IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jato Daniel
29th March, 2023

Outline

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

Executive Summary

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.

This capstone project is a research to determine and predict if the first stage of the SpaceX Falcon 9 rocket will land successfully.

The data was collected from web and SpaceX REST API. The data set was wrangled to deal with missing number, standardized and presented in a data frame for machine learning.

Visual exploratory data analysis was performed to better understand the trends of the features and to establish existing relationships.

Machine learning pipeline was built to predict the landing outcomes using logistic regression, support vector machine, decision tree and K-nearest neighbors (KNN)

The decision tree classifier proved to be the best model with a training accuracy of 88.9% and a test accuracy of 83.3%.

Introduction

- Project background and context

Companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Blue Origin manufactures sub-orbital and orbital reusable rockets. Spaces X's Falcon 9 launch like regular rockets. Perhaps the most successful is SpaceX. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises 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. Therefore, the aim of this project is to determine if the first stage will land with the aid of machine learning.

Problems you want to find answers

□ To determine if the first stage will land will land Successfully

□ To determine the price of each launch.

□ To determine the most important features that will determine a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected from the SpaceX REST API and web scraping from wikipedia
- Perform data wrangling
  - Data was wrangled to determine success rate
- Perform exploratory data analysis (EDA) using visualization and SQL

  Various graphs such as scatter charts and bar charts were created to trends and relationships between features.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

❑ The data was collected from web and SpaceX REST API

❑ Get request was performed using the requests library to obtain the launch data.

❑ The response data was in the form of json object which was converted to a dataframe and normalized.

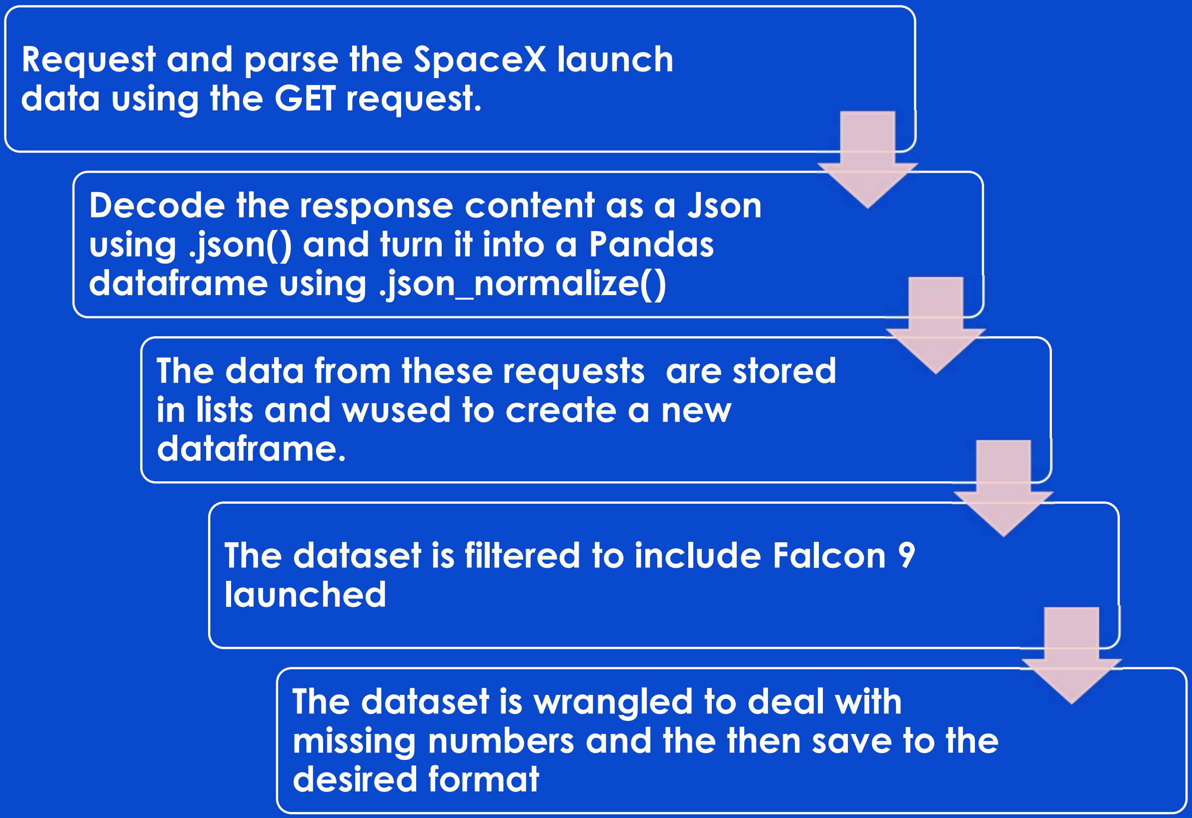
❑ Data was also obtained from Wikipedia using webscraping

❑ Columns and variable names were extracted from the HTML table header.

❑ These data were presented in dataframes

❑ The data was cleaned and presented in the desired format

Data Collection – SpaceX API



Request and parse the SpaceX launch data using the GET request.

Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()

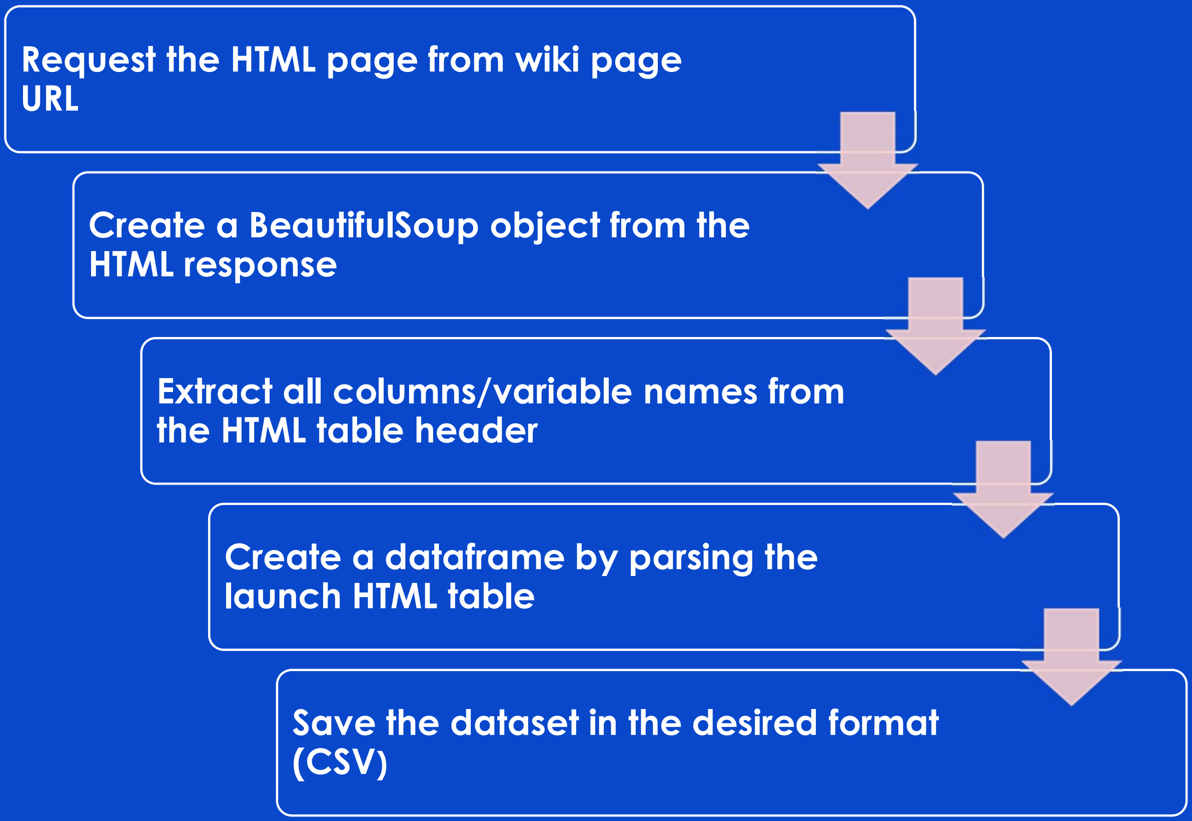
The data from these requests are stored in lists and wused to create a new dataframe.

The dataset is filtered to include Falcon 9 launched

The dataset is wrangled to deal with missing numbers and the then save to the desired format

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/Data%20Collection%20API.ipynb

Data Collection - Scraping



Request the HTML page from wiki page URL

Create a BeautifulSoup object from the HTML response

Extract all columns/variable names from the HTML table header

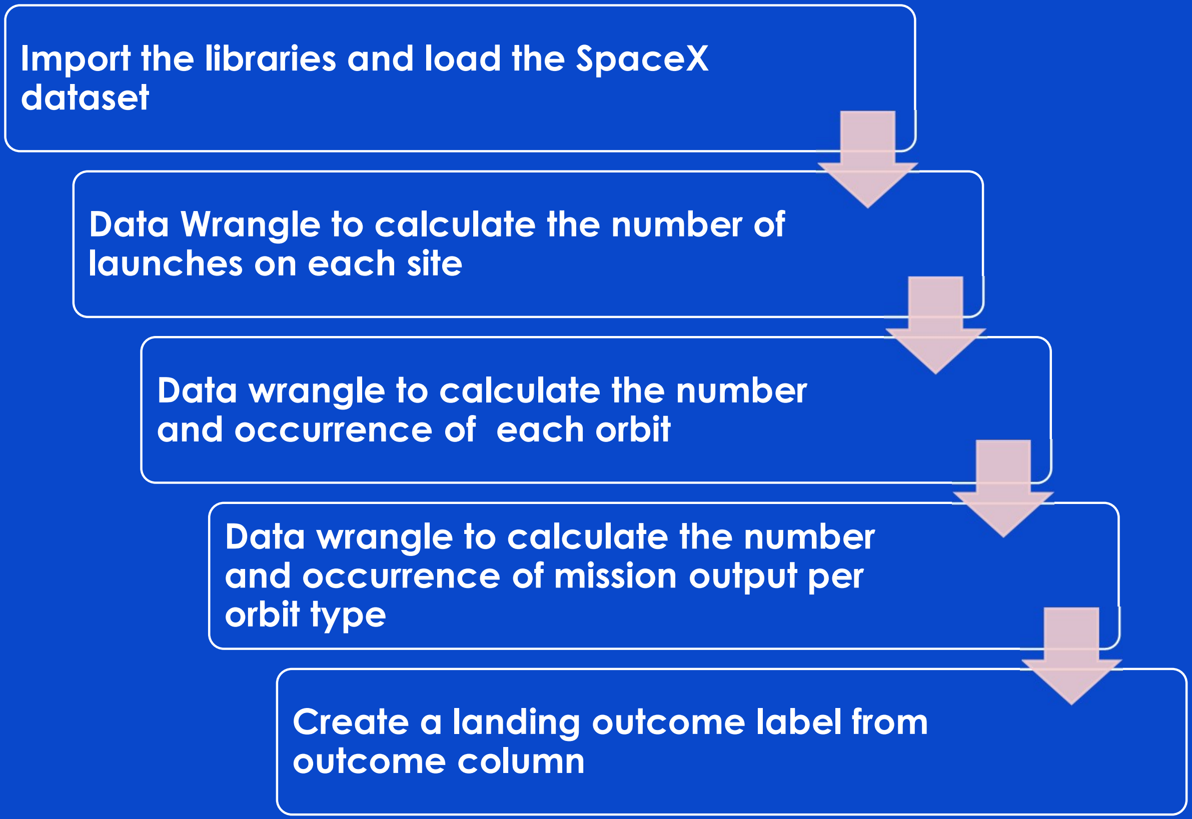
Create a dataframe by parsing the launch HTML table

Save the dataset in the desired format (CSV)

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/Webscraping.ipynb

Data Wrangling

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/Wrangling.ipynb



Import the libraries and load the SpaceX dataset

Data Wrangle to calculate the number of launches on each site

Data wrangle to calculate the number and occurrence of each orbit

Data wrangle to calculate the number and occurrence of mission output per orbit type

Create a landing outcome label from outcome column

EDA with Data Visualization

https://github.com/Dedonruks/Data-Science-Capstone-Project/blob/main/EDA%20Data%20Visualization.ipynb

- Visual Exploratory data analysis was performed to better understand the trends of the features and to establish the existing relationships in the dataset
- Catplot was used in order to visualize the relationship between flight number and launch site. We discovered that no rocket launched for heavy payload mass greater than 10000
- Bar chart was used to visualize the relationship between success rate of each orbit
- Scatter plots were used to visualize the relationship between flight number and orbit type, payload and orbit type.
- Line plot was used to visualize the success yearly trend

EDA with SQL

Summary of the SQL queries performed

□ Names of the unique launch site in the space mission

□ 5 launch sites that begin with the string CCA

□ Total payload mass carried by boosters launched by NASA (CRS)

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

□ Date when the first successful landing outcome in ground pad was achieved.

□ Names of boosters which have success in drone ship and have mass greater than 4000 but less than 6000

□ Total number of successful and failure mission outcomes

□ Rank of successful landing outcomes.

□ Records of month, failure landing outcomes, booster version and launch site for the year 2015

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/SQL%20EDA.ipynb

Build an Interactive Map with Folium

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/Launch%20Site%20Location%20with%20Folium.ipynb

❑ Folium Markers were used to mark all launch site on a map.

❑ Folium markers were used to mark the success/filed launches for each site on the map. Here we used a green marker for a successful launch and a red marker for a failed launch

❑ Folium markers were used to calculate the distances between a launch site to its proximities. Distances between coastlines and launch site were calculated.

❑ Folium circles to add a highlighted circle area with a text label on a specific coordinate and also used for each launch site on the map.

❑ Polylines between a launch site to the coastline point were drawn

Build a Dashboard with Plotly Dash

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/Plotly%20dash_interactivity.py

❑ The plotly dashboard was used to find more insights from the SpaceX dataset.

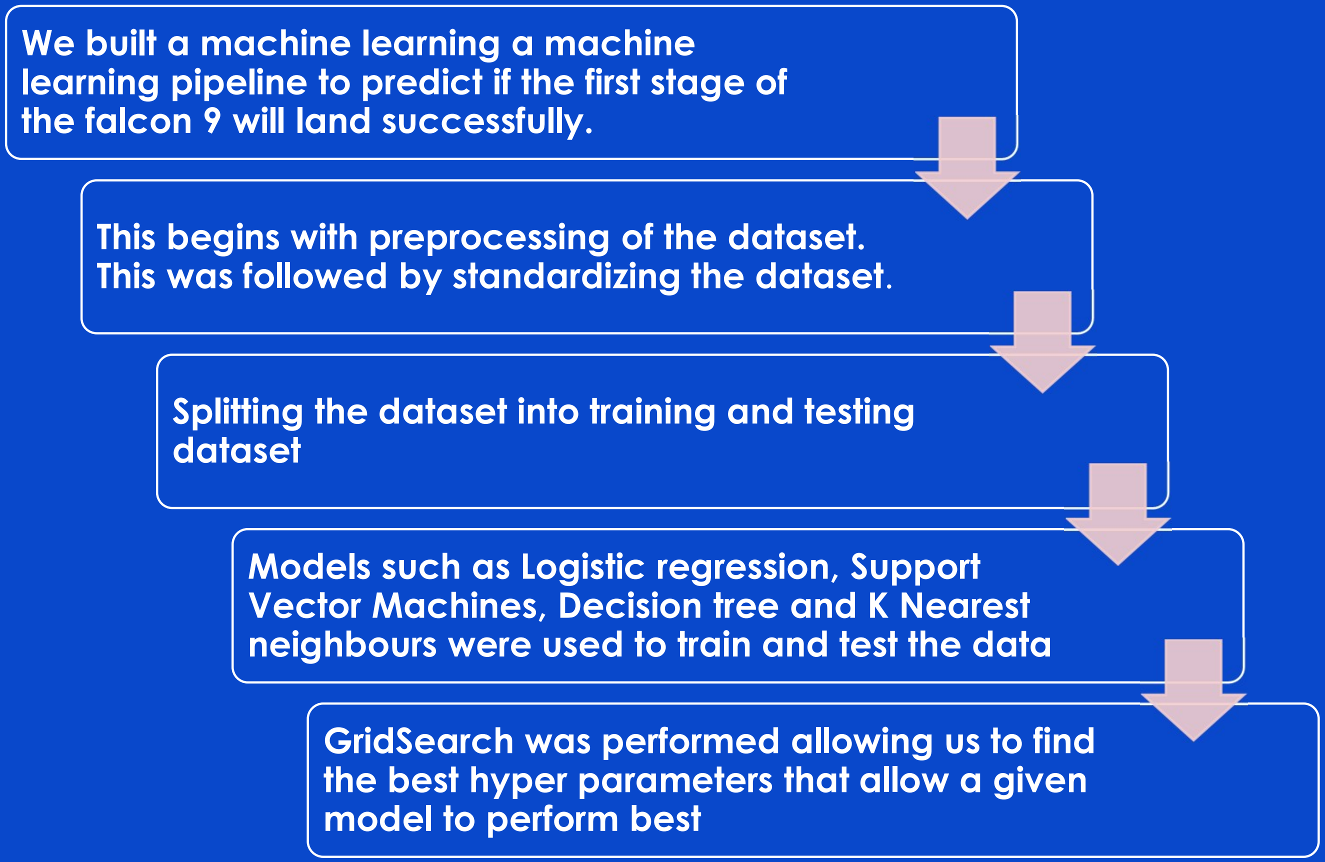
❑ Pie Chart was used to used to show the total successful launch by sites. This showed sites with largest successful launches and successful ratings.

❑ Scatter chart was also used to show the correlations between payload and success for all sites. This show payload range with highest and lowest success rate. Pie charts and scatter charts were used to visualize the launch records of SpaceX.

❑ Success rate of the various F9 booster version was also shown

Predictive Analysis (Classification)

https://github.com/Dedonrukks/Data-Science-Capstone-Project/blob/main/SpaceX%20Machine%20Learning%20Prediction.ipynb



We built a machine learning a machine learning pipeline to predict if the first stage of the falcon 9 will land successfully.

This begins with preprocessing of the dataset. This was followed by standardizing the dataset.

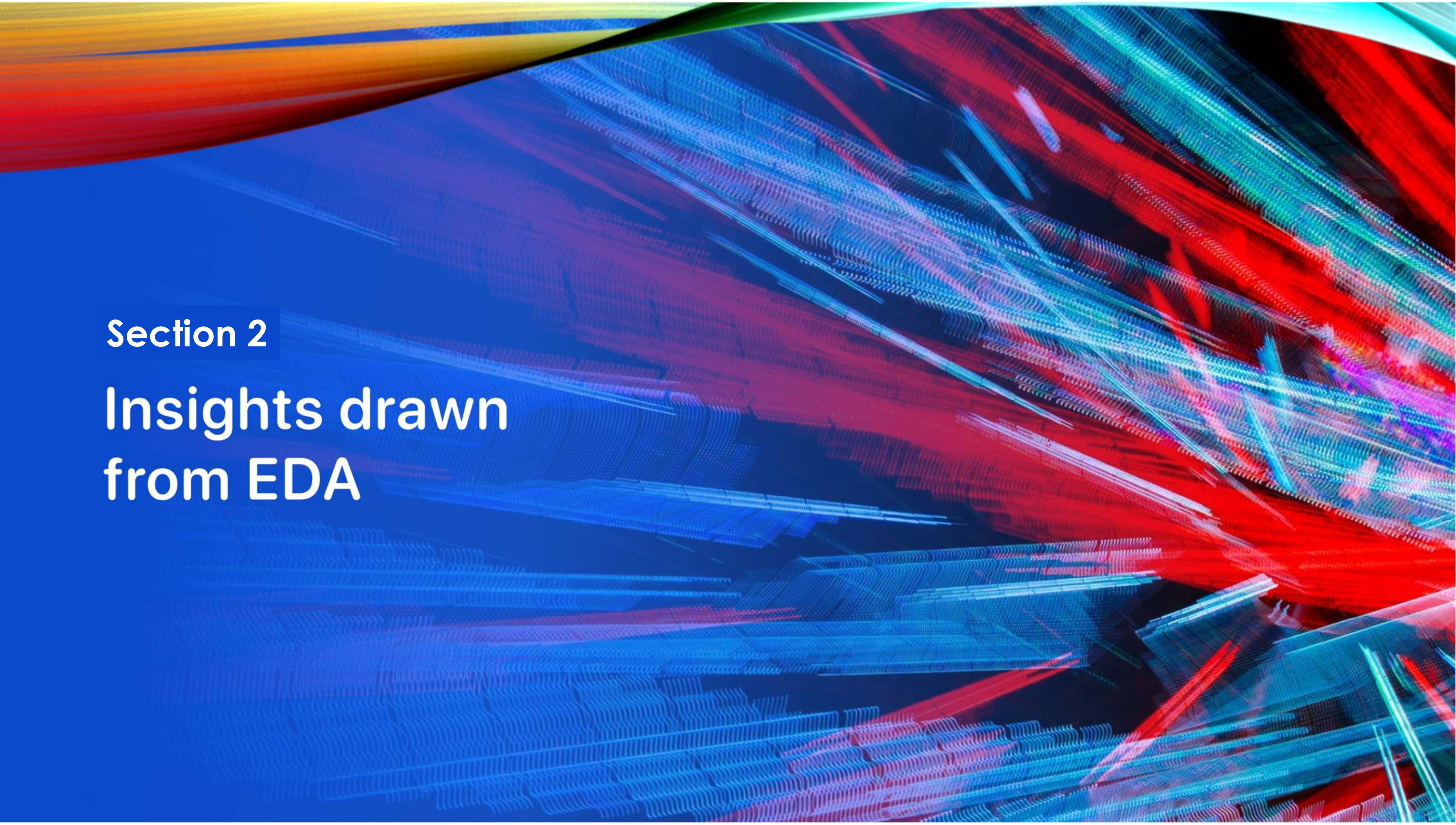
Splitting the dataset into training and testing dataset

Models such as Logistic regression, Support Vector Machines, Decision tree and K Nearest neighbours were used to train and test the data

GridSearch was performed allowing us to find the best hyper parameters that allow a given model to perform best

Results

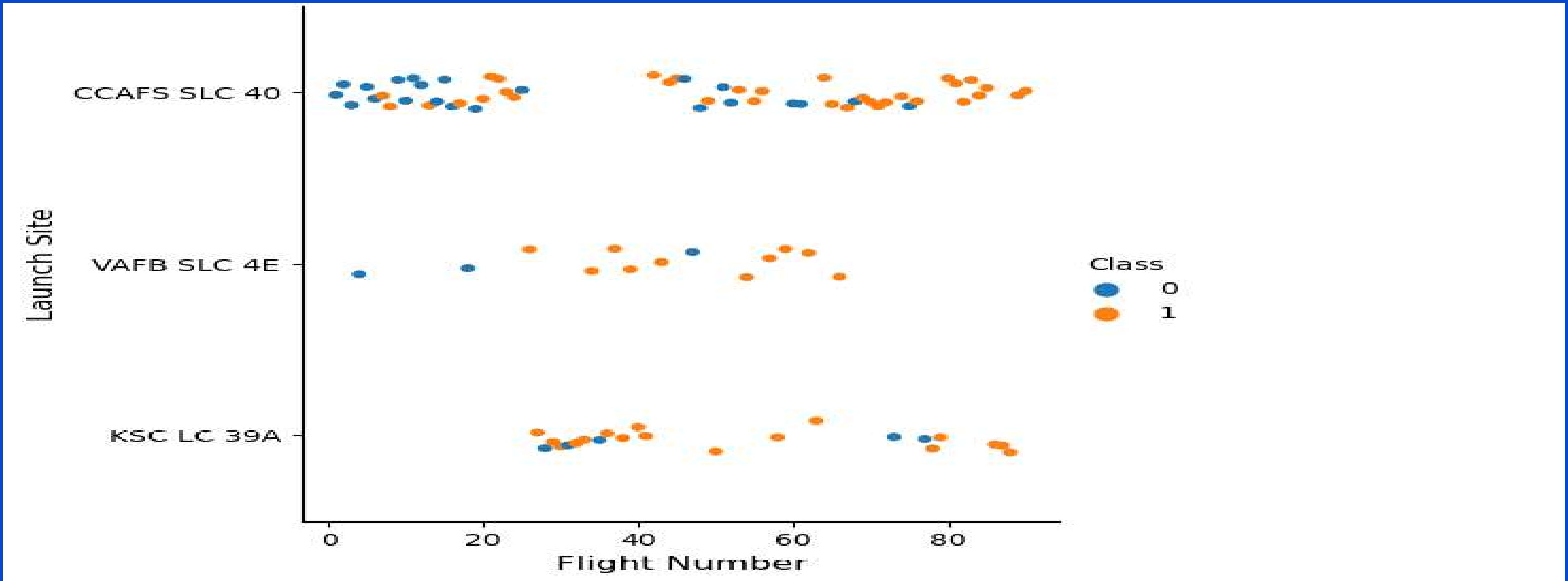
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results



Section 2

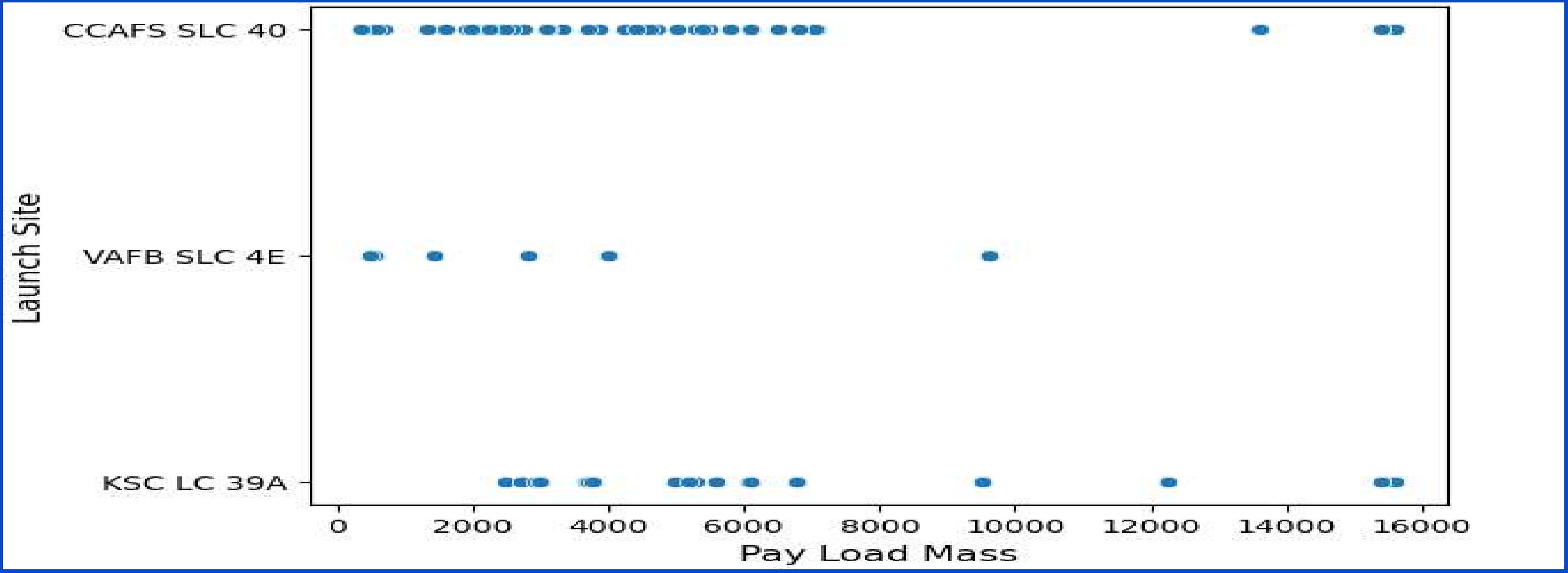
Insights drawn from EDA

Flight Number vs. Launch Site



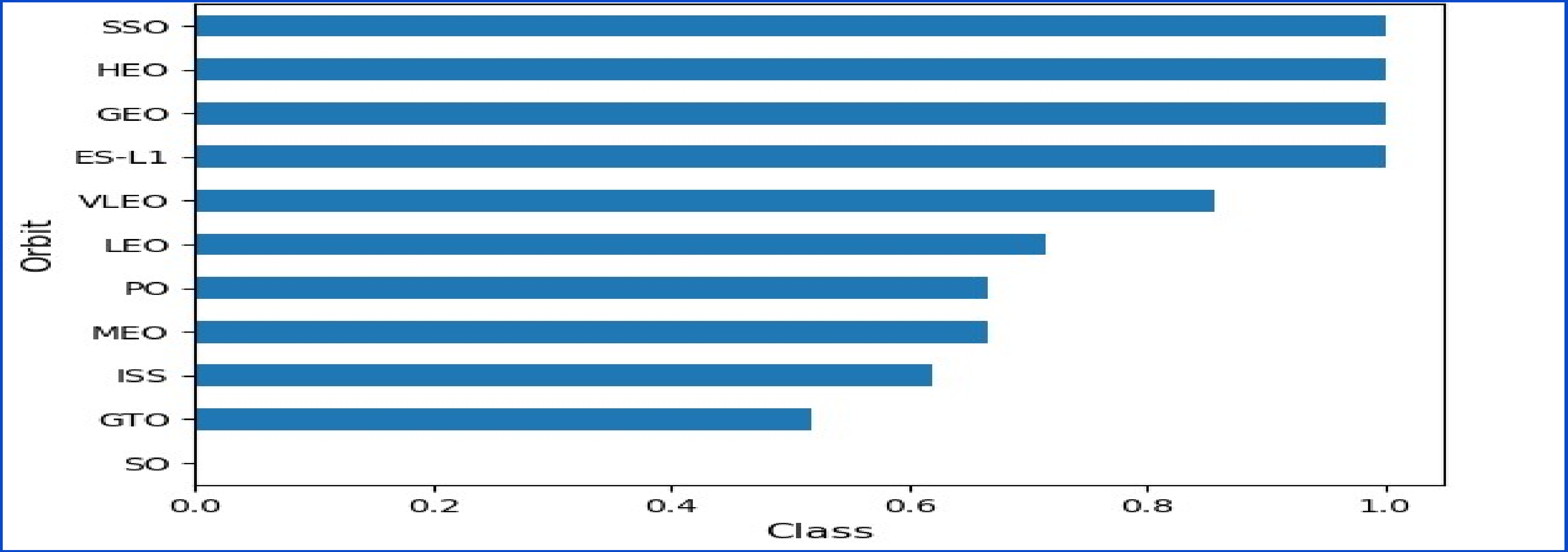
CCAFS SLC 40 appears to have the highest number of flight numbers and the highest number of success rate. Therefore, it seems the more the flight numbers, the greater the success rate.

Payload vs. Launch Site



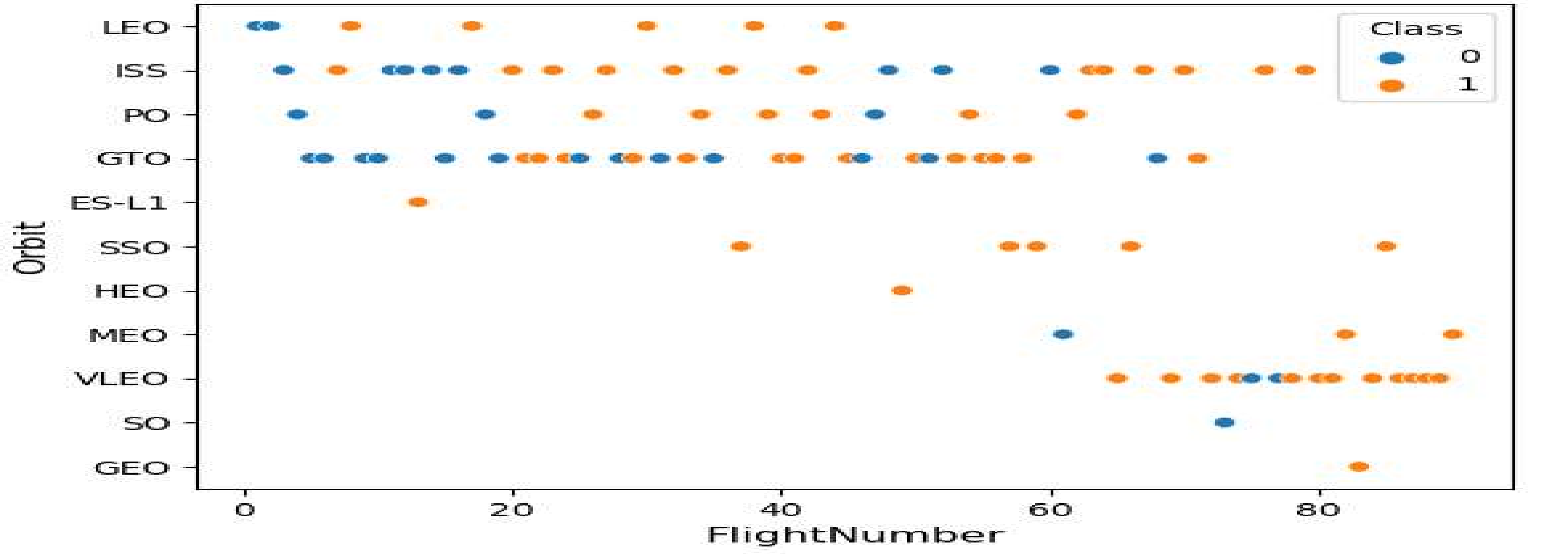
For the VAFB SLC 4E launch site, there are no rockets launch for heavy payload mass greater than 10000

Success Rate vs. Orbit Type



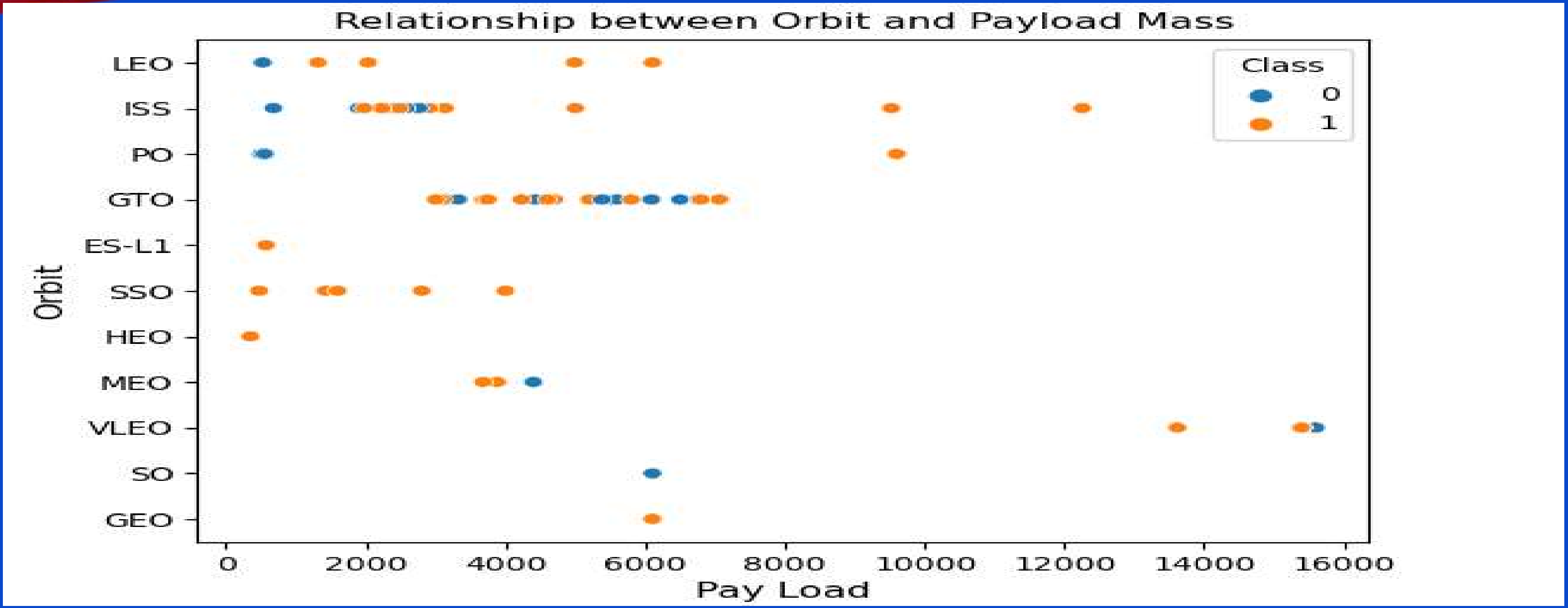
SSO, HEO, GEO and ES-L1 orbits had the highest success rate.

Flight Number vs. Orbit Type



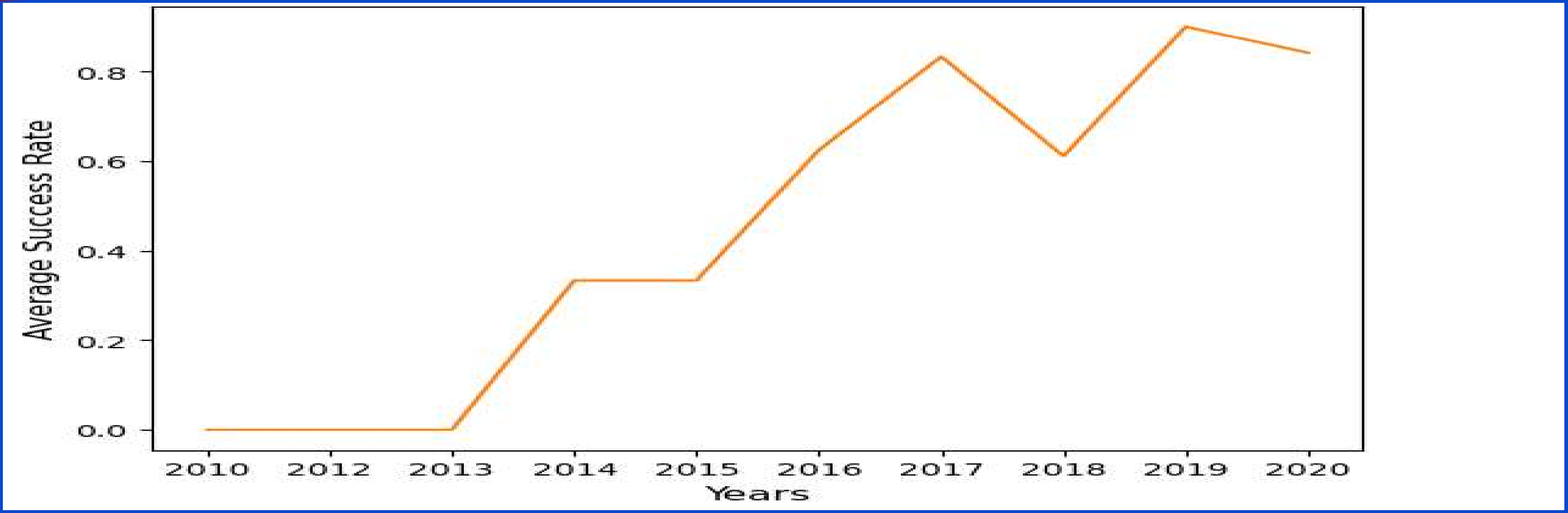
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



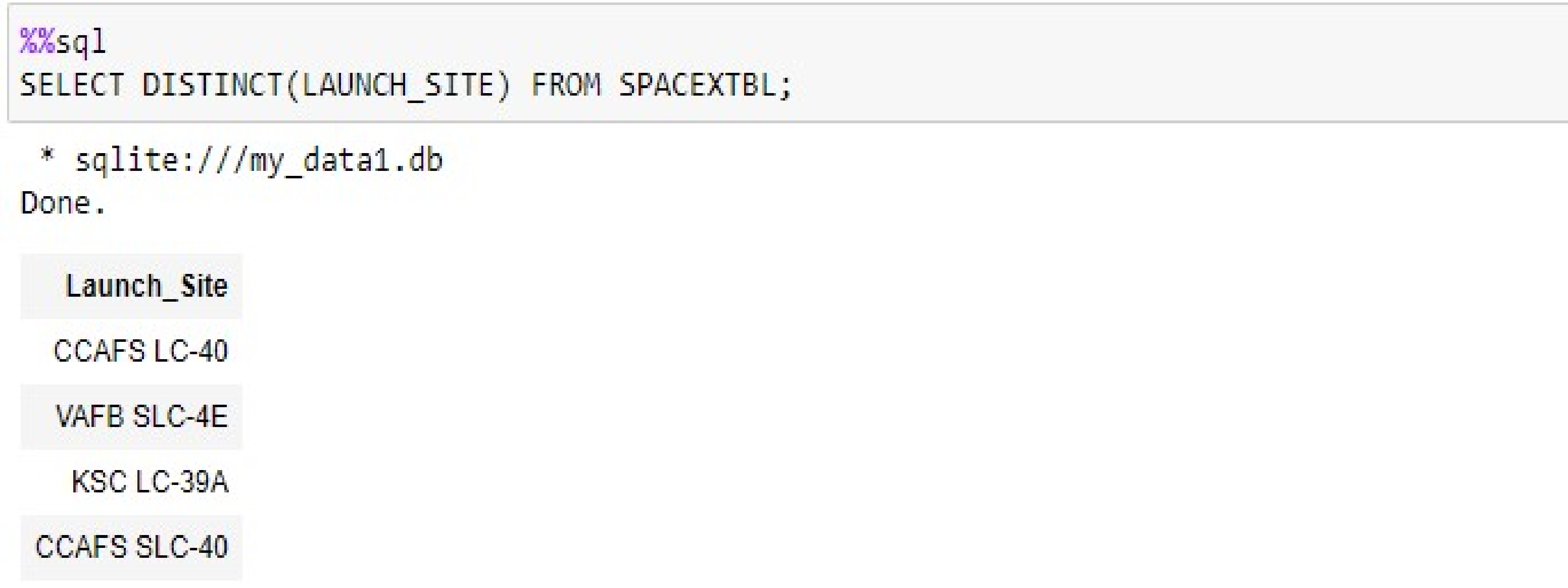
With heavy payloads the successful landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both successful landing rate and unsuccessful mission are both there here.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

All Launch Site Names



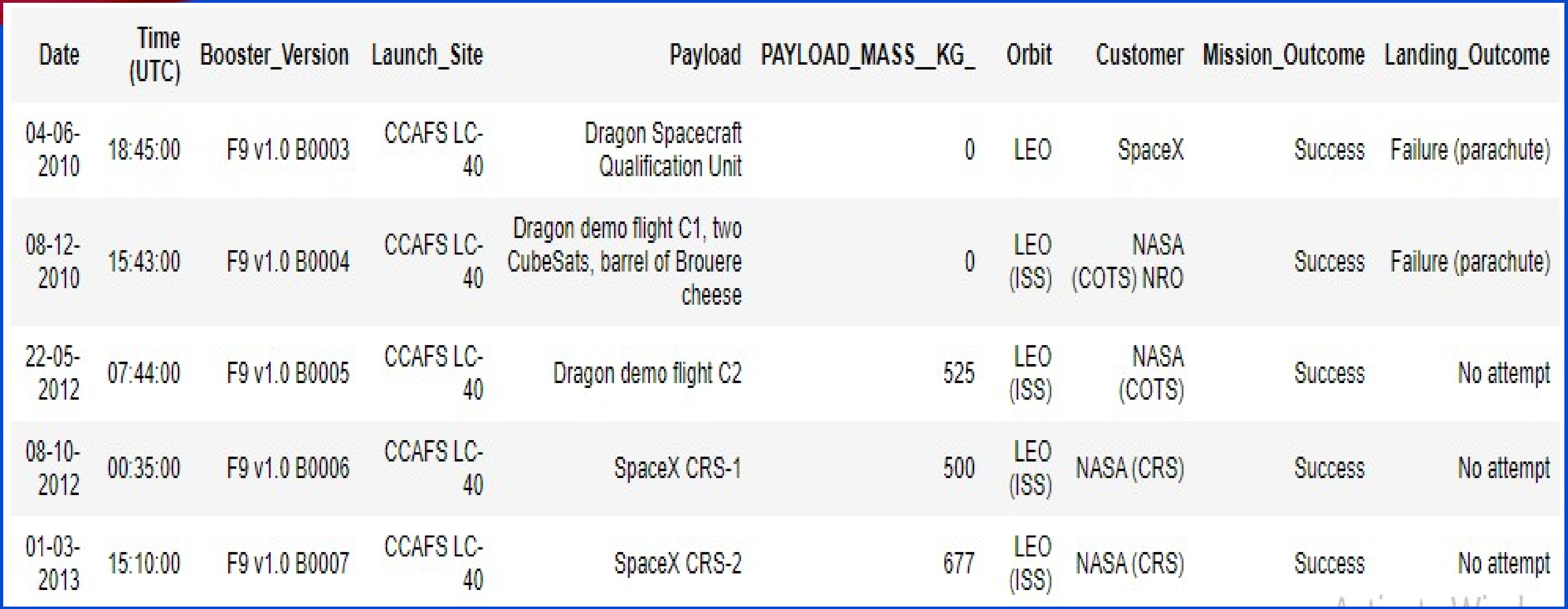
```
%%sql
SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

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

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

Query Explanation: The distinct keyword in the query statement was used to ensure that only unique launch site names were retrieved.

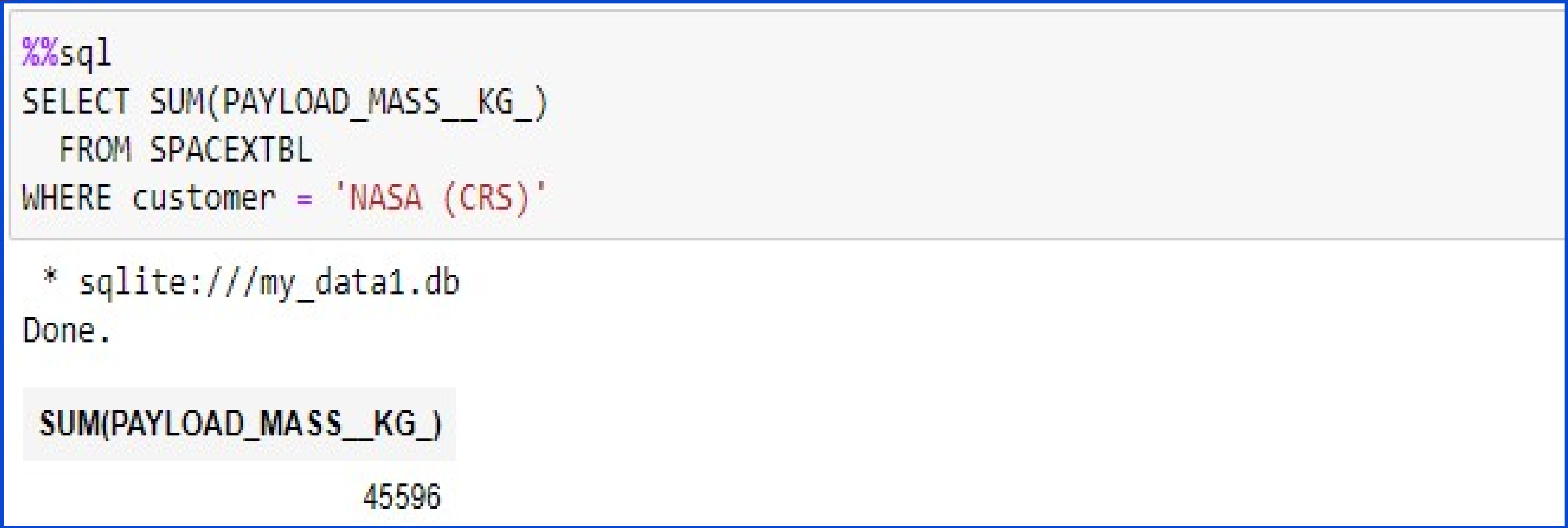
Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Query Explanation: We use the like, CCA and the limit keywords to ensure that only site names that begins with CCA and only first five records were retrieved.

Total Payload Mass



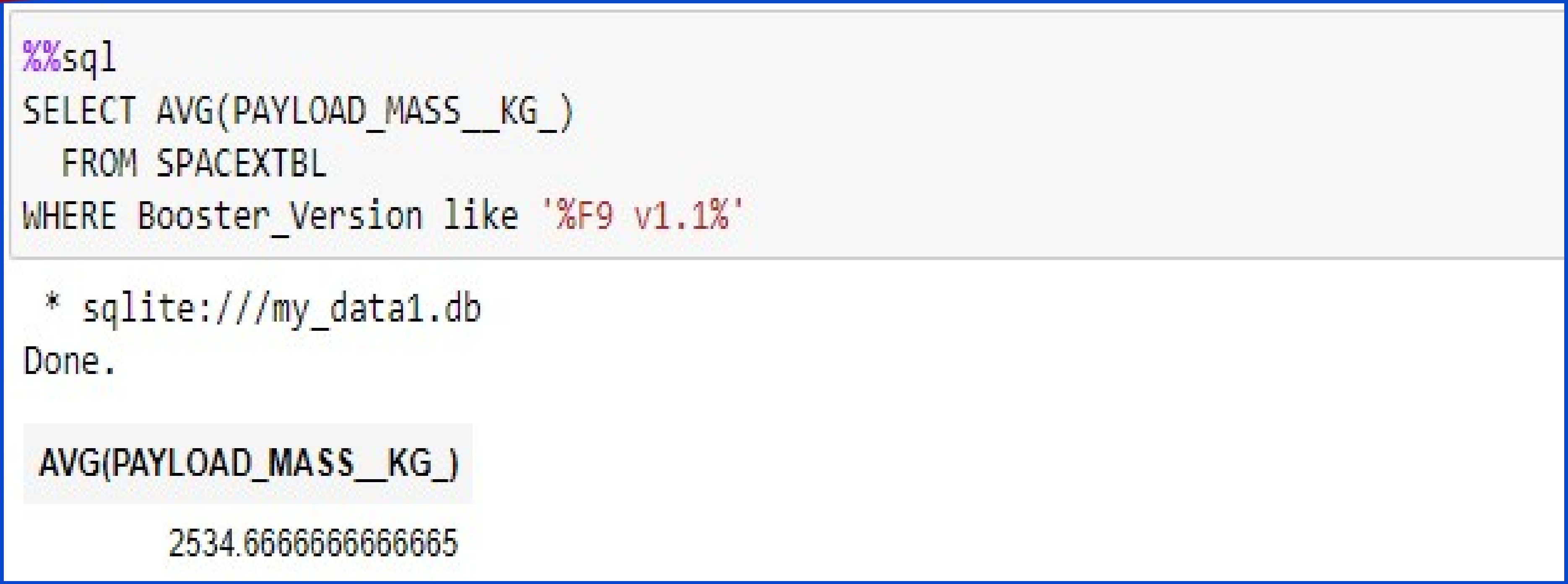
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
  FROM SPACEXTBL
WHERE customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596

Query Explanation: The sum function was used to aggregate the total payload mass while the where function was applied to filter the dataframe only for where customer is NASA (CRS)

Average Payload Mass by F9 v1.1



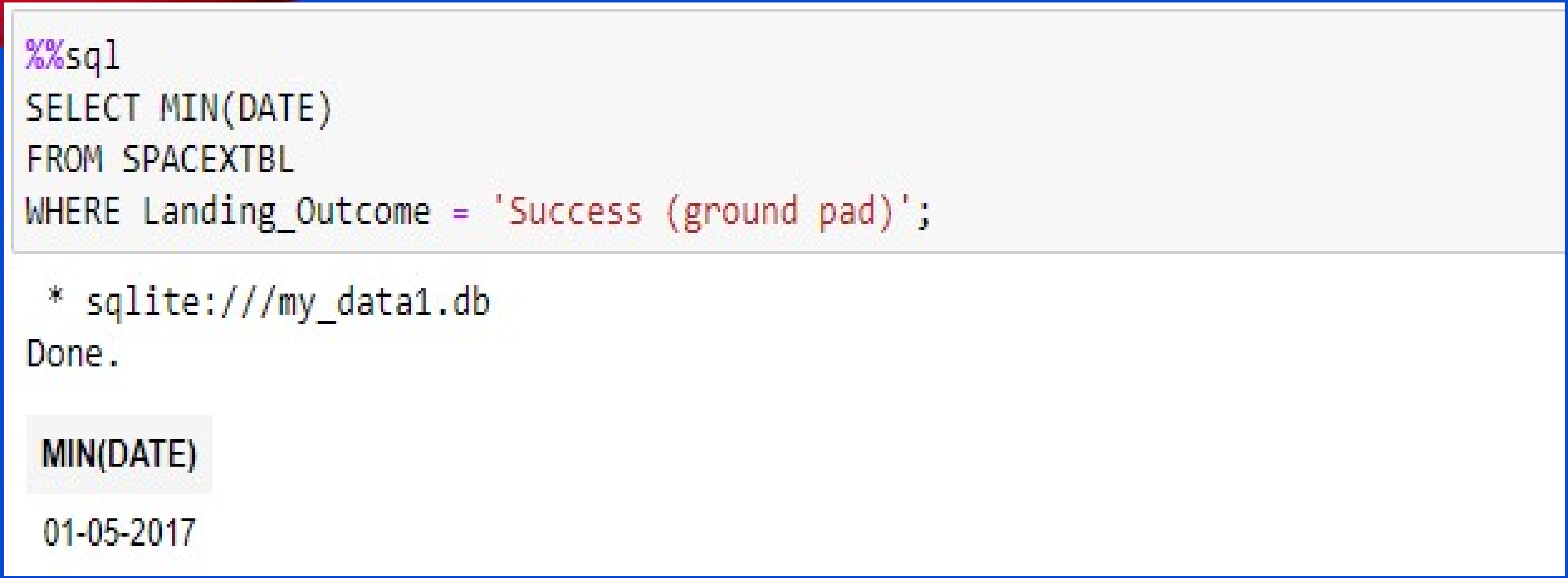
```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
  FROM SPACEXTBL
WHERE Booster_Version like '%F9 v1.1%'
```

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

Query Explanation: The AVG function was used to find the average payload mass while the where function was applied to filter the dataframe only where booster version contains F9 v1.1

First Successful Ground Landing Date



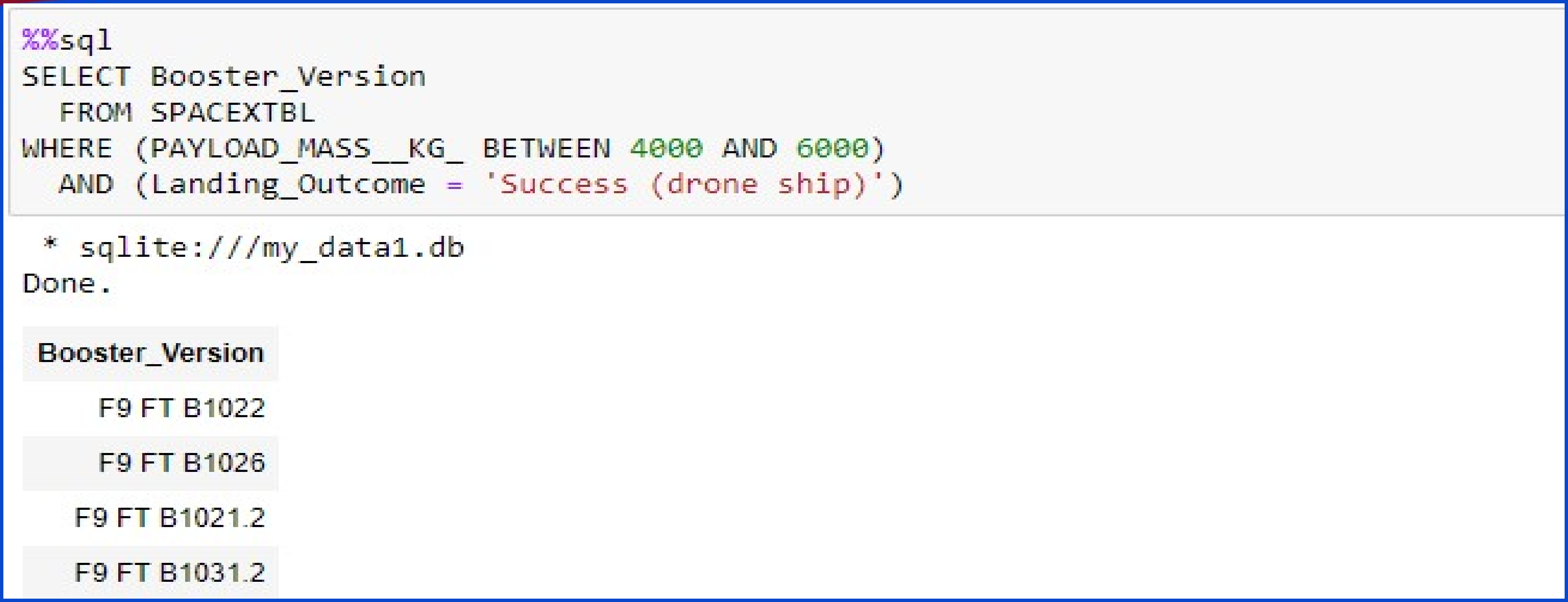
```
%%sql
SELECT MIN(DATE)
FROM SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)';
```

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

MIN(DATE)
01-05-2017

Query Explanation: The dates of the first successful landing outcome on ground pad was calculated by simply using the min function and then filtering the data frame using the where clause to obtain the data where landing outcome is success ground pad

Successful Drone Ship Landing with Payload between 4000 and 6000



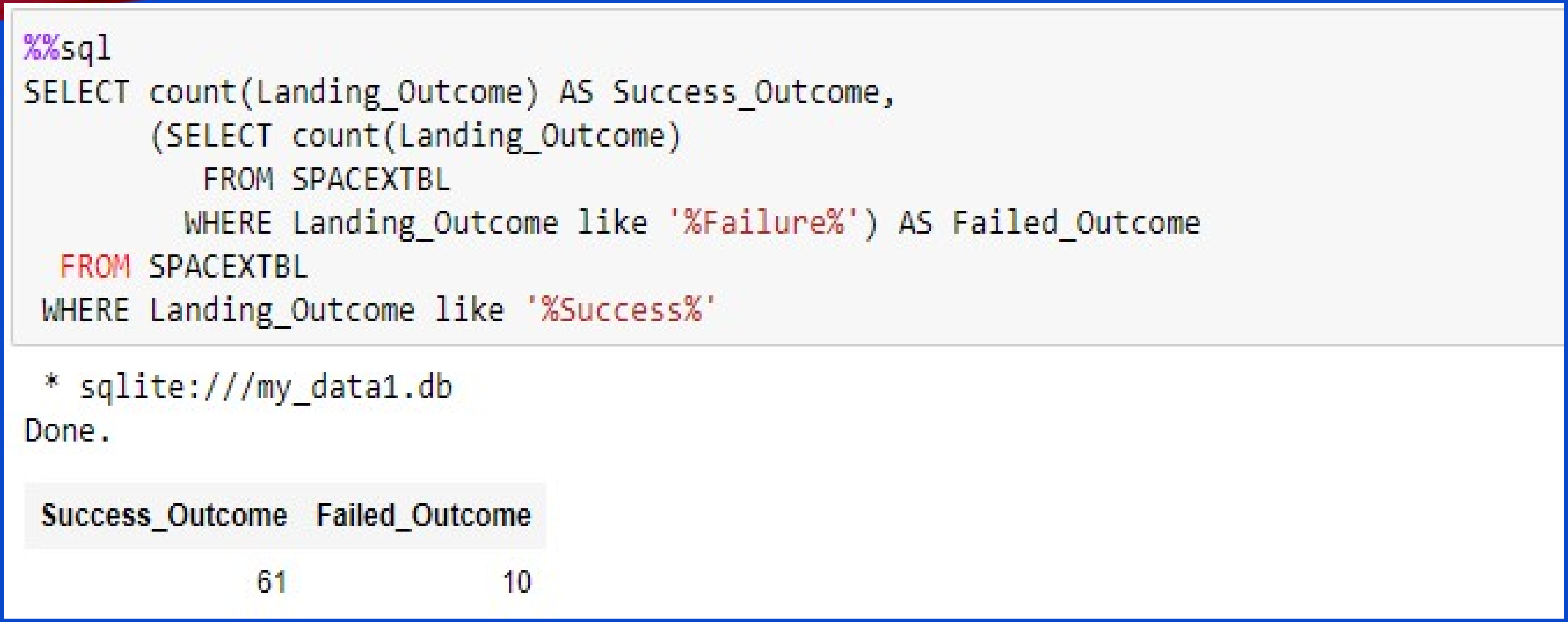
```
%%sql
SELECT Booster_Version
   FROM SPACEXTBL
WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
   AND (Landing_Outcome = 'Success (drone ship)')
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Query Explanation: The WHERE, BETWEEN & AND keywords were used in the query to list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



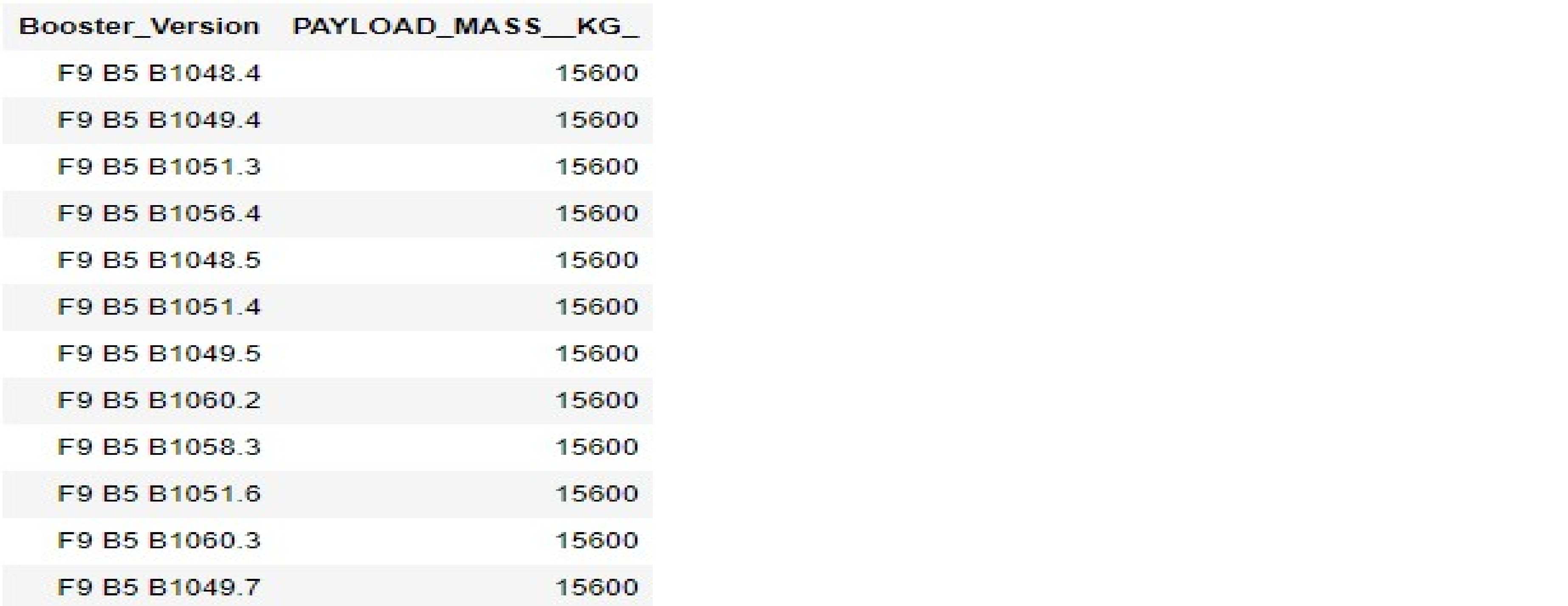
```
%%sql
SELECT count(Landing_Outcome) AS Success_Outcome,
       (SELECT count(Landing_Outcome)
          FROM SPACEXTBL
         WHERE Landing_Outcome like '%Failure%') AS Failed_Outcome
  FROM SPACEXTBL
 WHERE Landing_Outcome like '%Success%'
```

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

Success_Outcome	Failed_Outcome
61	10

Query Explanation: Sub query was used together with the count function to calculate the total number of successful and failure mission outcomes

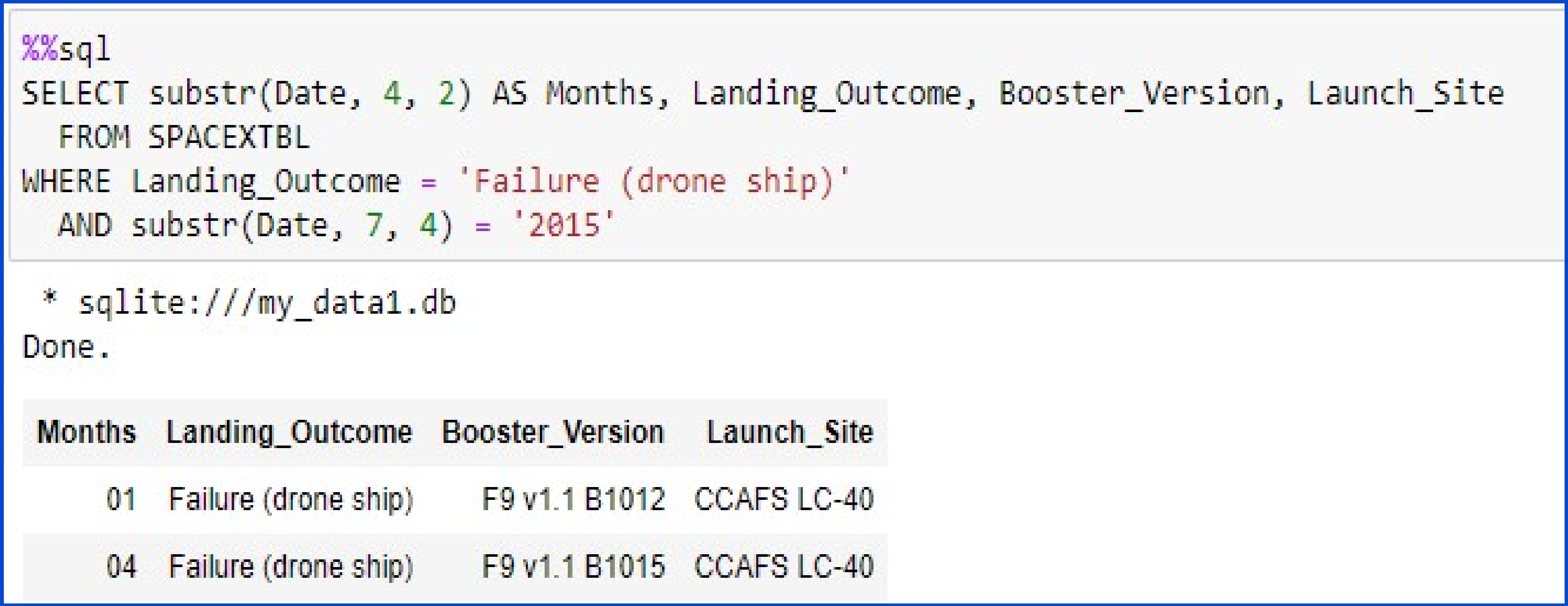
Boosters Carried Maximum Payload



Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Query Explanation: A subquery was used to find the maximum payload mass while the main query was used to retrieve is the names of the booster which have carried this maximum payload mass

2015 Launch Records



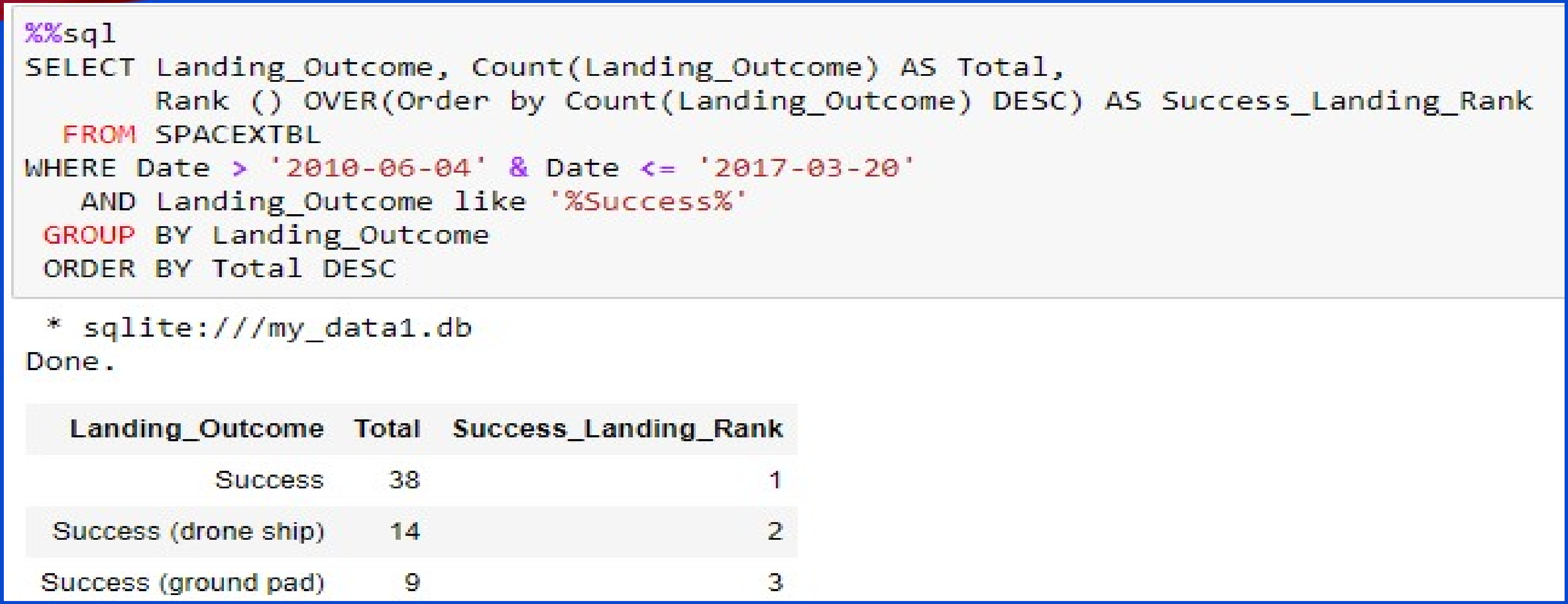
```
%%sql
SELECT substr(Date, 4, 2) AS Months, Landing_Outcome, Booster_Version, Launch_Site
  FROM SPACEXTBL
WHERE Landing_Outcome = 'Failure (drone ship)'
  AND substr(Date, 7, 4) = '2015'
```

```
 * 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

Query Explanation: The substr function was used to retrieve the date in 2015 and the WHERE clause was used to filter the data frame for failed landing outcomes in drone ship.

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



```
%%sql
SELECT Landing_Outcome, Count(Landing_Outcome) AS Total,
       Rank () OVER(Order by Count(Landing_Outcome) DESC) AS Success_Landing_Rank
  FROM SPACEXTBL
WHERE Date > '2010-06-04' & Date <= '2017-03-20'
   AND Landing_Outcome like '%Success%'
 GROUP BY Landing_Outcome
 ORDER BY Total DESC
```

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

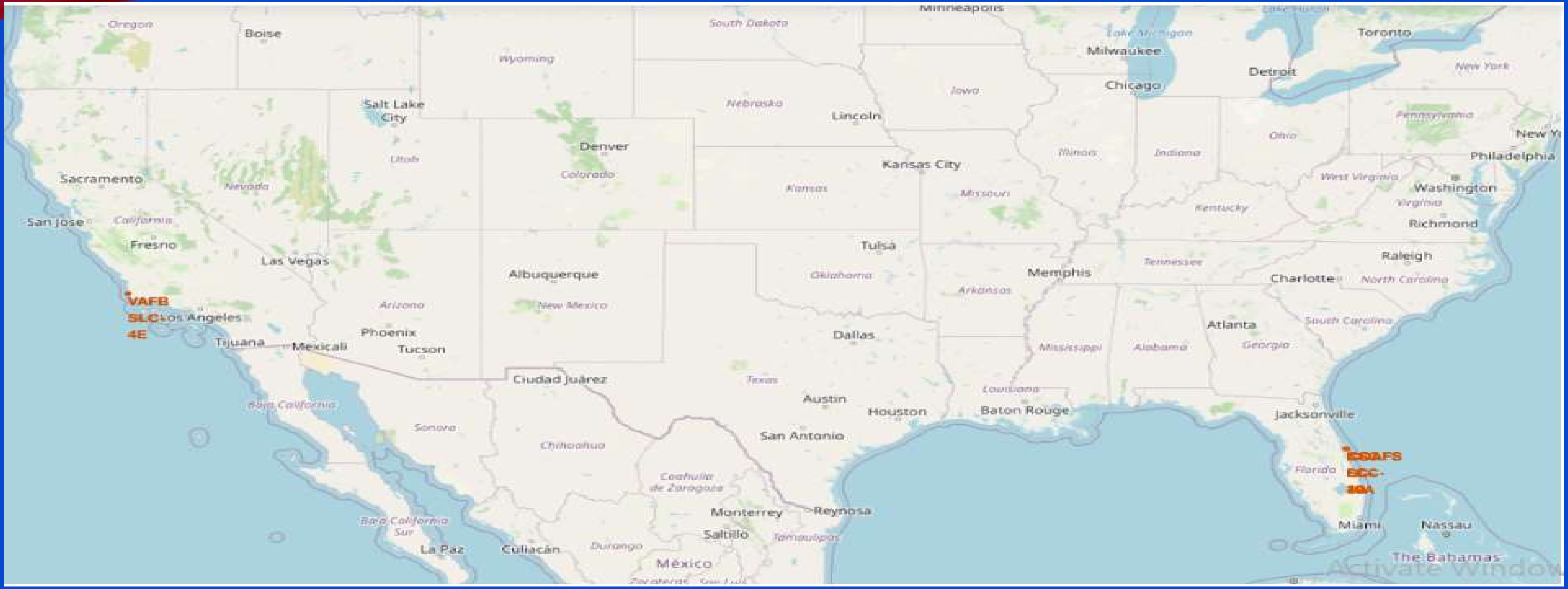
Landing_Outcome	Total	Success_Landing_Rank
Success	38	1
Success (drone ship)	14	2
Success (ground pad)	9	3

Query Explanation: The count function was first used to count the landing outcome between the date 2010-06-04 and 2017-03-20, in descending order and then the rank function was applied

Section 3

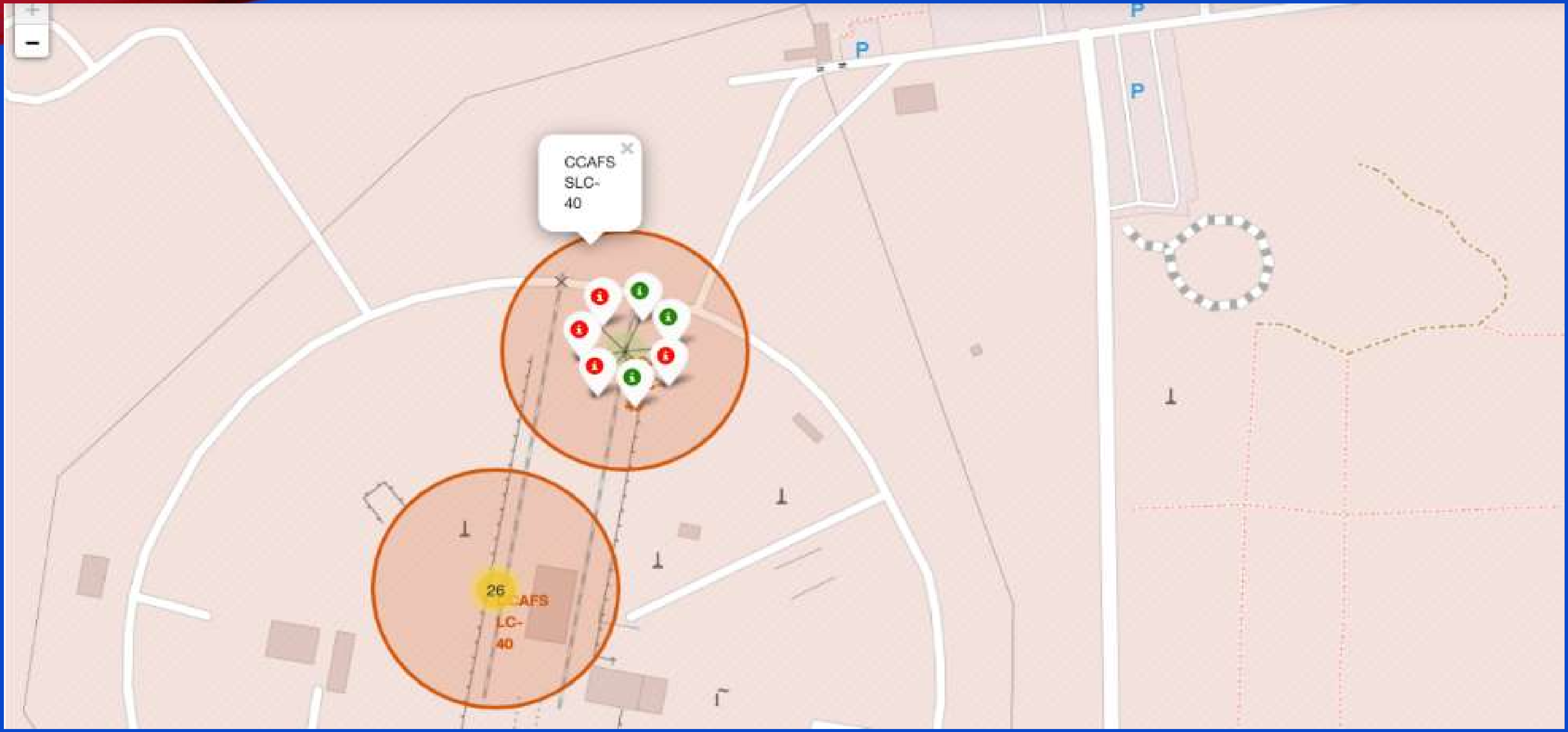
Launch Sites Proximities Analysis

Map showing Launch site Locations



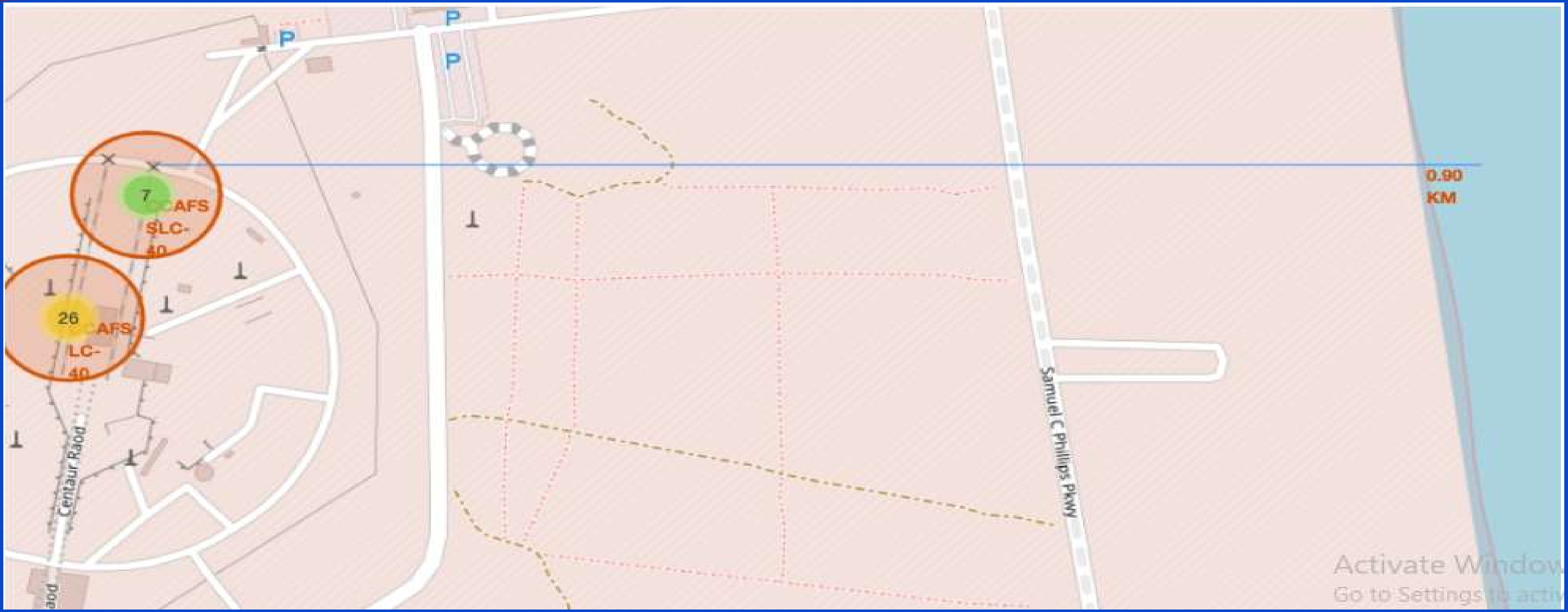
The locations of the launch sites from the map shows that the launch sites are in very close proximities to the coast lines and distance away from the equator

Success Rate of Launches for each site on the Map

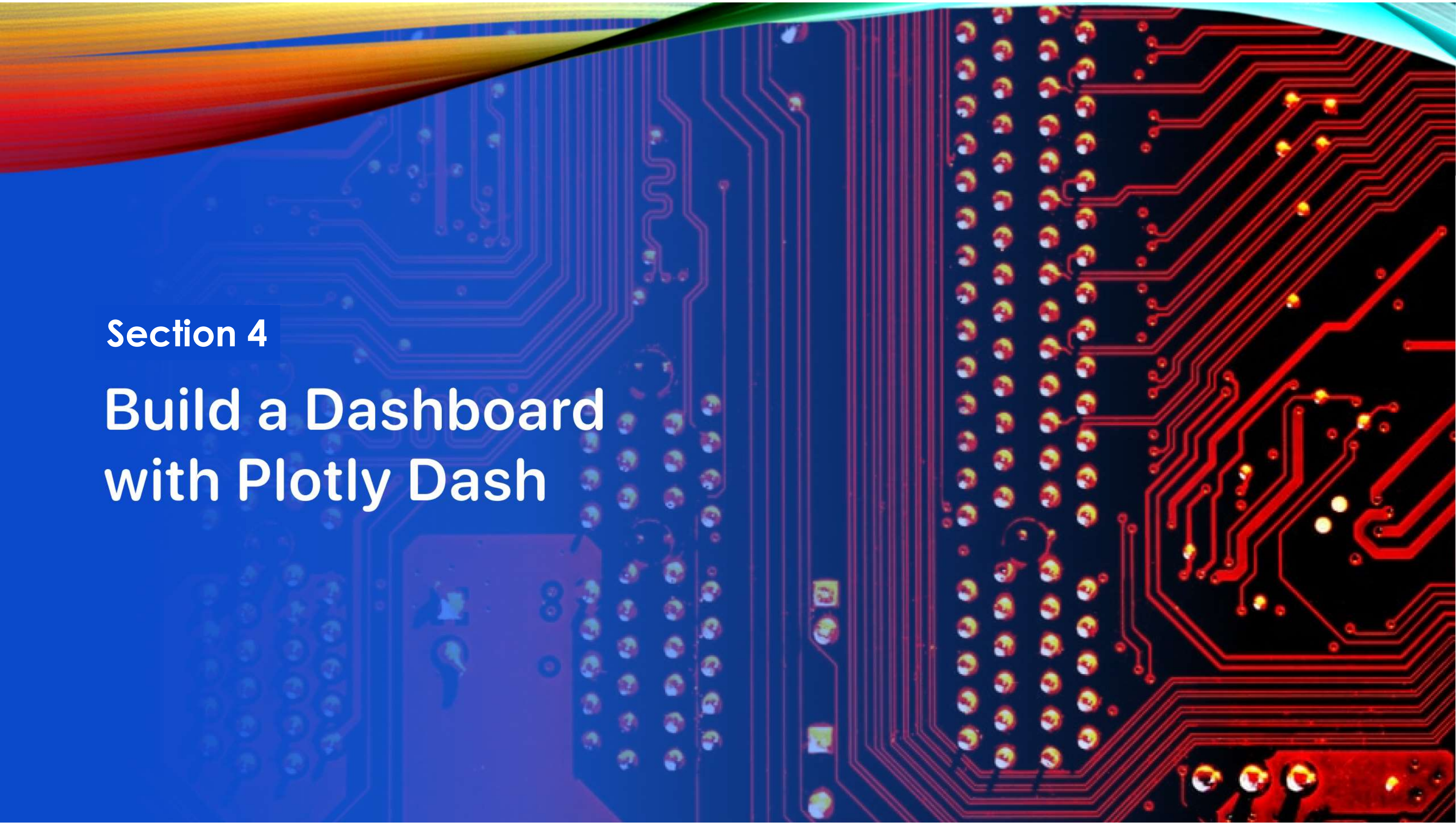


Green color stands for successful launch outcome while red color stands for a failed launch outcome

Map showing launch sites distance to proximities



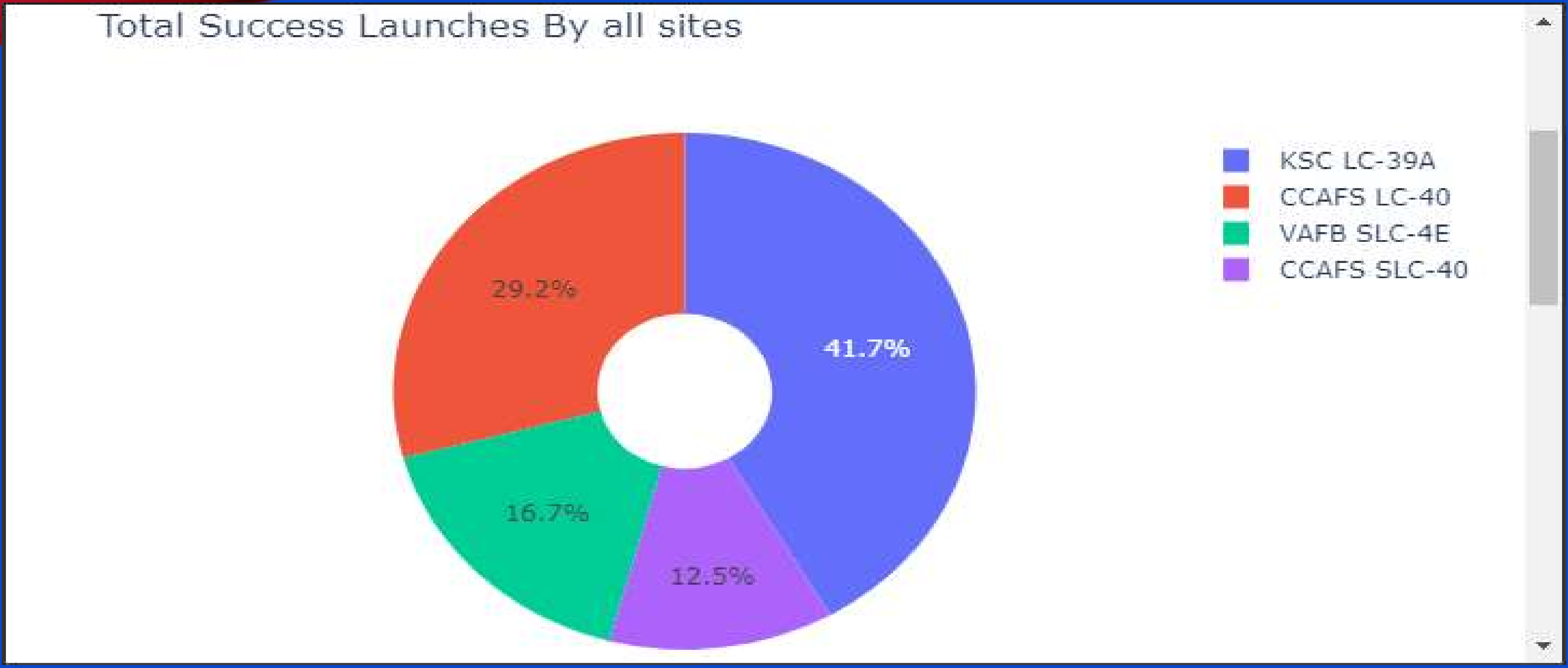
The launch sites appear to be close be in close proximities with railway, highway, coastline, with distance calculated as 0.9km



Section 4

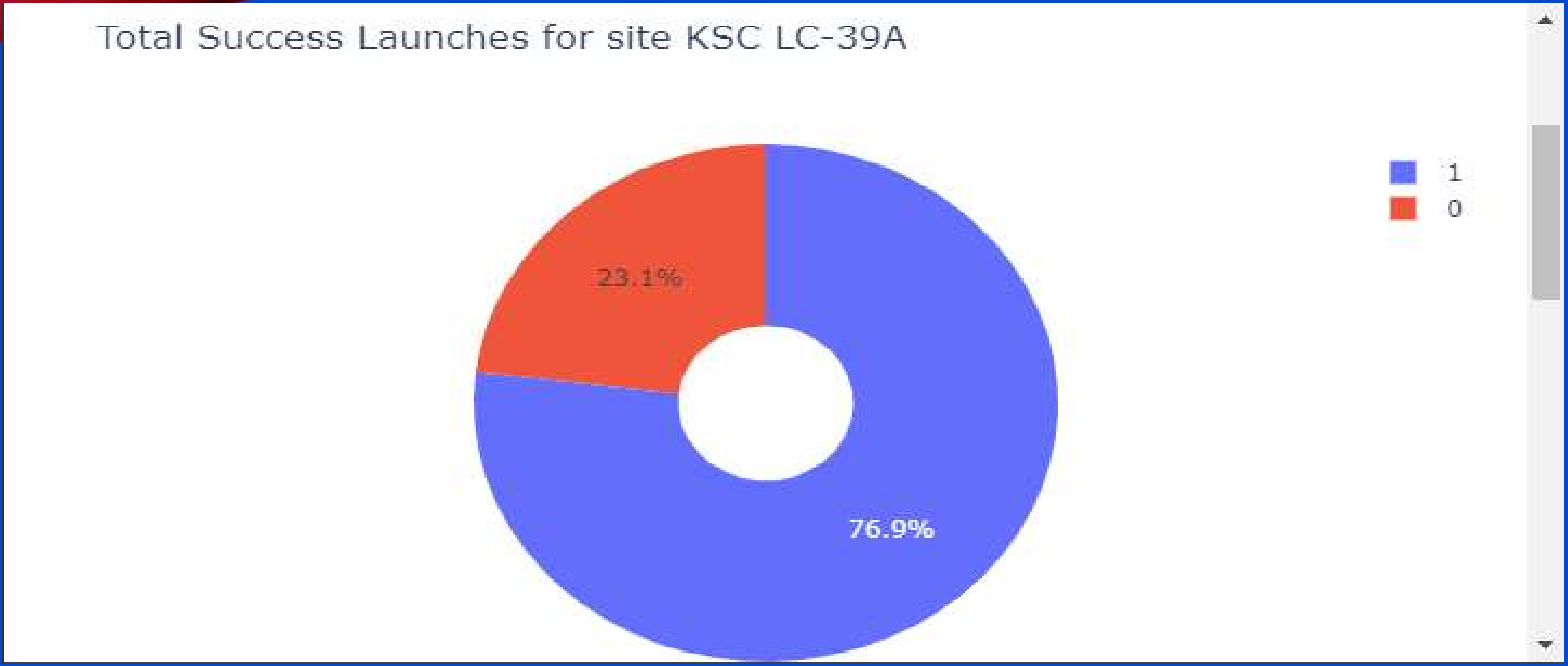
Build a Dashboard with Plotly Dash

Total Success launches by all sites



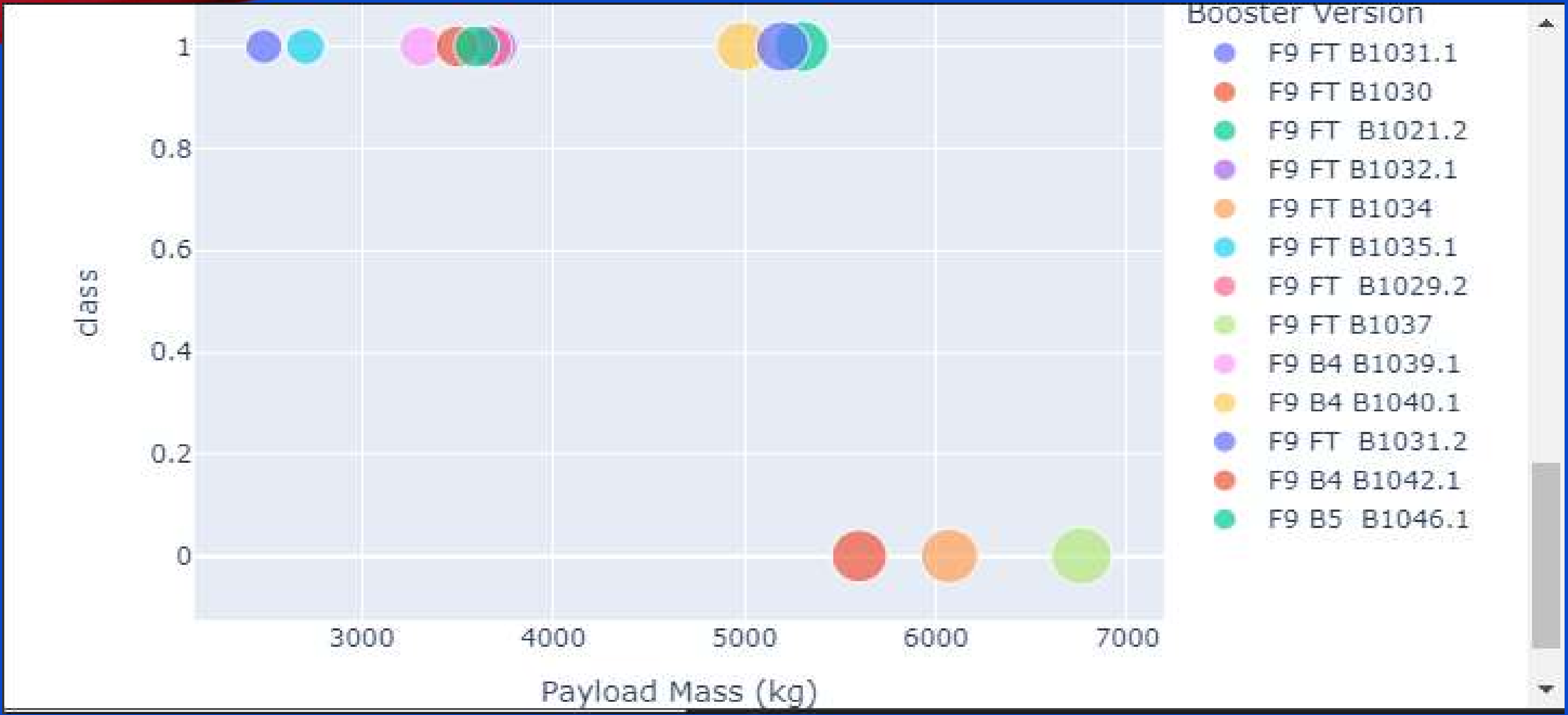
KSC LC-39A site has the largest successful launches as well the highest launch success rate.

Total Success launches for site KSC LC-39A



As shown from the chart, site KSC LC-39A recorded 76.9% success rate

Class vs. Payload Mass (kg)

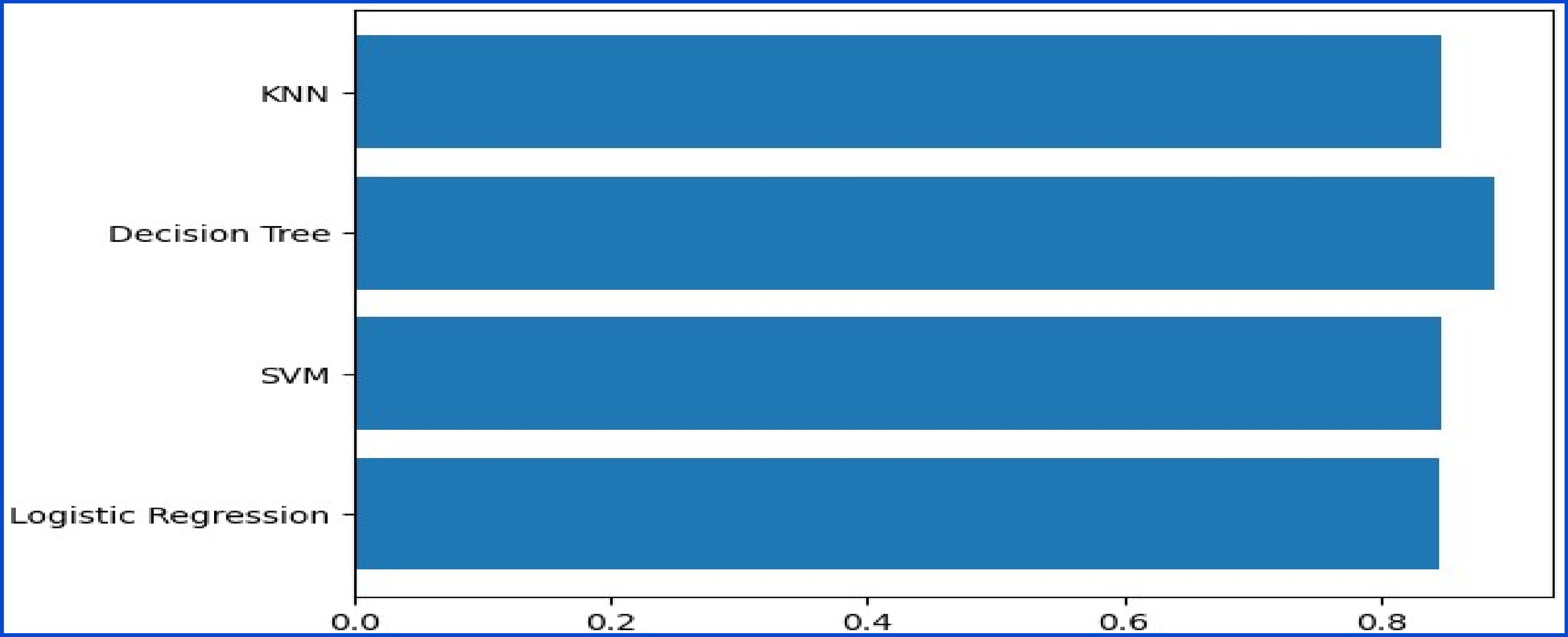


The payload range between 2000 to 6000kg seems to have the largest success rate

Section 5

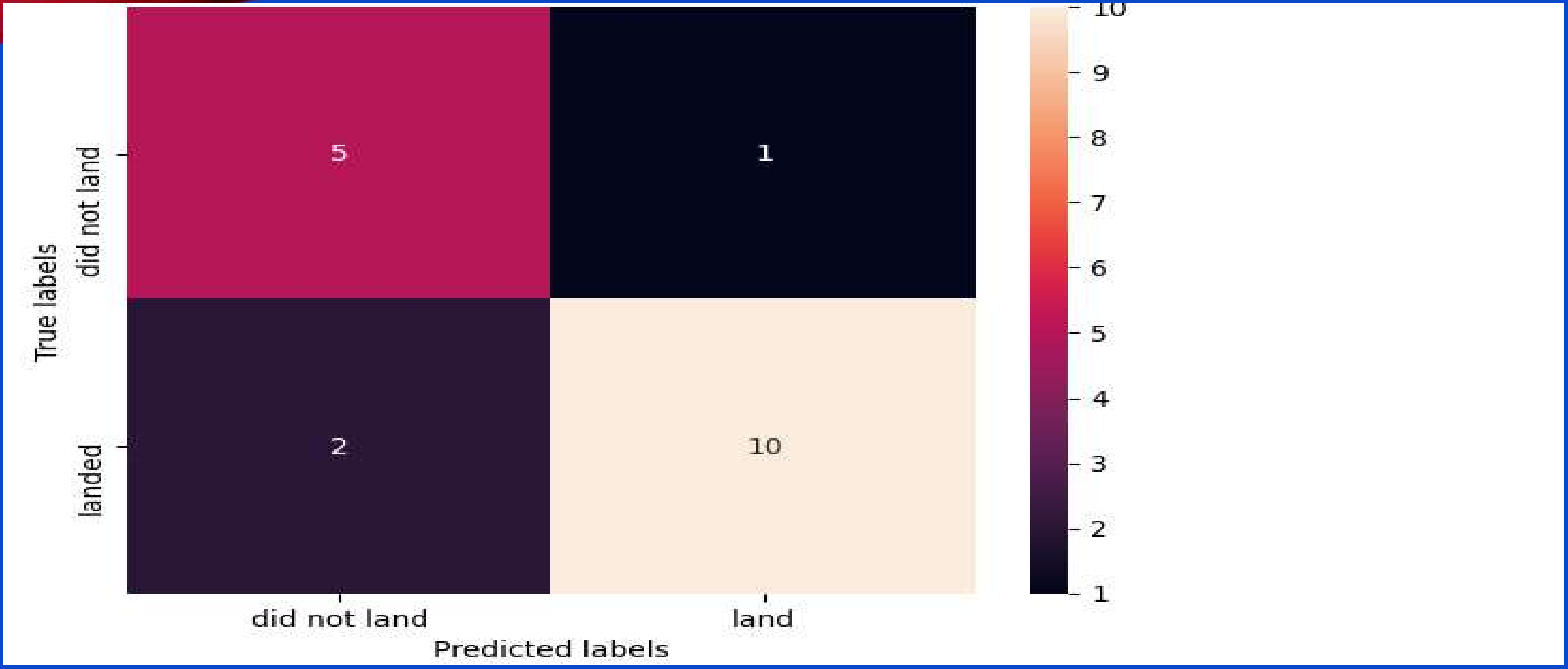
Predictive Analysis (Classification)

Classification Accuracy



Visualizing the built model accuracy for all built classification models in a bar chart shows that decision tree has the highest classification accuracy with an accuracy of 88.9%

Confusion Matrix



From the confusion matrix shown above the model correctly predicted 5 outcome of 6 not landing and correctly predicted 10 outcome landing from a possible 12.

Conclusions

- ❑ The Decision tree classifier has been proved to be the best model to be used for the prediction (classification) as it has the highest classification accuracy
- ❑ The decision tree classifier has an accuracy of 88.9%. The model correctly predicted 5 outcome of 6 not landing and correctly predicted 10 outcome landing from a possible 12.
- ❑ The launch sites are located near highways, coastlines and railways for ease of transportations.
- ❑ The trend analysis shows that there has been record of increasing success rate since 2013 till 2020
- ❑ The price of each launch can now be determined

.

Appendix

GitHub link of the project: https://github.com/Dedonrukks/Data-Science-Capstone-Project

Thank you!