
REACH GREATER HEIGHTS
WITH DATA SCIENE
HRIDAY SHAH
30/10/2022

1. Outline

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

2. 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

- Project background and context

  On its website, SpaceX promotes Falcon9 rocket launches for 62 million dollars; other companies charge upwards of 165 million dollars per. A large portion of the savings is due to SpaceX's ability to reuse the first stage. So, if we can figure out whether the first stage will land, we can figure out how much it will cost to launch. If a different business want to compete with spaceX for a rocket launch, it may use the information provided here. The project's objective is to build a machine learning pipeline that can forecast if the first stage will be successful.

- Problems you want to find answers

  -What factors determine if the rocket will land successfully?

  -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.


METHODOLOGY
SECTIO-1

3.Methodology

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform datawrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The information was gathered in numerous ways.

  Data was gathered by sending a get request to the SpaceX API. Next, we used the.json() function call to decode the response's content as JSON and the.json normalize function call to convert it to a pandas data frame ().

  -The data was cleansed, missing values were verified and filled in as needed.

  -Additionally, using BeautifulSoup, we scraped Wikipedia for information on Falcon 9 launch date.

  It was intended to extract the launch records as an HTML table, parse the information, and then transform the table into a pandas dataframe for further analysis.

Data Collection –SpaceX API

•We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

Data Collection -Scraping

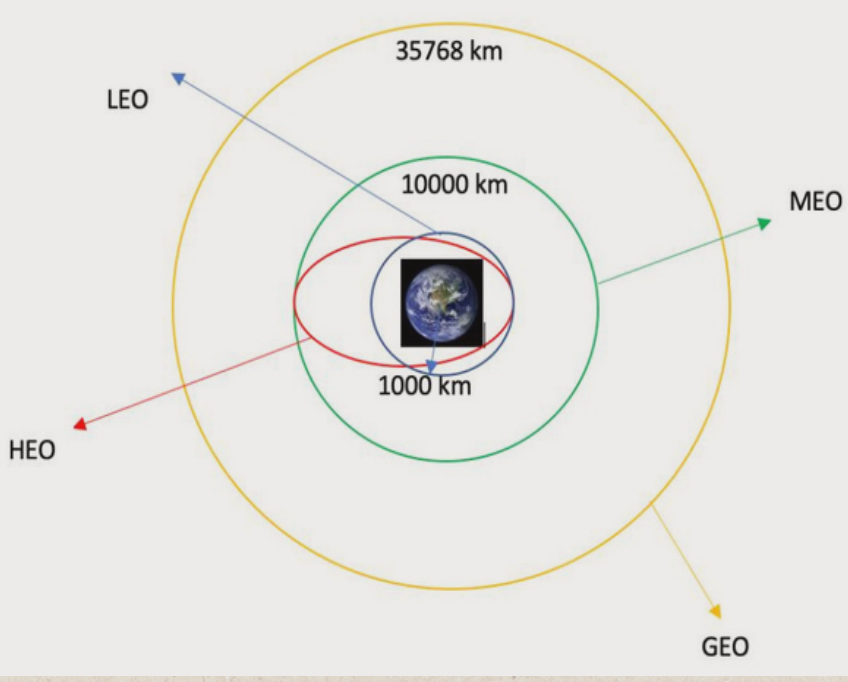
•We applied web scrapping to web scrap Falcon 9 launch records with BeautifulSoup
•We parsed the table and converted it into a pandas dataframe.

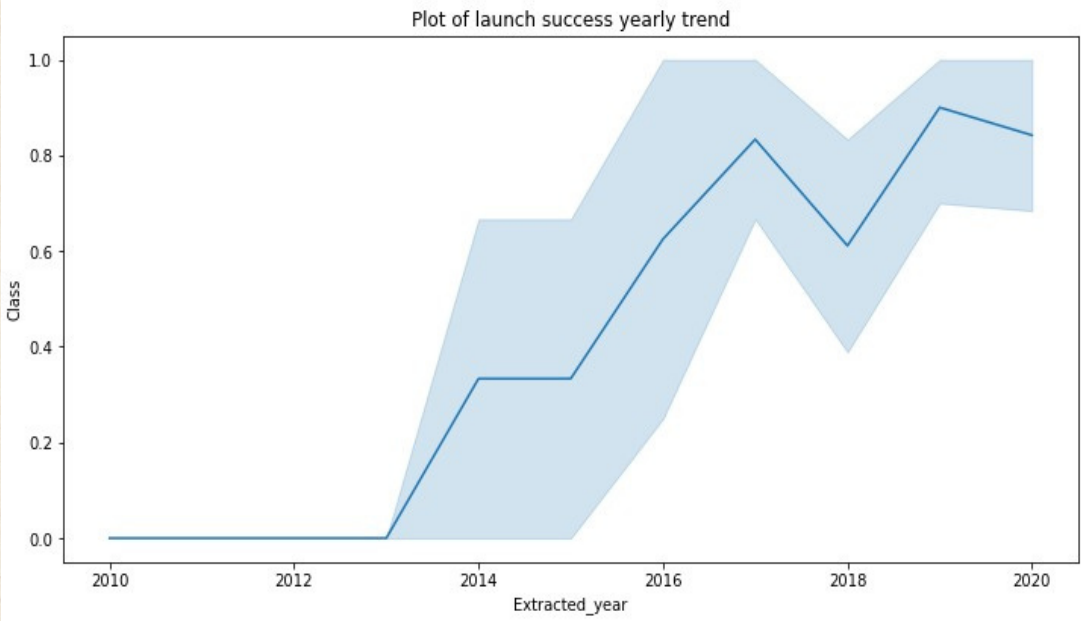
Data Wrangling

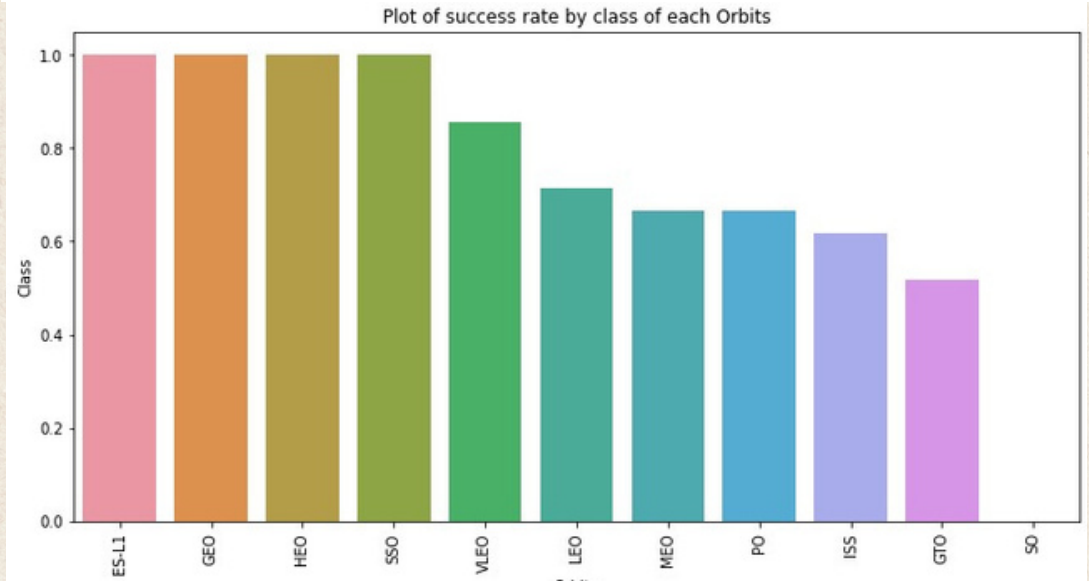
•Exploratory data analysis was done to establish the training labels.
We determined the number of launches at each location as well as the frequency and number of orbits.
•We used the outcome column to build the landing outcome label and saved the data to CSV.

EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.







EDA withcSQL

Without leaving the Jupyter Notebook, we imported the SpaceX dataset into the PostgreSQL database.
To gain understanding from the data, we used EDA along with SQL. We created queries to learn things like:

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

Build an Interactive Map with Folium

- On the folium map, we identified every launch point and added map elements like markers, circles, and lines to indicate whether a launch was successful or unsuccessful for each location. We categorise feature launch results (success or failure) into classes 0 and 1. 0 represents failure while 1 represents success.
- The launch sites with a comparatively high success rate were determined using the colour-labelled marker clusters.
- We measured the separations between a launch site and its environmentss.

We responded to various queries, such as:

-Are launch sites near railways, highways and coastlines.
-Do launch sites keep certain distance away from cities.

Build a Dashboard with Plotly Dash

Using Plotly dash, we created an interactive dashboard.
We created pie graphs that display all of the launches made by particular sites.
For each booster version, we created a scatter graph to highlight the link between the outcome and the payload mass (Kg).

Predictive Analysis (Classification)

•We used numpy and pandas to import the data, converted it, and divided it into training and testing sets.
•Using GridSearchCV, we constructed various machine learning models and adjusted various hyperparameters.
•We measured the performance of our model using accuracy, and we enhanced it using feature engineering and algorithm tweaking.
We identified the categorization model that performed the best.

4. Results

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


INSIGHTS
SECTION-2

Total Number of Success and Failures

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:

The total number of successful mission outcome is:

	successoutcome
0	100

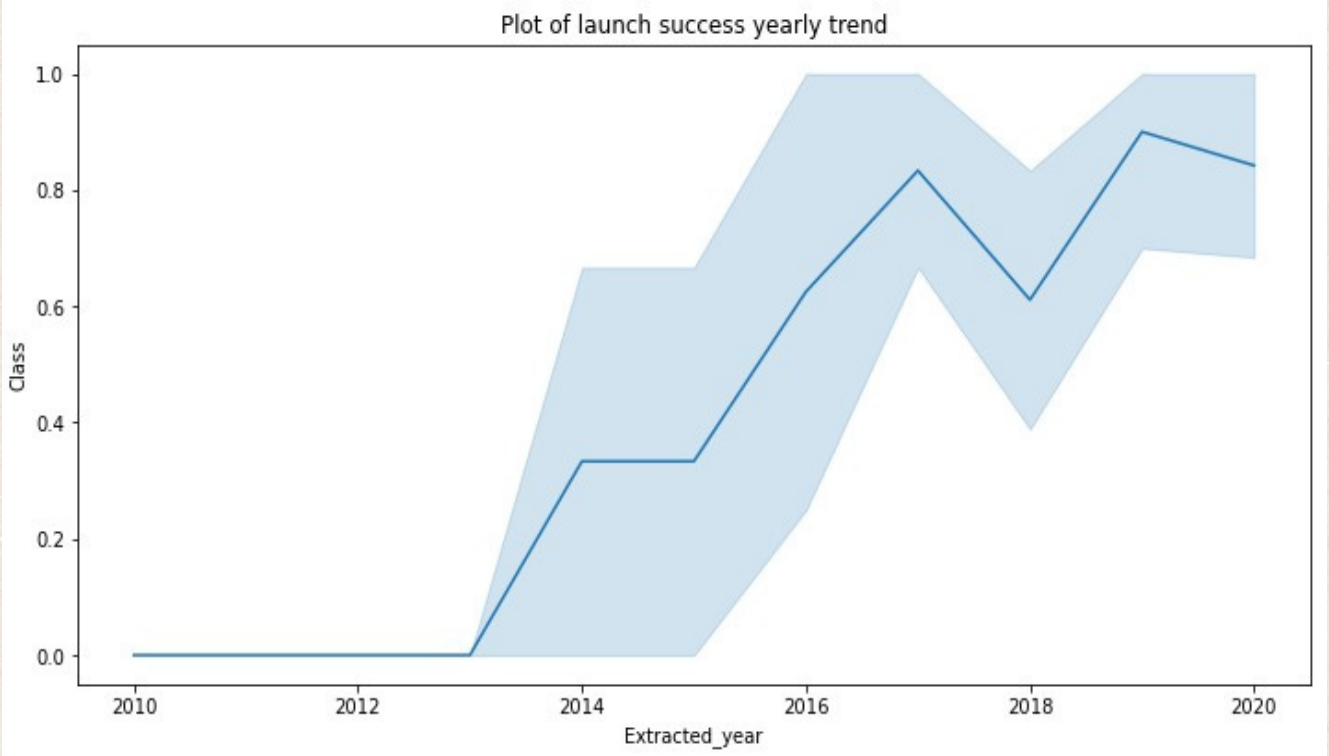
The total number of failed mission outcome is:

	failureoutcome
0	1

100:1 Success to Failure Ratio obtsained

Launch Success Yearly Trend

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



Average and Total payload mass by F9 v1.1

The average mass of the payload that booster version F9 v1.1 can carry was determined to be 2928.4.

Using the following query, we determined that NASA's boosters carried a total of 45596 kg of cargo.

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

In [13]:
```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

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

In [12]:
```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Boosters Carried Maximum Payload

Using a subquery in the WHERE clause and the MAX() method, we were able to identify the booster that had carried the most payload.

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

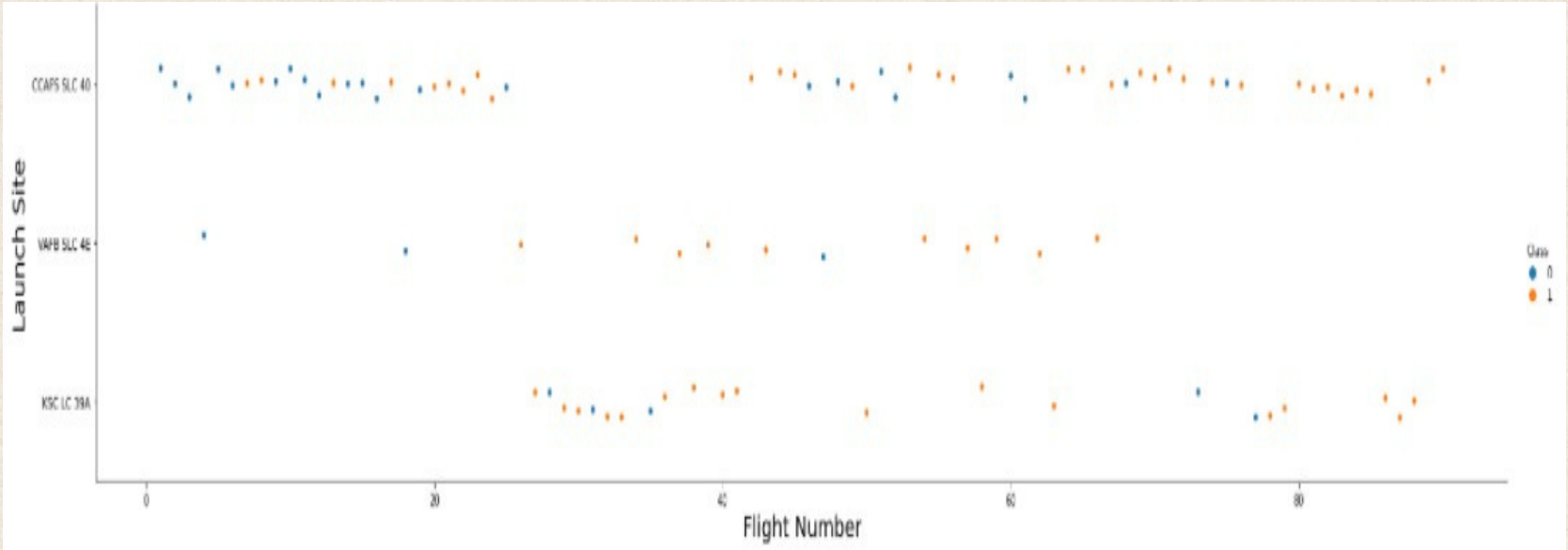
In [17]:
```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

Payload vs. Launch Site

The greater the payload mass for launch location CCAFS SLC 40, the better the success rate for the rocket



Successful Drone Ship Landing with Payload

In [15]:
```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

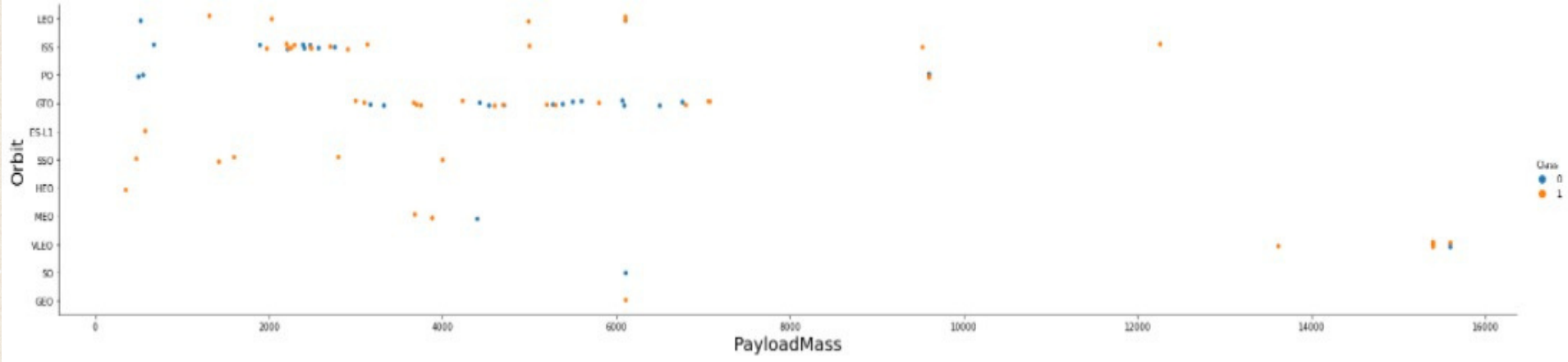
Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

- To find rockets that successfully landed on drone ships, we utilised the WHERE clause, and we used the AND condition to identify successful landings with payload masses larger than 4,000 but less than 6,000.

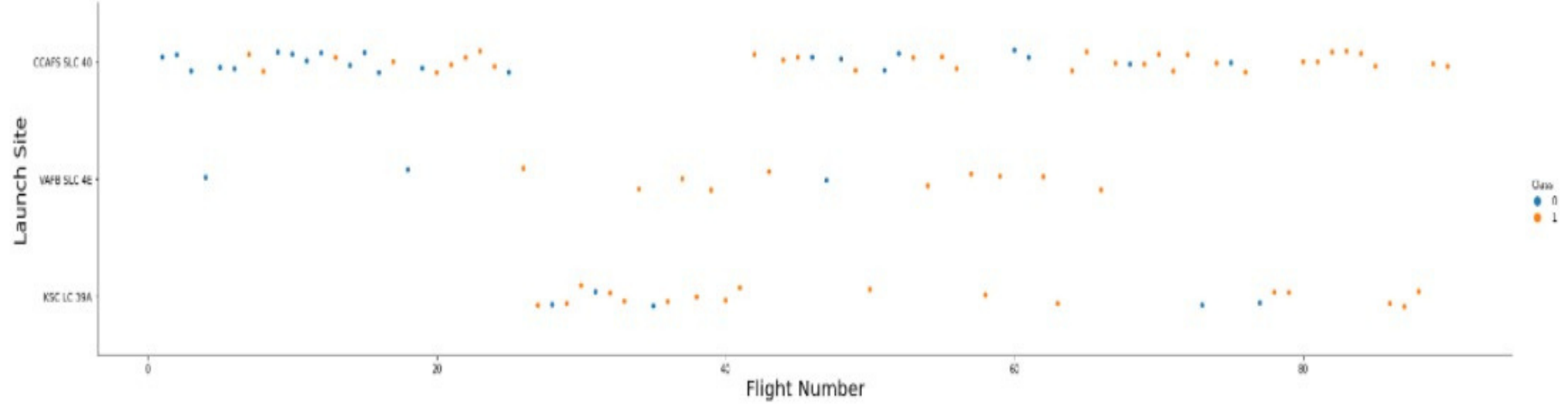
Payload vs. OrbitType

•With heavy payloads, we can see that successful landings are more often in PO, LEO, and ISS orbits.



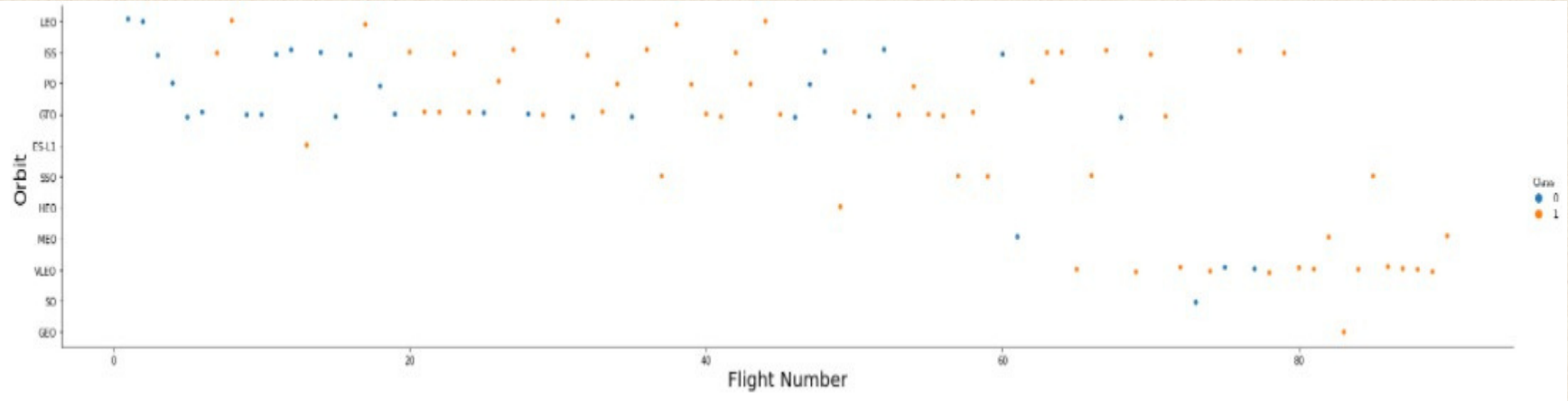
Flight Number vs. LaunchSite

The plot led us to the conclusion that a launch site's success rate increases with the size of the flight quantity.



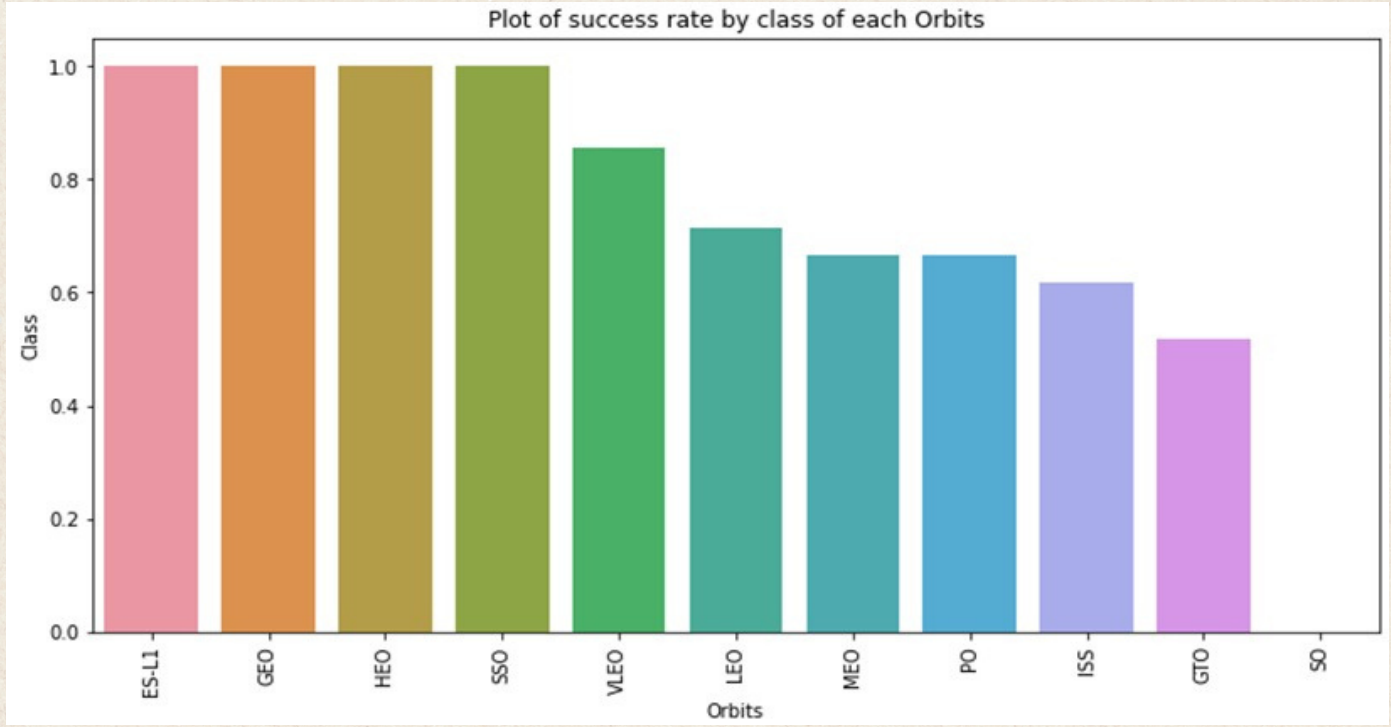
Flight Number vs. Orbit Type

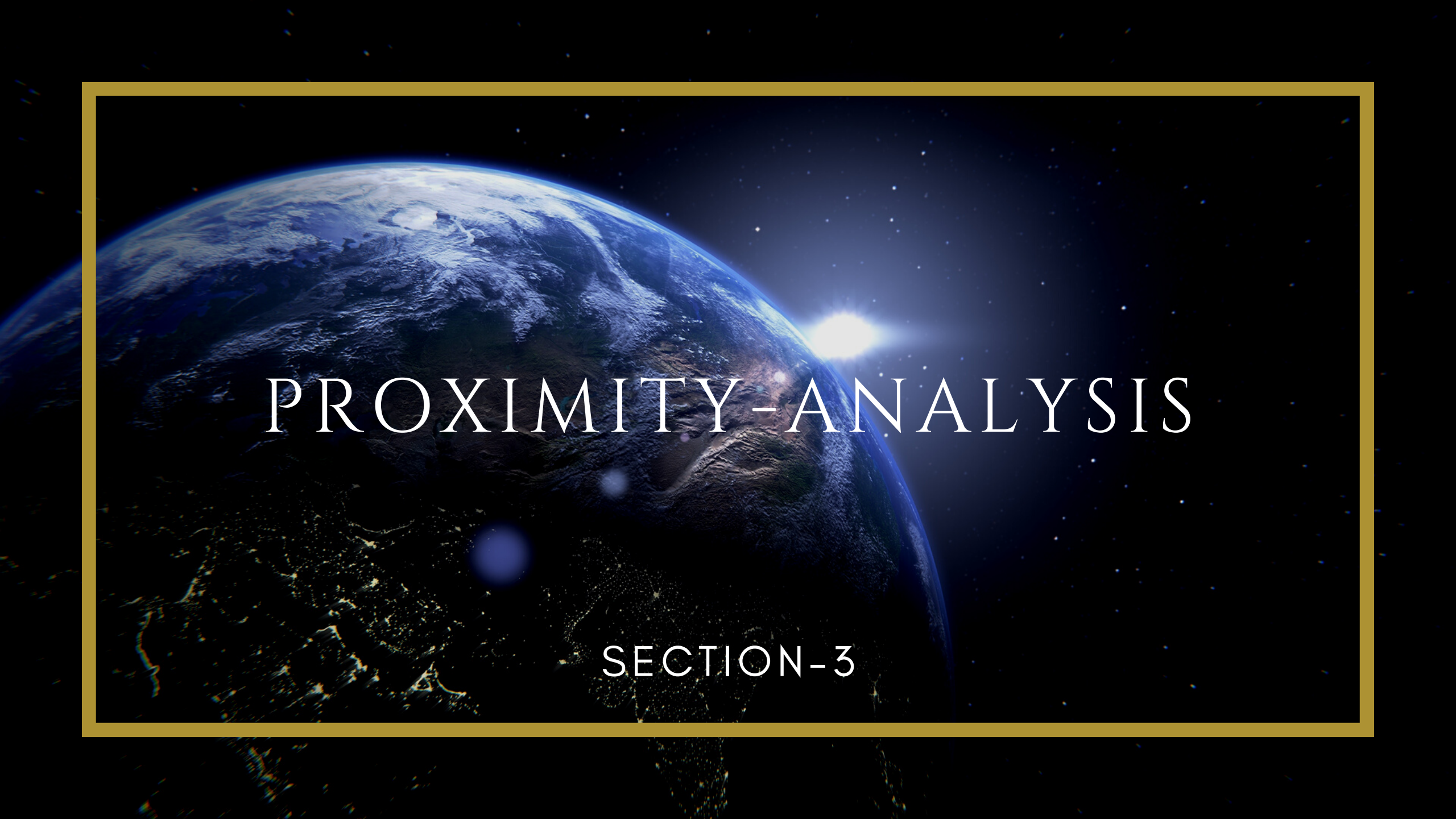
The graph below displays the Flight Number versus Orbit type.
We note that success in the LEO orbit is correlated with the number of flights, however there is no correlation between the number of flights and the GTO orbit.



Success Rate vs. Orbit Type

- According to the figure, ES-L1, GEO, HEO, SSO, and VLEO had the highest success rates.



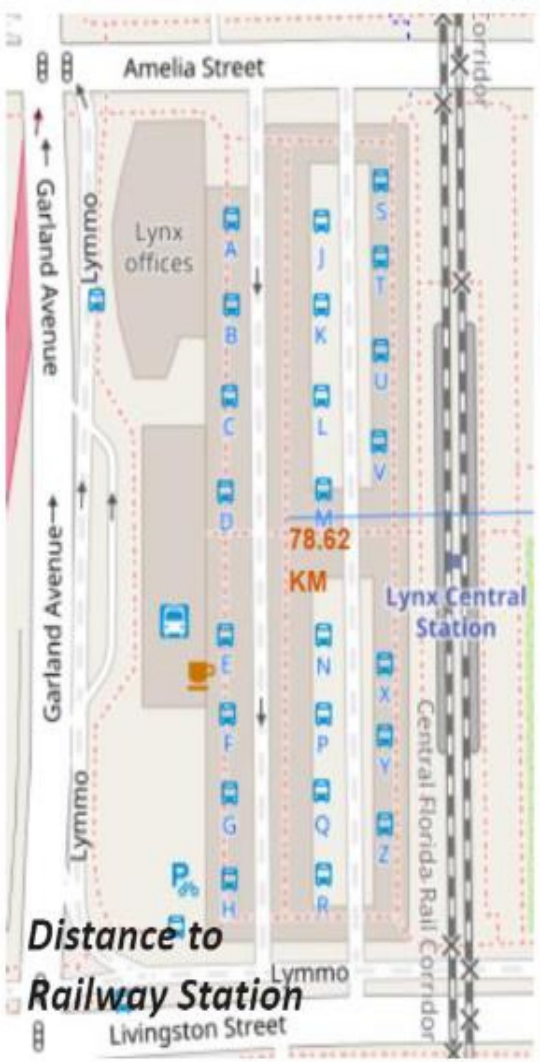

PROXIMITY-ANALYSIS
SECTION-3

All launch sites global map markers

We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

Markers showing launch sites with color labels

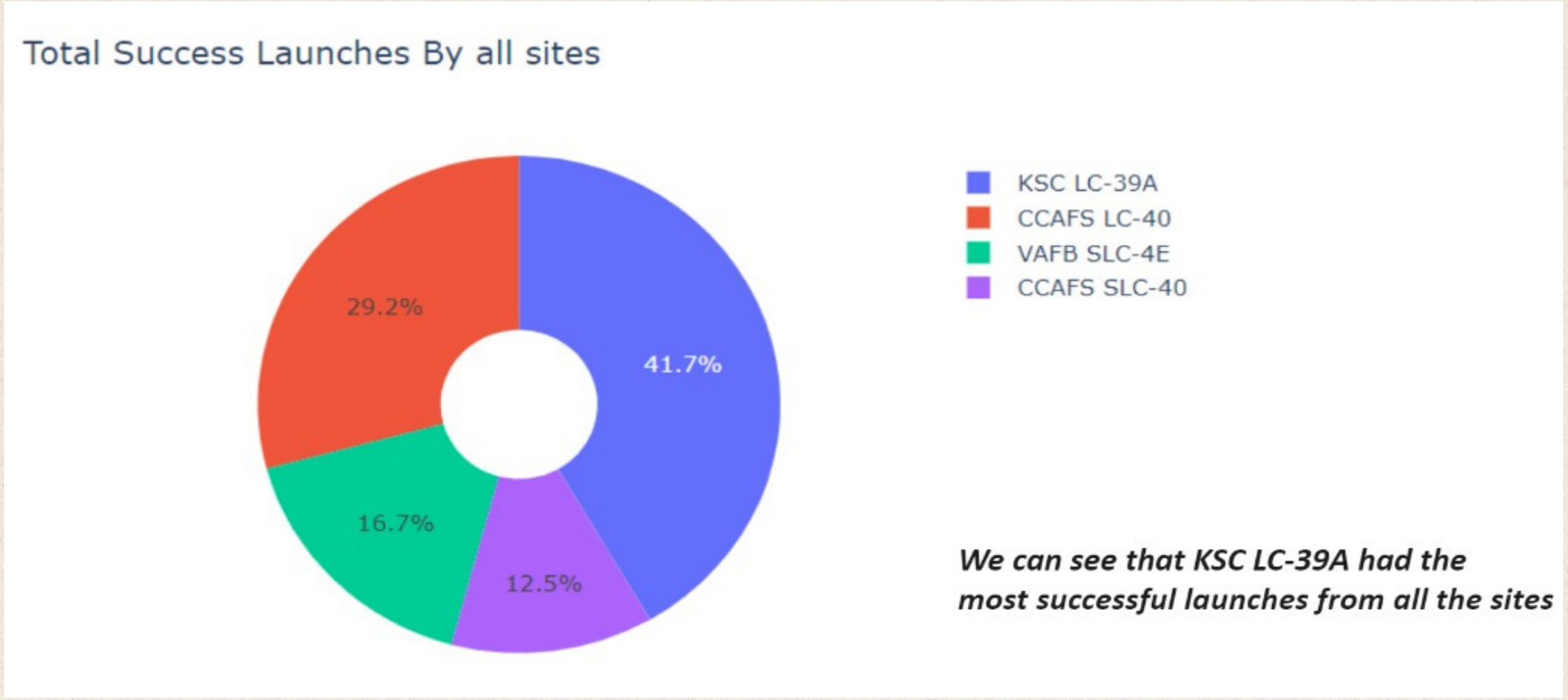
Launch Site distance to landmarks



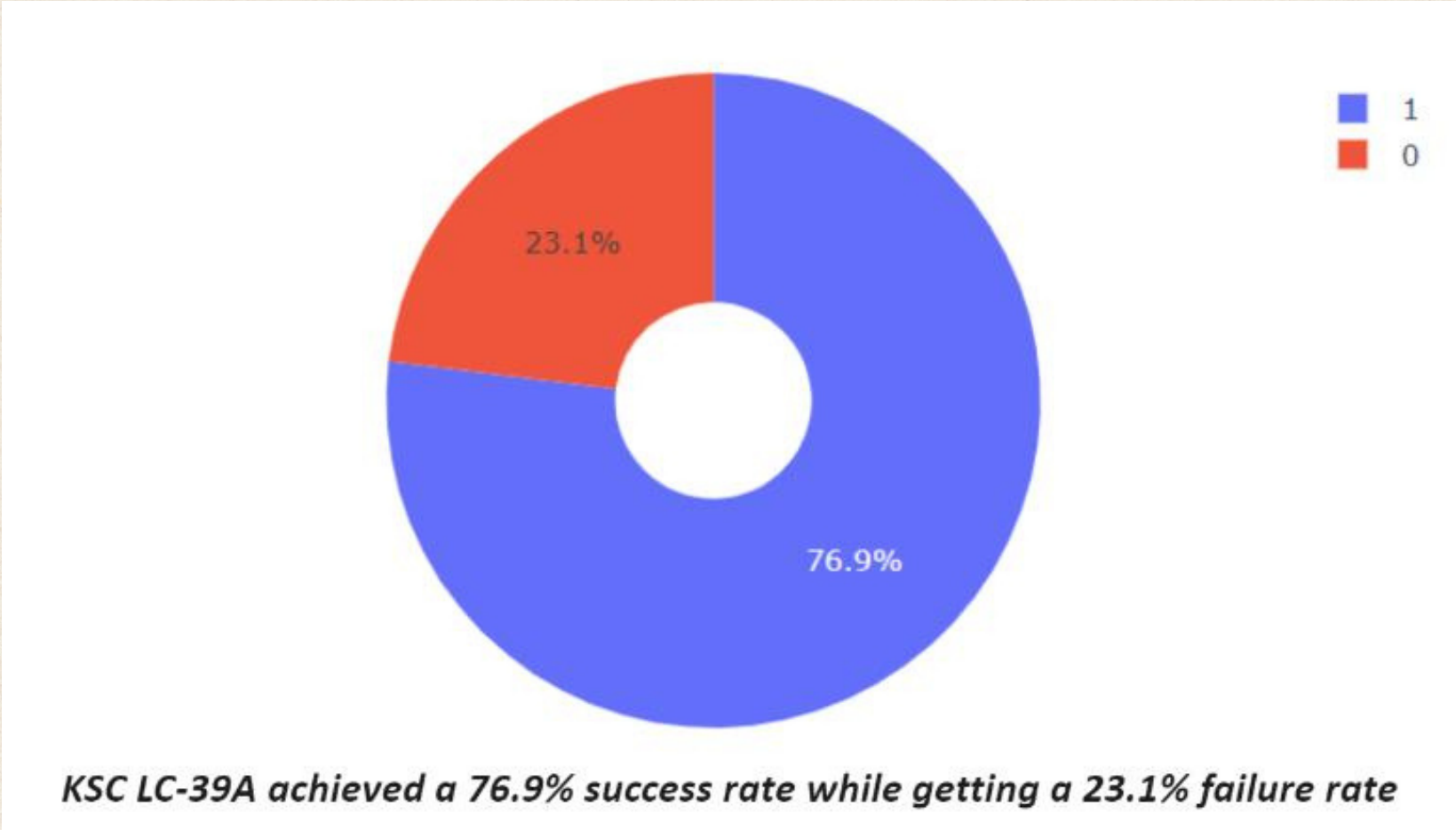
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


DASHBOARD
SECTION-4

Pie chart showing the success percentage achieved by each launch site

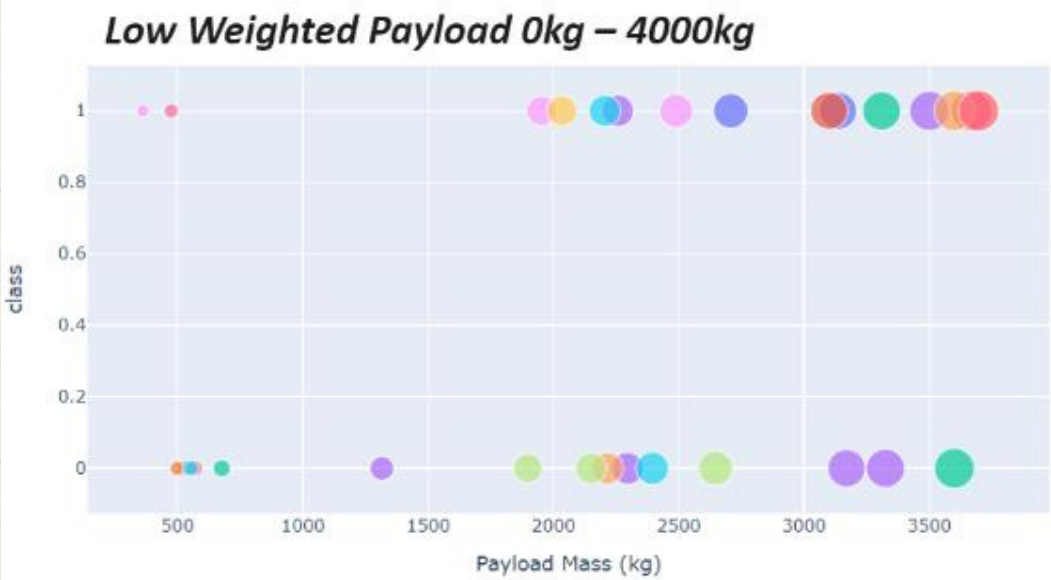


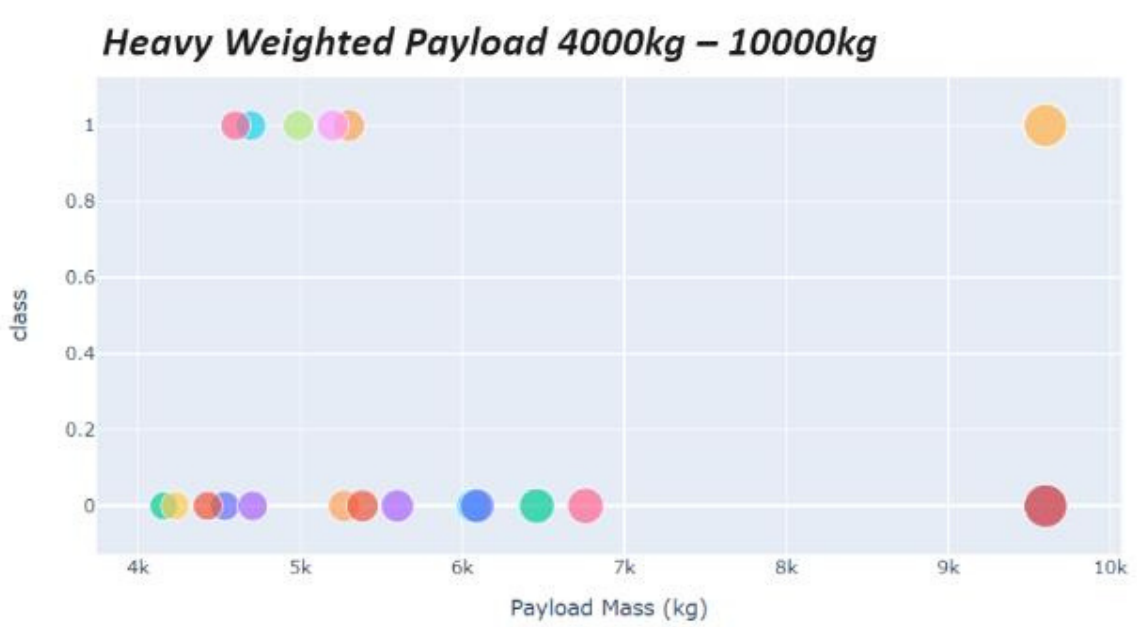
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads


PREDICTIONS
SECTION-5

Classification Accuracy

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

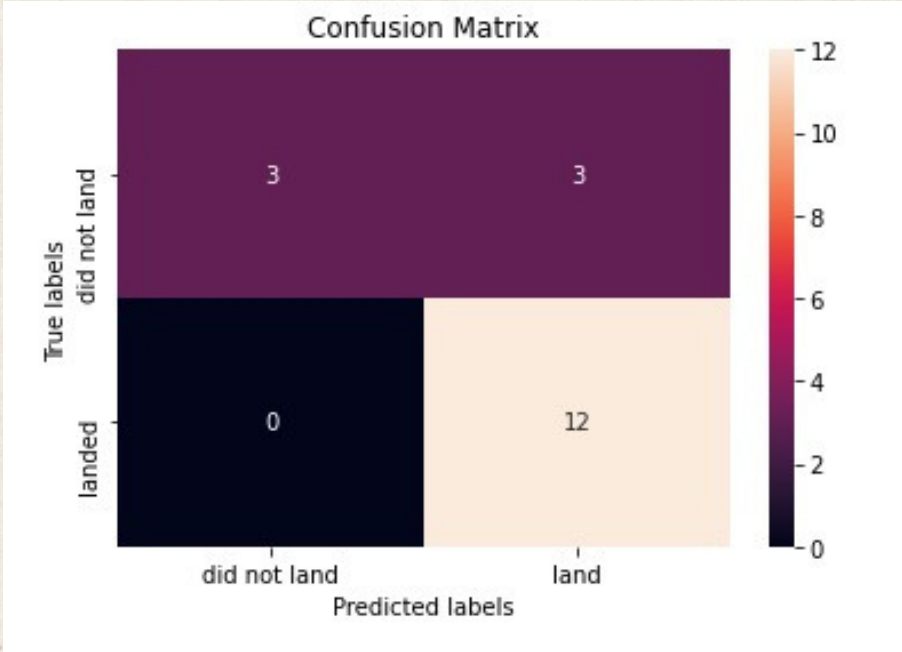
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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'}
```

Confusion Matrix

•The decision tree classifier's confusion matrix demonstrates that it is capable of differentiating between the various classes.
False positives are the main issue.
In other words, the classifier misclassified a failure landing as a successful one.



Conclusions

We can draw the following conclusion:

•T he success percentage at a launch location increases with the size of the flight quantity.

• The success rate of launches increased from 2013 to 2020.

• The highest success rate was in the ES-L1, GEO, HEO, SSO, and VLEO orbits.

• Of all the locations, KSC LC-39A had the most prosperous launches.

The most effective machine learning approach for this problem is the decision tree classifier.


THANK-YOU
HRIDAY SHAH