IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Roozbeh Khodayari
02/05/2025

Outline

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

Executive Summary

- **Summary of methodologies**

  - Data collection

  - Data wrangling

  - EDA-SQL

  - EDA-Data Visualization

  - Building an interactive analysis with Folium

  - Building a Dashboard with Plotly Dash

  - Predictive analysis (Classification)

- **Summary of all results**

  - Exploratory Data Analysis results

  - Interactive analytics in screenshots

  - Predictive analysis results

Introduction

- **Project background and context**

  In this capstone, we will 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.

- **Problems you want to find answers**

  -What factors are significant on a landing successful?

  -The interaction amongst various features that determine the success rate of a successful landing.

  -What is the best algorithm that can be used for predictive classification in this case?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology**

  - Data was collected using SpaceX API and web scraping.

- **Perform data wrangling**

  - Filtering the data

  - Dealing with missing values

  - Using One Hot Encoding to prepare the data for Machine Learning model

- **Perform exploratory data analysis (EDA) using visualization and SQL**

- **Perform interactive visual analytics using Folium and Plotly Dash**

- **Perform predictive analysis using classification models**

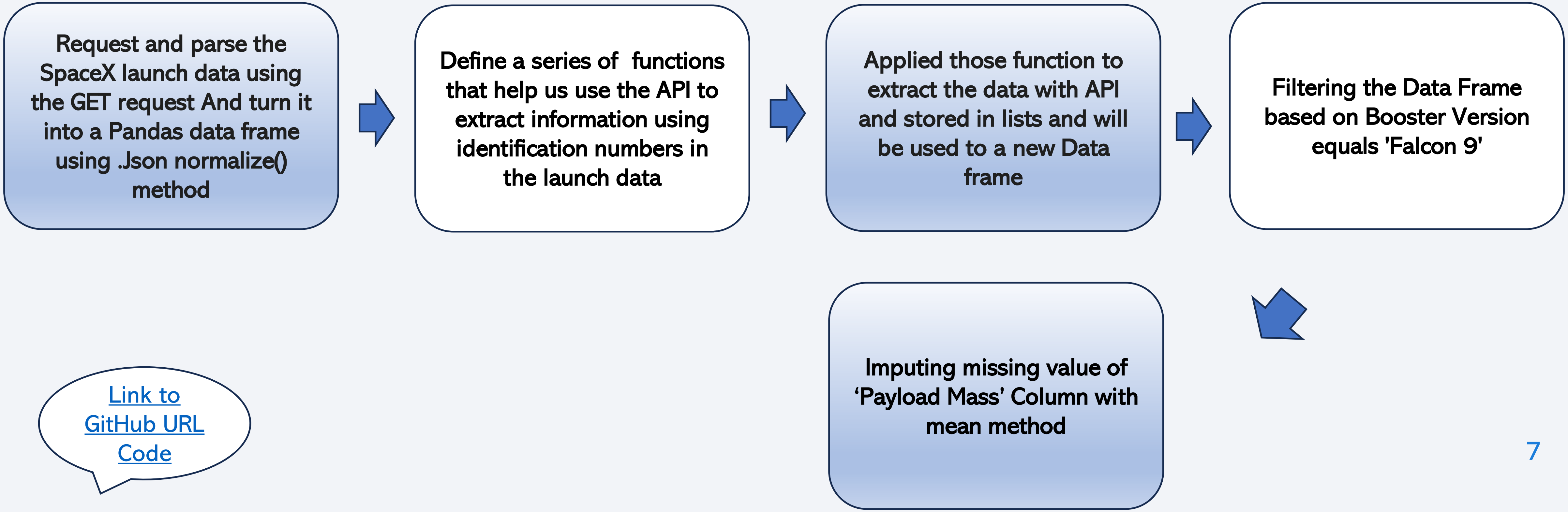
  - Predicting, Tuning and Evaluation of classification models to find the best model

Data Collection SpaceX API

- **The key difference between scraping and API**

API: You're using an official interface designed for programmatic access

Scraping: You're extracting data that wasn't necessarily meant to be accessed programmatically



Request and parse the SpaceX launch data using the GET request And turn it into a Pandas data frame using .Json normalize() method

Define a series of functions that help us use the API to extract information using identification numbers in the launch data

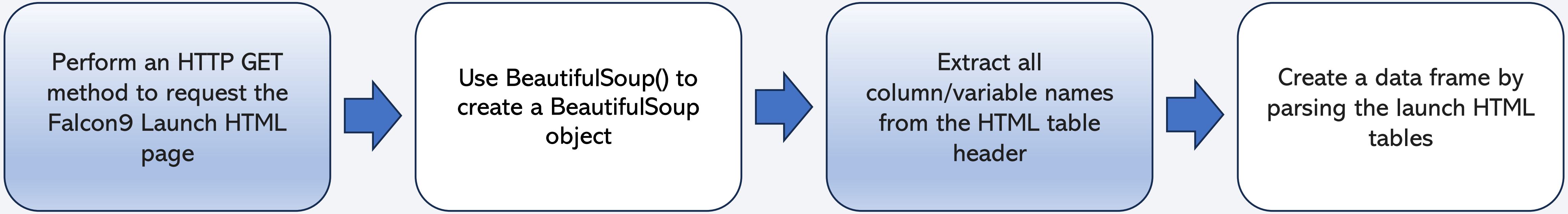
Applied those function to extract the data with API and stored in lists and will be used to a new Data frame

Filtering the Data Frame based on Booster Version equals 'Falcon 9'

Imputing missing value of 'Payload Mass' Column with mean method

Link to GitHub URL Code

Data Collection - Scraping



Perform an HTTP GET method to request the Falcon9 Launch HTML page

Use BeautifulSoup() to create a BeautifulSoup object

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Link to GitHub URL Code

Data Wrangling

- We need to check each categorical column to ensure not having the value-in-consistency
- There are some categories in landing outcome column so that need to collapsing them to 2 categories :1 represented did land successfully landed and 0 represented did not land successfully.
- For figure outing those cases as well as determining success rate, we should do these steps below:

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome of the orbits
4. Collapsing categories of landing outcome column to 2 categories (0 and 1)
5. Assign the result of before step to new column named class
6. Use mean() method on class column to determine success rate

Link to GitHub URL Code

EDA with Data Visualization

- We used seaborn package to visualize relationship between columns so that figure it out any pattern or relationship.
- Three kind of plot have plotted include: scatterplot, bar plot and line plot.

- **Scatter plot**
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass

- **Bar plot**
  - Success rate vs. Orbit

- **Line plot**
  - Success rate vs. Year

[Link to GitHub URL Code]

EDA with SQL

- **We applied EDA with SQL to achieve insight, these query are:**
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass
  - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
  - Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the 2 dates in descending order

Link to GitHub URL Code

Build an Interactive Map with Folium

Markers of all Launch Sites

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Colored Markers of the launch outcomes for each Launch Site

- Added colored Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities

- Added colored Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Link to GitHub URL Code

Build a Dashboard with Plotly Dash

Dashboard includes dropdown, pie chart, range slider and scatter plot components

- Dropdown helps users to choose the launch site.
- Pie chart demonstrates the total success and the total failure for the launch site chosen with the dropdown component.
- Range slider allows users to select a payload mass in a fixed range.
- Scatterplot shows the relationship between two variables, Success vs Payload Mass.

Link to GitHub URL Code

Predictive Analysis (Classification)

- Creating a NumPy array from the column Class in data, assigning to variable Y.
- Standardizing the data in X then assign it to the variable X.
- Splitting data into training and test sets.
- Selection of algorithms (logistic regression, support vector machine, decision tree classifier, k nearest neighbors).
- Setting parameters for each algorithm, then training models with GridSearchModel method.
- Getting the best hyperparameters for each type of model.
- Calculating and Computing accuracy for each model with test set.
- The model with the best accuracy will be reported.

Link to GitHub URL Code

Results

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

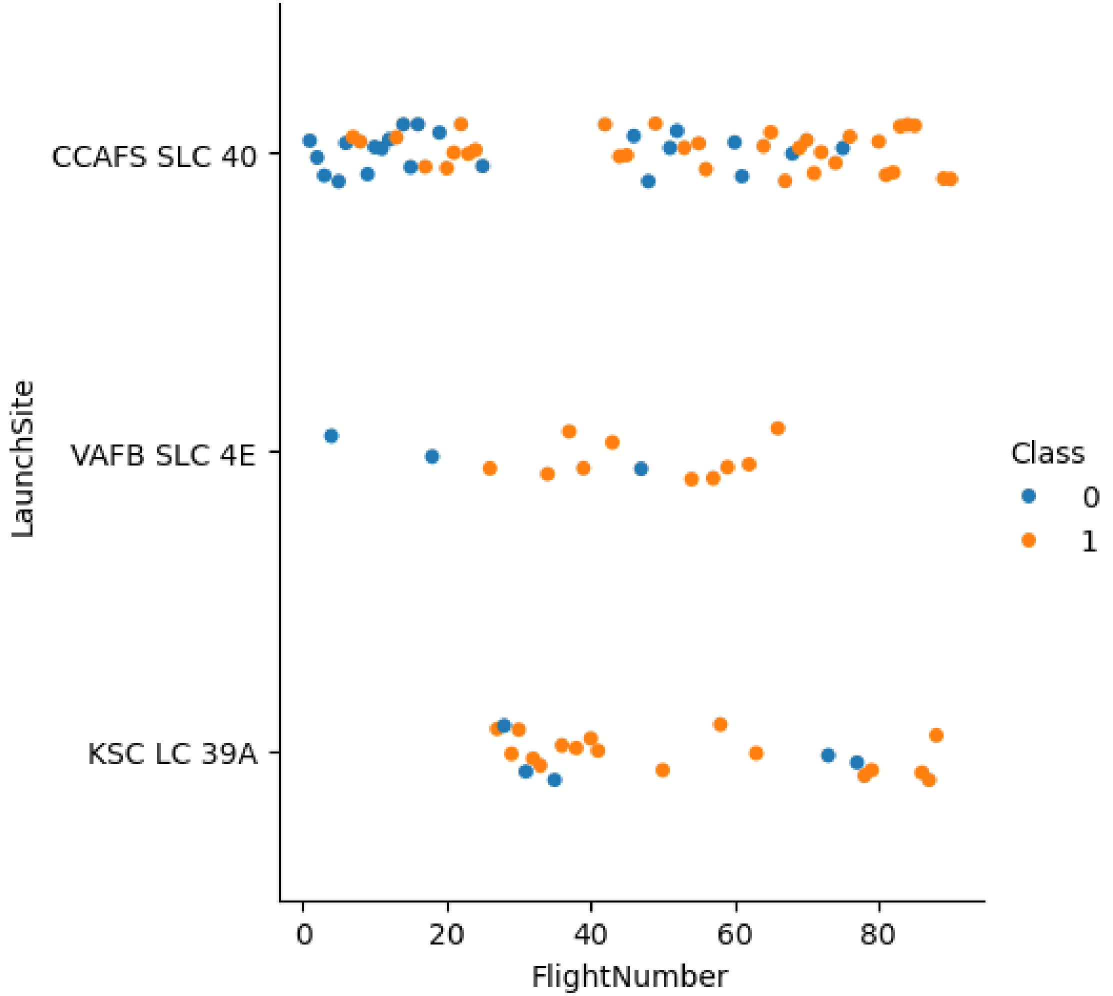
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

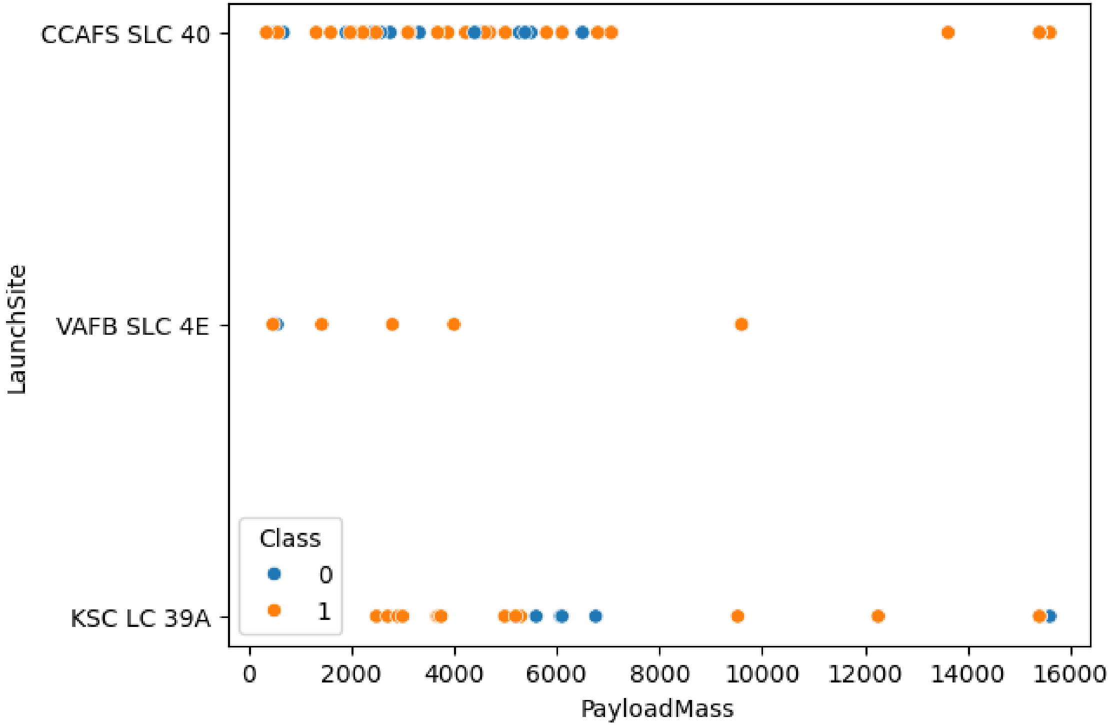
- VAFB SLC 4E and KSC LC 39A have higher success rates.

- The success rate increase by larger the flight number in each launch site.



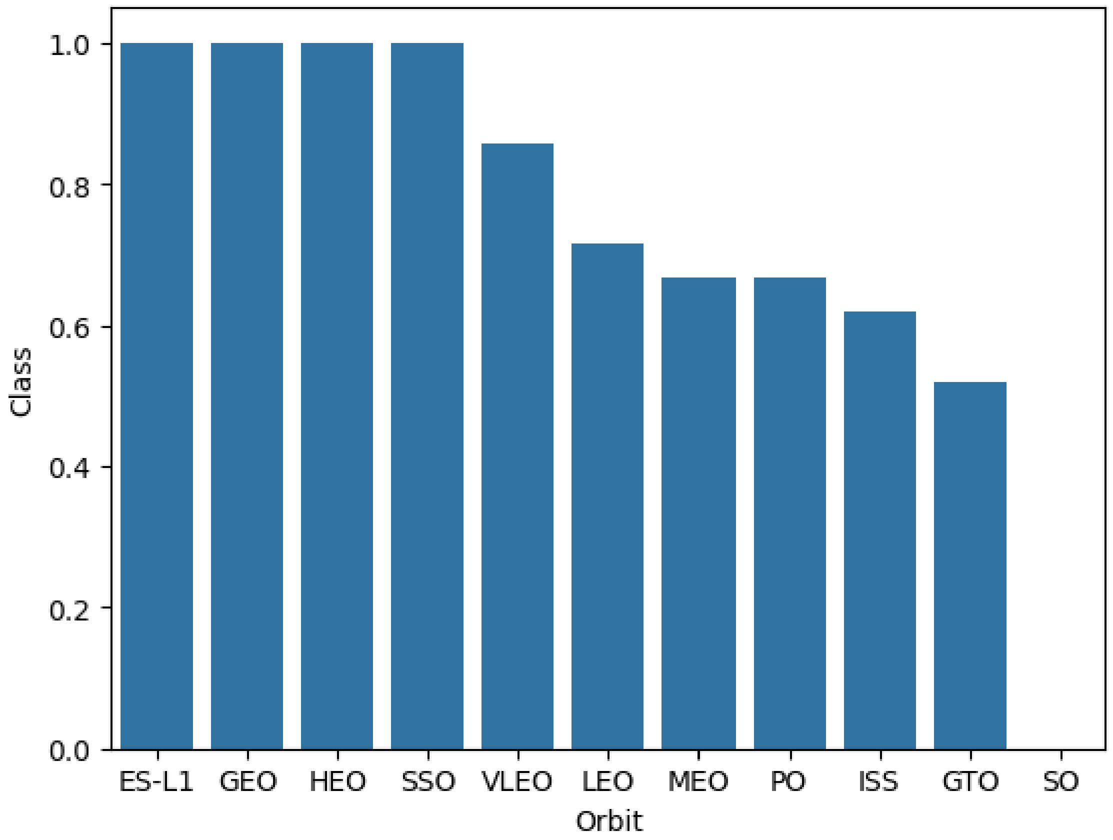
Payload vs. Launch Site

- For every launch site the higher payload mass, the higher success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.



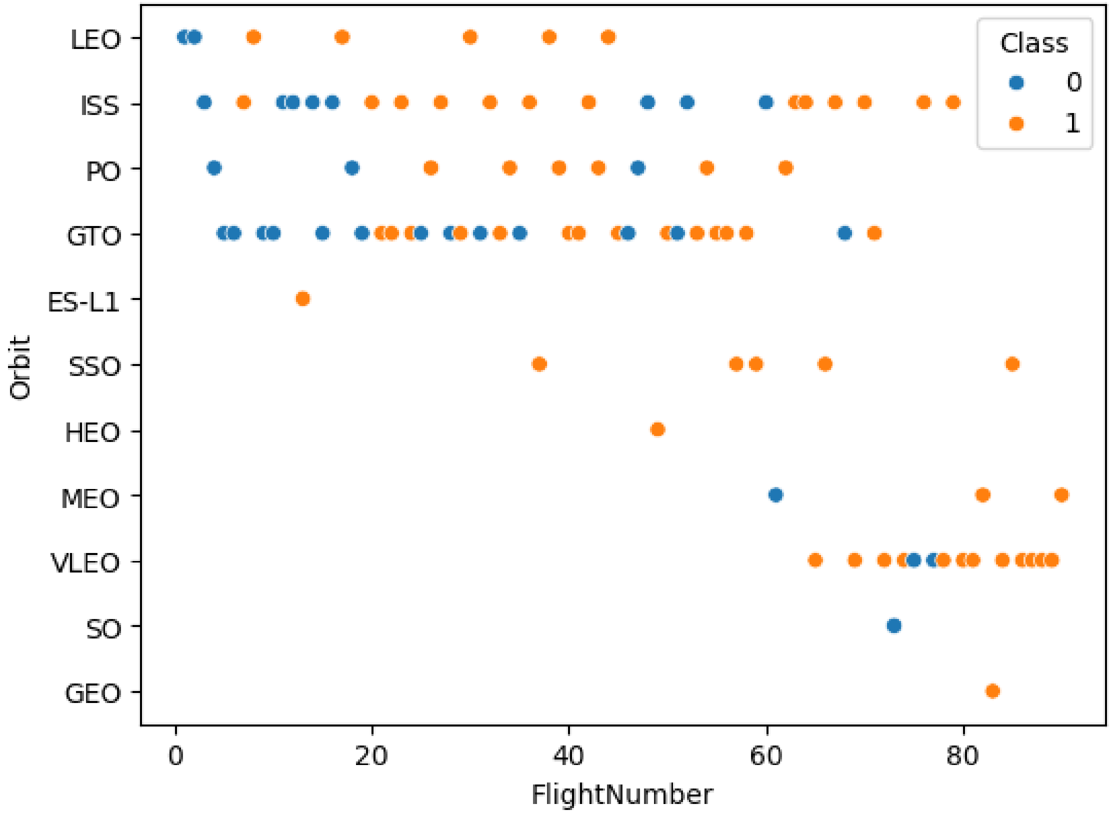
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO Orbits have success rate equal 100%
- SO Orbits has success rate equal 0%



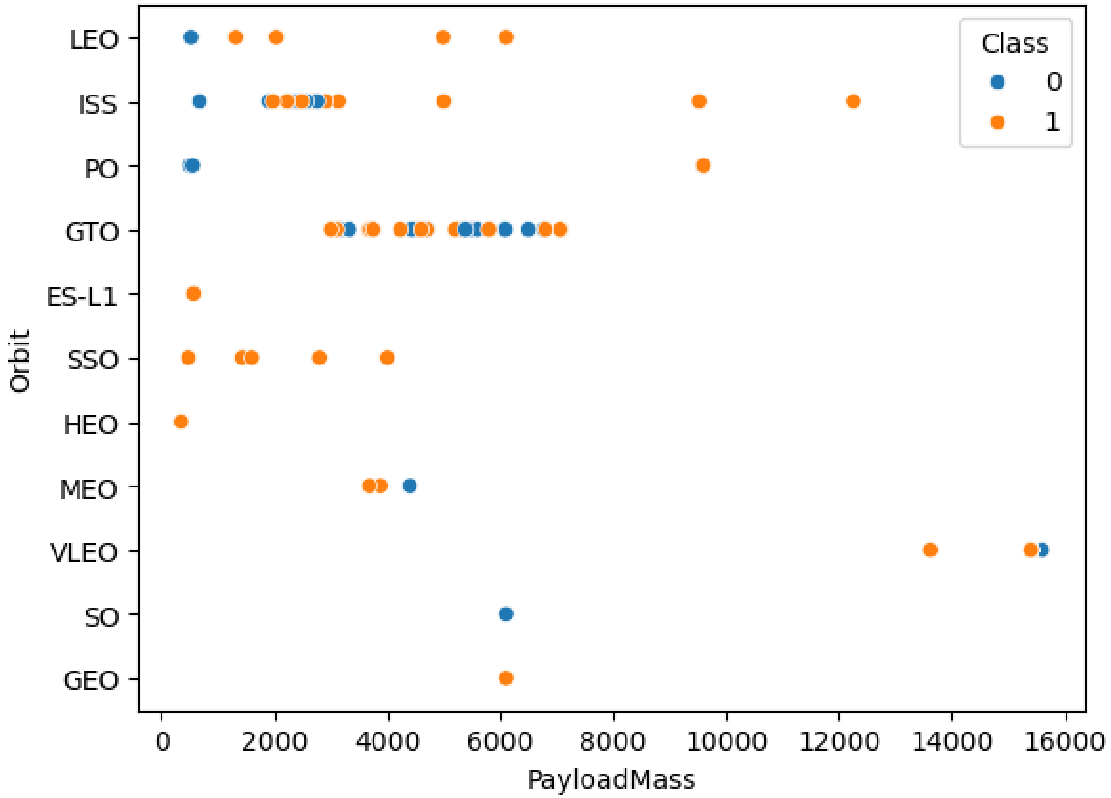
Flight Number vs. Orbit Type

- The LEO orbit, success rate has relationship with the number of flights, but in the GTO orbit, there is no relationship between success rate and flight number.



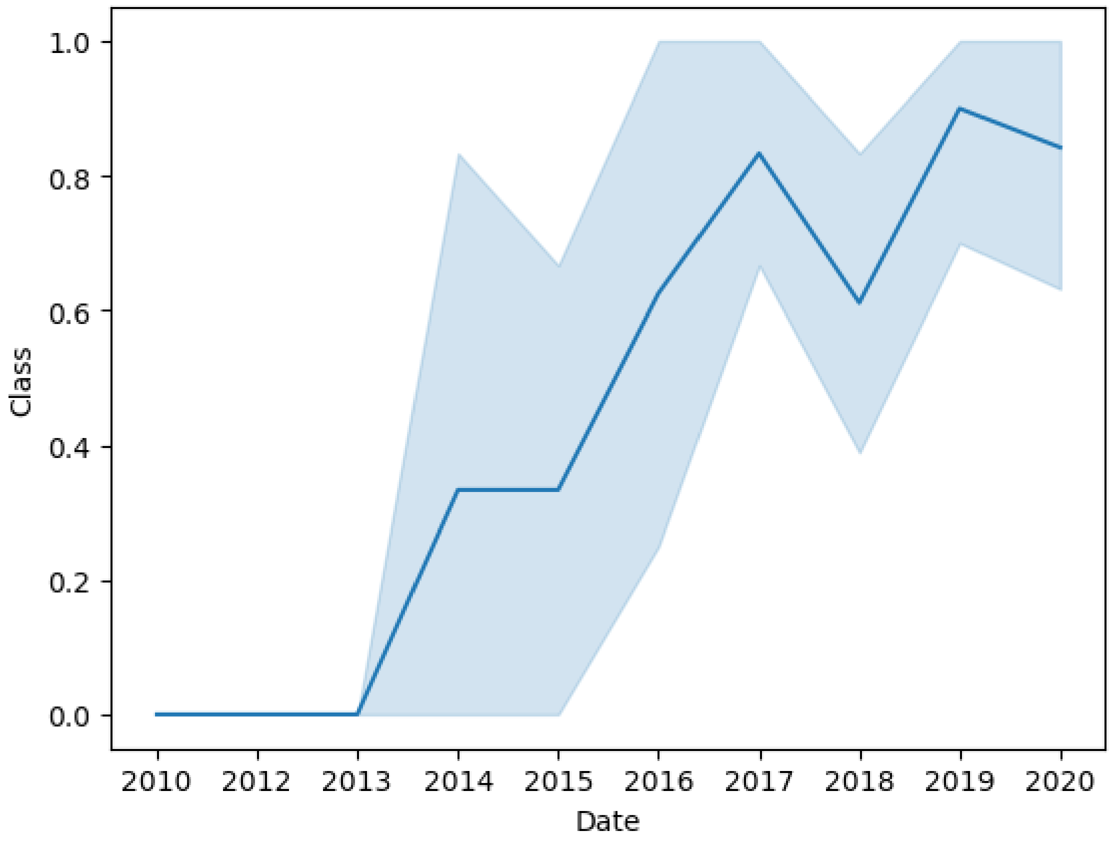
Payload vs. Orbit Type

- For the LEO orbit, Heavier payload Mass improve the success rate.
- For GTO, ES-L1, SSO,HEO orbit, the success rate are independent from payload Mass.



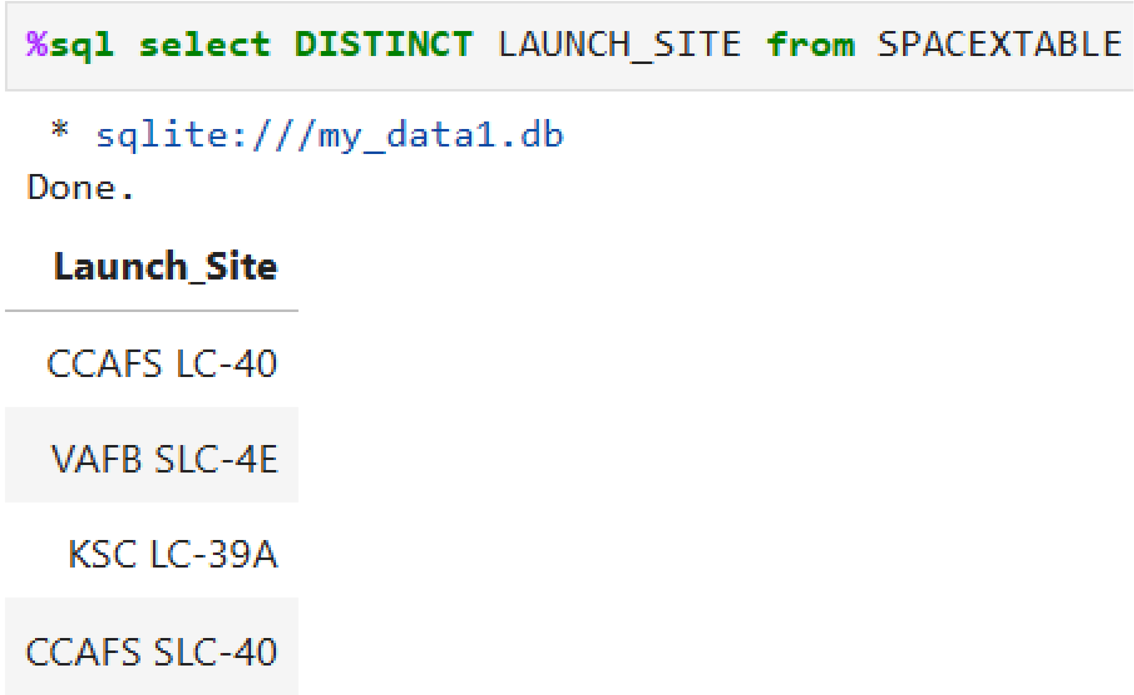
Launch Success Yearly Trend

- The success rate kept increasing since 2013.



All Launch Site Names

- Displaying the unique launch sites in the space mission with DISTINCT query.



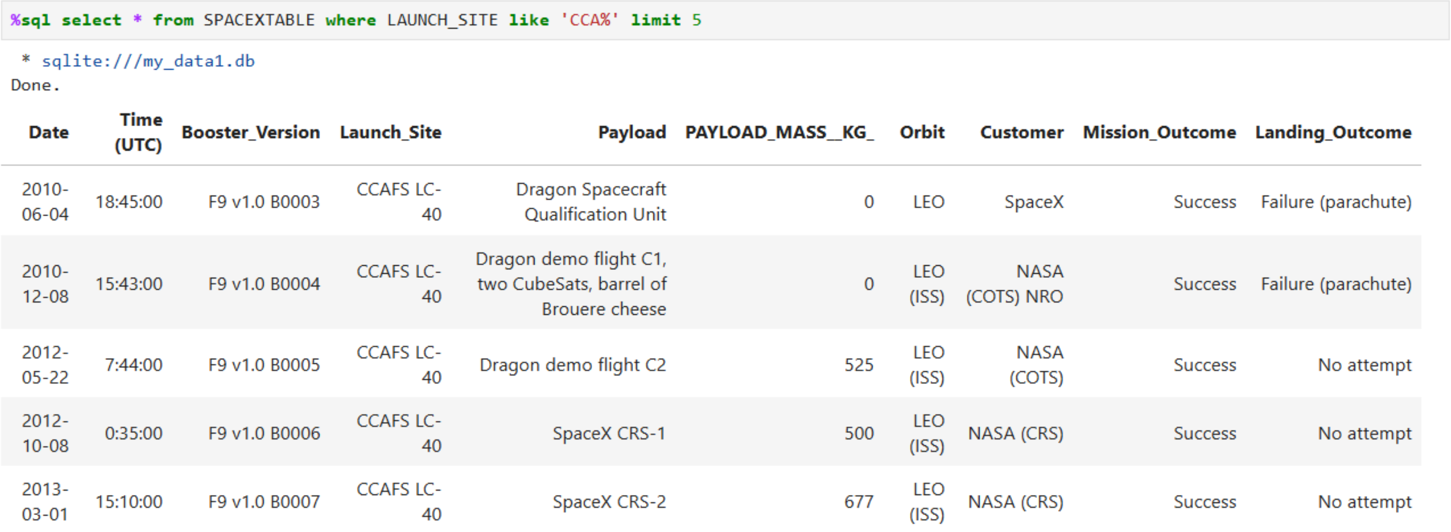
```
%sql select DISTINCT LAUNCH_SITE from SPACEXTABLE
```

```
 * 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'

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



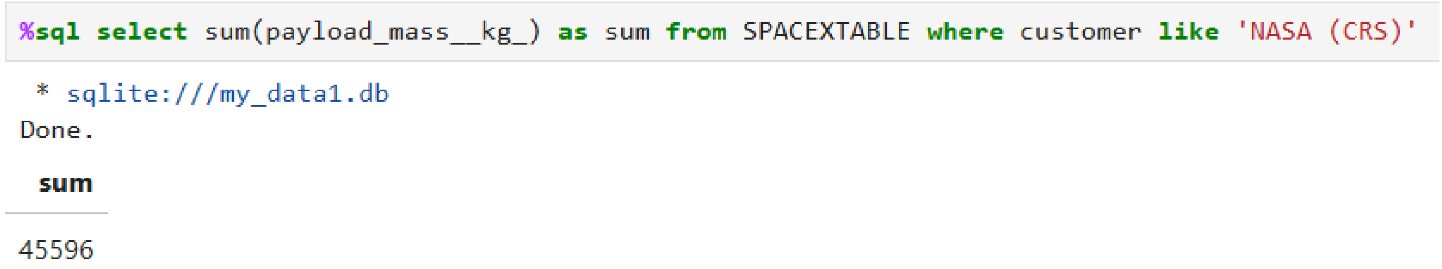
```
%sql select * from SPACEXTABLE where LAUNCH_SITE like 'CCA%' limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

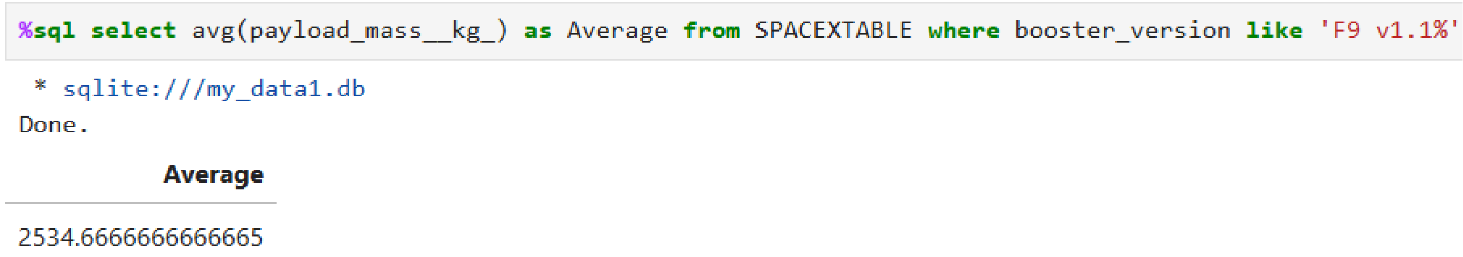


```
%sql select sum(payload_mass__kg_) as sum from SPACEXTABLE where customer like 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
  sum
45596
```

Average Payload Mass by F9 v1.1

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

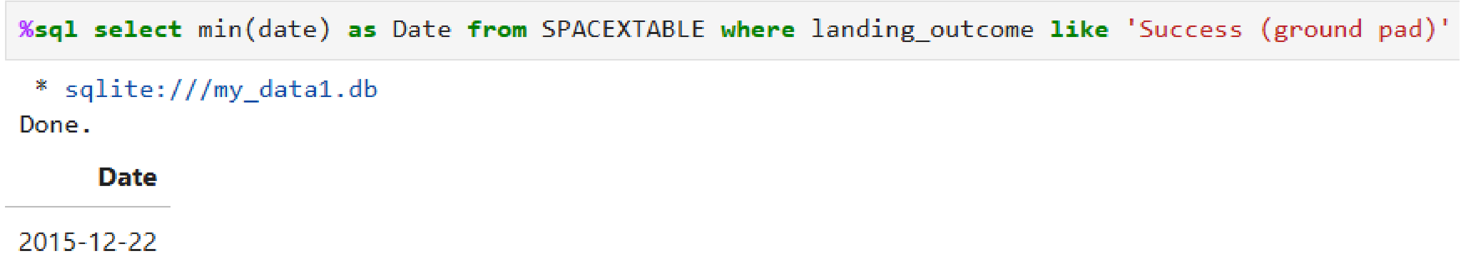


```
%sql select avg(payload_mass__kg_) as Average from SPACEXTABLE where booster_version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
           Average
2534.6666666666665
```

First Successful Ground Landing Date

- showing the first successful landing outcome’s date in ground pad



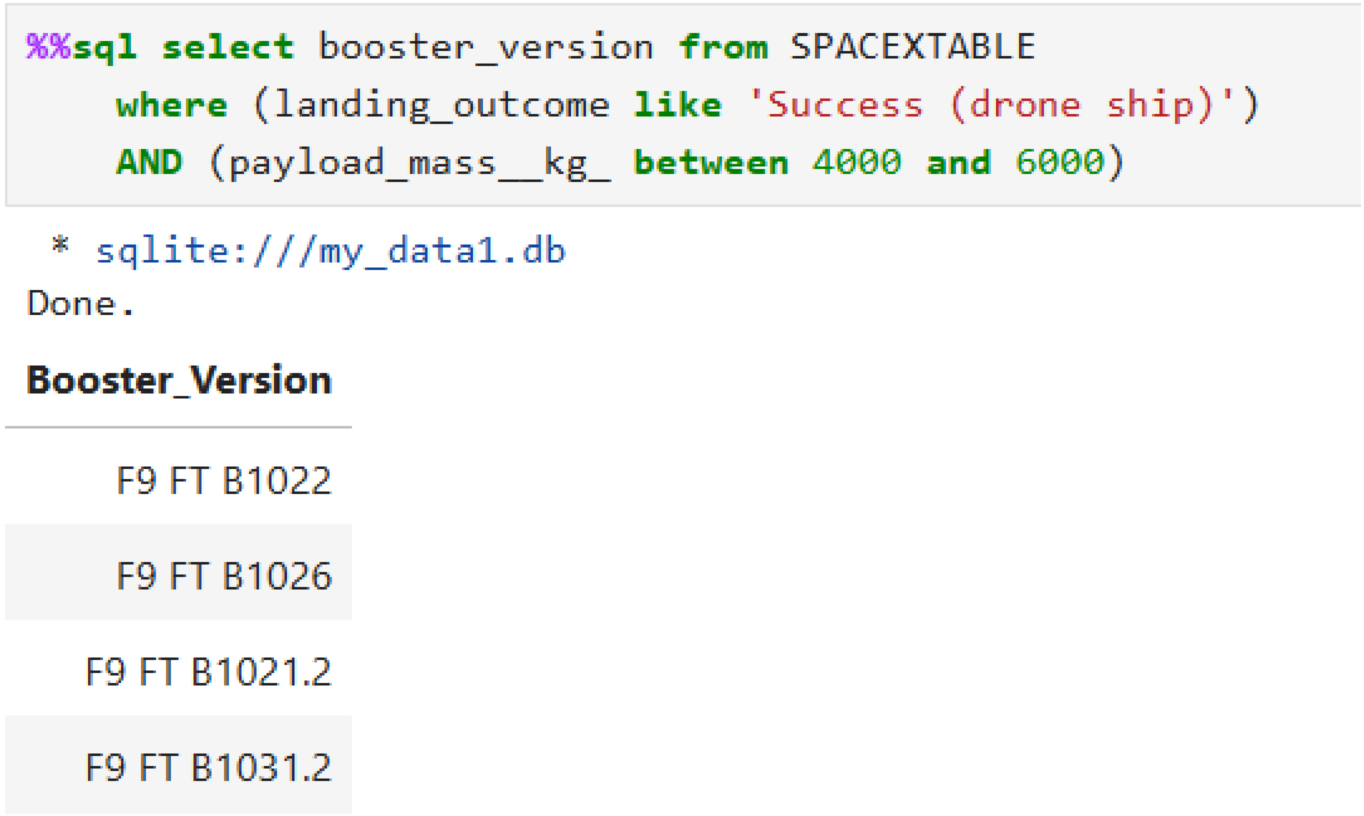
```
%sql select min(date) as Date from SPACEXTABLE where landing_outcome like 'Success (ground pad)'
```

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

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000



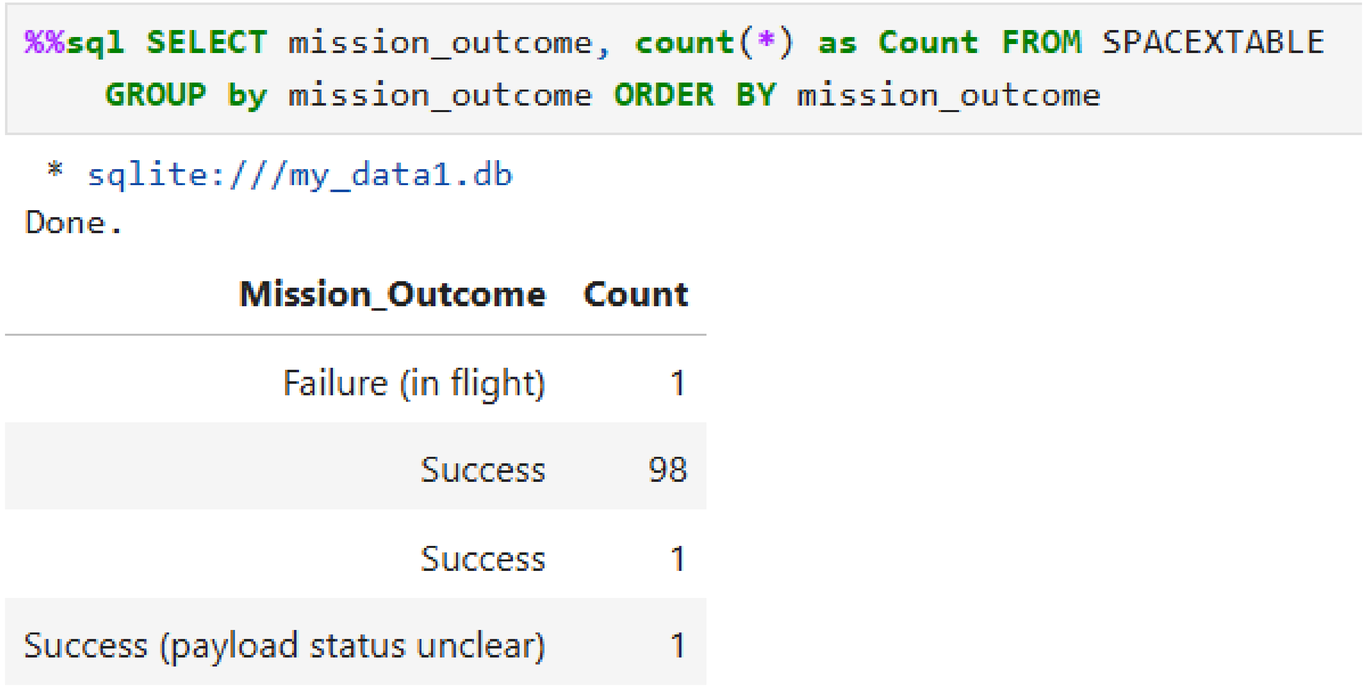
```
%%sql select booster_version from SPACEXTABLE
    where (landing_outcome like 'Success (drone ship)')
    AND (payload_mass__kg_ between 4000 and 6000)
```

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

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

Total Number of Successful and Failure Mission Outcomes

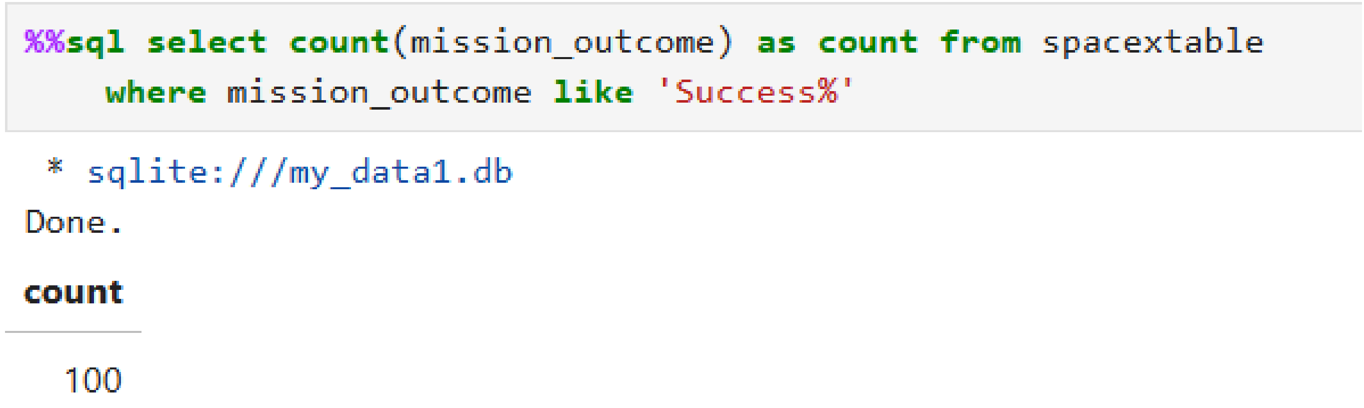
- Listing the total number of successful and failure mission outcomes



```
%%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTABLE
    GROUP by mission_outcome ORDER BY mission_outcome
```

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

Mission_Outcome	Count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1



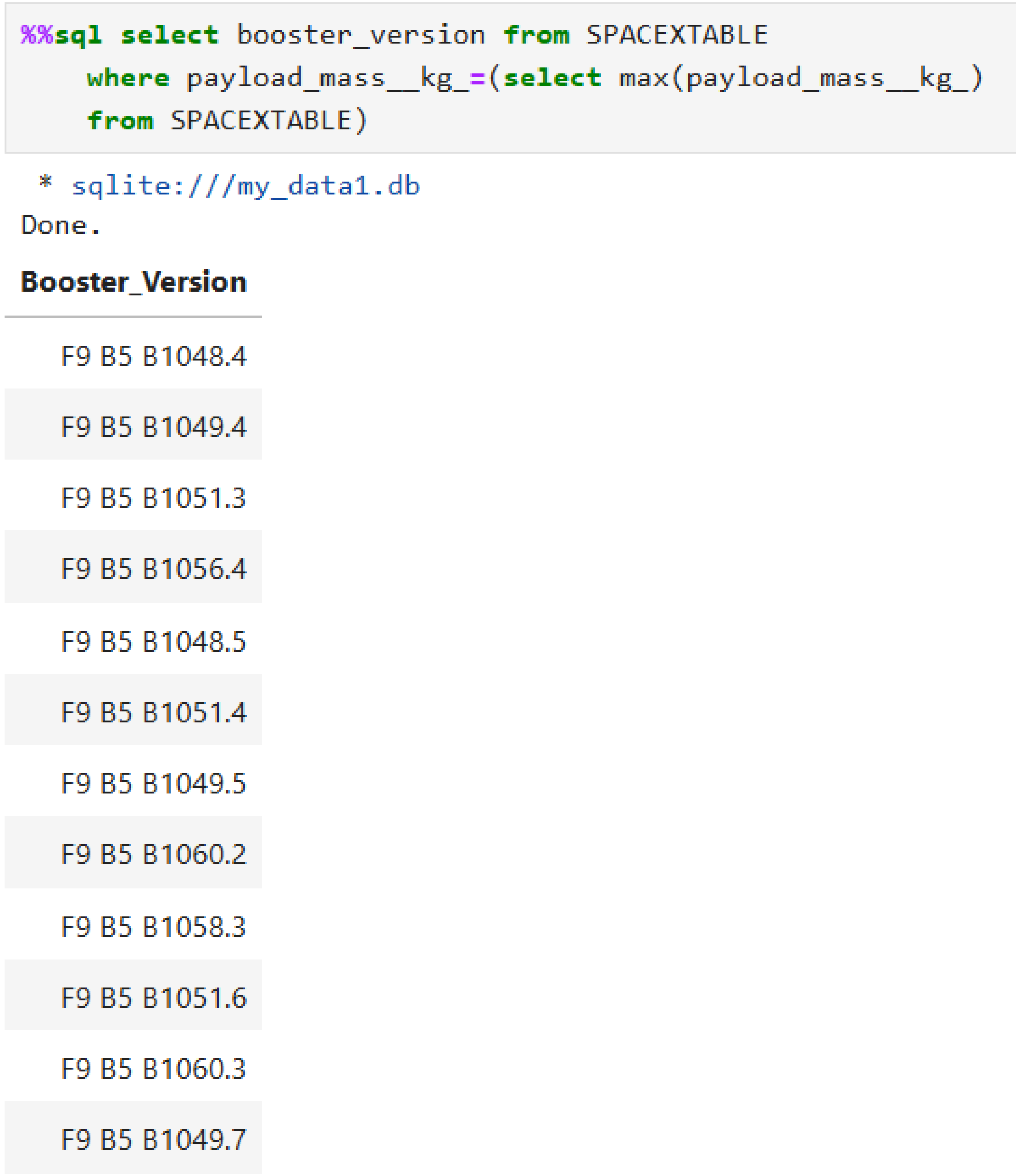
```
%%sql select count(mission_outcome) as count from spacextable
    where mission_outcome like 'Success%'
```

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

count
100

Boosters Carried Maximum Payload

- Listing the names of the booster which have carried the maximum payload mass



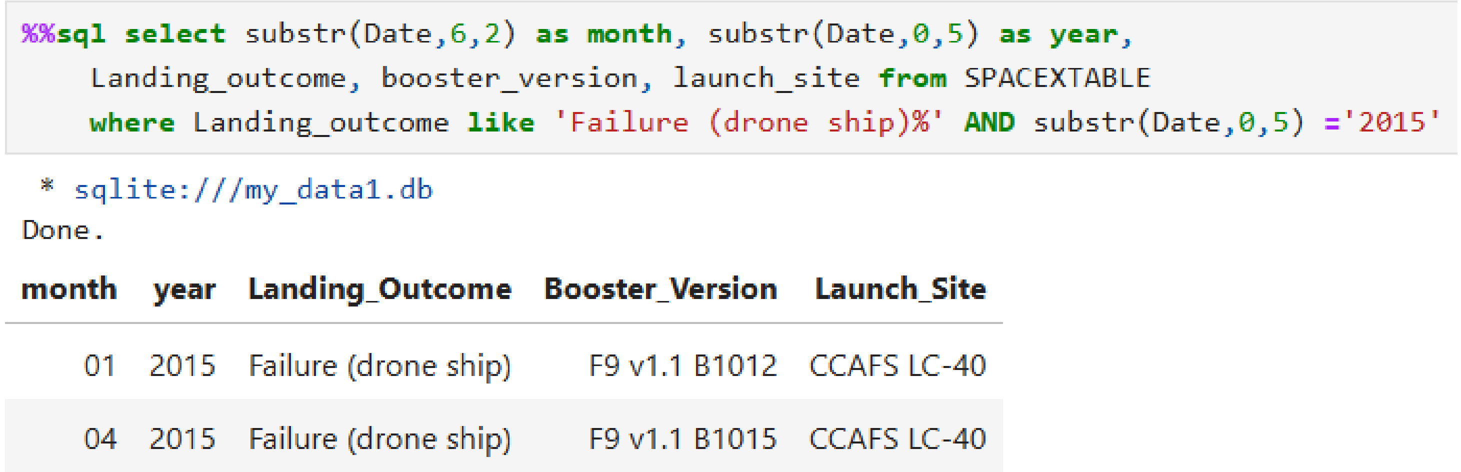
```
%%sql select booster_version from SPACEXTABLE
    where payload_mass__kg_=(select max(payload_mass__kg_)
    from SPACEXTABLE)
```

```
 * 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

2015 Launch Records

- Listing the failed landing_ outcomes in drone ship, their booster versions, and launch site names for in year 2015



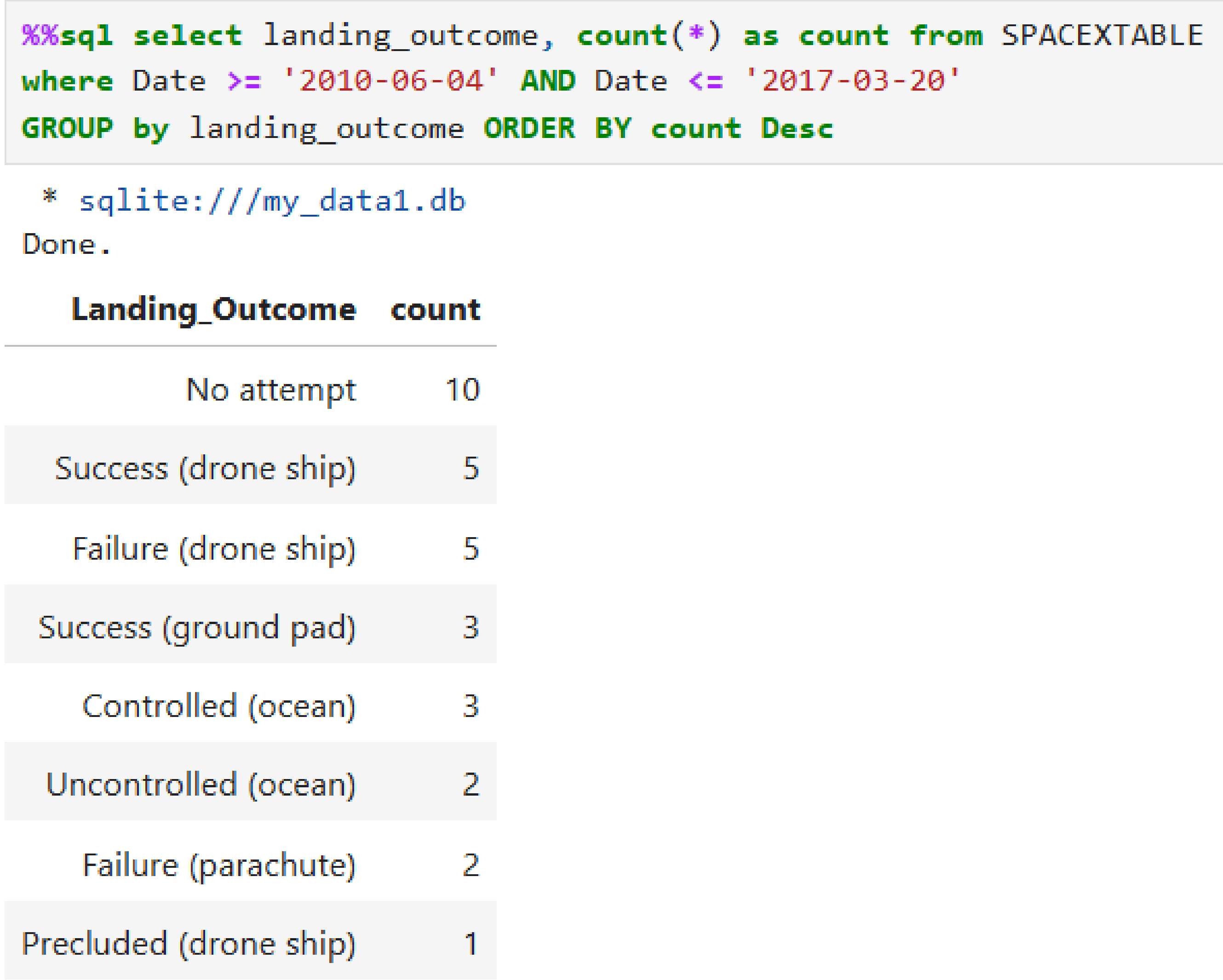
```
%%sql select substr(Date,6,2) as month, substr(Date,0,5) as year,
    Landing_outcome, booster_version, launch_site from SPACEXTABLE
    where Landing_outcome like 'Failure (drone ship)%' AND substr(Date,0,5) ='2015'
```

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

month	year	Landing_Outcome	Booster_Version	Launch_Site
01	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order



```
%%sql select landing_outcome, count(*) as count from SPACEXTABLE
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing_outcome ORDER BY count Desc
```

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

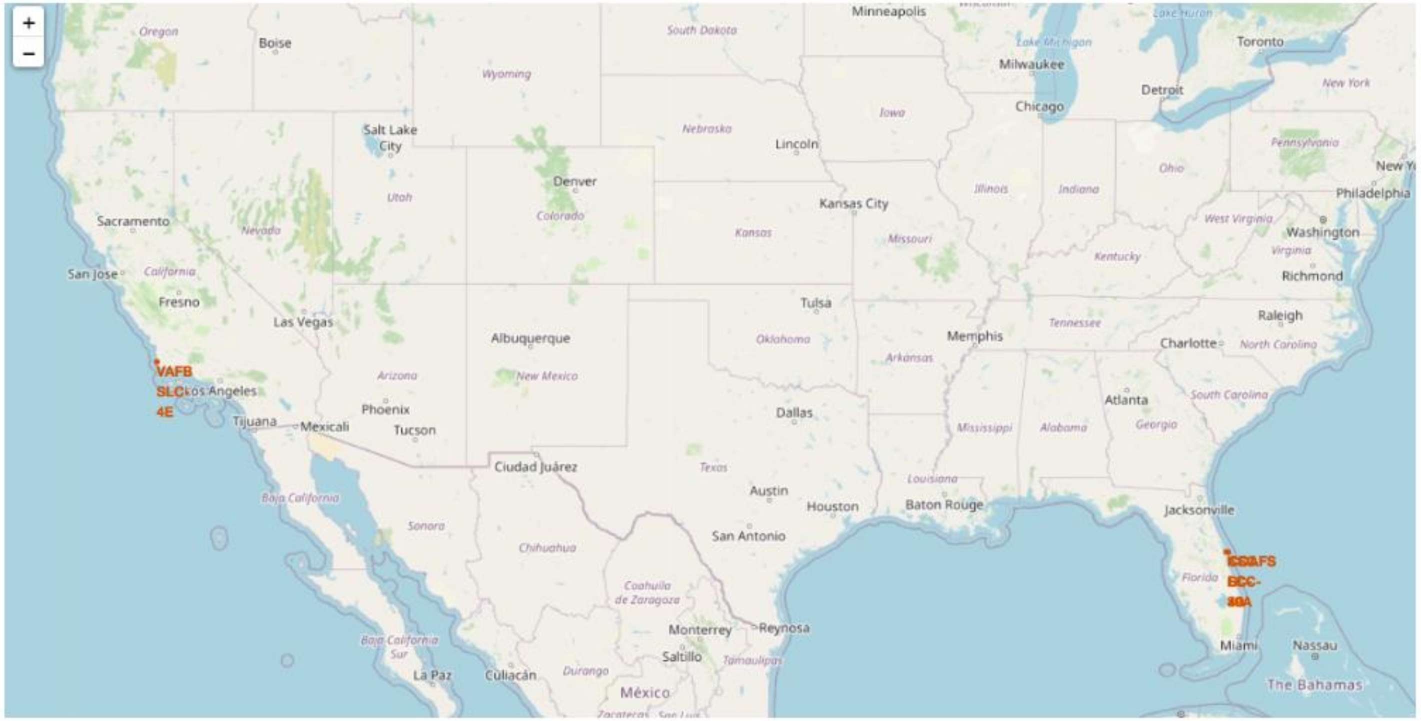
Landing_Outcome	count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

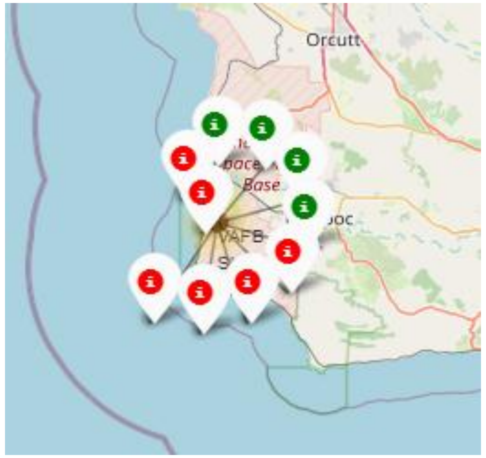
All launch sites global map markers

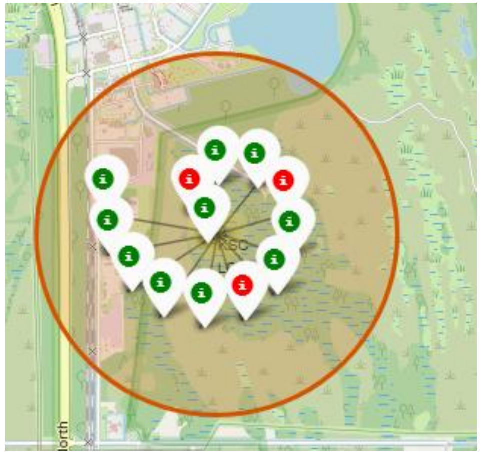
- We see that Space X launch sites are located on the United States, All launch sites are in very close proximity to the coast

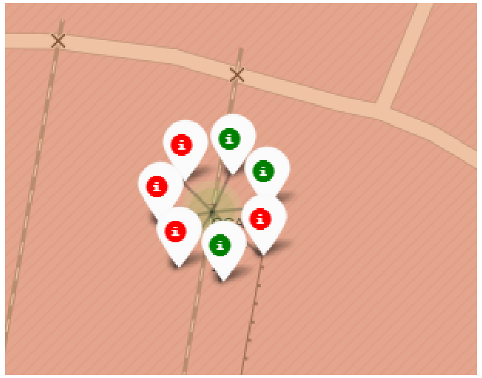


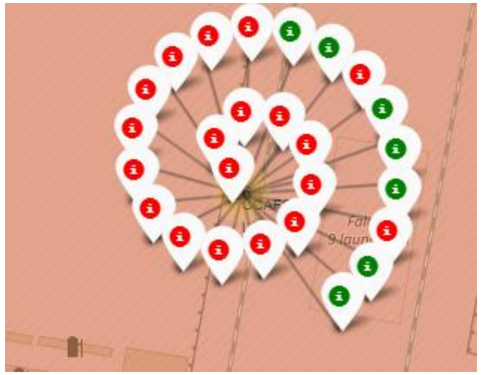
Colour-labeled launch records on the map

- Green marker represents successful launches.
- Red marker represents unsuccessful launches.
- Note that KSC LC-39A (Top right map) has a higher launch success rate.



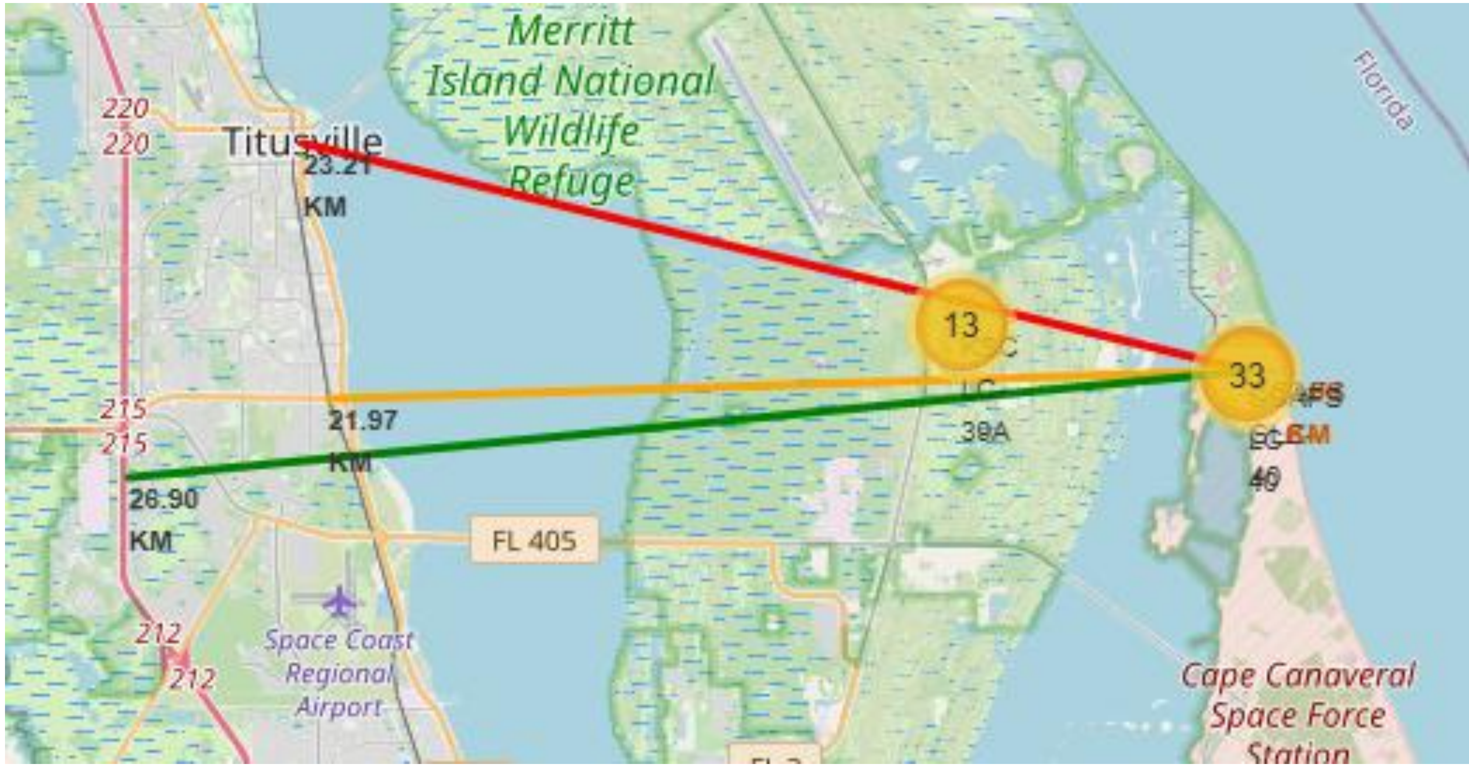


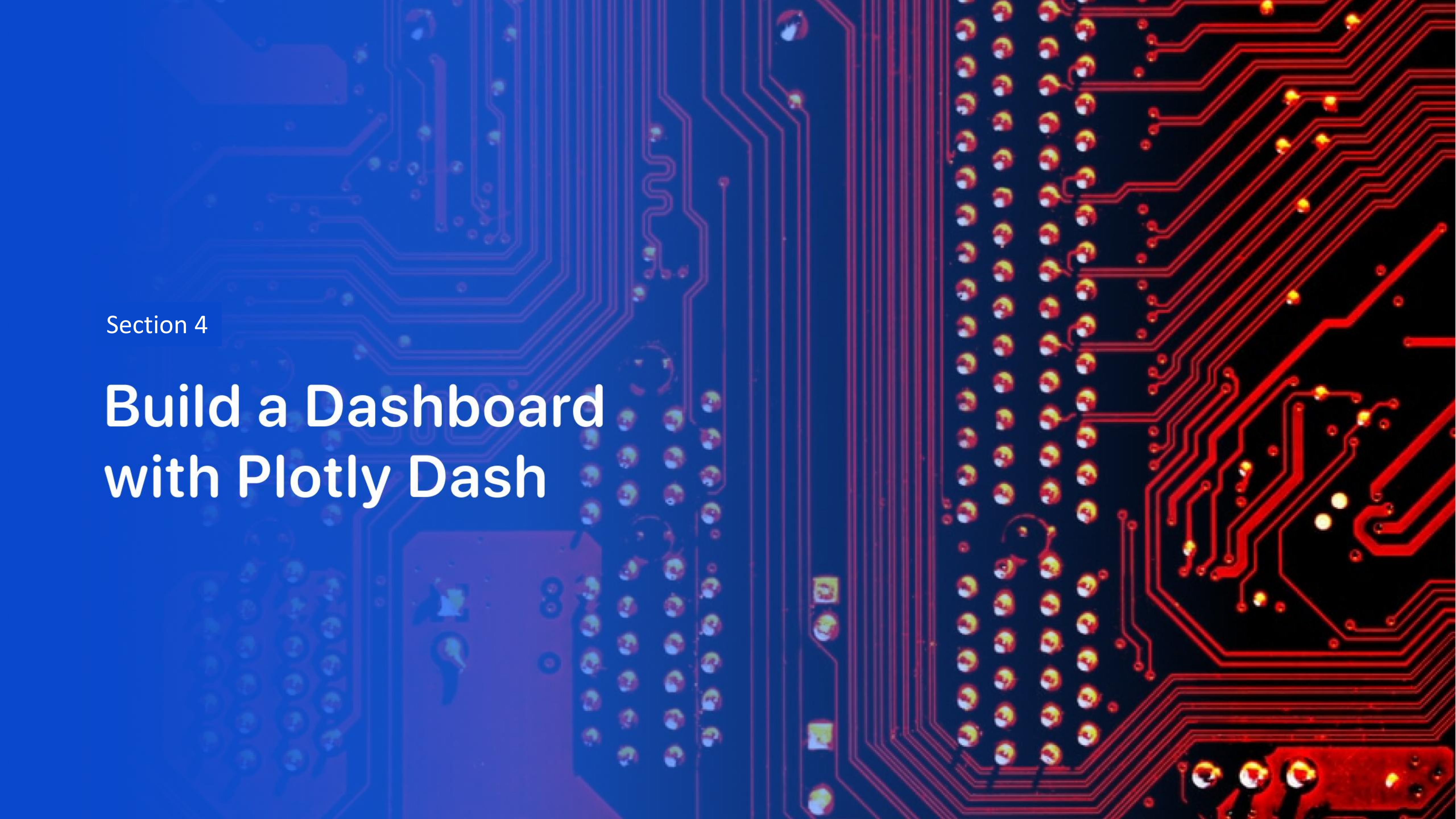




Distance from the launch site CCAFS SLC-40 to its proximities

- Relatively close to city (23.21 km)
- Relatively close to railway (21.97 km)
- Relatively close to highway (26.90 km)

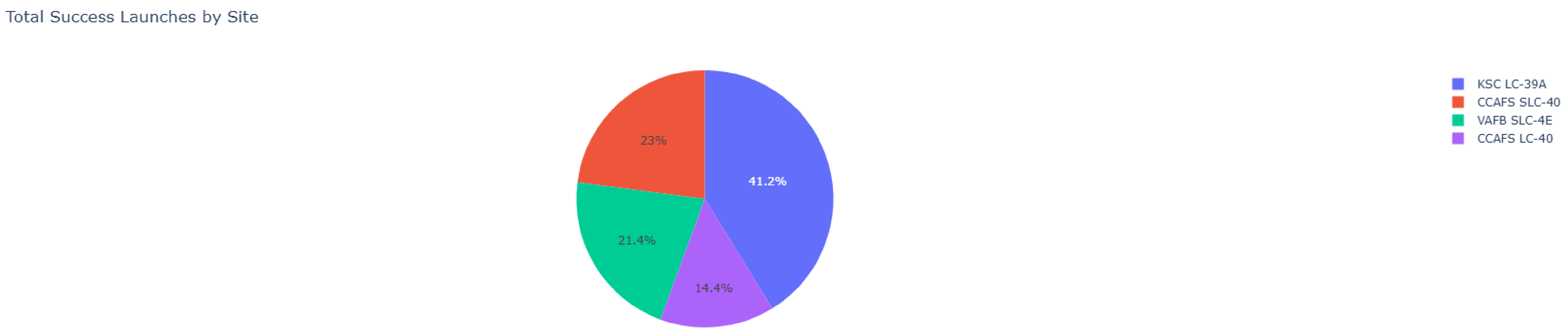

Merritt Island National Wildlife Refuge
Titusville
23.21 KM
21.97 KM
26.90 KM
13
33
39A
LC-40
FL 405
Space Coast Regional Airport
Cape Canaveral Space Force Station
Florida
220
215
212



Section 4

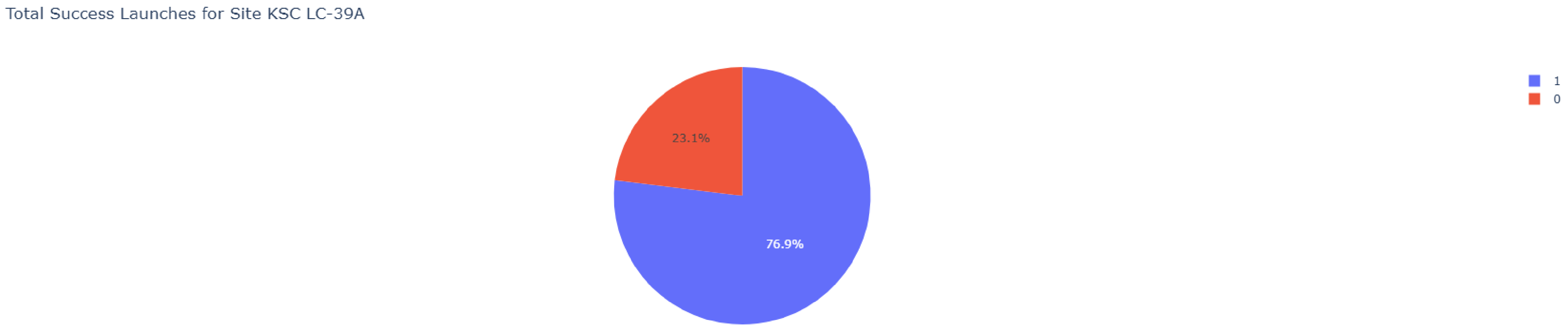
Build a Dashboard with Plotly Dash

Launch success count for all sites



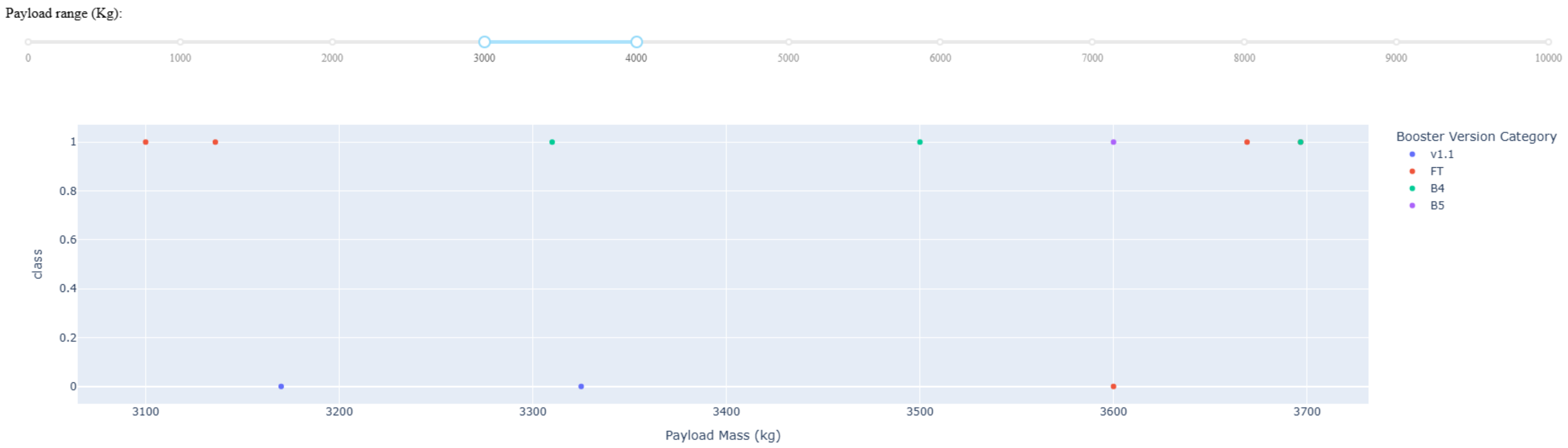
- We see Total launch success for all sites , KSC LC-39A has the most successful launches.

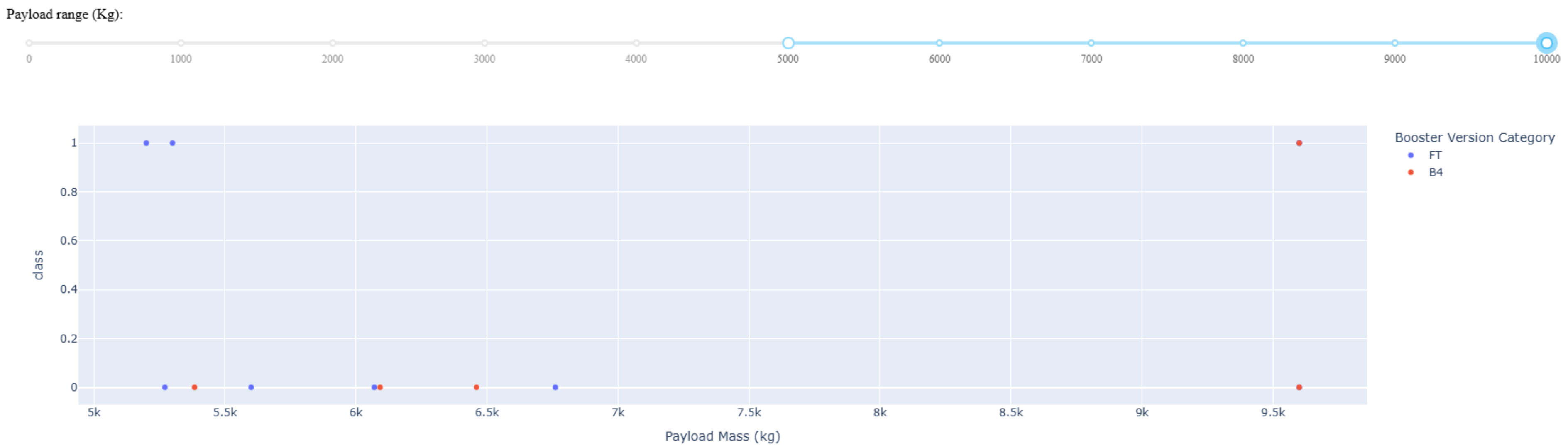
Launch site with highest launch success ratio


Total Success Launches for Site KSC LC-39A

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

Payload Mass vs. Launch Outcome scatter plot for all sites





- The greatest success rate occurs in launches with payloads between 3000-4000 kg.

- In overall, Low payloads mass have a better success rate than the heavy payloads mass.

- There are only FT and B4 among the Booster version categories in the payload mass range greater than 5000 kg.

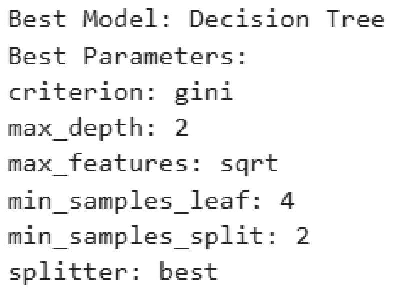
Section 5

Predictive Analysis (Classification)

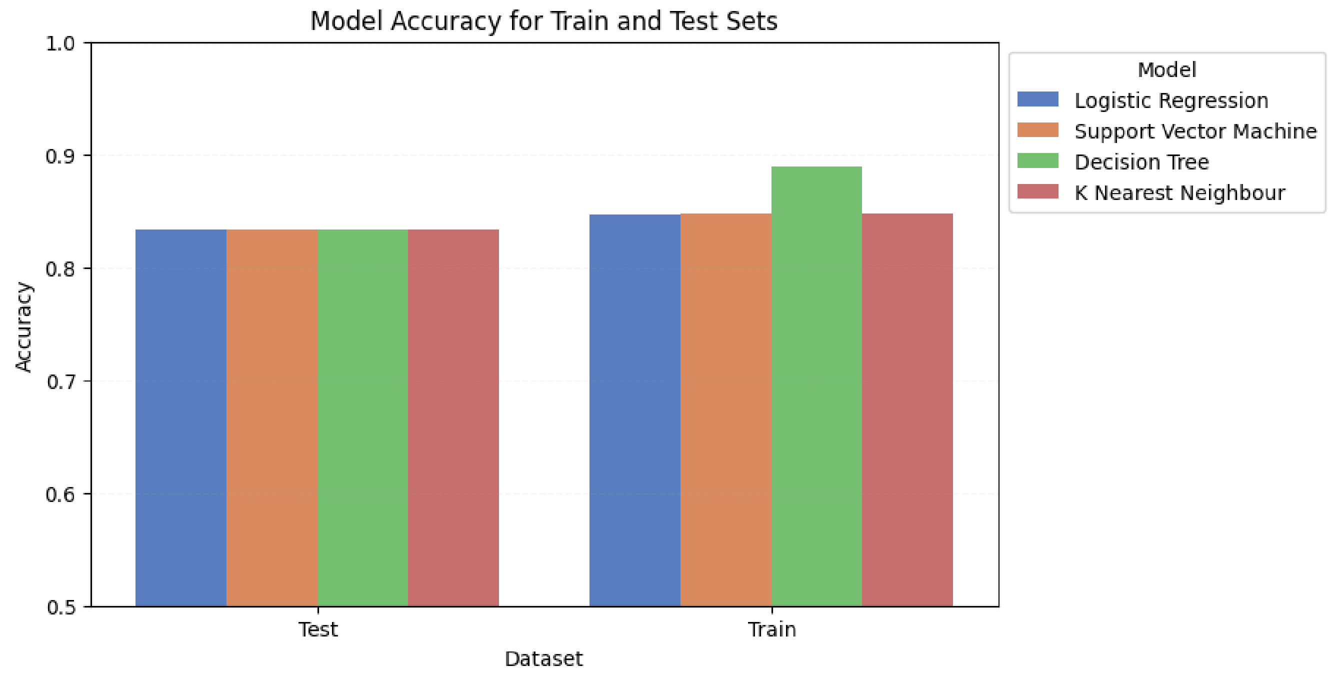
Classification Accuracy

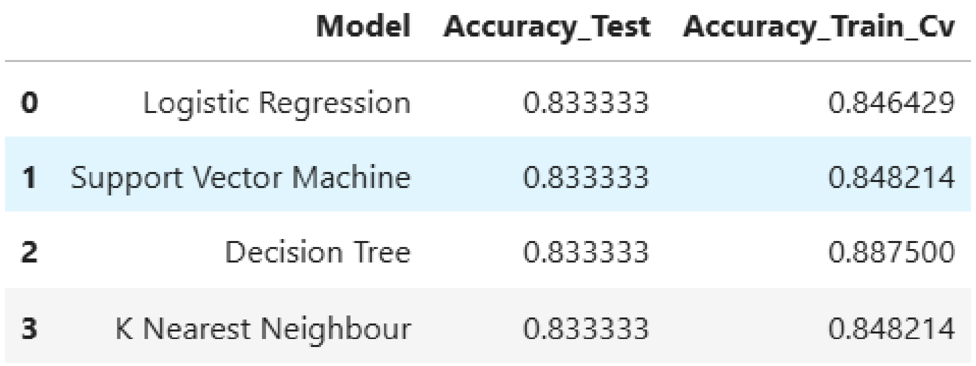
- We see that all models have the same accuracy in the test dataset; however, the Decision Tree model has the best accuracy in the train dataset.
- The same test set accuracy may be due to the small test sample size

- The best parameters of best model are:



```
Best Model: Decision Tree
Best Parameters:
criterion: gini
max_depth: 2
max_features: sqrt
min_samples_leaf: 4
min_samples_split: 2
splitter: best
```

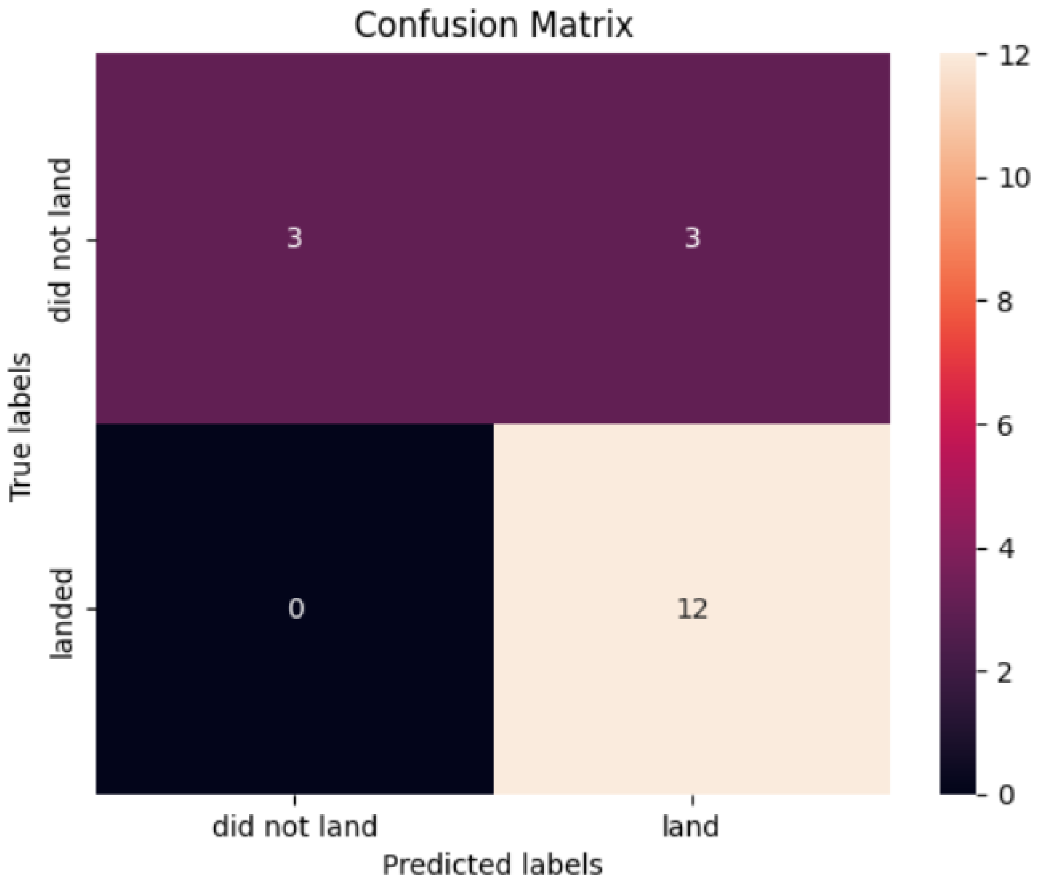




	Model	Accuracy_Test	Accuracy_Train_Cv
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.833333	0.887500
3	K Nearest Neighbour	0.833333	0.848214

Confusion Matrix

- As the test accuracy of all models are equal, the confusion matrices are also identical .The main problem of these models are false positives.



Conclusions

- The success rate of launches increases over the years.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Depending on the orbits, the payload mass can be impact to the success of a mission. Some orbits require a light or heavy payload mass.
- The greatest success rate occurs in launches with payloads between 3000-4000 kg.
- In overall, Low payloads mass have a better success rate than the heavy payloads mass.
- Finally ,we used some machine learning algorithms to learn the pattern to predict whether a mission will be successful or not based on the given features. Decision Tree Model is the best algorithm due to has the best train set accuracy. Although the test set accuracy between all the models used is identical.

Thank you!