IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Putera Muhammad Syaamil Bin Ismail
16 June 2024

Outline

Executive Summary

1. This research has been divided into 4 parts which are data collection, data analysis/wrangling, data visualization and modelling/predictive analysis. The primary focus would be on Falcon 9 Booster Version.

2. Summary of all results are:

1. Data Collection- There are 5 null values in the pay load mass column which have been replaced with the mean values with 26 null value rows in the Landing Pad columns.

2. Data Analysis- 4 unique launch sites, 45596kg of total payload mass by NASA CRS, average of 2534.67kg payload mass, GTO ISS and VLEO is the highest occurrence of orbit. Success rate of launching next mission is 67%.

3. Prediction- Using Logistic Regression, SVM, Decision Tree and KNN. Accuracy score is 83% for all models.

Introduction

- Project background and context. The project primarily focused on Falcon 9 Booster Version. There are 5 null values row from the pay load mass that have been replaced by mean values. The analysis and visualizations that have been used are the usuals graphs including folium map chart for better location visualizations.
- The business problem questions are what is the success rate of launching another rocket? Which launch sites have high success rate to launch a certain range of payload mass?Are launch sites in close proximity to railways? Are launch sites in close proximity to highways? Are launch sites in close proximity to coastline? Do launch sites keep certain distance away from cities?

Section 1

Methodology

Methodology

Data collection methodology:

  - SpaceX API
  - Web Scrap from Wikipedia page

- Perform data wrangling
  - Null values of pay load mass have been replaced with mean. Calculated the number of launches in each site. Calculated the occurrence of each orbit. Calculated the occurrence of mission outcomes of the orbits. Labelling landing outcome. Calculated success rate from the labels.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree and KNN. All acquired 83% accuracy percentage.

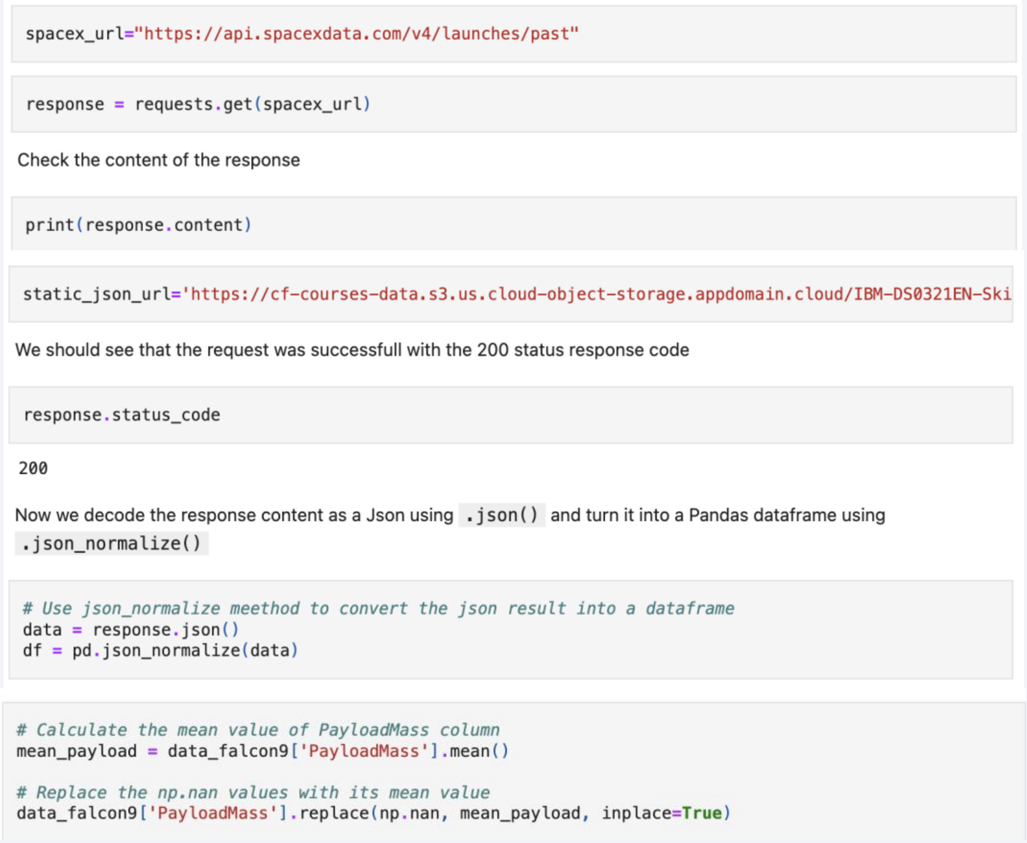
Data Collection

- Data sets have been collected via the SpaceX API and web scraping from Wikipedia page
- Acquired from SpaceX REST API. The API gave us data about launch sites, payload mass in kg, booster version, launching pad, landing outcomes and mission outcomes. Decoded response content with .json() and normalize it into a dataframe using .json_normalize().
- For web scraping, Beautiful Soup library have been used. Extracted launch records as HTML table, parse it and then convert it into a dataframe.

Data Collection – SpaceX API

- Requested from SpaceX API to collect the data, clean it and done some basic wrangling and formatting.

- The link to my notebook is https://github.com/puterasyaamil/finalproject/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Check the content of the response

```
print(response.content)
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-Ski
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
data = response.json()
df = pd.json_normalize(data)
```

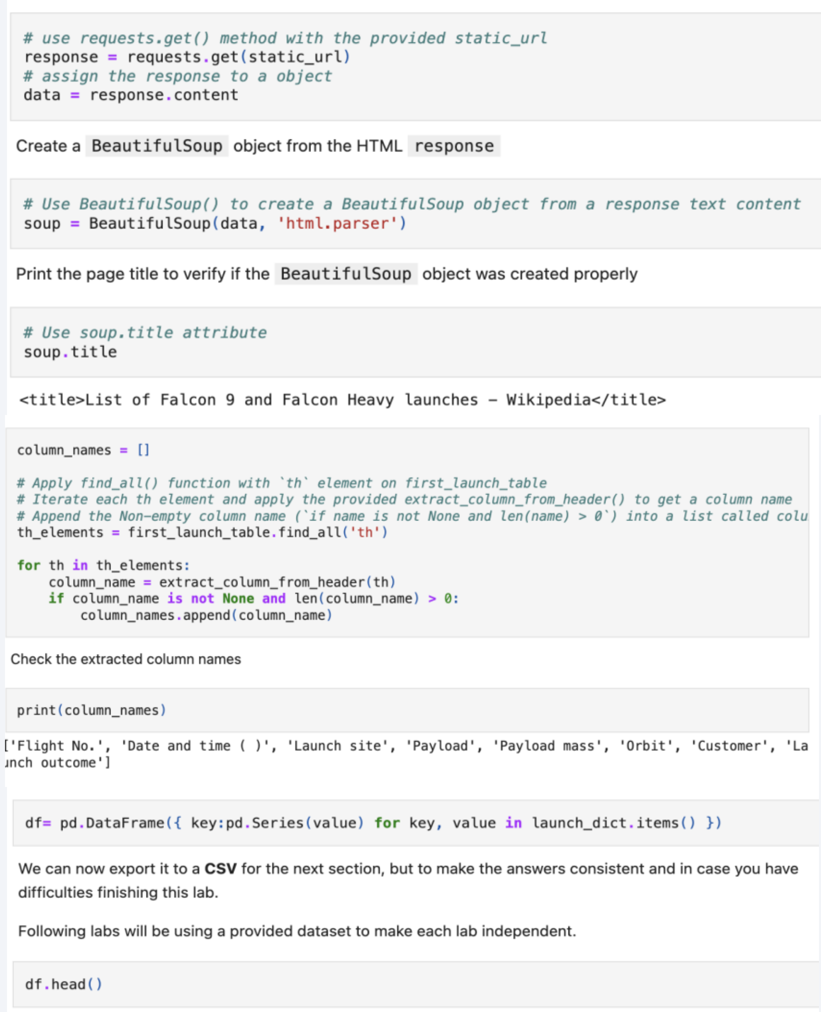
```
# Calculate the mean value of PayloadMass column
mean_payload = data_falcon9['PayloadMass'].mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan, mean_payload, inplace=True)
```

Data Collection - Scraping

- Using Beautiful Soup to extract res ponses as HTML
  table. Then parse it and insert into a dataframe.

- The link to my notebook is

https://github.com/puterasyaamil/finalproject/blob/main/jupyter-labs-webscraping.ipynb



```
# use requests.get() method with the provided static_url
response = requests.get(static_url)
# assign the response to a object
data = response.content
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(data, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

```
column_names = []

# Apply find_all() function with `th` element on first_launch_table
# Iterate each th element and apply the provided extract_column_from_header() to get a column name
# Append the Non-empty column name (`if name is not None and len(name) > 0`) into a list called colu
th_elements = first_launch_table.find_all('th')

for th in th_elements:
    column_name = extract_column_from_header(th)
    if column_name is not None and len(column_name) > 0:
        column_names.append(column_name)
```

Check the extracted column names

```
print(column_names)
```

```
['Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'La
unch outcome']
```

```
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

We can now export it to a **CSV** for the next section, but to make the answers consistent and in case you have difficulties finishing this lab.

Following labs will be using a provided dataset to make each lab independent.

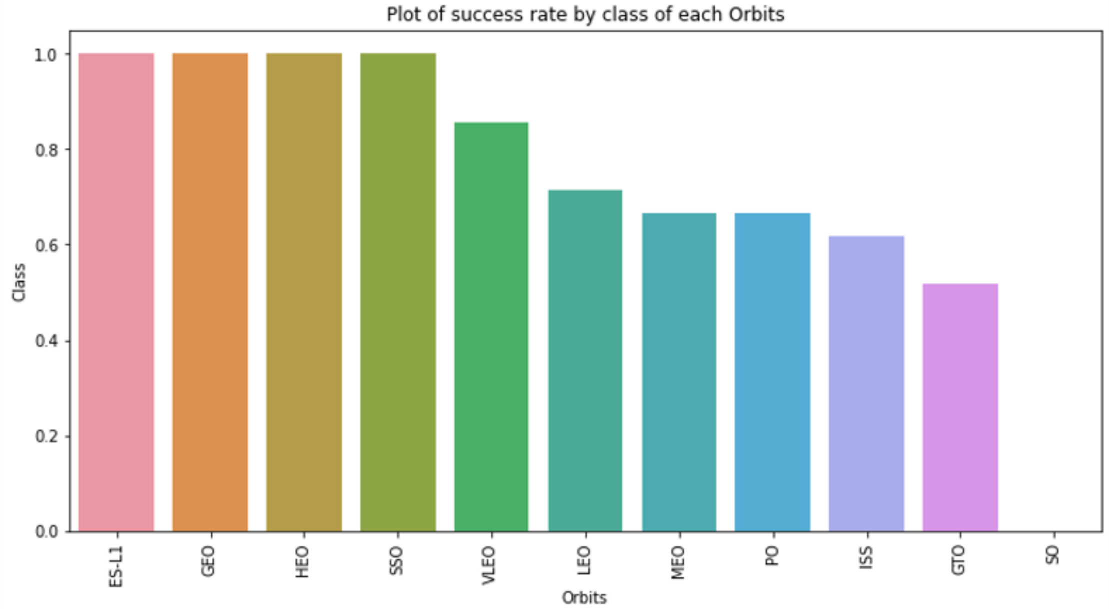
```
df.head()
```

Data Wrangling

- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launches at each site, and the number and occurrence of each orbits
- Created landing outcome label from outcome column and exported the results to csv.
- The link to my notebook https://github.com/puterasyaamil/finalproject/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We explored the data by visualizing the relationships among flight number and launch site, payload and launch site, success rates for each orbit type, flight number and orbit type, and the yearly trend in launch success.
- The link to my notebook is https://github.com/puterasyaamil/finalproject/blob/main/edadataviz.ipynb



EDA with SQL

We applied exploratory data analysis (EDA) using SQL to gain insights from the data. Specifically, we wrote queries to identify:

- The unique names of launch sites used in space missions.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by the booster version F9 v1.1.
- The total number of successful and failed mission outcomes.
- The failed landing outcomes on drone ships, including their booster versions and launch site names.

The link to my notebook
is https://github.com/puterasyaamil/finalproject/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marked all launch sites and added map objects such as markers, circles, and lines to indicate the success or failure of launches for each site on the folium map.
- Assigned launch outcomes to classes, with 0 for failure and 1 for success. Using color-labeled marker clusters, we identified which launch sites have relatively high success rates.
- Calculated the distances between each launch site and nearby proximities, addressing questions such as whether launch sites are near railways, highways, and coastlines, and if they maintain a certain distance from cities.
- The link to my notebook is https://github.com/puterasyaamil/finalproject/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Built an interactive dashboard using Plotly Dash App.
- In this dashboard, we included pie charts displaying the total number of launches at various sites and scatter plots illustrating the relationship between launch outcomes and payload mass (in kg) for different booster versions.
- The link to my notebook is https://github.com/puterasyaamil/finalproject/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- We loaded the data using NumPy and pandas, then transformed it and split it into training and testing sets.
- We built various machine learning models and tuned their hyperparameters using GridSearchCV.
- Accuracy was used as the evaluation metric for our models.
- Through feature engineering and algorithm tuning, we improved the model performance and identified the best performing classification model.
- The link to my notebook is https://github.com/puterasyaamil/finalproject/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

EDA Results

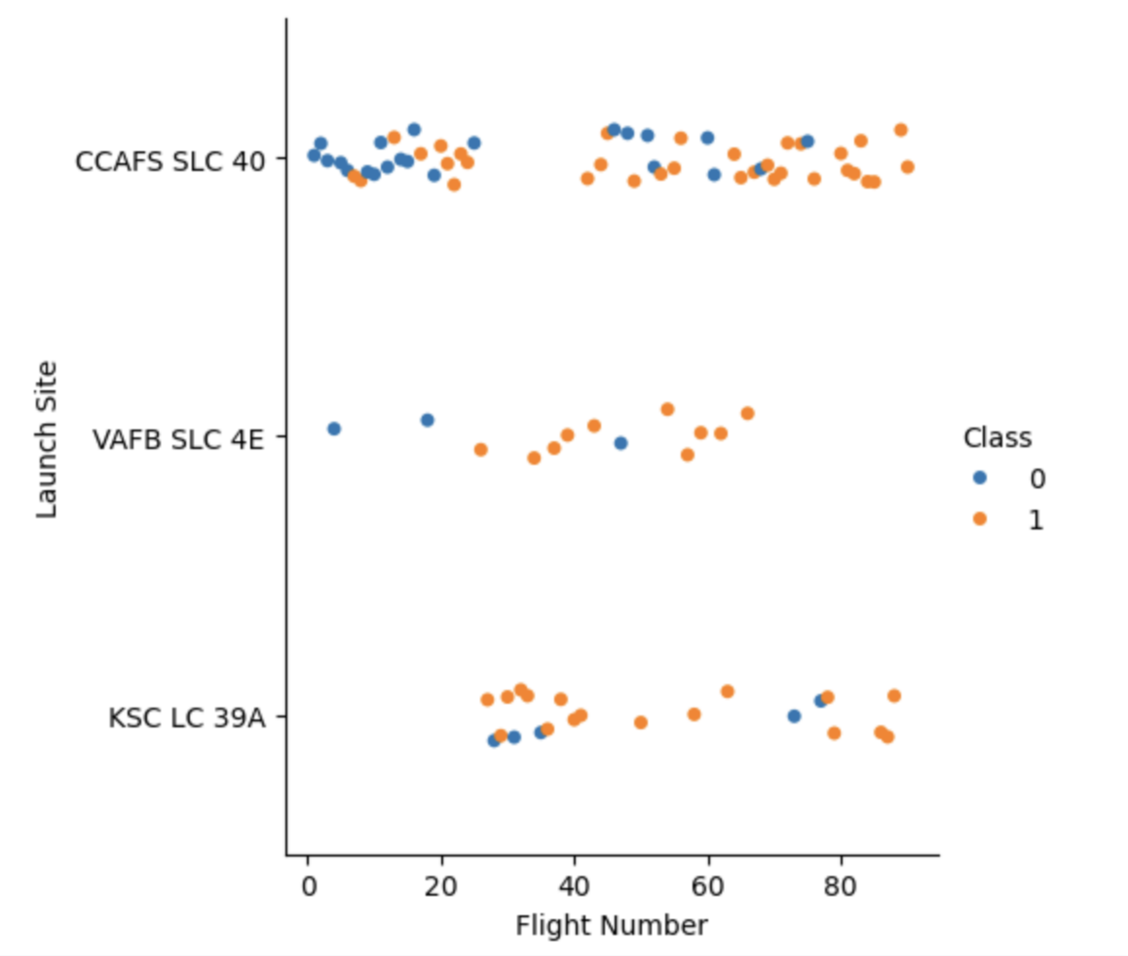
- **Launch Success Rates by Site:**
  - CCAFS LC-40 has a success rate of 60%.
  - KSC LC-39A and VAFB SLC 4E both have a success rate of 77%.
- **Most Successful Launch Site:**
  - KSC LC-39A had the highest number of successful launches among all sites.
- **Payload Mass Analysis:**
  - At the VAFB-SLC launch site, there are no rockets launched with a payload mass greater than 10,000 kg.
- **Orbit Success Rates:**
  - The orbits ES-L1, GEO, HEO, and SSO have the highest success rates.
- **Relationship Between Flight Number and Success:**
  - In LEO orbit, success appears to be related to the number of flights.
  - In GTO orbit, there seems to be no relationship between the number of flights and success.
- **Success Rate Trend Over Time:**
  - The success rate has been increasing steadily from 2013 to 2020.

Section 2

Insights drawn from EDA

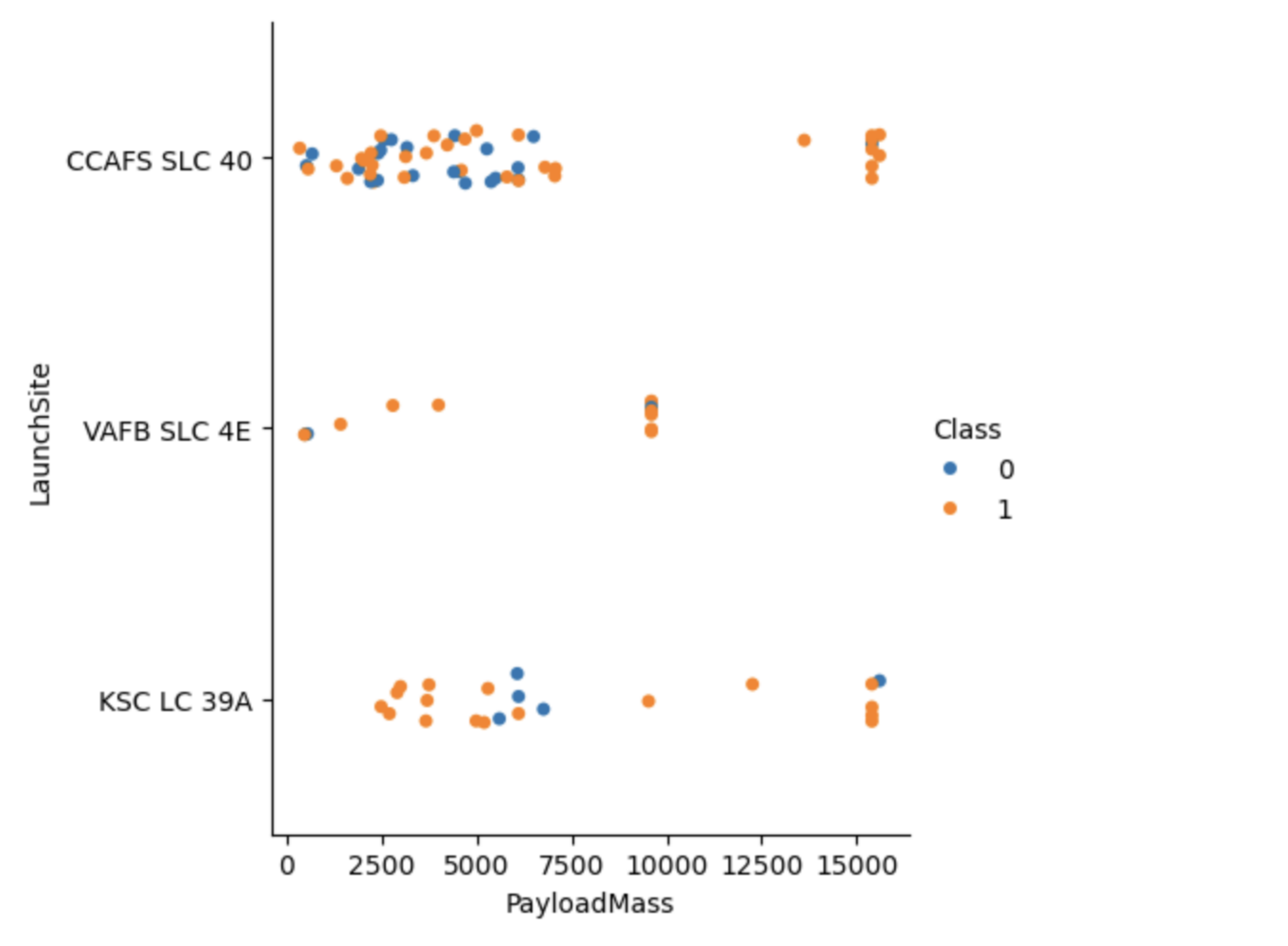
Flight Number vs. Launch Site

- Launch site CCAFS SLC 40, higher success rates for flight no of greater than 20.
- Launch site VAFB SLC 4E, higher success rates of flight no of greater than 20.
- Launch site KSC LC 39A, higher success rate of flight no of greater than 30.



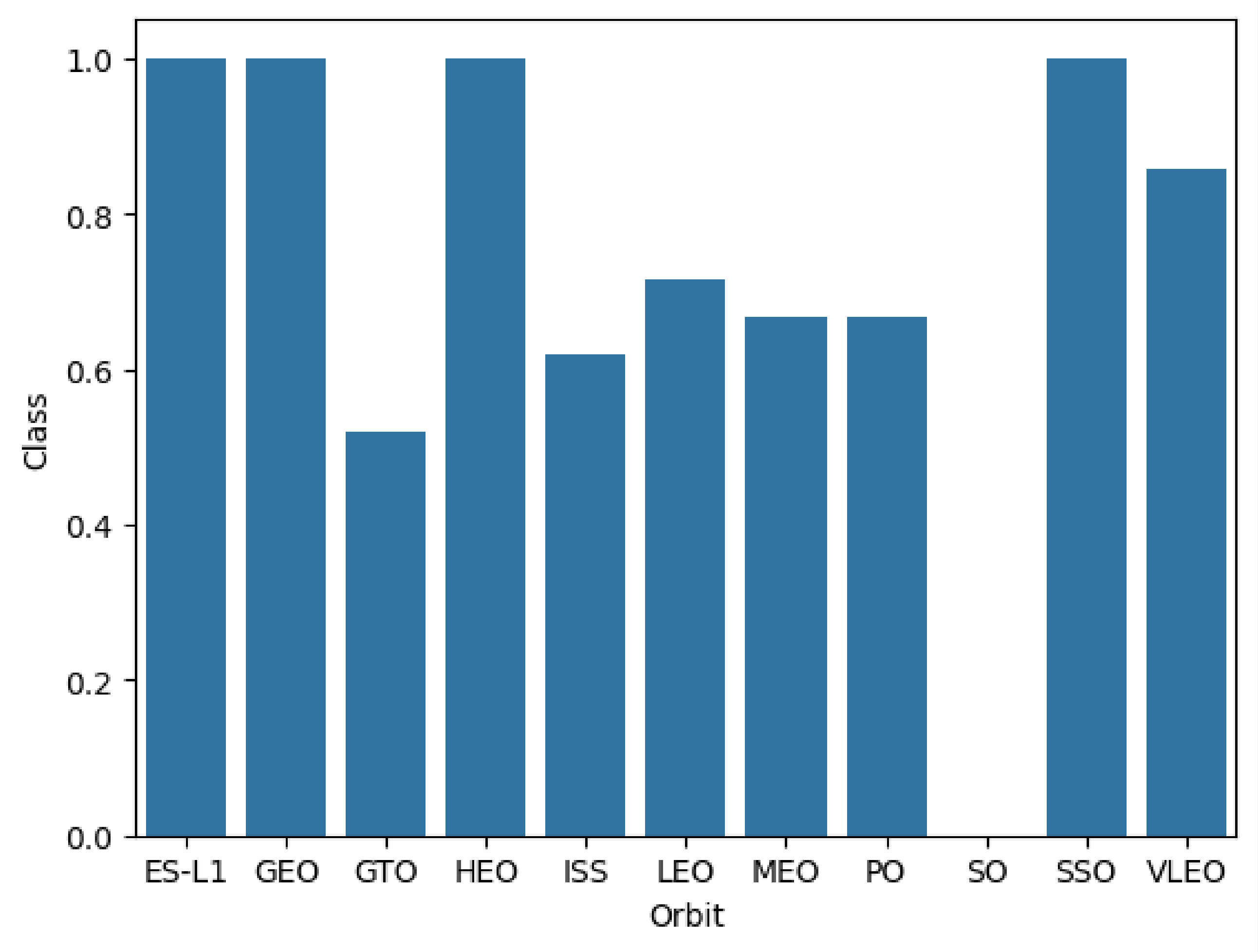
Payload vs. Launch Site

- For less than 10,000kg pay load, high success rate at VAFB SLC 4E launch site.
- Much balance trend is at the other 2 launch sites.



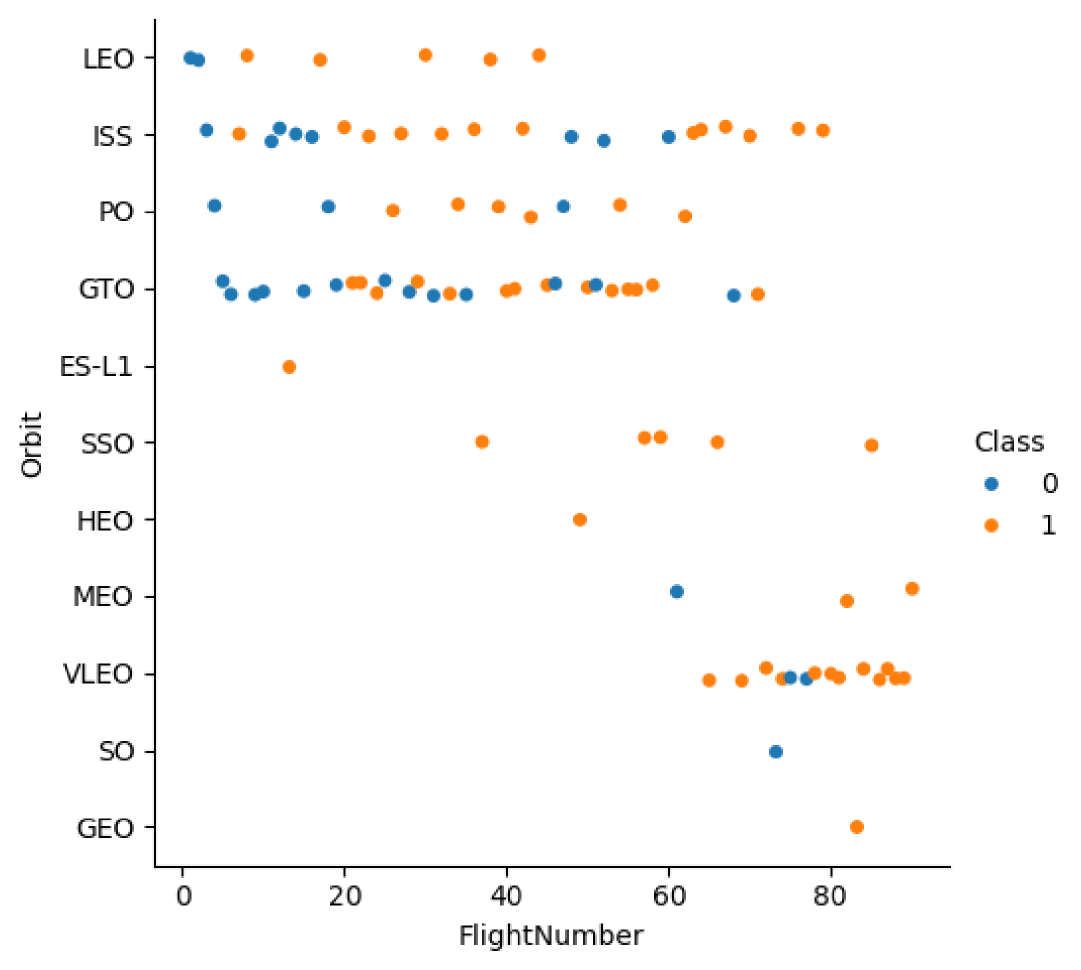
Success Rate vs. Orbit Type

- ESL1, GEO, HEO and SSO Orbit type have 100% success rate
- SO, GTO and ISS Orbit Type have the lowest success rates



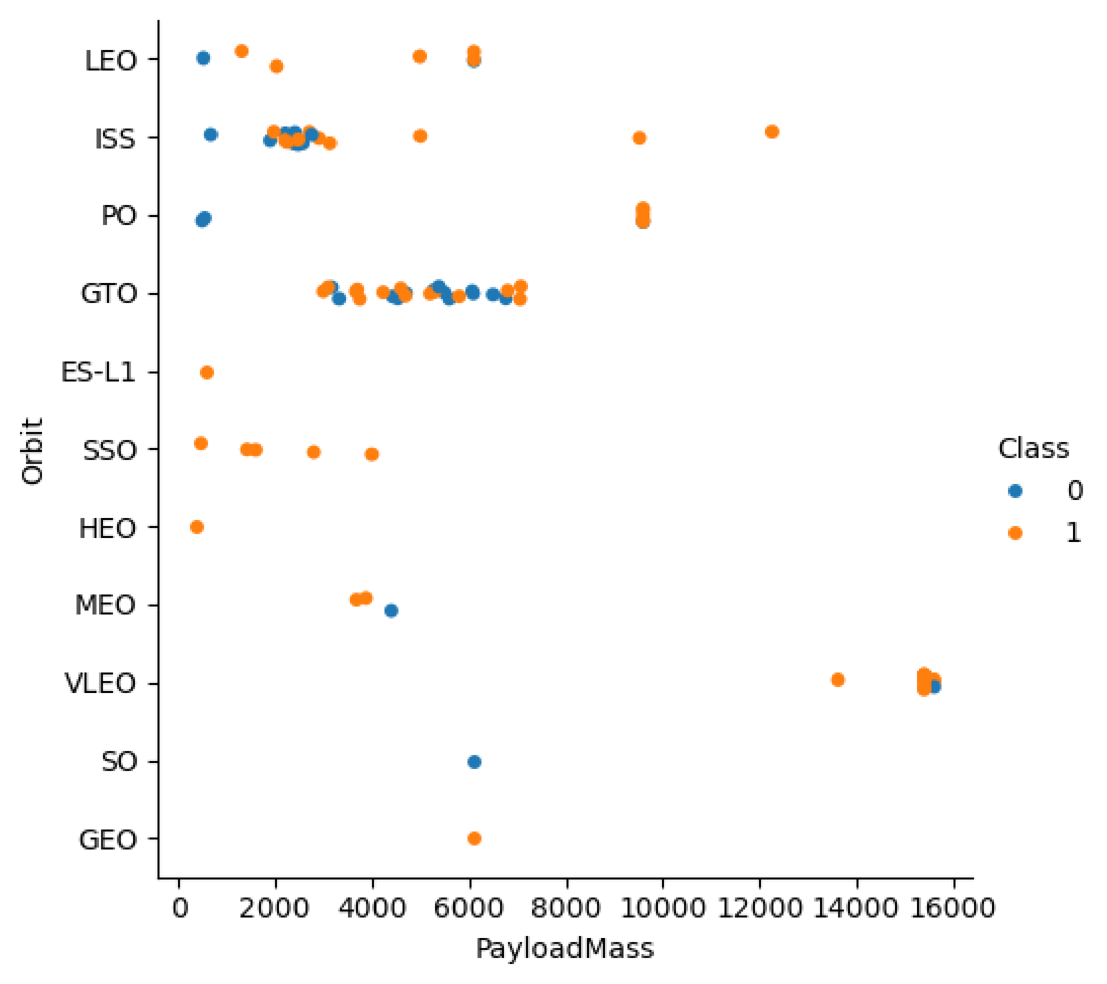
Flight Number vs. Orbit Type

- 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.



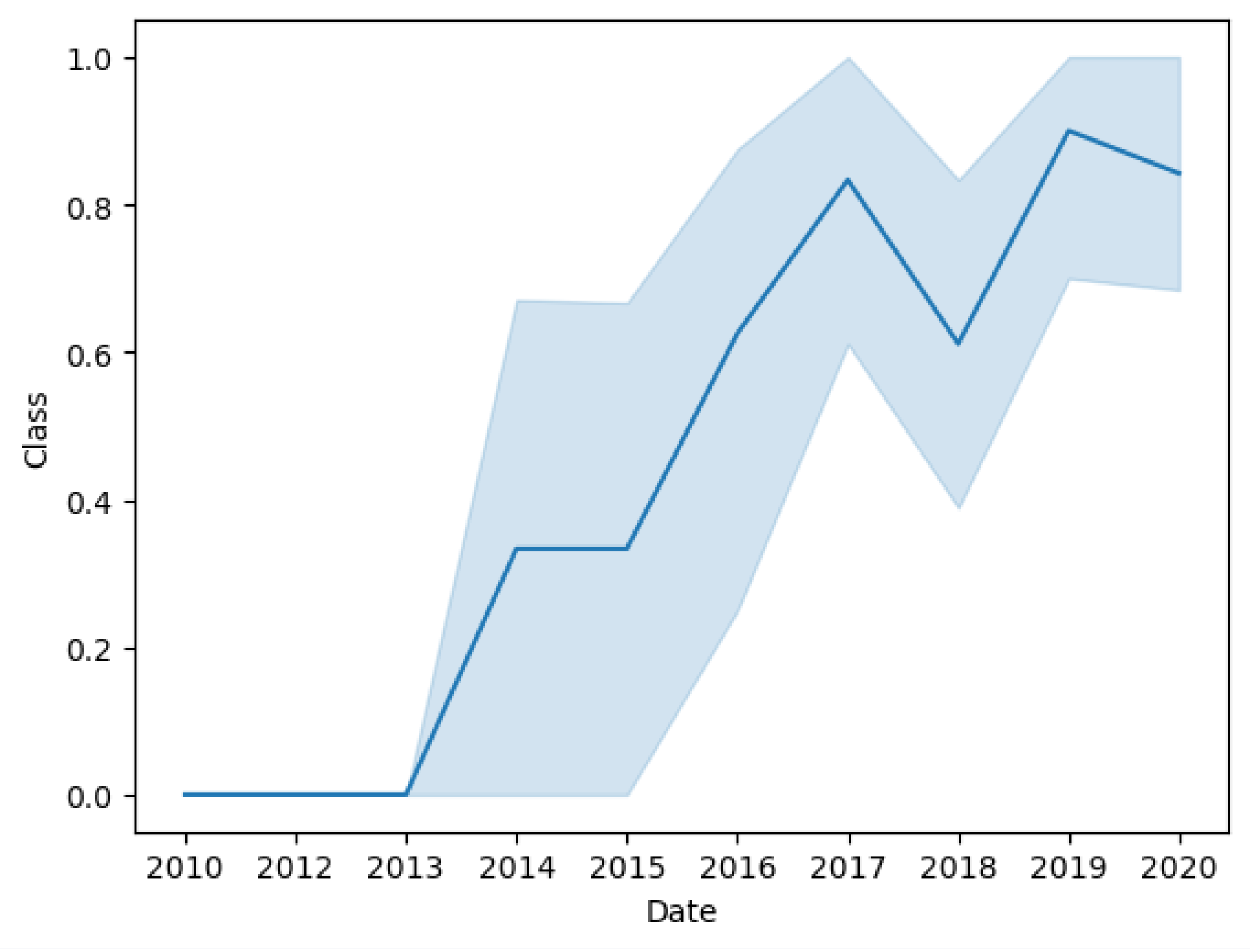
Payload vs. Orbit Type

- 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.



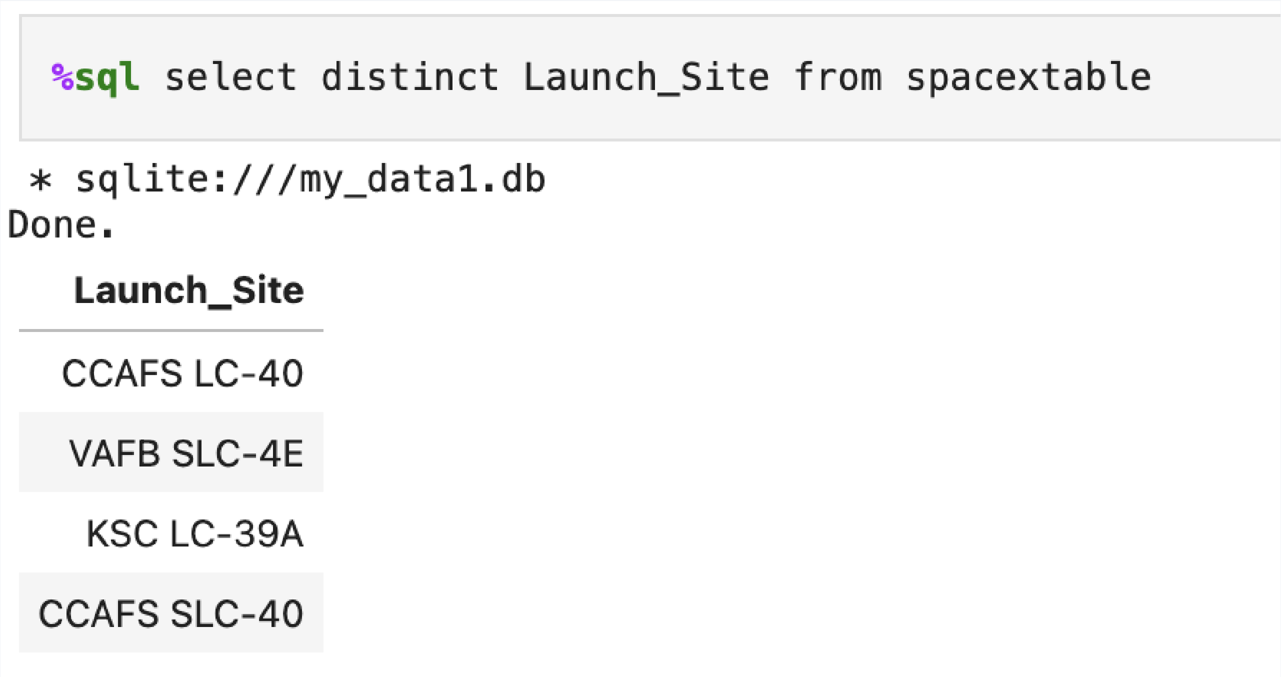
Launch Success Yearly Trend

- Sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- There are 4 unique launch sites which are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

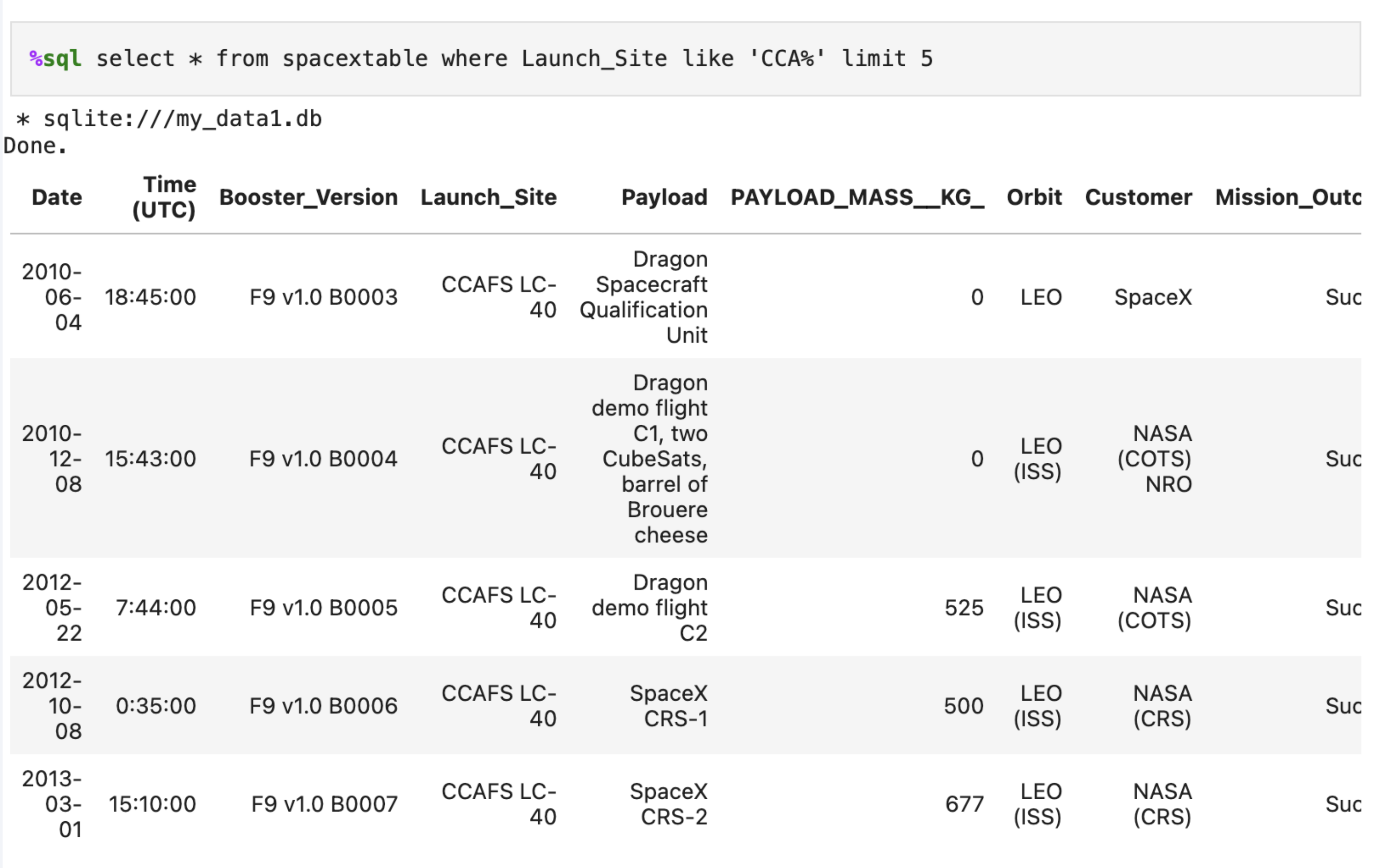


```
%sql select distinct Launch_Site from spacextable
```

```
 * 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'

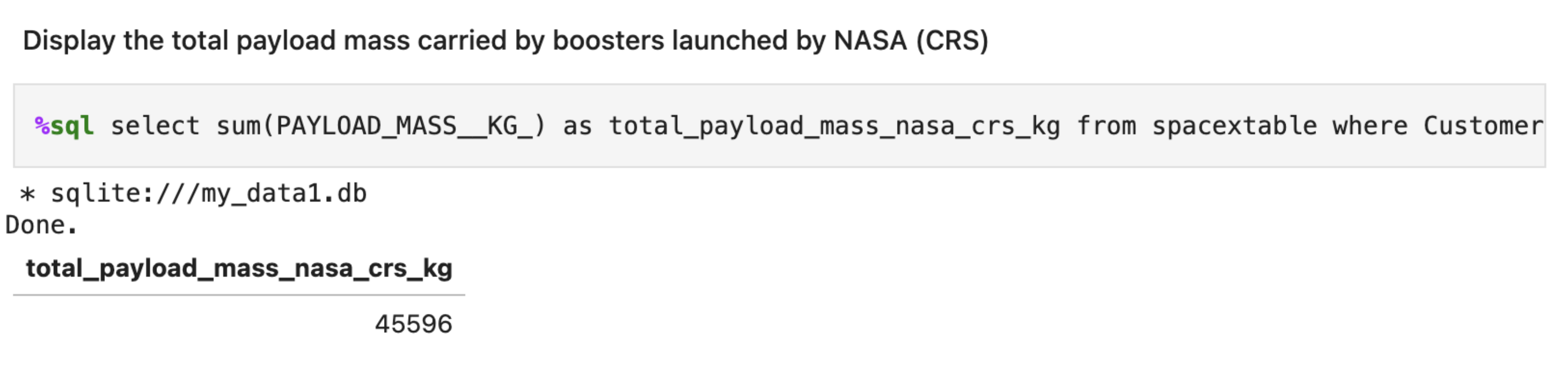


```
%sql select * from spacextable where Launch_Site like 'CCA%' limit 5
```

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

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

Total Payload Mass



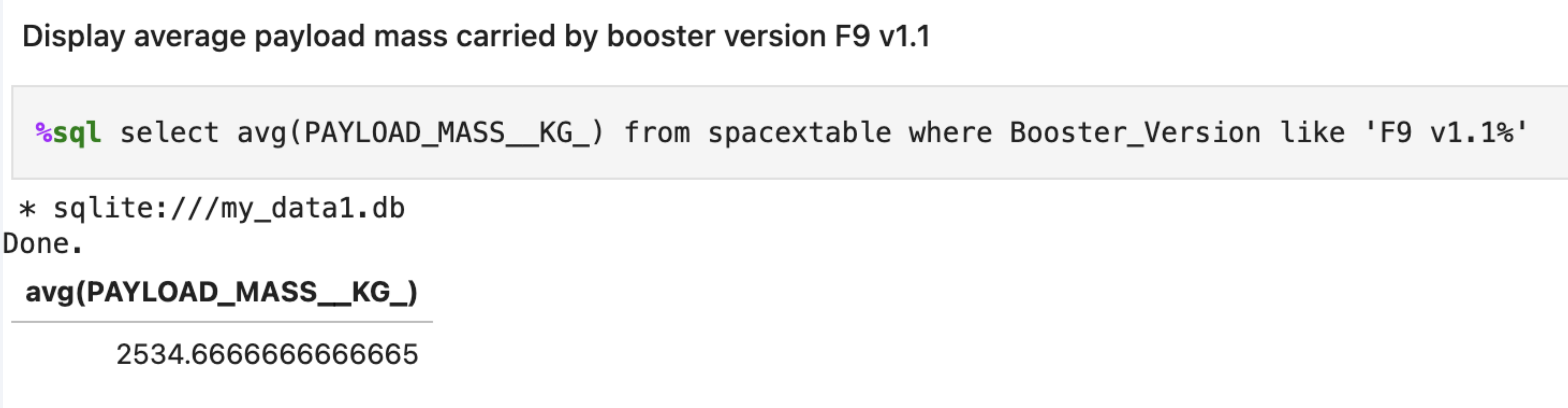
Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass_nasa_crs_kg from spacextable where Customer
```

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

total_payload_mass_nasa_crs_kg
45596

Average Payload Mass by F9 v1.1



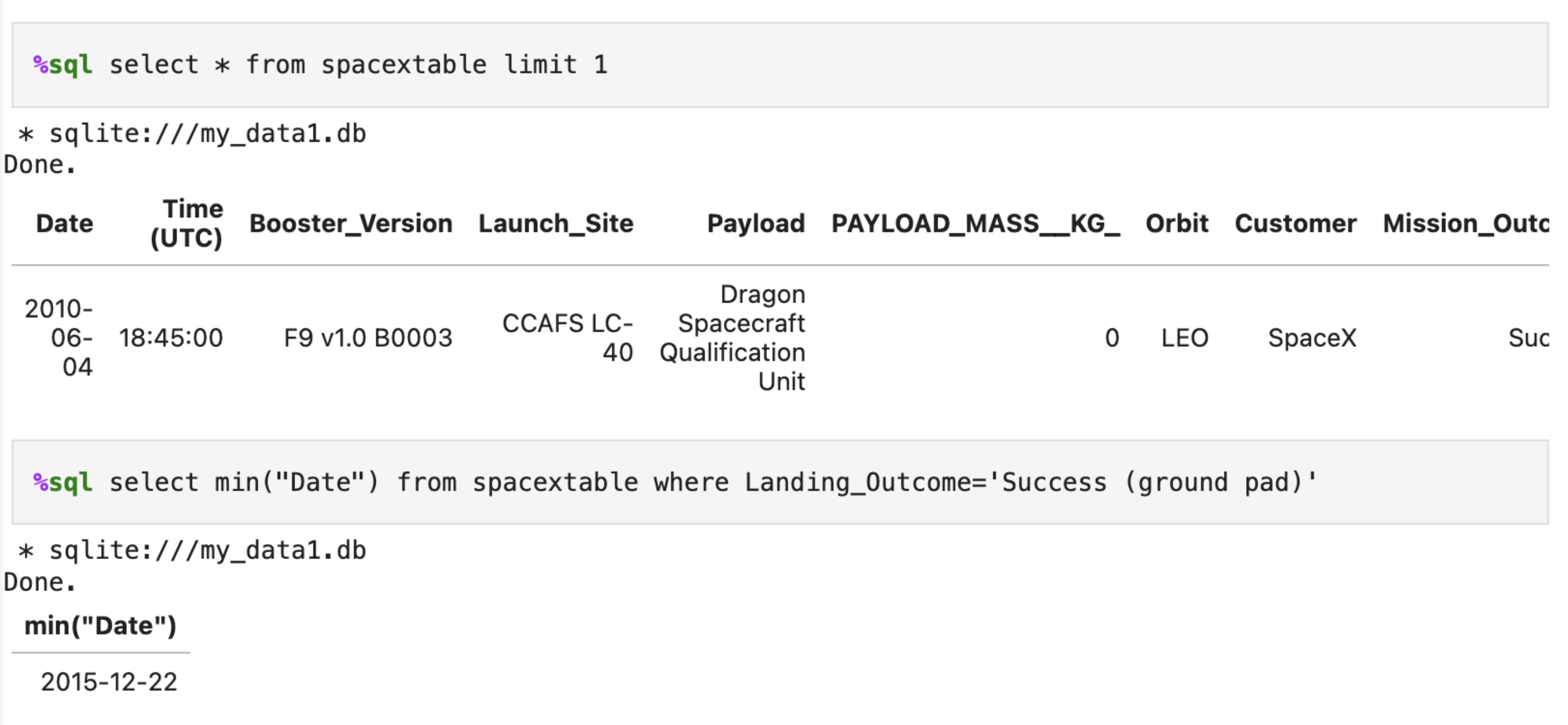
Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from spacextable where Booster_Version like 'F9 v1.1%'
```

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

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date



```
%sql select * from spacextable limit 1
```

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

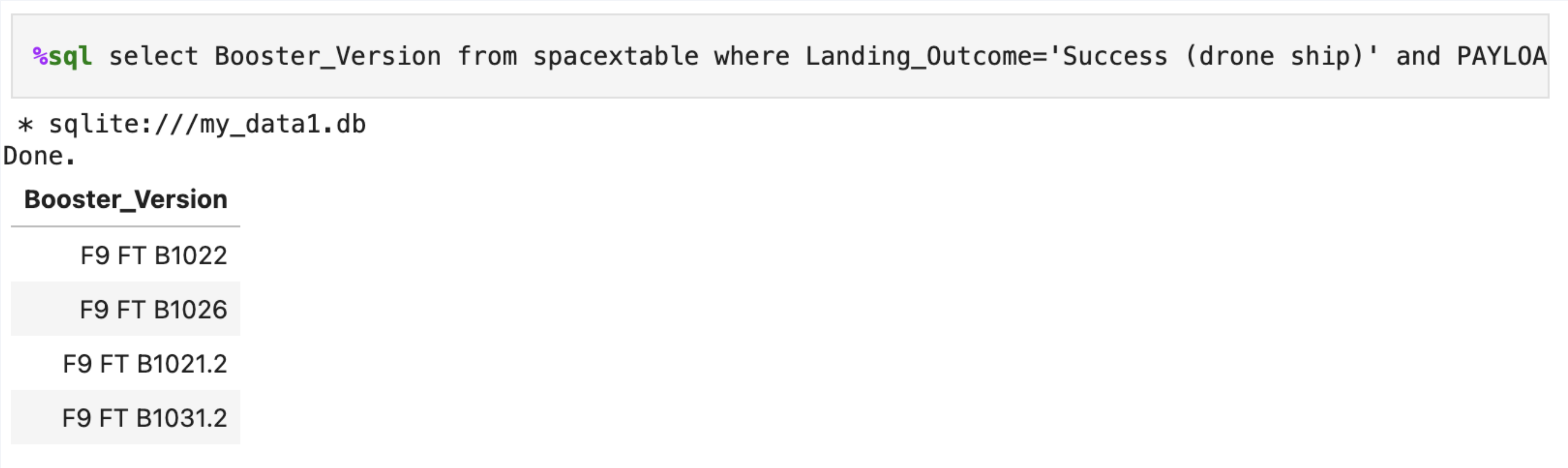
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Suc

```
%sql select min("Date") from spacextable where Landing_Outcome='Success (ground pad)'
```

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

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

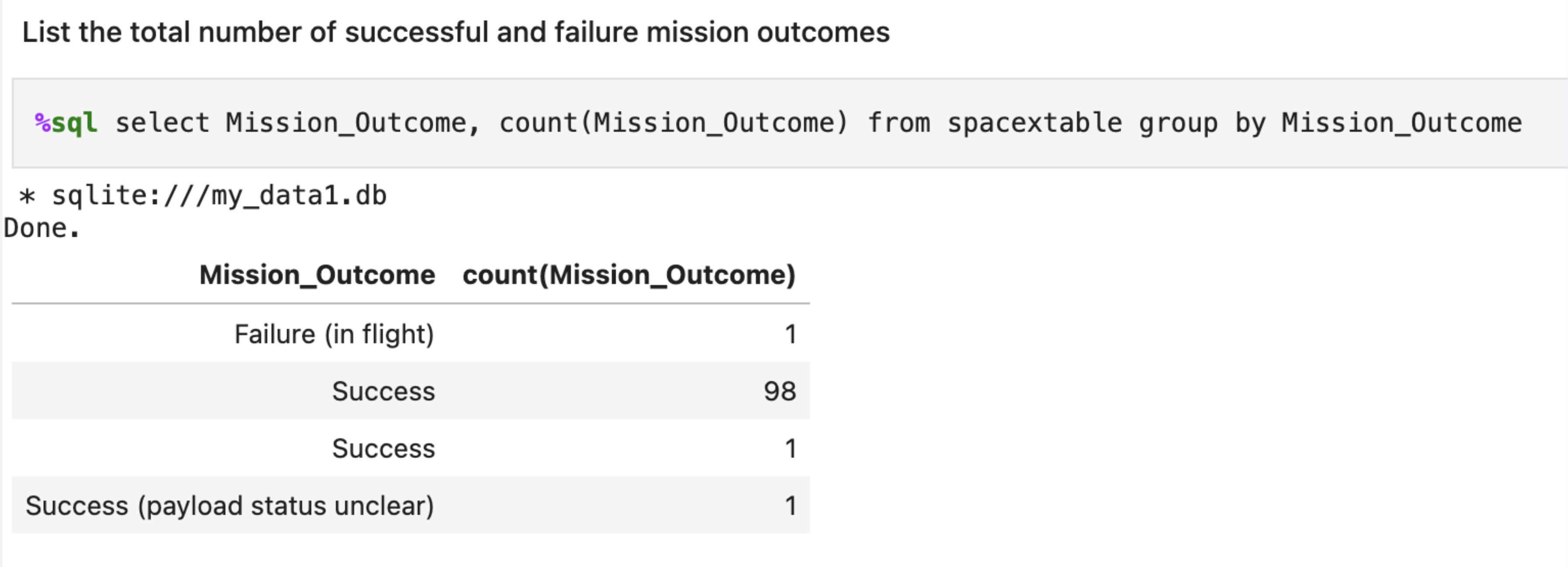


```
%sql select Booster_Version from spacextable where Landing_Outcome='Success (drone ship)' and PAYLOA
```

```
 * 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



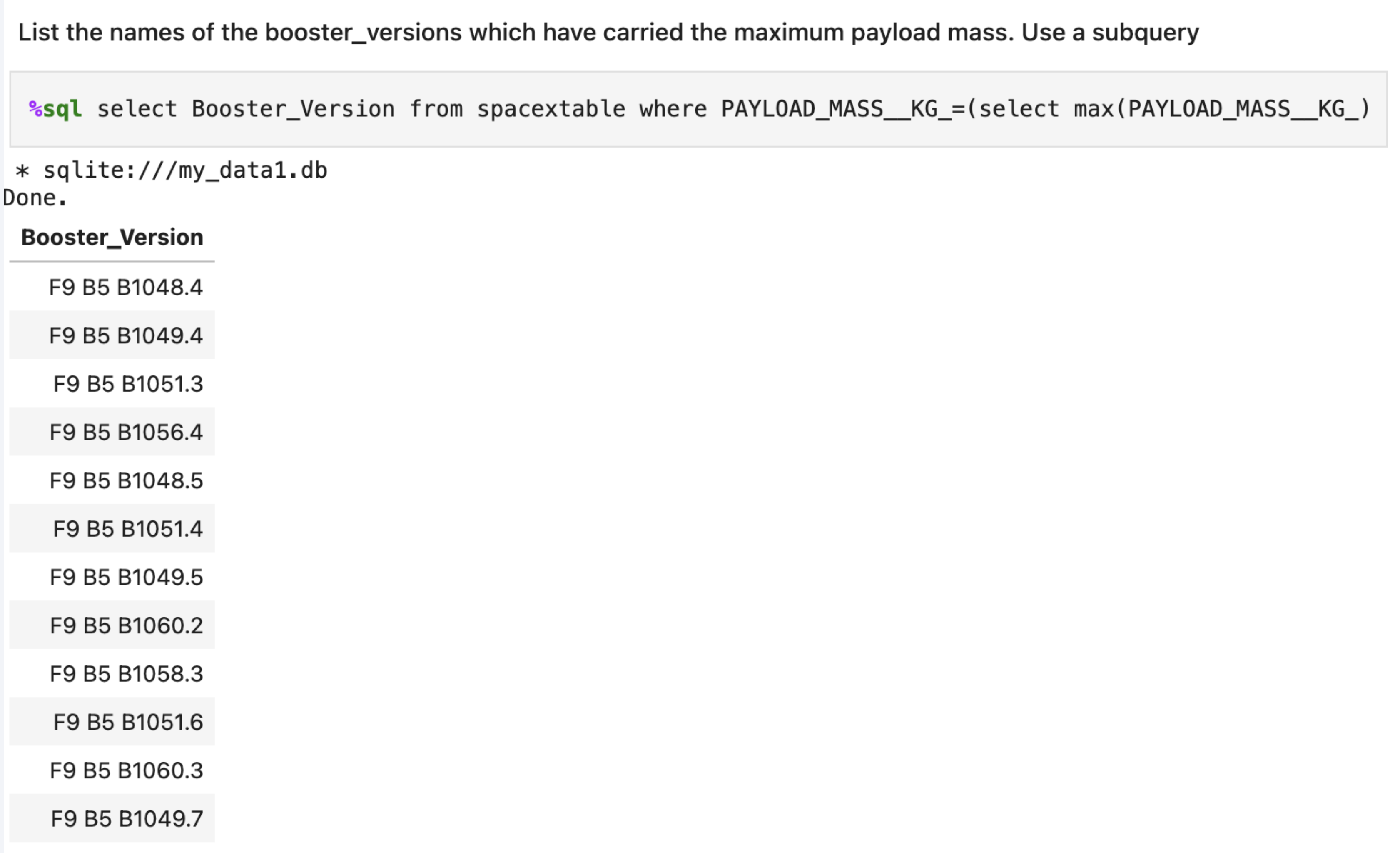
List the total number of successful and failure mission outcomes

```
%sql select Mission_Outcome, count(Mission_Outcome) from spacextable group by Mission_Outcome
```

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

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload



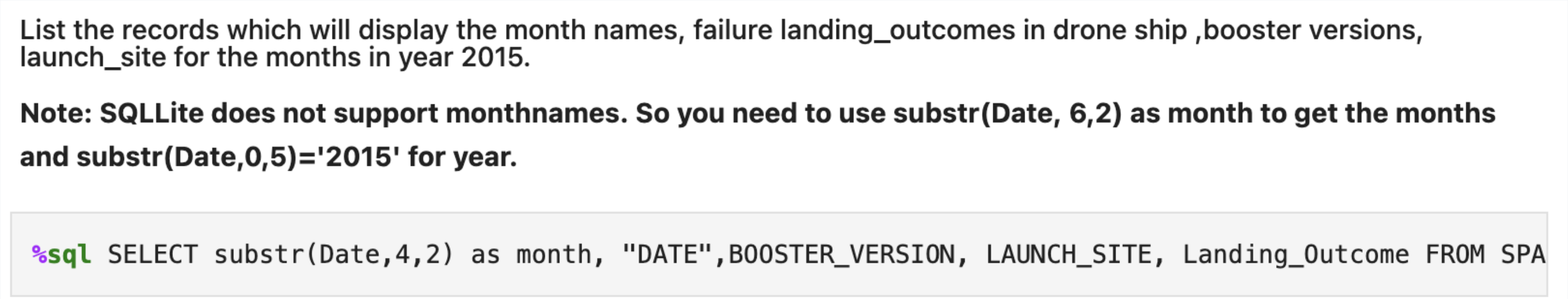
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

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

```
 * 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

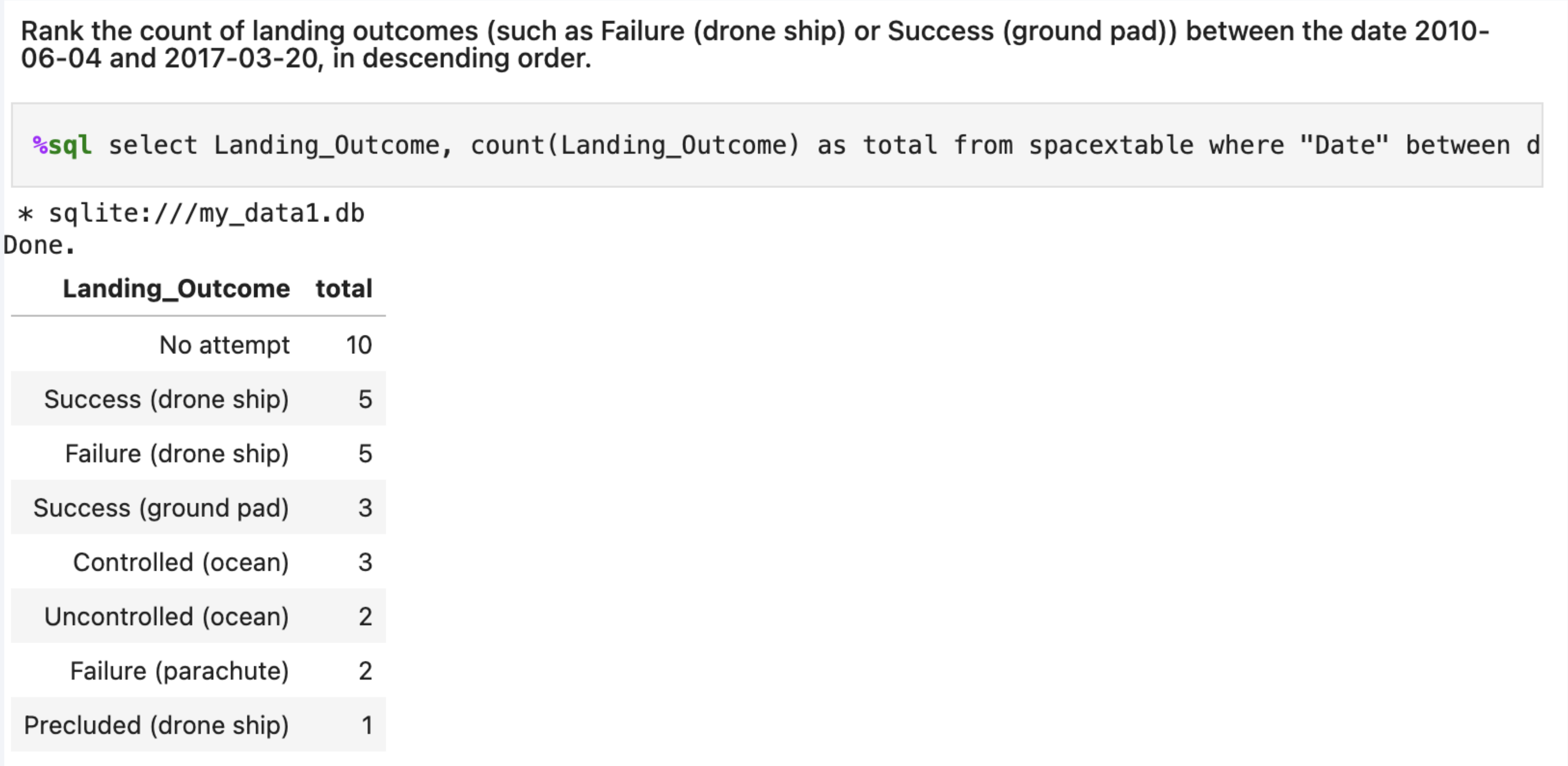


List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date,4,2) as month, "DATE",BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPA
```

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



Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

```
%sql select Landing_Outcome, count(Landing_Outcome) as total from spacextable where "Date" between d
```

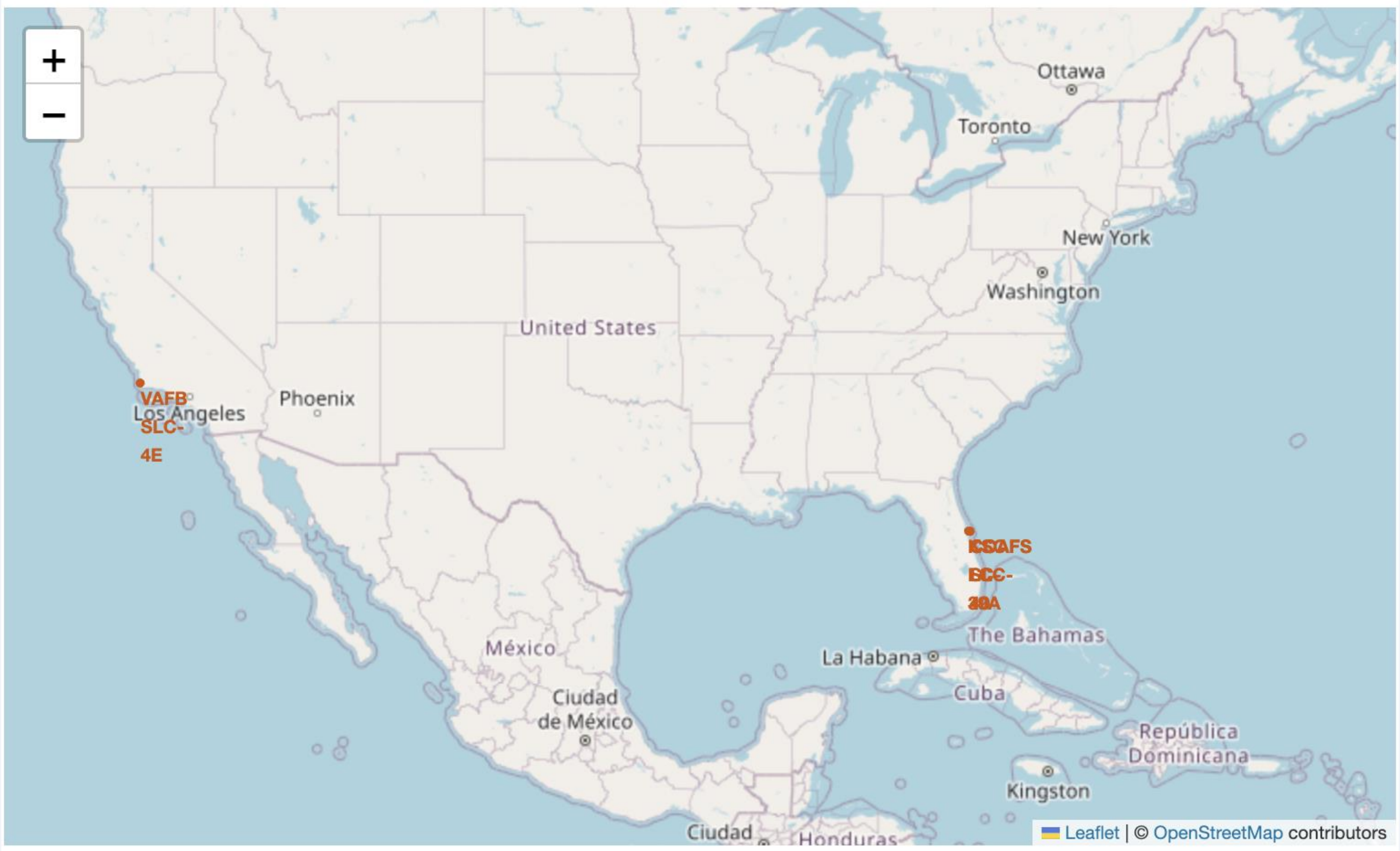
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	total
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

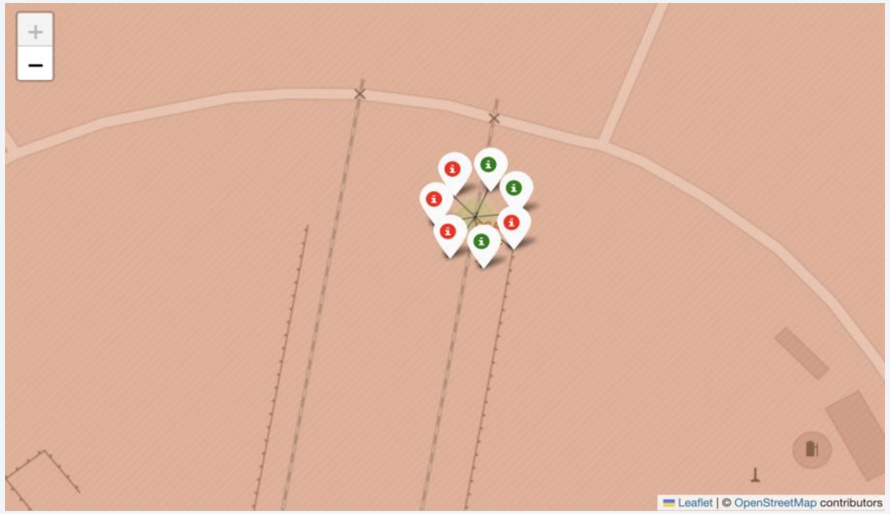
Section 3

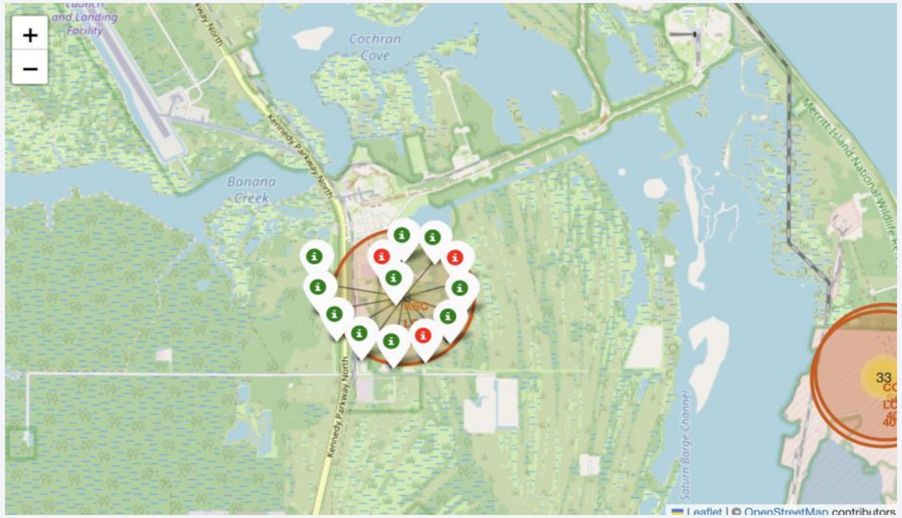
Launch Sites Proximities Analysis

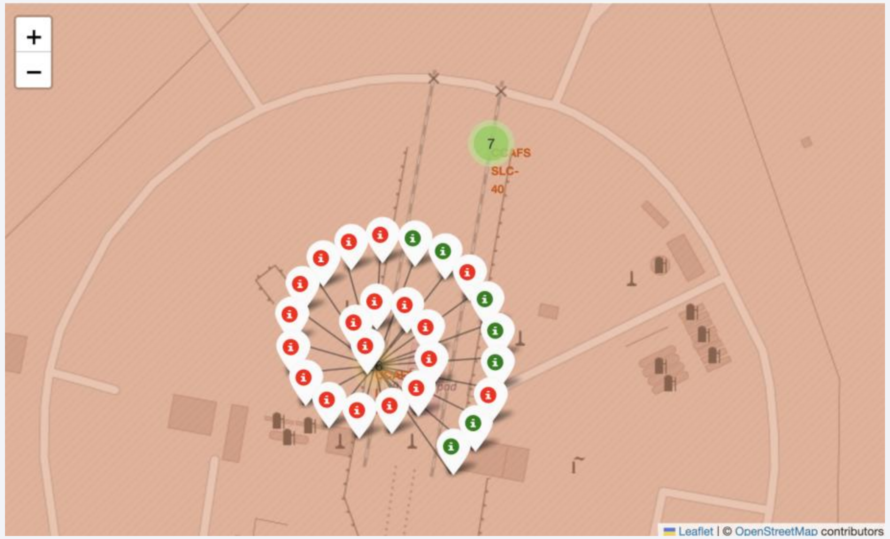
Launch Sites Map

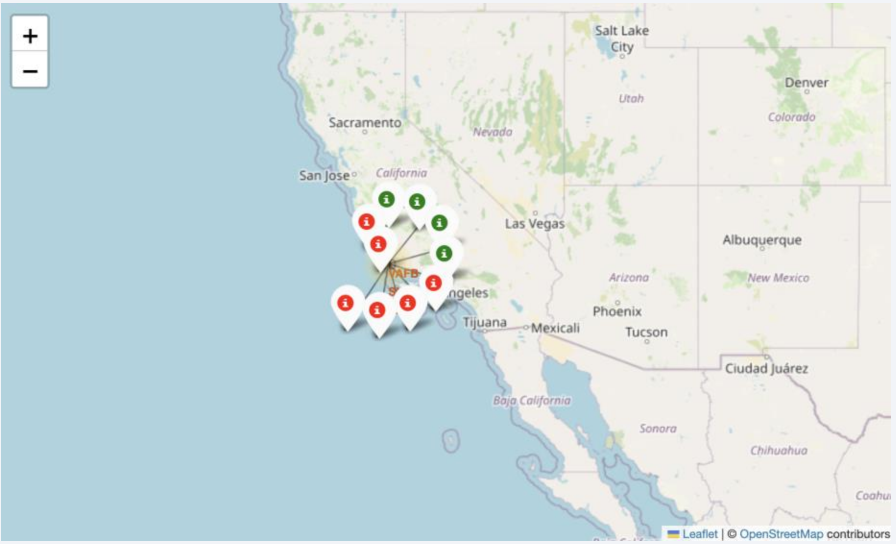


Launch Outcomes

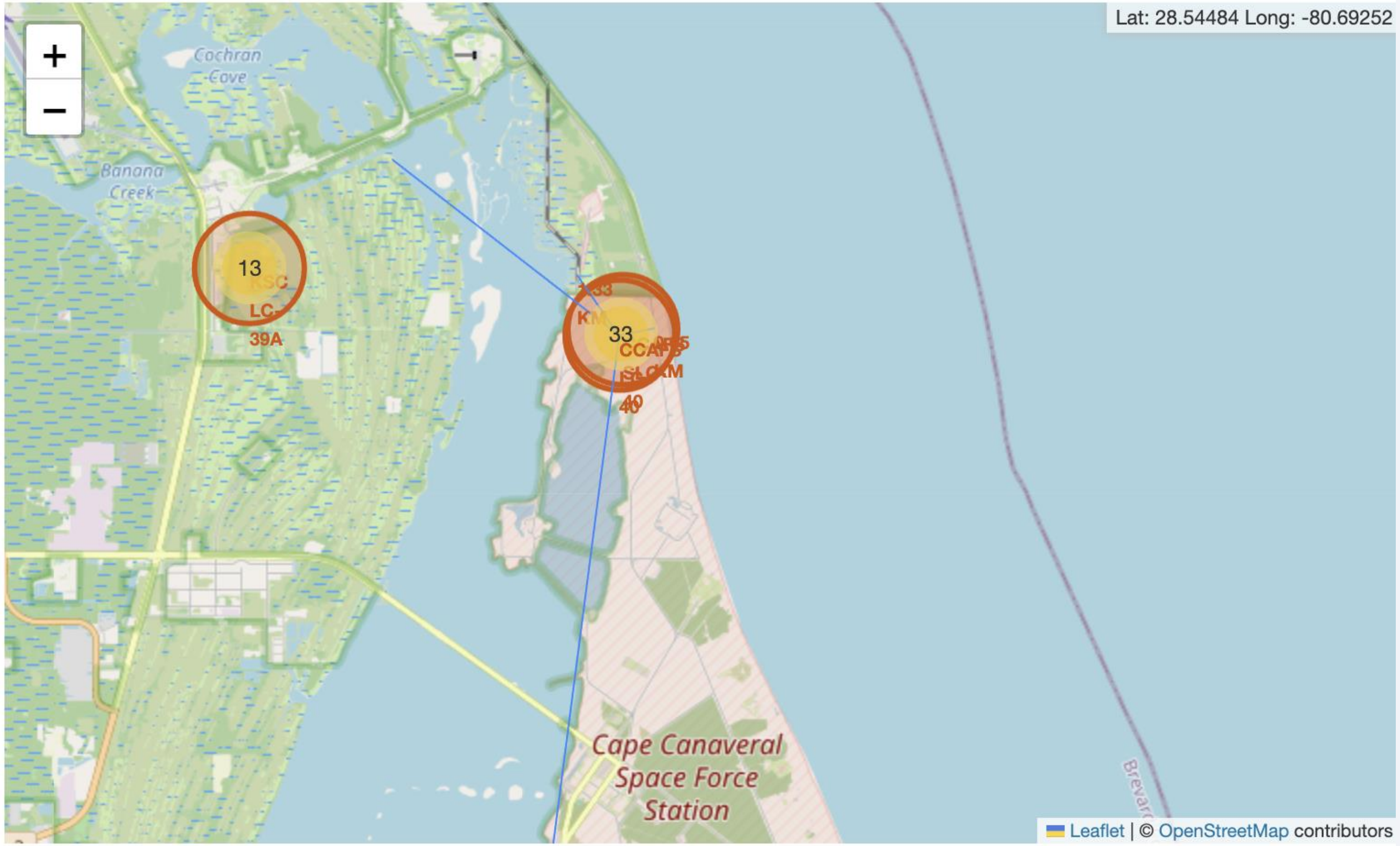


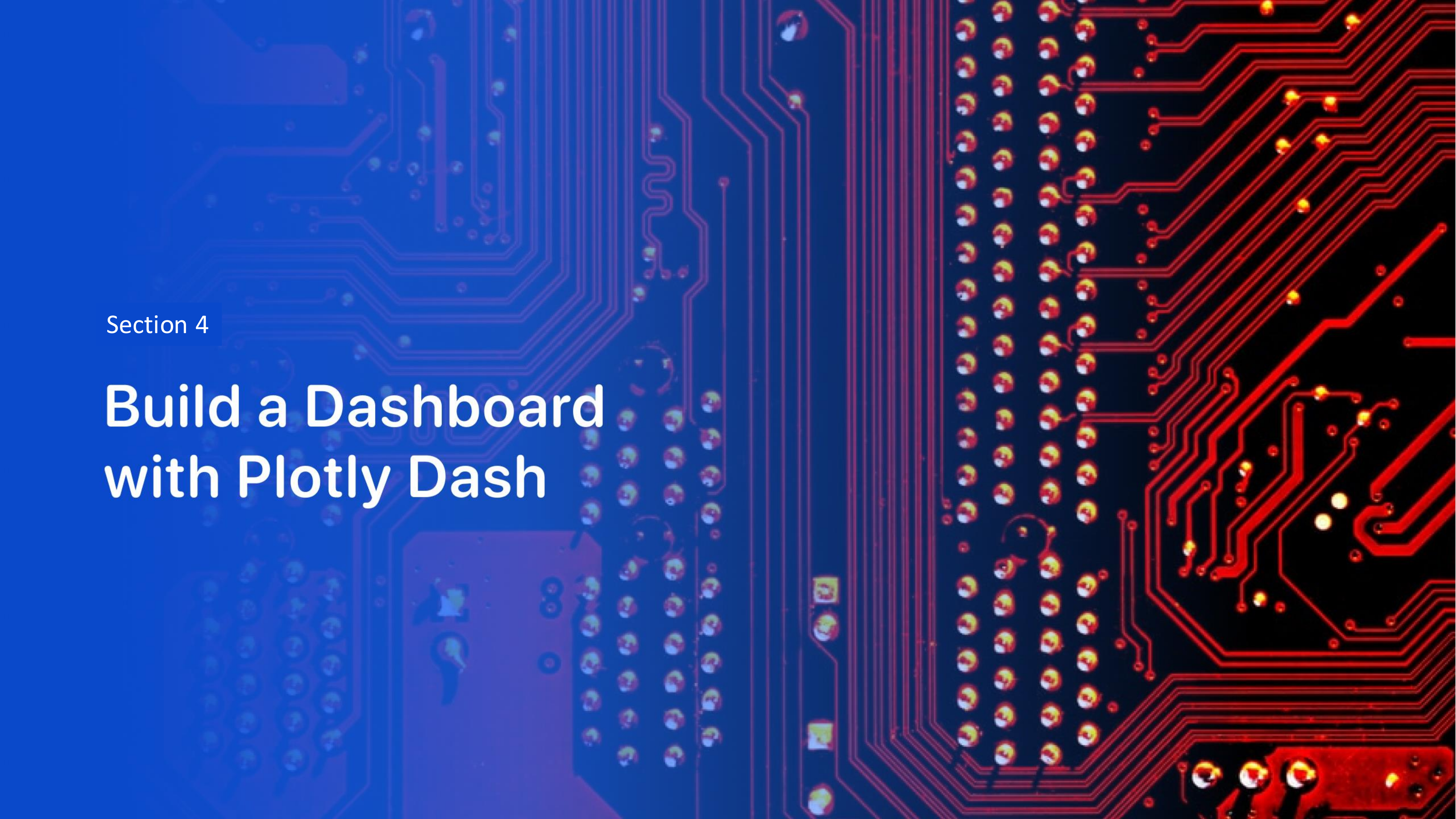

+
−
Launch and Landing Facility
Cochran Cove
Banana Creek
Kennedy Parkway North
KSC
33
Merritt Island National Wildlife
Saturn Barge Channel
Leaflet | © OpenStreetMap contributors





Launch Sites & Proximities

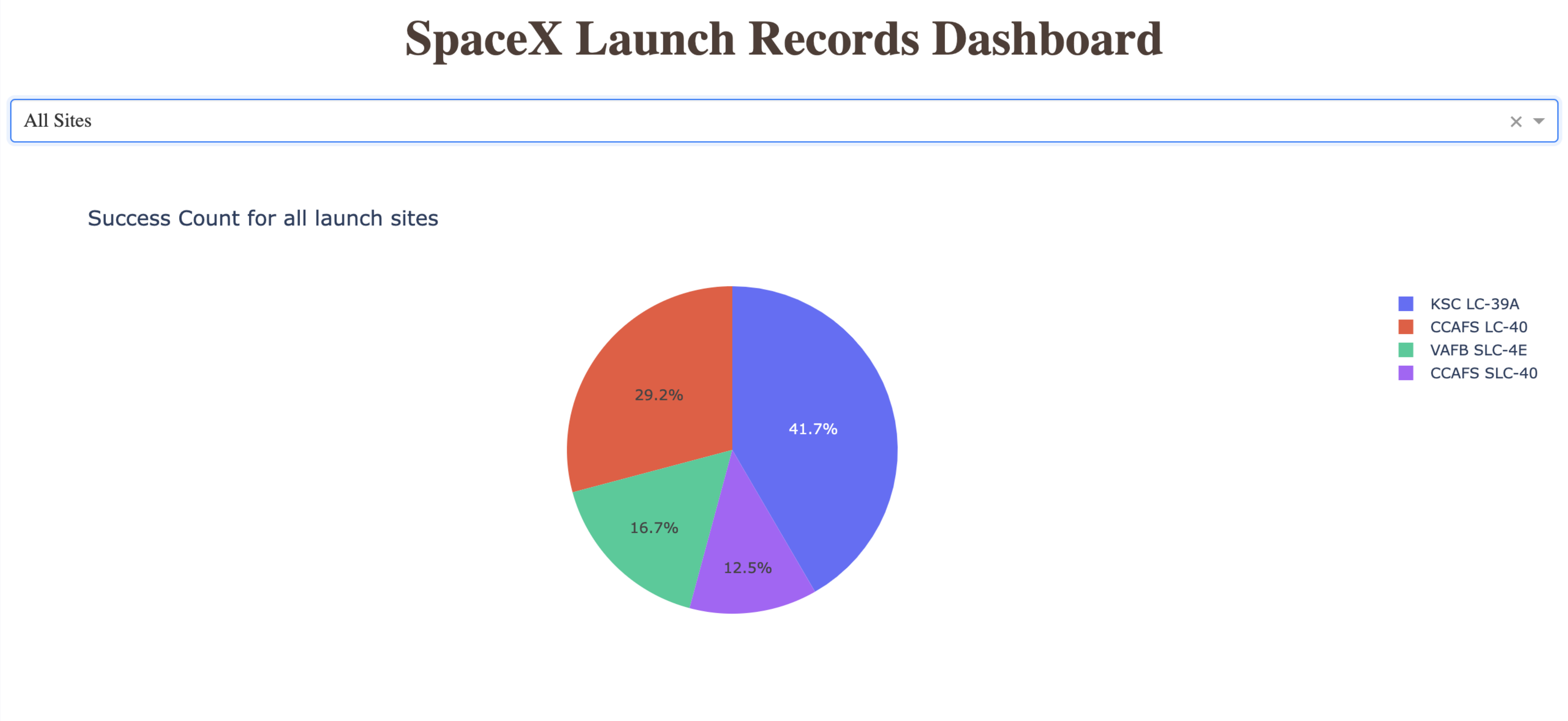

Lat: 28.54484 Long: -80.69252
+
−
Cochran Cove
Banana Creek
13
KSC
LC
39A
33
KM
CCAFS
40
Cape Canaveral Space Force Station
Brevard
Leaflet | © OpenStreetMap contributors



Section 4

Build a Dashboard with Plotly Dash

Launch Success Count for All Sites

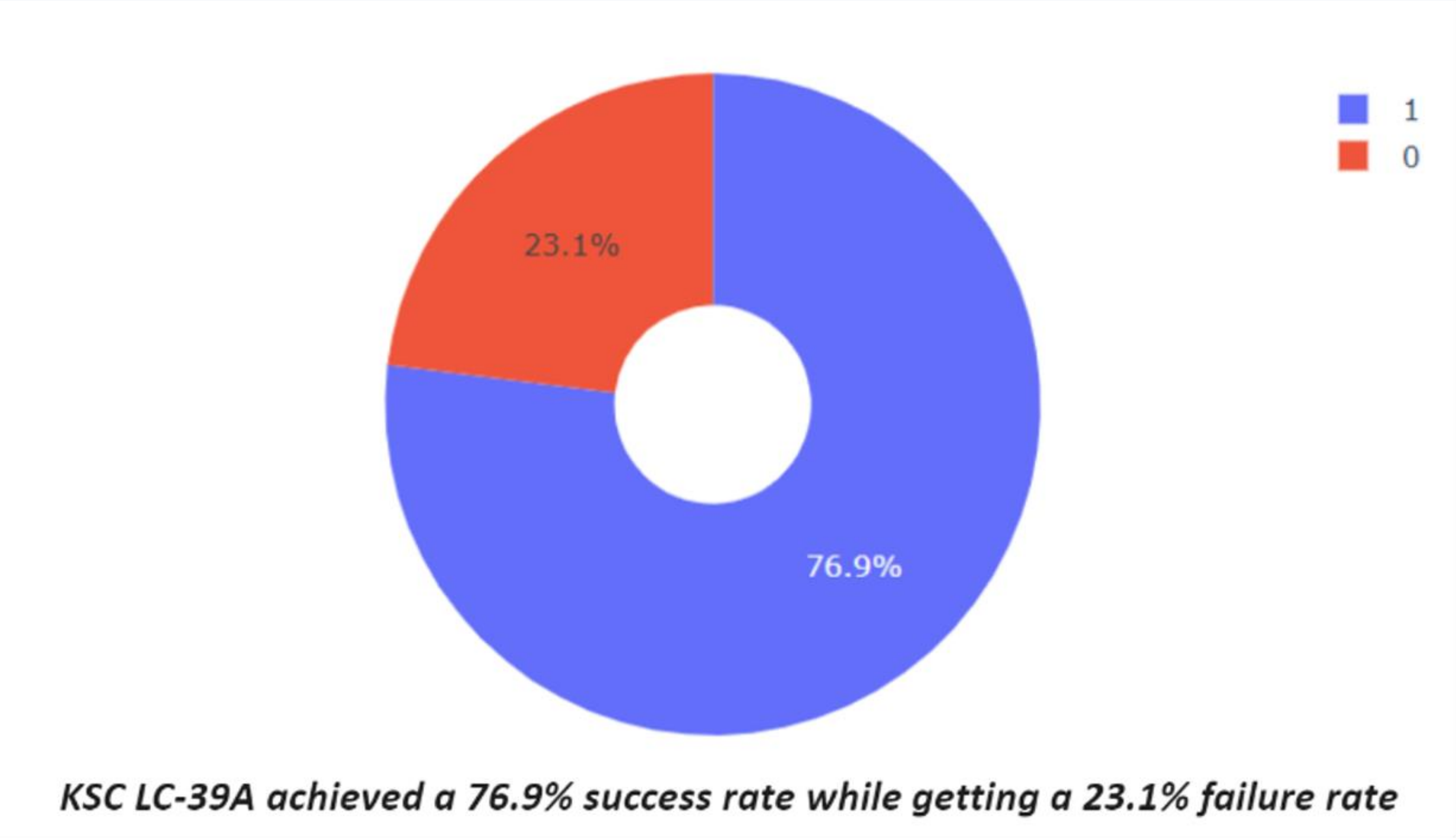


SpaceX Launch Records Dashboard

All Sites

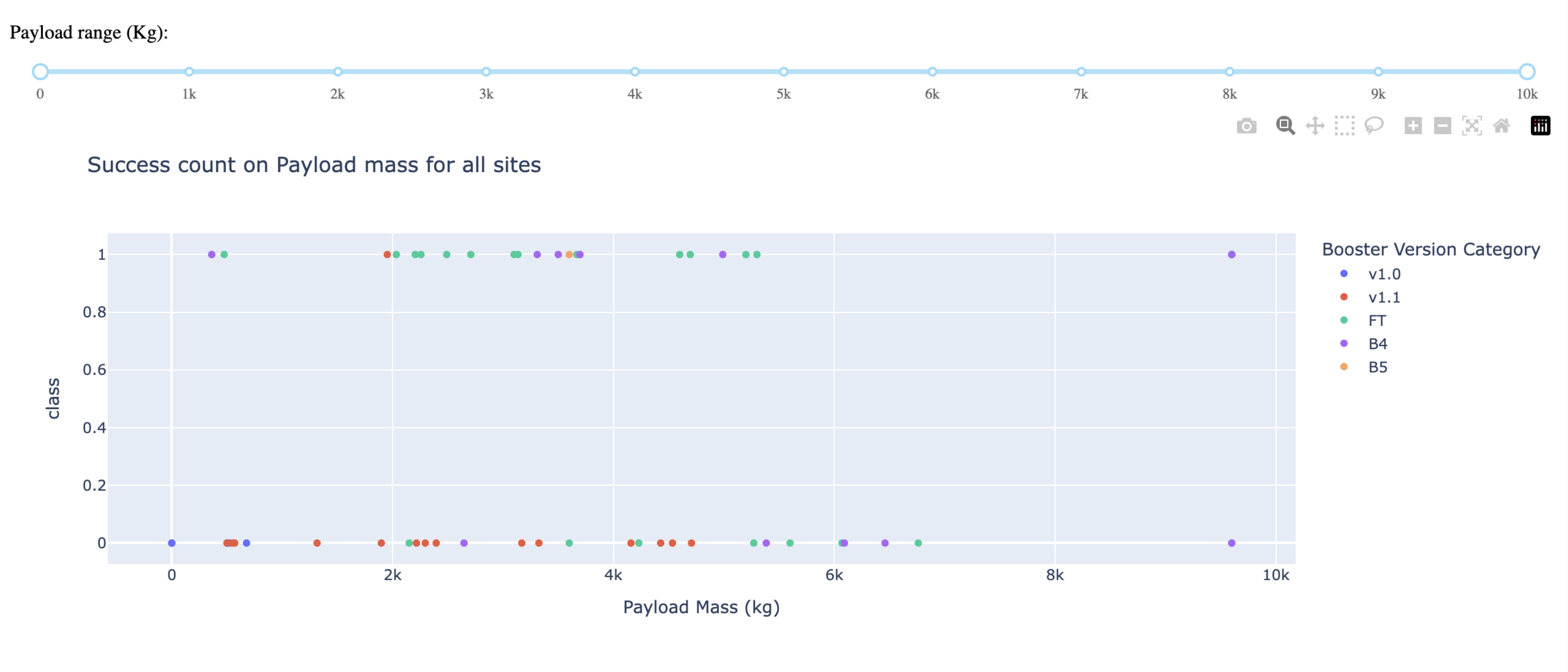
Success Count for all launch sites

Launch Success Rate by Launch Sites



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload VS Launch Outcome



Payload range (Kg):

Section 5

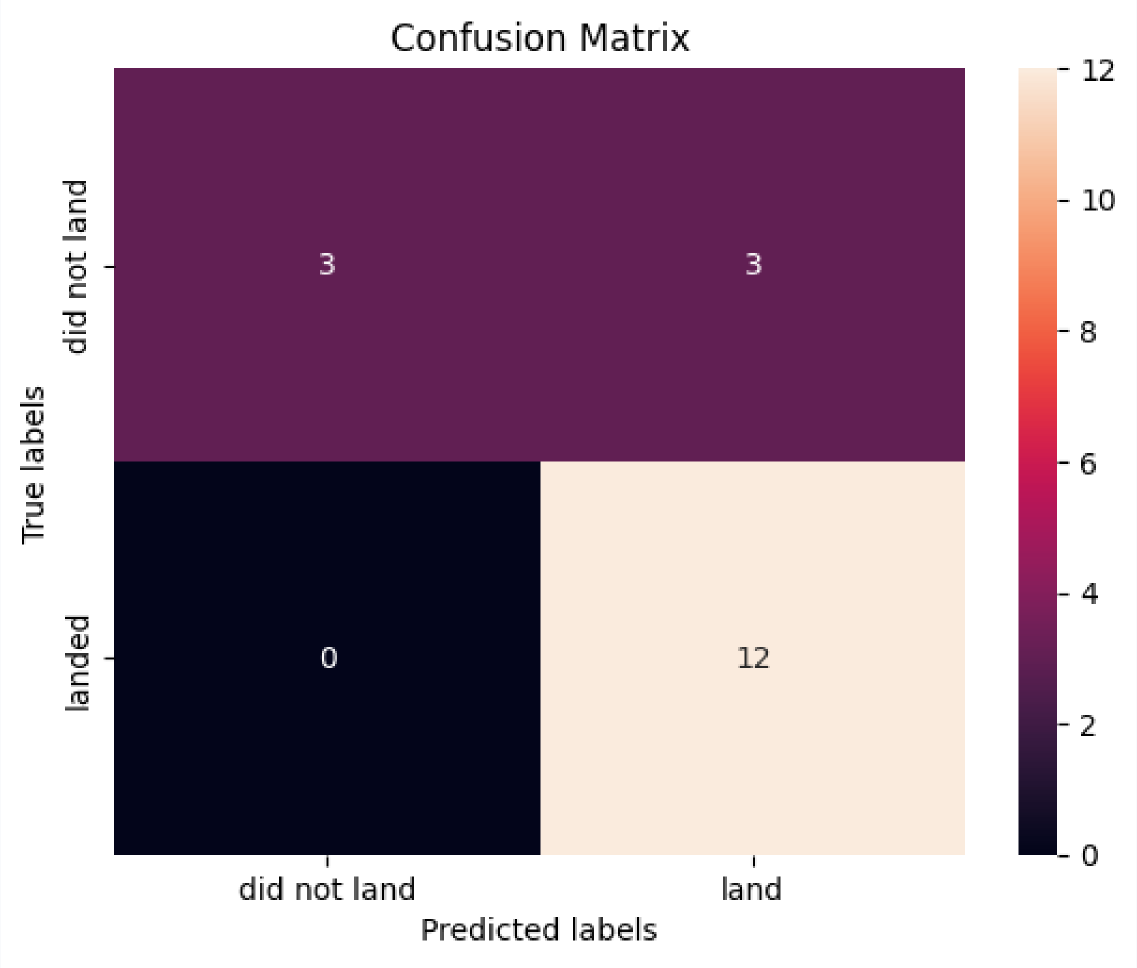
Predictive Analysis (Classification)

Classification Accuracy

- The decision tree classifier have the highest accuracy score which is 87% compared to other classifiers models KNN(84%), SVM(84%) and logistic regression(84%)

Confusion Matrix

- The confusion matrix for the Decision Tree classifier demonstrates its ability to differentiate between the various classes. However, a significant issue is the occurrence of false positives, where the classifier incorrectly labels unsuccessful landings as successful ones.



Conclusions

- A higher number of flights at a launch site correlates with an increased success rate.
- The success rate of launches began rising in 2013 and continued to improve until 2020.
- The orbits ES-L1, GEO, HEO, SSO, and VLEO achieved the highest success rates.
- KSC LC-39A recorded the most successful launches among all sites.
- For this task, the Decision Tree classifier is the most effective machine learning algorithm.

Thank you!