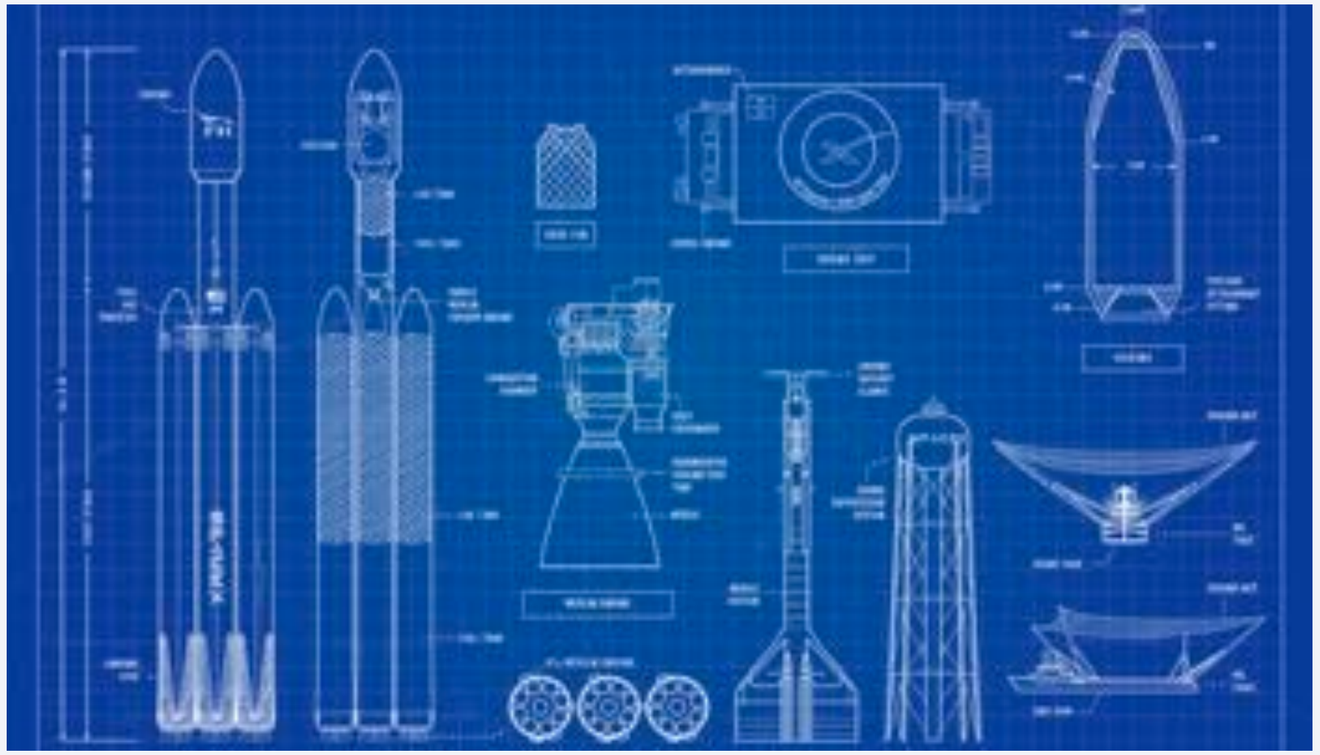
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mario Saraiva
Jan 07, 2023

Outline

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



Source: https://moonspaceway.com/falcon-9-blueprint/

Executive Summary

- The goal of this project is to predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.
- The process behind the prediction models used follow the Data Science methodology including **data collection** (API and Web Scrapping), **data wrangling** (Pandas and Numpy Libraries), **exploratory data analysis** (EDA with SQL), **data visualization** (Seaborn, Plotly/Dash, and Folium), **model development** (GridSearch) **and evaluation** (Accuracy scores and Confusion Matrix).
- The main findings are the following:
    1. Payload mass and launch site are key predictors. These two variables are extensively explored throughout this report.
    2. The best performing model was the **Decision Tree Classifier**;
    3. All machine learning models appear to be **overfitting the testing data due to a small sample size**.

Mario Saraiva - (January, 2022)

Introduction

- The purpose of this project is to determine the cost of a launch by predicting if the Falcon 9 first stage will land successfully.
- SpaceX (advertises) Falcon 9 rocket launches costs approximately 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore **if we can predict if the first stage will land, we can determine the cost of a launch**. This information may be important to investors and competitors considering the cost and future of rocket launches.

Mario Saraiva - (January, 2022)

Section 1

Methodology

Mario Saraiva - (January, 2022)

Methodology

Executive Summary

✓Data collection methodology: SpaceX API GET Request and Web Scrapping (bs4) Wikipedia

✓Perform data wrangling: The data wrangling consisted of two sections: (1) Exploratory Data Analysis; (2) Deciding on Training Labels

✓Perform exploratory data analysis (EDA) using visualization and SQL: Pandas and Numpy Libraries + SQL queries. Seaborn and Ploty data visualizations.

✓Perform interactive visual analytics using Folium and Plotly Dash

✓Perform predictive analysis using classification models: Models used are (1) Logistic Regression, (2) Sector Vector Machine, (3) Decision Tree Classifier, and (4) K-nearest neighbors.

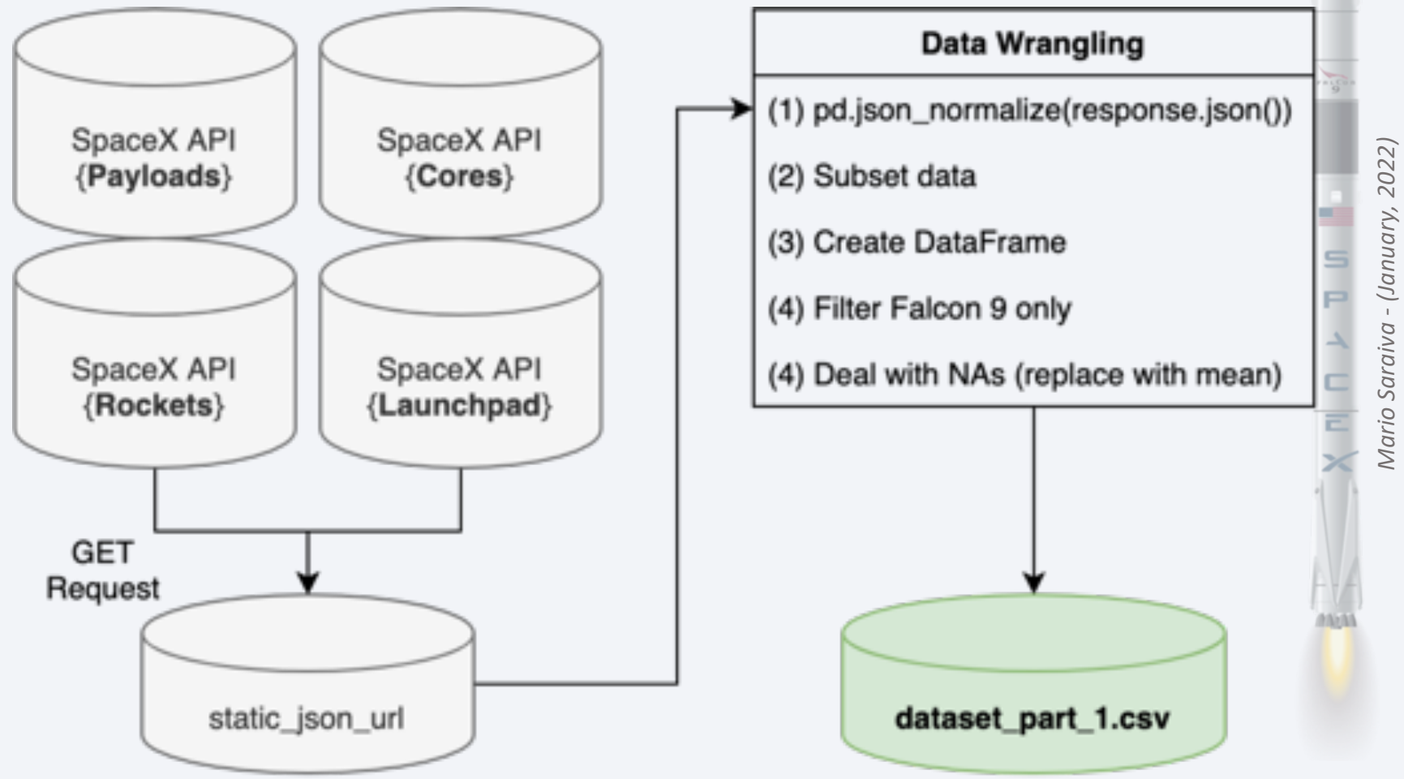
Data Collection

- Data was collected from SpaceX's API and scrapped from the Wikipedia (Falcon 9 and Falcon Heavy launches page).
- **From the API**, our request focused on getting the following data regarding: rockets, launchpad, payload, the outcome of the landing, the type of the landing, number of flights with that core, whether *gridfins* were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core. We limited the data to Falcon 9 launches.
- **From Wikipedia**, we scrapped Falcon 9 historical launch records from a (9th June 2021 page titled 'List of Falcon 9 and Falcon Heavy launches' . A GET Request with *BeaufitulSoup* (bs4) were used to extract the data.
- The final dataset was exported as .csv file.

Mario Saraiva - (January, 2022)

Data Collection – SpaceX API

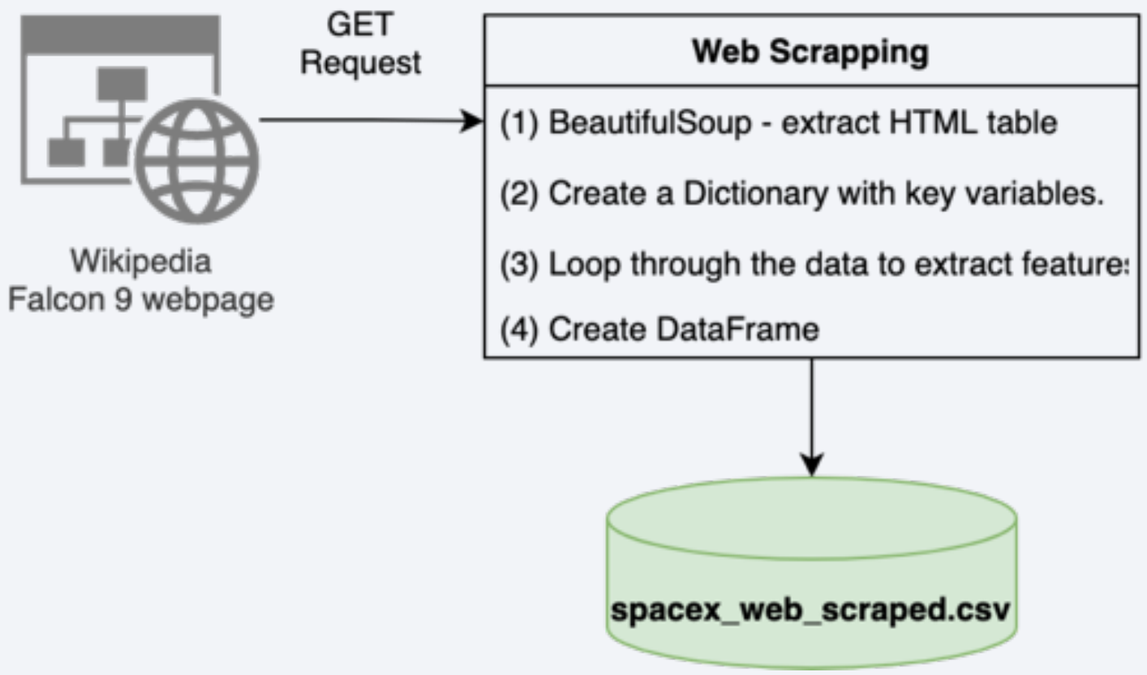
- We used the static Json provided to request the SpaceX data. Then we processed the Json, extracted desired features and created a DataFrame object.



[Click here for SpaceX API calls notebook on GitHub]

Data Collection - Scraping

- From Wikipedia, we scrapped Falcon 9 historical launch records from a (9th June 2021) page titled 'List of Falcon 9 and Falcon Heavy launches' .
- A GET Request with BeaufitulSoup (bs4) was used to parse the data.

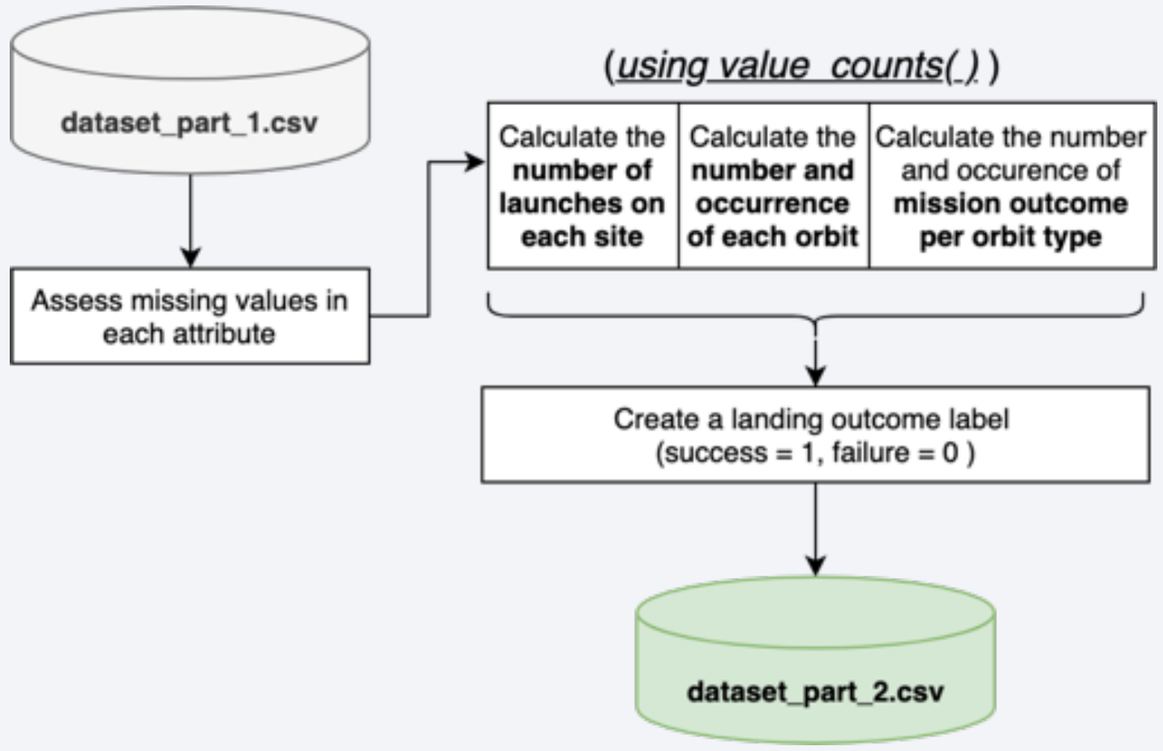


Click here for Web Scrapping notebook on GitHub

Data Wrangling

- Only the Pandas and Numpy libraries were used in the data wrangling script.
- Exploratory Data Analysis
  - 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.
- Deciding on Training Labels
  - Create a landing outcome label;
  - Success = 1 and Failure = 0.

Click here for data wrangling related notebooks



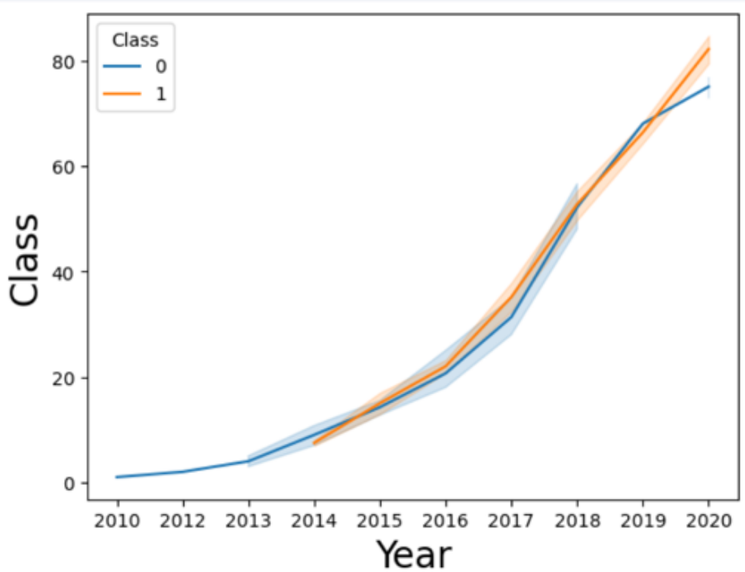
EDA with Data Visualization

Plot Summary: We used plots to understand the relationship between the variables and the success rate. This understanding is key to predicting the outcome.

- Plot 1: The Flight Number and Pay load Mass and overlay the outcome of the launch.
- Plot 2: Flight Number and Launch Site.
- Plot 3: Payload and Launch Site.
- Plot 4: Success rate per Orbit.
- Plot 5: Flight Number and Orbit.
- Plot 6: Launch success yearly trend.¶

Click here for EDA with data visualization notebook

Average Launch success yearly trend



EDA with SQL

- In total 10 queries were performed as part of the EDA seeking to understand:
  - **The Context**: The first four queries were to understand the distinct launch sites, maximum and average payload mass used.
  - **The Timeline**: A query was used to get the date of the first successful landing.
  - **The Payload mass**: A query was used to explore the relationship between successful drone ship landings with payload mass between 4.000 – 6.000 Kg. A query investigated the relationship between booster version and max payload mass.
  - **The Success rate**: A query was used to summarize total success and failure outcomes. Lastly, we ranked the count of landing outcomes (such as Failure or Success) between 2010-06-04 and 2017-03-20.

Click here for EDA with EDA with SQL notebook

Build an Interactive Map with Folium

- In our Folium map we:
  - Marked all launch sites on a map;
  - Marked the success/failed launches for each site on the map;
  - Calculated the distances between a launch site and the coastline

- The map is an important visualization to help assess geographical patterns between sites that affect the outcome of landings.

Click here for EDA with Folium Maps notebook

Build a Dashboard with Plotly Dash

- The Dashboard comprises of two visualizations – (1) a Pie Chart with the success ratio according to the site and (2) a scatter plot displaying the relationship between Payload mass, landing outcome and Booster version used.
- The goal was to identify which sites had the most (and least) successful landings and to further explore and quantify the ideal payload mass associated with successful outcomes.

Click here for the Dashboard script

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

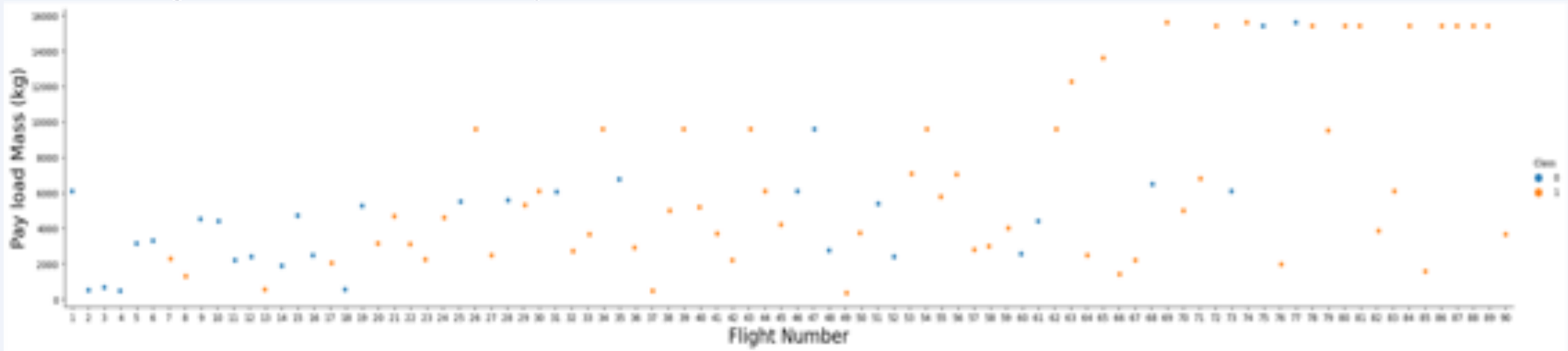
Click here for the Machine Learning notebook

Section 2

Insights drawn from EDA

Flight Number vs. Payload Mass

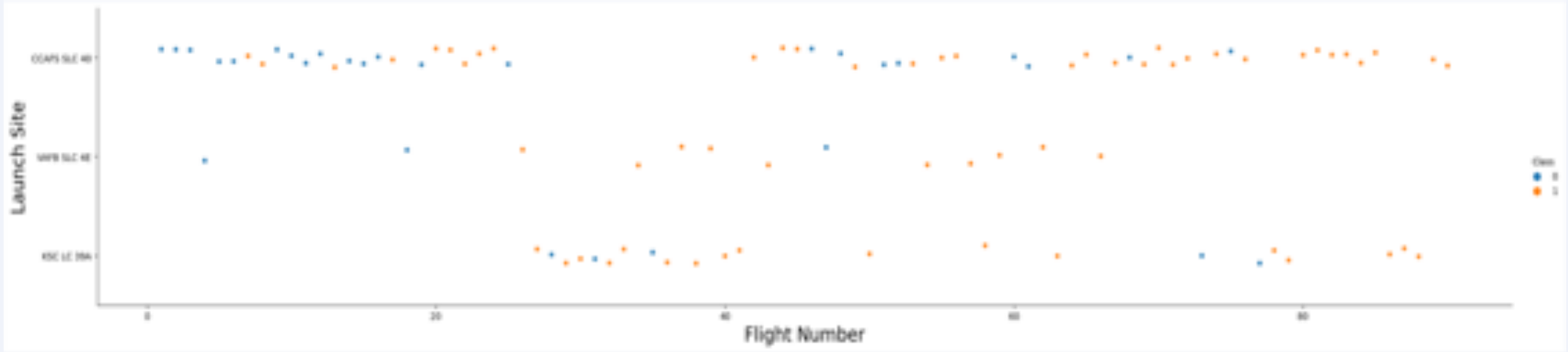
- Plot 1: ***Flight Number*** and ***Payload Mass*** and overlay the outcome of the launch.
  - We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return. **Both variables seem crucial to predict if the first stage will land successfully**.



Click here for EDA with data visualization notebook

Flight Number vs. Launch Site

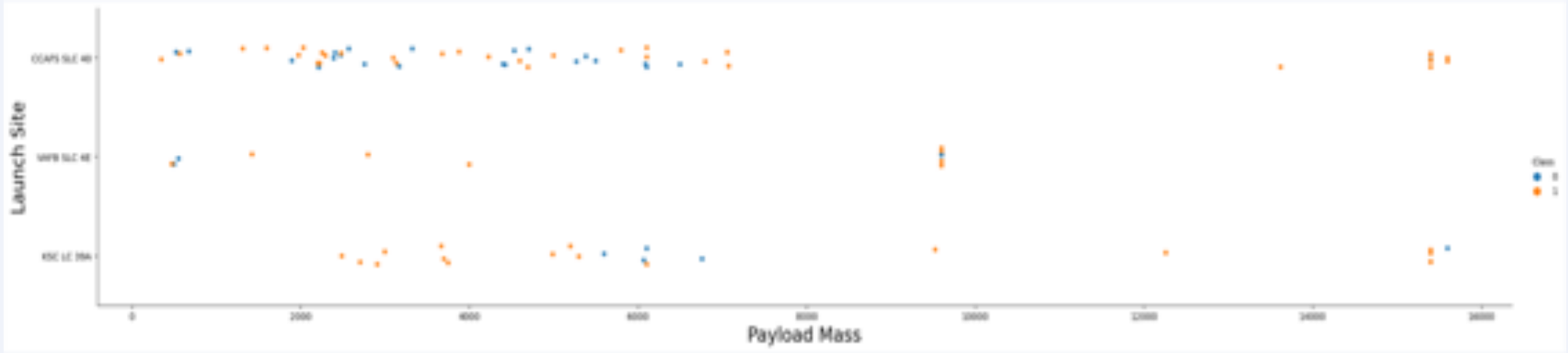
- Plot 2: ***Flight Number*** and ***Launch Site*** and overlay the outcome of the launch.
  - It is clear that as the number of flights increased the success rate also improved in all sites.
  - Some launch sites had a greater number of success cases than others. '*CCAFS LC-40*' site has a success rate of 60 %, while '*KSC LC-39A*' and '*VAFB SLC 4E*' of nearly 77%.



Click here for EDA with data visualization notebook

Payload vs. Launch Site

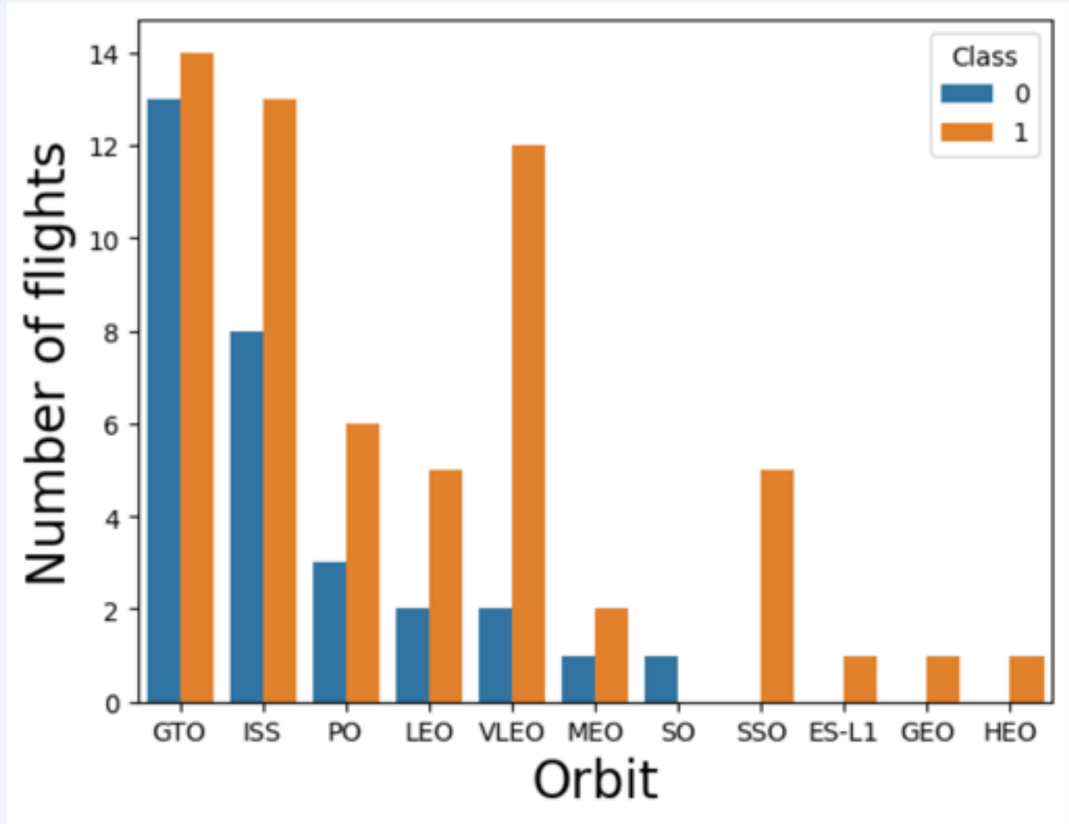
- Plot 3: ***Payload*** and ***Launch Site*** and overlay the outcome of the launch.
  - It appears that a greater Payload mass is associated with a successful first stage landing. Note that we have only 2 failures when the Payload Mass is greater than 9000 Kg for all sites.



[Click here for EDA with data visualization notebook]

Success Rate vs. Orbit Type

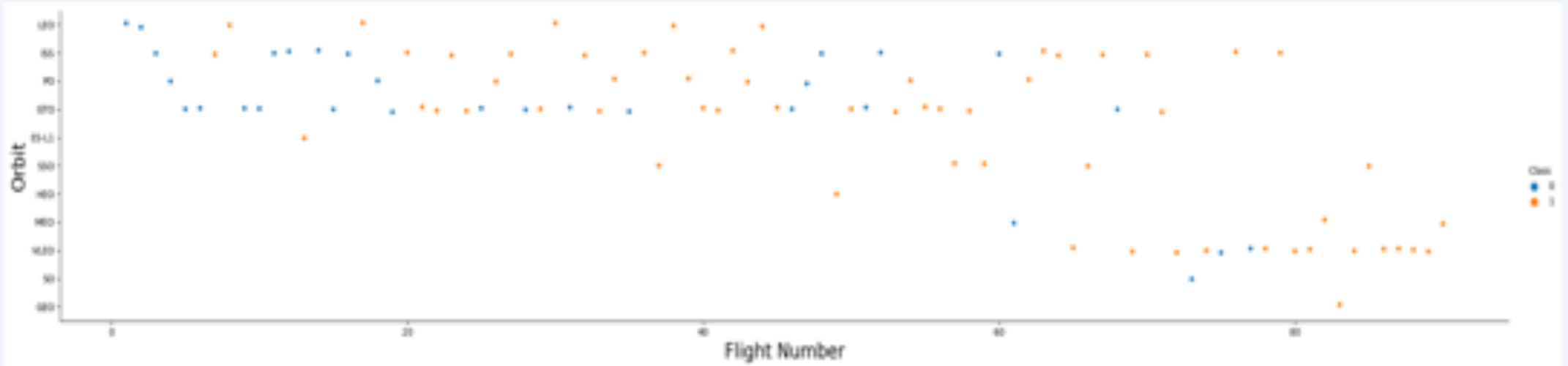
- Plot 4: **Success rate per Orbit type**.
  - Five Orbits – VLEO, SSO, ES-L1, GEO, HEO had a success rate of 80% or more.
  - Out of the five, four – SSO, ES-L1, GEO, HEO, had 100% success rate.



Click here for EDA with data visualization notebook

Flight Number vs. Orbit Type

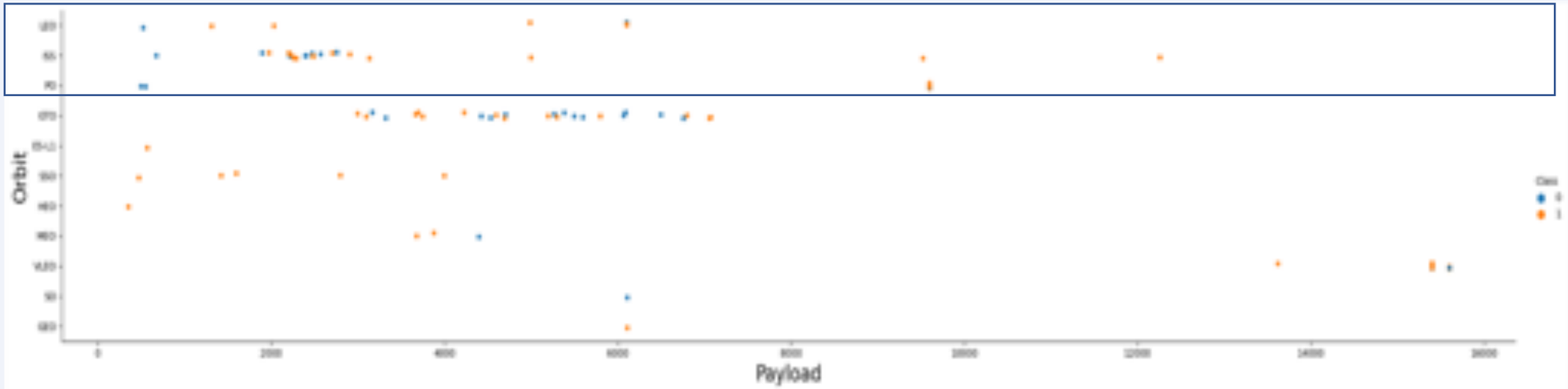
- Plot 5: ***Flight number*** and ***Orbit type*** and overlay the outcome of the launch.
  - It appears that the VLEO orbit success is related to the number of flights;
  - There seems to be no relationship between flight number when in GTO orbit. No clear pattern was observed for the other orbits.



Click here for EDA with data visualization notebook

Payload vs. Orbit Type

- Plot 6: ***Payload*** and ***Orbit type*** and overlay the outcome of the launch.
  - Greater payloads are associated with the successful landings for Polar, VLEO and ISS orbits. But this relationship does not appear to be strong enough (or valid) for other orbits.

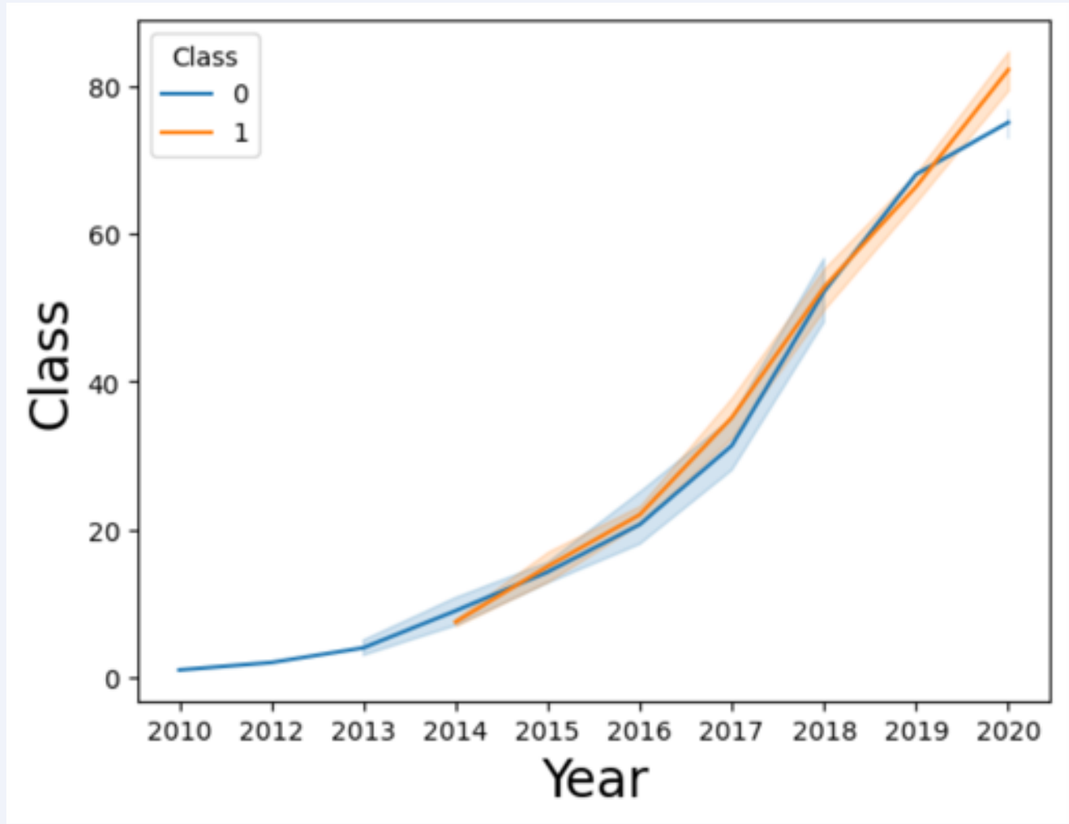


Click here for EDA with data visualization notebook

Launch Success Yearly Trend

- Plot 6: **Launch (average) success yearly trend**.
  - The success rate increased between 2013 and 2020.
  - Between 2014 and 2018, on average, there were more successes than failures. However, for a brief period between 2018 – 2019 that trend was reversed.
  - Lastly, the slope of the success (orange) curve is steeper than the failure curve (blue). Perhaps the technology is maturing.

Click here for EDA with data visualization notebook



All Launch Site Names

- There were four launch sites:
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- The picture shows the query for unique launch sites in the Spacex table.

```
%%sql

SELECT DISTINCT LAUNCH_SITE FROM Spacex
```

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

Mario Saraiva - (January, 2022)

Click here for EDA with SQL notebook

Launch Site Names Begin with 'CCA'

- There are only two sites that begin with ‘CCA’ - namely ‘*CCAFS LC-40*’ and ‘*CCAFS SLC-40*’.

- The query shown selects the first five launch sites that begin with the letters ‘CCA’ from the Spacex table. Note we used ‘%’ in our query to make sure we capture all occurrences that matched our condition.

```
%%sql
SELECT LAUNCH_SITE FROM Spacex WHERE LAUNCH_SITE LIKE '%CCA%' LIMIT(5)
```

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Click here for EDA with SQL notebook

Total Payload Mass

- The total payload mass carried by boosters from NASA was 45.596 Kg.
- The query sums the payload mass for each flight from the Spacex table for only customers that matched the condition of *'NASA (CRS)'*.

```
%%sql

SELECT SUM(payload_mass__kg_) FROM Spacex WHERE customer = 'NASA (CRS)'
```

1
45596

Click here for EDA with SQL notebook

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was 2.928 Kg.
- The query calculates the average payload mass from all launches from the Spacex table for the booster version ‘F9 v1.1’.

```
%%sql

SELECT AVG(payload_mass__kg_) FROM Spacex WHERE booster_version = 'F9 v1.1'
```

1
2928

Click here for EDA with SQL notebook

First Successful Ground Landing Date

- The first successful landing outcome on ground pad happened on December 22, 2015.
- The query selects the minimum date (oldest) from the Spacex table where a success landing on a ground pad occurred.

```
%%sql

SELECT min(DATE) FROM Spacex where landing__outcome = 'Success (ground pad)'
```

1
2015-12-22

Mario Saraiva - (January, 2022)

Click here for EDA with SQL notebook

Successful Drone Ship Landing with Payload between 4000 and 6000

- The boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 were:
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2
- The query selects the booster version from the Spacex table given three conditions: (1) payload mass > 4000, (2) payload mass < 6000, and (3) Success landing on drone ship.

```
%%sql

SELECT booster_version from Spacex
    where payload_mass__kg_ > 4000 AND
            payload_mass__kg_ < 6000  AND
            landing__outcome = 'Success (drone ship)'
```

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

Mario Saraiva - (January, 2022)

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were the following:
  - One failure (in flight) and 100 Success (including one case with unclear payload status).
- The query selects and counts mission outcome records from the Spacex table and it group the results by mission outcomes.

```
%%sql

SELECT mission_outcome, COUNT(*) from Spacex GROUP BY mission_outcome
```

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

Mario Saraiva - (January, 2022)

Click here for EDA with SQL notebook

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are presented in the image herein.
- his query contains a subquery in which we select a booster version from the Spacex table that used the max payload mass (where the max was calculated by the subquery).

[Click here for EDA with SQL notebook]

```
%%sql

SElECT booster_version from Spacex
where payload_mass__kg_ = (SELECT MAX(payload_mass__kg_)
                           FROM Spacex)
```

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

Mario Saraiva - (January, 2022)

2015 Launch Records

- The failed landing outcomes in drone ship, as well as, their booster versions, and launch site names for in year 2015 are given in the picture herein.

```
%%sql

SELECT booster_version, launch_site , landing__outcome
FROM Spacex
    WHERE landing__outcome = 'Failure (drone ship)'
    AND YEAR(DATE) > 2015
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1017	VAFB SLC-4E	Failure (drone ship)
F9 FT B1020	CCAFS LC-40	Failure (drone ship)
F9 FT B1024	CCAFS LC-40	Failure (drone ship)

- The query selects three variables (booster version, launch site and landing outcome) from the Spacex table where two conditions are met – (1) the landing outcome was a Failure (drone ship) and (2) it happened after 2015. Note we applied the YEAR function on the date variable.

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

- The ranking of landing outcomes or Success between 2010-06-04 and 2017-03-20 are presented in the image herein.
- The query counts the landing outcomes from the Spacex table given our date restrictions. The results are present by grouping the landing outcome and in descending order.

Click here for EDA with SQL notebook

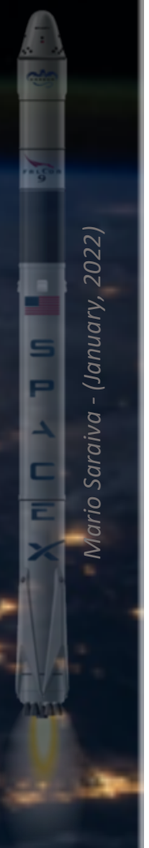
```
%%sql

SELECT  COUNT(*), landing__outcome from Spacex
    WHERE DATE > '2010-06-04' AND DATE < '2017-03-20'
    GROUP BY landing__outcome ORDER BY COUNT(*) DESC
```

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

Section 3

Launch Sites Proximities Analysis

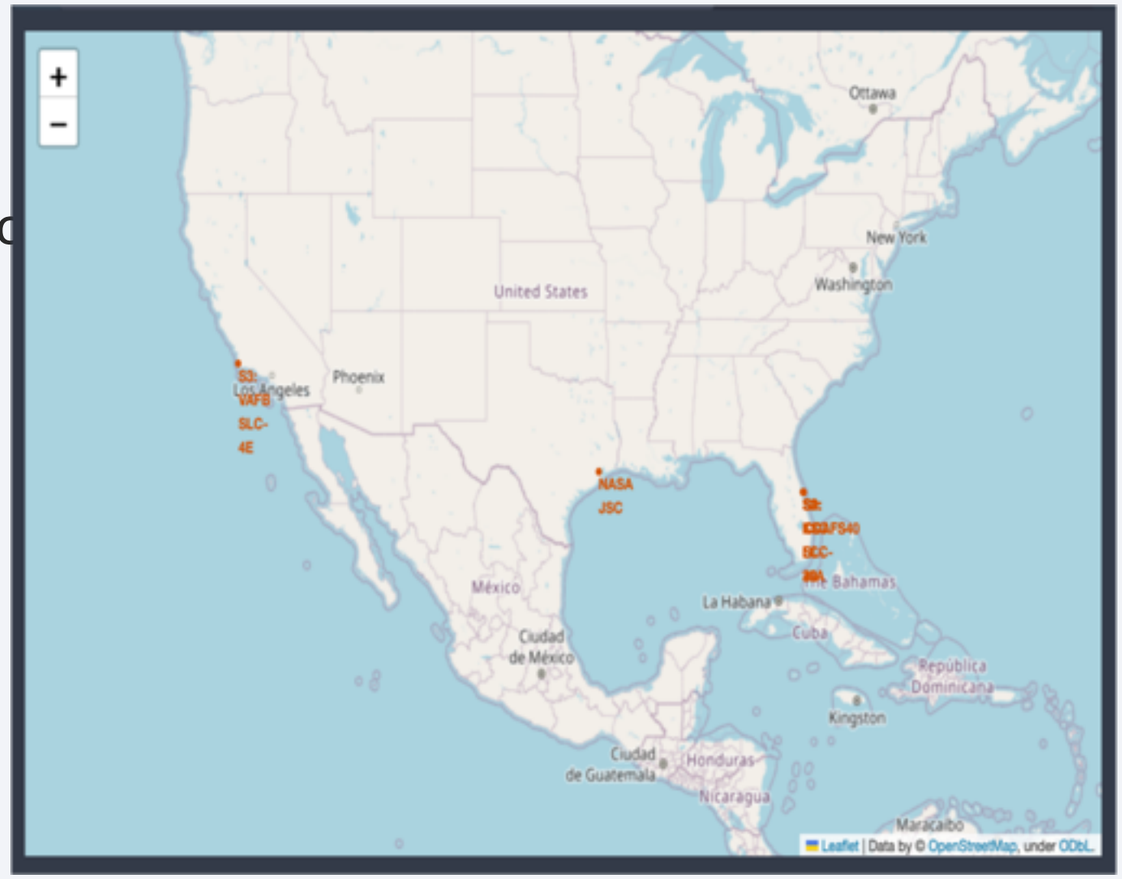


Mario Saraiva - (January, 2022)

Using Folium to map all Launch Sites

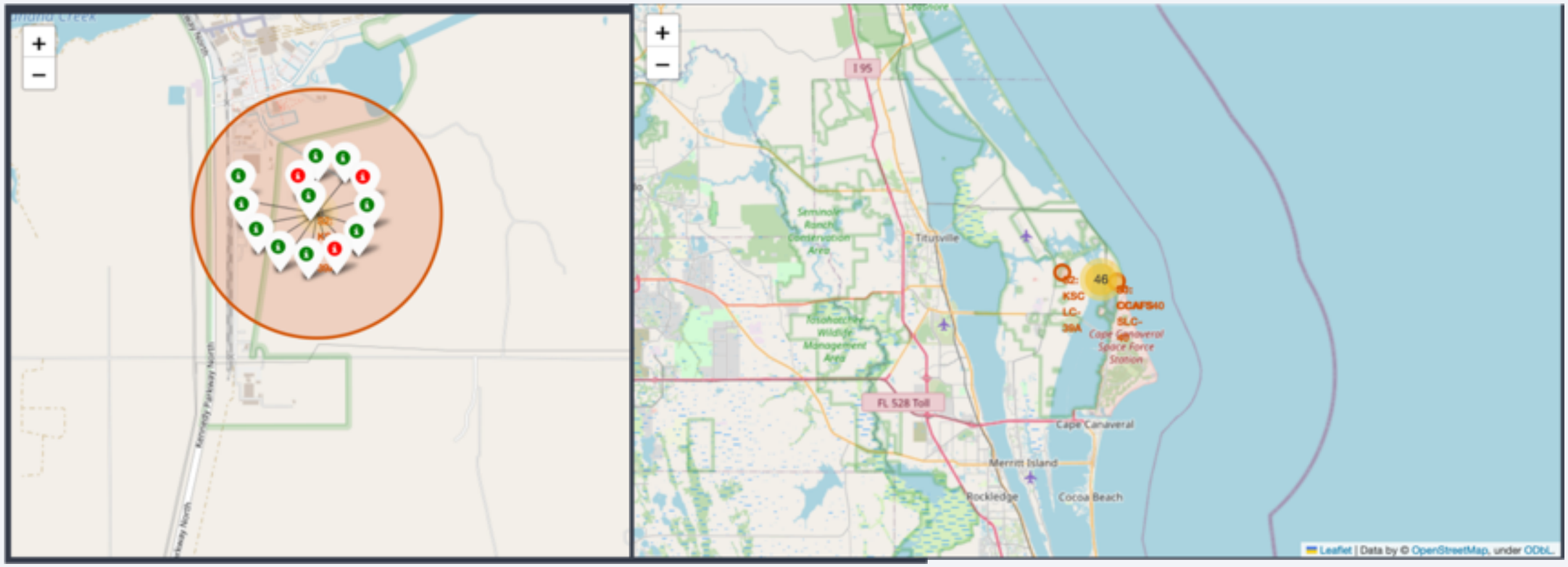
- The map shows that some sites are closer to the Tropic of Cancer than others.
- Nevertheless, all sites are relatively close to the Tropic of Cancer.
- In regards to predictions, it is possible that site location can increase the odds of a successful landing.

[Click here for EDA with Folium Maps notebook]



Mario Saraiva - (January, 2022)

Using Folium to map launch outcomes



- Once again we see that most launches at the KSC LC-39A site were successful.
- KSC LC-39A is more in-land than the other two (closest) sites (CCAFS40 and CCAFS SLC-40).

Using Folium to calculate distance from the coastline

- The picture confirms our previous observation that the CCAFS40 and CCAFS SLC-40 sites are significantly closer to the coastline than KSC LC-39A .

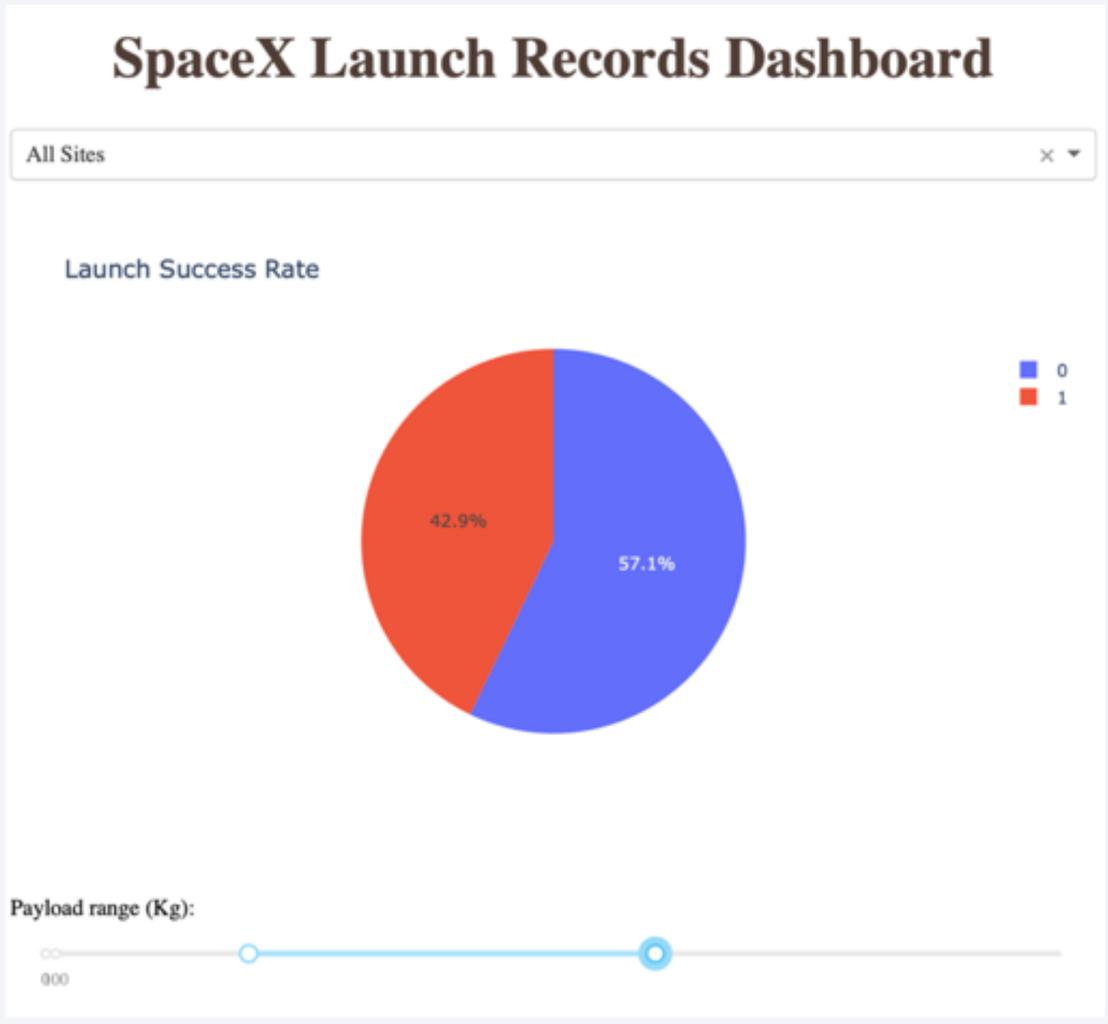
Mario Saraiva - (January, 2022)

Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Records Dashboard – All sites success Rate

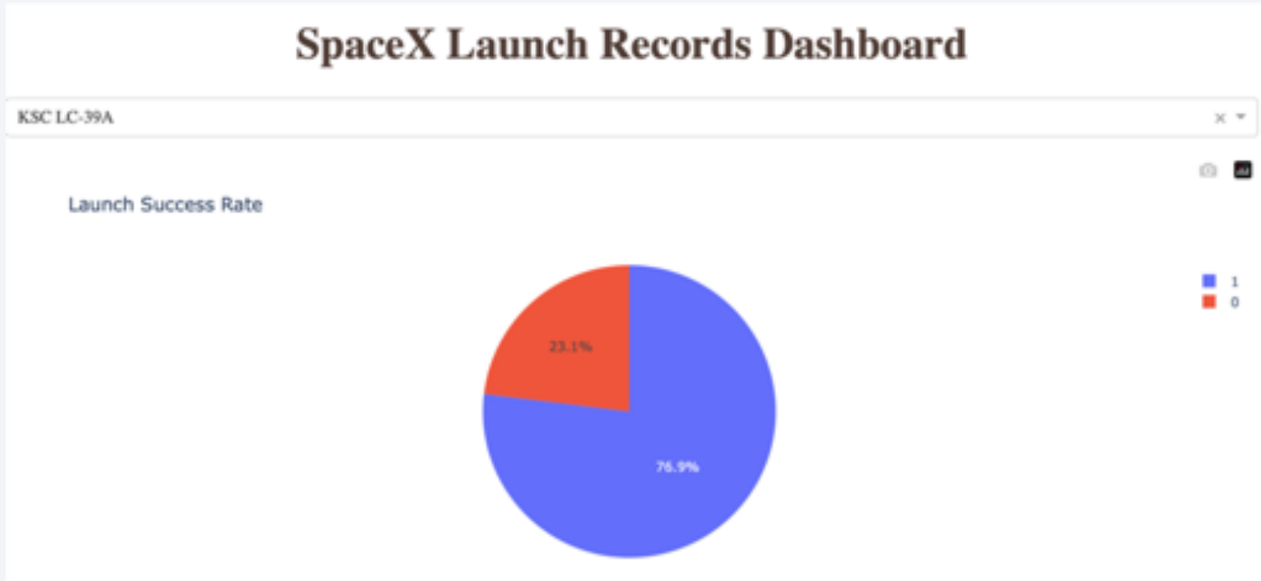
- When considering all sites, the launch success rate was only 42.9%.



Click here for the Dashboard script

SpaceX Launch Records Dashboard – Highest success

- The launch site with highest launch success ratio was KSC LC-39A with a success ratio of nearly 77%.
  - The second highest rate was at CCAFS LC-40 site, where the ratio was 73.1%.
- The launch site is likely a strong predictor of the landing outcome.



Click here for the Dashboard script

SpaceX Launch Records Dashboard – Payload and Success

- We note that most success cases happened with a payload mass between 1900 Kg and 5300 Kg. There are two outliers with payload mass below 500 Kgs and one outlier with over 9000 Kgs.
- When we zoom we see that many booster versions achieved success within this range.

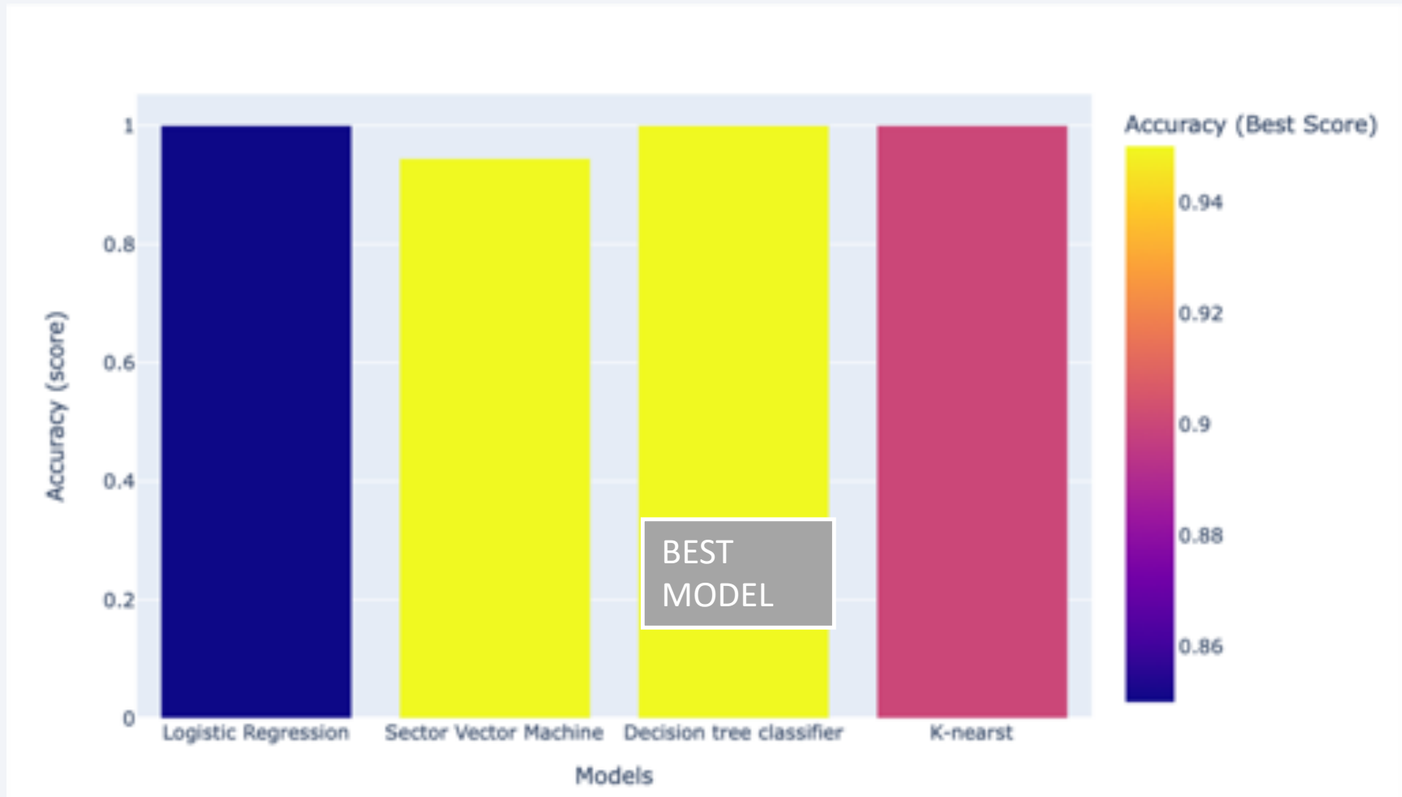
Section 5

Predictive Analysis (Classification)

Mario Saraiva - (January, 2022)

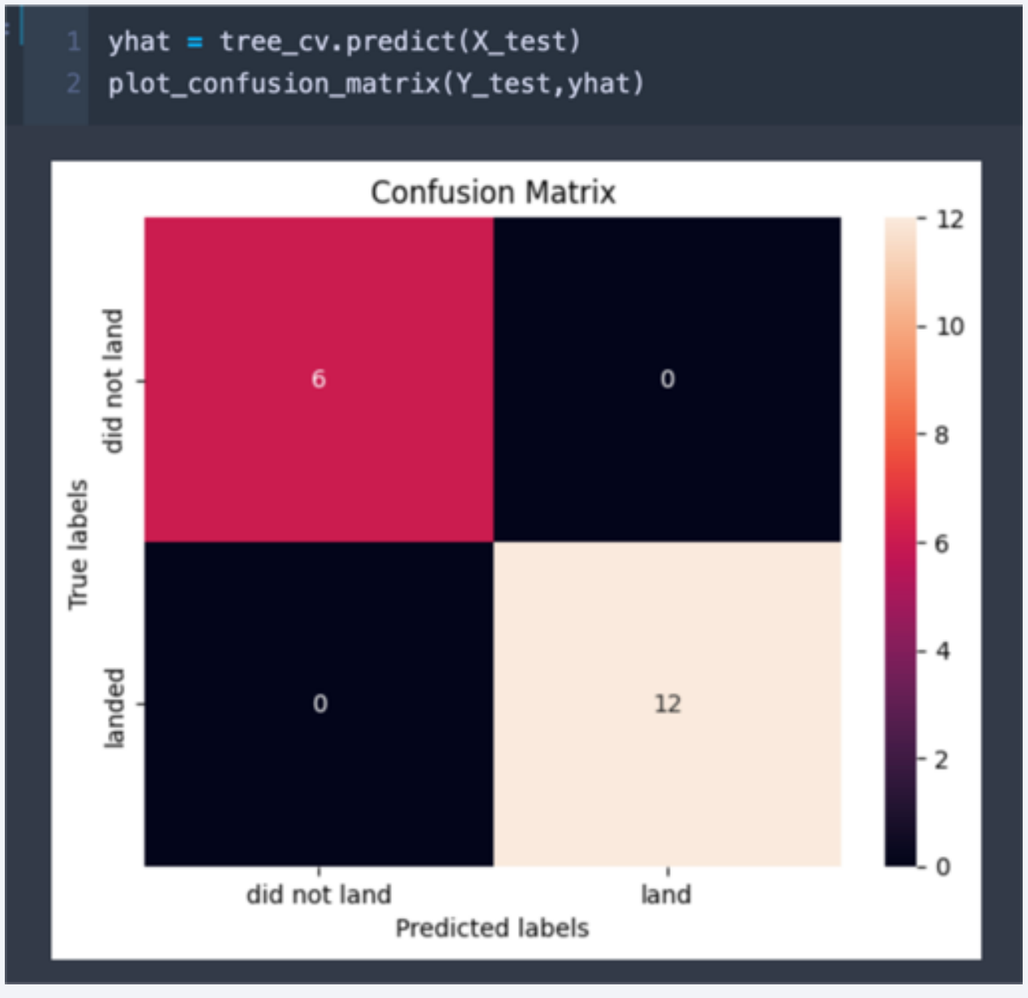
Classification Accuracy

- All models performed well in the test data. The results were similar when using *[MODEL].score(X_test, Y_Test),* as seen in the Y-axis.
- Thus, we decided to also use the Best Score accuracy using the [MODEL].best_score_ function to compare models.
- The model with the highest Accuracy score and best score is the **Decision Tree Classifier**.



Confusion Matrix

- As seen in the confusion matrix of our best performing model – the **Decision Tree Classifier** – the rate of true positives is 100%. The model classified correctly all instances.
- ***Caution***: Although our Tree model performed well on the test data, it is important to note that our testing sample was small. The results seen are likely a result of overfitting the model.



Conclusions

- The purpose of this project is to determine the cost of a launch by predicting if the Falcon 9 first stage will land successfully.
- As seen throughout this report, **payload mass and launch site** are very important factors associated with the landing outcome. Other variables such as booster version and orbit show mixed results in regards to its impact on the outcome.
- The position of the launching site relative to the **coastline and the Tropic of Cancer** are important variables that must be further explored.
- All our classification models performed well. Our best model was the Decision Tree Classifier. However, **our testing sample was very small** (n = 18) and the results are likely to be a classical example of **model overfitting**.
- The next step in this work is to test our models with out-of-sample data and to validate the relationship between proximity to coastline and the Tropic of Cancer.

Mario Sara - (January, 2022)

Appendix

- A list of all notebook links:

Click here for SpaceX API calls notebook on GitHub

Click here for EDA with EDA with SQL notebook

Click here for Web Scrapping notebook on GitHub

Click here for EDA with Folium Maps notebook

Click here for data wrangling related notebooks

Click here for the Machine Learning notebook

Click here for EDA with data visualization notebook

Click here for the Dashboard script

Thank you!

Mario Seriniva - (January, 2022)