IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arun Castromin Lawrance. S
16-May-2023

Outline

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

Executive Summary

Summary of methodologies

- Data Collection through APIs
- Data Collection via Web scraping
- Data Wrangling and Cleansing
- Exploratory Data Analysis with data visualization
- Exploratory Data Analysis with SQL
- Build an interactive map with Folium
- Interactive Visual Analytics and Dashboard
- Predictive Analysis with Machine Learning

Summary of all results

- Exploratory data analysis results
- Interactive analysis Summary
- Predictive Analysis results

Introduction

- **Project background and context**

Space X 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 Space X 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 space X for a rocket launch. The goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully

- **Problems you want to find answers**

What factors determine if the rocket will land successfully

What variables of various launch parameter combinations are best suited for successful landings

What parameters SpaceX has to tune to in order to have best probable results on landings

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX REST API and web scraping from Wikipedia
- Perform data wrangling
  - One-hot encoding was applied to categorical features.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Various visualization techniques were used such as Scatter Graphs and Bar Charts to study data structure and relationships
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Usage of Machine Learning algorithms to predict if the first stage of Falcon 9 will land successfully.

Data Collection

Data was gathered from

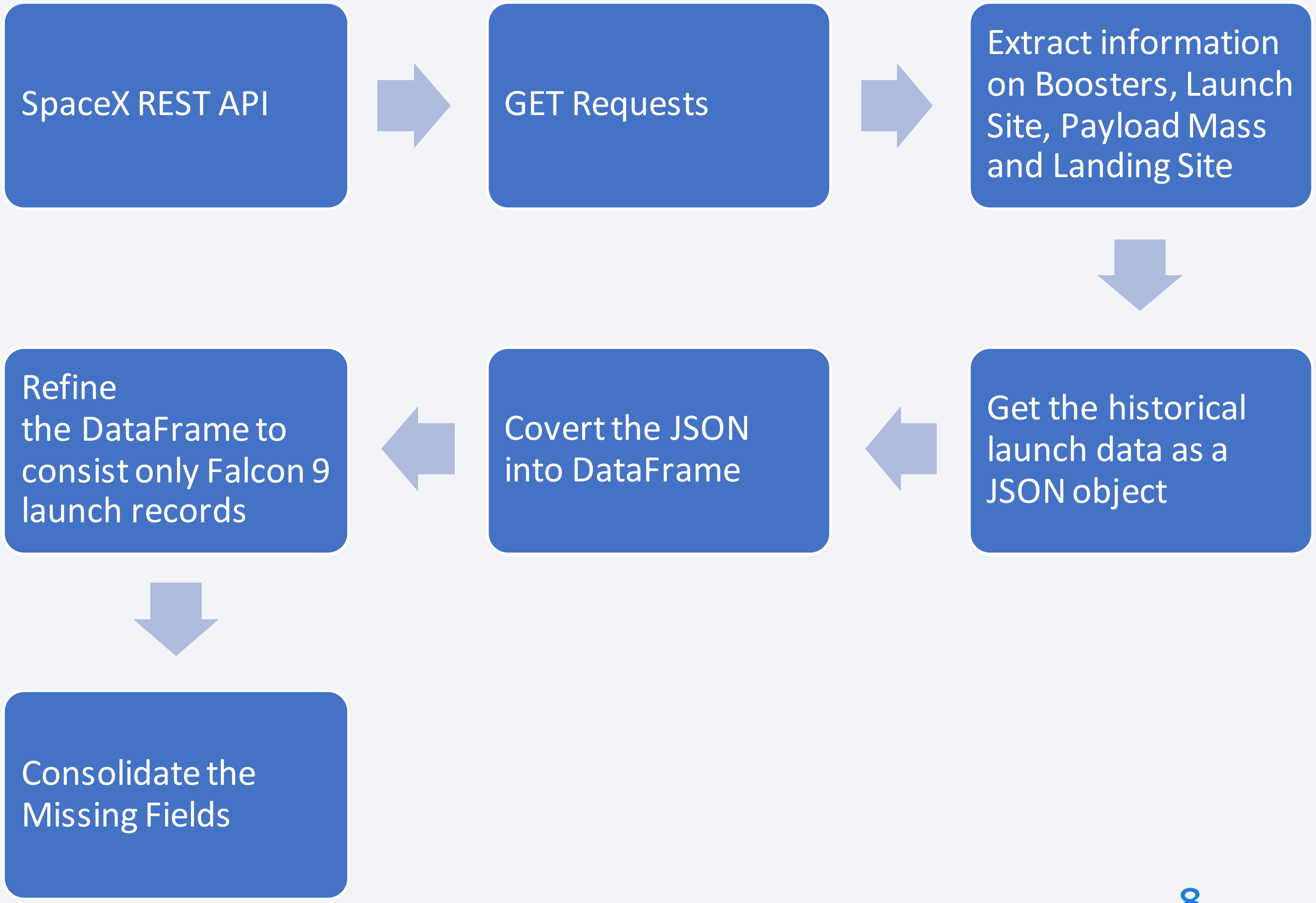
- SpaceX REST API (acquiring JSON responses and converting them to dataframes)
- WebScraping from wiki pages (acquiring the web scraped data into dataframes)

These thus acquired DataFrames were used for further processing

Data Collection – SpaceX API

- Connect via SpaceX REST API using GET requests to extract historical information on the available Falcon Launch and Landing details
- Convert the acquired JSON data stream into Dataframe

Data Collection – SpaceX API Notebook



SpaceX REST API

GET Requests

Extract information on Boosters, Launch Site, Payload Mass and Landing Site

Get the historical launch data as a JSON object

Covert the JSON into DataFrame

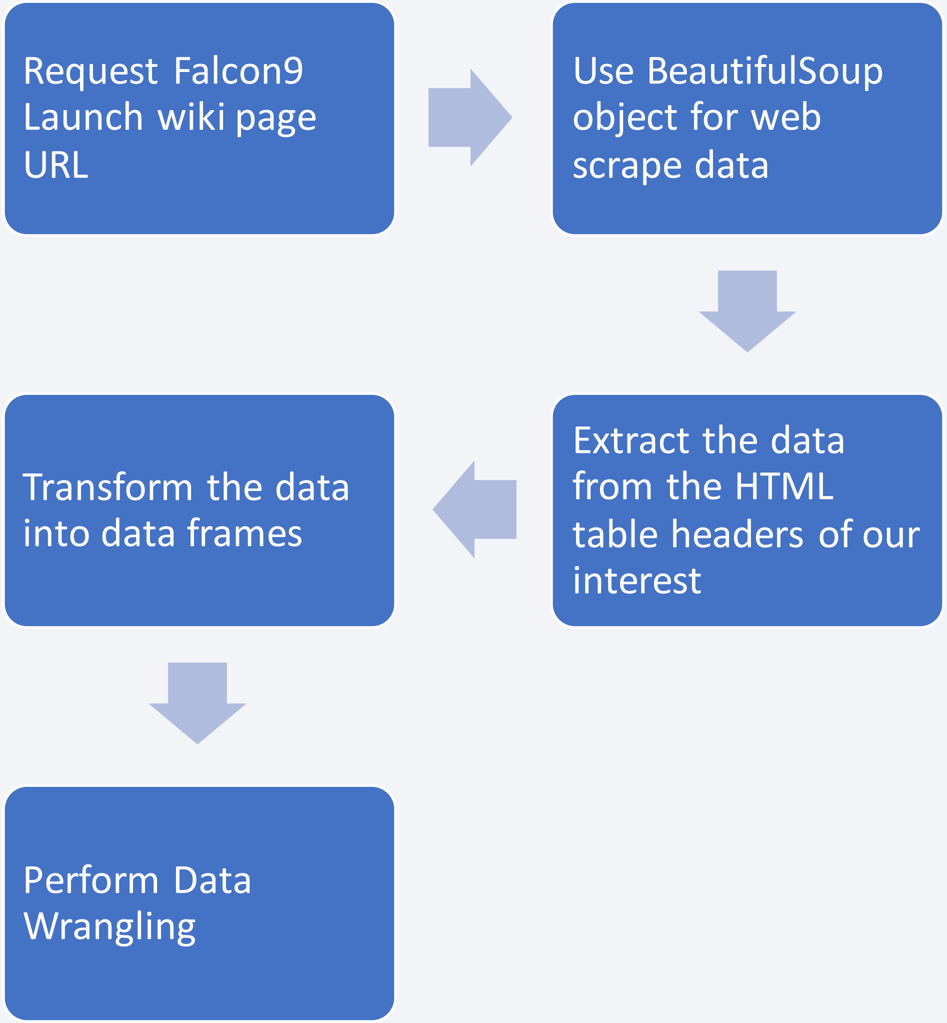
Refine the DataFrame to consist only Falcon 9 launch records

Consolidate the Missing Fields

Data Collection - Scraping

- Perform web scraping to collect Falcon 9 historical information on launch and landing details from Wiki page
- Repurpose the extracted data into dataframes

Data Collection - WebScraping Notebook



Request Falcon9 Launch wiki page URL

Use BeautifulSoup object for web scrape data

Extract the data from the HTML table headers of our interest

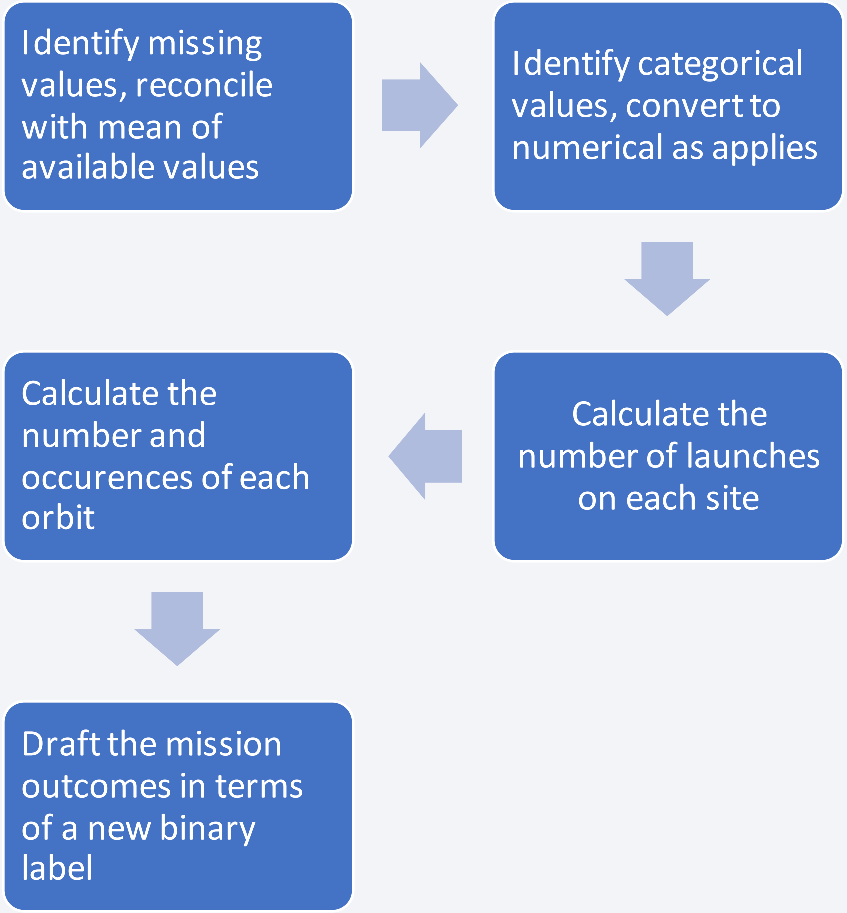
Transform the data into data frames

Perform Data Wrangling

Data Wrangling

- Performed Exploratory Data Analysis to find data relationship patterns
- Organize the records to be classified as a mission success of failure. This is done by assigning a new label 'Outcome' with values 0 and 1 indicating Failure and Success respectively

Data Wrangling Notebook



Identify missing values, reconcile with mean of available values

Identify categorical values, convert to numerical as applies

Calculate the number of launches on each site

Calculate the number and occurences of each orbit

Draft the mission outcomes in terms of a new binary label

EDA with Data Visualization

Charts for Visualization:

- Catplot to visualize the relationship between Flight Number and Payload
- Catplot to visualize the relationship between Flight Number and Launch Site
- Catplot to visualize the relationship between Payload and Launch Site
- Bar Chart to visualize the relationship between Success Rate of each Orbit types
- Catplot to visualize the relationship between Flight Number and Orbit Type
- Catplot to visualize the relationship between Payload and Orbit Type
- Line Chart to visualize the yearly trend of launch successes

EDA with Visualization Notebook

EDA with SQL

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

  select distinct Launch_Site from SPACEXTBL

- Display 5 records where launch sites begin with the string 'CCA'

  *select * from SPACEXTBL where launch_site like 'CCA%' limit 5*

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

  select sum(payload_mass__kg_) as "total payload mass for NASA (CRS)" from SPACEXTBL where Customer='NASA (CRS)'

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

  select avg(payload_mass__kg_) as "average payload mass carried by booster version F9 v1.1" from SPACEXTBL

  Where booster_version='F9 v1.1'

[EDA with SQL Notebook]()

EDA with SQL

- List the date when the first succesful landing outcome in ground pad was acheived.

  select min(Date) from SPACEXTBL where Landing_Outcome like '%Success%'

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

  Select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and

  PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000

- List the total number of successful and failure mission outcomes

  select count(mission_outcome) as "Total number of successfull missions" from SPACEXTBL where landing_outcome like '%Success%' select count(mission_outcome) as "Total number of failed missions" from SPACEXTBL where landing_outcome like '%Failure%'

[EDA with SQL Notebook]

EDA with SQL

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
select booster_version from spacextbl where payload_mass__kg_ = (select max (payload_mass__kg_) from spacextbl)
```

- 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.

```
select substr(Date,4,2) as 'month names',landing_outcome,booster_version,launch_site from spacextbl where

substr(Date,7,4)='2015' and landing_outcome ='Failure (drone ship)'
```

- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
Select landing_outcome,count(landing_outcome) cnt from spacextbl where landing_outcome like '%Success%' and date
between '04-06-2010' and '20-03-2017' group by landing_outcome order by cnt desc
```

EDA with SQL Notebook

Build an Interactive Map with Folium

Summary of map objects that were created and added to the Folium map:

Folium.Circle and **folium.Marker** to add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.

MarkerCluster object for simplify a map containing many markers having the same coordinate

MousePosition on the map to get coordinate for a mouse over a point on the map

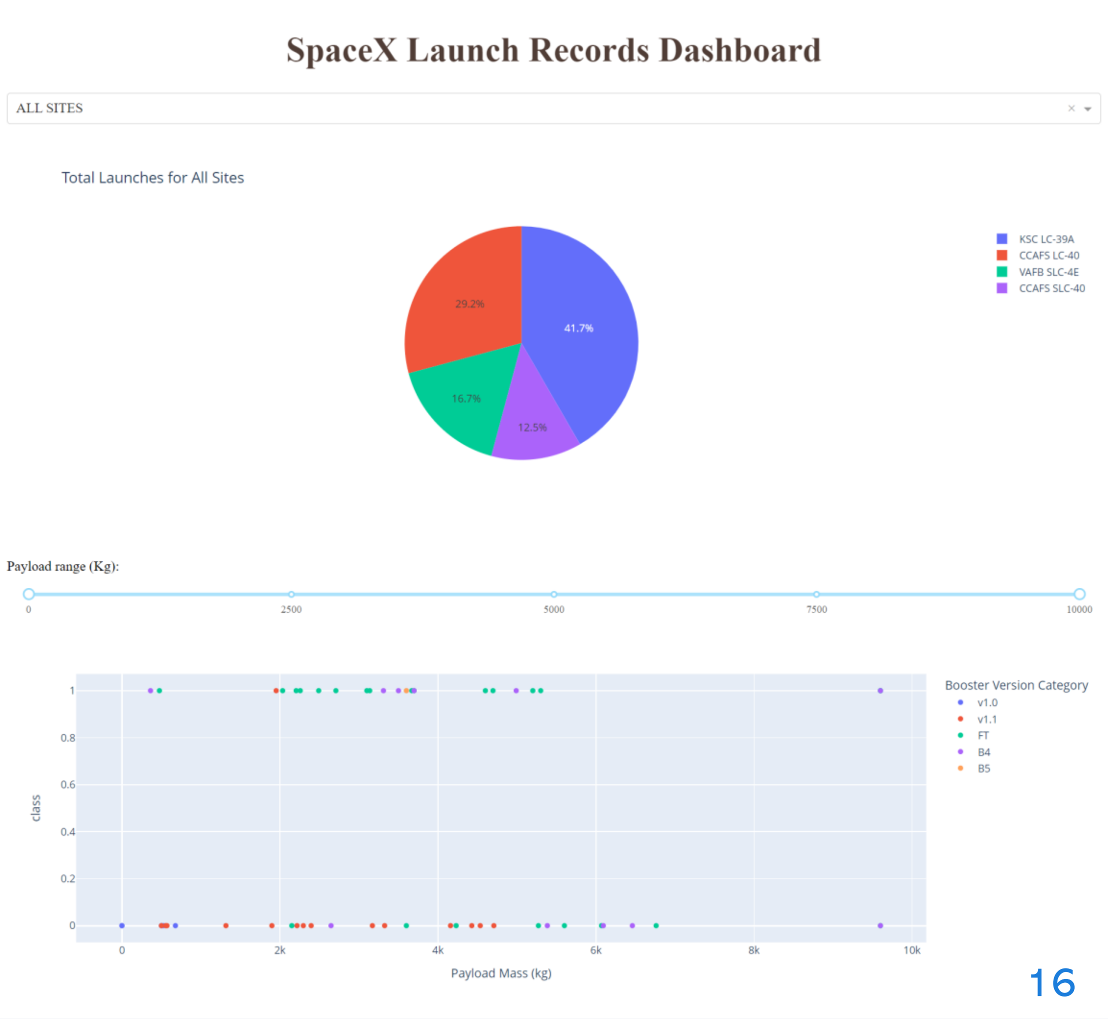
Folium.PolyLine object to draw a line between a launch site to its closest city, coastlines and highway

Interactive Map with Folium Notebook

Build a Dashboard with Plotly Dash

- An interactive Dashboard built showcasing a pie chart and a scatter graph
- Pie Chart shows the distribution of successful launches across launch sites with dropdown select options for the launch sites
- Scatter graph illustrates the success/failure instances by payload and booster versions with a range slider for selecting the payload mass

[Plotly Dashboard Notebook]



SpaceX Launch Records Dashboard

ALL SITES

Predictive Analysis (Classification)

Model used to predict the probability of first stage landing with the data collected and assessed from previous labs.

- NumPy array extraction from the column class in data
- Standardize the Data
- Split the data into Training and Test Data sets (80% and 20% respectively)
- Assess and evaluate the best Hyperparameters using various models – Logistic Regressions, SVM, Decision Tree and KNN with the help of
  - Confusion Matrix plotting
  - Accuracy Score Compilation
- Conclude with the thus best performing model with the help of Accuracy score compiled

Predictive Analysis Classification Notebook

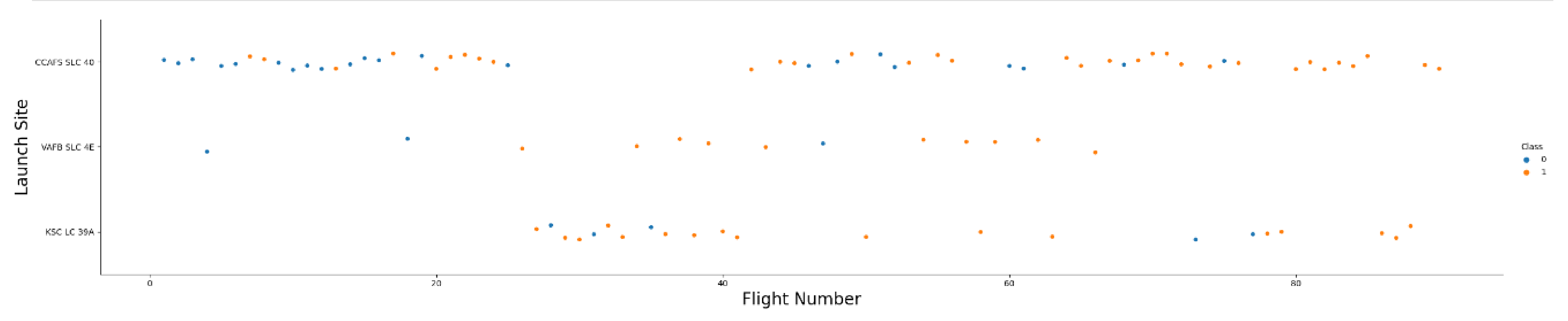
Results

- Exploratory data analysis results
  - Better Success rates observed for higher orbits
  - Launch success rates increases over time
- Interactive analytics demo in screenshots
- Higher success rates observed for higher payload mass
  - Higher success rates observed for booster versions FT, B4, B5
- Predictive analysis results
  - Best results observed with Decision Tree Model

Section 2

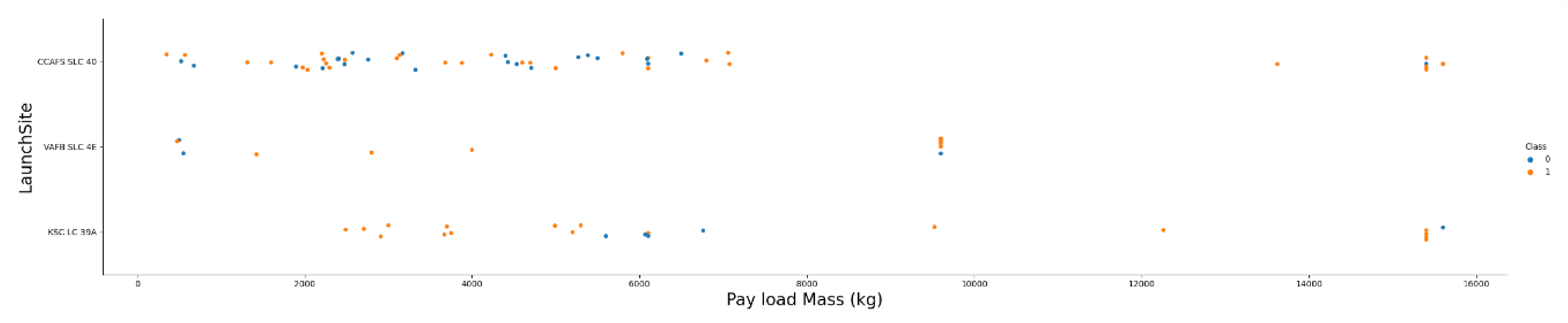
Insights drawn from EDA

Flight Number vs. Launch Site



- Success Rates can be seen increasing with time for every launch site. This is more so for CCAFS SLC 40
- Majority of launches are from launch site CCAFS SLC 40
- Higher Success Rates are observed for VAFB SLC 4E and KSC LC 39A

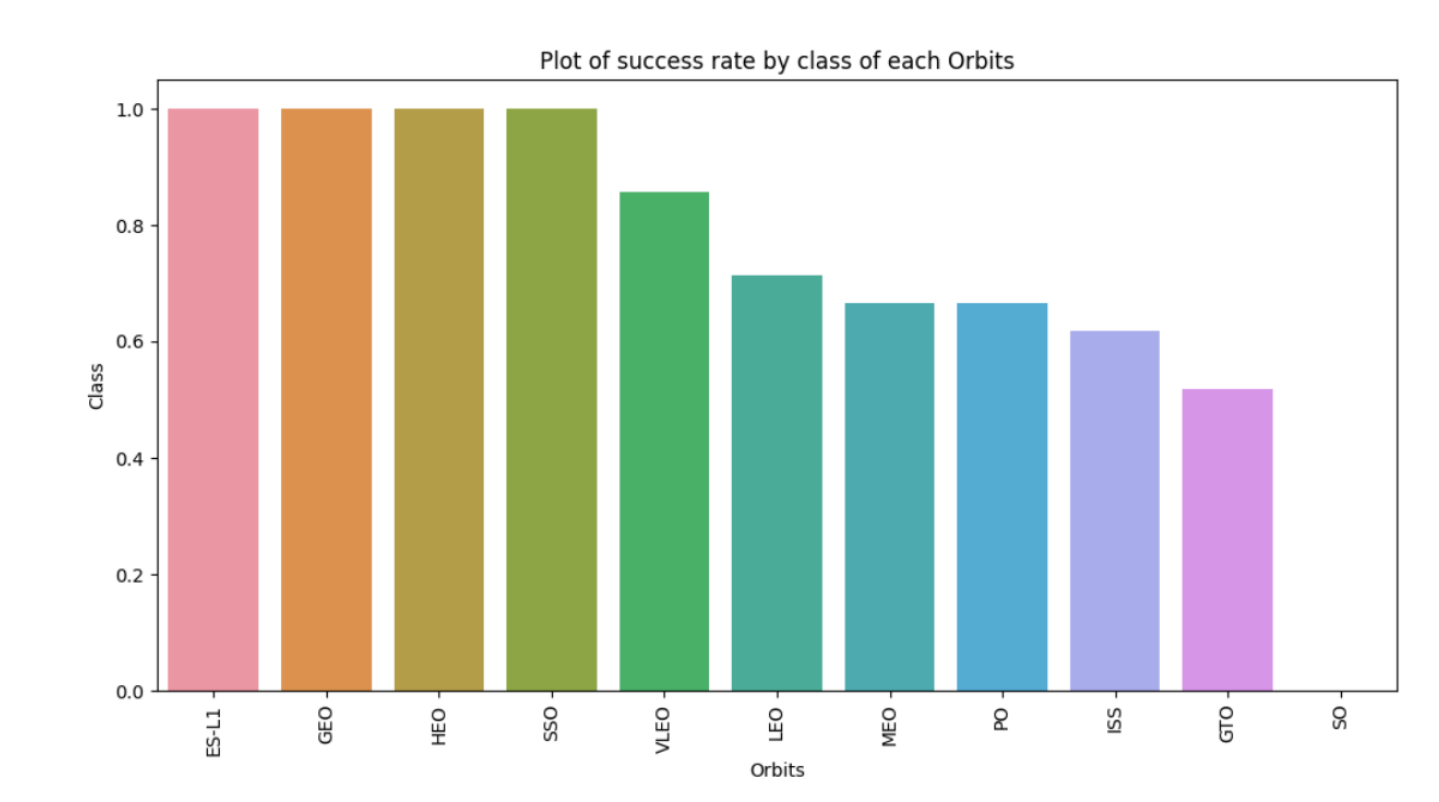
Payload vs. Launch Site



- No rockets launched for heavy payload mass > 10,000 kg at VAFB-SLC launch site
- No rockets launched for lower payload mass < 2,500 kg at KSC LC launch site
- Rockets launched with payload mass between7,500 and 13,000 kg at CCAFS SLC launch site

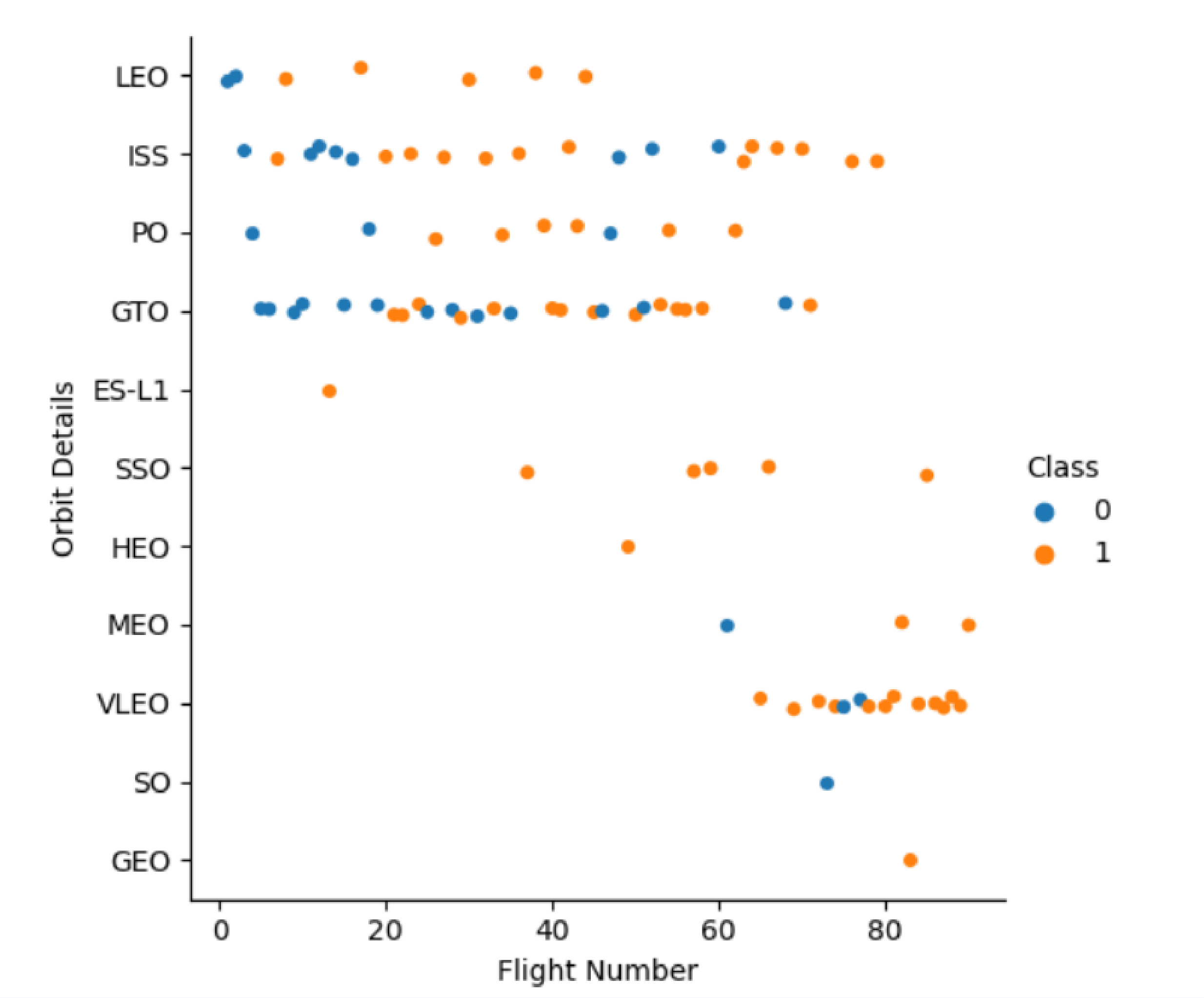
Success Rate vs. Orbit Type

- Higher Orbit types has better success rates
  - GEO
  - HEO
  - SSO
  - ES-L1



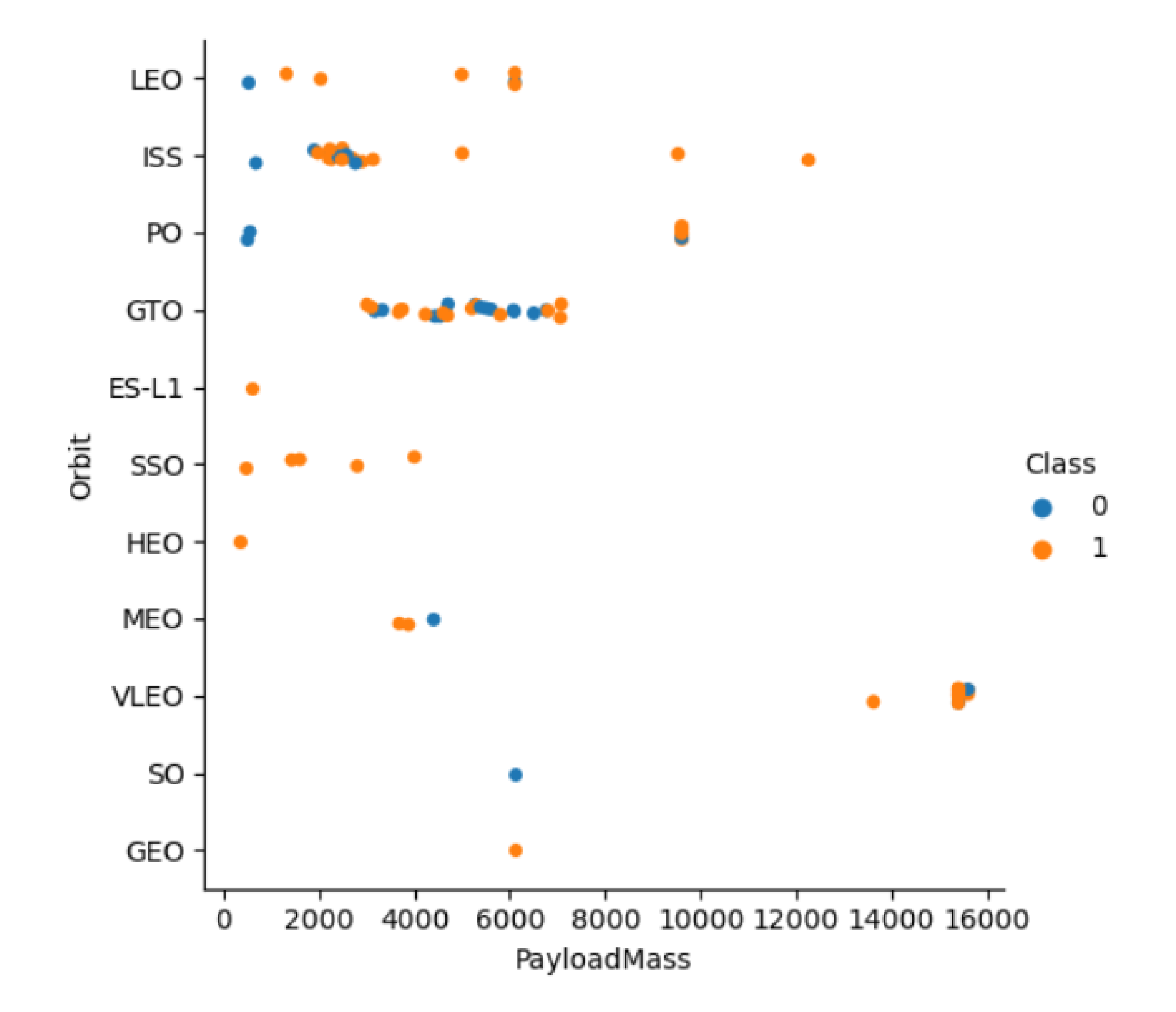
Flight Number vs. Orbit Type

- Clearly the higher oribit launches has a better success rate
- Success rates also improved over time
- More orbit types are seen added over time
- GTS and ISS orbits are heavily concentrated but with lower success rates – shows the demand and necessity for these orbit launches



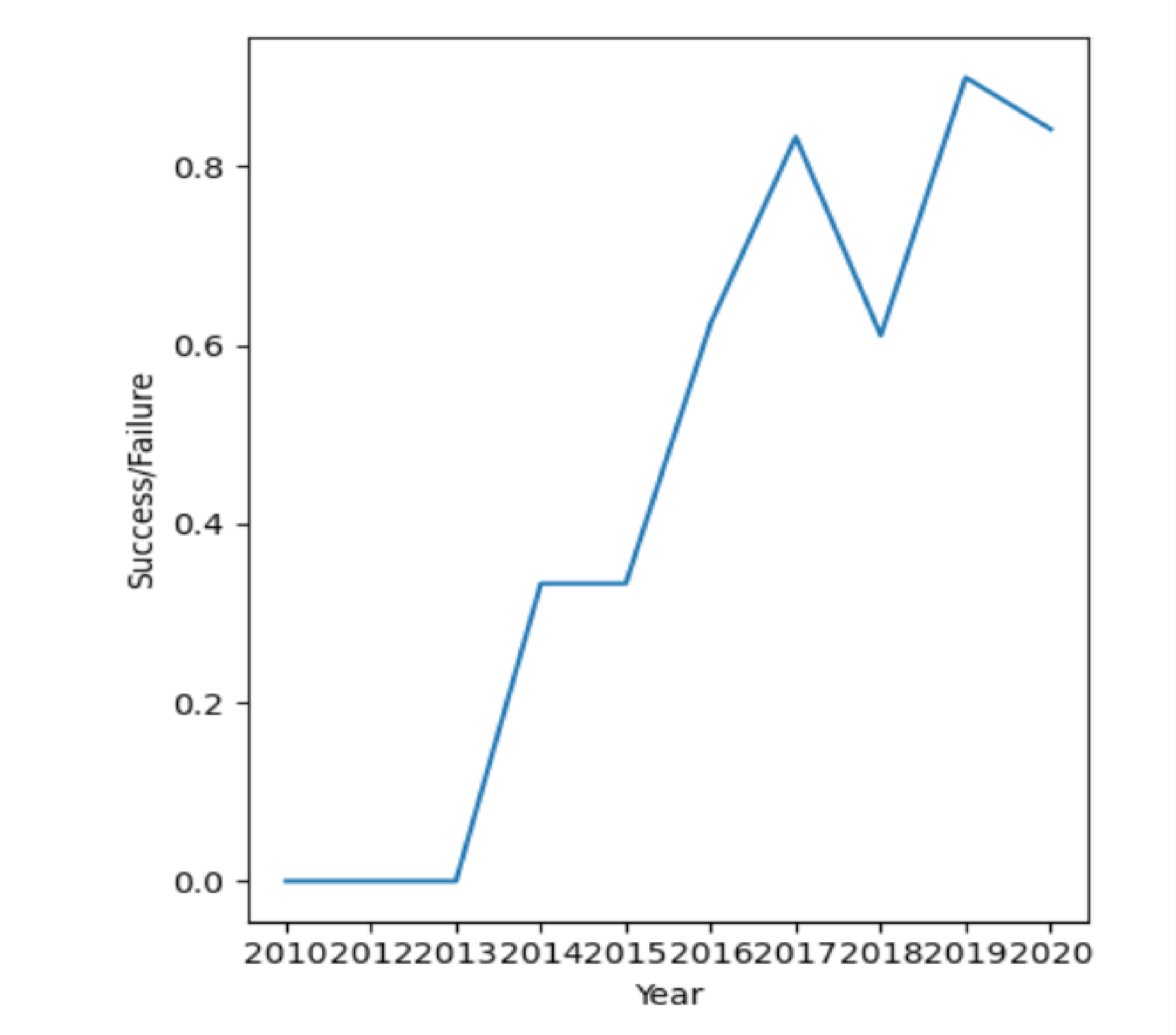
Payload vs. Orbit Type

- Most of the launches are within payloads less than 7,500 kgs
- With heavy payloads, successful landings are more for PO, LEO and ISS orbits



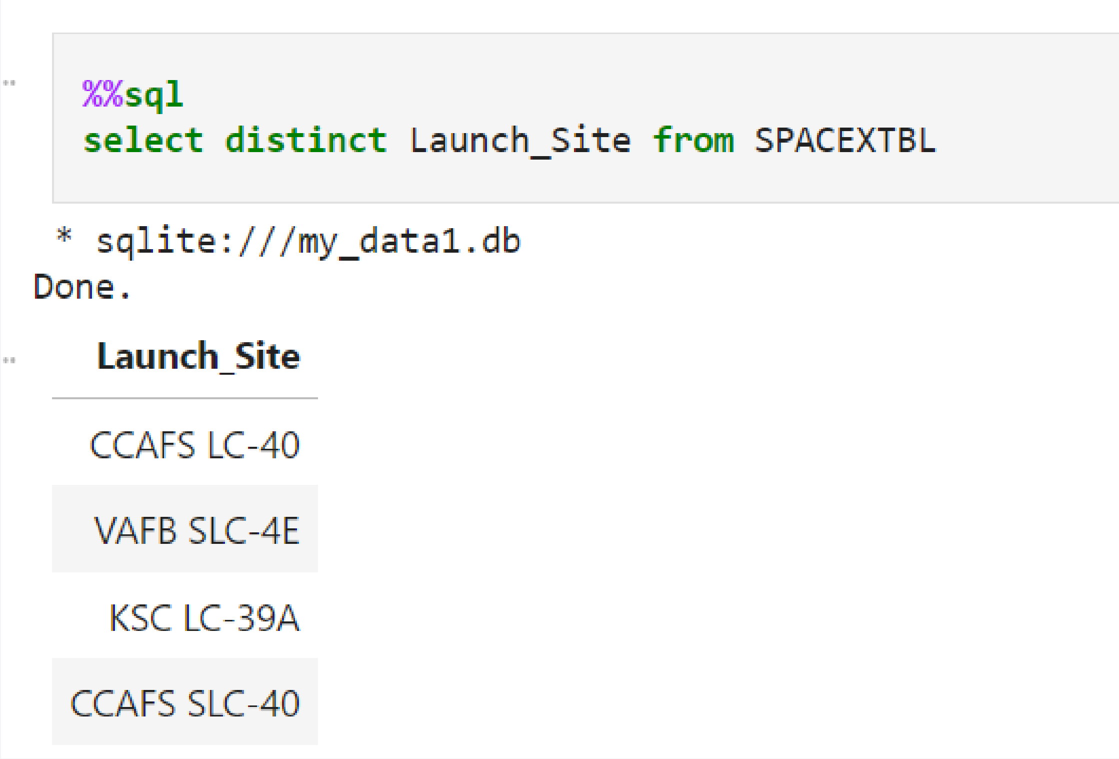
Launch Success Yearly Trend

- There is a steady incline of launch success rates since 2013



All Launch Site Names

- Used "DISTINCT" keyword on the spaceXtbl launch site column
- There are 4 unique launch sites as can be seen in the snippet



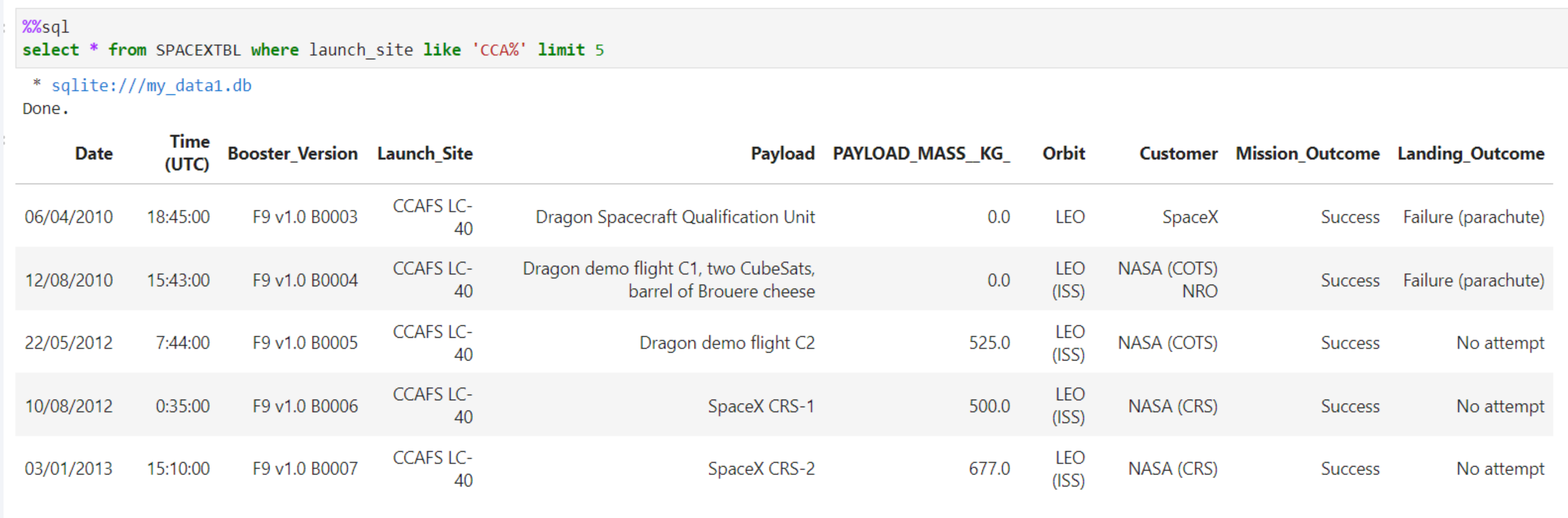
```
%%sql
select distinct Launch_Site from SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- The search criteria was placed in the query in the WHERE clause on SpaceXtbl as can be seen below:



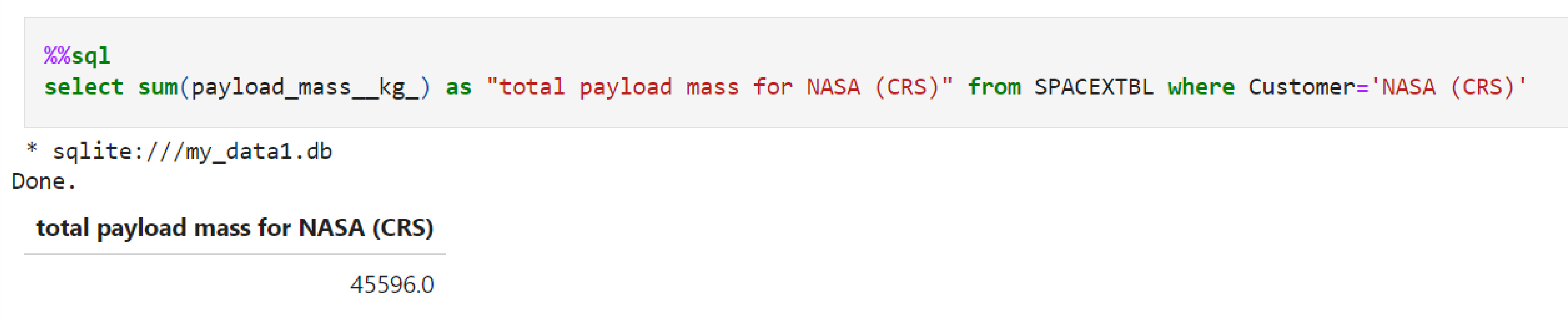
```
%%sql
select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- SUM keyword is used on payload_mass__kg_ column of SPACETBL along with a WHERE clause to filter out records for customer NASA (CRS)



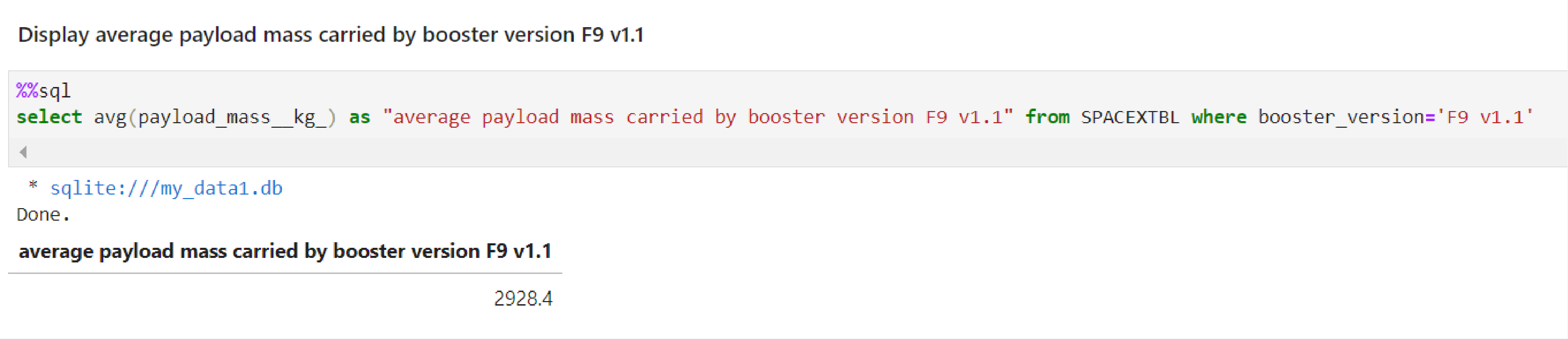
```
%%sql
select sum(payload_mass__kg_) as "total payload mass for NASA (CRS)" from SPACEXTBL where Customer='NASA (CRS)'
```

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

total payload mass for NASA (CRS)
45596.0

Average Payload Mass by F9 v1.1

- AVG keyword is used on SPACEXTBL along with a WHERE clause to filter out records belonging to booster version F9 v1.1 as below



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

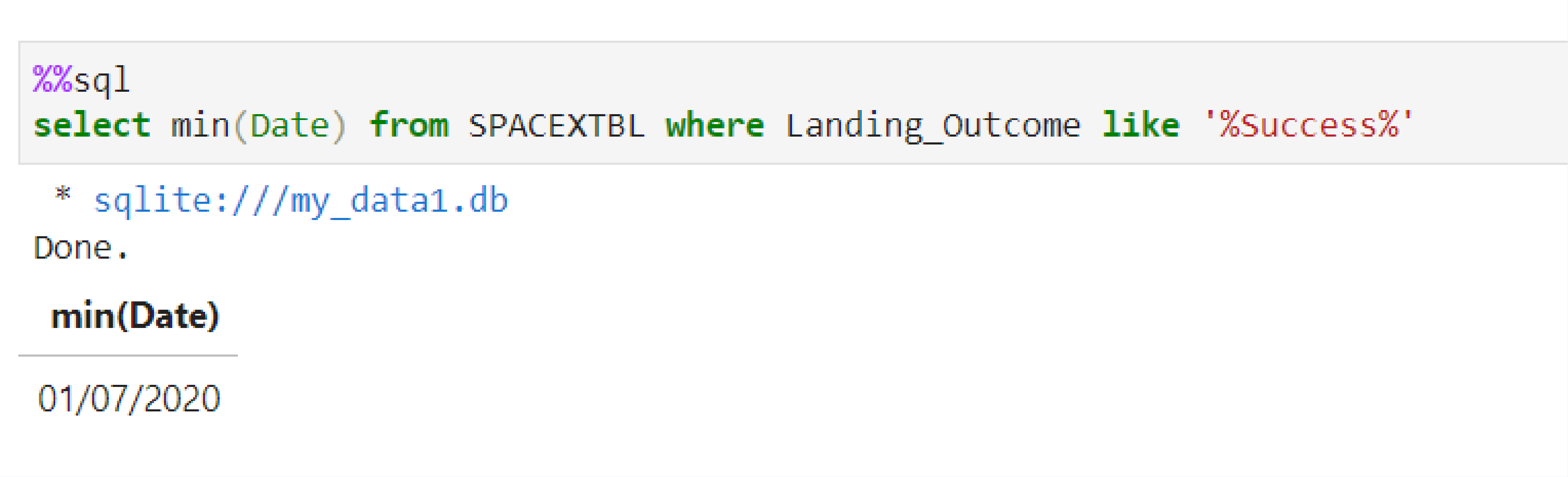
```
%%sql
select avg(payload_mass__kg_) as "average payload mass carried by booster version F9 v1.1" from SPACEXTBL where booster_version='F9 v1.1'
```

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

average payload mass carried by booster version F9 v1.1
2928.4

First Successful Ground Landing Date

- MIN function is used to getch the first successful landing outcome with a WHERE clause filter on successful landing



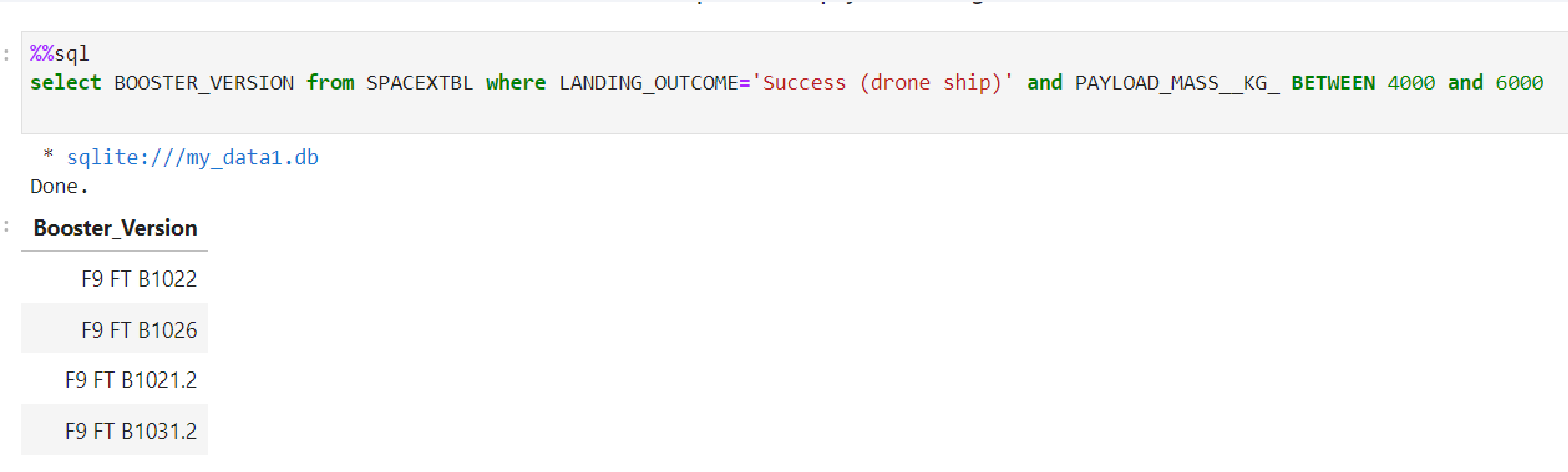
```
%%sql
select min(Date) from SPACEXTBL where Landing_Outcome like '%Success%'
```

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

min(Date)
01/07/2020

Successful Drone Ship Landing with Payload between 4000 and 6000

- WHERE clause filters applied on SPACEXTBL on successful landing outcome over drone ships with payload mass between 4000 and 6000 kgs



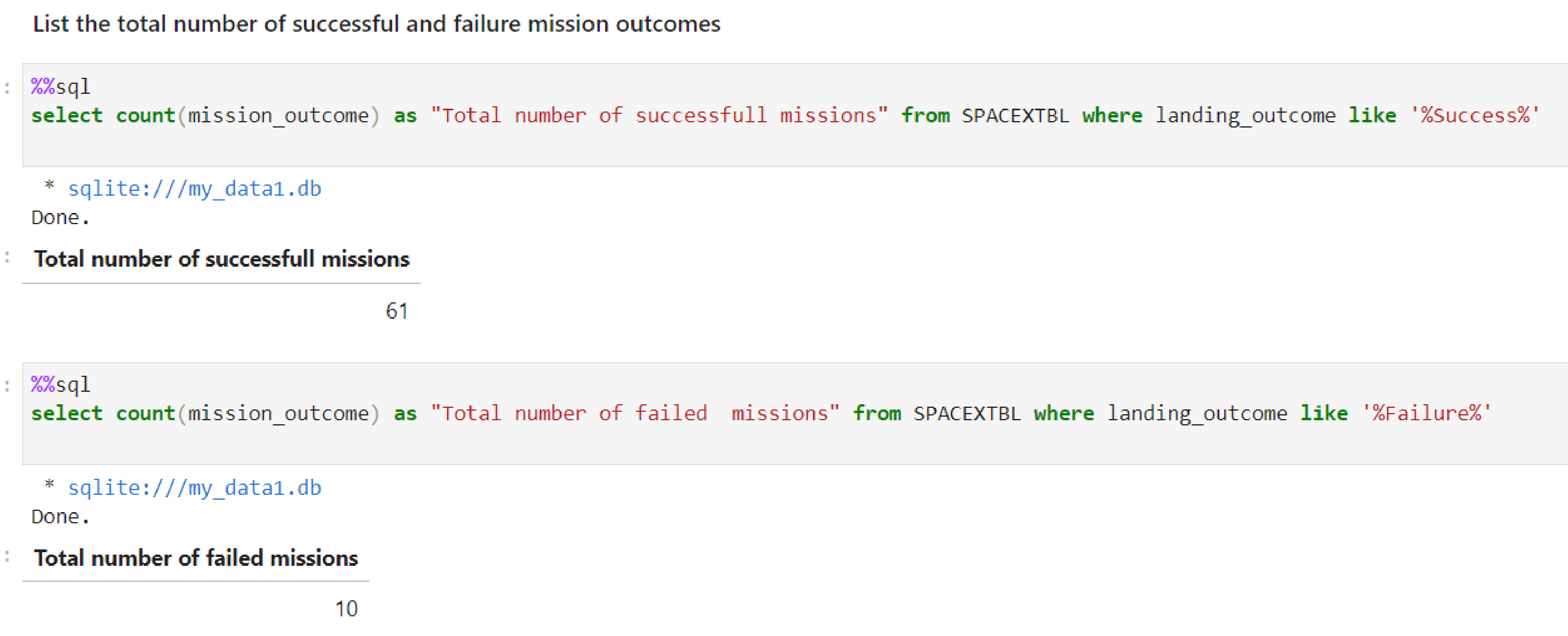
```
%%sql
select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000
```

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

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

Total Number of Successful and Failure Mission Outcomes

- COUNT function is used on the mission outcome column with WHERE clause filters on successful landing outcomes



List the total number of successful and failure mission outcomes

```
%%sql
select count(mission_outcome) as "Total number of successfull missions" from SPACEXTBL where landing_outcome like '%Success%'
```

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

Total number of successfull missions
61

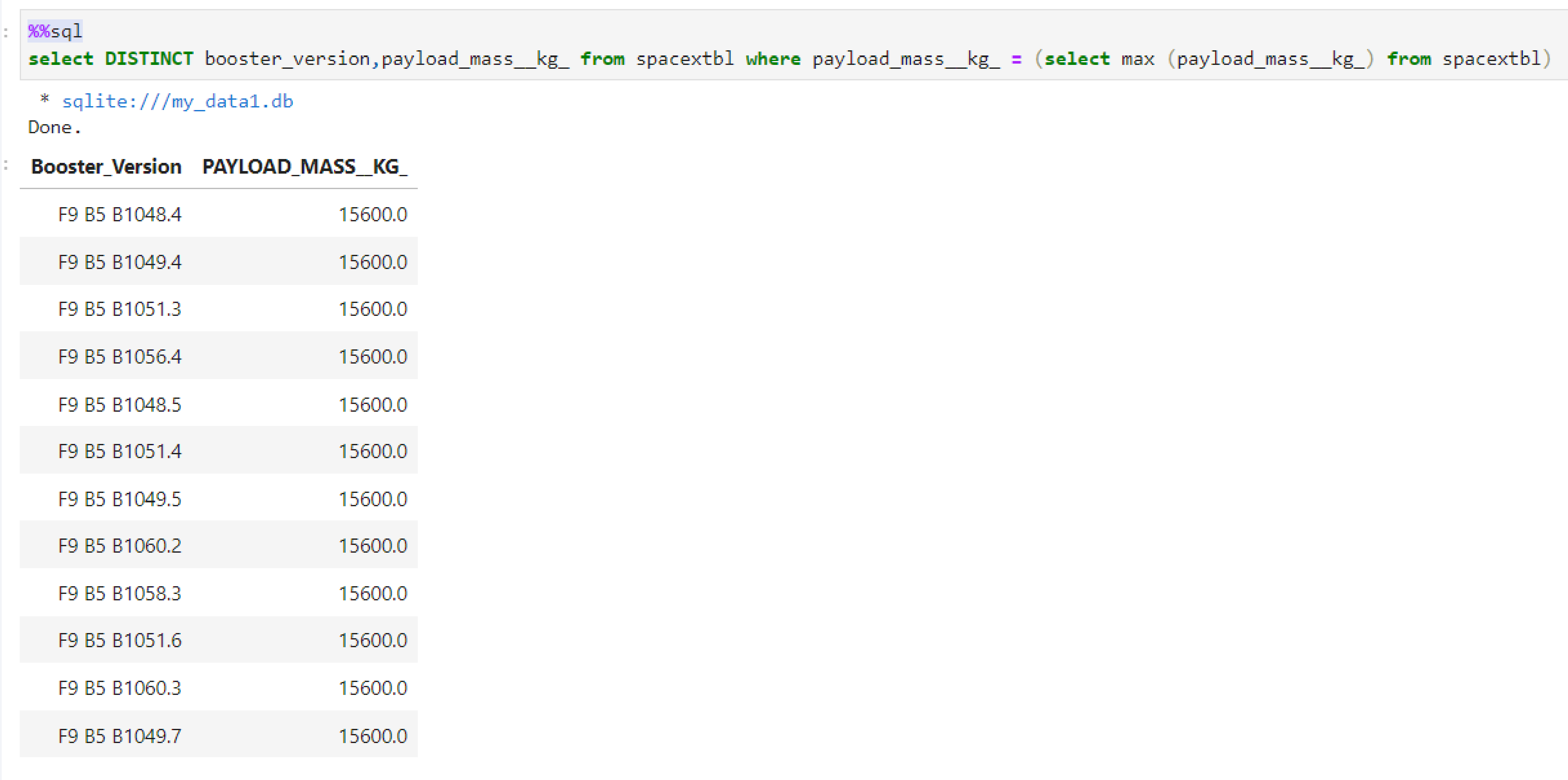
```
%%sql
select count(mission_outcome) as "Total number of failed  missions" from SPACEXTBL where landing_outcome like '%Failure%'
```

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

Total number of failed missions
10

Boosters Carried Maximum Payload

- A subquery is used with MAX function on payload mass to determine the boosters carrying maximum payloads



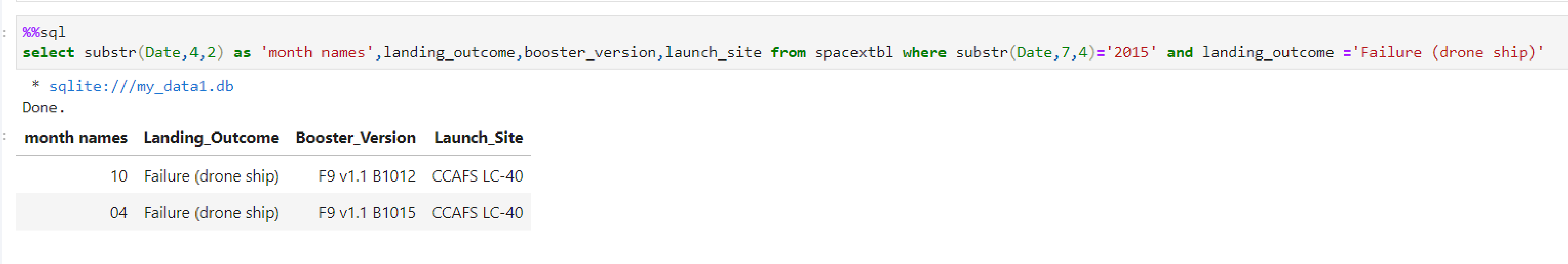
```
%%sql
select DISTINCT booster_version,payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max (payload_mass__kg_) from spacextbl)
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are prepared with a WHERE filter clause on Failed drone ship landing outcomes



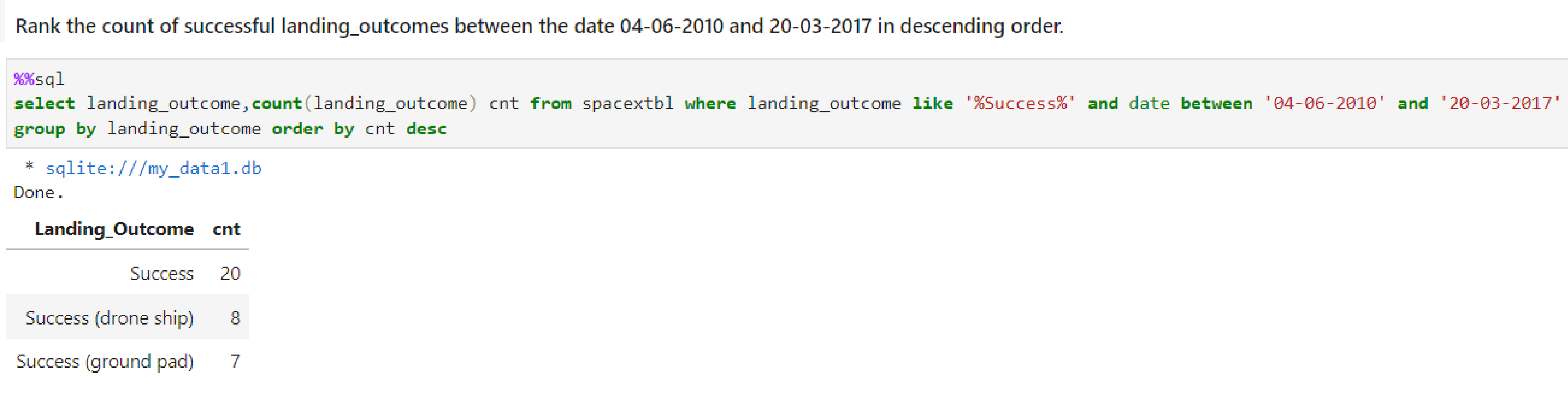
```
%%sql
select substr(Date,4,2) as 'month names',landing_outcome,booster_version,launch_site from spacextbl where substr(Date,7,4)='2015' and landing_outcome ='Failure (drone ship)'
```

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

month names	Landing_Outcome	Booster_Version	Launch_Site
10	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

- Ranking the count of landing outcomes (between the date 2010-06-04 and 2017-03-20, in descending order was fetched with usage of COUNT function on landing outcome over a grouped by landing outcome records as below.
- Where clause is applied to filter in Successful landing outcomes



Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql
select landing_outcome,count(landing_outcome) cnt from spacextbl where landing_outcome like '%Success%' and date between '04-06-2010' and '20-03-2017'
group by landing_outcome order by cnt desc
```

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

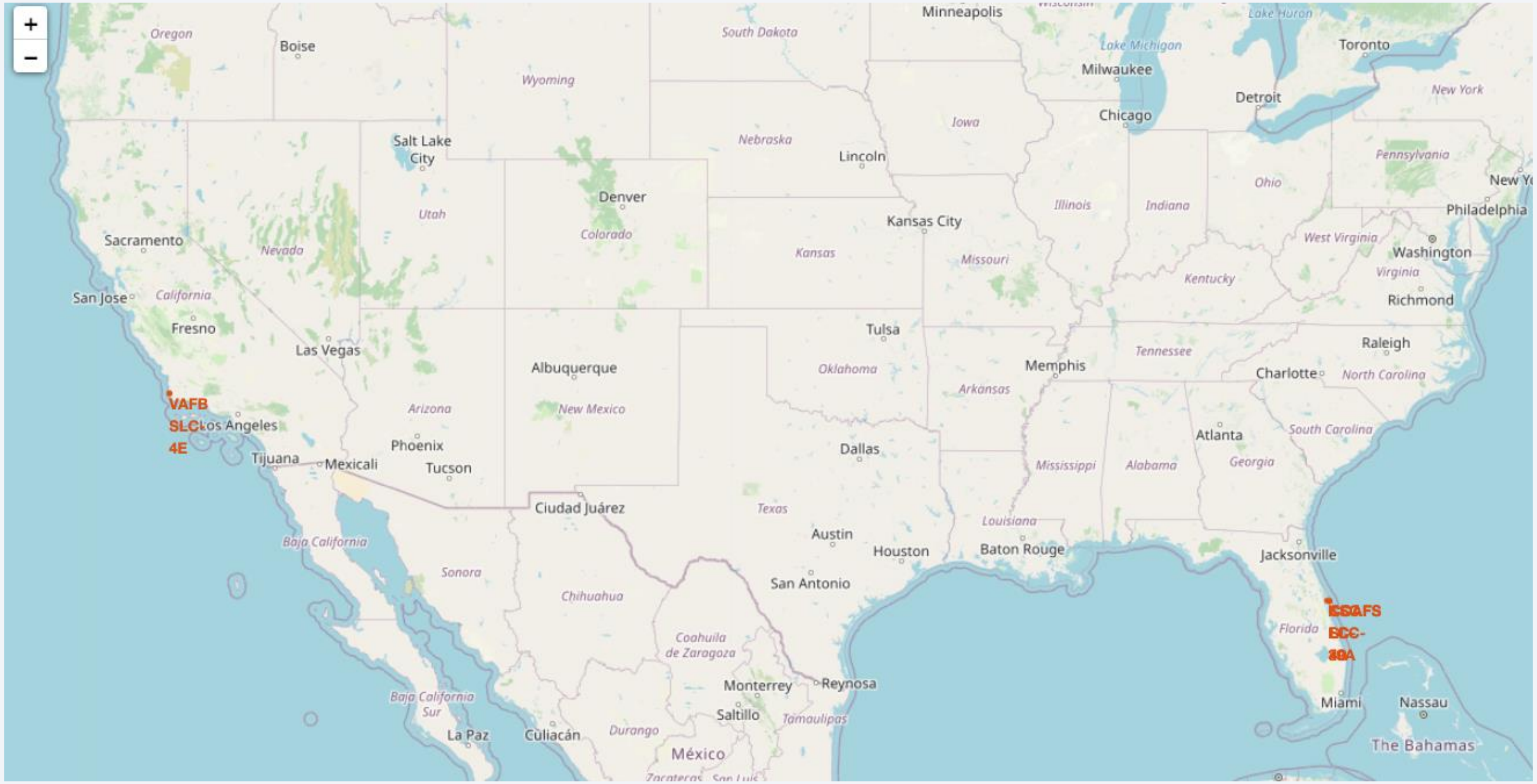
Landing_Outcome	cnt
Success	20
Success (drone ship)	8
Success (ground pad)	7

Section 3

Launch Sites Proximities Analysis

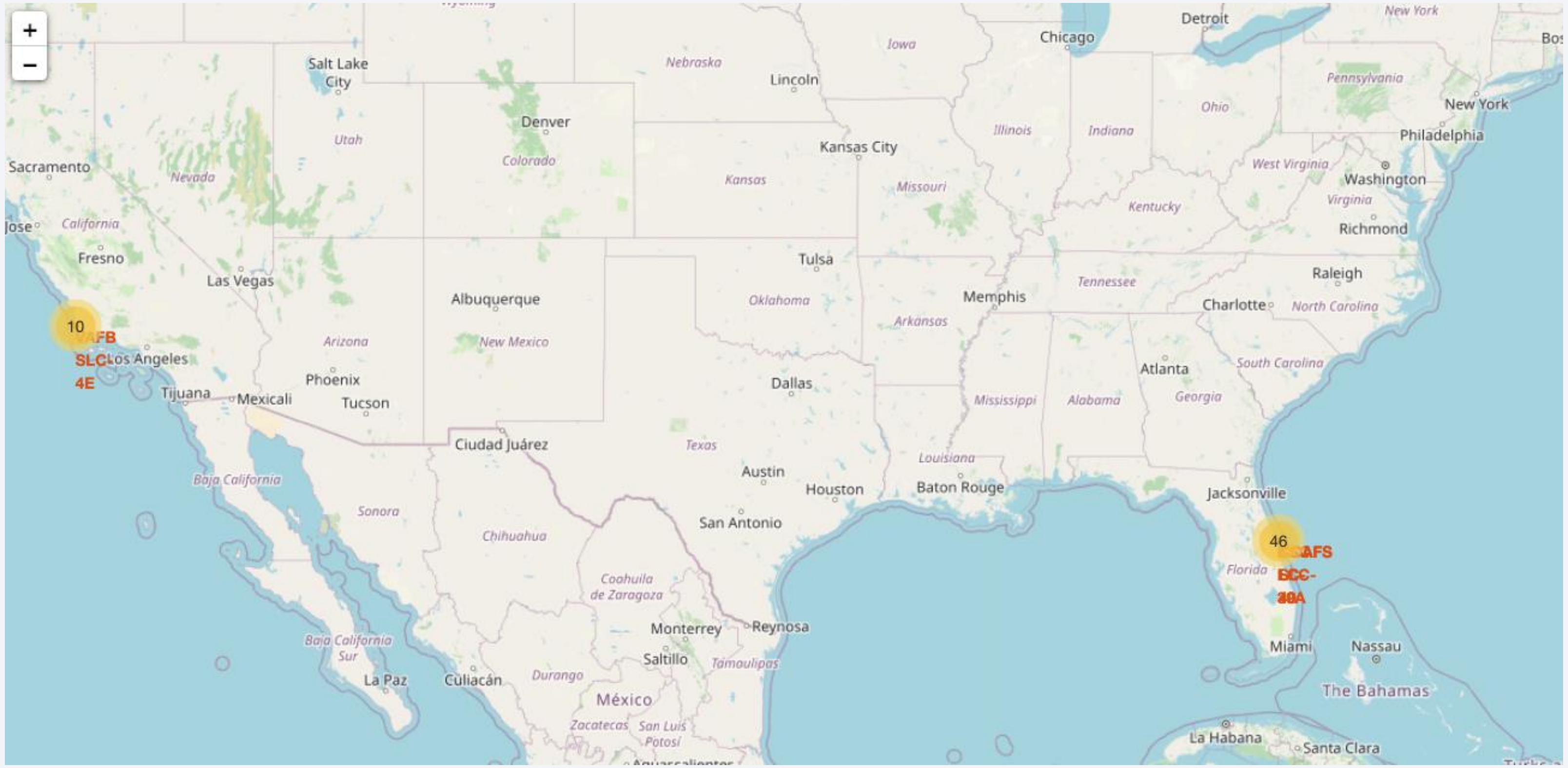
All Launch Sites global map markers

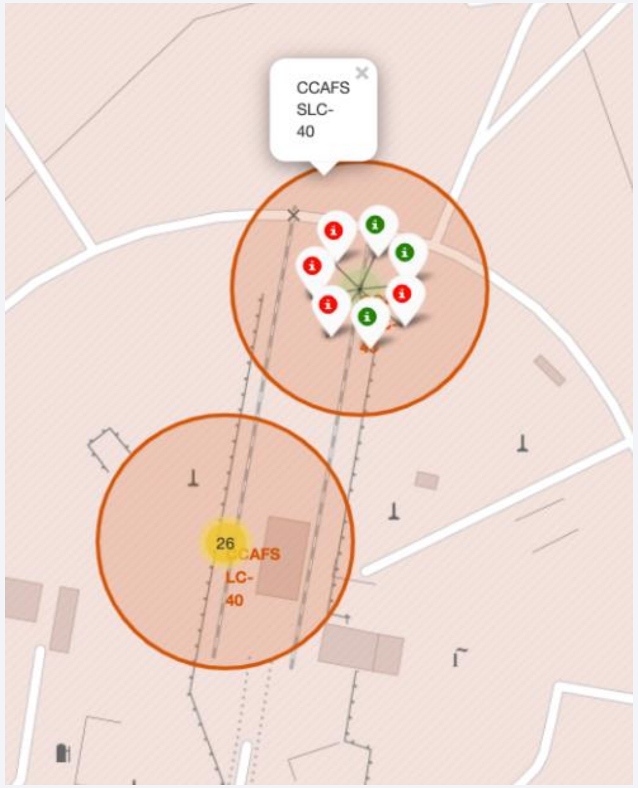
- All launch sites can be seen in the coastal areas of United States of America – Florida and California



Success/Failure Launches for each site

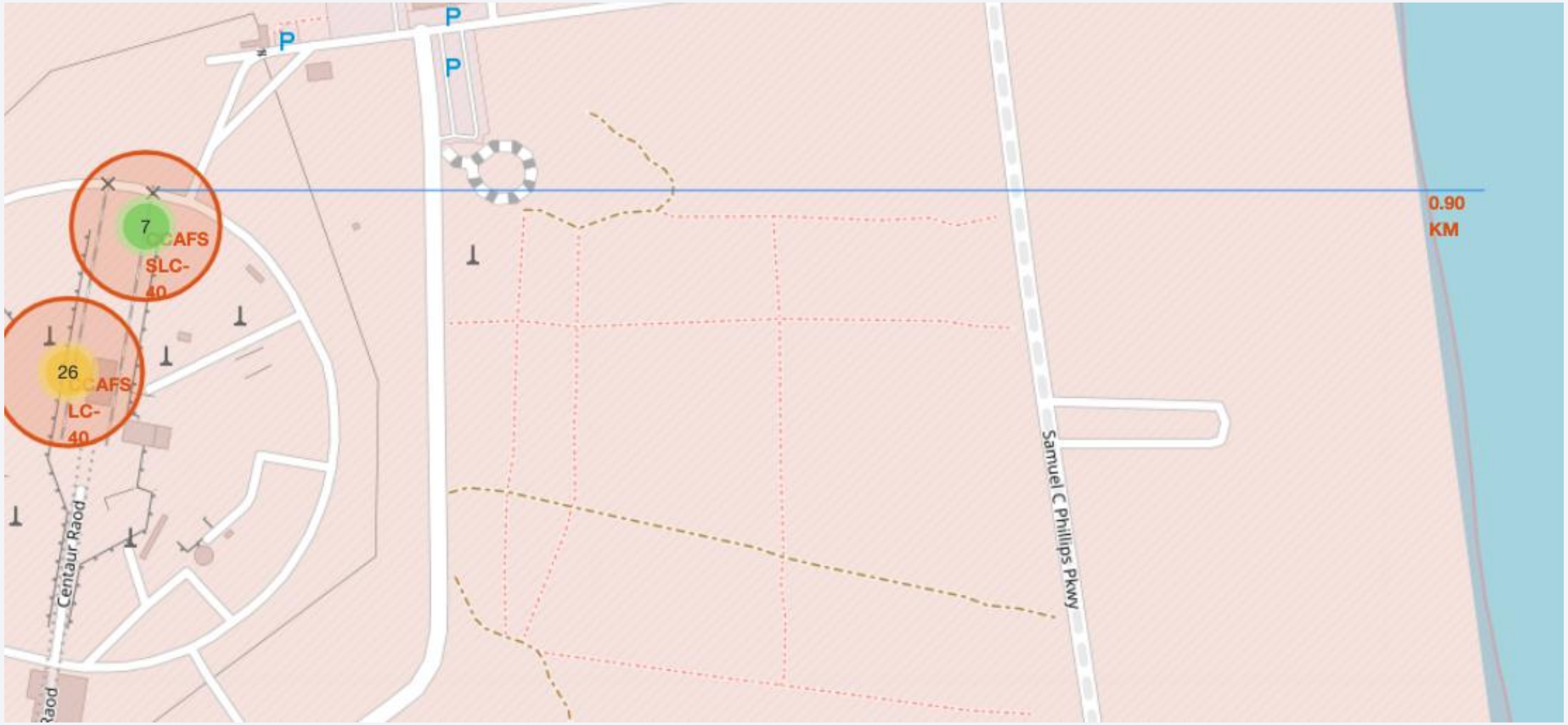
First map shows clusters for every launch site, second shows a green marker for successful and red marker for failed launch sites





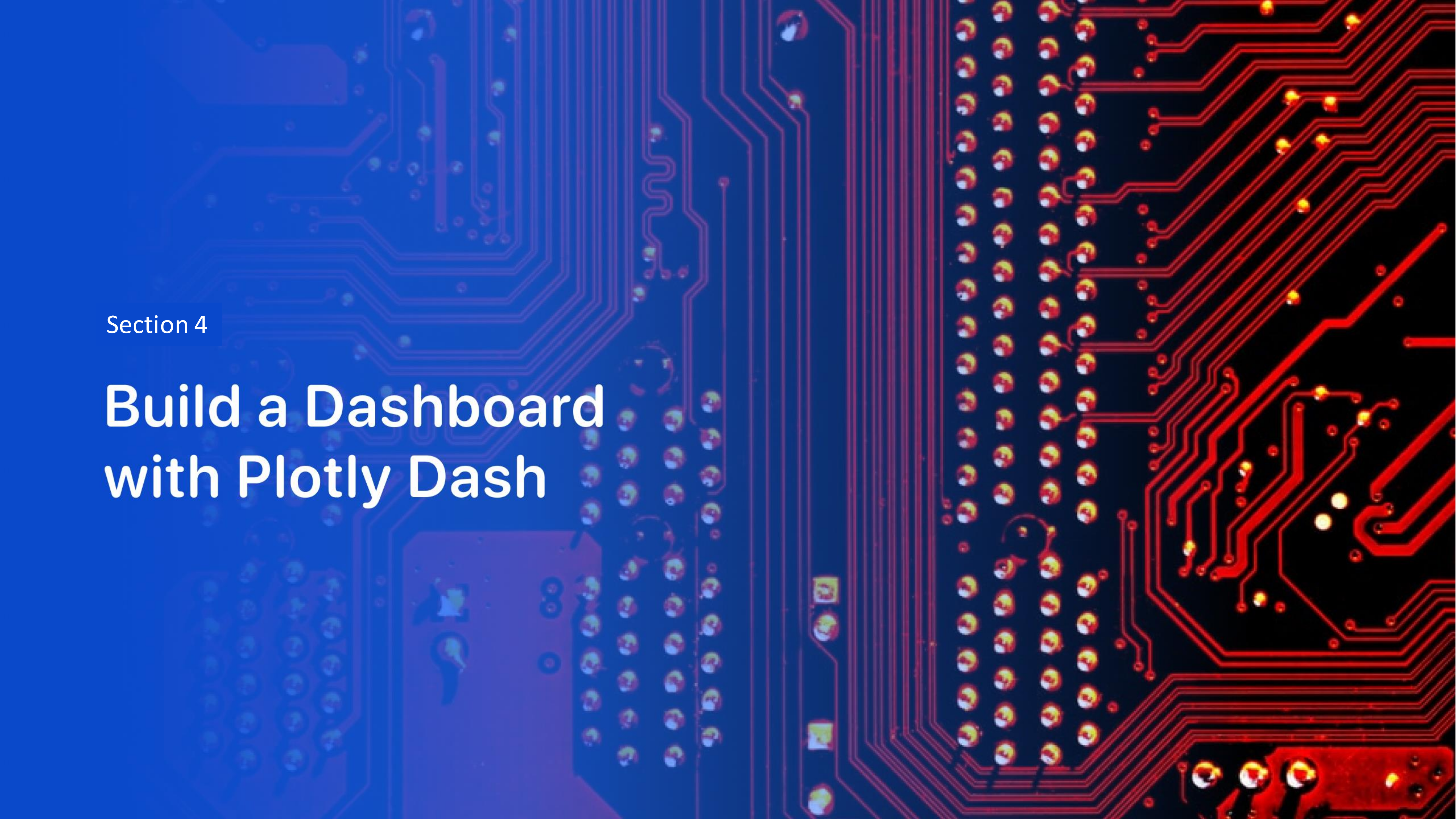
A Launch site and its Proximities

- Launch sites can be seen near to roads, highways and coastline. Also there is a minimum substantial safe distance from the cities.



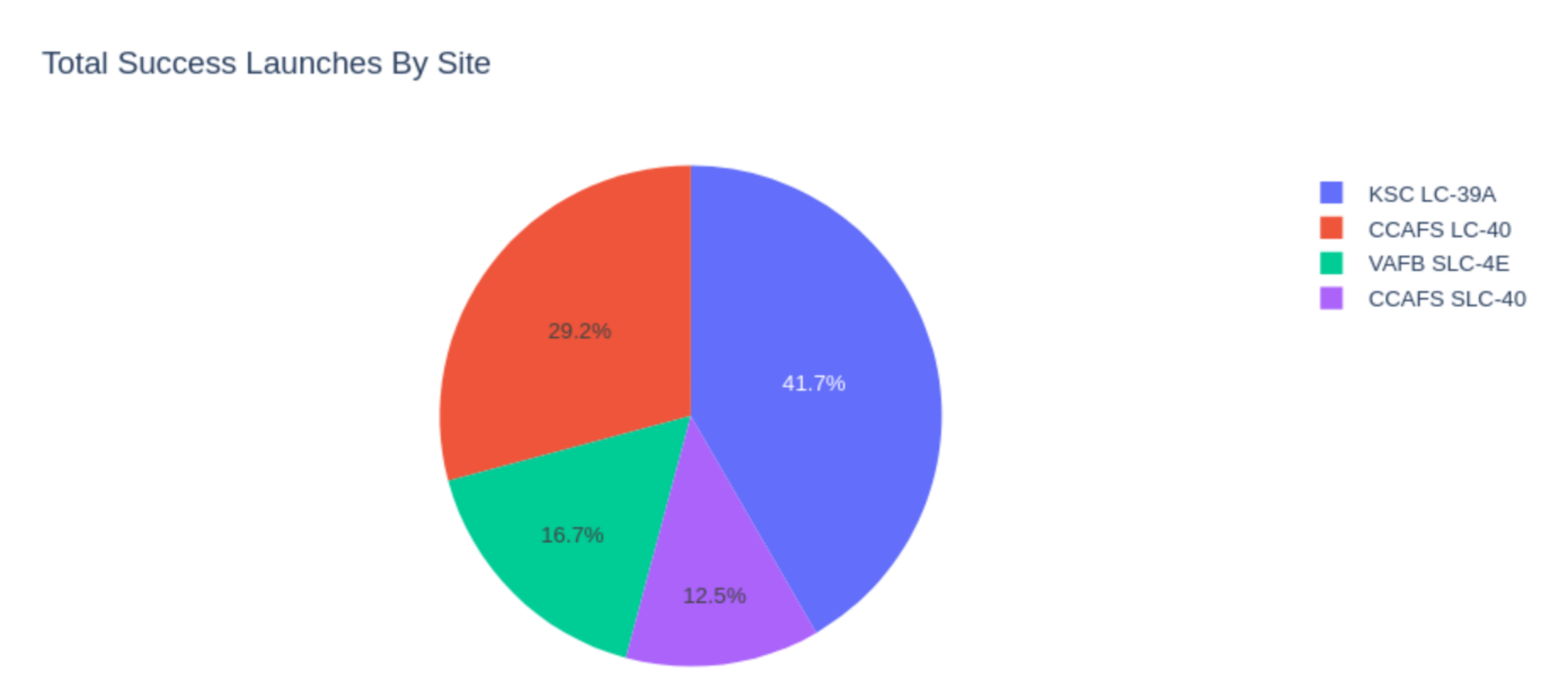
Section 4

Build a Dashboard with Plotly Dash



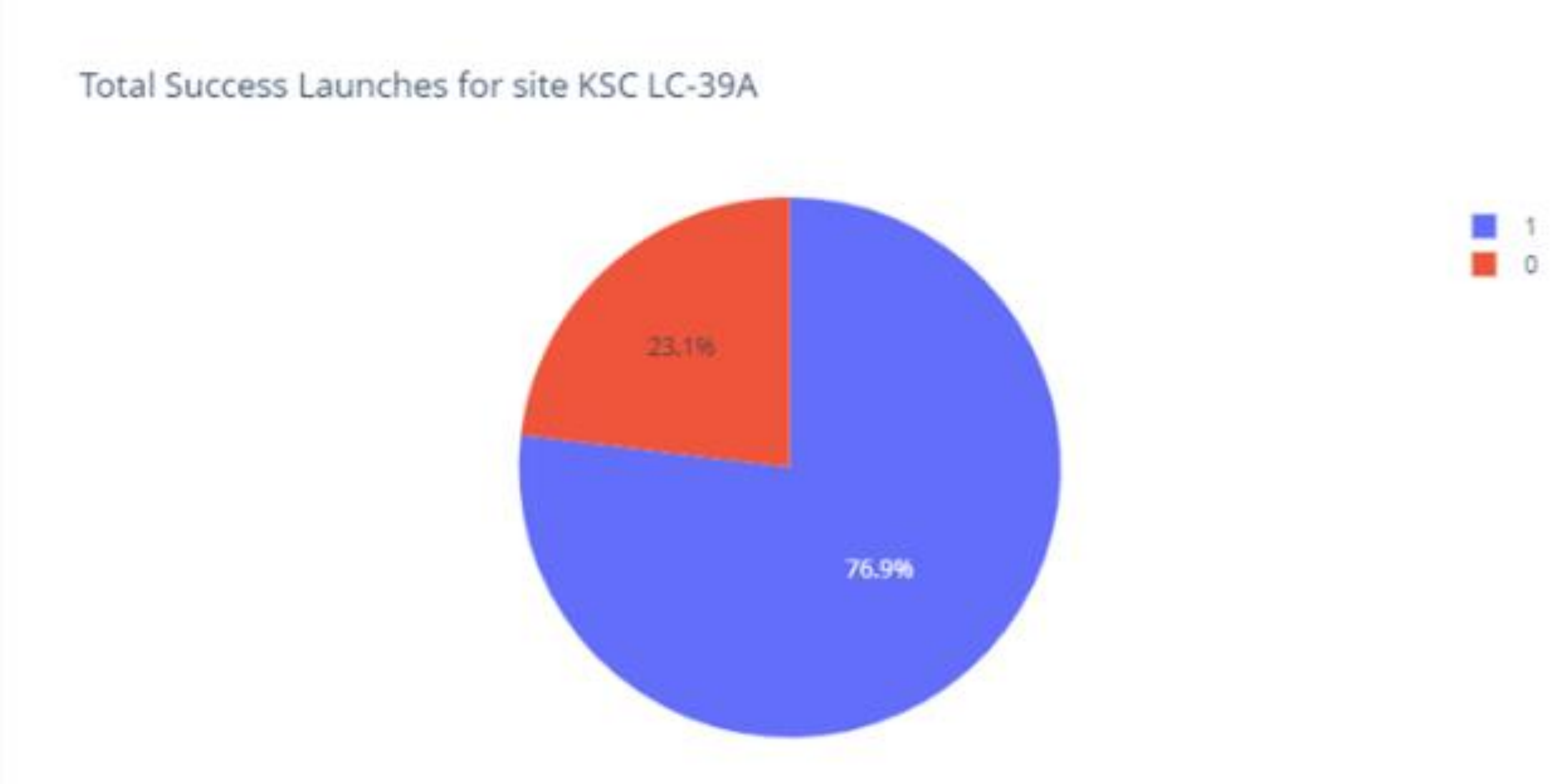
Total Successful Launches by Site

- SC LC-39 A has recorded a distinctive and highest successful launch events followed by the next distinct winner in CCAFS LC-40



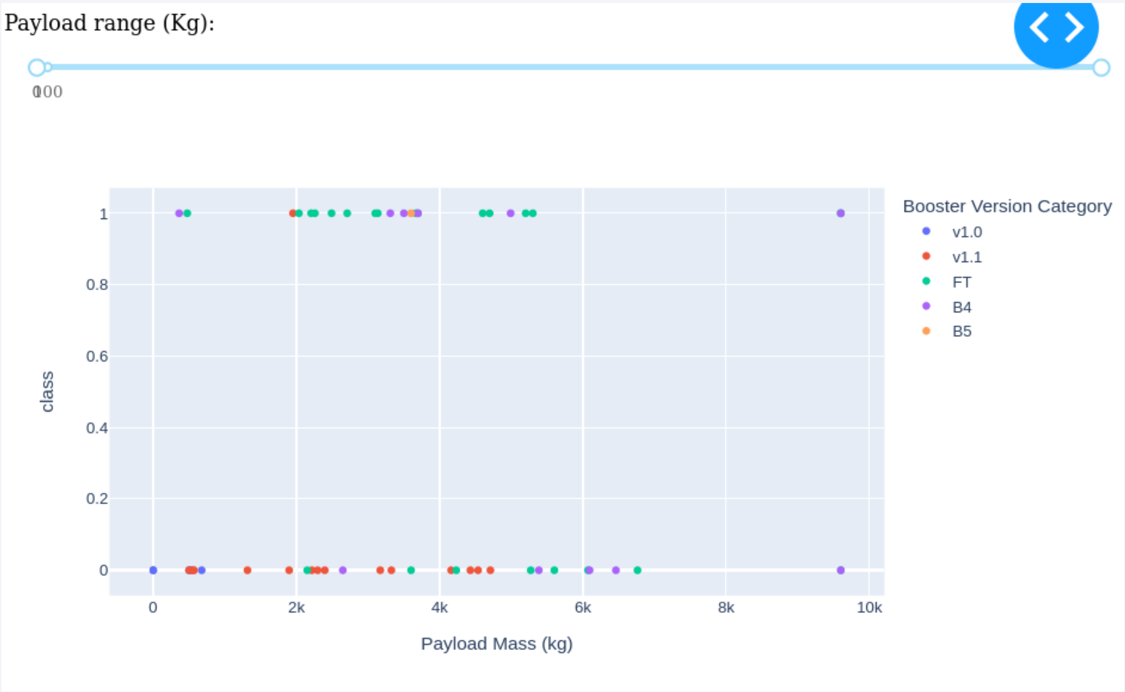
KSC LC-39A the highest successful launch site

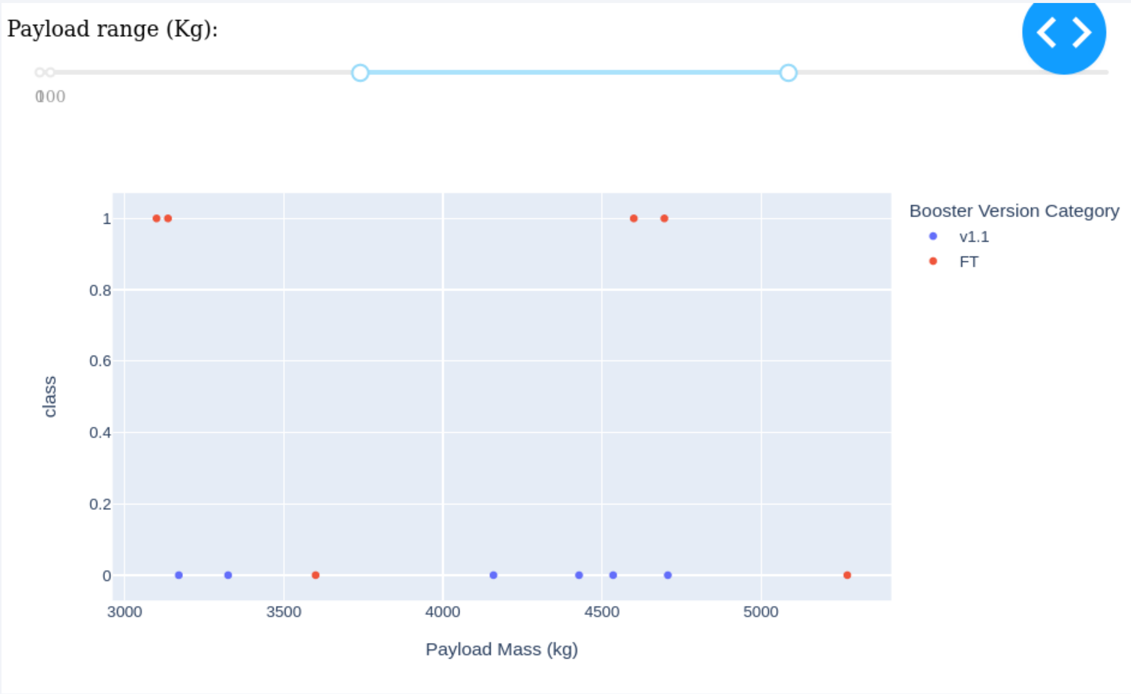
- Pie chart for launch site KSC LC-39A breakup of success to failure classes.



Payload Range Distribution

- Distribution of payload ranges against their Booster version categories
- Most payload concentration can be seen between 2000 to 7500 kgs



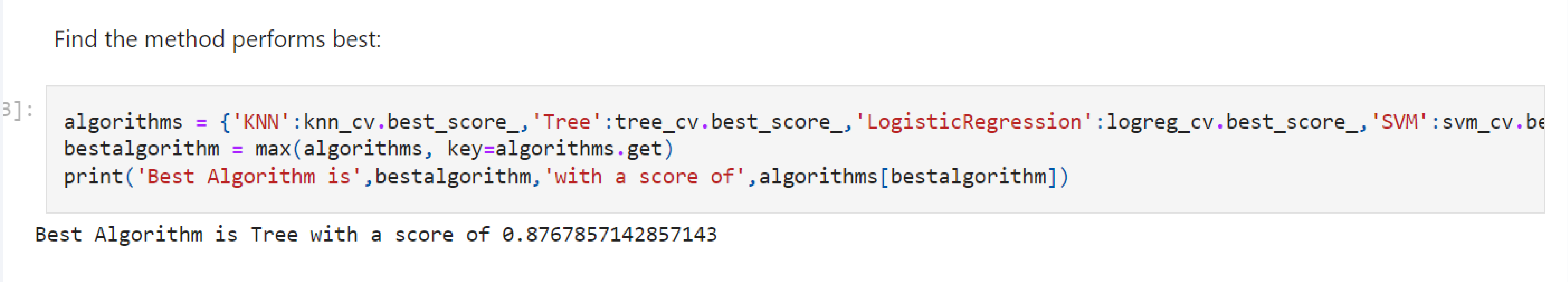


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree Classifier Model has the highest accuracy when compared to the rest



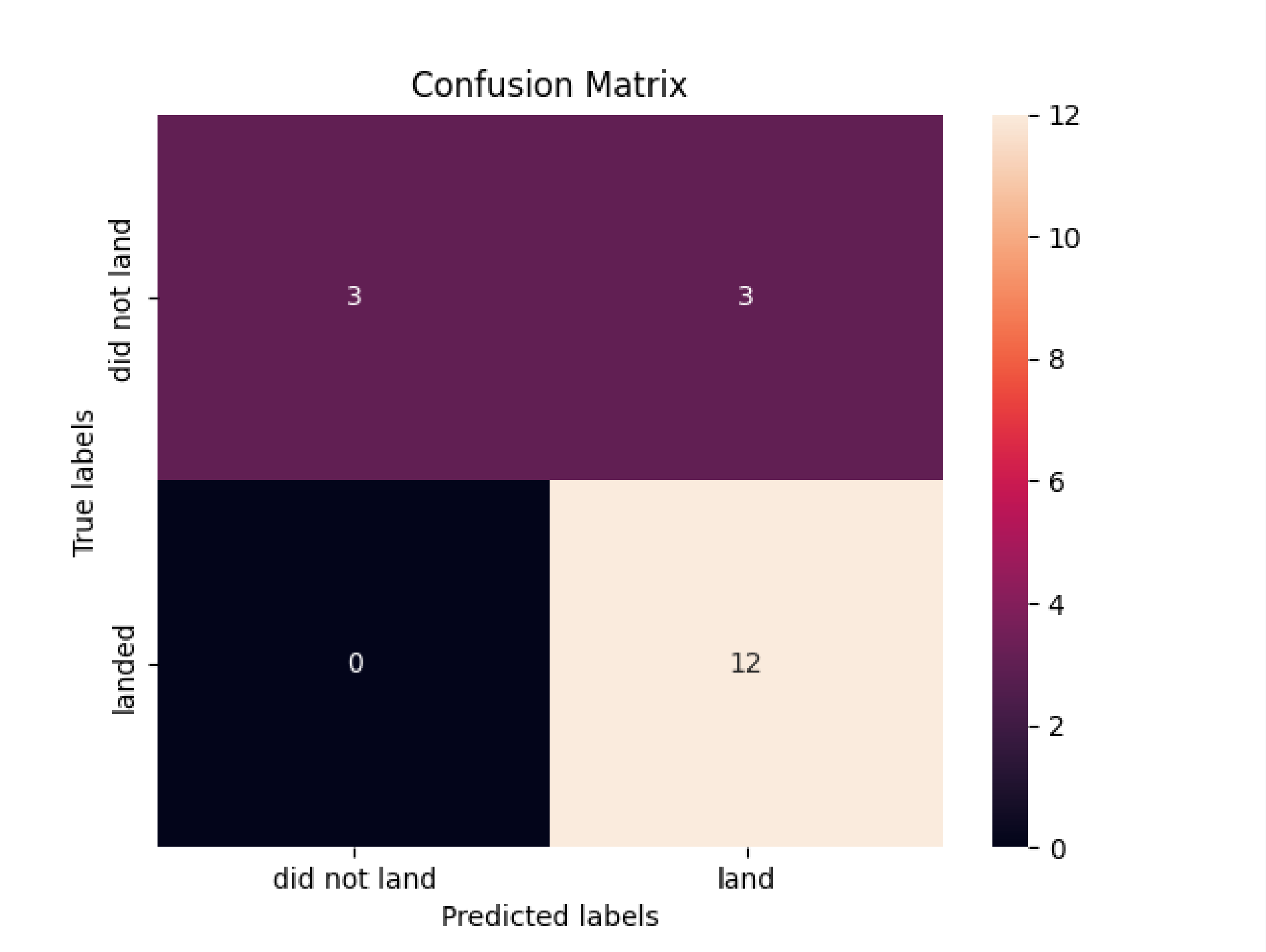
Find the method performs best:

```
algorithms = {'KNN':knn_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_,'SVM':svm_cv.be
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
```

```
Best Algorithm is Tree with a score of 0.8767857142857143
```

Confusion Matrix

- The Confusion Matrix for the decision tree classifier shows different classes. The major problem are the false positives which is the unsuccessful landing predicted as successful landings by the classifier.



Conclusions

- Heavier the payload mass of the flight, better are the success rates of launch
- Orbits ES-L1, GEO, HEO, SSO and VLEA had the most success rates
- Launch success rates steadily increased from 2013
- KSG LC-39A has the most successful launch records amongst all launch sites
- Decision Tree Classifier is the best Machine Learning model used to predict the successful landing outcomes
- False positives are the outliers of this exercise which is yet to be controlled by any model

Appendix

All the Notebooks and DataSets used in this project can be found in this GitHub Repository link

Applied Data Science Capstone Project

Thank you!