IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Rodrigues
18/04/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collect and explore data;
  - Visual Analytics;
  - Dashboard Report;
  - Predictive Analyses.
- Summary of all results
  - Relations between several parameters and landing success;
  - Predictive models to predict the success of a given mission.

Introduction

- Project background and context

  SpaceX developed the Falcon 9 rocket that can reuse the first stage. This technology decreases the waste left on orbit and can save a lot of money in each rocket launch.

- Problems you want to find answers

  - This project goal is to analyze data from previous Falcon 9 launches in order to determine which factors impact the rocket’s first stage to land successfully;

  - Predictive Models will also be developed in order to predict the success of landing the first stage on any given mission.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - SpaceX API;
    - Web Scraping.
- Perform data wrangling
    - Transformation of different types of outcomes into successful or not successful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Slip the data into Train and Test sets, fit the models with the training data set and then test the model with the testing data set.

Data Collection

Data was acquired in two different ways.

First we made a get request to the SpaceX API in order to extract the data.

Then we used the BeautifulSoup package to do web scrapping on a Wikipedia page in order to collect Falcon 9 historical launch records.

Filter the dataframe to only include Falcon 9 launches

Data Collection – SpaceX API

Data collection with the SpaceX REST API followed the next steps:

- Request and parse the SpaceX launch data using the GET request;
- Filter the DataFrame to only include Falcon 9 launches;
- Dealing with missing values;
- Save DataFrame into CSV file.

Github URL:

https://github.com/JPCR93/Capstone-Project/tree/main/Hands-on%20Lab:%20Complete%20the%20Data%20Collection%20API%20Lab

Data Collection - Scraping

Data collection using Web Scrapping followed the next steps:

- Request the Falcon9 Launch Wiki page from its URL;
- Create a Beautiful Soup object;
- Extract all column/variable names from the HTML table header;
- Create a data frame by parsing the launch HTML tables;
- Save the DataFrame into a CSV file.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/tree/main/Hands-on%20Lab:%20Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab

Data Wrangling

The Data Wrangling process was the following:

- Check for null values and data types in the DataFrame;
- Calculate the number of launches on each site;
- Calculate the number and occurrence of each orbit;
- Calculate the number and occurrence of mission outcome per orbit type;
- Create a landing outcome label from Outcome column
- Save the DataFrame into a CSV file.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/tree/main/Hands-on%20Lab:%20Data%20Wrangling

EDA with Data Visualization

In order to visualize how two variables would affect the launch outcome we plotted the following **scatter plots**, where a color code was used in each point to symbolize that launch outcome:

-> Payload Mass and Flight Number;

-> Launch Site and Flight Number;

- > Launch Site and Payload Mass ;

-> Orbit and Flight Number;

-> Orbit and Payload Mass;

We also plotted a **bar chart** to visualize the success ratio in each orbit type and a **line plot** to show the evolution of the success rate trough the years.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/blob/main/Hands%20on%20Lab:%20Complete%20the%20EDA%20with%20Visualization%20lab/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

The queries performed were:

- Display the names of the unique launch 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, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/blob/main/Hands-on%20Lab:%20Complete%20the%20EDA%20with%20SQL/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

First, we marked the different launch sites on the folium map with circles and added text markers with each launch site name.

Then we used colour labelled markers to show every launch record and its outcome, green for when it was successful and red when it failed. Since different launch records have the same location, their correspondent launch site, we used marked clusters to simplify the map visualization.

After this we added the mouse coordinates to the folium map, so that we could have the coordinates of any point just by pointing with the mouse cursor.

Lastly some lines and text were added in order to show the distance to some close points of interest, such as the coastline, railways, cities, etc..

GitHub URL:

https://github.com/JPCR93/Capstone-Project/blob/main/Hands-on%20Lab:%20Interactive%20Visual%20Analytics%20with%20Folium%20lab/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

In our dashboard we had a dropdown list where the user could select on of the different launch sites or all of them.

This option alone would create a pie chart. When we had all launch sites selected, the pie chart will show the number of successful launches from each site. When one launch site is selected in the dropdown list, the pie chart would show its success rate.

We also added a range slider where the user can select a payload range to investigate.

The sider plus the dropdown down list option will create a scatter plot to visualize the correlation between the chosen payload mass range and launch success from each launch site or all of them at once.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/blob/main/Hands-on%20Lab:%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash/spacex_dash_app.py

Predictive Analysis (Classification)

In order to predict the outcome of future launches we built some predictive classification models. We used logistic regression, support vector machine and tree decision models to find the best one. This was done through the following steps:

- Separate and standardize the independent from the dependent (target) variables;
- Separate our data into training and testing data sets:
- Create a GridSearchCV object to find the best performing parameters for each model. This was done for every model type, using their specific parameters;
- Fit the model with the training data set;
- Check which were the best parameters for each model;
- Evaluate each model checking its accuracy on the test data set;
- Then we compared the accuracy scores of each model in order to chose the one with the higher score as the best one.

GitHub URL:

https://github.com/JPCR93/Capstone-Project/blob/main/Hands-on%20Lab:%20Complete%20the%20Machine%20Learning%20Prediction%20lab/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

Results: Exploratory data analysis results

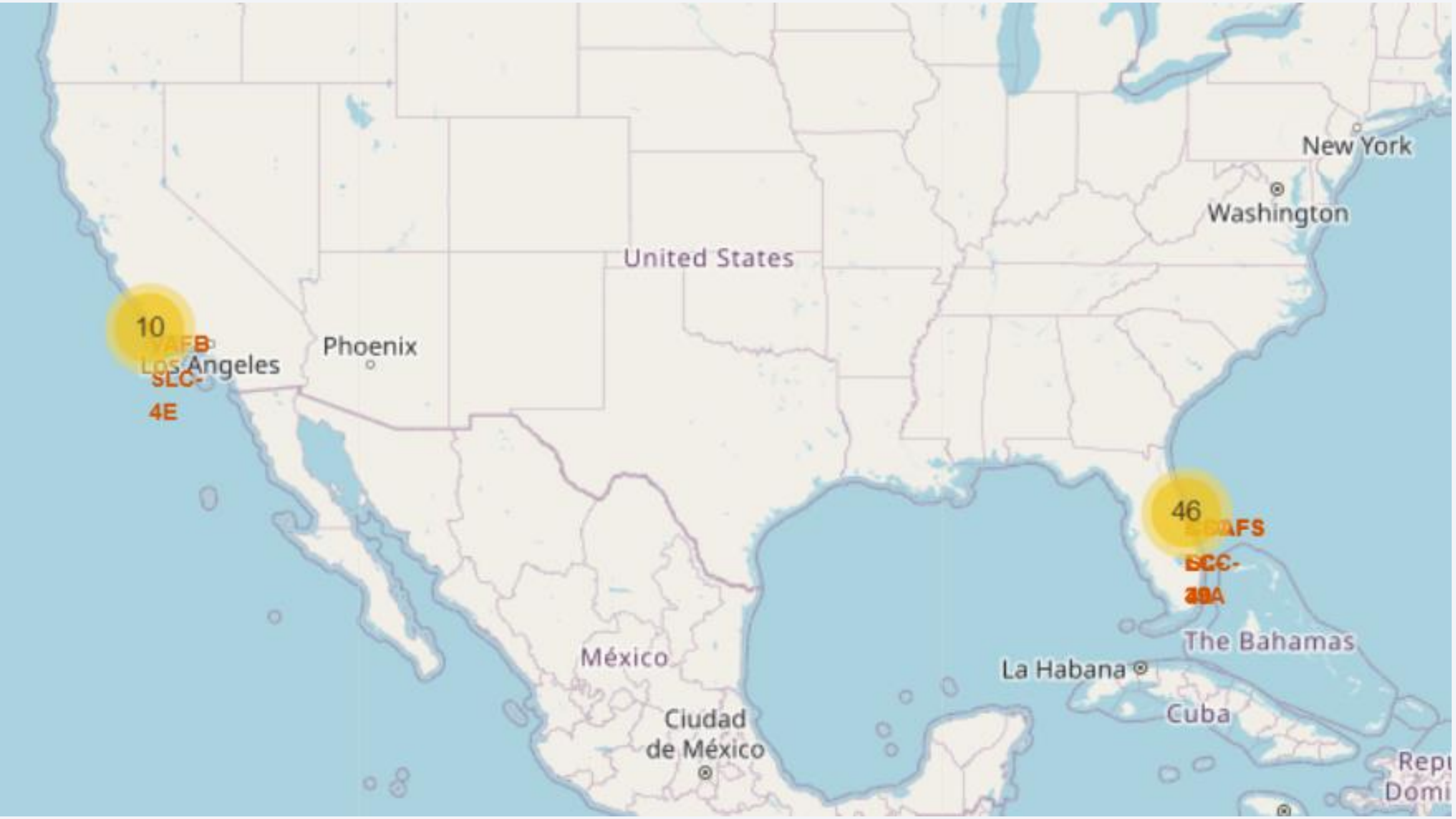
Using SQL we found out:

- There are 4 different launch sites;
- The total Payload Mass carried by any costumer;
- The average Payload Mass carried by any booster version;
- When the first successful landing outcome in ground pad was achieved;
- The total number of successful and failure mission outcomes;
- The names of the booster versions which have carried the maximum payload mass;
- The failure landing outcomes in drone ship for the year 2015.

Using visualization tools we fount out that:

- As the flight number increases, the first stage is more likely to land successfully;
- The more massive the payload, the less likely the first stage will return;
- Different launch sites have different success rates;
- For the VAFB-SLC launch site there are no rockets launched with a payload mass greater than 10000;
- The ES-L1, GEO, HEO and SSO orbits have a success rate of 100% and the SO orbit has a success rate of 0%;
- In LEO orbit the success appears related to the number of flights;
- In the GTO orbit the success appears to have no relationship with the flight number;
- The success rate has increased over the years.

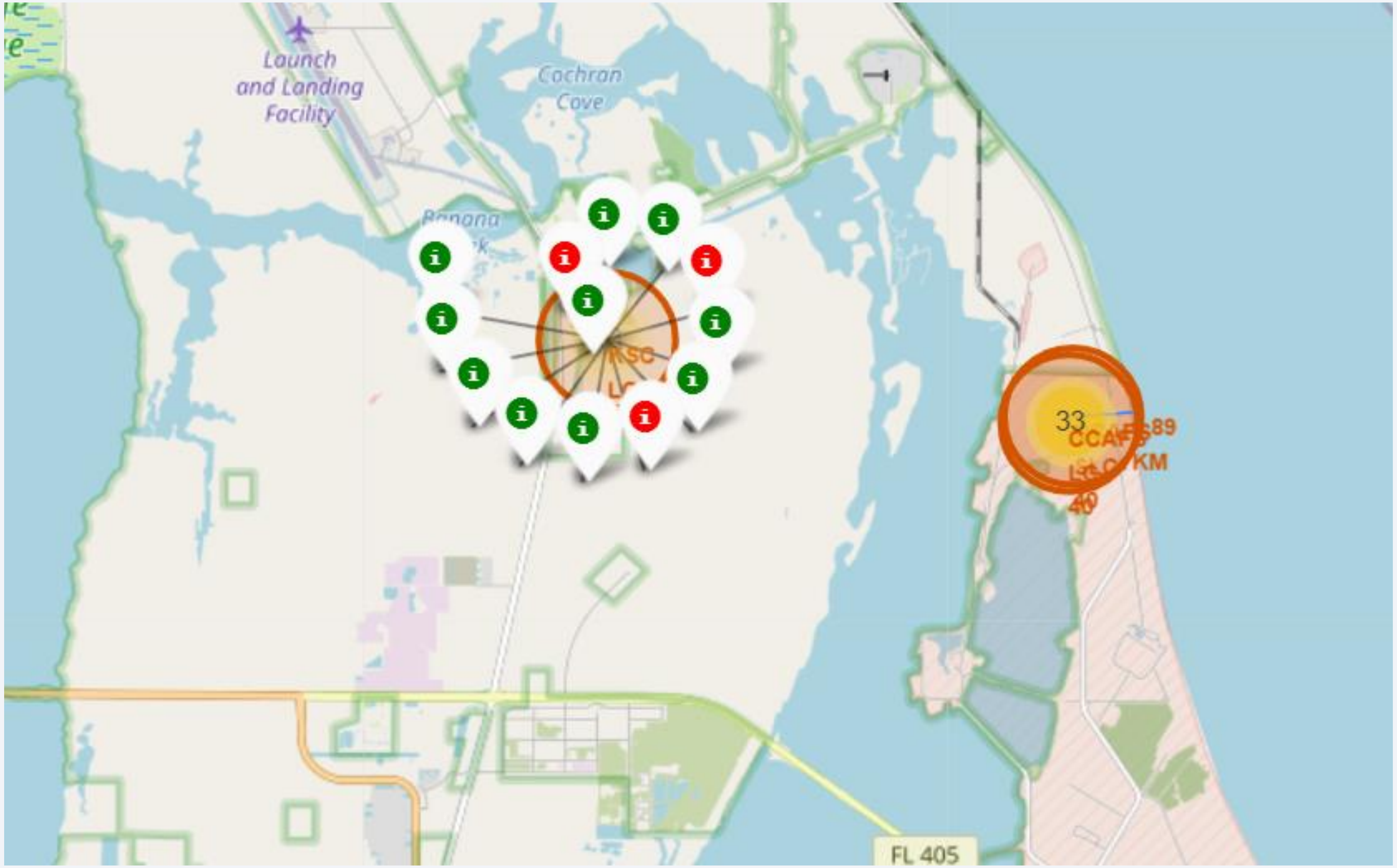
Results: Interactive analytics demo in screenshots



Here we see that the 4 launch sites are distributed in two locations, one on the west coast and the other on the east coast.

We can also see that from the 56 launches records analysed, 10 were deployed on the west coast and 46 on the east coast.

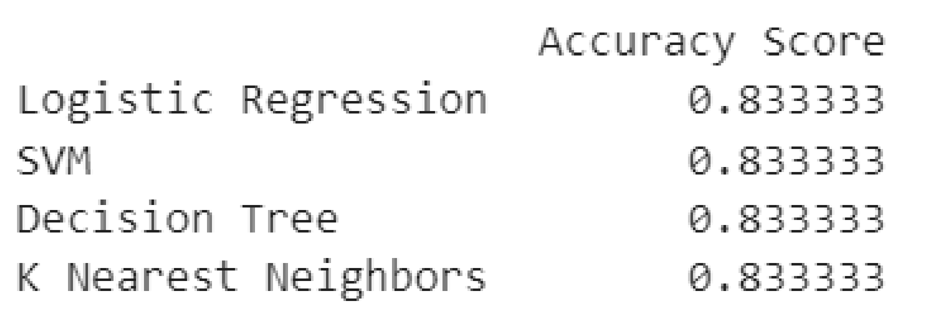
Results: Interactive analytics demo in screenshots



This is an approximation on the east side cluster from the last figure.

Here we can visualize the ratio from successful and unsuccessful missions from each launch site as well as the interest points in its proximities.

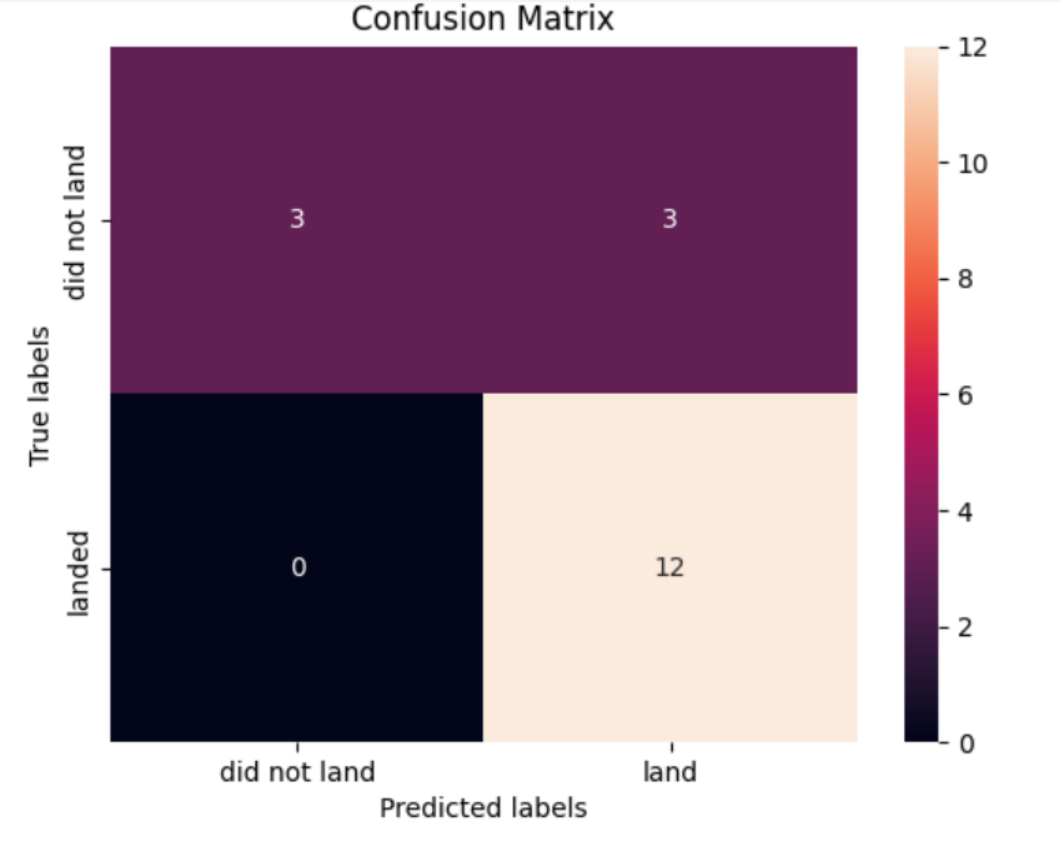
Results: Predictive analysis results



```
                     Accuracy Score
Logistic Regression        0.833333
SVM                        0.833333
Decision Tree              0.833333
K Nearest Neighbors        0.833333
```

From the classification models developed we found out that all of them has similar results.

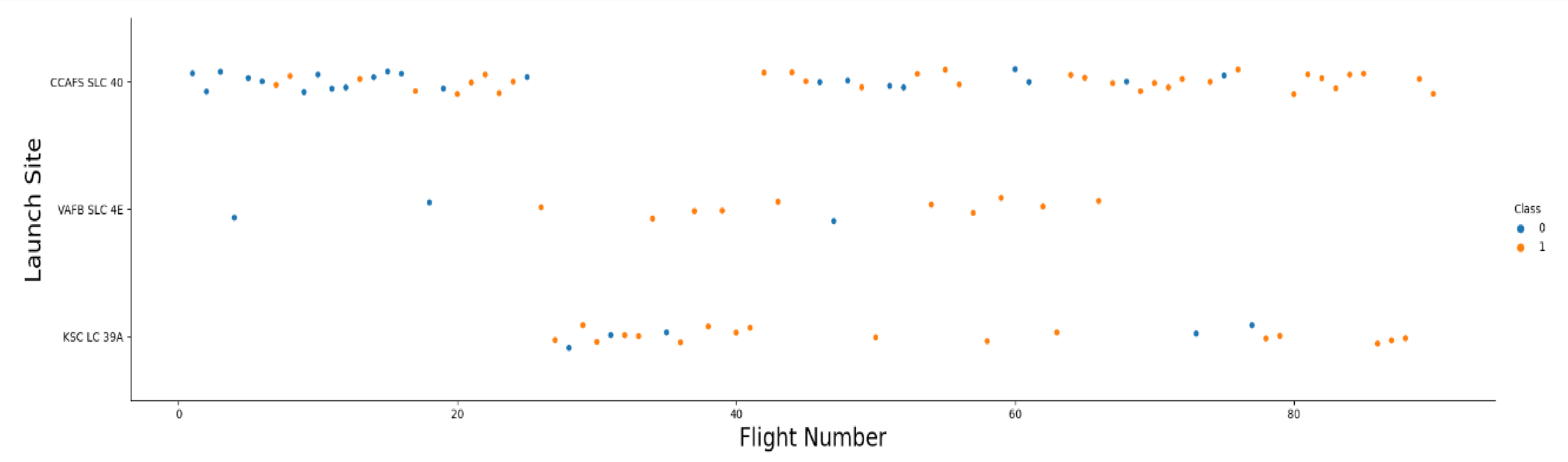
From analysing the confusion matrices produced when testing each model with the test data, we found out that the model's problem were false positive results. Our model sometimes predict a successful land when the first stage did not land.



Section 2

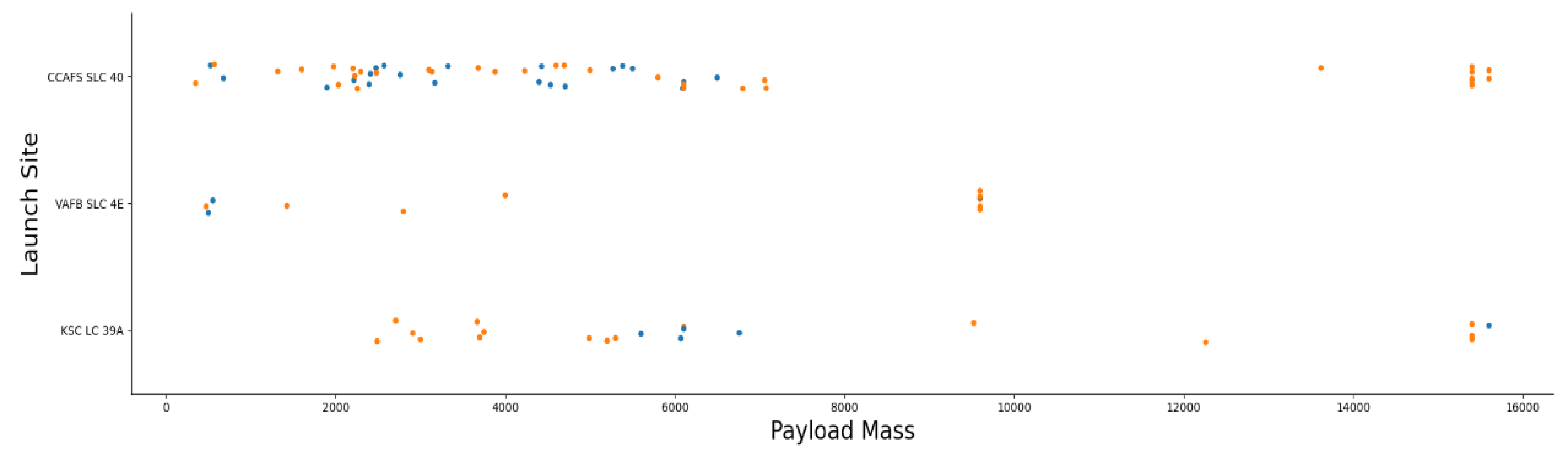
Insights drawn from EDA

Flight Number vs. Launch Site



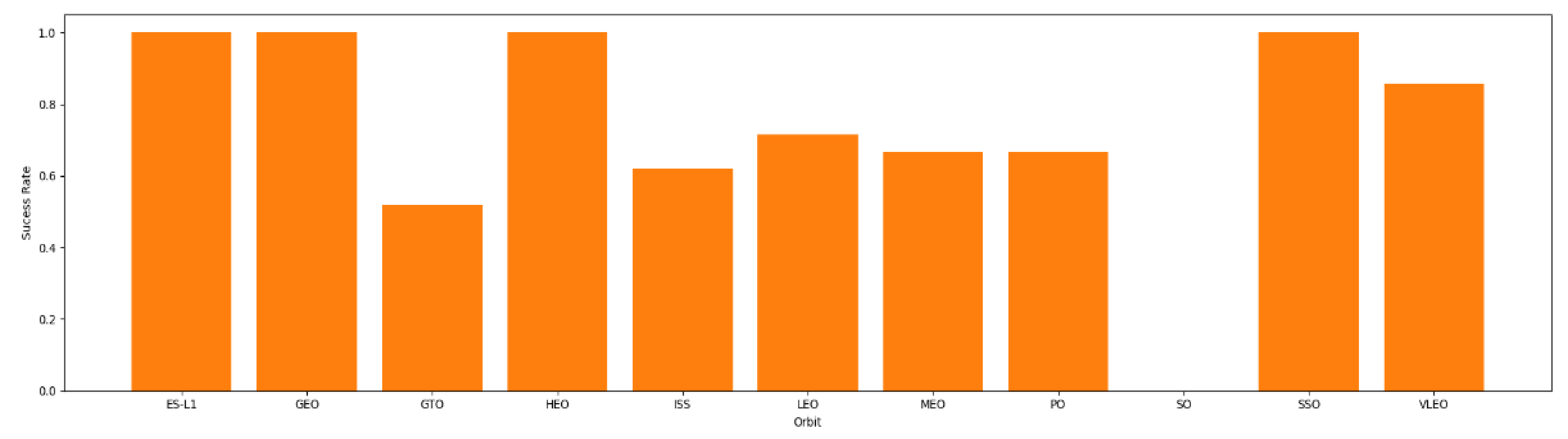
In the graph shows clearly that the success rate increases with the flight number when looking at each Launch Site separately. We can also see that the CCAFS SLC 40 launch site has much more launch records than the other any other launch site.

Payload vs. Launch Site



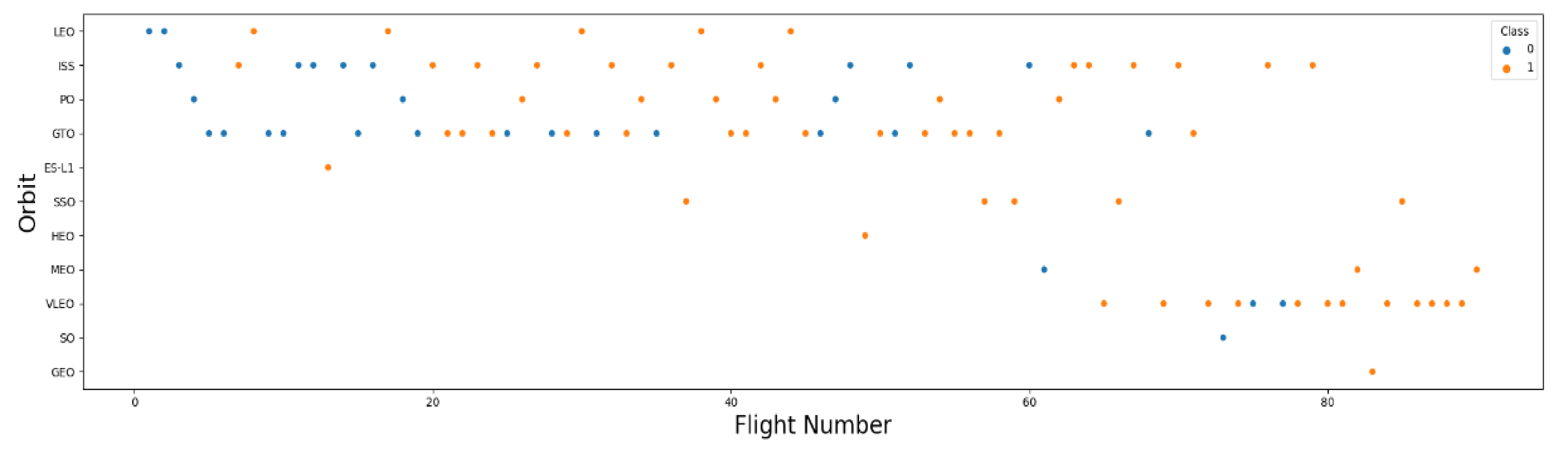
This graph shows that for KSC LC 39A lower payload masses have a higher success rate, whereas for the CCAFS SLC 40 is the other way around. We can also observe that the VAFB SLC 4E launch site does not deal with such heavier payload masses as the other launch sites.

Success Rate vs. Orbit Type



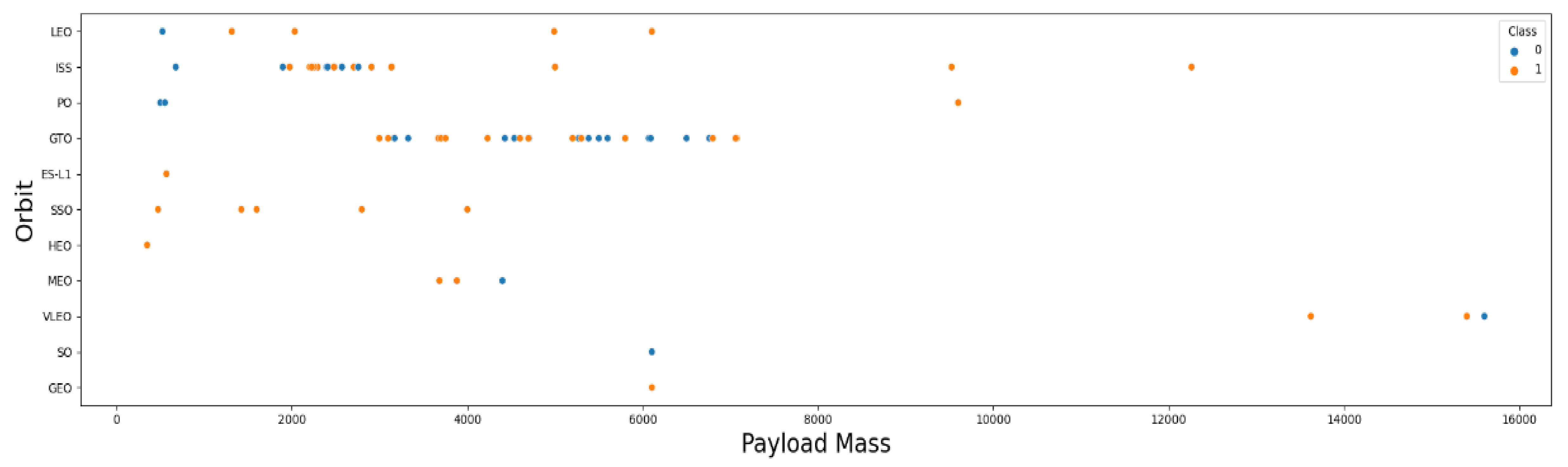
In this bar chart we can see that the ES-L1, GEO, HEO and SSO orbits have a success rate of 100% and the SO orbit has a success rate of 0%. The other orbits have a success rate that varies from 50% to 85%.

Flight Number vs. Orbit Type



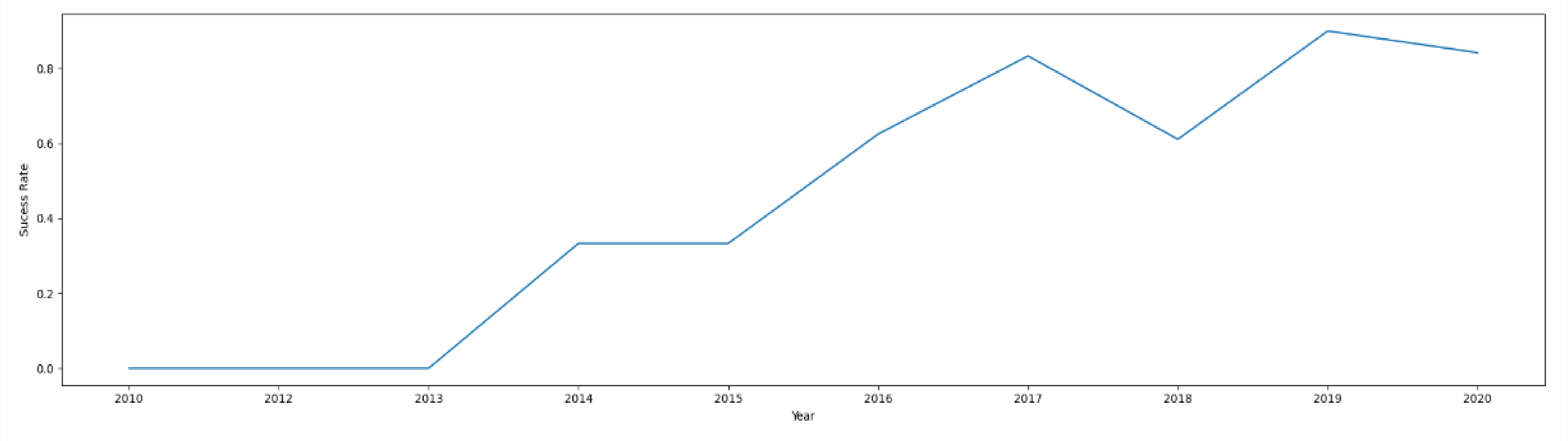
In this chart we can see the correlation between the orbit type, the flight number and the success of landing the first stage. It’s clear that some orbit’s have a lot more launch records than others, such as SO and GEO which have only 1 launch record. It’s also visible that the success rate varies a lot with the orbit type.

Payload vs. Orbit Type



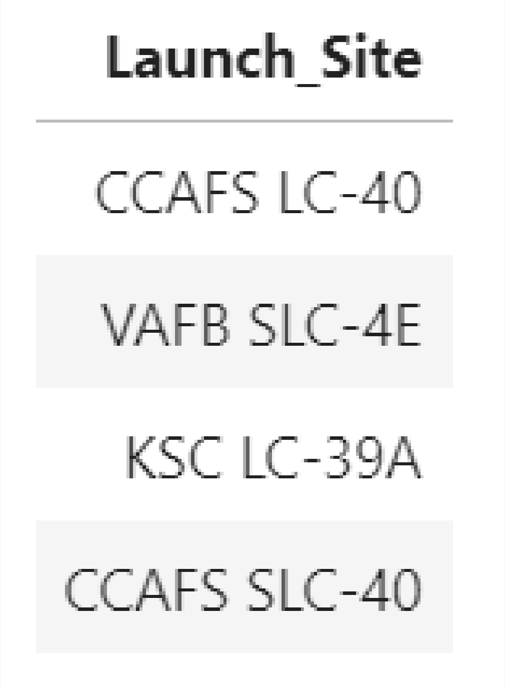
In this chart we can see the correlation between the orbit type, the payload mass and the success of landing the first stage for each launch. We can see clearly that missions to most orbit types do not have heavy payload mass. Only for the ISS, PO and SO orbits do we have payload masses greater than 8000Kg. It's also visible that for some orbits, like LEO, ISS and PO, the first stage landing is more successful for higher payload mass.

Launch Success Yearly Trend



In this chart we can see the relation between the years and the success rate of landing the first stage.
It’s clear that the success rate has a positive trend trough the years.

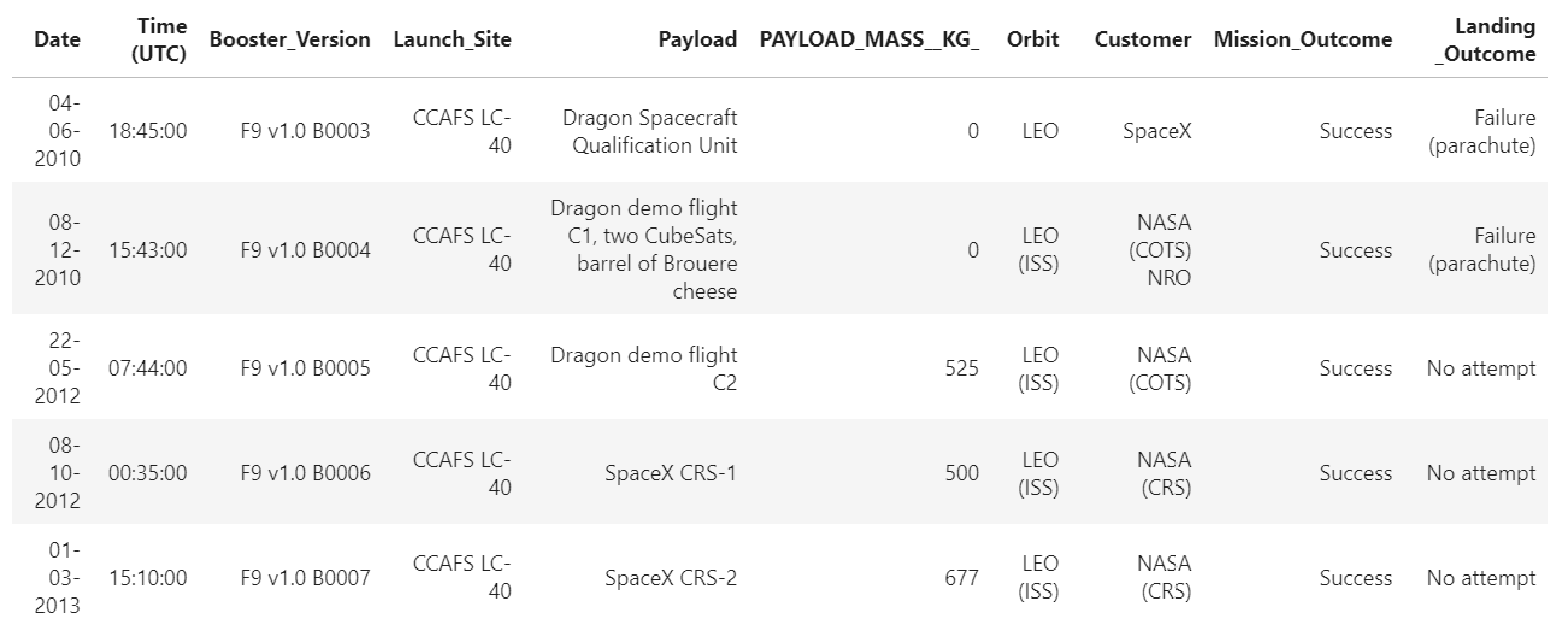
All Launch Site Names



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

There are 4 different launch sites.

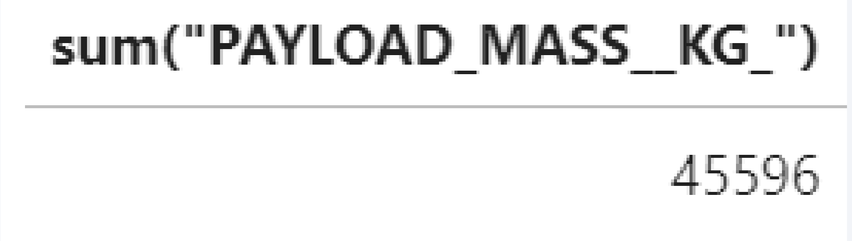
Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
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
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The first 5 launch records where launch site name began with ‘CCA’.

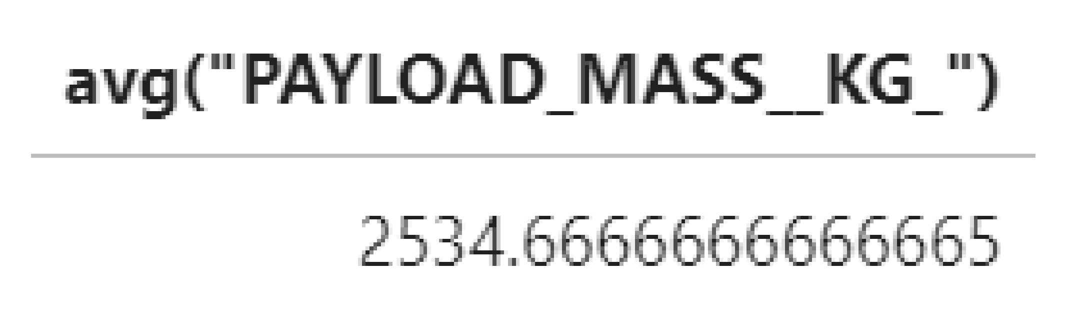
Total Payload Mass



sum("PAYLOAD_MASS__KG_")
45596

The total payload mass carried by boosters from NASA is 45596 Kg.

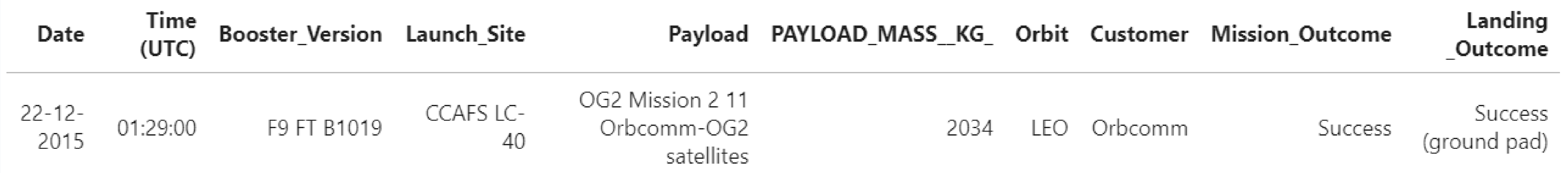
Average Payload Mass by F9 v1.1



avg("PAYLOAD_MASS__KG_")
2534.6666666666665

The average payload mass carried by booster version F9 v1.1 trough all launches is 2537 kg.

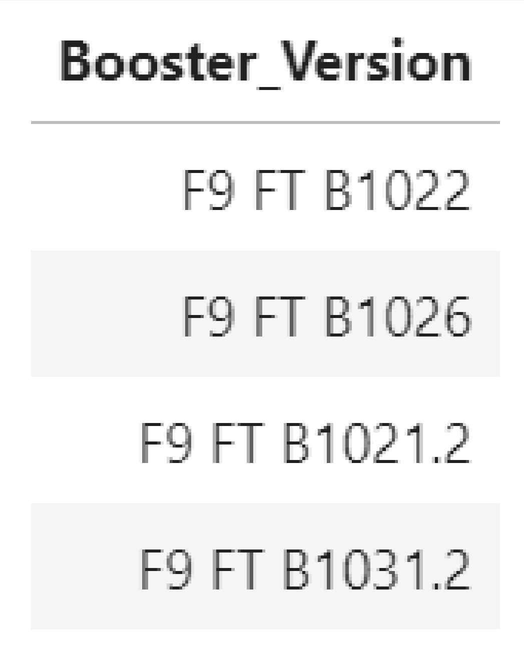
First Successful Ground Landing Date



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

The first successful first stage landing on a ground pad occurred on 22 of December of 2015. It had a payload of 2034kg and it was launch from the CCAFS LC-40 site.

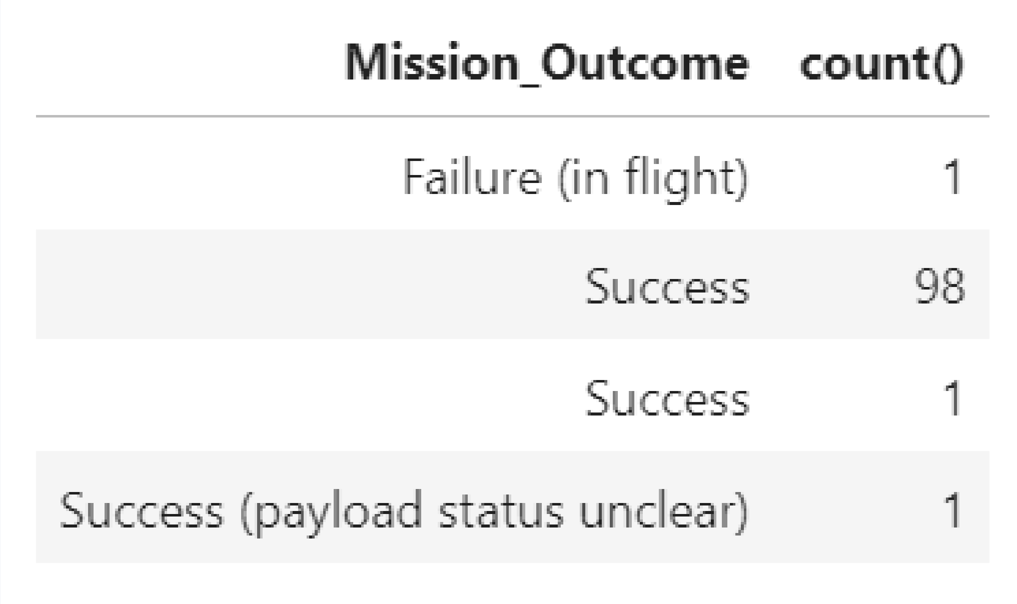
Successful Drone Ship Landing with Payload between 4000 and 6000



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

List of Booster Versions that successfully landed the rocket’s first stage on a drone shit with a payload between 4000 and 6000 kg.

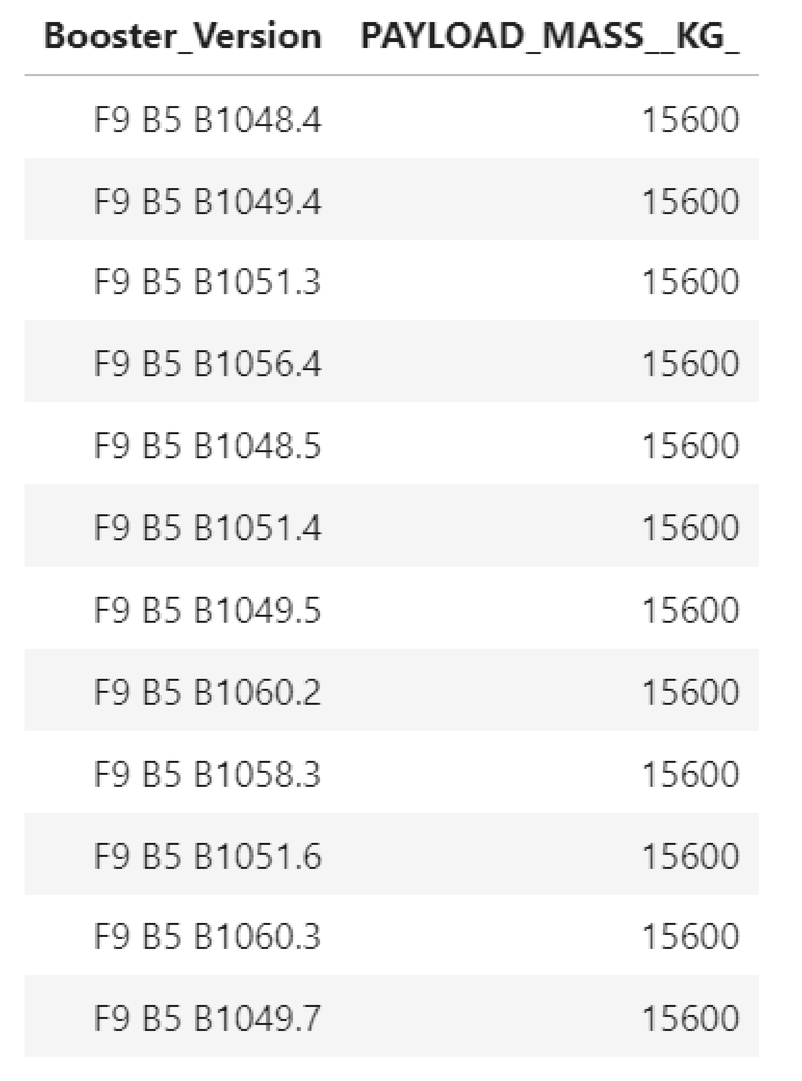
Total Number of Successful and Failure Mission Outcomes



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

There where 100 successful mission outcomes and only 1 failure.
On one of the successful missions, the payload status is unclear.

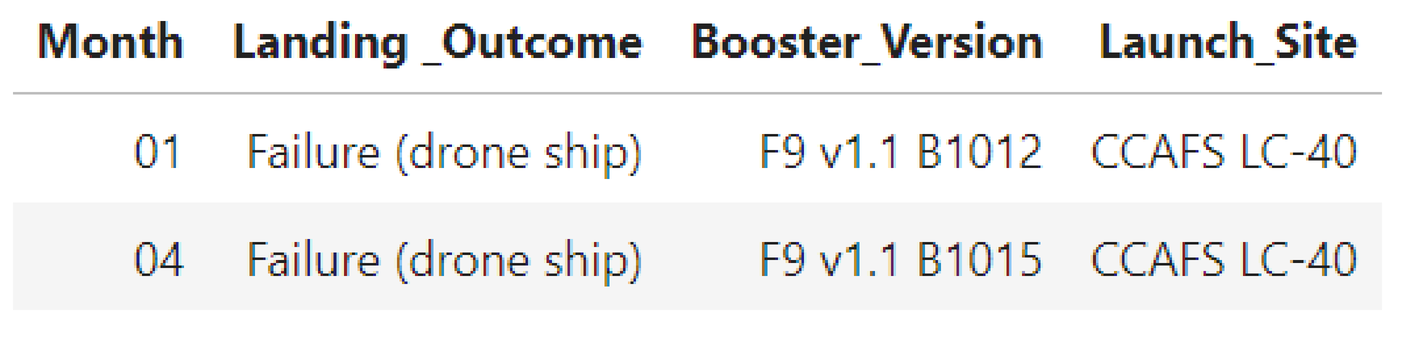
Boosters Carried Maximum Payload



Booster_Version	PAYLOAD_MASS__KG_
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

List of Booster Versions that carried the maximum payload of 15600 Kg

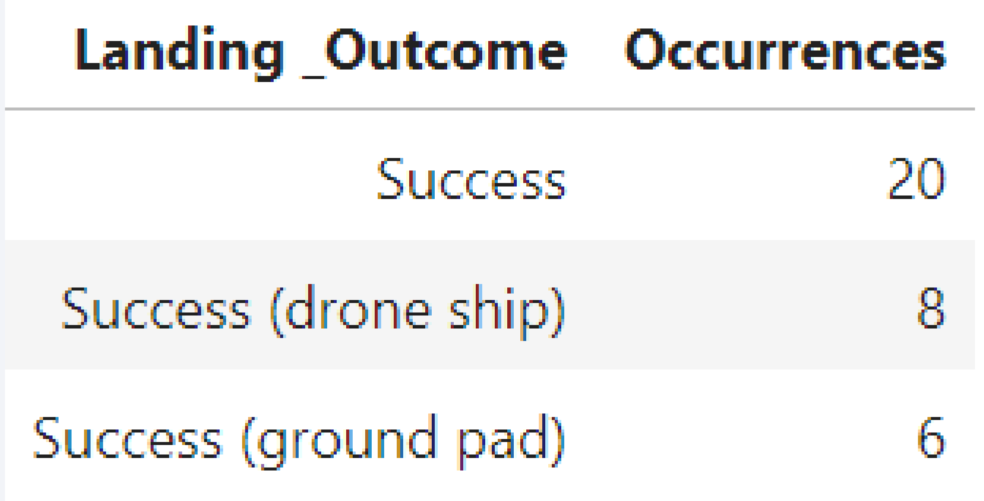
2015 Launch Records



Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

List of records from 2015 where the first stage failed the landing on drone shit.
Both mission launched from the CCAFS LC-40 site, one in January and the other in April.

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



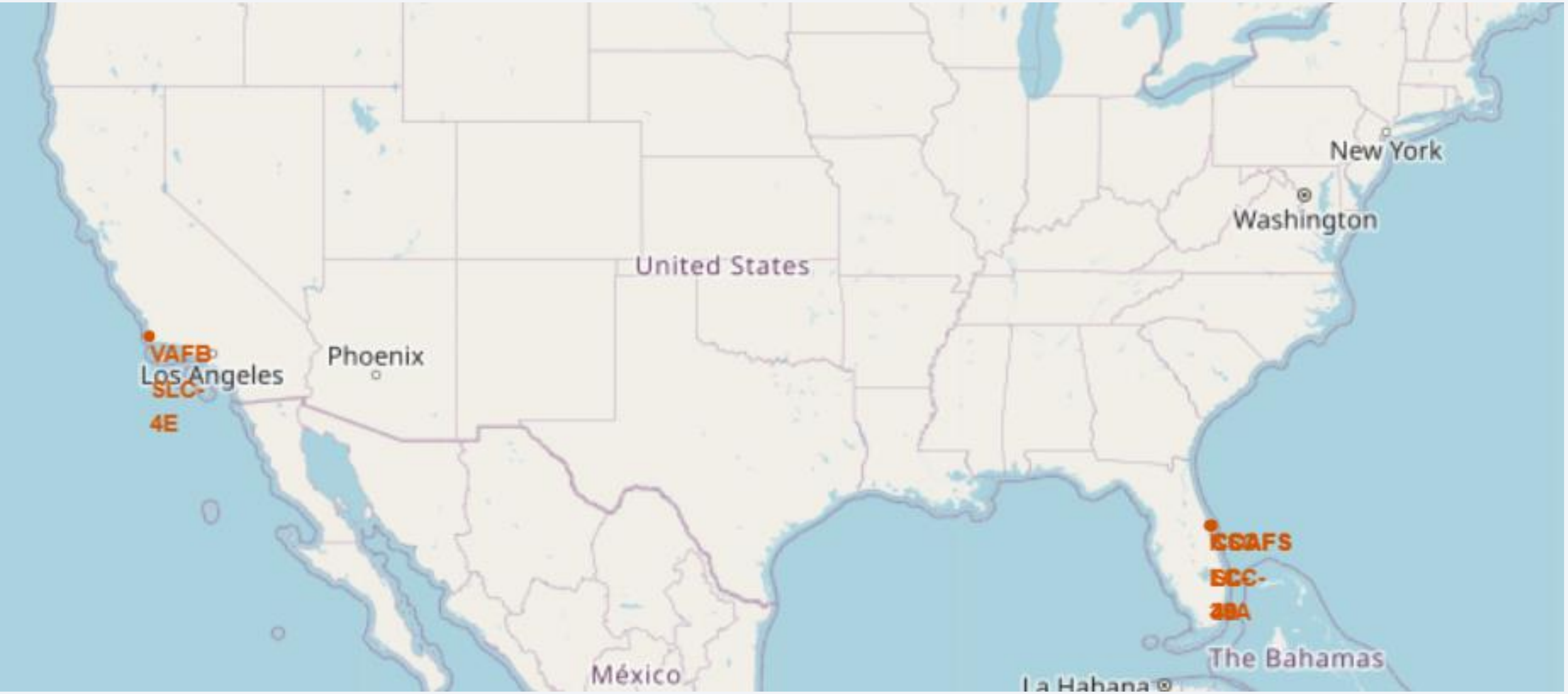
Landing _Outcome	Occurrences
Success	20
Success (drone ship)	8
Success (ground pad)	6

Every different successful landing outcome between 2010-06-04 and 2017-03-20 ranked by number of occurrences.

Section 3

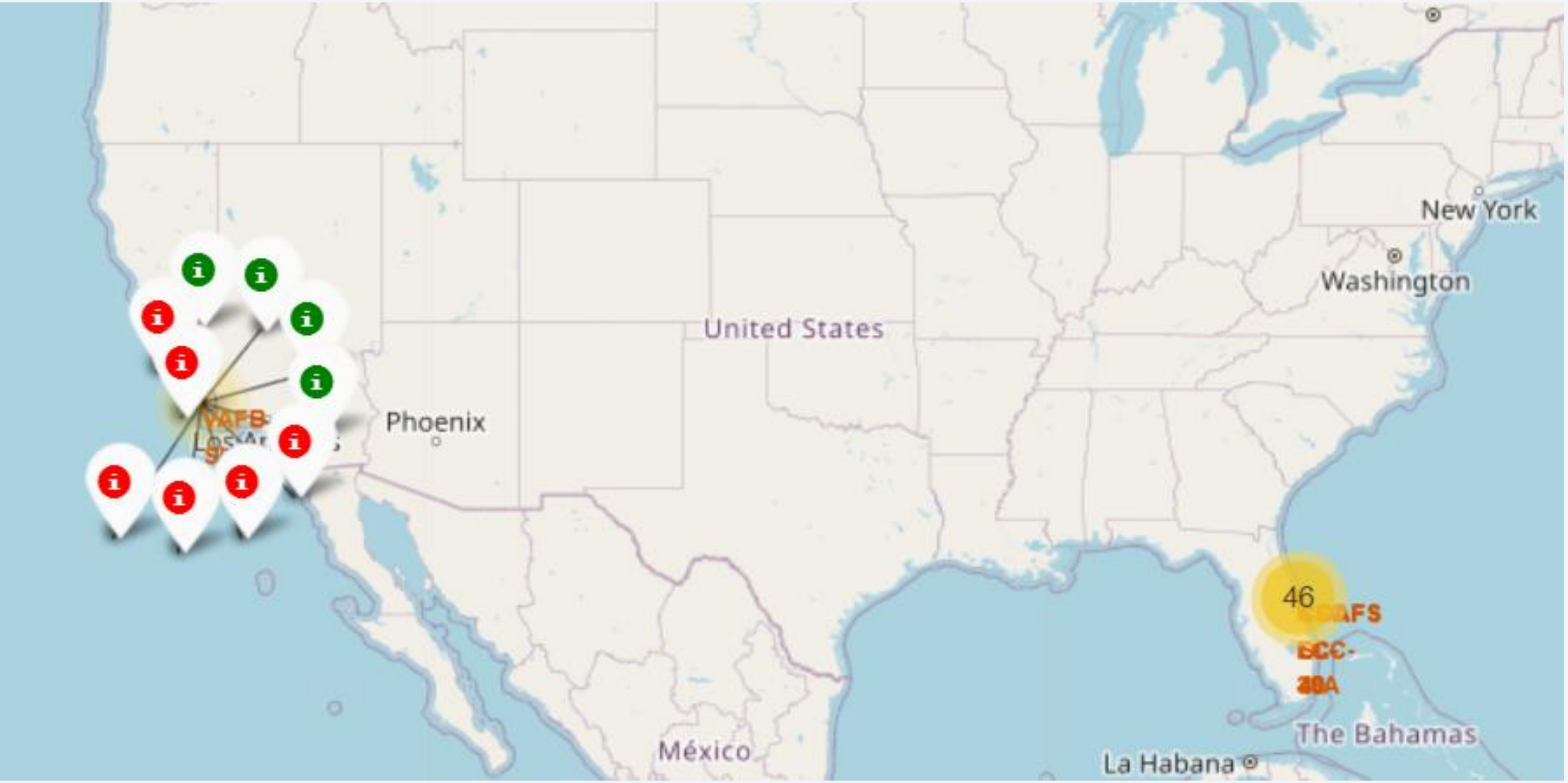
Launch Sites Proximities Analysis

Launch Site Locations



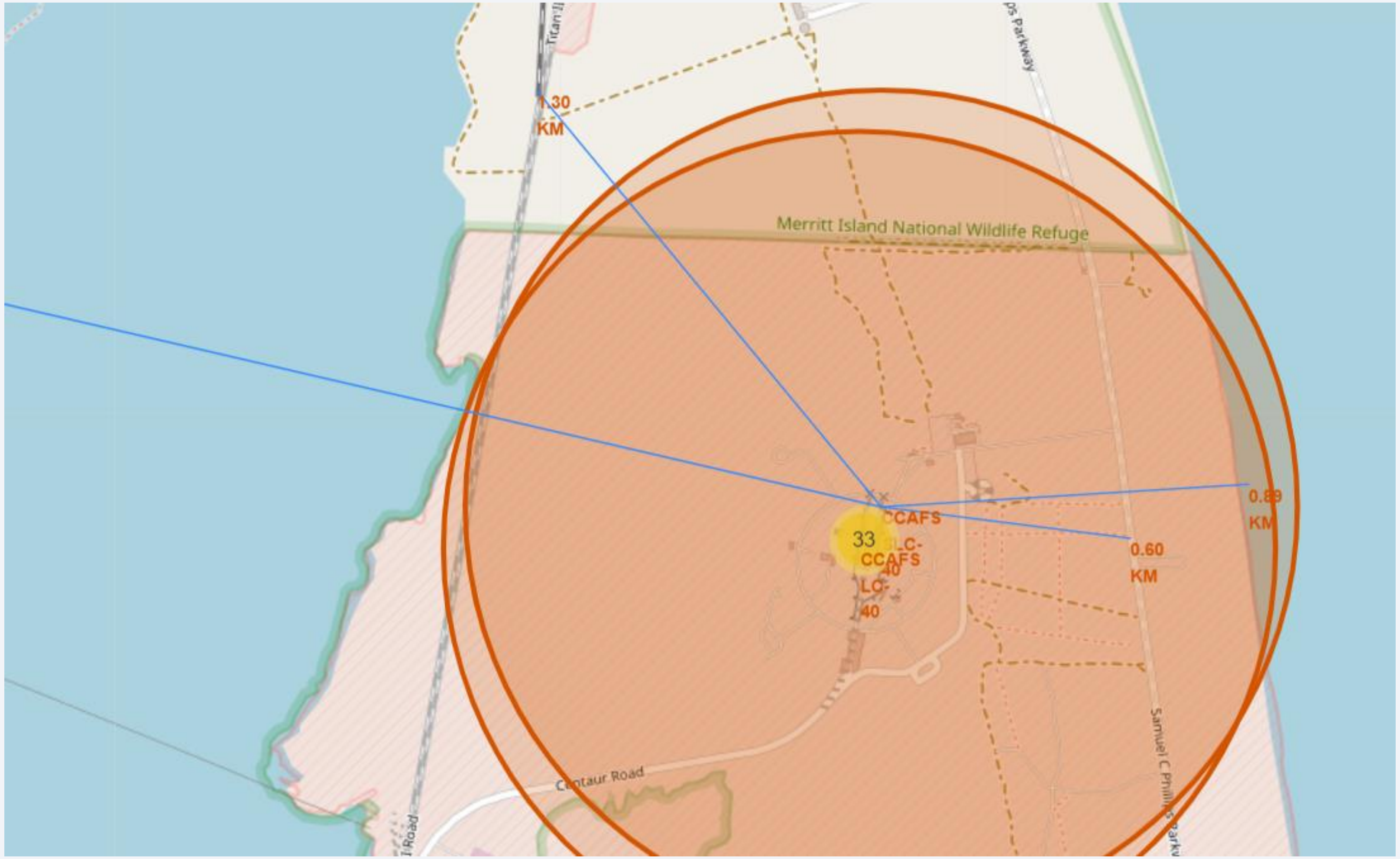
Here we see 4 launch sites locations and that they are distributed in two locations, one on the west coast and the other on the east coast.

Launch Records



In this map we have the launch records represented. On the right we can see a cluster of 46 launch records. On the left we have a exploded cluster of 10 launches from the VAFB SLC-4E launch site, with a color scheme represent the success of the landing.

Launch Site Proximities



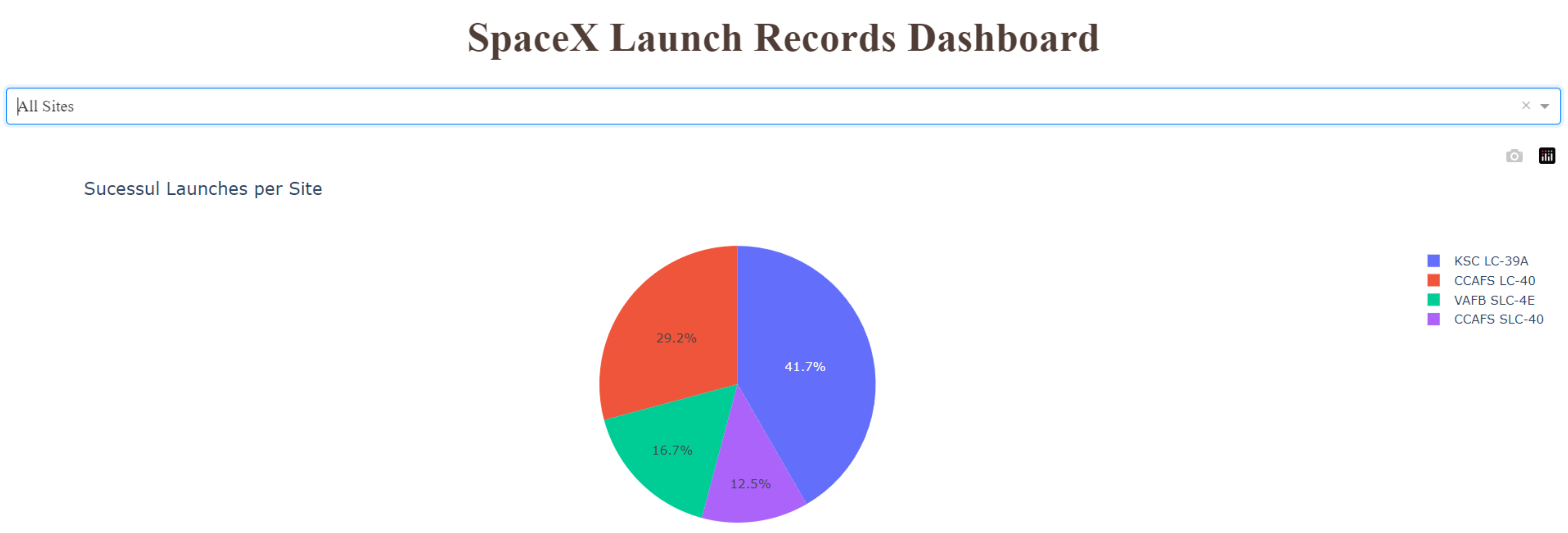
1.30
KM
Merritt Island National Wildlife Refuge
0.89
KM
0.60
KM
33
CCAFS
LC-
CCAFS
40
LC-
40
Centaur Road
Samuel C Phillips Parkway

In this image we can see a launch site and some lines pointing to its proximities, such as the coastline, a highway, railway and cities. The distance between point is also shown.

Section 4

Build a Dashboard with Plotly Dash

Dashboard – All Sites Pie Chart

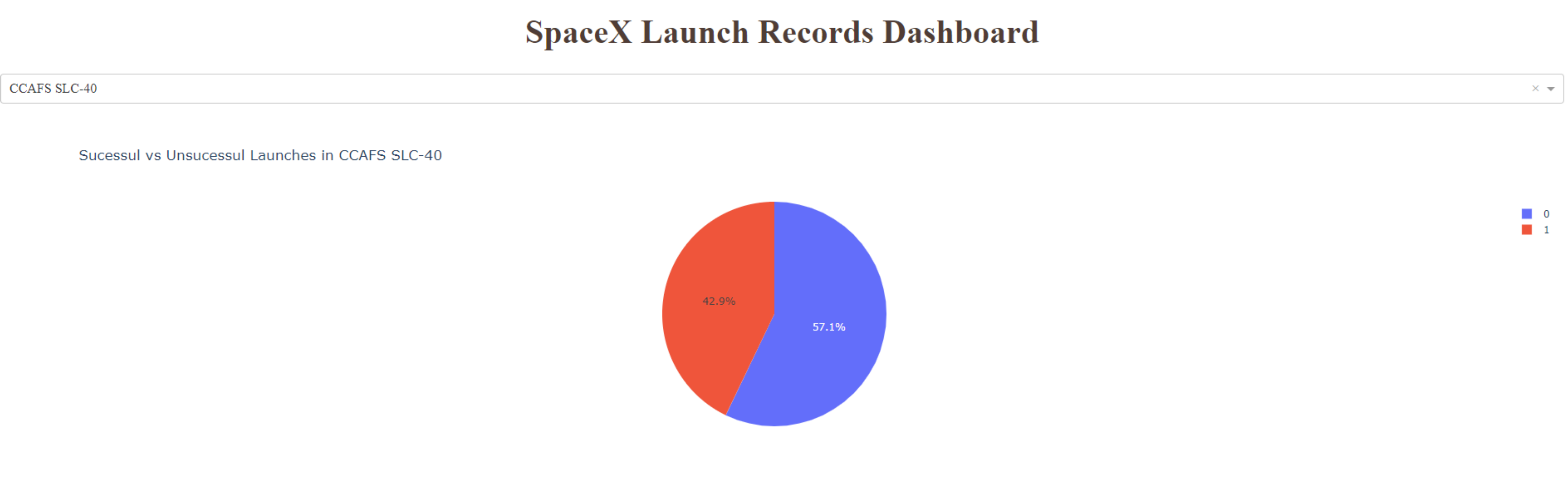


SpaceX Launch Records Dashboard

All Sites

Here we can see our dashboard. The dropdown option “All Sites” is selected and the pie chart is showing the success count for all sites.

Dashboard – One Site Success Rate



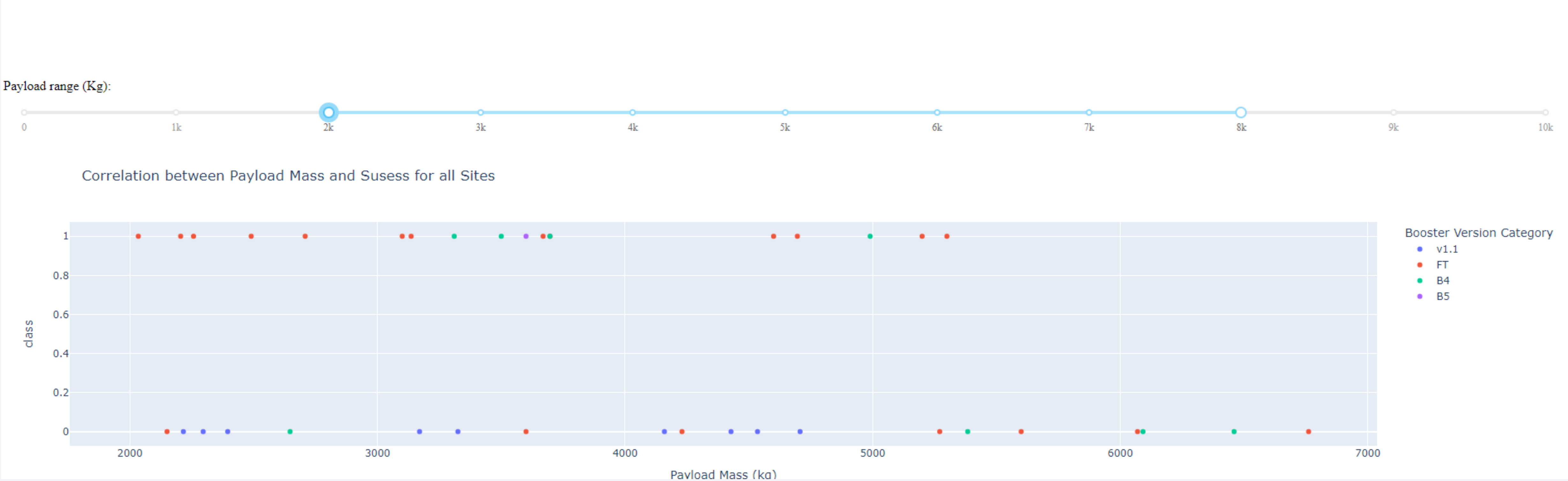
SpaceX Launch Records Dashboard

CCAFS SLC-40

Sucessul vs Unsucessul Launches in CCAFS SLC-40

Here we have selected the CCAFS SLC-40 launch site on the dropdown list and the pie chart is showing the success rate of that specific launch site. We can see in our records there were 43% of successful launches and 57% of unsuccessful ones.

Dashboard – Payload Mass VS Success Rate

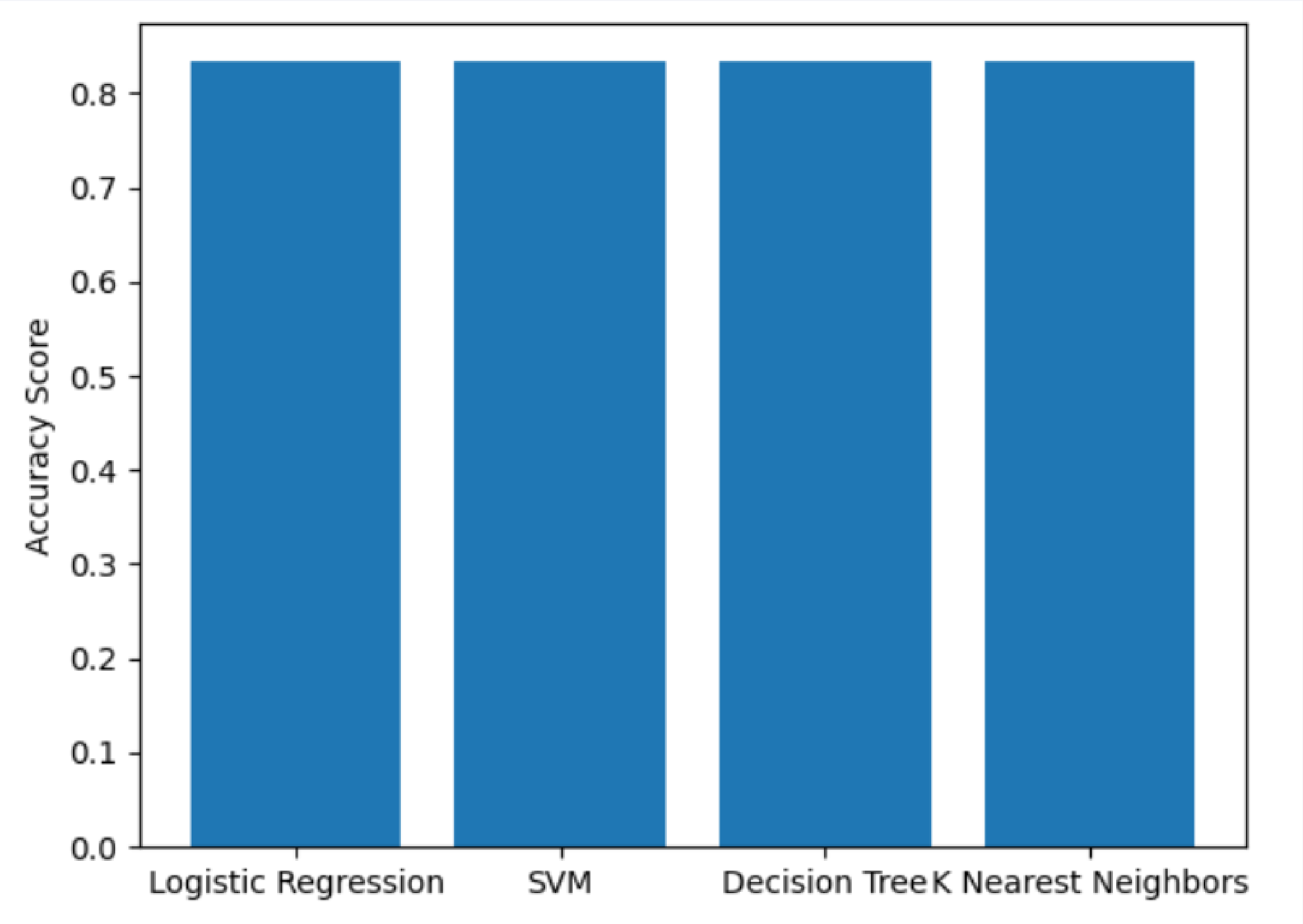


Here we have our Payload Mass range slider with a selected range of 2000kg to 8000kg. From this visual representation we can see that, for this payload range, we only records of 4 Booster Versions. By analysing the scatter plot we can also infer the success rate (Y axis) of the different Booster Versions. It’s clear that the Booster V1.1 has a much lower success rate than the FT one.

Section 5

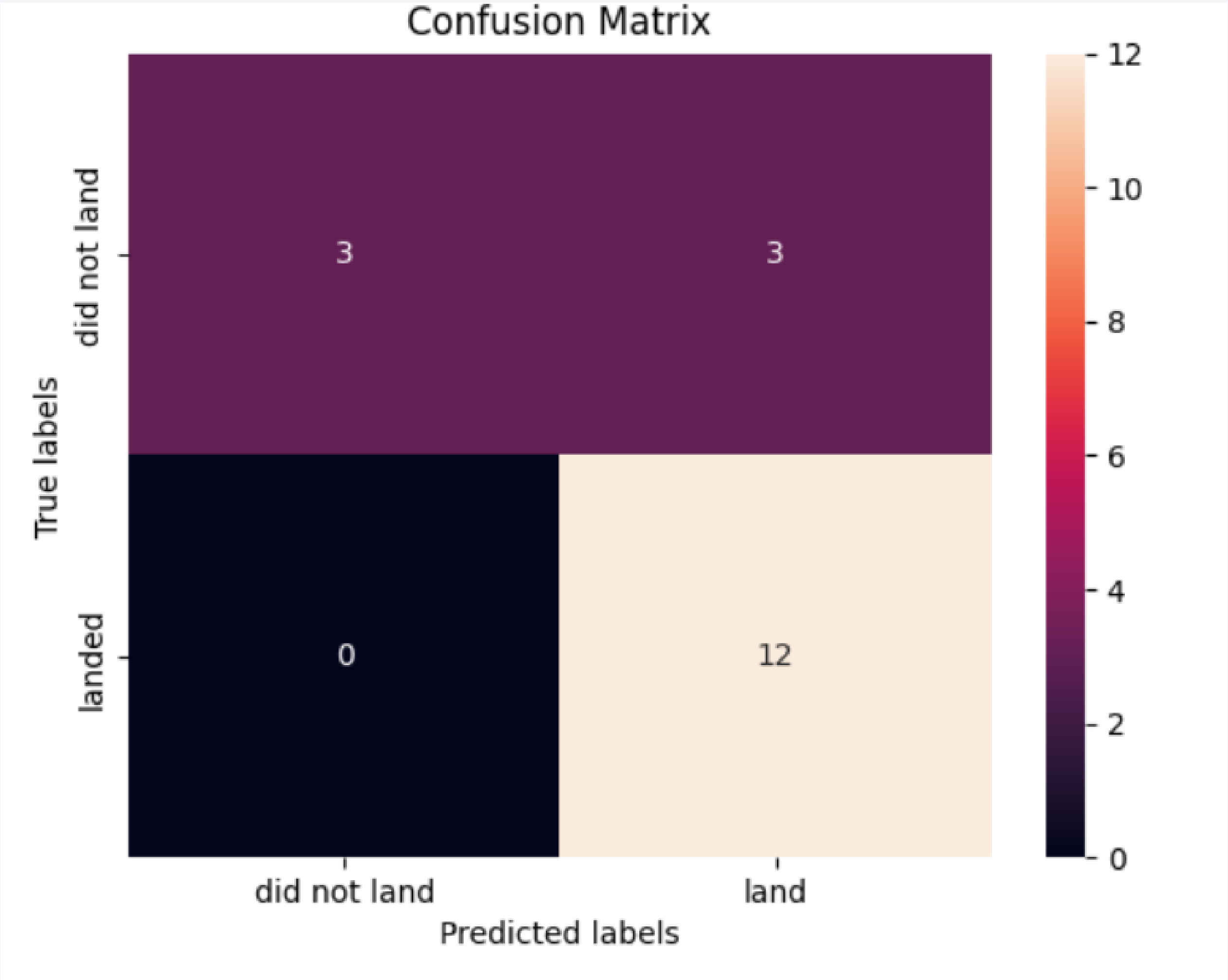
Predictive Analysis (Classification)

Classification Accuracy



Bar Chart showing the Accuracy Score for each classification model built. It's clear that all models had the same accuracy score.

Confusion Matrix



Every model produced the same confusion Matrix when tested with the test data. This shows that our models had a problem with false positive results, meaning that sometimes predicts a successful landing of the first stage, when it did not land successfully.

Conclusions

- In this worked we built a Dataset with records of rocket launches with relevant variables to the mission success and ready to be used to build classification models.
- We also determined what factors are more relevant to the success of landing the first stage of a Falcon 9 rocket launch.
- We developed models to predict if the first stage will land successfully with more than 80% of accuracy, which usually means that they are good enough to deploy.
- Even tough the produced models were good, some improvements could still be done. Our dataset was small, with a bigger one better relationships could have been discovered. Some of the variables that we use, will lose relevance with growth of the dataset. The ‘Flight Number’ is an example of this, because it’s expected that in the beginning more errors and failures will occur.
- More features that we did not analyze could also be relevant, such as the weather, of which we had no information.

Thank you!