IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

By Peter Papuli
12/14/2023

Outline

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

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

  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 successfully.

- 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.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to 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
  - How to build, tune, evaluate classification models

Data Collection

- The data was collected using various methods
  - Data collection was done using get request to the SpaceX API, they were kind enough to provide this data for free
  - We can create pandas dataframes from the collected JSON data
  - Once its in a pandas df, we can start to preprocess the data
    - We can perform feature engineering
    - Then, one hot encoding categorical variables is very important

Data Collection – SpaceX API

- A GET request was utilized to retrieve data from the SpaceX API, followed by meticulous data cleansing, basic wrangling, and formatting to ensure quality and usability.

- [ppapuli102/IBM-Data-Science (github.com)]

1. Get request for rocket launch data using API

In [6]:
```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

In [7]:
```
response = requests.get(spacex_url)
```

2. Use json_normalize method to convert json result to dataframe

In [12]:
```
# Use json_normalize method to convert the json result into a dataframe

# decode response content as json
static_json_df = res.json()
```

In [13]:
```
# apply json_normalize
data = pd.json_normalize(static_json_df)
```

3. We then performed data cleaning and filling in the missing values

In [30]:
```
rows = data_falcon9['PayloadMass'].values.tolist()[0]

df_rows = pd.DataFrame(rows)
df_rows = df_rows.replace(np.nan, PayloadMass)

data_falcon9['PayloadMass'][0] = df_rows.values
data_falcon9
```

Data Collection - Scraping

- We implemented web scraping techniques to extract Falcon 9 launch records using BeautifulSoup, subsequently parsing the data and transforming it into a structured pandas dataframe for analysis.
- [ppapuli102/IBM-Data-Science (github.com)]

1. Apply HTTP Get method to request the Falcon 9 rocket launch page

```
In [4]: static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

```
In [5]: # use requests.get() method with the provided static_url
        # assign the response to a object
        html_data = requests.get(static_url)
        html_data.status_code

Out[5]: 200
```

2. Create a BeautifulSoup object from the HTML response

```
In [6]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
        soup = BeautifulSoup(html_data.text, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
In [7]: # Use soup.title attribute
        soup.title

Out[7]: <title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

3. Extract all column names from the HTML table header

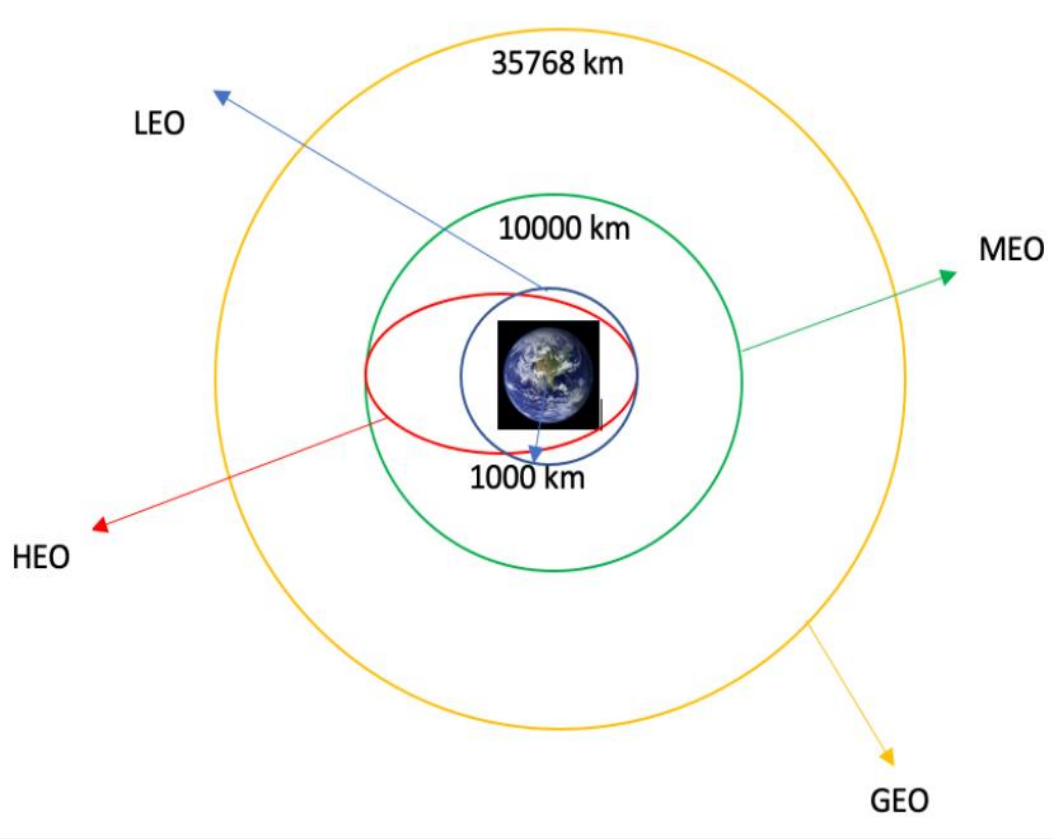
```
In [10]: column_names = []

         # Apply find_all() function with `th` element on first_launch_table
         # Iterate each th element and apply the provided extract_column_from_header() to get a column name
         # Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called column_names

         element = soup.find_all('th')
         for row in range(len(element)):
             try:
                 name = extract_column_from_header(element[row])
                 if (name is not None and len(name) > 0):
                     column_names.append(name)
             except:
                 pass
```

4. Create a dataframe by parsing the launch HTML tables
5. Export data to csv

Data Wrangling

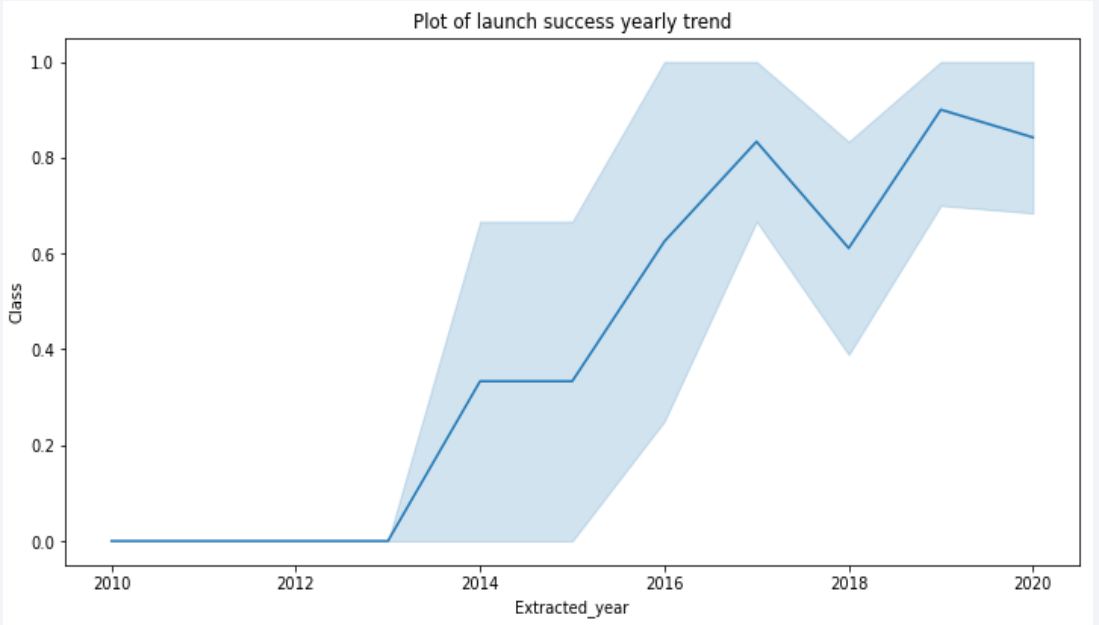


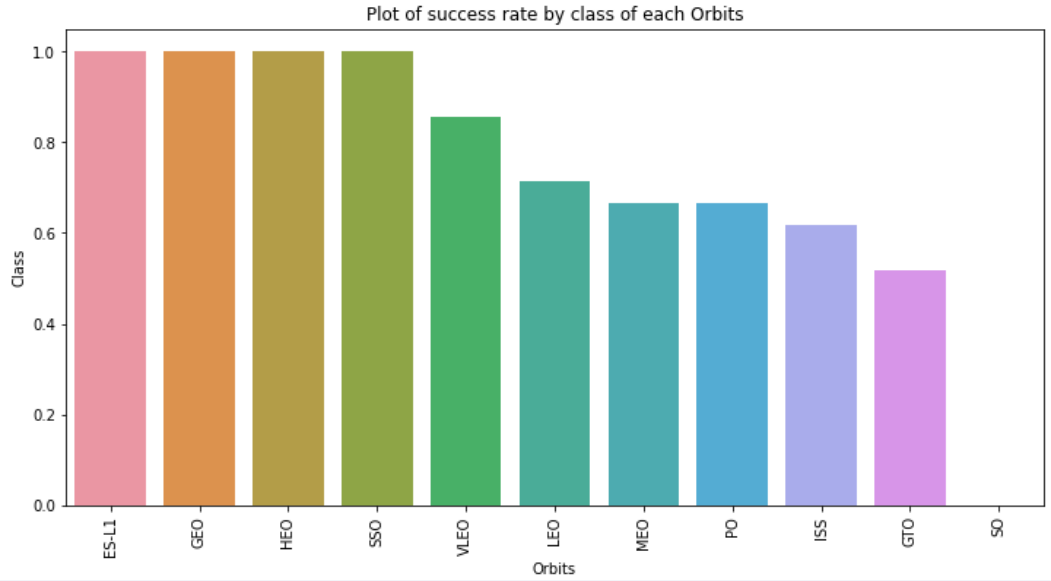
- An extensive exploratory data analysis was conducted to identify the training labels. This involved quantifying the number of launches per site and analyzing the frequency of various orbits. Additionally, we derived landing outcome labels from the outcome column and meticulously compiled the findings into a CSV file for further utilization

- [ppapuli102/IBM-Data-Science (github.com)]

EDA with Data Visualization

- We delved into the data by creating visual representations to understand the correlations between flight number and launch site, payload mass and launch site, as well as the success rate across different orbit types. Additionally, we examined the relationship between flight number and orbit type and observed the annual trend in launch success rates





- The link to the notebook is [ppapuli102/IBM-Data-Science (github.com)]

EDA with SQL

- Seamlessly integrated the SpaceX dataset into a PostgreSQL database directly from a Jupyter notebook. Employed SQL-based exploratory data analysis to derive insights, crafting queries to ascertain vital statistics such as unique launch site identifiers, aggregate payload masses for NASA (CRS) missions, mean payload mass for Falcon 9 v1.1 boosters, and comprehensive success and failure metrics of missions, including detailed breakdowns of failed drone ship landings by booster version and launch site

- ppapuli102/IBM-Data-Science (github.com)

Build an Interactive Map with Folium

- Enhanced a folium map with precise markers for each launch site, incorporating graphical elements to indicate launch outcomes. Classified launch successes and failures as binary features for analytical clarity. Utilized color-coded marker clusters to visually distinguish launch sites with higher success rates.
- Performed geospatial analysis to compute distances from launch sites to key infrastructures, enabling insights into the strategic placement of launch sites in relation to railways, highways, coastlines, and urban areas

Build a Dashboard with Plotly Dash

- "Developed an interactive dashboard utilizing Plotly Dash to showcase the dynamics of SpaceX launches. Integrated pie charts to depict launch distribution across various sites and implemented scatter plots to explore the correlation between payload mass and launch outcomes across different booster versions. The dashboard's interactivity provides a comprehensive and engaging user experience
- ppapuli102/IBM-Data-Science (github.com)

Predictive Analysis (Classification)

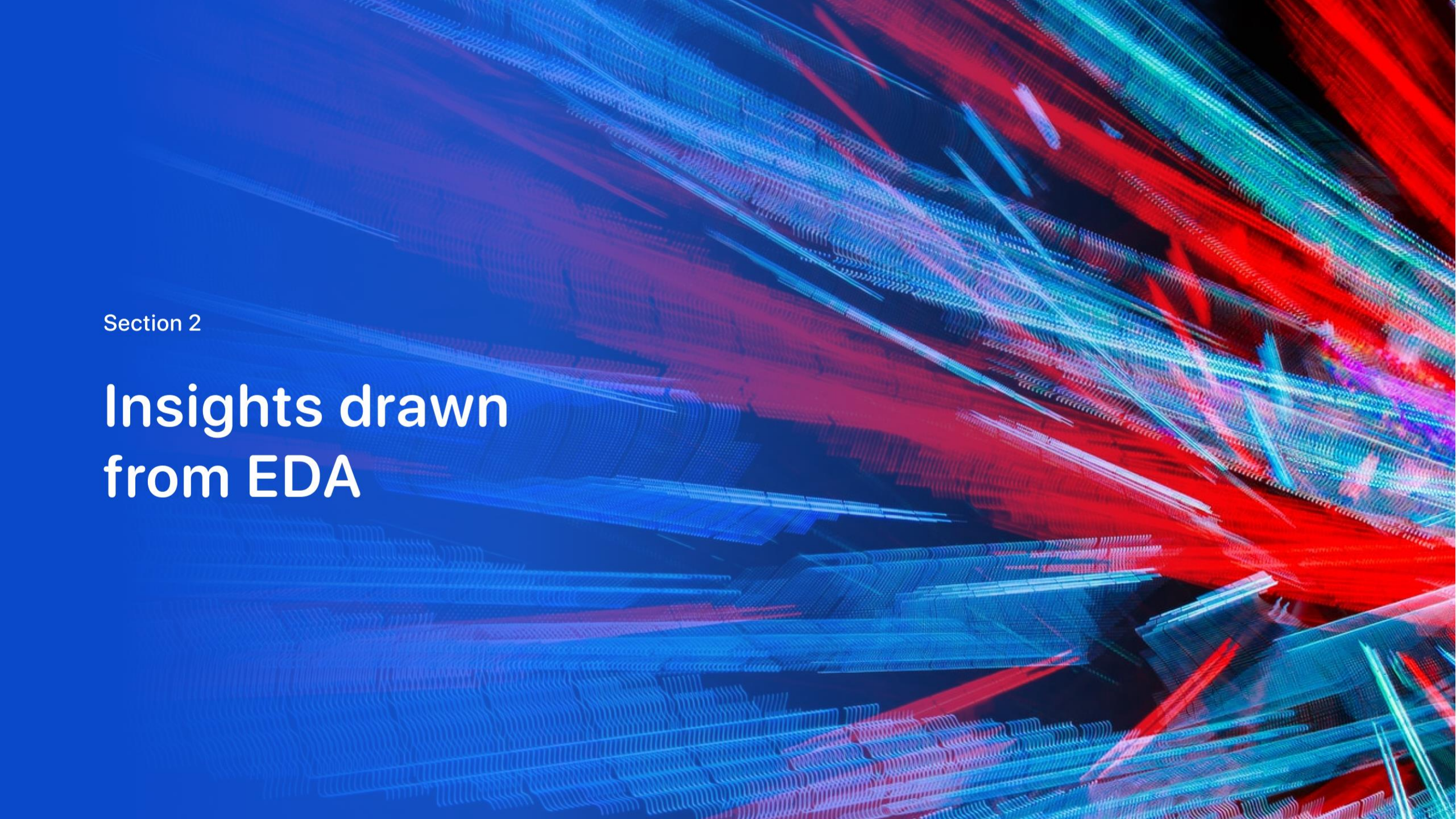
- Leveraged numpy and pandas for data ingestion and transformation, segmenting the dataset into training and testing subsets. Executed the construction of diverse machine learning models, meticulously optimizing hyperparameters via GridSearchCV. Adopted accuracy as the principal metric to refine our models through feature engineering and algorithm tuning, culminating in the identification of the most efficacious classification model
- ppapuli102/IBM-Data-Science (github.com)

Results

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

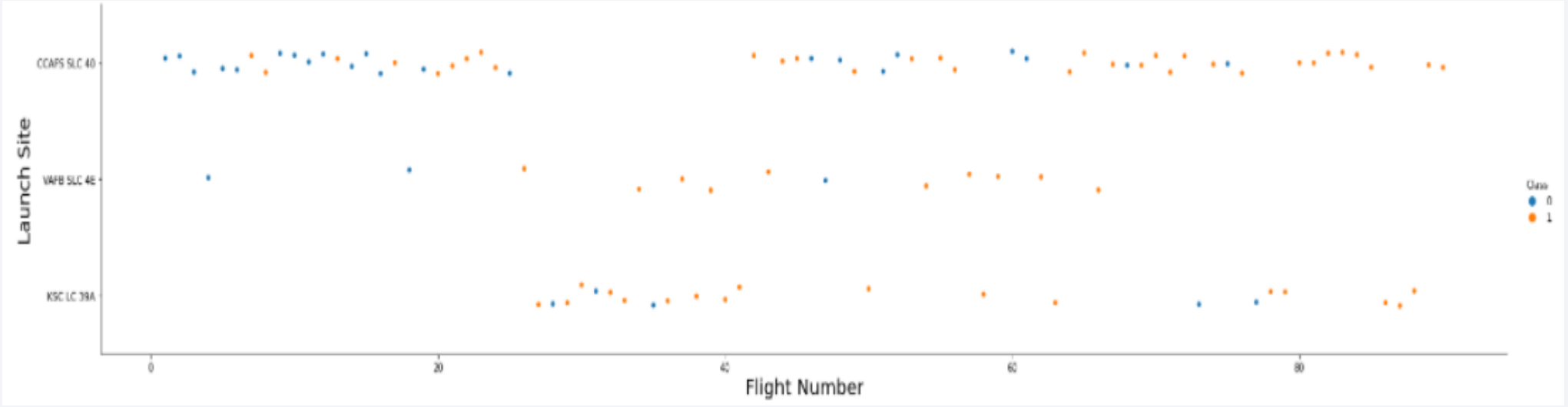
Section 2

Insights drawn from EDA



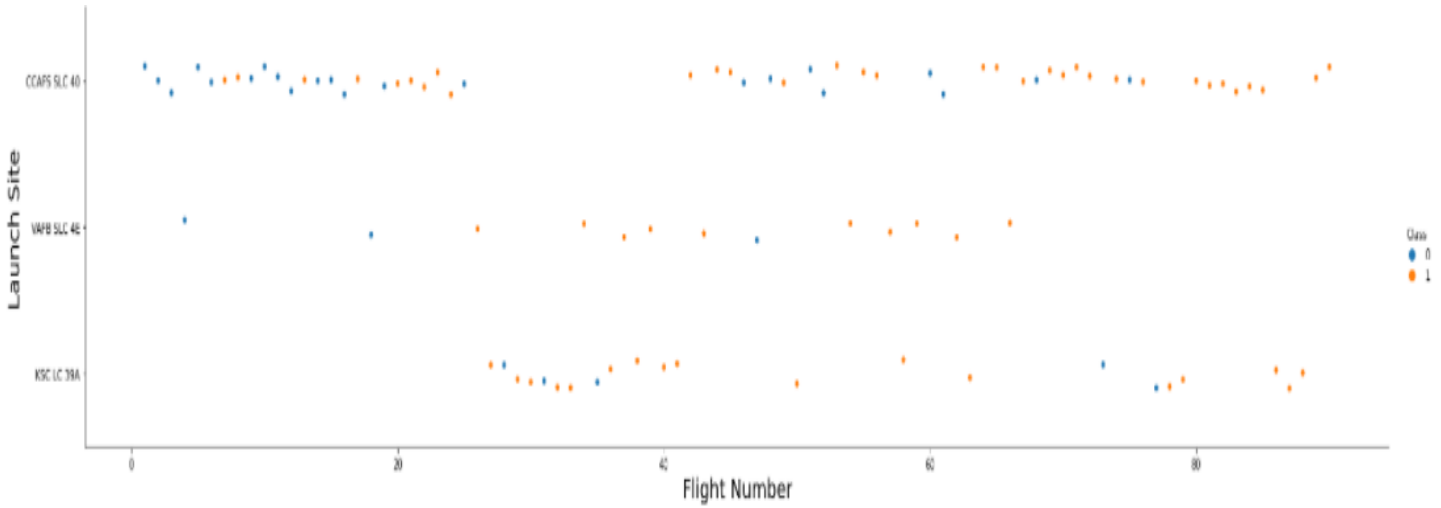
Flight Number vs. Launch Site

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



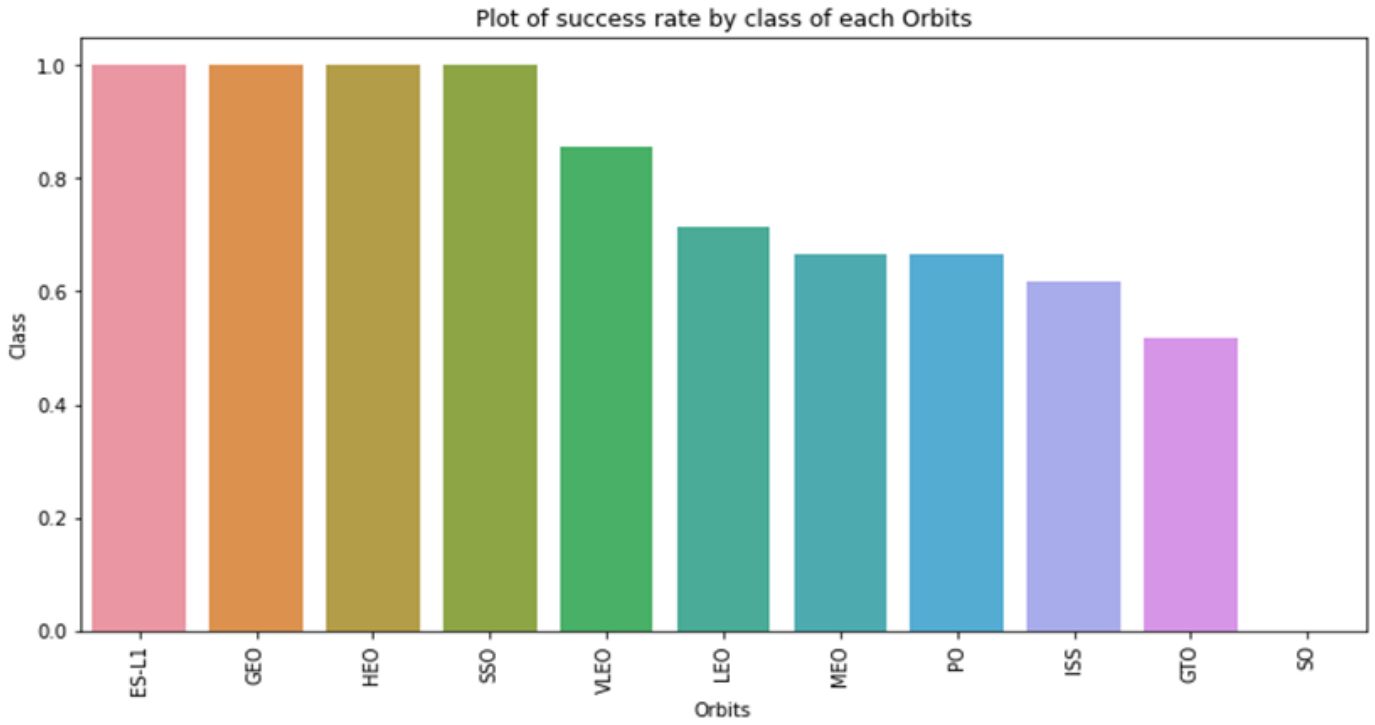
Payload vs. Launch Site

The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



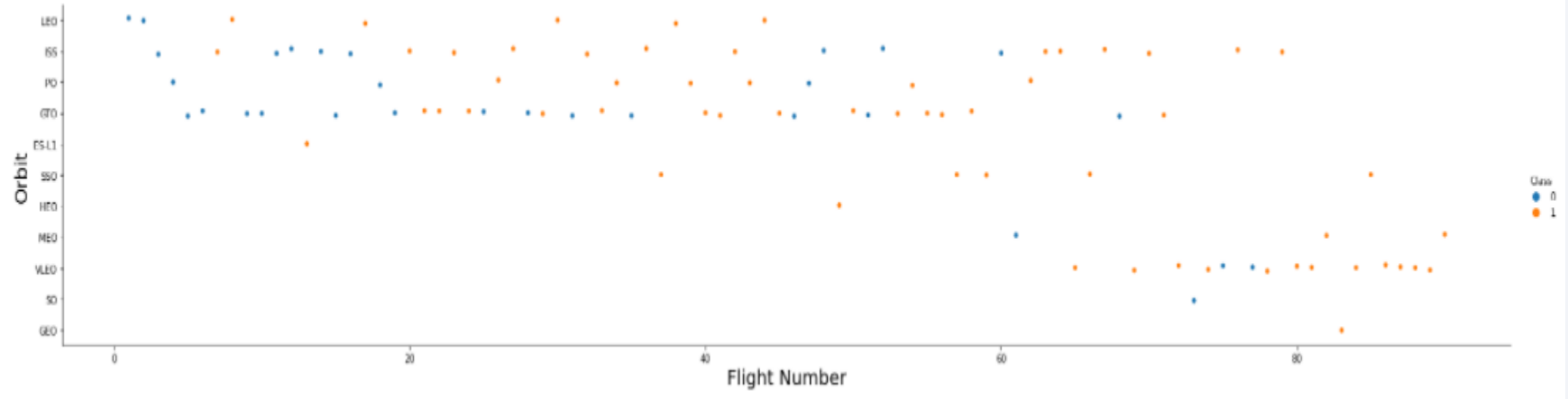
Success Rate vs. Orbit Type

- The analysis of the plot indicates that orbits ES-L1, GEO, HEO, SSO, and VLEO register the highest success rates, suggesting a pattern of reliability for missions targeting these specific orbital paths



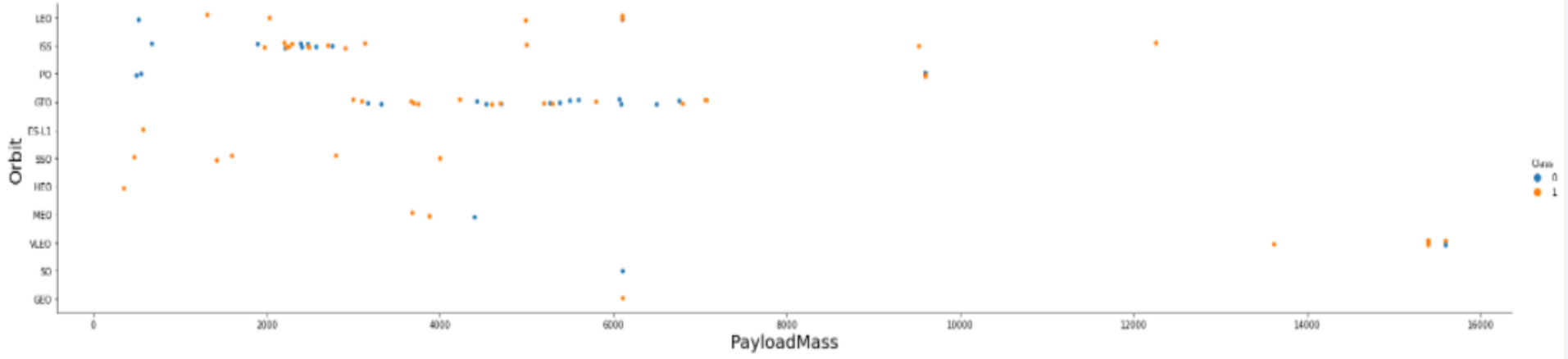
Flight Number vs. Orbit Type

- The visualization below presents the relationship between Flight Number and Orbit type, revealing a trend where successful missions within the Low Earth Orbit (LEO) appear correlated with increased flight frequency. Conversely, for Geostationary Transfer Orbit (GTO) missions, the data indicates no discernible correlation between the number of flights and success in that orbit category



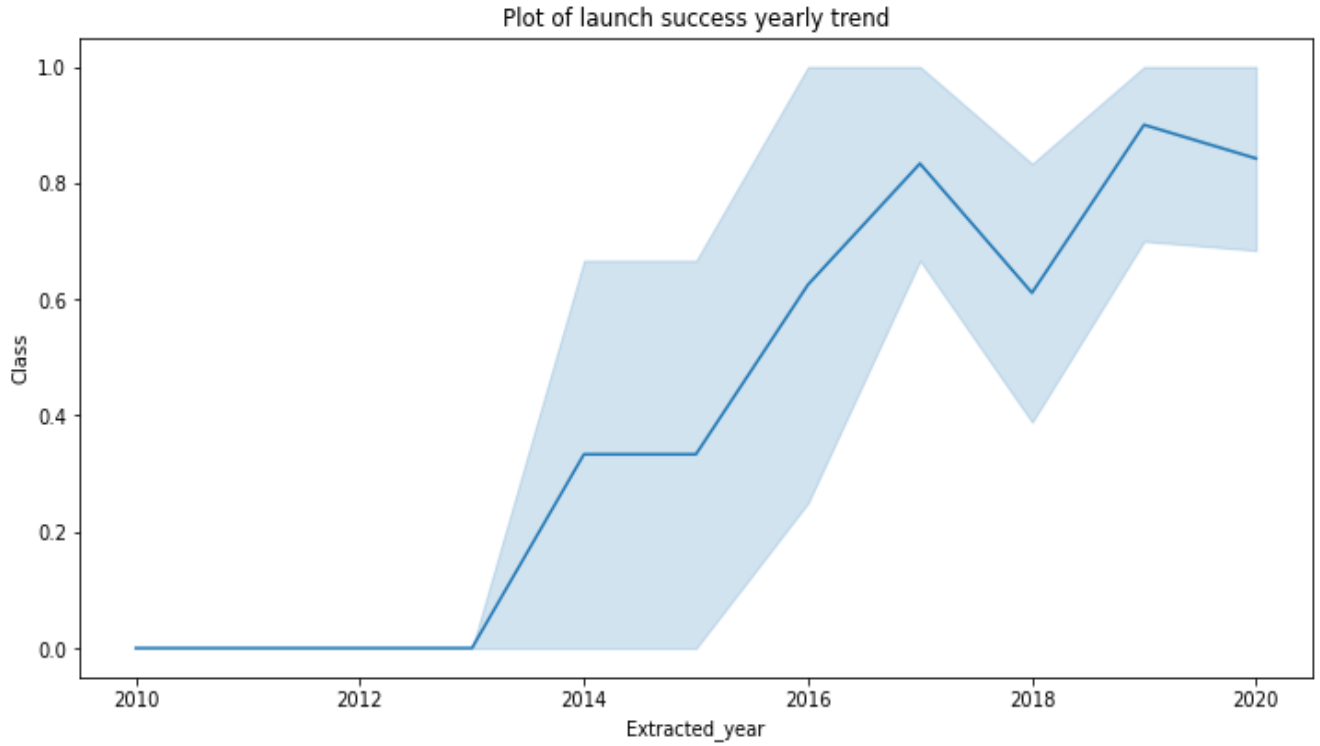
Payload vs. Orbit Type

- Upon examining the data, it is evident that heavier payloads correlate with a higher rate of successful landings in Polar, Low Earth, and International Space Station orbits



Launch Success Yearly Trend

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



All Launch Site Names

- We used the key word **DISTINCT** to show only unique launch sites

Display the names of the unique launch sites in the space mission

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'CCA'

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

In [11]:
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Using the query above we display 5 launch sites beginning with `CCA`

Total Payload Mass

- We computed the cumulative payload transported by NASA's boosters, amounting to a total of 4559

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

Average Payload Mass by F9 v1.1

- We determined that the average payload mass carried by booster version F9 v1.1 is 2928.4 units

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

First Successful Ground Landing Date

- We observed that the initial occurrence of a successful landing outcome on a ground pad took place on December 22, 2015

In [14]:
```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- In our SQL query, the WHERE clause was utilized to isolate instances of boosters that have achieved successful landings on a drone ship. Additionally, we employed the AND condition to refine this selection further, targeting only those successful landings where the payload mass was greater than 4000 kilograms and less than 6000 kilograms

Total Number of Successful and Failure Mission Outcomes

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)
```

The total number of successful mission outcome is:

	successoutcome
0	100

The total number of failed mission outcome is:

Out[16]:

	failureoutcome
0	1

- We employed a wildcard symbol, such as '%,' as a filtering mechanism to discern records in the WHERE clause where 'MissionOutcome' achieved either success or failure

Boosters Carried Maximum Payload

- We ascertained the booster responsible for carrying the utmost payload through the utilization of a subquery within the WHERE clause in conjunction with the MAX() function

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

2015 Launch Records

- We employed a combination of WHERE clauses, including LIKE, AND, and BETWEEN conditions, to selectively filter and retrieve information pertaining to failed landing outcomes on drone ships, their respective booster versions, and launch site names during the year 2015

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

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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))

In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

- We chose Landing outcomes and the COUNT of landing outcomes from the dataset, employing the WHERE clause to filter for landing outcomes falling within the timeframe of June 4, 2010, to March 20, 2010. Subsequently, we utilized the GROUP BY clause to categorize the landing outcomes and the ORDER BY clause to arrange the grouped landing outcomes in descending order

Section 4

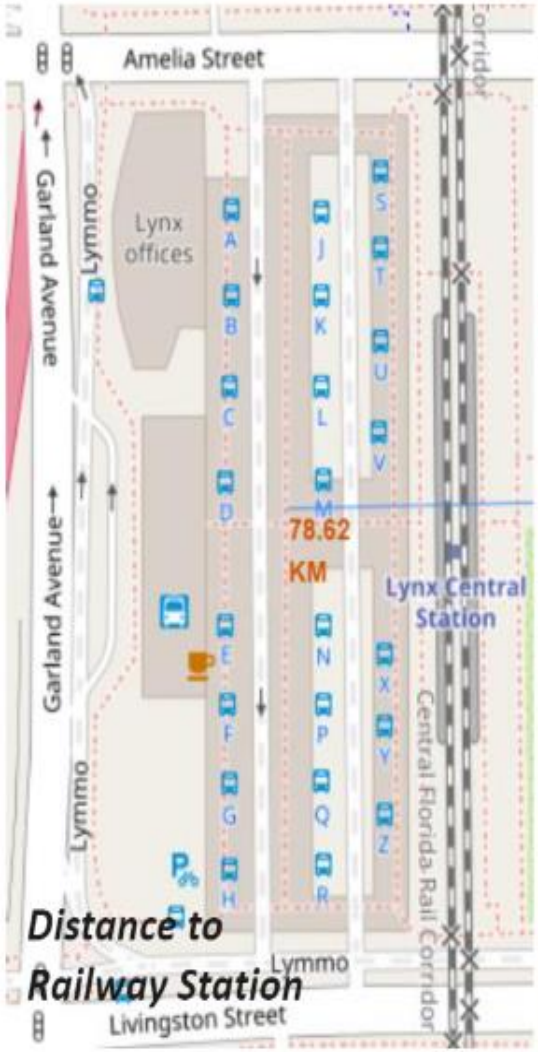
Launch Sites Proximities Analysis

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

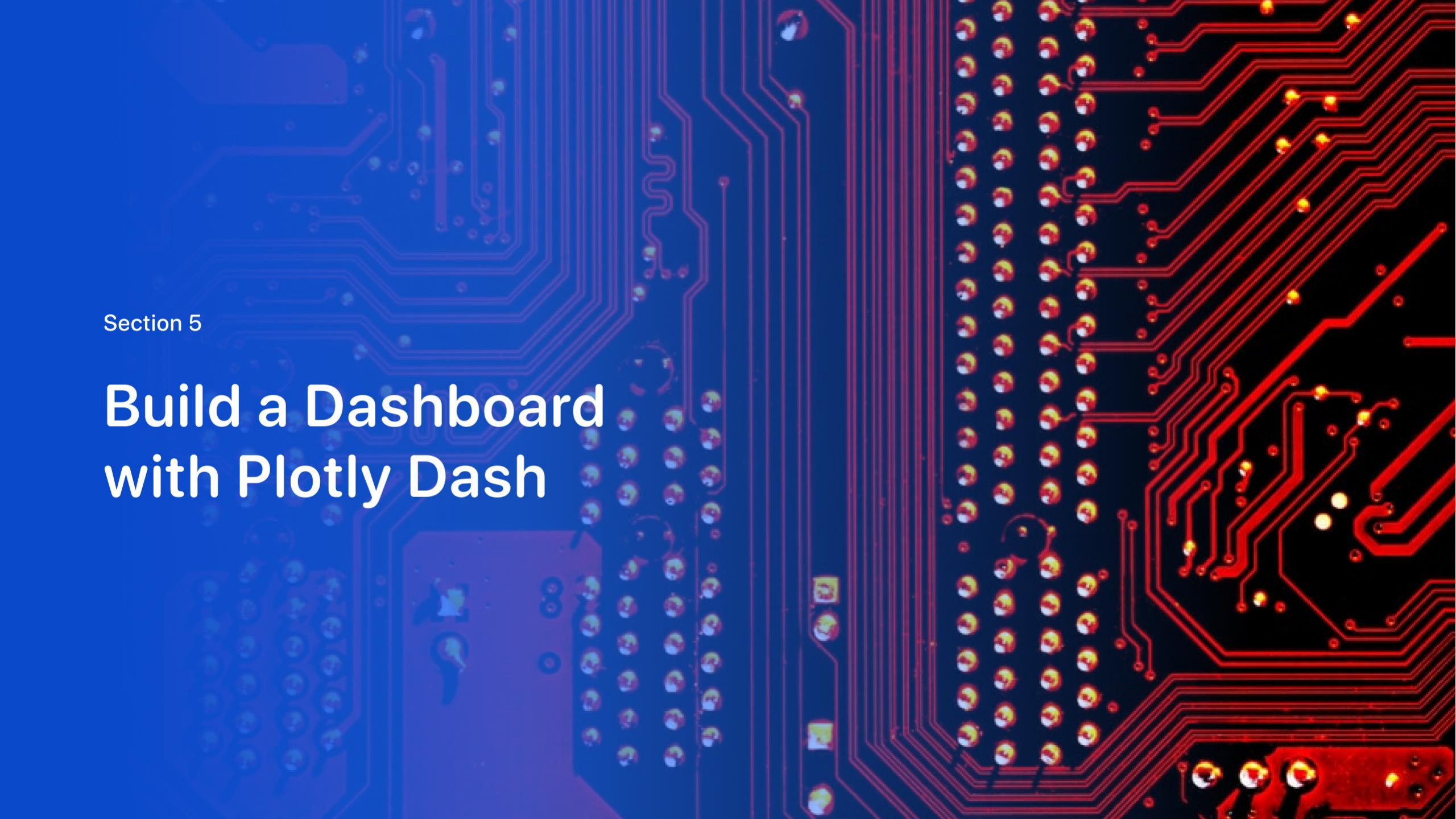
Launch Site distance to landmarks



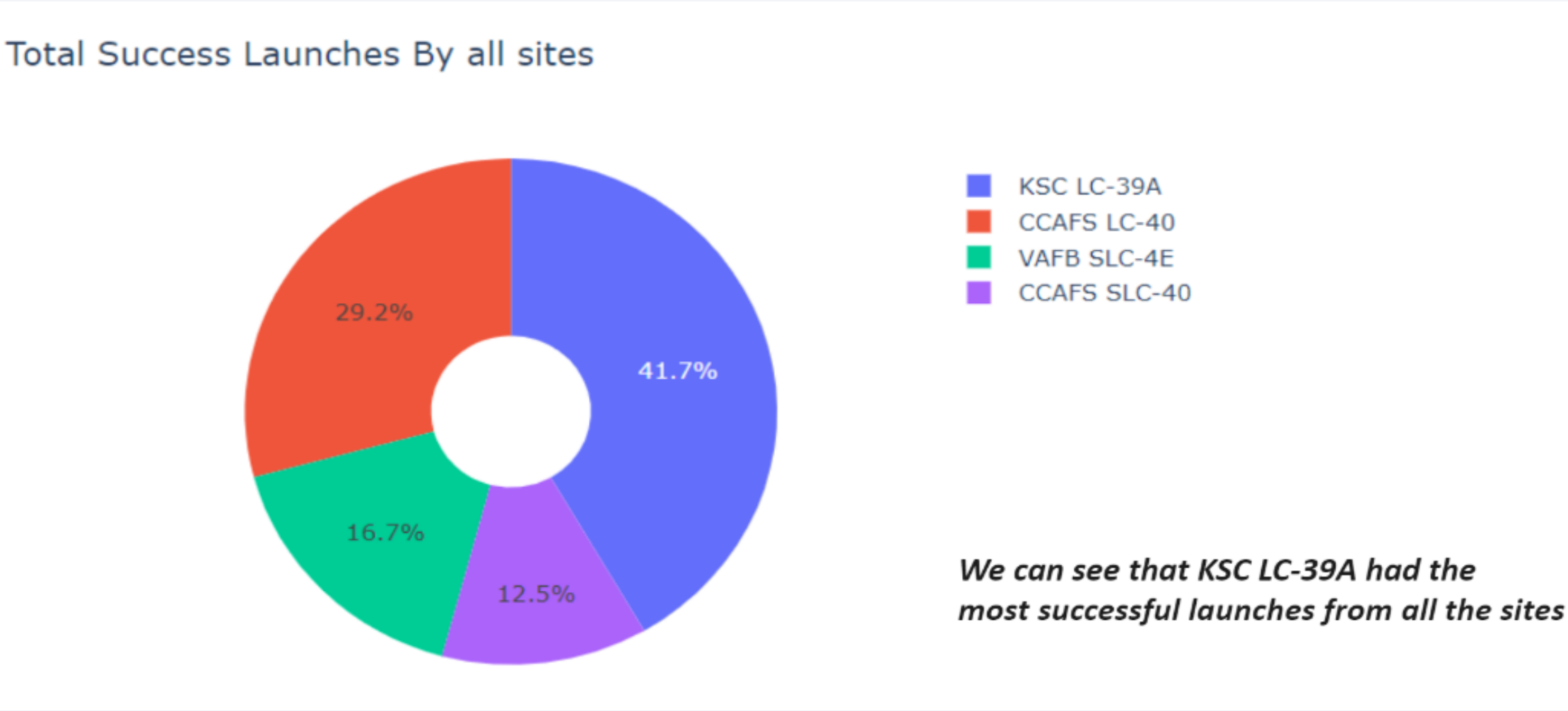
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 5

Build a Dashboard with Plotly Dash

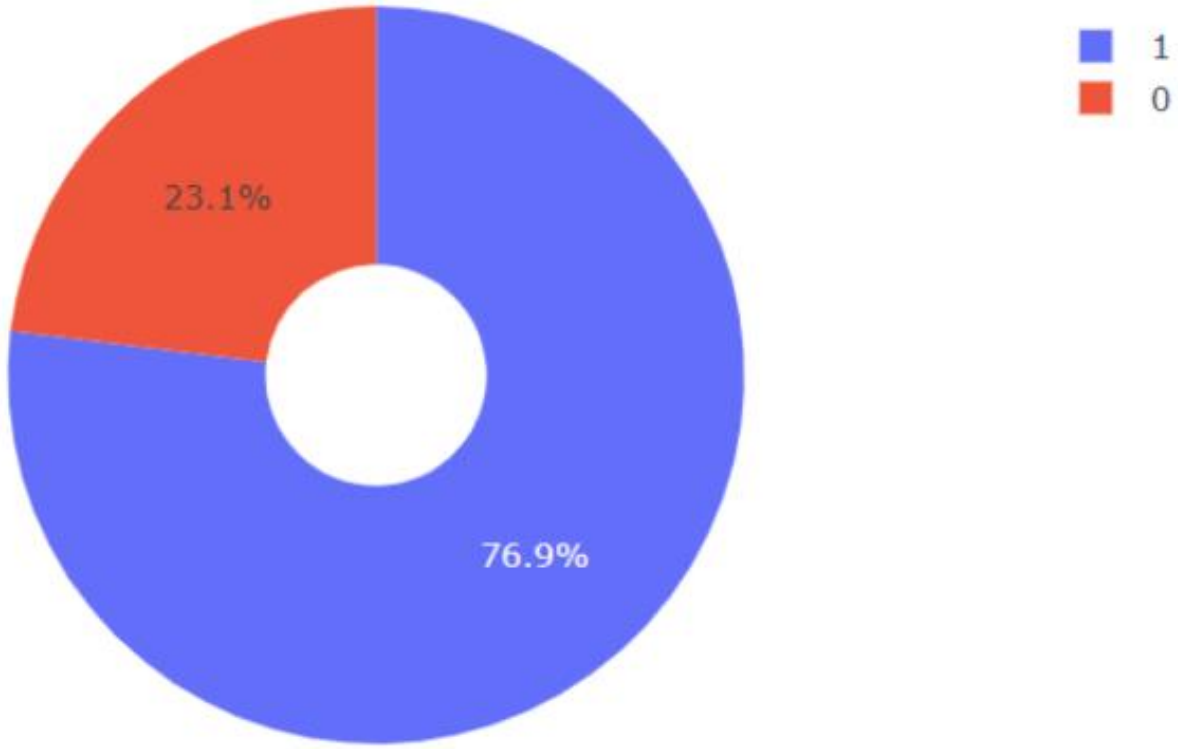


Pie chart showing the success percentage achieved by each launch site



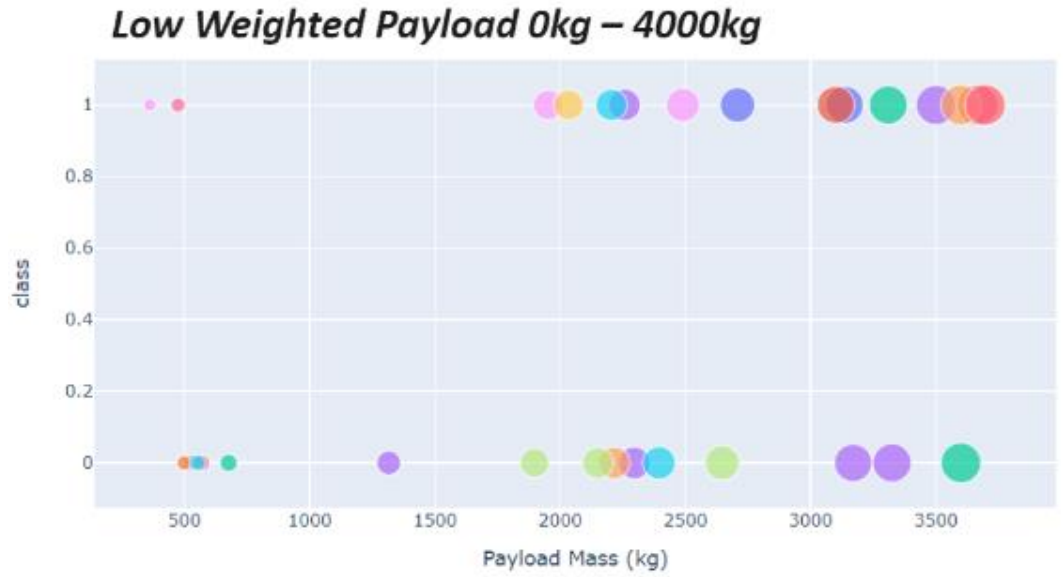
We can see that KSC LC-39A had the most successful launches from all the sites

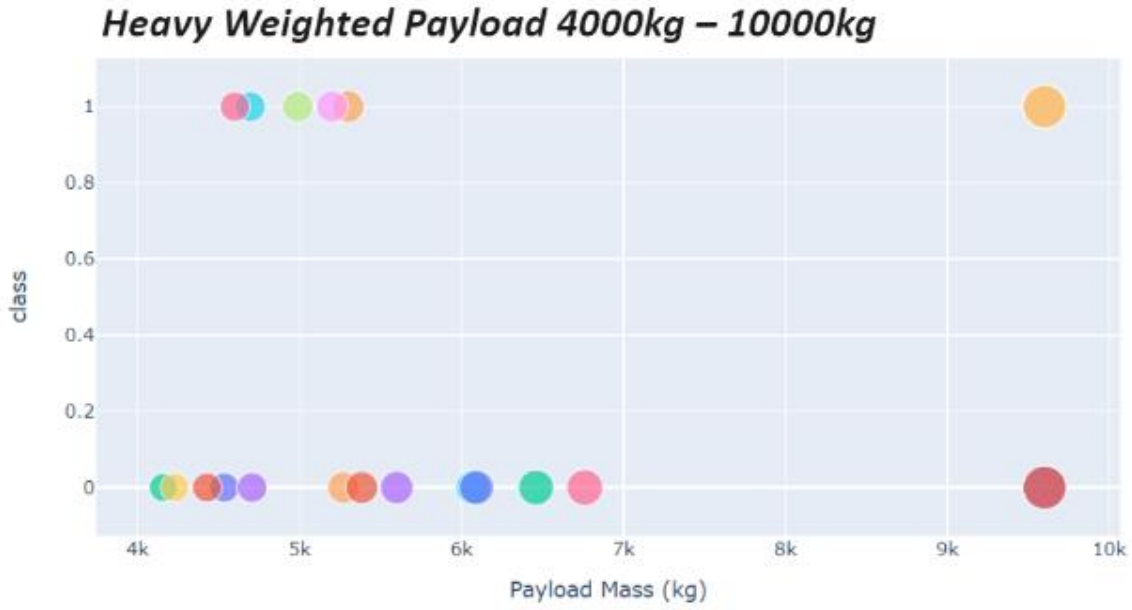
Pie chart showing the Launch site with the highest launch success ratio



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

Section 6

Predictive Analysis (Classification)

Classification Accuracy

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

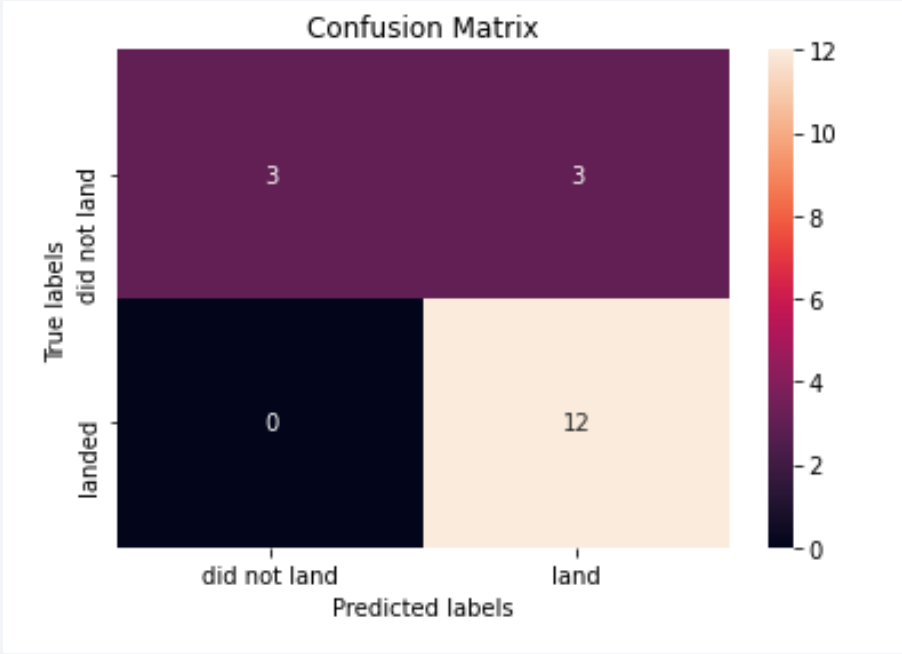
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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 confusion matrix generated for the decision tree classifier reveals its ability to differentiate between various classes. The primary issue lies in the occurrence of false positives, wherein the classifier misidentifies unsuccessful landings as successful ones



Conclusions

1. It can be deduced that there exists a positive correlation between the number of flights at a launch site and the success rate at that site.
2. The launch success rate exhibited an upward trend from 2013 through 2020.
3. Among the different orbital destinations, namely ES-L1, GEO, HEO, SSO, and VLEO, these exhibited the highest success rates.
4. KSC LC-39A stands out as the launch site with the highest number of successful launches.
5. Based on the evaluation, the Decision Tree classifier emerges as the most suitable machine learning algorithm for this specific task.

Thank you!