IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

SUMMARY OF METHODOLOGIES

THE DATA WAS FIRST COLLECTED THROUGH TWO METHODS: API AND WEBSCRAPING, BOTH VIA PUBLIC SOURCES.

THEN THE DATA HAS BEEN ANALYSED VIA METHODS SUCH AS EDA, DATA WRANGLING AND INTERACTIVE VISUAL ANALYTICS TO GENERATE INSIGHTS, AS WELL AS MACHINE LEARNING.

THE PROCESS SHOWED DATA CAN BE COLLECTED IN DIFFERENT WAYS.

ALSO, NOT ALL MACHINE LEARNING ALGORITHMS ARE EQUALLY PERFORMANT, AS WE FOUND OUT THAT DECISION TREES WAS THE BEST PERFORMER TO PREDICT IF A ROCKET WOULD LAND SUCCESSFULLY OR NOT.

Introduction

- Project background and context

- The objective of the work was to find ways to predict the successful landing or not of SpaceX rockets, using ML instead of rocket science. We also generated further insights regarding the context surrounding SpaceX landings, such as maps of best location for launches for instance.

- All of this would help assess costs encountered by SpaceX for launches, to see if SpaceY could compete.

- Problems you want to find answers

- In the notebooks, we tried to figure out what algorithm was the most performant to predict the outcome of the rocket's landing.

Section 1

Methodology

Methodology

Data collection methodology:

Data was collected through the SpaceX public API and through webscraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches).

Perform data wrangling

After cleaning and analysis of the features, the data was labelled (successful or not).

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

After the data wrangling, the data was split into test and train sets. Four machine learning algorithms were then applied and compared, using classification methods.

Data Collection

SpaceX API v4: https://api.spacexdata.com/v4/rockets It contains data on the launch of rockets. The data was retrieved using a get request.

Data was also collected by using webscraping of this wikipedia link: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches.

Data Collection. - SpaceX API

Source: https://github.com/sina-ali/capstone_spaceX/blob/main/1_Data_collection_API.ipynb

Data retrieved through get request of the link

Filter the dataframe to only include Falcon 9 launches

Data wrangling: remove missing values

Data Collection - Wikipedia webscraping

Extract a Falcon 9 launch records HTML table from Wikipedia

Using beautifoulsoup and get requests

Extract all column/variable names From header

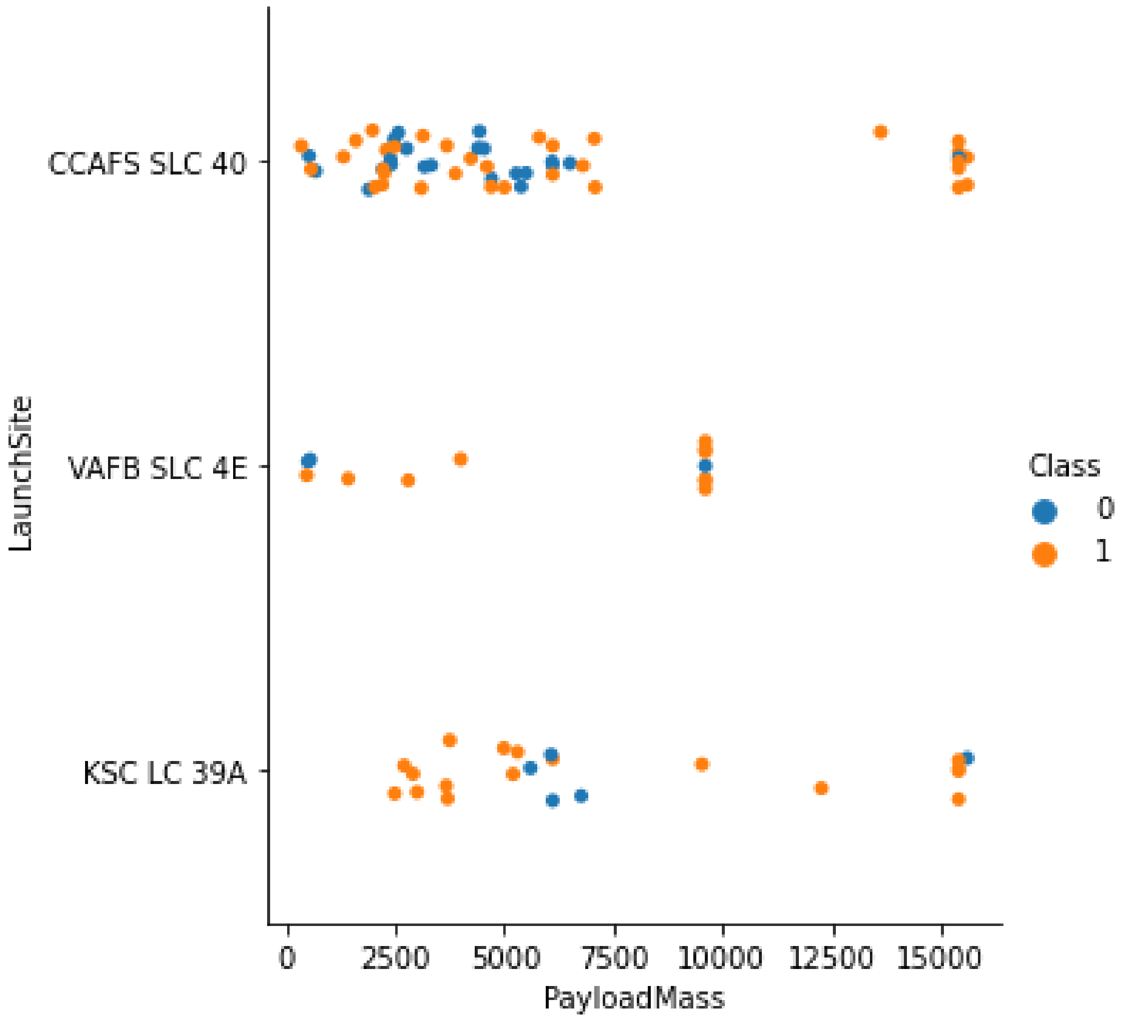
Create a data frame by parsing the launch HTML tables

Source: https://github.com/sina-ali/capstone_spaceX/blob/main/1B_Data_Collection_Webscraping.ipynb

Data Wrangling

- Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models

- https://github.com/sina-ali/capstone_spaceX/blob/main/1C_Daya_Wrangling.ipynb



EDA with Data Visualization

- The objective was to study the relationship between pairs of features, so barcharts and scatterplots were mainly used for this. For instance, this below is a scatterplot of the relationship between LaunchSite and PayloadMass.

- Source: https://github.com/sina-ali/capstone_spaceX/blob/main/2B_EDA_Visualisation_Lab.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Build an Interactive Map with Folium

On this Folium Interactive Map, markers show launch locations, circles show specific highlighted locations, lines show the distance between two coordinates, and finally, marker clusters show multiple markers at once, e.g. when multiple launch sites happen in one location.

These objects give us explanations of events happening according to their geographical location and help make the map interactive.

Source: https://github.com/sina-ali/capstone_spaceX/blob/main/3_Interactive_Visualisation_Folium.ipynb

Predictive Analysis (Classification)

- To classify our data, four ML models were built: logistic regression, SVM, decision tree, KNN. Their hyperparameters were then tuned with GridSearch to improve their individual accuracy score. In the end, their score were compared and decision trees was the best performer.

Data preparation + standardization + split into test/train → Building of models, and their confusion matrix and accuracy → Comparison of models' performance based on accuracy score

- https://github.com/sina-ali/capstone_spaceX/blob/main/4_Machine_Learning.ipynb

Results — EDA

Exploratory data analysis results:

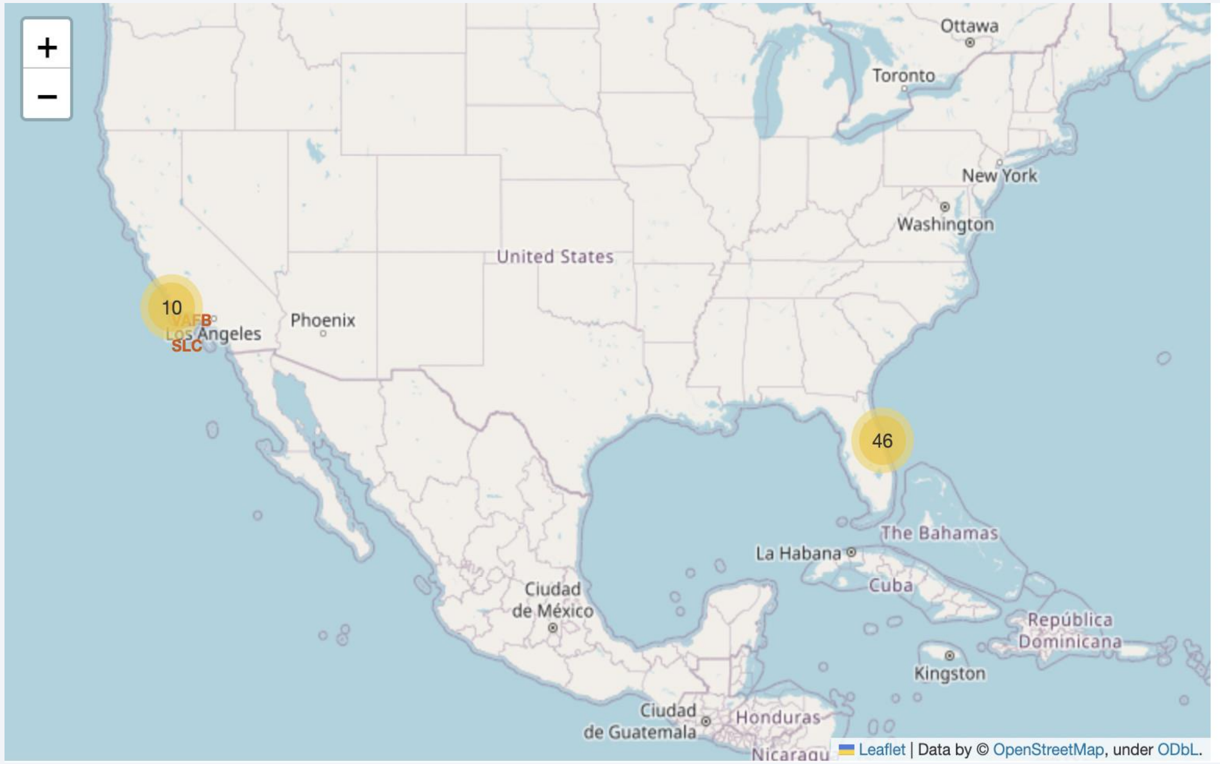
The first launch of SpaceX was in 2015.

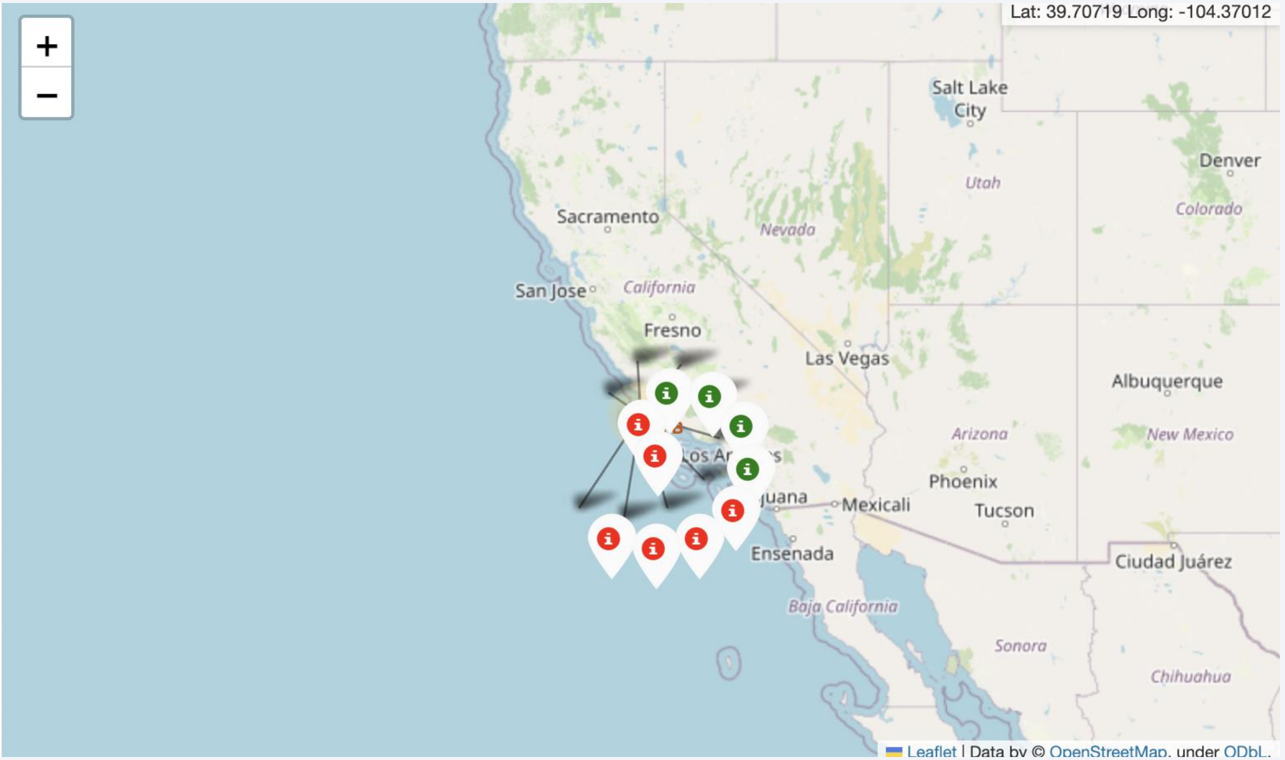
Almost all missions have successful outcomes: only two failures in 2015.

They are realized mainly near coastlines.

The average payload of f9 v1.1 is 2,298 kg.

Results – Interactive analytics





There are more launches on the East cost than the West coast (46 vs 10).

Multiples launches happen at the same location.

Results — Predictive analysis

- The classification applied on our data indicates that the best performer out of all four ML models was decision trees, with an accuracy score of 0.875.

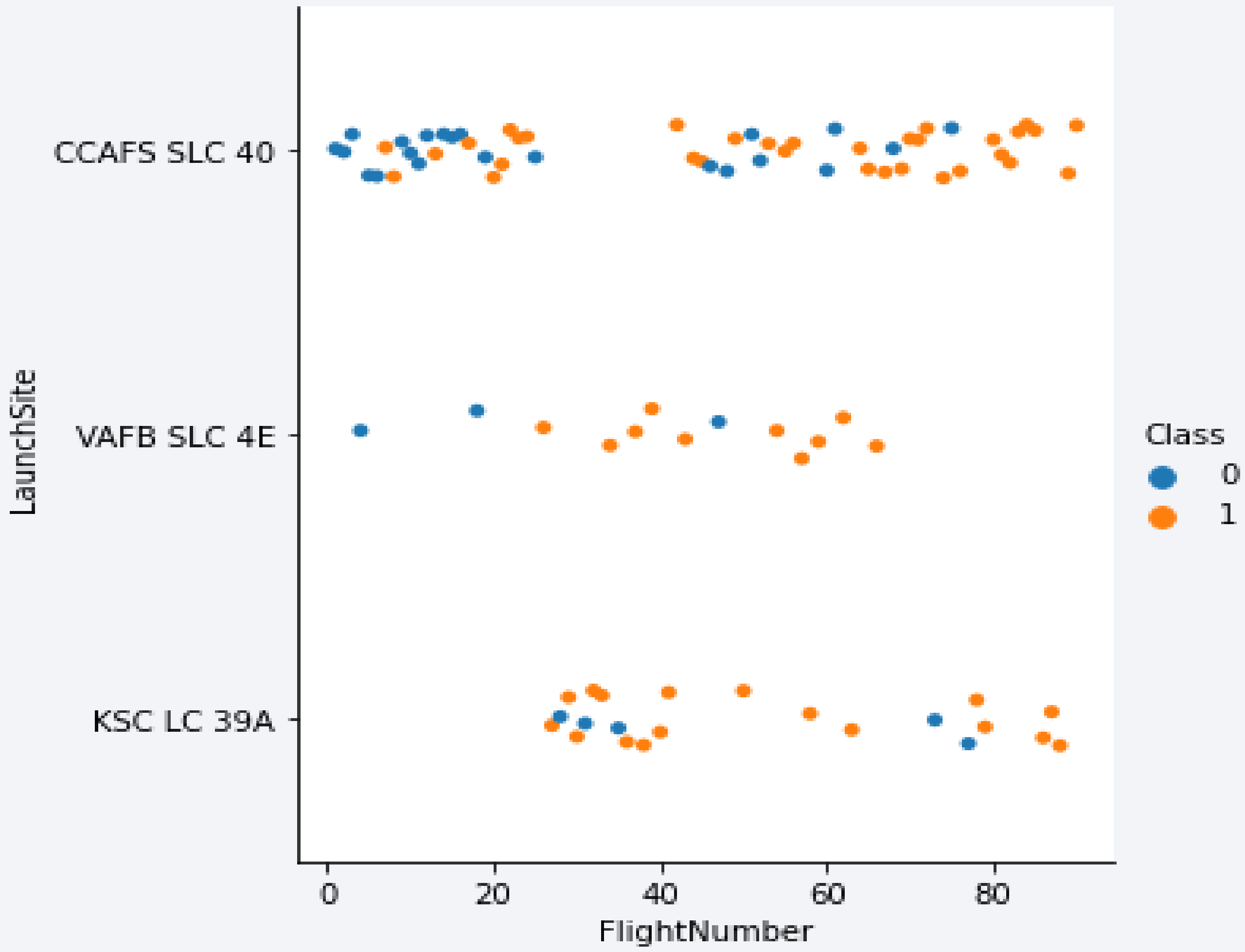
```
Logistic Regression Accuracy: 0.8464285714285713
SVM Accuracy: 0.8482142857142856
Decision Tree Accuracy: 0.875
KNN Accuracy: 0.8482142857142858
Decision Tree is the best performing method.
```

Section 2

Insights drawn from EDA

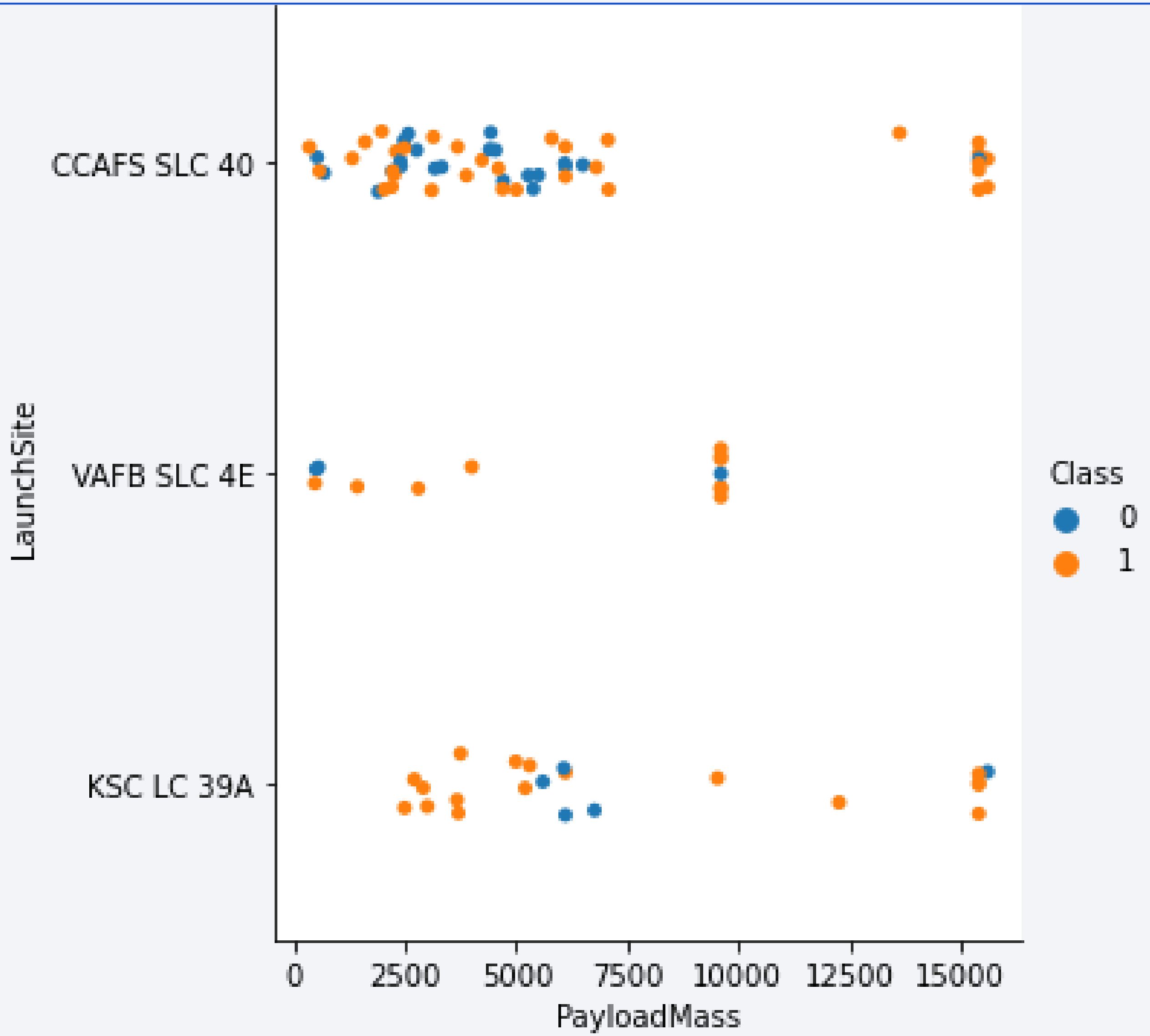
Flight Number vs. Launch Site

- This scatterplot shows that there is a correlation between FlightNumber and LaunchSite: the success rates increases with the number of launches made.



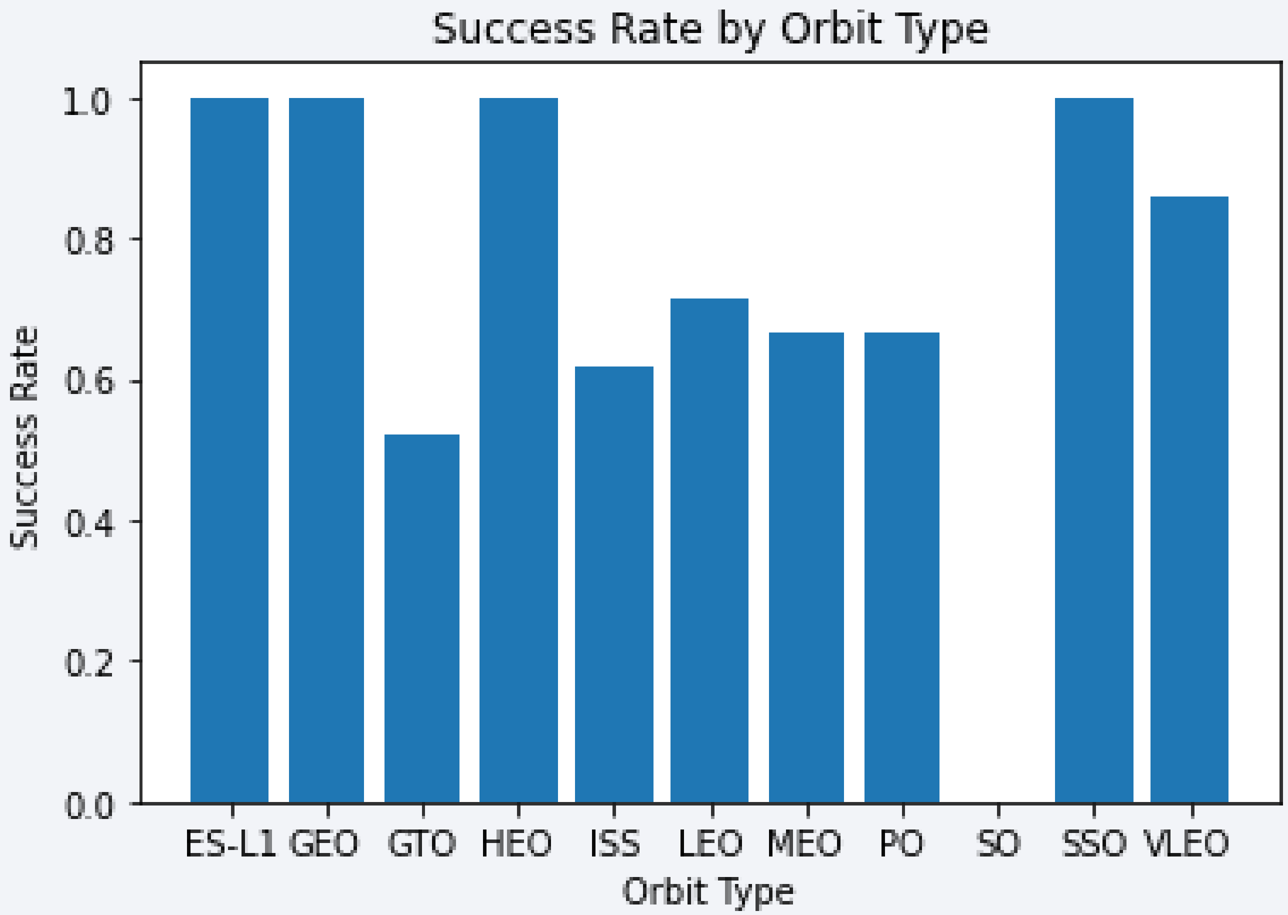
Payload vs. Launch Site

- It seems that the heavier the mass of the payload, the more success the launch will have in all three locations. It is a positive correlation.



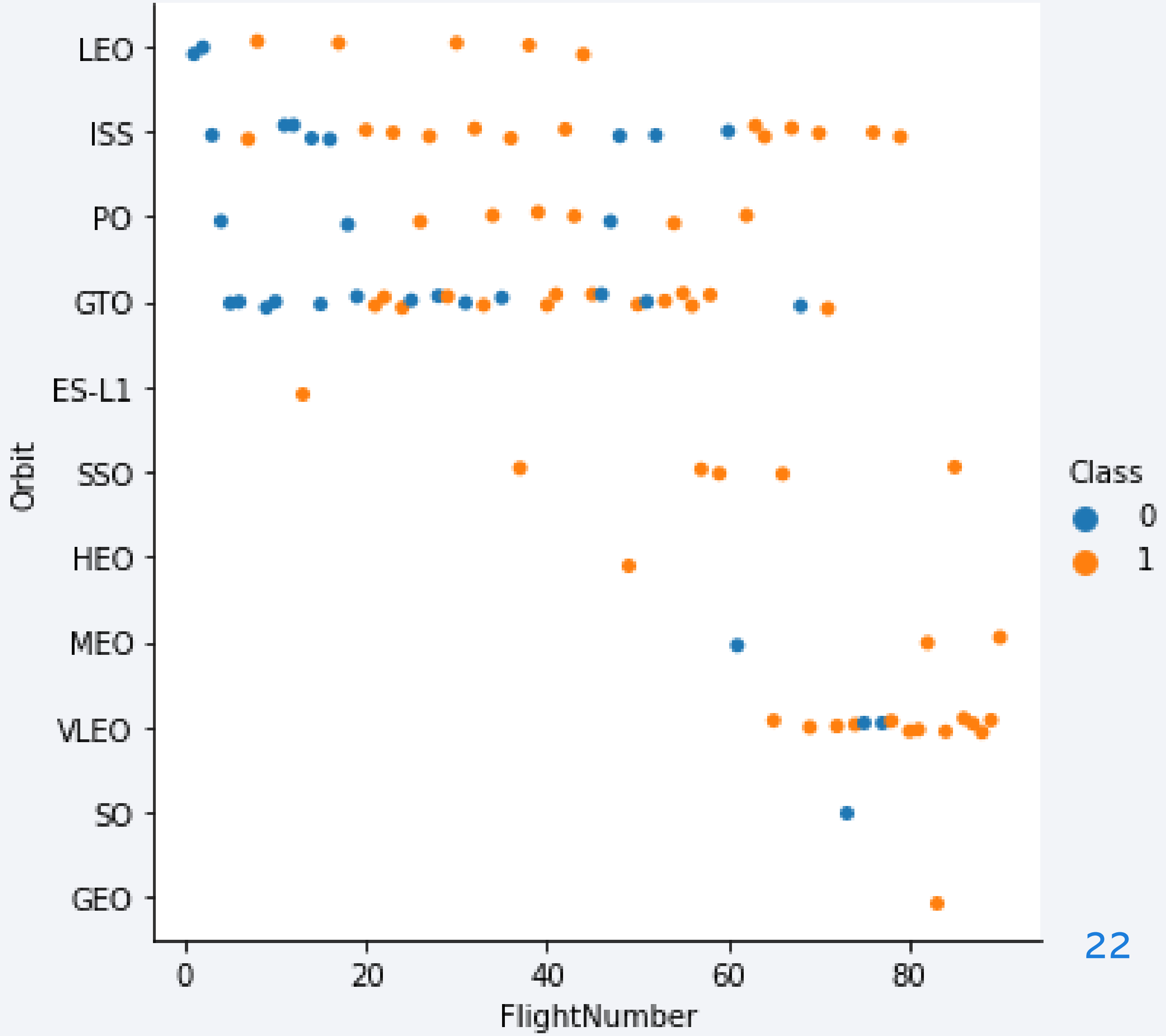
Success Rate vs. Orbit Type

- This barchart shows that the success rate will largely depend on the type of orbit, they're not equal: ES-L1, GEO, HEO, SSO and VLEO have the highest success rates.



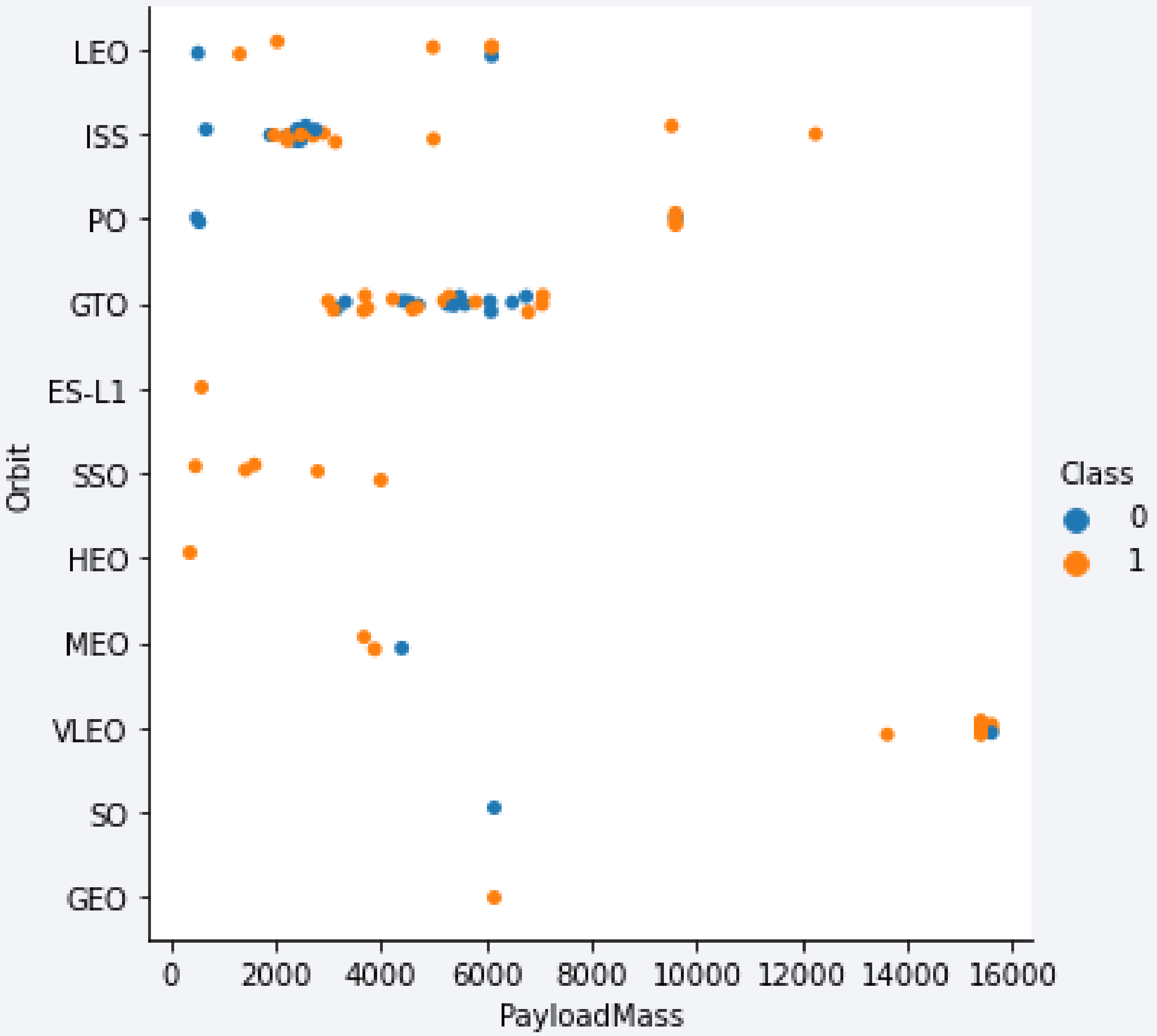
Flight Number vs. Orbit Type

- This scatterplot displays the fact that the more launches are made the more it will succeed and
  this, independently of the type of orbit.



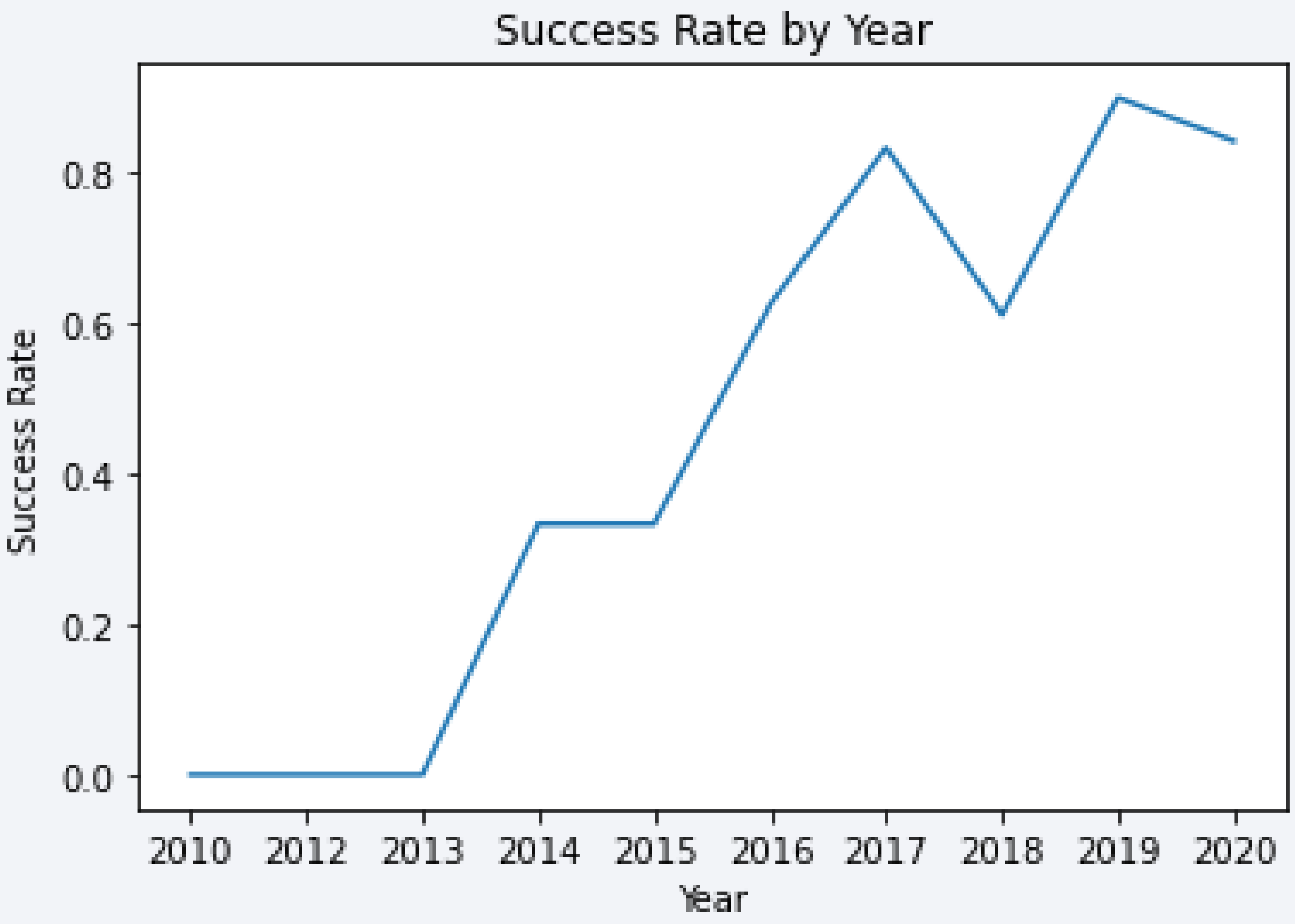
Payload vs. Orbit Type

- Once again, the heavier the payload the more chances of success the launch will have, except for GTO and MEO where the odds seem unaffected by the PayloadMass.
- For some orbits such as HEO, SO or GEO there are very few launches.



Launch Success Yearly Trend

- The success rate of the launches positively correlates with time: as the years go by, there is more success.
- There has been a dip in 2019.

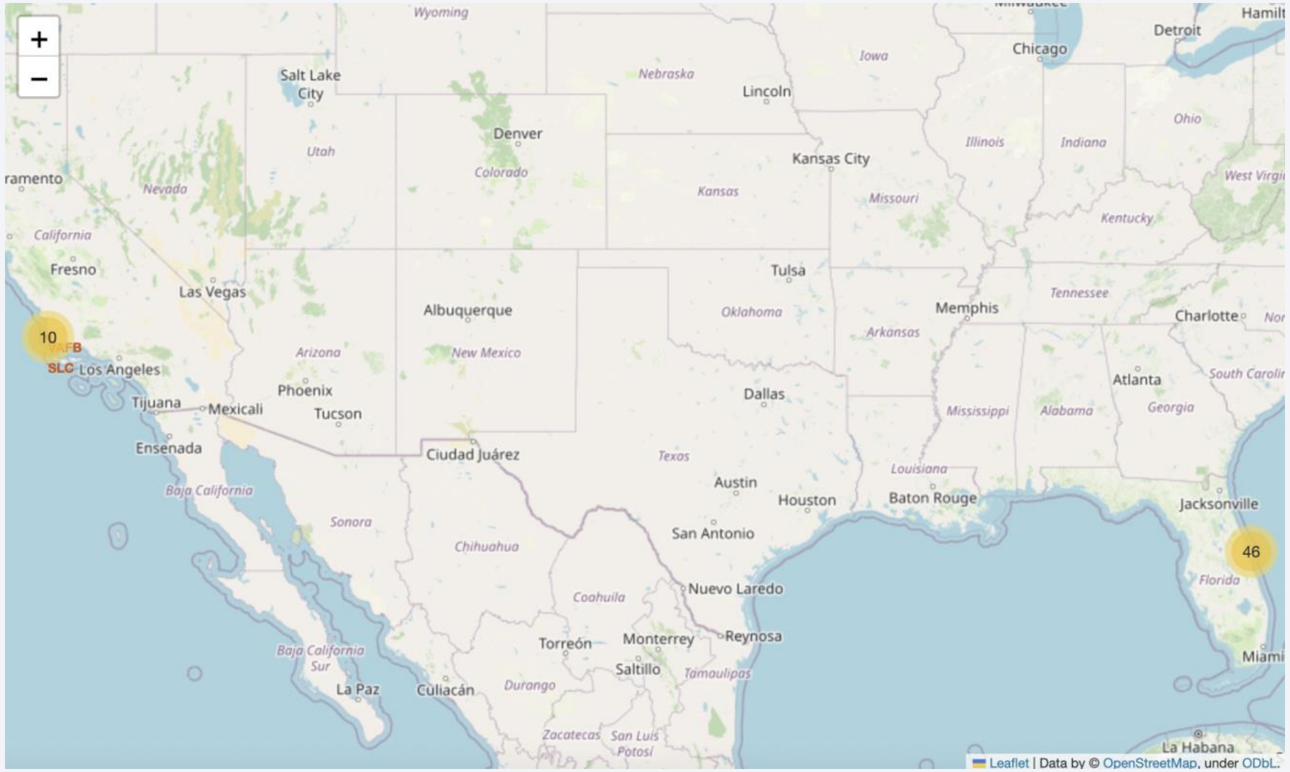


Section 3

Launch Sites Proximities Analysis

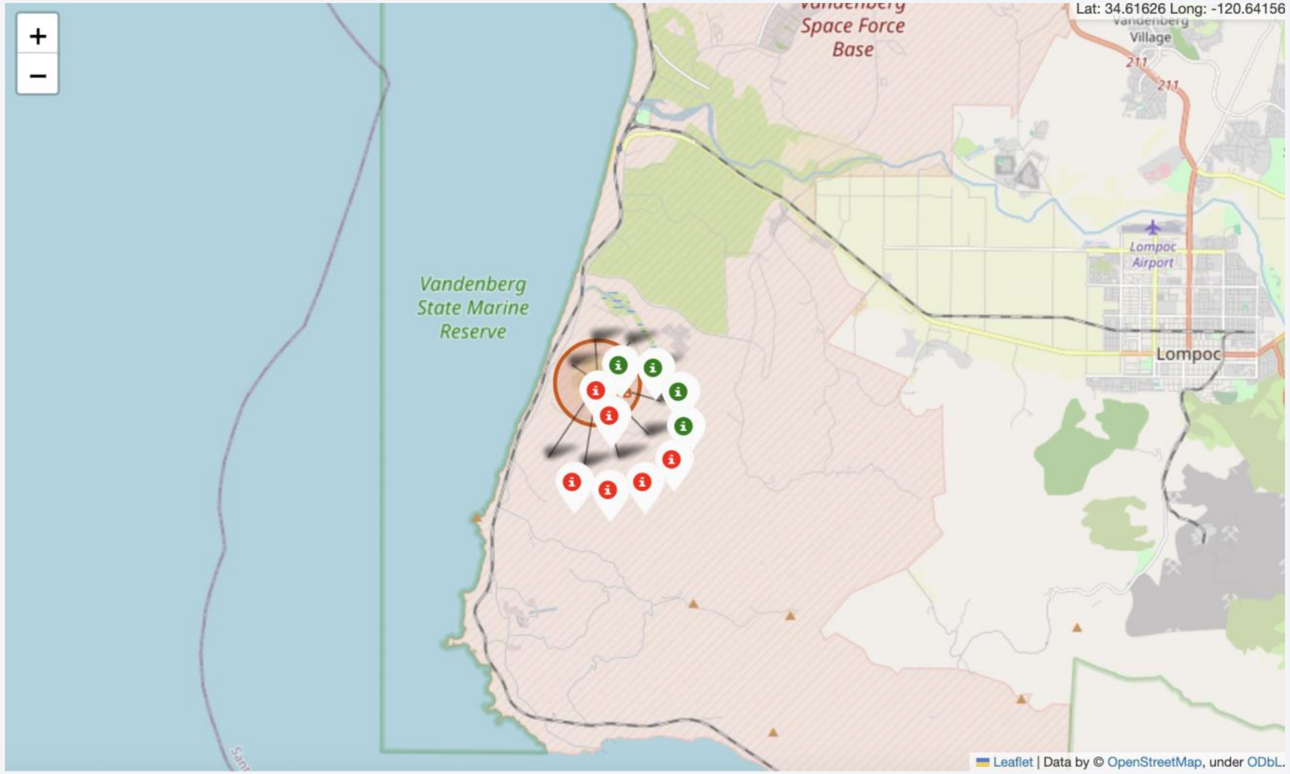
Map: Launch sites - markers

- This interactive map shows all of SpaceX's launch sites' locations.
- There are 10 locations on the west coast and 46 on the east coast of the United States, as shown by the markers.
- They're all near the sea, for safety.



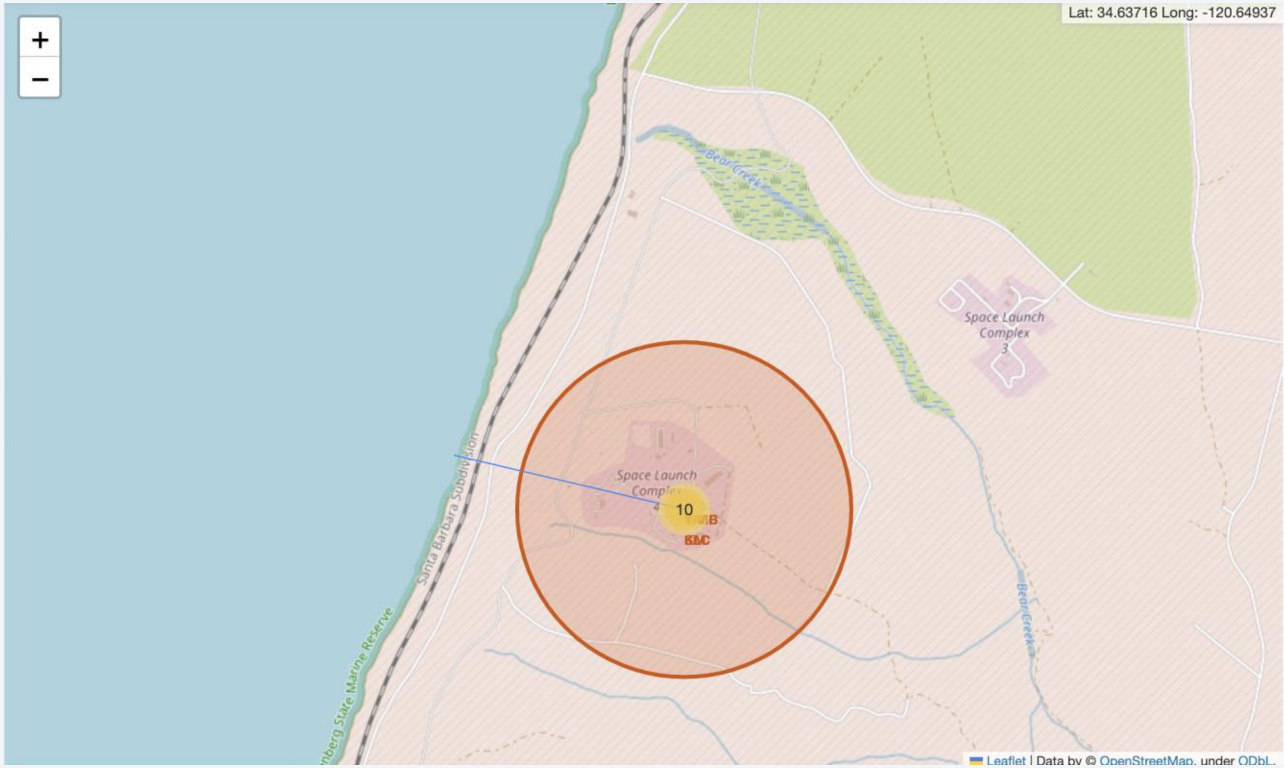
Map: Outcome of launches - labels

- The red and green labels on this map respectively represent the unsuccessful and successful launches sites.
- In this particular case, there were more failures than successes.



Map: Distance from other locations - lines

- Here, the point of the code was to generate a line that would display the distance between a launch site and a particular location, in this case its distance from the coastline.
- This is useful to visualize distances between coordinates.



Section 5

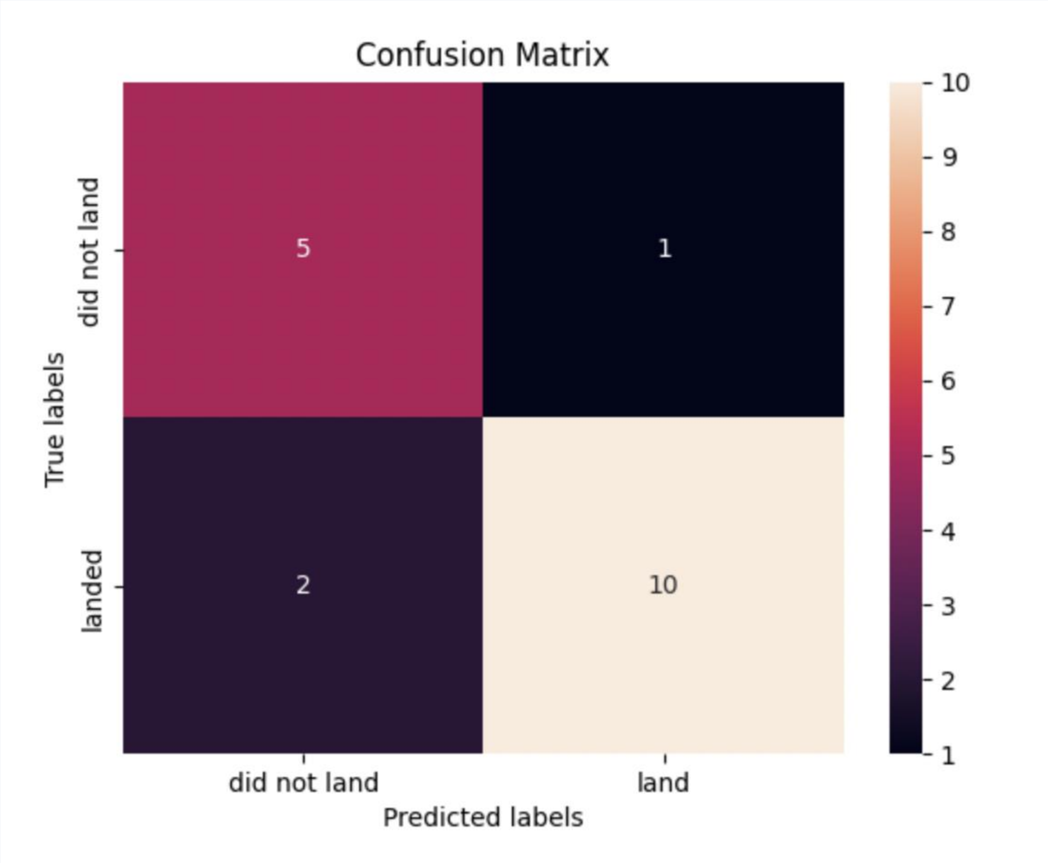
Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart
- Find which model has the highest classification accuracy

Confusion Matrix

- This image represents the confusion matrix of the best performing model, decision trees.
- It has only one false positive and 2 false negatives, which are really low numbers.
- Moreover, there is a significant number of true positives and negatives.



Conclusions

- To choose an orbit, it is crucial to estimate what factors we deem to be the most essential, since the success rates of orbits varies based on different features.
- KSC LC 39A is the best location to launch from, except when the payload is too high.
- Maps are the most efficient way to visualize which factors make a launch location strategic, to then decide which ones to focus the strategy on.
- Decision trees is the trusted model of choice to predict the chances of success for futures launches.

Thank you!