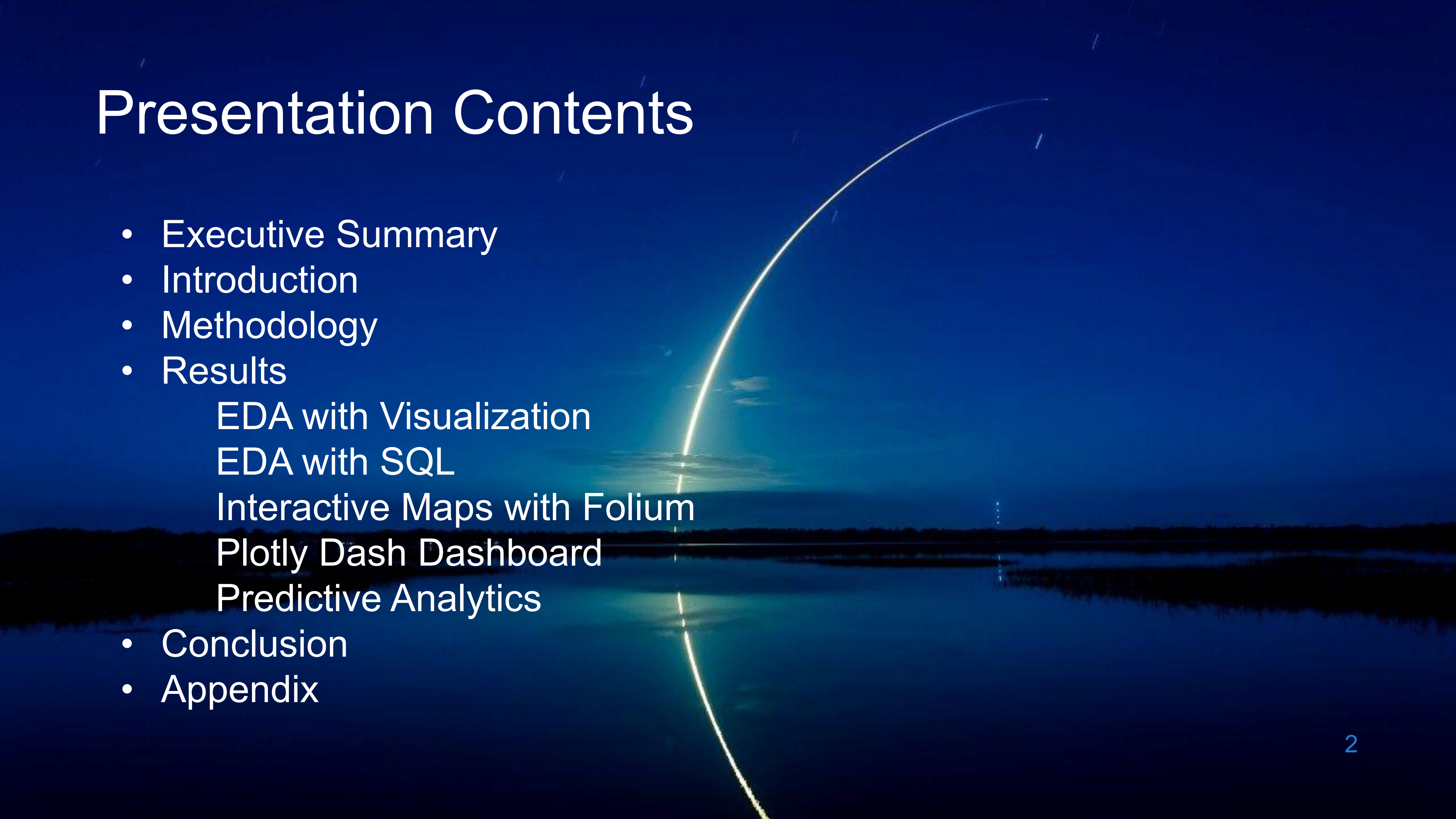
Winning Space Race with Data Science

Echeverria Ignacio
01/2024



Presentation Contents

- Executive Summary
- Introduction
- Methodology
- Results
  - EDA with Visualization
  - EDA with SQL
  - Interactive Maps with Folium
  - Plotly Dash Dashboard
  - Predictive Analytics
- Conclusion
- Appendix

Executive Summary

Summary of methodologies

- Data Collection: gathered information through SpaceX REST API and web scraping techniques.
- Data Wrangling: formulated a success/failure classification from the acquired data.
- Data Exploration: employed data visualization to delve into factors like payload, launch site, flight number, and yearly trends.
- Statistical Analysis: Leveraged SQL to compute statistics, including total payload, payload range for successful launches, and counts of successful and failed outcomes.
- Launch Site Evaluation: Investigated launch site success rates and their proximity to geographical markers.
- Visual Representation: Utilized visuals to illustrate launch sites with the highest success rates and successful payload ranges.
- Model Development: Constructed predictive models using logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbor (KNN) algorithms to forecast landing outcomes.

Summary of all results

- Most launch sites are near the equator, and all are close to the coast
- Over time, there's a noticeable enhancement in launch success.
- KSC LC 39A emerges as the leading landing site with the highest success rate.
- Orbits ES L1, GEO, HEO, and SSO exhibit a flawless 100% success rate.
- Across the test set, all models displayed comparable performance. However, the decision tree model showcased a slight edge in performance.

Introduction

Background and context

- In this capstone project, our objective is to predict the successful landing of the Falcon 9 first stage. SpaceX promotes Falcon 9 rocket launches on its platform at a price of 62 million dollars, significantly lower than other providers whose costs can go up to 165 million dollars per launch. This cost efficiency primarily stems from SpaceX's ability to reuse the first stage. Hence, by determining the success of the first stage landing, we can estimate the overall launch cost. Such insights could be invaluable if a competing company aims to bid against SpaceX for a rocket launch.

This project aims to answer these questions

- How do payload mass, launch site, flight number, and orbits impact the success of the first stage landing?
- How has the rate of successful landings evolved over time?
- What is the most effective predictive model for determining successful first stage landings using binary classification?

Methodology

Methodology

Data collection methodology

- Data was collected for the SpaceX capstone via the SpaceX REST API and web scraping from Wiki pages. The SpaceX REST API offered launch specifics, including rockets, payloads, and landing outcomes. Web scraping using BeautifulSoup extracted Falcon 9 launch records from Wiki pages. The collected data underwent wrangling, filtering out Falcon 1 launches, and addressing NULL values to prepare a clean dataset for analysis.

Perform data wrangling

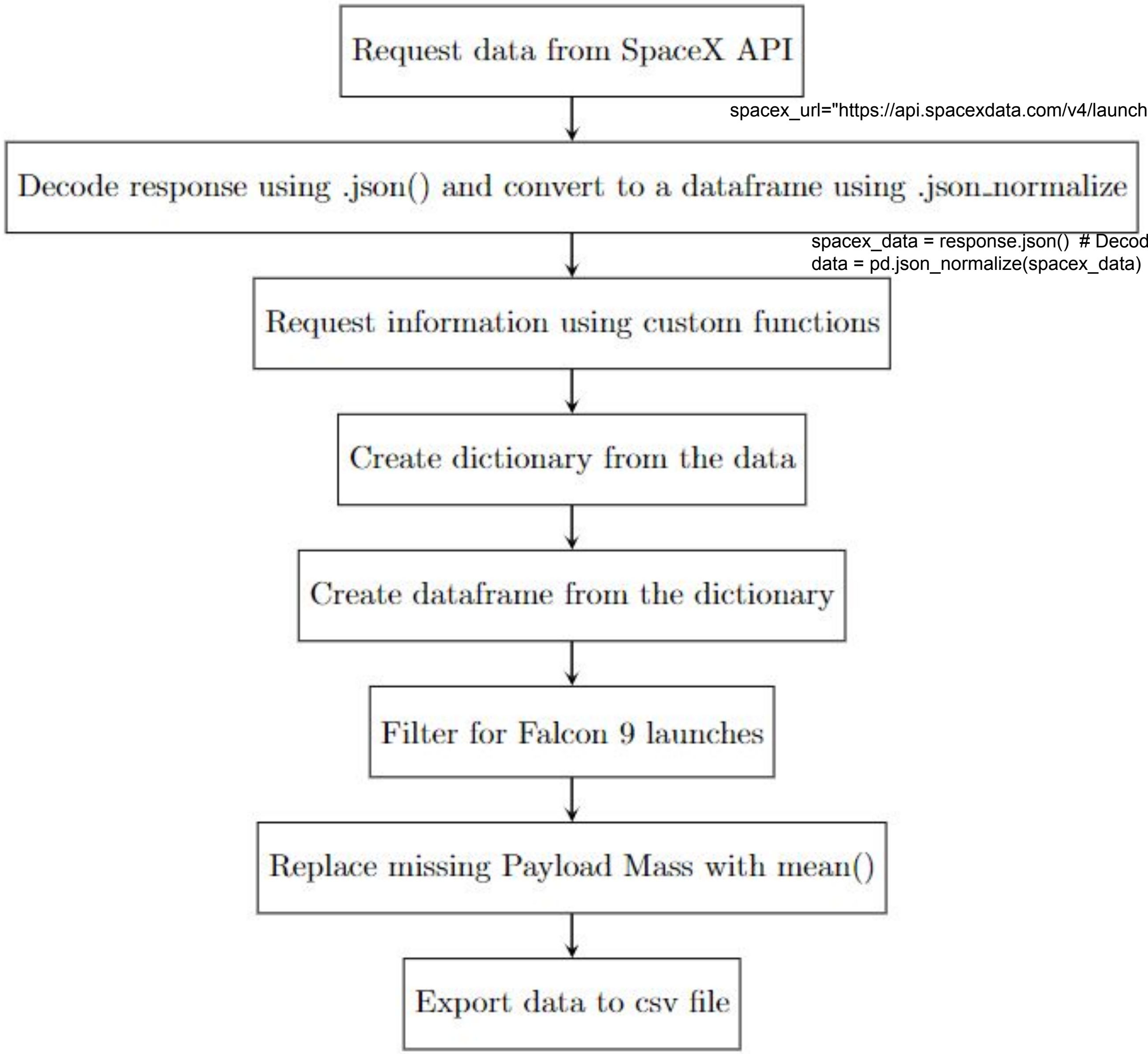
- Data was processed by cleaning attributes like Flight Number, Date, Booster version, and Payload mass. Launch sites such as Vandenberg AFB, Kennedy Space Center, and CCAFS SLC 40 were categorized. Orbits like Low Earth orbit (LEO) and geosynchronous orbit (GTO) were identified. Outcomes indicating successful or unsuccessful landings were converted into binary classes (0 or 1) for analysis. Perform exploratory data analysis (EDA) using visualization and SQL

Perform exploratory data analysis (EDA) using visualization and SQL.

Perform interactive visual analytics using Folium and Plotly Dash.

Perform predictive analysis using classification models.

Data Collection - SpaceX API



Request data from SpaceX API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

Decode response using .json() and convert to a dataframe using .json_normalize

```
spacex_data = response.json()  # Decode the JSON content
data = pd.json_normalize(spacex_data)  # Convert JSON to DataFrame
```

Request information using custom functions

Create dictionary from the data

Create dataframe from the dictionary

Filter for Falcon 9 launches

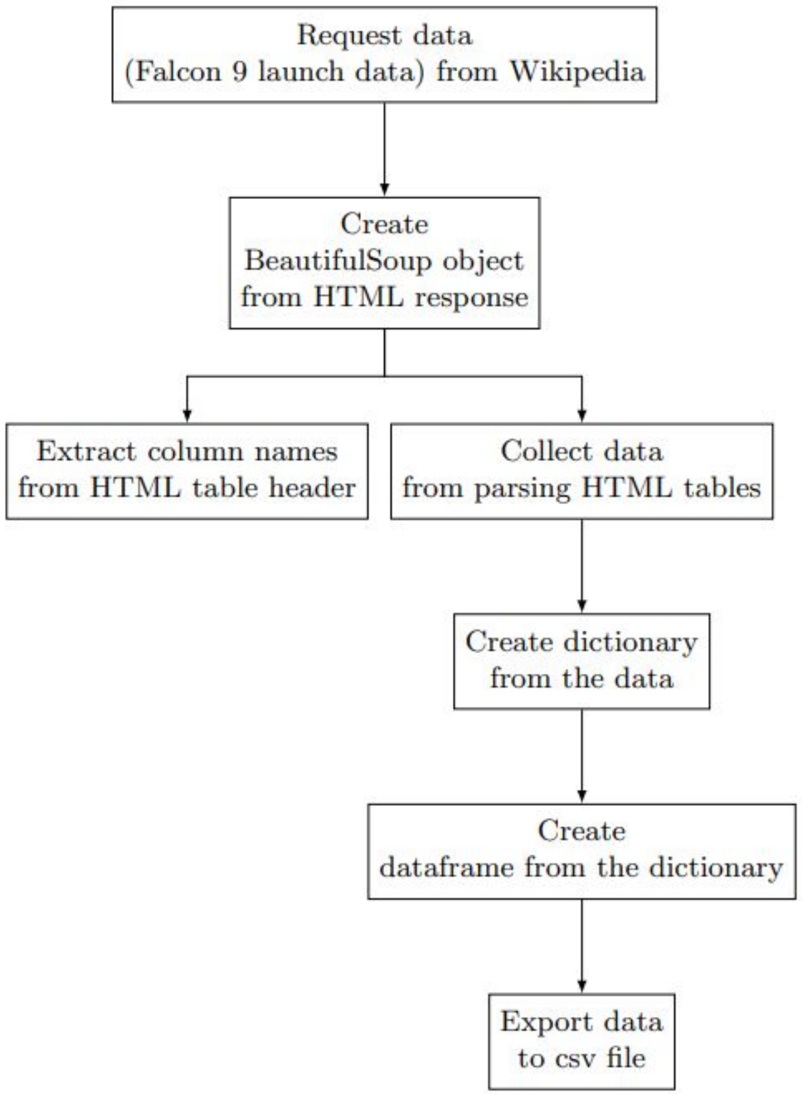
Replace missing Payload Mass with mean()

Export data to csv file

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

GitHub [Data Collection]

Data Collection - Web Scraping



Request data (Falcon 9 launch data) from Wikipedia

Create BeautifulSoup object from HTML response

Extract column names from HTML table header

Collect data from parsing HTML tables

Create dictionary from the data

Create dataframe from the dictionary

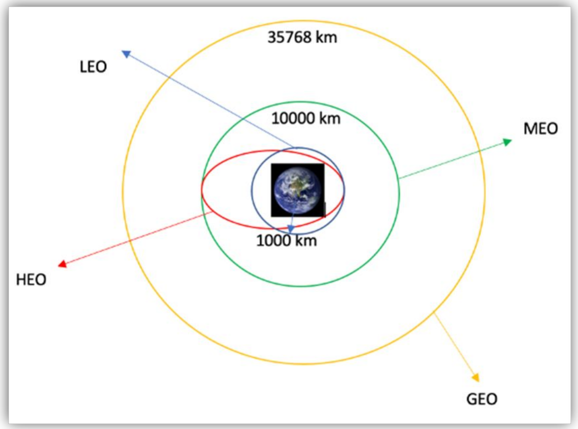
Export data to csv file

Data Collection

Data Wrangling

Process of cleaning, restructuring, and transforming raw data for analysis.

- Data Cleaning: Addressing missing or incorrect data, managing outliers.
- Data Transformation: Restructuring for analysis, converting types, creating new variables.
- Data Enrichment: Combining datasets for enhanced insights.
- Handling Missing Data: Imputation or deletion based on context.
- Dealing with Outliers: Identification and management for accurate analysis.
- Data Normalization: Scaling or transforming for standardized comparisons.
- Data Formatting: Ensuring consistency in formats and units.



Steps

- Determine Launch Distribution: Explore launches across SpaceX facilities.
- Analyze Orbital Types: Examine the prevalence of different orbits.
- Evaluate Mission Outcomes: Analyze outcomes based on orbit types.
- Binary Classification: Establish success/failure labels for launches.
- Success Rate Calculation: Evaluate overall SpaceX launch success.
- Export Data to CSV: Save processed data for future analysis.

GitHub Data Wrangling

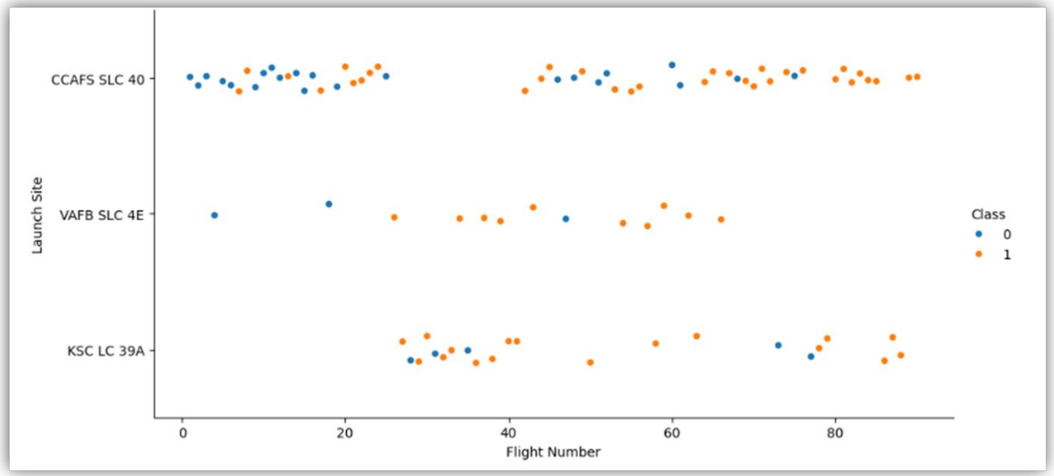
EDA with Data Visualization

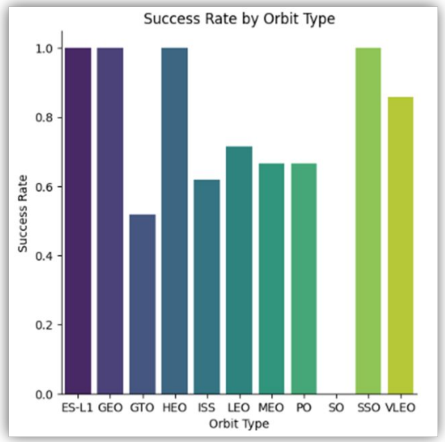
We initiated the analysis using scatter graphs to unveil relationships between attributes:

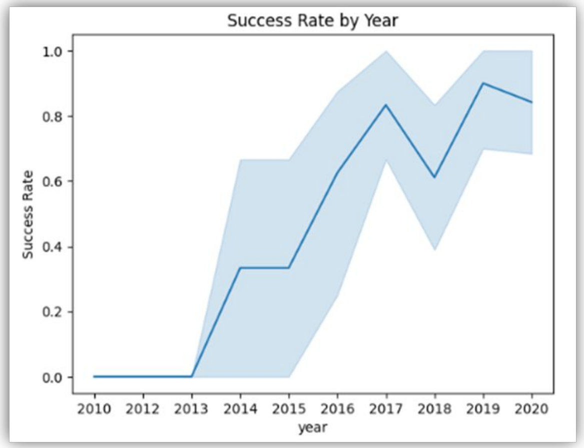
- Payload and Flight Number.
- Flight Number and Launch Site.
- Payload and Launch Site.
- Flight Number and Orbit Type.
- Payload and Orbit Type.

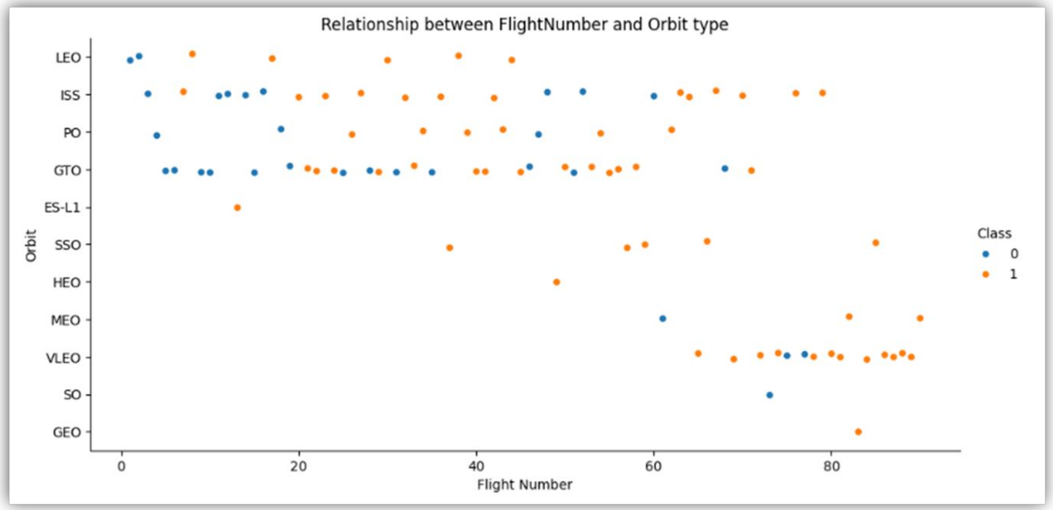
Scatter plots reveal attribute dependencies, aiding in understanding factors that impact landing success.

After exploring relationships with scatter plots, we'll utilize bar graphs for attribute interpretation, focusing on orbits' success probability. Line graphs will illustrate attribute trends over time, revealing yearly launch success patterns. Feature Engineering involves creating dummy variables for categorical columns, enhancing future success prediction modules.









EDA with Data Visualization

EDA with SQL

Utilizing SQL, we executed several queries to gain a comprehensive understanding of the dataset, including:

- Retrieving the names of all launch sites.
- Displaying 5 records where launch sites start with the string ‘CCA’.
- Summarizing the total payload mass carried by boosters launched under NASA's CRS program.
- Calculating the average payload mass carried by booster version F9 v1.1.
- Listing the date of the first successful landing outcome on a ground pad.
- Identifying boosters with success in drone ship landings and payload mass between 4000 and 6000.
- Providing the total count of successful and failed mission outcomes.
- Listing booster_versions associated with the maximum payload mass.
- Detailing failed drone ship landings, including booster versions and launch site names, in the year 2015.

GitHub EDA with SQL

Build an Interactive Map with Folium

Launch Site Markers:

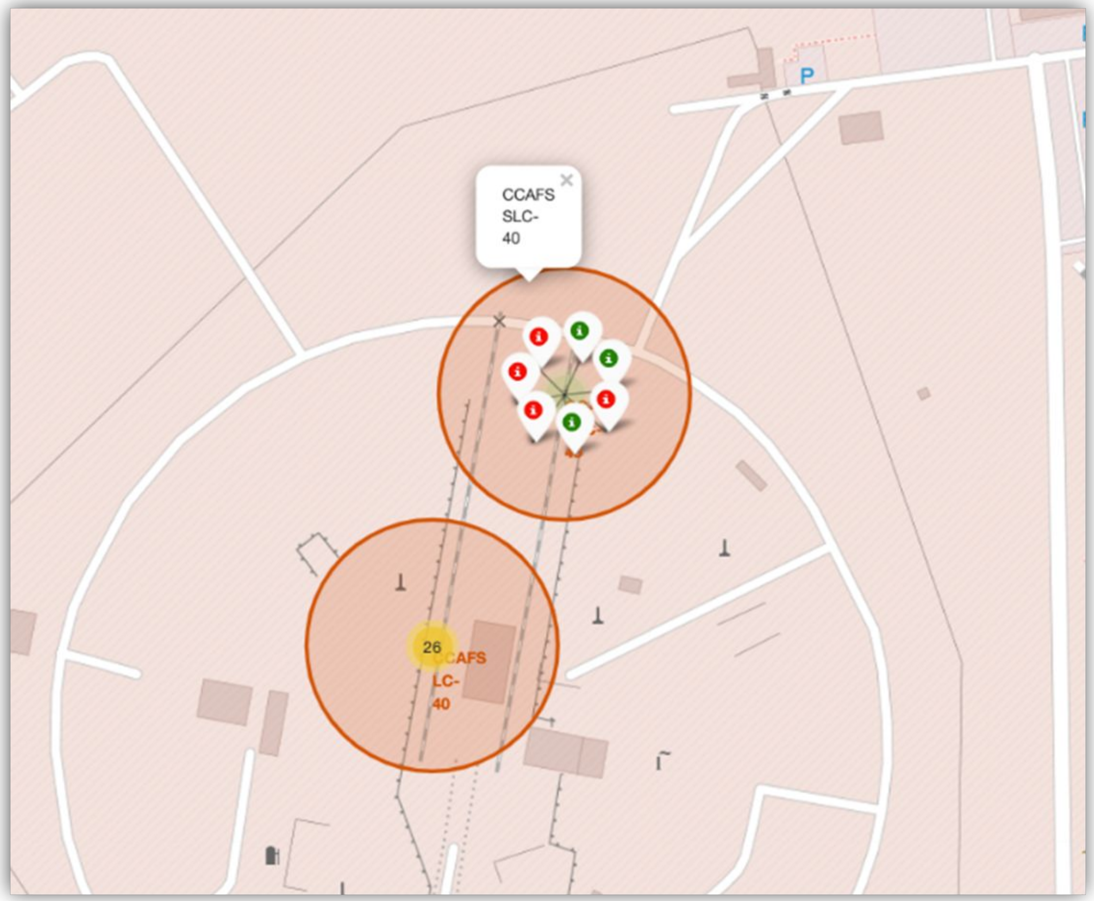
- Introduced a distinctive blue circle at the coordinates of NASA Johnson Space Center, accompanied by a popup label that displays its name using latitude and longitude coordinates.
- Implemented red circles at the coordinates of all launch sites, featuring popup labels indicating their names through precise latitude and longitude coordinates.

Launch Outcomes Visualization:

- Incorporated colored markers, utilizing green for successful and red for unsuccessful launches at each site, providing a visual representation of success rates.

Proximity Distances:

- Introduced colored lines to illustrate the distance between launch site CCAFS SLC 40 and nearby features, including the nearest coastline, railway, highway, and city.



GitHub Folium

Build a Dashboard with Plotly Dash

Launch Site Dropdown:

- Enable users to choose between all launch sites or a specific launch site from a dropdown list.

Plotly Dash Dashboard Features:

- Payload Mass Range Slider: Facilitate user selection of payload mass range through an interactive slider.
- Successful Launches Pie Chart: Provide a percentage breakdown of successful and unsuccessful launches for enhanced insight.
- Payload Mass vs. Success Rate Scatter Chart (by Booster Version): Visualize the correlation between payload mass and launch success, allowing users to explore patterns by booster version.

SPACEX

Launch Dashboard

Dashboard

Predictive Analysis (Classification)

Initial Stage:

Constructing the Model

- Load the dataset using NumPy and Pandas
- Preprocess the data, followed by division into training and testing sets.
- Determine the appropriate type of machine learning to employ.
- Specify the parameters and algorithms for GridSearchCV and apply it to the dataset.

Assessing the Model

- Examine the accuracy for each model.
- Obtain optimized hyperparameters for each algorithm type.
- Visualize the confusion matrix.

Improving the Model

- Employ Feature Engineering and Algorithm Tuning.

Determining the Optimal Model

- Identify the model with the highest accuracy score as the most effective performing model.

GitHub Predictive Analysis

Results

Data Analysis Highlights:

- Over time, there is a noticeable improvement in launch success rates.
- KSC LC 39A emerges as the landing site with the highest success rate.
- Orbits ES L1, GEO, HEO, and SSO consistently achieve a 100% success rate.

Visual Analytics Summary:

- Most launch sites are strategically located near the equator and in close proximity to the coast.
- Launch sites strike a balance, being far enough from potential damage sources (city, highway, railway) in case of a failed launch, yet still accessible for necessary support.

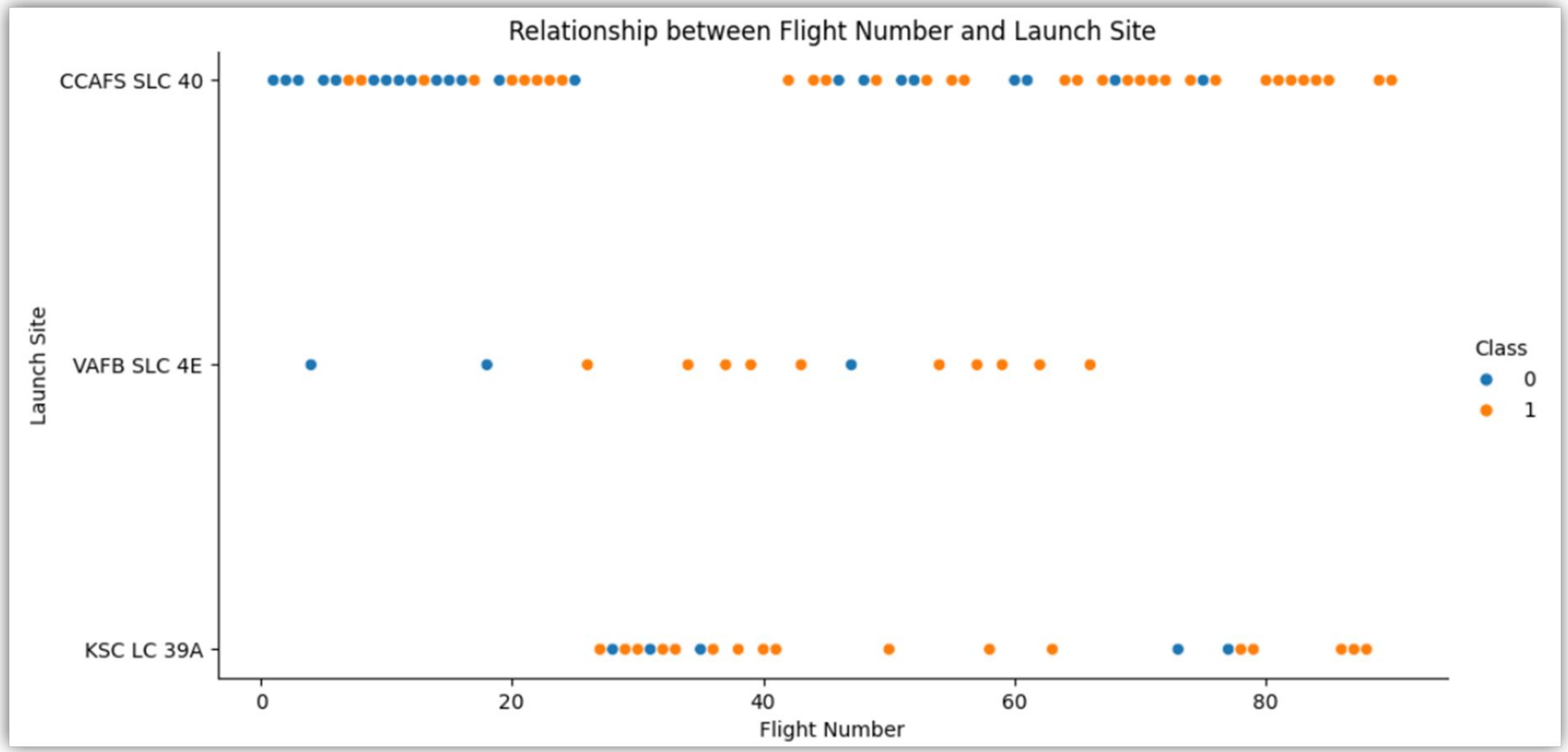
Predictive Analytics Insight:

- The Decision Tree model stands out as the most effective predictive model for the dataset.

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

- ★ Earlier flights: lower success rate **blue = fail**
- ★ Later flights: higher success rate **orange = success**
- ★ VAFB SLC 4E and KSC LC 39A exhibit elevated success rates.
- ★ The CCAFS SLC 40 launch site emerged as the most frequently utilized location

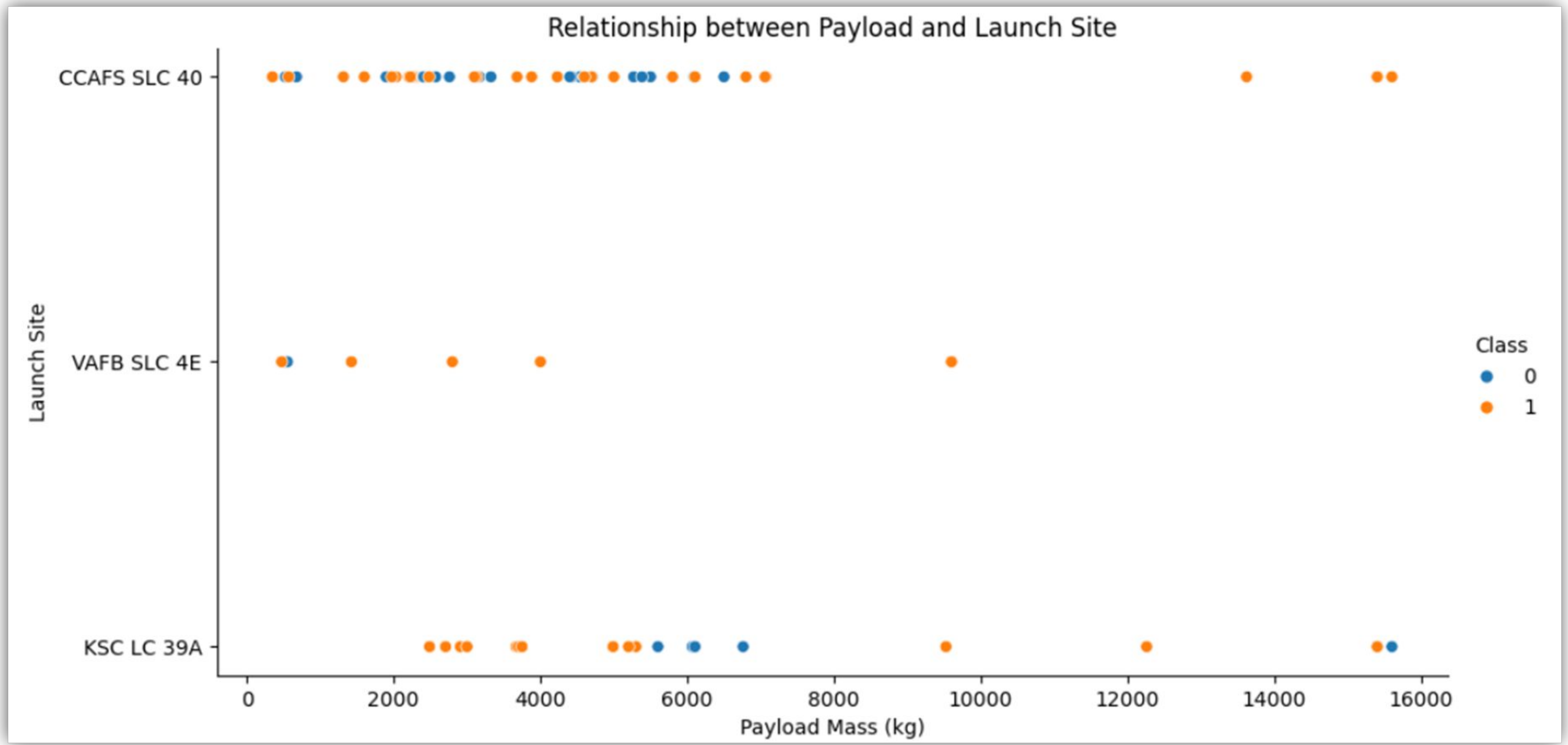


Blue = Fail
Orange = Success

Payload vs. Launch Site

A significant majority of launches involving a payload exceeding 7,000 kg have achieved success.
Launches from KSC LC 39A boasting a payload less than 5,500 kg exhibit a flawless 100% success rate.
However, there is no clear pattern to say the launch site is dependent to the payload mass for the success rate.

Blue = Fail
Orange = Success



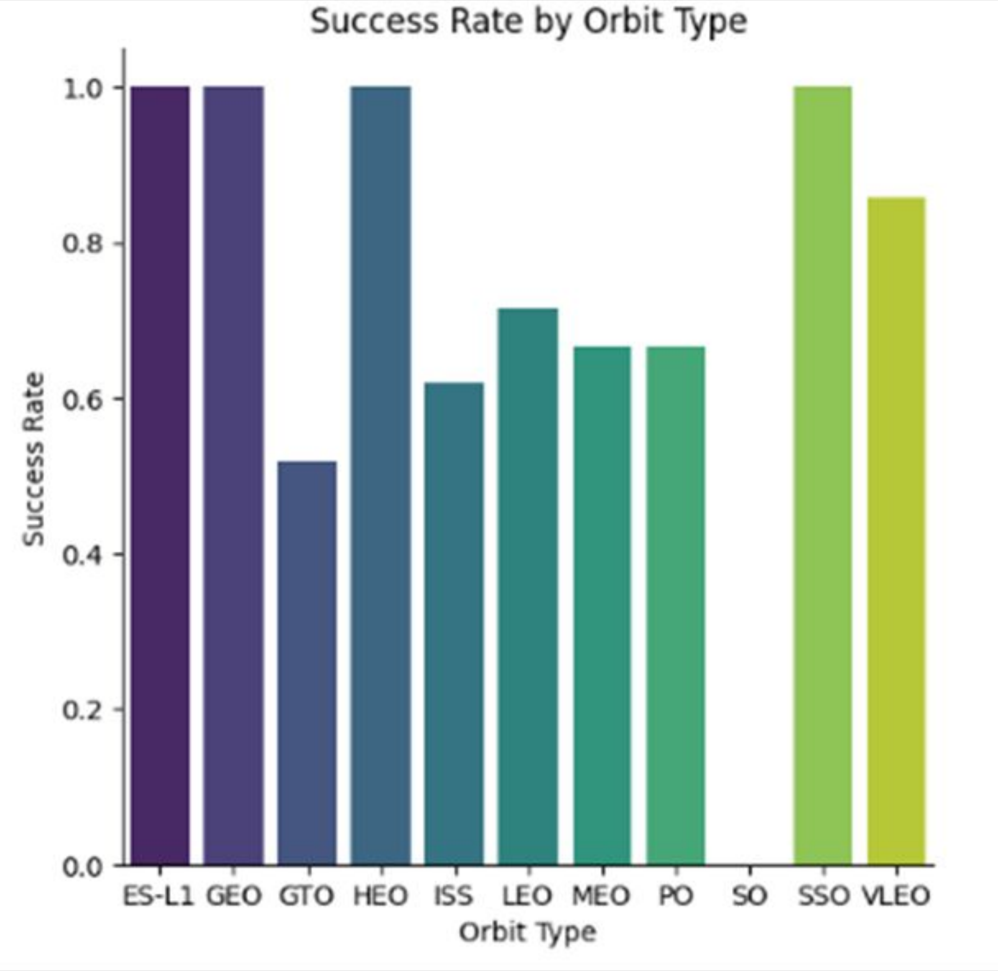
Success Rate vs. Orbit Type

Success Rate

100% = ES L1, GEO, HEO, SSO

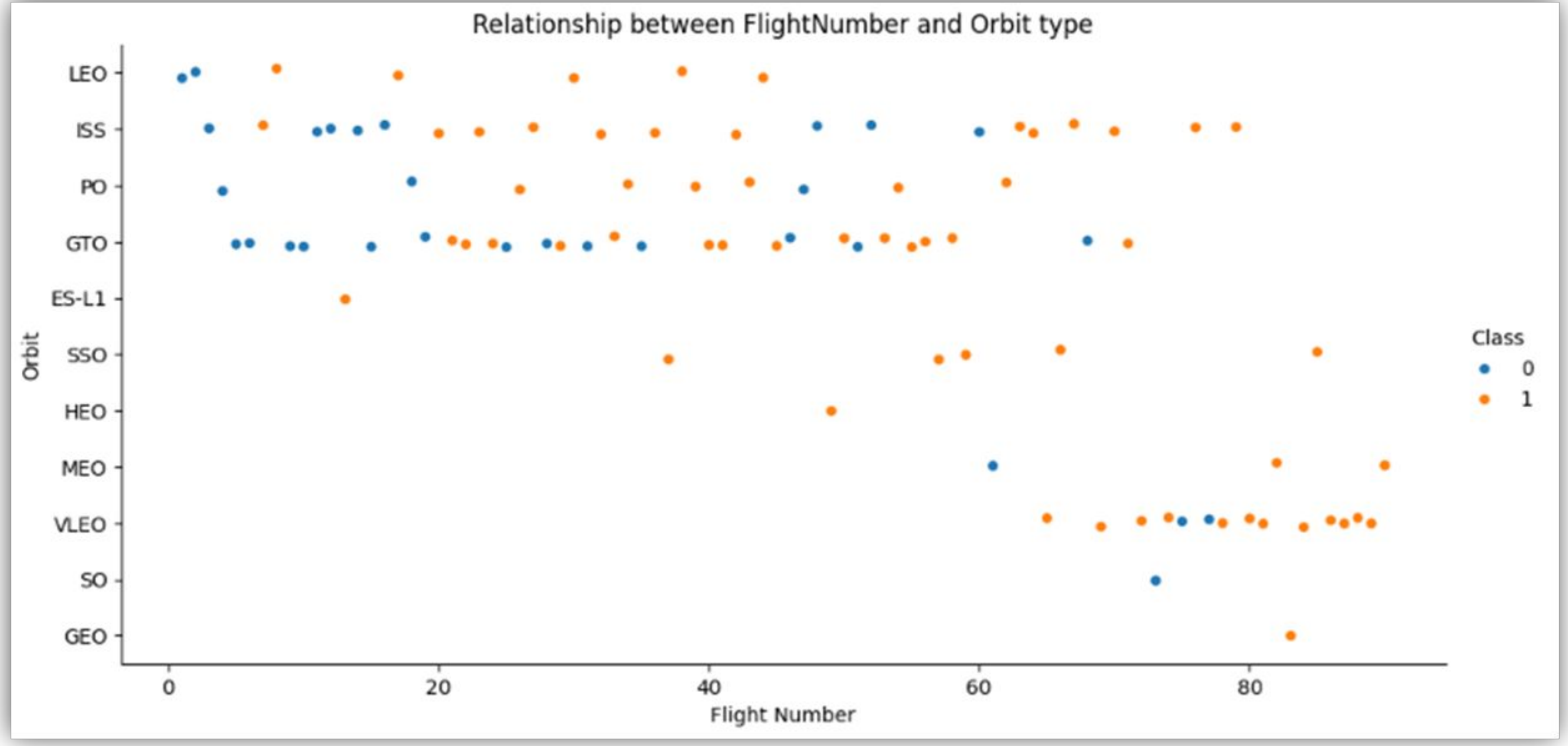
50%-85% = GTO, ISS, LEO, MEO, PO

0% = SO



Flight Number vs. Orbit Type

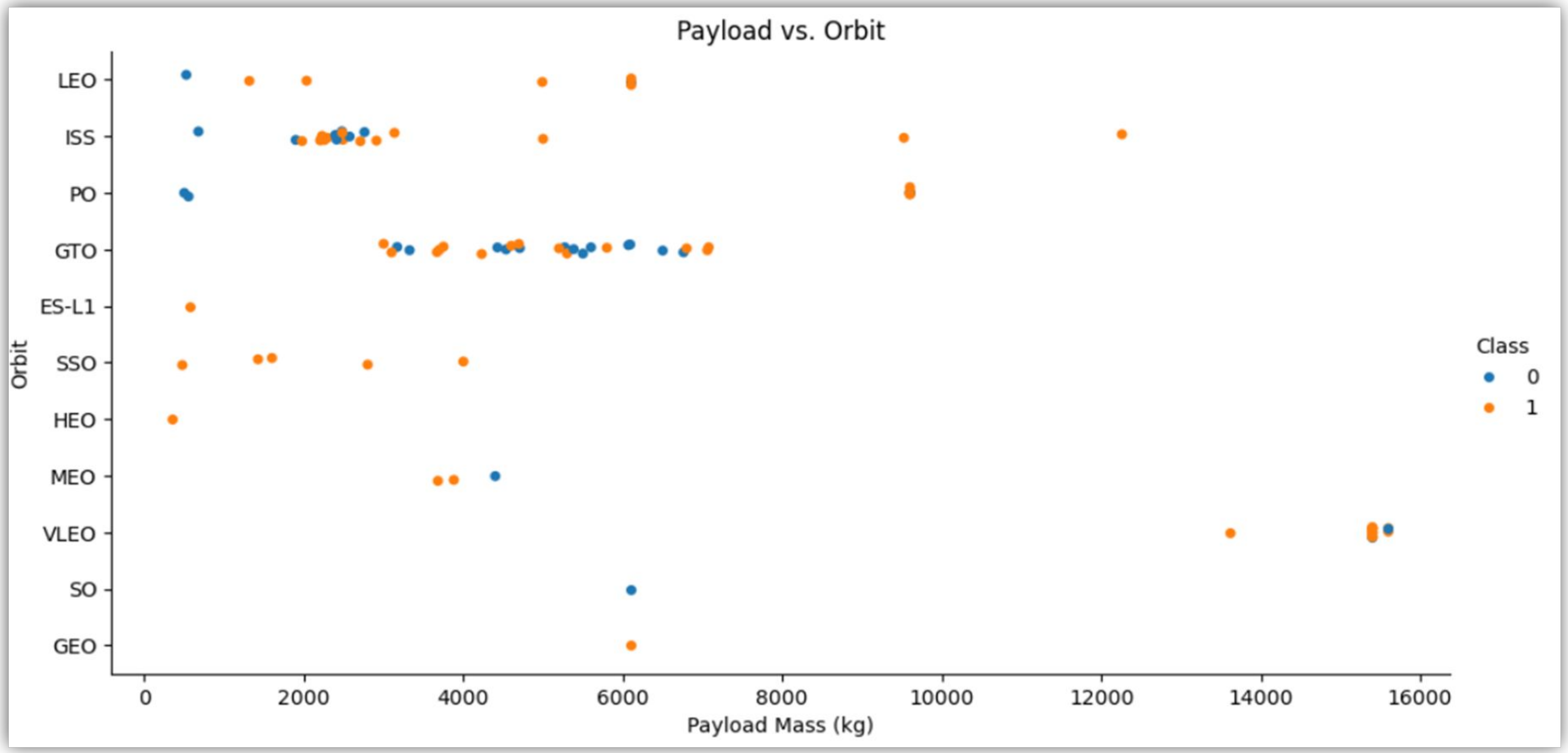
The presented scatter plot illustrates a prevailing trend: as the flight number increases for each orbit, the corresponding success rate tends to elevate, with a particularly noticeable impact in the case of (LEO). However, the (GTO) stands out by showcasing no discernible relationship between the two attributes.



Blue = Fail
Orange = Success

Payload vs. Orbit Type

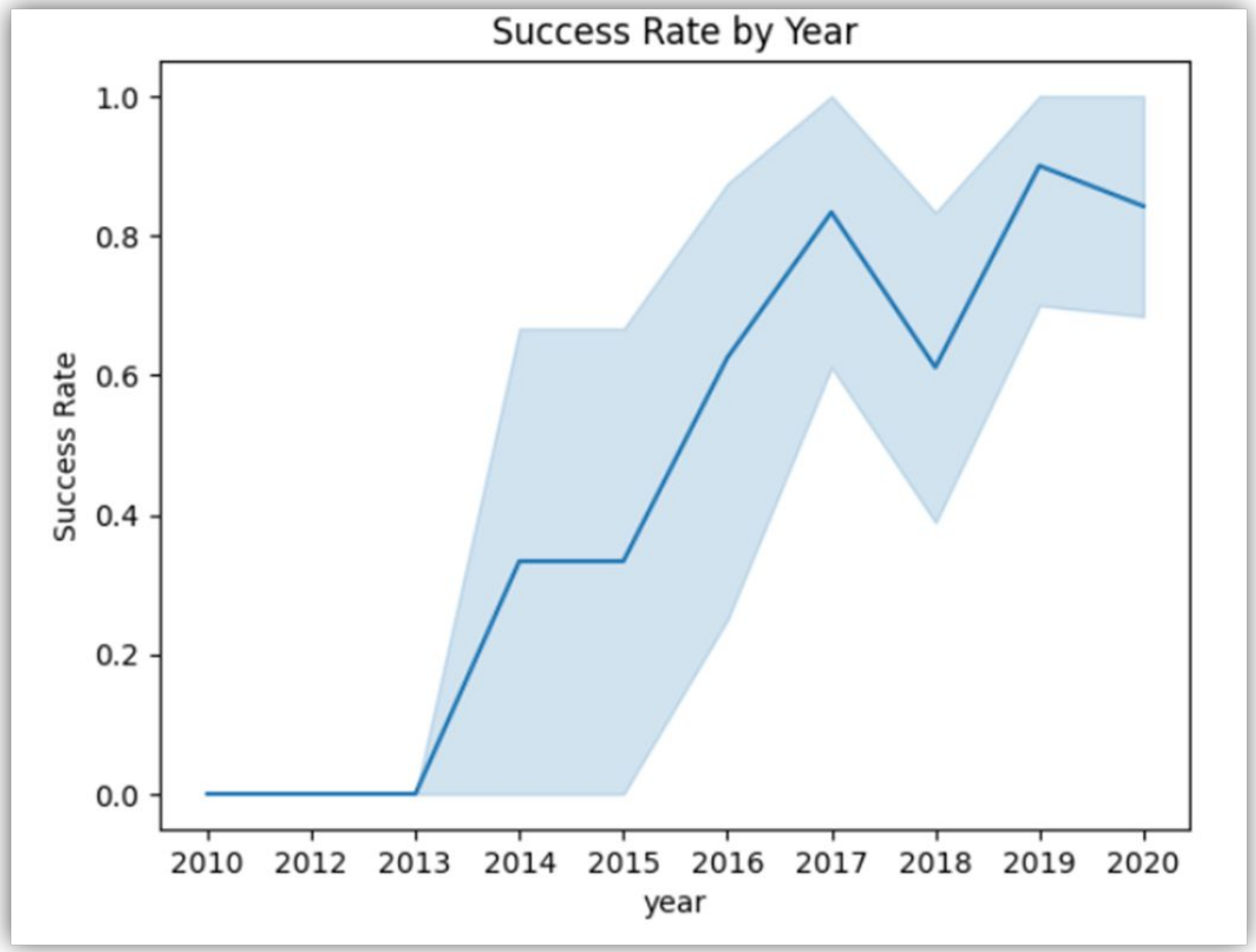
Increased payload weight demonstrates a favorable influence on (LEO), (ISS), and PO orbit. In contrast, it exerts an adverse effect on (MEO) and (VLEO). The (GTO) appears to lack a discernible relationship between the respective attributes. Meanwhile, for (SO), (GEO), and (HEO), a more extensive dataset is required to identify any underlying patterns or trends.



Blue = Fail
Orange = Success

Launch Success Yearly Trend

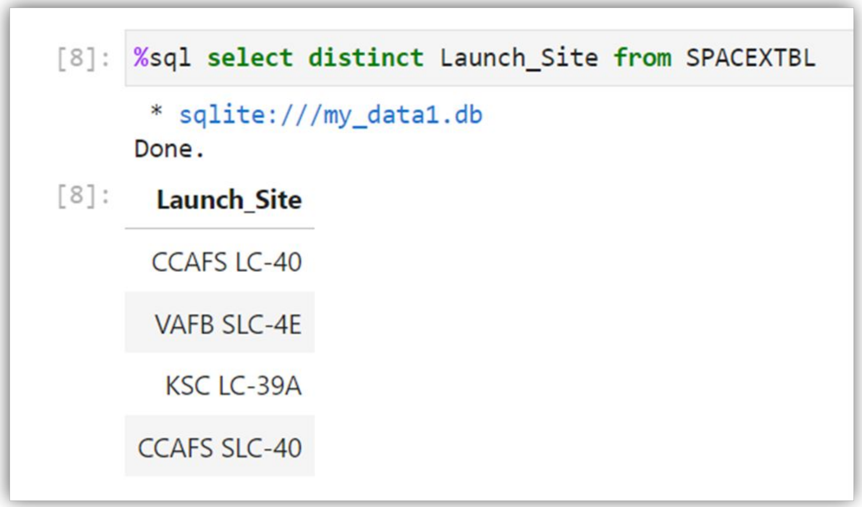
These figures clearly illustrate a rising trend from 2013 to 2020. If this trend persists in the coming years, the success rate will continue to climb steadily, ultimately reaching a 1/100% success rate.



Launch Site Information

We employed the **DISTINCT** keyword to display exclusively the unique launch sites from the SpaceX data.

All Launch Site Names



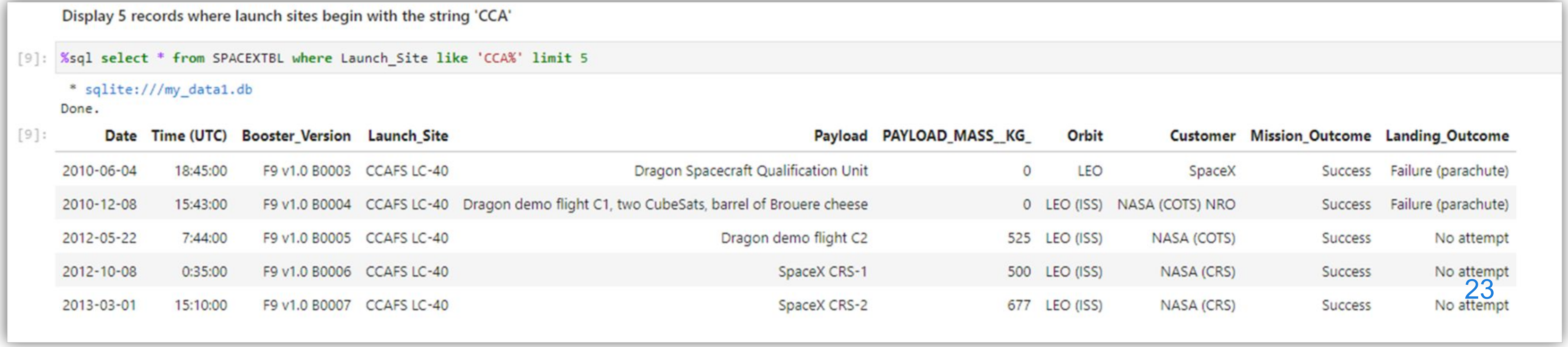
```
[8]: %sql select distinct Launch_Site from SPACEXTBL
```

```
 * 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'



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

```
[9]: %sql select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5
```

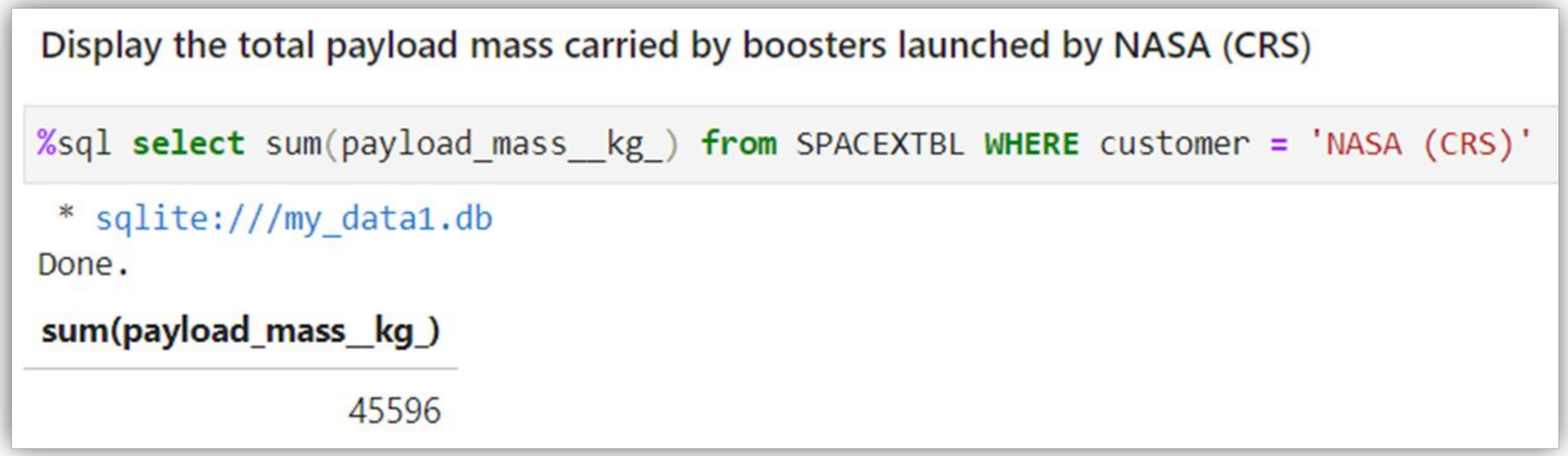
```
 * sqlite:///my_data1.db
Done.
```

[9]:

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

Payload Mass

Total Payload Mass



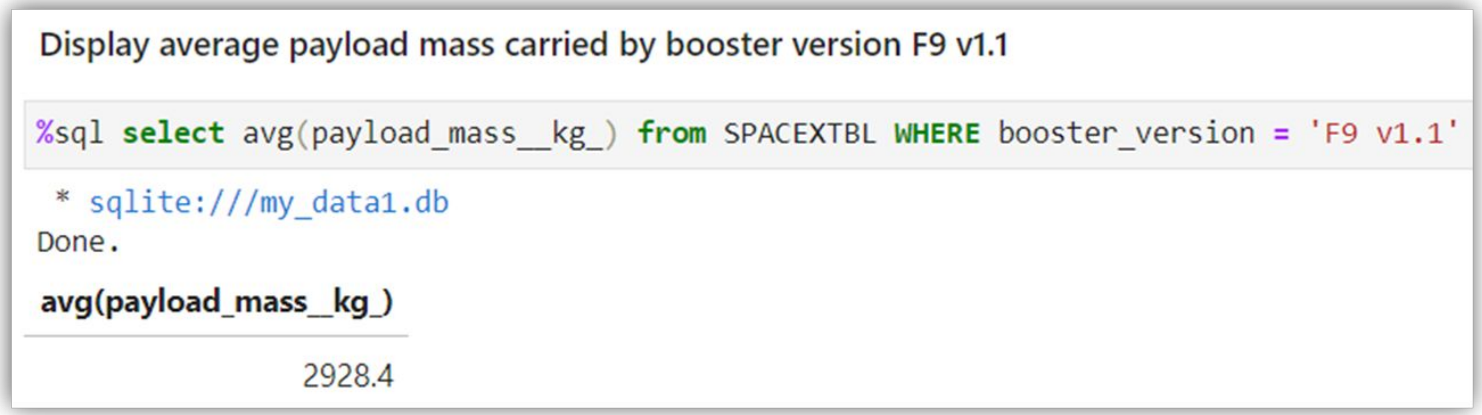
Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) from SPACEXTBL WHERE customer = 'NASA (CRS)'
```

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

sum(payload_mass__kg_)
45596

Average Payload Mass



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

```
%sql select avg(payload_mass__kg_) from SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

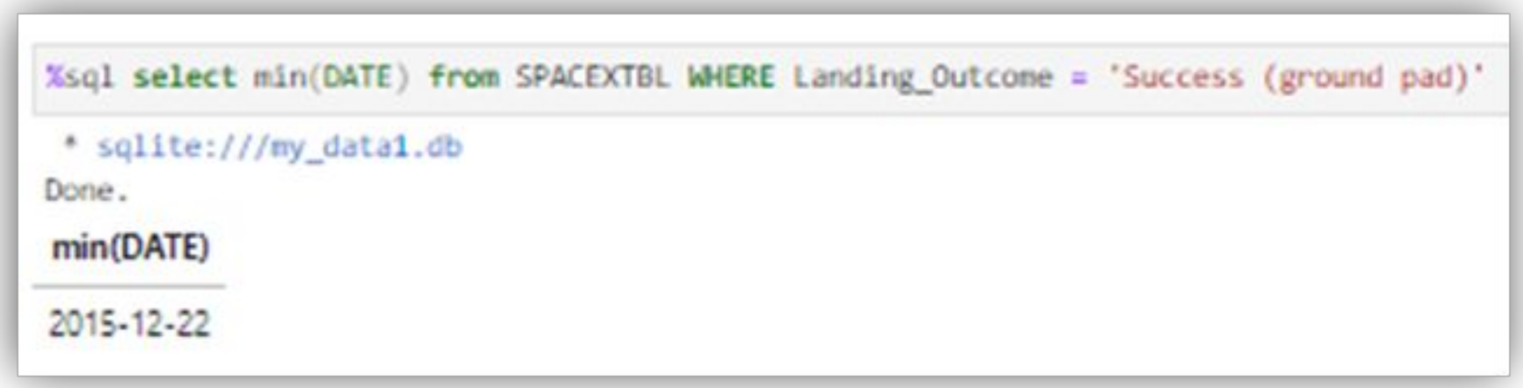
```
 * sqlite:///my_data1.db
Done.
```

avg(payload_mass__kg_)
2928.4

Landing & Mission Info

First Successful Landing in Ground Pad

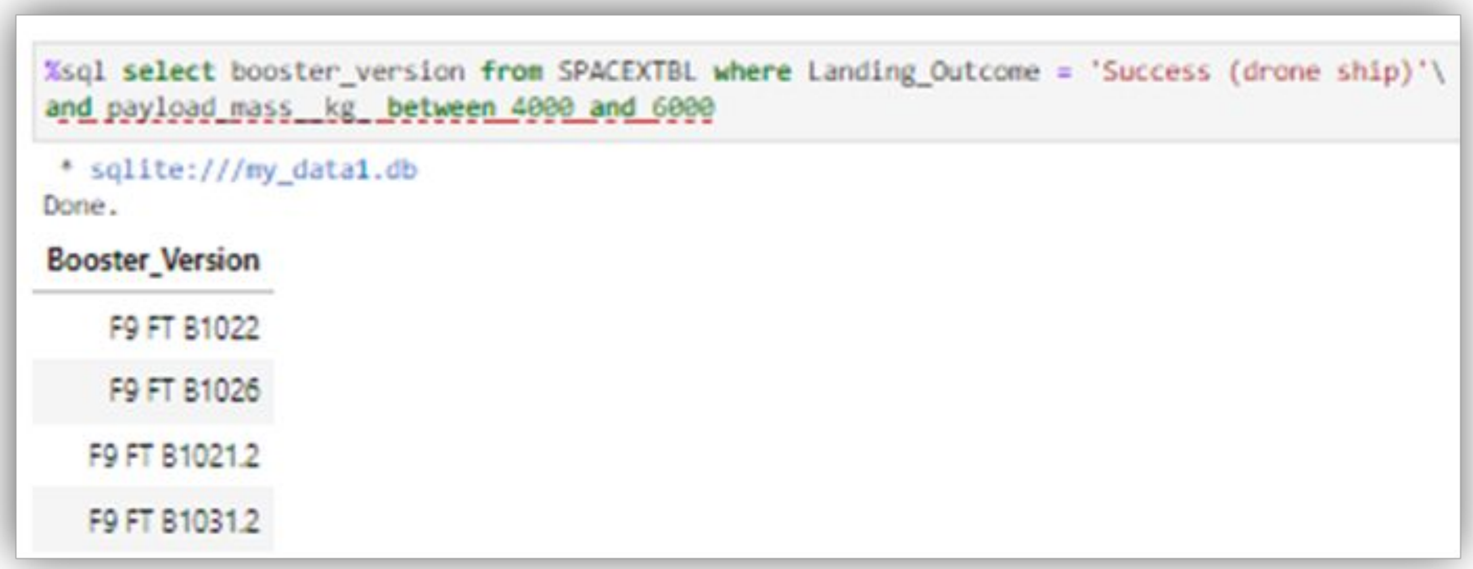
★ *12/22/2015*



```
%sql select min(DATE) from SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

min(DATE)
2015-12-22



```
%sql select booster_version from SPACEXTBL where Landing_Outcome = 'Success (drone ship)'\
and payload_mass__kg_ between 4000 and 6000
```

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

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

Booster Drone Ship Landing

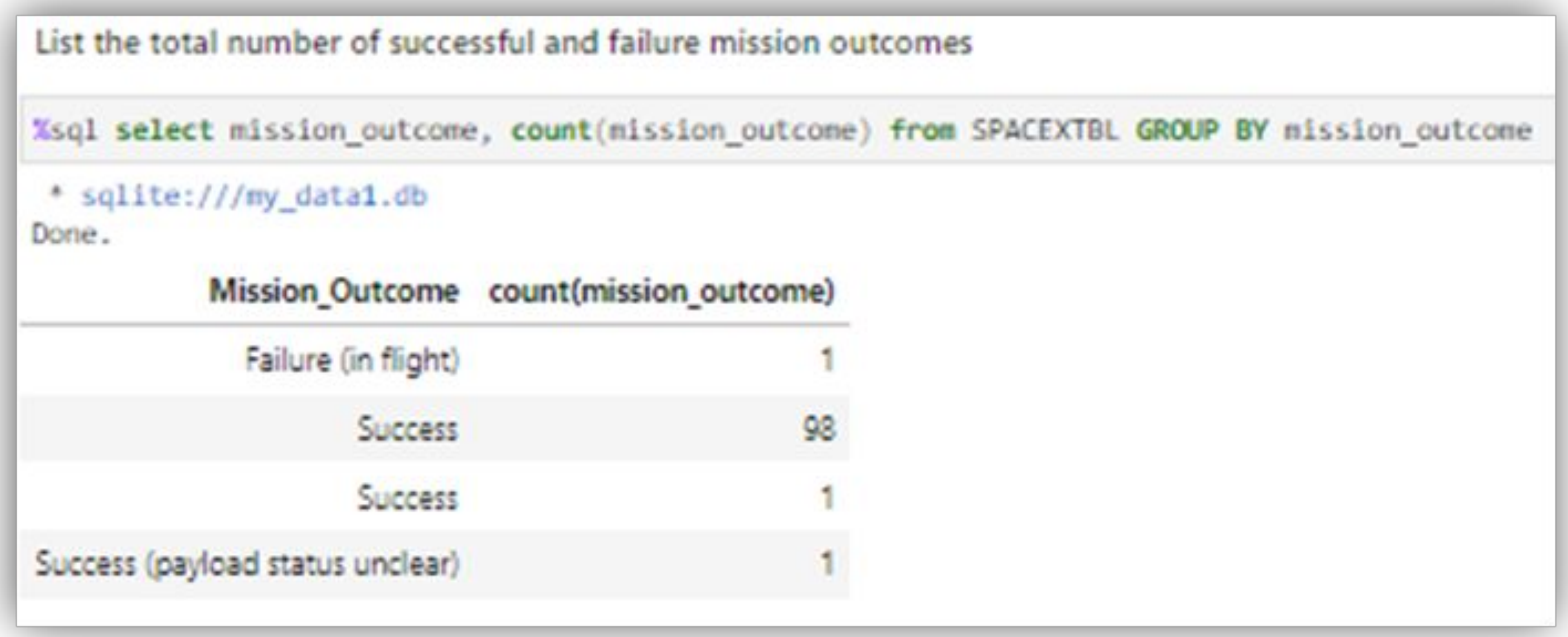
Booster mass greater than 4,000 but less than 6,000

Total Number of Successful and Failure Mission Outcomes

★ *1 Failure in Flight*

★ *99 Success*

★ *1 Success (payload status unclear)*



List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome
```

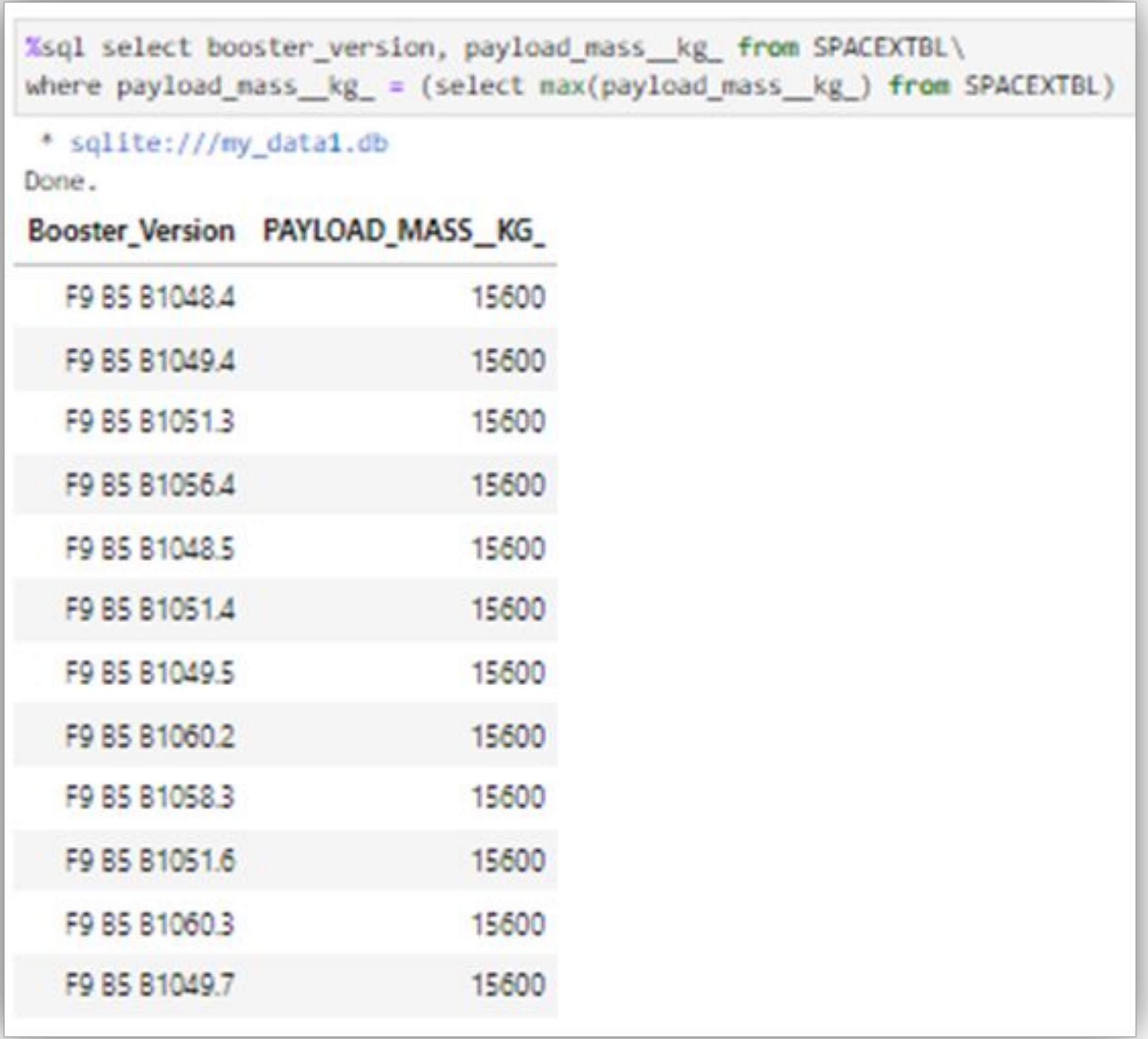
```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	count(mission_outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Carrying Max Payload

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



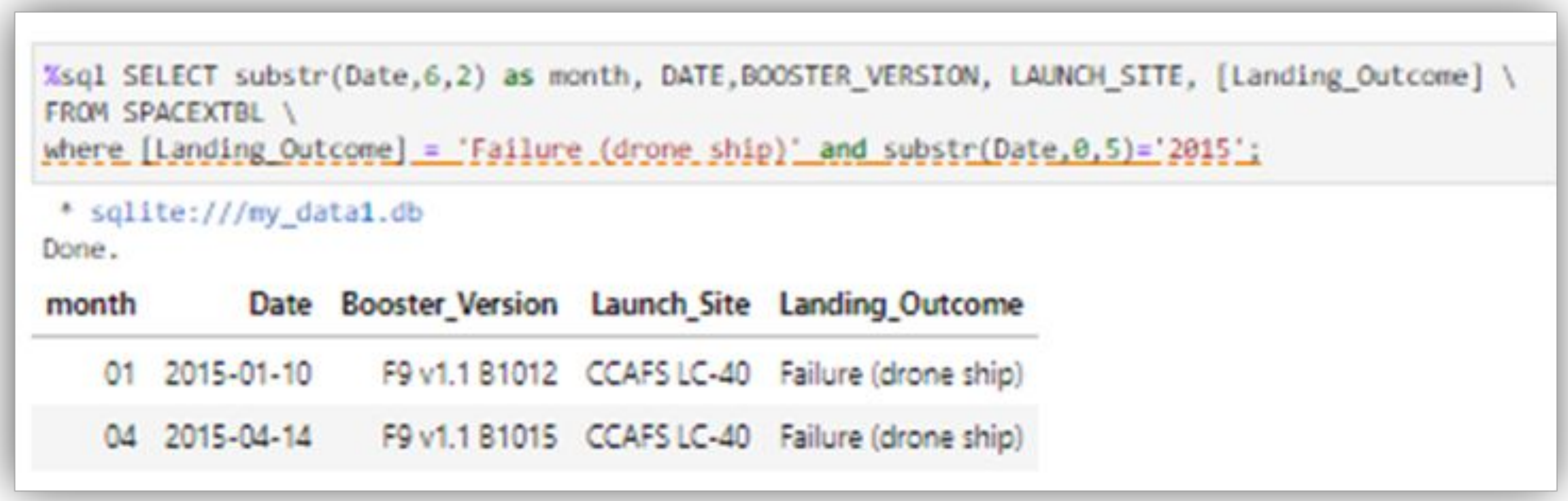
```
%sql select booster_version, payload_mass__kg_ from SPACEXTBL\
where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Showing month, date, booster version, launch site and landing outcome

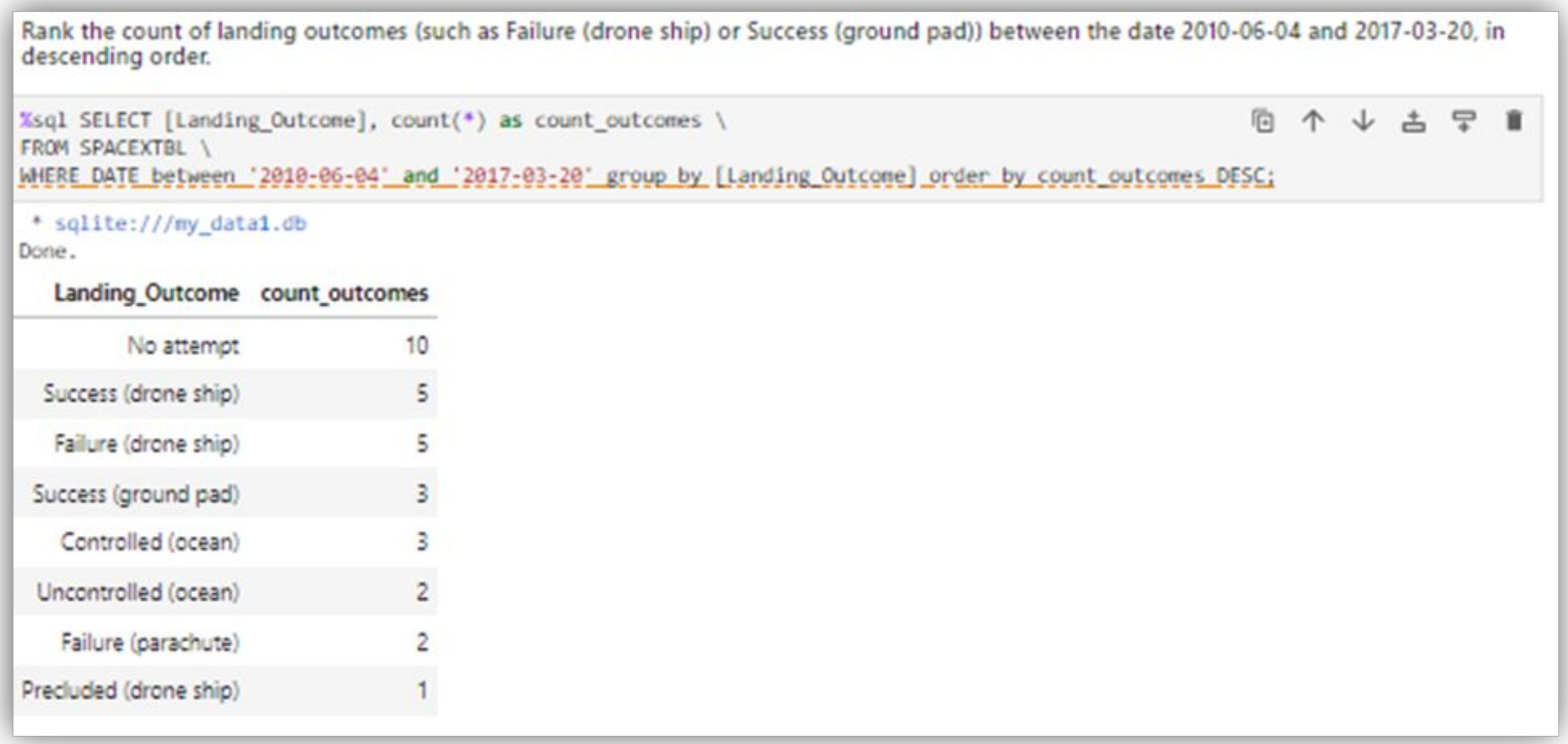


```
%sql SELECT substr(Date,6,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing_Outcome] \
FROM SPACEXTBL \
where [Landing_Outcome] = 'Failure (drone ship)' and substr(Date,0,5)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	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)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql SELECT [Landing_Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

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

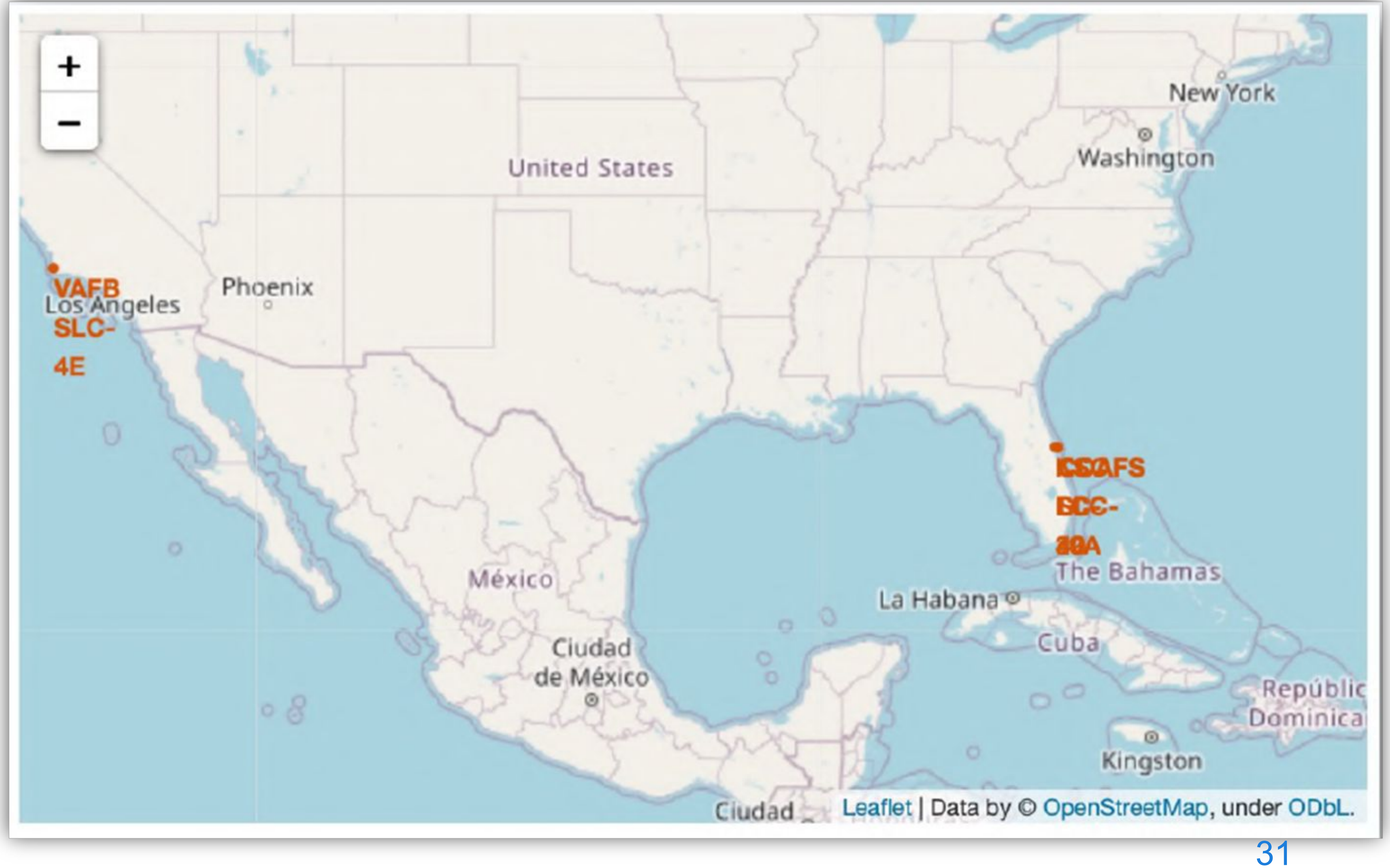
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3
Launch Sites
Proximities Analysis

Location of all the Launch Sites

SpaceX operates three orbital launch platforms:

- Cape Canaveral Air Force Station,
- John F. Kennedy Space Center
- Vandenberg Air Force Base.



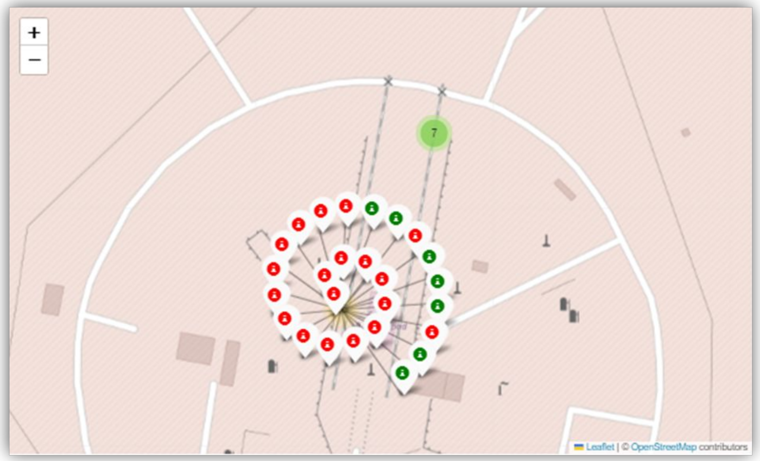
Location of all the Launch Sites

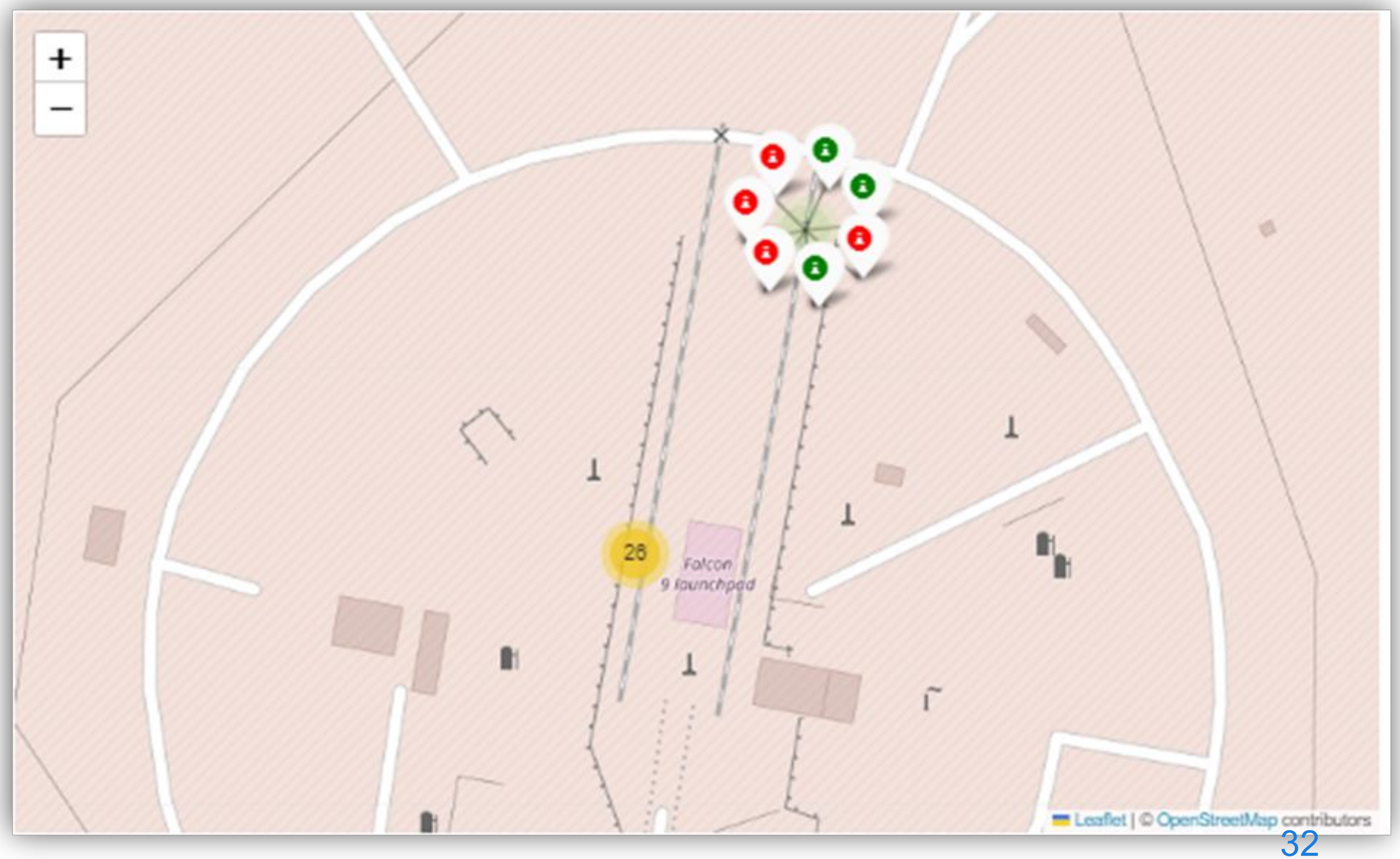
At Each Launch Site Outcomes

Green: markers for successful launches

Red: markers for unsuccessful launches

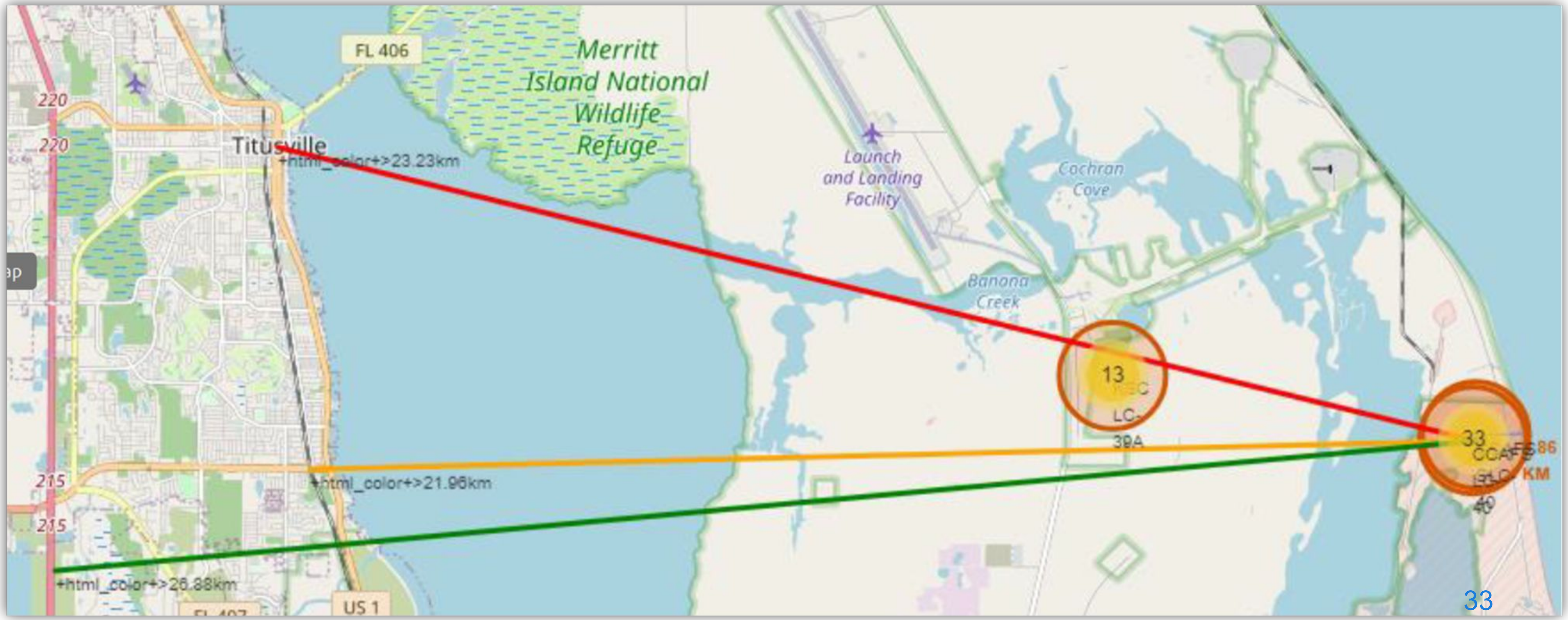
Launch site: CCAFS SLC 40 has a 3/7 success rate 42.9%)





Distance to Proximities

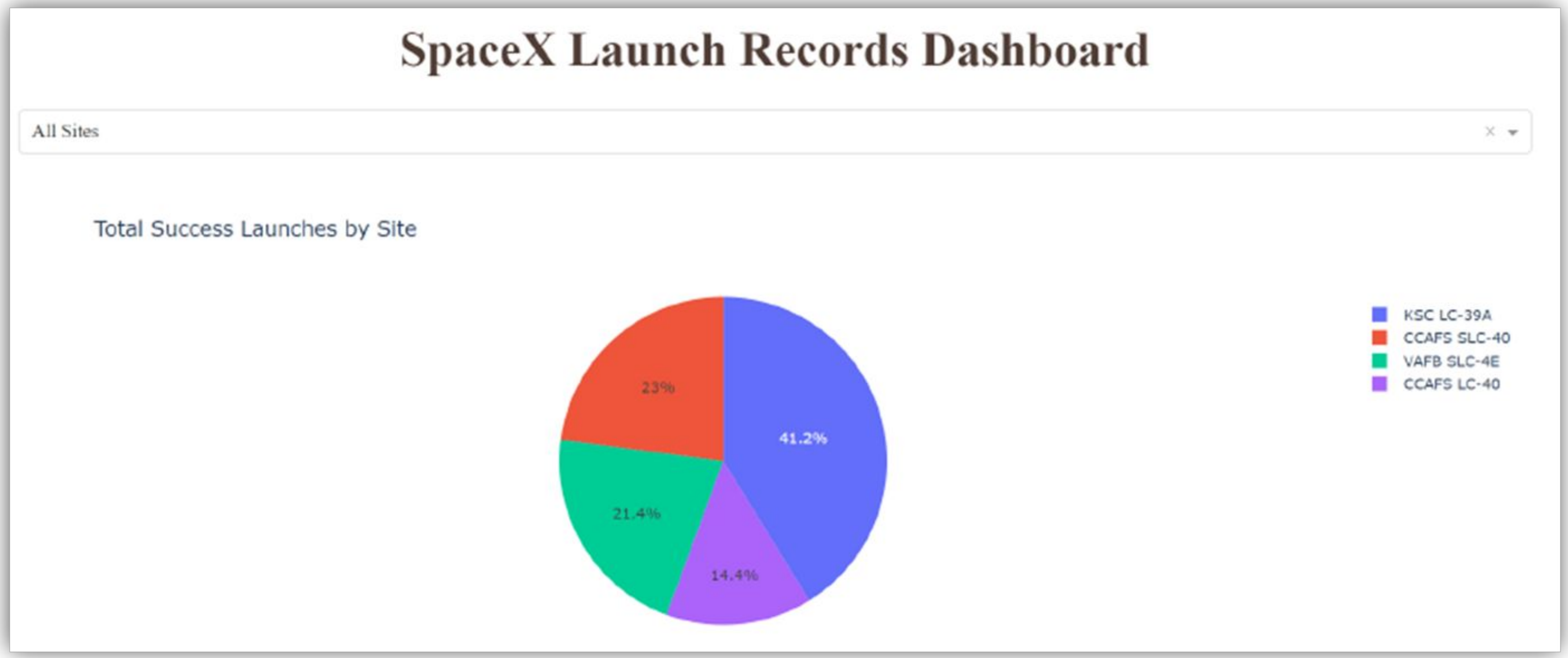
CCAFS SLC 40



Section 4
Build a Dashboard with Ploty Dash

Success percentage by each sites.

Kennedy Space Center Launch Complex 39A (KSC LC-39A) boasts the highest success rate among launch sites, achieving a remarkable 41.2% success rate.

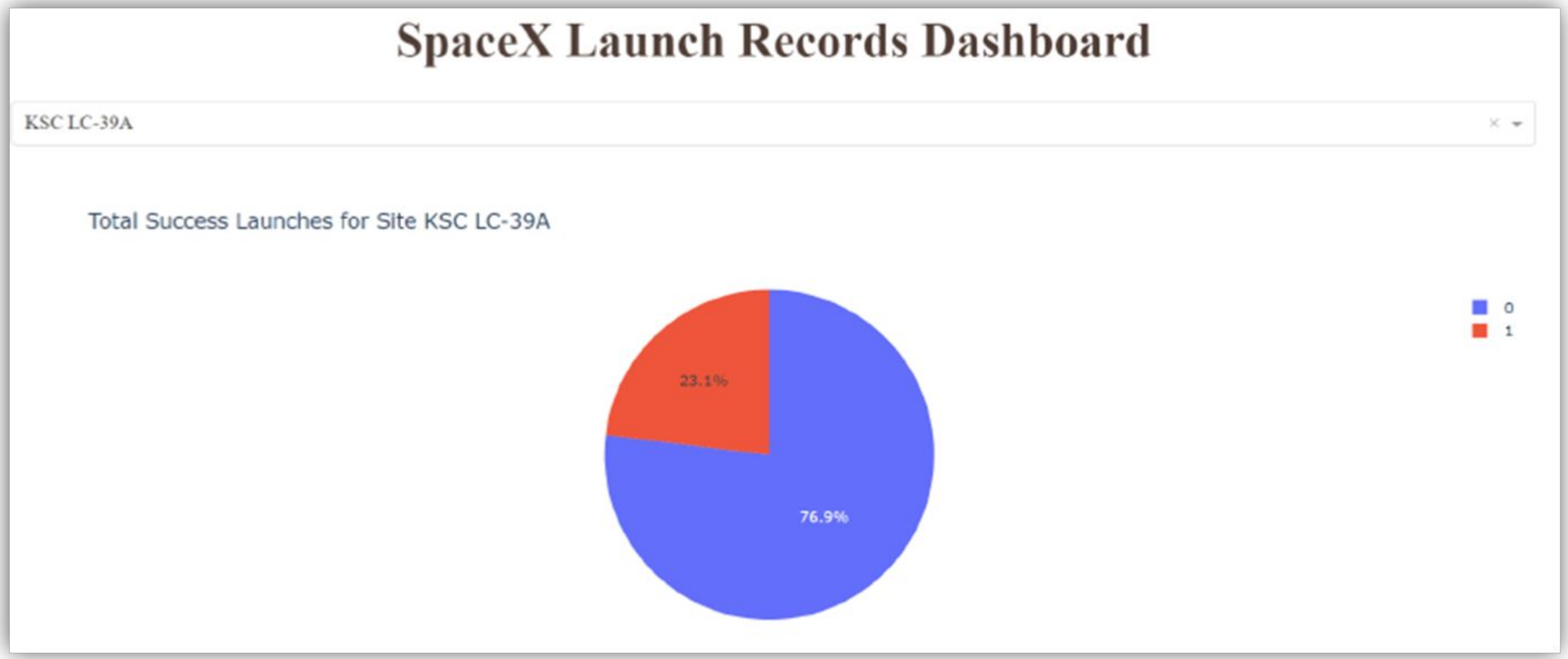


SpaceX Launch Records Dashboard

All Sites × ▾

Highest launch-success ratio

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



SpaceX Launch Records Dashboard

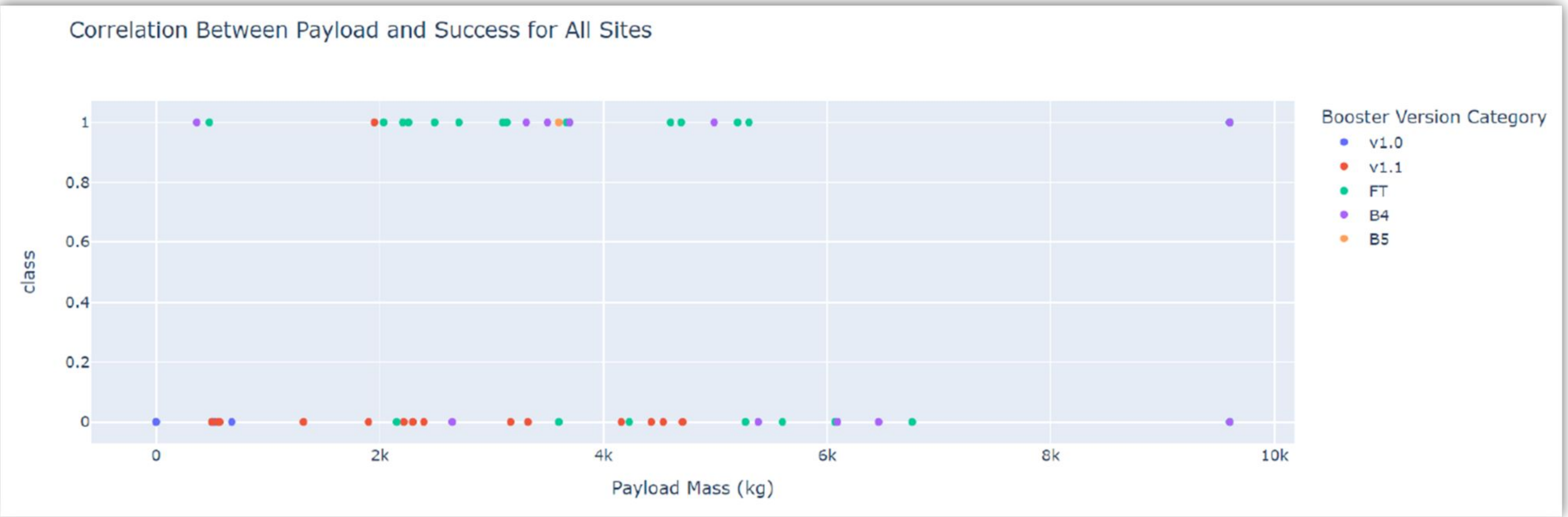
KSC LC-39A

Total Success Launches for Site KSC LC-39A

Payload vs. Launch Outcome

The most successful missions occur with payloads weighing between 2,000 kg and 5,000 kg. rate.

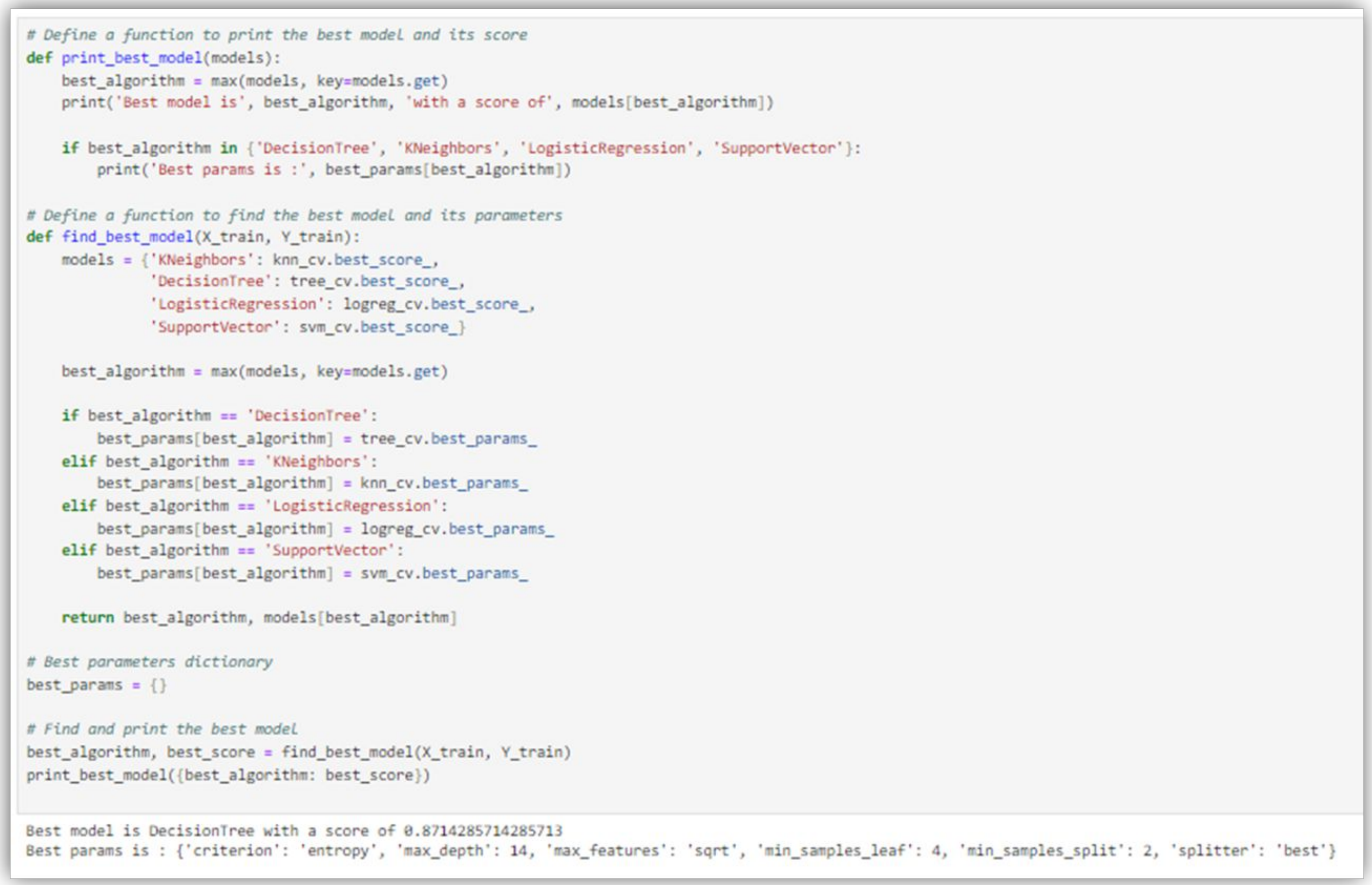
0 = Fail
1 = Success



Section 5
Predictive Analysis (Classification)

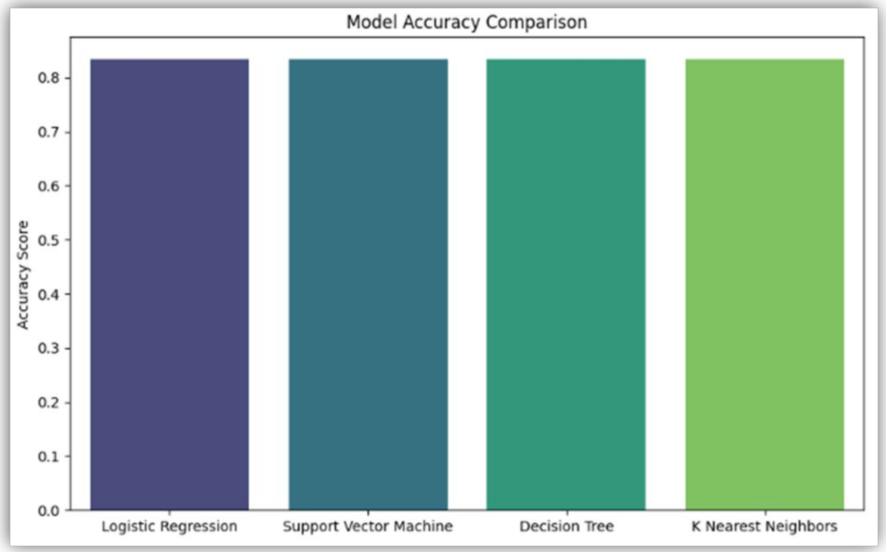
Classification Accuracy

Using the code provided below, we can observe that the Tree Algorithm achieved the highest classification accuracy, identifying it as the best-performing algorithm.



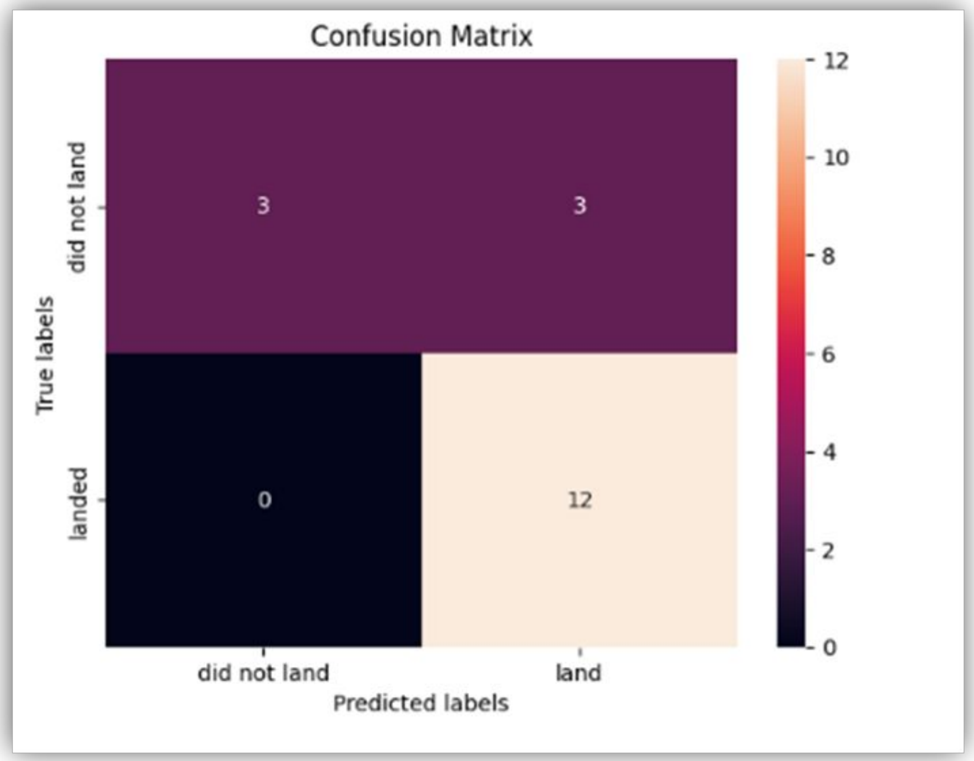
```python
# Define a function to print the best model and its score
def print_best_model(models):
    best_algorithm = max(models, key=models.get)
    print('Best model is', best_algorithm, 'with a score of', models[best_algorithm])

    if best_algorithm in {'DecisionTree', 'KNeighbors', 'LogisticRegression', 'SupportVector'}:
        print('Best params is :', best_params[best_algorithm])

# Define a function to find the best model and its parameters
def find_best_model(X_train, Y_train):
    models = {'KNeighbors': knn_cv.best_score_,
              'DecisionTree': tree_cv.best_score_,
              'LogisticRegression': logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

    best_algorithm = max(models, key=models.get)

    if best_algorithm == 'DecisionTree':
        best_params[best_algorithm] = tree_cv.best_params_
    elif best_algorithm == 'KNeighbors':
        best_params[best_algorithm] = knn_cv.best_params_
    elif best_algorithm == 'LogisticRegression':
        best_params[best_algorithm] = logreg_cv.best_params_
    elif best_algorithm == 'SupportVector':
        best_params[best_algorithm] = svm_cv.best_params_

    return best_algorithm, models[best_algorithm]

# Best parameters dictionary
best_params = {}

# Find and print the best model
best_algorithm, best_score = find_best_model(X_train, Y_train)
print_best_model({best_algorithm: best_score})
```

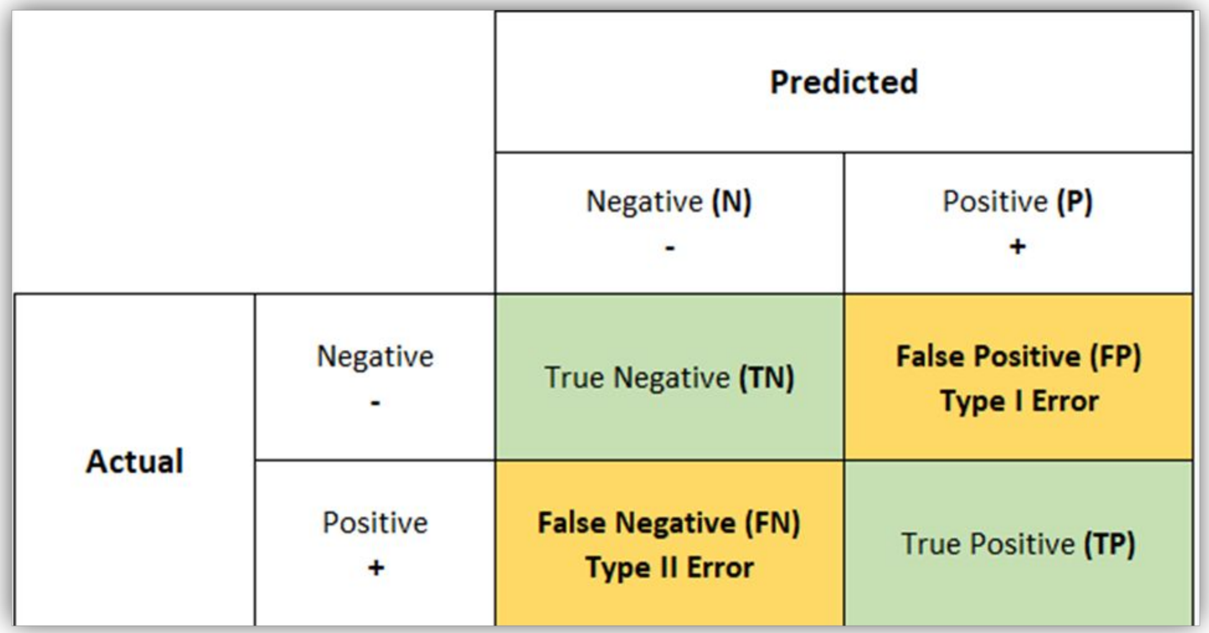
```
Best model is DecisionTree with a score of 0.8714285714285713
Best params is : {'criterion': 'entropy', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'splitter': 'best'}
```



Classification Accuracy

The confusion matrix of the decision tree classifier reveals its ability to differentiate between various classes. However, a notable challenge lies in the occurrence of false positives, where instances of unsuccessful landings are incorrectly identified as successful landings by the classifier.





		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Conclusion

Key Findings:

- Launch Success Trends: Over time, there is a notable increase in launch success rates.
- KSC LC 39A Site: This launch site exhibits the highest success rate 76.9%, particularly for launches below 5,500 kg.
- Orbit Success Rates: ES L1, GEO, HEO, and SSO orbits consistently achieve a 100% success rate.
- Payload Mass Impact: Higher payload masses correlate with higher success rates across all launch sites.

Considerations for Future Research:

- Dataset Expansion: A larger dataset could enhance predictive analytics and generalize findings.
- Feature Analysis: Conducting additional feature analysis or principal component analysis may improve model accuracy.

Key Conclusions:

- Best Algorithm: The Tree Classifier Algorithm proves to be the most effective for this dataset. In analyzing the performance of various machine learning models, the decision tree model emerged as a slightly superior choice on the test set. The geographical distribution of launch sites, with a focus on proximity to the equator and coast, contributes to the overall success of launches.
- Success Rate Trend: SpaceX's success rate has shown consistent improvement since 2013, indicating a positive trajectory for future launches.

Thank you!