Anastasia Chatzitriantafyllou
20th December 2022

IBM DS0720EN - Data Science and Machine Learning Capstone Project

Outline

Executive Summary

Methodologies

- Data Collection using API
- Data Collection using Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction (Classification)

Results

- Exploratory Data Analysis results
- Interactive analytics results
- Predictive Analytics results

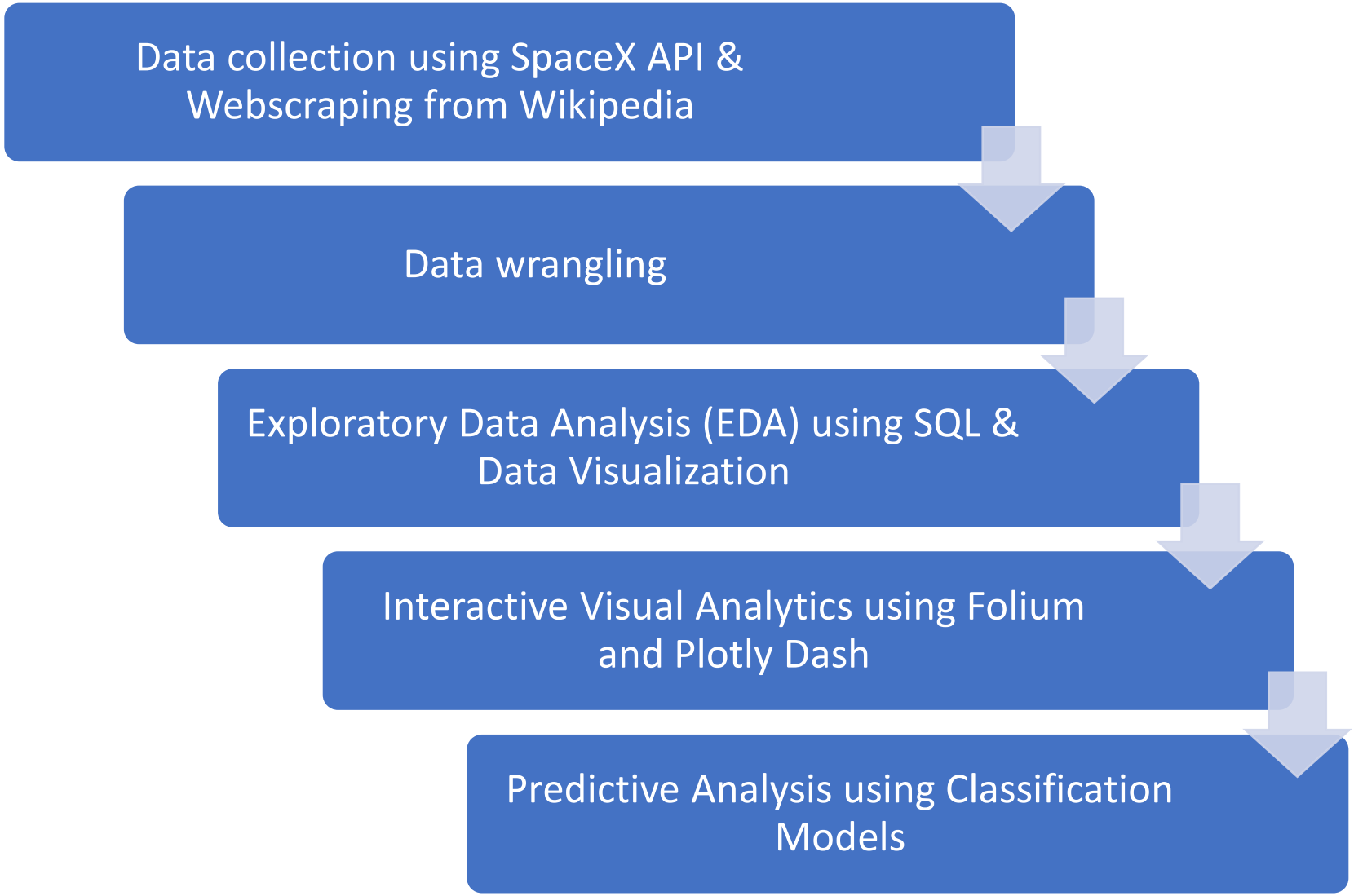
Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.

Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

The goal of the project is to create a machine learning pipeline to **predict if the first stage will land successfully**.

Methodology



Data Collection – SpaceX API

- Using the GET request we requested launch data from the SpaceX API.
- We cleaned the requested data.
- We performed data wrangling and formatting.

Notebook link:

https://nbviewer.org/github/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Data%20Collection%20API.ipynb

Data Collection – Web scraping

- We performed web scrapping using BeautifulSoup to acquire Falcon 9 launch data from Wikipedia.
- We parsed the table and converted it into a pandas dataframe.

Notebook link:
https://nbviewer.org/github/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Webscraping.ipynb

Data Wrangling

- We performed Exploratory Data Analysis (EDA) and determined the training labels.
- We calculated the number of launches for each launching site, the number and occurrence of each orbit.
- We created the landing outcome label.

Notebook link:
https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Data%20Wrangling.ipynb

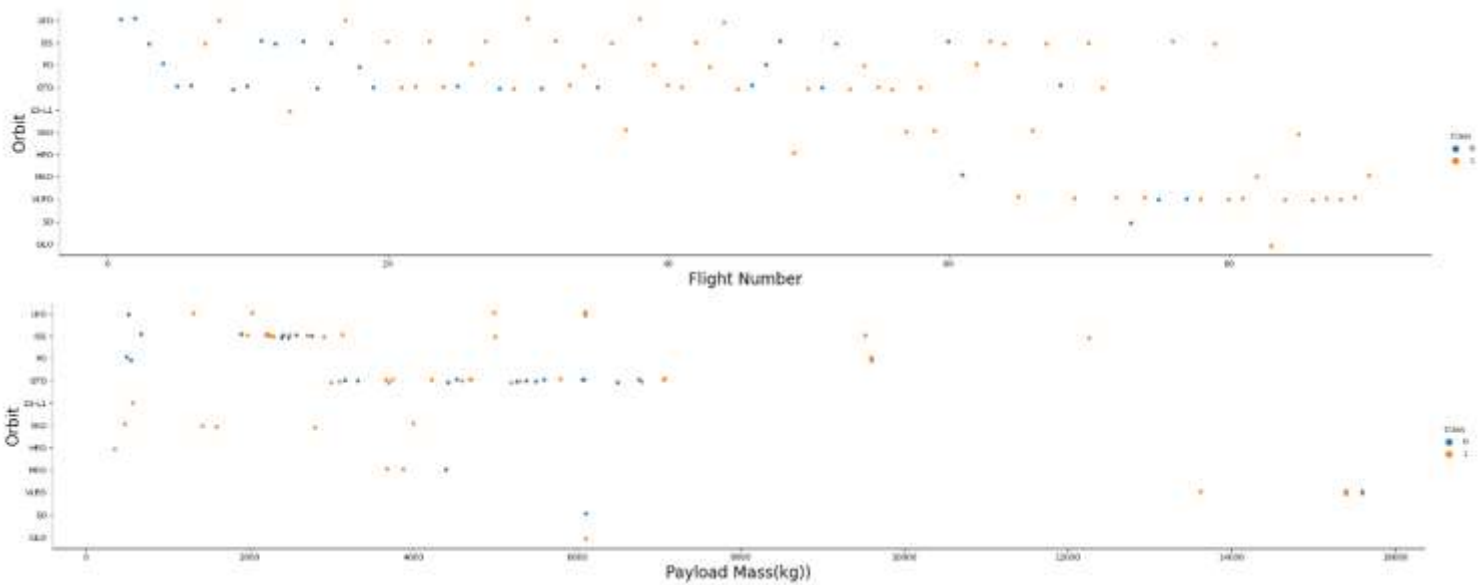
Exploratory Data Analysis – SQL

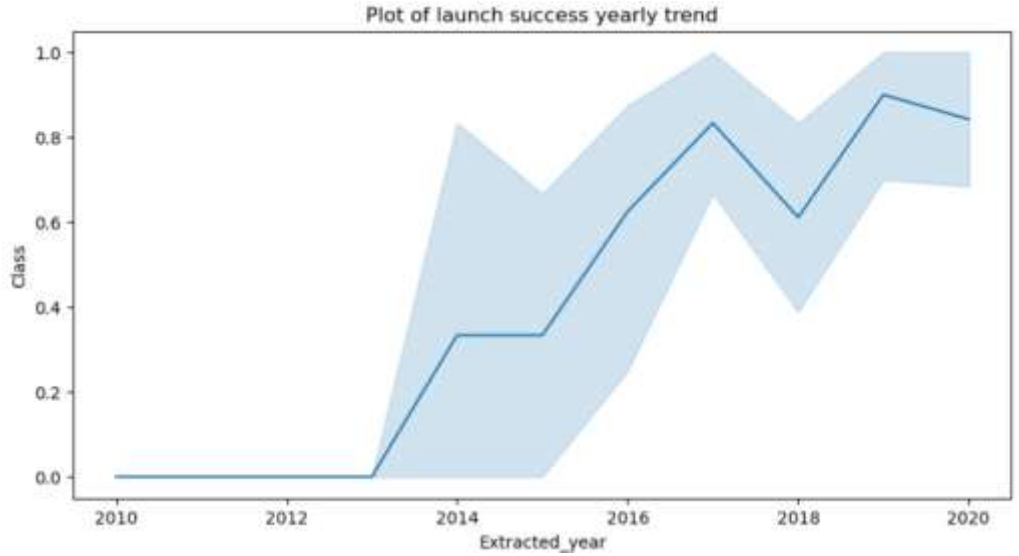
- We performed EDA using SQL to get insights from the data.

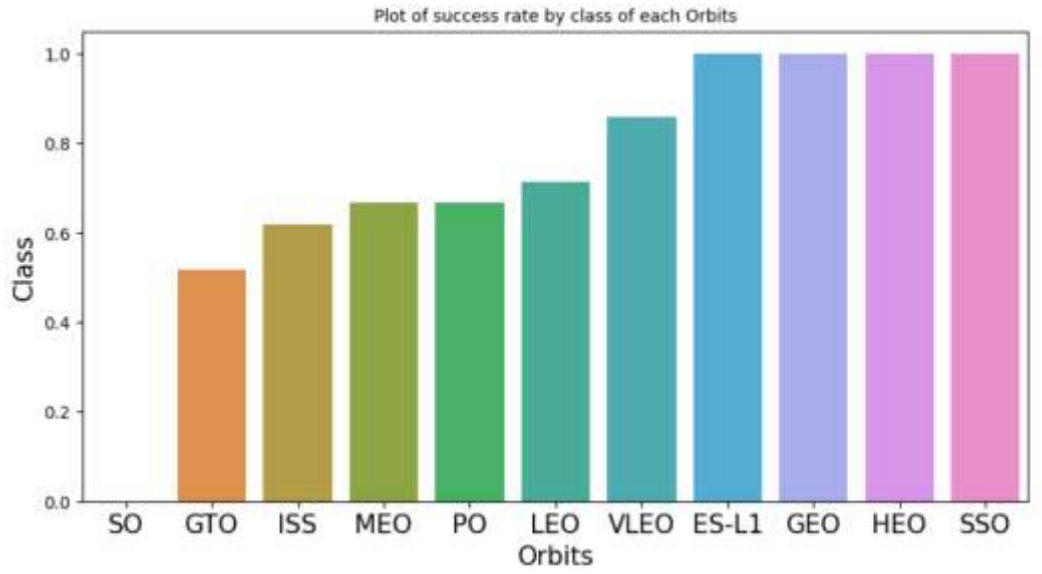
Notebook link:
https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/SQL%20EDA.ipynb

Exploratory Data Analysis – Data Visualization

- We performed EDA by visualizing the data and, more specifically, the relationships between certain variables, such as launch site, flight number, payload, orbits' success rate, orbit type, yearly launch success.



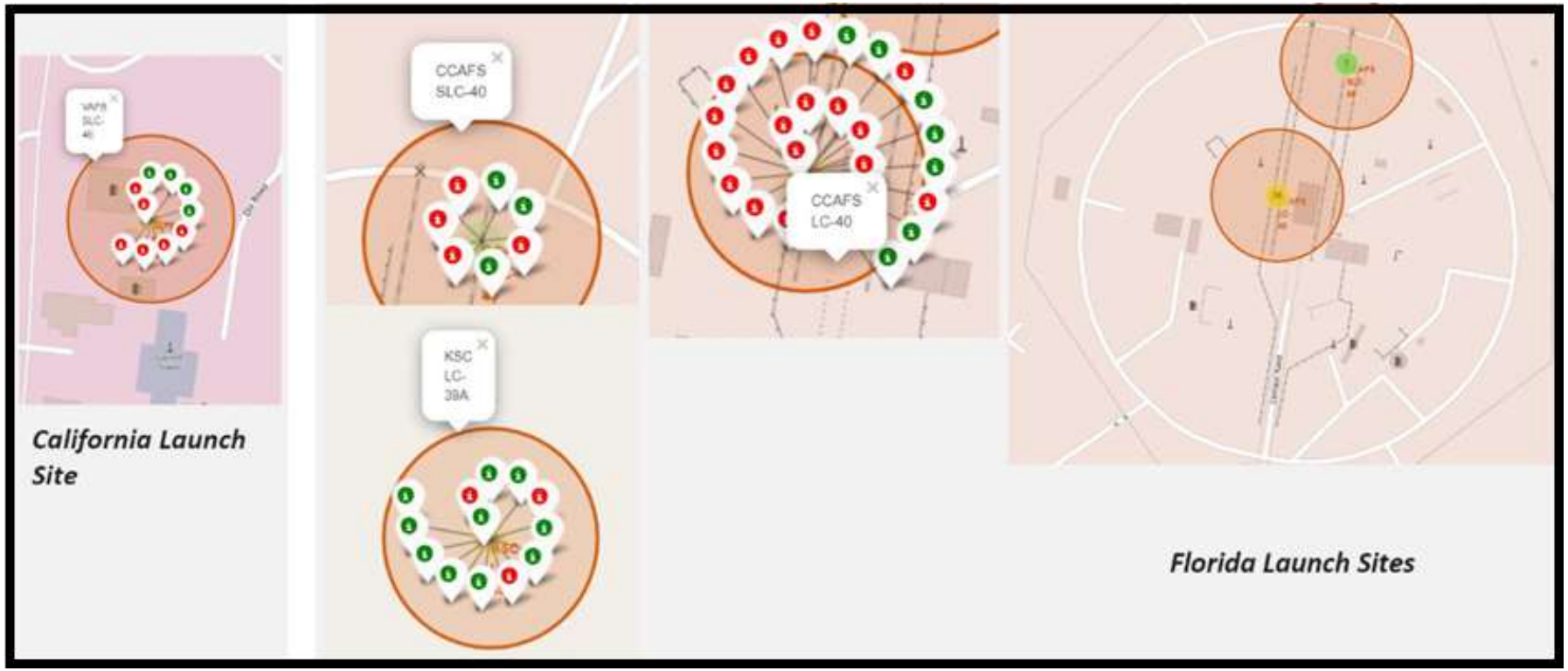




Notebook link:
https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Data%20Visualization%20EDA.ipynb

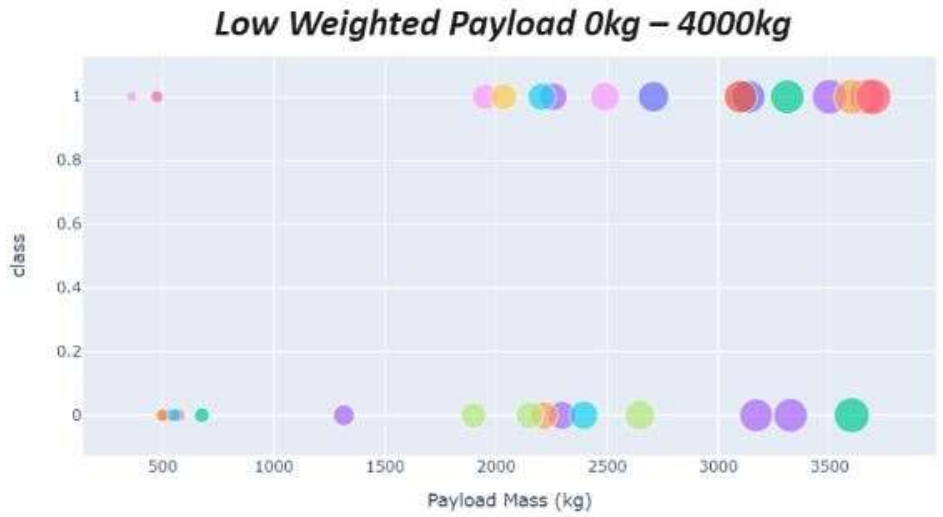
Interactive Visual Analytics – Folium

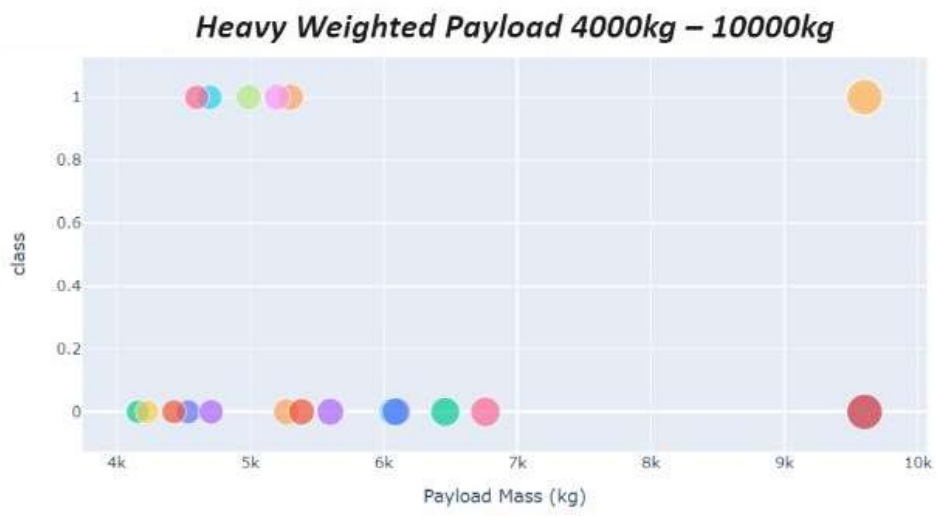
- On the folium map, we marked all of the launch sites, added map objects and marked the success or failure of launches for each site.
- Notebook link: https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

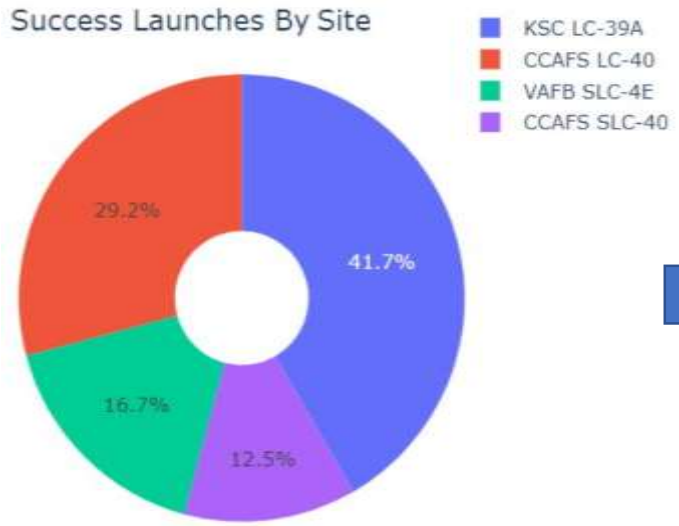


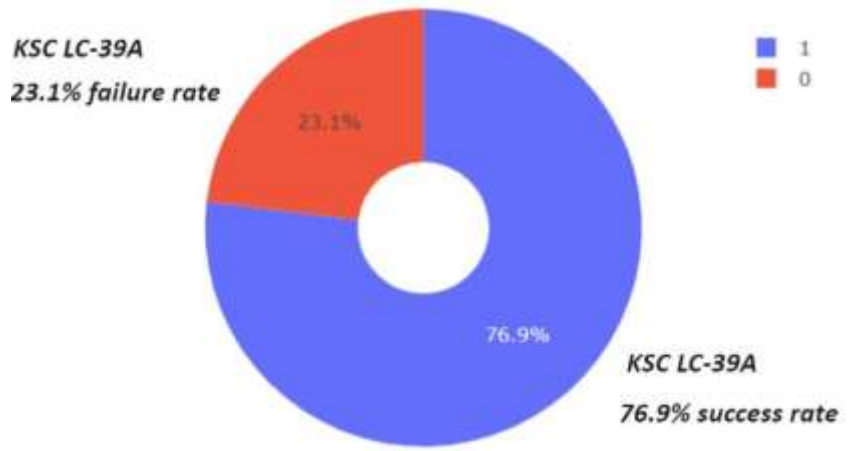
Interactive Visual Analytics – Plotly Dashboard

- On the plotly dashboard, we plotted the relationship between the outcome and the payload mass.
- We also plotted pie charts to show the success rate of the launch sites.
- Notebook link: https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/plotly_dash





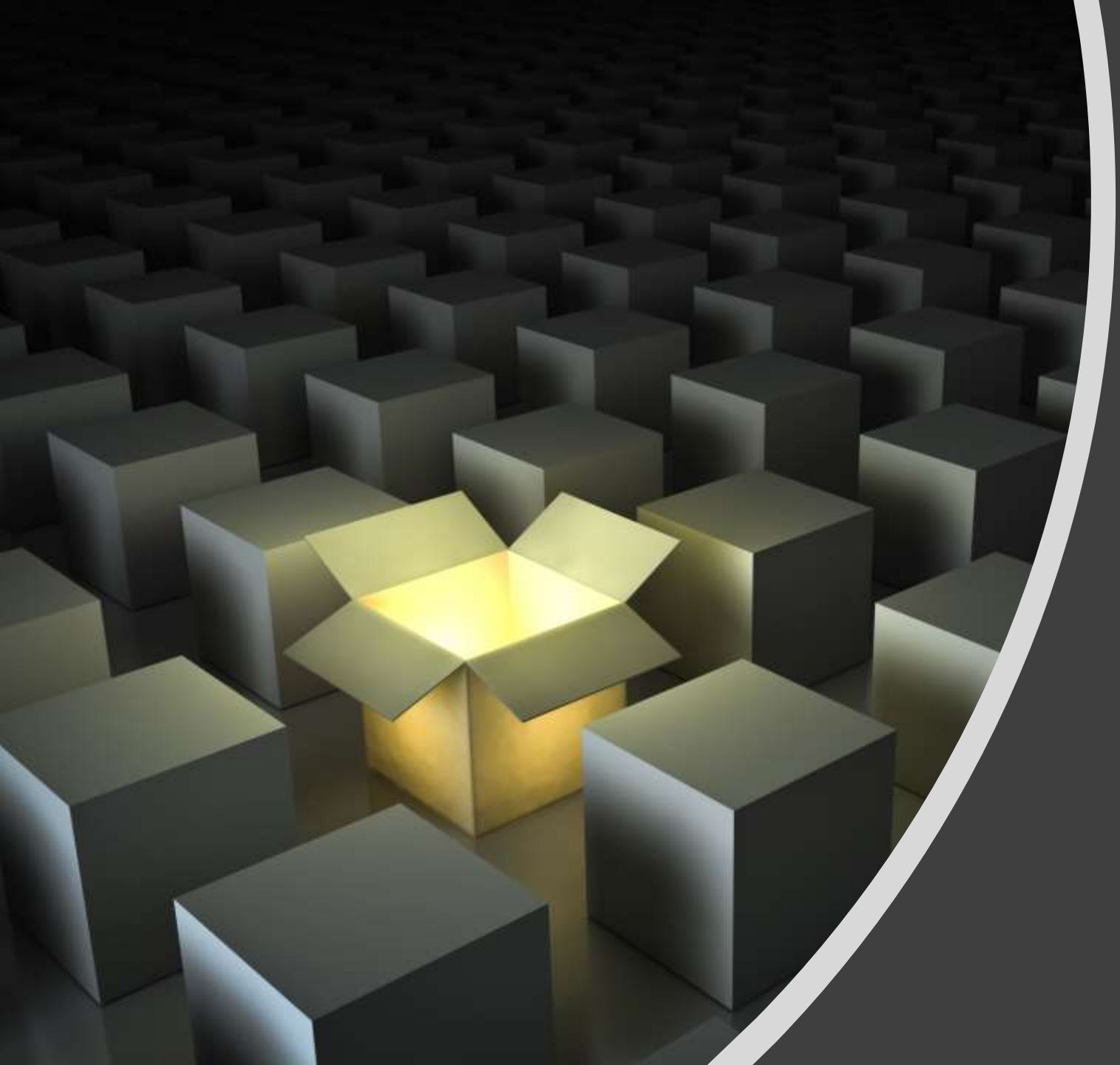




Predictive Analysis – Classification Models

- We built various machine learning models, tuned different hyperparameters and used accuracy as the metric for our model.
- We improved the model using feature engineering and algorithm tuning and we found the best performing classification model.

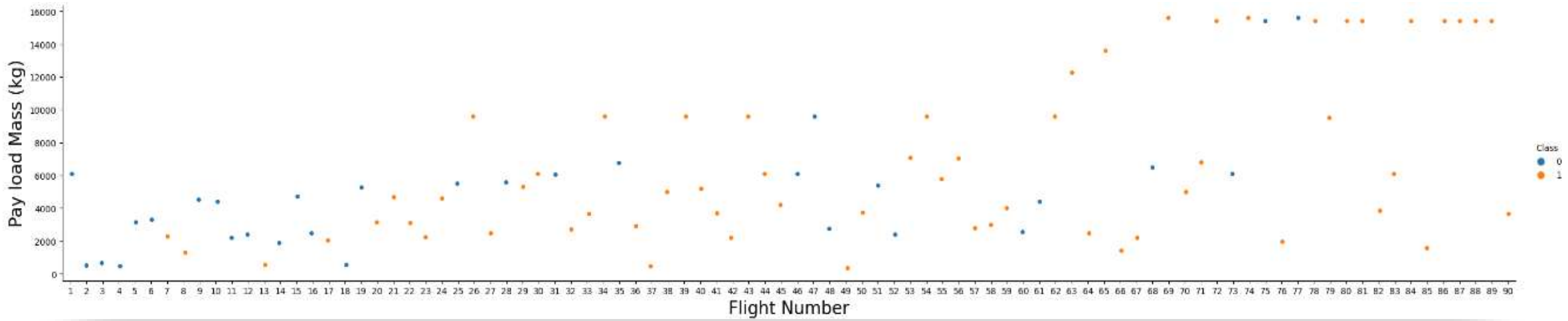
Notebook link:
https://github.com/AnastasiaChatzi/IBM_DataScience_Capstone_Project/blob/master/Machine%20Learning%20(Classification).ipynb



Results

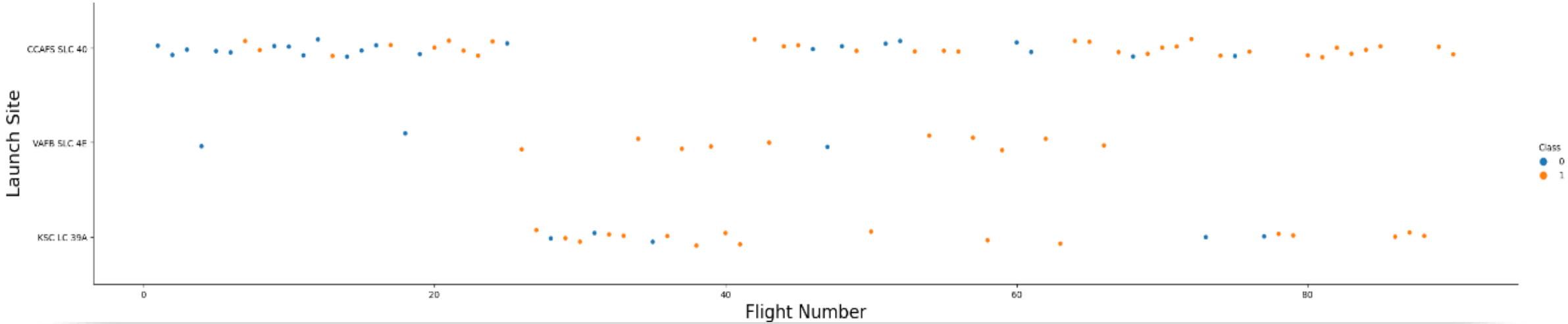
Payload Mass Vs Flight Number

- As the flight number increases, the first stage is more likely to land successfully.
- The more massive the payload mass, the less likely the first stage will return



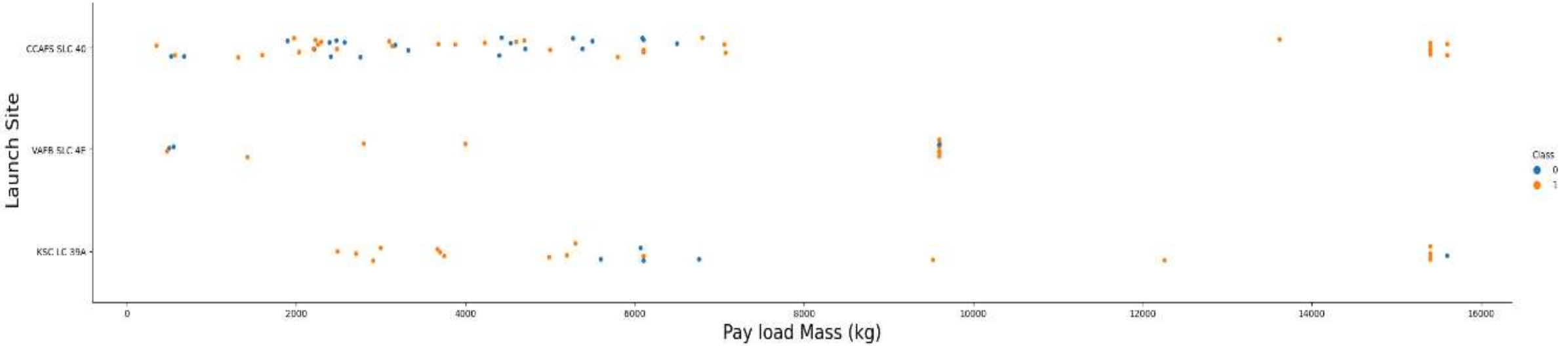
Flight Number Vs Launch Site

- The larger the flight number at a launch site, the greater its success rate.



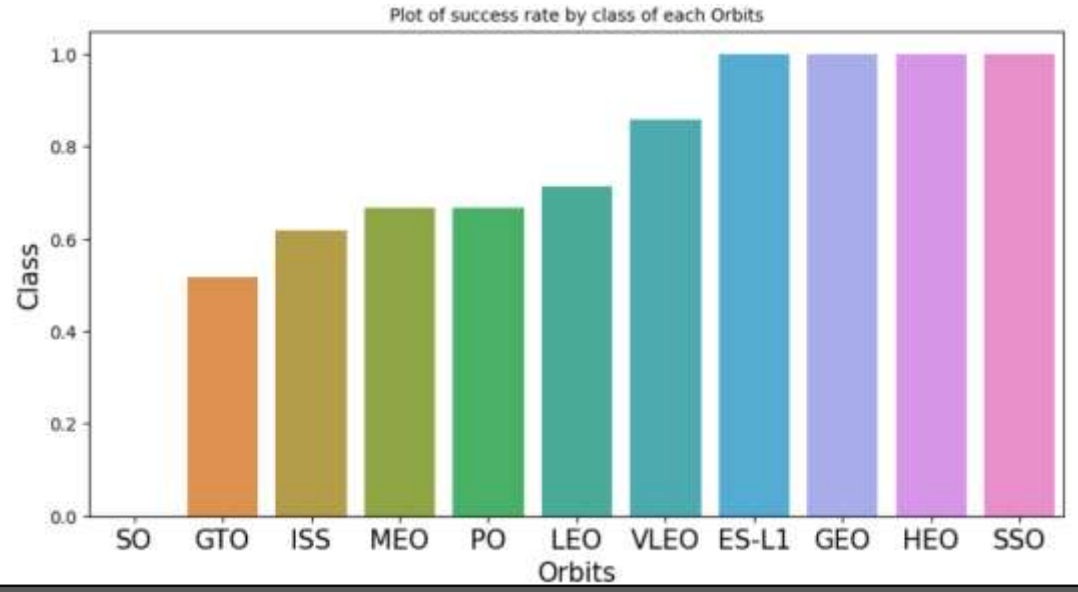
Flight Number Vs Launch Site

- The greater the payload mass at the CCAFS SLC 40 launch site, the greater its success rate.
- This seems to be applicable to the KSC LC 39A launch site as well, however only up to a certain payload mass amount.



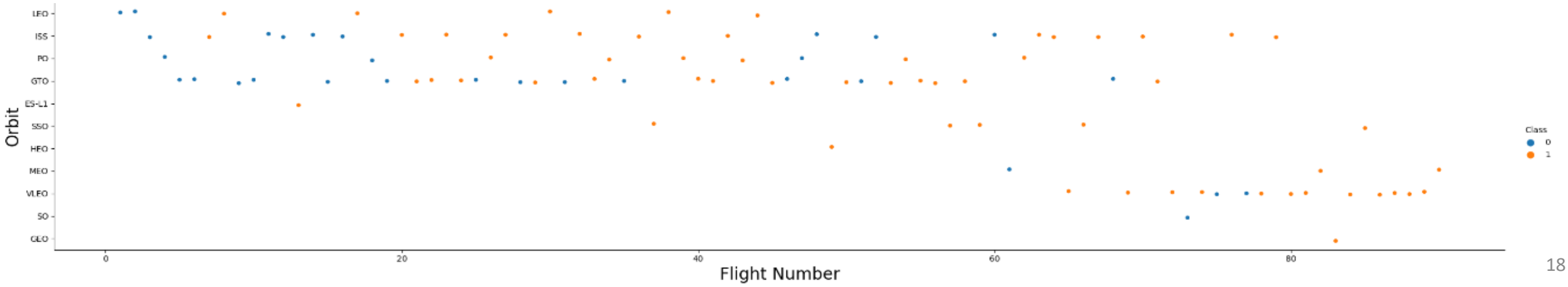
Success Rate Vs Orbit

- The orbits with the highest success rates are: SSO, HEO, GEO and ES-L1.



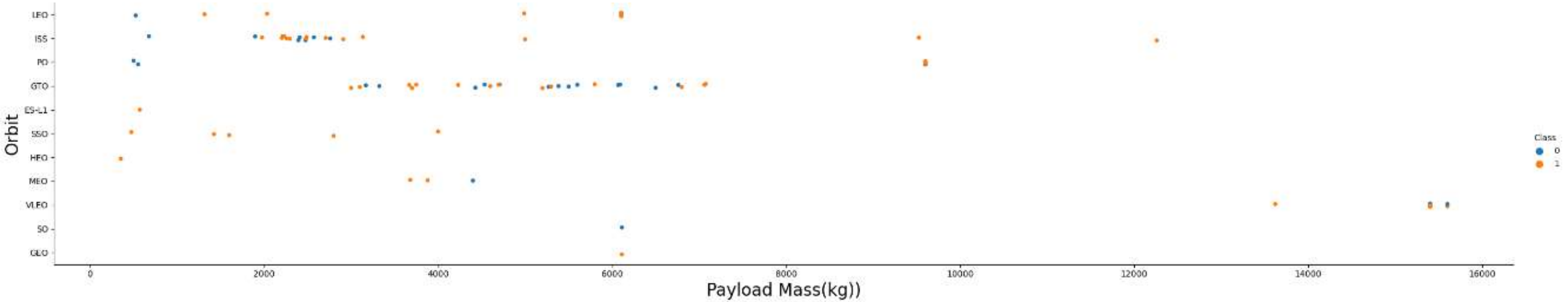
Flight Number Vs Orbit

- For the LEO and VLEO orbits, the success seems to be related to the number of flights.



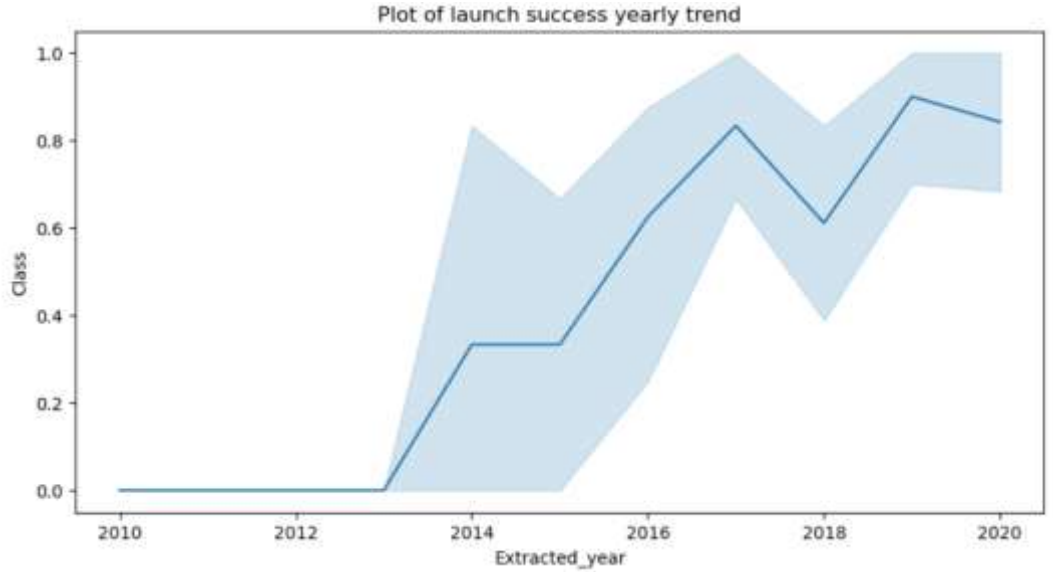
Payload Mass Vs Orbit

- The orbits LEO and ISS appear to have a higher success rate for heavy payload mass.



Success Rate Vs Orbit

- There seems to be an increasing trend for the launch success rate from 2013 until 2020.



SQL

```
SELECT DISTINCT launchsite
FROM spacex;
```

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

- Names of the unique launch sites.

```
SELECT SUM(payloadmasskg) AS Total_Payload_Mass_kg
FROM spaceX
WHERE customer='NASA(CRS)';
```

	Total_Payload_Mass_kg
0	45596

- Total payload mass carried by boosters launched by NASA (CRS).

SQL

```
SELECT AVG(payloadmasskg) AS Average_Payload_Mass_kg
FROM spaceX
WHERE boosterversion='F9 c1.1';
```

Average_Payload_Mass_kg
2928.4

- Average payload mass carried by booster version F9 v1.1

```
# Failed missions
SELECT COUNT(missionoutcome) AS Failed_Missions
FROM spaceX
WHERE missionoutcome='Failure';

# Successful missions
SELECT COUNT(missionoutcome) AS Successful_Missions
FROM spaceX
WHERE missionoutcome='Success';
```

	Failed_Missions
0	1

The total number of failed mission outcome is:

Successful_Missions
100

- Total number of successful and failure mission outcomes.

SQL

```
SELECT boosterversion, payloadmasskg
FROM spaceX
WHERE payloadmasskg=(
                     SELECT MAX(payloadmasskg)
                     FROM spaceX
                     )
ORDER BY boosterversion;
```

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

- Names of the booster versions that have carried the maximum payload mass.

```
SELECT COUNT(landingoutcome)
FROM spaceX
WHERE landingoutcome LIKE 'Success%' AND (date BETWEEN '2010-06-04' AND '2017-03-20')
ORDER BY COUNT(landingoutcome) DESC;
```

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- Landing outcomes in descending order between 2010-06-04 and 2017-03-20.

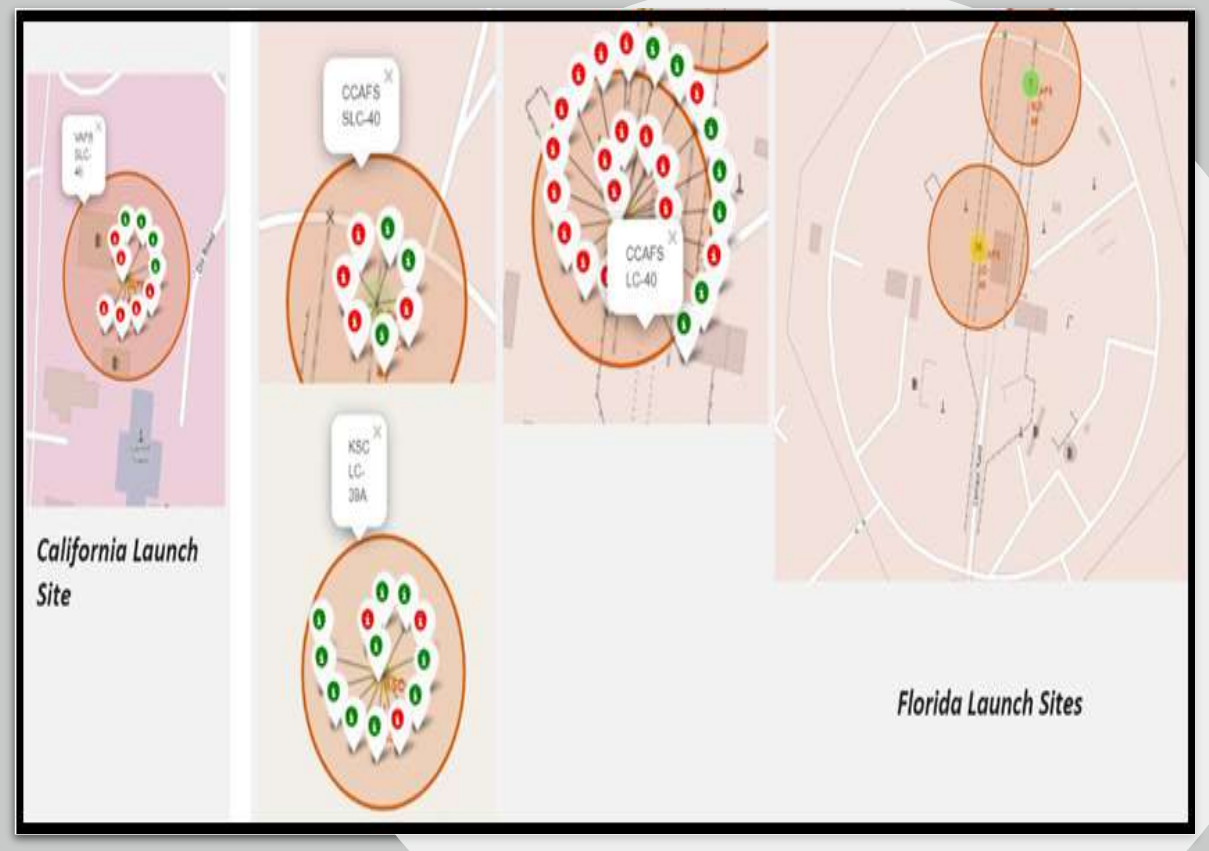
Launch Sites Proximity Analysis

SpaceX Launch Sites

They are located in the USA coasts, Florida and California.

Launch sites' colour labels:

- Green for successful launches
- Red for failed launches



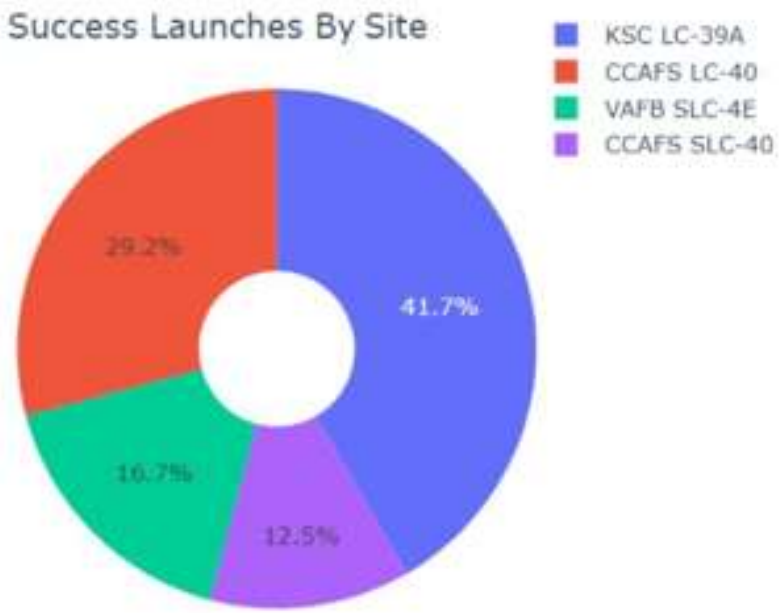
Launch Site distance to landmarks

Launch sites are:

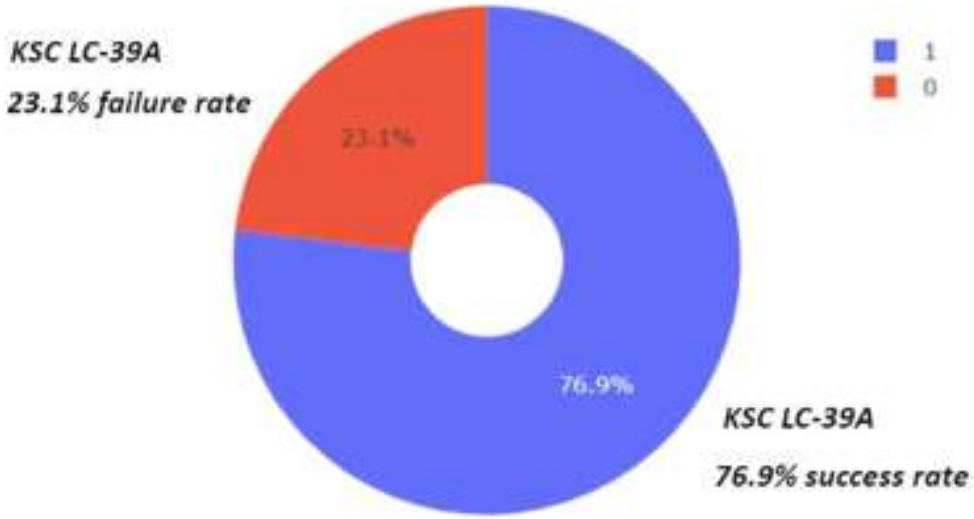
- **Not in close proximity to railways**
- **Not in close proximity to the highways**
- **In close proximity to the coastlines**
- **In certain distance from the cities**

Plotly Dash Dashboard

Success rate of each launch site

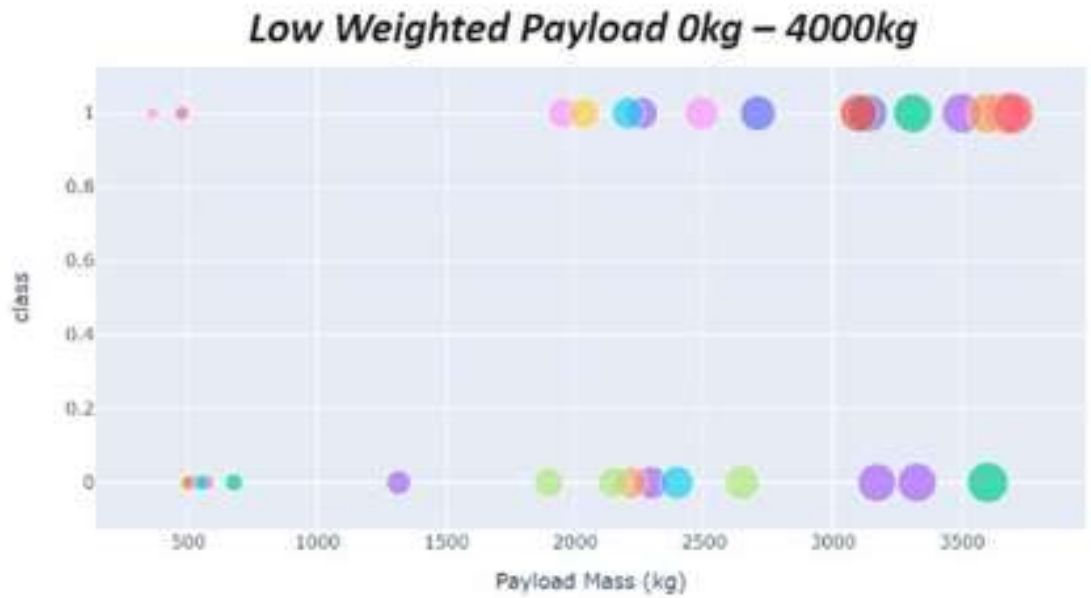


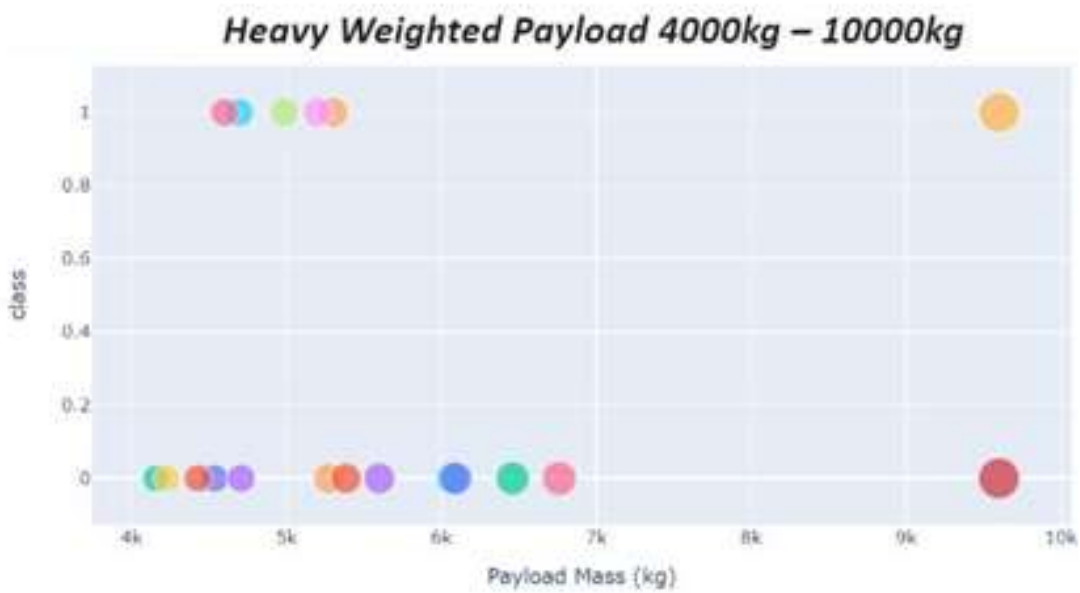
KSC LC-39A launch site appears to have the highest percentage of successful launches



Launch success ratio of the KSC LC-39A site

Payload Vs Launch Outcome





The success rate appears to be higher when the payload is lighter.

Predictive Analysis (Classification)

Classification Accuracy

The model with the highest classification accuracy is the **decision tree**:

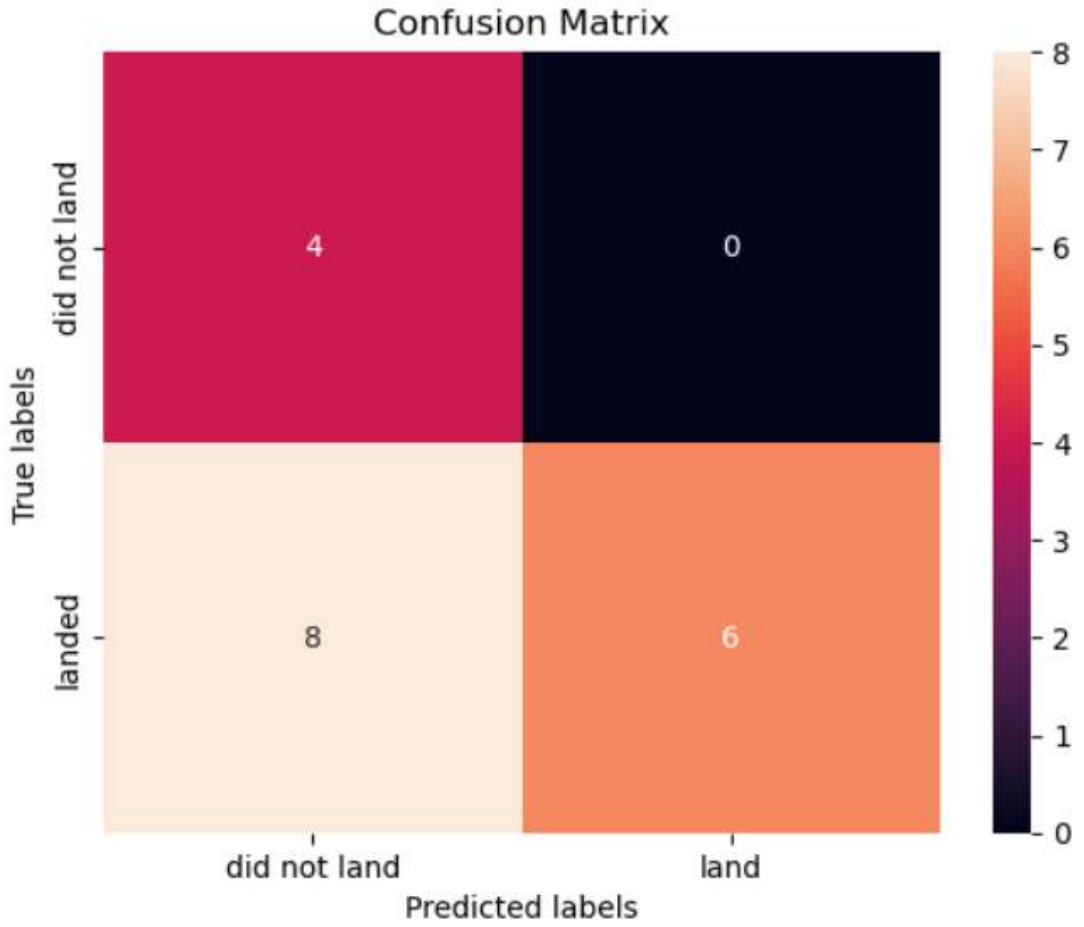
```
all_models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(all_models, key=all_models.get)
print('Best model is', bestalgorithm,'with a score of', all_models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142857
Best params is : {'criterion': 'entropy', 'max_depth': 2, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

The confusion matrix appears to have some issues regarding the false positives and negatives:

- False Negative error (Type II) : incorrectly predicted failed landings as successful.
- False Positive error (Type I) : inability to predict successful landings



Conclusion

- The launch success rate has increased from 2013 to 2020.
- The larger the flight amount of a launch site, the greater its success rate.
- The more massive the payload mass, the less likely the first stage will return
- Orbits SSO, HEO, GEO, ES-L1 appear to have the biggest success rate.
- KSC LC-39A is the launch site with the most successful launches.
- The most accurate machine learning algorithm to predict the success of the first stage is the Decision tree classifier.

Thank you