IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason Caron
22JUN2024

Outline

Executive Summary

- Summary of methodologies
  - Collected data using the SpaceX API and Data scraping of the SpaceX Launch Wikipedia page with BeautifulSoup
  - Data wrangling to clean up null data and some features engineering to standarize the data
  - Created multiple graphs and charts to analize the data
  - Created a folium map to see the relationship of the launch sites to other locations
  - Created a dashboard to analize successful launch data
- Summary of all results
  - Built a predictive model to determine the likely success rate of future launches

Introduction

- In this project we will determine the success rate of the Falcon 9 rocket and develop a model to predict if the rocket will land successfully.
- The problems we need to address are finding the different factors that contribute to the success of the launch such as the location, payload mass, rocket version, type of orbid, etc
- Then take this information and build a model that will predict if a future launch will be successful

Section 1

Methodology

Methodology

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 models

Data Collection

- Data was collected from two sources:
  - SpaceX API
  - Launch data on the SpaceX Wikipedia page
- The following slides will illustrate the process used to collect the data

Data Collection – SpaceX API

Used a "requests.get" call to retreive the JSON data

Decoded the response into JSON data with "json()"

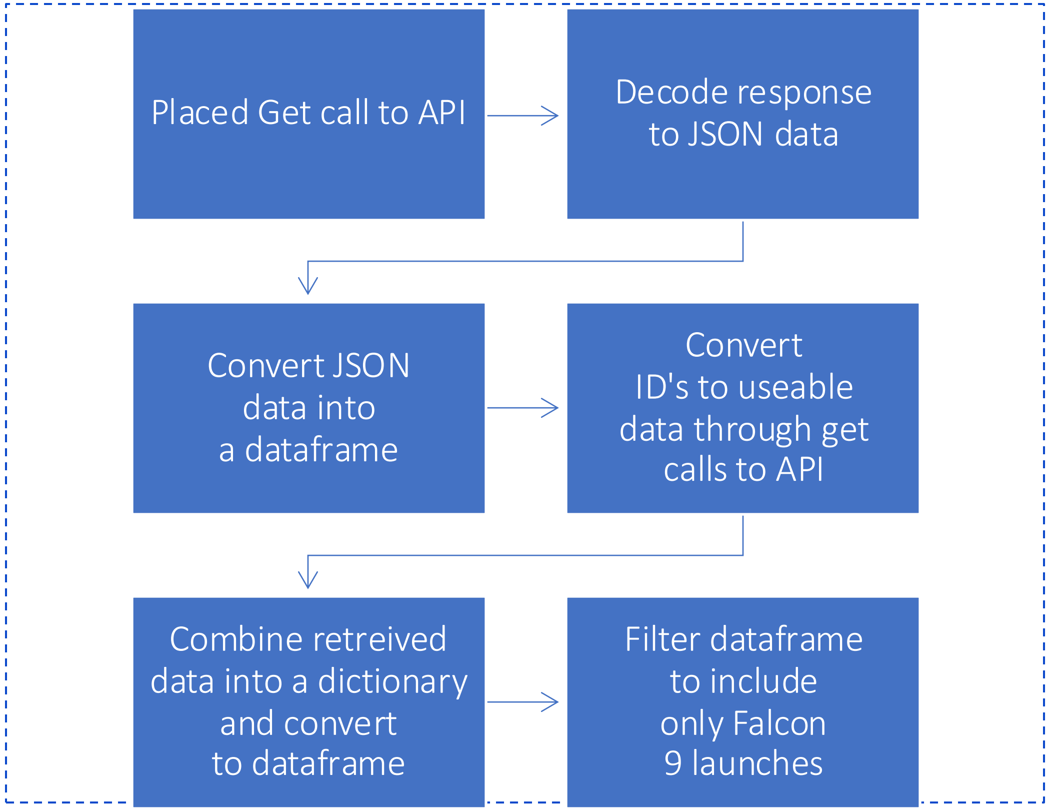
Converted JSON data to a dataframe
utilizing "json_normalize"

Used functions to make get requests to the API to
convert ID's in the dataset to useable data

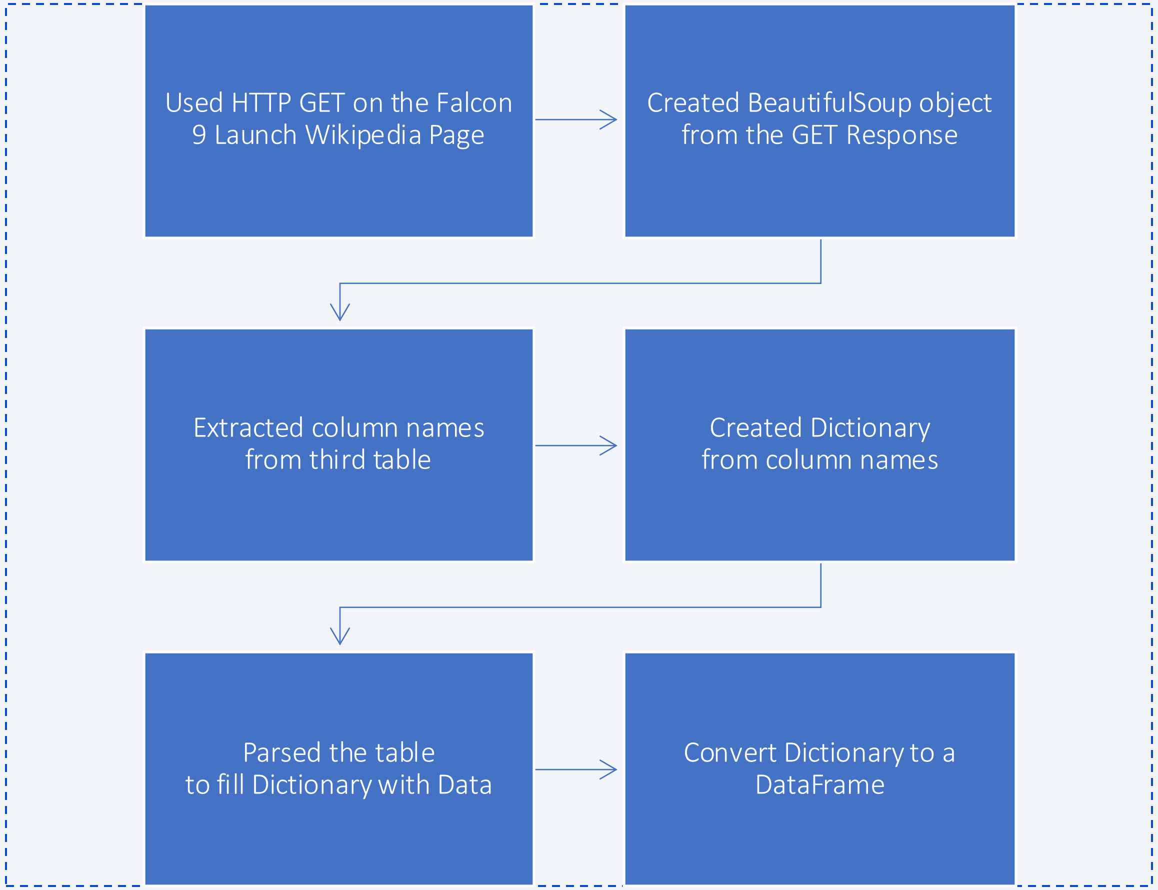
Combined data from the original dataframe and
the retreived data from ID's into a dictionary and converted
the dictionary into a dataframe

Finally, filtered the dataframe to include only Falcon
9 launches

GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/1-Data-Collection-From-API.ipynb



Data Collection - Scraping



- Extracted HTML data from the Falcon 9 Launch Wikipedia Page using "requests.get()"
- Passed the response to a BeautifulSoup object
- Found all the tables using "find_all('table')"
- Extracted the column names from the third table with a loop and "find_all('th')"
- Created a Dictionary using the Column names and utilized
- Parsed the table and extrated the data from each column
- Converted the Dictionary to a DataFrame
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/2-Data-Collection-From-Webscraping.ipynb

Data Wrangling

- Found where data was missing using "isnull().sum()"
- It was discovered that PayloadMass and LandingPad had null values
- LandingPad will retain the None values to indicate when landing pads were not used
- Calculated the mean value of the PayloadMass and replaced all the null values with the mean using the "mean()" and "replace()" dataframe functions
- Checked the null values again to confirm that they had all been replaced
- Converted categorical values to numbers for use in model building
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/1-Data-Collection-From-API.ipynb

EDA with Data Visualization

- Multiple Scatter Point Charts were created:
  - FlightNumber VS PayloadMass and their success rate. This showed that the higher flight numbers were more successful as well as the smaller payloads
  - FlightNumber VS LaunchSite and their success rate. This again showed that the later flights were more successful and that after the first 20 flights, the success rate was similar among all of the launch sites
  - PayloadMass VS LaunchSite and their success rate. You could see that at VAFB-SLC nothing was launched with more than 10k LBS payload
  - FlightNumber VS Orbit and their success rate. LEO showed that the success rate improved over more launches while the other orbits do not show a consistent relationship to the number of flights
  - PayloadMass VS Orbit and their success rate. LEO, ISS, and Polar orbits all have higher success rates with heavier payloads.
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/3-EDA-Visualization.ipynb

EDA with Data Visualization (Cont)

- A bar chart was created to show the success rate of each type of orbit. It showed that ES-L1, GEO, HEO, and SSO has a 100% success rate
- Plotted a line chart to show the success rate year over year. It shows that as time passed the success rate improved
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/3-EDA-Visualization.ipynb

EDA with SQL

- Multiple SQL calls were used to pull information about the launches
  - The names of the unique launch sites
  - 5 records where the launch site started with CCA
  - Total payload mass carried by boosters launched by NASA(CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Date when the first successful landing was acheived
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/4-EDA-SQL.ipynb

EDA with SQL (Cont)

- Multiple SQL calls were used to pull information about the launches
  - 6. Boosters that have success on a drone ship and payload between 4000 and 6000
  - 7. Total of successful and failed mission outcomes
  - 8. Booster versions that have carried max payload mass
  - 9. Month, Booster Version and Launch Site for failed Drone Ship landings in 2015
  - 10. List landing outcomes in descending order between 2010-06-04 and 2017-03-20
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/4-EDA-SQL.ipynb

Build an Interactive Map with Folium

- Created a map to show information about each launch site
  - Markers with Circles to show each launch site
  - Clusters at each site to indicate successful and failed launches
  - Lines connecting launch sites to nearest coastline, rail, highway and city with a marker indicating the distance
- These objects were added to show the relationship of the launch sites with their surrounding areas
- GitHub link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/5-Folium-Launch-Site-Locatoins.ipynb

Build a Dashboard with Plotly Dash

- Created a Dashboard with the following graphs and interactions
  - Dropdown that allows you to select each launch site as well as an option to select all of them
  - Pie Chart that will show the success and failure of each launch site selected and a comparison of the successes if all launch sites are selected
  - Scatterplot of the payload mass vs success rate for each booster version. Shown for selected launch site or for all sites depending on what is selected
- These graphs/plots were selected to show the relationship of payload mass to booster version as well as showing the most successful launch sites
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/6-Dashboard-Launch-Success-Rates.py

Predictive Analysis (Classification)

- Using the StandardScaler transform, standarized the data for X and used the Success Rate for Y. Then utilizing train_test_split created an 80/20 split of training and testing data. With the small data set, this resulted in only 18 test samples
- Using GridsearchCV, trained an tested the data with the following models:
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K Nearest Neighbor
- Calculated the accuracy of each model using the Score method and plotting a Confusion Matrix
- Each model had similar accuracy and the same results on the Confusion Matrix with an issue with false positives. We may have seen more of a difference in the models accuracy with a larger dataset
- GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone/blob/main/7-Machine-Learning-Predictive-Analysis.ipynb

Predictive Analysis (Classification) Cont

Applied StandardScalar transform to standarize the data for X

Set the Success Rate to Y

Used train_test_split
to create an 80/20 split of training and testing data

Utilizing GridsearchCV trained and
tested LogReg, SVM, Decision Tree, and KNN on the data

Calcualted the accuracy using Score for each method

Created a Confusion Matrix for each method

Results

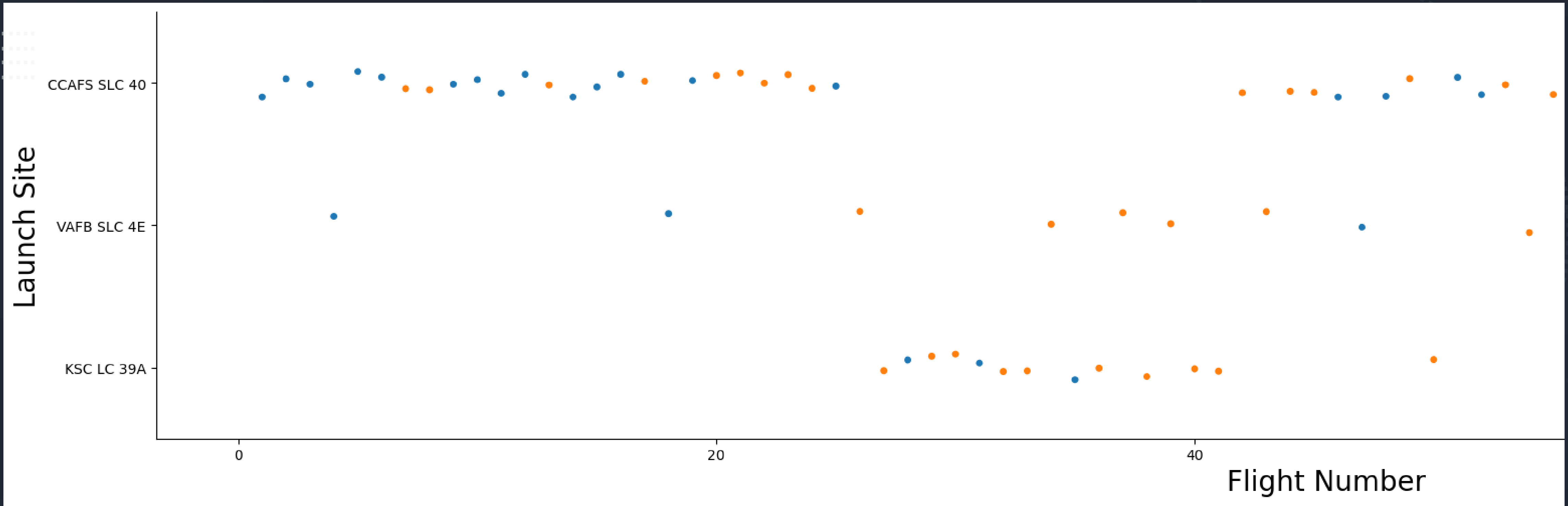
- The following slides will show the results by showing the plots and screenshots:
  - Exploratory data analysis results using various plots and graphs
  - Interactive analytics demo in screenshots of the dashboard that was created
  - Displays of the various maps created with Folium
  - Predictive analysis results

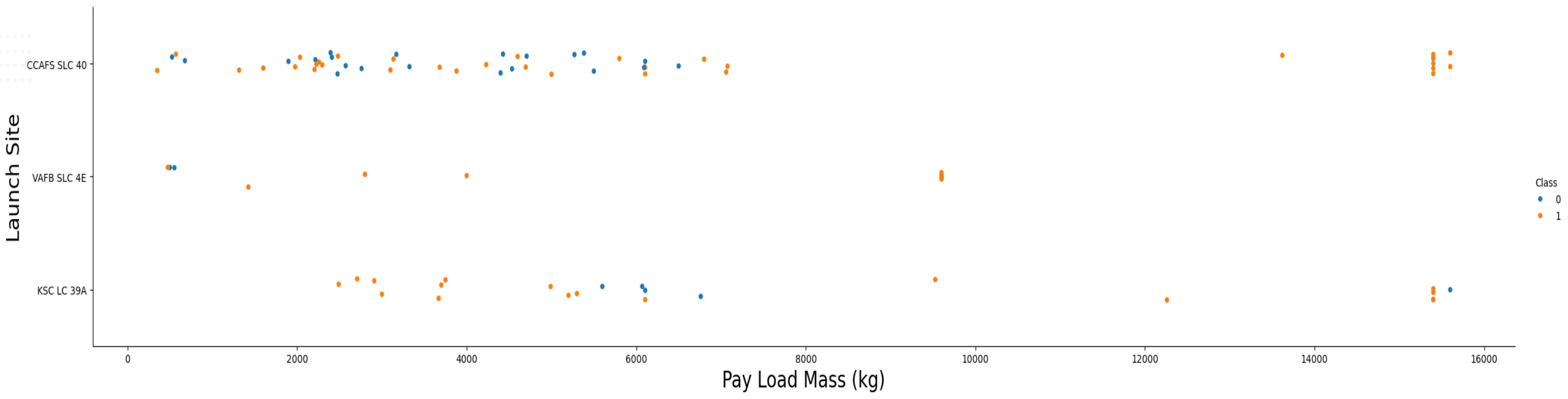
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

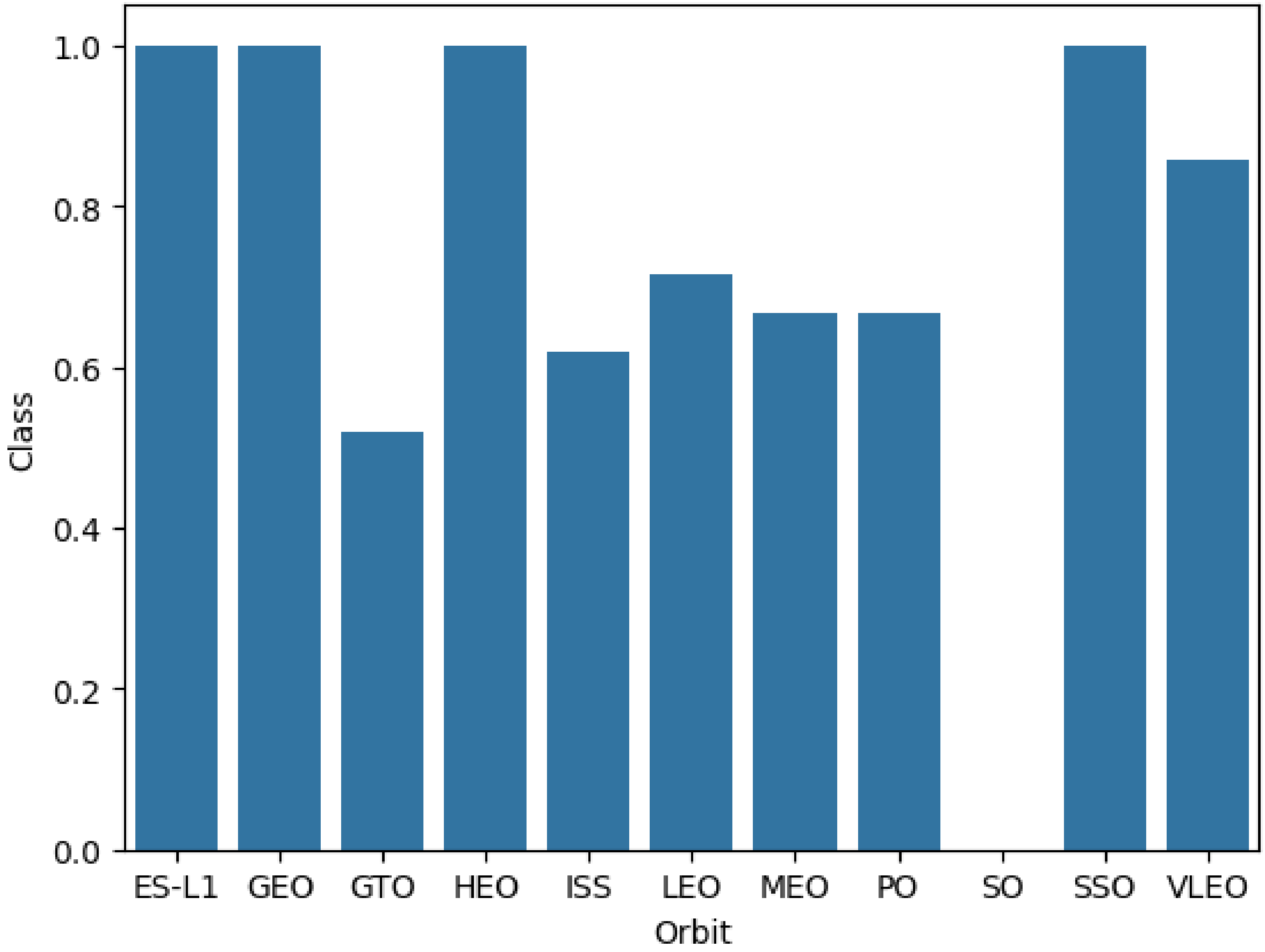
- This plot shows that as the flight number increases the success rate also increases at all three launch sites.





Payload vs. Launch Site

- As the payload mass increases, the success rate also increases
- For payloads over 9000kg there was only one failure

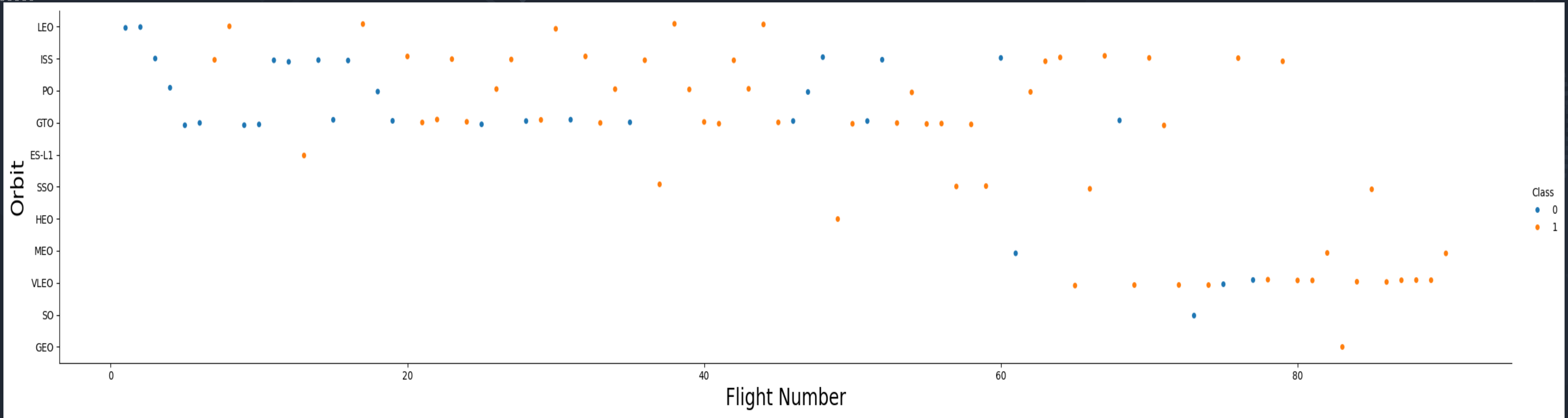


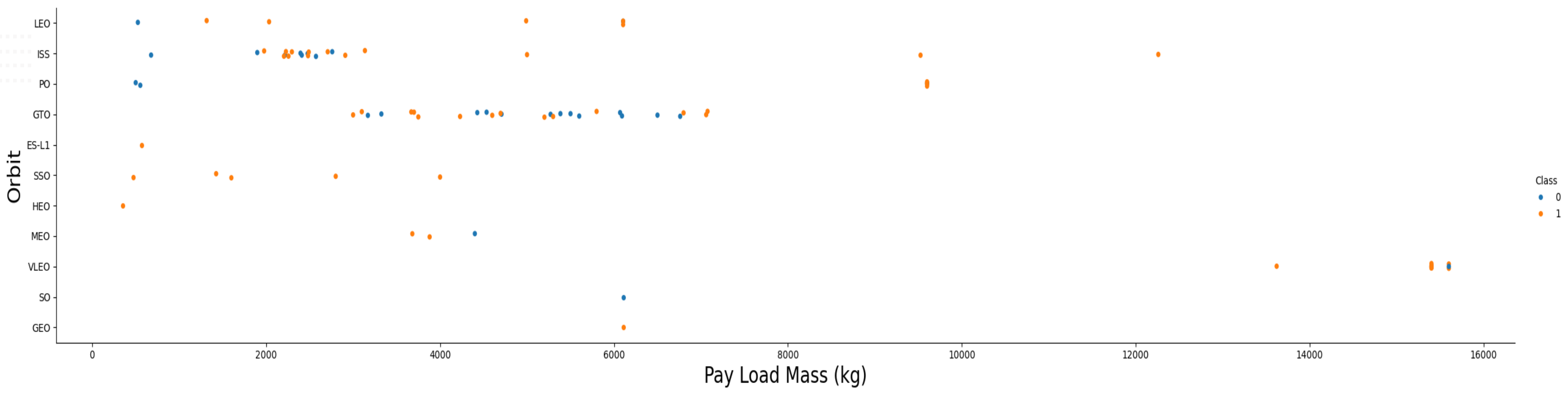
Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, and SSO orbits all had a 100% success rate
- The SO orbit had no successful launches

Flight Number vs. Orbit Type

- For the LEO and VLEO orbits, as the flight numbers increased, the success rate improved.
- There was less of a correlation between flight number and success for the other orbits



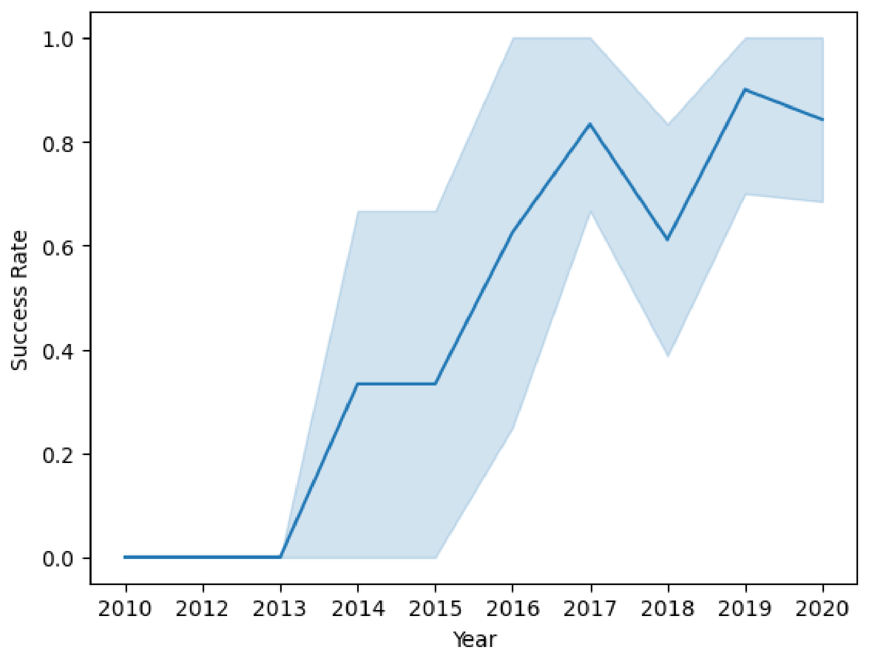


Payload vs. Orbit Type

- For LEO, PO, and ISS the heavier the payload the more successful the launch
- The other orbits had a mix of success and failures regardless of the payload mass

Launch Success Yearly Trend

- As time progressed the rate of success improved with a small dip in 2018



All Launch Site Names

Using the following query we found the names of the launch sites:

```
%sql select distinct Launch_Site from SPACEXTABLE
```

This query pulls the data in the Launch_Site column and with the "distinct" command it shows only the unique names

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

Launch Site Names Begin with 'CCA'

Using the following query, we found the first five records for launch sites beginning with CCA

```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

This query pulls all the data where Launch_Site starts with CCA and limited the results to five records

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The following query pulls the total payload mass for boosters launched from NASA(CRS):

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

This query sums the data in PAYLOAD_MASS_KG_ where the customer was NASA (CRS)

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

The following query calculates the average payload mass for the F9 v1.1 booster

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = "F9 v1.1"
```

This query finds the average of the PAYLOAD_MASS_KG_ column where the booster version is F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

The following query finds the date of the first successful landing on a ground pad:

```
%sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
```

This query finds the minimum(earliest) date where the landing outcome was successful on the ground pad

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The following query finds the names of the boosters that have landed on a drone ship with a payload mass between 4000 and 6000 kg

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

This query finds the Booster_Version where the Landing_Outcome is successful on the drone ship and the payload mass is between 4000 and 6000

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

Total Number of Successful and Failure Mission Outcomes

The following query finds the total number of each type of mission outcome:

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome
```

This query finds the mission outcome and the total of each type of outcome grouped together

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

Boosters Carried Maximum Pay load

The following query finds all of the boosters that carried the maximum payload:

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

This query finds the Booster_Version where the payload mass is the max(largest) mass

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

- The following query will show the month, booster version and landing site for failed drone ship landings in 2015

```
%sql SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5)='2015' AND Landing_Outcome='Failure (drone ship)'
```

- This query extracts the date using the substr function, then pulls the booster and launch sites where the outcome is Failure (Drone Ship) in 2015.

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

The following query will rank the landing outcomes between 2010-06-04 and 2017-03-20 in descending order

```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) AS Total FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY Total DESC
```

This query lists the landing outcome and it's count, between the dates listed above in desceding order

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

Section 3

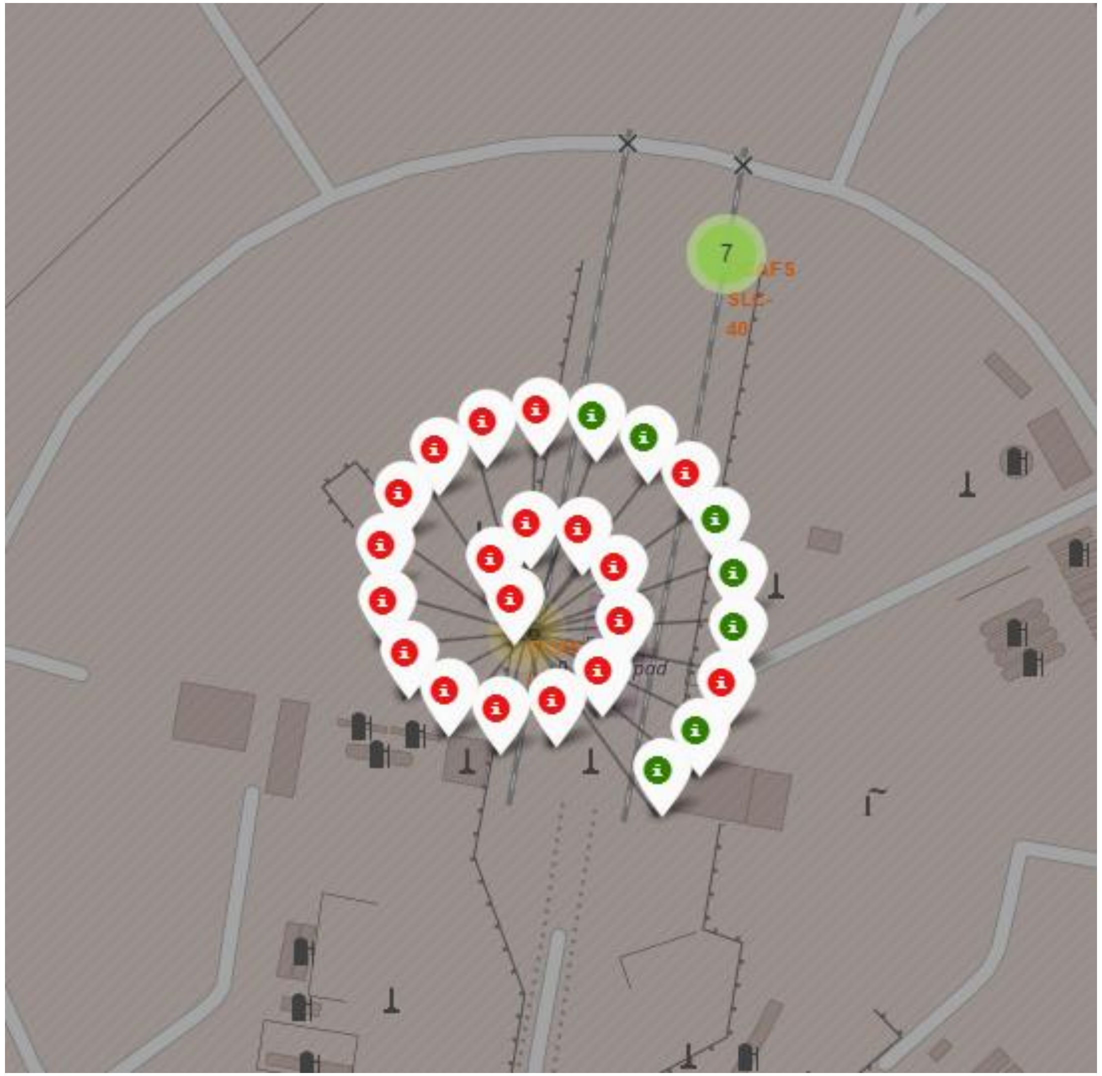
Launch Sites Proximities Analysis

Launch Site Locations

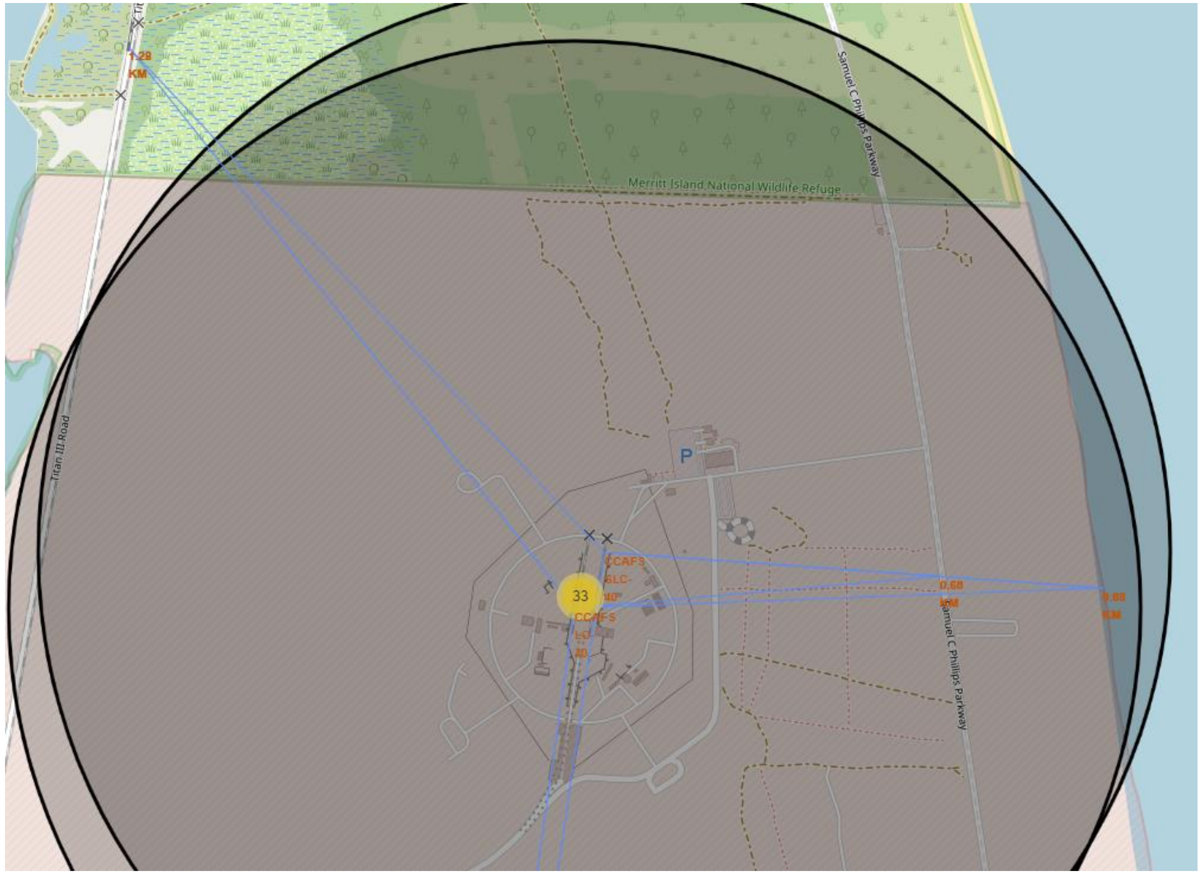
- This map shows the locations of the launch sites. One on the coast of California and the other three on the coast of Florida.

Launch outcomes at CCAFS LC-40

- This map shows the launch outcomes at launch site CCAFS LC-40
- The Green bubbles indicate a successful launch and the red indicate failure



Distance to Rail, Highway and Coastline for CCAFS LC-40 and SLC-40

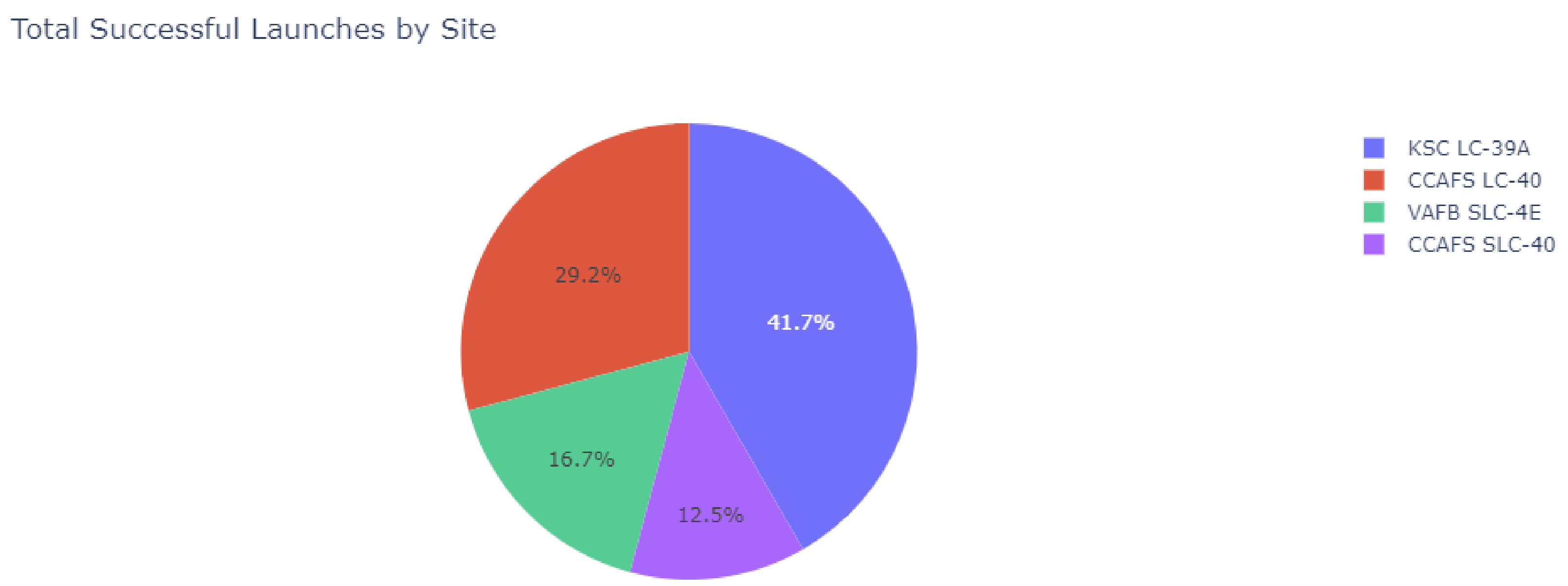


This map shows the distance of the two launch sites to the nearest Rail, Highway and Coastline with the distance

Section 4

Build a Dashboard with Plotly Dash

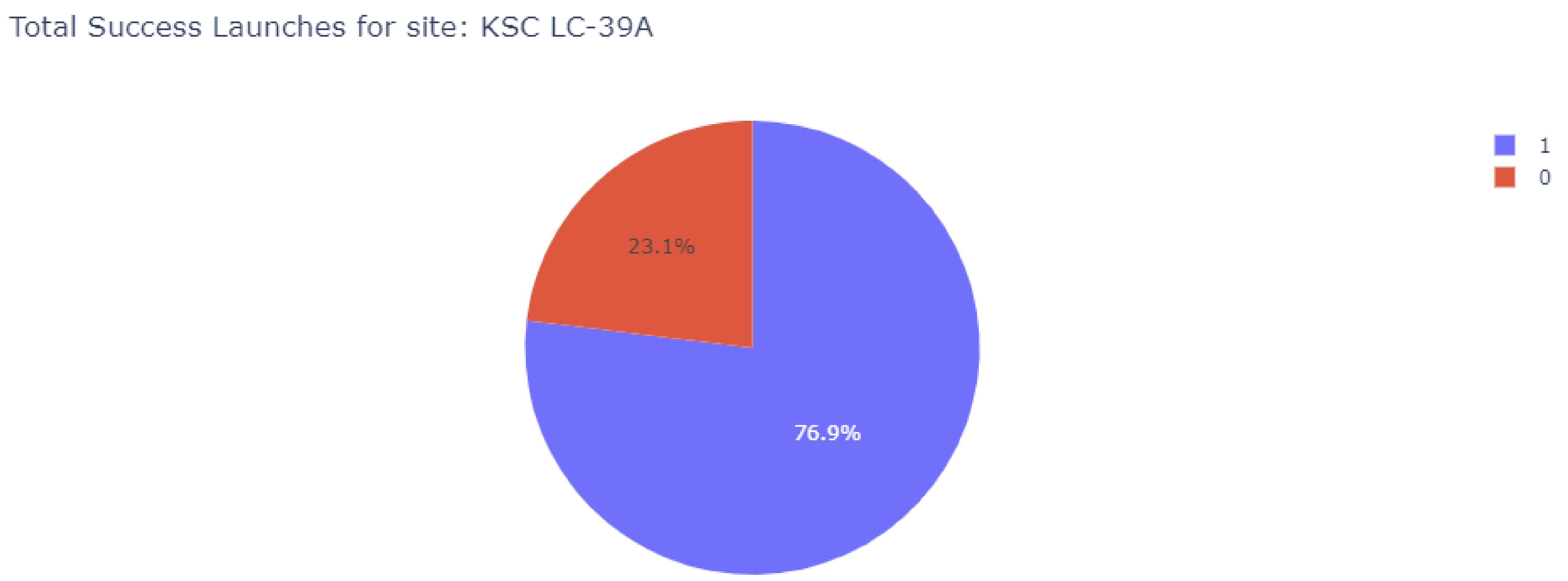
Percentage of Successful Launches by Site



This pie chart shows the percentage of total successful launches for all launch sites. We can see that KSC LC-39A has the highest percentage of successes

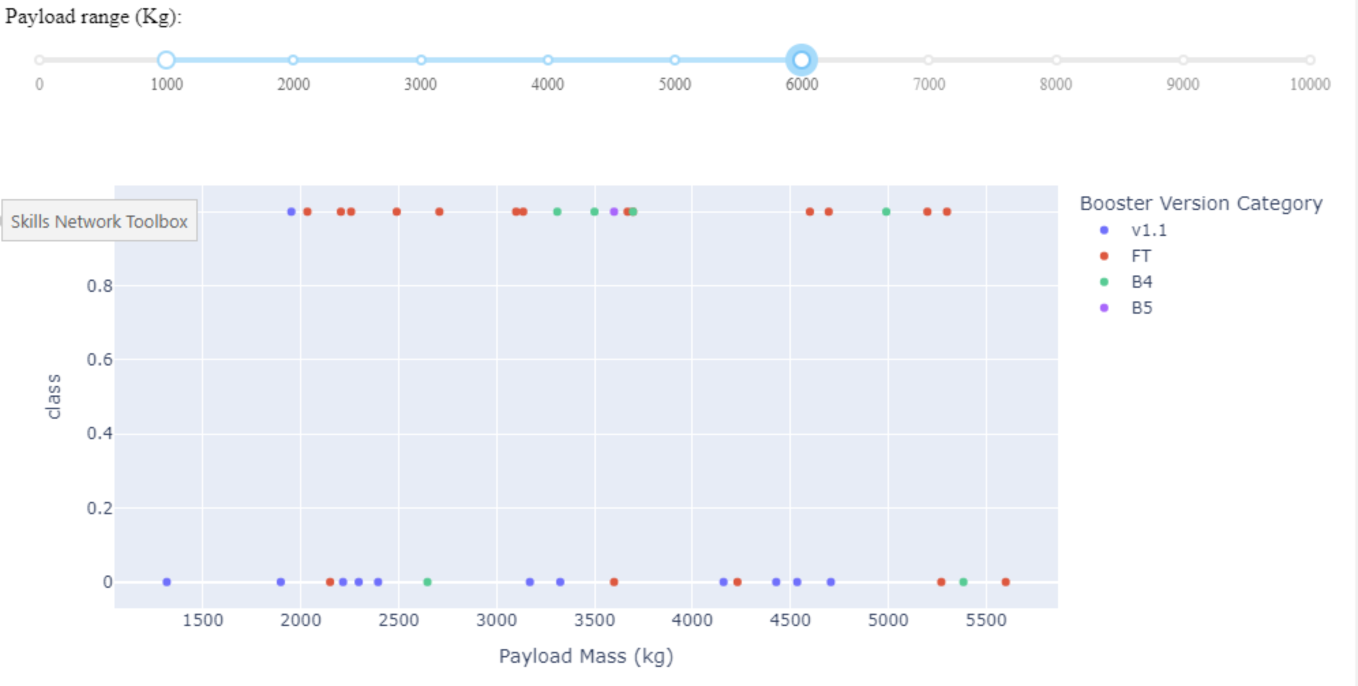
Total Successful Launches for KSC LC-39A

This pie chart shows the success rate of the most successful launch site



Success rate of booster version vs payload mass

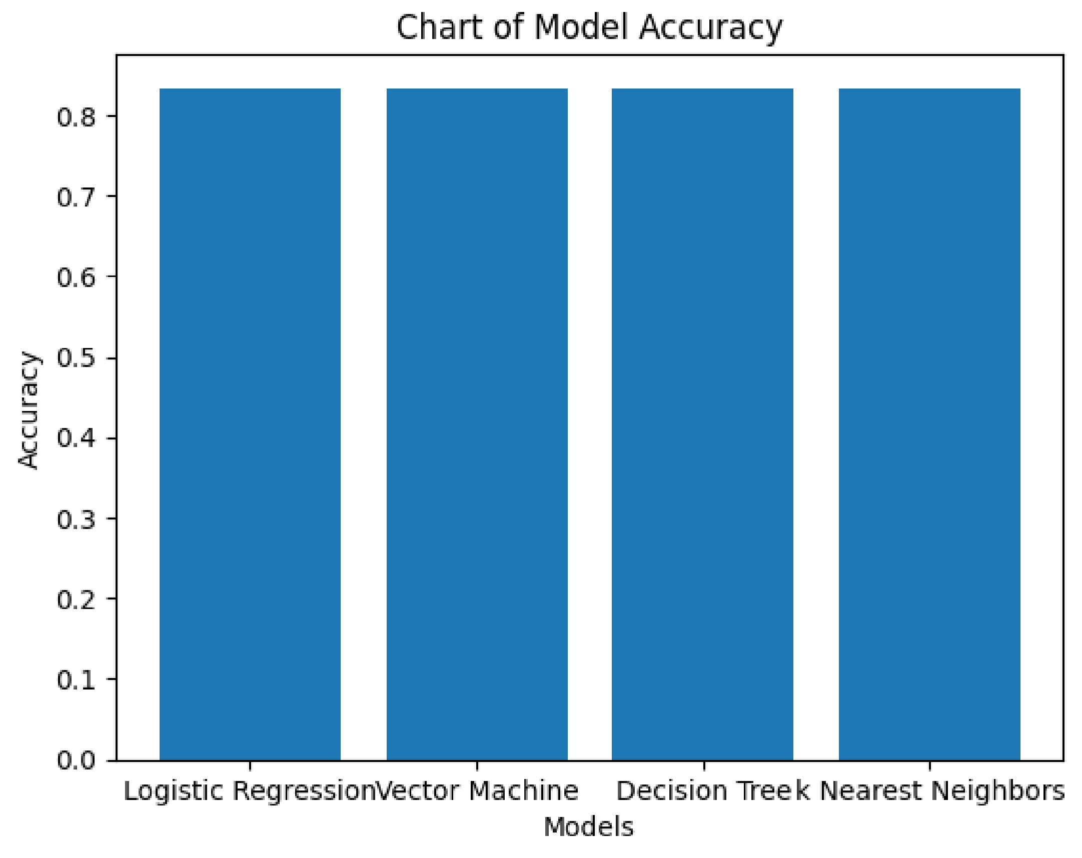
- This scatter plot shows that the FT booster version has the most successful launches with a payload mass between 1000 and 6000 kg



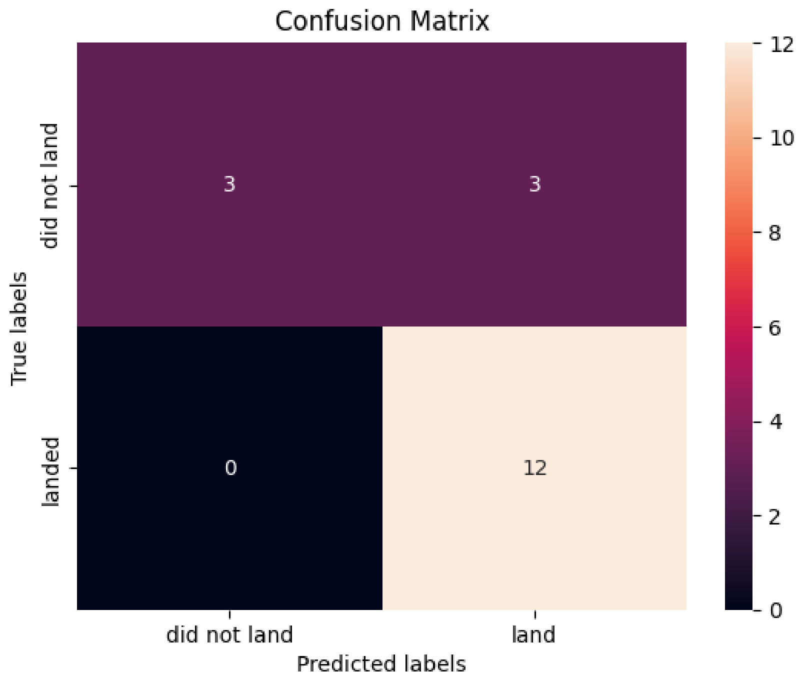
Section 5

Predictive Analysis (Classification)

Classification Accuracy



This bar chart shows that each of the models have
the same accuracy: approx 0.8333



Confusion Matrix

- A confusion matrix was created for each of the different models. They all resulted in the same matrix. It shows that there is a possibility of false positives but no false negatives.

Conclusions

- We have found that the FT version of the booster has the most success and that the KSC-L39A launch site has the best success rate for launches
- We developed a model to predict the future success of launches with a ~0.8334 accuracy. However there is still a chance of reporting a false positive from these models.
- With a larger data set we could tune the models to be more accurate in the future.

Appendix

GitHub Link: https://github.com/Jacaron75/Data-Science-Capstone

Thank you!