SpaceX Race

Anmol Tripathi

28-May-2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Summary of methodologies
- - Data Collection through API
- - Data Collection with Web Scraping
- - Data Wrangling
- - Exploratory Data Analysis with SQL
- - Exploratory Data Analysis with Data Visualization
- - Interactive Visual Analytics with Folium

INTRODUCTION

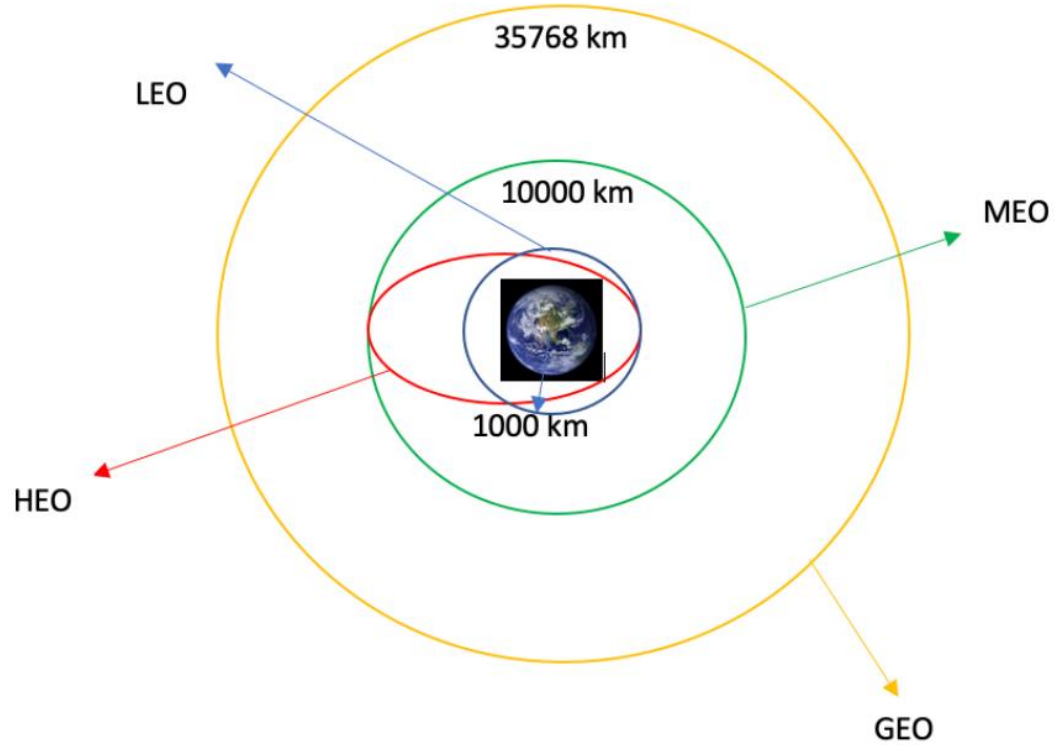
- SpaceX is a revolutionary company who has disrupt the space industry by offering a rocket launches specifically Falcon 9 as low as 62 million dollars; while other providers cost upward of 165 million dollar each. Most of this saving thanks to SpaceX astounding idea to reuse the first stage of the launch by re-land the rocket to be used on the next mission. Repeating this process will make the price down even further. As a data scientist of a startup rivaling SpaceX, the goal of this project is to create the machine learning pipeline to predict the landing outcome of the first stage in the future. This project is crucial in identifying the right price to bid against SpaceX for a rocket launch. The problems included: • Identifying all factors that influence the landing outcome. • The relationship between each variables and how it is affecting the outcome. • The best condition needed to increase the probability of successful landing.

METHODOLOGY

- Data collection methodology:
- • Data was collected using SpaceX REST API and web scrapping from Wikipedia
- • Perform data wrangling
- • Data was processed using one-hot encoding for categorical features
- • Perform exploratory data analysis (EDA) using visualization and SQL
- • Perform interactive visual analytics using Folium and Plotly Dash
- • Perform predictive analysis using classification models
- • How to build, tune, evaluate classification model

Data Wrangling

Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA). We will first calculate the number of launches on each site, then calculate the number and occurrence of mission outcome per orbit type.

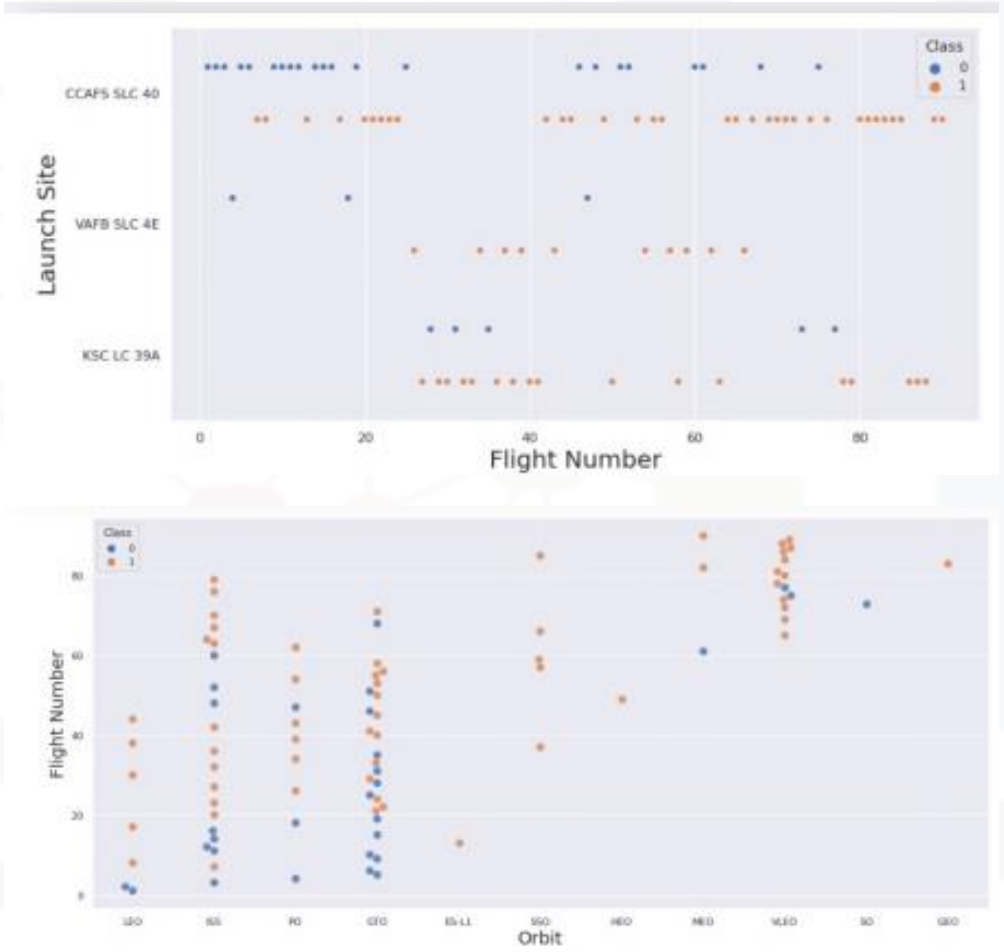


EDA with Data Visualization

We first started by using scatter graph to find the relationship

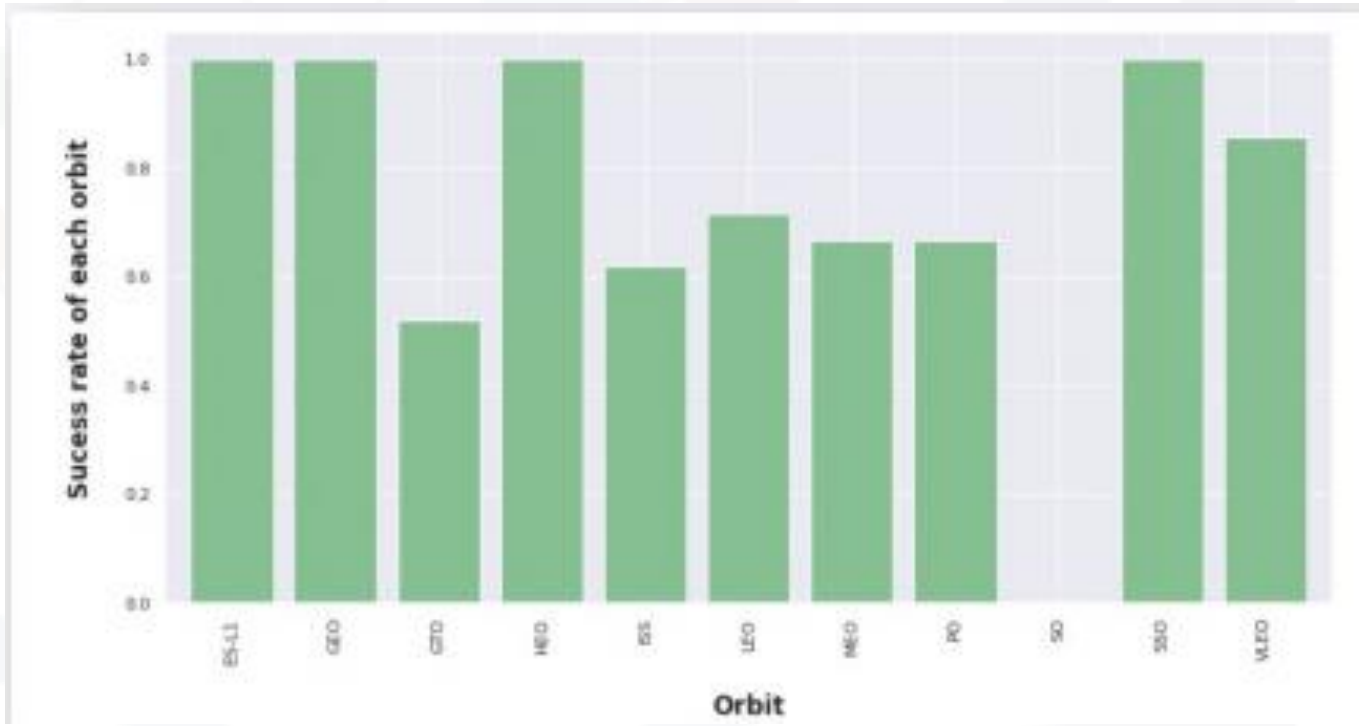
between the attributes such as between:

- Payload and Flight Number.
- Flight Number and Launch Site.
- Payload and Launch Site.
- Flight Number and Orbit Type.
- Payload and Orbit Type.



EDA with Data Visualization

Once we get a hint of the relationships using scatter plot. We will then use further visualization tools such as bar graph and line plots graph for further analysis. Bar graphs is one of the easiest way to interpret the relationship between the attributes. In this case, we will use the bar graph to determine which orbits have the highest probability of success.



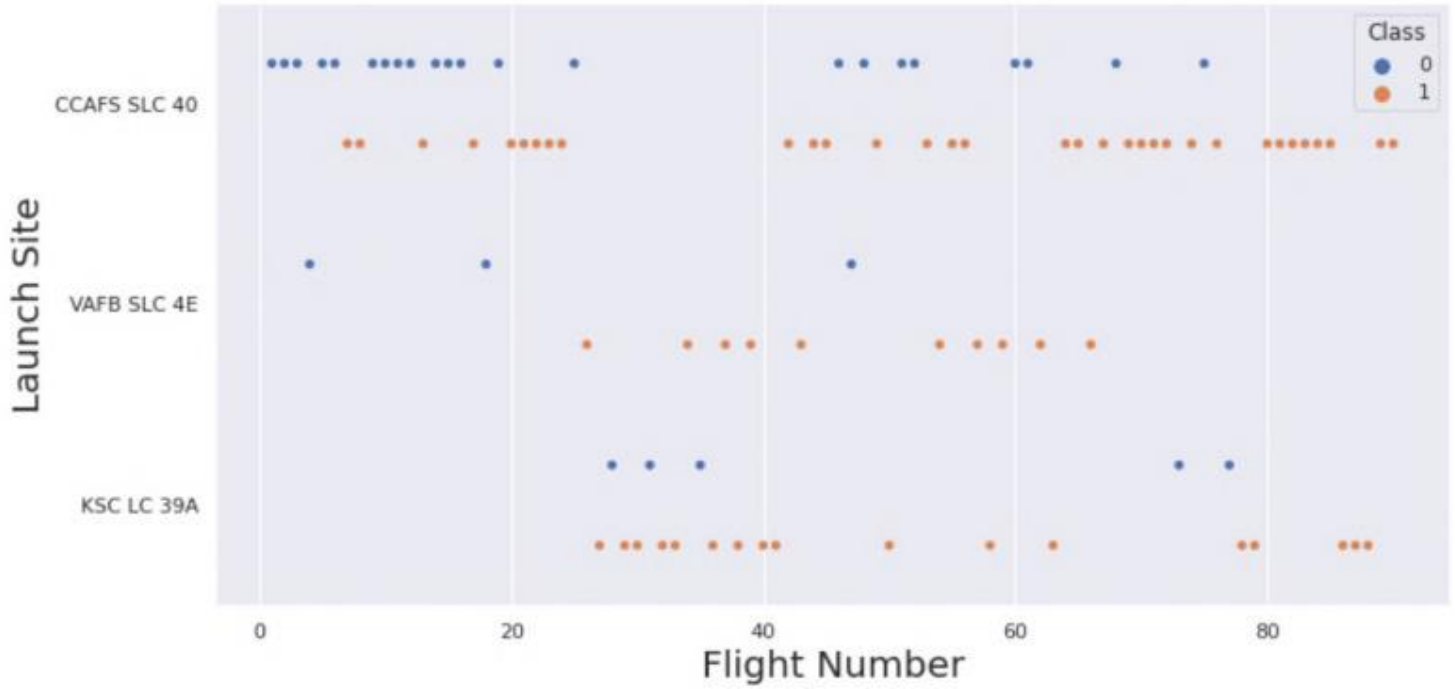
EDA with SQL

Using SQL, we had performed many queries to get better understanding of the dataset, Ex:

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

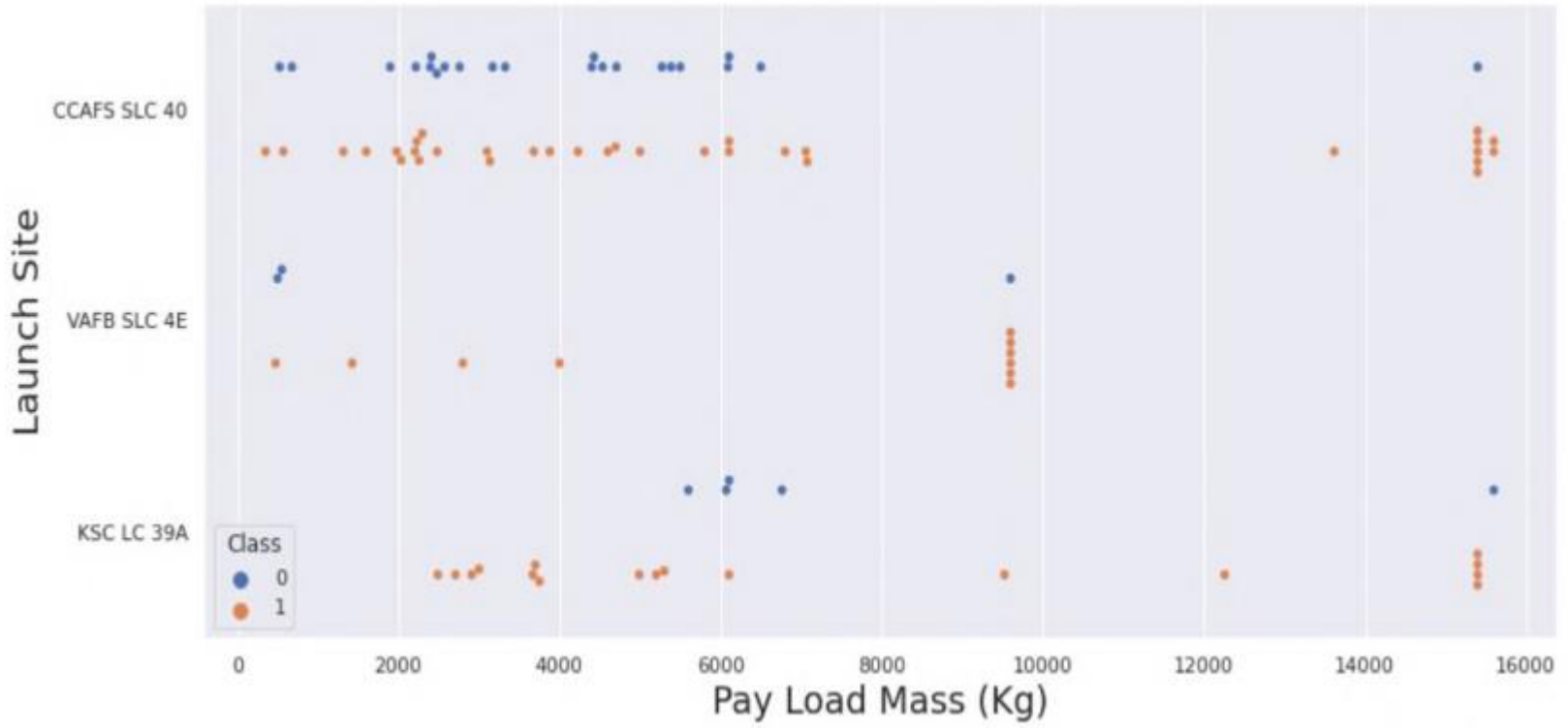
Flight Number vs. Launch Site

This scatter plot shows that the larger the flights amount of the launch site, the greater the the success rate will be.



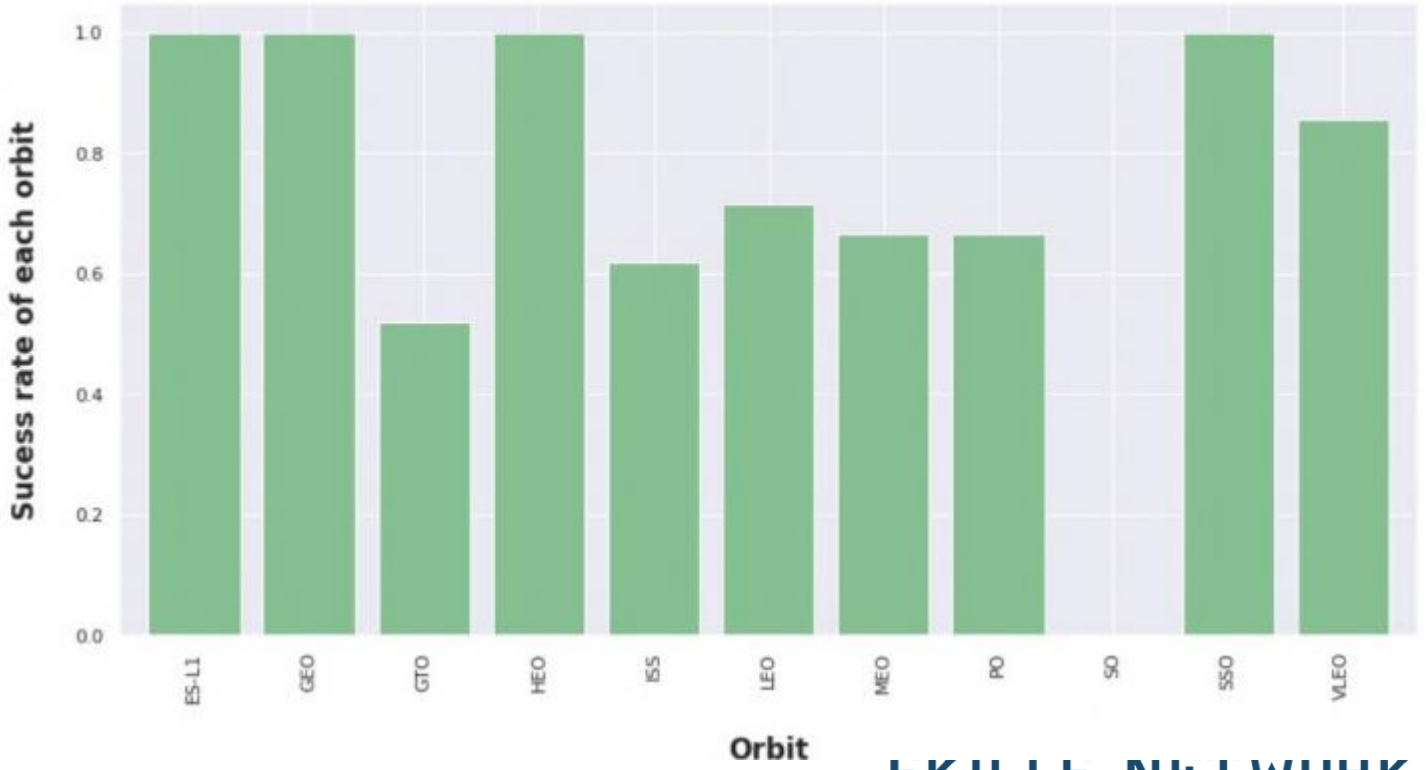
Payload vs. Launch Site

This scatter plot shows once the pay load mass is greater than 7000kg, the probability of the success rate will be high increased.



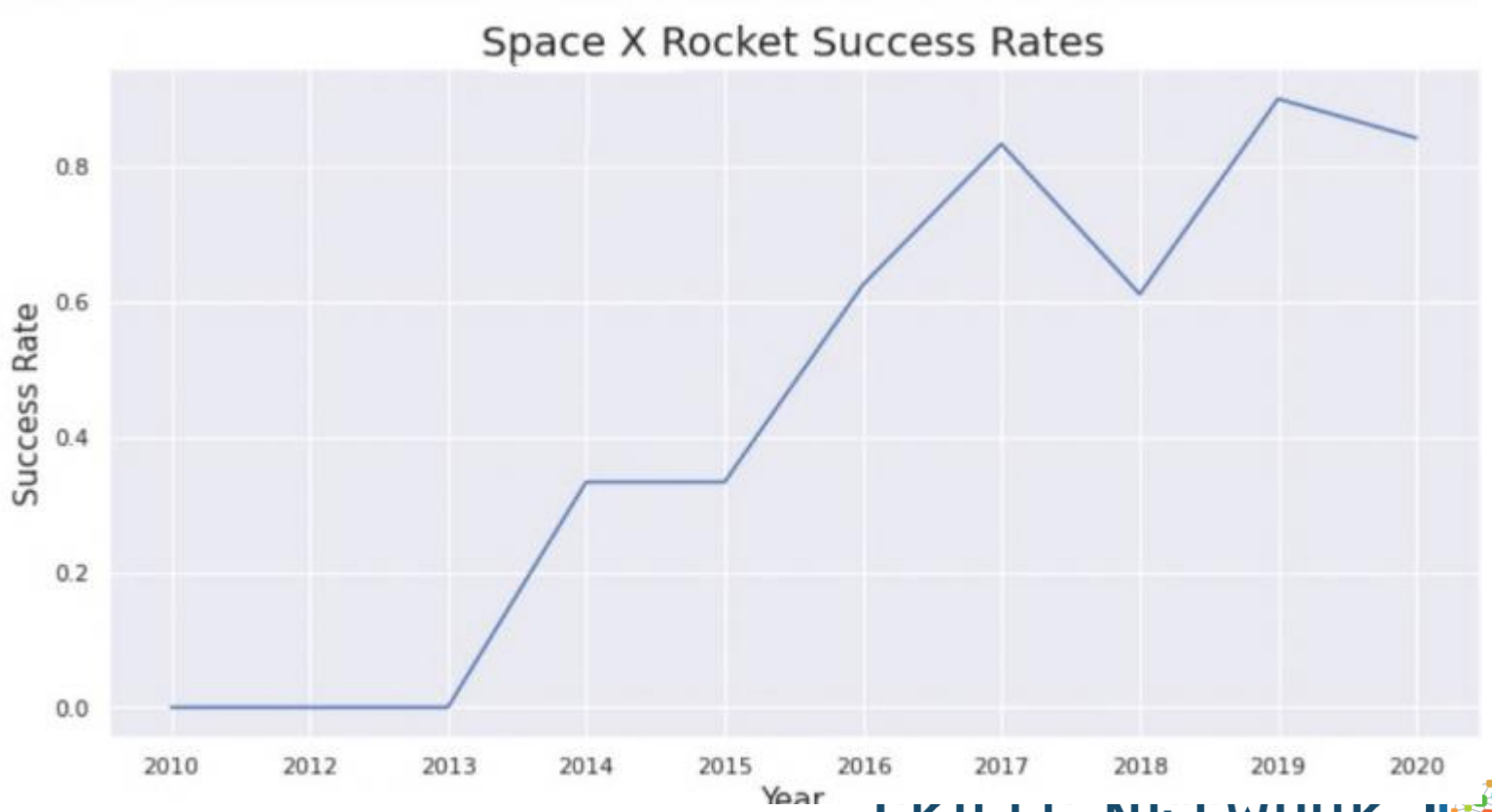
Success Rate vs. Orbit Type

This figure depicted the possibility of the orbits to influences the landing outcomes as some orbits has 100% success rate such as SSO, HEO, GEO AND ES-L1 while SO orbit produced 0% rate of success



Launch Success Yearly Trend

This figures clearly depicted and increasing trend from the year 2013 until 2020.



Build an Interactive Map with Folium

To visualize the launch data into an interactive map. We took the latitude and longitude coordinates at each launch site and added a circle marker around each launch site with a label of the name of the launch site.

We then assigned the dataframe launch_outcomes(failure,success) to classes 0 and 1 with Red and Green markers on the map in MarkerCluster().

We then used the Haversine's formula to calculated the distance of the launch sites to various landmark to find answer to the questions of:

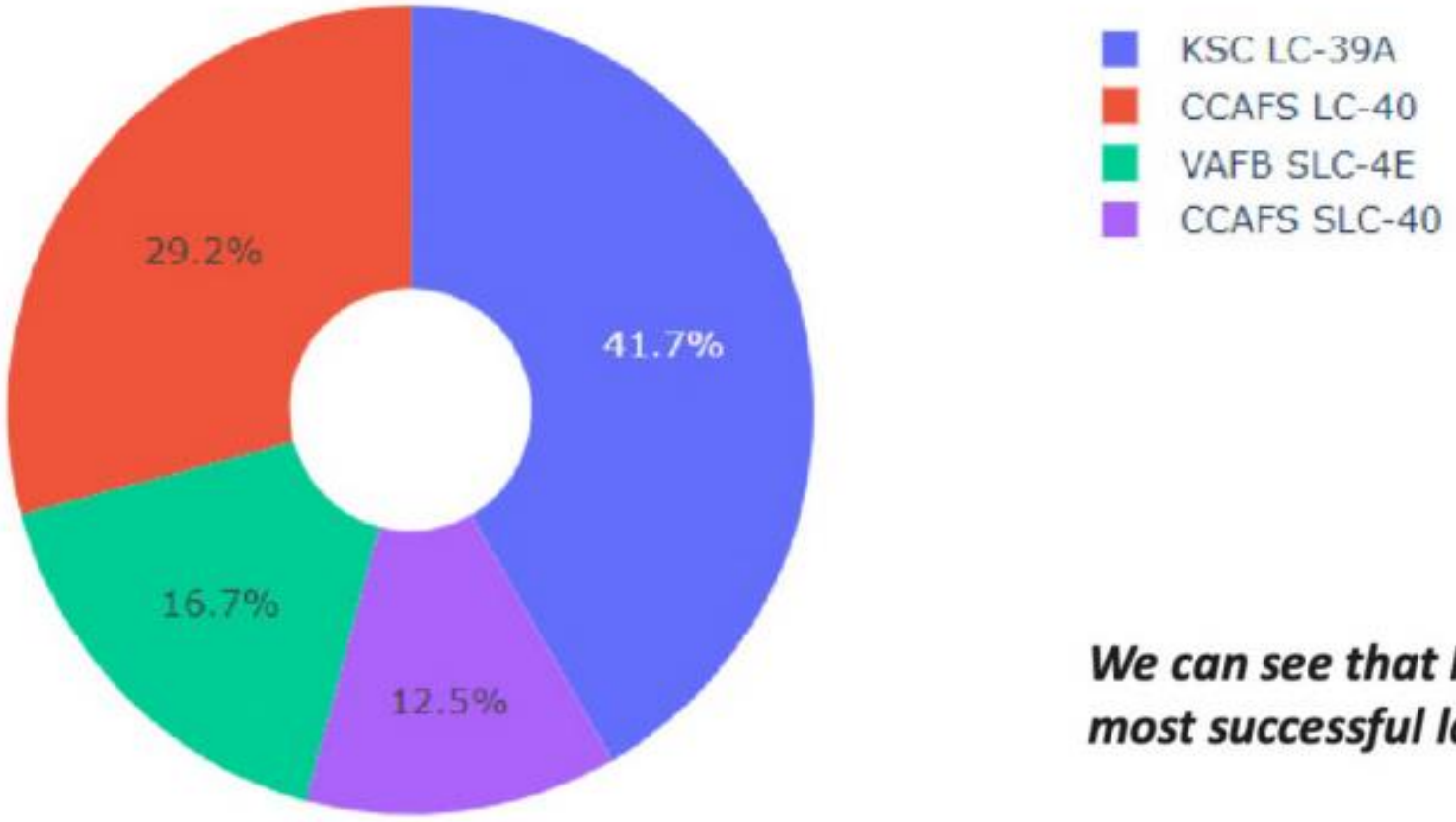
- How close the launch sites with railways, highways and coastlines?
- How close the launch sites with nearby cities?

DASHBOARD

We built an interactive dashboard with Plotly dash which allowing the user to play
around with the data as they need.

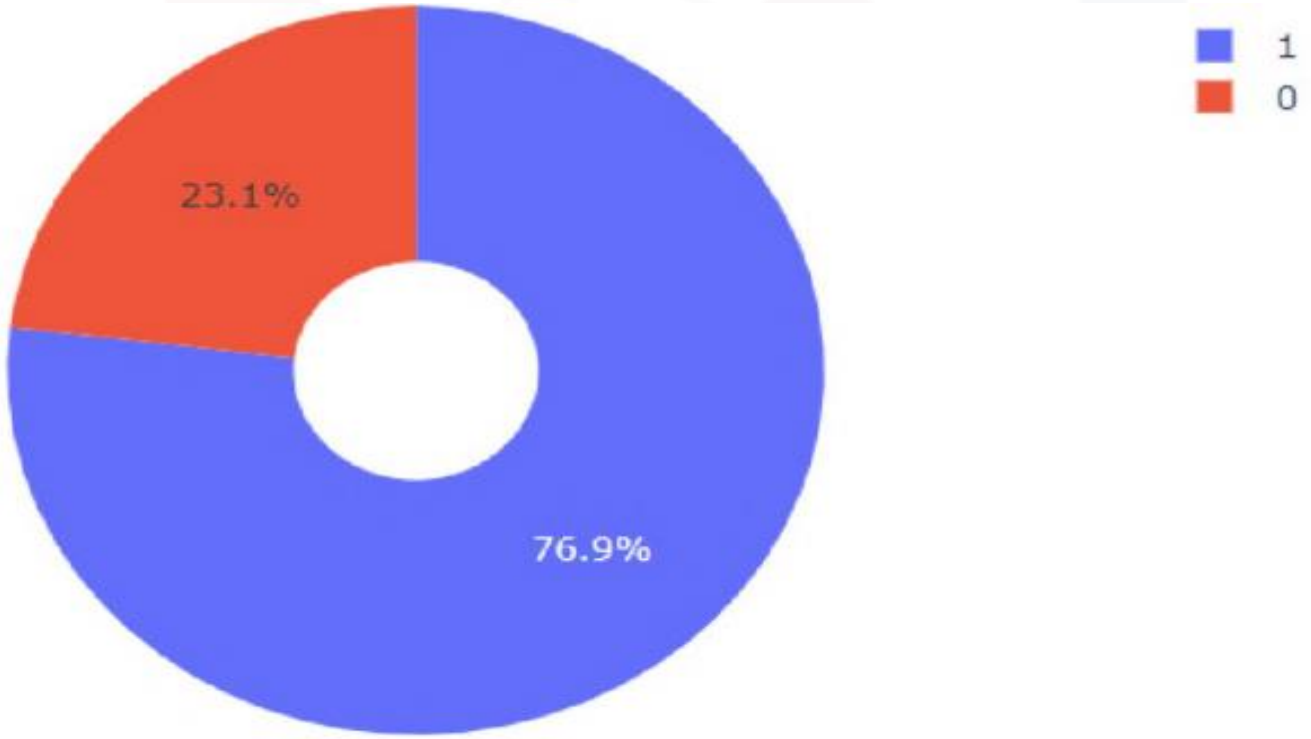
- We plotted pie charts showing the total launches by a certain sites.
- We then plotted scatter graph showing the relationship with Outcome and Payload
Mass (Kg) for the different booster version.

The success percentage by each site



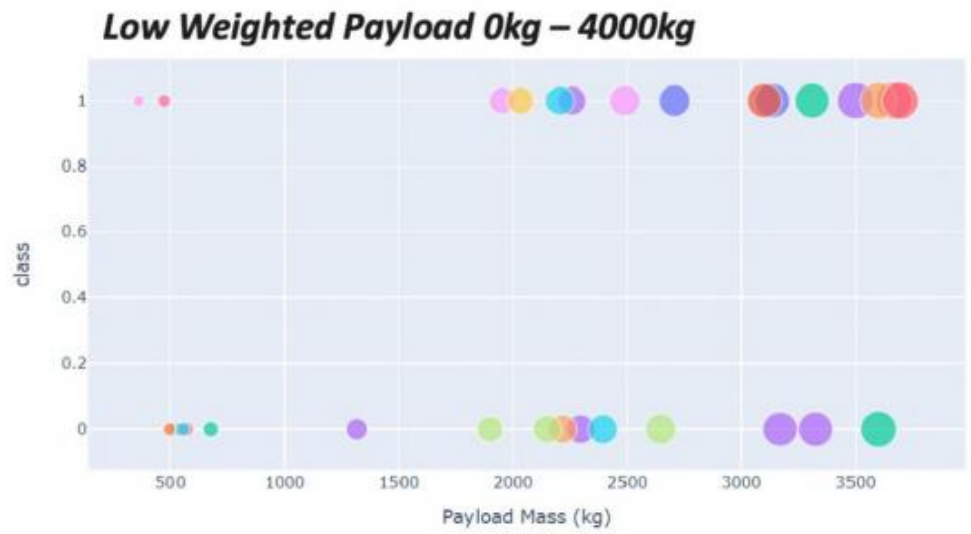
We can see that KSC LC-39A had the most successful launches from all the sites

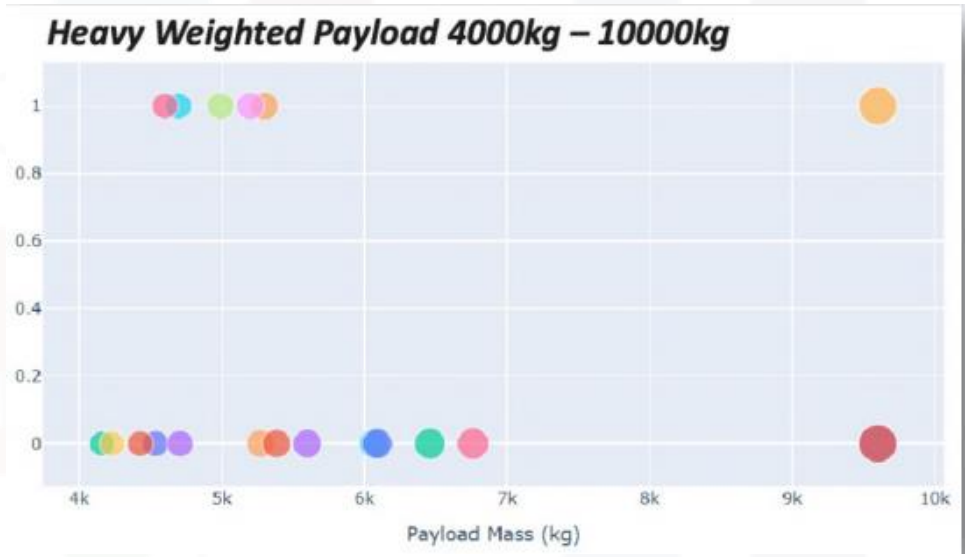
The highest launch-success ratio: KSC LC-39A



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

DASHBOARD TAB Payload vs Launch Outcome Scatter Plot 3





Classification Accuracy

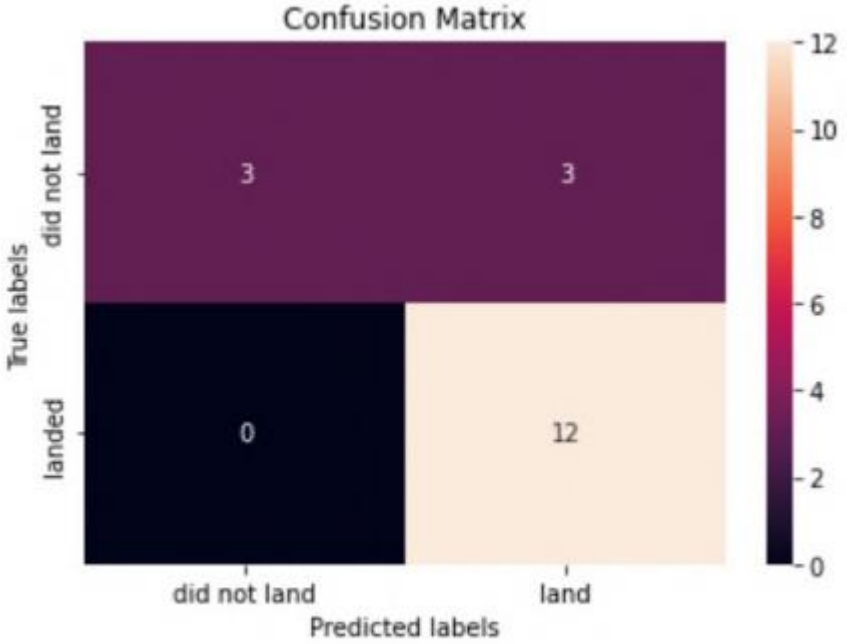
As we can see, by using the code as below: we could identify that the best algorithm to be the Tree Algorithm which have the highest classification accuracy

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.9017857142857142
Best Params is : {'criterion': 'entropy', 'max_depth': 10, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_sampl
es_split': 10, 'splitter': 'random'}
```

OVERALL FINDINGS & IMPLICATIONS

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier



CONCLUSION

- We can conclude that:
- • The Tree Classifier Algorithm is the best Machine Learning approach for this dataset.
- • The low weighted payloads (which define as 4000kg and below) performed better
- than the heavy weighted payloads.
- • Starting from the year 2013, the success rate for SpaceX launches is increased,
- directly proportional time in years to 2020, which it will eventually perfect the
- launches in the future.
- • KSC LC-39A have the most successful launches of any sites; 76.9%
- • SSO orbit have the most success rate; 100% and more than 1 occurrence.