IBM Developer
SKILLS NETWORK

Data Science Capstone Project

BY SAMIKSHA BHARGAVA

02/05/2023

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

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

- Project background and context-

Space X 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 Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers-

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program?

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 coding was applied to categorial features.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data Collection was done using get request to SpaceX API.
- Then we decoded the response content as a json using .json() and turn it into a Pandas DataFrame using .json_normalize().
- Next, we filtered the DataFrame.
- We also cleaned the data and dealt with missing values.
- Other method we used to collect data was web scraping from Wikipedia.
- In this we request the Falcon9 launch wiki page from its URL
- Next, we created a BeautifulSoup object from the html response.
- We created a data frame by parsing the launch html tables.

Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request.
- Filtered the DataFrame.
- Data wrangling and formatting

> Here’s the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/1.Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- Request the Falcon9 launch wiki page from its URL.
- Created a BeautifulSoup object from the HTML response.
- Extracted all the columns or variable names from the HTML table header .
- Created a data frame by parsing the launch HTML tables.

> Here’s the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/Web%20Scraping.ipynb

Data Wrangling

Performed some Exploratory Data Analysis (EDA) to find some patterns in data to determine what would be the label for training supervised models.

- Calculated the number of launches on each site.
- Calculated the number and occurrence of each orbit and mission outcome of the orbits.
- Created a landing outcome label

> Here's the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

Performed Exploratory Data Analysis and

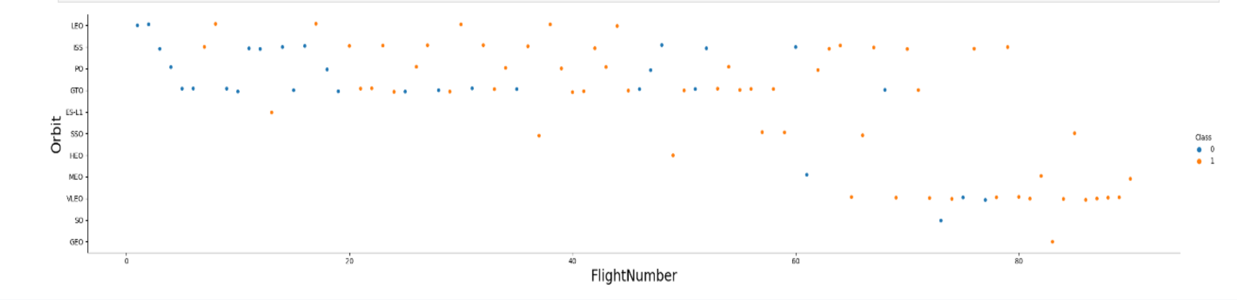
Feature Engineering using Pandas and Matplotlib.

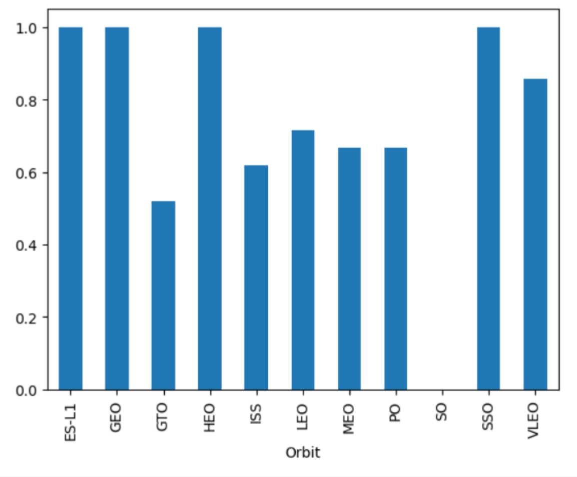
> Here' some of the visualizations :

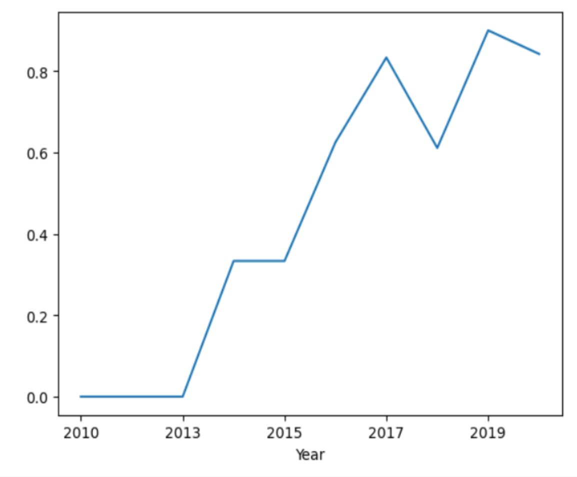
○ The line chart visualizes the launch success yearly trend.

○ The bar graph shows the relationship between success rate of each orbit type.

○ The Scatter plot shows the relationship between Flight Number and Orbit type.







> Here’s the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/EDA%20with%20Visulaization.ipynb

EDA with SQL

SQL Queries performed to find the following:

- Names of the unique launch sites in the space mission.
- 5 records where launch sites begin with the string 'CCA’.
- Total payload mass carried by boosters launched by NASA(CRS).
- Average payload mass carried by booster version F9 v1.1.
- Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass (using a subquery).
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

> Here’s the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/EDA%20SQL.ipynb

Build an Interactive Map with Folium

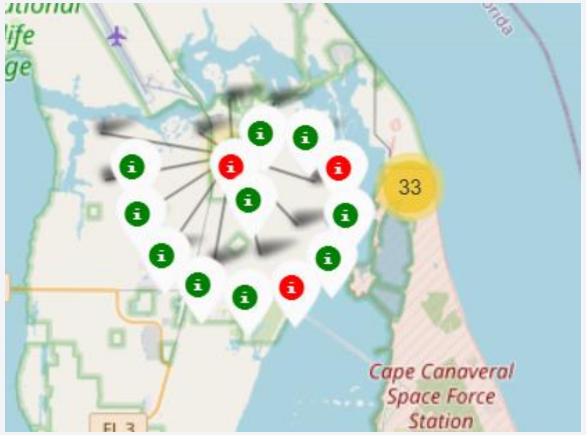
- We marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.

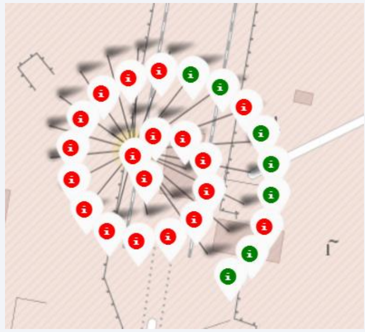
- We assigned the feature launch outcomes (failure or success) to class 0 and 1.i.e., 0 for failure, and 1 for success.

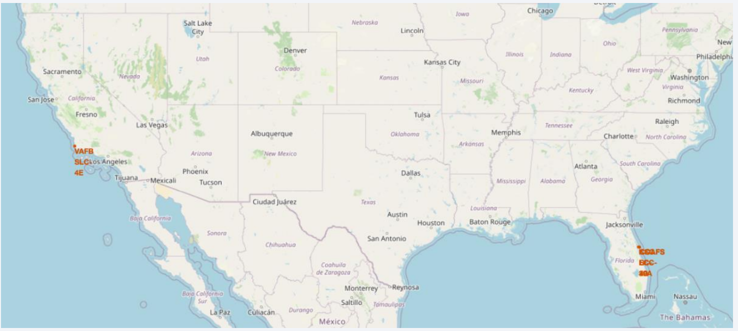
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.

> Here's the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb





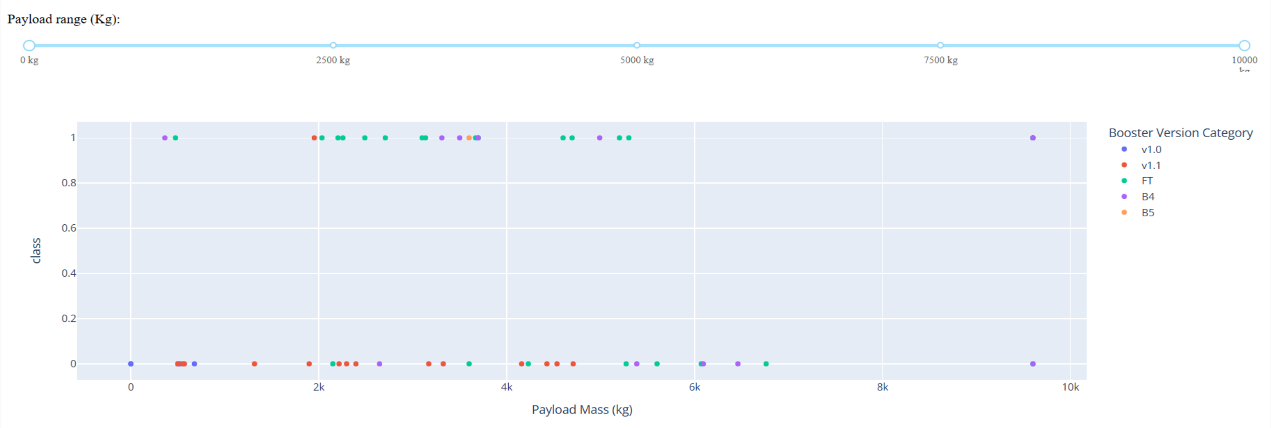


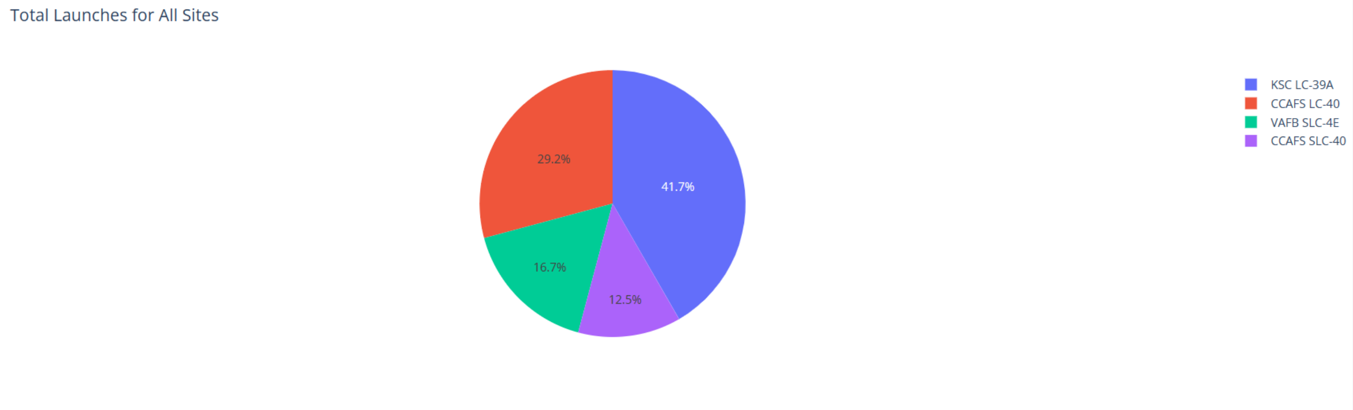
Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie chart showing the total launches by each different sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

> Here’s the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/spacex_dash_app.py





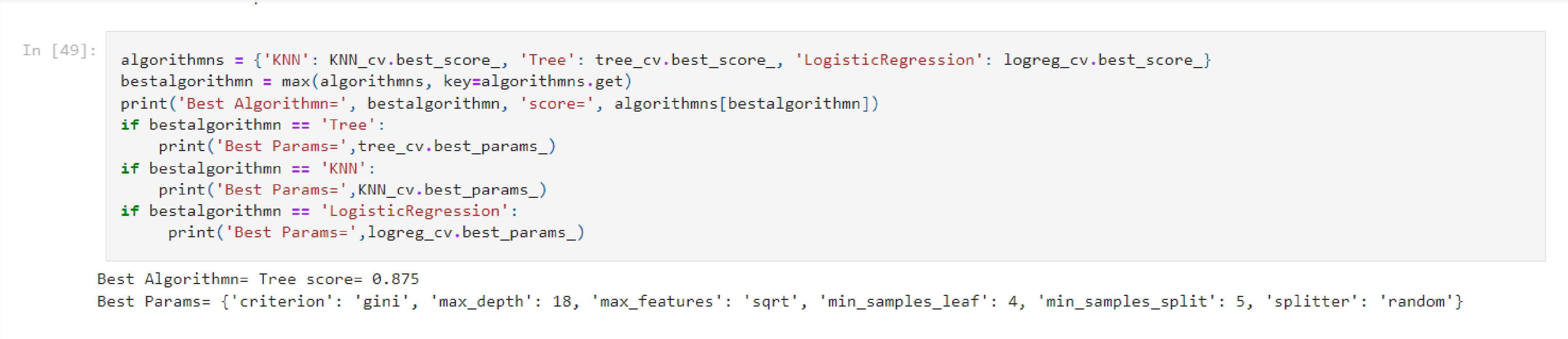
Predictive Analysis (Classification)

- First, we import libraries like panda, NumPy, matplotlib etc., and loaded the DataFrame.
- Then, use the function train_test_split to split the data .Built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used the accuracy as the metric of the model.
- At last, we found the best performing classification model.

>Here's the link to the notebook:

https://github.com/SamikshaBhargava16/Data-Science-Capstone-falcon/blob/main/Machine%20Learning.ipynb

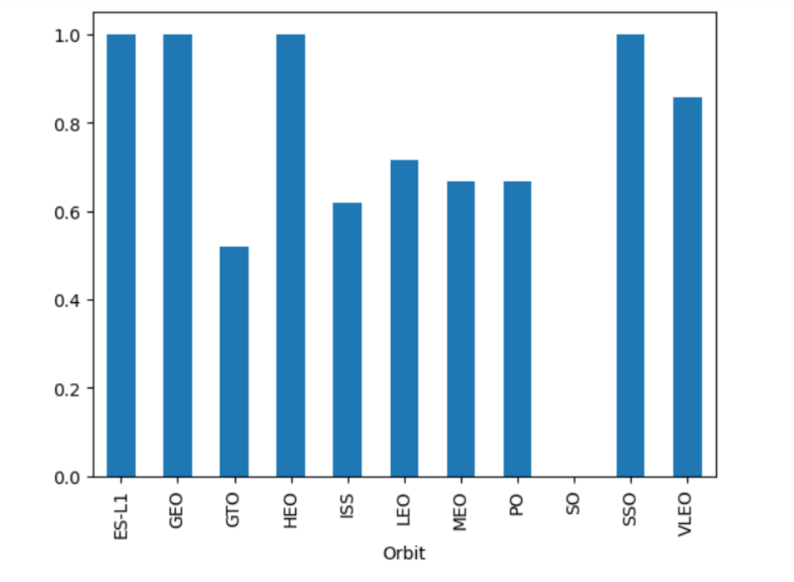
Results

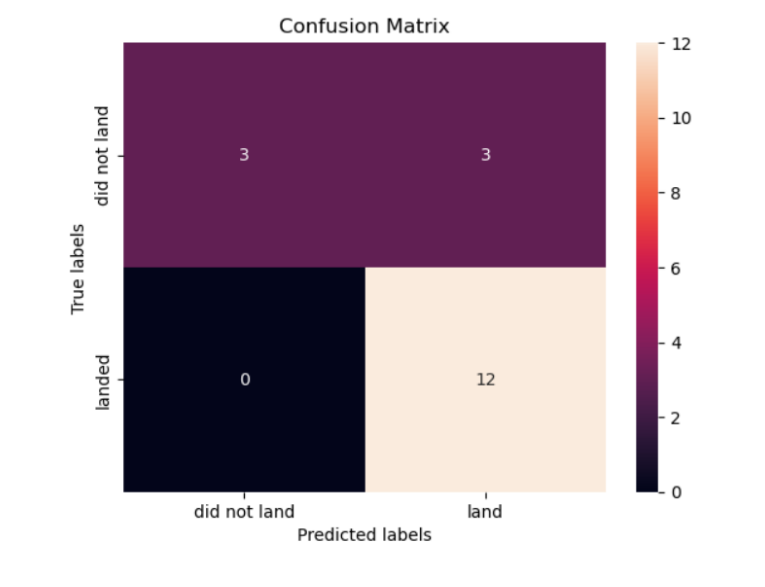


```
In [49]:
algorithmns = {'KNN': KNN_cv.best_score_, 'Tree': tree_cv.best_score_, 'LogisticRegression': logreg_cv.best_score_}
bestalgorithmn = max(algorithmns, key=algorithmns.get)
print('Best Algorithmn=', bestalgorithmn, 'score=', algorithmns[bestalgorithmn])
if bestalgorithmn == 'Tree':
    print('Best Params=',tree_cv.best_params_)
if bestalgorithmn == 'KNN':
    print('Best Params=',KNN_cv.best_params_)
if bestalgorithmn == 'LogisticRegression':
     print('Best Params=',logreg_cv.best_params_)
```

```
Best Algorithmn= Tree score= 0.875
Best Params= {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
```

- Query shows decision tree classifier as best machine learning algorithm
- Bar graph shows the orbits with higher success rates.



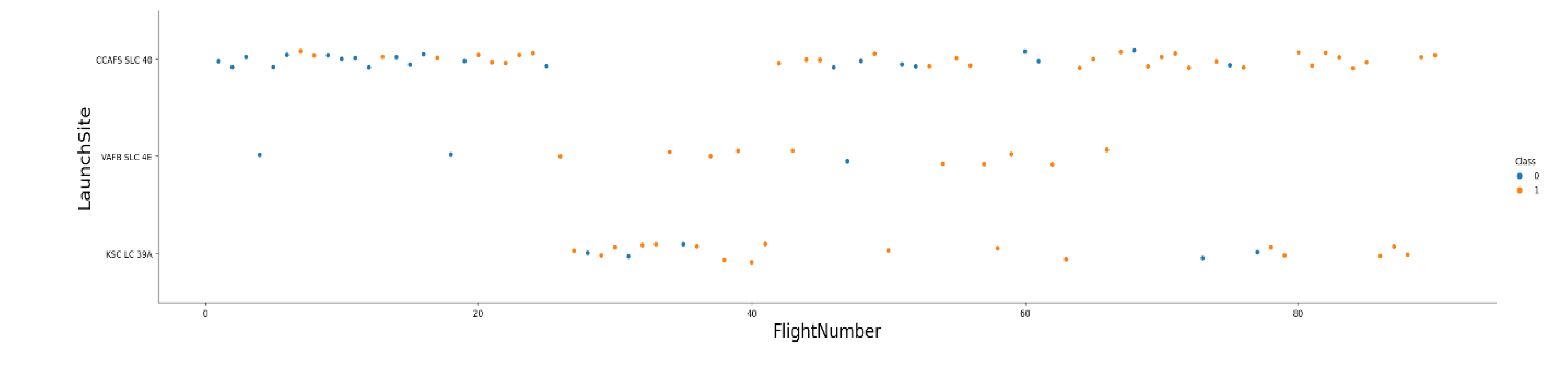


Section 2

Insights drawn from EDA

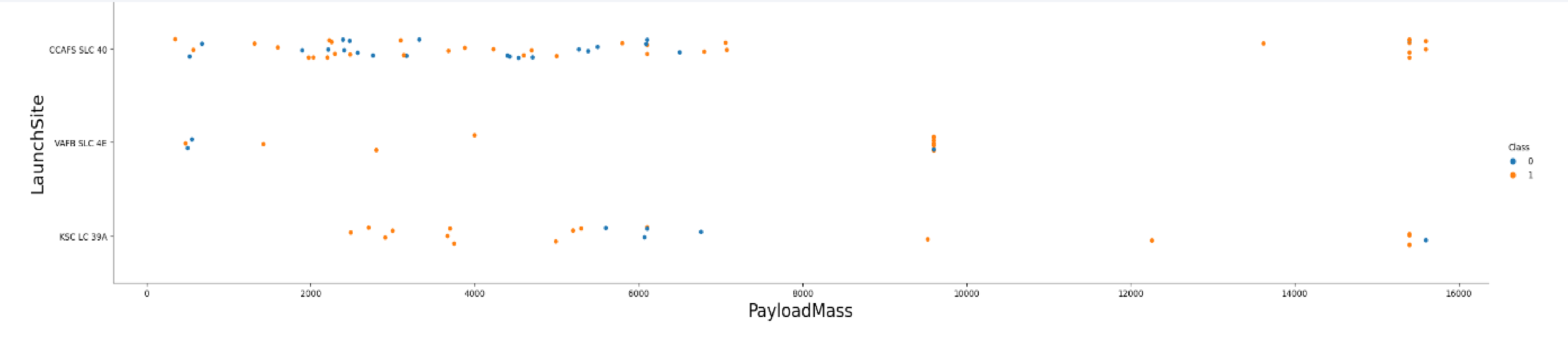
Flight Number vs. Launch Site

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



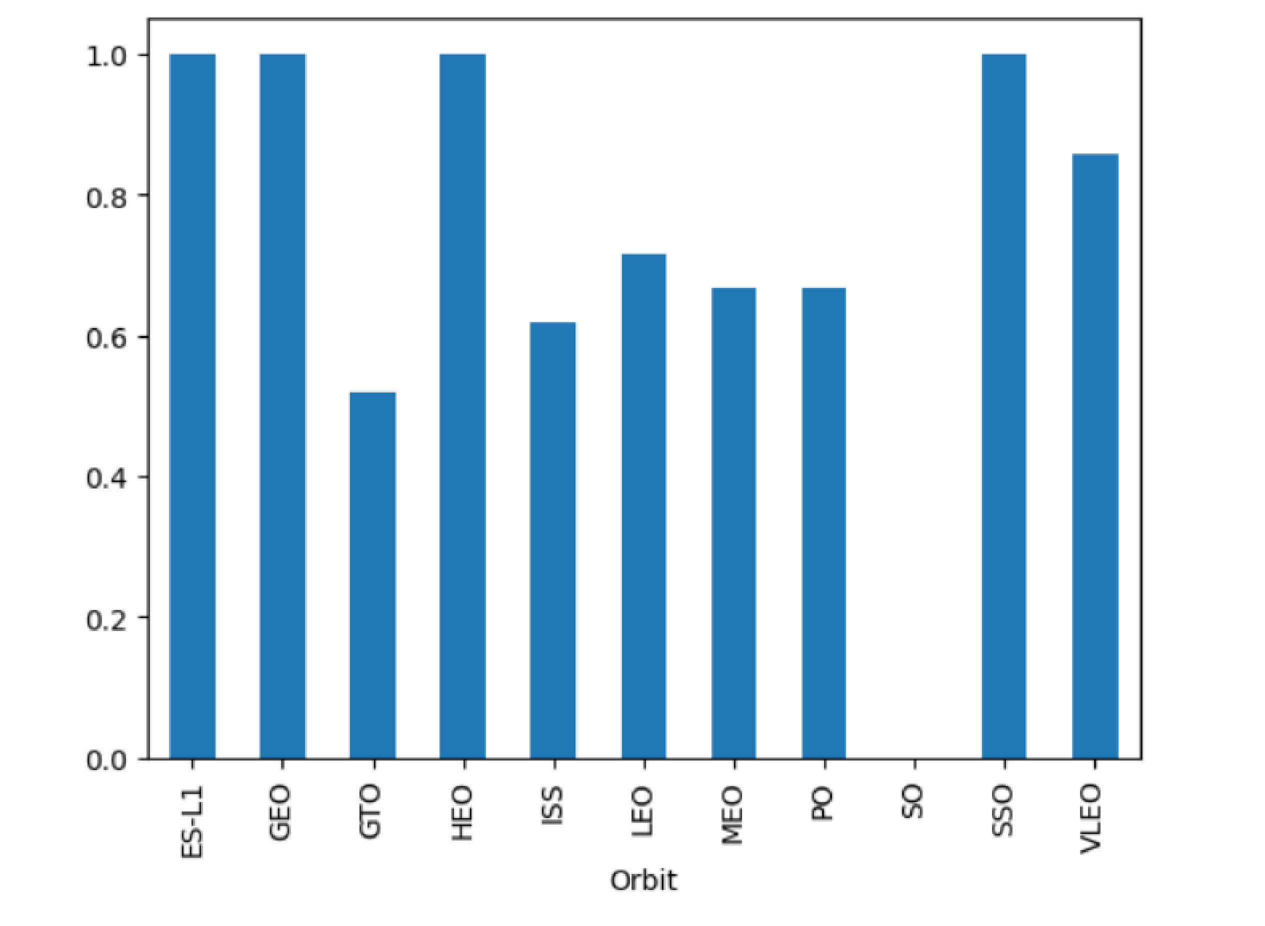
Payload vs. Launch Site

For the VAFB-SLC Launch site there are no rockets launched for heavy payload mass(greater than 10000).



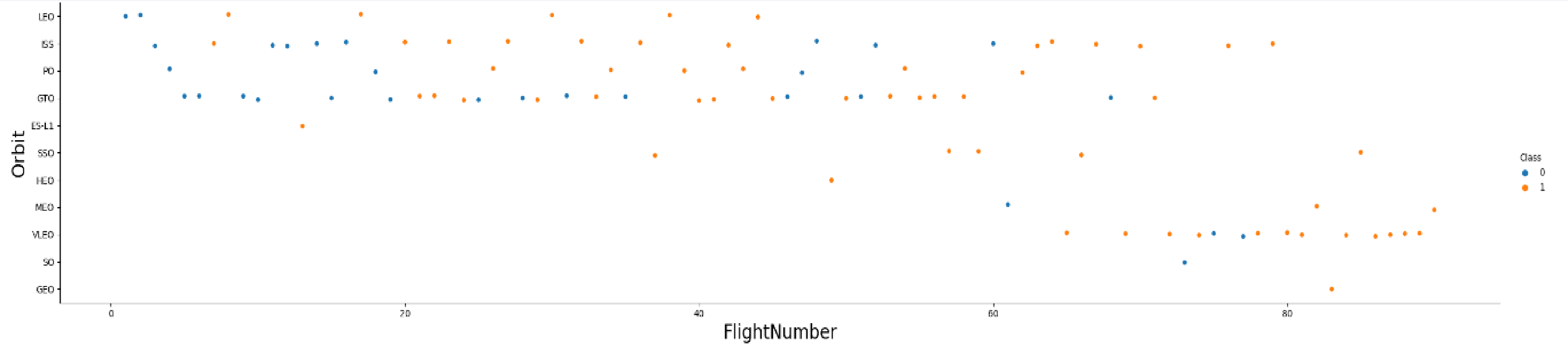
Success Rate vs. Orbit Type

In the bar graph it can be seen that four Orbits ES-L1, GEO, HEO and SSO have the highest success rate.



Flight Number vs. Orbit Type

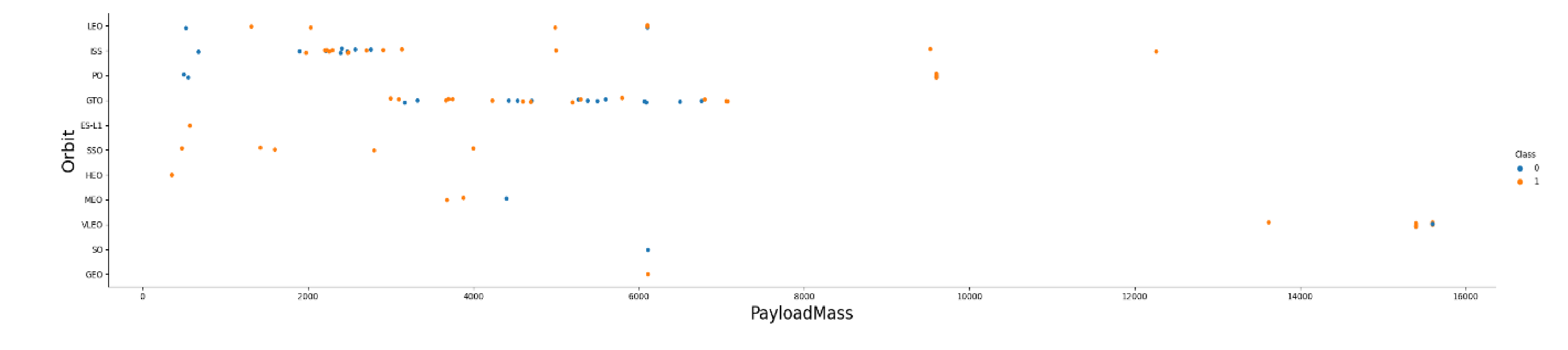
In the scatter plot we can see that in the LEO orbit the Success appears related to the Number of Flights, on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload vs. Orbit Type

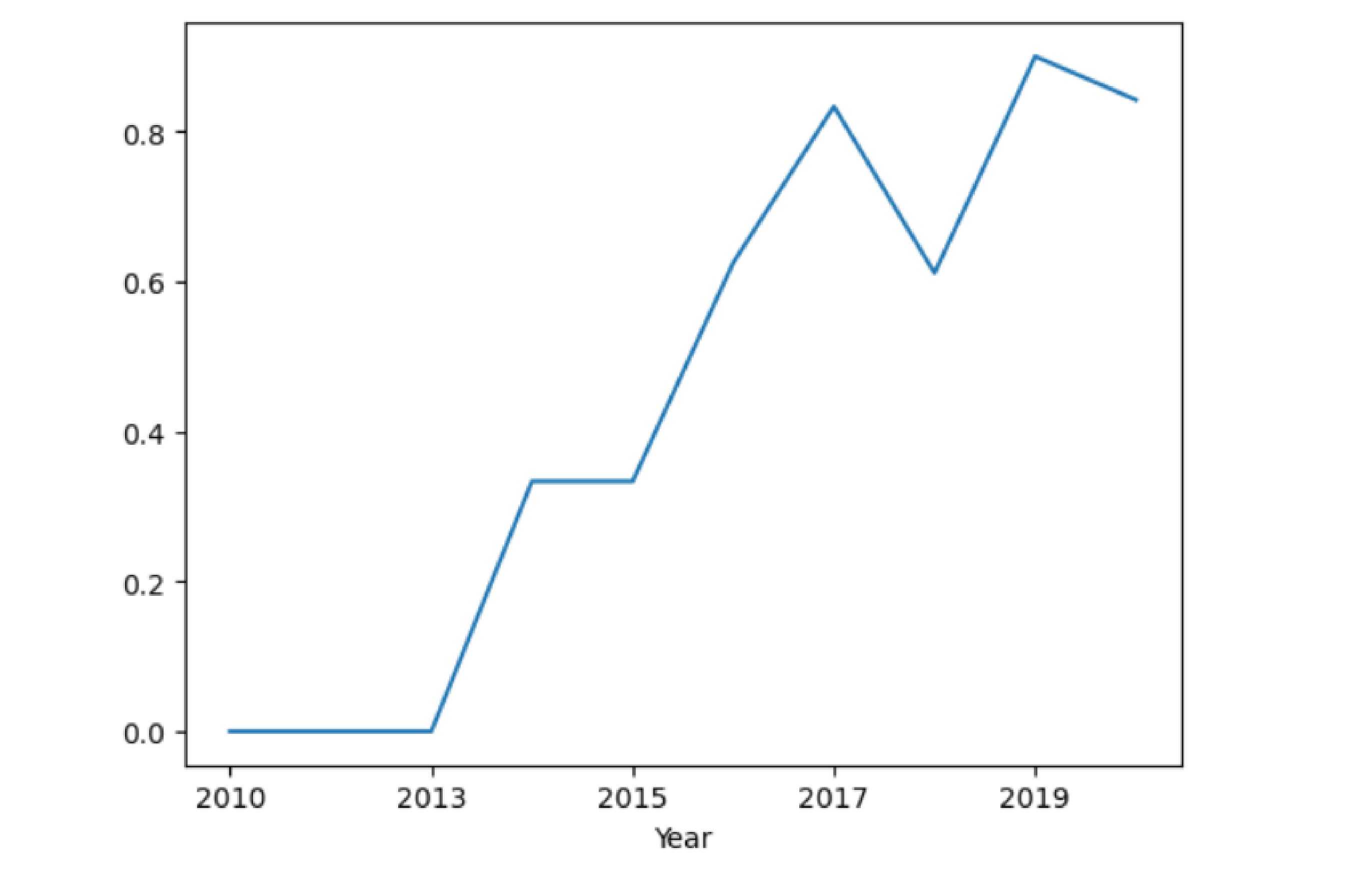
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



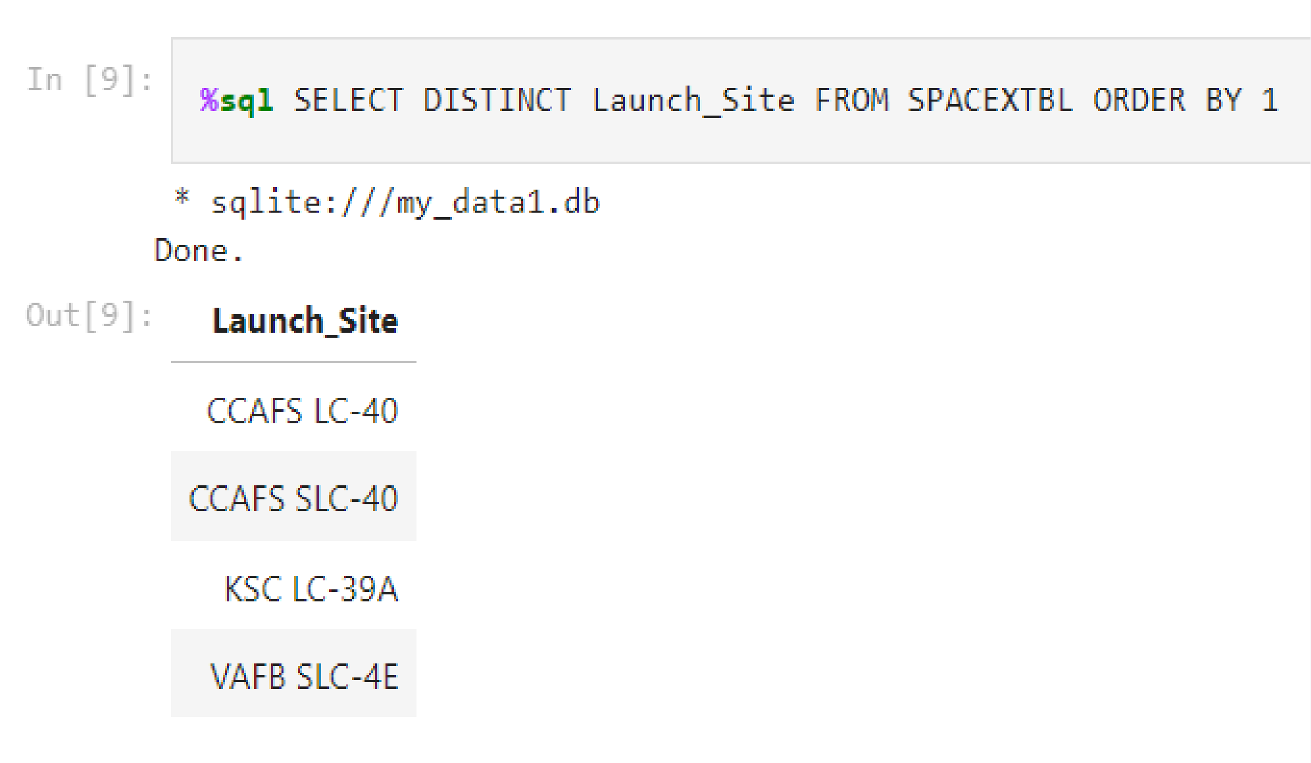
Launch Success Yearly Trend

In the line chart we can see that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

In this SQL query we used DISTINCT function to show unique sites from SpaceX data.



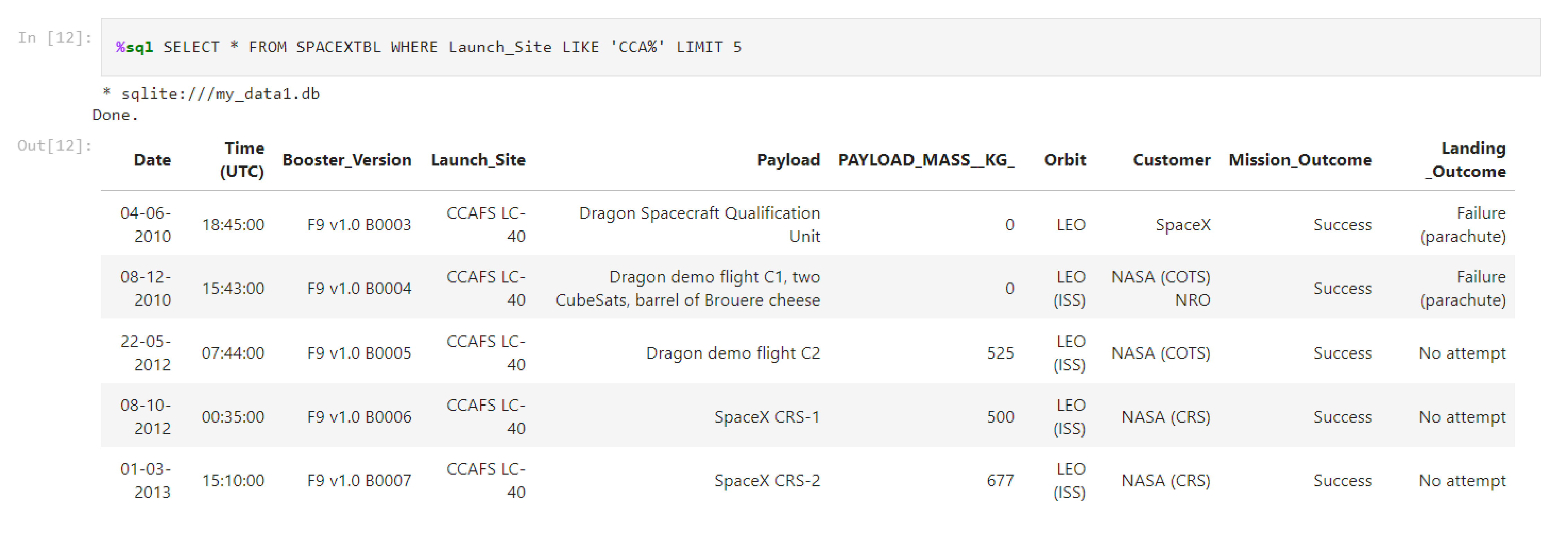
```
In [9]:  %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL ORDER BY 1

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

Out[9]:

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

Launch Site Names Begin with 'CCA'

In this query we display the 5 records that begin with ‘CCA’



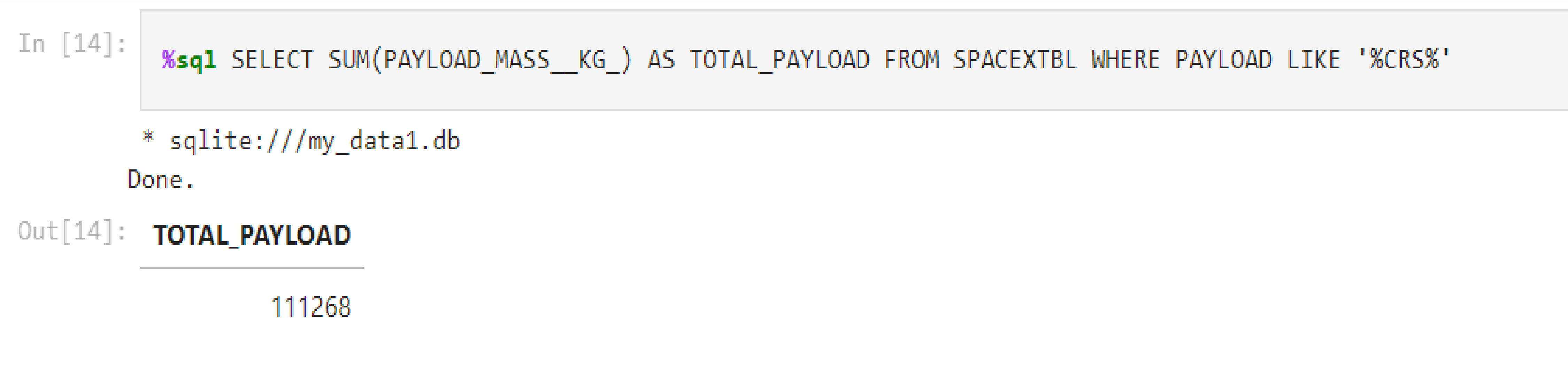
```
In [12]:  %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5

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

Out[12]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

In this query we calculated the total payload mass carried by boosters launched by NASA (CRS).

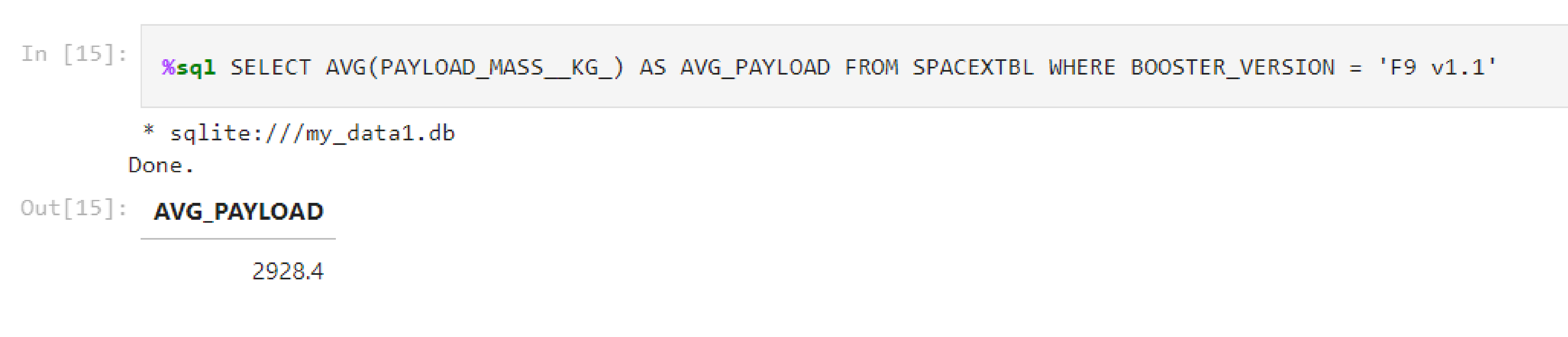


```
In [14]:  %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%'

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

```
Out[14]: TOTAL_PAYLOAD
                111268
```

Average Payload Mass by F9 v1.1

In this query we calculated the Average Payload Mass by F9 v1.1



In [15]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'
```

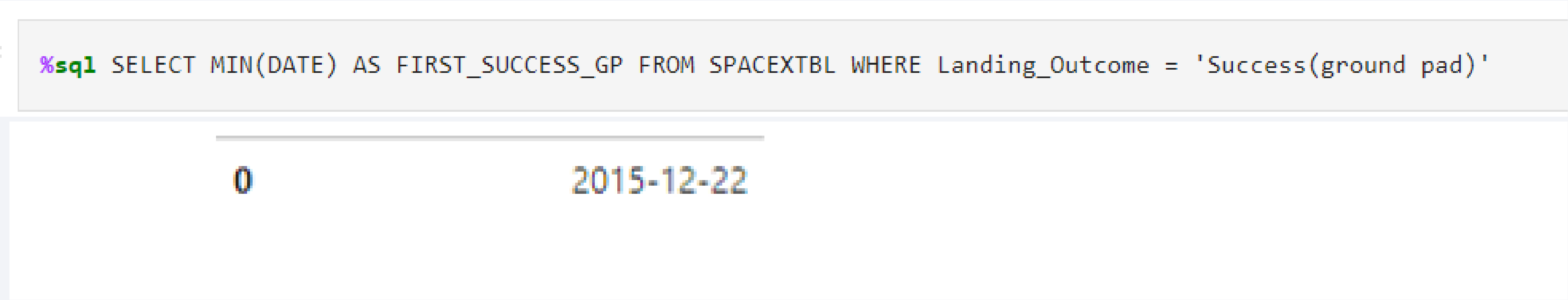
```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- In this query we determined the date when the first successful landing outcome in ground pad was achieved.

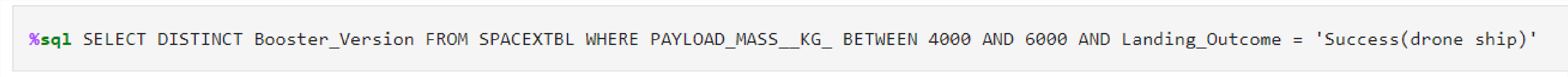


```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE Landing_Outcome = 'Success(ground pad)'
```

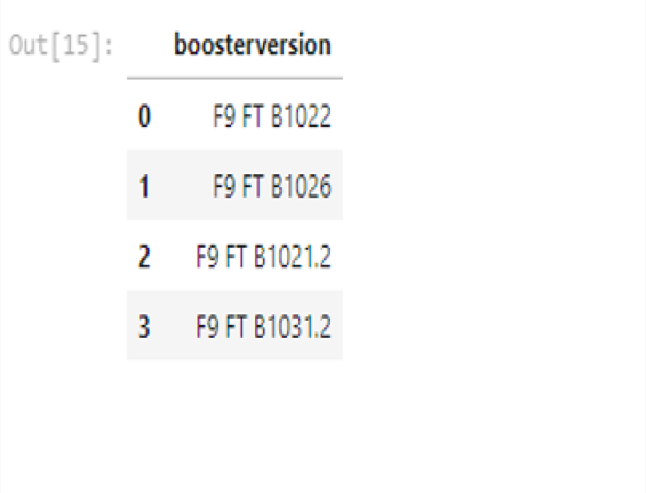
```
0            2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

In this query we 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 DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND Landing_Outcome = 'Success(drone ship)'
```

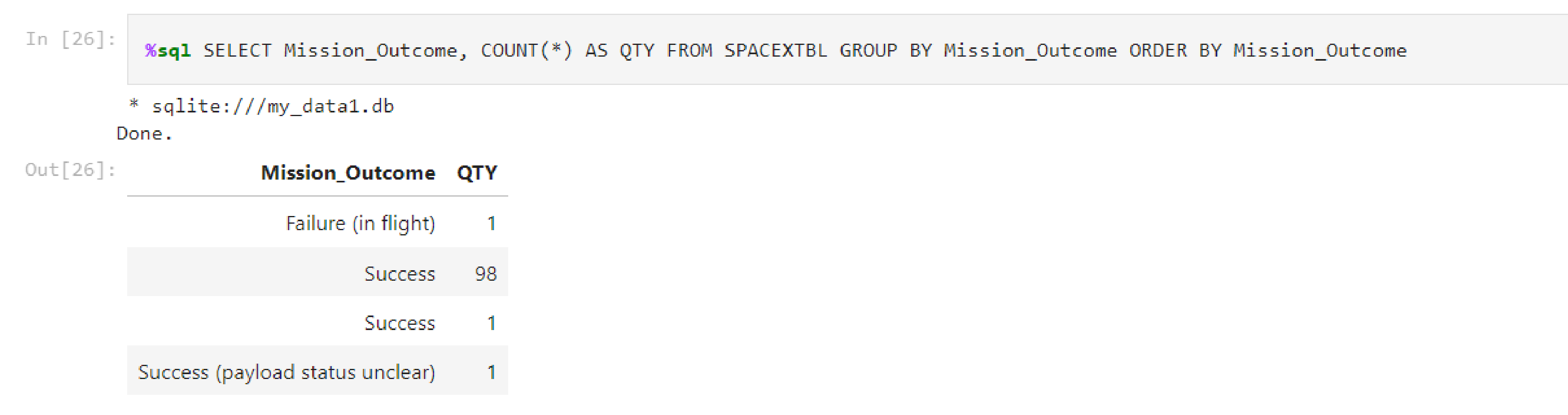


Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

In this query we list the total number of successful and failure mission outcomes.



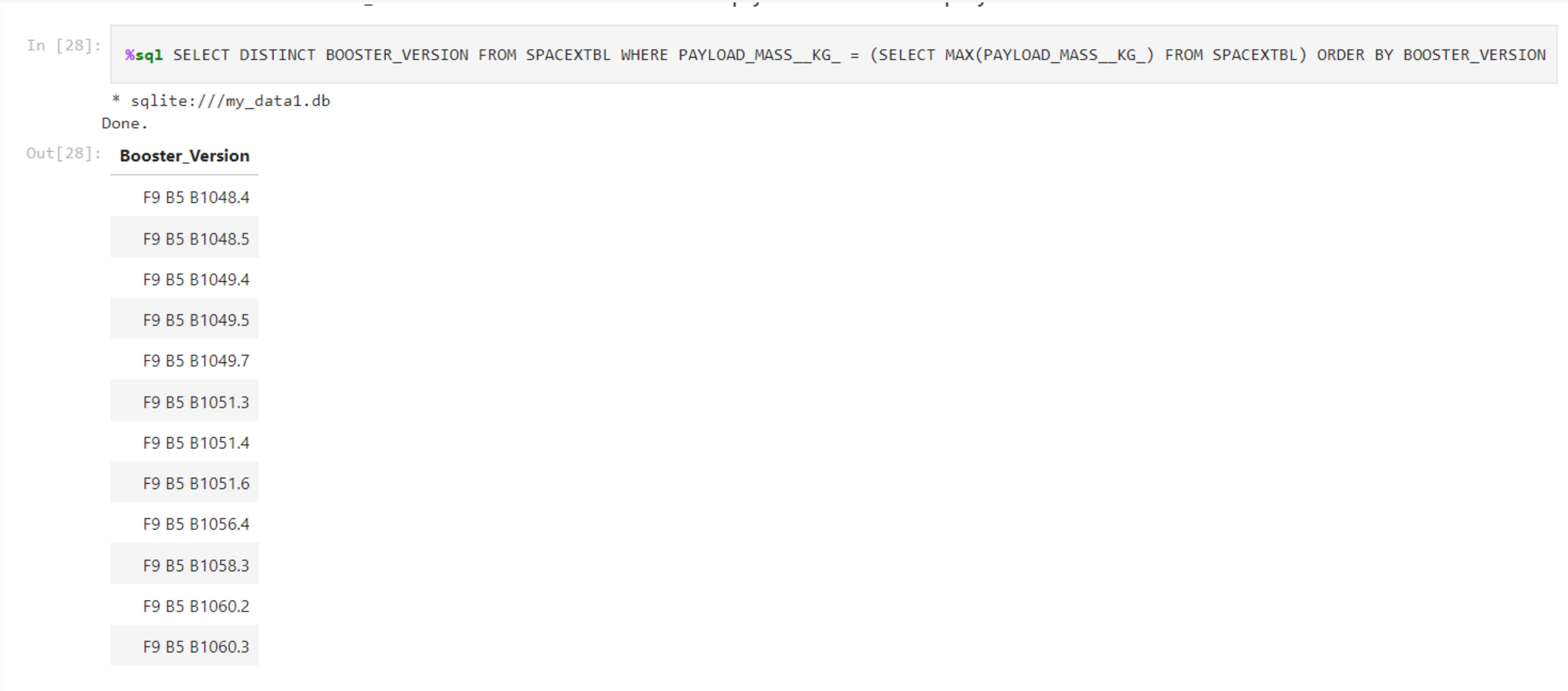
```
In [26]:  %sql SELECT Mission_Outcome, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY Mission_Outcome ORDER BY Mission_Outcome

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

Out[26]:

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

Boosters Carried Maximum Payload

In this query we list the names of the booster versions which have carried the maximum payload mass using a sub query.



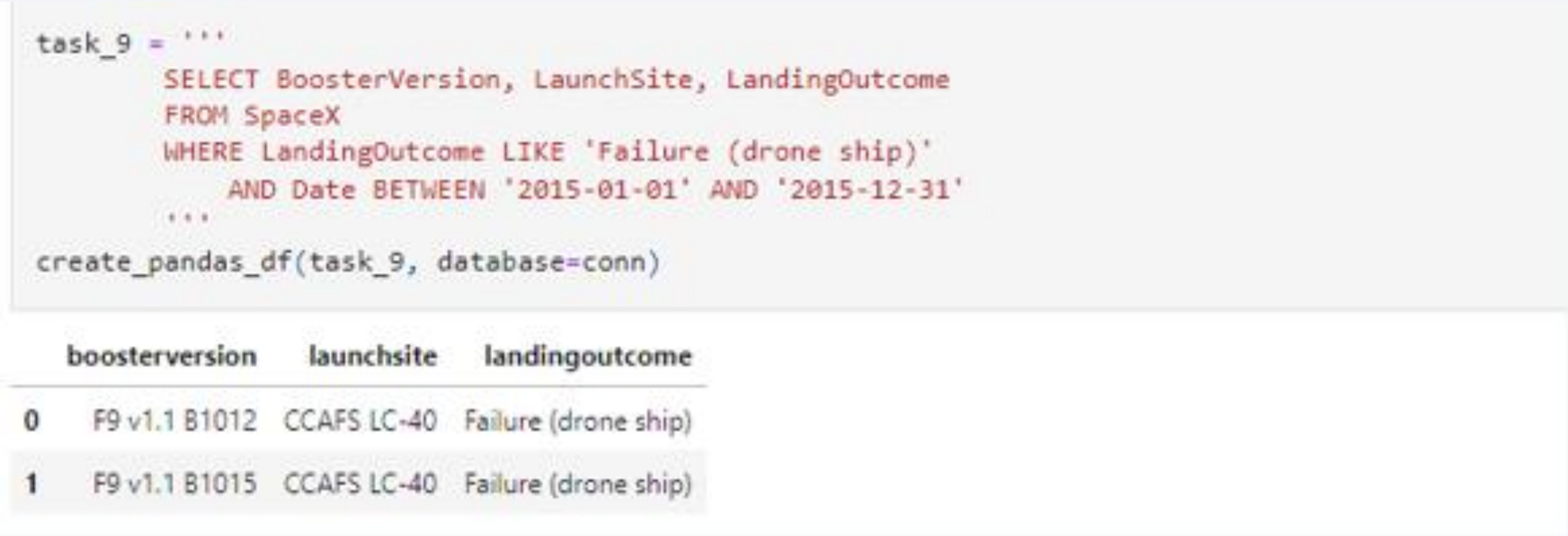
```
In [28]:  %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION

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

Out[28]:

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

2015 Launch Records

In this query we 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.

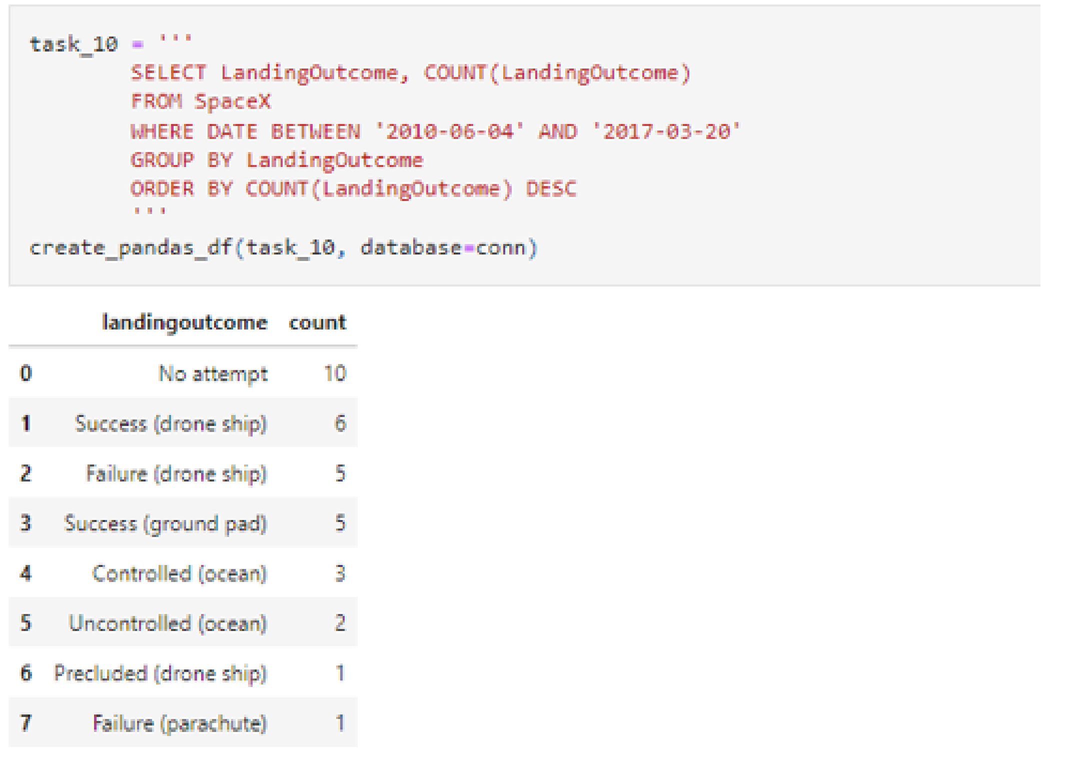


```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

In this query we rank the count of successful Landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.



```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

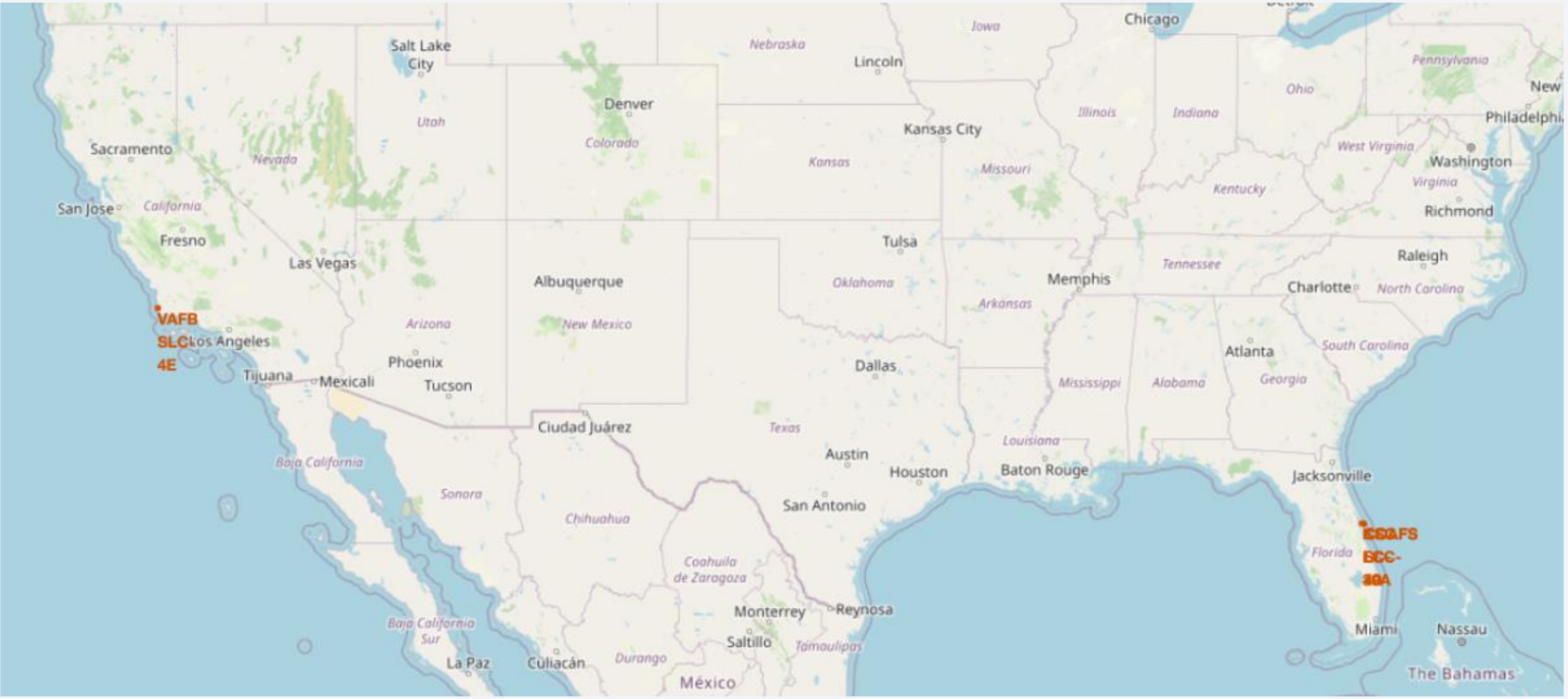
	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

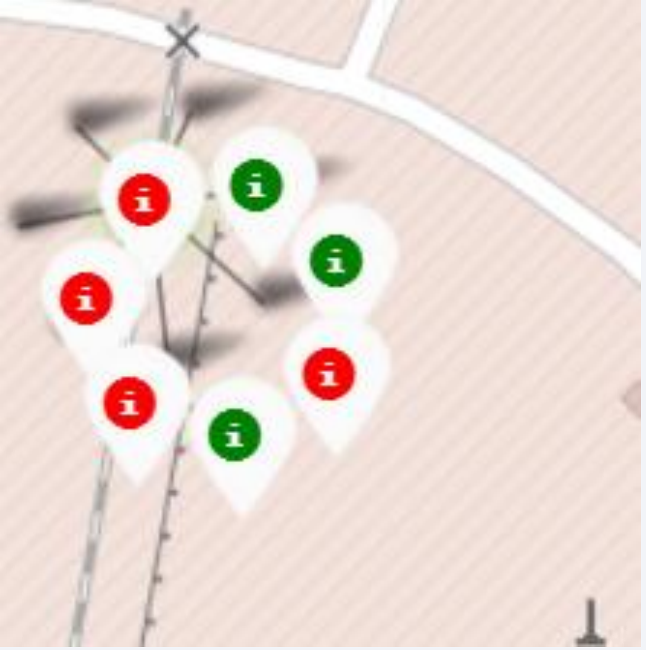
Marked Launch Sites

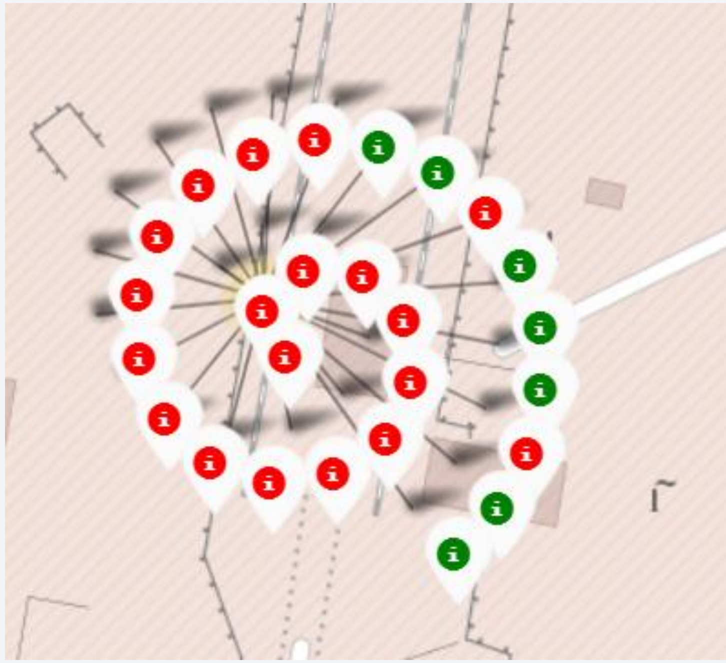
- All launch sites are in very close proximity to the coast

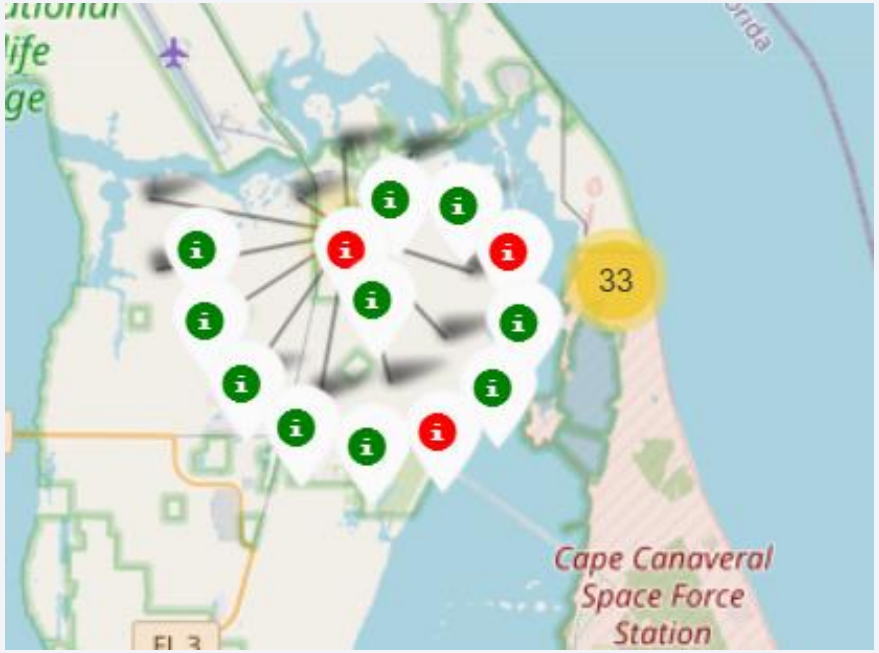


Success/Failed Launches

Green Markers shows successful launches and Red Markers shows Failure

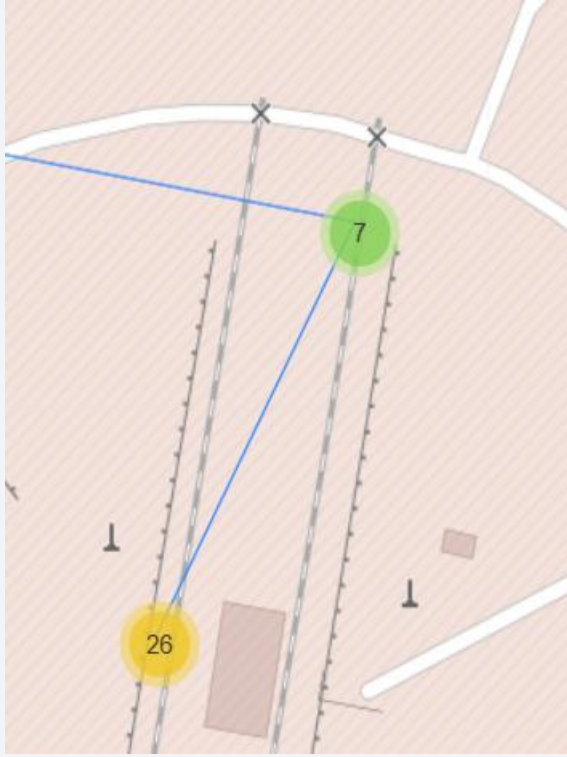


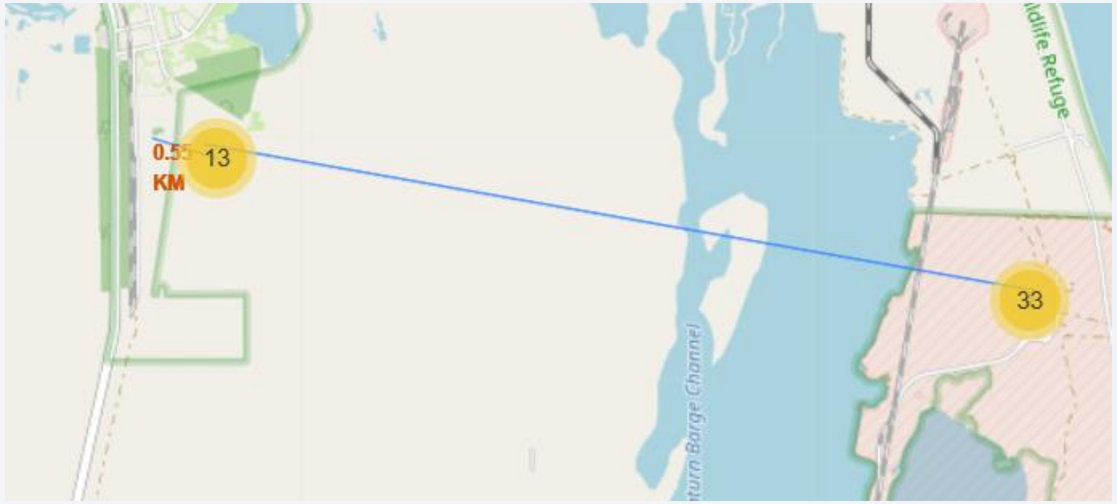


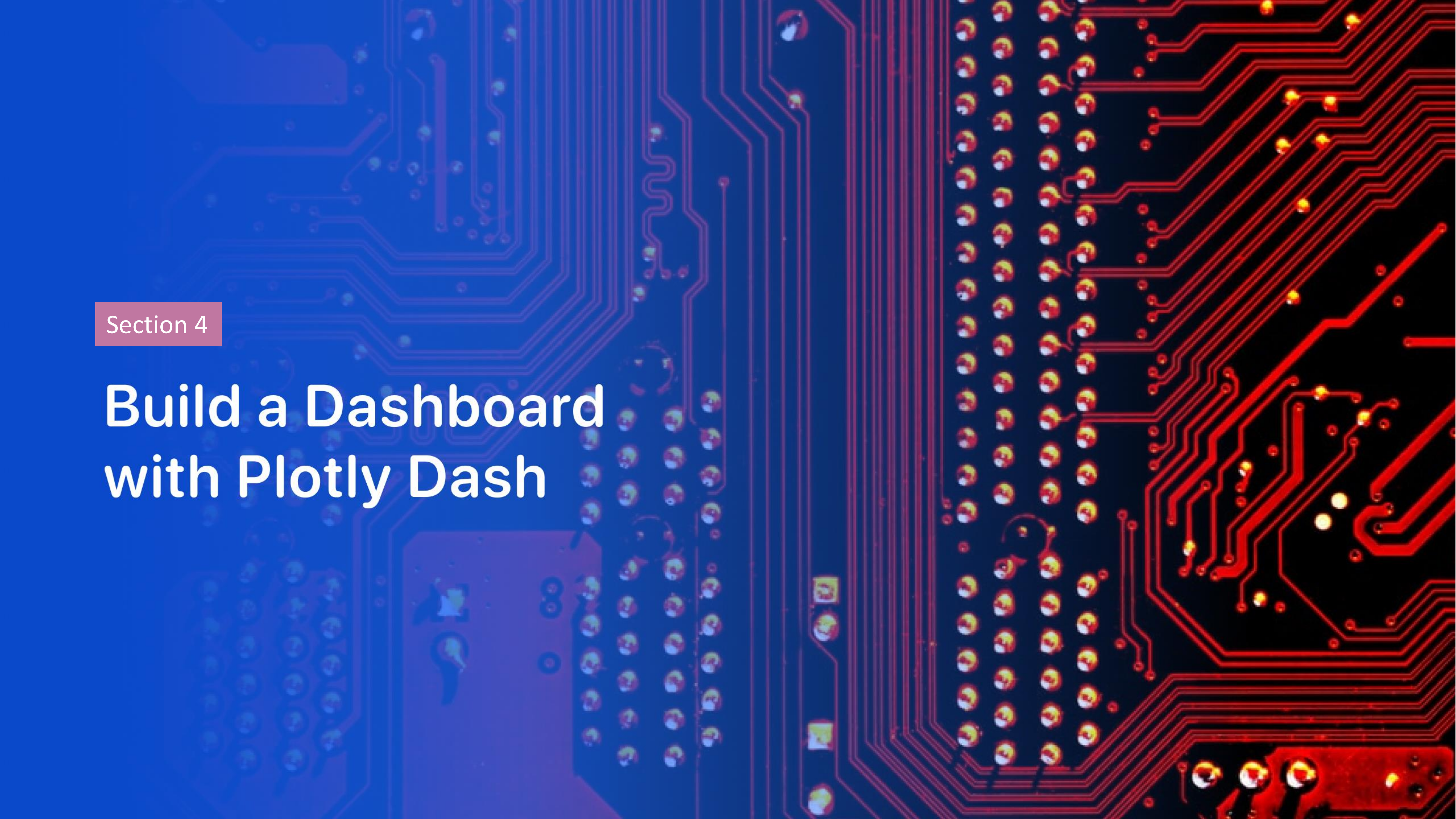


Launch Sites Distance

Map shows the distance line between two launch sites







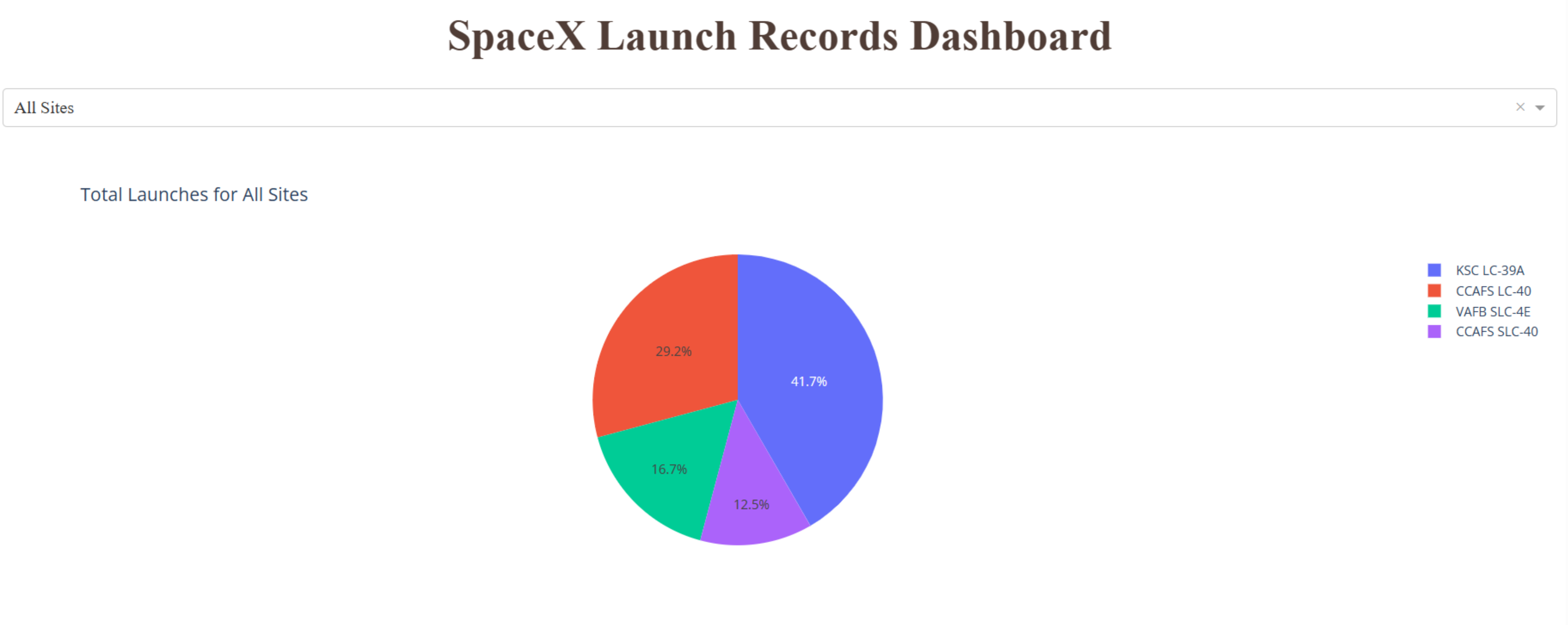
Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

- The pie chart shows the ratio of all sites

- It shows that KSC LC-39A had the most successful launches.



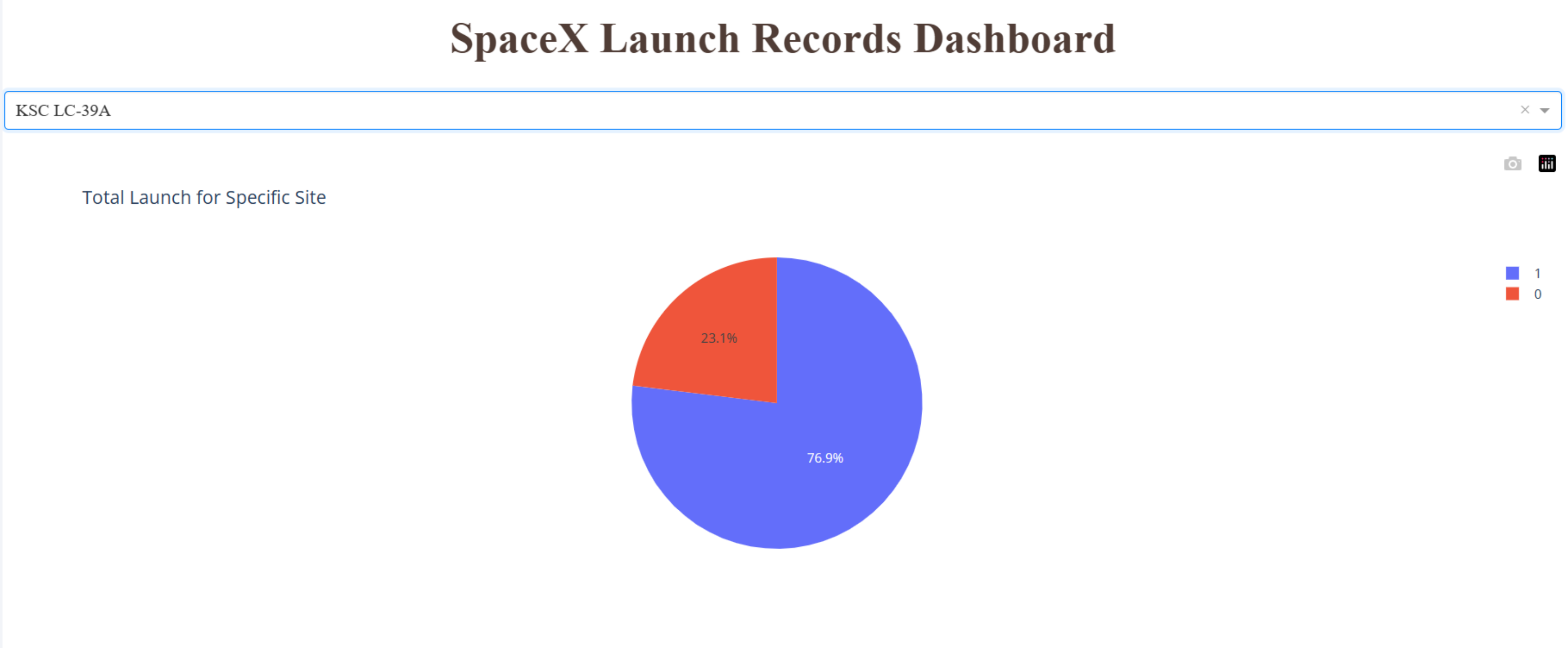
SpaceX Launch Records Dashboard

All Sites

Total Launches for All Sites

Highest Launch Success Ratio

KSC LC-39A had the highest success ratio

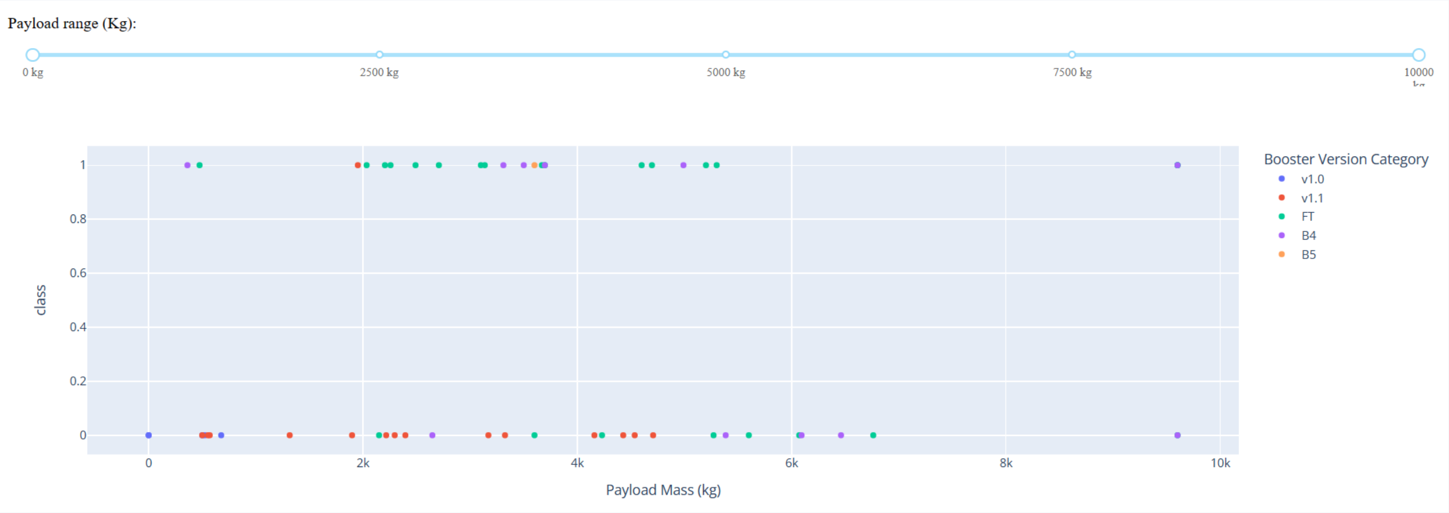


SpaceX Launch Records Dashboard

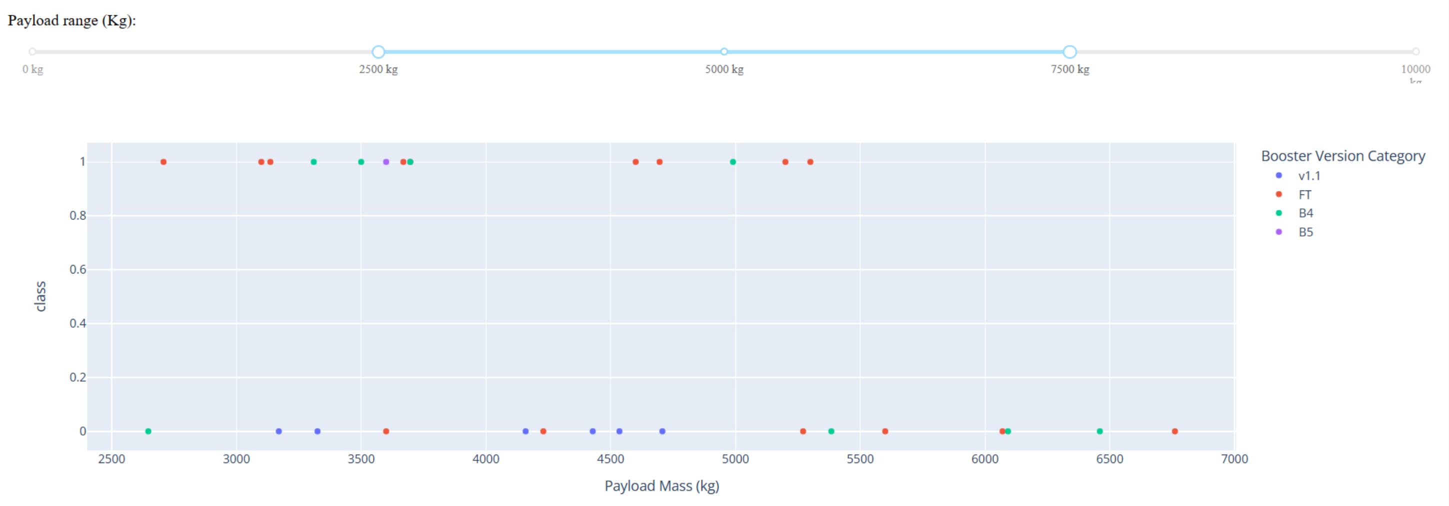
KSC LC-39A

Payload vs. Launch Outcome

First scatter plot shows the payload range of all sites weighted 0kg to 10000 kg



Second scatter plot shows the payload range of all sites weighted 2500kg to 7500 kg

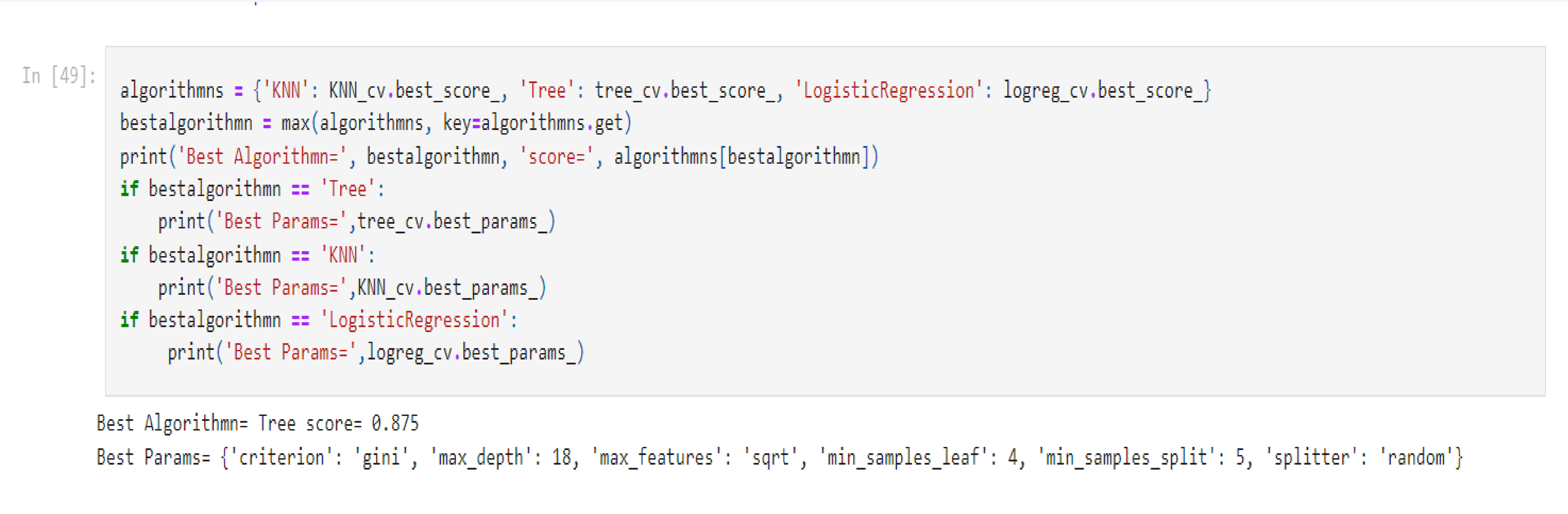


Section 5

Predictive Analysis (Classification)

Classification Accuracy

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

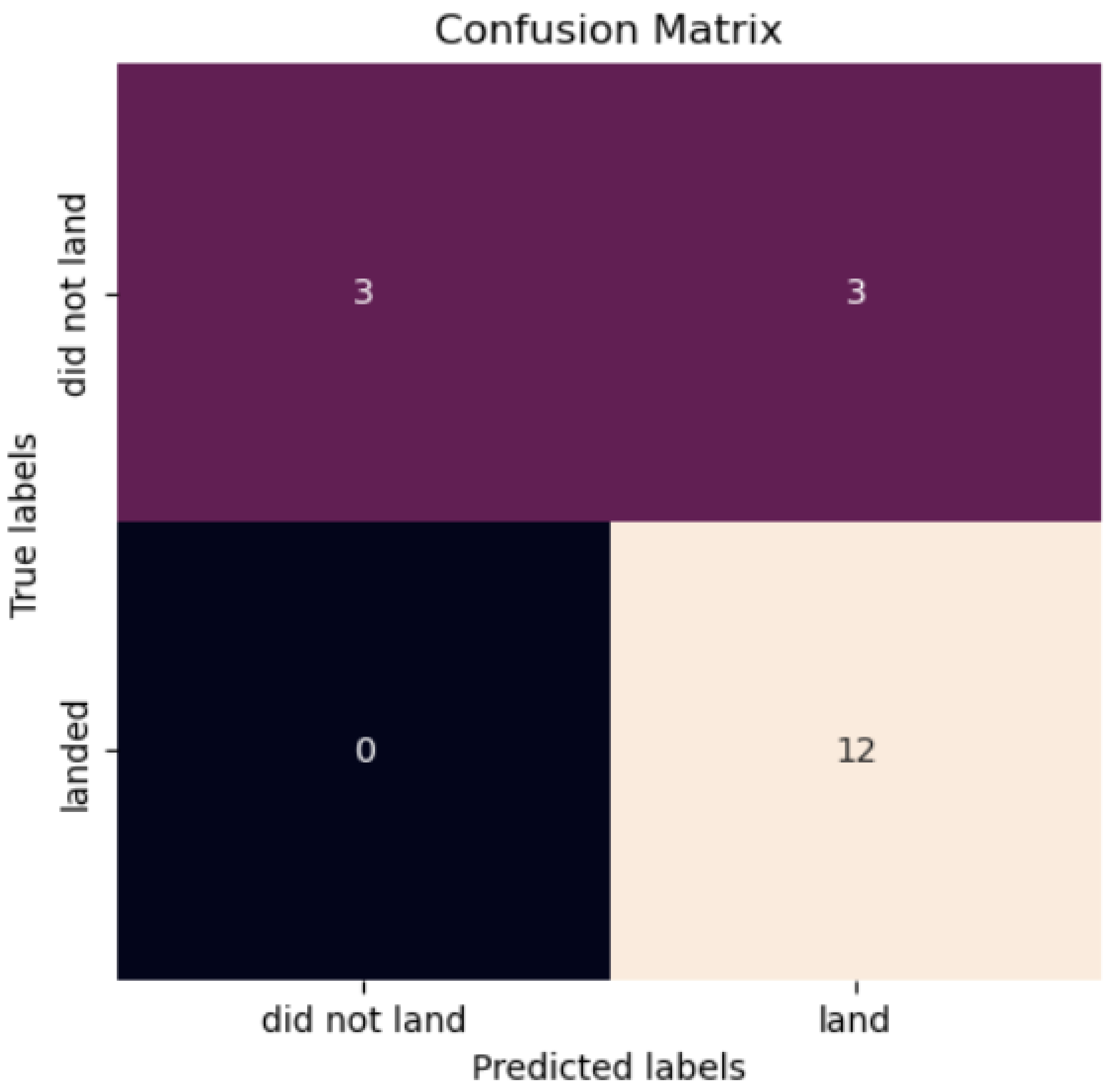


```
algorithmns = {'KNN': KNN_cv.best_score_, 'Tree': tree_cv.best_score_, 'LogisticRegression': logreg_cv.best_score_}
bestalgorithmn = max(algorithmns, key=algorithmns.get)
print('Best Algorithmn=', bestalgorithmn, 'score=', algorithmns[bestalgorithmn])
if bestalgorithmn == 'Tree':
    print('Best Params=',tree_cv.best_params_)
if bestalgorithmn == 'KNN':
    print('Best Params=',KNN_cv.best_params_)
if bestalgorithmn == 'LogisticRegression':
     print('Best Params=',logreg_cv.best_params_)
```

```
Best Algorithmn= Tree score= 0.875
Best Params= {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 4, '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

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at the launch site.
- Launch success rate started to increase since 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO has the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The decision tree classifier is the best machine learning algorithm for this task.

Thank you!