Winning Space Race with Data Science

Olisamedua Okafor
September 9th, 2023

Outline

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

Executive Summary

- **Overview of Methodologies**: A comprehensive summary of the techniques and approaches used throughout the project.
  - **Data Collection Methods**: Utilized SpaceX API and web scraping techniques to gather relevant data.
  - **Data Preprocessing and EDA**: Conducted data wrangling to clean and format the raw data, followed by Exploratory Data Analysis to uncover trends and patterns.
  - **Machine Learning Models**: Employed various classification algorithms, including decision trees, to make predictions.
- **Results Summary**: A consolidated report of all findings, analyses, and predictions made during the project.
  - **Key Determinants of Success**: Identified grid fins, launch sites, and payload mass as significant factors contributing to successful launches.
  - **Predictive Modeling**: Utilized a Decision Tree Classifier to accurately predict the outcomes of SpaceX launches.

Introduction

- The first stage of the rocket is a critical factor for successful launches.
  - Being considerably larger and more expensive than the second stage, the first stage represents a significant investment.
  - SpaceX gains a cost advantage over other providers by reusing the first stage.
- Accurately predicting the success of each first stage landing could result in substantial cost savings.
- Therefore, my objectives are as follows:
  - Identify the key factors that contribute to the success and cost of each launch.
  - Utilize this information to predict the outcomes of future SpaceX launches.

Section 1

Methodology

Methodology

Executive Summary

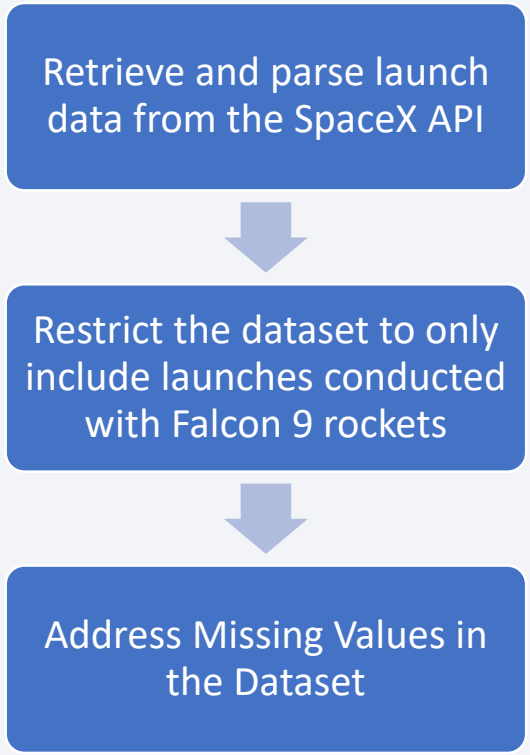
- Data collection methodology:
  - Acquired comprehensive datasets through web scraping techniques and by leveraging the SpaceX API for targeted data extraction.
- Perform data wrangling
  - Eliminated extraneous data, addressed missing values, and engineered a new column specifically for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Divide the data into training and test sets, then calculate the performance score of each model to identify the best-fitting one.

Data Collection

- Data was gathered from two primary sources: the SpaceX API (https://api.spacexdata.com/v4/rockets/) and a Wikipedia page listing Falcon 9 and Falcon Heavy launches (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches). The collection methods employed API calls for SpaceX data and web scraping techniques for the Wikipedia information.

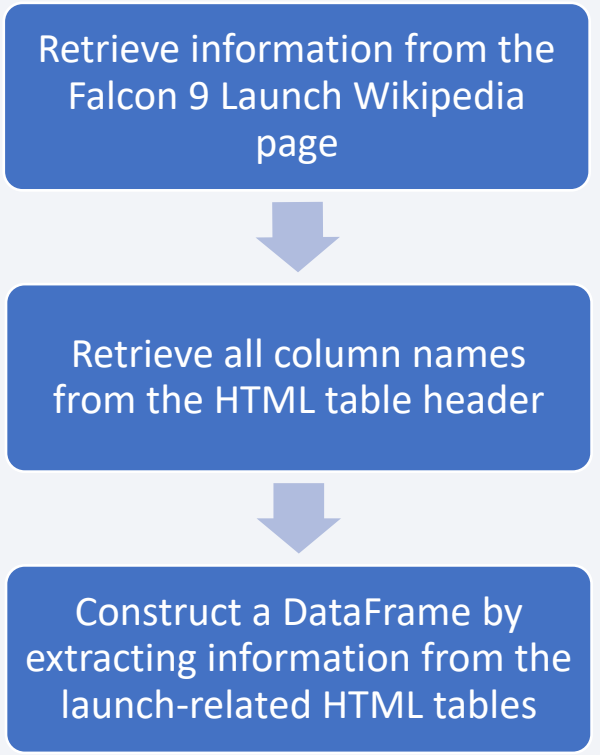
Data Collection – SpaceX API

- SpaceX provides a public API that allows for the convenient retrieval and utilization of data.
- The API was utilized as outlined in the adjacent flowchart, after which the data was saved for long-term storage.
- Source code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%201/jupyter-labs-spacex-data-collection-api.ipynb



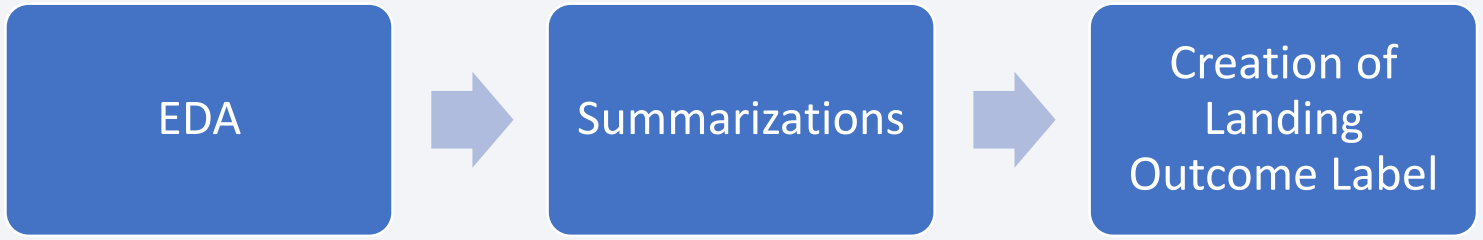
Data Collection - Scraping

- Information on SpaceX launches is also accessible through Wikipedia.
- Data is sourced from Wikipedia following a specific flowchart and is then stored for long-term use.
- Source code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%201/jupyter-labs-webscraping.ipynb



Data Wrangling

- Initially, I conducted exploratory data analysis (EDA) on the dataset.
- Subsequently, we calculated the number of launches per site, the frequency of each orbit type, and the distribution of mission outcomes for each orbit category.
- Ultimately, the 'landing outcome' label was generated from the 'Outcome' column.



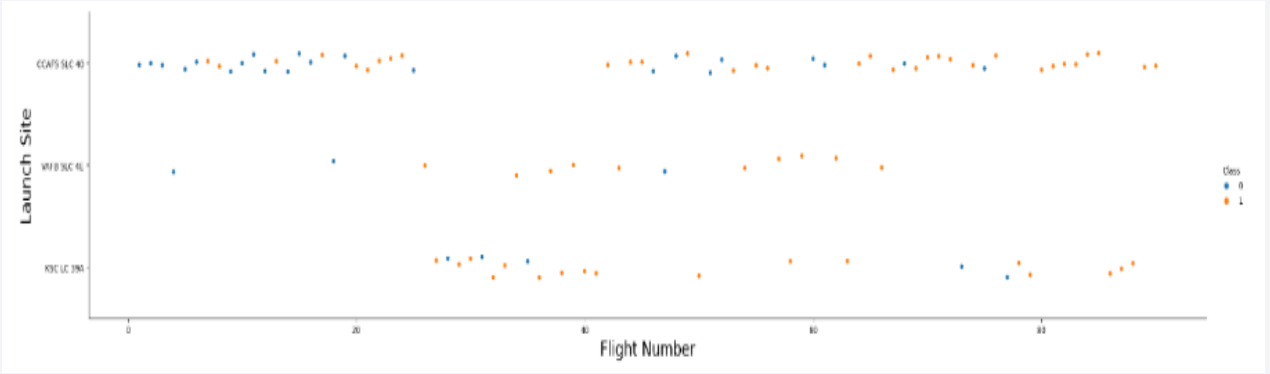
- Source code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%201/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with SQL

- **SQL Queries Performed:**
  - Retrieved names of unique launch sites involved in the space mission.
  - Identified the top 5 launch sites with names starting with 'CCA'.
  - Calculated the total payload mass carried by NASA (CRS) boosters.
  - Computed the average payload mass for boosters of version F9 v1.1.
  - Determined the date of the first successful ground pad landing.
  - Listed names of boosters successful in drone ship landings with payload mass between 4,000 and 6,000 kg.
  - Tallied the total number of successful and failed mission outcomes.
  - Named the booster versions that have carried the maximum payload mass.
  - Highlighted failed drone ship landings in 2015, including booster versions and launch site names.
  - Ranked the frequency of landing outcomes (e.g., Failure (droneship), Success (ground pad)) between June 4, 2010, and March 20, 2017.
- Source Code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%202/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

- To explore the data, various types of visualizations were employed, specifically scatterplots and bar plots.
- Relationships between pairs of features were examined as follows:
  - Payload Mass and Flight Number
  - Launch Site and Flight Number
  - Launch Site and Payload Mass
  - Orbit and Flight Number
  - Payload and Orbit



- Source code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%202/jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

- Utilized various mapping elements in Folium Maps, including markers, circles, lines, and marker clusters.
  - **Markers**: Used to pinpoint specific locations, such as launch sites.
  - **Circles**: Highlight areas around specific coordinates, like the NASA Johnson Space Center.
  - **Marker Clusters**: Group together events at each coordinate, such as multiple launches at a single launch site.
  - **Lines**: Indicate distances between two coordinates, aiding in spatial understanding.
- Source Code: https://github.com/okaforoa/ibm-data-science-capstone/blob/main/Week%203/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- Developed an interactive dashboard using Plotly Dash.
- Visualized total launches by specific sites through pie charts.
- Illustrated the relationship between 'Outcome' and 'Payload Mass (Kg)' across different booster versions using scatter plots.
- The dashboard code is accessible at this GitHub link.

Predictive Analysis (Classification)

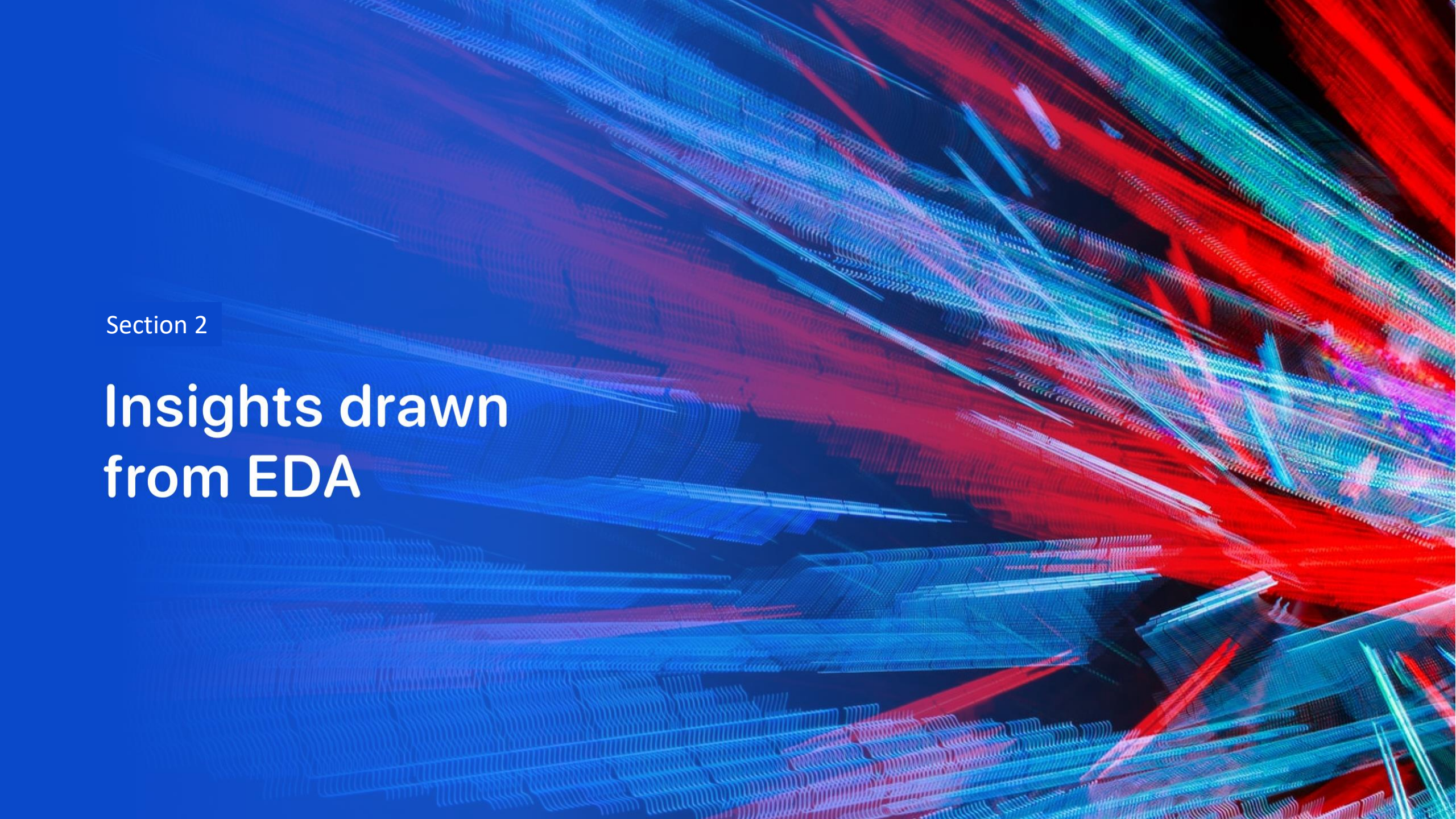
- Loaded the data into the environment using Numpy and Pandas libraries.
- Transformed and preprocessed the raw data for analysis.
- Split the dataset into training and testing sets for model validation.
- Built multiple machine learning models and fine-tuned their hyperparameters using GridSearchCV.
- Utilized accuracy as the evaluation metric for model performance.
- Enhanced the model's performance through feature engineering and algorithm tuning.
- Identified the best-performing classification model for our dataset.
- For further details, refer to the GitHub Notebook.

Results

- **Launch Sites**: SpaceX utilizes four different launch sites.
- **Initial Partners**: The first launches were conducted for SpaceX itself and NASA.
- **Average Payload**: The F9 v1.1 booster has an average payload capacity of 2,928 kg.
- **First Successful Landing**: The first successful landing occurred in 2015, five years after the initial launch.
- **Above-Average Payloads**: Many versions of the Falcon 9 booster successfully landed on drone ships while carrying payloads above the average.
- **Mission Success Rate**: Nearly 100% of missions have been successful.
- **Failed Landings in 2015**: Two booster versions—F9 v1.1 B1012 and F9 v1.1 B1015—failed to land on drone ships in 2015.
- **Improving Landing Success**: The rate of successful landings has improved over the years.

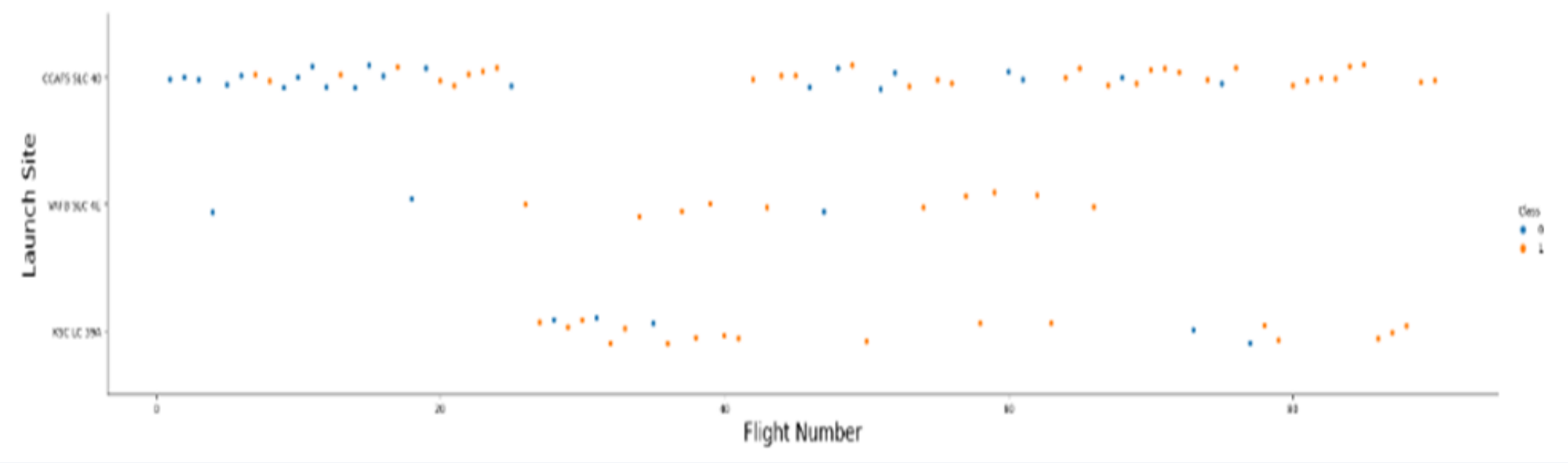
Results

- Through interactive analytics, it was observed that launch sites are typically located in safe areas, often near the sea.
- These sites are also supported by robust logistical infrastructure.
- A majority of launches occur at launch sites located on the East Coast.


Section 2
Insights drawn from EDA

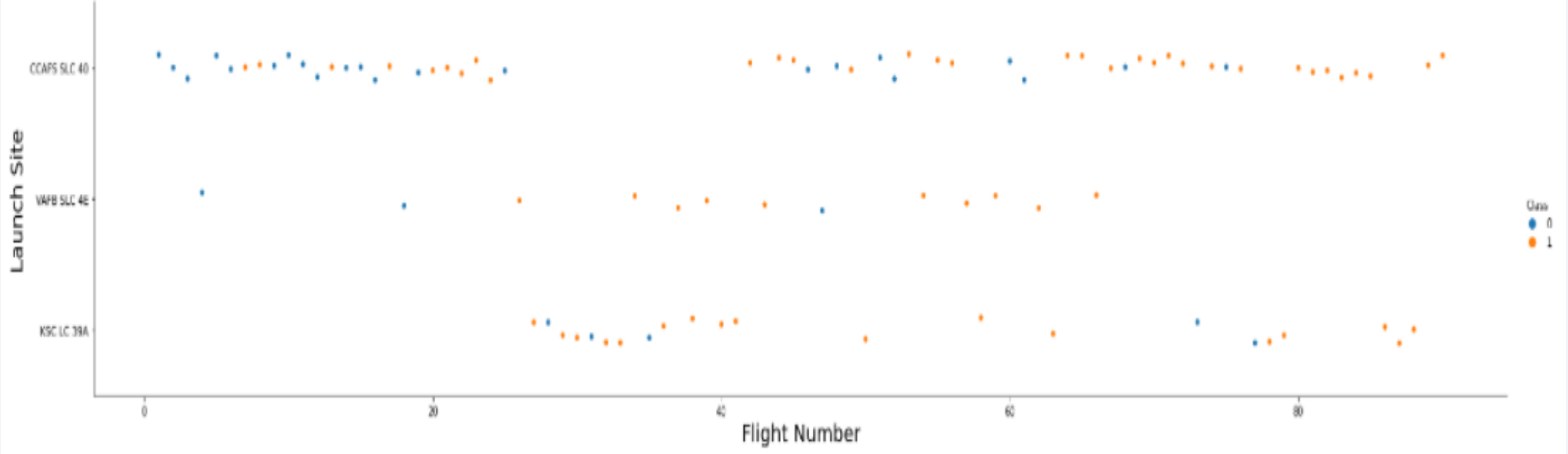
Flight Number vs. Launch Site

- Analysis of the plot reveals a positive correlation between the volume of flights at a launch site and its corresponding success rate.



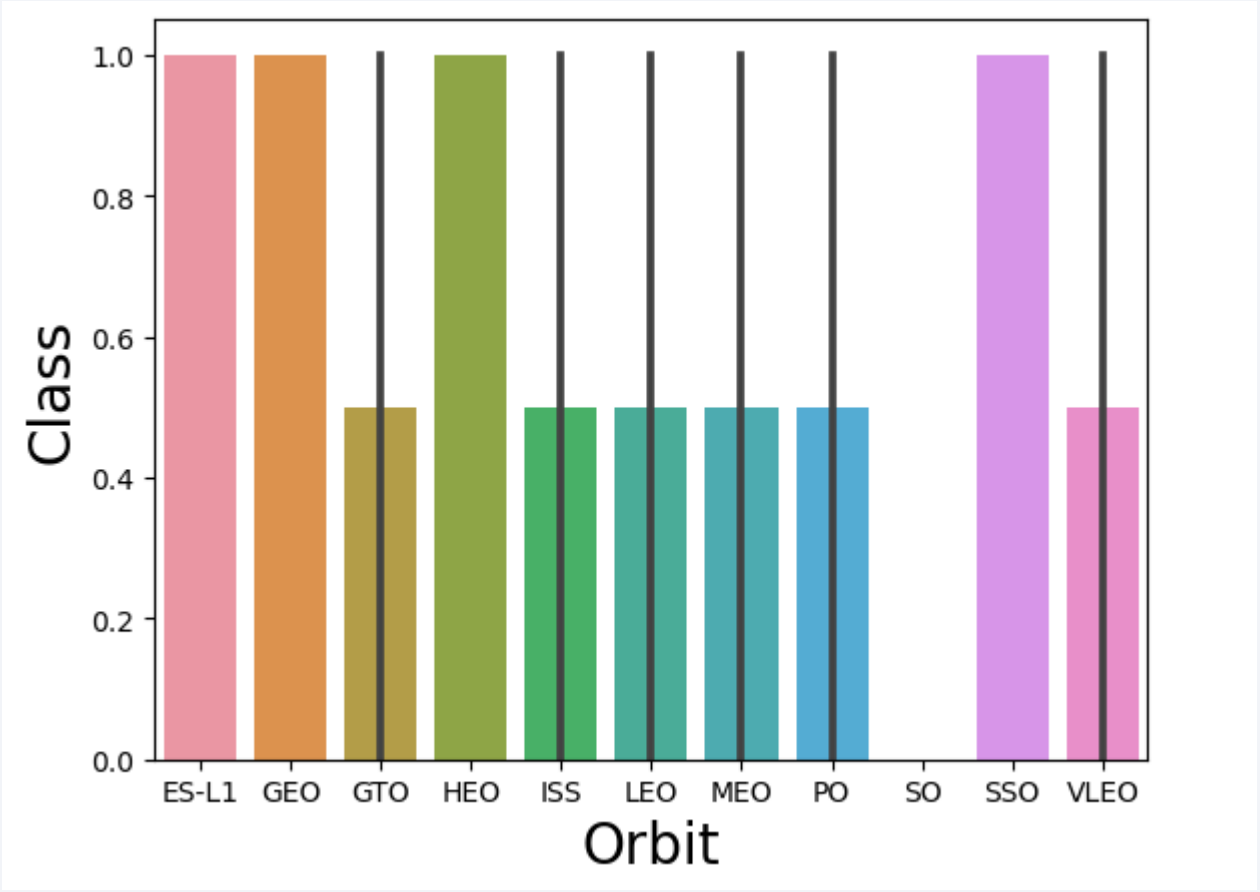
Payload vs. Launch Site

- The success rate of rockets launched from site CCAFS SLC 40 increases with greater payload mass.



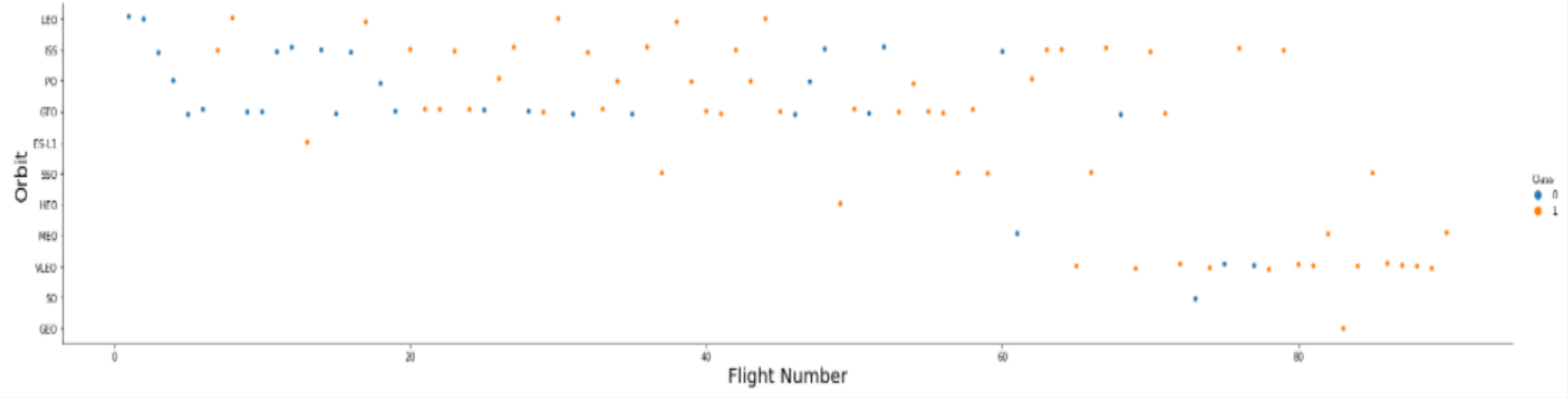
Success Rate vs. Orbit Type

- Based on the graph, it is evident that ES-L1, GEO, HEO, SSO, and VLEO exhibit the highest success rates.



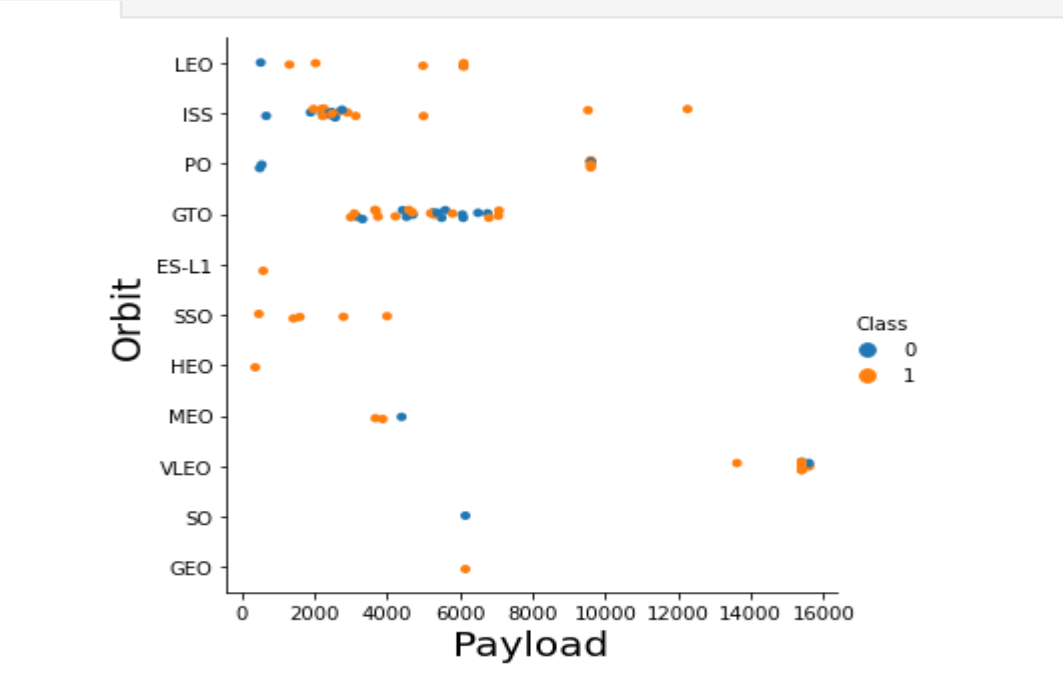
Flight Number vs. Orbit Type

- The chart below illustrates the relationship between Flight Number and Orbit Type. It becomes apparent that for missions targeting Low Earth Orbit (LEO), there is a correlation between the number of flights and mission success. However, for Geosynchronous Transfer Orbit (GTO) missions, the data suggests no discernible relationship between the number of flights and the likelihood of mission success.



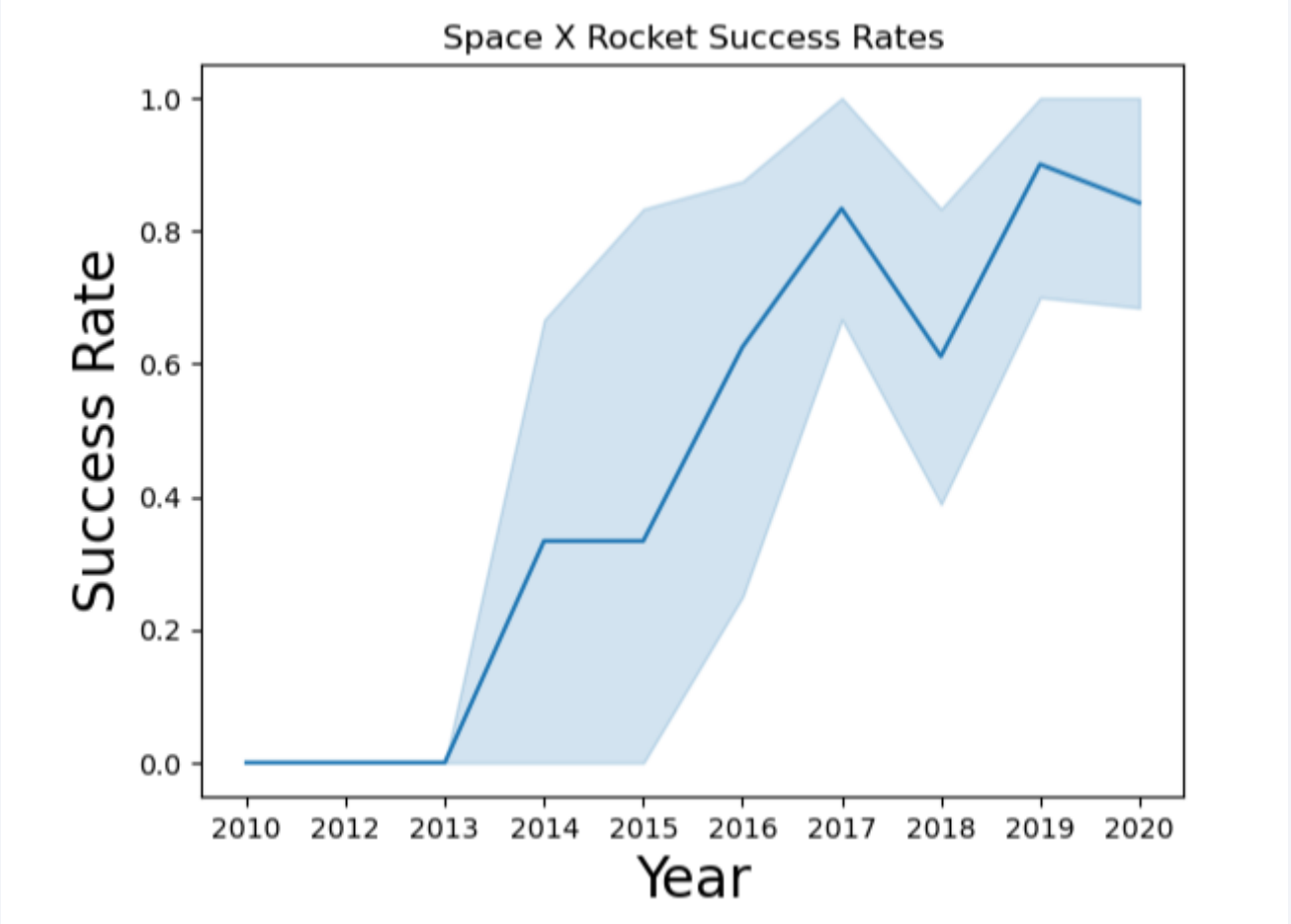
Payload vs. Orbit Type

- We observe that for heavy payloads, successful landings are more frequent in PO, LEO, and ISS orbits.



Launch Success Yearly Trend

- Based on the graph, it is evident that the success rate has consistently risen from 2013 through 2020.



All Launch Site Names

- We employed the 'DISTINCT' keyword to display only unique launch sites from the SpaceX dataset.

Task 1

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

In [10]:
```
%sql select distinct(LAUNCH_Site) from SPACEXTBL;
```

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

Out[10]:

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

Launch Site Names Begin with 'CCA'

Task 2

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

```
1]: %sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
 * sqlite:///my_data1.db
Done.
```

1]:

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

- We executed the aforementioned query to retrieve and display five records featuring launch sites whose names begin with 'CCA'.

Total Payload Mass

- We determined the total payload capacity of NASA's rocket boosters to be 45,596 units, as obtained through the following query:

Task 3

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

In [12]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)';
```

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

Out[12]:

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

Task 4

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

In [13]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION LIKE '%F9 v1.1';
```

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

Out[13]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- We noted that the first successful landing on a ground pad took place on December 22, 2015.

Task 5

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

Hint:Use min function

In [13]:
```
%sql select min(DATE) from SPACEXTBL where Landing_Outcome = 'Success (ground pad)';
```

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

Out[13]:

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We employed the WHERE clause to filter boosters that have successfully landed on drone ships. Additionally, we used the AND condition to further refine our search to include only those landings with a payload mass between 4,000 and 6,000 kilograms.

Task 6

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

In [14]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Out[14]:

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 employed the wildcard character '%' in our SQL query to filter records where the Mission_Outcome was either a success or a failure.

Task 7

List the total number of successful and failure mission outcomes

In [15]:
```
%sql select count(MISSION_OUTCOME) from SPACEXTBL where MISSION_OUTCOME = 'Success' or \
MISSION_OUTCOME = 'Failure (in flight)'
```

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

Out[15]:

count(MISSION_OUTCOME)
99

Boosters Carried Maximum Payload

- We identified the booster that has carried the maximum payload by utilizing a subquery within the WHERE clause along with the MAX() function.

Task 8

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

In [16]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

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

Out[16]:

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

2015 Launch Records

- We employed a combination of SQL conditions—specifically the WHERE clause, LIKE operator, AND, and BETWEEN— to filter for unsuccessful drone ship landings, corresponding booster versions, and launch site names for the year 2015.

Task 9

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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [17]:
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND \
LANDING_OUTCOME = 'Failure (drone ship)';
```

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

Out[17]:

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- We extracted the 'Landing Outcomes' and their respective counts from the dataset. To focus on a specific time frame, we employed the WHERE clause to filter records between March 20, 2010, and June 4, 2010. Subsequently, we used the GROUP BY clause to aggregate the data based on the landing outcomes. Finally, the ORDER BY clause was applied to sort the grouped landing outcomes in descending order.

Task 10

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.

```
%sql select Count(LANDING_OUTCOME) AS "Rank success count between 2010-06-04 and 2017-03-20" from SPACEXTBL \
where LANDING_OUTCOME like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by date desc
```

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

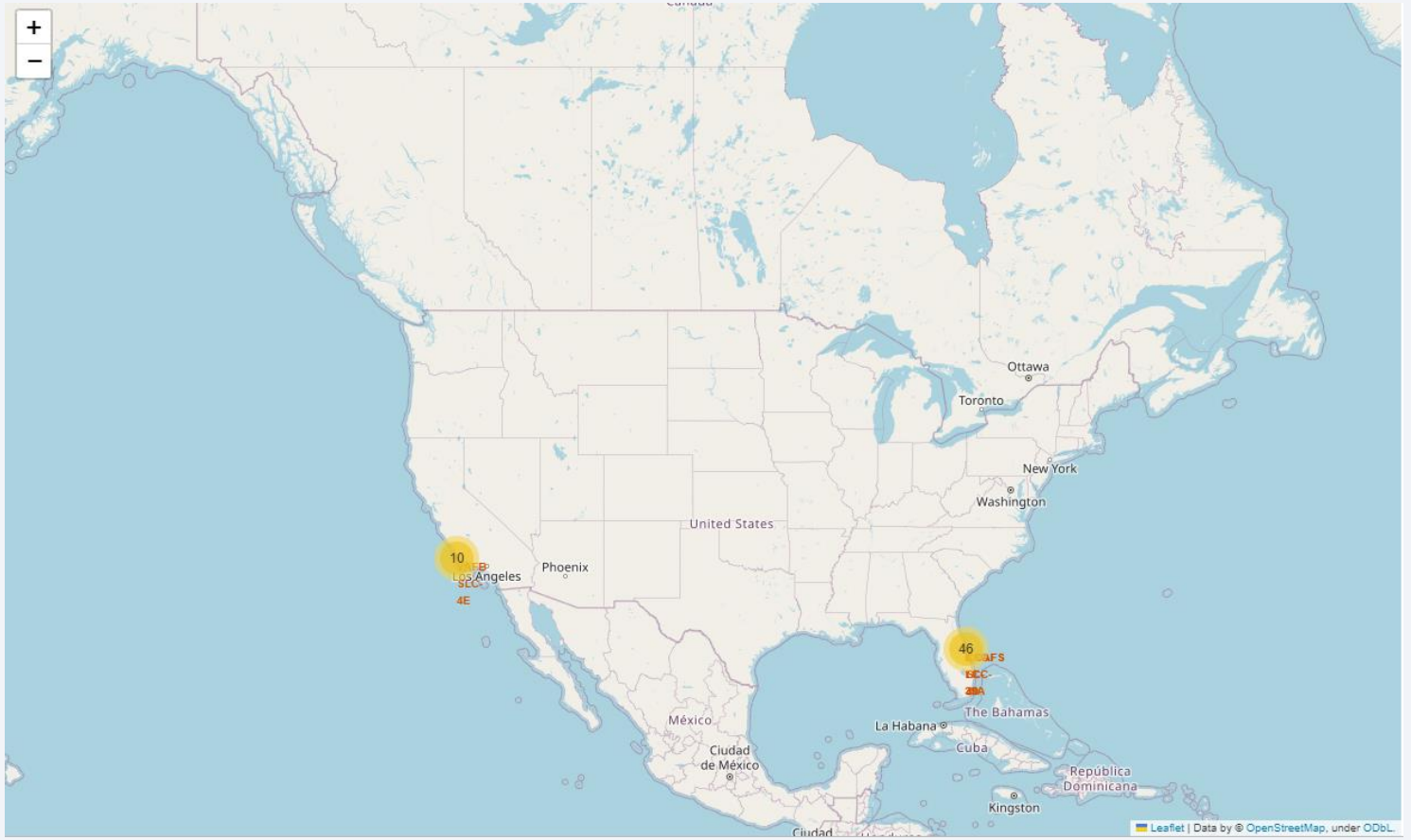
Rank success count between 2010-06-04 and 2017-03-20
10

Section 3

Launch Sites Proximities Analysis

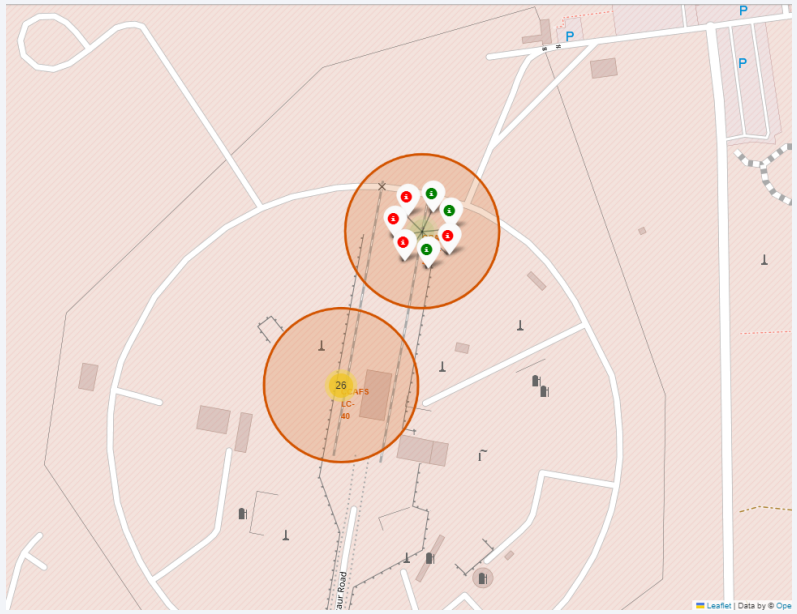
All Launch Sites in the United States

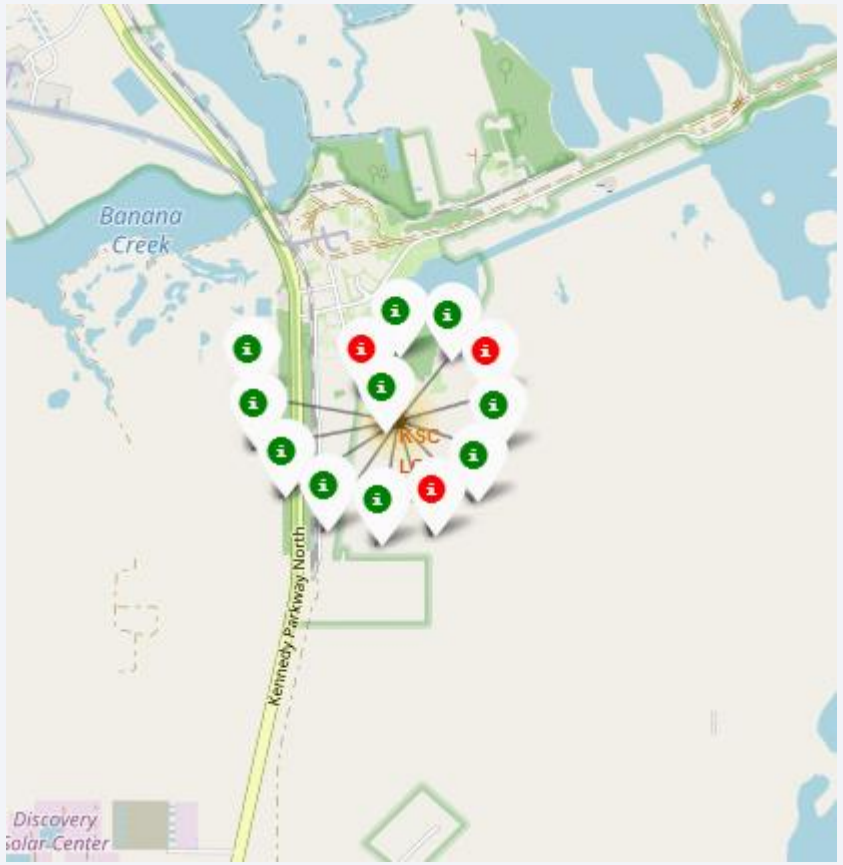
- According to the map, SpaceX's launch sites are situated at Cape Canaveral's Space Launch Complex 40 (SLC-40) in Florida, USA, as well as at Vandenberg Space Force Base in California, USA.

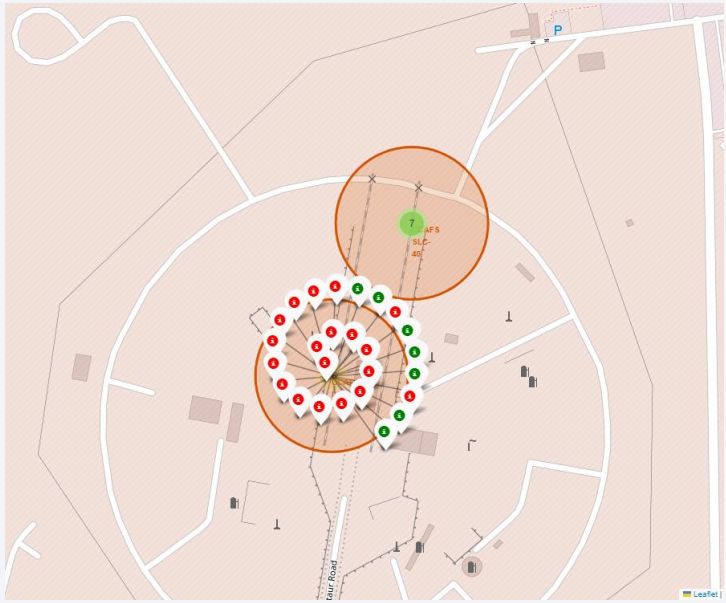


Color-coded markers indicating launch site locations

- The green markers indicates successful launches, while the red markers signifies failures.
- Florida Sites:



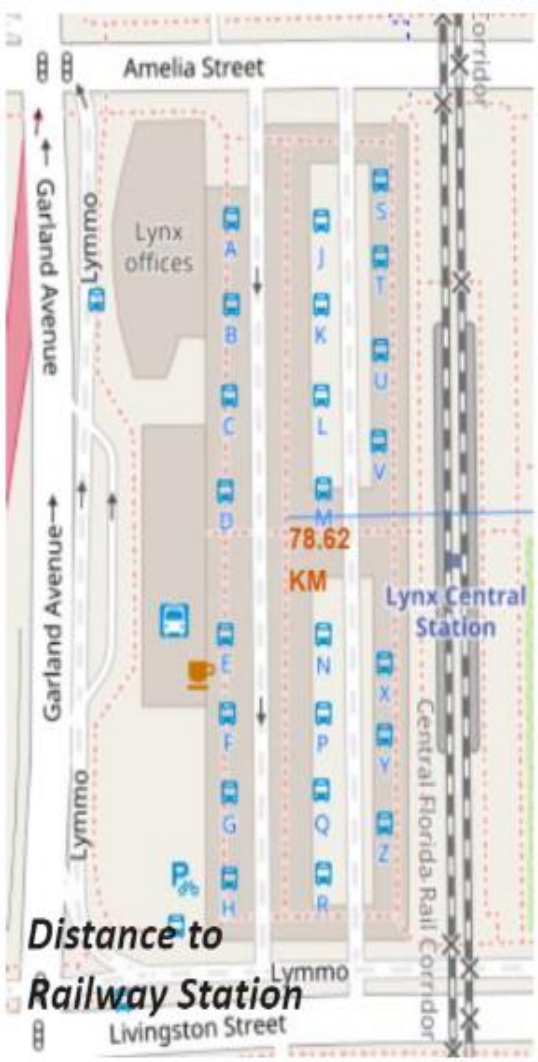




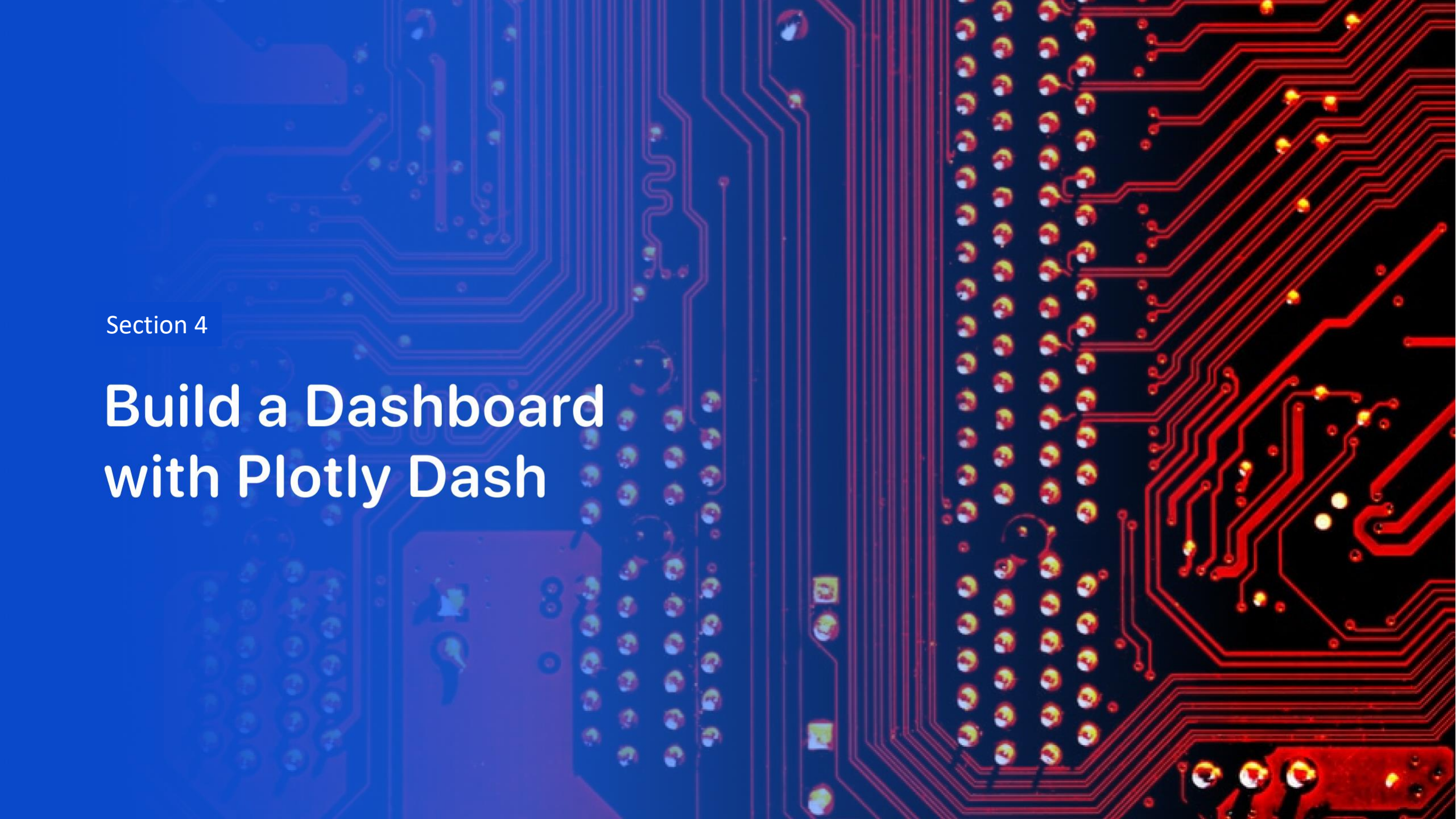
Color-coded markers indicating launch site locations

- The green markers indicates successful launches, while the red markers signifies failures.
- California Site:

Distance from Launch Site to Nearby Landmarks

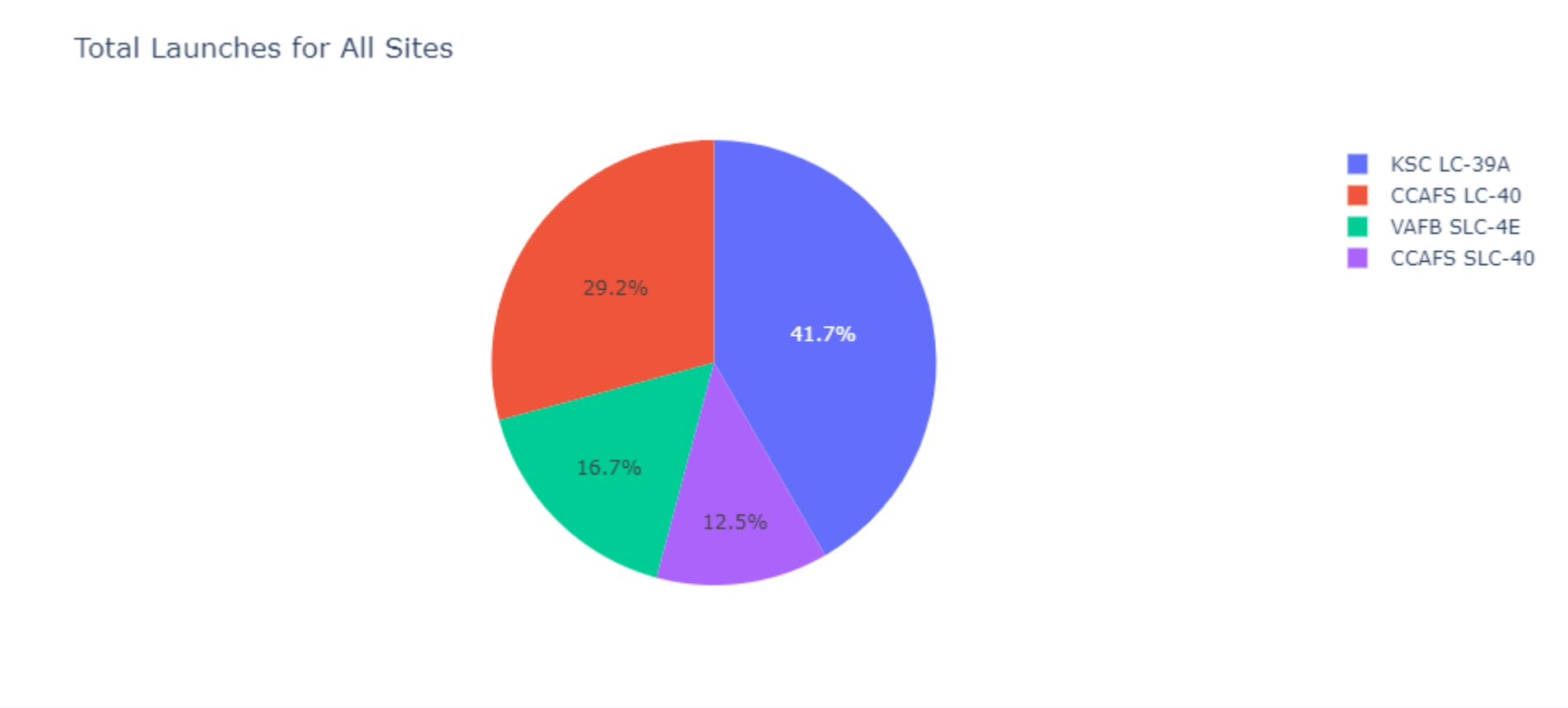


- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

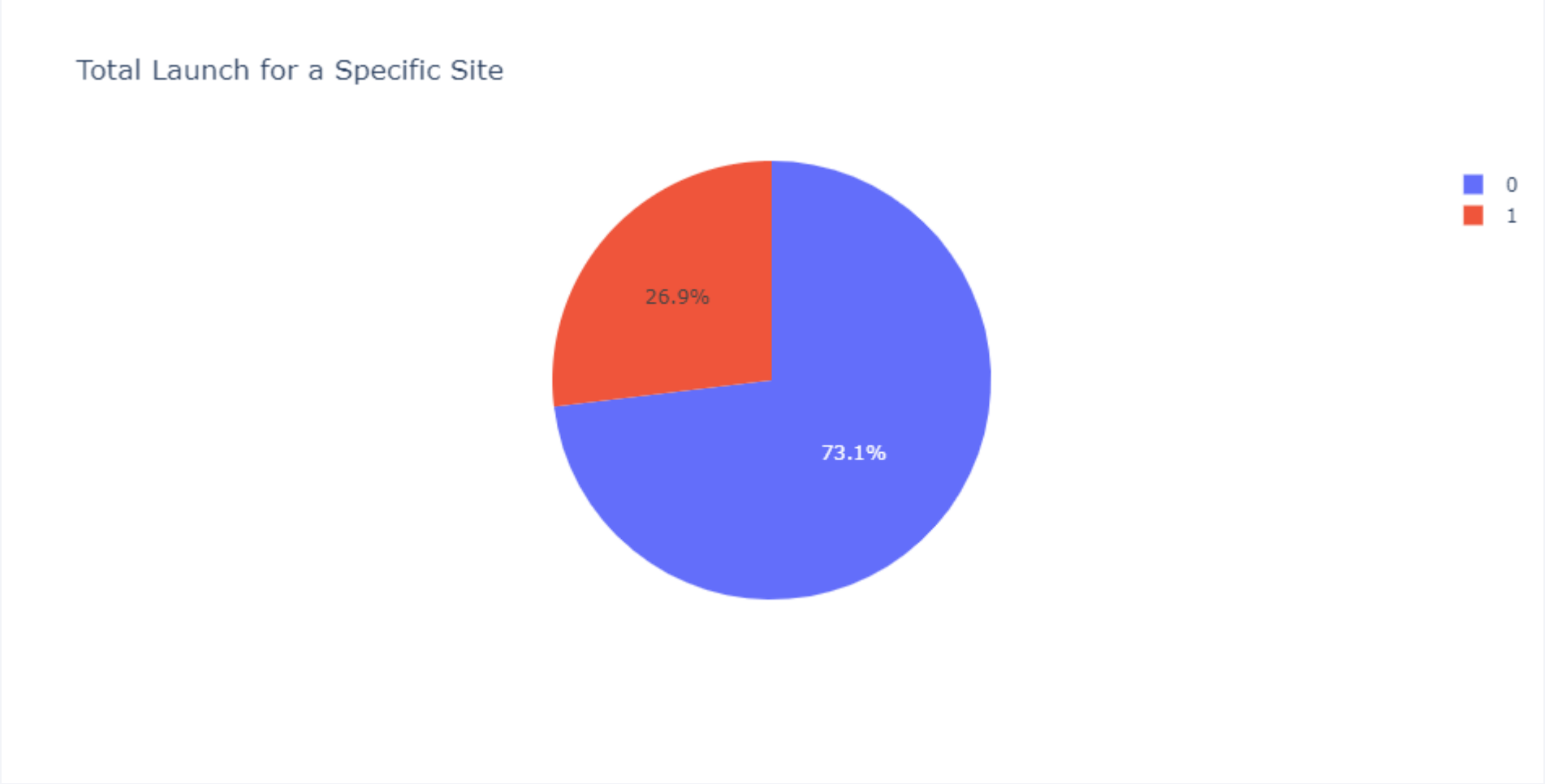
Pie Chart Illustrating the Success Rates of Various Launch Sites

- KSC LC-39A had the most successful launches.



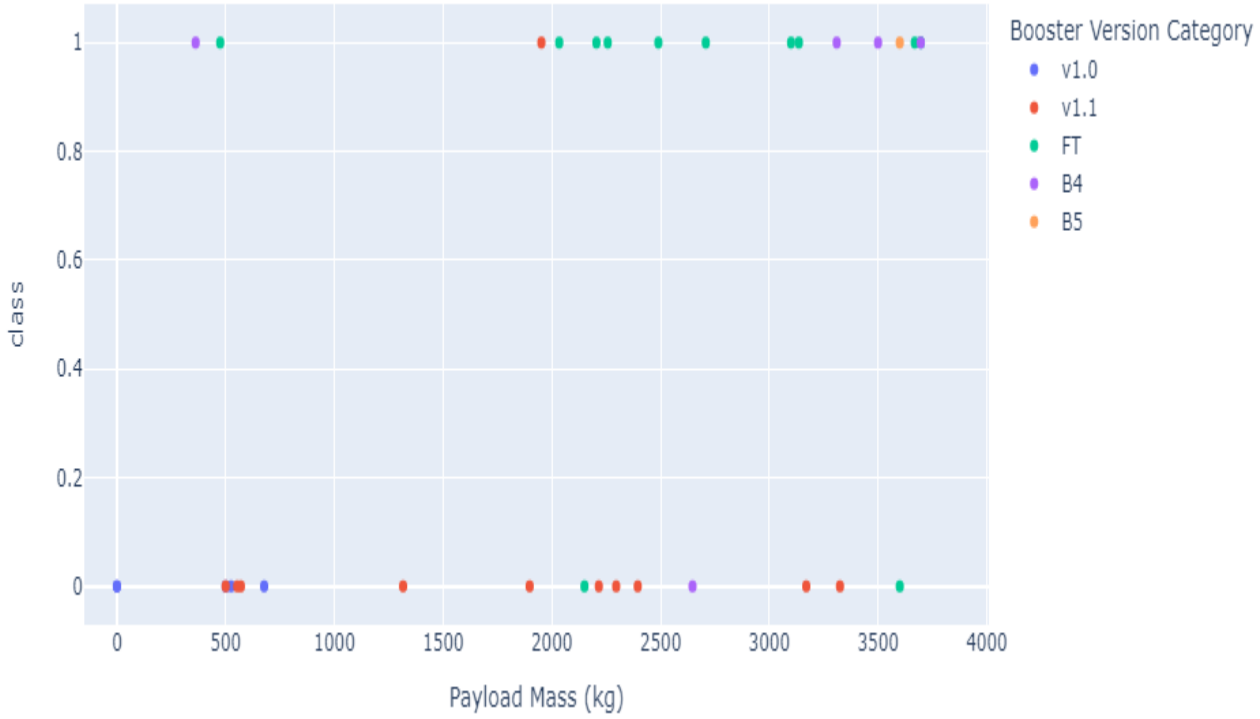
Pie Chart Illustrating the Launch Site with the Highest Success Rate for Rocket Launches

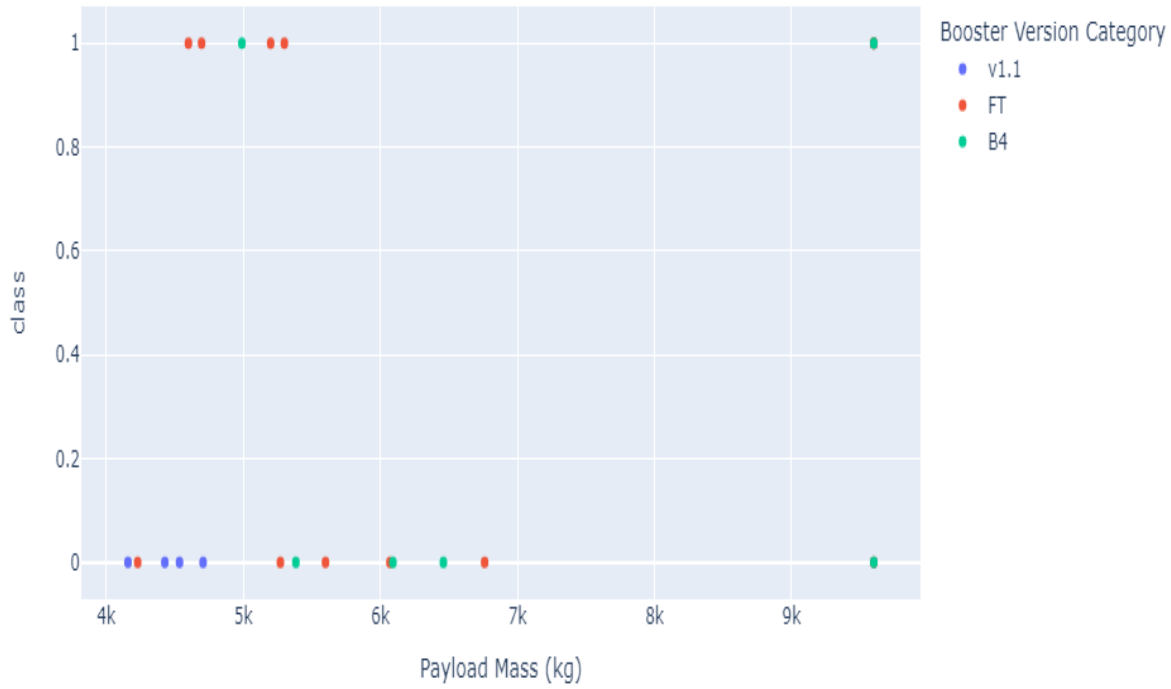
- KSC LC-39A had a 73.1% successful rate and a 26.9% failure rate.



Interactive Scatter Plot Comparing Payload Weight to Launch Outcome Across All Launch Sites, Featuring a Range Slider for Payload Selection

- The success rate for payloads with lower weight is higher compared to those with heavier weight.





Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier emerges as the model delivering the highest level of classification accuracy.

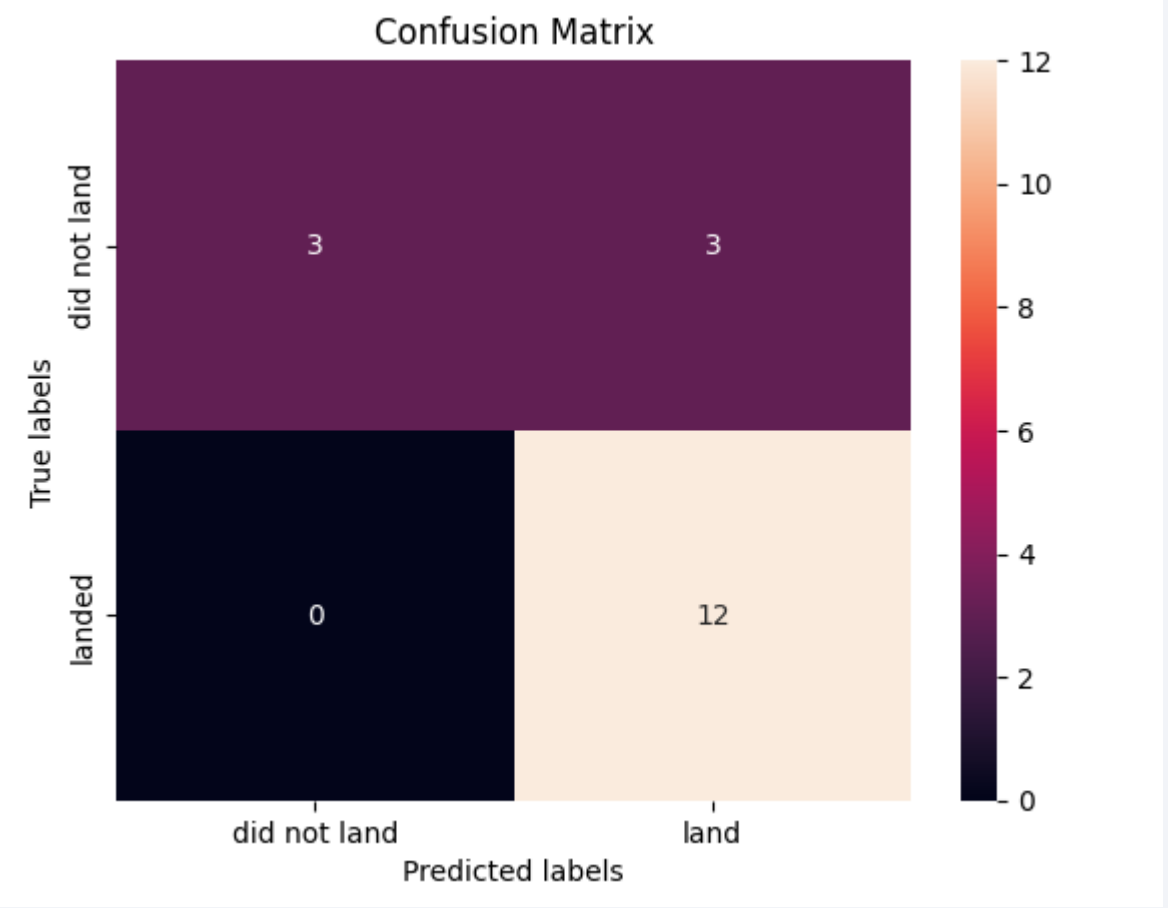
In [67]:

```
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 0.8892857142857142
Best Params is : {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split':
10, 'splitter': 'best'}
```

Confusion Matrix

- The confusion matrix for the Decision Tree classifier indicates that the model is generally capable of differentiating between various classes. However, a significant issue lies in its tendency to produce false positives—specifically, instances where the classifier incorrectly labels unsuccessful landings as successful.



Conclusions

We can conclude that:

- A higher number of flights at a launch site correlates with a greater launch success rate.
- The overall launch success rate has been on the rise from 2013 to 2020.
- The orbits ES-L1, GEO, HEO, SSO, and VLEO have the highest success rates.
- KSC LC-39A stands out as the launch site with the most successful launches.
- The Decision Tree Classifier is the most effective machine learning algorithm for predicting launch success.

Thank you!