SpaceX Data Science

Bertan Pank

05/08/2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Summary of methodologies
- Data Collection via API, Web Scraping
- Exploratory Data Analysis (EDA) with Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboards with Plotly Dash
- Predictive Analysis

- Summary of all results
- Exploratory Data Analysis results
- Interactive maps and dashboard
- Predictive results

INTRODUCTION

- Project background and context
- The aim of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9 rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The price difference is explained by the fact that SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch.

- Problems you want to find answers
- What are the main characteristics of a successful or failed landing?
- What are the effects of each relationship of the rocket variables on the success or failure of a landing?
- What are the conditions which will allow SpaceX to achieve the best landing success rate?

METHODOLOGY

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- 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 METHODOLOGY

- Datasets are collected with Rest SpaceX API and webscrapping from Wikipedia
  - The information obtained by the API are rocket, launches, payload information.
    - The Space X REST API URL is api.spacexdata.com/v4/
- The information obtained by the webscrapping from Wikipedia are launches, landing, payload information.
  - URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

DATA WRANGLING METHODOLOGY

- In the dataset, there are several cases where the booster did not land successully.
  - True Ocean, True RTLS, True ASDS means the mission has been successful.
  - False Ocean, False RTLS, False ASDS and Nones mean that the mission was a failure.
- We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

EDA WITH DATA VISUALIZATION METHODOLOGY

Scatter Graphs

Flight Number vs. Payload Mass

Flight Number vs. Launch Site

Payload vs. Launch Site

Orbit vs. Flight Number

Payload vs. Orbit Type

Orbit vs. Payload Mass

Scatter plots show relationship between variables. This relationship is called the correlation

Bar Graph

Success rate vs. Orbit

Bar graphs show the relationship between numeric and categoric variables

Line Graph

Success rate vs. Year

Line graphs show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.

EDA WITH SQL METHODOLOGY

- We performed SQL queries to gather and understand data from dataset:
- Displaying the names of the unique lauunch 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.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, faiilure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

BUILDING AN INTERACTIVE MAP WITH FOLIUM

Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).

Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker, folium.features.DivIcon).

The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).

Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing. (folium.map.Marker, folium.Icon).

Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them. (folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

These objects are created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.

BUILD A DASHBOARD WITH PLOTLY DASH

Dashboard has dropdown, pie chart, rangeslider and scatter plot components

Dropdown allows a user to choose the launch site or all launch sites (dash_core_components.Dropdown).

Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component (plotly.express.pie).

Rangeslider allows a user to select a payload mass in a fixed range (dash_core_components.RangeSlider).

Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass (plotly.express.scatter)

PREDICTIVE ANALYSIS

Data preparation

Load dataset

Normalize data

Split data into training and test sets.

Model preparation

Selection of machine learning algorithms

Set parameters for each algorithm to GridSearchCV

Training GridSearchModel models with training dataset

Model evaluation

Get best hyperparameters for each type of model

Compute accuracy for each model with test dataset

Plot Confusion Matrix

Model comparison

Comparison of models according to their accuracy

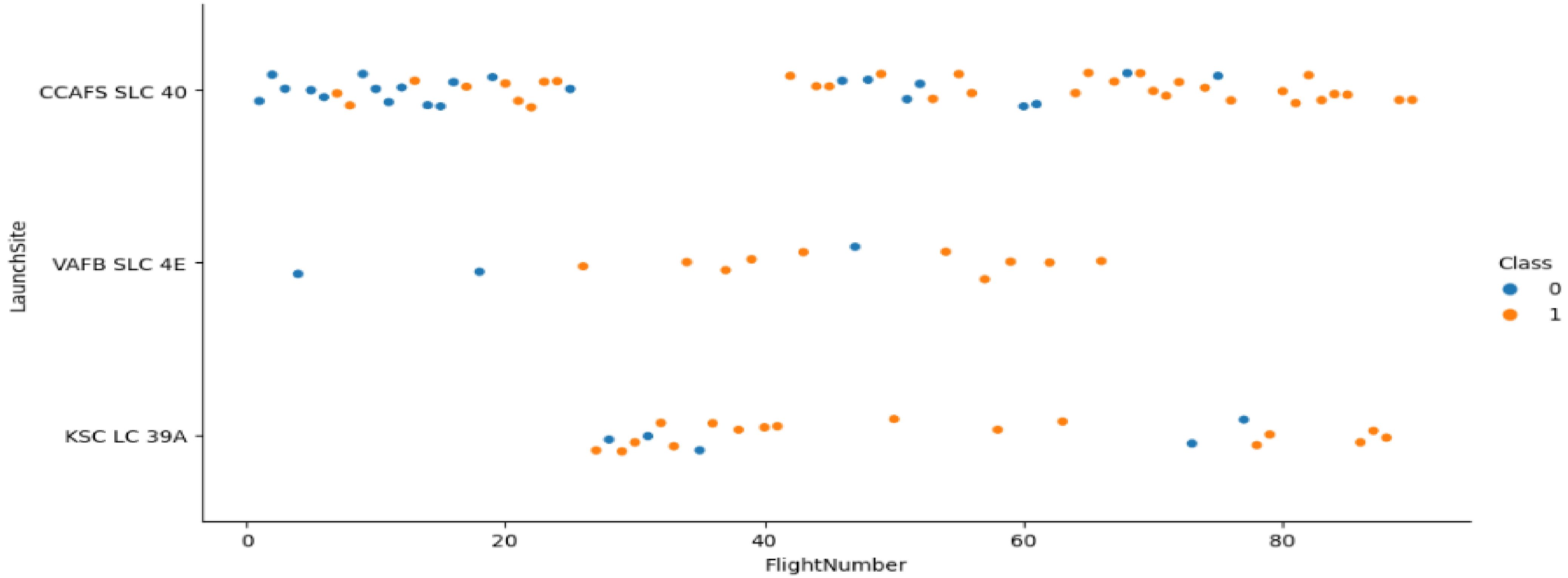
The model with the best accuracy will be chosen

RESULTS

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

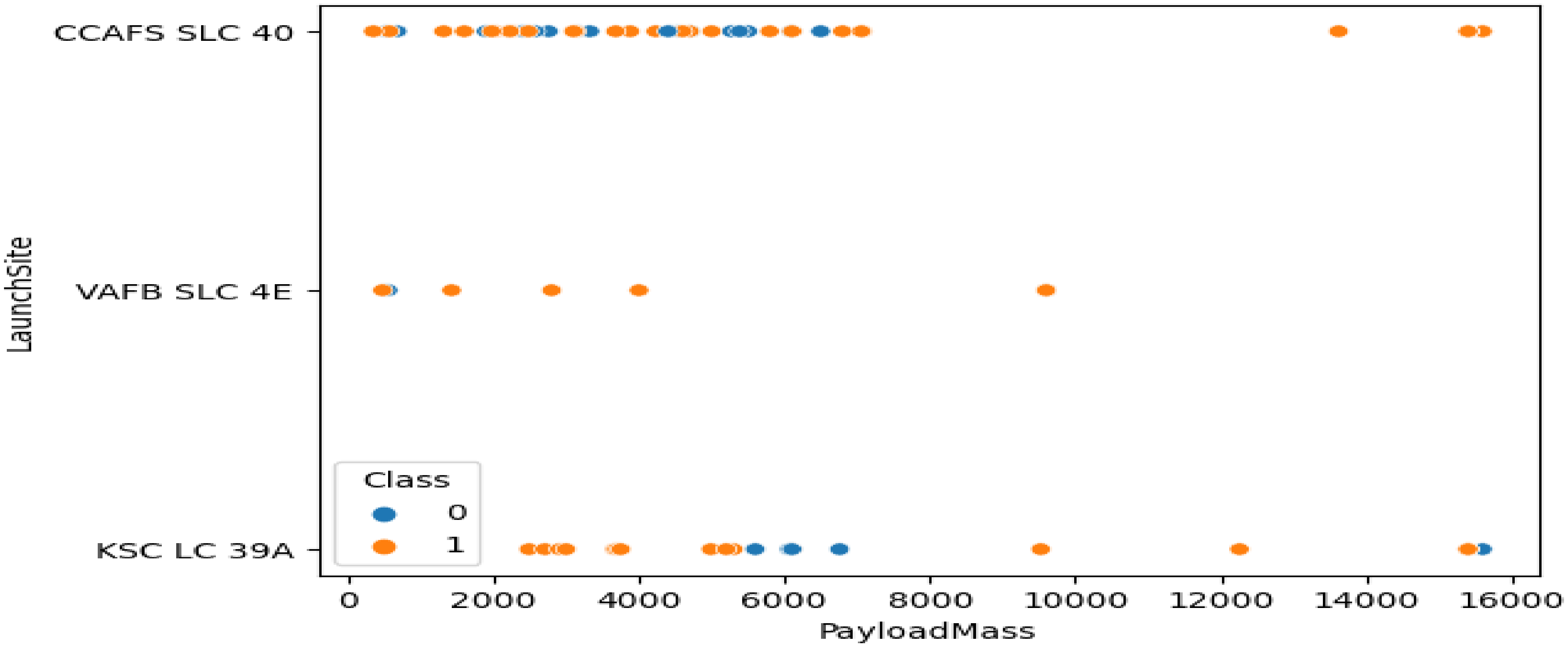
EDA WITH VISUALIZATION

Flight Number vs. Launch Site



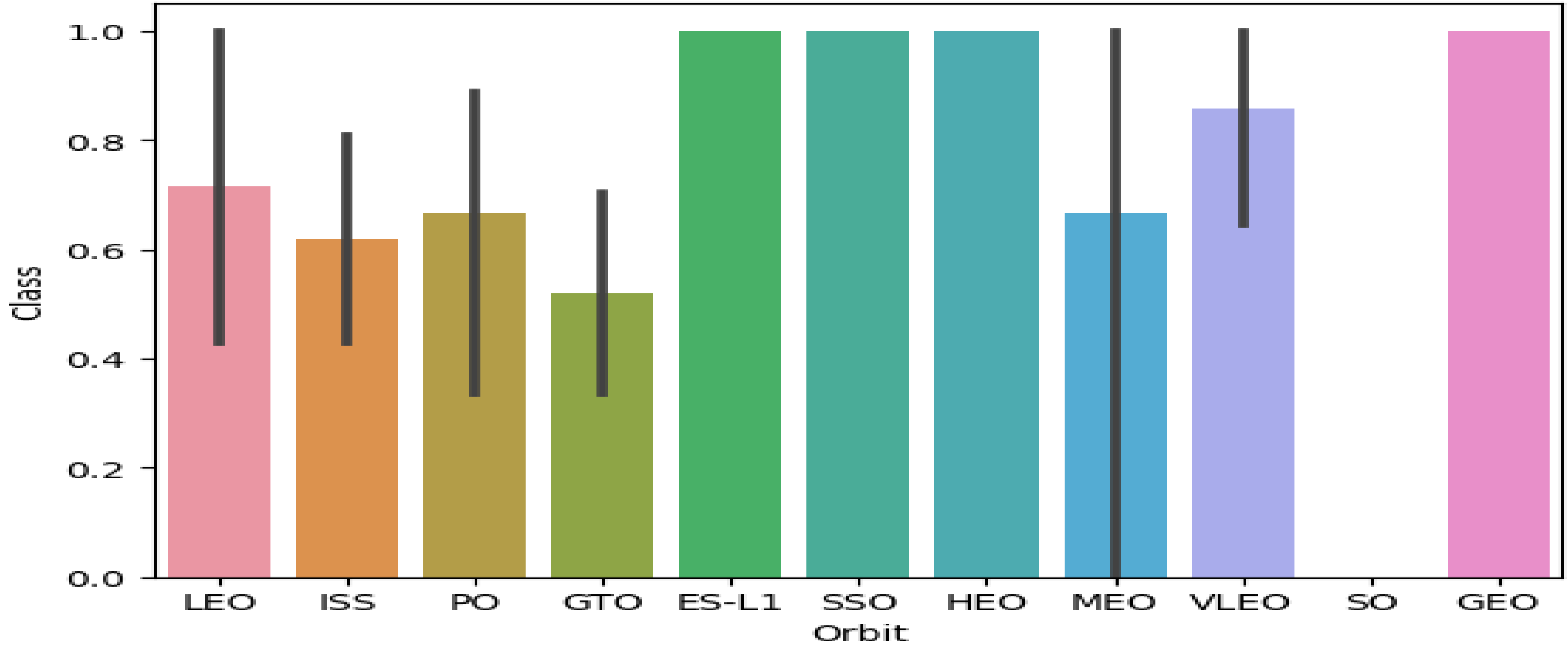
EDA WITH VISUALIZATION

Payload vs. Launch Site



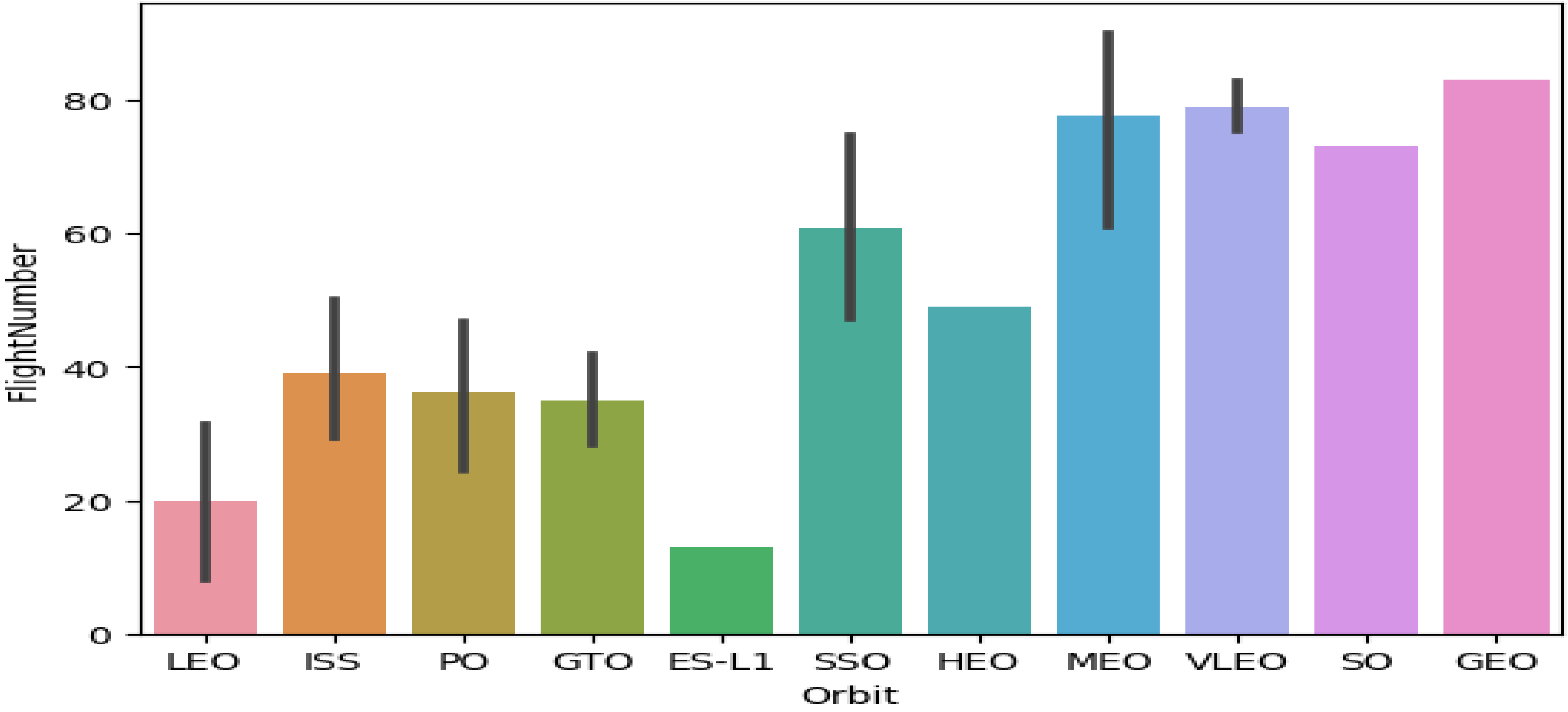
EDA WITH VISUALIZATION

Orbit vs. Success Rate



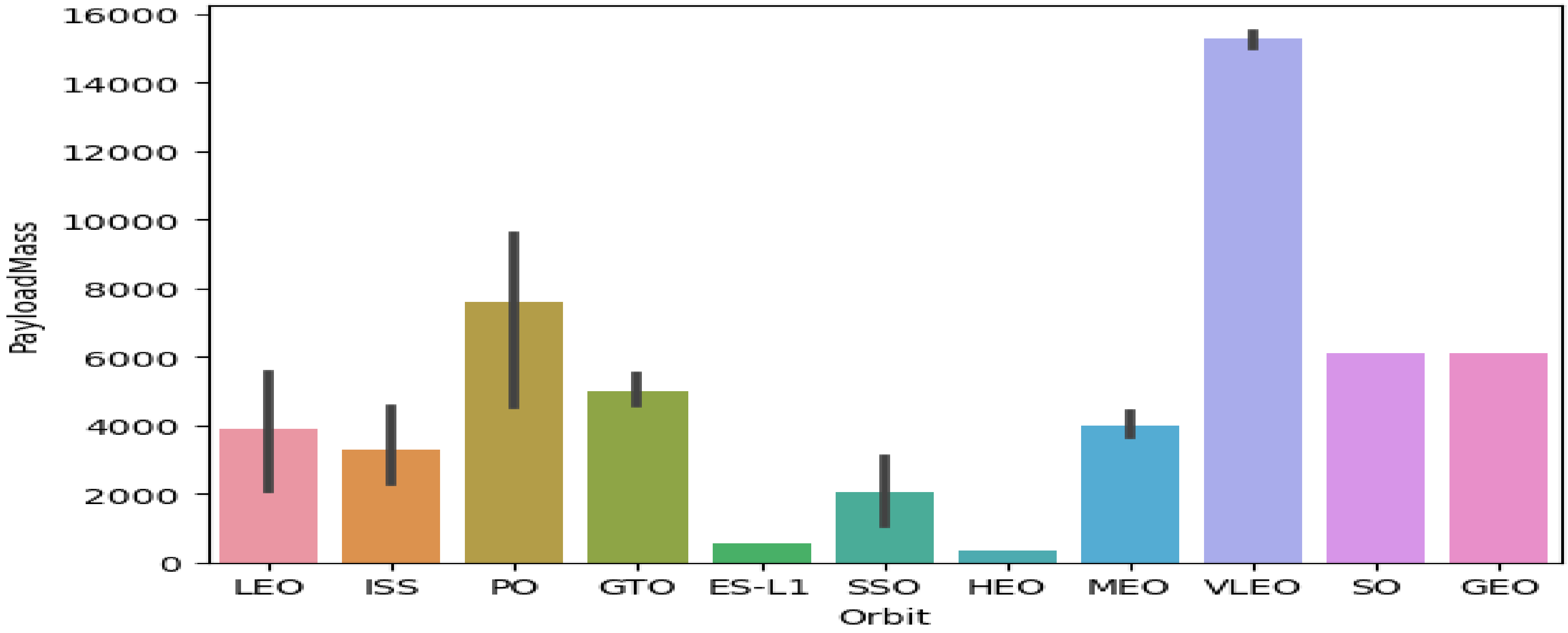
EDA WITH VISUALIZATION

Flight Number vs. Orbit



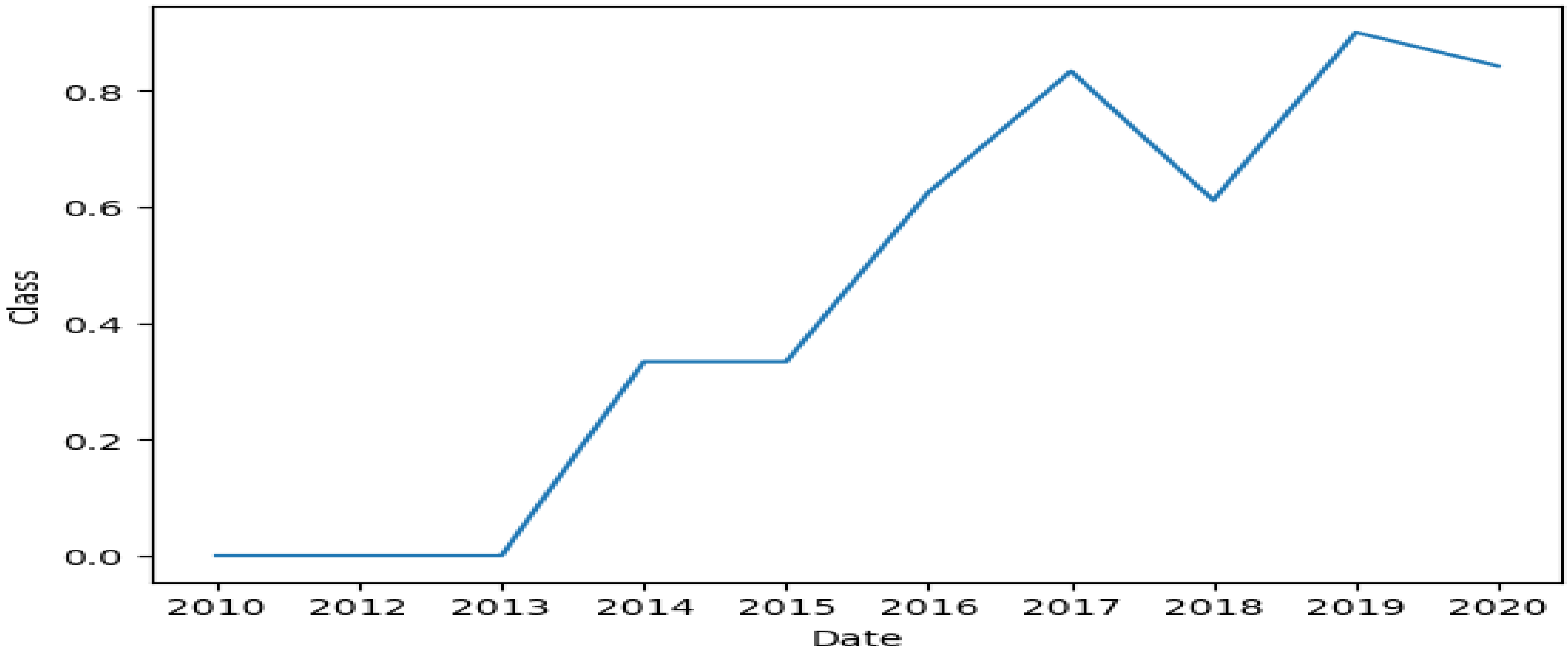
EDA WITH VISUALIZATION

Payload vs. Orbit



EDA WITH VISUALIZATION

Launch Success Yearly Trend



EDA WITH SQL

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

```
%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

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_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

EDA WITH SQL

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

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

```
%sql SELECT TOTAL("PAYLOAD_MASS__KG_") as "Total_Payload_Mass_CRS" FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)';
```

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

Total_Payload_Mass_CRS
45596.0

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

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

AVG_MASS
2928.4

EDA WITH SQL

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

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

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

First_Successful_Ground_Pad_Landing_Date
2015-12-22

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_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

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

EDA WITH SQL

List the total number of successful and failure mission outcomes

```
%%sql SELECT "Mission_Outcome", COUNT(*) as "Total_Count"
FROM SPACEXTABLE
GROUP BY "Mission_Outcome";
```

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

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

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

```
%%sql SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "PAYLOAD_MASS__KG_" = (
    SELECT MAX("PAYLOAD_MASS__KG_")
    FROM SPACEXTABLE
);
```

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

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

EDA WITH SQL

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.

```
%%sql SELECT
    CASE
        WHEN substr("Date", 6, 2) = '01' THEN 'January'
        WHEN substr("Date", 6, 2) = '02' THEN 'February'
        WHEN substr("Date", 6, 2) = '03' THEN 'March'
        WHEN substr("Date", 6, 2) = '04' THEN 'April'
        WHEN substr("Date", 6, 2) = '05' THEN 'May'
        WHEN substr("Date", 6, 2) = '06' THEN 'June'
        WHEN substr("Date", 6, 2) = '07' THEN 'July'
        WHEN substr("Date", 6, 2) = '08' THEN 'August'
        WHEN substr("Date", 6, 2) = '09' THEN 'September'
        WHEN substr("Date", 6, 2) = '10' THEN 'October'
        WHEN substr("Date", 6, 2) = '11' THEN 'November'
        WHEN substr("Date", 6, 2) = '12' THEN 'December'
    END AS "Month",
    "Landing_Outcome",
    "Booster_Version",
    "Launch_Site"
FROM SPACEXTABLE
WHERE substr("Date", 1, 4) = '2015'
AND "Landing_Outcome" = 'Failure (drone ship)';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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.

```
%%sql SELECT "Landing_Outcome", COUNT(*) AS "Count"
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY COUNT(*) DESC;
```

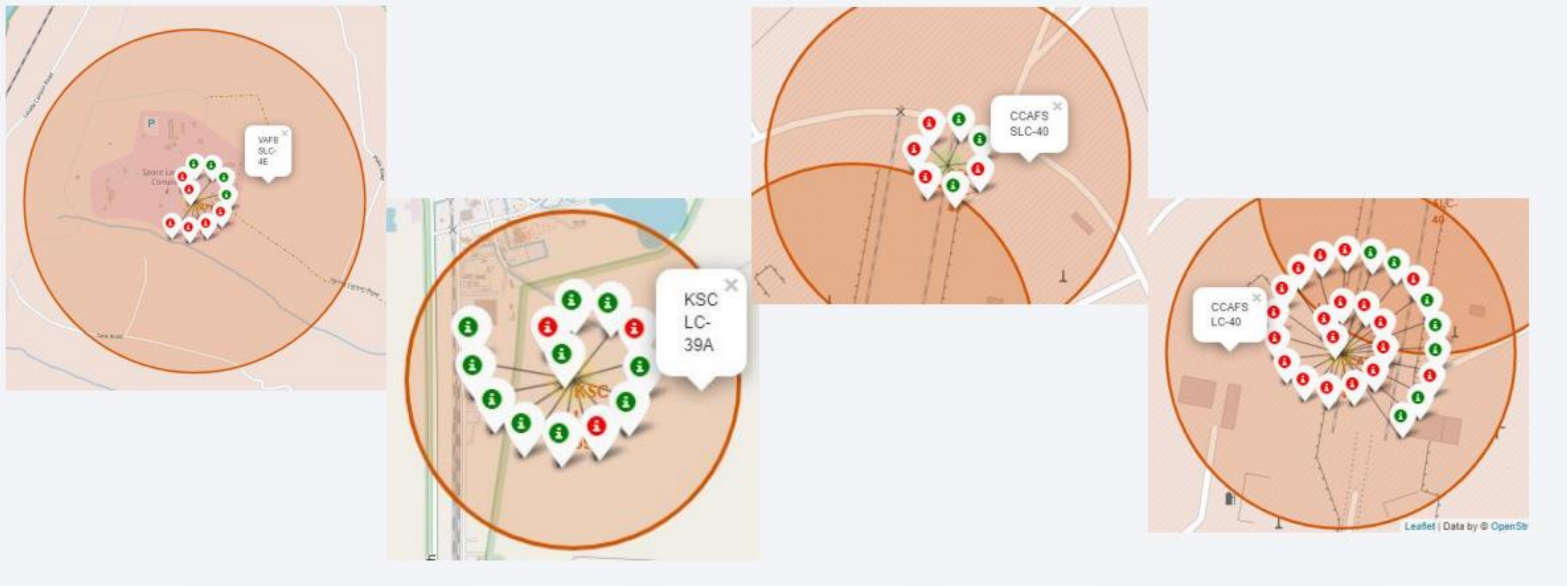
```
* sqlite:///my_data1.db
Done.
```

Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

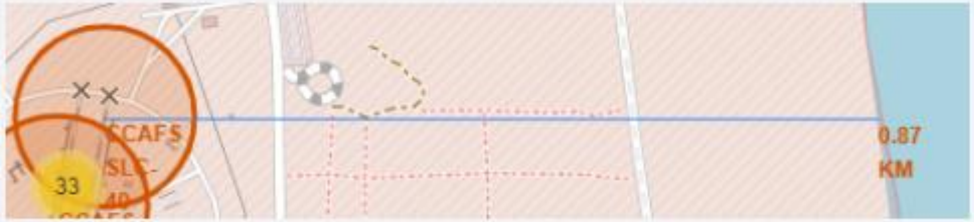
INTERACTIVE MAP WITH FOLIUM

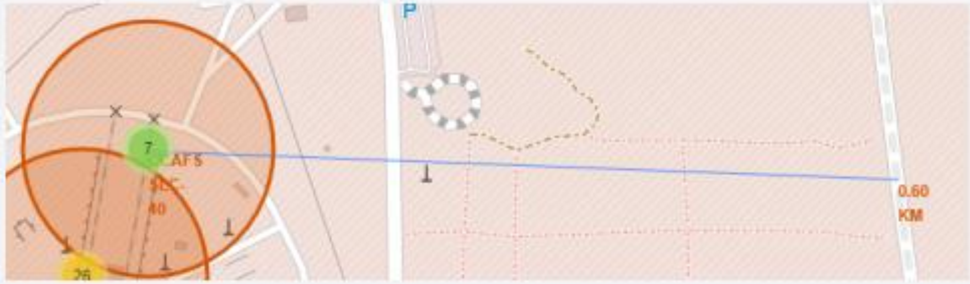
Ground Stations

COLORED MARKERS



DISTANCES BETWEEN CCAFS SLC-40 AND ITS PROXIMITIES

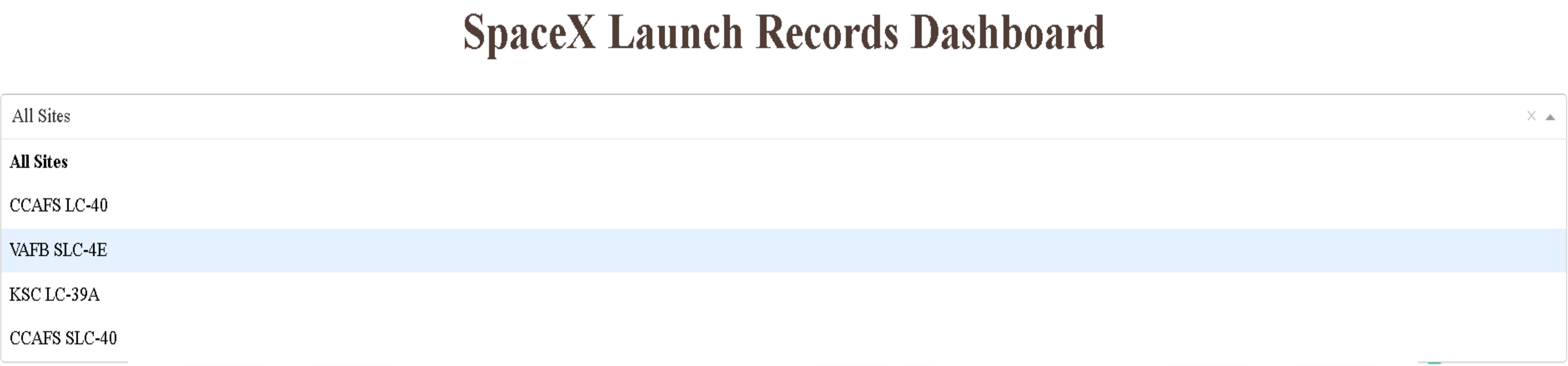




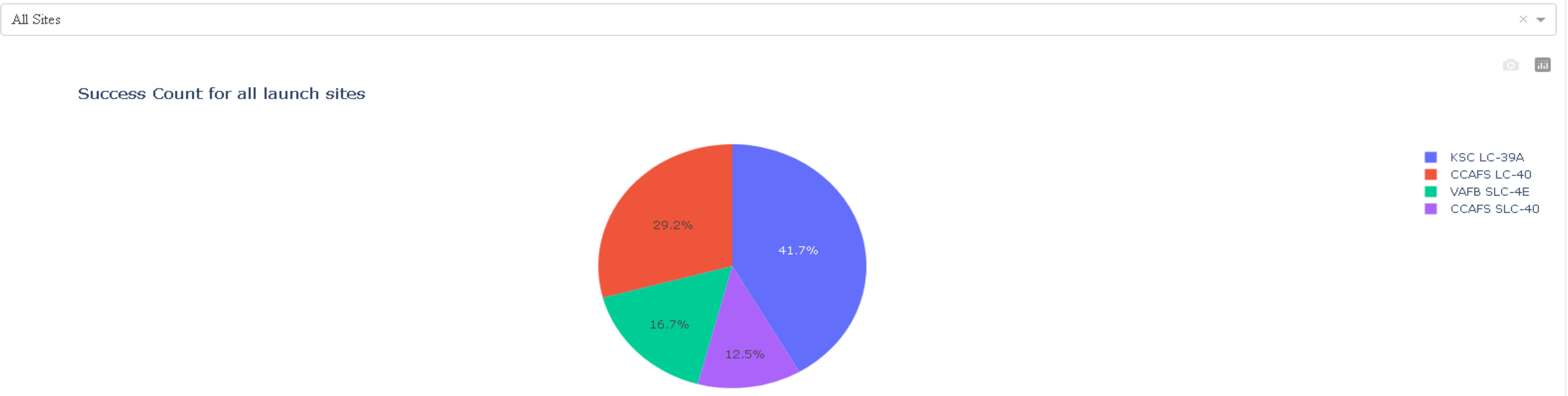
ANSWER OF QUESTIONS

- Is CCAFS SLC-40 in close proximity to railways ? Yes
- Is CCAFS SLC-40 in close proximity to highways ? Yes
- Is CCAFS SLC-40 in close proximity to coastline ? Yes
- Do CCAFS SLC-40 keeps certain distance away from cities ? No

PLOTLY DASHBOARD – OPTIONS

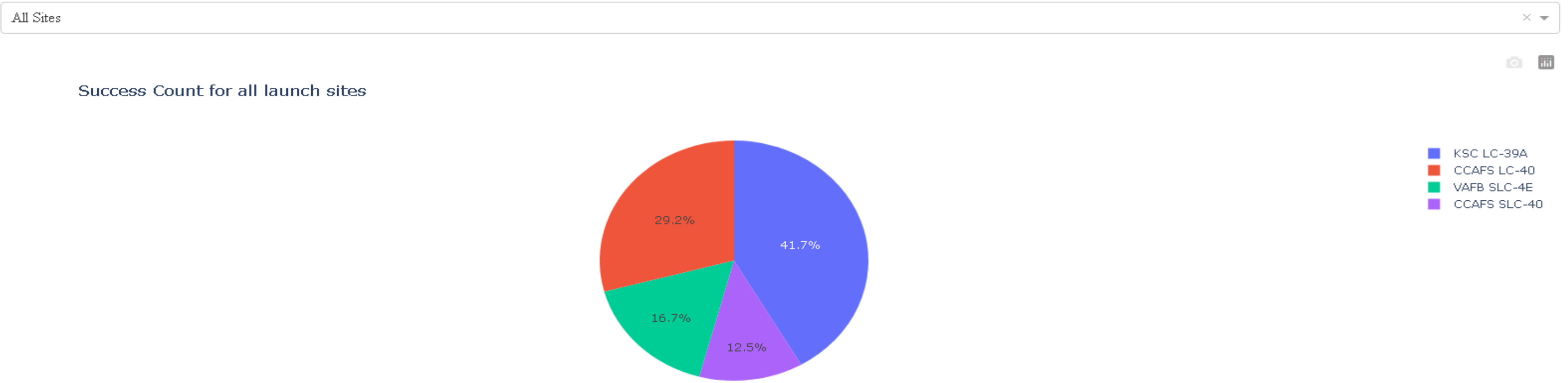


PLOTLY DASHBOARD – TOTAL SUCCESS BY SITE



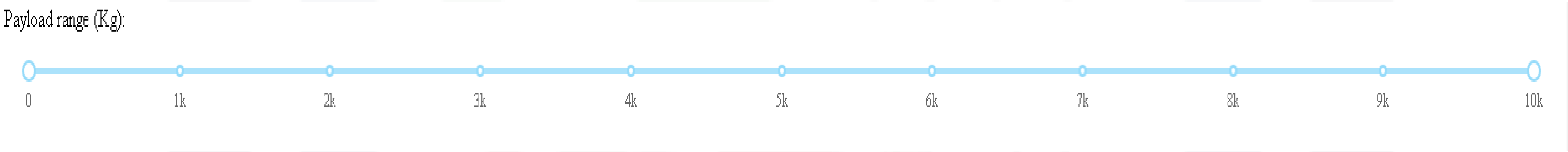
KSC LC-39A has the largest successfull launch amount.

PLOTLY DASHBOARD – SUCCESS RATE

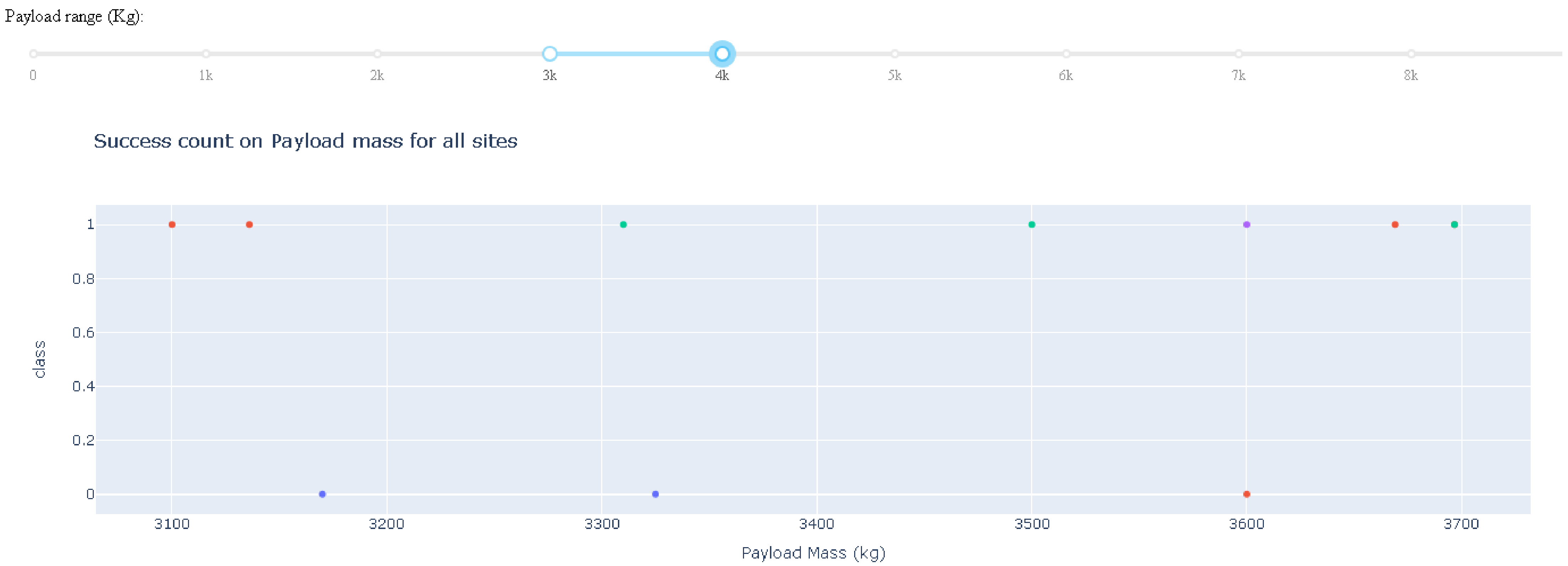


KSC LC-39A has the best lauch success rate.

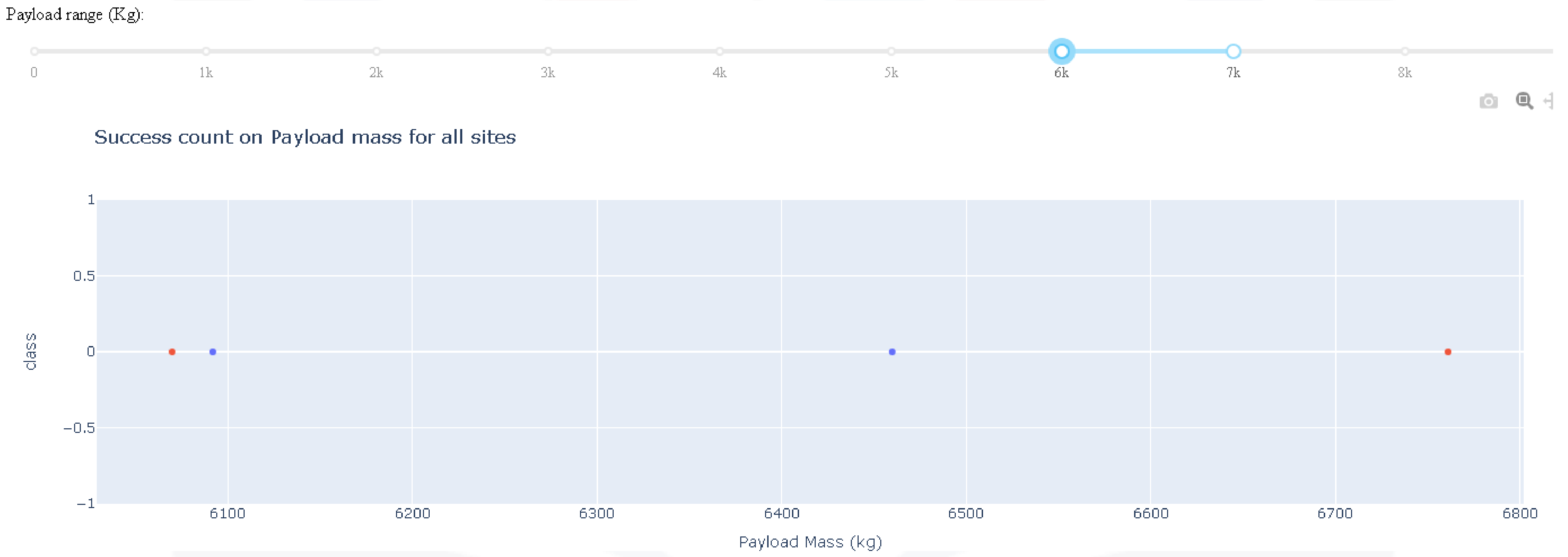
PLOTLY DASHBOARD – RANGE OPTION BUTTON



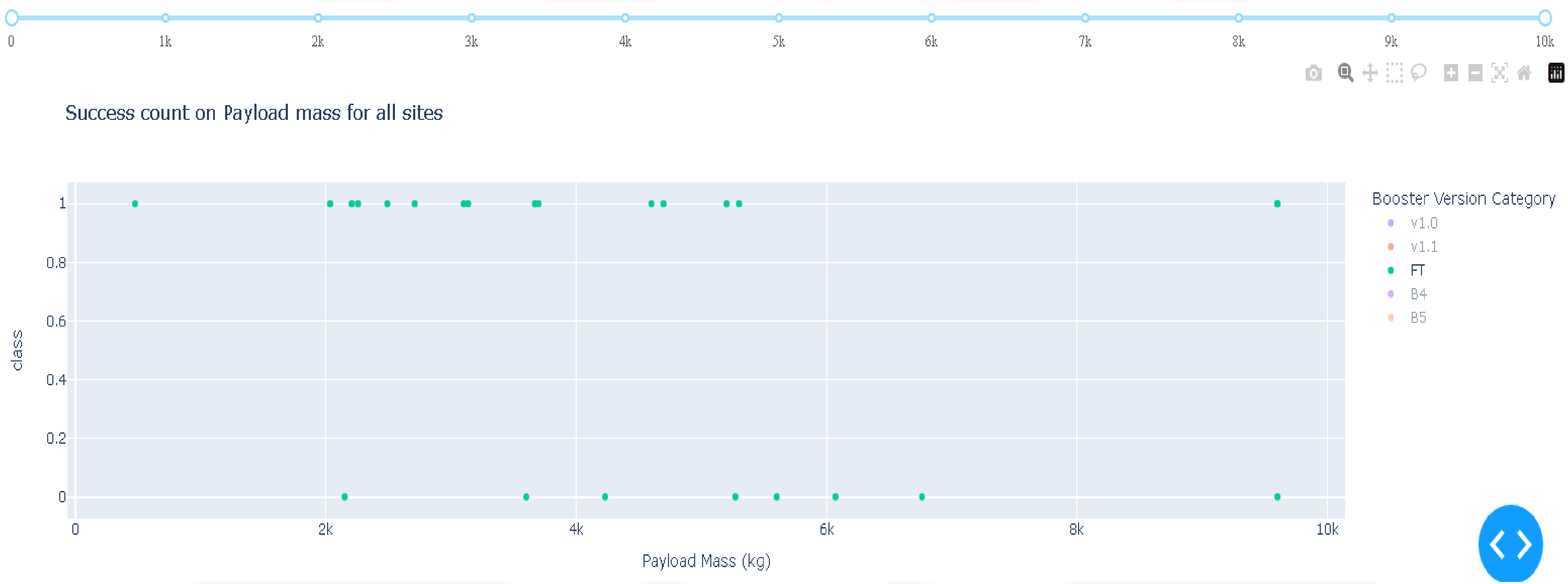
PLOTLY DASHBOARD – BEST PAYLOAD RANGE



PLOTLY DASHBOARD – WORSE PAYLOAD RANGE

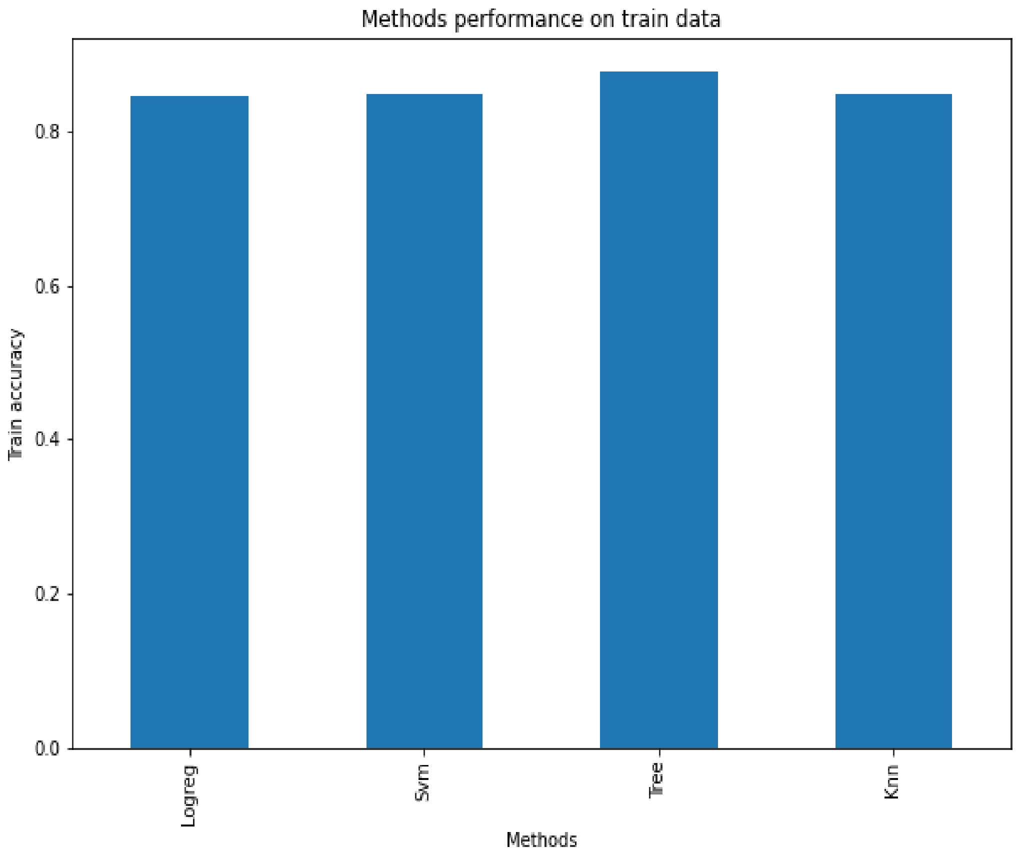


PLOTLY DASHBOARD – BEST BOOSTER



PREDICTIVE ANALYSIS - ACCURACIES

	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333



PREDICTIVE ANALYSIS – CONFUSION MATRIX

CONCLUSION

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

THANK YOU!