IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Nabila Rahmah
December 2024

# Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - Insights Drawn from EDA with Pandas and SQL
  - Launch Sites Proximity Analysis with Folium
  - Dashboard with Plotly Dash
  - Predictive Analysis (Classification) with Machine Learning
- Conclusion
- Appendix

# Executive Summary

- We used various methods of data exploration (static data visualization, SQL, and interactive map and dashboard using Folium and Dash) to understand more about relevant past launches of Space X Falcon 9 rockets. We also applied several Machine Learning algorithms (Logistic Regression, SVM, Decision Tree, and KNN) to predict the launch outcome of the rockets.
- The result shows that several variables are to be considered to achieve a successful launch, namely launch site, payload, orbit type, flight date, and booster. We found out that the launch site KSC LC-39A has the highest success rate, with the payload range of 2500-5000kg andF9 Booster "FT" contributing to the optimal launch success. For further predictive analysis, the Decision Tree algorithm performs best in classifying launch outcome using the aforementioned variables as input.

# Introduction

**Project Background**

In this age of commercial space age, companies are trying to make space travels affordable to customers. One of the most successful company in this field is Space X, which to this date has some notable achievements (sending spacecrafts to the ISS, Starlink satelite, and sending manned missions to space).

SpaceX advertises Falcon 9 rockets that costs less than other competitor, because it has the advantage of the ability to recover and reuse the first stage of the rocket.

Our company, SpaceY, is a new company trying to get into this field. Therefore, we need to understand more informations about relevant past launches from our competitor, including the **cost of launches** and the **first stage reuse data** of SpaceX Falcon 9 rockets.

# Introduction

### Main Problem

What can we learn from past launches of Space X Falcon 9 rockets?

### Research Problems

1. What factors impact the success of the Space X Falcon 9 rockets landing?
2. How does the factors interact with each other to influence the success of a landing?

2. What other insights can we get from the Space X Falcon 9 rockets data?

Section 1

# Methodology

# Methodology

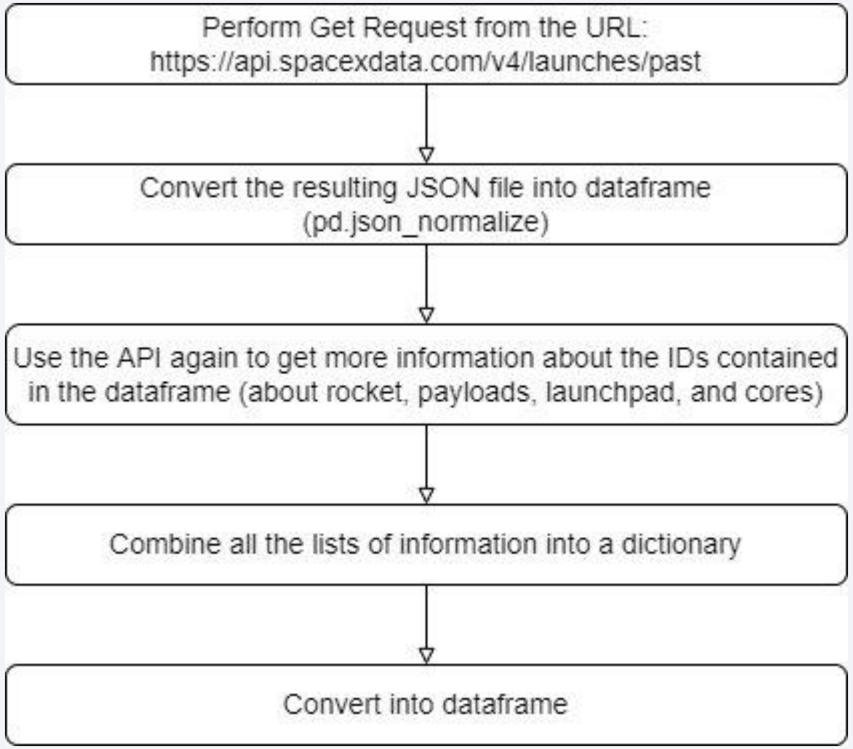
## Executive Summary

- Data collection methodology:
  - SpaceX Rest API and web-scraping methods
- Perform data wrangling
  - Data cleaning, formatting, creating new feature (landing outcomes classified into successful and unsuccessful)
- 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 models are built and tuned with GridSearch to predict landing outcome based on various factors, then evaluated based on accuracy score and the confusion matrix

# Data Collection – SpaceX API

- We'll collect data by making REST API calls to SpaceX's API and extract important information using identification numbers in the launch data
- Then, we convert the resulting json into pandas dataframe. We also filter the dataframe to include only relevant Falcon 9 data and deal with the missing values
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



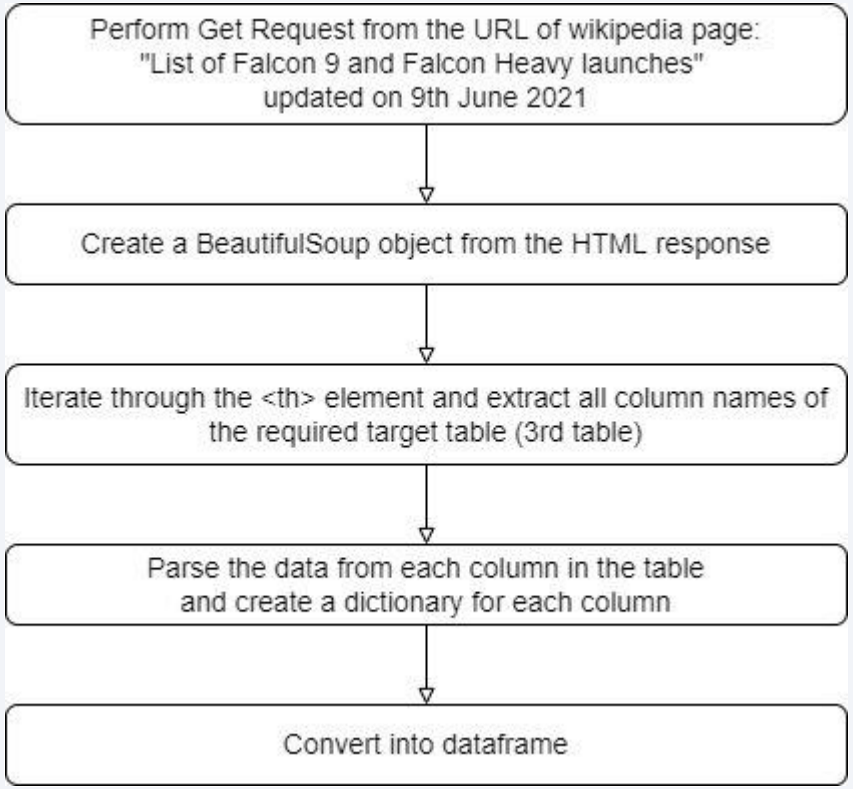
Perform Get Request from the URL: https://api.spacexdata.com/v4/launches/past
Convert the resulting JSON file into dataframe (pd.json_normalize)
Use the API again to get more information about the IDs contained in the dataframe (about rocket, payloads, launchpad, and cores)
Combine all the lists of information into a dictionary
Convert into dataframe

flowchart of SpaceX API calls

# Data Collection - Scraping

- We'll collect data by webscraping the Wikipedia page “List of Falcon 9 and Falcon Heavy launches” dated 9th June 2021
- Then, we convert the HTML response into BeautifulSoup object, extract column information, and create a pandas dataframe
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Perform Get Request from the URL of wikipedia page: "List of Falcon 9 and Falcon Heavy launches" updated on 9th June 2021

Create a BeautifulSoup object from the HTML response

Iterate through the <th> element and extract all column names of the required target table (3rd table)

Parse the data from each column in the table and create a dictionary for each column

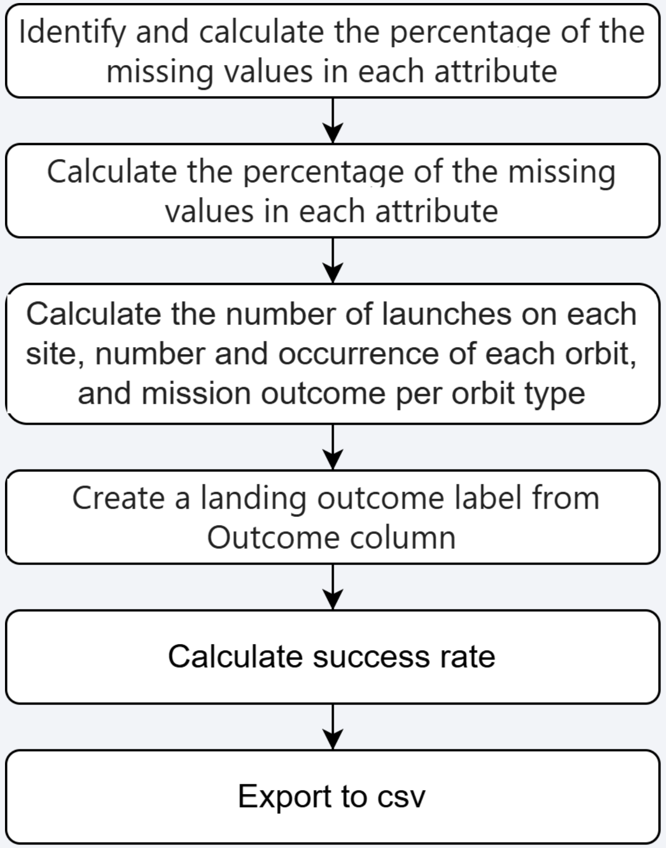
Convert into dataframe

flowchart of SpaceX data webscraping

# Data Wrangling

- We'll inspect the variable types, see value occurences, and identify missing values for each column in the dataframe.
- Then, we create a new target variable that categorizes successful and unsuccessful launches, and export the result into csv file
- See the code used for more details:

[https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/labs-jupyter-spacex-data%20wrangling_jupyterlite.ipynb](https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/labs-jupyter-spacex-data%20wrangling_jupyterlite.ipynb)


Identify and calculate the percentage of the missing values in each attribute
Calculate the percentage of the missing values in each attribute
Calculate the number of launches on each site, number and occurrence of each orbit, and mission outcome per orbit type
Create a landing outcome label from Outcome column
Calculate success rate
Export to csv

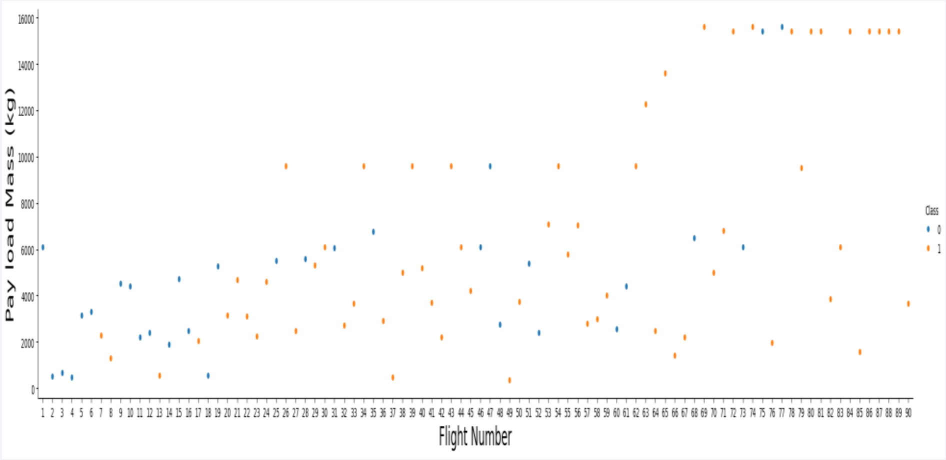
flowchart of data wrangling process

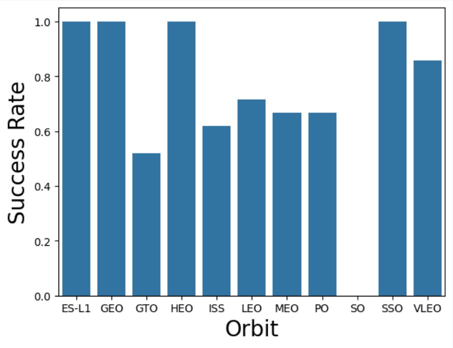
# EDA with Data Visualization

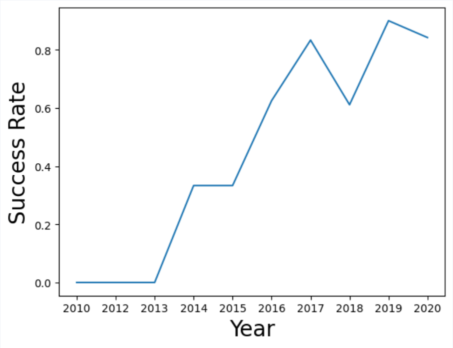
- Scatterplots hued with the outcome variable are used to see the relationship between Flight Number, Launch Site, Orbit Type, and Payload Mass as they relate to the Outcome of the Launch

- Bar chart is used to see relationship between success rate and orbit type

- Line chart is used to see relationship between the year of launch and the launch success rate







# EDA with Data Visualization

- The resulting knowledge about variables that might influence the outcome variable is then used to build the feature for subsequent analysis, and perform additional data formatting steps (creating dummy variables and ensuring float data type for numerical variables)
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/edadataviz.ipynb

# EDA with SQL

- We connect with the database and load the dataset
- Using SQL queries, we explore:
    - Unique launch sites names and records with certain strings or that fulfills certain conditions
    - Calculating total number of successful and failure mission outcomes, total payload mass carried by boosters launched by NASA (CRS), and average payload mass carried by booster version F9 v1.1
    - The date where the successful landing outcome in drone ship was achieved
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

# Build an Interactive Map with Folium

- We'll perform more interactive visual analytics with Folium to answer more exploratory questions
- We use markers and circles to mark all the launch sites on the map, complete with the success/failed launches for each site, distinguished by color. This is to see which sites have high success rates.
- Then, we calculate the distances between a launch site to its proximities using polylines.
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

# Build a Dashboard with Plotly Dash

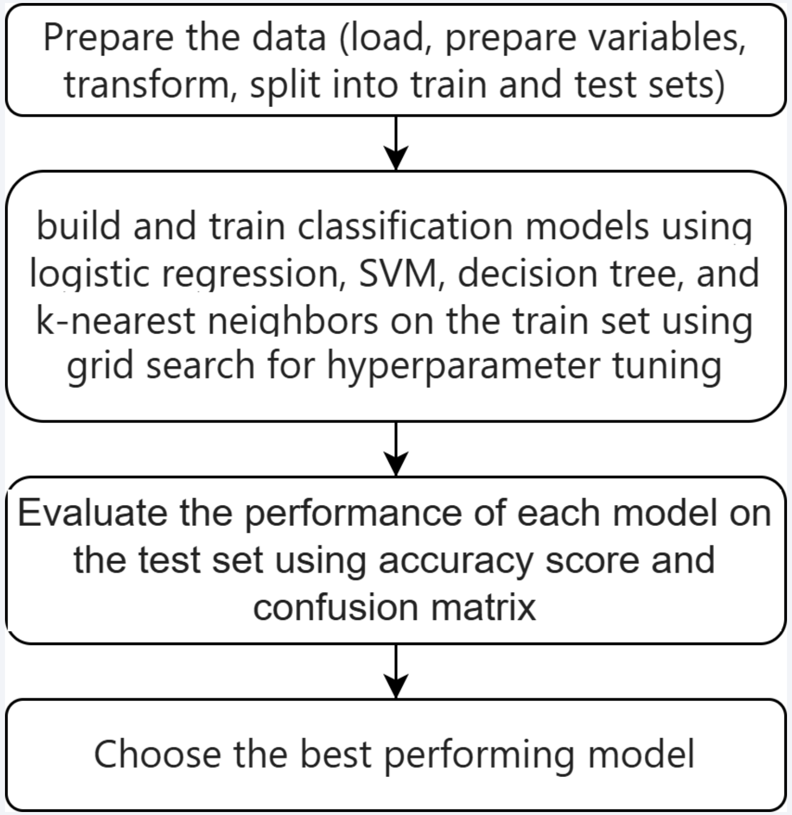
- We build a dashboard with Plotly Dash to further explore the data in a more interactive way
- We add a pie chart to display the successful launches by site, which can be altered by choosing the appropriate sites/all sites option from the dropdown menu
- We also added a scatterplot that displays how successful launches differ by payload range, with a sliding range for payload, and color coded by the booster names
- These graphs may help us answer the question as to which site, payload range, and booster has the most successful landing and success rate
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/spacex_dash_app.py

# Predictive Analysis (Classification)

- We will create and compare classification models with machine learning to predict if the first stage will land given the data
- First, we prepare the data by loading, perform data transformation (standardizing) and splitting the data into train and test sets.
- Then, we create and train several classification models, which are logistic regression, support vector machine, decision tree, and k-nearest neighbors. The models are trained on train set data with grid search to find the most optimal hyperparameters.
- We evaluate each model by scoring how well the model classify the test data set. Accuracy and confusion matrix are taken into considerations when choosing the best performing model.
- See the code used for more details:

https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Prepare the data (load, prepare variables, transform, split into train and test sets)

build and train classification models using logistic regression, SVM, decision tree, and k-nearest neighbors on the train set using grid search for hyperparameter tuning

Evaluate the performance of each model on the test set using accuracy score and confusion matrix

Choose the best performing model

flowchart of SpaceX landing success classification

# Results

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

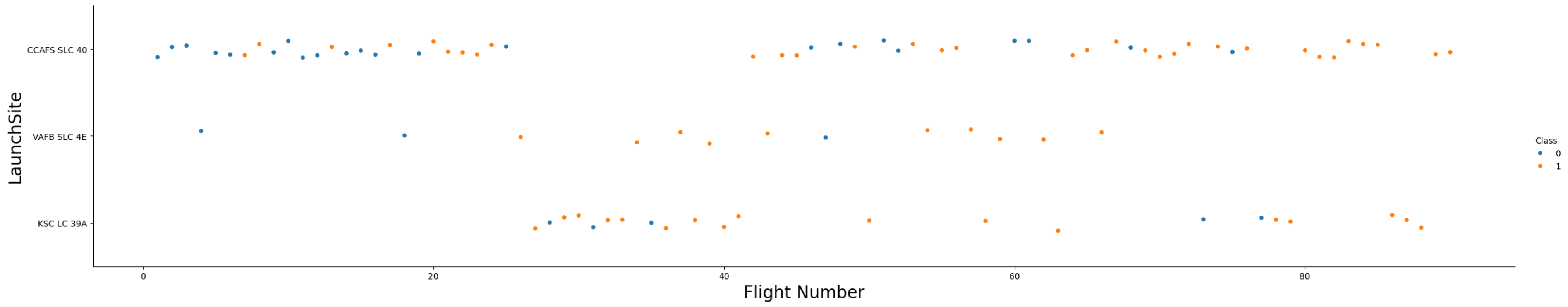
Section 2

# Insights drawn from EDA

Part 1

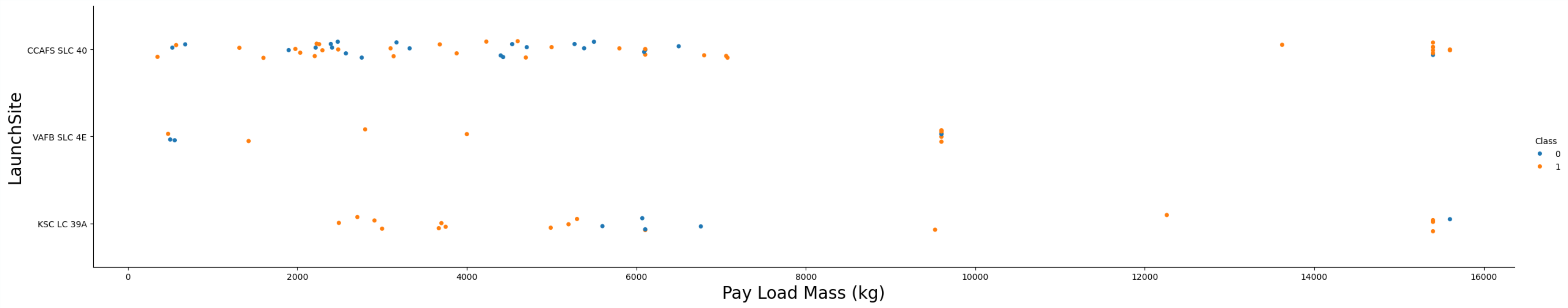
# Python EDA with Visualization

# Flight Number vs. Launch Site



- In general, the higher the flight number (dating from earlier to later), the higher the success rate of the landing, regardless of the launch site.
- The CCAFS SLC 40 has the largest number of flights, with significant improvement in landing success since the flight number 40

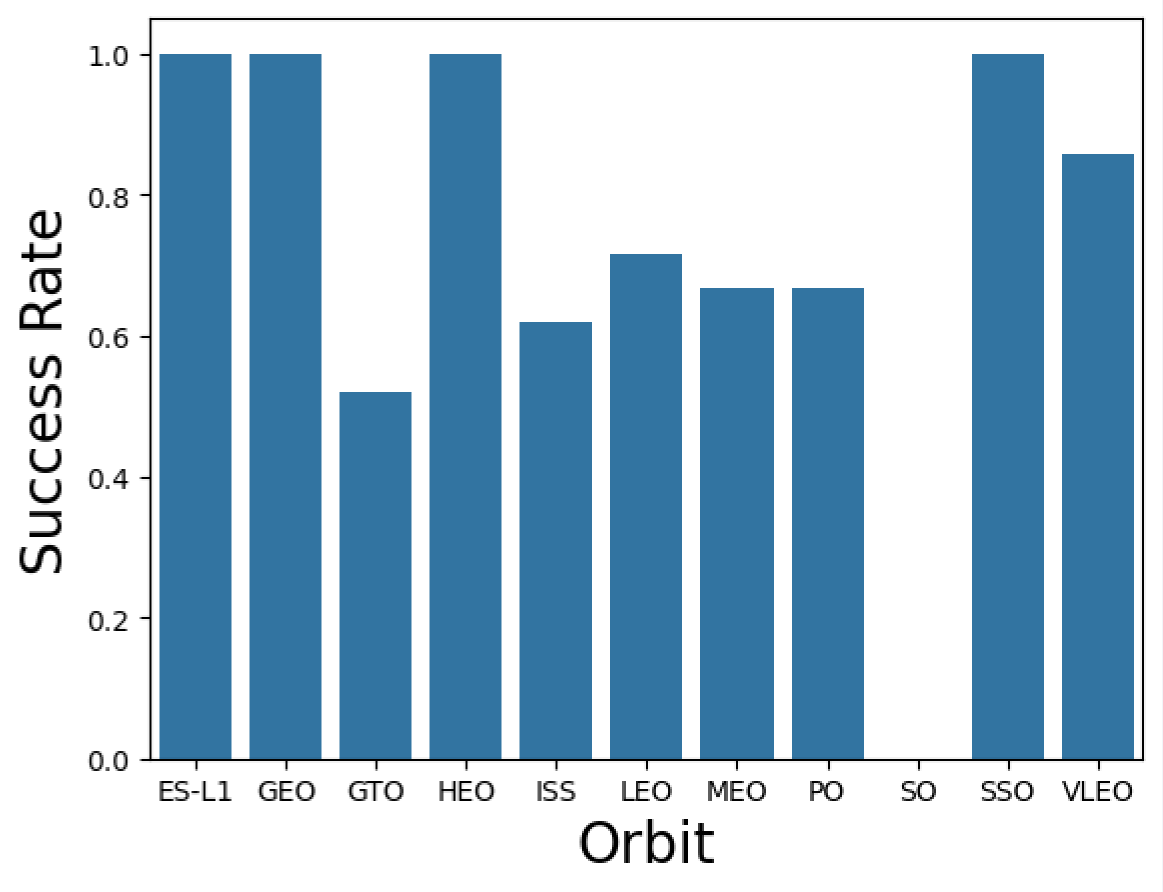
# Payload vs. Launch Site



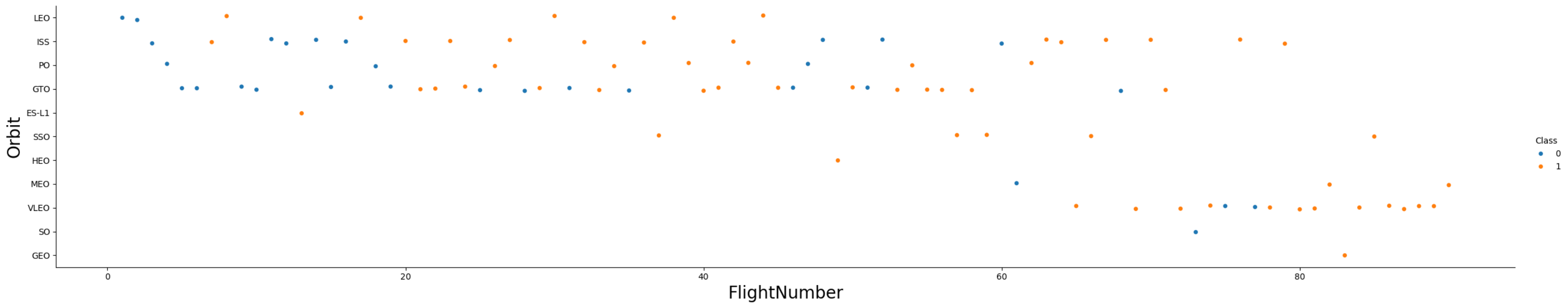
- At the CCAFS SLC 40 launch site, the heavier payload mass tend to have more success rate than the lighter payload mass
- The VAFB SLC 4E has no flight records for the heavier payload mass (>10000 kg)

# Success Rate vs. Orbit Type

- The ES-L1, GEO, HEO, and SSO have the highest and maximum success rate of 1
- SO only has 1 flight with success rate of 0

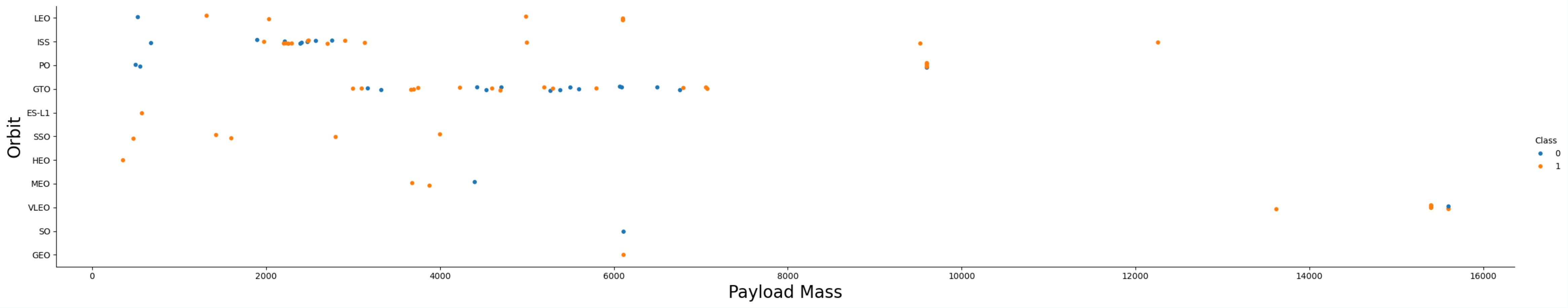


# Flight Number vs. Orbit Type



- In the LEO and VLEO orbit, success seems to be related to the number of flights, with the latest flights having more success rate than the earlier flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success

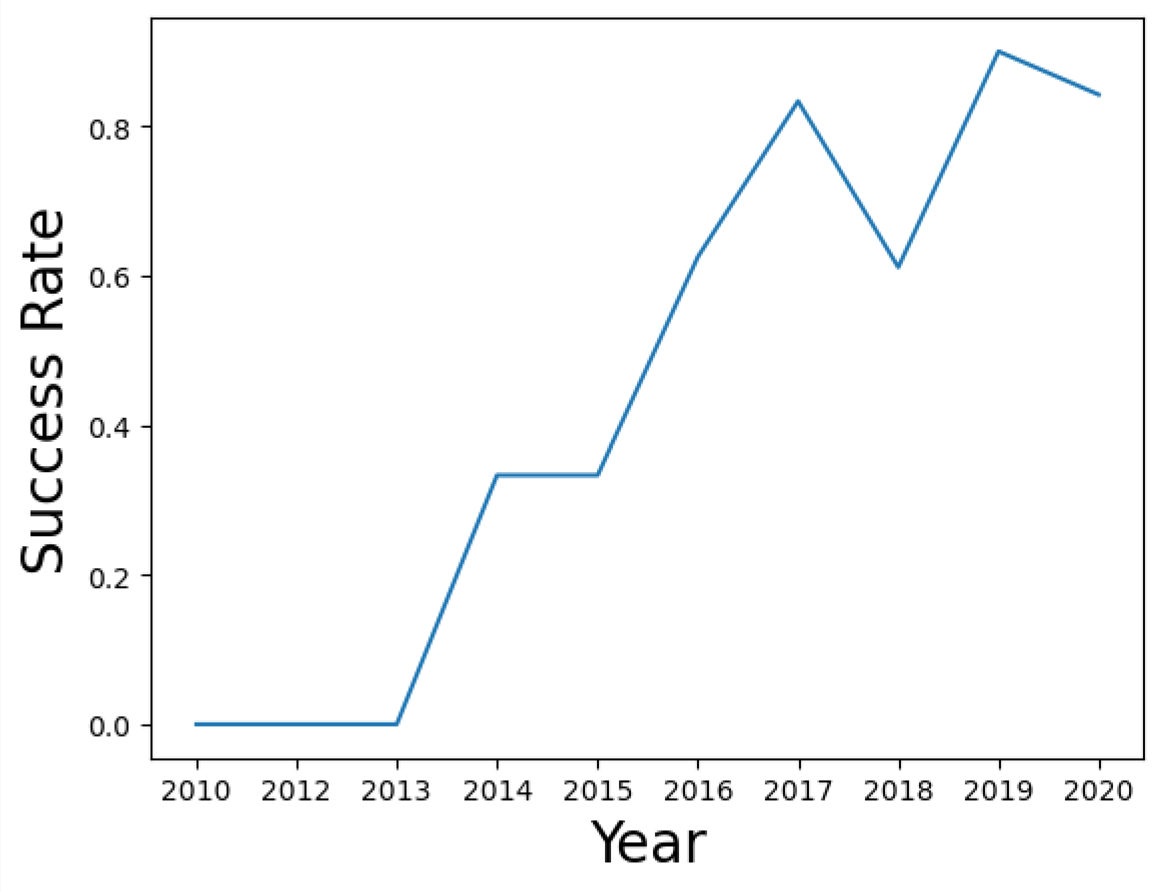
# Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are higher for the Polar, LEO and ISS orbits
- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

# Launch Success Yearly Trend

- The yearly average success rate tends to rise from the year of 2013-2017, gets lower in 2018, and then rise again to its peak in 2019.
- Overall, it has an increasing trend for the time range of 2010-2020

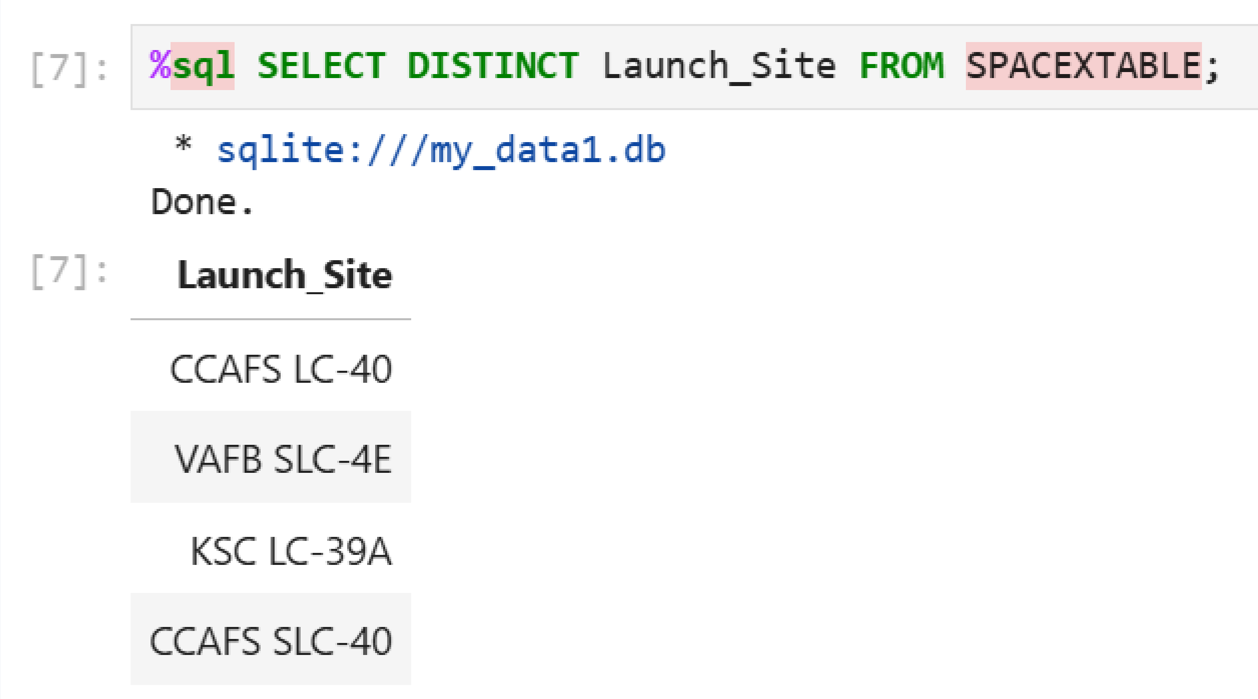


Part 2

# EDA with SQL

# All Launch Site Names

- We use SELECT and DISTINCT in the query to explore all distinct launch site names



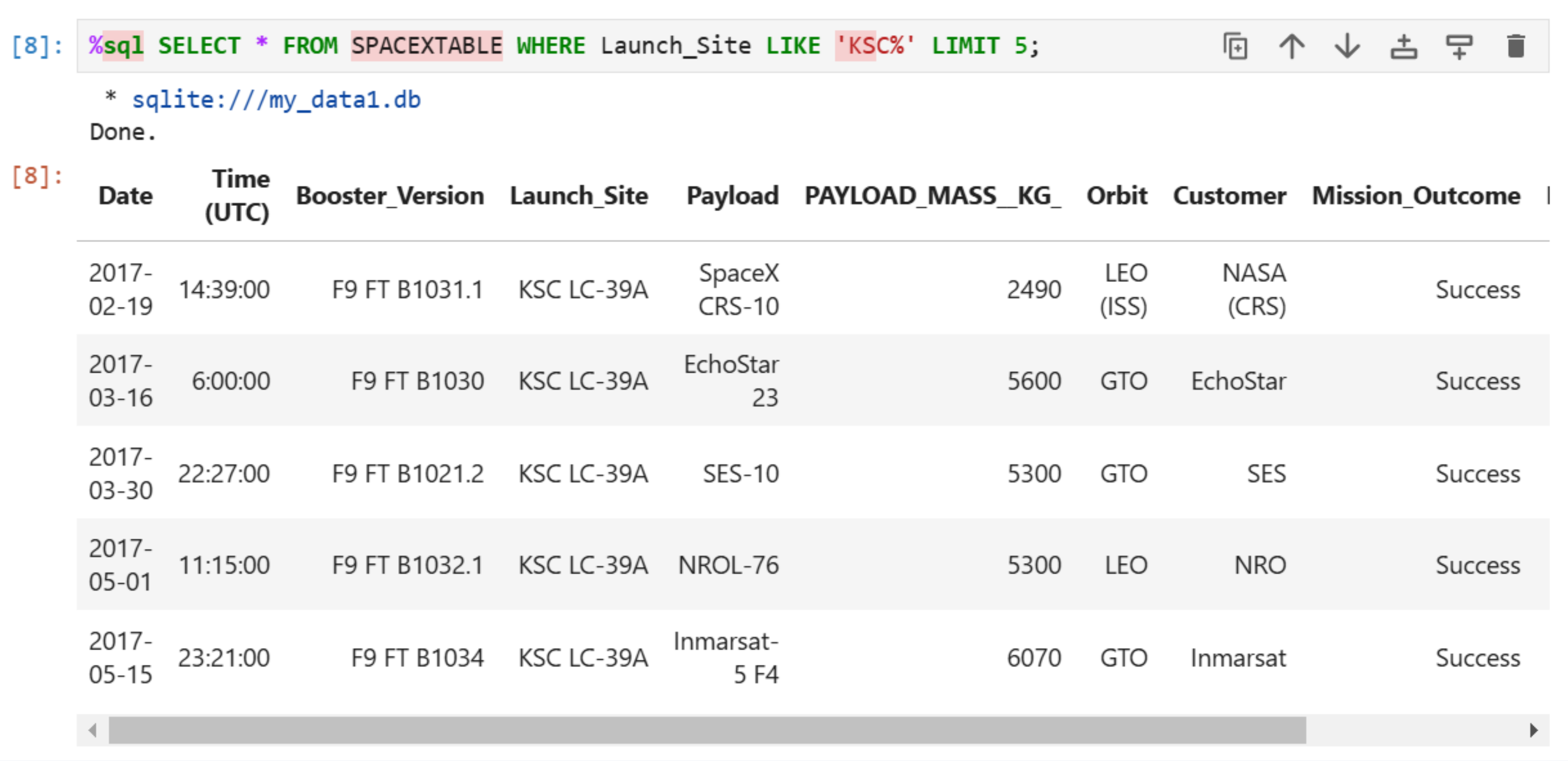
```
[7]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

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

[7]:

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

# Launch Site Names Begin with 'KSC'



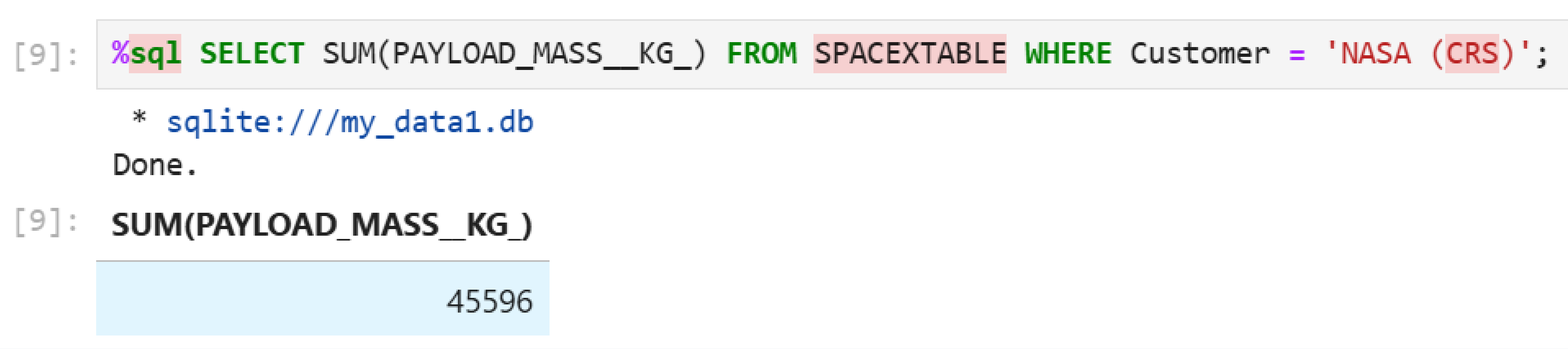
```
[8]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'KSC%' LIMIT 5;

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

[8]:

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome |
|---|---|---|---|---|---|---|---|---|
| 2017-02-19 | 14:39:00 | F9 FT B1031.1 | KSC LC-39A | SpaceX CRS-10 | 2490 | LEO (ISS) | NASA (CRS) | Success |
| 2017-03-16 | 6:00:00 | F9 FT B1030 | KSC LC-39A | EchoStar 23 | 5600 | GTO | EchoStar | Success |
| 2017-03-30 | 22:27:00 | F9 FT B1021.2 | KSC LC-39A | SES-10 | 5300 | GTO | SES | Success |
| 2017-05-01 | 11:15:00 | F9 FT B1032.1 | KSC LC-39A | NROL-76 | 5300 | LEO | NRO | Success |
| 2017-05-15 | 23:21:00 | F9 FT B1034 | KSC LC-39A | Inmarsat-5 F4 | 6070 | GTO | Inmarsat | Success |

- To find 5 records where launch sites' names start with `KSC`, we use SELECT, WHERE, LIKE, and LIMIT in the query

# Total Payload Mass



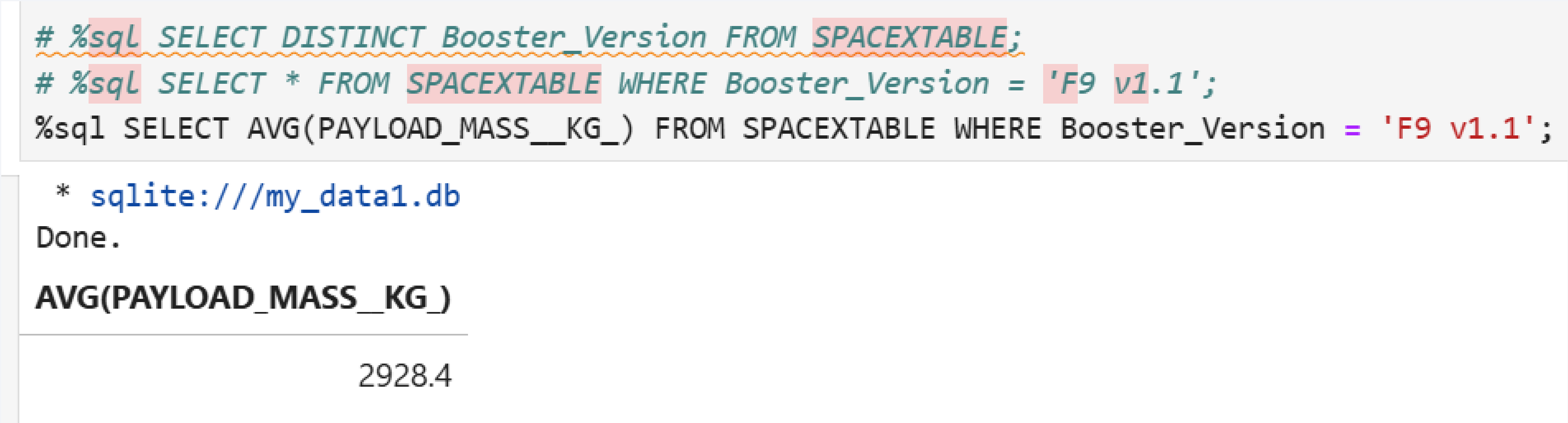
```
[9]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';

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

[9]:

| SUM(PAYLOAD_MASS__KG_) |
|---:|
| 45596 |

- Using the query showed above, we obtain the total payload carried by boosters from NASA which is 45596 kg

# Average Payload Mass by F9 v1.1



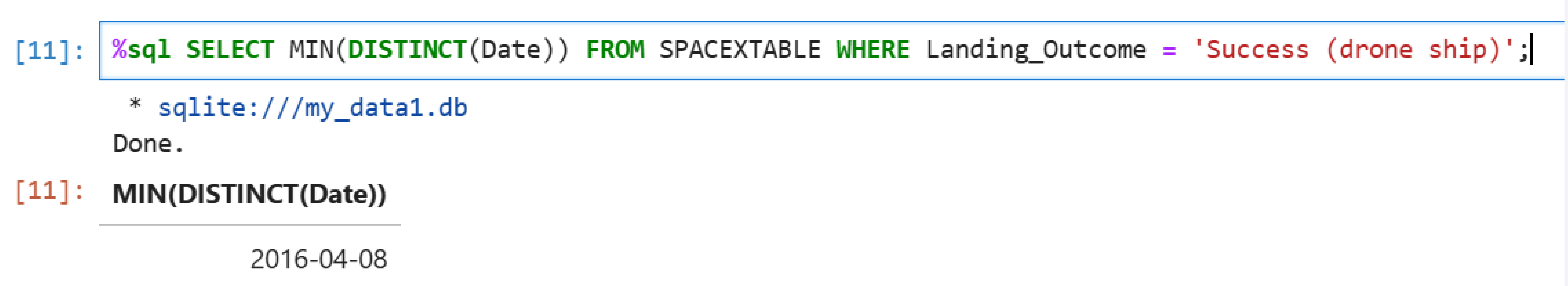
```
# %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE;
# %sql SELECT * FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

| AVG(PAYLOAD_MASS__KG_) |
|---:|
| 2928.4 |

- Using the query showed above, we obtain the average payload mass carried by booster version F9 v1.1 as 2928.4 kg

# First Successful Landing Date



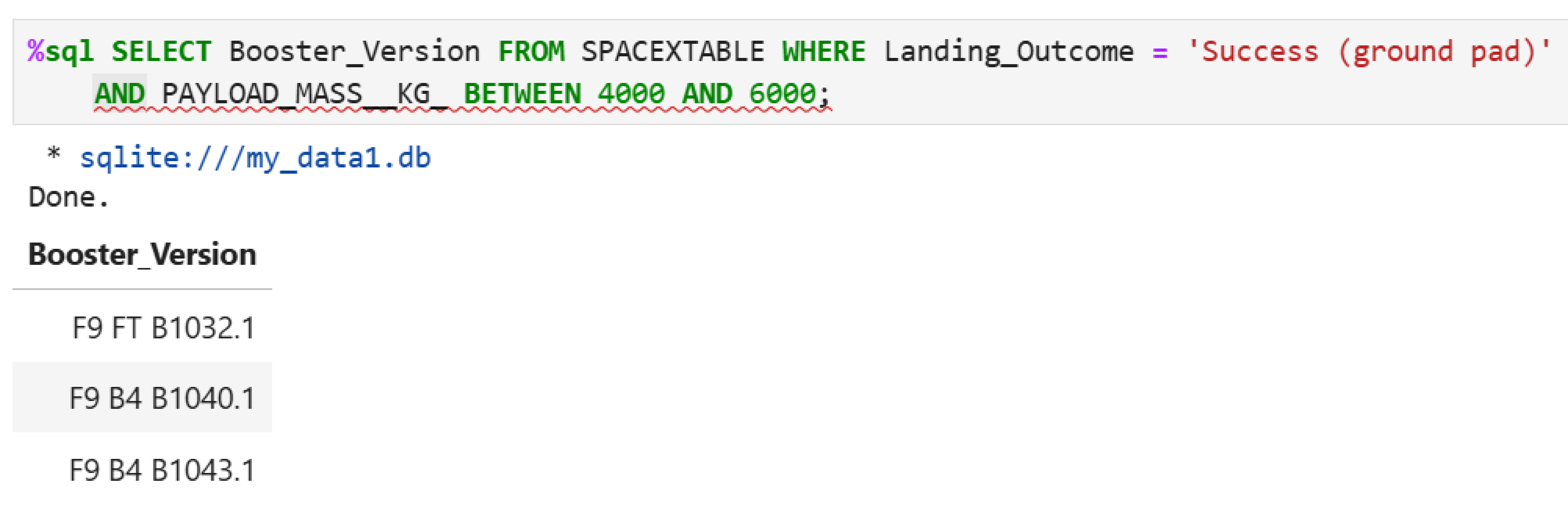
```
[11]: %sql SELECT MIN(DISTINCT(Date)) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)';
```

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

```
[11]: MIN(DISTINCT(Date))
                2016-04-08
```

- We found that the 8th of April, 2016 was the the date of the first successful landing outcome on a drone ship using above query

# Successful Ground Pad Landing with Payload between 4000 and 6000



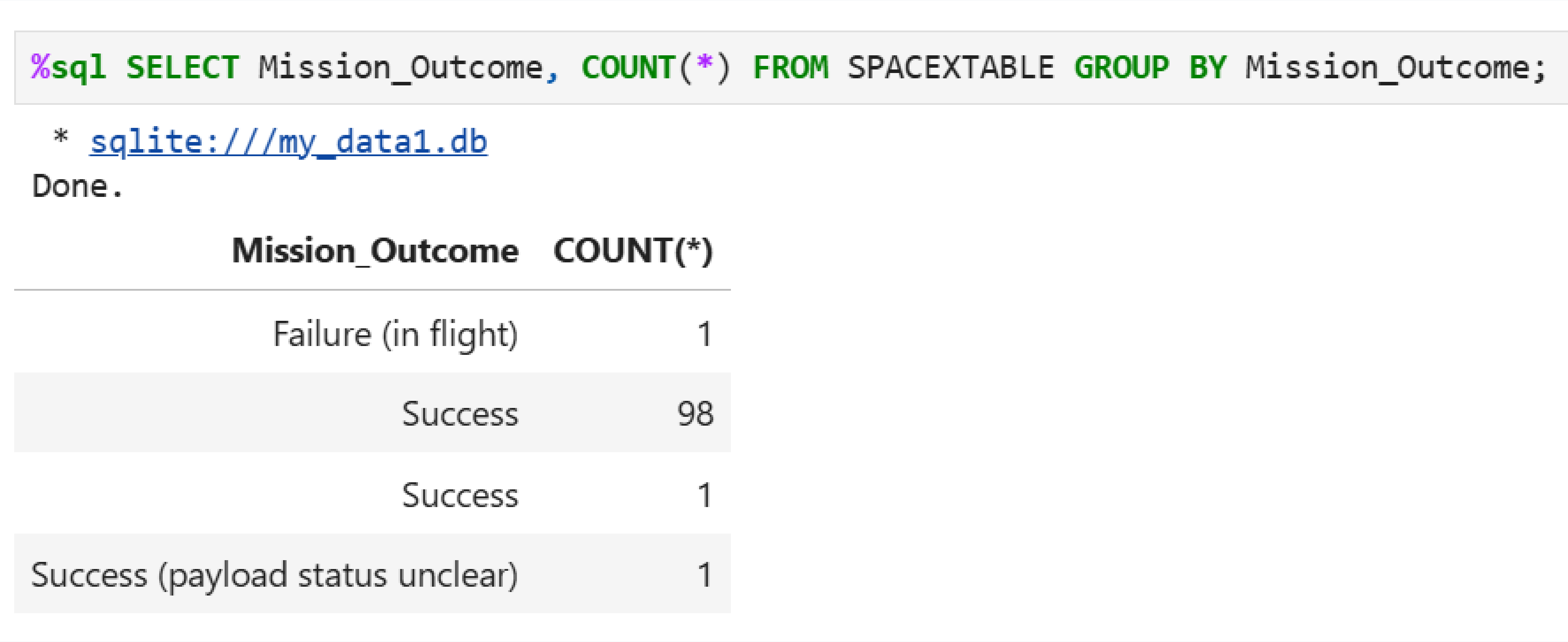
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'
    AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

| Booster_Version |
|---|
| F9 FT B1032.1 |
| F9 B4 B1040.1 |
| F9 B4 B1043.1 |

- We present the name of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 by using above query

# Total Number of Successful and Failure Mission Outcomes



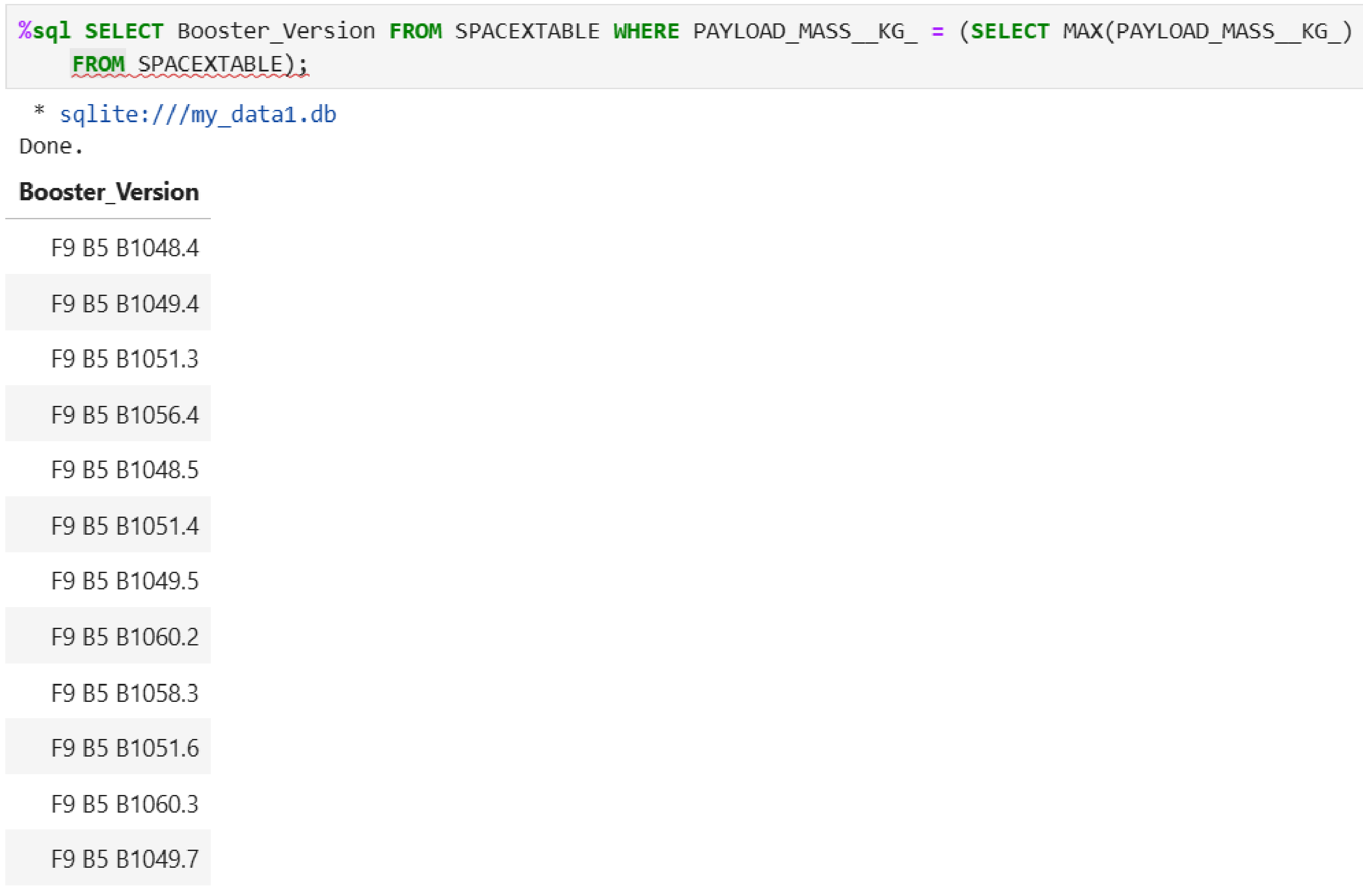
```
%sql SELECT Mission_Outcome, COUNT(*) FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

| Mission_Outcome | COUNT(*) |
|---:|---:|
| Failure (in flight) | 1 |
| Success | 98 |
| Success | 1 |
| Success (payload status unclear) | 1 |

- We can see that there are 1 failure in flight and 100 successes by using above query

# Boosters Carried Maximum Payload



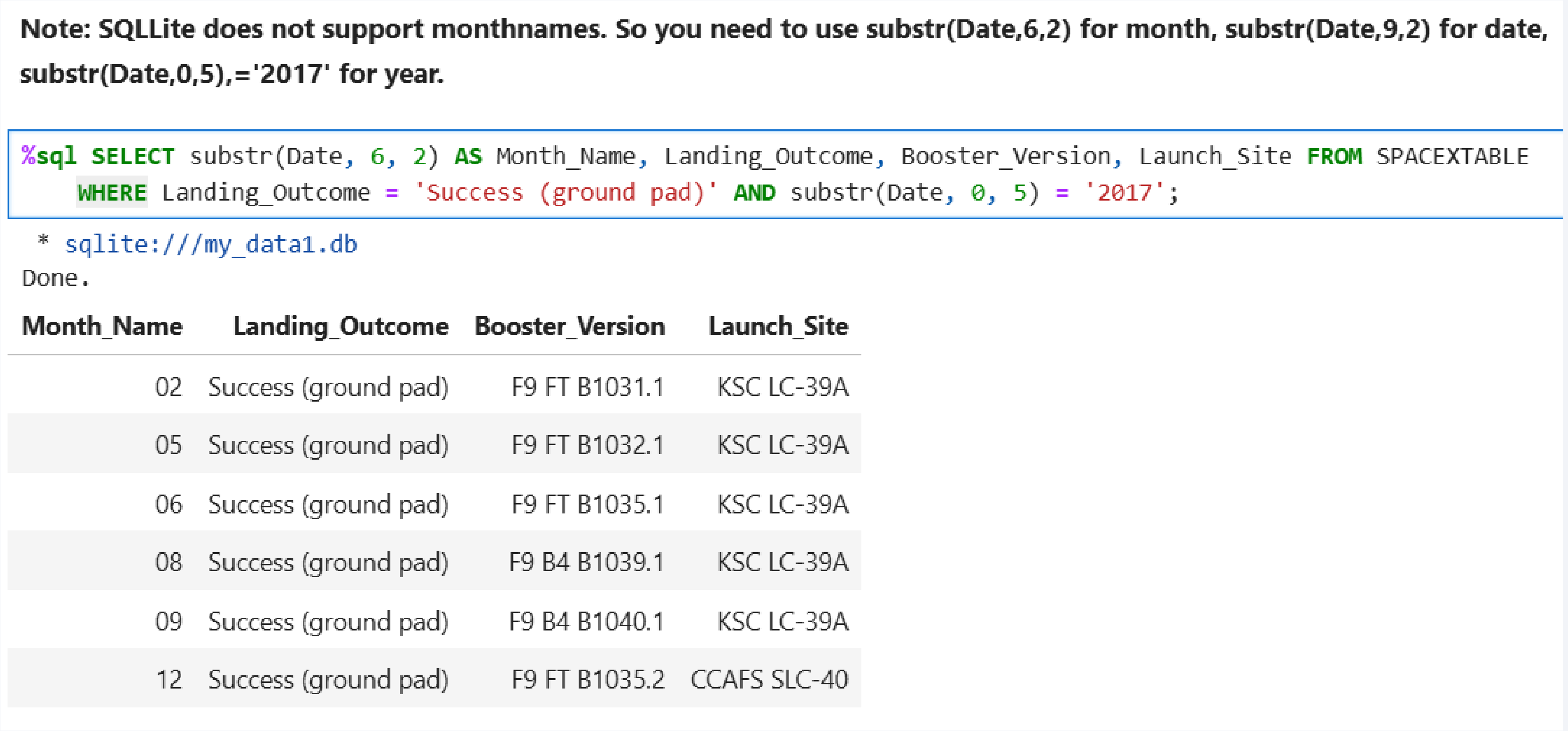
```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE);
```

```
 * 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 |

- We can see the names of the booster which have carried the maximum payload mass by using this query

# 2017 Launch Records



**Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.**

```
%sql SELECT substr(Date, 6, 2) AS Month_Name, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
    WHERE Landing_Outcome = 'Success (ground pad)' AND substr(Date, 0, 5) = '2017';
```

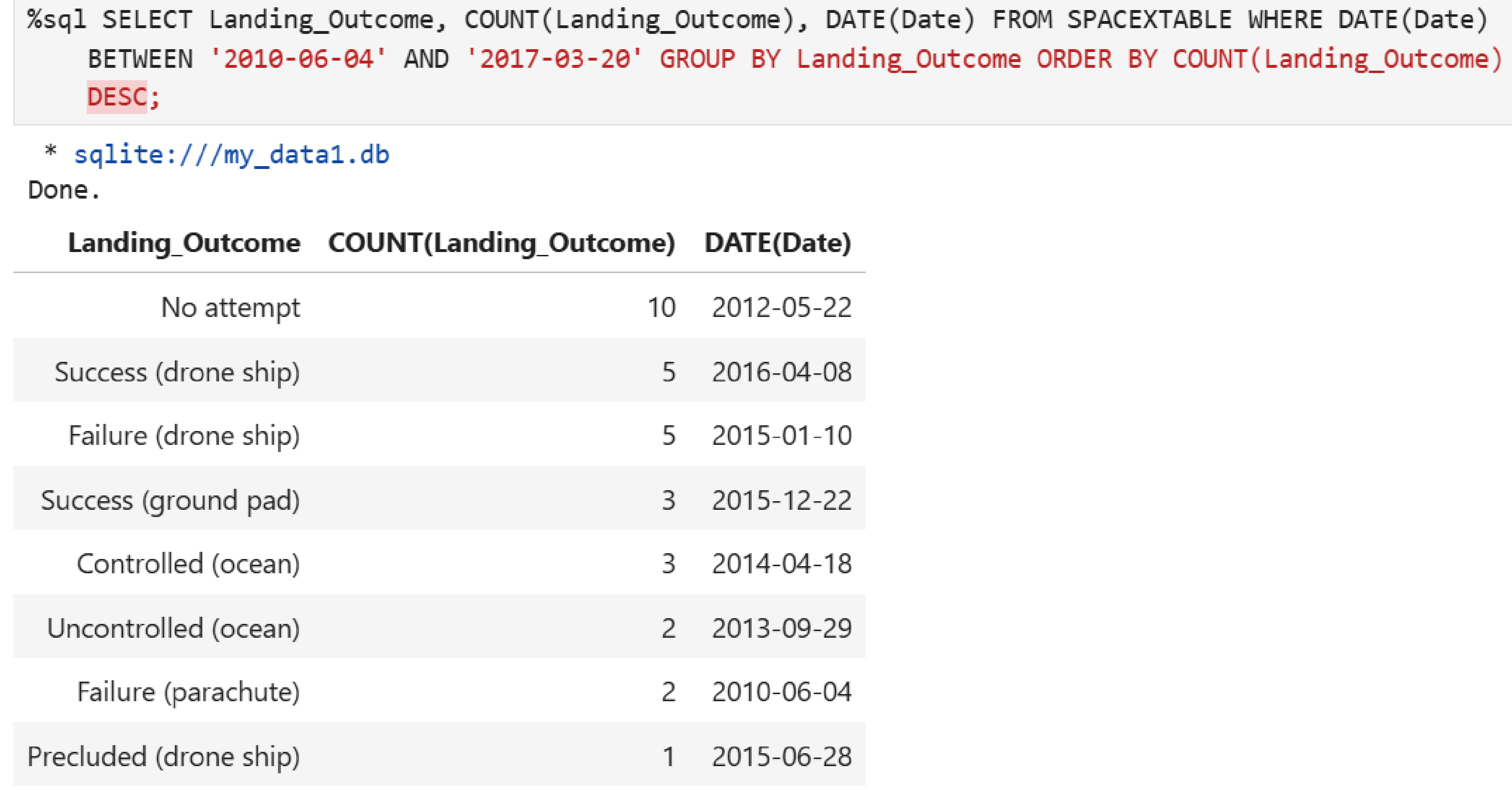
```
 * sqlite:///my_data1.db
Done.
```

| Month_Name | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|
| 02 | Success (ground pad) | F9 FT B1031.1 | KSC LC-39A |
| 05 | Success (ground pad) | F9 FT B1032.1 | KSC LC-39A |
| 06 | Success (ground pad) | F9 FT B1035.1 | KSC LC-39A |
| 08 | Success (ground pad) | F9 B4 B1039.1 | KSC LC-39A |
| 09 | Success (ground pad) | F9 B4 B1040.1 | KSC LC-39A |
| 12 | Success (ground pad) | F9 FT B1035.2 | CCAFS SLC-40 |

- We use SELECT and WHERE with the above requirements to print the list of the records for the months in the year 2017.
- We can see that most records are from the launch site KSA-LC-39A

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

- We obtain the 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 by using the following query



```
%sql SELECT Landing_Outcome, COUNT(Landing_Outcome), DATE(Date) FROM SPACEXTABLE WHERE DATE(Date)
    BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome)
    DESC;
```

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

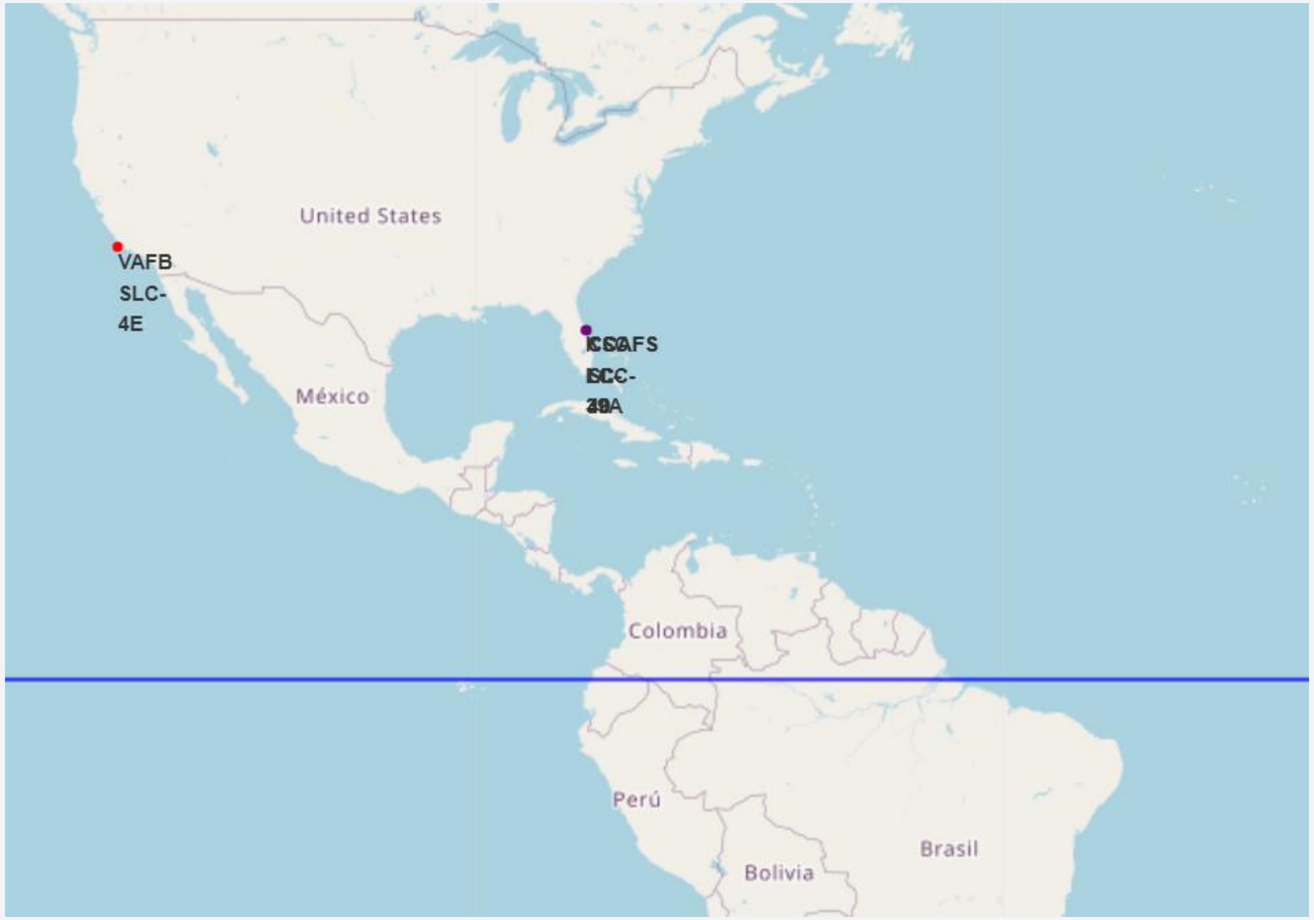
| Landing_Outcome | COUNT(Landing_Outcome) | DATE(Date) |
|---:|---:|---:|
| No attempt | 10 | 2012-05-22 |
| Success (drone ship) | 5 | 2016-04-08 |
| Failure (drone ship) | 5 | 2015-01-10 |
| Success (ground pad) | 3 | 2015-12-22 |
| Controlled (ocean) | 3 | 2014-04-18 |
| Uncontrolled (ocean) | 2 | 2013-09-29 |
| Failure (parachute) | 2 | 2010-06-04 |
| Precluded (drone ship) | 1 | 2015-06-28 |

Section 3

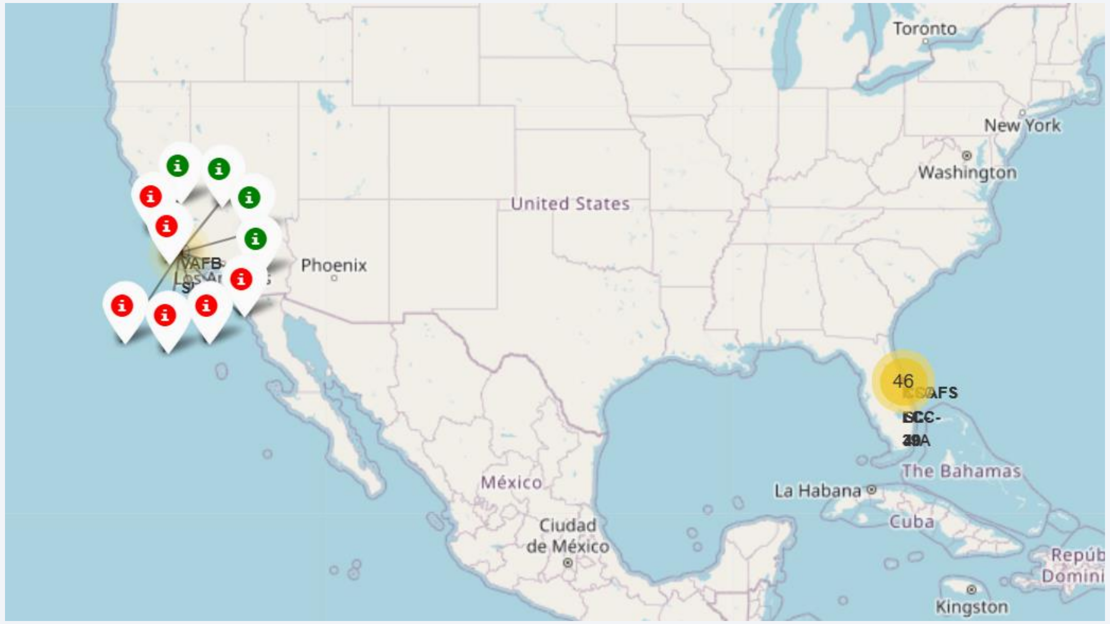
# Launch Sites Proximities Analysis

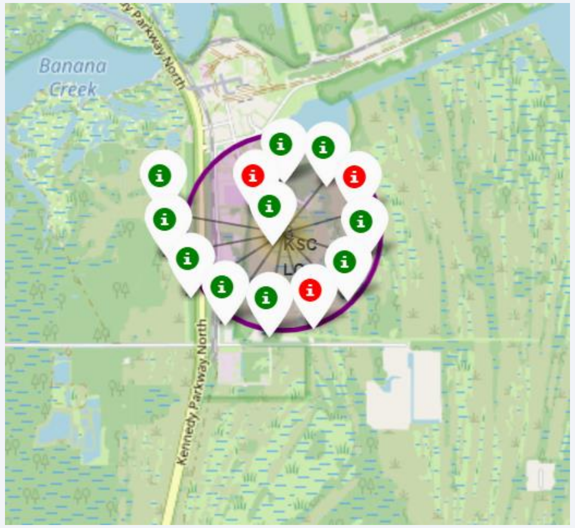
# Launch Sites Location

- All four launch sites seemed to be located on the proximity of United States of America coasts. This might be for safety reasons
- All four launch sites are also located fairly close to the equator line



# Launch Success Markers



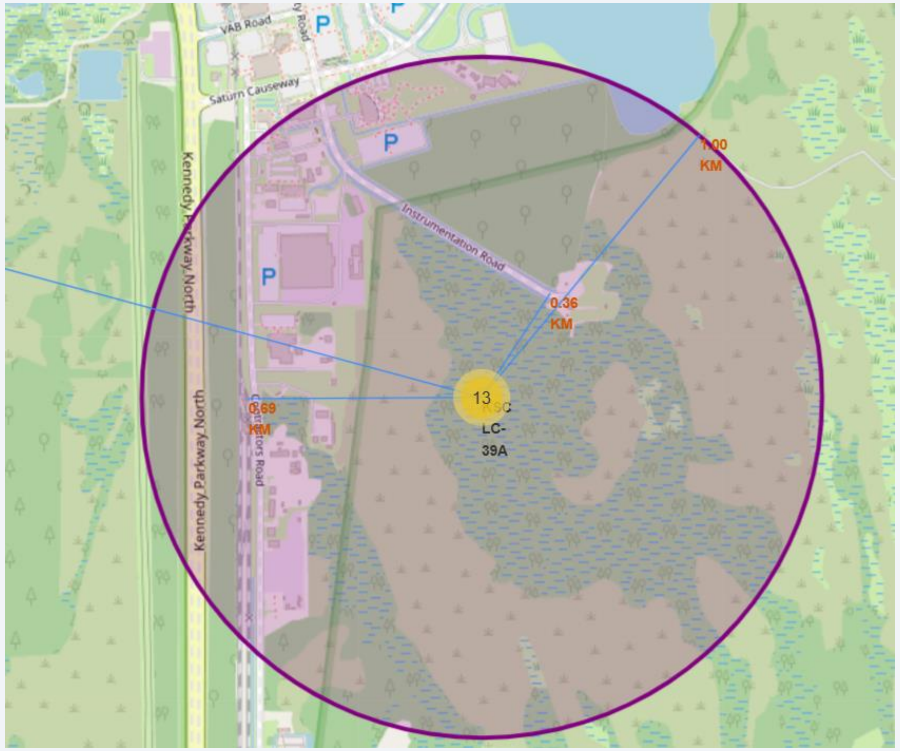


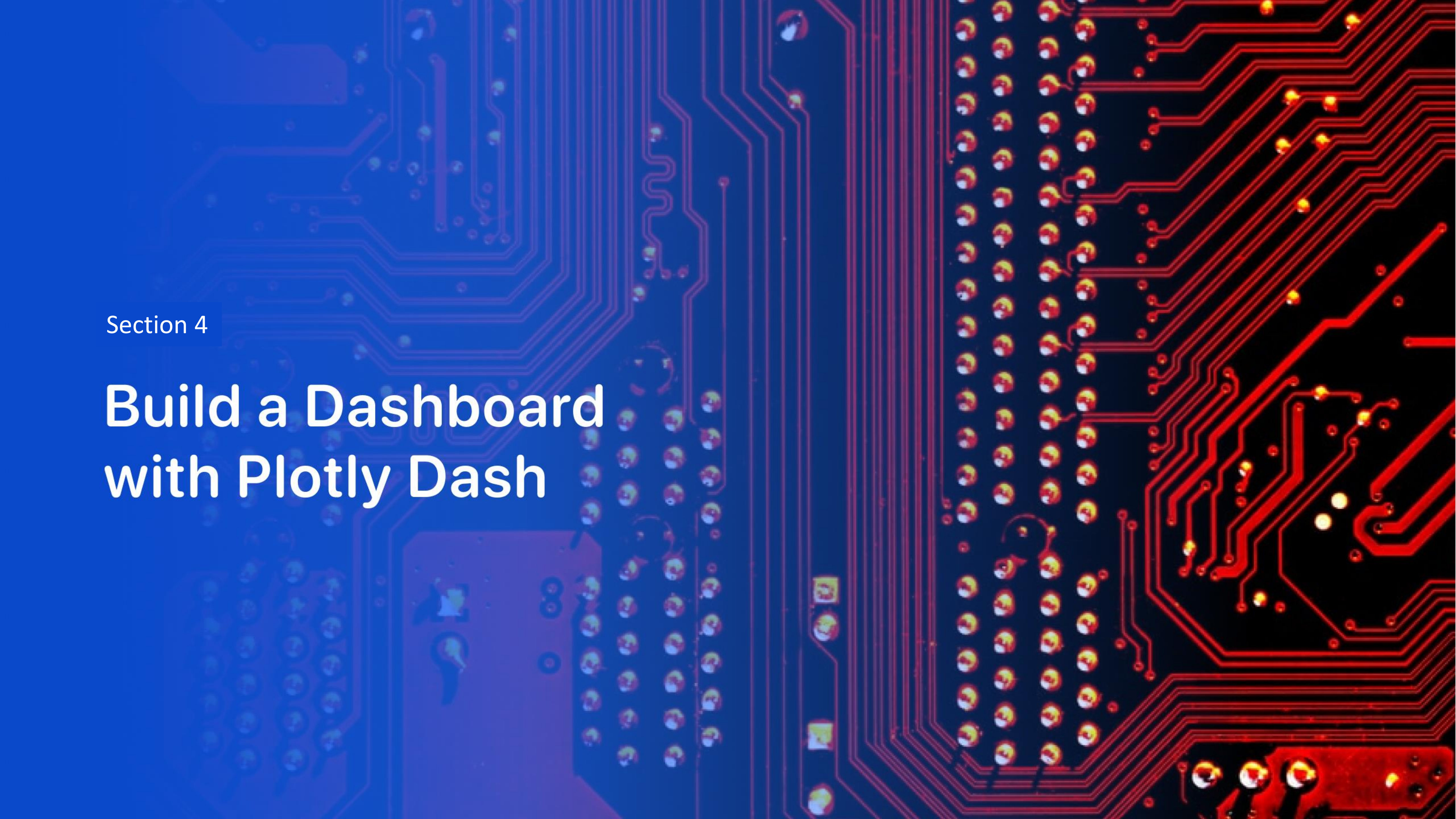
- We add color coded launch markers to each sites to see which sites have high success rates (GREEN for successful launches, RED for unsuccessful launches)

- The KSC LC 39A launch site seemed to have a high success rate

# Launch Sites Proximities

- We further explore the proximities of the chosen launch site, the KSC LC 39A
- The KSC LC 39A:
  - Is in very close proximity to railroad (0.69 km) and highway (0.36 km), for ease of access
  - is in close proximity to coastline (1.004 km), presumably for safety reasons
  - keeps a certain distance away from nearest city or highly populated area (16.48 km), for safety reasons
- We find out that these observations can also be generalized to other three launch stations.



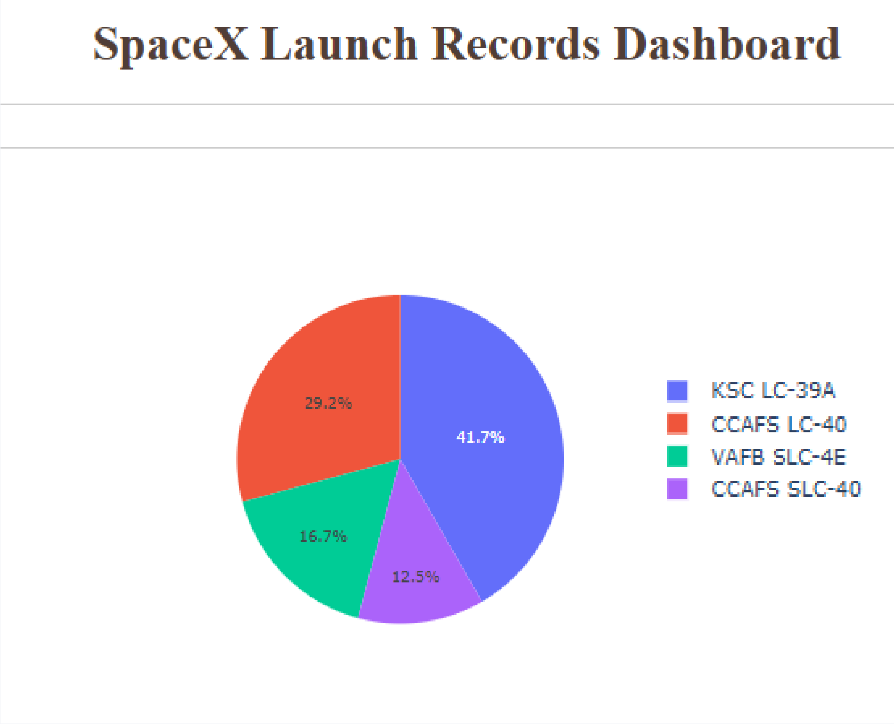


Section 4

# Build a Dashboard with Plotly Dash

# Launch Success Count for All Sites

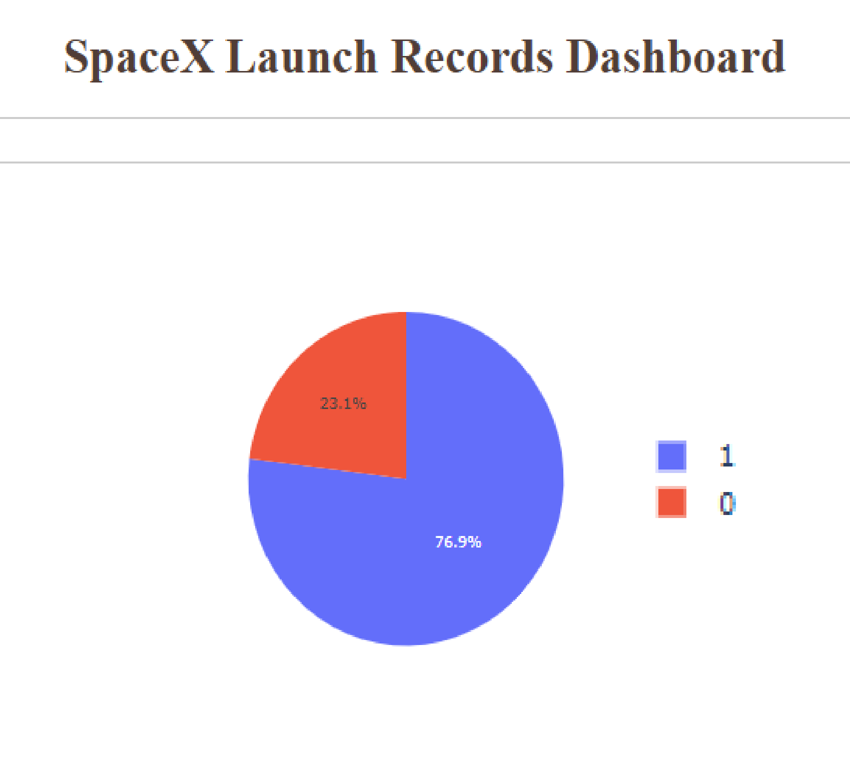
- Exploring the dashboard created with Plotly Dash, we can see that the KSC LC-39A launch site has the largest successful launches and also the highest launch success rate



SpaceX Launch Records Dashboard

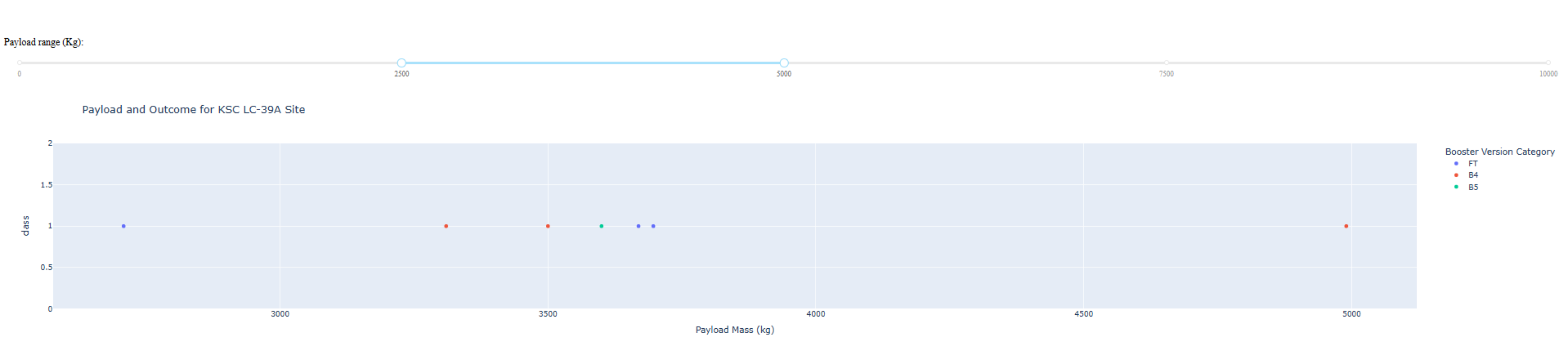
# Launch Site with the Highest Launch Success Ratio

- We can see that the KSC LC-39A launch site also has the highest launch success rate (76.9%)



SpaceX Launch Records Dashboard

# Most Optimal Payload for Successful Launch Outcome (KSC Launch Site)



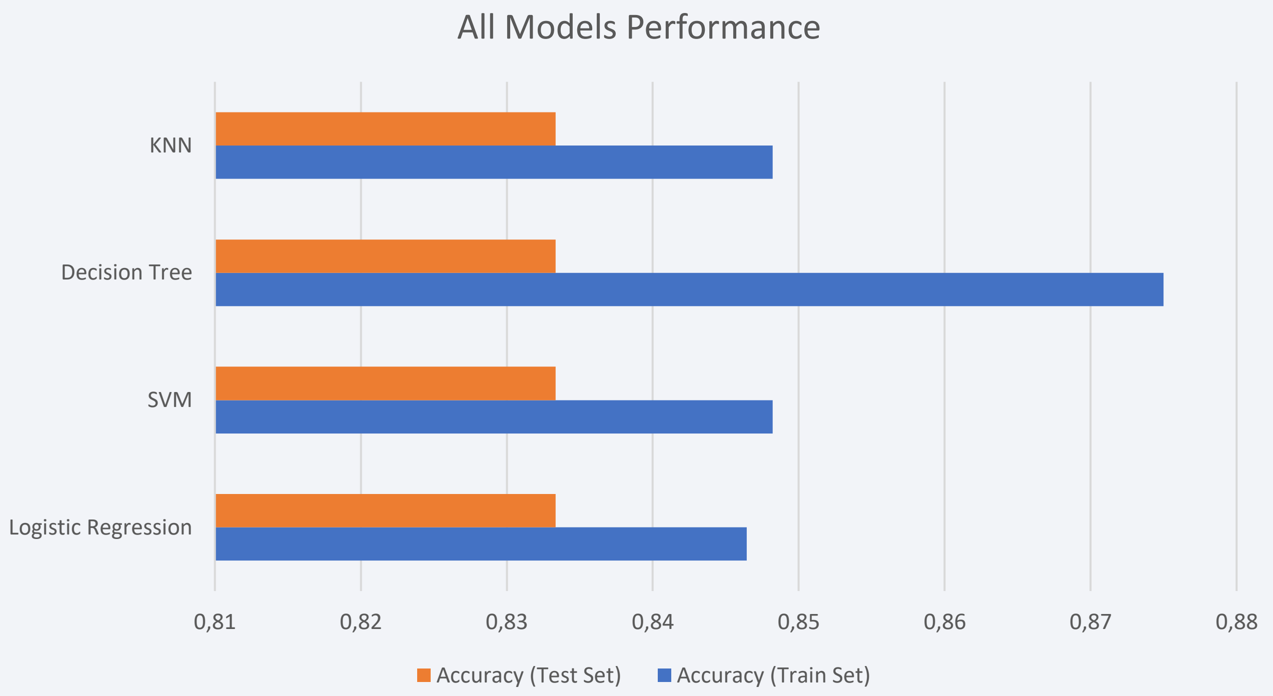
- For the chosen Launch Site (KSC LC 39A):
  - Payload range 2500-5000 has the highest launch success rate (100%)
  - while payload range 5000-7500 appears to have the lowest launch success rate
  - F9 Booster "FT" seemed to have the highest launch success rate

Section 5

# Predictive Analysis (Classification)

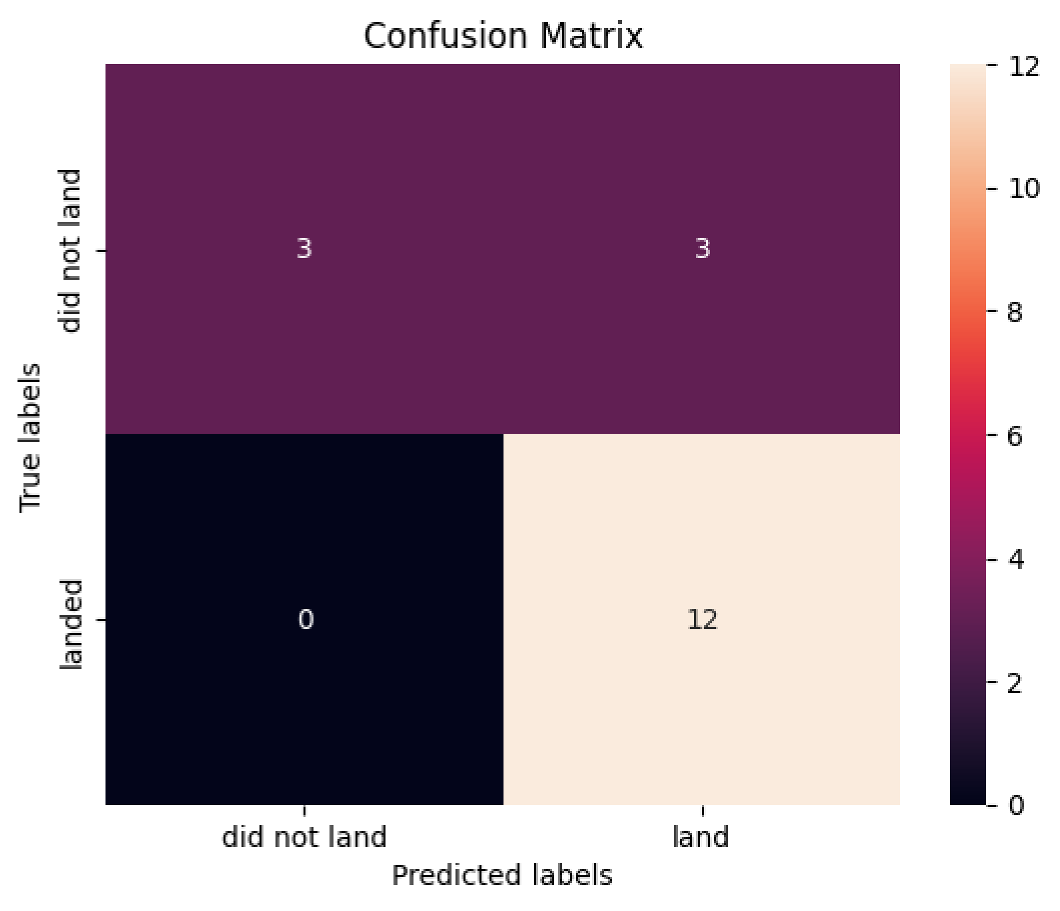
# Classification Accuracy

- On train data, decision tree classifier performs the best
- Considering the models' performances on test data, all models (Logistic Regression, SVM, Decision Tree, KNN) performs at the same level.
- The researcher chooses the Decision Tree model as the best model as it has the best performance on test data, and also performs the best on train data



# Confusion Matrix

- The decision tree model can classify the landing outcome relatively well
- However, there are still some cases of false positives found, even with the hyperparameter tuning stage



Summary

# CONCLUSION

# Conclusions

- There are several variables that seemed to influence launch outcome: flight number (date), launch site, payload, orbit type, and booster
- The yearly average success rate has an increasing trend in general for the time range of 2013-2019
- Launch sites tend to keep a certain distance away from cities (for safety reasons), but is usually in close proximity to railways and highways (for easy access), and also located in a very close proximity to coastline (for safety reasons)
- Out of all launch sites, KSC LC-39A has the highest total successful launches and success rate. The payload range of 2500-5000kg with the F9 Booster "FT" seemed to be most optimal for this launch site
- The Decision Tree Classifier is the best performing method found to predict the launch outcome

# Appendix

- All resources and files regarding this project can be found at https://github.com/nabila-rahmah/IBM-DS-Capstone-Project/tree/main

# Thank you!