IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nir Perelshtein
07.02.24

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization (Pandas & Matplotlib)
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly Dash
  - ML Prediction
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive Visual Analytics & Dashboard screenshots
  - ML Prediction results

Introduction

- Project background and context

  SpaceX was founded in 2002 by Elon Musk with one of the goals to reducing the cost of space transportation.

  SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

  We will predict if the Falcon 9 first stage will land successfully based on ML and analytical approaches. This information will help us to identifying the price to bid against SpaceX for a rocket launch.

- The problems we want to find answers are:

  - The explanatories variables that has impact on the chance of first stage successfully landing.
  - The correlation between variables and the chance of success of first stage landing.
  - What conditions must be met to improve the chance of a successful landing.

Section 1

Methodology

Methodology

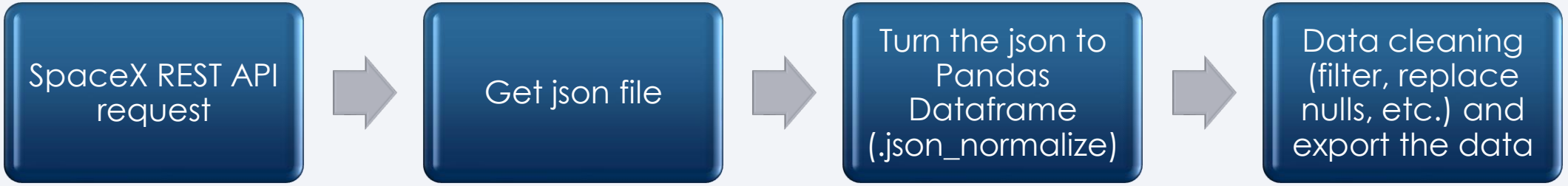
Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Determining the analysis population
  - Dealing with missing values and determining how to deal with them
  - Using one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build, tune, evaluate classification models to get the most accurate result

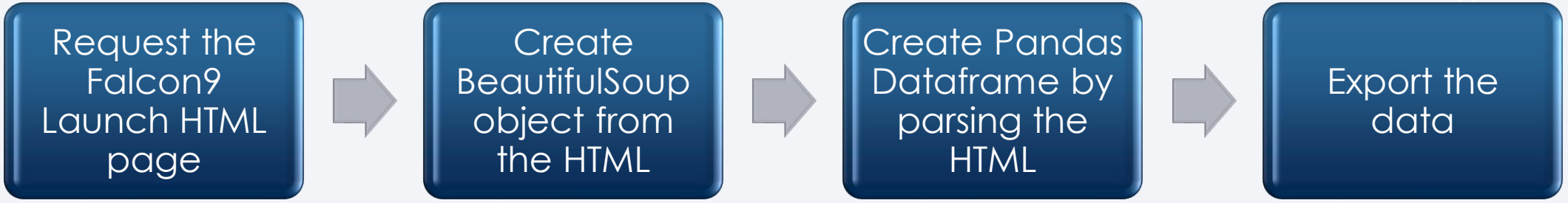
Data Collection

Data sets were collected by SpaceX REST API and web scrapping from a Wikipedia page.

SpaceX REST API - https://api.spacexdata.com/v4/



Web Scrapping- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches



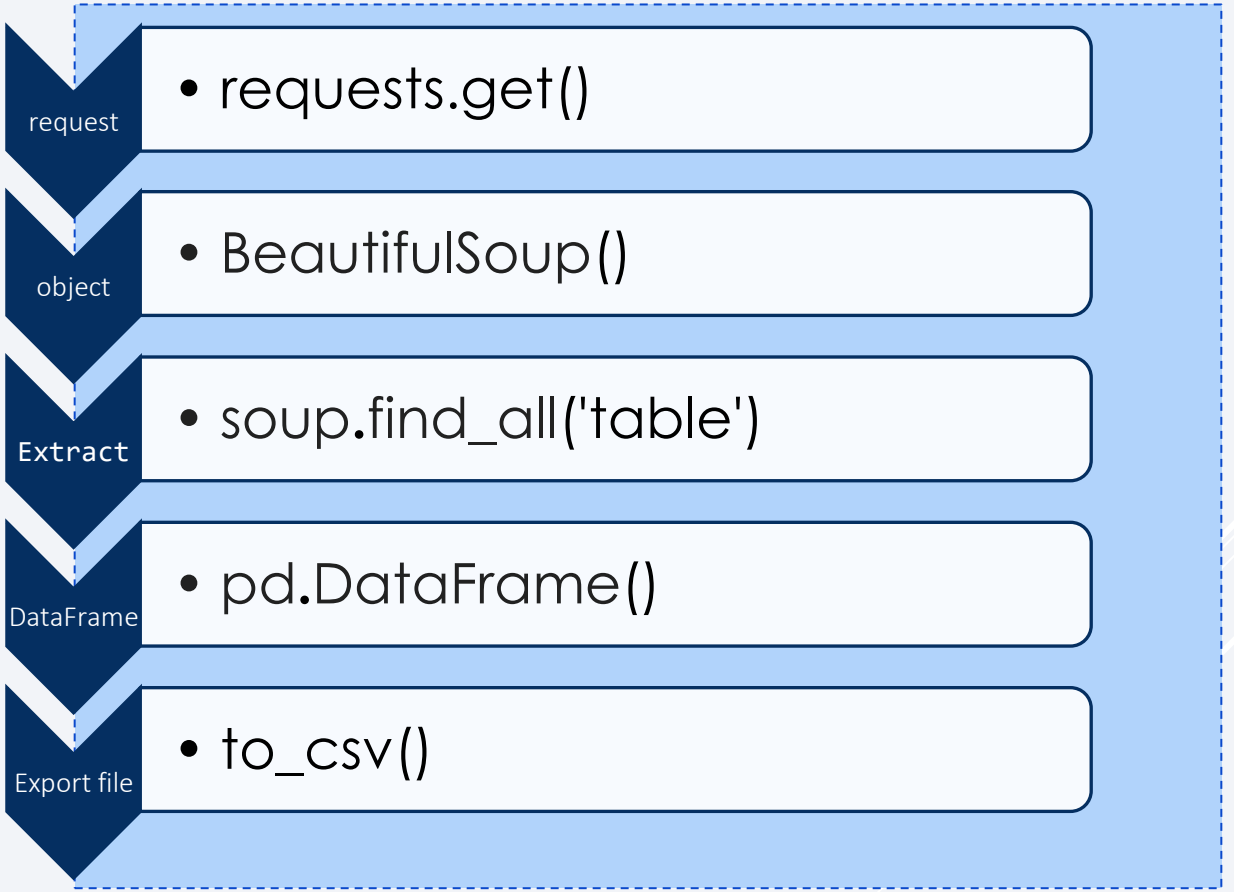
Data Collection – SpaceX API

- We used requests.get() method to get SpaceX data from the API.
- .json() to decode the response content as a Json file.
- .json_normalize() to turn it in Dataframe.
- We used df manipulation techniques to get only the relevant data for us and we handle the null data.
- We export the data to csv for future analysis.

GitHub URL: Click Here

Data Collection - Scraping

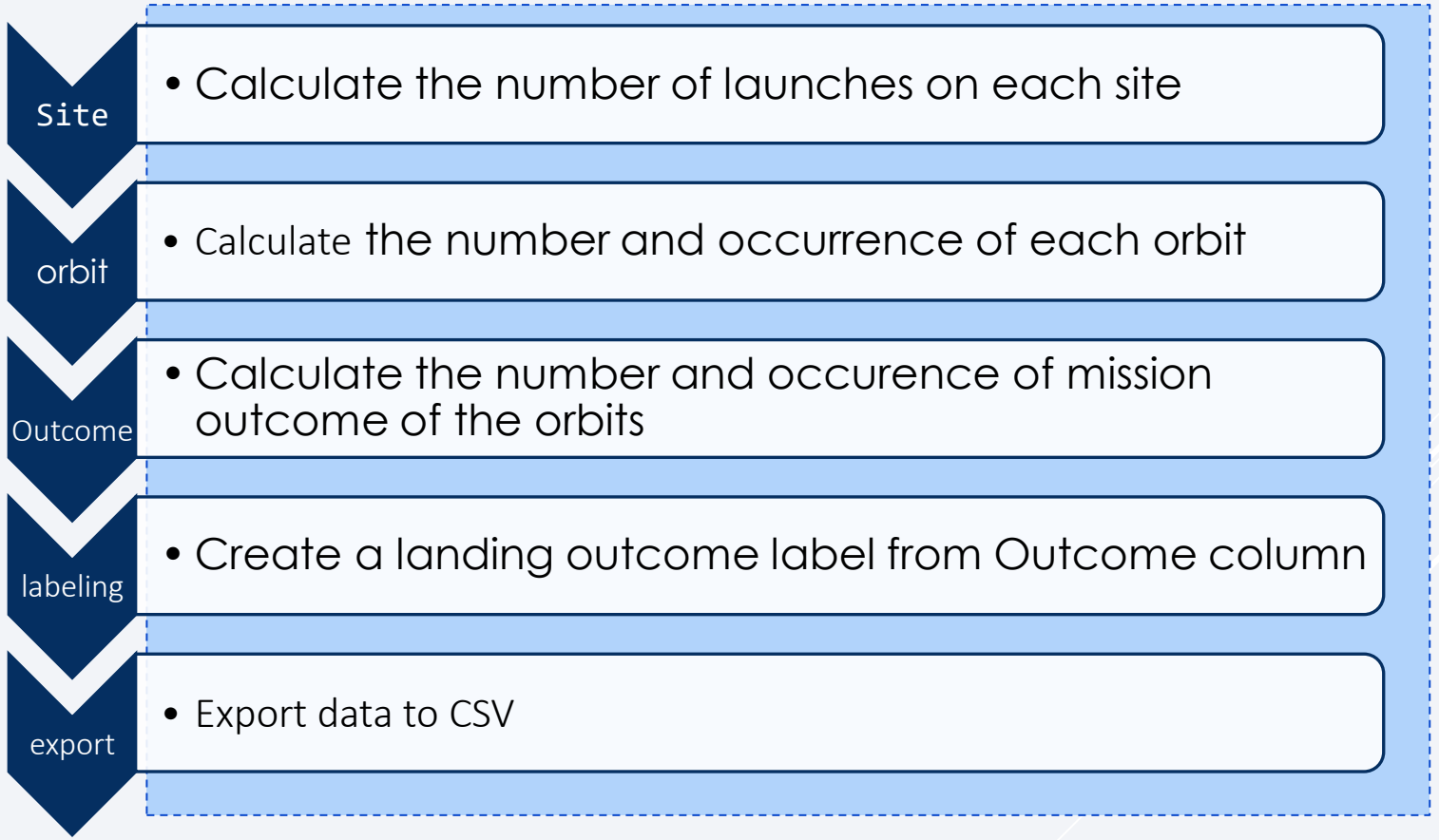
- We used requests.get() method to get Falcon9 Launch Wiki page.
- Create BeautifulSoup object from the HTML response
- Extract all columns from the HTML table
- Create Dataframe by parsing the launch HTML table
- We export the data to csv for future analysis.



GitHub URL: Click Here

Data Wrangling

We performed exploratory Data Analysis (EDA) for our classification model and labeled our landing outcome to 0 if the landing failed, or 1 if the landing succeeded and added the calculate column to our Dataframe.



GitHub URL: Click Here

EDA with Data Visualization

Scatter Plot

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Launch Site vs. Payload Mass
- Flight Number vs. Orbit type
- Payload Mass vs. Orbit type

Scatter plots present the relationship between two variables.

Bar chart

- Orbit type vs. Success rate

Bar charts present relationship between categorical and numeric variables.

Line chart

- Success rate over the years

Line charts present trends, relationships in data over time or other intervals.

GitHub URL: Click Here

EDA with SQL

The SQL queries we performed:

- Displaying the names of the unique launch sites in the space mission.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by boosters launched by NASA (CRS).
- Displaying average payload mass carried by booster version F9 v1.1.
- Listing the date when the first successful landing outcome in ground pad was achieved.
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listing the total number of successful and failure mission outcomes.
- Listing the names of the booster versions which have carried the maximum payload mass.
- Listing the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranking 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.

GitHub URL: Click Here

Build an Interactive Map with Folium

- First we created a folium map object with NASA Johnson Space Center at Houston, TX.
- We added highlighted circle area with text label for the launch sites.
- Marked the success(green)/ failed(red) launches for each site on the map.
- Calculated distances between a launch site to railways, highways, coastline and cities and then plotting lines for each of them.

We added those object in order to understand better how geographical locations of launch sites and proximity to selected areas affect our success of a landing and to get better visualization understanding about the company environment and how close the launch sites to those areas.

GitHub URL: Click Here

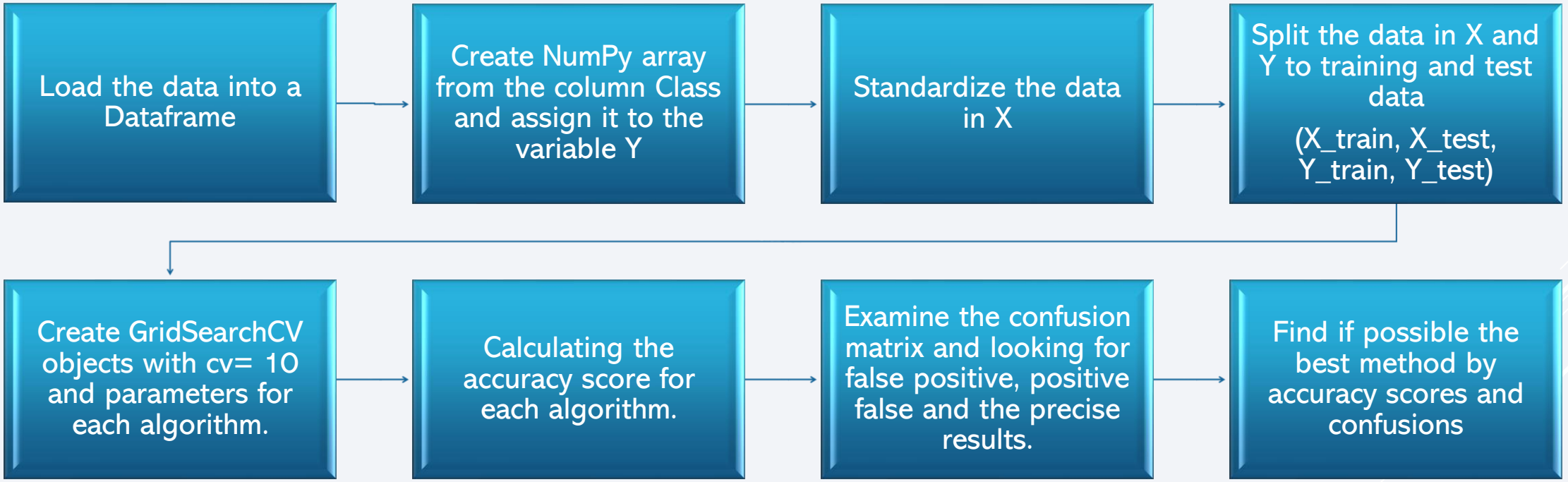
Build a Dashboard with Plotly Dash

We built a Dashboard with the following functionality:

- A dropdown with Launch site selection.
- A pie chart showing total success launches portion by site when we choose “All” in dropdown or success vs failed for specific launch site for visualizing launch success.
- Slider to select Payload Mass (0 to 10k with 1k steps)
- Scatter plot to visually observe how payload may be correlated with mission outcomes for selected site or all of them.

GitHub URL: Click Here

Predictive Analysis (Classification)



GitHub URL: Click Here

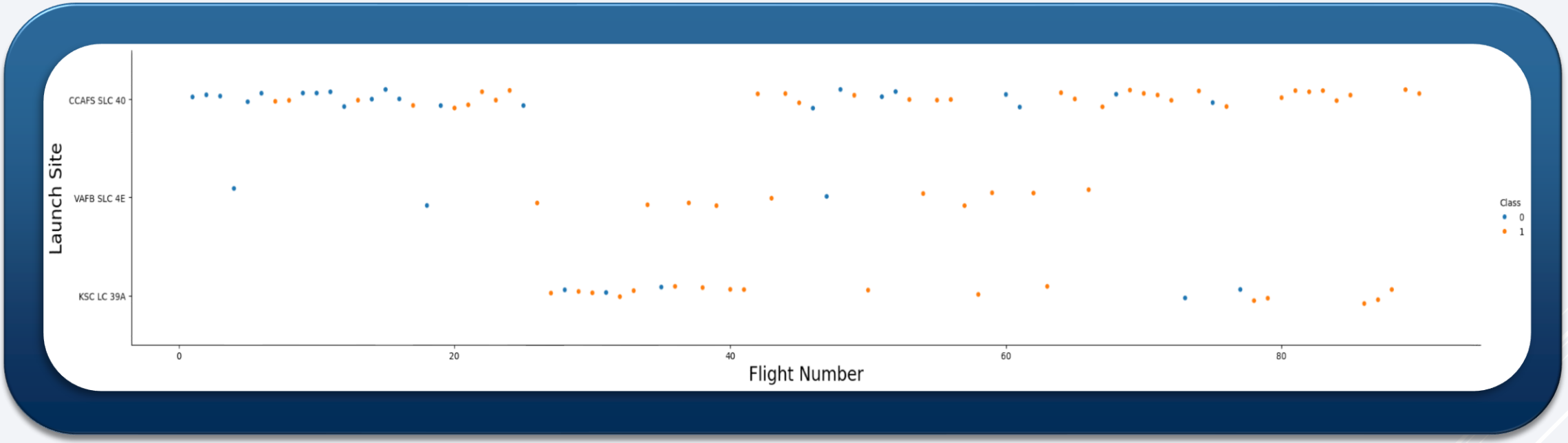
Results

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

Section 2

Insights drawn from EDA

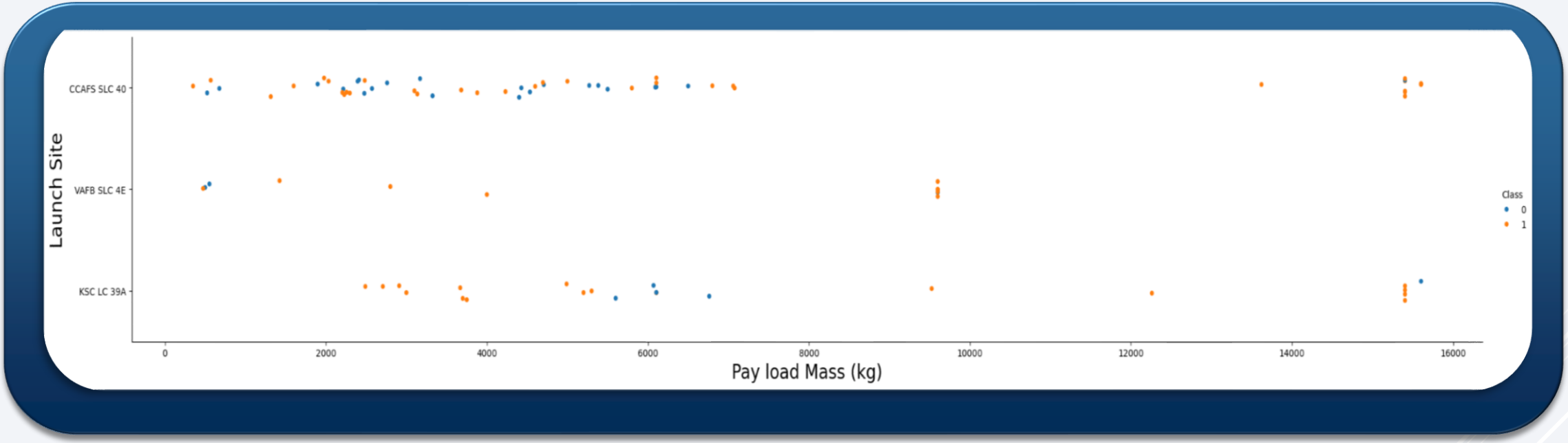
Flight Number vs. Launch Site



Explanation:

- The Scatter plot shows that when the flight number increasing the success rate increasing.
- CCAFS CLS-40 has the largest number of flights and the first launches start at this site.

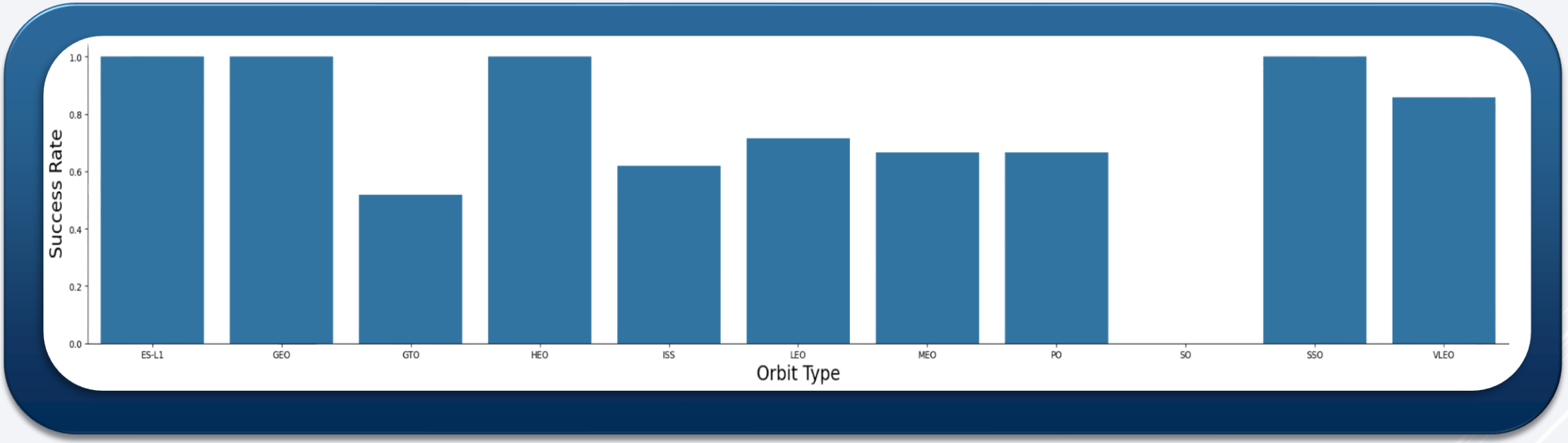
Payload vs. Launch Site



Explanation:

- The Scatter plot shows that when the Pay load Mass is higher the success rate increasing.
- VAFB SLC 4E has a maximum Pay load Mass below 10,000Kg, the other sites have a higher Pay load mass maximum.

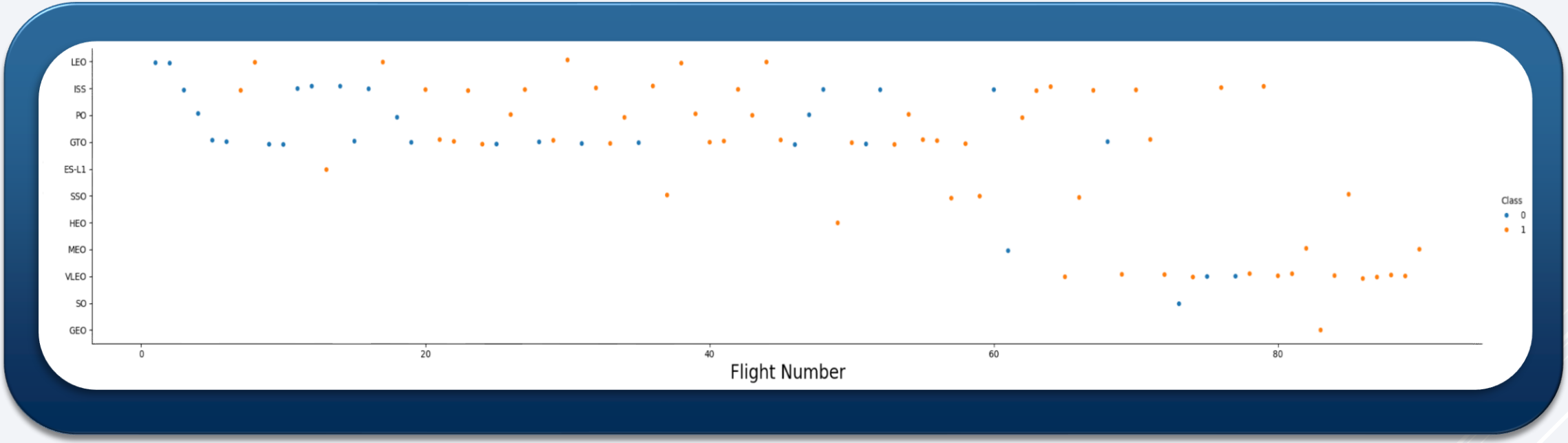
Success Rate vs. Orbit Type



Explanation:

- The Bar chart shows that ES-L1, GEO, HEO, SSO have a perfect success rate but if we look at the number of flights we learn that only SSO orbit has a flight number above 1.
- SO has no success at all, but the flight number is also 1 so we can't draw a conclusion.
- All the other orbits shows a Success rate above 50%.

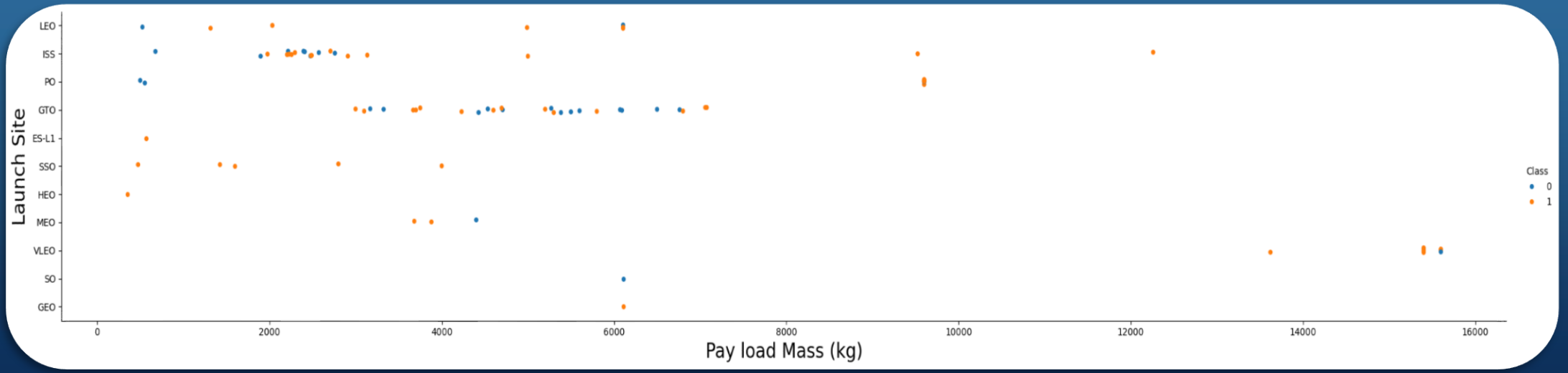
Flight Number vs. Orbit Type



Explanation:

- The Scatter plot shows that most flights occurred in GTO, ISS, LEO, VLEO orbits
- VLEO, LEO and SSO have a great success rate (HEO, MEO and GEO too but they lack of minimal number of observations)
- We can also see here that as the flight number increases, so does the success rate. An exception is GTO where we can't see this trend.

Payload vs. Orbit Type



Explanation:

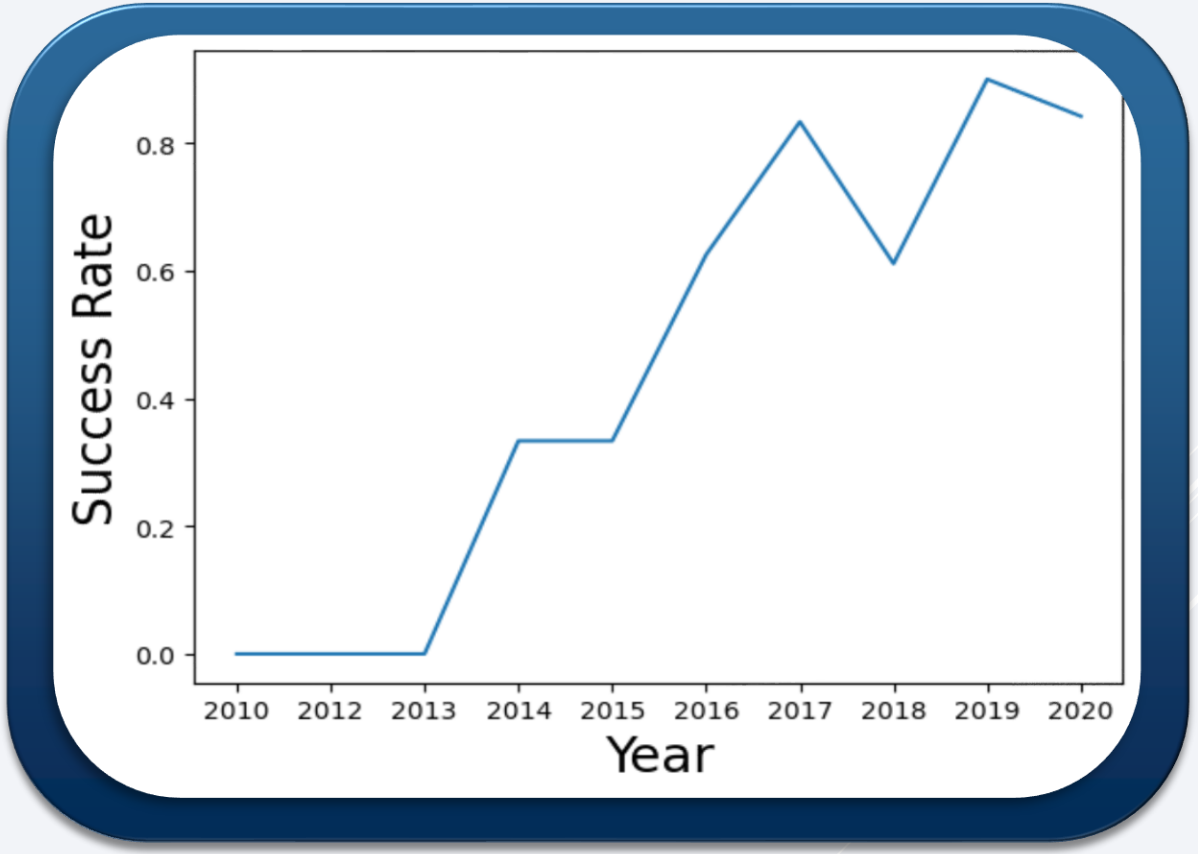
- The Scatter plot shows that pay load mass over 8k kg has high success rate.
- Pay load mass between 0 to 1K and 4k to 8k has low success rate.
- In GTO, the higher the pay load mass the lower the success rate.

Launch Success Yearly Trend

Explanation:

The line chart shows that from 2013 there is a strong increase on success rate over the years.

This trend is consistent with our previous analyzes which showed that the new flights have a greater chance of a successful landing.



All Launch Site Names

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

Explanation:

SQL query: `SELECT DISTINCT launch_site FROM SPACEXTABLE`

We use in our query the DISTINCT clause to display each launch site once.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Explanation:

SQL query: `SELECT * FROM SPACEXTABLE WHERE launch_site LIKE 'CCA%' LIMIT 5`

The LIKE clause filter Launch site that begin with ‘CCA’ and LIMIT 5 shows maximum 5 records from the table.

Total Payload Mass

total_payload_mass
45596

Explanation:

SQL query: `SELECT SUM(payload_mass__kg_) AS total_payload_mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'`

The total payload mass of NASA (CRS) as a customer is 45,596 kg

Average Payload Mass by F9 v1.1

avg_payload_mass
2534.6666666666665

Explanation:

SQL query: `SELECT AVG(payload_mass__kg_) AS avg_payload_mass FROM SPACEXTABLE WHERE booster_version LIKE 'F9 v1.1%'`

The average payload mass where the booster version starts with 'F9 v1.1'

First Successful Ground Landing Date

first_landing_on_ground_pad
2015-12-22

Explanation:

SQL query: `SELECT MIN(date) AS first_landing_on_ground_pad FROM SPACEXTABLE WHERE landing_outcome = 'Success (ground pad)'`

The first successful ground landing occurred in 22 DEC 2015. We got this answer using the MIN function and the WHERE clause.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Explanation:

SQL query: `SELECT DISTINCT booster_version FROM SPACEXTABLE WHERE payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000 AND landing_outcome = 'Success (drone ship)'`

This query filter with WHERE clause booster versions with payload mass between 4k and 6k and by using DISTINCT to display each booster once.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	count_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Explanation:

SQL query: `SELECT mission_outcome, COUNT(*) AS count_outcome FROM SPACEXTABLE GROUP BY trim(mission_outcome)`

As we can see 100 out of 101 missions the outcome was successful.

Boosters Carried Maximum Payload

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

Booster_Version
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Explanation:

SQL query: `SELECT DISTINCT booster_version FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXTABLE)`

The query display boosters where the payload mass is the max value in our data. We used a subquery technique to retrieve the max value of payload mass.

2015 Launch Records

month	Landing_Outcome	Date	Booster_Version	Launch_Site
01	Failure (drone ship)	2015-01-10	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	2015-04-14	F9 v1.1 B1015	CCAFS LC-40

Explanation:

SQL query: `SELECT substr(Date, 6,2) AS month, landing_outcome, date, booster_version, launch_site FROM SPACEXTABLE WHERE landing_outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015'`

The query displays failed landing that occurred in 2015. We used the substr function and the where clause.

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

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

Explanation:

SQL query:

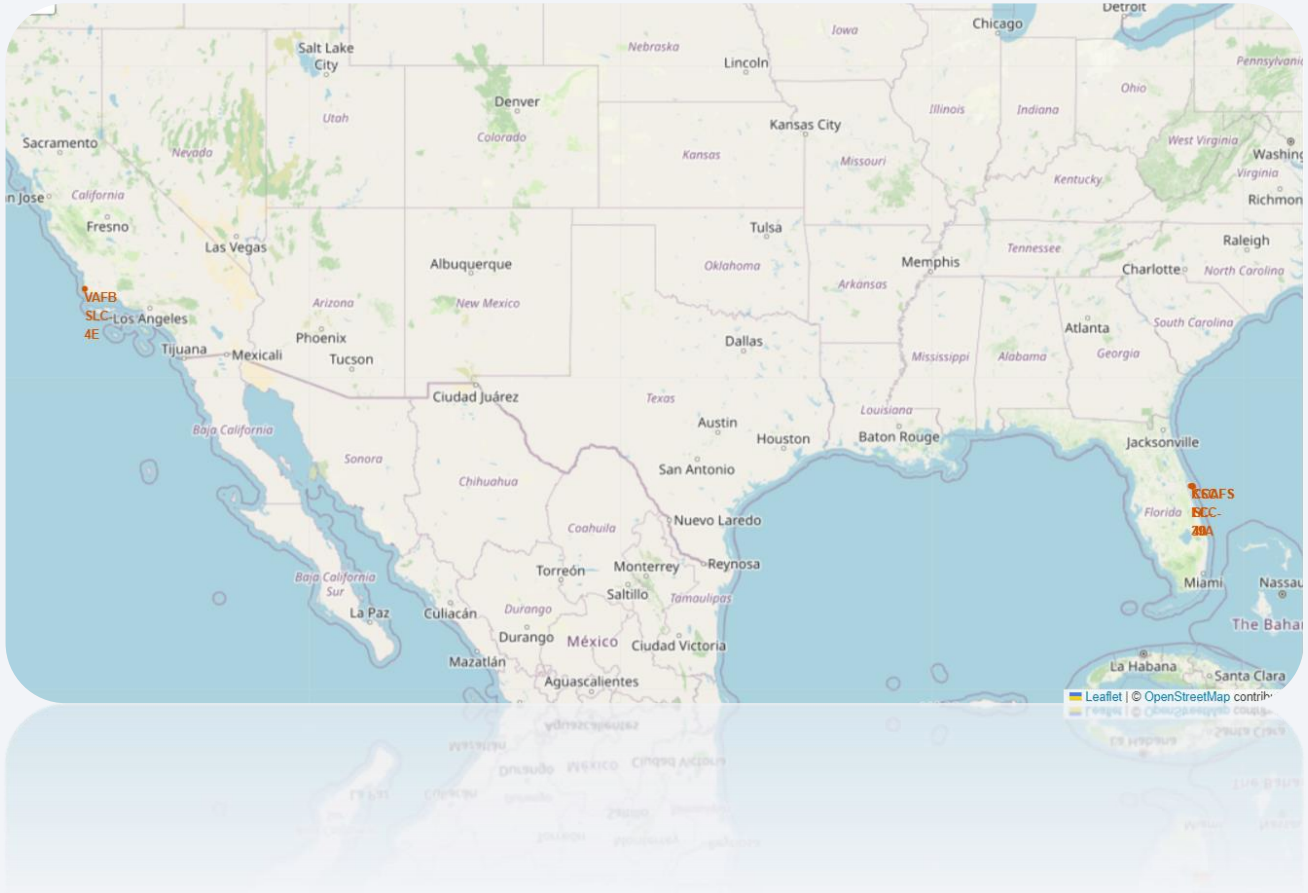
```
SELECT landing_outcome, count(*) AS count_outcomes FROM SPACEXTABLE WHERE Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP BY landing_outcome ORDER BY count_outcomes DESC
```

The query displays the ranking of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order. We used the WHERE clause to filter, ORDER BY to show records in descending order, Group by and count function to display for each landing outcome the amount of records.

Section 3

Launch Sites Proximities Analysis

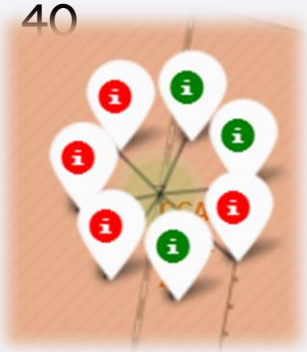
Folium Map – Launch sites



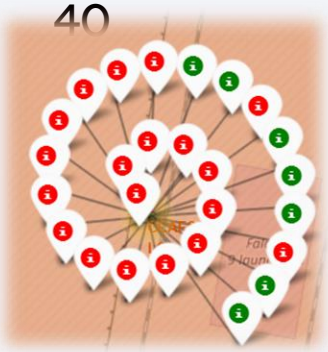
- We can learn that all Launch sites are very close to the coastline
- CCAFS SLC-40 and CCAFS LC-40 are very close to each other
- VAFB SLC-4E is relatively far away from the other launch sites.

Folium Map – Launch outcome by sites

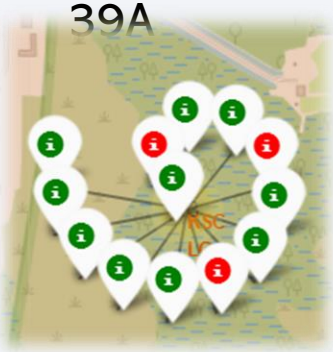
CCAFS SLC-40



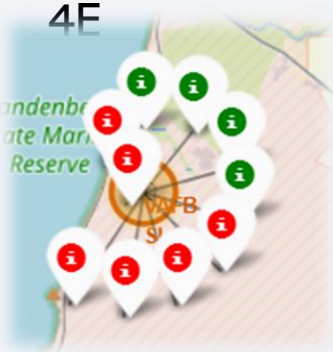
CCAFS LC-40



KSC LC-39A



VAFB SLC-4E

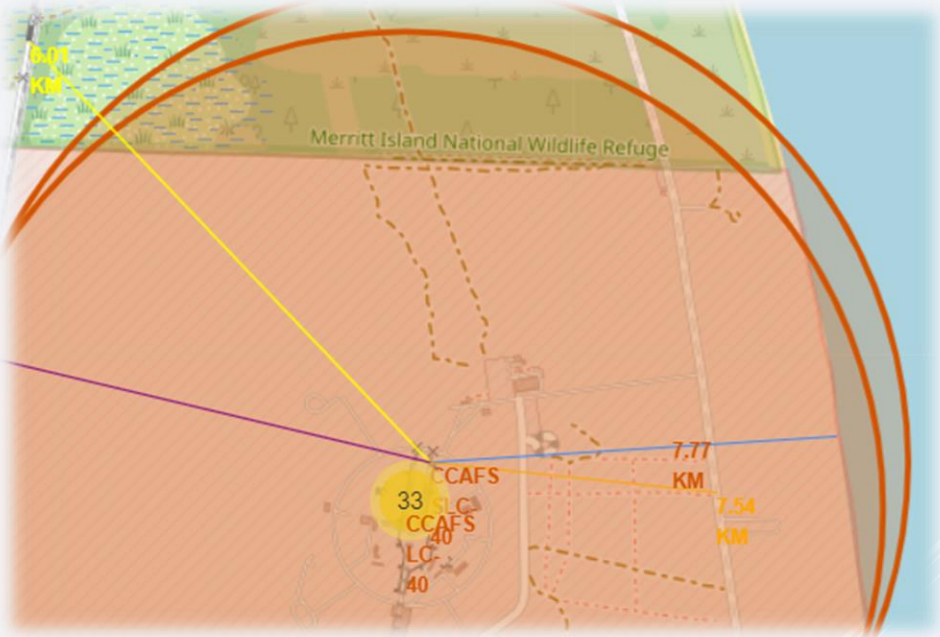


Green marker – Successful launch

Red marker – Unsuccessful launch

- We can see that CCAFS SLC-40 has the fewest launches and CCAFS LC-40 has the most launches.
- KSC LC-39A has the best ratio of successfully launches and CCAFS LC-40 has the worst ratio.

Folium Map – Distance from Launch site to selected landmarks

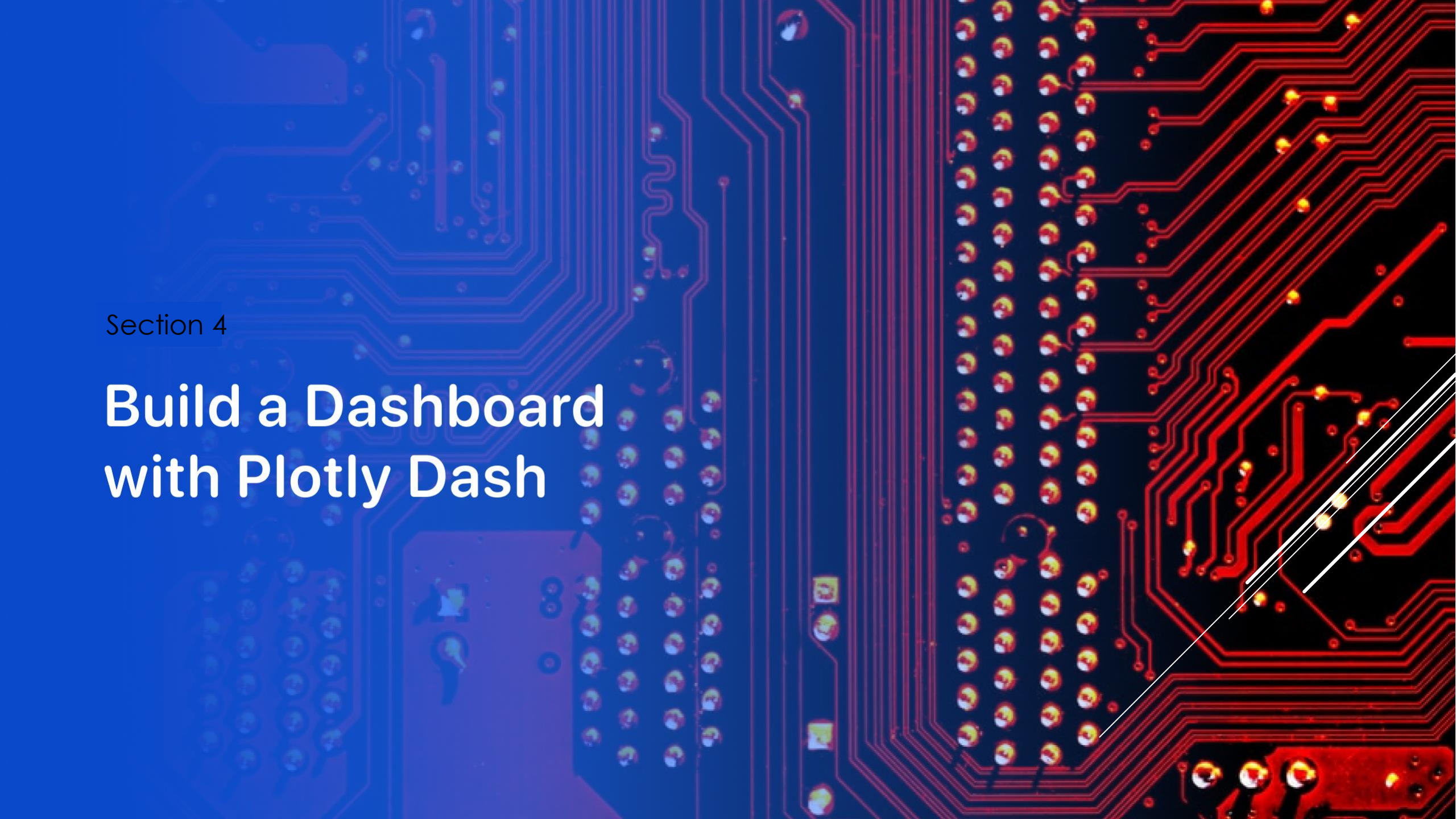


From the Folium map analysis we found out the following facts about CCAFS SLC-40:

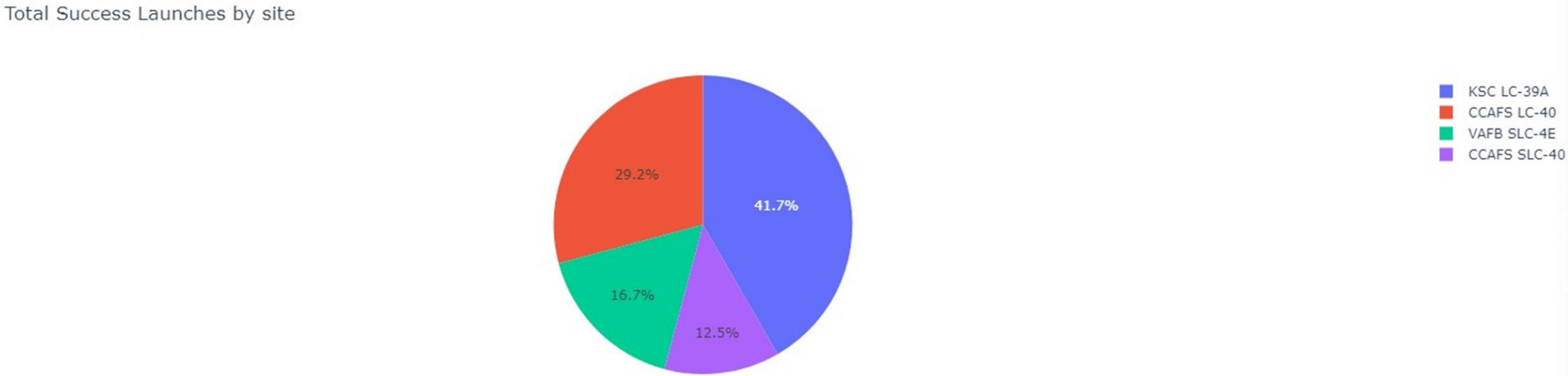
- The site is 16.27 km from closest city.
- The site is 6.01 km from railway.
- The site is 7.54 km from highway.

Section 4

Build a Dashboard with Plotly Dash

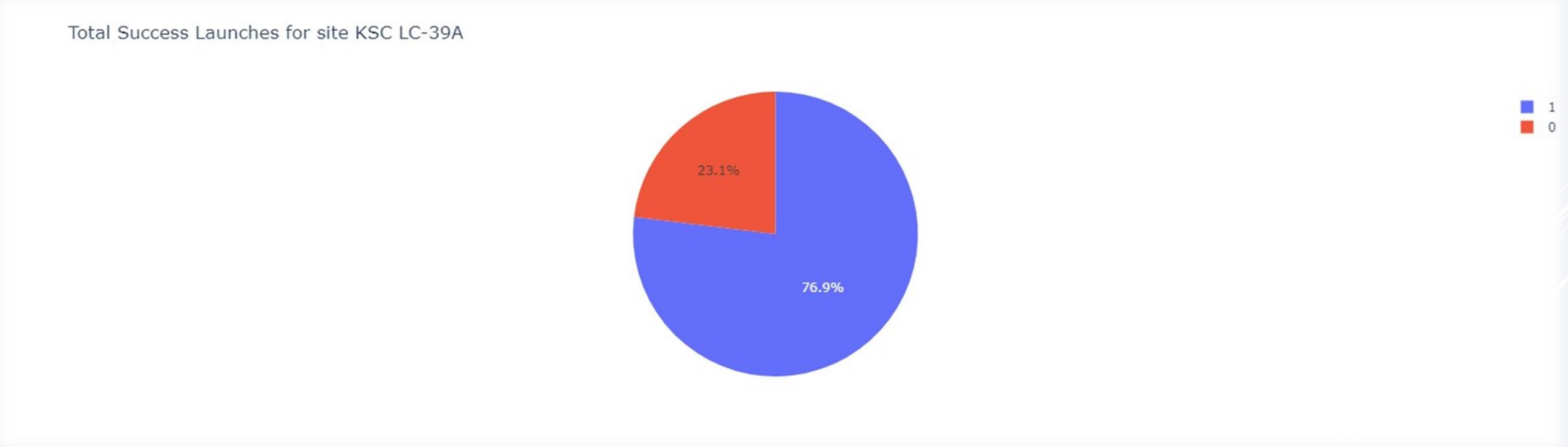


Dashboard – Success launch of each site



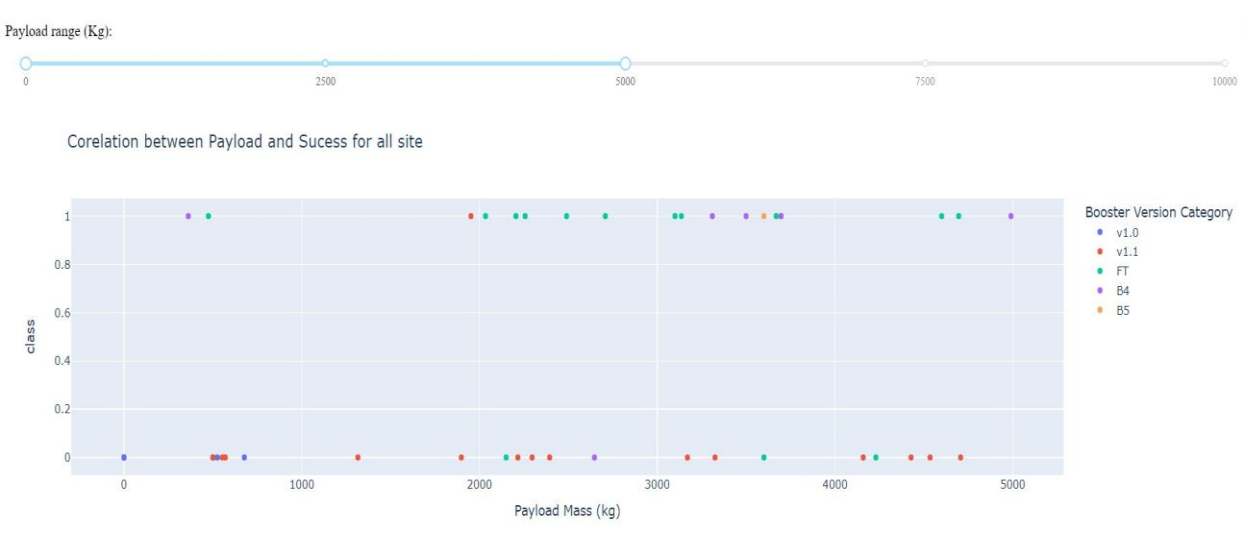
We can see that KSC LC-39A has the largest count of successful launches of all sites.

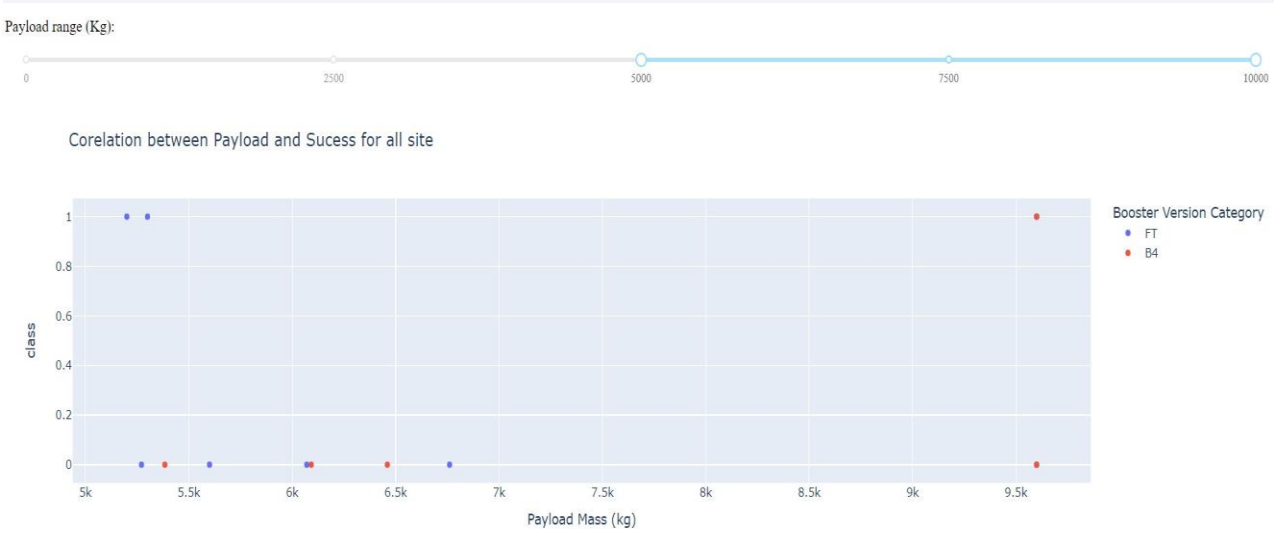
Dashboard – Success launches of KSC LC-39A



We can learn that the ratio of successful launches for the site KSC LC-39A is 76.9%

Dashboard – Payload and Booster Version vs. Launch outcome for all sites



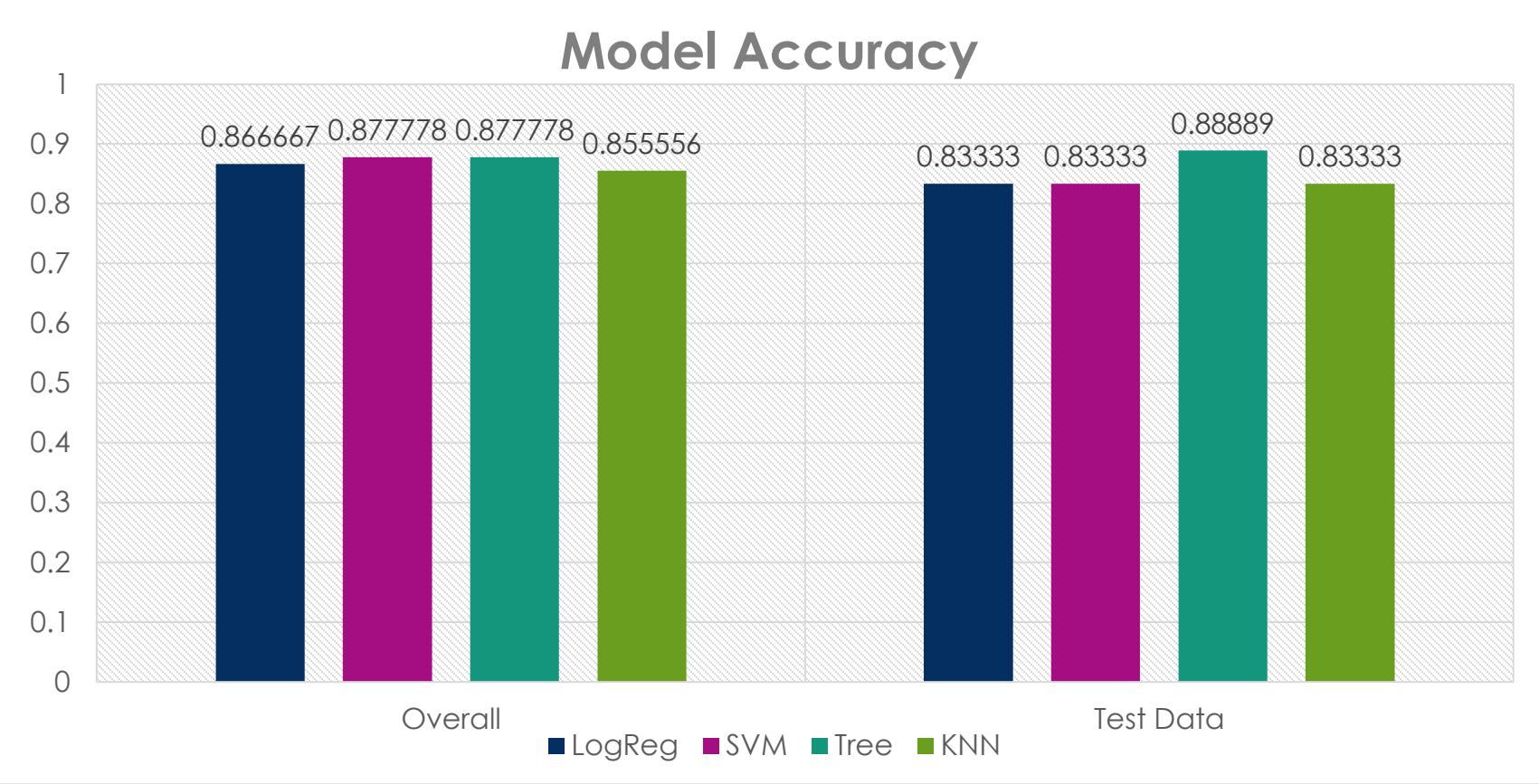


- When weighted payload is lower than 5k kg the success rate is higher than over 5k kg
- Booster version 1.1 has the worst success rate of successful landing.
- Booster version FT has the largest amount of of successful landing.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



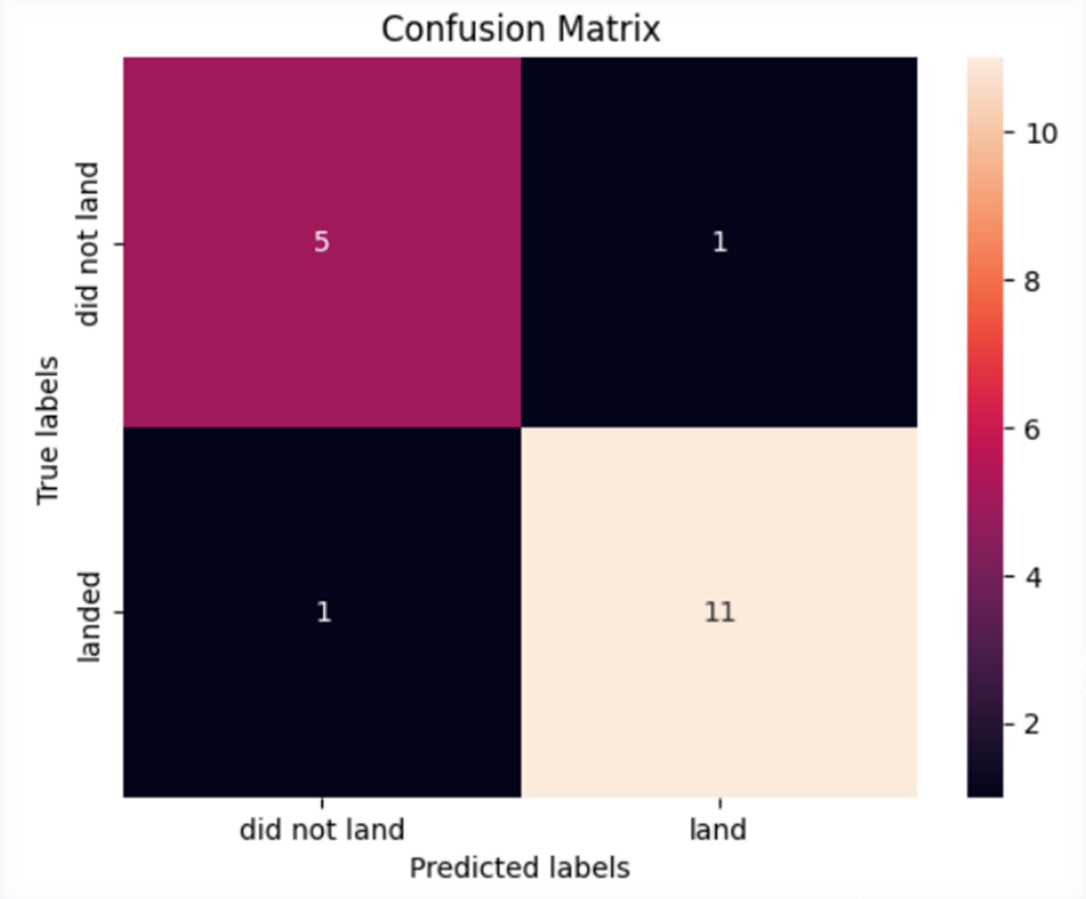
	LogReg	SVM	Tree	KNN
overall	0.866667	0.877778	0.877778	0.855556
test	0.833333	0.833333	0.888889	0.833333

- According to our analysis we found that all the models have classification accuracy result that close to each other with a slightly better result for the Tree model, to get more precise result we need a larger data set.

Confusion Matrix

From the confusion matrix of the Tree model we can learn:

- We have 1 False Positive case it mean the Launch didn't land but the predicted value is a successful landing.
- We have 1 False Negative case it mean the Launch successfully landed but the predicted value is did not land.
- All the other cases marked correct (True Positive and True Negative)



Conclusions

- There is an improvement in successful launches over the years.
- Payload Masses above 8k kg have a significant higher chance of successful launches.
  The KSC LC 39A also has a high chance on Payload mass below 5.5K kg.
- VLEO, LEO and SSO Orbits have a great success rate (also HEO, MEO and GEO but they lack of minimal number of observations so we can't draw a conclusion about them).
- All launch sites are very close to the coastline and equator line.
- The tree model give us the best accurate result. In order to obtain an adequate level of confidence, we will repeat the calculation of the models when the number of observations increases.

Thank you!