IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

D.Laurens
18/01/2025

Outline

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

Executive Summary

The primary objective of this project is to identify the key factors contributing to a successful launch. Our approach involved several critical steps:

1. **Data Collection:** Gathering relevant data.
2. **Exploratory Data Analysis (EDA):** Conducting a thorough analysis to uncover patterns and insights.
3. **Model Building:** Developing predictive models to forecast the outcomes of future launches.

These topics will be discussed in detail during the presentation.

Key Takeaways:

Achieved an accuracy rate of 83.3% in predicting launch outcomes on the test set.

Introduction

SpaceX, a pioneering force in the space industry, is dedicated to making space travel accessible and affordable for all. SpaceX is known for activities such as **Delivering spacecraft** to the International Space Station / **Deploying a satellite constellation** to provide global internet access **/ Executing manned space missions.**

SpaceX's cost-effective rocket launches, priced at $62 million per launch, are made possible by the innovative reuse of the first stage of its Falcon 9 rocket. In contrast, other providers, unable to reuse the first stage, face costs exceeding $165 million per launch.

By predicting the likelihood of a successful first-stage landing, we can estimate the launch cost. This can be accomplished using publicly available data and machine learning models to forecast whether SpaceX or a competitor can reuse the first stage.

Exploration Focus:

- Analyzing how payload mass, launch site, number of flights, and orbits influence first-stage landing success.
- Tracking the rate of successful landings over time.
- Determining the most effective predictive model for successful landings.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data have been collected using SpaceX REST API and WebScrapping.
- Perform data wrangling
  - Scope the data, handling missing values and transforming categorical features to numerical (One Hot Encoding) in order to prepare for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

In this section, we will delve into the data collection techniques employed in the project.

The primary methods include:

- Utilizing the SpaceX REST API.

- Extracting data from the Wikipedia Falcon 9 page.

The objective of the data collection step is to:

- Acquire relevant and accurate data.

- Ensure data quality.

- Build a comprehensive dataset to facilitate future analysis.

Data Collection – SpaceX API

- Request data from SpaceX API
- Decode response and Transform to a dataframe (use of .json() and .json_normalize())
- Request information about the launches from SpaceX API using custom functions to decode IDs and store in a dictionary
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file

Link to GitHub hosted notebook

Data Collection - Scraping

- Request data (Falcon 9 launch data) from Wikipedia using requests librairy
- Create BeautifulSoup object from requests.get() method
- Populate a dictionary using HTML Tables (Header (th) and data (td))
- Create dataframe from the dictionary
- Export data to csv file

Link to GitHub hosted Notebook

Data Wrangling

Steps :

1. Data Loading : import and read of the csv file
2. Data Analysis : Missing data proportion, type of features, exploration of Launch Sites and Orbits to better understand data
3. Mapping of Outcome Feature : Create a numerical feature from a categorical one (if bad outcome 0 else 1)

Link to GitHub hosted Notebook

EDA with SQL

EDA with SQL has been divided into 2 types of queries :

1. Display information of the table :

   Unique launch sites / 5 records where launch site begins with ‘CCA’ / Total payload mass carried by boosters launched by NASA (CRS) / Average payload mass carried by booster version F9 v1.1.

2. Explore Data by creating new field :

   Date of first successful landing on ground pad / Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000 / Total number of successful and failed missions / Names of booster versions which have carried the max payload / Failed landing outcomes on drone ship, their booster version and launch site for the months in the year 2015 / Count of landing outcomes between 2010-06-04 and 2017-03-20

Link to GitHub hosted Notebook

EDA with Data Visualization

During EDA we plotted different chart to discover bivariate relationship.

1. FlightNumber vs. PayloadMass (scatter plot for relationship)
2. Flight Number vs. Launch Site (scatter plot for relationship)
3. Payload Mass vs. Launch Site (scatter plot for relationship)
4. SuccessRate of each Orbit Type (bar plot for relationship)
5. FlightNumber vs. Orbit type (scatter plot for relationship)
6. Payload Mass vs. Orbit type (scatter plot for relationship)
7. Launch Success Trend (Year) (line plot for trend)

Understanding relationship between features is key for developing a machine learning model.

Link to GitHub hosted Notebook

Build an Interactive Map with Folium

First of all, we marked all launch sites on a map and added a popup with some info in it.

Then we marked the success/failed launches for each site on the map (green if it was a success and red a failure) which gives us a good overview of high success site.

Finally, we calculated the distances between a launch site to its proximities such as coastline, railway, highway, and city

Link to GitHub hosted Notebook

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

Allow user to navigate through launch sites to scope their analysis

Slider of Payload Mass Range

Enable users to choose a range for the payload mass.

Pie Chart Showing Successful Launches

Enable users to view the outcome of launches as a percentage of total number of launches.

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

Show the relationship between these features to users

Link to GitHub hosted py file

Predictive Analysis (Classification)

- Create Dependant variable from Class column (using NumPy array)
- Standardize the data with StandardScaler. Fit and transform the data then split the data using train_test_split, we will use train data to train our models.
- For each algorithms LogisticRegression(), SVC(), DecisionTreeClassifier(), KNeighborsClassifier() :
  - Initialise the model and create a parameter dictionnary
  - Create a GridSearchCV object with 10 folds and parameters created before
  - Fit GridSearchCV object on train data
  - Calculate accuracy on the test data using .score() and assess the confusion matrix
- Identify the best model based on accuracy

Link to GitHub hosted Notebook

Results

Exploratory Data Analysis

- Launch success has improved over time with a peak at 0.9
- KSC LC-39A has the highest success rate among landing sites 77.2%
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are close to the equator and the coast
- Launch sites are distant enough to public installation (City, Railway etc) in case of a bad outcome, while still close enough to supply launch activities

Predictive Analytics

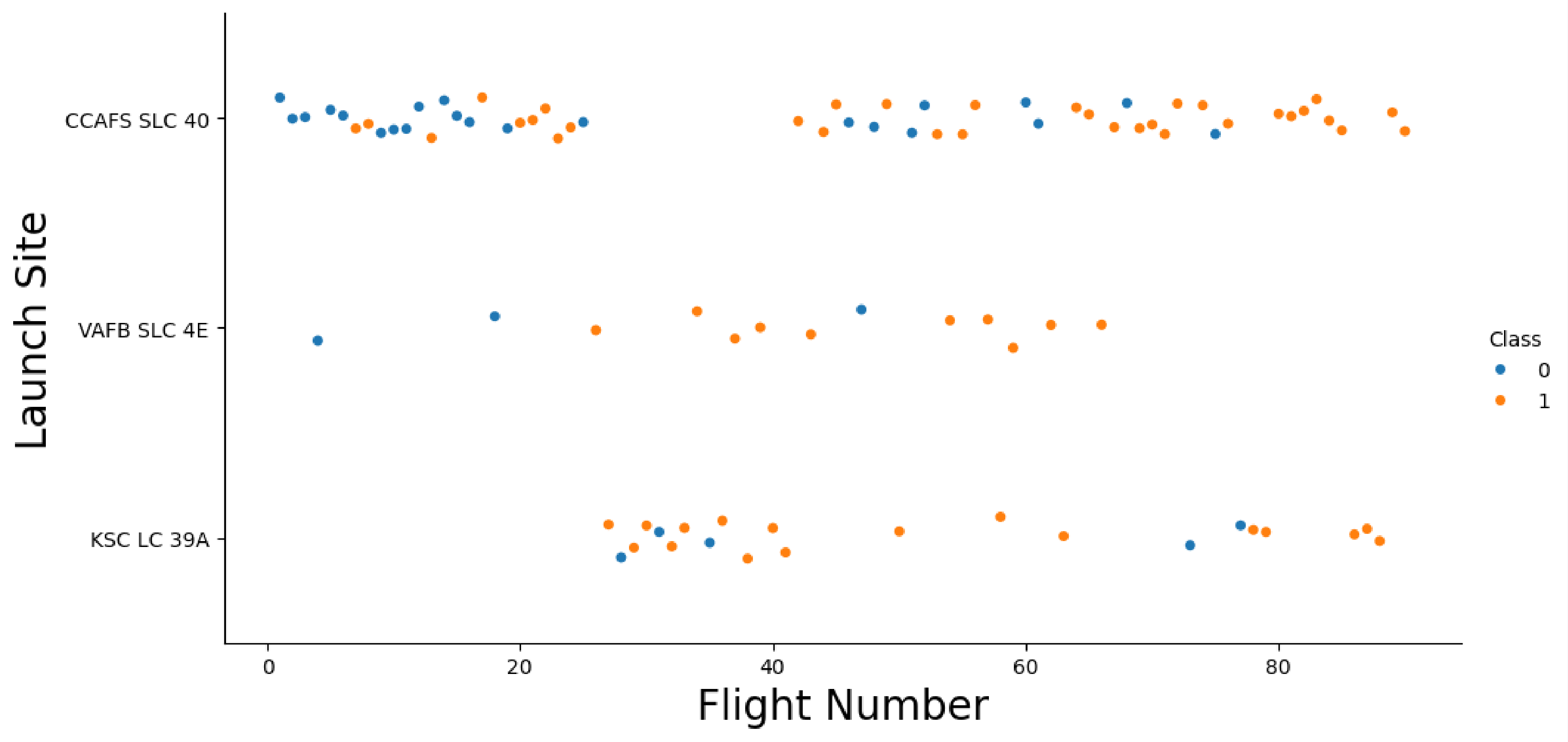
- Logistic Regression / Support Vector Classifier and KNN are the best model on test data

Section 2

Insights drawn from EDA

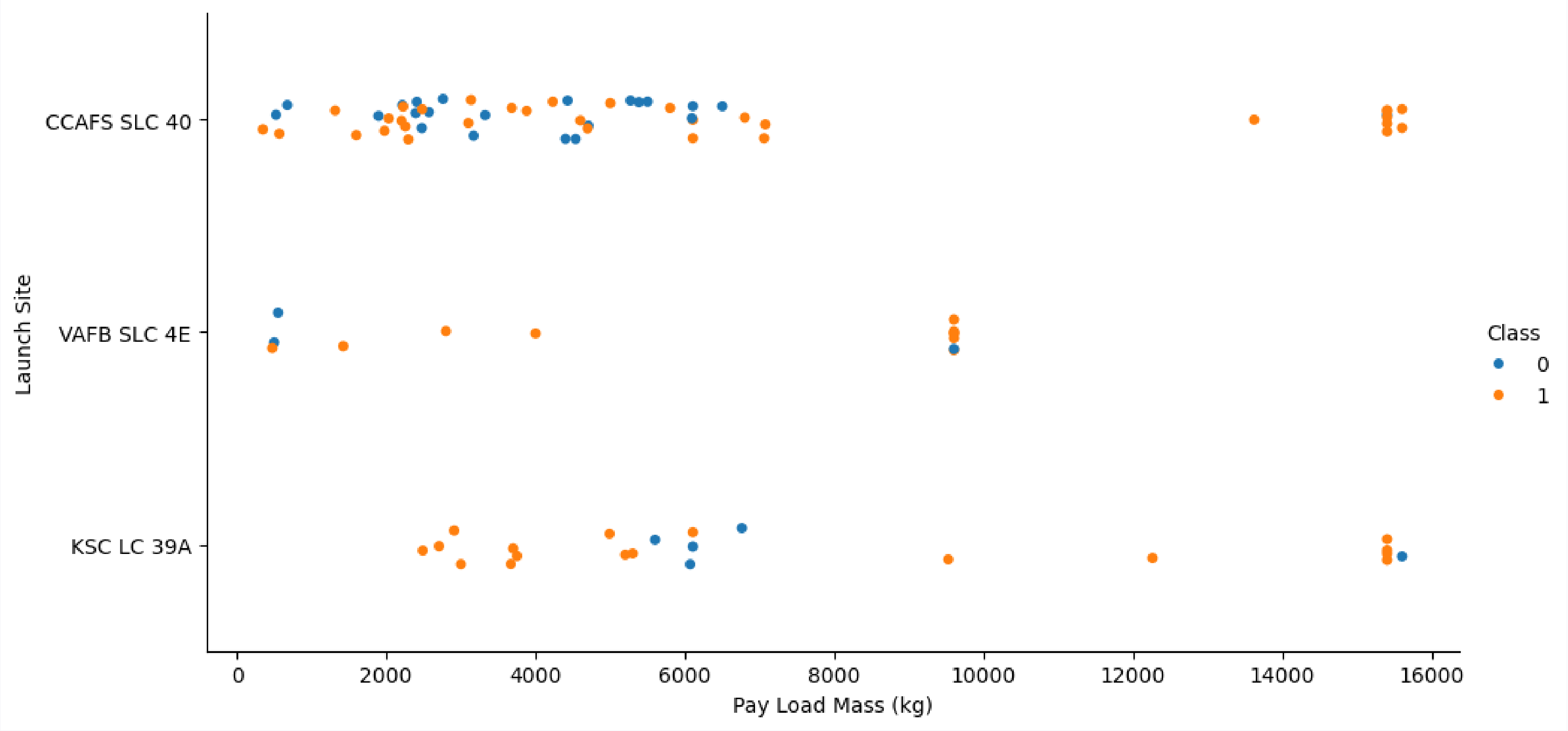
Flight Number vs. Launch Site

- Blue dot are failure and Orange are success
- CCAF5 have half of success while VAFB and KSC have higher success rate (around 70%)



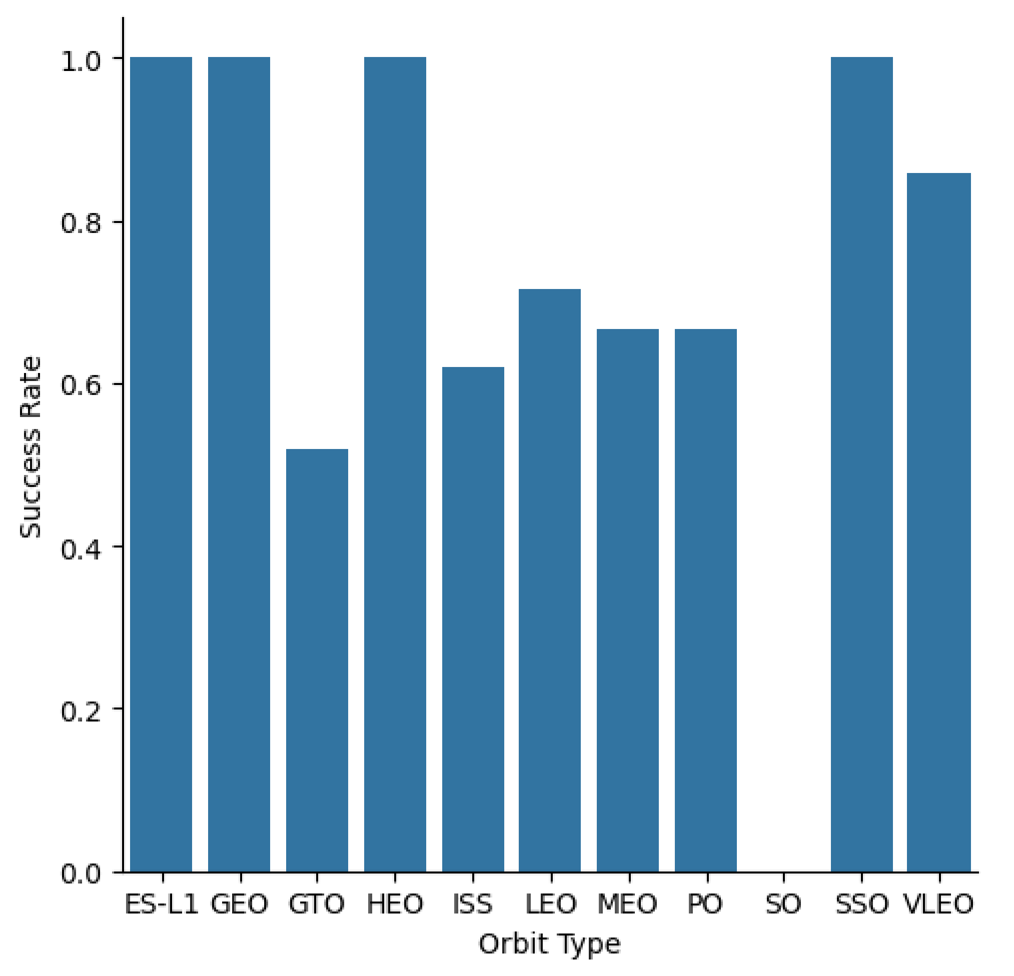
Payload vs. Launch Site

- The higher the payload mass is, higher is the success rate
- Almost all launch with a payload greater than 7000 kg were successful
- KSC LC has a 100% success rate for launches less than 5500 kg
- VAFB has not launched anything greater than 10 000 kg



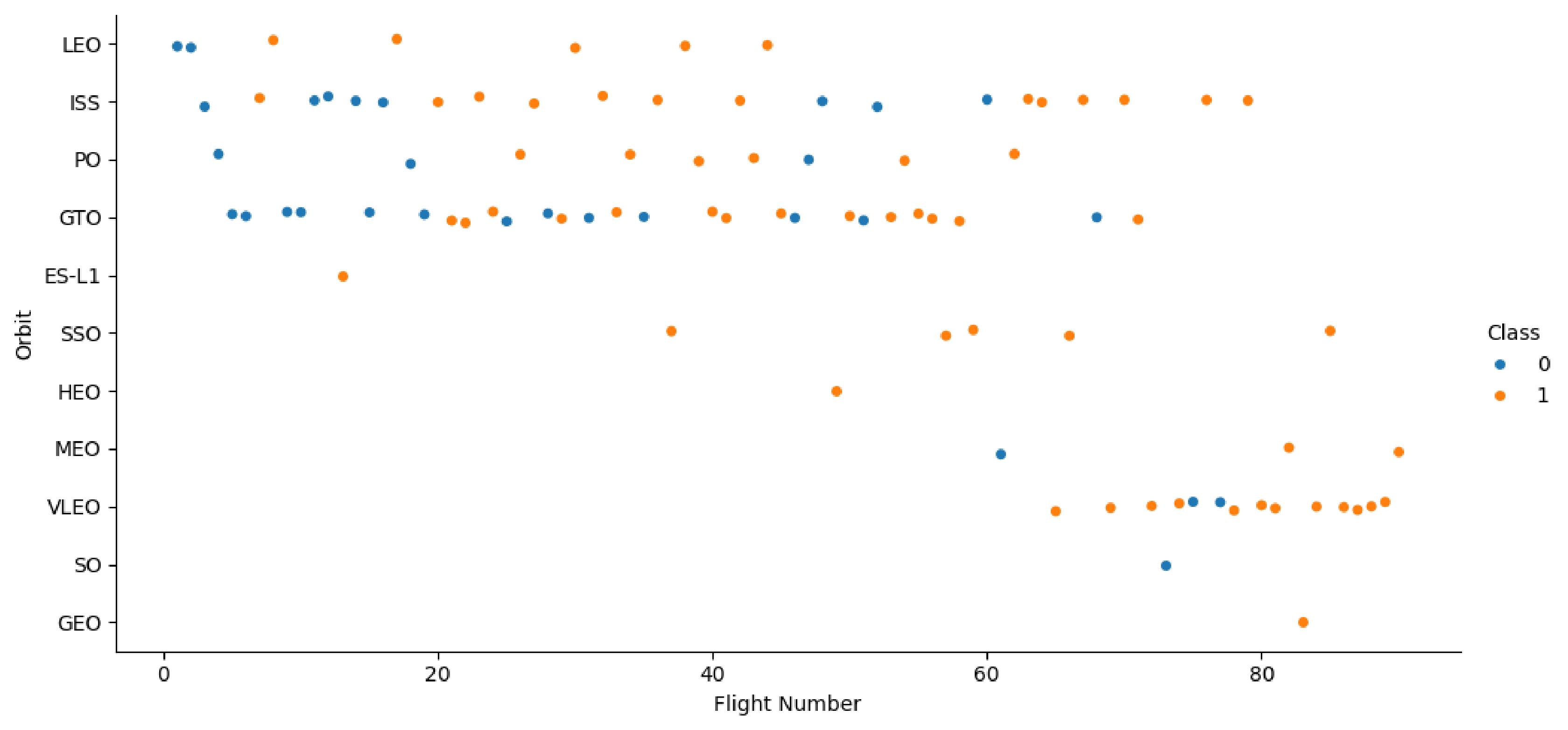
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO has perfect success rate
- SO has 0% success rate
- GTO, ISS, LEO, MEO, PO are in the range 50-70% of success



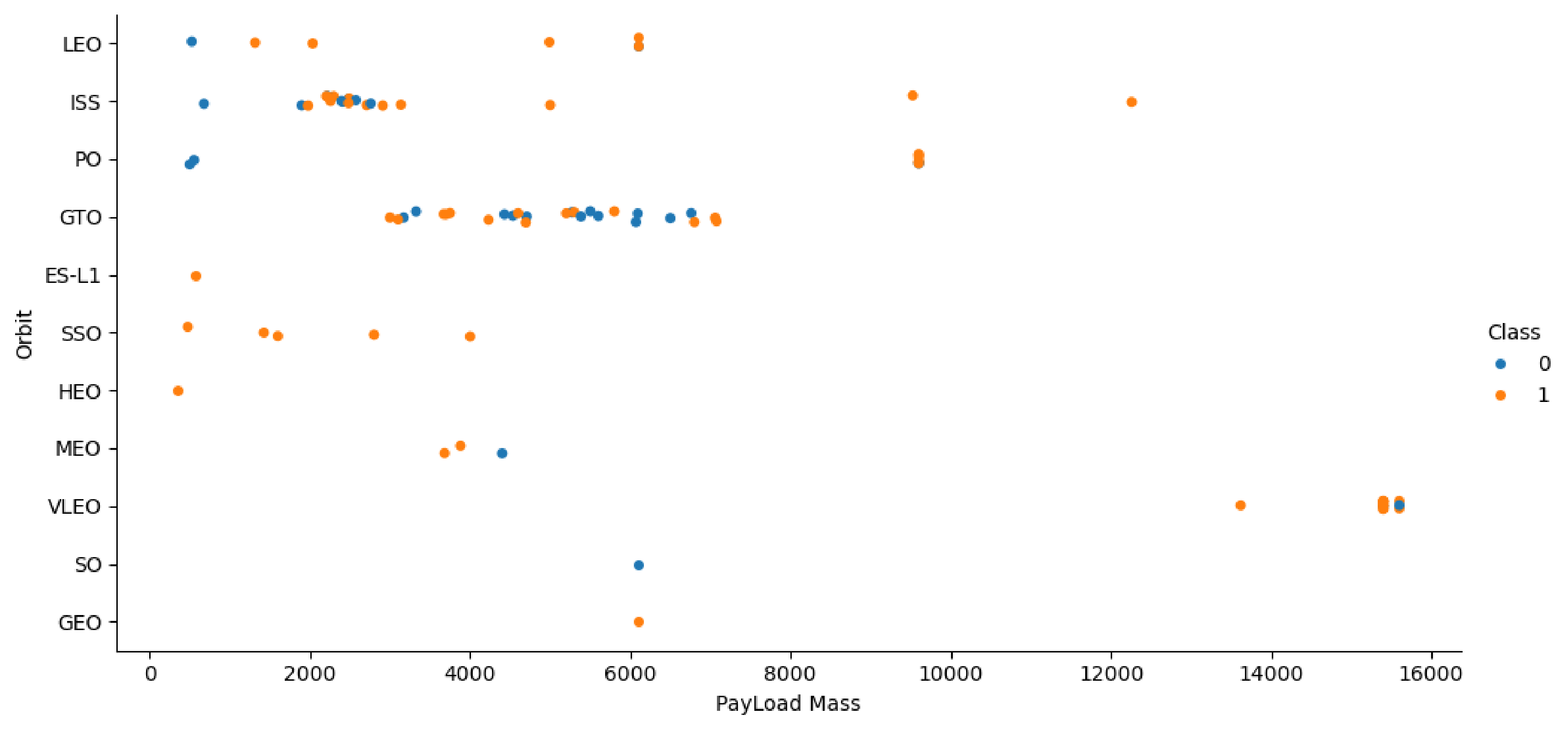
Flight Number vs. Orbit Type

- It seems like there is a positive correlation between these variables



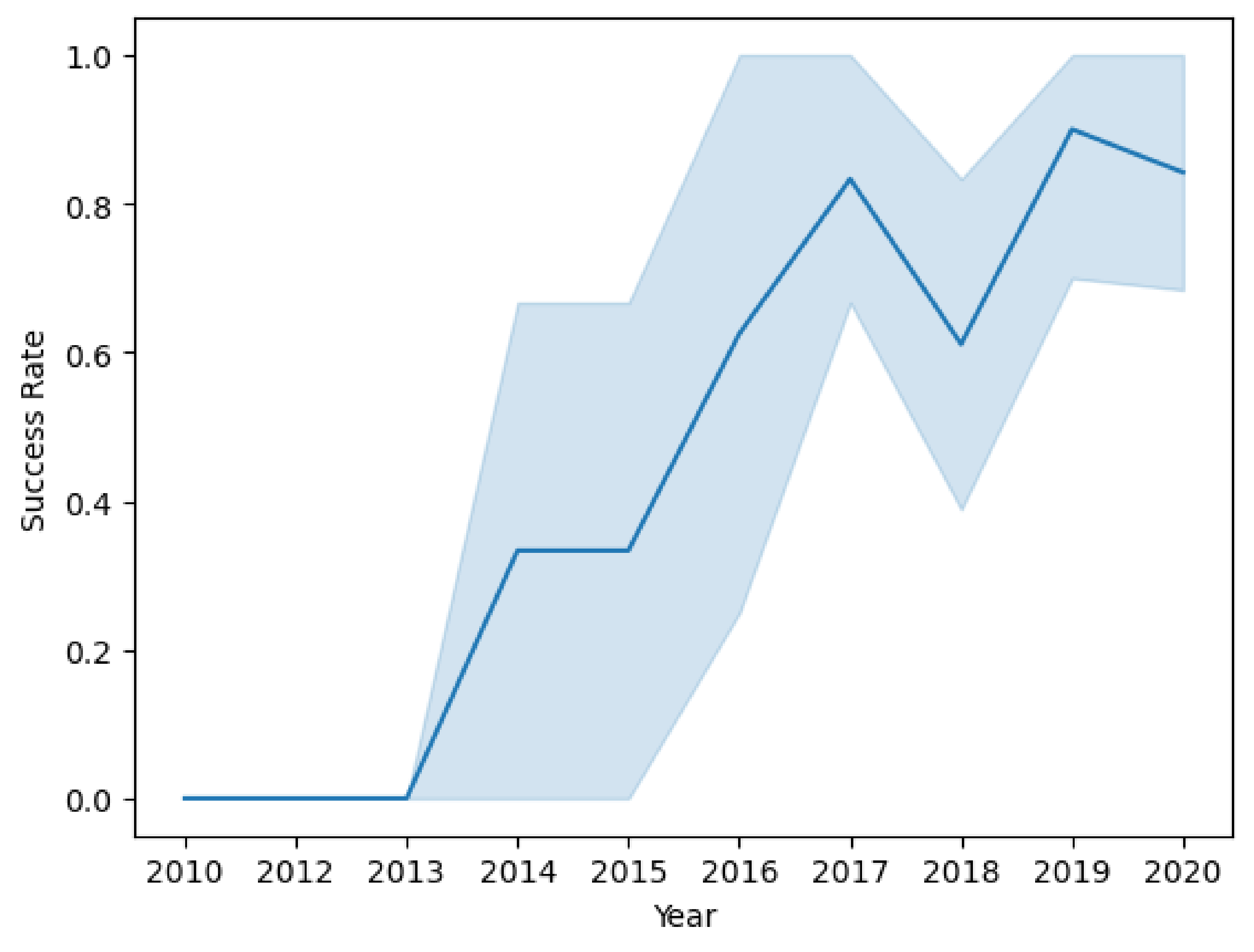
Payload vs. Orbit Type

- Except GTO orbit, it seems that higher payload have higher success rate.



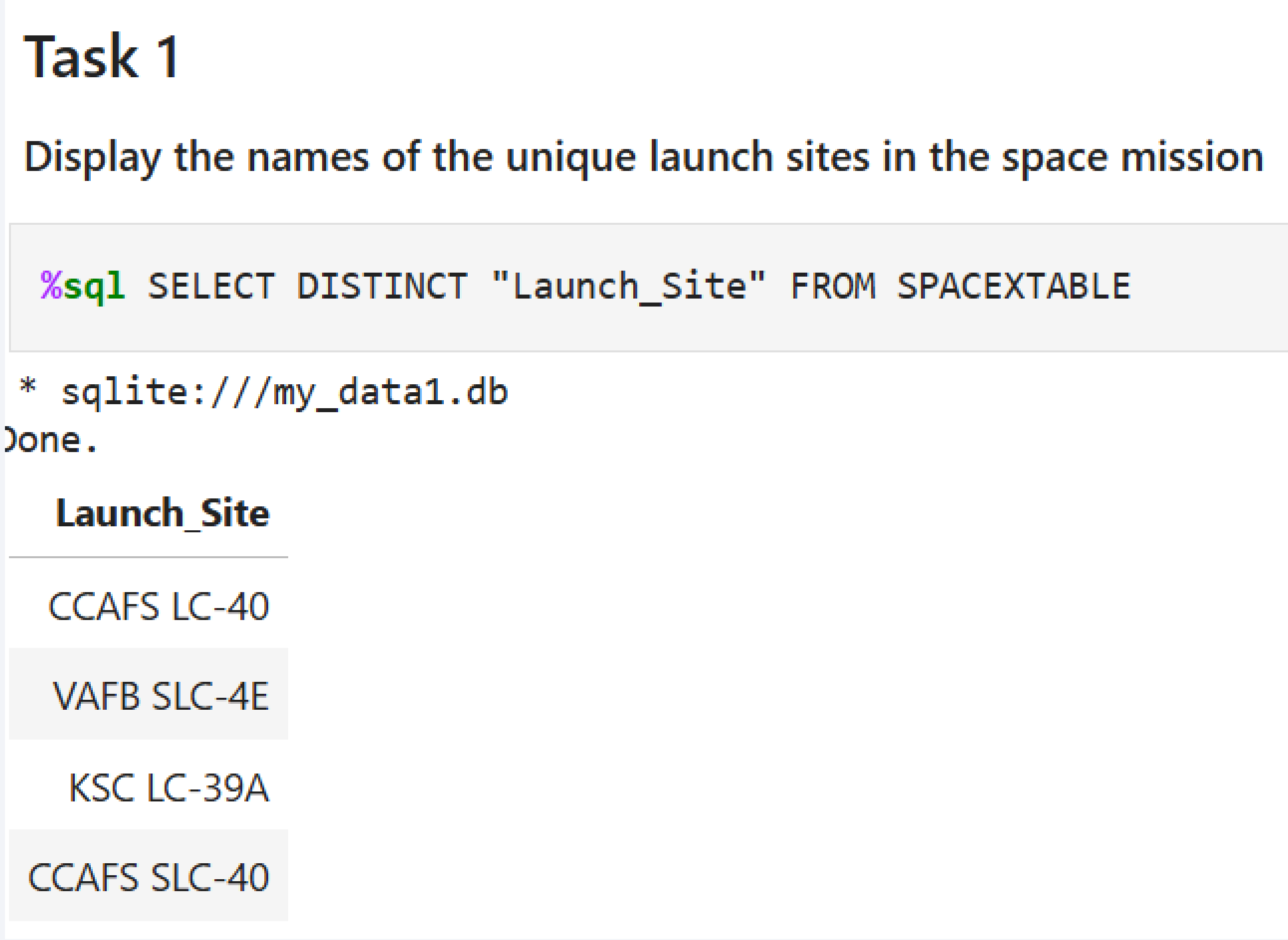
Launch Success Yearly Trend

- High increase from 2013 to 2017 (0 to ~0.8)
- Then decrease to ~0.6 from 2016 to 2018
- Increase again from 2018 to 2019 (to ~0.9)
- Finally, down trend to 0.8



All Launch Site Names

- SELECT DISTINCT is used to get unique values of a column



Task 1

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

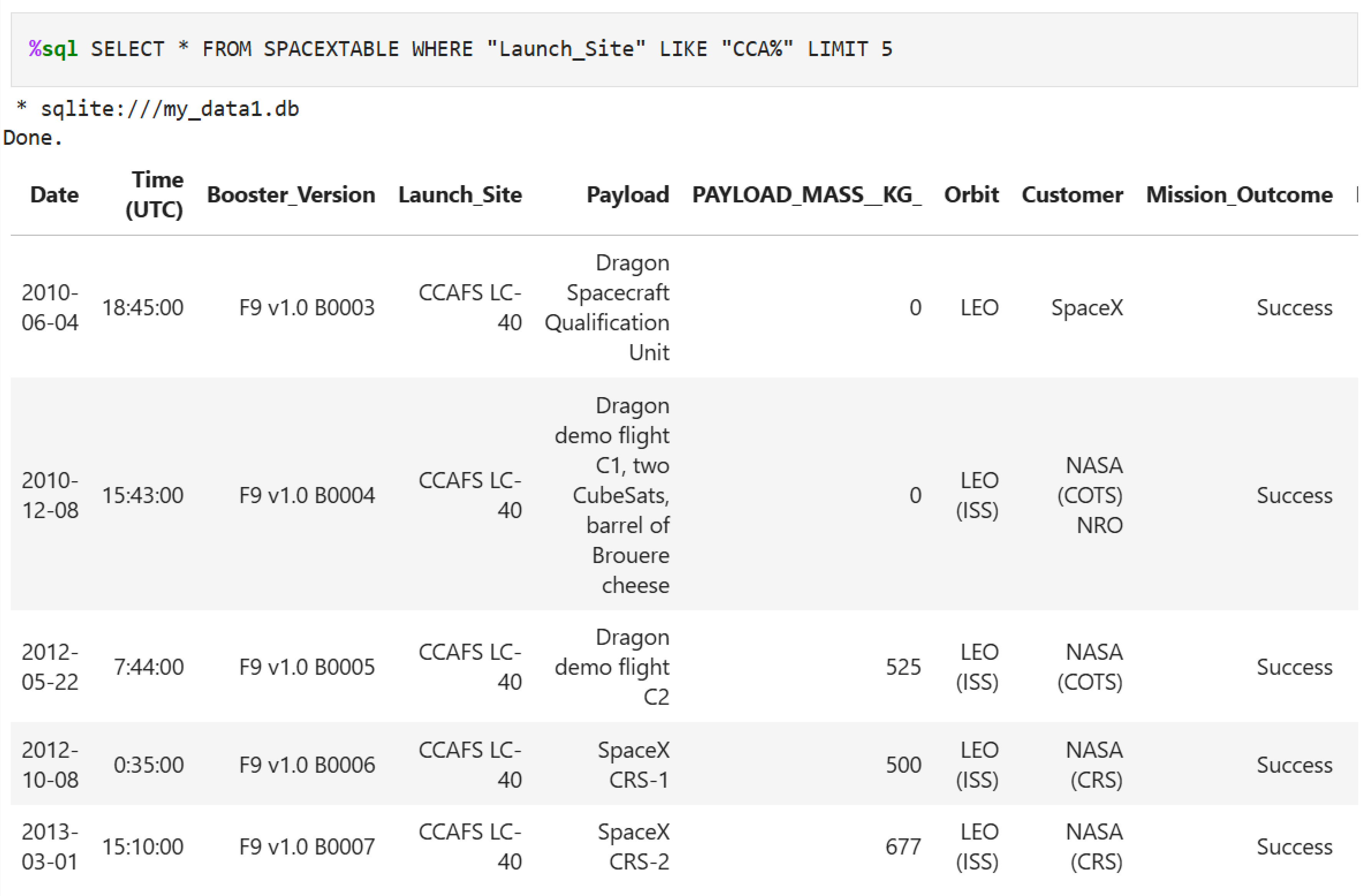
```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
```

```
 * 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'

- LIMIT is used to select a certain number of records



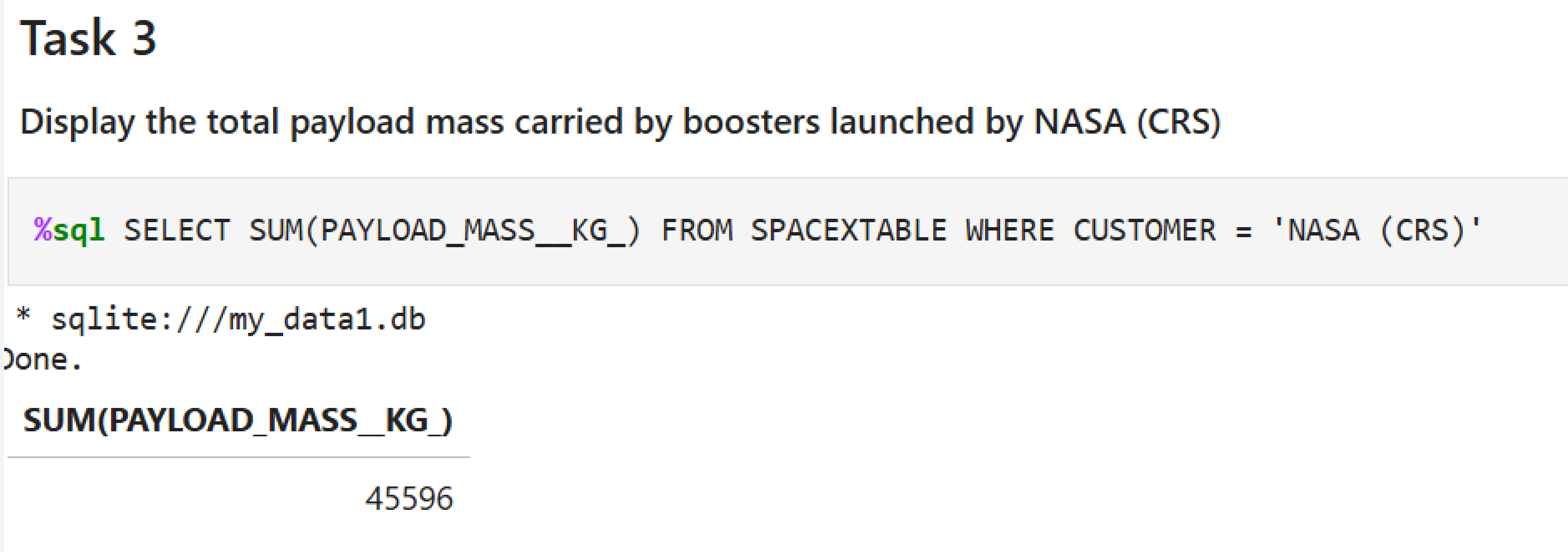
```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE "CCA%" LIMIT 5
```

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

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

Total Payload Mass

- SUM() returns the sum of all records in the column, here a subset has been done thanks to **WHERE**



Task 3

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

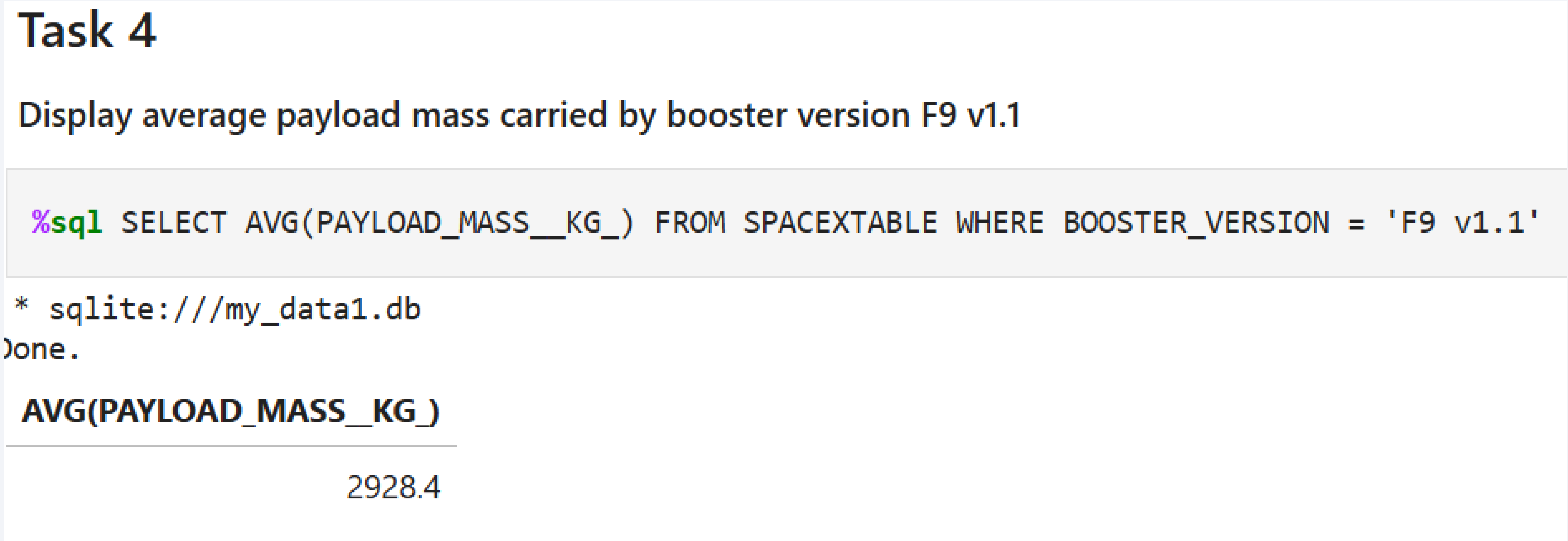
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Same as before but instead of SUM() we use AVG() to get the mean



Task 4

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

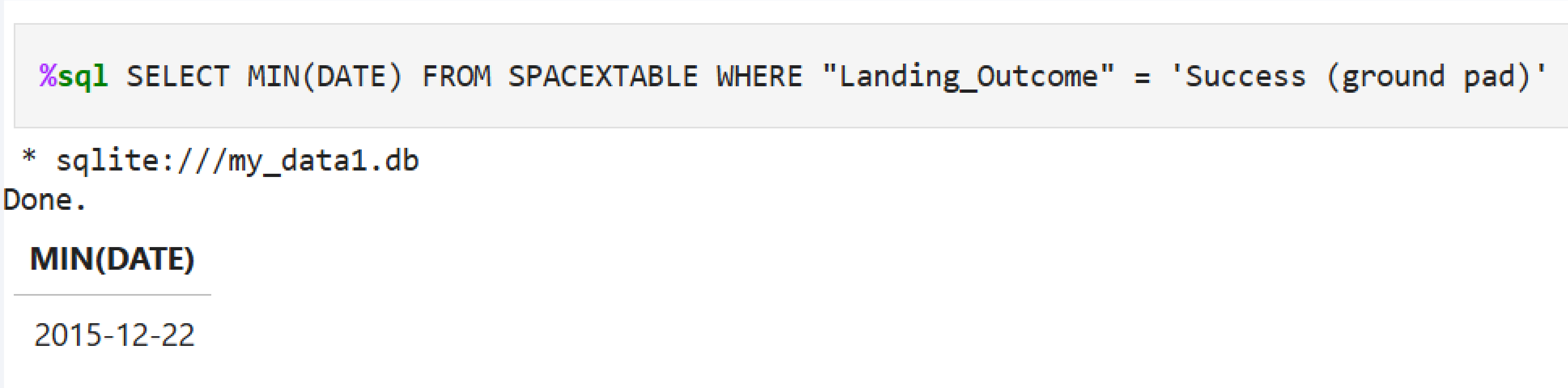
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION = 'F9 v1.1'
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- MIN() get the minimum value of a column



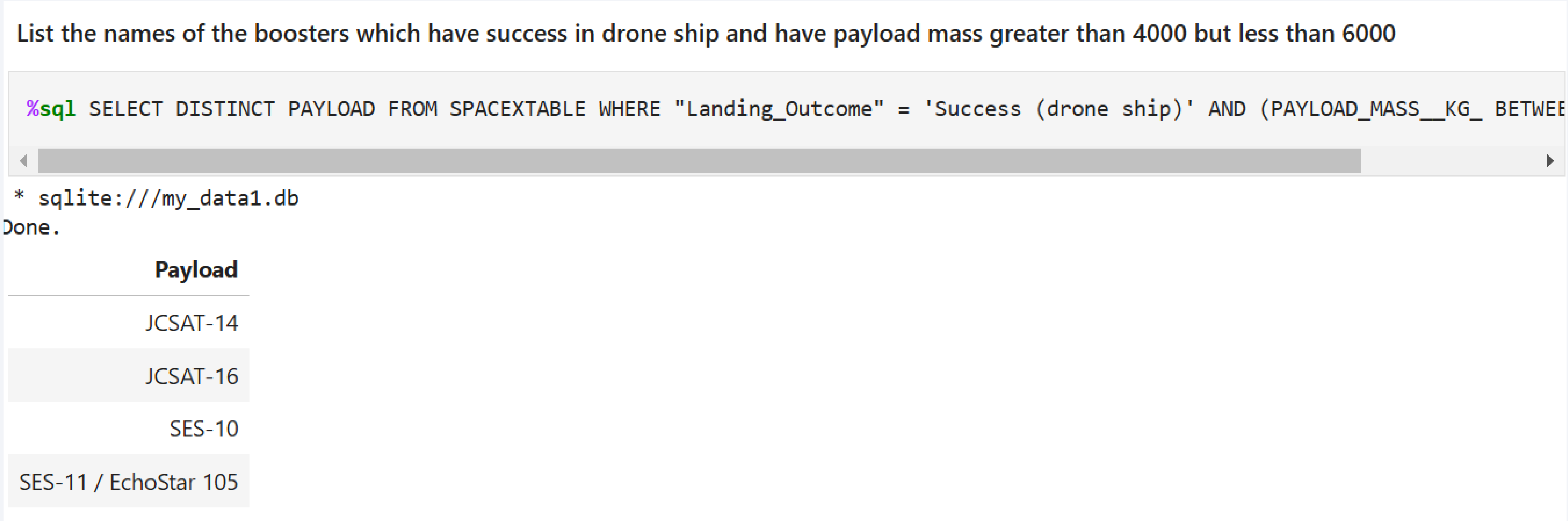
```
%sql SELECT MIN(DATE) FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)'
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query has been cropped, payload is filtered using BETWEEN keyword



List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

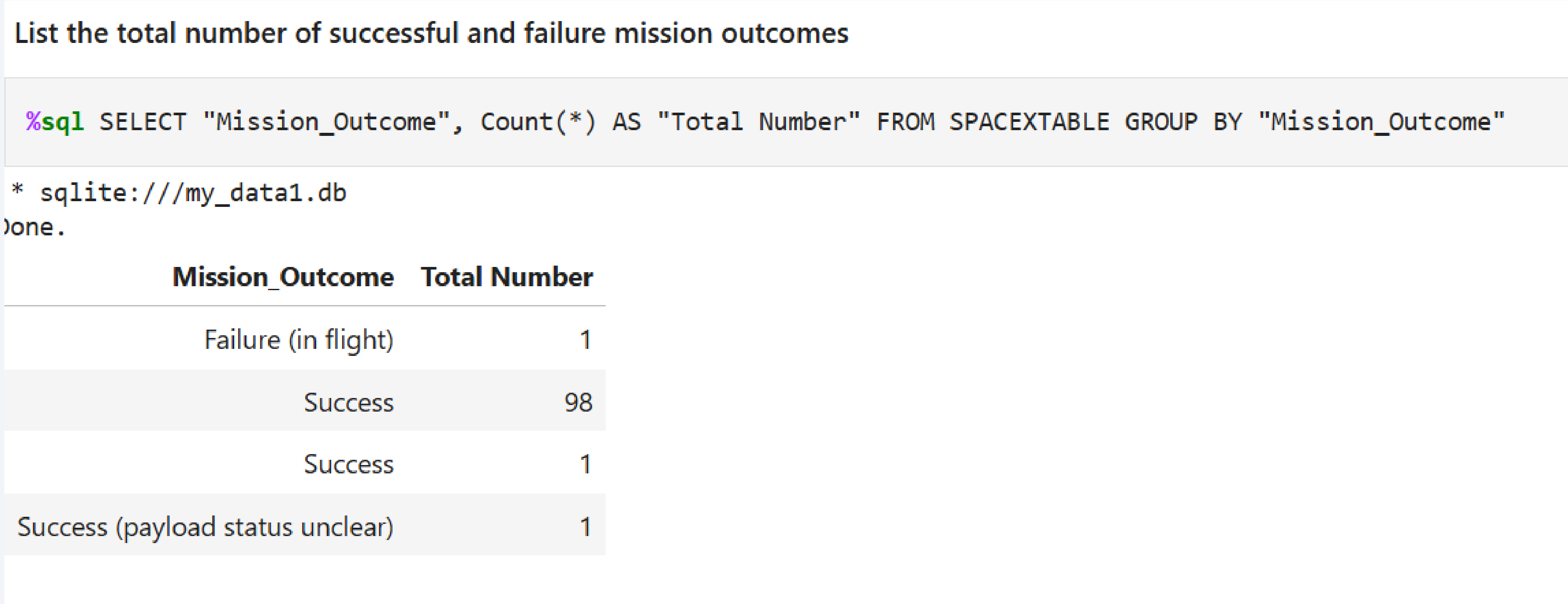
```
%sql SELECT DISTINCT PAYLOAD FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (drone ship)' AND (PAYLOAD_MASS__KG_ BETWEE
```

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

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- Creation of a new column using AS keyword to rename



List the total number of successful and failure mission outcomes

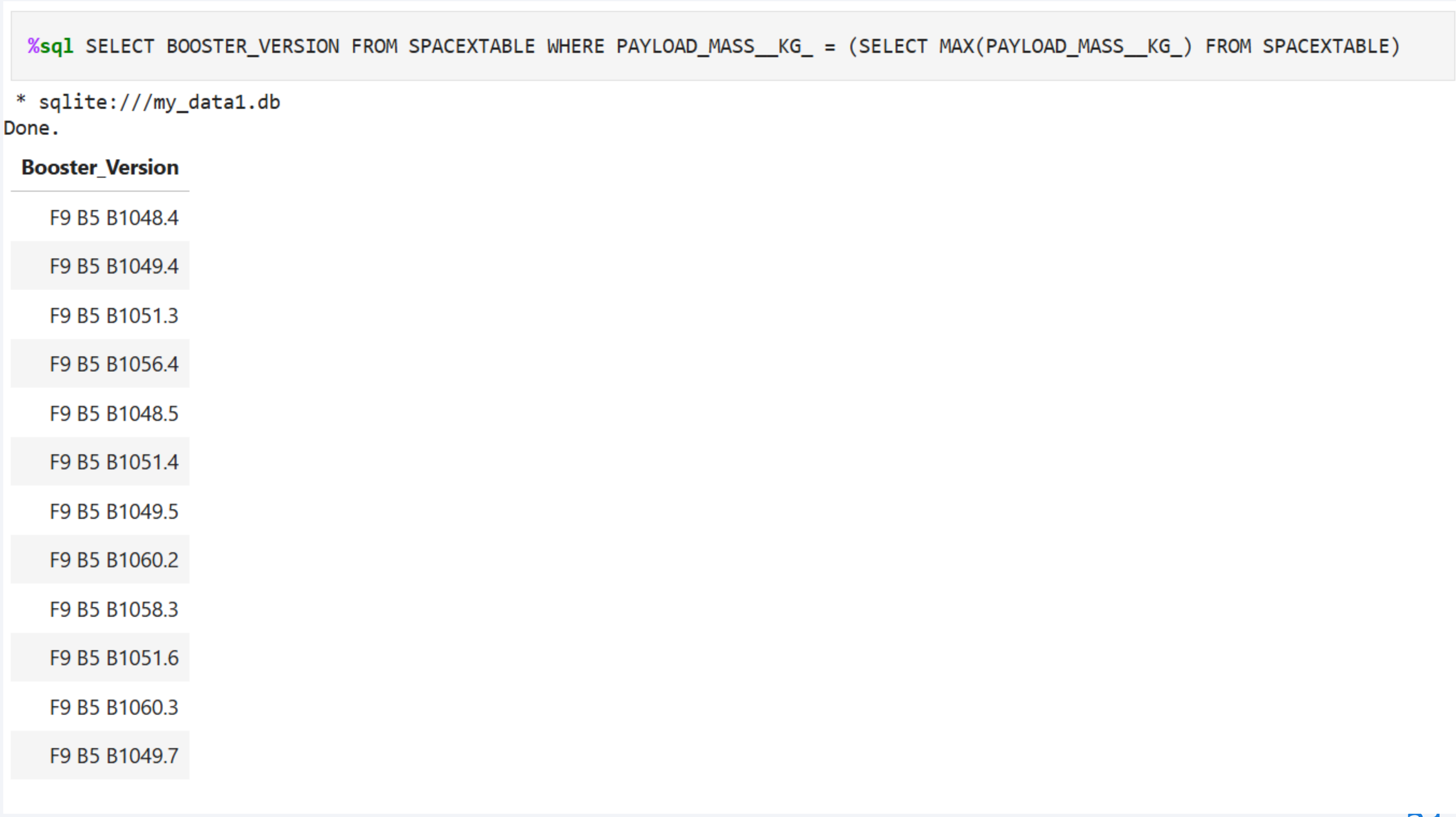
```
%sql SELECT "Mission_Outcome", Count(*) AS "Total Number" FROM SPACEXTABLE GROUP BY "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
)one.
```

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

Boosters Carried Maximum Payload

- Use of a subquery to filter the main table



```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

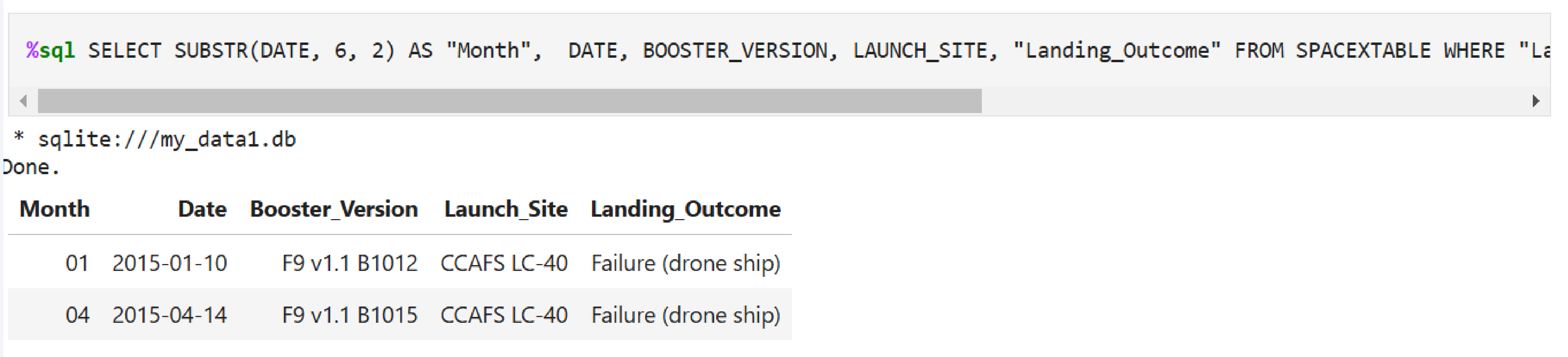
```
 * sqlite:///my_data1.db
Done.
```

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

- Query is :

%sql SELECT SUBSTR(DATE, 6, 2) AS "Month", DATE, BOOSTER_VERSION, LAUNCH_SITE, "Landing_Outcome" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND SUBSTR(DATE, 0, 5) = '2015'



```
%sql SELECT SUBSTR(DATE, 6, 2) AS "Month",  DATE, BOOSTER_VERSION, LAUNCH_SITE, "Landing_Outcome" FROM SPACEXTABLE WHERE "La
```

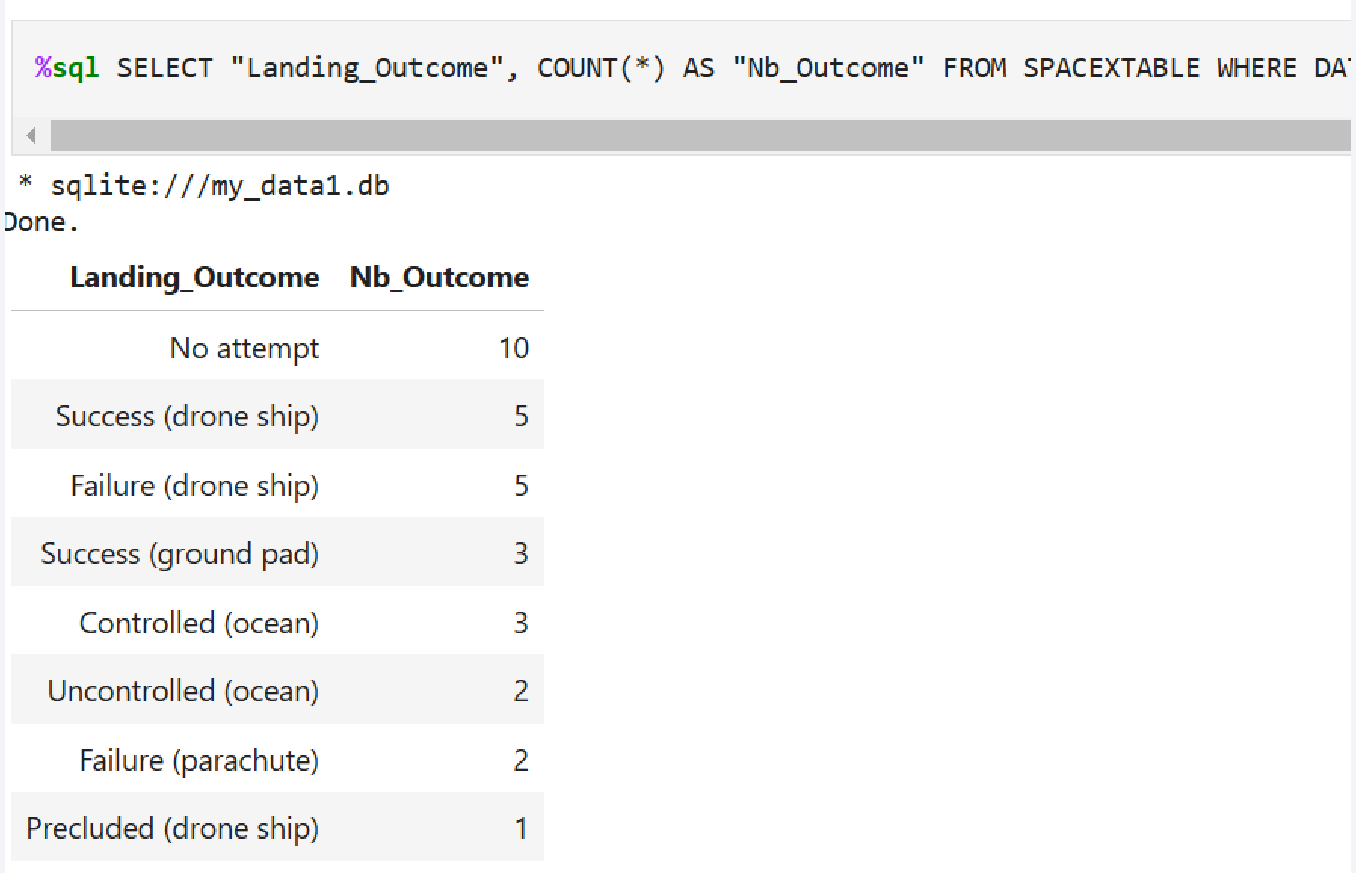
```
 * sqlite:///my_data1.db
Done.
```

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Query :

%sql SELECT "Landing_Outcome", COUNT(*) AS "Nb_Outcome" FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY "Nb_Outcome" DESC



```
%sql SELECT "Landing_Outcome", COUNT(*) AS "Nb_Outcome" FROM SPACEXTABLE WHERE DA
```

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

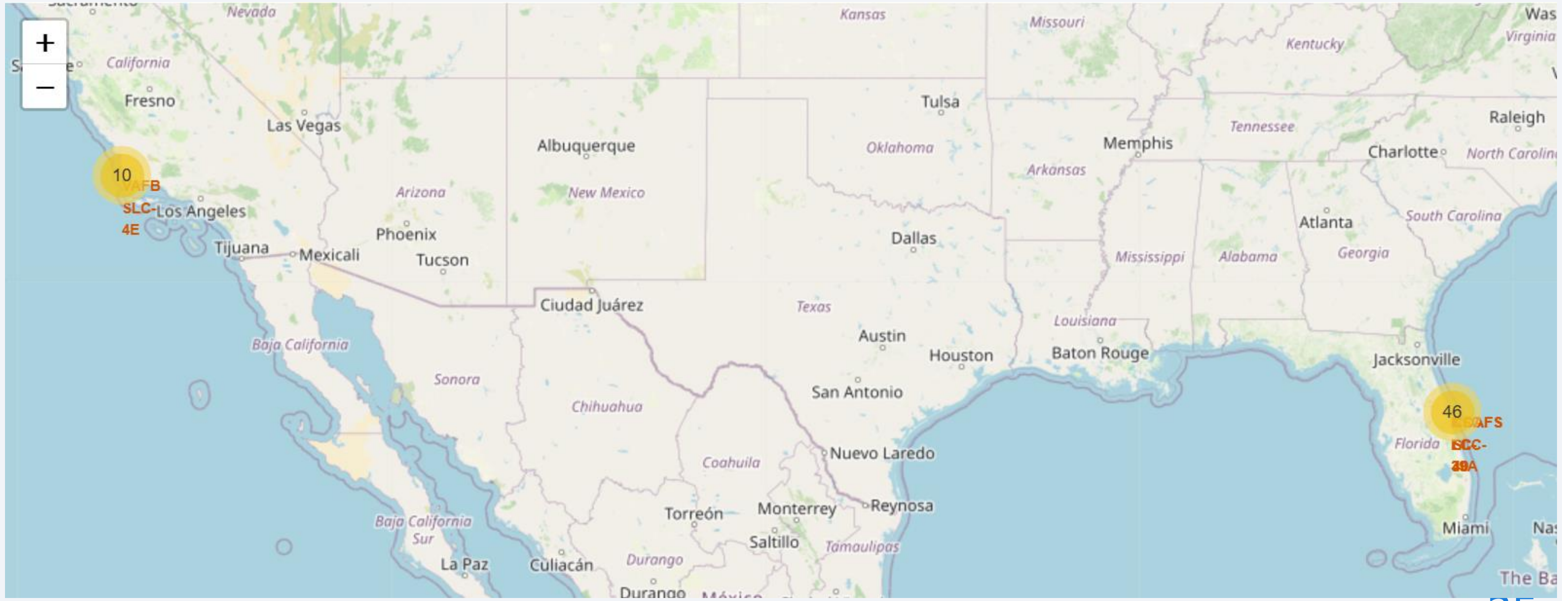
Landing_Outcome	Nb_Outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Localization of Launch Site

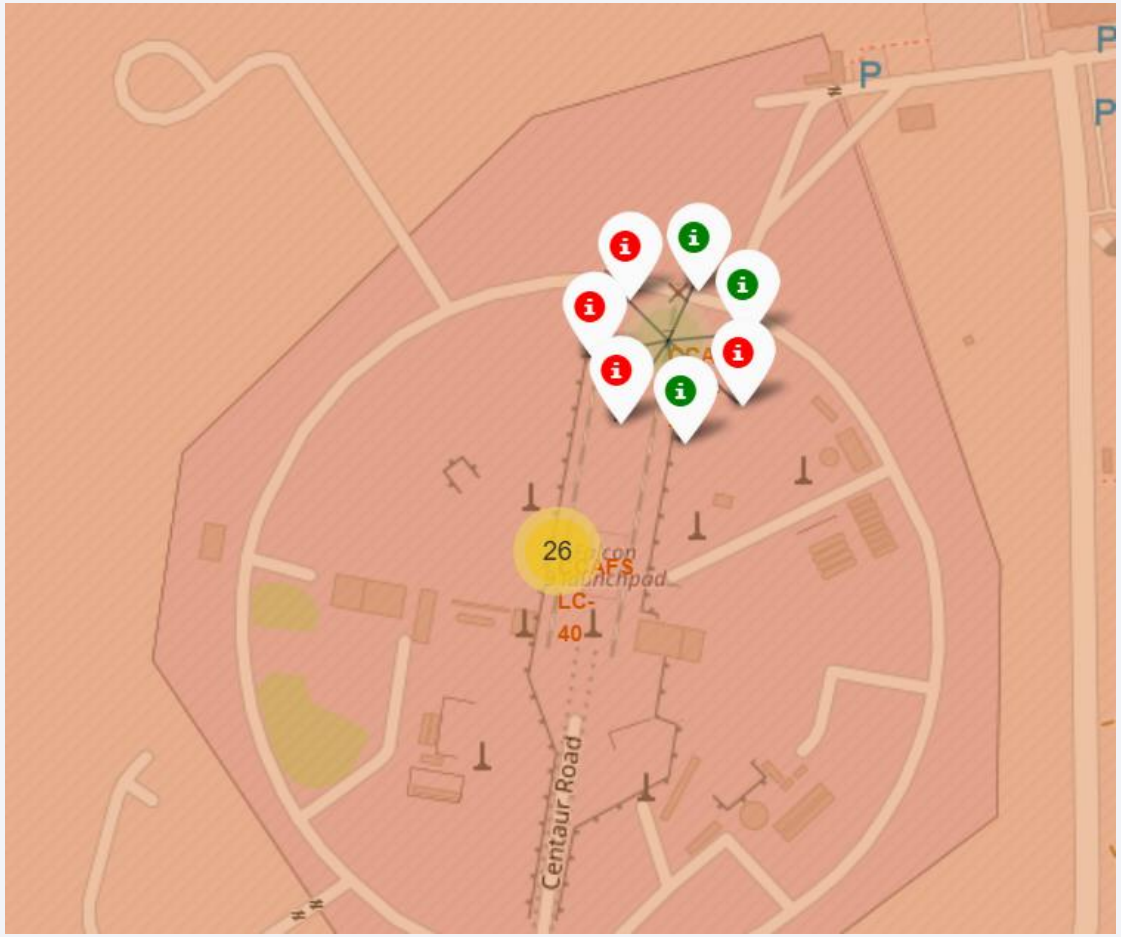
- We can see Launch Site name and their localization.
- All are near coast, one east one west.



Launch Site | Launch Outcome

- Green markers for **successful** launches
- Red markers for **unsuccessful** launches

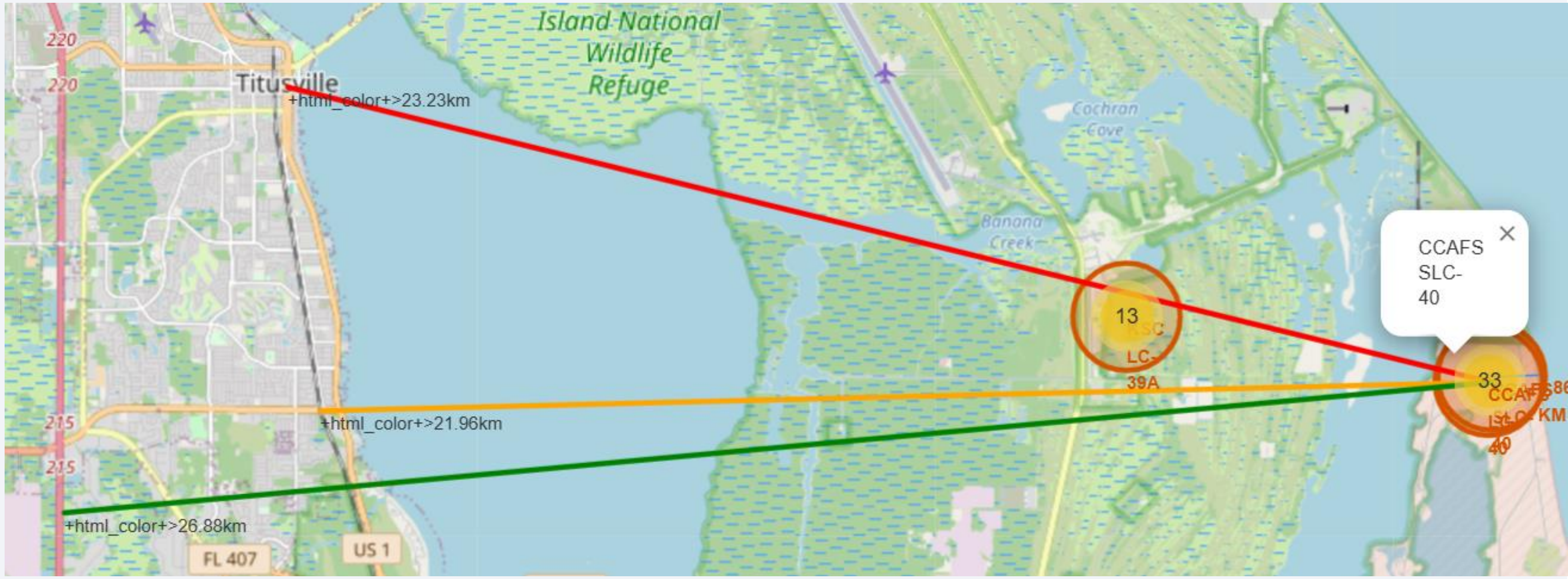
- On this screenshot we can see that CCA have a really good success rate.

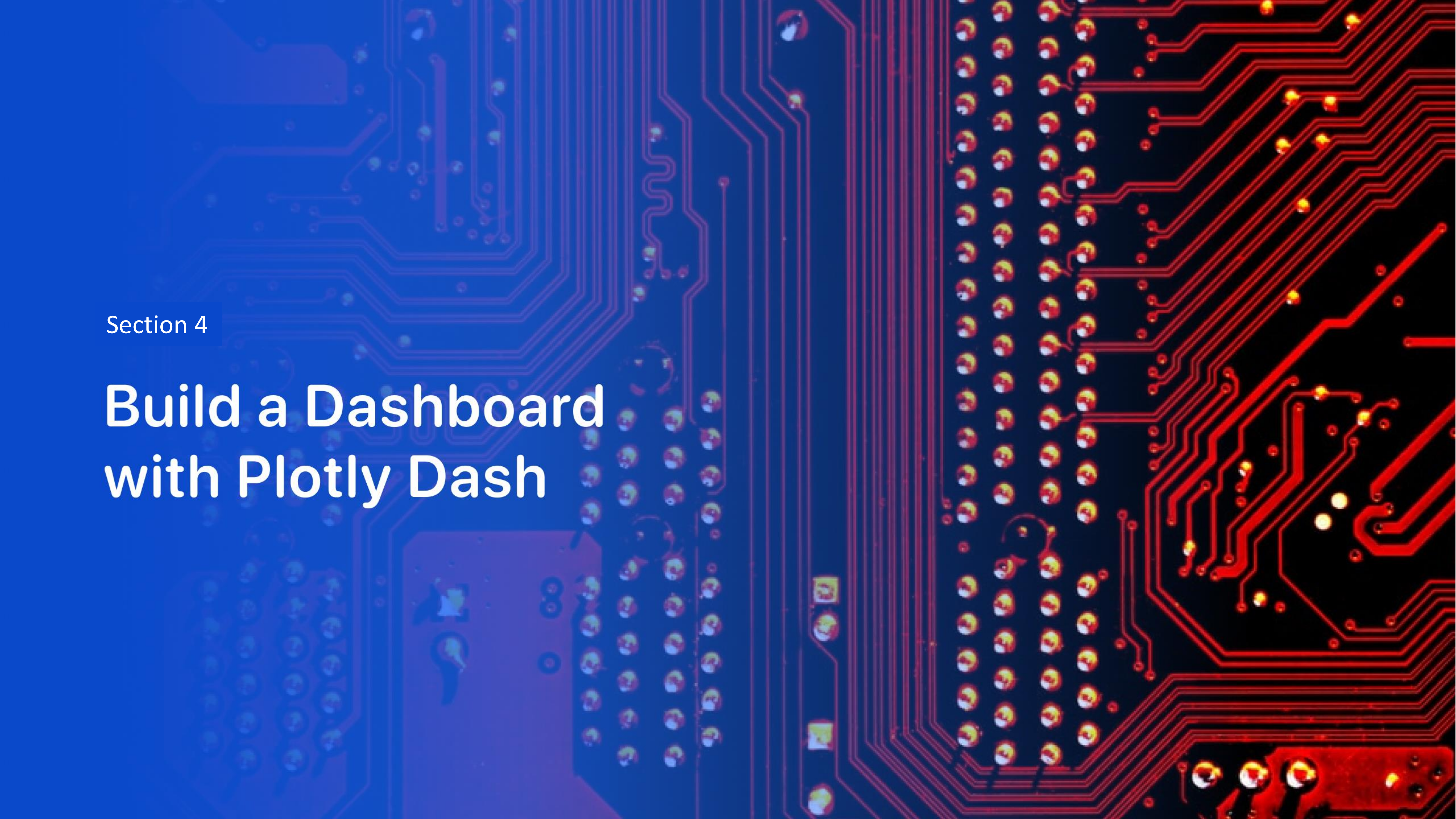


Distance to public installation

- City Distance 23.23 km
- Railway Distance 21.96 km
- Highway Distance 26.88 km
- Coastline Distance 0.86 km

Close to the coast but away from the public installation in case of a failure





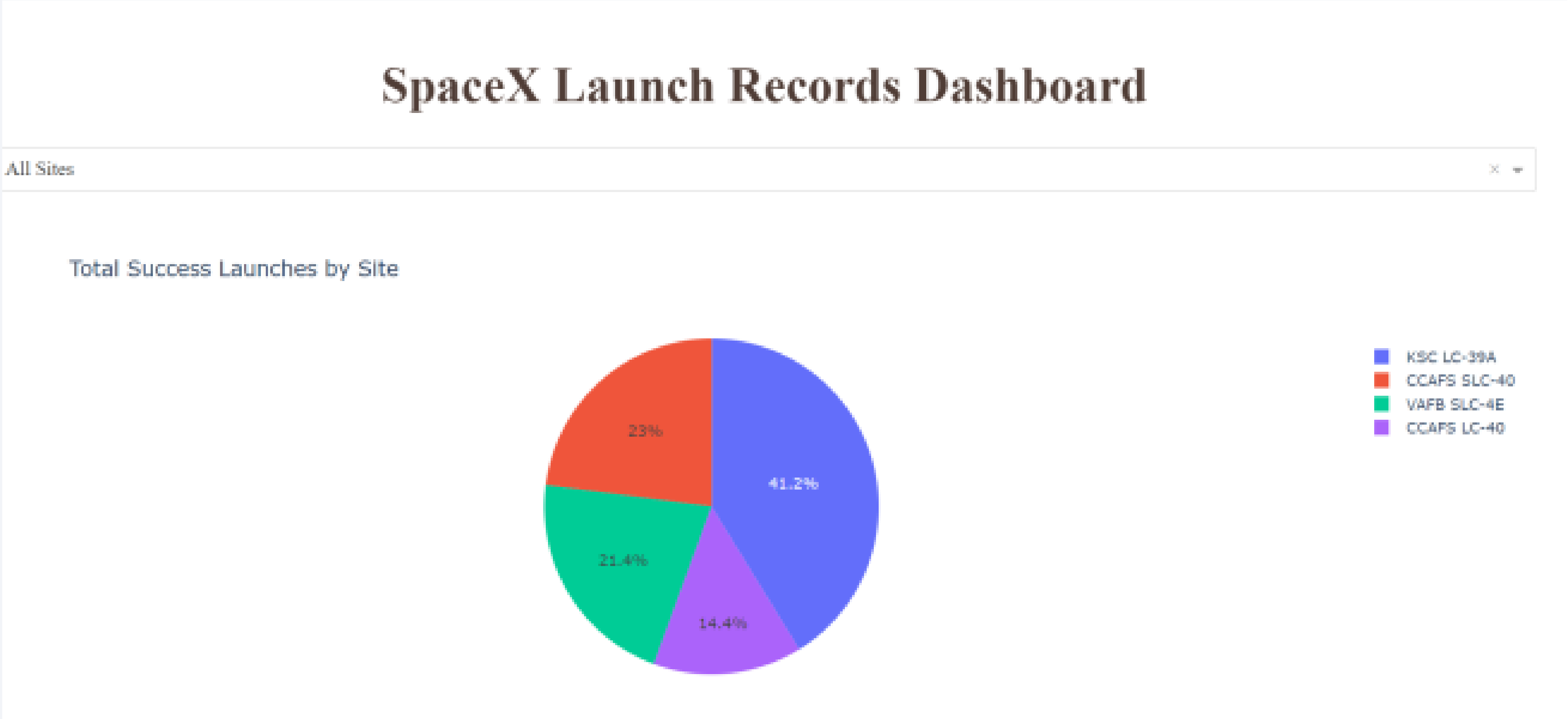
Section 4

Build a Dashboard with Plotly Dash

Full scope (all launch site) success rate

- KSC : 41,2%
- CCAPS SLC- : 23%
- CCAPS LC : 14,4%
- VAFB : 21.4 %

KSC is the best launch site while CCAPS LC is the worst



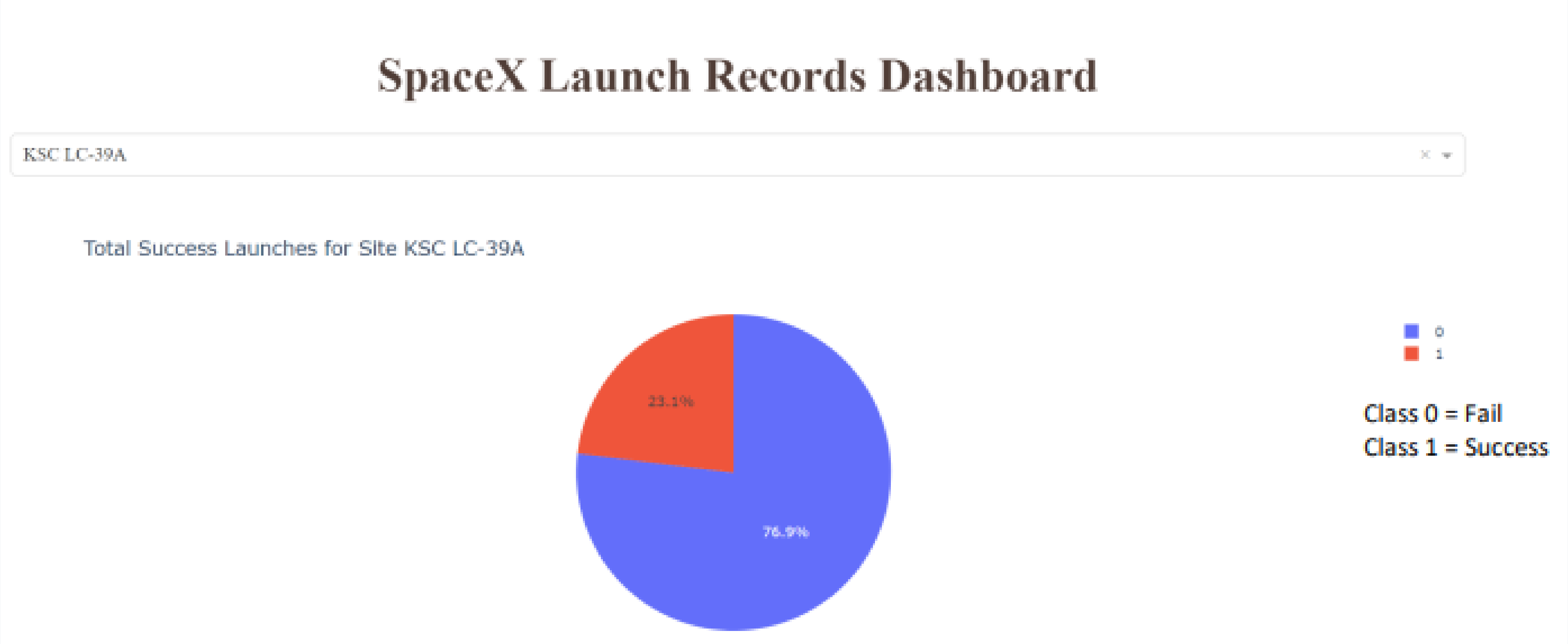
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

Success Rate

- KSC site launch has the highest success rate as we already said, with almost 77%

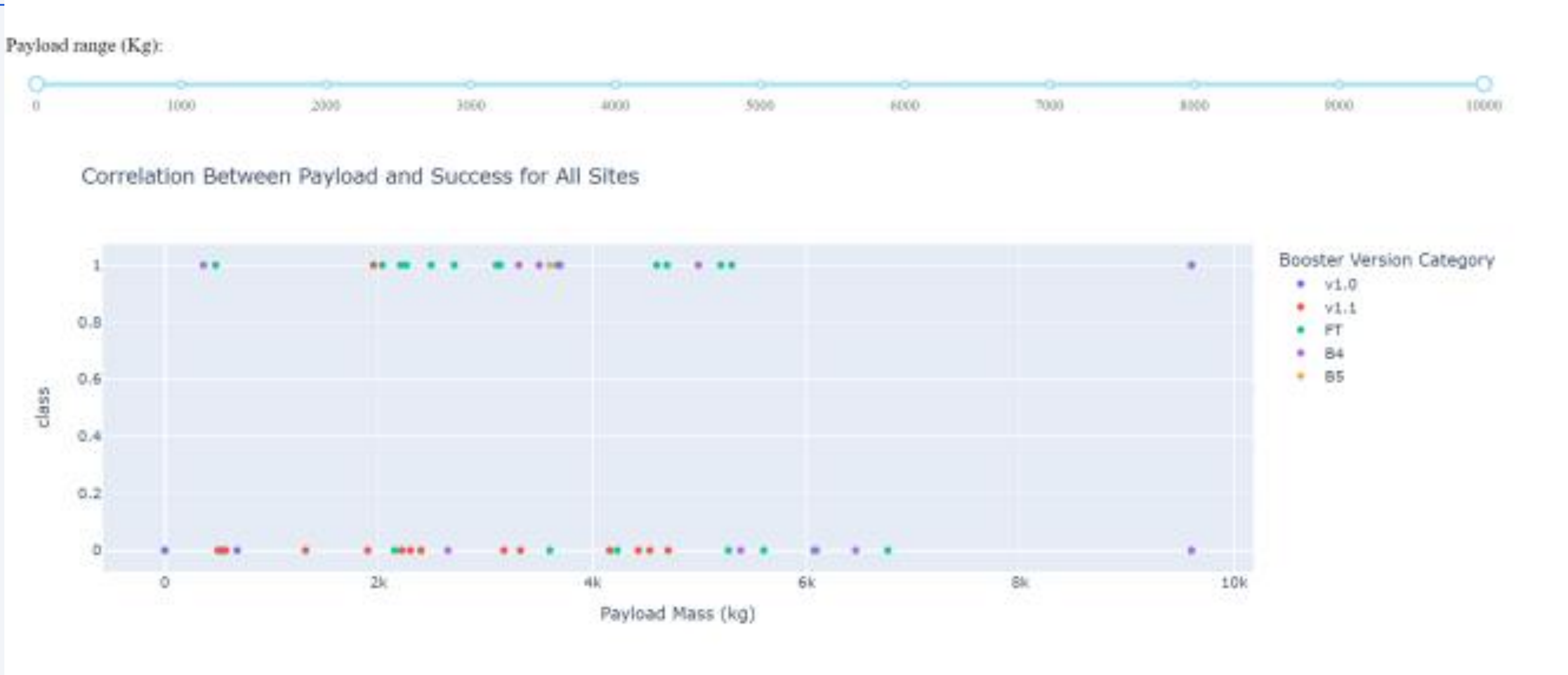


SpaceX Launch Records Dashboard

KSC LC-39A

Class 0 = Fail
Class 1 = Success

Payload vs. Launch Outcome



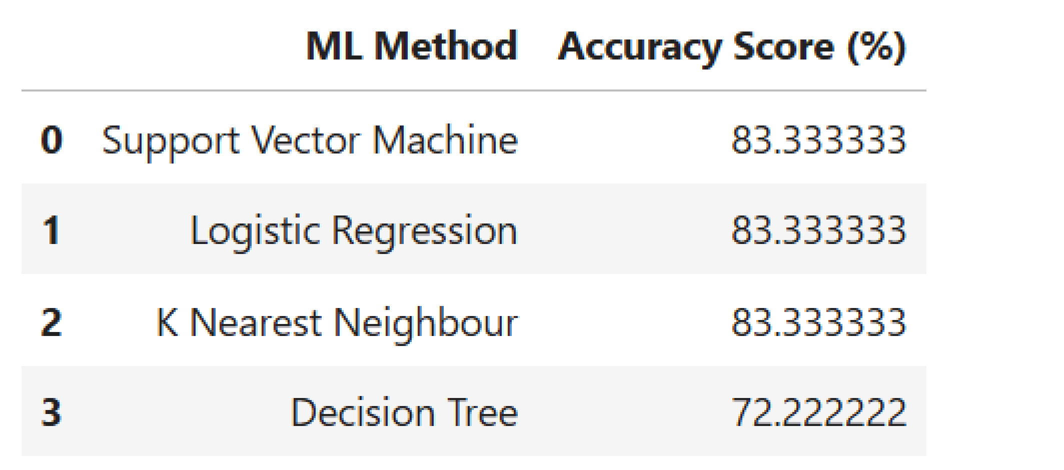
- Payloads between 2000 kg and 5000 kg have the highest success rate and most of them are FT booster version. v1.1 booster almost always failed

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 3 models performed at the same level and had the same scores and accuracy.
- This is likely due to the small training set and test set



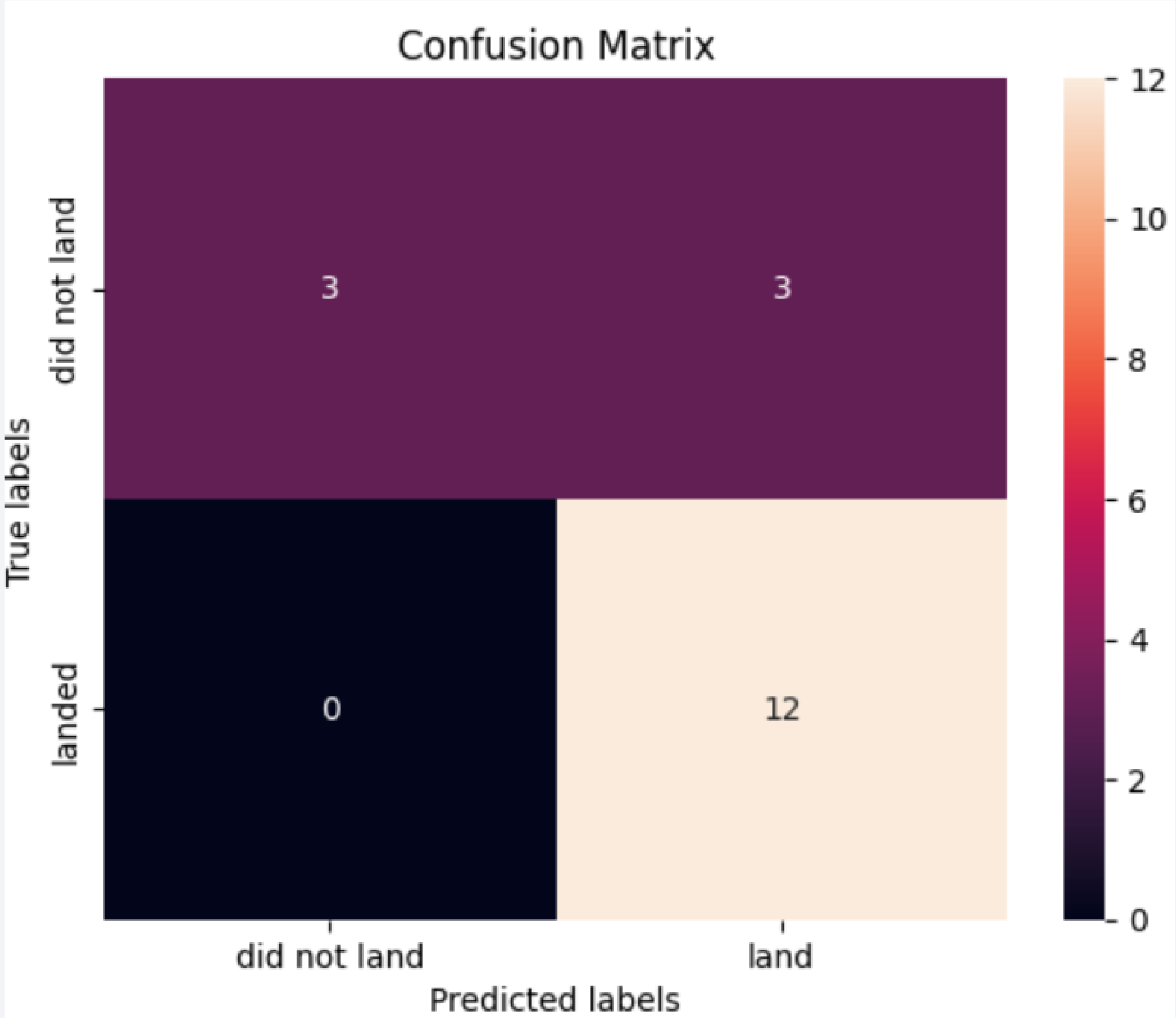
	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	72.222222

Confusion Matrix

- Model here successfully predicted landed records (12 out of 12) but miss predicted 3 records in did not land category.

- 12 True positive / 3 True negative / 3 False positive / 0 False Negative

- We can calculate Precision (TP / (TP + FP)) / Recall (TP / (TP + FN)) / F1_Score (2 * (Precision * Recall) / (Precision + Recall)) and Accuracy ((TP + TN) / (TP + TN + FP + FN))



Conclusions

- Model Performance:
  - The models performed similarly on the test set with decision tree classifier slightly underperforming
- Coast:
  - All the launch sites are close to the coast
- Launch Success:
  - Increases over time
- KSC LC-39A:
  - Has the highest success rate among launch sites. Has a perfect success rate for launches low payload (<5500 kg)
- Orbits:
  - ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass:
  - Across all launch sites, the higher the payload mass (kg), the higher the success rate

Future Improvement

- In order to have better result, we need to have more records in our base dataset.
- We can explore more feature and some dimension reduction techniques such as PCA / UMAP / T-SNE
- Explore algorithm such as DBSCAN or even deep learning solution with neural networks
- We used decision tree, we could explore Random Forest and XGBoost model

Thank you!