IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Antonio Marrero Rodríguez
August 25th, 2022

Outline

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

Executive Summary

- The following methodologies were used to analyze data:
  - Data collection via the SpaceX API and Web scraping.
  - Exploratory Data Analysis (EDA) with SQL, Pandas, and Matplotlib.
  - Interactive Visual Analytics and Dashboards.
  - Machine Learning Predictive Analysis.
- Summary of all results
  - EDA Results
  - Interactive Analytics
  - Predictive Analytics

Introduction

- Project background and context
  - 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 information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - Can the successful landing of the first stage of SpaceX's Falcon 9 rocket be predicted? If the likelihood of the first stage rocket landing successfully can be accurately predicted, the cost of a launch can be determined.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API (https://api.spacexdata.com/v4/rockets/)
  - Web Scrapping (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and data cleaning of null values and irrelevant columns.

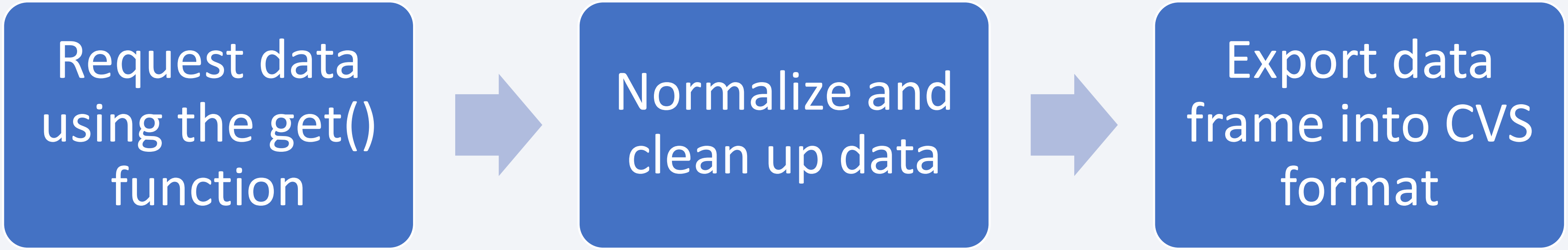
Methodology

Executive Summary (Cont.)

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Models such as: logistic regression, support vector machine, decision tree, and k nearest neighbors were built and evaluated to obtained the better classifier.

Data Collection

- Data was collected from two sources, the SpaceX Rest API and Wikipedia.
- The steps followed for both sources using Python can be summarized as the following:



Request data using the get() function

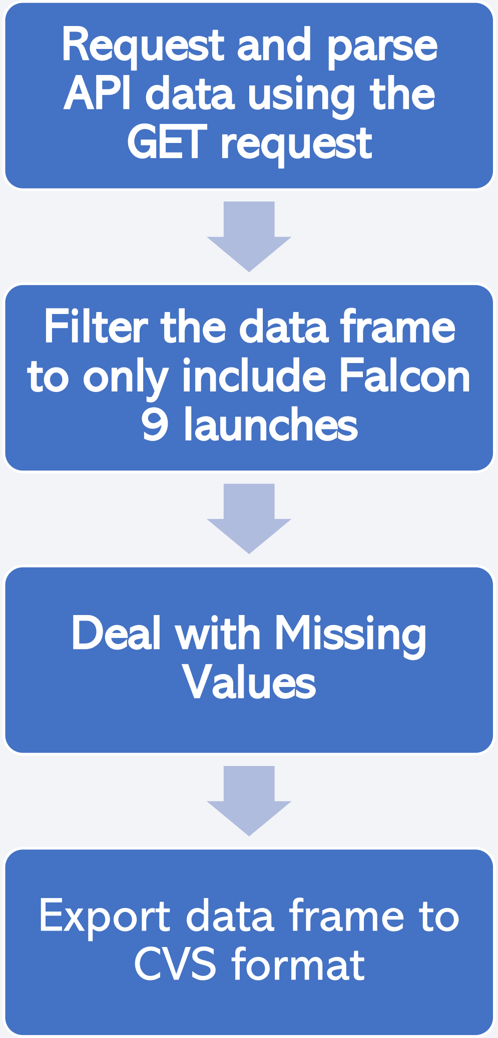
Normalize and clean up data

Export data frame into CVS format

Data Collection – SpaceX API

- SpaceX offers a public API which contains the data that will be used.
- At first the data seems like nonsense, but it can be decoded as a json file using the .json_normalize() command.
- Not all data provided is needed, therefore the data is cleaned up to only include what is needed, for example, data for rockets that are not the Falcon 9 are removed.
- Some of the data for the payload mass was missing. To deal with this problem, the average value of the payload mass was calculated and then replace the missing values.
- Finally, the data is exported into a CVS file.

Source Code:
https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/46e76e307023c9944b4c0d1b863bc156dc772e80/Data%20Collection%20API.ipynb


Request and parse API data using the GET request

Filter the data frame to only include Falcon 9 launches

Deal with Missing Values

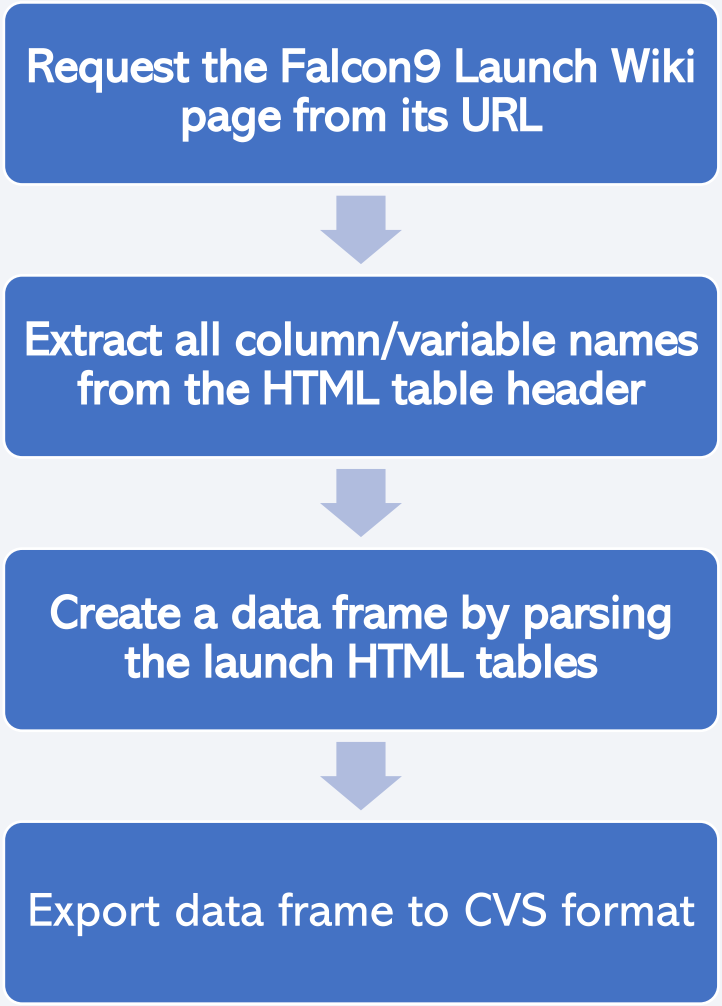
Export data frame to CVS format

Data Collection - Scraping

- Data from SpaceX launches are also obtainable from Wikipedia.
- This time, the data is obtained in html format.
- The process is described in the following flowchart.

Source Code:

https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/07471cf0744e8f5c8b92fac8f7b44baa17406c79/Webscraping.ipynb



Request the Falcon9 Launch Wiki page from its URL

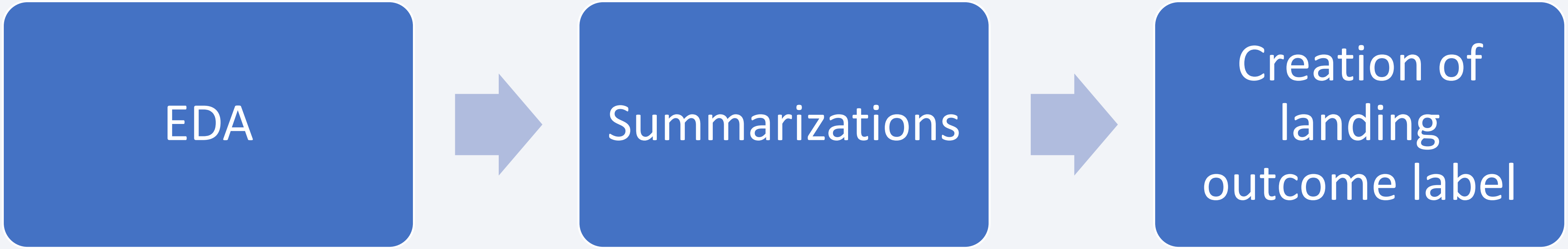
Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Export data frame to CVS format

Data Wrangling

- An exploratory data analysis (EDA) of the data obtained is performed to determine training labels.
- Summaries for launches per launching site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Then, a landing outcome label from the outcome column was created.



EDA

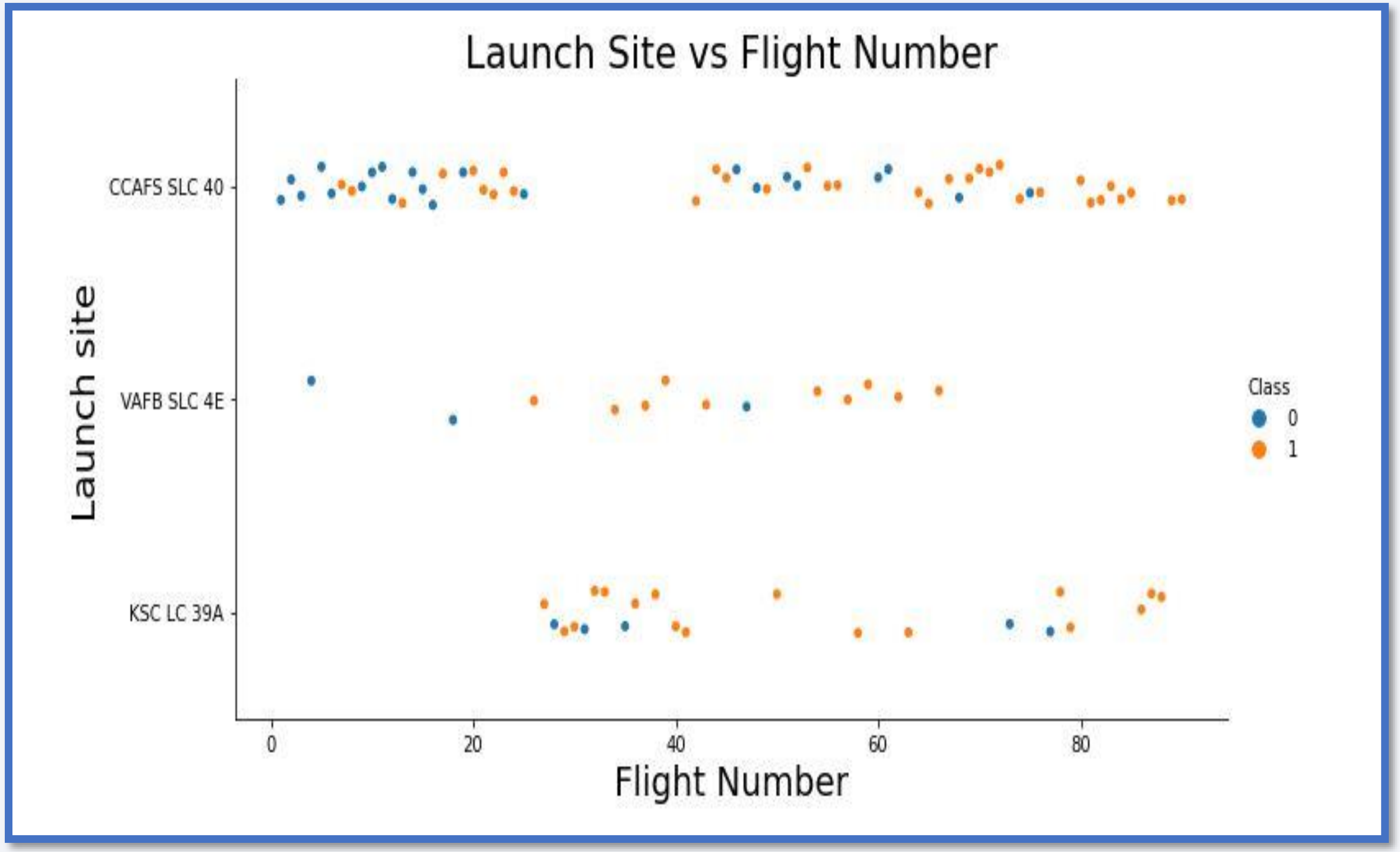
Summarizations

Creation of landing outcome label

Source Code:
https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/1722036cb670a449bece67987611a27834b177da/Data%20wrangling.ipynb

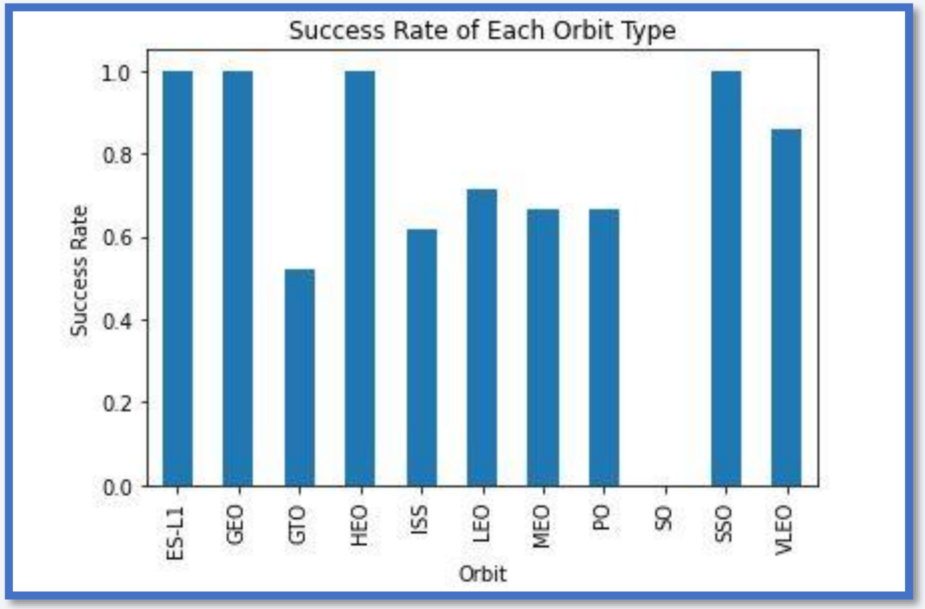
EDA with Data Visualization

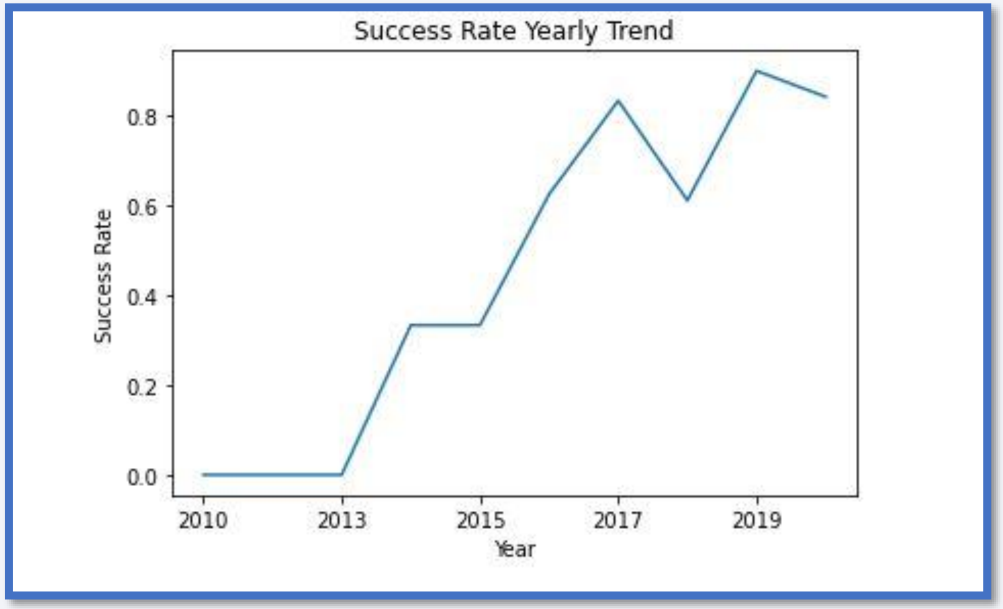
- To better understand and visualize data, the use of scatter plots, bar plots and line plots were employed.
- Scatter plots were used to visualize the following relationships:
  - Payload Mass vs Flight Number
  - Launch Site vs Flight Number
  - Launch Site vs Payload Mass
  - Orbit vs Flight Number
  - Orbit vs Payload Mass



EDA with Data Visualization

- A bar plot was used to visualize the success rate in each orbit type.
- And finally, a line plot was used to visualize the yearly trend of the success rate.





https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/85702ccdc977e8f2c3d39f8c9f1137d0de8cdde0/EDA%20with%20Visualization.ipynb

EDA with SQL

- The following SQL queries were performed:
  - Display the names of the unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string 'CCA’.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass.
  - List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, and launch site for the months in year 2015.
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/55ec4a3e1d9ccea8cd058c3efd382b0b2252e234/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/c36fd25d4a4fa847185f31963dc15d838b47aa82/lab_jupyter_launch_site_location.ipynb

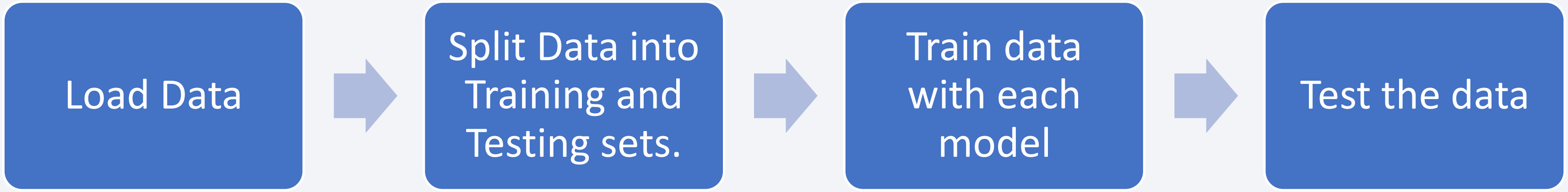
Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data:
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify the best place to launch according to payloads.

https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/f252d19e27a1f60fde7df44323c9d6158c4a36c6/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.



Load Data

Split Data into Training and Testing sets.

Train data with each model

Test the data

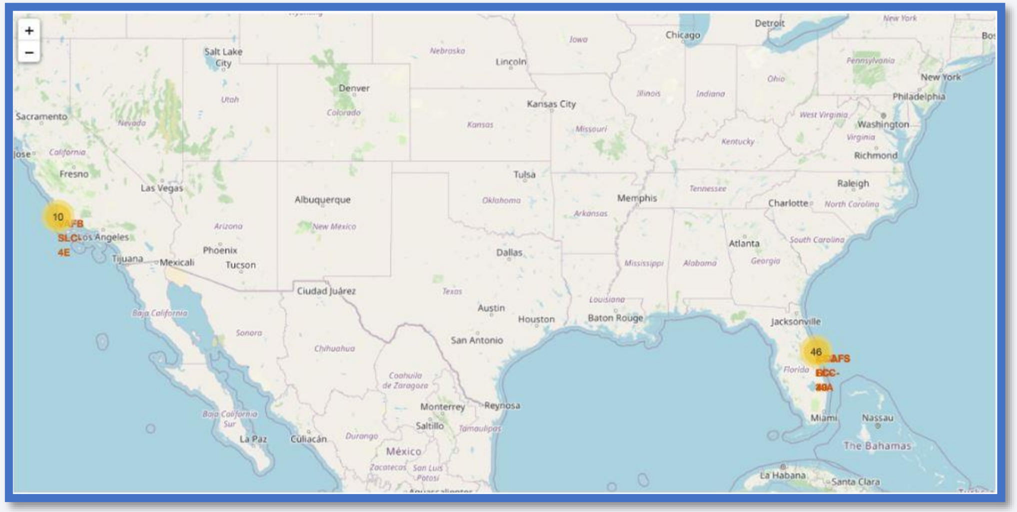
https://github.com/ArtificerLucian/Applied-Data-Science-Capstone/blob/09769962b8275739eaac09378d2331093e9bbe40/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

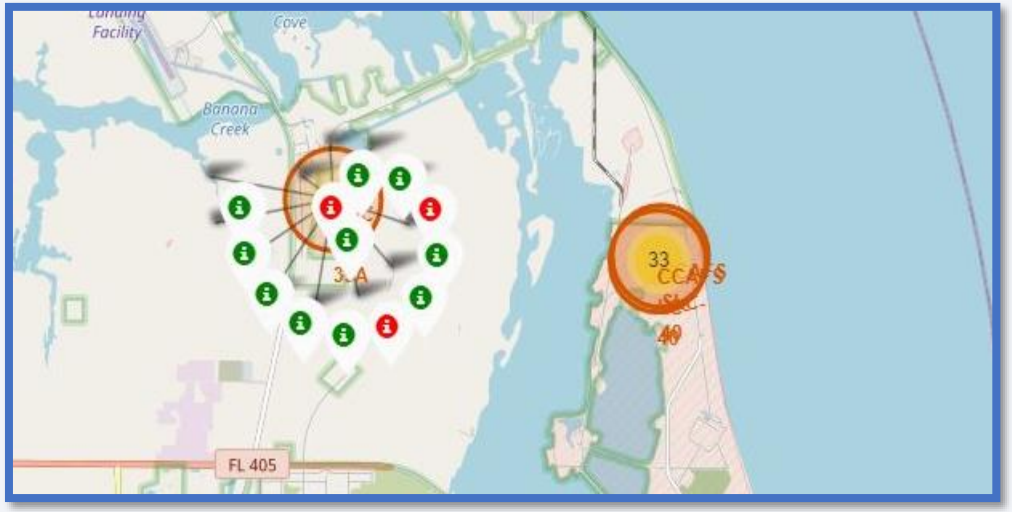
Results

- Exploratory data analysis results
  - Space X has used 4 different launch sites: one in Kennedy Space Center, two in Cape Canaveral, and one on Vandenberg Space Force Base.
  - The average payload of F9 v1.1 booster is 2,928 kg.
  - The first success landing outcome happened in 2015 fiver year after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

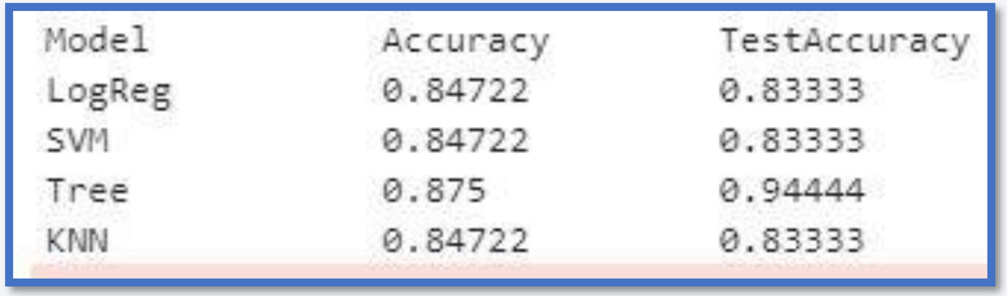
- Utilizing interactive analytics, it was possible to identify the location of the launch sites which are near the coast and away from cities.



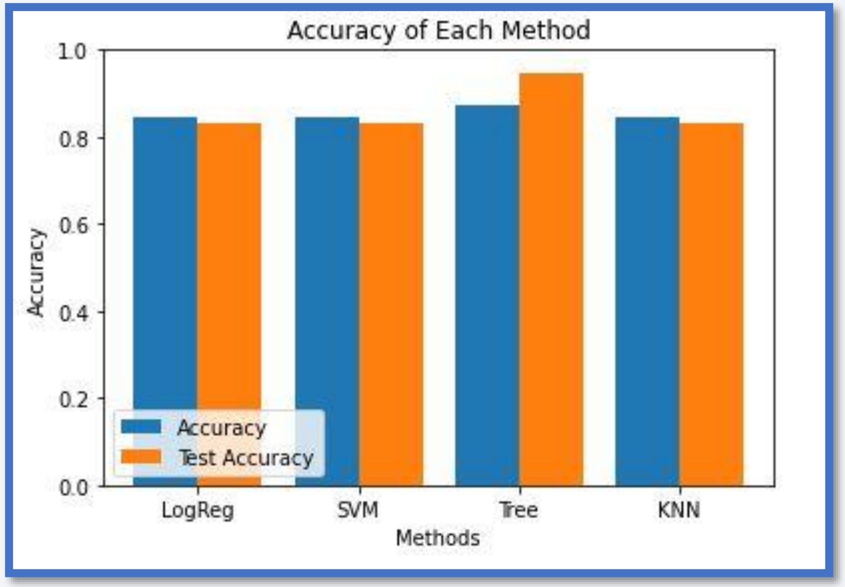


Results

- Predictive analysis results show that the decision tree classifier model is the best model for this project at it has slightly better training and testing accuracies with a training accuracy of 87.5% and a testing accuracy of 94.44%.



Model	Accuracy	TestAccuracy
LogReg	0.84722	0.83333
SVM	0.84722	0.83333
Tree	0.875	0.94444
KNN	0.84722	0.83333



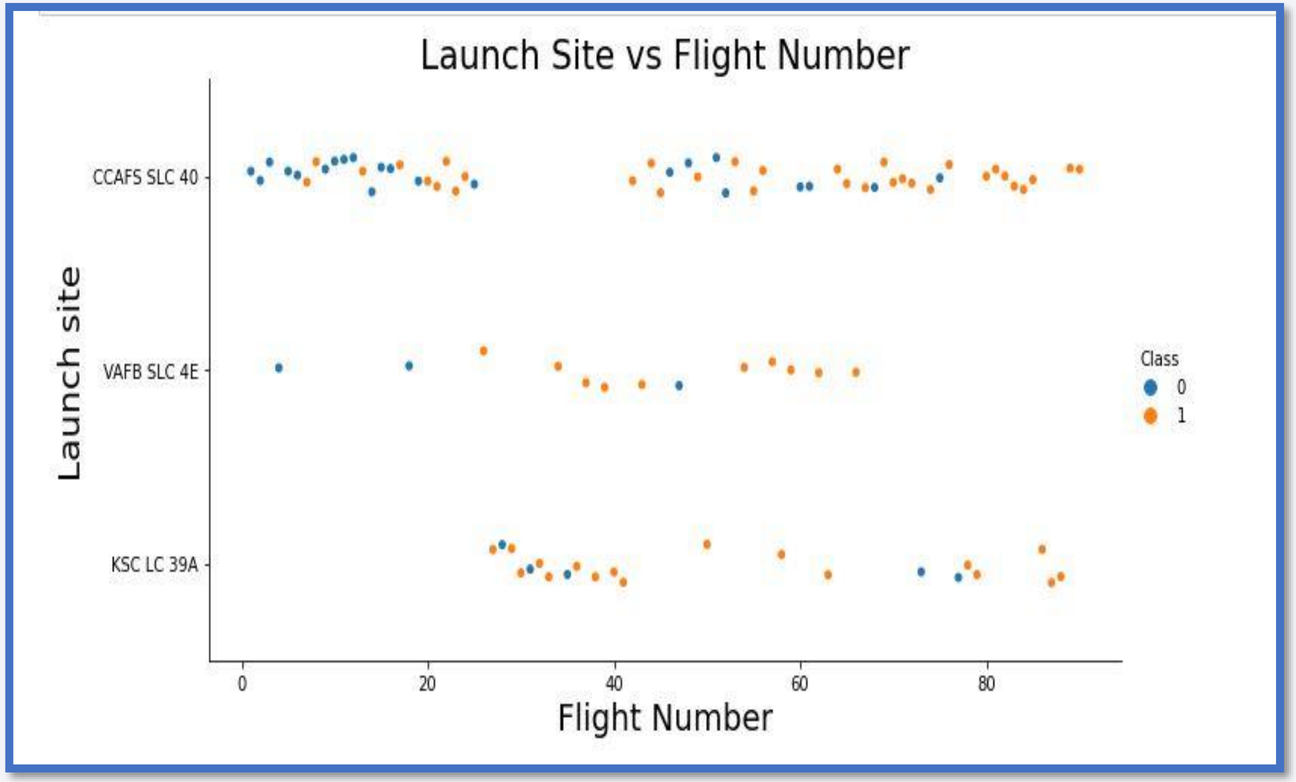
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

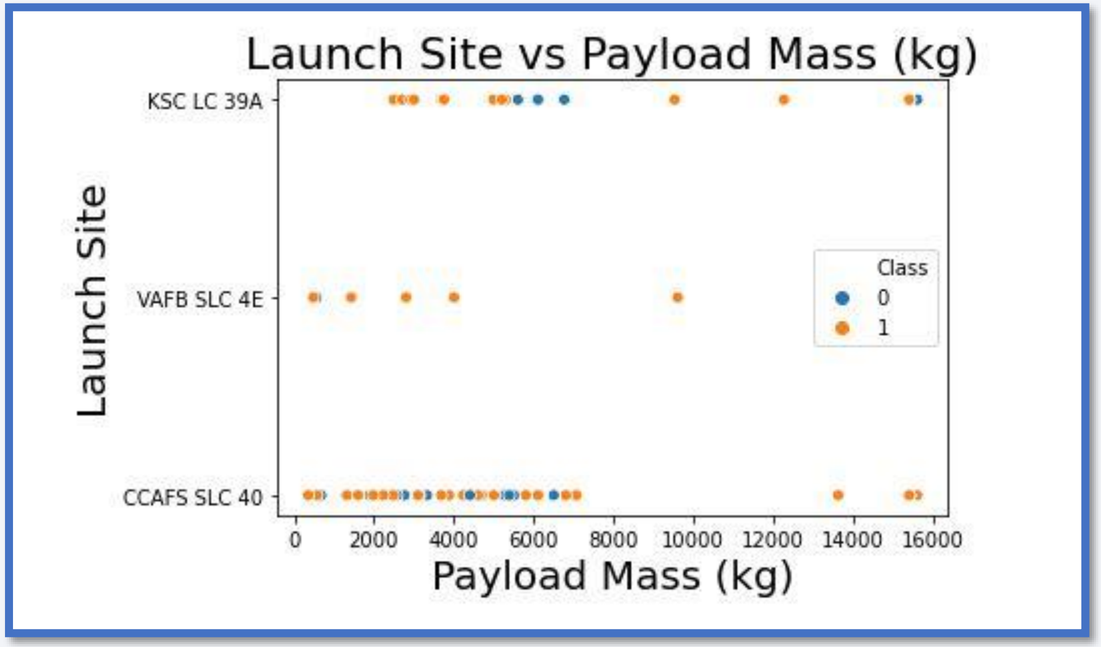
From this plot, it can be observed that Cape Canaveral has had the most launches while Kenney Space Center has the best success rate.

It can also be observed that as the flight number increases there are less failures.



Payload vs. Launch Site

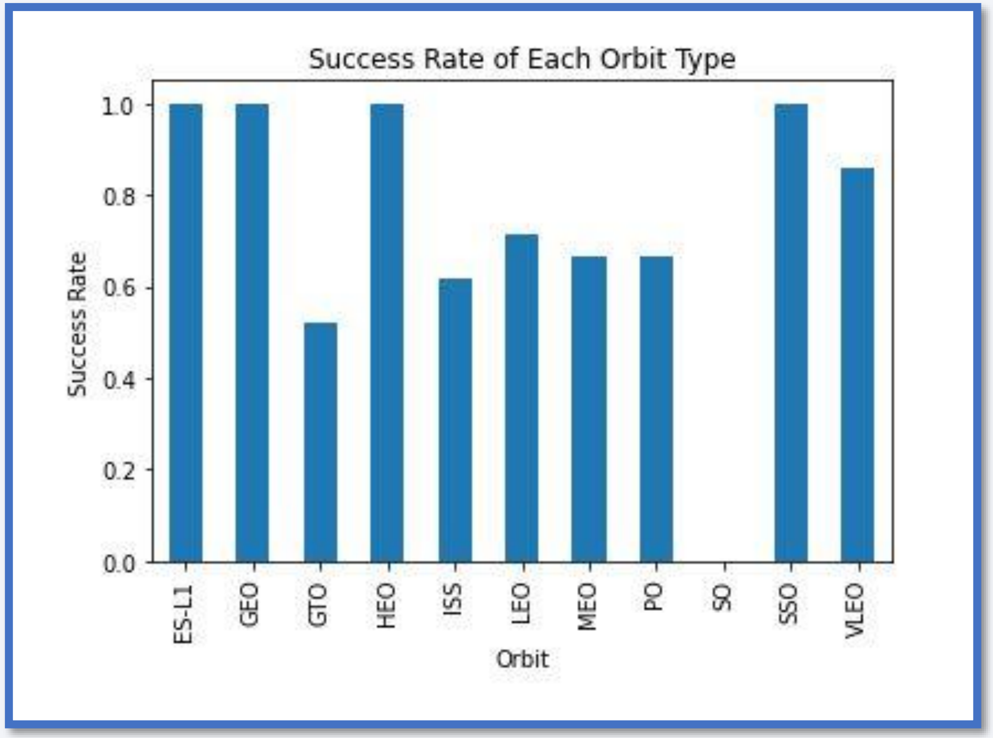
Most launches have had a payload mass below 8000 kg (about 17636.96 lb). There have only been 8 launches with payloads above 8000 kg with only one failed launch at around 16000 kg. It should also be noted that Vandenberg Space Force Base has not had a launch with a payload above 10000 kg.



Success Rate vs. Orbit Type

The ES-L1, GEO, HEO, and SSO have had an impeccable success rate of 100%, with VLEO reaching a success rate of around 85.7%.

It should also be noted that the SO orbit did not have a successful launch meaning its success rate is 0%.

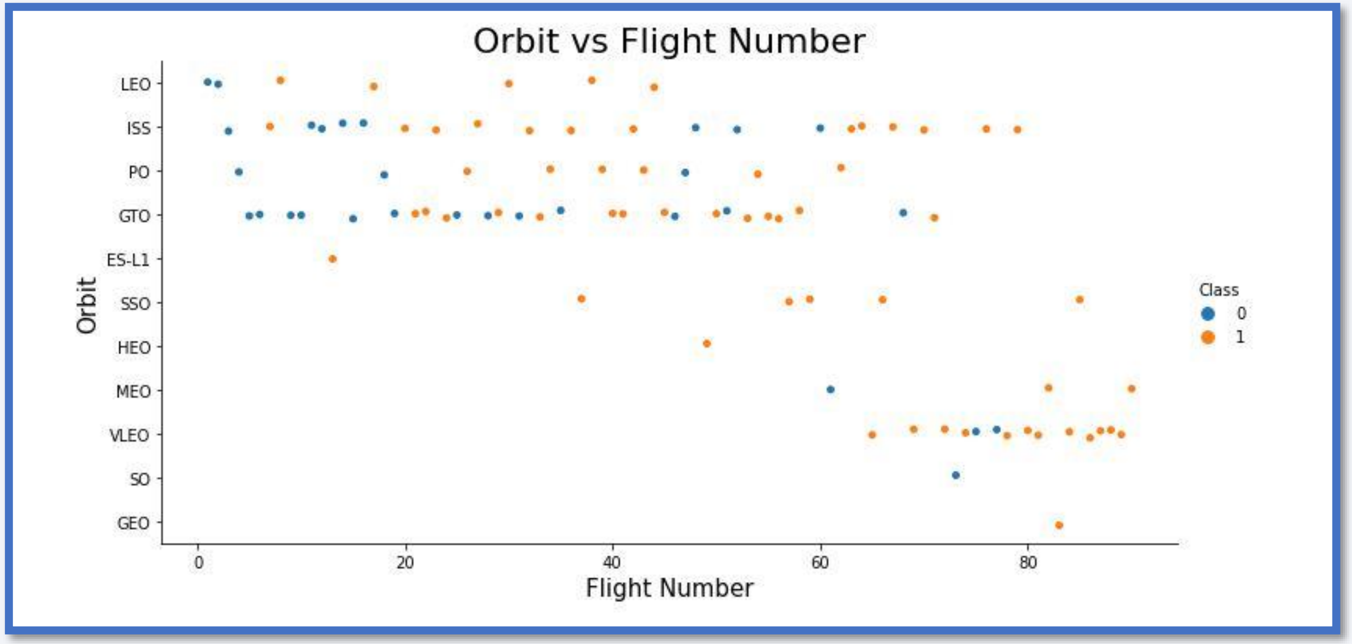


Flight Number vs. Orbit Type

This plot show an increase in successful launches as the flight number increases.

An interesting note when taking the previous plot into consideration, while the ES-L1, HEO, and Geo orbits have a perfect launch rate, there has only been one launch in each of these orbits.

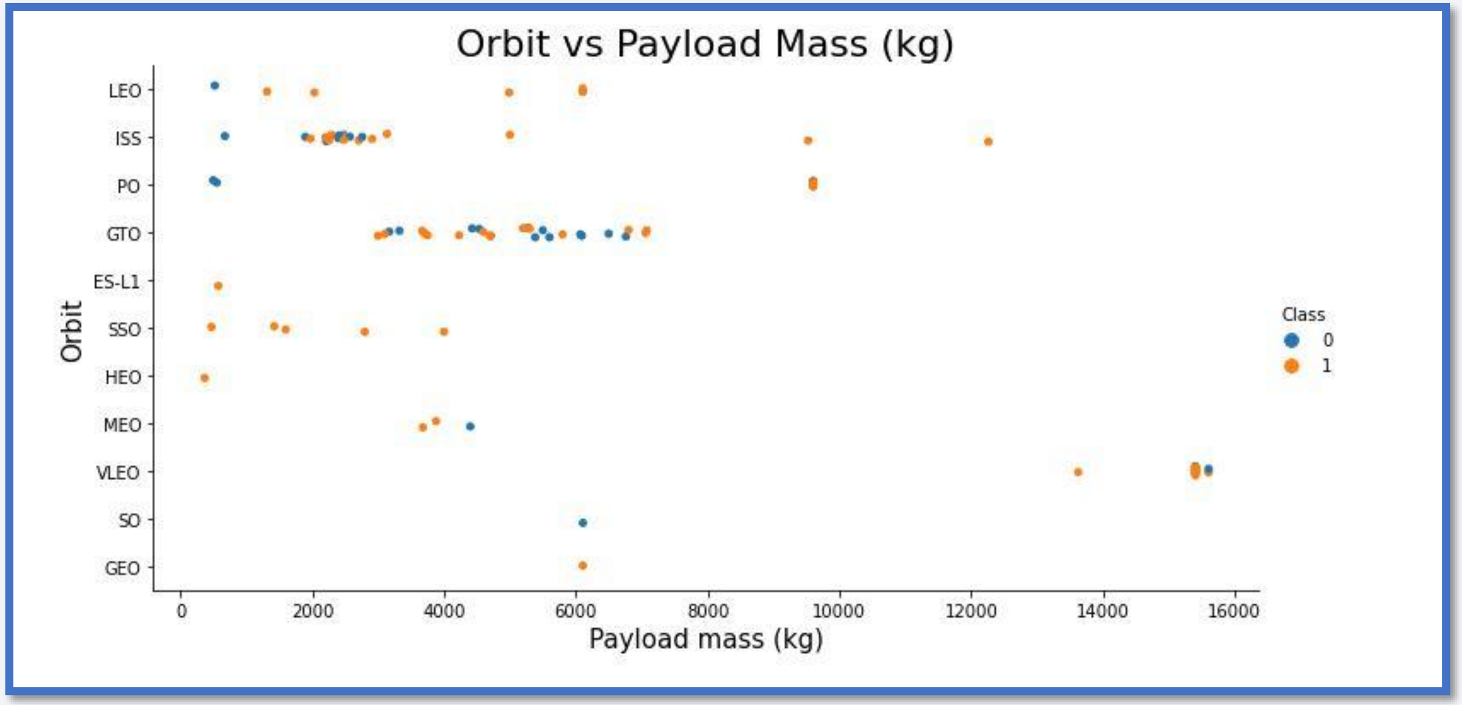
The same can be said of the SO orbit. It 0% success rate is due to only having one launch that just so happened to have failed.



Payload vs. Orbit Type

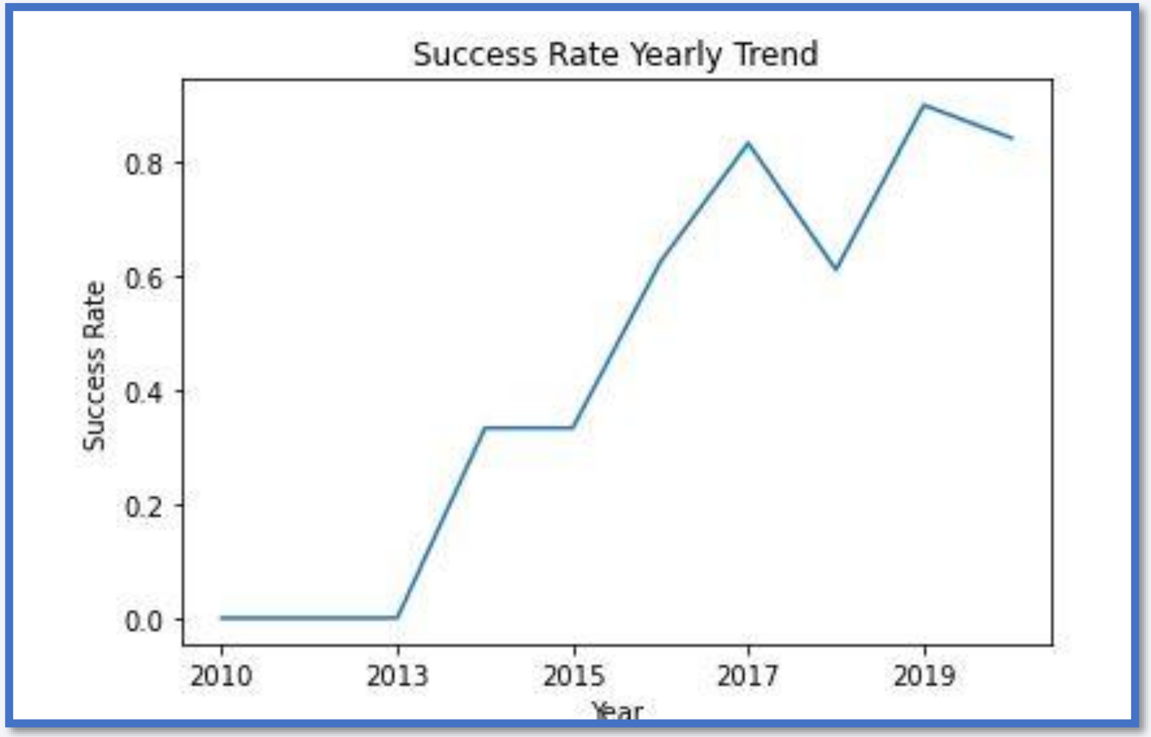
This plot shows the relation ship between the payload mass and the orbits. It can be observed that the majority of launches have had a payload mass between 2000 kg and 8000 kg.

The orbit with the largest range is ISS with payload masses between 1000 kg and 13000 kg.

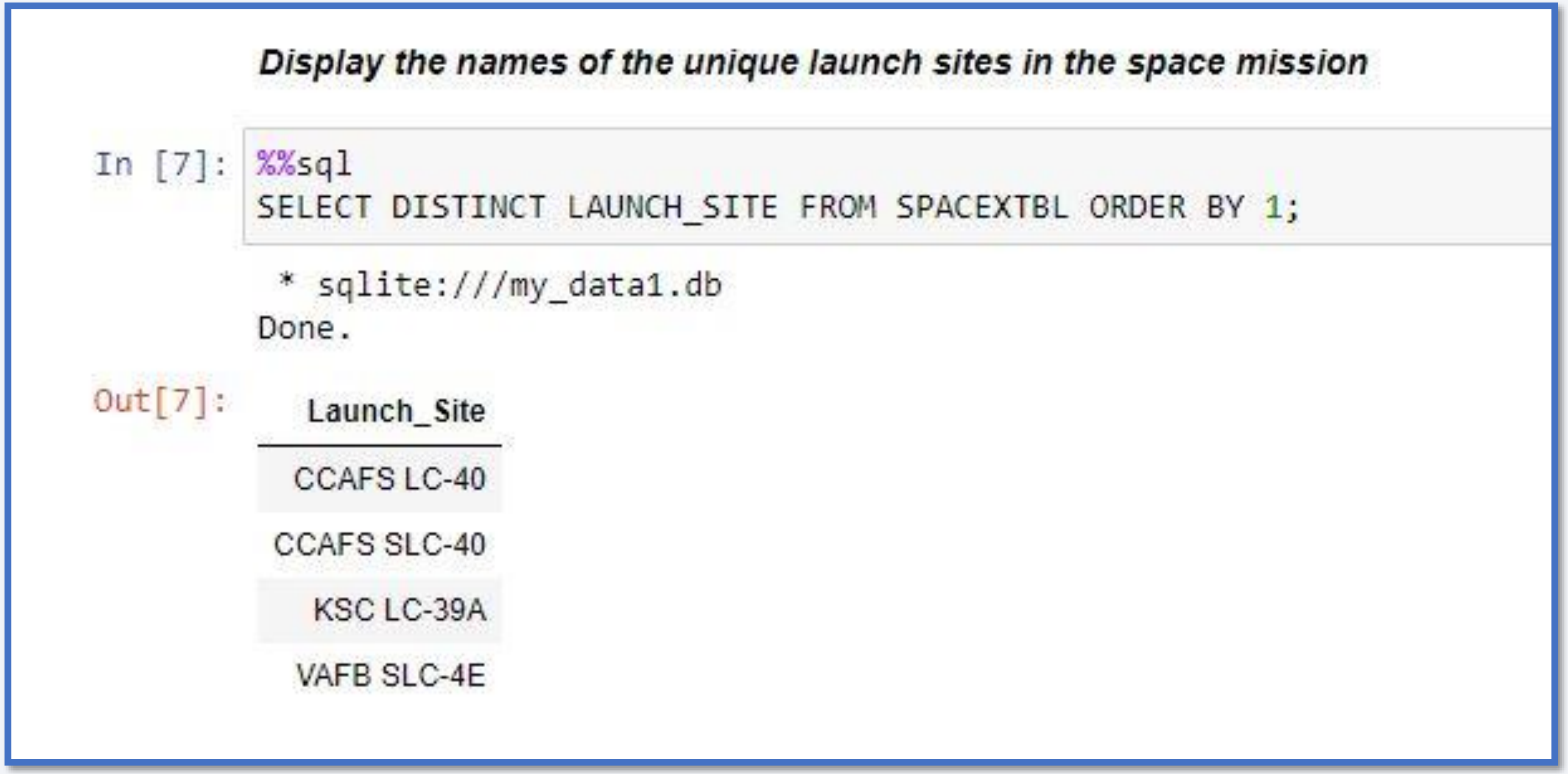


Launch Success Yearly Trend

This plot shows the yearly trend of the success rate. It can be observed that from 2013 and 2020 there has been a constant increase in the success rate with only a slight dip in 2018.



All Launch Site Names



Display the names of the unique launch sites in the space mission

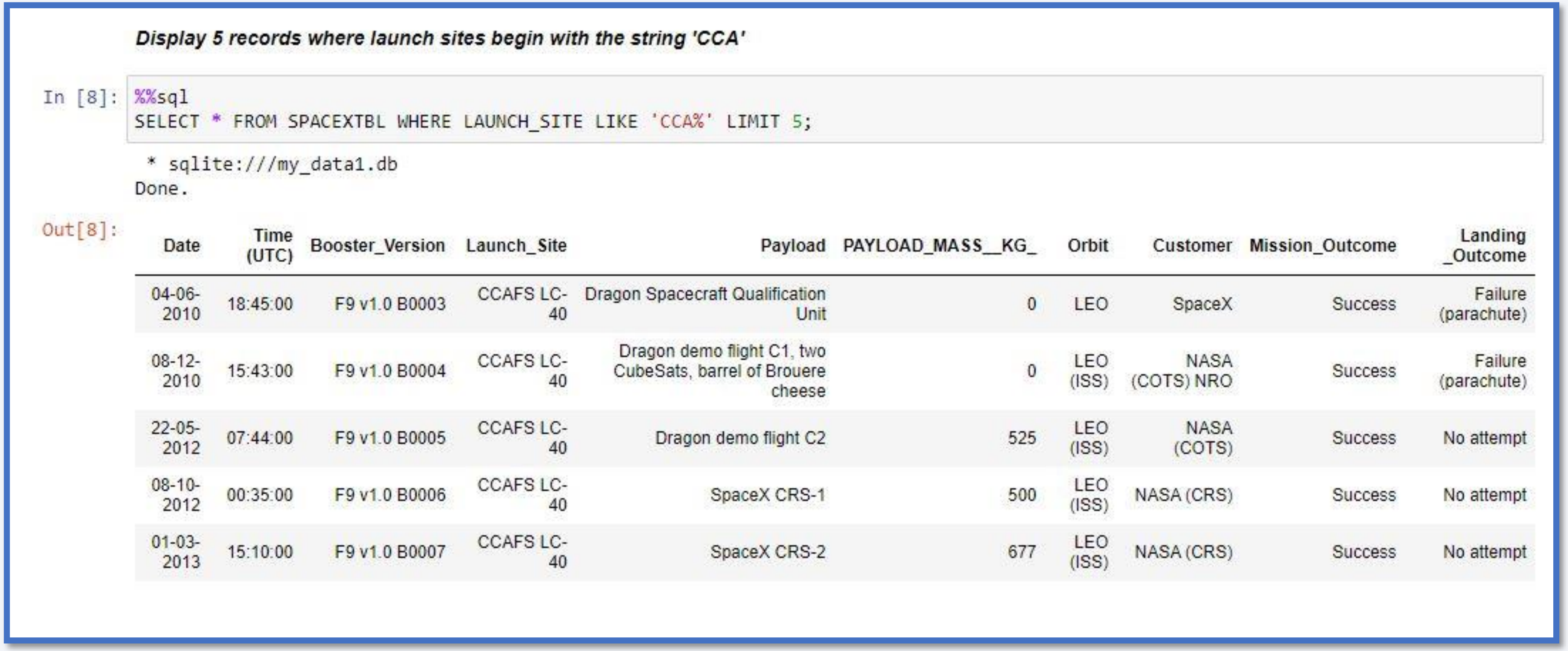
```
In [7]: %%sql
        SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

Out[7]:

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

Launch Site Names Begin with 'CCA'



Display 5 records where launch sites begin with the string 'CCA'

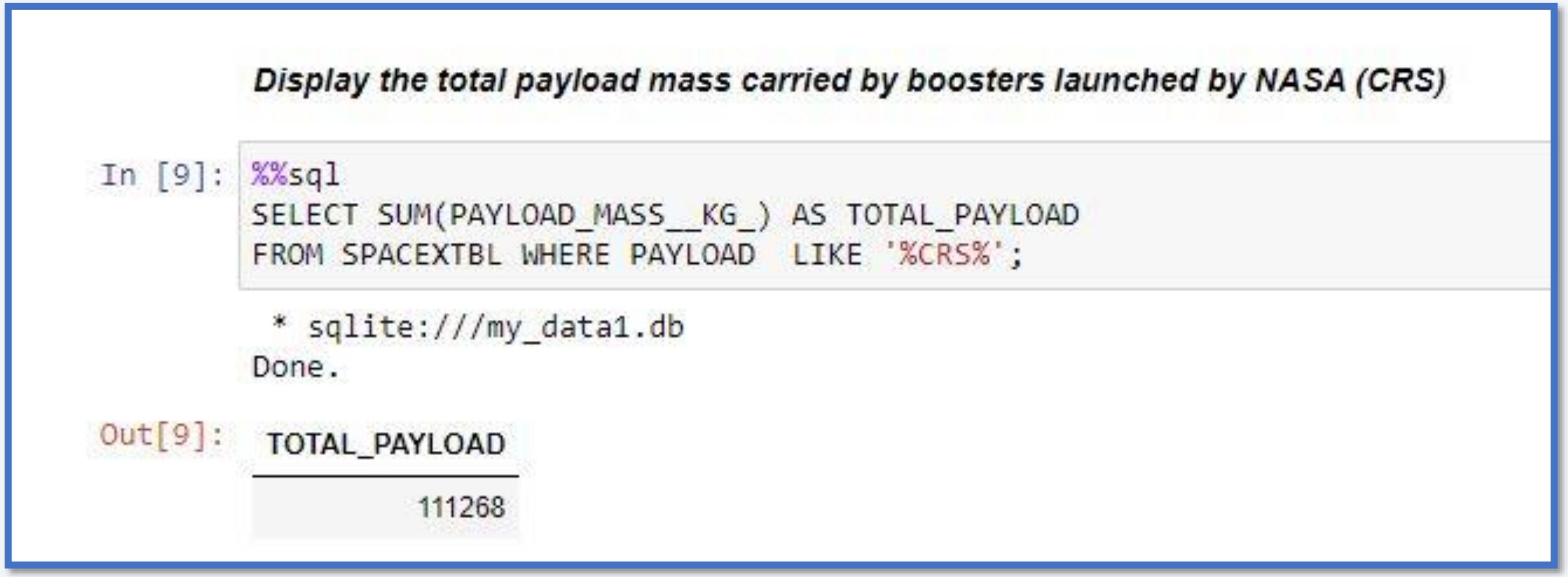
```
In [8]: %%sql
        SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

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

Out[8]:

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



Display the total payload mass carried by boosters launched by NASA (CRS)

In [9]:

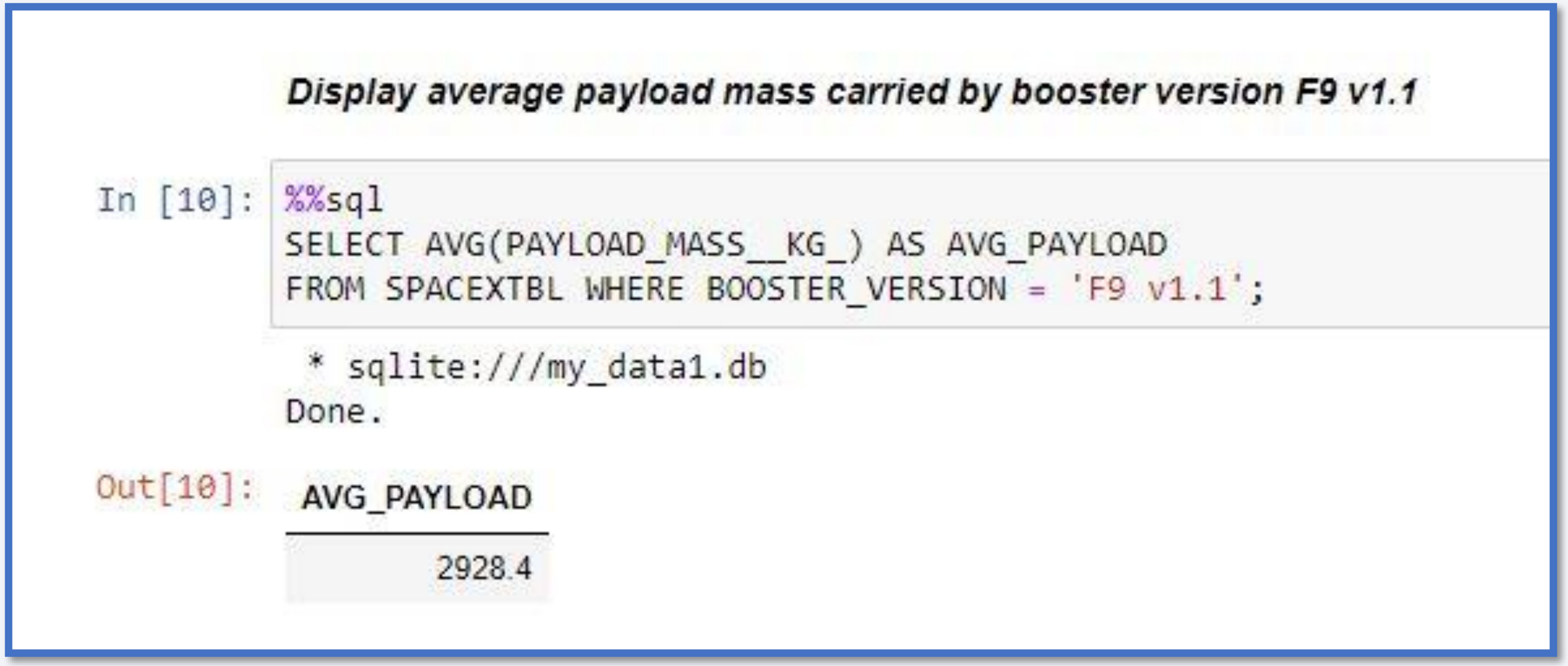
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD
FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

Out[9]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1



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

In [10]:

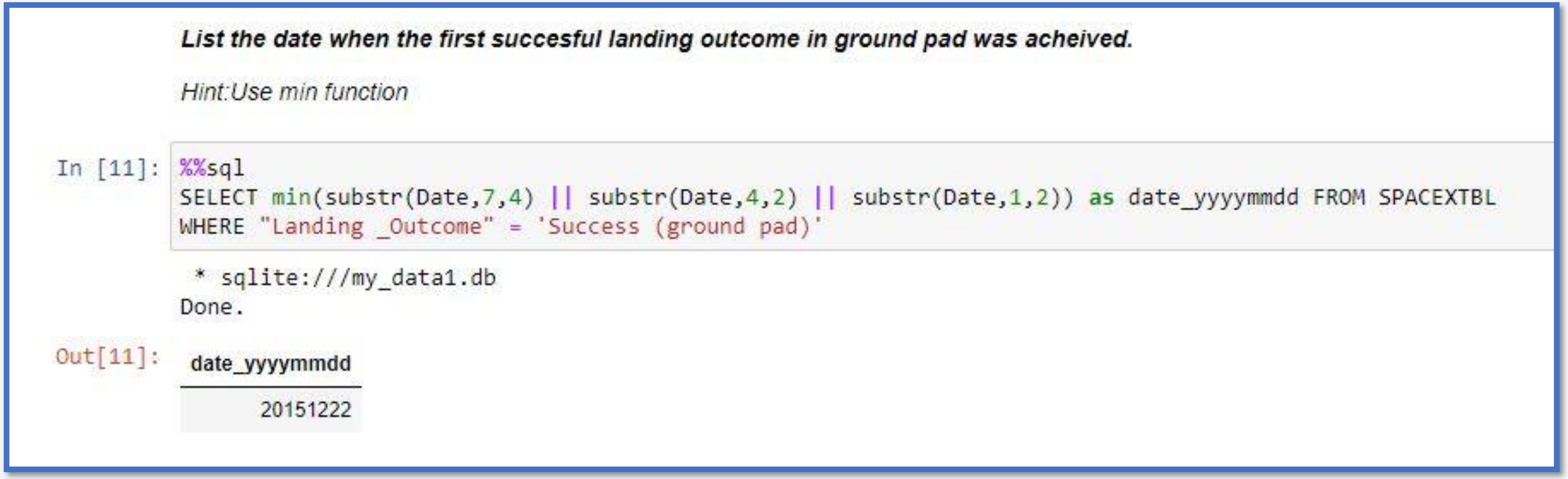
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD
FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

Out[10]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date



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

Hint:Use min function

In [11]:

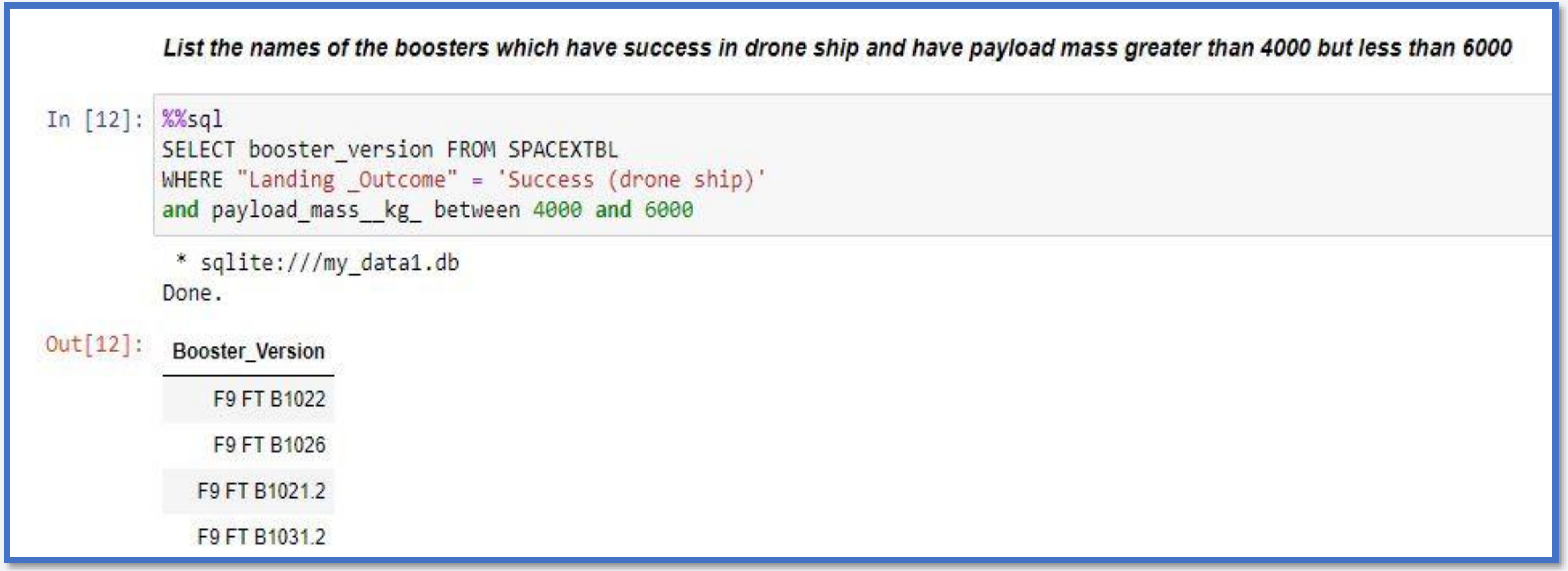
```
%%sql
SELECT min(substr(Date,7,4) || substr(Date,4,2) || substr(Date,1,2)) as date_yyyymmdd FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (ground pad)'
```

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

Out[11]:

date_yyyymmdd
20151222

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

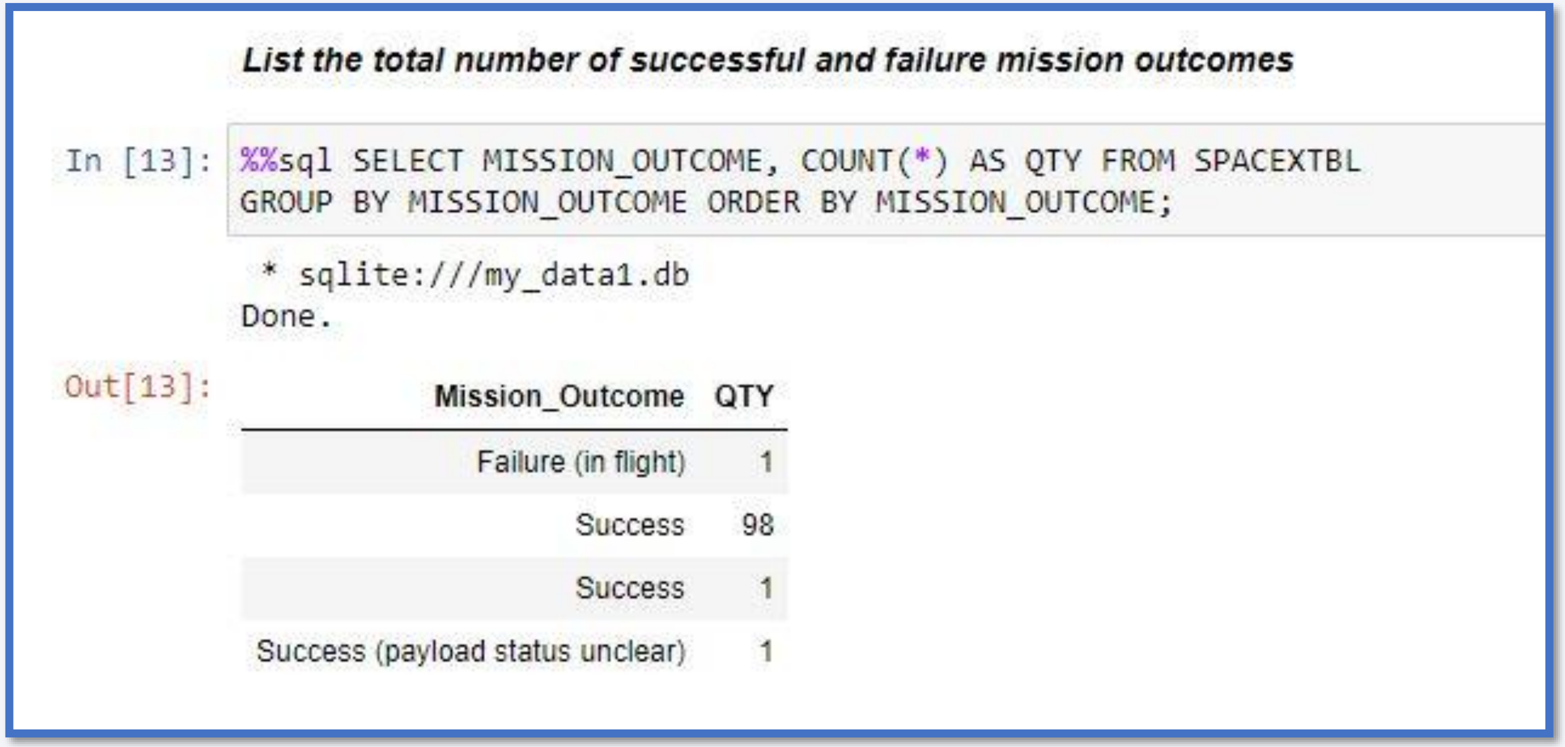
In [12]:
```
%%sql
SELECT booster_version FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success (drone ship)'
and payload_mass__kg_ between 4000 and 6000
```

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

Out[12]:

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

Total Number of Successful and Failure Mission Outcomes



List the total number of successful and failure mission outcomes

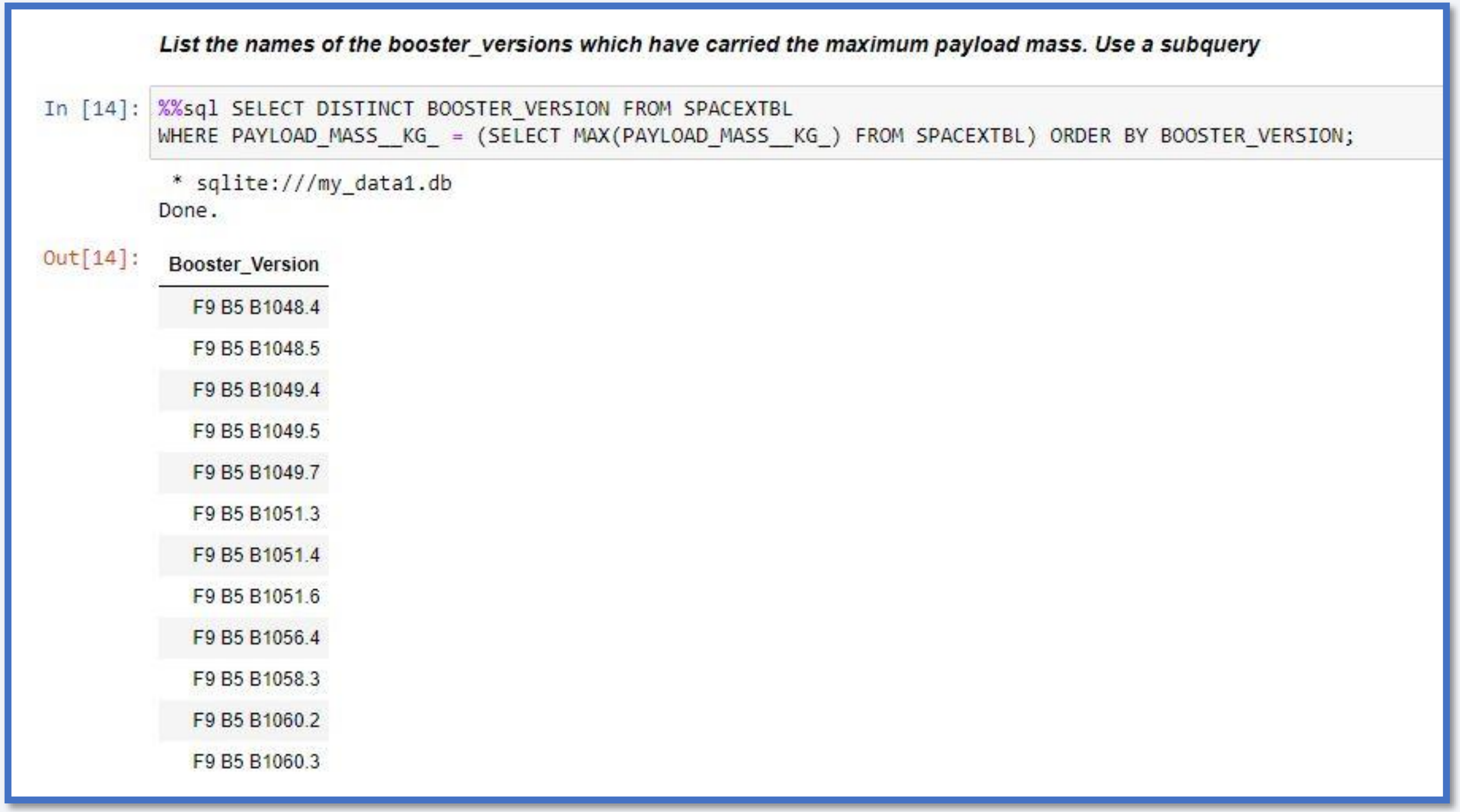
```
In [13]: %%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
         GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[13]:

Mission_Outcome	QTY
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

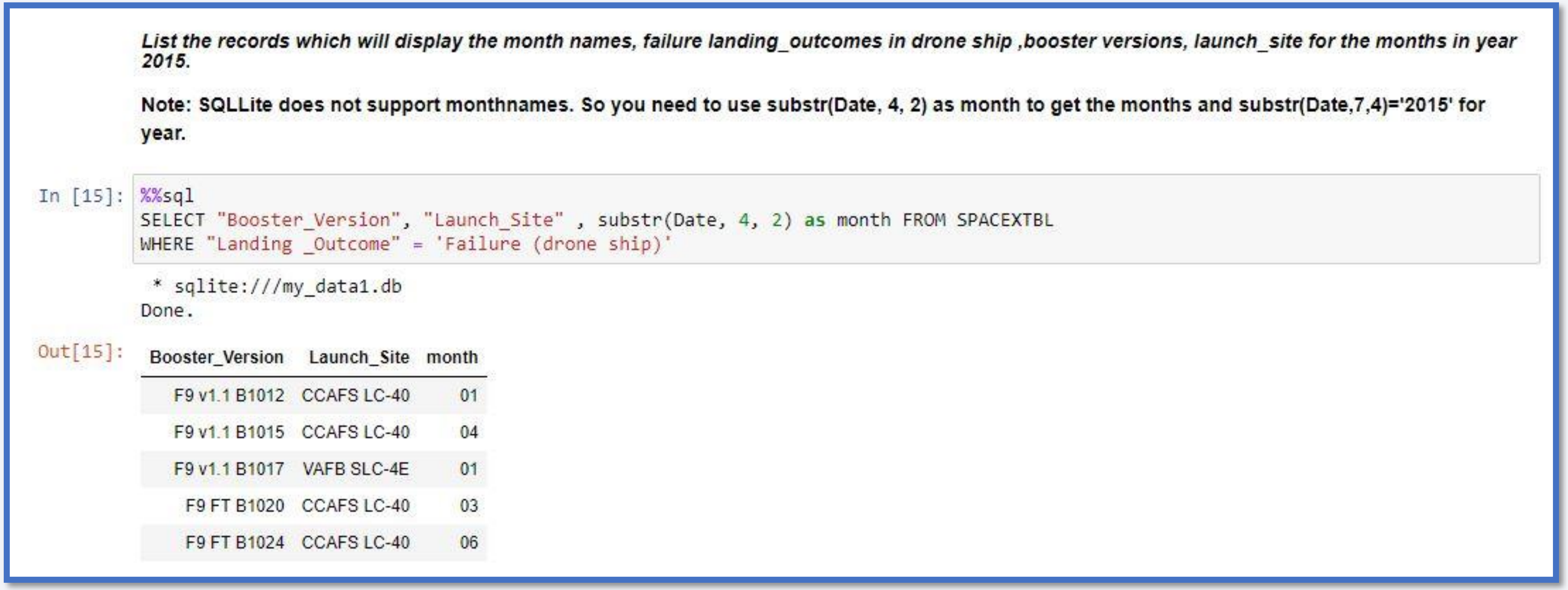
In [14]:
```
%%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;
```

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

Out[14]:

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



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.

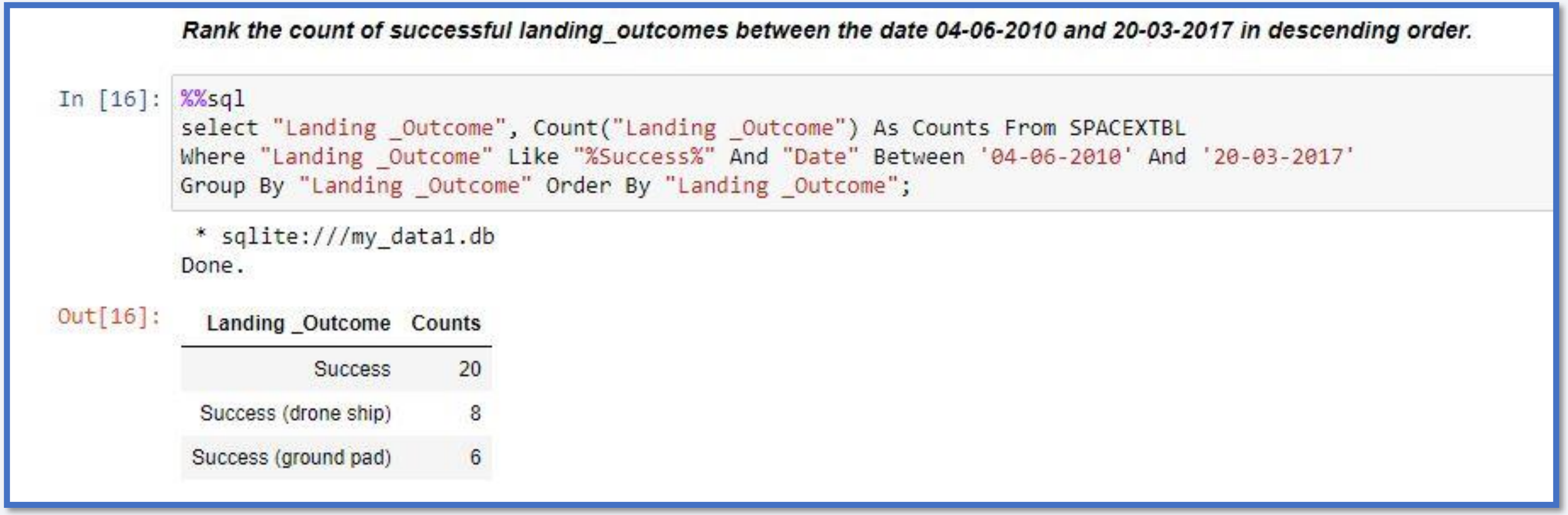
```
In [15]: %%sql
         SELECT "Booster_Version", "Launch_Site" , substr(Date, 4, 2) as month FROM SPACEXTBL
         WHERE "Landing _Outcome" = 'Failure (drone ship)'

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

Out[15]:

Booster_Version	Launch_Site	month
F9 v1.1 B1012	CCAFS LC-40	01
F9 v1.1 B1015	CCAFS LC-40	04
F9 v1.1 B1017	VAFB SLC-4E	01
F9 FT B1020	CCAFS LC-40	03
F9 FT B1024	CCAFS LC-40	06

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



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

In [16]:

```
%%sql
select "Landing _Outcome", Count("Landing _Outcome") As Counts From SPACEXTBL
Where "Landing _Outcome" Like "%Success%" And "Date" Between '04-06-2010' And '20-03-2017'
Group By "Landing _Outcome" Order By "Landing _Outcome";
```

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

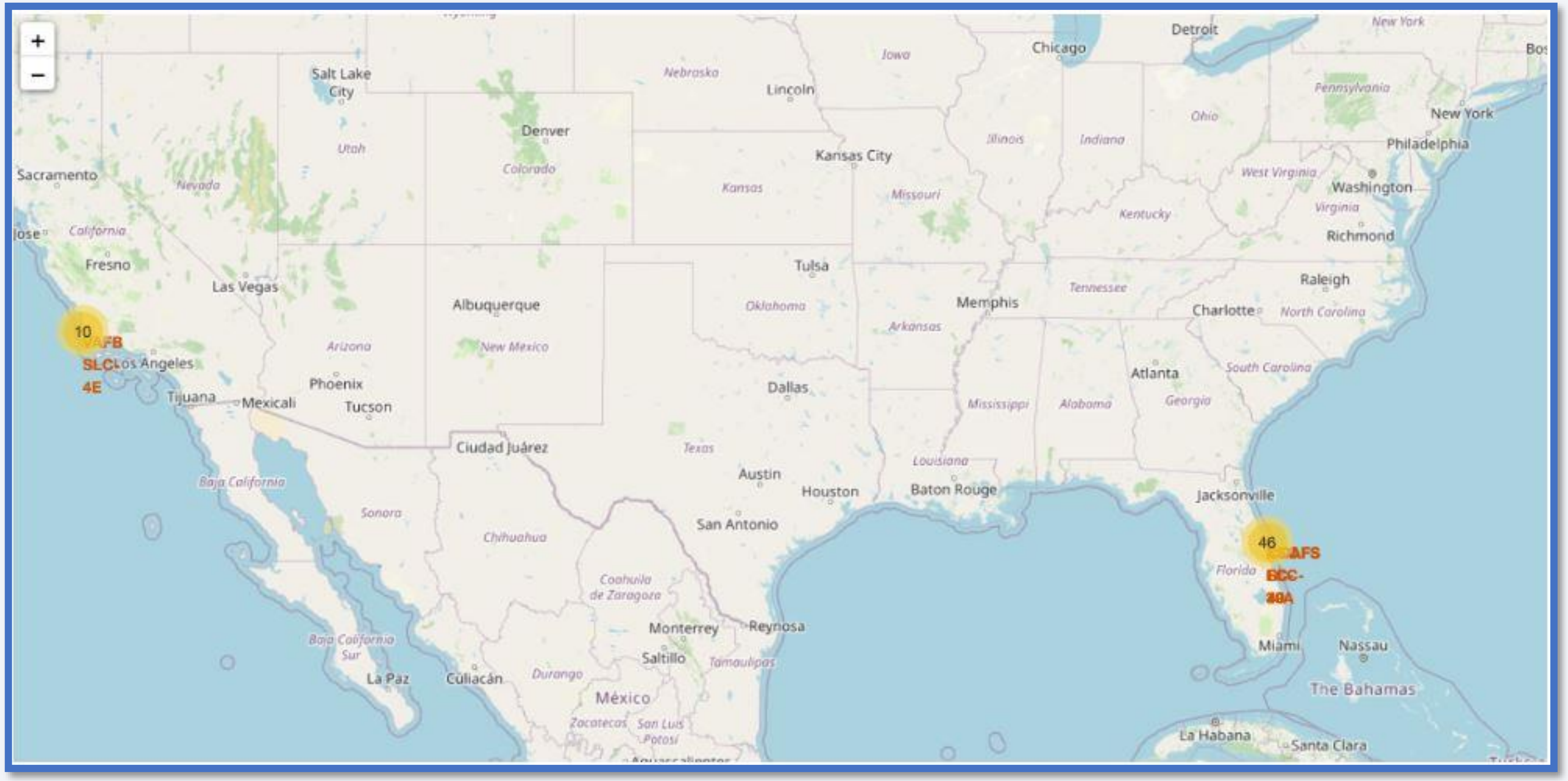
Out[16]:

Landing _Outcome	Counts
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

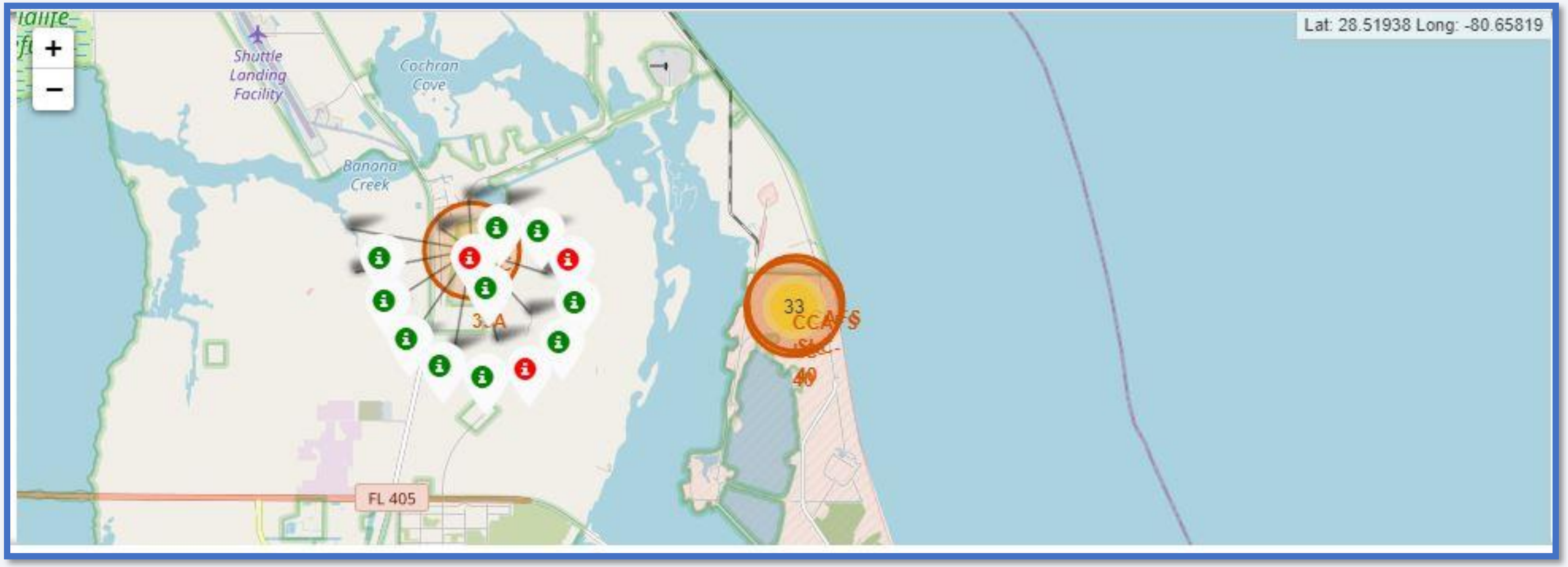
Launch Sites Proximities Analysis

Launch Sites Locations



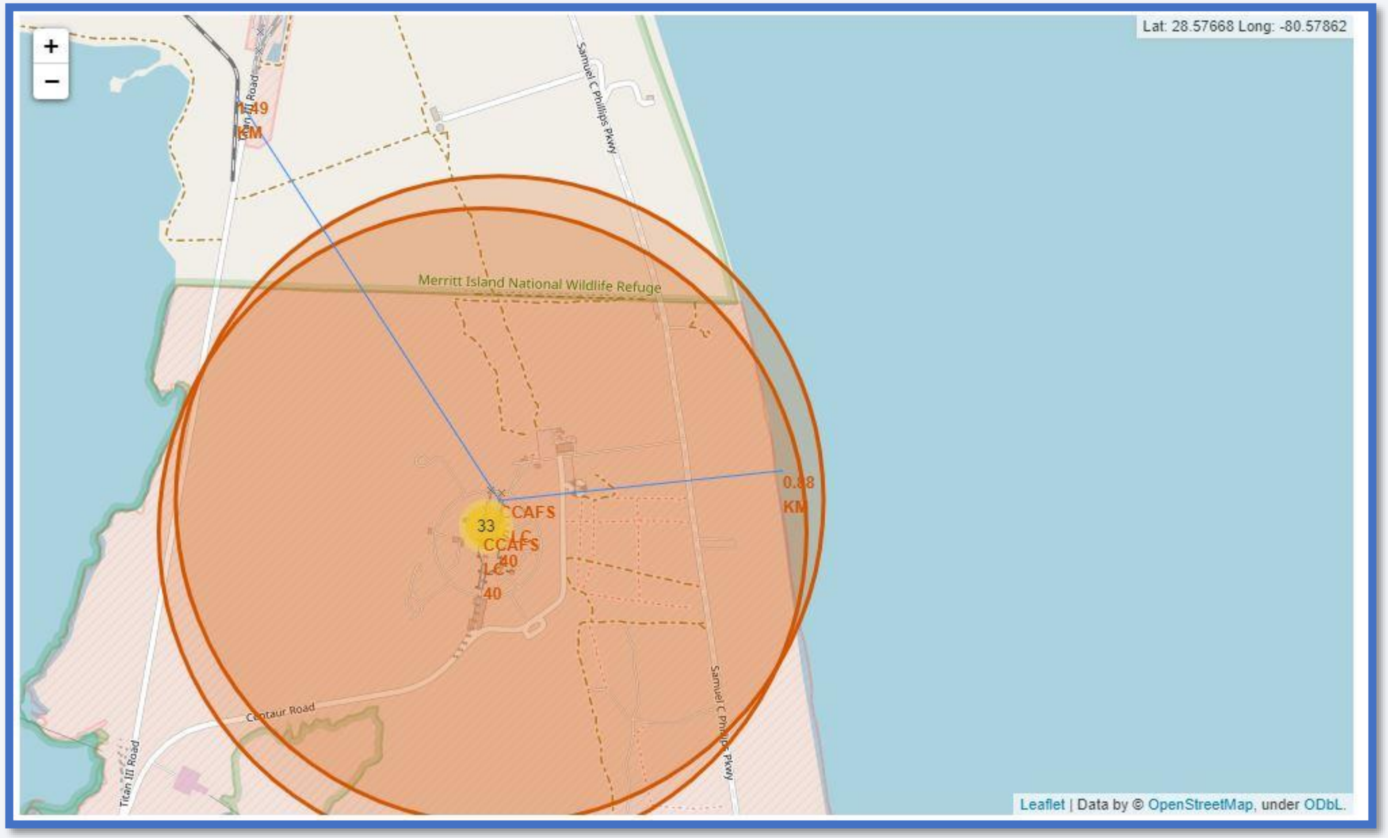
Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Launch Outcome Markers



Green markers indicate successful and red ones indicate failure.

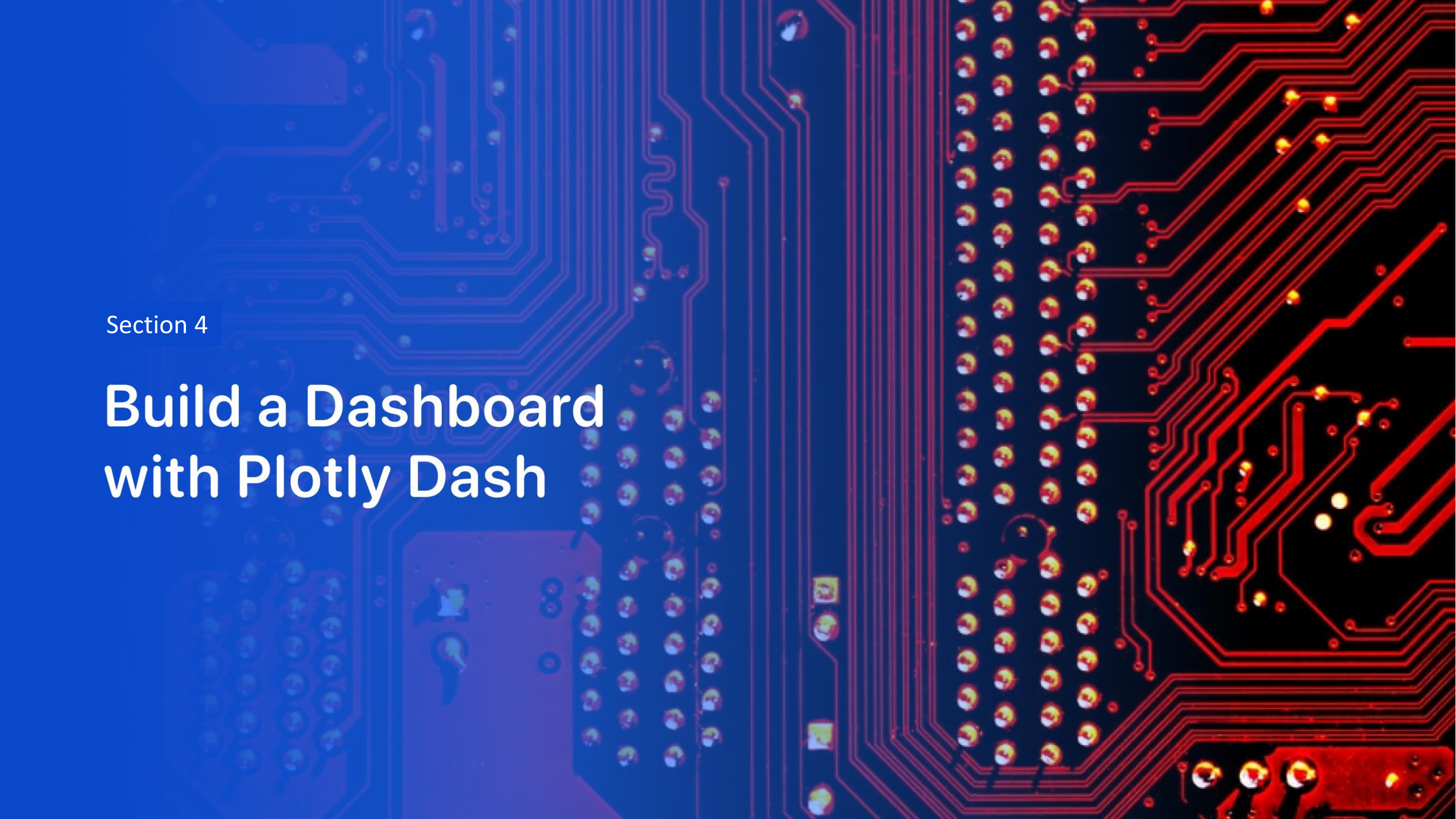
Distance Markers


+
−
Lat: 28.57668 Long: -80.57862
1.49
KM
Samuel C Phillips Pkwy
Merritt Island National Wildlife Refuge
0.88
KM
33
CAFS
CCAFS
LC
40
LC
40
Centaur Road
Titan III Road
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

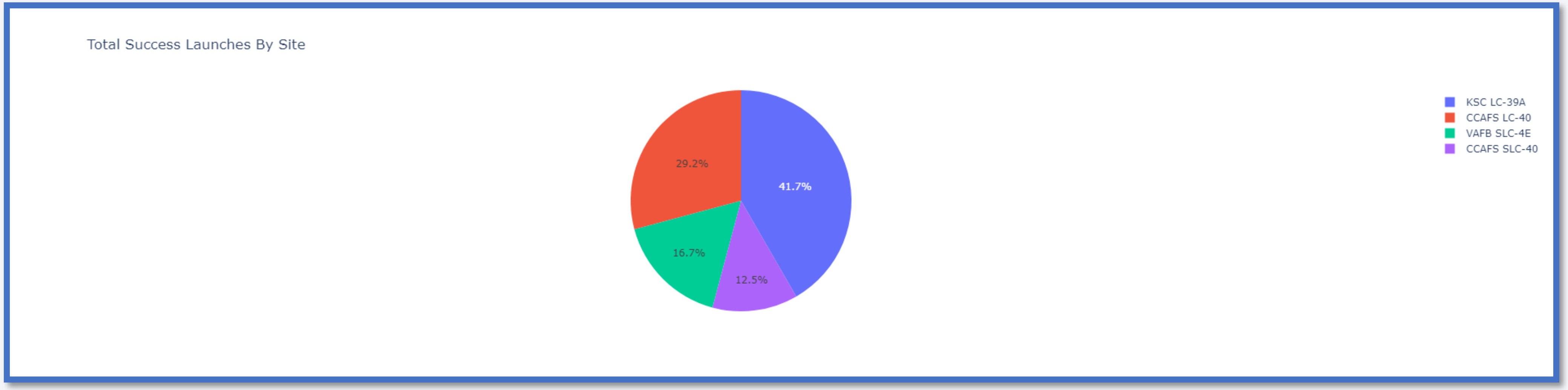
The launch sites are relatively close from the coast and railroads while keeping a fair distance from cities for safety reasons.

Section 4

Build a Dashboard with Plotly Dash

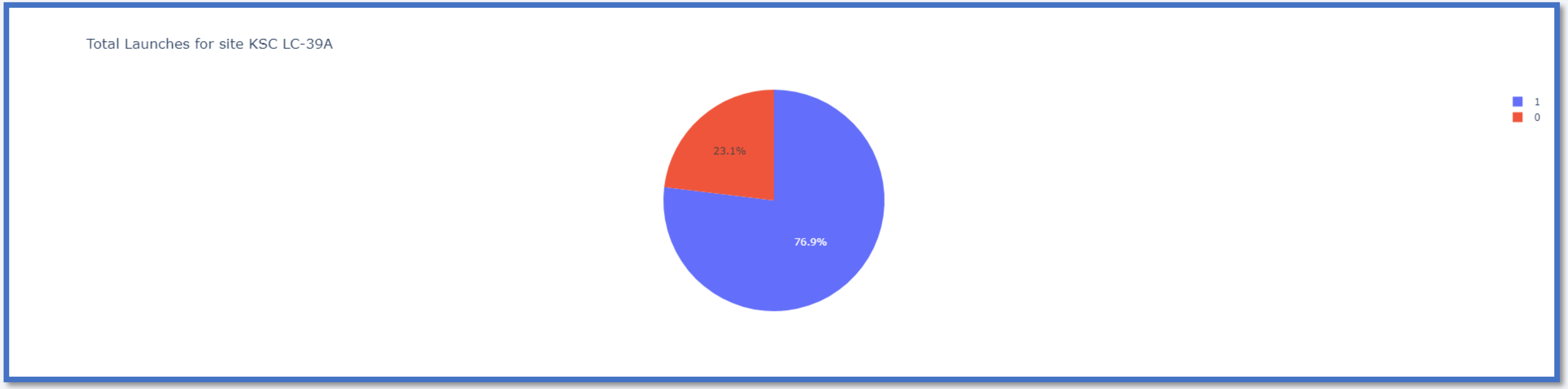


Successful Launches by Site



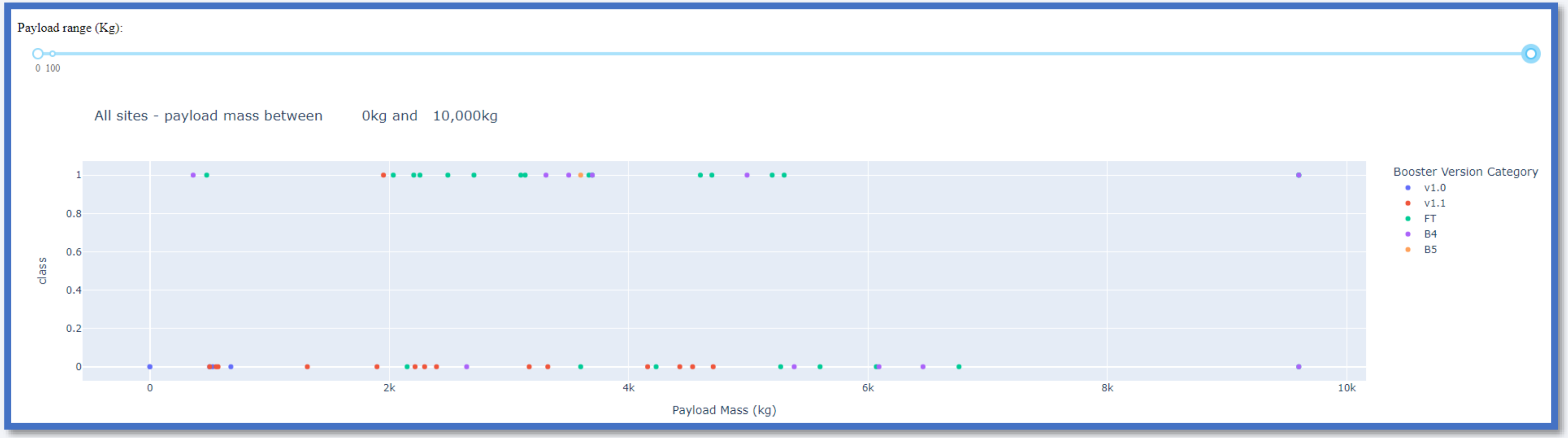
The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A

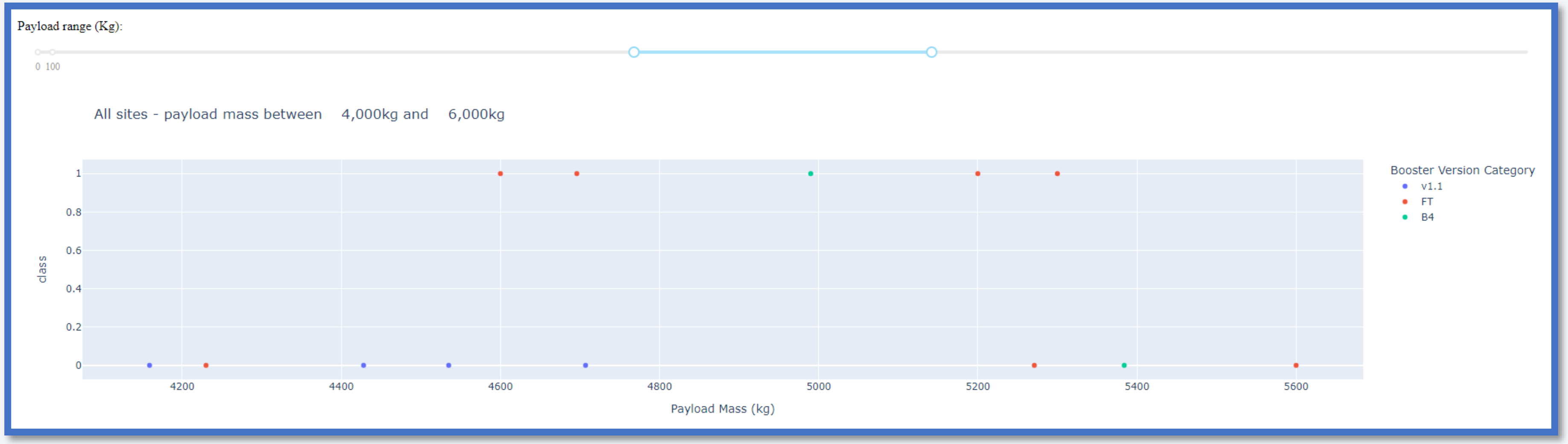


The launch site with the highest success rate is KSC LC-39A with 76.9% of its launches being successful.

Payload Mass Range Slider


Payload range (Kg):

All sites - payload mass between 0kg and 10,000kg


Payload range (Kg):

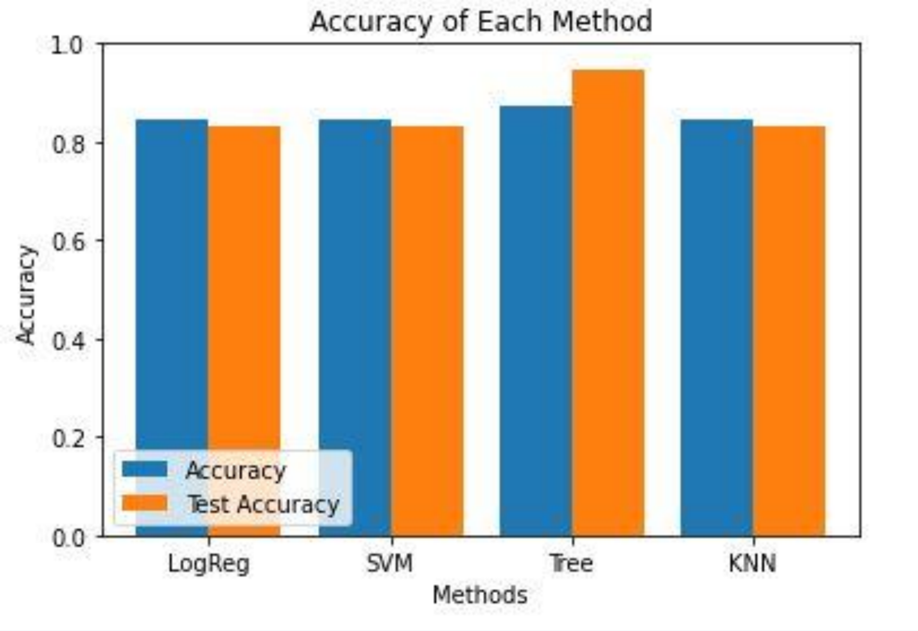
All sites - payload mass between 4,000kg and 6,000kg

Section 5

Predictive Analysis (Classification)

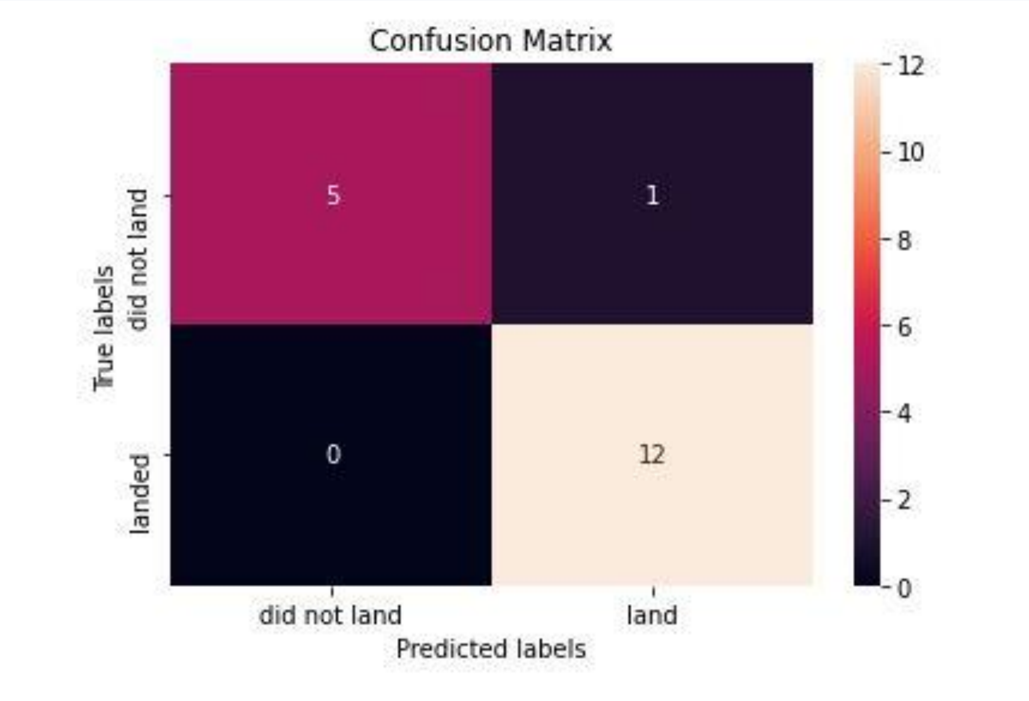
Classification Accuracy

- Four classification models were tested, and their accuracies were plotted into a bar plot.
- While all models give respectable accuracies, the model with the highest classification accuracy is Decision Tree Classifier.



Confusion Matrix

The confusion matrix for the decision tree model proves its accuracy by showing that it incorrectly predicted 1 outcome out of 18.



Conclusions

- Different data sources were analyzed, refining conclusions along the process.
- The best launch site is KSC LC-39A.
- Launches above 7,000kg are less risky.
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!