Conquering the Space with Data Science

Verónica Andrade

10 April 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results & Discussion
- Conclusion

Executive Summary

- ✓ Summary of methodologies
- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

- ✓ Summary of all results
- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

Background and Context

There are several companies which provide rocket lunches for different proposes and their prices are in a range of 60 to 200 millions dollars. One company in particular, SpaceX, offers rocket lunches (Falcon 9) with a cost of 62 million dollars. This is possible because they can reuse the first stage.

Consequently, predicting if the rocket in first stage will land is an assent to determine the cost of the lunch.

Objective

This project looks forward to create a machine learning pipeline to predict if the first stage will land successfully.

Questions

Which factors are determining for the rocket to land successfully?

Which variables are relevant to determine the success rate of a successful landing?

What are the operating conditions to ensure a qualified landing program?

TOOLS FOR THE CONQUERING

Methodology

Data Collection

We want to extract launch records as HTML table, analize the table and convert to a panda dataframe.

i. Data was collected using SpaceX API and web scraping from Wikipedia.

ii. Decoding the response content as Json using . json() function call and turn it into a pandas dataframe using .json_normalize().

```
In [6]: spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
In [7]: response = requests.get(spacex_url)
```

Check the content of the response

```
In [8]: print(response.content)
```

```
In [9]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_
```

We should see that the request was successfull with the 200 status response code

```
In [10]: response.status_code
Out[10]: 200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
In [11]: # Use json_normalizre meethod to convert the json result into a dataframe

response.json()

data=pd.json_normalize(response.json())
```

https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Methodology

Data Collection

iii. Clean the data and check for missing values

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

In [36]:
```
data_falcon9 = df[df['BoosterVersion']!='Falcon 1']
```

Now that we have removed some values we should reset the FlgihtNumber column

In [37]:
```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

iv. Wrangling and formatting data

In [39]:
```
# Calculate the mean value of PayloadMass column
payloadmassmean = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, payloadmassmean, inplace=True)
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages/pandas/core/generic.py:6619: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return self._update_inplace(result)
```

You should see the number of missing values of the `PayLoadMass` change to zero.

In [40]:
```
#to see missing values
data_falcon9.isnull().sum()
```

https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Methodology

Data Collection

v. Web scrapping

```
In [59]:
# use requests.get() method with the provided static_url
# assign the response to a object

response = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

```
In [60]:
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content

soup = BeautifulSoup(response)
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
In [61]:
# Use soup.title attribute
soup.title

Out[61]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

vi. Wrangling and formatting data

```
In [69]:
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

```
In [70]:
df=pd.DataFrame(launch_dict)
df.head()
```

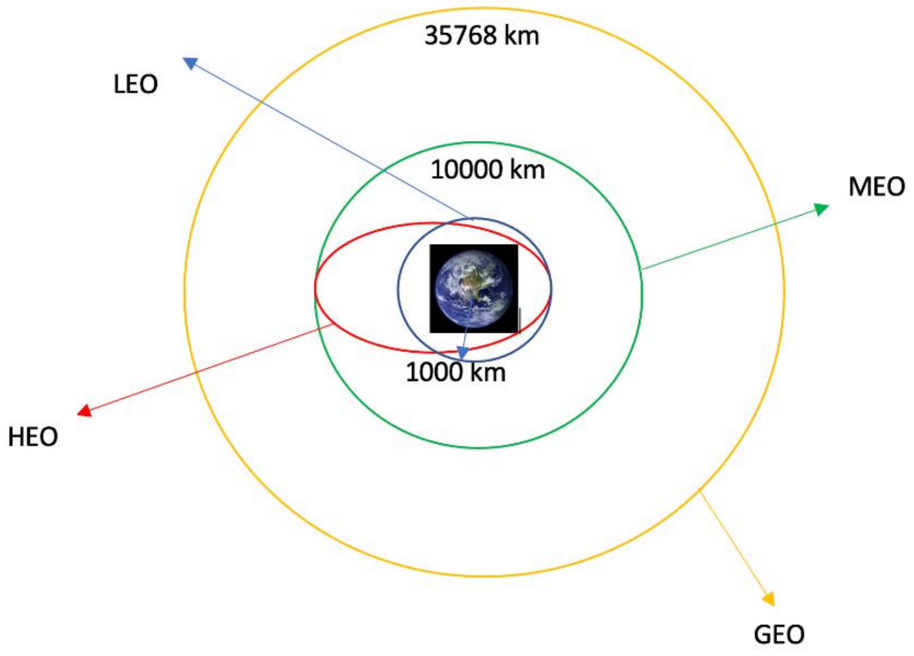
https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-webscraping.ipynb

Methodology

Data Wrangling

We want to perform exploratory data analysis and determine training labels.

- Calculations of the number of launches on each site where perform.
- Calculations of the number and occurrence of each orbit.
- Calculations of the number and occurrence of mission outcome of the orbits.
- Creation of a landing outcome label from Outcome column.



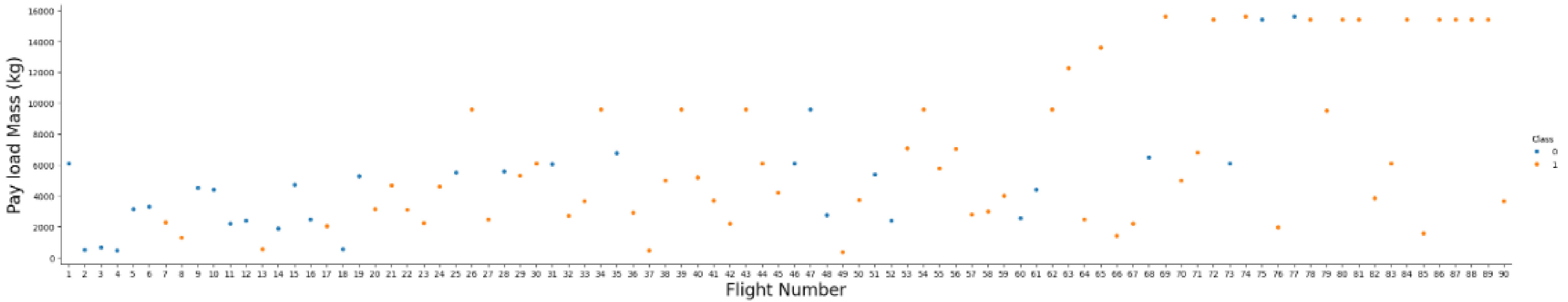
https://github.com/veronicaestefania/The-Certificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

Methodology

Exploratory data analysis (EDA)

Important to read the SpaceX dataset into a Pandas dataframe and see how the FlightNumber (indicating the continuous launch attempts.) and Payloavariables would affect the launch outcome.

In [7]:
```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



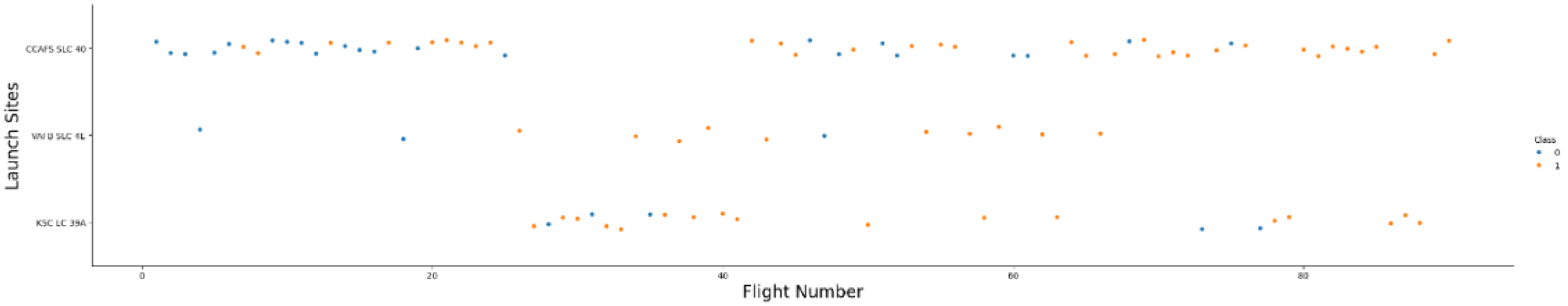
https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Methodology

Exploratory data analysis (EDA)

Visualize the relationship between Flight Number and Launch Site

In [8]:
```
### TASK 1: Visualize the relationship between Flight Number and Launch Site
sns.catplot(x="FlightNumber",y="LaunchSite",hue='Class',data=df, aspect=5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Sites",fontsize=20)
plt.show()
```



https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Methodology

EDA with SQL

EDA with SQL was apliyed to get insight from the data. We wrote different queries to find it out

- Names of unique launch sites in the space mission.
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Total number of successful and failure mission outcomes
- Failed landing outcomes in drone ship, their booster version and launch site names.

https://github.com/veronicaestefania/The-Certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Methodology

Interactive MAP with Folium

Visualizing optimal location for building a launch site to discover some of the factors by analyzing the existing launch site locations.

In [4]:

```
# Download and read the `spacex_launch_geo.csv`
from js import fetch
import io

URL = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/spacex_laun
resp = await fetch(URL)
spacex_csv_file = io.BytesIO((await resp.arrayBuffer()).to_py())
spacex_df=pd.read_csv(spacex_csv_file)
```

Now, you can take a look at what are the coordinates for each site.

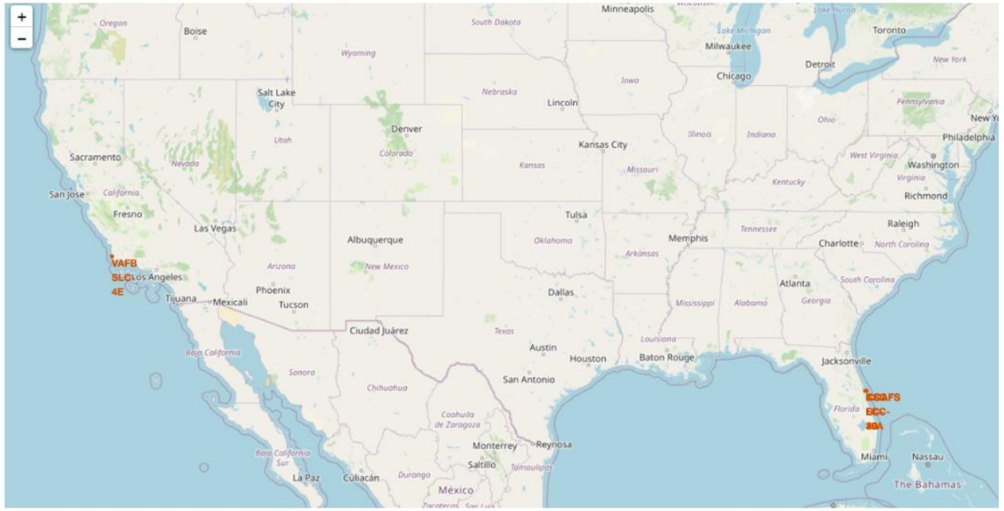
In [5]:

```
# Select relevant sub-columns: `Launch Site`, `Lat(Latitude)`, `Long(Longitude)`, `class`
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
launch_sites_df
```

In [8]:

```
# Initial the map
site_map = folium.Map(location=nasa_coordinate, zoom_start=5)
# For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name
```

The generated map with marked launch sites should look similar to the following:

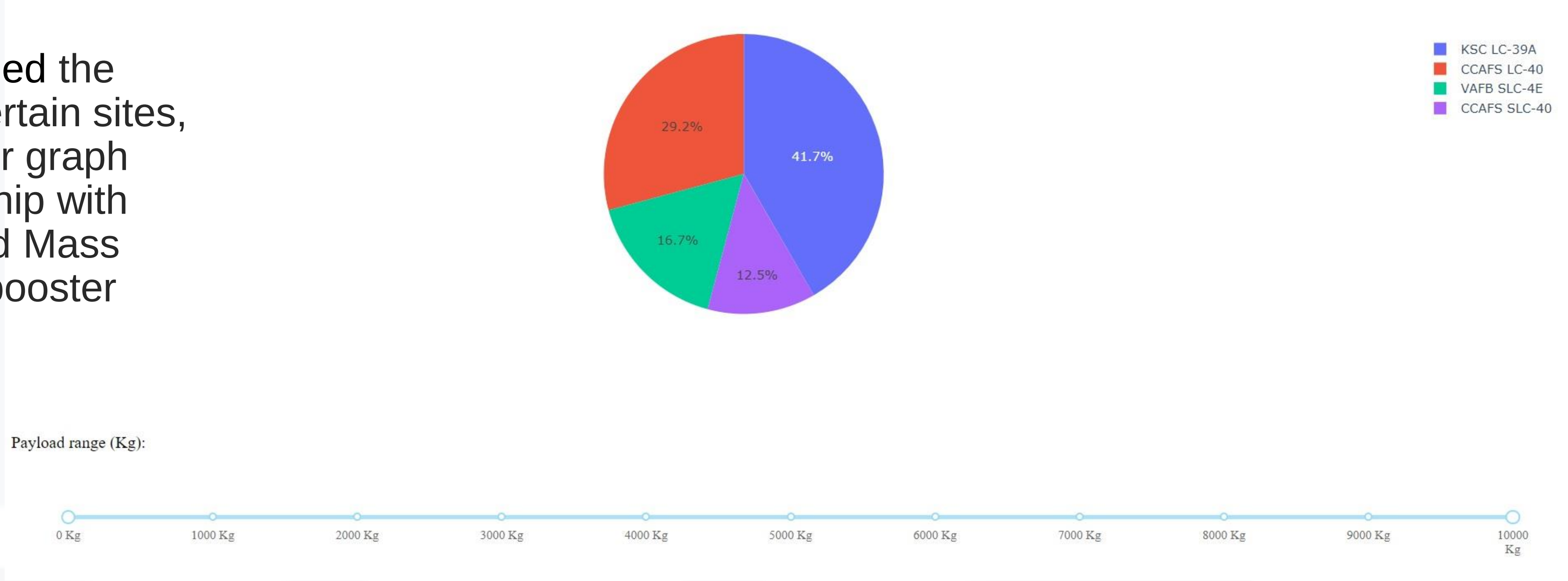


https://github.com/veronicaestefania/The-Certificate/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Methodology

Dashboard with Plothy Dash

- We wanted to visualized the total launches by a certain sites, we also plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.



https://github.com/veronicaestefania/The-Certificate/blob/main/Build%20a%20Dashboard%20Application%20with%20Plotly%20Dash

Methodology

Predictive Analysis

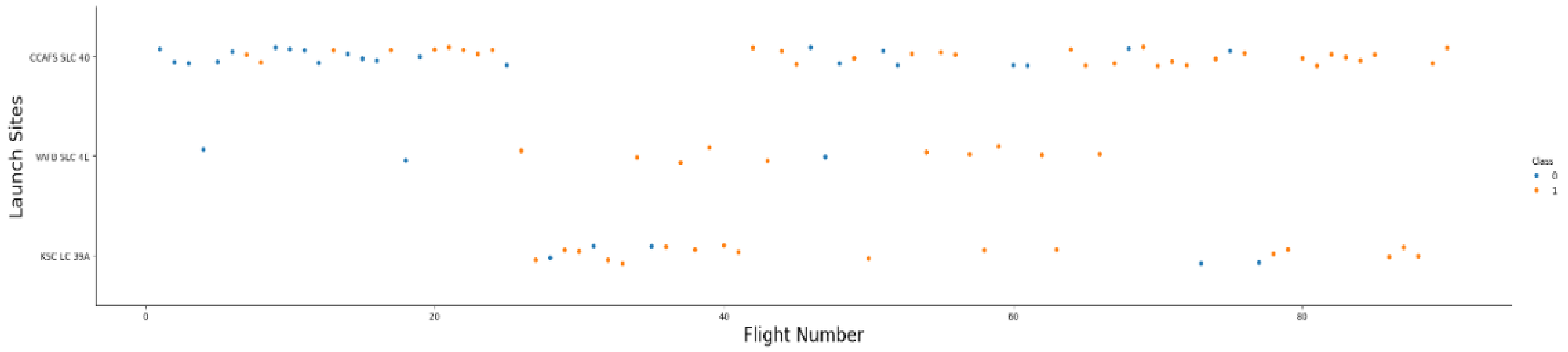
- We used GridSearchCV to stablish different hyperparameters to help us in machine learning.
- Accuracy was the metric used for the model and consequently we found the best performing classification model.

FINDINGS

Results and Discussion

Exploratory Data Analysis (EDA)

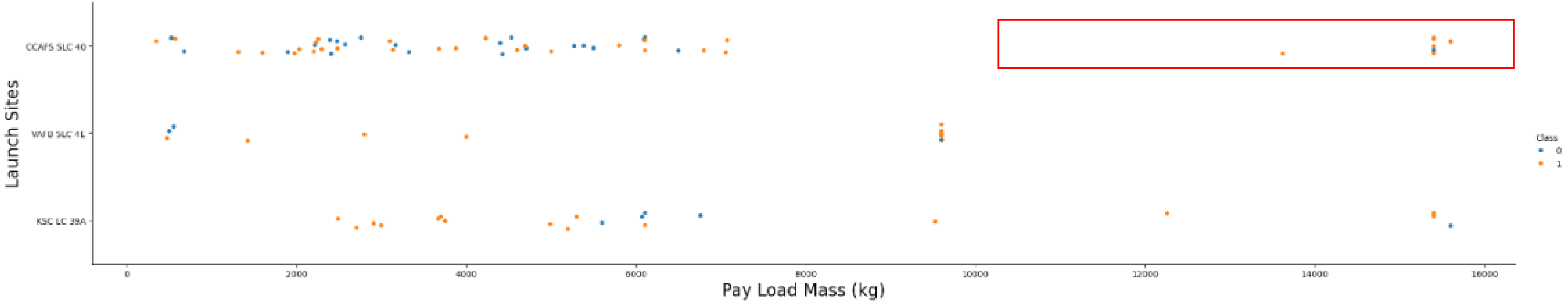
- Successful information of launches IDs from API data (columns rocket, payloads, launchpad, and cores).
- **The larger the amount of flights are, the grater the success rate is, at a determined launch site.**



Results and Discussion

Exploratory Data Analysis (EDA)

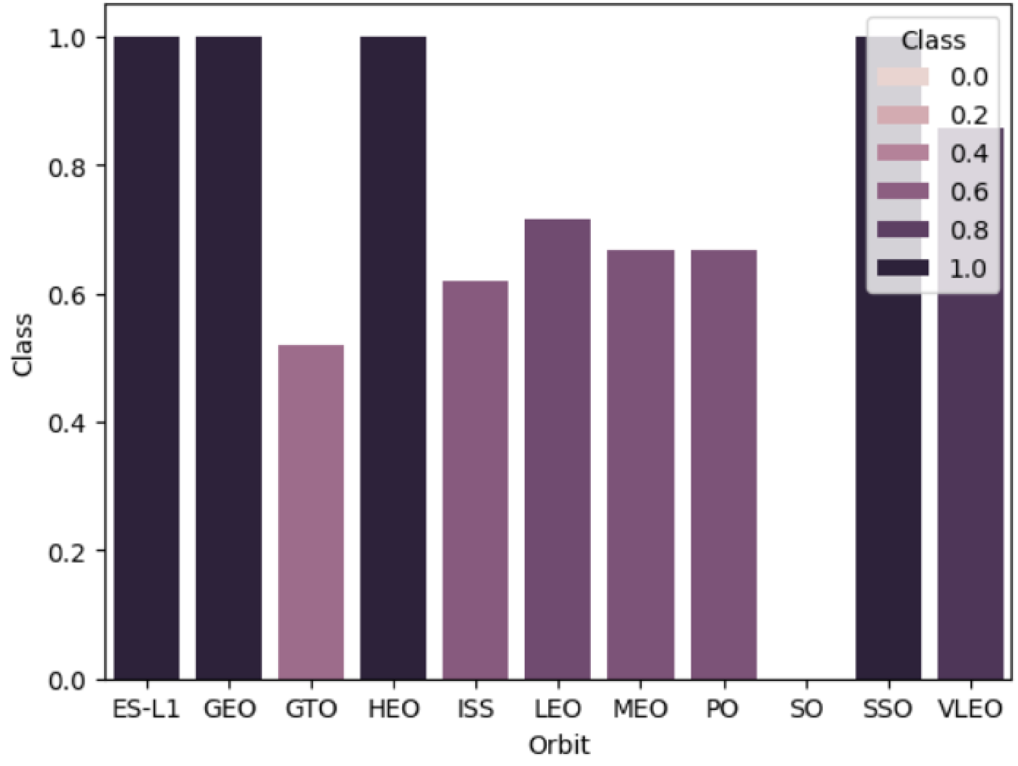
- **The greater the payload mass is for launch site CCAFS SLC 40, the better is the success rate for the rokect.**



Results and Discussion

Exploratory Data Analysis (EDA)

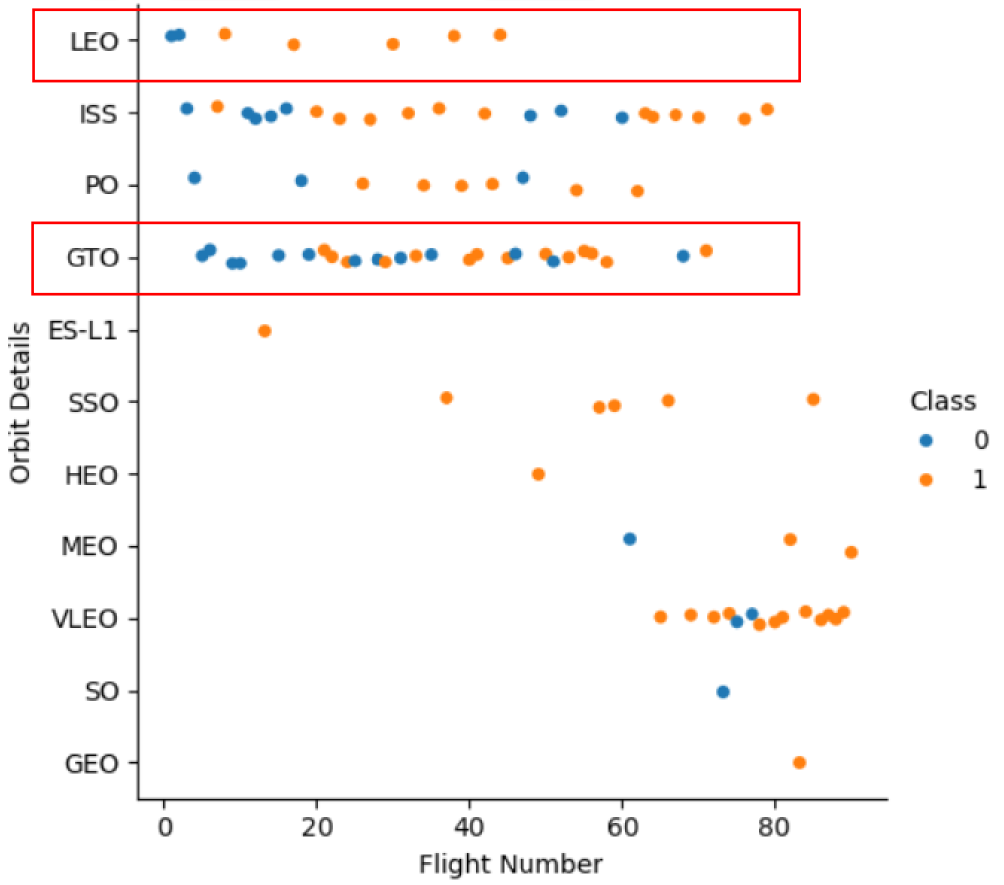
- **Orbits ES-L1, GEO, HEO and SSO have the most successful rates followed by VLEO. The orbit type is a determinate factor for launch success Falcon 9.**



Results and Discussion

Exploratory Data Analysis (EDA)

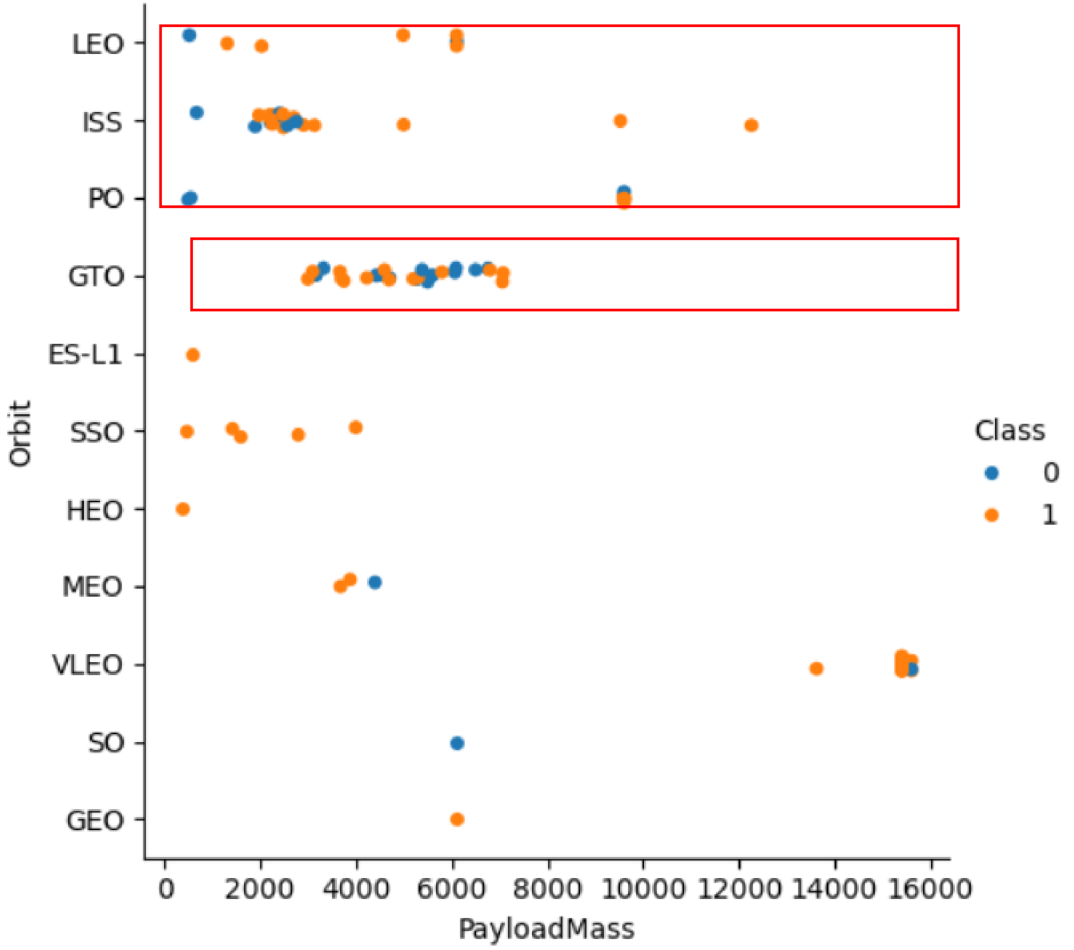
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Results and Discussion

Exploratory Data Analysis (EDA)

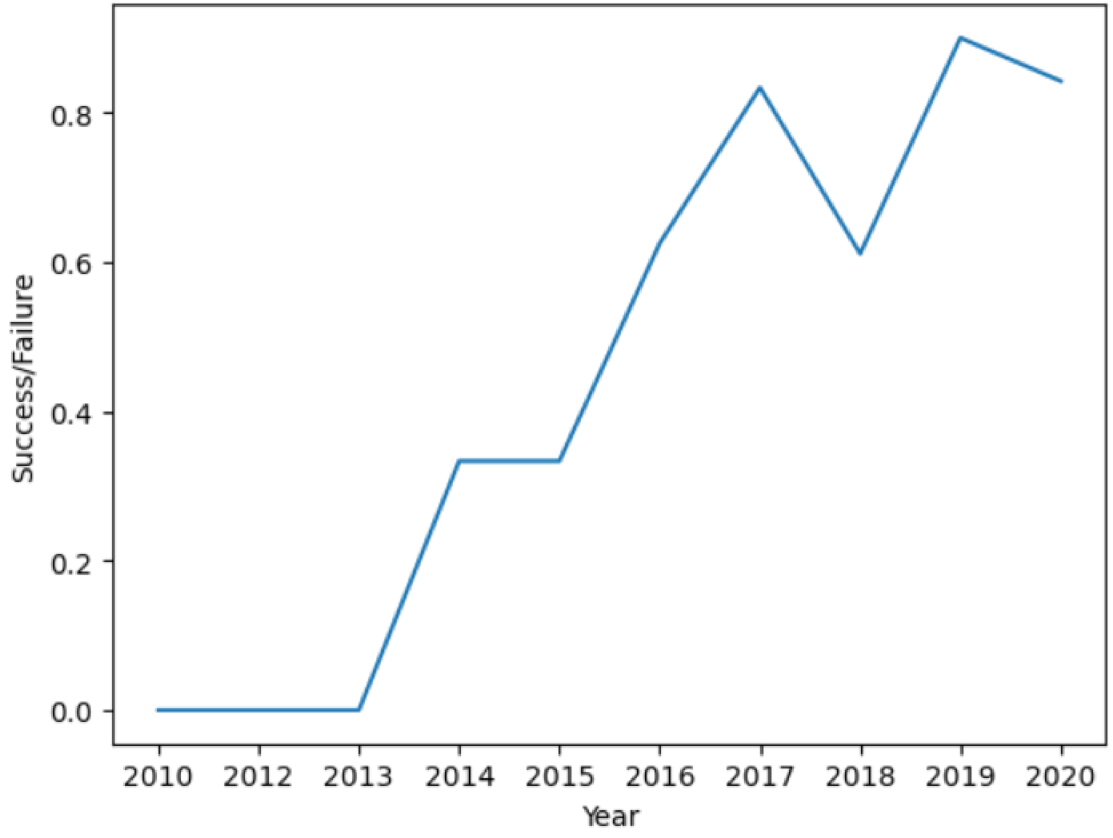
- Positive landing with heavy payloads rate are more for Polar, LEO and ISS.
- GTO unsuccessful mission.



Results and Discussion

Exploratory Data Analysis (EDA)

- Looking up the trend between year and average success rate, we can observe that from 2013 to 2020 success rate keeps increasing. This shows that some of the variables explored are suitable to determine successful launches.



Results and Discussion

Exploratory Data Analysis (EDA) with SQL

1. Unique launch sites in the space mission.

```
In [11]: %sql select "Launch_Site" from SPACEXTBL;

 * sqlite:///my_data1.db
Done.
```

Out[11]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

2. Launch sites begin with the string 'CCA'

```
In [12]: %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;

 * sqlite:///my_data1.db
Done.
```

Out[12]:

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Results and Discussion

Exploratory Data Analysis (EDA) with SQL

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

In [13]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer= 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[13]:

sum(PAYLOAD_MASS__KG_)
45596

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

In [14]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version='F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

avg(PAYLOAD_MASS__KG_)
2928.4

Results and Discussion

Exploratory Data Analysis (EDA) with SQL

5. **First successful landing outcome in ground pad achieved**

In [15]:
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome='Success (ground pad)'
```
```
 * sqlite:///my_data1.db
Done.
```
Out[15]:

min(DATE)
2015-12-22

6. **Boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**

In [17]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 a
```
```
 * sqlite:///my_data1.db
Done.
```
Out[17]:

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

7. **Total number of successful and failure mission outcome**

In [29]:
```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```
```
 * sqlite:///my_data1.db
Done.
```
Out[29]:

missionoutcomes
1
98
1
1

Results and Discussion

Exploratory Data Analysis (EDA) with SQL

8. **Booster versions which have carried the maximum payload mass**

In [32]:

```
%sql select distinct Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL
```

```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

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

9. **Month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.**

In [41]:

```
%sql select substr(DATE,6,2),LANDING_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE from SPACEXTBL where substr(DATE,0,5)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[41]:

substr(DATE,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

Results and Discussion

Exploratory Data Analysis (EDA) with SQL

10. Landing outcomes (Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

```
In [42]: %sql select LANDING_OUTCOME, count(*) from SPACEXTBL where DATE BETWEEN '2010-06-04' and '2017-03-20' group by LANDING_OUTC
```

```
 * sqlite:///my_data1.db
Done.
```

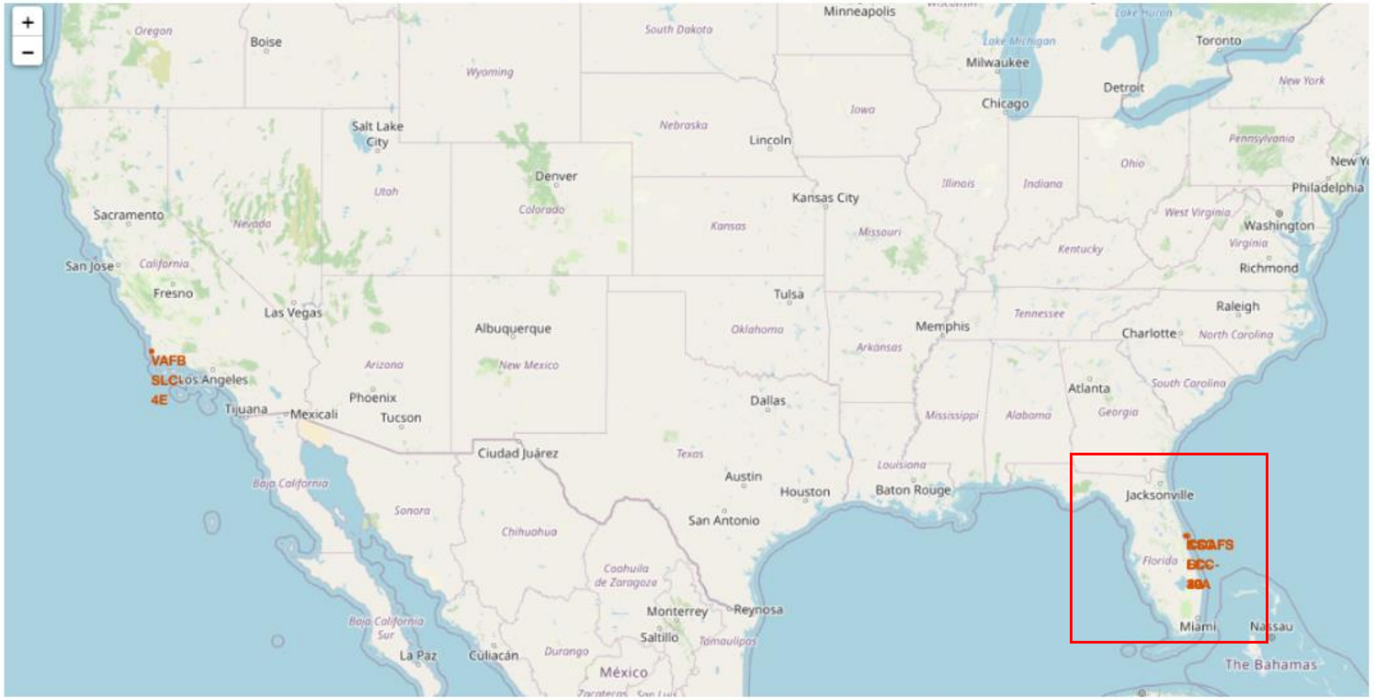
Out[42]:

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

Results and Discussion

Launch locations and proximities

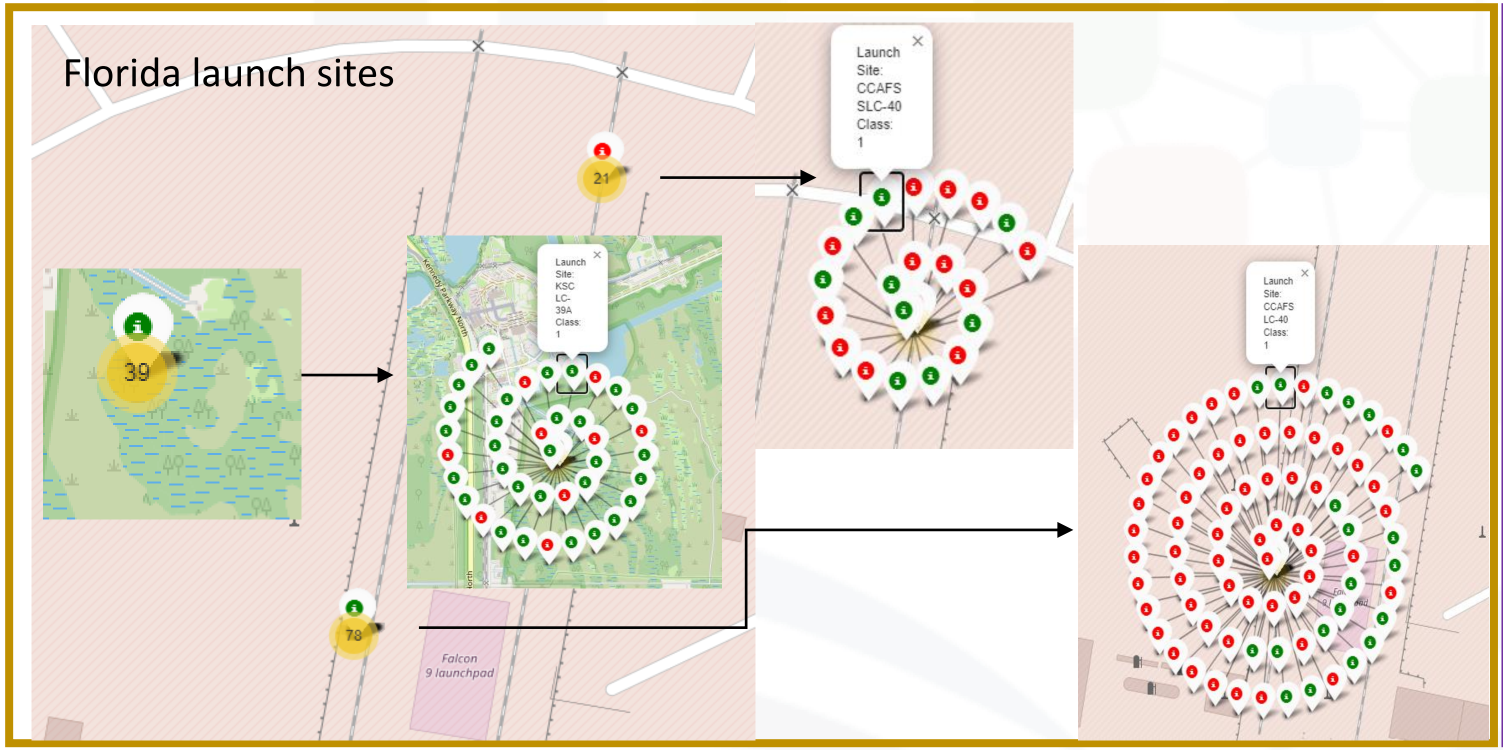
Marked launch sites are in Unites States of America, all of them closed to the coast but not closed to the equator witch means the rockets won’t have an additional extra velocity boost in general but maybe a little bit more the ones launched from Florida.

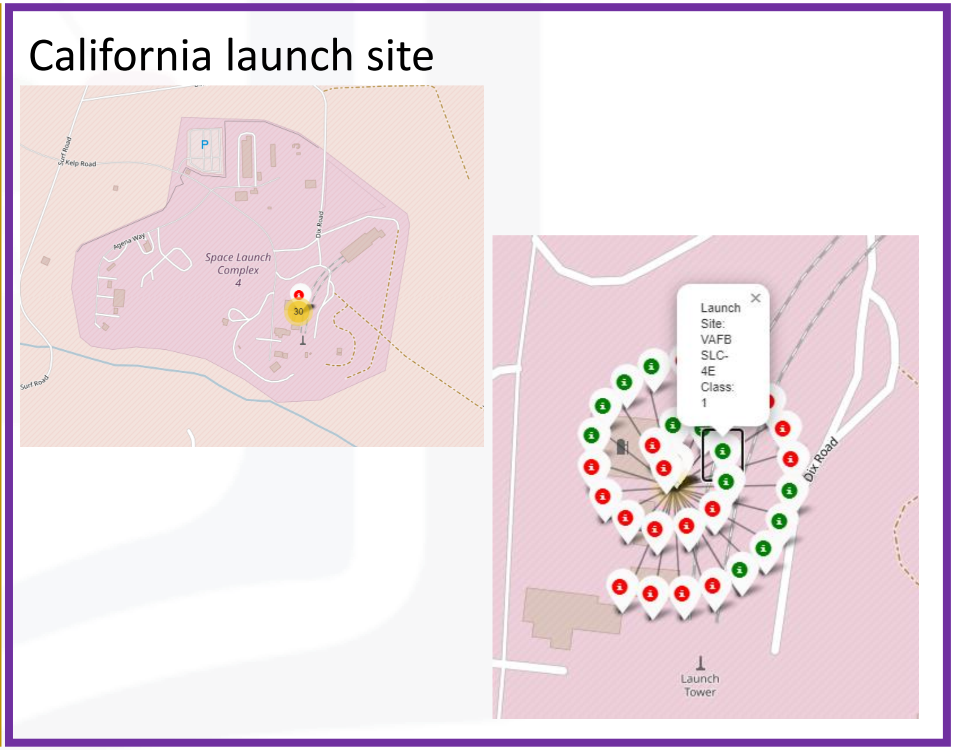


Results and Discussion

Launch locations marked

Green= success Red= Fail

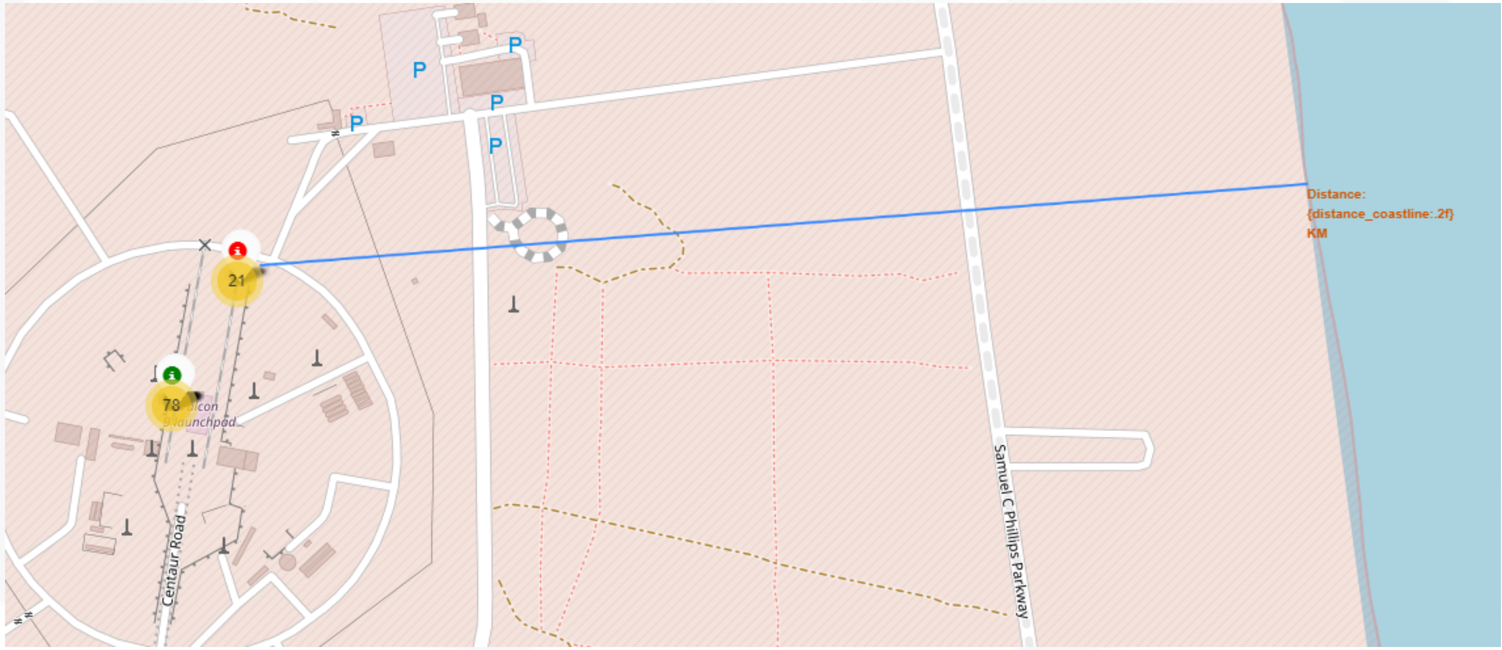




Results and Discussion

Launch locations distances to landmarks

All the launch locations are not in close proximity of railways, highways nor cities because of safety reasons. In the other hand, launch locations are closed to the costal area in case there is an accident, the rocket can crash towards the sea.



Results and Discussion

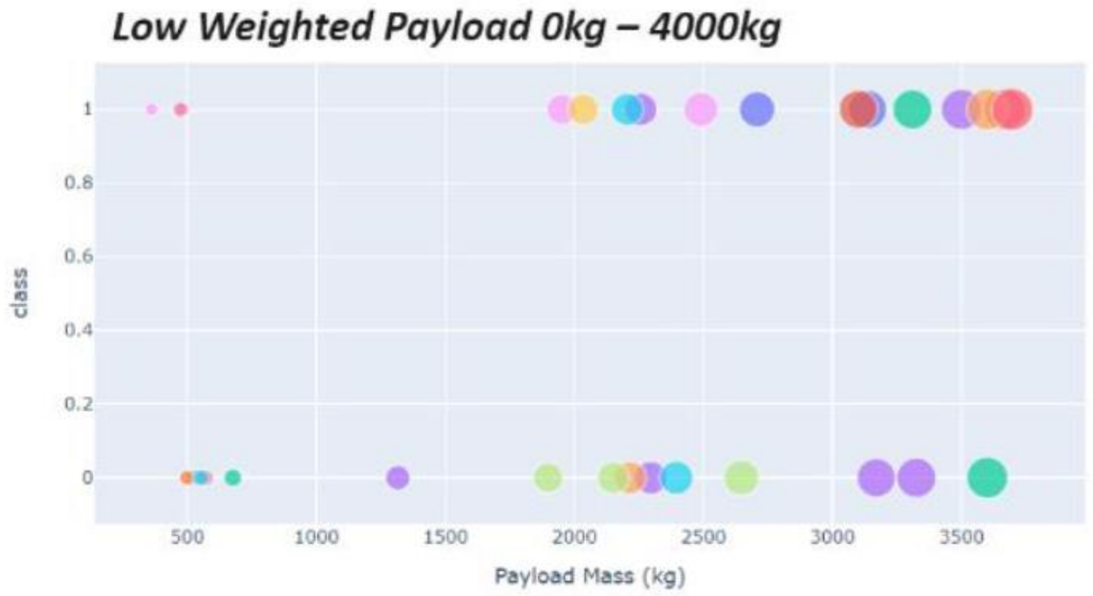
Launch sites success percentage

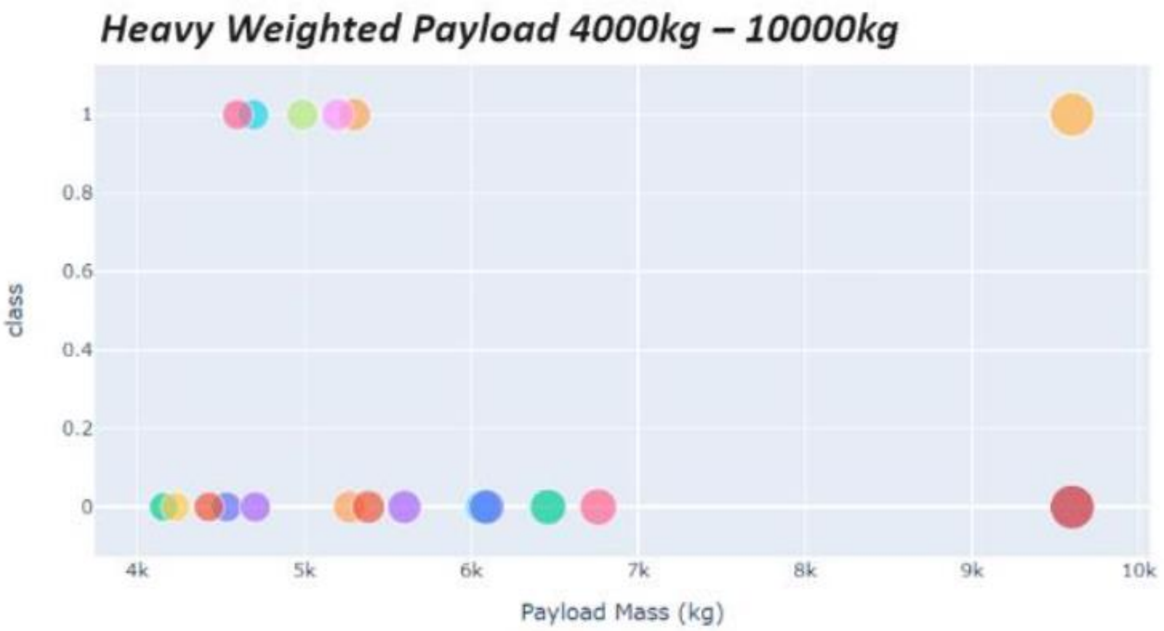
KSC LC-39 A is the site with most successful launches

https://github.com/veronicaestefania/The-Certificate/blob/main/Build%20a%20Dashboard%20Application%20with%20Plotly%20Dash

Results and Discussion

Payload vs. Launch Outcome for all sites





The success rates for low weighed payloads are higher than the heavy ones.

Results and Discussion

Classification and Predictive Analysis

- **Decision tree** performs well on training data with an accuracy of 86.25%
- Decision tree **confusion matrix** shows that the classifier can classify between different classes. However false positives are not taken in consideration, some unsuccessful landings are being counted as successful ones.

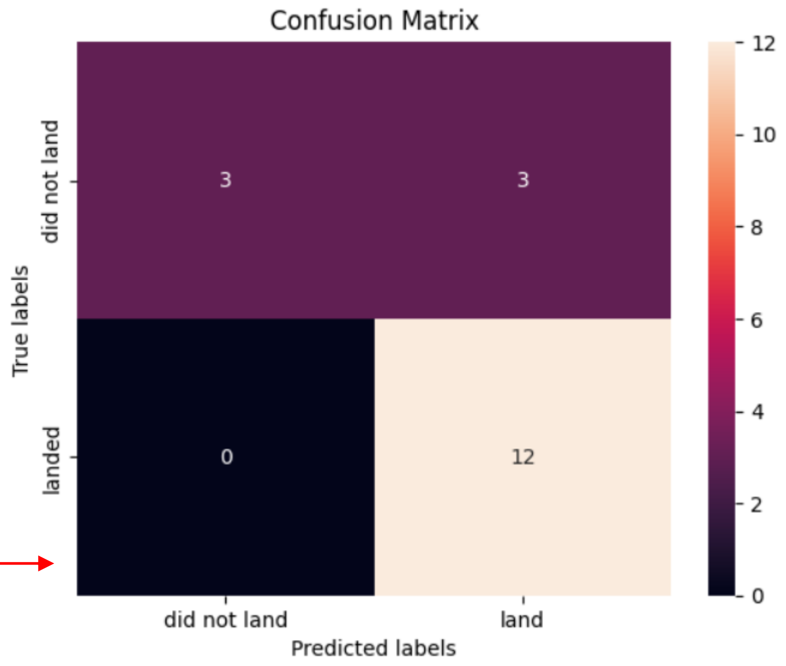
In [43]:
```
accuracy = tree_cv.score(X_test, Y_test)
print("Accuracy on Test Data:", accuracy)
```

```
Accuracy on Test Data: 0.8333333333333334
```

We can plot the confusion matrix

In [44]:
```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

All successful landings???




LEARNINGS

Conclusions

To predict the landing in the first stage of a Falcon 9 launch:

- ✓ Payload and Orbit type are important variables to consider that may affect the mission outcome.
  - ✓ Orbits ES-L1, GEO, HEO, SSO, VLEO had the most successful rate.
  - ✓ KSC LC-39A had the most successful launches of all the sites.
- ✓ Over different predictive models, decision tree algorithm performed the best for this project.
- ✓ The more a falcon9 type rocket flights in a certain launch site, the greater the success rate at that site is.


THANK YOU!!