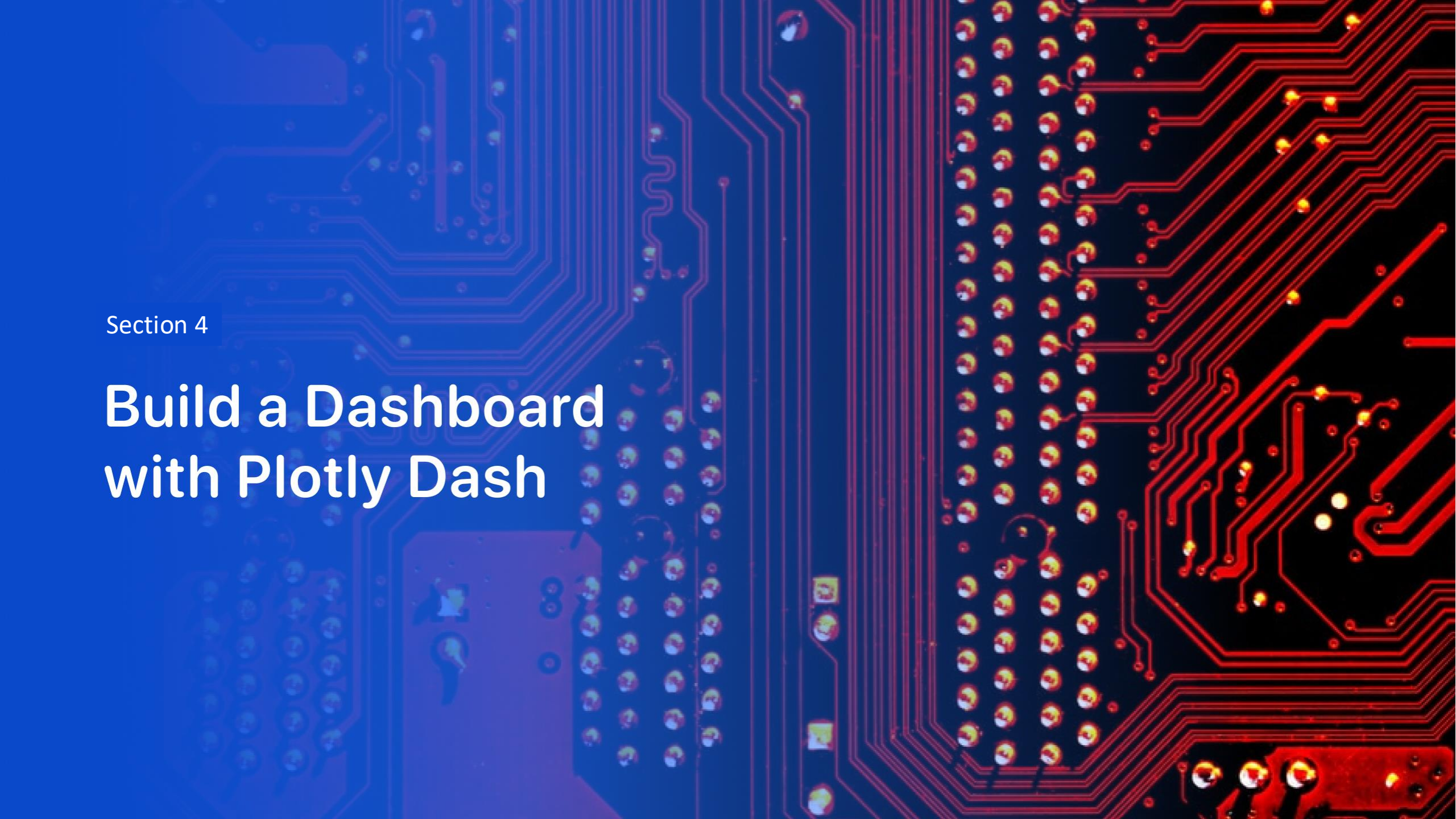
- Using the (PolyLine) function, we plot a line connecting a launch site to its nearest coastline and calculate the distance between them

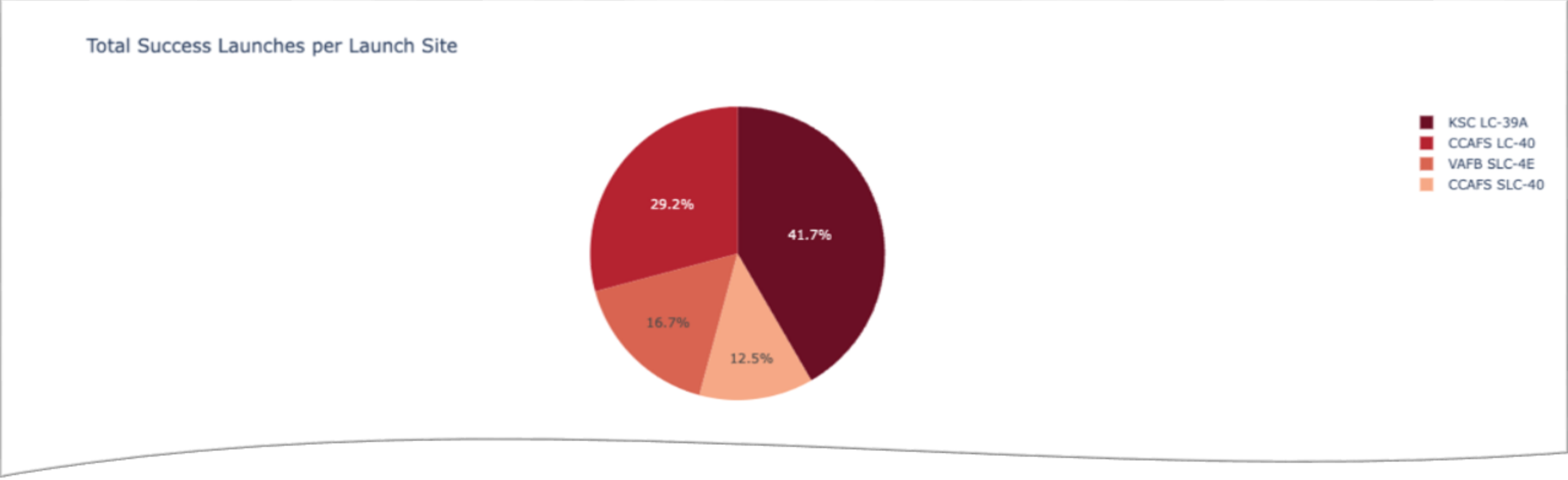


Visualizing the Distance Between Launch Site and Highway



- Using the (PolyLine) function, we plot a line connecting a launch site to its nearest highway and calculate the distance between them


Section 4
Build a Dashboard with Plotly Dash



Total Successful Launches per Launch Site

- The launch site (KSC LC-39A) has the highest number of successful launches.

KSC LC-39A

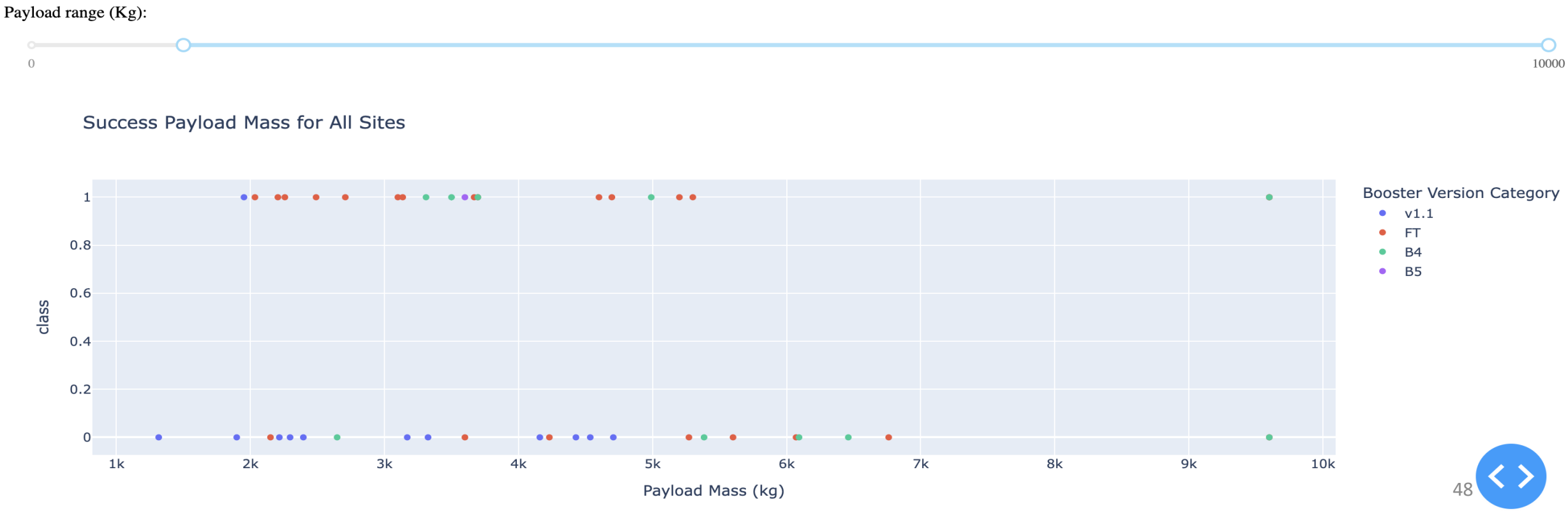
Launch Success Rate for KSC LC-39A

The Most Successful Launch Site

- This site has the highest success rate, with 10 successful launches and 3 failed launches.

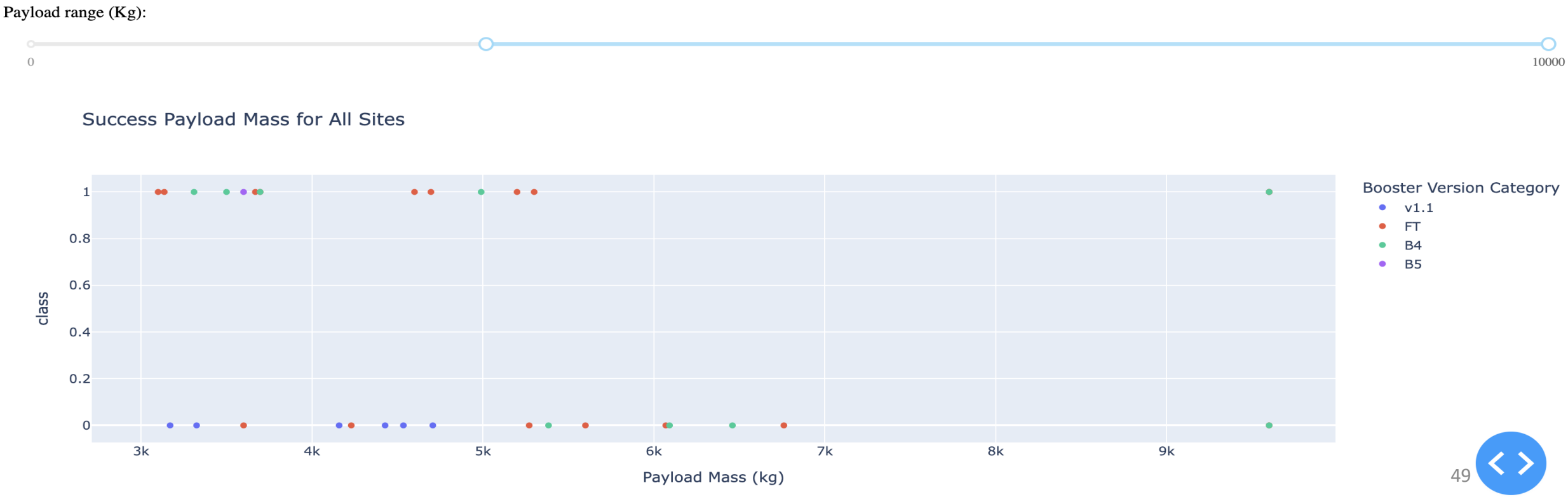
The Influence of Payload Mass on Booster Version Categories

We observe that the success rate of the FT booster version category increases as payload mass rises from 2,000 to 7,000 kilograms, while the success rate of the v1.1 category decreases.

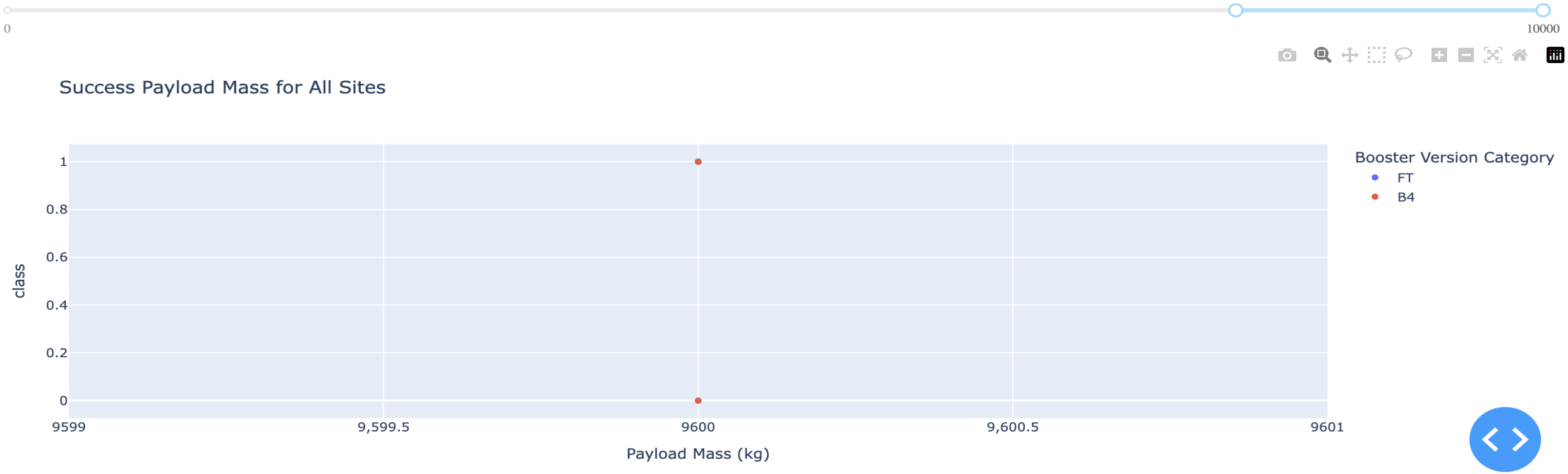


The Influence of Payload Mass on Booster Version Categories

We observe that the success rates of the FT and B4 booster version categories increase as payload mass rises from 3,000 to 5,500 kilograms. However, they both decrease when payload mass exceeds 5,500 kilograms. In contrast, the success rate of the v1.1 booster version consistently decreases regardless of payload mass.



Payload range (Kg):



The Influence of Payload Mass on Booster Version Categories

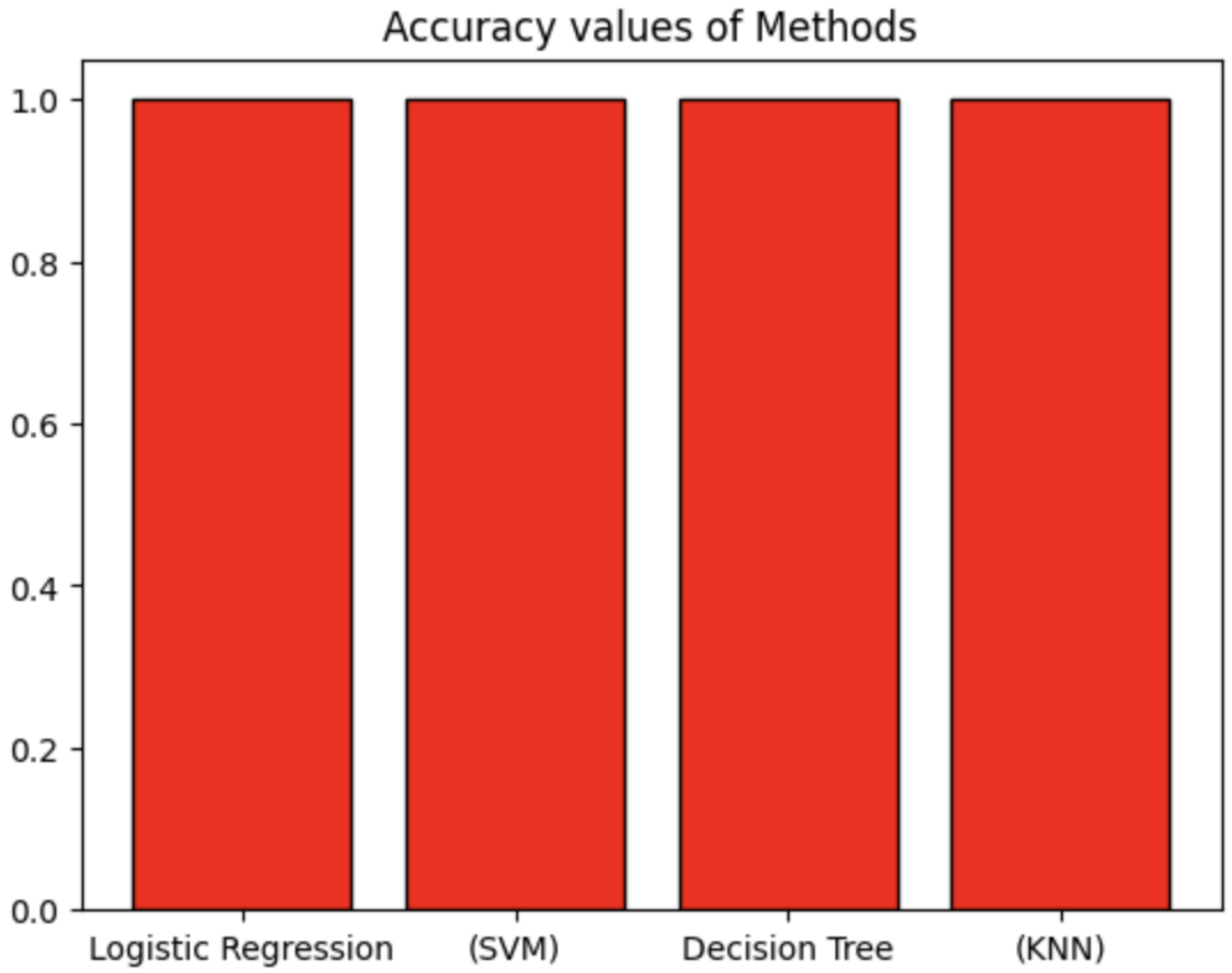
- We observe that the success rates of the FT booster version category increase with payload mass up to 9600 kilograms, while the success rates of the B4 booster version category decrease at the same payload mass

Section 5

Predictive Analysis (Classification)

Classification Accuracy

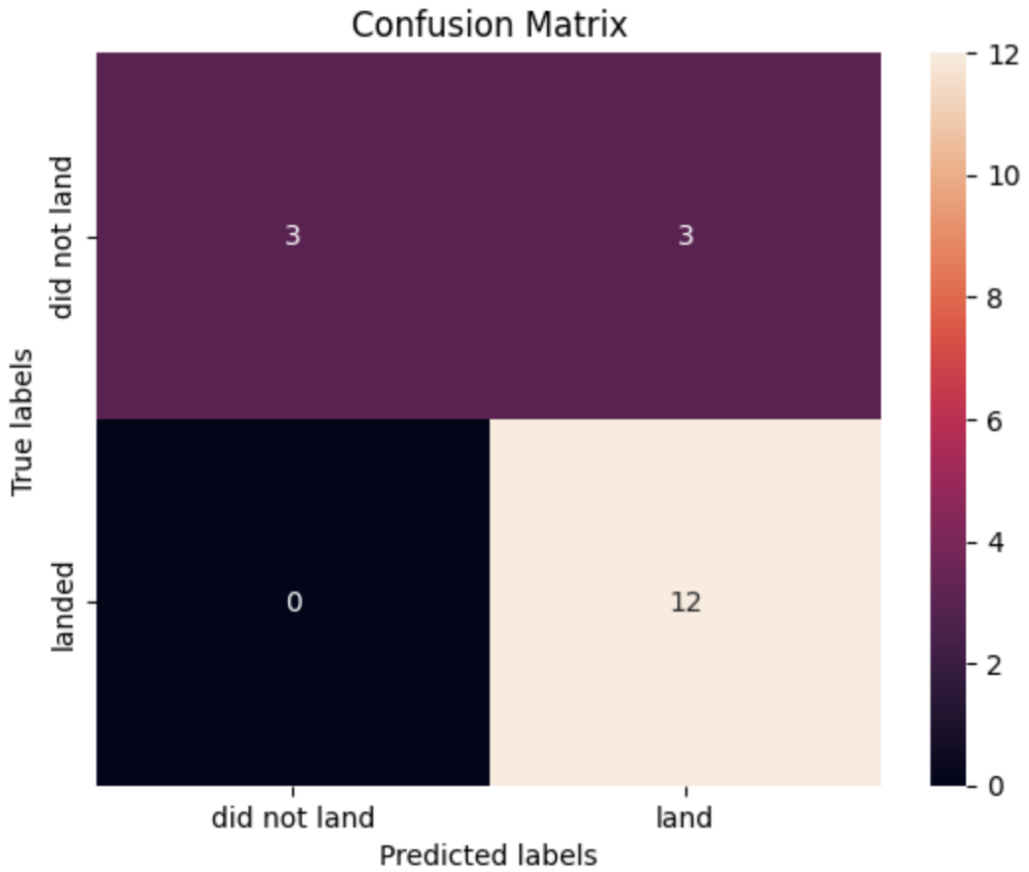
- According to the bar chart, all the models have identical accuracy scores of 1.0 .



Among the four methods (logistic regression, SVM, Decision Tree, and KNN), logistic regression, SVM, and KNN achieve the highest performance,

as demonstrated by their identical best confusion matrix, presented here.

Confusion Matrix



Conclusions

- By leveraging various machine learning techniques and training on public data, we predicted the likelihood of successful first-stage landings. Additionally, our analysis indicated that SpaceX does not reuse the first stage of their rockets.

Appendix

- GitHub URL of completed Machine Learning Prediction here
- Github URL of completed SQL Notebook here

Thank you!