IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>黃梃家
<Date>2022/7/27

Outline

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

Executive Summary

Methodologies:

- In the project, I perform data collection, data wrangling, using SQL to get information, using folium to make maps, and making a dash online, and then using machine learning techniques to build a predictive model.

Results:

- Through the project, I found that there are several practical insights from the data, and they all came from data methodologies.

Introduction

- The project is to get to know how to predict the success rate of a launch of the rocket through the data of SpaceX.
- The problem I want to discover is that in which condition the rocket launch has a higher success rate.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

In this stage, I perform SpaceX API and web scrapping in order to get the information I need. SpaceX API can help me get the rough information of the website, and web scrapping can help filter unneeded information.

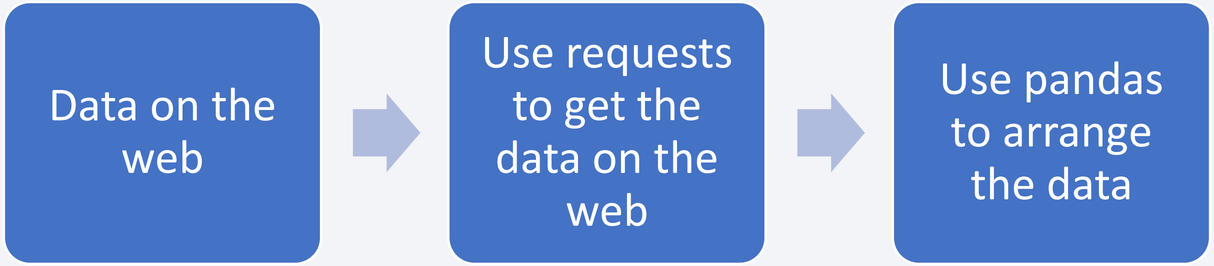
SpaceX API

Web scrapping

data

Data Collection – SpaceX API

- Use **get** statement in requests to obtain information in SpaceX website.
- Use **.json()** and **pd.json_normalize()** to arrange the data
- Use methods of pd.dataFrame to filter the data
- the code is on my github



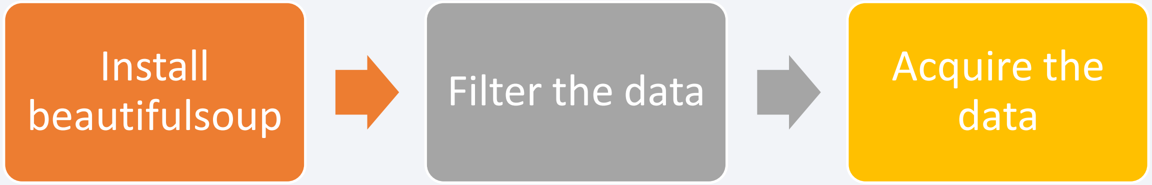
Data on the web

Use requests to get the data on the web

Use pandas to arrange the data

Data Collection - Scraping

- In the stage, I use the package called beautifulsoup to acquire the aimed data.
- I use methods such as find_all and title in the package to filter the data.
- the code is on my github



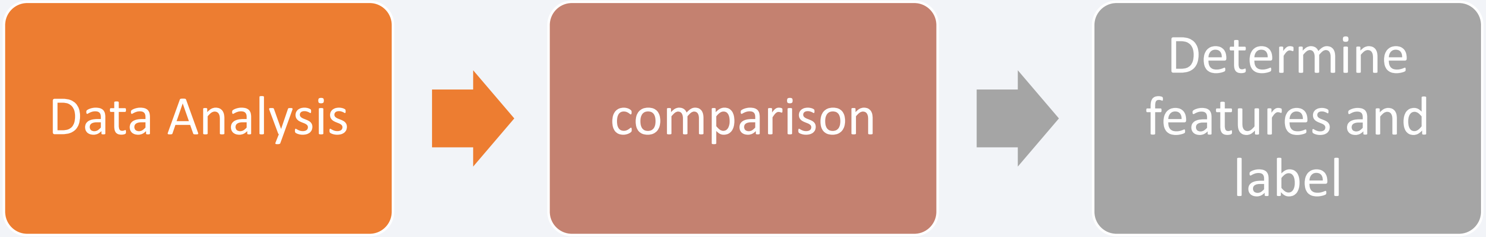
Install beautifulsoup

Filter the data

Acquire the data

Data Wrangling

- I performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.
- The code is on my github.



Data Analysis

comparison

Determine features and label

EDA with Data Visualization

- With a view to understanding the data further, I use different kinds of charts to present the data, such as bar chart, scatter chart, line chart, and pie chart. All of them were created using python's module, matplotlib and seaborn.
- The code is on my github.

EDA with SQL

- I use SQL to get information about launch site , customer, date, booster version then compare them with the landing outcome and count the number of successful cases.
- The code is on my github.

Build an Interactive Map with Folium

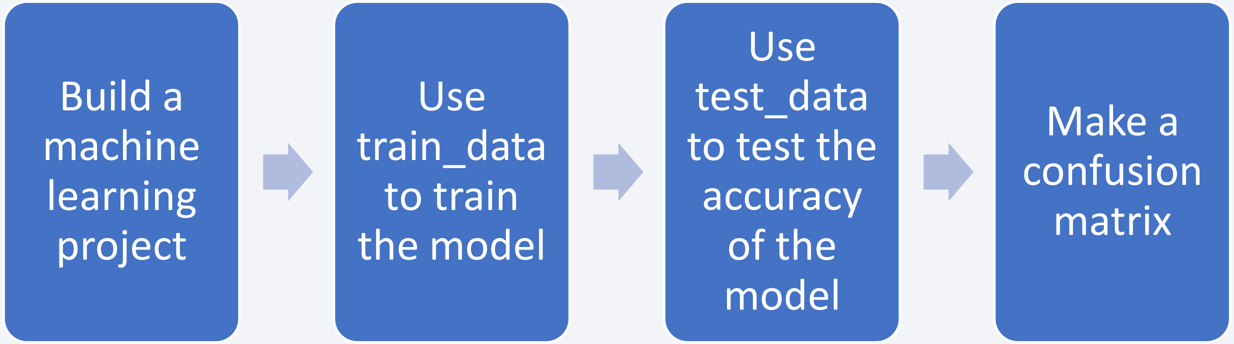
- Throughout the stage, I create circle, marker, MarkerCluster, and MousePosition objects.
- I create them in order to do some tags to distinguish target points from the map.
- The code is on my github.

Build a Dashboard with Plotly Dash

- In the stage, I use the success count of every launch-site to do a scatter diagram and then use it to build a interactive pie plot with dash packages.
- With the dashboard, I am allowed to understand the success proportion of every case.
- My code is on my github.

Predictive Analysis (Classification)

- Throughout the case, I use KNN, decision tree, logistic regression, and SVC to train the data and then use test_data to test its accuracy and made a confusion matrix to observe the result.
- I use sklearn library to train the model. The process can be presented as belows.
- My code is on my github.



Build a machine learning project

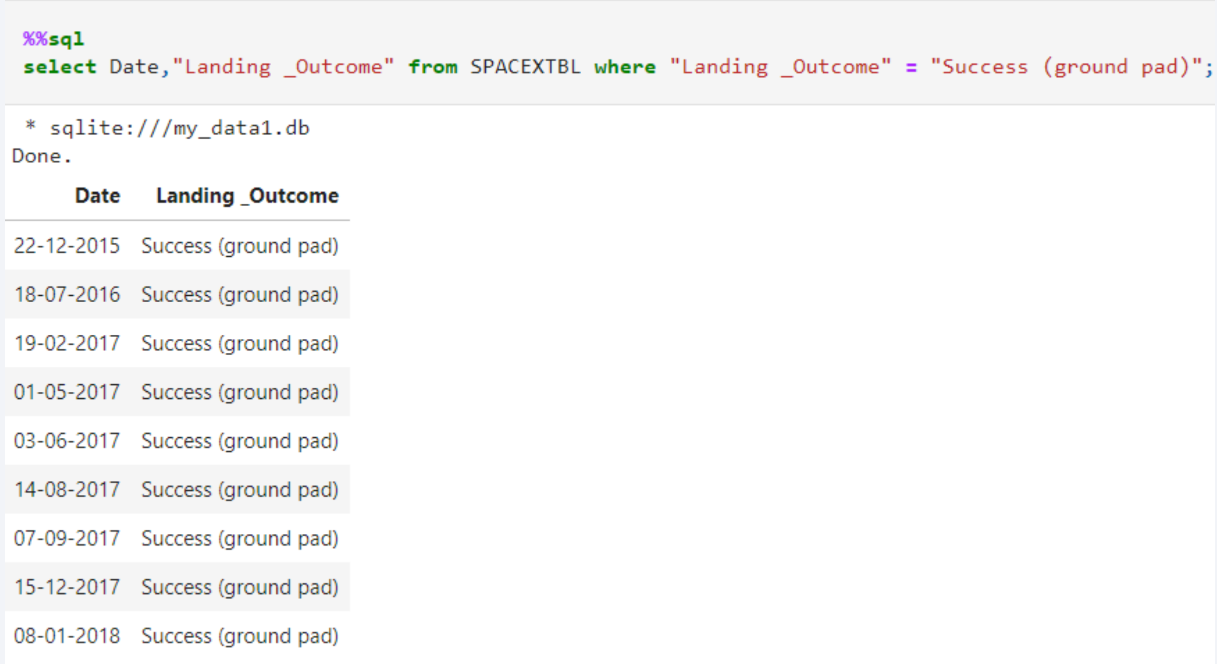
Use train_data to train the model

Use test_data to test the accuracy of the model

Make a confusion matrix

Results

- Through exploratory data analysis, I can get a set of charts that show the relation of every parameters.
- The accuracy of the model I built is about 80%.



```
%%sql
select Date,"Landing _Outcome" from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)";
```

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

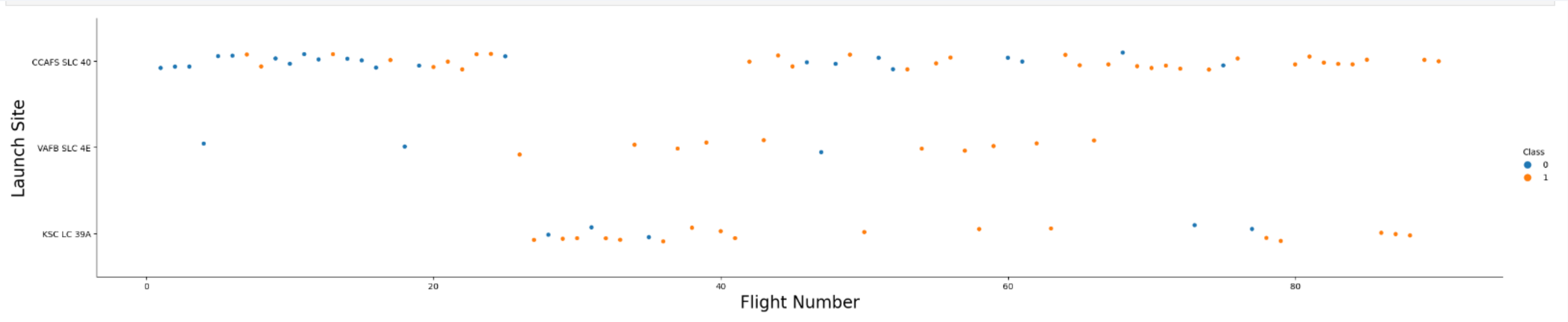
Date	Landing _Outcome
22-12-2015	Success (ground pad)
18-07-2016	Success (ground pad)
19-02-2017	Success (ground pad)
01-05-2017	Success (ground pad)
03-06-2017	Success (ground pad)
14-08-2017	Success (ground pad)
07-09-2017	Success (ground pad)
15-12-2017	Success (ground pad)
08-01-2018	Success (ground pad)

The screenshot shows how I did interactive analysis.

Section 2

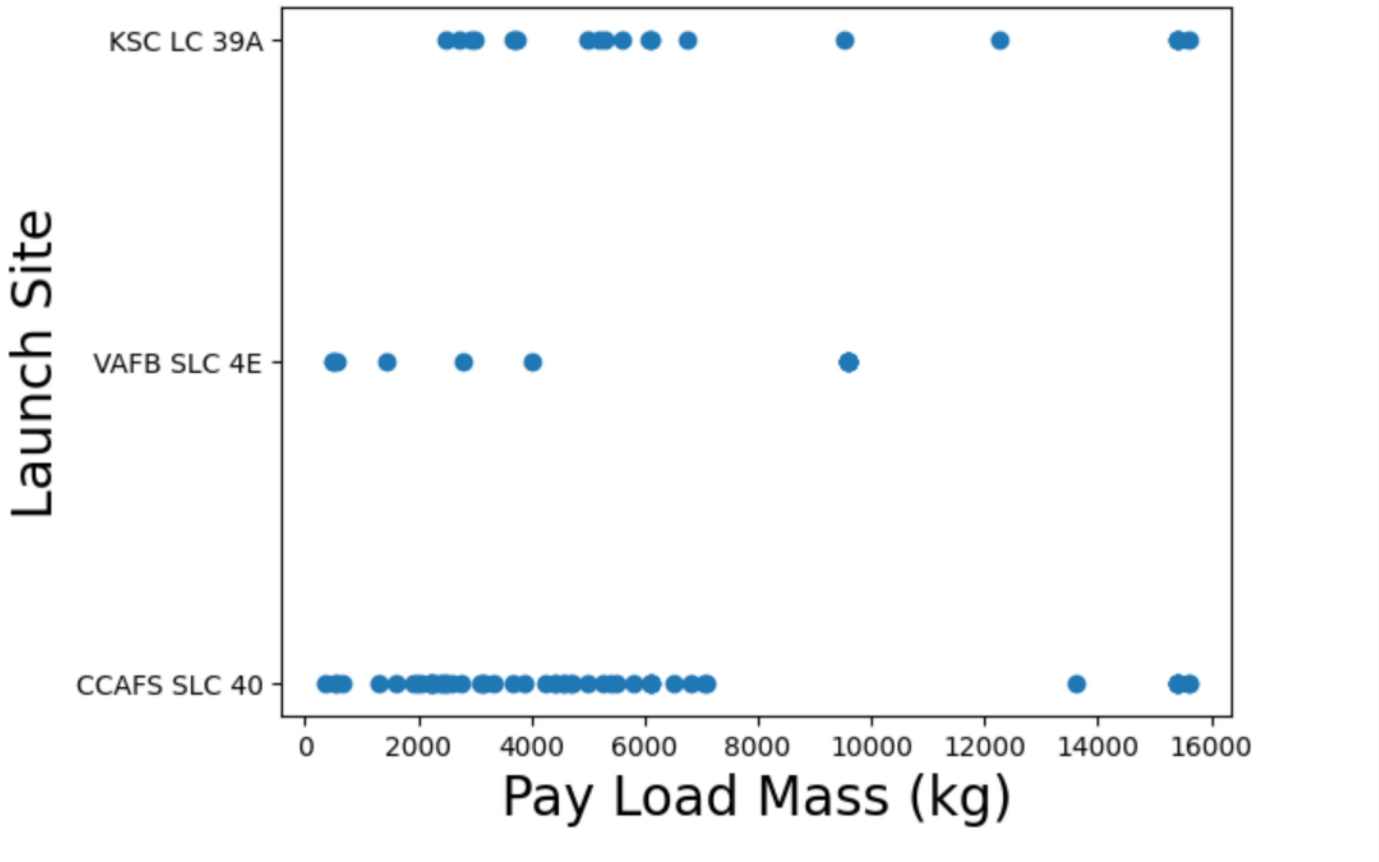
Insights drawn from EDA

Flight Number vs. Launch Site



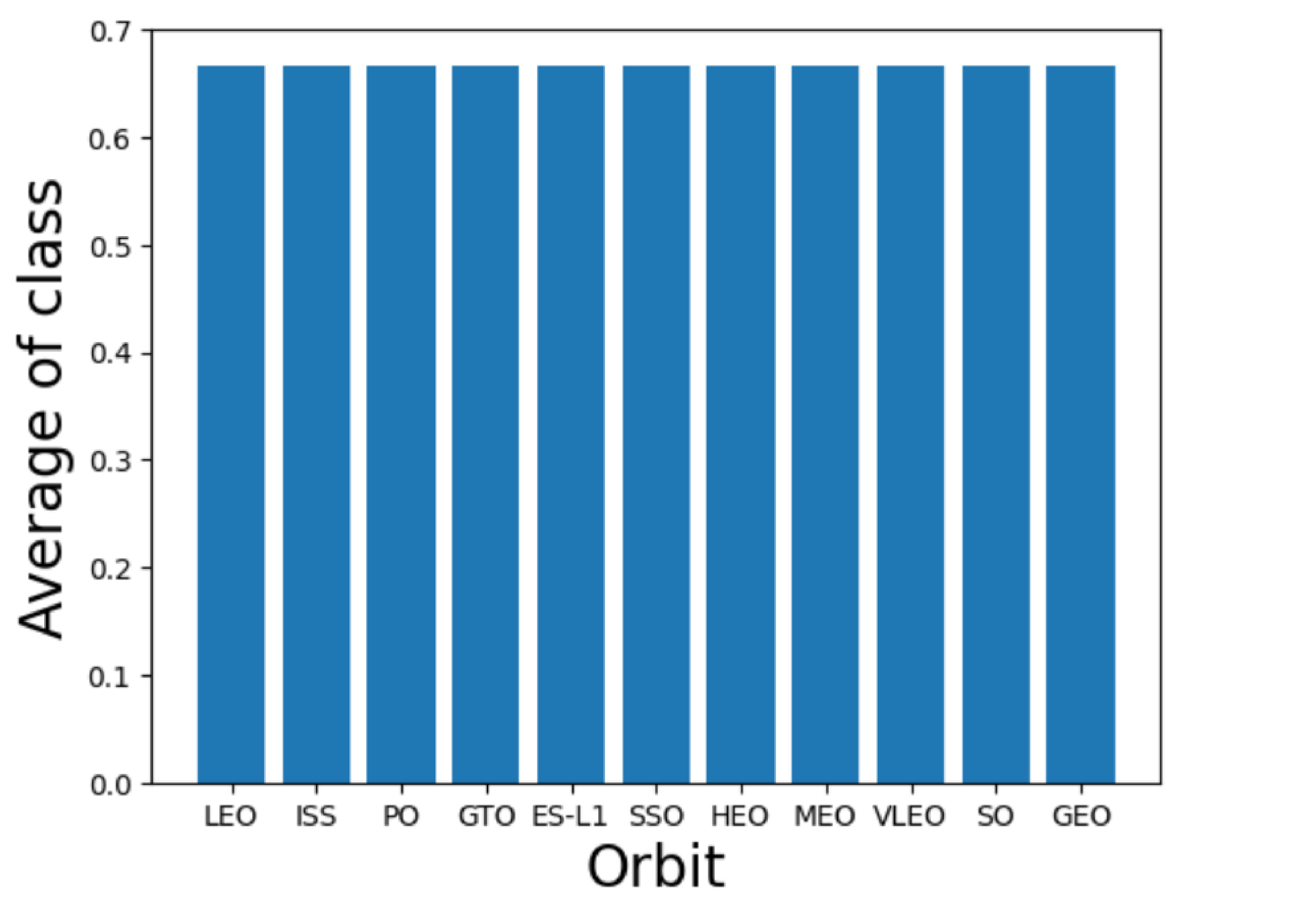
With above chart, we can observe that flight number in KSC LC 39A station is always more than twenty. Then, VAFB SLC 4E is a station that the flight number is spread evenly in the chart. Finally, CCAFS SLC 40 is a station that there is no flight number between 25 and 40.

Payload vs. Launch Site



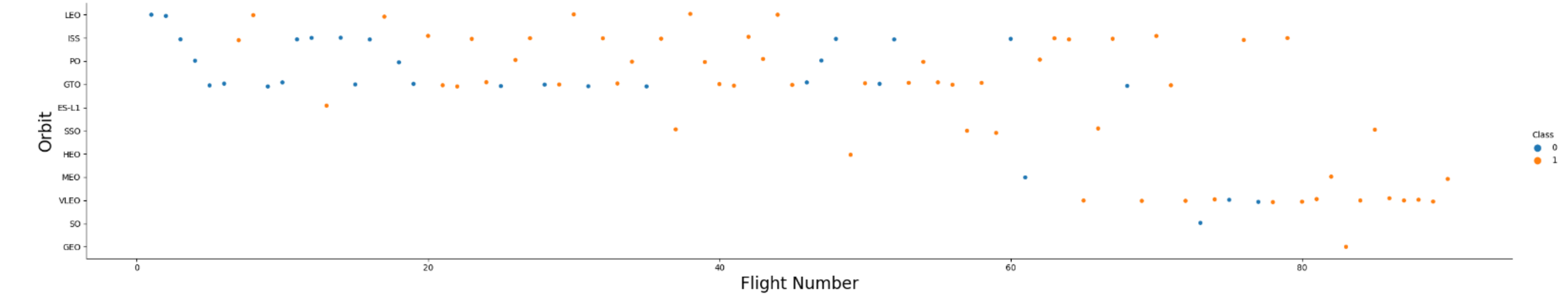
We can observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



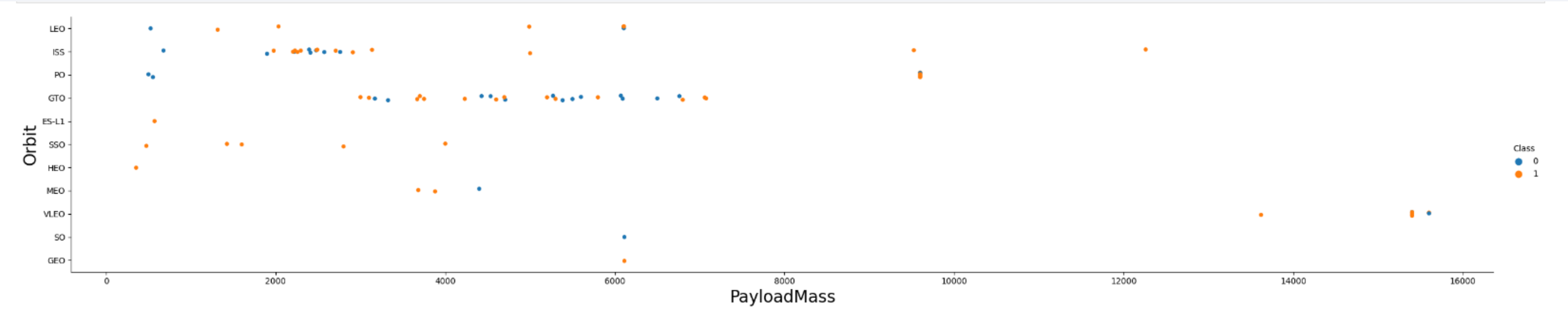
The success rate of every orbit is close.

Flight Number vs. Orbit Type



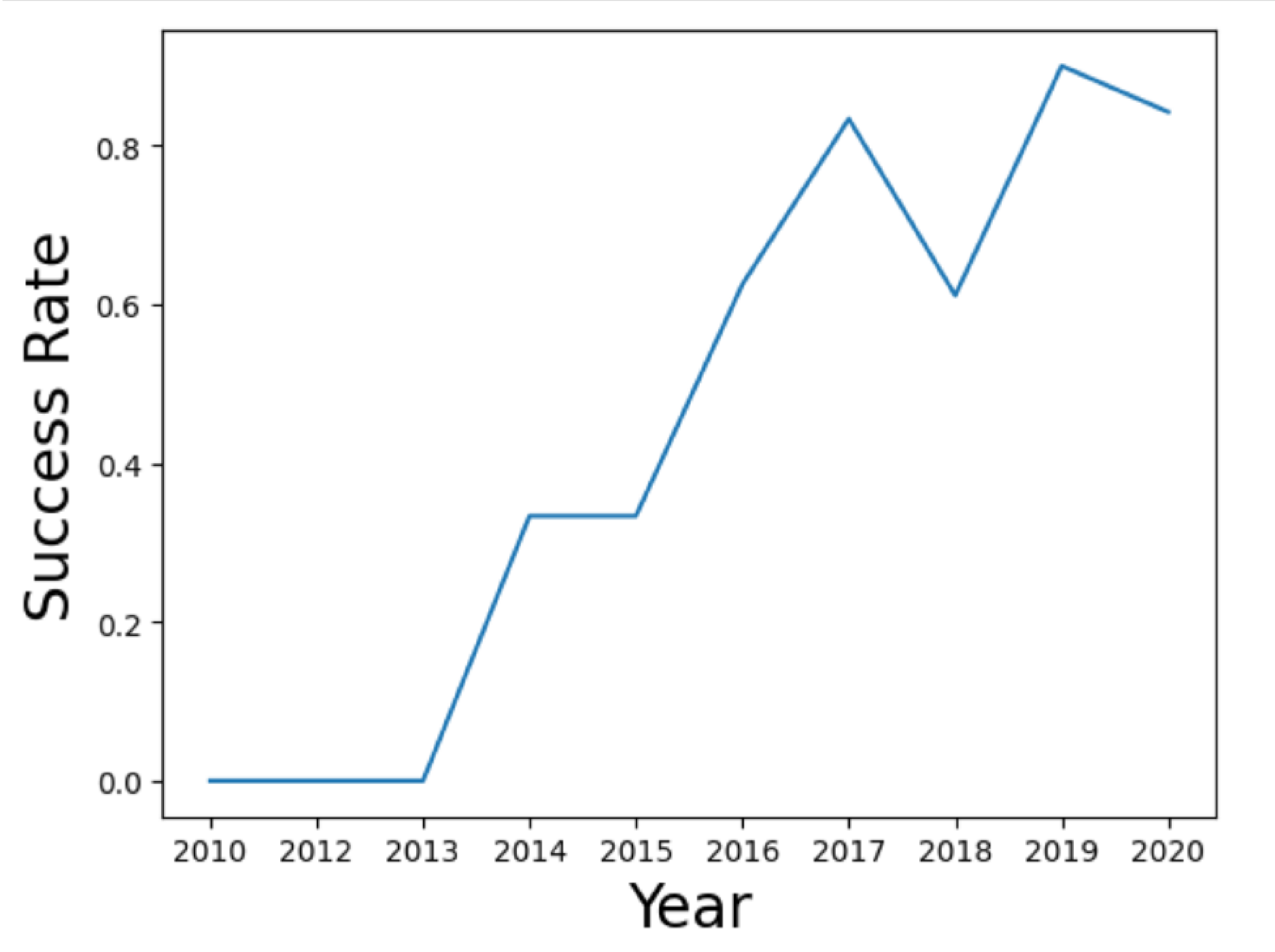
We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



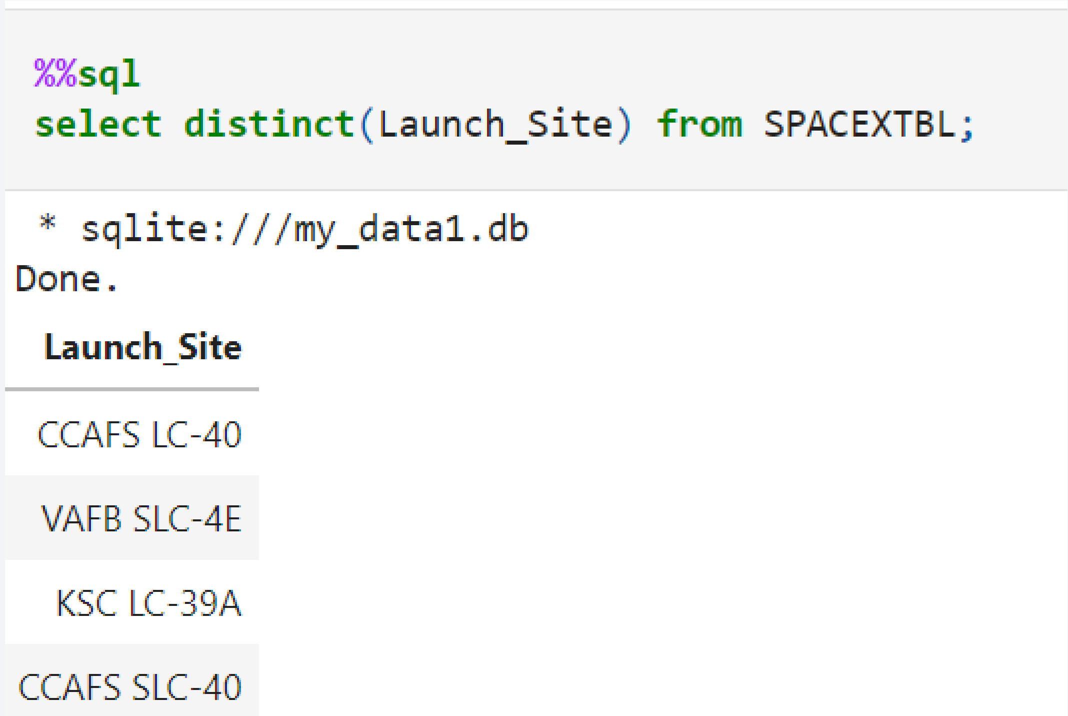
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



We can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names



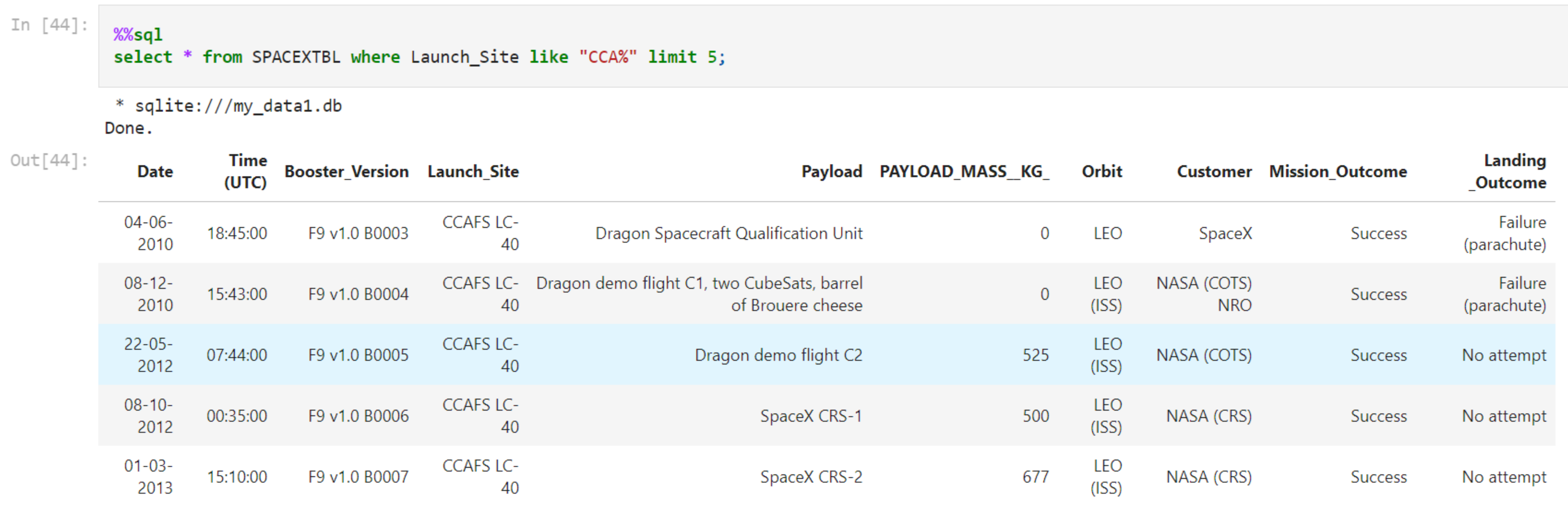
```
%%sql
select distinct(Launch_Site) from SPACEXTBL;
```

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

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

We can know that there are four launch sites.
CCAFS LC-40,VAFB SLC-4E,KSC LC-39A,and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'



```
In [44]: %%sql
         select * from SPACEXTBL where Launch_Site like "CCA%" limit 5;
```

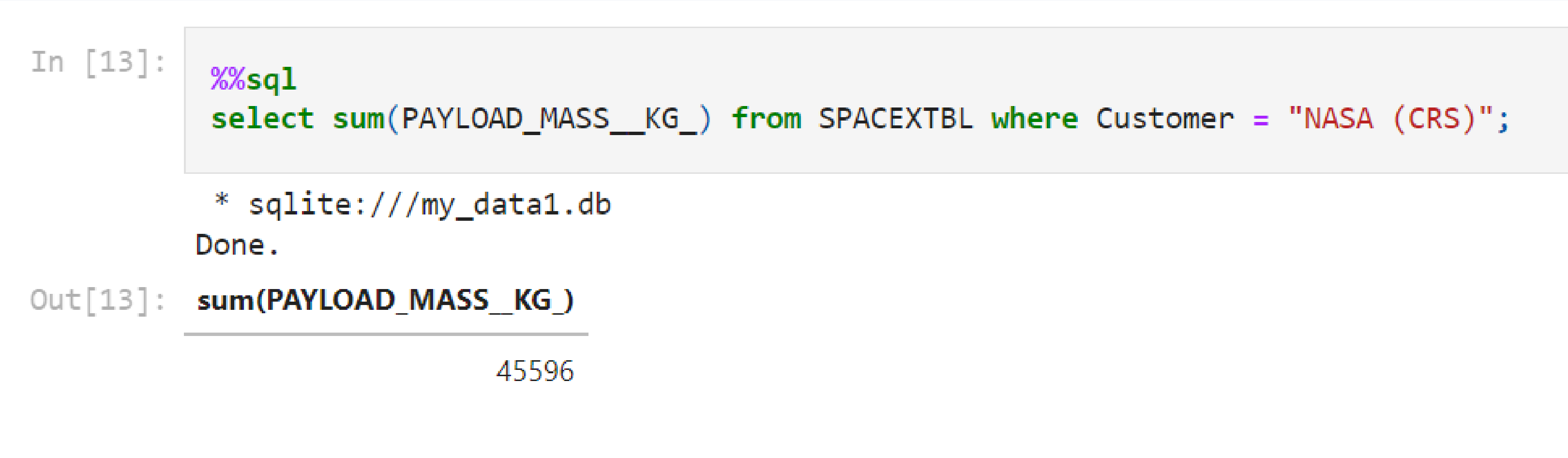
```
 * sqlite:///my_data1.db
Done.
```

Out[44]:

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

I use SQL query to get launch sites begin with the string 'CCA'

Total Payload Mass



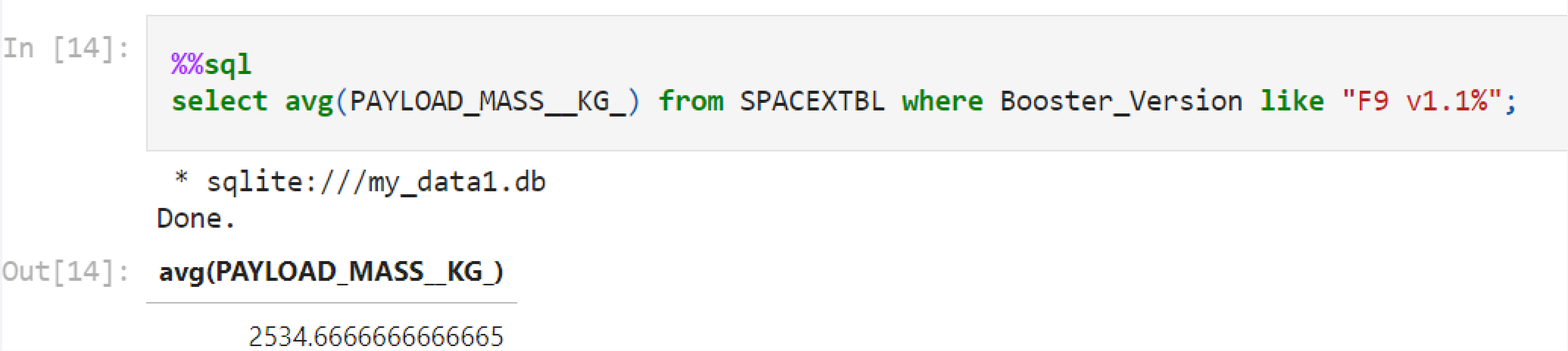
```
In [13]: %%sql
         select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where Customer = "NASA (CRS)";

          * sqlite:///my_data1.db
         Done.

Out[13]: sum(PAYLOAD_MASS__KG_)
                          45596
```

We can realize that sum of payload is 45596 kilograms(NASA) by using SQL.

Average Payload Mass by F9 v1.1



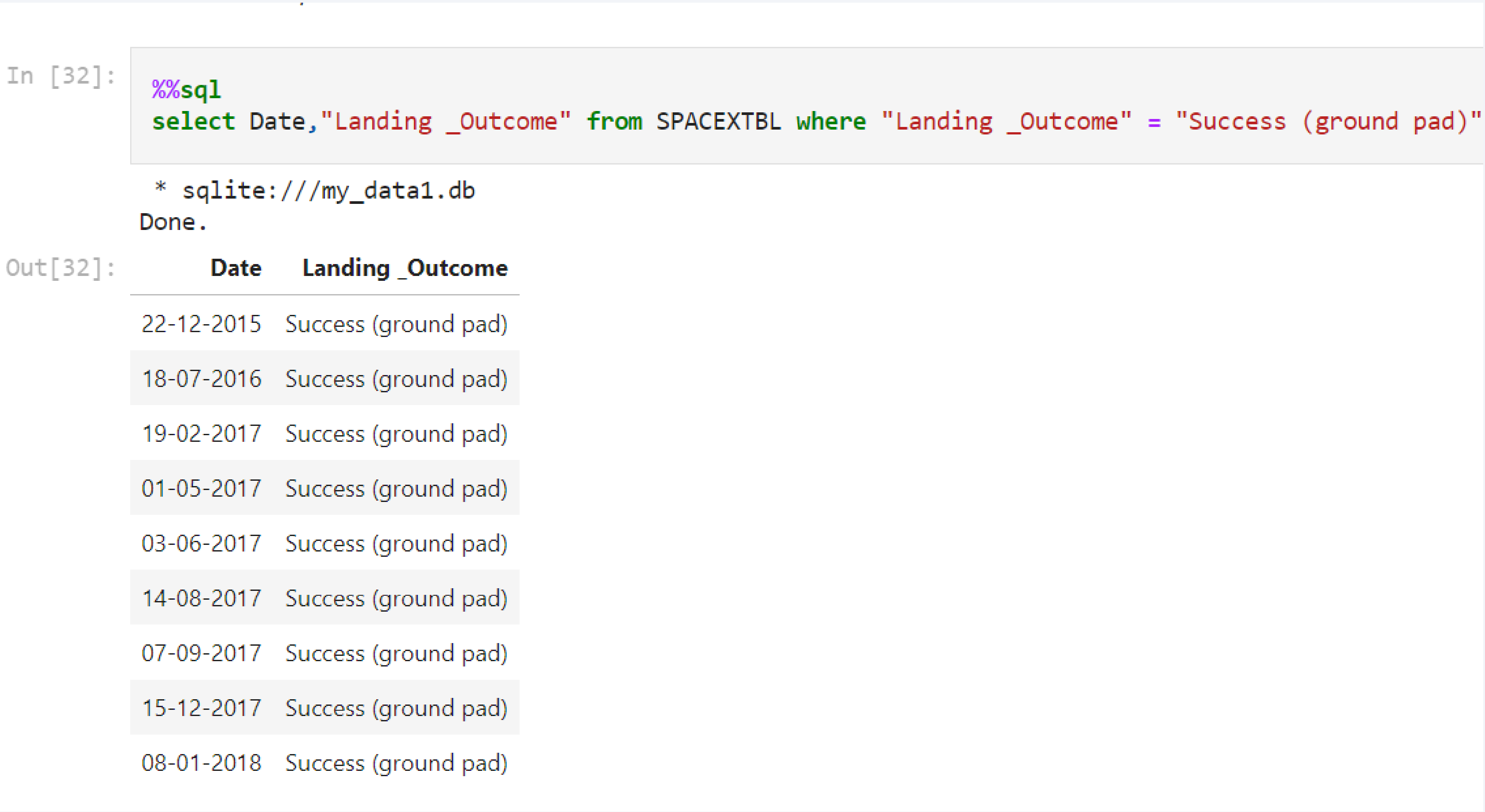
```
In [14]:  %%sql
          select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where Booster_Version like "F9 v1.1%";

           * sqlite:///my_data1.db
          Done.

Out[14]:  avg(PAYLOAD_MASS__KG_)
                2534.6666666666665
```

We can realize that average of payload is 2534.7 kilograms(F9 v1.1) by using SQL.

First Successful Ground Landing Date



In [32]:

```
%%sql
select Date,"Landing _Outcome" from SPACEXTBL where "Landing _Outcome" = "Success (ground pad)"
```

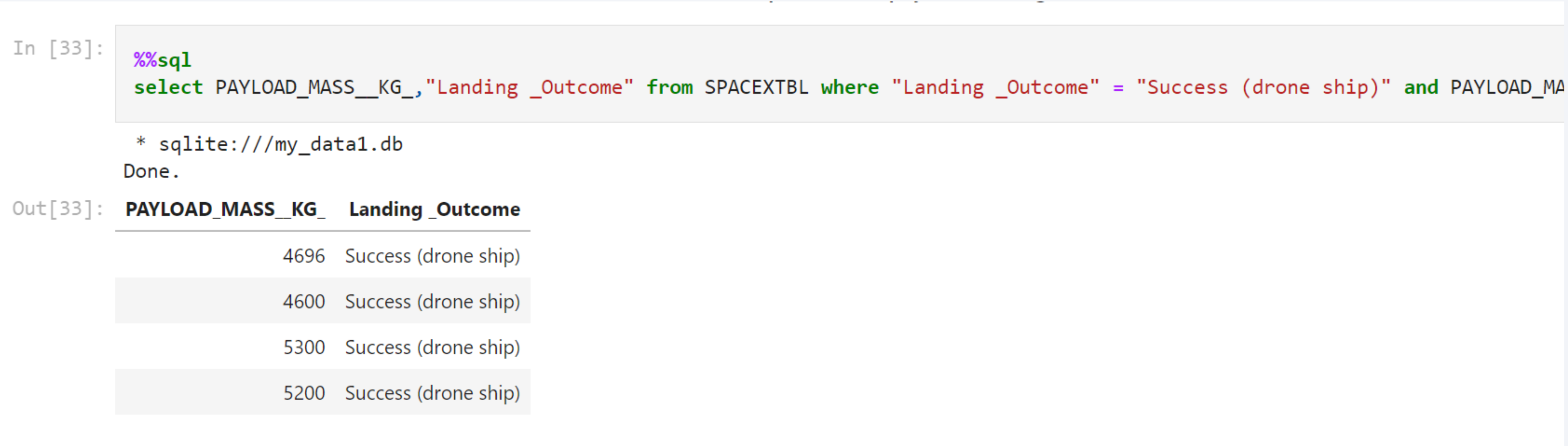
```
 * sqlite:///my_data1.db
Done.
```

Out[32]:

Date	Landing _Outcome
22-12-2015	Success (ground pad)
18-07-2016	Success (ground pad)
19-02-2017	Success (ground pad)
01-05-2017	Success (ground pad)
03-06-2017	Success (ground pad)
14-08-2017	Success (ground pad)
07-09-2017	Success (ground pad)
15-12-2017	Success (ground pad)
08-01-2018	Success (ground pad)

We can know that first successful landing date is 2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000



```
In [33]:
%%sql
select PAYLOAD_MASS__KG_,"Landing _Outcome" from SPACEXTBL where "Landing _Outcome" = "Success (drone ship)" and PAYLOAD_MA
```

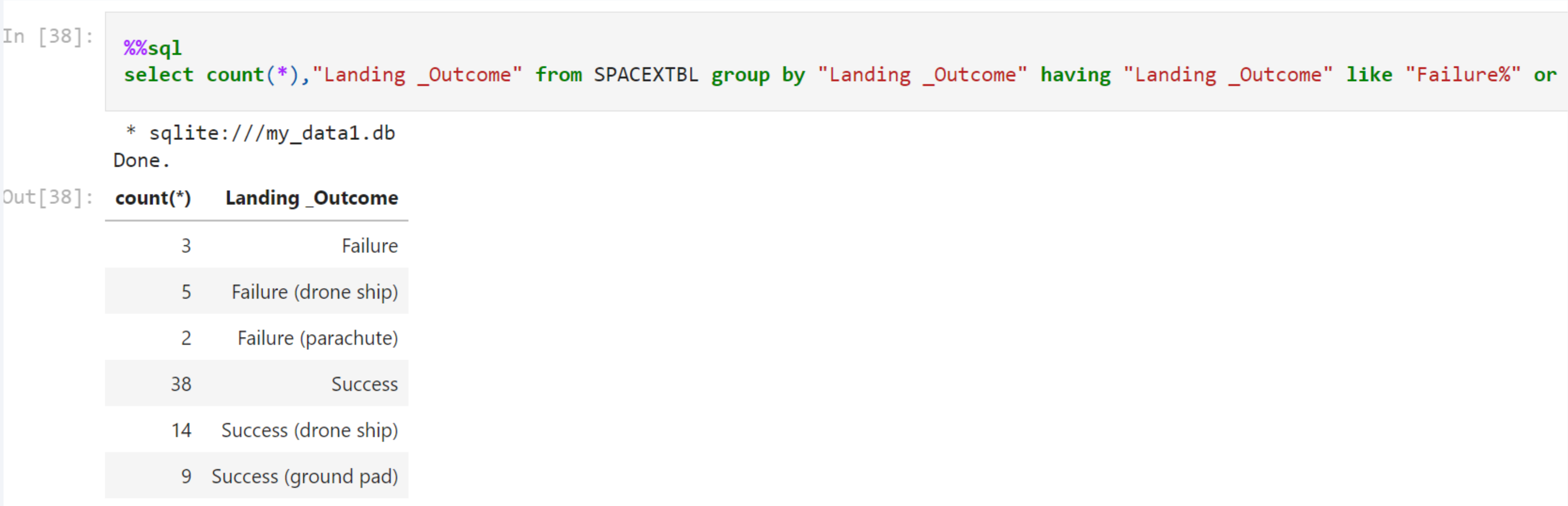
```
 * sqlite:///my_data1.db
Done.
```

Out[33]:

PAYLOAD_MASS__KG_	Landing _Outcome
4696	Success (drone ship)
4600	Success (drone ship)
5300	Success (drone ship)
5200	Success (drone ship)

We can see there are four cases that payload between 4000kg and 6000kg landed successfully.

Total Number of Successful and Failure Mission Outcomes



In [38]:

```
%%sql
select count(*),"Landing _Outcome" from SPACEXTBL group by "Landing _Outcome" having "Landing _Outcome" like "Failure%" or
```

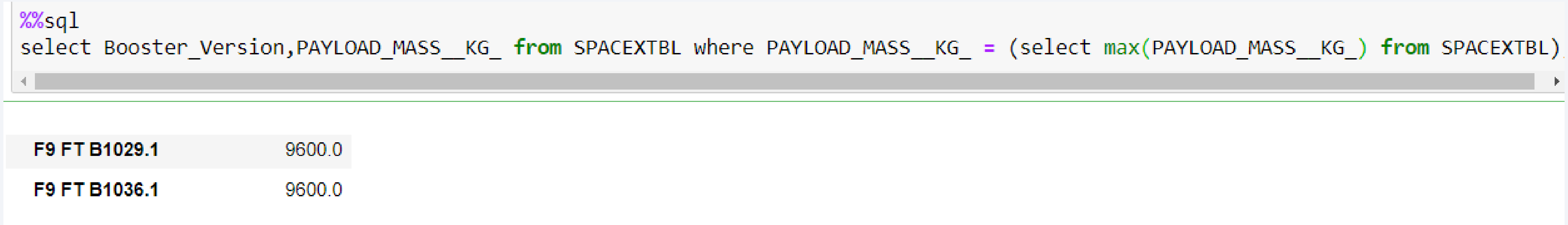
```
 * sqlite:///my_data1.db
Done.
```

Out[38]:

count(*)	Landing _Outcome
3	Failure
5	Failure (drone ship)
2	Failure (parachute)
38	Success
14	Success (drone ship)
9	Success (ground pad)

According to above query, there are 61 successful cases and 10 failure cases.

Boosters Carried Maximum Payload

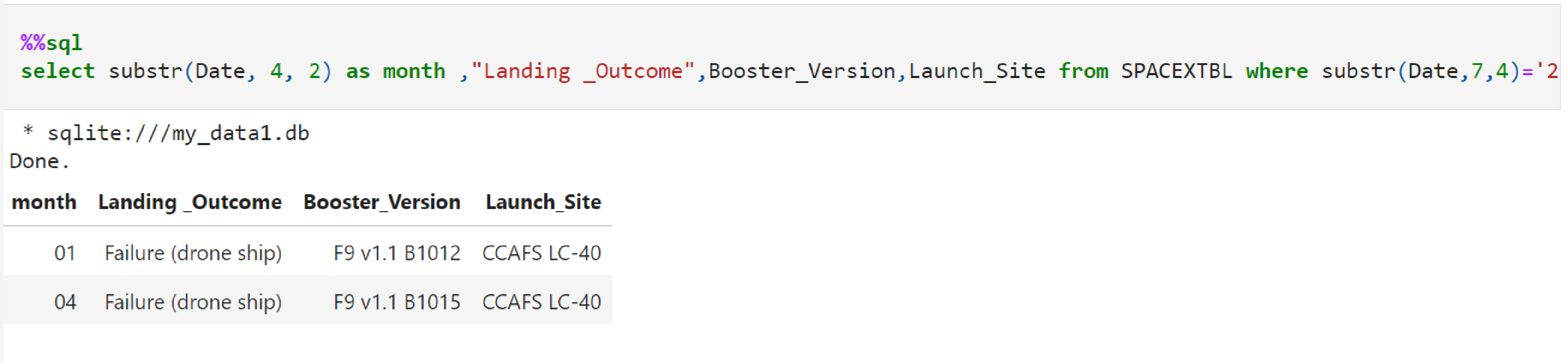


```
%%sql
select Booster_Version,PAYLOAD_MASS__KG_ from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

F9 FT B1029.1	9600.0
F9 FT B1036.1	9600.0

We can observe that the maximum payload is 9600 kg

2015 Launch Records



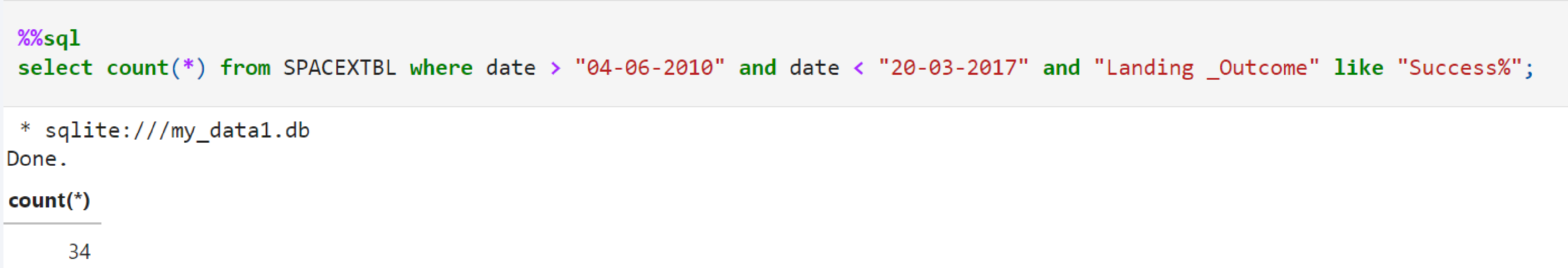
```
%%sql
select substr(Date, 4, 2) as month ,"Landing _Outcome",Booster_Version,Launch_Site from SPACEXTBL where substr(Date,7,4)='2
```

```
 * 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

We can get to know that there are two fail launch in 2015.

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



```
%%sql
select count(*) from SPACEXTBL where date > "04-06-2010" and date < "20-03-2017" and "Landing _Outcome" like "Success%";
```

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

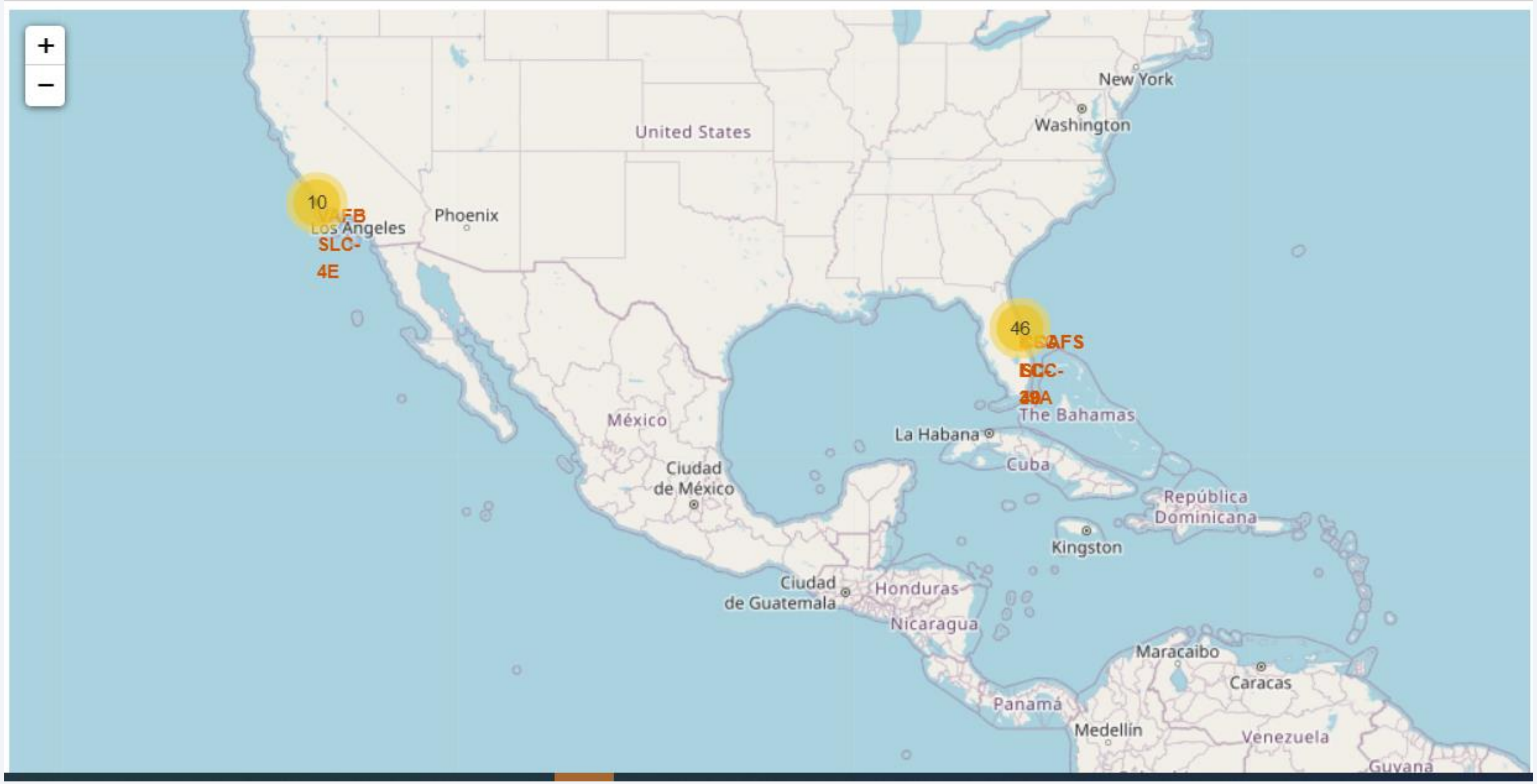
count(*)
34

There are 34 successful cases between 2010-06-04 and 2017-03-20.

Section 3

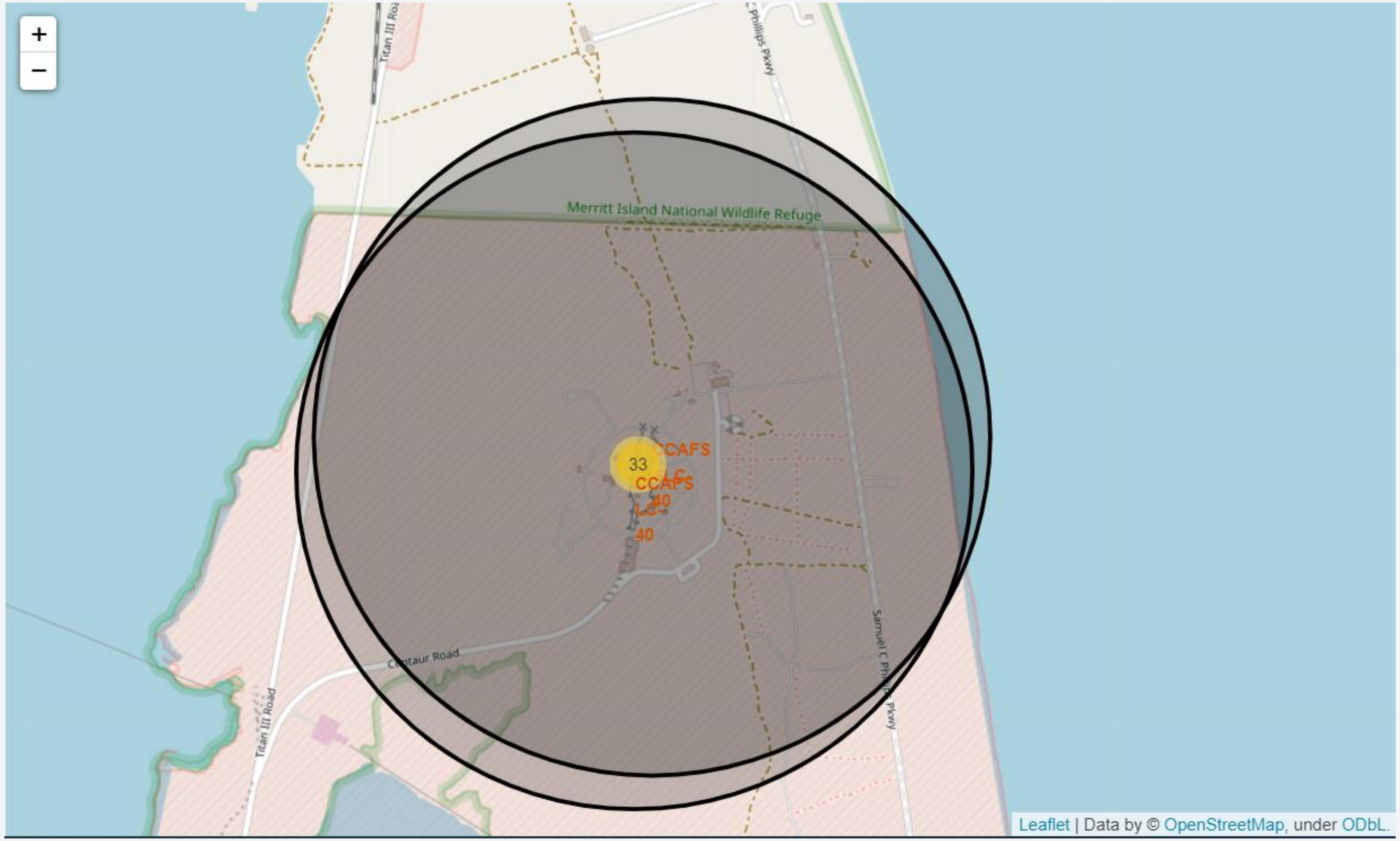
Launch Sites Proximities Analysis

<Approximate location of launch sites>



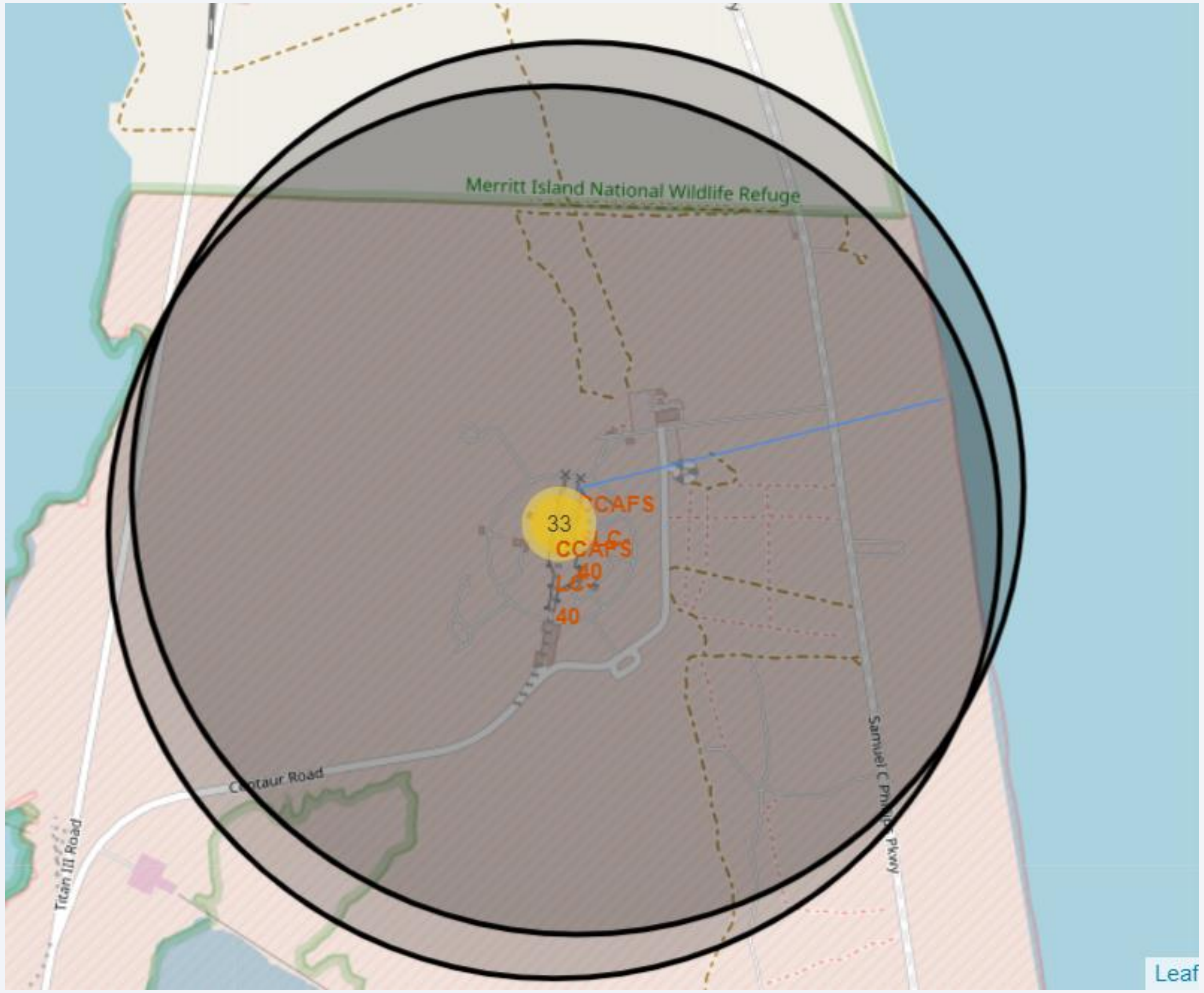
According to the map, the launch site are all near coastline.

<Proximity of launch sites>

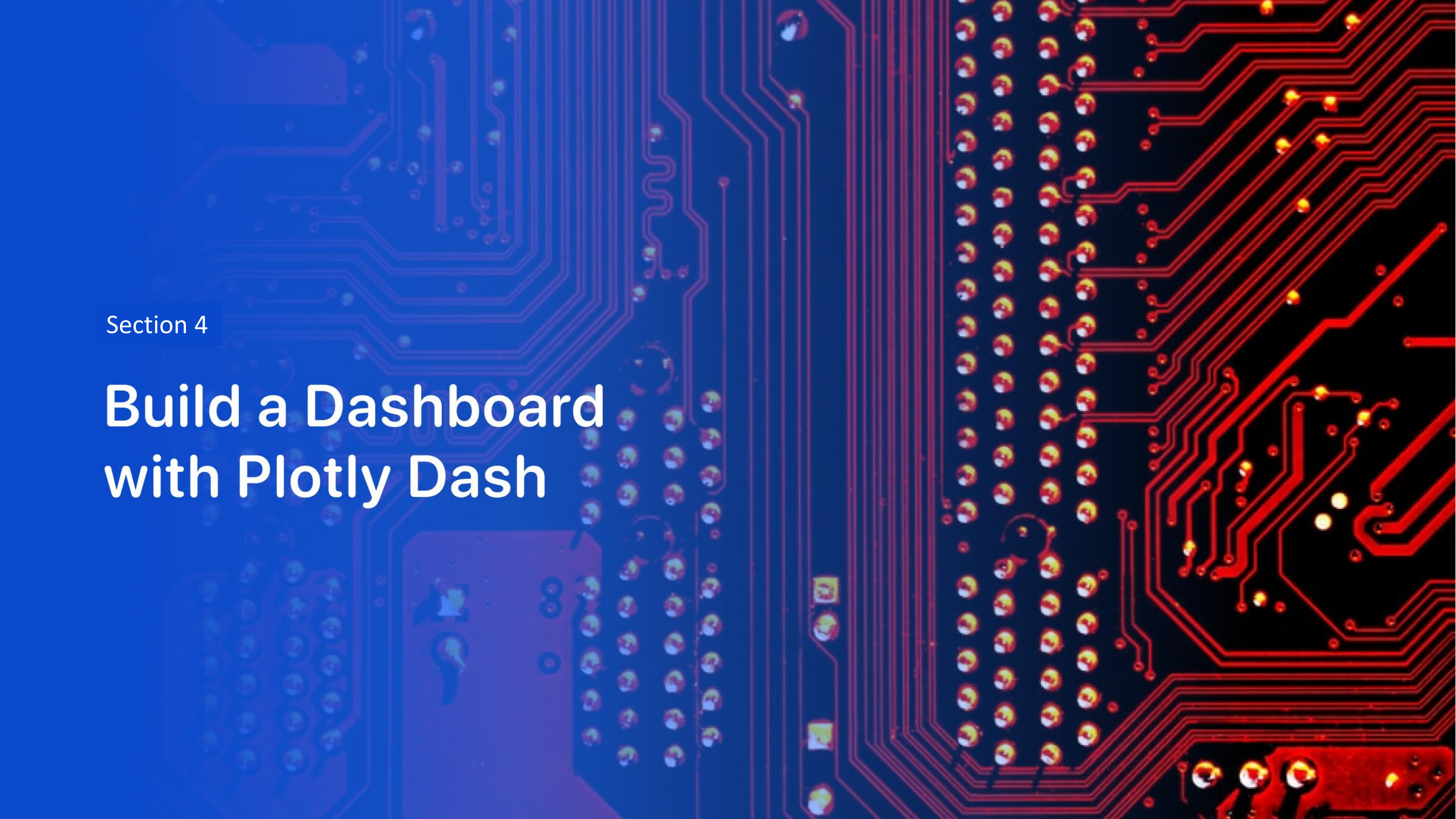


Take a closer look at the proximity of the launch site. We can observe that there are few large facilities near there.

<distance to the coastline>



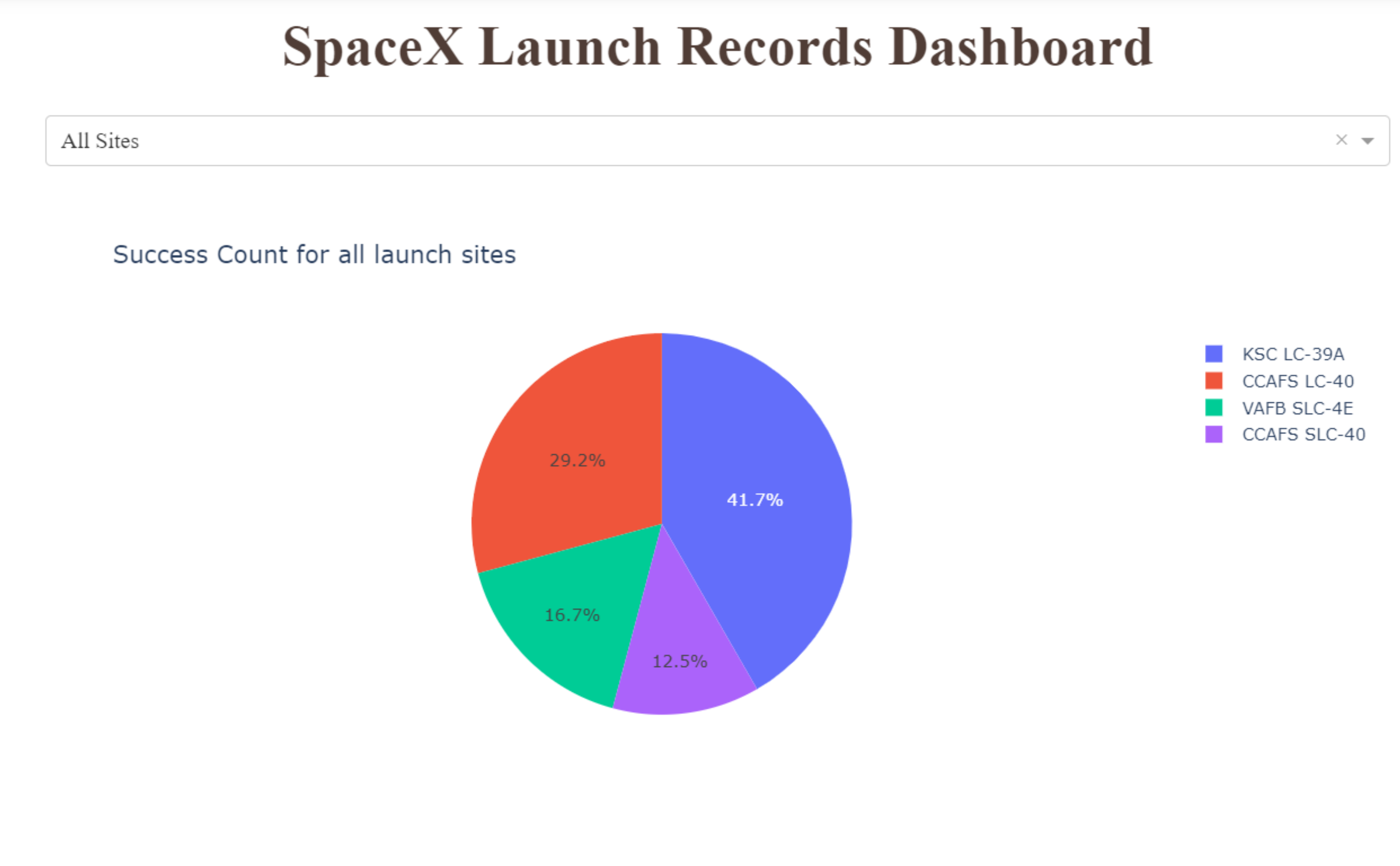
After calculation, the distance from the launch site to the closest coastline is only 0.85 kilogram.



Section 4

Build a Dashboard with Plotly Dash

< launch success count for all sites >

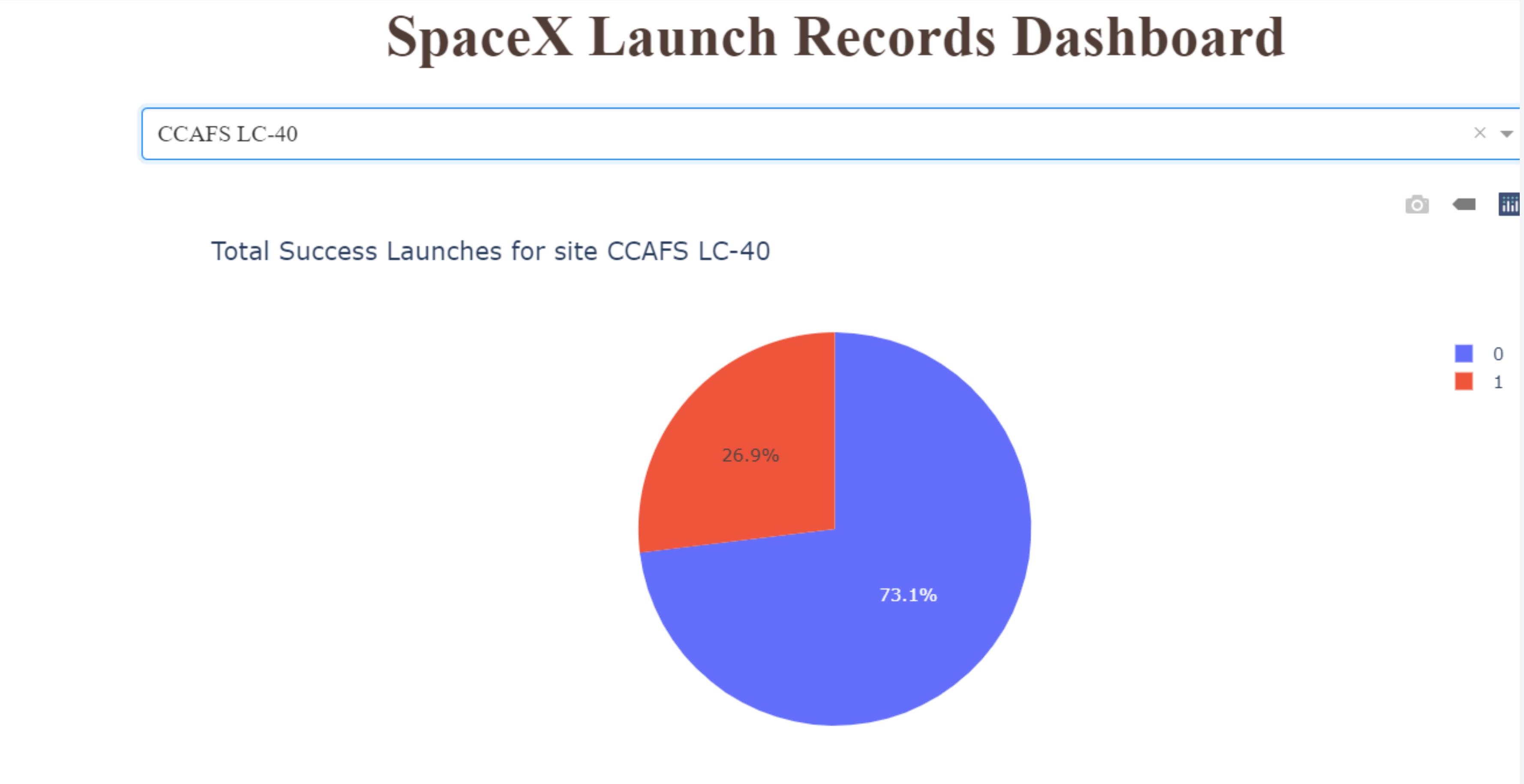


SpaceX Launch Records Dashboard

All Sites

The chart shows the success count of four launch sites. The site with highest success count is KSC LC-39A.

<launch site with highest success rate>



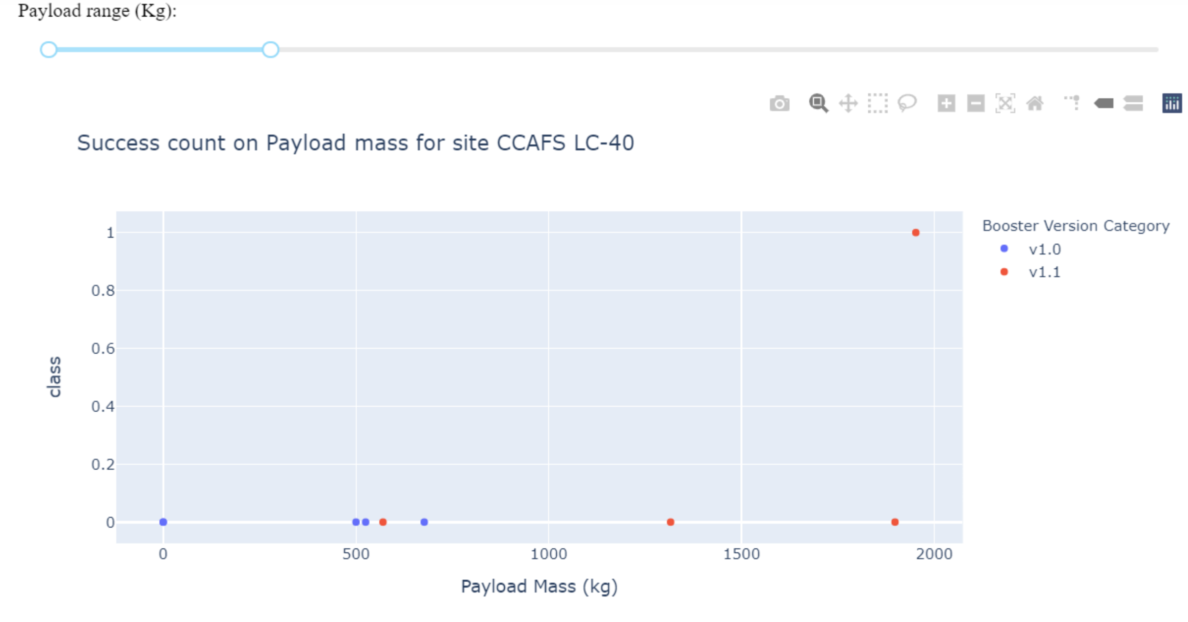
SpaceX Launch Records Dashboard

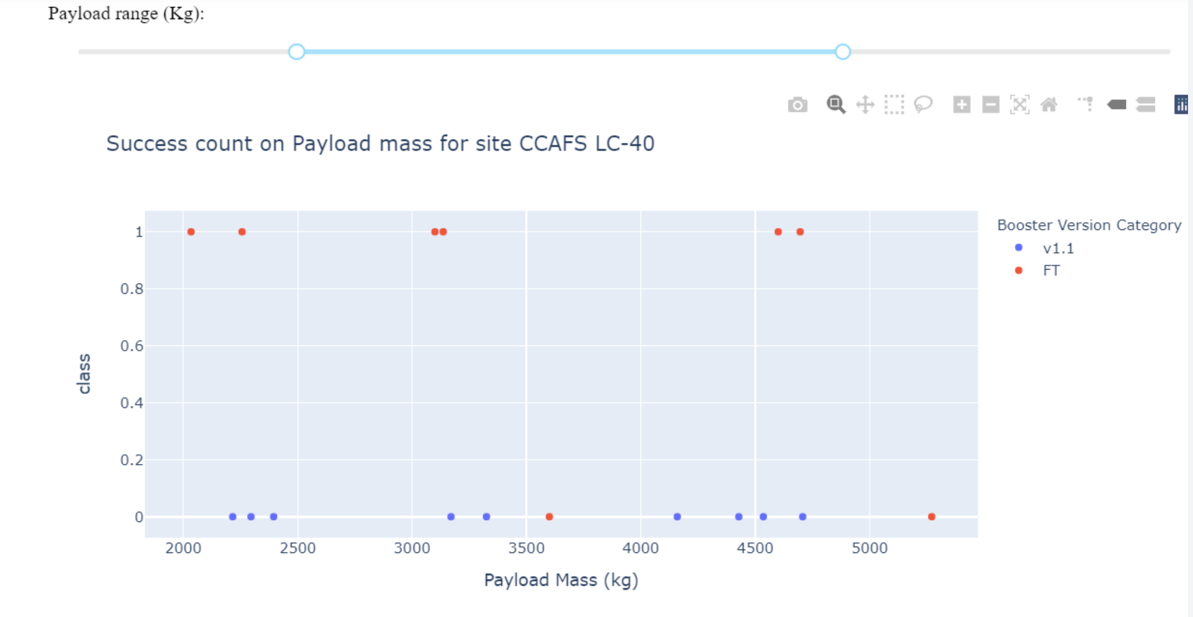
CCAFS LC-40

Total Success Launches for site CCAFS LC-40

The site with highest success rate is CCAFS LC-40 but the site with highest success count is KSC LC-39A.

<payload comparison>



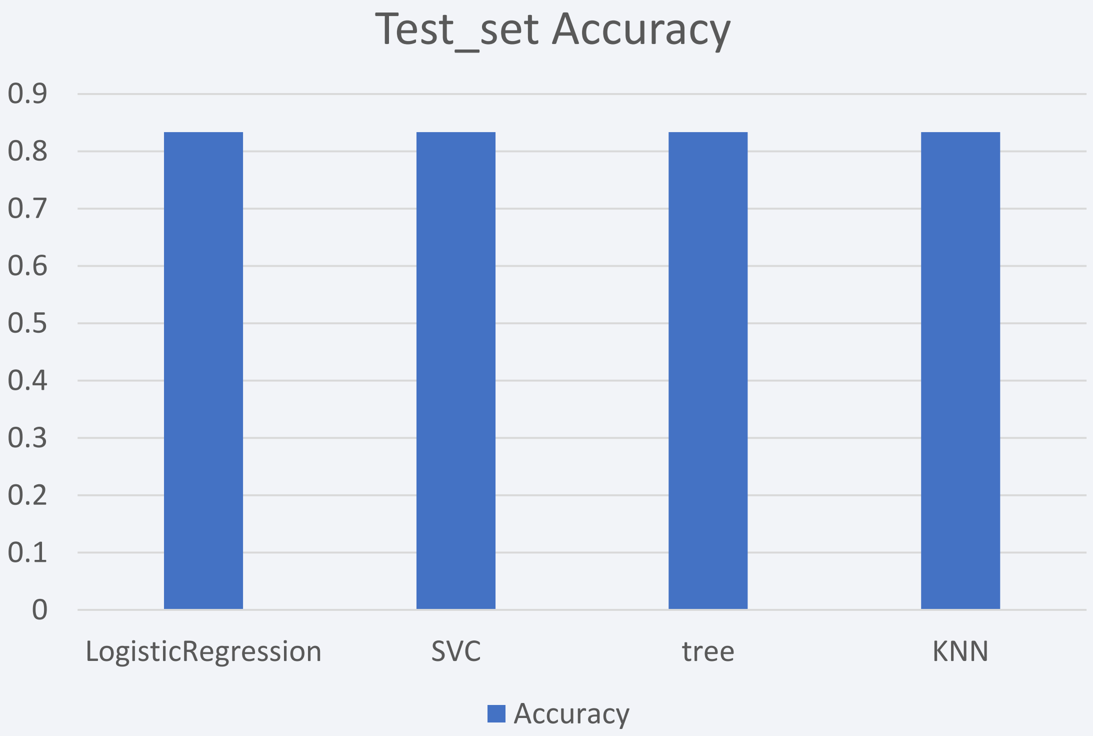


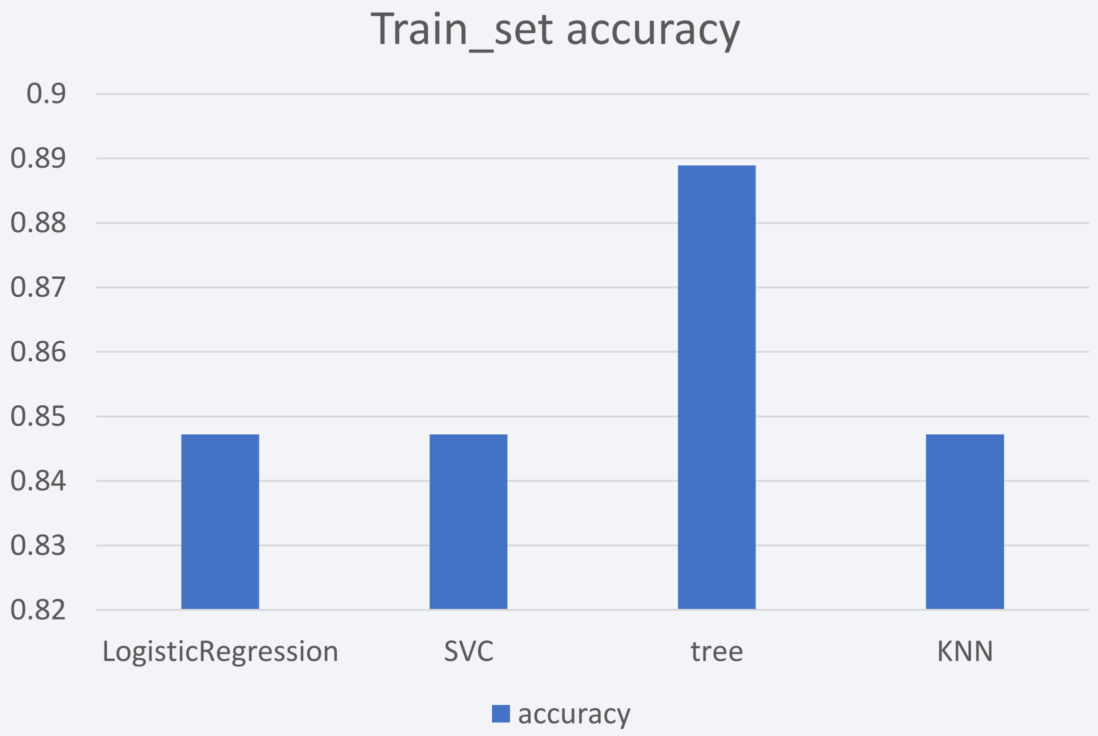
From the above two charts, we can know that the success rate is low for both booster. As the payload higher, the success rate for the FT booster raises dramatically.

Section 5

Predictive Analysis (Classification)

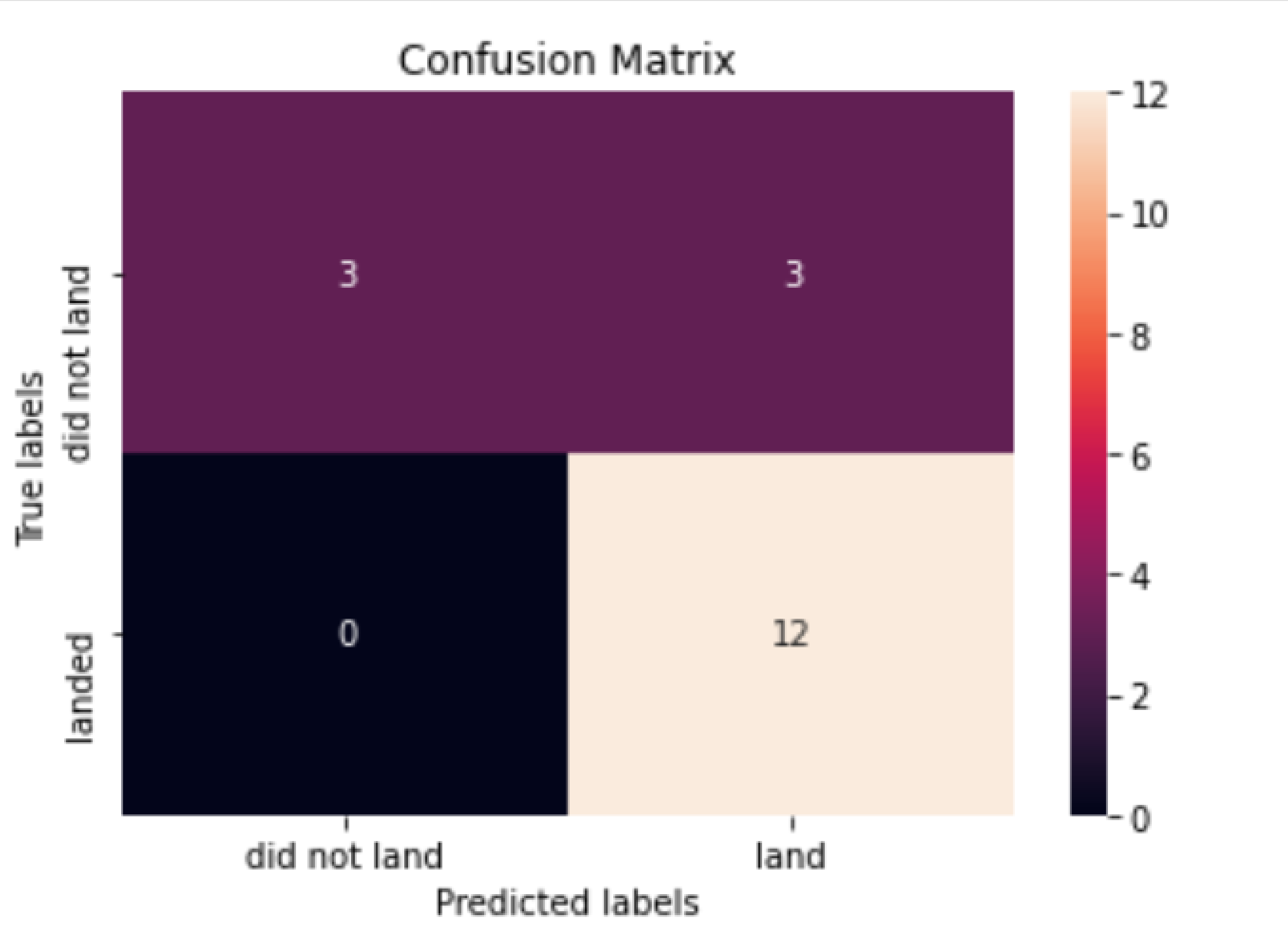
Classification Accuracy





With the chart, we can know that tree algorithum’s accuracy on train set is higher

Confusion Matrix



The chart can clearly show that the model’s accuracy is not bad, with 15 correct and 3 incorrect.

Conclusions

- Decision tree algorithum owns the highest accuracy among the four algorithums
- The success rate is low for both booster. As the payload higher, the success rate for the FT booster raises dramatically.
- SpaceX launch sites are all close to coastline.
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

Appendix

- In the project, I also used pandas to check my SQL results.
- The code is on my github.

Thank you!