IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Victor Manke
November 8th 2022

Outline

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

Executive Summary

- Summary of the methodologies:
  - Data was collected using both web scraping and the SpaceX REST API;
  - An exploratory data analysis of the data was conducted using various methods to clean up, and visualize different aspects of the gathered data;
  - A set of machine learning techniques were tested to try and predict whether or not a SpaceX launch would result in a successful landing or not.
- Summary of all results:
  - The exploratory data analysis let to the definition of which part of our dataset had some correlation to the landing success metric we are trying to predict;
  - All of the tested techniques resulted in a similar prediction score of around 83.3% of the test dataset being correctly predicted.

Introduction

- This project has the objective of determining the price of each launch from SpaceX, by analyzing when launches have successful landings;
- The main problem we are focusing on is:
  - To find the best way to predict when a launch will have a successful landing, and therefore be able to reuse part of it's rocket, reducing the costs of the next one.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was initially scraped from the Falcon 9 Wikipedia page;
  - After that we used the open SpaceX REST API to get even more data.
- Perform data wrangling
  - The different data objects we gathered were joined to form a single table that described a single launch per row;
  - A classification label was created to indicate when a landing was successful based on a column that carried that information.

Methodology

Executive Summary

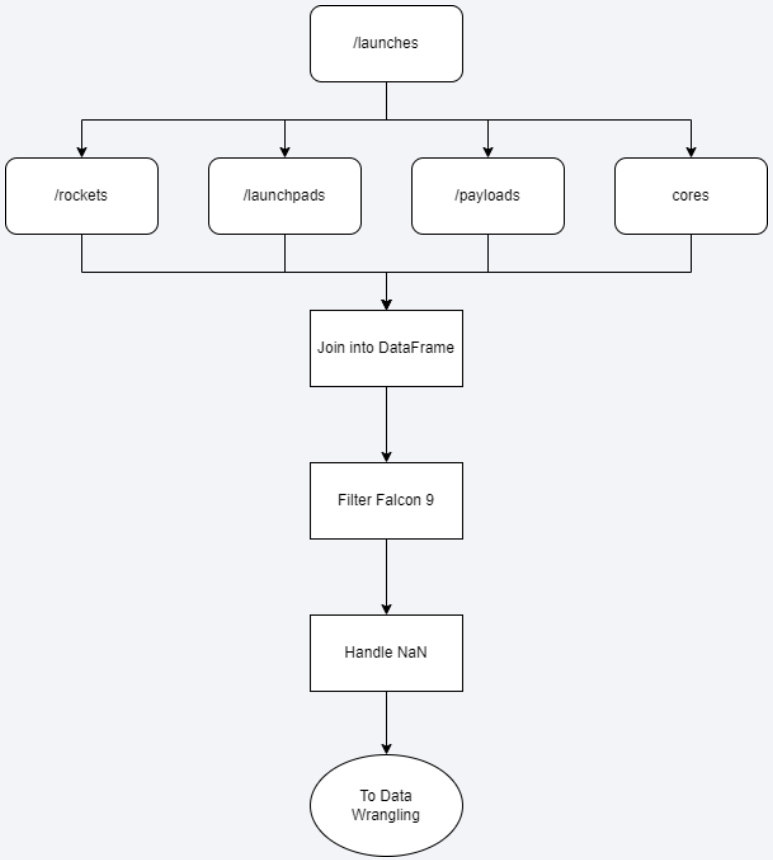
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - In this step the data was divided into training and testing slices;
  - This data was used to train and evaluate four different models:
    - Logistic Regression;
    - Support Vector Machine;
    - Decision Tree;
    - k Nearest Neighbors.

Data Collection

- Data was collected from two different sources:
  - The Falcon 9 launches Wikipedia page (Link)
    - Data was scraped from the tables present on the page;
  - The SpaceX REST API (Link)
    - Data was requested from multiple endpoints and then joined as a single table.
- The parsed data from each source was turned into pandas dataframes for ease of manipulation.

Data Collection – SpaceX API

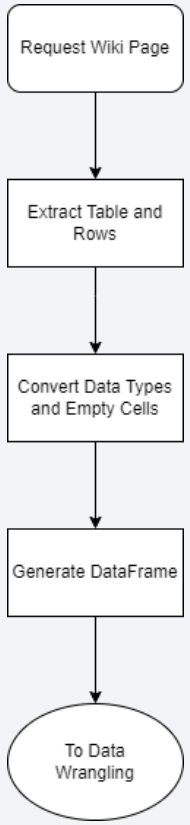
- SpaceX has a public API which has a lot o data from their launches and rockets that we can use;
- For example, this route returns a list of launches and their data;
- The data from other routes was then aggregated to the launch route in order to substitute the field ids with their representation:
  - Rockets, launch pads, payloads and cores.



GitHub Link

Data Collection - Scraping

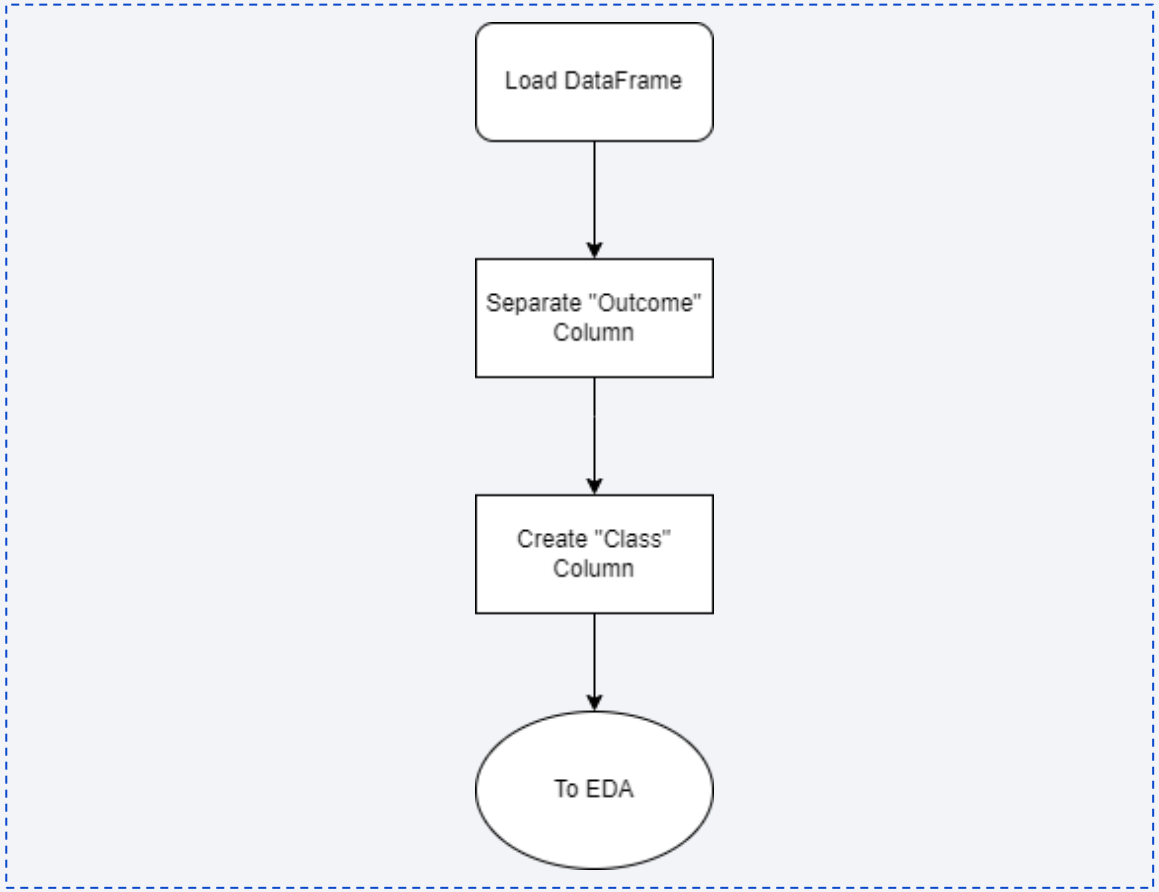
- There exists a Wikipedia page where each SpaceX launch and some metadata is registered.
- The page can be accessed through this link;
- The data was scraped from the tables present on the page by parsing the HTML with the BeautifulSoup library.



GitHub Link

Data Wrangling

- With a dataframe created a little analysis was conducted on the categorical columns in order to understand each value occurrence on the dataset;
- The “Outcome” column carried both information from the success of the launch and where the landing was conducted (ocean, ground pad, drone ship);
- A separate column “Class” was created to represent solely the success part of the outcome to simplify the subsequent analysis.



GitHub Link

EDA with Data Visualization

- For the EDA several plots were generated, including scatter, bar and line charts comparing different columns;
- On the scatter plots, the previously defined “Class” was used as the hue of the points for extra insight;
  - Payload Mass x Flight Number
  - Launch Site x Flight Number
  - Launch Site x Payload Mass
  - Orbit x Flight Number
  - Orbit x Payload Mass
  - Success Rate x Orbit
  - Success Rate x Year

GitHub Link

EDA with SQL

- With the data saved in an SQL database it was also possible to obtain insight directly with SQL queries:
  - Names of the unique launch sites;
  - Launches from launch sites beginning with “CCA”;
  - Total payload mass carried by boosters launched by NASA;
  - Average payload mass carried be booster “F9 v1.1”;
  - Date of the first successful ground pad landing;
  - Boosters with successful drone ship landings with mass between 4000 and 6000 kg;
  - Total number of successful and failures;
  - Booster versions that have carried the maximum recorded payload mass;
  - Failed drone ship landings of 2015;
  - Rank of the most frequent landing outcomes from 2010-06-04 to 2017-03-20.

GitHub Link

Build an Interactive Map with Folium

- A map plot from Folium was generated from the data with a number of indicators for insight:
  - Markers indicating the launch sites and clusters for the launches;
  - Circles to hightlight the area around the launch sites;
  - A line was created to indicate the distance from the launch site “CCAFS SLC-40” to the coastline.

GitHub Link

Build a Dashboard with Plotly Dash

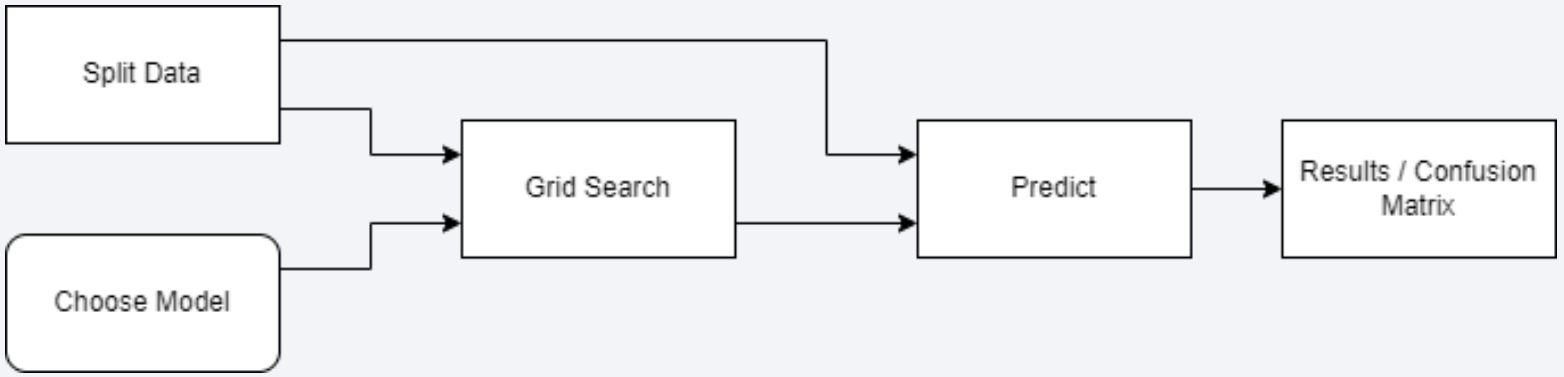
- A plotly dash dashboard was created with the following visualizations:
  - Total success launches
  - Correlation between payload and success
- Each graph was made to change according to a selector in order to filter the results by launch site;

- The Correlation graph was also filtered with a slider for the payload mass.

GitHub Link

Predictive Analysis (Classification)

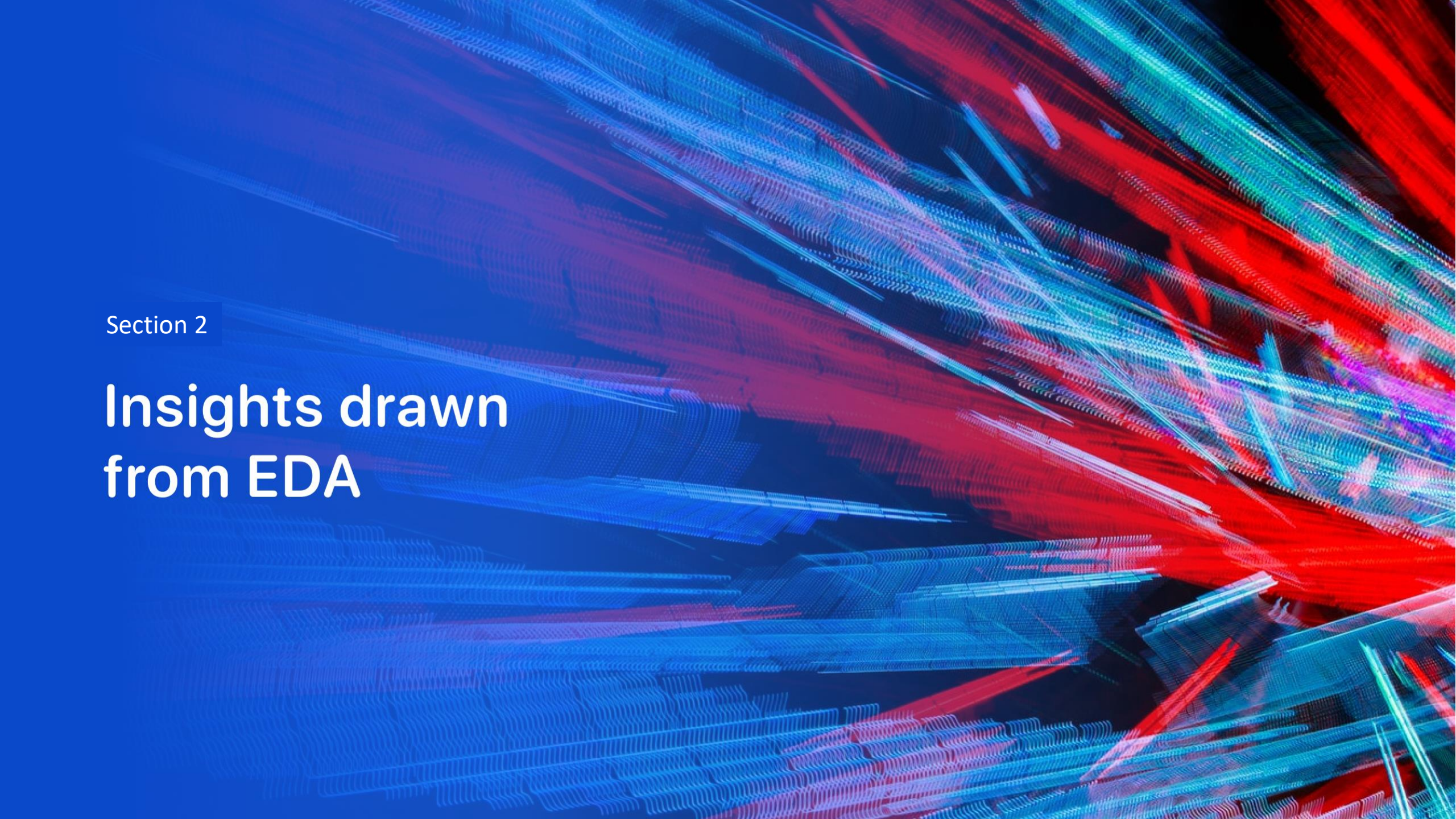
- The data was first separated into separate train and test data sets;
- The four classification models that were tested were trained with the train data set, and also using a Grid Search wrapper to optimize the hyperparameters;
- To evaluate the model, a prediction was made on the test data set and a confusion matrix was generated.



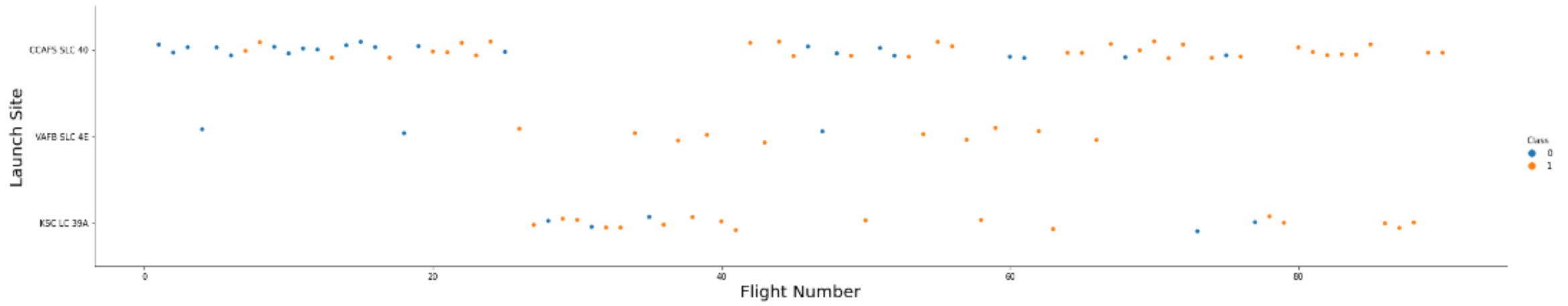
GitHub Link

Results

- The success rate of the landings has been increasing over time, even with a dip in 2018;
- VLEO seems to be the new go-to orbit for launches, and has a high landing success rate;
- There are four launch sites currently, and they are all located near the shore and far from the general population;
- Even though the landings are not always successful the missions are still 99% successful;
- All models have resulted in a similar prediction for the test data, however, there is a low number of datapoints to test.

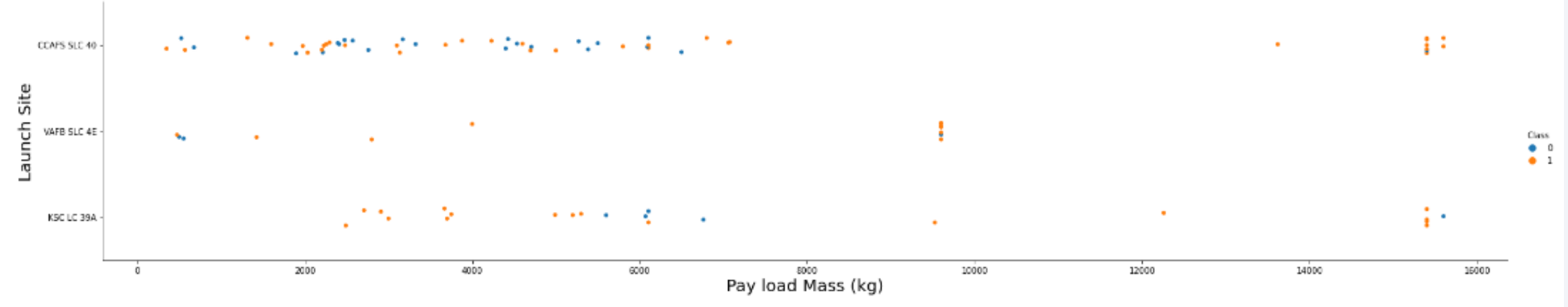

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



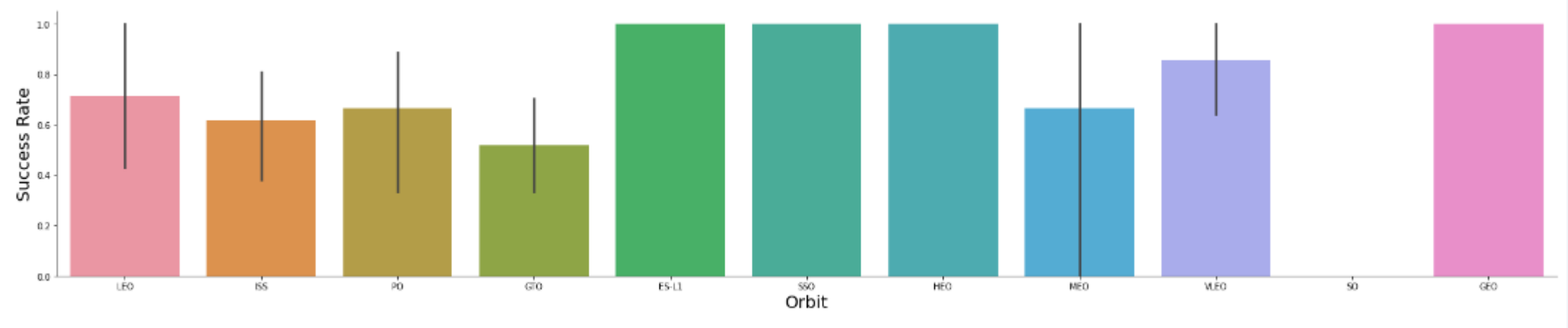
- Just from the classification and flight number comparison, we can see that the success rate has been improving over time;
- CCAFS SLC 40 is the most used launch site by far, with VAFB being the least used and more sparsely used as well.

Payload vs. Launch Site



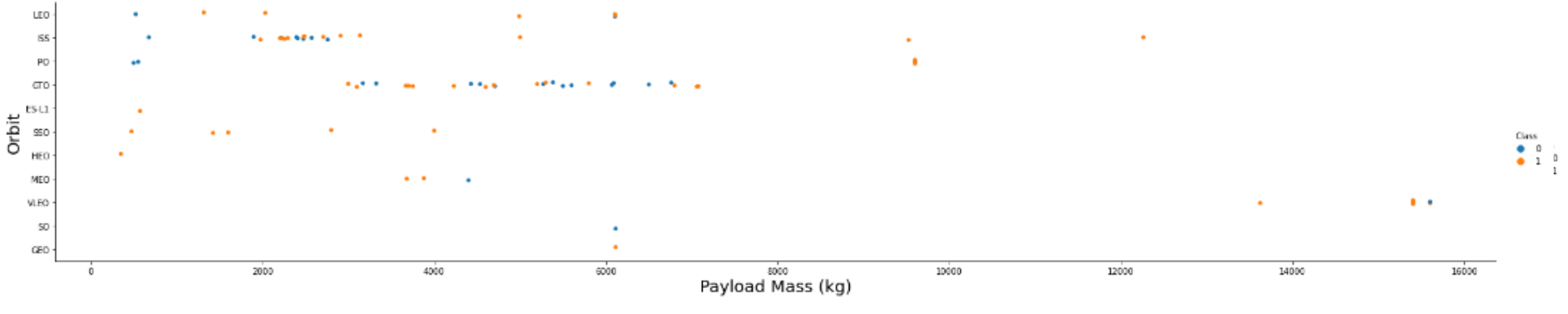
- Payloads over 8000 kg have a very high success rate, almost all between 9000 and 10000 kg launched from VAFB SLC 4E, the others are all newer launches;
- Smaller payloads have a lower success rate, generally because the earlier launches were all with these smaller payloads.

Success Rate vs. Orbit Type



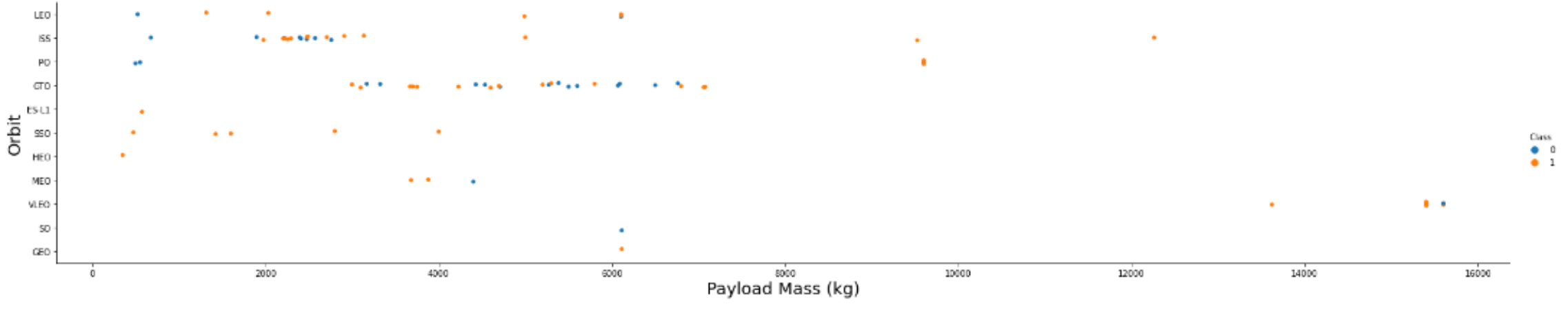
- All of the orbits have more than 50% success rate;
- The most successful orbits, with 100% success rate are: ES-L1, SSO, HEO and GEO;
- The only other orbit with over 80% success rate is VLEO.

Flight Number vs. Orbit Type



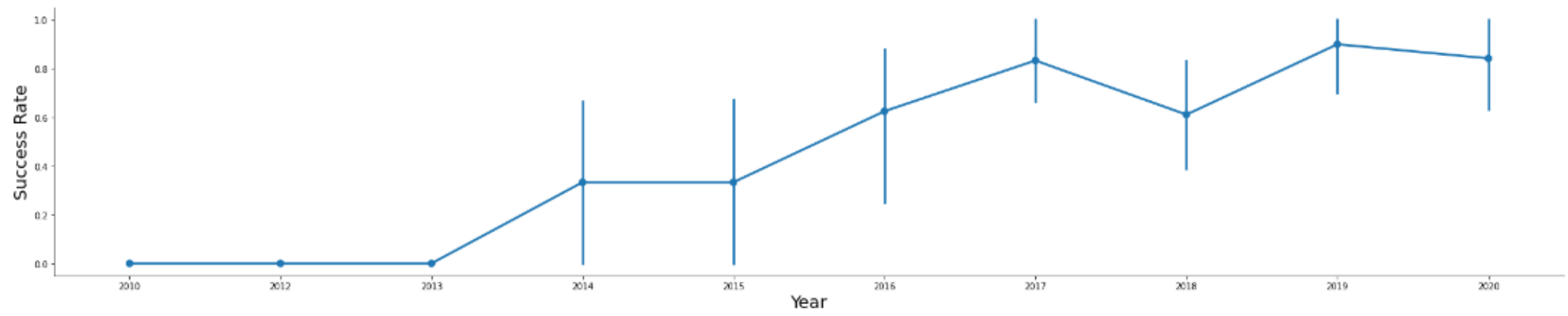
- LEO, ISS, PO and GTO were used the most in early launches, but are not being used as frequently anymore;
- From around flight 60 and onward a preference to the VLEO orbit seems to be forming.

Payload vs. Orbit Type



- Most orbits are pretty versatile when it comes to payload mass launches, having a pretty big range of launched masses;
- The very large payloads seem to prefer the VLEO orbit, but it might be just that there are so few of those launches still.

Launch Success Yearly Trend



- The success rate is trending up, but it looks to be slowing down since 2017;
- A dip in success rate in 2018 might be due to new orbits or technology being tested;
- From 2010 to 2013 are likely the experimental years where the concepts were being tested and validated.

All Launch Site Names

- The data we had showed four different launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC039A
  - VAFB SLC-4E

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

- These were the distinct values on the “launch_sites” column of our database;

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	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

- We searched for 5 records from the “CCA” launch sites;
- We can see that all of them use the same orbit;
- All of them were successful missions, but none landed successfully;

Total Payload Mass

- By aggregating the payload masses from the customer “NASA (CRS)” we found that the boosters launched by them totaled 45596 kg;

- There are other “NASA” customers, but this analysis was made exclusively for the “(CRS)” variant.

	1
45596	

Average Payload Mass by F9 v1.1

- By averaging the payload masses from boosters “F9 v1.1” we found that these boosters launched an average of 2928 kg;

- The same could be queried for different booster versions.

1
2928

First Successful Ground Landing Date

- By querying the minimal date for record with value “Success (ground pad)” on the “landing__outcome” column, we found that the first successful ground pad landing was in December 22nd 2015;

	1
2015-12-22	

Successful Drone Ship Landing with Payload between 4000 and 6000

- By creating a filter on the landing outcome **and** on the payload mass columns, we were able to retrieve which booster versions successfully landed on Drone Ship with payload between 4000 and 6000 kg:
  - F9 FT B1022;
  - F9 FT B1026;
  - F9 FT V1021.2;
  - F9 FT V1031.2.

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

Total Number of Successful and Failure Mission Outcomes

- By grouping the records by mission outcome, we can extract the total number of records for each different outcome:
  - Success: 99;
  - Success (payload status unclear): 1;
  - Failure (in flight): 1.

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

- We can see that even if the landings are not always successful, the missions are over 99% successful.

Boosters Carried Maximum Payload

- By querying the maximum payload and subsequently the missions carrying this payload size, we can list all the booster version that have carried maximum payload:
  - 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

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

2015 Launch Records

- During the year 2015 only two missions had failure landing in drone ships, both were launched from the CCAFS LC-40 launch site, with the following booster versions:
  - F9 v1.1 B1012
  - F9 v1.1 B1015

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

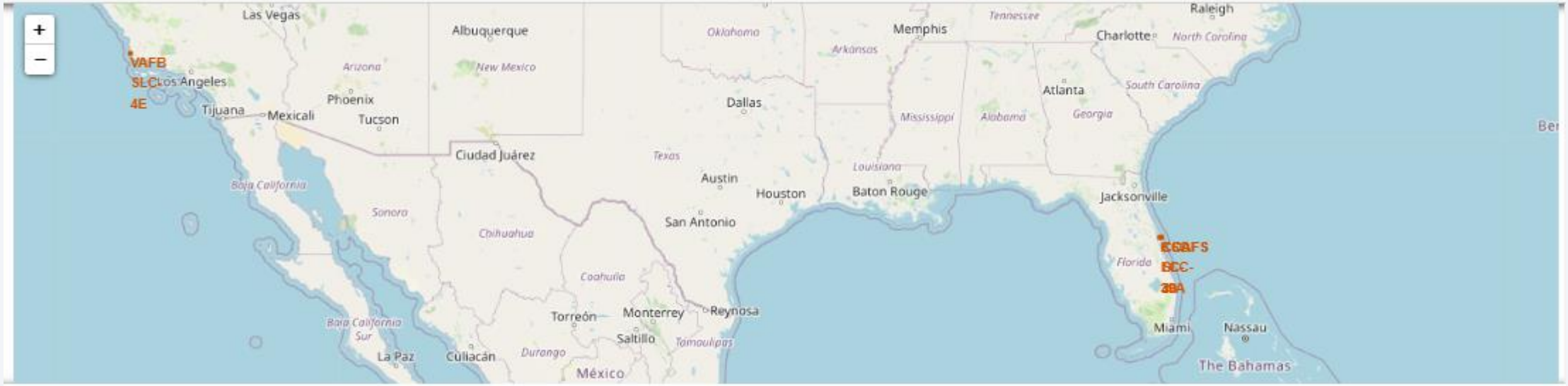
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Grouping by landing outcome, and filtering by date we can rank the most common landing outcomes within the period of June 6th 2010 to March 3rd 2017:
  - No attempt: 10 missions;
  - Failure (drone ship): 5;
  - Success (drone ship): 5;
  - Controlled (ocean): 3;
  - Success (ground pad): 3;
  - Uncontrolled (ocean): 2;
  - Failure (parachute): 1;
  - Precluded (drone ship): 1;

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


Section 3
Launch Sites
Proximities Analysis

Location of Launch Sites

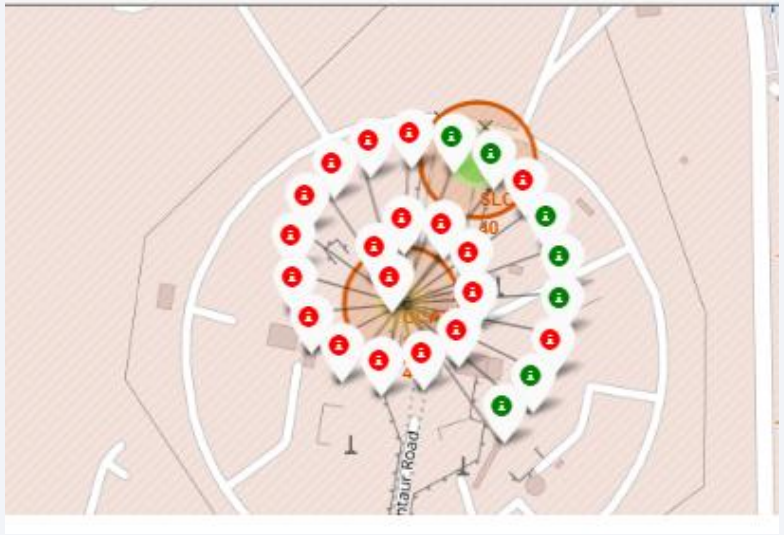


- Other than VAFB SLC-4E, which is located in California, all the other launch sites are located in Florida;

- In both cases they are located near the shore.

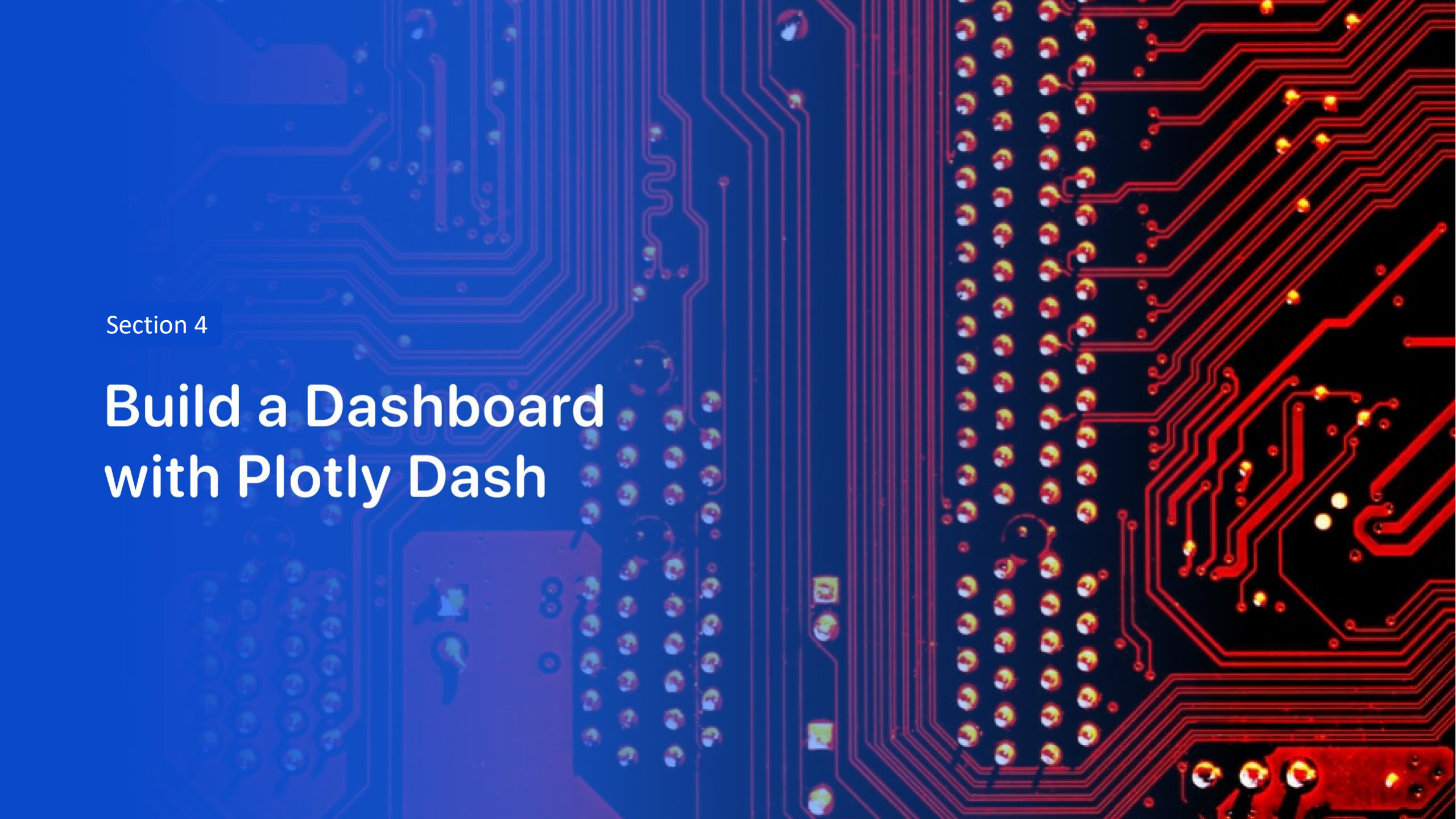
Landing Outcomes Marks

- On each launch site we can now see each launch marked, as well as the classification of success (green) and failure (red) for the landing outcome;
- The picture indicating the missions for the CCAFS LC-40 site.



Launch Site Proximities

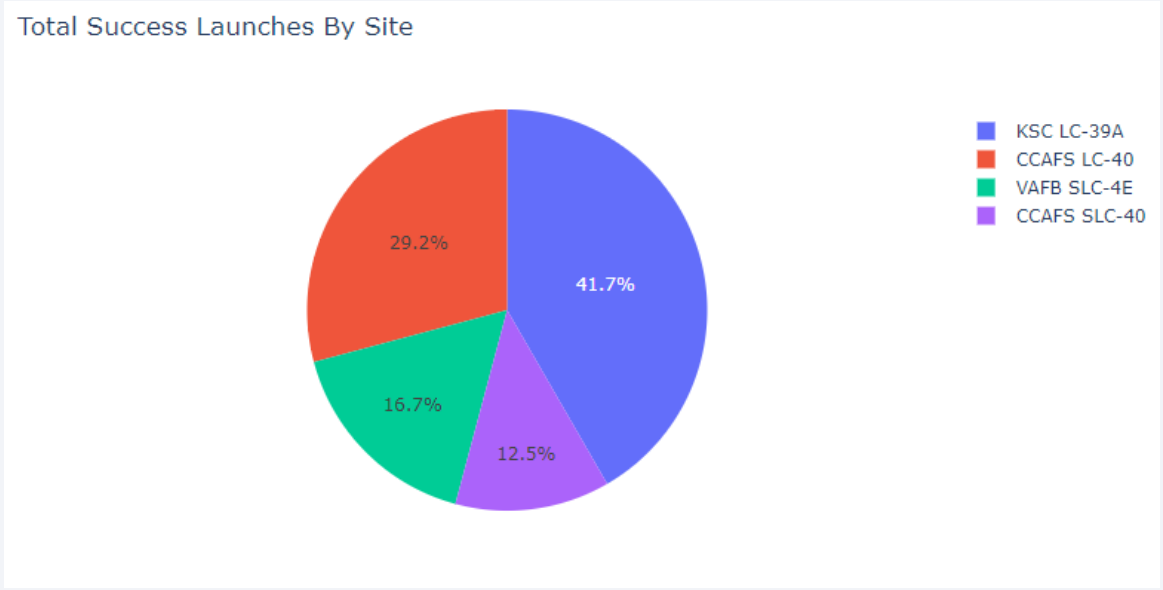
- Here we have exampled the distance of site CCAFS SLC-40 to the shore, we can see the labeled distance of 0.86 km;
- All other sites are distanced similarly to the shore, but much farther to inhabited areas;


Section 4
Build a Dashboard with Plotly Dash

Total Successful Launches

- The following chart shows that the KSC LC-39A site is the one with the most successful launches;

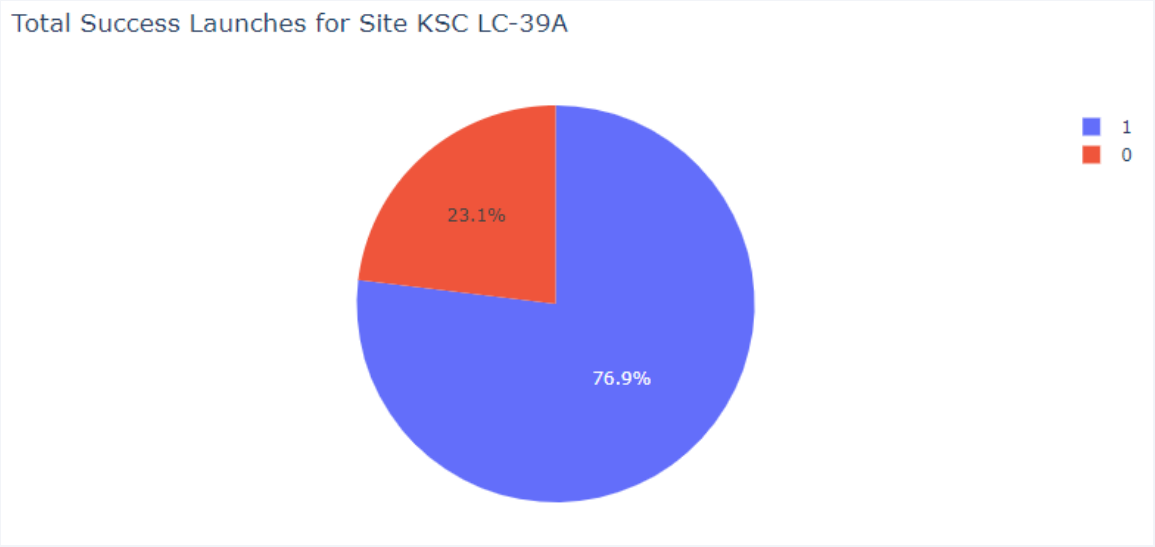
- It is interesting that the CCAFS's have a huge difference in total successful launches.



Success Rate for Site KSC LC-39A

- The following chart shows the success rate for the set KSC LC-39A, which is the one with the most successful launches;

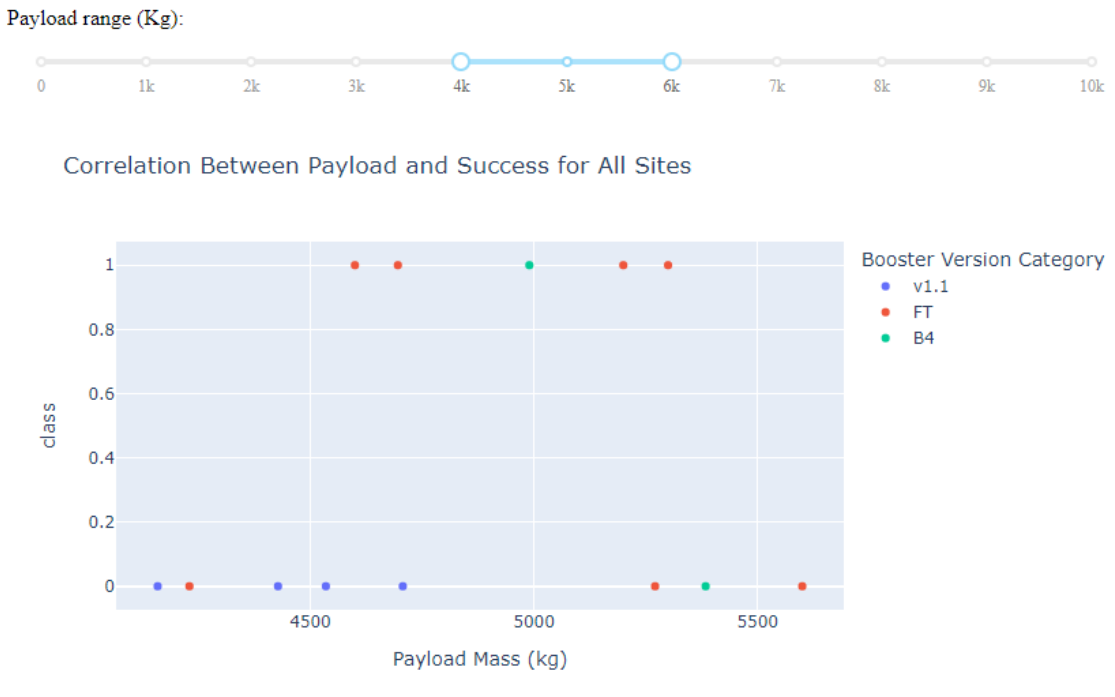
- We can see that it also have a high success rate of 76.9%.



Correlation Between Payload and Success Rate

- Within the range of 4000kg to 6000kg, the FT boosters were the most successful;

- Within the same range all the successes are placed between the 4500kg and 5500kg range.



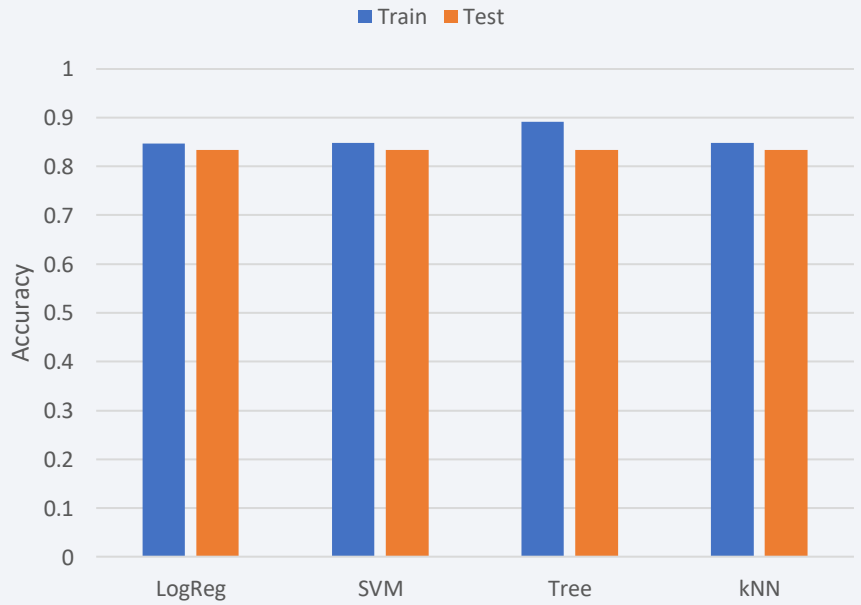
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- In the following chart are the accuracies of the four models trained;

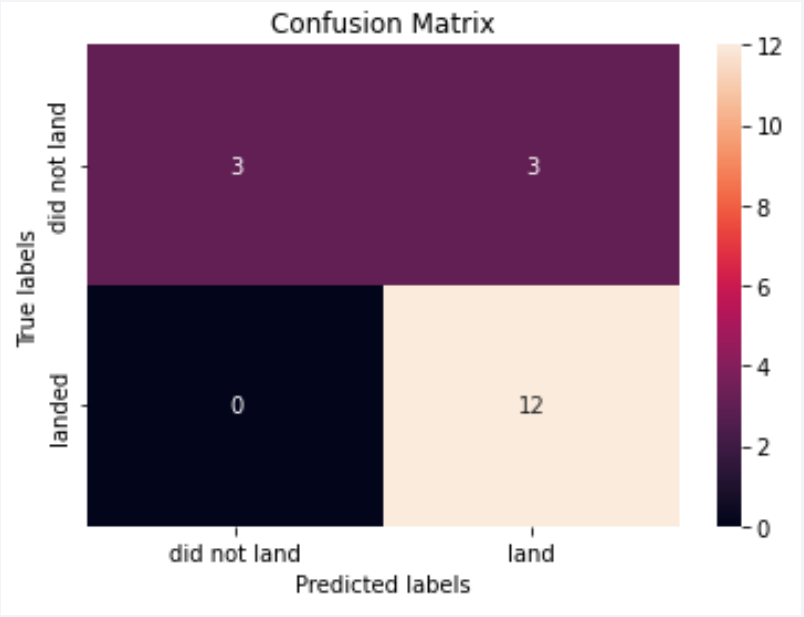
- The model with the highest train accuracy was the decision tree, however all of the models performed the same on the test dataset.



Confusion Matrix

- In the following figure is the confusion matrix for the decision tree model;

- It shows a few false positives (3) and no false negatives, with a high accuracy for correct predictions (15).



Conclusions

- The success rate of the landings has been increasing over time, even with a dip in 2018;
- VLEO seems to be the new go-to orbit for launches, and has a high landing success rate;
- There are four launch sites currently, and they are all located near the shore and far from the general population;
- Even though the landings are not always successful the missions are still 99% successful;
- All models have resulted in a similar prediction for the test data, however, there is a low number of datapoints to test.

Thank you!