Spyros Rigas
9/19/2021

IBM Developer
SKILLS NETWORK

SpaceX Falcon 9:
First Stage Landing Prediction

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions
- Appendix

Executive Summary

- **Methodology**:

  After collection, the data was preprocessed (data wrangling) in order to be prepared for EDA (exploratory data analysis). EDA was performed via SQL querying, as well as data visualizations. This included interactive visualizations via Folium Maps and a Dashboard with Plotly Dash. Finally, machine learning methods (classification) were applied in order to determine the best classifier for the problem at hand.

- **Results**:

  The main results of the EDA process, including the interactive analytics, correspond to the best scoring launch site, the correlation between payload mass and successful landings, the optimal orbits and the specifics of the geographic locations for the launch sites. A yearly trend is also provided for the average success score. As far as predictive analysis is concerned, the best algorithm for classification is determined, including the optimal values for several parameters thereof.

Introduction

- **Project background**

The project is an analysis of SpaceX Falcon 9 rockets' first stage landing procedure. These rockets are designed in such a fashion that their first stage can be recycled and recommissioned in future missions, since, shortly after the launch, it returns back on earth and lands smoothly in order to be recycled.

- **Goal**

Being able to reuse the first stage significantly reduces the production cost of each rocket. As a result, the factors that influence whether the first stage lands successfully or not need to be investigated, so that models which predict future successful/failed landings can be deployed.

Methodology

Methodology

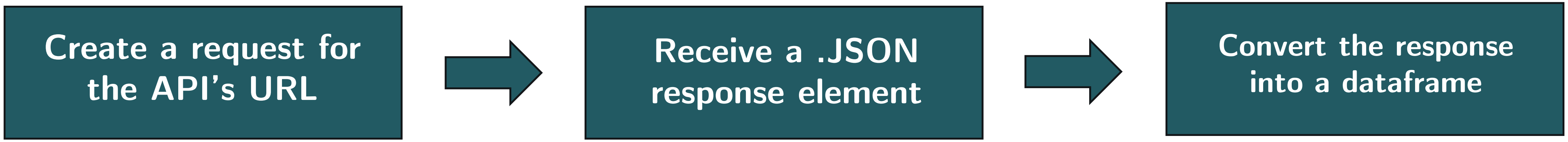
Executive Summary

- Data collection:
  - SpaceX Rest API, Web Scraping from Wikipedia
- Data wrangling:
  - Dealt with missing values, dropped irrelevant columns, performed one-hot encoding
- Exploratory Data Analysis (EDA):
  - SQL querying and data visualization using Seaborn and Matplotlib.
  - Interactive visualizations using Folium and Plotly Dash
- Predictive Analysis (Machine Learning):
  - Tuned and evaluated classification models in order to find the optimal algorithm

Data Collection

The datasets were built by collecting information via two methods:

- SpaceX REST API

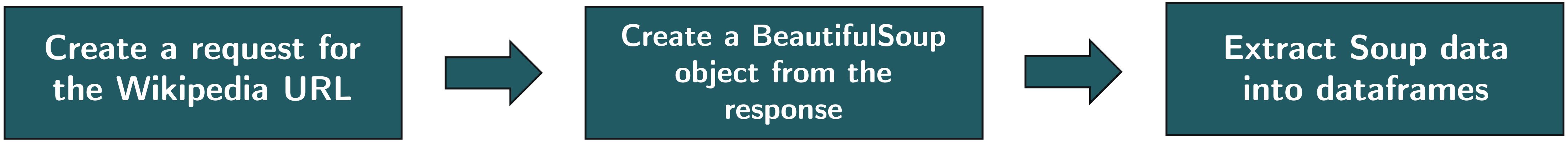


Create a request for the API's URL

Receive a .JSON response element

Convert the response into a dataframe

- Web Scraping from Wikipedia



Create a request for the Wikipedia URL

Create a BeautifulSoup object from the response

Extract Soup data into dataframes

Data Collection – SpaceX API

A simplified flowchart of the process:

- Create request and get response from the API.
- Convert the response into a .json file and subsequently normalize it into a pandas dataframe.
- Keep only a subset of the dataframe and create more requests to expand the dataframe.
- Perform some very basic data cleaning and export the dataframe into a .csv formatted file.

The full process can be found in the corresponding Notebook:

jupyter

Click here for the GitHub link

Data Collection – Web Scraping

A simplified flowchart of the process:

- Create request and get response for the static Wikipedia URL.
- Create a BeautifulSoup object from the HTML response.
- Exctract all the columns and variable names from the table header using the Soup object.
- Parse the HTML content into a dictionary and use the dictionary to create a dataframe.
- Export the dataframe into a .csv formatted file.

The full process can be found in the corresponding Notebook:

Click here for the GitHub link

Data Wrangling

Using the datasets acquired via the aforementioned methods:

- Missing values were replaced by the corresponding column's mean value.
- The values of columns such as Launch Sites (locations where the launches take place) and Orbits (the dedicated orbit to which each launch aims) were calculated.
- Depending on the outcome of each landing (success/failure) a column was created with a categorical value one-hot-encoded into 0 (failure) and 1 (success).

The full procedure can be found in the corresponding Notebook:

Click here for the GitHub link

EDA with SQL

With the preprocessed datasets, SQL queries were performed to determine:

- the unique launch sites.
- five records where the launch sites begin with the string ‘CCA’.
- the total payload mass carried by boosters launched by NASA.
- the average payload mass carried by booster version F9 v1.1.
- the date when the first successful landing outcome in ground pad was achieved.
- the names of the boosters which succeeding in drone ship landings and had payload mass between 4000kg and 6000kg.
- the total number of successful and failure mission outcomes.

EDA with SQL

With the preprocessed datasets, SQL queries were performed to determine:

- the names of the booster versions which carried the maximum payload mass in the dataset.
- the failed landing outcomes in drone ships, their booster versions and the corresponding launch site names, for the year 2015.
- the count of each landing outcome between the dates 2010-06-04 and 2017-03-20, in descending order.

The exact queries is further analyzed in the upcoming Results section and can also be found in the corresponding SQL file:

Click here for the GitHub link

EDA with Data Visualization

Apart from performing SQL queries, Exploratory Data Analysis was performed by creating data visualizations, such as:

- **Scatter Plots**, in order to visualize the impact of one variable to another and get therefore study their correlation.
- **Bar Graphs**, in order to show comparisons among discrete categories (categorical data). While the one axis depicts the specific categories to be compared, the other represents the occurences of each category in the dataset.
- **Line Chart**, in order to clearly depict trends that arise from the correlation between the data.

All of the charts will be seen in the upcoming Results section and can also be found in the corresponding Notebook:

Click here for the GitHub link

Interactive Folium Map

- Using SQL queries, we concluded that there were only 4 unique launch locations. We therefore created an interactive Folium Map, in order to depict these locations by adding **circles** and **markers** to the corresponding coordinates.

- In order to get an understanding of how the launch site impacts the success of the first stage's landing, we created **marker clusters**, with the colors of the markers indicating whether the landing was successful (green) or not (red).

- After calculating the distance to various locations (railways, coastline, cities, etc.), **lines** were drawn on the map in order to depict said distances.

The Folium maps will be presented in the upcoming Results section and can also be found in the corresponding Notebook:

Click here for the GitHub link

Interactive Dashboard

Being able to interact with visualizations is of upmost importance, since it allows us to draw conclusions regarding aspects such as the correlation between variables, without having to deploy more specialized coding algorithms. The Dashboard (created by Plotly Dash) contains two main graphs:

- A **Pie Chart**, depicting either the ratio of successful launches per site to the total successful launches, or the success/failure percentages of each launch site.
- A **Scatter Plot**, showing the relationship between the Payload's mass and the final outcome, for each booster version.

Screenshots of the Dashboard will be presented in the upcoming Results section, while the code can also be found in the following Python file:

Click here for the GitHub link

Predictive Analysis (Machine Learning)

The problem at hand was a classification problem: training a model to be able to determine whether future launches will result in successful landings or not. To that end:

- The data was standardized and split into training (80%) and test (20%) datasets.
- Four different GridSearchCV objects were created, in order to perform a fine tuning of several parameters for four different classification algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (Tree) and k-Nearest Neighbours (KNN).
- After determining the optimal parameters for each model, they were trained for these values of the parameters and evaluated by their accuracy on the test datasets (via the score method, as well as the confusion matrix method).

All of the models will be discussed in the upcoming Results section and can also be found in the corresponding Notebook:

jupyter [Click here for the GitHub link]

Results

Results

Executive Summary

- Exploratory Data Analysis results:
    - Results from SQL Queries
    - Presentation of Data Visualizations
- Interactive Analytics demo in screenshots:
    - Launch Sites Proximities Analysis (Folium)
    - Dashboard with Plotly Dash
- Predictive Analysis Results

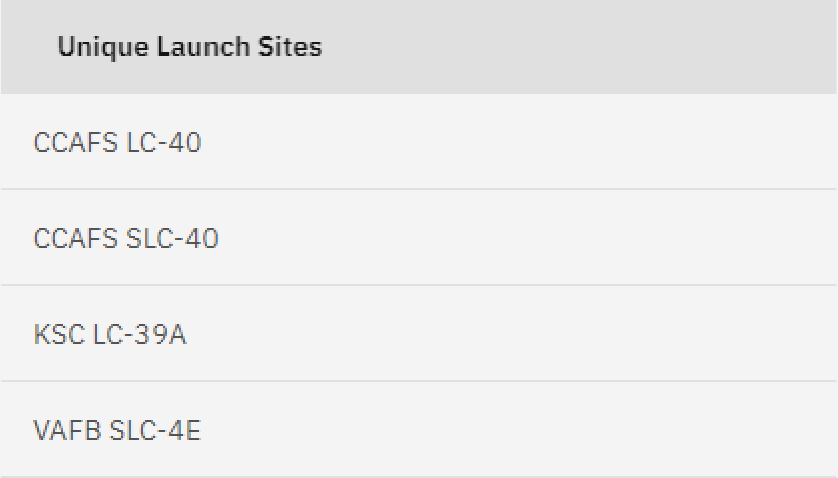
Results from SQL queries

SQL Queries Results

All Launch Site Names

As the 1st task, the following query creates a table with all the distinct launch site names.

```
SELECT DISTINCT Launch_Site
AS "Unique Launch Sites"
FROM SPACEXTBL;
```



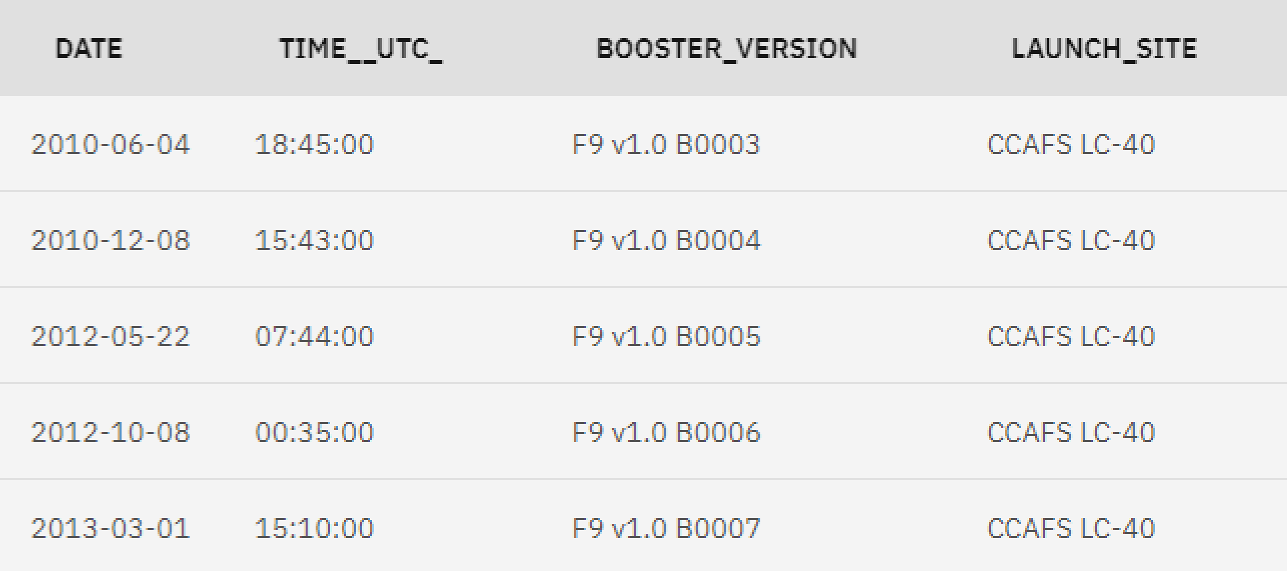
Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

SQL Queries Results

Launch Site Names Begin with ‘CCA’

As the 2nd task, the following query displays 5 records where the launch sites begin with ‘CCA’.

```
SELECT * FROM SPACEXTBL
WHERE Launch_Site
LIKE 'CCA%' LIMIT 5;
```



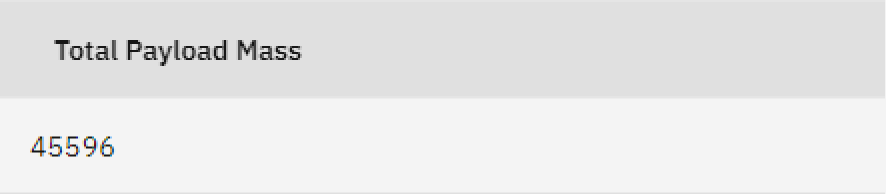
DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

SQL Queries Results

Total Payload Mass

As the 3rd task, the following query displays the total payload mass carried by boosters launched by NASA (CRS).

```
SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass"
FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```



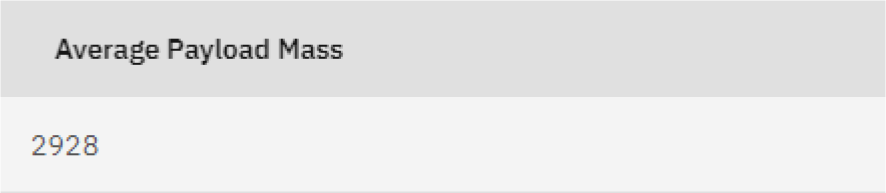
Total Payload Mass
45596

SQL Queries Results

Average Payload Mass by F9 v1.1

As the 4th task, the following query displays the average payload mass carried by booster version F9 v1.1.

```
SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass"
FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
```



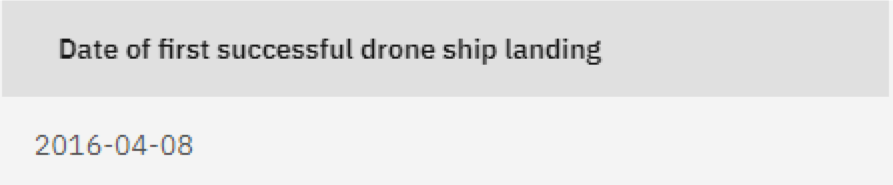
Average Payload Mass
2928

SQL Queries Results

First Successful Ground Landing Date

As the 5th task, the following query lists the date when the first successful landing outcome in ground pad was achieved.

```
SELECT MIN(Date) AS "Date of first successful drone ship landing"
FROM SPACEXTBL WHERE Landing__Outcome = 'Success (drone ship)';
```



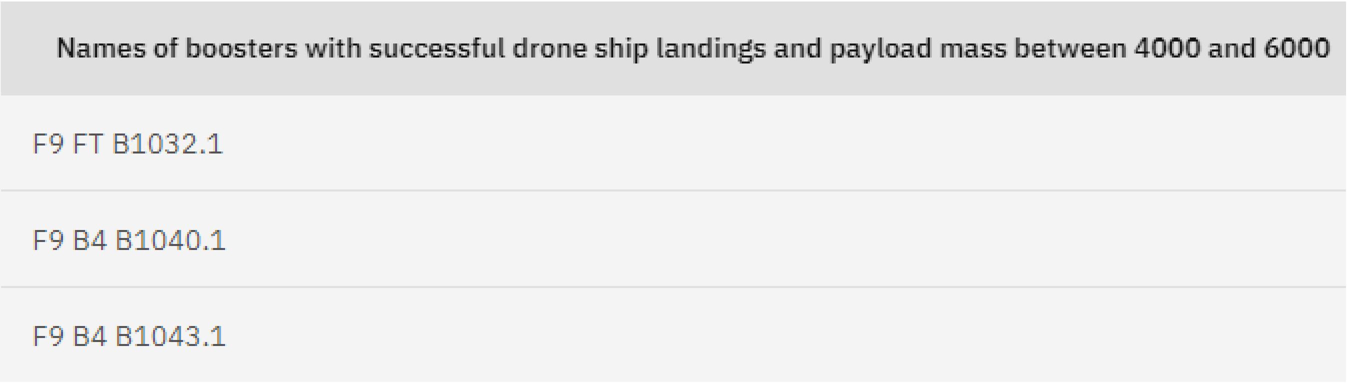
Date of first successful drone ship landing
2016-04-08

SQL Queries Results

Successful Drone Ship Landing with Payload between 4000kg and 6000kg

As the 6th task, the following query lists the names of the boosters which have successful drone ship landings and payload mass between 4000kg and 6000kg.

```
SELECT Booster_Version
AS "Names of boosters with successful drone ship landings and
payload mass between 4000 and 6000"
FROM SPACEXTBL WHERE Landing__Outcome = 'Success (ground pad)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```



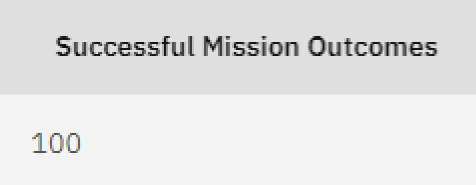
Names of boosters with successful drone ship landings and payload mass between 4000 and 6000
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

SQL Queries Results

Total Number of Successful and Failure Mission Outcomes

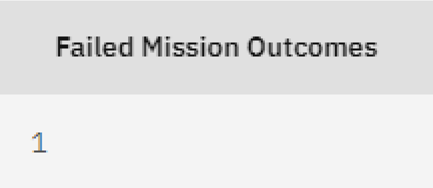
As the 7th task, the following queries list the total number of successful and failure mission outcomes.

```
SELECT COUNT(Mission_Outcome)
AS "Successful Mission Outcomes"
FROM SPACEXTBL
WHERE (Mission_Outcome LIKE '%Success%');
```



Successful Mission Outcomes
100

```
SELECT COUNT(Mission_Outcome)
AS "Failed Mission Outcomes"
FROM SPACEXTBL
WHERE (Mission_Outcome LIKE '%Failure%');
```



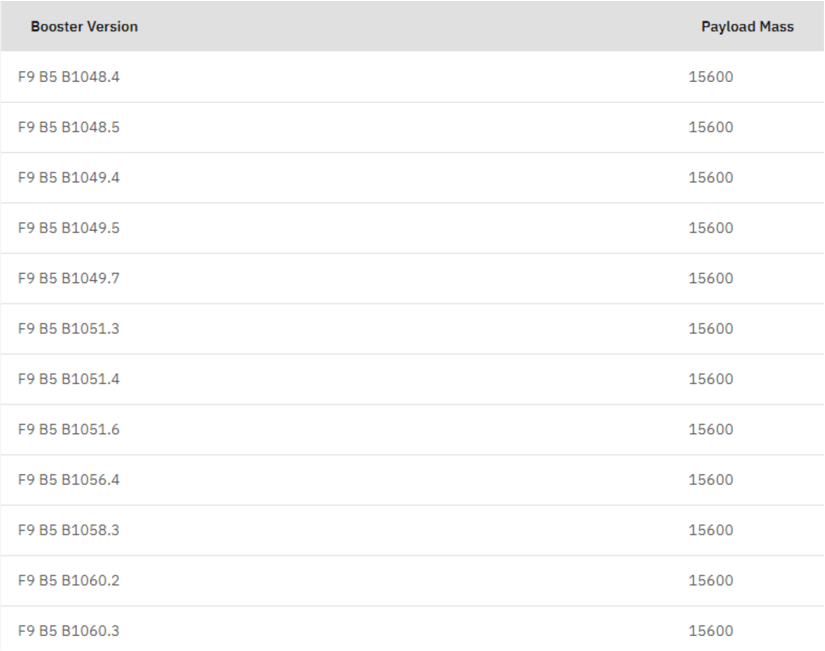
Failed Mission Outcomes
1

SQL Queries Results

Boosters Carried Maximum Payload

As the 8th task, the following query uses a subquery to list the names of the booster versions which have carried the maximum payload mass.

```
SELECT DISTINCT Booster_Version AS "Booster Version",
PAYLOAD_MASS__KG_ AS "Payload Mass"
FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```



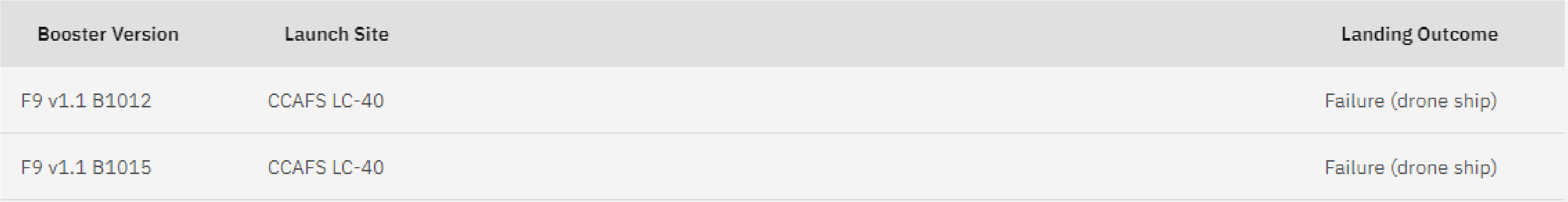
Booster Version	Payload Mass
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

SQL Queries Results

2015 Launch Records

As the 9th task, the following query lists the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015.

```
SELECT Booster_Version AS "Booster Version",
Launch_Site AS "Launch Site", Landing__Outcome AS "Landing Outcome"
FROM SPACEXTBL WHERE (Landing__Outcome LIKE '%Failure%')
AND DATE LIKE '%2015%';
```



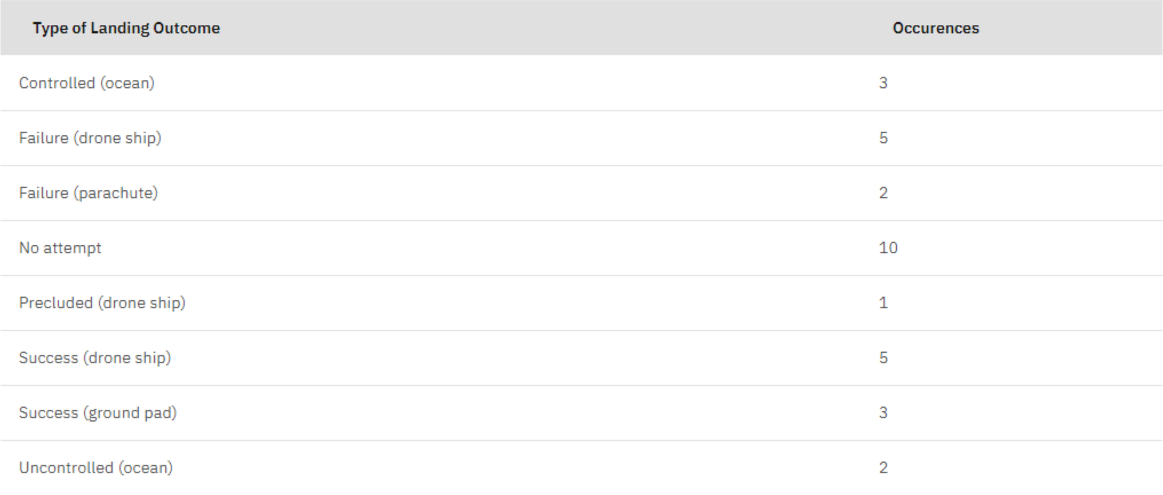
Booster Version	Launch Site	Landing Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL Queries Results

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

As the 10th task, the following query ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
SELECT Landing__Outcome
AS "Type of Landing Outcome",
COUNT(Landing__Outcome)
AS "Occurences"
FROM SPACEXTBL
WHERE (Date >= '2010-06-04')
AND (Date <= '2017-03-20')
GROUP BY Landing__Outcome
ORDER BY "Occurences" DESC;
```

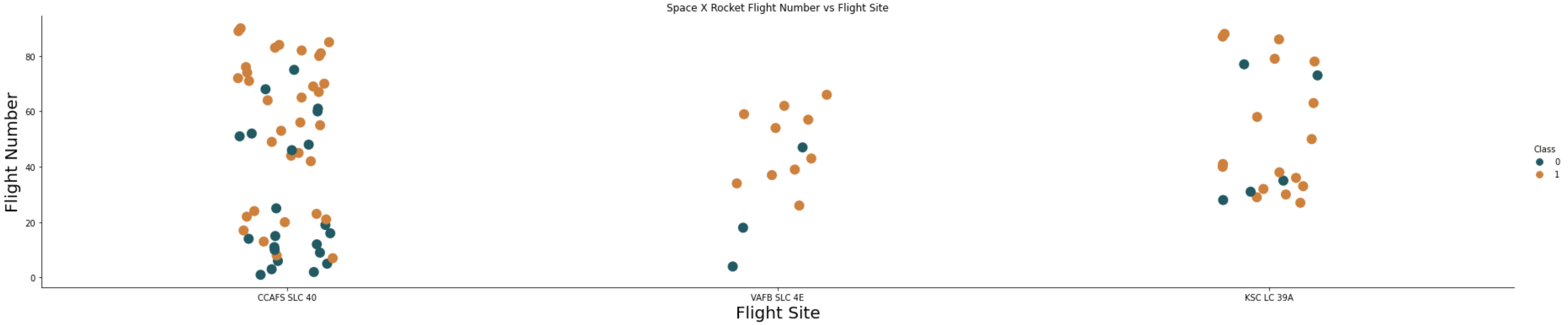


Type of Landing Outcome	Occurences
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

Presentation of Data Visualizations

Data Visualizations

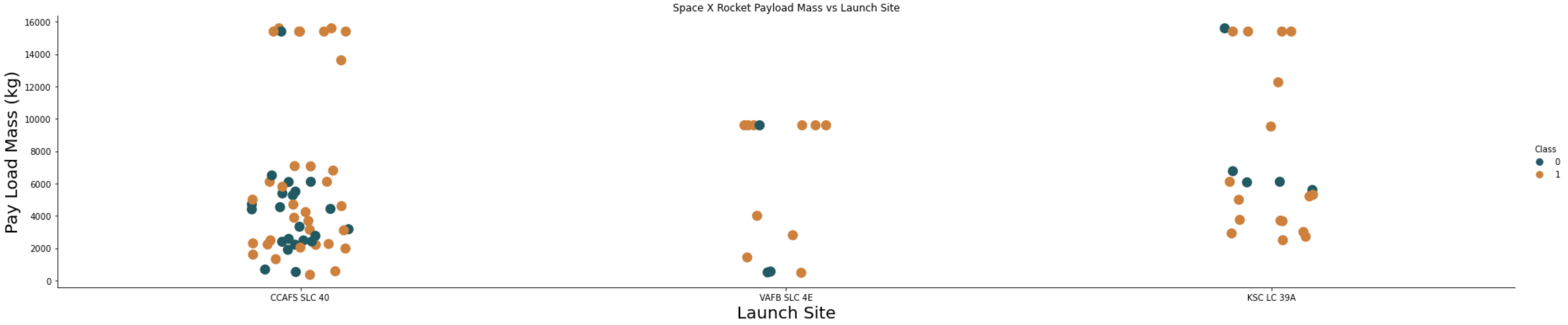
Flight Number vs. Launch Site



From this graph it becomes evident that significantly more launches have been performed from the CCAFS SLC 40 launch site, compared to the other two. However, the other two launch sites appear to have a better score of successful first stage landings.

Data Visualizations

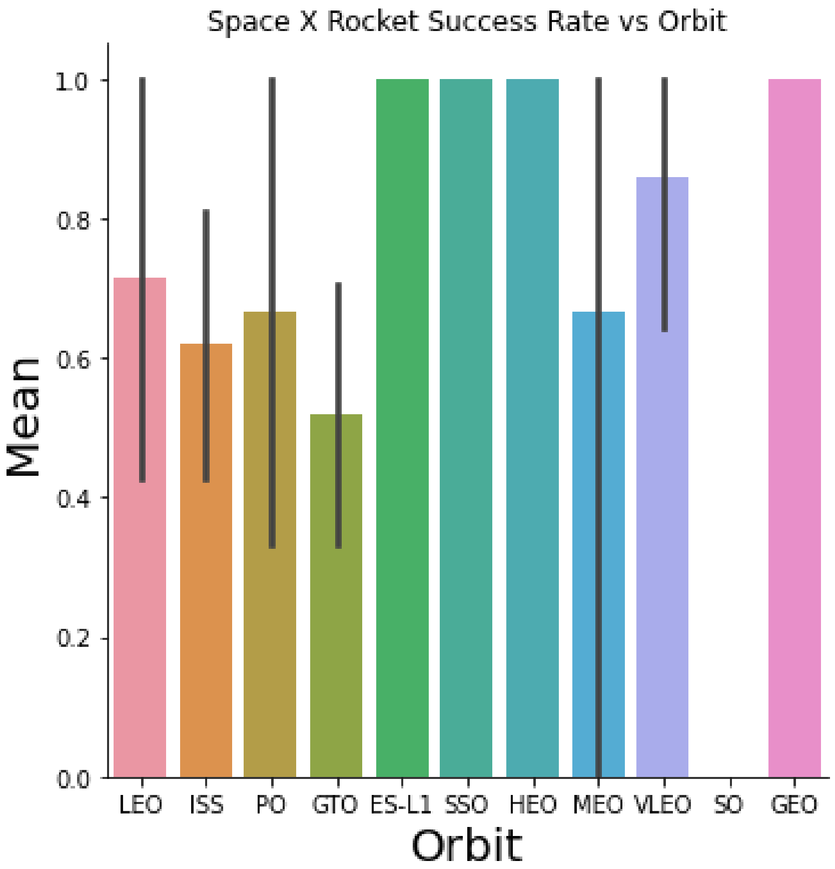
Payload vs. Launch Site



The insight acquired from the above graph is that relatively higher Payload masses tend to have a positive correlation with successful landings. This is very clear specially for the first launch site, whose class contains more elements compared to the other two.

Data Visualizations

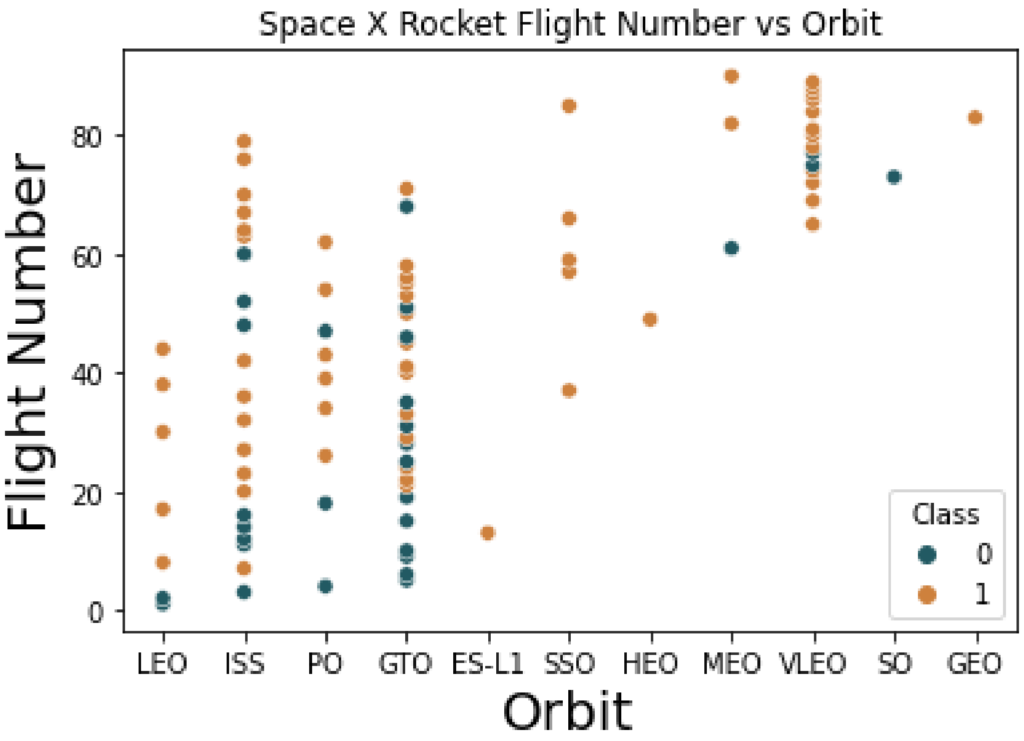
Success Rate vs. Orbit Type



This bar graph indicates that the success rate is maximized in the case of the ES-L1, SSO, HEO and GEO Orbits, while the exact opposite is true for the SO Orbit. The results for the remaining Orbits are mixed and the variance significantly varies, however it’s safe to say that the fact that no bar graph falls below 50% is encouraging.

Data Visualizations

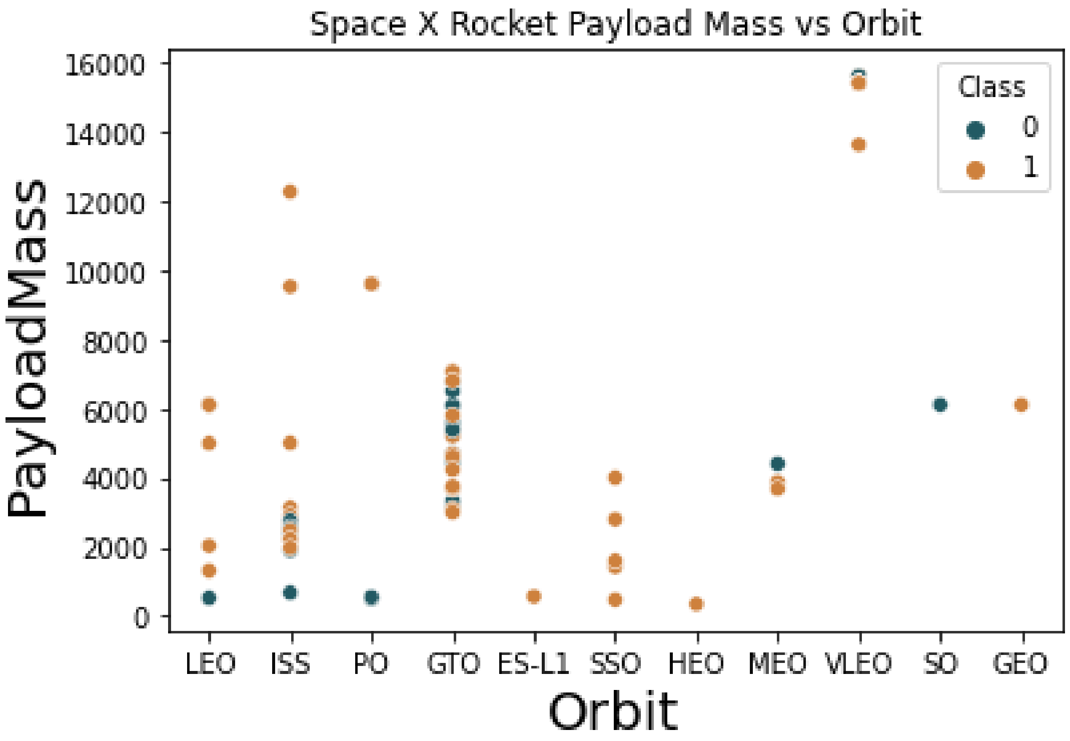
Flight Number vs. Orbit Type



This Scatter Plot gives us more insight regarding the conclusions we previously drew, based on the Bar Chart. As far as the ES-L1, HEO and GEO Orbits are concerned, the fact that only one data point belongs to each of them indicates that we have a significantly high uncertainty regarding their efficiency. On the other hand, the SSO Orbit remains a very good candidate, since it has a 100% success rate for a total of 5 missions – not a rather large number, but certainly larger than 1.

Data Visualizations

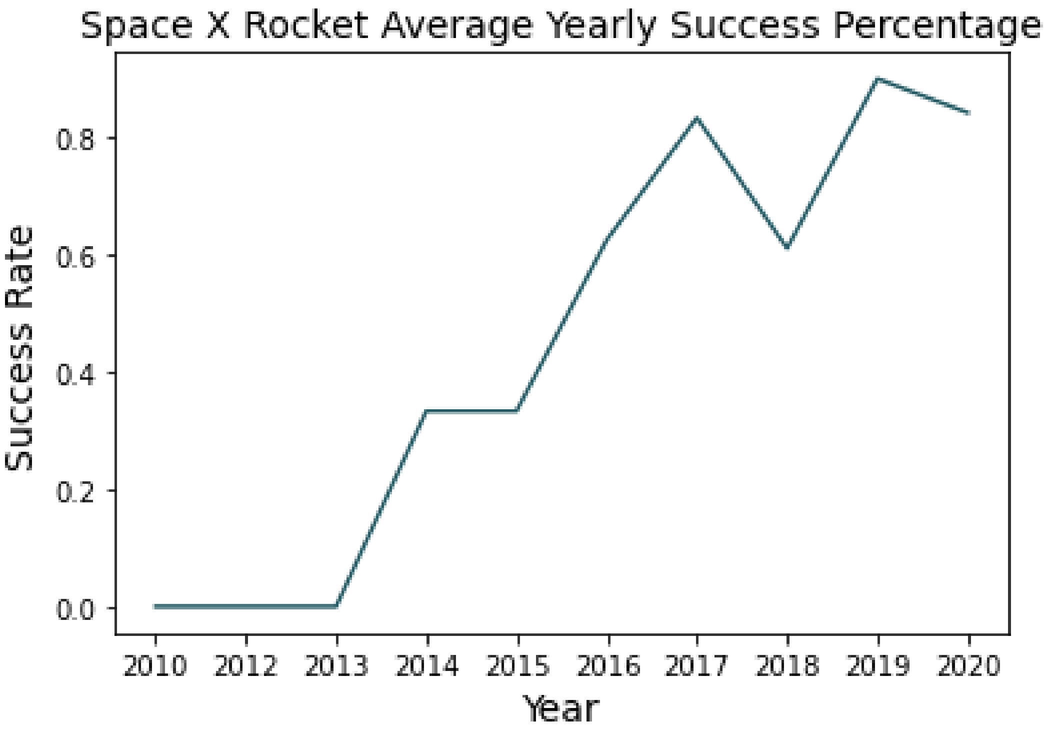
Payload vs. Orbit Type



We previously observed that higher payload masses were positively correlated with successful landings, a trend which is reflected by this Scatter Plot as well. Notice, however, that there is an exception, namely the GTO Orbit, which seems to score better for lower payload mass values.

Data Visualizations

Launch Success Yearly Trend

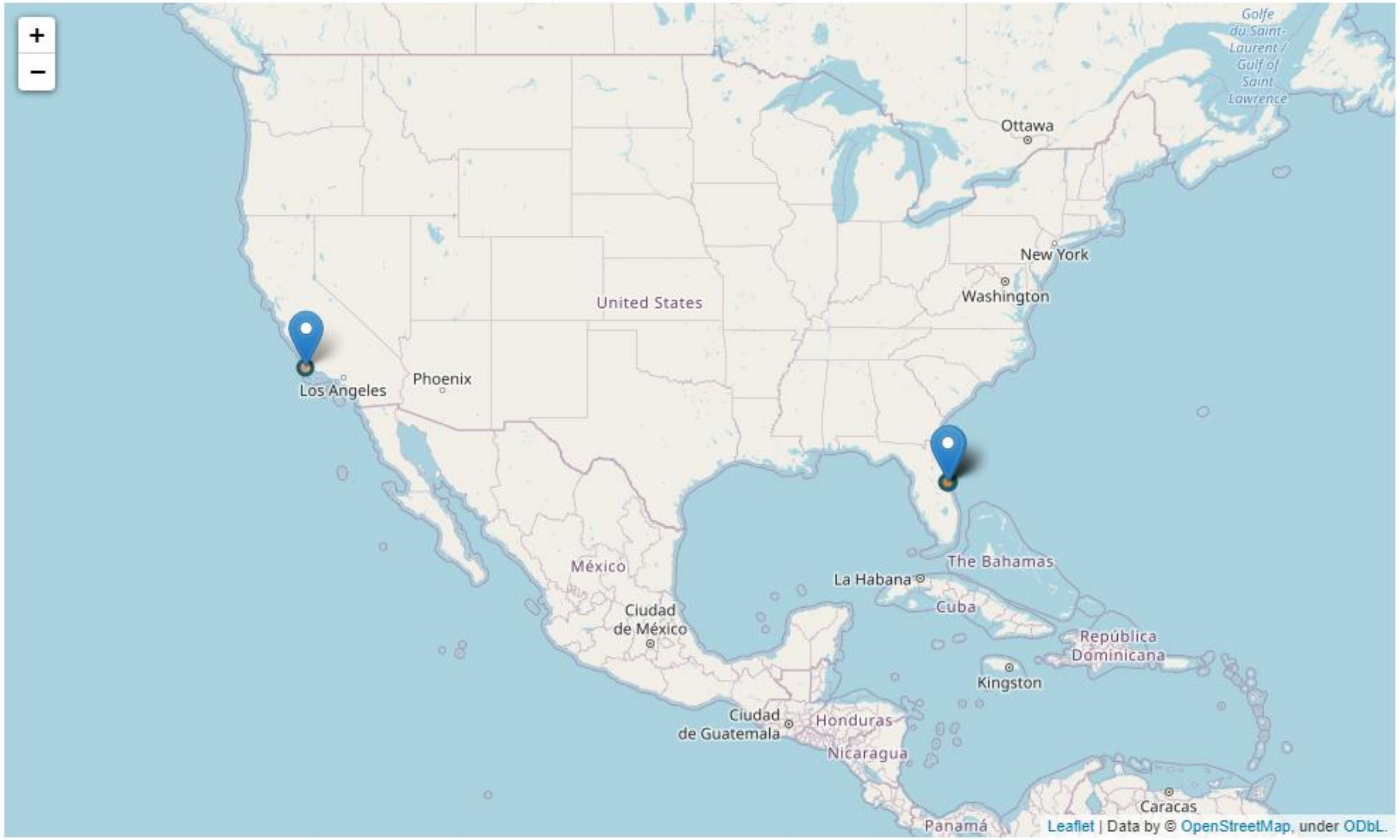


This final visualization gives us some insight on the general trend of the SpaceX First Stage Landing Success Percentage. While not strictly monotone, there is a clear overall increase in the average yearly success of the landings, with 2019 holding the best results so far.

Launch Sites Proximities Analysis (Folium)

Folium Maps

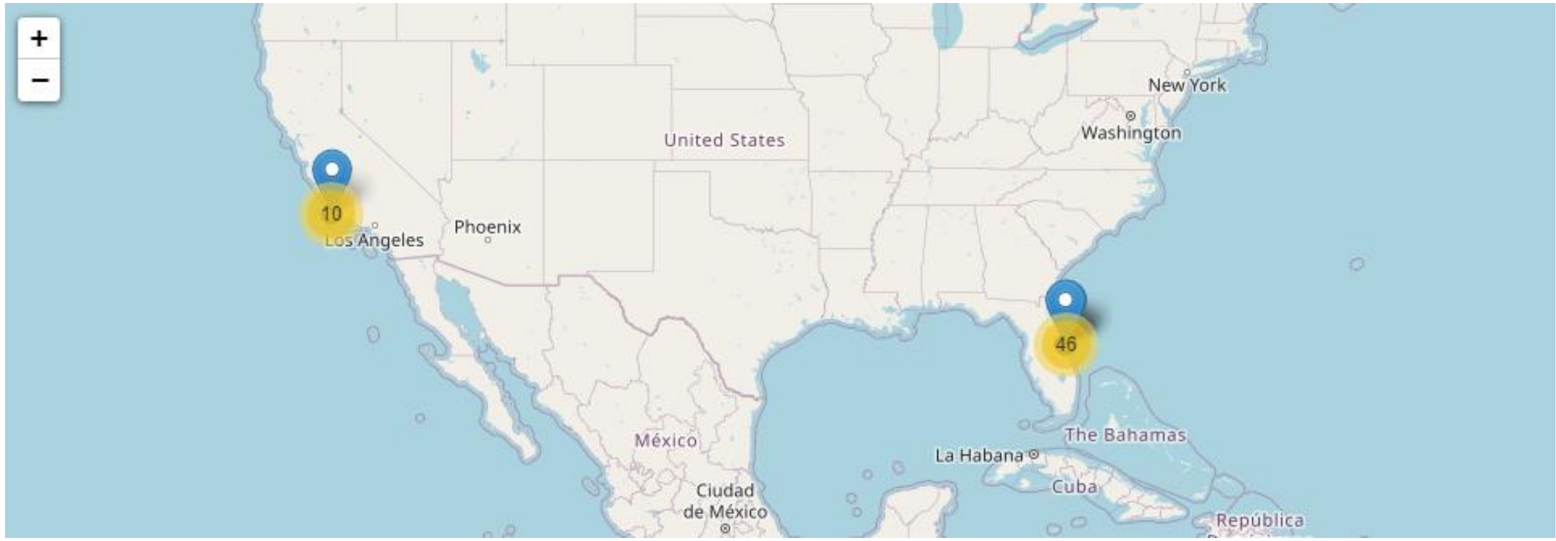
Launch Site Locations on Folium Map

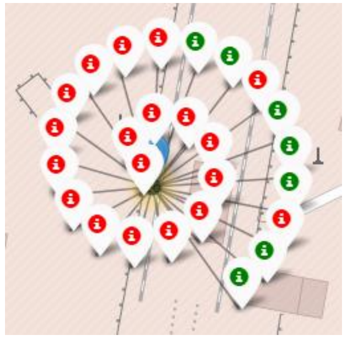


All four launch sites can be seen in coastal areas in USA. One of them (VAFB SLC-4E) is located in California, while the remaining three (KSC LC-39A, CCAFS SLC-40, CCAFS LC-40) are located in Florida.

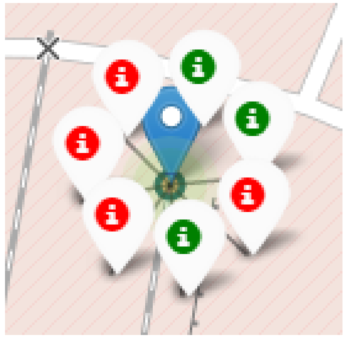
Folium Maps

Launch Sites Clustered and Labelled depending on Success

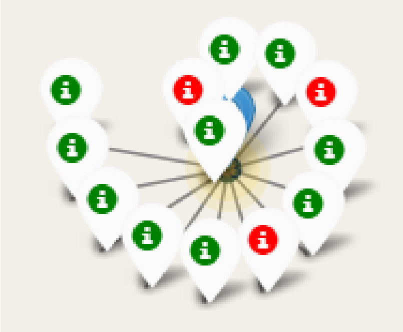




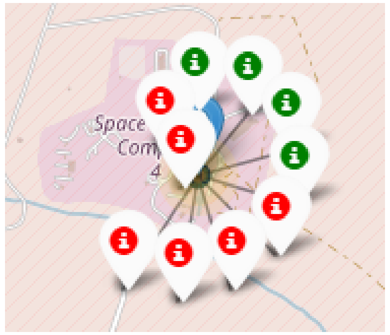
CCAFS LC-40



CCAFS SLC-40



KSC LC-39A

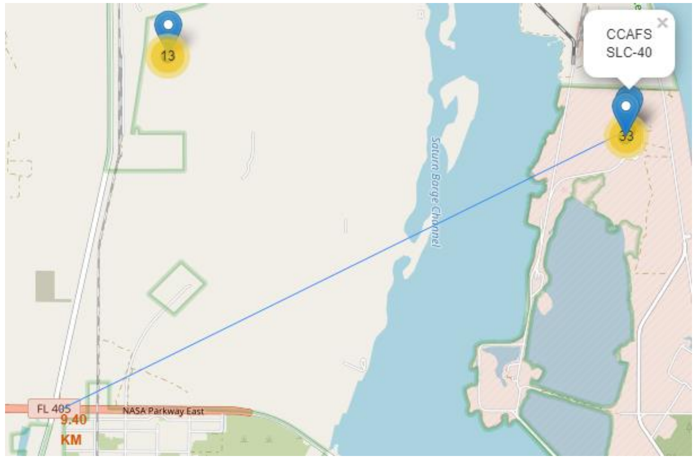


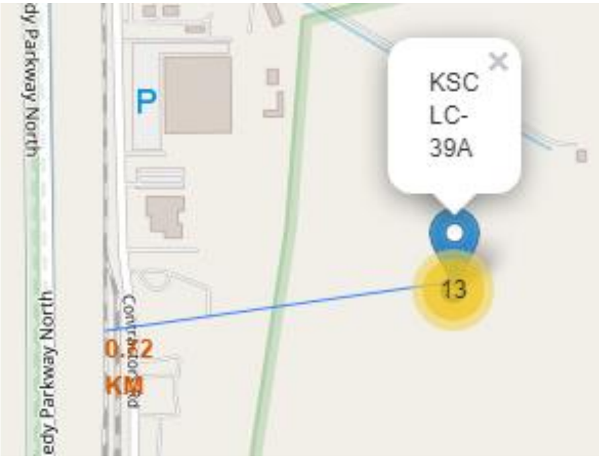
VAFB SLC-4E

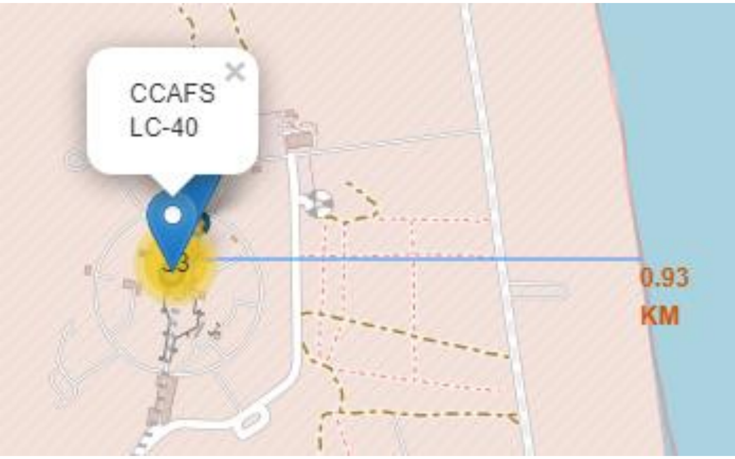
The screenshot of the map depicts how the launch sites have been clustered, depending on location. The following four screenshots depict the number of successful (green) and failed (red) first stage landings. The best score clearly belongs to , with only 3 failures out of 13 total launches.

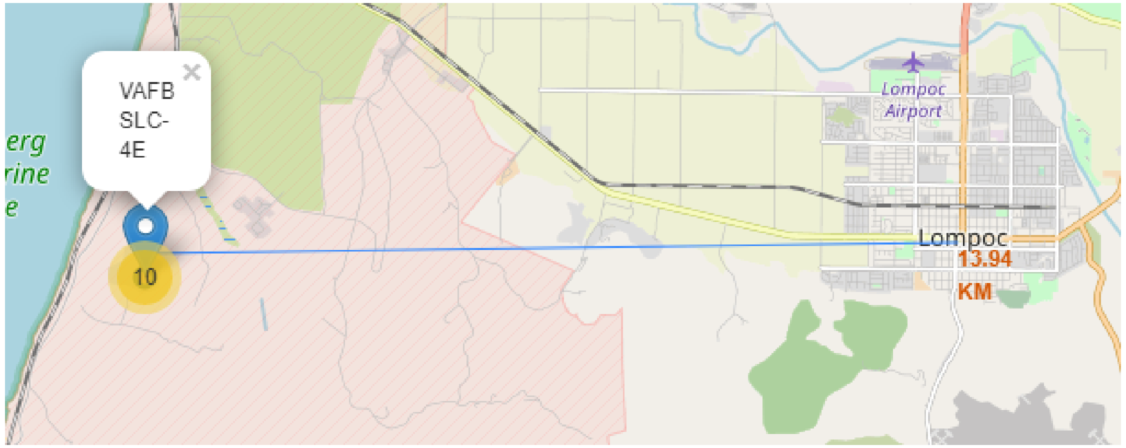
Folium Maps

Launch Site Distance from various Locations









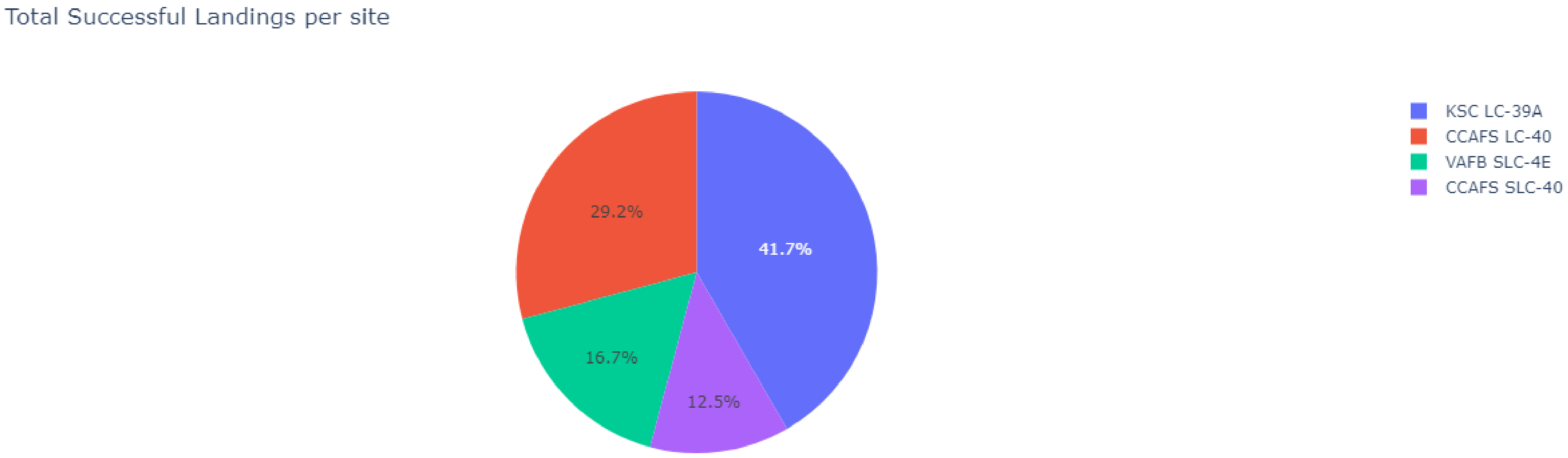
Each of the screenshots depicts the distance of one launch site from a location of importance (highway, railway, coast, city). It's clear that launch sites are chosen so that they are as far away from cities and highways as possible. On the other hand, they appear to be located very close to the coast.

```
<script>
    import sidebarController from './sidebarController.js'
    import mementoes from '../../store.js'
    import './sidebar.scss';
</script>

<div class="sidebar">
    <button class="add-memento" on:click=

    <ul class="mementoes">
        {#each $mementoes as memento}
            <li class="memento-item"
                on:click={() => sidebarController
                {memento.title}
            </li>
        {/each}

</div>
```

Dashboard with Plotly Dash

Dashboard with Plotly Dash

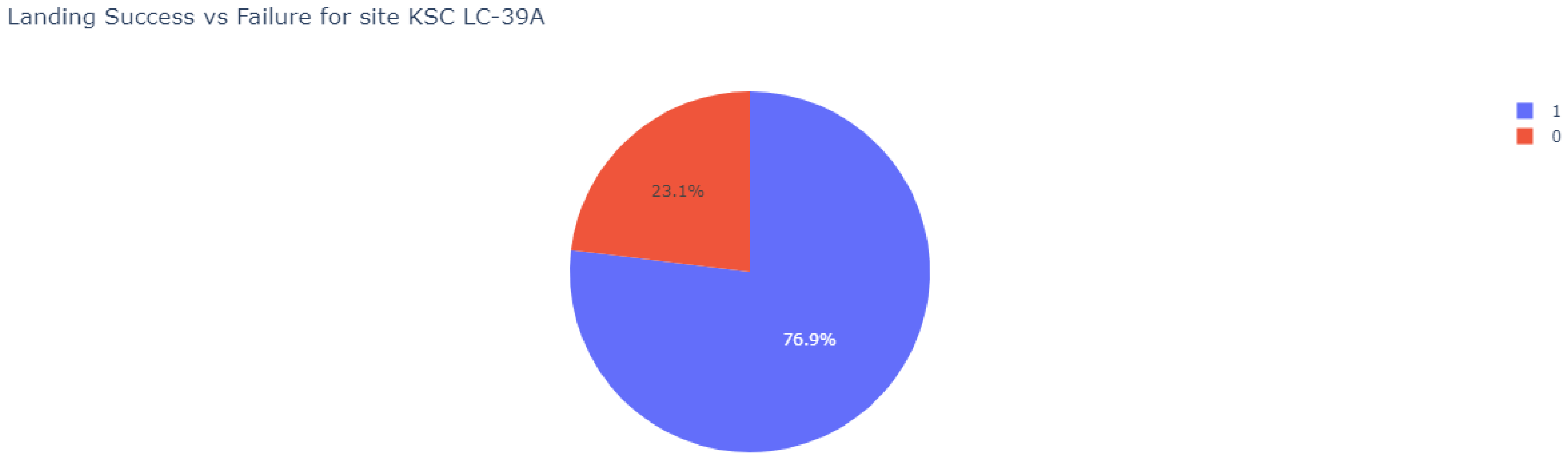
Pie Chart with launch success count for all sites



It's clear that the launch site with the most successful landings is **KSC LC-39A**. Of course, we need to further study the sites separately in order to determine which one corresponds to the highest launch success ratio (the high percentage here could simply imply that many more launches were performed in the **KSC LC-39A** site compared to other sites).

Dashboard with Plotly Dash

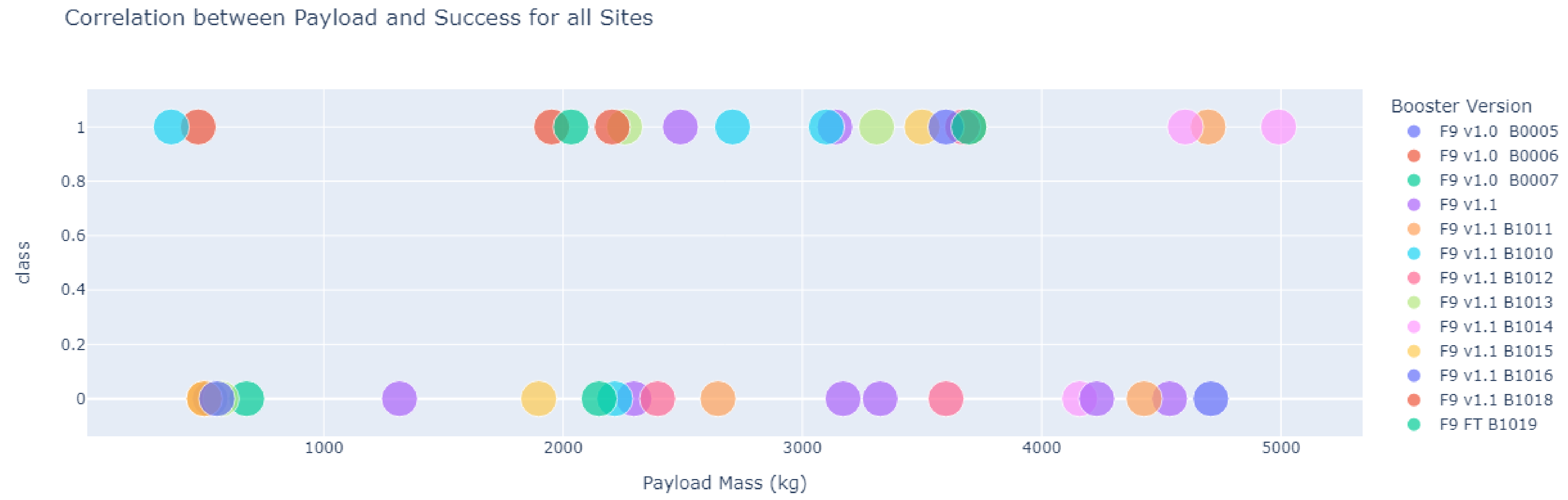
Pie Chart for the launch site with the highest launch success ratio (KSC LC-39A)



After observing the resulting pie chart for each site, it appears that **KSC LC-39A** indeed is The site with the highest landing success score, at a 76.9% percentage, compared to **CCAFS SLC-40** (with a 57.1% success score), **VAFB SLC-4E** (with a 60% success score) and **CCAFS LC-40** (with a 73.1% success score).

Dashboard with Plotly Dash

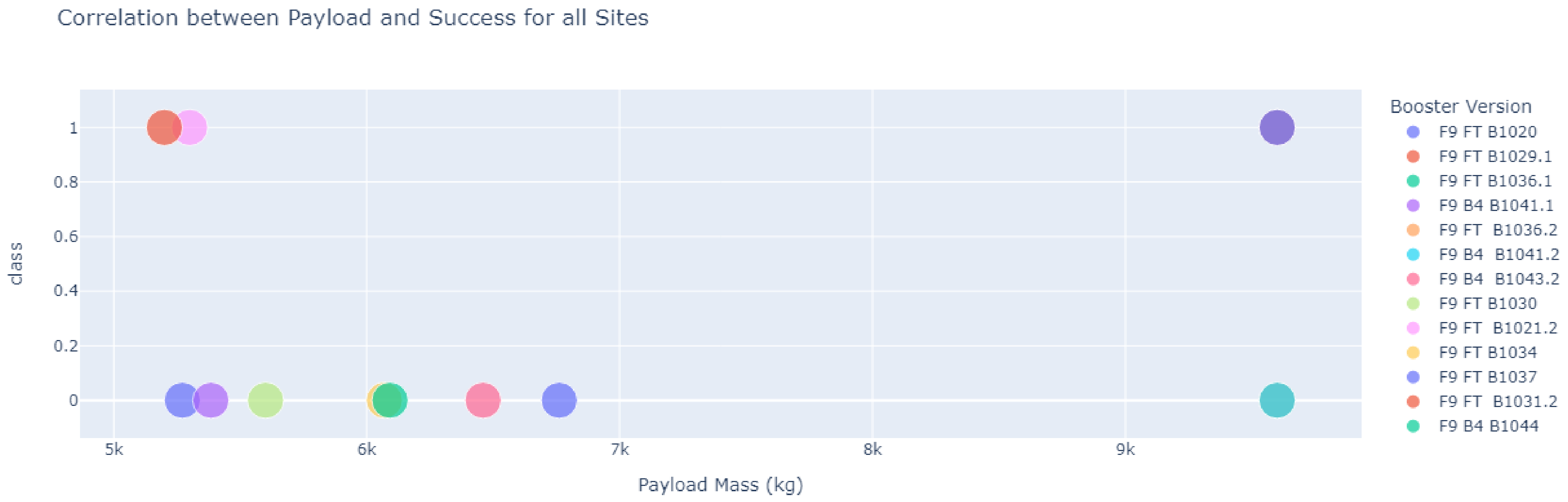
Scatter Plot of Payload vs. Launch Outcome (low payloads)



For low payloads (masses between 0 and 5000 kg) the successful landings appear to be near the 2000kg - 4000kg payload mass range, with an overall trend not being clear.

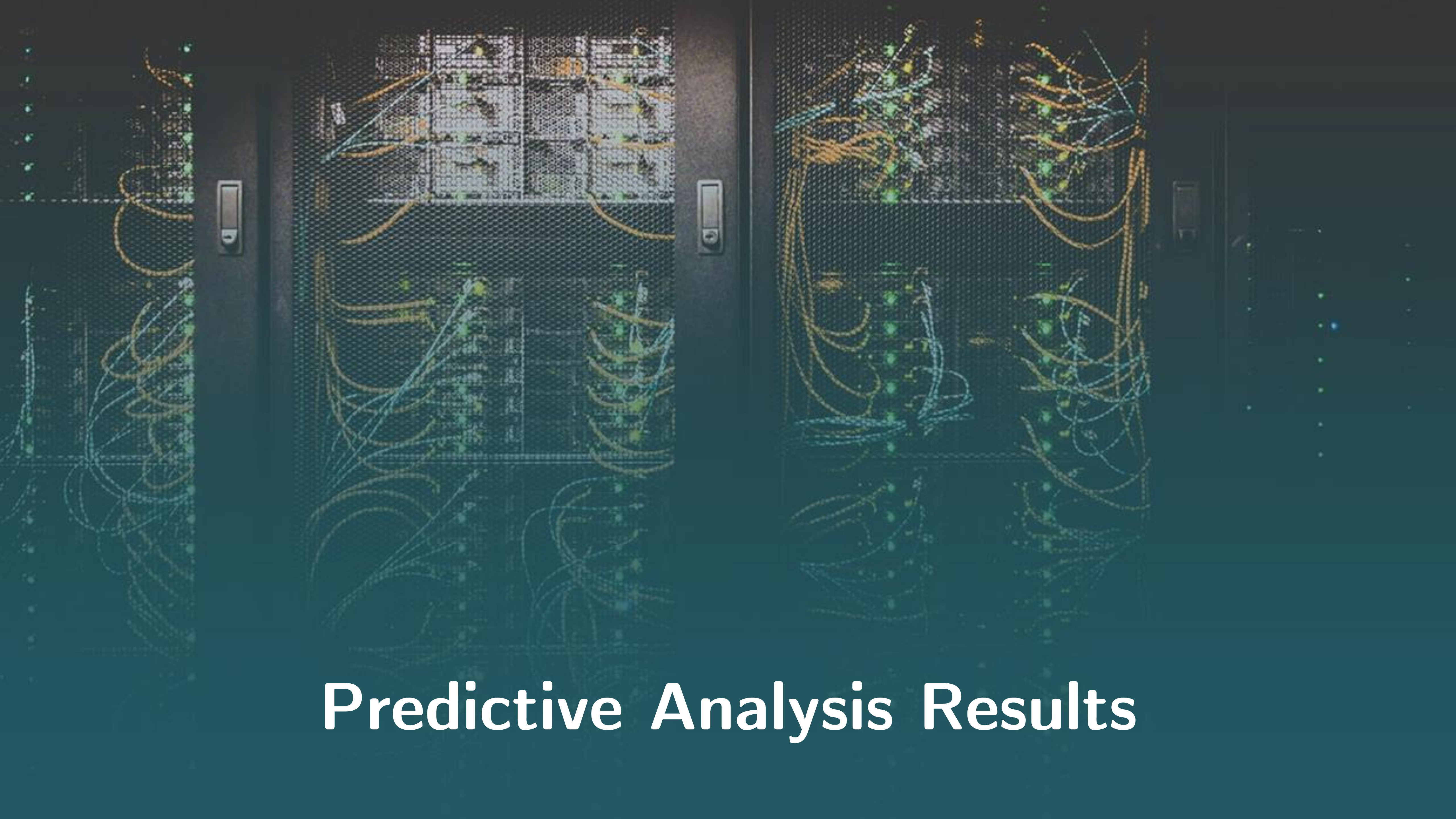
Dashboard with Plotly Dash

Scatter Plot of Payload vs. Launch Outcome (high payloads)



For high payloads (masses between 5000kg and 10000kg) the the landings tend to fail more often, Especially after crossing the 6000kg threshold (with a singular exception).

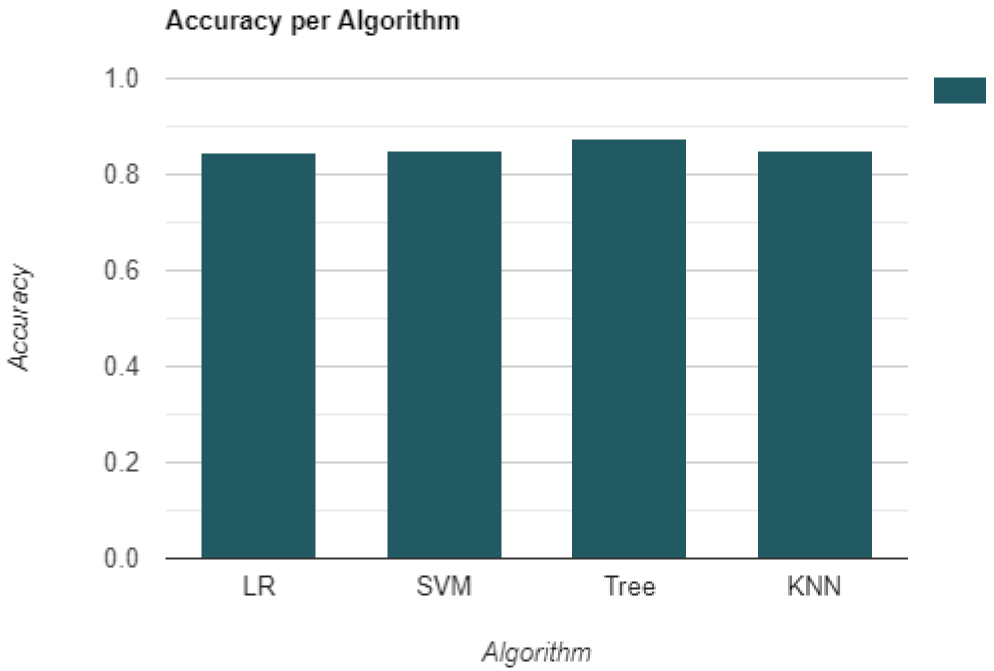
Predictive Analysis Results



Predictive Analysis Results

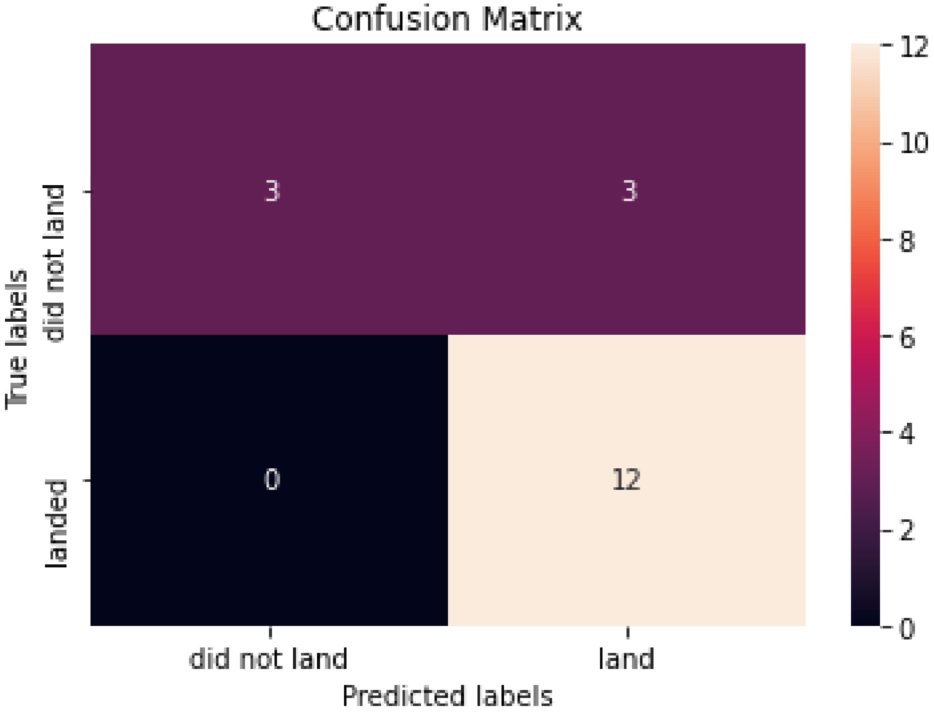
Classification Accuracy

After tuning each model's parameters to their corresponding optimal values, the accuracy achieved by each model has been plotted in the bar chart where LR stands for Logistic Regression, SVM for Support Vector Machine, Tree for Decision Tree and KNN for k-Nearest Neighbours. While all algorithms score values close to 0.84, the highest scoring algorithm is the Decision Tree with an accuracy of ≃ 0.877.



Predictive Analysis Results

Confusion Matrix



Since the Decision Tree appears to be the highest-scoring algorithm, we evaluate it by plotting the Confusion Matrix corresponding to its predictions on the test data. While the Tree appears to be struggling a bit with the False Positive values, it successfully predicts all of the true landings.

Conclusions

Conclusions

Before closing this report, we present some of the main Conclusions drawn from our Analysis.

- While the CCAFS SLC-40 site was chosen for more launches, the KSC LC-39A had the highest success rate for first stage landings.

- At first sight, the SO Orbit appears to yield bad results, however the data is not sufficient to draw such conclusions. Similarly, the GEO, HEO and ES-L1 Orbits appear to yield good results, but the data is insufficient in these cases as well. There is higher confidence that the SSO and LEO orbits perform well.

- More successful landings occur when the payload masses are not relatively high. Especially after crossing the 6000kg threshold, failed landings are very common.

Conclusions

Before closing this report, we present some of the main Conclusions drawn from our Analysis.

- While the trend is not strictly monotone, as the years pass it appears that the first stage landings tend to score better compared to past years.

- Judging from their locations, launch sites are usually chosen to be near the coastline and simultaneously far from main highways or cities.

- The Decision Tree classifier appears to be the best model to deploy in the future in order to predict successful and failed landings for future missions. Of course, the other three algorithms had accuracies similar to the Decision Tree's.

Appendix

Note at this point that in order to perform SQL queries using IBM's Db2 instance, the data had to have specific date and time formats in order to be uploaded to the servers. At this point, a script was written in order to transform the dates into a format suitable for upload. The script can be accessed in the following link.

jupyter

Click here for the GitHub link

Thank you!