IBM Developer
SKILLS NETWORK

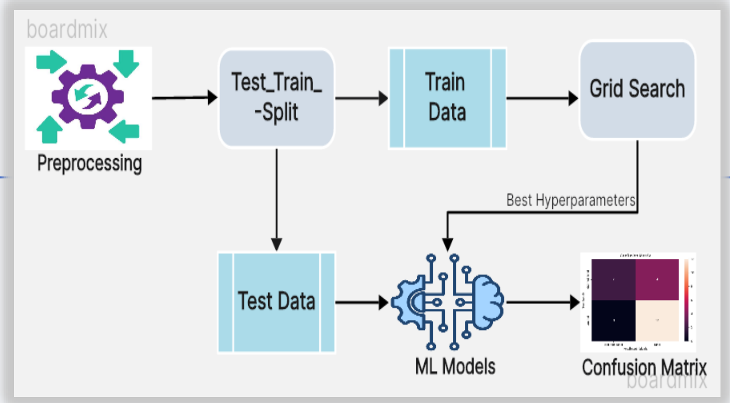
Winning Space Race with Data Science

Steshi
18 April 2024

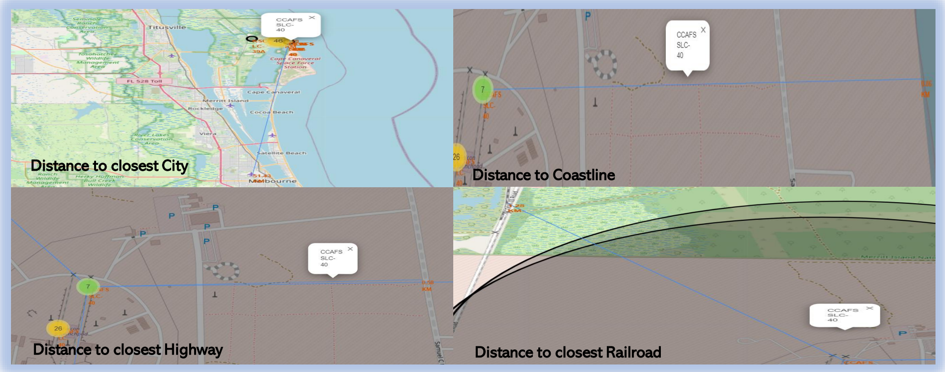
Outline

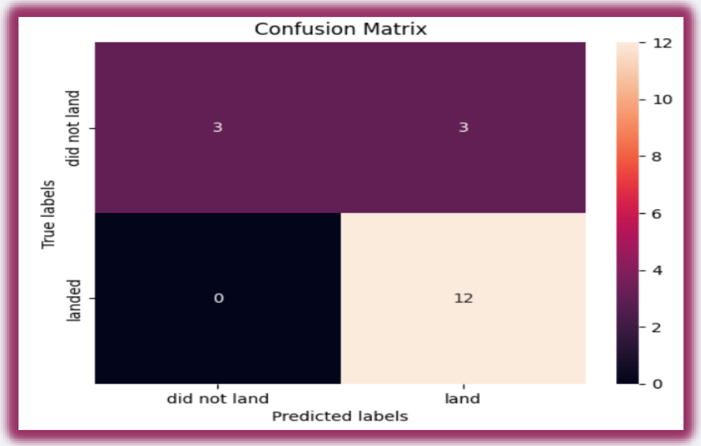
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

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

The commercial space age has arrived, with companies making space travel more accessible than ever. Among them, SpaceX stands out, promoting Falcon 9 rocket launches at a significantly lower cost than competitors, thanks to its innovative approach to reusing the first stage.

In this capstone, we embark on a journey to predict the success of Falcon 9 first stage landings using advanced Machine Learning models. We aim to provide valuable insights for cost estimation in the fiercely competitive space launch industry by:

- Utilizing advanced Machine Learning models to predict the success of Falcon 9 first stage landings.
- Identifying and analyzing key factors that influence the success of Falcon 9 first stage landings.
- Providing valuable insights into the space launch industry by understanding the relationship between various factors and landing success.

Section 1

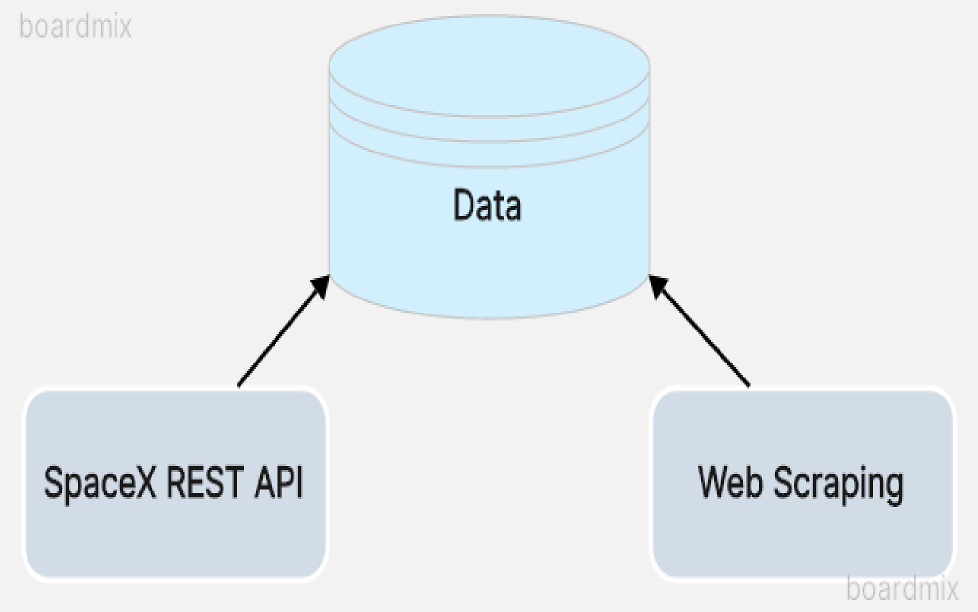
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data Collection has been performed using SpaceX REST API and Web Scraping related wiki pages.
- Perform data wrangling
  - Transitioned Landing Outcomes to classes 1(Successful landing) and 0(Unsuccessful Landing).
- Perform exploratory data analysis (EDA) :
  - Derived insights using visualization and SQL.
- Perform interactive visual analytics :
  - Used Folium and Plotly Dash to perform visual analysis.
- Perform predictive analysis using classification models
  - Used models like SVM , Logistic Regression , KNN , etc. along with GridSearchCV to find the best parameters for creating the models.

Data Collection

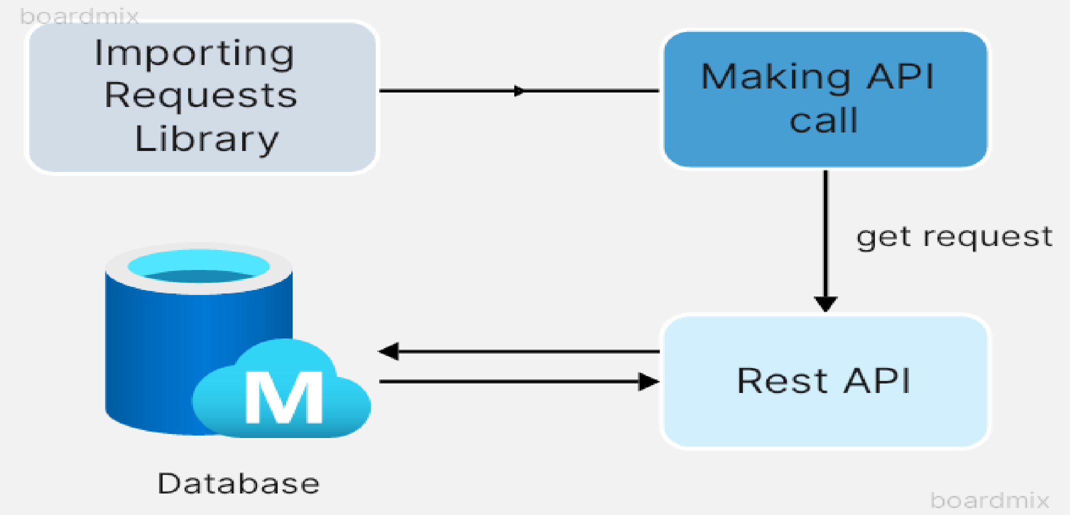


- In the data collection methodology, two robust sources are integrated: the SpaceX API and Web Scraping techniques.
- Through the SpaceX API, structured data is accessed directly from SpaceX's servers, ensuring reliability and accuracy.
- Additionally, web scraping is employed to extract valuable insights from relevant web pages, enriching the dataset with comprehensive information.

Data Collection – SpaceX API

In this section, we'll utilize a targeted API endpoint to obtain historical launch data vital for our analysis.The following are the key steps:

- **Targeted API Endpoint:** Utilizing a specific URL to access the desired data endpoint.
- **GET Request Execution:** Employing the requests library to execute a GET request, ensuring seamless data retrieval.
- **JSON Transformation:** Converting the retrieved response into a JSON format, facilitating easy interpretation and manipulation.



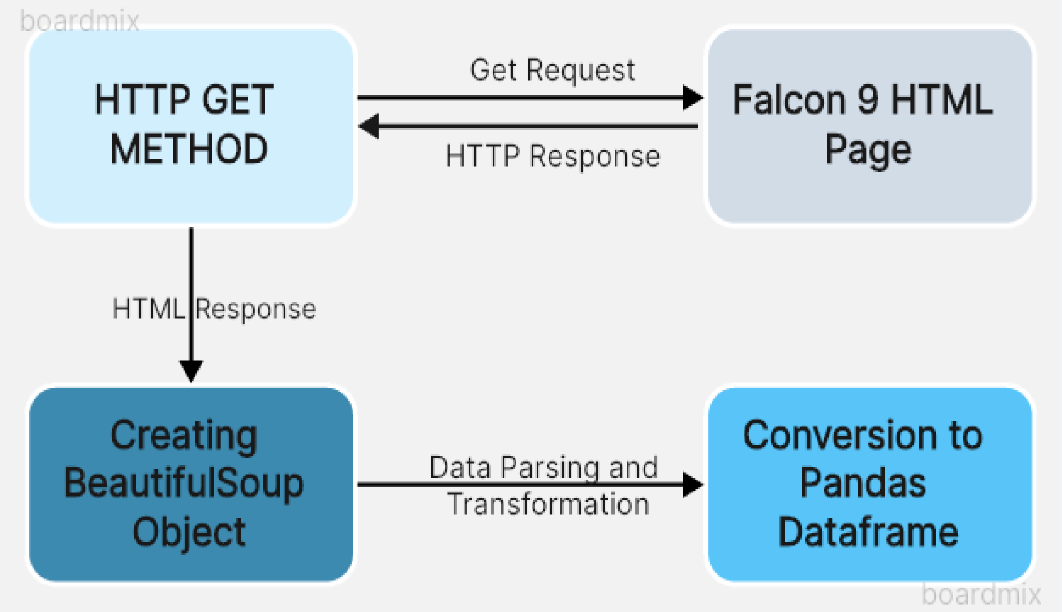
GitHub URL of the related notebook:

https://github.com/Steshitaya28/SpaceX-Data-Collection-using-API

Data Collection - Scraping

In this segment, we'll employ Python's BeautifulSoup library to scrape HTML tables containing Falcon 9 launch records.Following are the key objectives:

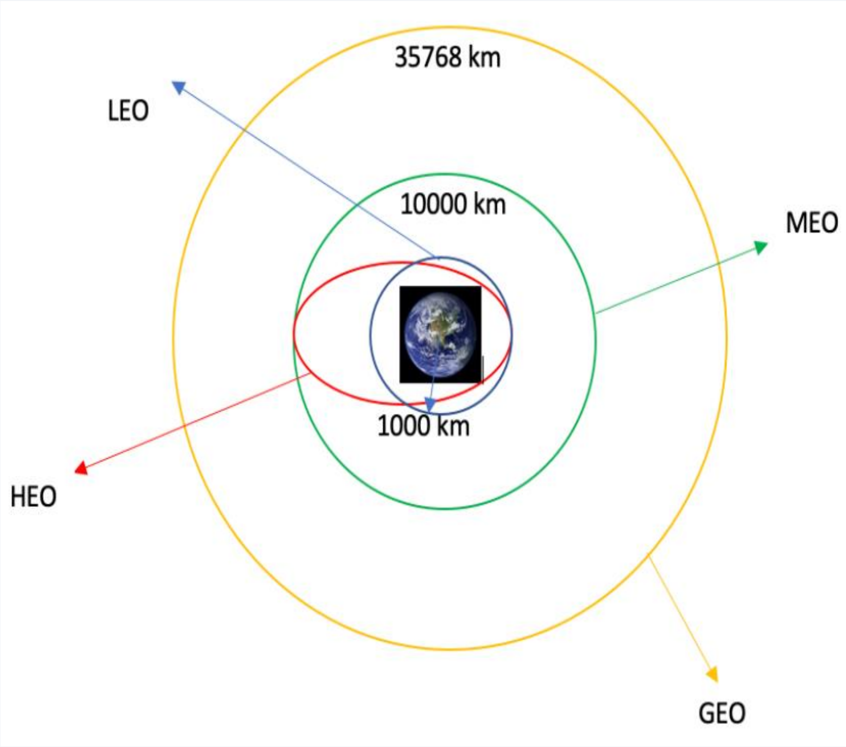
- **Web Scraping with BeautifulSoup:** Leveraging BeautifulSoup to extract data from HTML tables containing Falcon 9 launch records.
- **Data Parsing and Transformation:** Parsing the scraped data and converting it into a structured format suitable for further analysis.


HTTP GET METHOD
Get Request
HTTP Response
Falcon 9 HTML Page
HTML Response
Creating BeautifulSoup Object
Data Parsing and Transformation
Conversion to Pandas Dataframe
boardmix
boardmix

- **Creating a Clean Dataset:** Converting the parsed data into a Pandas DataFrame.

GitHub URL of the related notebook:
https://github.com/Steshitaya28/SpaceX-Applied-Data-Science-Capstone/blob/main/SpaceX-Data-Collection-using-WebScraping.ipynb

Data Wrangling



During the data wrangling phase, we performed the following tasks:

- Checked the dataset for null values to ensure data integrity.
- Verified the correct data types for all features in the dataset.
- Calculated the frequency of launches on each site to understand launch distribution.
- Determined the occurrences of orbits to gain insights into mission destinations.
- Converted various landing outcomes into binary values, assigning 0 for unsuccessful landings and 1 for successful ones, thereby creating a classification variable.

GitHub URL of the related notebook:

https://github.com/Steshitaya28/SpaceX-Applied-Data-Science-Capstone/blob/main/Data-Wrangling-of-SpaceX-Dataset.ipynb

EDA with Data Visualization

During our exploratory data analysis, enhanced by visualizations, we undertook the following tasks:

- **Scatter Plots**: Examined relationships between flight number, payload, launch site, and orbit using scatter plots to assess their impact on success rate.
- **Bar Chart:** Created a bar chart to compare success rates for each orbit, aiming to uncover associations.
- **Line Chart:** Plotted a line chart showing success rates over the years, offering insight into their evolution.
- **One-Hot Encoding:** Implemented one-hot encoding on features, converting categorical variables into numerical representations for analysis.

GitHub URL of the related notebook:
https://github.com/Steshitaya28/EDA-of-SpaceX-Dataset-with-Visualization

EDA with SQL

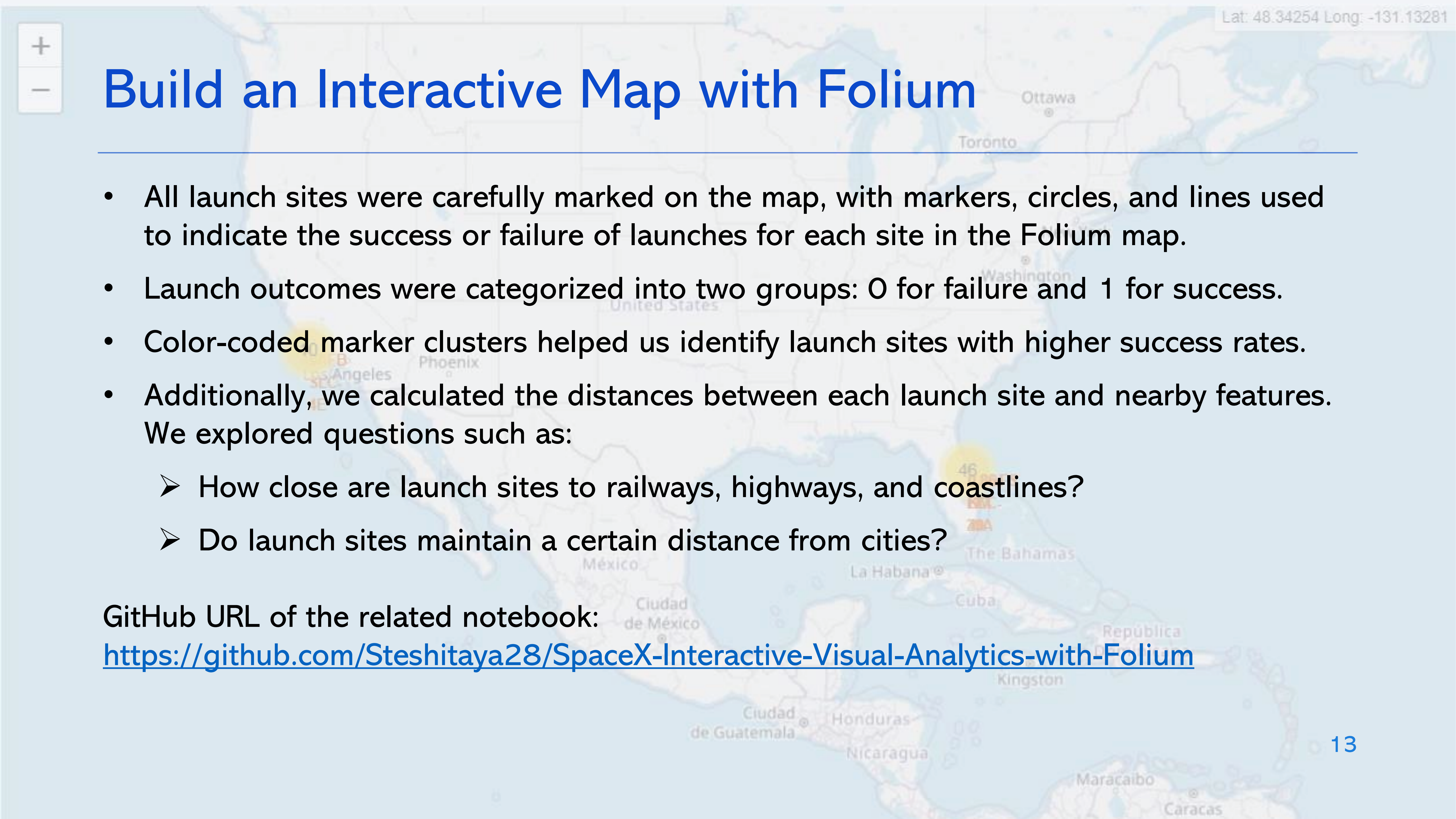
The SpaceX dataset was imported into a PostgreSQL database directly within the Jupyter Notebook environment.

Exploratory data analysis (EDA) was then carried out using SQL to extract insights from the dataset. Queries were formulated to uncover:

- The names of different launch sites involved in the space mission.
- The total payload mass carried by boosters launched by NASA's Commercial Resupply Services (CRS).
- The average payload mass carried by booster version F9 v1.1.
- The total number of successful and failed mission outcomes.
- Details on failed landing outcomes on drone ships, including their booster version and launch site names.

GitHub URL of the related notebook:

https://github.com/Steshitaya28/EDA-of-SpaceX-Dataset-using-SQL



Build an Interactive Map with Folium

- All launch sites were carefully marked on the map, with markers, circles, and lines used to indicate the success or failure of launches for each site in the Folium map.
- Launch outcomes were categorized into two groups: 0 for failure and 1 for success.
- Color-coded marker clusters helped us identify launch sites with higher success rates.
- Additionally, we calculated the distances between each launch site and nearby features. We explored questions such as:
  - ➢ How close are launch sites to railways, highways, and coastlines?
  - ➢ Do launch sites maintain a certain distance from cities?

GitHub URL of the related notebook:
https://github.com/Steshitaya28/SpaceX-Interactive-Visual-Analytics-with-Folium

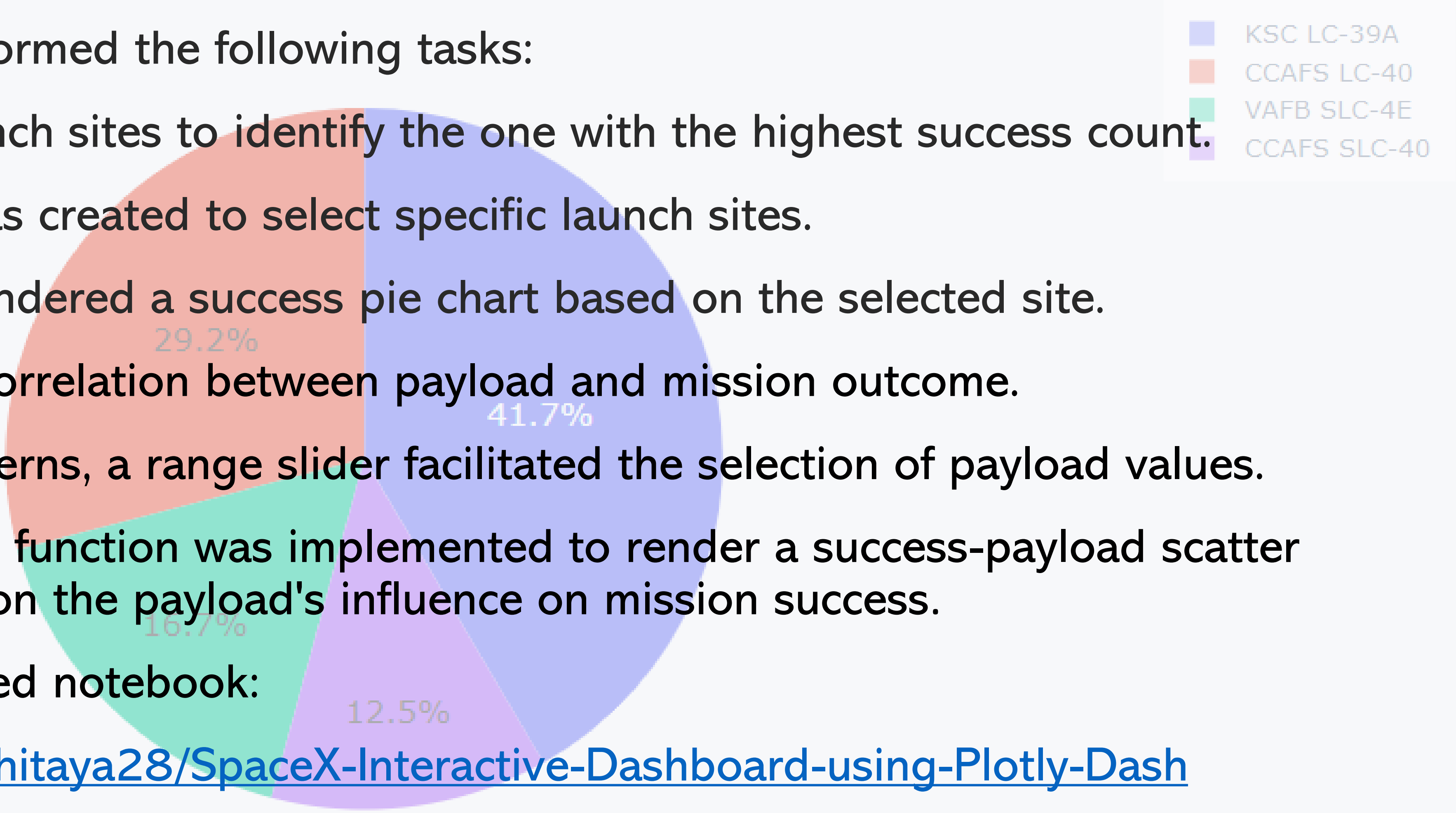
Build a Dashboard with Plotly Dash

In this section , we performed the following tasks:

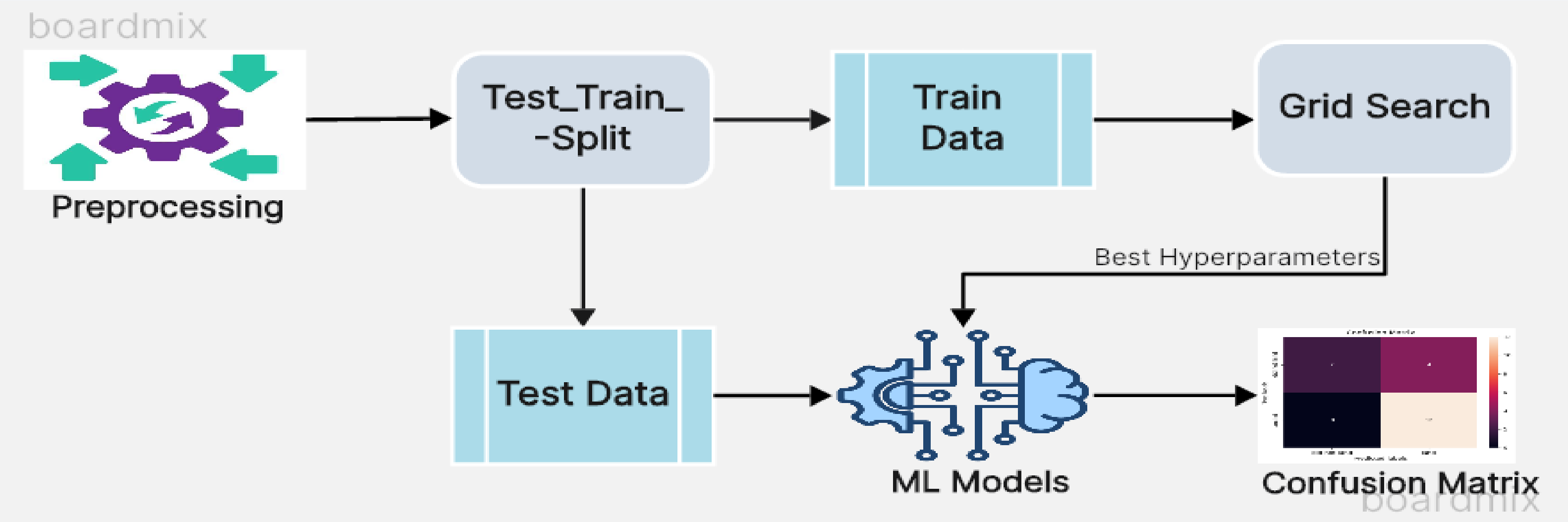
- We assessed four launch sites to identify the one with the highest success count.
- A dropdown menu was created to select specific launch sites.
- A callback function rendered a success pie chart based on the selected site.
- We investigated the correlation between payload and mission outcome.
- To analyze visual patterns, a range slider facilitated the selection of payload values.
- An additional callback function was implemented to render a success-payload scatter chart, shedding light on the payload's influence on mission success.

GitHub URL of the related notebook:

https://github.com/Steshitaya28/SpaceX-Interactive-Dashboard-using-Plotly-Dash



Predictive Analysis (Classification)

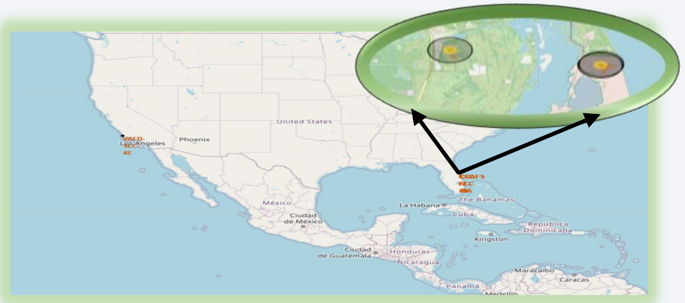


- Data preprocessing standardized the data.
- Train_test_split divided the data into training and testing sets.
- We trained the models (Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest Neighbors) and conducted Grid Search to find optimal hyperparameters for each algorithm.
- Using the best hyperparameter values, we determined the model with the highest accuracy on the training data.
- Finally, we generated and analyzed the confusion matrix.

GitHub URL of the related notebook:
https://github.com/Steshitaya28/SpaceX-Applied-Data-Science-Capstone/blob/main/SpaceX-Machine-Learning-Prediction.ipynb

Results

- **Exploratory data analysis results:**
  - Heavy Payloads are associated with higher frequency of successful landings.
  - Orbits ES-L1, GEO, HEO and SSO demonstrate the highest success rates.
  - There is a Positive correlation between the number of flights conducted at a launch site and the success rate of launches at that site.
  - The success rate has been on a steady rise since 2013, with the exception of a noticeable drop in 2018.
  - The first successful landing on a ground pad occurred on December 22, 2015.
- **Interactive analytics results and demo:**
  - The launch sites are located along the coasts of the United States, specifically in California and Florida.
  - The launch site KSC LC-39A has the highest number of successful launches, while launch site CCAFS SLC-40 has the fewest successful launches.

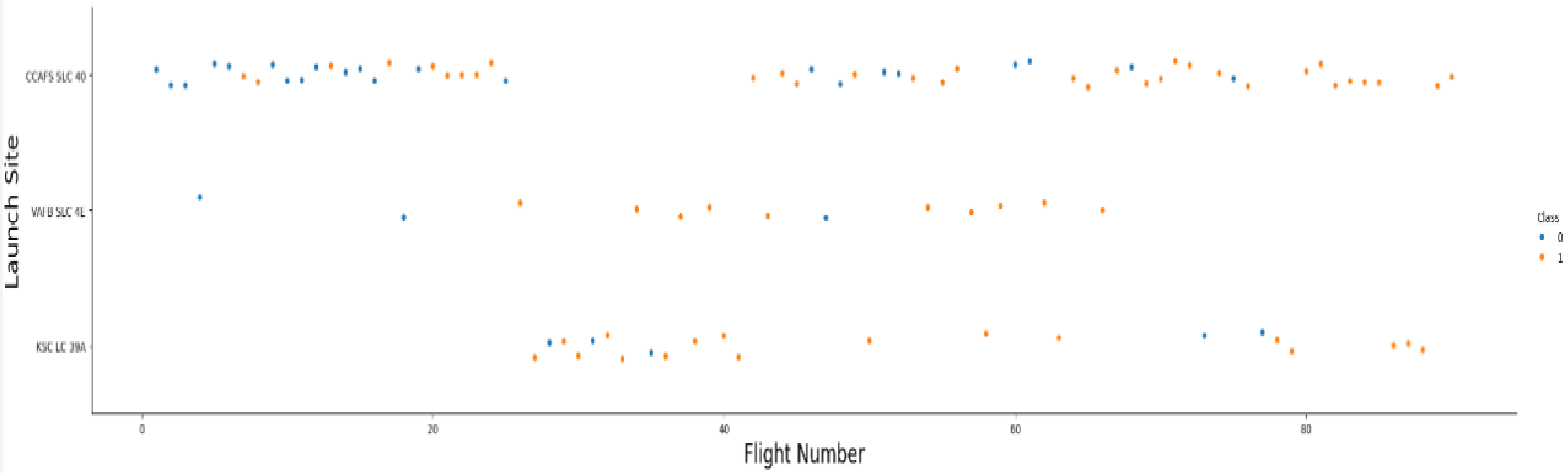


- **Predictive analysis results:**
  - The logistic regression model exhibits the highest accuracy (0.833), while the k-nearest neighbors (KNN) model demonstrates the lowest accuracy(0.611).

Section 2

Insights drawn from EDA

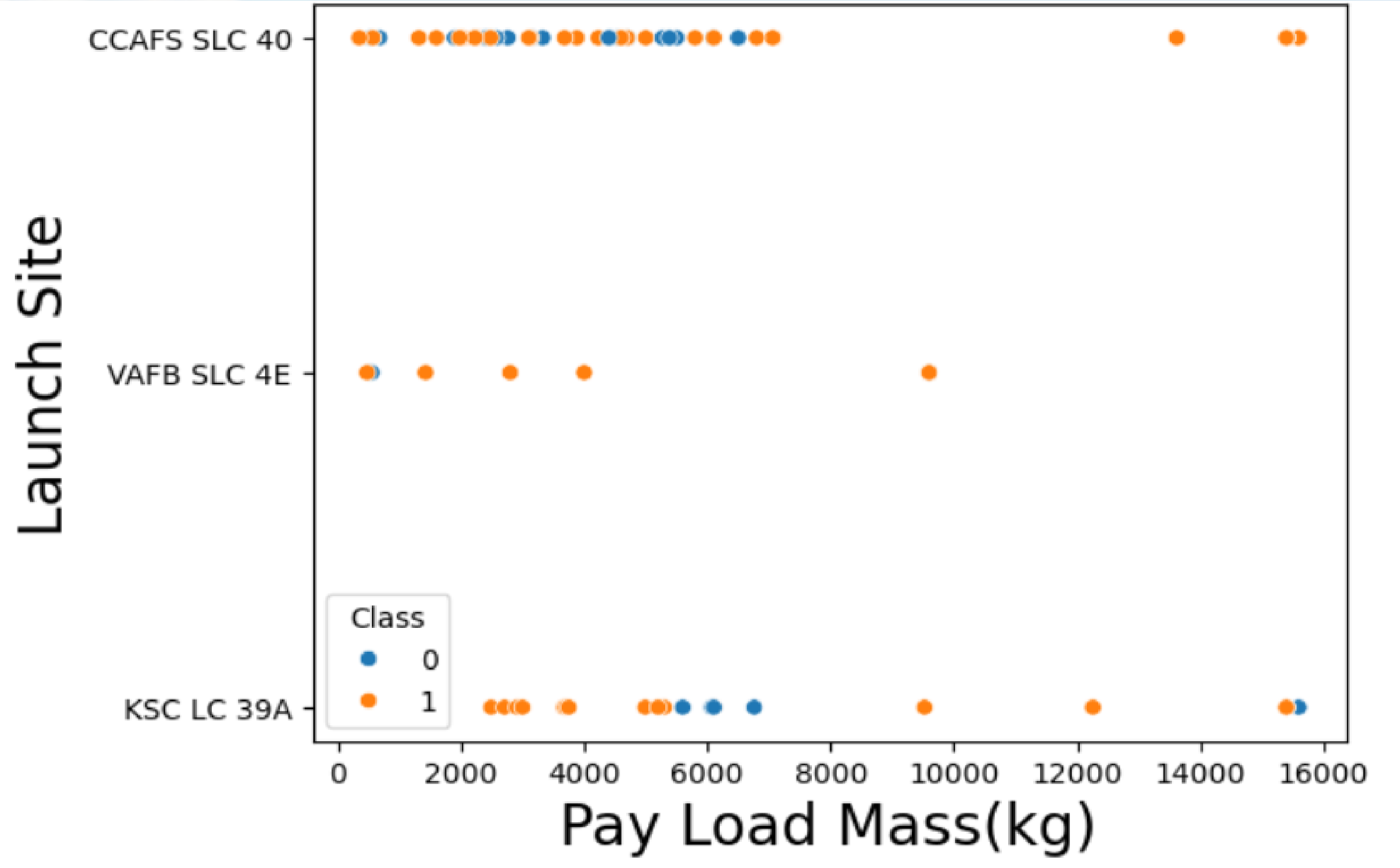
Flight Number vs. Launch Site



Based on the plot analysis, it appears that there is a positive correlation between the number of flights conducted at a launch site and the success rate of launches at that site.

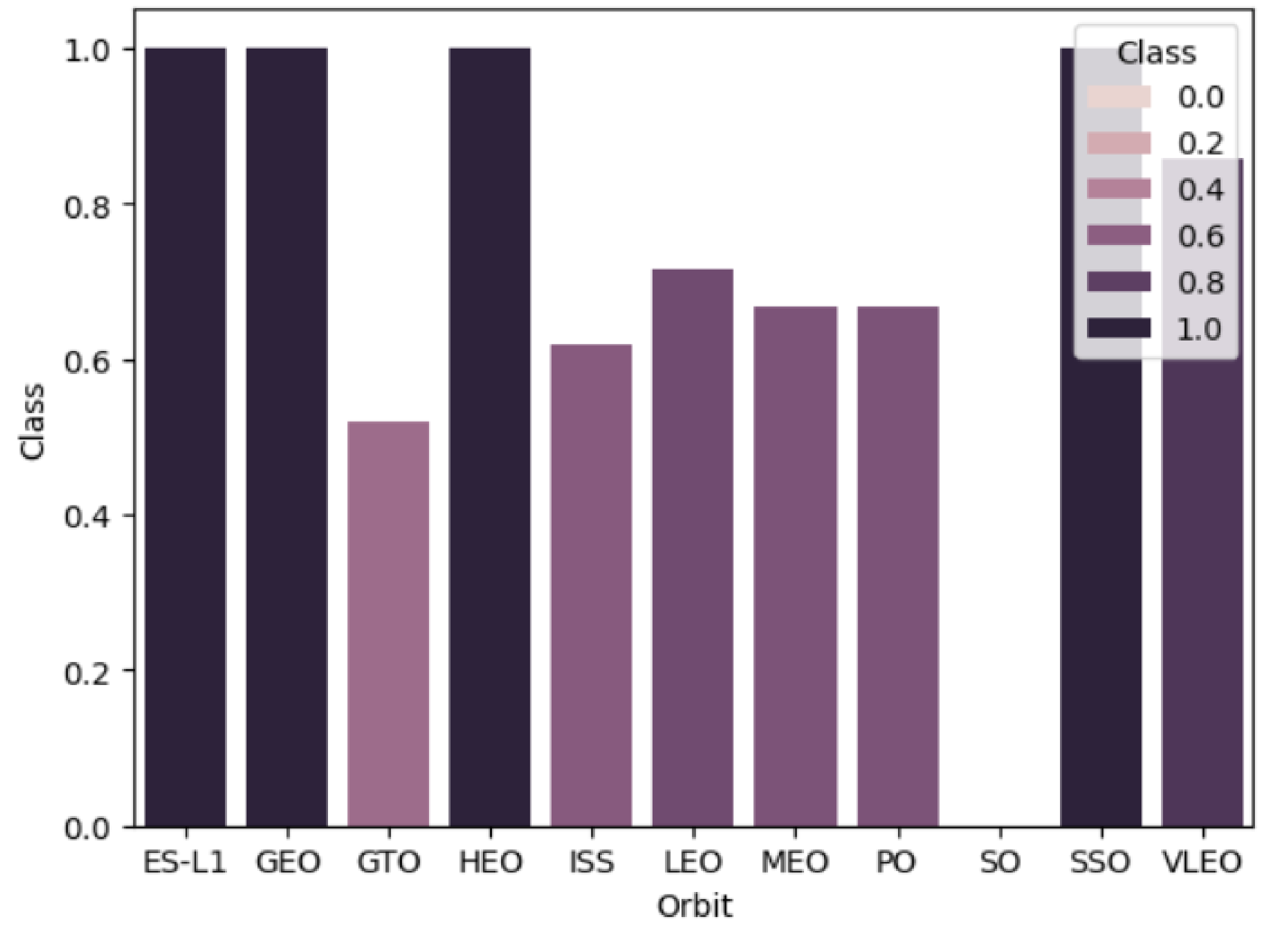
Payload vs. Launch Site

Upon examining the plot, it becomes evident that as the payload mass increases, so does the likelihood of a successful launch from the CCAFS SLC 40 site.



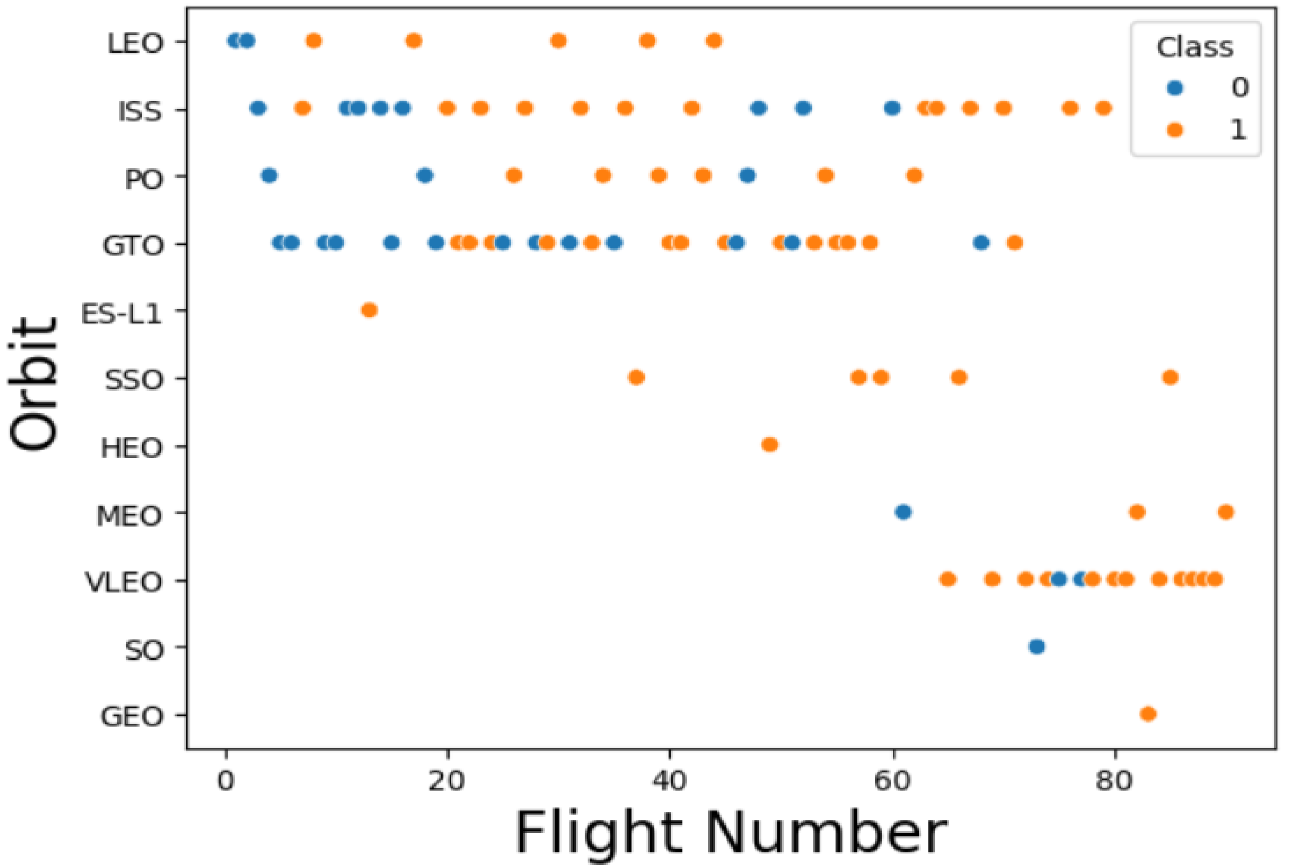
Success Rate vs. Orbit Type

The data clearly indicates that orbits ES-L1, GEO, HEO and SSO demonstrate the highest success rates, while the GTO orbit exhibits the lowest success rate among them.



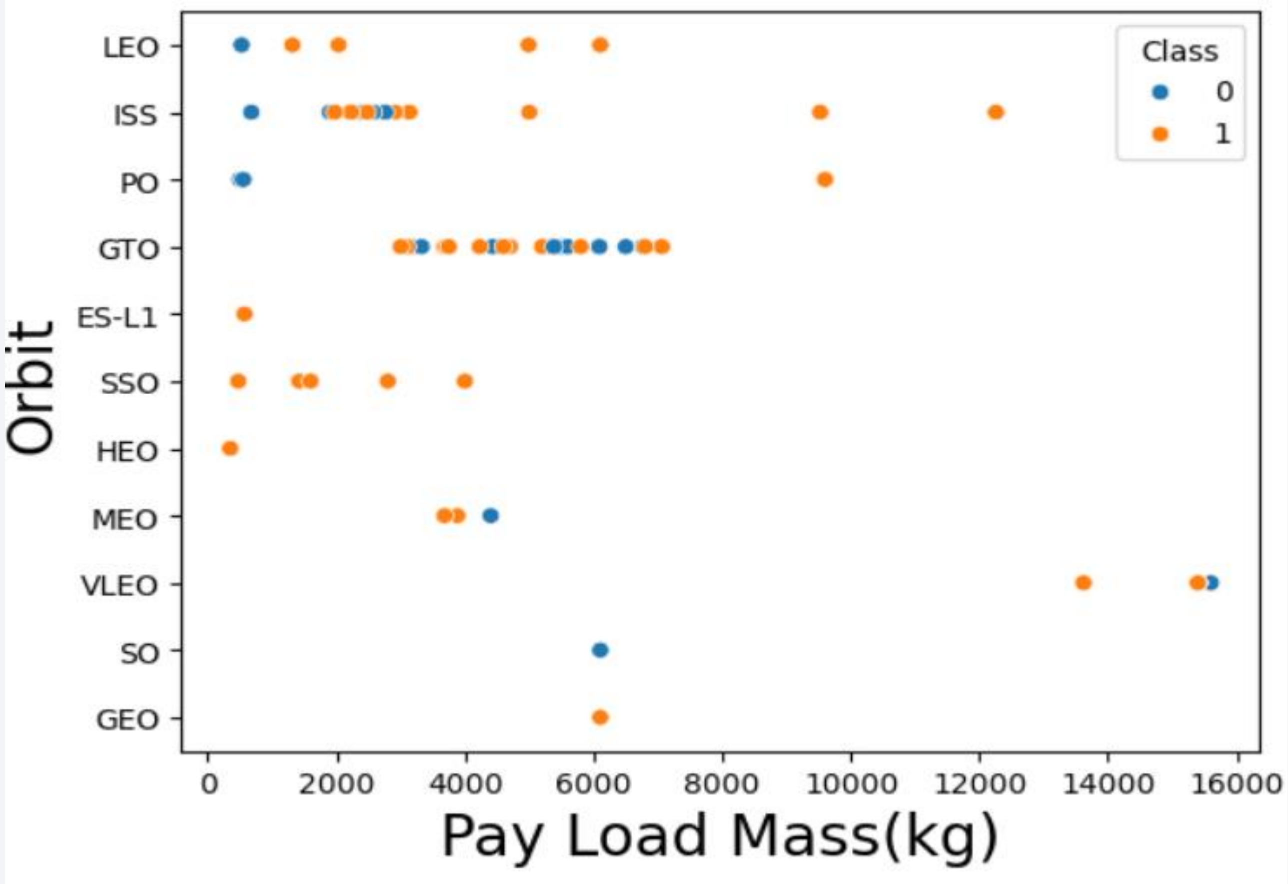
Flight Number vs. Orbit Type

The success rate of launches into GTO orbit appears to be unrelated to the flight number. In contrast, for orbits such as LEO and MEO, there is a noticeable trend of increasing success rates with higher flight numbers



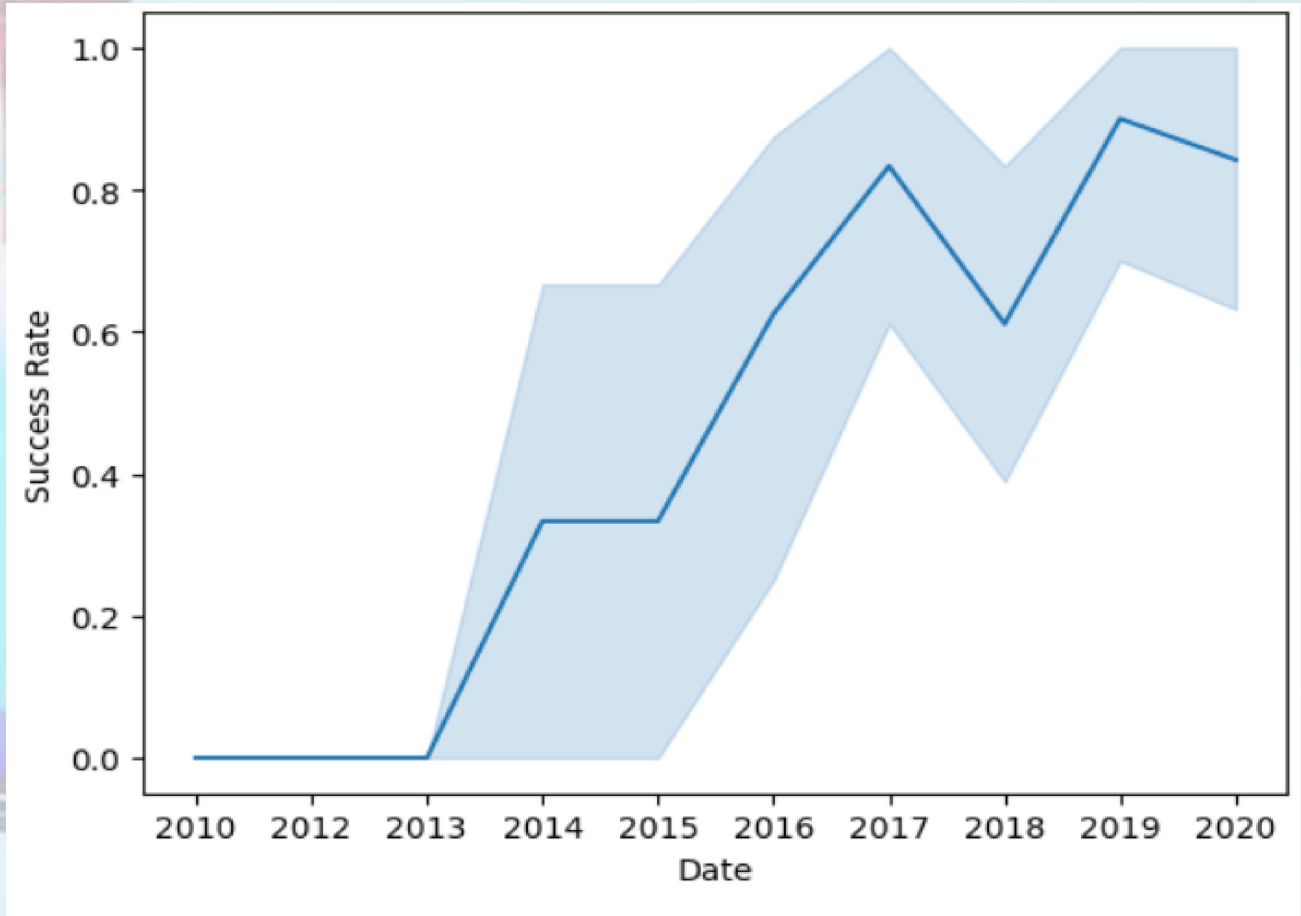
Payload vs. Orbit Type

It's evident from our observation that heavier payloads are associated with a higher frequency of successful landings in PO, LEO, and ISS orbits.



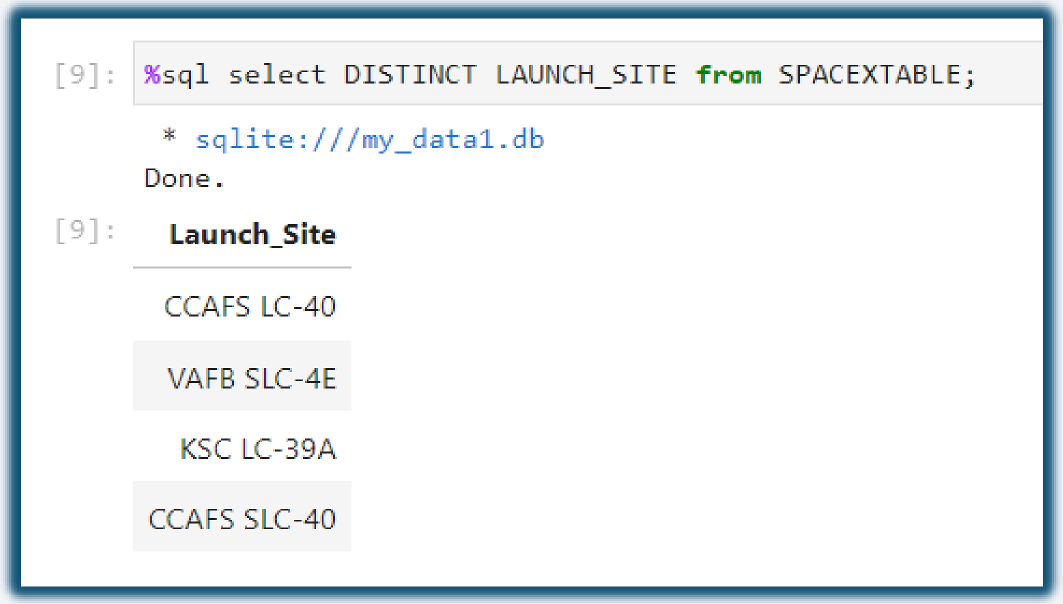
Launch Success Yearly Trend

From the data depicted in the plot, it's apparent that the success rate has been on a steady rise since 2013, with the exception of a noticeable drop in 2018.



All Launch Site Names

We used the DISTINCT keyword to eliminate duplicate entries and focus solely on distinct launch site names stored within the SPACEXTABLE.



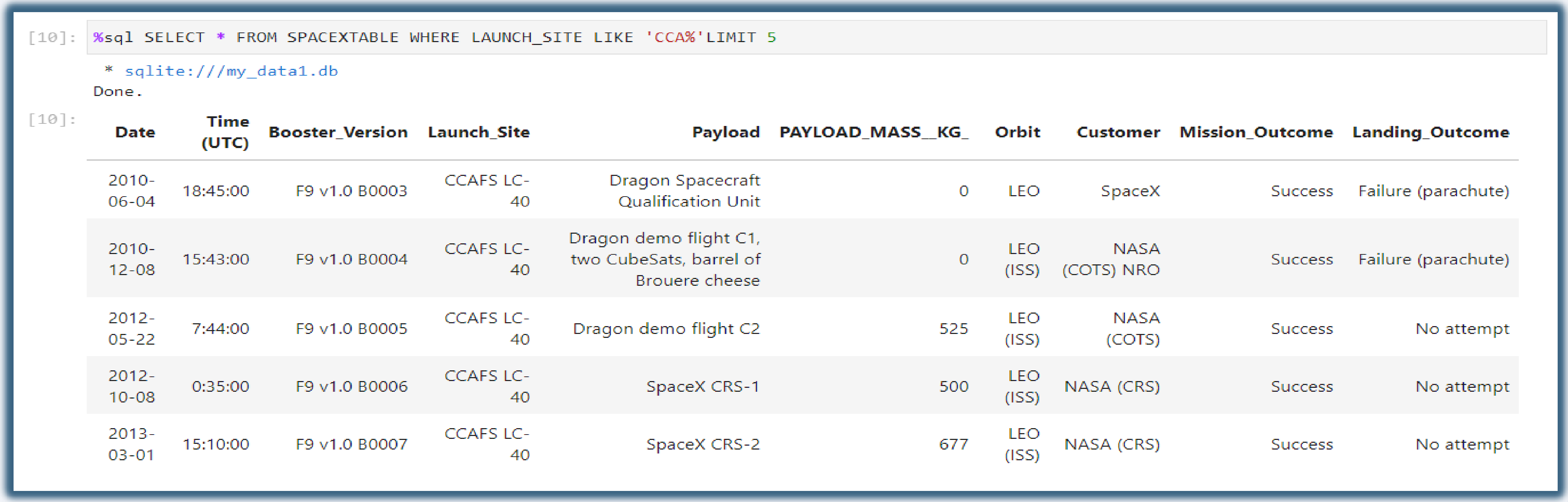
```
[9]: %sql select DISTINCT LAUNCH_SITE from SPACEXTABLE;

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

[9]:

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

Launch Site Names Begin with 'CCA'

We utilized the keywords LIKE and LIMIT to filter and display only the first five records from the database where the launch site name begins with “CCA“.



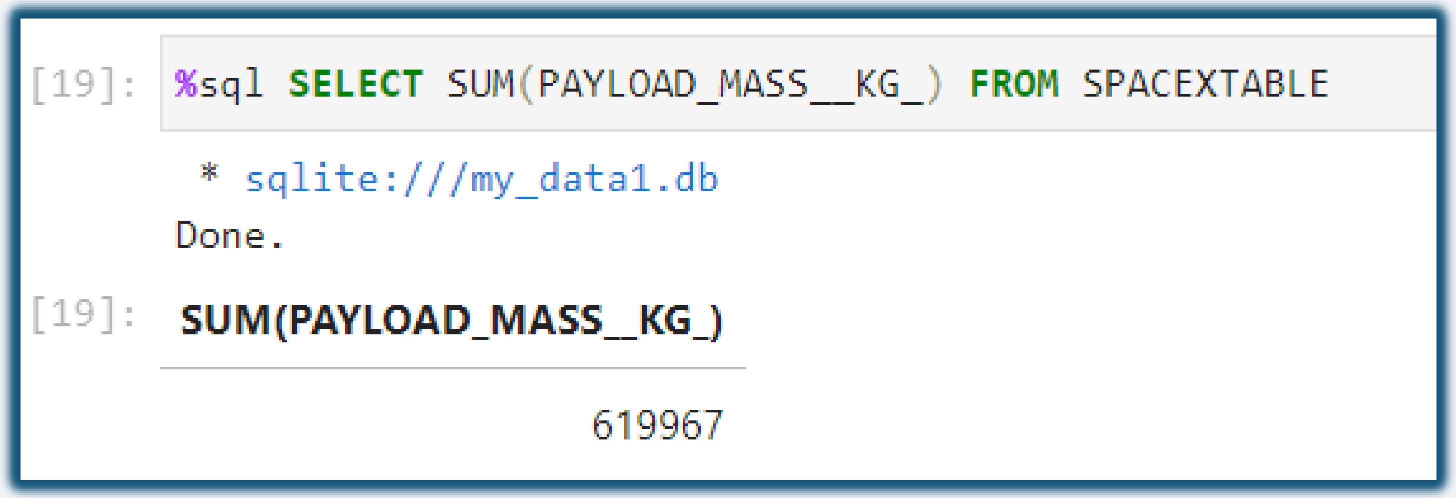
```
[10]: %sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%'LIMIT 5

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

[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

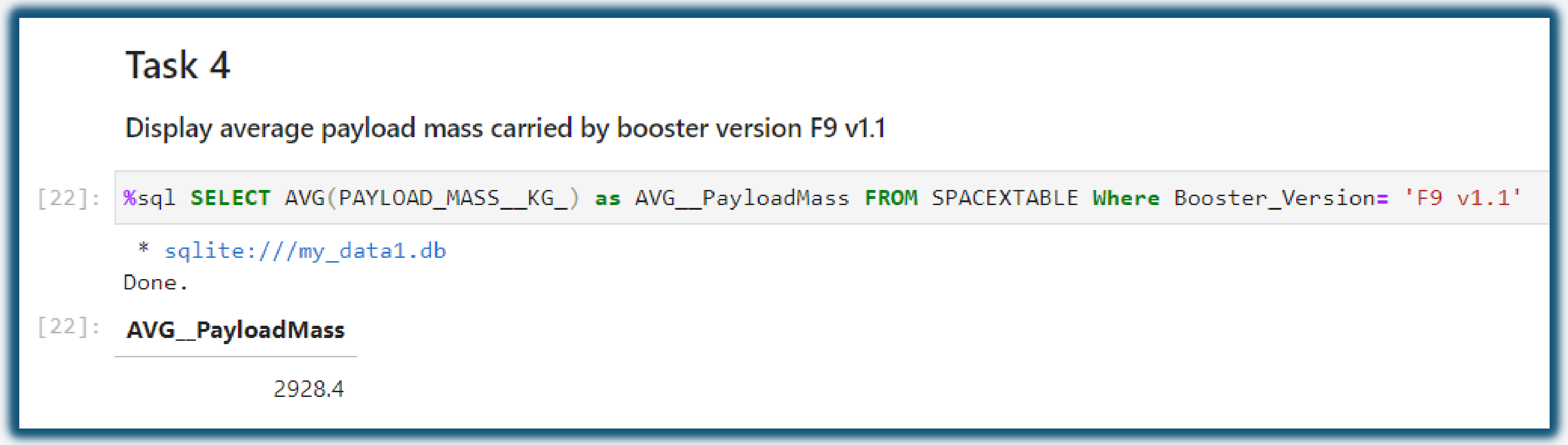
The aggregate function SUM was used to calculate the total payload carried by boosters from NASA.



```
[19]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE

 * sqlite:///my_data1.db
Done.

[19]: SUM(PAYLOAD_MASS__KG_)

                    619967
```

Average Payload Mass by F9 v1.1

We utilized the aggregate function AVG with a suitable predicate in the WHERE clause to calculate the average payload mass carried by booster version F9 v1.1.



Task 4

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

```
[22]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as AVG__PayloadMass FROM SPACEXTABLE Where Booster_Version= 'F9 v1.1'
```

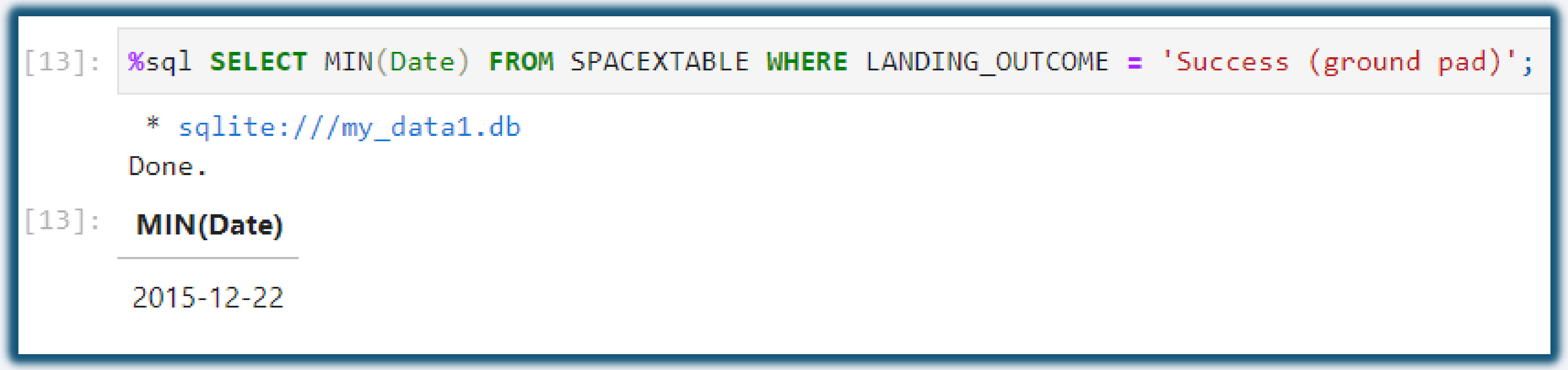
```
 * sqlite:///my_data1.db
Done.
```

[22]:

AVG__PayloadMass
2928.4

First Successful Ground Landing Date

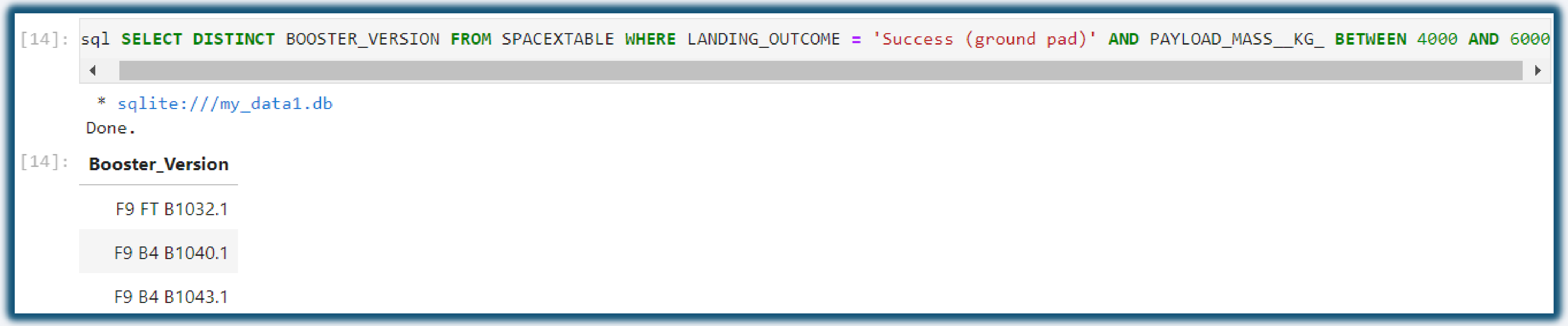
It was observed that the first successful landing on a ground pad occurred on December 22, 2015.



```
[13]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)';

       * sqlite:///my_data1.db
      Done.

[13]:  MIN(Date)

       2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

We employed the WHERE clause to filter for boosters that have achieved successful landings on drone ships. Additionally, we utilized the BETWEEN operator, along with the AND operator, to specify successful landings with a payload mass greater than 4000 but less than 6000.



```
[14]: sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

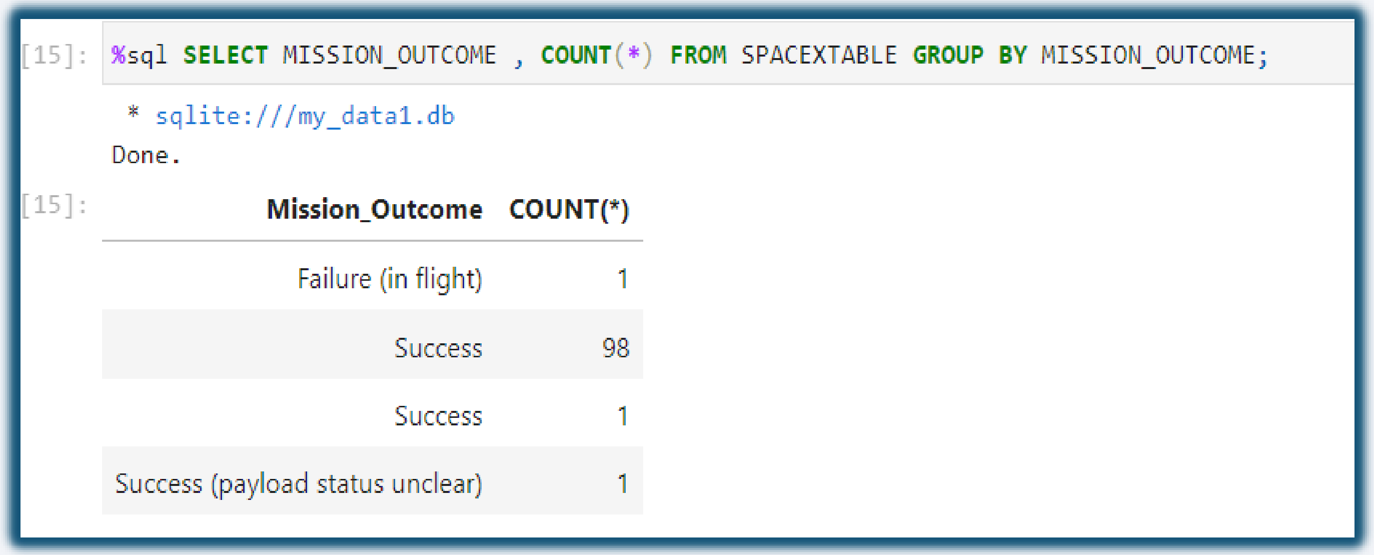
```
 * sqlite:///my_data1.db
Done.
```

[14]:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

We employed the COUNT function extensively to compute the total number of successful and unsuccessful landings.



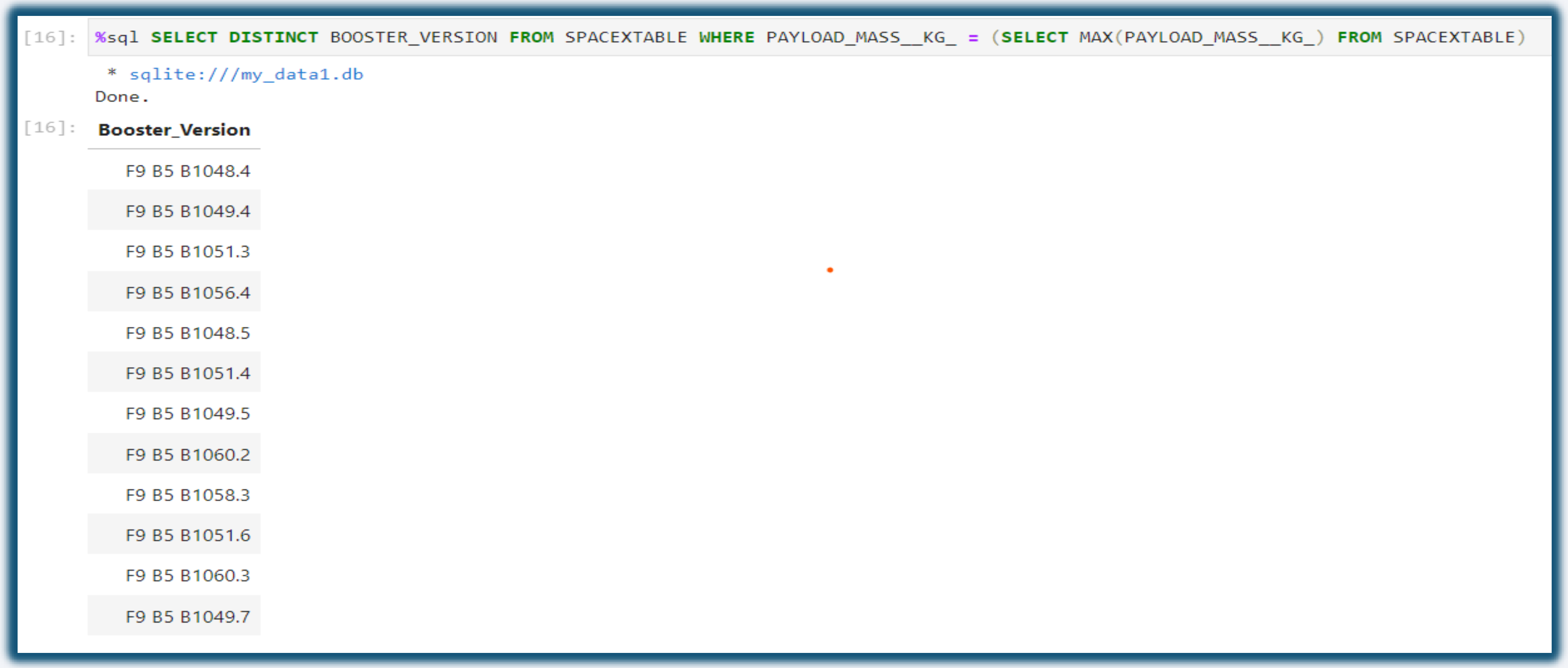
```
[15]: %sql SELECT MISSION_OUTCOME , COUNT(*) FROM SPACEXTABLE GROUP BY MISSION_OUTCOME;

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

[15]:

Mission_Outcome	COUNT(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

We employed a subquery within the WHERE clause alongside the MAX() function to determine the boosters that have carried the maximum payload.



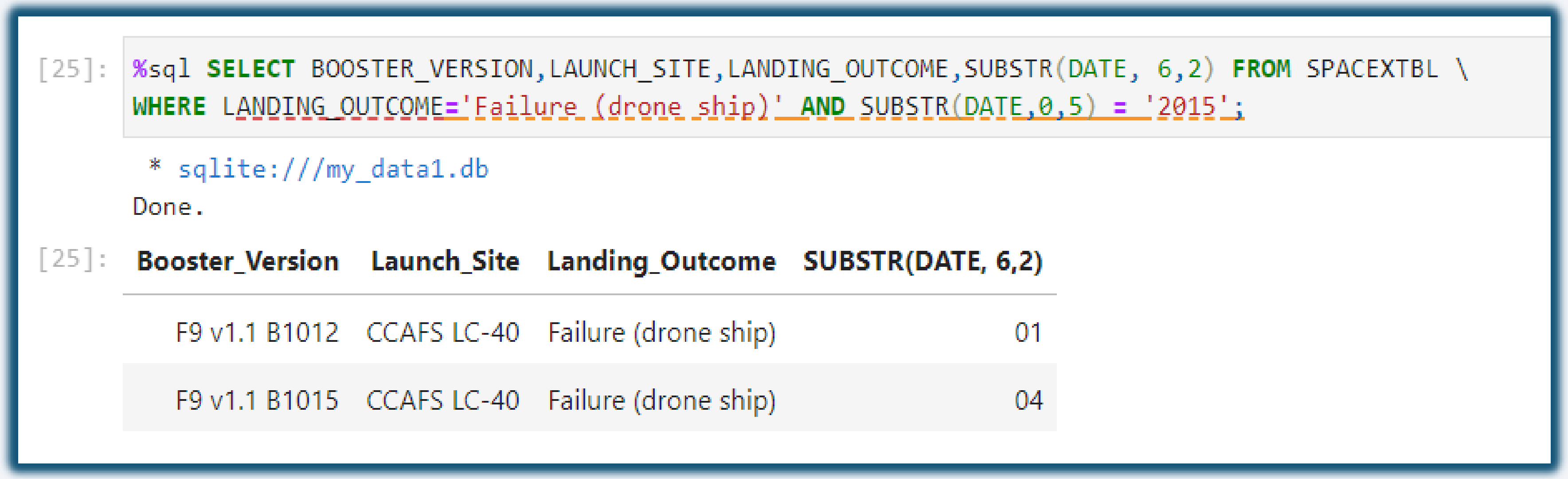
```
[16]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)

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

[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 conditions using the WHERE clause, including AND and SUBSTR, to filter for failed landing outcomes on drone ships, along with their corresponding booster versions and launch site names for the year 2015.



```
[25]: %sql SELECT BOOSTER_VERSION,LAUNCH_SITE,LANDING_OUTCOME,SUBSTR(DATE, 6,2) FROM SPACEXTBL \
      WHERE LANDING_OUTCOME='Failure (drone ship)' AND SUBSTR(DATE,0,5) = '2015';

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

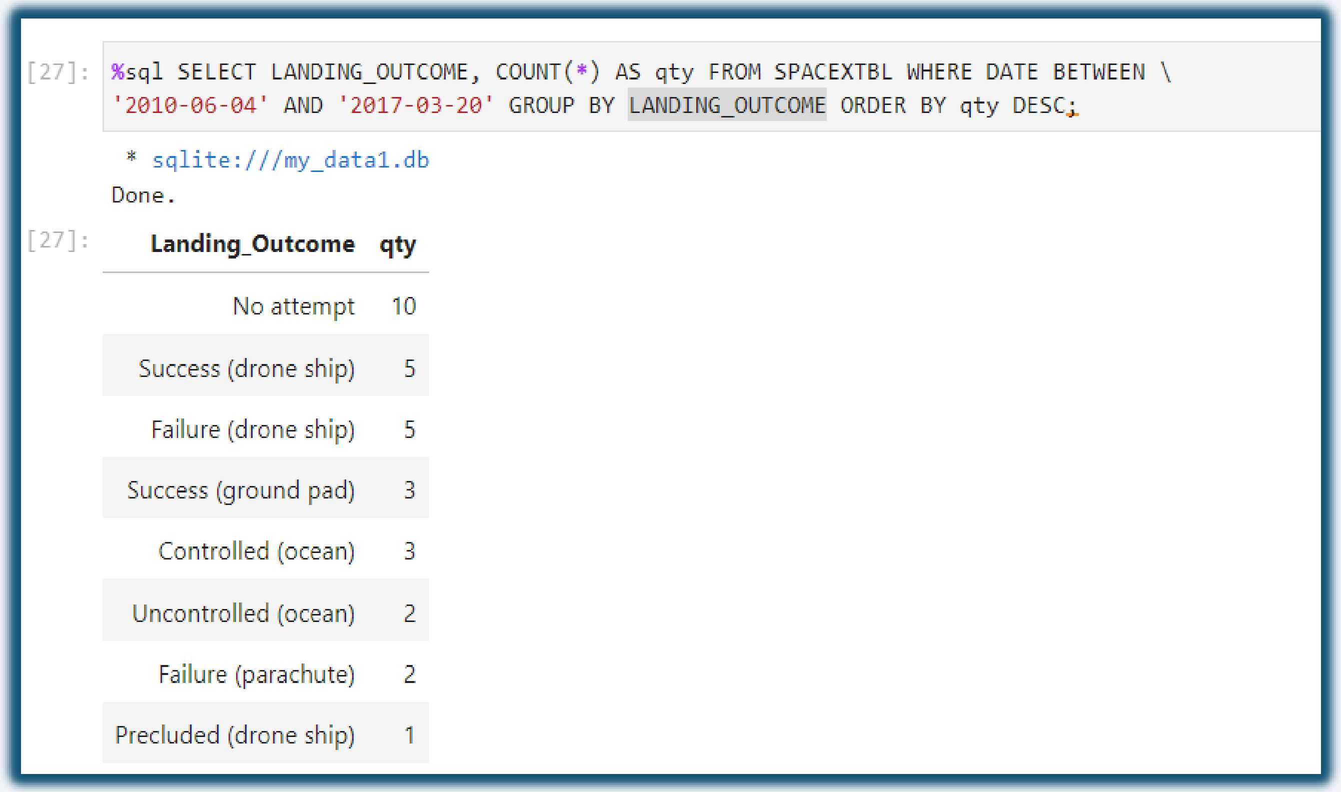
[25]:

Booster_Version	Launch_Site	Landing_Outcome	SUBSTR(DATE, 6,2)
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	01
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	04

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

- We selected landing outcomes and their corresponding counts from the dataset, using the WHERE clause with BETWEEN AND to filter for landing outcomes between June 4, 2010, and March 20, 2010.

- Then, we applied the GROUP BY clause to group the landing outcomes and the ORDER BY clause to arrange the grouped landing outcomes in descending order.



```
[27]: %sql SELECT LANDING_OUTCOME, COUNT(*) AS qty FROM SPACEXTBL WHERE DATE BETWEEN \
      '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY qty DESC;

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

[27]:

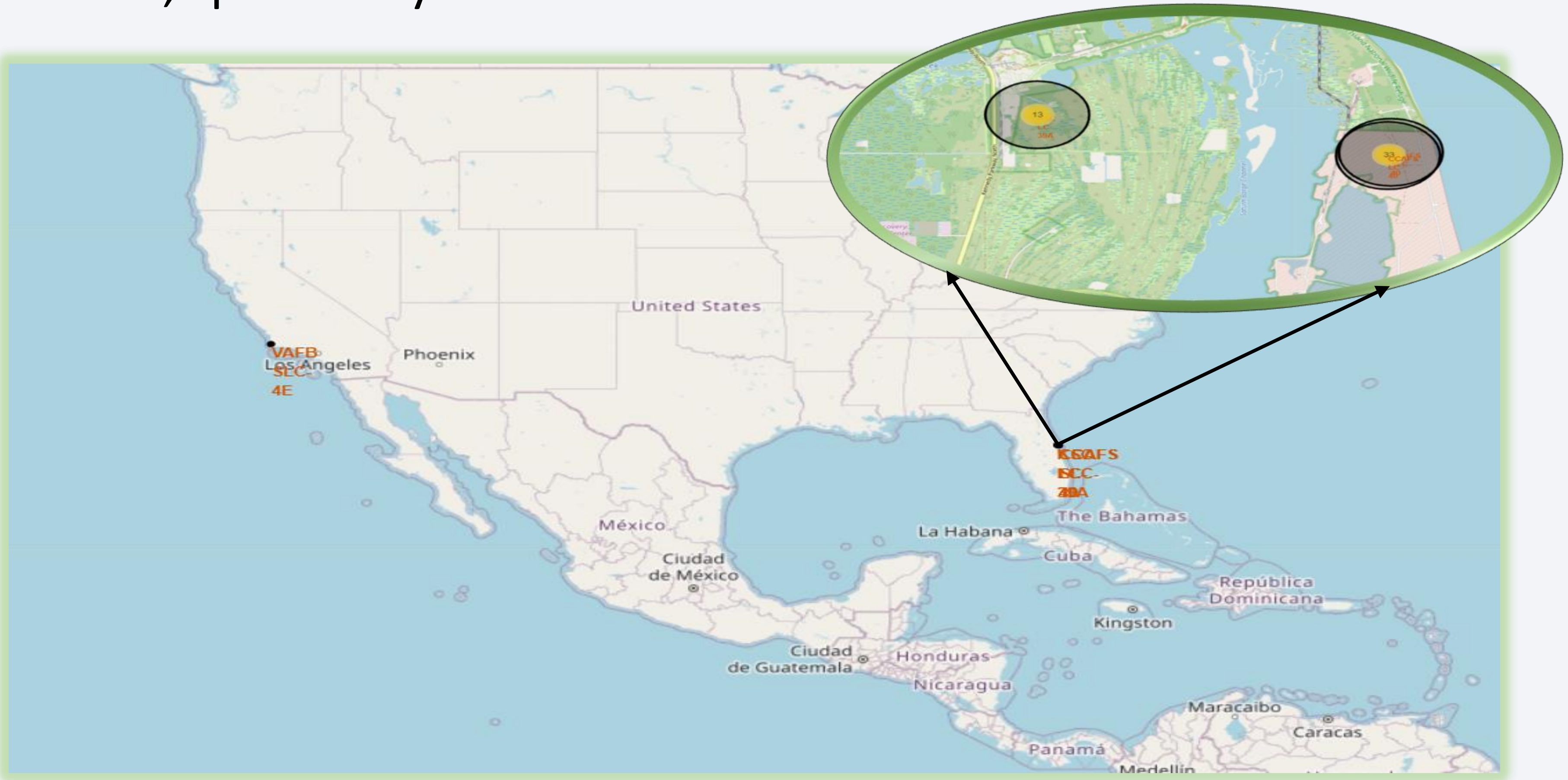
Landing_Outcome	qty
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

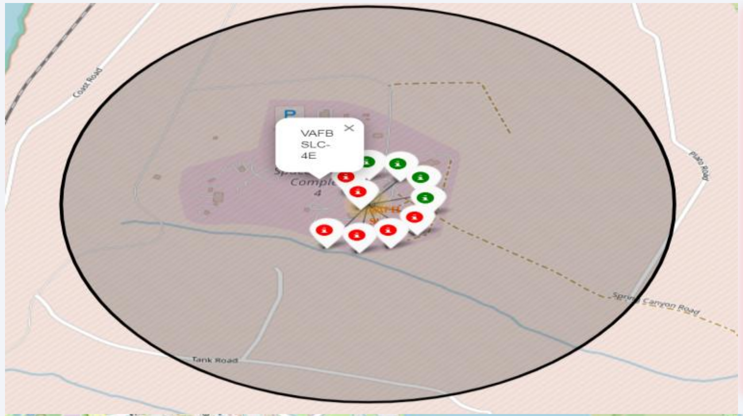
Launch Sites Proximities Analysis

All Launch Sites Global Map Markers

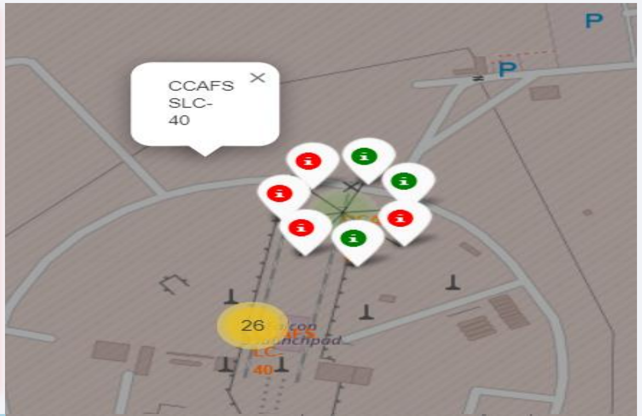
As observed from the map ,the launch sites are located along the coasts of the United States, specifically in California and Florida.

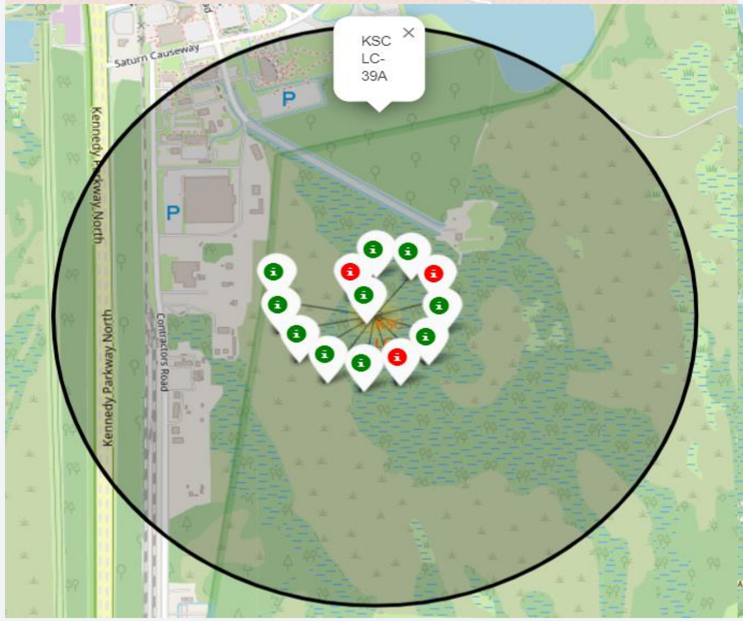


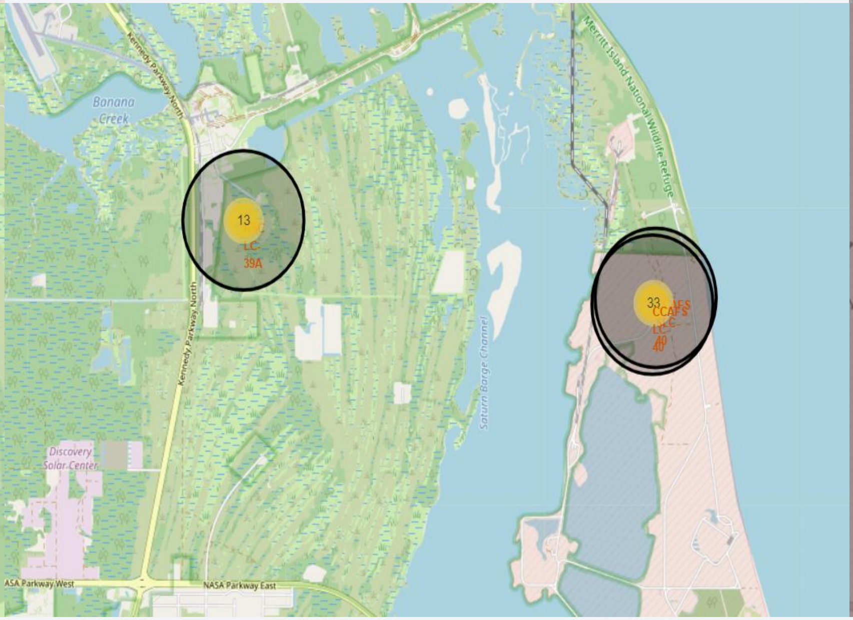
Markers showing launch sites with color labels

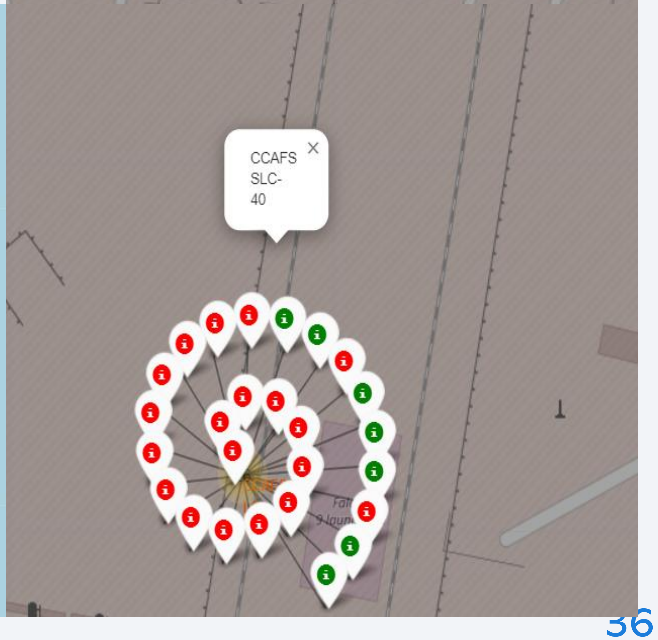


The green markers on the map indicate successful landings, whereas the red markers denote unsuccessful landings.

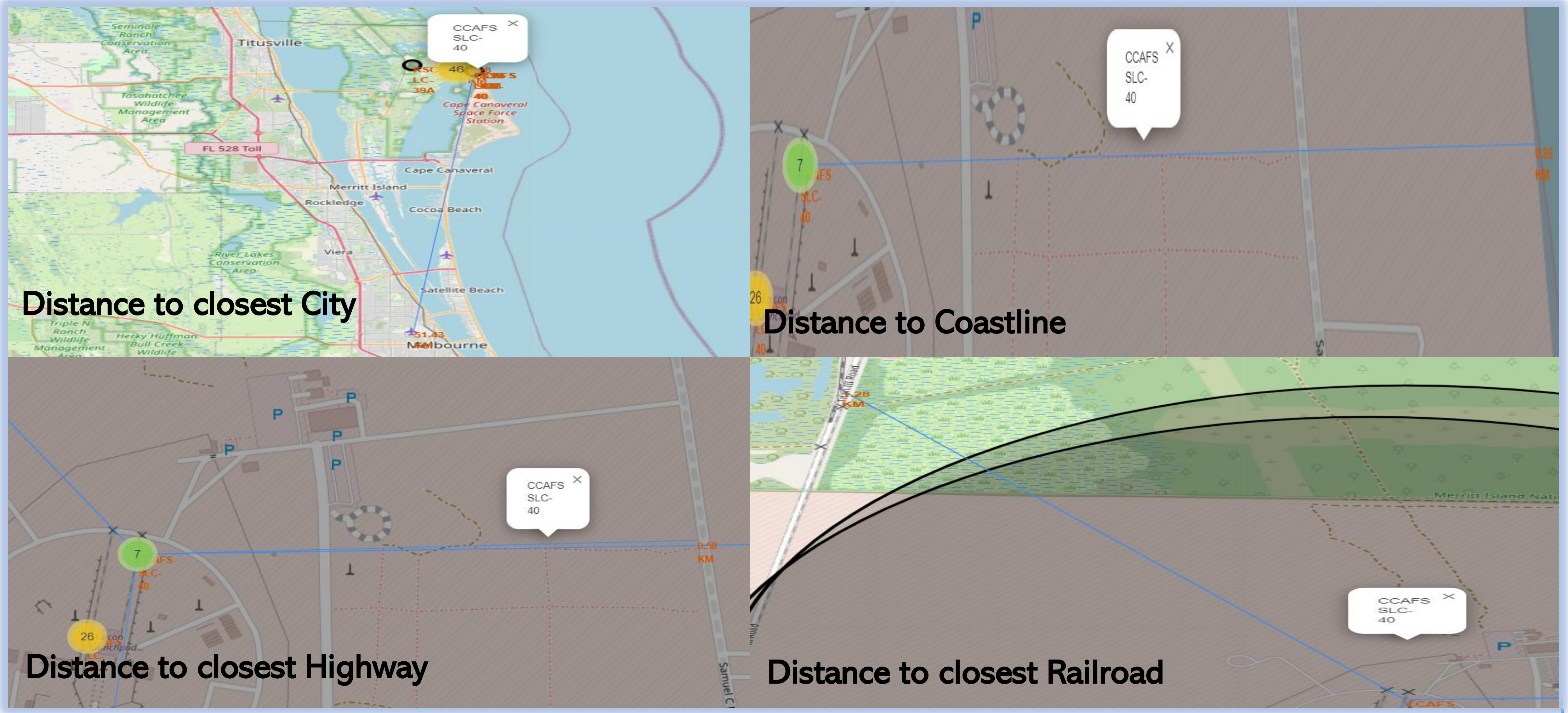








Launch Site distance to landmarks



Distance to closest City

Distance to Coastline

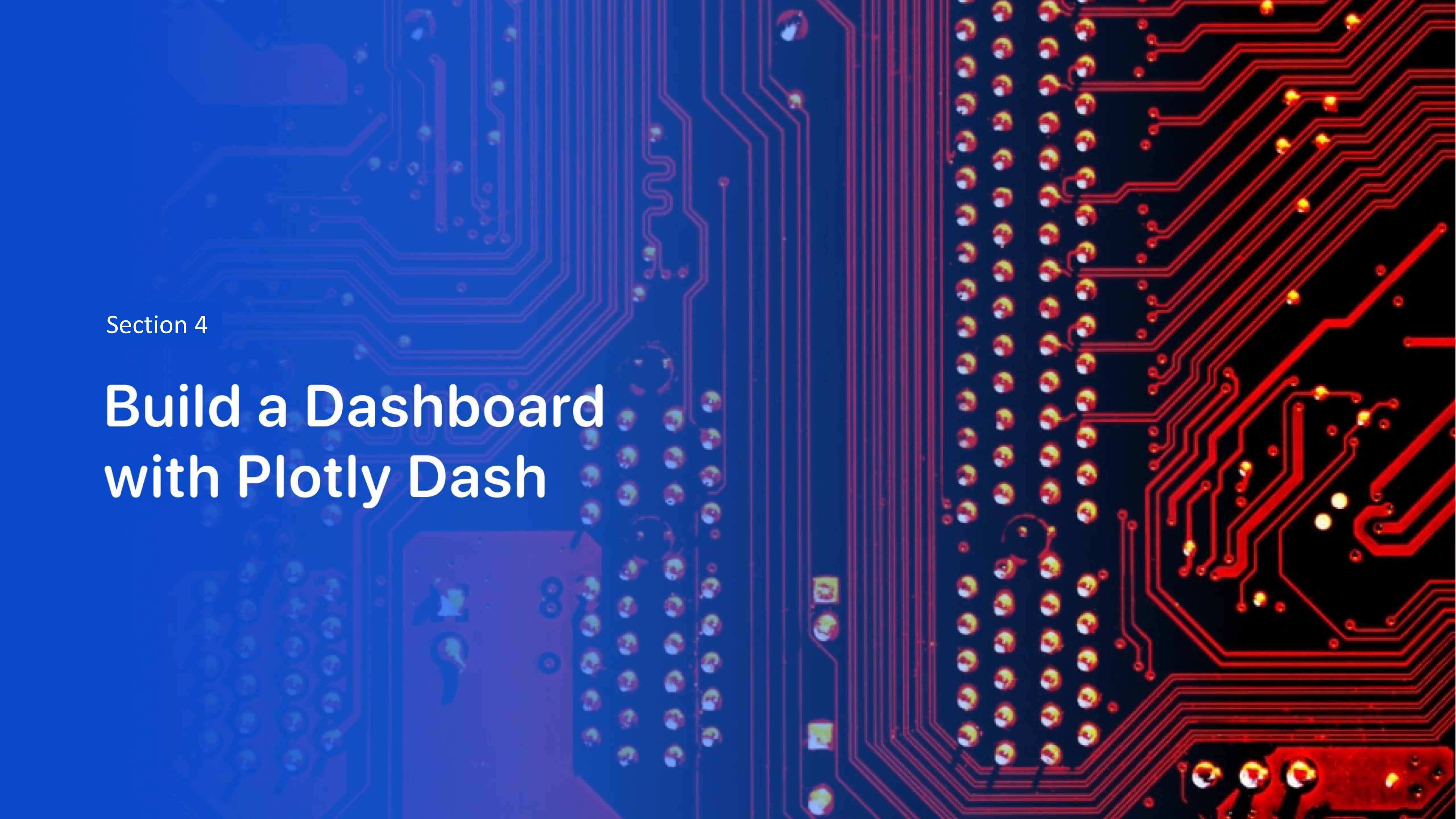
Distance to closest Highway

Distance to closest Railroad

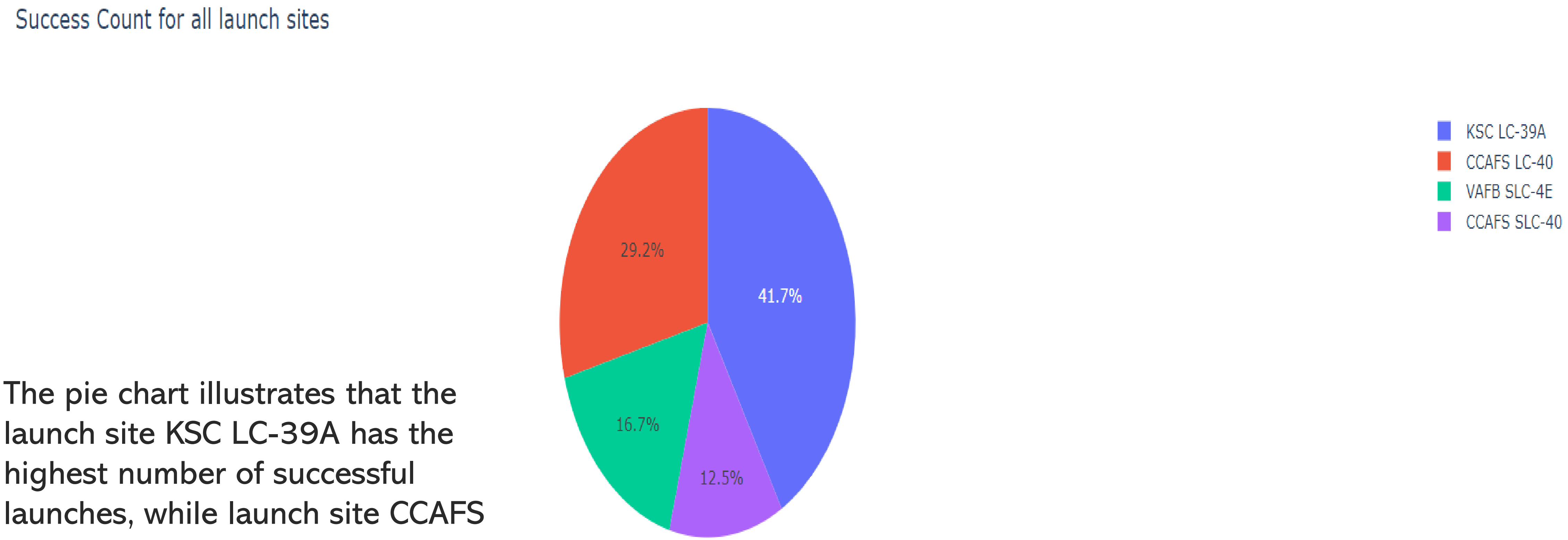
CCAFS SLC-40
Titusville
Seminole Ranch Conservation Area
Tosohatchee Wildlife Management Area
FL 528 Toll
KSC LC-39A
46
Cape Canaveral Space Force Station
Cape Canaveral
Merritt Island
Rockledge
Cocoa Beach
Viera
River Lakes Conservation Area
Satellite Beach
Triple N Ranch Wildlife Management Area
Herky Huffman Bull Creek Wildlife
51.43 KM
Melbourne
7
26
0.86 KM
0.58 KM
Samuel C
1.28 KM
Merritt Island Nat

Section 4

Build a Dashboard with Plotly Dash

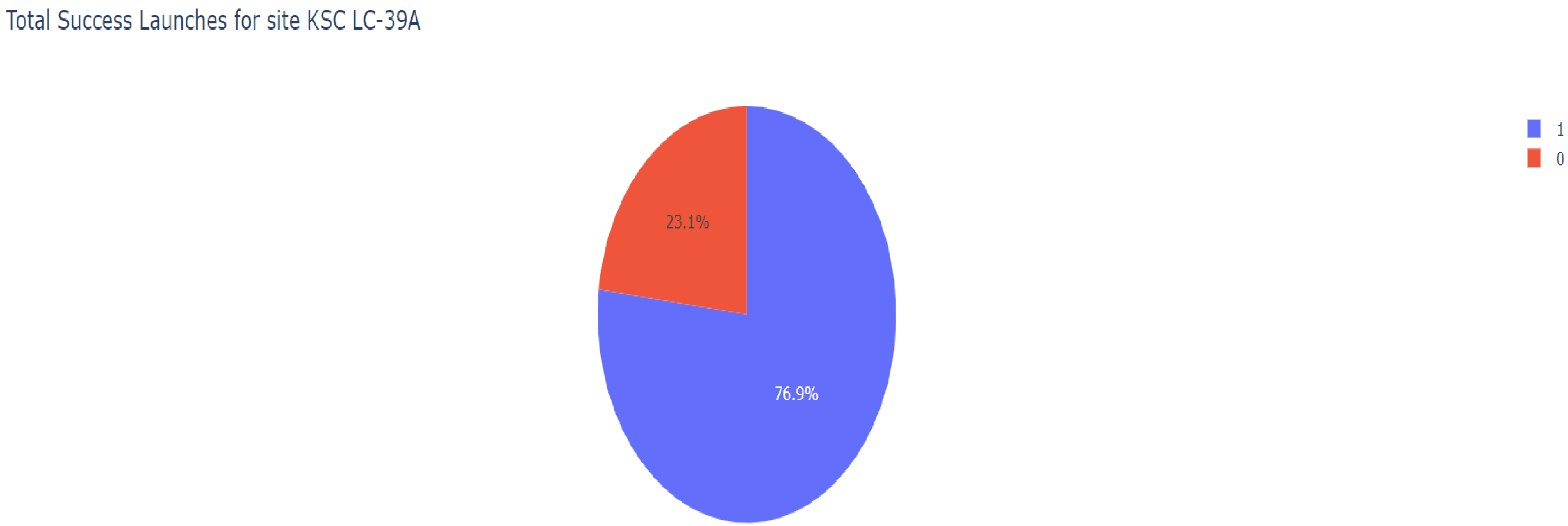


Pie chart showing the success percentage achieved by all Launch Sites



The pie chart illustrates that the launch site KSC LC-39A has the highest number of successful launches, while launch site CCAFS SLC-40 has the fewest successful launches.

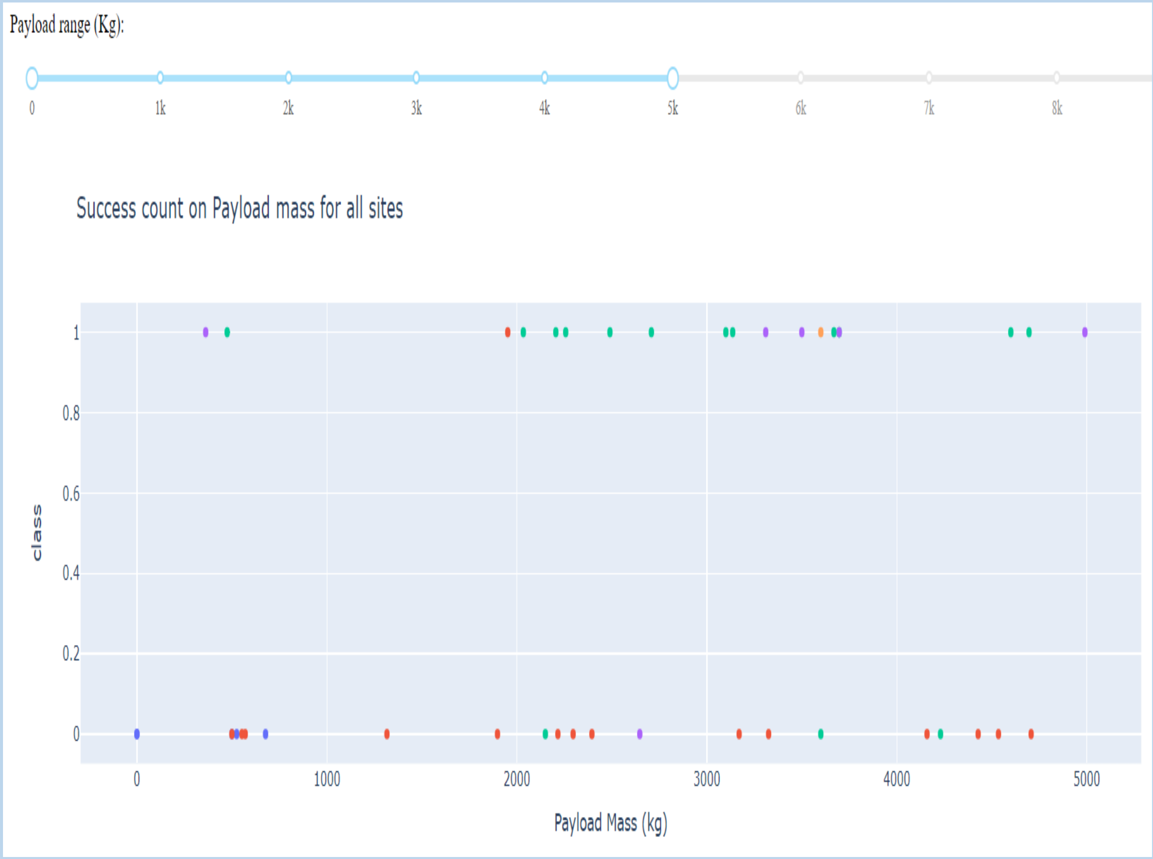
Pie chart showing the Launch site with the highest launch success ratio



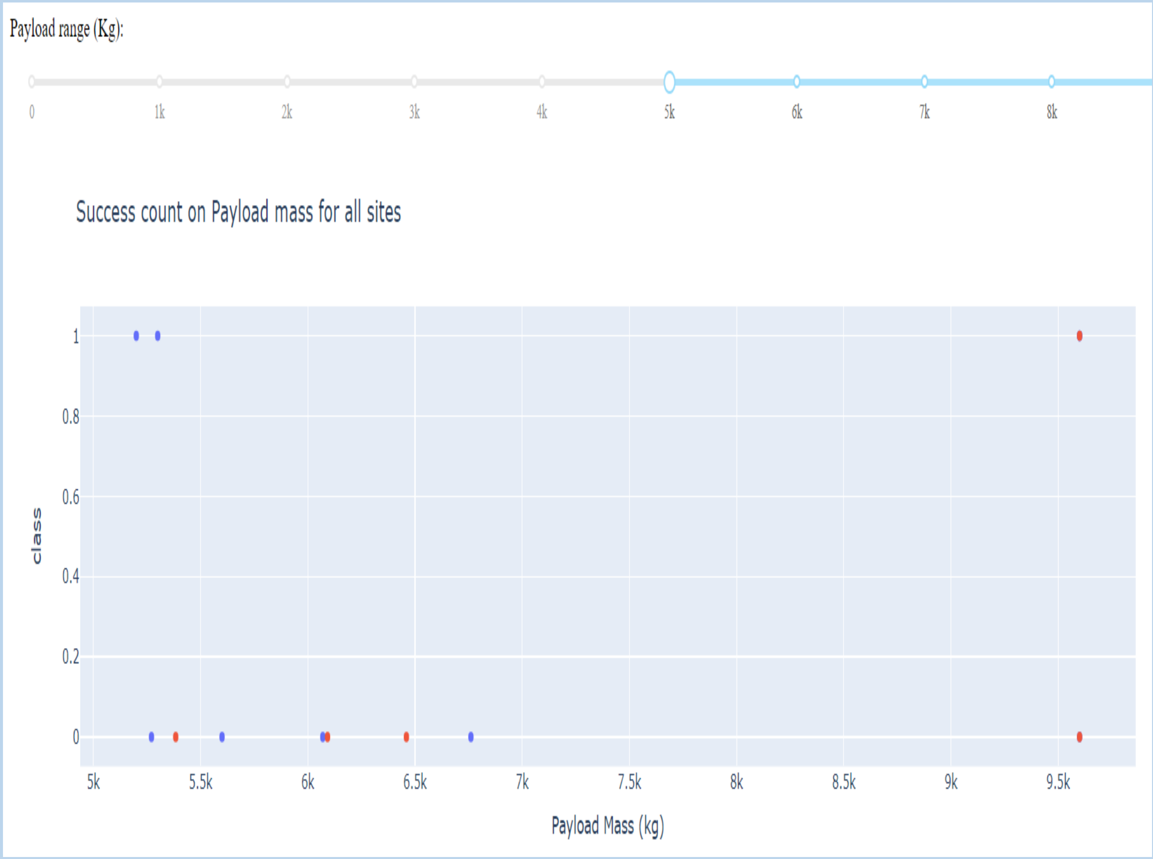
Launch site KSC LC-39A achieved the highest success rate among all launch sites, with a success rate of x% and a failure rate of y%.

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

Low weighted Payload (0-5000)kg



Heavy weighted Payload (5000-10000)kg



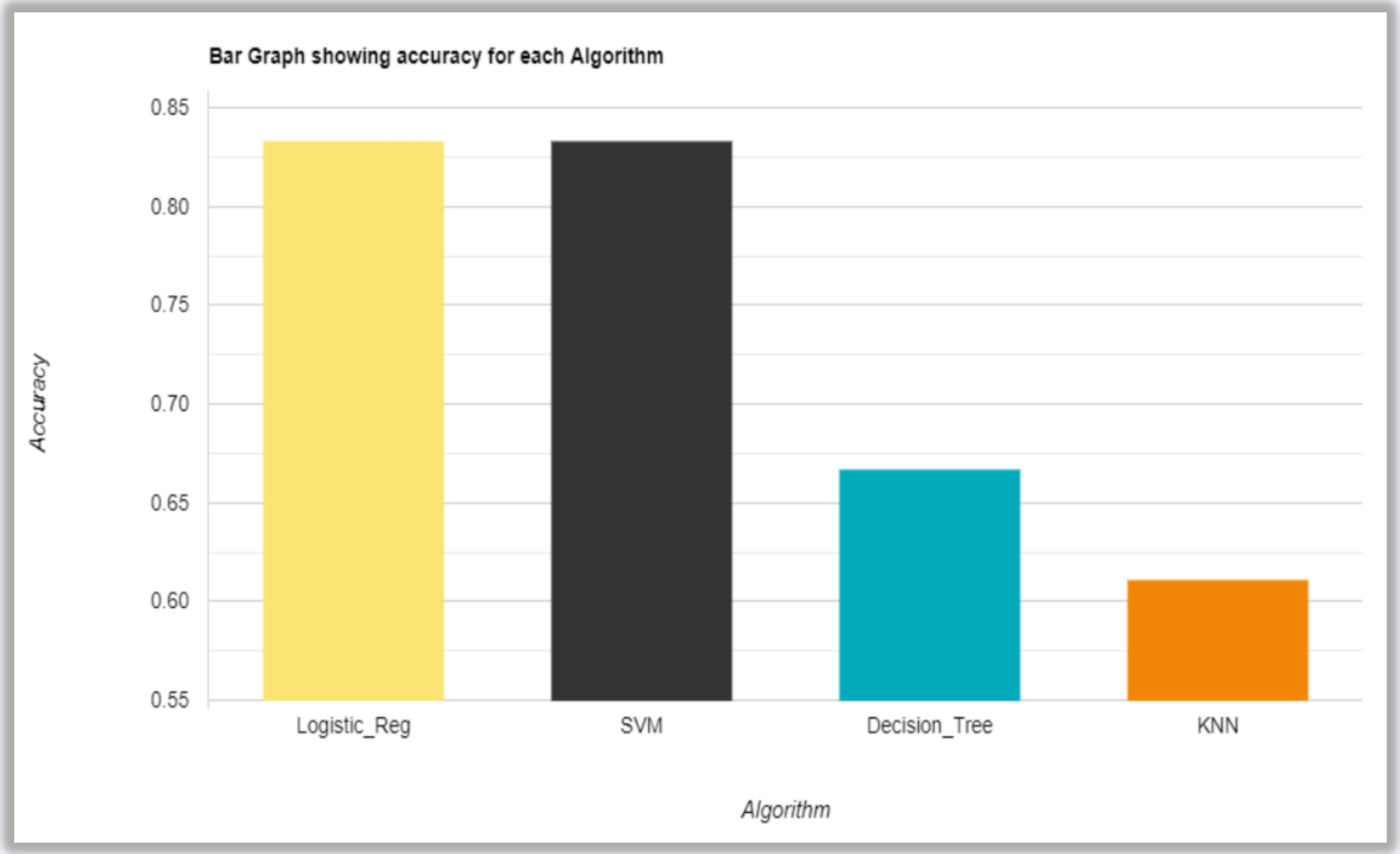
The Success rate for Low weighted Payload is higher as compared to High weighted Payload.

Section 5

Predictive Analysis (Classification)

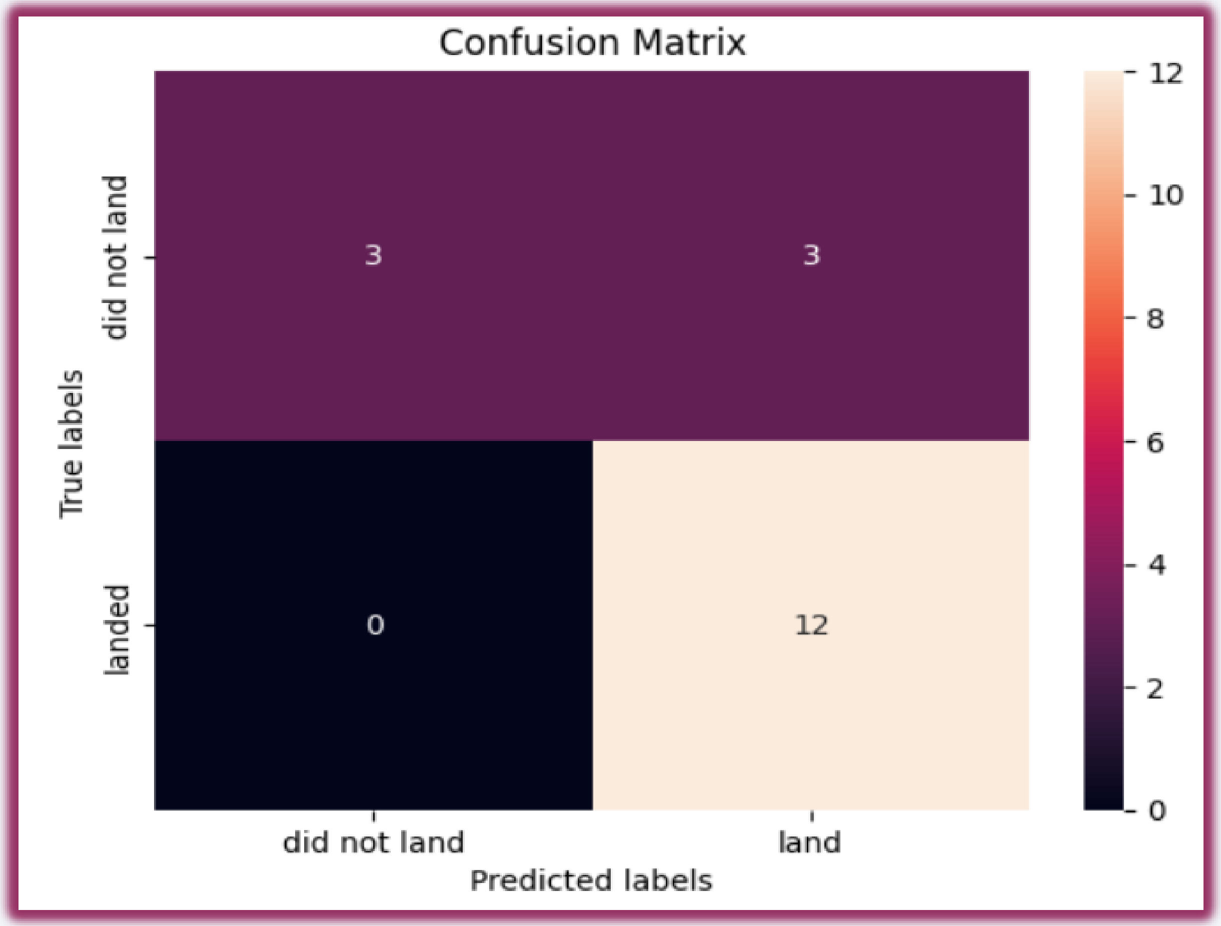
Classification Accuracy

From the bar chart observation, it's evident that the logistic regression model exhibits the highest accuracy (0.833), while the k-nearest neighbors (KNN) model demonstrates the lowest accuracy(0.611).



Confusion Matrix

- The confusion matrix for the logistic regression model indicates that the model effectively distinguishes between different classes.
- However, a notable issue lies in false positives, where unsuccessful landings are incorrectly classified as successful by the model.



Conclusions

- Heavy Payloads are associated with higher frequency of successful landings.
- Orbits ES-L1, GEO, HEO and SSO demonstrate the highest success rates.
- There is a Positive correlation between the number of flights conducted at a launch site and the success rate of launches at that site.
- The success rate has been on a steady rise since 2013, with the exception of a noticeable drop in 2018.
- The first successful landing on a ground pad occurred on December 22, 2015.
- The launch site KSC LC-39A has the highest number of successful launches, while launch site CCAFS SLC-40 has the fewest successful launches.
- Logistic Regression is the best machine learning algorithm for this task.
- KNN model demonstrates the lowest accuracy for this task.
- Other companies could use these factors to compete with SpaceX.

Appendix

All the codes are available on my GitHub:

https://github.com/Steshitaya28/SpaceX-Applied-Data-Science-Capstone

Thank you!