APPLIED DATA SCIENCE CAPSTONE PROJECT: RACE TO SPACE

OM SINGH

7/09/2021

TABLE OF CONTENT

EXECUTIVE SUMMARY

Summary of Methodologies:

- Data Collection
- Data Wrangling
- EDA (Exploratory Data Analysis) with SQL and Data Visualization
- Visually Analyzing Data with Folium
- Building a Dashboard to present data with Plotly Dash
- Predictive Analysis

Summary of Results:

- Success rate has increased over time and is currently at its highest.
- We were able to determine factors that affected the success rate of the launch
- Prediction of Flacon 9 rocket's first stage landing is 83.333%
- Proximities of launch sites to other areas can affect the success rate of launch

INTRODUCTION

Context and Background Information:

Space X is a company at the forefront of mission to Space in our modern age. We have been hired by clients at a company, Space Y that wishes to compete with Space Y. Space X is mainly successful due to their ability to reuse the first stage of their Falcon 9 Rockets, which we will be using our data science techniques to explore

Question to be Answered:

The Problems being answered in the report are as follows. We first determine the price of each launch. We will also gather information through the Space X API and determine whether Space X will reuse the first stage of the Falcon 9 rockets.

METHODOLOGY

DATA COLLECTION

- The Data collection we undergo requires both Web scraping Wikipedia and collecting Data from the Space X API.
- We collect data from both these sources for a more detailed analysis and to get a more complete picture of the information that we are working with.
- For example, while we get Longitude and Latitude of launches from the Space X REST Api, we don’t get this from Wikipedia, while Booster Landing is only available from Wikipedia.
- We append the data together into a single data frame that we export to a CSV, which will be worked upon later.

DATA COLLECTION- SPACE X API

- To collect data we will mainly be utilizing the Space X API.
- We then use the data collected through this API in a json format and transform it into a Pandas Data frame.
- This data is now cleansed, to get rid of irrelevant information such as data about Falcon 1 rockets and to to appropriately handing missing values within the data.
- Through multiple requests, we collected information pertaining to Booster Version, Launch sites of the falcon 9 rockets, Payload Data and Core Data
- GitHub URL: New-Capstone/DATA COLLECTION API LAB.ipynb at main · Oms27ct/New-Capstone (github.com)

SPACE X API DATA COLLECTION FLOW CHART

Request Rocket Launch Data from Space X Api

Convert the resulting JSON to a data frame

Request Extra Information through specific calls

Add this data into the data frame

Filter data frame to include only Falcon 9 launches

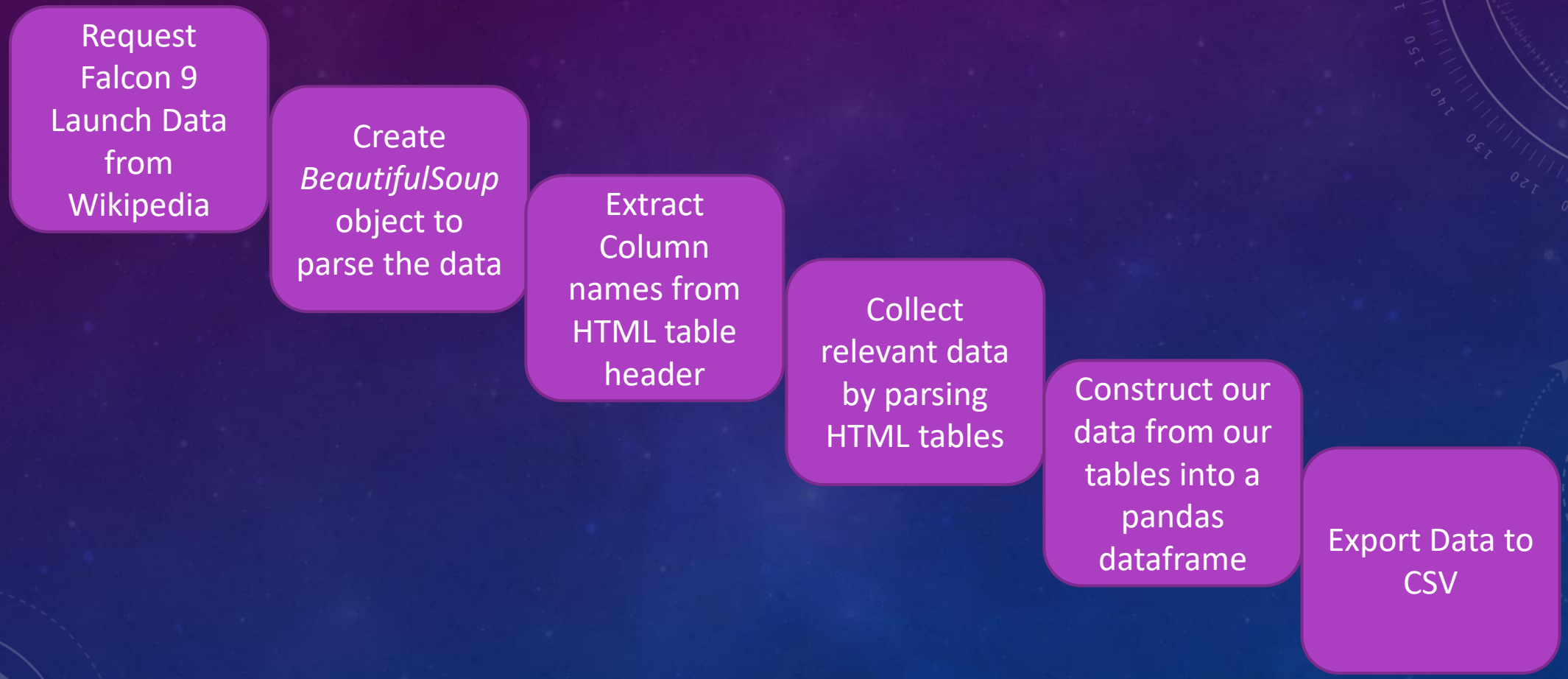
Appropriately deal with missing values

Export Data to CSV

DATA COLLECTION- WEB SCRAPING

- We use the *Beautiful Soup* package within Python to effective scrape data from Wikipedia tables from List of Falcon 9 and Falcon Heavy launches – Wikipedia.
- From This webpage, we collect Falcon 9 historical launch records and store them within a Pandas data frame format.
- GitHub URL: New-Capstone/Data Collection with webscraping.ipynb at main · Oms27ct/New-Capstone (github.com)

WEB SCRAPING FLOWCHART



DATA WRANGLING

- Here we check and appropriately applied methods [such as finding the average of values] to replace missing values of the data
- We identified the different data types of different variables, and fixed them to what they should be and:
- We created a variable class, which assigned values where class = 0 [in case of mission failure], or class =1 [in case of success]
- We also computed the number of launches on each site and the different type of orbital missions, and their occurrence.
- GitHub URL: New-Capstone/Data Wrangling.ipynb at main · Oms27ct/New-Capstone (github.com)

EDA WITH DATA VISUALIZATION

Here we used:

- Scatter charts were also used to identify the relationships between different variables such as:
  - Flight Number Vs Payload Mass
  - Launch Site Vs Flight Number
  - Payload Vs Launch Site
  - Flight Number Vs Orbit Type
  - Payload Vs Orbit Type
- A Line Plot was also used to get the average success trend over time of the launches.
- GitHub URL: New-Capstone/EDA WITH VISUALISATION.ipynb at main · Oms27ct/New-Capstone (github.com)

EDA WITH SQL

- Numerous SQL queries were performed. Through these we found out:
  - Names of unique launch sites
  - Launch sites that begin with the string ‘CCA’
  - Total Payload Mass carried by boosters launched by ‘NASA’
  - Average Payload Mass carried by Boosters in Falcon 9 rockets
  - Date of First Successful landing, and number of successful and failure missions
  - Names of successful boosters with payload mass between 4000 and 6000 kilograms
  - Booster versions and launch sites with failed landing outcomes in drone ship
  - The different landing outcomes between 04/06/2010 and 20/03/2017
- GitHub URL: New-Capstone/Exploratory Data Analysis (1).ipynb at main · Oms27ct/New-Capstone (github.com)

INTERACTIVE MAP WITH FOLIUM

- Here we created a map Displaying the different launch sites and their relative locations to areas such as the coast etc.
- We highlighted all the different launch sites and used colour coded markers for successful and failed launches
- We also calculated the distances each of these launch sites were from different locations such as the coastline to see if there were any trends between successful launches and the location where they are performed.
- GitHub URL: New-Capstone/Data visualisation with folium.ipynb at main · Oms27ct/New-Capstone (github.com)

DASHBOARD WITH PLOTLY DASH

- We added a dropdown menu for the different launch sites
- A pie chart was presented which returned the success rate for the different launch site
- This helped us identify the best launch sites
- We also added a scatterplot between Payload Mass and the outcome of each mission for the different sites for each booster version. This helped to identify relationships between Payload Mass, Booster Version Category and the landing outcome.

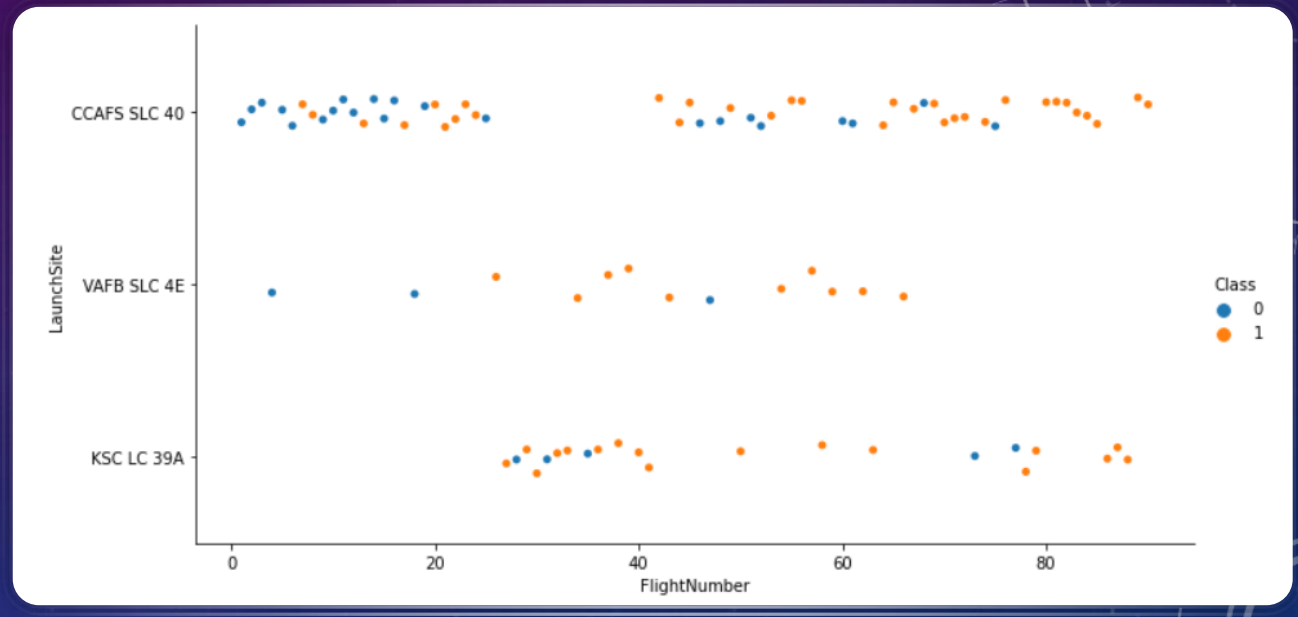
PREDICTIVE ANALYSIS(CLASSIFICATION)

- First we splitted the data into training and test sets
- Here we used different data analysis methods such as Logistic regression, SVM (Support Vector Machines), Decision trees and KNN (K Nearest Neighbours)
- We figured out the model with the best parameters using the GridSearchCV method, and we then fitted the best model.
- Then we tested the accuracy of the model and from the model with the best parameters, drew a confusion matrix.
- GitHub URL: New-Capstone/Machine Learning Prediction Lab.ipynb at main · Oms27ct/New-Capstone (github.com)

INSIGHTS DRAWN FROM EDA

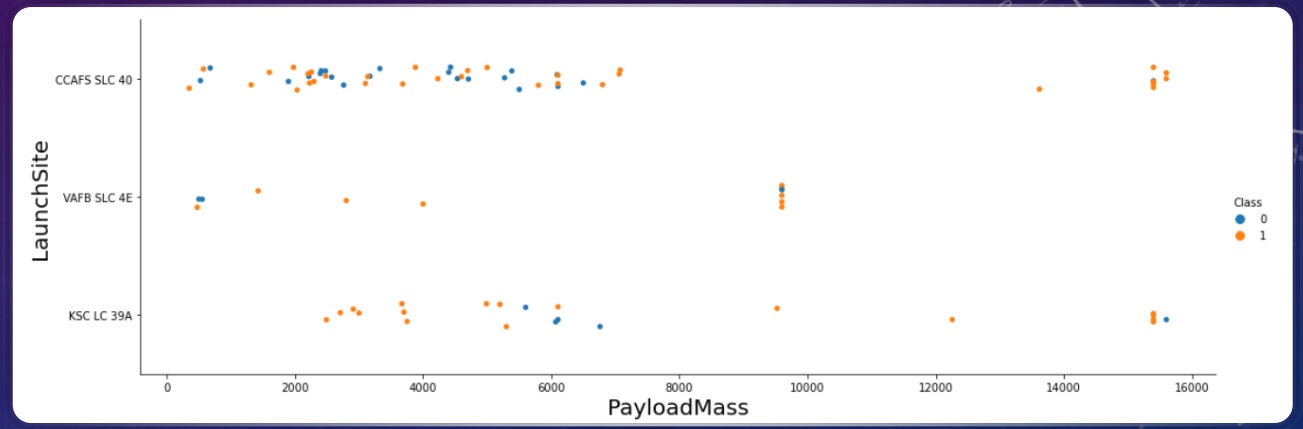
FLIGHT NUMBER VS LAUNCH SITE

- In all these graphs the trend is realized that as the flight number increases the chance of success is increased



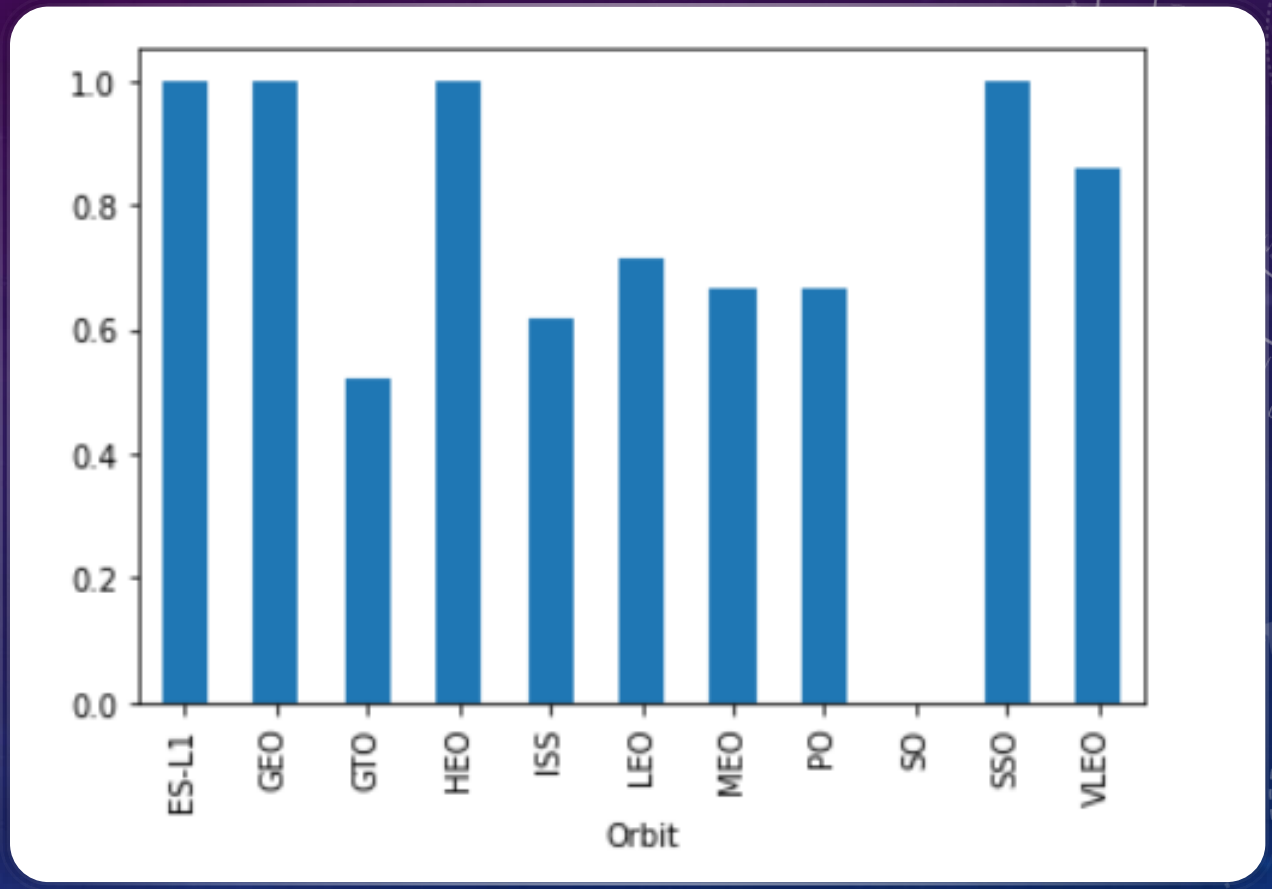
PAYLOAD VS LAUNCH SITE

- It can be observed that a payload mass above 8000 is optimal for all these different launches and results in a higher chance of success.



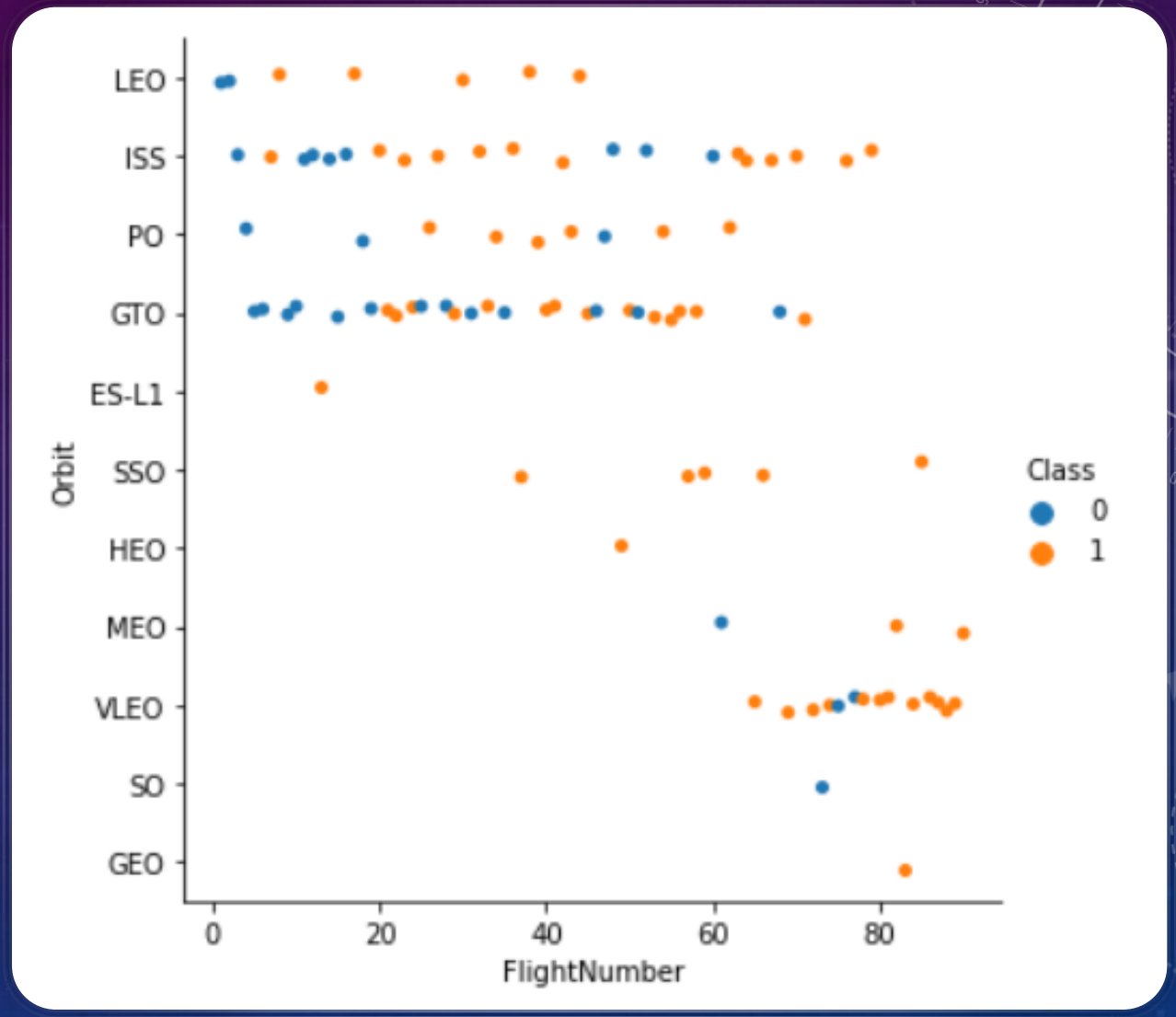
SUCCESS RATE OF EACH ORBIT

- It can be assumed that the orbit has an impact on the outcome of the landing as there are 4 orbits with perfect success rates, however there is one orbit with a zero percent success rate. However, the reliability of this graph is very low as some of these orbits have had very few rocket launches, a greater sample would be optimal for this graph.



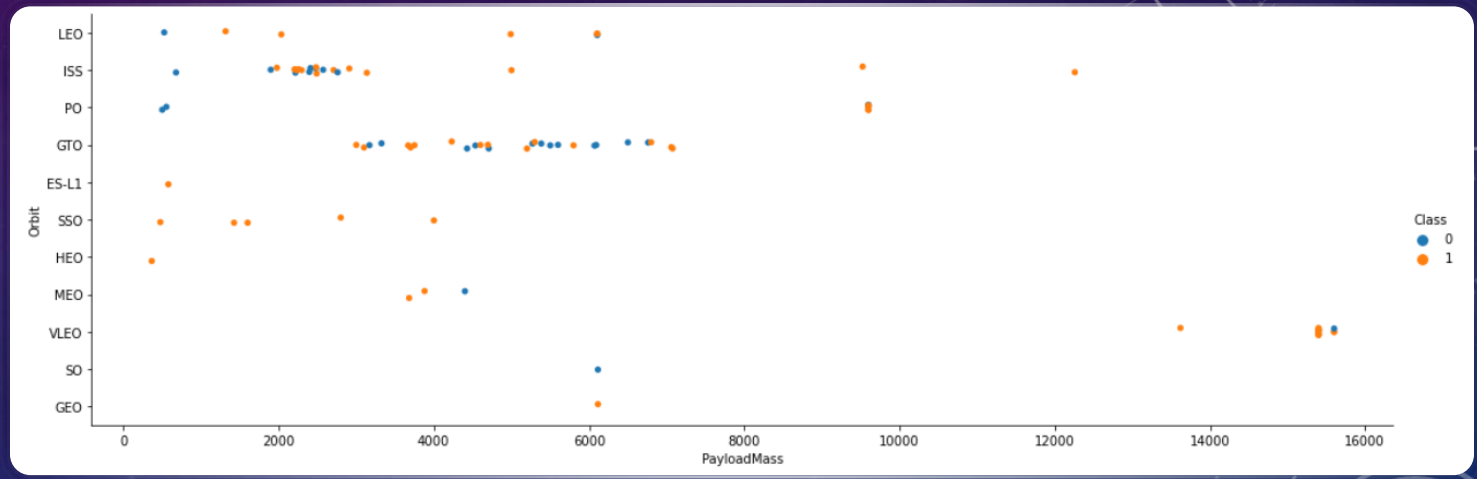
FLIGHT NUMBER VS ORBIT TYPE

- From Here we can see that some orbits have had a very low number of missions. Hence, we can gather no reliable trend from them.
- The data seems to varied for the other orbits however. Indeed the only main correlation that can be observed is for the LEO Orbit which see a greater success rate with the flight number increase.
- The GTO orbit and the ISS are the orbits with a low chance of success based on their large amounts of missions, while the VLEO and PO and LEO have a relatively high chance of success



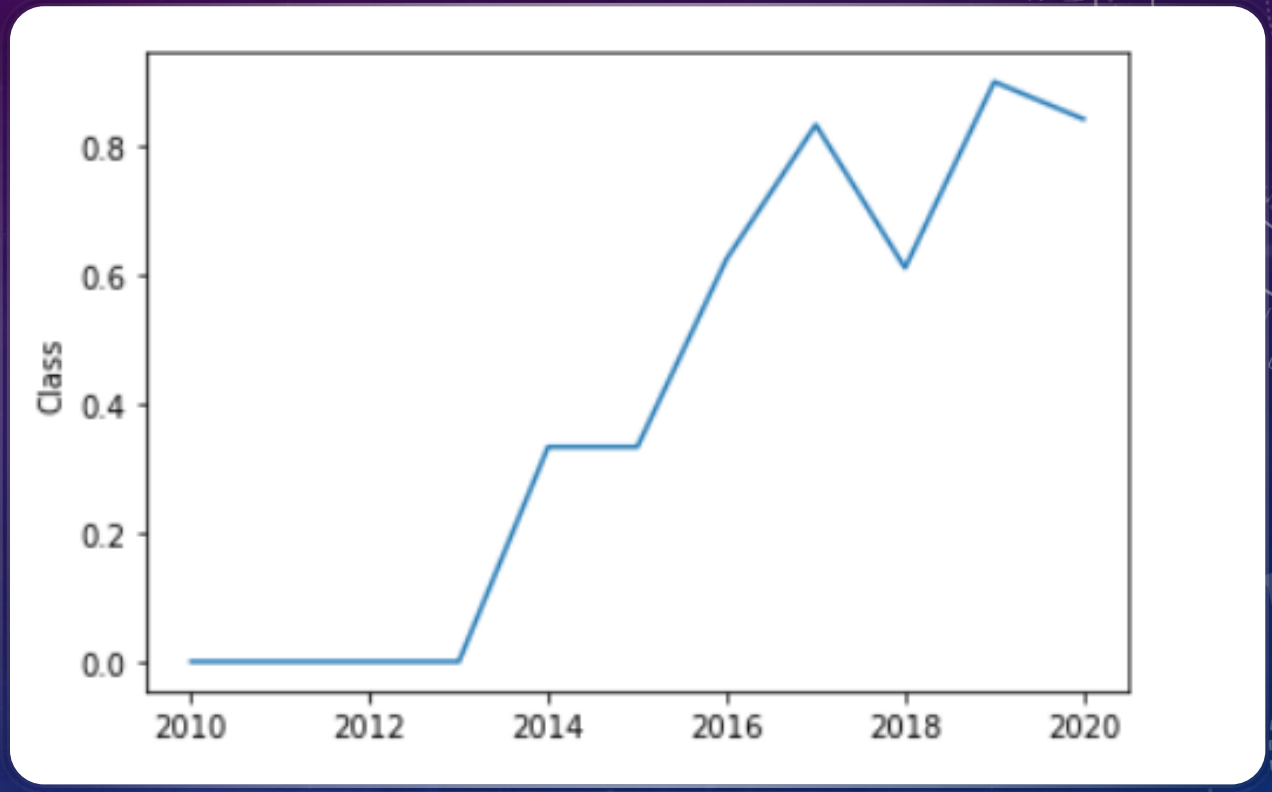
PAYLOAD VS ORBIT TYPE

- Here we can observe that a heavy payload has a negative influence on GTO orbit
- A heavy payload, however, has a positive influence on Leo and ISS orbits.



LAUNCH SUCCESS YEARLY TREND

- In general the success rate of launches is increasing. This is most likely due to the year on year improved in science and technology humans are making.



LAUNCH SITES

Unique Launch Sites:

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

Launch Sites beginning with ‘CCA’:

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

PAYLOAD MASS:

Total Payload Mass:

1
45596

Average Payload Mass:

1
2928

LAUNCH OUTCOMES

- The first successful landing outcome was on:
- 22-12-2015.
- There were 4 Boosters which successfully landed with a payload mass between 4000 and 6000 kg. These are on the right
- There are overall 99 successful mission outcomes with 1 failure and 1 where the payload status was unclear. This does not depict the number of failed landings as some of these failed landings were planned. The table from the SQL query is on the right.

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

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

BOOSTERS CARRYING MAX PAYLOAD

The boosters displayed on the right are all those that have carried max payload.

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

- These are the 2 booster versions that failed the landing out come in 2015

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

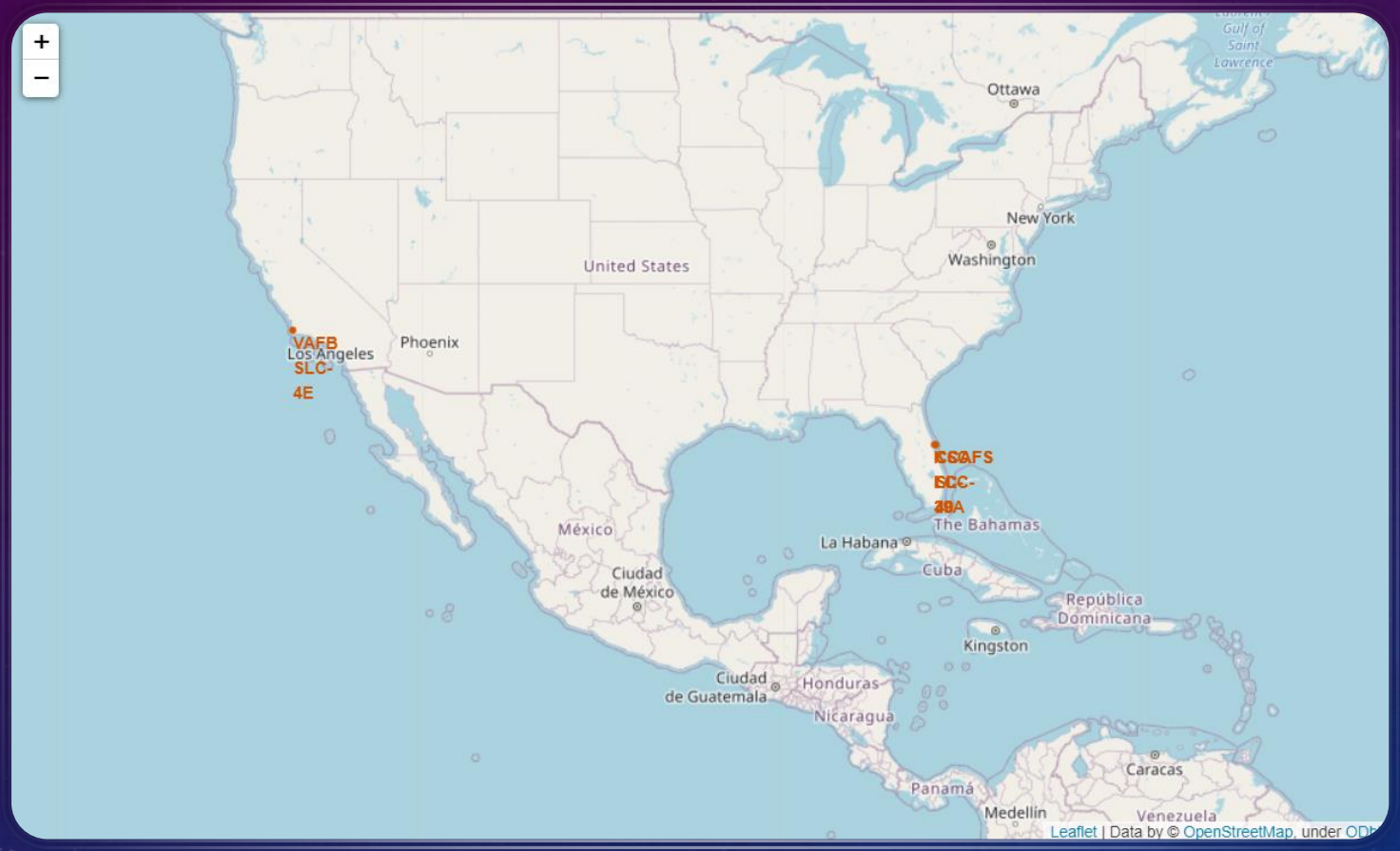
RANK THE COUNT OF LANDING OUTCOMES BETWEEN 04-06-2010 AND 20-03-2017

- This gives a count of the different forms of successes and failures that were undergone based on landing outcomes during 04-06-2010 and 20-03-2017

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

LAUNCH SITES PROXIMITY ANALYSIS

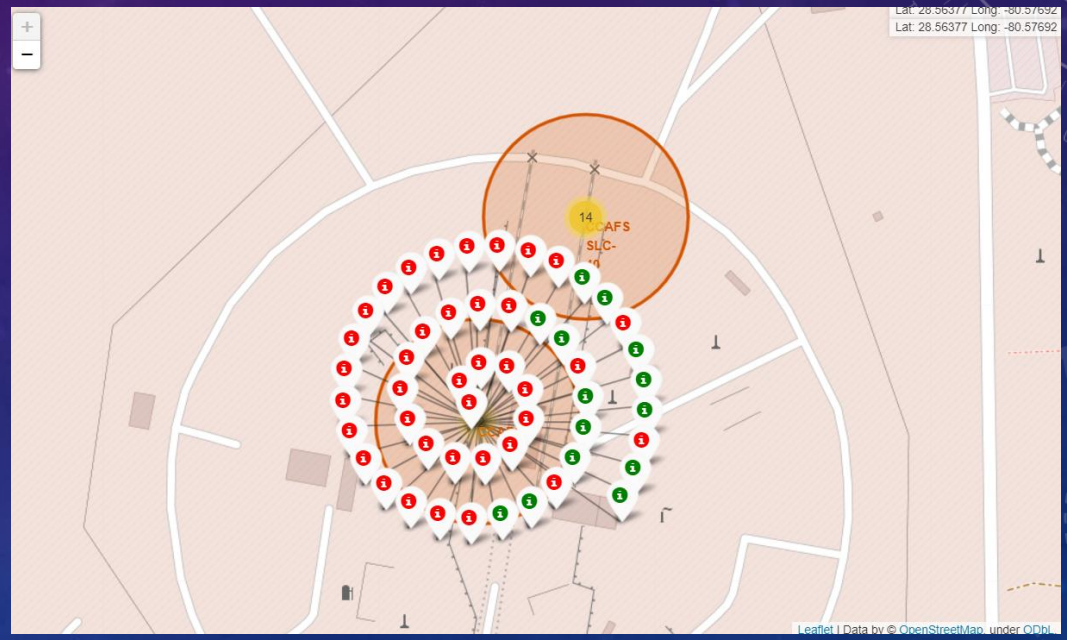
ZOOMED OUT LOCATION OF ALL LAUNCH SITES



- The Launch Sites are all near the coastlines
- It is hard to make out but upon zooming in there are actually there launch sites all close to each other on the east coast of Florida

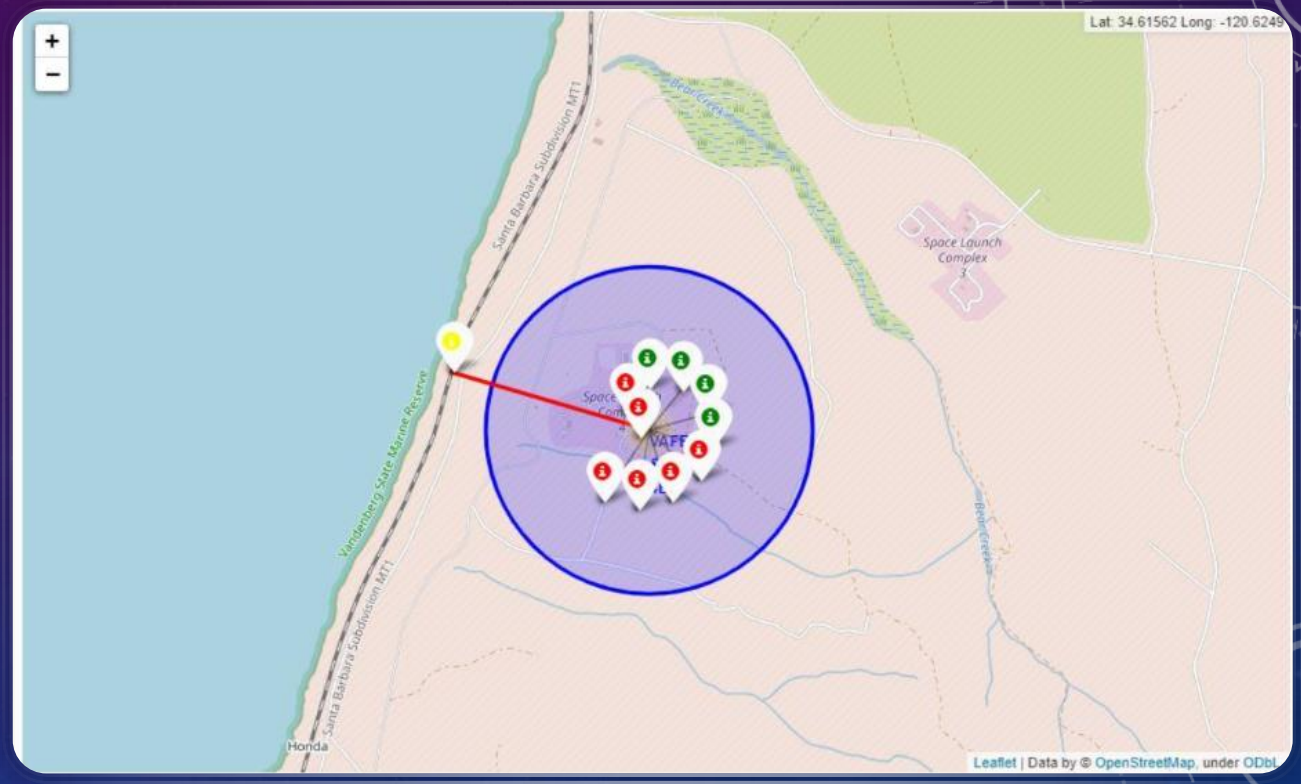
ZOOMED IN VIEW

- Here we can observe the icons. They are colour coded to depict success or failure of the launches
- Green represent success while red represent failure.



LAUNCH SITE LOCATIONS AND PROXIMITIES:

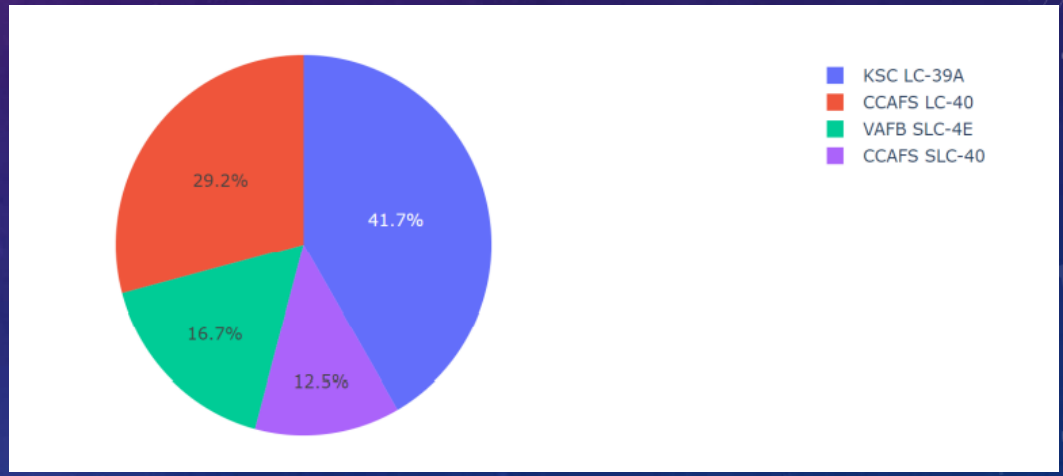
- This line reveals the proximity of one launch site to a railway line, but also reveals that the launch site is very close to the coast. No city can be seen nearby which is to be expected due to safety reasons.



BUILD A DASHBOARD WITH PLOTLY DASH

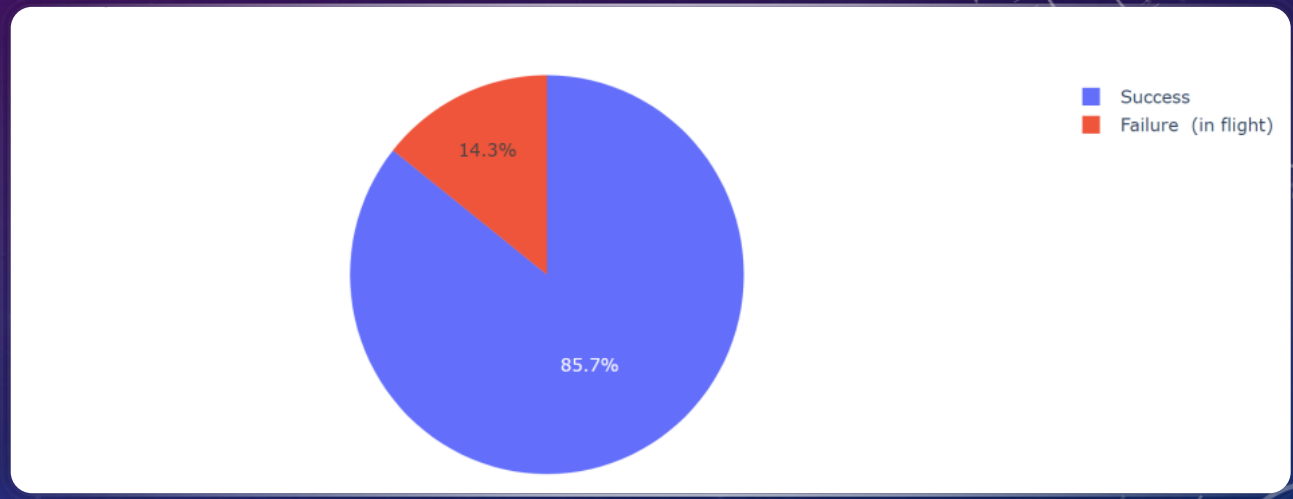
SUCCESS COUNT FOR ALL SITES

- This bar graph depicts the different success rates for the different launch sites.
- It can be seen that KSC LC-39A has the most successful landing outcomes, while CCAFS SLC-40 has the least



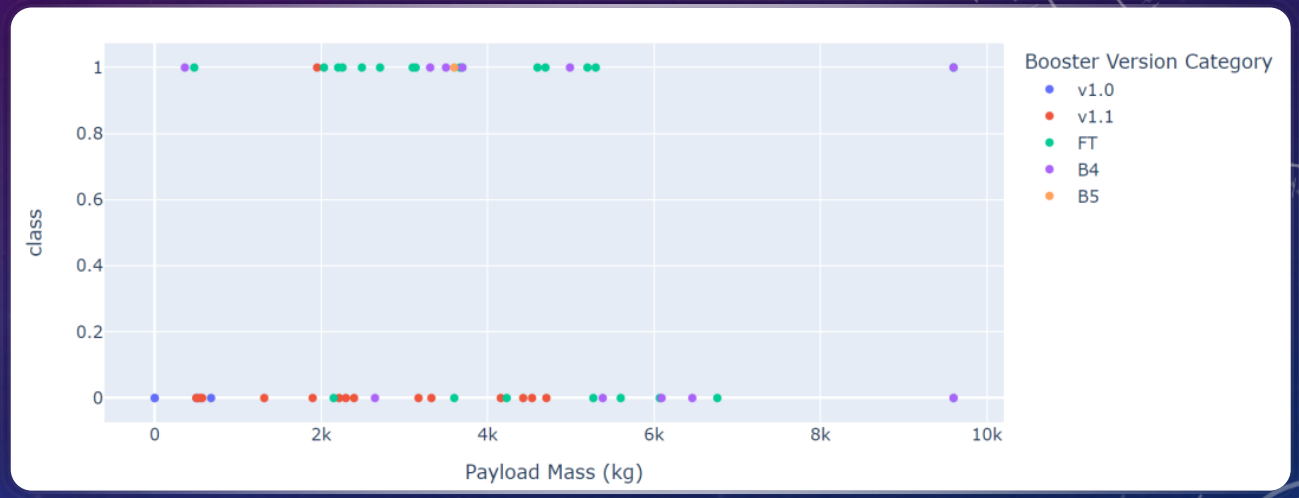
BEST LAUNCH SITE FOR A SUCCESS RESULT

- The proportion of successful mission outcome for CCAFS LC-40 is 85.7%
- This shows this launch site has the highest launch to success ratio



PAYLOAD VS LAUNCH OUTCOME SCATTERPLOT AND RELATIONSHIP WITH BOOSTER VERSION

- This graph shows no results as there are numerous failures and successes through the graph.
- The FT booster, however, does have a positive result in general between2-4.5k loads, and is the best booster for a success in general.



PREDICTIVE ANALYSIS (CLASSIFICATION)

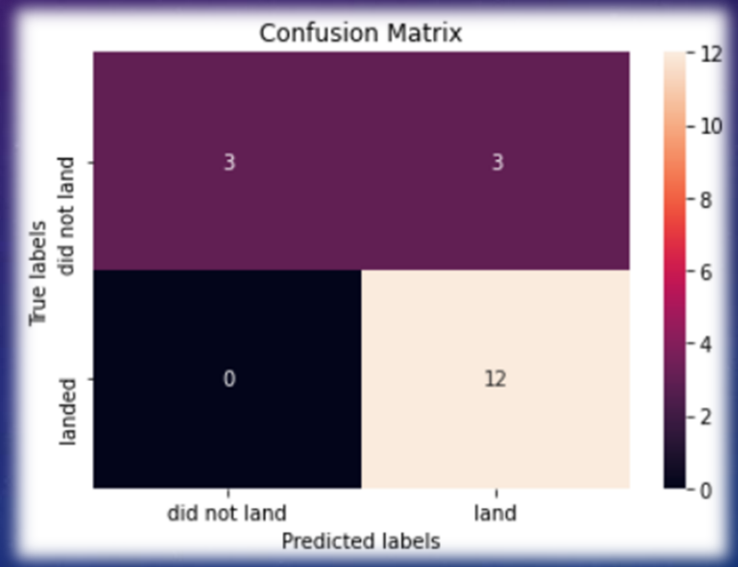
CONFUSION MATRIX

- All models had the same accuracy of 83.3333
- However, the Decision tree model had the best score accuracy of 0.877 so that is what we will use for the confusion matrix

	SVM	TREE	KNN	LOGISTIC REGRESSION
Accuracy	83.3333	83.3333	83.3333	83.3333
Best Score	0.848	0.877	0.848	0.846

CONFUSION MATRIX

- This model reveals that, while the rocket landed every time we predicted it to land, 3 times the mission failed even though it was predicted to succeed. The high rate of false positives is probably a point of worry for Space X.



CONCLUSION

- It was discovered that there are numerous parameters that can be an indicator of success of the stage 1 of the falcon 9 rocket returning successfully.
- These are, in order of most to least important:
  - Booster version [The FT Boosters clearly have the highest success rate]
  - Orbit [Certain orbits have higher success rates]
  - Payload Mass [In general, lighter payload masses have lower chances of success]
  - However, for orbit and payload mass, more data would have been useful, as some of the variables within these parameters do not have enough data for reliable conclusions to be drawn.
- Machine Learning Models are also useful to predict the outcome of the launches, as can be seen through their high accuracy
- However, they have a problem with False Positives, which means machine learning models should be improved upon.
- From the current trend being seen, it has been observed that the success rate has grown higher throughout the past years, and considering rapid growths in tech and science sector and looking at the trend, we can predict success rates will go higher still.
- Thus for Space Y to compete with Space X, they will have to focus on specific orbits, such as Geo and Leo orbits, have optimal amounts of Payload Mass, and Booster Versions similar in design to the FT boosters. They will also need to obtain similar rockets to the falcon 9, and improve upon machine learning techniques so that they are more successful and not have false positives, which may be a temporary investment, but will see prices similar and competitive to that which Space X currently has.