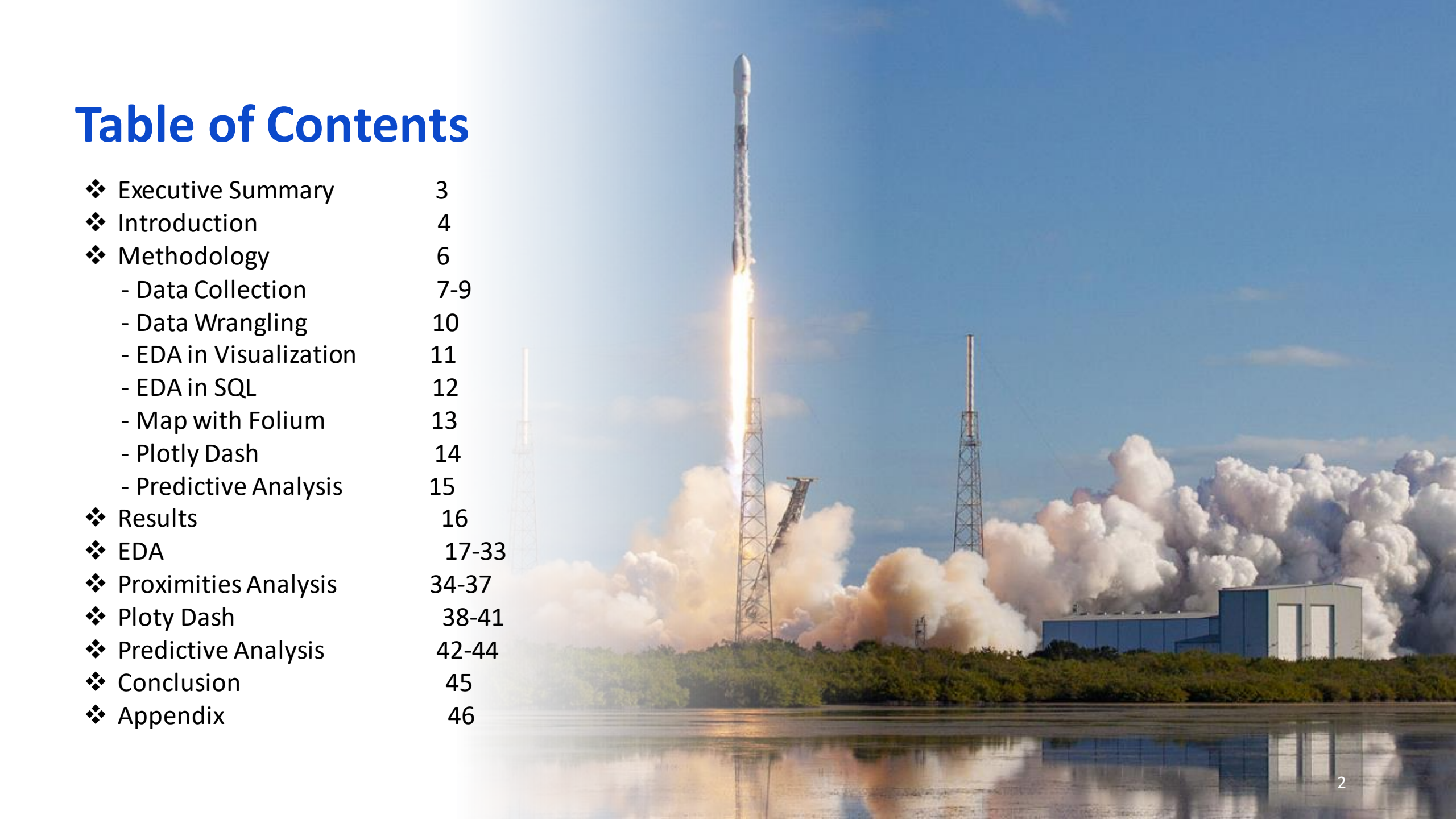
The Victory of SpaceX with Data Engineering

IBM Developer
SKILLS NETWORK

Ramazan Turkmen
12-09-2021

Table of Contents



Executive Summary

❖ Summary of methodologies

- Data Collection with APIs, web scrapping (BeautifulSoup), SQL connections
- Data Wrangling with Pandas, Numpy, SQL
- Data Visualization with Plotly, Seaborn, Dash, Folium
- Predictive Analysis with Sklearn

❖ Summary of all results

- Thanks to all of this projects we were able to have a better understanding of the success rate of space launch.
- But also to find the launching pads in the US.

Introduction

- It is aimed to use all phases of data science learned in the course, with the project in the last section, which consists of 9 chapters. Using the data of the space company, especially the Falcon 9 rocket has been launched on the space.

- The price of each launch is to gather and display information about Space X with dashboard. If SpaceX can reuse the Falcon 9 in the first stage, it will be realized with the success of machine learning.

Metodology

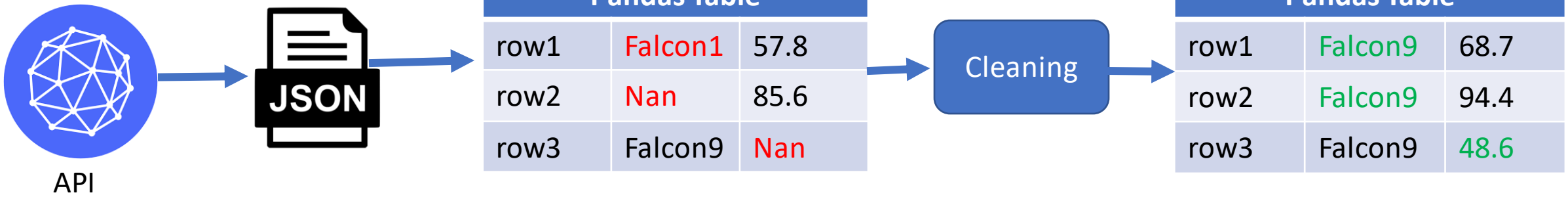
Methodology

Executive Summary

- Data collection methodology:
  - We used web scrapping, SQL queries and APIs
- Perform data wrangling
  - The data were most of the time standardized and processed with NumPy and Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were build using Sklearn, we used a GridSearch with 10 fold of cross validation, in the end we use the accuracy to determine the best classifier

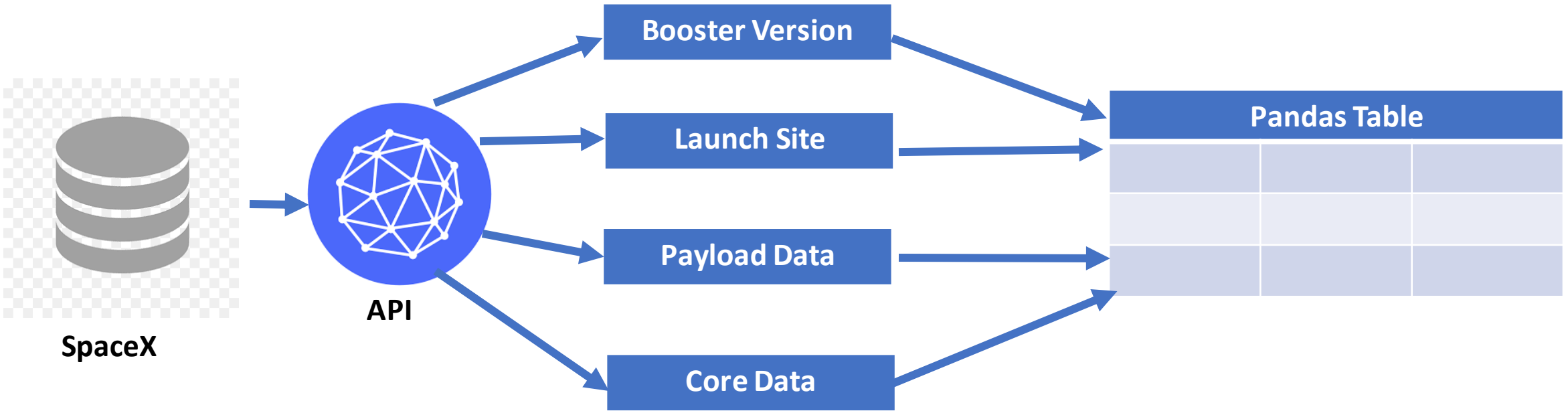
Data Collection

- We used the SpaceX data **API** for data collection
- Thanks to the SpaceX Api, we were able to collect lauches data in **JSON** format and then we converted it to **Pandas** DataFrame. After that, we **cleaned** the missing dataset and Falcon1 data that will not be useful to us.



Data Collection – SpaceX API

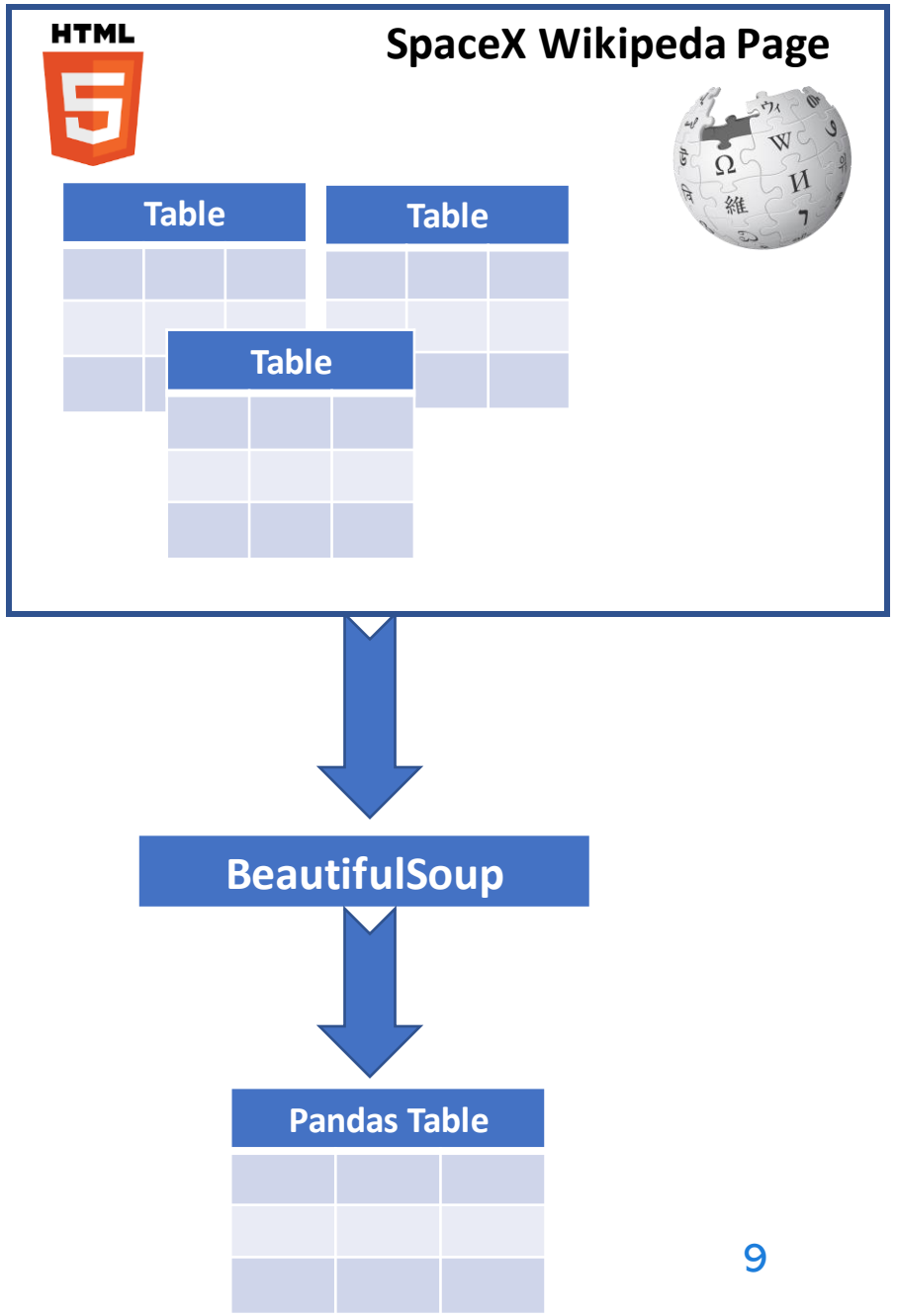
Thanks to multiple requests we were able to collect the **Booster Version**, the **Launch Site**, the **Payload data** and the **Core data**



rmzturkmen/github

Data Collection - Scraping

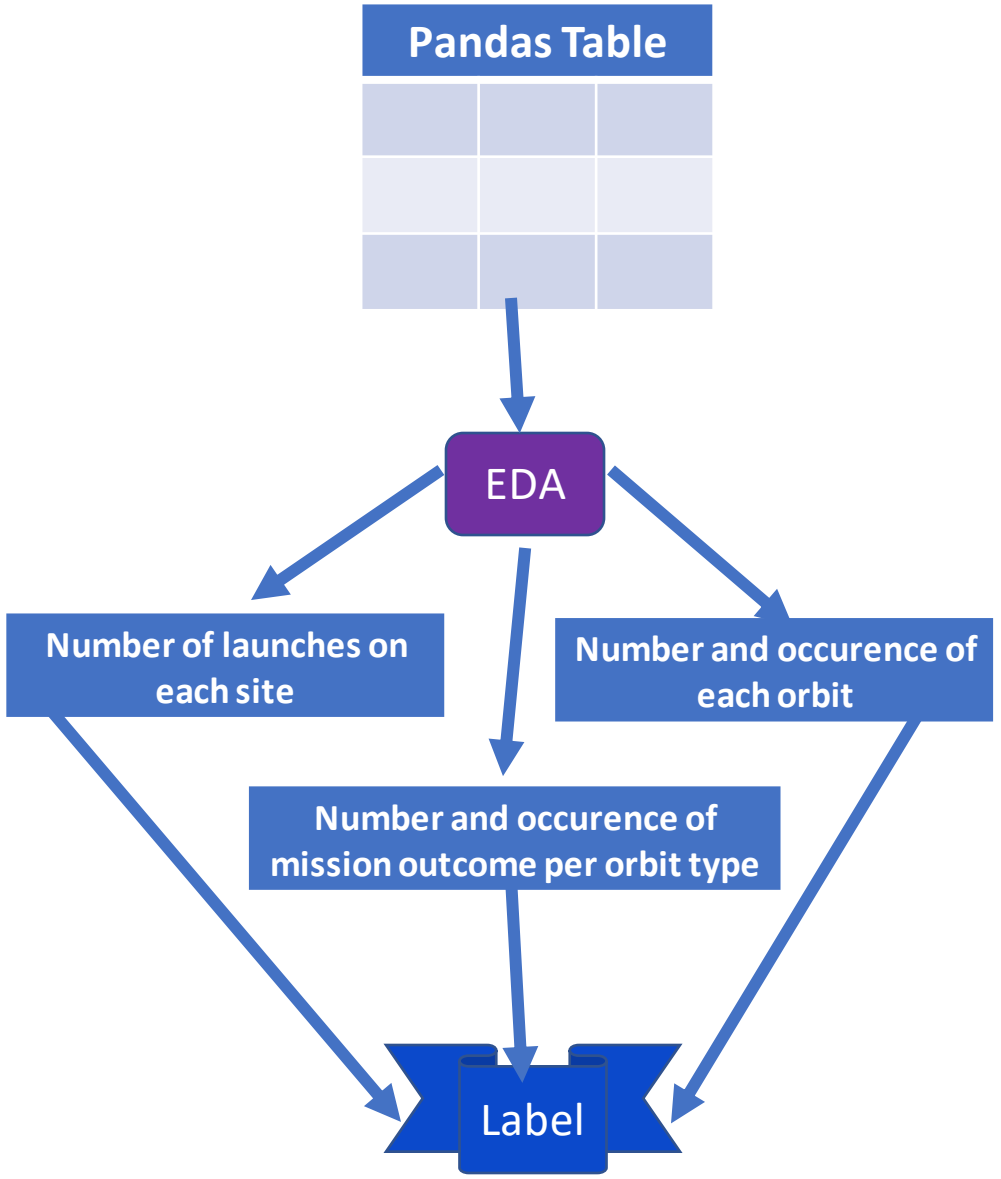
- Thanks to BeautifulSoup it is easy to do web scrapping and especially collect tables from HTML.
- Here we used **BeautifulSoup** to extract the header of tables and then populate the table with the rows collected



rmzturkmen/github

Data Wrangling

- Thanks to Exploratory data Analysis we were able to determine the training labels.
- We computed the number of launches on each site, the number of occurrence of each orbit, the number and occurrence of mission outcome per orbit type and landing outcome **label** from Outcome column



rmzturkmen/github

EDA with Data Visualization

- A scatter plot of **FlightNumber vs. PayloadMass** based on the outcome to see if as the flight number increases, the first stage is more likely to land successfully and if there is an impact of PayloadMass
- Then a second scatter plot of **LaunchSite vs. FlightNumber** based on the outcome to understand if the LaunchSite does not become a problematic element following the number of rockets which are launched there (FlightNumber).
- A third scatter plot to observe if there is any relationship between **LaunchSite and their PayloadMass**
- A bar chart to visually check if there are any relationship between the **Success Rate and the Orbit Type**
- Another scatter plot to visualize the relationship between **FlightNumber and Orbit type**
- A last scatter plot to reveal the relationship between **Payload and Orbit type**
- And a line plot to get the **average launch success trend**

EDA with SQL

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster_versions which have carried the maximum payload mass
- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

rmzturkmen/github

Build an Interactive Map with Folium

- In this part we marked the NASA Johnson Space Center with a blue circle and a Marker. Then we marked with Circle and Marker each launch which results in marking launching sites. Thanks to Marker we have marked successful and failed launches with green and red colors. In the end we add a Marker and a Line to the closest railway.

- All of those Markers Lines and Icons helped us understand the geography of launch sites and their proximity to cities, railways, coast etc..

Build a Dashboard with Plotly Dash

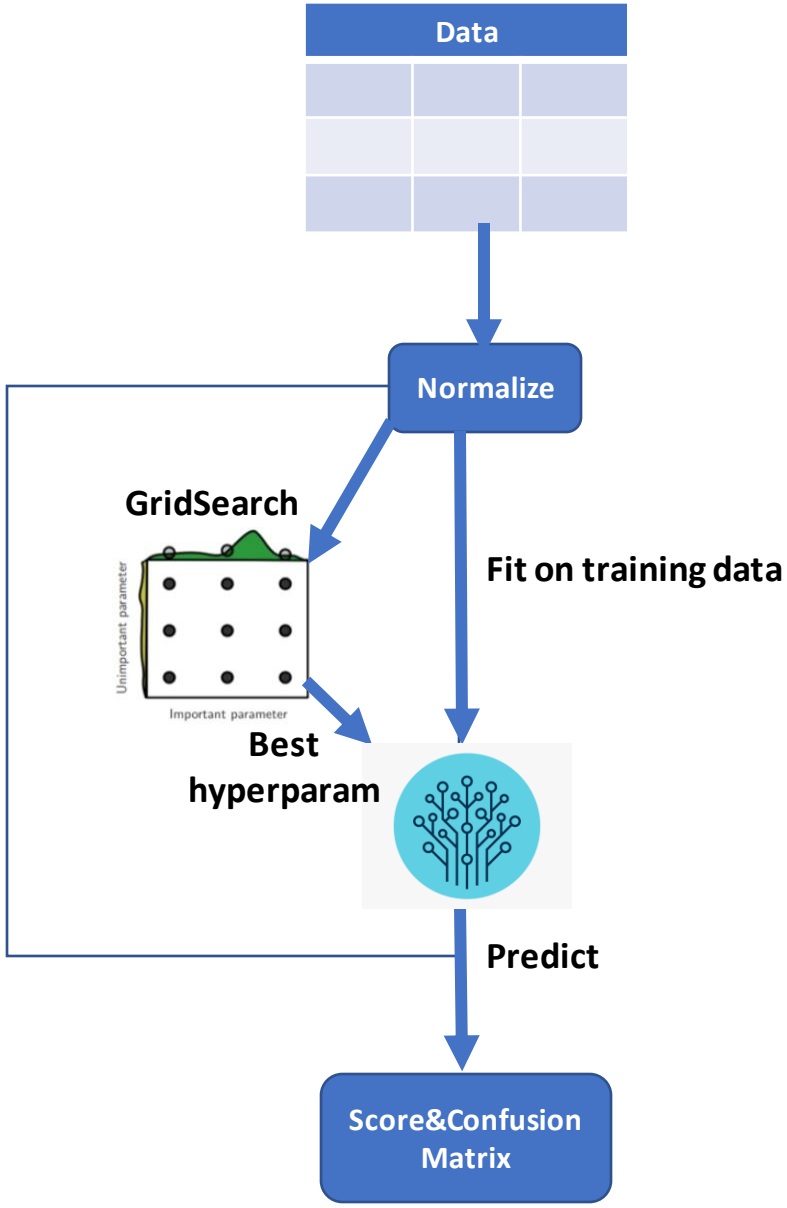
- In the Dashboard we add a pie chart of success rate for each launch site and also a scatterplot of the payload mass by the outcome of the mission (class) all this based on Booster versions

- The first plot helps to understand the success rate on each launching site, the second demonstrate which booster as the higher success rate and that it is related to the Payload mass.

rmzturkmen/github

Predictive Analysis (Classification)

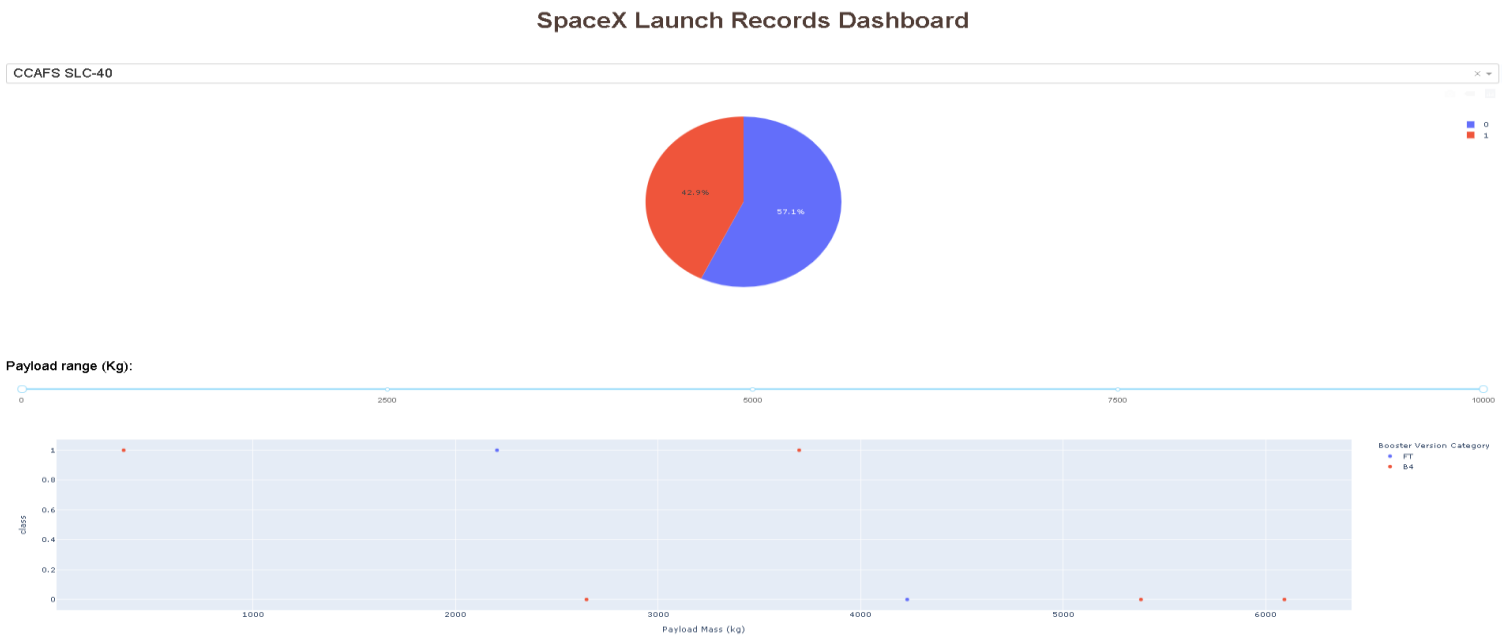
Models were built using **Sklearn**, data were previously **normalized** and models **hyperparameters** were found using a **GridSearch** with a 10 fold cross validation, in the end the best performing model has been selected based on accuracy.



rmzturkmen/github

Results

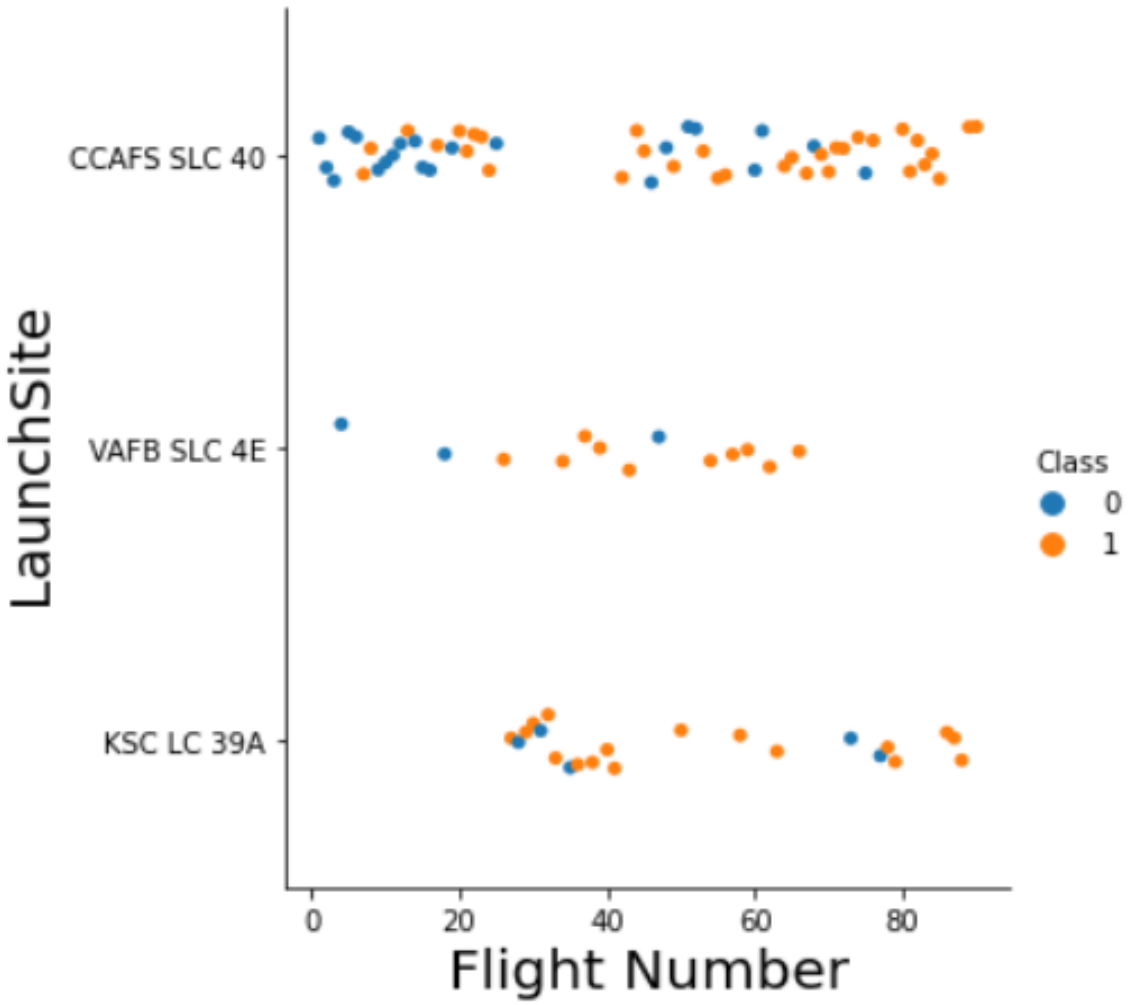
- The EDA shows the importance of the payload mass, but also the booster version on the success rate of the launch missions
- All machine learning models studied were equally performing



Insights Drawn from EDA

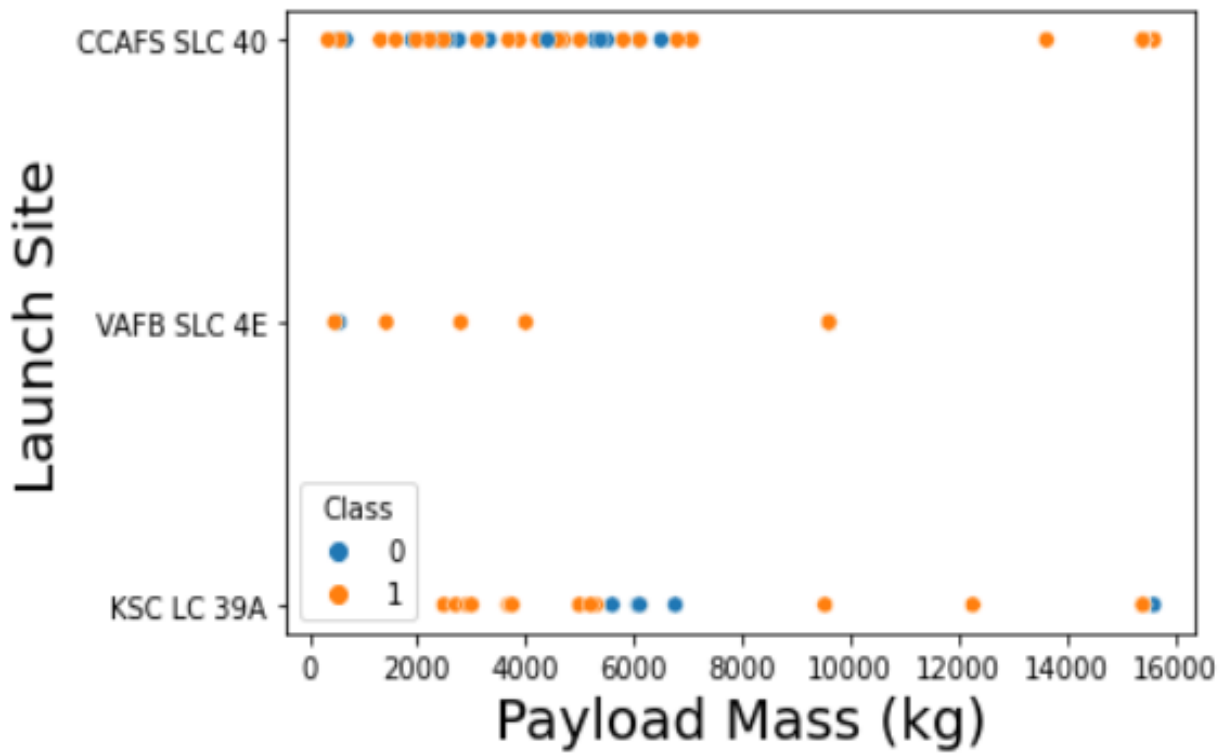
Flight Number vs. Launch Site

This chart indicates that a “young” launching site will probably a lower success rate than one which had a lot of rocket launched from.



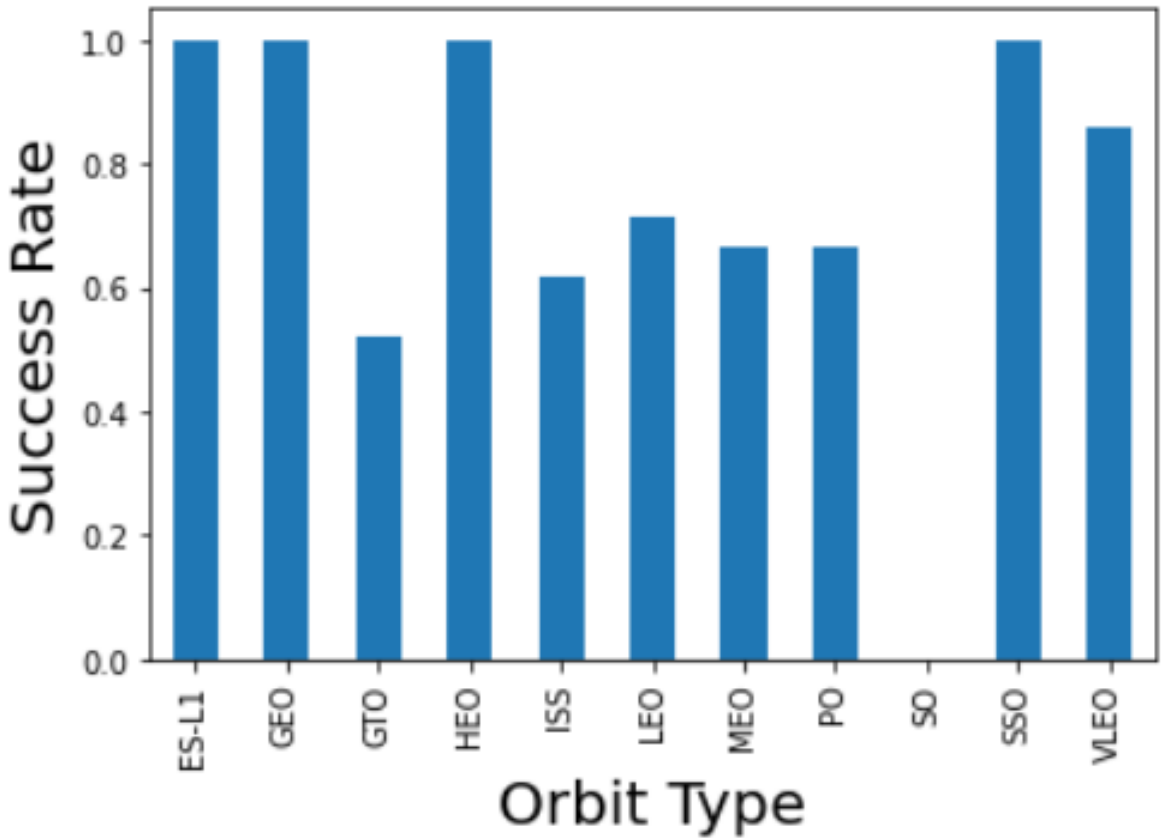
Payload vs. Launch Site

It seems like a lot of rocket launched had a payload between 500kg and 6000kg. Also the launching site VAFB SLC 4E seems to be a site where there are not that much rocket launched. An impact of the payload could be possible but it will need further analysis.



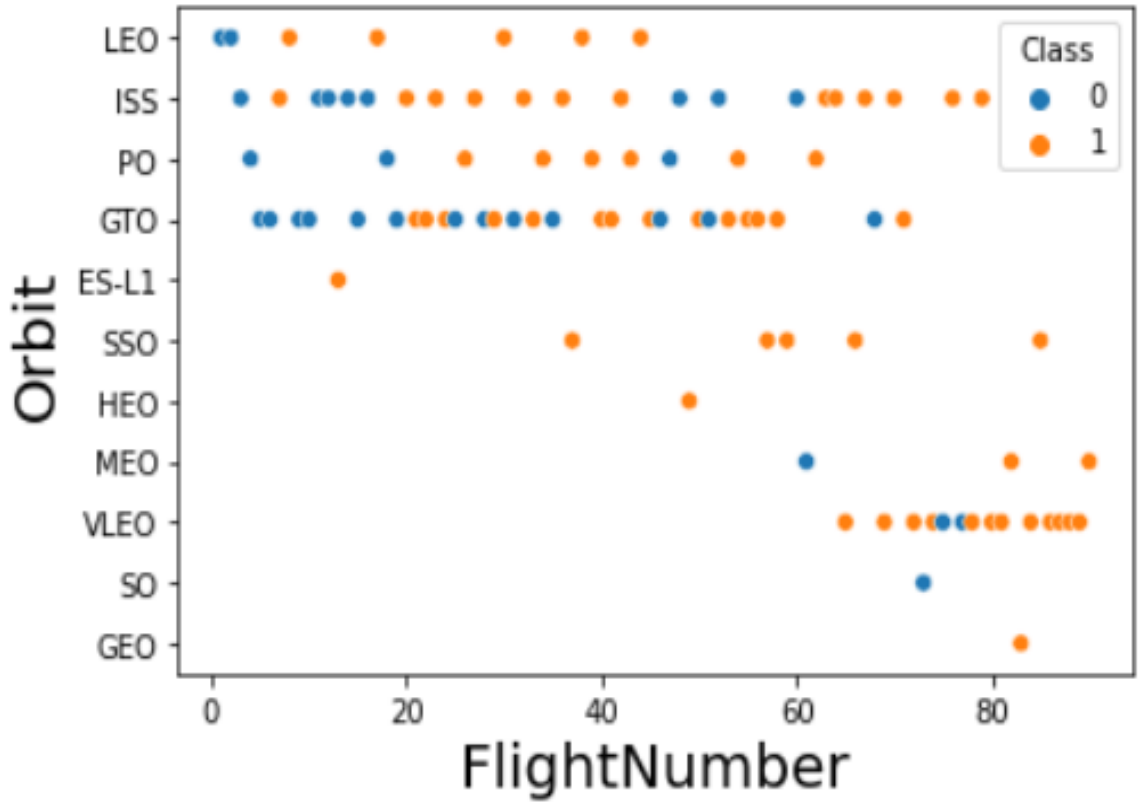
Success Rate vs. Orbit Type

There is a strong correlation between these two indeed as we can observe the SO or GTO Orbit Type are quite risky as the success rate is below 0.6. However, some Orbit Type provide a 1.0 success rate which is perfect but can hide suspicious data. Indeed if for this Orbit type only one rocket has been launched the reliability of this hypothesis is null.



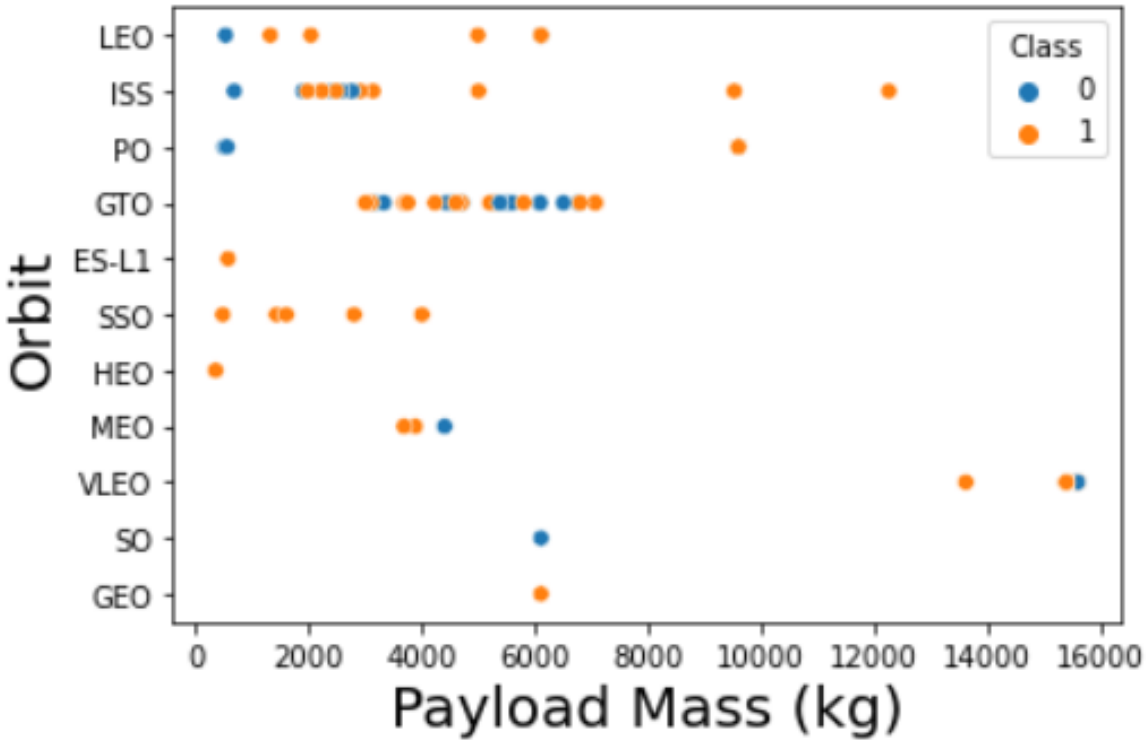
Flight Number vs. Orbit Type

This chart confirmed what has been said before, some Orbit type have only couples of Flights in their history and thus make data quite confusing. However for the GTO,VLEO and ISS it seems like there are enough data to be confident on those data.



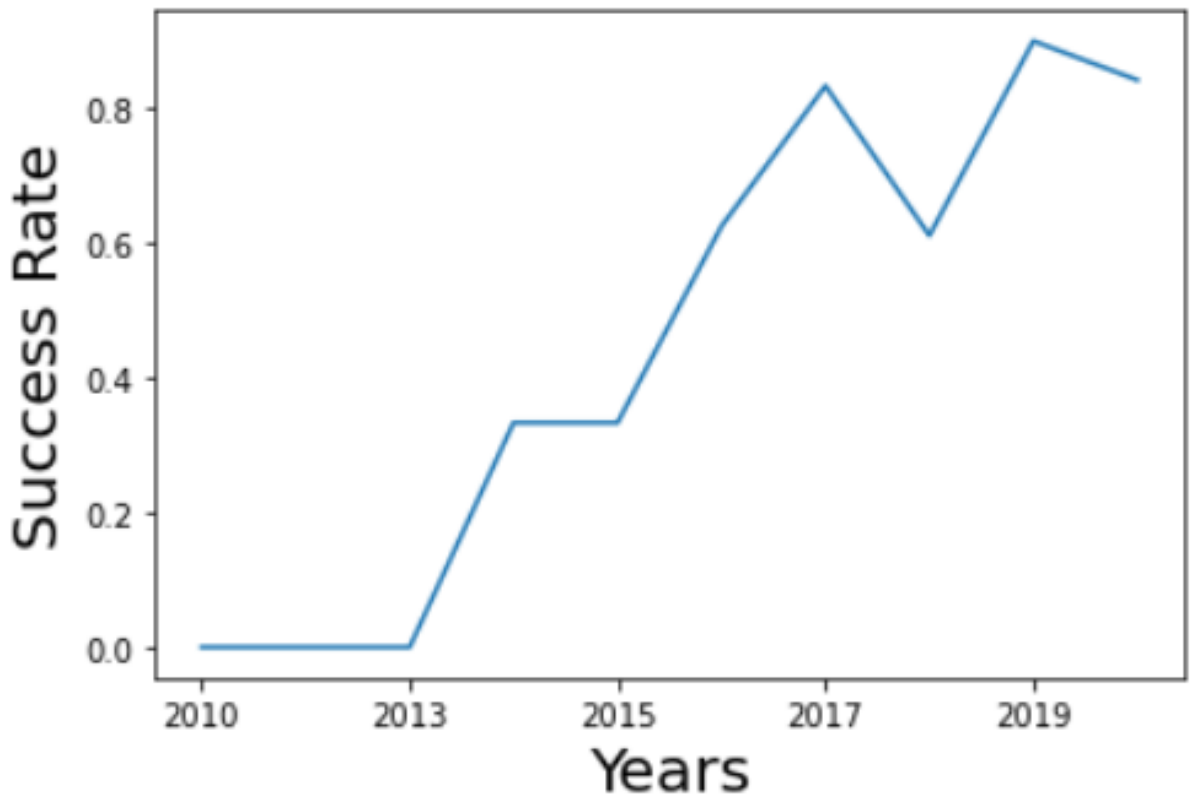
Payload vs. Orbit Type

Here we can observe that certain sites have a strong relation with the payload mass, for example the GTO and ISS.



Launch Success Yearly Trend

Here the chart demonstrates that as Humans learn more and more through the years thanks of Sciences,it results in a significant rocket launches success rate increasing.



All Launch Site Names

❖ Here are the launch site names obtained by **SQL** query:

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

Launch Site Names Begin with 'CCA'

❖ Here are 5 records where launch sites begin with `CCA` names obtained by **SQL** query:

Launch Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- The total payload carried by boosters from NASA is **99.980 kg.**

- This has been obtained by **SQL** query.

Average Payload Mass by F9 v1.1

❖ Hier is the average payload mass carried by booster version F9 v1.1:

- **2.534 kg**

❖ This has been obtained by **SQL** query.

First Successful Ground Landing Date

❖ Hier is the dates of the first successful landing outcome on ground pad:

- 2015-12-22

❖ This has been obtained by **SQL** query.

Successful Drone Ship Landing with Payload between 4000 and 6000

❖ Hier is the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

Booster Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

❖ This has been obtained by **SQL** query.

Total Number of Successful and Failure Mission Outcomes

❖ Hier is the total number of successful and failure mission outcomes:

Mission Outcome	Count
Failure(in flight)	1
Success	99
Success (payload status unclear)	1

❖ This has been obtained by **SQL** query.

Boosters Carried Maximum Payload

- Here is the names of the **booster** which have carried the maximum **payload mass:**

- This has been obtained by **SQL** query.

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

- Hier is the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

- This has been obtained by **SQL** query.

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

- Hier is the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between the date 2010-06-04 and 2017-03-20, in descending order.

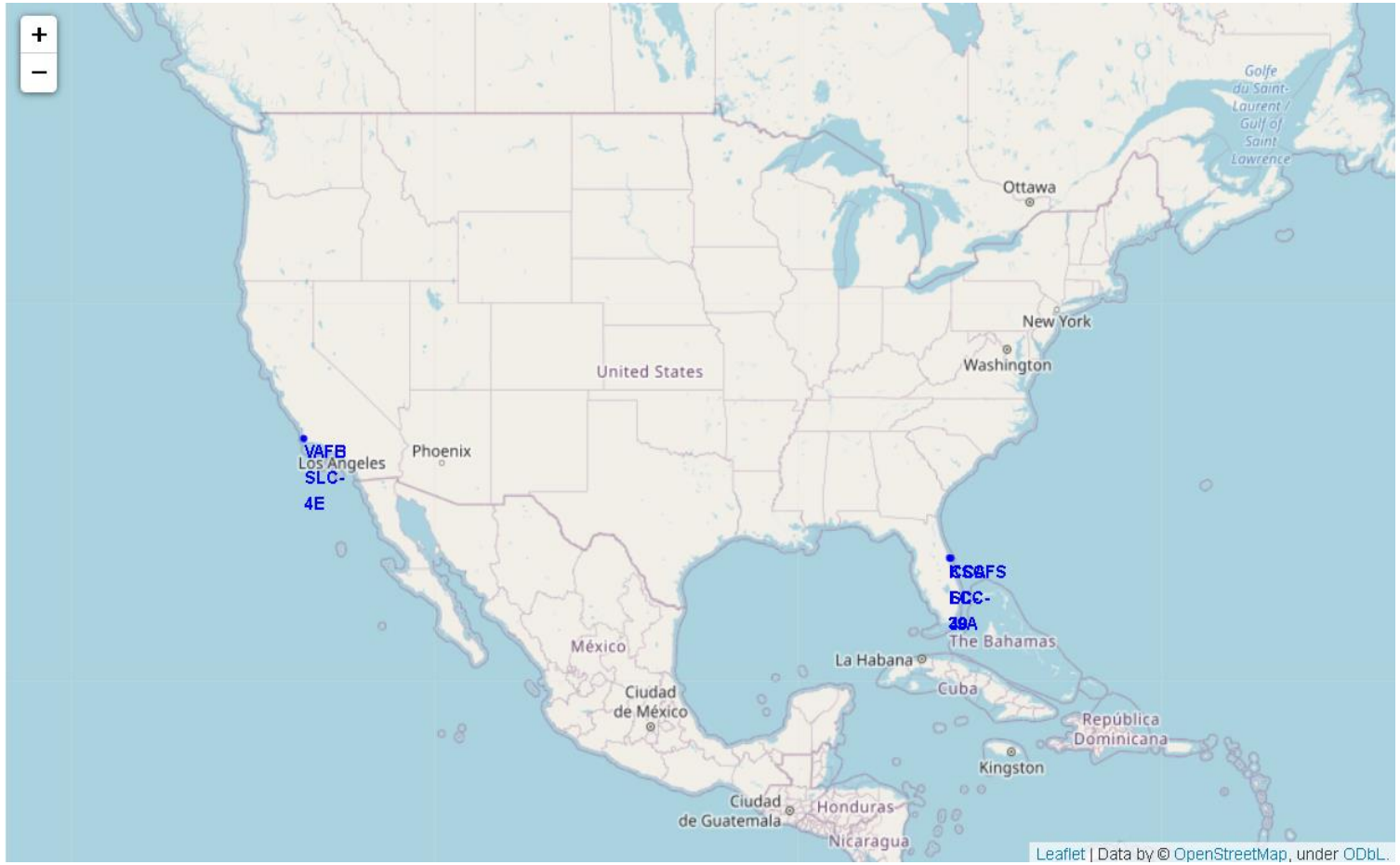
- This has been obtained by **SQL** query.

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

Launch Sites Proximities Analysis

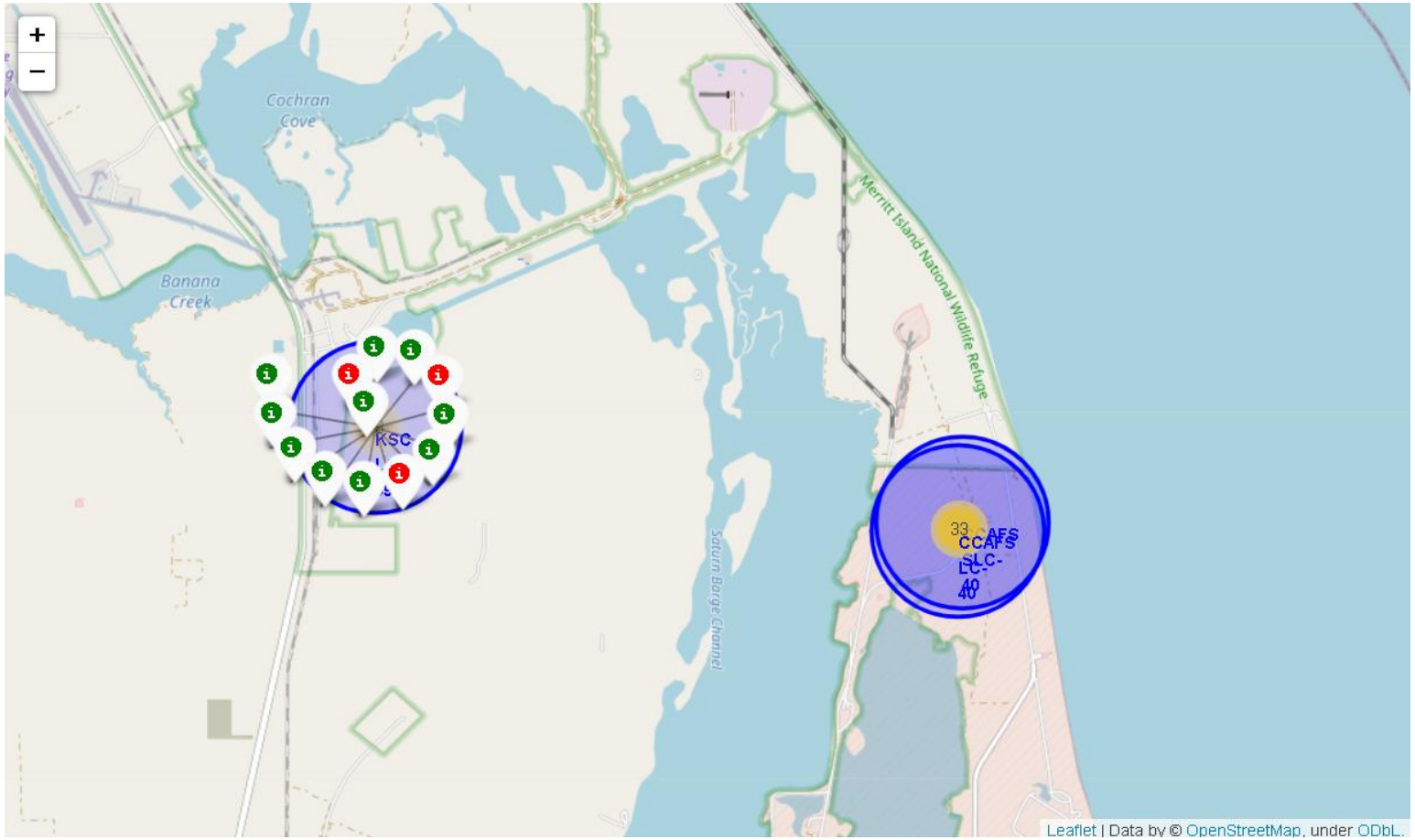
Launch Sites in the US

Here we can see launching sites in the US with blue colored signs. However, in Florida, the three sites cannot be seen due to their close proximity.



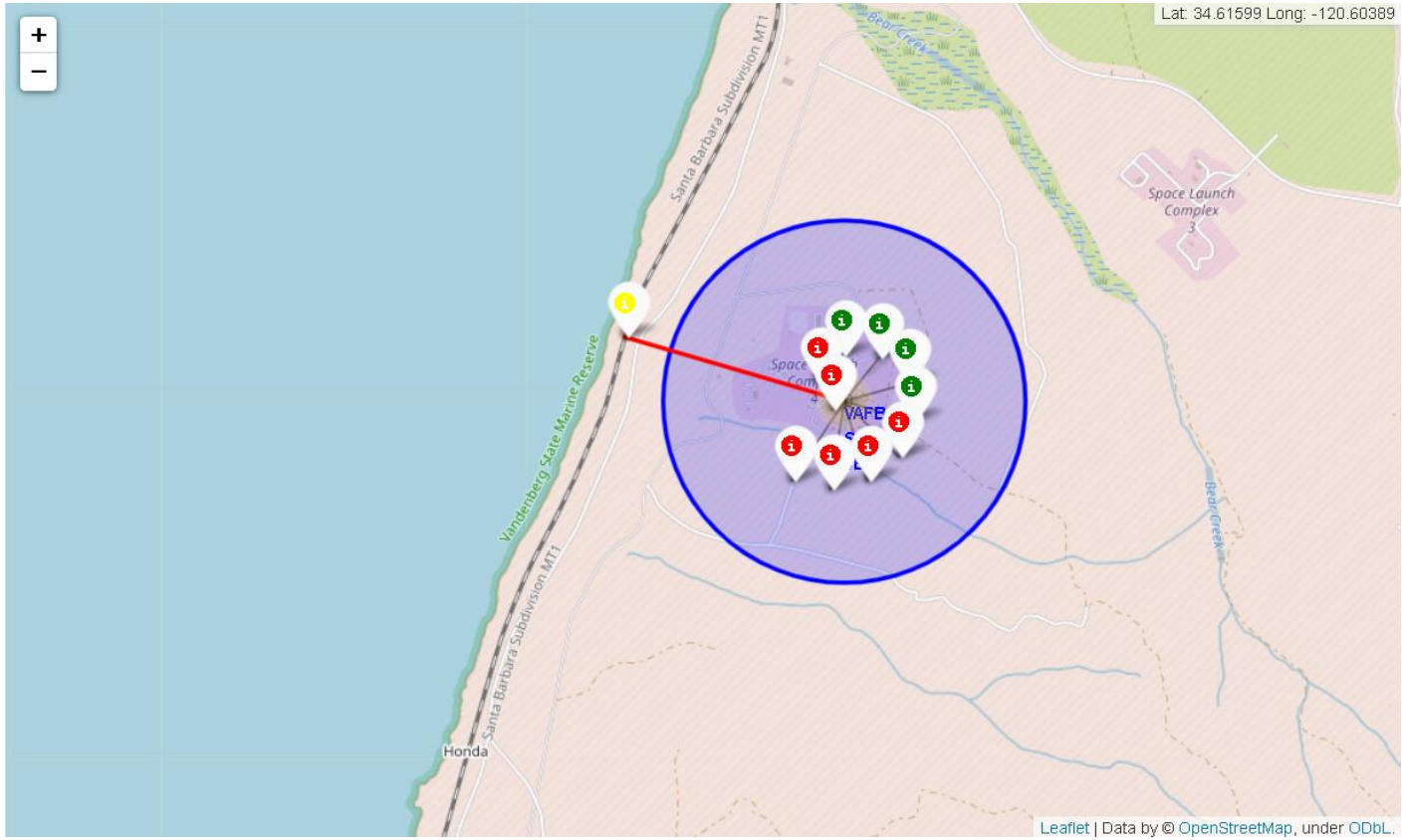
Success/Failure Marker

Here we have, of the three sites in Florida, there are two sites on the right and the other on the left. On the left, the success icon is green and the failure icon is red.



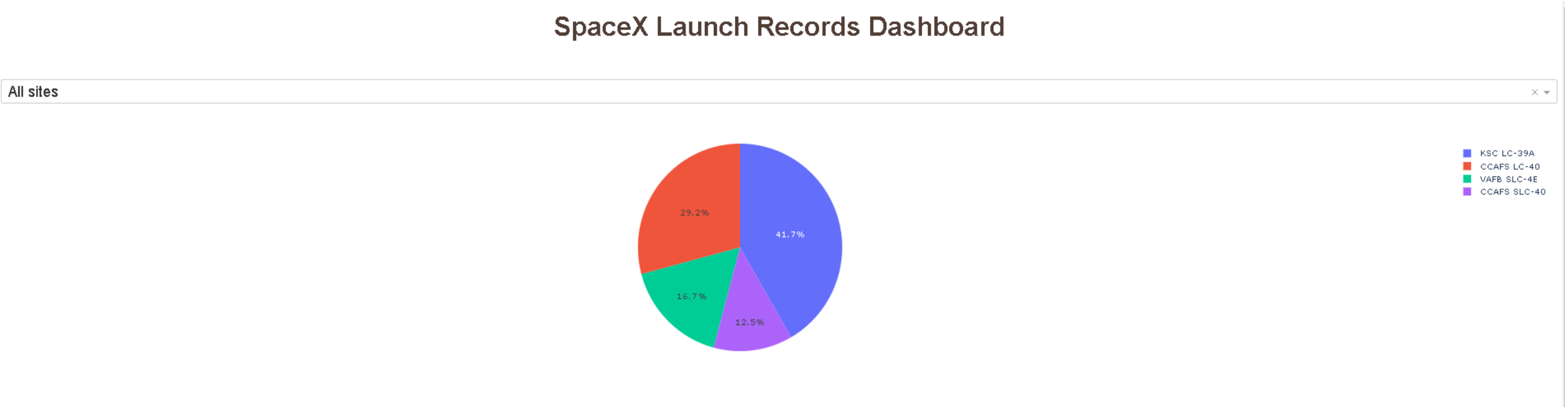
Launch Site Proximities

Here we see the proximity of the launching site to the coast. In addition, the trainline is marked with green and red colors.



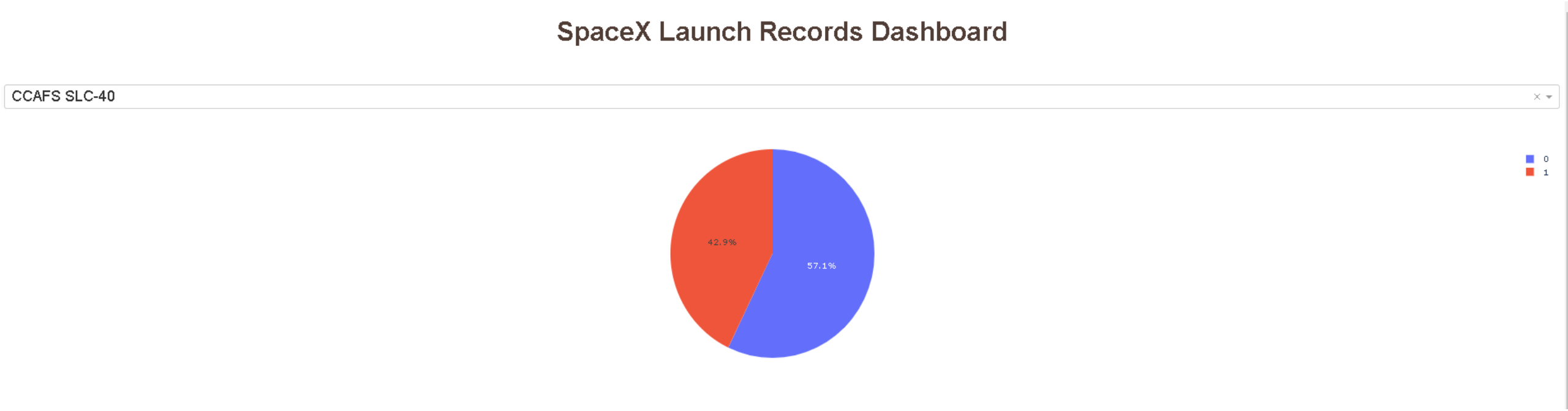
Build a Dashboard with Plotly Dash

Success Rate of All Sites



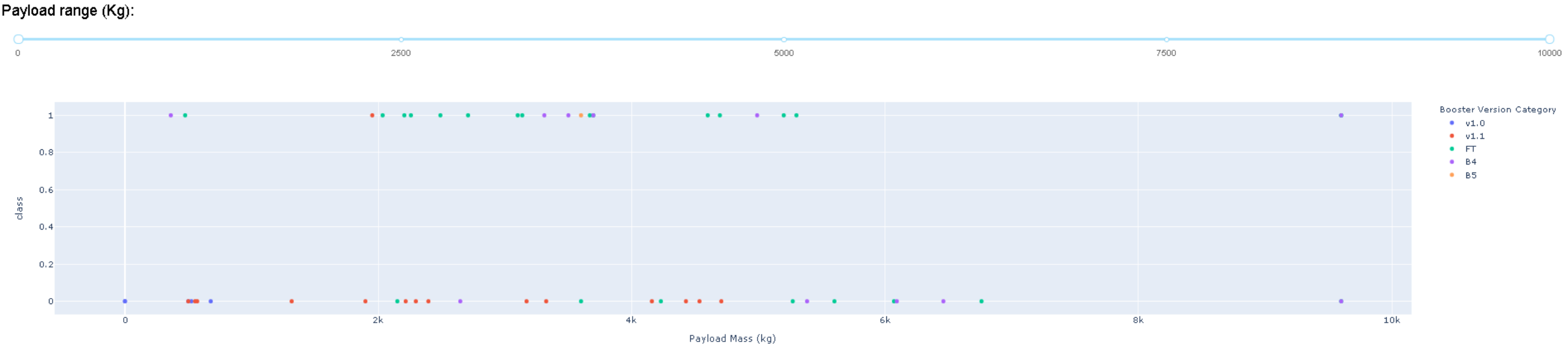
Here we can observe the different success rate for each launching site.

Highest Launch Success Rate (CCAFS SLC-40)



Here we can see that the success rate of this site is 42.9%

Payload vs. Launch Outcome scatter plot for all sites

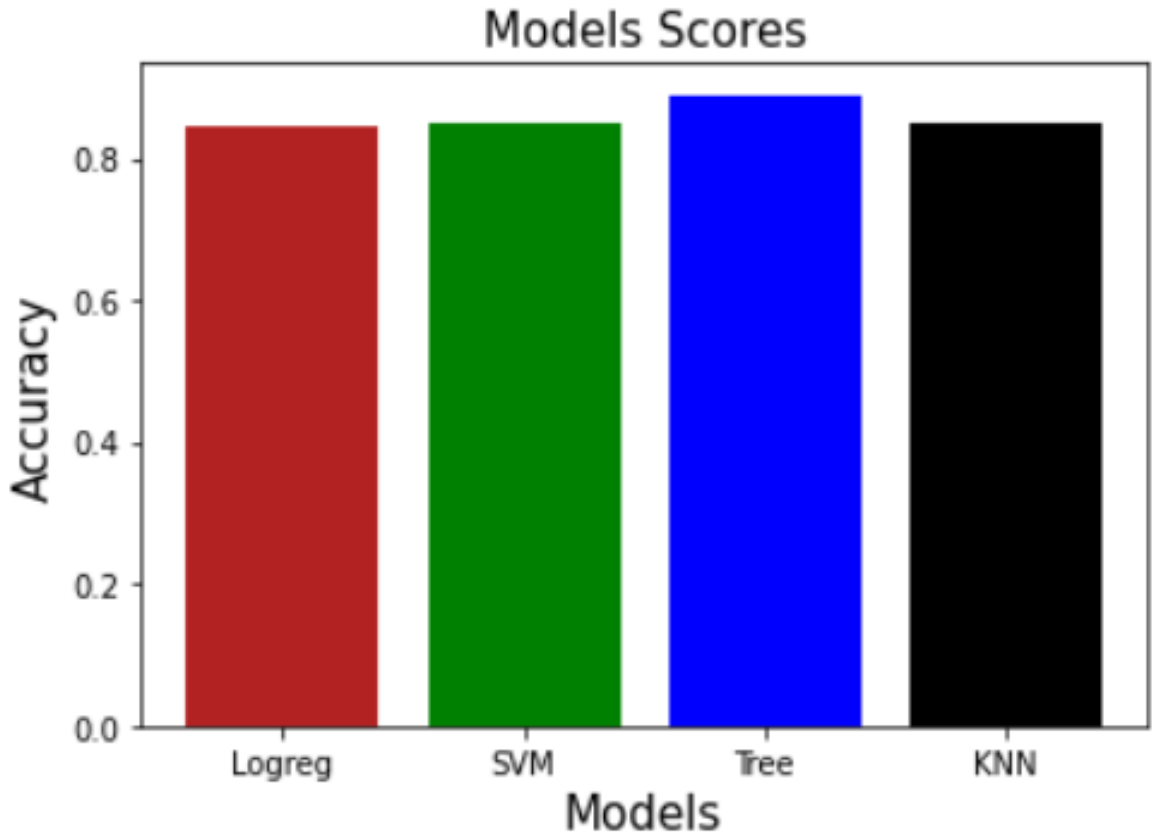


- Here we can see that the payload range with the best success rate is between 1.500 and 3.800 kg
- It is also clear that the FT Booster version is the best version

Predictive Analysis (Classification)

Classification Accuracy

The best model based on the accuracy is a **Decision Tree** with a score of **0.8892**

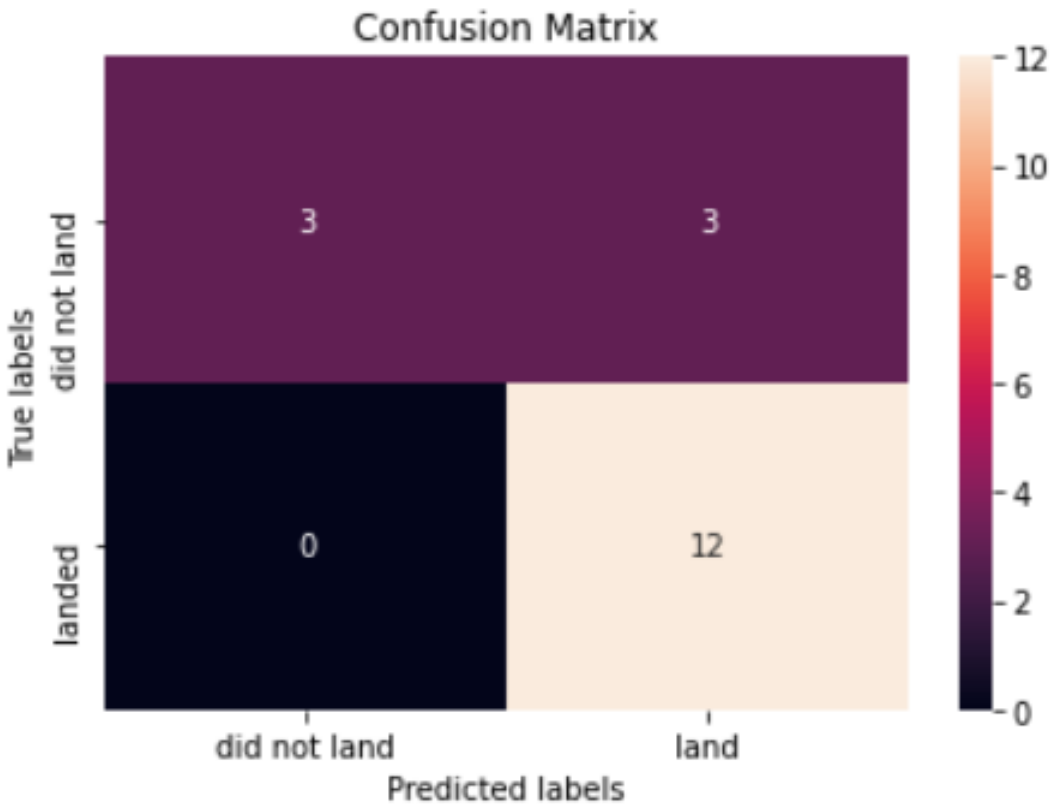


Confusion Matrix

❖ The confusion matrix of the Decision Tree

True/Predicted	Number
True Positive	12
False Negative	0
True Negative	3
False Positive	3

❖ The model is quite interesting as it predicts a lot of times the good labels, however 3 times it predicted the success of the mission and the mission failed. Reducing the amount of False Positive would be a good idea to avoid spending Millions and years of work. It could be done using Boosting or maybe look at a model with a lower accuracy but a better precision.



Conclusions

❖ There are many parameters when considering launching rockets in space

- The Booster version is definitely one of this essential parameter

- The Orbit, Payload Mass are also important

- Machine Learning models can really helps to understand if a mission will be a success or a failure as it will learn from data of all previous launches. As we saw a model is able to predict with a high accuracy the reliability of a mission.

❖ However more data would be useful to have better, I have no doubt that engineer and scientists use these data in their predictions

Appendix

All noteboeks, datasets are available in rmzturkmen/github.

Thanks to IBM and Coursera for this course!