IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alisson Moreira
Alissonmoreira99@Outlook.com
08/12/22

Outline

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

Introduction

Rocket launches are expensive, on average 165 million dollars, but there is a company called spaceX that managed to reuse the first stage of a rocket, causing the launch cost to drop to 62 million dollars.

Because of this, the company's rocket launch data was analyzed to determine whether or not a rocket will land. And with this information it is possible to define the price of the releases.

So the main purpose of this analysis is to determine whether the first stage of a rocket will land, so that we can determine the cost of a launch.

Executive Summary

For data collection, the spaceX API was used, and after that, some rows and columns were removed, as they contained null values. After the database is ready, the model training process is developed. As a result of this whole process, it was possible to predict, with a high rate of accuracy, whether or not a rocket would land

Section 1

Methodology

Methodology

Data collection methodology:

- The data was collected from an API through http request

- Perform data wrangling
    - In order to make the dataset consistent, some rows and columns were dropped, because of missing values.
- 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 order to build the machine learning model, it was necessary to separate the dependent attribute from the rest. Next, is necessary to standardize the features, split the data set and train the model

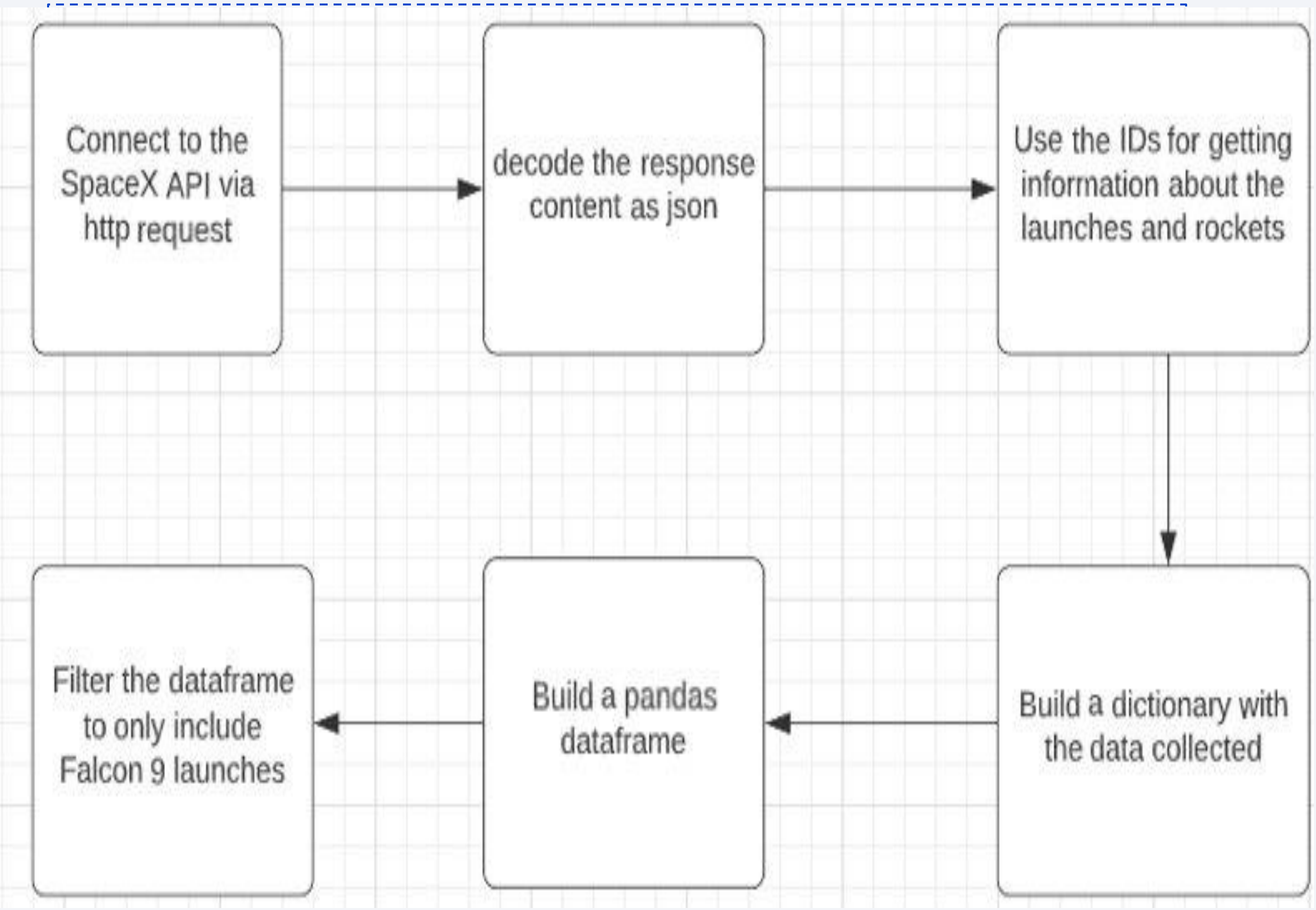
Data Collection – SpaceX API

Two call for the SpaceX API were necessary:

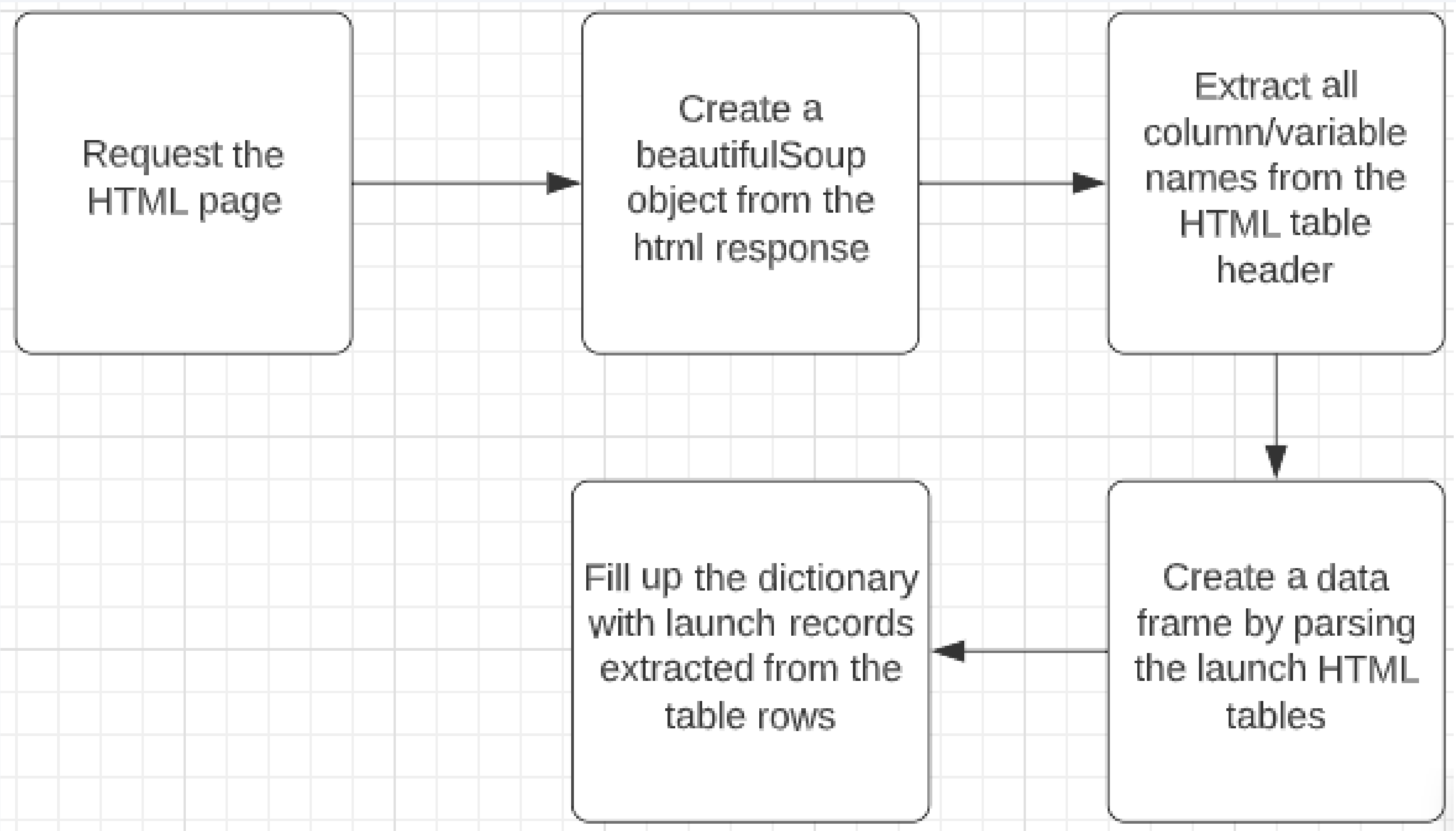
1. Get the table of Ids for the features;
2. Get the information from the API based on the Id's

Full code at:

https://github.com/DevMoreira99/Tasks/blob/main/Capstone/Extract%20data.ipynb



Data Collection - Scraping



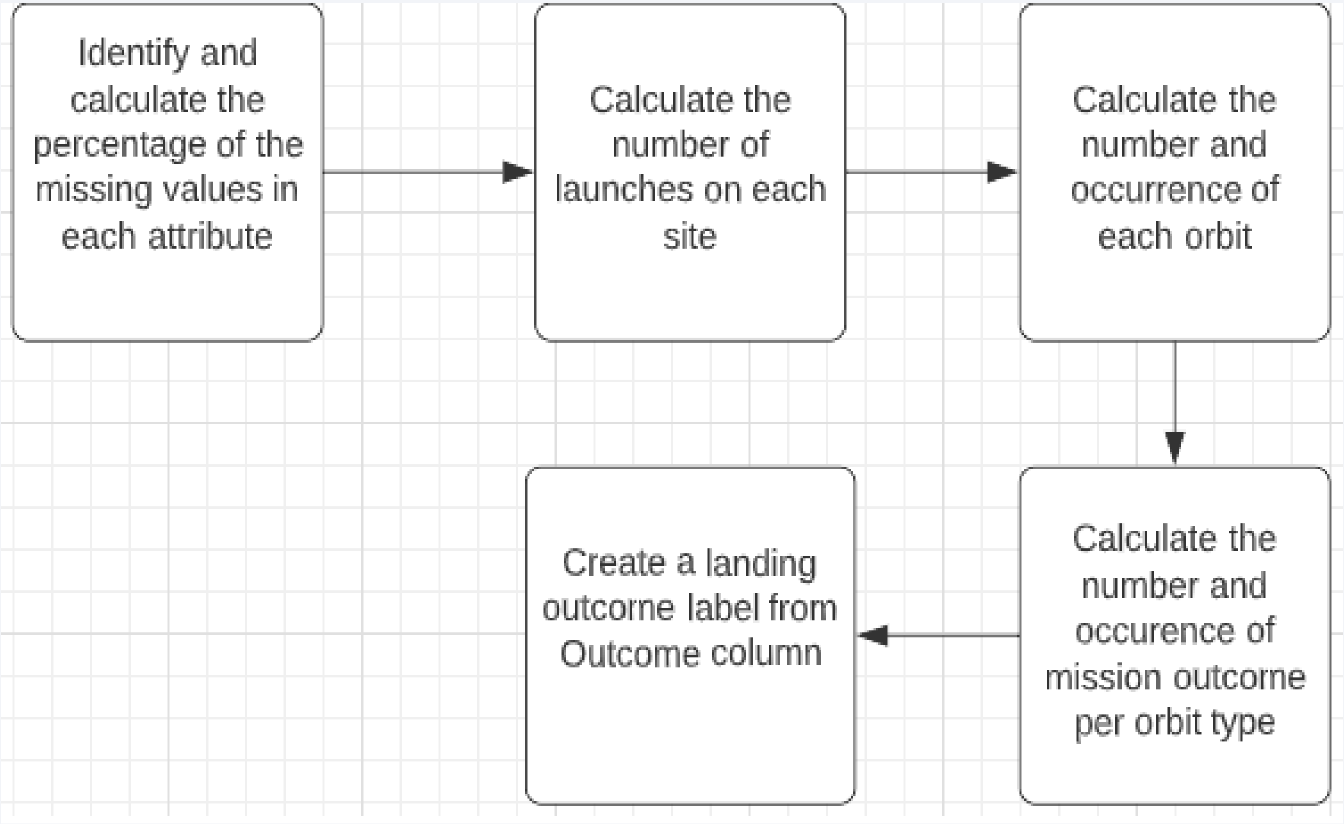
Full code at:
https://github.com/DevMoreira99/Tasks/blob/main/Capstone/webscraping.ipynb

Data Wrangling

In this phase, missing values were trated and a new landing outcome label was created.

Full code at:

https://github.com/DevMoreira99/Tasks/blob/main/Capstone/Data%20wrangling.ipynb



EDA with Data Visualization

Scatter Plot

The scatter plot was used to analyse the relationship between some feature variables and the dependent variable. This is the best chart for this because it's possible to see tendencies in case the two variables were related.

Bar chart

The bar chat was used to see the succes rate of each orbit. This chart is great for representing quantities by categories.

Line chart

The line chart the succes rate by year. That's the best option for representing time data.

Full code at:
https://github.com/DevMoreira99/Tasks/blob/main/Capstone/Exploring%20data.ipynb

EDA with SQL

- **Display the names of the unique launch sites in the space mission**
- **Display 5 records where launch sites begin with the string 'CCA'**
- **Display the total payload mass carried by boosters launched by NASA (CRS)**
- **Display average payload mass carried by booster version F9 v1.1**
- **List the date when the first succesful landing outcome in ground pad was acheived.**
- **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
- **List the total number of successful and failure mission outcomes**
- **List the names of the booster_versions which have carried the maximum payload mass**

Full code at:
https://github.com/DevMoreira99/Tasks/blob/main/Capstone/Exploring%20with%20sql.ipynb

Build a Dashboard with Plotly Dash

Pie Chart

The pie chart was used to divide the success rate by the platforms.
It's an interactive graph that responds to menu inputs.

Scatter plot

This graph was used to demonstrate the relationship between the success rate and the payload mass.

Full code at:

https://github.com/DevMoreira99/Tasks/blob/main/Capstone/spacex_dash_app.py

Predictive Analysis (Classification)

In order to find the best classification model, the following steps were applied:

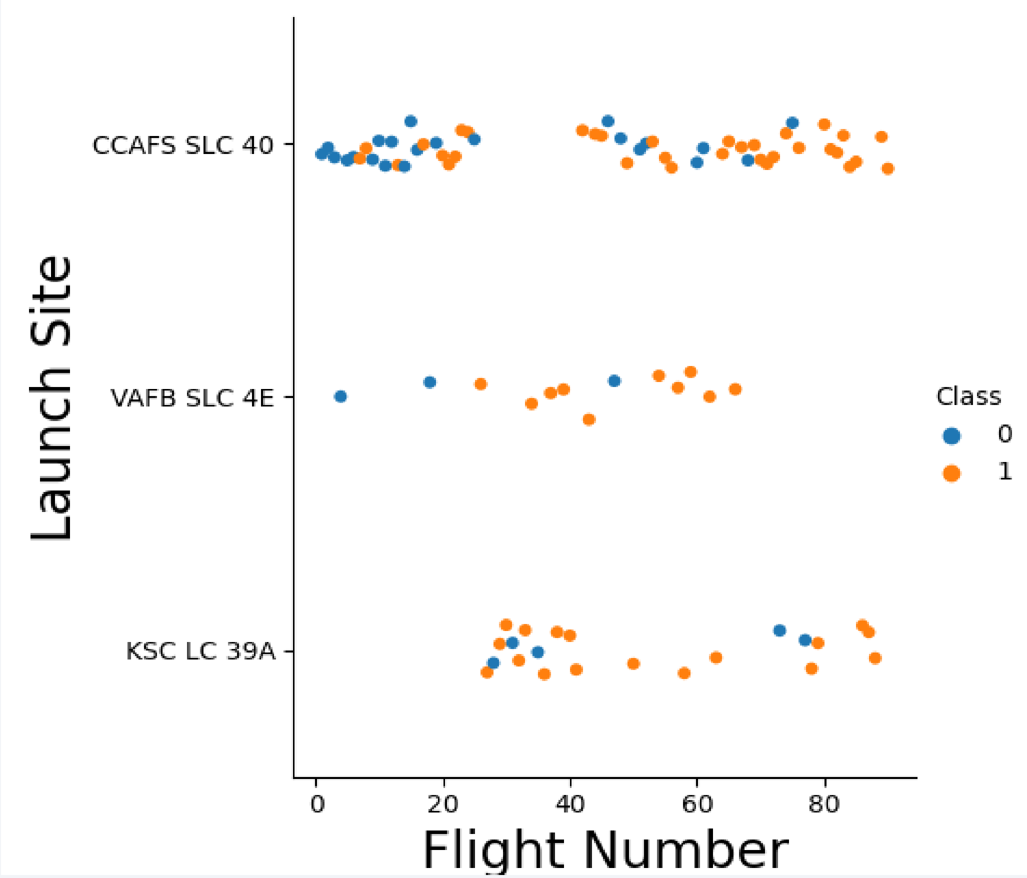
- Standardize the features.
- Split the data into test and training data.
- Then the models are trained and hyperparameter s are selected using the function GridSearchCV.
- Evaluate the score of the algorithm.

Full code:

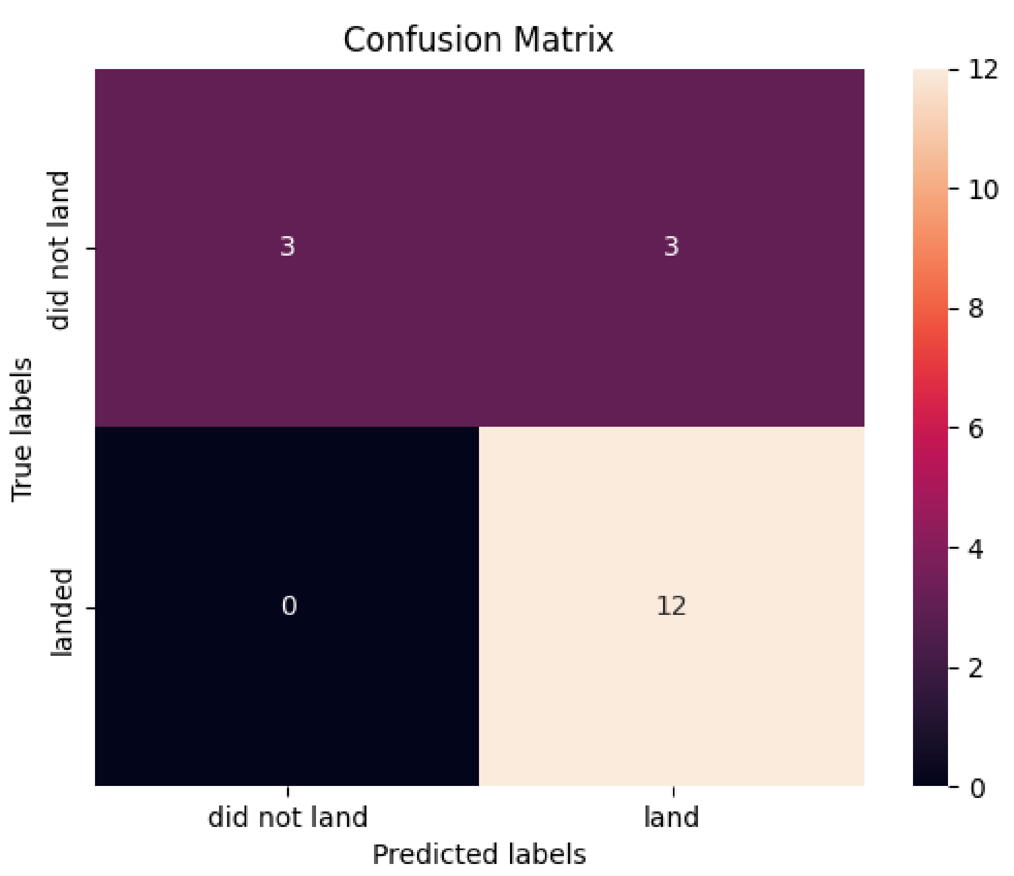
https://github.com/DevMoreira99/Tasks/blob/main/Capstone/machine%20learning.ipynb

Results

Success rate considering the platform and number of flights

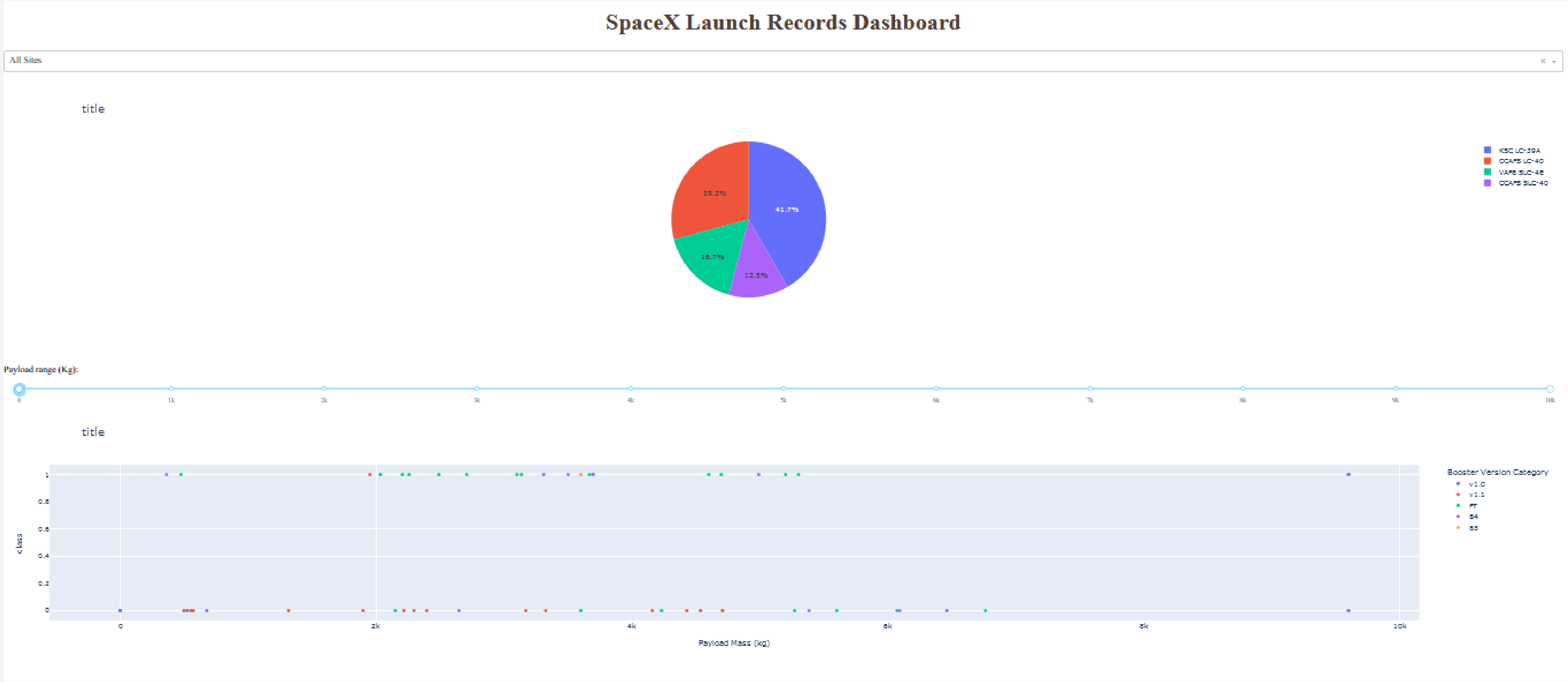


Results of decision tree algorithm, the one with the highest accuracy



Results

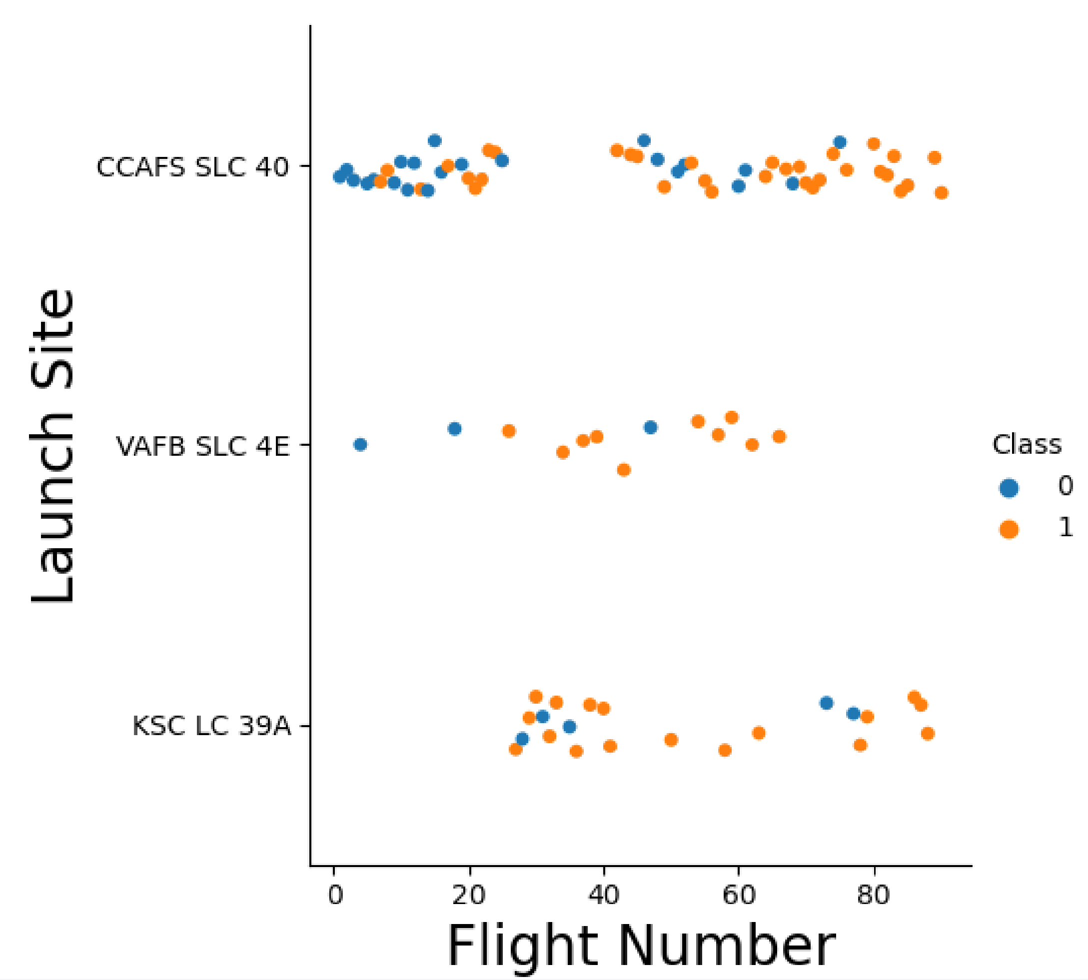
Dashboard for success rate analysis



Section 2

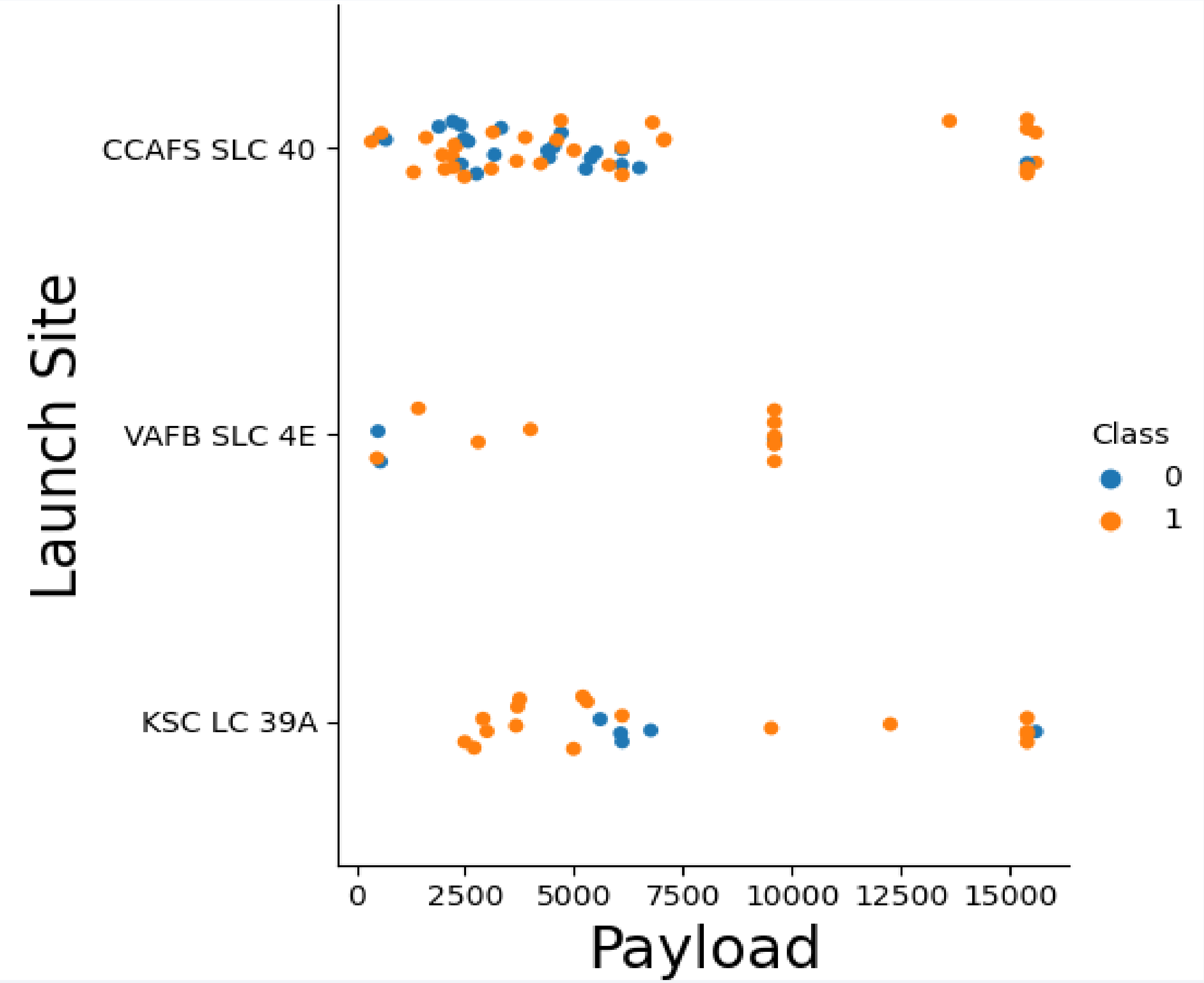
Insights drawn from EDA

Flight Number vs. Launch Site



This graph demonstrates that the more launch attempts are made, the higher the success rate.

Payload vs. Launch Site

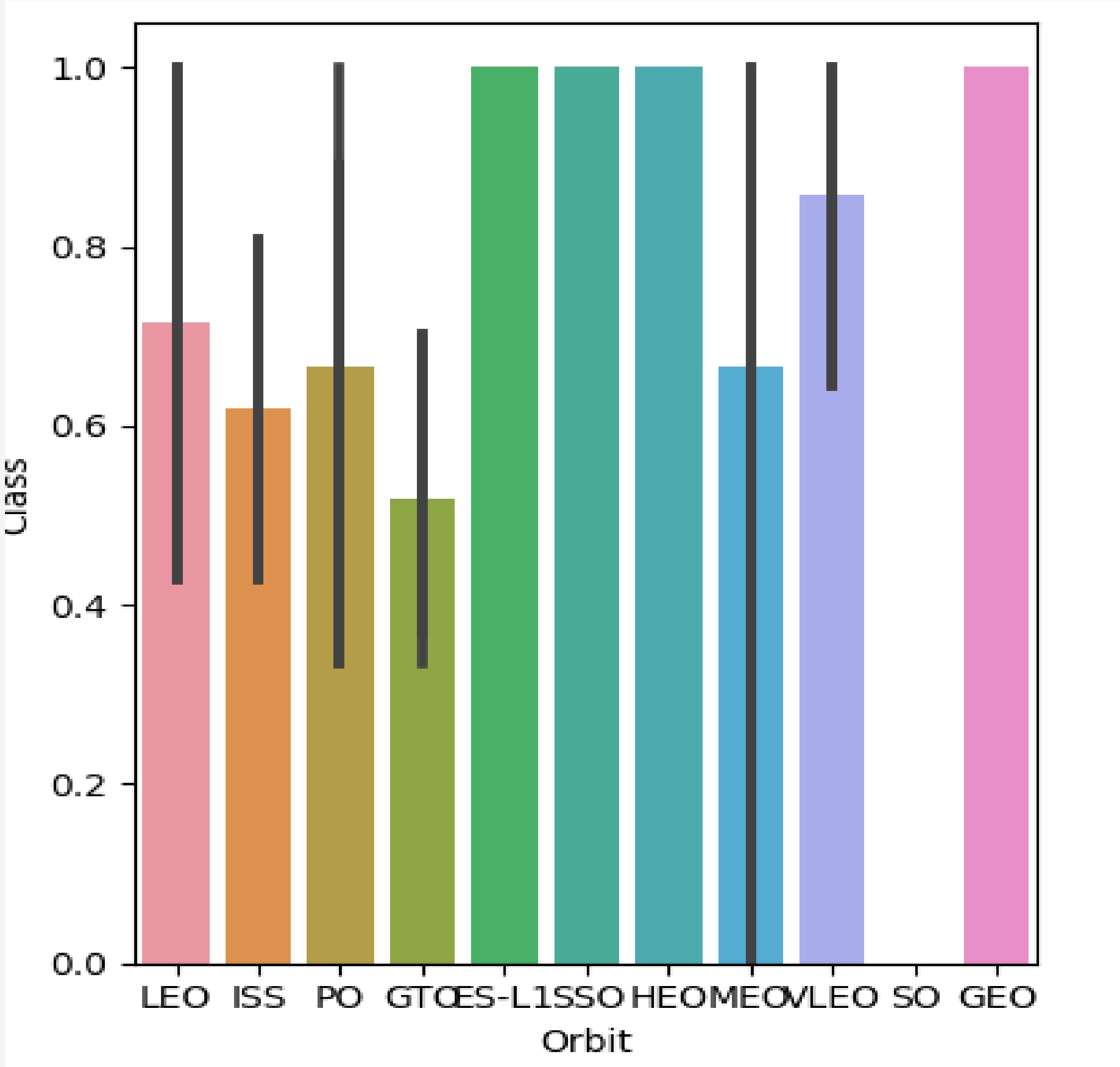


In this graph, it's possible to extract the following insights:

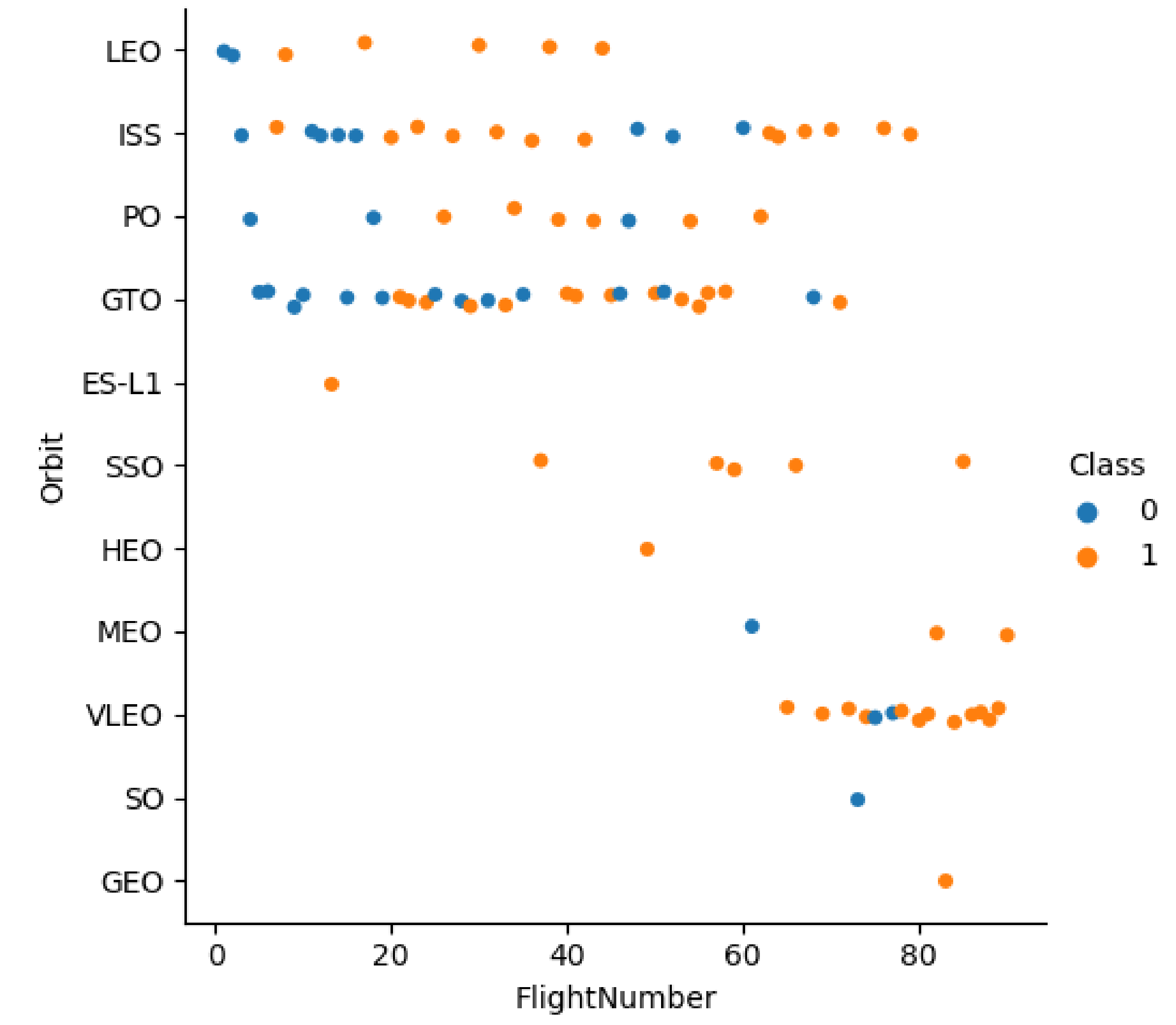
On the VAFB SLC 4E site, there was no rockets launched for heavypayload mass.

For the KSC LC, a low payload mass garantess more rate success;

Success Rate vs. Orbit Type

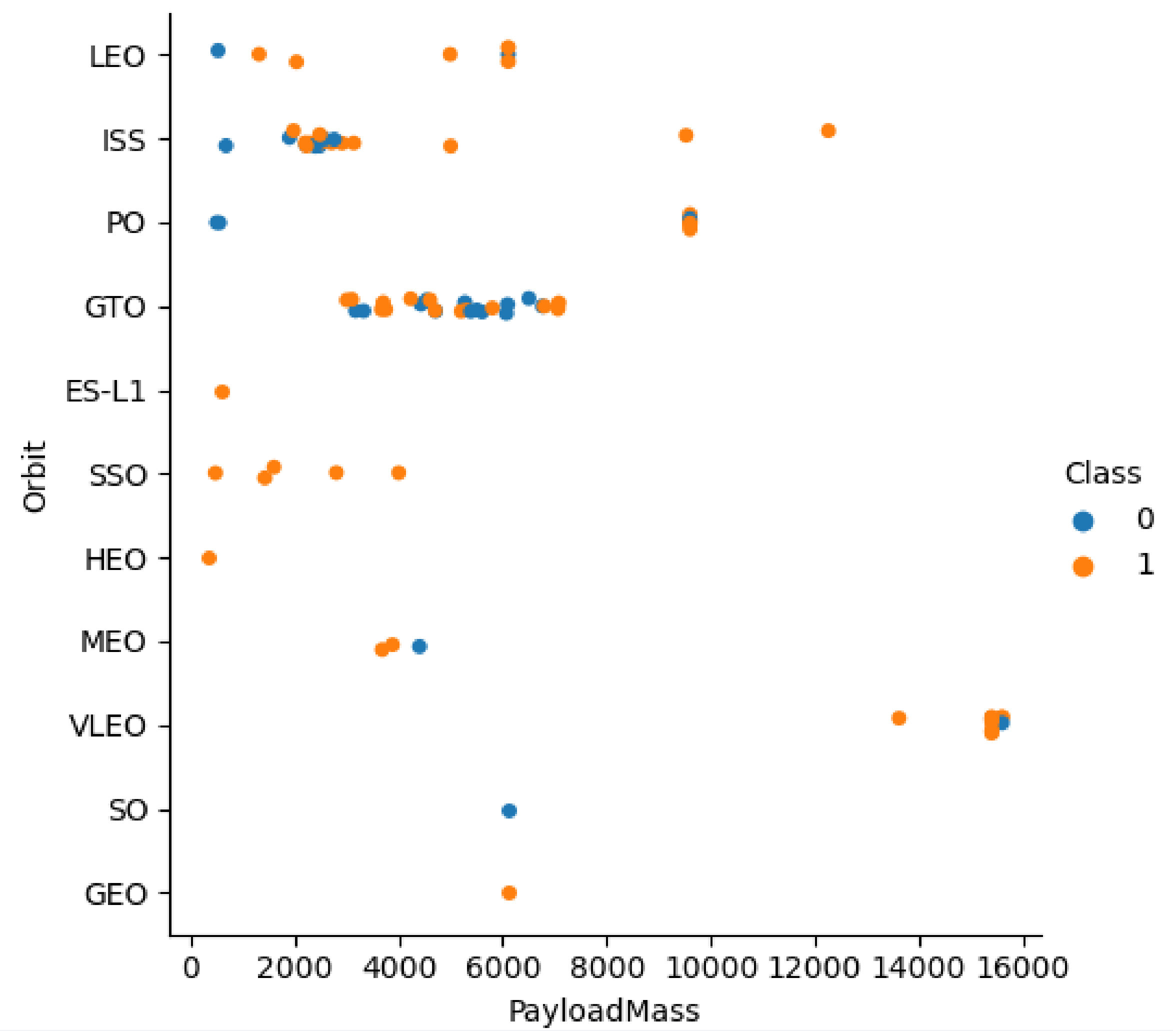


Flight Number vs. Orbit Type



In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

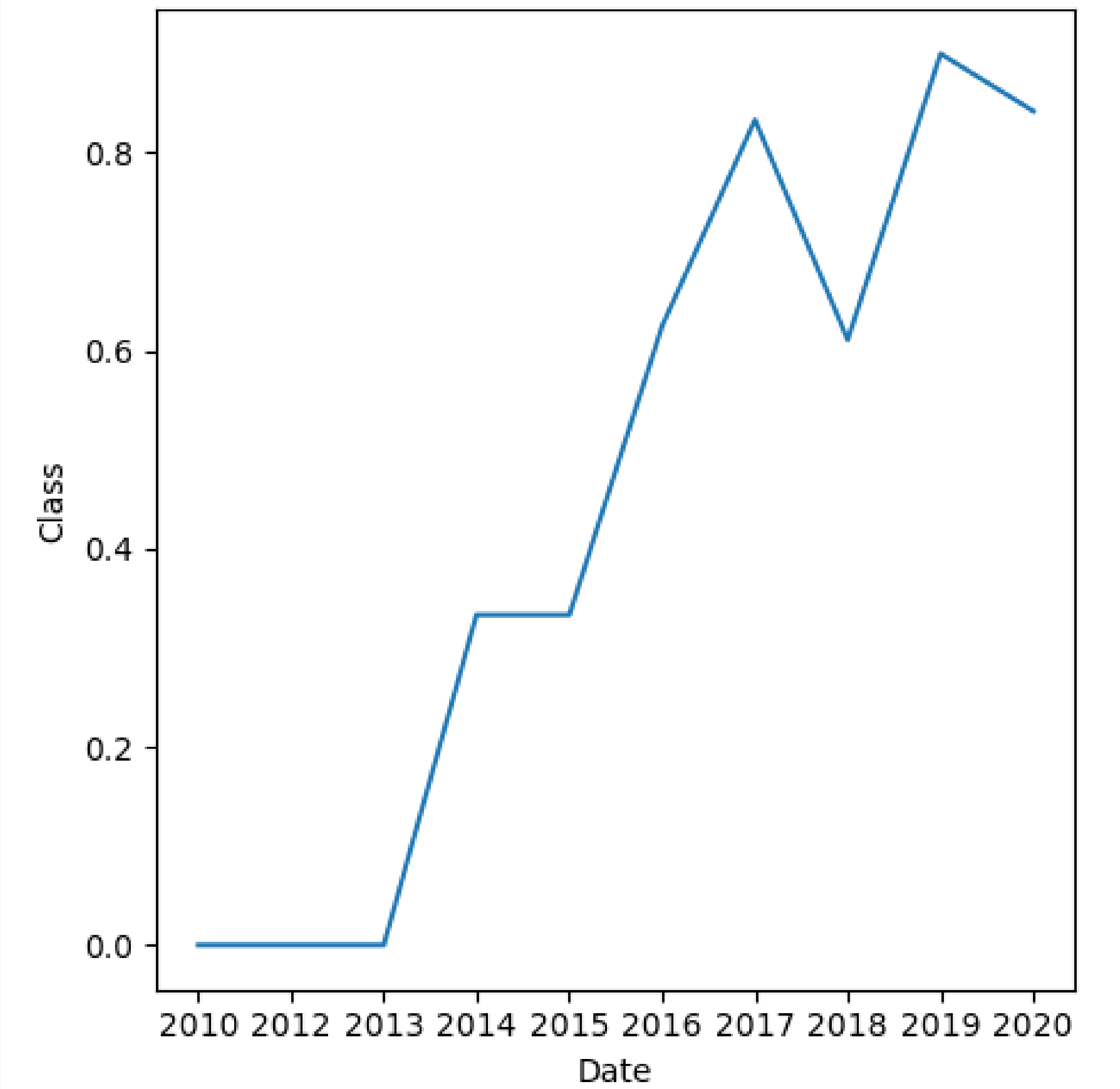
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



The success rate started and continues to grow s since 2013

All Launch Site Names

```
%sql select distinct("Launch_Site") from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select "Launch_Site" from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql select sum("PAYLOAD_MASS__KG_") from SPACEXTBL where Customer = "NASA (CRS)"
```

```
 * sqlite:///my_data1.db
Done.
```

sum("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

```
%sql select avg("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Booster_Version" = "F9 v1.1"
```

```
 * sqlite:///my_data1.db
Done.
```

avg("PAYLOAD_MASS__KG_")
2928.4

First Successful Ground Landing Date

```
%sql select * from SPACEXTBL as S where "Landing _Outcome" = "Success (ground pad)" order by Date(S.Date) asc limit 1
```

```
 * sqlite:///my_data1.db
Done.
```

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

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTBL where "Landing _Outcome" = "Success (drone ship)" and "PAYLOAD_MASS__KG_" between 40
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

```
%sql select "Mission_Outcome", COUNT("Mission_Outcome") \
from SPACEXTBL GROUP BY "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version" from SPACEXTBL where "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL)
```

```
 * sqlite:///my_data1.db
Done.
```

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

```
%sql select "Booster_Version", substr(Date, 4, 2) as month, substr(Date,7,4) as year, "Landing _Outcome", "Launch_Site"  from SPACEXTBL \
where substr(Date,7,4)='2015' and "Landing _Outcome" = "Failure (drone ship)"
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	month	year	Landing _Outcome	Launch_Site
F9 v1.1 B1012	01	2015	Failure (drone ship)	CCAFS LC-40
F9 v1.1 B1015	04	2015	Failure (drone ship)	CCAFS LC-40

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

```
%sql select "Landing _Outcome", count("Landing _Outcome") as CountOutcome, date("Date") FROM SPACEXTBL \
WHERE "Landing _Outcome" LIKE "Success%" \
group by "Landing _Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	CountOutcome	Date
Success	38	22-07-2018
Success (drone ship)	14	08-04-2016
Success (ground pad)	9	22-12-2015

Section 3

Launch Sites Proximities Analysis

Launch Site Location

There are four launch sites, three of them are very near to each other

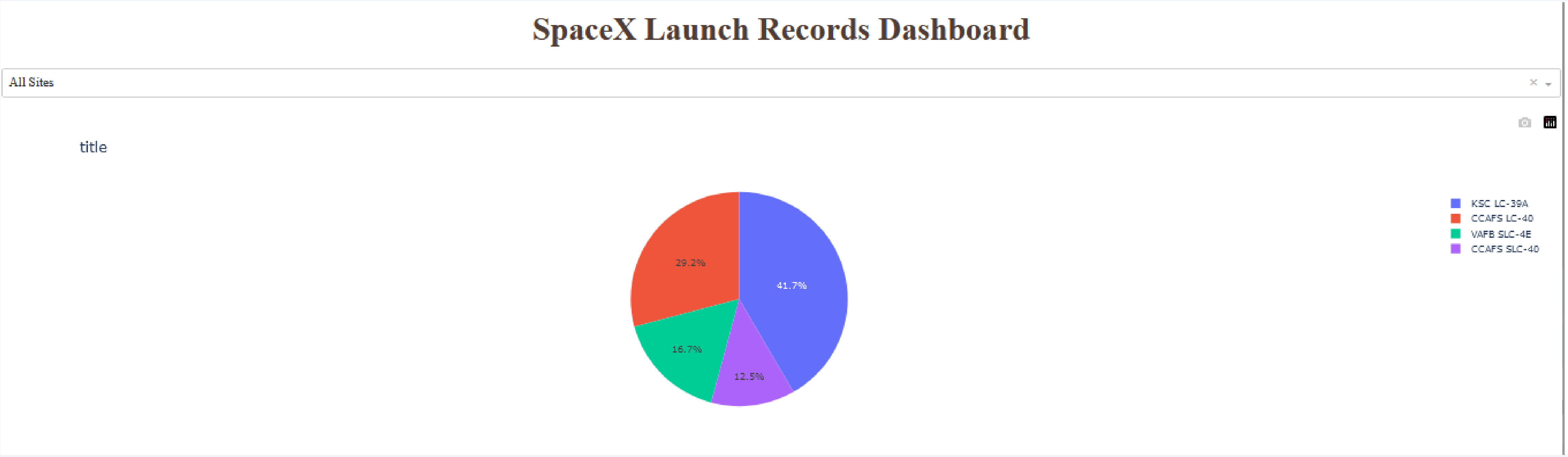
Locations Colored

Quantity of Success and Failures by location

Section 4

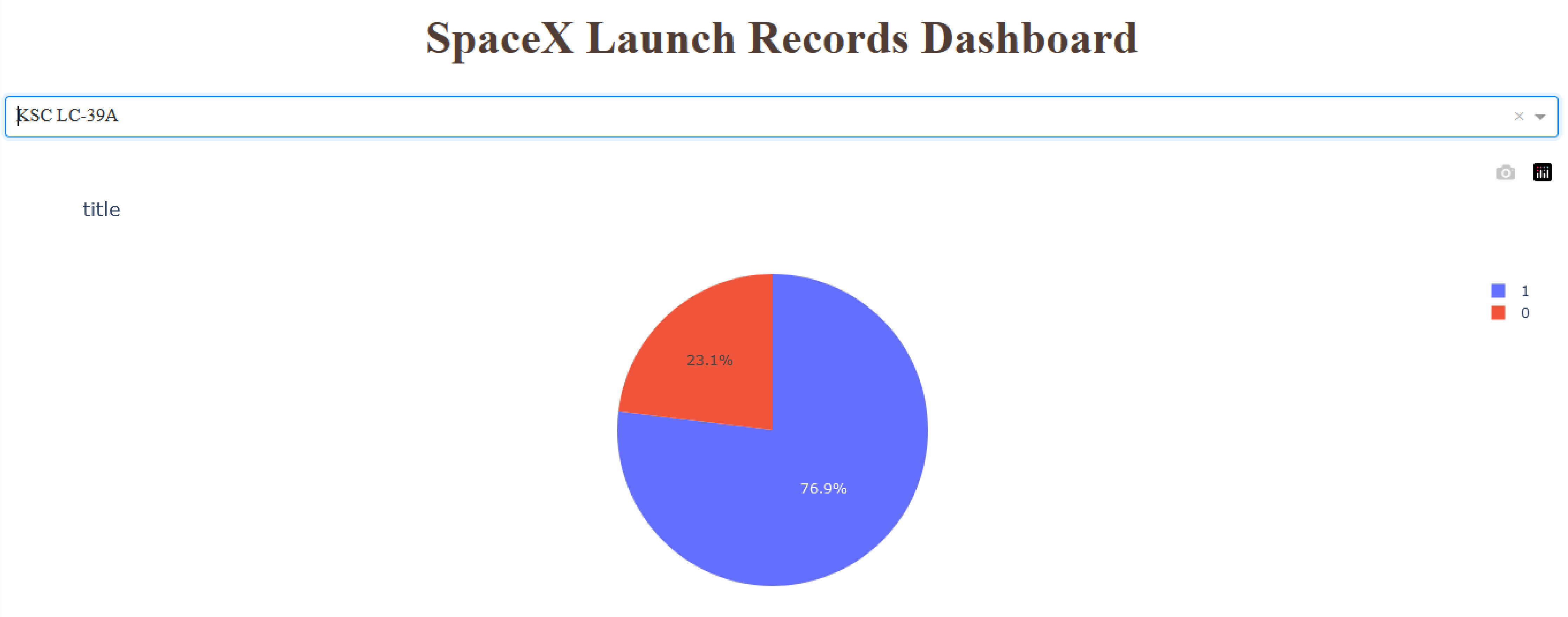
Build a Dashboard with Plotly Dash

Lauch success for all sites



In this graph is possible to see that the KSC LC site is the one with more success rate. This is even clear when we see only its class distribution.

Lauch site with highest launch success



The KSC LC is the launch site with more success rate

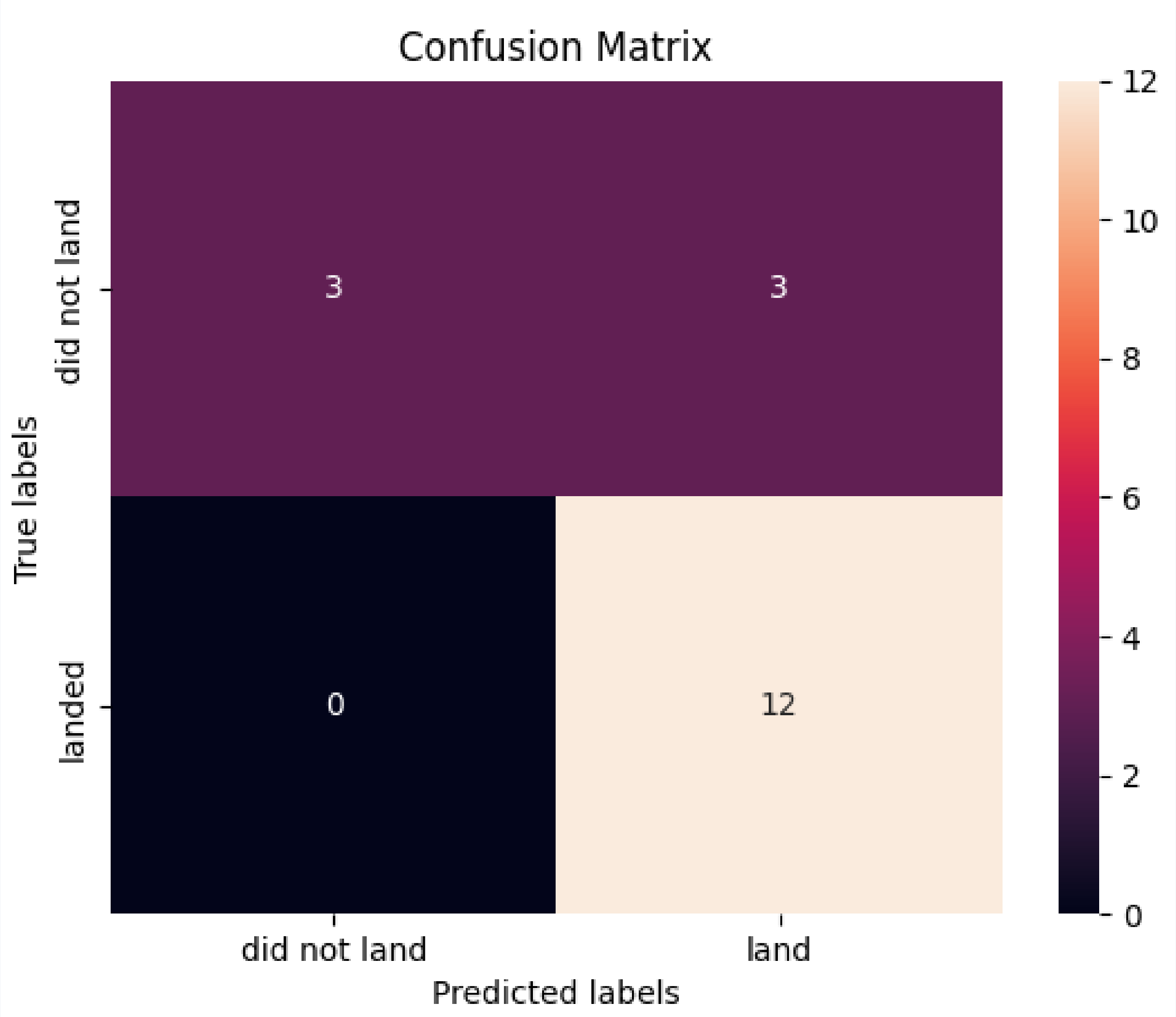
Payload vs Launch Outcome

The graph shows that there are only two booster categories that loaded above 5k payload, and of those two, only FT was successful between 5 and 8k. However, for a mass of 9600, only B4 was successful.

Section 5

Predictive Analysis (Classification)

Confusion Matrix



The best algorithm, which was the decision tree with 88% accuracy, had only 3 false positives

Conclusions

- The more launches are made, the more success rate will increase;
- The payload mass influences on the launch;
- Some Orbits have a higher success rate than others;
- The decision tree was the bast model with 88% of accuracy;
- Some Booster Version categories don't support heavy payloads mass;

Appendix

Datasets used

https://api.spacexdata.com

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json

https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Thank you!