IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ivan Pronin

16/06/2023

Outline

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

Executive Summary

- The project aims to calculate the probability of landing the first stage of the Falcon 9 rocket by analyzing data on Python and using machine learning and classification methods to forecast the likelihood of success. This report outlines the methodology, data sources, and models used in the analysis and provides insights into the accuracy and reliability of the predictions. The results of this project will inform decision-making processes for future rocket launches and contribute to the ongoing development of space exploration technology.
- The result of this work shows that by choosing right Launch Site and Payload mass you can reduce variety of failure for landing of Falcon 9 booster and as a result of that bit-by-bit company can reduce launch costs and what is more important is that could be provided with new development of technologies.

Introduction

- In this project, will be predicted if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore, if it could be determined if the first stage will land successfully, we can determine the cost of a launch. So, this is the main problem of whole project, what provides best success rate of each launch, what does it relate on? Well, let’s find out

Section 1

Methodology

Methodology

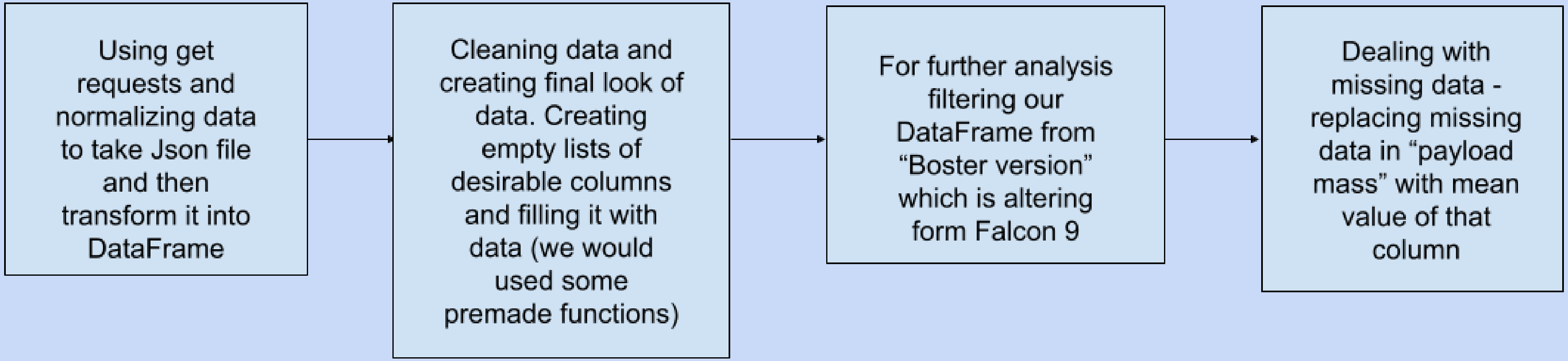
Executive Summary

- Data collection methodology:
  - Using get request to the SpaceX API for CSV file and web scraping from Wikipedia for some sensitive data about each launch
- Perform data wrangling
  - Cleaning the data from missing values and prepare data by adding class values for further machine learning
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Finding best Hyperparameter for SVM, Classification Trees and Logistic Regression and the method performs best using test data

Data Collection

- Sometimes the data is ready to go (e.g., company has been collecting data). In our case SpaceX has a huge amount of date about each launch, collecting is based on get requests to the SpaceX API and making data frame with sensitive data i.e., Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome with details, Flights, Grid Fins, Reused Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude. During this process, the data was cleaned with Nan values, null values, replaced missing values, some data was transformed to appropriate formats.
- Sometimes you need to collect it by your own. Well, I used web scrapping and Beautiful Soup (i.e., Python Libraries for collecting and web scraping data) from Wikipedia like enthusiasts from whole world would record and collect it. The results was saved into data frame with additional data i.e., Payload, Customer, Outcome in general, Version Booster, Booster landing, Time. During this process, the data was cleaned with Nan values, links, some data was transformed to appropriate formats.

Data Collection – SpaceX API



Using get requests and normalizing data to take Json file and then transform it into DataFrame

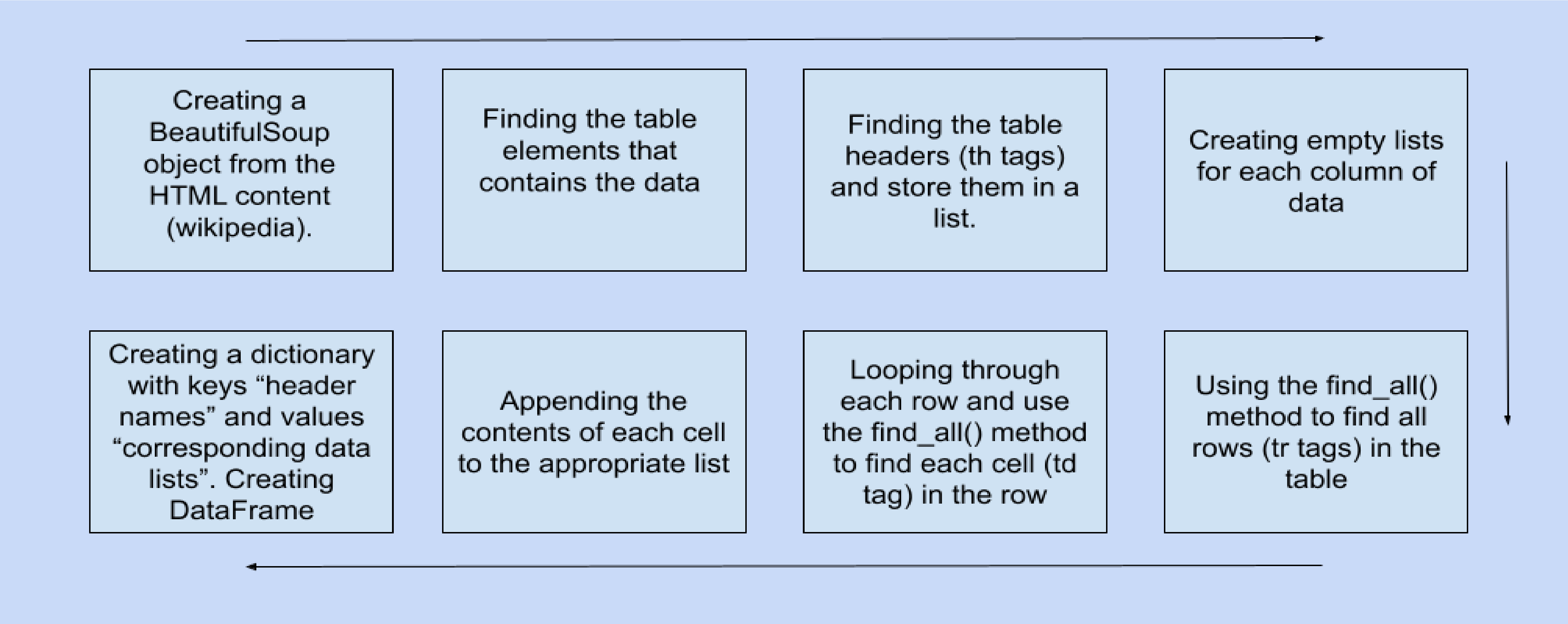
Cleaning data and creating final look of data. Creating empty lists of desirable columns and filling it with data (we would used some premade functions)

For further analysis filtering our DataFrame from “Boster version” which is altering form Falcon 9

Dealing with missing data - replacing missing data in “payload mass” with mean value of that column

GitHub - SpaceX Data Collection API

Data Collection - Scraping



Creating a BeautifulSoup object from the HTML content (wikipedia).

Finding the table elements that contains the data

Finding the table headers (th tags) and store them in a list.

Creating empty lists for each column of data

Using the find_all() method to find all rows (tr tags) in the table

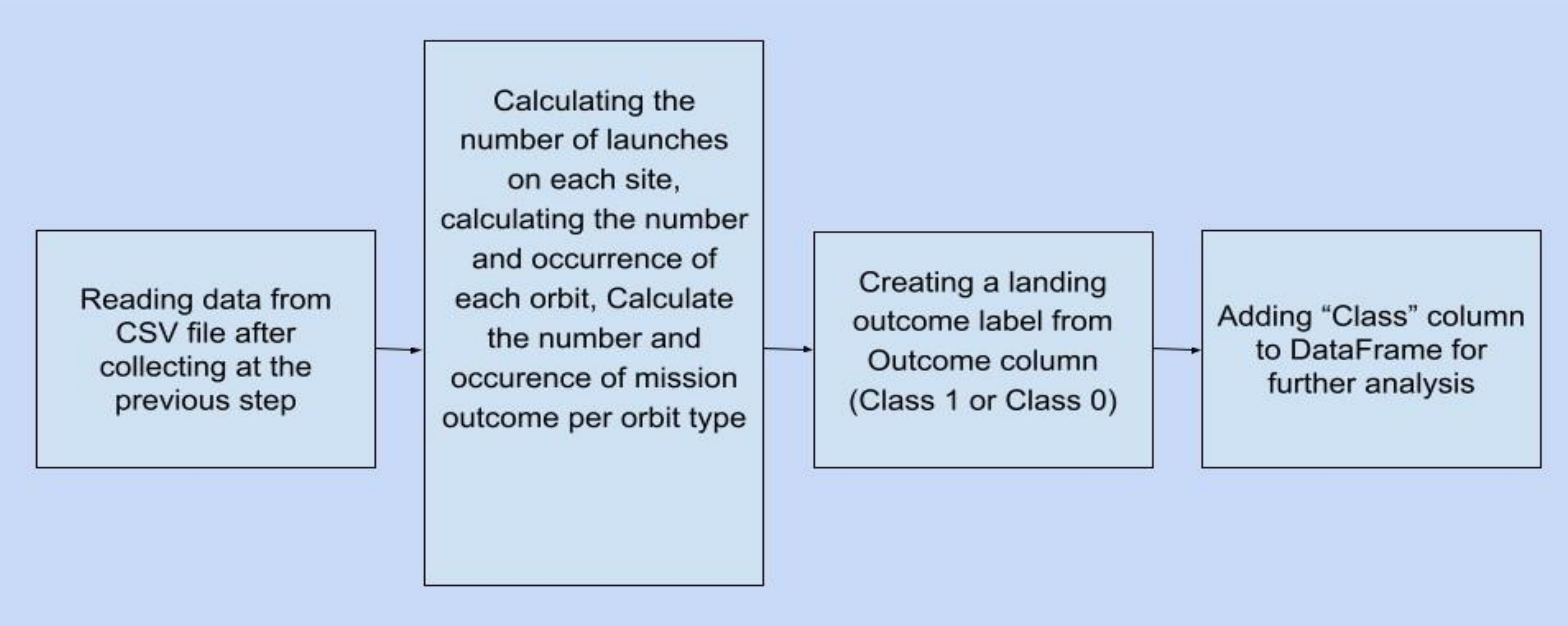
Looping through each row and use the find_all() method to find each cell (td tag) in the row

Appending the contents of each cell to the appropriate list

Creating a dictionary with keys “header names” and values “corresponding data lists”. Creating DataFrame

GitHub – Web scraping

Data Wrangling



Reading data from CSV file after collecting at the previous step

Calculating the number of launches on each site, calculating the number and occurrence of each orbit, Calculate the number and occurence of mission outcome per orbit type

Creating a landing outcome label from Outcome column (Class 1 or Class 0)

Adding “Class” column to DataFrame for further analysis

GitHub - Data Wrangling

EDA with Data Visualization

- Scatter plots and cat plots were used to analyze the relationships between Flight Number and Launch Site, Payload and Launch Site, Flight Number and Orbit type, and Payload and Orbit type because these types of plots are effective in visually representing the distribution and correlation of two variables.
- A Bar chart was used to analyze the relationship between success rate of each orbit type because it is an effective way to compare the values of different categories.
- A line chart was used to analyze the trend of successful launches because it is a useful tool for showing changes over time and identifying patterns or trends.

GitHub - Data Visualization

EDA with SQL

- - Selecting distinct launch sites from SPACEXTBL
- - Selecting launch site from SPACEXTBL where ‘Launch_Site’ starts with 'CCA' and limiting the result to 5
- - Selecting the sum of ‘PAYLOAD_MASS__KG_’ where Customer is 'NASA (CRS)'
- - Selecting the average of ‘PAYLOAD_MASS__KG_’ where ‘Booster_Version’ is 'F9 v1.1'
- - Selecting the minimum ‘Date’ where ‘Landing_Outcome’ is 'Success (ground pad)'
- - Selecting ‘Booster_Version’ from SPACEXTBL where ‘Landing_Outcome’ is 'Success (drone ship)' and ‘PAYLOAD_MASS__KG_’ is between 4000 and 6000
- - Grouping ‘Mission_Outcome’ and counting the frequency of each outcome in SPACEXTBL
- - Selecting ‘Booster_Version’ and ‘PAYLOAD_MASS__KG_’ from SPACEXTBL, ordering by ‘PAYLOAD_MASS__KG_’ in descending order, and limiting to the top 20 results
- - Selecting month, year, ‘Landing_Outcome’, ‘Booster_Version’, and ‘Launch_Site’ from SPACEXTBL where ‘Landing_Outcome’ is 'Failure (drone ship)' and year is '2015', and limiting to the top 20 results
- - Grouping ‘Mission_Outcome’ and counting the frequency of each outcome in SPACEXTBL where the Date is between '04/06/2010' and '20/03/2017' and ‘Mission_Outcome’ is 'Success', ordering by frequency in descending order.

GitHub - SQL

Build an Interactive Map with Folium

- The folium map created included markers for all launch sites, as well as markers indicating the success or failure of launches at each site. And it’s interactive. Additionally, circles were added to the map to show the proximity of each launch site to nearby coastlines, railroads, highways, and cities. Also, lines with distance values were added.
- These objects were added to provide additional context and information about each launch site, including potential logistical challenges and environmental factors that may impact launch operations.

GitHub - Interactive Map with Folium

Build a Dashboard with Plotly Dash

- The dashboard visualizes the SpaceX launch records using different plots and interactions.
- Pie charts are used to compare the launch sites and their success rates, which show how each site contributes to the total number of launches and how successful they are.
- A scatter plot is used to explore the correlation between the payload mass and the launch success for each site, which can reveal some patterns or outliers in the data.
- A slider is added to make the scatter plot more interactive, which allows the user to change the input payload mass and see how it affects the launch success. This way, the user can experiment with different scenarios and get instant feedback.
- GitHub - Dashboard with Plotly Dash

Predictive Analysis (Classification)

- In the data preparation step, the data was transformed, standardized using preprocessing 'StandardScaler' due to reduce influences of big numbers in general, and split into training and test sets.
- In the modeling step, four classification algorithms were applied to the training data: logistic regression, support vector machine, decision tree, and k nearest neighbors. GridSearchCV was used to perform hyperparameter tuning and cross-validation for each algorithm and find the best parameters of the accuracy score.
- In the evaluation step, the models were compared and ranked based on their accuracy scores on the cross-validation data. The best performing model was selected and tested on the unseen test data to validate its generalization ability and confirm its accuracy score.
- GitHub - Predictive Analysis (Classification)

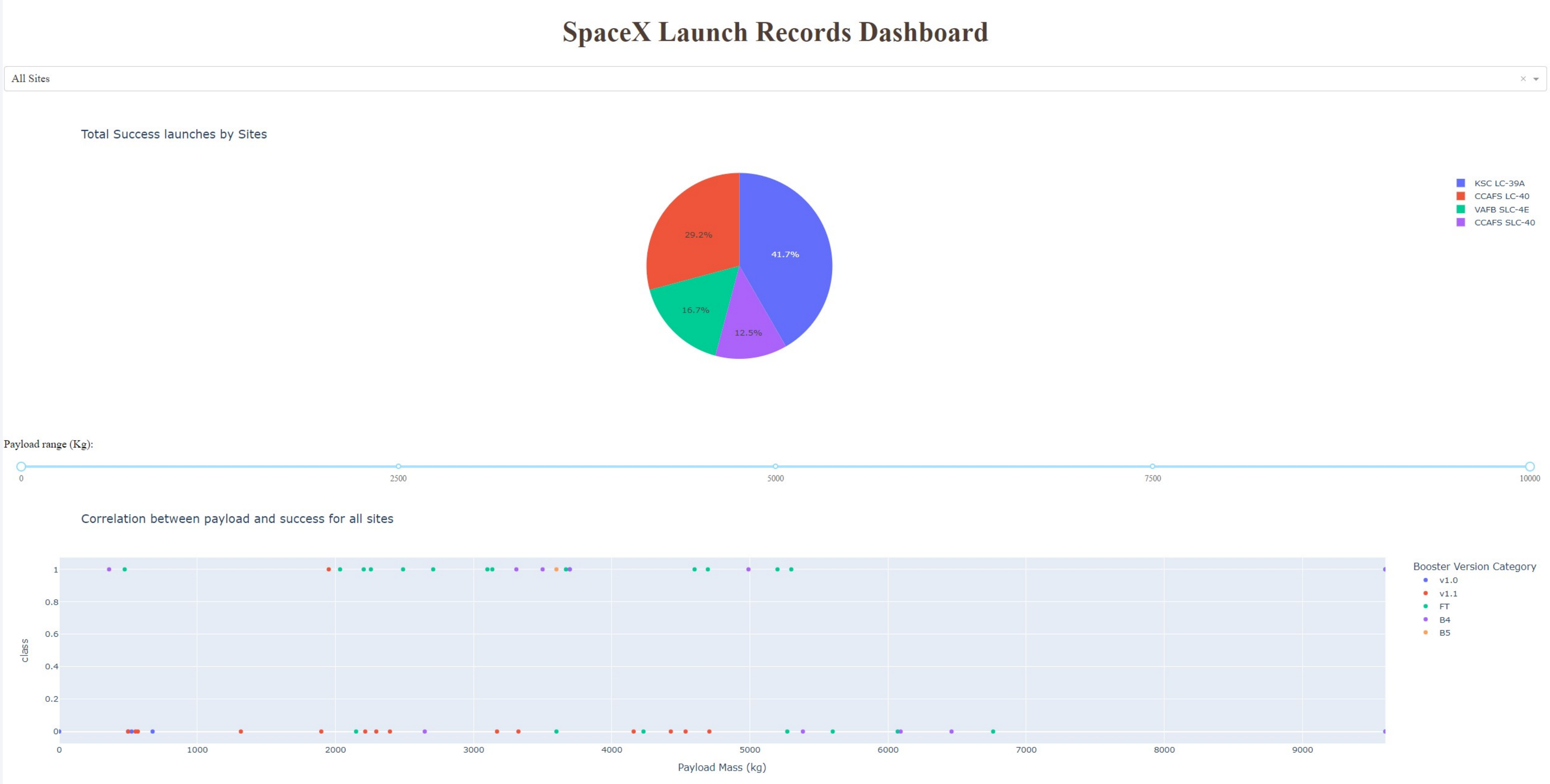
Results

Exploratory data analysis results:

This exploratory data analysis used various data sources and methods to gain insights into SpaceX launches and their success rates. Data were collected from the SpaceX API and Wikipedia, and data wrangling and data visualization were performed using scatter plots, cat plots, bar charts and line charts. SQL queries were also used to query the data and find some interesting facts. The main goal of this stage was to prepare the data for further analysis, such as creating a dashboard to display the key metrics, making a folium map to show the launch locations, and training a classification model to predict the launch outcomes. The results of this analysis revealed some patterns and trends in the data, such as the increasing success rate over time, the variation in success rate among launch sites, and the positive correlation between payload mass and success rate.

Results

Interactive analytics demo in screenshots:



SpaceX Launch Records Dashboard

All Sites

Results

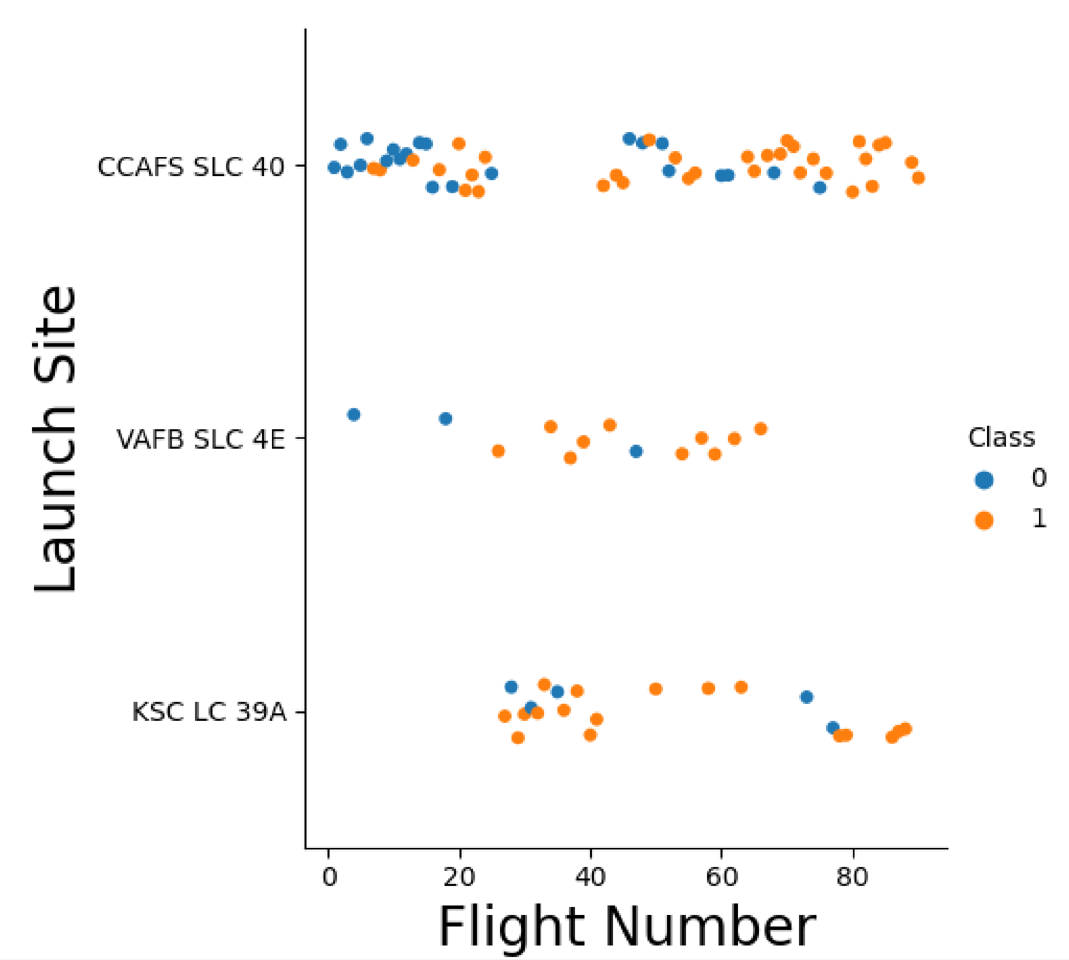
Prediction analysis results:

This stage involved prediction analysis using various supervised machine learning models to predict the launch outcomes. The data were normalized using ‘StandardScaler’ and split into train and test sets with a ratio of 80:20. The models used were support vector machine (SVM), decision tree, k-nearest neighbors (KNN) and logistic regression. Grid search was used to find the best hyperparameters for each model, such as the kernel and C for SVM, the max depth and criterion for decision tree, the number of neighbors and distance metric for KNN, and the solver and penalty for logistic regression. The best R2 score achieved was 83.3% by all models on the test set, but best accuracy of model achieved was 88.75%. The decision tree model was chosen based on its simplicity and interpretability. A confusion matrix was generated to evaluate its performance, showing the true positives, true negatives, false positives and false negatives.

Section 2

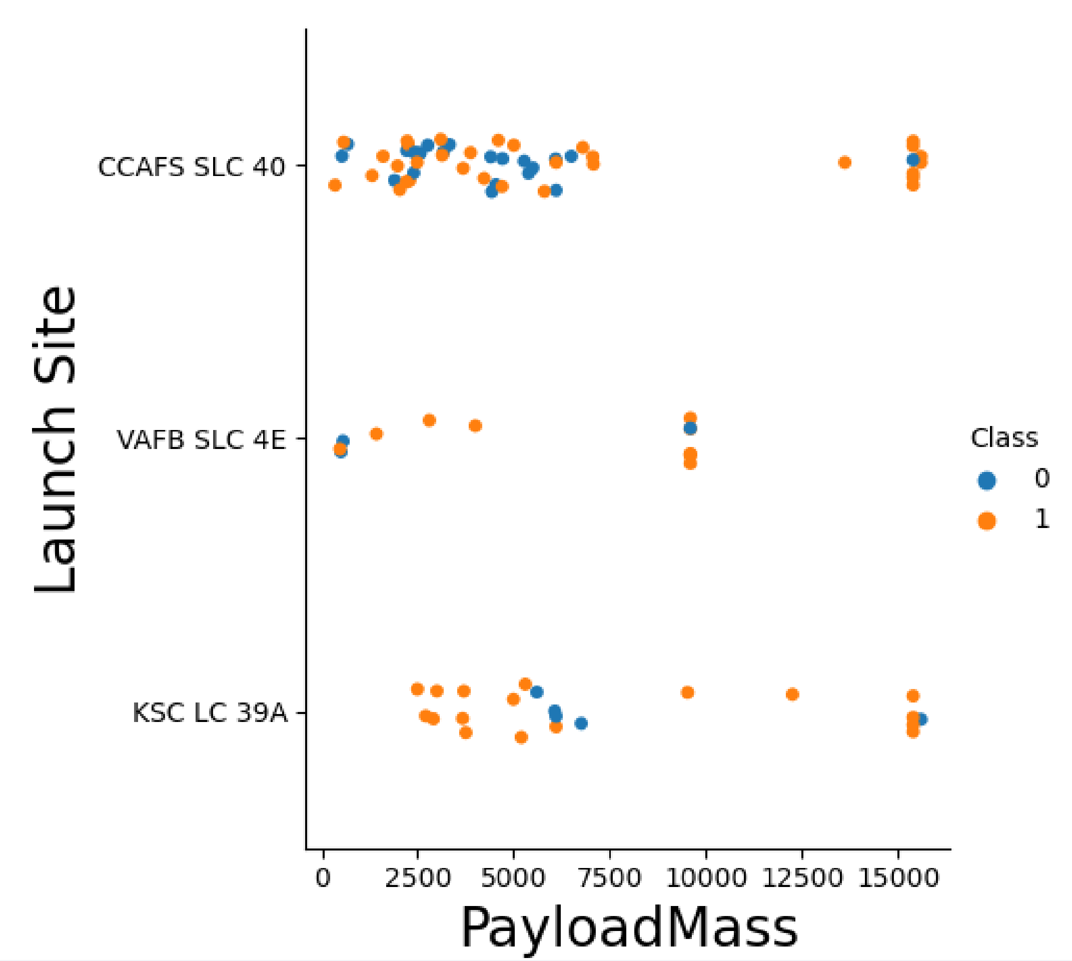
Insights drawn from EDA

Flight Number vs. Launch Site



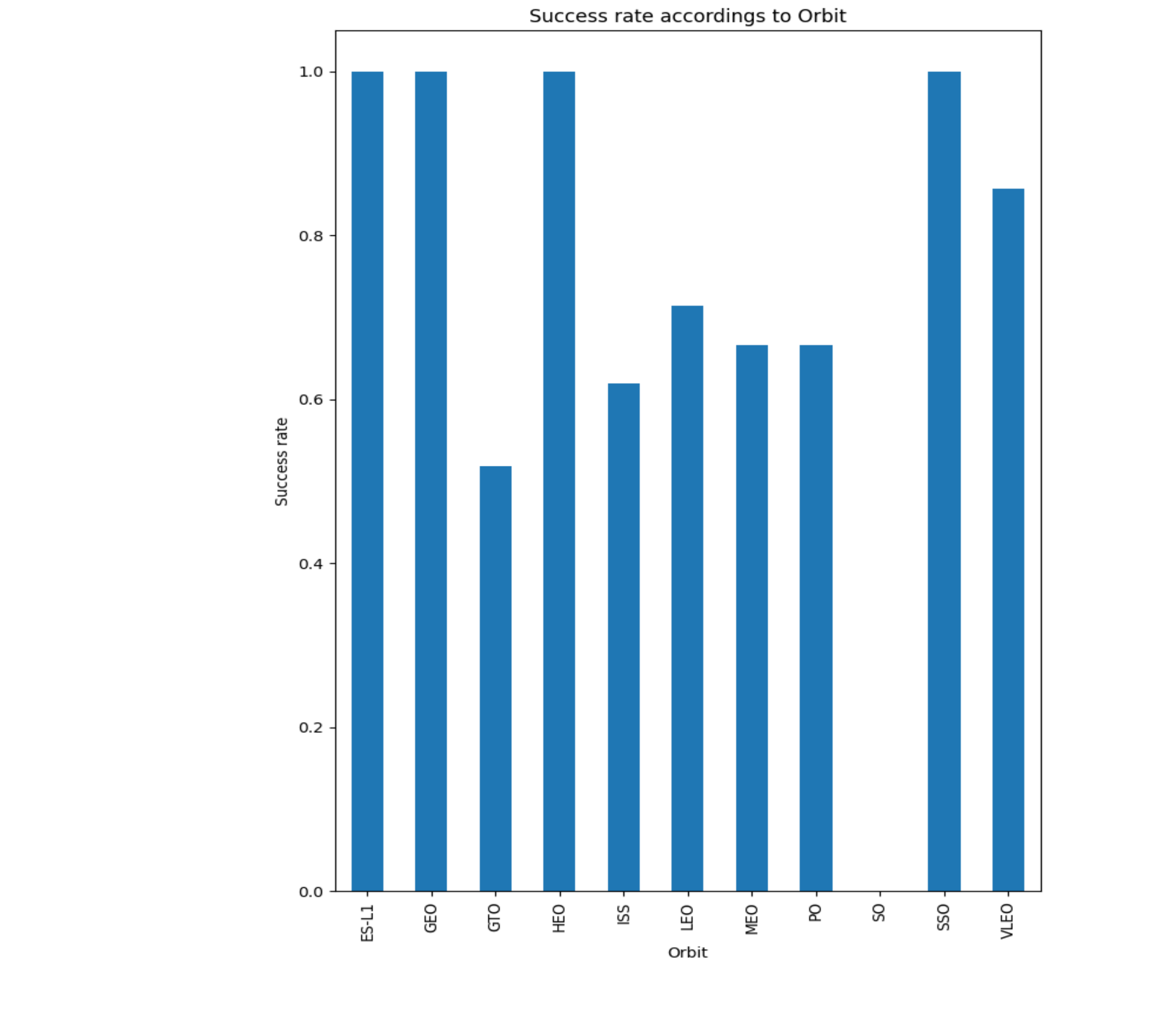
This plot shows that Launch Site calls CCAFS SLC 40 has more flights than others.

Payload vs. Launch Site



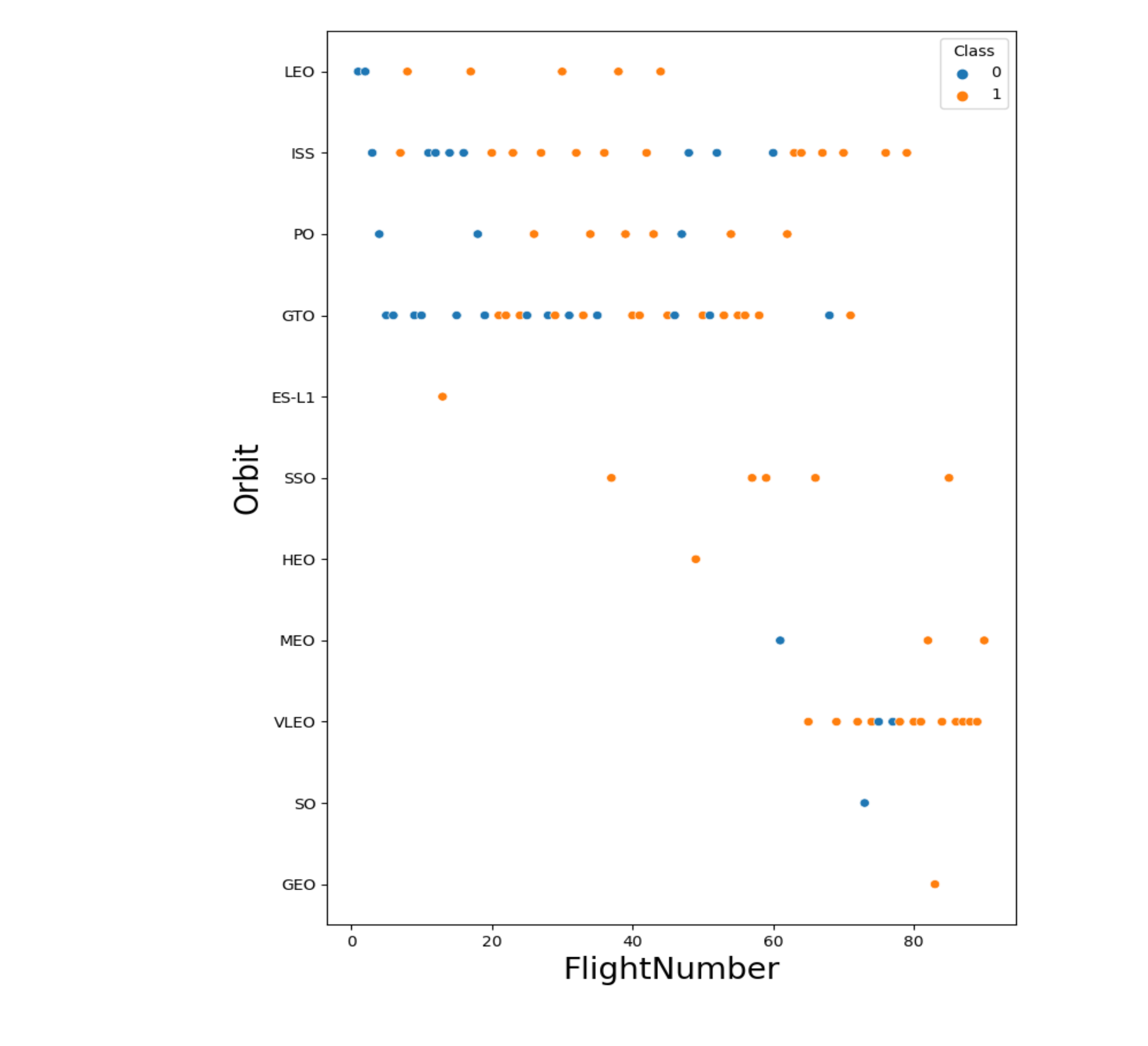
This plot shows that Launch Site calls VAFB SLC 4E doesn’t use with payload mass more than 10 000 KG.

Success Rate vs. Orbit Type



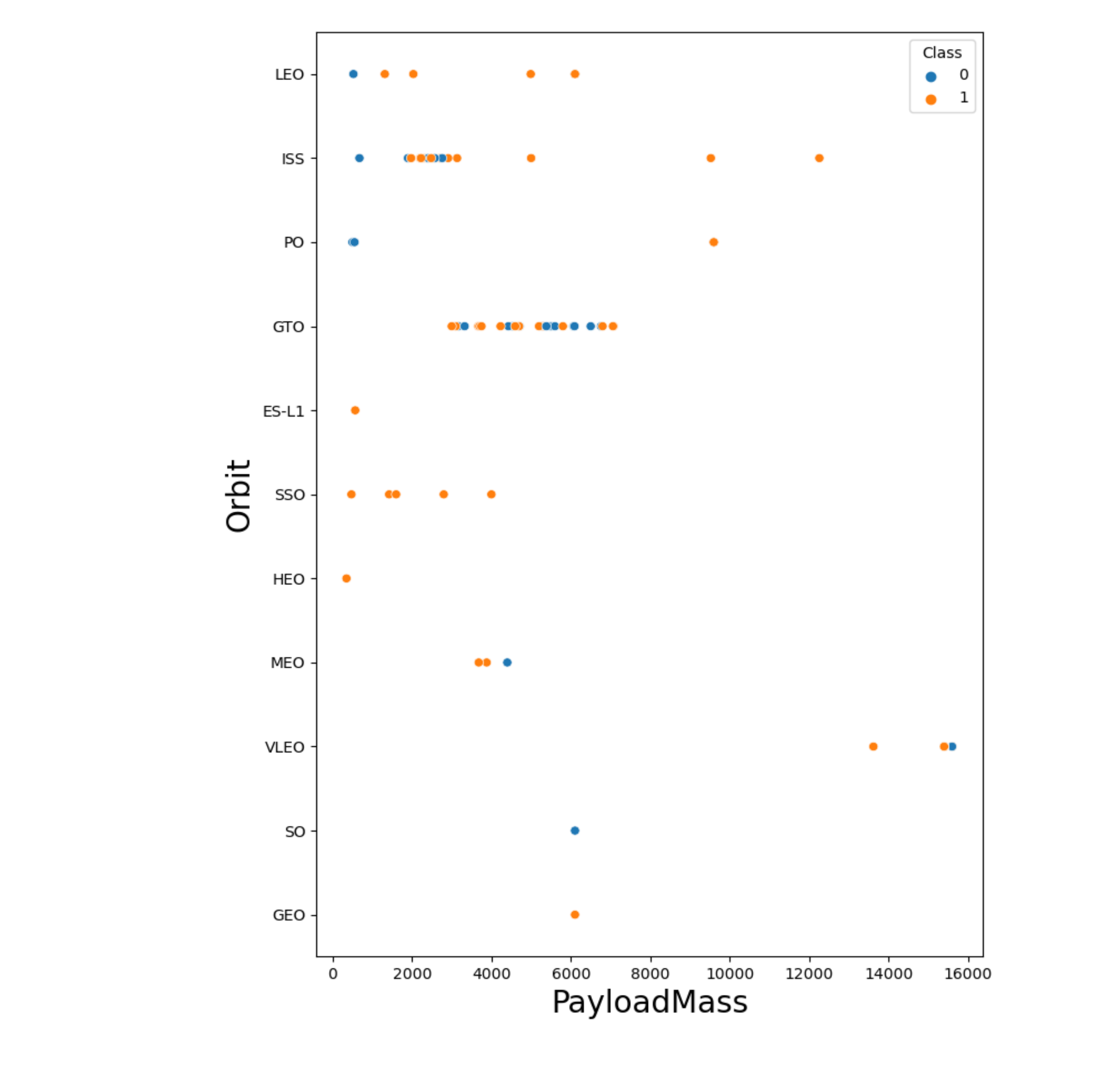
As you can see, the most successful orbit type are ES-L1, GEO, HEO, SSO with 1.0 result at all, and the least successful orbit type are SO with 0 result, GTO with 0.5 result and so on.

Flight Number vs. Orbit Type



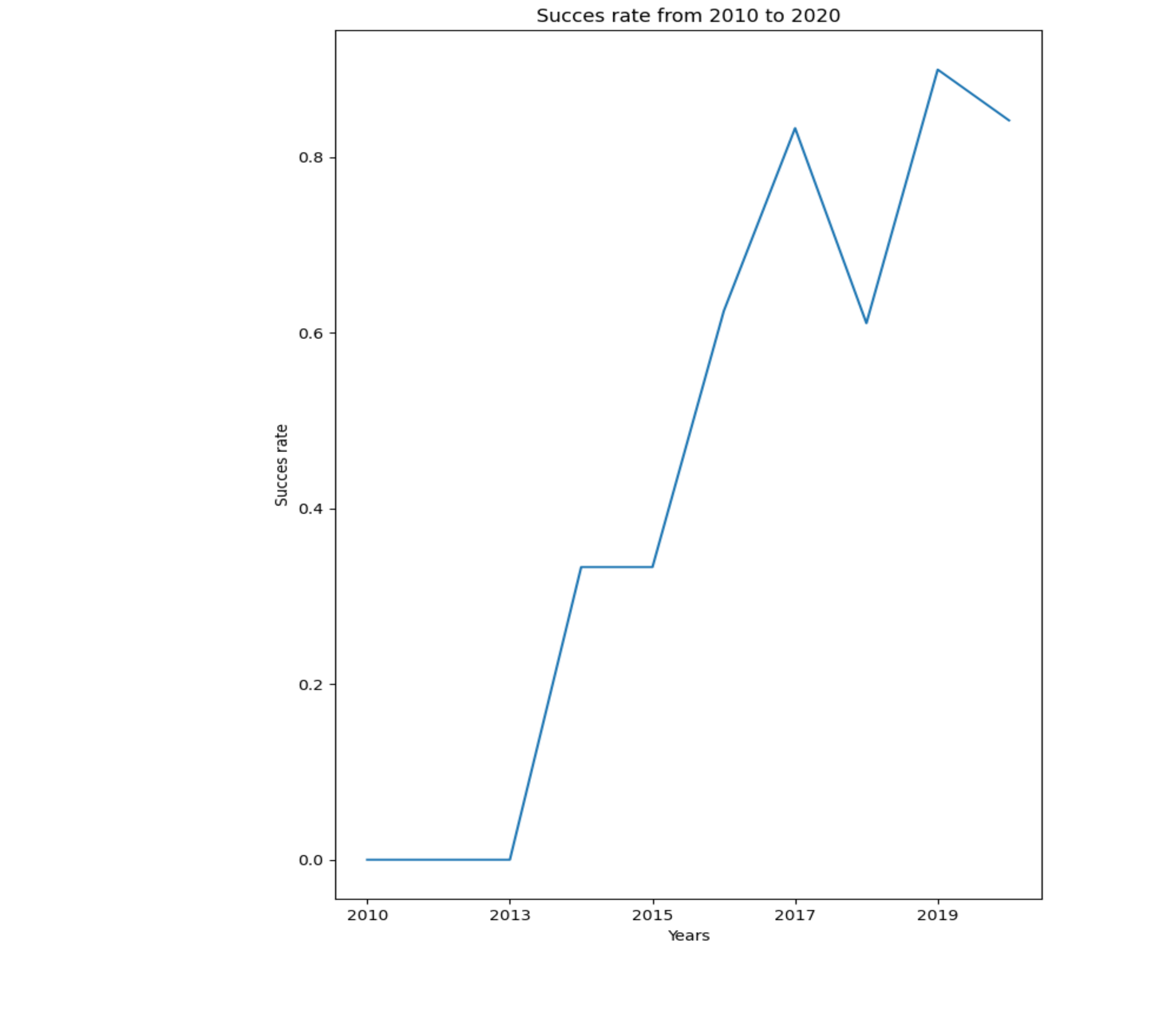
This plot shows more preferable type of orbit among all launches and there are LEO, ISS, Po, GTO, VLEO. Others are less usable.

Payload vs. Orbit Type



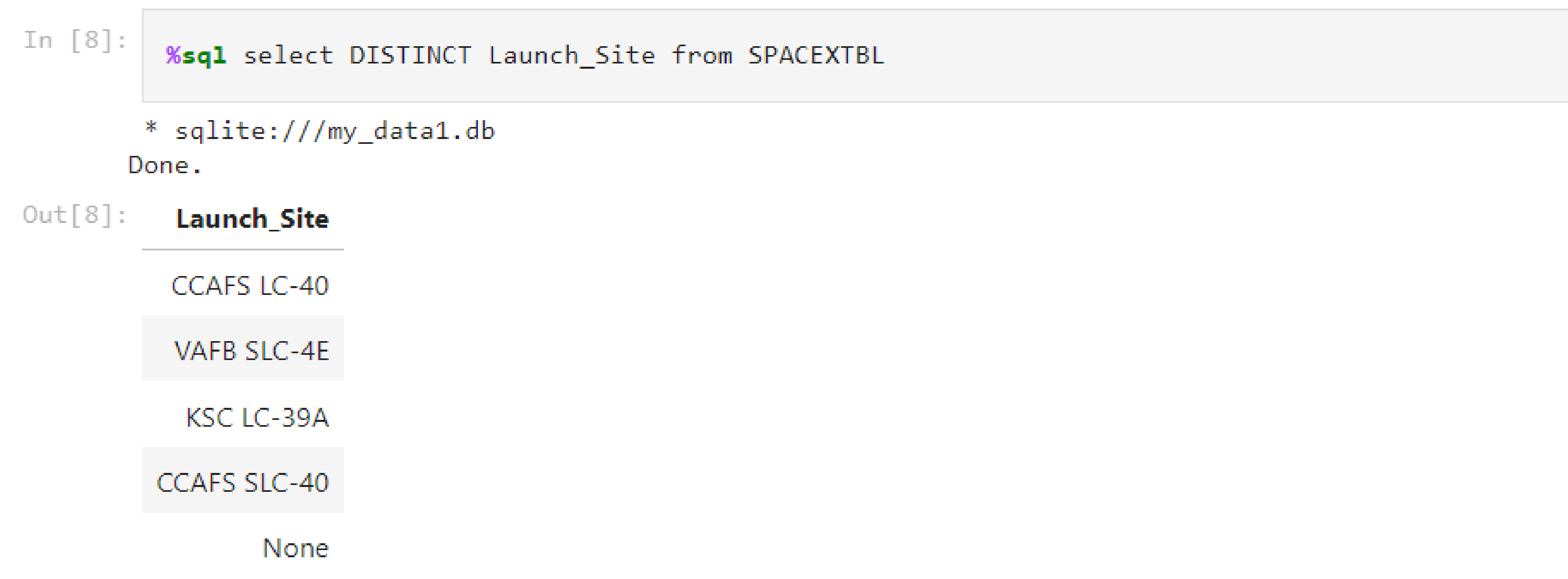
This plot shows that launches with type of orbit – SSO and payload mass less 5 000 KG have 100% successful rate.

Launch Success Yearly Trend



Since SpaceX realized their program of returning first stage of Falcon 9 in 2010, successful rate gradually increases, and now overall successful rate of all launches is above 80%.

All Launch Site Names



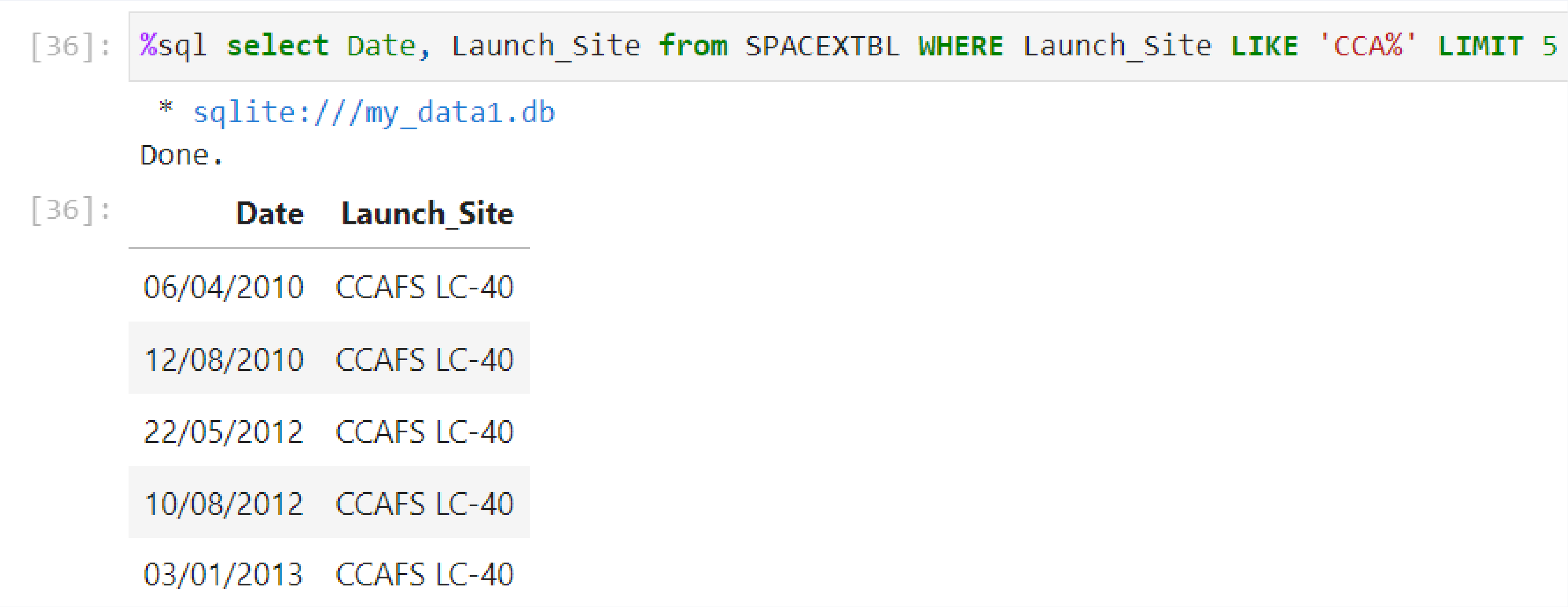
```
In [8]:  %sql select DISTINCT Launch_Site from SPACEXTBL

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

Out[8]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

As a result, there are 4 Launch Sites that have been using for launch Falcon 9

Launch Site Names Begin with 'CCA'



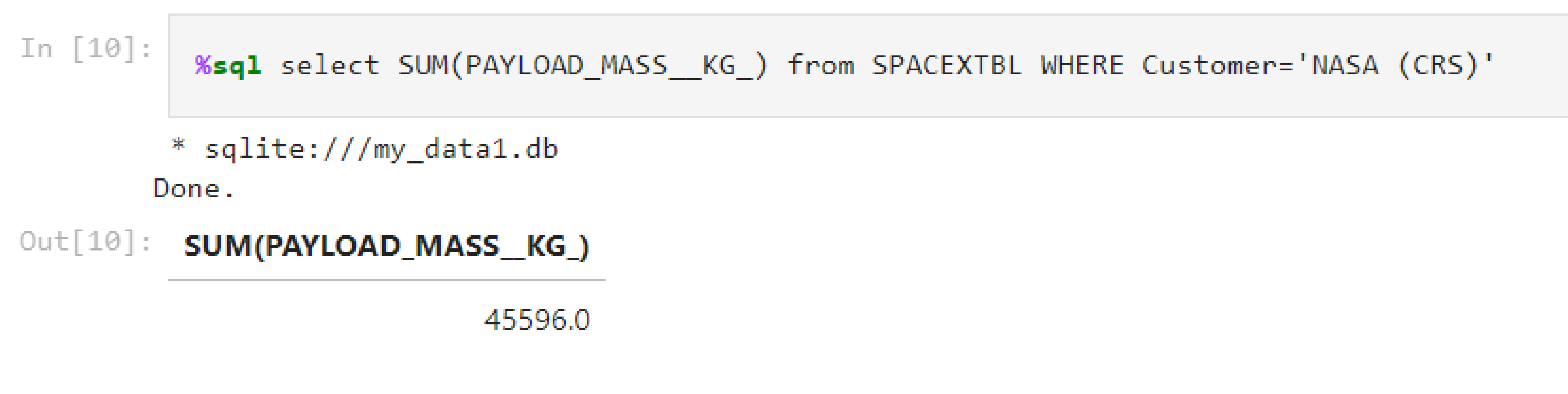
```
[36]: %sql select Date, Launch_Site from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

[36]:

Date	Launch_Site
06/04/2010	CCAFS LC-40
12/08/2010	CCAFS LC-40
22/05/2012	CCAFS LC-40
10/08/2012	CCAFS LC-40
03/01/2013	CCAFS LC-40

Four records with Site Name begin with ‘CCA’

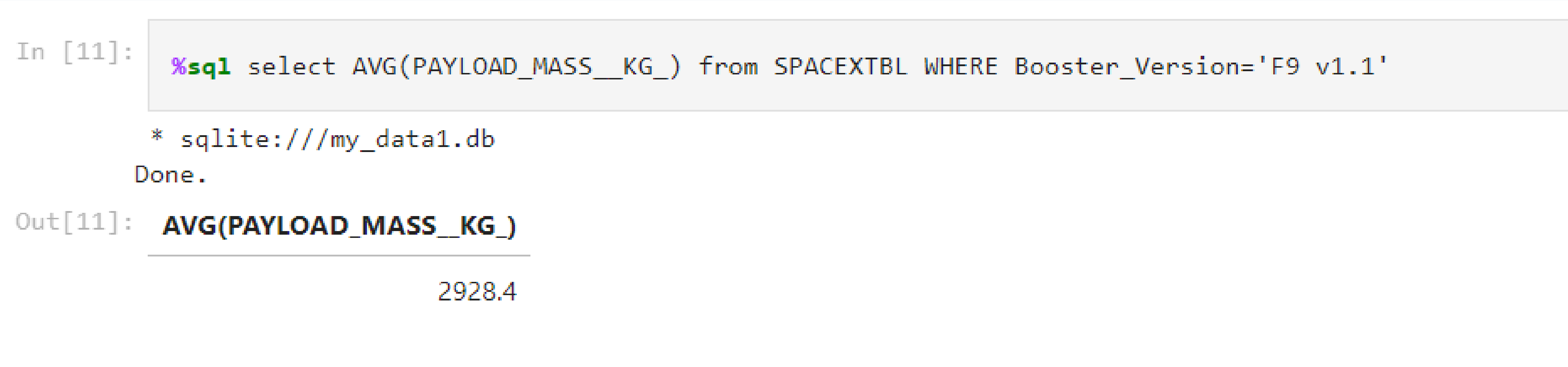
Total Payload Mass



```
In [10]:  %sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Customer='NASA (CRS)'

           * sqlite:///my_data1.db
          Done.

Out[10]:  SUM(PAYLOAD_MASS__KG_)
          ----------------------
                         45596.0
```

Total payload mass carried by Falcon 9 is 45 596.0 KG

Average Payload Mass by F9 v1.1



```
In [11]: %sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE Booster_Version='F9 v1.1'
```

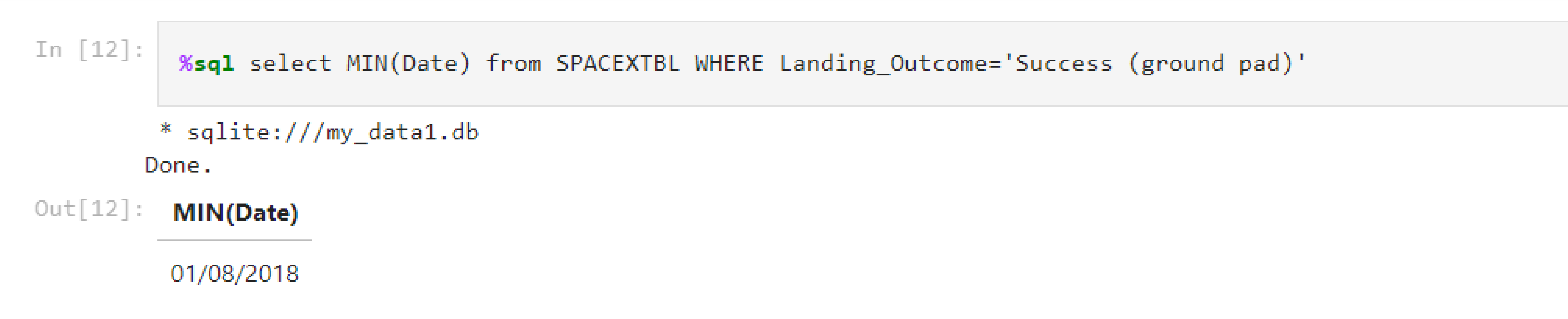
```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

AVG(PAYLOAD_MASS__KG_)
2928.4

Total payload mass carried by Falcon 9 v 1.1 is 2928.4 KG

First Successful Ground Landing Date



```
In [12]:  %sql select MIN(Date) from SPACEXTBL WHERE Landing_Outcome='Success (ground pad)'
```

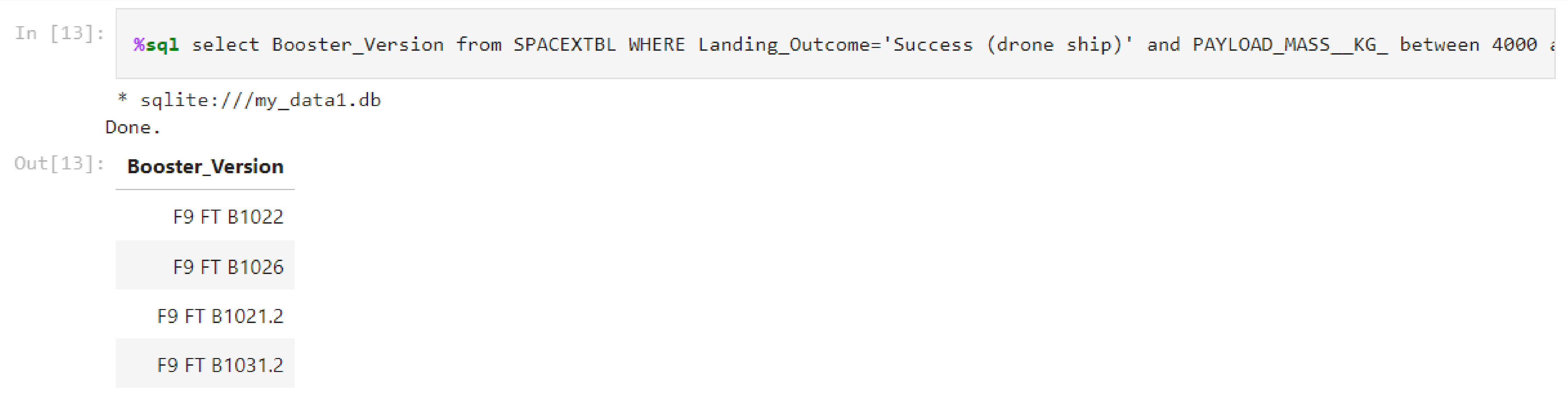
```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

MIN(Date)
01/08/2018

First Date of successful ground landing is 01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000



In [13]:

```
%sql select Booster_Version from SPACEXTBL WHERE Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 a
```

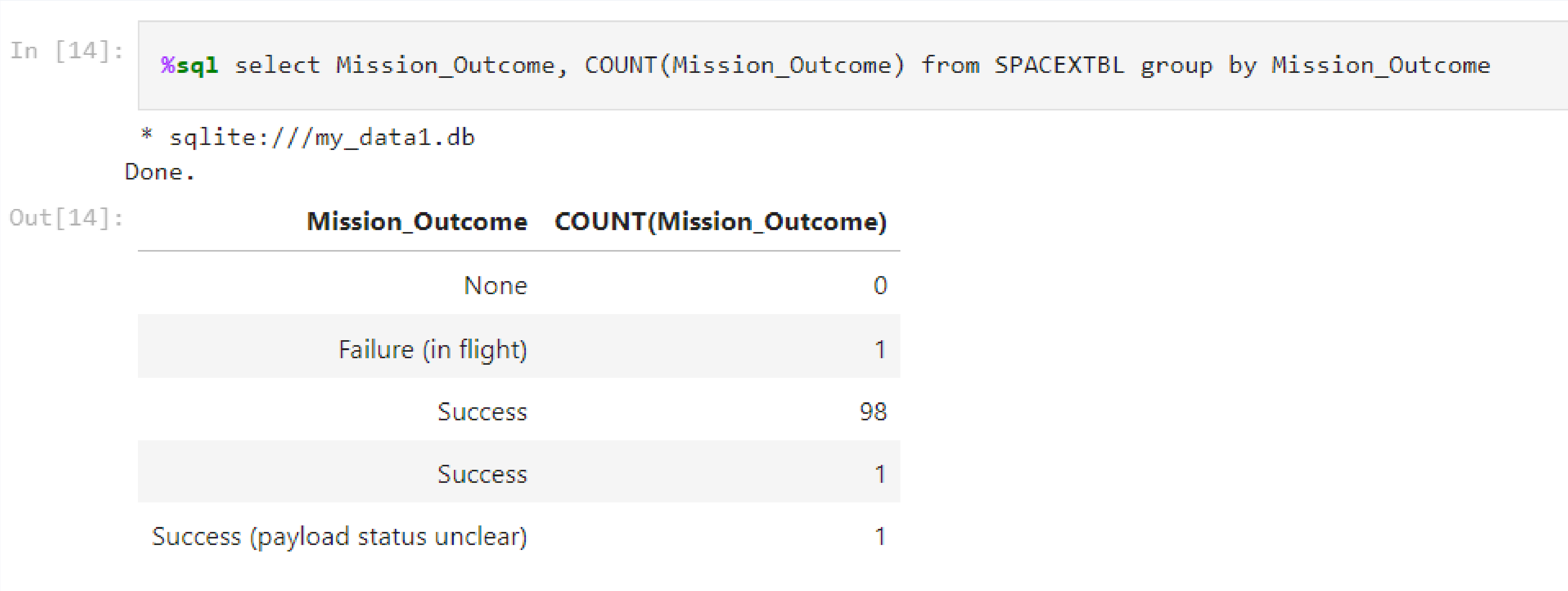
```
 * sqlite:///my_data1.db
Done.
```

Out[13]:

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

There are four boosters of Falcon 9, which had a successful drone ship landing with payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes



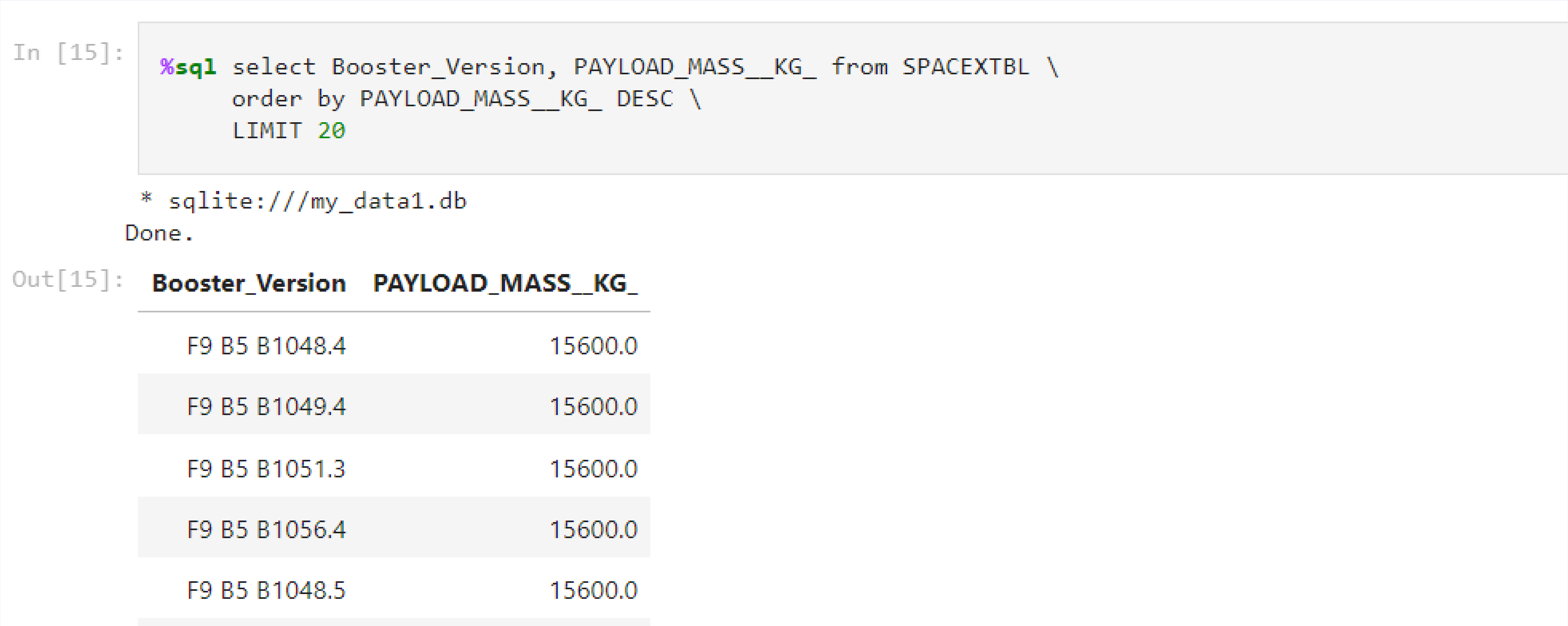
```
In [14]: %sql select Mission_Outcome, COUNT(Mission_Outcome) from SPACEXTBL group by Mission_Outcome

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

Out[14]:

Mission_Outcome	COUNT(Mission_Outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

As a result, we have got 100 missions with success outcome and 1 mission with failure outcome.

Boosters Carried Maximum Payload



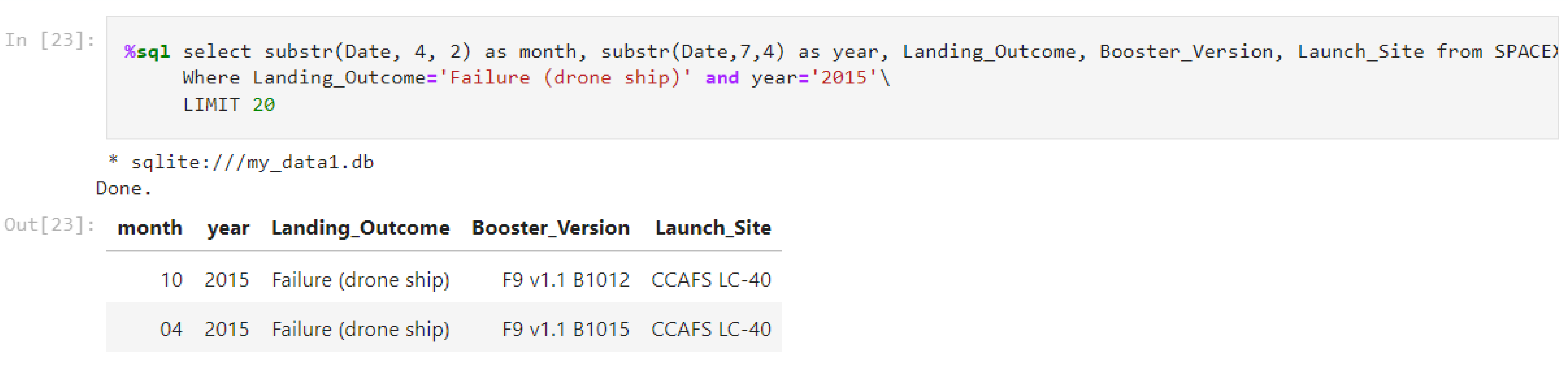
```
In [15]:  %sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTBL \
              order by PAYLOAD_MASS__KG_ DESC \
              LIMIT 20

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

Out[15]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0

As a result, we found out that maximum payload mass of 15 600 KG is taken to booster version F9 B5 B1048.4, F9 5 B1049.4, F9 B5 B1051.3 and so on.

2015 Launch Records



In [23]:

```
%sql select substr(Date, 4, 2) as month, substr(Date,7,4) as year, Landing_Outcome, Booster_Version, Launch_Site from SPACEX
     Where Landing_Outcome='Failure (drone ship)' and year='2015'\
     LIMIT 20
```

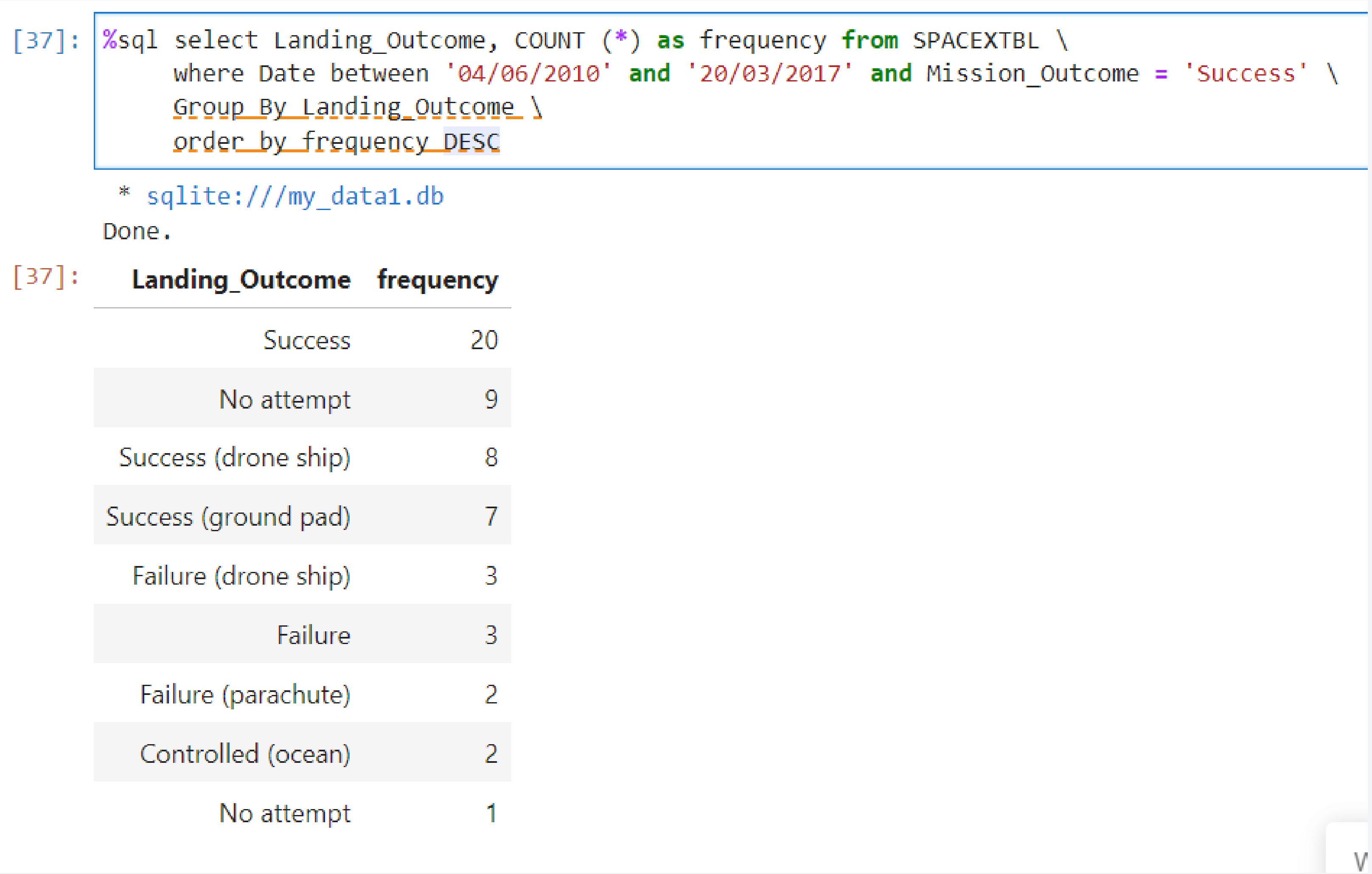
```
 * sqlite:///my_data1.db
Done.
```

Out[23]:

month	year	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

As a result, we see two mission with failure outcome in 2015

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



```
[37]: %sql select Landing_Outcome, COUNT (*) as frequency from SPACEXTBL \
          where Date between '04/06/2010' and '20/03/2017' and Mission_Outcome = 'Success' \
          Group By Landing_Outcome \
          order by frequency DESC

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

[37]:

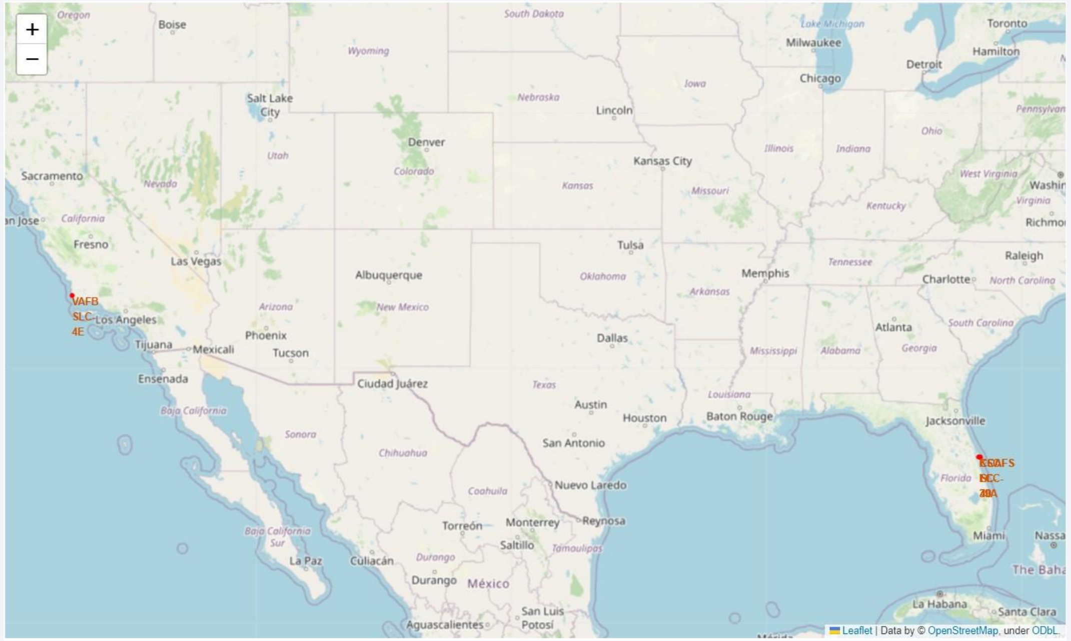
Landing_Outcome	frequency
Success	20
No attempt	9
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

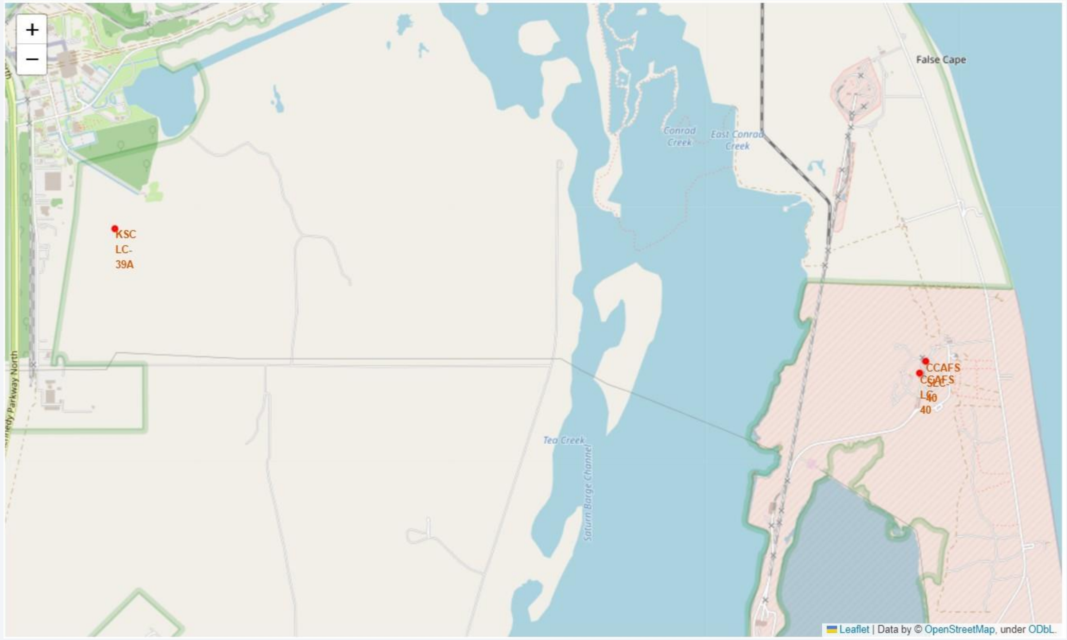
As a result, we have got the most frequently result is 20 times of ‘Success’ with no details, other variables have more details.

Section 3

Launch Sites Proximities Analysis

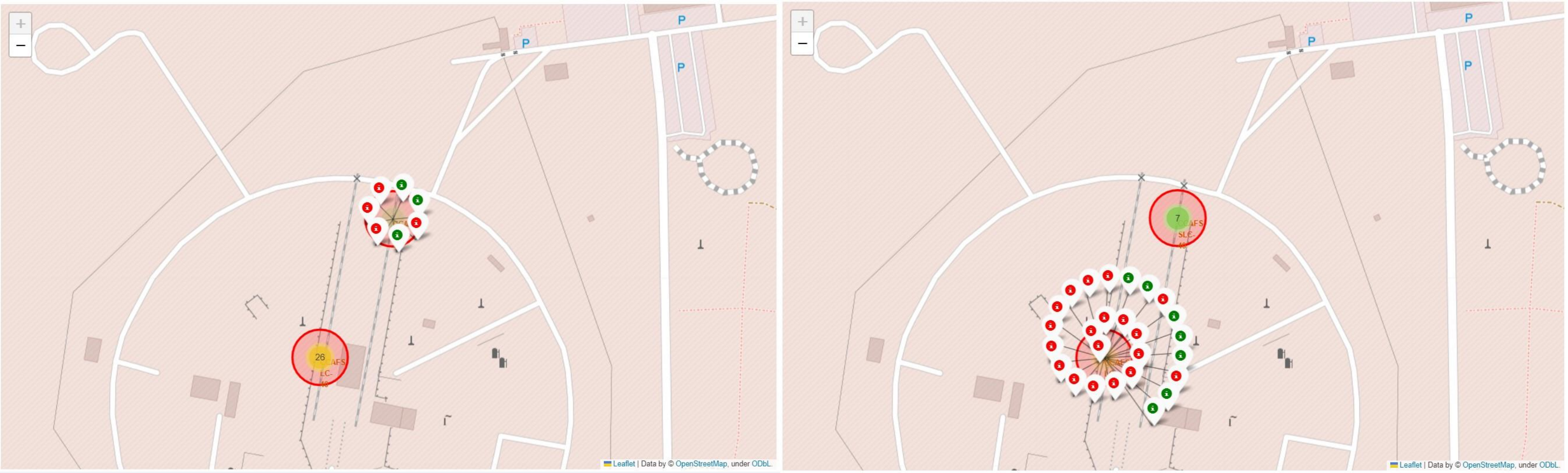
Site map with markers of Launch Site





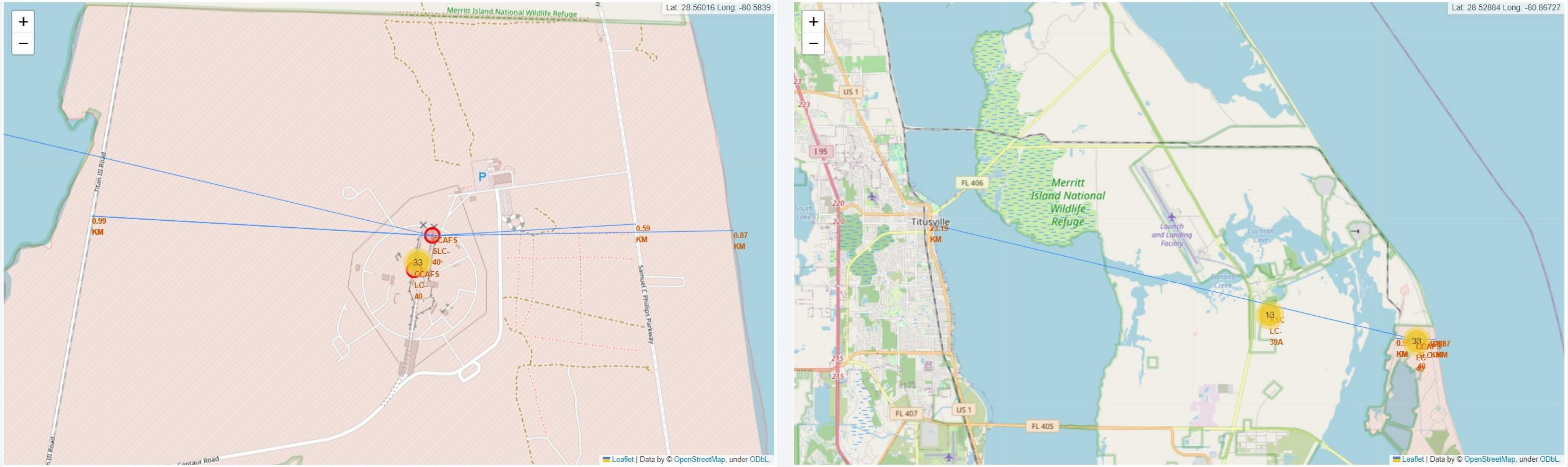
This maps show that we have 4 Launch Sites here, three of them on east coast in Florida and one on the west coast in California.

Site map with colored outcomes of Launch sites


+
−
P
26
7
Leaflet | Data by © OpenStreetMap, under ODbL.
Leaflet | Data by © OpenStreetMap, under ODbL.

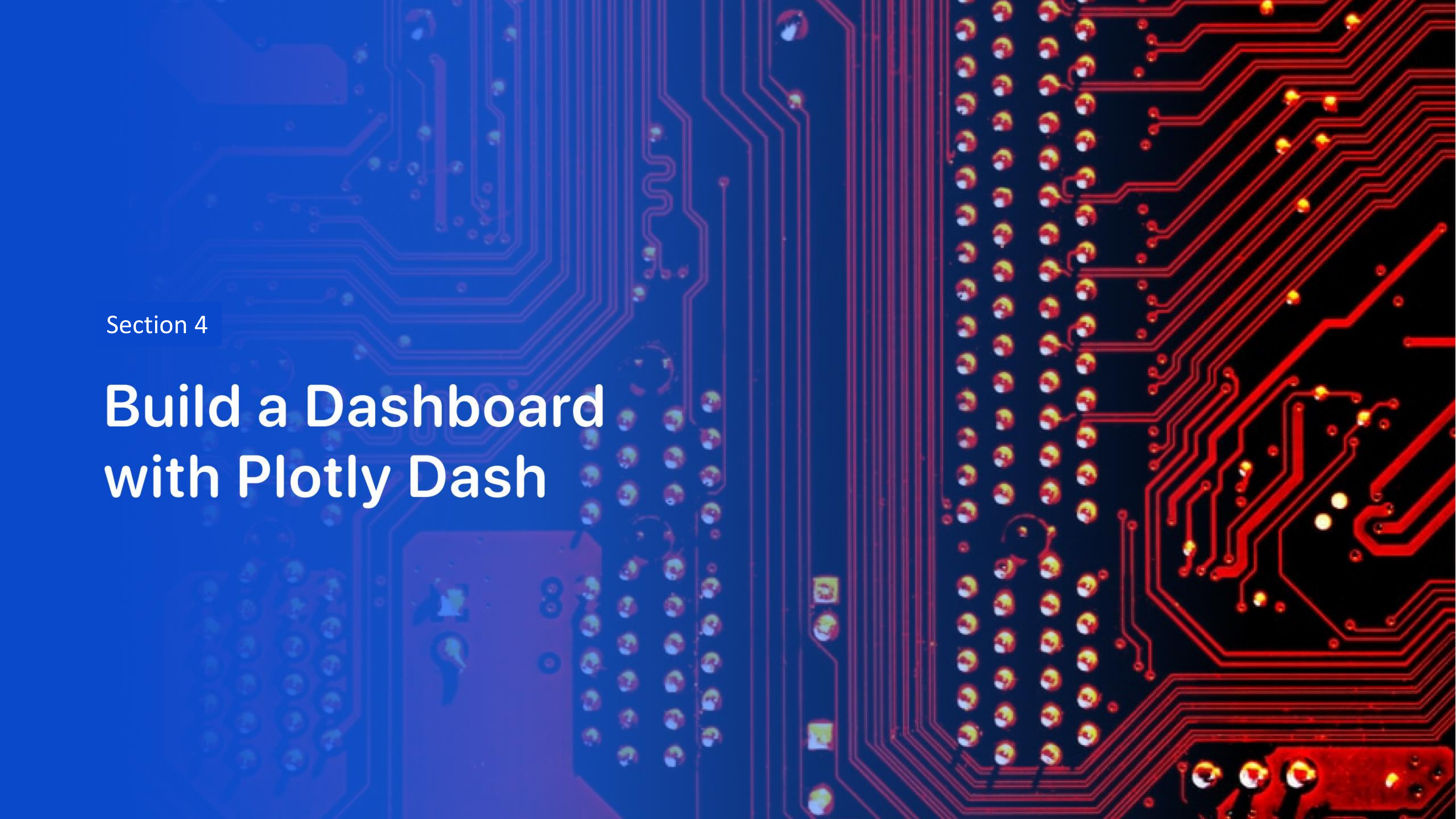
As you can see Launch site CCAFS SLC-40 has only 7 launches with 4 of them was successful and CCAFS LC-40 has 26 launches, and they were mostly not successful (19 of them are failed and 7 of them are successful)

Site map with markers and lines to its proximities



Lat: 28.56016 Long: -80.5839
Merritt Island National Wildlife Refuge
Titan III Road
0.99 KM
0.59 KM
0.87 KM
CCAFS SLC-40
33
CCAFS LC-40
Samuel C Phillips Parkway
Centaur Road
Leaflet | Data by © OpenStreetMap, under ODbL.
Lat: 28.52884 Long: -80.86727
US 1
I 95
FL 406
Merritt Island National Wildlife Refuge
Titusville
23.19 KM
Launch and Landing Facility
13
LC-39A
33
FL 407
FL 405
Leaflet | Data by © OpenStreetMap, under ODbL.

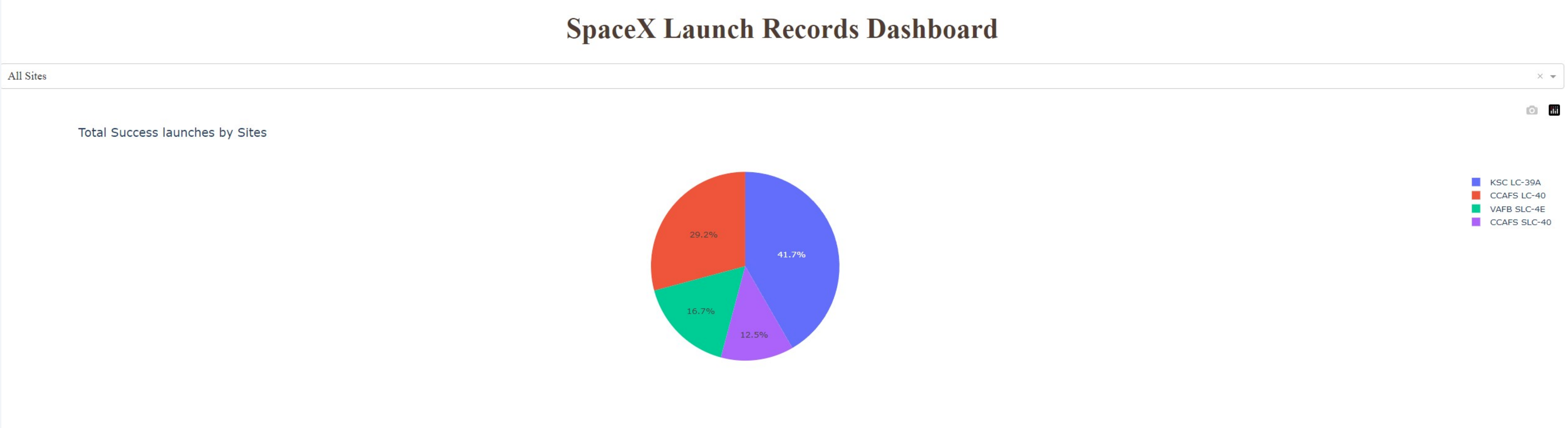
Launch sites CCAFS SLC-40 and CCAFS LC-40 are located close to coastline (0.87 KM), railroad (0.99 KM), highway (0.59 KM) and city (23.19 KM) The reason is logistic, you can't provide and develop your research without ongoing logistic. So, location near city is a way to workers stay home and feel motivated.



Section 4

Build a Dashboard with Plotly Dash

Dashboard for All Sites with launch success count

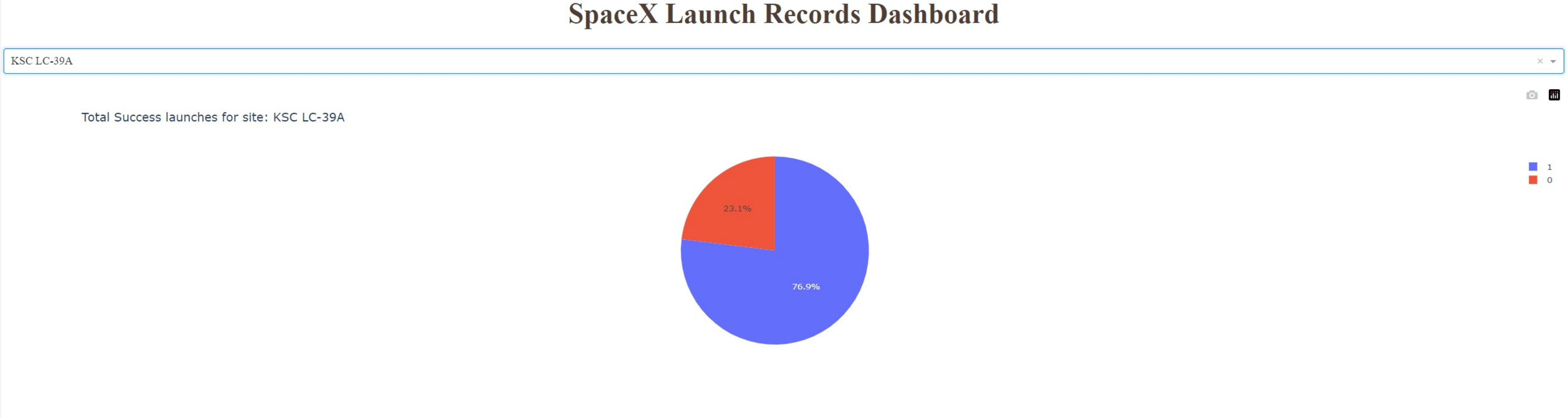


SpaceX Launch Records Dashboard

All Sites

As you can see the most successful launch site is KSC LC-39A and among all success launches its share counts 41.7%

Dashboard with the most successful Launch Site



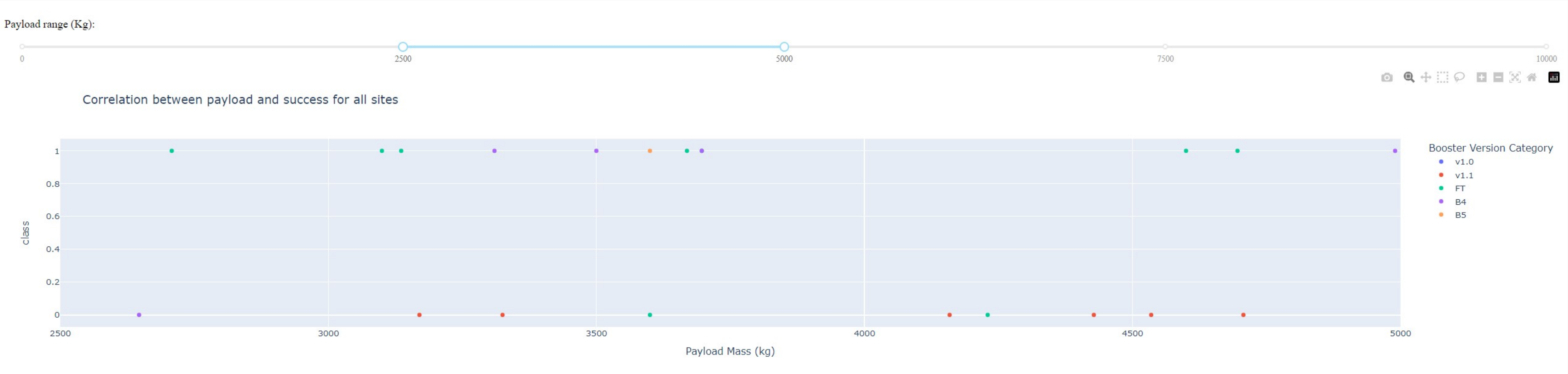
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success launches for site: KSC LC-39A

As you can see Launch Site KSC LC-39A has a 76,9% successful rate over its launches

Dashboard with different payload mass and successful rates for all sites

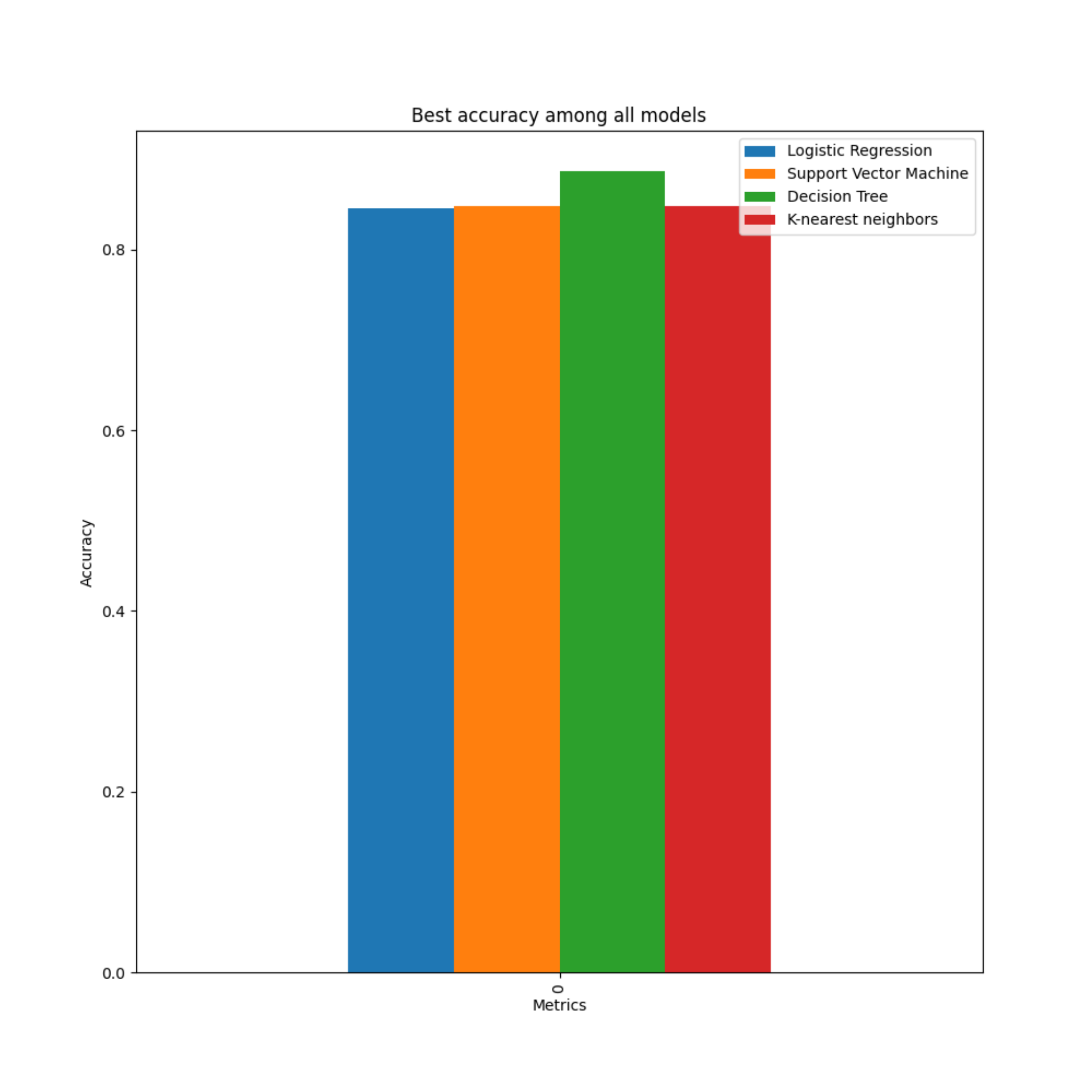


As you can see, booster version v 1.1 with payload mass between 2 500 and 5 000 KG has only failed outcomes.

Section 5

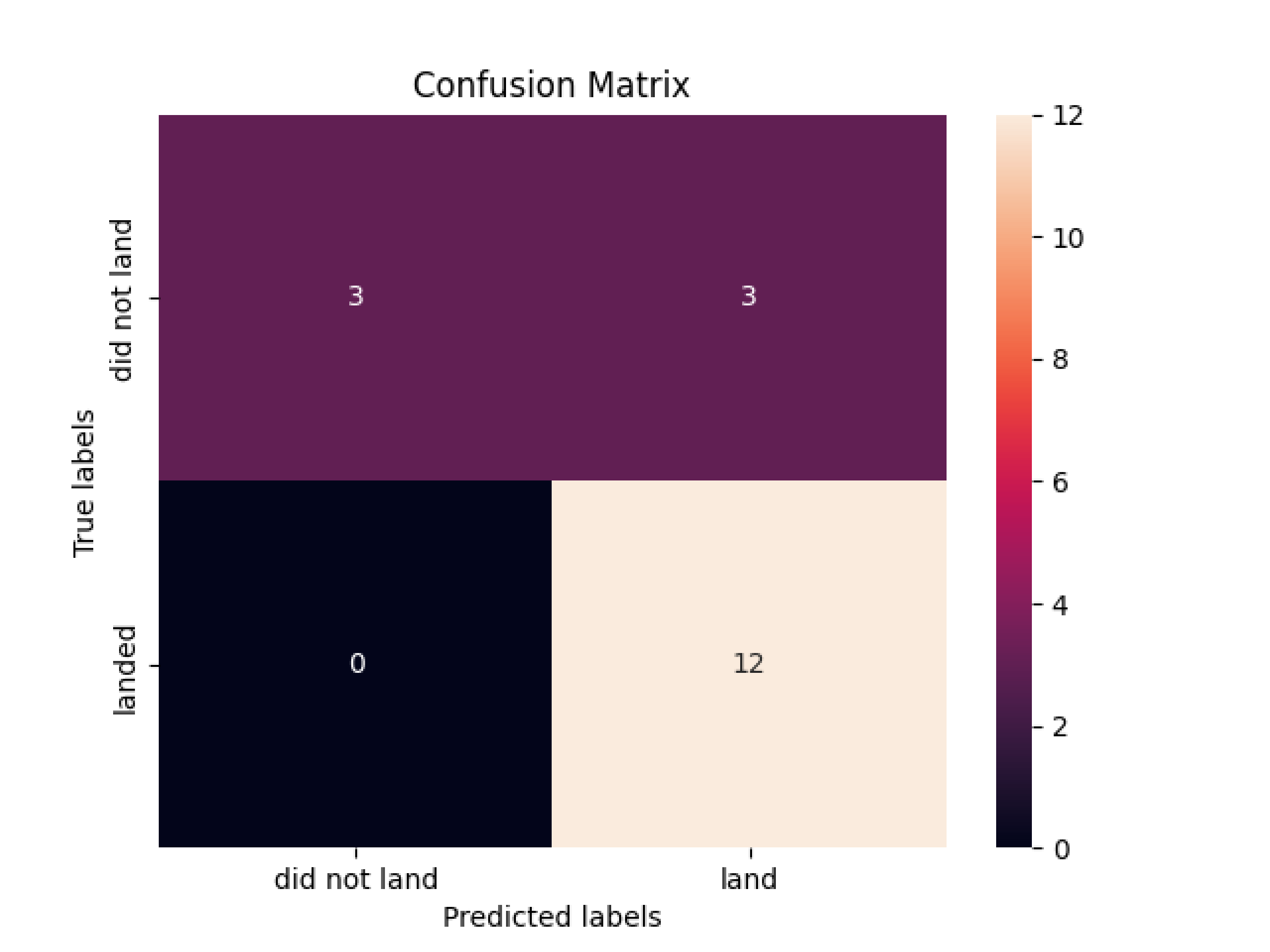
Predictive Analysis (Classification)

Classification Accuracy



As you can see decision tree model has best accuracy among all models

Confusion Matrix



As you can see the confusion matrix has next parameters:

TP (true positive): 3

FN (false negative): 3

FP (false positive): 0

FN (false negative): 12

Conclusions

The results achieved with decision tree model:

1. R2 score: The R2 score of 83.3% indicates that decision tree model explains 83.3% of the variance in the dependent variable using the independent variables in the model. This is a good score and suggests that model is performing well.
2. Best accuracy: The best accuracy of 88.75% indicates that model correctly classified 88.75% of the samples in the test dataset. This is also a good score and suggests that model is performing well.
3. True positives (TP): The number of true positives (3) indicates that model correctly predicted 3 positive cases out of all positive cases in the test dataset.
4. False positives (FP): The number of false positives (0) indicates that model did not incorrectly predict any negative cases out of all negative cases in the test dataset.
5. False negatives (FN): The number of false negatives (3) indicates that model incorrectly predicted 3 negative cases out of all positive cases in the test dataset.
6. True negatives (TN): The number of true negatives (12) indicates that model correctly predicted 12 negative cases out of all negative cases in the test dataset.

Based on these results, it seems like decision tree model is performing well overall but could benefit from further tuning to reduce the number of false positives.

Appendix

- GitHub

Thank you!