# DATA SCIENCE IN SPACE

PRESENTED BY: MOHAMAD JAMIL JAMMOUL

APRIL 30TH 2023

# OUTLINE

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

# 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

- The context and background of the project involve Space X advertising Falcon 9 rocket launches on its website for $62 million, which is much cheaper than other providers who charge upwards of $165 million per launch. This cost difference is due to Space X's ability to reuse the first stage of the rocket. To compete with Space X, an alternate company needs to know if the first stage will land successfully to determine the cost of a launch. Therefore, the objective of this project is to develop a machine learning pipeline that can predict whether the first stage will land successfully.
- The project aims to answer the following questions: What are the factors that determine the success rate of a successful landing? How do various features interact to determine the success rate of a successful landing? What operating conditions need to be in place to ensure a successful landing program?

# METHODOLOGY

- The following executive summary outlines the methodology and process used to collect, analyze and visualize data related to SpaceX rocket launches.
- Data was collected through a combination of SpaceX's API and web scraping from Wikipedia. The data was then wrangled, and categorical features were encoded using one-hot encoding. Exploratory data analysis (EDA) was conducted using SQL and visualization techniques. Interactive visual analytics was performed using Folium and Plotly Dash.
- Finally, predictive analysis was performed using classification models, which were built, tuned, and evaluated to ensure their accuracy.

# DATA COLLECTION

- To obtain the data required for analysis, several methods were used. Firstly, data collection was achieved by sending a get request to the SpaceX API. The response content was then decoded using the .json() function, and converted into a pandas dataframe using .json_normalize(). The data was then cleaned, missing values were identified and filled where necessary.
- In addition to using the SpaceX API, web scraping was also employed to gather data from Wikipedia. Specifically, BeautifulSoup was used to extract launch records from the Falcon 9 table on Wikipedia. The objective was to convert the extracted HTML table into a pandas dataframe for future analysis.

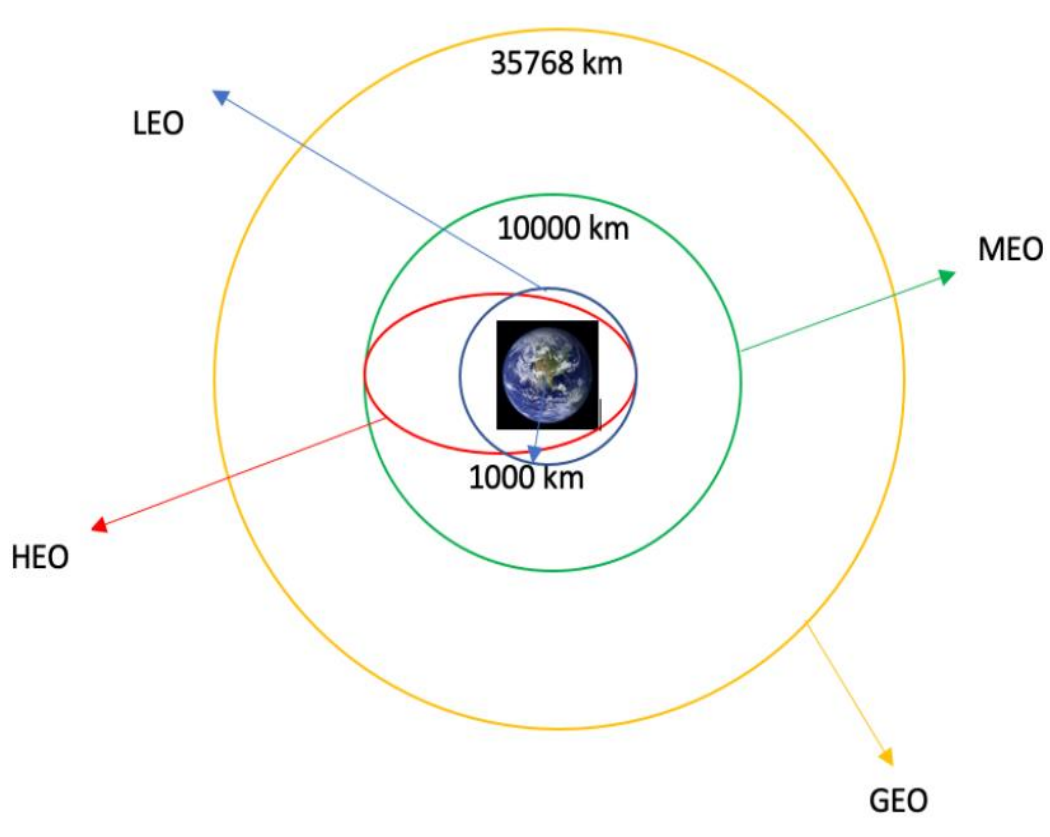
# DATA COLLECTION THROUGH AN API

- Data collection from the SpaceX API was initiated by sending a get request. The data obtained was then subjected to cleaning and basic wrangling and formatting. A link to the notebook containing these processes can be found at https://github.com/moejamul/IBMDSCapstone/blob/main/Data%20Collection%20API.ipynb.

# DATA COLLECTION THROUGH WEB SCRAPPING

- Web scraping was utilized to extract Falcon 9 launch records from a webpage, using BeautifulSoup. The extracted data was then parsed and converted into a pandas dataframe. More information on this process can be found in the following notebook:
- https://github.com/moejamul/IBMDSCapstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb.

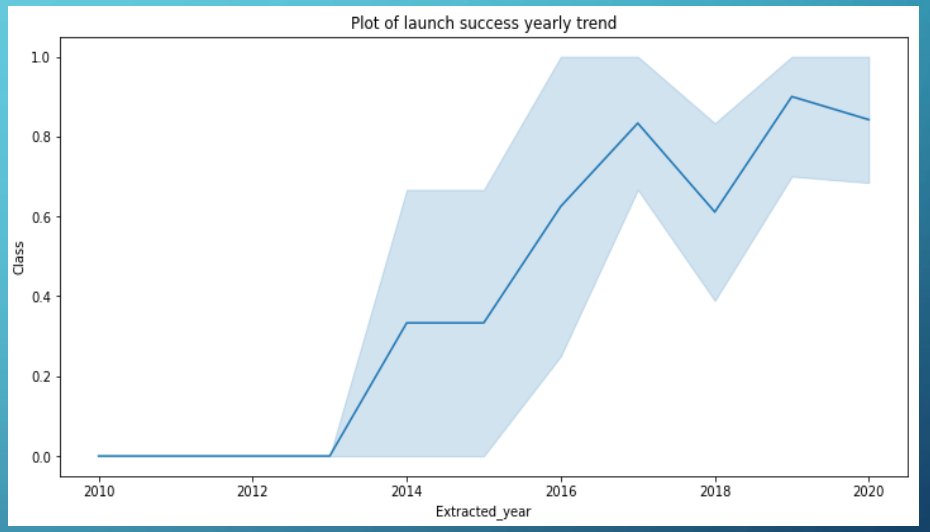
# DATA WRANGLING




- We performed exploratory data analysis and determined the training labels.
- We calculated the number of launches at each site, and the number and occurrence of each orbits
- We created landing outcome label from outcome column and exported the results to csv.
- The link to the notebook is.
- https://github.com/moejamul/IBMDSCapstone/blob/main/Data%20Wrangling.ipynb

# DATA VISUALIZATION

- We conducted exploratory data analysis by creating visualizations to examine the relationships between different variables. Specifically, we visualized the relationship between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, and the launch success trend over the years.




The link to the notebook is https://github.com/moejamul/IBMDSCapstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

# SQL

- The SpaceX dataset was loaded into a IBM db2 SQL database within the Jupyter notebook environment. We conducted exploratory data analysis using SQL to extract insights from the data.
- Queries were written to extract information such as the names of unique launch sites, the total payload mass carried by NASA-launched boosters, the average payload mass carried by booster version F9 v1.1, the total number of successful and failed mission outcomes, and the failed landing outcomes in drone ships, along with their booster versions and launch site names.
- For more information on this process, please visit: https://github.com/moejamul/IBMDSCapstone/blob/main/SQL%20JUPYTER.ipynb

# MAP WITH FOLIUM

- We utilized Folium to create a map and marked all launch sites on it. To distinguish between successful and failed launches at each site, we added markers, circles, and lines to represent the success or failure of launches. Launch outcomes were assigned to class 0 and 1, with 0 representing failure and 1 representing success. Using color-labeled marker clusters, we identified launch sites with relatively high success rates.
- We also calculated the distances between launch sites and their proximities to answer questions such as whether launch sites are located near railways, highways, and coastlines, and whether they maintain a certain distance from cities.

# CLASSIFICATION

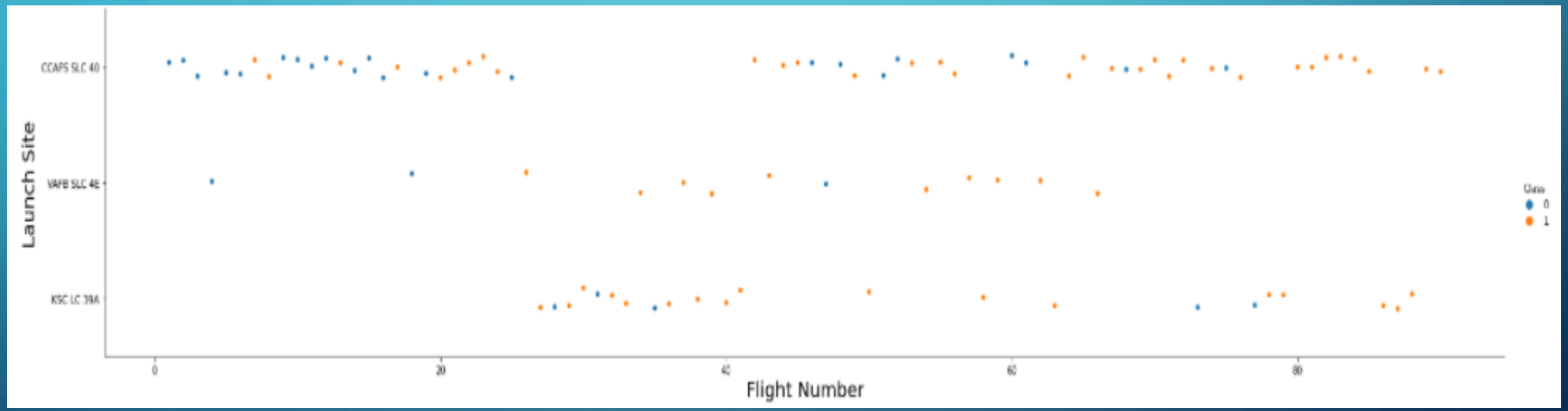
- We loaded the data with the help of numpy and pandas, performed data transformation, and then split it into training and testing sets. After that, we built multiple machine learning models and fine-tuned various hyperparameters by employing GridSearchCV.
- We used accuracy as the metric for evaluating our models and improved their performance by using feature engineering and algorithm tuning. Eventually, we identified the best-performing classification model.
- You can access the notebook at https://github.com/moejamul/IBMDSCapstone/blob/main/Machine%20Learning%20Prediction.ipynb.

# RESULTS

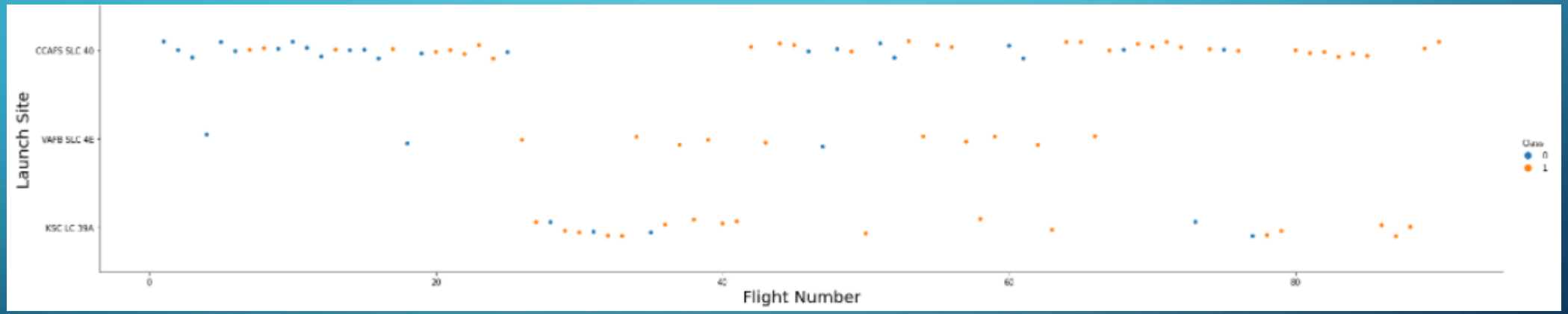
- EDA results
- Screenshots
- Predictive analysis results

# FLIGHT NUMBER AND LAUNCH SITE

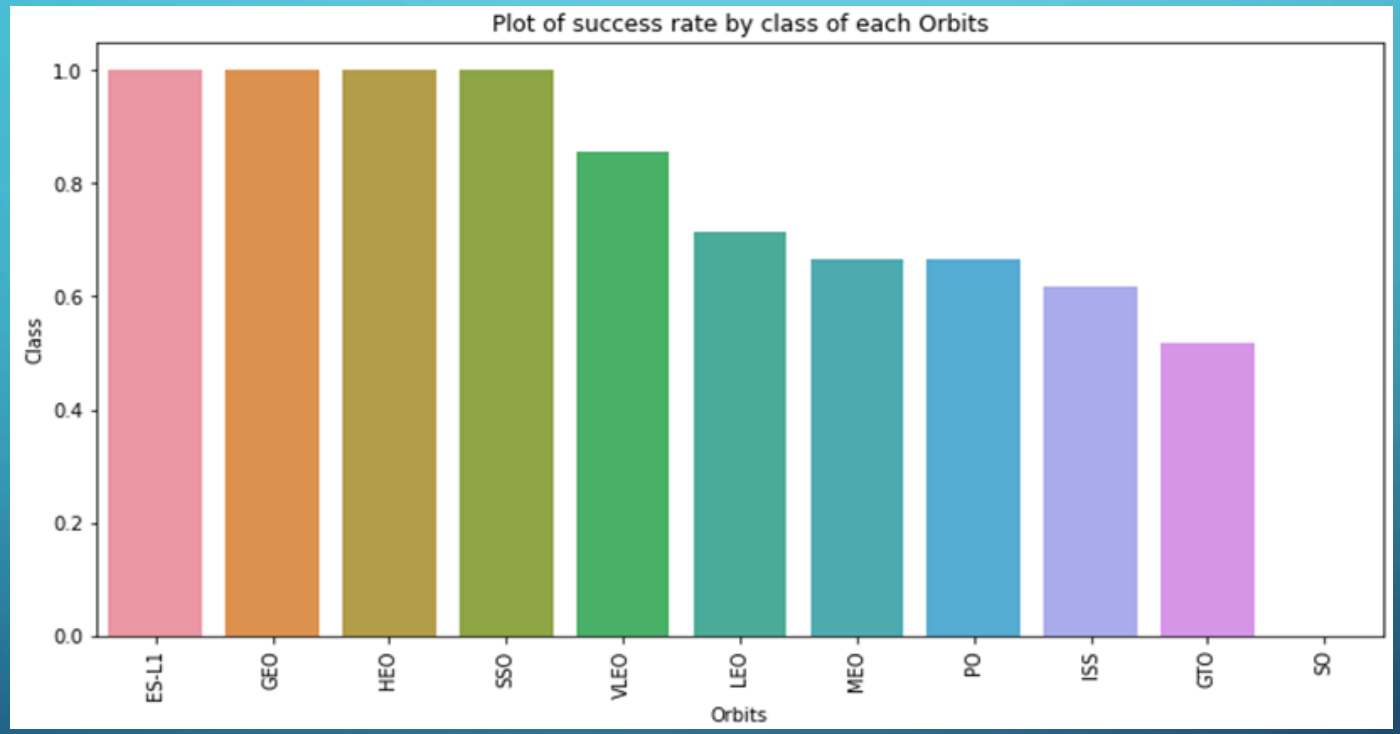
- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



# FLIGHT NUMBER AND LAUNCH SITE

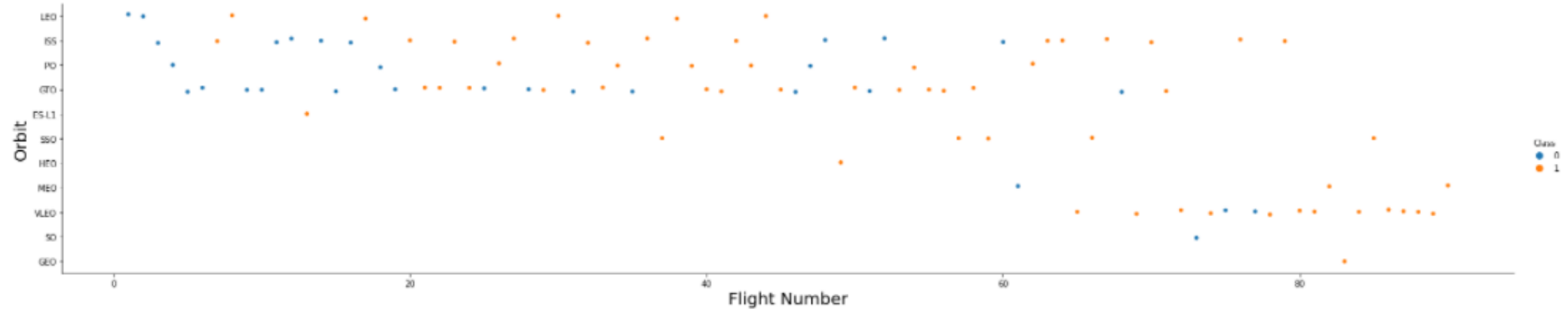


# SUCCESS RATE AND ORBIT TYPE



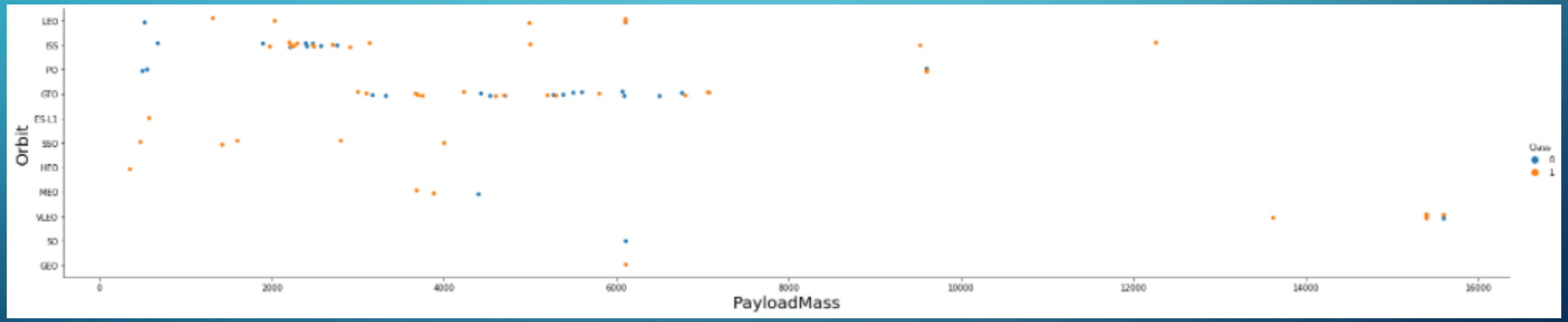
# FLIGHT NUMBER VS. ORBIT TYPE

The following plot displays the relationship between Flight Number and Orbit type. It can be observed that for the LEO orbit, there is a correlation between the number of flights and the success rate, whereas for the GTO orbit, there is no correlation between flight number and the orbit's success rate.

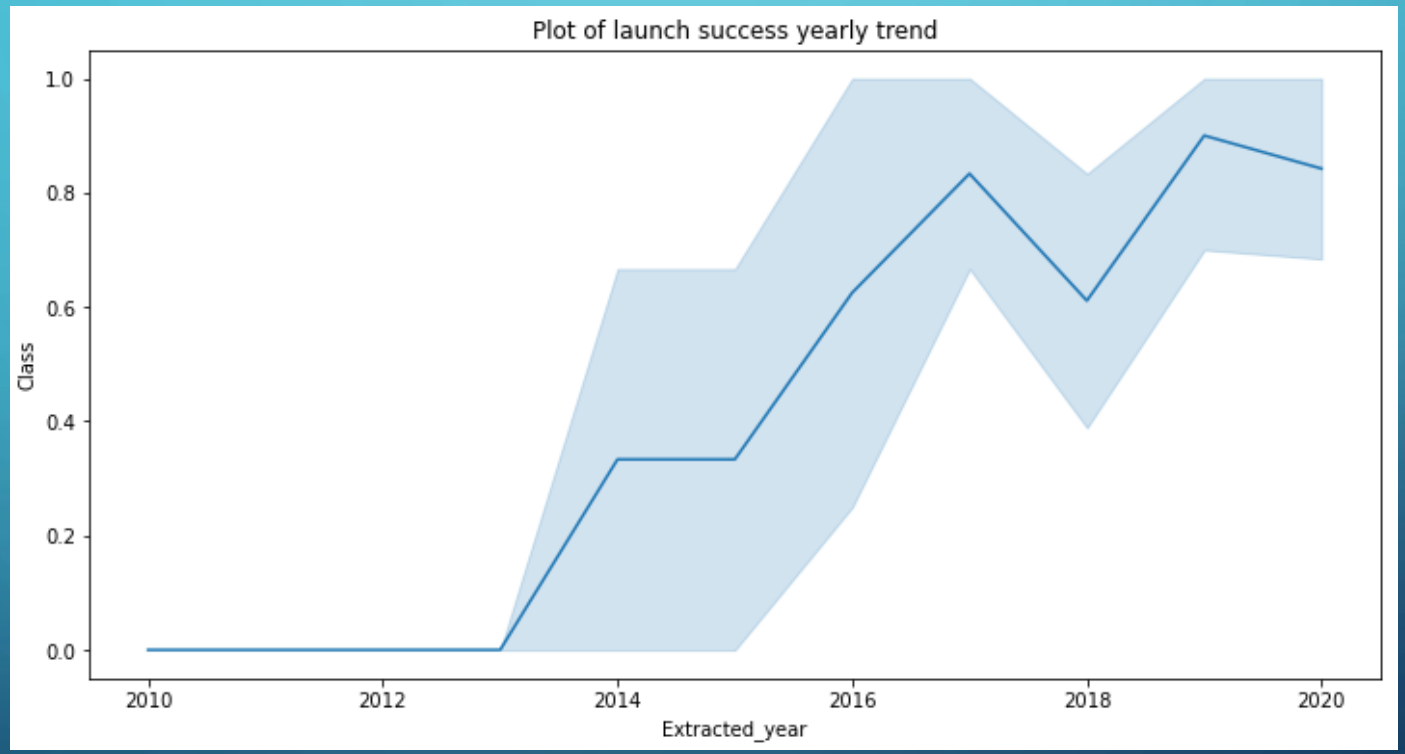


# PAYLOAD AND ORBIT TYPE

It can be noted that for PO, LEO, and ISS orbits, there is a higher success rate for landings when heavier payloads are carried.



# LAUNCH SUCCESS YEARLY TREND



# All Launch Site Names

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

```
%sql select Unique(LAUNCH_SITE) from SPACEX;
```

[9]

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS SLC-40 |
| KSC LC-39A |
| VAFB SLC-4E |

# LAUNCH SITE NAMES BEGIN WITH 'CCA'

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

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT LAUNCH_SITE from SPACEX where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| launch_site |
| --- |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

Task 3

# TOTAL PAYLOAD MASS

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

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[11] %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEX
```

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| payloadmass |
| --- |
| 256163 |

# AVERAGE PAYLOAD MASS BY F9

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEX
```

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| payloadmass |
| --- |
| 5692 |

# FIRST SUCCESSFUL GROUND LANDING DATE

List the date when the first successful landing outcome in ground pad was acheived.

*Hint:Use min function*

```
%sql select min(DATE) from SPACEX
```

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| 1 |
| --- |
| 2010-04-06 |

# SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select BOOSTER_VERSION from SPACEX where LANDING__OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;
```

[16]

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| booster_version |
| --- |
| F9 FT B1022 |
| F9 FT B1031.2 |

# TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

List the total number of successful and failure mission outcomes

```
[18] %sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEX GROUP BY MISSION_OUTCOME;
```

```
 * ibm_db_sa://snv10338:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

| missionoutcomes |
|---|
| 44 |
| 1 |

# CLASSIFICATION ACCURACY

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

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', 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 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

# SUMMARY

- In summary, the analysis shows that:
- Launch sites with higher flight volume have higher success rates.
- Launch success rates have been increasing since 2013 until 2020.
- Orbits such as ES-L1, GEO, HEO, SSO, and VLEO have higher success rates.
- KSC LC-39A is the launch site with the highest number of successful launches.
- The decision tree classifier algorithm is the most suitable for this task.

# THANK YOU