IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Xavier Bravo
14/01/2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
    - Data collection from Api
    - Data collection with Web Scraping
    - Data wrangling
    - EDA with SQL
    - EDA & Data Visualization with Python
    - Machine Learning
- Summary of all results
    - EDA Result
    - Interactive analytics with screenshot
    - Machine Learning results

Introduction

- **Project background and context**

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. This goal of the project is to create a machine learning pipeline to predict if the first stage will land succesfully.

- **Problems you want to find answers**
- What factors determine if the rocket will land succesfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

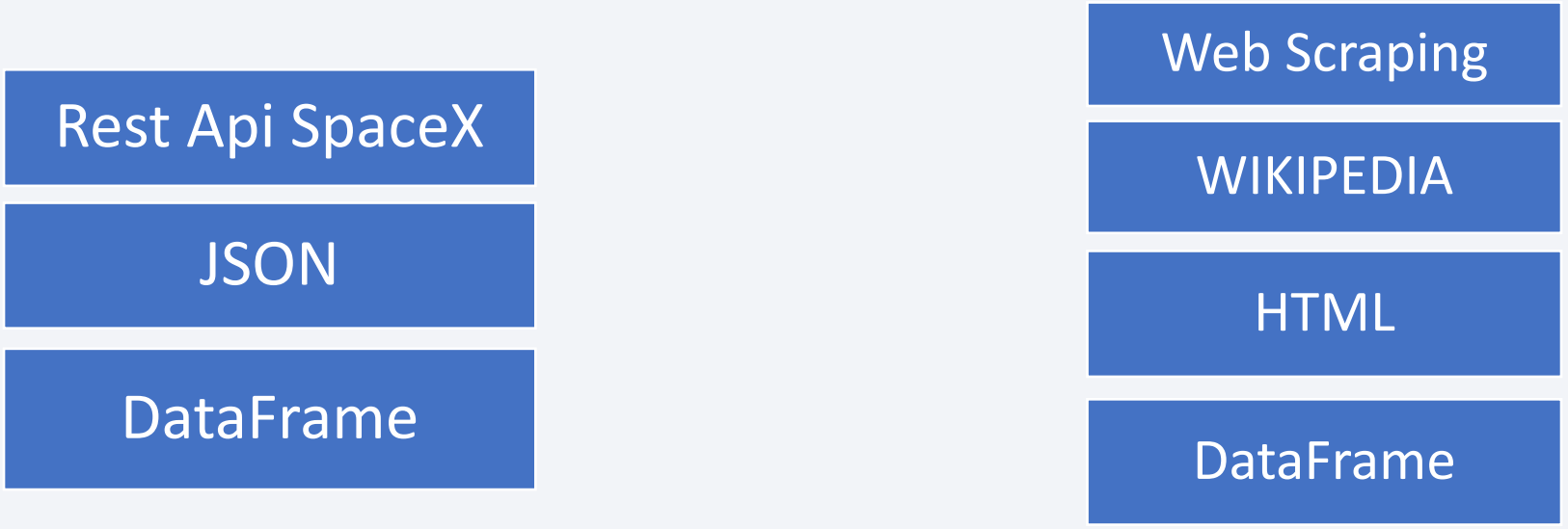
Executive Summary

- Data collection methodology:
    - DatafromSpaceXwasobtainedfrom2sources:
    - SpaceXAPI(https://api.spacexdata.com/v4/rockets/) • WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_l aunches)
    - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_l aunches)
- Perform data wrangling
    - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features

Data Collection

- Describe how data sets were collected.

- Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.



Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and then used;
- This API was used according to the flowchart beside and then data is persisted.
- https://github.com/Xavierongo/ibm_data_science_capstone

```
Import requests
Import pandas as pd
url = https://api.spacexdata.com/v4/launches/past
response = requests.get(url)
data = pd.json_normalize(response.json())
```

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.

Getting response from API

↓

Converting response to a json file

↓

Apply clean data functions

↓

Assign list to dict and then créate df

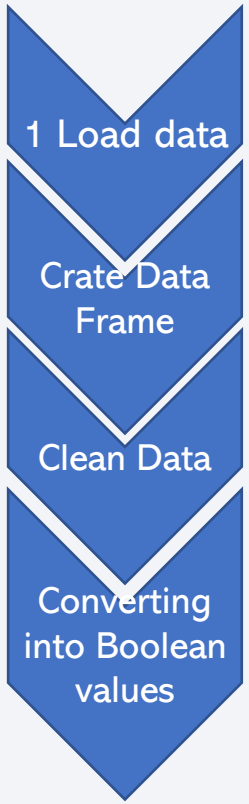
↓

Filter df and then export it

```
Import requests
From bs4 import BeautifulSoup
static_url =
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_
Falcon_Heavy_launches&oldid=1027686922
response = requests.get(static_url)
html_content  = response.content
soup = BeautifulSoup(html_content, 'html.parser')
```

Data Wrangling

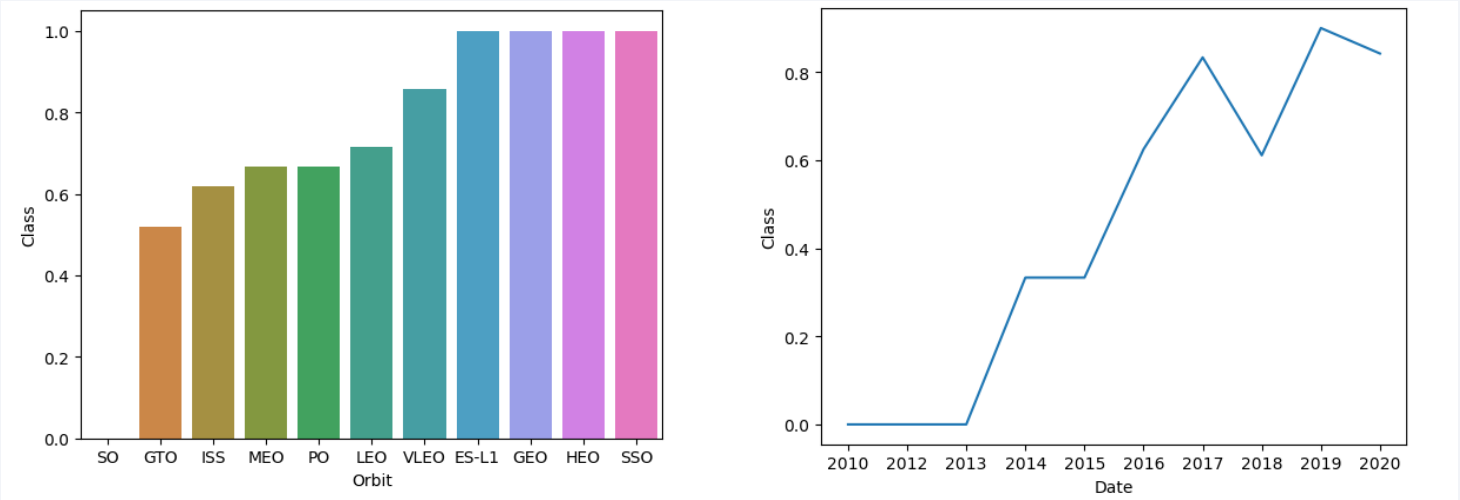
Once the data has been collected, we examine it for missing values and verify the data types. Next, we take steps to clean it, such as filling in missing values using averages or other techniques, adjusting data types as required, and transforming categorical variables into numerical formats like integers, floats, or through one-hot encoding.



EDA with Data Visualization

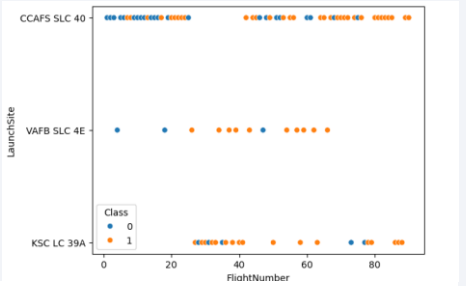
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.

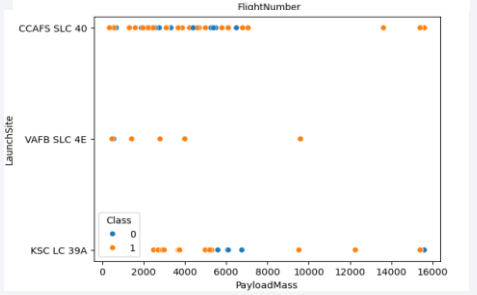


Bar plot: Succes rate of each orbit

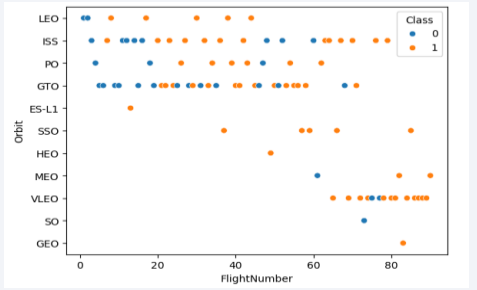
Line plot: To get yearly average launch succes trend



Scatter plot: flightnumber vs Launchsite



Scatter plot: Payloadmass vs Orbit



Scatter plot: flightnumber vs Orbit

EDA with SQL

The following SQL queries were performed:

- Names of the unique launch sites in the space mission.
- Top 5 launch sites whose name begins with the string 'CCA’.
- Total pay load mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1.
- Date when the first successful landing outcome in ground pad was achieved.
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg.
- Total number of successful and failure mission outcomes.
- Names of the booster versions which have carried the maximum payload mass.
- Failed landing out comes in droneship, their booster versions, and launch site names for in year 2015 .
- Rank of the count of landing outcomes (such as Failure (droneship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Explain why you added those objects
- https://github.com/Xavierongo/ibm_data_science_capstone

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- https://github.com/Xavierongo/ibm_data_science_capstone

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.

- We built different machine learning models and tune different hyperparameters using GridSearchCV.

- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.

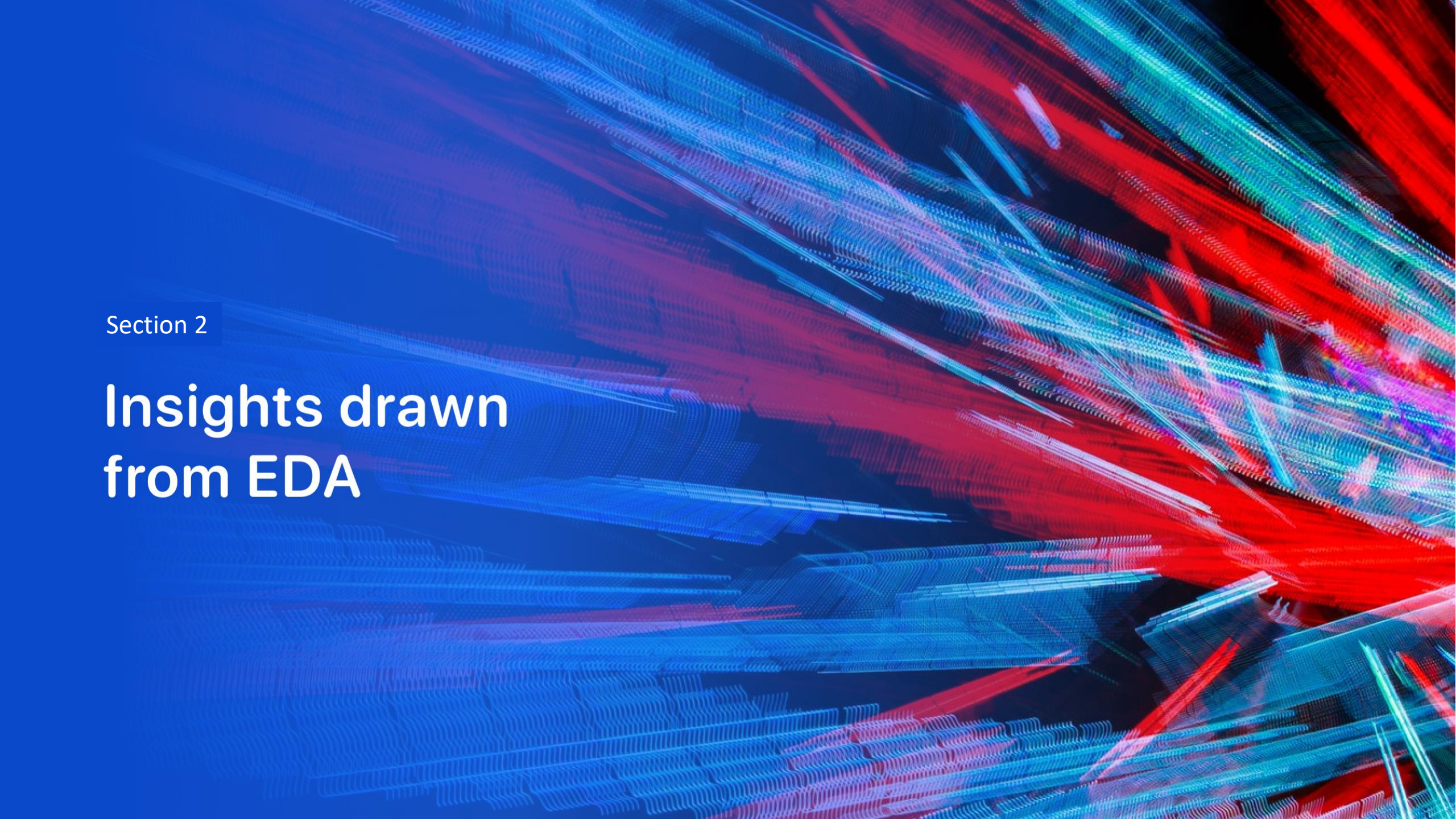
- We found the best performing classification model

Results

EDA Results

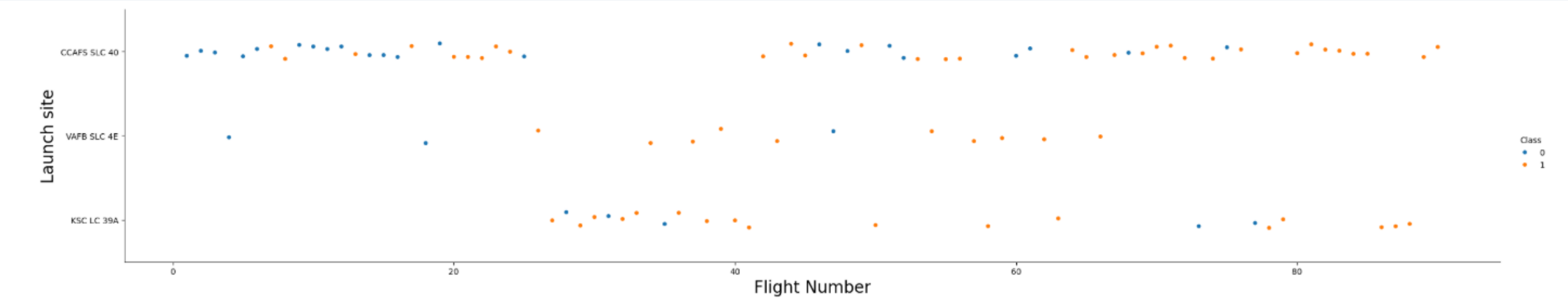
- Exploratory data analysis results.
- Space X uses 4 different launch sites.
- The first launches were done to Space X itself and NASA.
- The average payload of F9 v1.1 booster is 2,928 kg.
- The first success landing outcome happened in 2015 fiver year after the first launch.
- Many Falcon 9 booster versions were successful at landing in drone ships having.

payload above the average.


Section 2
Insights drawn from EDA

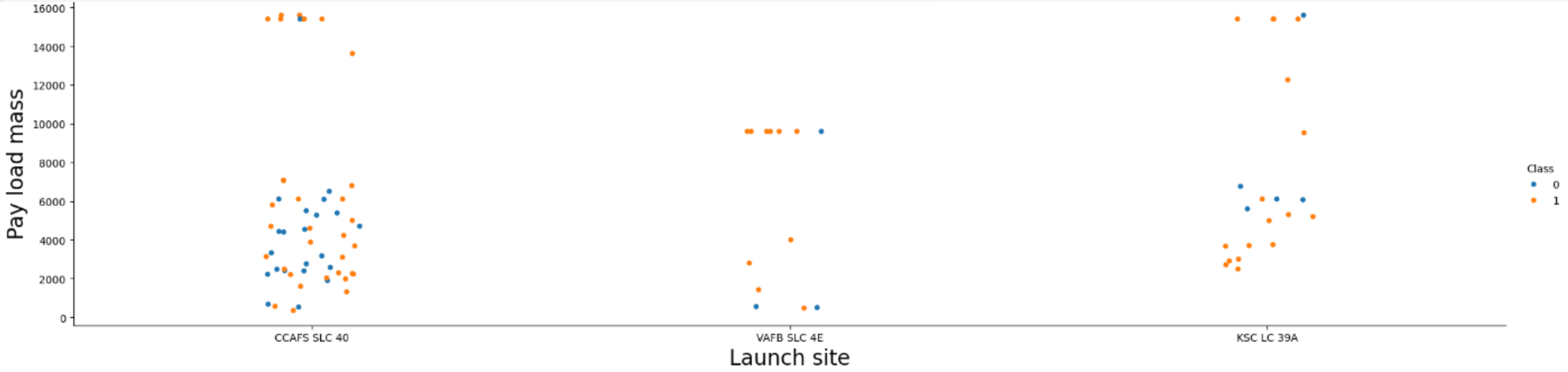
Flight Number vs. Launch Site

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch site",fontsize=20)
plt.show();
```

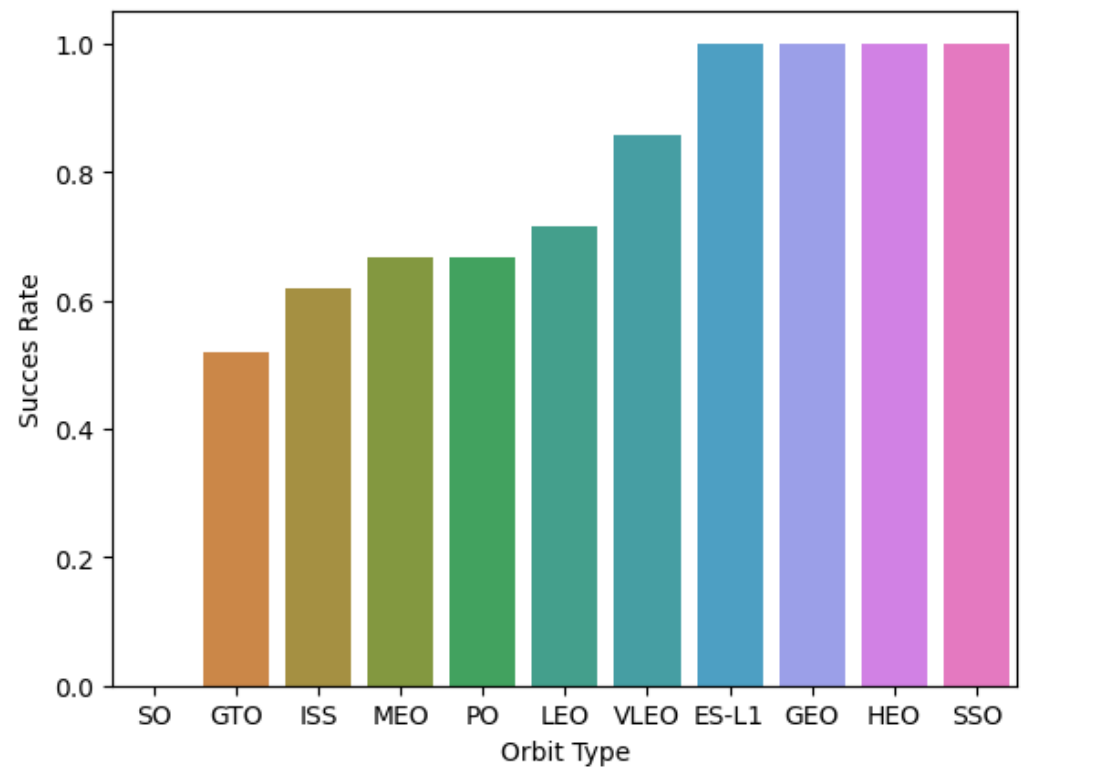


Payload vs. Launch Site

```
sns.catplot(y="PayloadMass", x="LaunchSite", hue="Class", data=df, aspect = 4)
plt.xlabel("Launch site",fontsize=20)
plt.ylabel("Pay load mass",fontsize=20)
plt.show();
```



Success Rate vs. Orbit Type

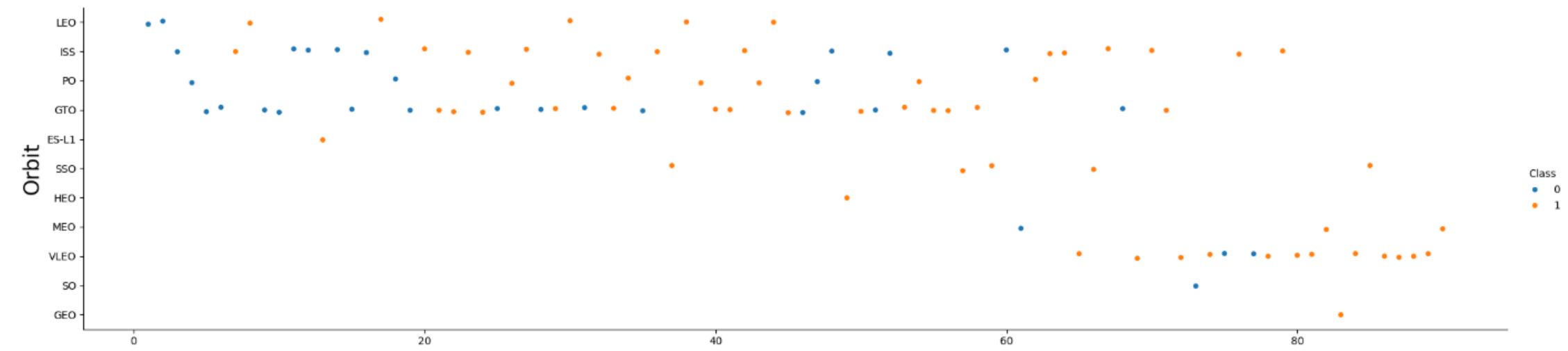


From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

```
# HINT use groupby method on Orbit column and get the mean of Class column
df_gouped = df.groupby("Orbit")["Class"].mean().sort_values().reset_index()
sns.barplot(data = df_gouped, x = "Orbit", y = "Class", hue = "Orbit")
plt.xlabel("Orbit Type")
plt.ylabel("Succes Rate")
plt.show();
```

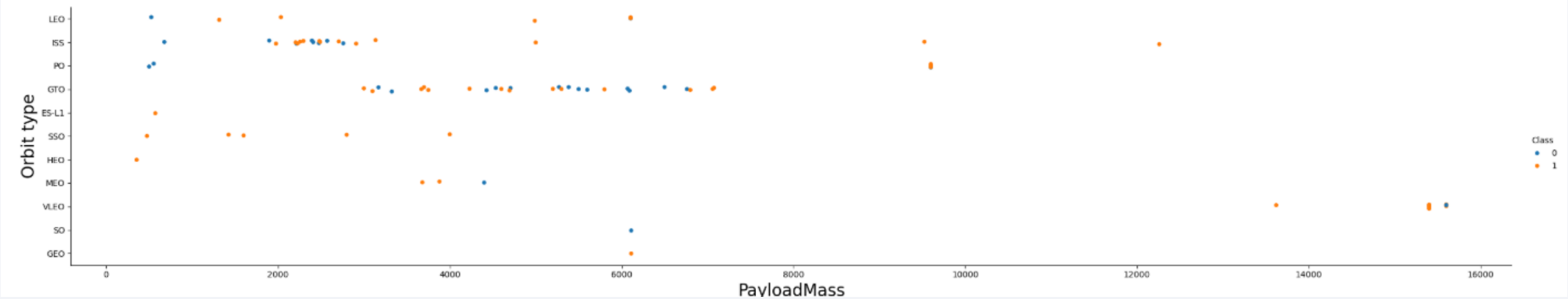
Flight Number vs. Orbit Type

The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



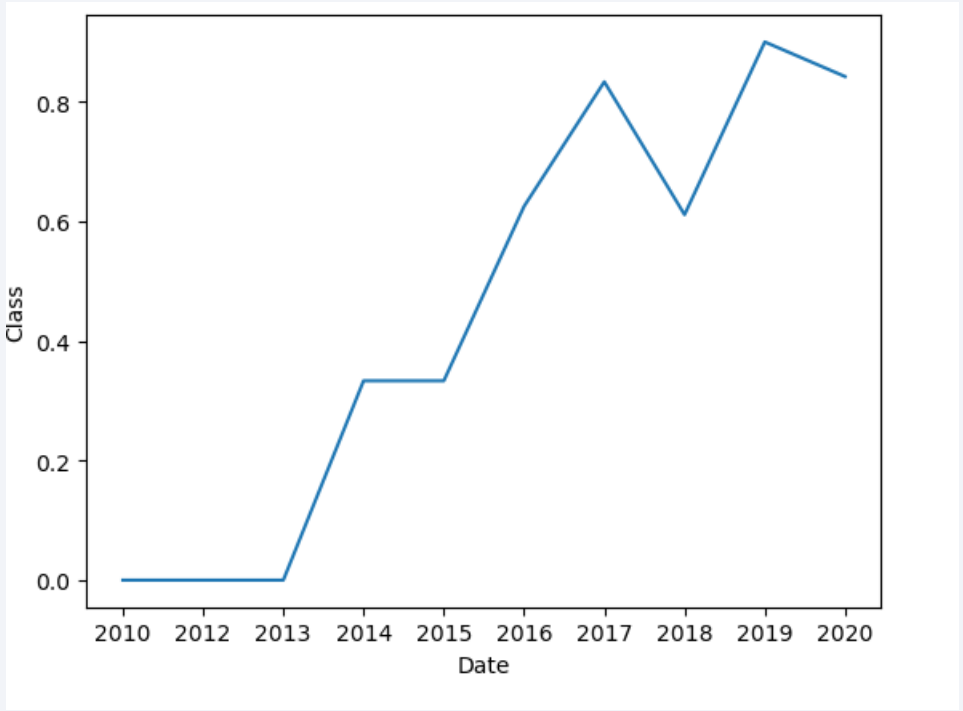
Payload vs. Orbit Type

We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



Launch Success Yearly Trend

From the plot, we can observe that success rate since 2013 kept on increasing till 2020.




```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
df_ag_line = df.groupby("Date")["Class"].mean().reset_index()
sns.lineplot(data = df_ag_line, x = "Date", y= "Class")
```

All Launch Site Names

We used the key word DISTINCT to show only unique launch sites from the SpaceX data.

```
%sql  SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

We used the query above to display 5 records where launch sites begin with `CCA`

```
%sql SELECT Launch_site FROM SPACEXTBL WHERE  Launch_site  LIKE 'CCA%' LIMIT 5;
```

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

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

Total Payload Mass

We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass, Customer FROM SPACEXTBL WHERE Customer = 'NASA (CRS)';
```

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

total_payload_mass	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_carrier FROM SPACEXTBL WHERE Booster_Version  = 'F9 v1.1'
```

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

avg_carrier
2928.4

First Successful Ground Landing Date

We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
%sql SELECT Date, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)';
```

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

Date	Landing_Outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcome

```
%sql SELECT  Mission_Outcome, COUNT(*) AS n_missions_succes FROM SPACEXTBL WHERE Mission_Outcome LIKE 'S%' ;
```

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

Mission_Outcome	n_missions_succes
Success	100

```
%sql SELECT  Mission_Outcome, COUNT(*) AS n_missions_succes FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Fai%' ;
```

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

Mission_Outcome	n_missions_succes
Failure (in flight)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

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

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

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT strftime('%m', Date) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL WHERE Landing_Outcome LIKE '%drone ship%' AND substr(Date, 0, 5) = '2015' AND substr(Date, 6, 2) BETWEEN '01' AND '12';
```

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

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
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40

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

- 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

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

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

Landing_Outcome	Outcome_Count
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

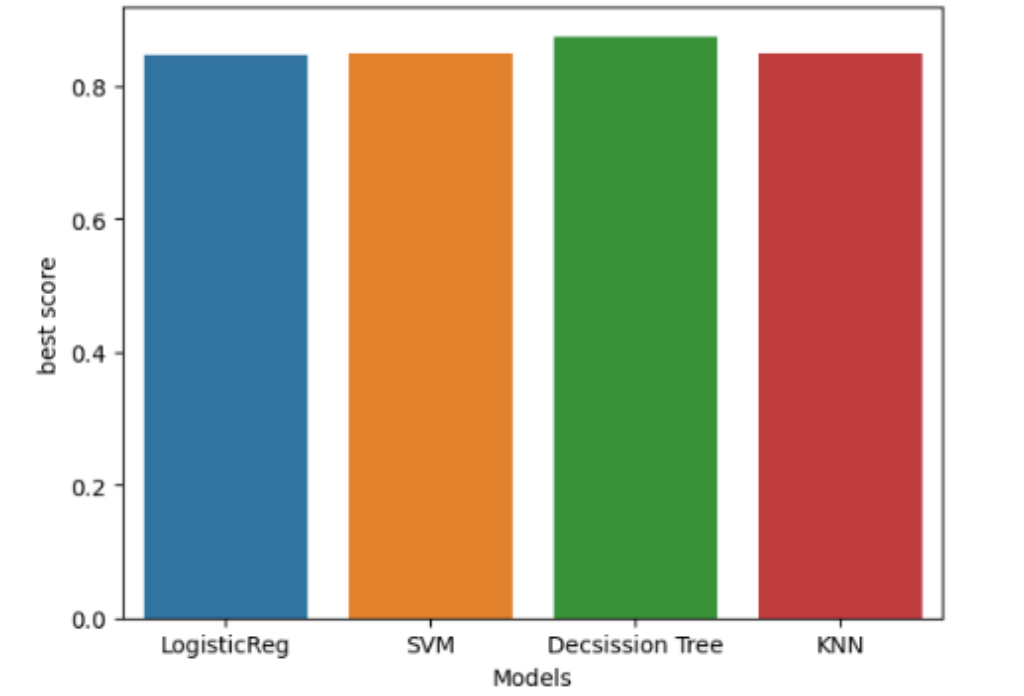
<Folium Map Screenshot 1>

There are 2 zones of launches, California and Florida.

Section 5

Predictive Analysis (Classification)

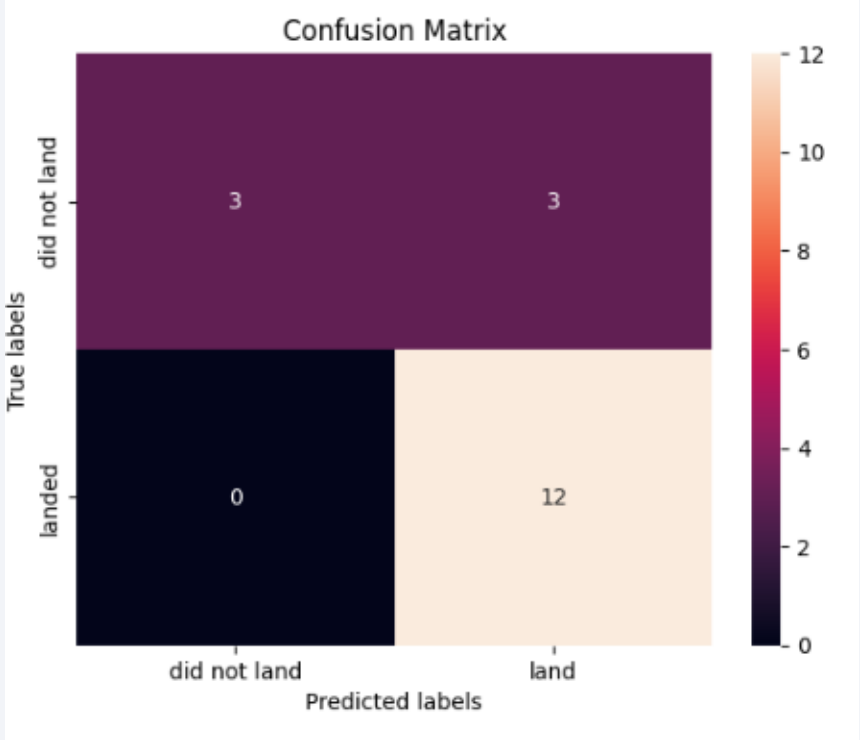
Classification Accuracy



The decision tree classifier is the model with the highest classification accuracy

Confusion Matrix

The decision tree classifier's confusion matrix demonstrates its ability to differentiate between various classes. However, the main issue lies in the false positives, where the classifier incorrectly identifies unsuccessful landings as successful ones.



Conclusions

- **Best Machine Learning Algorithm**: The Decision Tree classifier outperformed other models for this dataset.

- **Top Launch Site**: KSC LC-39A recorded the highest number of successful launches, with higher flight counts correlating to greater success rates.

- **Successful Orbits**: Orbits such as ES-L1, GEO, HEO, SSO, and VLEO showed the highest success rates.

- **Improved Success Over Time**: Launch success rates have steadily increased, particularly between 2013 and 2020.

Thank you!