IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Gergely Katona
14.03.2024.

# Outline

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

# Executive Summary

## - Methodology Summary:

**Data Collection through API**: Gathering data from various sources using Application Programming Interfaces (APIs) to access and retrieve structured information.

**Data Collection with Web Scraping**: Extracting data from websites by programmatically navigating web pages and scraping relevant content.

**Data Wrangling**: Preprocessing and cleaning collected data to ensure its quality and compatibility with analysis tools.

**Exploratory Data Analysis with SQL**: Utilizing SQL (Structured Query Language) to explore and analyze datasets stored in relational databases, performing queries for insights.

**Exploratory Data Analysis with Data Visualization**: Visual exploration of data patterns and relationships using charts, graphs, and other visualizations to gain insights and identify trends.

**Interactive Visual Analytics with Folium**: Creating interactive maps and visualizations using Folium, a Python library for leaflet.js mapping.

**Machine Learning Prediction**: Applying machine learning algorithms to trained models for making predictions or classifications based on the analyzed data.

## - Summary of Results:

**Exploratory Data Analysis Result**: Findings and insights derived from exploring and analyzing the dataset, including trends, patterns, and anomalies discovered.

**Interactive Analytics in Screenshots**: Visual representations of interactive analytics using screenshots, showcasing dynamic data exploration and insights.

**Predictive Analytics Result**: Outcome of predictive modeling, including accuracy metrics, predictions, and any actionable insights derived from machine learning predictions.

# Introduction

**Project Background and Objectives:**

Space X offers Falcon 9 rocket launches on its website at a significantly lower cost of 62 million dollars compared to other providers, whose costs exceed 165 million dollars per launch. The primary reason for these savings is Space X's ability to reuse the first stage of the rocket. Thus, determining the success of the first stage landing can directly impact the overall cost of a launch. This information holds significance for potential competitors bidding against Space X for rocket launch contracts. The main objective of this project is to develop a machine learning pipeline capable of predicting the success of the first stage landing.

**Questions to Address:**

- What factors influence the successful landing of the rocket's first stage?

- How do various features interact to affect the likelihood of a successful landing?

- What operational conditions are necessary to ensure a successful landing program?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - SpaceX API
  - WEB scraping
- Perform data wrangling
  - Data cleaning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

# Data Collection – SpaceX API

1. **SpaceX API Access**:
    1. Data collection was initiated by sending GET requests to the SpaceX API.
    2. The response content was decoded as JSON using the .json() function call.
    3. JSON data was normalized into a pandas dataframe using .json_normalize().
2. **Data Cleaning**:
    1. Data cleaning procedures were implemented to address any inconsistencies or missing values.
    2. Missing values were identified and filled in where necessary to ensure data completeness.
3. **Web Scraping from Wikipedia**:
    1. Web scraping techniques were applied to extract Falcon 9 launch records from Wikipedia.
    2. The launch records were extracted as an HTML table using BeautifulSoup.
    3. The HTML table was parsed and converted into a pandas dataframe for subsequent analysis.

https://api.spacexdata.com/v4/payloads/

https://github.com/katonagergely/datascientist/blob/main/EDA-dataviz.ipynb

# Data Collection - Scraping

The data was scraped from the following

URL: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

This website specifically contains information about Falcon 9 launches.

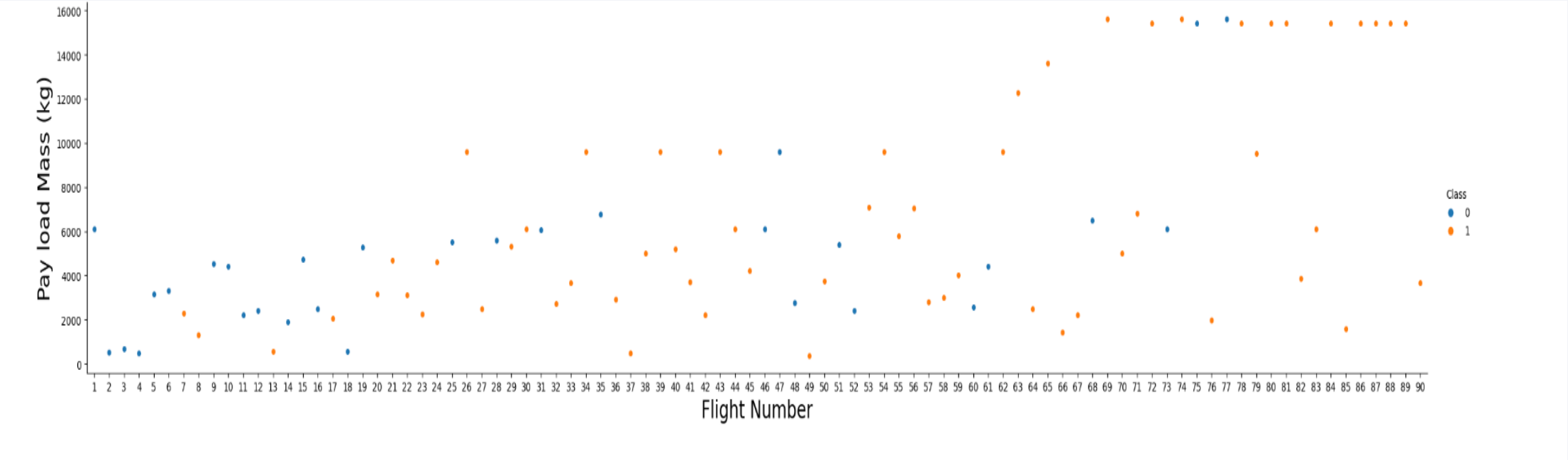
https://github.com/katonagergely/datascientist/blob/main/Webscraping.ipynb

# Data Wrangling

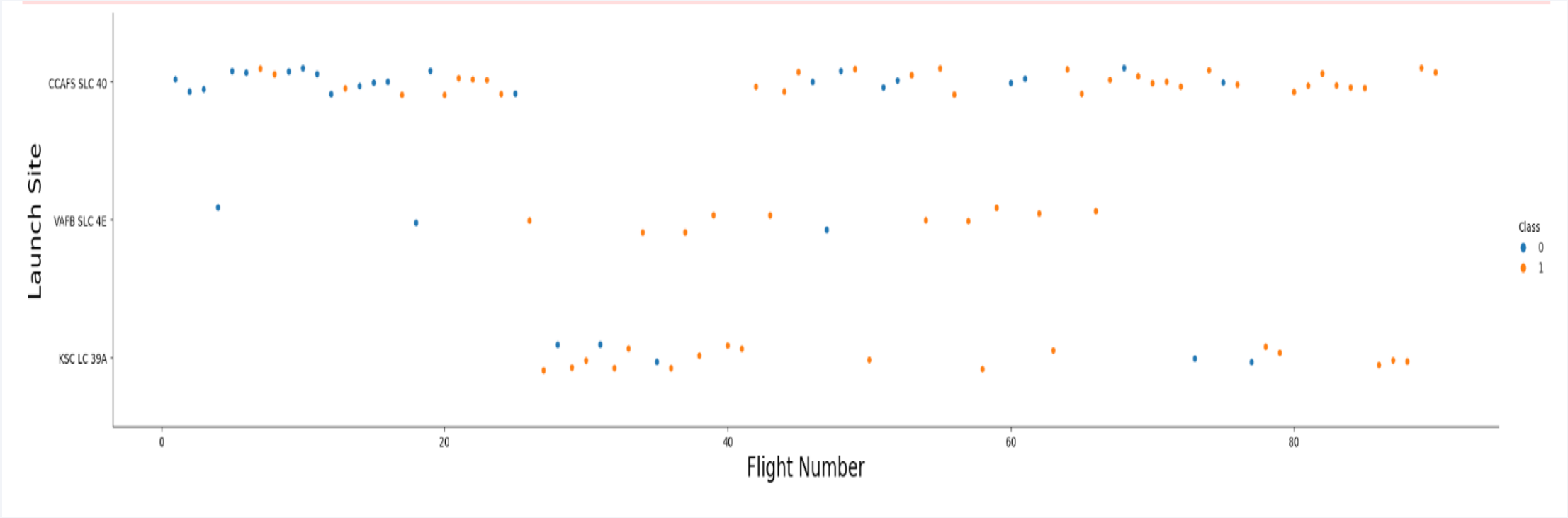
https://github.com/katonagergely/datascientist/blob/main/Spacex-Data%20wrangling.ipynb

# EDA with Data Visualization

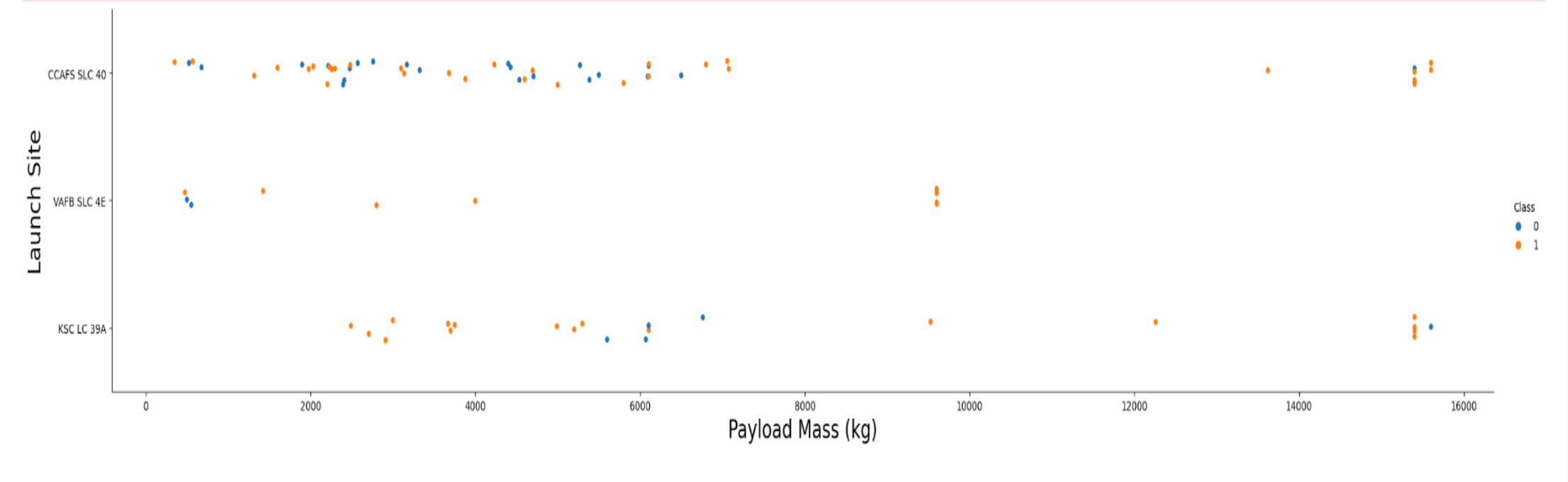
https://github.com/katonagergely/datascientist/blob/main/EDA-dataviz.ipynb



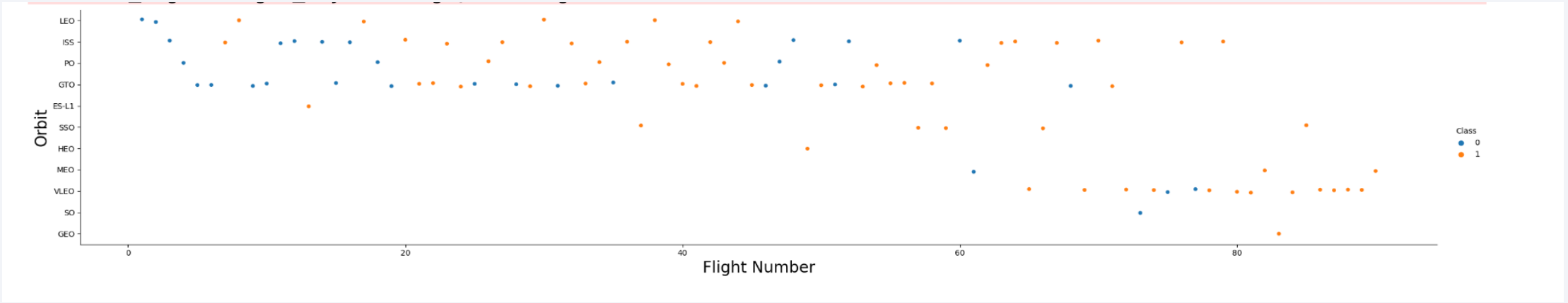
# EDA with Data Visualization

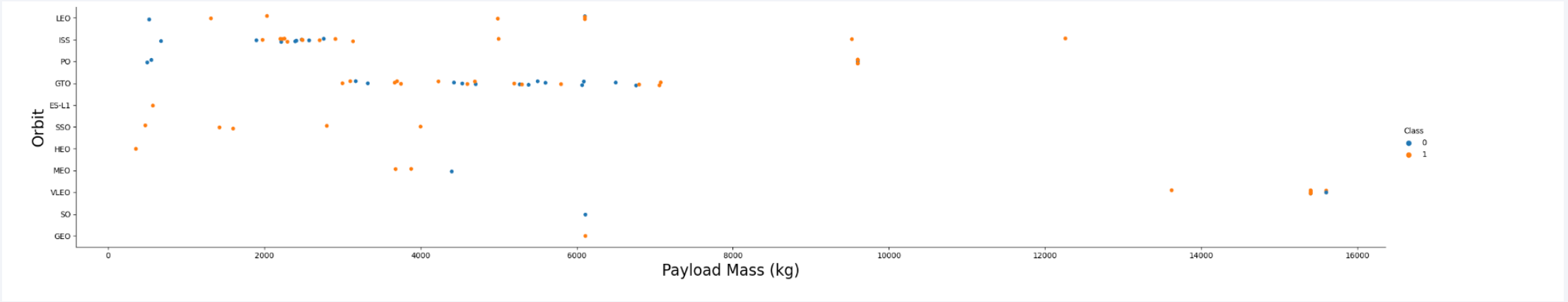


# EDA with Data Visualization

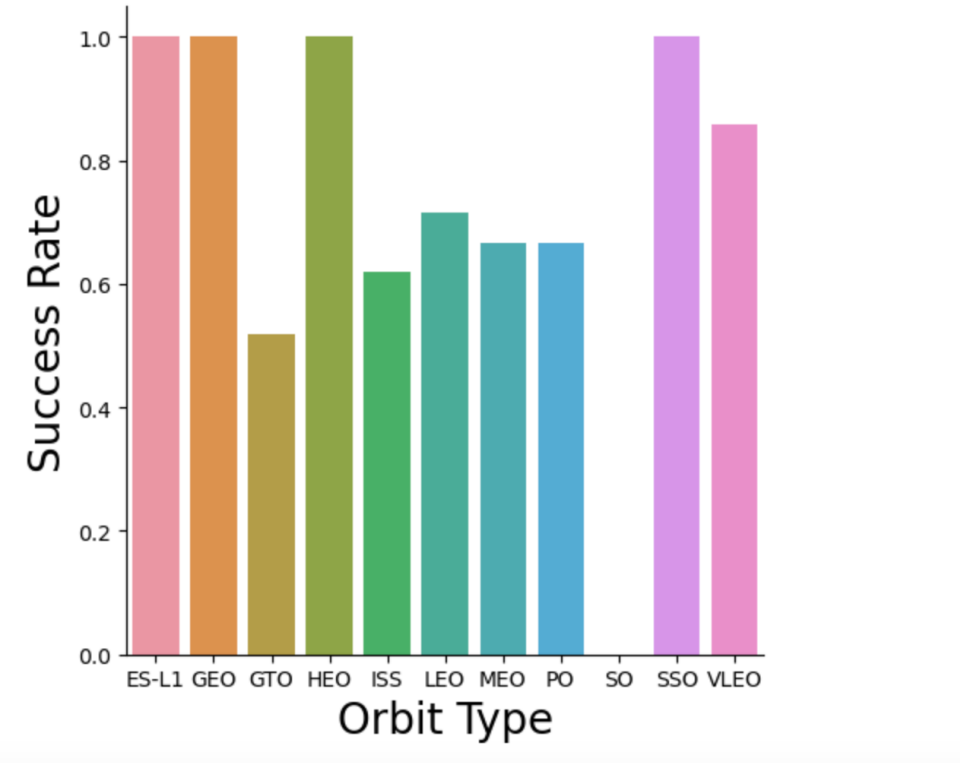


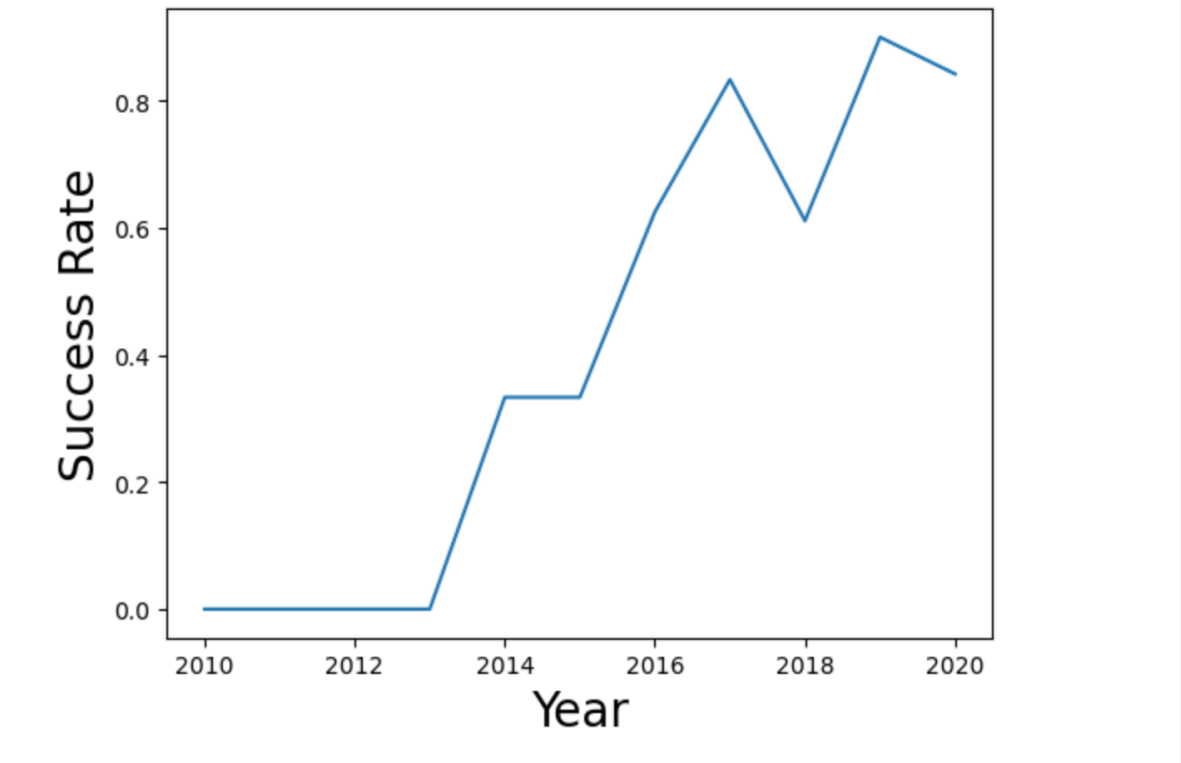
# EDA with Data Visualization



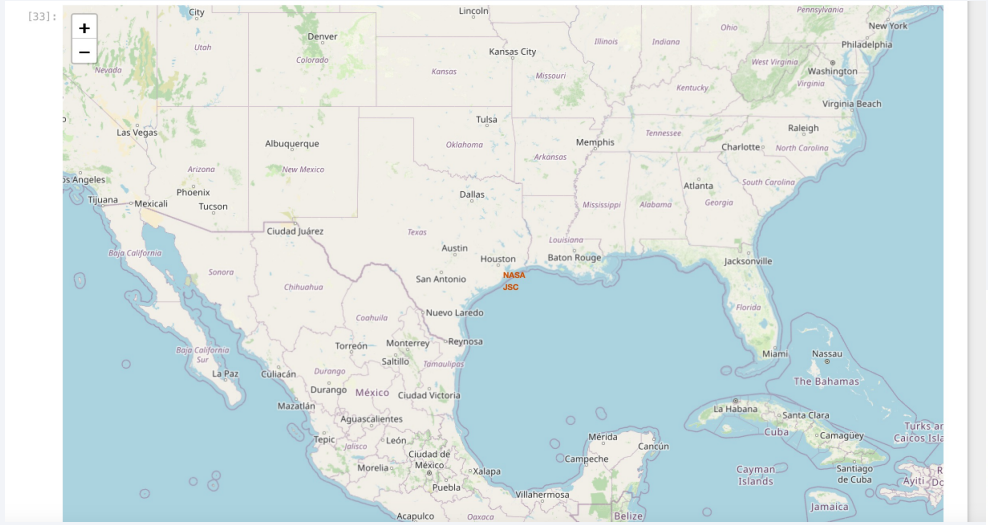



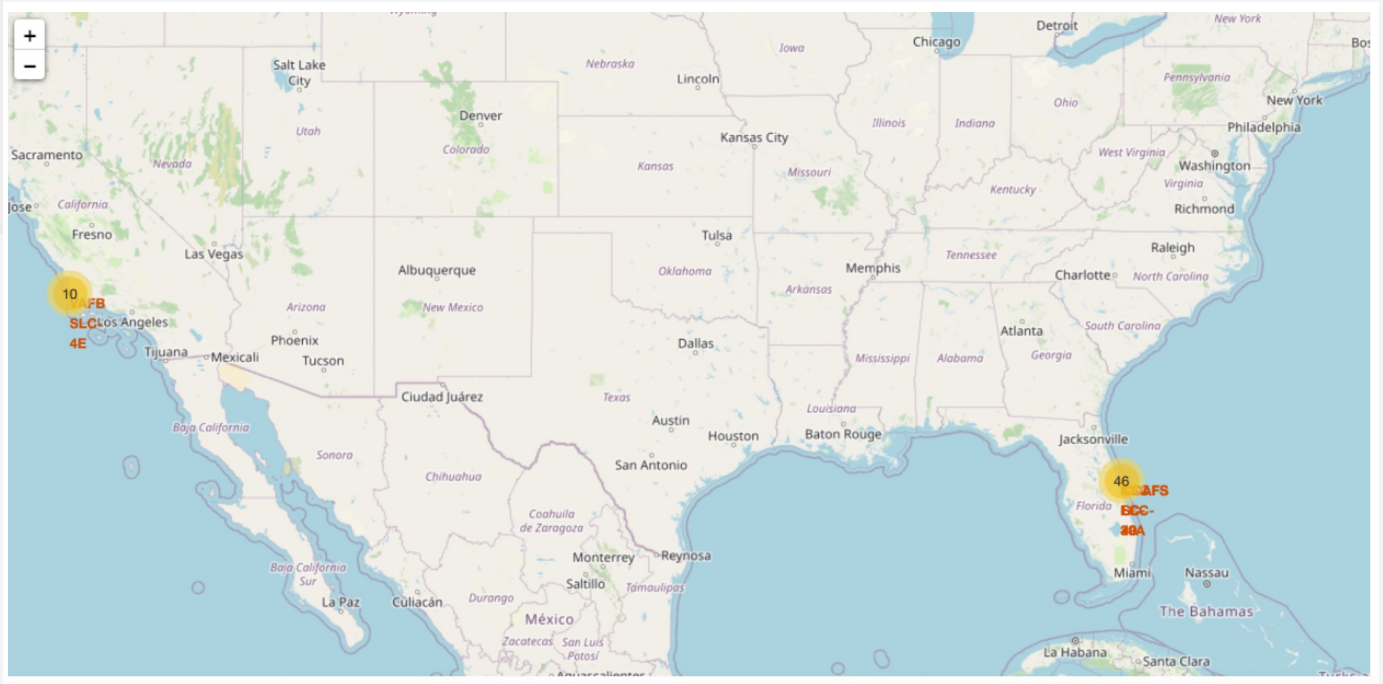
# EDA with Data Visualization



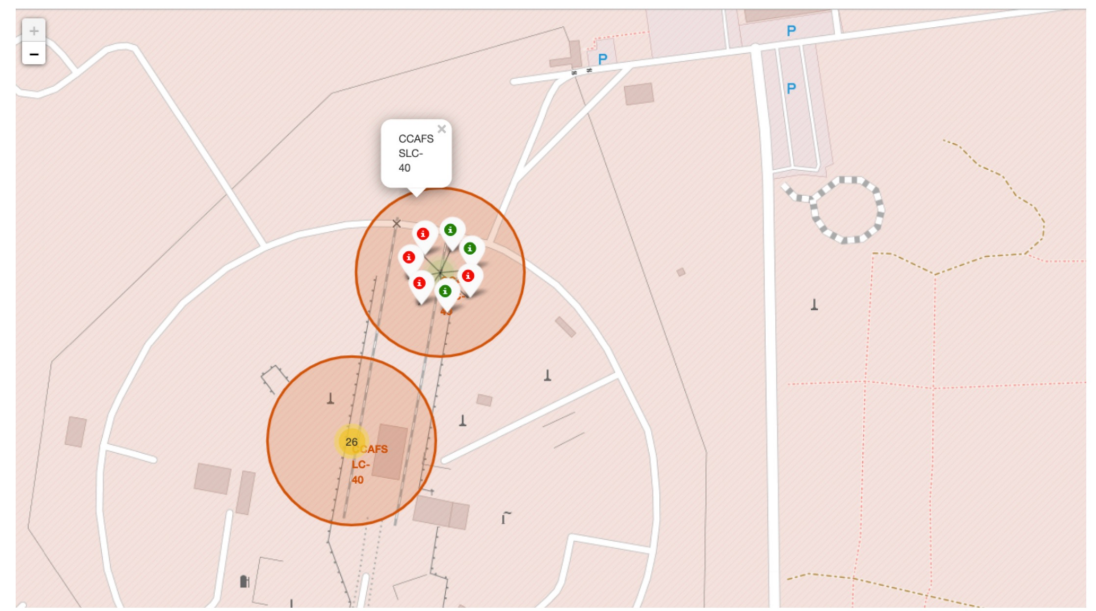



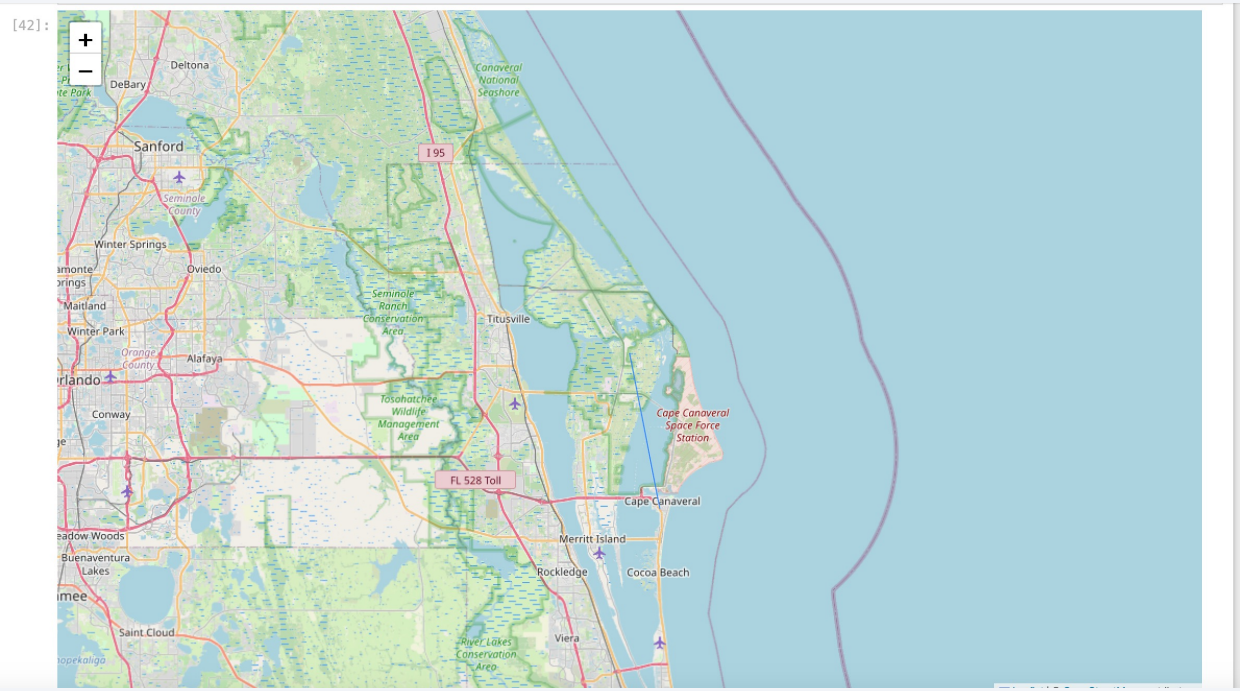
# Build an Interactive Map with Folium



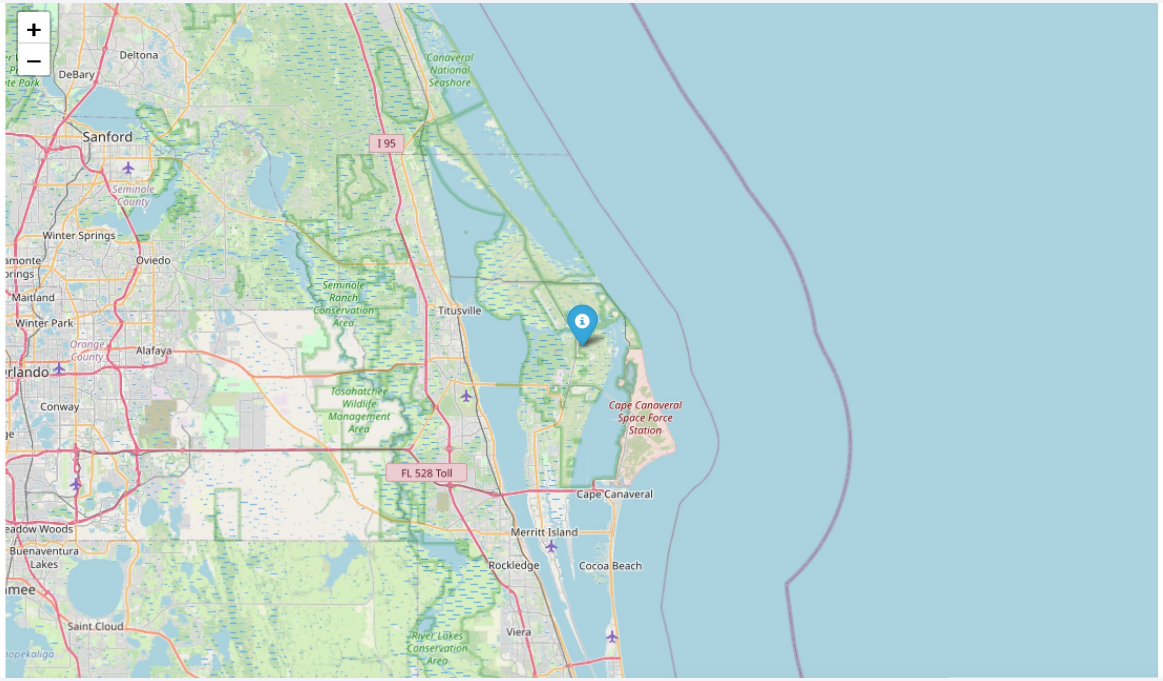
# Build an Interactive Map with Folium



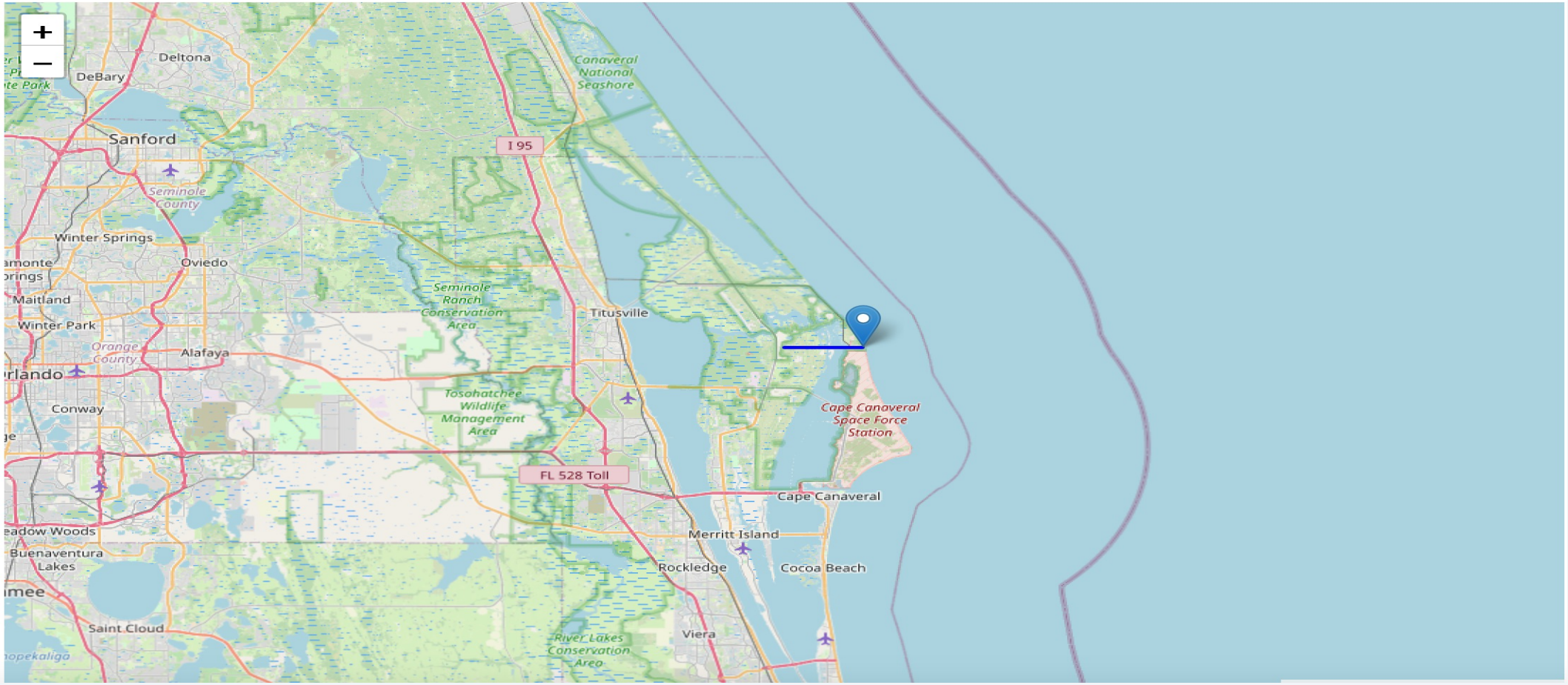
# Build an Interactive Map with Folium


[51]:

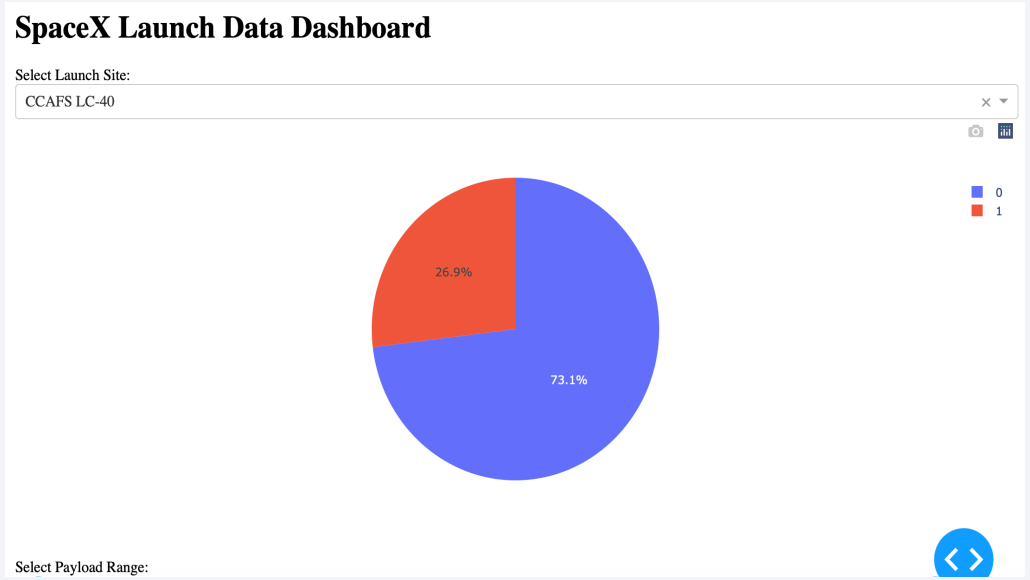


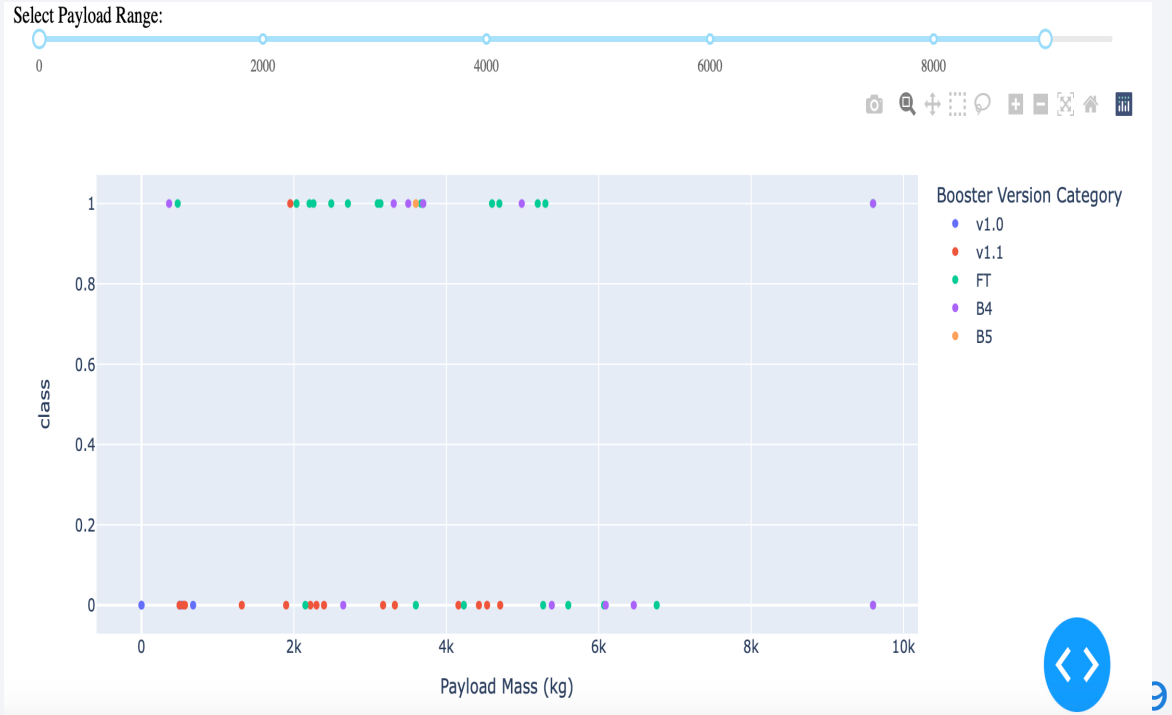
# Build an Interactive Map with Folium

[56]:



# Build a Dashboard with Plotly Dash



# Predictive Analysis (Classification)

Logistic regression

SVM (Support Vector Machine)

Decision tree

KNN (K nearest neighbors)

https://github.com/katonagergely/datascientist/blob/main/SpaceX_Machine_Learning_Prediction_Part.ipynb

# Predictive Analysis (Classification)

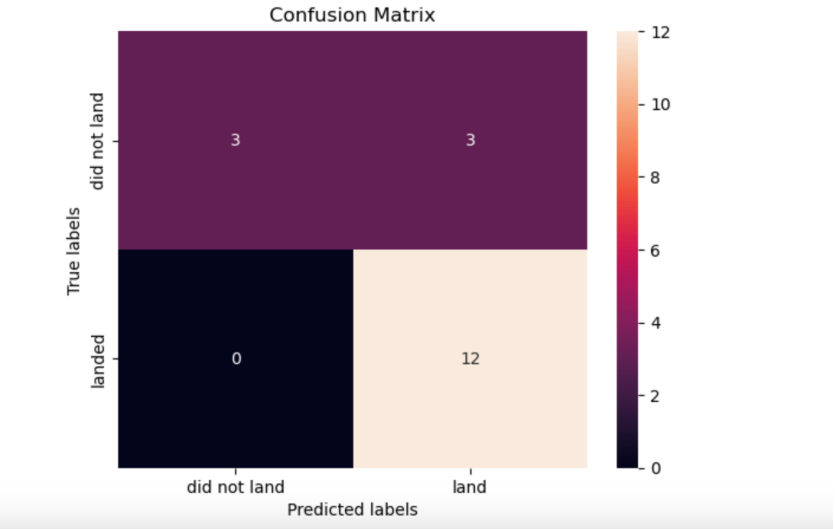
Logistic regression

```
[26]: lr_accuracy = logreg_cv.score(X_test, Y_test)
      lr_accuracy
```

```
[26]: 0.8333333333333334
```

Lets look at the confusion matrix:

```
[27]: yhat=logreg_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



# Predictive Analysis (Classification)

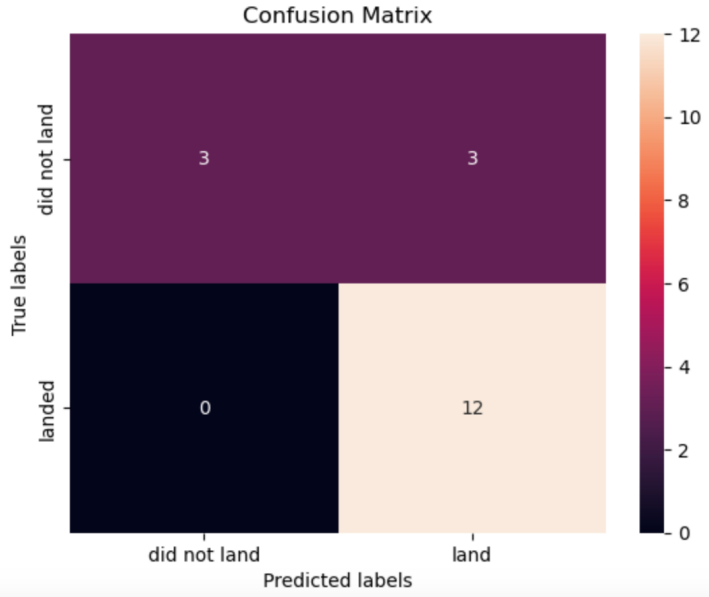
## SVM (Support Vector Machine)

```
[32]: svm_accuracy = svm_cv.score(X_test, Y_test)
      svm_accuracy
```

```
[32]: 0.8333333333333334
```

We can plot the confusion matrix

```
[33]: yhat=svm_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



# Predictive Analysis (Classification)

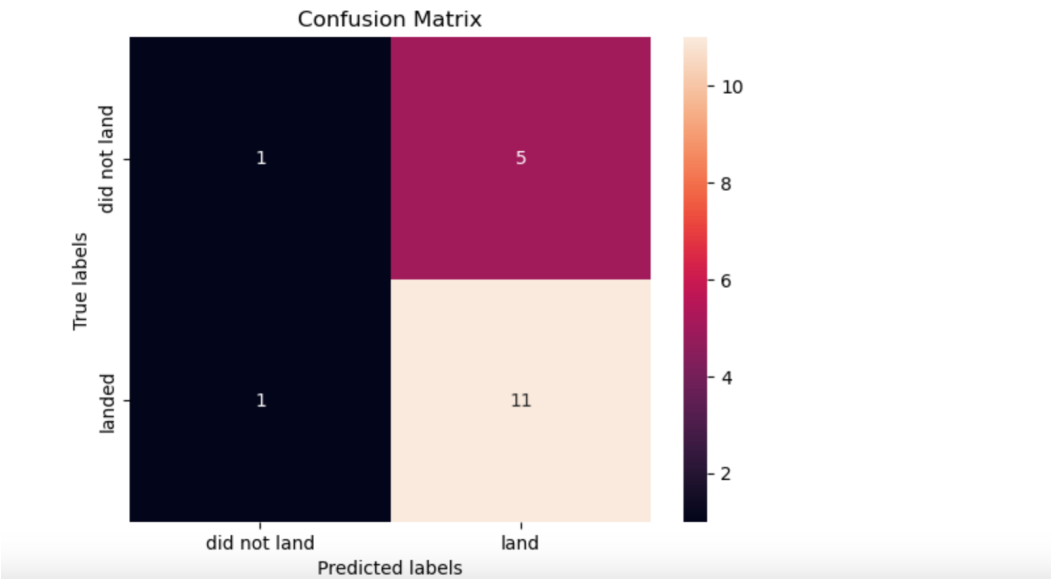
Decision tree

```
[39]: tree_accuracy = tree_cv.score(X_test, Y_test)
      tree_accuracy
```

```
[39]: 0.6666666666666666
```

We can plot the confusion matrix

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



# Predictive Analysis (Classification)

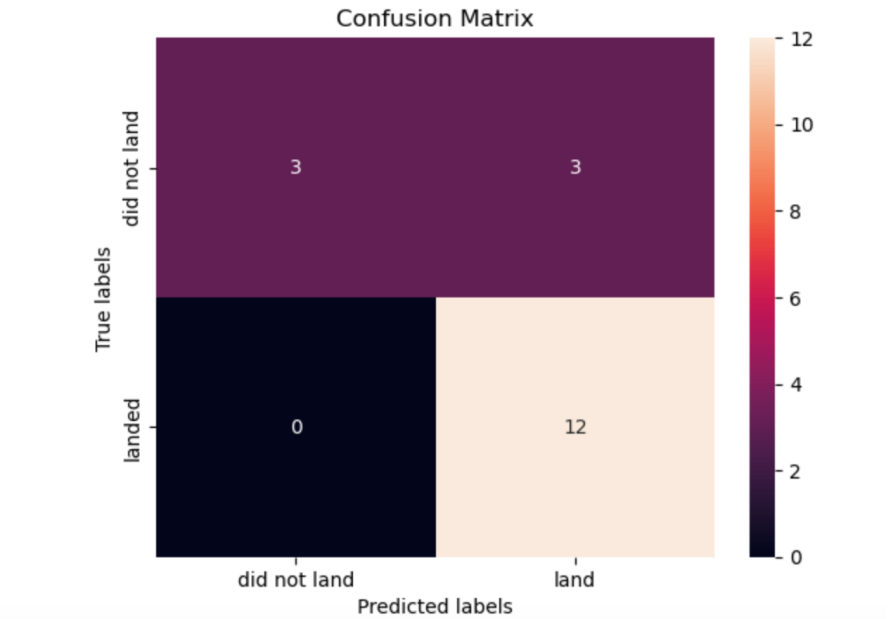
KNN (K nearest neighbors)

```
[44]: accuracy = knn_cv.score(X_test, Y_test)
      print("Accuracy:", accuracy)
```

```
Accuracy: 0.8333333333333334
```

We can plot the confusion matrix

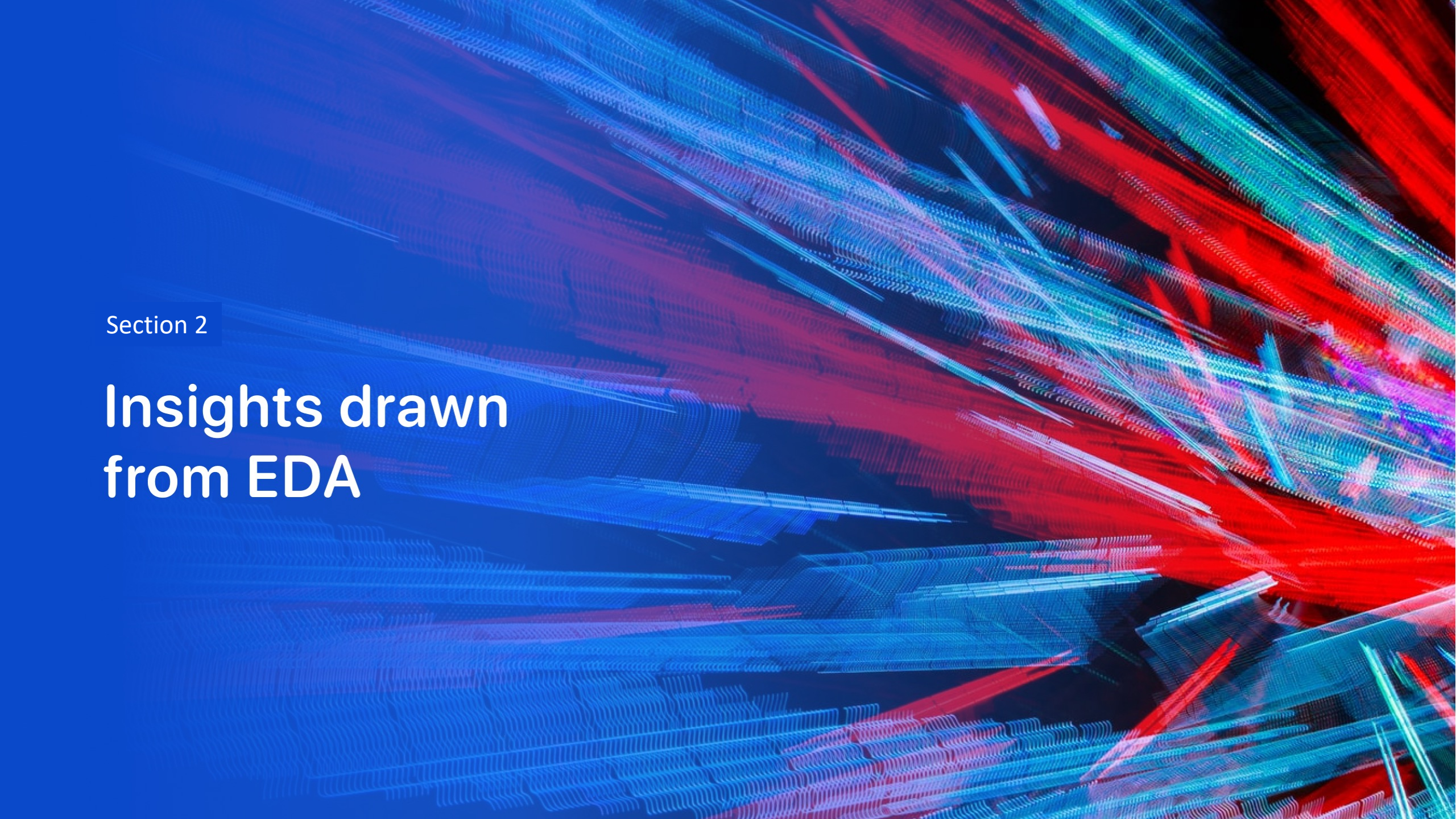
```
[45]: yhat = knn_cv.predict(X_test)
      plot_confusion_matrix(Y_test,yhat)
```



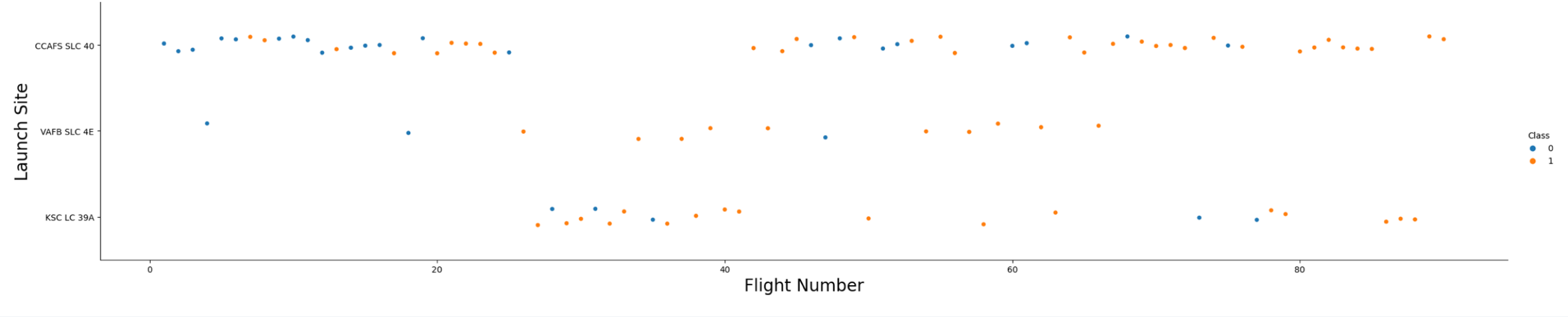
# Results

When comparing the results of all four models side by side, it becomes evident that they exhibit identical accuracy scores and confusion matrices during testing on the test set. As a result, their rankings are determined based on their GridSearchCV best scores. Ranked in descending order of GridSearchCV best scores, the models are as follows:

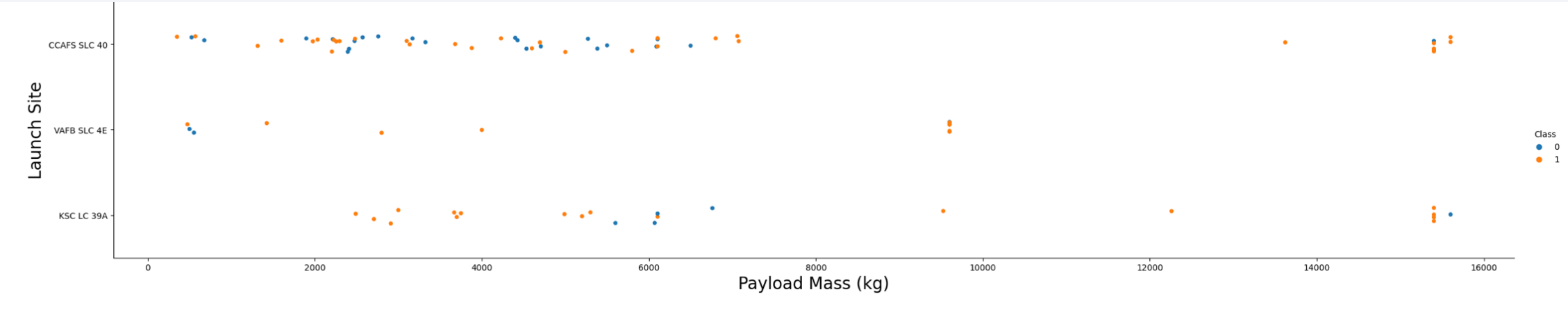
| | Best scores |
|---|---|
| Logistic regression | 0.846429 |
| SVM | 0.848214 |
| Decision tree | 0.889286 |
| KNN | 0.848214 |


Section 2
Insights drawn from EDA

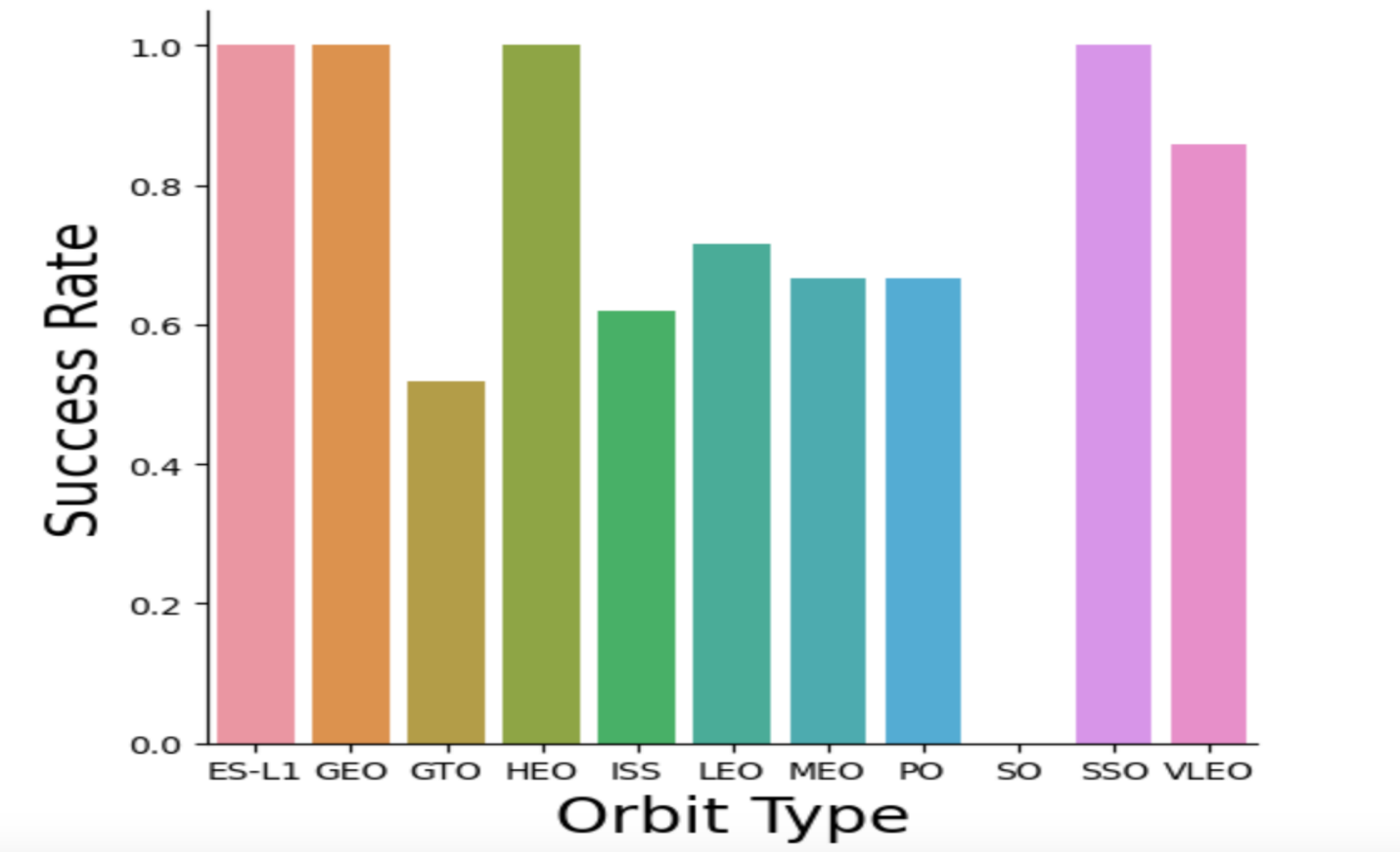
# Flight Number vs. Launch Site



# Payload vs. Launch Site



# Success Rate vs. Orbit Type

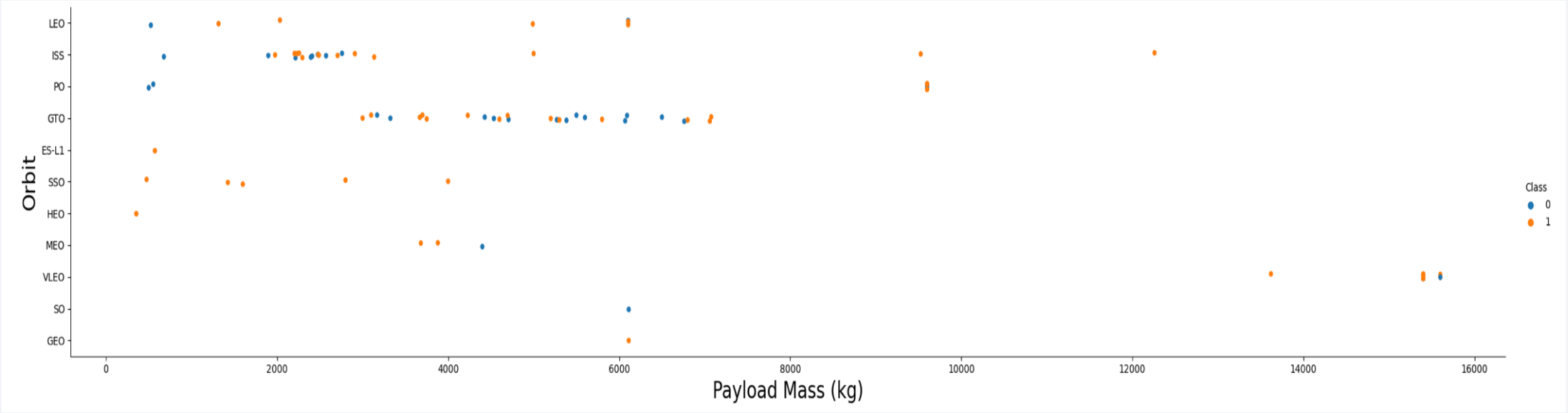


# Flight Number vs. Orbit Type

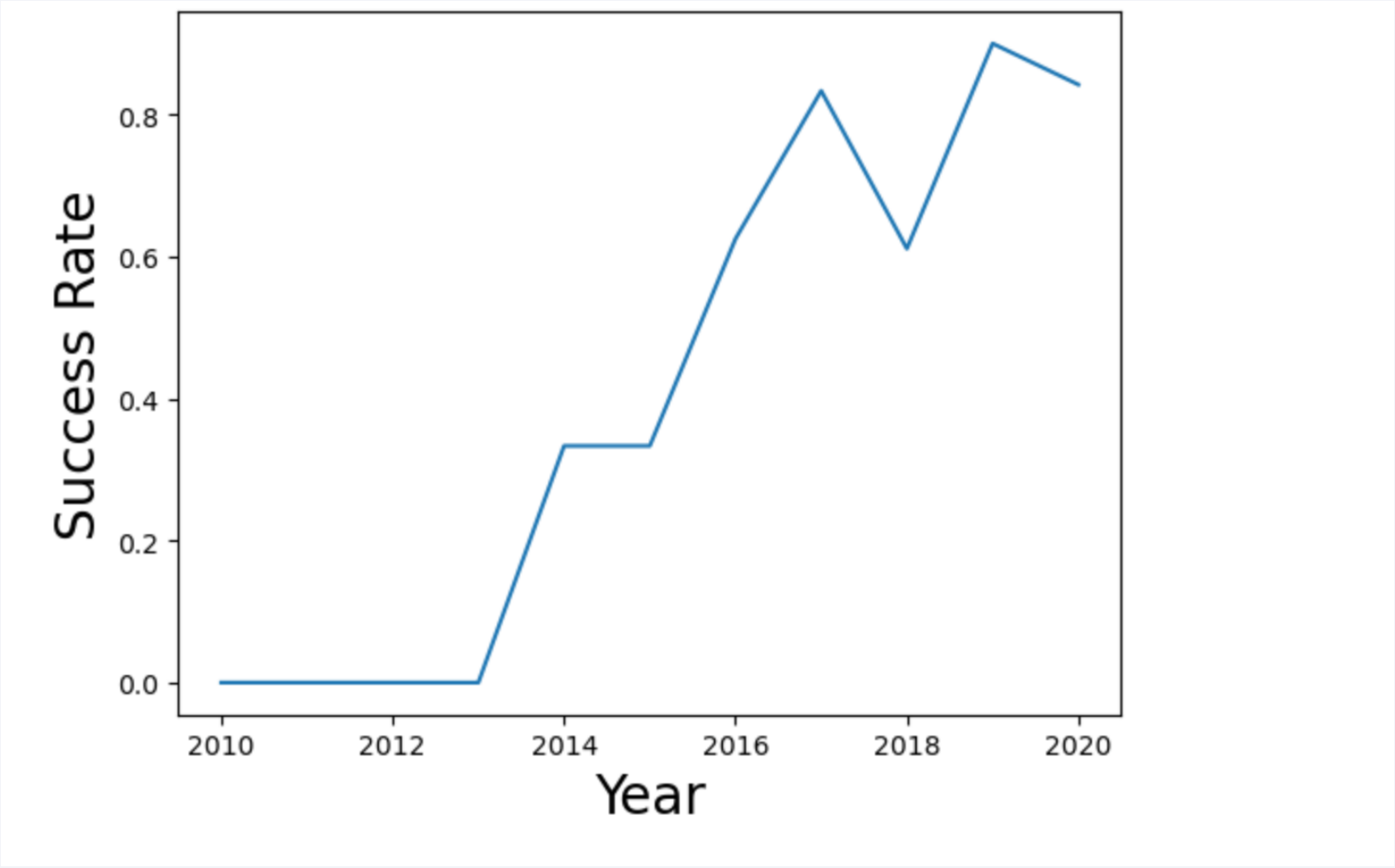
- Show a scatter point of Flight number vs. Orbit type

- Show the screenshot of the scatter plot with explanations

# Payload vs. Orbit Type



# Launch Success Yearly Trend



# All Launch Site Names

[32]:

| | FlightNumber | PayloadMass | Orbit | LaunchSite | Flights | GridFins | Reused | Legs | LandingPad | Block | ReusedCount | Serial |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 6104.959412 | LEO | CCAFS SLC 40 | 1 | False | False | False | NaN | 1.0 | 0 | B0003 |
| **1** | 2 | 525.000000 | LEO | CCAFS SLC 40 | 1 | False | False | False | NaN | 1.0 | 0 | B0005 |
| **2** | 3 | 677.000000 | ISS | CCAFS SLC 40 | 1 | False | False | False | NaN | 1.0 | 0 | B0007 |
| **3** | 4 | 500.000000 | PO | VAFB SLC 4E | 1 | False | False | False | NaN | 1.0 | 0 | B1003 |
| **4** | 5 | 3170.000000 | GTO | CCAFS SLC 40 | 1 | False | False | False | NaN | 1.0 | 0 | B1004 |

# Conclusion

This project aims to forecast whether the first stage of a Falcon 9 launch will successfully land, which in turn influences the overall cost of the launch. Each aspect of a Falcon 9 launch, including payload mass and orbit type, can potentially impact the mission's outcome. Multiple machine learning algorithms are utilized to analyze historical Falcon 9 launch data and develop predictive models capable of anticipating launch outcomes. Among the four machine learning algorithms utilized, the decision tree algorithm yielded the most effective predictive model.

# Thank you!