Winning Space Race with Data Science

Ghenu Malina-Nicoleta

28.10.2022

Outline

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

Executive Summary

Methodologies used for Data Analysis

- Data Collection using the SpaceX API and Webscraping & Data cleaning
- Exploratory Data Analysis (EDA), data wrangling, data visualization & interactive visual analytics
- Machine Learning Prediction

Summary of all results

- Machine Learning Prediction
- Collected data from public website
- Identified which features are the best at predicting successful outcomes
- By applying some basic statistical analysis and data visualization, I was able to see how variables are related to each other.
- Tested various models to predict which one performs best on the dataset

Introduction

- Our objective is to evaluate the viability of a new company, SpaceY, by analyzing existing company, SpaceX.
- By gathering information about SpaceX launches and determining which factors lead to a successful landing of the first stage, we can predict the price of launch for company SpaceY.

Methodology

Section 1

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX data was collected from two sources:
    - the SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - the SpaceX Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - The purpose of data wrangling was to find some patterns in the data and determine what would be the label for training supervised models and also convert the outcomes into Training Labels with '1' for successful landing and '0' for unsuccessful.

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Load the Spacex dataset
  - Sort and select the useful data for further analysis
- Perform interactive visual analytics using Folium and Plotly Dash
  - By displaying the data in a useful way, we can see which variable is useful for our purpose
  - With interactive maps we can mark the successful/failed launches for each location
- Perform predictive analysis using classification models
  - After standardizing the data and splitting it into train and test sets, we find the method that performs best on out test set

Data Collection

SpaceX data was collected from two sources:

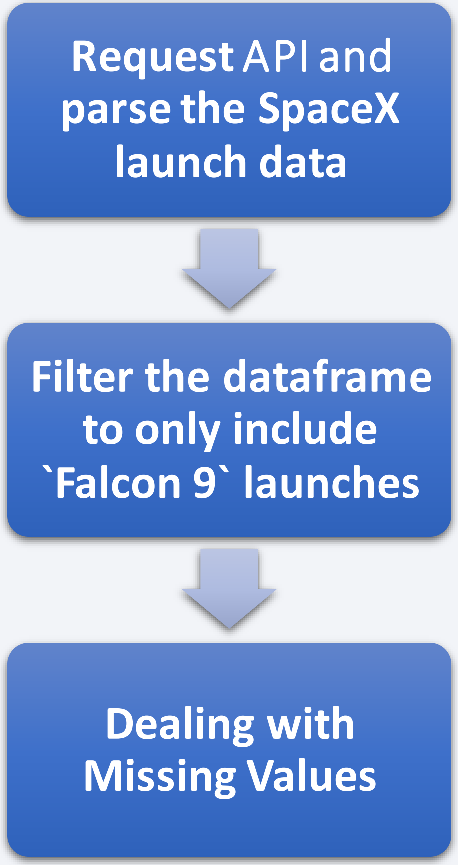
- The SpaceX API (https://api.spacexdata.com/v4/rockets/).
- The SpaceX Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

After requesting rocket launch data from SpaceX API with the URL, we clean the data, create a Dataframe and filter it to only include Falcon 9 launches.

Data Collection – SpaceX API

- SpaceX provides a public API, where data can be obtained and used.
- After requesting and parsing the SpaceX launch data using the GET request, the data was filtered and turned into a Pandas dataframe.
- The missing values in column "PayloadMass" were replaced with the column average value.

GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/1.%20Data%20Collection%20API.ipynb



Request API and parse the SpaceX launch data

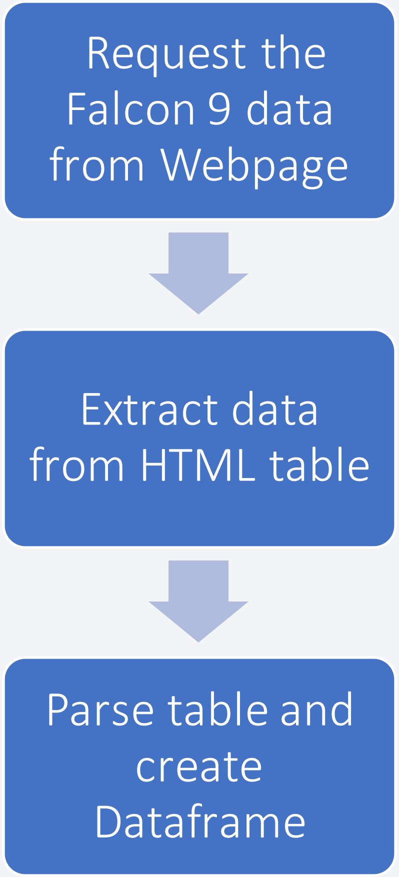
Filter the dataframe to only include `Falcon 9` launches

Dealing with Missing Values

Data Collection - Scraping

- The data was collected from the SpaceX Wikipedia page (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- The Falcon 9 launch records were extracted from a HTML table from Wikipedia.
- After parsing the table, I converted it into a Dataframe.

GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/2.%20jupyter-labs-webscraping.ipynb



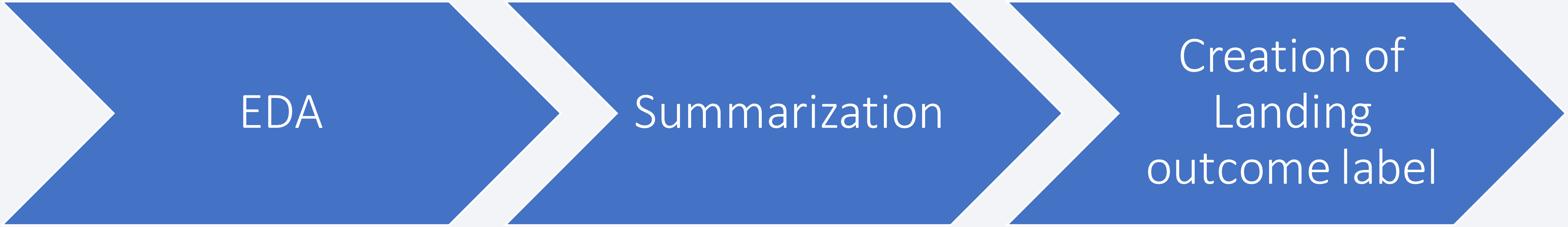
Request the Falcon 9 data from Webpage

Extract data from HTML table

Parse table and create Dataframe

Data Wrangling

- First, I dealt with the missing values by replacing them with the column average.
- I performed some Exploratory Data Analysis (EDA) on the dataset, to find some patterns in the data and determine what would be the label for training supervised models.
- Next, I converted the outcomes into Training Labels with `1` meaning the booster successfully landed, and `0` that it was unsuccessful.



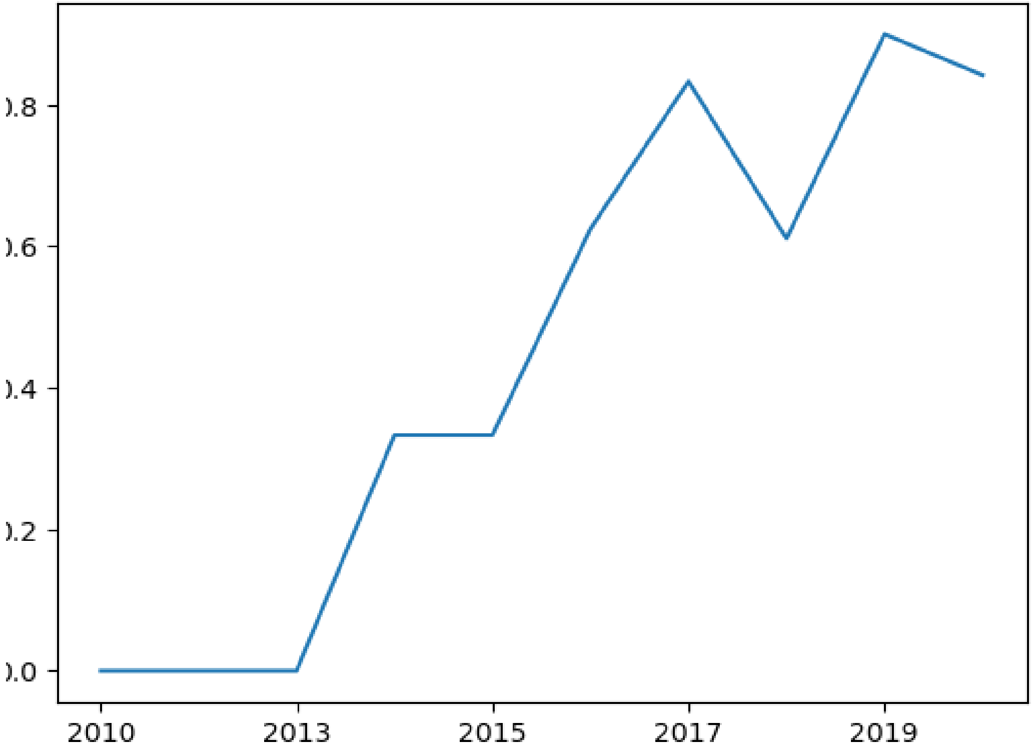
GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/3.%20labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

To explore data, scatter plots, line charts and barplots were used to visualize the relationship between pair of features:

- Flight Number & Payload
- Flight Number & Launch Site
- Payload & Launch Site
- Success rate & Orbit Type
- Flight number & Orbit type
- Payload & Orbit Type
- The Launch success over the years

- GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/5.%20jupyter-labs-eda-dataviz.ipynb



EDA with SQL

- The following SQL queries were performed:
  - Displayed the names of the unique launch sites
  - Top 5 records where launch sites begin with the string 'CCA'
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Date when the first succesful landing outcome in ground pad was acheived.
  - List of the successful booster names in drone ship that have payload mass between 4000 and 6000
  - List of successful and failed mission outcomes
  - List of boosters which have carried the maximum payload mass
  - Months in 2015 of failure landing outcomes in drone ship, booster versions and launch site
  - Rank of the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017.

GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/4.%20jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- When building an interactive map with Folium, I used markers, circles, lines and marker clusters.
  - Markers indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
  - Marker Clusters indicate groups of events in each coordinate
  - Lines are used to indicate distance

GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/6.%20lab_jupyter_launch_site_location%20(2).ipynb

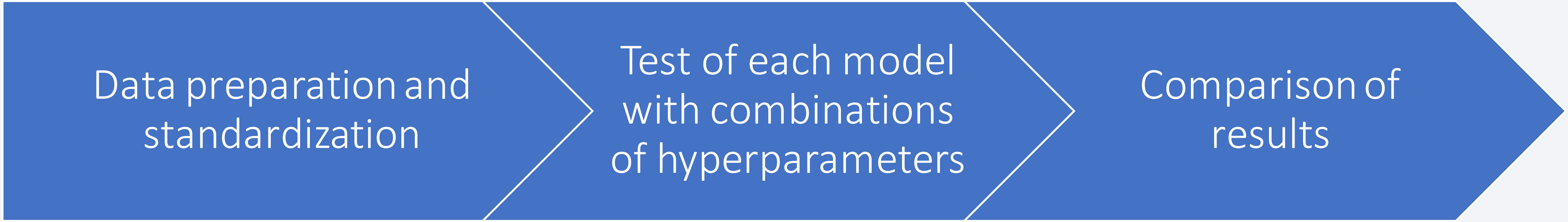
Build a Dashboard with Plotly Dash

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

GitHub URL: https://github.com/malinaghenu/Coursera/blob/4c91970d80d4e4f0bf2fb3ac54d623d3bbabc278/7.%20spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared:
  - Logistic Regression
  - Support Vector Machine (SVM)
  - Decision tree
  - K-nearest Neighbour



Data preparation and standardization

Test of each model with combinations of hyperparameters

Comparison of results

GitHub URL: https://github.com/malinaghenu/Coursera/blob/5c6411522ccbb8e7807e062a2f16001fd52f9295/8.%20SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb

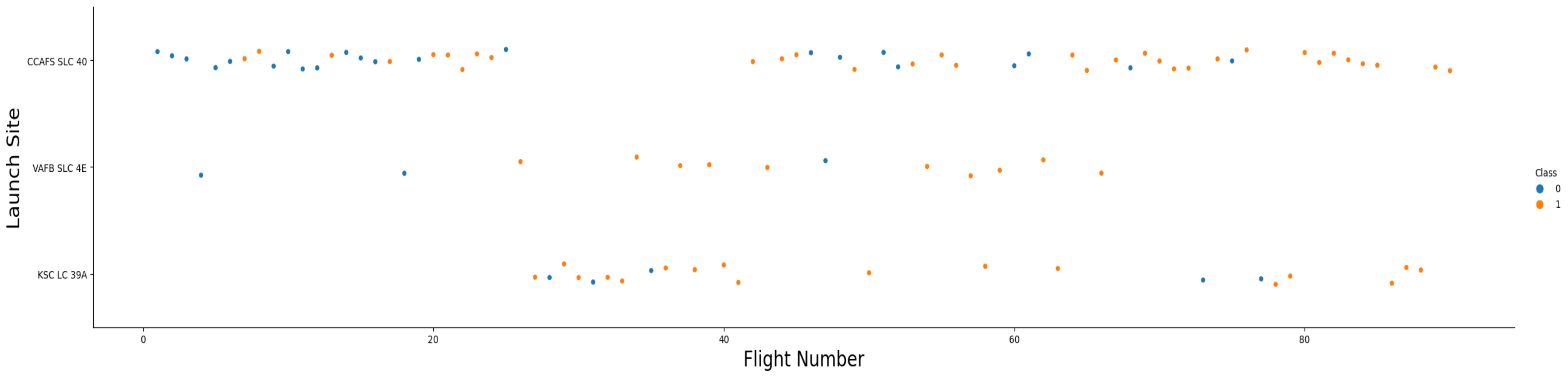
Results

- Exploratory data analysis results
  - SpaceX uses four different launch sites. In the maps created we can observe all of them being near the Ocean. We calculated and displayed:
    - distance to coastline,
    - distance to railways,
    - distance to highways and
    - distance to nearest towns.
  - The first successful landing was in 2015, five years after first launch
  - Almost 100% of mission outcomes were successful & two booster versions failed at landing
  - The number of successful outcomes continually improved

Insights drawn from EDA

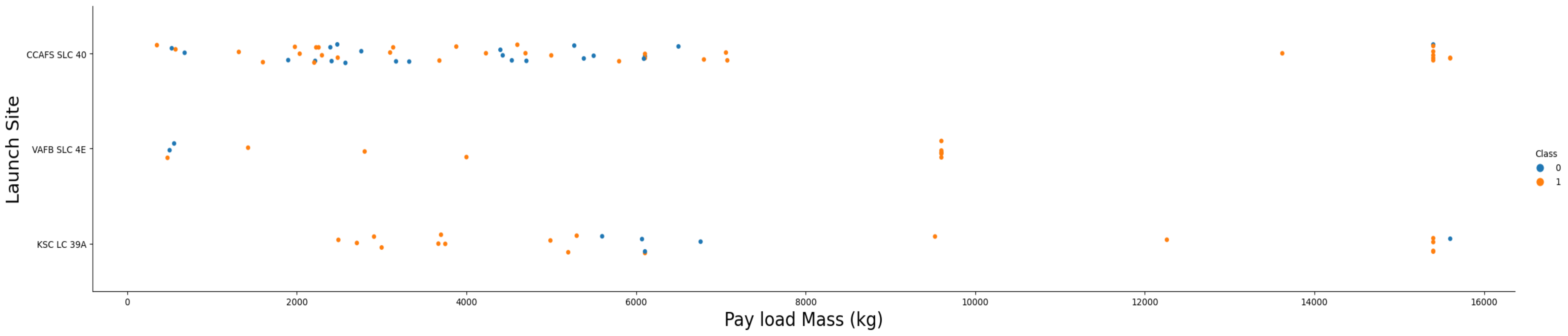
Section 2

Flight Number vs. Launch Site



- According to the scatter plot above, the most successful launch site is CCAFS SLC 40
- We can also observe that the success rate has improved over time

Payload vs. Launch Site

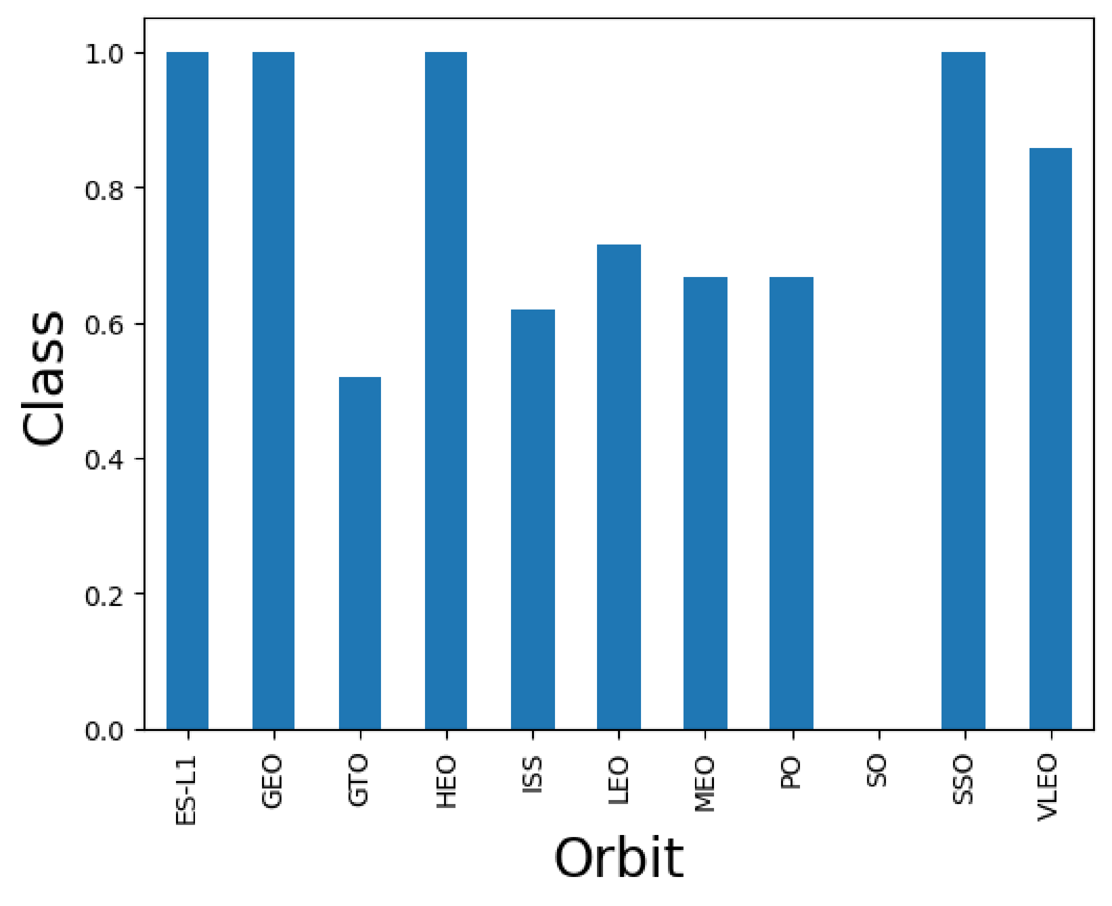


- In the scatter plot above we can see that higher payloads have better success rate.
- The success rate of payload mass above 8000kg is almost 100%.
- At the VAFB-SLC launchsite there were no rockets launched with a mass greater that 10 000kg.

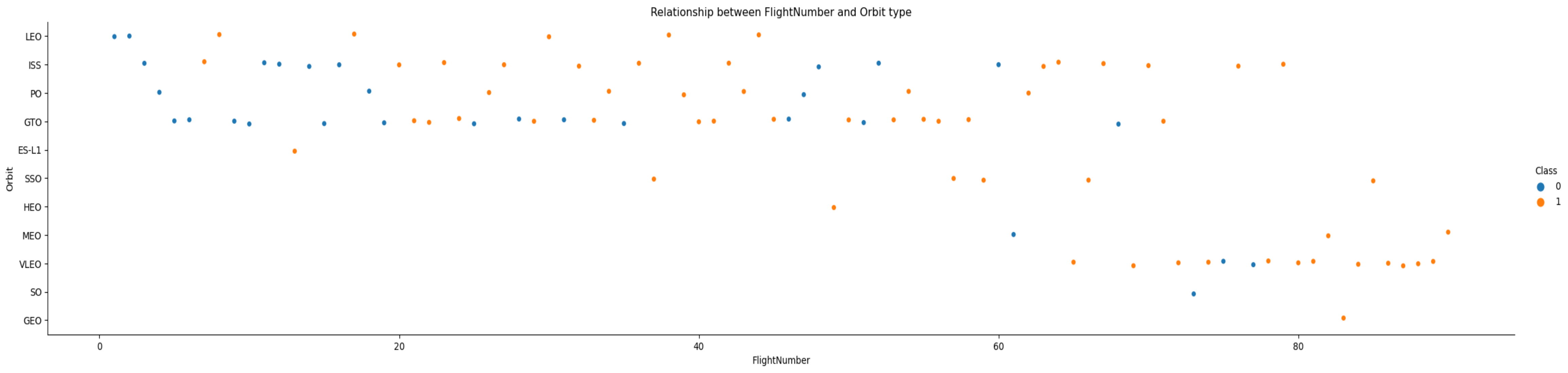
Success Rate vs. Orbit Type

The best success rate is for orbits ES-L1, GEO, HEO and SSO.

For orbit SO there was no successful landing.

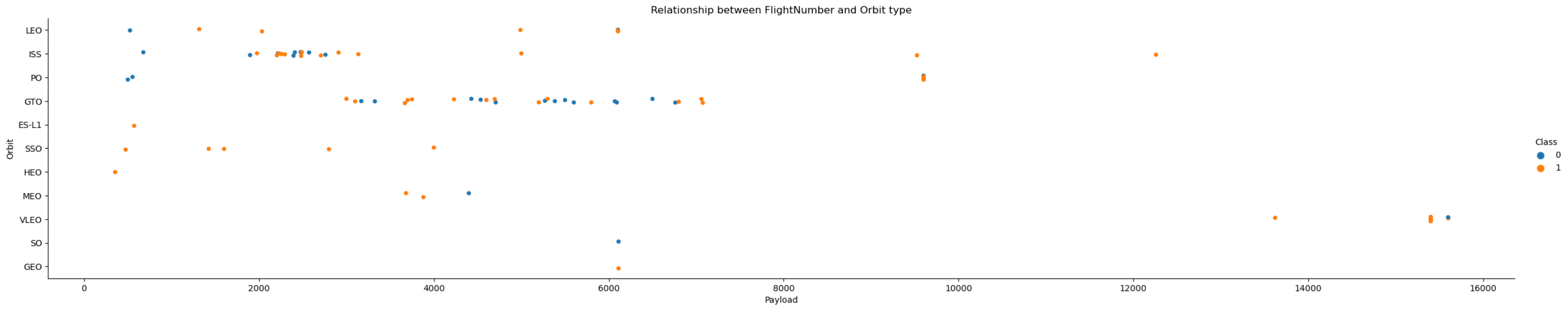


Flight Number vs. Orbit Type

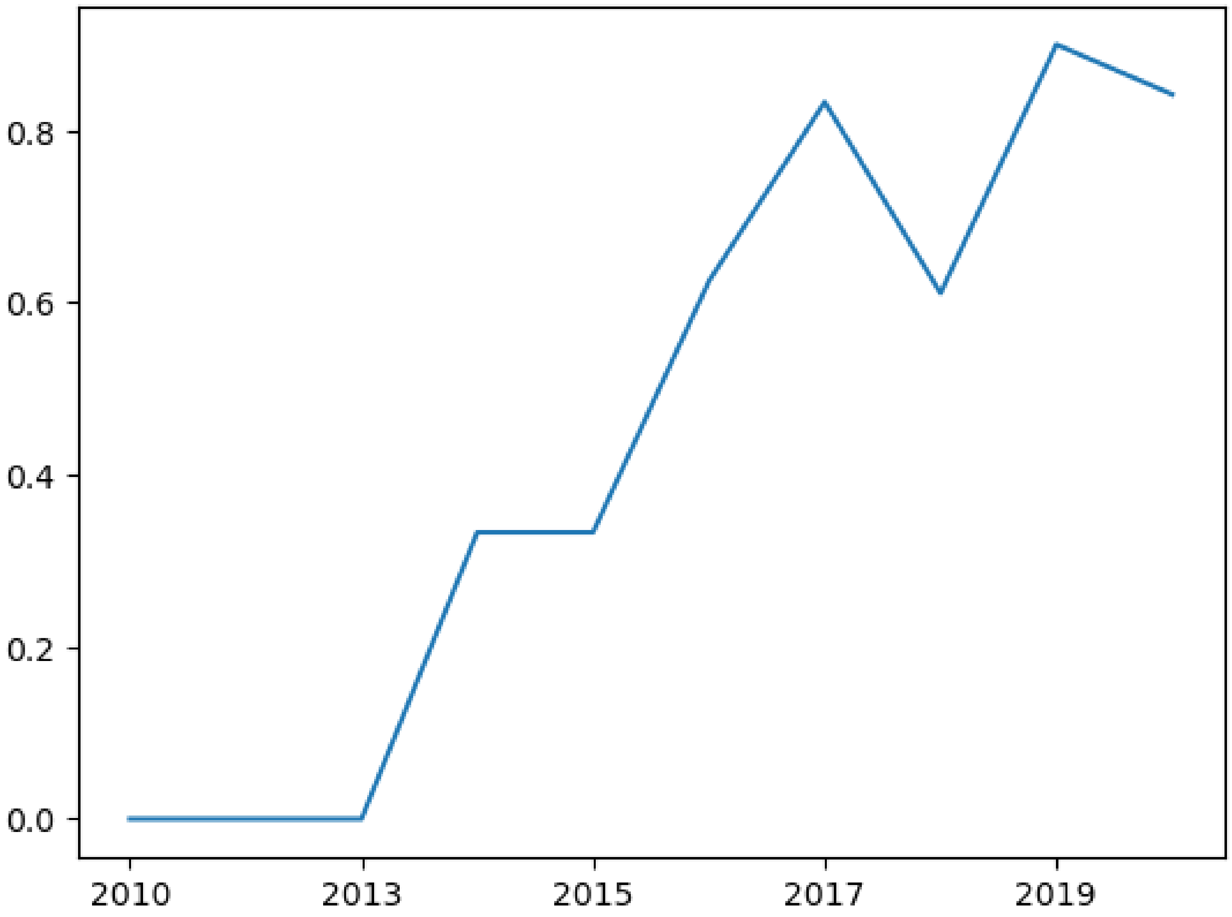


- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

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.

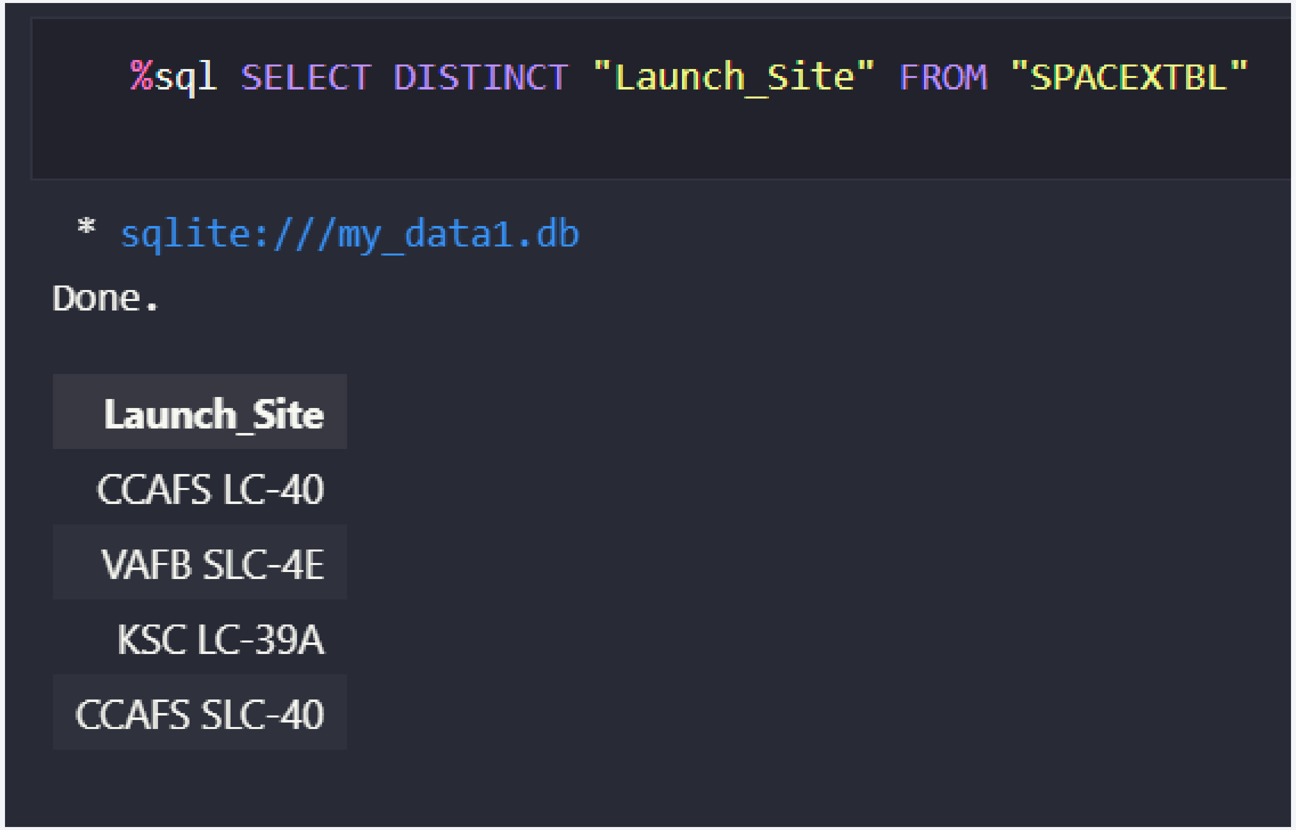


Launch Success Yearly Trend

- From the plot, we can oserve that the success rate continually increased since 2013.

All Launch Site Names

- I used the KeyWord DISTINCT, to only show the unique SpaceX Launch Sites.

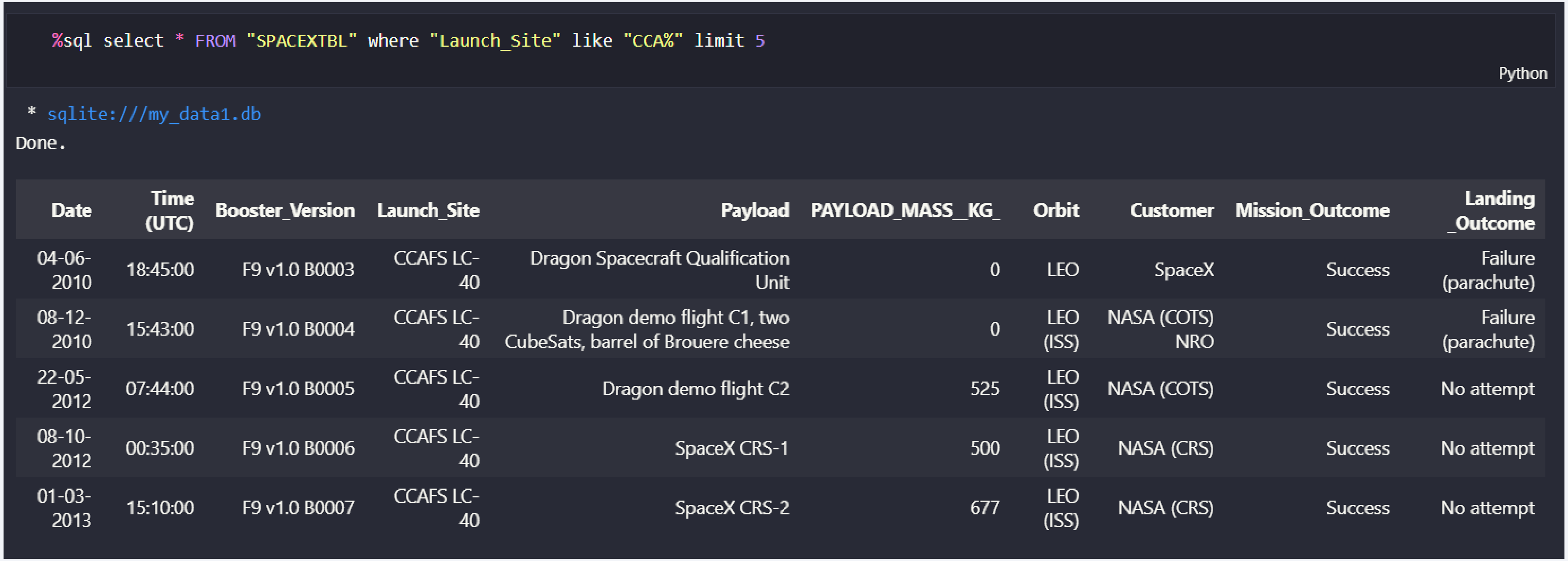


```
%sql SELECT DISTINCT "Launch_Site" FROM "SPACEXTBL"
```

```
 * 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'



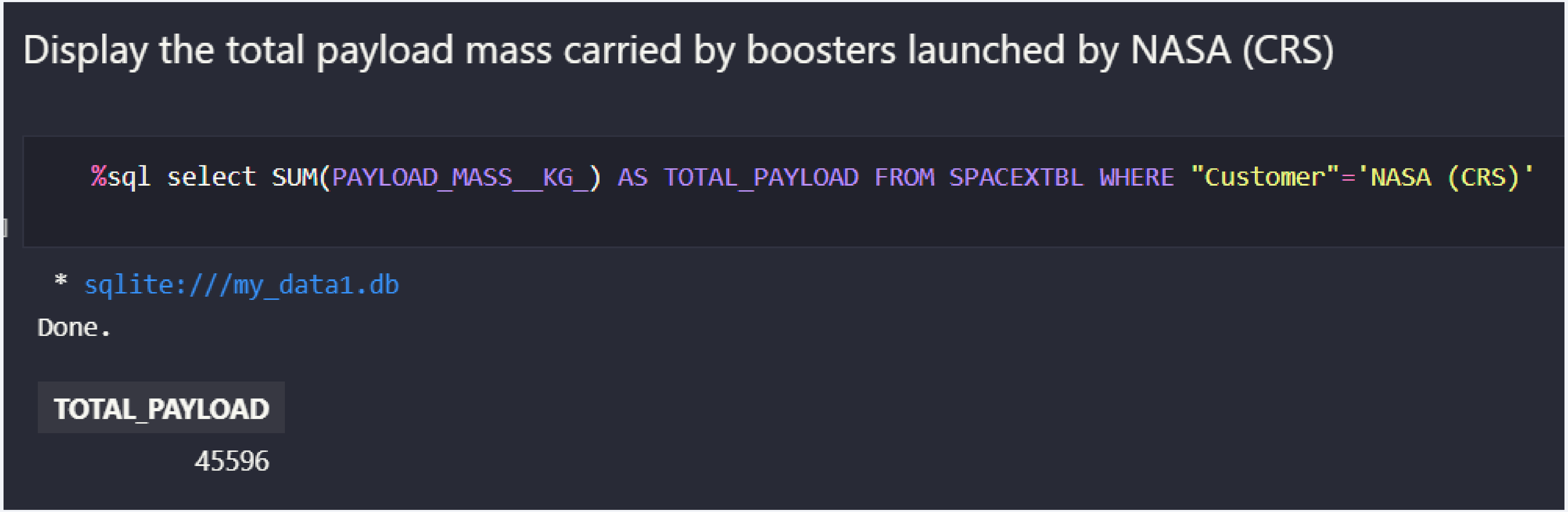
```
%sql select * FROM "SPACEXTBL" where "Launch_Site" like "CCA%" limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
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

- Find 5 records where launch sites begin with `CCA

Total Payload Mass



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

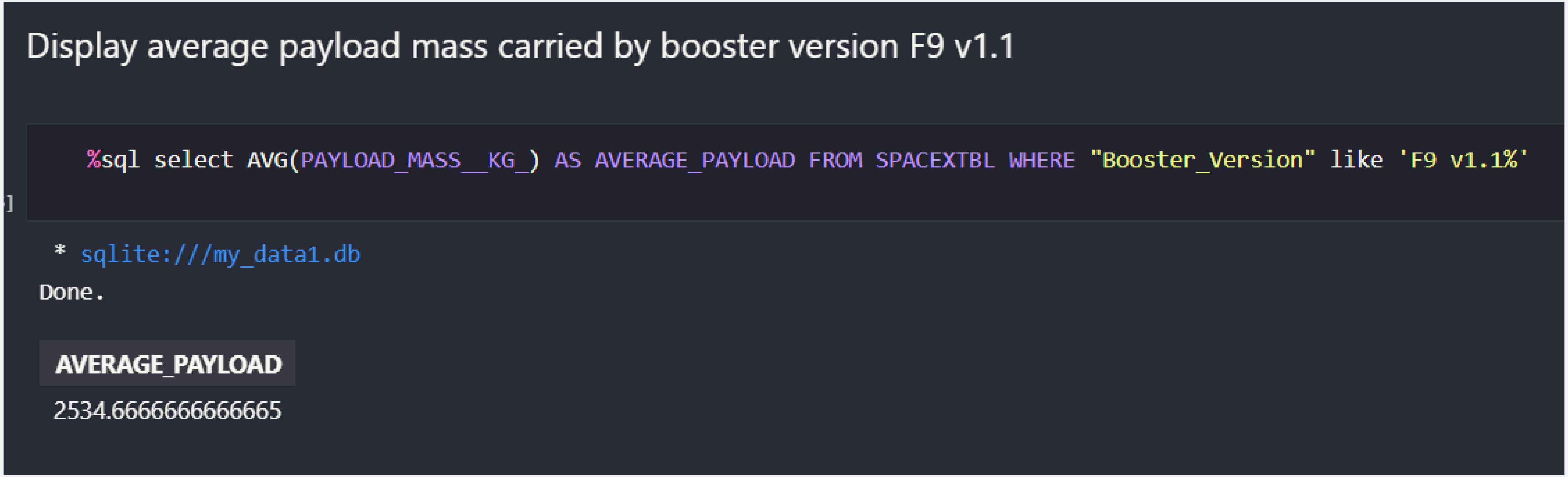
```
%sql select SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE "Customer"='NASA (CRS)'
```

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

TOTAL_PAYLOAD
45596

- The total payload carried by boosters from NASA is 45596.

Average Payload Mass by F9 v1.1



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

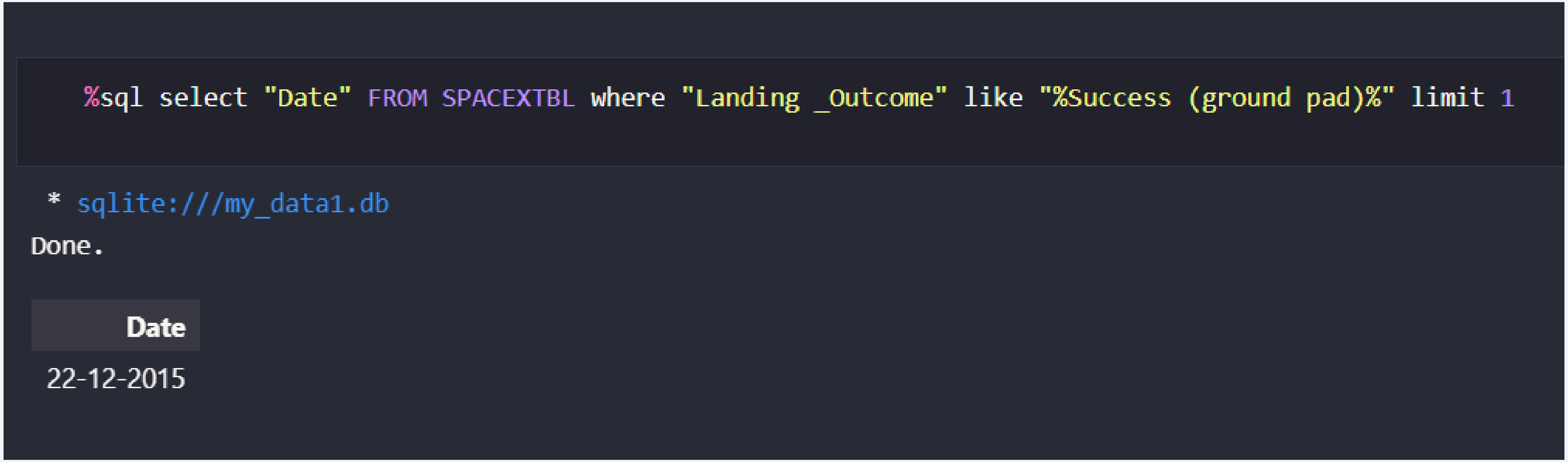
```
%sql select AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD FROM SPACEXTBL WHERE "Booster_Version" like 'F9 v1.1%'
```

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

AVERAGE_PAYLOAD
2534.6666666666665

- The average payload mass carried by booster version F9 v1.1 is 2534.6

First Successful Ground Landing Date



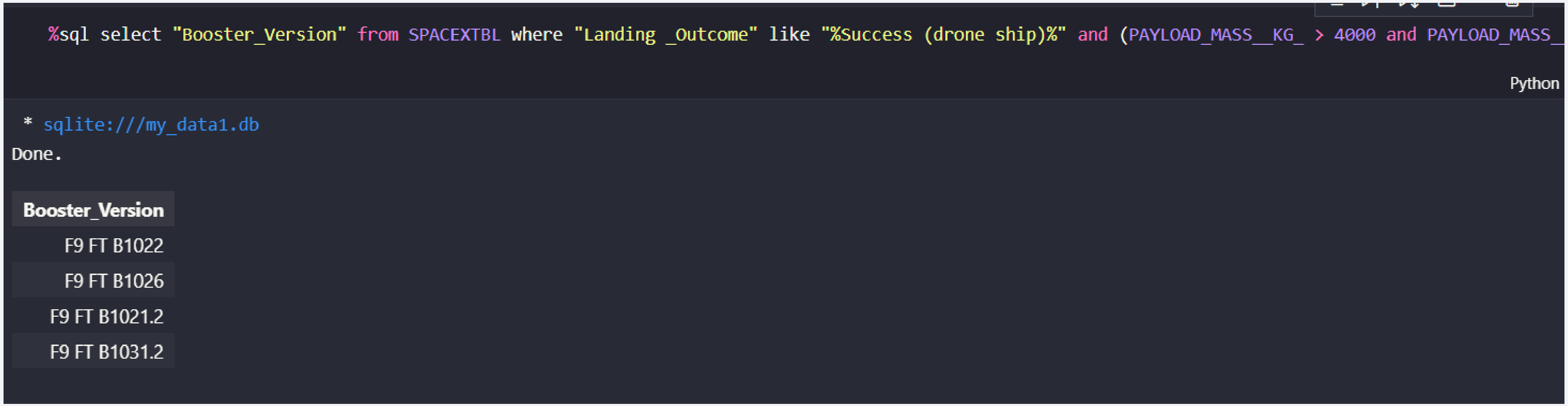
```
%sql select "Date" FROM SPACEXTBL where "Landing _Outcome" like "%Success (ground pad)%" limit 1
```

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

Date
22-12-2015

- The date of the first successful landing outcome on ground pad was 22-12-2015

Successful Drone Ship Landing with Payload between 4000 and 6000



```
%sql select "Booster_Version" from SPACEXTBL where "Landing _Outcome" like "%Success (drone ship)%" and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS_
```

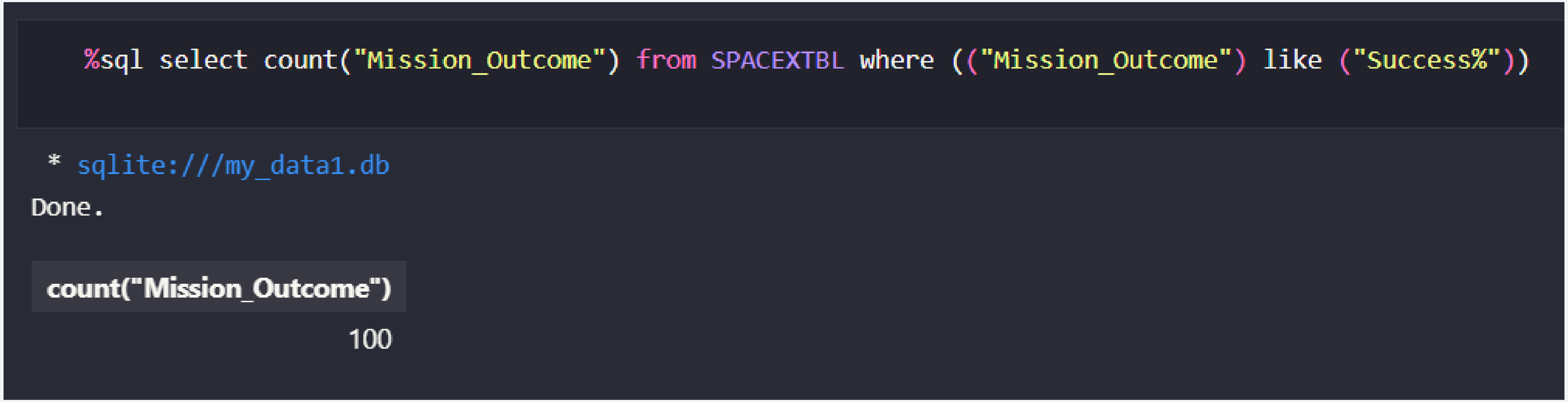
```
 * sqlite:///my_data1.db
Done.
```

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

- List of the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000. Full code:

%sql select "Booster_Version" from SPACEXTBL where "Landing _Outcome" like "%Success (drone ship)%" and (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)

Total Number of Successful and Failure Mission Outcomes



```
%sql select count("Mission_Outcome") from SPACEXTBL where (("Mission_Outcome") like ("Success%"))
```

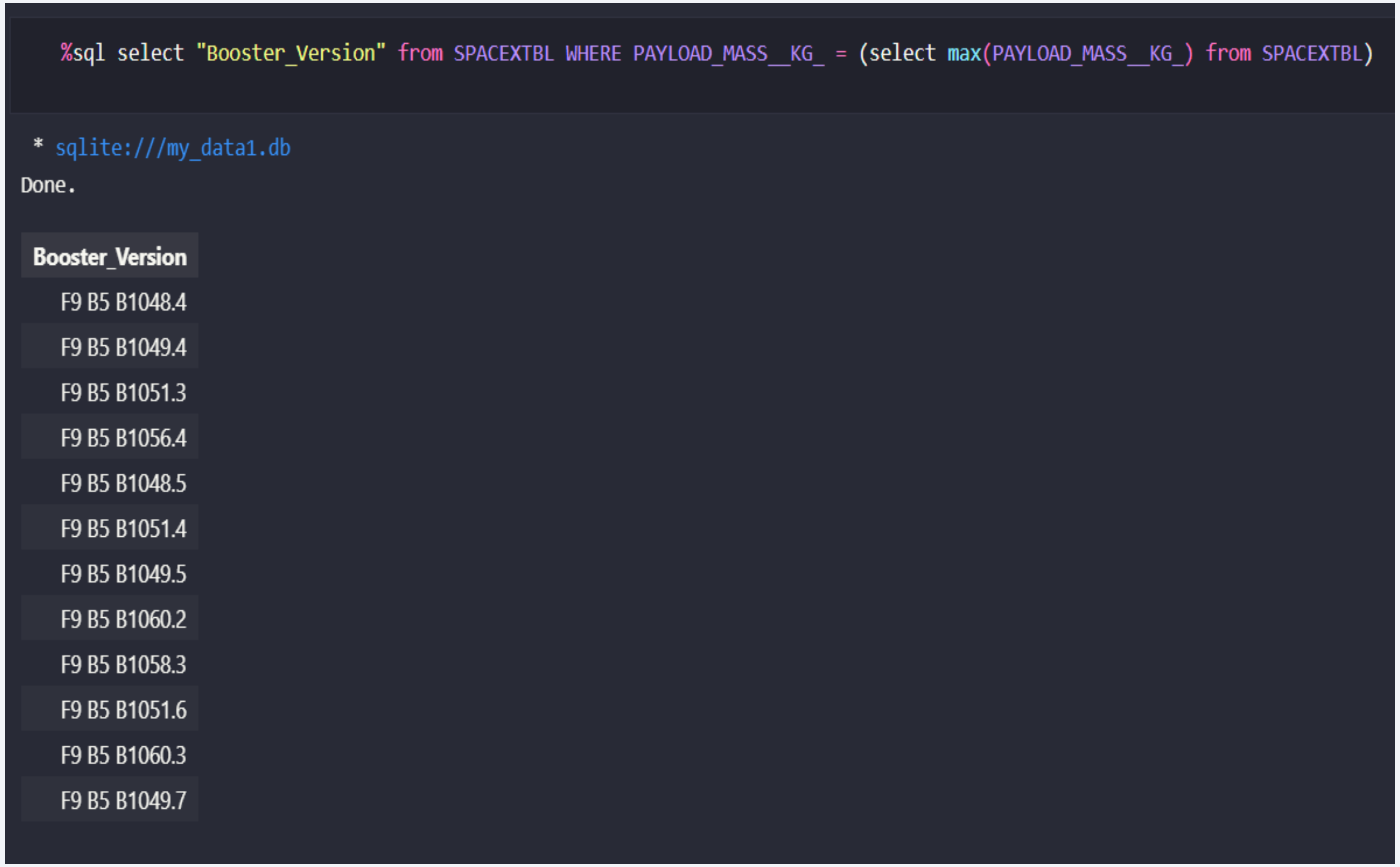
```
 * sqlite:///my_data1.db
Done.
```

count("Mission_Outcome")
100

- There was a total of 100 successful mission outcomes and only one failure.

Boosters Carried Maximum Payload

- We determined the list of the names of the booster which have carried the maximum payload mass using a subquery and the MAX() function.

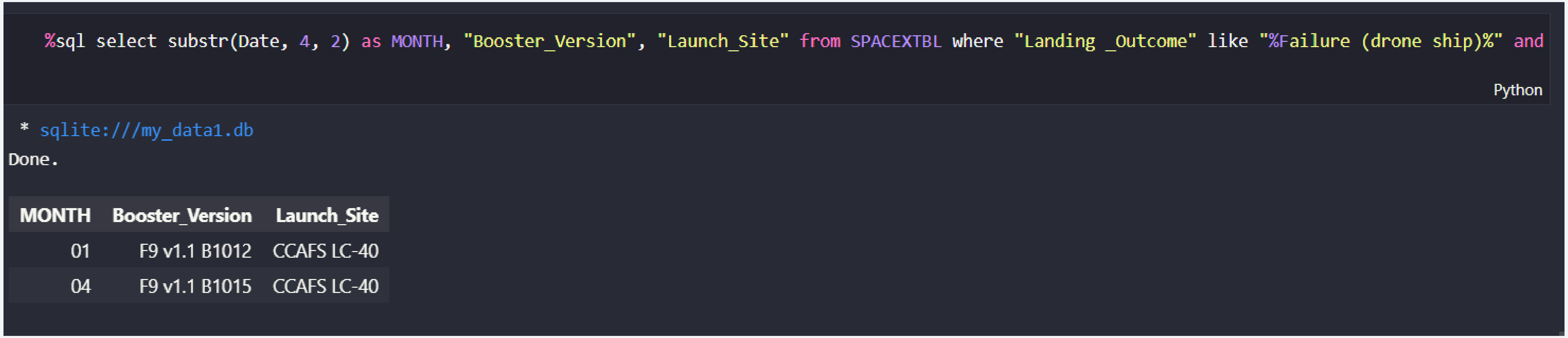


```
%sql select "Booster_Version" from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

```
 * 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



```
%sql select substr(Date, 4, 2) as MONTH, "Booster_Version", "Launch_Site" from SPACEXTBL where "Landing _Outcome" like "%Failure (drone ship)%" and
```

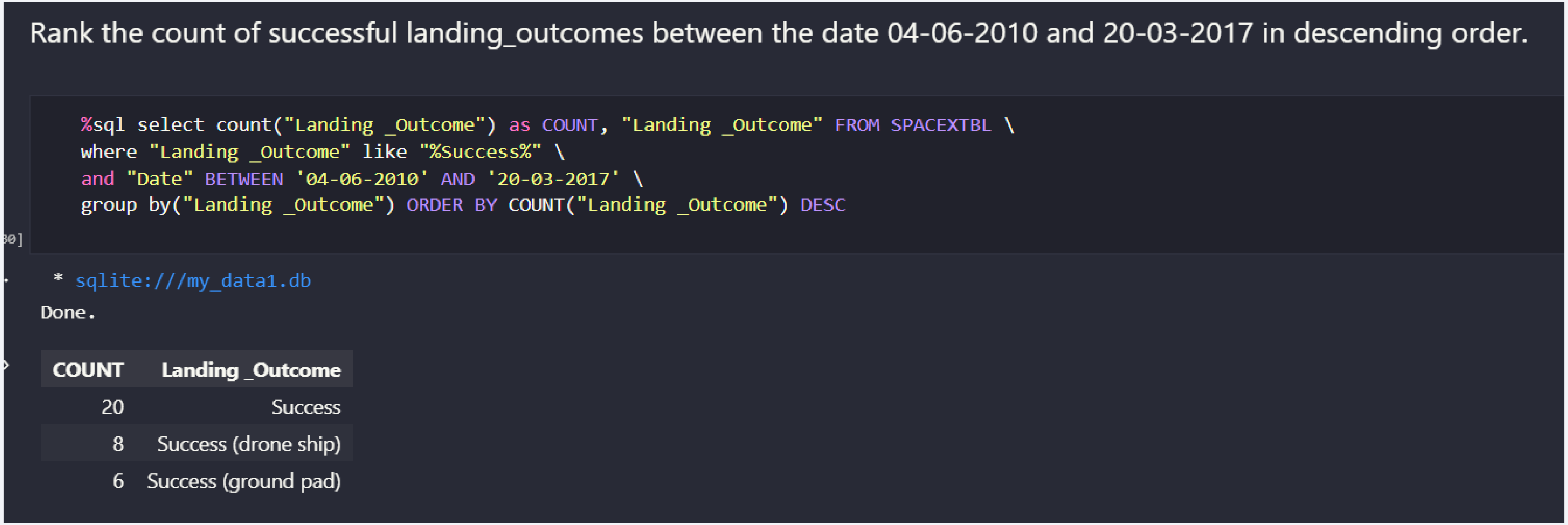
```
 * sqlite:///my_data1.db
Done.
```

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- I determined the list of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, using the WHERE clause, LIKE, AND and BETWEEN conditions, to filter the results. Full code is:

%sql select substr(Date, 4, 2) as MONTH, "Booster_Version", "Launch_Site" from SPACEXTBL where "Landing _Outcome" like "%Failure (drone ship)%" and substr(Date,7,4)='2015'

Rank Successful 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.

```
%sql select count("Landing _Outcome") as COUNT, "Landing _Outcome" FROM SPACEXTBL \
where "Landing _Outcome" like "%Success%" \
and "Date" BETWEEN '04-06-2010' AND '20-03-2017' \
group by("Landing _Outcome") ORDER BY COUNT("Landing _Outcome") DESC
```

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

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

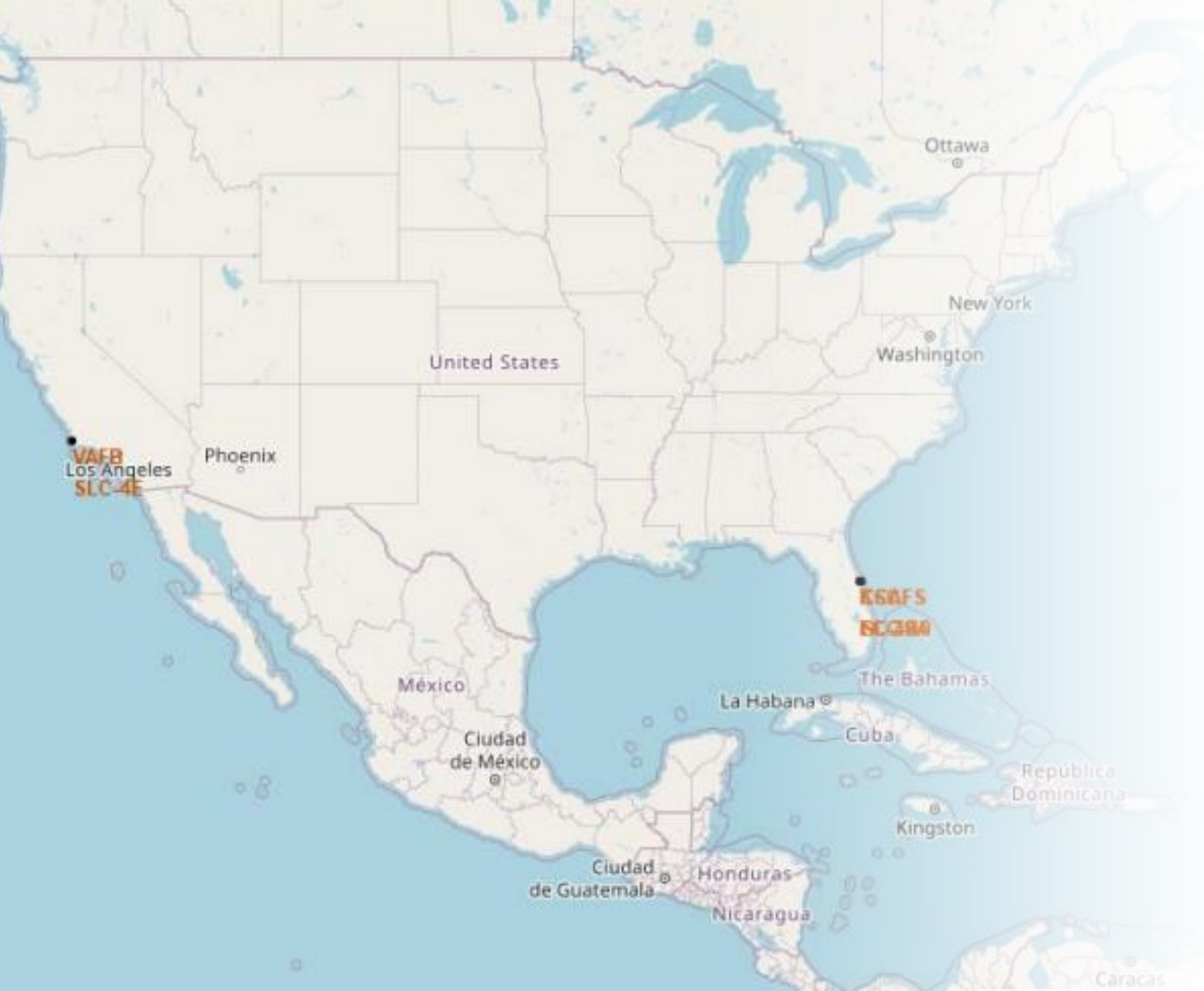
I ranked the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, by creating a new column named "COUNT" that used the function count from Landing_Outcome column. Then, I filtered the results to show only the successful outcome between the desired dates, grouped the outcomes and sorted them in descending order.

*task asked for successful landing outcomes

Launch Sites Proximities Analysis

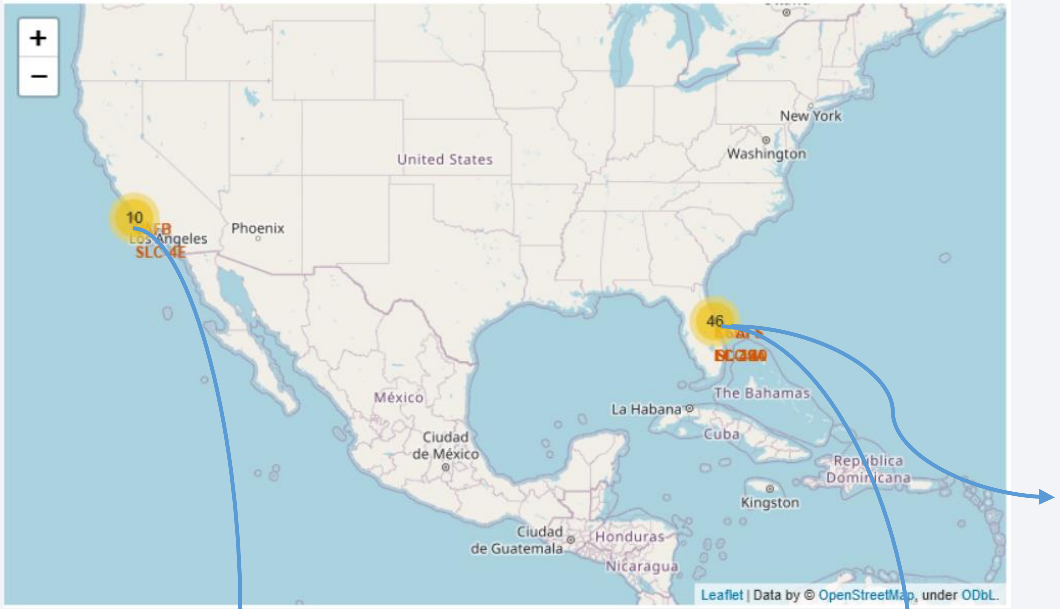
Section 3

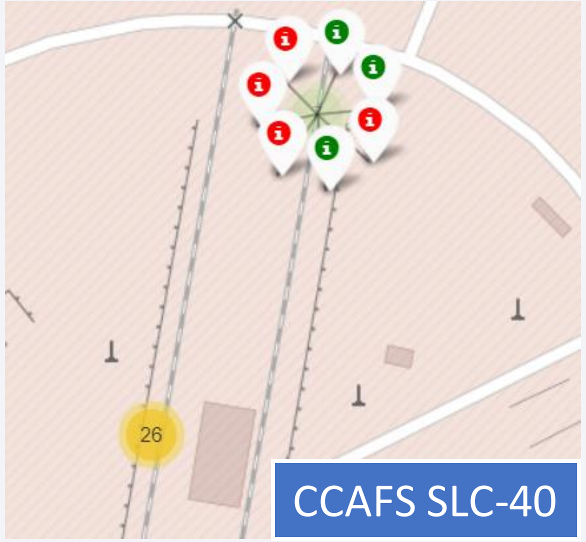
Launch Sites



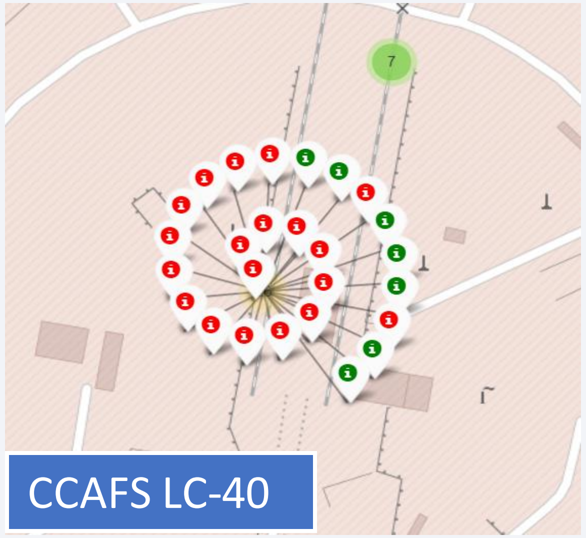
We can see that all Launching Sites are in the USA, on the coast.

SUCCESS/FAILED LAUNCHES FOR EACH SITE

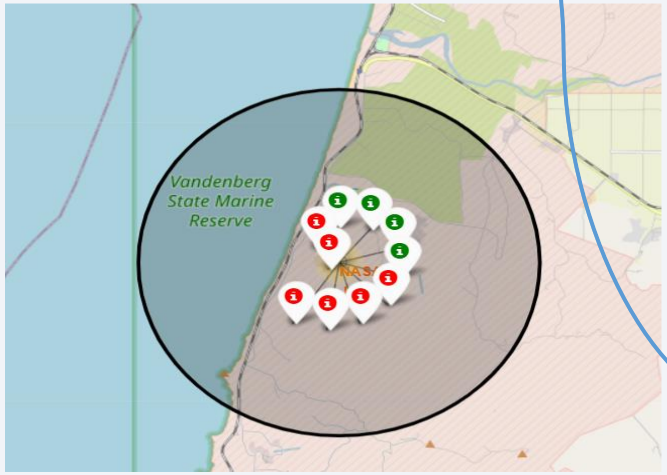




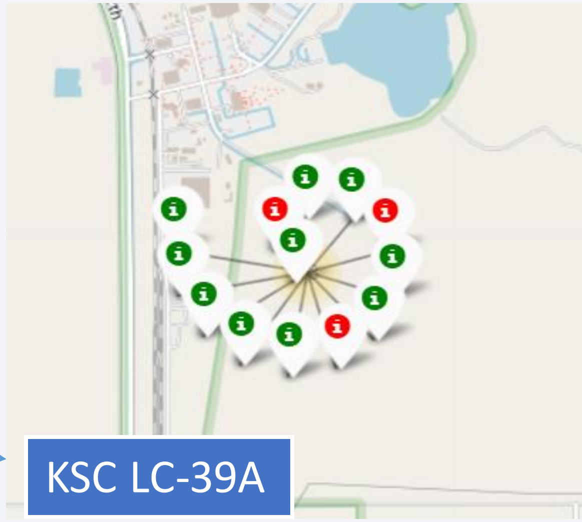
CCAFS SLC-40



CCAFS LC-40



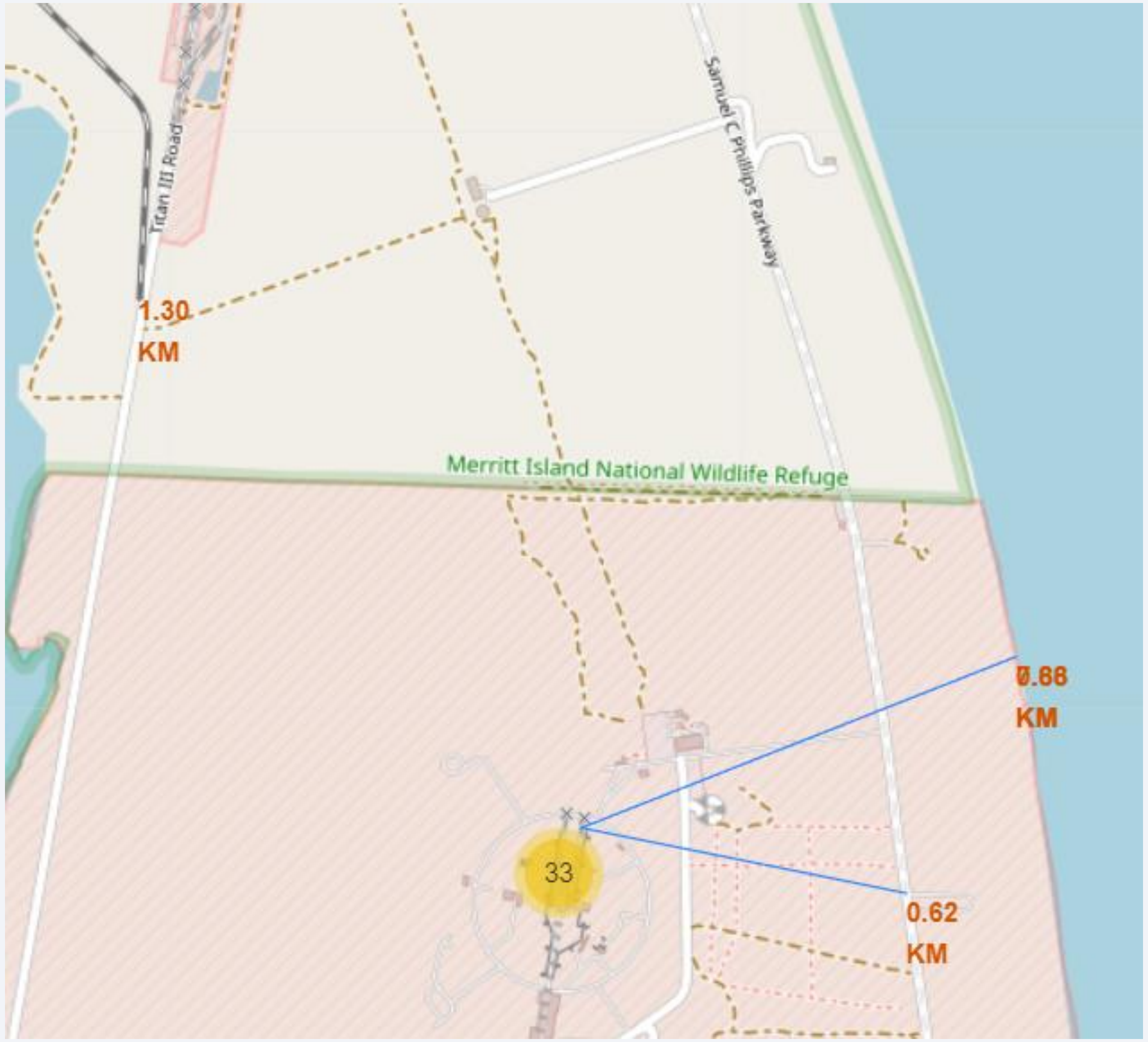
VAFB SLC-4E



KSC LC-39A

Launches have been grouped into clusters and annotated with green markers for successful landings and red markers for unsuccessful.

PROXIMITY OF LAUNCH SITES TO OTHER POINTS OF INTEREST

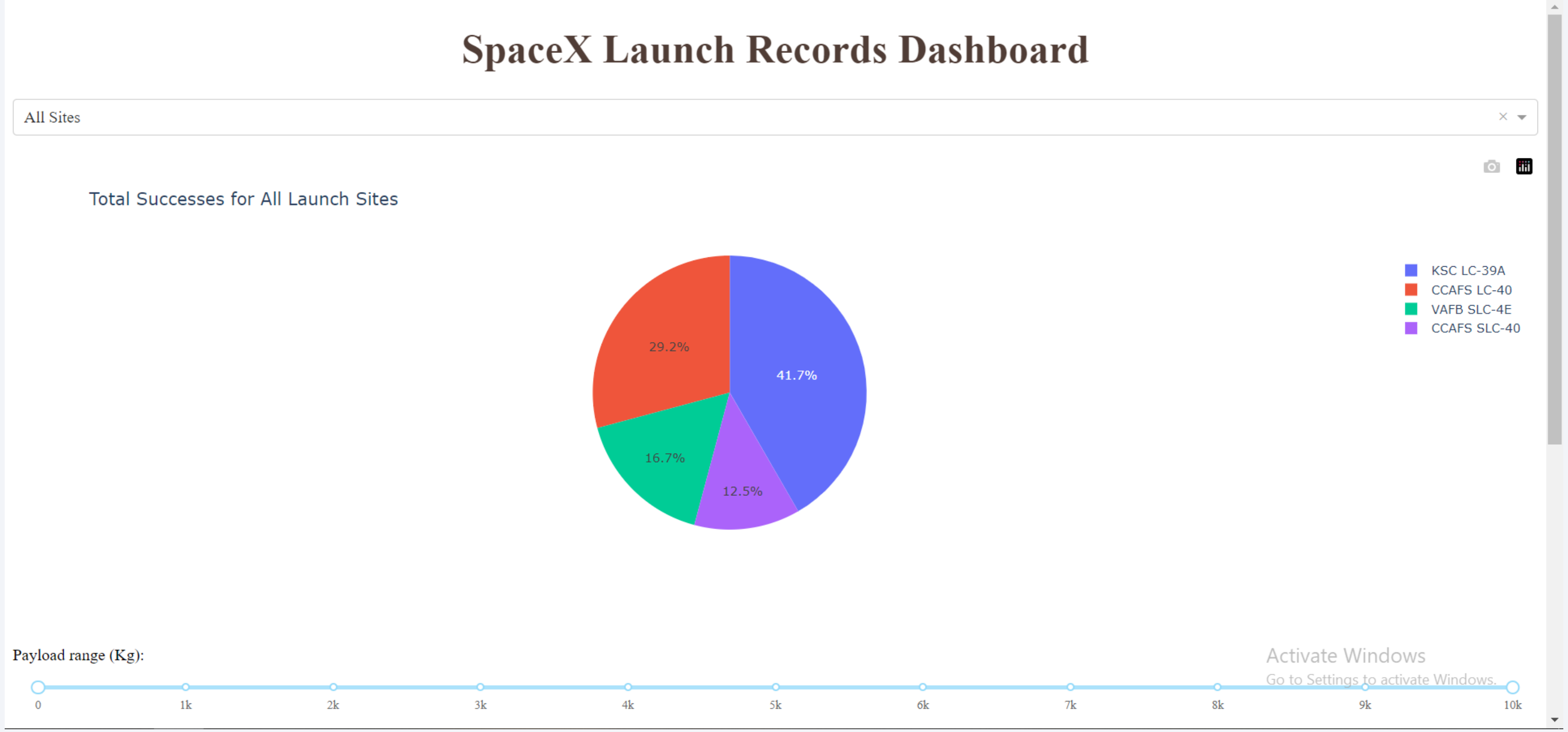

Titan III Road
Samuel C Phillips Parkway
1.30 KM
Merritt Island National Wildlife Refuge
0.66 KM
33
0.62 KM

- Using the CCAFS SLC-40 launch site as an example site, we can understand more about the placement of launch sites.
- On the we can see that the CCAFS SLC-40 distance to coastline is 0.66 km, to highway 0,62km and to the closest railway, 1,30km.
- By creating interactive maps, we found out that Launch sites are near coastline, highway and railway but keep a certain distance from the cities.
- The CCAFS distance to Cape Canaveral is 18.2 km.

Build a Dashboard with Plotly Dash

Section 4

LAUNCH SUCCESS COUNT FOR ALL SITES



SpaceX Launch Records Dashboard

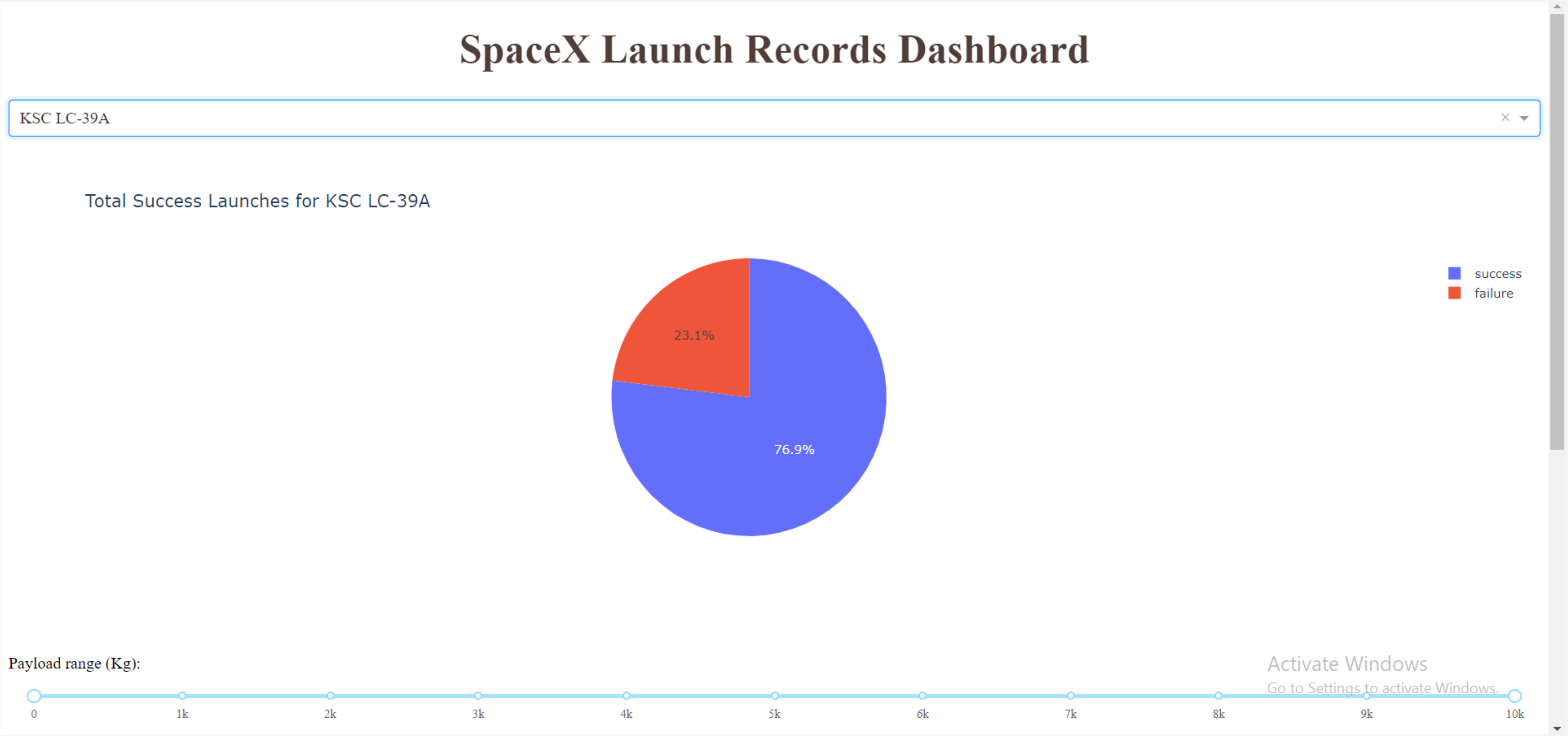
All Sites

Total Successes for All Launch Sites

Payload range (Kg):

The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches.

PIE CHART FOR THE LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

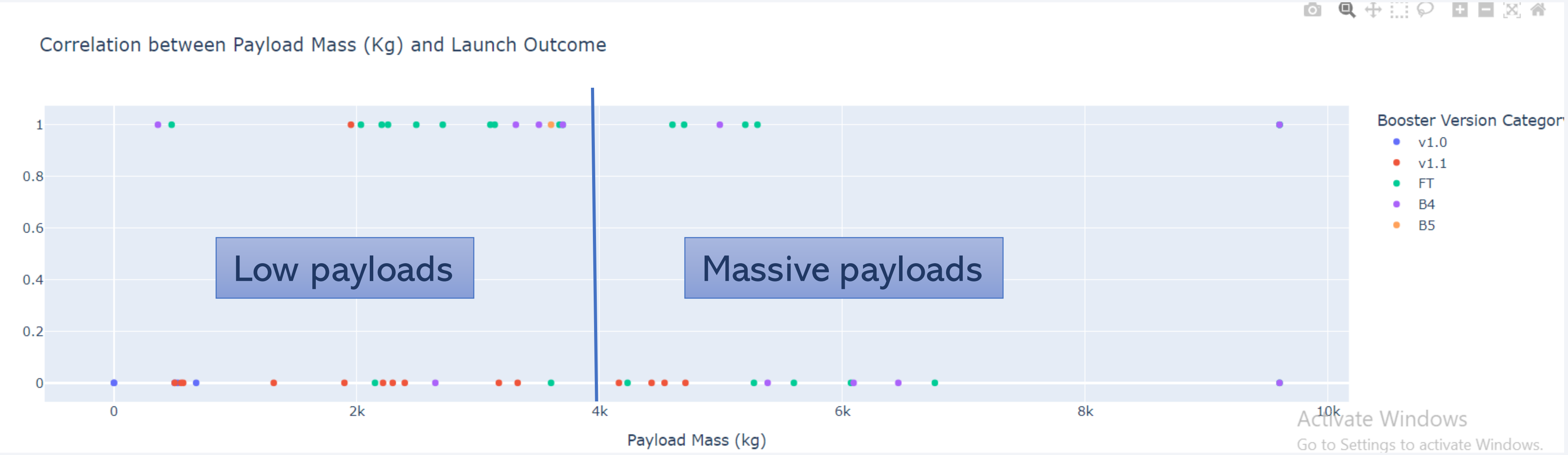


SpaceX Launch Records Dashboard

KSC LC-39A

- The launch site KSC LC-39 A also had the highest rate of successful launches, with a 76.9% success rate.

LAUNCH OUTCOME VS. PAYLOAD SCATTER PLOT FOR ALL SITES



Plotting the launch outcome vs. payload for all sites shows a gap around 4000 kg, so it makes sense to split the data into 2 ranges:

- 0 – 4000 kg (low payloads)
- 4000 – 10000 kg (massive payloads)

From these 2 plots, it can be shown that the success for massive payloads is lower than that for low payloads.

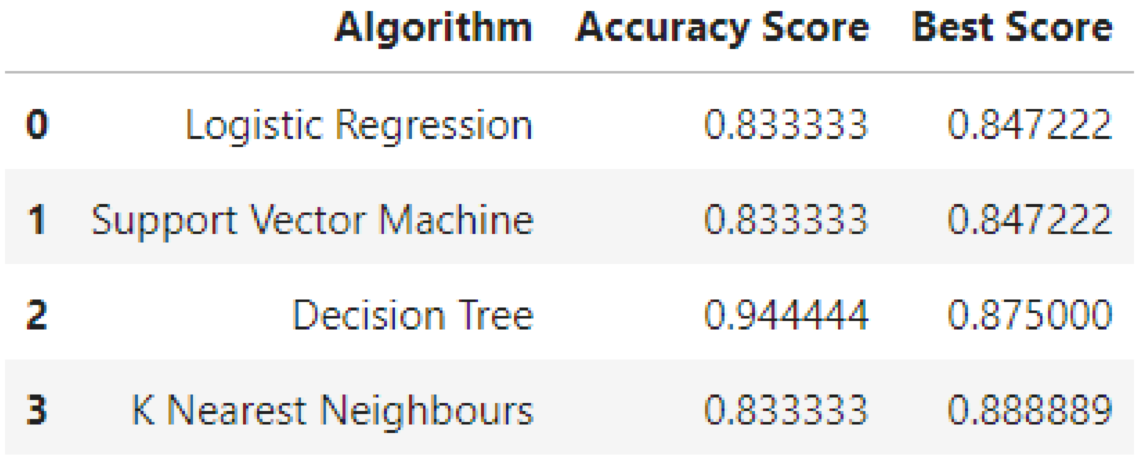
It is also worth noting that some booster types (v1.0 and B5) have not been launched with massive payloads.

Predictive Analysis (Classification)

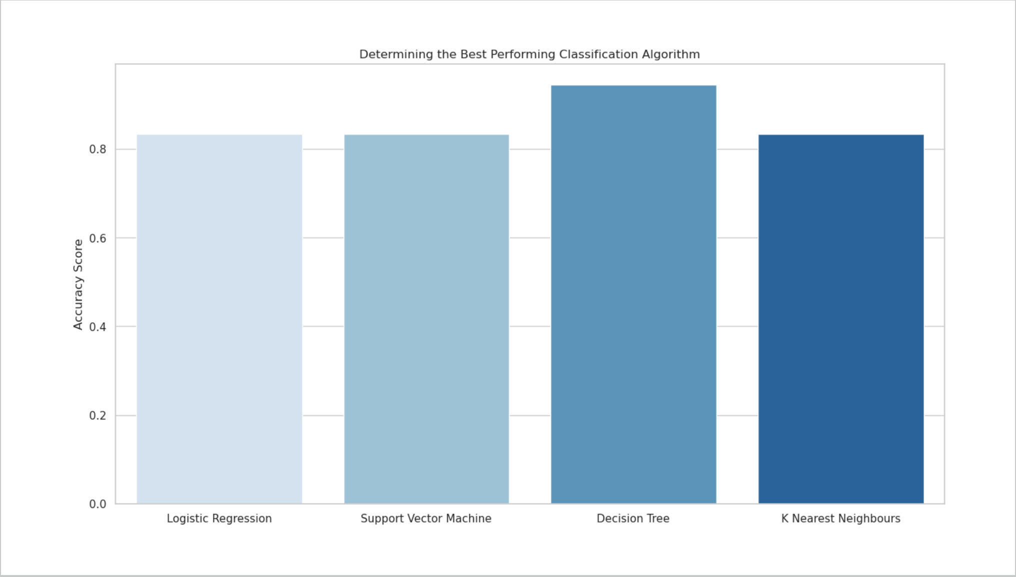
Section 5

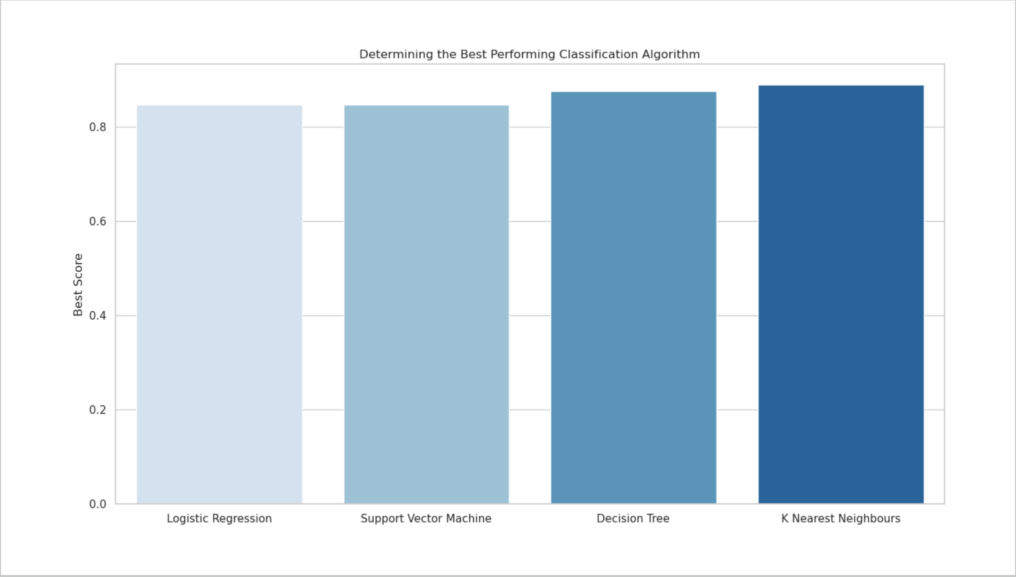
Classification Accuracy

- Plotting the Accuracy Score and Best Score for each classification algorithm produces the following result:
  - The Decision Tree model has the highest classification accuracy
  - The Accuracy Score is 94.44%
  - The Best Score is 90.36%



	Algorithm	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.847222
1	Support Vector Machine	0.833333	0.847222
2	Decision Tree	0.944444	0.875000
3	K Nearest Neighbours	0.833333	0.888889



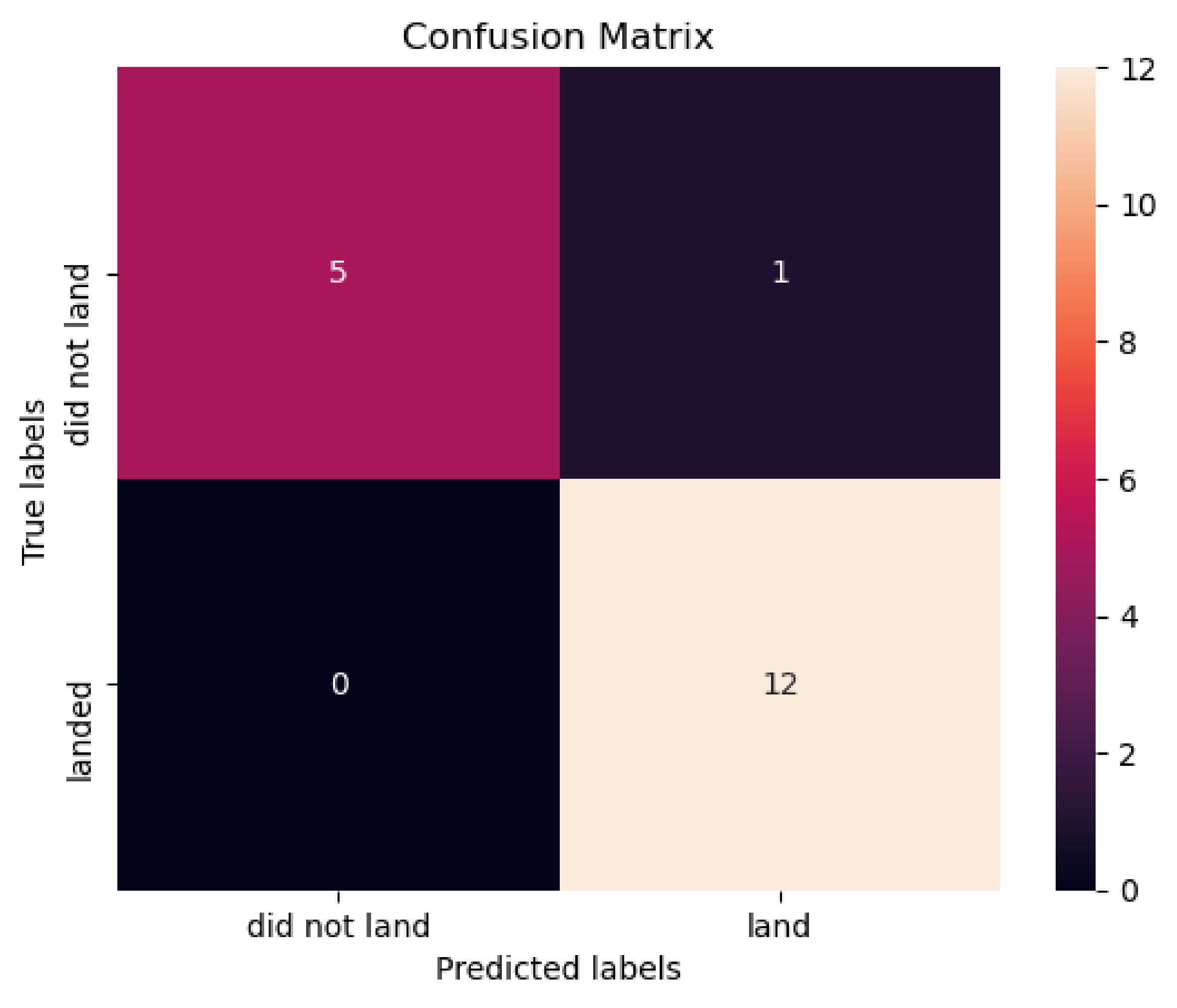


Confusion Matrix

As shown previously, best performing classification model is the Decision Tree model, with an accuracy of 94.44%.

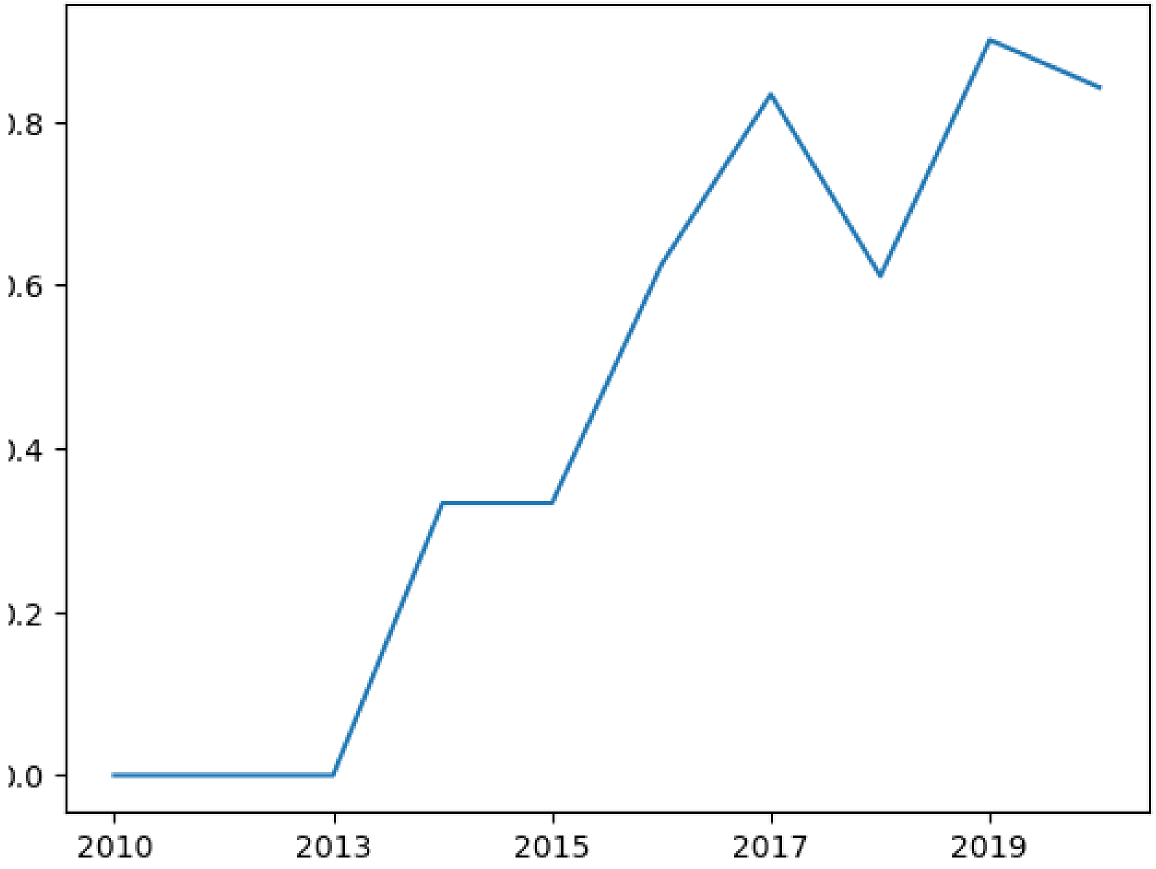
This is explained by the confusion matrix, which shows only 1 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).

The other 17 results are correctly classified (5 did not land, 12 did land).



Conclusions

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful. I.e. with more experience, the success rate increases.
  - Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
  - After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
  - After 2016, there was always a greater than 50% chance of success



Conclusions

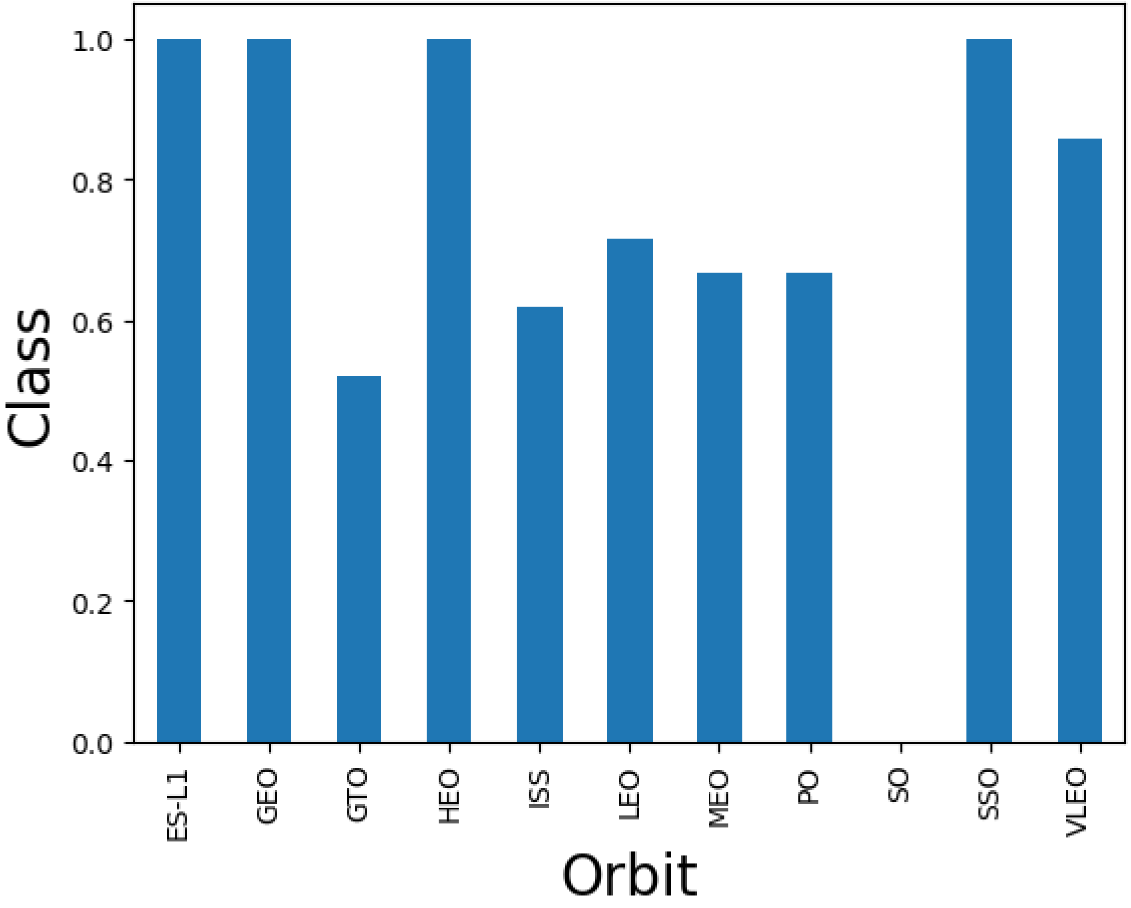
- Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.

- The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.

- The 100% success rate in SSO is more impressive, with 5 successful flights.

- The orbit types PO, ISS, and LEO, have more success with heavy payloads:

- VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.



Conclusions

- The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.
- The success for massive payloads (over 4000kg) is lower than that for low payloads.
- The best performing classification model is the Decision Tree model, with an accuracy of 94.44%.

Appendix

- Folium didn’t show maps on Github, so I took screenshots.

Thank you!