IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Rosemary Medige
29/08/2023

# Outline

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

# Executive Summary

In this capstone project, we will predict if the Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This will be achieved using different machine learning algorithms.

The Methodologies followed will include Data Collection, Data Wrangling, Exploratory Data Analysis, Data Visualization and Machine Learning.

During our investigation, the results of our analysis indicate that there are some features of rocket launches that have correlation with Success and Failure launches.

In the end we conclude that the Decision Tree may be the best Machine Learning algorithm for this problem.

# Introduction

The main goal of this project is to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

This brings us to the main question that we are trying to answer: For a given set of features about a Falcon9 rocket launch, will the first stage of rocket land successfully?

Section 1

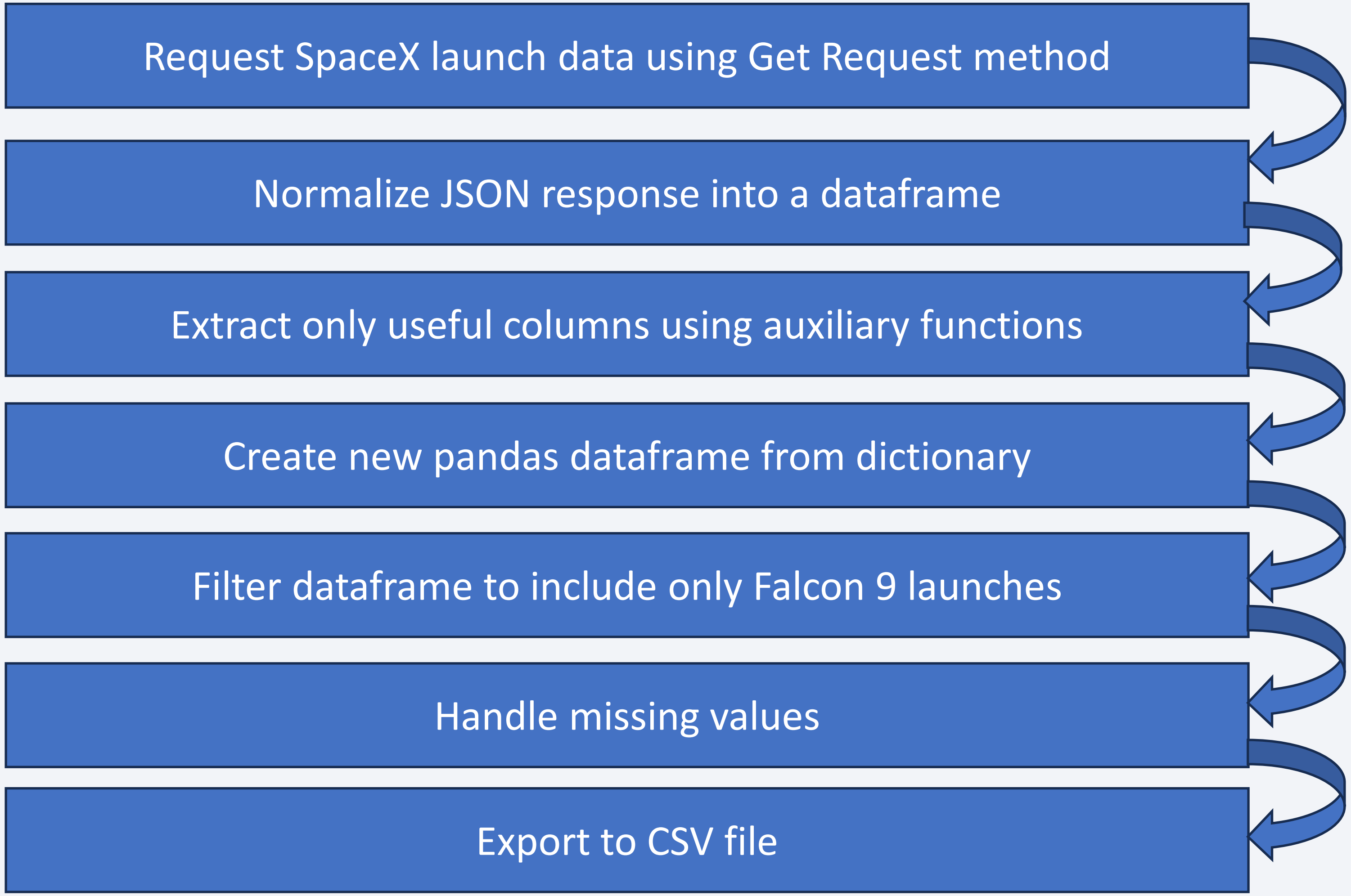
# Methodology

# Methodology

## Executive Summary

- Data was collected through 2 methods: by requesting data from the SpaceX API and other way was web scraping launch data from Wikipedia.
- Data wrangling was performed after data collection. By using Python’s Pandas library, we transform and clean the data.
- Then we performed Exploratory Data analysis using Visualization tools such matplotlib and seaborn libraries. We also analyzed data using SQL queries. For Creating maps, we used Folium and for creating interactive data visualizations we used Plotly Dash.
- We used four different Machine learning algorithms for predictive analysis. They are logistic regression, support vector machines, k-nearest neighbour and decision tree classifier. Each model was trained, tuned and evaluated to find the best one.

# Data Collection



Request SpaceX launch data using Get Request method

Normalize JSON response into a dataframe

Extract only useful columns using auxiliary functions

Create new pandas dataframe from dictionary

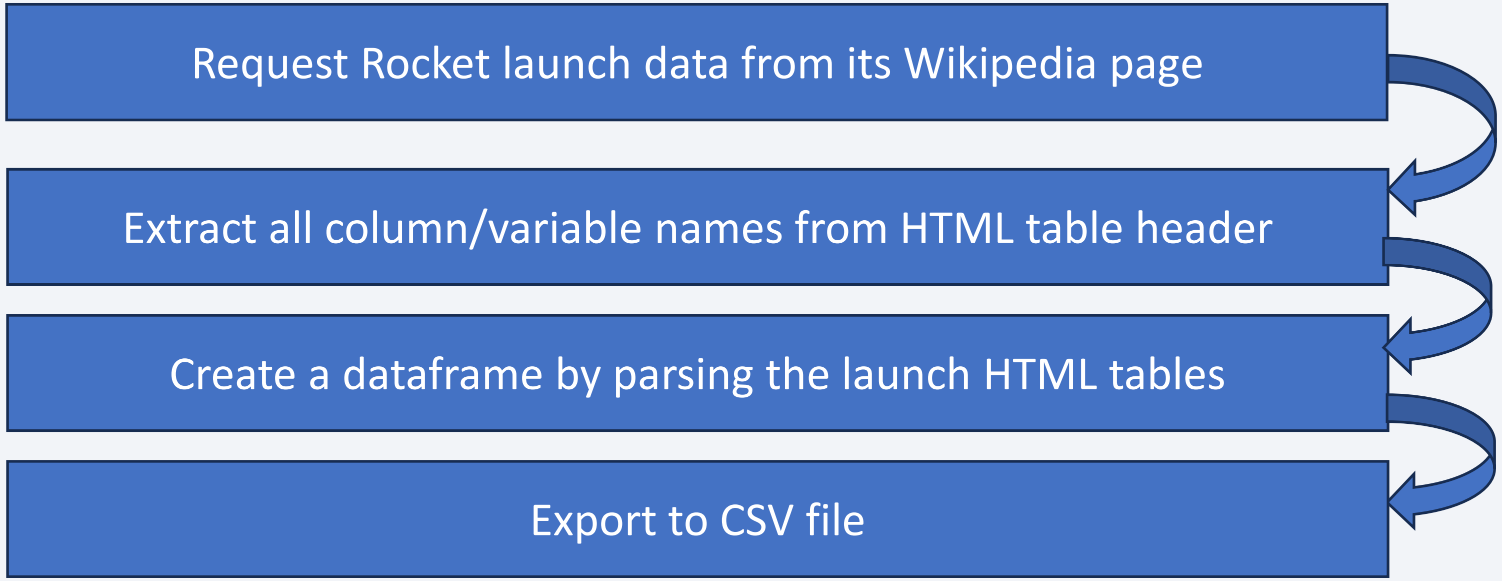
Filter dataframe to include only Falcon 9 launches

Handle missing values

Export to CSV file

GitHub link:
DataCollection

# Data Collection - Scraping



Request Rocket launch data from its Wikipedia page

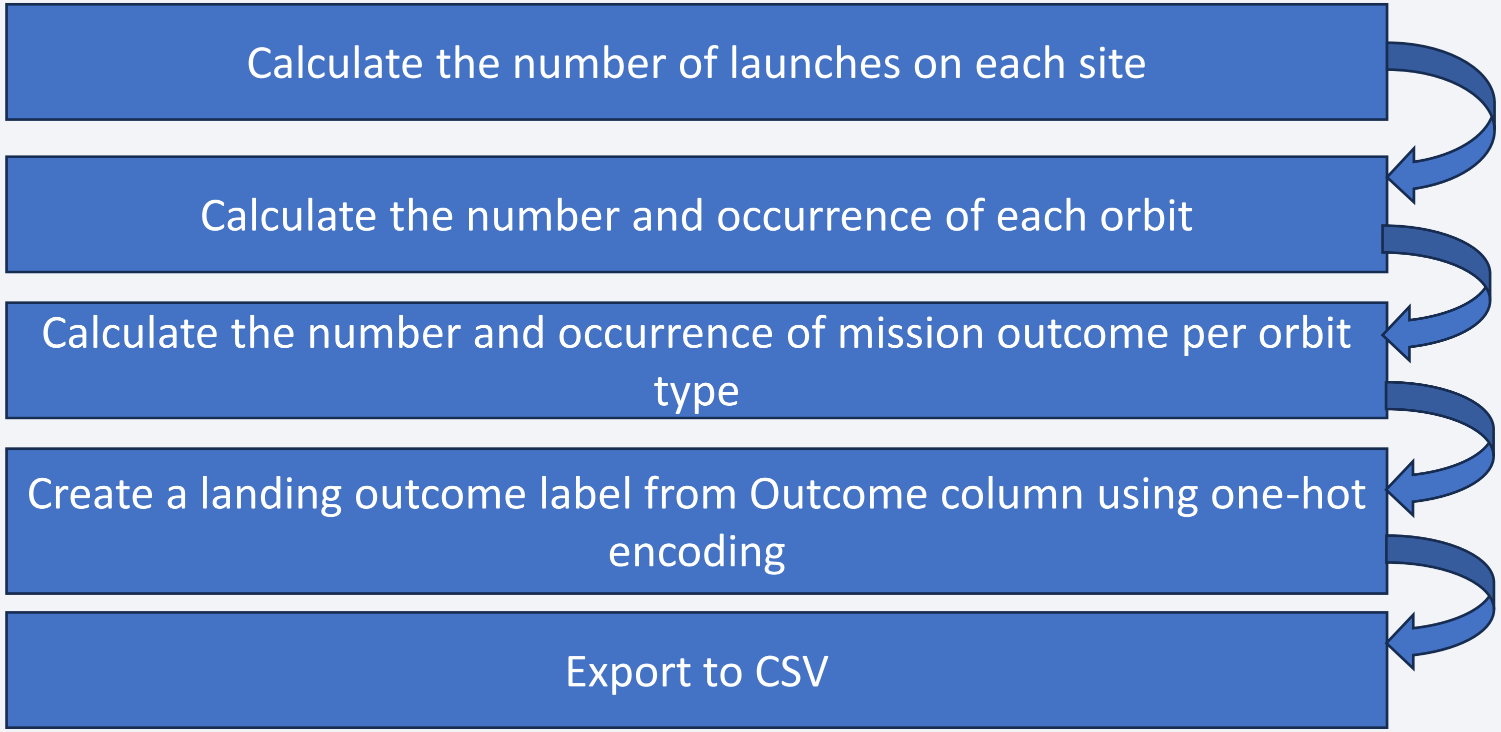
Extract all column/variable names from HTML table header

Create a dataframe by parsing the launch HTML tables

Export to CSV file

GitHub link:
Webscraping

# Data Wrangling



Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column using one-hot encoding

Export to CSV

GitHub link:
Data Wrangling

# EDA with Data Visualization

- Scatter plots: These are used to represent the relationship between two variables. Here using scatter plots, we compare different features such as Flight number vs launch site, Payload vs launch site, Flight number vs orbit type and Payload vs orbit type
- Bar charts: These are used to make it easy to compare values between multiple groups at a time. Bar charts were used to compare the Success Rate for different Orbit types
- Line charts: These are helpful for showing data trends over time. A line chart was used to show Success Rate over certain number of Years.

**GitHub link**:Data Visualization

# EDA with SQL

- SQL queries performed on the dataset are listed below:

- 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
- To List the date when the first succesful landing outcome in ground pad was acheived.
- To List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- To List the total number of successful and failure mission outcomes
- To List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- To 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.

**GitHub link**:
EDA with SQL

# Build an Interactive Map with Folium

- Objects were created and added to Folium map. Marker objects were used to show all launch sites on a map as well as the success/fail launches for each site on the map. Line objects were used to calculate the distances between a launch site and its proximities.
- By adding these objects, following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Are launch sites kept away from cities? Yes

**GitHub link**: Interactive Map with Folium

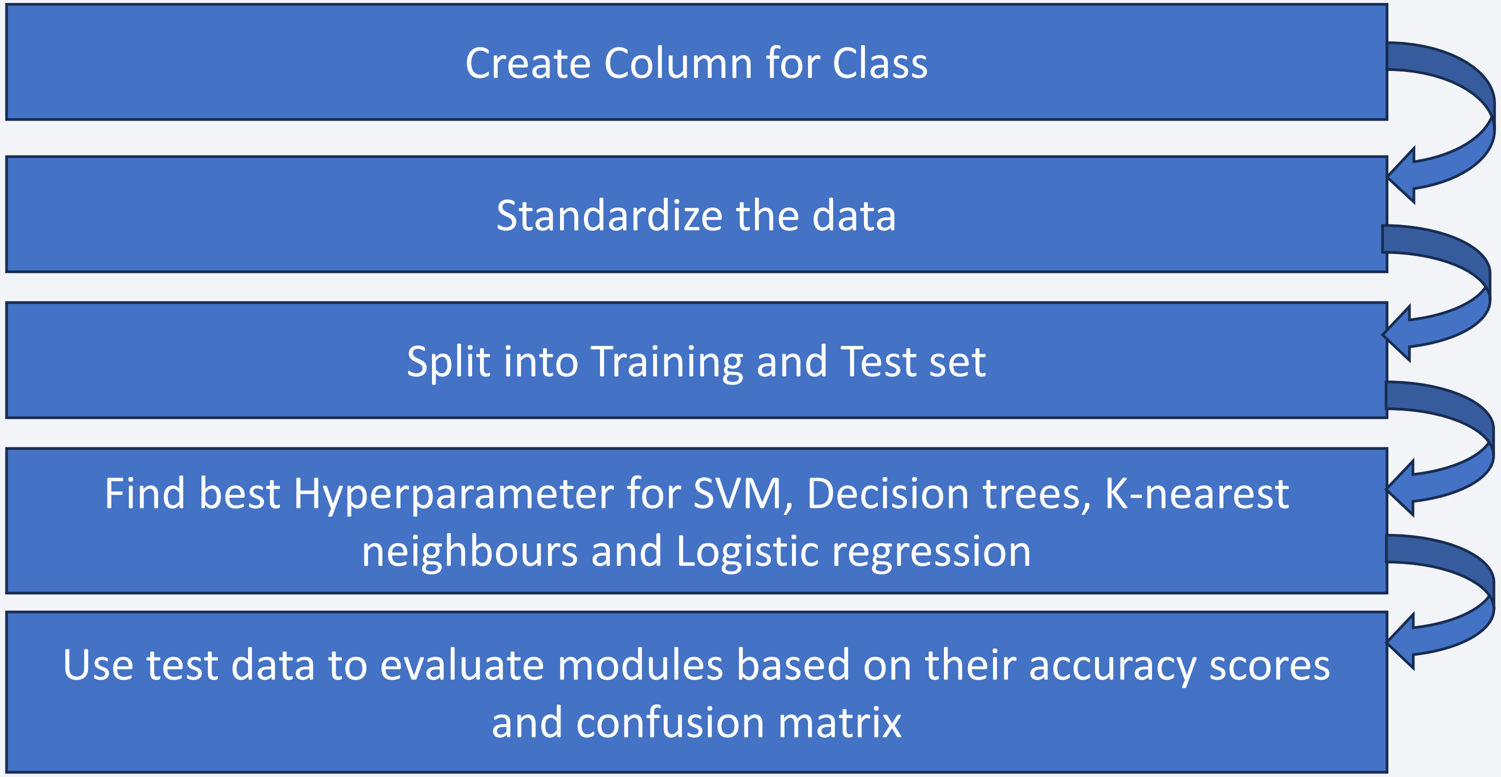
# Build a Dashboard with Plotly Dash

- The Dashboard contains two charts:

1. A pie chart which shows successful launch by each site. This chart shows the distribution of landing outcomes across all launch sites or show the success rate of launches on individual sites.
2. A scatter chart shows the relationship between landing outcomes and payload mass of different boosters. The dashboard takes two inputs, sites and payload mass.

**GitHub link**: Dashboard

# Predictive Analysis (Classification)



Create Column for Class

Standardize the data

Split into Training and Test set

Find best Hyperparameter for SVM, Decision trees, K-nearest neighbours and Logistic regression

Use test data to evaluate modules based on their accuracy scores and confusion matrix

**GitHub link**:
Predictive Analysis

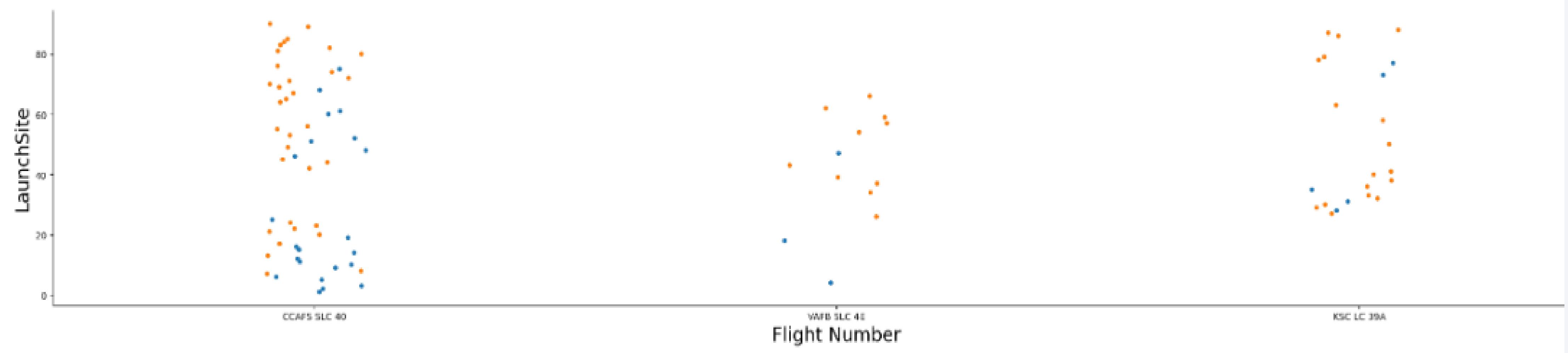
# Results

- Exploratory data analysis results: The success rate of Falcon9 landings was 66%.
- Predictive analysis results: The Decision tree algorithm was the best classification method with an accuracy of 88%

Section 2

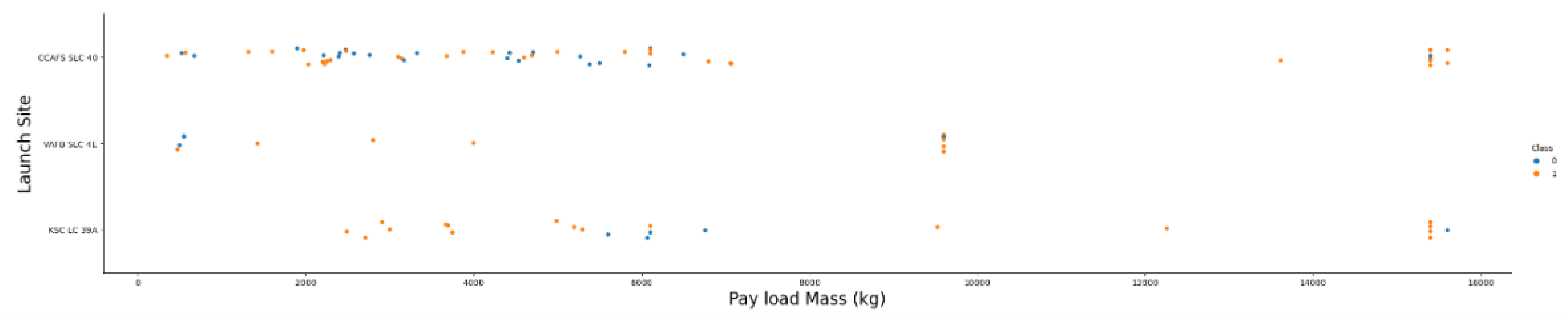
# Insights drawn from EDA

# Flight Number vs. Launch Site



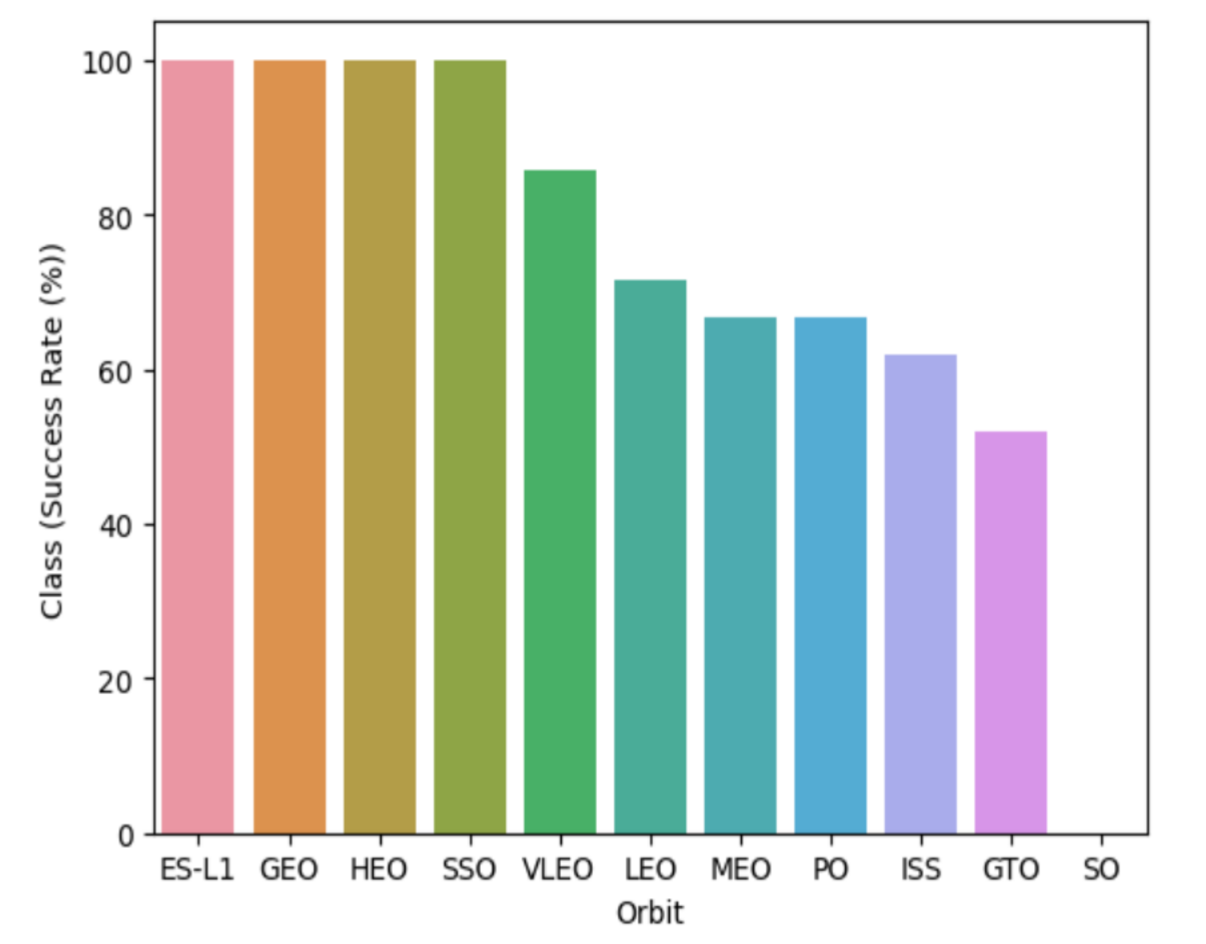
- This figure shows that success rate increased as the number of flights increased

# Payload vs. Launch Site



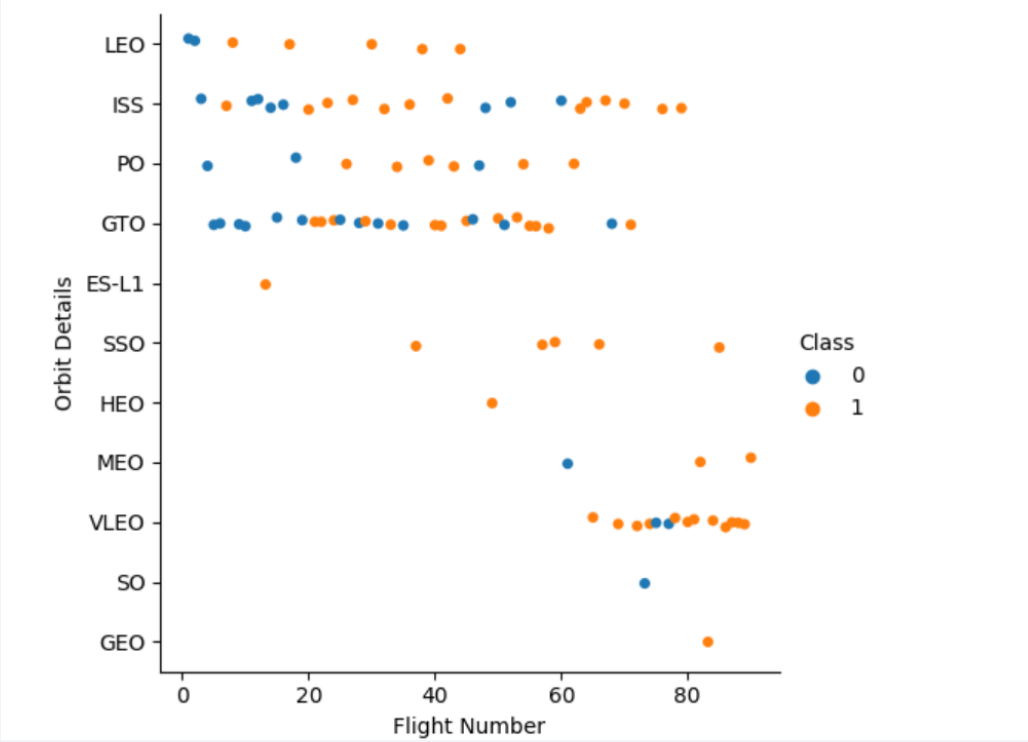
- There seems to be weak correlation between payload and launch site and so decisions cannot be made using this.

# Success Rate vs. Orbit Type



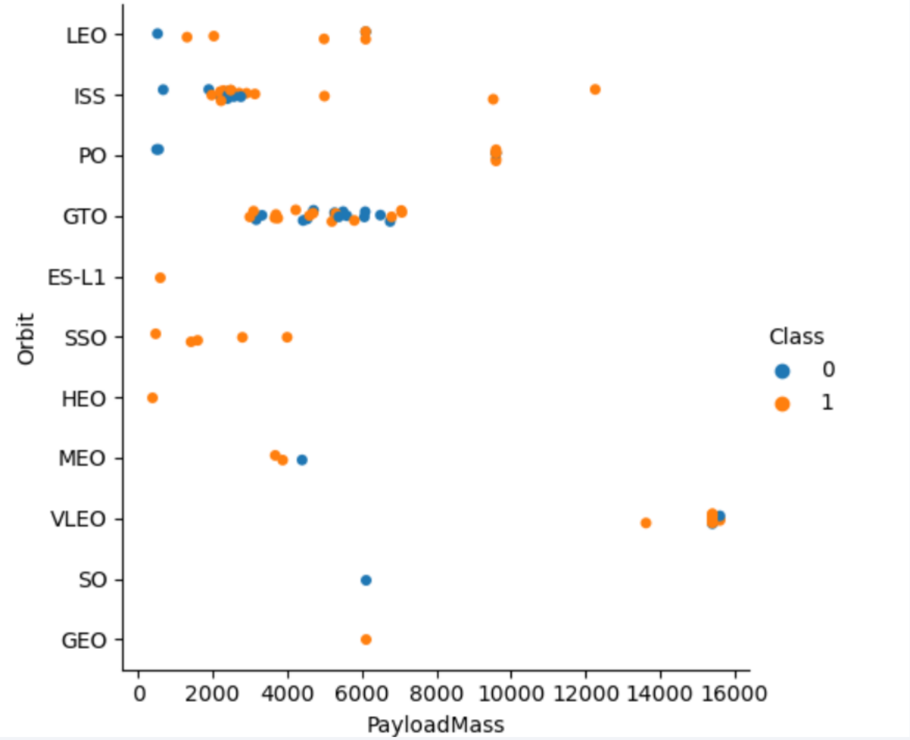
- Orbits ES-L1, GEO, HEO and SSO have 100% success rates.

# Flight Number vs. Orbit Type



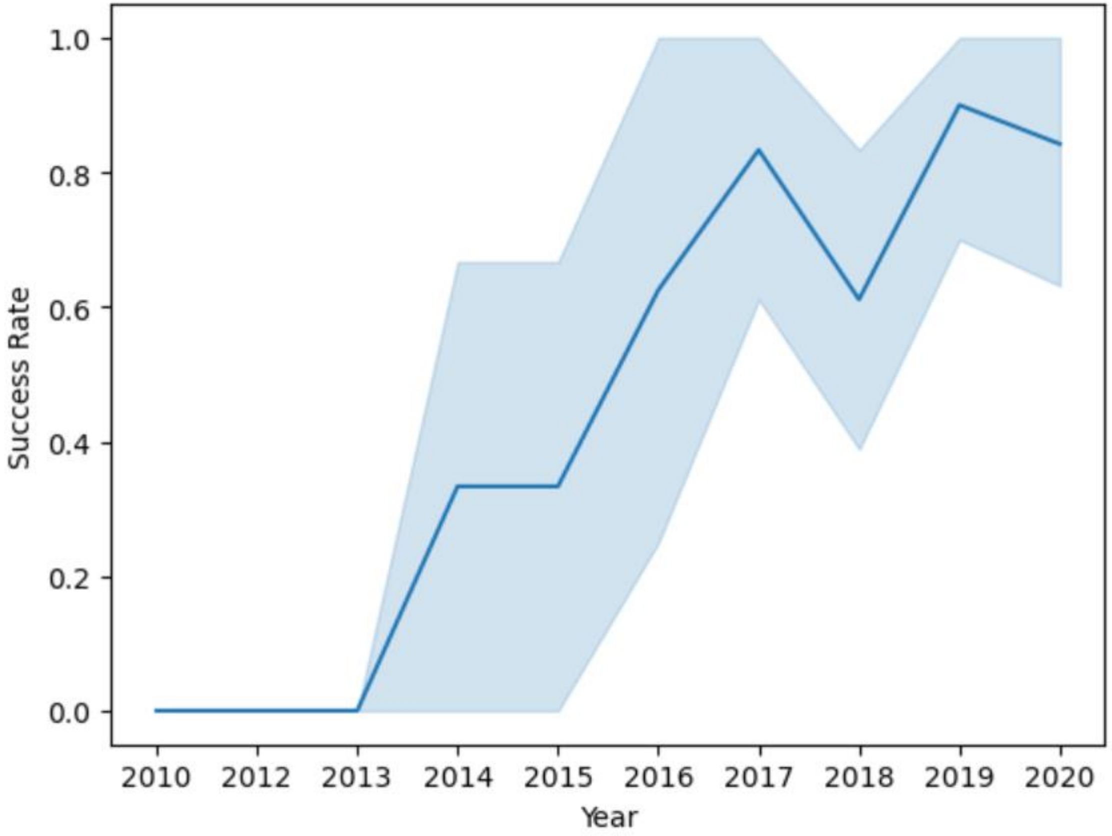
- There seems to be no correlation in GTO orbit
- In LEO orbit, the success is positively correlated with number of flights
- The SSO orbit has 100% success rate however with fewer flights than other orbits
- Flight numbers greater than 40 have high success rate

# Payload vs. Orbit Type



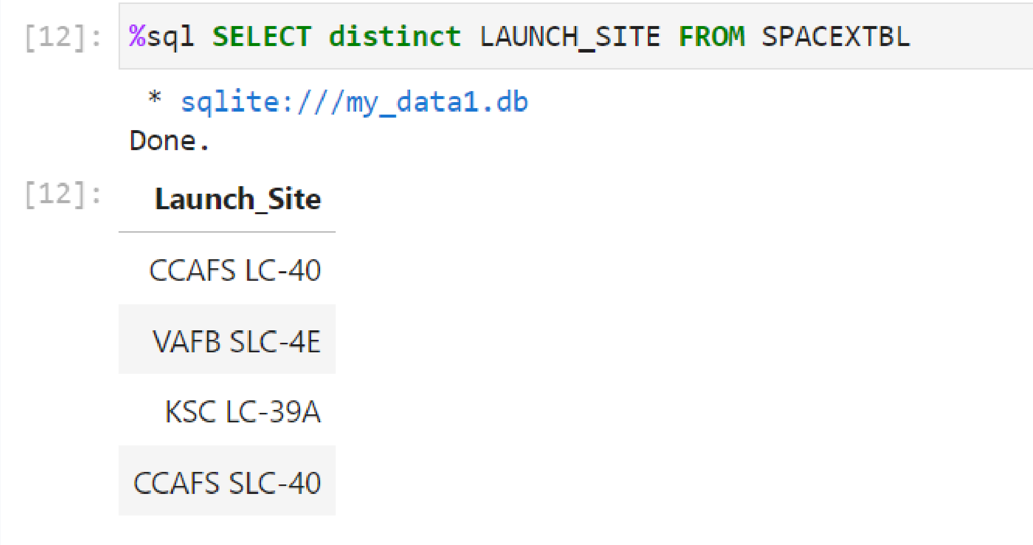
- There is no correlation in GTO orbit
- In LEO orbit, the success is positively correlated with number of flights

# Year vs. Success Rate



- The Success rate increased from 2013 and had its peak at 2019.

# Unique Launch Sites in Space mission



```
[12]: %sql SELECT distinct LAUNCH_SITE FROM SPACEXTBL
```

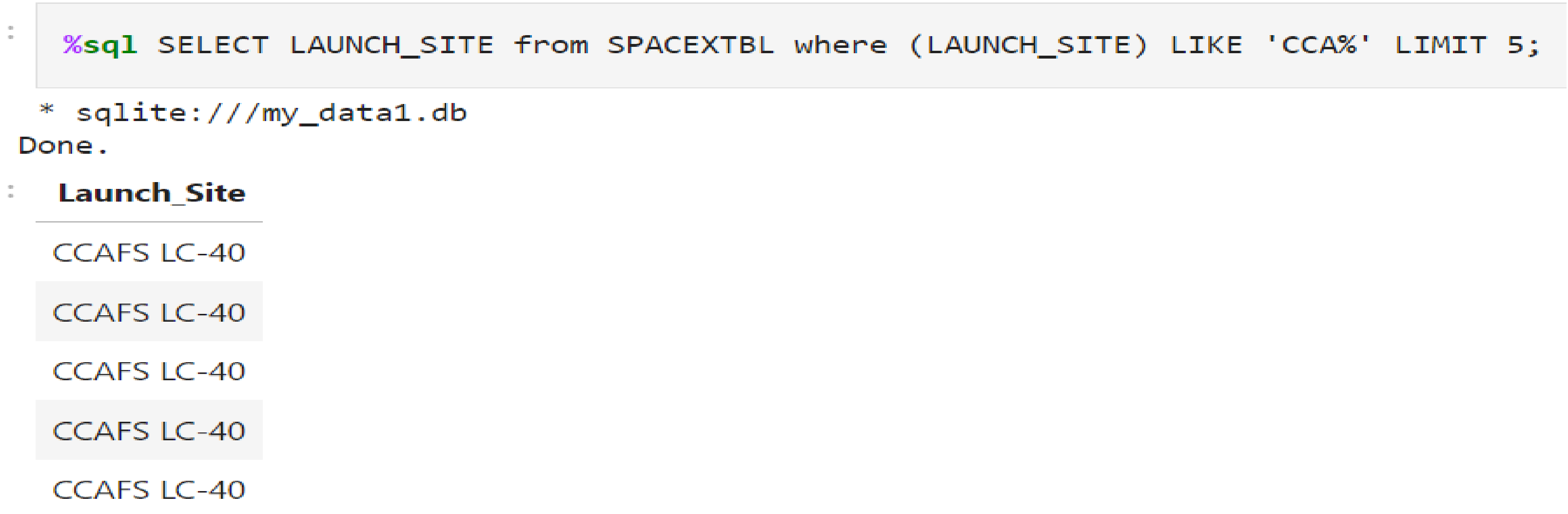
```
 * sqlite:///my_data1.db
Done.
```

[12]:

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

- These are the 4 Unique Launch sites in Space mission

# Launch Site Names Begin with 'CCA'



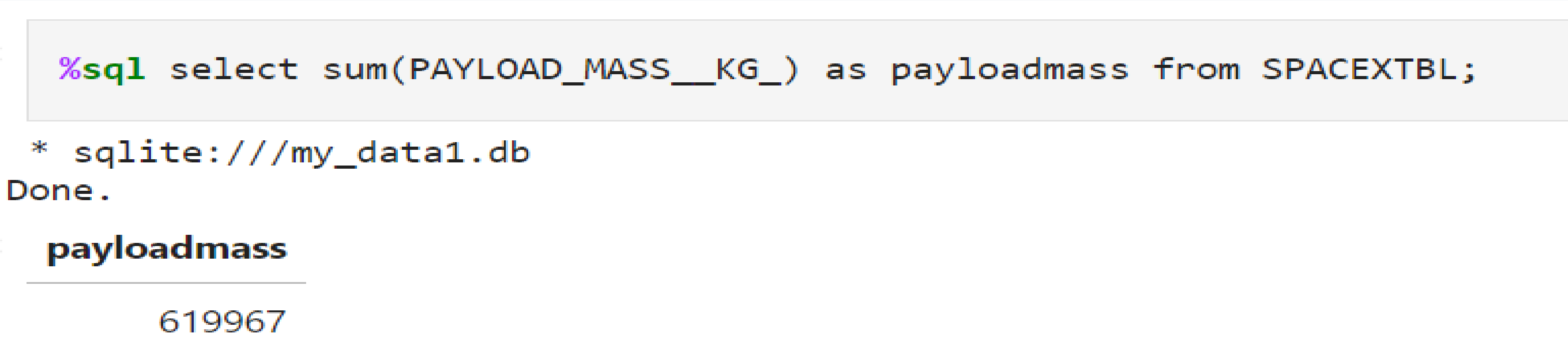
```
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

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

| Launch_Site |
|---|
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

- These are the 5 Launch sites that begins with ‘CCA’

# Total Payload Mass

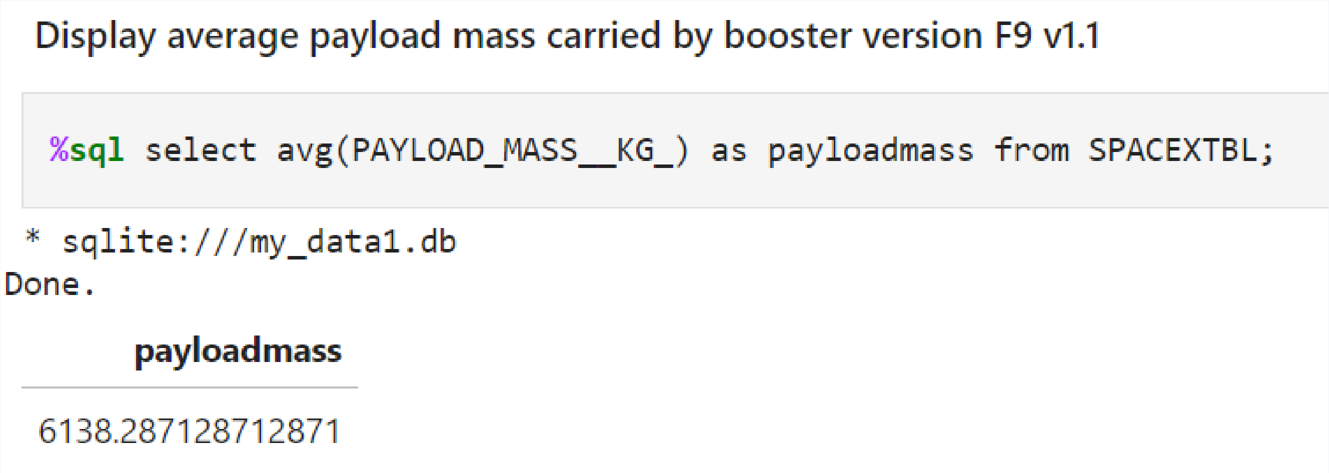


```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
   payloadmass
        619967
```

- The total payload carried by boosters from NASA is 619967

# Average Payload Mass by F9 v1.1



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

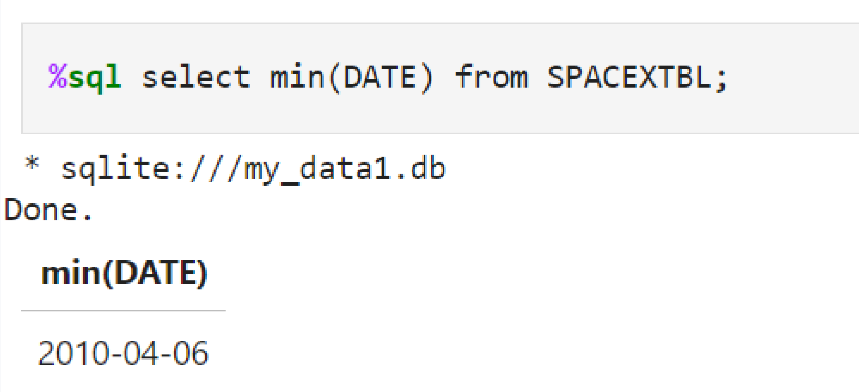
```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

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

| payloadmass |
|---|
| 6138.287128712871 |

- The average payload mass carried by booster version F9 v1.1 is 6138.29 kg

# First Successful Ground Landing Date



```
%sql select min(DATE) from SPACEXTBL;
```

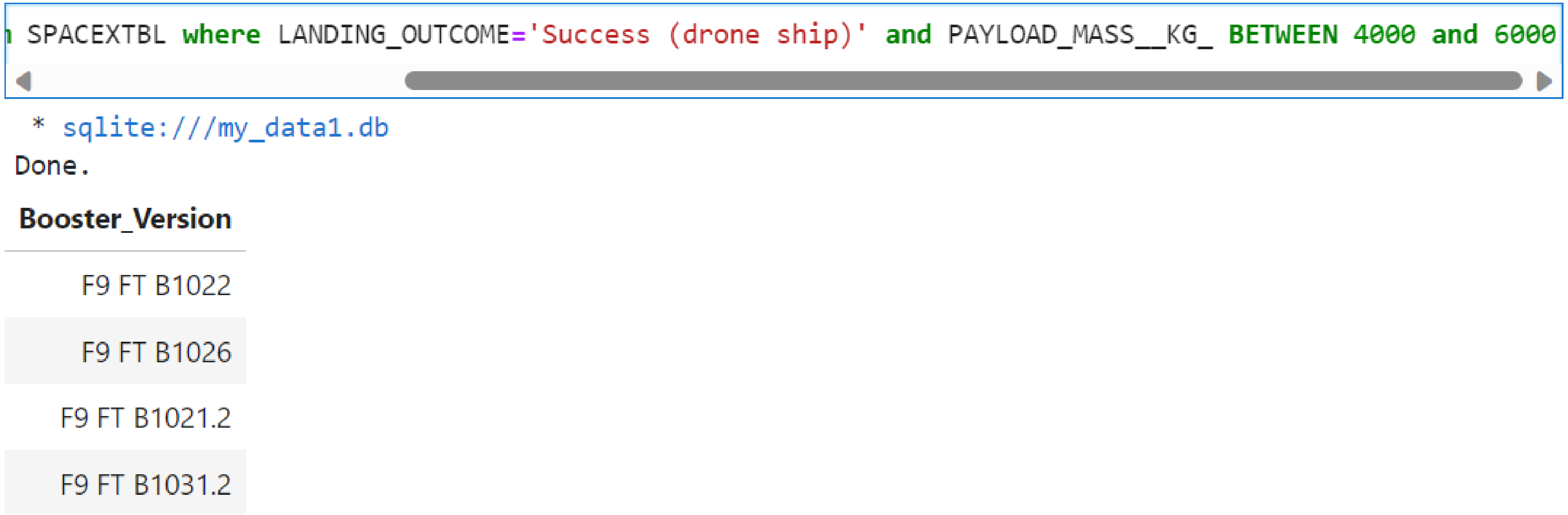
```
 * sqlite:///my_data1.db
Done.
```

| min(DATE) |
|---|
| 2010-04-06 |

- The date of the first successful landing outcome on ground pad is 2010-04-06

# 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='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000



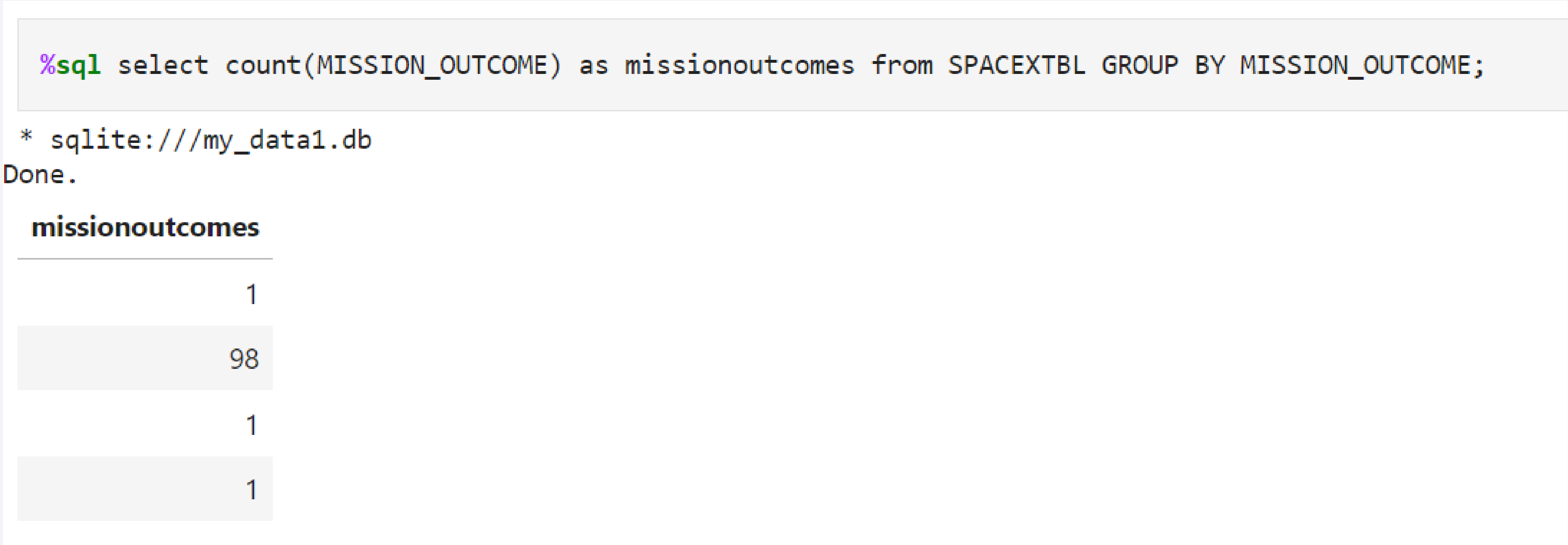
```
SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

| Booster_Version |
|---:|
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

- There are 4 booster versions which qualifies above condition

# Total Number of Successful and Failure Mission Outcomes



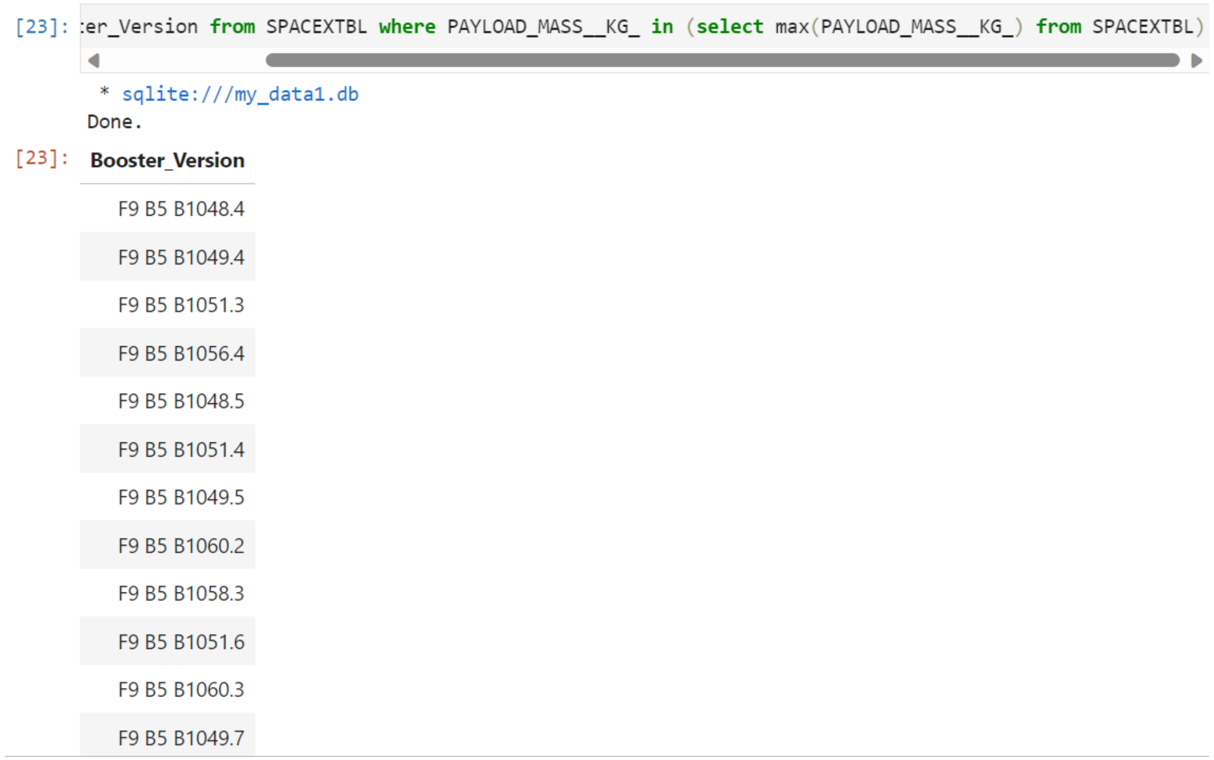
```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

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

| missionoutcomes |
|---:|
| 1 |
| 98 |
| 1 |
| 1 |

# Names of the booster_versions which have carried the maximum payload mass

%sql select Booster_Version from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)



```
[23]: :er_Version from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

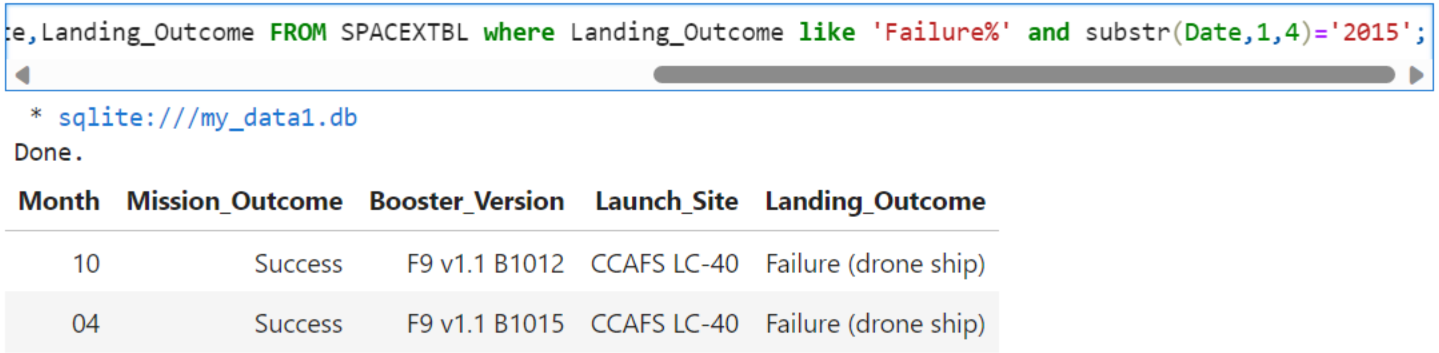
```
 * sqlite:///my_data1.db
Done.
```

[23]:

| 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 |

# Records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

%sql SELECT substr(Date, 6, 2) as Month,Mission_Outcome,Booster_Version,Launch_Site,Landing_Outcome FROM SPACEXTBL where Landing_Outcome like 'Failure%' and substr(Date,1,4)='2015’;

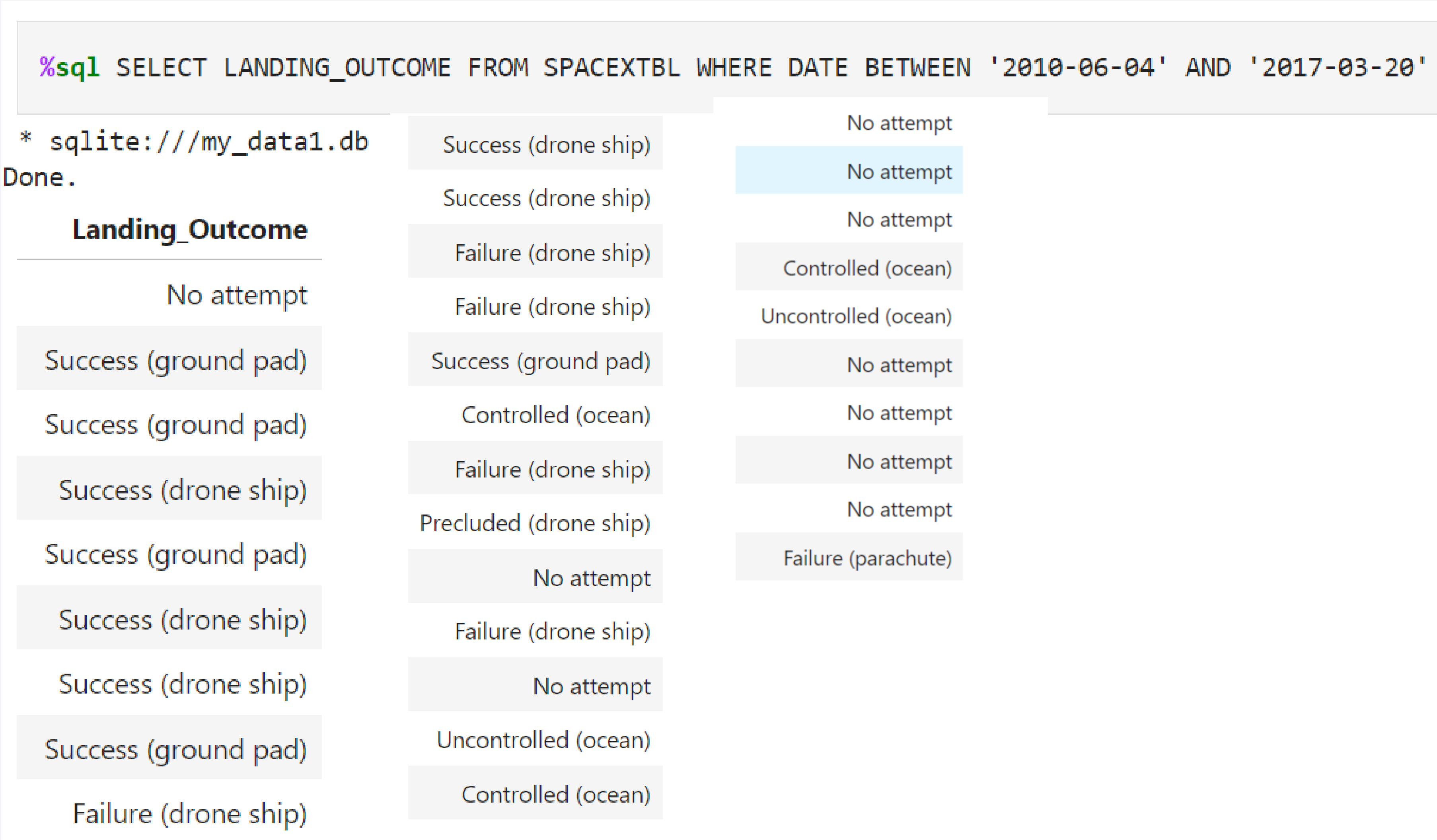


```
e,Landing_Outcome FROM SPACEXTBL where Landing_Outcome like 'Failure%' and substr(Date,1,4)='2015';
```

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

| Month | Mission_Outcome | Booster_Version | Launch_Site | Landing_Outcome |
|---|---|---|---|---|
| 10 | Success | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 04 | Success | F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

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



```
%sql SELECT LANDING_OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
```

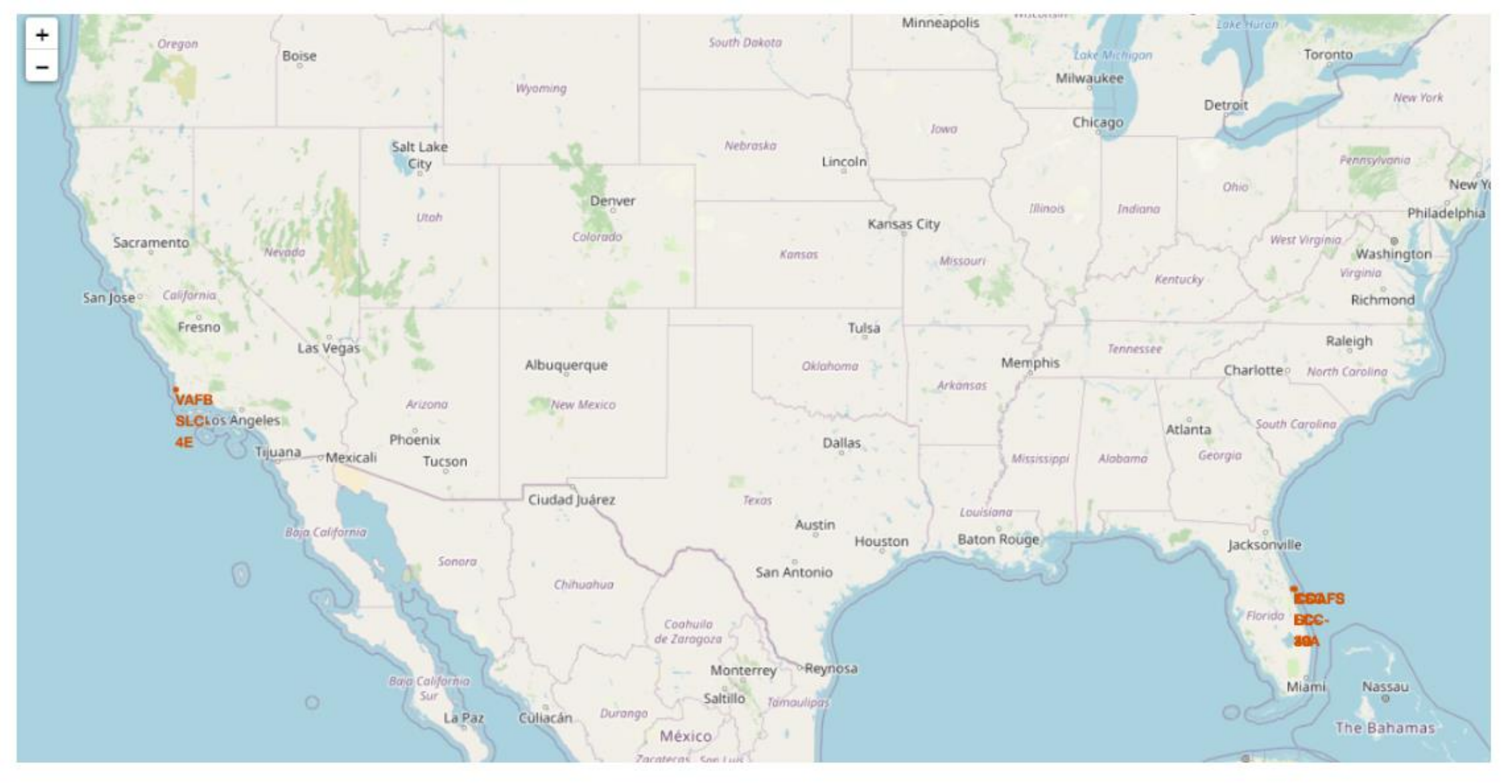
```
 * sqlite:///my_data1.db
Done.
```

| Landing_Outcome |
|---:|
| No attempt |
| Success (ground pad) |
| Success (ground pad) |
| Success (drone ship) |
| Success (ground pad) |
| Success (drone ship) |
| Success (drone ship) |
| Success (ground pad) |
| Failure (drone ship) |
| Success (drone ship) |
| Success (drone ship) |
| Failure (drone ship) |
| Failure (drone ship) |
| Success (ground pad) |
| Controlled (ocean) |
| Failure (drone ship) |
| Precluded (drone ship) |
| No attempt |
| Failure (drone ship) |
| No attempt |
| Uncontrolled (ocean) |
| Controlled (ocean) |
| No attempt |
| No attempt |
| No attempt |
| Controlled (ocean) |
| Uncontrolled (ocean) |
| No attempt |
| No attempt |
| No attempt |
| No attempt |
| Failure (parachute) |

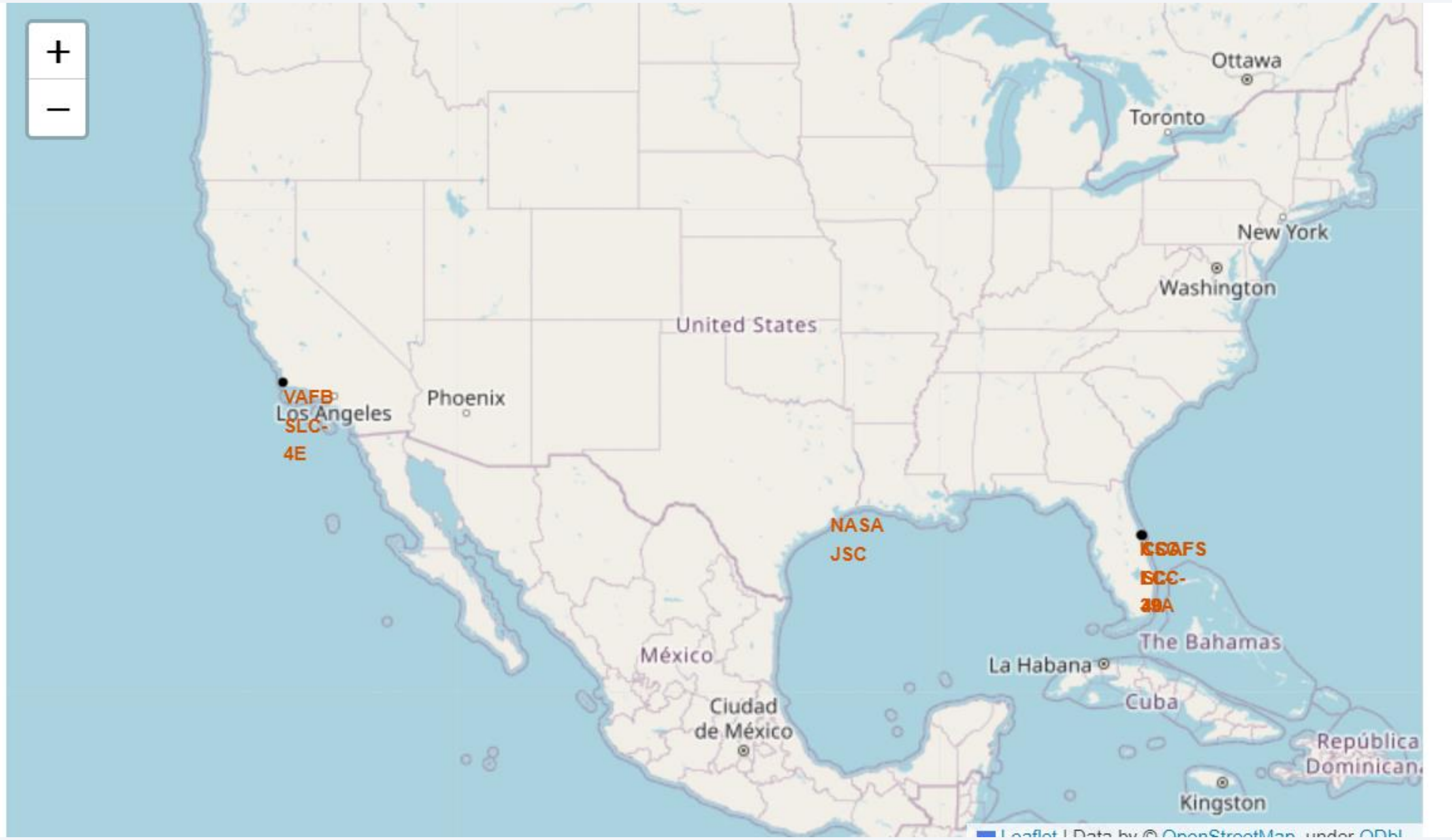
Section 3

# Launch Sites Proximities Analysis

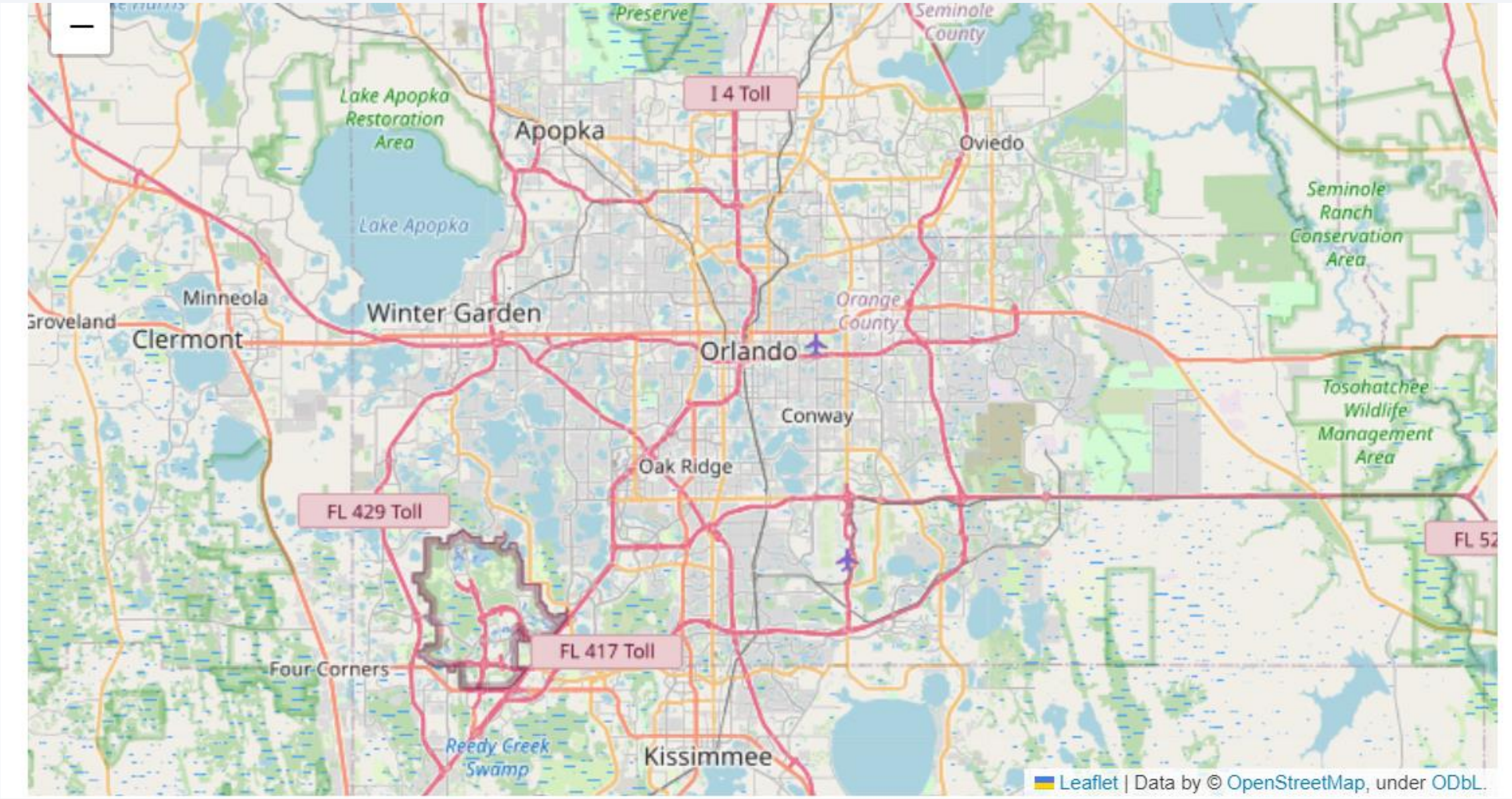
# Folium Map 1



# Folium Map 2


+
−
Ottawa
Toronto
New York
Washington
United States
Phoenix
VAFB
Los Angeles
SLC-
4E
NASA
JSC
CCAFS
SLC-
39A
The Bahamas
La Habana
Cuba
México
Ciudad de México
República Dominicana
Kingston
Leaflet | Data by © OpenStreetMap, under ODbL

# Folium Map 3


−
Preserve
Seminole County
I 4 Toll
Lake Apopka Restoration Area
Apopka
Oviedo
Lake Apopka
Seminole Ranch Conservation Area
Minneola
Groveland
Winter Garden
Clermont
Orange County
Orlando
Conway
Tosohatchee Wildlife Management Area
Oak Ridge
FL 429 Toll
FL 52
FL 417 Toll
Four Corners
Reedy Creek Swamp
Kissimmee
Leaflet | Data by © OpenStreetMap, under ODbL.

Section 5

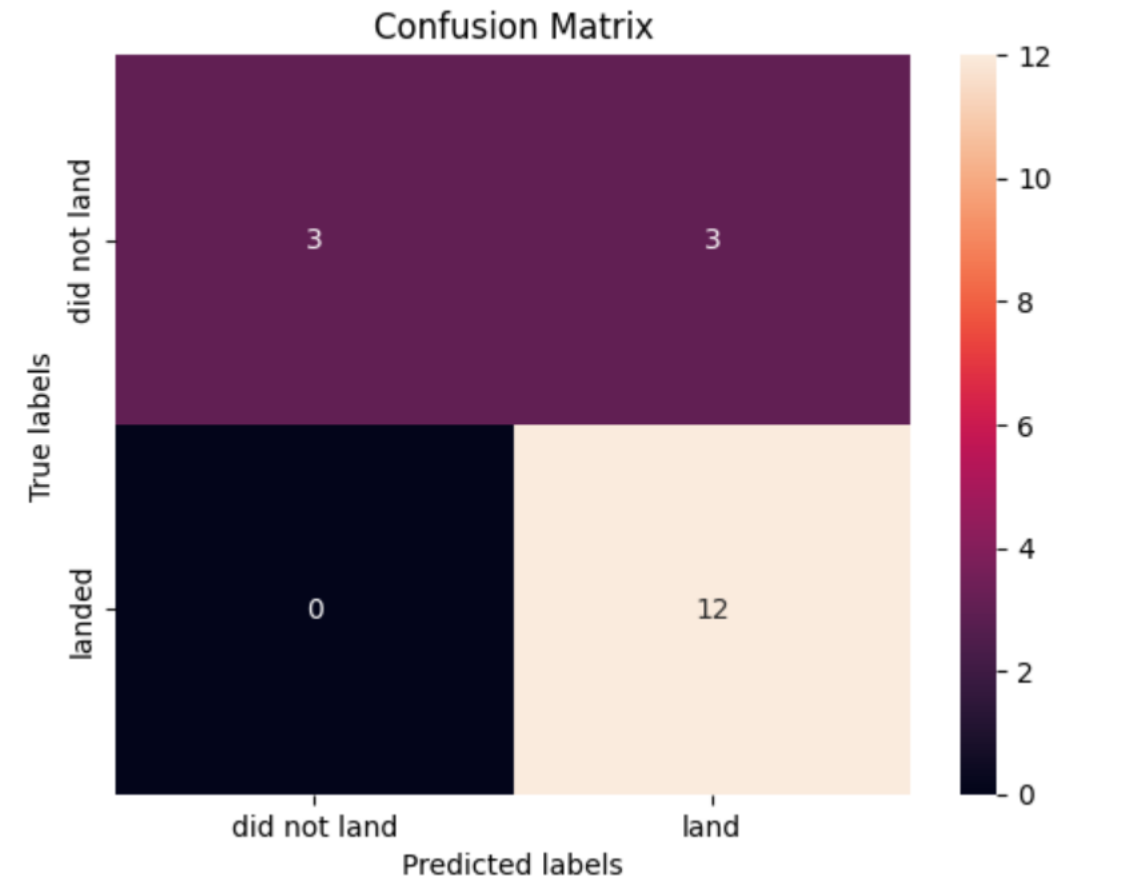
# Predictive Analysis (Classification)

# Classification Accuracy

- At the end, we found Classification Accuracy for different models as below:

    1. Logistic Regression test data accuracy: 0.83333%
    2. SVM test data accuracy: 0.84821%
    3. Decision tree data accuracy: 0.88928%
    4. K-neighbours data accuracy: 0.848215%

# Confusion Matrix



- This is the confusion matrix of Decision tree.
- As we see Decision tree method have highest test data accuracy, i.e, 88.9% we consider this as the best performing model.

# Conclusions

- To Summarize, we identify the problem that we need to predict if the Falcon 9 first stage will land successfully.
- Then we collected the data using Webscraping and SpaceX URL.
- We preprocessed and analyzed data using Exploratory data analysis and SQL queries.
- We trained and tested using different Machine learning algorithms.
- Finally, we concluded that Decision tree is the best model for this problem.
- Hence with 88% Success accuracy, we can conclude that Falcon 9 first stage will land successfully.

# Thank you!