IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Syed Bokhari
28th September 2022

Table of Contents

Outline

- Executive Summary
- Introduction
- Methodology
- Insights Drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusion
- Appendix (with Github URL)

Executive Summary

Space has been referred to as the “final frontier” by Captain James T. Kirk from the science fiction franchise Star Trek. The early decades of the 21st century appear to confirm this sentiment as companies compete to commercialize space travel for the public. This goal has spurred many innovations. One such innovation is by the company SpaceX which has managed to find ways to reduce the cost of space flights by finding ways to reuse the first stage of the rocket launch. The first stage of a rocket launch is very expensive and the ability to reuse it brings down the cost of a space launch significantly. However, the first stage does not always land successfully either due to technical issues or mission parameters.

The project aims to find a way to classify landing outcomes for the first stage of a space launch beforehand. If this can be done successfully, then the cost of a launch can be calculated before the launch is attempted. For this purpose, the project makes use of publicly available data on launches of SpaceX from the web and from the REST API SpaceX API. This data is collected, cleaned and processed for further analysis.

Executive Summary

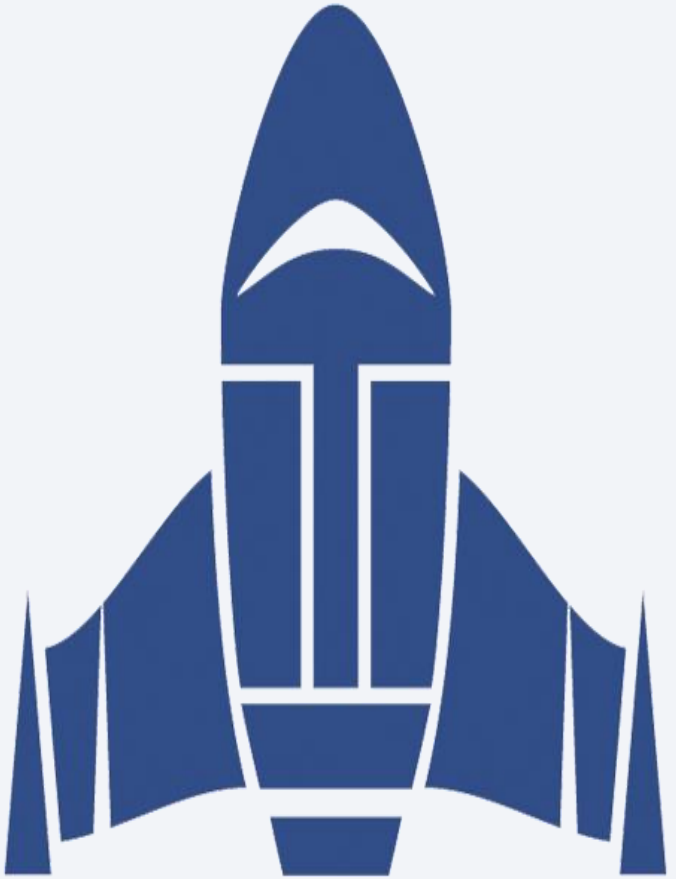
An Exploratory Data Analysis (EDA) of the dataset was conducted. The EDA identified several important features in the dataset that may impact the success of obtaining a landing outcome. These features were Payload Mass carried by the launch, Orbit type targeted as goal of the launch, Booster Version used for the launch and Launch Site. Feature Engineering was attempted on the dataset to transform these and other features in a suitable format for the next stage of the project: Machine Learning.

Machine Learning, more specifically classification techniques were used to determine the likelihood that the landing outcome for a launch will be successful. 4 different classification techniques were used to classify landing outcomes for launches. Each of the techniques was used to evaluate the data separately. The results were then used to determine the accuracy of classifications and a confusion matrix was plotted for each technique. The results were used to draw important conclusions.

Introduction

The age of space exploration is upon us! As technology advances, the dream of humankind exploring space is increasingly getting closer to becoming a reality. One company that is pushing the boundaries of space exploration is SpaceX. SpaceX has managed to make a name for itself in space launches. It has managed to reduce the total cost of a space launch by finding ways to reuse the first stage of the space launch. However, the first stage does not always land successfully, either due to some error or because mission parameters might require it.

The goal of this project is to determine, based on publicly available data about SpaceX's space launches, whether for a given launch, the first stage will land successfully or not. If this can be predicted beforehand, the total cost of the launch can be determined early on. The project aims to single out various features specific to launches that impact the success or failure of a landing outcome. It then uses Machine Learning classification techniques on the identified features to classify whether a launch will be successful or not in landing the first stage. The results can be further refined in the future as more launches occur.



Section 1

Methodology

Methodology

- Data collection methodology:
  - Described how data was collected
- Performed data wrangling
  - Described how data was processed
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- For the project, data was collected from two sources.
- The first source was Past Launches table from the wiki page titled List of Falcon 9 and Falcon Heavy launches. The link for the page is: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- The second source was an API call to the SpaceX API, a REST API which provides access to data about SpaceX launches. The link for the API is: https://api.spacexdata.com/v4/launches/past
- The data from both sources was collected into two separate datasets in CSV format.
- The collection process has been outlined for both data sources using flowcharts in the following sections.

Data Collection – SpaceX API

- For the SpaceX API, the requests library was used for obtaining data from the API and pandas and numpy libraries were used for processing.
- A number of auxiliary functions for collecting the data were defined for obtaining data specific to certain aspects of the launches such as rockets, launch pads, payloads and cores.
- The requests library was used to obtain the data from the API. The data was in JSON format.
- The JSON data was normalized into a pandas dataframe for further processing.
- A number of key value pairs were defined for storing the processed data in a dictionary format in later stages.

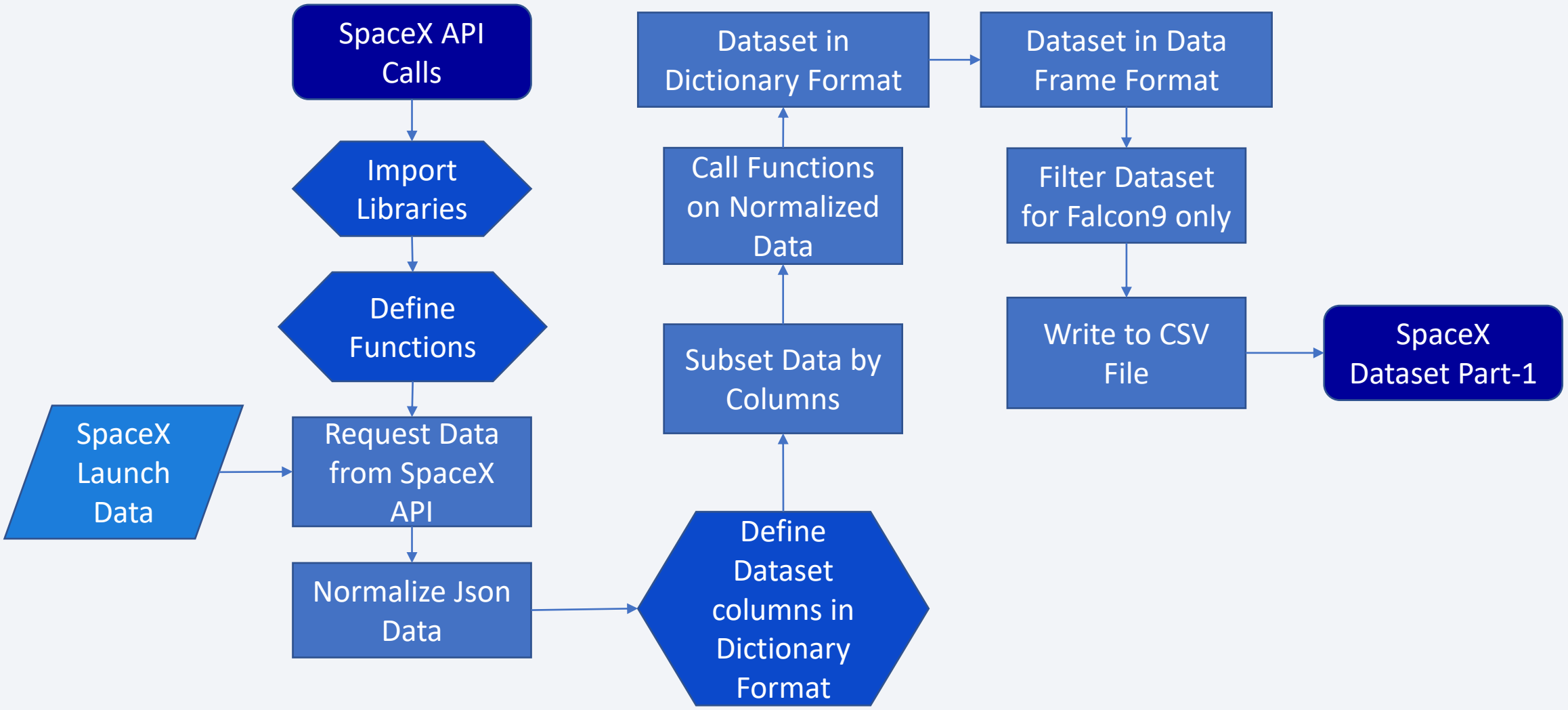
Data Collection – SpaceX API

- The data was subsetted so that only columns of interest for the project: rocket, payloads, launch pads, cores, flight numbers and date in utc format would remain.
- Data for dates after 13th November 2020 was discarded. Data for multiple cores and payloads was also discarded.
- Cores and payloads were normalized to contain 1 value only afterwards. Previously, they were in list format as some rows had multiple values. The data was prepared for further processing now.
- The auxiliary functions defined previously were called on the data. The values returned by the function were stored in the dictionary created previously.
- The dataset in dictionary format was converted into a pandas dataframe.

Data Collection – SpaceX API

- The dataset in pandas dataframe was filtered so that only values in the Booster Version column corresponding to ‘Falcon 9’ were kept. The rest were discarded.
- The filtered dataset was converted from a pandas dataframe into a CSV file.
- The CSV file was named SpaceX Dataset Part-1.

Data Collection – SpaceX API Flowchart



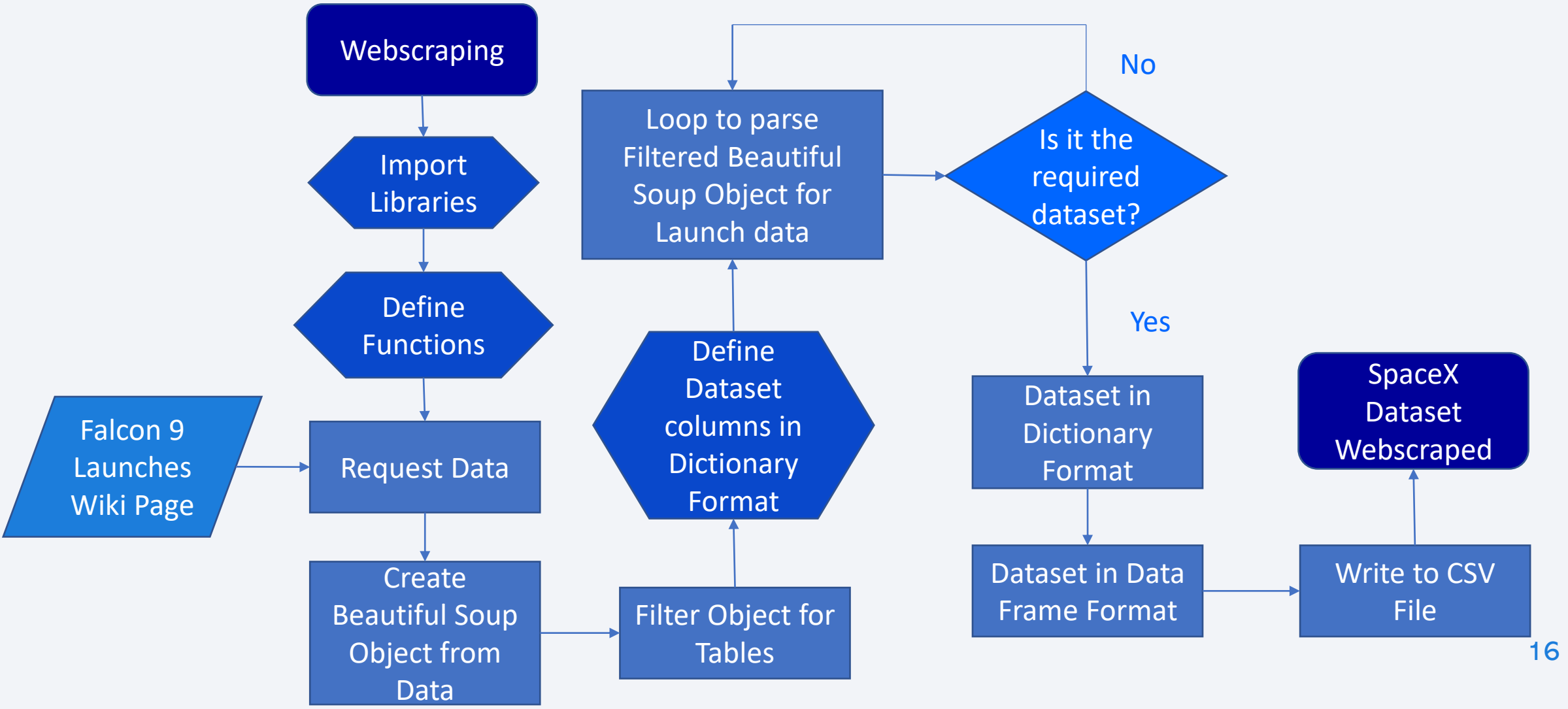
Data Collection - Scraping

- For Webscraping, the requests library was used for obtaining data from the API and pandas and bs4 libraries were used for processing. BeautifulSoup from bs4 played an integral role.
- A number of auxiliary functions for extracting data from the tables to be scrapped were defined beforehand.
- The requests library was used to request the data from the wiki page. The data was stored in a response variable.
- The response variable was parsed with html5lib to create a BeautifulSoup object that stored data in html format.
- The soup object was filtered for tables and the tables in html format were stored separately. The rest of the data was discarded.

Data Collection - Scraping

- A number of key value pairs were defined for storing the processed data in a dictionary format in later stages.
- A for loop was created to parse through the tables row by row and extract the required data from each row and save it into different key-value pairs in the dictionary defined previously. The auxiliary functions also defined previously were used for that purpose. The loop repeated this for each row.
- The data in dictionary format was obtained. It was converted into a pandas dataframe.
- The pandas dataframe was converted into the dataset in CSV format.
- The dataset was named SpaceX dataset webscraped.

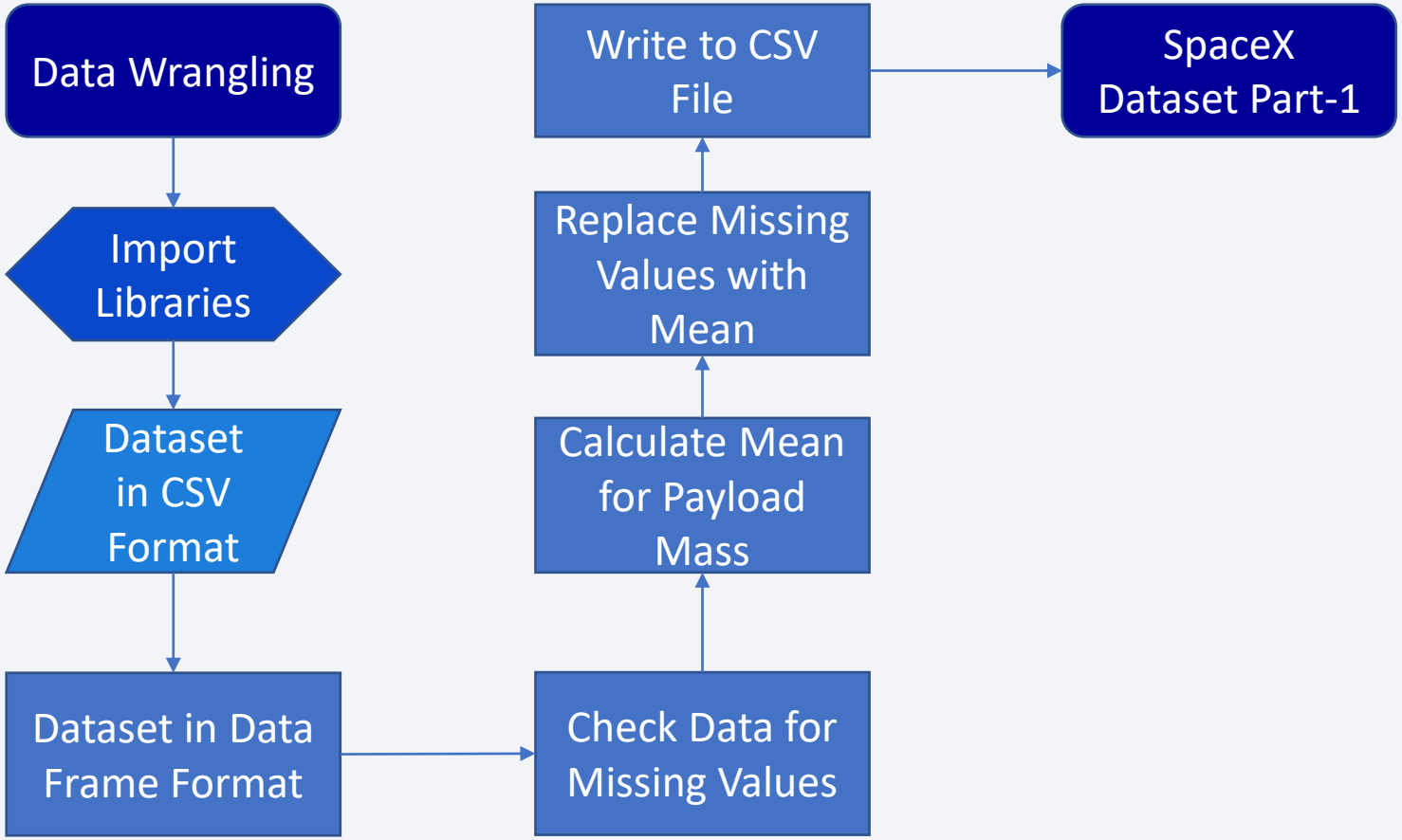
Data Collection – Scraping Flowchart



Data Wrangling – SpaceX API

- The data from the SpaceX API needed to be wrangled.
- The data was checked for missing values. The missing values were in the Payload Mass column and Landing Pad column.
- It was decided to leave the missing values in Landing Pad column. They would represent when landing pads were not used.
- It was decided to interpolate a value for Payload Mass and replace the missing values in the column with it.
- The mean of Payload Mass column was chosen as the interpolated value and was calculated.
- The missing values in the column were replaced with the mean and the data was written into CSV format.

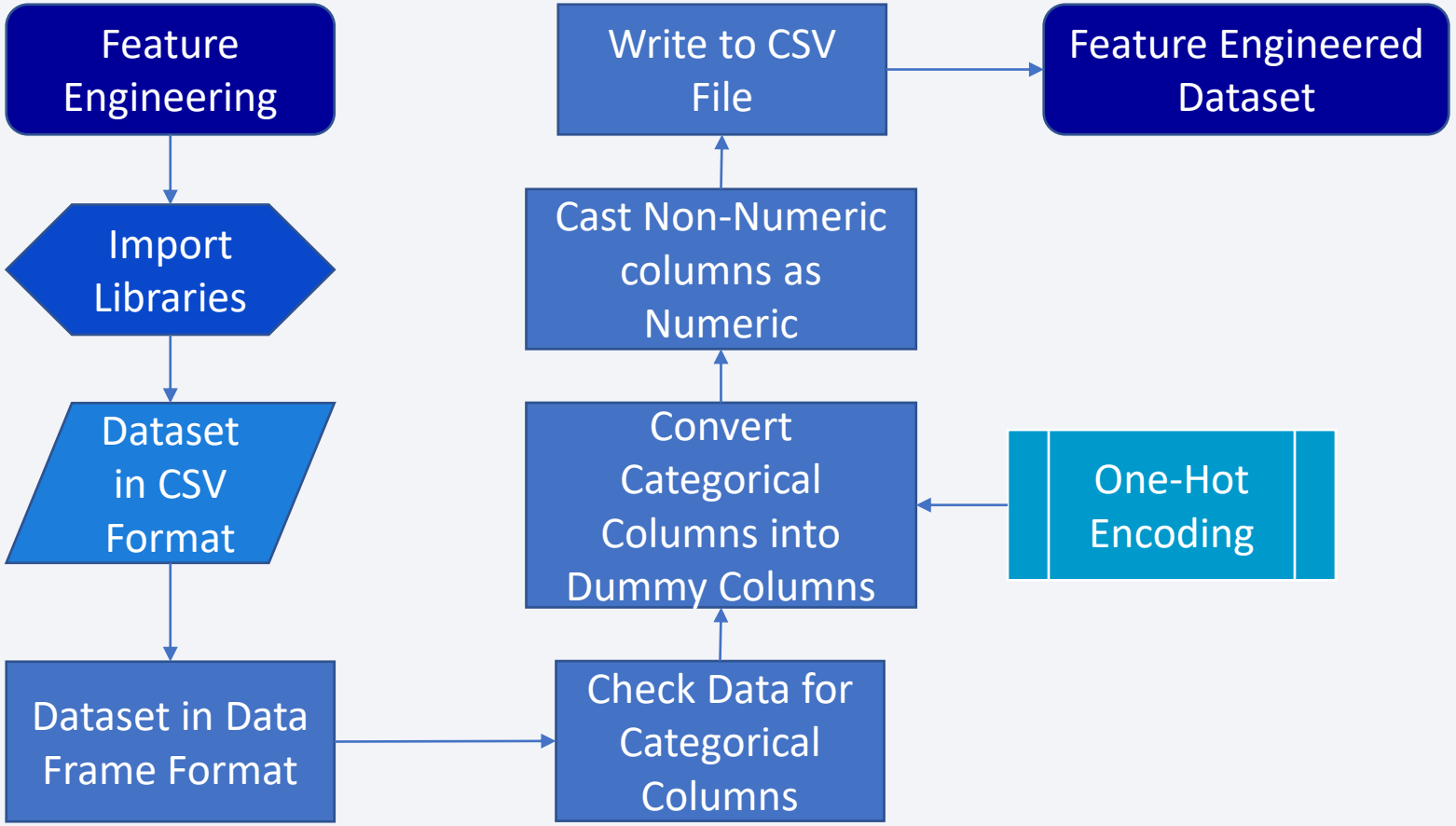
Data Wrangling – SpaceX API Flowchart



Data Wrangling – Feature Engineering

- The data from the dataset needed to be made more suitable for predictive analysis.
- This was necessary because features from the dataset needed to be clear for subsequent stages of predictive analysis.
- Consequently, feature engineering was attempted on the dataset so that the features were in a clearly readable format for machine learning models in predictive analysis.
- Categorical columns were converted into dummy columns using one-hot encoding technique.
- All dummy columns and other features were converted into numerical columns for machine learning models.

Data Wrangling – Feature Engineering Flowchart



EDA with Data Visualization

- Several charts were plotted for the purposes of Exploratory Data Analysis to better understand the dataset.
- Flight Numbers were plotted against Launch Sites while also displaying Launch Outcomes on a scatter plot. This was significant since it would display how launch outcomes changed for each site as more and more flights were conducted from them.
- Payloads were plotted against Launch Sites while also displaying Launch Outcomes on a scatter plot. This was done to see if certain Payloads were used only for certain Launch sites and whether this impacted the outcome of the launches.

EDA with Data Visualization

- Success rates of different orbits were plotted on a bar chart. This would tell at a glance how successful launches for each orbit type had been.
- Orbit Types were plotted against Flight Numbers while also displaying Launch Outcomes on a scatter plot. This was done to see if there was any relationship between certain Orbits and Launch Outcomes as well as to see any preferences in flights over time for certain orbits.
- Payloads were plotted against Orbit Types while also displaying Launch Outcomes on a scatter plot. This was to see any relationship between certain Payloads and Orbits as well as the Launch outcomes for these flights.
- Line chart for success rate of flights over time was plotted to observe the trend.

EDA with SQL

- Several SQL queries were performed during Exploratory Data Analysis which are listed below:
- List of all launch sites for space missions
- 5 records where the name of launch site for space mission starts with ‘CCA’
- Total payload mass carried by all the boosters launched on behalf of NASA
- Average payload mass carried by the F9 v1.1 booster version
- Date when the first successful landing outcome for ground pad was achieved
- List of all the boosters that achieved successful landing outcome for drone ship while carrying a payload mass between 4000 and 6000 Kg.

EDA with SQL

- List the total number of all successful and failed mission outcomes
- List of all the booster versions that carried the maximum payload mass mentioned in the dataset
- List of records with month names, failed landing outcomes for drone ships, booster versions and launch site for the year 2015
- Ranking of all the successful landing outcomes between the dates 04-06-2010 and 20-03-2017 in descending order

Build an Interactive Map with Folium

- An interactive map with Folium was built to better understand the locations of the launch sites for the space program and any relevant features pertaining to them.
- In the first instance, all the locations of the launch sites were marked on a Folium map along with labels indicating the names of the sites.
- This was done by making use of Markers and Circles at the coordinates of the sites.
- Then, Marker Clusters were used with different colors for markers to indicate different launch outcomes for each site. Green colored markers indicated successful launches and red colored markers indicated unsuccessful launches.

Build an Interactive Map with Folium

- The Marker Clusters would indicate the number of launches from each launch site when the map is looked at from above.
- Upon zooming in on any location, the breakdown of those flights and their outcomes would become visible to the viewer with different colored markers.
- Finally, points of interest close to the sites were noted and these points were marked on the map by using markers and by drawing lines from the launch sites to these markers.
- Furthermore, the distance of these points from the launch site was labelled clearly on the map next to the point of interest.

Build a Dashboard with Plotly Dash (Pie Chart)

- A dashboard was built by making use of Plotly Dash in Python.
- The dashboard consists of two graphs, a pie chart and a scatter plot and a dropdown menu which has all Launch Site names and All Sites as options.
- The pie chart displays the proportion of successful launch outcomes.
- When All Sites is selected in the dropdown menu it displays the proportion of successful launch outcomes for each site.
- When a specific site is selected, it displays the proportion of successful and unsuccessful launch outcomes for that specific site in the form of a pie chart. The title of the chart changes to reflect the selected site.

Build a Dashboard with Plotly Dash (Scatter Chart)

- The second graph on the dashboard is a scatter chart that displays the relationship between payload mass and success of launch outcomes. T
- There is a slider input above the graph where the minimum and the maximum payload mass displayed on the graph can be inputted by the user. The payloads range from 0 to 10K with steps of 1K.
- The graph also indicates the different boosters used for each launch by using markers of different colors.
- The graph displays the outcome of launches from all sites when All Sites is selected. When a specific site is selected, it displays the launch outcomes of that site. The title of the chart changes to reflect the selected site.
- The different boosters for launch outcomes can be unselected by clicking their icons in the legend. Clicking their icon again selects them again.

Build a Dashboard with Plotly Dash

- The dashboard includes two graphs, both of which represent important relationships between launch outcomes and their successes.
- There is a clear relationship between Launch Site and Launch Outcome. Outcomes from one site tend to be successful most of the time while outcomes from either sites are mixed. The first graph represents that.
- There is also a clear relationship between Payload Mass and Launch Outcome. Launches carrying Payload Mass between a certain threshold tend to be more successful. The second graph tries to capture that.
- The second graph also tries to look at the relationship between Booster Version and Launch Outcome by noting the Booster Version for each Launch. Some Boosters are more successful than others.

Predictive Analysis (Classification)

- Machine Learning was used in order to classify the launch outcomes as landed or not landed based on the cleaned dataset.
- In order to get best results, a number of different classification techniques were implemented on the data set.
- The techniques used were Logistic Regression (LR), Support Vector Machine (SVC), Decision Tree Classifier and K-Nearest Neighbors (KNN) Algorithm.
- Each of the techniques was implemented using the Sci-Kit Learn library in Python which contains objects for each of these classification methods.
- The data set for the project was divided into the features and the labels set. The features set was standardized to eliminate bias due to differences in value ranges.

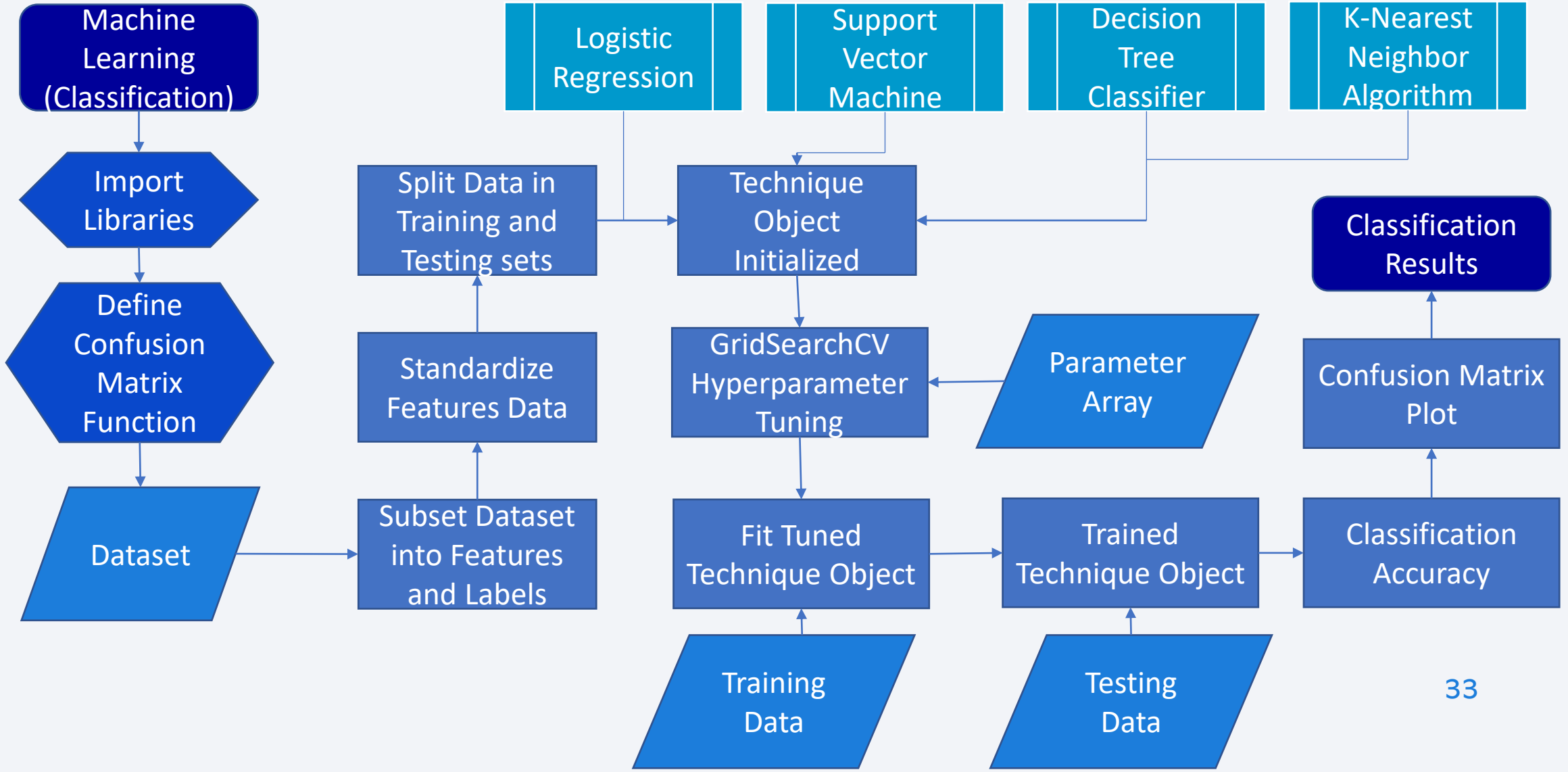
Predictive Analysis (Classification)

- The data set for the project was then split into a training and a testing set. Only 20% of the set was assigned for testing purposes.
- The objects for each technique were initialized. As many of the classification techniques take multiple parameters to be initialized, these parameeters had to be selected.
- For optimal results, the GridSearchCV function was sued to optimize the choice of parameters. The GridSearchCV function was fed a set of parameters and was programmed to choose the set of parameters that maximized the 'accuracy' of the results.
- The GridSearchCV function was also programmed to make use of ten cv folds for each technique.

Predictive Analysis (Classification)

- The resulting optimized parameters, referred to as tuned hyperparameters, were then used for each classification technique object, which in turn was fitted to the training data.
- After the classification object for each technique had been fitted and trained on the training data, its accuracy for predicting values on the training data was measured.
- Afterwards, the trained and fitted classification object for each technique was fed the testing data and its predictions were compared with the actual values in the testing set.
- The accuracy of each classification technique on testing data was measured.
- A confusion matrix for each classification technique was made as well.

Predictive Analysis (Classification) Flowchart

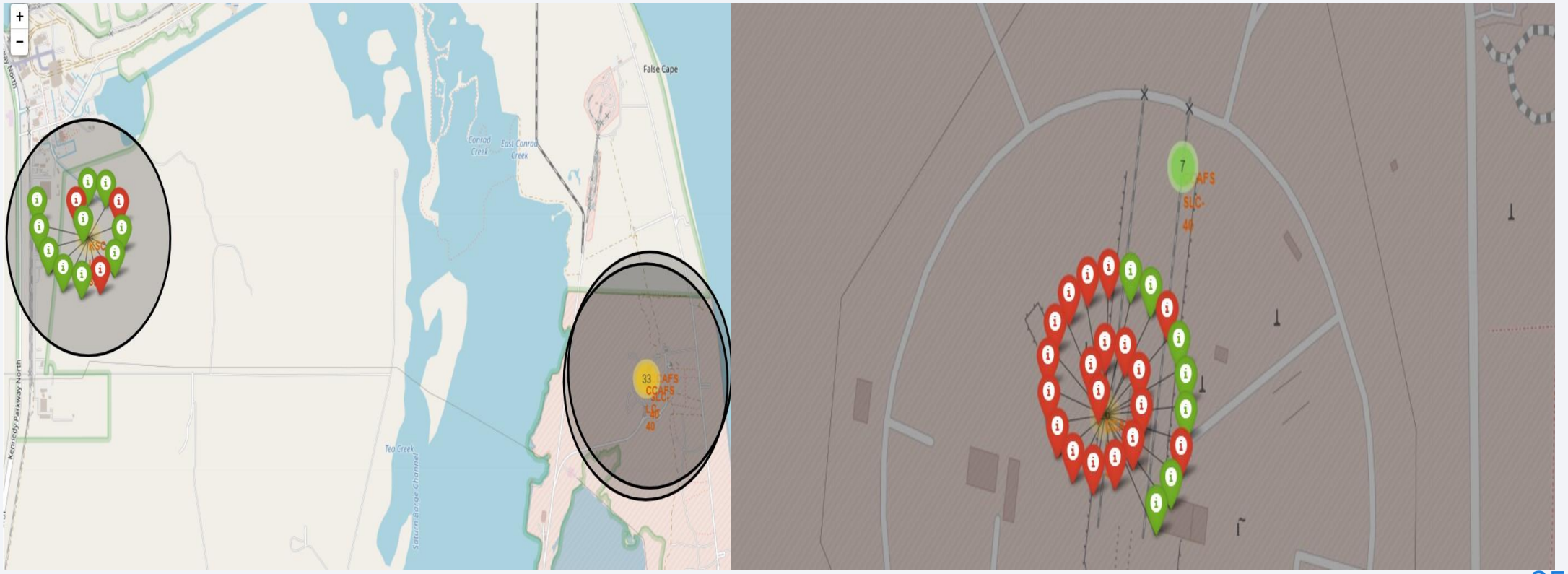


Results (EDA)

- The results from EDA indicated that launch outcome success rate has generally seen an increase as more time passes.
- There is a correlation between payload mass and launch outcome success. Very low masses and very high masses result in launch failures more often than launch successes.
- Launch site also impacts launch outcome success as some launch sites have produced more successful launches than other sites.
- Orbits and launch outcome success are also correlated. Some orbits see more successful launches than others.

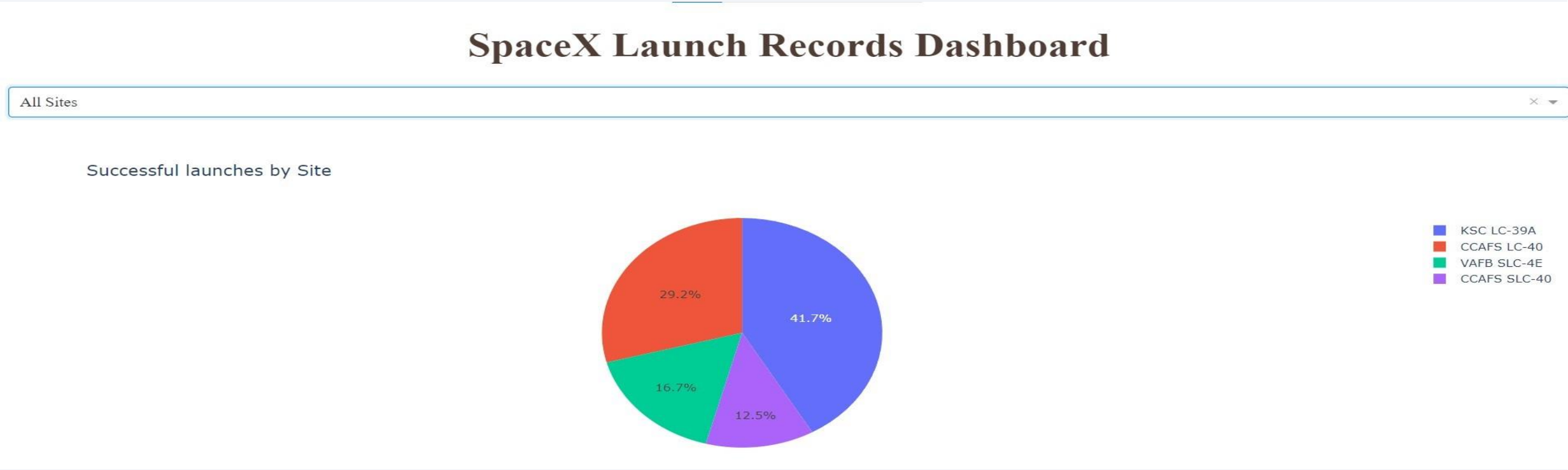
Results (Interactive Analytics)

Interactive Analytics confirmed some results from EDA. Some launch sites do produce more successful outcomes than others.



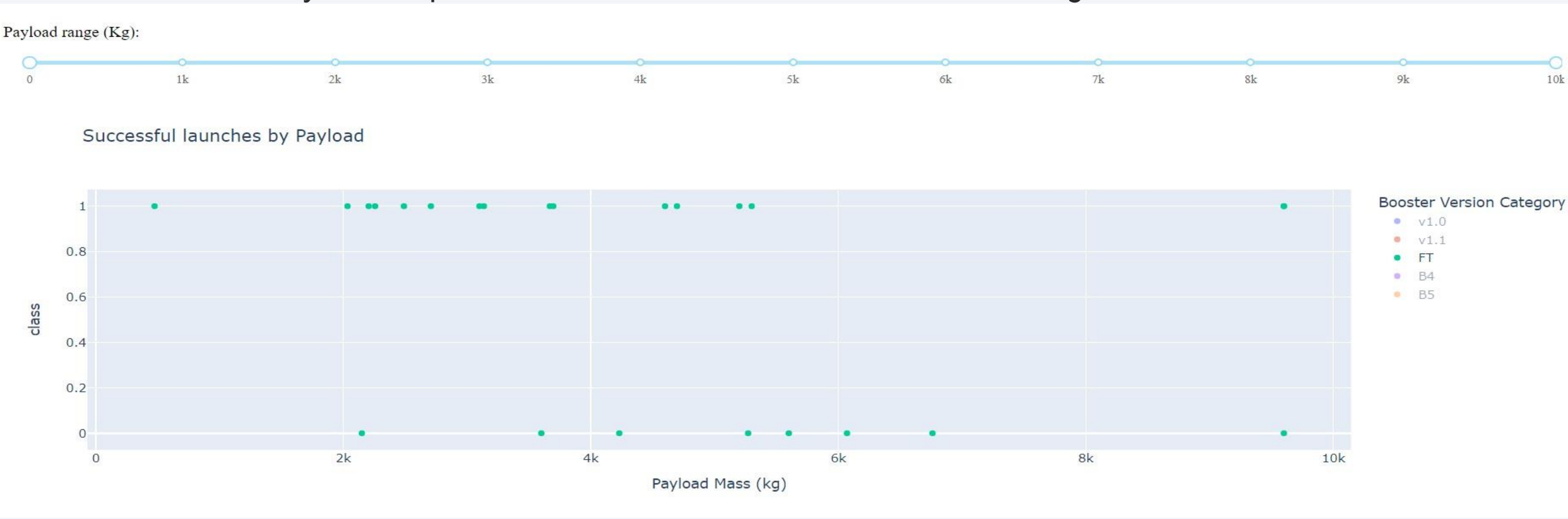
Results (Interactive Analytics)

Interactive Analytics confirmed some results from EDA. Some launch sites do produce more successful outcomes than others.



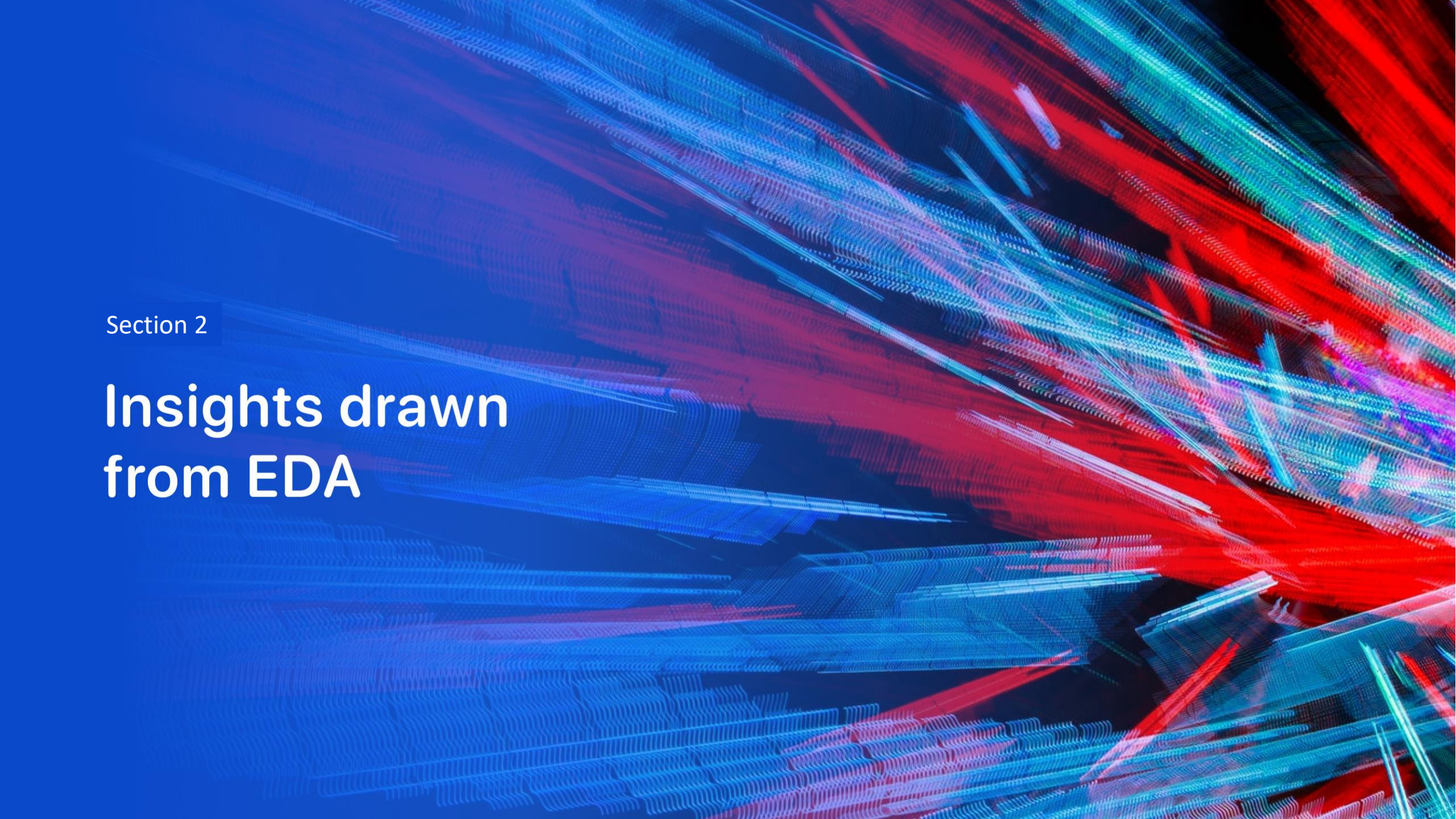
Results (Interactive Analytics)

Interactive Analytics also produced new results such as some boosters being more successful.

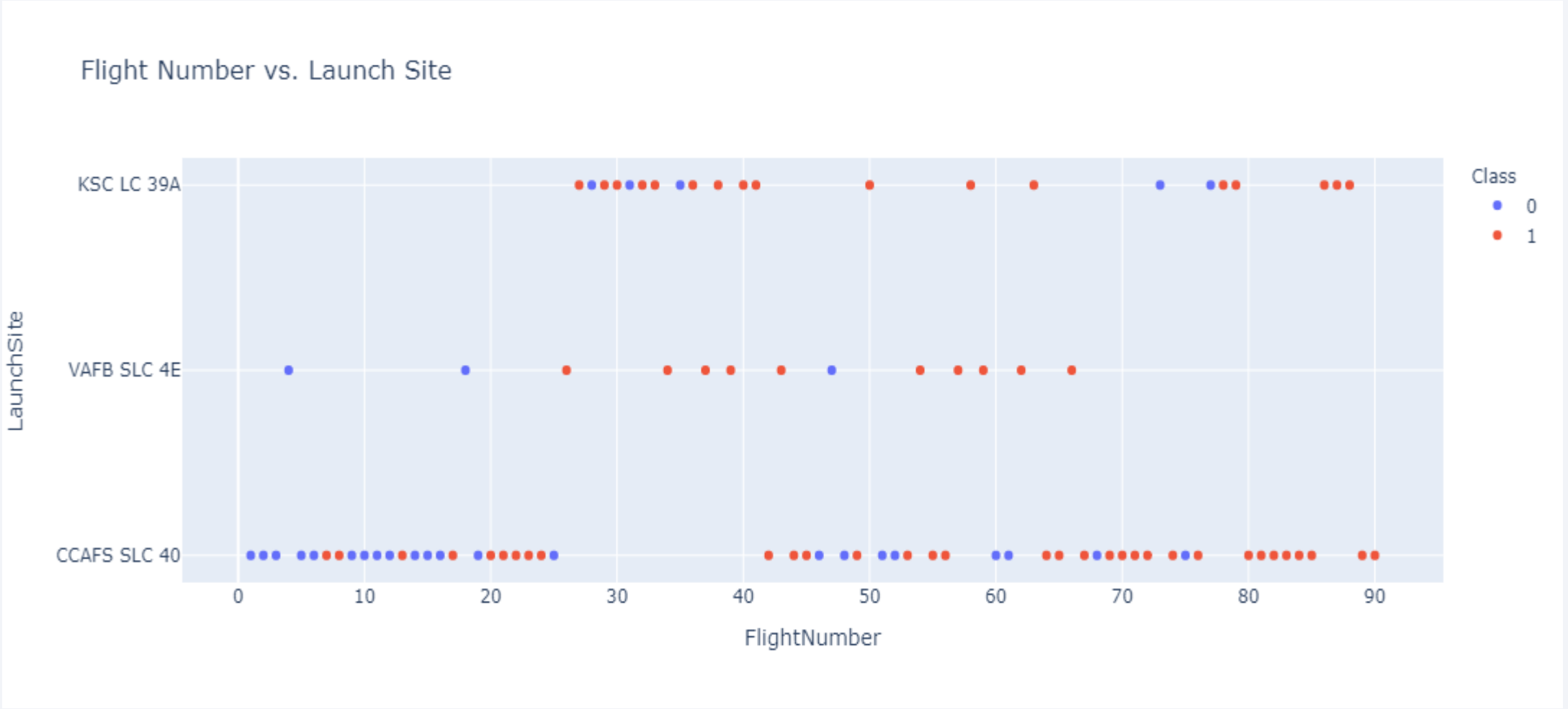


Results (Predictive Analysis)

- The primary predictive analytics technique used in the project was classification with machine learning.
- The techniques used for classification all produced trained models that had the same accuracy for classification on the testing set.
- The results may appear surprising at first but in fact, this is simply because the dataset used for testing was very small (18 samples only).
- The models may be trained with larger datasets and tested on larger sets as well to produce much more interesting results.
- The confusion matrix for each technique gave the same results. False positives emerged as a concern but no false negatives.


Section 2
Insights drawn from EDA

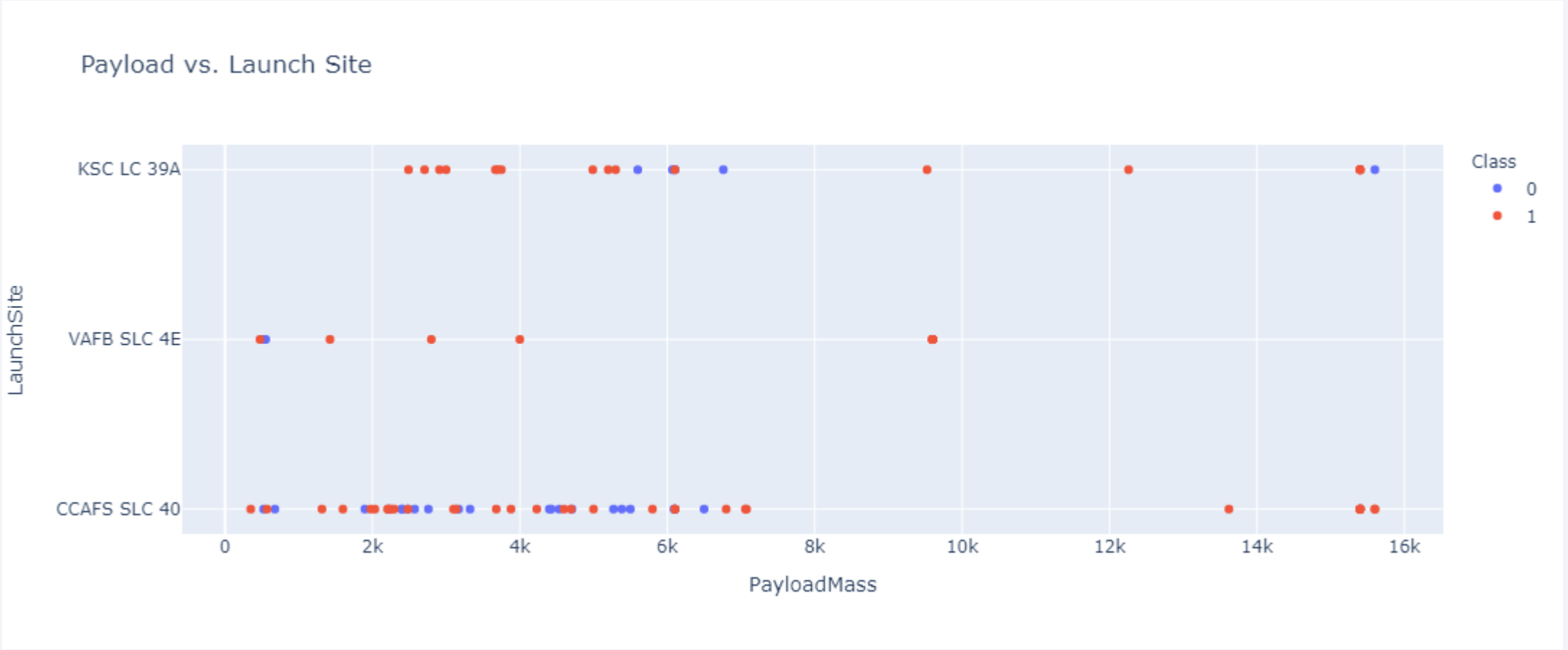
Flight Number vs. Launch Site



Flight Number vs. Launch Site

- The plot shows the correlation between flight numbers and launch site.
- It is clear from the plot that most flights are from CCAFS SLC 40 site. KSC LC 39A was the second most used site while the least flights were from VAFB SLC 4E.
- The plot also reveals that most flights from KSC LC 39A and VAFB SLC 4E have successful landing outcomes.
- Although flights from CCAFS SLC 40 seem to have a mixed record of success, this changes if only recent flights are considered. Then, the record of flights from CCAFS SLC 40 also shows more successes.
- More recent flights in general tend to be more successful regardless of site.

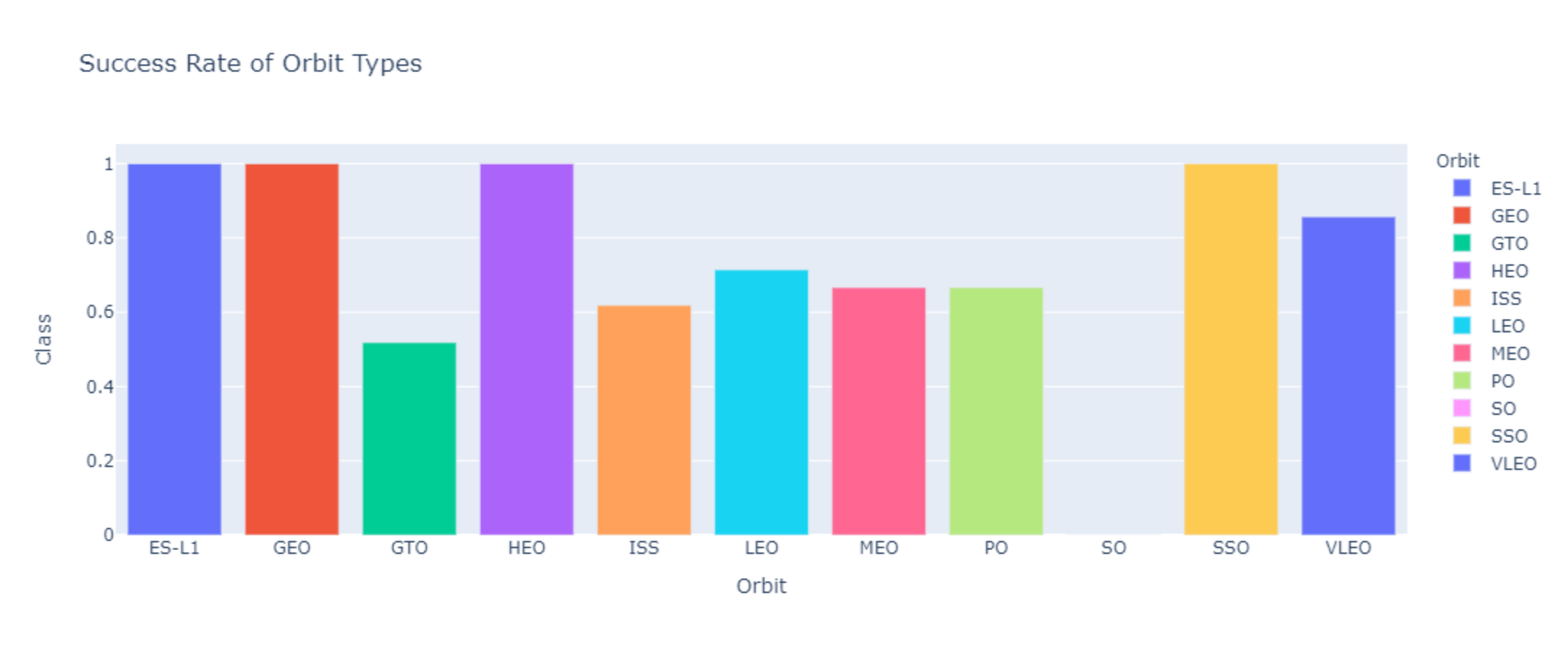
Payload vs. Launch Site



Payload vs. Launch Site

- The plot shows the correlation between payload and launch site.
- The plot clearly indicates that flights with more payload tend to be more successful than flights with lesser payloads.
- VAFB SLC 4E site has not been used for flights with more than 10K payloads.
- Flights from VAFB SLC 4E and KSC LC 39A even with low payloads between 0 to 6K tend to be more successful.
- CCAFS SLC 40 has a mixed record of success for landing outcomes for flights with payloads less than 6K.

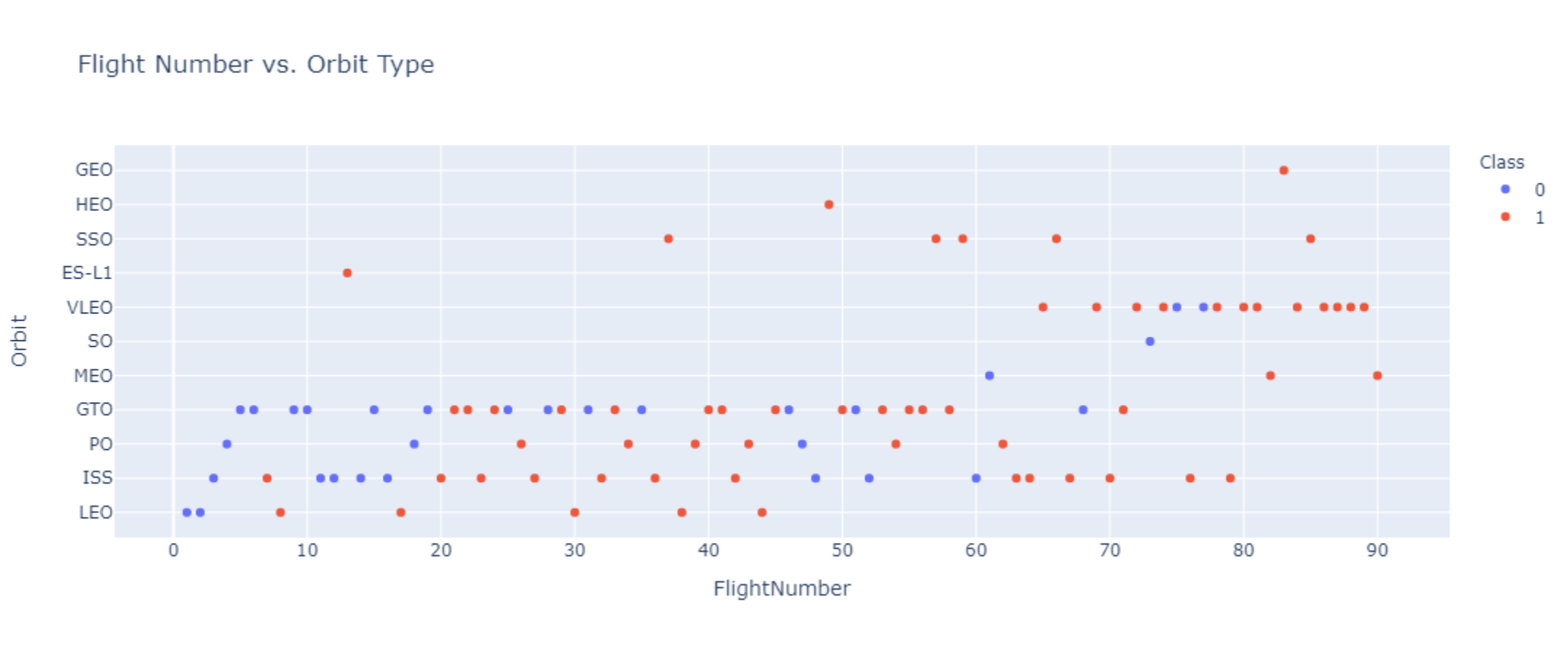
Success Rate vs. Orbit Type



Success Rate vs. Orbit Type

- The plot shows the success rate of launch outcomes for different orbits.
- It is clearly observed from the plot that ES-L1, GEO, HEO and SSO orbits have a success rate of 1. Launches aimed for these orbits tend to be successful.
- VLEO orbit does not have a success rate of 1, but it has a higher success rate than the remaining orbits.
- ISS, LEO, MEO and PO orbits tend to have a success rate that is higher than 60% but lower than 80%.
- GTO orbit has a low success rate close to 50% only.
- SO orbit has a success rate of 0. Very few launches for this orbit were attempted that were unsuccessful.

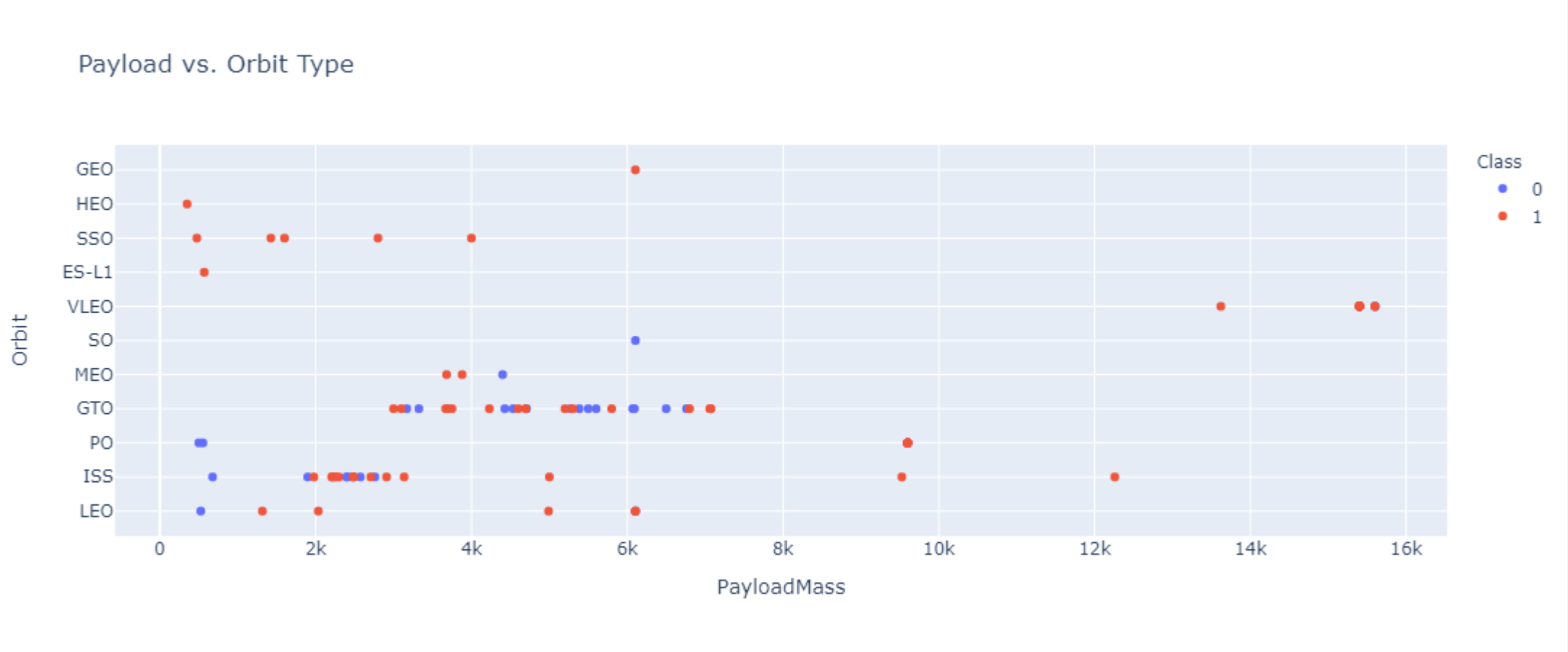
Flight Number vs. Orbit Type



Flight Number vs. Orbit Type

- The plot indicates the correlation between flight number and orbit type.
- It is clear that higher flight numbers are generally more successful regardless of orbit type. Recent flights, above flight number 80, are all successful.
- Recent flights have avoided LEO orbit.
- Most flights have focused on GTO orbit overall, and more recent flights have focused on VLEO orbit.
- A lot of flights have focused on ISS as well. It remains second most frequented orbit both recently and overall.
- Very few flights targeted GEO, HEO, SSO and ES-L1 orbits and were all successful. Few flights targeted MEO and SSO orbits too, mostly unsuccessfully.

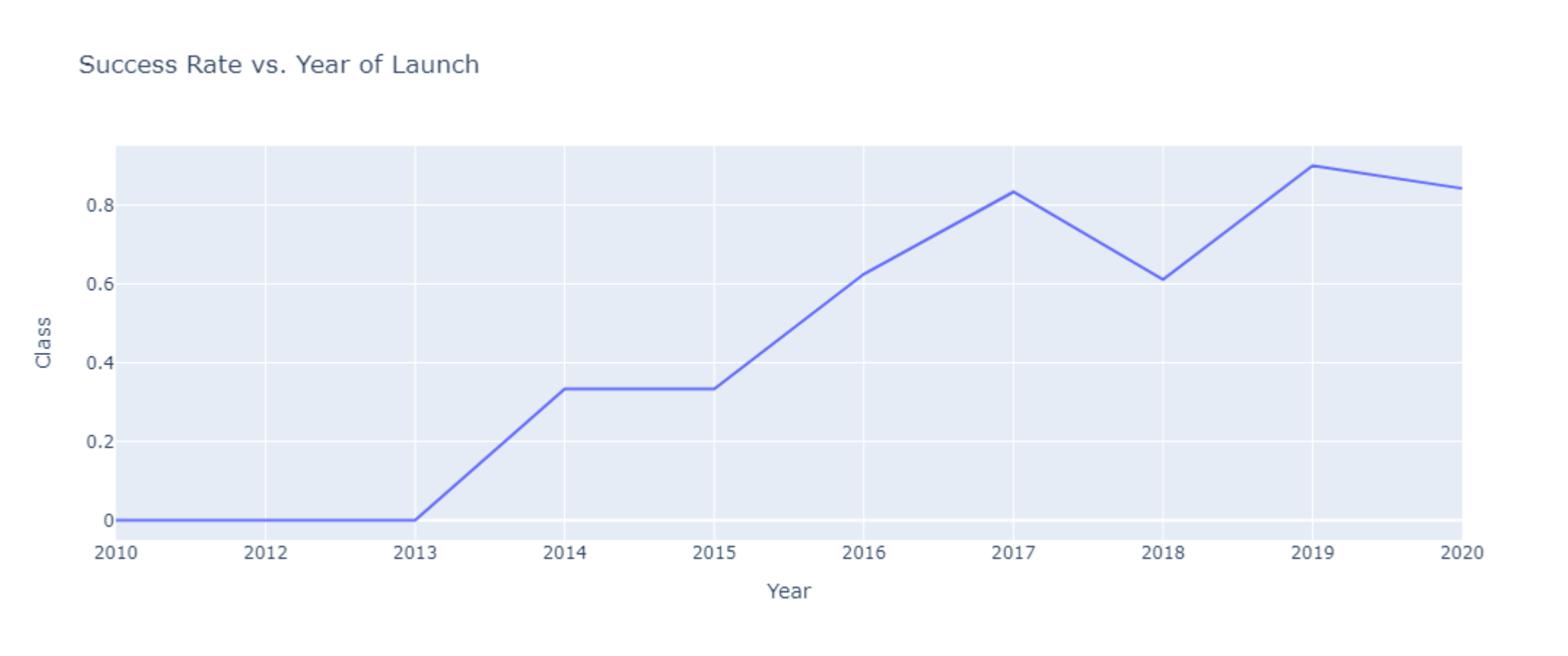
Payload vs. Orbit Type



Payload vs. Orbit Type

- The plot indicates the correlation between payload and orbit type.
- It can be clearly observed that payloads greater than 8K have been delivered successfully, regardless of the targeted orbit.
- VLEO orbit has only been targeted by payloads greater than 8K.
- Payloads for GEO, HEO, SSO and ES-L1 have generally been low, with most less than 4K and all were delivered successfully.
- Lower payloads for PO were unsuccessful but high payload flight was successful.
- High payloads of more than 4K for ISS were successful but lower payloads present mixed results. Results for GTO are mixed regardless of payload.

Launch Success Yearly Trend



Launch Success Yearly Trend

- The graph plots Launch Success Rate by Year.
- It can be clearly observed that as years pass, the success rate of launches has gone up substantially.
- Peak success rate was achieved in 2019. Ever since the success rate has gone down a little.
- The success rate rose from 0 in 2013 sharply till 2014, stayed constant between 2014 and 2015, then rose up sharply for both 2016 and 2017.
- The success rate dipped from 2017 to 2018 then rose to the peak in 2019. It may rise up to a new peak in the future.

Insights from Plots

- The graph plots yielded many important insights on their own but together, they provide some information about overall trends and observations as well.
- First, as flight numbers increased, the probability of launch outcomes being successful seems to increase. This may be because as SpaceX progresses forward, it gains more experience and is more confident of making successful launches.
- Second, there seems to be a relationship between payload mass and launch success. Payload masses that are too low or too high don’t result in many successful launch outcomes.
- Third, certain orbits are more prone to successful launch outcomes than others.

Insights from Plots

- The second and third points together yield the insight that both weight carried by the flight and the orbit targeted by the flight are significant for launch's landing outcome success. Orbit targeted can be thought of as distance to be traveled.
- And finally, some sites produce more successful launch outcomes than others. This could be due to a number of reasons. The more successful sites may be favored for less risky launches or may use more successful technologies etc. Without further data, this remains simply an observation.

All Launch Site Names

- Find the names of the unique launch sites

Task 1

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

```
[10]: Query1="SELECT DISTINCT Launch_Site FROM SPACEXTBL"
pd.read_sql_query(Query1,con)
```

[10]:

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

The query results present the names of all the launch sites mentioned in the dataset.

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`

The query results include 5 records where launch site begins with ‘CCA’,

Task 2

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

[11]:
```
Query2="SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5"
pd.read_sql_query(Query2,con)
```

[11]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
0	04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

Task 3

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

```
Query3="SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_for_NASA_CRS FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'"
pd.read_sql_query(Query3,con)
```

	Total_Payload_Mass_for_NASA_CRS
0	45596

The query calculates the total payload carried by boosters for NASA and presents the sum in the result.

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

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

```
Query4="SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Mass_Carried_By_F9_v1_1 FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'"
pd.read_sql_query(Query4,con)
```

	Average_Mass_Carried_By_F9_v1_1
0	2534.666667

The query calculates the average mass of all the payload carried by Booster Version F9 v1.1 and presents the average in the results.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

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

Hint:Use min function

```
Query5='SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (ground pad)"'
pd.read_sql_query(Query5,con)
```

	MIN(Date)
0	01-05-2017

The query calculates the date of the first successful landing on ground pad by the rocket.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

```
Query6='SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000'
pd.read_sql_query(Query6,con)
```

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

The query lists the boosters that managed to successfully land on drone ships with their payload mass greater than 4000 and less than 6000.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
Query7="SELECT COUNT(*), Mission_Outcome FROM SPACEXTBL GROUP BY Mission_Outcome"
pd.read_sql_query(Query7,con)
```

	COUNT(*)	Mission_Outcome
0	1	Failure (in flight)
1	98	Success
2	1	Success
3	1	Success (payload status unclear)

The query lists the total number of successful and failure mission outcomes. There is an additional Success outcome in the results which may be due to data formatting issues and another Success outcome which is listed as having an unclear payload status.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

```
Query8="SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_= (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)"
pd.read_sql_query(Query8,con)
```

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

The query lists all the boosters which carried the maximum payload mass recorded in the dataset.

2015 Launch Records

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

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
Query9='SELECT substr(Date,7,4) AS year,CASE WHEN substr(Date,4,2)="01" THEN "January" WHEN substr(Date,4,2)="04" THEN "April" END AS month_name,
        "Landing _Outcome",Booster_Version,Launch_Site FROM SPACEXTBL WHERE substr(Date,7,4)="2015" AND "Landing _Outcome"= "Failure (drone ship)"'
pd.read_sql_query(Query9,con)
```

	year	month_name	Landing _Outcome	Booster_Version	Launch_Site
0	2015	January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	2015	April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The query lists the failed landing outcomes for drone ships along with the booster versions of the rockets used and their launch site names as well as the months of those flights for the year 2015. The query is split for readability.

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
Query10='SELECT COUNT(*) AS Succeses, RANK () OVER (ORDER BY COUNT(*) DESC) AS Rank,"Landing _Outcome" FROM SPACEXTBL
        WHERE "Landing _Outcome" LIKE "Success%" AND Date BETWEEN "04-06-2010" AND "20-03-2017" GROUP BY "Landing _Outcome"'
pd.read_sql_query(Query10,con)
```

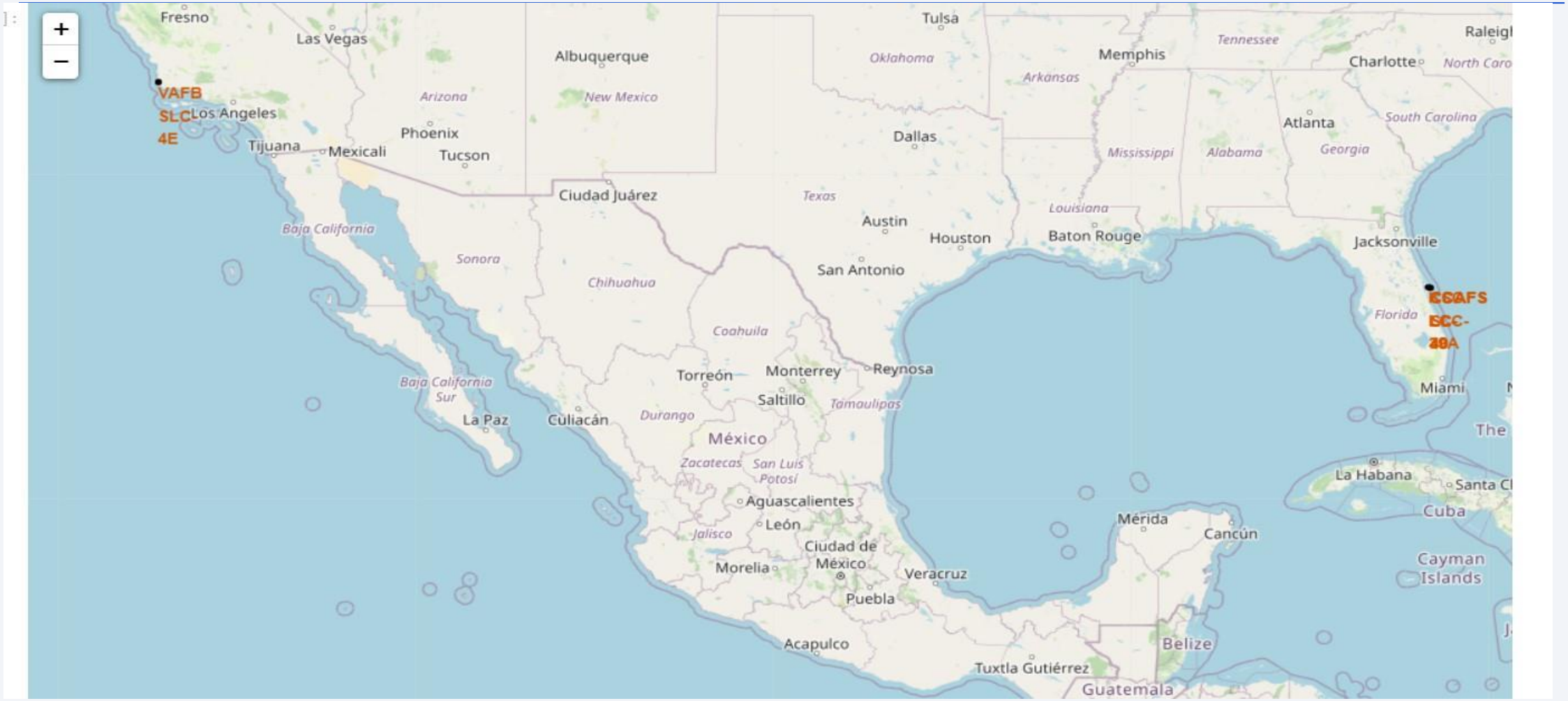
	Succeses	Rank	Landing _Outcome
0	20	1	Success
1	8	2	Success (drone ship)
2	6	3	Success (ground pad)

The query ranks the count of successful landing outcomes between 2010-06-04 and 2017-03-20 in descending order. The query is split for readability. It makes use of RANK () OVER function in sqlite3.

Section 3

Launch Sites Proximities Analysis

Location of Launch Sites Mapped in Folium

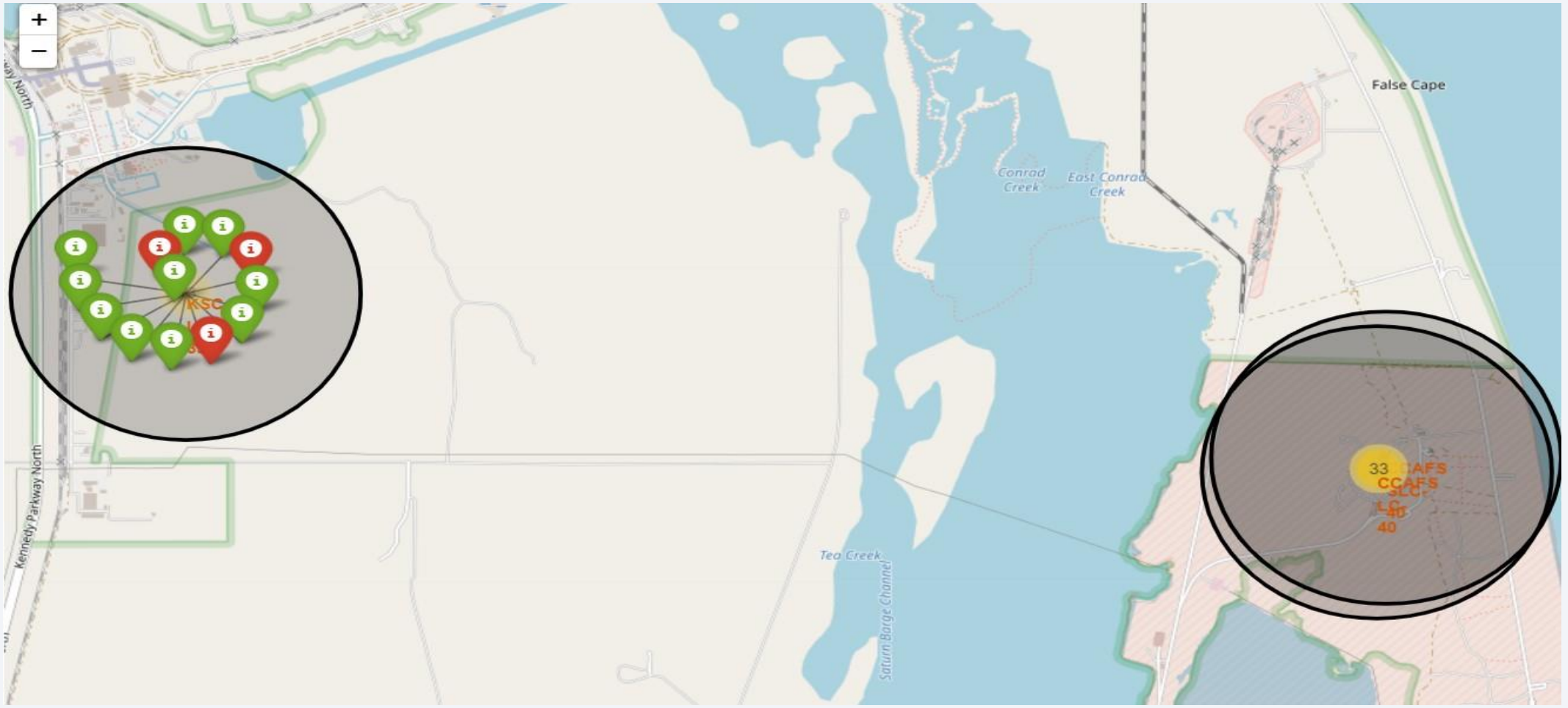


Location of Launch Sites Mapped in Folium

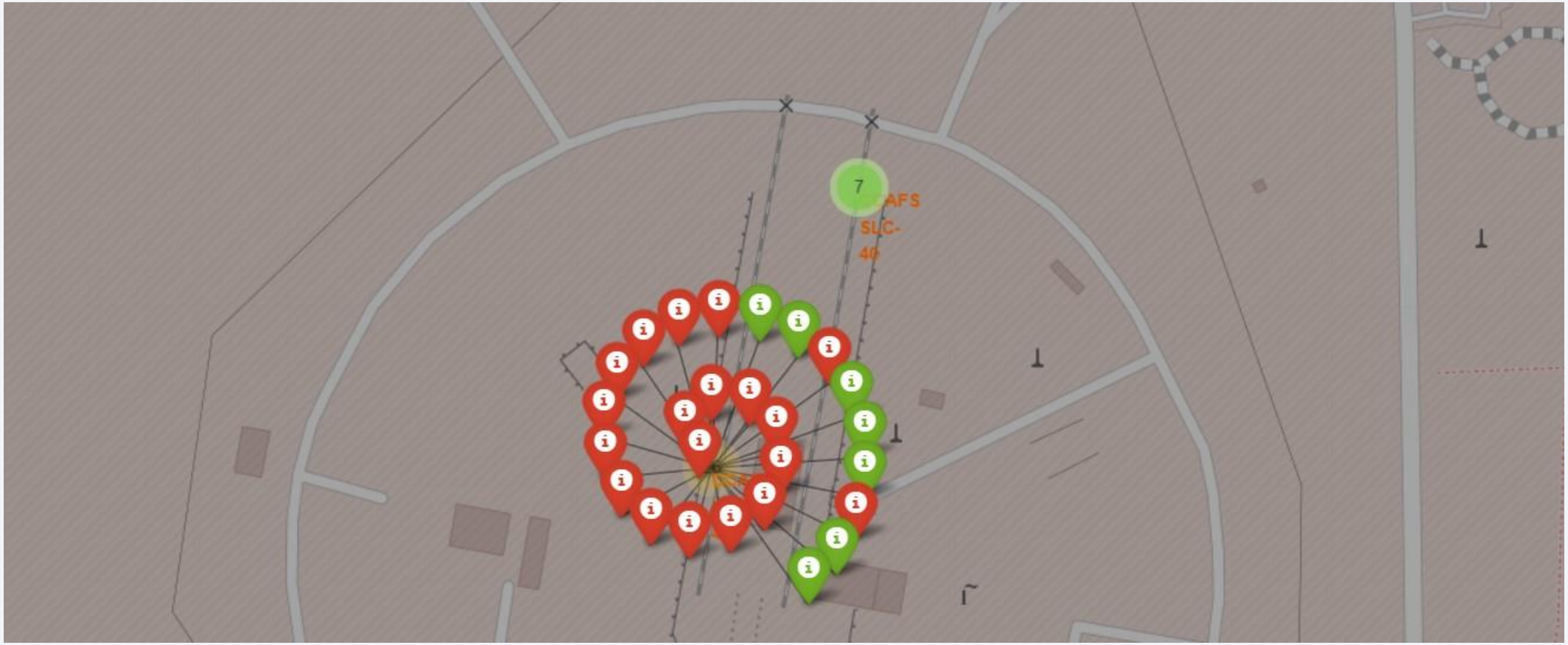
- The folium map marks the location of all the launch sites used by Space X for rocket launches.
- It can be observed from the map that all the sites are located close to the equator. This is because land at the equator is moving at the fastest speed of about 1670 km/hour. Thus, spacecraft launched closer to the equator get a speed boost.
- The launch sites are also located at coastal locations. This makes it easier to transport recovered rockets back to the space launch sites. The launched rockets may also crash in the ocean if a failure occurs instead of inhabited land.
- Most sites are located on the east coast. This is because rockets traveling eastwards get a speed boost from the earth's natural spin from west-to-east.

Launch Outcomes for Sites marked on Folium map

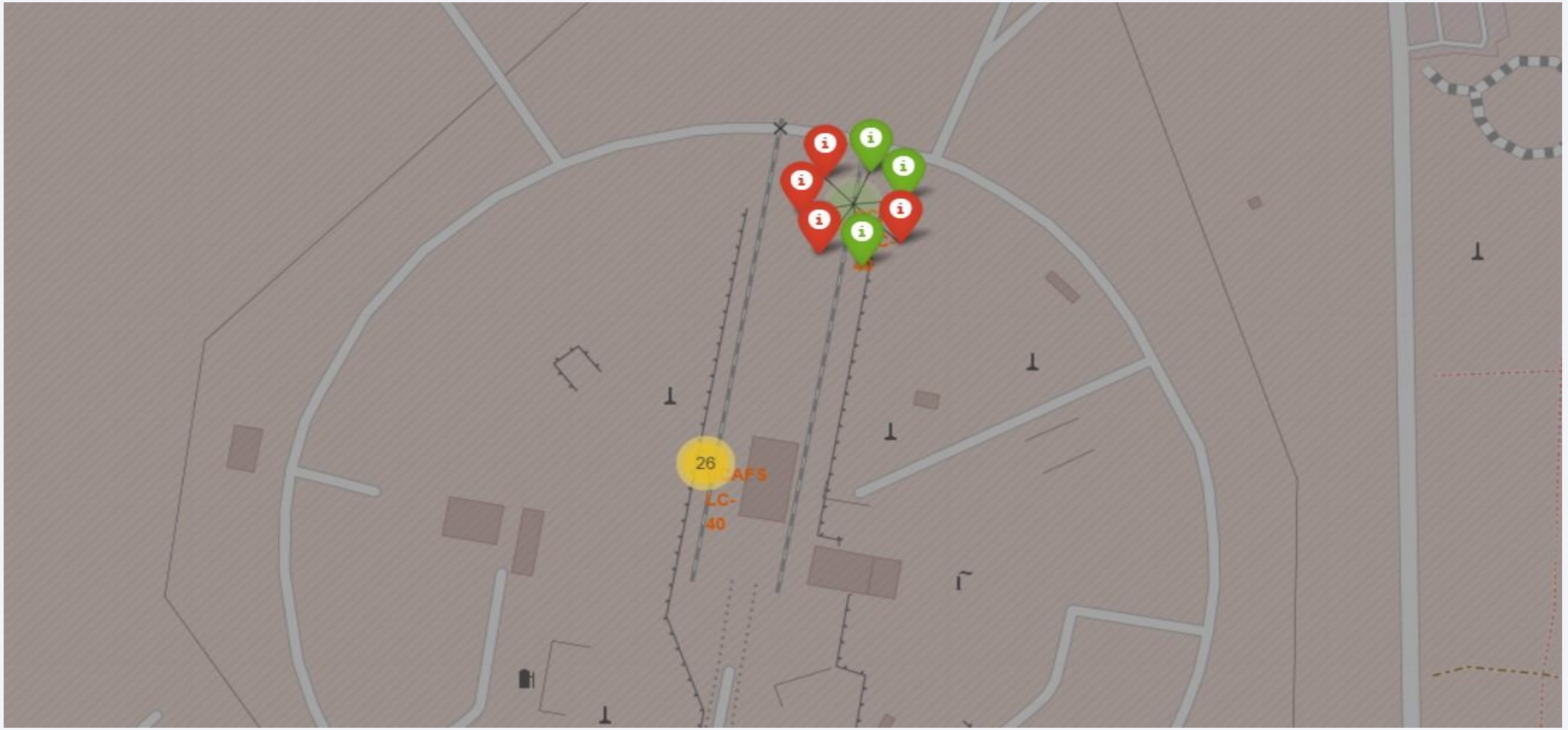
Launch Outcomes for Sites marked on Folium map



Launch Outcomes for Sites marked on Folium map



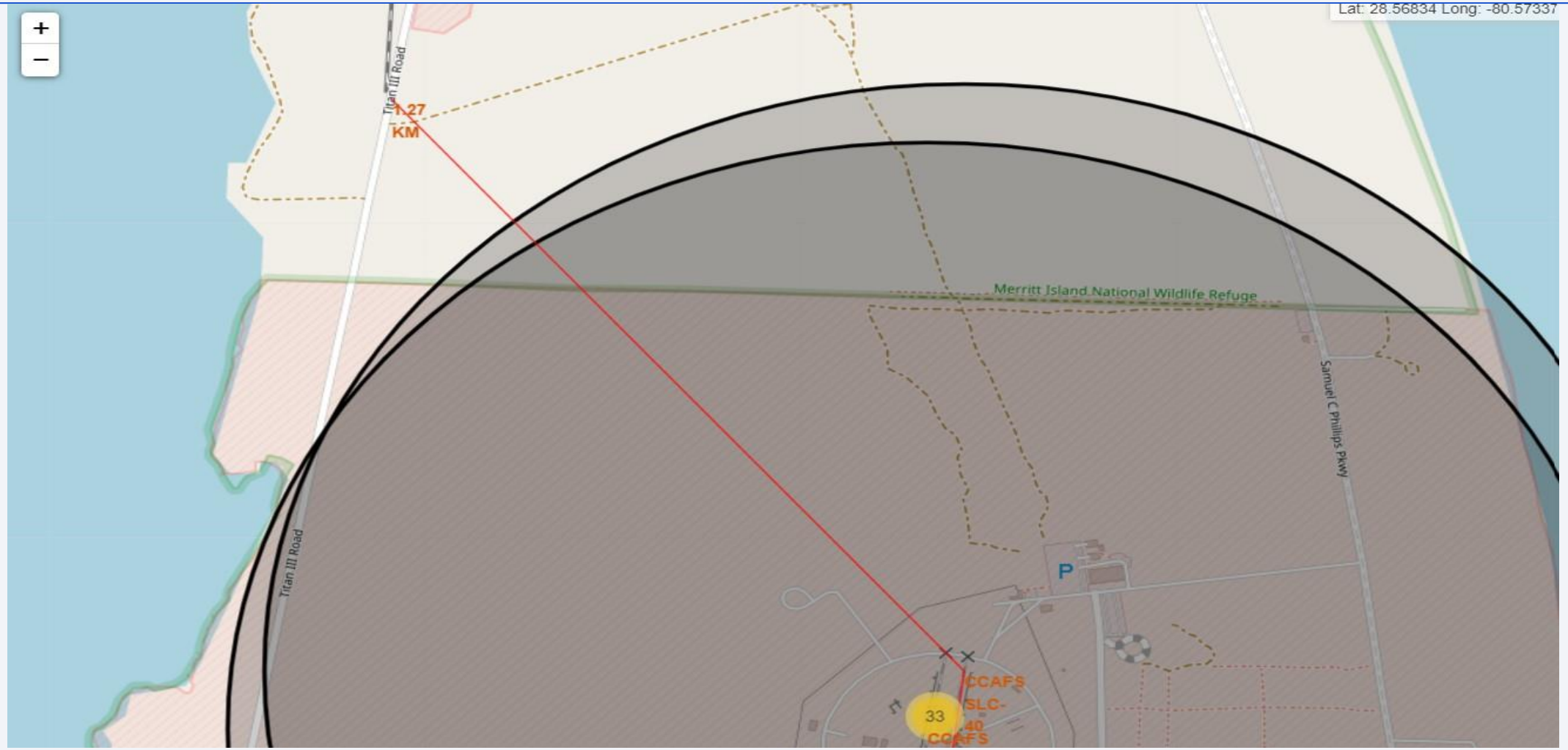
Launch Outcomes for Sites marked on Folium map



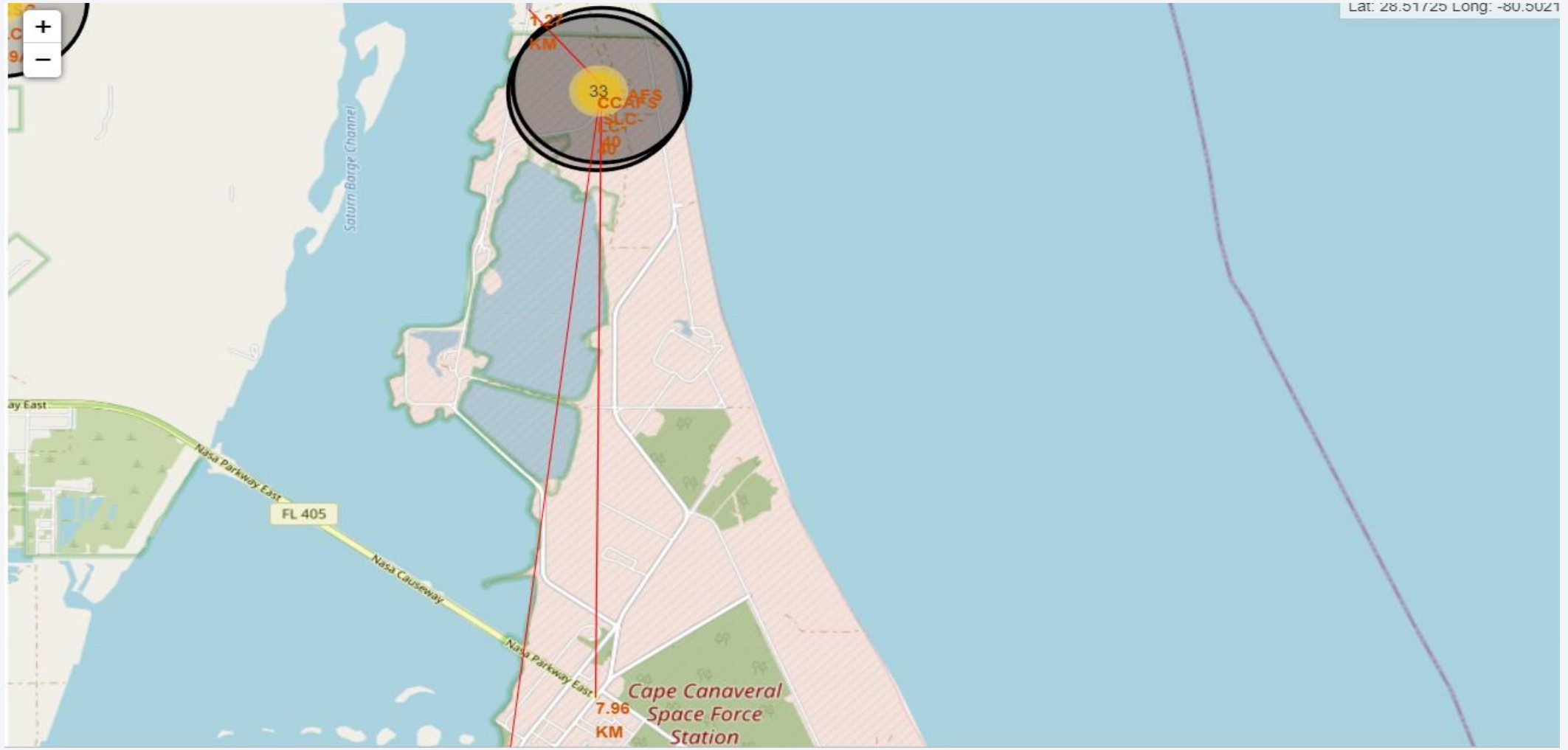
Launch Outcomes for Sites marked on Folium map

- The folium map marks the launch outcomes for all the launch sites used by Space X for rocket launches.
- For greater ease, launch outcomes for each of the sites were recorded separately.
- From the launch outcomes marked on the map, it can be observed that the launch outcomes for most of the sites are not particularly encouraging, with the number of failures (in red) exceeding the number of successes (in green).
- The sole exception to this is the KSC LC 39A site, which has a much better record, with the number of successes (in green) far outweighing the number of failures (in red).

Points of Interest close to Launch Sites on the Folium map

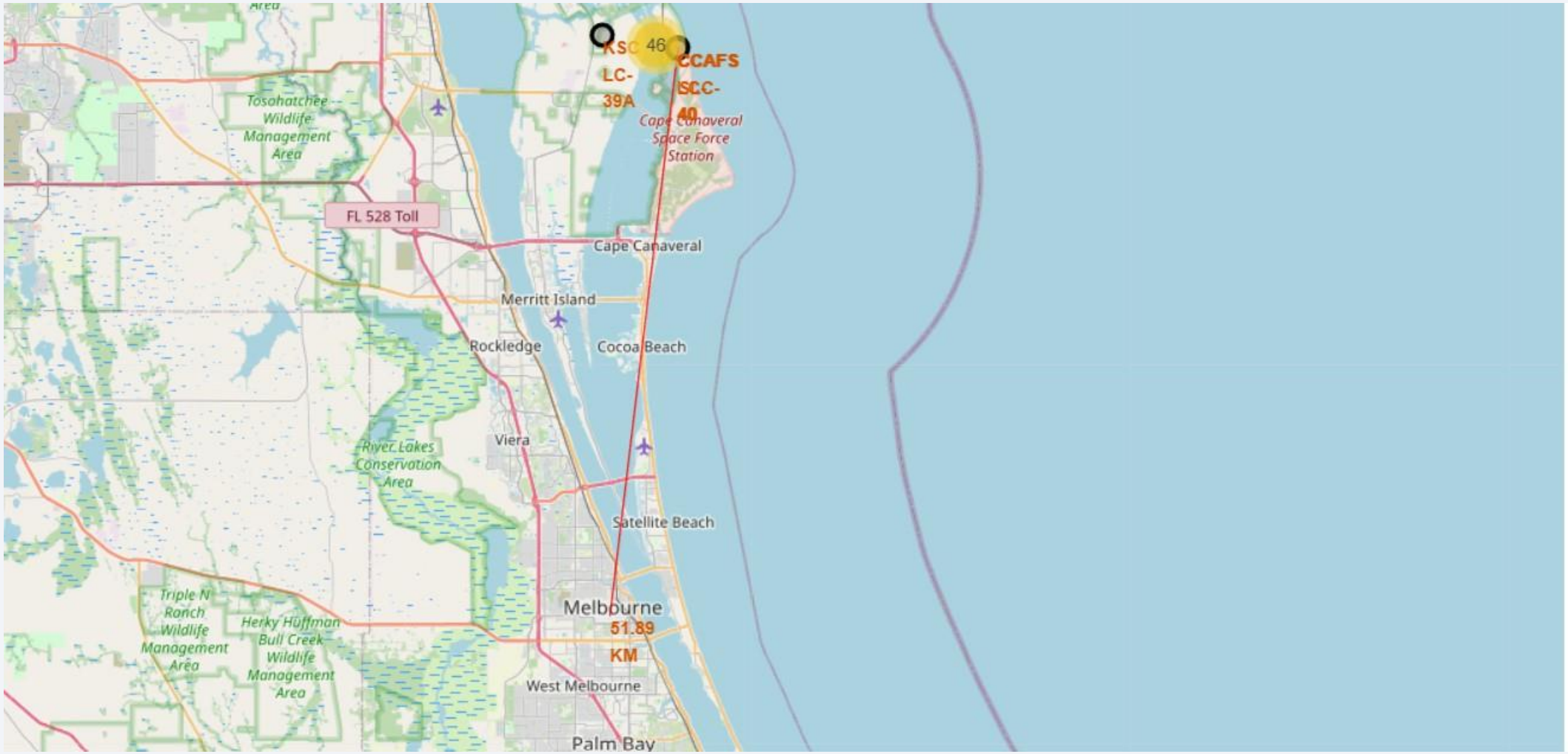


Points of Interest close to Launch Sites on the Folium map



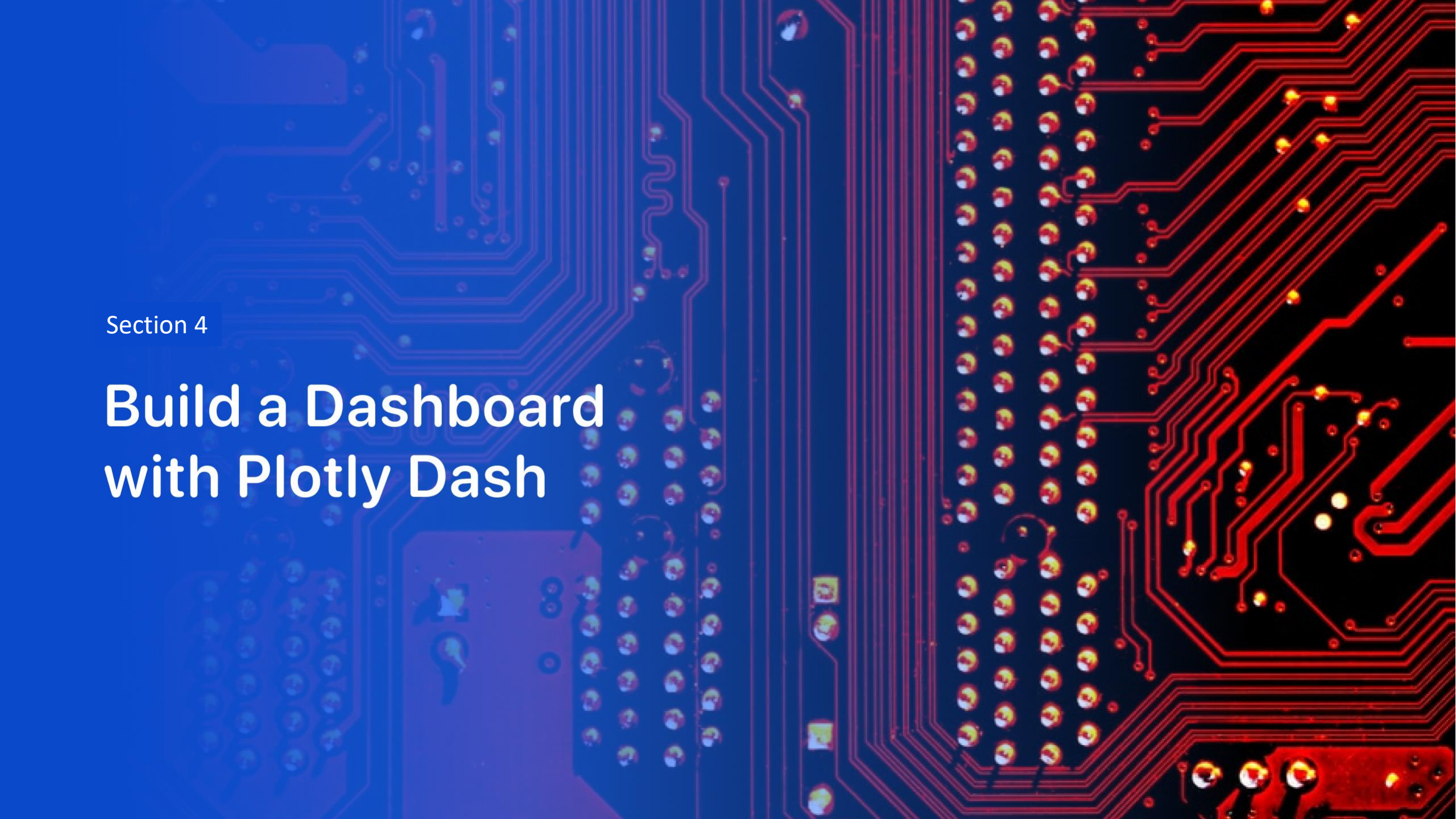
Points of Interest close to Launch Sites on the Folium map

Points of Interest close to Launch Sites on the Folium map

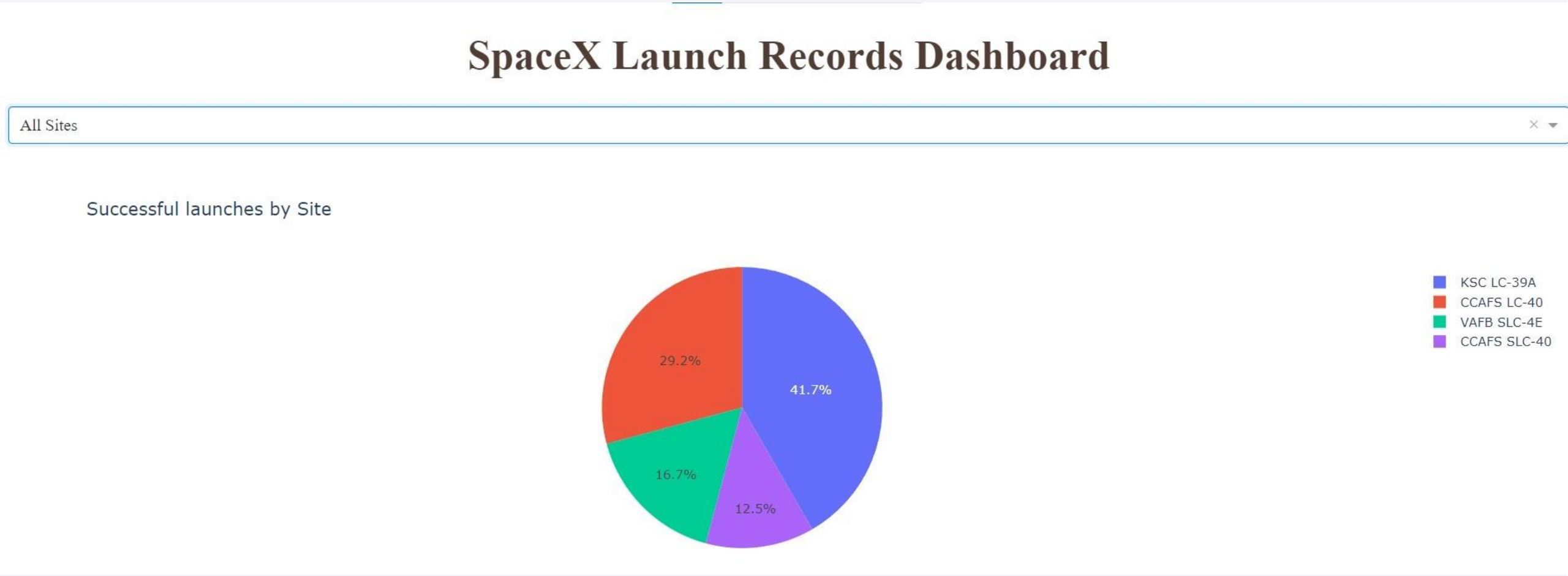


Points of Interest close to Launch Sites on the Folium map

- The folium map marks some points of interest close to the launch sites.
- The map gives the distance to these points from the launch sites in numbers as well as by plotting lines to these points from the launch sites.
- The points chosen indicate that the launch sites are placed close to different kinds of infrastructure such as highways and railroads. That makes sense since it makes it easier to manage logistics for these launch sites.
- The sites are also located far from inhabited places. This is also sensible since space launches can be noisy and release fumes, which would be a source of disturbance for any nearby habitants.


Section 4
Build a Dashboard with Plotly Dash

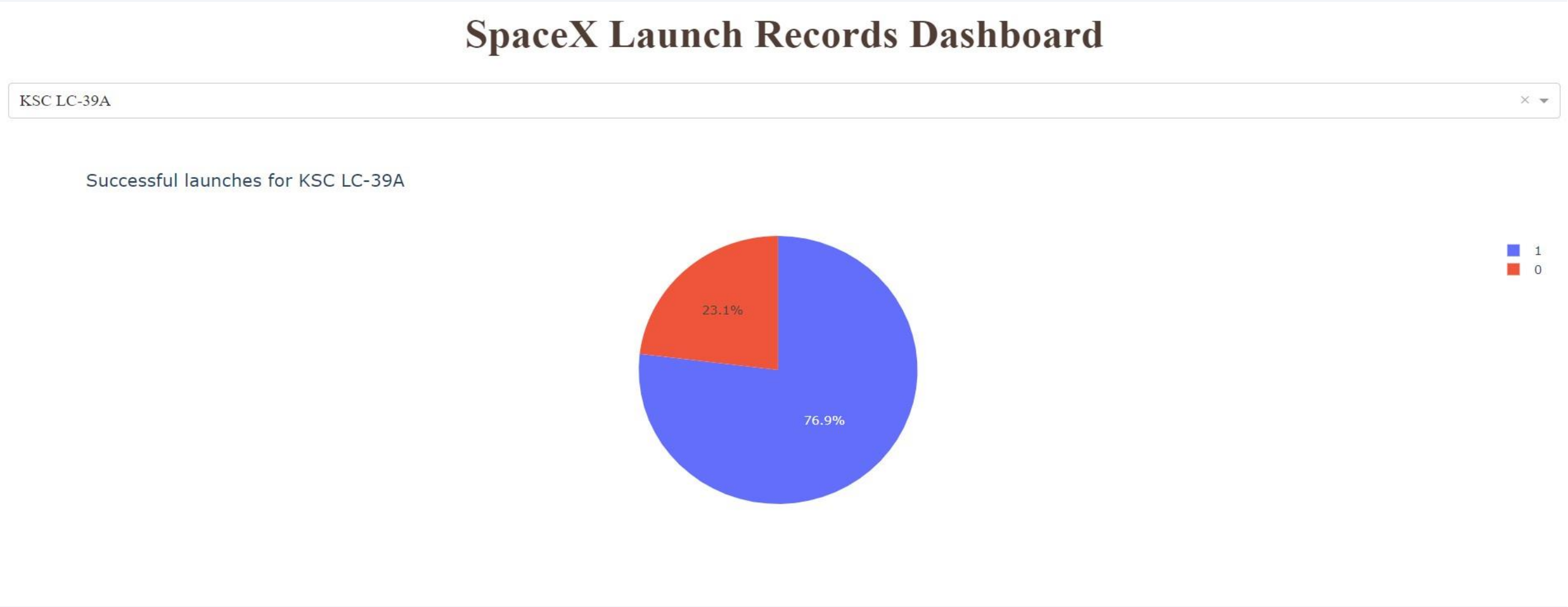
Launch Success for All Sites



Launch Success for All Sites

- The dashboard displays a pie chart with ALL selected from the menu.
- The pie chart records the success rate of landing outcomes for all launch sites as a portion of the total successes for all launches.
- It can be seen from the pie chart that KSC LC-39A is the site which has the greatest portion of successful landing outcomes of all the successful launches, accounting for 41.7% of all the successes.
- CCAFS LC-40 is the second most successful site, with 29.2% of all the successful landing outcomes taking place at this site.
- The other two sites are not remarkable, accounting for small portions of successful landing outcomes below 20%.

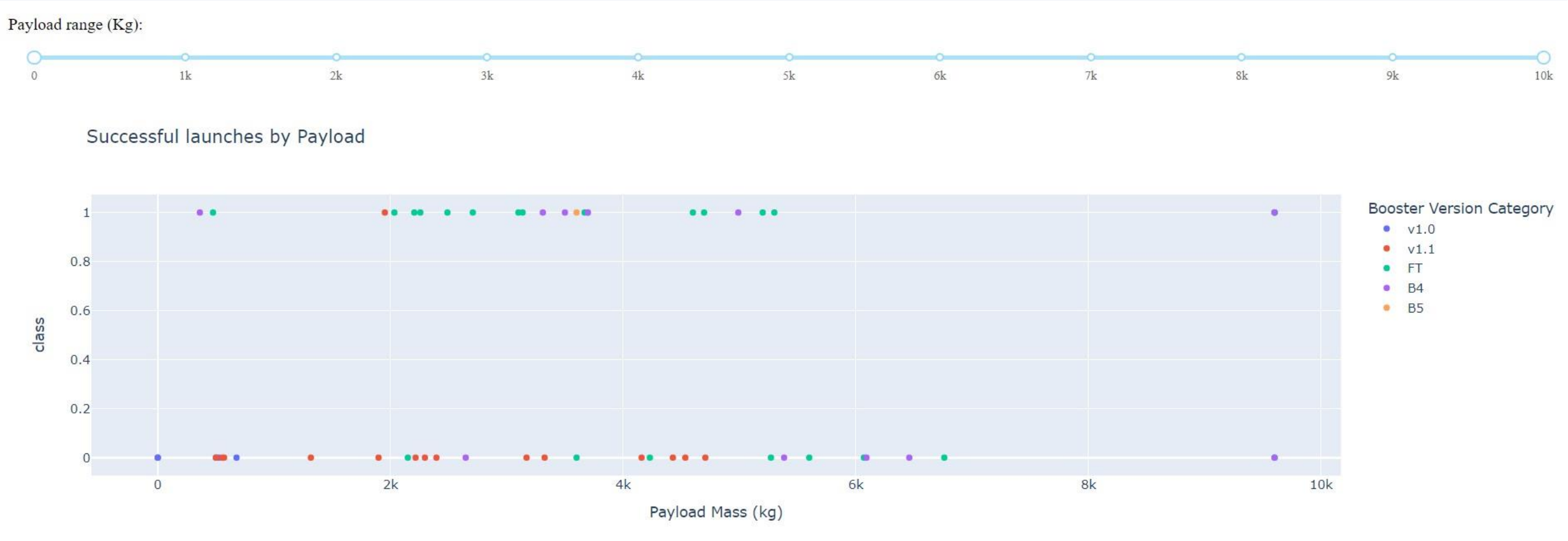
Launch Site with Highest Success Rate



Launch Site with Highest Success Rate

- The dashboard displays a pie chart with KSC LC-39A selected from the dropdown menu. The title of the pie chart changes to reflect that.
- The pie chart displays the overall rate of success for all the launches made from the launch site, in this case KSC LC-39A.
- From the pie chart, it can be clearly observed that the rate of success for KSC LC-39A is 76.9%, which is an impressive figure. Only 23.1% of the launches from this site failed to get a positive landing outcome.
- KSC LC-39A has the highest number of successful launches % age wise and the highest rate of successful launches for landing outcomes.

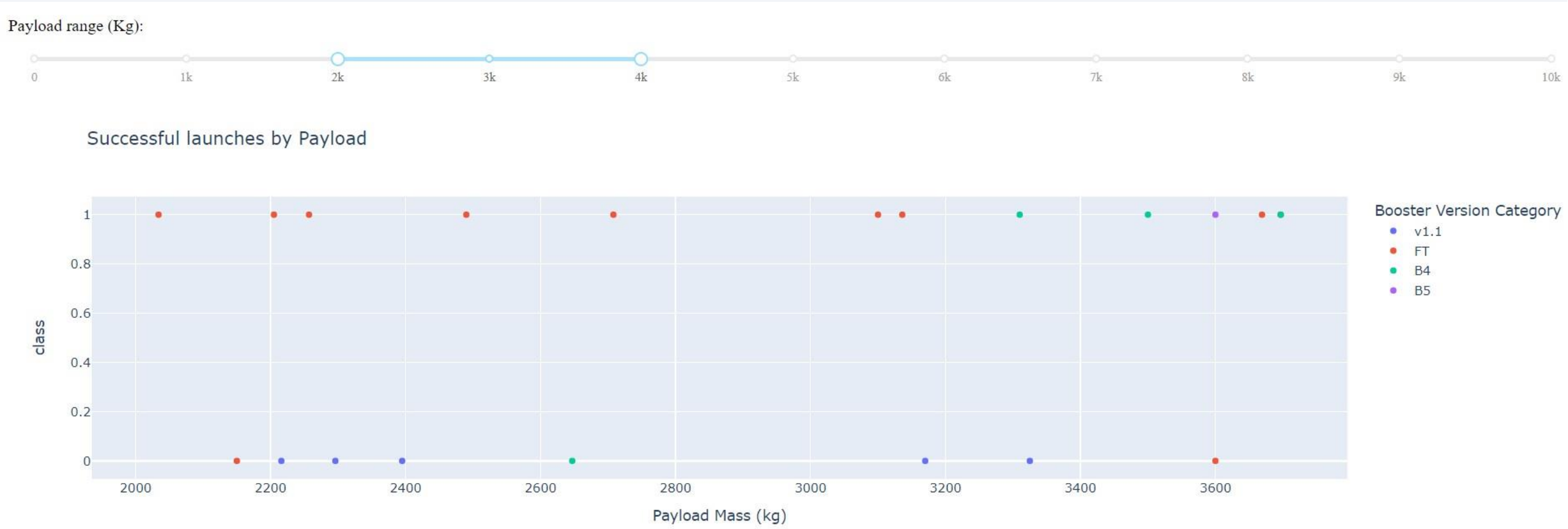
Payloads and Launch Outcomes



Payloads and Launch Outcomes

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- On the graph, all the launches for all the recorded payloads, their success status and the booster used for those launches are presented.
- The graph makes it clear that payloads which are below 6K tend to be more successful than payloads above 6K.
- Another point of interest is that most successful landing outcomes tend to be with the FT booster.

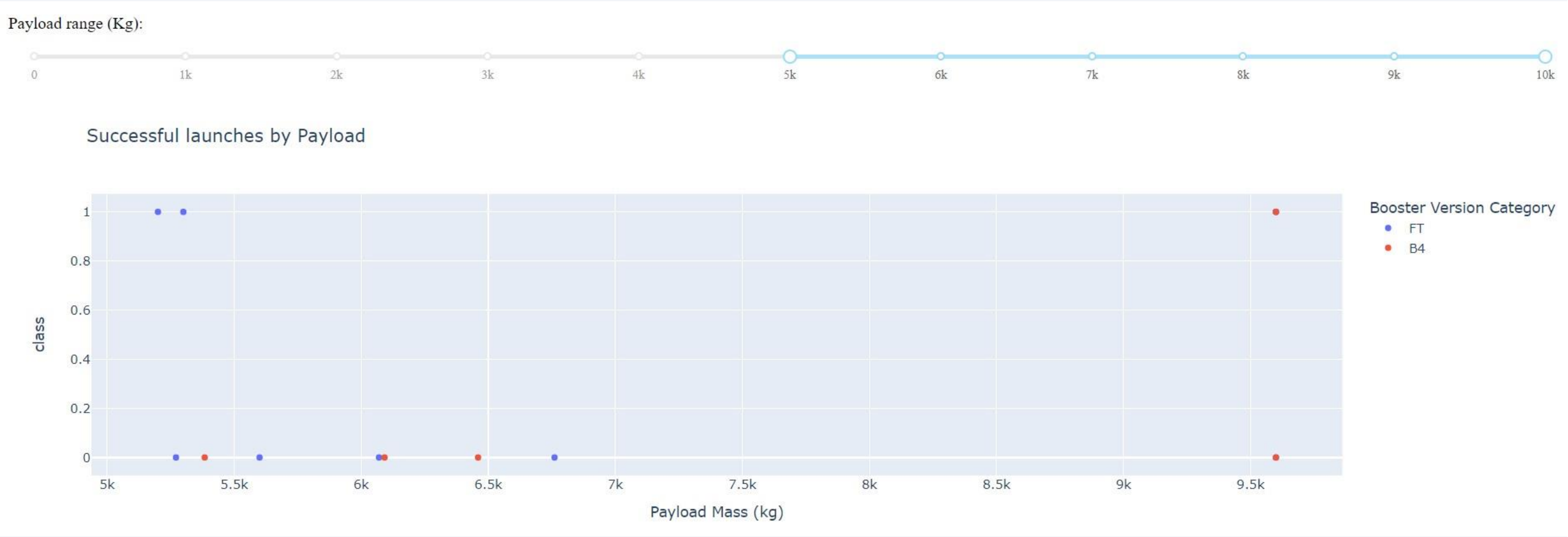
Most Successful Payload Range



Most Successful Payload Range

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- From the graph, the range of payloads falling between 2K and 4K was deemed the most successful payload range.
- As can be seen from the graph, most launches with payloads in this range are successful.
- The FT Booster seems to have the largest number of successful landing outcomes compared to other boosters.

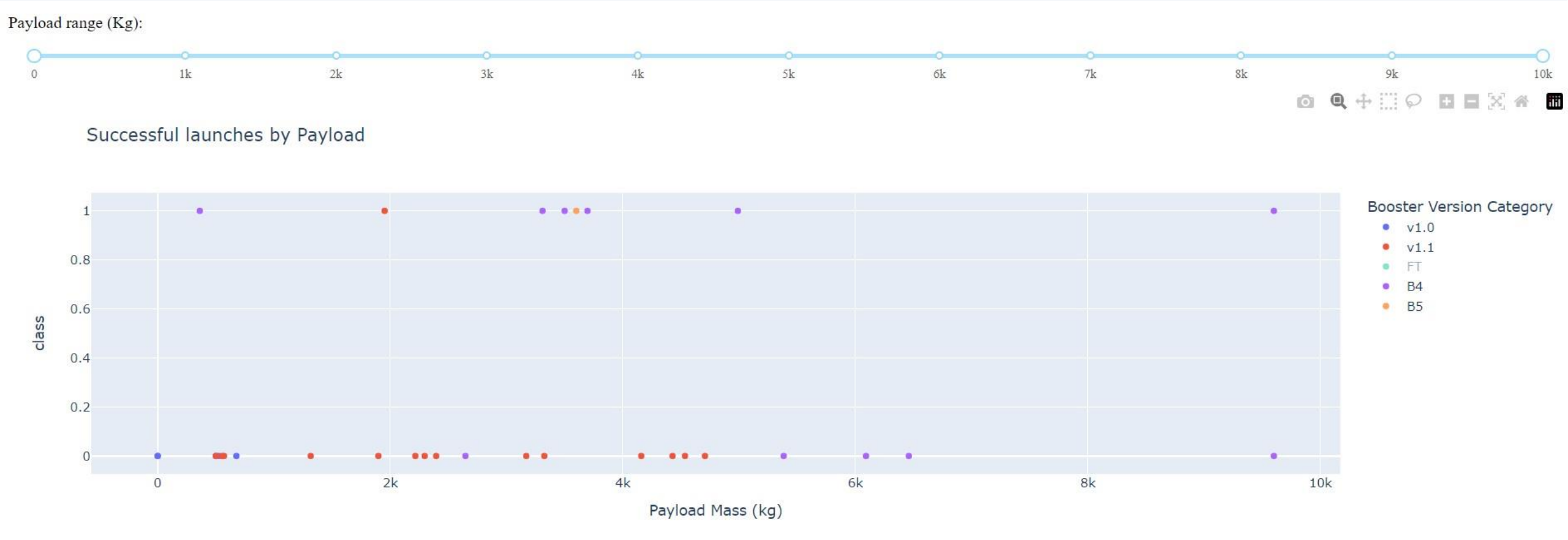
Payloads and Launch Outcomes



Least Successful Payload Range

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- From the graph, the range of payloads falling between 5K and 10K was deemed the least successful payload range.
- As can be seen from the graph, most landing outcomes with payloads in this range are failures.
- The FT Booster seems to be more successful in this range as well compared to the other booster, B4.

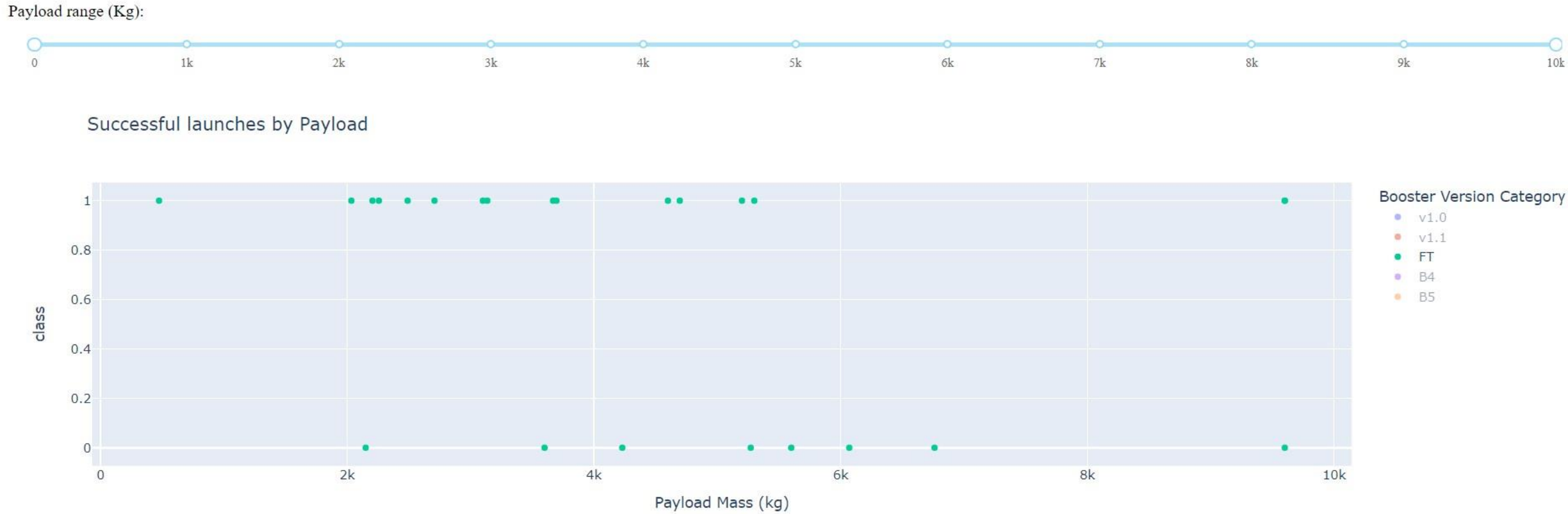
Payloads and Launch Outcomes without FT Booster



Payloads and Launch Outcomes without FT Booster

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- To better understand the role of Boosters in landing outcome success, all launches with Boosters other than the FT Booster were plotted on the graph.
- Without FT Booster, the B4 booster seemed to be the most successful booster overall.
- However, its fail success ratio seems close to 50% on simple observation.

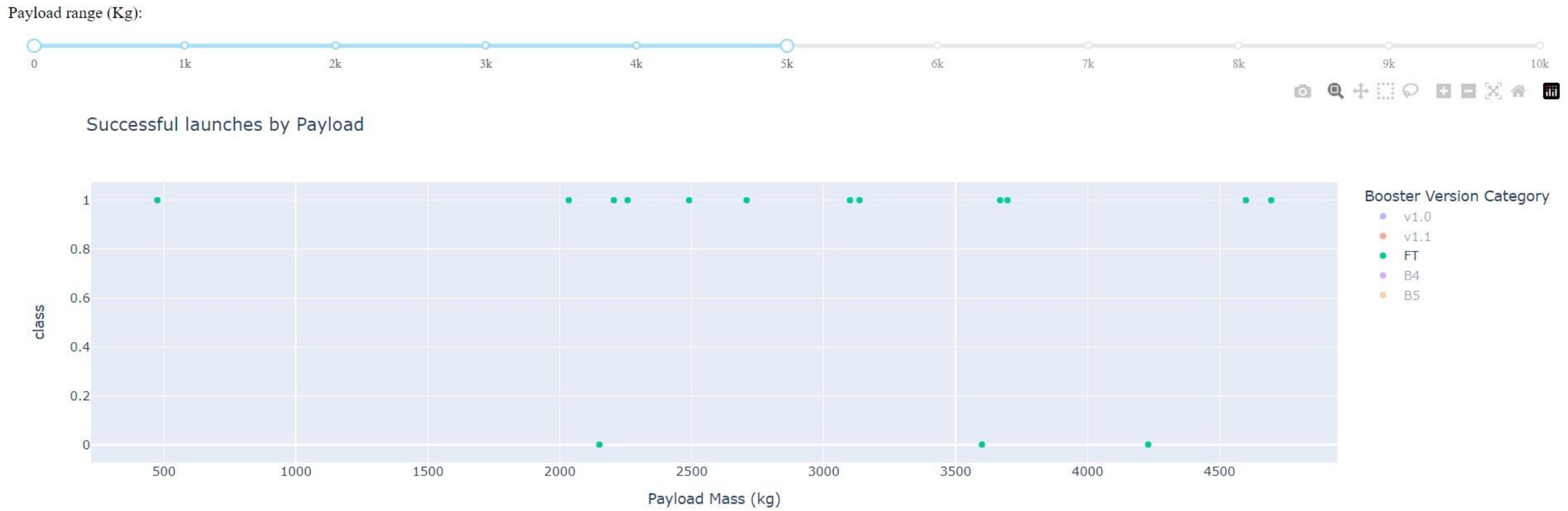
Payloads and Launch Outcomes with FT Booster only



Payloads and Launch Outcomes with FT Booster only

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- To better understand the role of FT Booster in landing outcome success, only launches with FT Booster were plotted on the graph.
- The FT Booster seems to enjoy a good rate of success, with a large number of its landing outcomes being successful. On observation, it seems to have a 66% rate of success which is good.
- However, there appears to be a clear bias in its successes with most of them occurring for launches with lower payloads.

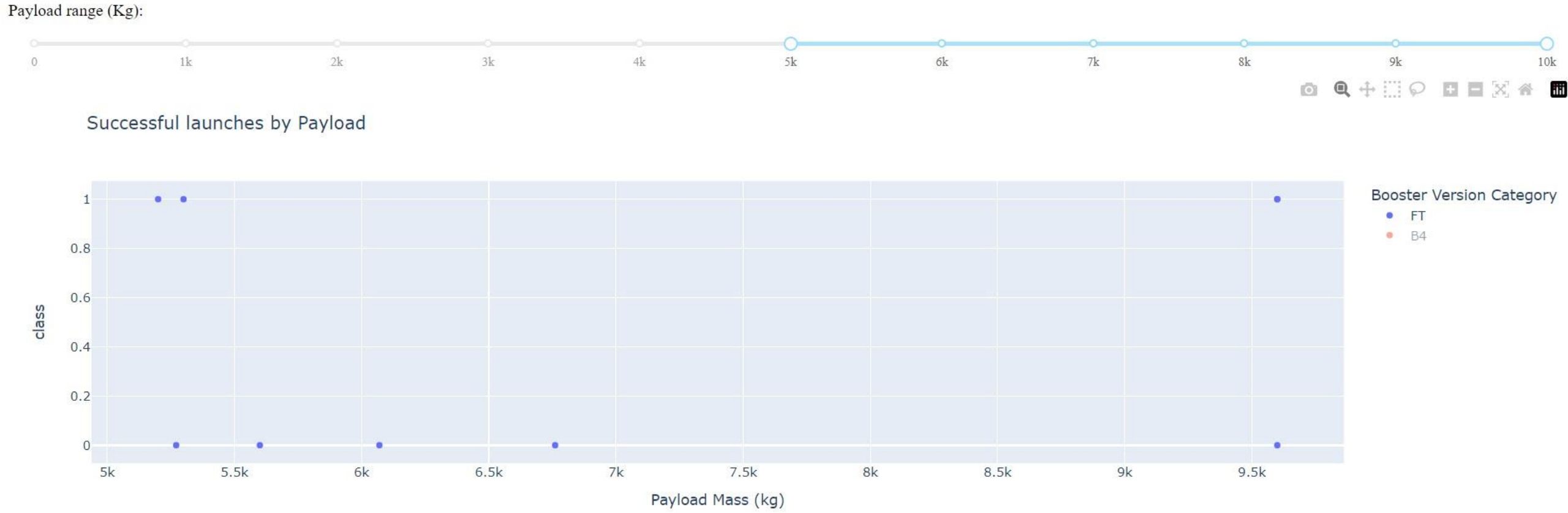
Small Payloads and Launch Outcomes with FT Booster only



Small Payloads and Launch Outcomes with FT Booster only

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- To better understand the role of payloads in landing outcome successes of FT Booster, only launches with payloads less than 5K for FT Booster were plotted on the graph.
- The success rate of landing outcomes with FT Booster jumps up considerably if only payloads less than 5K are considered. On observation, it appears to be around 80% which is very impressive!

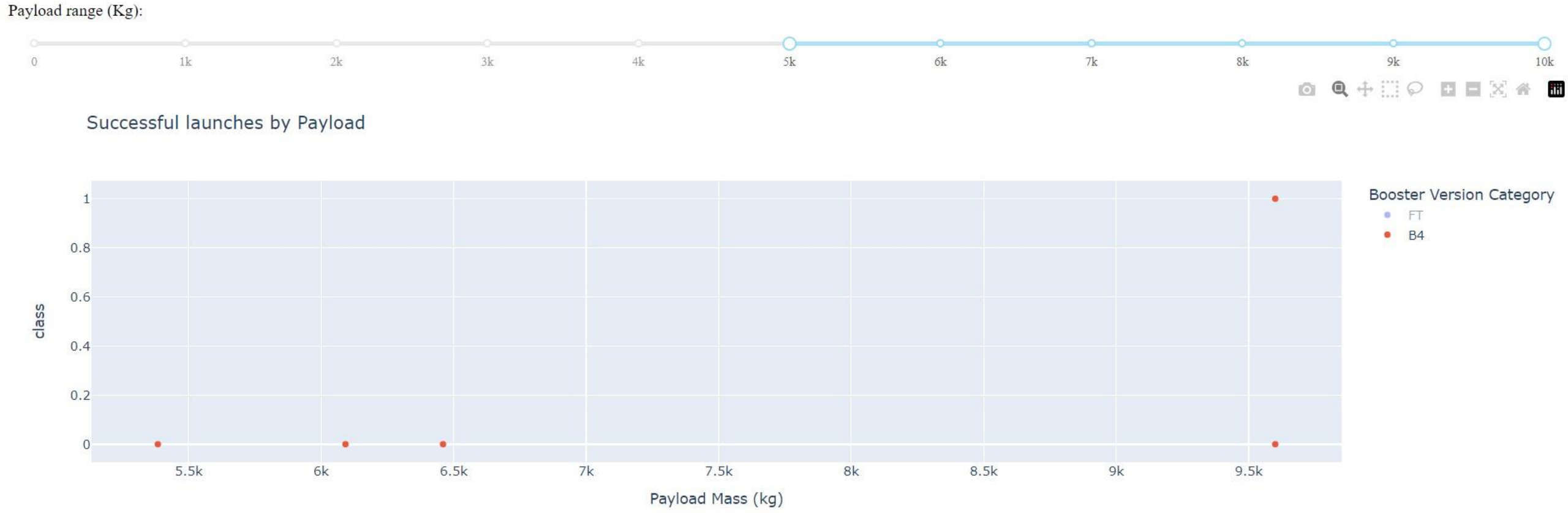
Large Payloads and Launch Outcomes with FT Booster only



Large Payloads and Launch Outcomes with FT Booster only

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes of the launches. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- To better understand the role of payloads in landing outcome successes of FT Booster, only launches with payloads greater than 5K for FT Booster were plotted on the graph.
- The success rate of landing outcomes with FT Booster drops substantially for payloads greater than 5K. On observation, it appears to be around 40% which is only half of its previous success rate.

Large Payloads and Launch Outcomes with B4 Booster only



Large Payloads and Launch Outcomes with B4 Booster only

- The dashboard displays a scatter chart where payloads are plotted against the success and failure of the landing outcomes of the launches. The range of payloads to be considered can be selected through the payload slider shown above the graph.
- In order to determine best booster for high payloads, launches with B4 booster only, for payloads greater than 5K were plotted. B4 is the second most successful booster after FT as determined previously.
- The results indicate that launches with B4 booster for high payloads are not very successful either. On observation, the success rate is about 20% which is about half of FT Booster.
- FT Booster is the best booster overall for all payloads, high and low.

Section 5

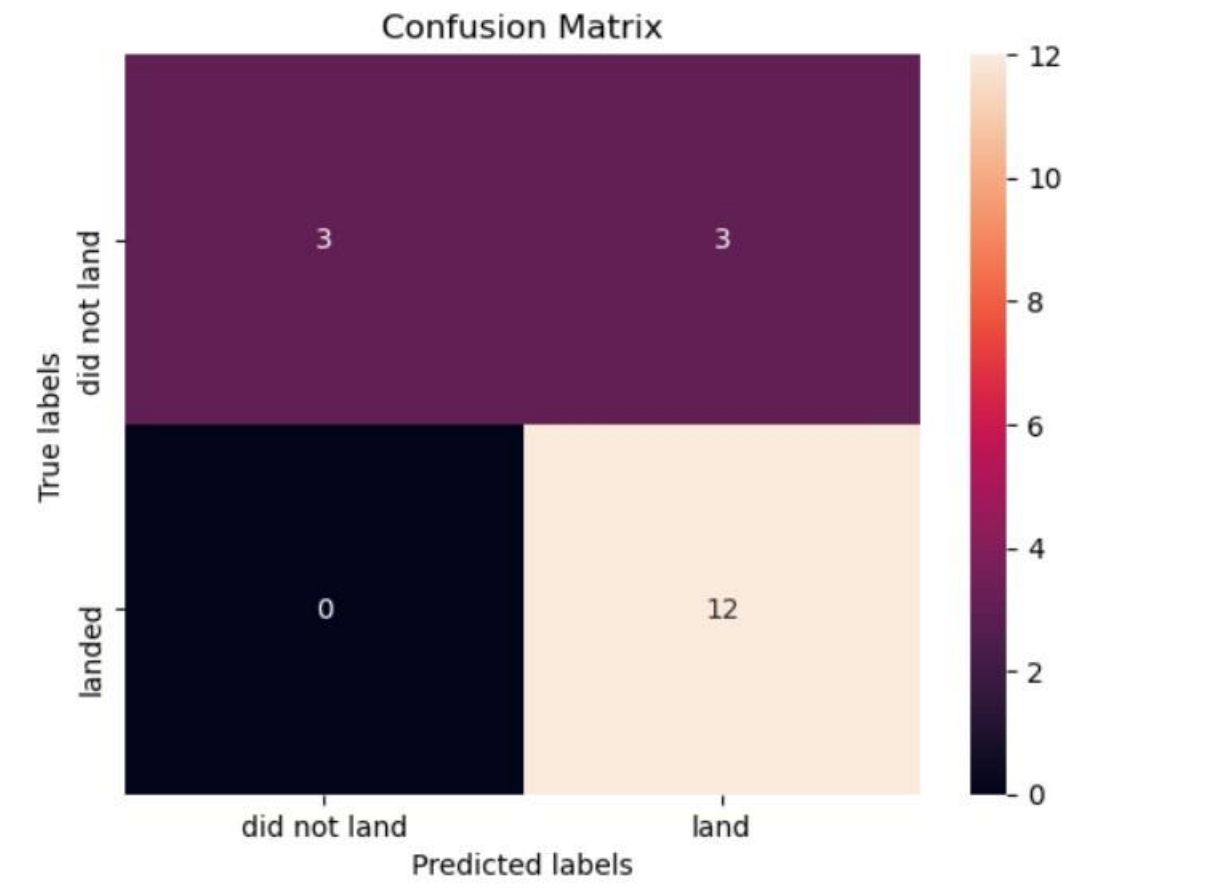
Predictive Analysis (Classification)

Classification Accuracy

Classification Accuracy

- The models used for classification of landing outcomes of rocket launches were Logistic Regression, Support Vector Machine, Decision Tree and K-Nearest Neighbor.
- The classification accuracy of all the models used for classifying landing outcomes of rocket launches was recorded in a bar chart against the names of each of the model.
- The bar chart shows clearly that all the models had approximately the same amount of accuracy in classifying rocket launch landing outcomes correctly. The accuracy was around 83.33%.
- This was due to small number of samples in testing data.

Confusion Matrix

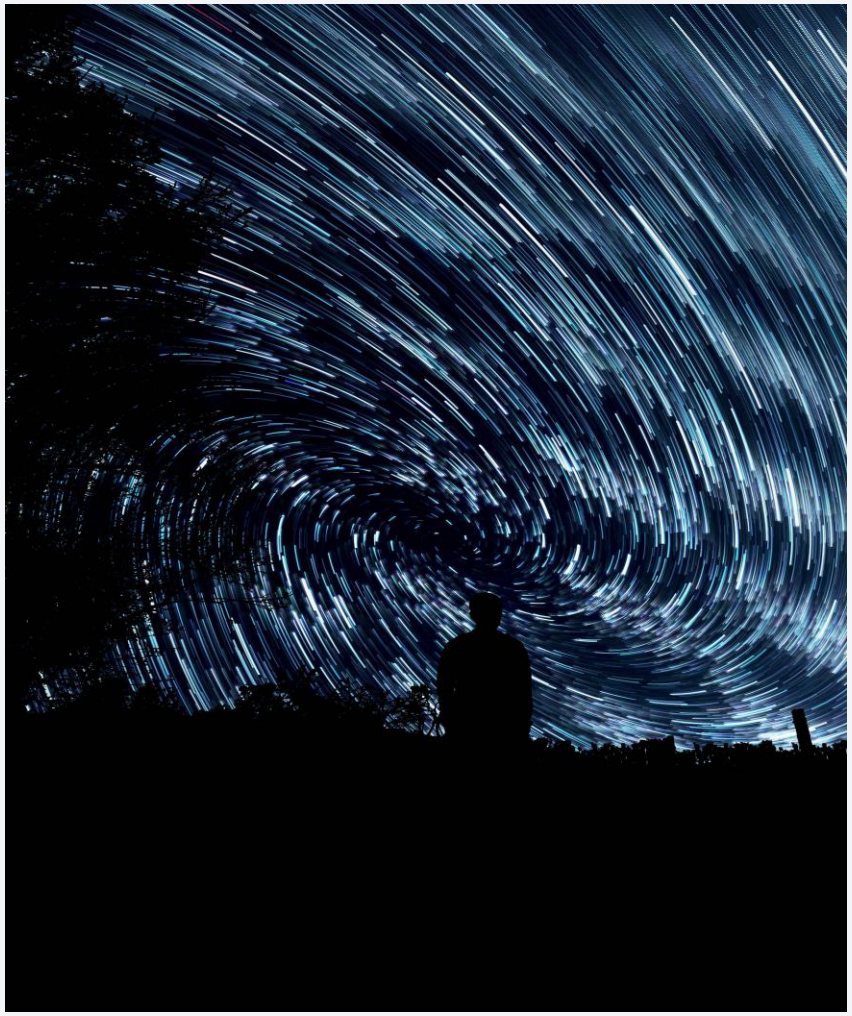


The confusion matrix for each of the models used for classification gave similar results. From the matrix, it can be clearly seen that all the classification models managed to successfully identify landed outcomes as landed outcomes.
However, all the models had trouble identifying did not land outcomes. They misclassified half the non-landed outcomes as landed outcomes. In other words, the primary issue they faced had to do with False Positive results, that is negative results being “falsely” identified as “positive” results.

Conclusions

The project made use of publicly available data on SpaceX launches to determine if a space launch will succeed in landing the first stage of the launch successfully or not. The first stage of a launch is very expensive but if landed successfully can be re-used for other launches. Thus, if it can be determined beforehand whether a launch will successfully land the first stage or not, then the cost of the launch becomes known prior to attempting it which can be very significant.

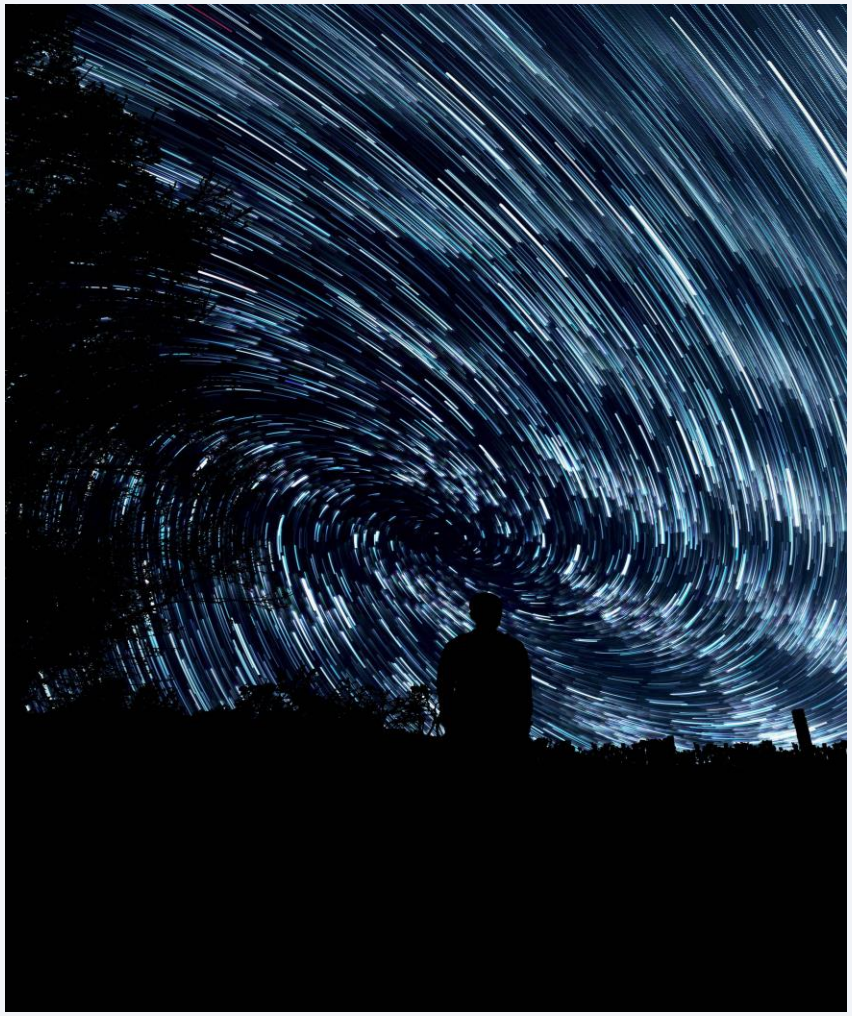
The project made use of Machine Learning techniques to determine the success of a landing outcome for any given launch. It used multiple Classification techniques for this purpose: Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree Classifier and K-Nearest Neighbors Algorithm (KNN). All 4 techniques were applied on the training



Conclusions

dataset with selected features such as orbit, payload mass, booster version and launch site and the resulting Classification Models (CM) were used on the testing data to determine their accuracy in classifying landing outcomes for launches.

The resulting CMs all managed to accurately classify 83.33% of the test data's landing outcomes correctly. The coincidence appeared odd so a confusion matrix for each CM was plotted. It turned out that every CM had gotten the same results on the confusion matrix. This was because the testing data was too small. Furthermore, the confusion matrix revealed that the CMs had trouble with False Positives i.e. Negative landing outcomes that were incorrectly identified as Positive. Given that space launches are still occurring, additional data can be added to the models in the future which could improve model results.



Conclusions

A summary of the conclusions for the project is as follows:

- 4 Machine Learning Classification techniques gave an accuracy of 83.33% while classifying landing outcomes for launches.
- False Positives emerged as the biggest error for classification models.
- The data for the labels set is too small which is why all 4 techniques had the same accuracy.
- As launches continue more data may be added to the models for better results.
- Some important features determined through EDA were Orbit, Booster Version, Flight Number, Payload Mass and Launch Site.

Appendix

Github URL: https://github.com/AmaltaasKhabti/Applied-Data-Science-Project

Thank you!