IBM Data Science Capstone Project – Space X

TABLE OF CONTENTS

INTRODUCTION

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers (which usually cost upwards of $165m), and much of the savings are because SpaceX can land, and then re-use the first stage of the rocket.

- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether or not an alternate company should bid and SpaceX for a rocket launch.

- This project will ultimately predict if the Space X Falcon 9 first stage will land successfully.

METHODOLOGY

1. Data Collection Data Wrangling
2. EDA with SQL
3. EDA with Visualization
4. Interactive Visual Analytics
5. Predictive Analysis

METHODOLOGY

1. Data Collection:

WEB SCRAPING :

Web scraping spacex launches from wikipedia and converting it into a dataframe .

DATA COLLECTION

Step 1:

```
# use requests.get() method with the provided static_url
# assign the response to a object
req=requests.get(static_url)
req=req.text
req
```

Get Contents from webpage.

Step 2:

```
column_names = []
labels = first_launch_table.find_all('th')
for label in labels:
    name = extract_column_from_header(label)
    if name != None:
      if len(name) > 0:
          column_names.append(name)
```

Loop through all ‘th’ elements in the html file and extract the column names.

Step 3:

Pull the data for each column and convert the data into a Data Frame.

Step 4:

```
df.to_csv('spacex_web_scraped.csv', index=F
```

Save the data into CSV file

DATA WRANGLING

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

We will mainly convert outcomes into training labels with 1 means the booster landed successfully and 0 means booster landing was unsuccessful.

1

```
df.isnull().sum()/df.shape[0]*100
```

Find any column has null values and replace them.

2

Using value_counts() method, calculate:

1. No. of launches on each site
2. No and occurrence of each orbit
3. Mission Outcome occurrence

3

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class=[]
for i in range(0,90):
    outcome=df.iloc[i,6]
    if outcome in bad_outcomes:
        landing_class.append('0')
    else:
        landing_class.append('1')

landing_class
```

Create a landing class column with the outcome as 0 and 1 for unsuccessful and successful landing respectively

4

Save the dataframe into a csv

```
df.to_csv("dataset_part_2.csv", index=False)
```

EXPLORATORY DATA ANALYSIS USING SQL

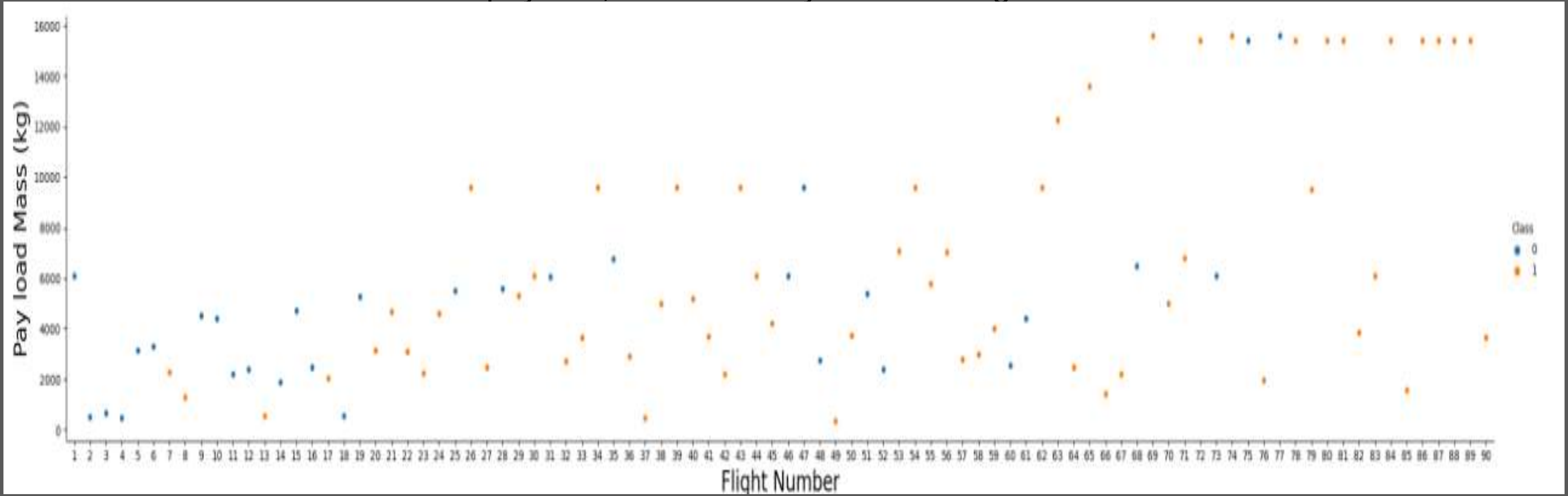
To gather some information about the dataset, some SQL queries were performed.

The SQL queries performed on the data set were used to:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display the average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome on a ground pad was achieved
6. List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
7. List the total number of successful and failed mission outcomes
8. List the names of the booster versions which have carried the maximum payload mass
9. List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

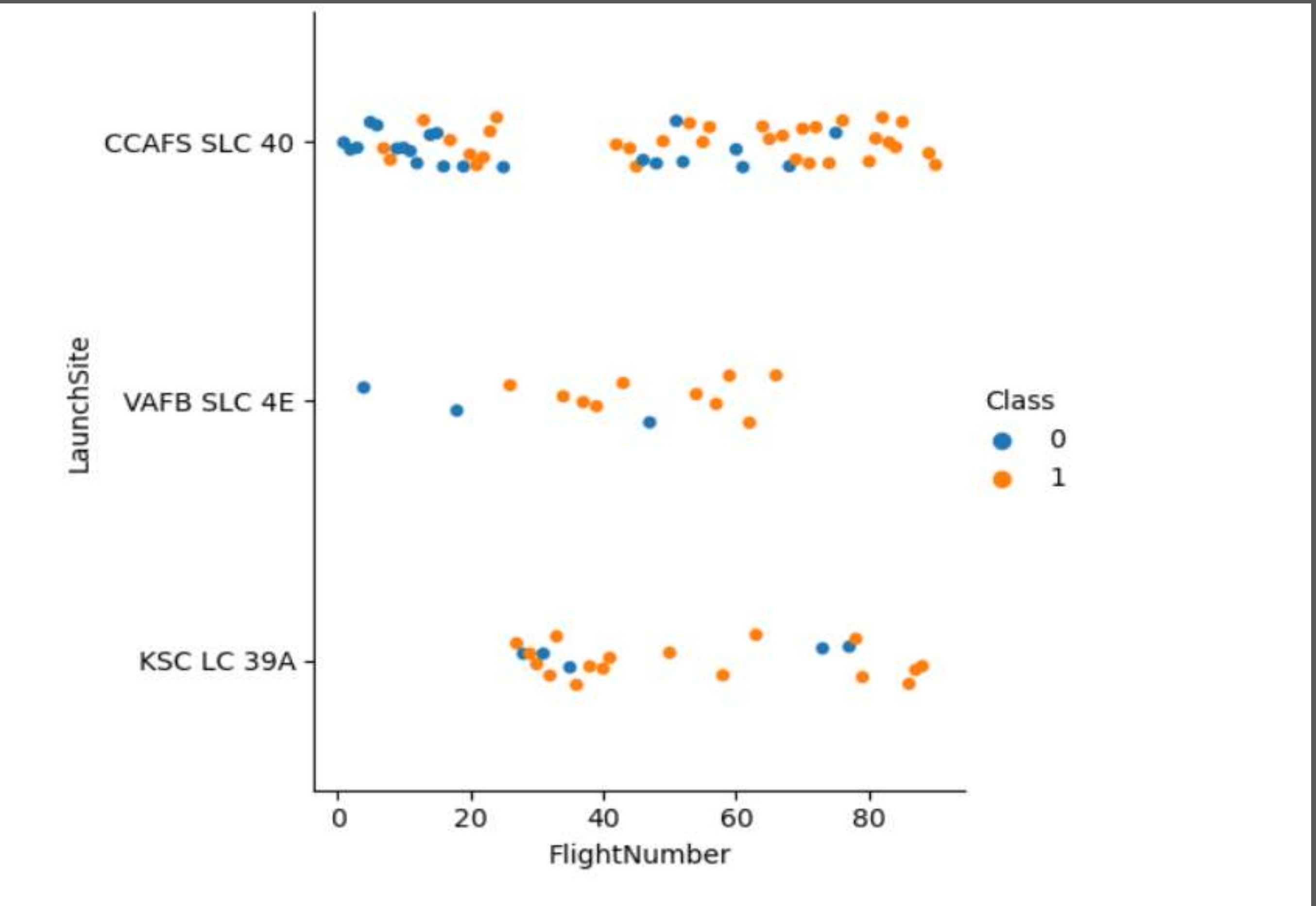
EXPLORATORY DATA ANALYSIS USING VISUALIZATION

Plot out the FlightNumber vs. PayloadMassand overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important;
it seems the more massive the payload, the less likely the first stage will return.



PLOT BETWEEN LAUNCH SITE AND FLIGHT NUMBER:

```
sns.catplot(y='LaunchSite',x='FlightNumber',hue='Class',data=df)
plt.xlabel('FlightNumber')
plt.ylabel('LaunchSite')
plt.show()
```

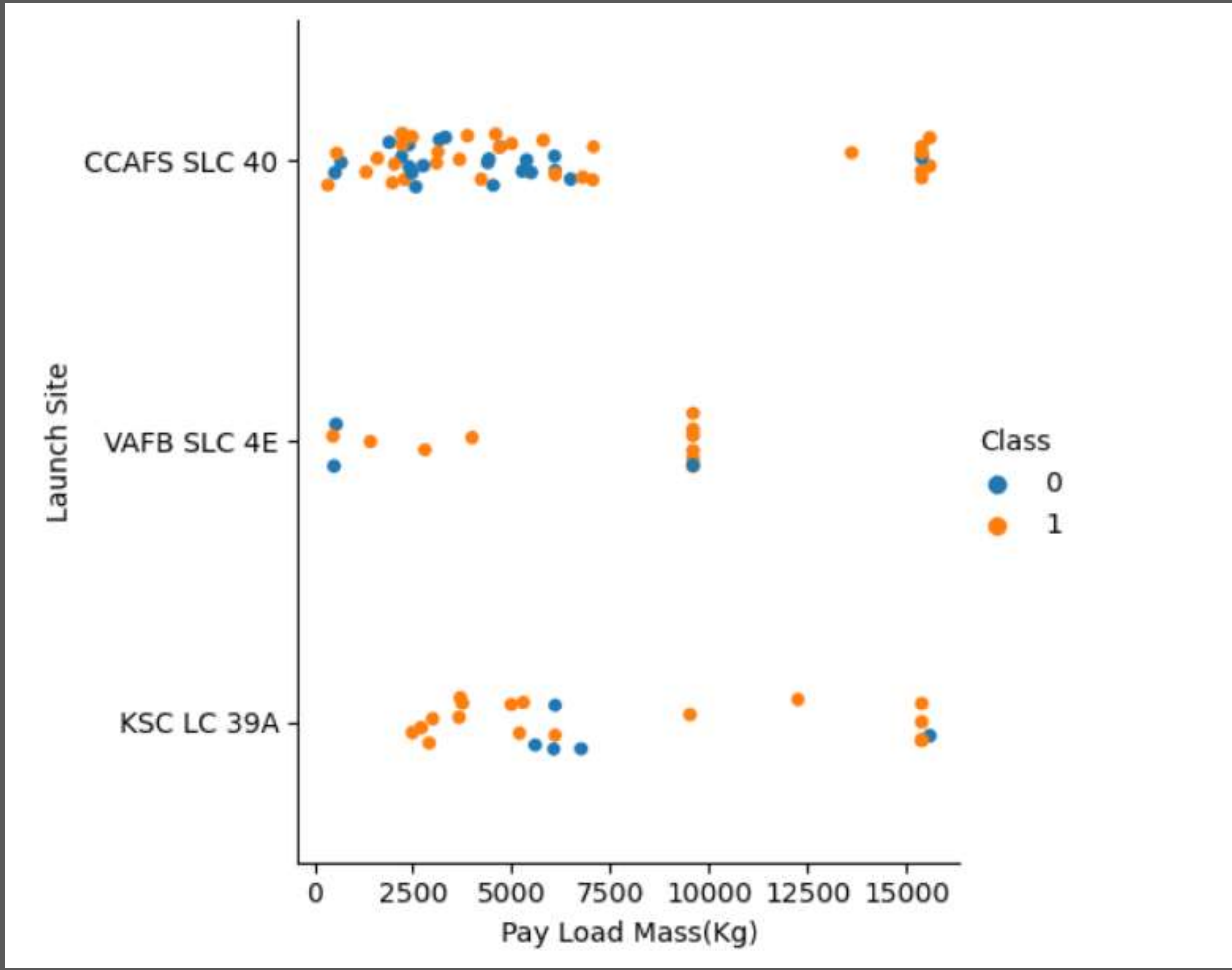


Inferences from plot:

- As the number of flights increases, the rate of success at a launch site increases.
- Most of the early flights (flight numbers < 30) were launched from CCAFS SLC 40, and were generally unsuccessful.
- The flights from VAFB SLC 4E also show this trend, that earlier flights were less successful.
- No early flights were launched from KSC LC 39A, so the launches from this site are more successful.
- Above a flight number of around 30, there are significantly more successful landings (Class = 1).

PLOT BETWEEN LAUNCH SITE AND PAYLOADMASS:

```
sns.catplot(x='PayloadMass',y='LaunchSite',data=df,hue='Class')
plt.xlabel('Pay Load Mass(Kg)')
plt.ylabel('Launch Site')
plt.show()
```



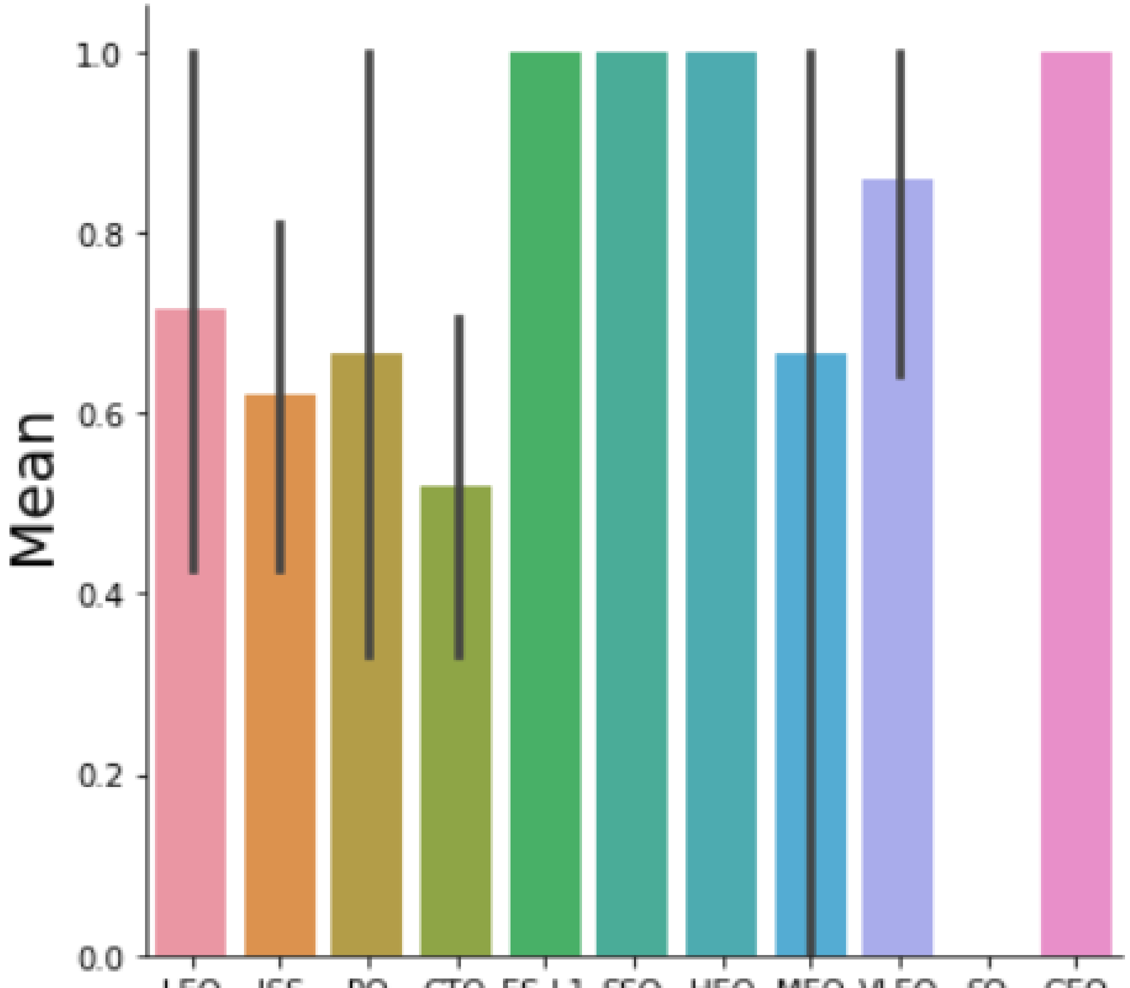
Inferences from plot:

The scatter plot of Launch Site vs. Payload Mass shows that:

- Above a payload mass of around 7000 kg, there are very few unsuccessful landings, but there is also far less data for these heavier launches.
- There is no clear correlation between payload mass and success rate for a given launch site.
- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 being comparatively lighter payloads (with some outliers).

SUCCESS RATE OF ORBIT:

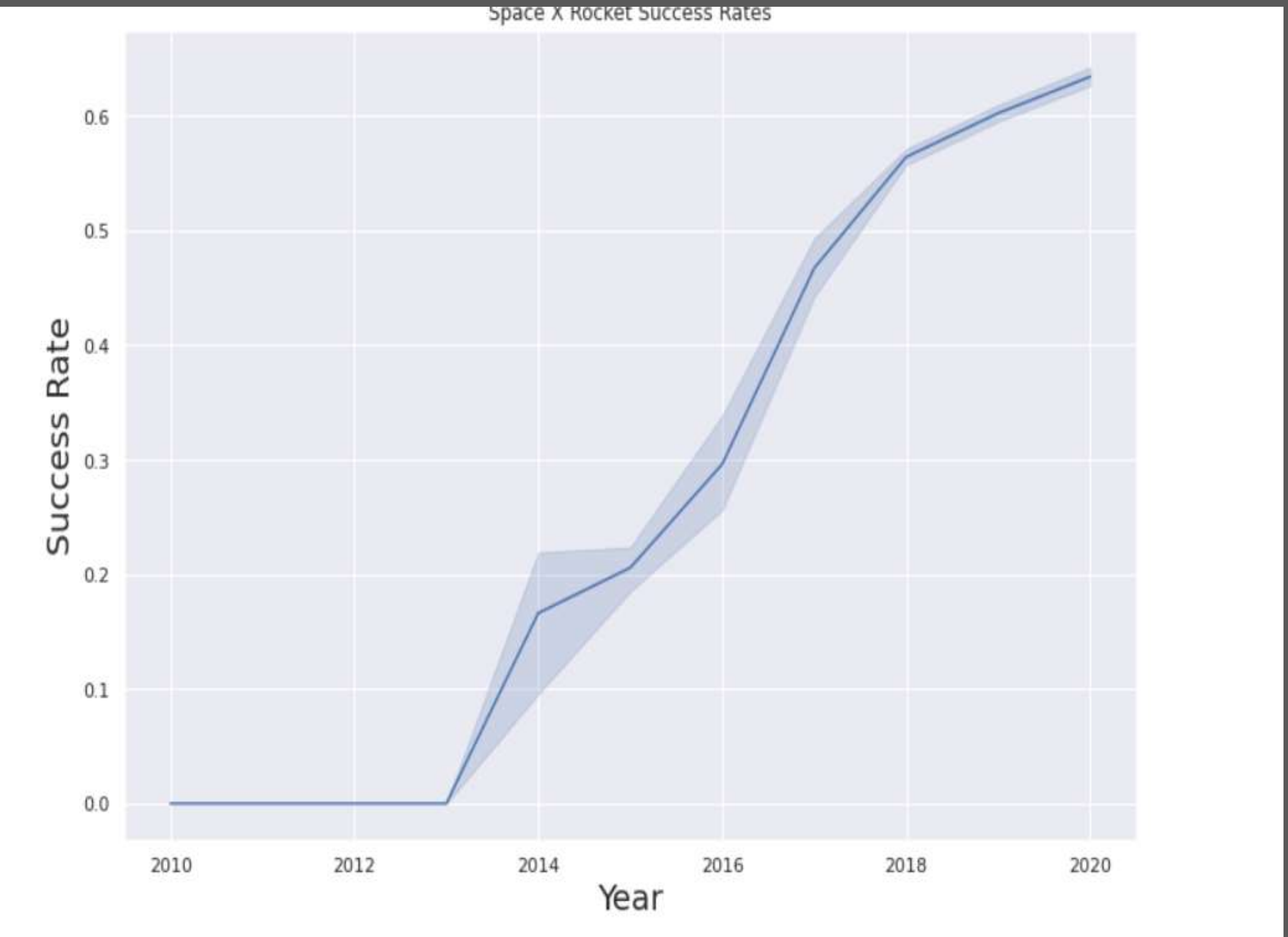
```
df.groupby(['Orbit']).mean()
sns.catplot(x="Orbit",y="Class", kind="bar",data=df)
plt.xlabel("Orbit",fontsize=20) plt.ylabel("Mean",fontsize=20)
plt.show()
```



Inferences from plot:

From the visualization we can conclude that, GEO,HEO,Sso and ES-L1 orbits have higher success rates than other orbits.

SUCCESS RATE OF ROCKETS:



Inferences from plot:

From the visualization, it is evident that Rate of success is increasing over time .

LOCATION ANALYSIS USING FOLIUM

Here we visualize:

1. Launch sites on the map.
2. Mark success and failure launches for each site.
3. Cal distance between launch sites and proximities.

LOCATION OF LAUNCH SITES

VISUALIZE LAUNCHES ON MAP

PREDICTIVE ANALYSIS

Here we perform predictive Analysis for the data.
We are using the following Machine Learning models to predict data:

1. Logistic Regression
2. Decision Tree
3. SVM
4. K-Nearest Neighbours

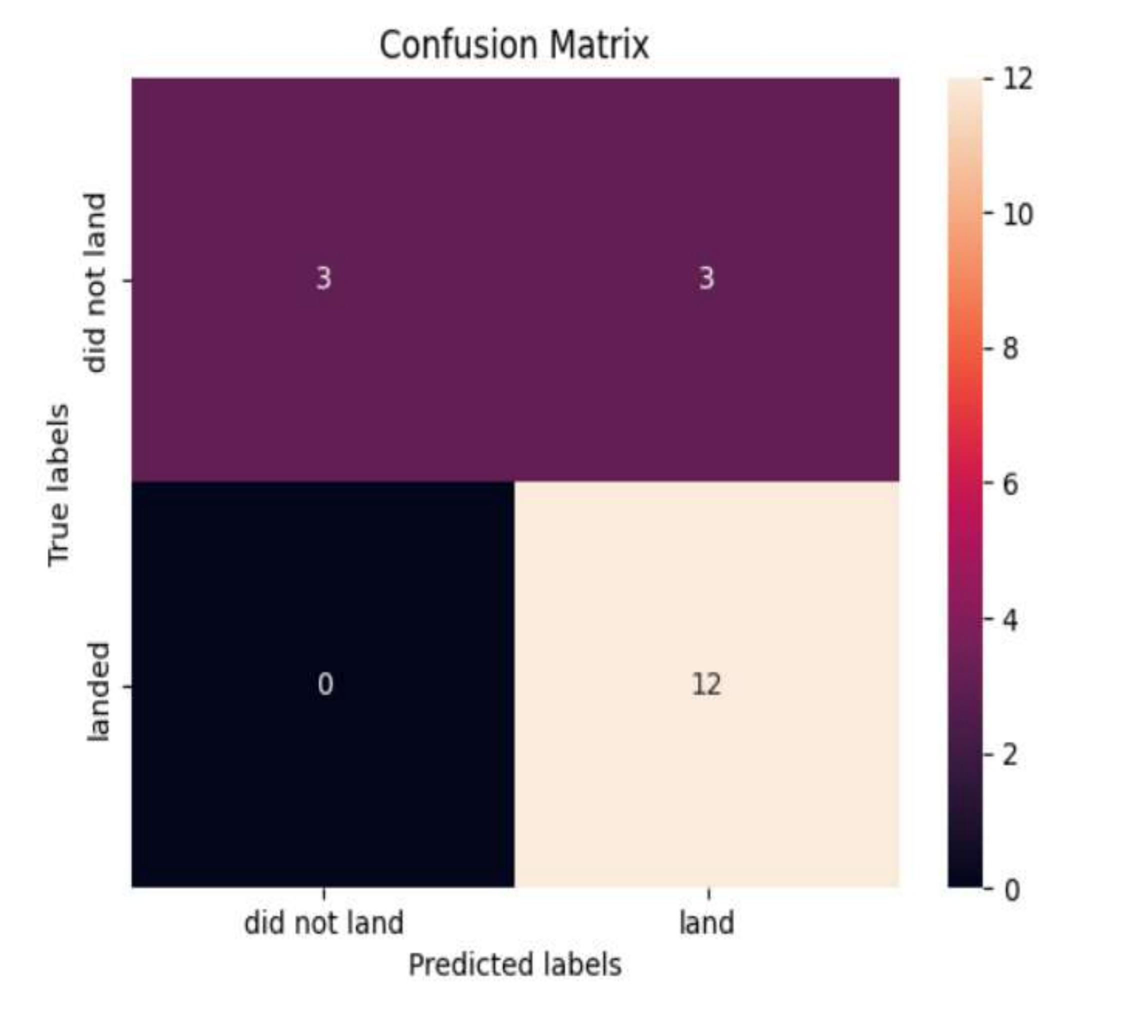
Models are tuned for best performance provide insights on the probability of a launch being a success or a failure.

BEST MODEL

From the different models used we see that KNN, Logistic and SVM models have a accuracy of 95% to predict the data.

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()

knn_cv=GridSearchCV(KNN,parameters,cv=10)
knn_cv.fit(X,Y)

print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)

yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

CONFUSION MATRIX
KNN MODEL



CONCLUSION

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success.

- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.

  The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
  The 100% success rate in SSO is more impressive, with 5 successful flights.
  The orbit types PO, ISS, and LEO, have more success with heavy payloads:
  VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.

- The success for massive payloads (over 4000kg) is lower than that for low payloads.

- The best-performing classification model is the KNN model, with an accuracy of 94.44%.