IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Luu Vi

25 February 2025

Outline

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

Executive Summary

In this capstone project, we will use various machine learning classification algorithms to predict whether the SpaceX Falcon 9 first stage will land successfully. The key steps in this project are:

1. **Data Collection, Wrangling, and Formatting**
2. **Exploratory Data Analysis**
3. **Interactive Data Visualization**
4. **Machine Learning Prediction**

Our analysis reveals that certain features of the rocket launches are correlated with the success or failure of the landings. Based on the findings, we conclude that the decision tree algorithm might be the most effective machine learning model for predicting whether the Falcon 9 first stage will land successfully.

Introduction

- In this capstone project, the goal is to predict whether the first stage of a Falcon 9 rocket will land successfully. SpaceX lists the cost of a Falcon 9 launch at $62 million, significantly lower than other providers, whose launches start at over $165 million. A major factor in this cost difference is SpaceX’s ability to reuse the first stage. By predicting whether the first stage will land, we can estimate the cost of the launch. This information can be valuable for other companies bidding against SpaceX for rocket launch contracts.
- It’s important to note that most unsuccessful landings are planned events. Occasionally, SpaceX may carry out a controlled landing in the ocean.
- The primary question we aim to answer is: given a set of features, such as payload mass, orbit type, launch site, and more, can we predict whether the first stage of the Falcon 9 rocket will land successfully?

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
  - Standard scaler was applied to numerical features for normalization
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K-nearest neighbors (KNN)

Data Collection

Data Collection Process:

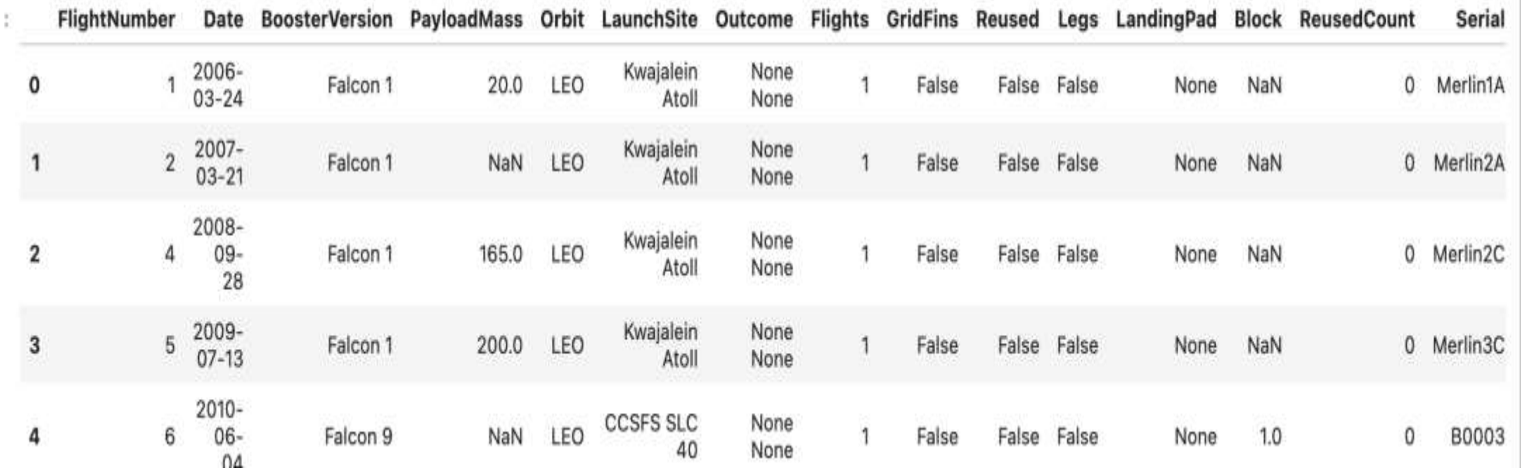
- **SpaceX API Data Retrieval:**
    - Data was collected using a GET request to the SpaceX API.
    - The response content was decoded as JSON using the .json() function.
    - The JSON data was converted into a pandas dataframe using .json_normalize().
- **Data Cleaning:**
    - The dataset was cleaned by checking for missing values.
    - Missing values were filled in as needed.
- **Web Scraping:**
    - Web scraping was performed on Wikipedia for Falcon 9 launch records using BeautifulSoup.
    - The objective was to extract the launch records from an HTML table.
    - The table was parsed and converted into a pandas dataframe for future analysis.

Data Collection

SpaceX API:

- **API Used:** https://api.spacexdata.com/v4/rockets/
- The API provides data on various SpaceX rocket launches.
- The data is filtered to focus only on Falcon 9 launches.
- Missing values are filled with the mean of the respective column.
- The dataset consists of:
  - **90 instances (rows)**
  - **17 features (columns)**

The image below shows the first few rows of the data.



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003

Github

https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B1%5Djupyter-labs-spacex-data-collection-api-v2.ipynb

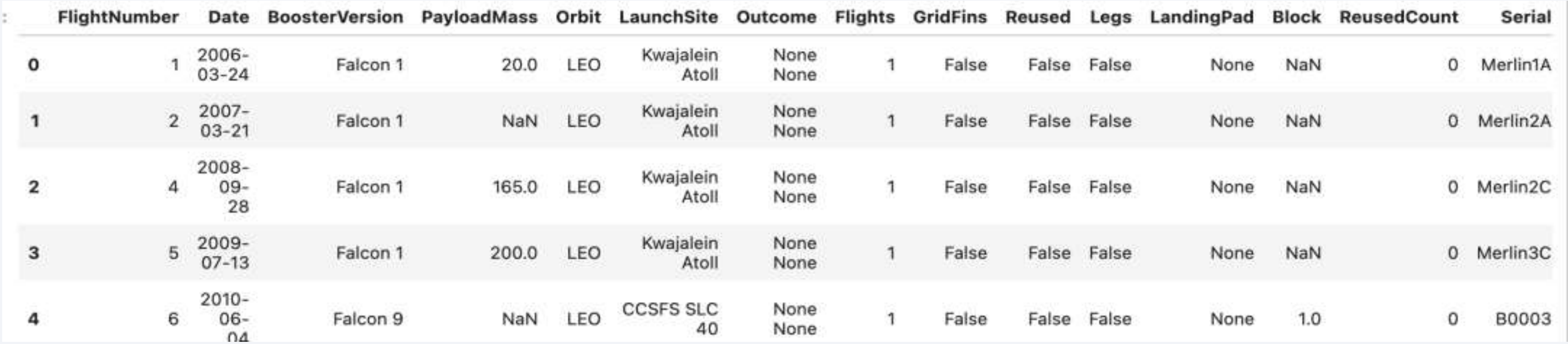
Data Collection – SpaceX API

SpaceX API:

- **API Used:** https://api.spacexdata.com/v4/rockets/
- The API provides data on various SpaceX rocket launches.
- The data is filtered to focus only on Falcon 9 launches.
- Missing values are filled with the mean of the respective column.
- The dataset consists of:
    - **90 instances (rows)**
    - **17 features (columns)**

The image below shows the first few rows of the data.

Github https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B1%5Djupyter-labs-spacex-data-collection-api-v2.ipynb



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003

Data Collection - Scraping

Web Scraping:

•**Source:** Data was scraped from [Wikipedia - Falcon 9 Launches]

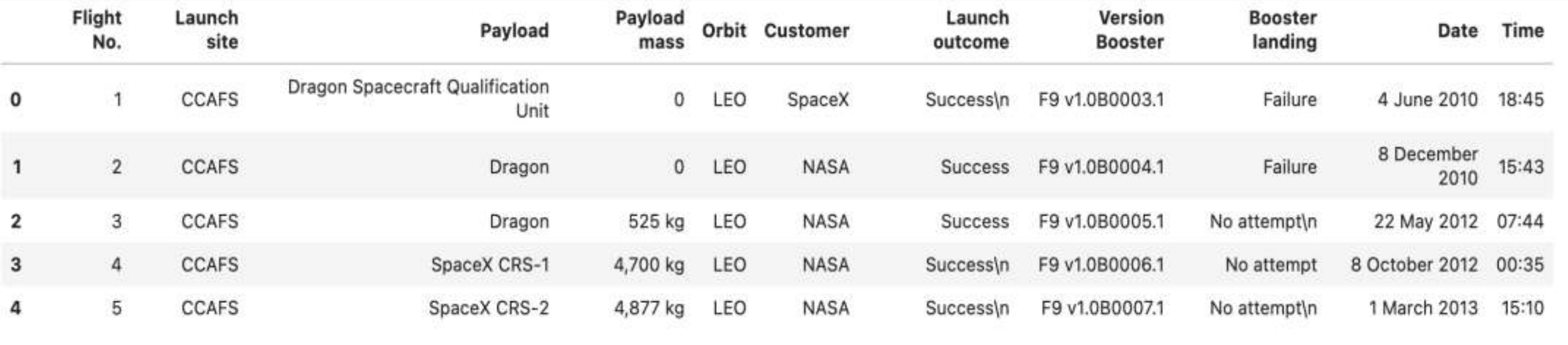
•The website contains data only about Falcon 9 launches.

•The dataset consists of:

- **121 instances (rows)**
- **11 features (columns)**

The image below shows the first few rows of the data.

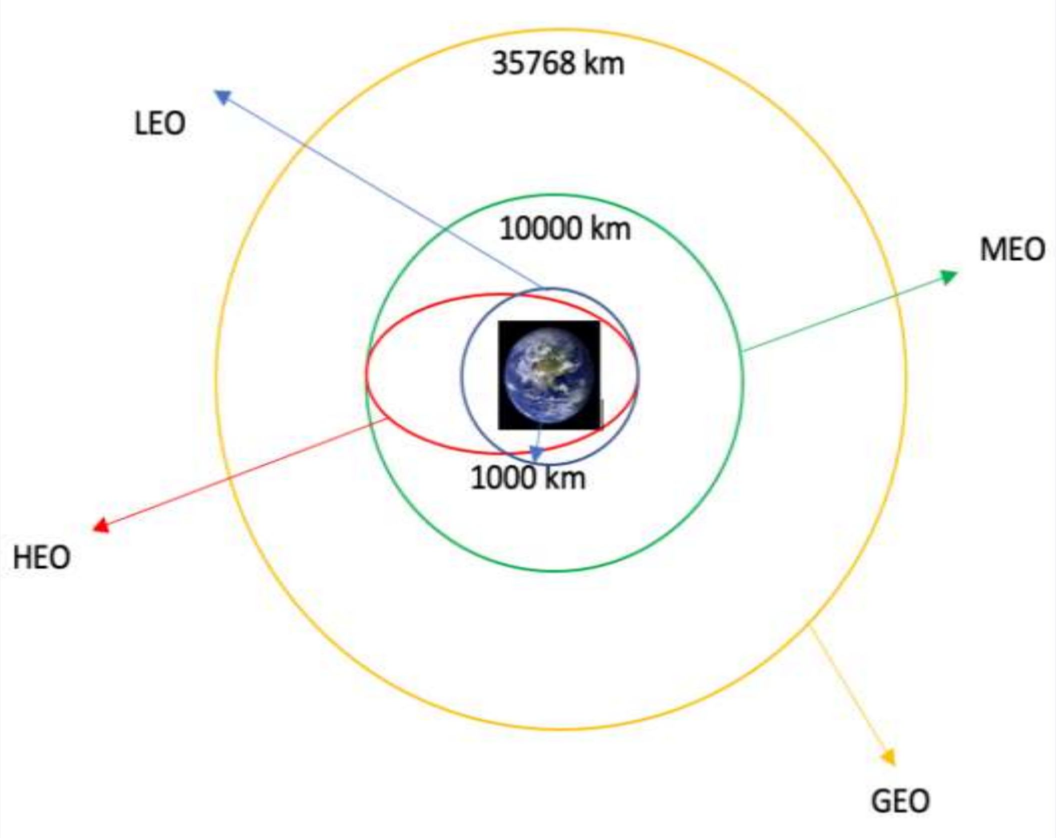
Github
https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B2%5Djupyter-labs-webscraping.ipynb



	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.
- Github: https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B3%5Dlabs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with SQL

Pandas and NumPy:

•Functions from **Pandas** and **NumPy** were used to derive basic information about the data, including:

- Number of launches at each launch site.
- Occurrence of each orbit type.
- Frequency of each mission outcome.

SQL:

•SQL queries were used to answer key questions, such as:

- The names of unique launch sites in the space mission dataset.

Github:
https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B4%5Djupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with Data Visualization

Matplotlib and Seaborn:

- **Matplotlib** and **Seaborn** were used to visualize the data through various charts:
    - Scatterplots, bar charts, and line charts were created to understand relationships between features, including:
        - The relationship between flight number and launch site.
        - The relationship between payload mass and launch site.
        - The relationship between success rate and orbit type.

Github:
https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B5%5Djupyter-labs-eda-dataviz-v2.ipynb

Folium:

- Functions from the **Folium** library were used to create interactive maps:
    - Mark all launch sites on the map.
    - Mark successful and failed launches for each site.
    - Show the distances from launch sites to nearby cities, railways, or highways.

Additional Data Insights:

- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by the F9 v1.1 booster version.

Github:
https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B6%5Dlab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Dash:

- Functions from Dash are used to create an interactive site with:
  - **Dropdown menu** for toggling input.
  - **Range slider** for adjusting data.

Visualizations:

- **Pie Chart** shows the total successful launches from each launch site.
- **Scatterplot** illustrates the correlation between payload mass and mission outcome (success or failure) for each launch site.
- Github: https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B7%5D%20dashboard.py

Predictive Analysis (Classification)

Scikit-learn Library:

•Functions from Scikit-learn are used to create machine learning models.

Machine Learning Prediction Phase:

1.Data Standardization:

   1. Standardizing the data before training.

2.Data Splitting:

   1. Splitting the data into training and test sets.

3.Creating Models:

   1. Logistic Regression
   2. Support Vector Machine (SVM)
   3. Decision Tree
   4. K Nearest Neighbors (KNN)

4.Model Training:

   1. Fit the models on the training set.

5.Hyperparameter Tuning:

   1. Find the best combination of hyperparameters for each model.

6.Model Evaluation:

   1. Evaluate models based on accuracy scores and confusion matrix.

Github:
https://github.com/LuuVi2911/IBM-Data-Science-Capstone/blob/main/%5B8%5DSpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

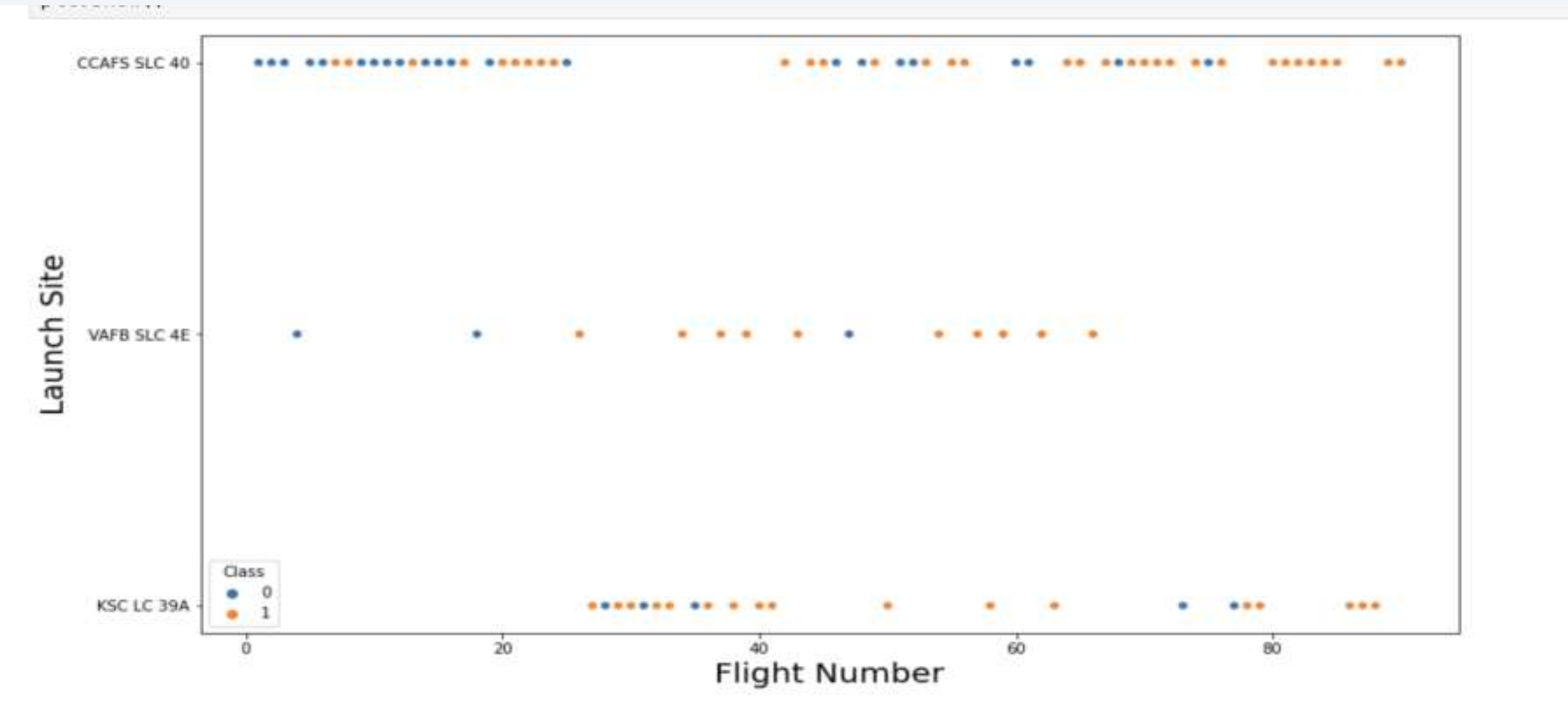
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

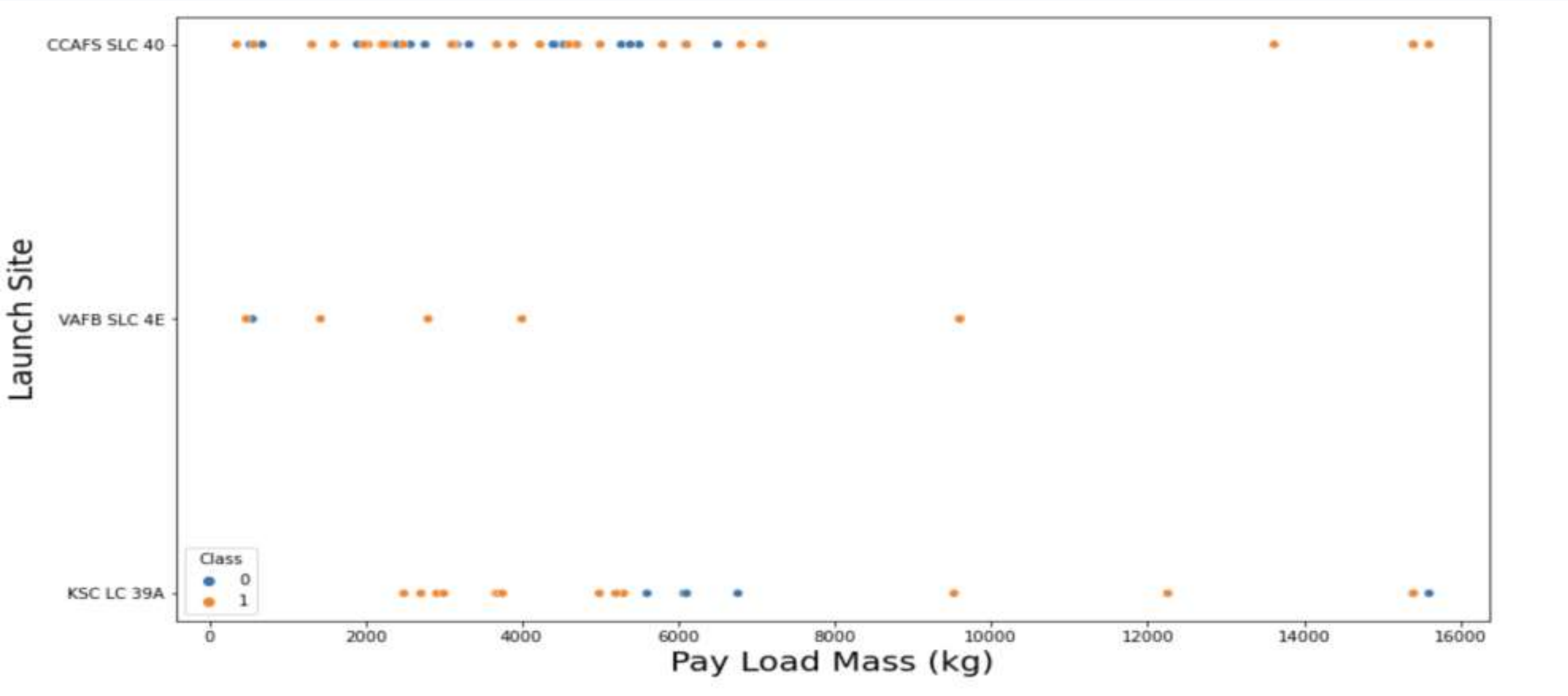
Flight Number vs. Launch Site

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



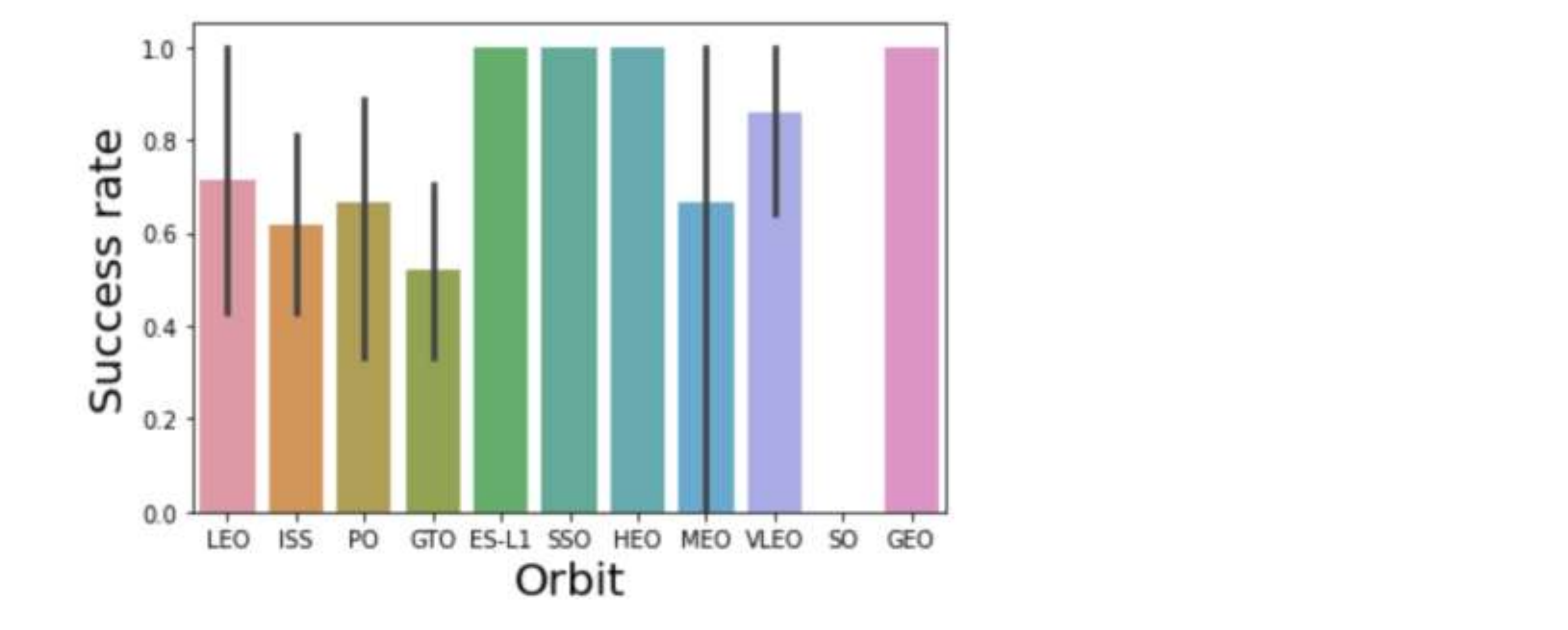
Payload vs. Launch Site

- The greater the payload mass for launches at **CCAFS SLC 40**, the higher the success rate of the rocket.



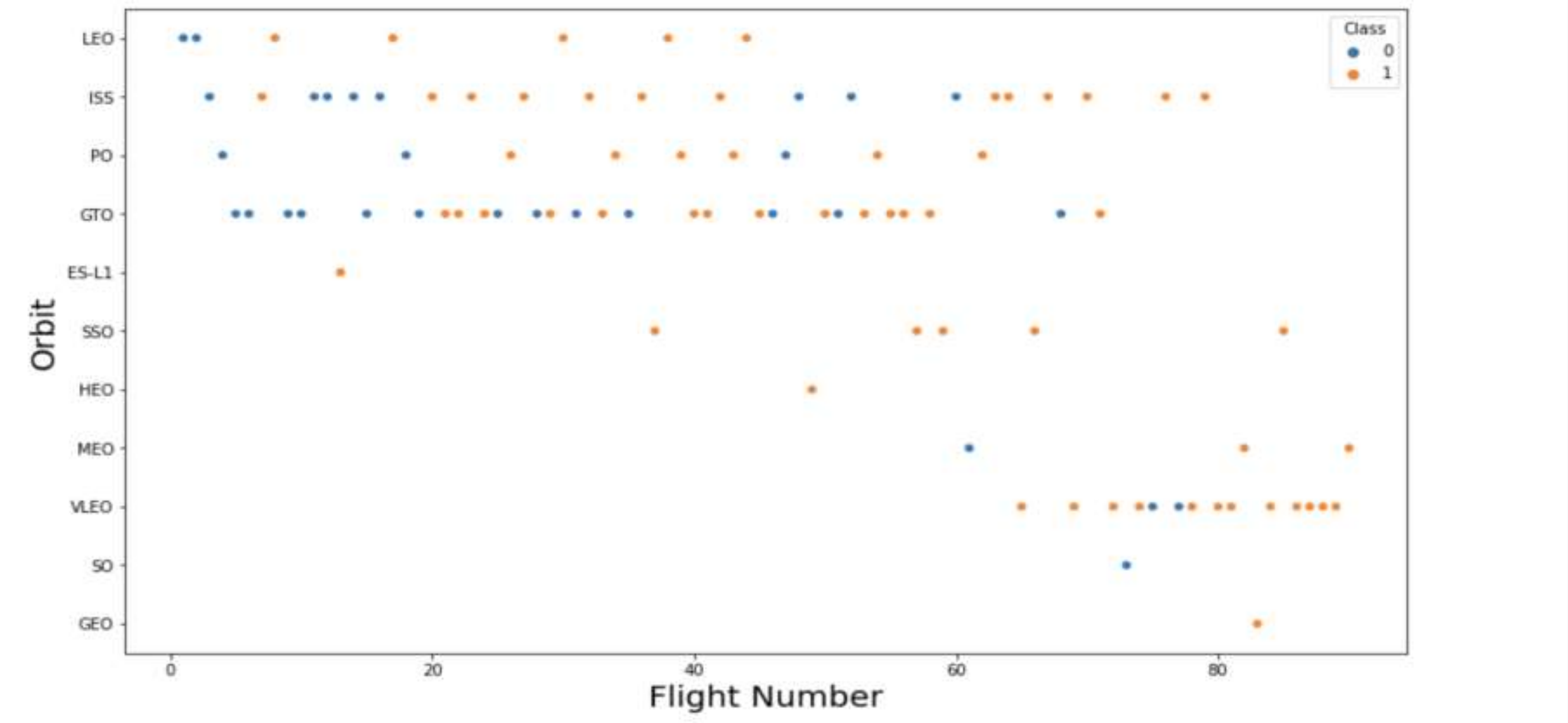
Success Rate vs. Orbit Type

- The relationship between success rate and orbit type
- VLEO had the most success rate.



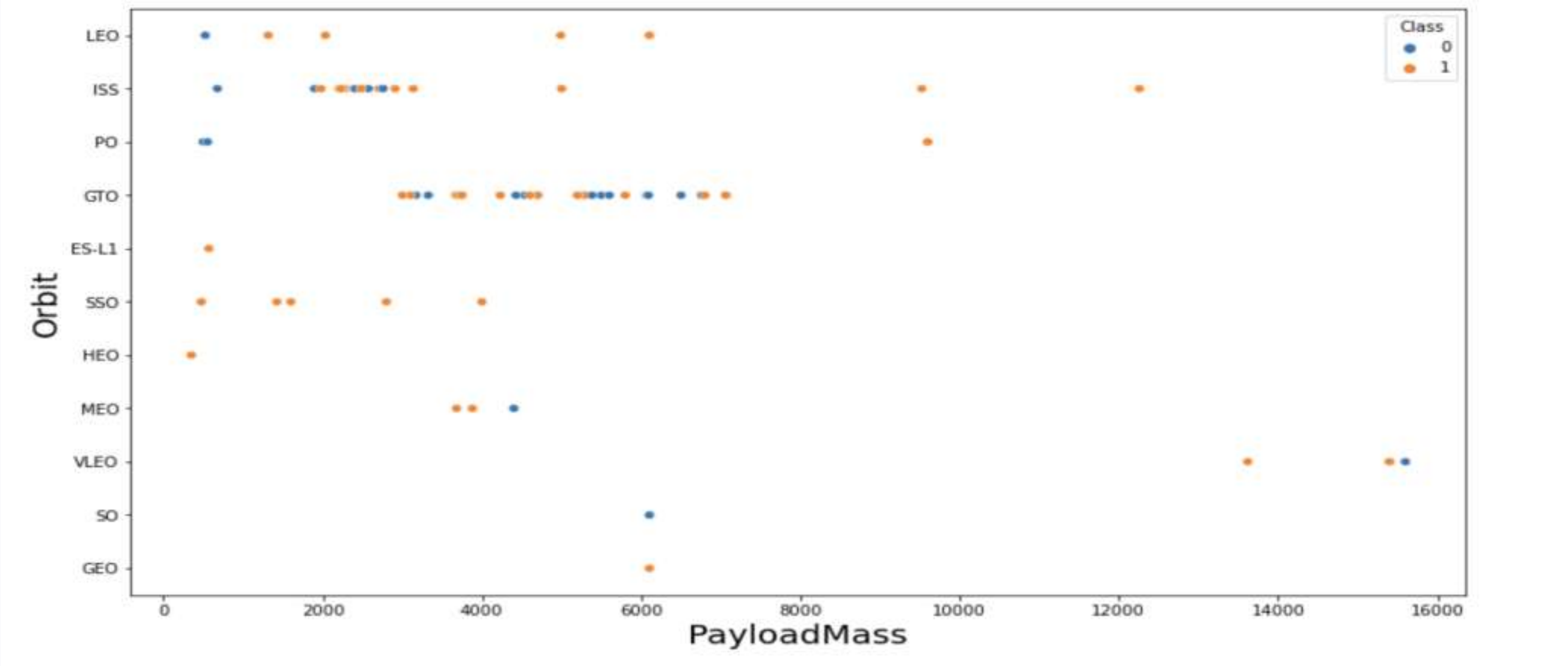
Flight Number vs. Orbit Type

- 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.



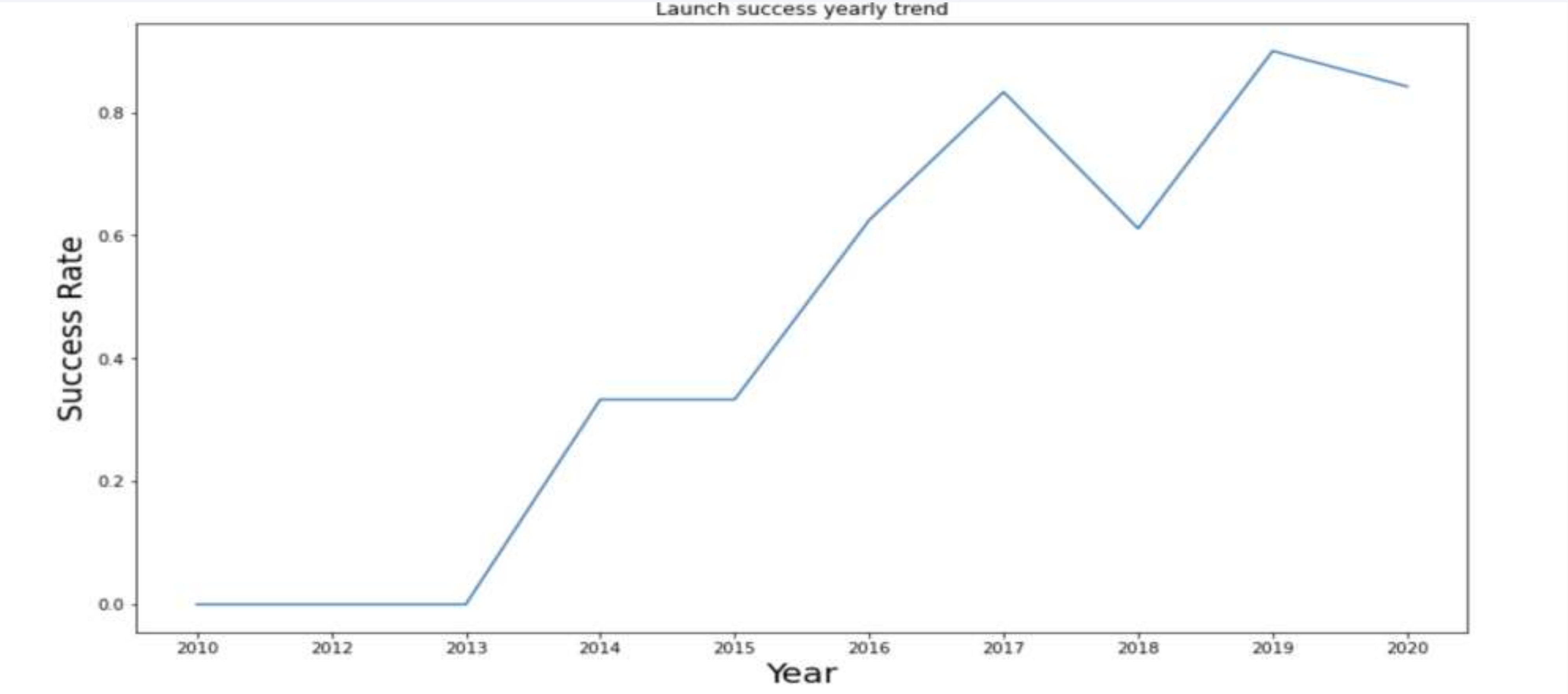
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.20



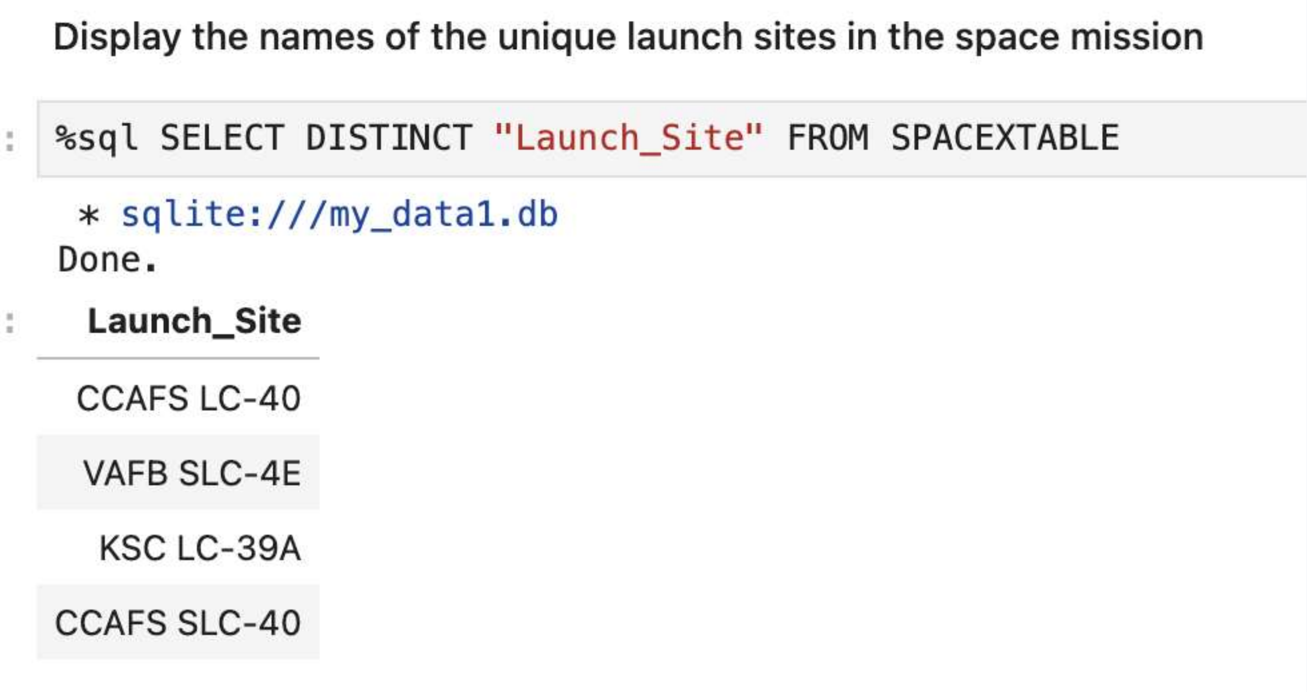
Launch Success Yearly Trend

- 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

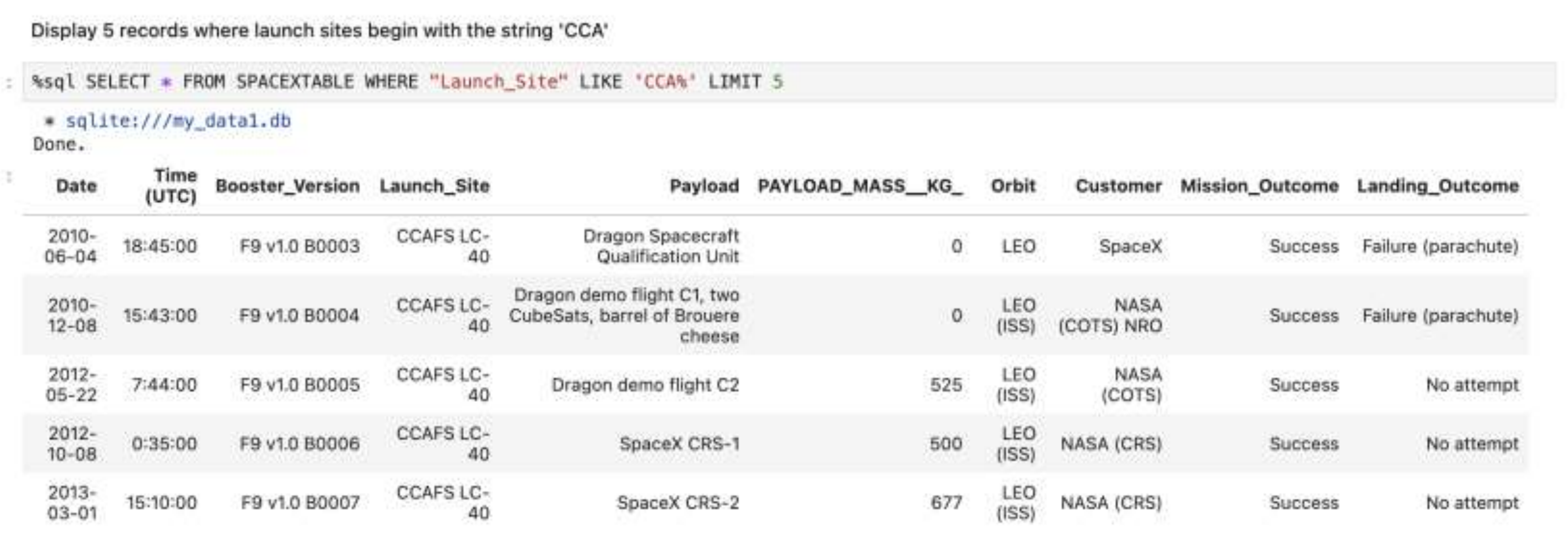
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

- This SQL query retrieves the first 5 records from the **SPACETABLE** where the **Launch_Site** starts with "CCA". It displays details like the launch date, payload mass, orbit type, and mission outcome.



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

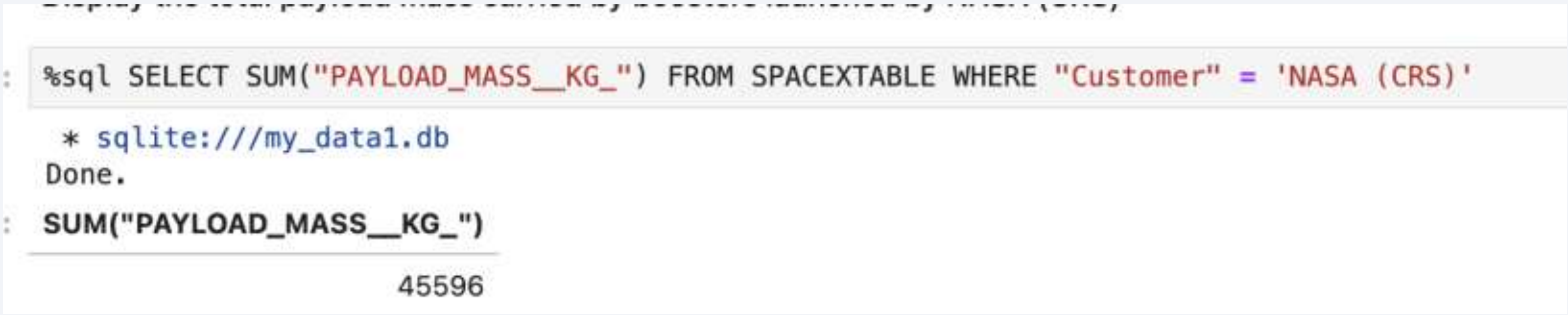
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

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

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

- This SQL query calculates the total payload mass carried by boosters launched for **NASA (CRS)**. The result of the query is **45,596 KG**, representing the sum of the payload mass for all such launches.



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

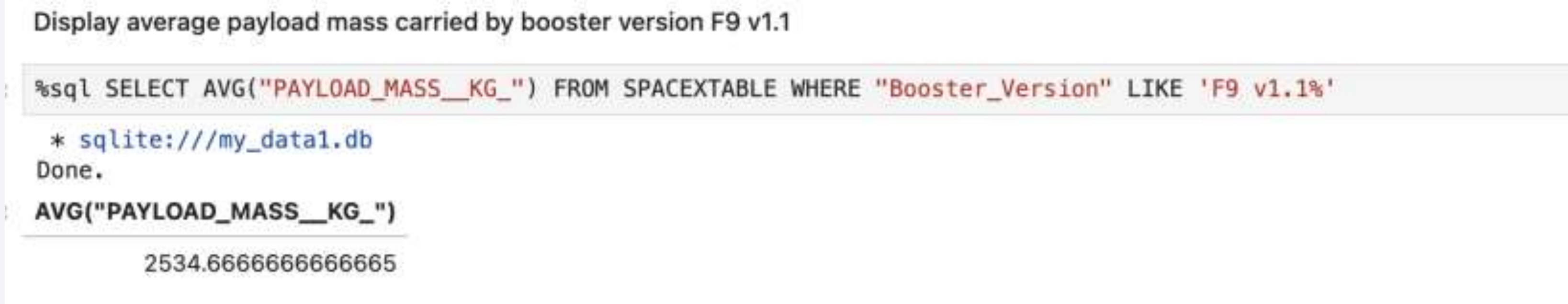
```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

SUM("PAYLOAD_MASS__KG_")
                   45596
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1



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

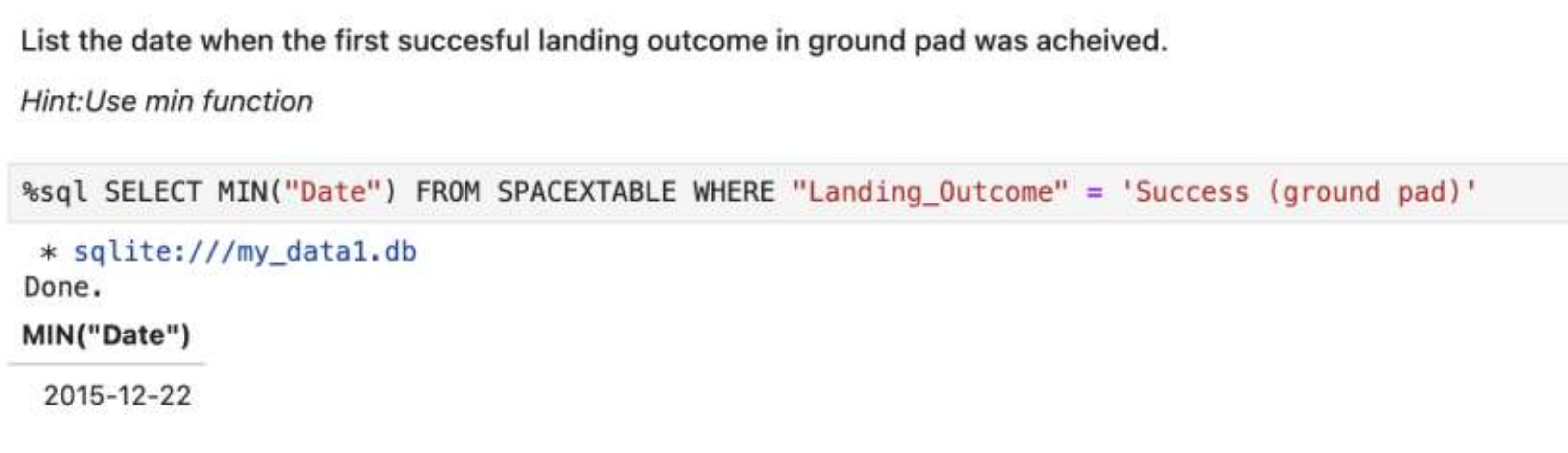
```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" LIKE 'F9 v1.1%'
```

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

AVG("PAYLOAD_MASS__KG_")
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad



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

Hint:Use min function

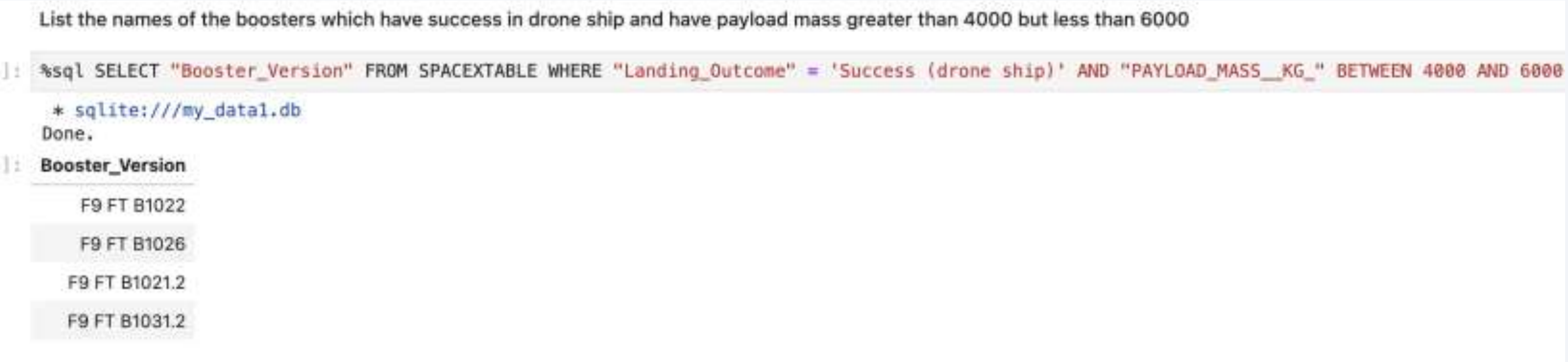
```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

MIN("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had 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

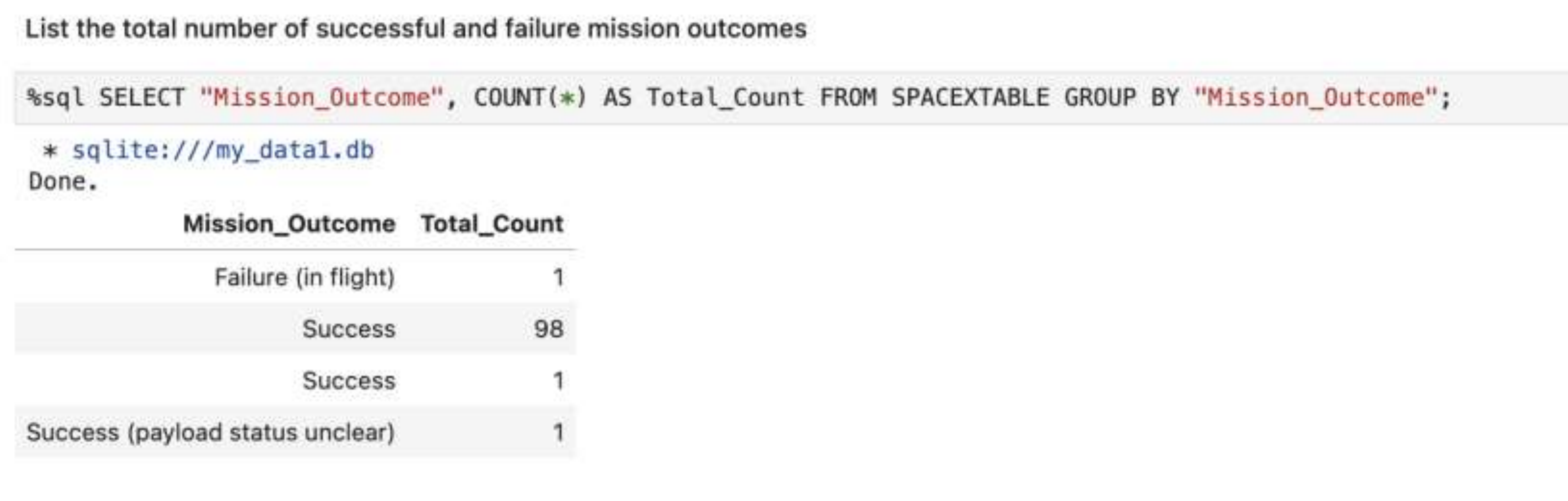
```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes



List the total number of successful and failure mission outcomes

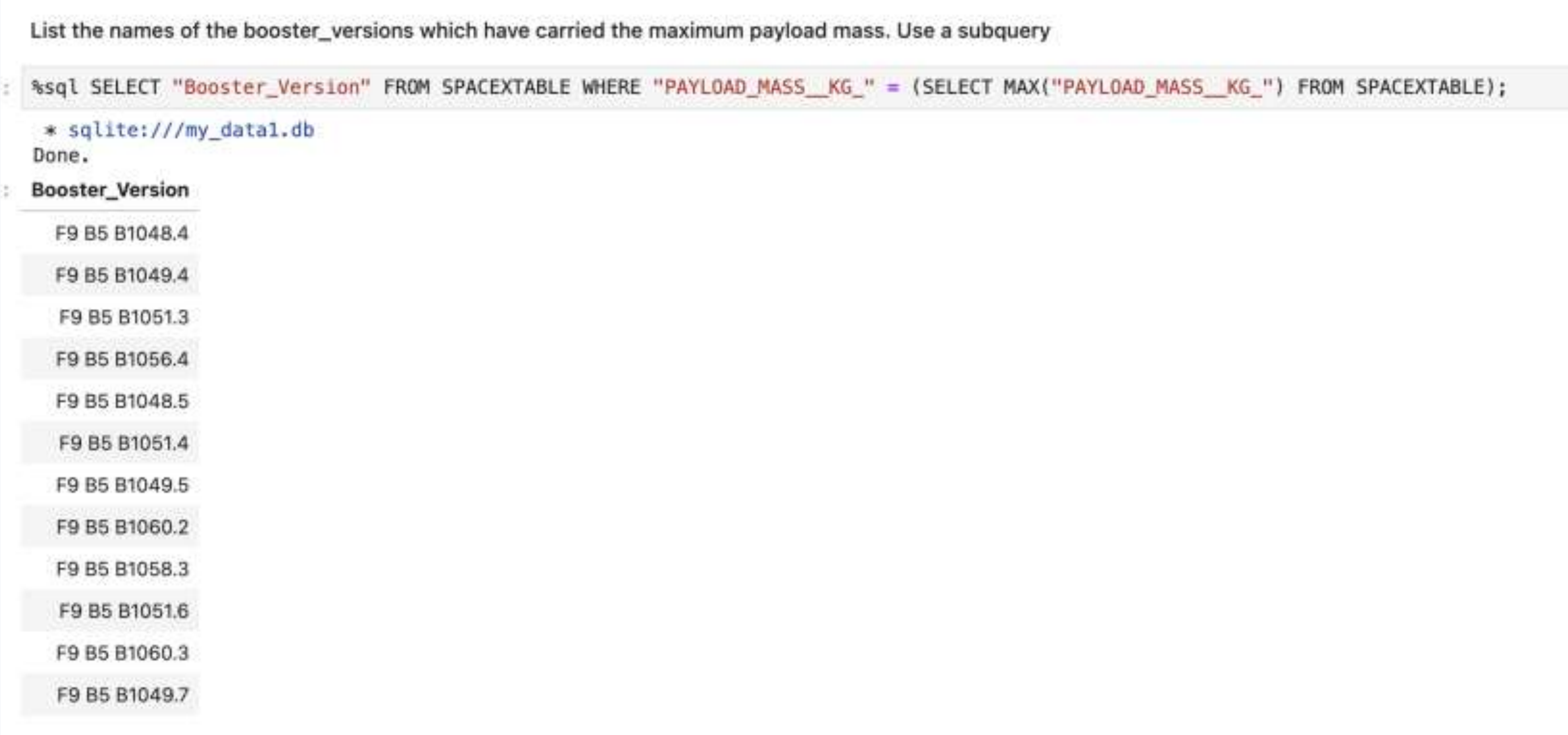
```
%sql SELECT "Mission_Outcome", COUNT(*) AS Total_Count FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass



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

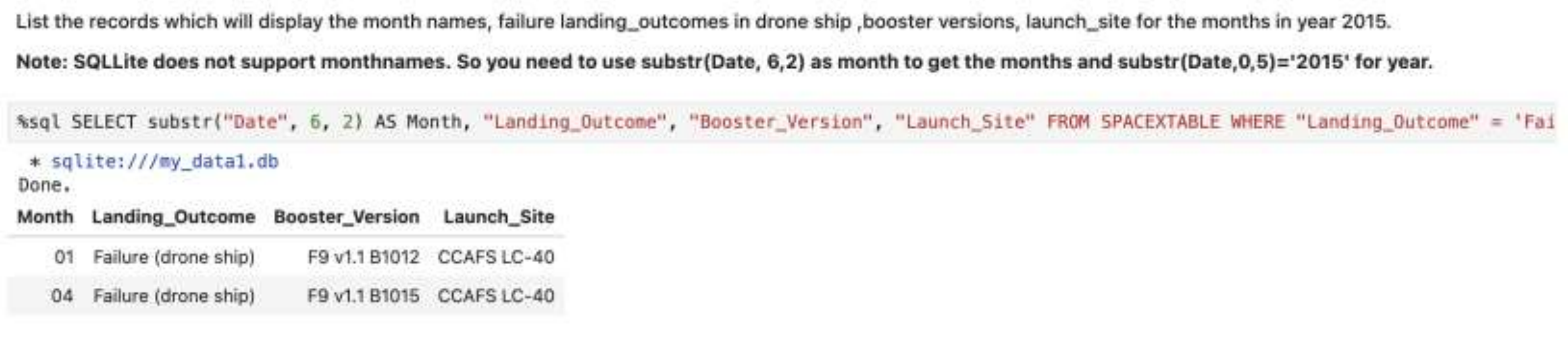
```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
```

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

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

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



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.

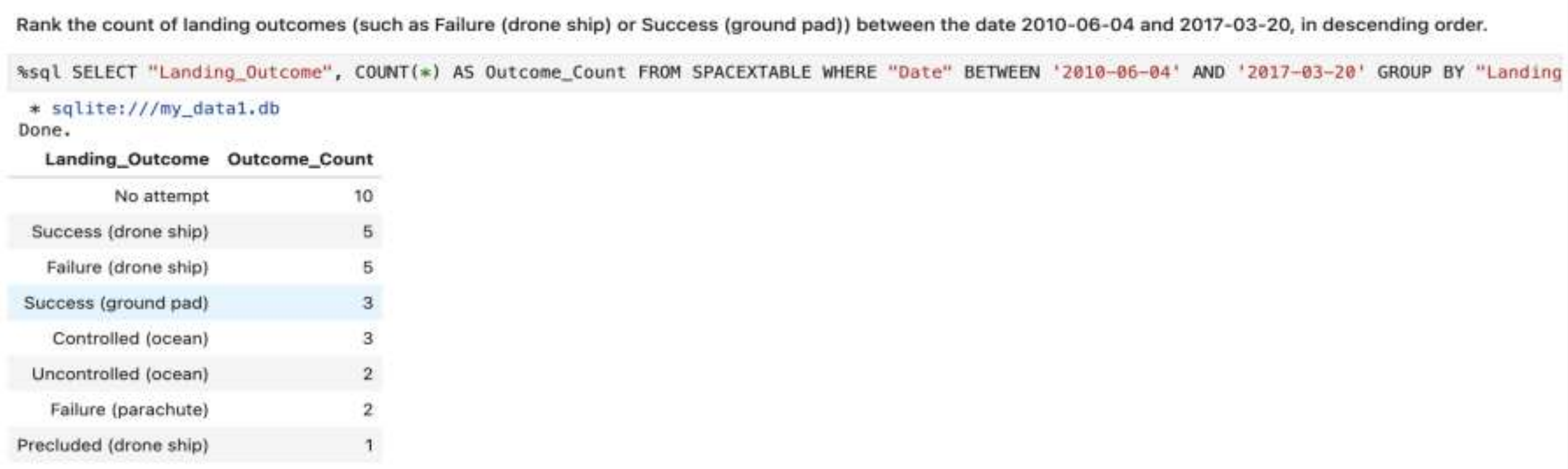
```
%sql SELECT substr("Date", 6, 2) AS Month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Fai
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
01	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

- 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



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 Outcome_Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing
```

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

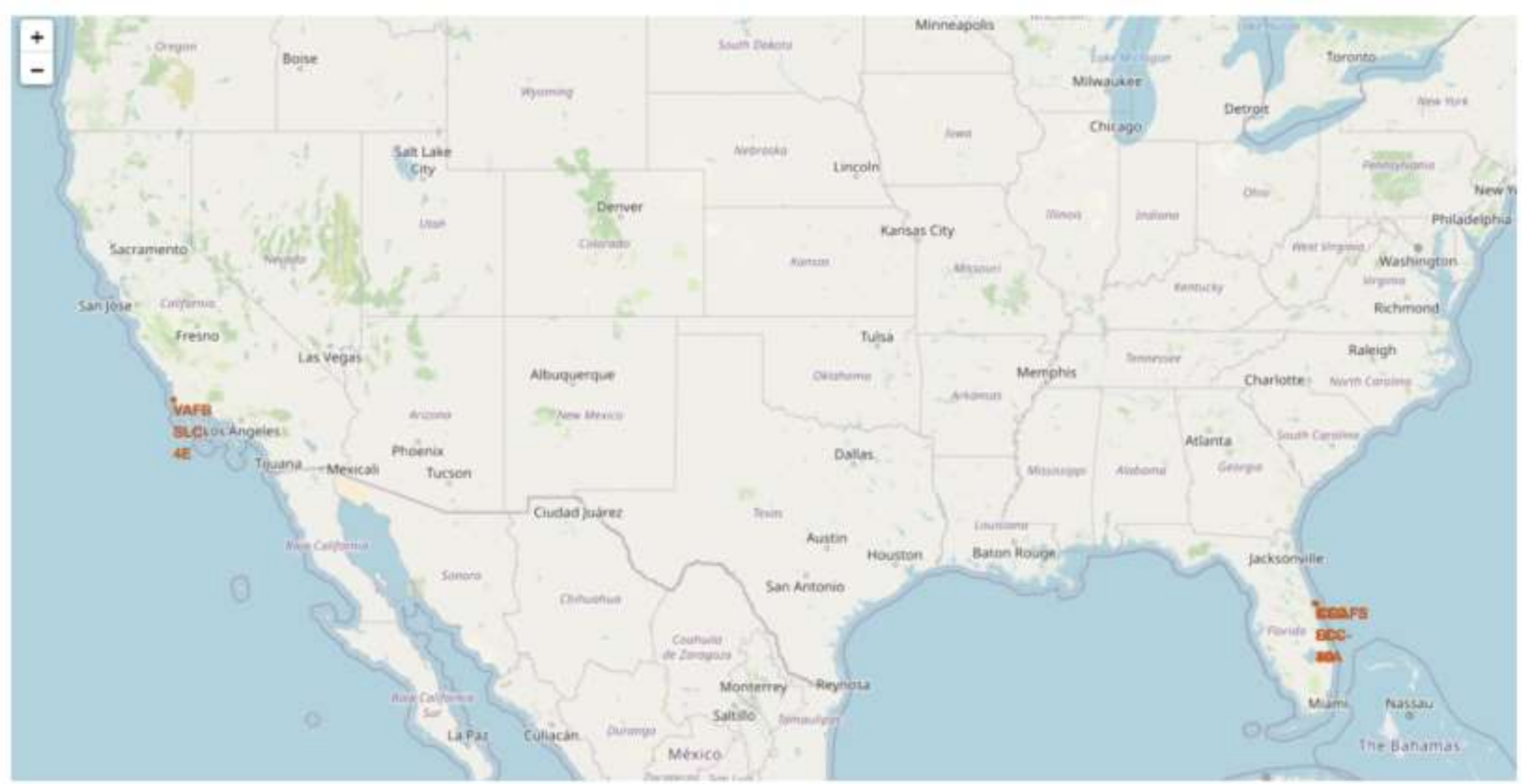
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

All launch sites

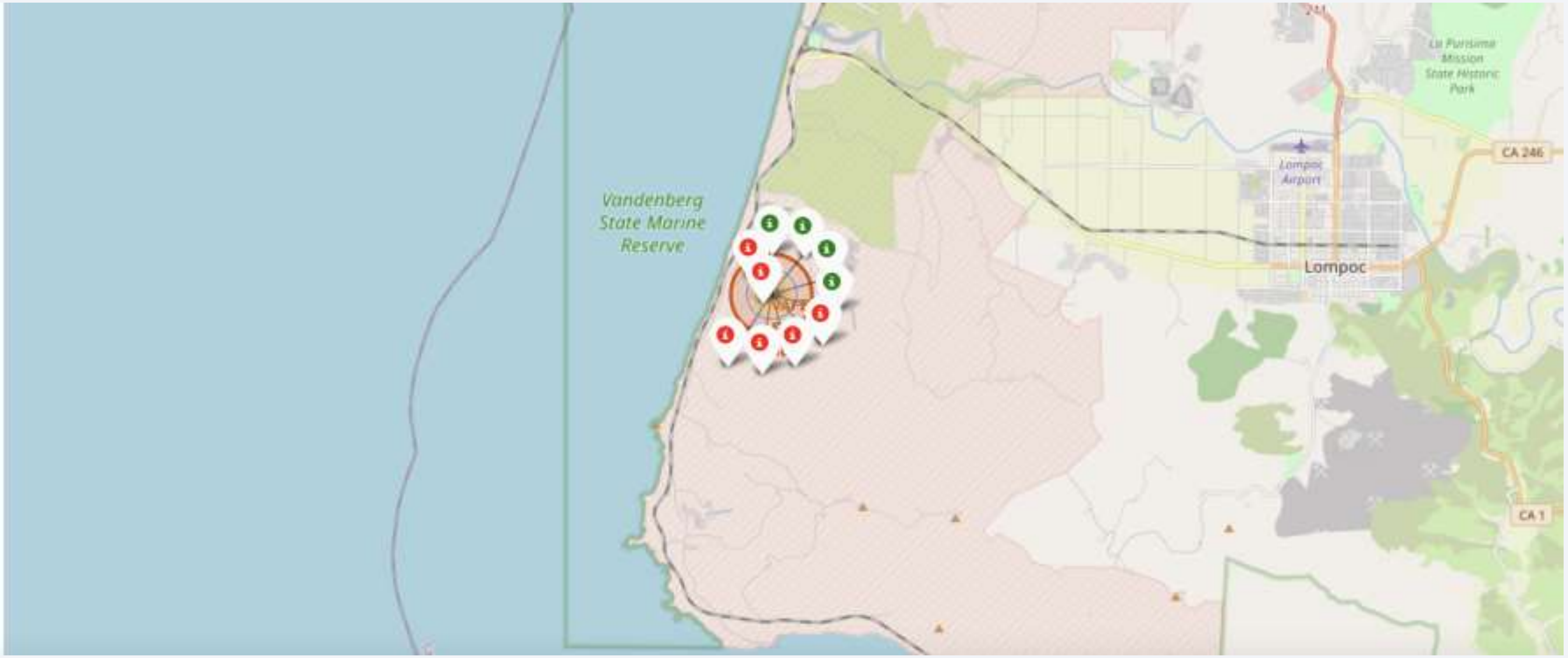
- All launch sites on map



The succeeded launches and failed launches

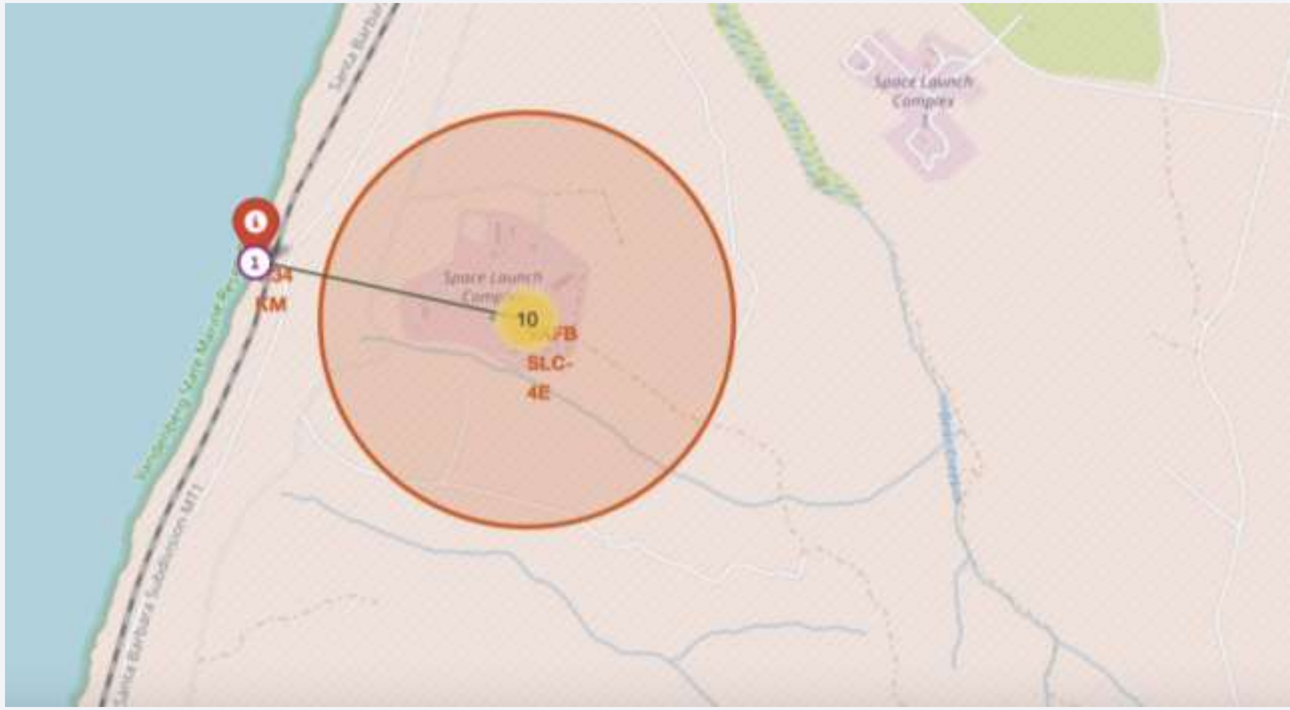
The succeeded launches and failed launches for each site on map

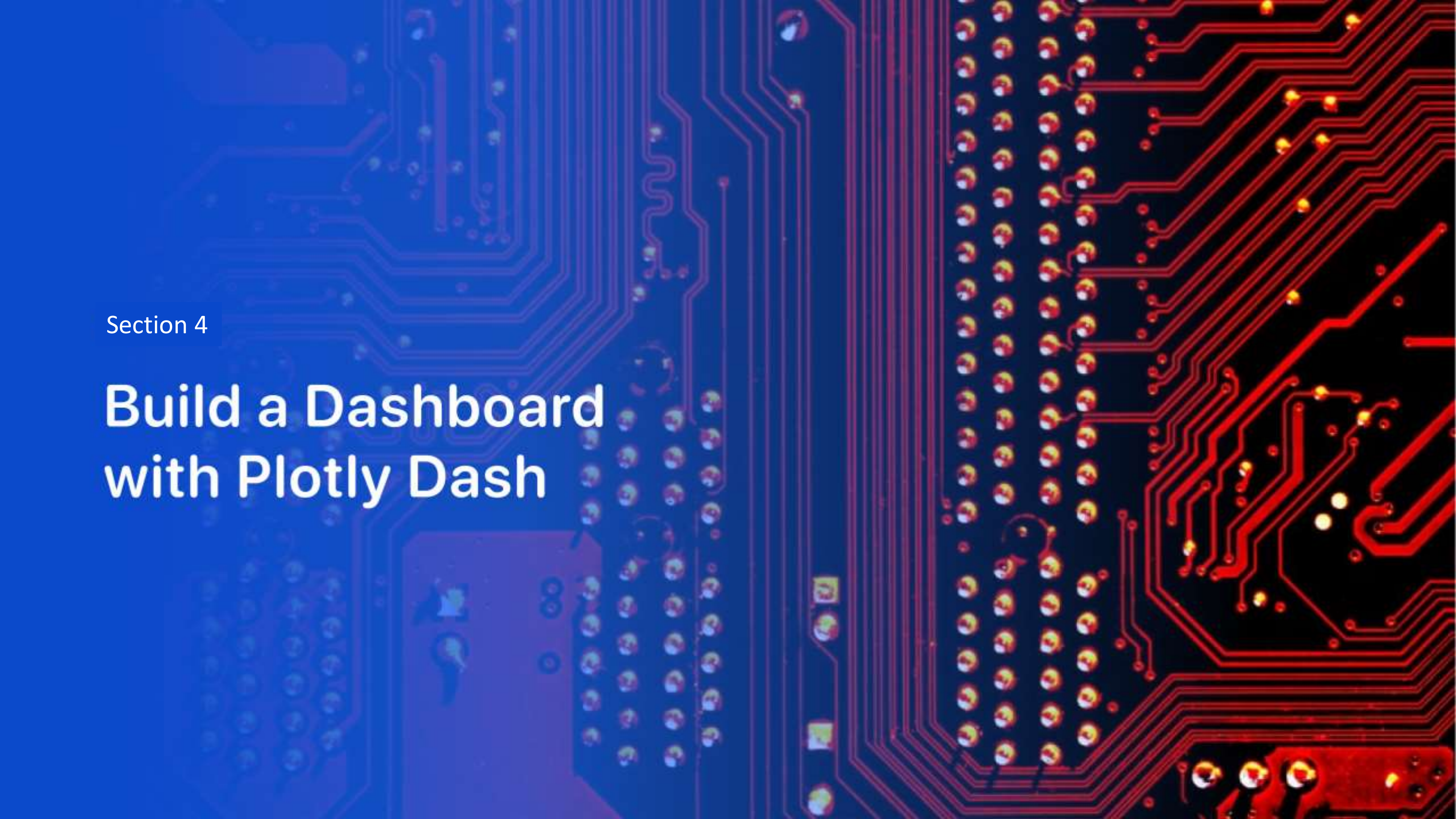
If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch


Vandenberg State Marine Reserve
La Purisima Mission State Historic Park
Lompoc Airport
Lompoc
CA 246
CA 1

The distance

- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



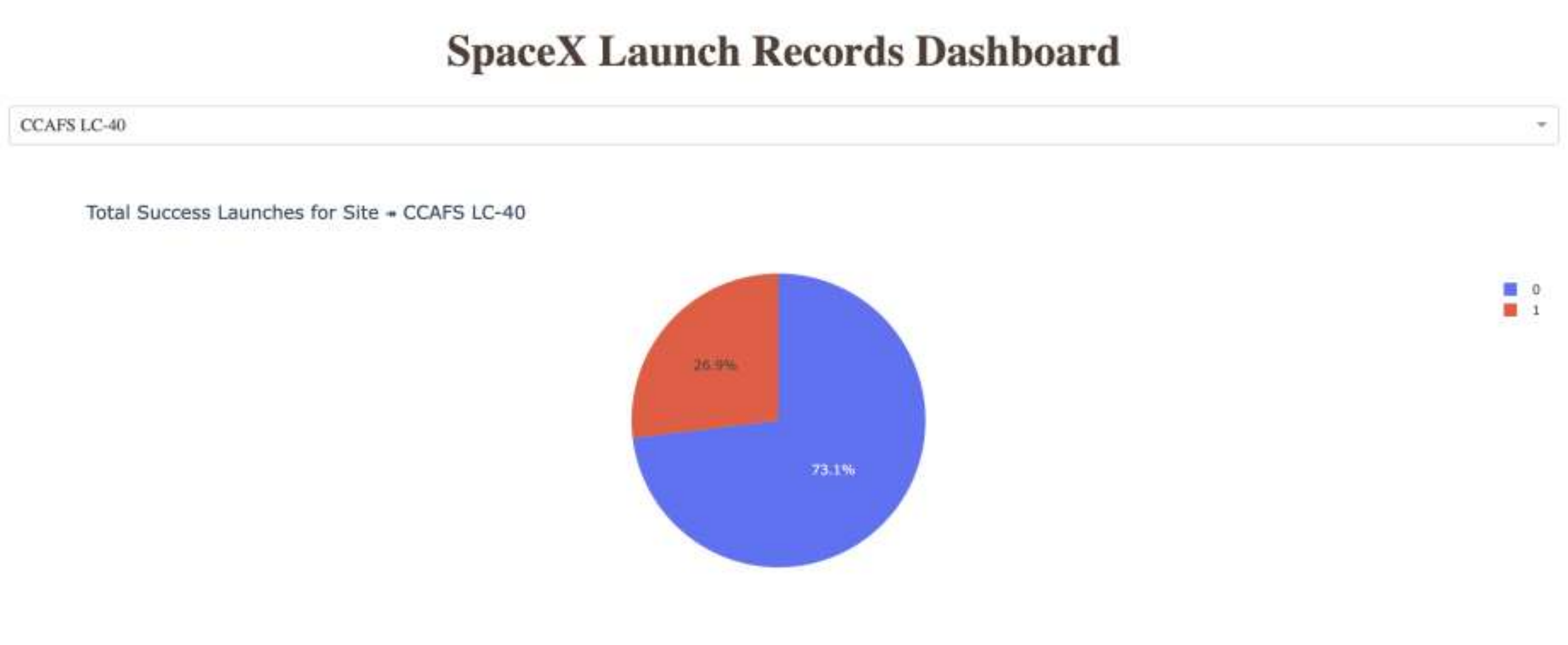


Section 4

Build a Dashboard with Plotly Dash

Dashboard

- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



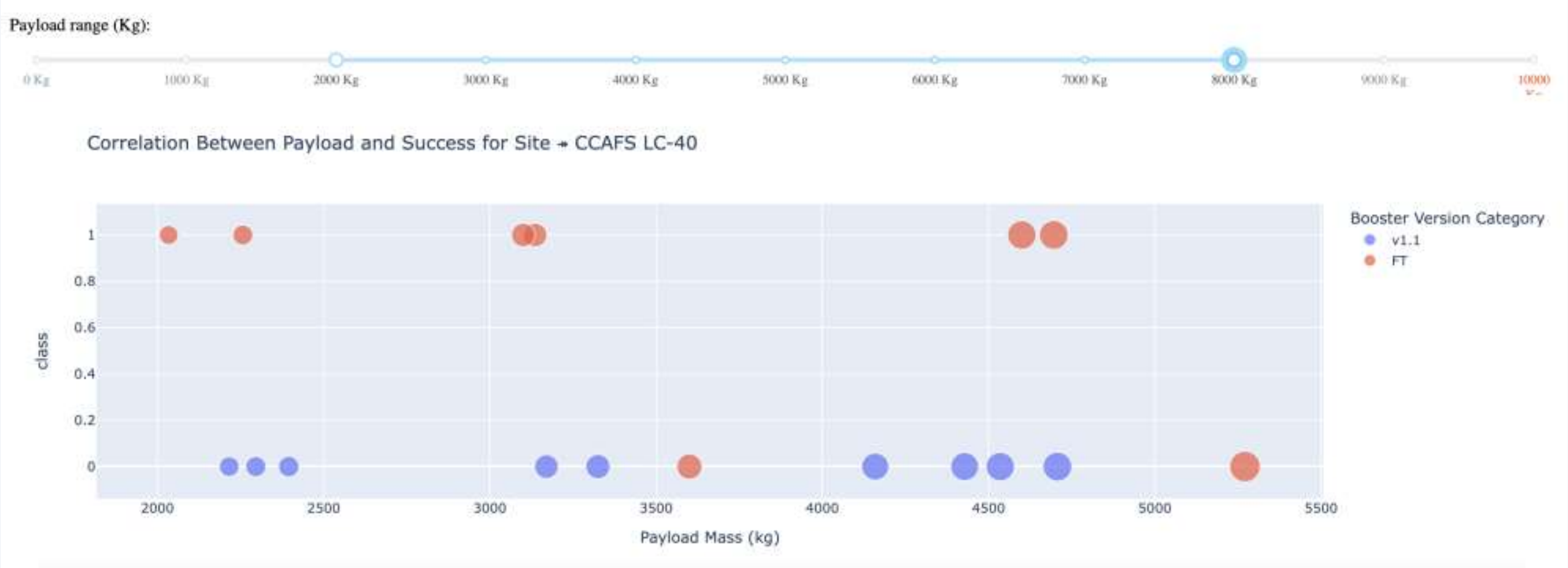
SpaceX Launch Records Dashboard

CCAFS LC-40

Total Success Launches for Site → CCAFS LC-40

Dashboard

- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

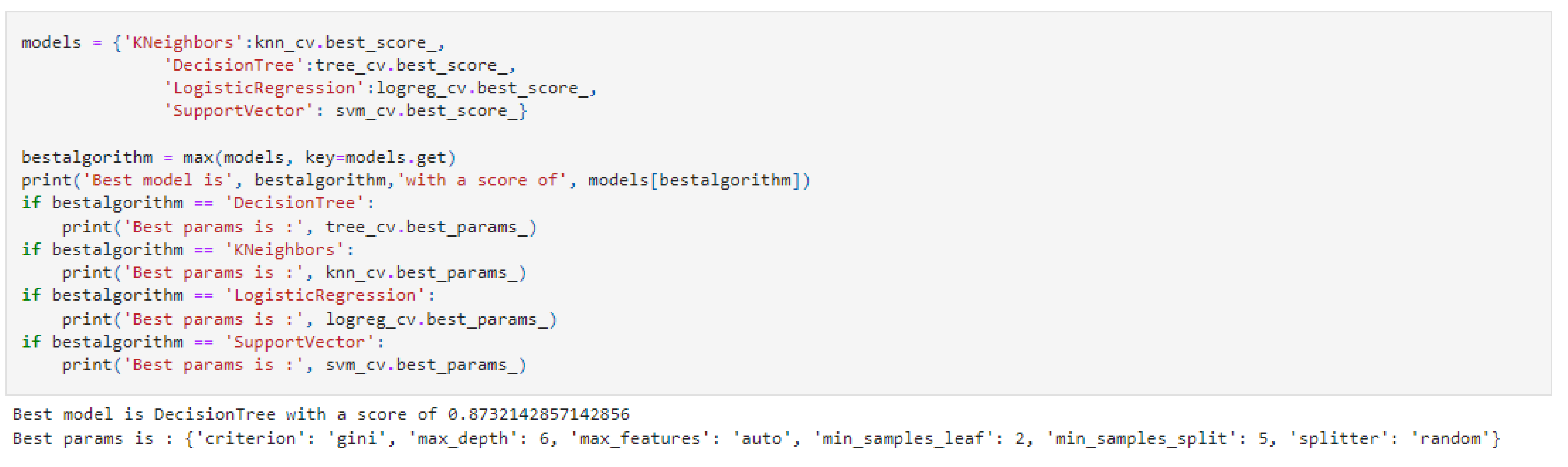

Payload range (Kg):

Section 5

Predictive Analysis (Classification)

Classification Accuracy

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



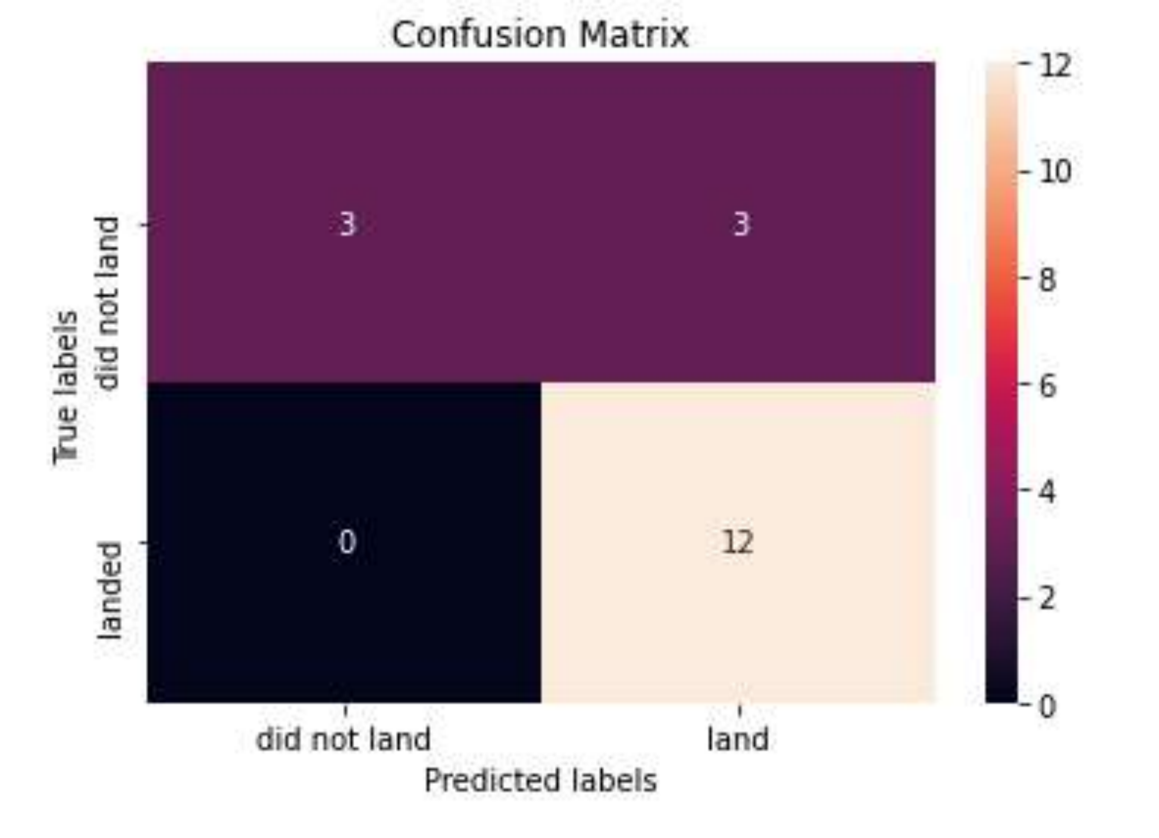
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

- **Project Goal:**
Predict if the first stage of a Falcon 9 launch will land to determine the cost of a launch.
- **Features:**
Key features like payload mass and orbit type influence the mission outcome.
- **Machine Learning Models:**
Several machine learning algorithms are used to analyze past launch data and build predictive models.
- **Best Performing Model:**
The **Decision Tree** algorithm outperformed the other 3 models in predicting launch outcomes.

Thank you!