IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Leila Fabiola Ferreira
10-09-2021

Outline

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

Complete Project shared on GitHub
https://github.com/leilaff89/applied_data_science_capstone

Executive Summary

This project will cover a complete study about the Falcon 9 first stage.

The main objective is to predict whether the first stage of Falcon 9 will land successfully.

To accomplish this, some prior processes are necessary to prepare and understand the data being worked on.

This is the Business Understanding.

So let's get started!

Introduction

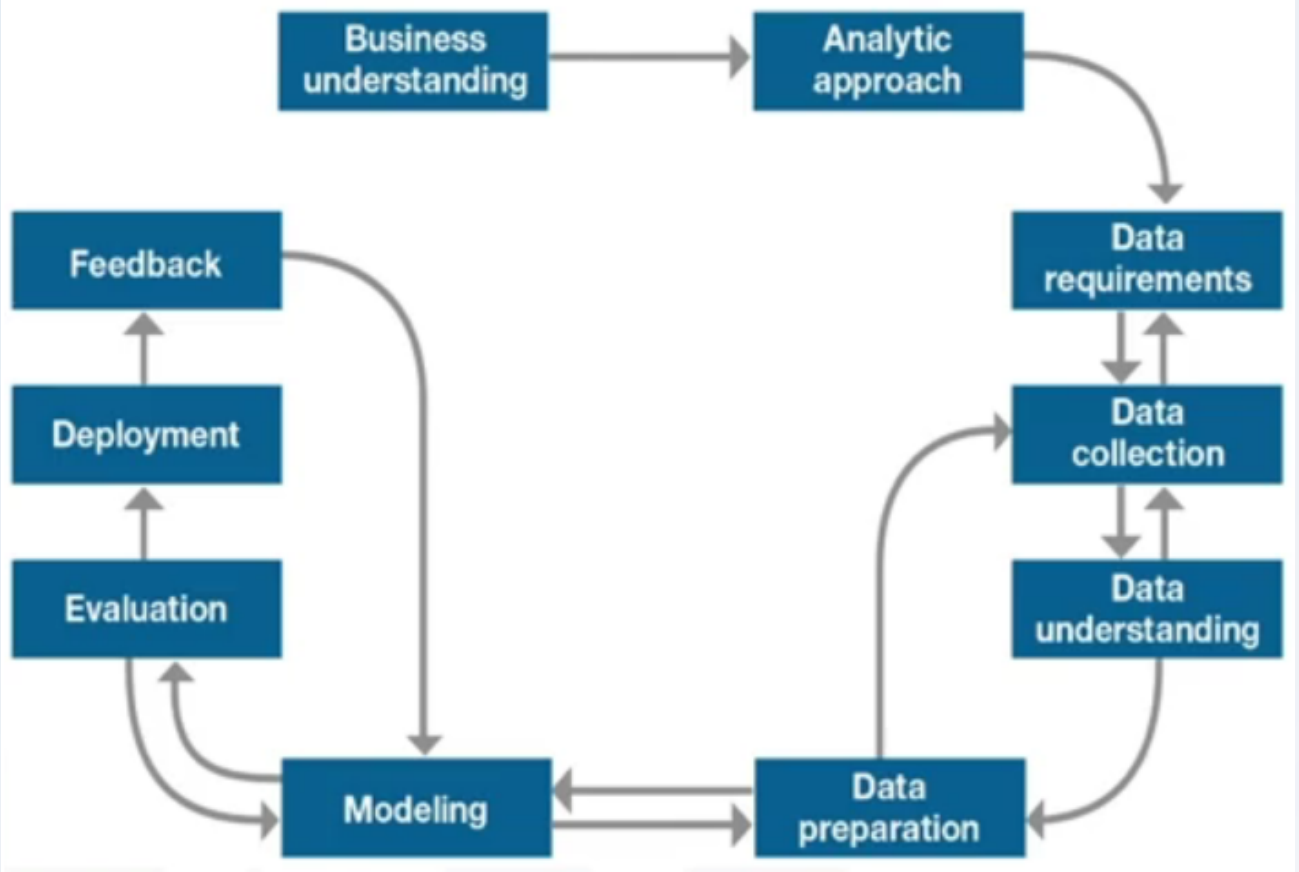
As SpaceX can reuse the first stage, the Falcon 9 rocket launch was announced at a cost of 62 million dollars, while other vendors cost more than 165 million dollars each. Thus, from the processing, analysis and modeling of data related to Falcon 9, we can predict whether the first stage will land successfully and hence the cost of each launch (Analytic Approach). To do this, some data about Falcon 9 are required (Data Requirements).

Section 1

Methodology

Data Science Methodology in a Nutshell

This flowchart shows the complete cycle of a Data Science project. In this final project, we reach the stage of evaluating machine learning models, applied to our prediction problem. The steps were mentioned throughout this presentation.



Fonte: IBM

Methodology

- Data collection methodology:
  - Data were collected through web scraping from Wikipedia and request to the SpaceX API. Once collected, the data were cleaned and filtered.
- Performing data wrangling
  - In this step, analyzing the data from the Payload Mass column some missing values were observed and this column is important for future modeling. Therefore, we completed these missing values with the arithmetic mean of this column.

Methodology

- Exploratory data analysis (EDA) using visualization and SQL
  - In this phase we get the insights about the data using some tools for visualization and SQL queries.
- Interactive visual analytics using Folium and Plotly Dash
  - Here, we can see data distribution into a map of launch sites available in the dataset to visualize successful and unsuccessful landings at each launch site. Also, we analyze the relation between the features of dataset using Seaborn library.
- Predictive analysis using classification models
  - Finally, after all these steps, we prepare the data for modeling. Then, some supervised machine learning algorithms are applied and their results compared.

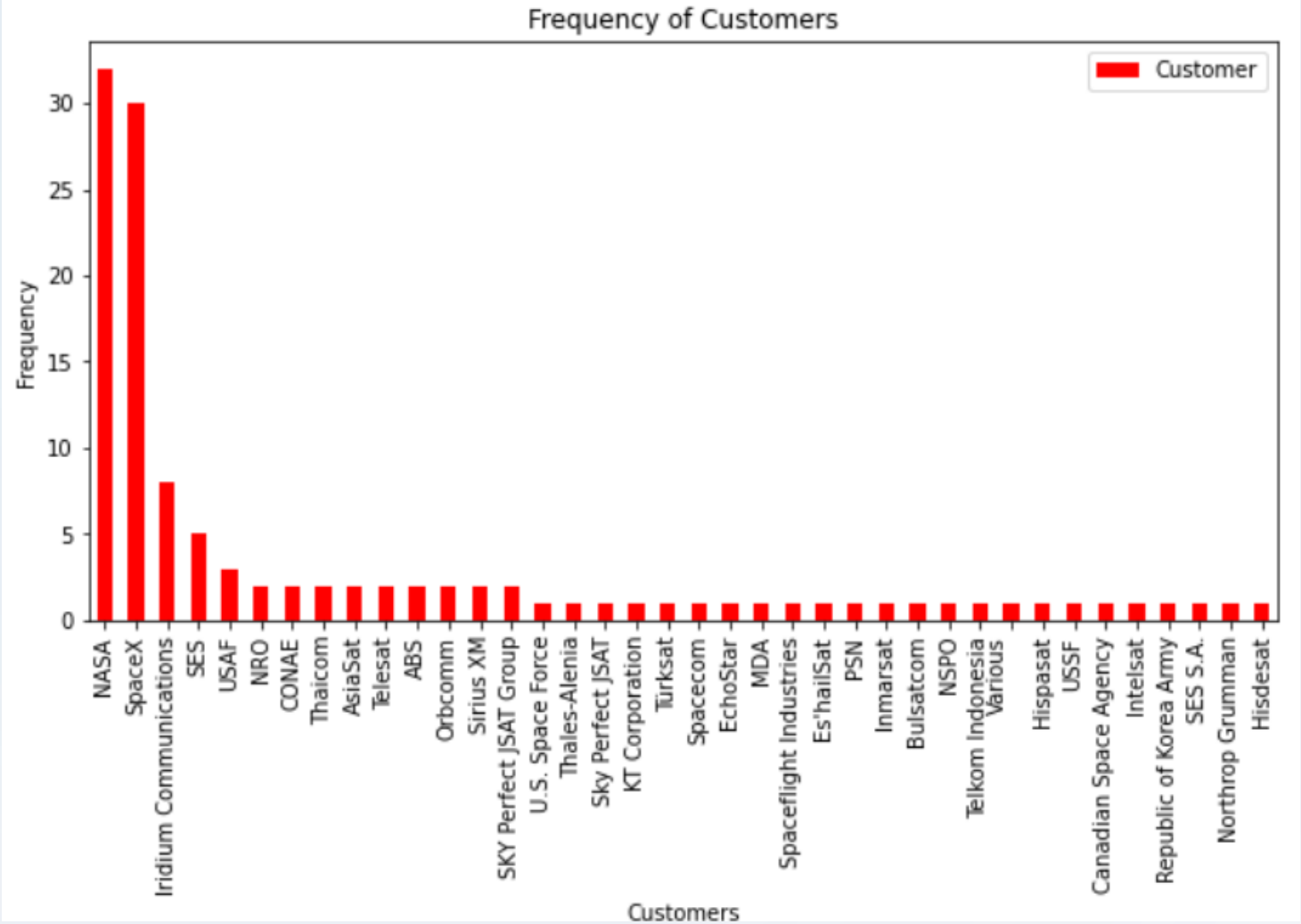
Data Collection - Scraping

The first dataset was collected through web scraping on Wikipedia and then cleaned. The Falcon 9 launch records were collected using Beautifulsoup library.

From this dataset we can analyze, for example, the most common Falcon 9 customers as shown in this chart. NASA and SpaceX are the two main customers that launch Falcon 9.

Shared notebook on GitHub:

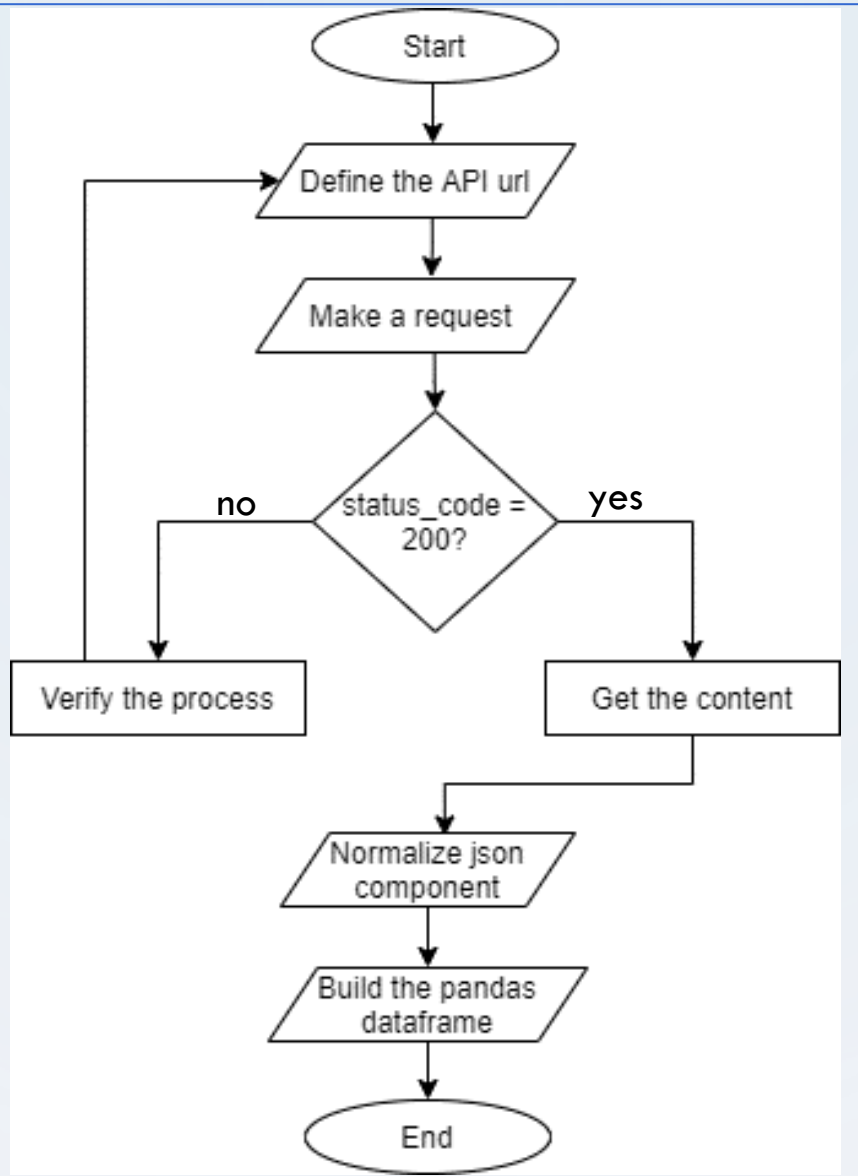
https://github.com/leilaff89/applied_data_science_capstone/blob/main/data_collection_web_scraping.ipynb



Data Collection – SpaceX API

To do the request correctly, some helper functions and Python libraries were used.

So the API was accessed via specific URL. Then, the response object content was decoded as a Json file and transformed into a Pandas dataframe using Json file normalization.



Once the dataframe has been created and cleaning, an exploration and modeling can be used as long as the data has the necessary resources.

If data is not enough for modeling, consideration should be given to collect more data or changing the source.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/data_collection.ipynb

Data Wrangling

In some cases, exploring the data characteristics, it is noticed that some data is missing or are in inappropriate formats.

In this dataset, some missing values in ‘payload mass’ have been replaced by the mean of all values reported in this feature. This process is useful for more reliable modeling.

There are other methods for data wrangling that, in this case, were not needed.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/data_collection.ipynb

Section Data Wrangling in the bottom of this notebook.

EDA with Data Visualization

Data visualization is a fundamental step to know more about the data, besides allowing us to visualize through graphs the relationship between two or more resources, and this analysis helps us to carry out a good modeling of the data later on.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/eda_data_visualization.ipynb

EDA with SQL

In many cases, the dataset to be analyzed is available as a .csv (comma separated values) file, perhaps on the internet, or even kept in a specific database.

Therefore, as a data scientist, knowing how to query a database is often necessary.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/eda_with_sql.ipynb

Build an Interactive Map with Folium

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. Also, it depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations. To analyze this, we can use Folium to build interactive maps to better visualize the data distribution at different locations on the map.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/visual_analytics.ipynb

Build a Dashboard with Plotly Dash

Dash is a Python framework created by Plotly to create interactive web applications.

The plots and interactions have been added then, users can perform interactive visual analysis on SpaceX launch data in real-time.

This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter plot chart.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/interactive_visual_analytics_dashboard.ipynb

Predictive Analysis (Classification)

To do this step, some processes are needed.

Firstly, we have to create a column called ‘class’ which is our target. The data in this column is set to '0' (zero) if the first stage landing was not successful or it is set to 1 (one) if is a successful landing.

So, we must to standardize all data values until they have the same weight for modeling.

And lastly, the data is divided into training data (80%) and testing data (20%). Training data is used to train the model and testing data is used to evaluate the model.

Then, the models are trained and hyperparameters are selected using the function GridSearchCV. This function make exhaustive search over specified parameter values for an estimator. The estimator’s parameters used to apply these methods are optimized by cross-validated grid-search over a parameter grid.

Shared notebook on GitHub:

https://github.com/leilaff89/applied_data_science_capstone/blob/main/machine_learning_prediction.ipynb

Results

After this explanation of all these steps, let's see the results obtained from them. The next slides show images from each phase of this project. All these results were generated in the notebooks provided in the previous slides (links).

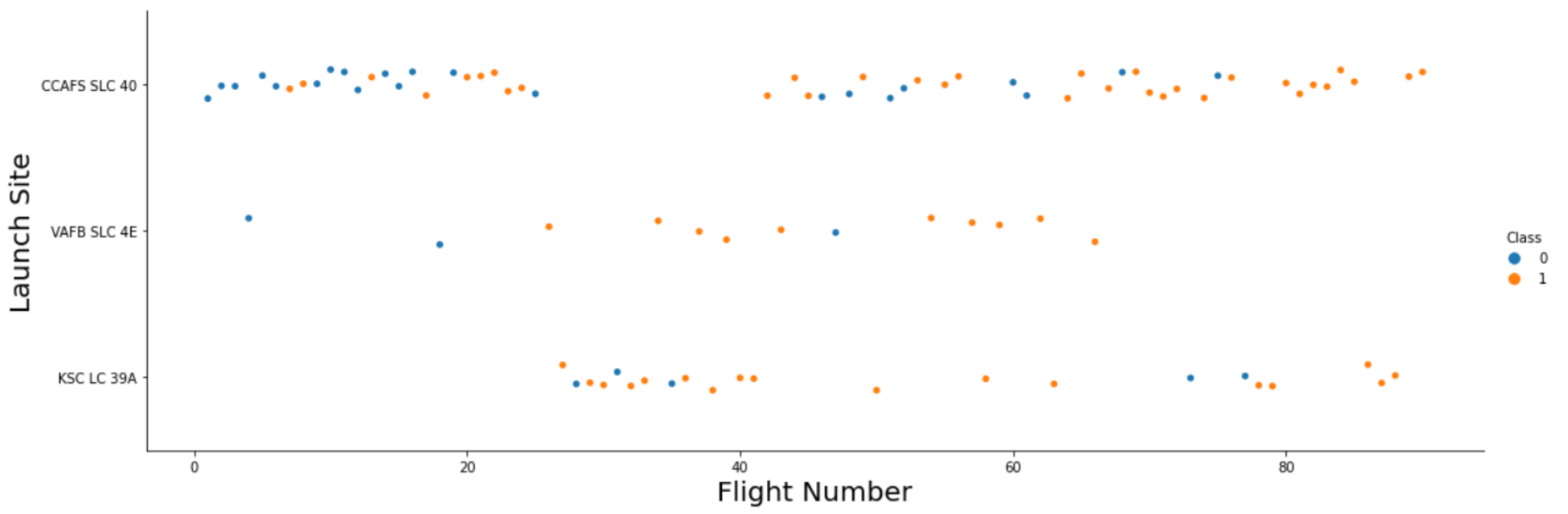
These are the next topics covered.

- Insights drawn from EDA;
- Launch Sites: Proximities Analysis;
- Build a Dashboard with Plotly Dash;
- Predictive analysis (Classification) results.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site



In this graph, it is possible to visualize the relationship between Launch Sites and the success rate over the number of flights. It is clear that the success cases represented by the class 1 increase in same ratio as the number of flights. It can represent improvements made on Launch Sites or in another feature.

Payload vs. Launch Site

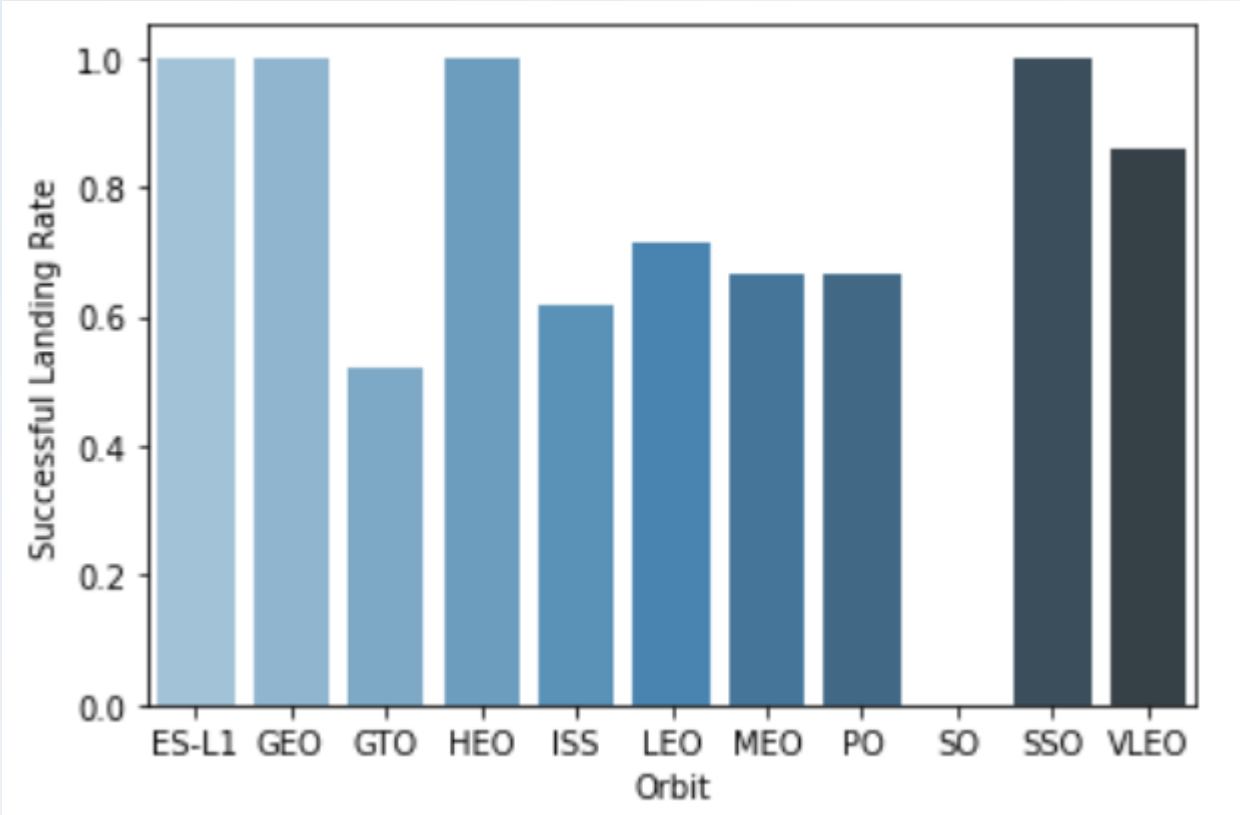
In the CCAFS SLC 40 the distribution of successful or unsuccessful landing is balanced between payload mass from 0 to 8000, but is more successful in higher payload mass. In the VAFB SLC 4E Launch Site, most of landing was successful regardless of payload mass. As well as on the VAFB SLC 4E launch site, KSC LC 39A also had the most successful landings, but there is a detail when the payload mass is close to 6000 kg, because in this range the landings were mostly unsuccessful.

Success Rate vs. Orbit Type

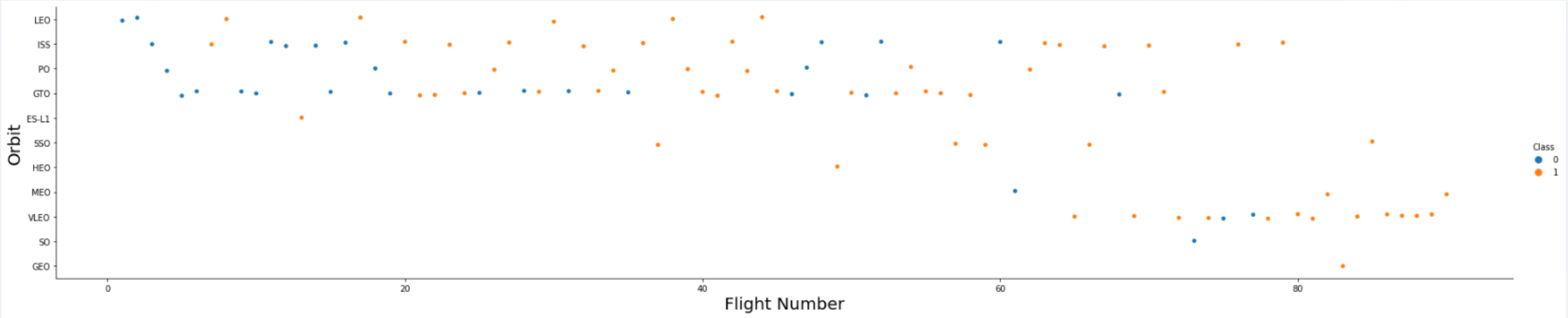
In this graph, we can visualize the most and least successful orbits.

ES-L1, GEO, HEO and SSO had the best success rates in this case, and GTO the worst rate.

The SO orbit brought no significant results.

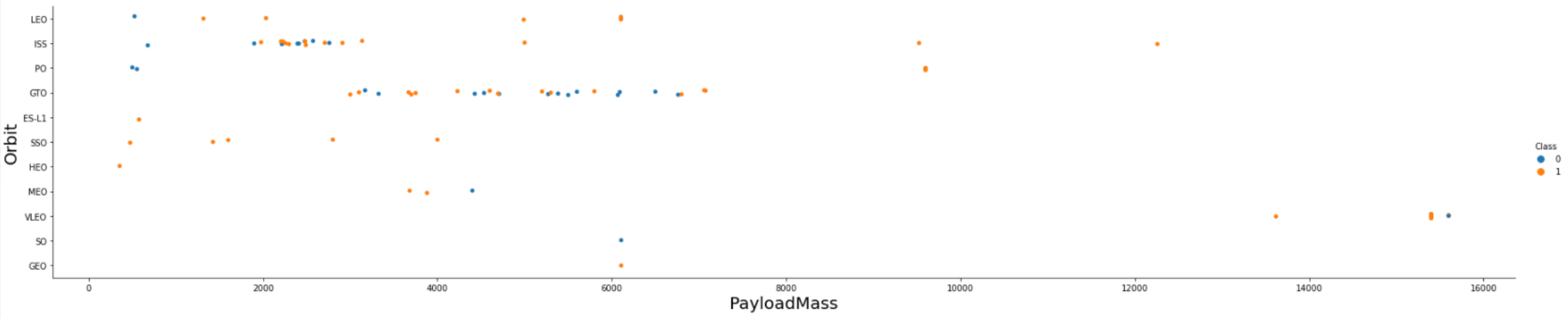


Flight Number vs. Orbit Type



It is possible to see that in LEO orbit the success rate appears related to the number of flights, on the other hand, there seems to be no relationship between flight number when in GTO orbit.

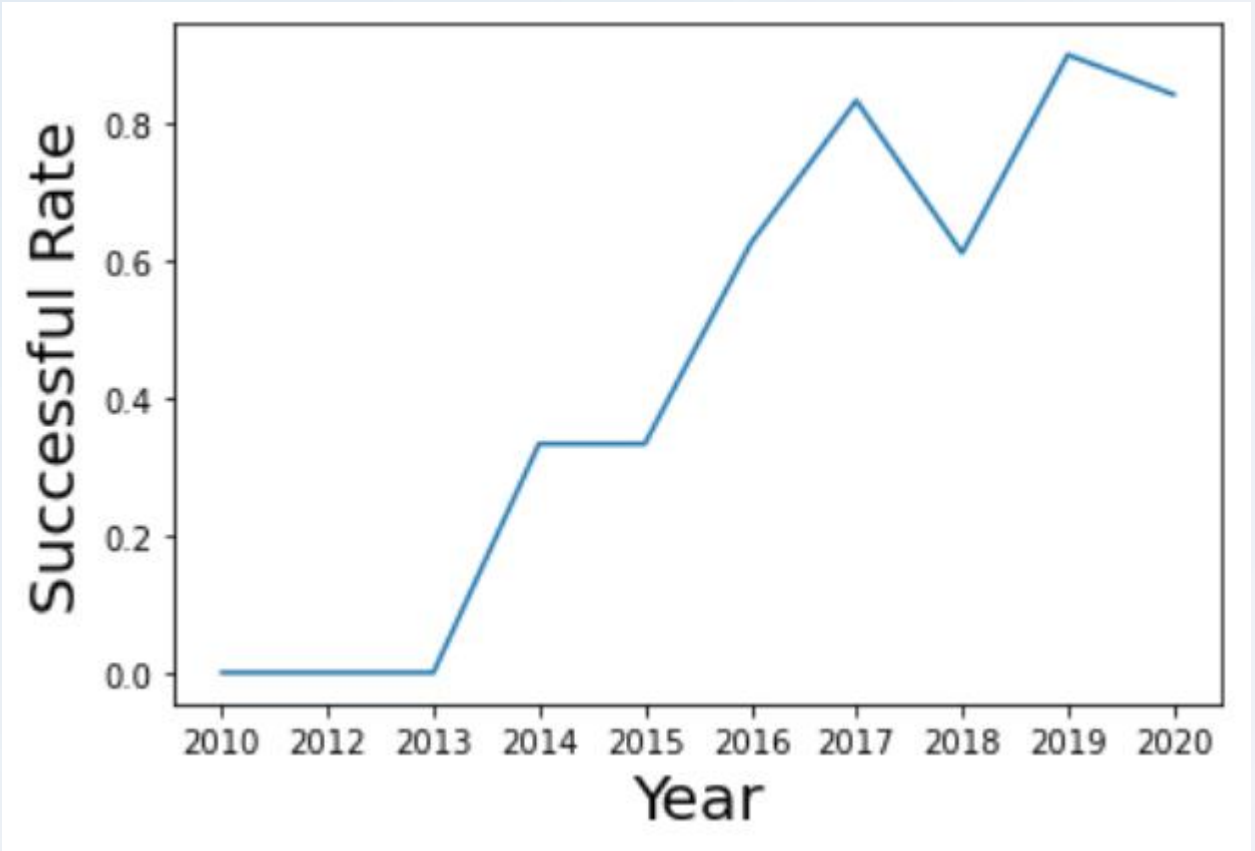
Payload vs. Orbit Type



We can observe that heavy payloads have negative influence on GTO orbits and positive on PO and LEO orbits. However, this feature is indifferent to ES-L1 and SSO orbits.

Launch Success Yearly Trend

We can visualize that the success rate since 2013 kept increasing till 2020. But, there is a significant decrease between the years 2017 and 2018, and it started to increase again from 2018 to 2019.



EDA with SQL – Results

SQL stands for “Structured Query Language”, or “Structured Query Language” in English. Briefly, it is a programming language for dealing with relational (table-based) databases. It was created for multiple developers access and modify a company's data simultaneously in an uncomplicated and unified way.

This allows us to perform custom queries, changes, inserts and remove data. With this tool it is possible to explore data from one or more tables in a database.

The next slides shows exploratory data analysis done on the SpaceX dataset maintained in IBM DB2.

In this step, using different queries, we can explore features to better understand the dataset.

EDA with SQL – Results

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

Returns five records where launch sites begin with the string 'CCA'

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

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

avg_payload_mass_f9
2928

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

1
45596

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

EDA with SQL – Results

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

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

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

List the total number of successful and failure mission outcomes

1
2015-12-22

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

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

EDA with SQL – Results

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

Display the names of the booster versions which have carried the maximum payload mass.

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

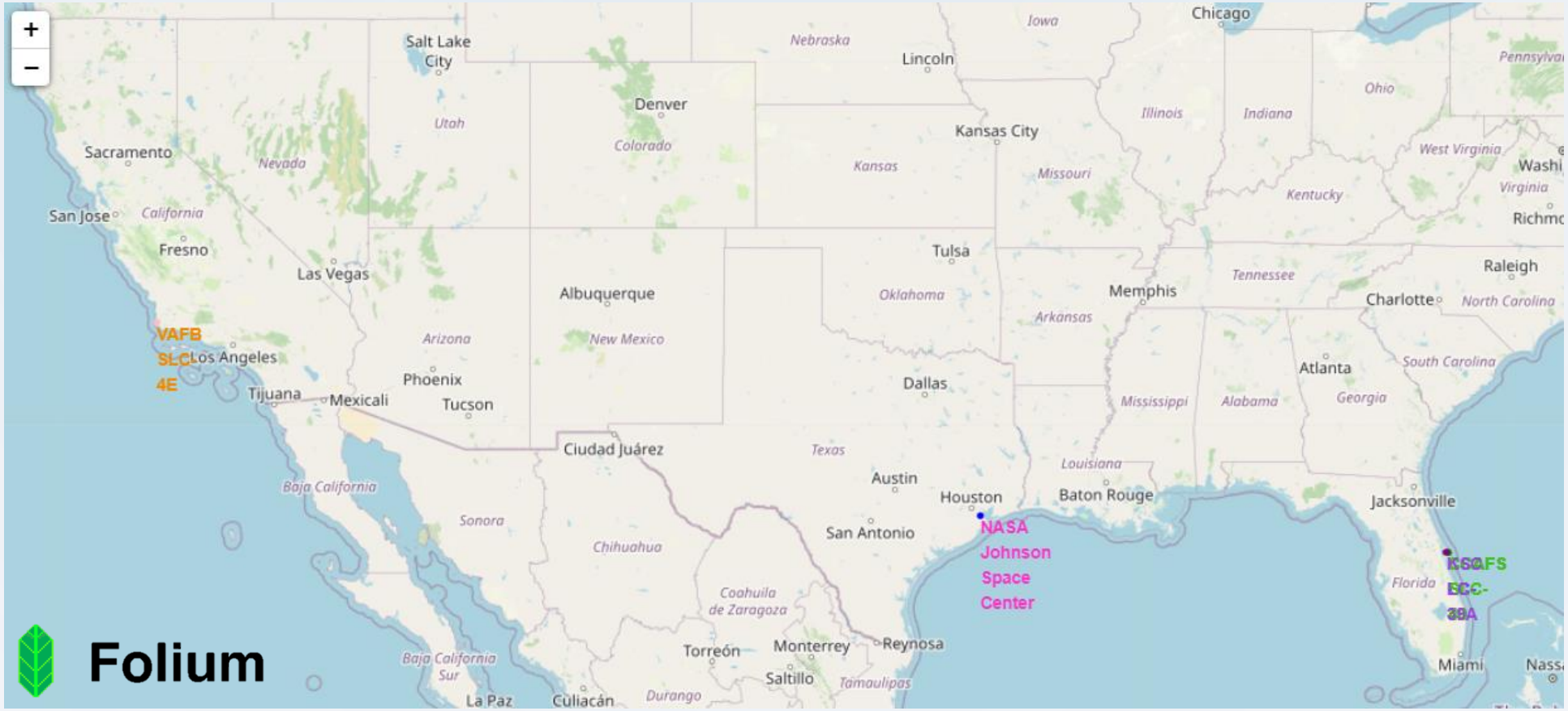
Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

Section 4

Launch Sites Proximities Analysis

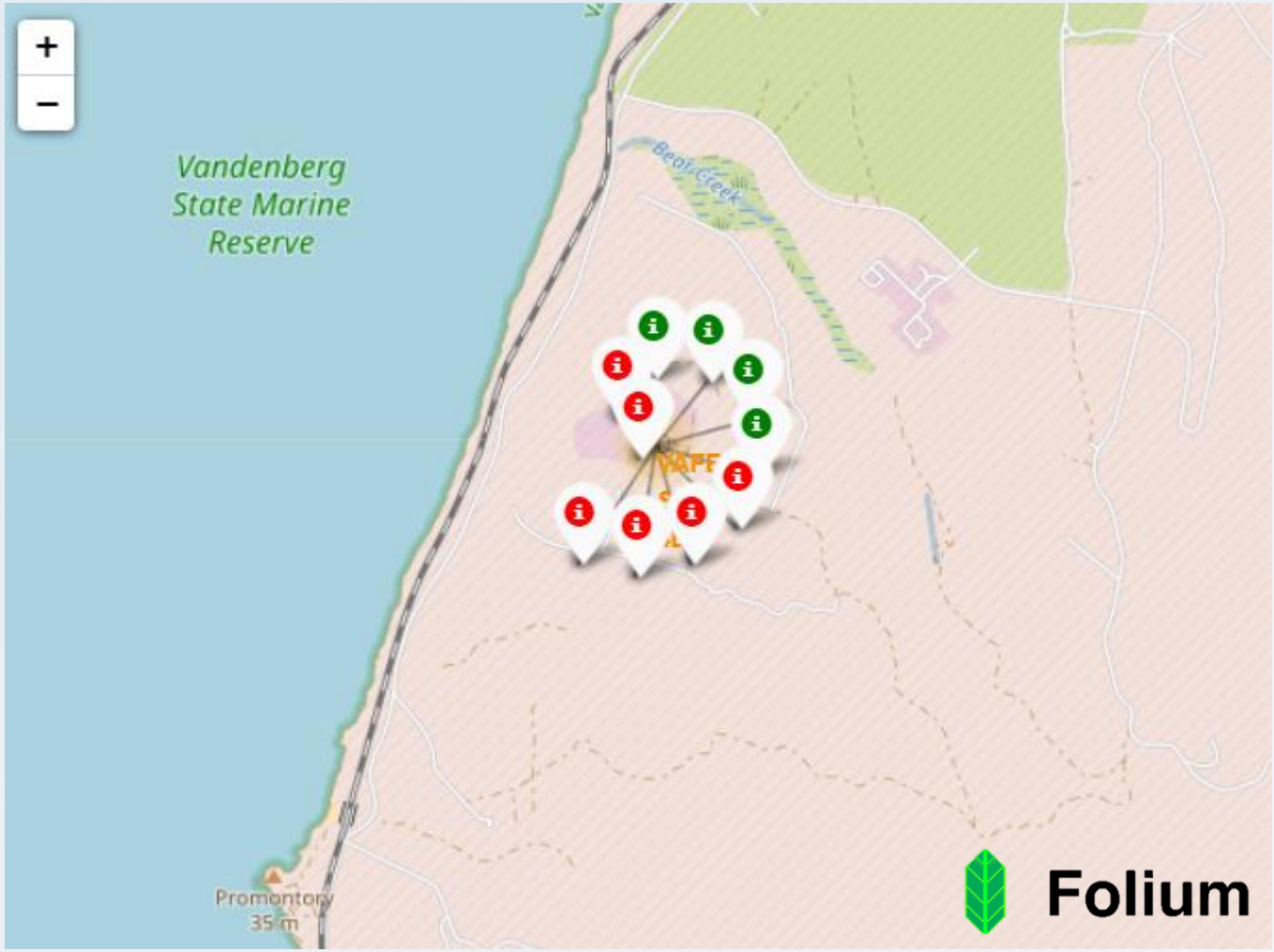
Launch Sites on the Map

Geographic coordinates are just numbers that can not provide any intuitive insight into where the launch locations are, unless you are very good at geography. Therefore, creating a map we can visualize these places through their coordinates. On the map below, the launch sites are marked.

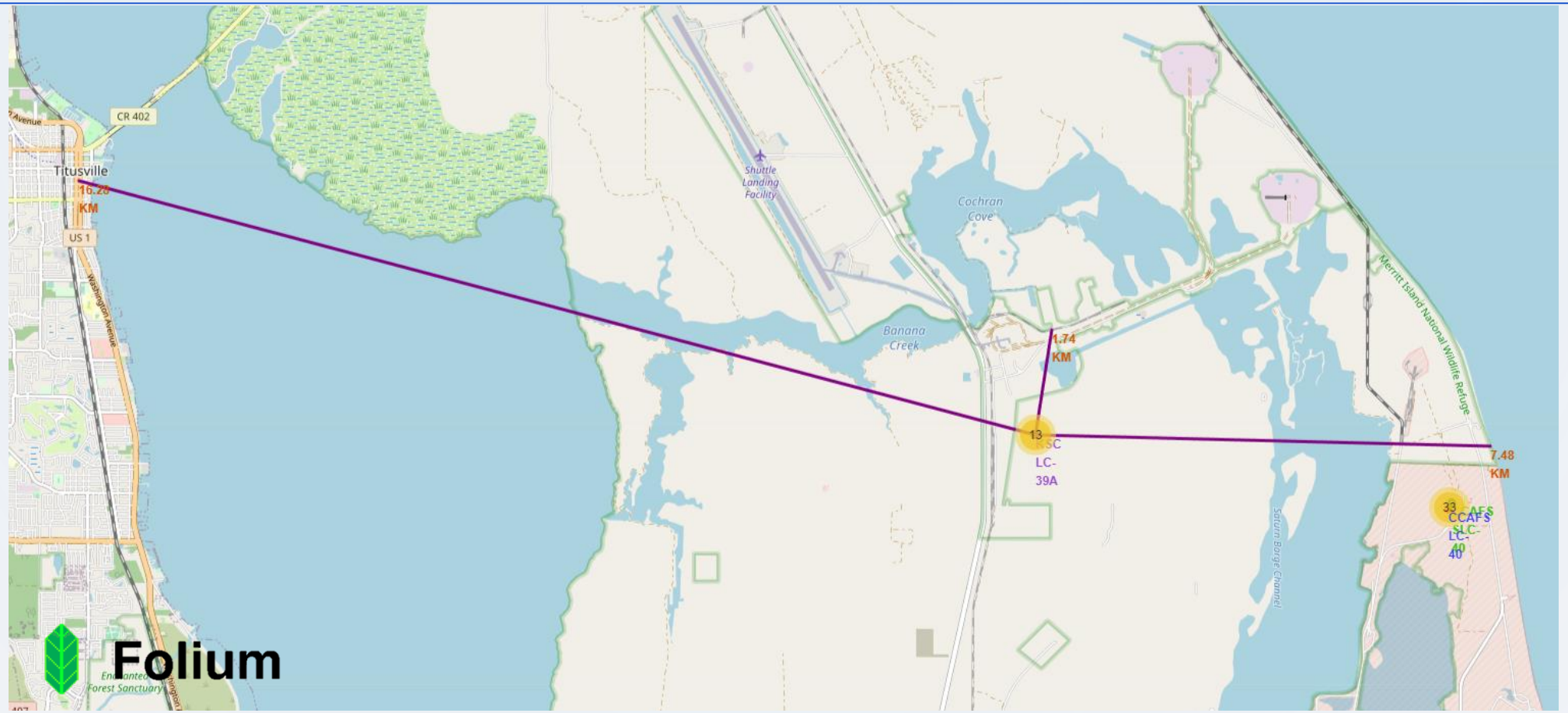


Launch clusters

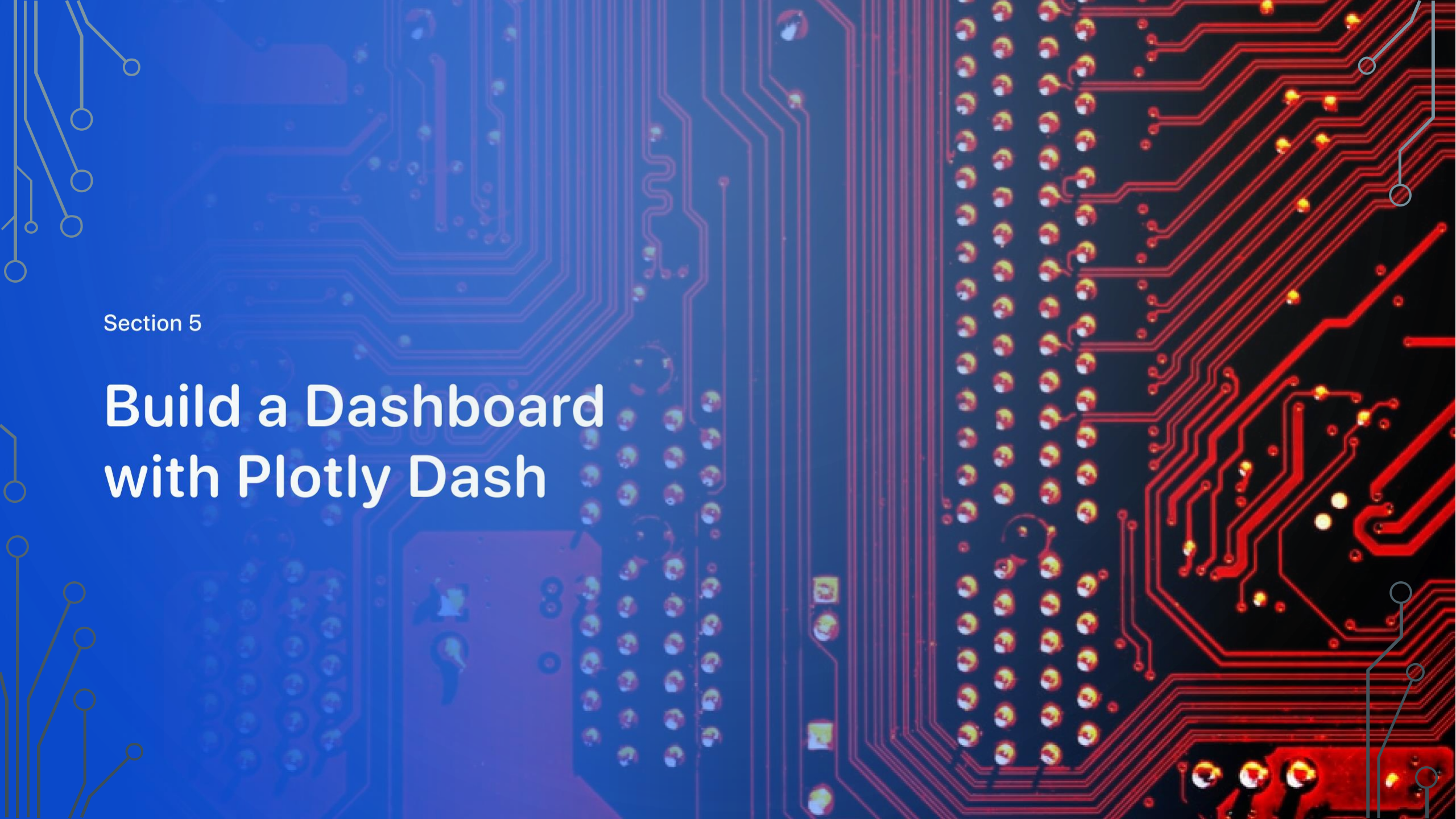
In this map, we can visualize the successful and unsuccessful launches through the color of pins in the cluster of each launch site. The green color is successful launches and the red ones are unsuccessful.



Measuring Distances

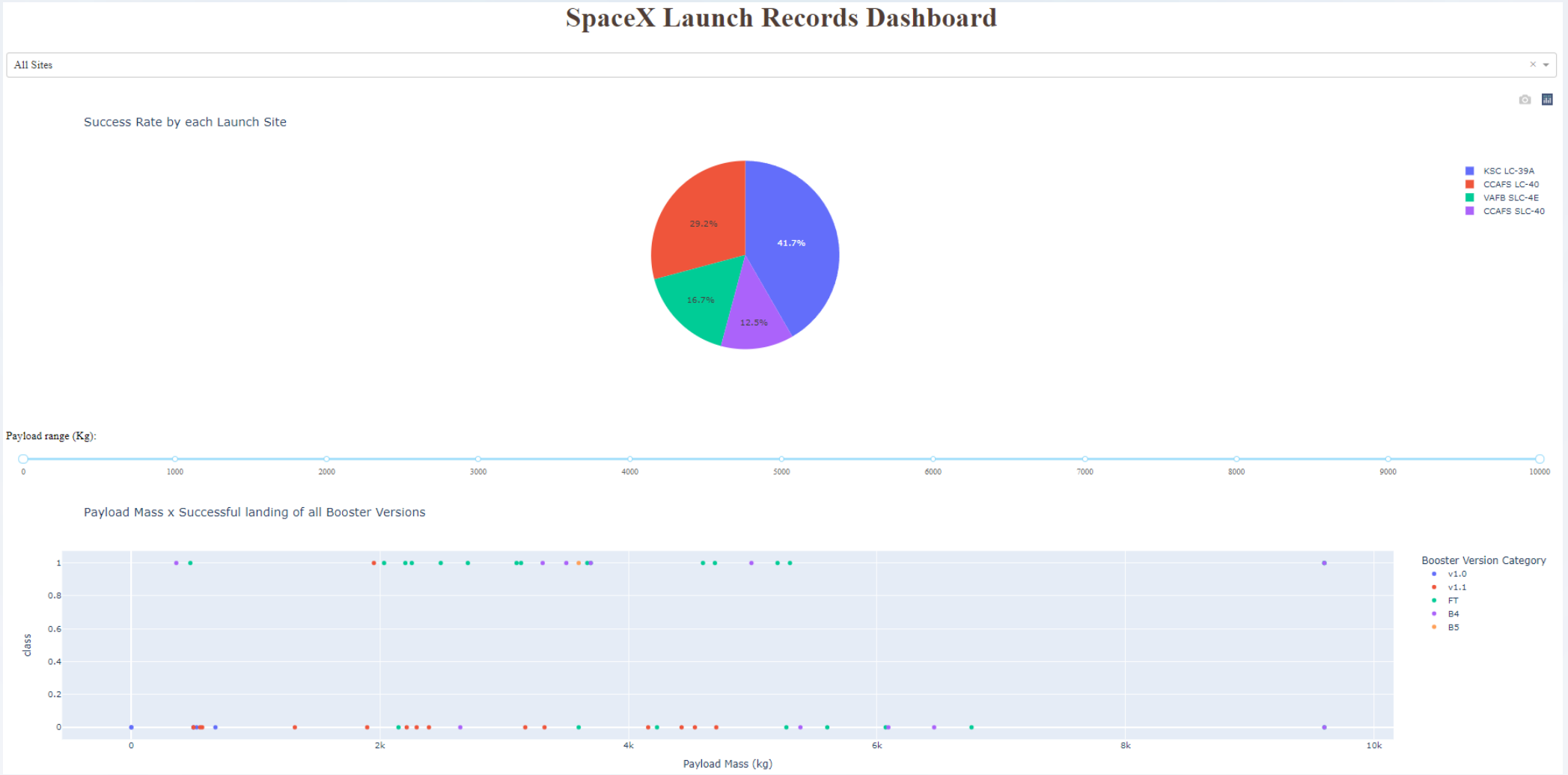


In this map, using Folium's polyline function, we can explore and analyze the proximities of each launch site. We can see that this specific launch site is close to the railroad and the coast, but far from the cities.


Section 5
Build a Dashboard with Plotly Dash

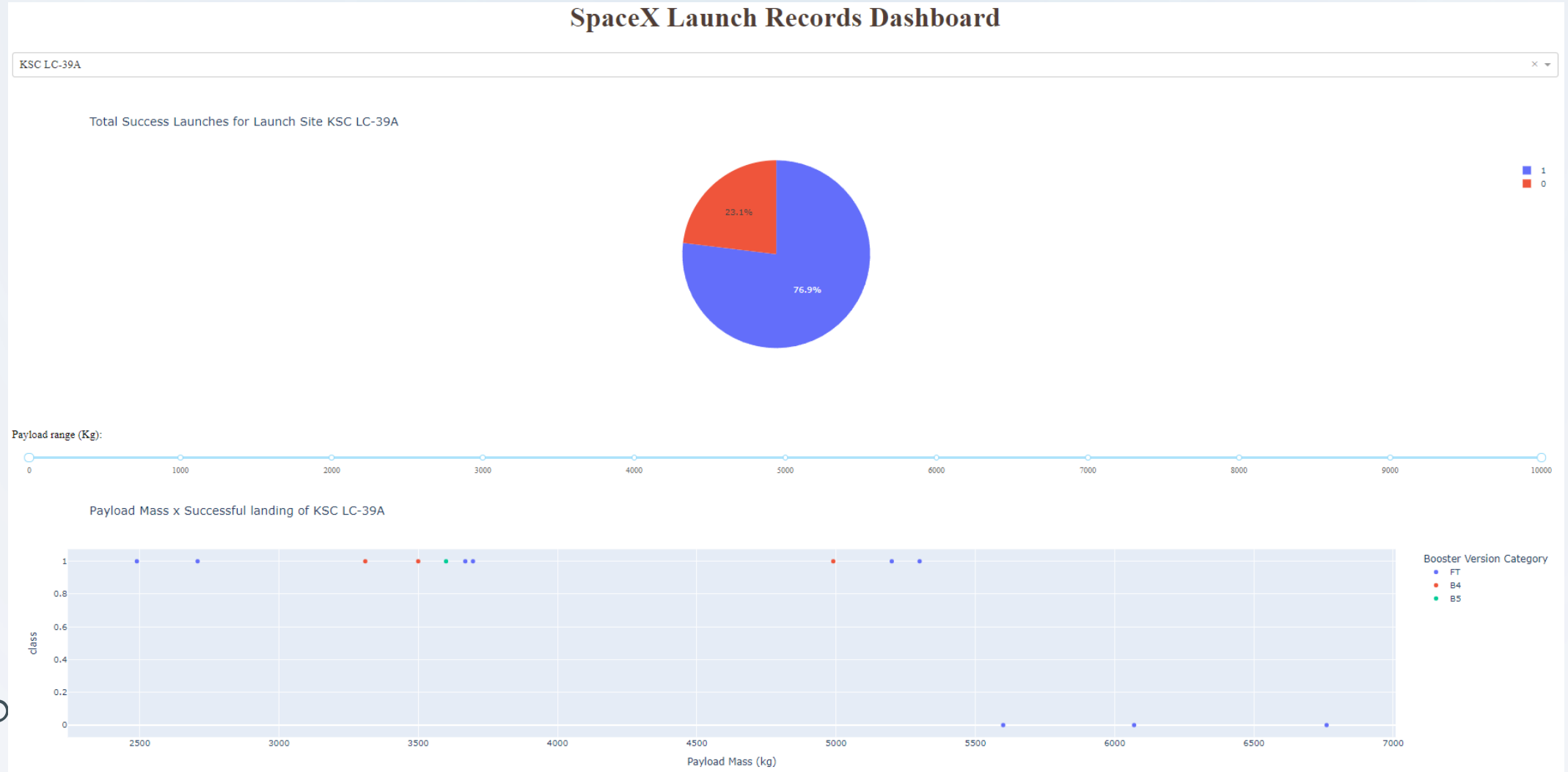
SpaceX Launch Records Dashboard – All Launch Sites

In this dashboard you can view an analysis of all launch sites. The first pie chart shows the success rates for each launch site and below, you can select a specific payload mass range to check the successful launches of each booster version.



Dashboard – Launch Site with highest launch success ratio

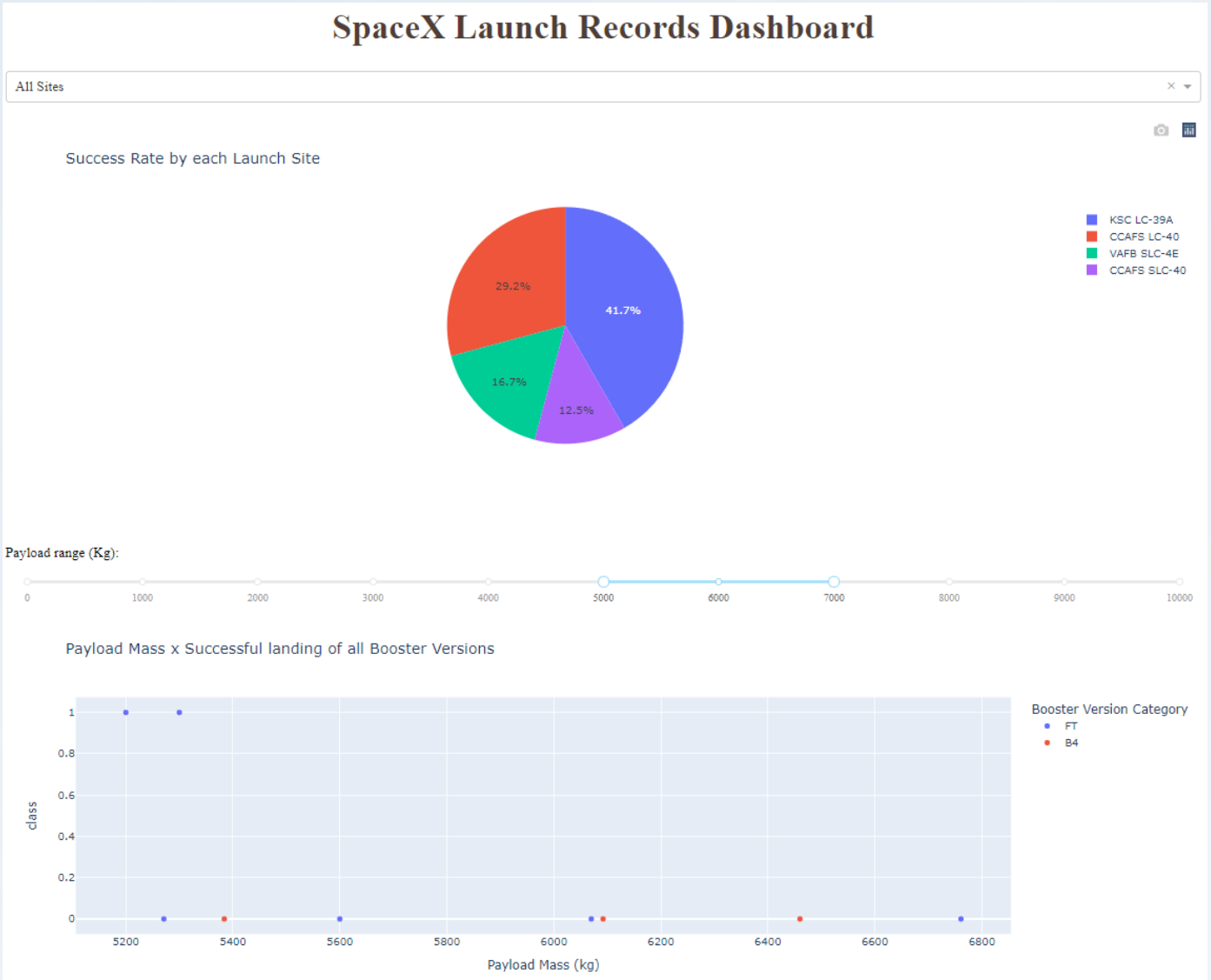
Now, in the same dashboard, we filter the results to the launch site with the highest success rate (KSC LC - 39A)



Dashboard – Filtering by Payload Mass

Now, let's see the results of all launch sites, but filtering with payload mass range slider.

This function allow us to analyze different ranges of payload mass of each or all launch sites.



Section 6

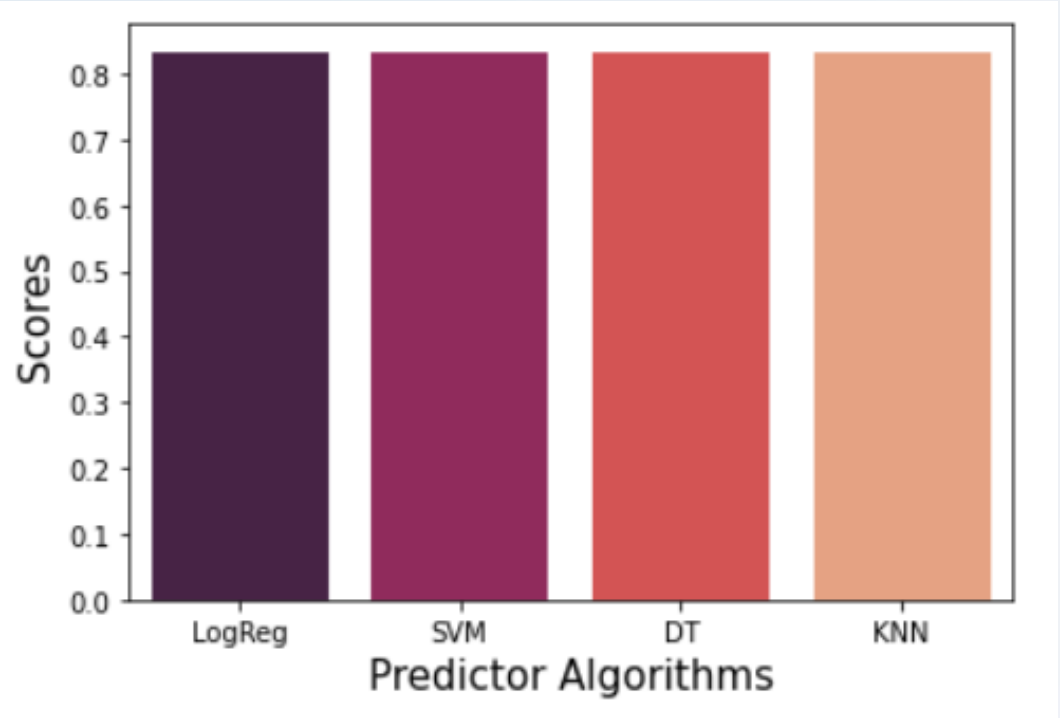
Predictive Analysis (Classification)

Classification Accuracy

Before applying a machine learning algorithm to predict successful or unsuccessful launch, it is necessary to split the data into training and testing data to validate the model. Then, the models are trained and the hyperparameters are selected using the GridSearchCV function.

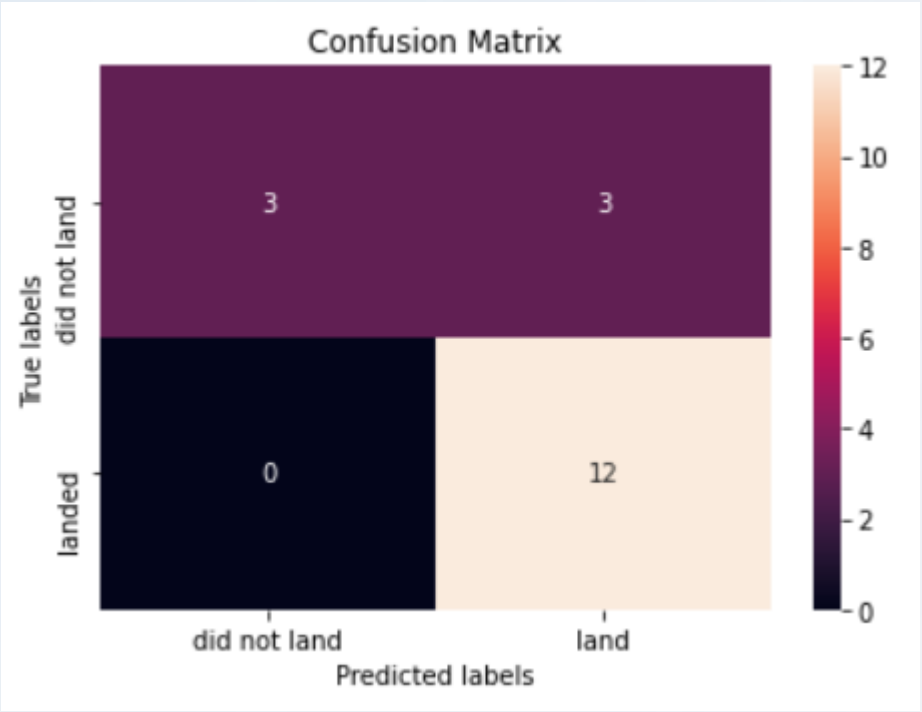
Logistic Regression, SVM, Decision Trees and KNN algorithm were applied to find the best model. The scores for each algorithm are shown in the bar graph on the right.

In this case, all models had the same score.



Confusion Matrix

According to the confusion matrix, in the eighteen actual launch attempts, six times the launch was unsuccessful and twelve times it was successful. The algorithms predicted that three times the attempt failed and fifteen times it was successful. There is a small error, but the prediction score was good for all models.

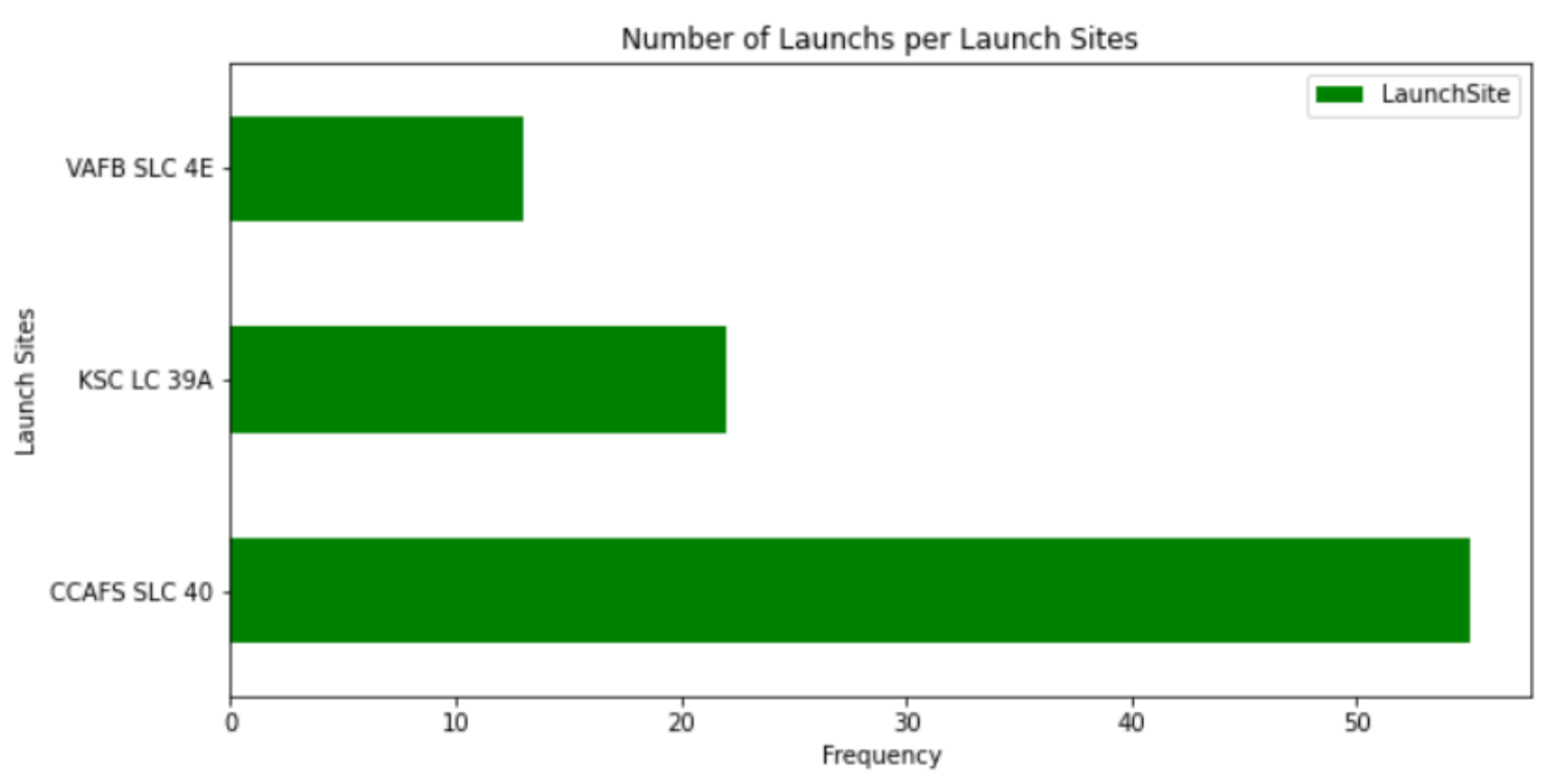


Conclusions

- It is important to follow the sequence of steps to achieve the main objective of the project, for example ETL, EDA, modeling and implementation;
- There are many ways to get a dataset, such as scraping websites, downloading from the Internet, querying a database or even doing it from scratch;
- Doing proper the data wrangling and exploratory data analysis steps, using visualizations such as charts, maps, or arranging a more complete visualization on a dashboard is essential to better understand the data;
- The four algorithms applied in this project to predict the successful launching of Falcon 9 scored well in this case.

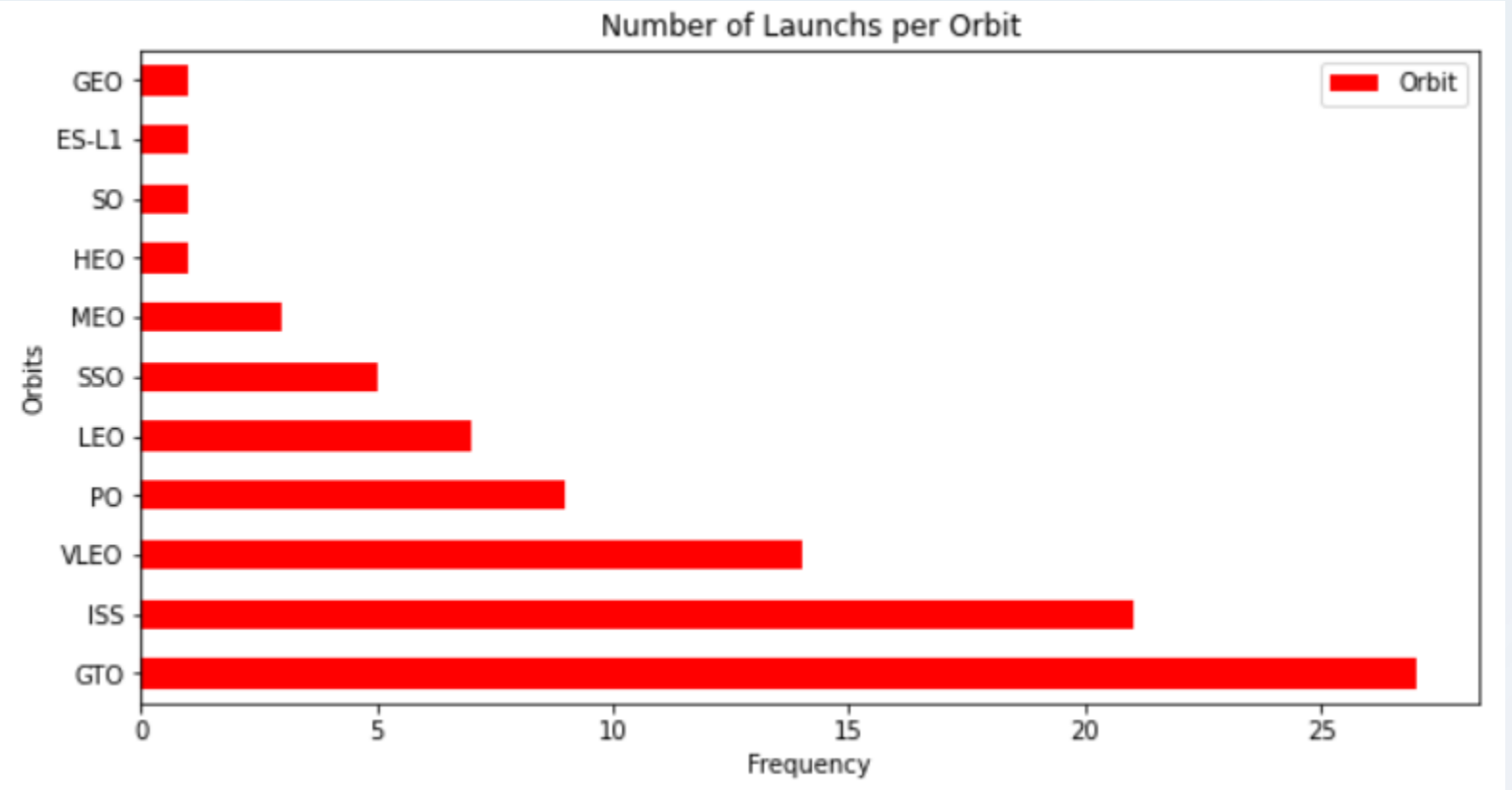
Appendix

Some extra bar plots obtained in the exploratory data analysis.



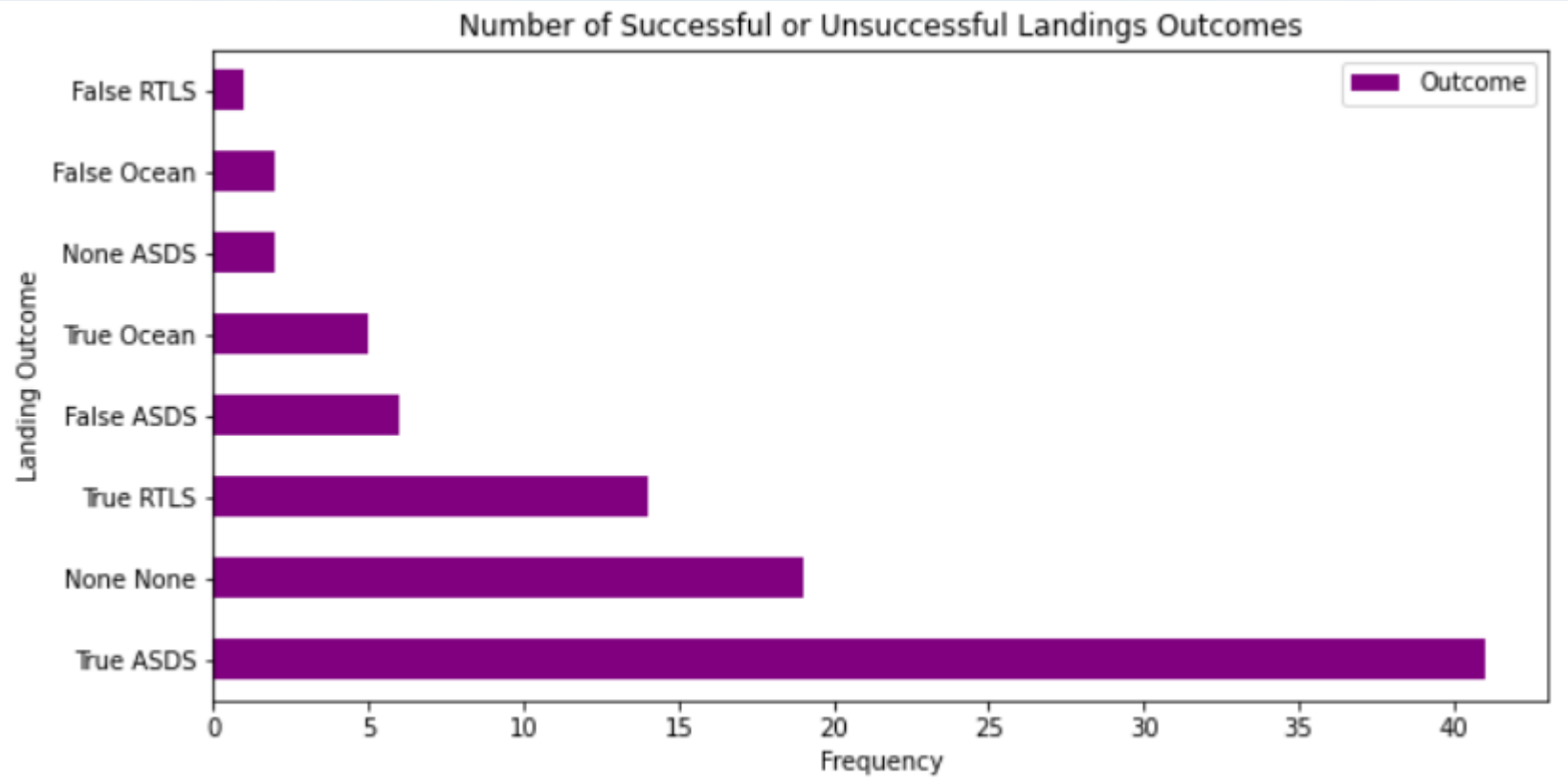
Appendix

Some extra bar plots obtained in the exploratory data analysis.



Appendix

Some extra bar plots obtained in the exploratory data analysis.



Thank you!