IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Estefânia Carvalho

5th August 2024

Outline

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

Executive Summary

In this capstone project, our objective was to predict the likelihood of a successful landing of the Falcon 9 first stage. To achieve this, we employed several methodologies. We began with data collection, gathering comprehensive data on Falcon 9 launches, including various parameters affecting the first stage's landing success, from multiple sources using webscraping and a web API. After data wrangling, Exploratory Data Analysis (EDA) was conducted using data visualization techniques to understand the distribution and relationships within the data. Further exploration was done using SQL queries to facilitate complex aggregations and enable deeper insights. An interactive map was created using Folium and a dashboard built with Plotly Dash for interactive analytics, allowing dynamic exploration of the data. Finally, predictive analysis was carried out to determine the likelihood of a successful landing of the Falcon 9 first stage.

Our analyses revealed some features to be correlated with launch outcomes and all machine learning algorithms used to predict the Falcon 9 first stand landing have similar high accuracies.

Introduction

SpaceX's Falcon 9 rocket has transformed space travel by significantly reducing launch costs through the reuse of its first stage. While other providers charge around $165 million per launch, SpaceX offers a competitive price of $62 million, primarily due to the ability to successfully land and reuse the first stage.

We aim to develop a predictive model to estimate the likelihood of a Falcon 9 first stage landing successfully. This information can help other companies formulate competitive bids against SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scraping Wikipedia tables
  - SpaceX Rest API
- Perform data wrangling
  - One Hot Encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linear Regression, K-Nearest Neighbors, Support Vector Machine and Decision Tree were tested and compared

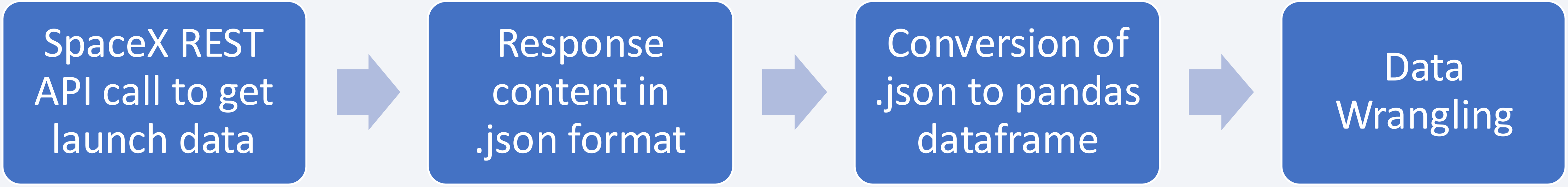
Data Collection

Data were collected using web scraping and a Rest API.

The SpaceX Rest API was used to gather information of launch data that included several features like type of rocket used, payload delivered and landing outcome. Data retrieved this way were initially in .json format but were converted to a pandas dataframe.

Other Falcon 9 launch data was retrieved by web scraping Wikipedia using Beautiful Soup.

Data Collection – SpaceX API



SpaceX REST API call to get launch data

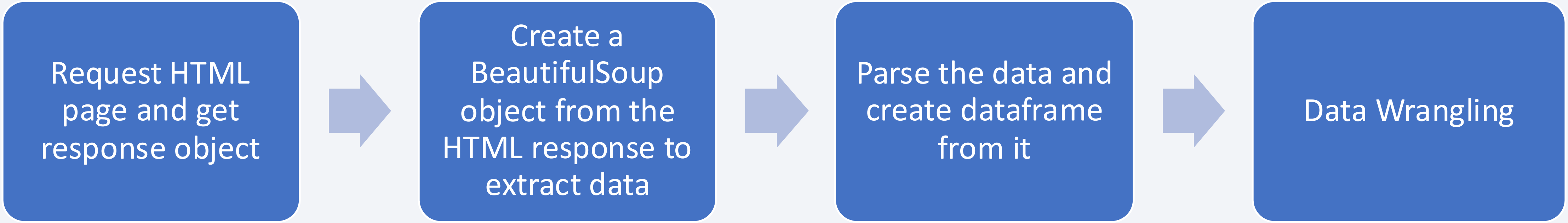
Response content in .json format

Conversion of .json to pandas dataframe

Data Wrangling

Completed SpaceX API calls notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Request HTML page and get response object

Create a BeautifulSoup object from the HTML response to extract data

Parse the data and create dataframe from it

Data Wrangling

Completed Web Scraping notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb

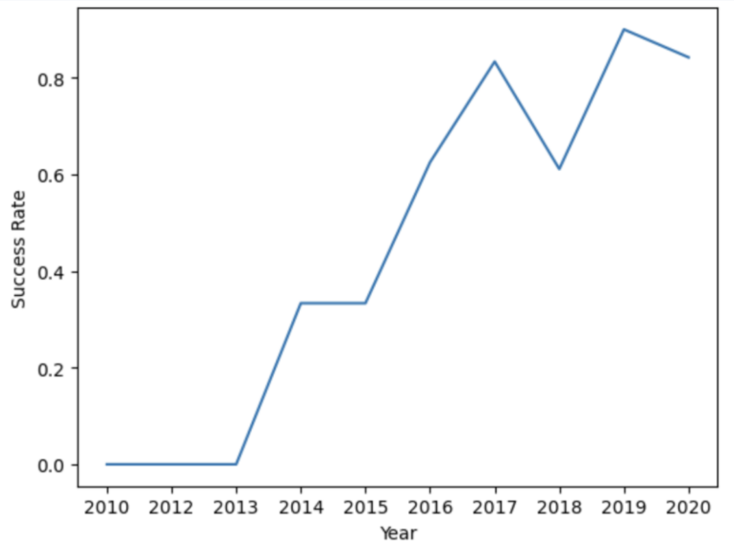
Data Wrangling

- Data collected via SpaceX REST API was filtered to only include Falcon 9 Launches. Web scraped data didn't need filtering as the table parsed only included Falcon 9 launches.
- The "Payload Mass" column had 5 missing values thar were replaced by the mean payload mass.
- "Landing Pad" column's missing values were not changed as they represented landing pads that were not used.
- A "Class" column was added to the dataset that represented the outcome of each launch being the value 0 for an unsuccessful first stage landing and value 1 for a successful landing.
- Complete Data Wrangling notebook available at: https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data wrangling.ipynb

EDA with Data Visualization

- Several scatter plots were used to analyse the relationship between different variables to determine which were more correlated with a successful outcome.
- A bar chart was used to examine the success rate of each orbit.
- A line chart was used to see the trend in successful first stage landings over time.

- Complete EDA with Data Visualization notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/edadataviz.ipynb



EDA with SQL

- To perform EDA with SQL we used several queries to:
  - Display the names of unique launch sites in the space mission.
    CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40
  - Display the total payload mass carried by boosters launched by NASA (CRS)
    45596
  - Display average payload mass carried by booster version F9 v1.1
    2534.67
  - List the date when the first successful landing outcome in ground pad was achieved
    2015-12-22
  - List the names of the booster versions which have success in drone ship and have payload mass greater than 4000 but less than 6000 and have carried the maximum payload mass.
  - List the total number of successful and failure mission outcomes
    1 failure and100 successes
- Complete EDA with SQL notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- The Folium library is used to create interactive maps for data visualization. We used Folium to:
  - Mark all launch sites on a map
  - Mark the successful and failed launches for each site on the map (using MarkerCluster) where green markers represents successful launches and red markers represents failed launches
  - Mark and calculate the distances between a specific launch site to its proximities (coastline and city)

- Complete Interactive Map with Folium notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Our dashboard consisted of a pie chart and a scatter plot controlled by a dropdown menu and a rangeslider.
- The dropdown menu options were the launch sites.
- The pie chart displayed total successful launches by site if the ”All Sites” option of the dropdown was selected or total successful launches for the specific site selected.
- The rangeslider represented the payload range in kg and ranged from 0 to 10000.
- The scatterplot represented the correlation between payload and success for the site selected (or for all sites).

- Complete Plotly Dash notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- To create Machine Learning Models, Scikit-Learn library was used.
- The ”class” column was transformed into a Pandas series and assigned to our Y variable. Remaining column were assigned to our X variable.
- Data in X was standardized and all data was split into training and test data.
- We used GridSearchCV to find the best hyperparameters for various classification algorithms: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors
- Each model is trained and evaluated, with their accuracy compared to determine the best performing method

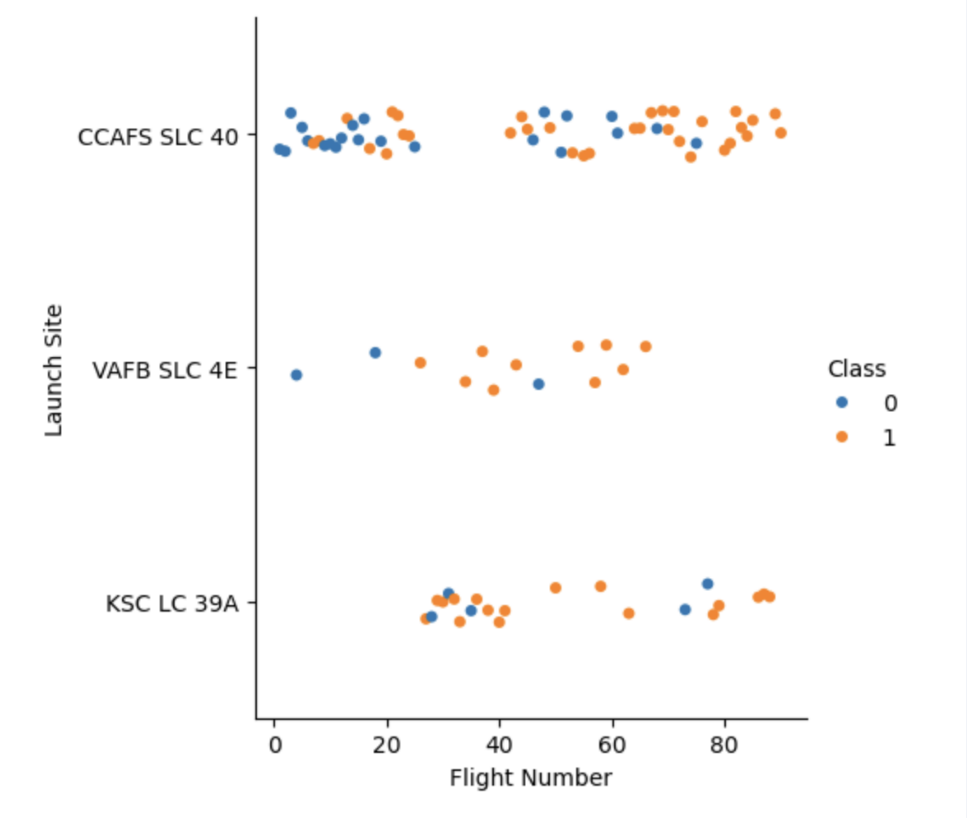
- Complete Machine Learning notebook available at https://github.com/scc131/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine Learning Prediction_Part_5.ipynb

Section 2

Insights drawn from EDA

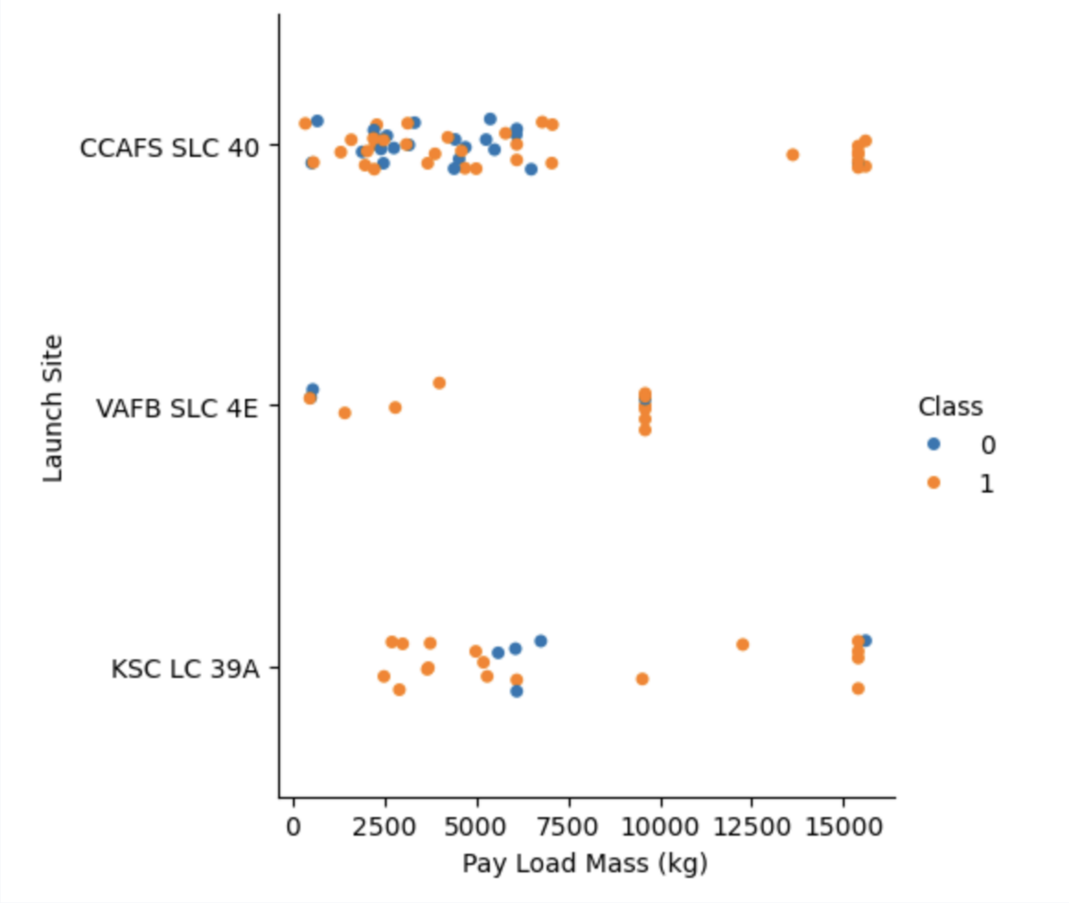
Flight Number vs. Launch Site

- Site CCAFS SLC 40 has the most launches.
- Site VAFB SLC 4E has the least amount of launches.
- As flight number increases so does success probability.



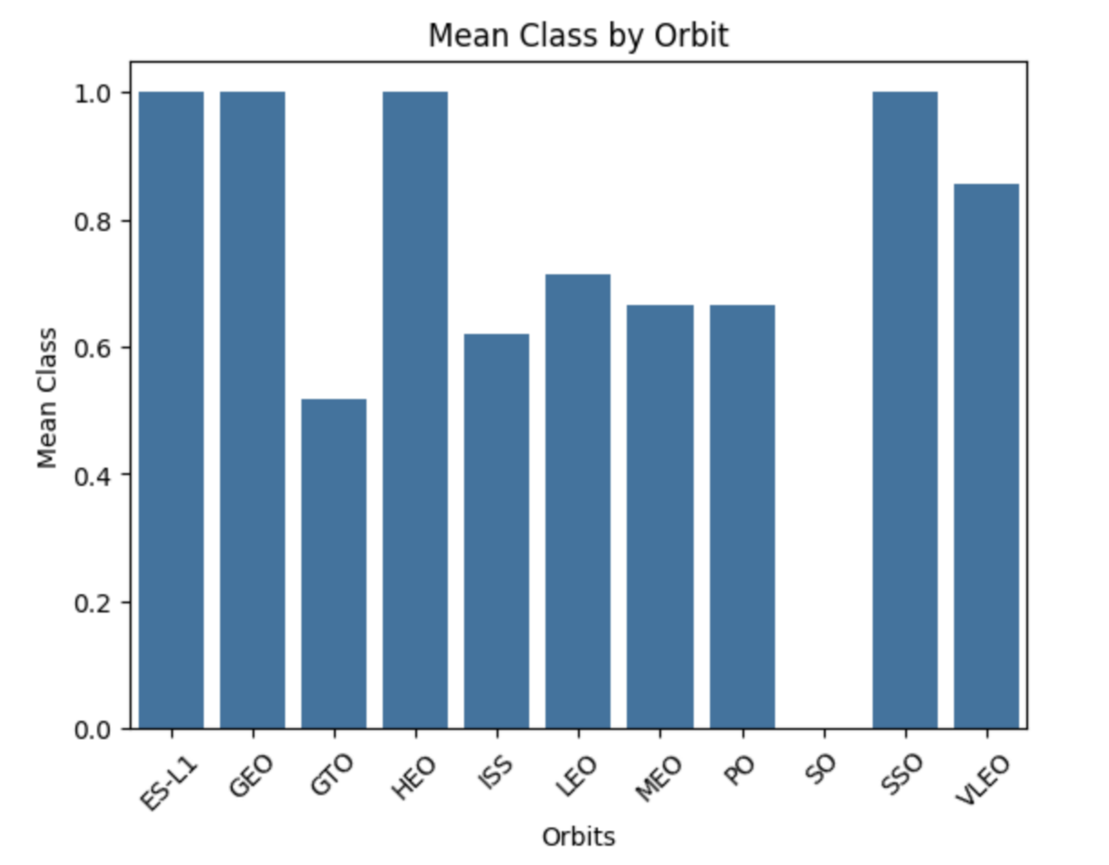
Payload vs. Launch Site

- Few launches have had heavy payloads (>10000kg) but nearly all of those were successful
- VAFB SLC 4E doesn't have any launches with a payload of over 10000kg
- Heavier payloads appear to be correlated with increased success rate



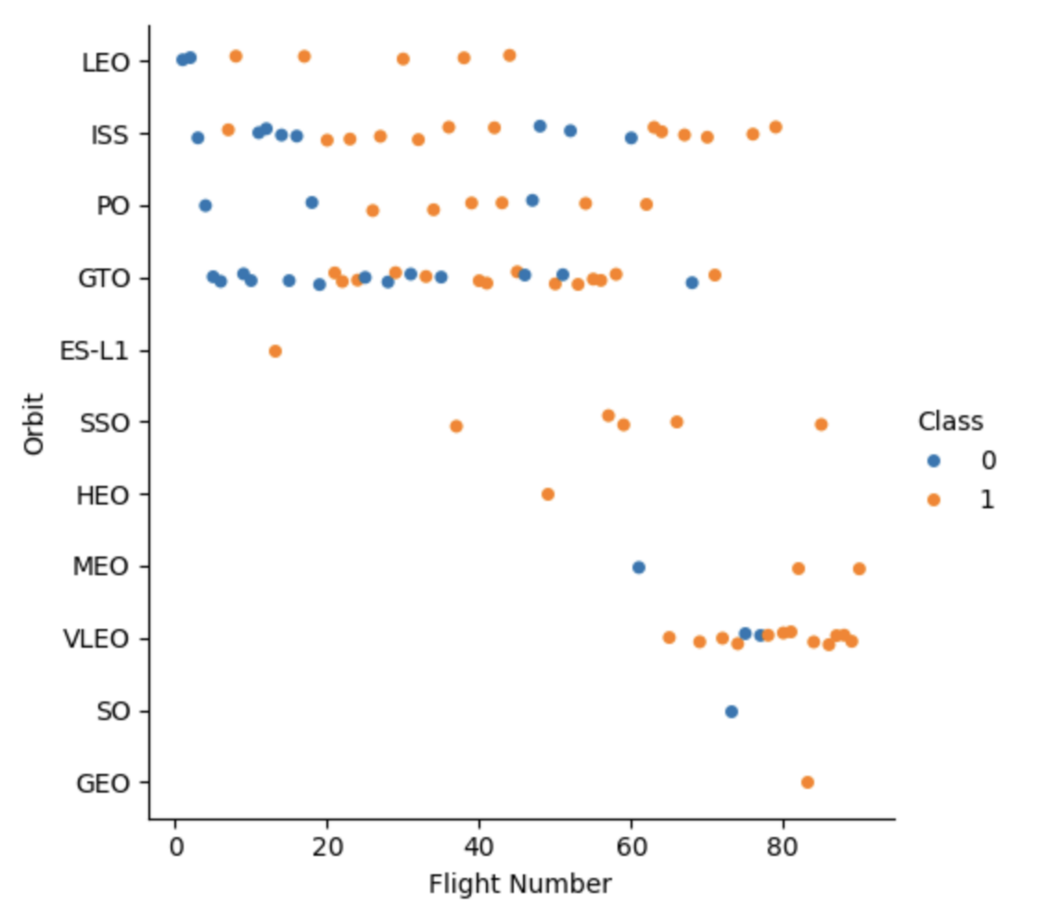
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO and SSO have always been successful
- Orbit SO has never been successful
- Remaining orbits have varying levels of success with most over 0.5 success rate (meaning more launches were successful than unsuccessful)



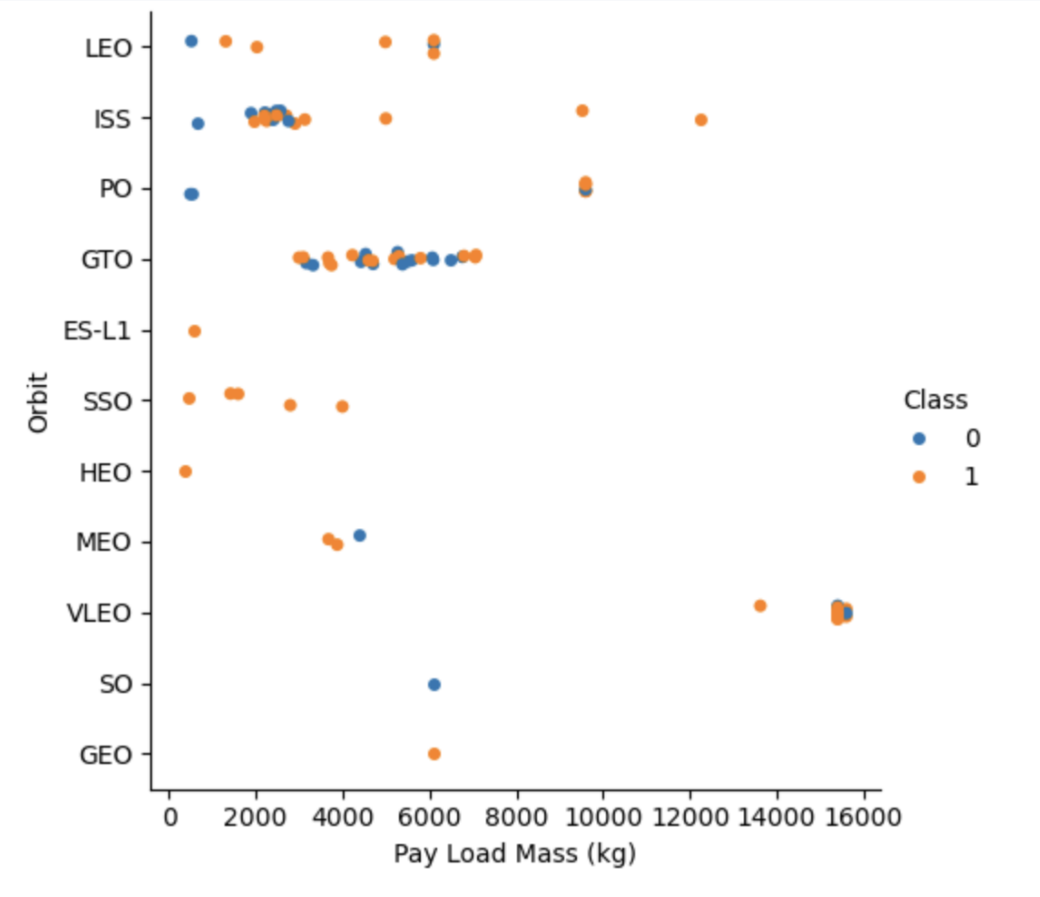
Flight Number vs. Orbit Type

- It appears that success rate improves as flight number goes up.
- Several orbit types were only tested after around 40 flights or more



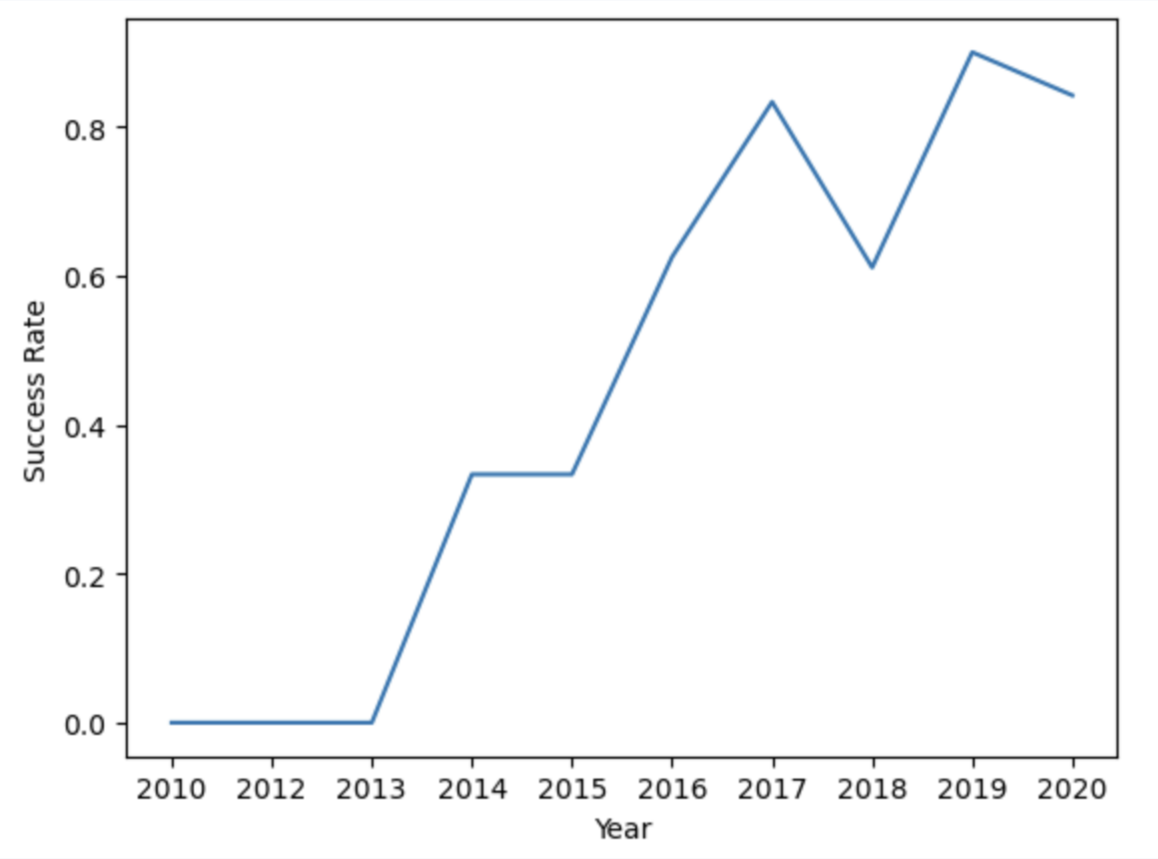
Payload vs. Orbit Type

- Orbit GTO shows no correlation between payload and successful landing.
- In general, orbits seem to be more successful as the payload gets heavier



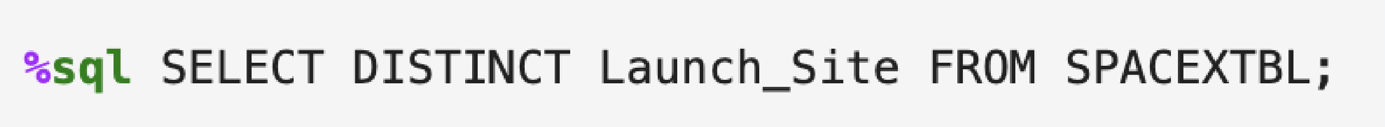
Launch Success Yearly Trend

- As of 2013 the success rate has been increasing

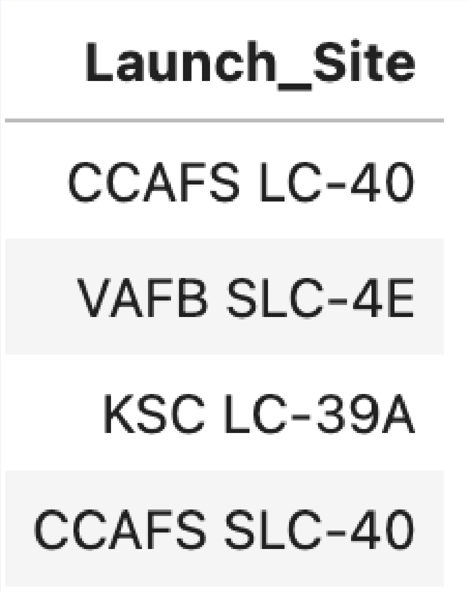


All Launch Site Names

- These are the names of the four sites that have launched Falcon 9 rockets



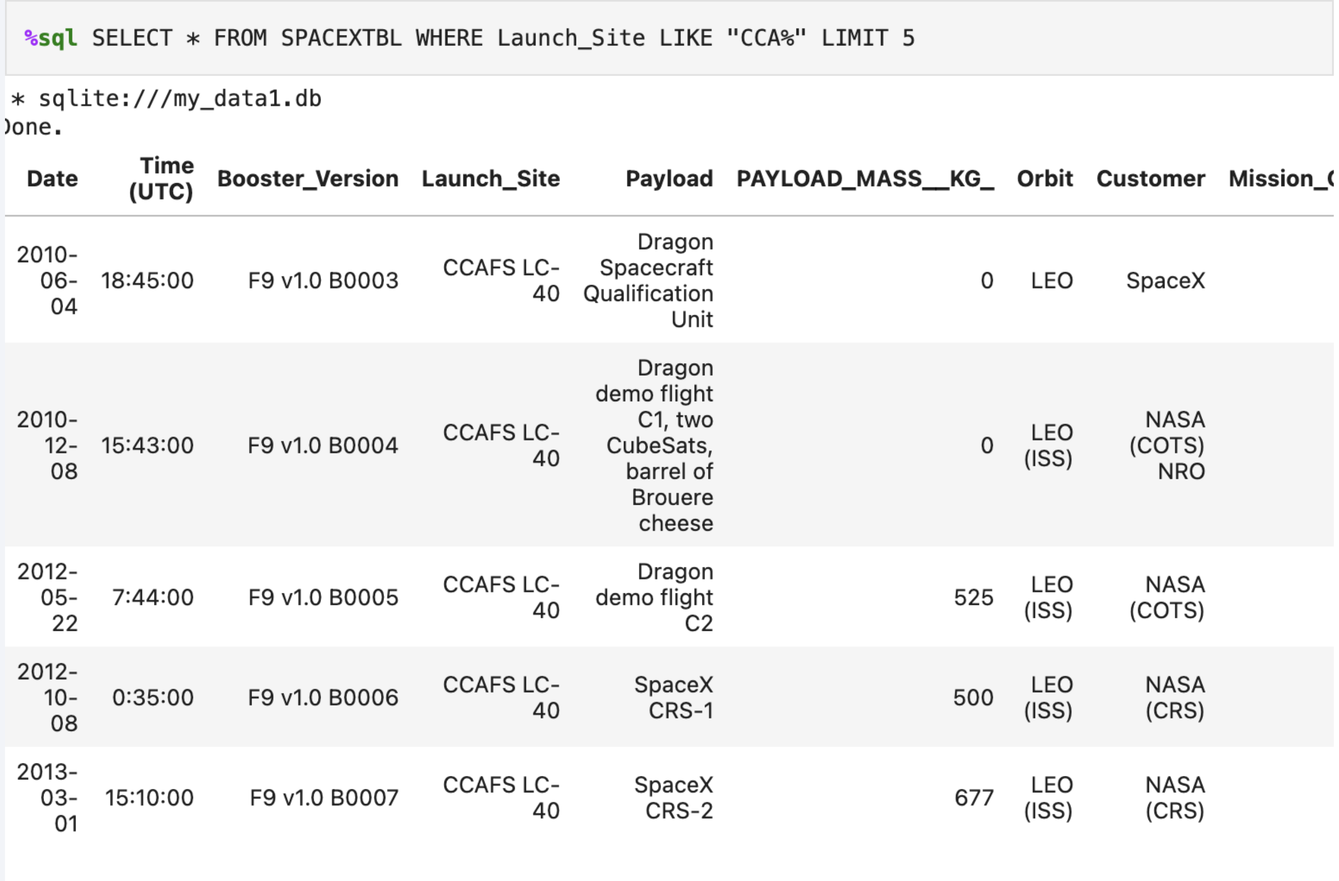
```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
```



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

Launch Site Names Begin with 'CCA'

- This query was used to display 5 Launch Sites whose names begin with “CCA”



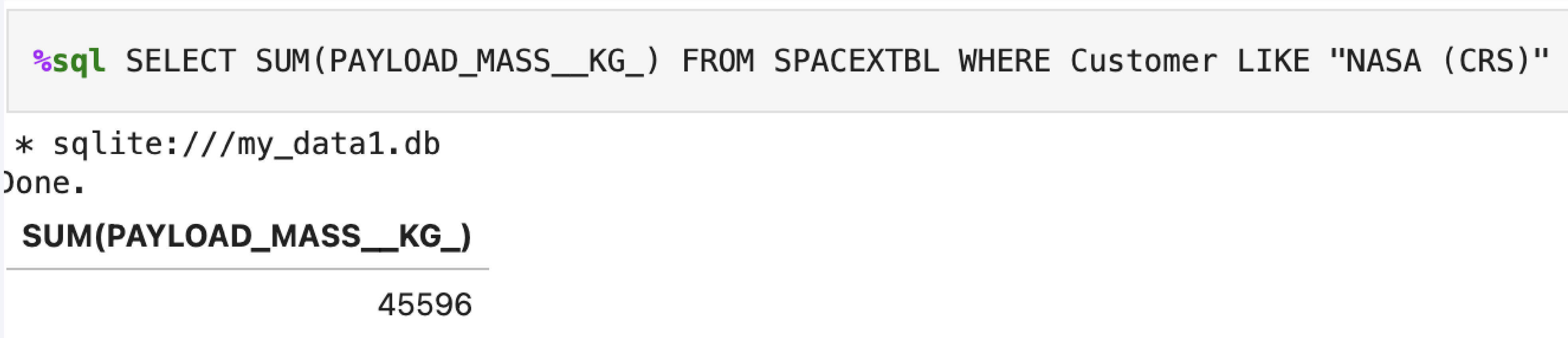
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE "CCA%" LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_(
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
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	
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

- This query was used to display the total payload mass carried by boosters launched by NASA (CRS)



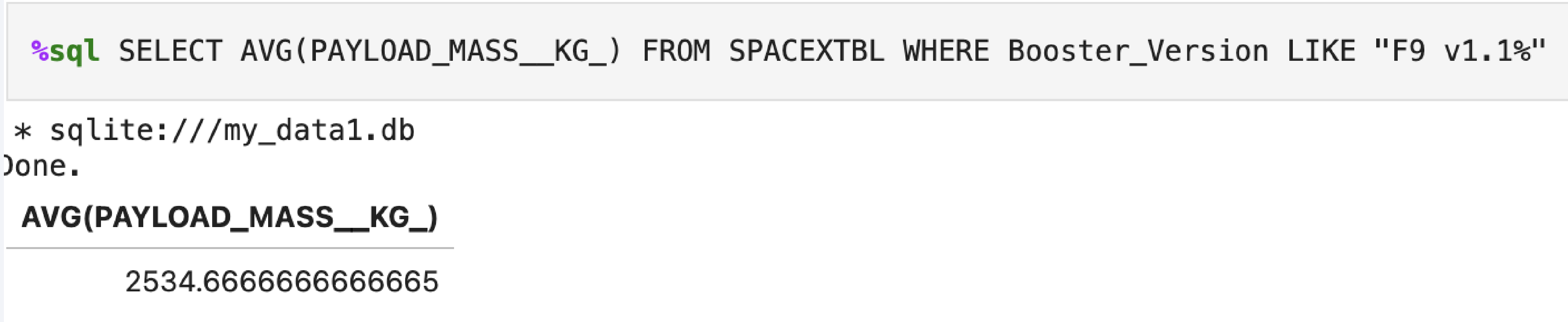
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- This query was performed to display average payload mass carried by booster version F9 v1.1



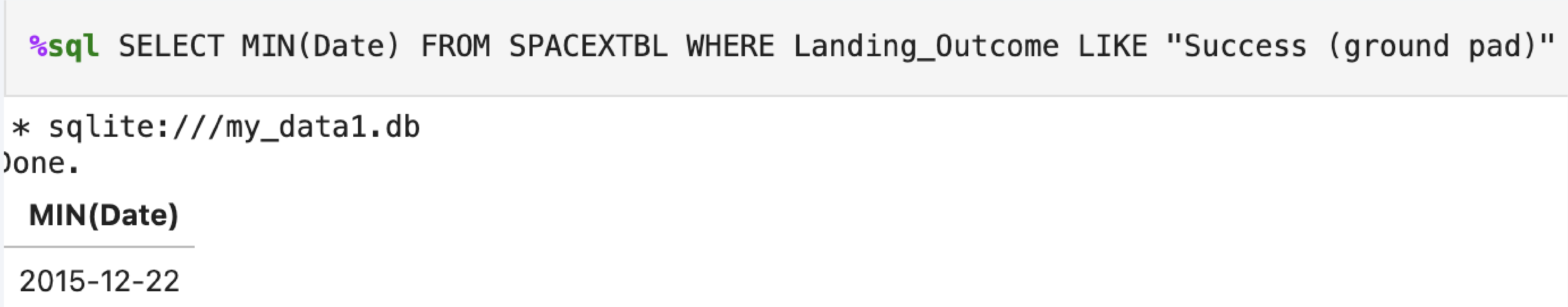
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE "F9 v1.1%"
```

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

AVG(PAYLOAD_MASS__KG_)
2534.666666666665

First Successful Ground Landing Date

- This query lists the date when the first succesful landing outcome in ground pad was acheived.



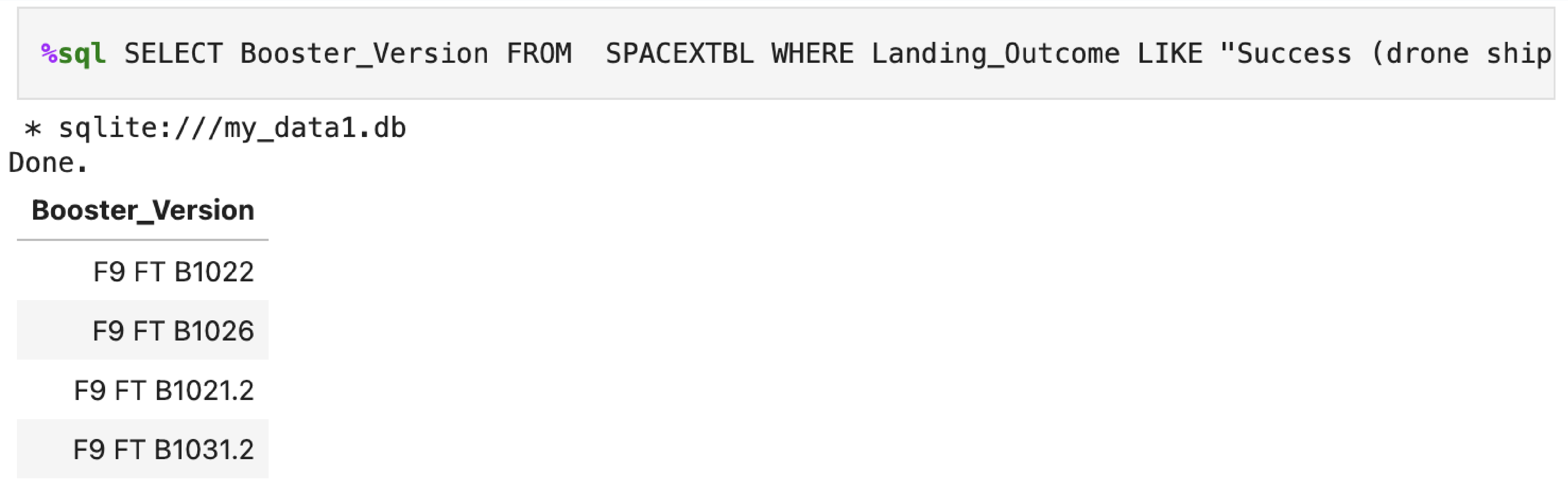
```
%sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE "Success (ground pad)"
```

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

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000



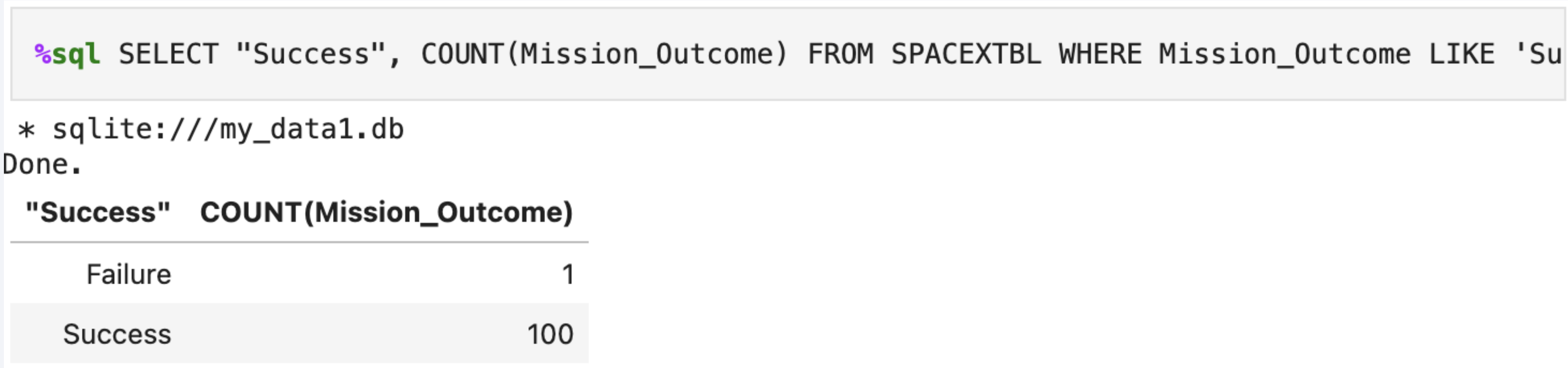
```
%sql SELECT Booster_Version FROM  SPACEXTBL WHERE Landing_Outcome LIKE "Success (drone ship
```

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

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

Total Number of Successful and Failure Mission Outcomes

- This query calculates the total number of successful and failure mission outcomes

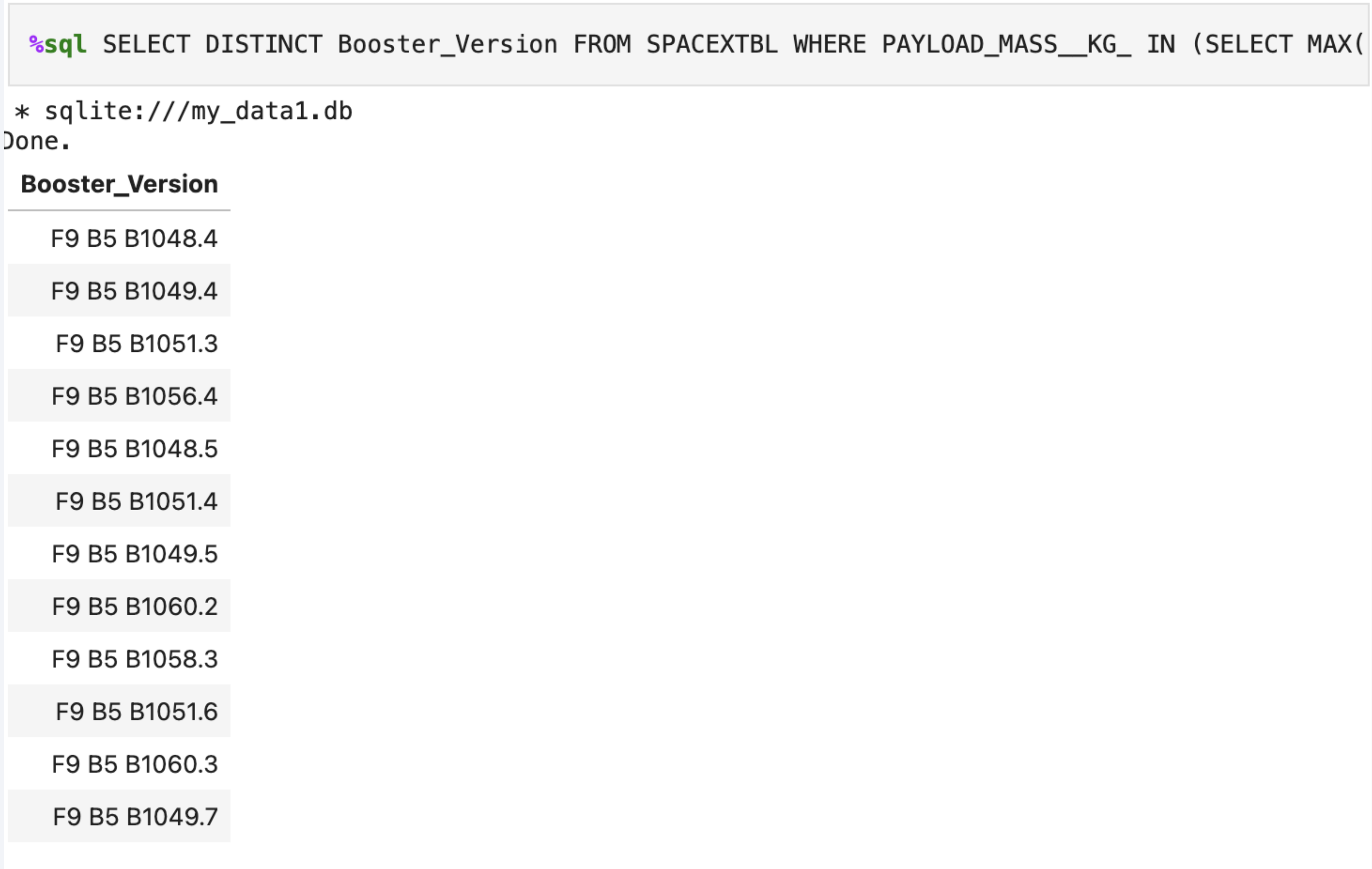


```
%sql SELECT "Success", COUNT(Mission_Outcome) FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Su
```

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

"Success"	COUNT(Mission_Outcome)
Failure	1
Success	100

Boosters Carried Maximum Payload



```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(
```

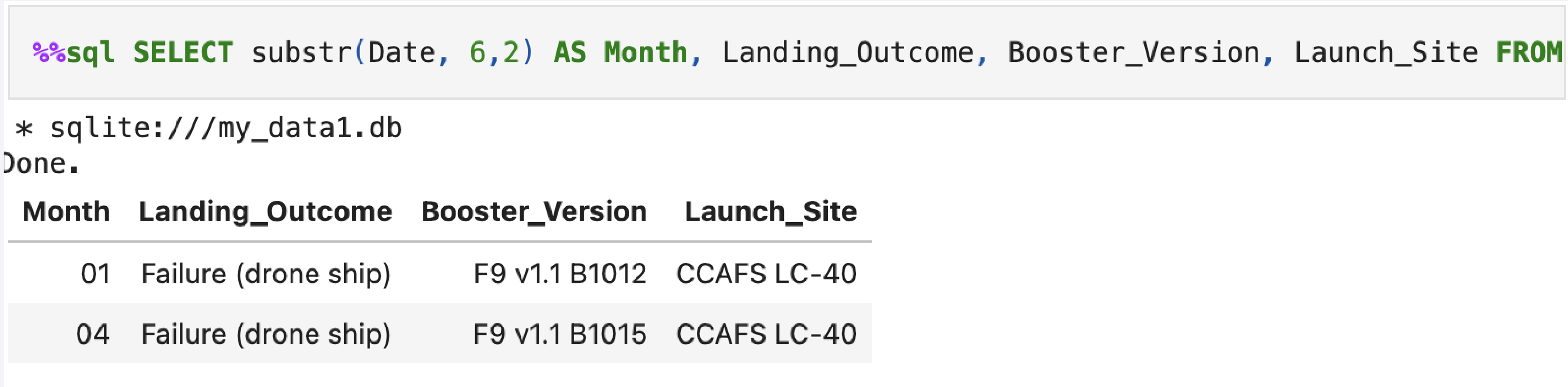
```
 * sqlite:///my_data1.db
Done.
```

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

- This query lists the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

- This query lists the failed landing outcomes in drone ship, along with their booster versions, the month, and launch site names for the year 2015.



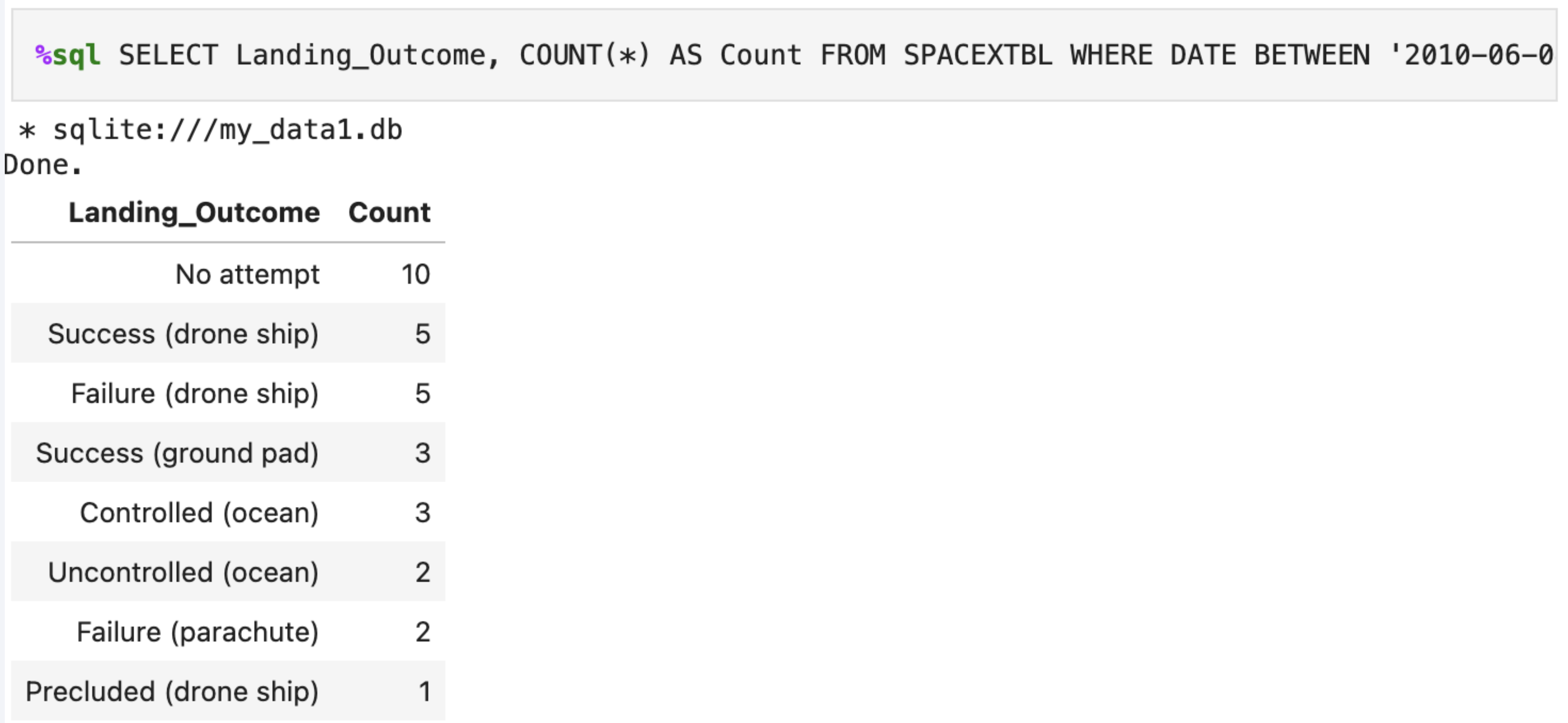
```
%%sql SELECT substr(Date, 6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM
```

```
* 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

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

- This query gives us a rank of the count of landing outcomes 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 '2010-06-0
```

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

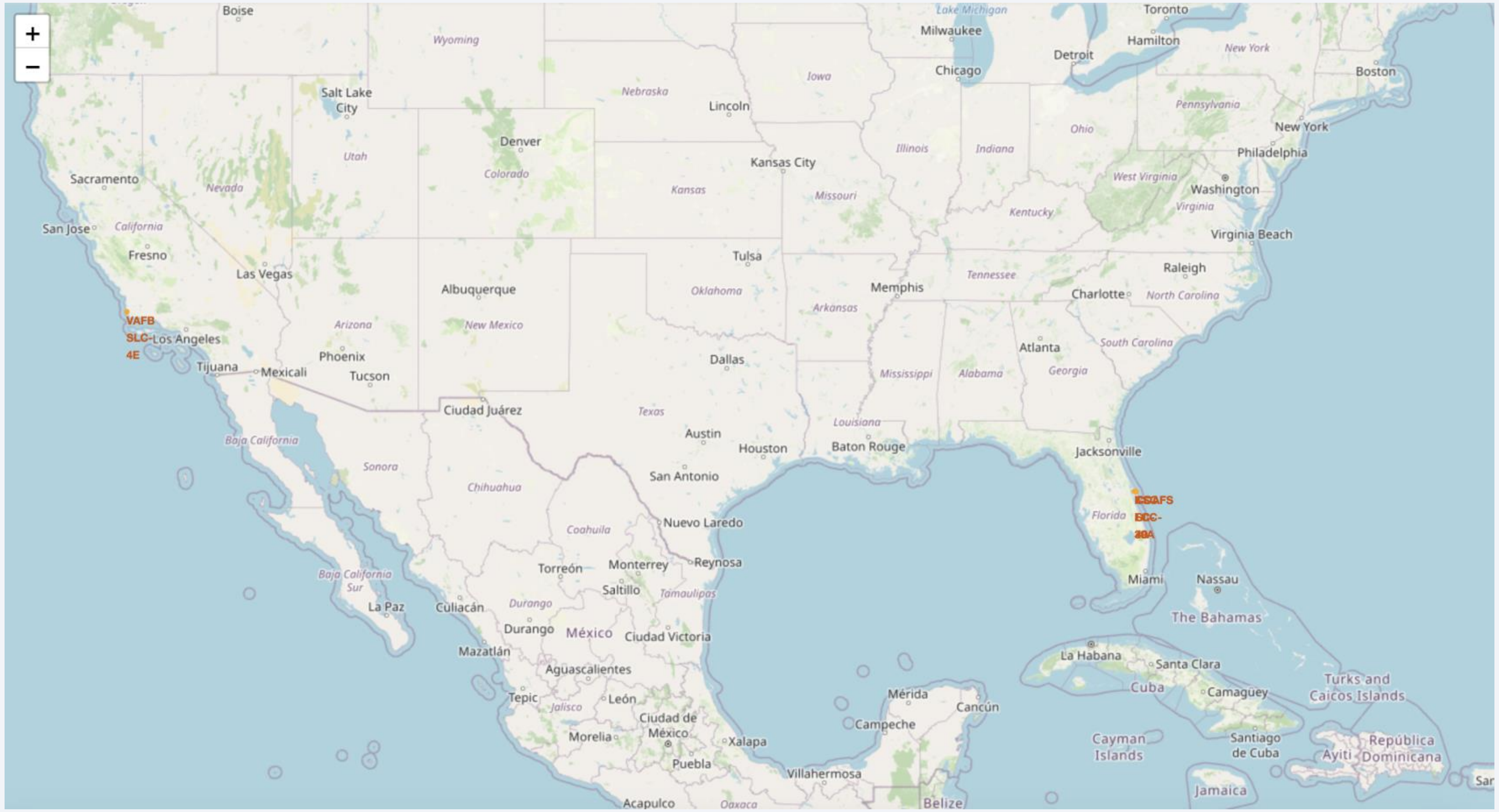
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

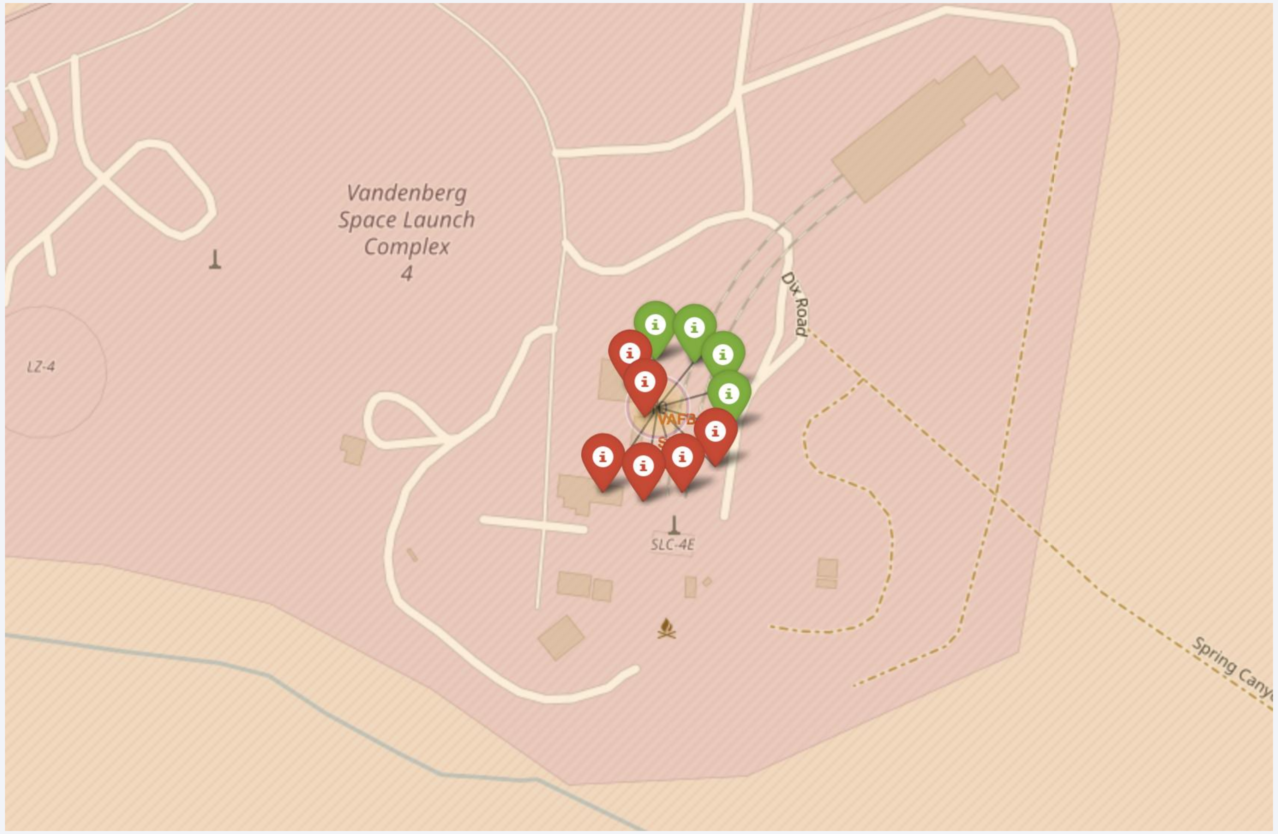
Map showing all launch sites

- This map shows all Launch Sites.
- Intriguingly, all launch sites are near the coast and near water.



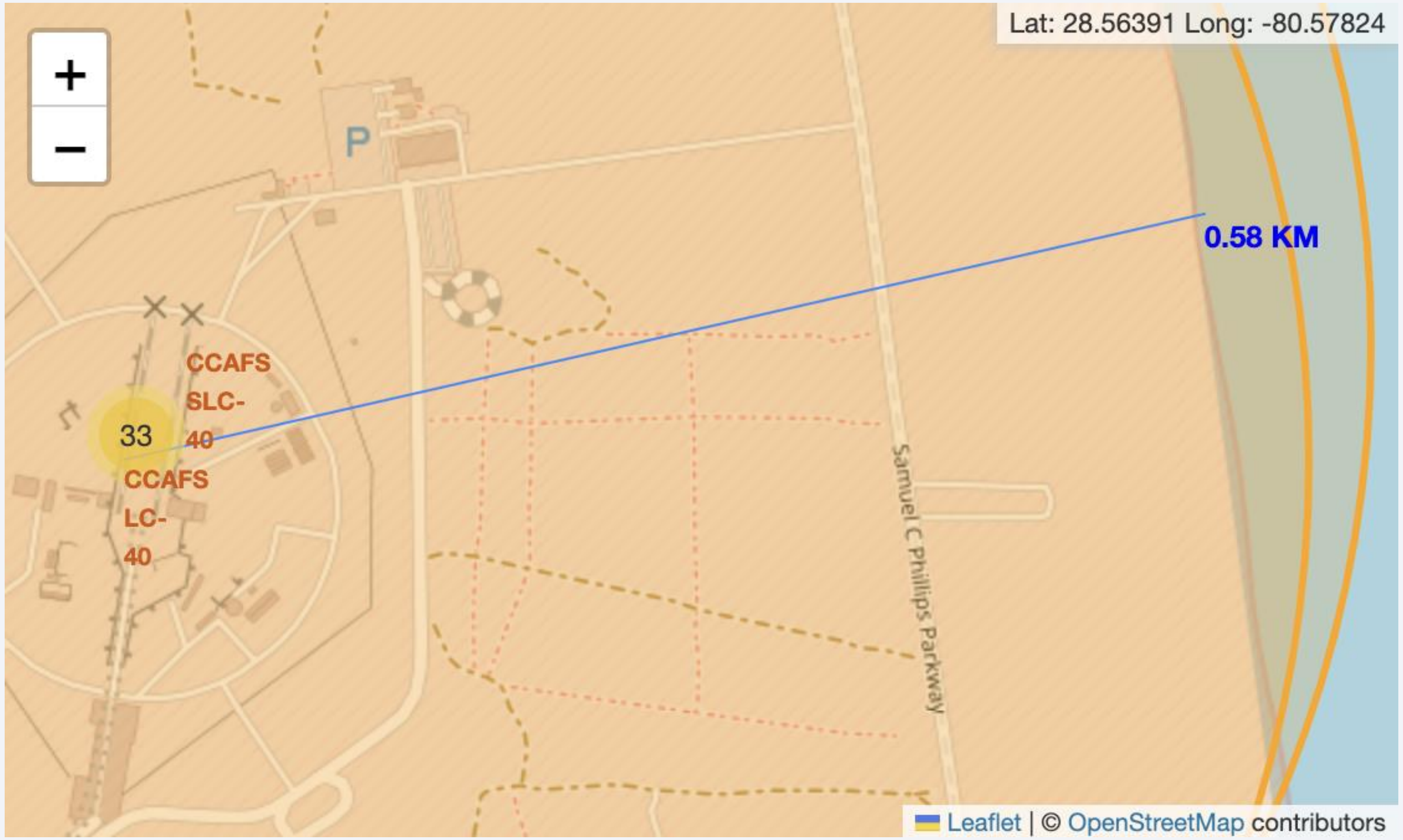
Map showing launch outcomes

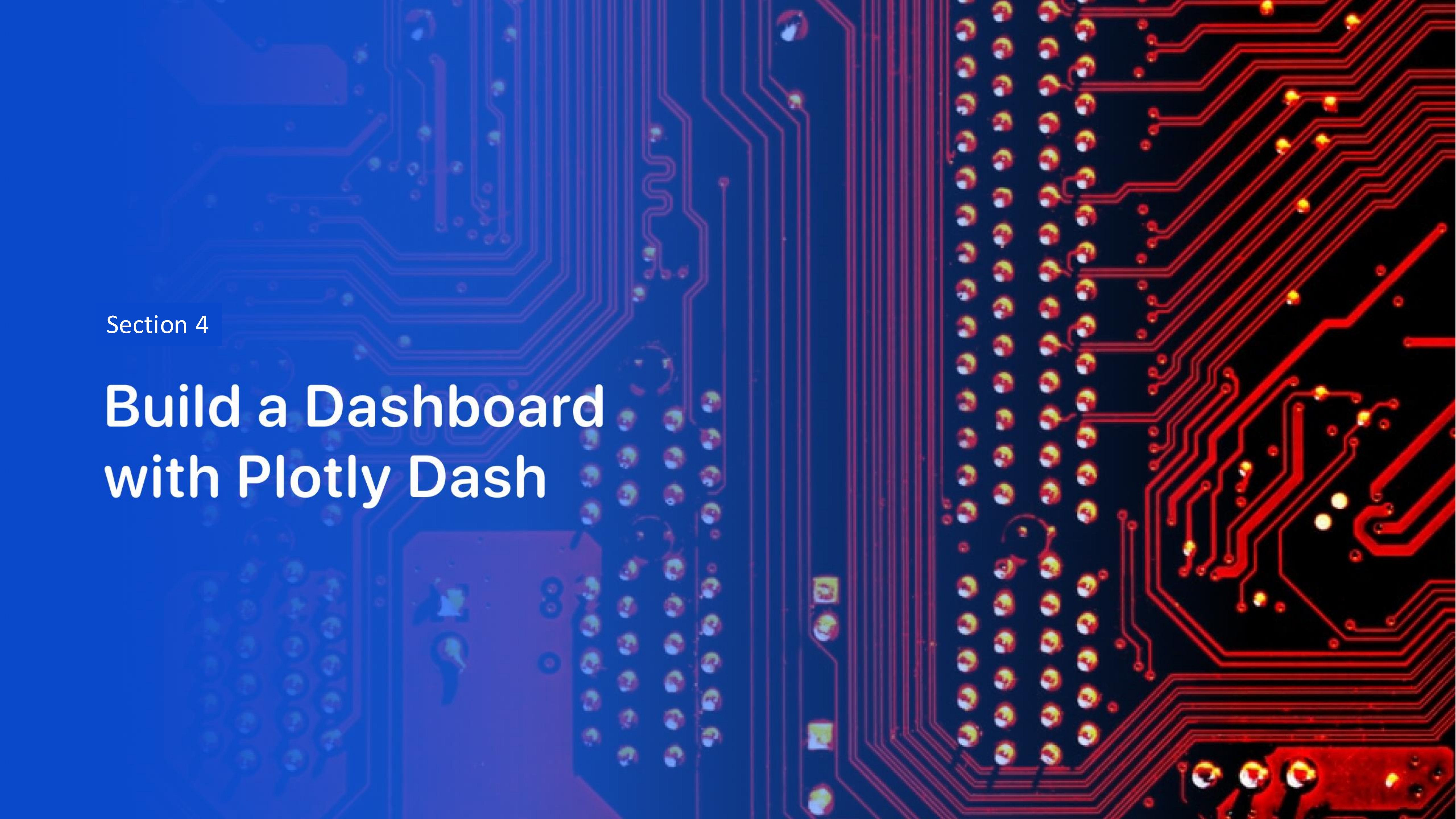
- This map shows the launch outcomes color coded. Red symbolizes failed landings and green symbolizes successful landings.
- This is an easy and visual way to see the success/failure of landings in each launch sites.



Map showing proximities of launch sites

- This map shows the distance of one of the launch sites to the coastline.


+
−
Lat: 28.56391 Long: -80.57824
P
0.58 KM
CCAFS
SLC-
40
33
CCAFS
LC-
40
Samuel C Phillips Parkway
Leaflet | © OpenStreetMap contributors

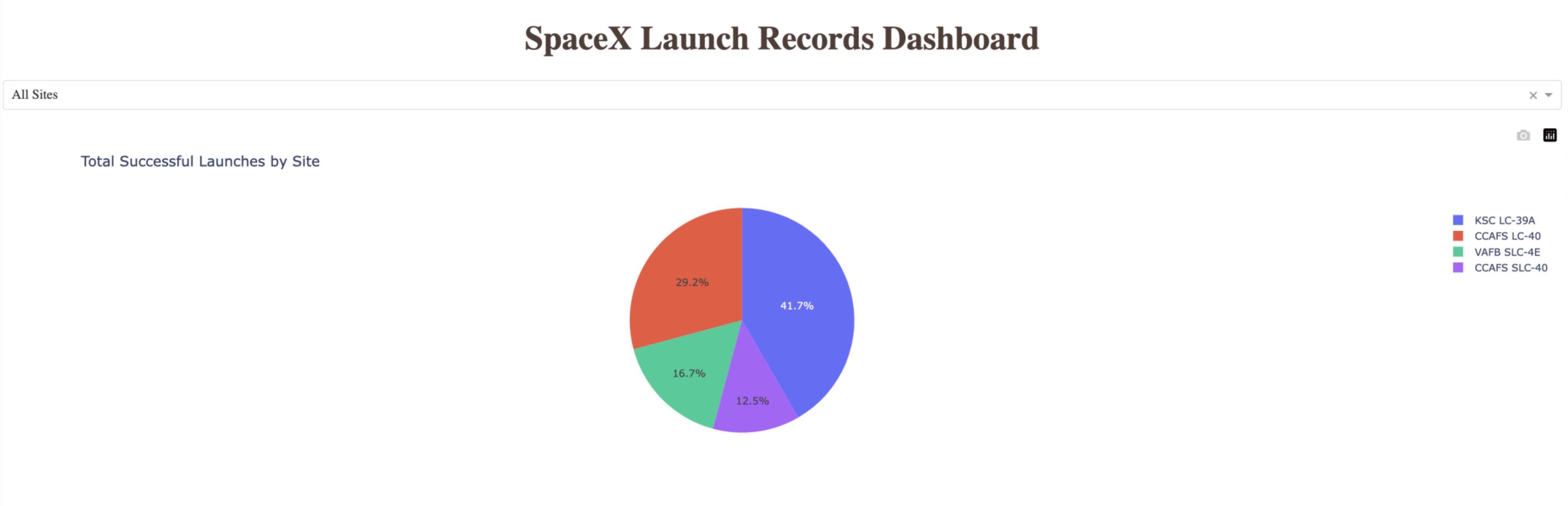


Section 4

Build a Dashboard with Plotly Dash

Total successful launches by site

- Pie chart showing total successful launches by site in the interactive dashboard.
- Launch site KSC LC-39A has the highest success rate while CCAFS SLC-40 has the lowest.



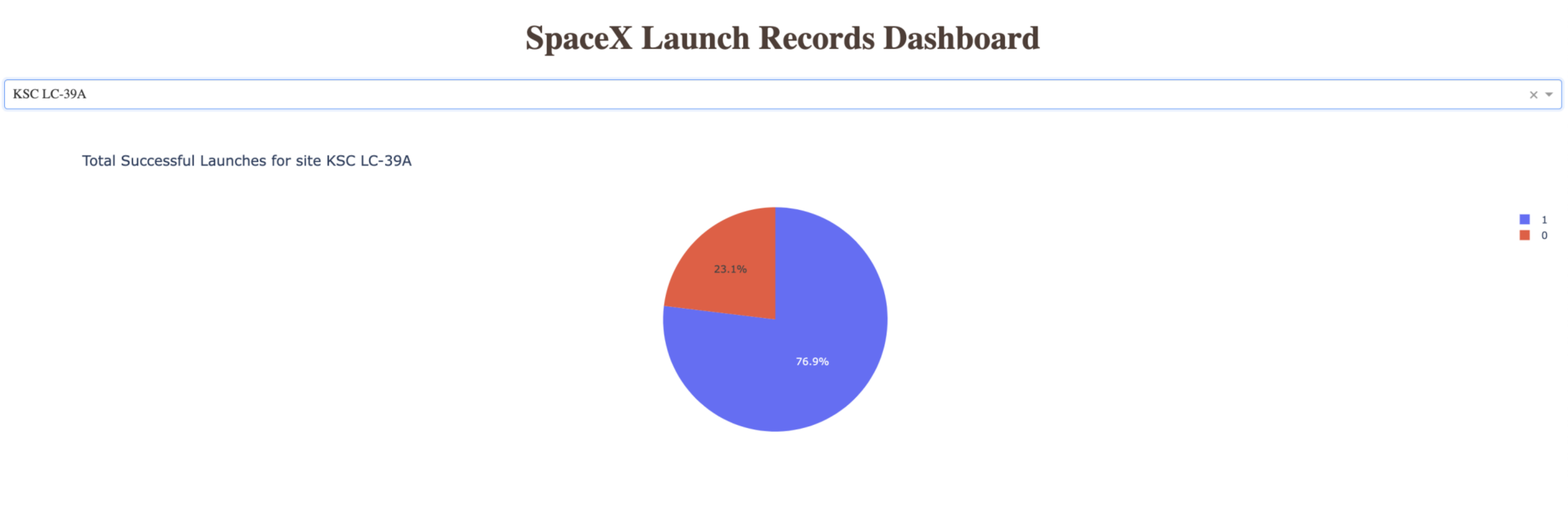
SpaceX Launch Records Dashboard

All Sites

Total Successful Launches by Site

Pie chart of the launch site with highest success rate

- This pie chart shows the launch site with highest launch success rate - KSC LC-39A
- In this site, 76.9% of launches are successful whereas only 23.1% of launches fail

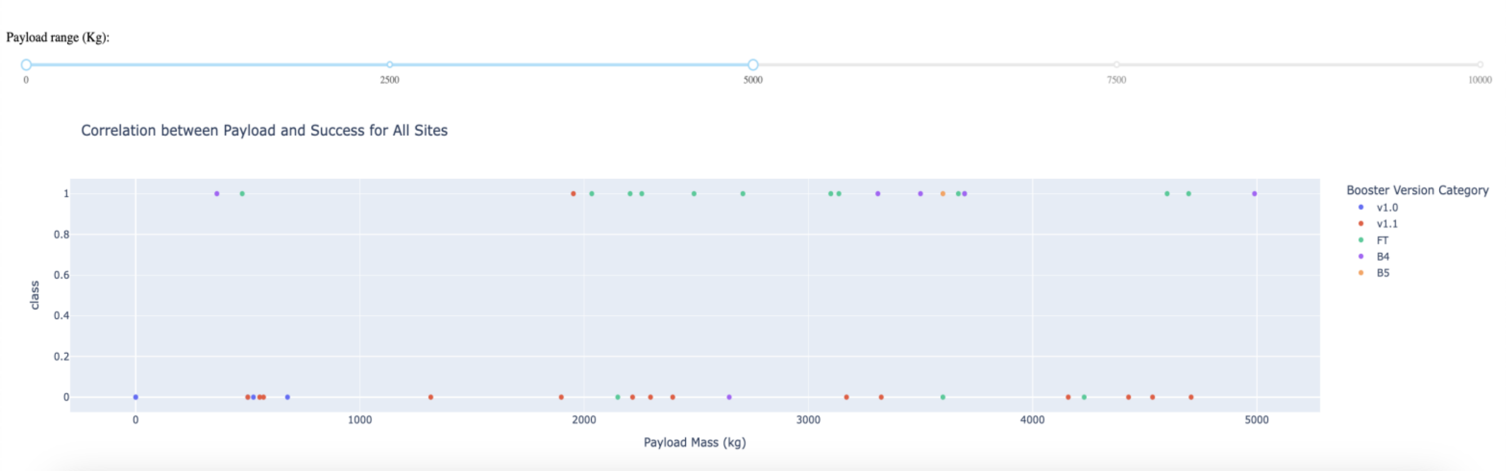


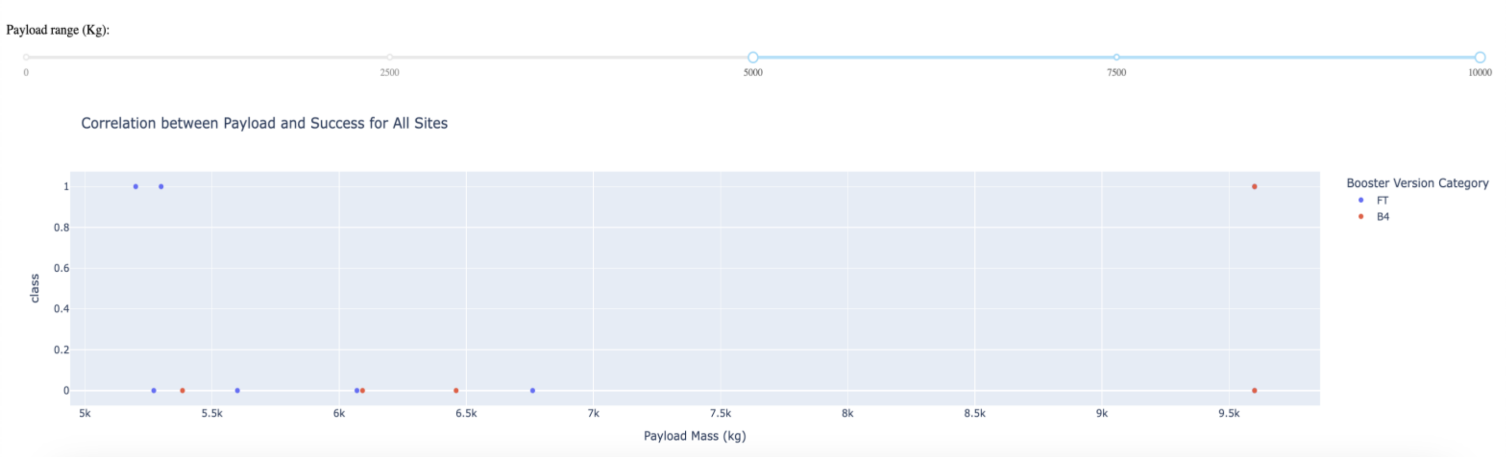
SpaceX Launch Records Dashboard

KSC LC-39A

Total Successful Launches for site KSC LC-39A

Payload vs. Launch Outcome





- Below are two scatter plots showing payload vs launch outcome where the first plot has payload between 0kg and 5000kg and the second plot has payload between 5000kg and 10000kg.

- It seems that payloads under 5000kg are correlated with more successful launches.

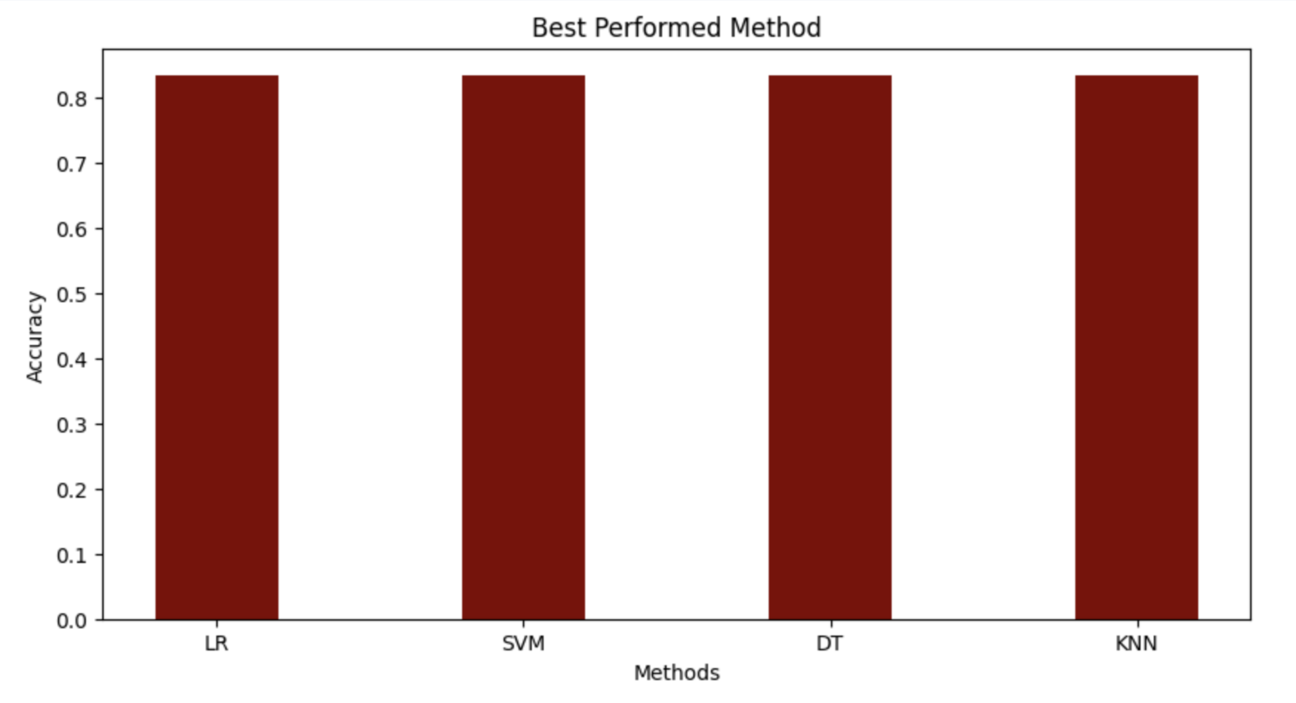
- FT booster version also seems to be correlated with higher success rate .

Section 5

Predictive Analysis (Classification)

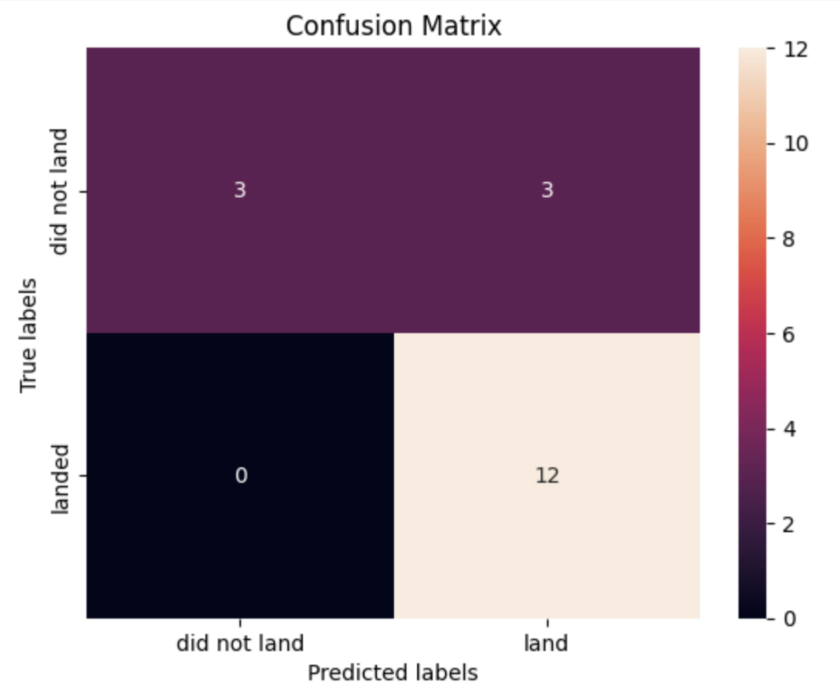
Classification Accuracy

- When we compare the accuracy of all tested methods, we can see they all have the same accuracy: 83.33%



Confusion Matrix

- This is the confusion matrix of the Decision Tree method.
- We can see that 3 landings were incorrectly predicted as having landed when they didn’t – there is 3 false positives.
- However, the model correctly predicted all successful landings giving us no false negatives.



Conclusions

In this project, we aimed to predict the success of the rocket's first stage landing.

Our analysis, based on SpaceX's data, revealed the following points:

- The KSC LC-39A site has the highest success rate among all observed sites.
- Launch success rate has been increasing since 2013.
- The higher the number of flights, the higher the success rate.
- Orbits ES-L1, GEO, HEO and SSO had the highest success rates.
- To make accurate predictions, we applied various machine learning models. The overall accuracy of these models was 83.33%.

Thank you!