Winning the Space Race with Data Science

Charalambos Tsioutis
20-Mar-2022

IBM Developer
SKILLS NETWORK

Outline

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

Executive Summary

- Methodology:

  Using data wrangling methods we managed to collect data from SpaceX API and SpaceX Wikipedia page. At the next stage, we proceeded to EDA (exploratory data analysis) via SQL querying, as well as data visualizations. In this phase, we presented interactive visualizations from Folium Maps and a Dashboard via Plotly Dash. We concluded our report applying machine learning methods to determine the best classifier for the task.

- Results:

  According to EDA results, we managed to determine the optimal scoring launch site, the correlation between payload mass and successful landings as well as the optimal orbits and geographic locations for the launch sites. Although the best classification algorithm for classification is determined, all models produced similar results and accuracy rates.

Introduction

- Project background:

  This project is focused on the analysis of SpaceX Falcon 9 rocket's first stage. SpaceX designed the rockets in a way that the first stage can be recycled and reused in future procedures. The first stage of the rocket after the launch can land back to earth and be reused.

- Problem:

  The ability to reuse the first stage of the rocket launch puts SpaceX ahead of the competition due to the fact that the cost of production is reduced significantly. Although the recycling of the first age is a revolutionary development, the first stage does not always successful. Our goal is to examine the factors that affect the success of the landing and deploy models that predict with accuracy the result of the launching.

Methodology

Section 1

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.

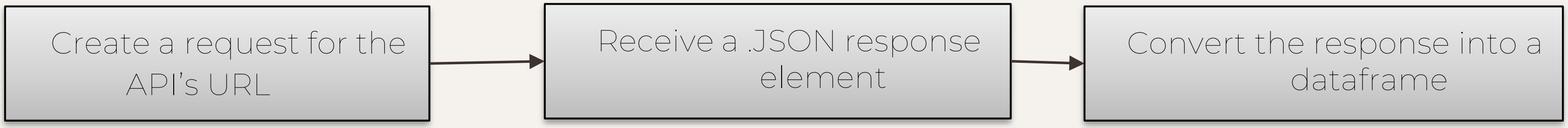
  Interactive visualizations using Folium and Plotly Dash

- Predictive Analysis (Machine Learning):

  Tuned and evaluated classification models in order to find the optimal algorithm

Data Collection

SpaceX REST API

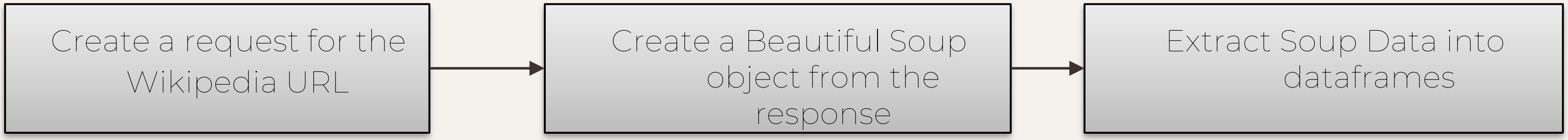


Create a request for the API's URL

Receive a .JSON response element

Convert the response into a dataframe

Web Scraping from Wikipedia



Create a request for the Wikipedia URL

Create a Beautiful Soup object from the response

Extract Soup Data into dataframes

Data Collection – Web Scraping

The process can be described with the following flowchart:

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

A link to the process Notebook:

Github Link

Data Collection – Data Wrangling

The method we used is presented in the following flowchart:

- 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).

A link to the process Notebook:

Github Link

EDA with SQL

With acquired datasets we performed SQL queries 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 the ground pad was achieved.
- The names of the boosters which succeeded in drone ship landings and had payload mass between 4000kg and 6000kg.
- The total number of successful and failure mission outcomes.
- 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.

A link to the SQL file:

jupyter

Github Link

EDA with Data Visualization

We utilize Exploratory Data Analysis by creating data visualizations, such as:

- Scatter Plots, with the purpose of correlating the two variables in hand and exploring the impact of one variable on another.
- Bar Graphs, in order to compare discrete categorical data. One axis represents the categories that we want to compare, while the other represents the occurrences of each category in the dataset.
- Line Chart, to present the trends that arise from the correlation between the data.

A link to the Notebook file:

Github Link

Interactive Folium Map

- Utilizing SQL queries we find that there were four launch locations. Then we created an interactive Folium Map and illustrated these locations with circles and markers to the corresponding coordinates.
- Creating marker clusters, with green and red colors for successful or unsuccessful landings respectively, we showed how the launch location can impact the success of the first stage landing.
- We illustrated the distances from various locations such as railways, coastlines, and cities by drawing lines on the corresponding map.

A link to Folium Maps file:

Github Link

Interactive Dashboard

Visualizations allow us to draw conclusions regarding the correlation between variables, without having to deploy more specialized coding algorithms. The Plotly Dashboard contains two main graphs:

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

A link to the Plotly Dashboard code:

Github Link

Predictive Analysis (Machine Learning)

In this section, we deployed a classification problem. We trained a model to be able to determine the result of future launches. To accomplish that we:

- Standardized and split the data into training and test datasets (80% and 20% respectively).
- Created four GrindSearchCV objects and performed a fine-tuning of several parameters for four different classification algorithm: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (Tree) and k-Nearest Neighbours (KNN).
- Determined the optimal parameters for each model and then we trained the models using the values of the obtained parameters and evaluated their accuracy on the test datasets.

A link to the ML code:

Github Link

Results

Section 2

Results

Executive Summary

- Exploratory Data Analysis results
    - SQL Queries results
    - Data Visualization presentation
- Interactive Analytics demo in screenshots:
    - Folium Maps, launch site analysis
    - Plotly Dash dashboard
- Predictive Analysis Results

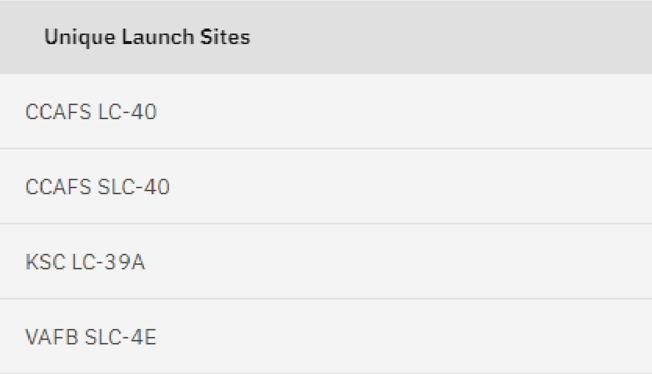
Results from SQL queries

SQL Queries Results

All Launch Site Names

1st task: Create a table with all distinct launch site names

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



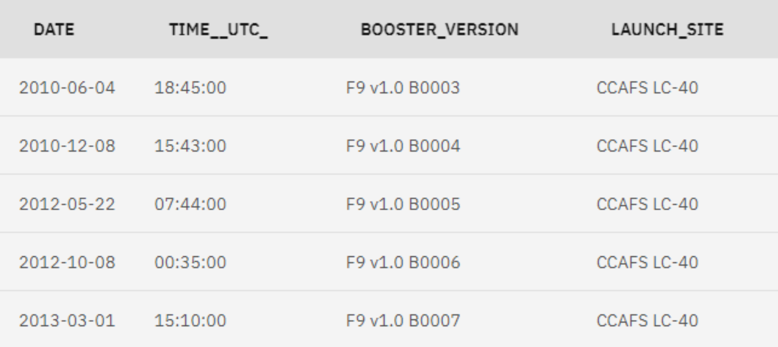
Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

SQL Queries Results

Launch Site Names Begin with ‘CCA’

2nd task: Query that displays 5 records where the launch sites begin with ‘CCA’.

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



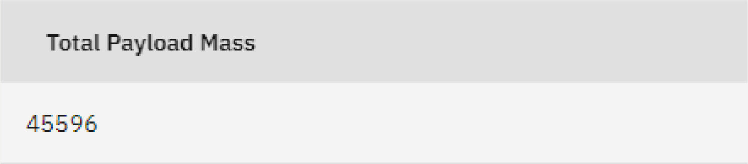
DATE	TIME__UTC_	BOOSTER_VERSION	LAUNCH_SITE
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

SQL Queries Results

Total Payload Mass

3th task: Query that displays 5 records where the launch sites begin with ‘CCA’.

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



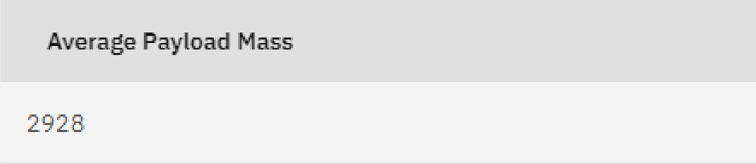
Total Payload Mass
45596

SQL Queries Results

Average Payload Mass by F9 v1.1

4th task: Query that 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';
```



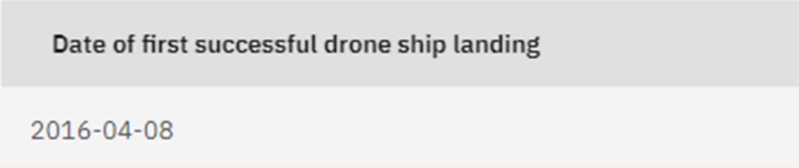
Average Payload Mass
2928

SQL Queries Results

First Successful Ground Landing Date

5th task: Query that lists the date of the first successful landing on the ground pad.

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



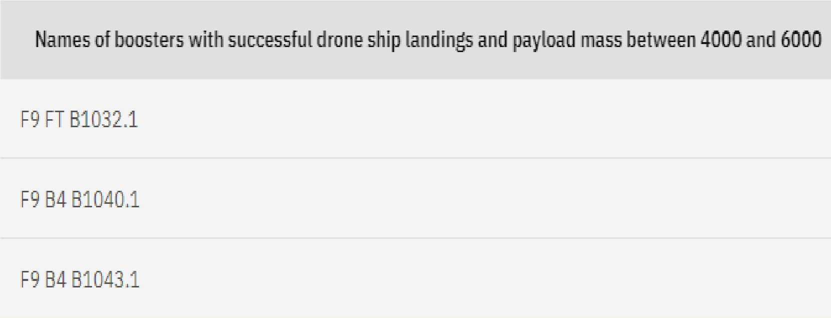
Date of first successful drone ship landing
2016-04-08

SQL Queries Results

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

6th task: Query that 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;
```



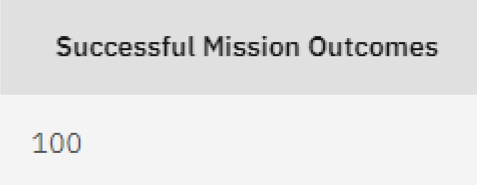
Names of boosters with successful drone ship landings and payload mass between 4000 and 6000
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

SQL Queries Results

Total Number of Successful and Failure Mission Outcomes

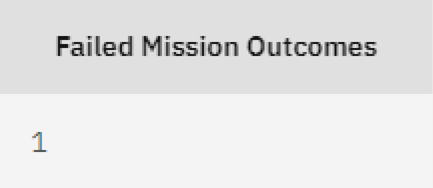
7th task: Query that lists the total number of successful and failed mission outcomes.

```
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%');
```



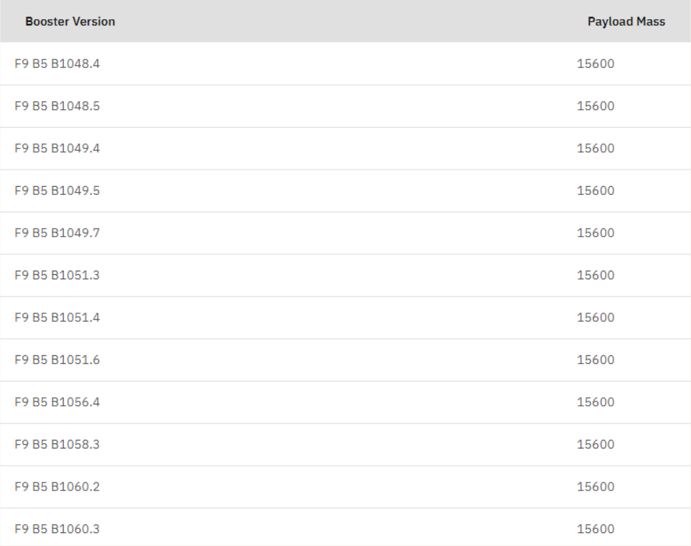
Failed Mission Outcomes
1

SQL Queries Results

Boosters Carried Maximum Payload

8th task: Query that lists 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);
```



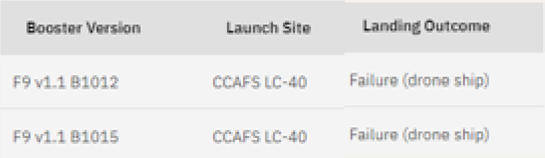
Booster Version	Payload Mass
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

SQL Queries Results

2015 Launch Records

9th task: Query that lists the failed landing outcomes in drone ships, their booster versions, and launch site names for the 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%';
```



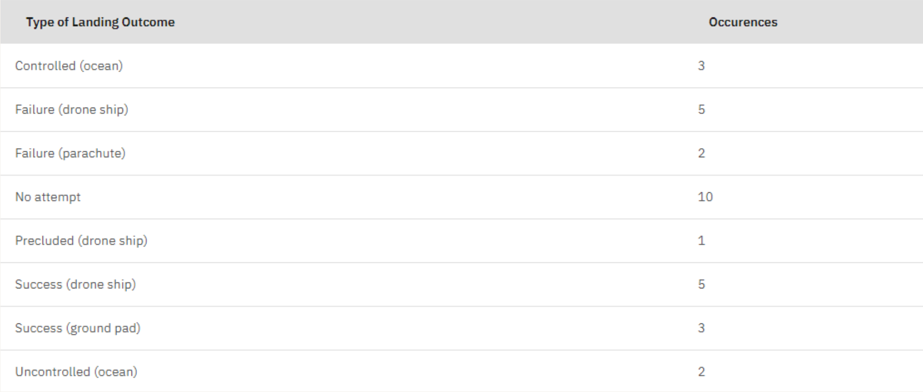
Booster Version	Launch Site	Landing Outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL Queries Results

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

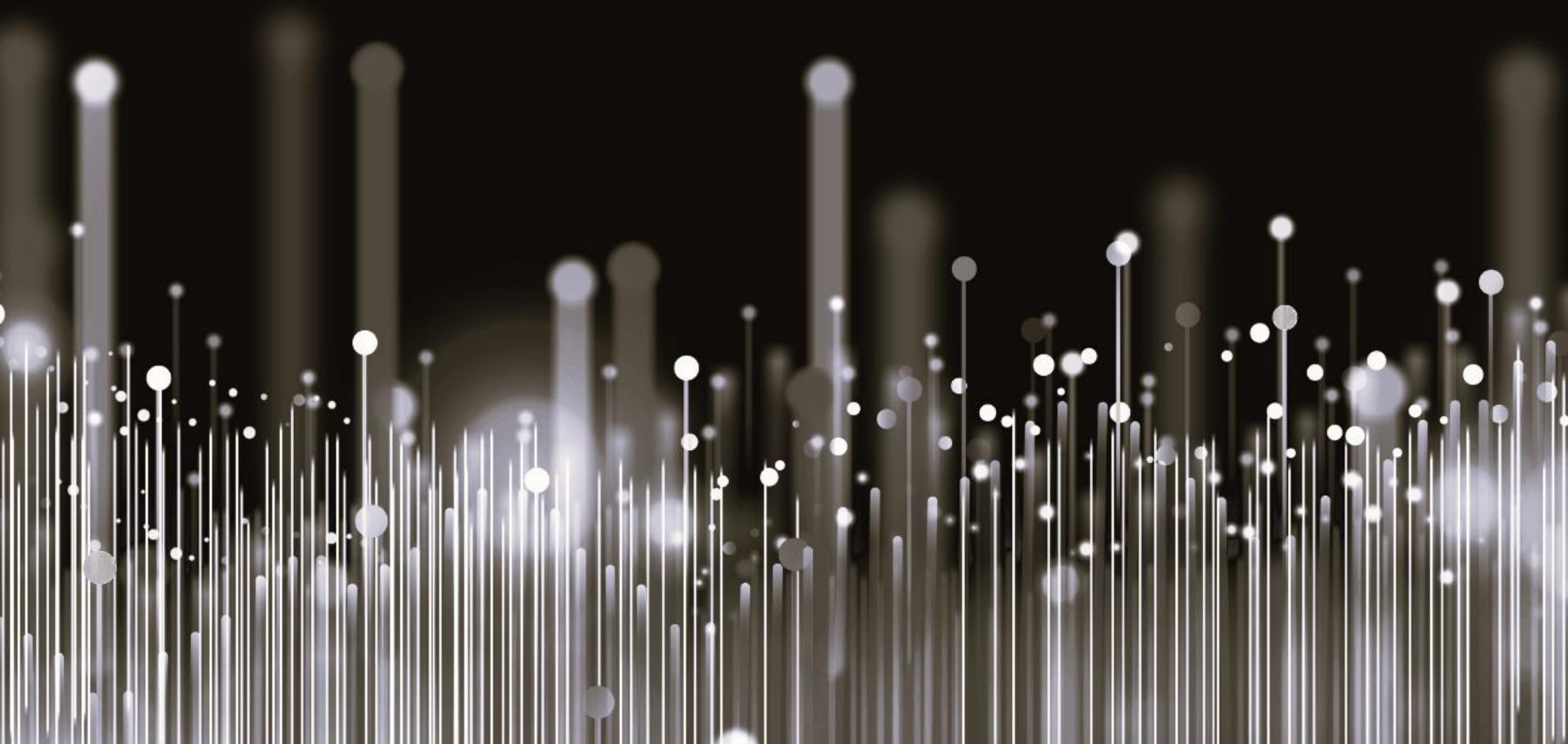
10th task: Query that ranks the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order.

```
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;
```



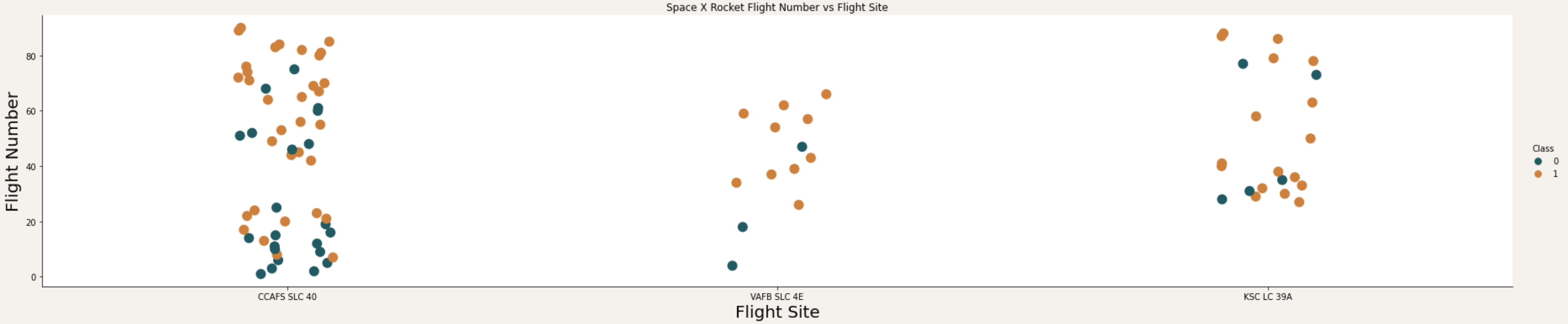
Type of Landing Outcome	Occurences
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1
Success (drone ship)	5
Success (ground pad)	3
Uncontrolled (ocean)	2

Data Visualization Presentation



Data Visualizations

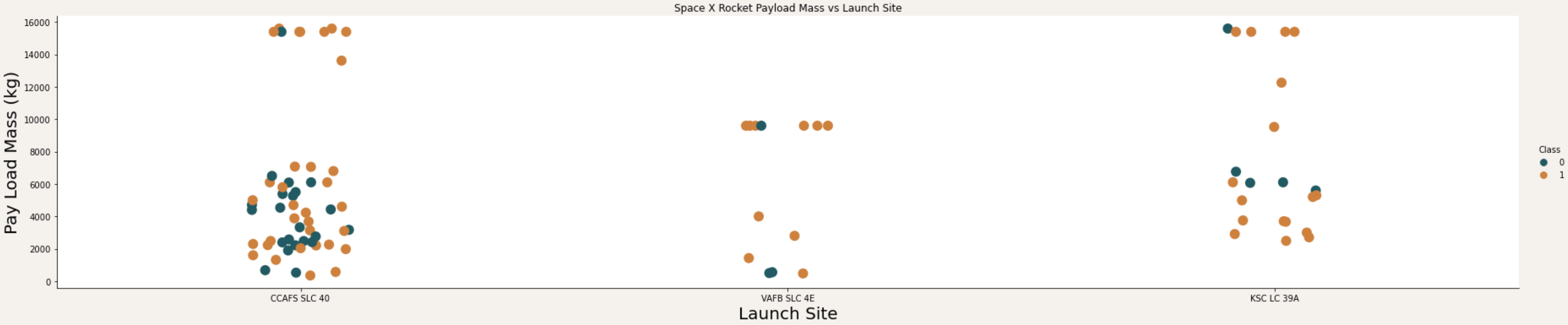
Flight Number vs. Launch Site



The graph shows us that from CCAFS SLC 40 launch site we had significantly more launches, compare to the other two launch sites. Although the other sites had performed fewer launches, they have higher percentage of successful fist stage landings.

Data Visualizations

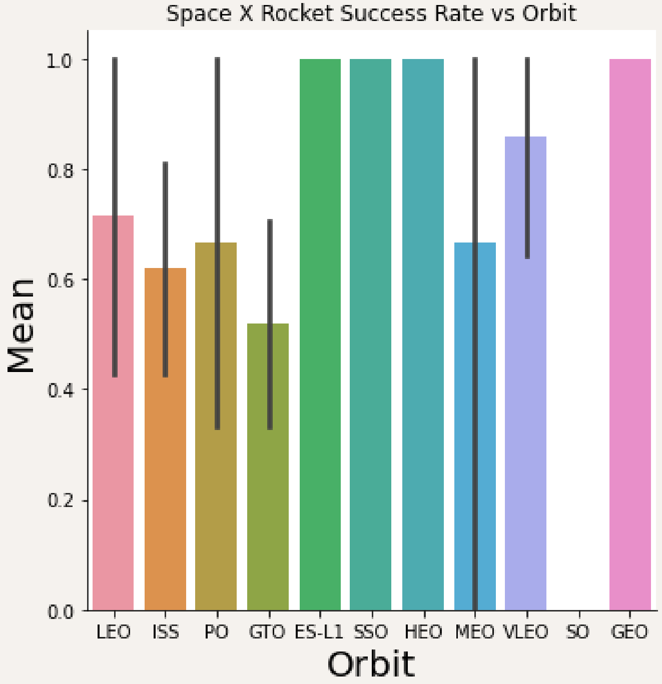
Payload vs. Launch Site



The graph indicates that relatively higher Payload masses tend to have a positive correlation with successful landings. The first launch site is evidently higher due to the number of elements compared to the other two.

Data Visualizations

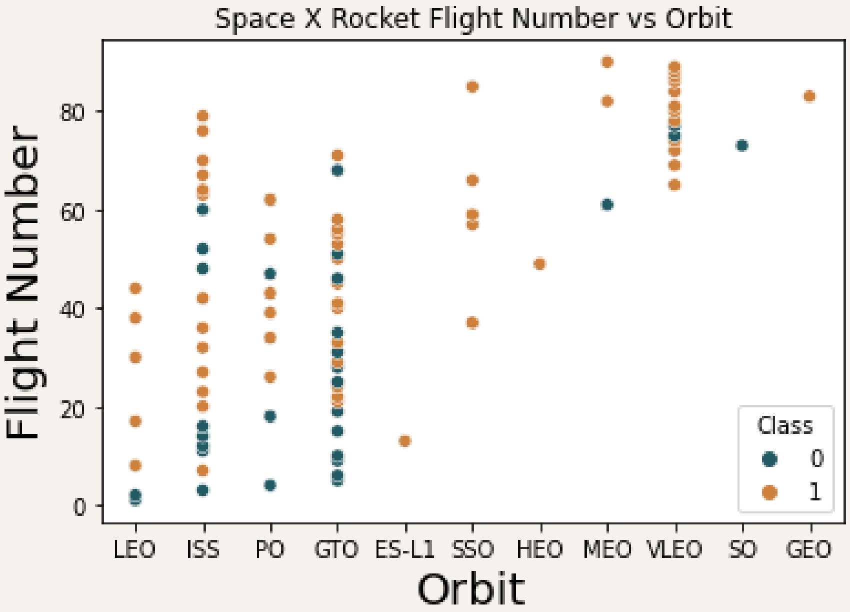
Success Rate vs. Orbit Type



The bar graph illustrates that the success rate is maximized for the case of the ES-L1, SSO, HEO and GEO Orbits, while the exact opposite is true for the SO Orbit. Although the results for the remaining Orbits have a variety of mean values, the bar graph does not fall below 50%.

Data Visualizations

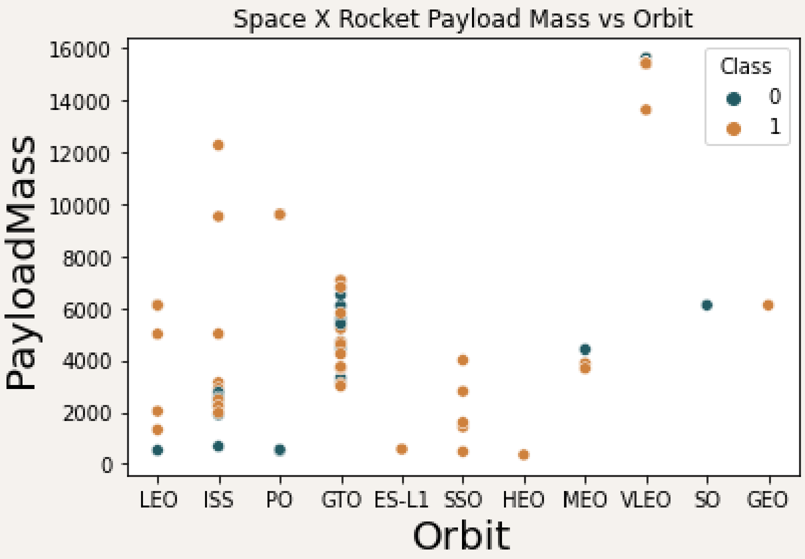
Flight Number vs. Orbit Type



Between the ES-L1, HEO, and GEO Orbits we have only one data point that belongs to each of them. This fact indicates that we have a significantly high uncertainty regarding their efficiency. This fact does not apply to SSO Orbit since it has a 100% success rate for a total of 5 missions.

Data Visualizations

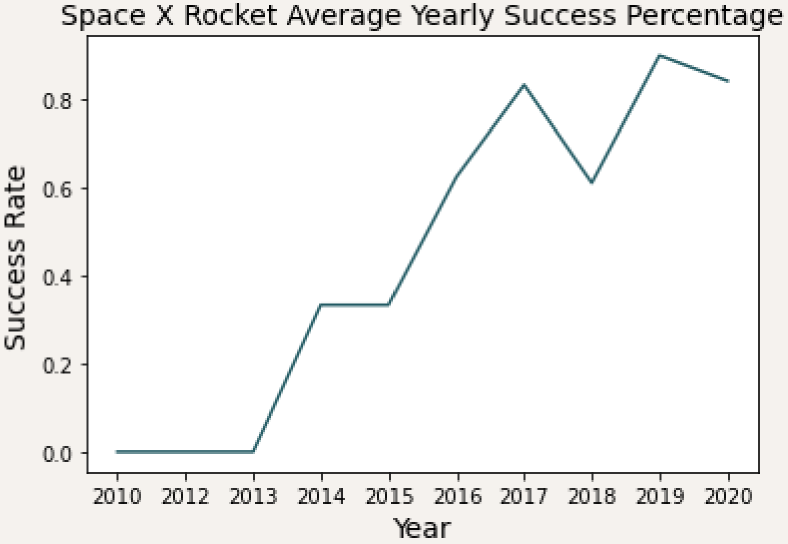
Payload vs. Orbit Type



Previously we showed a positive correlation between higher payload masses and successful landings, a trend which is reflected by this Scatter Plot as well. However, in this plot, we depict that the GTO Orbit seems to score better for lower payload mass values.

Data Visualizations

Launch Success Yearly Trend

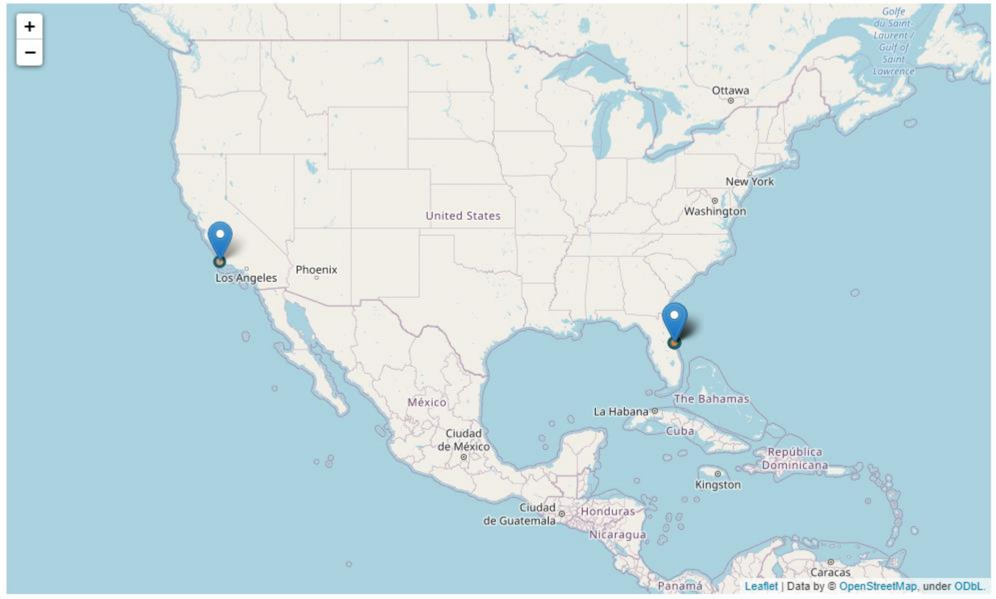


In this graph, we present a general trend of the SpaceX First Stage Landing Success Percentage. Overall there is a clear increase in the average yearly success of the landings, with 2019 holding the best results so far.

Launch Sites Proximities Analysis

Folium Maps

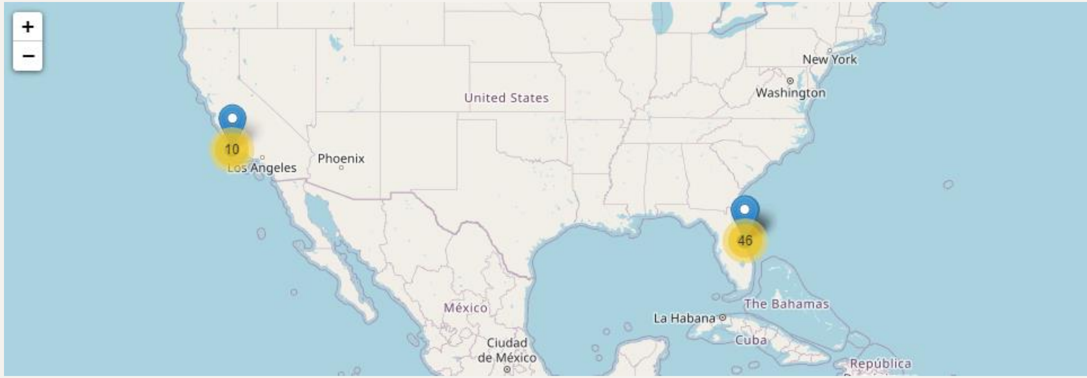
Launch Sites Locations on Folium Maps

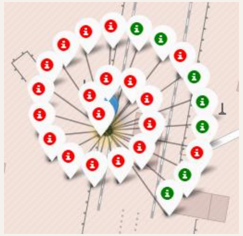


All four launch sites can be seen in coastal areas in USA. One of them (VAFB SLC-4E) is located in California, while the remaining three (KSC LC-39A, CCAFS SLC-40, CCAFS LC-40) are located in Florida

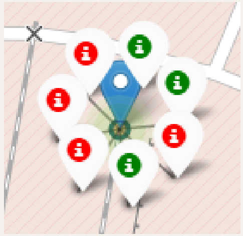
Folium Maps

Launch Sites Clustered and Labelled depending on Success

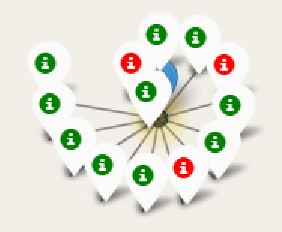




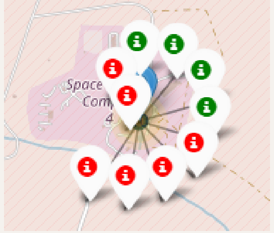
CCAFS LC-40



CCAFS SLC-40



KSC LC-39A



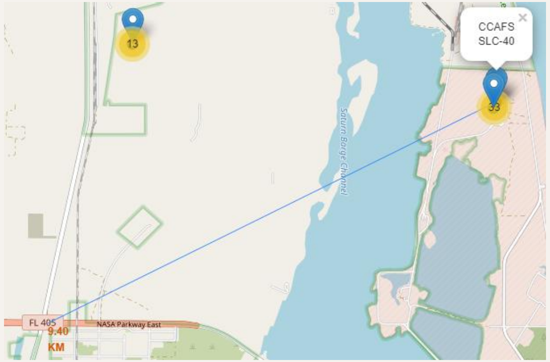
VAFB SLC-4E

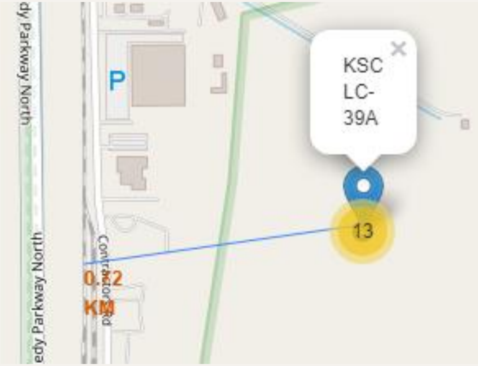
The screenshot of the map depicts how the launch sites have been clustered, depending on location. The following four screenshots depict the number of successful (green) and failed (red) first stage landings. The best score clearly belongs to , with only 3 failures out of 13 total launches.

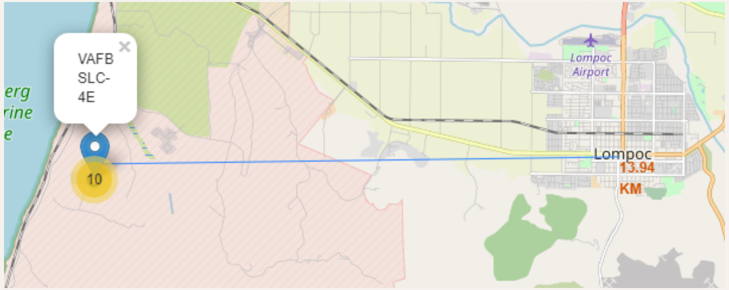
Folium Maps

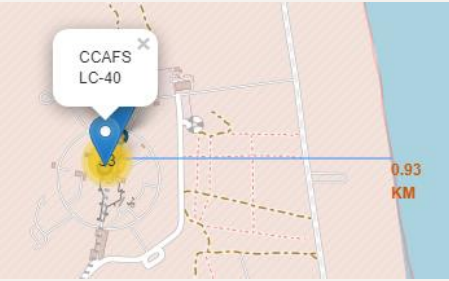
Launch Site Distance from various Locations

The screenshots present the distance of one launch site from a highway, railway, coast and city. Launch sites are as far away from cities and highways as possible. On the other hand, they appear to be located very close to the coast.





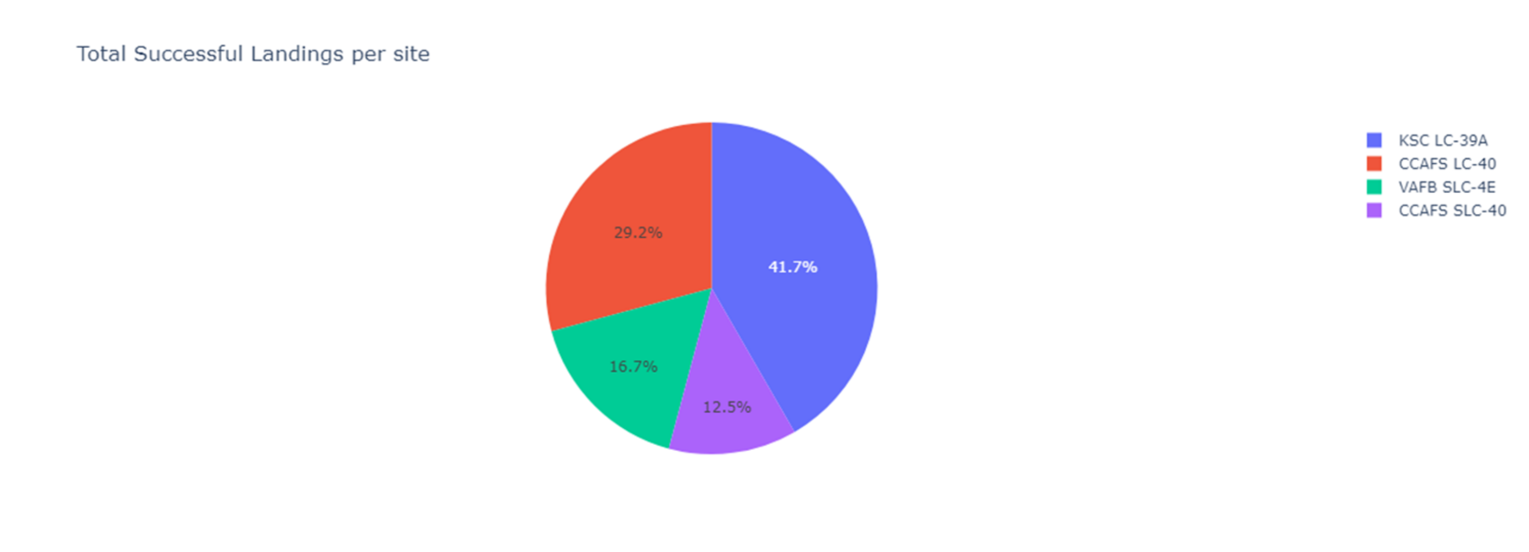




Dashboard with Plotly Dash

Dashboard with Plotly Dash

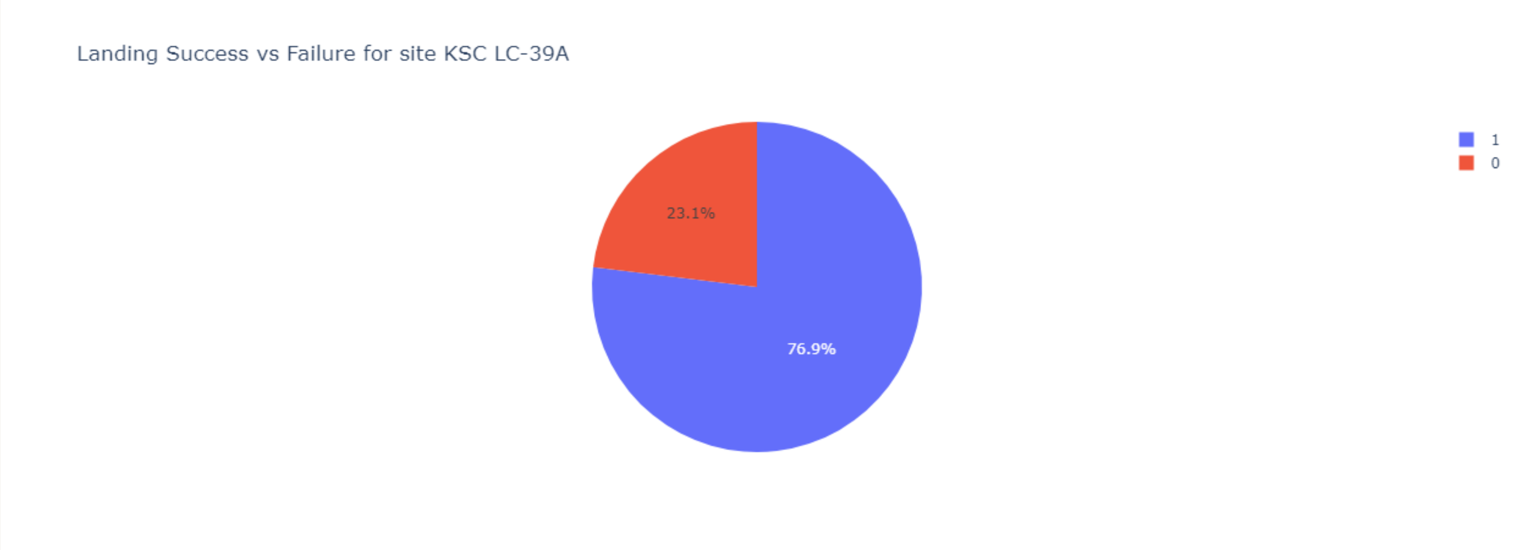
Pie Chart with launch success count for all sites



Although KSC LC-39A is the site with the most successful landings, we need to further study the sites separately in order to determine which one corresponds to the highest launch success ratio.

Dashboard with Plotly Dash

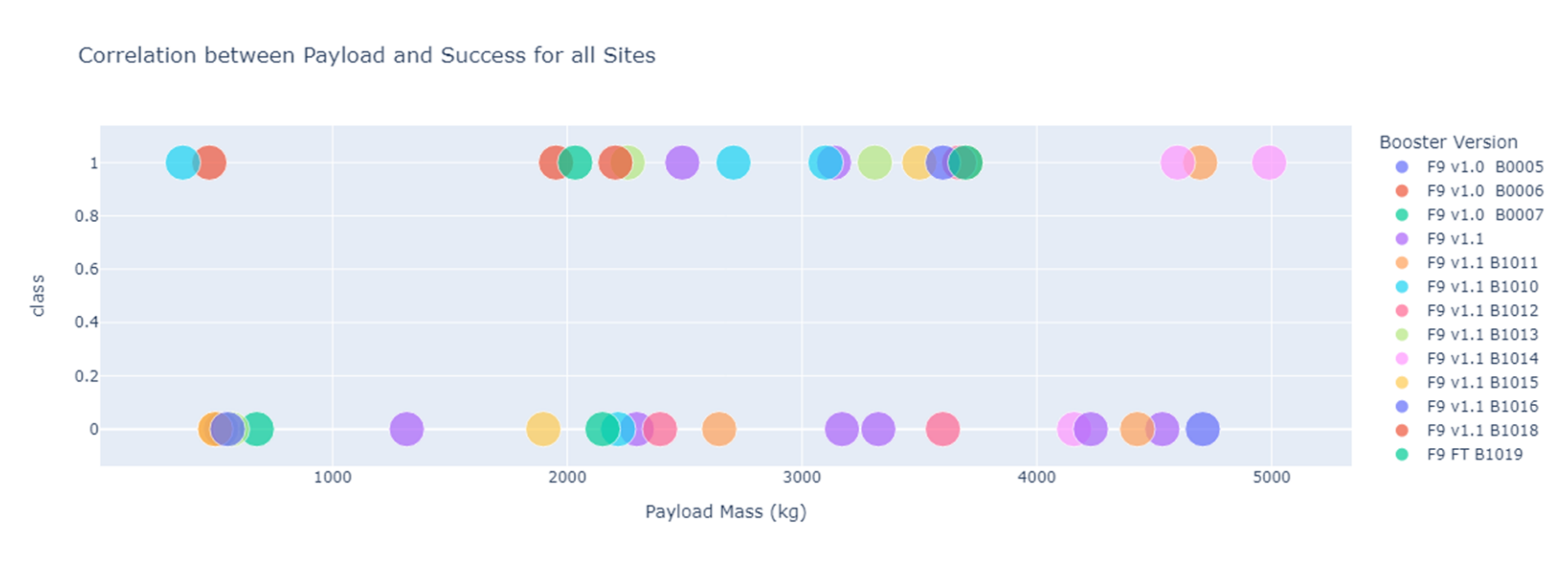
Pie Chart for the launch site with the highest launch success ratio KSC LC-39A



KSC LC-39A is also 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

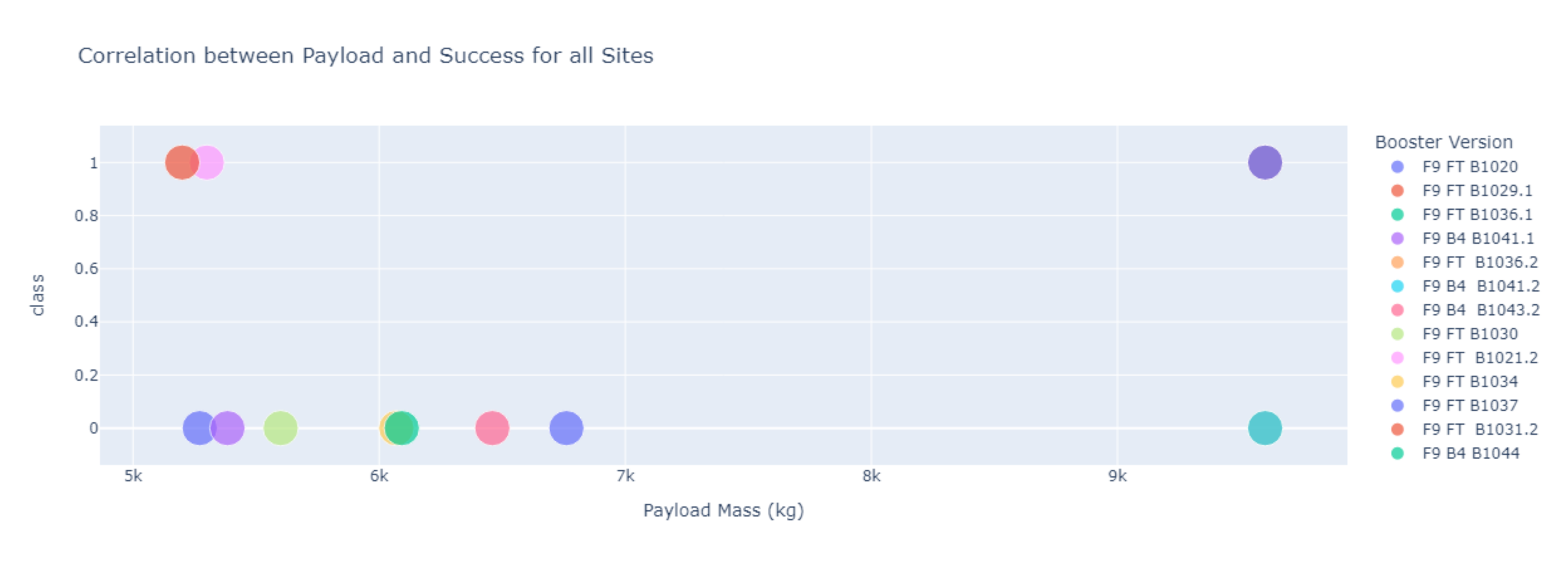
Scatter Plot of Payload vs. Launch Outcome (low payloads)



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)



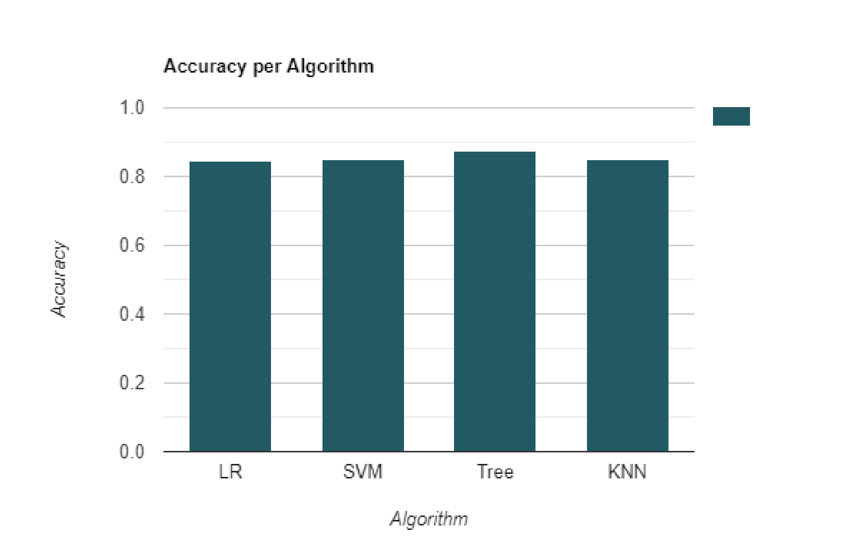
For high payloads (masses between 5000kg and 10000kg) the the landings tend to fail more often, Especially after crossing the 6000kg threshold (with a singular exception).

Predictive Analysis Results

Predictive Analysis Results

Classification Accuracy

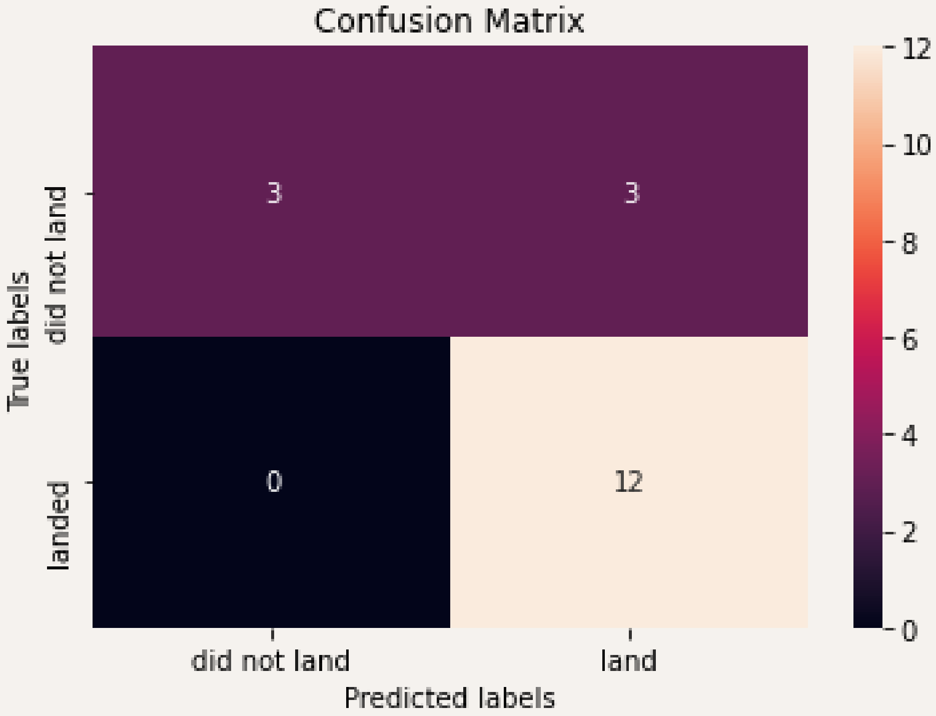
We firstly tuned each model's parameters to their corresponding optimal values and then we plotted the accuracy that each model has achieved. LR stands for Logistic Regression, SVM for Support Vector Machine, Tree for Decision Tree, and KNN for k-Nearest Neighbors. While all algorithms score values close to 0.84, the highest-scoring algorithm is the Decision Tree with an accuracy of ≃ 0.877.



Predictive Analysis Results

Confusion Matrix

Previously we showed that Decision Tree is the highest-scoring algorithm therefore we evaluate it by plotting the Confusion Matrix corresponding to its predictions on the test data. Decision Tree presents a deviation with the False Positive values, it predicts successfully all the true landings.



Conclusions

Conclusions

In this section, we present some of the main conclusions drawn from our analysis.

- The site that was chosen for more launches (CCAFS SLC-40) was not the site that had the highest success rate for first-stage landings (KSC LC-39A).
- 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

In this section, 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.

Thank you!