IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shaw
2022/6/20

Outline

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

Executive Summary

In this project, we predict if the Falcon 9 first stage will land successfully. To achieve this goal, we collected data from SpaceX REST API, and conducted an ETL process to make data accessible and actionable. We explored data using both python and SQL, then visualized data features with a user interactive web dashboard. After data understanding, several machine learning models were introduced to solve the classification problem. Finally, a Decision Tree model with 85% accuracy rate was selected as the best model to predict if the Falcon 9 first stage will land successfully.

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. This can be used if an alternate company wants to bid against SpaceX for a rocket launch.

In this study, we want to find answer for questions like which launch sites have high success landing rate, what factors determine a successful first stage landing, and what machine learning model has the best performance in predicting the first stage landing.

Section 1

Methodology

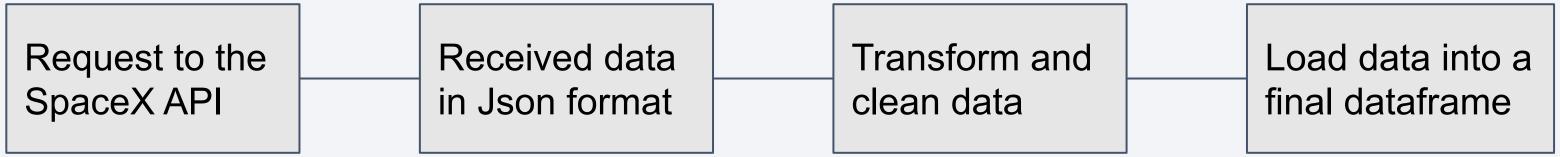
Methodology

Executive Summary

- Data collection methodology:
  - Extract from SpaceX REST API and webscripting
- Perform data wrangling
  - Impute missing value, modify data formats, encoding categorical data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build KNN, Decision Tree, Logistic Regression, and SVM classifier
  - tuning model with GridSearchCV, evaluate with accuracy, F-1, and jaccard score

Data Collection

- API Method:



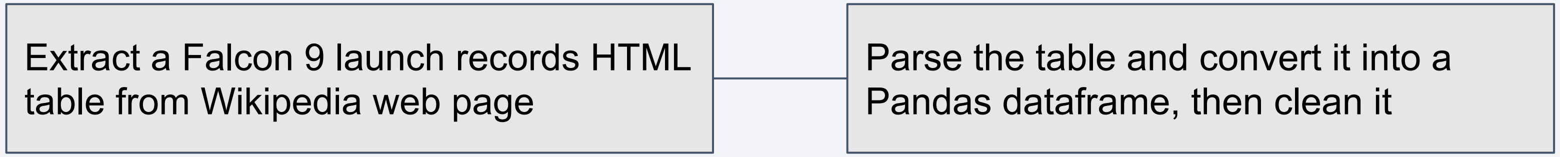
Request to the SpaceX API

Received data in Json format

Transform and clean data

Load data into a final dataframe

- Webscripting Method:



Extract a Falcon 9 launch records HTML table from Wikipedia web page

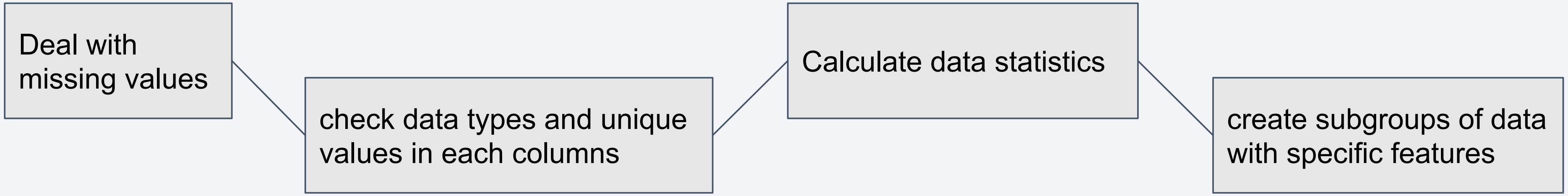
Parse the table and convert it into a Pandas dataframe, then clean it

Two methods are equally effective to the data collect process.

- https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/data_collection.ipynb
- https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/Webscripting.ipynb

Data Wrangling

With loaded data, we imputed missing values, checked data types and unique values, calculate data stats, create subgroups of data with specific features.



Deal with missing values

check data types and unique values in each columns

Calculate data statistics

create subgroups of data with specific features

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/Data_Wrangling.ipynb

EDA with Data Visualization

In order to understand variable relationships, scatter charts were utilized:

- visualize the relationship between flight number and launch site, we found that different launch sites have different success rates
- visualize the relationship between payload and launch site, we saw that it is obvious that payload mass positively impacts the success rate
- visualize the relationship between flight number and orbit type, we noticed that in some orbits the Success appears related to the number of flights
- visualize the relationship between payload and orbit type, we discovered that with heavy payloads, the positive landing rate is more for only some launch sites

To see the distribution of success rate for all orbit types, a bar chart was plotted, and we detected that four sites have a success rate of 1. Other sites are around 0.6, and the lowest one has only 0.5 success rate. Moreover, a line chart was plotted to visualize the launch success yearly trend, and we observed that the success rate since 2013 kept increasing till 2020.

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/EDA_with_Python.ipynb

EDA with SQL

We completed following tasks using SQL:

1. Display the names of the unique launch sites in the space mission
2. Display records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was achieved
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. 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.
10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/EDA_with_SQLite.ipynb

Build an Interactive Map with Folium

- To see all launch sites on a map, we added markers according to launch sites' map coordination, and set ratio around each location marker to display range circles, which makes the map visualization of each site easier to see.
- In addition, the success/failed launches for each site were marked within the range circle of each site in the map.
- What's more, we calculated the distances between a launch site to its proximities. A random coastline point near one of launch site was picked to approximate the real distance, and a line was drawn between a coastline point and a launch site to visualize the distance.

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/Launch_site_map_visualization.ipynb

Build a Dashboard with Plotly Dash

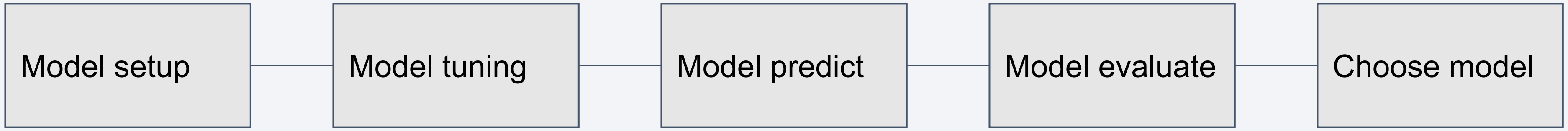
Dashboard is consisting of two main visualization techniques, one is pie chart and another is scatter chart. User can interact with a dropdown list and a range slider to see different visualizations. By choosing different sites in the dropdown list, the pie chart and scatter chart will display different contexts according to the option specified. By adjusting the minimum and maximum value in the ranger slider, scatter will show variable relationship only within that controlled range.

- Pie chart in this dashboard shows the total success launches for all sites, or total success and failure launches for a specific site if dropdown list is specified.
- Scatter chart in this dashboard demonstrates the relationship between payload mass (kg) and number of success launches for all sites or a specific site.

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/plotly_dashboard.py

Predictive Analysis (Classification)

KNN, Decision Tree, Logistic Regression, and SVM classifier were constructed in this part, model building process is illustrated as below:



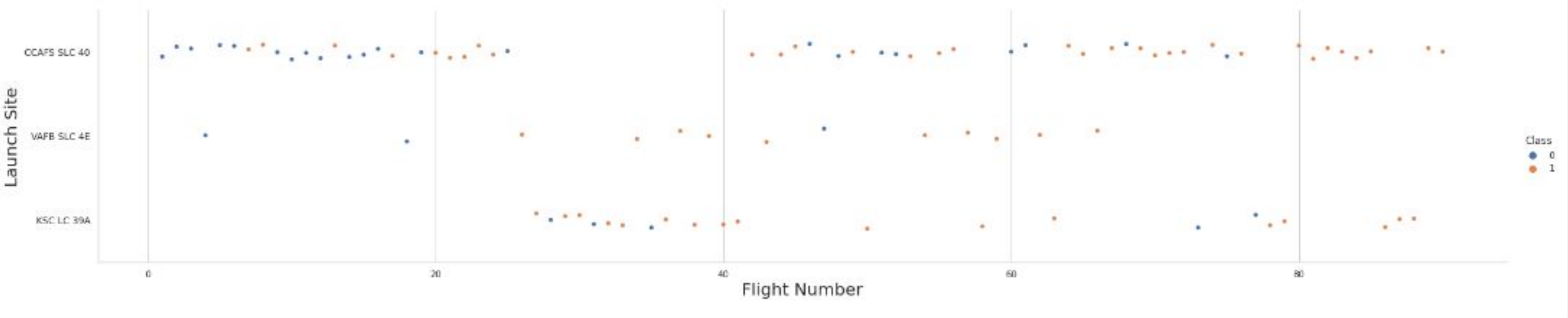
In the model tuning stage, GridSearchCV method was used. In the Model evaluation stage, we applied four indicators: accuracy, F-1, log loss, and jaccard score. A confusion matrix was also established to help understand model performance.

https://github.com/zw2791/Space-X-Falcon-9-First-Stage-Landing-Prediction/blob/main/ML_Predictive_Analysis.ipynb

Section 2

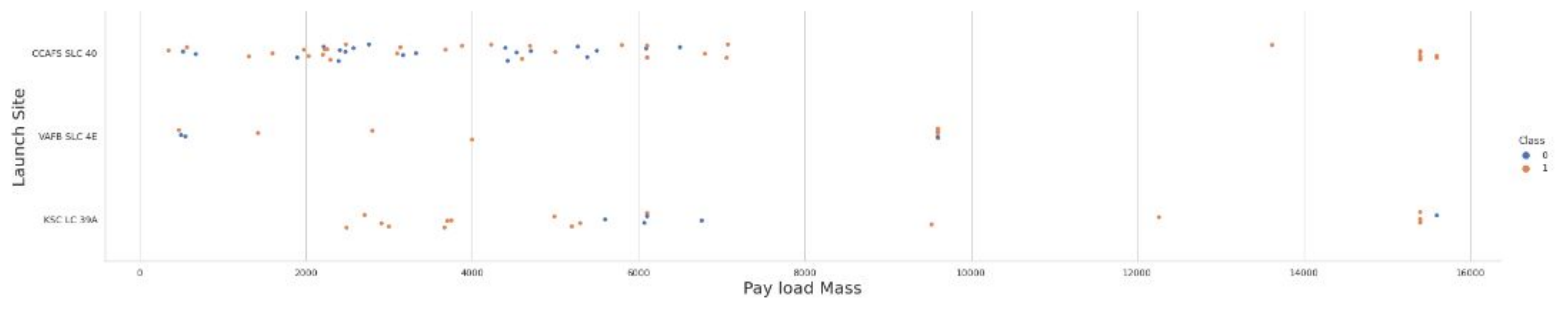
Insights drawn from EDA

Flight Number vs. Launch Site



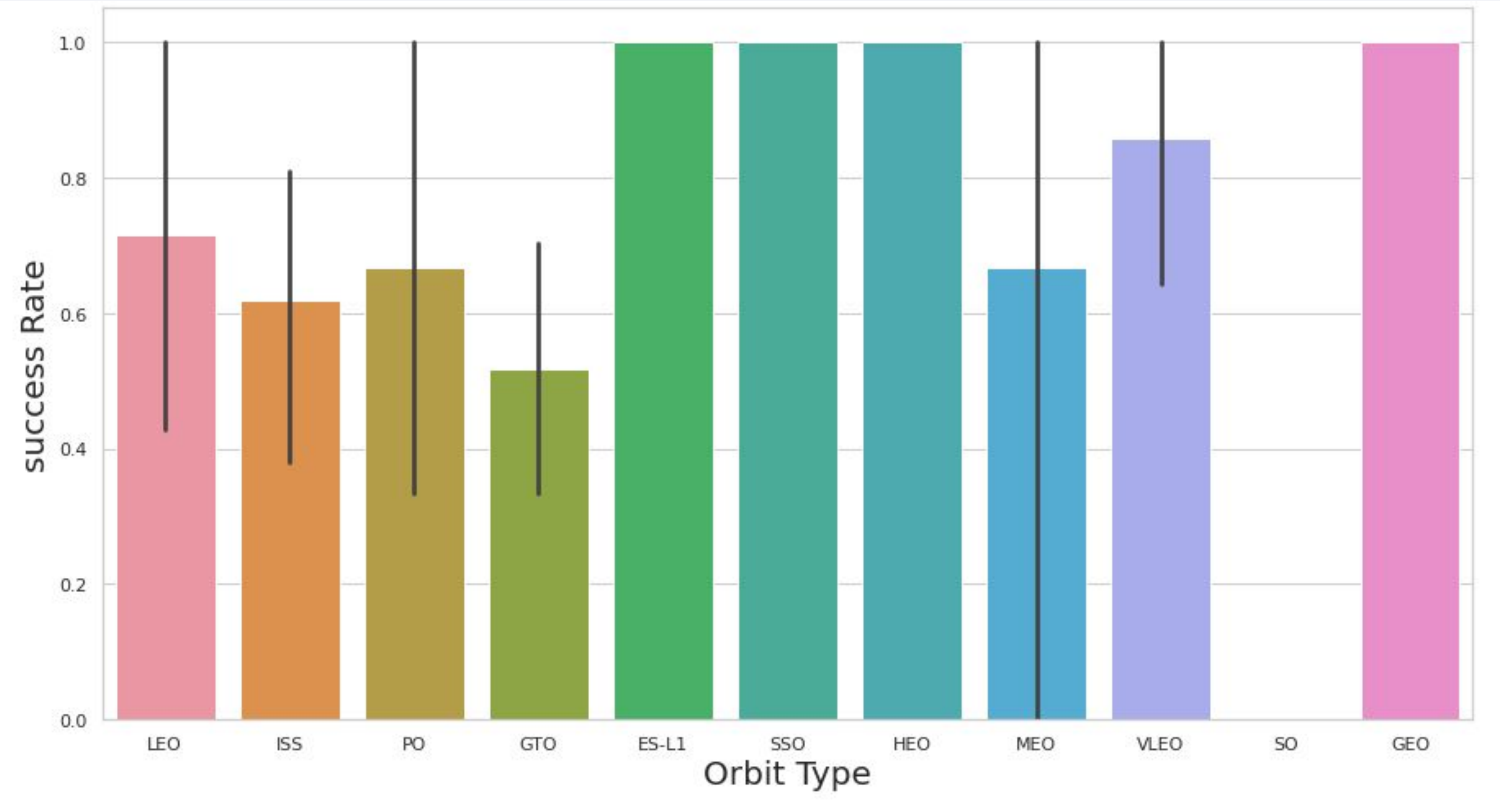
We detect that success rates are obviously different in different locations, KSC LC 39A and VAFB SLC 4E are better launch sites based on the plot. Flight number can affect the success rate but the relationship is not very strong.

Payload vs. Launch Site



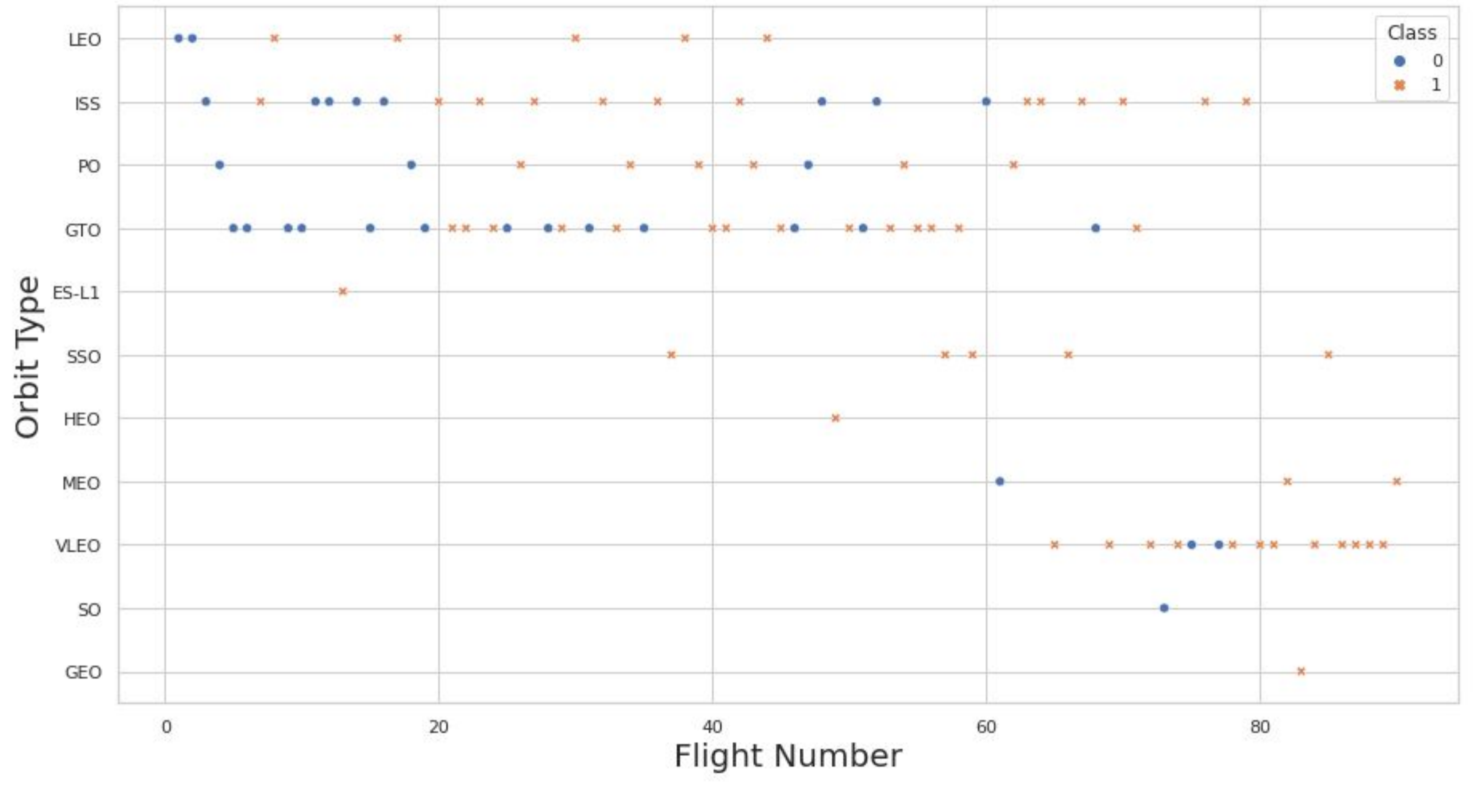
It is obvious that payload mass positively impacts the success rate. We can also find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type



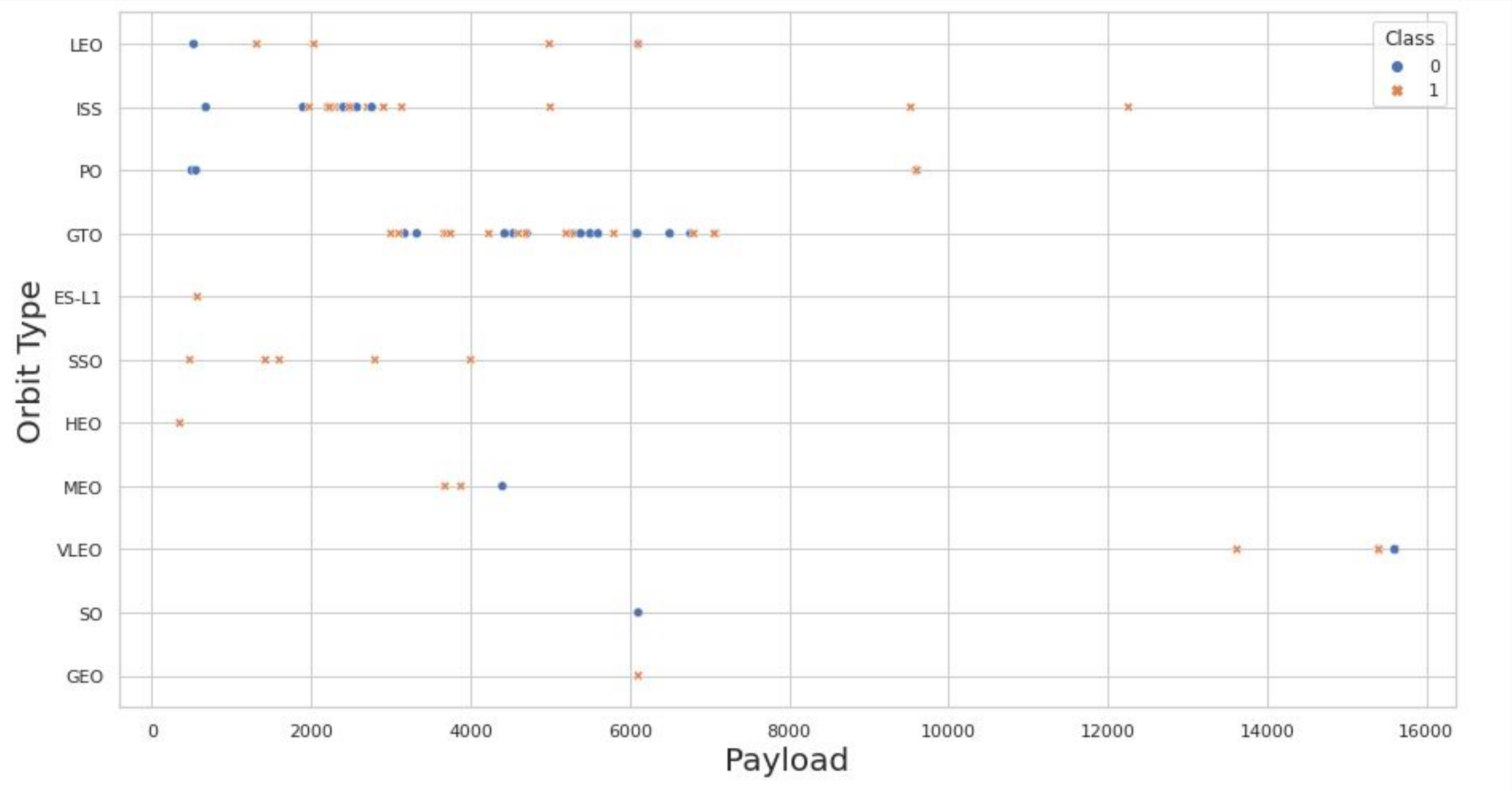
ES-L1, SSO, HEO, and GEO have a success rate of 1. Other sites are around 0.6, the lowest one is about 0.5 success rate at GTO.

Flight Number vs. Orbit Type



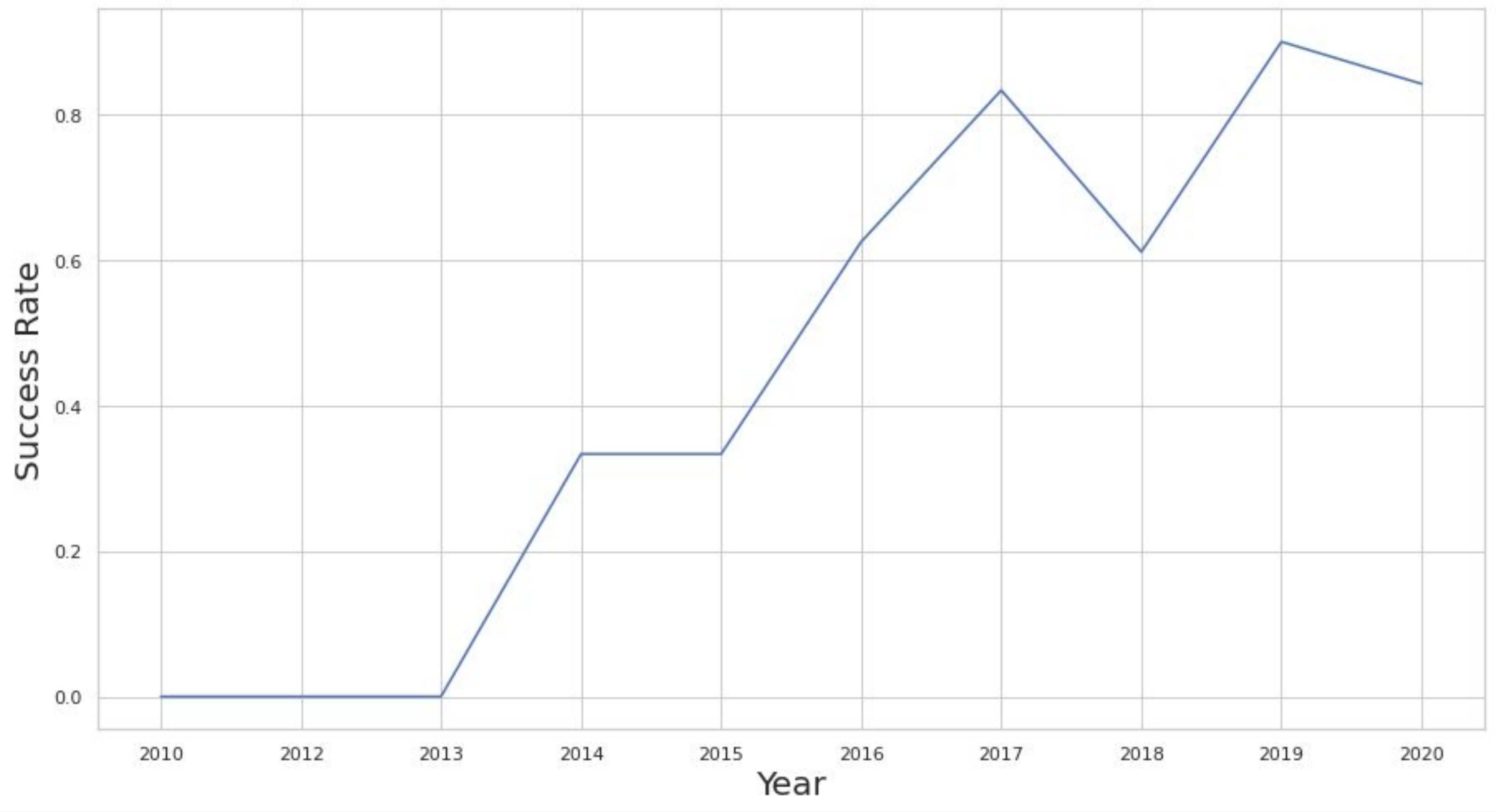
We can see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit. Moreover, orbit types in the upper side of our plot usually have less than 60 flights, however, orbits types in the bottom side always have more than 40 flights.

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

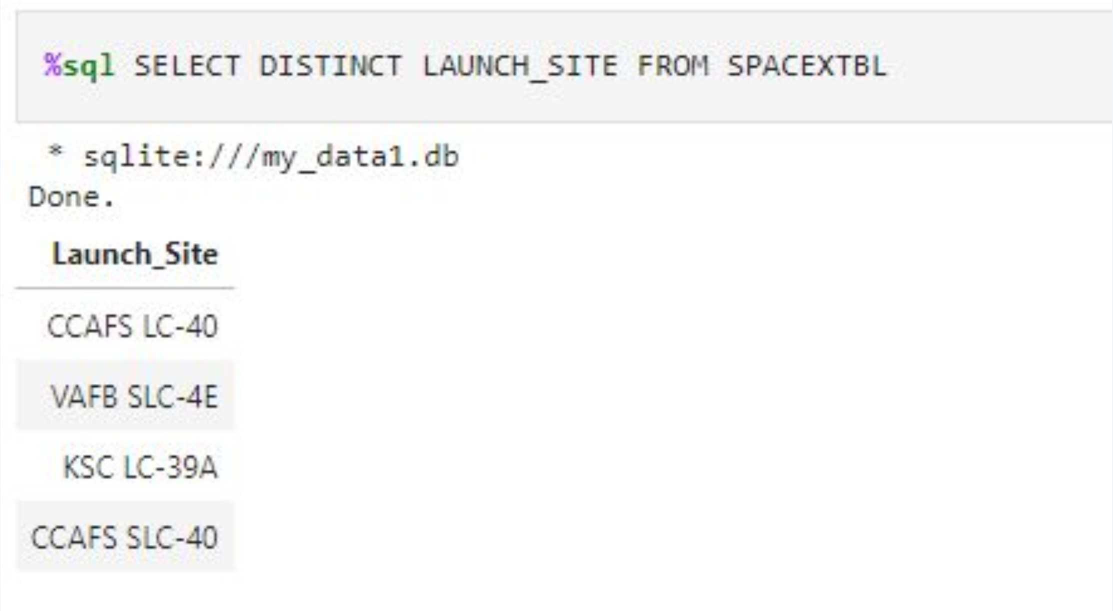


We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

4 unique launch sites are obtained from sql querying, they are:

1. CCAFS LC-40
2. VAFB SLC-4E
3. KSC LC-39A
4. CCAFS SLC-40

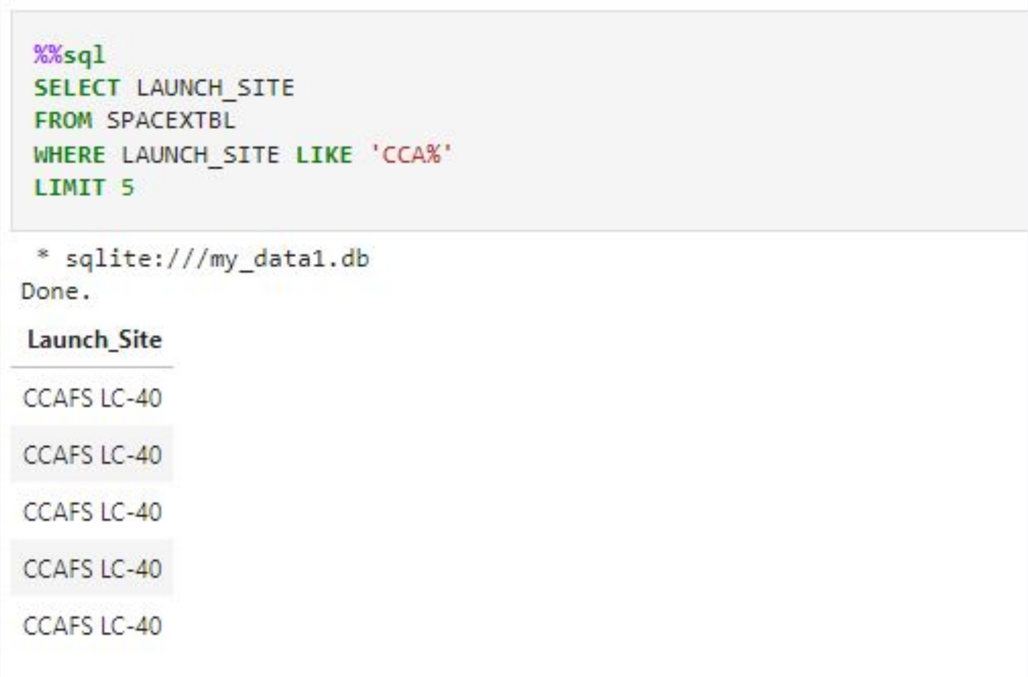


```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

```
 * 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'

We obtain 5 CCAFS LC-40 using the SQL query below



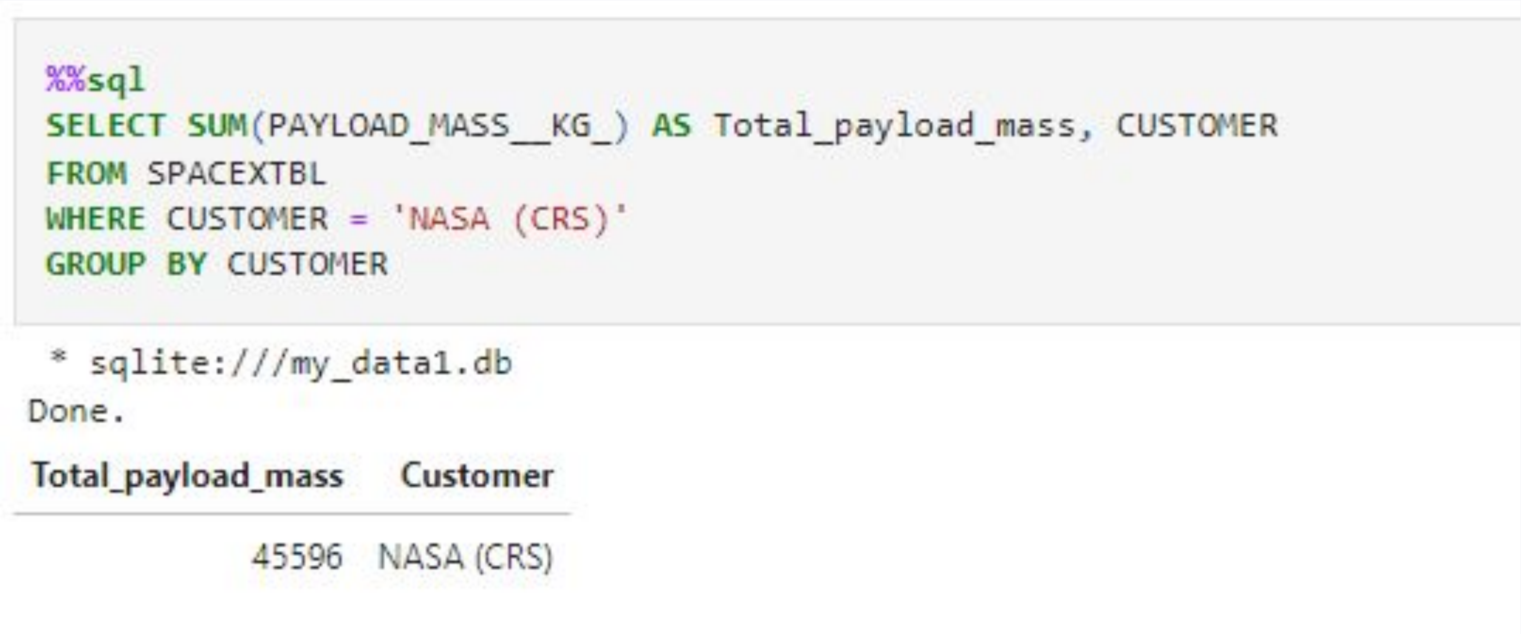
```
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

We calculate the total payload mass carried by boosters launched by NASA (CRS) to be 45596



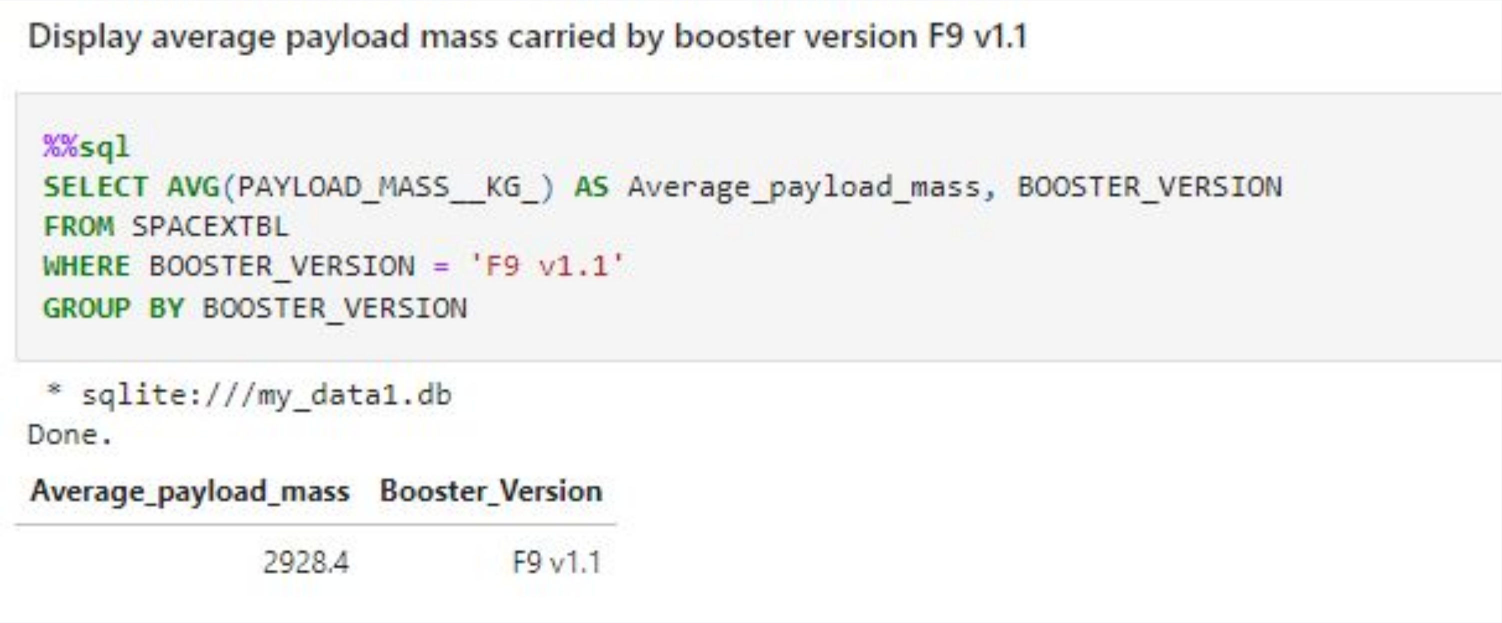
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_payload_mass, CUSTOMER
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)'
GROUP BY CUSTOMER
```

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

Total_payload_mass	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

Average payload mass calculated is 2928.4 for F9 V1.1



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

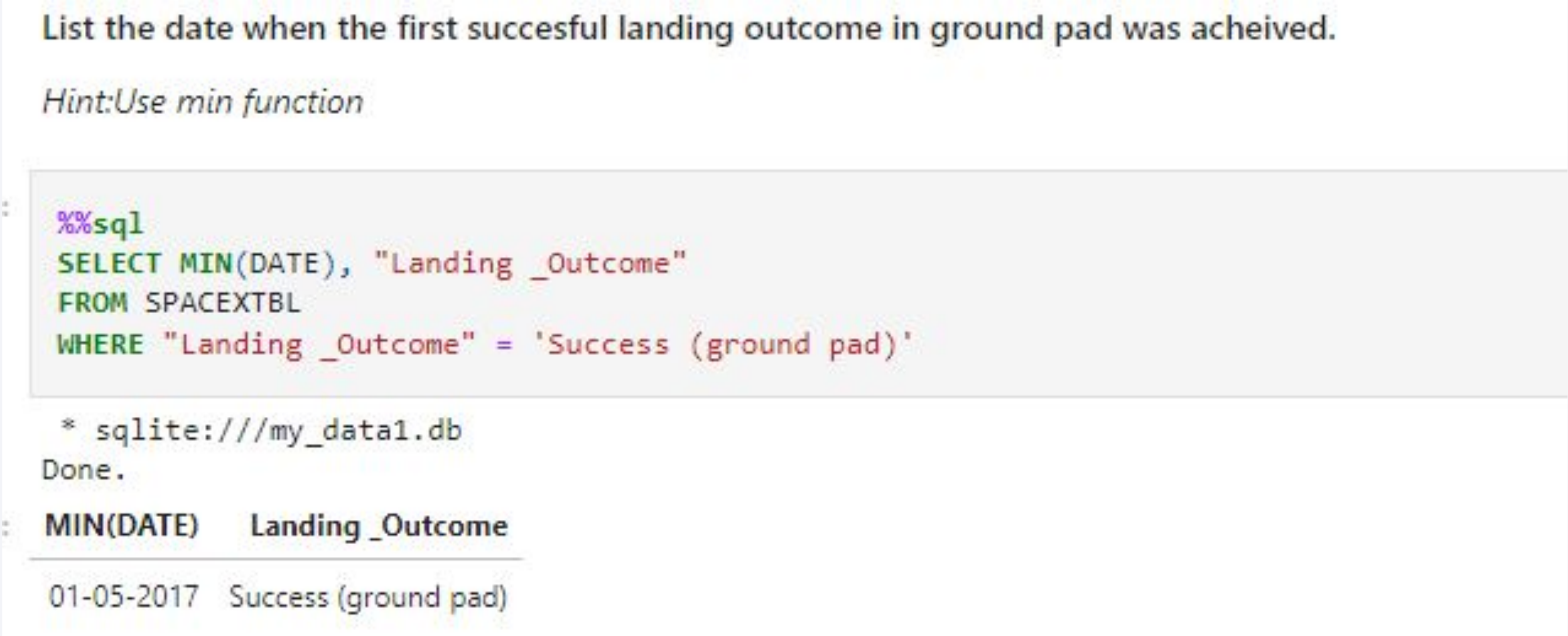
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_payload_mass, BOOSTER_VERSION
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
GROUP BY BOOSTER_VERSION
```

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

Average_payload_mass	Booster_Version
2928.4	F9 v1.1

First Successful Ground Landing Date

We see the first successful grounding landing date is 01/02/2017



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

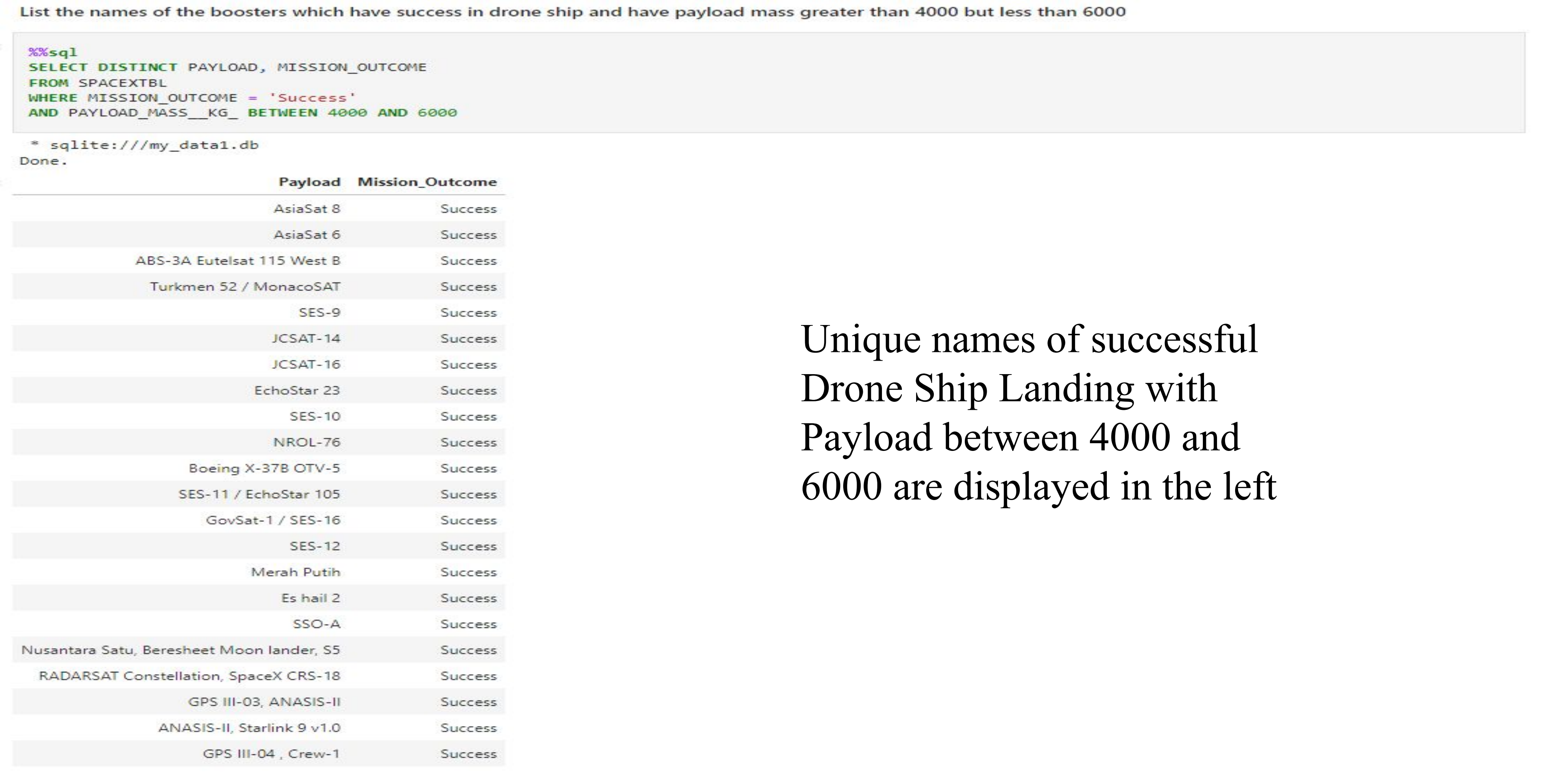
Hint:Use min function

```
%%sql
SELECT MIN(DATE), "Landing _Outcome"
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (ground pad)'
```

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

MIN(DATE)	Landing _Outcome
01-05-2017	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 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 DISTINCT PAYLOAD, MISSION_OUTCOME
FROM SPACEXTBL
WHERE MISSION_OUTCOME = 'Success'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

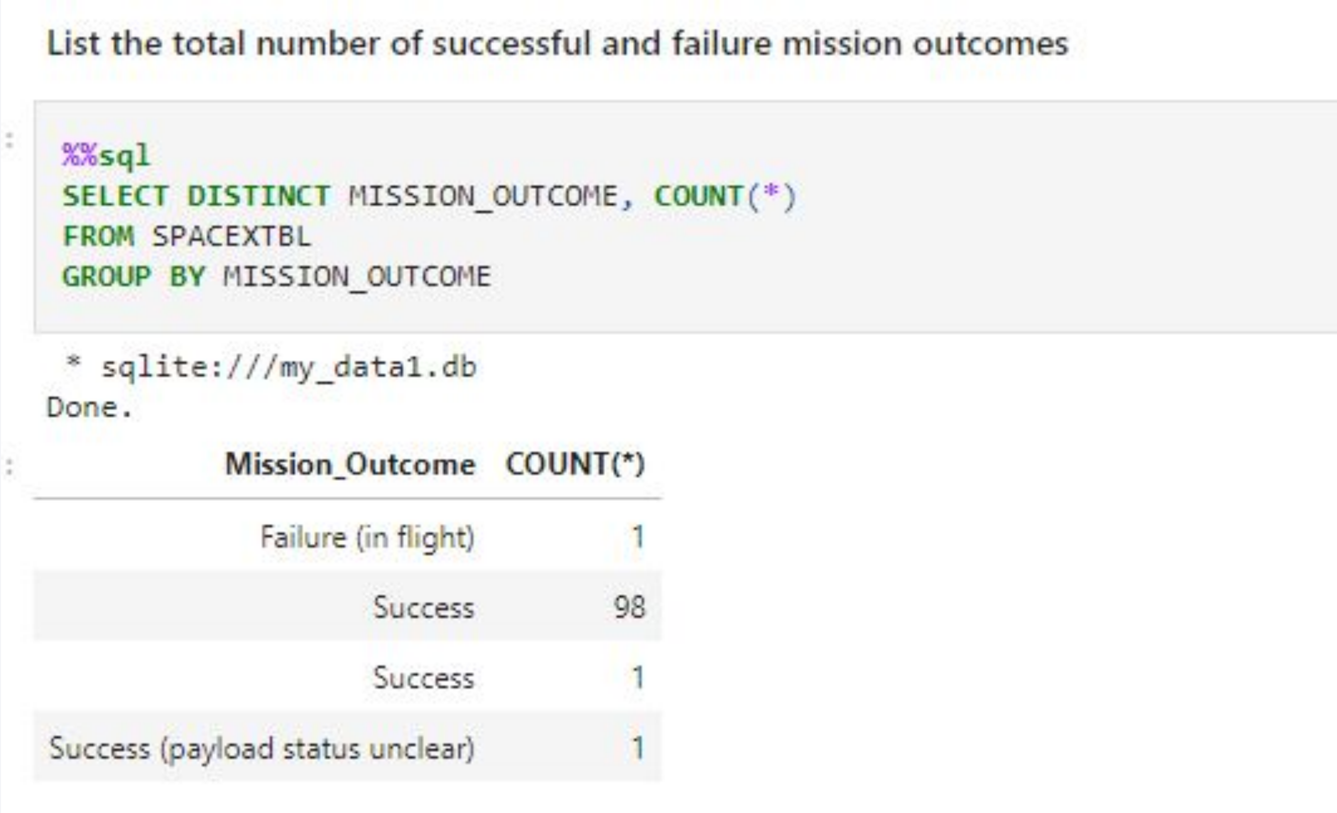
```
 * sqlite:///my_data1.db
Done.
```

Payload	Mission_Outcome
AsiaSat 8	Success
AsiaSat 6	Success
ABS-3A Eutelsat 115 West B	Success
Turkmen 52 / MonacoSAT	Success
SES-9	Success
JCSAT-14	Success
JCSAT-16	Success
EchoStar 23	Success
SES-10	Success
NROL-76	Success
Boeing X-37B OTV-5	Success
SES-11 / EchoStar 105	Success
GovSat-1 / SES-16	Success
SES-12	Success
Merah Putih	Success
Es hail 2	Success
SSO-A	Success
Nusantara Satu, Beresheet Moon lander, S5	Success
RADARSAT Constellation, SpaceX CRS-18	Success
GPS III-03, ANASIS-II	Success
ANASIS-II, Starlink 9 v1.0	Success
GPS III-04 , Crew-1	Success

Unique names of successful Drone Ship Landing with Payload between 4000 and 6000 are displayed in the left

Total Number of Successful and Failure Mission Outcomes

We count the number of different mission outcomes



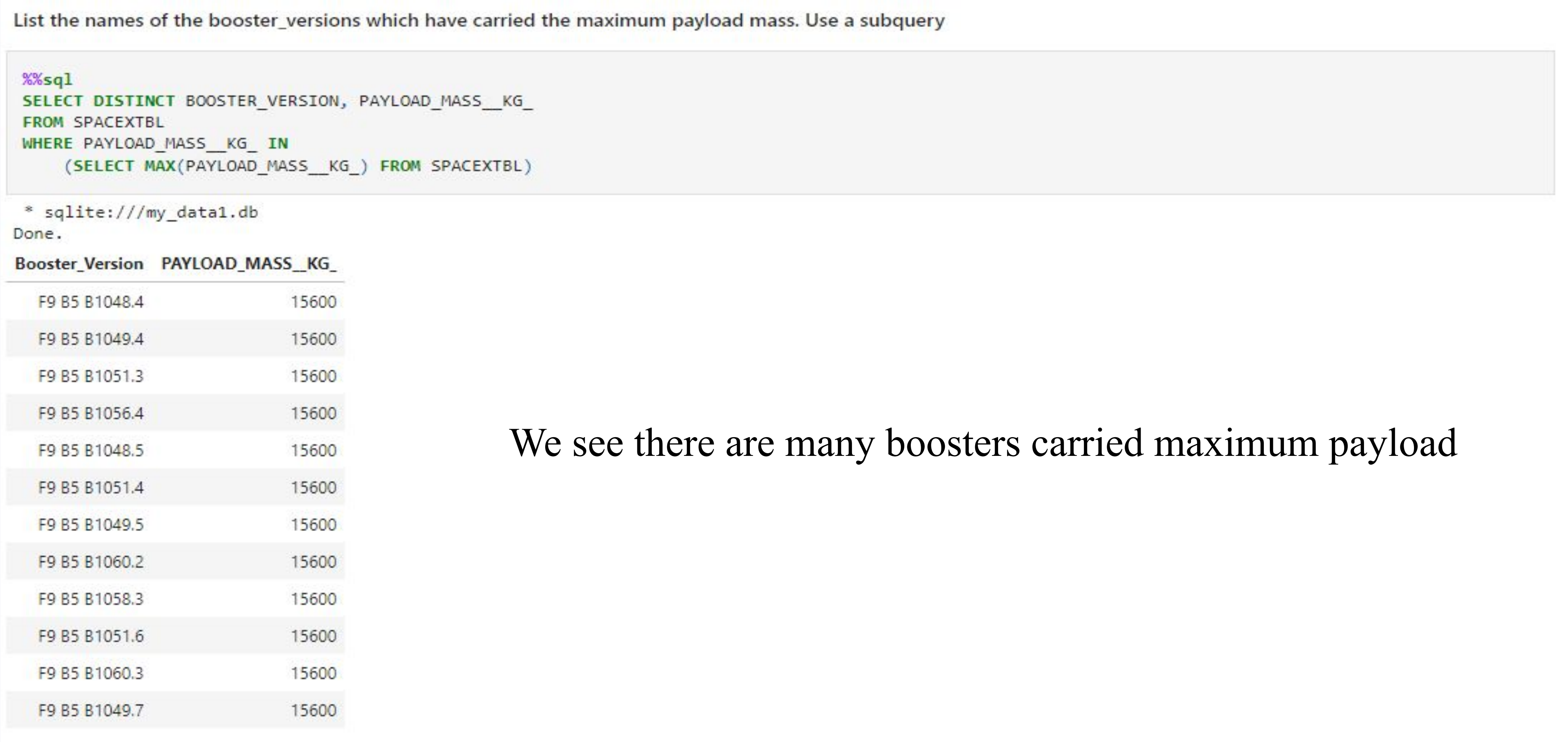
List the total number of successful and failure mission outcomes

```
%%sql
SELECT DISTINCT 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



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

```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ IN
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

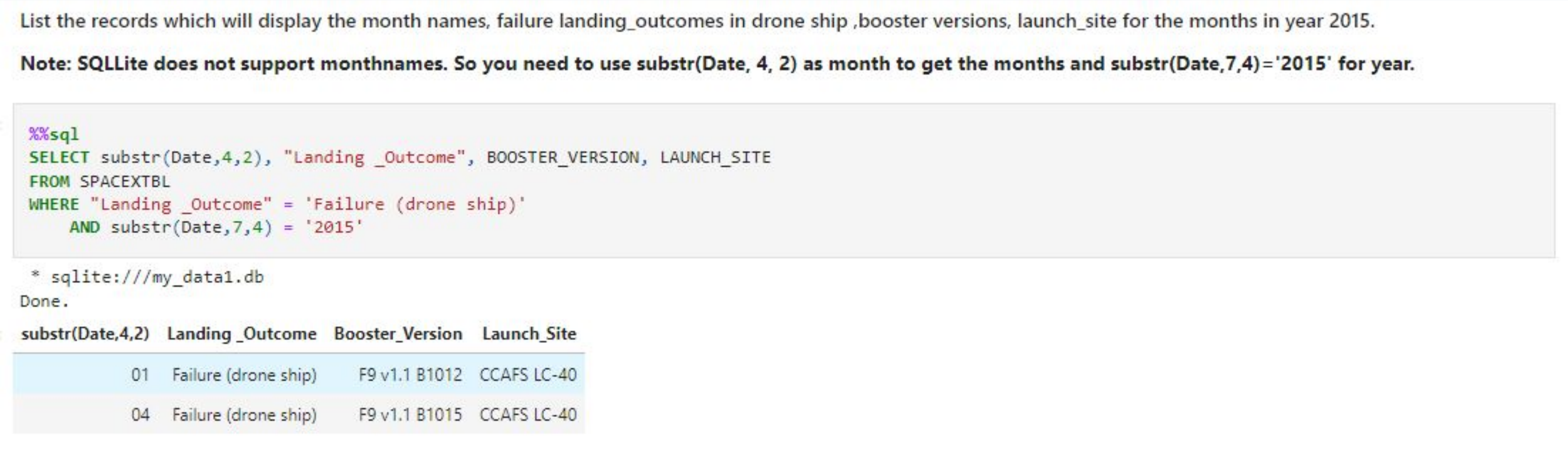
```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

We see there are many boosters carried maximum payload

2015 Launch Records

We find there are two failure landing outcomes 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, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

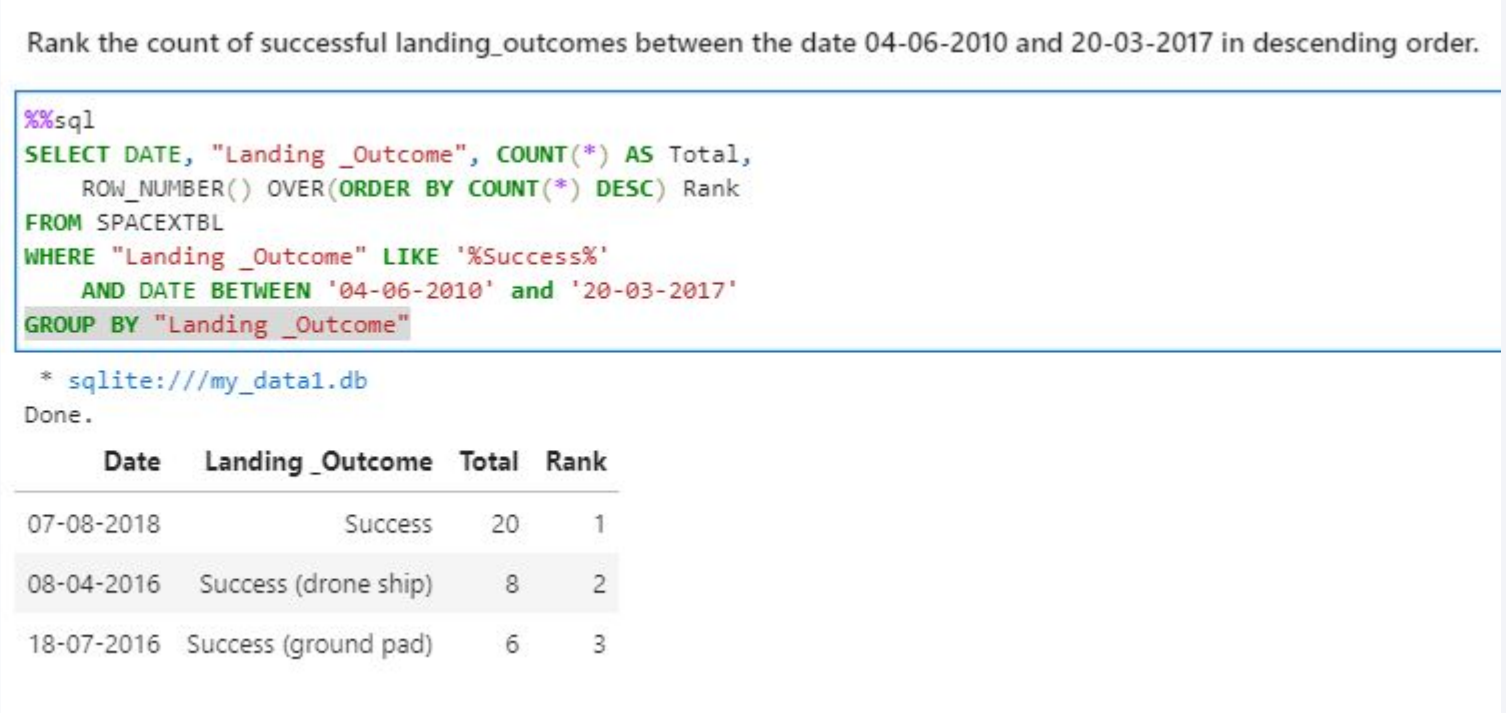
```
%%sql
SELECT substr(Date,4,2), "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Failure (drone ship)'
    AND substr(Date,7,4) = '2015'
```

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

substr(Date,4,2)	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

Success in 7/8/2018 is the most frequent landing outcome



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
SELECT DATE, "Landing _Outcome", COUNT(*) AS Total,
    ROW_NUMBER() OVER(ORDER BY COUNT(*) DESC) Rank
FROM SPACEXTBL
WHERE "Landing _Outcome" LIKE '%Success%'
    AND DATE BETWEEN '04-06-2010' and '20-03-2017'
GROUP BY "Landing _Outcome"
```

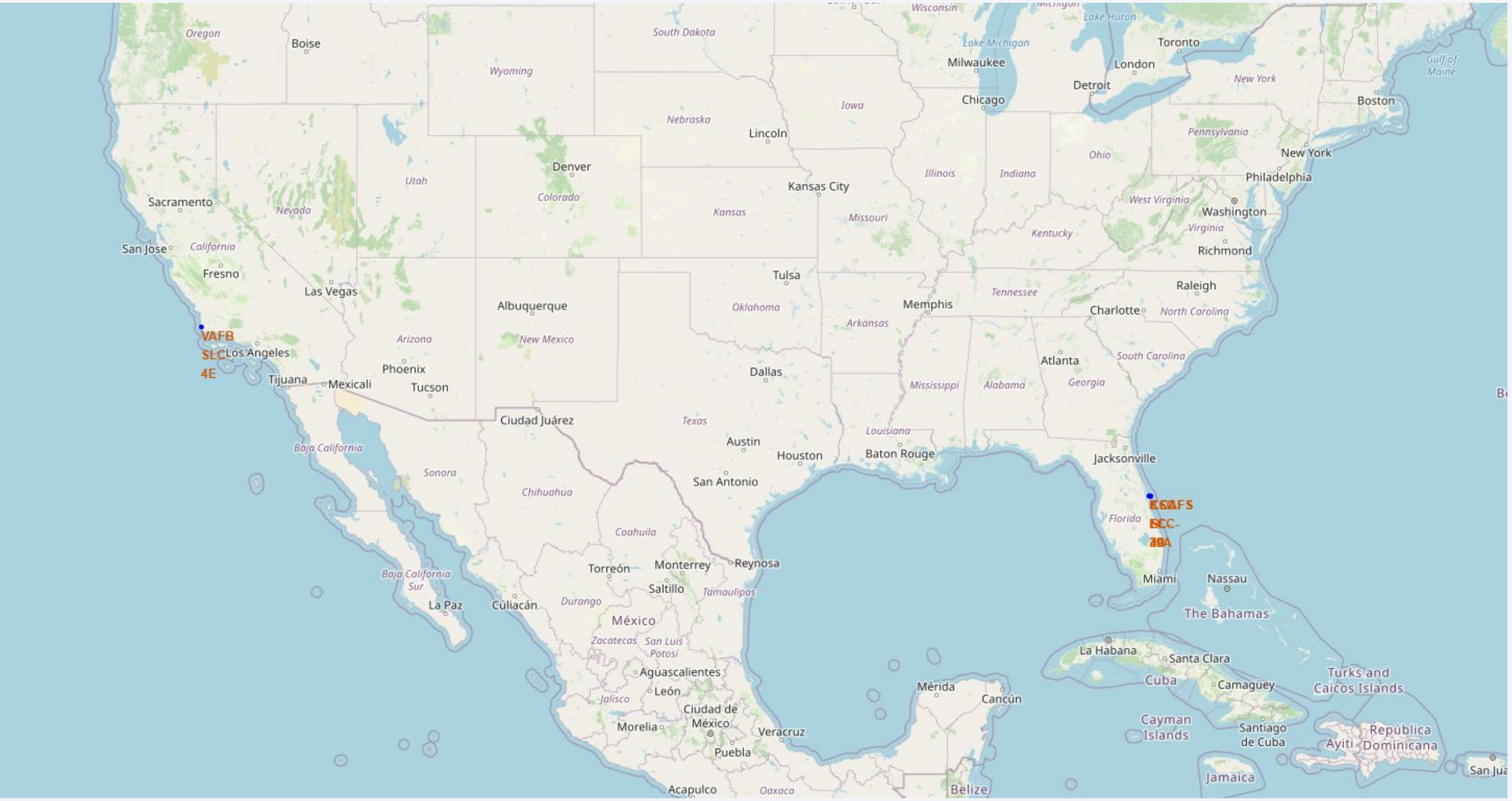
```
 * sqlite:///my_data1.db
Done.
```

Date	Landing _Outcome	Total	Rank
07-08-2018	Success	20	1
08-04-2016	Success (drone ship)	8	2
18-07-2016	Success (ground pad)	6	3

Section 3

Launch Sites Proximities Analysis

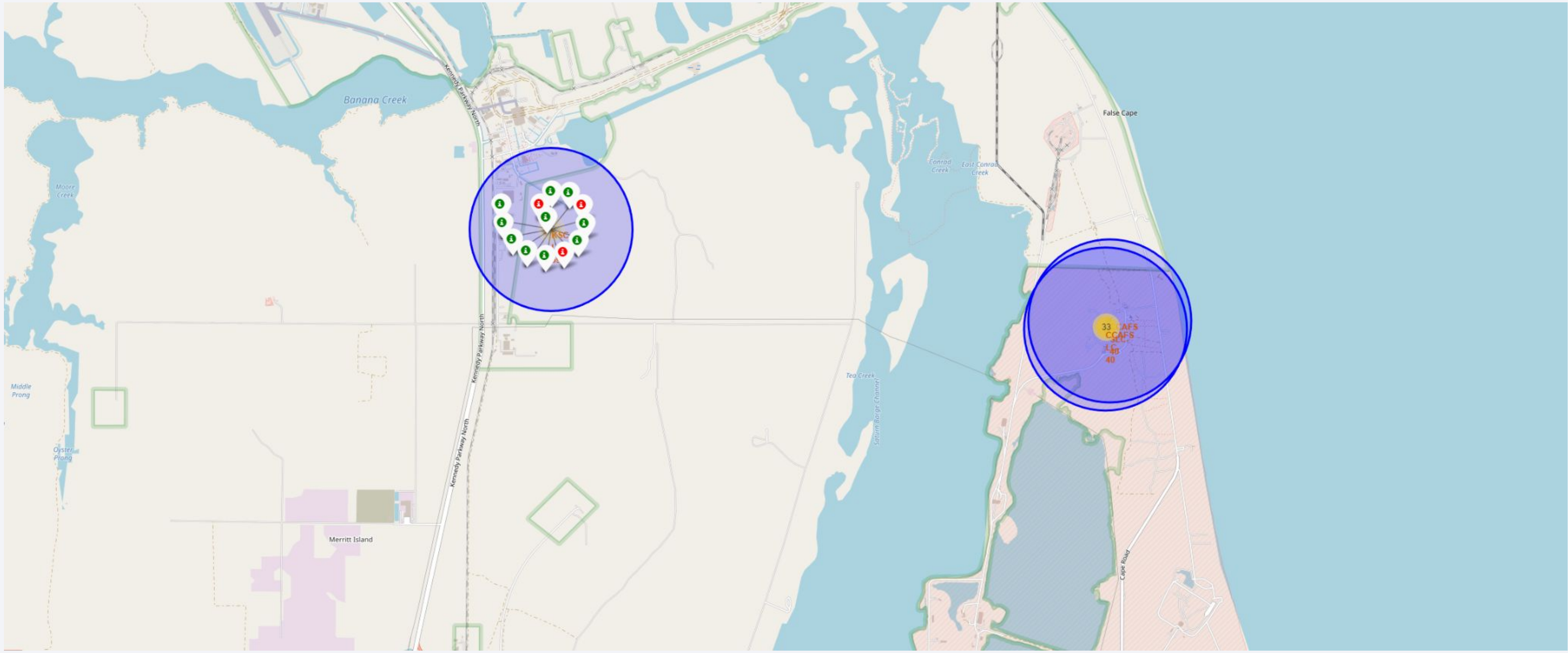
Launch Sites on the Map



We can see all the launch sites are distributed in two locations only

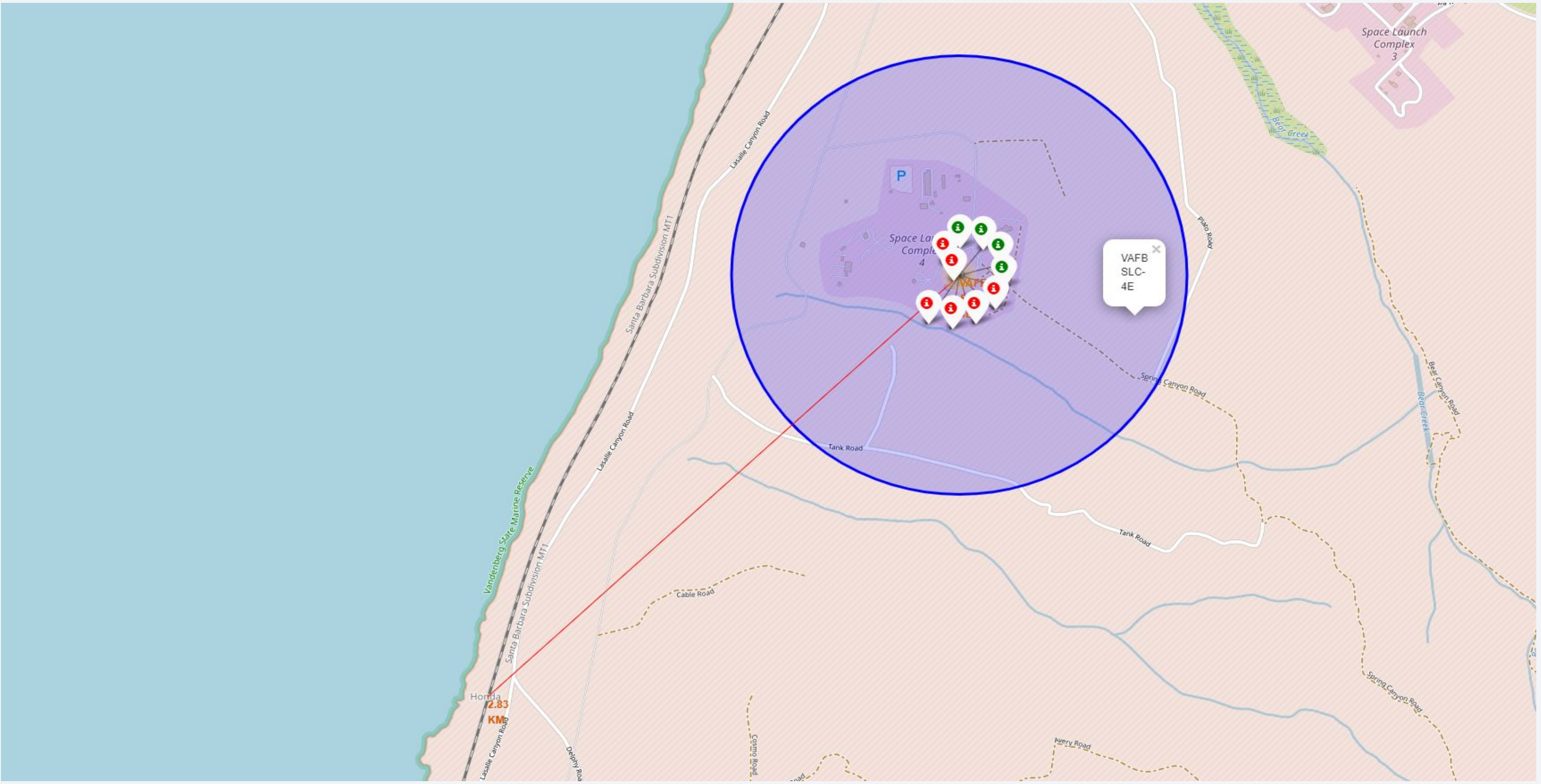
Succeeded / Failed Launches in each site

Succeeded launches are marked as green and Failed launches are marked as red



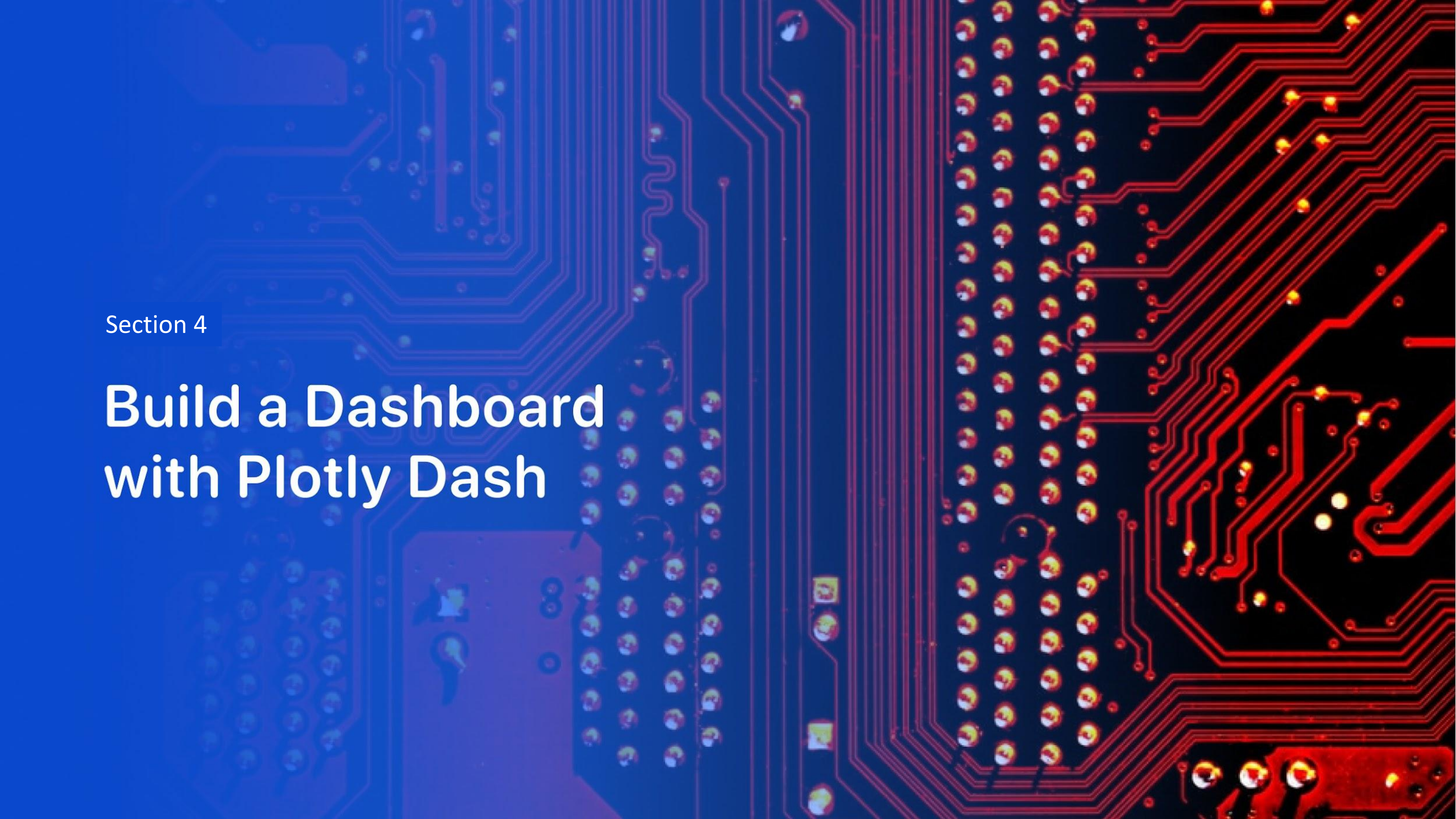
Calculate the distances between a launch site to its proximities

A line is drawn between a launch site to the selected coastline point to show the approximate distance

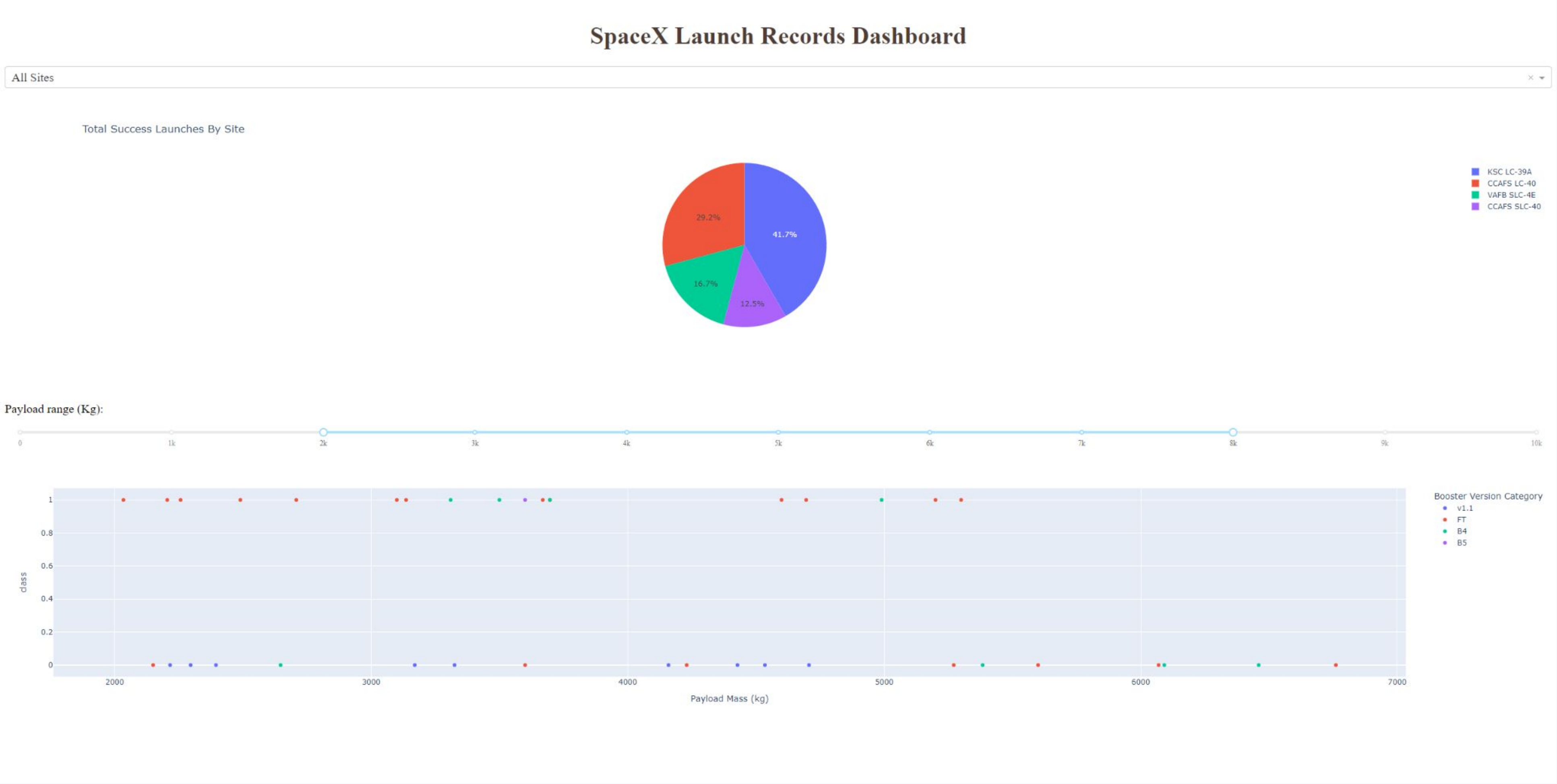


Section 4

Build a Dashboard with Plotly Dash



Dashboard



SpaceX Launch Records Dashboard

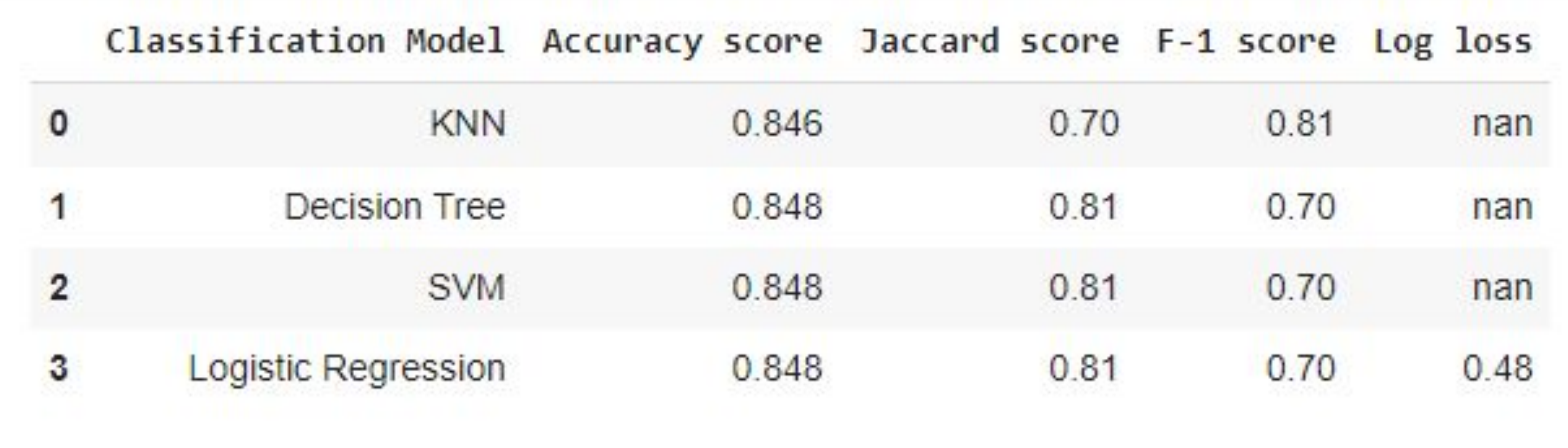
All Sites

Section 5

Predictive Analysis (Classification)

Classification Accuracy

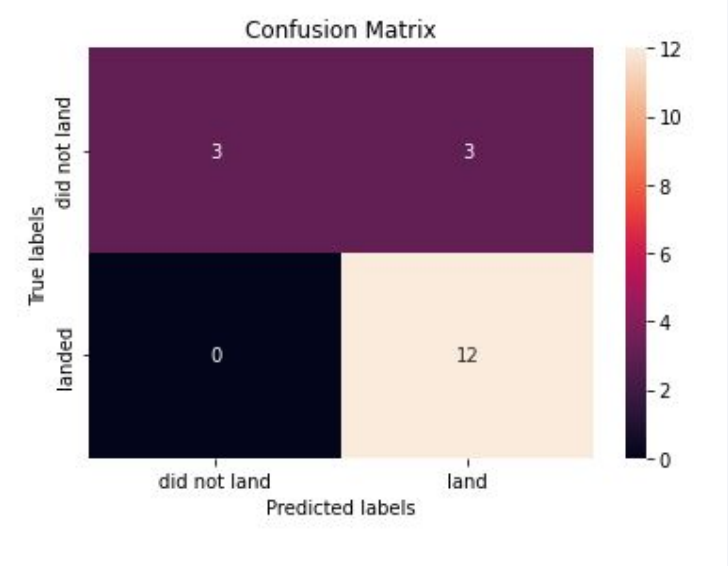
According to the table in below, KNN may have the worst performance. Other three models have similar accuracy, jaccard, and F-1 score. I think this is because the classification issue is not enough complex to make a real difference among models, or we may need a larger dataset.



	Classification Model	Accuracy score	Jaccard score	F-1 score	Log loss
0	KNN	0.846	0.70	0.81	nan
1	Decision Tree	0.848	0.81	0.70	nan
2	SVM	0.848	0.81	0.70	nan
3	Logistic Regression	0.848	0.81	0.70	0.48

Confusion Matrix

I choose Decision Tree Model as the best model since Decision Tree model is excellent at dealing with potential large datasets.



The confusion matrix for Decision Tree model is shown in the left side, we can see that Decision Tree Model does not do very well in predicting False positive, but in other aspects it is fine.

Conclusions

- ES-L1, SSO, HEO, and GEO have a success rate of 1. Other sites are around 0.6, the lowest one is about 0.5 success rate at GTO
- Payload mass positively impacts the success rate.
- 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.
- Success rate since 2013 kept increasing till 2020
- Decision Tree is the best model to predict first stage landing success

Thank you!