IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marcelo Bin Resende da Silva

01/02/2023

Outline

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

Executive Summary

In this study, we collected SPACEX data both via API and scraping information from Wikipedia.

We processed the data using Python and were able to obtain information using SQL as well.

We created maps and dashboards to assist us in data visualization and analysis, and we also created machine learning models to predict the success outcome of a launch.

Introduction

The aim of this study is to use data science to predict whether the first stage of the Falcon 9 will land successfully.

The Falcon 9 has a cost of $62 million for its launch. A much lower value than its competitors (approximately $165 million). This savings presented by SpaceX is that it can reuse the first stage.

Section 1

Methodology

Methodology

Summary

Data collection methodology:

- API REST / WebScraping

Run Data Wrangling

- Data Processing and Cleansing using Python primarily

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analysis using Folium and Plotly Dash

Perform predictive analysis using classification models

How to build, adjust, evaluate classification models

Data Collection

- The collected data was extracted using a Space X REST API and also a webscraping of a Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

API REST
•Flight Number •Date •Booster Version •Payload Mass •Orbit •LaunchSite •Outcome •Flights •GridFins •Reused •Legs •LandingPad •Block •Reused Count •Serial •Longitude - Latitude

WebScraping
•Data Collected: •Flight No., •Date and time, •Launch site, •Payload, •Payload mass, •Orbit, •Customer, •Launch outcome

Data Collection – SpaceX API

- Using the SpaceX API we did the extraction of various release information, such as:
- Rocket,
- Payloads,
- Launchpad,
- Etc
- API: https://api.spacexdata.com/v4/launches/past

Data Collection - Scraping

- The Scraping Process was conducted from the Wikipedia site which also contains launch information for the Falcon 9 rocket.
- Libraries such as BeautifulSoup and Requests were used to assist in data collection.

Data Wrangling

- After collection, it's time to analyze the data. For this stage, Python and its Pandas and Numpy libraries were used. Through the analysis we were able to identify:
- The number of launches on each site;
- The number and occurrence of each orbit;
- The number and occurence of mission outcome per orbit type;
- Create a landing outcome label from Outcome column;

EDA with Data Visualization

- In the study, some graphs were made to help visualize and interpret the data. Some examples created in the study:
  - Scatter Plots
  - Bar Chart
  - Line Chart

EDA with SQL

Using SQL we are able to gain interesting insights such as:

- Identifying launch locations;
- Show the total payload mass carried by boosters launched by NASA (CRS);
- Show the average payload mass carried by F9 v1.1 booster version;
- Listing the date of the first successful landing outcome on a ground pad;
- Listing the names of boosters that have succeeded in a drone ship and have a payload mass between 4000 and 6000

Build an Interactive Map with Folium

- Markers have been created on an interactive map generated by the Folium library. We created circles to mark the launch sites, as well as popups to aid in identification. Icons were also added with launch data, whose color is green for successful cases or red for failed cases

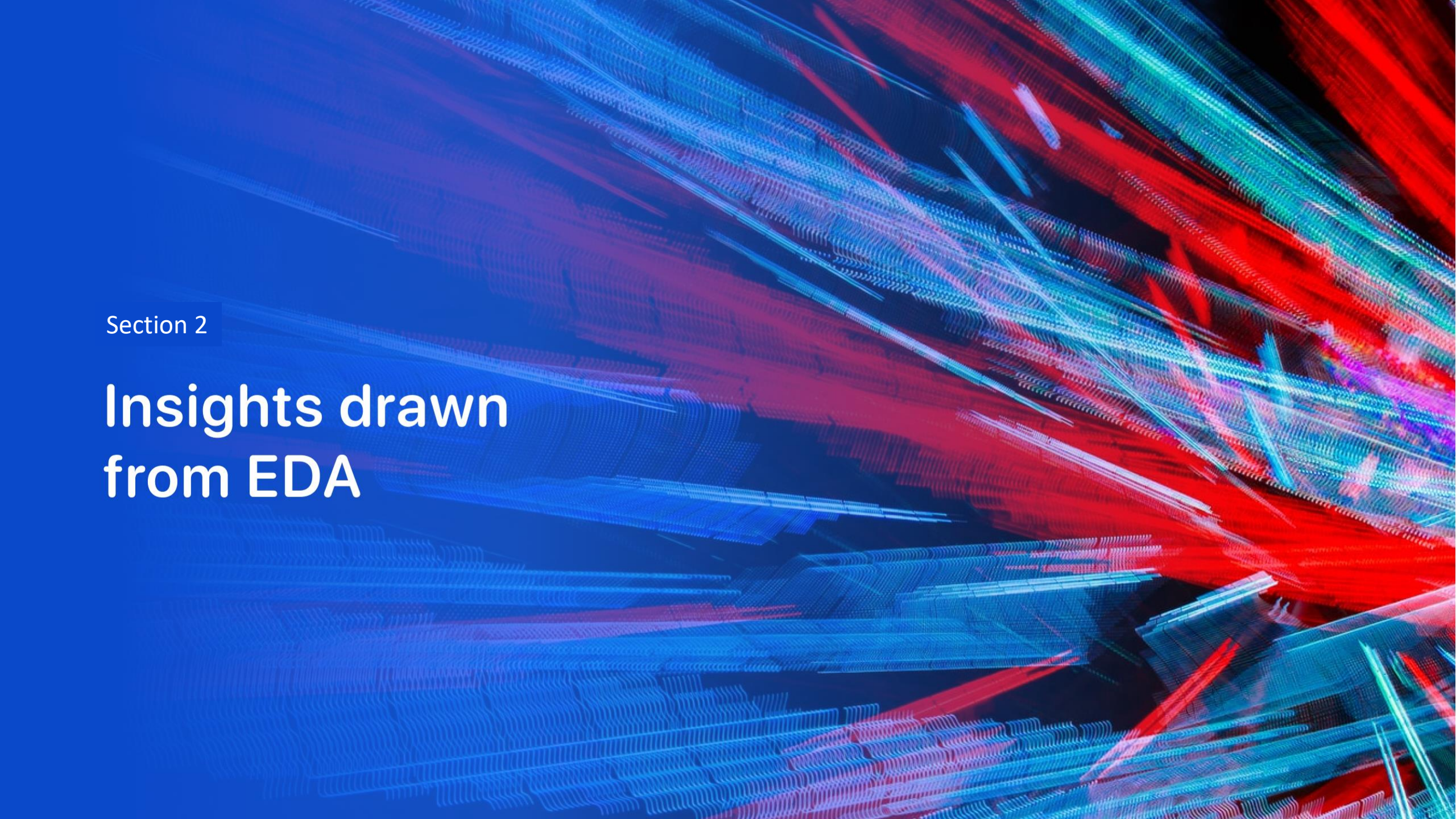
Build a Dashboard with Plotly Dash

- In our interactive dashboard, we use two types of charts:
  - Pie chart, to visualize the number of launches by location and also the success rate of the launch by location;
  - Scatter chart, where we visualize the relationship between the success of the launch and the rocket's payload, and we can also see the models of boosters used

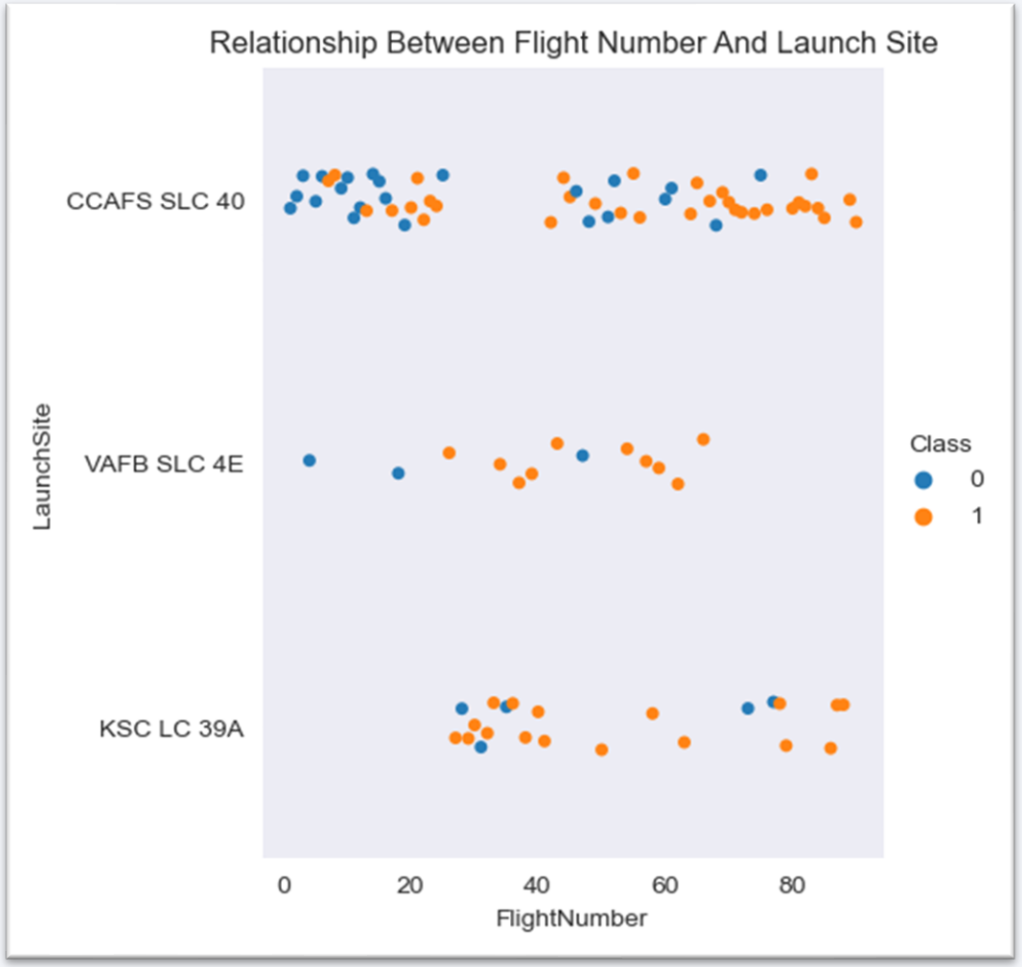
Predictive Analysis (Classification)

- For predictive analysis we tested four models: Logistic Regression, SVM, Decision Tree and KNN;
- For each algorithm we create hyperparameters and use GridSearch to test them and find the best performance.

Results

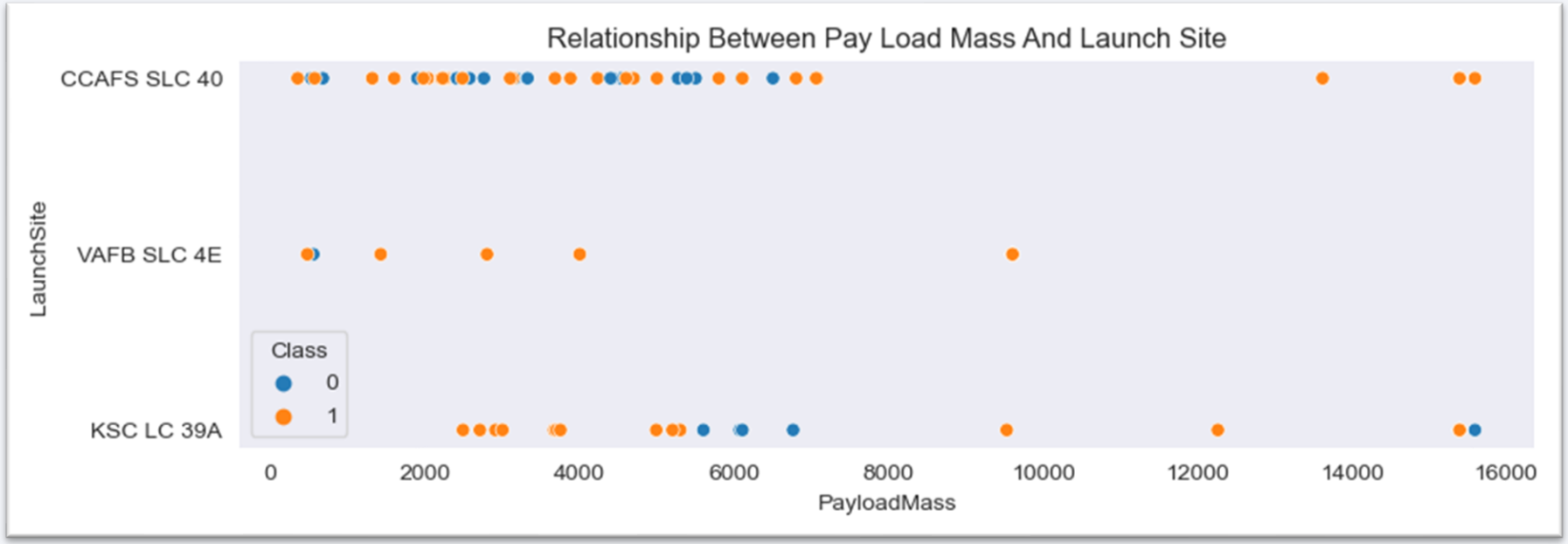

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



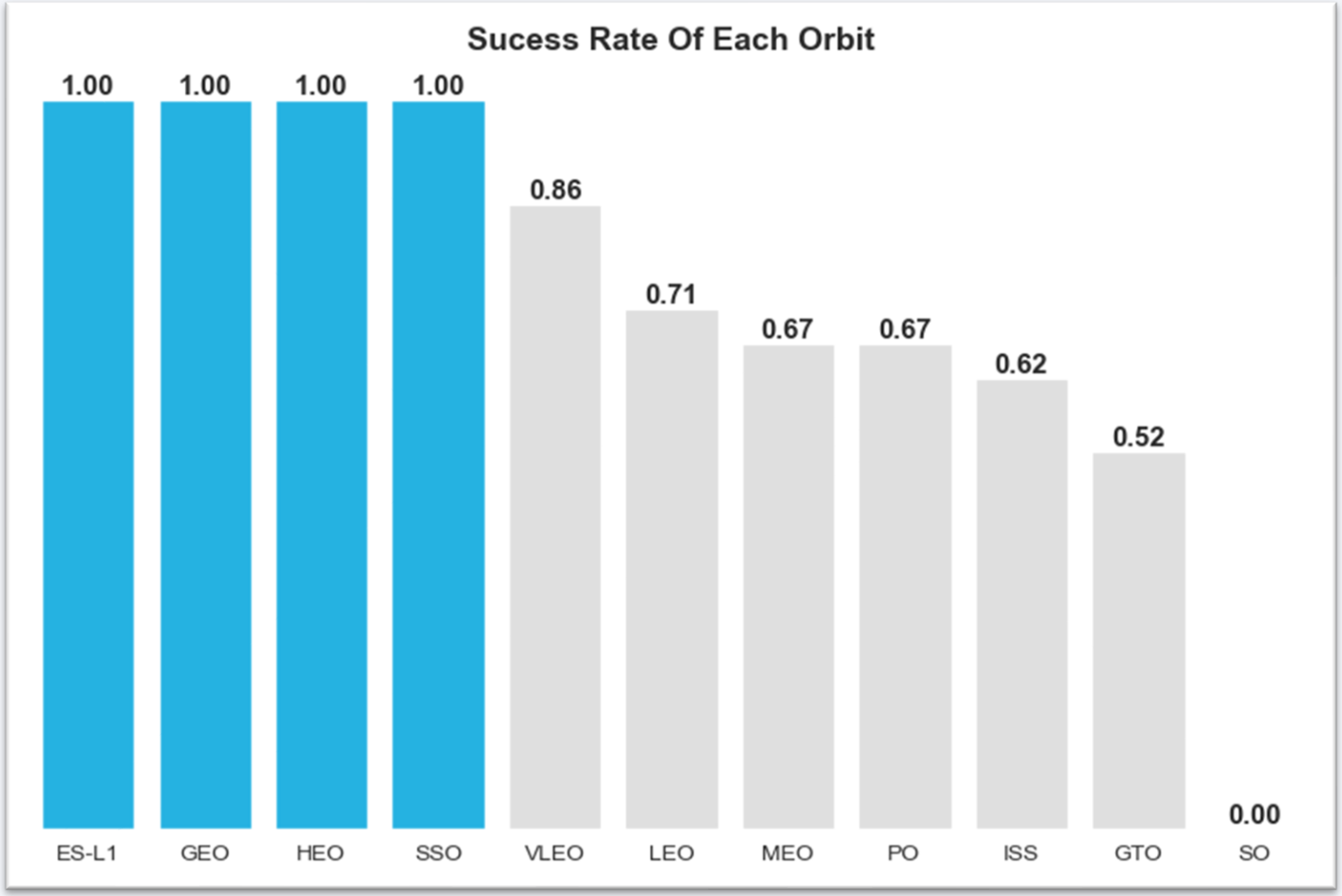
Payload vs. Launch Site

Below we can verify that there are more releases with Payload Mass below 8,000 kg and mostly made at the CCAFS SLC 40 station.



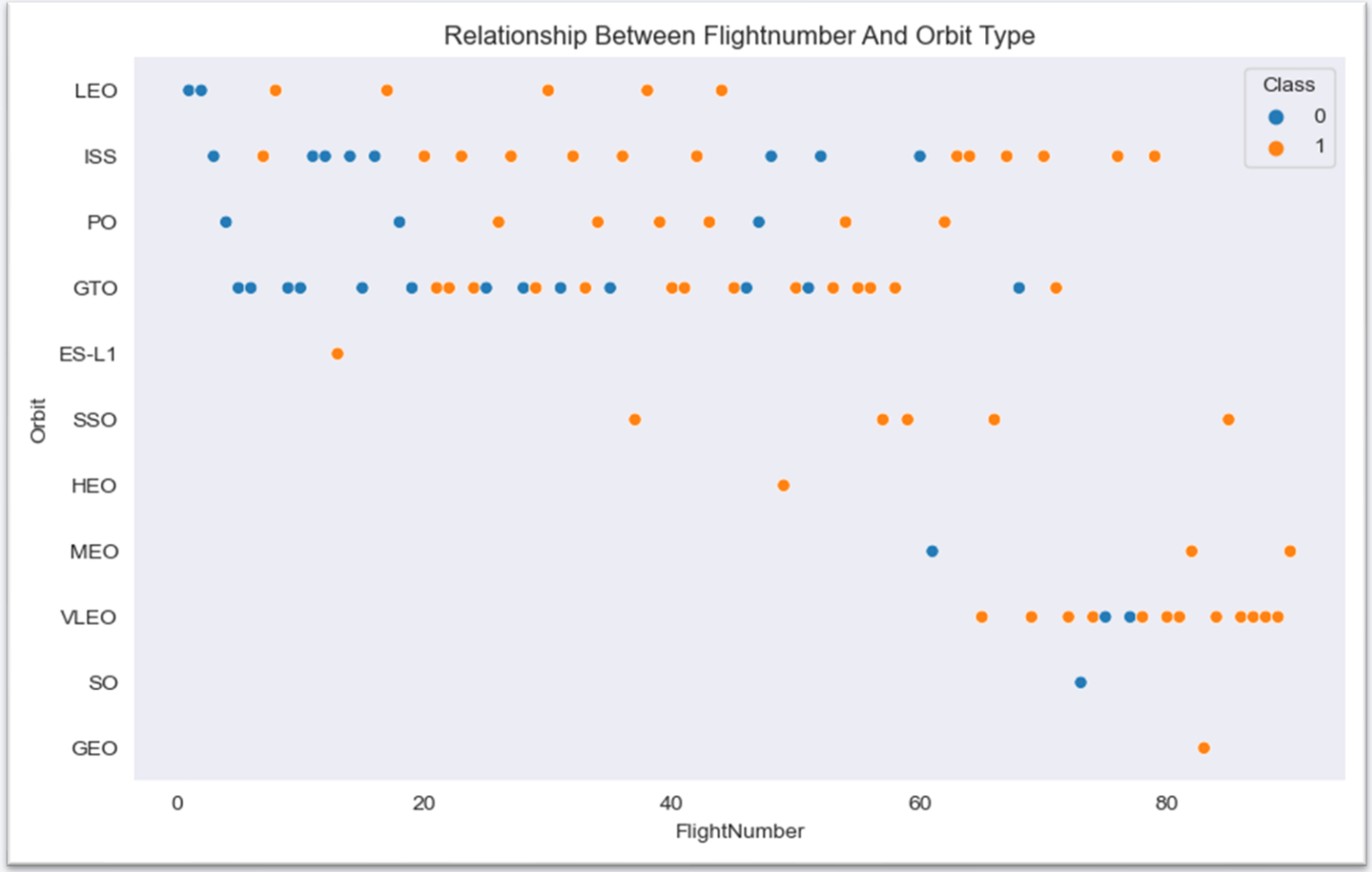
Success Rate vs. Orbit Type

Four orbits stand out in relation to the success of the launch: ES-L1, GEO, HEO, SSO



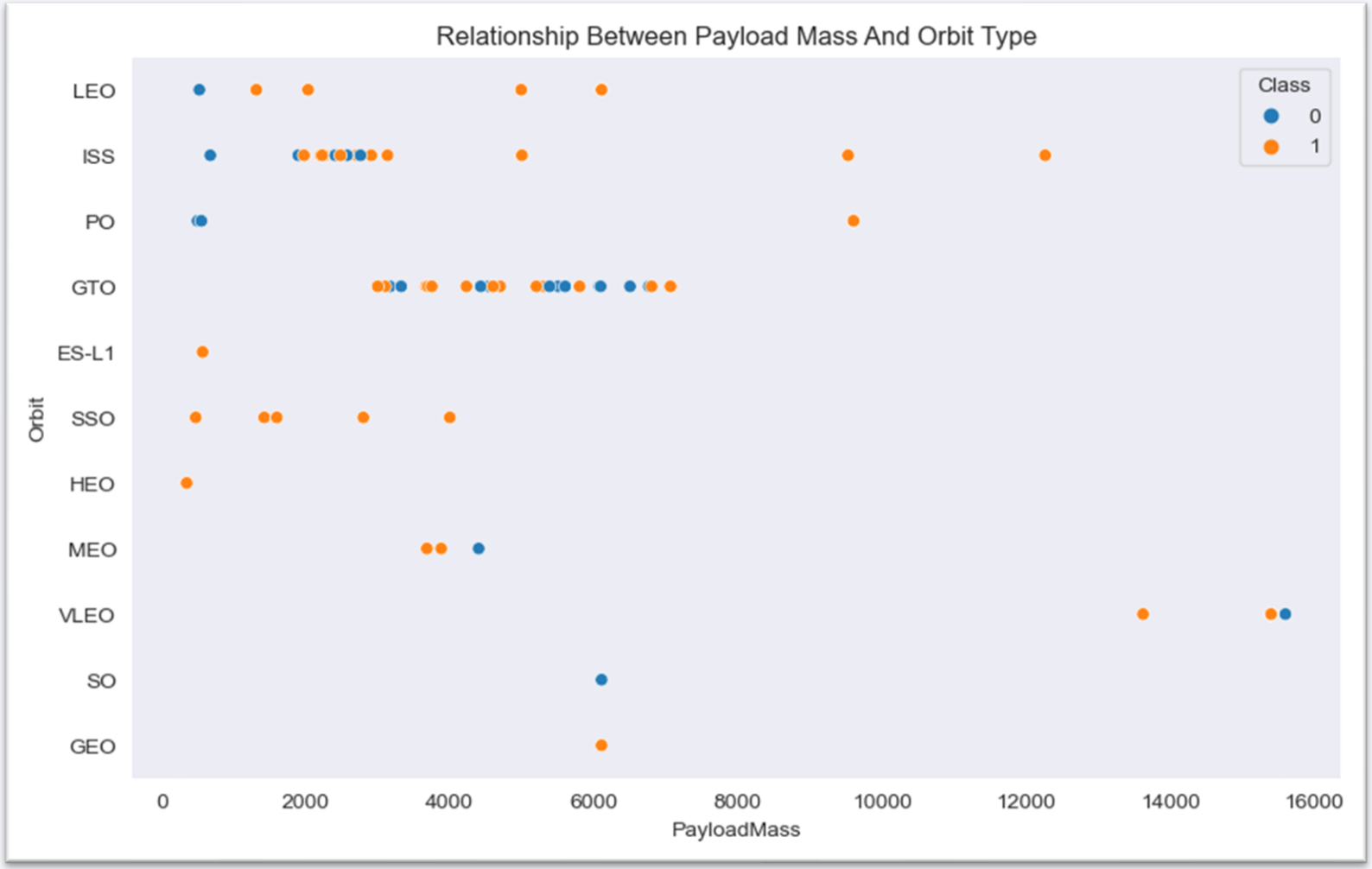
Flight Number vs. Orbit Type

Until flight number 60, we were able to see a concentration of launches in two orbits: ISS and GTO. From that number, there was a concentration of launches in the VLEO orbit.



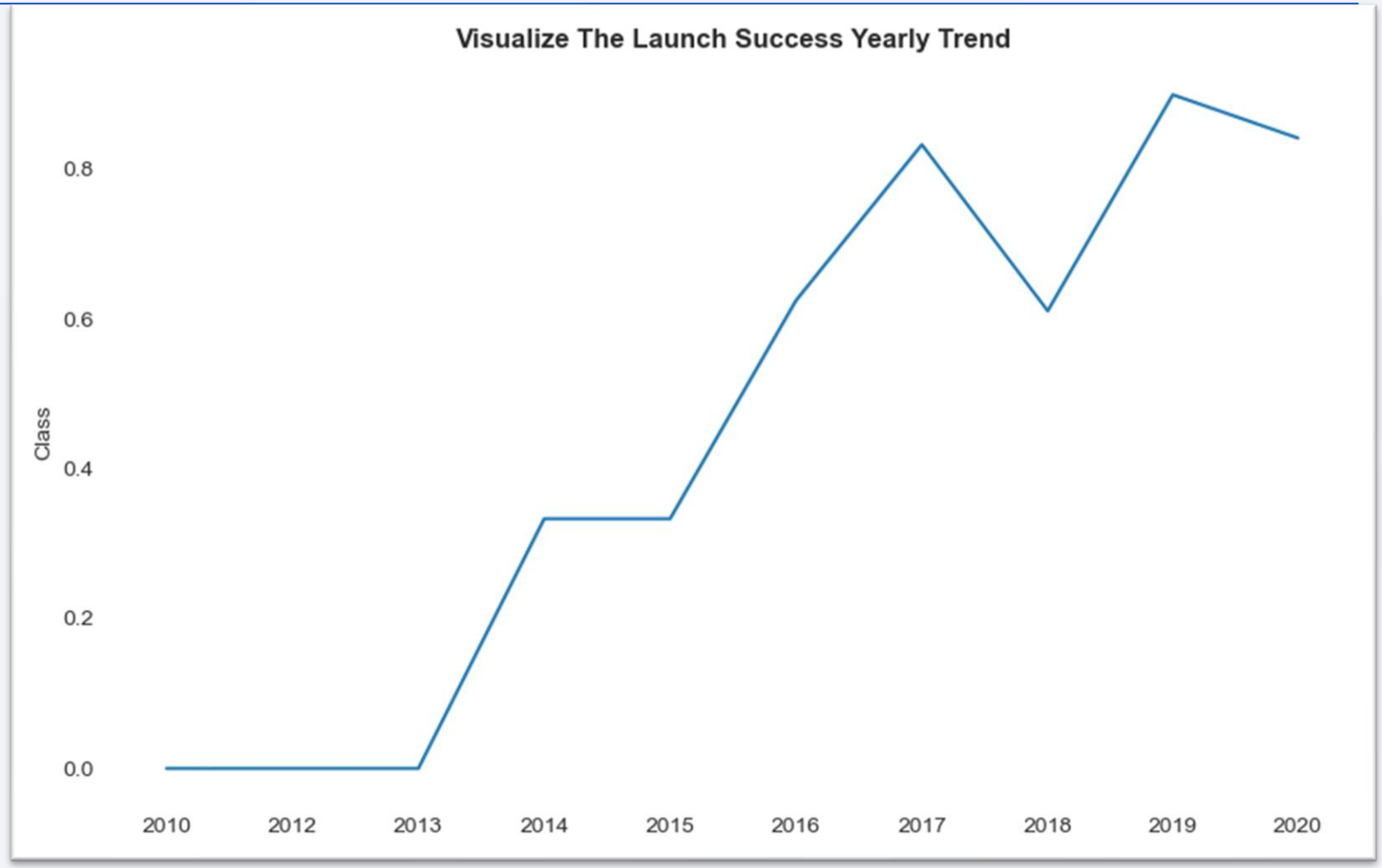
Payload vs. Orbit Type

- Up to 8,000 kg, the most used orbits are GTO and ISS
- For the loads of 10,000 kg and above the most used orbits are VLEO and ISS.



Launch Success Yearly Trend

- The chart on the side demonstrates the evolution of success in falcon 9 launches.



All Launch Site Names

- The exclusive Launch locations are:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

Esses dados foram extraídos utilizando SQL conforme código abaixo:

```
SELECT DISTINCT(launch_site)
FROM SPACEX
```

Launch Site Names Begin with 'CCA'

- Using SQL we make the following query with your result:

```
%%sql
Select * from SPACEX
WHERE launch_site LIKE 'CCA%' limit 5
```

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

```
%%sql
SELECT SUM(payload_mass__kg_) AS TOTAL
FROM SPACEX
WHERE customer LIKE 'NASA (CRS)%'
```

total
48213

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(payload_mass__kg_) as AVG
FROM SPACEX
WHERE booster_version LIKE 'F9 v1.1'
```

AVG
2928

First Successful Ground Landing Date

- Using the SQL query, we can identify that the first successful landing on the ground was on 12/22/2015.

```
%%sql
SELECT MIN(DATE) as First_Success_land_date
FROM SPACEX
WHERE landing__outcome = 'Success (ground pad)'
```

first_success_land_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The next query shows us the four booster versions successfully used for Payload between 4000 and 6000.

```
%%sql
SELECT DISTINCT(booster_version)
FROM SPACEX
WHERE landing__outcome = 'Success (drone ship)' and
payload_mass__kg_ > 4000 and
payload_mass__kg_ < 6000
```

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT COUNT(mission_outcome) AS Success,
(select count(mission_outcome)
    from SPACEX where mission_outcome like 'Failure%') AS Failure
from SPACEX
WHERE mission_outcome = 'Success'
```

success	failure
99	1

Boosters Carried Maximum Payload

```
%%sql
select DISTINCT(booster_version), payload_mass__kg_
from SPACEX
where payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) as max_payload from SPACEX)
```

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

2015 Launch Records

- In 2015 there were two failed drone ship. Next to us we view booster versions and launch location

```
%%sql
select landing__outcome,booster_version, launch_site
from SPACEX
where landing__outcome = 'Failure (drone ship)' and
YEAR(DATE) = 2015
```

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The query next to it shows that between 06/2010 and 03/2017, nine landings were successful on ground pad and five failed on the drone ship.

```
%%sql
select landing__outcome, count(landing__outcome) as landing_counts
from spacex
where DATE between '2010-06-04' AND '2017-03-20' and
landing__outcome = 'Failure (drone ship)' or
landing__outcome = 'Success (ground pad)'
group by landing__outcome
order by landing_counts DESC
```

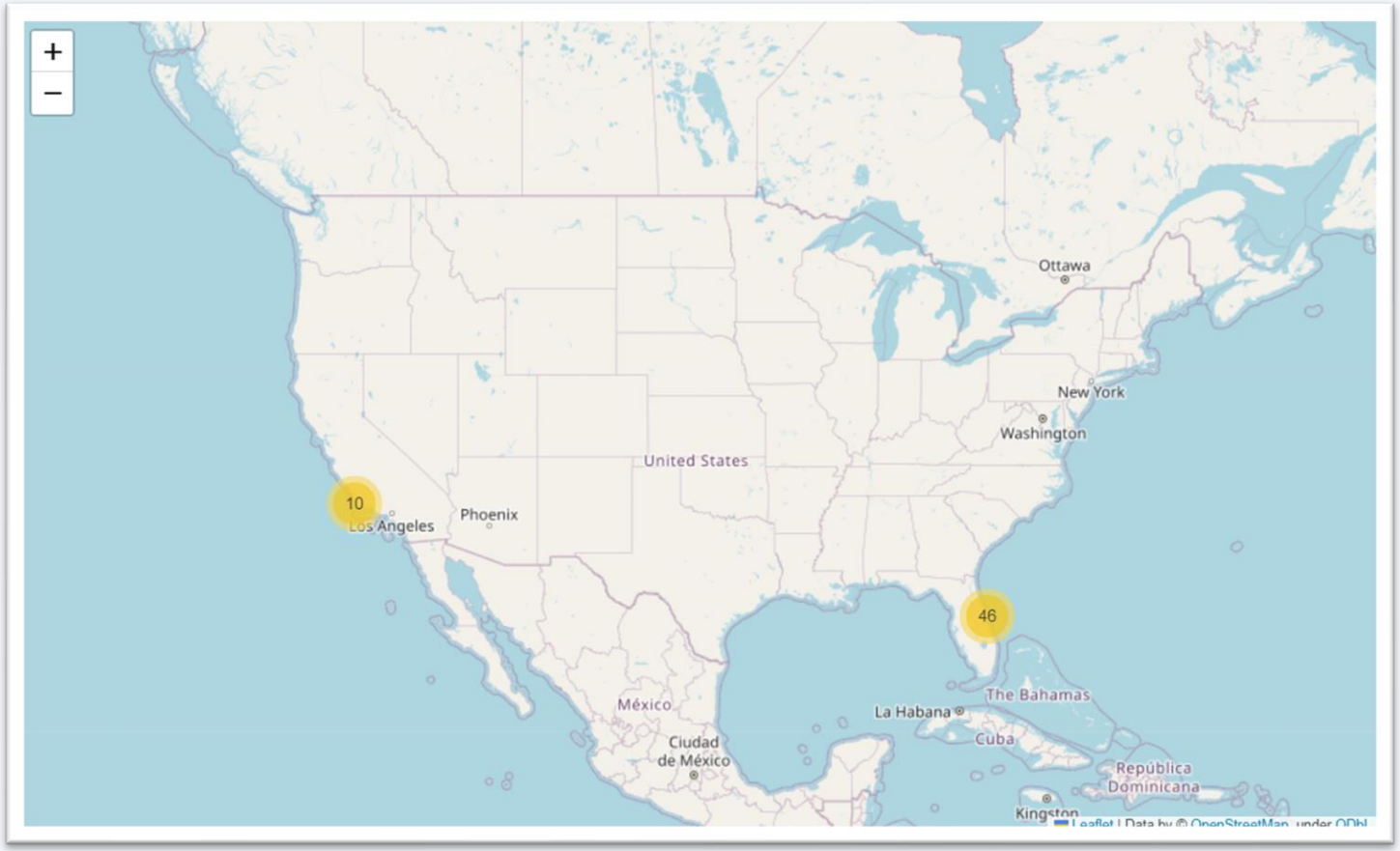
landing__outcome	landing_counts
Success (ground pad)	9
Failure (drone ship)	5

Section 3

Launch Sites Proximities Analysis

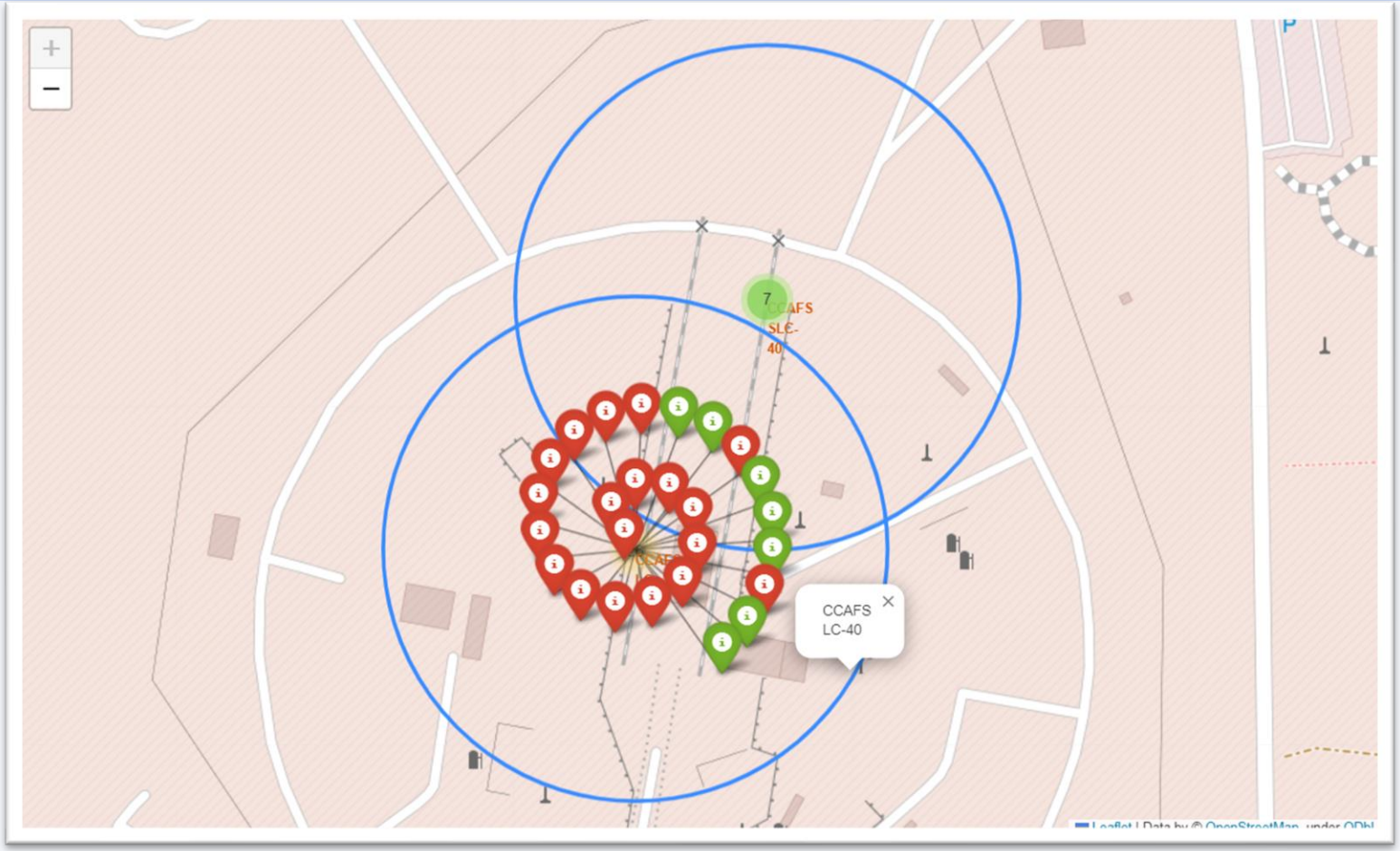
Falcon 9 Launch Locations

- On the map, we can observe the Falcon 9 launch locations. 46 launches were carried out on the east coast and 10 on the west coast.



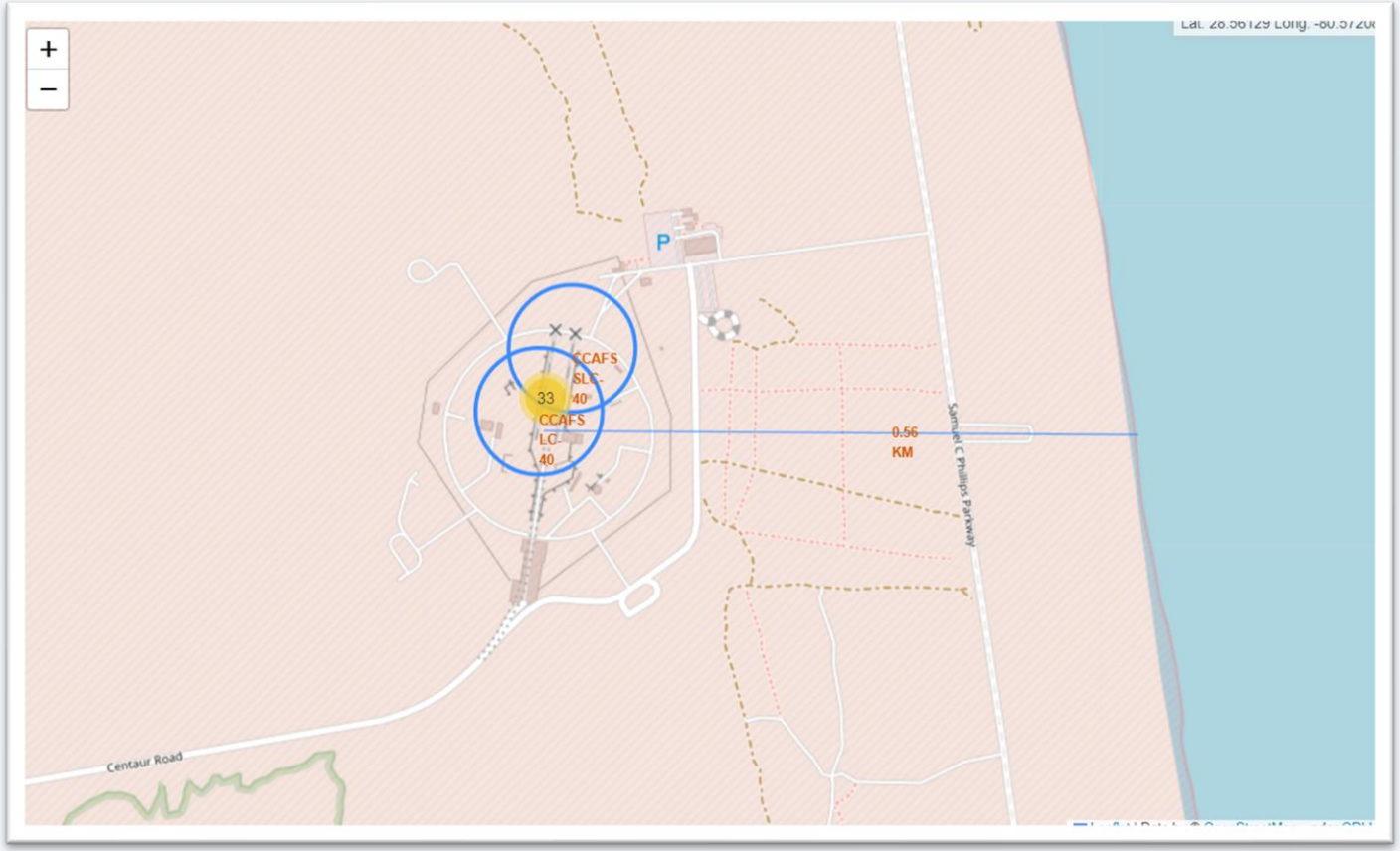
Launch Details

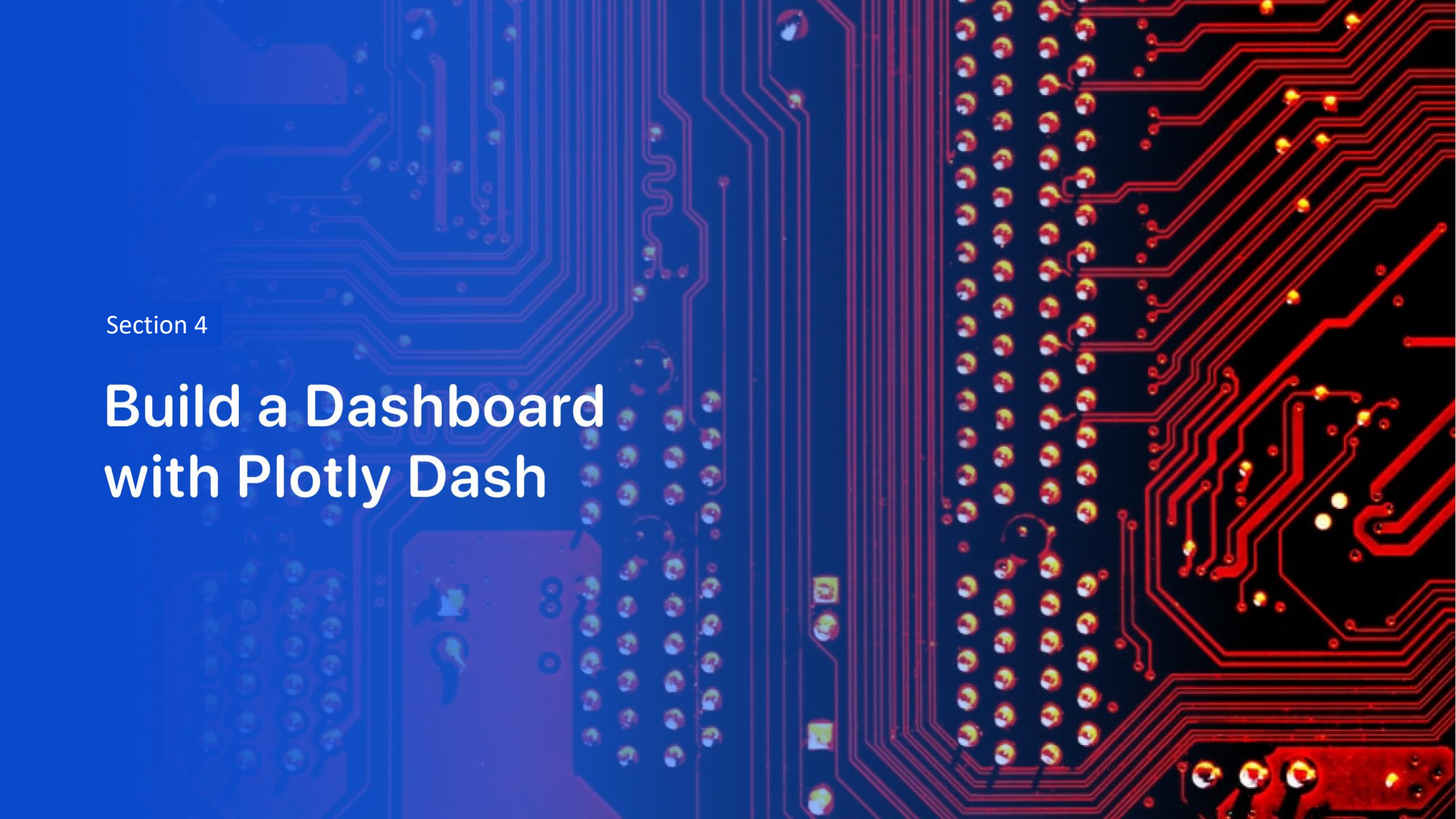
The map on the side shows us the launch locations and also the result of each launch: Green for the successful launches, Red for the launches that failed.



Map Details

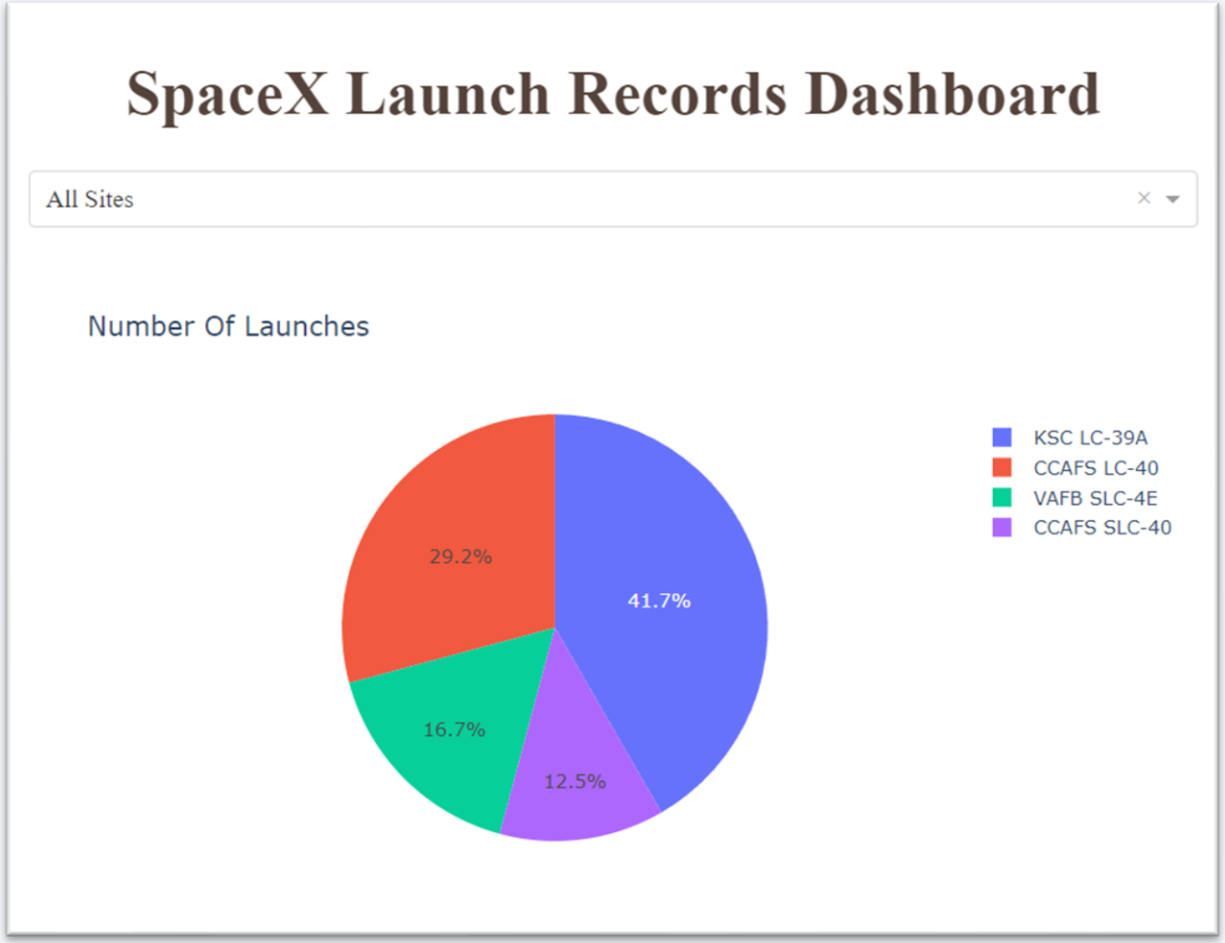
- As seen on the map, we observe that the launch sites, in general, are close to the coast. In the example next to it, the CCAFS Station is approximately 500 meters from the coast




Section 4
Build a Dashboard with Plotly Dash

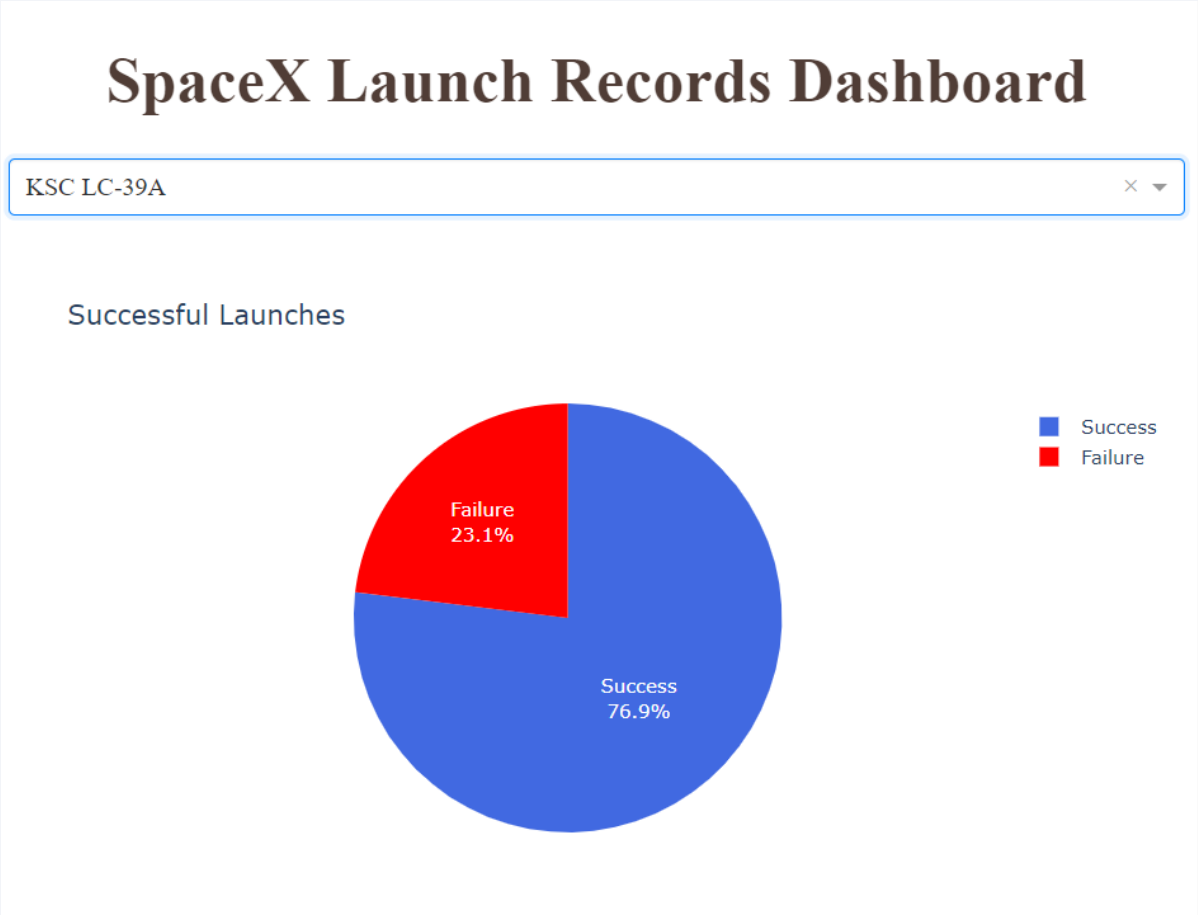
Launch Records

- In the chart to the side, we visualize the number of Falcon 9 launches per station. From it, we can identify that the KSC LC station had the highest number of launches (41.7%).



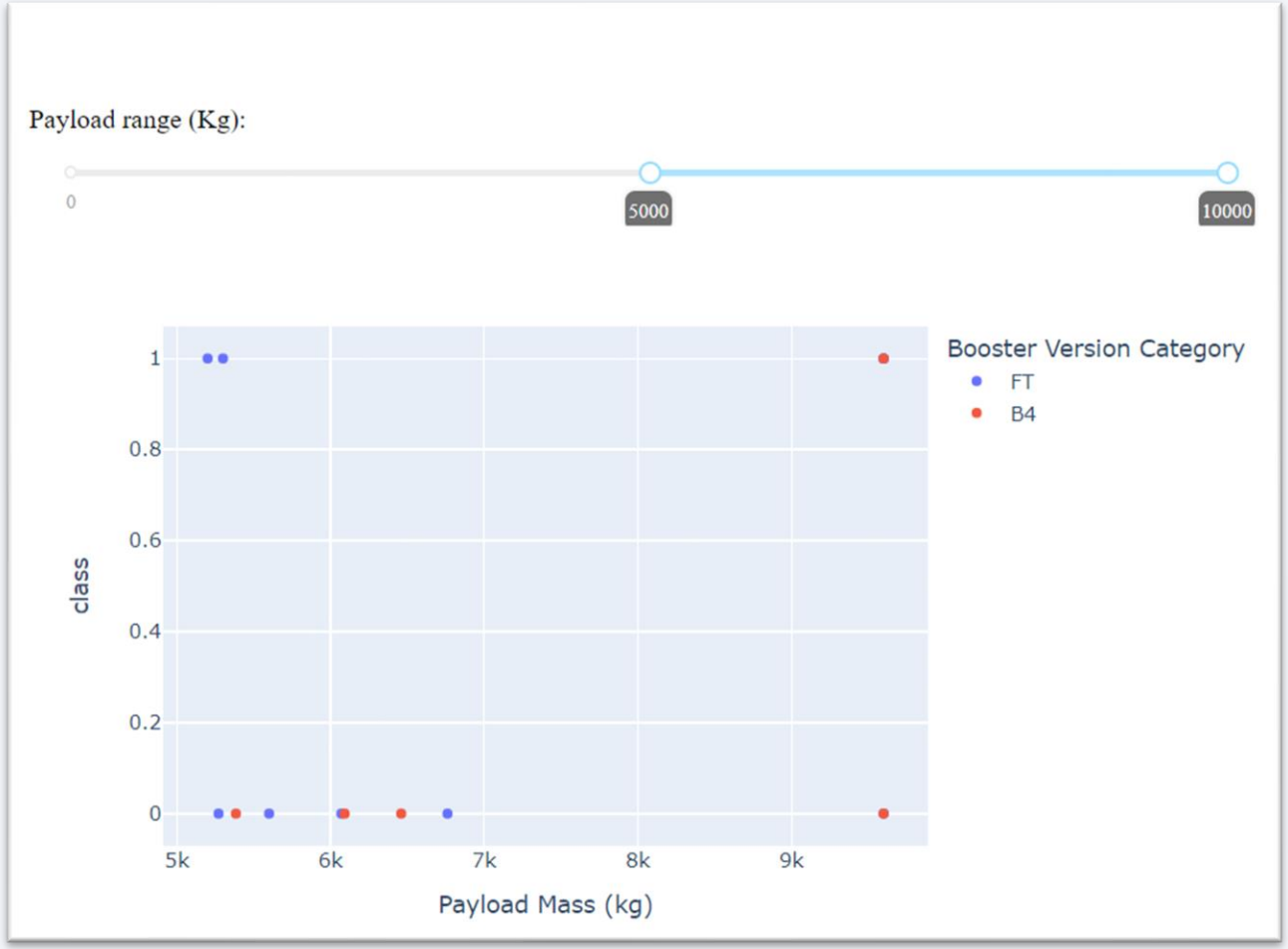
Launch Records – KSC LC

- In addition to having the highest number of launches, KSC LC station also has the highest success rate in launches, approximately 77%



Payload x Success

- The chart to the side shows the relationship between launch success and the Payload used in the rockets. We can see that above 5 tons of payload we have more failures than success in launches. In addition, we can visualize that only two versions of boosters are used: FT and B4

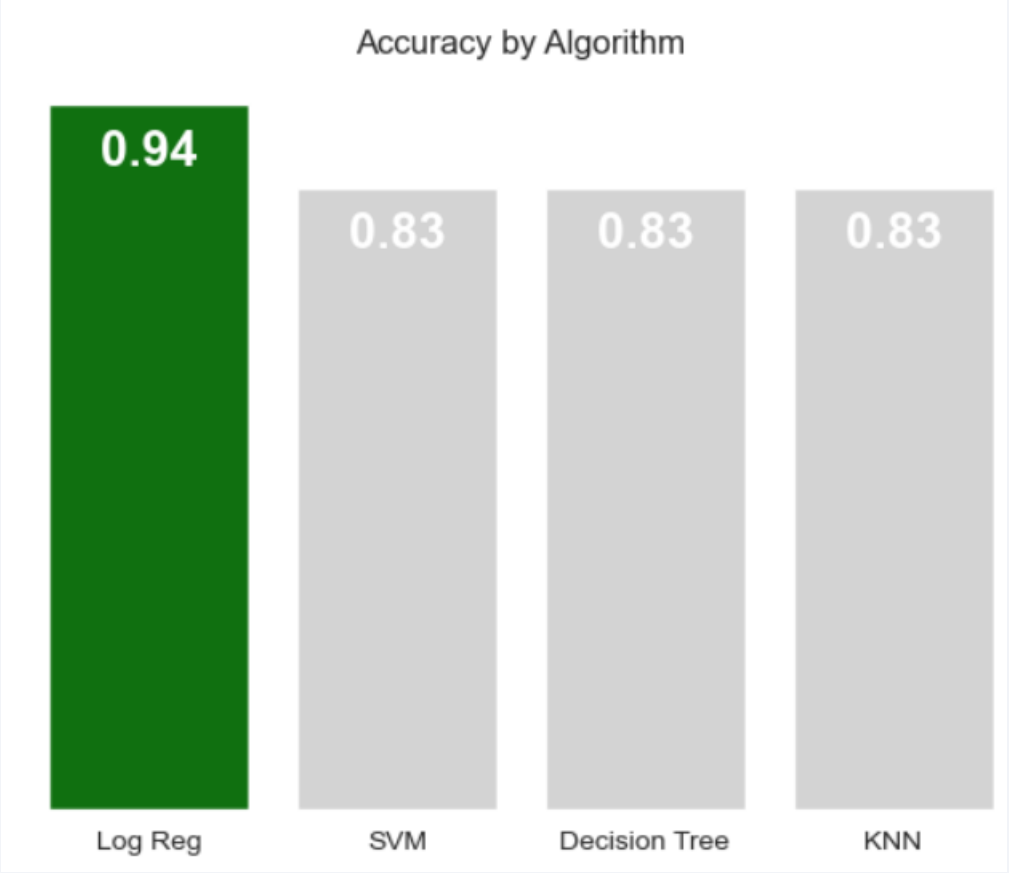


Section 5

Predictive Analysis (Classification)

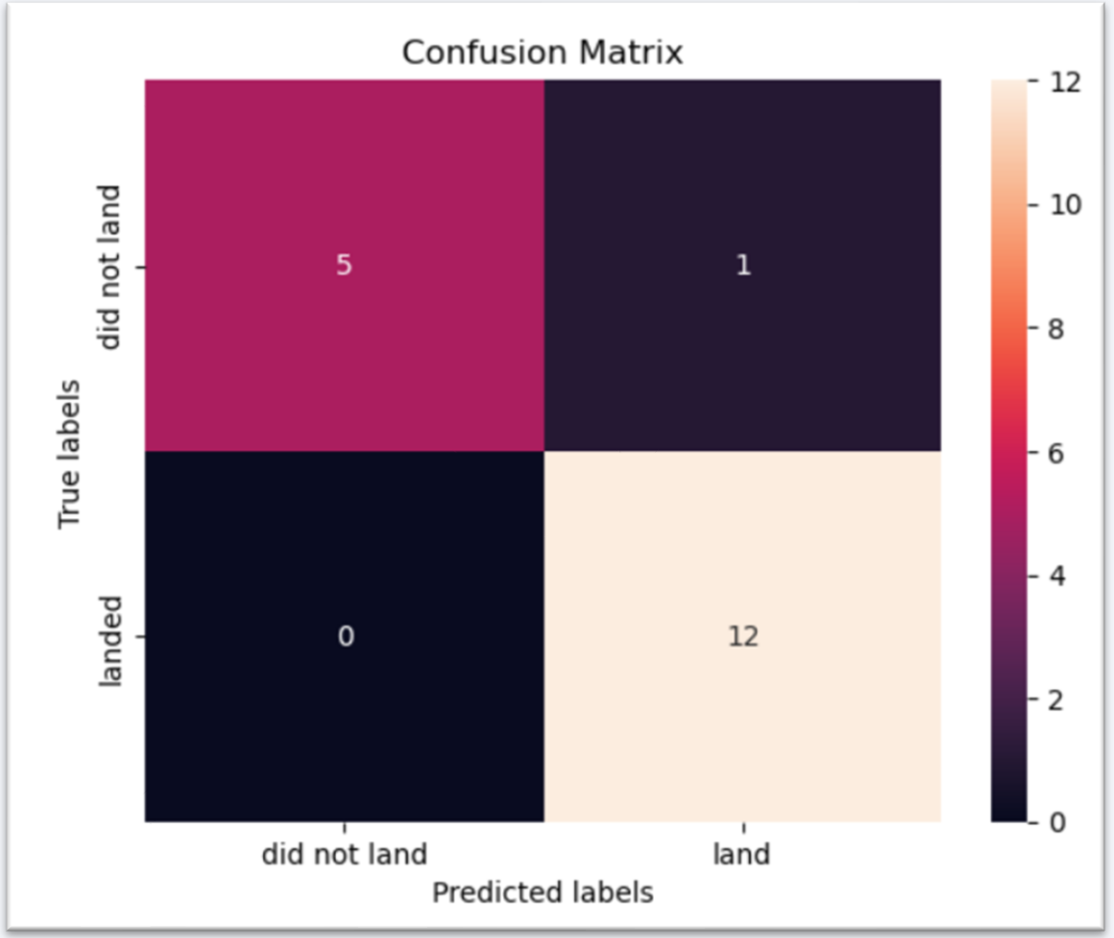
Classification Accuracy

- After some simulations using the Logistic Regression, SVM, Decision Tree and KNN algorithms, Logistic Regression presented the best performance in all simulations, reaching the accuracy of 0.94.



Confusion Matrix

- In the Confusion Matrix, we noticed that Logistic Regression performed well, with only one error.



Appendix

- Notebooks:
- Collecting the data.ipynb
- Data wrangling.ipynb
- EDA-SQL.ipynb
- EDA-data-Viz.ipynb
- SpaceX_Machine Learning Prediction.ipynb
- Web scraping Falcon 9.ipynb

Thank you!