IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Domingo Herrera
October 16 2024

Outline

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

Executive Summary

- **Summary of methodologies**
    - Data Collection from API (Application Programming Interface) and Website using Web Scraping
    - Data Wrangling
    - Exploratory Data Analysis with Data Visualization
    - Exploratory Data Analysis with SQL
    - Interactive Visual Analytics with Folium Library
    - Predictive Analytics with Machine Learning
- **Summary of all results**
    - Exploratory Data Analysis
    - Interactive Analytics with Obtained Results
    - Predictive Analytics Model

Introduction

- Project background and context
  - SpaceX can do rockets launches relatively inexpensive. SpaceX Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.
- Problems you want to find answers
  - How can we determine the price of each launch ?
    - We will do this by gathering information about Space X and creating dashboards for the team.
  - Can we predict if SpaceX will reuse the first stage ?
    - We will train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We collected the data from two sources: Space's REST API and Wikipedia Falcon 9 Heavy Launches Website.
- Perform data wrangling
  - We downloaded data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome from Spacex's API, transforming them from JSON to dataframe. Later, we combined this data with Wikipedia's Falcon 9 Launches (parsing them from HTML to table), dealing with null values to create a dataset for data analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Determine if the Falcon 9's first stage will land, identifying some attributes that can be used to determine if the first stage can be reused. We can then use these features with machine learning to automatically predict if the first stage can land successfully.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Build Dashboards and Maps for stakeholders to interact with visual analytics, explore and manipulate data to find more insights from the SpaceX dataset more easily than with static graphs.
- Perform predictive analysis using classification models
  - Build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully using Preprocessing data stage, Train, Test and Split the dataset, find hyperparameters with Grid Search, determine the model with the best accuracy for Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.

Data Collection

- Describe how data sets were collected.

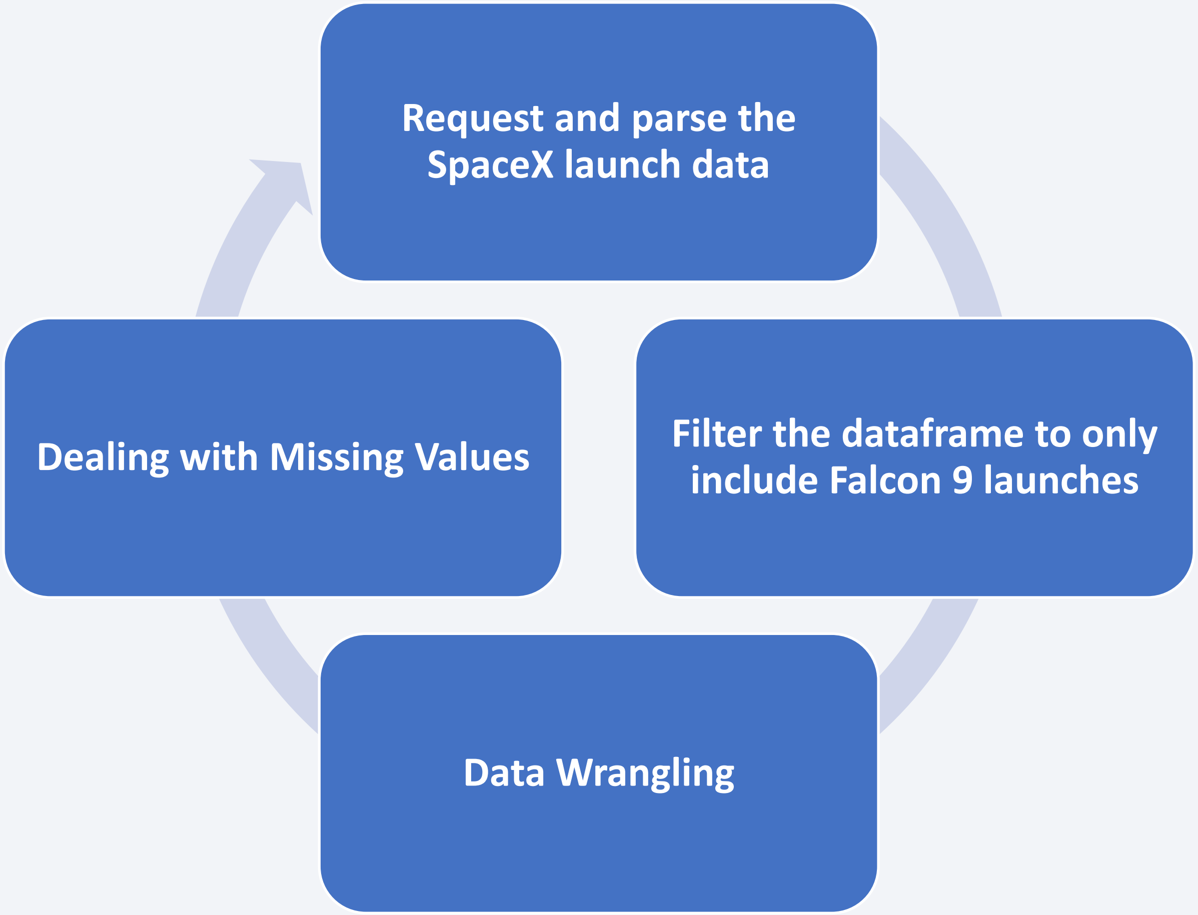
  Data was collected from two different sources: SpaceX Rest API and Wikipedia Falcon X launch information history. We merge both dataset into one to be used to train, test and built machine learning model.

- You need to present your data collection process use key phrases and flowcharts
  - Process:
    1. Connect to data sources: API and Website
    2. Get data from data sources into two dataframes
    3. Clean datasets and merge datasets into one to be to create machine learning models.

Data Collection – SpaceX API

- For data collection, we request and parse the SpaceX launch data, filter the dataframe to only include Falcon 9 launches, we check missing values (data wrangling) and values counts and handled missing values.

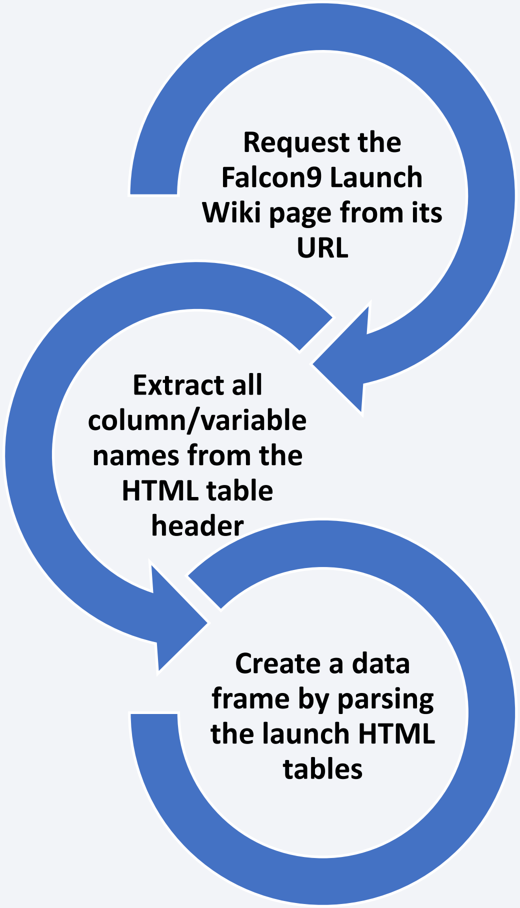
- The GitHub URL of the completed SpaceX API calls notebook is https://github.com/domingohr/Projects_DS/blob/main/jupyter-labs-spacex-data-collection-api.ipynb.


Request and parse the SpaceX launch data
Filter the dataframe to only include Falcon 9 launches
Data Wrangling
Dealing with Missing Values

Data Collection - Scraping

- We extracted a Falcon 9 launch records HTML table from Wikipedia and parse the table and convert it into a Pandas data frame.

- The GitHub URL of the completed web scraping notebook is https://github.com/domingohr/Projects_DS/blob/main/jupyter-labs-webscraping.ipynb.



Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

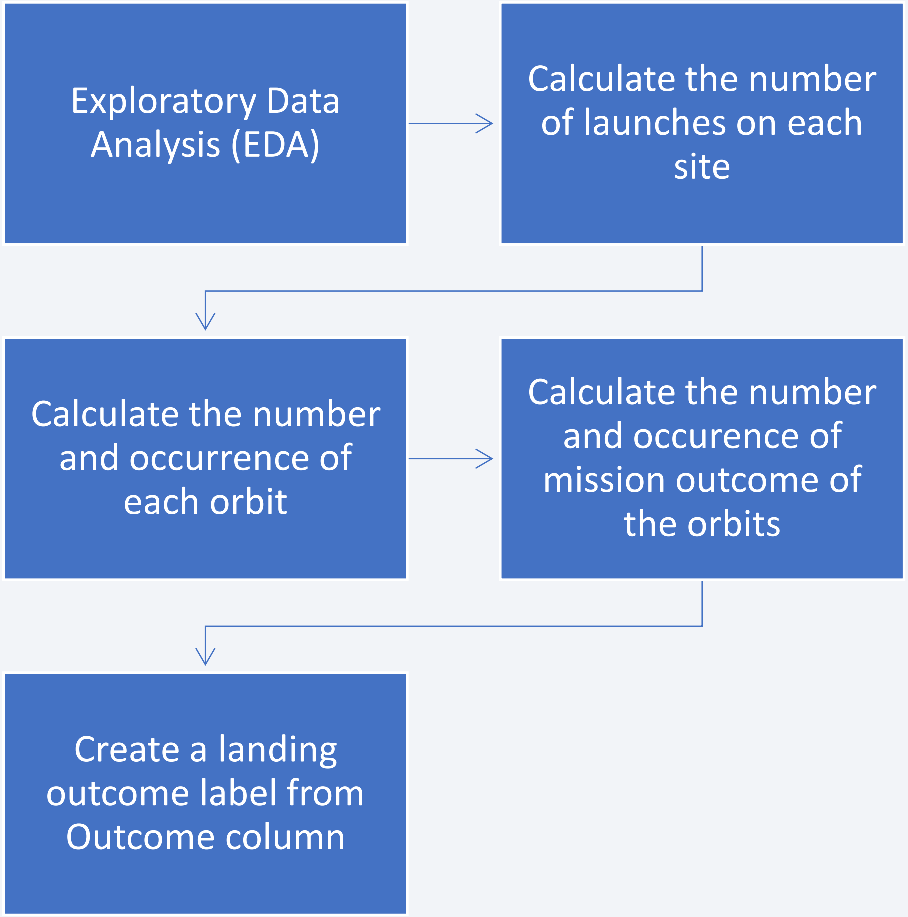
Create a data frame by parsing the launch HTML tables

Data Wrangling

- We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and determine what would be the label for training supervised models.

- We calculate the number of launches on each site, later calculate the number and occurrence of each orbit, estimate the number and occurrence of mission outcome of the orbits and finally, we create a landing outcome label from outcome column.

- The GitHub URL of your completed data wrangling is https://github.com/domingohr/Projects_DS/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb


Exploratory Data Analysis (EDA)
Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate the number and occurence of mission outcome of the orbits
Create a landing outcome label from Outcome column

EDA with Data Visualization

- We plotted chats for visualize the relationship between a couple of variables like Flight Number and Launch Site, Payload Mass and Launch Site, Flight Number and Orbit type, and Payload Mass and Orbit type. Also plot chats to see variable trends, with variable like success rate of each orbit type, launch success yearly trend and others.

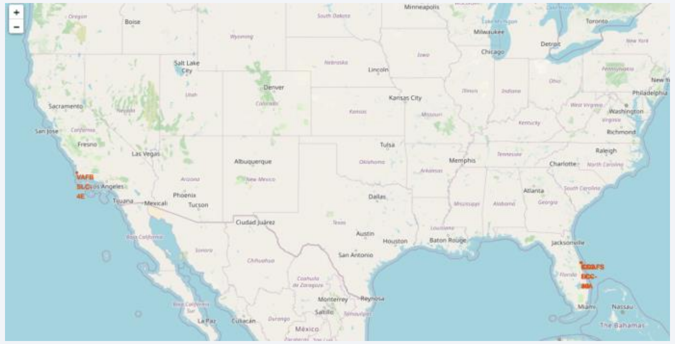
- The GitHub URL of your completed EDA with data visualization notebook is https://github.com/domingohr/Projects_DS/blob/main/edadataviz.ipynb

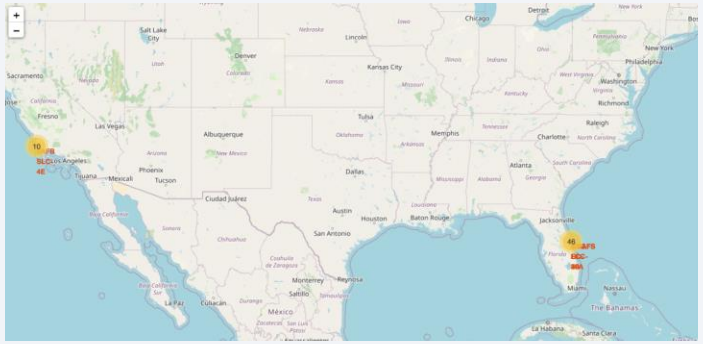
EDA with SQL

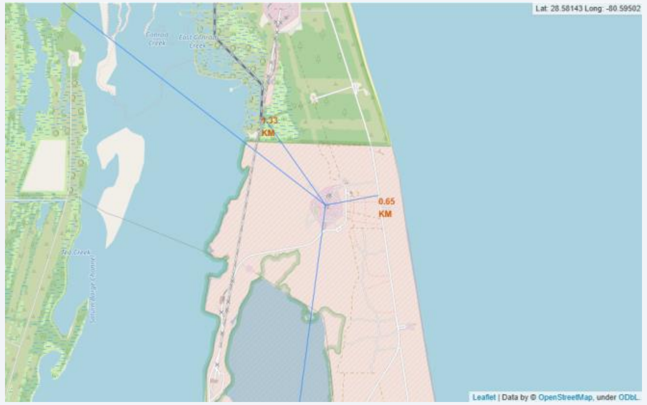
- The summarize the SQL queries you performed is:
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string 'CCA'**
  - **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - **Display average payload mass carried by booster version F9 v1.1**
  - **List the date when the first succesful landing outcome in ground pad was acheived.**
  - **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. Use a subquery**
  - **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.**
  - **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.**
- The GitHub URL of your completed EDA with SQL notebook is https://github.com/domingohr/Projects_DS/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- To find some geographical patterns about launch sites, we mark all launch sites on a map, mark the success/failed launches for each site on the map and Calculate the distances between a launch site to its proximities.
- The GitHub URL of your completed interactive map with Folium map is https://github.com/domingohr/Projects_DS/blob/main/lab_jupyter_launch_site_location.ipynb

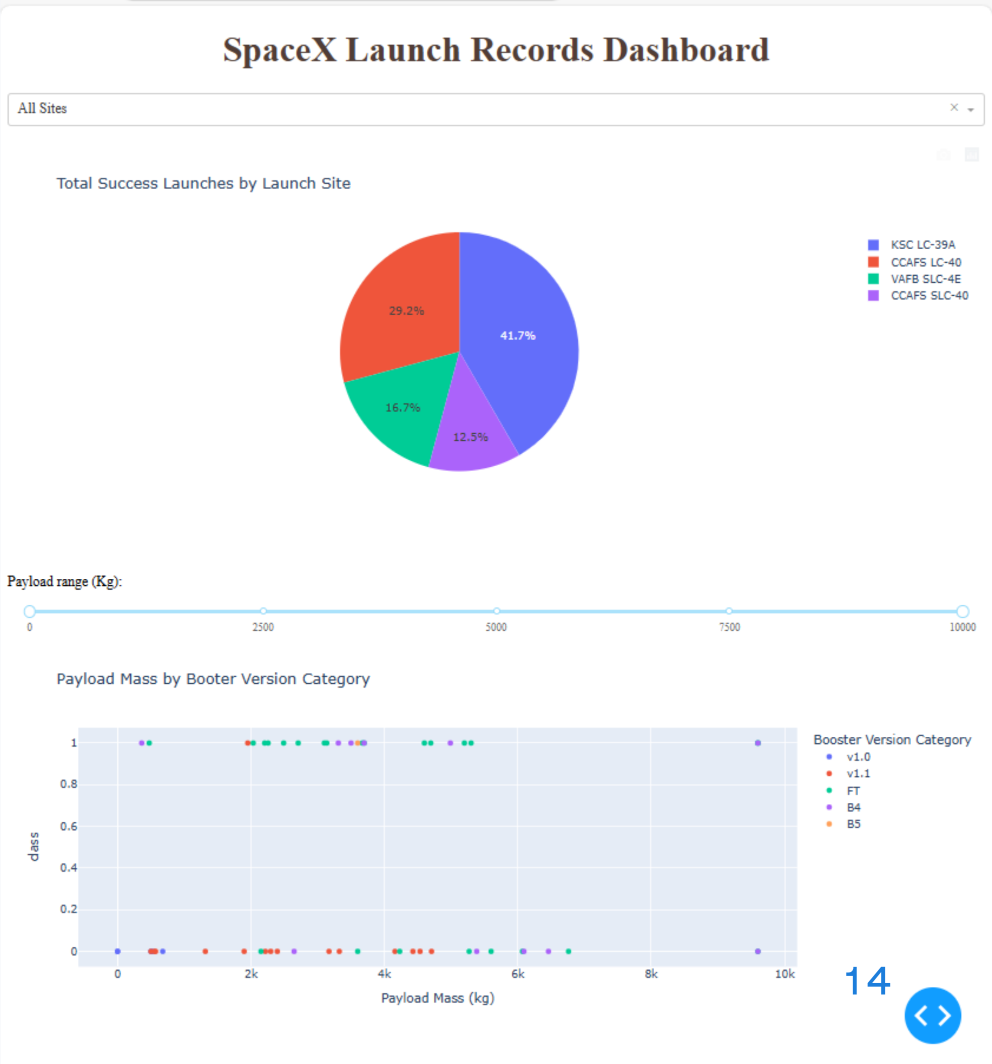





Lat: 28.58143 Long: -80.59502
1.33 KM
0.65 KM
Leaflet | Data by © OpenStreetMap, under ODbL.

Build a Dashboard with Plotly Dash

- We builded a Plotly Dash application for users to perform interactive visual analytics on SpaceX launch data in real-time. This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- The GitHub URL of your completed Plotly Dash lab is https://github.com/domingohr/Projects_DS/blob/main/spacex_dash_app.py



SpaceX Launch Records Dashboard

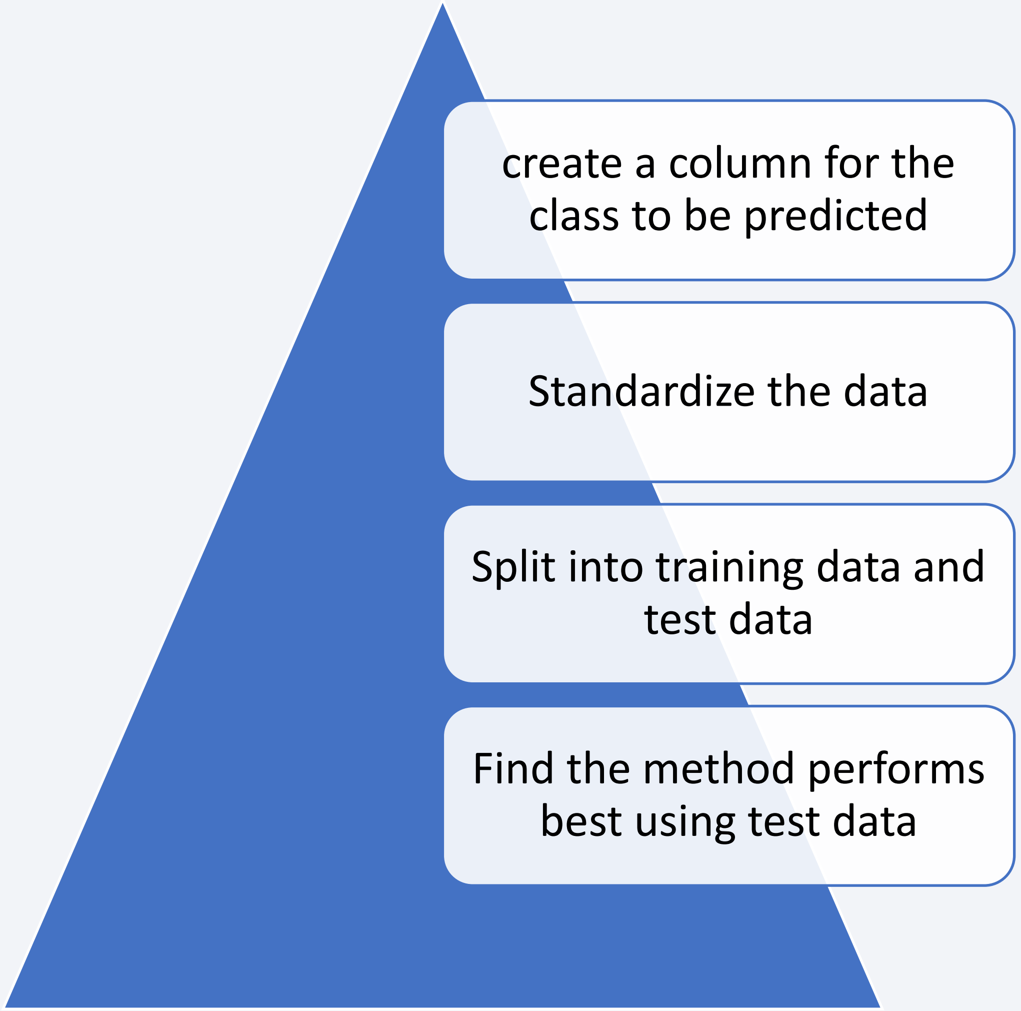
All Sites

Payload range (Kg):

Predictive Analysis (Classification)

- We builted our predictive analysis, loading the data into pandas *dataframes*, split the data into training and test set, create different model and evaluated each one through cross validation techniques to found the best performing classification model.

- The GitHub URL of completed predictive analysis lab is https://github.com/domingohr/Projects_DS/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



create a column for the class to be predicted

Standardize the data

Split into training data and test data

Find the method performs best using test data

Results

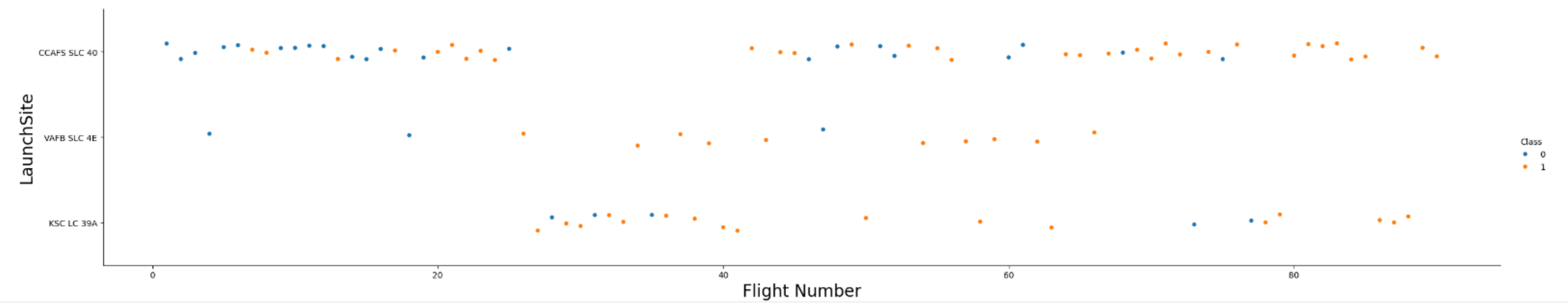
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

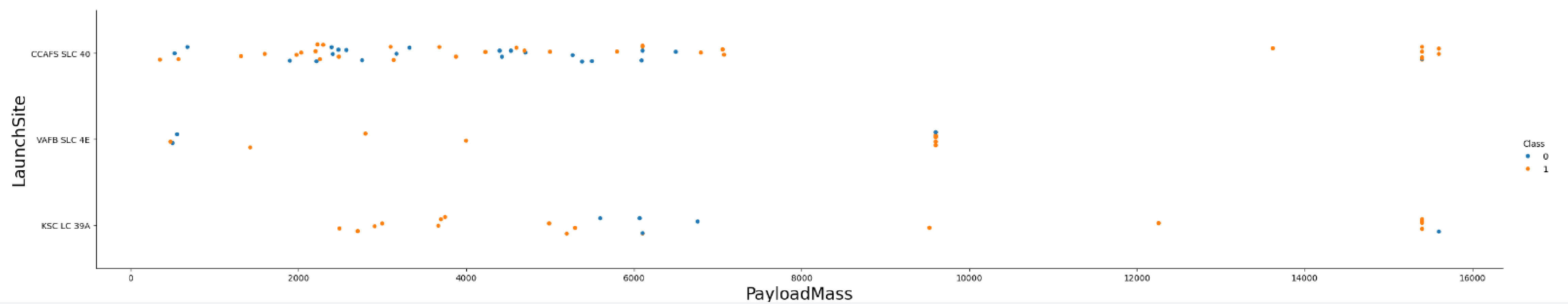
Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



Payload vs. Launch Site

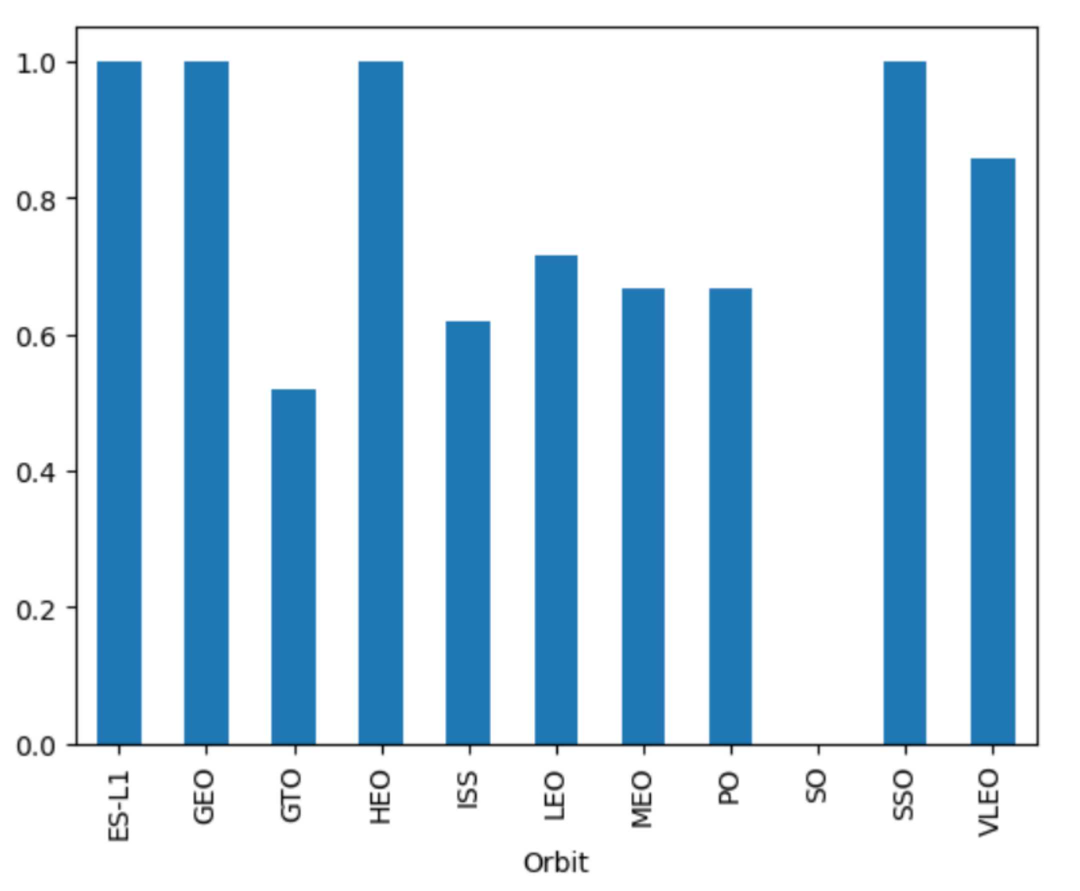
- Scatter plot of Payload vs. Launch Site



The scatter point chart show the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).

Success Rate vs. Orbit Type

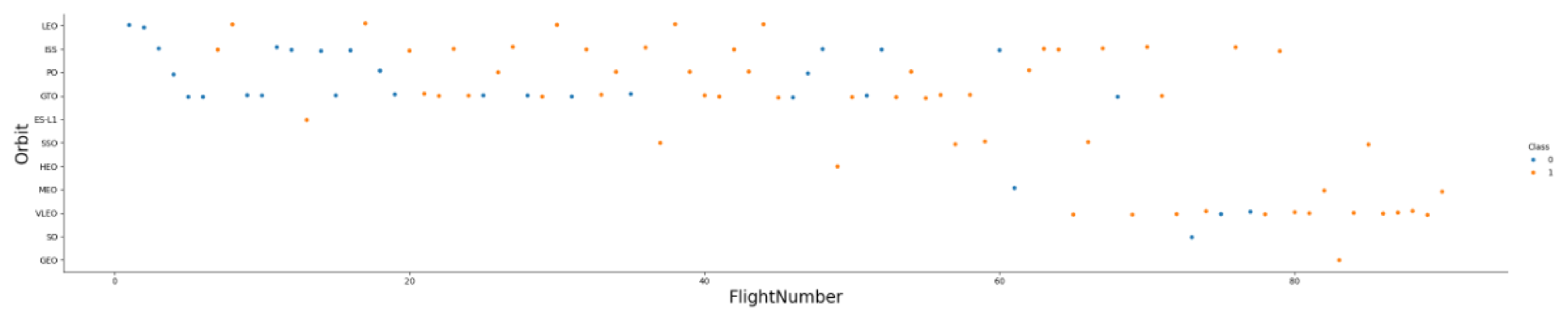
- Show a bar chart for the success rate of each orbit type



The scatter plot show that orbit type GTO has the lower rate of launch success but ES-L1, GEO, HEO and SSO has the highest success rate.

Flight Number vs. Orbit Type

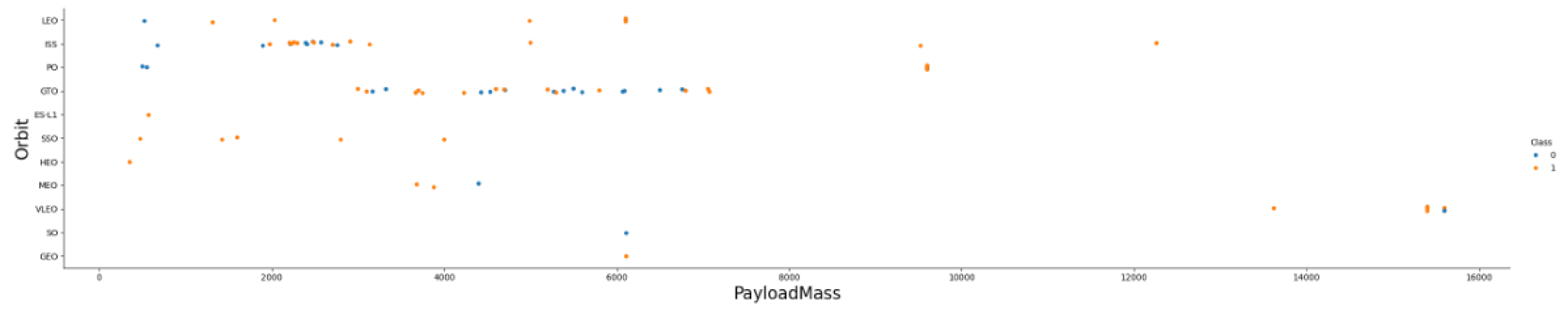
- Show a scatter point of Flight number vs. Orbit type



The scatter plot show that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

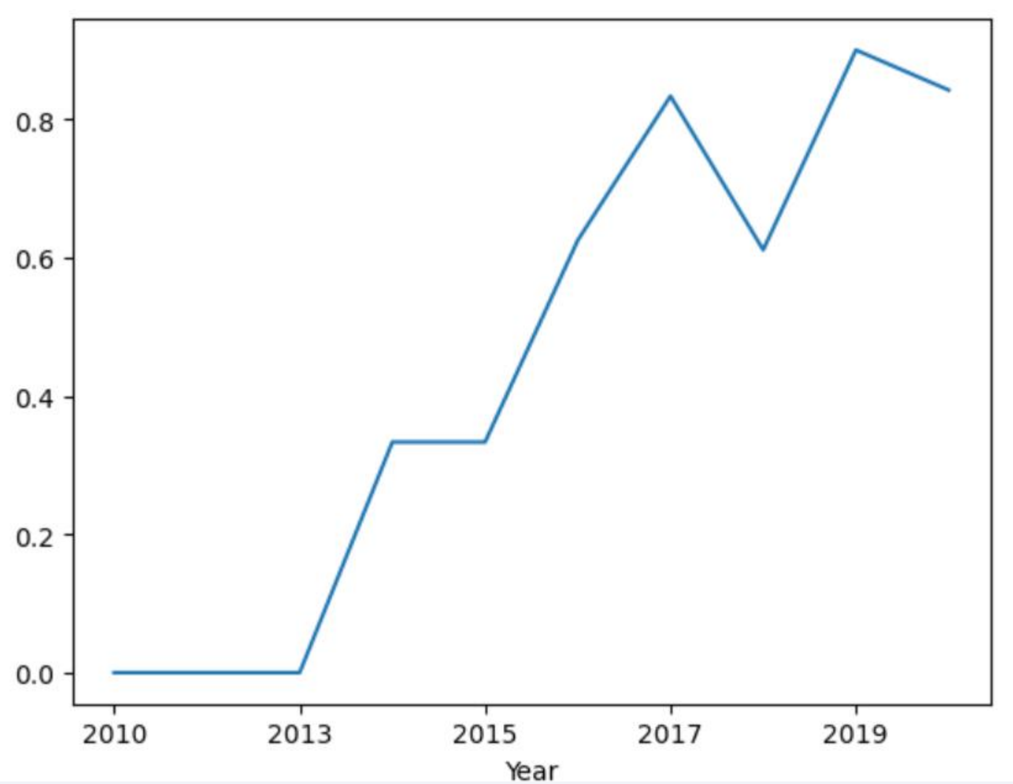


With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend

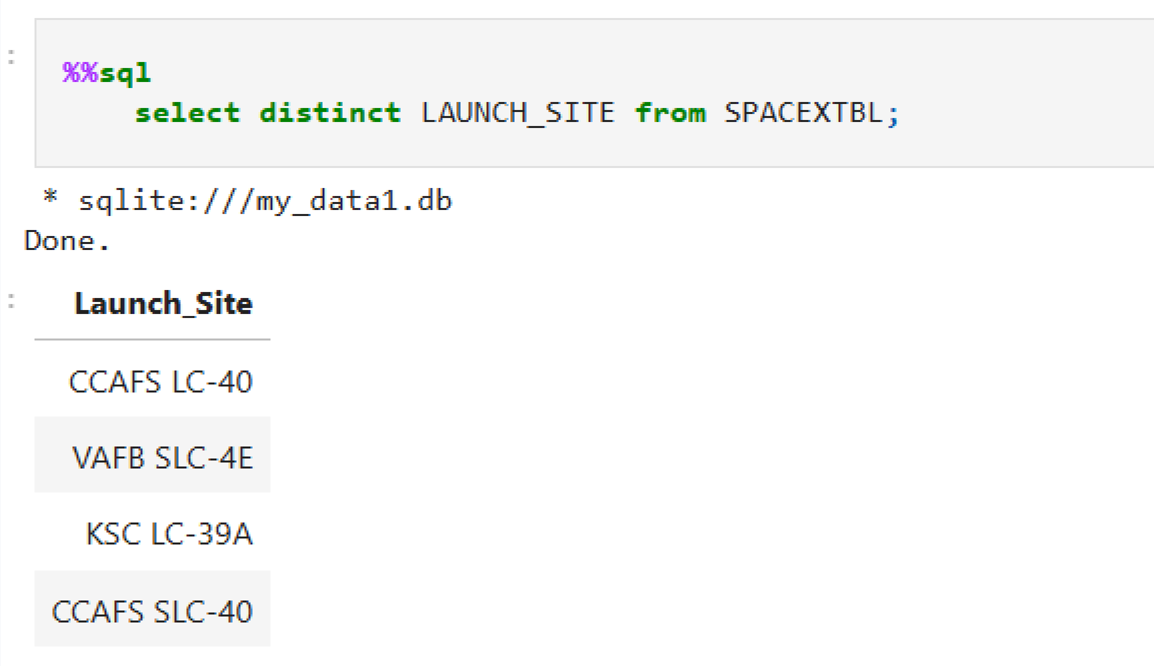
- Show a line chart of yearly average success rate



The sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

- Names of the unique launch sites



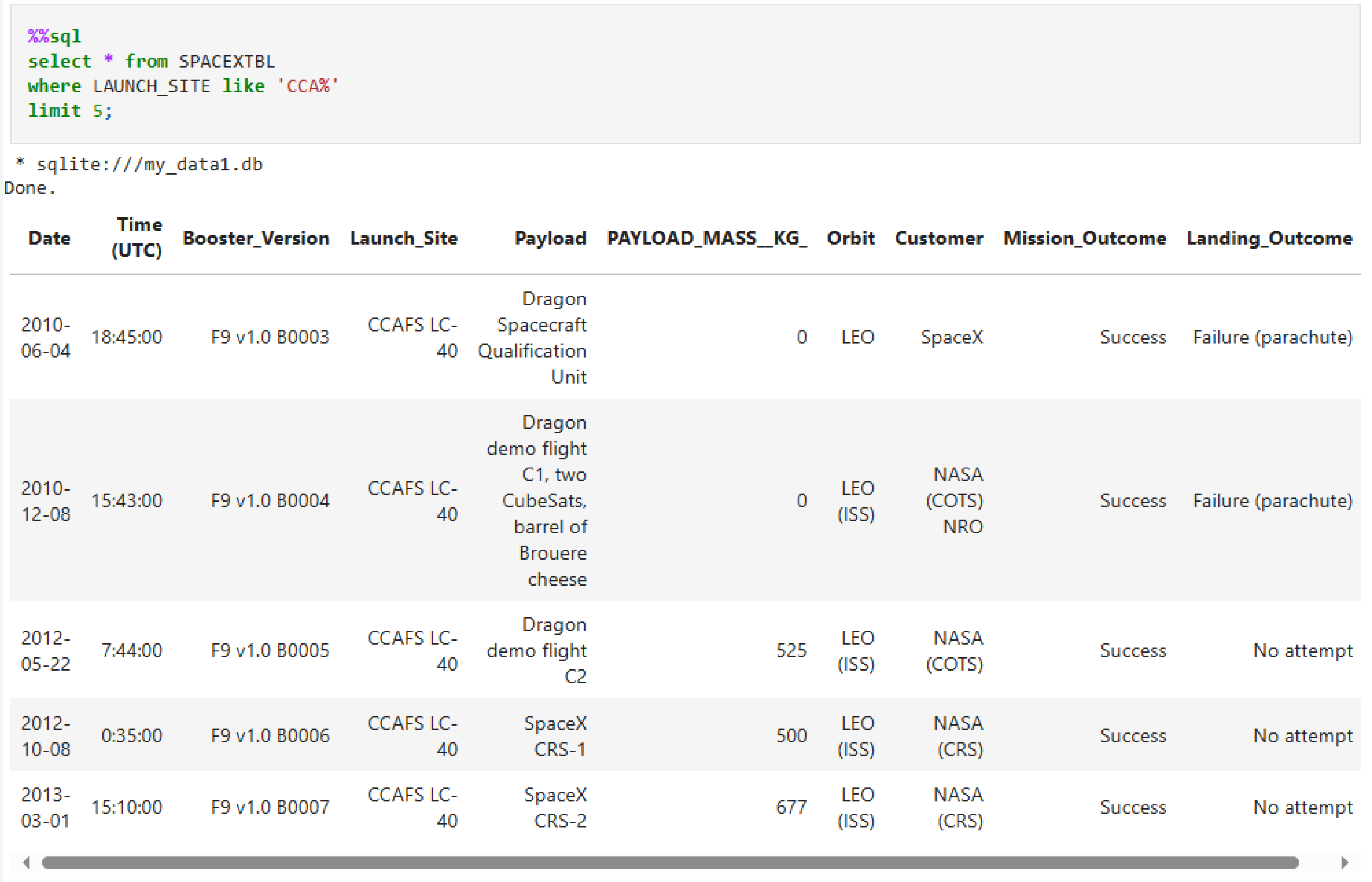
```
%%sql
    select distinct LAUNCH_SITE from SPACEXTBL;
```

```
 * 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'

- 5 records where launch sites begin with `CCA`



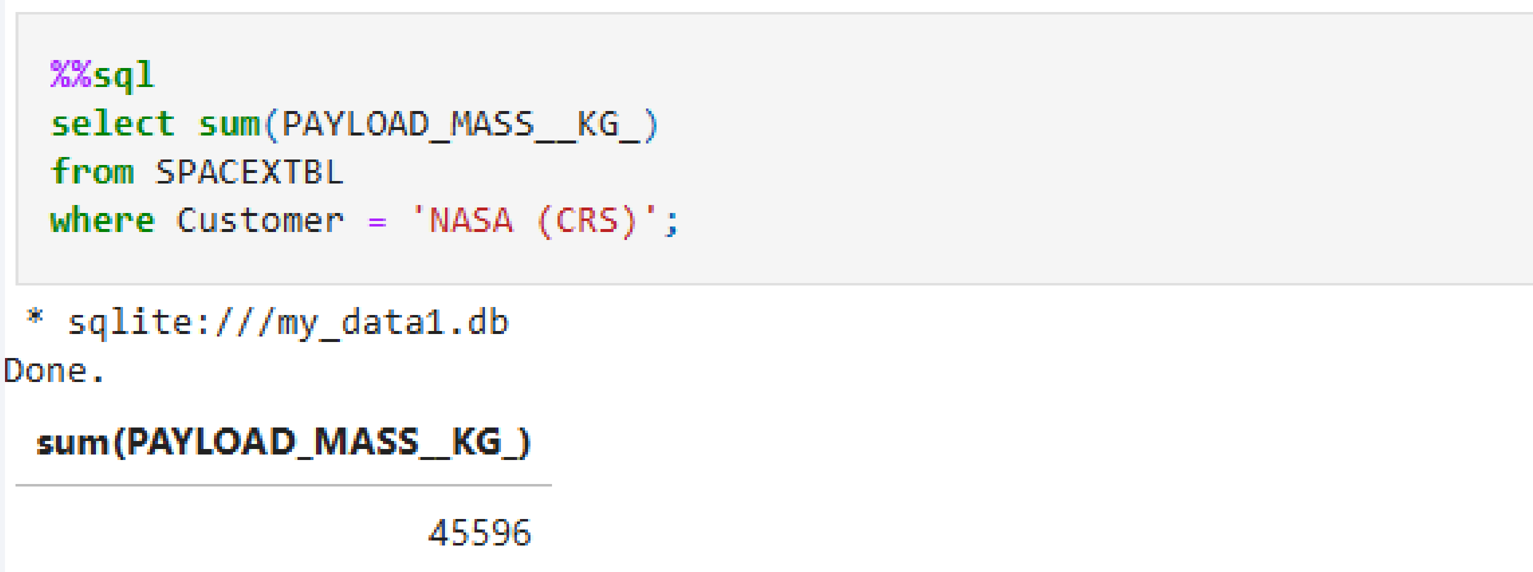
```
%%sql
select * from SPACEXTBL
where LAUNCH_SITE like 'CCA%'
limit 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The total payload carried by boosters from NASA



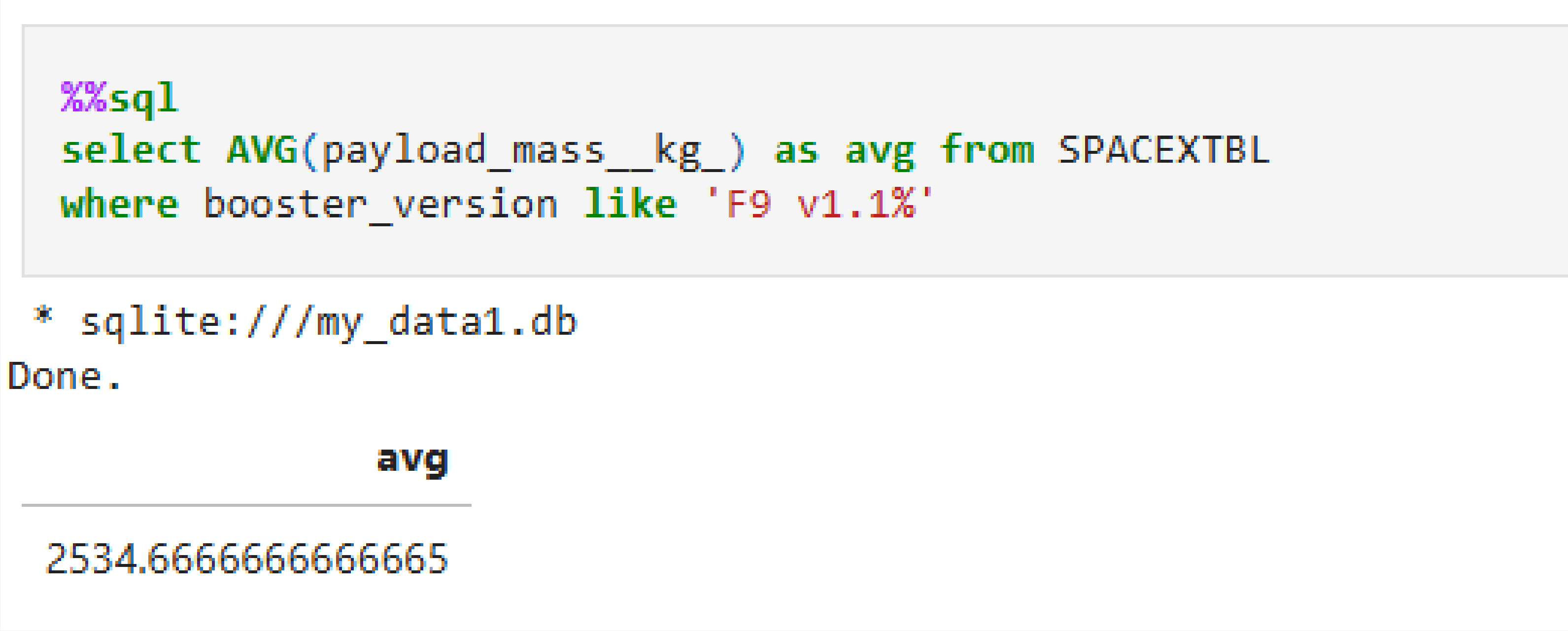
```
%%sql
select sum(PAYLOAD_MASS__KG_)
from SPACEXTBL
where Customer = 'NASA (CRS)';
```

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

```
sum(PAYLOAD_MASS__KG_)
                 45596
```

Average Payload Mass by F9 v1.1

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



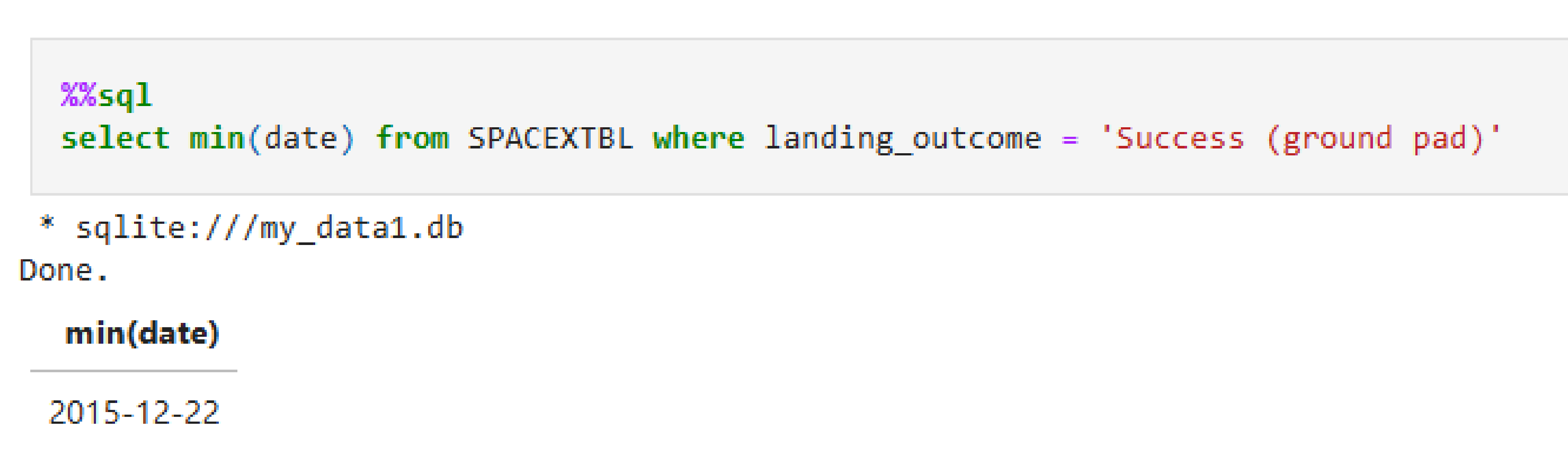
```
%%sql
select AVG(payload_mass__kg_) as avg from SPACEXTBL
where booster_version like 'F9 v1.1%'
```

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

```
                avg
 2534.6666666666665
```

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad



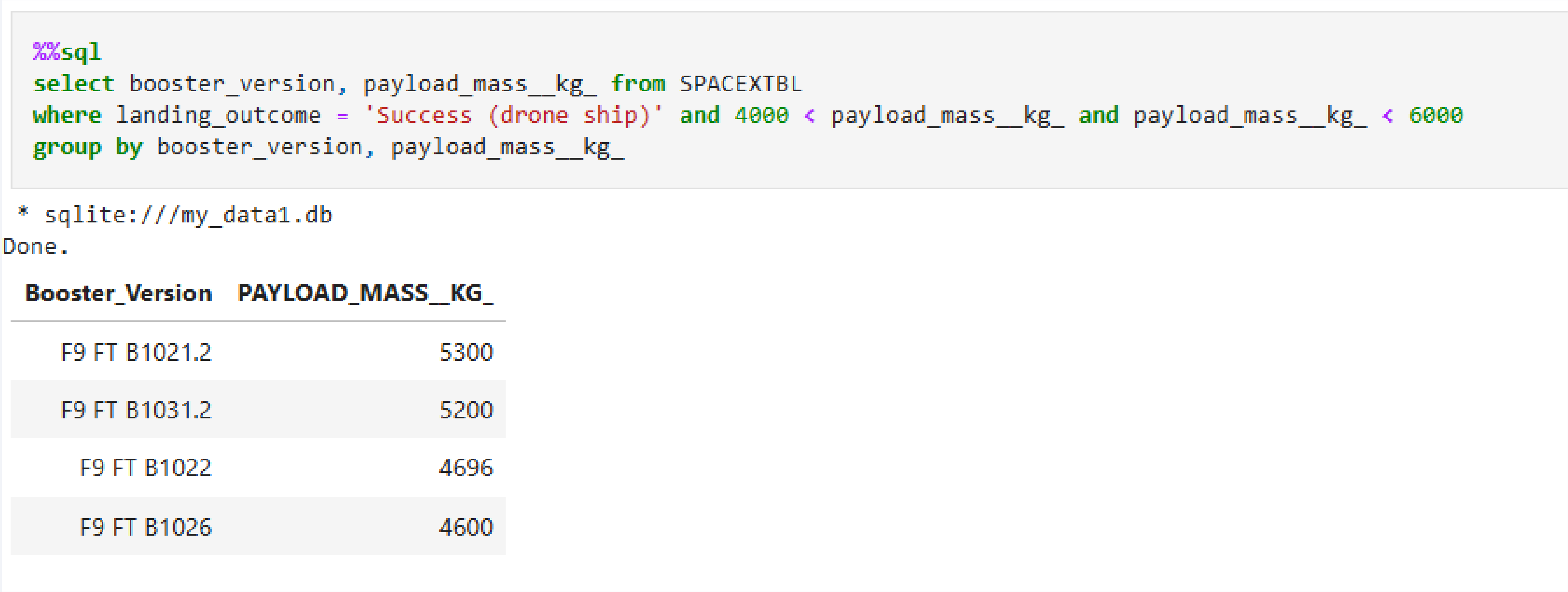
```
%%sql
select min(date) from SPACEXTBL where landing_outcome = 'Success (ground pad)'
```

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

min(date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



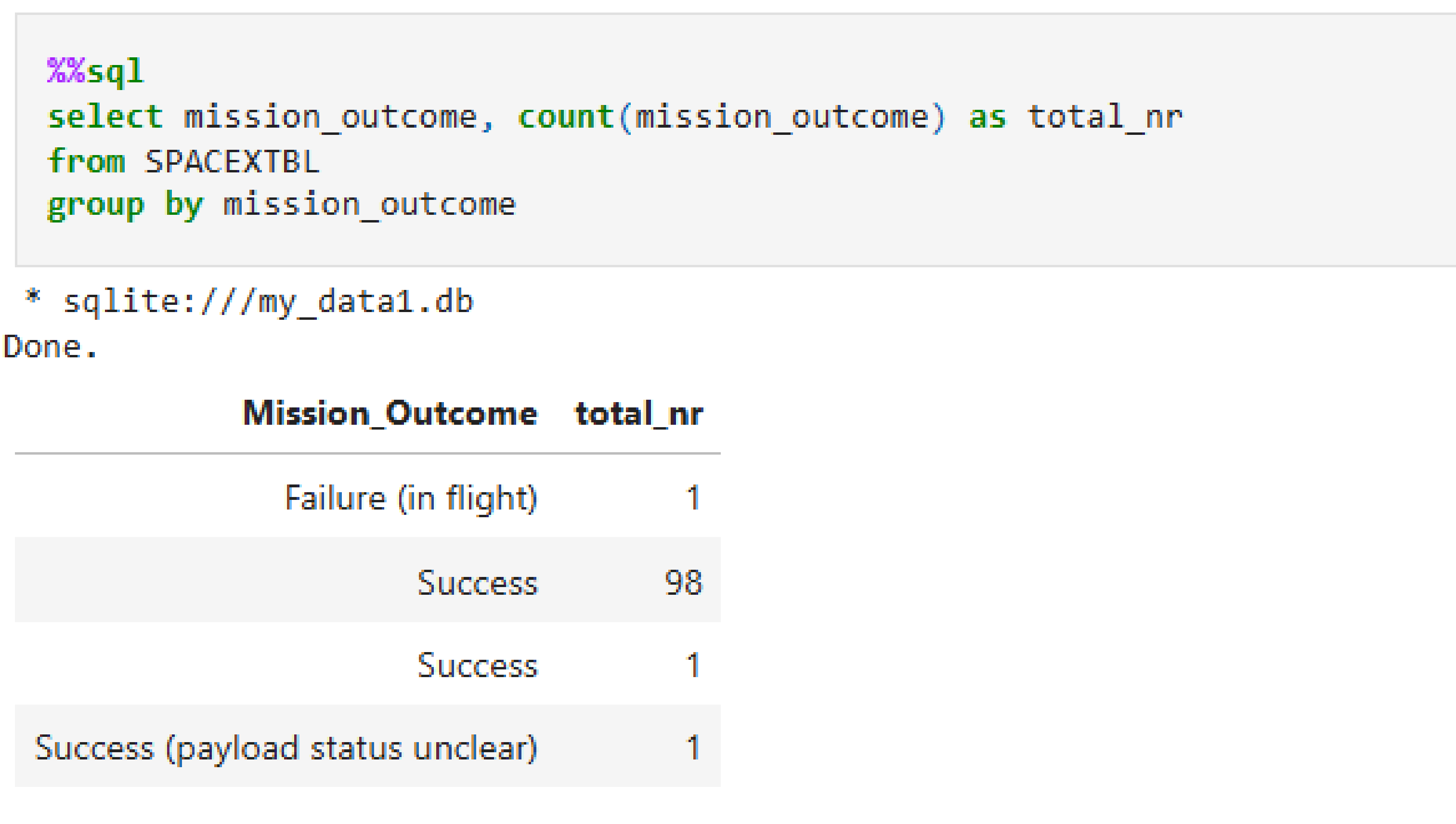
```
%%sql
select booster_version, payload_mass__kg_ from SPACEXTBL
where landing_outcome = 'Success (drone ship)' and 4000 < payload_mass__kg_ and payload_mass__kg_ < 6000
group by booster_version, payload_mass__kg_
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1021.2	5300
F9 FT B1031.2	5200
F9 FT B1022	4696
F9 FT B1026	4600

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes



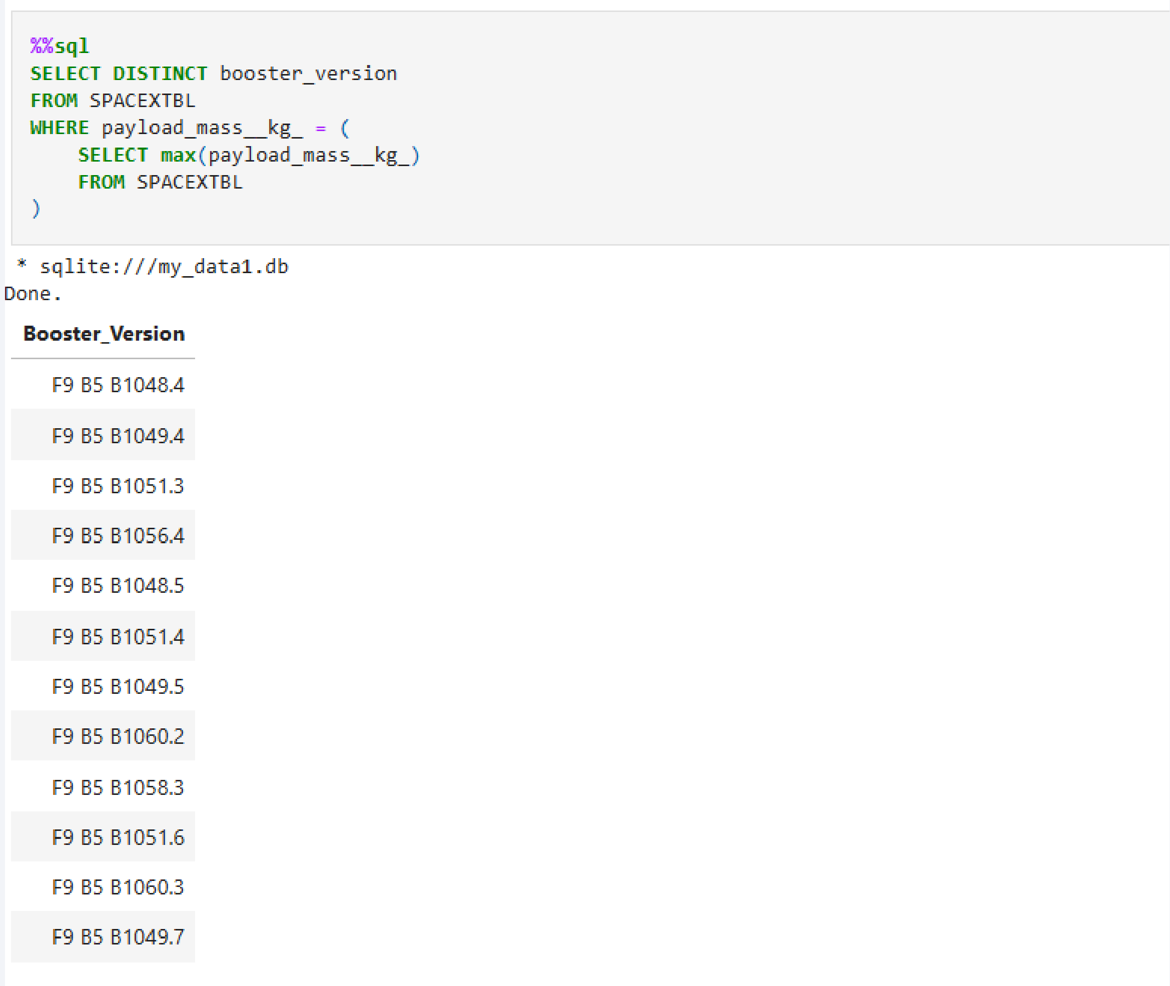
```
%%sql
select mission_outcome, count(mission_outcome) as total_nr
from SPACEXTBL
group by mission_outcome
```

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

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass



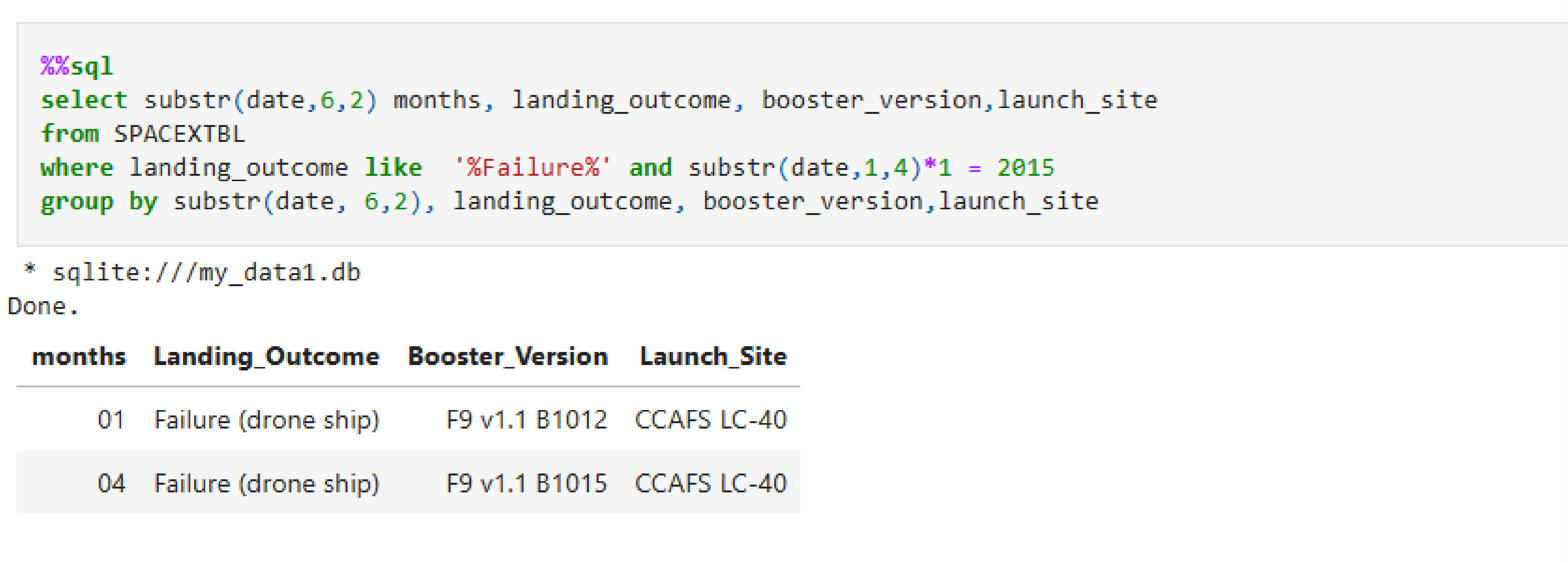
```
%%sql
SELECT DISTINCT booster_version
FROM SPACEXTBL
WHERE payload_mass__kg_ = (
    SELECT max(payload_mass__kg_)
    FROM SPACEXTBL
)
```

```
 * 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

- Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015



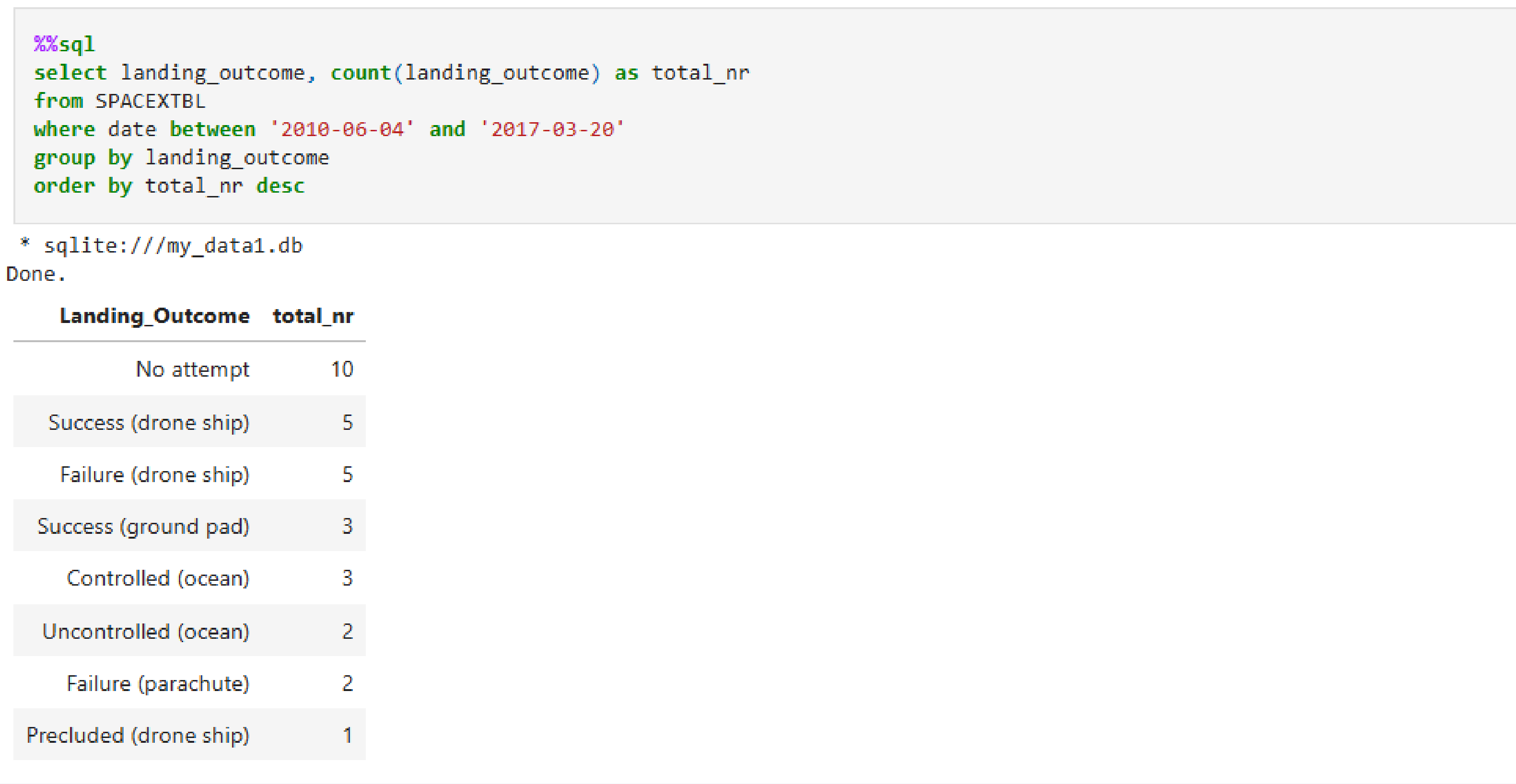
```
%%sql
select substr(date,6,2) months, landing_outcome, booster_version,launch_site
from SPACEXTBL
where landing_outcome like  '%Failure%' and substr(date,1,4)*1 = 2015
group by substr(date, 6,2), landing_outcome, booster_version,launch_site
```

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

months	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

- 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_nr
from SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by landing_outcome
order by total_nr desc
```

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

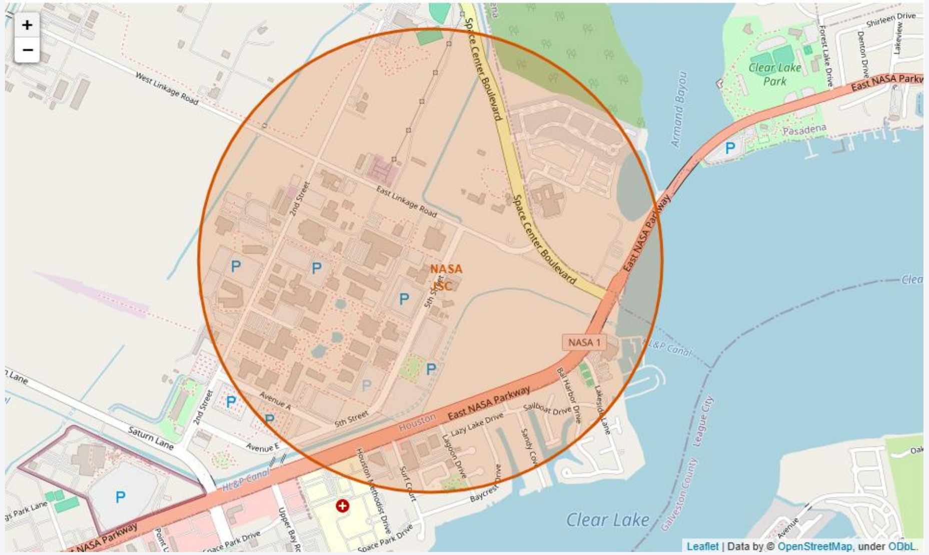
Landing_Outcome	total_nr
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

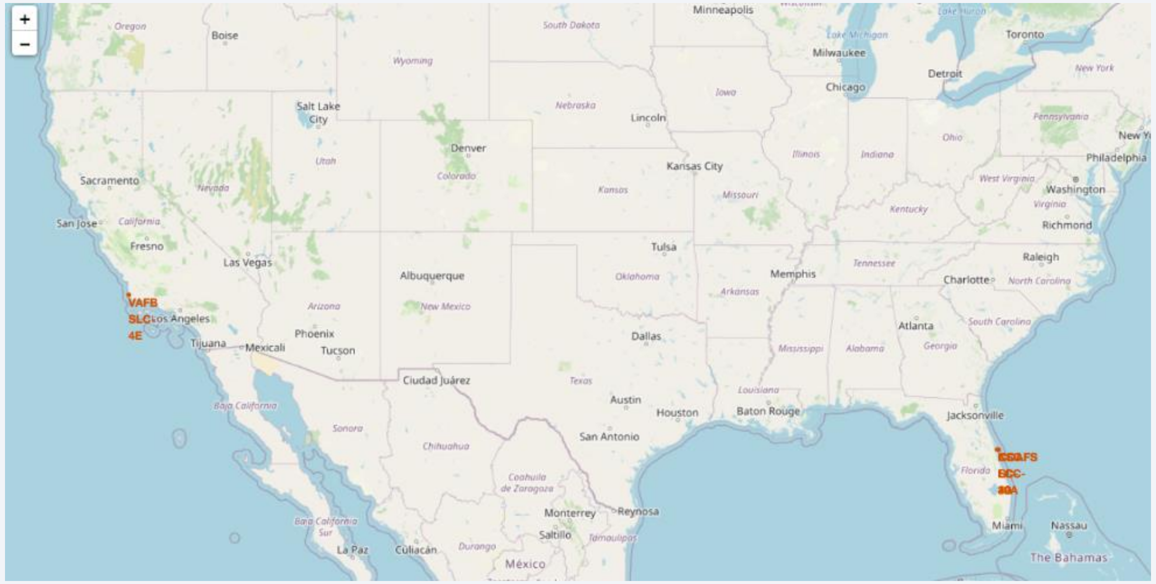
Launch Sites Proximities Analysis

NASA Johnson Space Center

- NASA Johnson Space Center

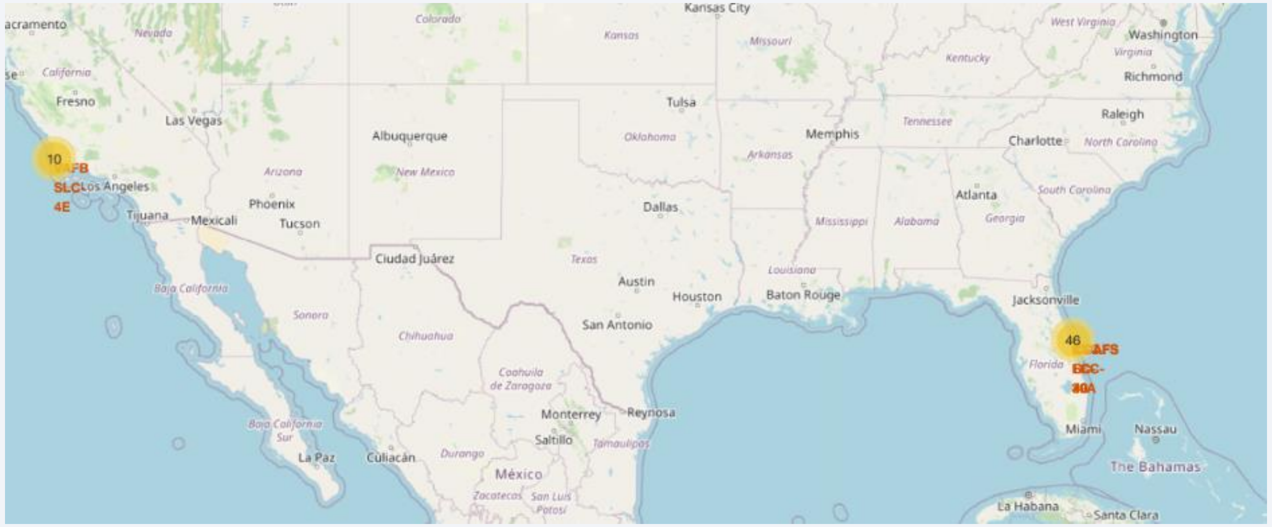


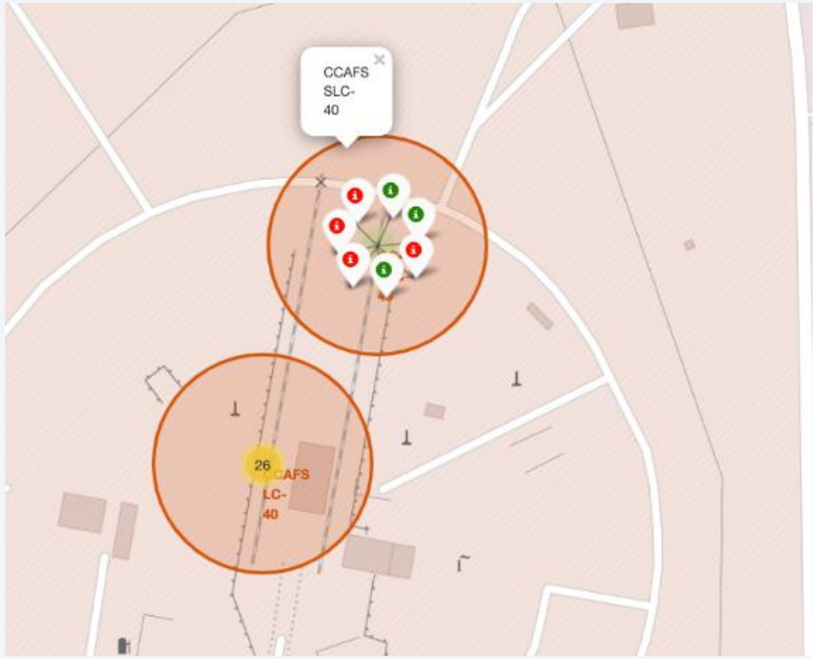
All launch sites' location markers on a global map



Sites locations

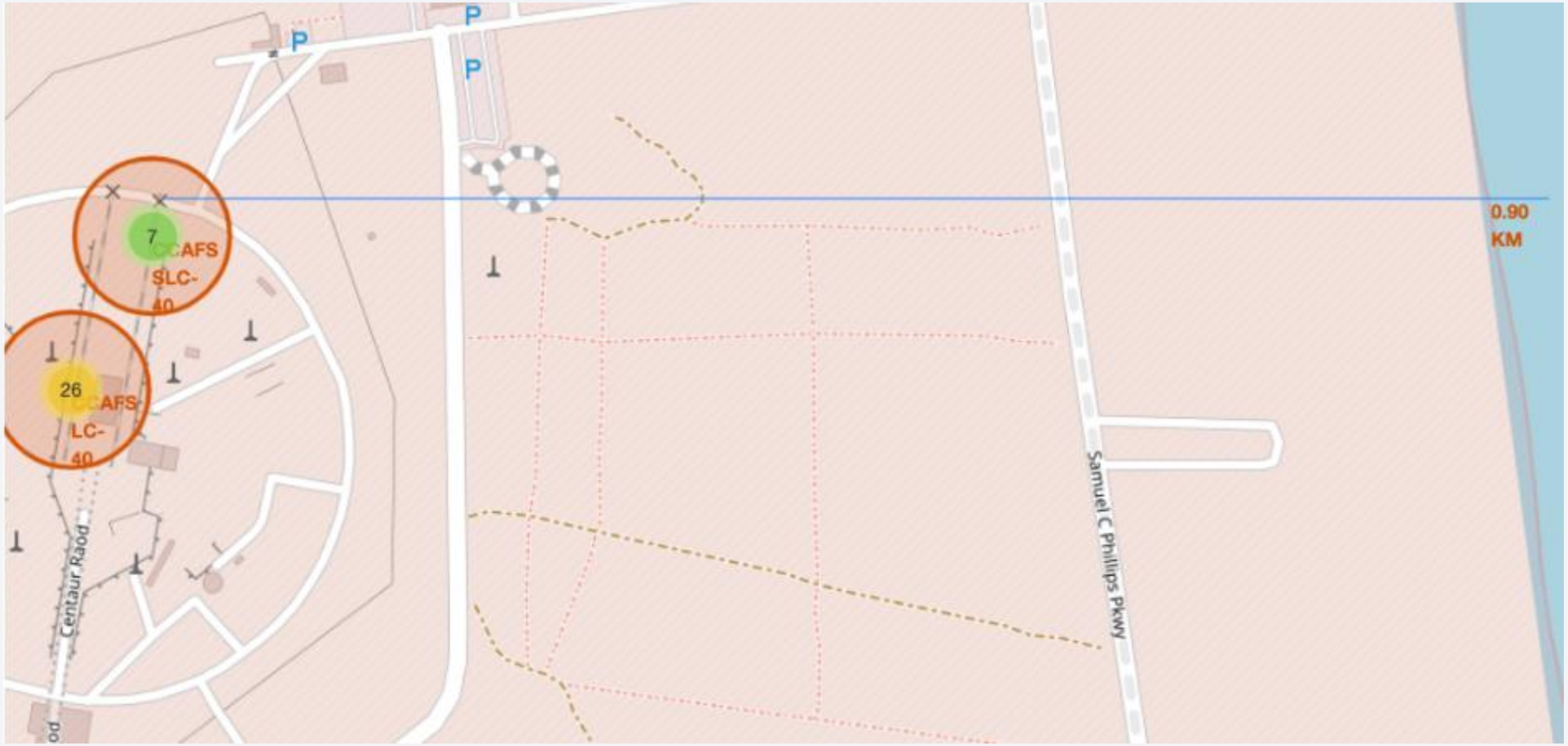
- Sites locations

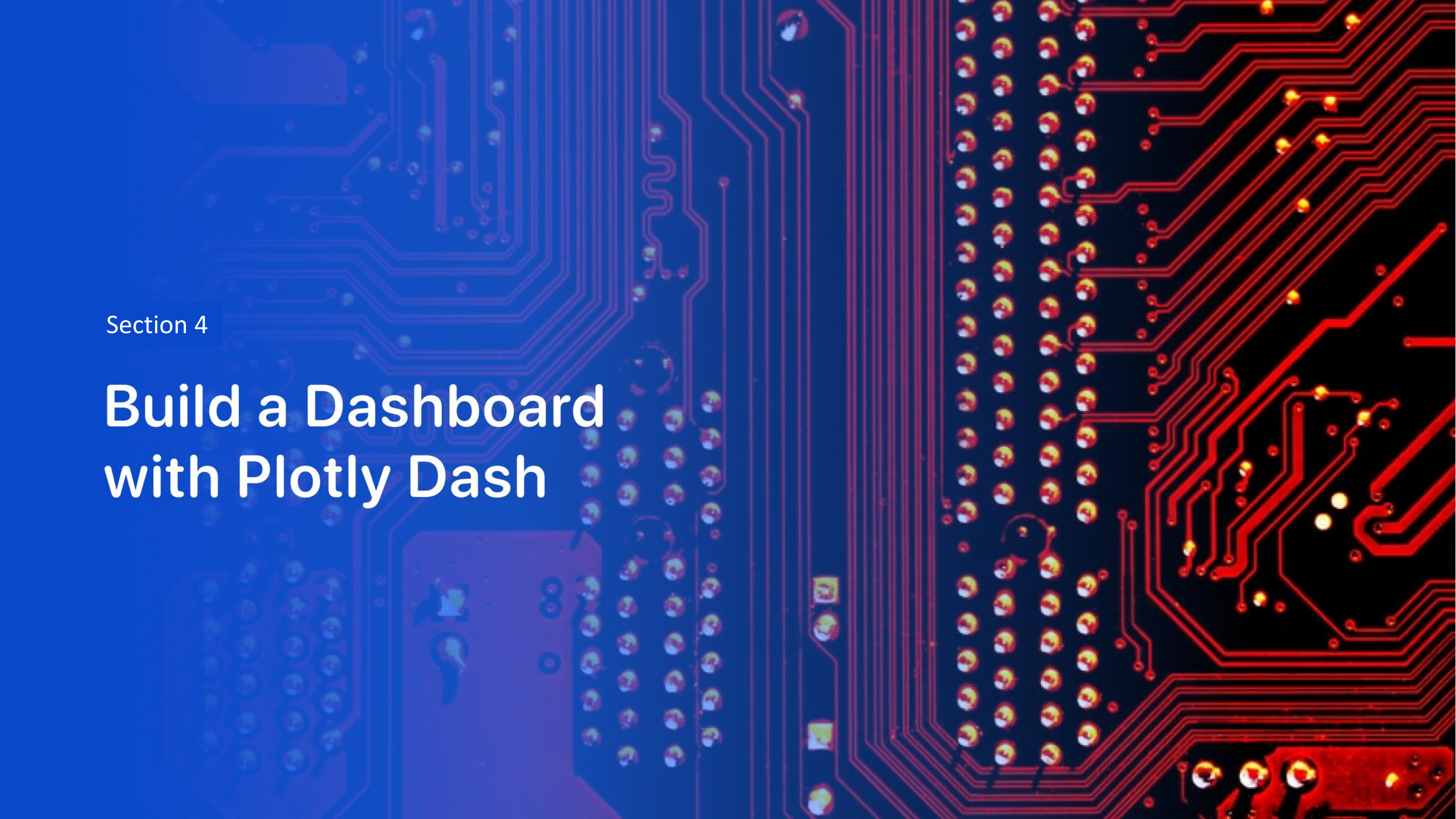
- 



Launch site to its proximities such as railway, highway, coastline

- Launch site to its proximities such as railway, highway, coastline, with distance calculated



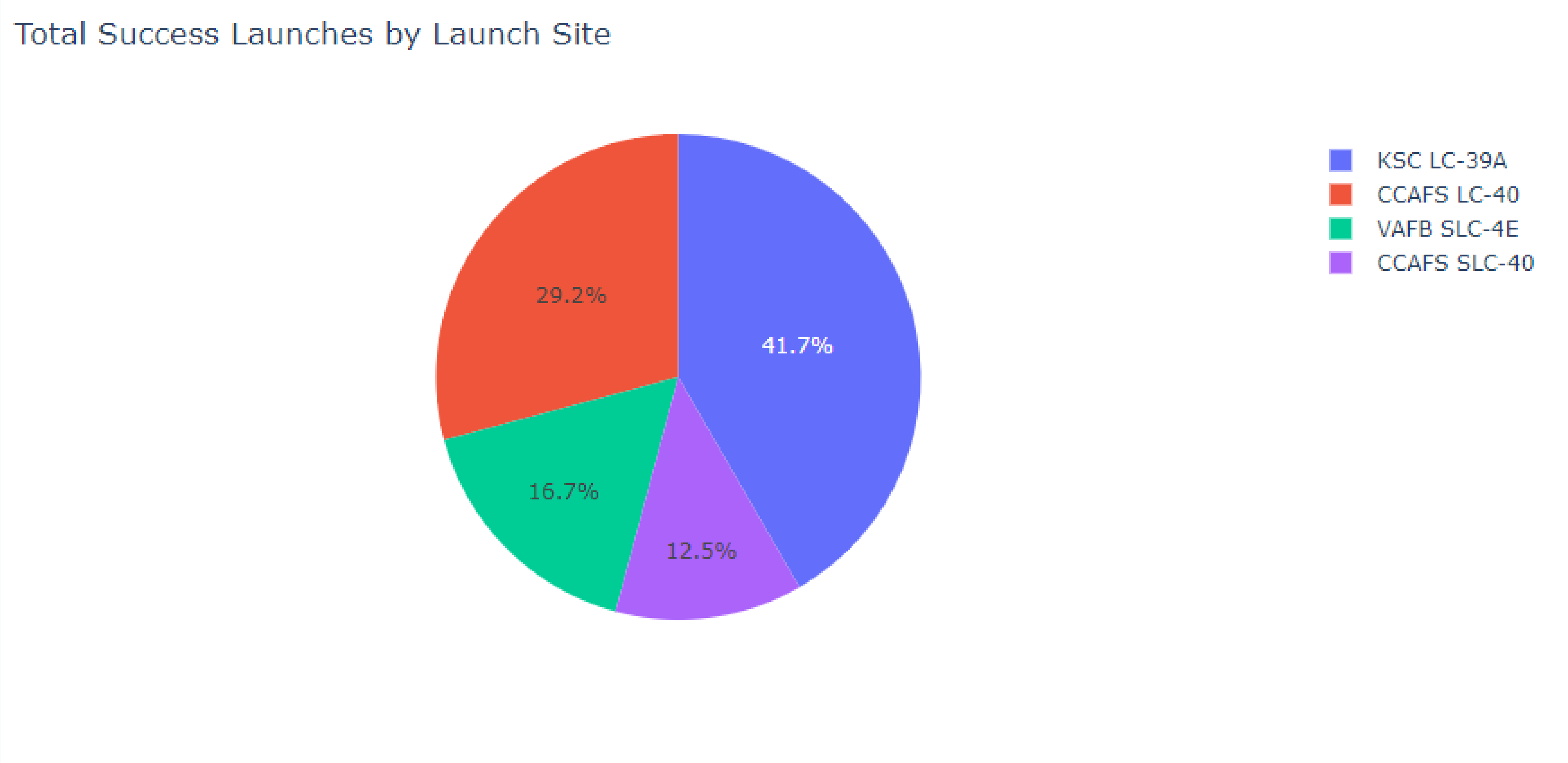


Section 4

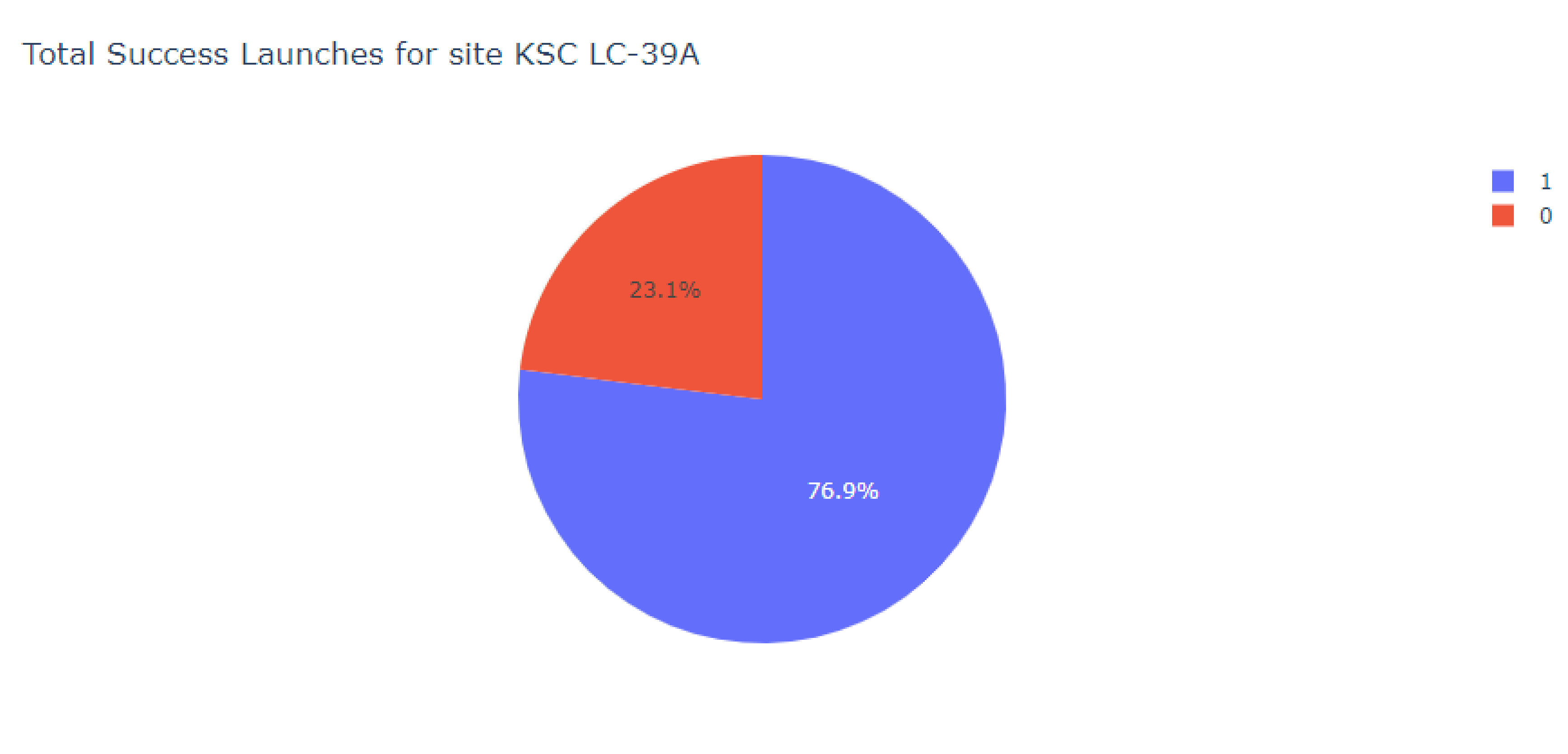
Build a Dashboard with Plotly Dash

Launch success count for all sites

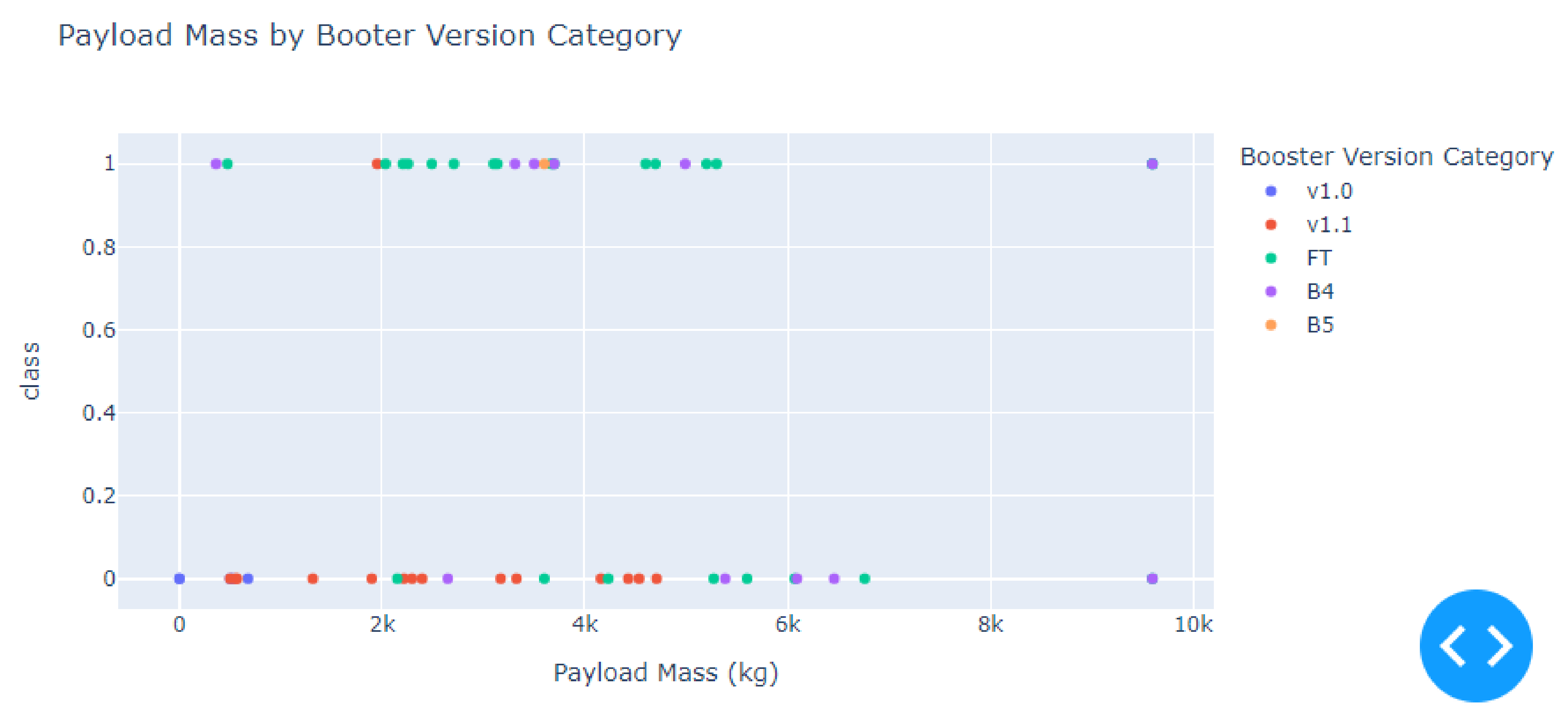
- KSC LC-39A has the highest launch success percentage of all sites



KSC LC-39A



Payload vs. Launch Outcome scatter plot for all sites



Section 5

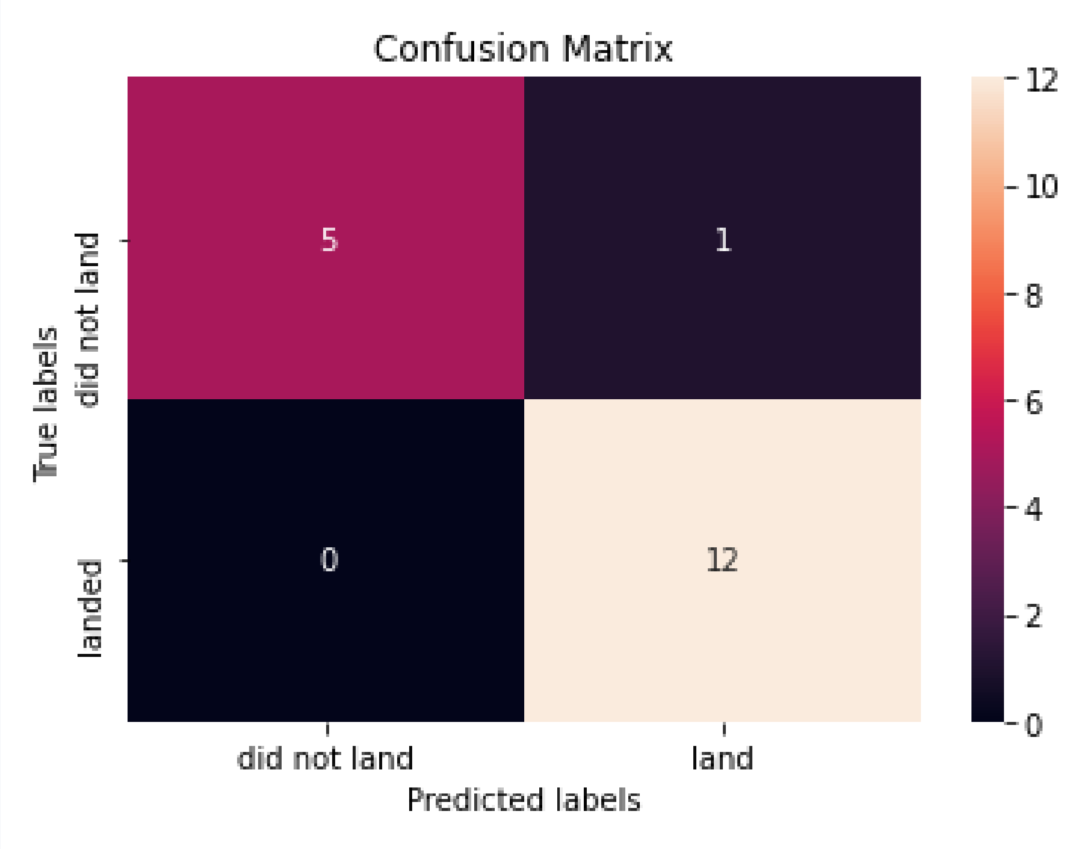
Predictive Analysis (Classification)

Model accuracy for all built classification models

- The highest classification accuracy was **DecisionTree**

Confusion Matrix

- Confusion matrix of the best performing model



Conclusions

- The scatter plot show that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.
- Average Payload Mass by F9 v1.1 was 2,534.66 kg
- The dates of the first successful landing outcome on ground pad was 2015-12-22.
- KSC LC-39A has the highest launch success percentage of all sites
- The highest classification accuracy was *DecisionTree*

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!