IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Diego Gomez
11-25-2025

Outline

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

Executive Summary

This report presents a comprehensive analysis of the SpaceX first stage's capability to land successfully, including different predicting models the study systematically evaluates the success rate of SpaceX's space missions.

The study focuses on data collection and processing through web scraping and wrangling techniques. Data was prepared for analysis using an interactive dashboard, as well as its use in different predictive models to determine the probability of successful first-stage landings.

The data analysis provided detailed insights into the effectiveness of SpaceX's landing strategies. The predictive models demonstrated high accuracy in identifying key factors influencing successful landings, in addition, significant patterns in the data were identified that allow for more confident predictions of future mission success.

Introduction

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 we can determine if the first stage will land, we can determine the cost of a launch.

First, the correlation of variables is important to understand how one can change other, which is the most important variable, which is the least important variable. If we can see everything in a dashboard, what insights this graphics could show us. How well can a determine predicting model predict the outcome of a rocket and which of these models is the best for this specific study.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology(how data was collected)
- Perform data wrangling(how data was processed)
- 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)

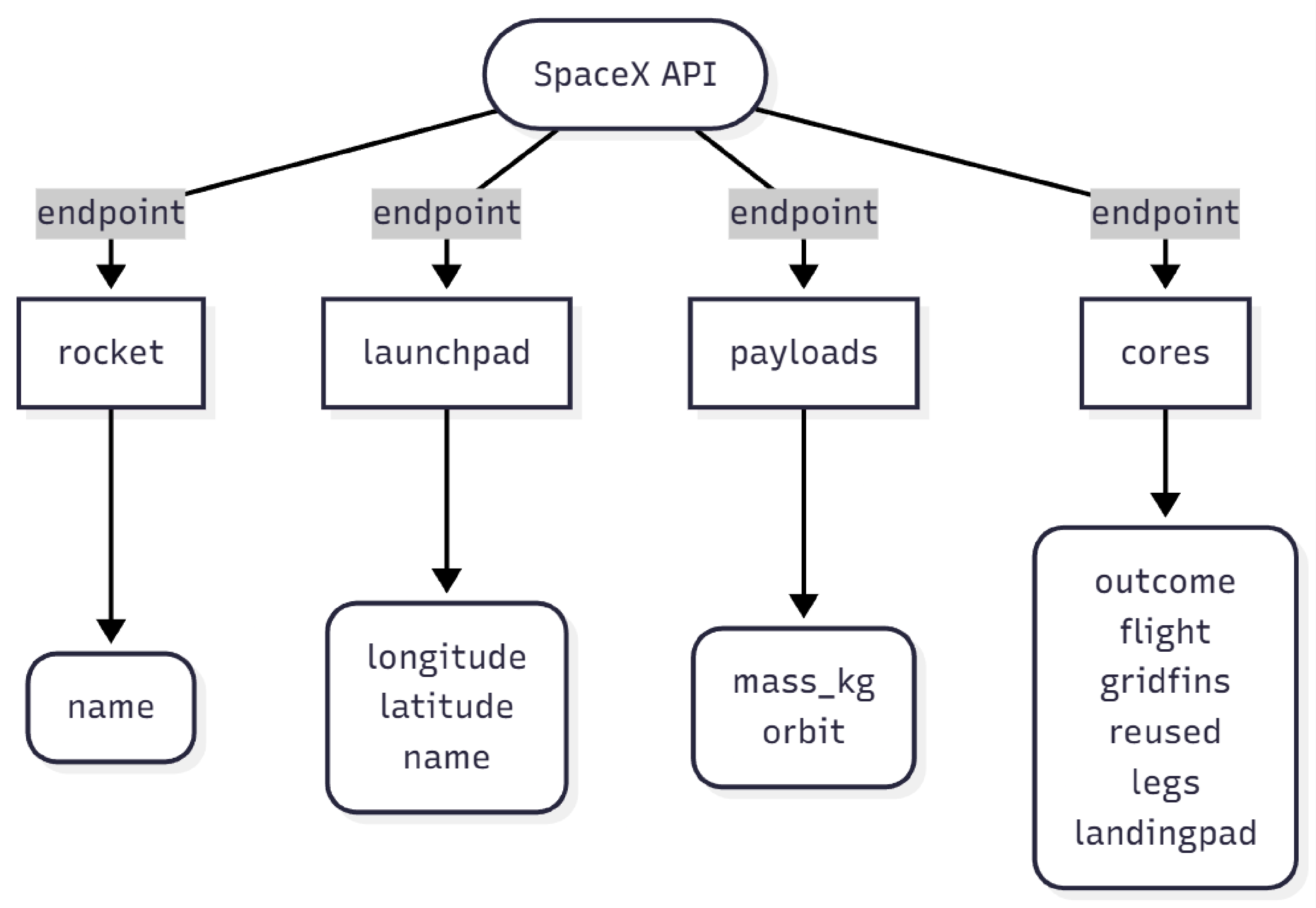
Data Collection

First the use of the API was important to obtain all the data use in the study, so getting the JSON was the primary step to obtain the structured data. Then the JSON was normalized and using to obtain important features such as the type of rocket, launchpad and cores.

Finally, from these features we obtain the most useful such as booster version, payload mass, orbit, launch site, and our target feature outcome.

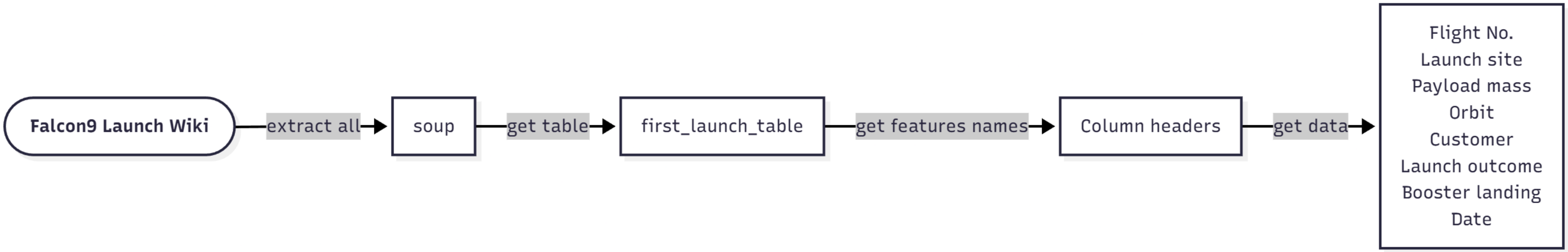
For the second part, we use Wikipedia to recollect the launch records, parsing the HTML table and creating a data frame of the required columns of the table.

Data Collection – SpaceX API



link here

Data Collection - Scraping



Falcon9 Launch Wiki

extract all

soup

get table

first_launch_table

get features names

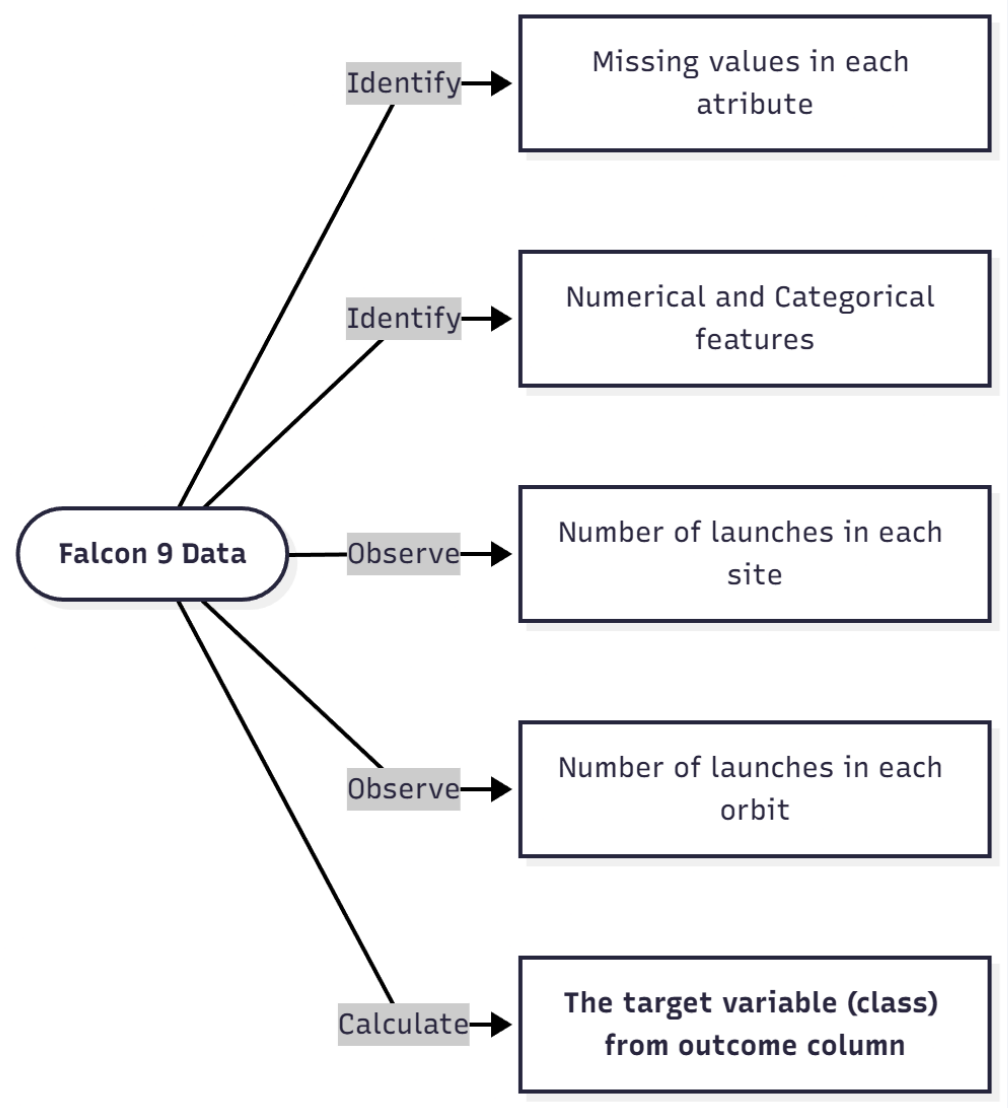
Column headers

get data

Flight No.
Launch site
Payload mass
Orbit
Customer
Launch outcome
Booster landing
Date

link here

Data Wrangling


Falcon 9 Data
Identify
Missing values in each atribute
Identify
Numerical and Categorical features
Observe
Number of launches in each site
Observe
Number of launches in each orbit
Calculate
The target variable (class) from outcome column

link here

EDA with Data Visualization

Understand the relationship between the variables and the target (class) is essential before starting creating models, we must know how Y change when we change X_1,X_2, X_3, etc.

Another important step is to have the same weight in all the variables; the last model we want to have is a model that incorrectly predicts the target variable due to incorrect weights in the data, so one-hot encoding the categorical variables and casting the numeric ones is another relevant part of EDA.

link here

EDA with SQL

The SQL queries used in this part were:

- The distinct launch sites.
- The first 5 records where launch sites begin with “CCA”.
- The average payload mass carried by booster version F9 v1.1.
- The list of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- The total number of successful and failure mission outcomes
- More…

link here

Build an Interactive Map with Folium

- The first map is to locate an initial center such as NASA Johnson Space Center at Houston, Texas.
- Then we must locate all the launch sites, filtering the data to obtain unique launch sites and passing their coordinates to our map.
- The next step is mapping all the success/failed launches for each site (green for success and red for failed)
- The last step is calculating the distance between a launch site and different position in the map, such a railway, highway, coastline, etc.

link here

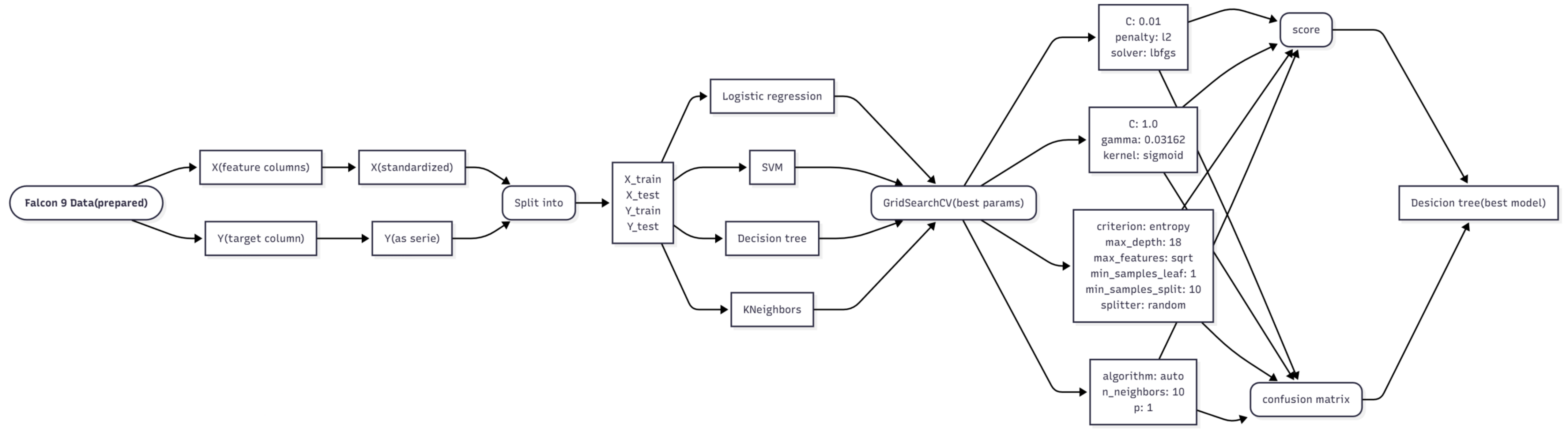
Build a Dashboard with Plotly Dash

To have a useful dashboard is essential to create not a static one but also an interactive one that is user-friendly and correctly shows the graphics in an easy way to understand:

- The first part has the selection of launch site where the user can select if show the graphics by all the sites or just one, then it appears a payload range where the user can change the minimum and maximum of the payload mass.
- The second part is the two graphics(pie chart and scatter plot) that shows the total success launches and the correlation between payload and success for and between the two previous selections the user chose.

link here

Predictive Analysis (Classification)



link here

Results

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

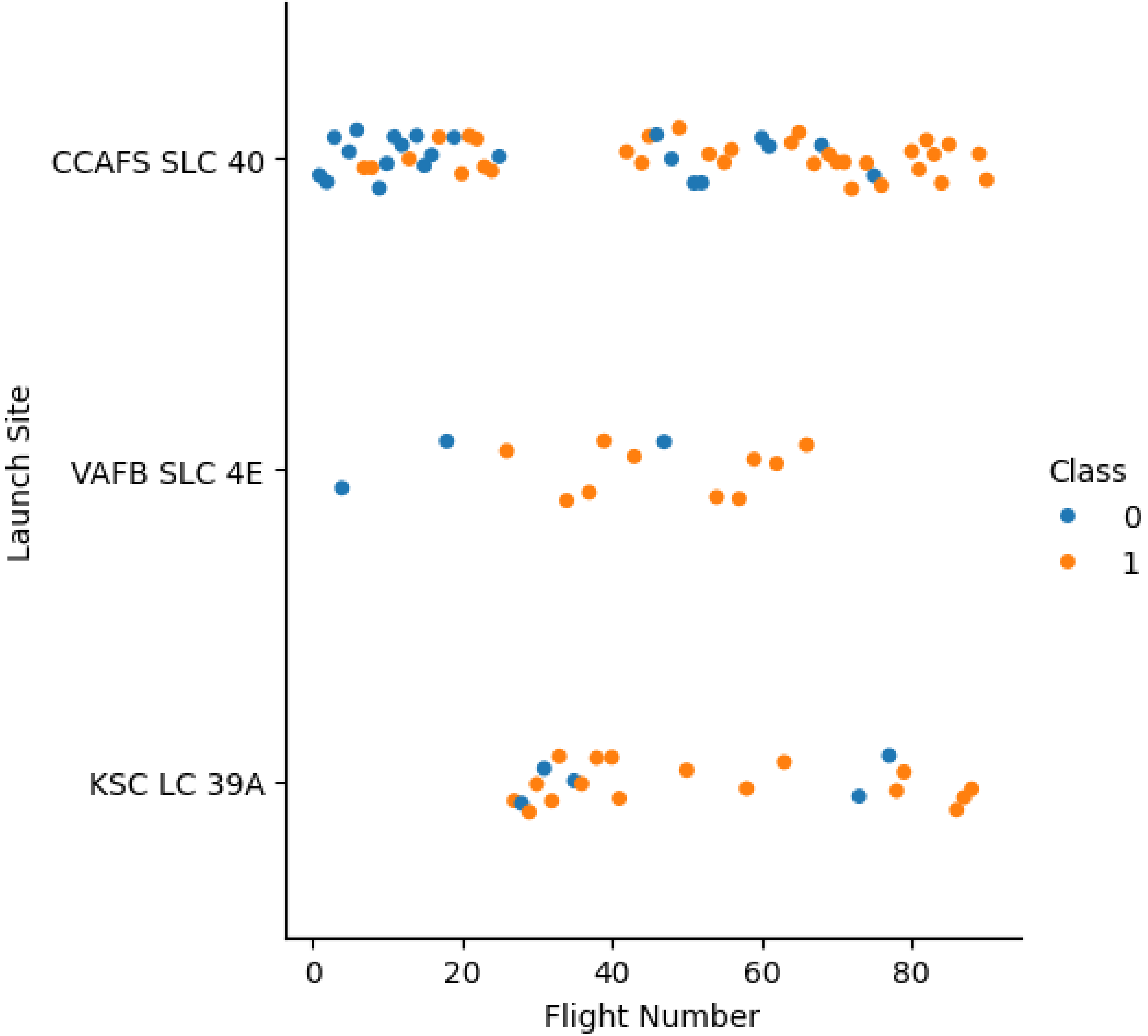
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- In this chart we can see that in an elevate number of flights the success rate increase, in this case for the three launch sites.

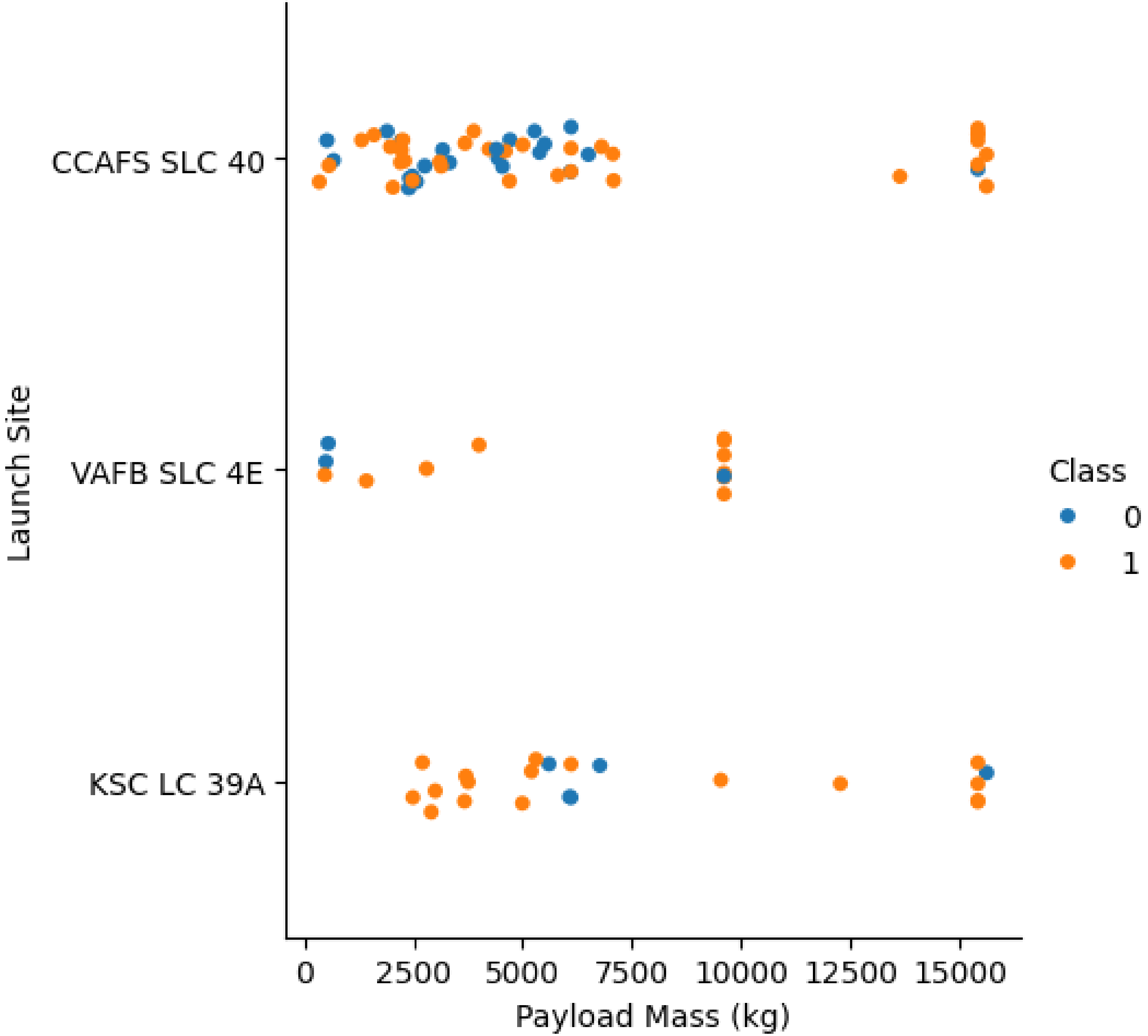
- Additionally, there are many more flights for CCAFS SLC 40 site than the other sites.



Payload vs. Launch Site

- In this case it looks like the same case, when the payload mass increases, the success rate also do it.

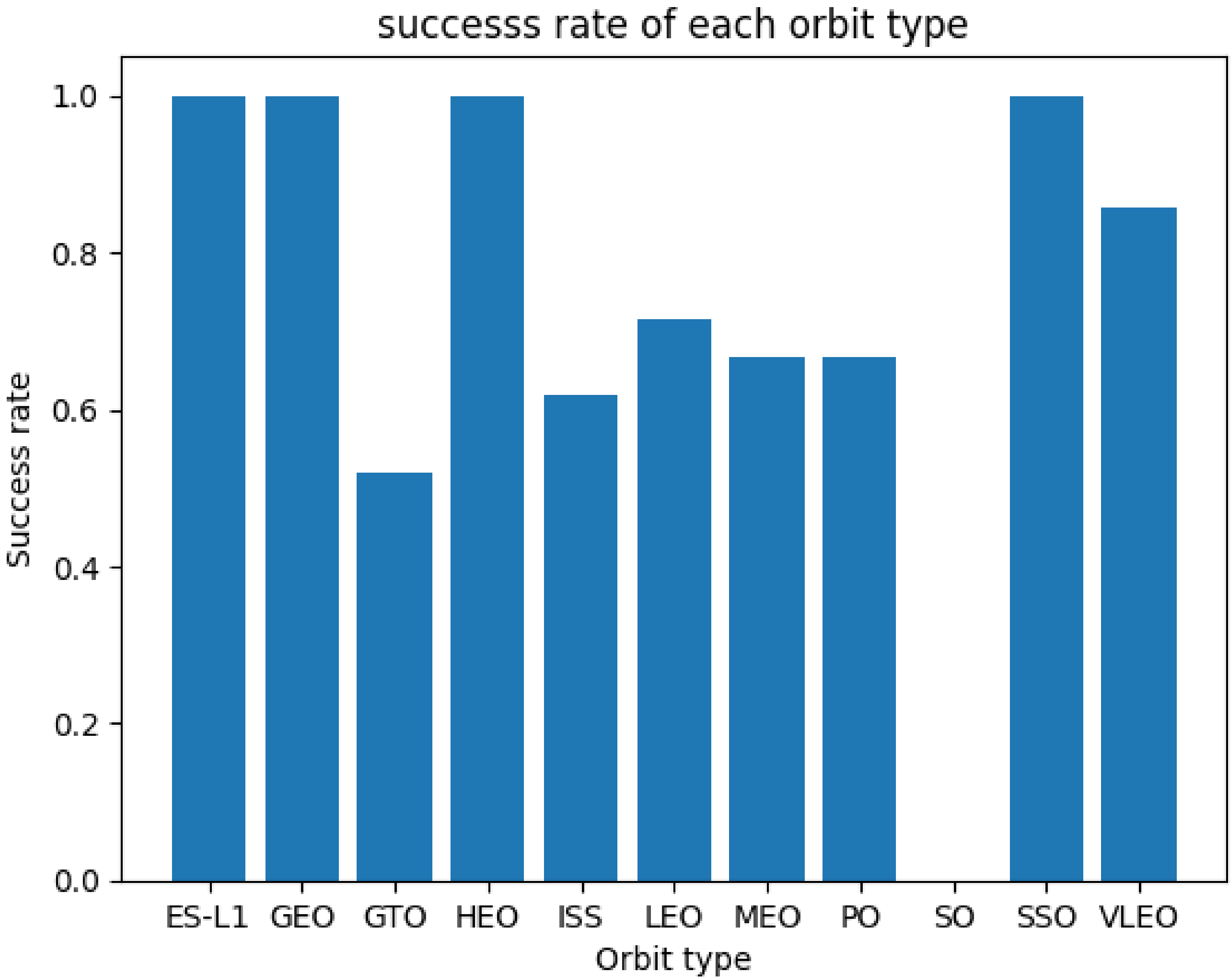
- For VAFB SLC 4E there are no rockets launched for heavy payloads(greater than 10000).



Success Rate vs. Orbit Type

- We can observe that the best orbits with the highest success rate is ES-L1, GEO, HEO, SSO.

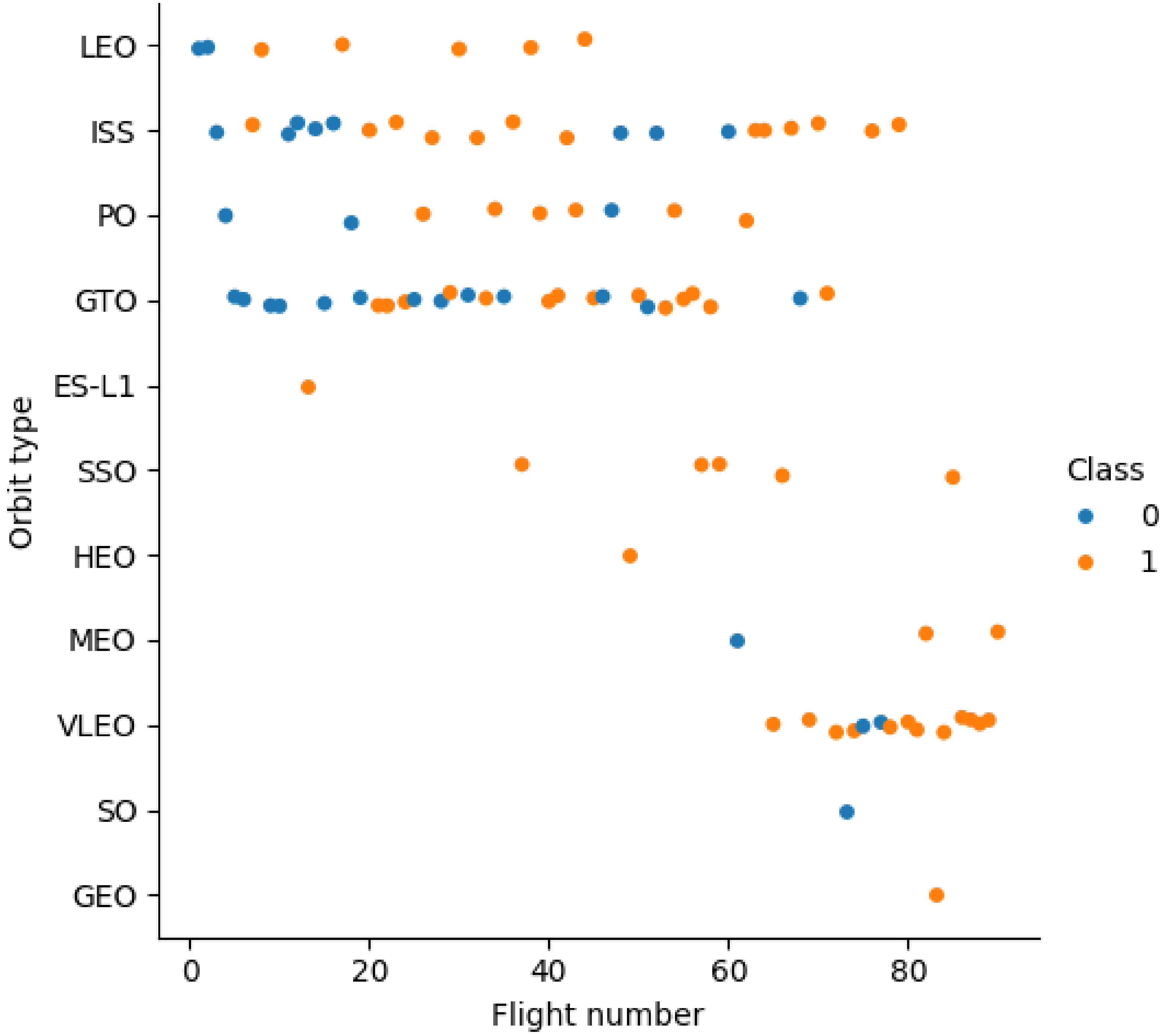
- On the other hand, the worst orbits with the lowest success rate are GTO and ISS.



Flight Number vs. Orbit Type

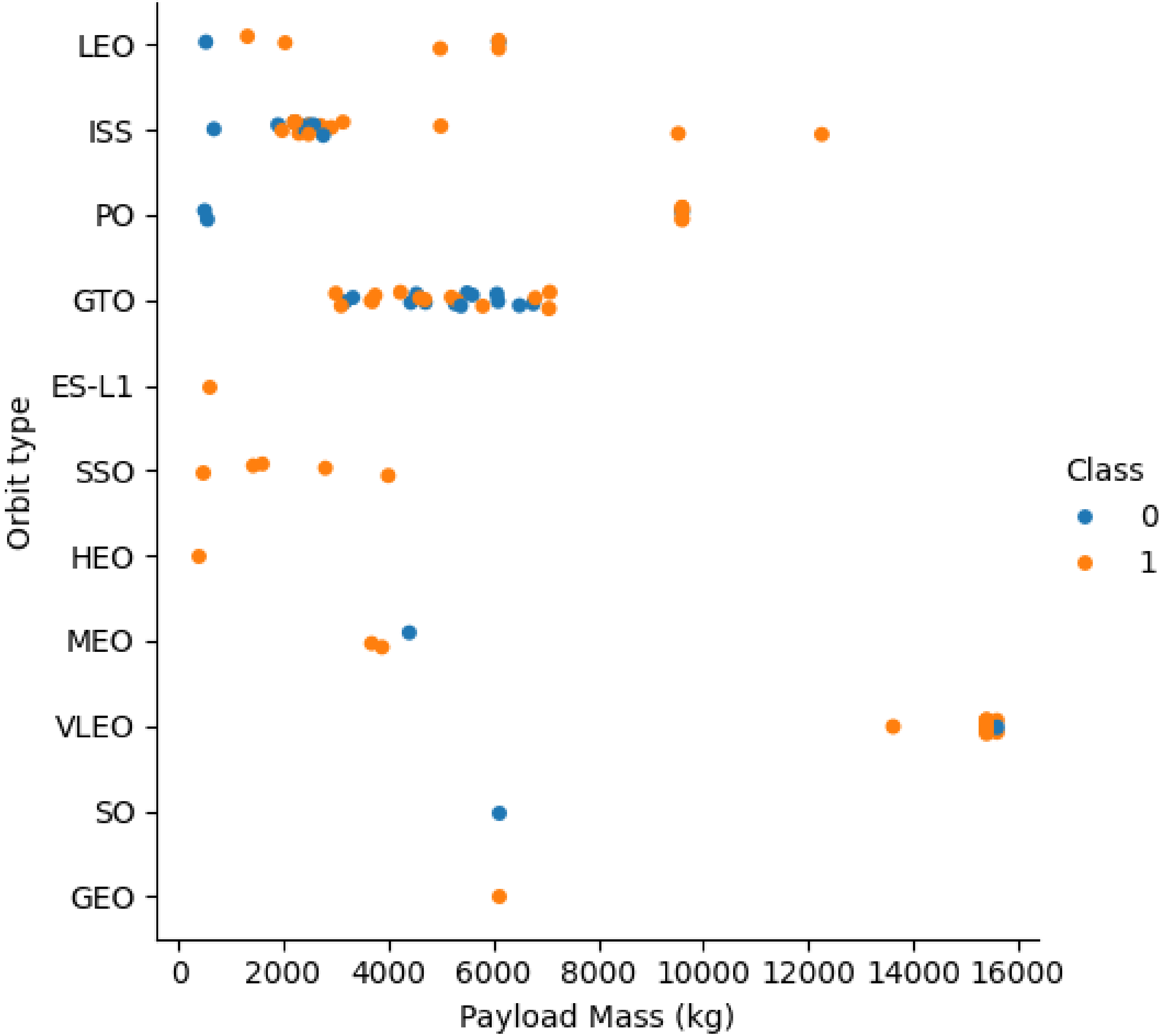
- It seems that some orbits have no failed rate, for example ES-L1, SSO, HEO and GEO.

- In the case of LEO, we can say that the number of flights affect the success rate, but this is not an applicable for all the orbit types as we can see in GTO for example.



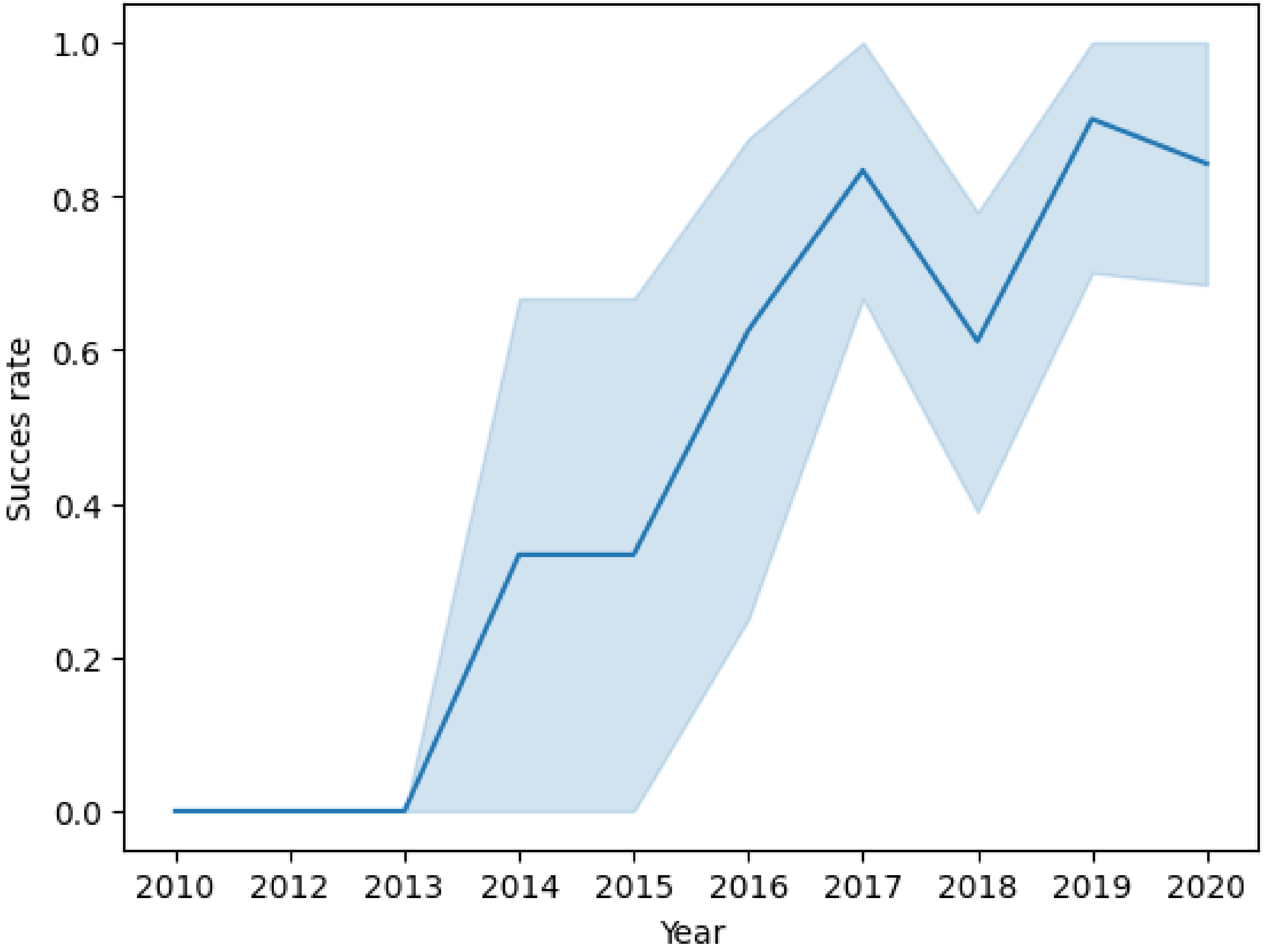
Payload vs. Orbit Type

- In the case of LEO, ISS and PO is eident that with a heavy payload, the success rate increases, but this is not the rule for the rest of the orbits as we can see in GTO, MEO or maybe VLEO.



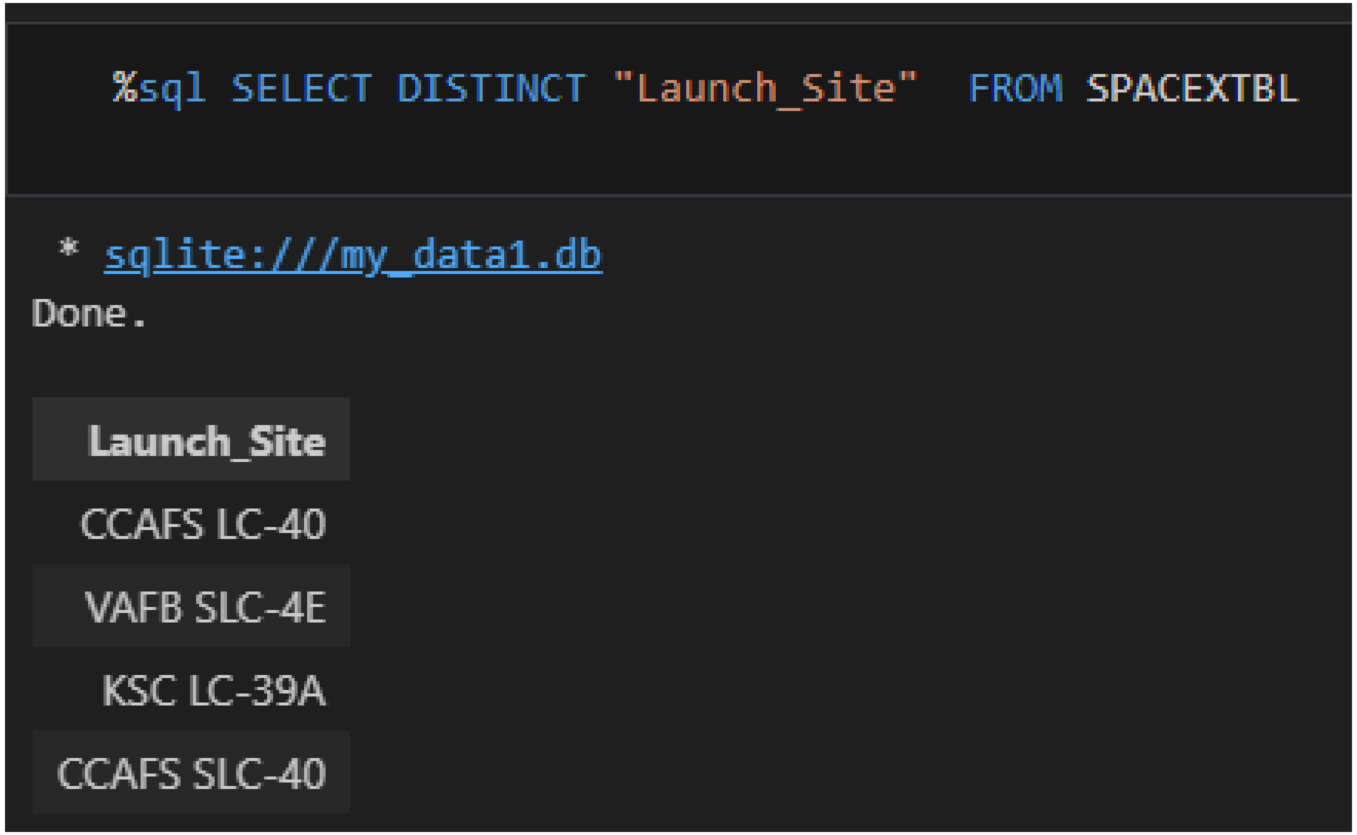
Launch Success Yearly Trend

- Since 2013 we can observe that the success rate kept increasing until 2020.



All Launch Site Names

This is the query and the response for the distinct launch sites.



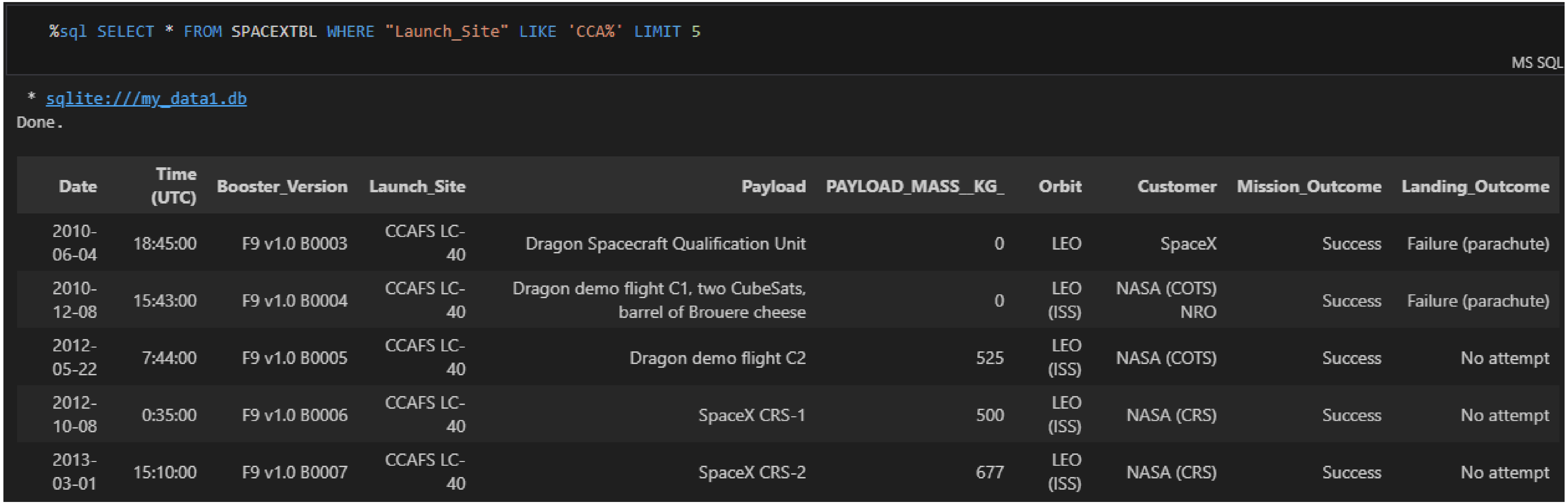
```
%sql SELECT DISTINCT "Launch_Site"  FROM SPACEXTBL

 * 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'

Query of the first five launch sites with “CCA” at the beginning of the name.



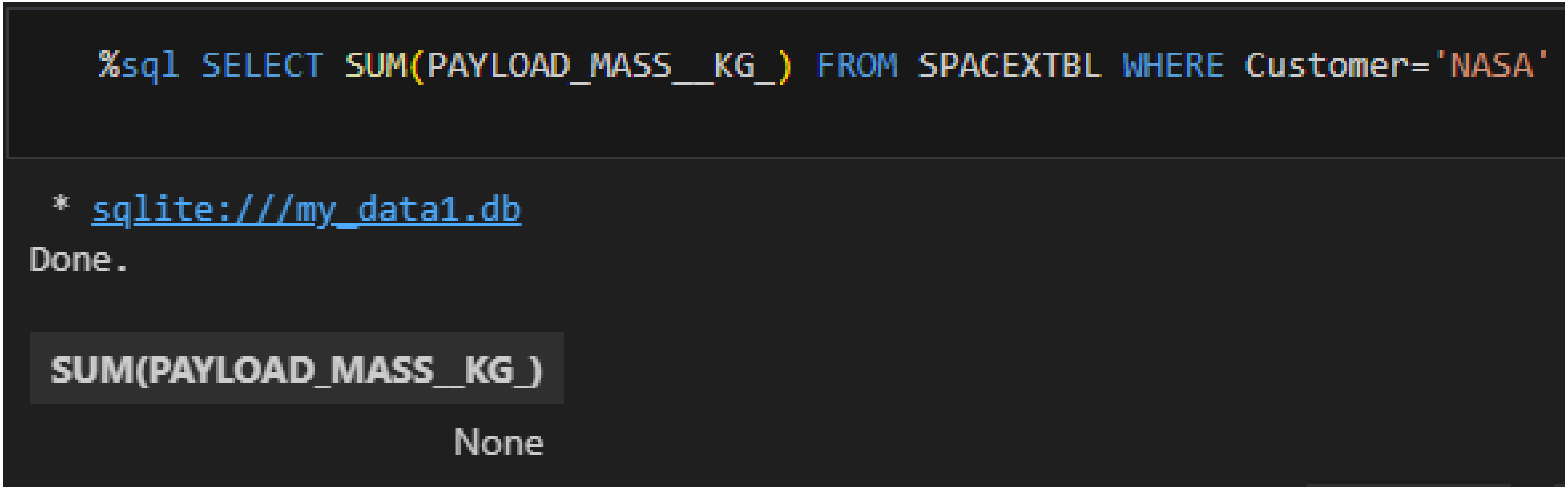
```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
```

```
 * 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

In this query the total payload mass from the customer “NASA” is zero.



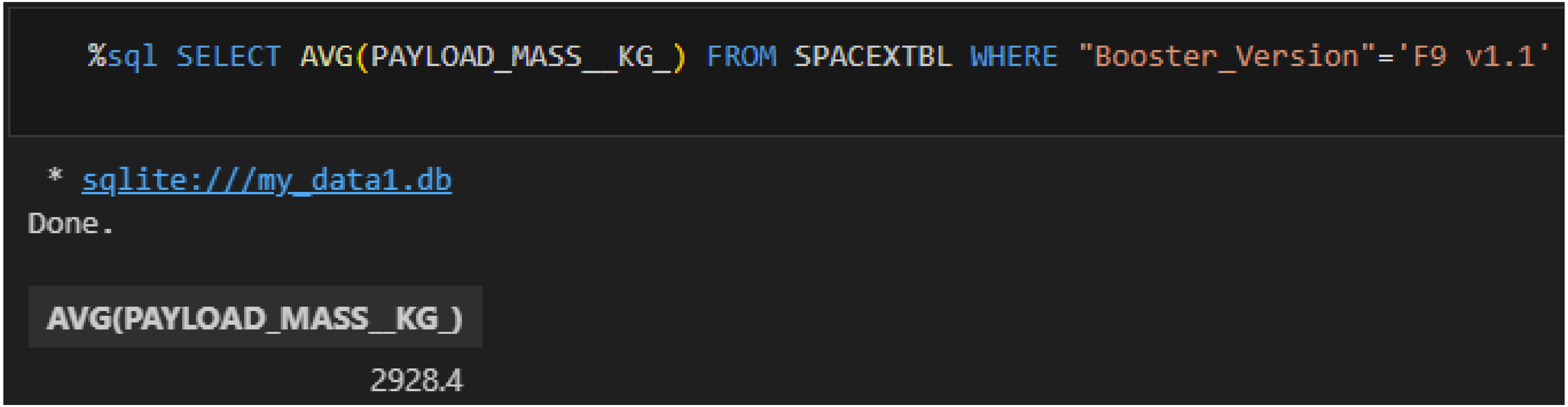
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer='NASA'

 * sqlite:///my_data1.db
Done.

SUM(PAYLOAD_MASS__KG_)
                  None
```

Average Payload Mass by F9 v1.1

This query show us the average payload mass carried by booster version F9 v1.1.



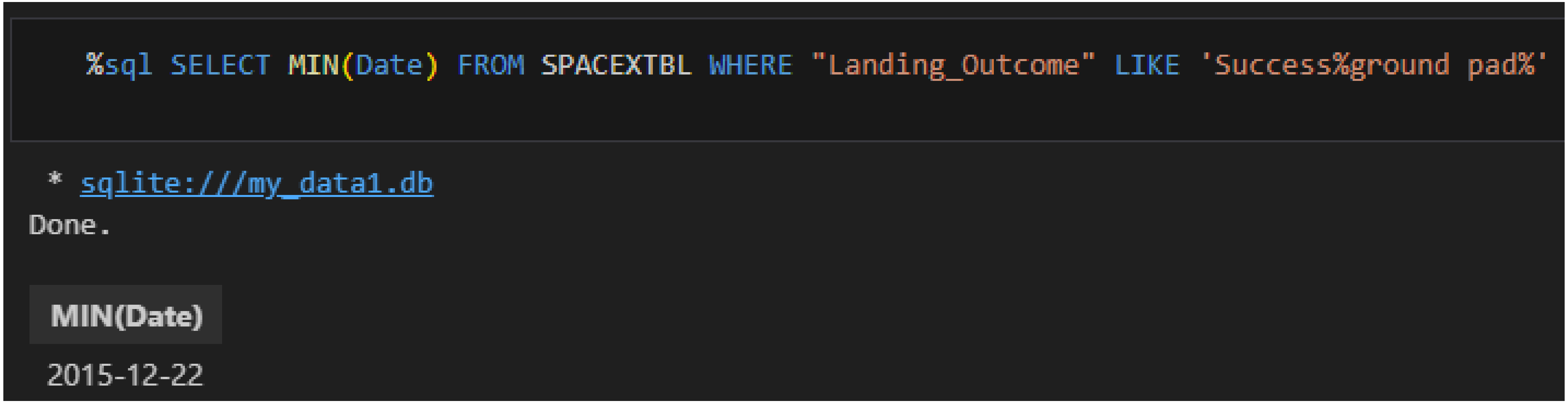
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version"='F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
                2928.4
```

First Successful Ground Landing Date

This query show us the first successful landing outcome in ground pad.



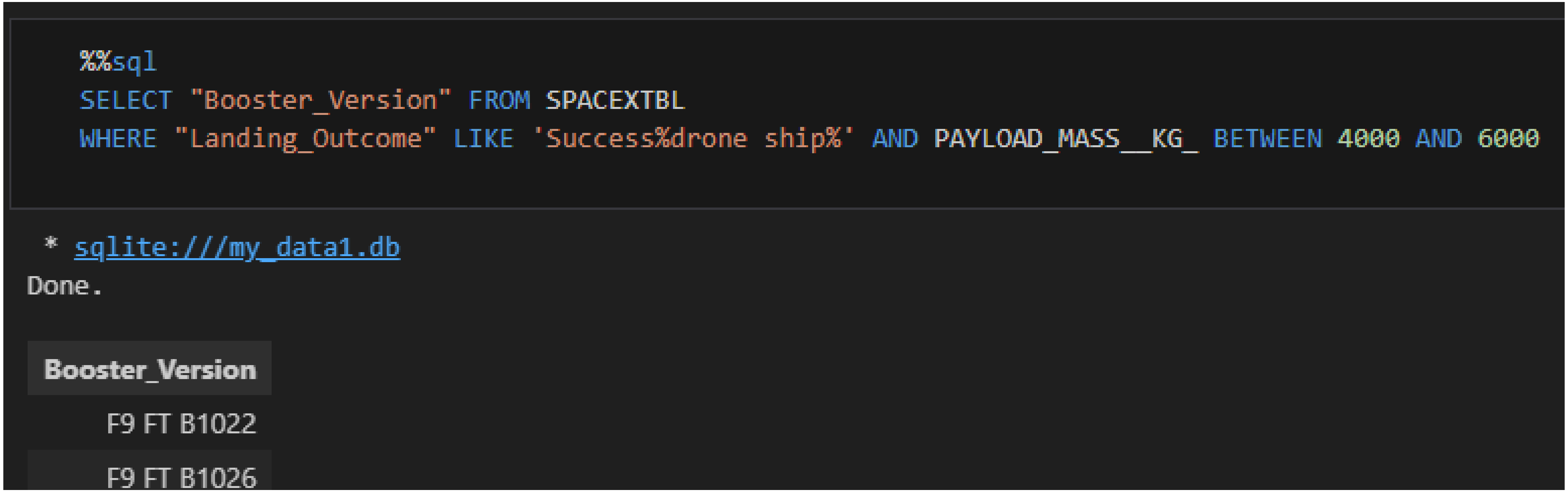
```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE "Landing_Outcome" LIKE 'Success%ground pad%'
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

In this case the query demonstrate the list of the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.



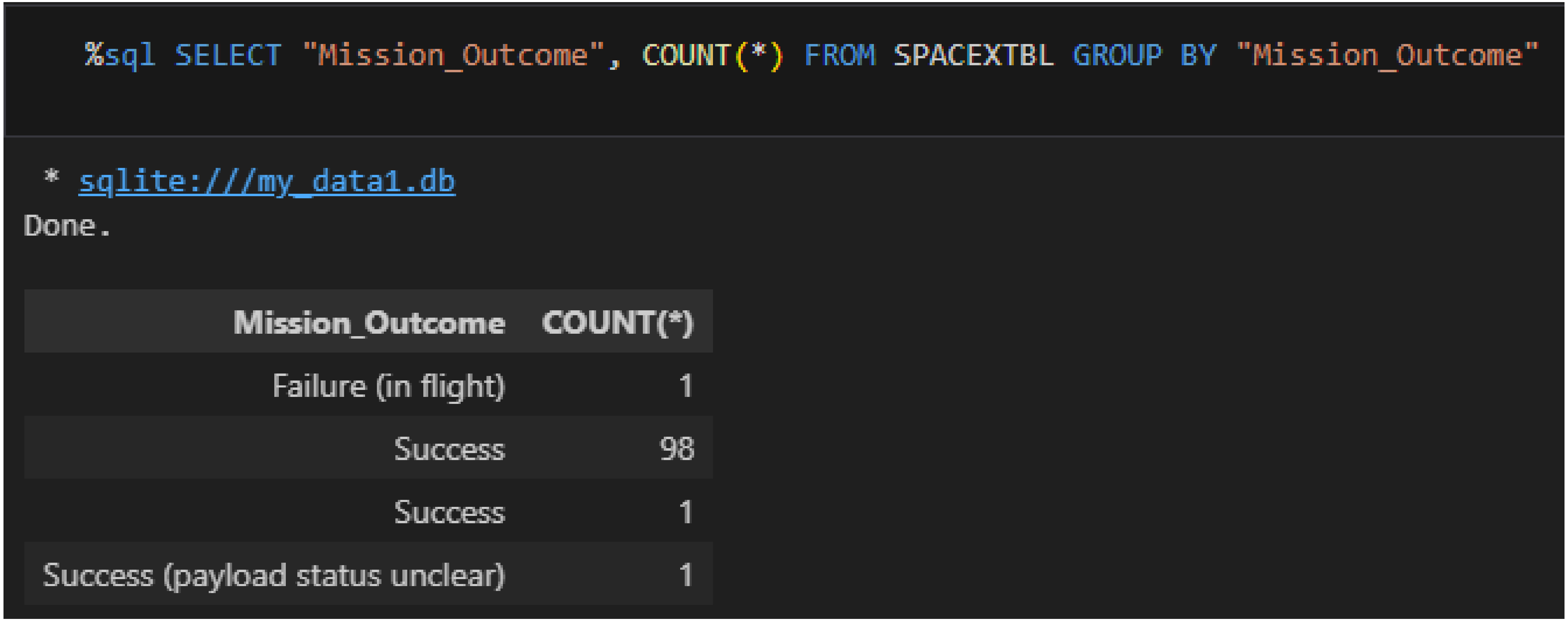
```
%%sql
SELECT "Booster_Version" FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE 'Success%drone ship%' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

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

Booster_Version
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

The query show us a list of the total number of successful and failure mission outcomes.



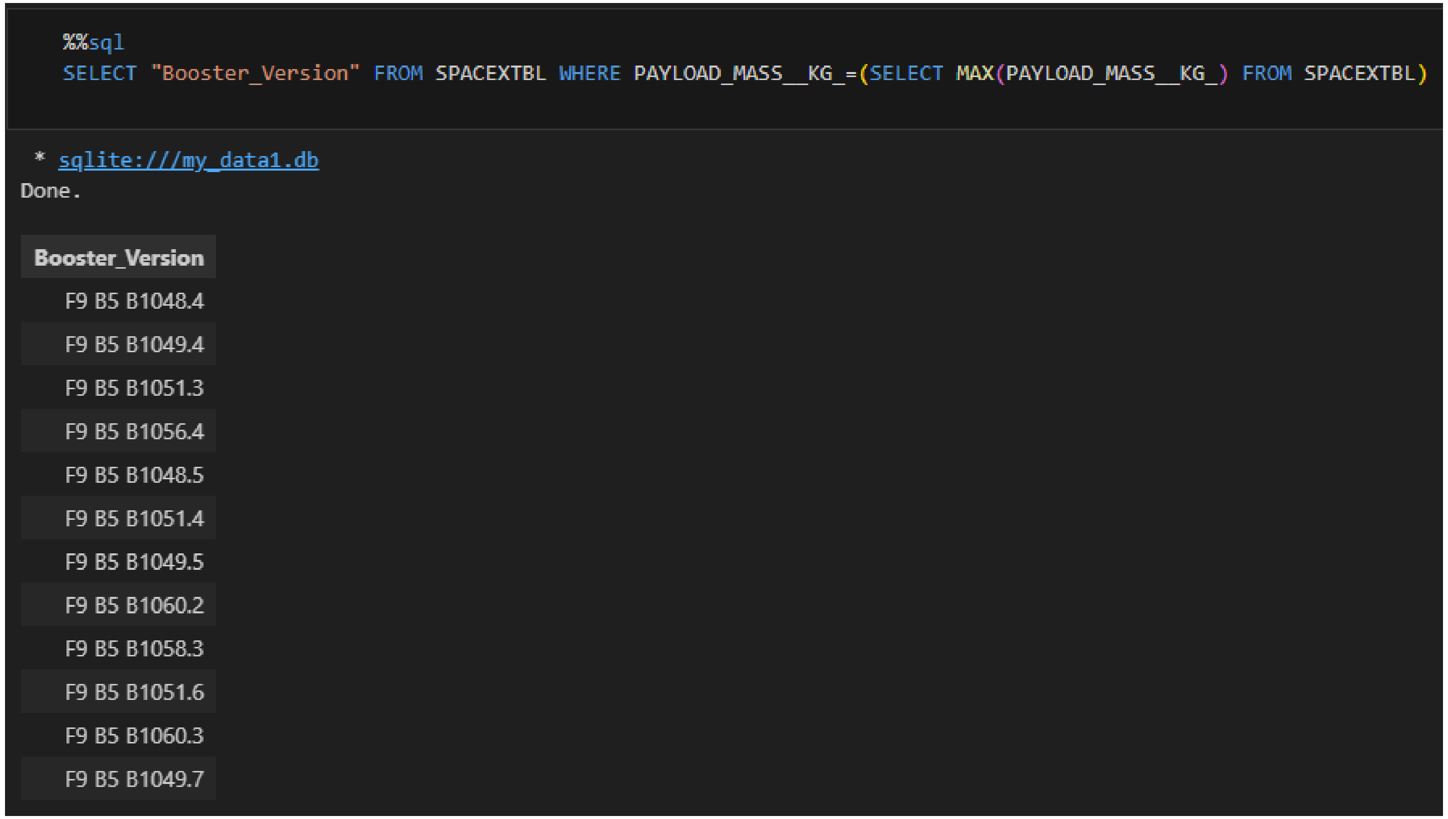
```
%sql SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTBL GROUP BY "Mission_Outcome"
```

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

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

Boosters Carried Maximum Payload

The query show us a list of all the booster versions that have carried the maximum payload mass.



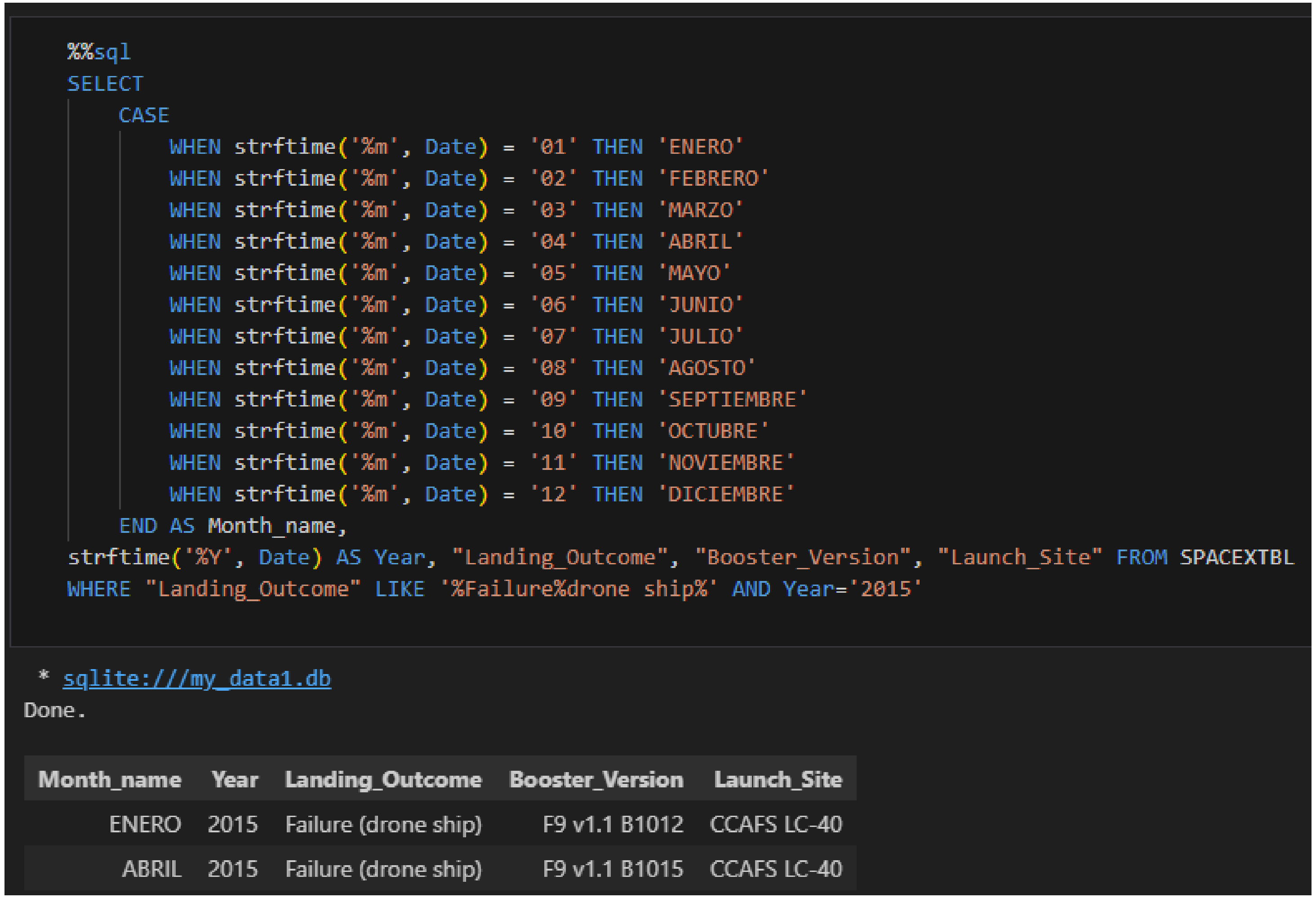
```
%%sql
SELECT "Booster_Version" FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

- This is a query that show us a list of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015(with their month names).



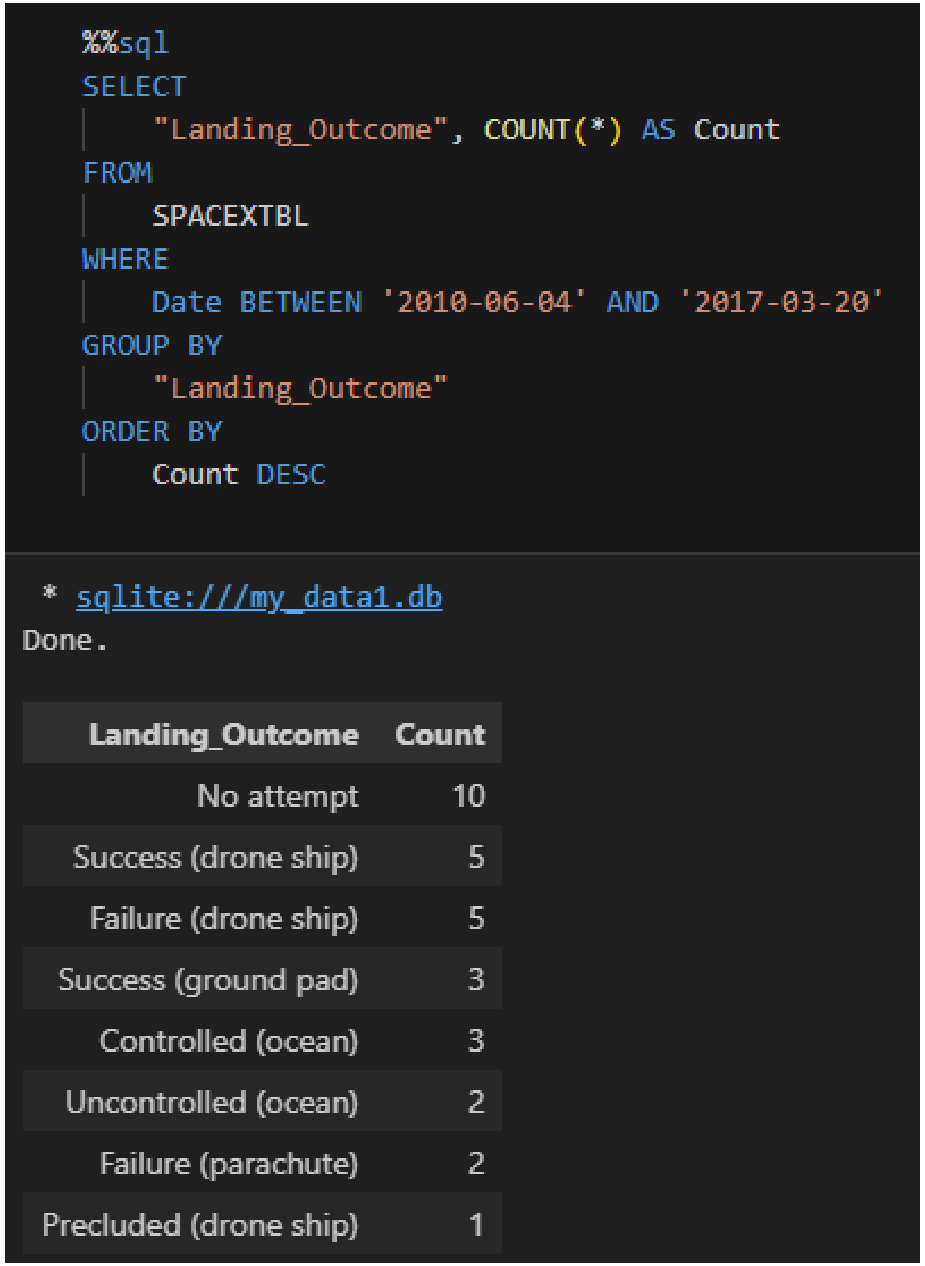
```
%%sql
SELECT
    CASE
        WHEN strftime('%m', Date) = '01' THEN 'ENERO'
        WHEN strftime('%m', Date) = '02' THEN 'FEBRERO'
        WHEN strftime('%m', Date) = '03' THEN 'MARZO'
        WHEN strftime('%m', Date) = '04' THEN 'ABRIL'
        WHEN strftime('%m', Date) = '05' THEN 'MAYO'
        WHEN strftime('%m', Date) = '06' THEN 'JUNIO'
        WHEN strftime('%m', Date) = '07' THEN 'JULIO'
        WHEN strftime('%m', Date) = '08' THEN 'AGOSTO'
        WHEN strftime('%m', Date) = '09' THEN 'SEPTIEMBRE'
        WHEN strftime('%m', Date) = '10' THEN 'OCTUBRE'
        WHEN strftime('%m', Date) = '11' THEN 'NOVIEMBRE'
        WHEN strftime('%m', Date) = '12' THEN 'DICIEMBRE'
    END AS Month_name,
strftime('%Y', Date) AS Year, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL
WHERE "Landing_Outcome" LIKE '%Failure%drone ship%' AND Year='2015'
```

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

Month_name	Year	Landing_Outcome	Booster_Version	Launch_Site
ENERO	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
ABRIL	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank of 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
FROM
    SPACEXTBL
WHERE
    Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
    "Landing_Outcome"
ORDER BY
    Count DESC
```

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

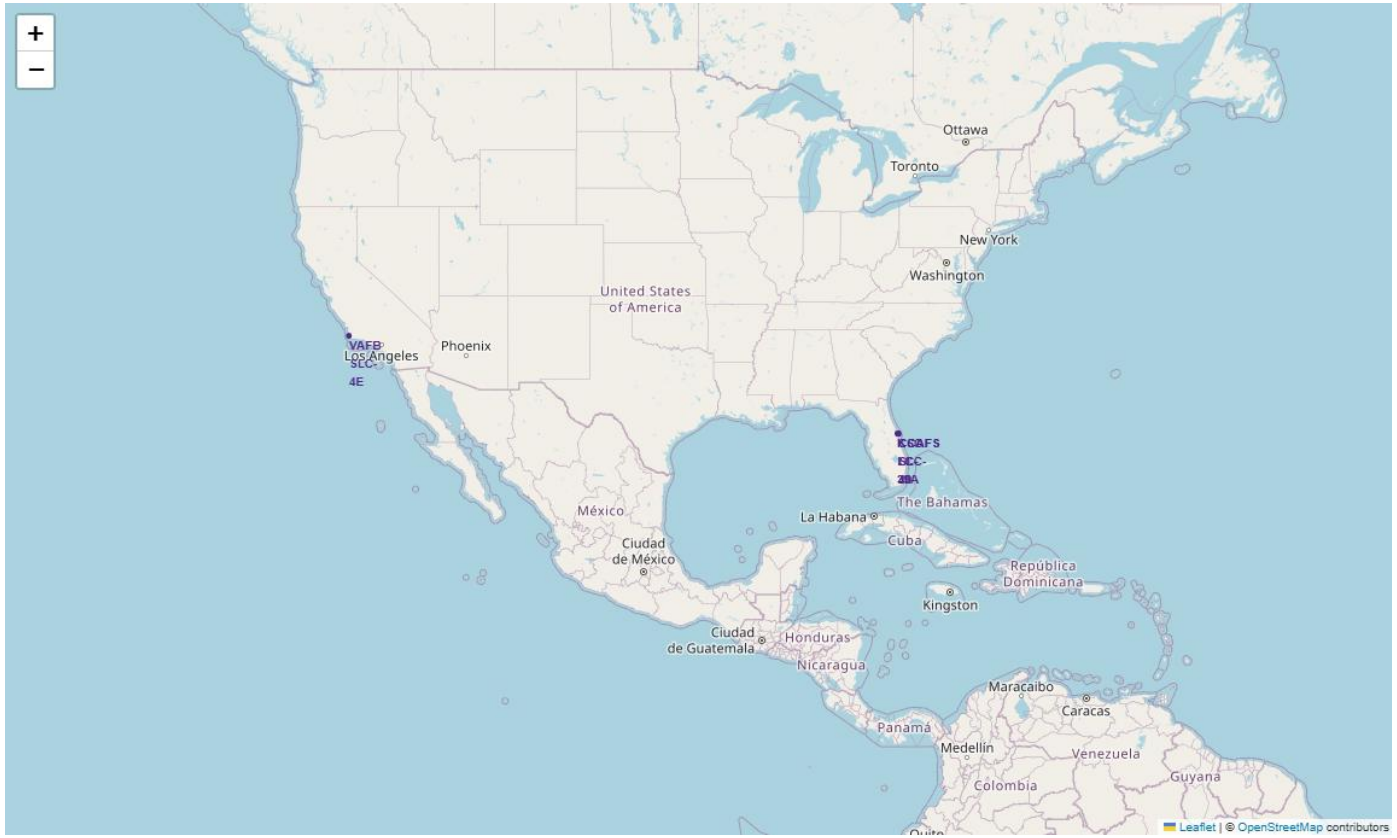
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

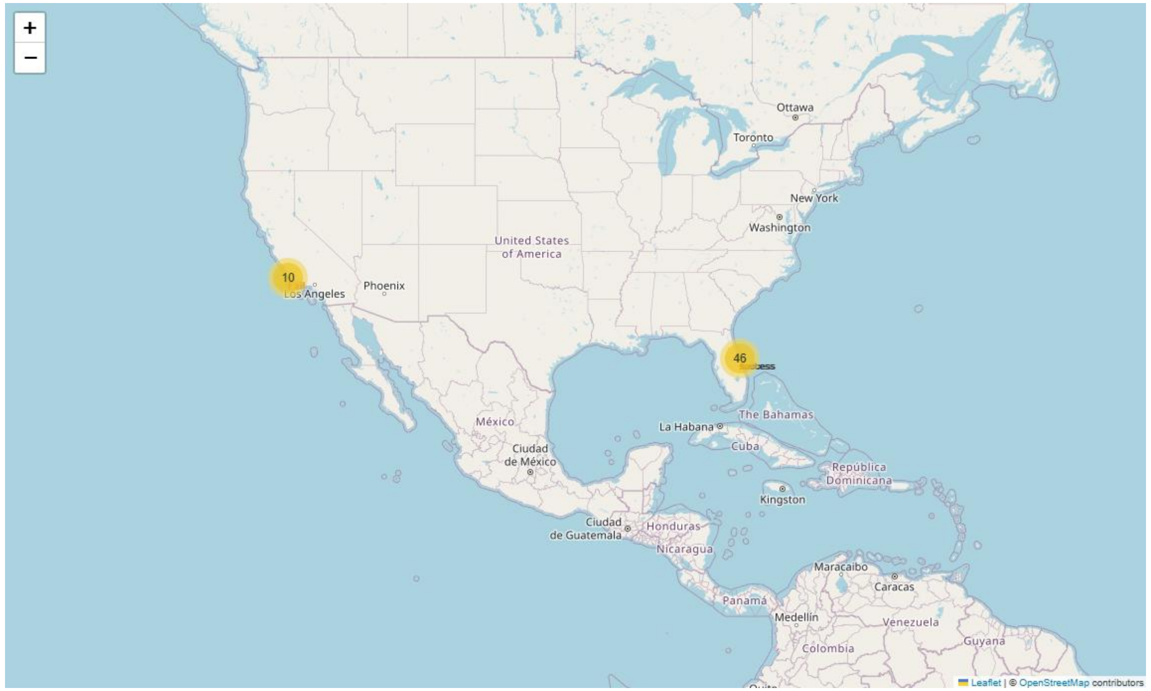
Location of all Launch Sites

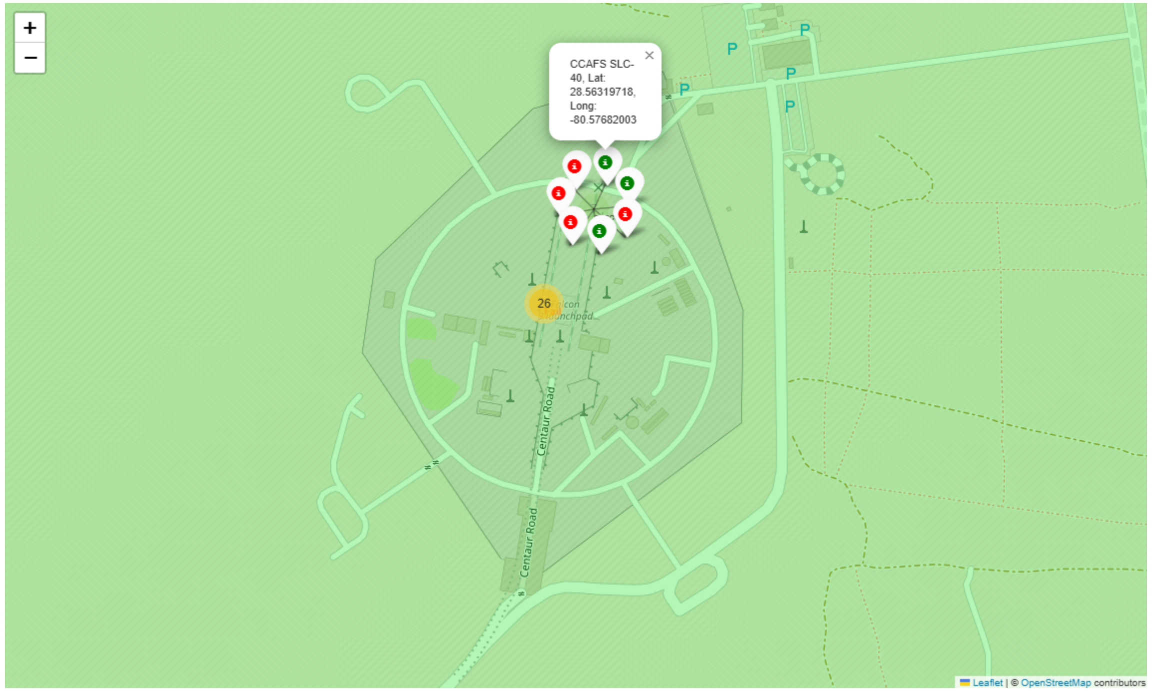
In this map we can see the four locations of the launch sites. One attribute we can say from this map is that the four sites are close to the coast



Success/Failed launches for each site

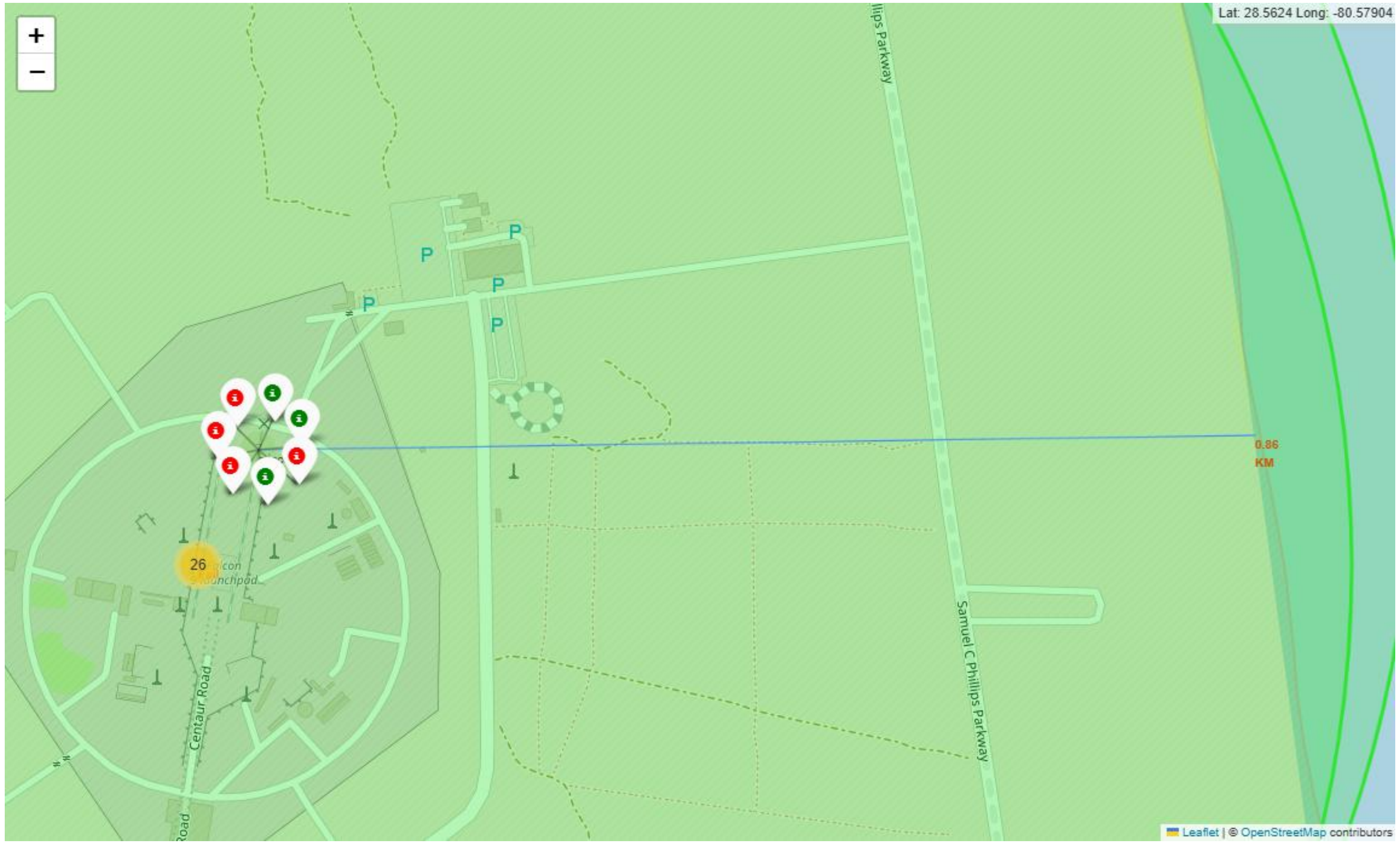
In this case we can see if a launch was a success or a failure in each site



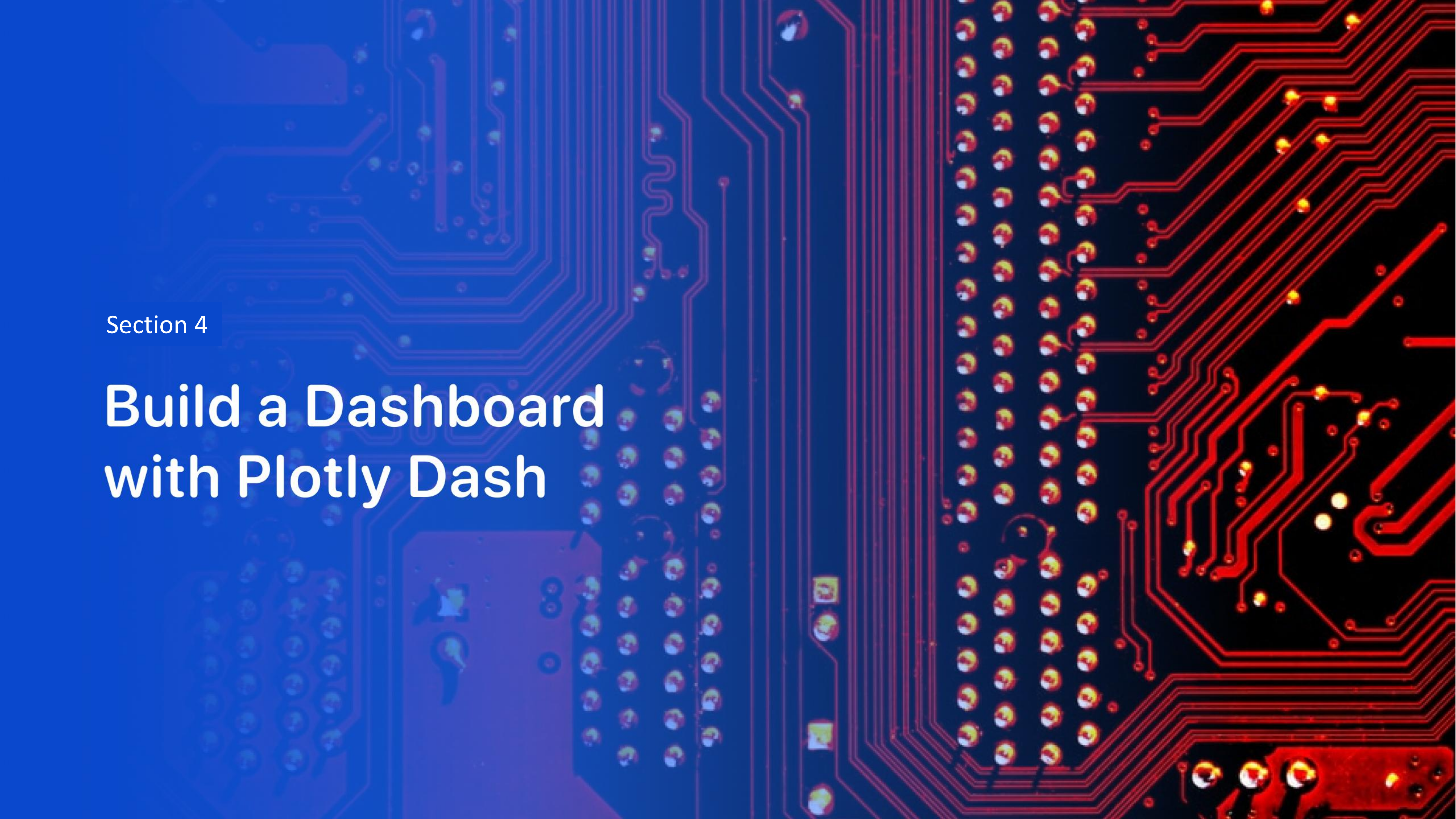

+
−
CCAFS SLC-40, Lat: 28.56319718, Long: -80.57682003
26
Centaur Road
Centaur Road
P
Leaflet | © OpenStreetMap contributors

Proximity between launch site and coastline

As we see in the past the four launch sites are close to coastlines, in this case **860m**



Lat: 28.5624 Long: -80.57904
+
−
0.86 KM
26
P
Centaur Road
Samuel C Phillips Parkway
llips Parkway
Leaflet | © OpenStreetMap contributors

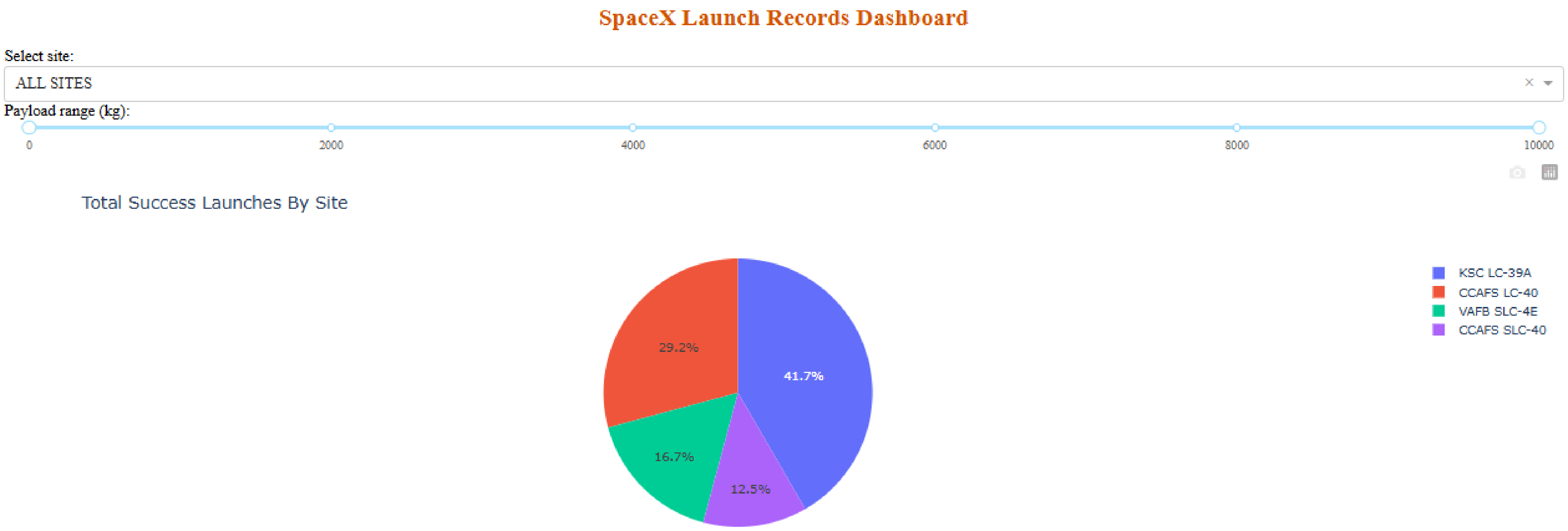


Section 4

Build a Dashboard with Plotly Dash

Success rate for all the sites

We can obtain different insights from this pie chart such as the site with the highest success rate(KSC LC-39A) and the site with the lowest success rate(CCAFS SLC-40).



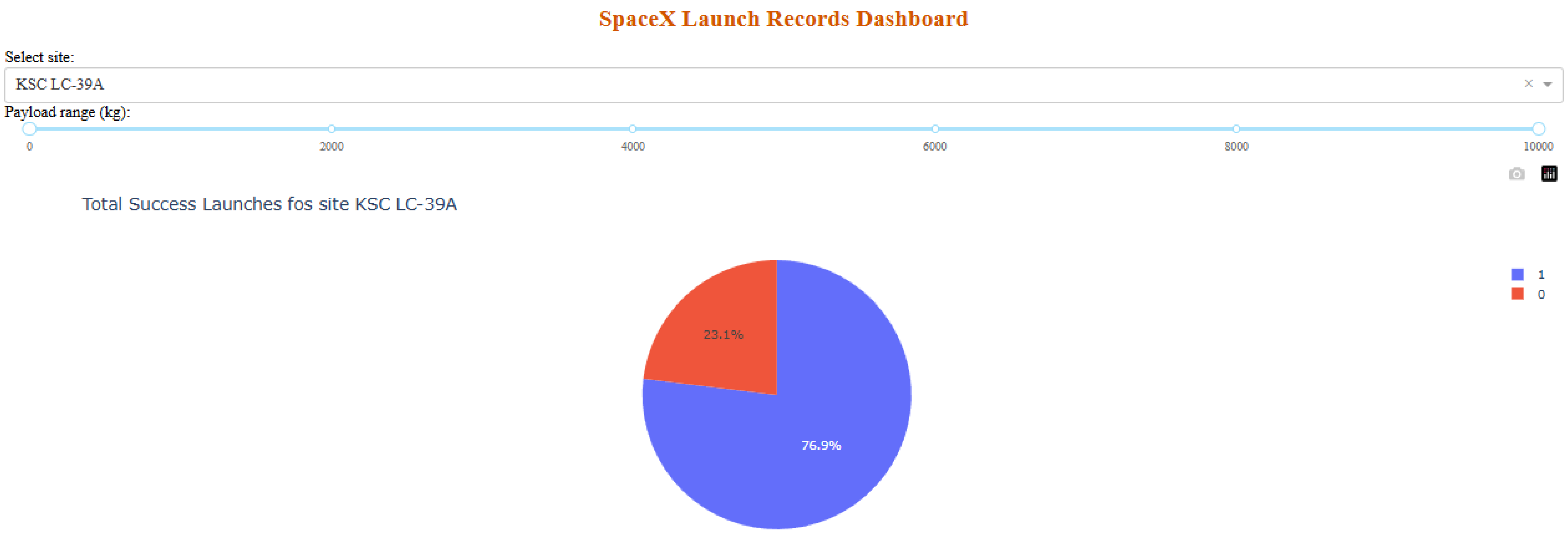
SpaceX Launch Records Dashboard

Select site: ALL SITES

Payload range (kg):

Total success/failure launches for KSC LC-39A

Now, for the highest launch site, we can see that 76.9% of his launches is success.



SpaceX Launch Records Dashboard

Select site:

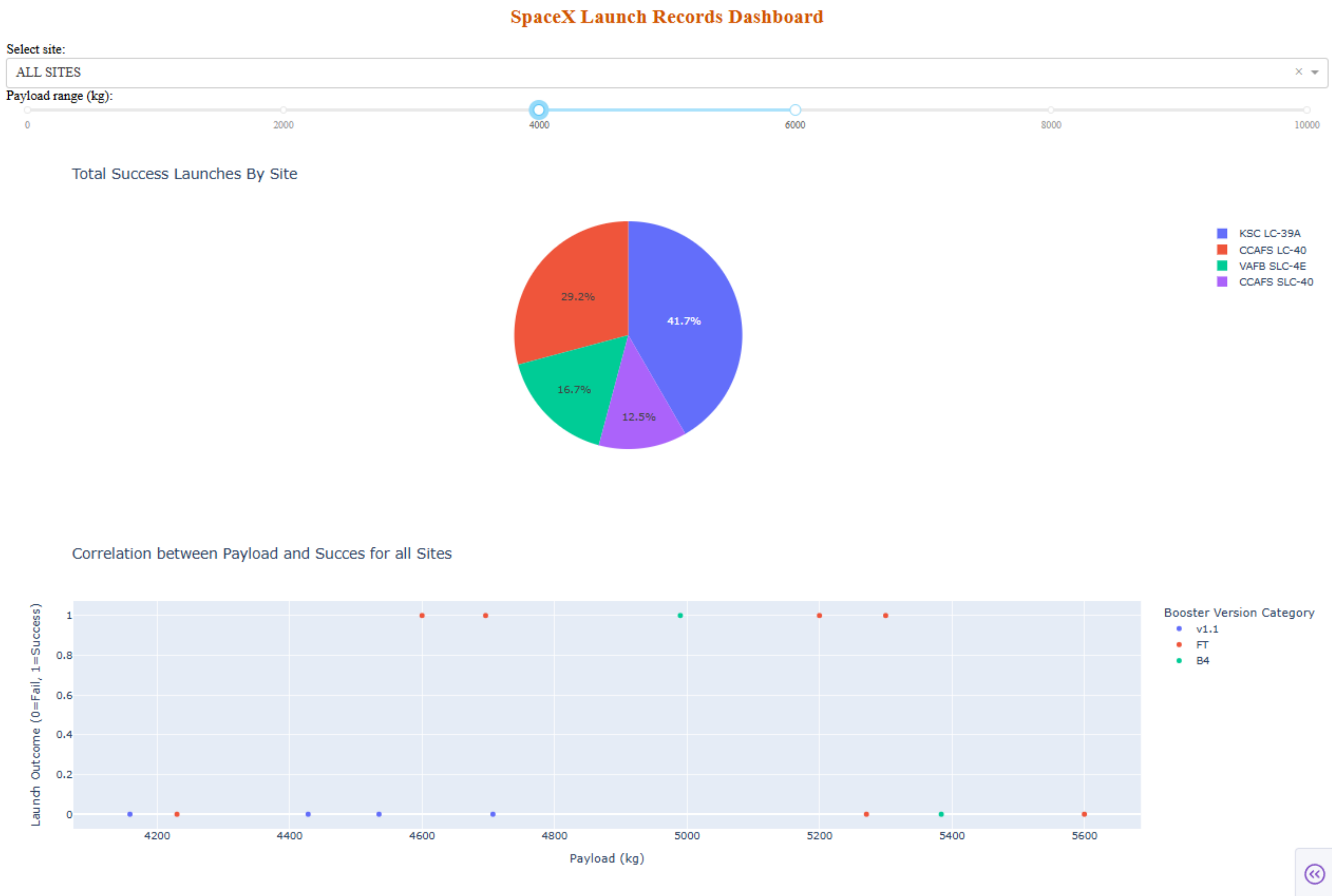
KSC LC-39A

Payload range (kg):

Total Success Launches fos site KSC LC-39A

Payload vs. Launch Outcome

We can see in this case from payload mass between 4000 and 6000 the best booster version is FT, while the worst is v1.1.



SpaceX Launch Records Dashboard

Select site:

ALL SITES

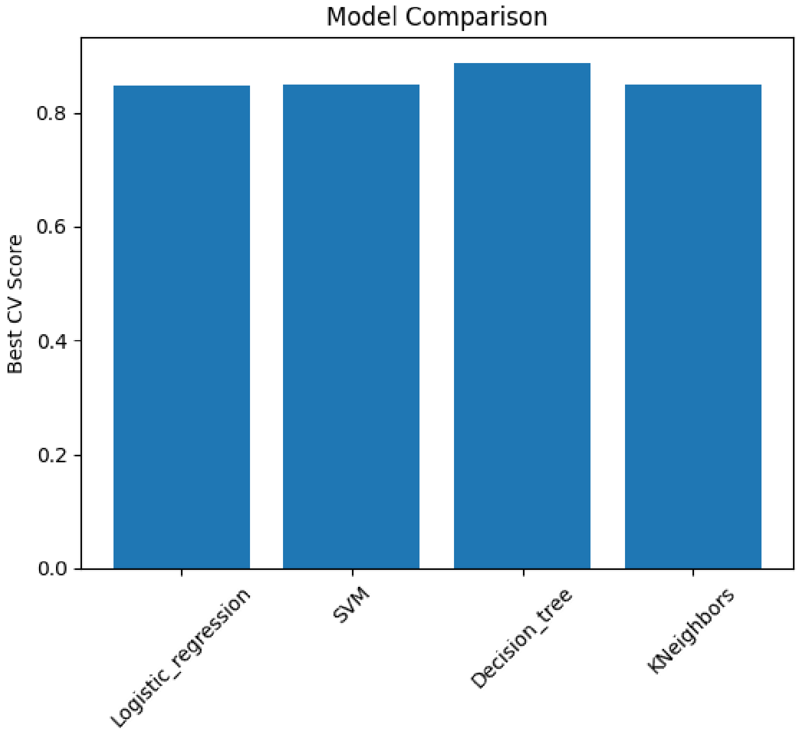
Payload range (kg):

Section 5

Predictive Analysis (Classification)

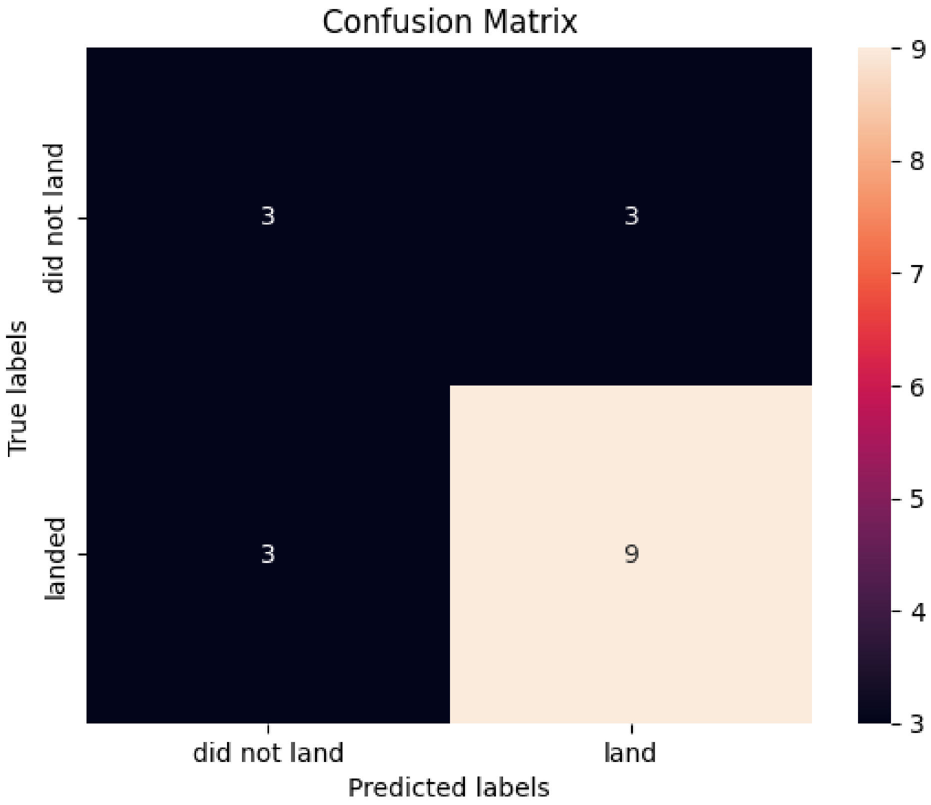
Classification Accuracy

The best model is decision tree



Confusion Matrix

We can see that the confusion matrix for decision tree is not the best, but it has the best accuracy.



Conclusions

After completing all the path to find different answers such as the principal features that affect the success landing, all the launch sites for some reason are nearby to a coastline, the best model to predict if a launch is going to be success is the decision tree and if we want to find out other answer we can use the interactive dashboard to figure out other insights.

Appendix

- Data collection API
- Data web scraping
- Data wrangling
- EDA
- EDA with SQL
- Site location
- Dashboard
- Machine Learning

Thank you!