IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

UGWUTE CHARLES OGBONNA
4th Dec 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Visual Analysis with Folium and Dash
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result (Machine Learning Lab)

Introduction

- **Project background and context**

  SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

- **Problems you want to find answers**
  - What features determine the landing outcomes?
  - What are the relationships between those features?
  - What state of these features determine the best outcome?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected through SpaceX RESTful API and Web Scraping from Wikipedia
- Perform data wrangling
  - Data processed through one-hot encoding for categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data split into training testing data.
  - Training Data used with different classification models.
  - Hyperparameter grid search applied for parameter tuning

Data Collection

Data collection is the process of gathering data for use in business decision-making, strategic planning, research and other purposes. It's a crucial part of data analytics applications and research projects: Effective data collection provides the information that's needed to answer questions, analyze business performance or other outcomes, and predict future trends, actions and scenarios.

Data Collection – SpaceX API

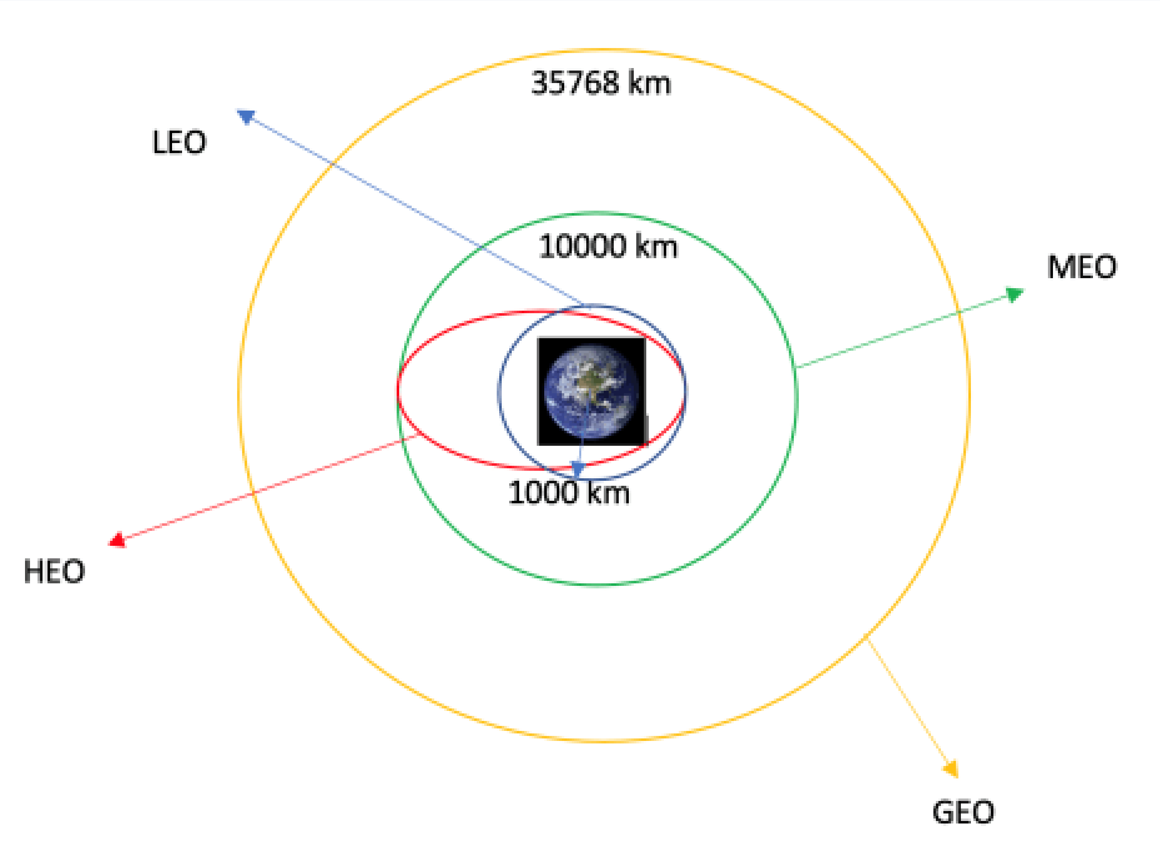
- Get request for rocket launch data using API
- Use json_normalize() method to convert JSON result to Dataframe

Data Collection - Scraping

- Request the Falcon9 Launch Wikipedia page from url
- Create a BeautifulSoup from the HTML response
- Extract all column/variable names from the HTML header

Data Wrangling

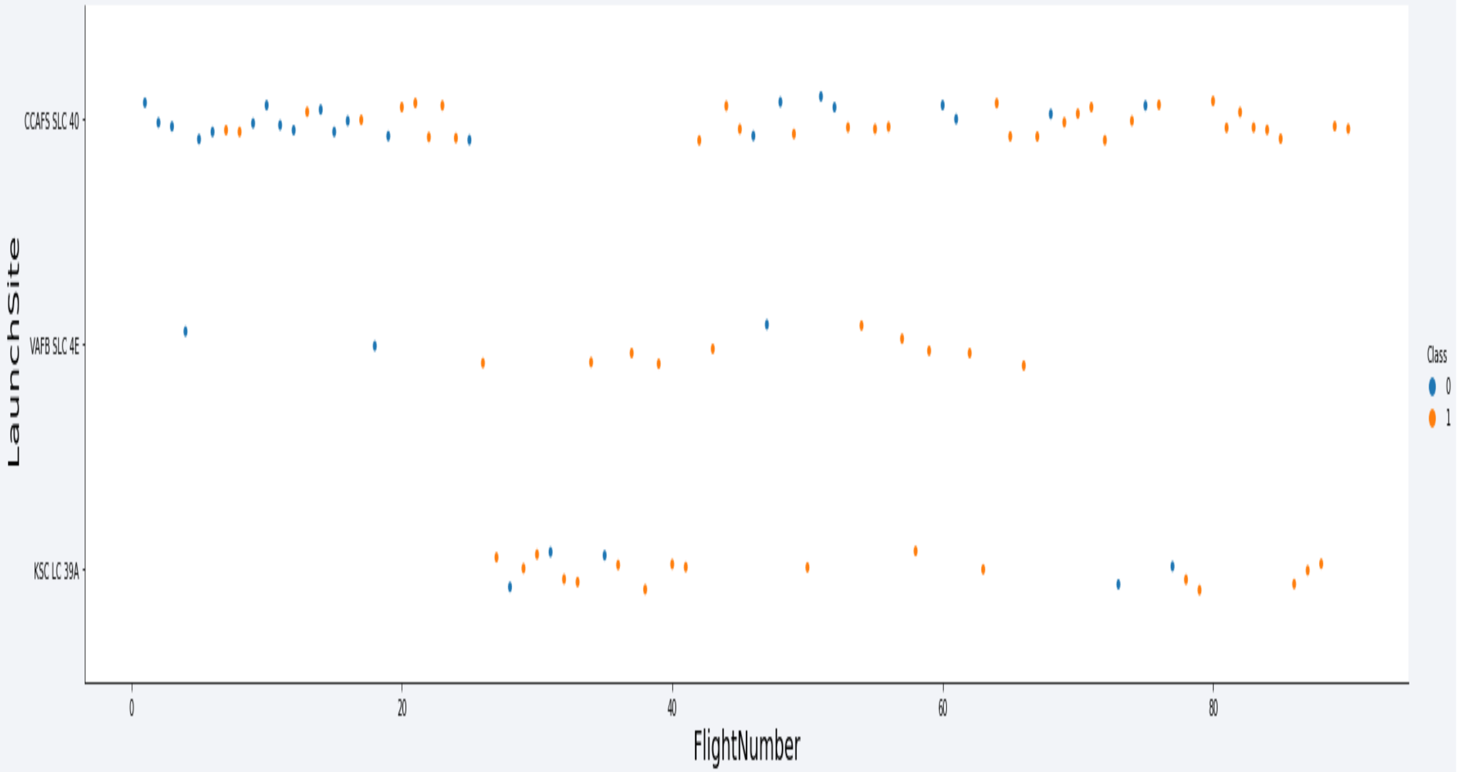
Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis.



EDA with Data Visualization

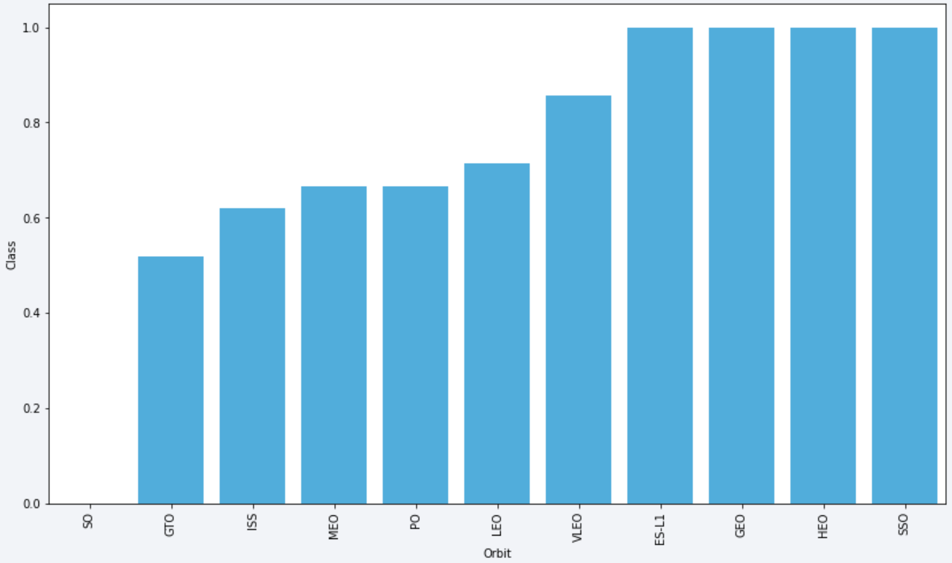
- I used scatter graph to find the relationship between the attributes such as between:
  - Payload and Flight Number.
  - Flight Number and Launch Site.
  - Payload and Launch Site.
  - Flight Number and Orbit Type.
  - Payload and Orbit Type.

Scatter plots show dependency of attributes on each other. Once a pattern is determined from the graphs. It’s very easy to see which factors affecting the most to the success of the landing outcomes.

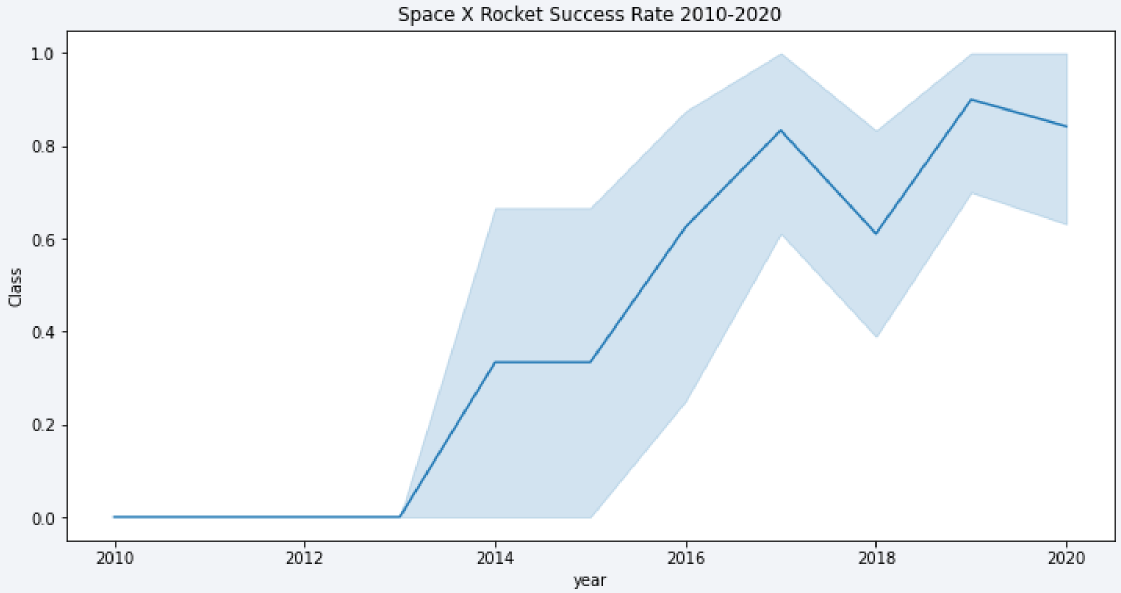


EDA with Data Visualization

- Bar graphs is one of the easiest way to interpret the relationship between the attributes. In this case, we used a bar chart to determine which orbits have the highest probability of success.



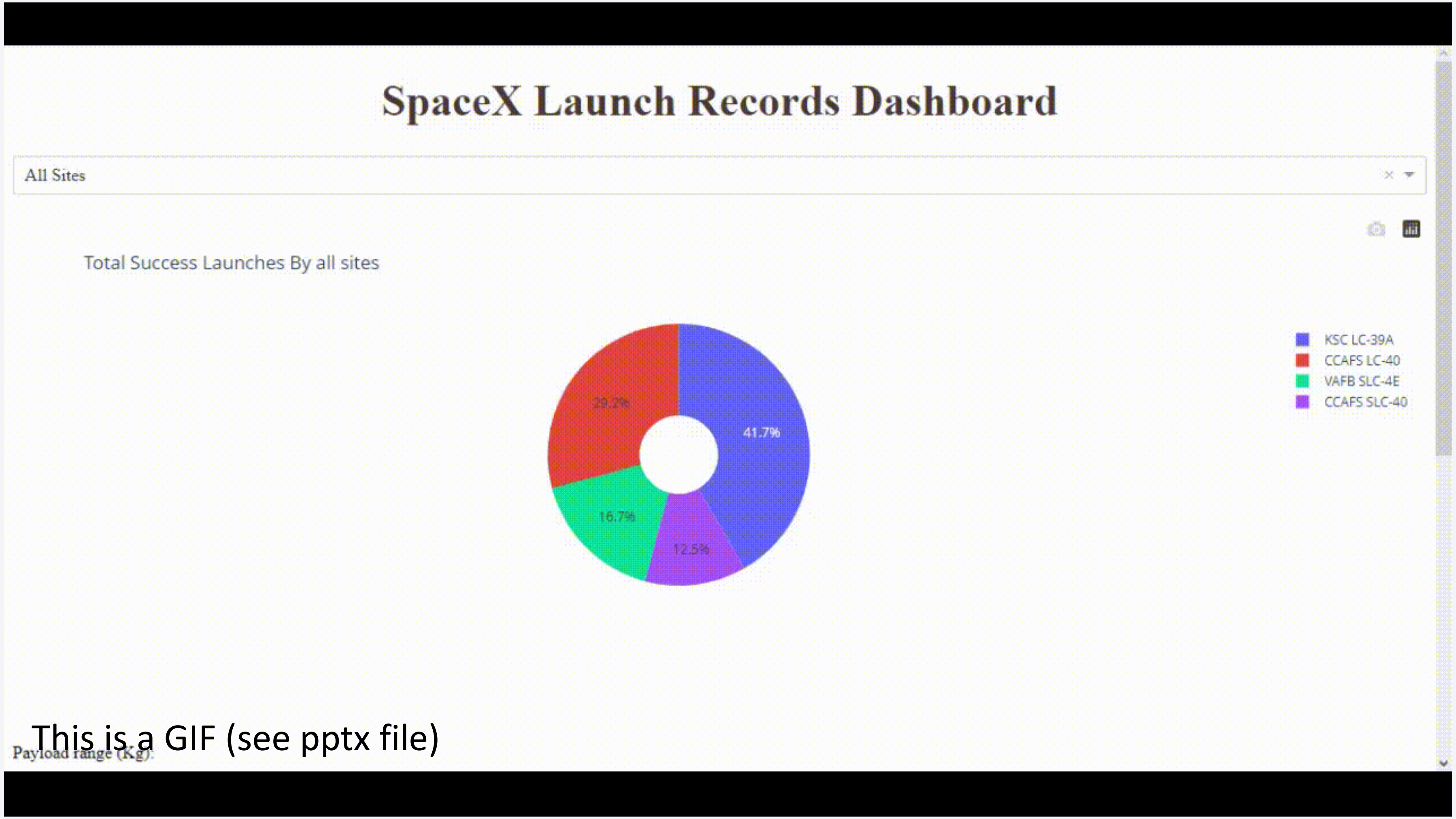
I used a line graph to show a trends or pattern of the attribute over time which in this case, is used for see the launch success yearly trend.



Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash which allowing the user to play around with the data as they need.
- We plotted pie charts showing the total launches by a certain sites.
- We then plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- Explain why you added those plots and interactions

Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard

All Sites

Total Success Launches By all sites

Payload range (Kg):

This is a GIF (see pptx file)

Predictive Analysis (Classification)

- We loaded the data using numpyand 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.

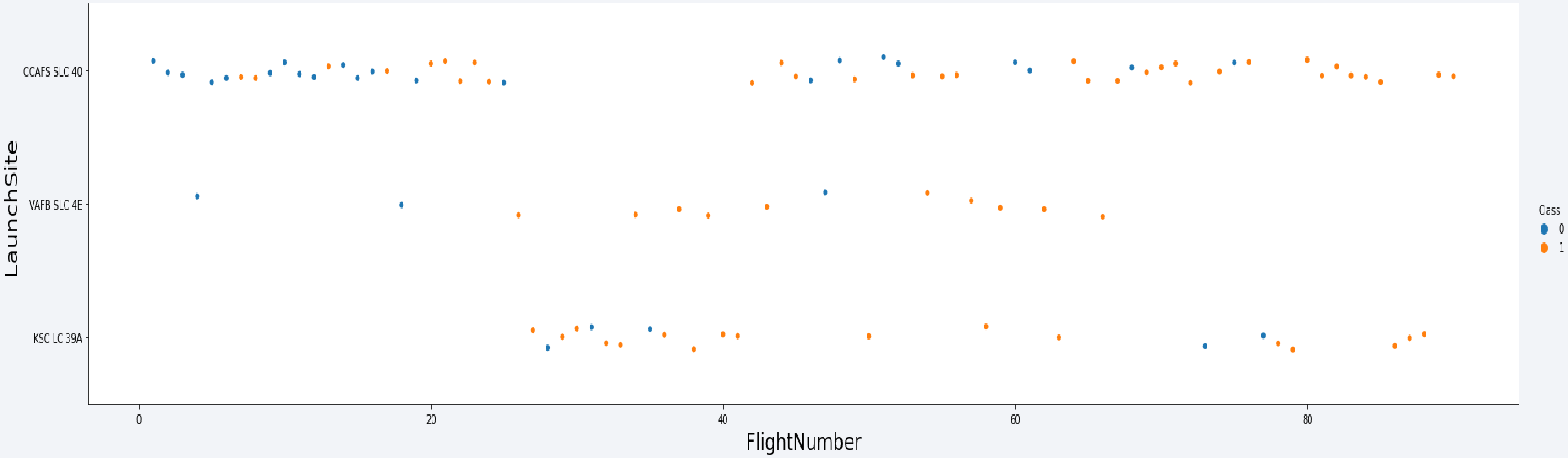
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

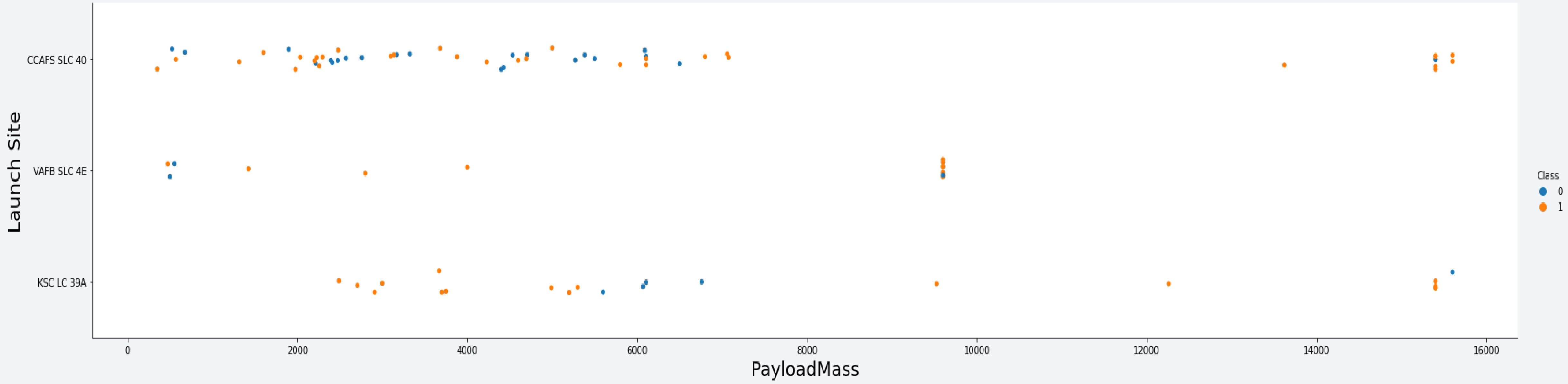
Insights drawn from EDA

Flight Number vs. Launch Site



This plot shows that the larger the flights number of each launch site, the greater the success rate will be.

Payload vs. Launch Site

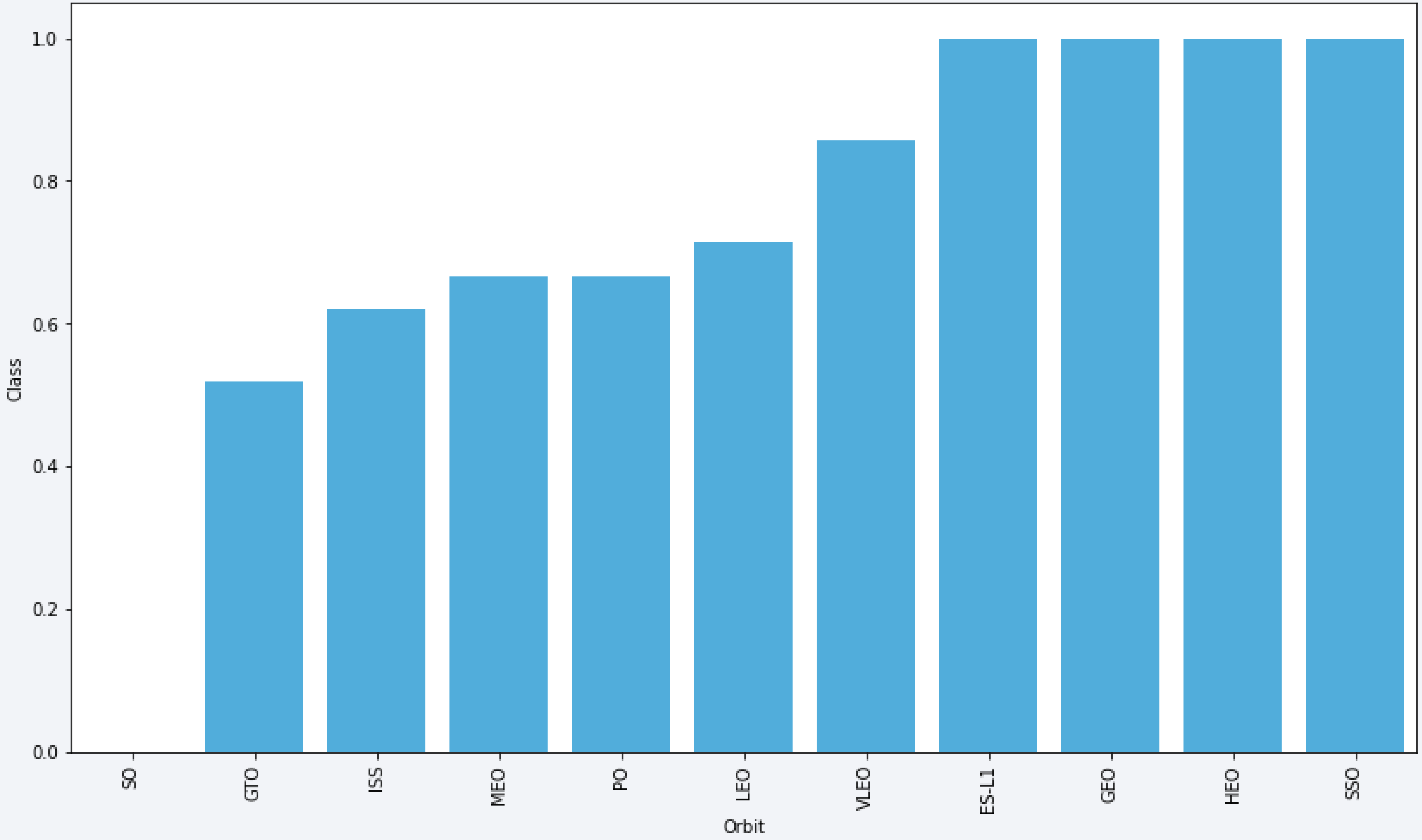


This scatter plot shows once the pay load mass is greater than 7000kg, the probability of the success rate will be highly increased.

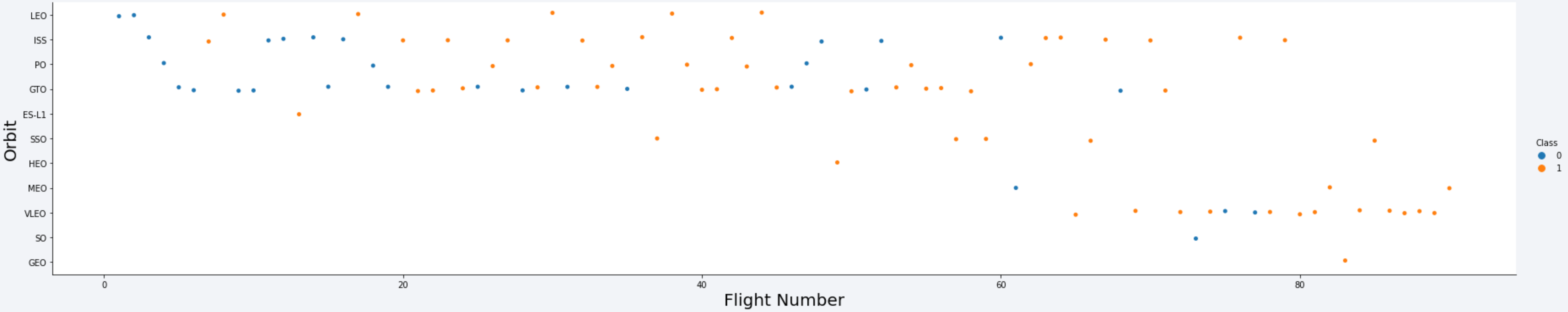
However, there is no clear pattern to say the launch site is dependent to the pay load mass for the success rate.

Success Rate vs. Orbit Type

This chart depicted the possibility of the orbits to influences the landing outcomes as some orbits has 100% success rate such as SSO, HEO, GEO AND ES-L1 while SO orbit produced 0% rate of success.

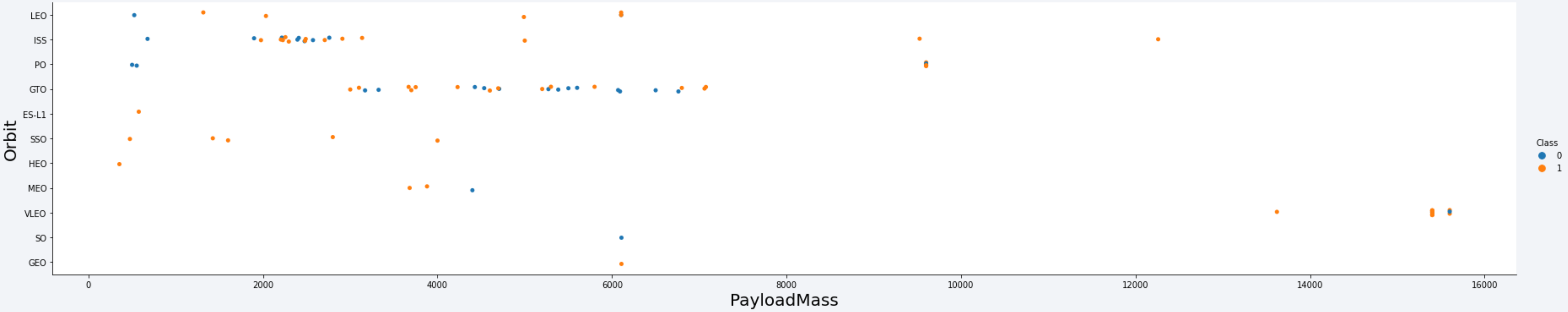


Flight Number vs. Orbit Type



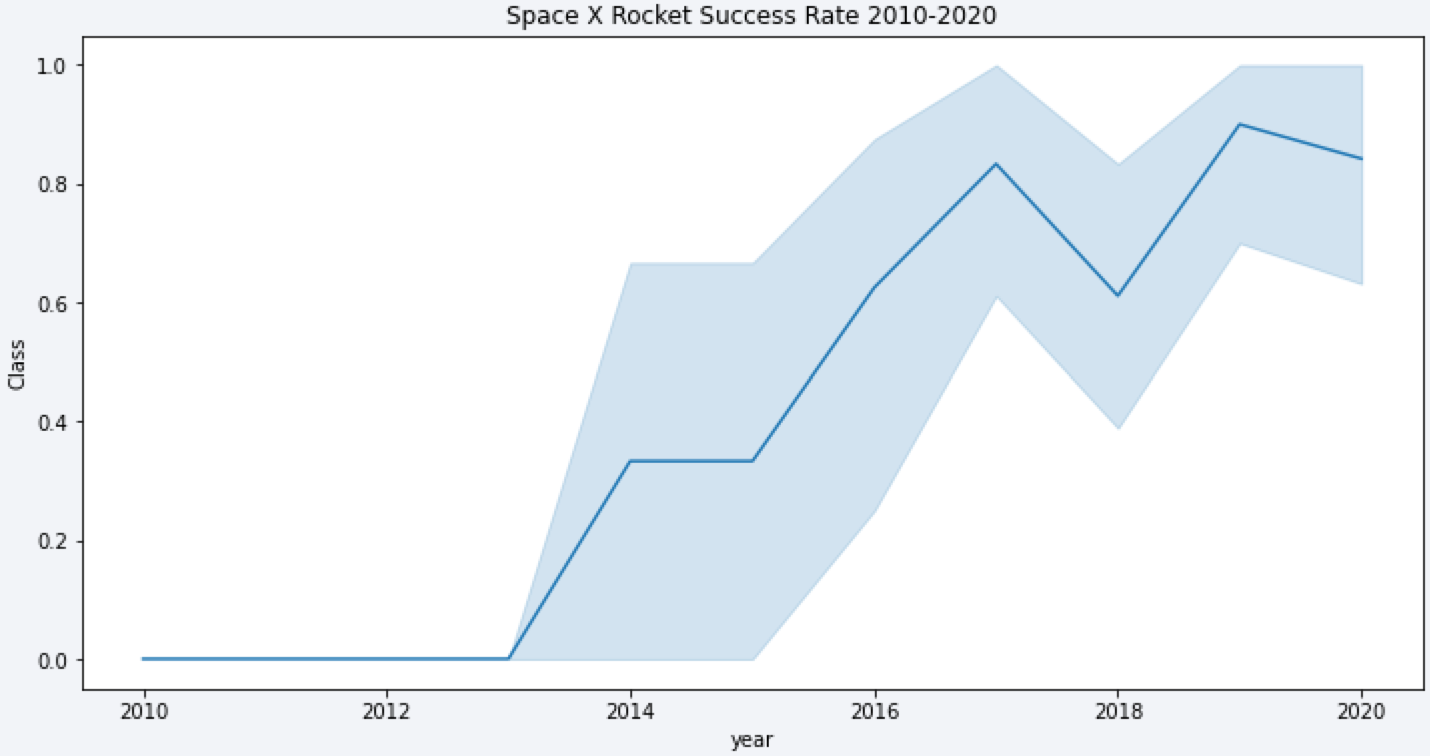
We see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



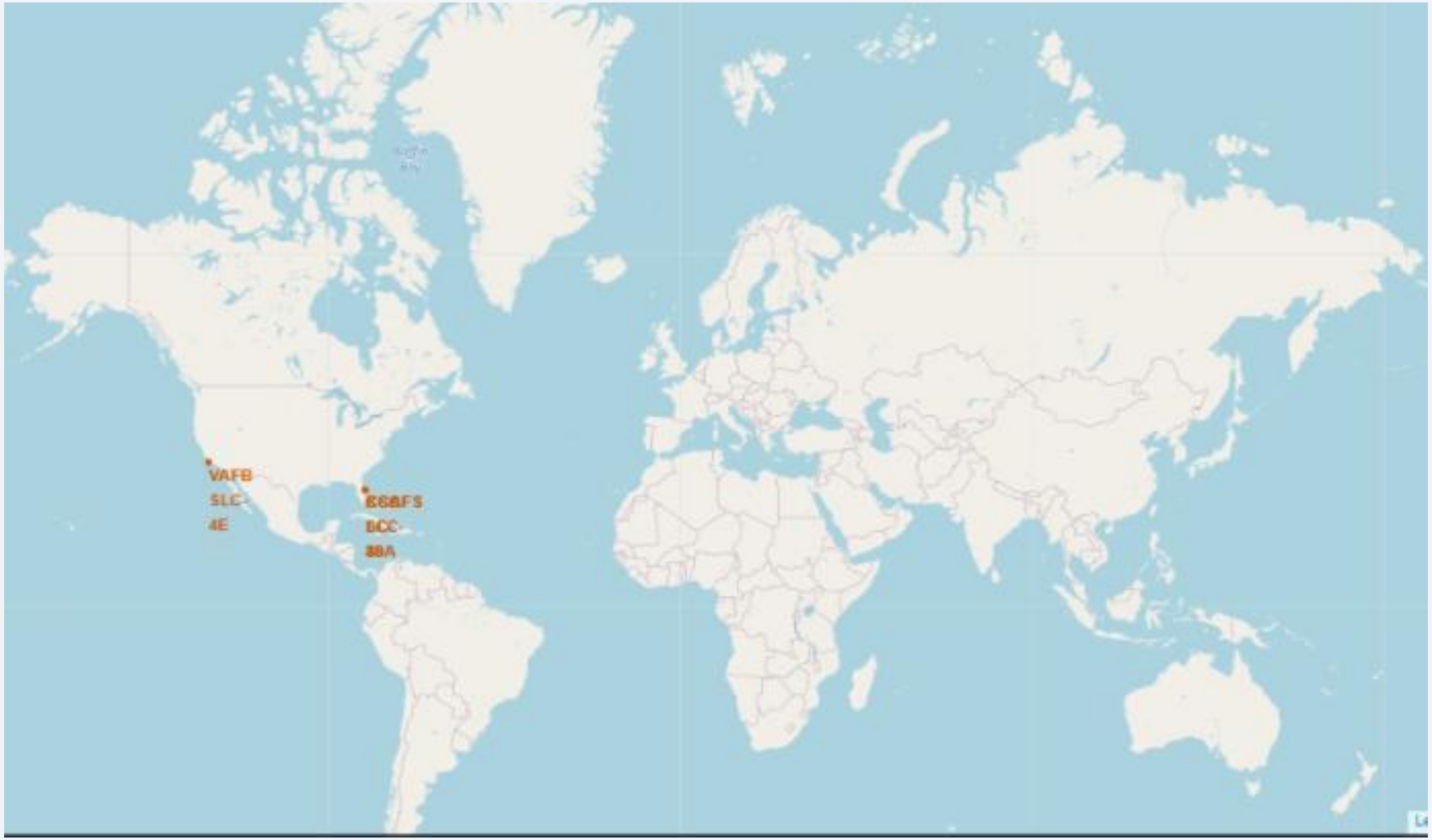
you can observe that the success rate since 2013 kept increasing till 2020

Section 3

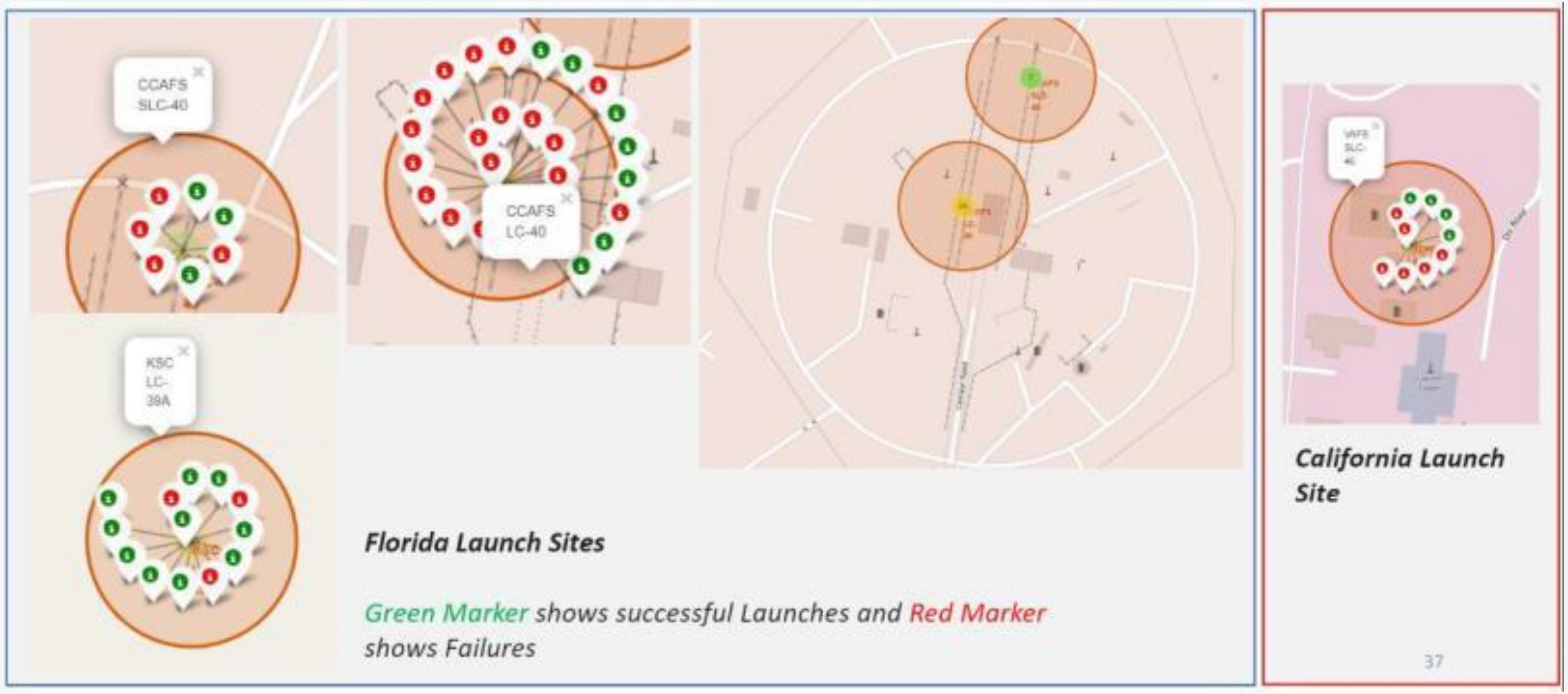
Launch Sites Proximities Analysis

Location of all the Launch Sites

We can see that all the SpaceX launch sites are located inside the United States



Markers showing launch sites with color labels



Florida Launch Sites

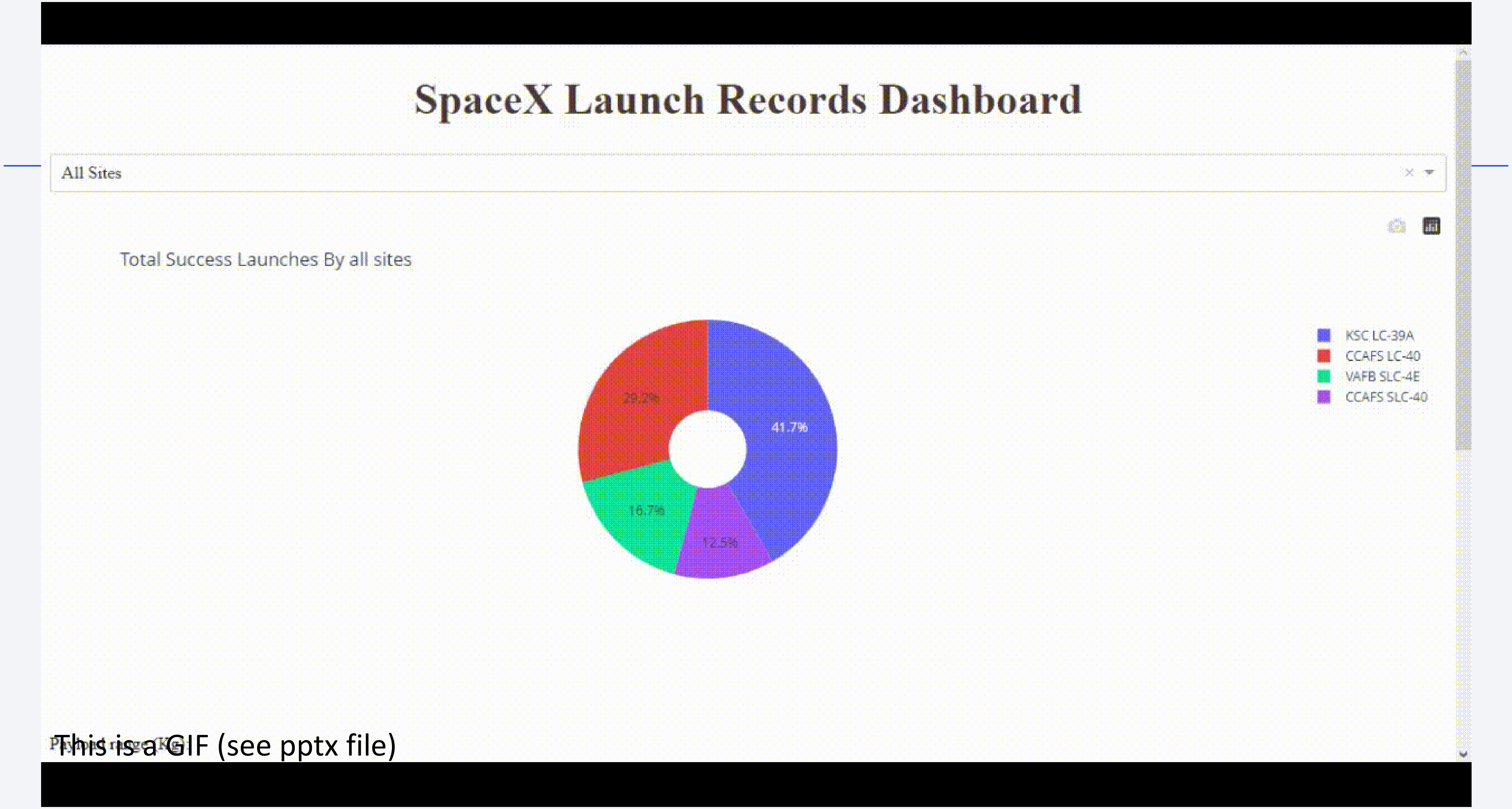
Green Marker shows successful Launches and Red Marker shows Failures

California Launch Site

37

Section 4

Build a Dashboard with Plotly Dash



SpaceX Launch Records Dashboard

All Sites

Total Success Launches By all sites

Payload range (Kg):

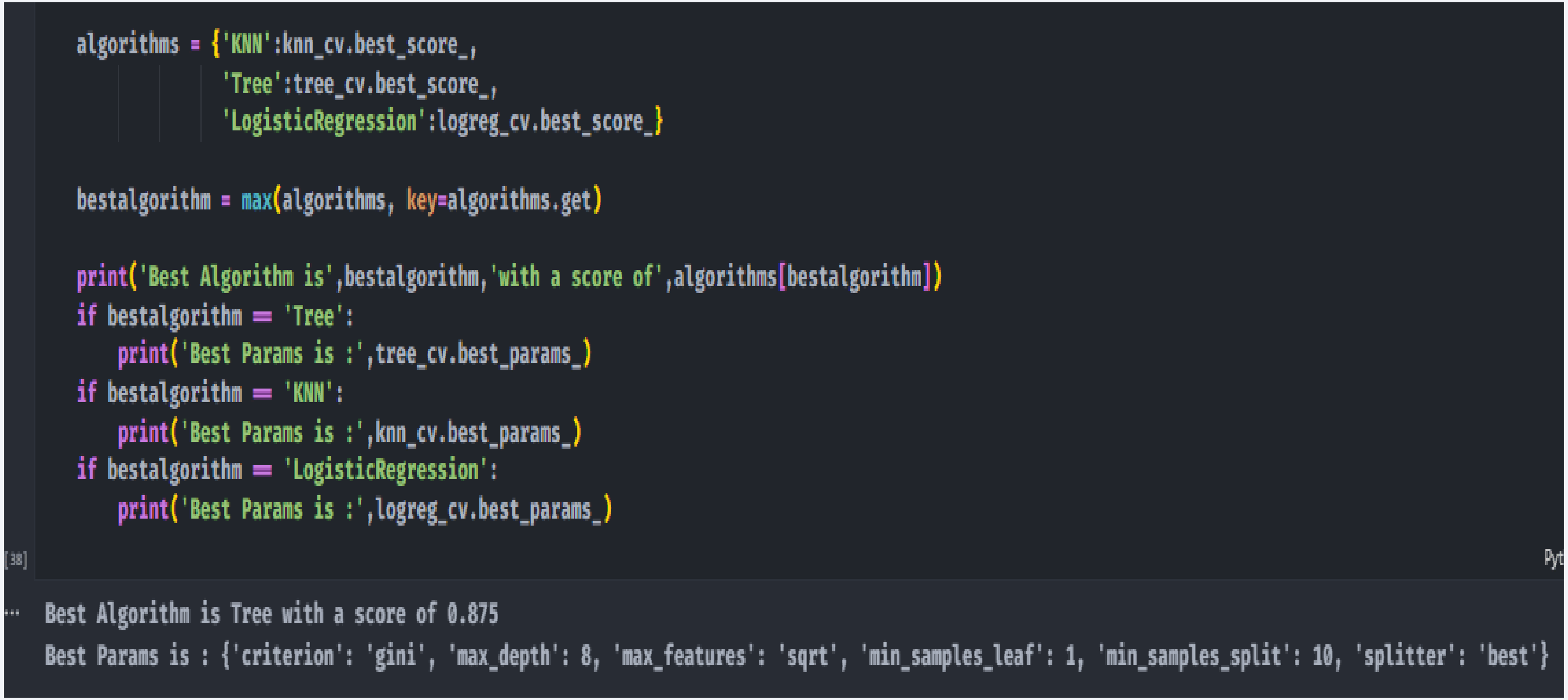
This is a GIF (see pptx file)

- We can see the success percentage by each sites.
- We can see Payload vs Launch Outcome Scatter Plot

Section 5

Predictive Analysis (Classification)

Classification Accuracy



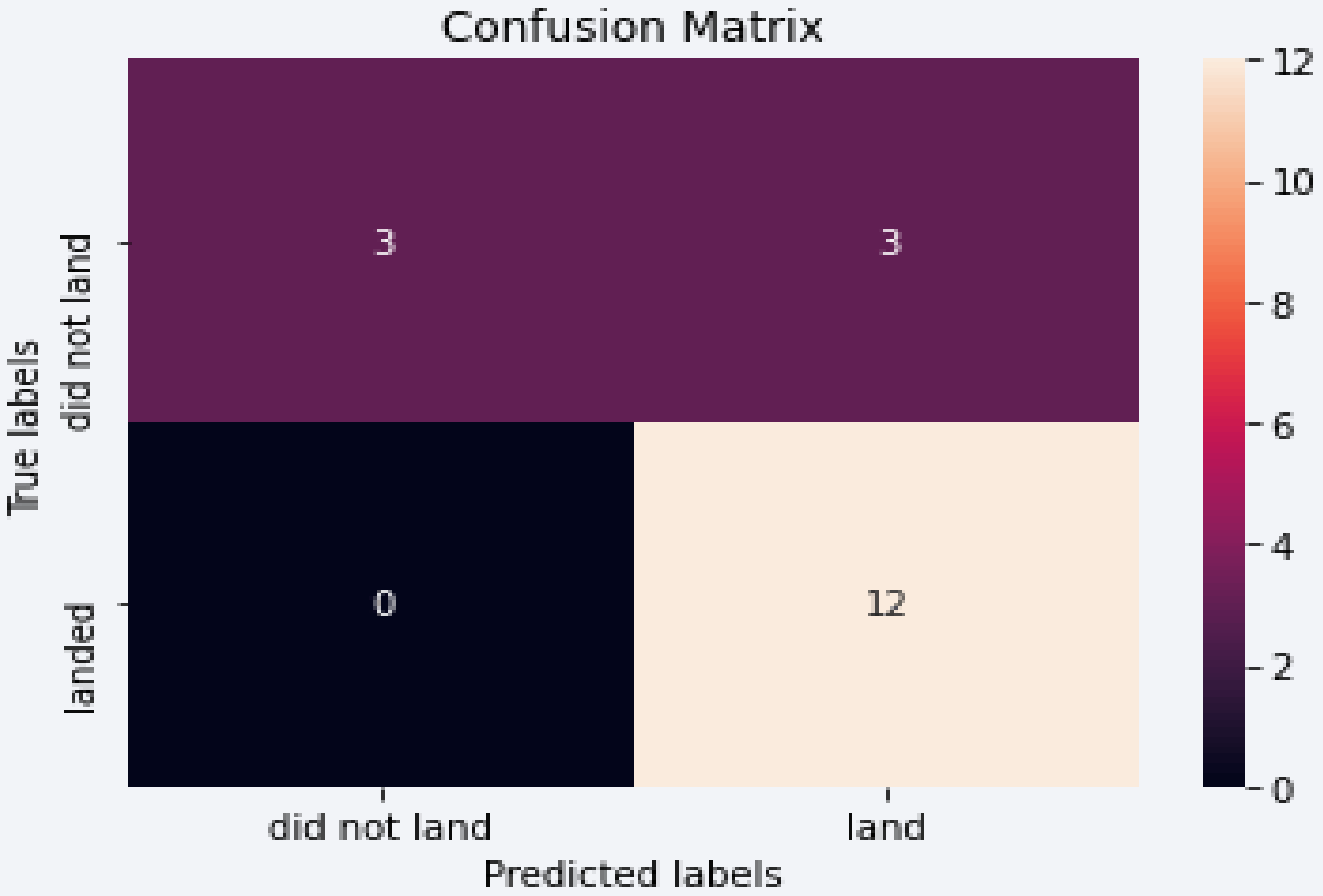
```
algorithms = {'KNN':knn_cv.best_score_,
              'Tree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_}

bestalgorithm = max(algorithms, key=algorithms.get)

print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.875
Best Params is : {'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'best'}
```

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

Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC 39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!