IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vikas Movva
2023-12-12

Outline

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

Executive Summary

Summary of methodologies

- In pursuit of identifying the determinants of a successful rocket landing, this research employs the following methodologies:
- Data collection through SpaceX REST API and web scraping techniques.
- Data wrangling to establish a success/fail outcome variable.
- Data exploration utilizing visualization techniques, considering payload, launch site, flight number, and yearly trends.
- Data analysis through SQL, encompassing calculations such as total payload, payload range for successful launches, and the overall count of successful and failed outcomes.
- Investigation into launch site success rates and their proximity to geographical markers.
- Visualization of launch sites showcasing the highest success rates and successful payload ranges.
- Development of predictive models for landing outcomes, employing logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN).Summary of all results

Results

Exploratory Data Analysis:

- Orbits ES-L1, GEO, HEO, and SSO have the highest success rates
- The KSC LC-39A launch site has the highest success rate
- The success rate of launches has gone up over time

Visualization

- All launch sites are in the southern united states (closest to the equator) and close to the coast

Predictive Analytics

- All machine learning models preformed similarly on the test data set however the decision tree model as a marginally better accuracy.

Introduction

Background

SpaceX stands out as the foremost successful entity in the commercial space era, revolutionizing space travel through cost-effectiveness. Highlighted on its website are Falcon 9 rocket launches, priced at 62 million dollars, in stark contrast to other providers whose costs soar above 165 million dollars per launch. The substantial savings stem from SpaceX's innovative practice of reusing the first stage. Consequently, forecasting the successful landing of the first stage becomes pivotal in determining the overall launch cost. Leveraging publicly available information and machine learning models, our objective is to predict whether SpaceX will opt to reuse the first stage.

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?

Section 1

Methodology

Methodology

Data collection methodology:

- Data was collected from the SpaceX **REST API** and Wikipedia's **List of Falcon 9 Launches**

Perform data wrangling

- Data was filtered looking for missing values and categorical values which were then encoded using one hot encoding

Perform exploratory data analysis (EDA) using visualization and SQL

- Data was visualized using the **Seaborn** library
- Calculations on the data was made through **SQL** queries

Perform interactive visual analytics using Folium and Plotly Dash

- Both an interactive map with launch sites and successful/failed launches and key statistical relationships were presented using **Folium** and **Dash** respectively

Perform predictive analysis using classification models

- Using the dataset created through the previous steps various machine learning models were created using **Grid-search** algorithms to find the best hyperparameters for each model

Data Collection

Describe how data sets were collected.

- Data was collected from the SpaceX REST API as well as Wikipedia’s List of Falcon 9 Launches webpage.

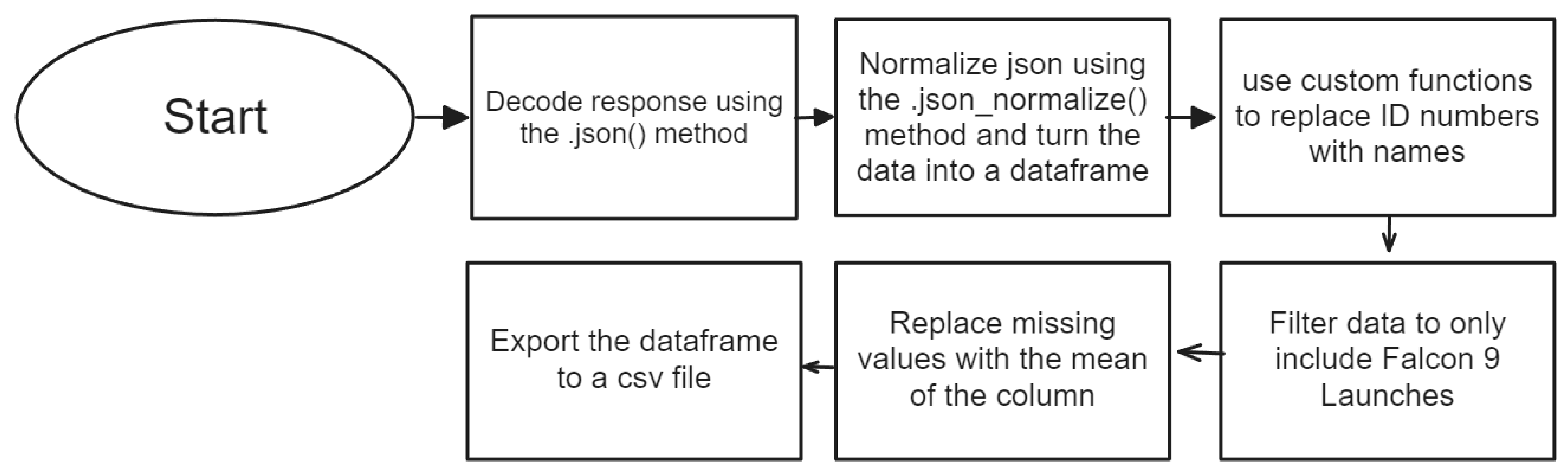
Data collected through the SpaceX REST API:

- Flight number, date, Booster version, Payload mass, Orbit, Launch site, Outcome, Flight, Grid fins, Reused, Legs, Landing pad, Block, Reused count, Serial, Longitude, Latitude.

Data collected through web scraping Wikipedia:

- Flight number, Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Booster version, Booster landing, Date, Time

Data Collection – SpaceX API



Start

Decode response using the .json() method

Normalize json using the .json_normalize() method and turn the data into a dataframe

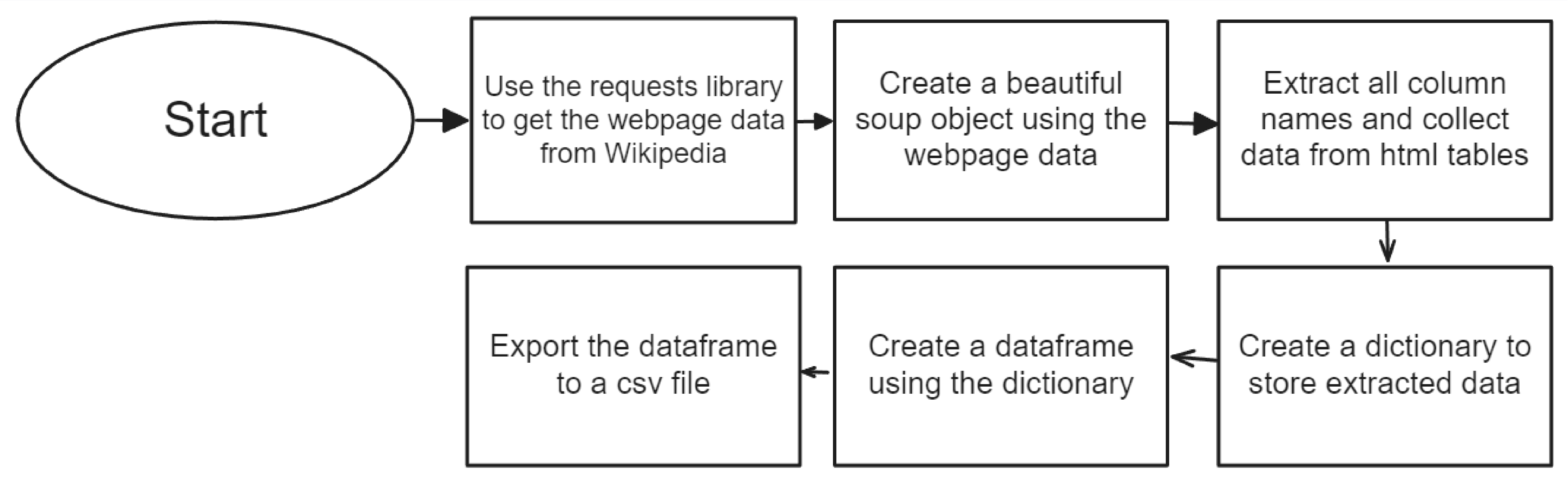
use custom functions to replace ID numbers with names

Filter data to only include Falcon 9 Launches

Replace missing values with the mean of the column

Export the dataframe to a csv file

Data Collection - Scraping



Start
Use the requests library to get the webpage data from Wikipedia
Create a beautiful soup object using the webpage data
Extract all column names and collect data from html tables
Create a dictionary to store extracted data
Create a dataframe using the dictionary
Export the dataframe to a csv file

Data Wrangling

Perform EDA and determine data labels

Calculate:

- # of launches for each site
- # and occurrence of orbit
- # and occurrence of mission outcome per orbit type]

Create binary landing outcome column (dependent variable)

Export data to csv file

Landing Outcomes

- Landing was not always successful
- True Ocean: mission outcome had a successful landing to a specific region of the ocean
- 2023 Landing Outcome Cont.
- False Ocean: represented an unsuccessful landing to a specific region of ocean
- True RTLS: meant the mission had a successful landing on a ground pad
- False RTLS: represented an unsuccessful landing on a ground pad
- True ASDS: meant the mission outcome had a successful landing on a drone ship
- False ASDS: represented an unsuccessful landing on drone ship
- Outcomes converted into 1 for a successful landing and 0 for an
- unsuccessful landing

EDA with Data Visualization

Charts:

Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

These charts show the relationship between variables and if they exist could be used as a feature in a machine learning model

EDA with SQL

SQL queries preformed:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA’
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

Build an Interactive Map with Folium

Markers for launch site:

- Added blue circles at JSC with a popup label showing the name and coordinates
- Added red circles at all launch site coordinates with a popup label showing its name using its name using its coordinates

Colored Markers of Launch Outcomes

- Added colored markers of successful and unsuccessful launches at each launch site

Added Distances between launch site to nearest coastline, railway, highway, and city

Build a Dashboard with Plotly Dash

Dashboard elements:

- Dropdown list with launch sites
- Pie chart showing successful launches
- Slider of payload mass range
- Scatter plot showing payload mass vs. Success rate by booster version

Predictive Analysis (Classification)

Created a NumPy array from the Class Column

Standardized the data with the StandardScalar object. Fitted and transformed the data

Split the data into a training set and a testing set using the train_test_split method

Created and used a GridSearchCV object on each machine learning algorithm for hyperparameter optimization

Calculated the accuracy of each model and created a confusion matrix

Identified the best model using different evaluators like theJaccard score, F1 score, Accuracy, etc.

Results

Exploratory Data Analysis

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

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

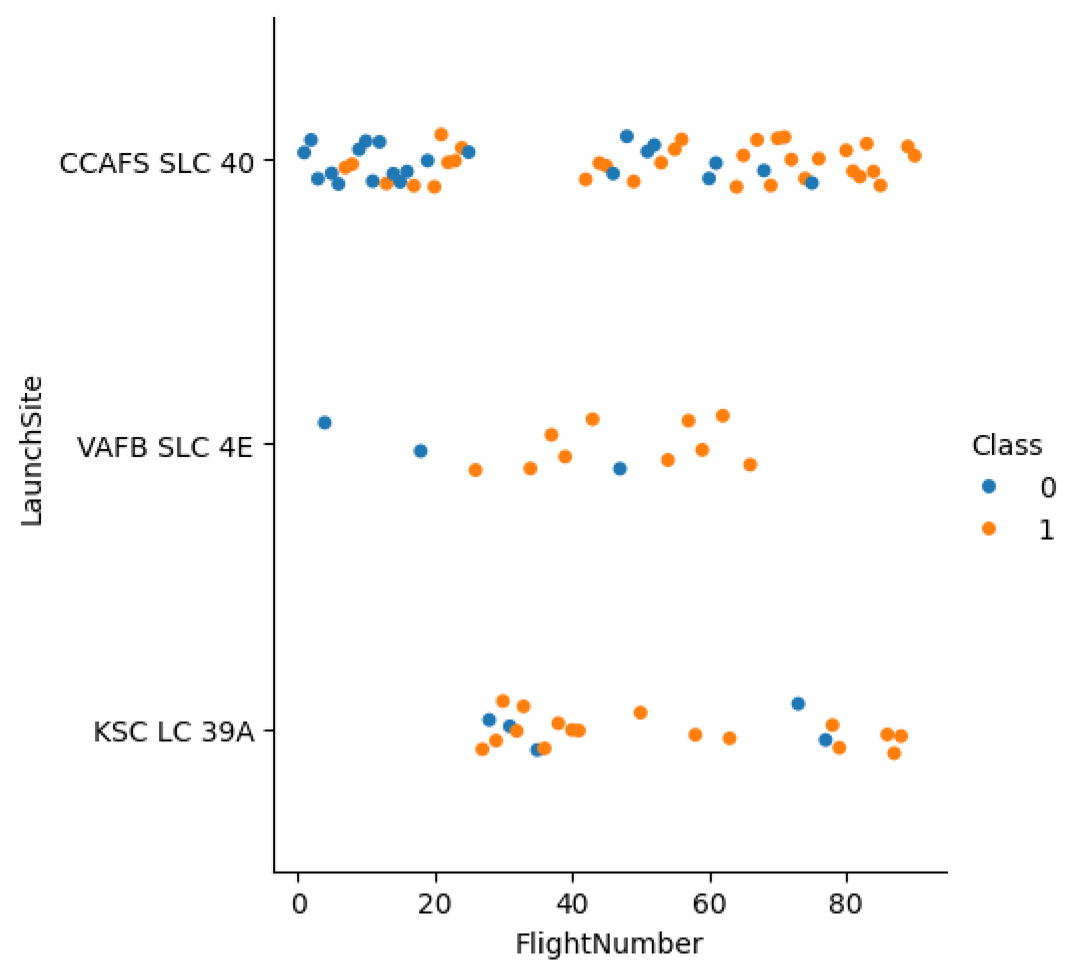
- Decision Tree model is the best predictive model for the dataset

Section 2

Insights drawn from EDA

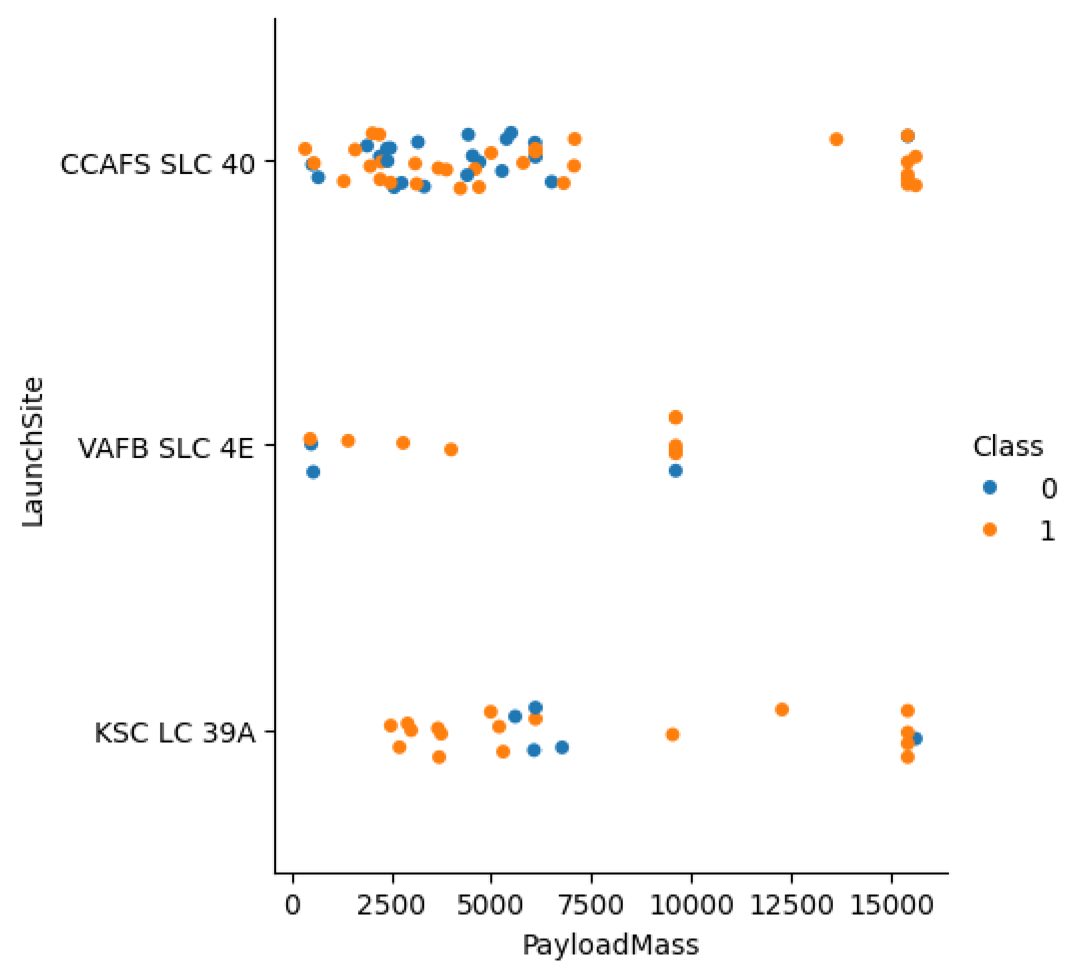
Flight Number vs. Launch Site

- Success rate increases as the flight number increases
- Majority of launches were from the CCAFS SLC 40 launch site
- The other two launch sites have a higher success rate
- The data shows that flight number and success rate has a positive relationship



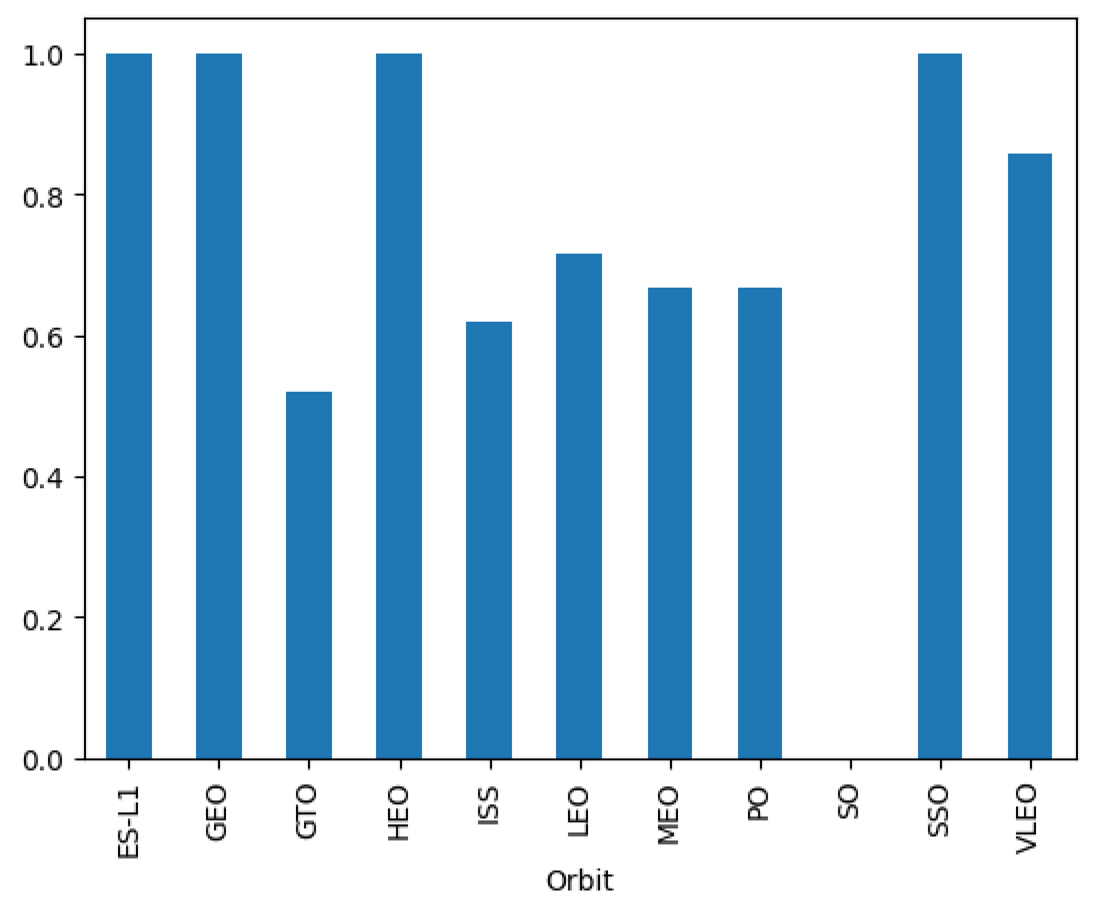
Payload vs. Launch Site

- The data shows that the higher the payload mass the higher the success rate.
- A payload above 12500 at the CCAFS SLC 40 launch site has a 100% success rate
- Payload greater than 10000 was never launched at the VAFB SLC 4E launch site



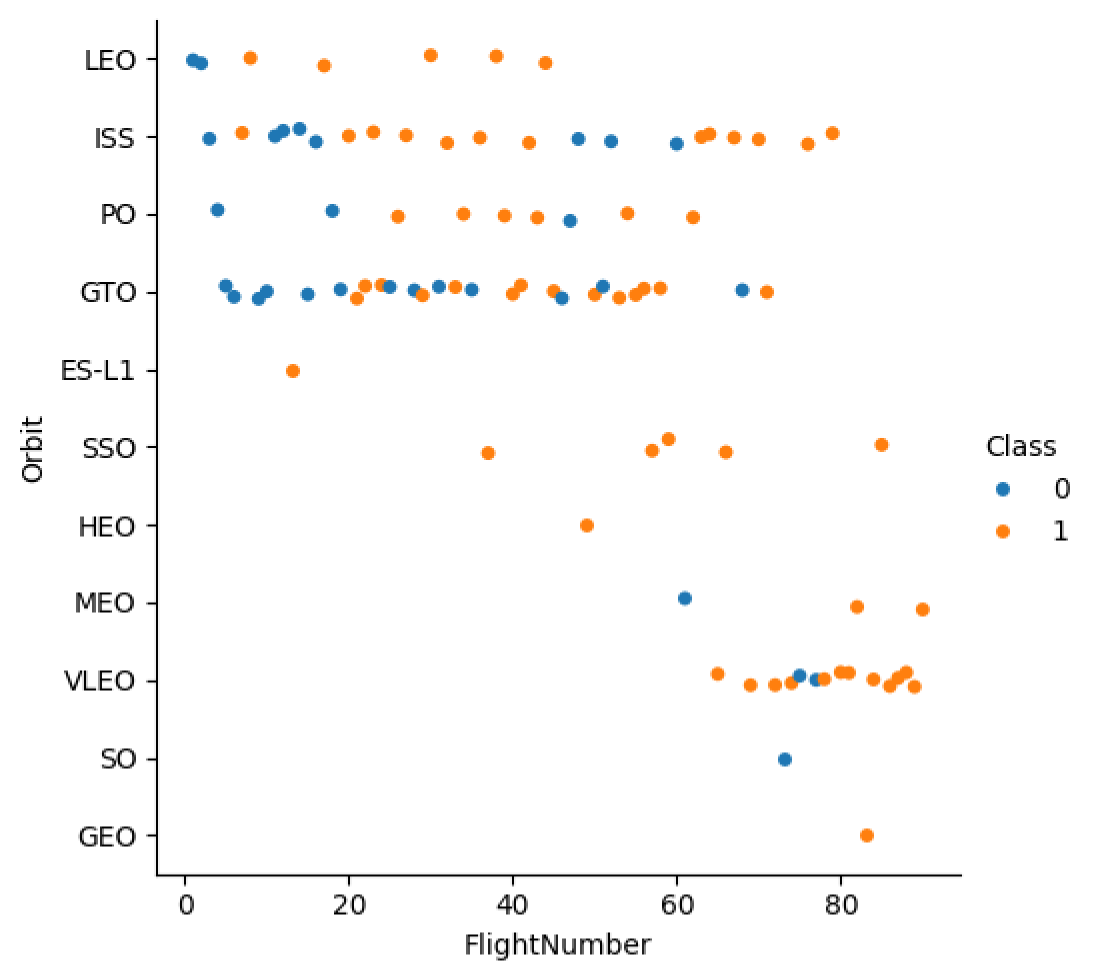
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have a 100% success rate
- GTO, ISS, LEO, MEO, and PO have a success rate that is between 50% - 80%
- SO has a 0% success rate



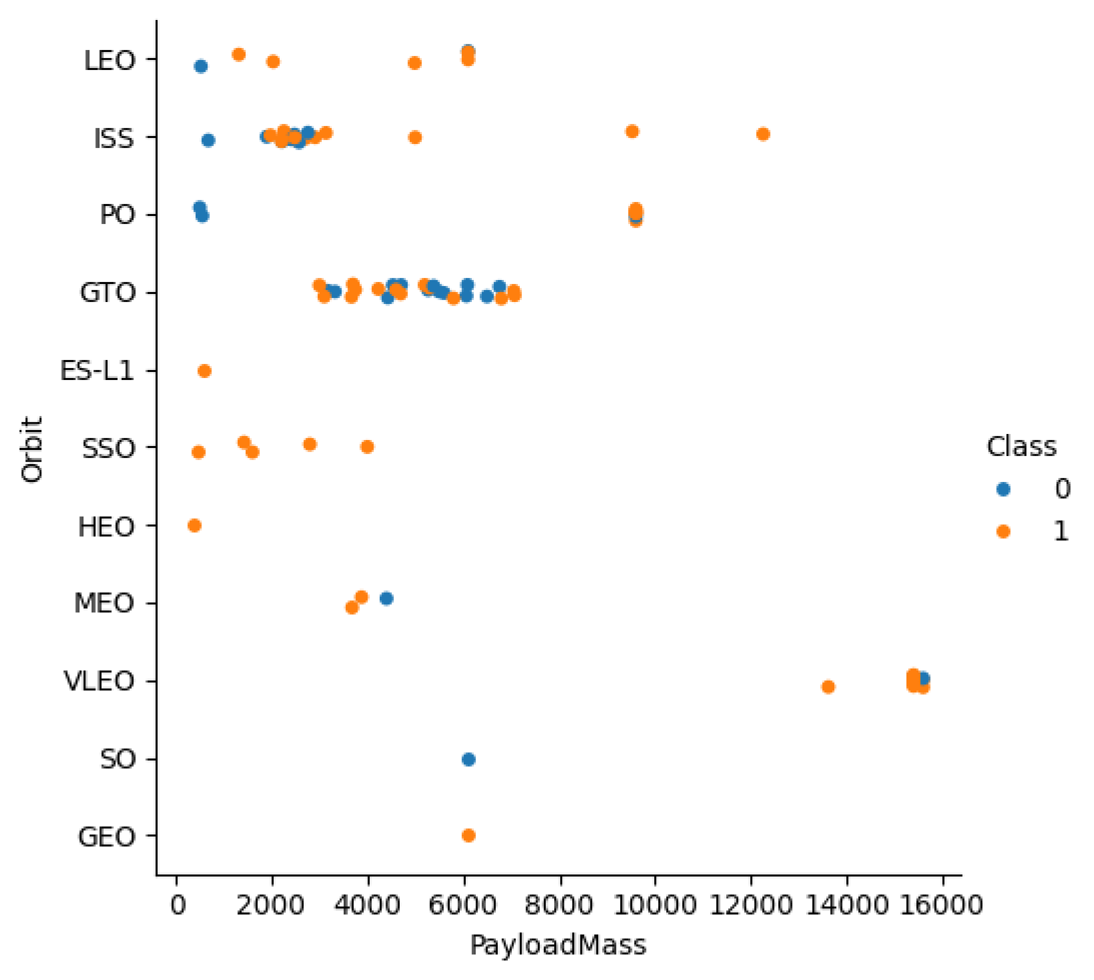
Flight Number vs. Orbit Type

- The success rate increases as the flight number increases
- An exception is GTO where this trend is not evident



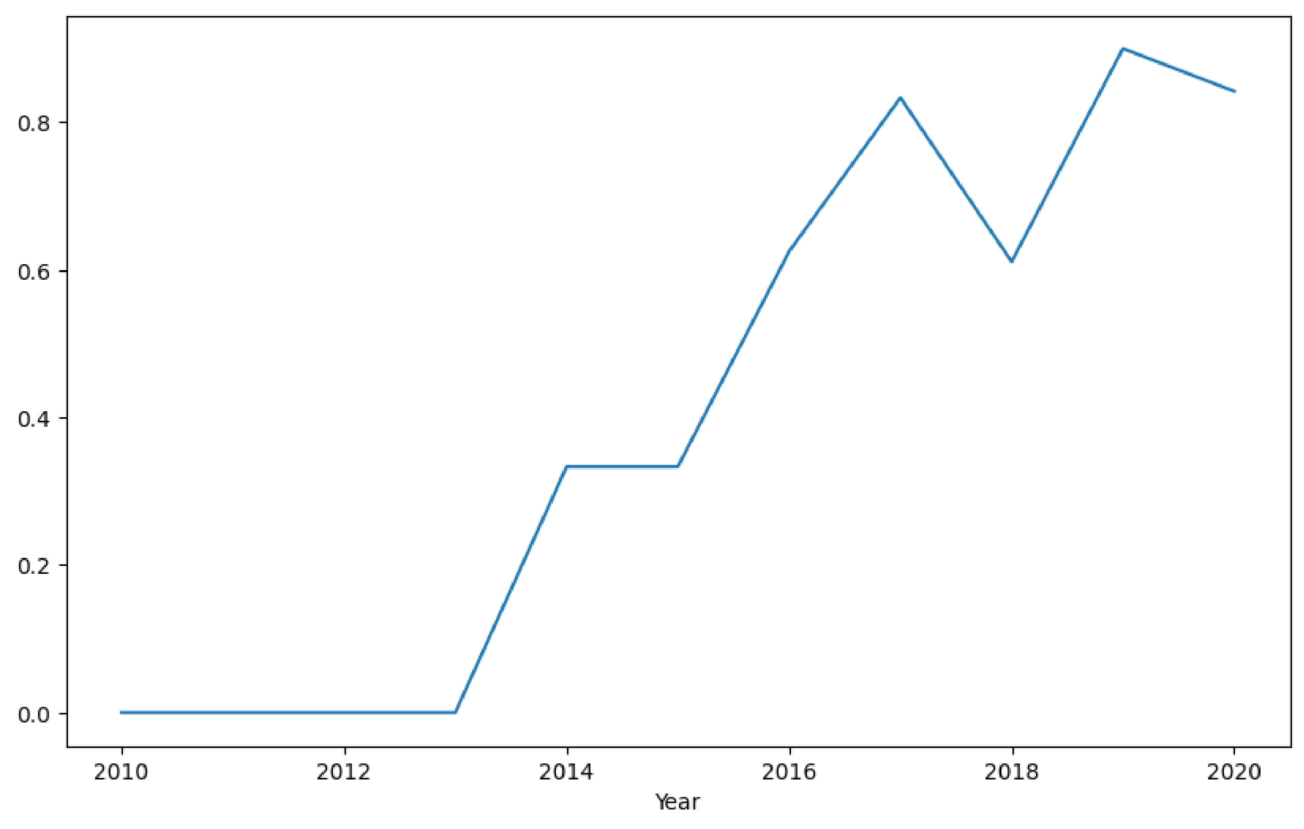
Payload vs. Orbit Type

- Higher payloads show greater success for LEO, ISS, and PO orbits
- GTO success is mixed at all displayed payloads



Launch Success Yearly Trend

- Overall the success rate improves over time
- The only time the success rate went down was during 2017 – 2018 and 2019 to 2020



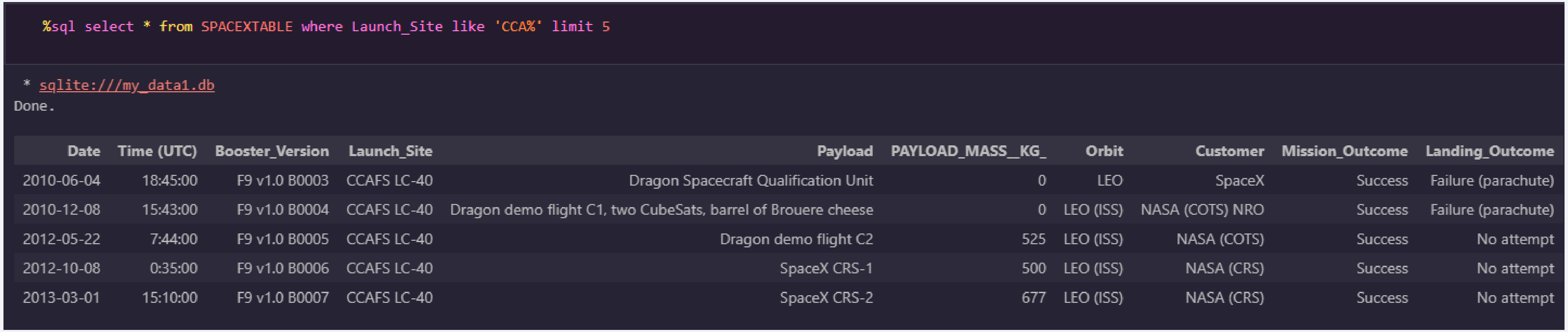
All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Launch Site Names Begin with 'CCA'

Launch sites:

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



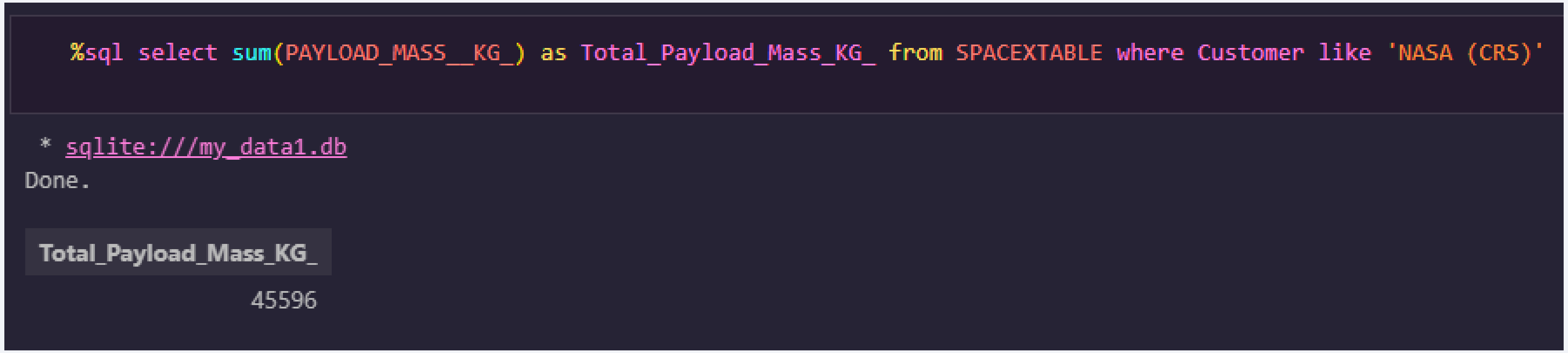
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

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

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

Payload Mass

Total payload mass

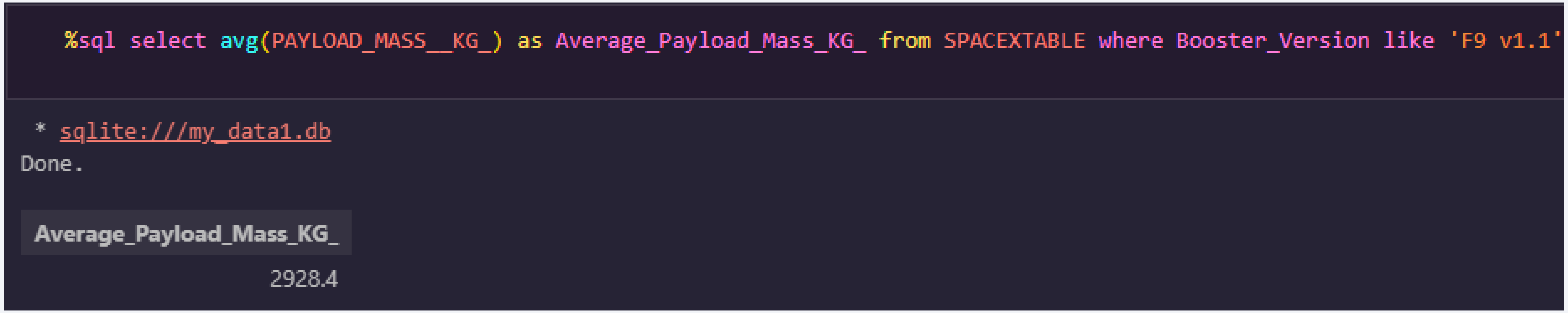


```
%sql select sum(PAYLOAD_MASS__KG_) as Total_Payload_Mass_KG_ from SPACEXTABLE where Customer like 'NASA (CRS)'
```

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

Total_Payload_Mass_KG_
45596

Average payload mass

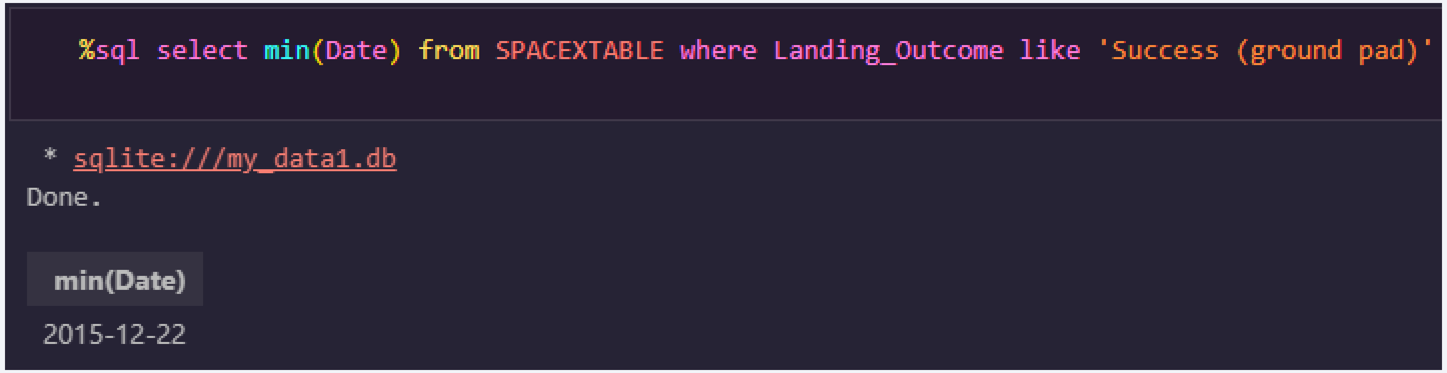


```
%sql select avg(PAYLOAD_MASS__KG_) as Average_Payload_Mass_KG_ from SPACEXTABLE where Booster_Version like 'F9 v1.1'
```

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

Average_Payload_Mass_KG_
2928.4

First Successful Ground Landing Date

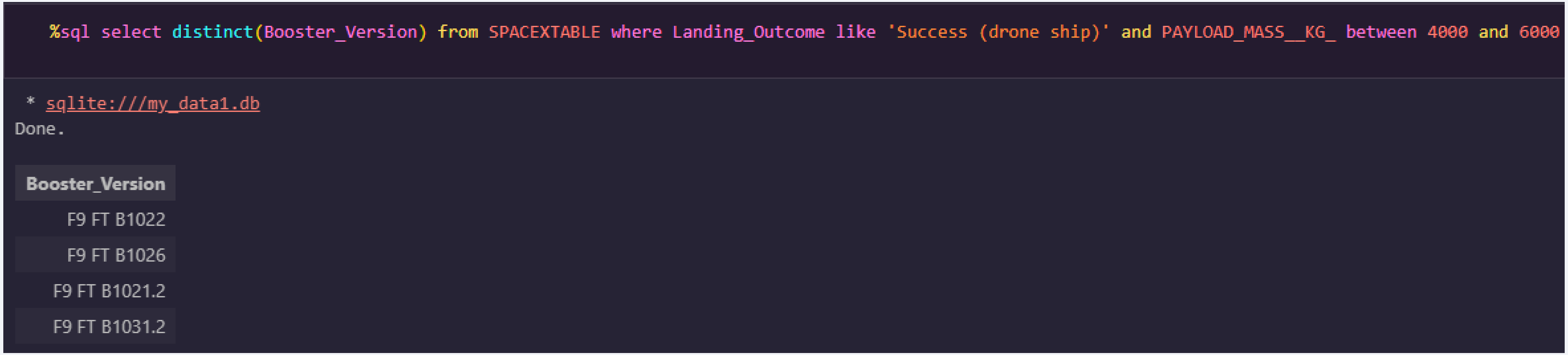


```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)'
```

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

min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

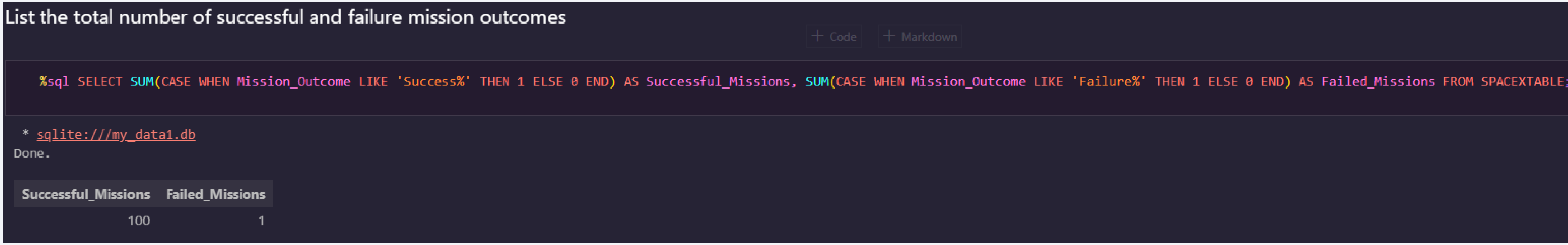


```
%sql select distinct(Booster_Version) from SPACEXTABLE where Landing_Outcome like 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * 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



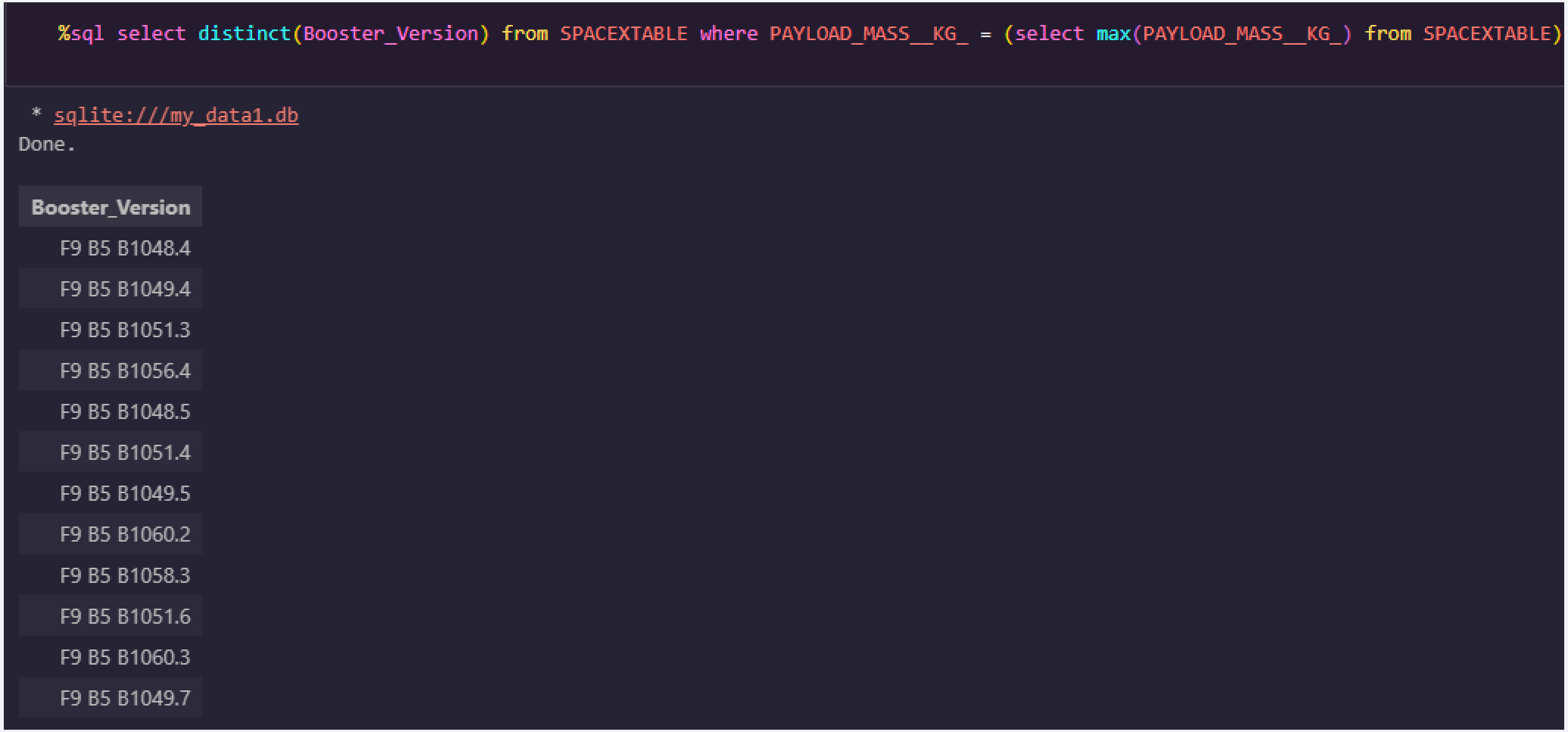
List the total number of successful and failure mission outcomes

```
%sql SELECT SUM(CASE WHEN Mission_Outcome LIKE 'Success%' THEN 1 ELSE 0 END) AS Successful_Missions, SUM(CASE WHEN Mission_Outcome LIKE 'Failure%' THEN 1 ELSE 0 END) AS Failed_Missions FROM SPACEXTABLE;
```

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

Successful_Missions	Failed_Missions
100	1

Boosters Carried Maximum Payload

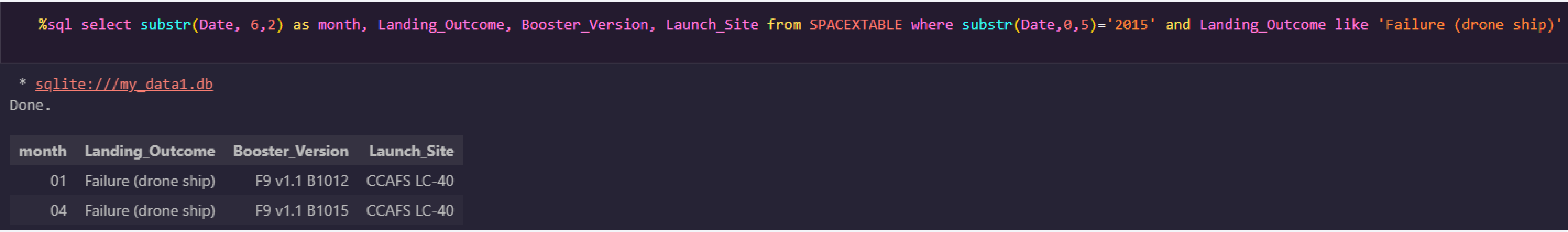


```
%sql select distinct(Booster_Version) from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * 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

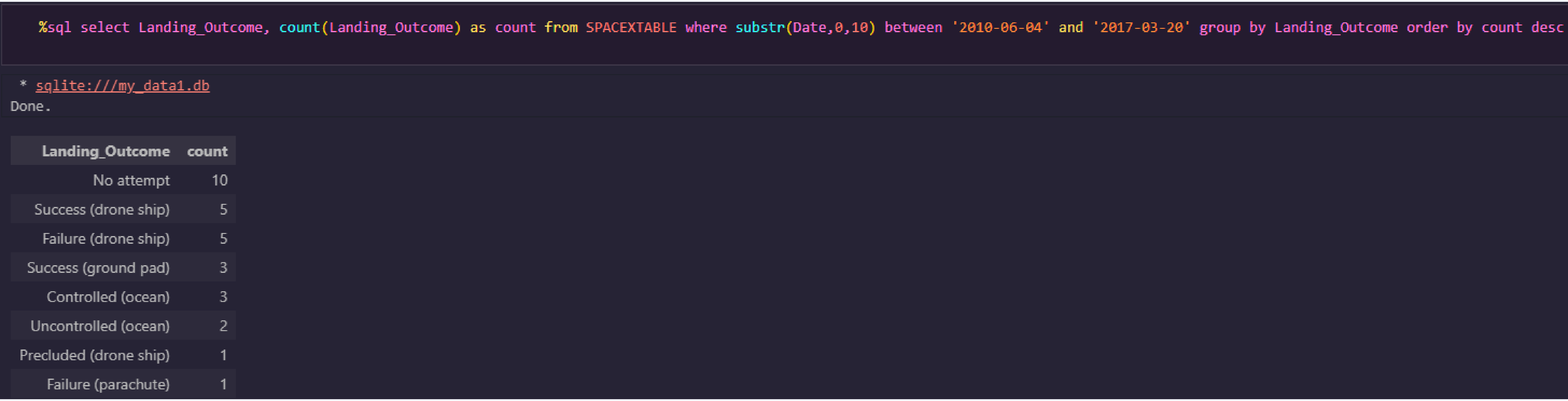


```
%sql select substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where substr(Date,0,5)='2015' and Landing_Outcome like 'Failure (drone ship)'
```

```
 * 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



```
%sql select Landing_Outcome, count(Landing_Outcome) as count from SPACEXTABLE where substr(Date,0,10) between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by count desc
```

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

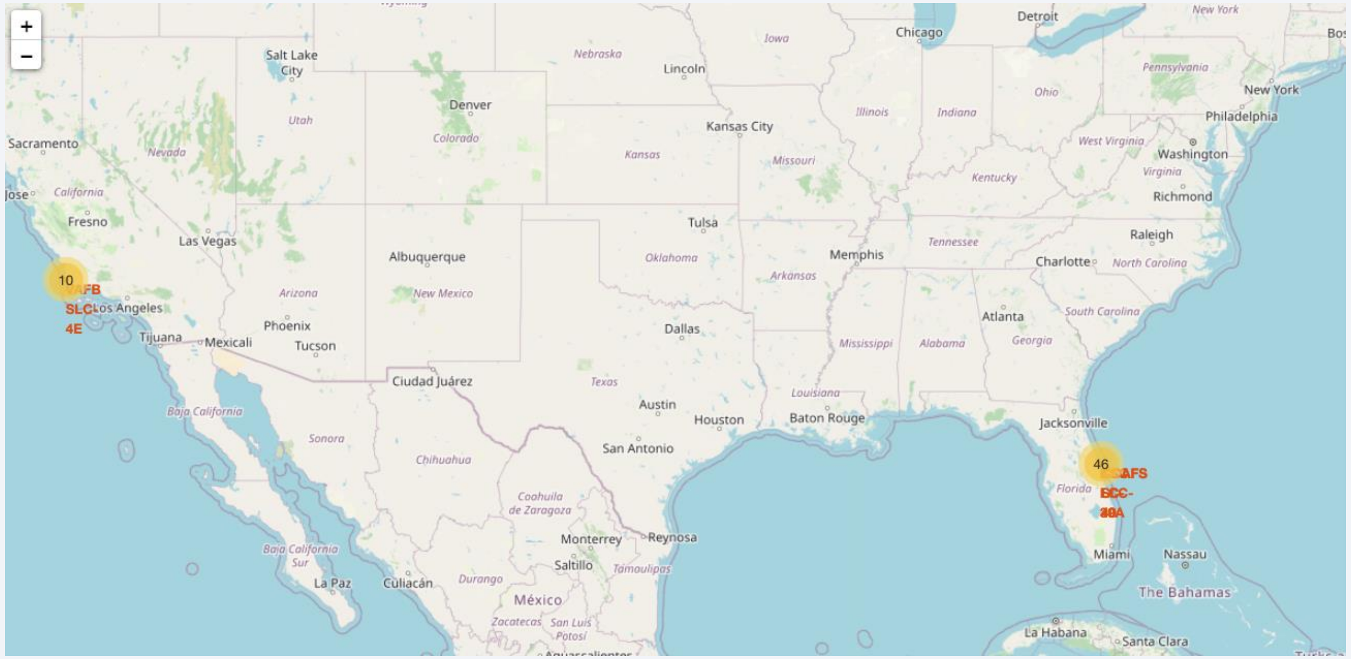
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites

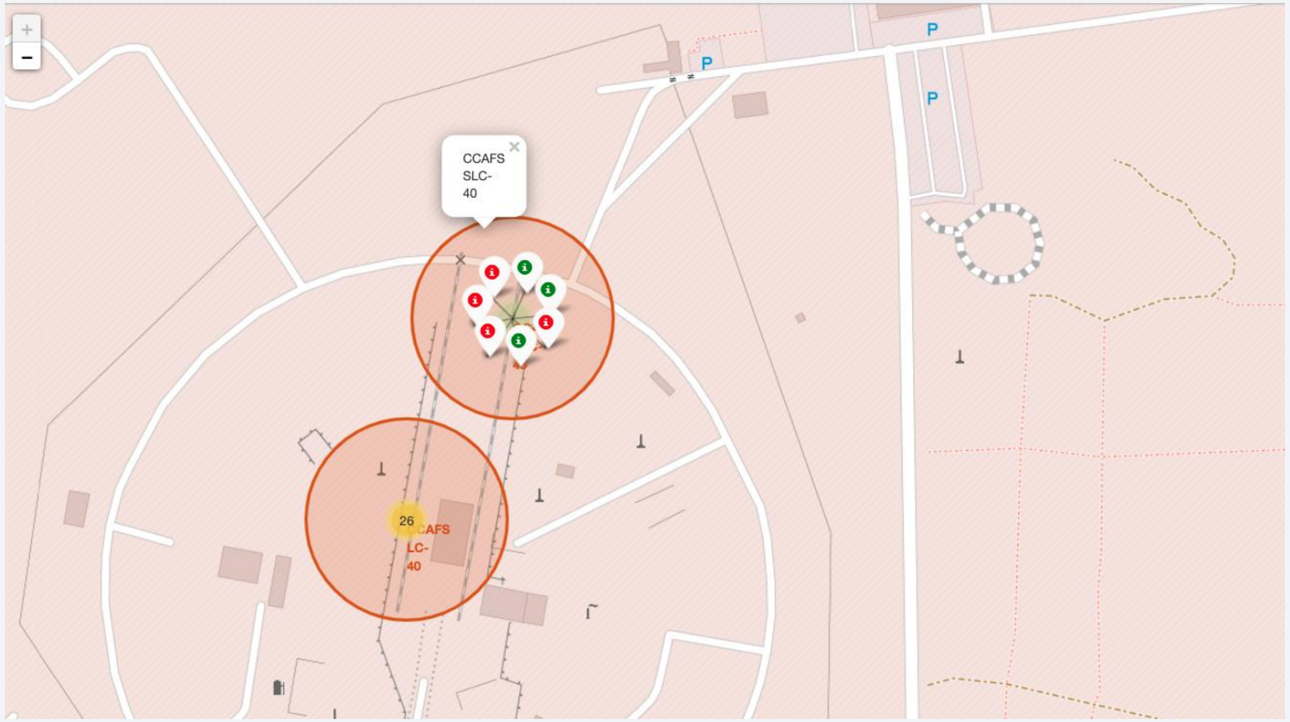
All launch sites are in the southern US in order to be close to the equator. This is because the rockets use the angular momentum of the earth's rotation to help with the launch. This saves the cost of loading additional fuel into the boosters



Launch Outcomes

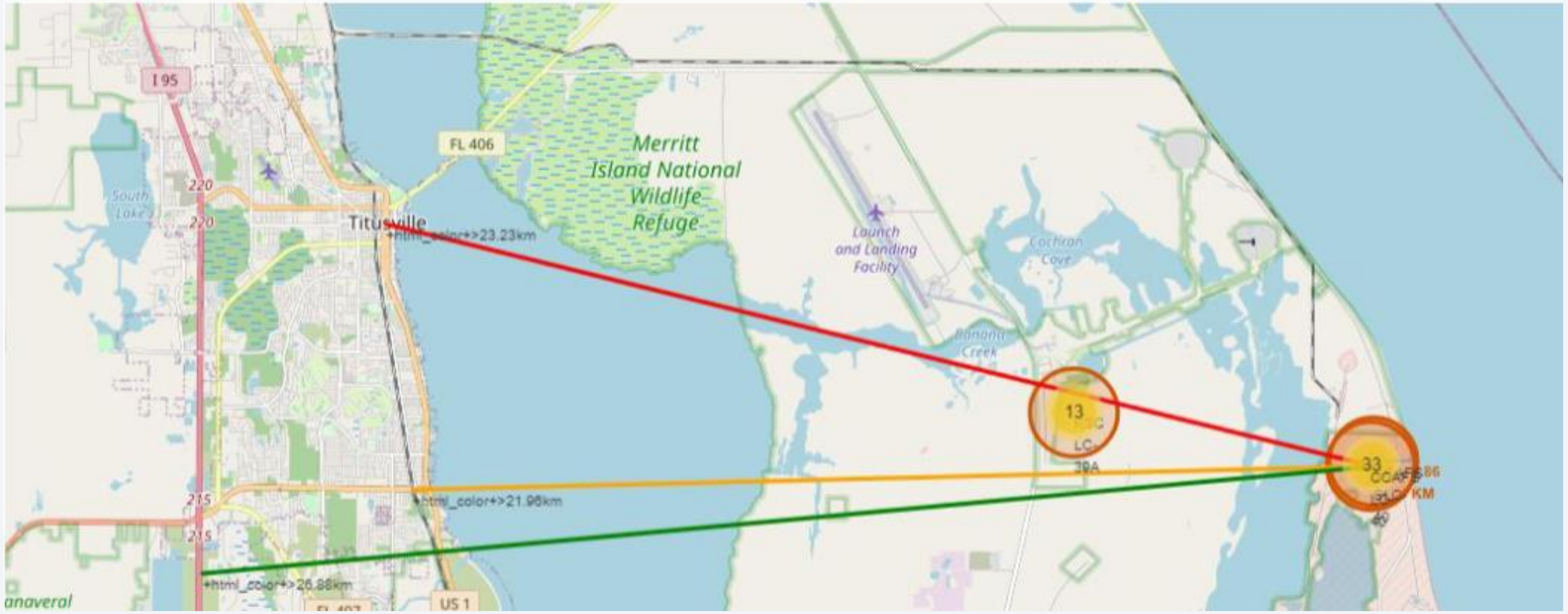
Launch outcome markers:

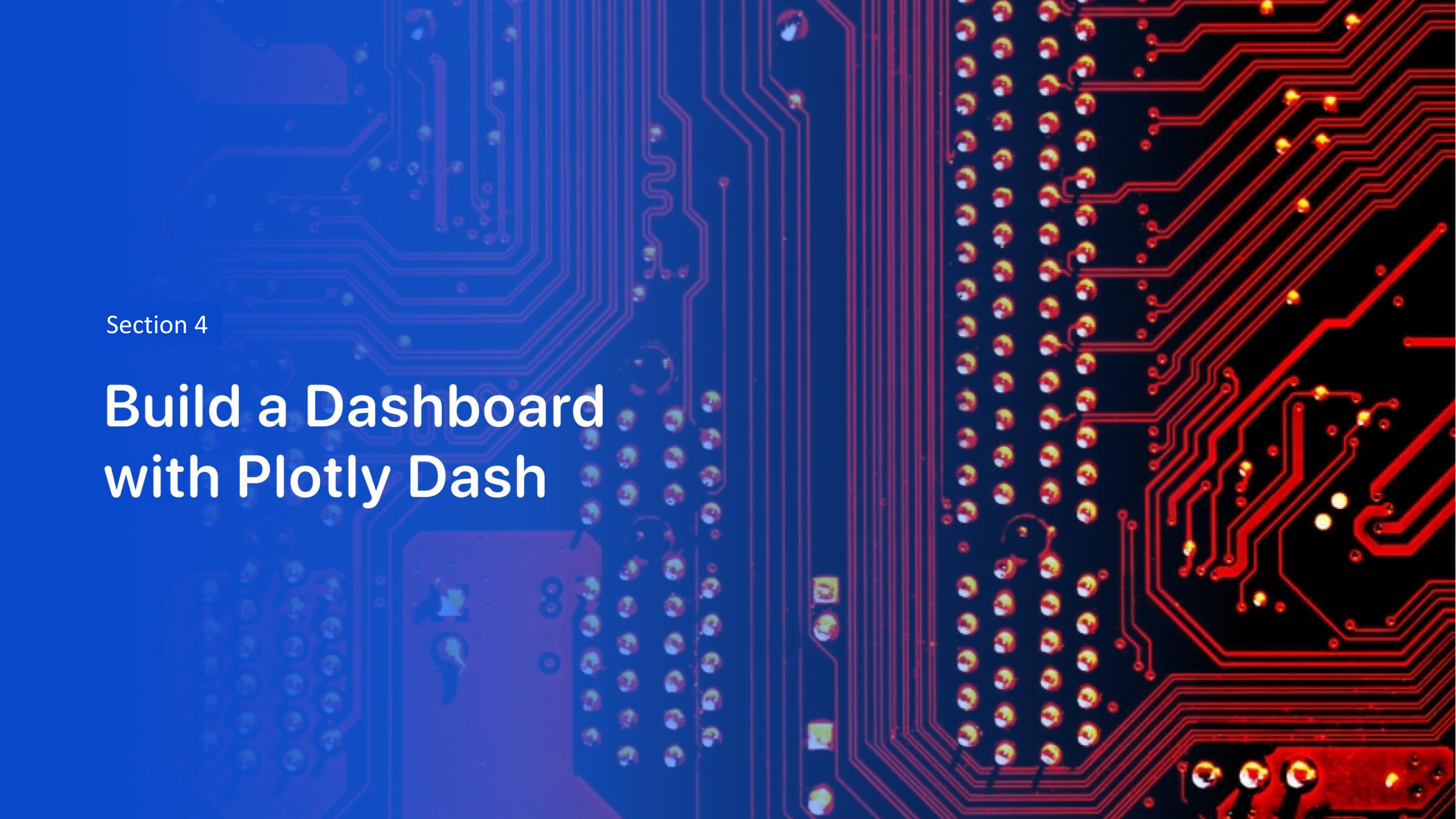
- Green – Successful launches
- Red – Failed launches



Distance to points of interest

Distance from nearest coastline: 0.86km
Distance from nearest railway: 21.96km
Distance from nearest city: 23.33km
Distance from nearest highway:26.88km



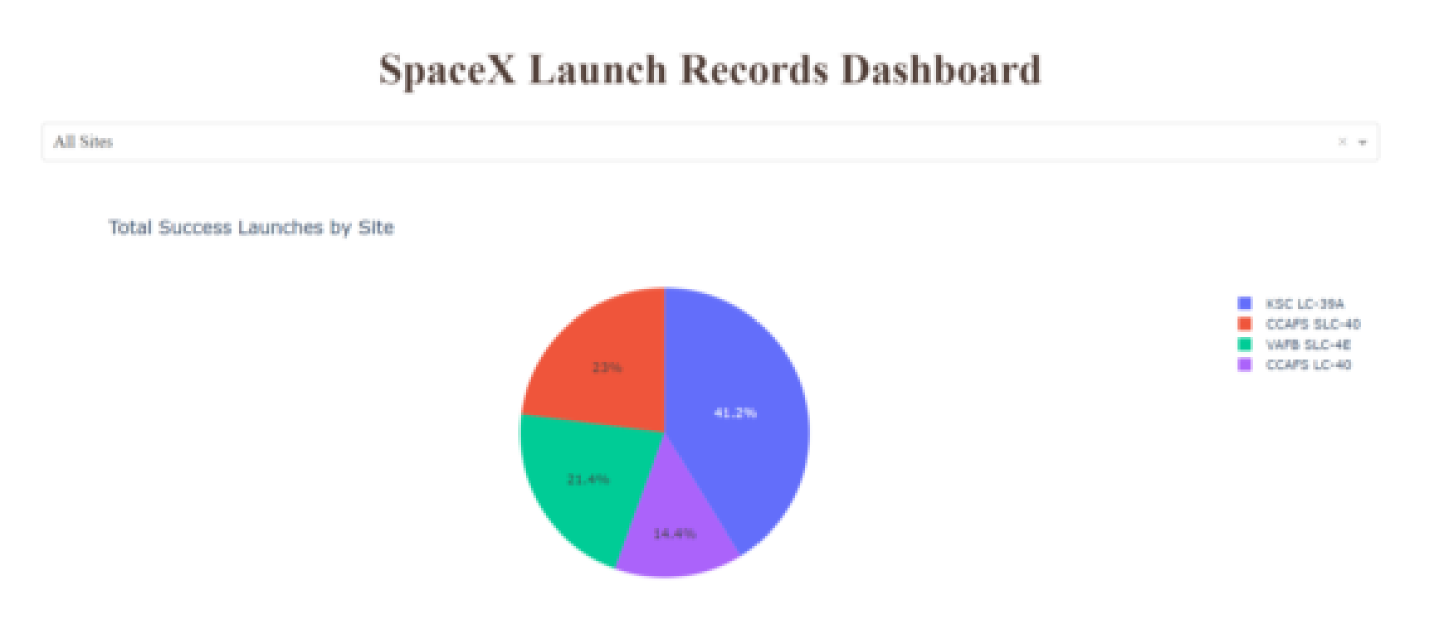


Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site

KSC LC-39A has the greatest number of successful launches



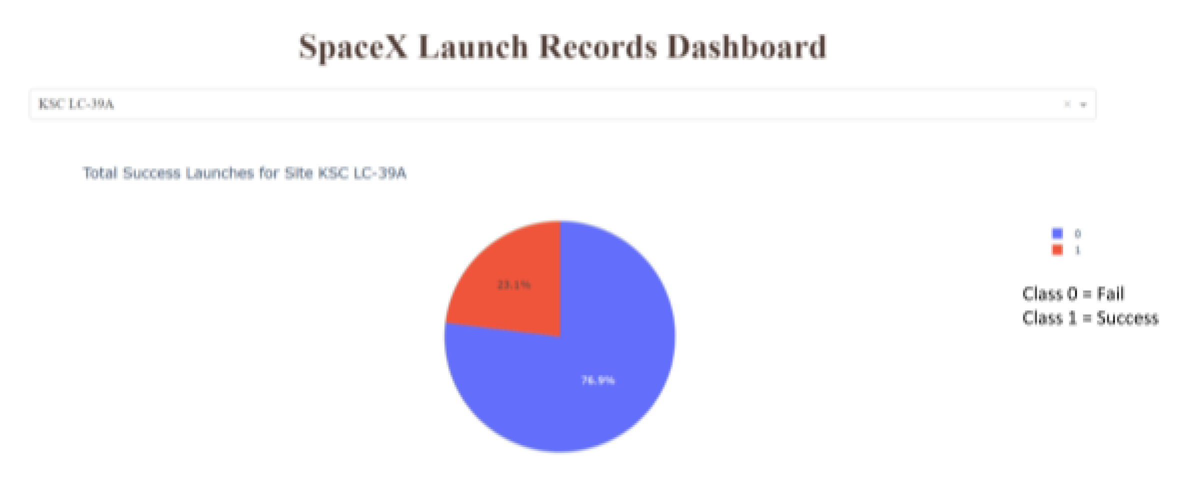
SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site

Launch Success of Launch site KSC LC-29A

KSC LC-29A has a launch success rate of 76.9%



SpaceX Launch Records Dashboard

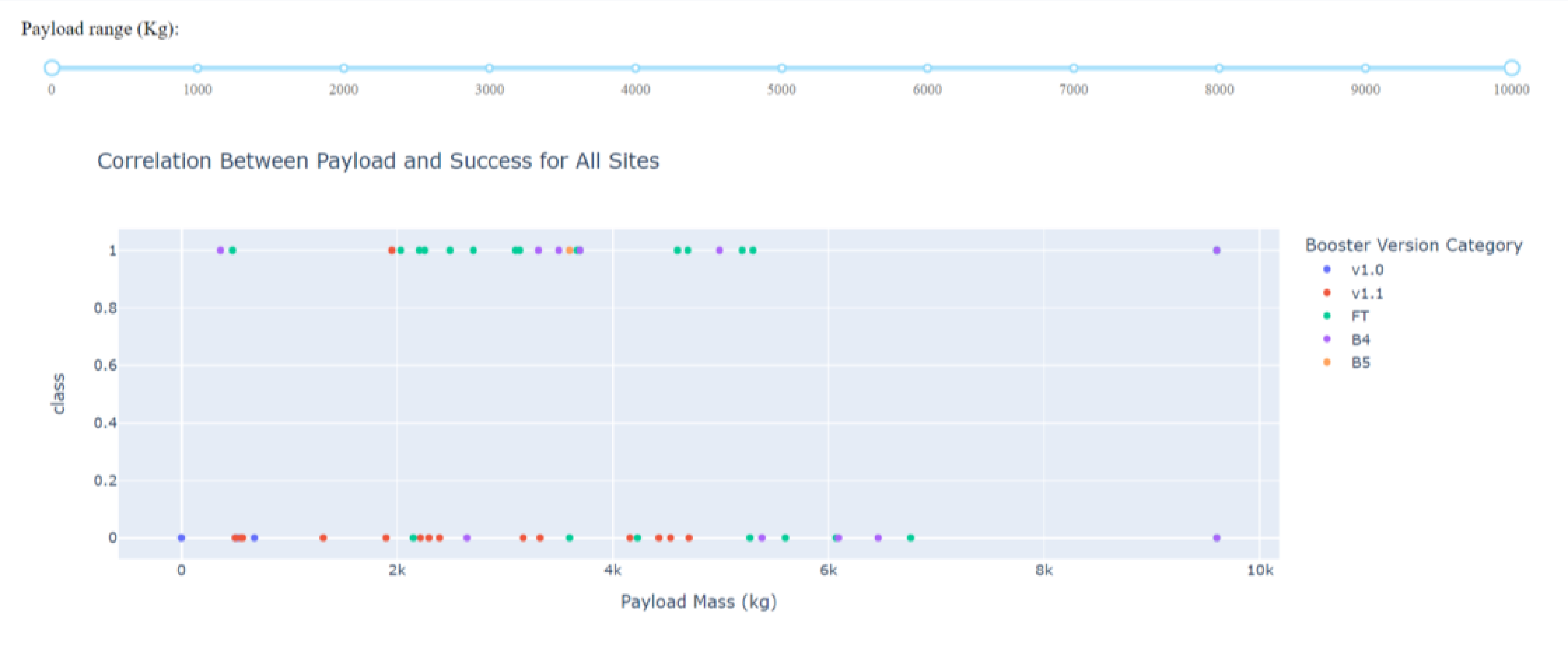
KSC LC-39A

Total Success Launches for Site KSC LC-39A

Class 0 = Fail
Class 1 = Success

Payload Mass in Relation to Launch Success

Payloads between 2000kg and 5000kg have the highest success rate

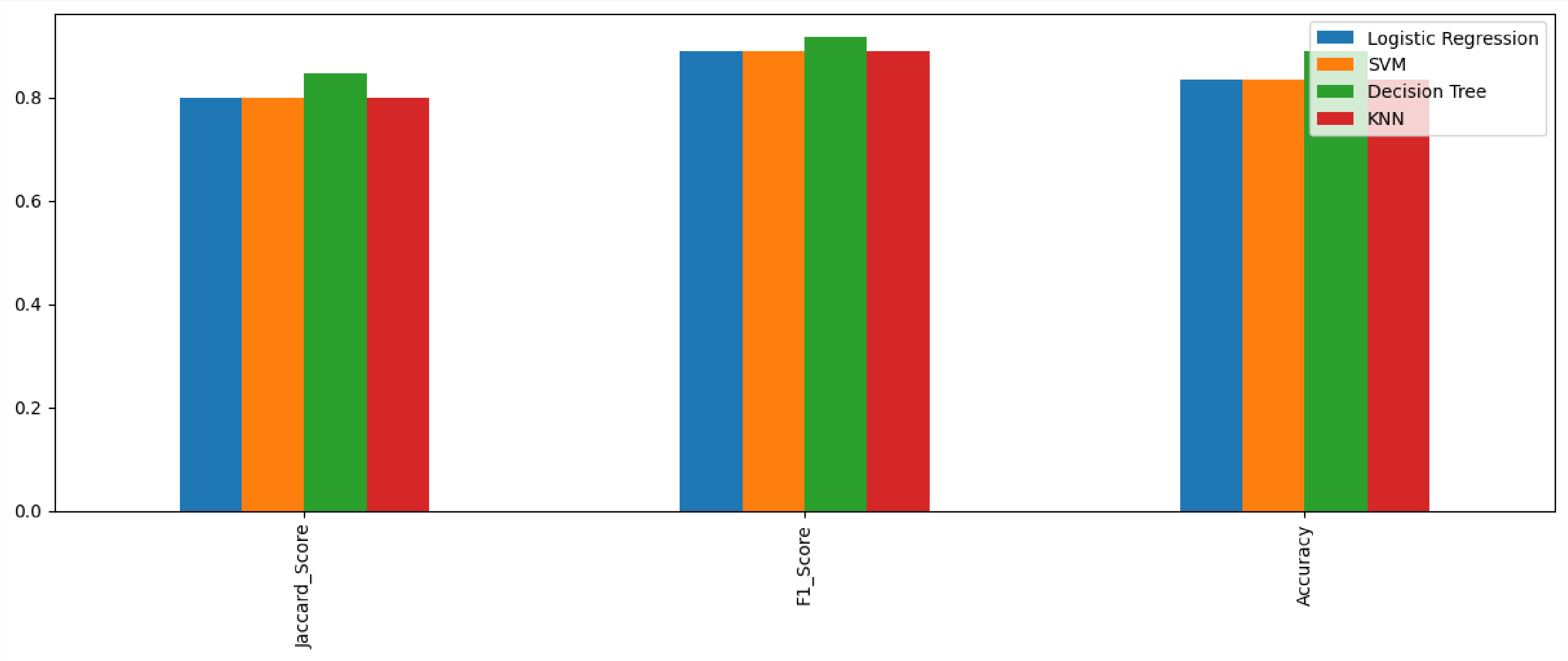


Section 5

Predictive Analysis (Classification)

Classification Accuracy

The accuracy scores are similar for all models however the decision tree model seems to have a marginally higher score



Confusion Matrix

The confusion matrix of the decision tree model is like all the other models even though it is the best performing one.

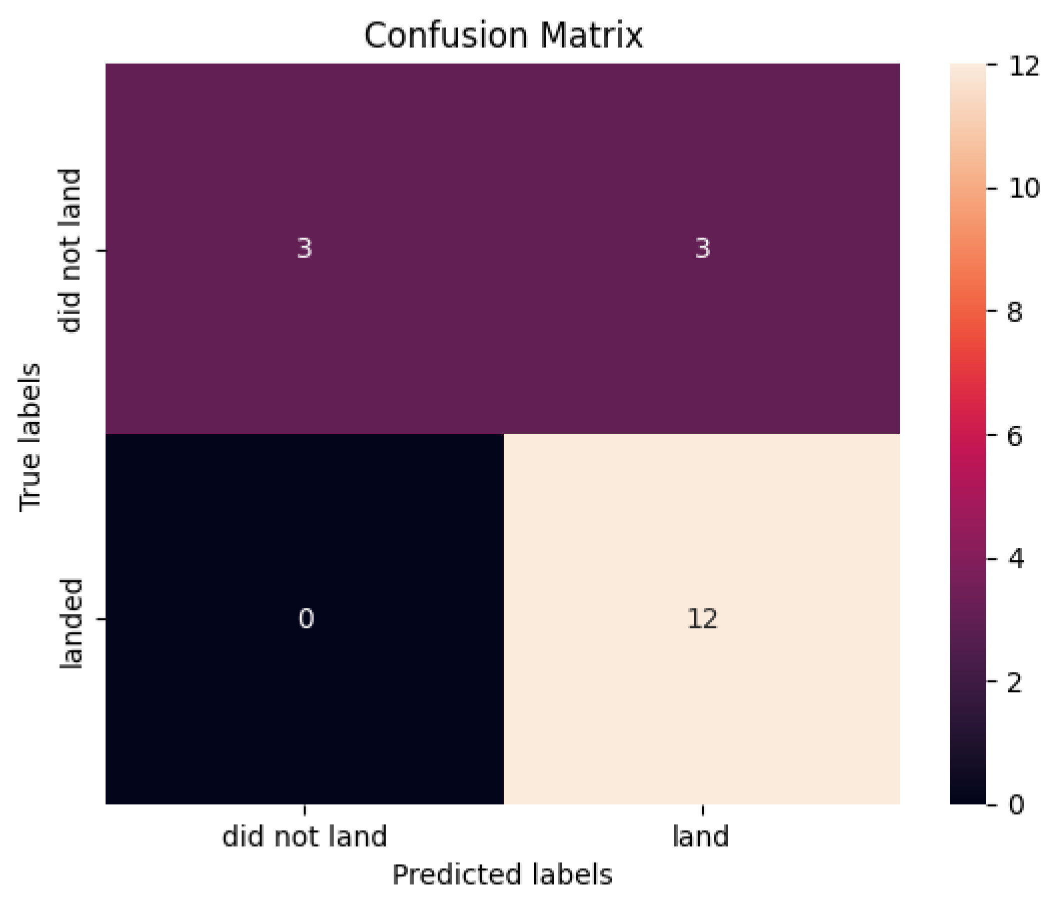
Confusion matrix outputs:

TP: 12

TN: 3

FP: 3

FN: 0



Conclusions

Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming

Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters

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 100% success rate for launches less than 5,500 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

Thank you!