IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mostafizur Rahman
December 2, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - API Data Collection
  - Web Scraping for Data Acquisition
  - Data Wrangling and Cleaning
  - Exploratory Data Analysis using SQL
  - Data Exploration through Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning for Predictive Analysis
- Summary of all results
  - Exploratory Data Analysis Insights
  - Screenshots of Interactive Analytics
  - Predictive Analytics Outcome

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 play a pivotal role in determining the success of a rocket landing?
  - How do various features interact to influence the success rate of a landing?
  - What operating conditions must be in place to ensure a successful rocket 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

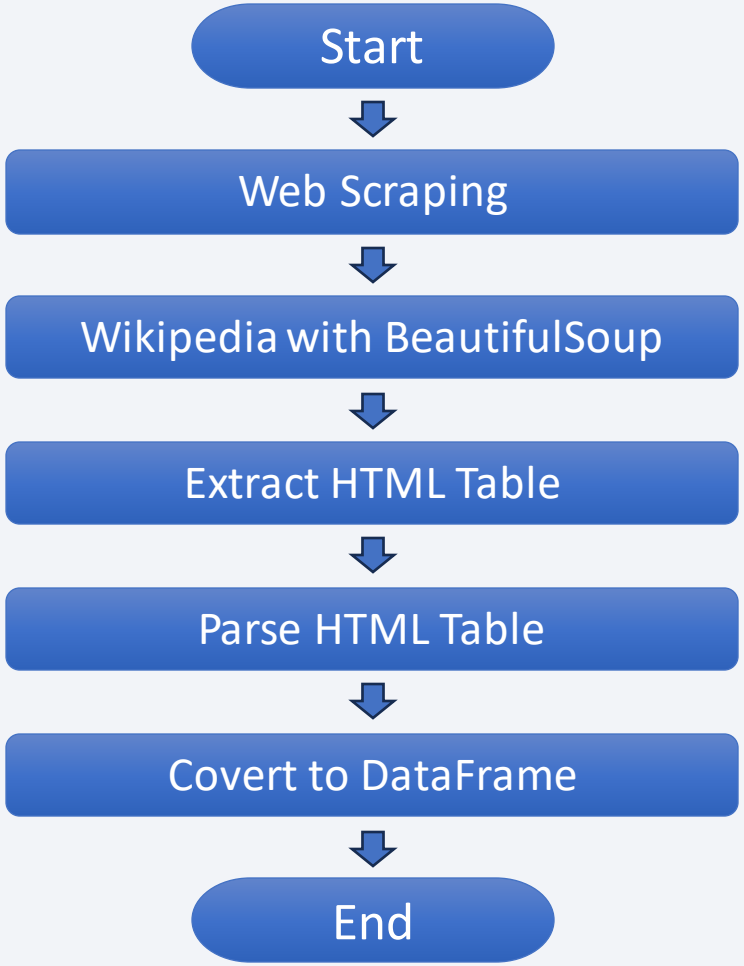
Data Collection

Various methods were employed for data collection:

- SpaceX API was utilized through a GET request to retrieve data.
- The response content was decoded into JSON format using the .json() function and subsequently transformed into a pandas dataframe via .json_normalize().
- Data cleaning processes were implemented, including the identification and handling of missing values.
- Web scraping from Wikipedia using BeautifulSoup was conducted to obtain Falcon 9 launch records.
- The goal was to extract launch records from an HTML table, parse the table, and convert the information into a pandas dataframe for subsequent analysis.

Data Collection - Scraping

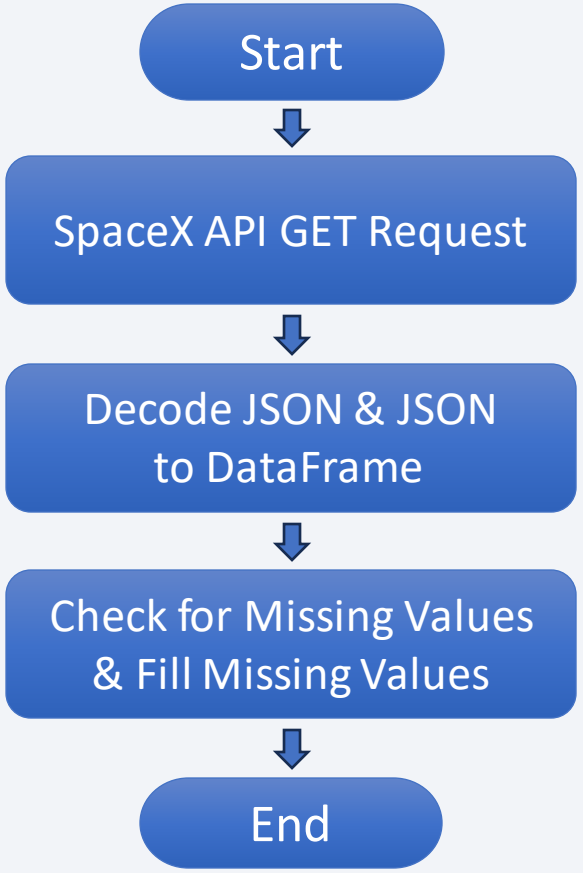
- We applied web scraping to extract Falcon 9 launch records from a webpage using BeautifulSoup. Subsequently, we parsed the table and converted the retrieved information into a pandas dataframe.
- The link to the notebook is M1.2.jupyter-labs-webscraping.ipynb



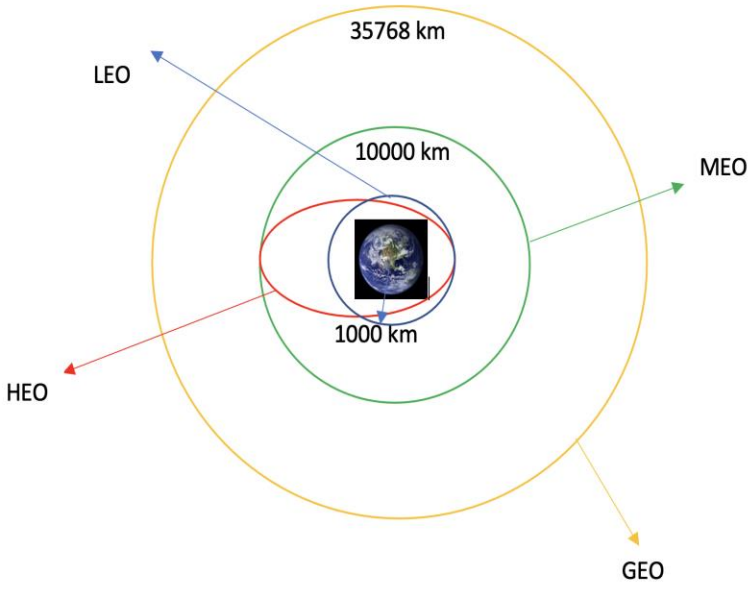
Data Collection – SpaceX API

- We utilized a GET request to the SpaceX API to retrieve data, performed data cleaning on the acquired information, and conducted essential data wrangling and formatting.

- The link to the notebook is M1.1.jupyter-labs-spacex-data-collection-api.ipynb



Data Wrangling



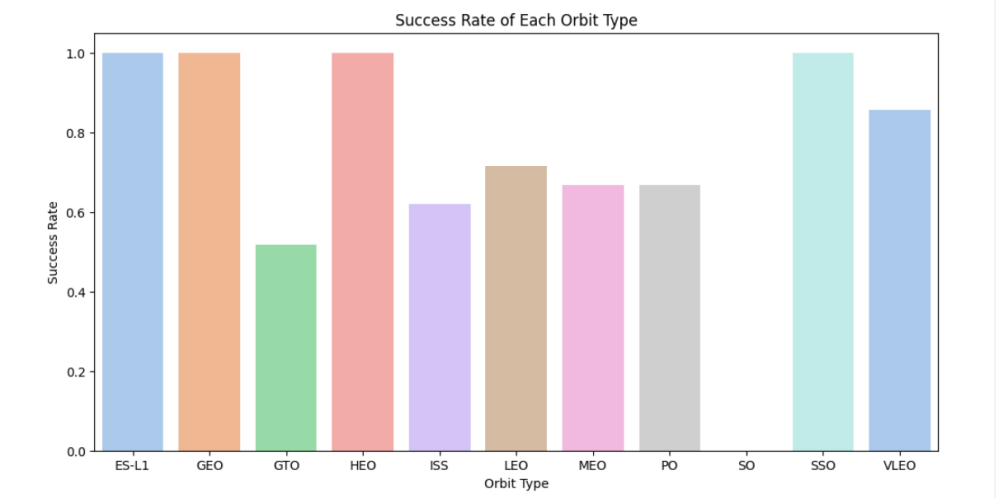
- We Calculated the number of launches on each site
- We Calculated the number and occurrence of each orbit
- We Calculated the number and occurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- The link to the notebook is M1.3.labs-jupyter-spacex-Data wrangling.ipynb

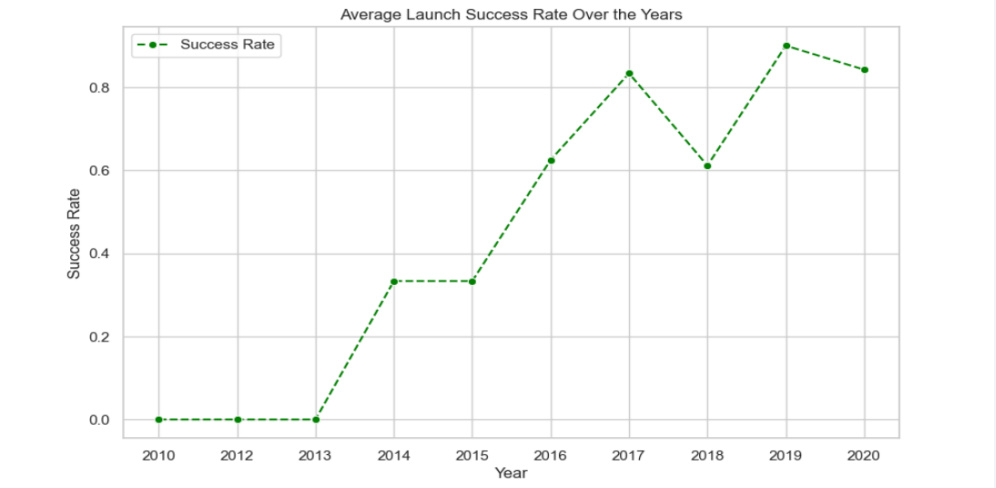
EDA with Data Visualization

Conducted an in-depth exploration of the data through visualizations examining relationships such as:

- Flight number and launch site
- Payload and launch site
- Success rate of each orbit type
- Flight number and orbit type
- Launch success yearly trend

The link to the notebook is M1.3.labs-jupyter-spacex-Data_wrangling.ipynb





EDA with SQL

Utilized Exploratory Data Analysis (EDA) with SQL to extract meaningful insights from the dataset. Formulated queries to unveil key information, such as:

- ❖ Identifying unique launch site names in the space mission.
- ❖ Determining the total payload mass carried by boosters launched by NASA (CRS).
- ❖ Calculating the average payload mass carried by booster version F9 v1.1.
- ❖ Analyzing the total number of successful and failed mission outcomes.
- ❖ Investigating failed landing outcomes on drone ships, including details on booster versions and launch site names.

The link to the notebook is M2.1.jupyter-labs-eda-sql-coursera_sqllite.ipynb

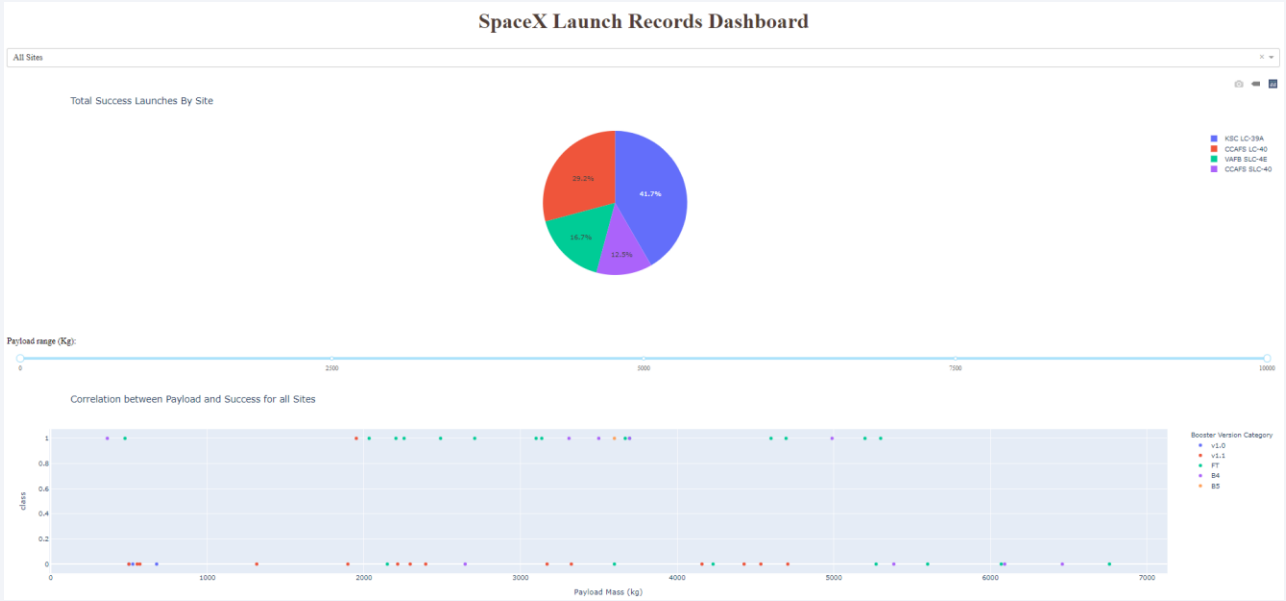
Build an Interactive Map with Folium

- Geospatial Visualization: Marked all launch sites on a folium map, incorporating map objects like markers, circles, and lines to signify the success or failure of launches at each site.
- Feature Assignment: Assigned launch outcomes (failure or success) to classes 0 and 1, with 0 representing failure and 1 representing success.
- Success Rate Identification: Utilized color-labeled marker clusters to pinpoint launch sites with relatively high success rates.
- Proximity Analysis: Calculated distances between launch sites and their surroundings, addressing questions such as:
  - Are launch sites situated near railways, highways, and coastlines?
  - Do launch sites maintain a specific distance from cities?

The link to the notebook is M3.1.lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Developed an interactive dashboard using Plotly Dash.
- Created informative visualizations, including:
  - Pie charts illustrating the success and failure percentages for specific sites.
  - Scatter graphs depicting the relationship between Outcome and Payload Mass (Kg) for various booster versions.
- The link to the notebook is M3.2.Build a Dashboard.ipynb



Predictive Analysis (Classification)

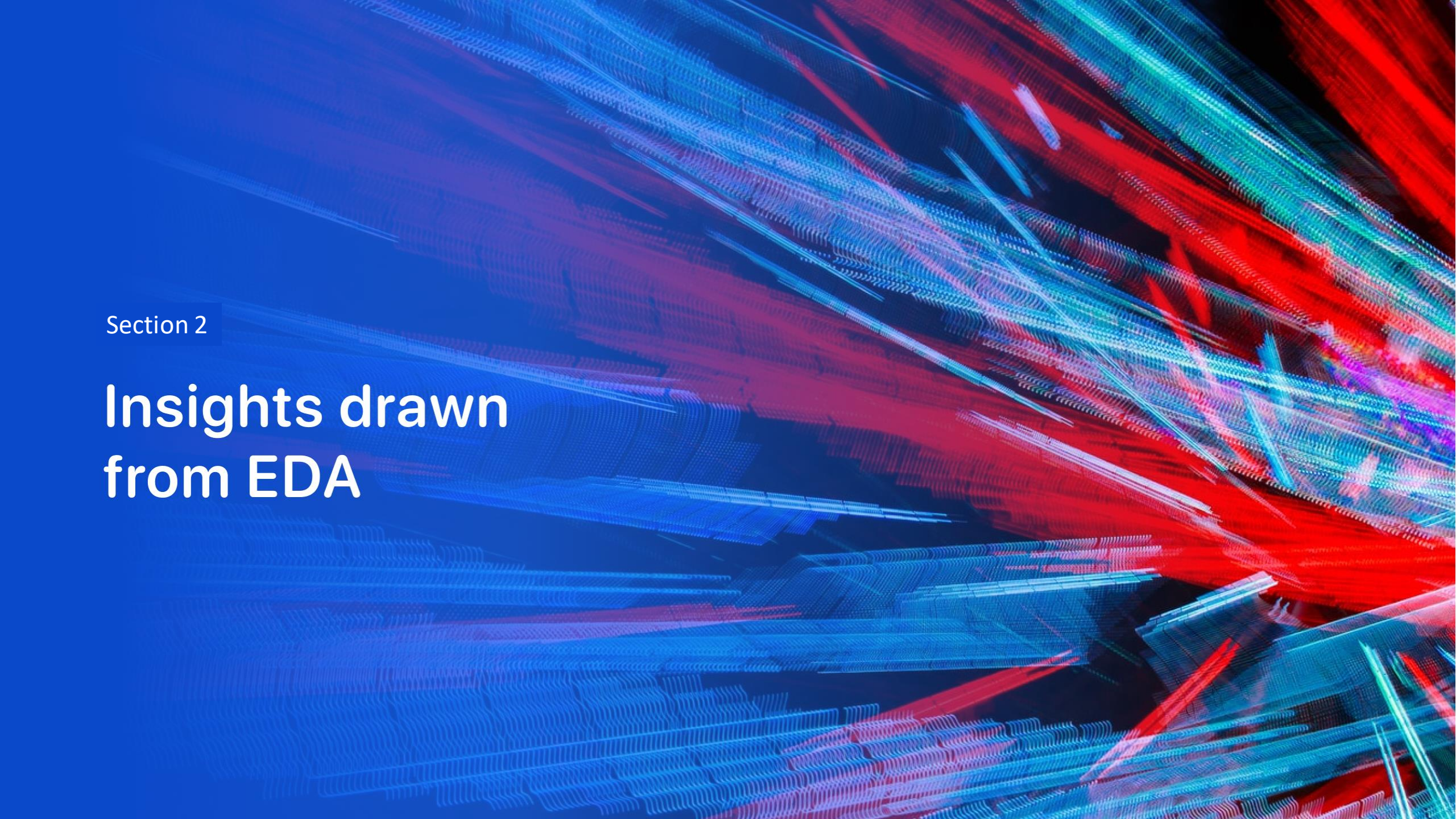
- Employed numpy and pandas for data loading, transformation, and splitting into training and testing sets.
- Constructed diverse machine learning models, optimizing hyperparameters through GridSearchCV.
- Evaluated models based on accuracy, continually enhancing performance through feature engineering and algorithm tuning.
- Identified the most effective classification model.
- The link to the notebook is M4.1.SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

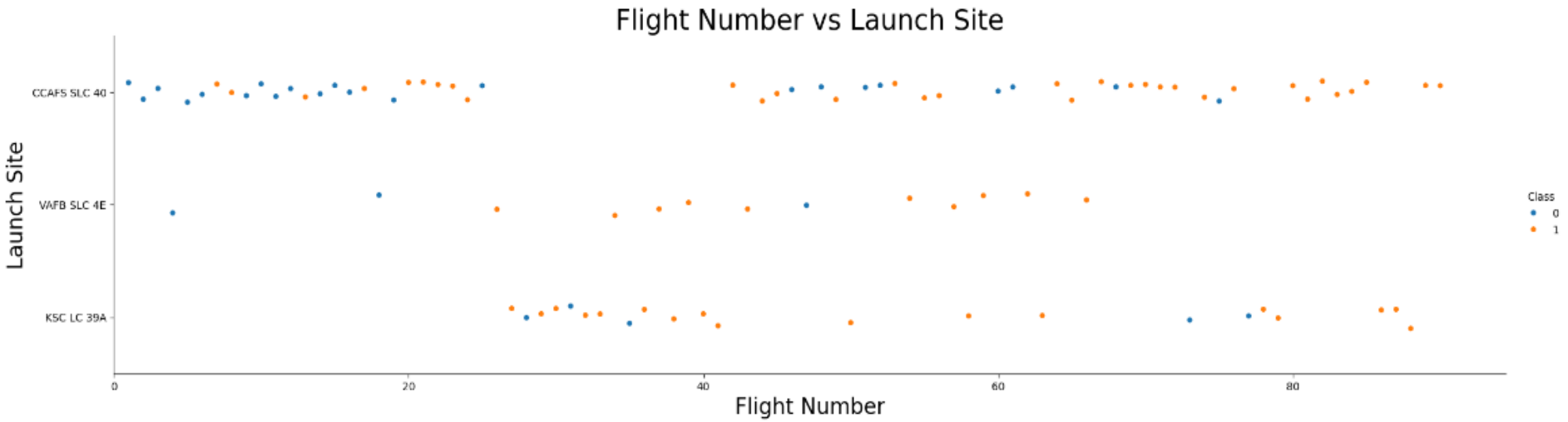
Results

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

Section 2

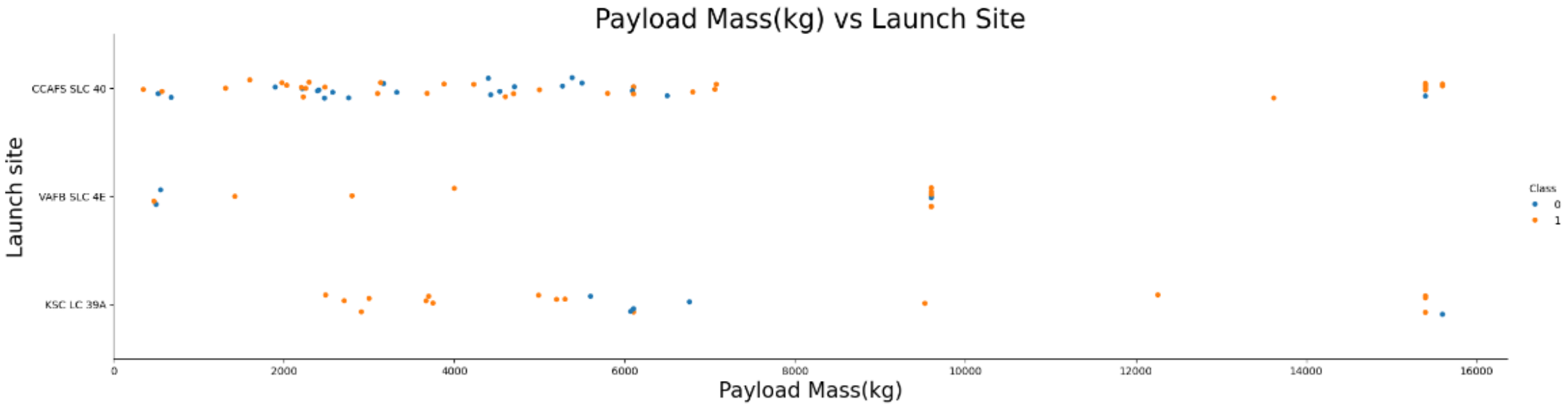
Insights drawn from EDA





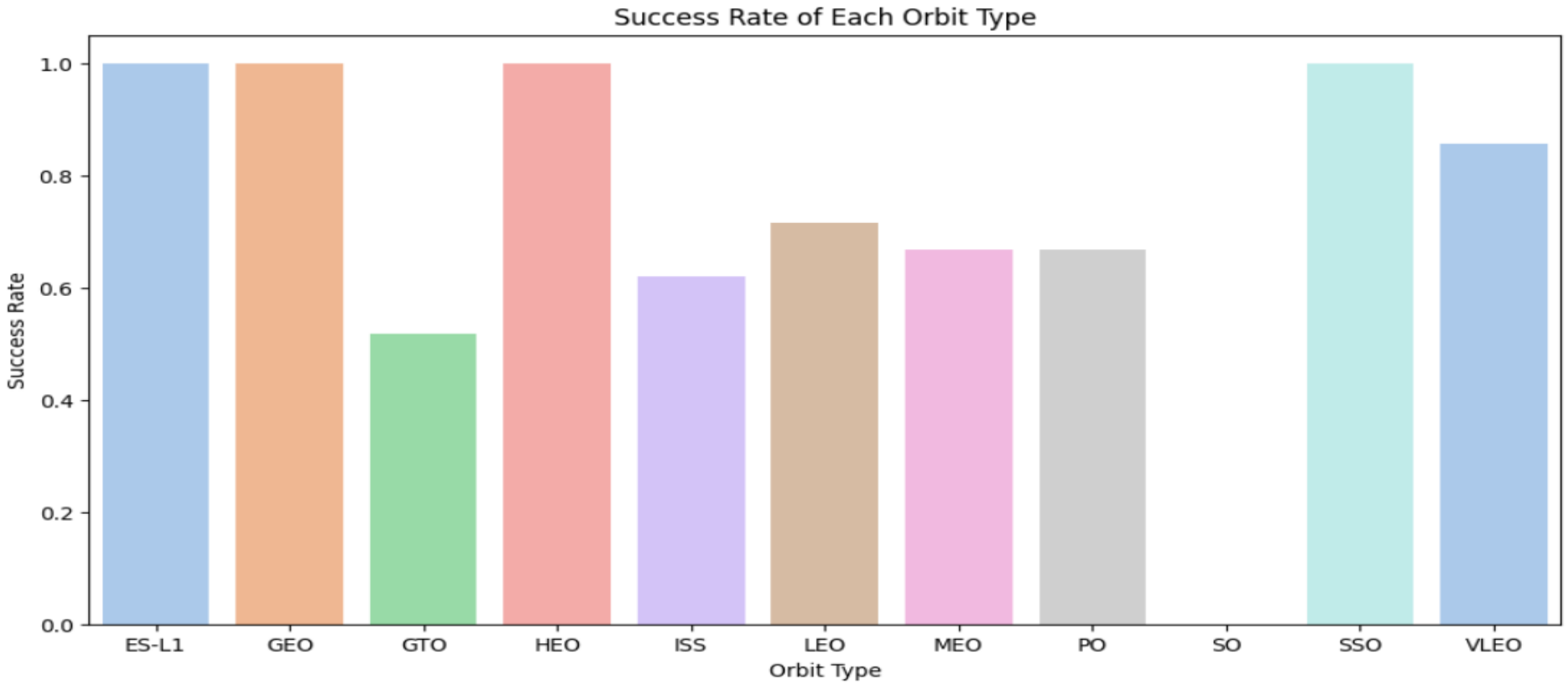
Flight Number vs. Launch Site

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



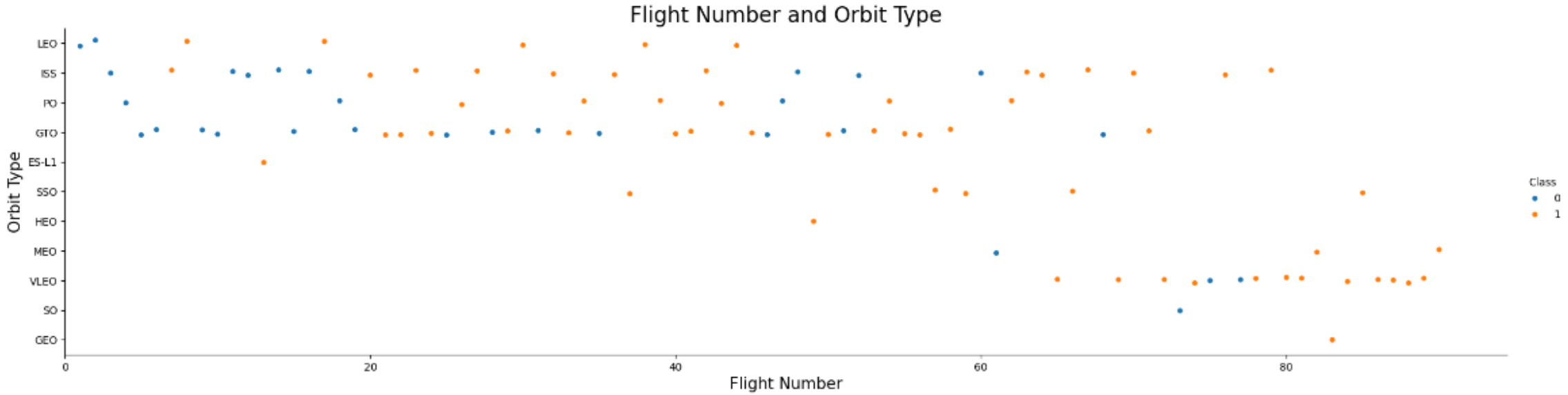
Payload vs. Launch Site

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



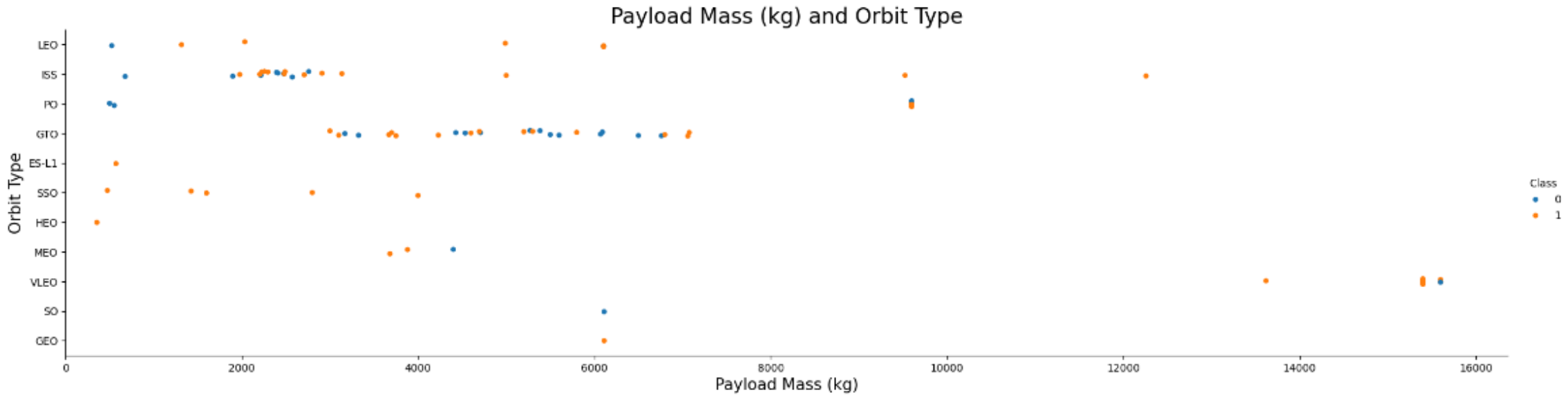
Success Rate vs. Orbit Type

- From the plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



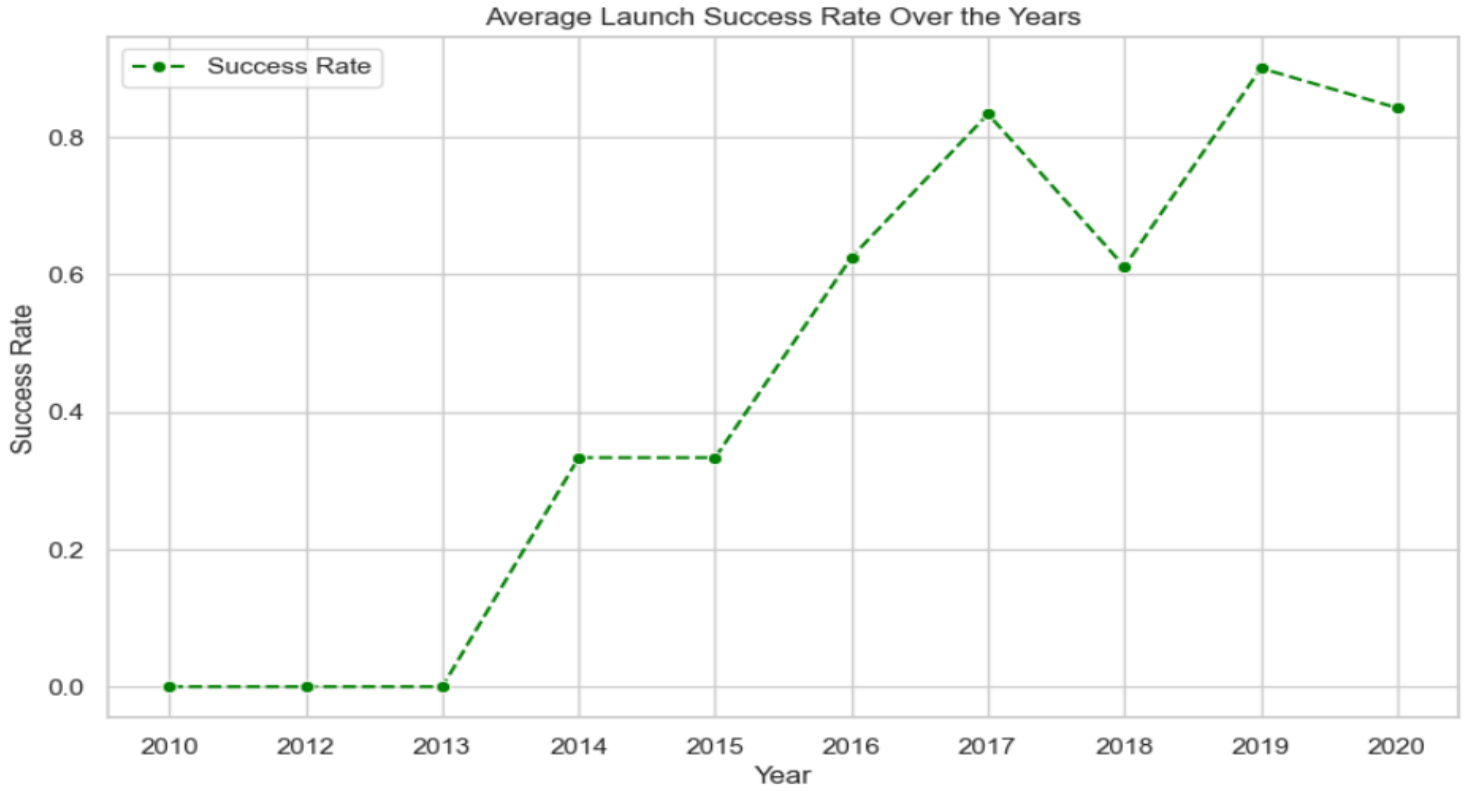
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS 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 from the SpaceX data.

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

[8]:
```
%%sql
select distinct "Launch_Site" from SPACEXTABLE;
```

```
 * sqlite:///my_data1.db
Done.
```

[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

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

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

[14]:
```
%%sql
select * from SPACEXTABLE
where "Launch_Site" like 'CCA%'
limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

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)

[15]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

[15]:

Total_Payload_Mass
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

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

[16]:
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

[16]:

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

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

[17]:
```
%%sql
SELECT MIN(Date) AS First_Successful_Landing_Date
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

First_Successful_Landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the WHERE clause to filter for boosters which have successfully landed on drone ship and applied the AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

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

[18]:
```
%%sql
SELECT DISTINCT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- We observed that there were total 98 successful mission.

List the total number of successful and failure mission outcomes

[21]:
```
%%sql
SELECT COUNT(Mission_Outcome) AS Successful_Mission
FROM SPACEXTABLE
WHERE Mission_Outcome = 'Success';
```

```
 * sqlite:///my_data1.db
Done.
```

[21]:

Successful_Mission
98

Boosters Carried Maximum Payload

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

[22]:
```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

```
 * sqlite:///my_data1.db
Done.
```

[22]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function

2015 Launch Records

- We used substr() function to get months and year.

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

[23]:
```
%%sql
SELECT
    strftime('%m', Date) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2015'
    AND Landing_Outcome = 'Failure (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

[23]:

Month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- We applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to order the grouped landing outcome in descending order.

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

[26]:
```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

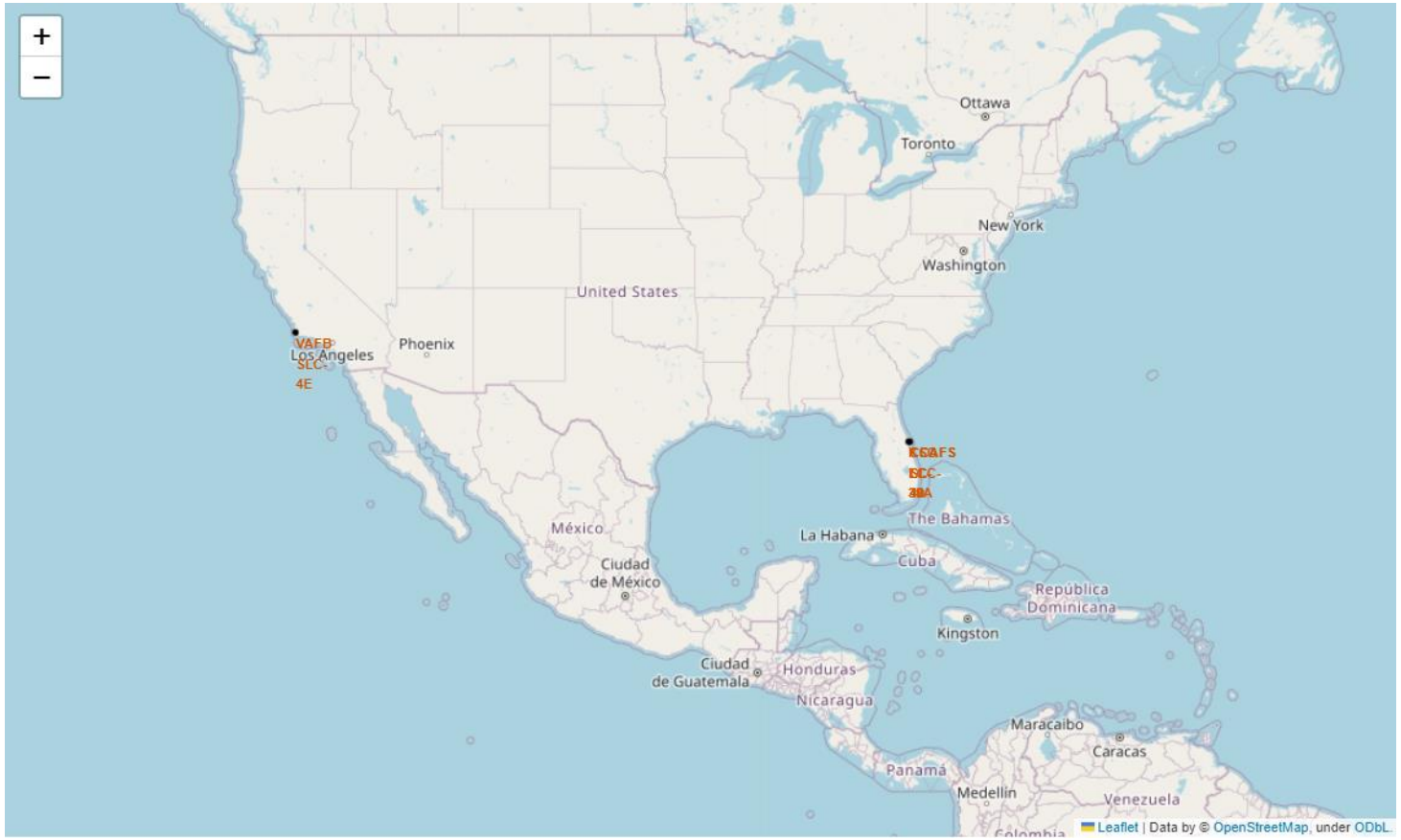
[26]:

Landing_Outcome	Count
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1


Section 3
Launch Sites
Proximities Analysis

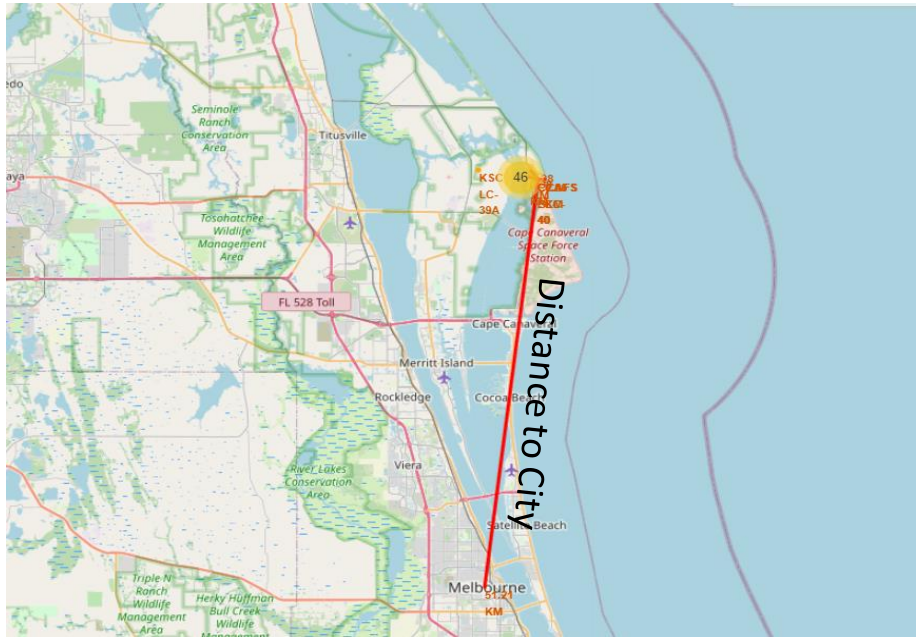
Global Map Markers for All Launch Sites

The SpaceX launch sites are prominently located along the coastlines of the United States, specifically in California and Florida.



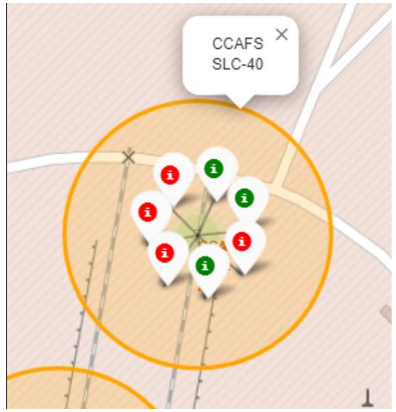
Distances from Launch Sites to Landmarks

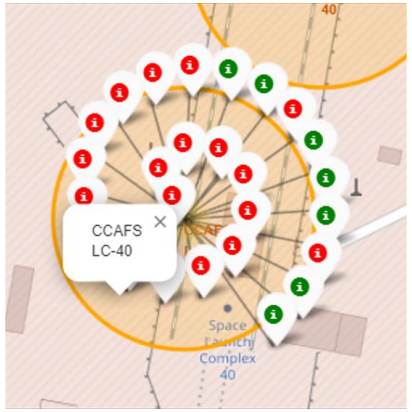
- Are launch sites in close proximity to railways? Yes
- Are launch sites in close proximity to highways? Yes
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

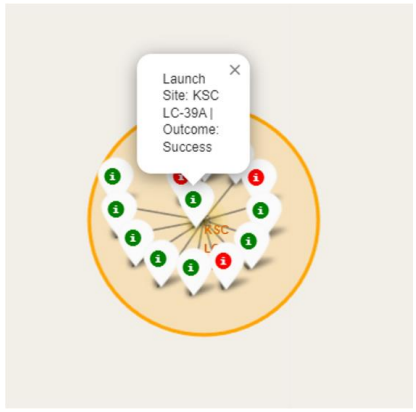


Markers displaying launch sites with color-coded labels

Green Marker shows Successful lunches and ***Red Marker*** shows Failure.

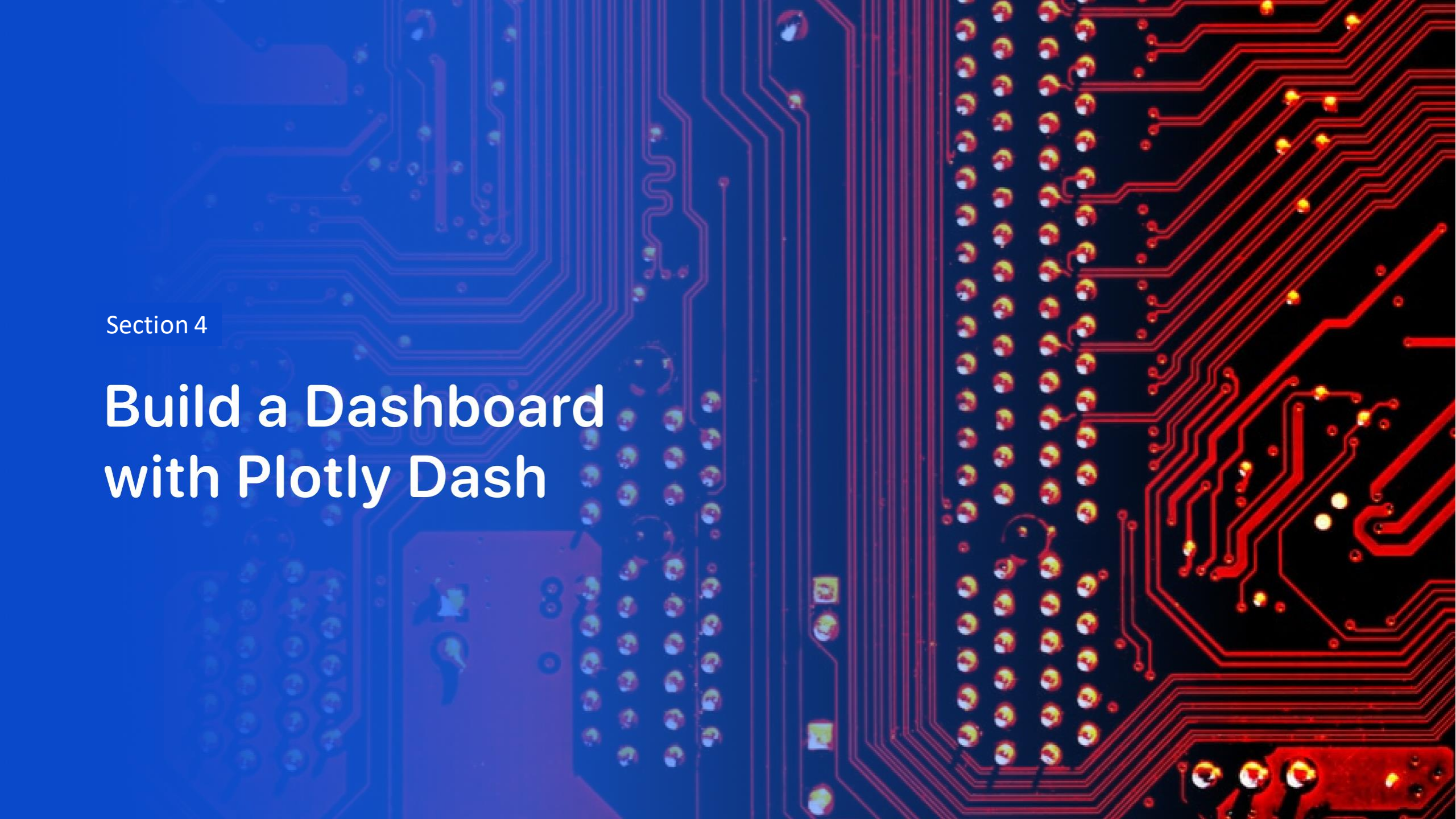




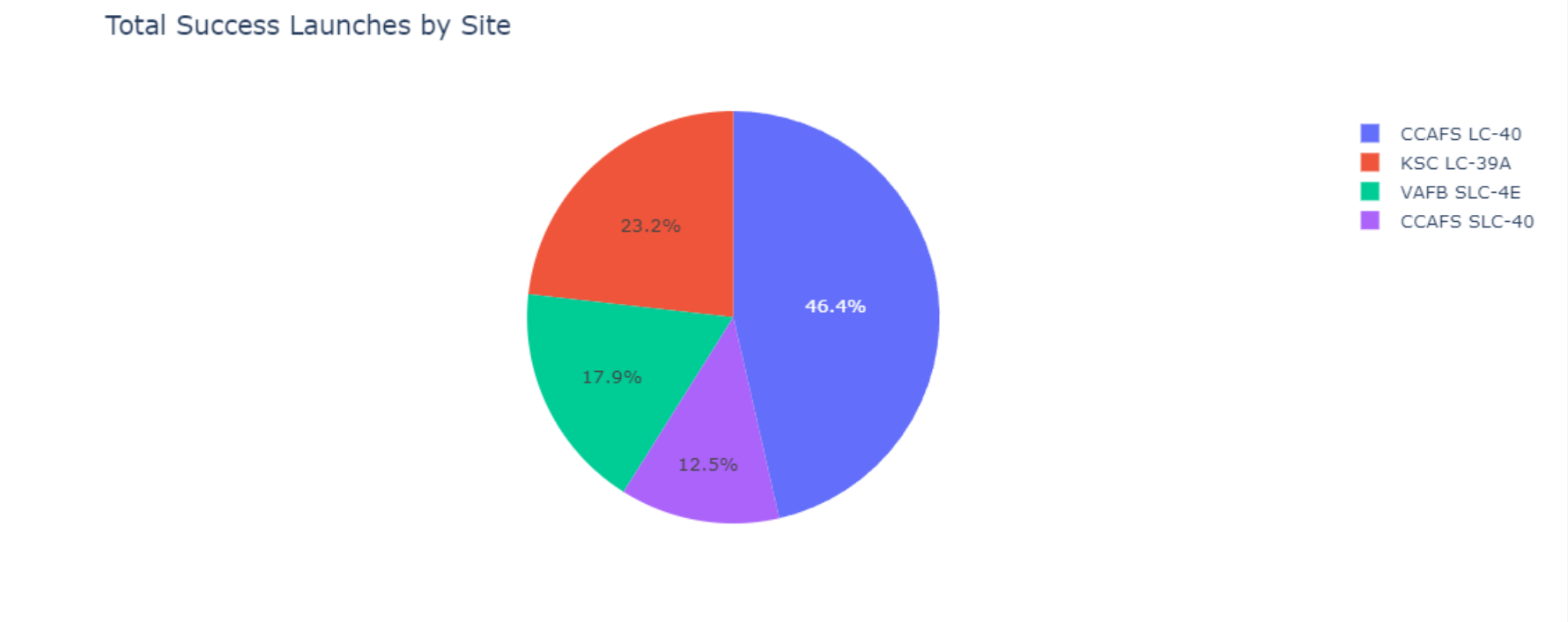


California Launch Sites

Florida Lauch Site

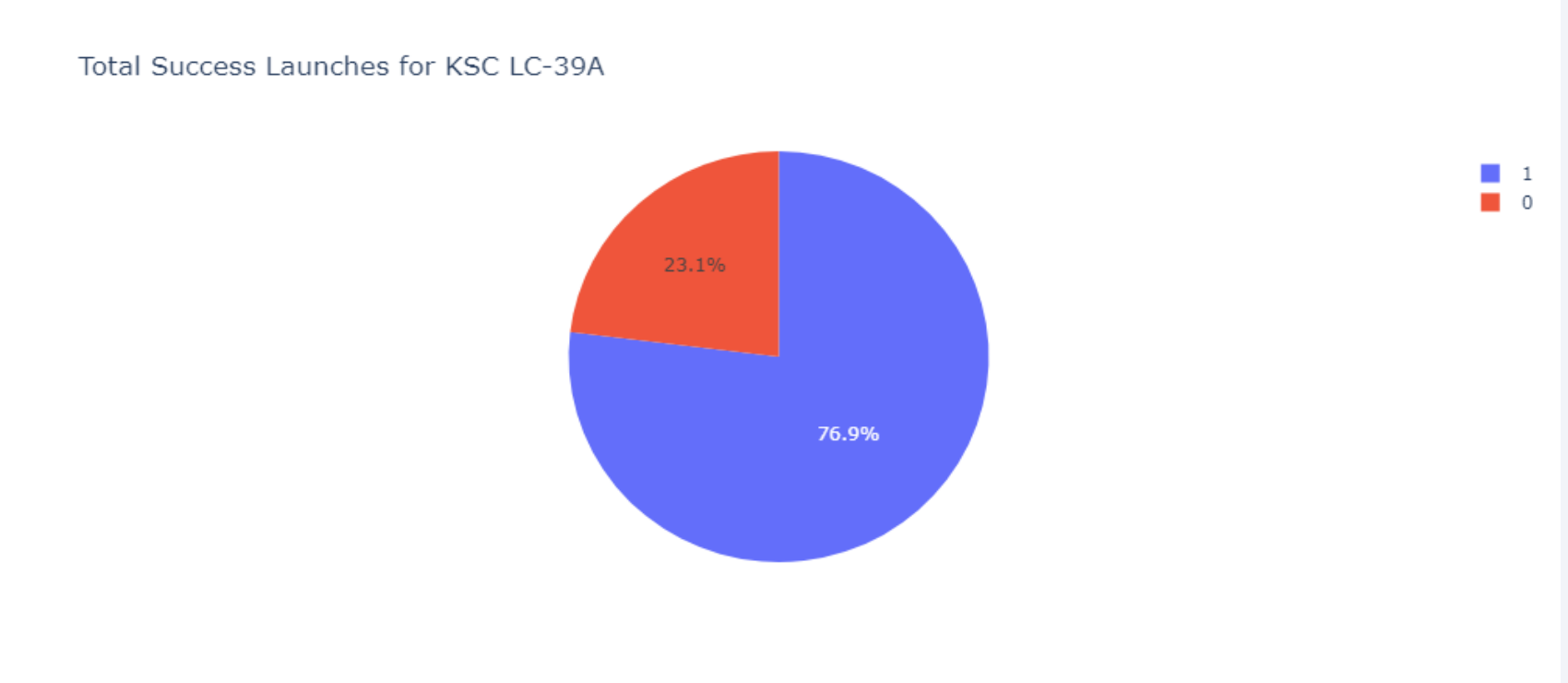

Section 4
Build a Dashboard with Plotly Dash

Pie Chart Illustrating Success Percentage for each Launch Site



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

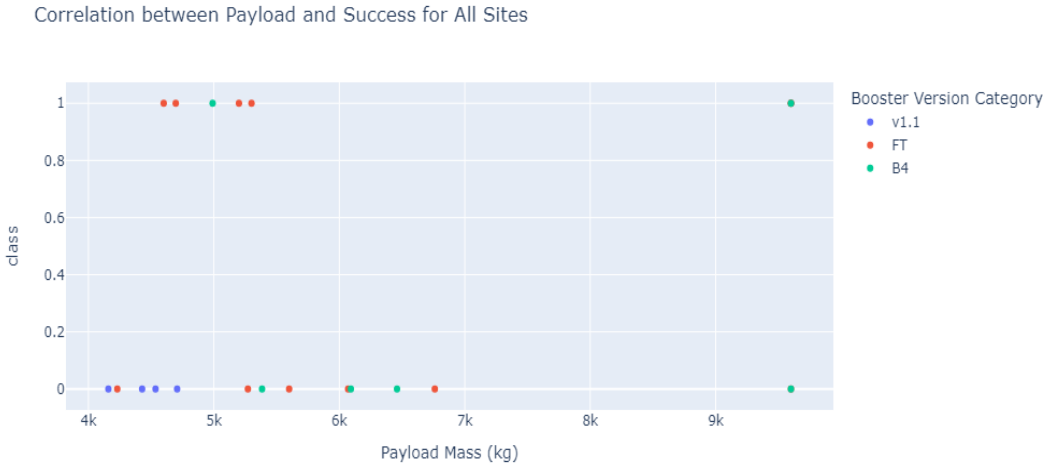
Pie Chart Illustrating Launch Site with the Highest Launch Success Ratio



KSC LC-39A attained a success rate of 76.9% and experienced a failure rate of 23.1%.

Scatter Plot of Payload vs Launch Outcome for All Sites, with Various Payloads Selected in the Range Slider

Low Weighted Payload (0-4000 KG)



Heavy Weighted Payload (0-4000 KG)

We can see that the success rate for low weighted payloads is higher than the heavy weighted payloads.

Section 5

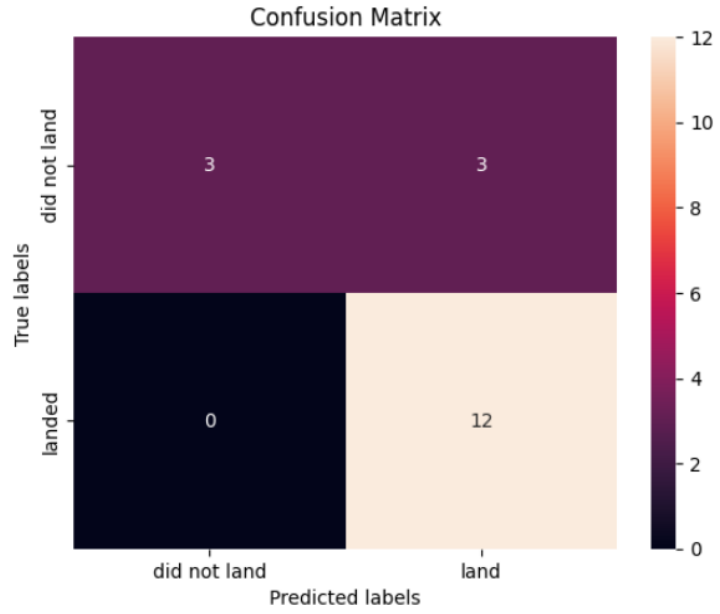
Predictive Analysis (Classification)

Classification Accuracy

```python
# Calculate accuracy scores
accuracy_lr = logreg_cv.score(X_test, Y_test)
accuracy_svm = svm_cv.score(X_test, Y_test)
accuracy_tree = tree_cv.score(X_test, Y_test)
accuracy_knn = knn_cv.score(X_test, Y_test)

# Create a dictionary to store the accuracy scores
accuracy_scores = {
    'Logistic Regression': accuracy_lr,
    'SVM': accuracy_svm,
    'Decision Tree': accuracy_tree,
    'KNN': accuracy_knn
}

# Find the method that performs best
best_method = max(accuracy_scores, key=accuracy_scores.get)
best_accuracy = accuracy_scores[best_method]

print("Accuracy Scores:")
for method, accuracy in accuracy_scores.items():
    print(f"{method}: {accuracy:.4f}")

print(f"The best-performing method is {best_method} with an accuracy of {best_accuracy:.4f}.")
```

```
Accuracy Scores:
Logistic Regression: 0.8333
SVM: 0.8333
Decision Tree: 0.7222
KNN: 0.8333
The best-performing method is Logistic Regression with an accuracy of 0.8333.
```

Classification Accuracy



- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

Based on our analysis, we can draw the following conclusions:

- There appears to be a positive correlation between the number of flights at a launch site and the success rate.
- The launch success rate exhibited an upward trend from 2013 to 2020.
- Orbits ES-L1, GEO, HEO, SSO, and VLEO demonstrated the highest success rates.
- KSC LC-39A emerged as the launch site with the highest number of successful launches.
- Logistic regression proved to be the most effective machine learning algorithm for this specific task.

Thank you!