Data Science Capstone project

Faruk Ahmad

2021/08/19

Outline

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

Executive Summary

- Summary of methodologies

  The SpaceX data has been preprocessed using some data wrangling methods to understand the details of the dataset, then the data has been normalized. Feature engineering has been done for converting categorical data to one hot encoding. Then different machine learning algorithms has been trained & tested on the dataset.

- Summary of all results

  From the machine learning predictive modeling it is evident that almost **89%** of the launches can be predicted if the first stage will be a success or failure.

Introduction

- Project background and context

  SpaceX is a space mission conducting company that developed Falcon 9 & demands that it can save up to 2/3 of the space mission cost since it can reuse the first stage of Falcon 9. Having the historical data of Falcon 9 's last missions & outcomes if the first stage was landed successfully or not. We will be analyzing that data to justify if the advertisement done by SpaceX is supposed to be true & what is the possibility of reusing the first stage of Falcon 9.

- Problems that I wanted to find answers

  a. What are the parameters for a successful landing of the first stage of Falcon 9, e.g. launch site, payload mass or anything else?

  b. What is the possibility or success rate of Falcon 9 to be reusing the first stage using some predictive modeling?

Methodology

- Data collection methodology:
  - Describe how data were collected
- Perform data wrangling
  - Describe how data were processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Methodology

Data collection

- Data has been collected in two ways
  - Data collection API namely - https://api.spacexdata.com/v4
  - Web scraping wiki page using BeautifulSoup

Data collection – SpaceX API

We have used REST API to the bellow endpoint for collecting SpaceX data-

Endpoint: https://api.spacexdata.com/v4

Github URL of
data collection API notebook

Data Collection API Notebook

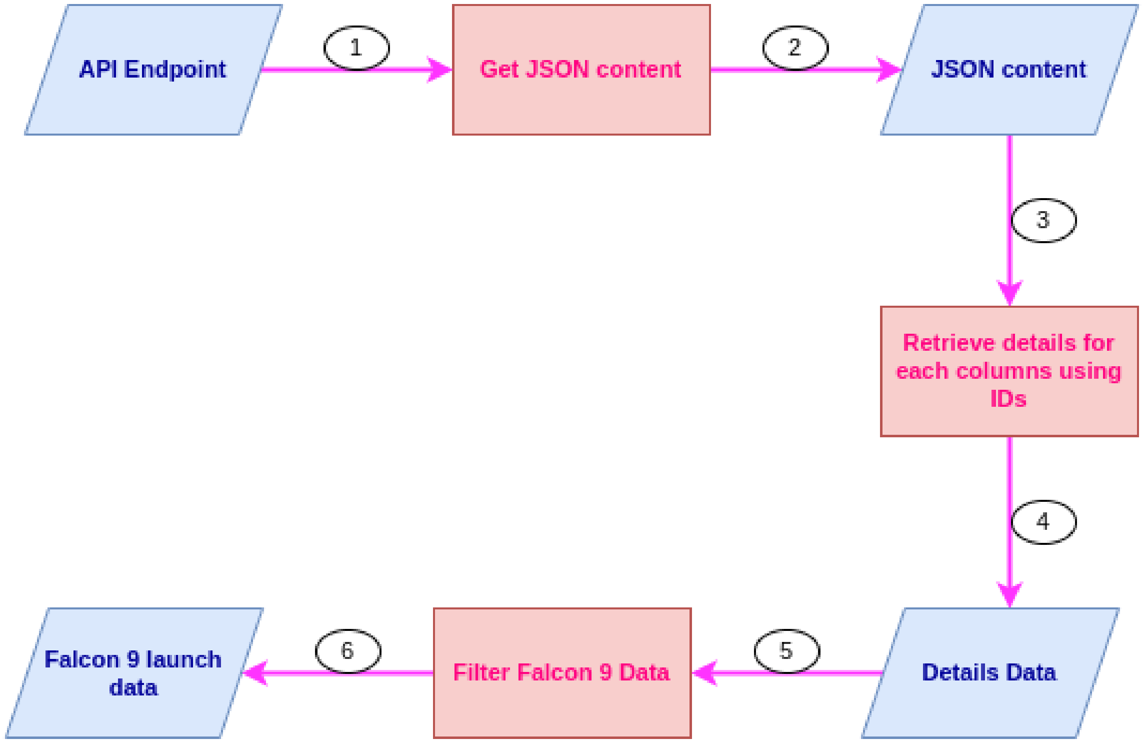


Fig: Data Collection API Flowchart

Data collection – Web scraping

We have used request module to get text content from wikipedia page. Then used BeautifulSoup for extracting table data from textual data for Falcon 9

Github URL of data collectiong web scraping notebook

Data Collection web scraping notebook

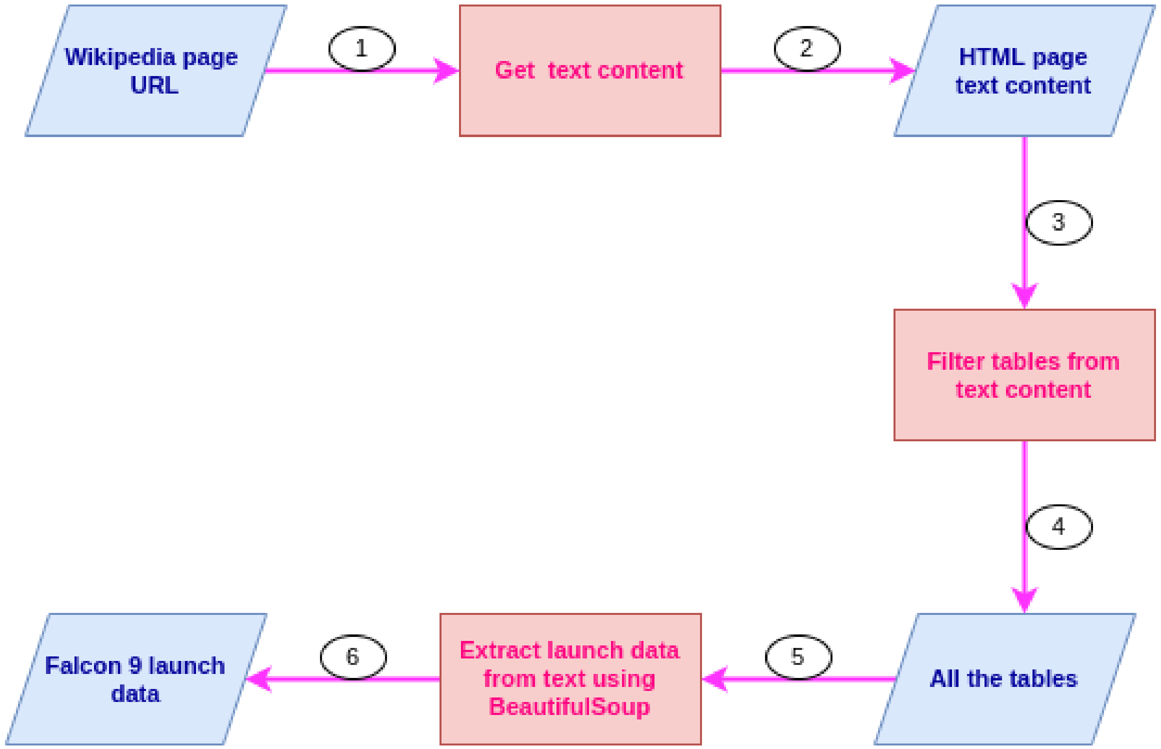
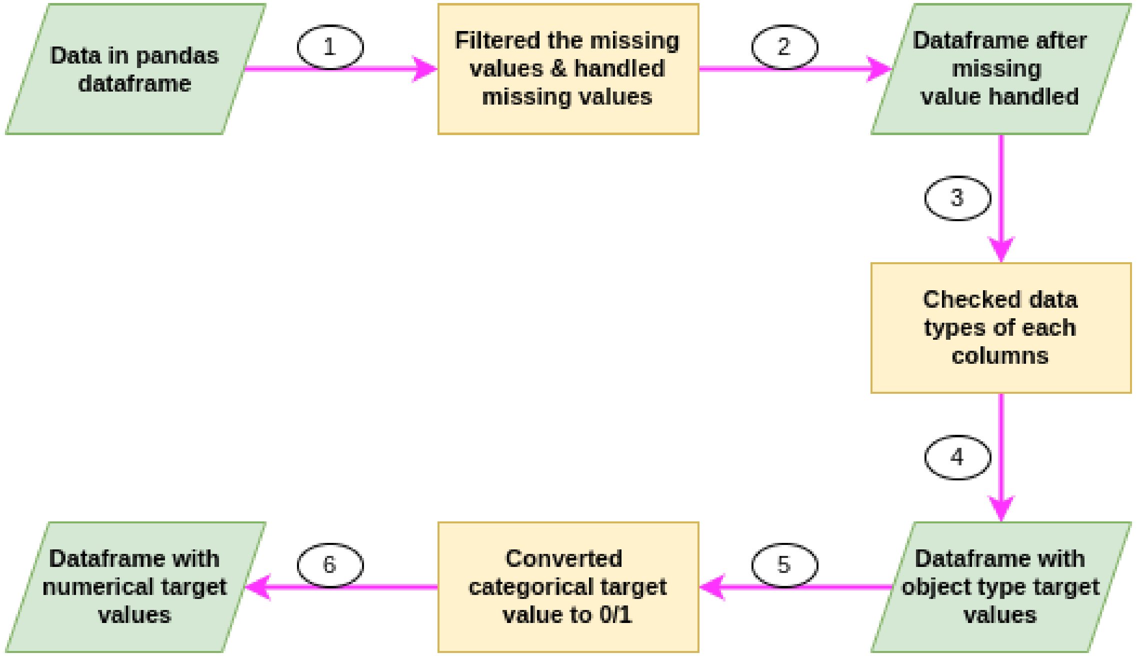


Fig: Data Collection Web Scraping Flowchart

Data wrangling

Data wrangling steps:

1. Filtered the missing values, handled the missing value if possible
2. Checked the data types of each column, cast datatypes if needed
3. Conveted the categorical object type target column to 0/1 neumerical values



Data in pandas dataframe
1
Filtered the missing values & handled missing values
2
Dataframe after missing value handled
3
Checked data types of each columns
4
Dataframe with object type target values
5
Converted categorical target value to 0/1
6
Dataframe with numerical target values

Fig: Data wrangling Flowchart

Github URL of data wrangling notebook

Data wrangling notebook

EDA with data visualization

SL No.	Chart Type	Purpose
01	Flight no. Vs payload mass scatter point chart	To visualize the relationship between flight no. & palyload mass
02	Flight no. Vs launch site scatter point chart	To visualize how many flight is launching from which launch site
03	Launch site Vs payload mass scatter plot	To visualize what is the ratio of different payload mass launched from different sites
04	Orbit Vs success rate bar chart	To visualize the success rate of different orbits
05	Orbit Vs flight no. Scatter plot	To visualize the relation between no. Of flights in each orbit
06	Orbit Vs payload mass scatter plot	To visualize the relation between payload mass with each orbit type
07	Year Vs success rate line plot	To visualize the yearly success trend over time

Data visualization notebook link

EDA with SQL

Performed SQL queries

- SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXDATASET;
- SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
- SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer='NASA (CRS)';
- SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1%';
- SELECT MIN(DATE) FROM SPACEXDATASET WHERE landing__outcome='Success (ground pad)';
- SELECT booster_version FROM SPACEXDATASET WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 6000;
- SELECT mission_outcome, COUNT(*) **as** "Total Number" FROM SPACEXDATASET GROUP BY mission_outcome;
- SELECT booster_version AS "Booster Version with Maximum Payload Mass" FROM SPACEXDATASET WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);
- SELECT MONTHNAME(DATE), landing__outcome, booster_version, launch_site FROM SPACEXDATASET WHERE landing__outcome='Failure (drone ship)' AND DATE LIKE '**%2015%**';
- SELECT landing__outcome, COUNT(*) AS "Count" FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome LIKE '%Success%' GROUP BY landing__outcome;

SQL Query notebook link

Build an interactive map with Folium

SL No.	Map Objects	Purpose
01	Marker	To mark each launch site & success or failure in each launch
02	Circle	To mark the launch sites in a cluster
03	Line	To display the distance between launch site & other geo locations

Visualization map with Folium notebook link

Build a Dashboard with Plotly Dash

SL No.	Chart Type	Purpose
01	Pie Chart	To visualize the success count of each launch site or all launch site
02	Scatter plot	To visualize the launch success status with respect to payload mass

Interactive dashboard in plotly - script link

Predictive analysis (Classification)

List of predictive models

1) K Nearest neighbors algorithm
2) Decision tree classifier
3) Support vector machine algorithm
4) Logistic regression model

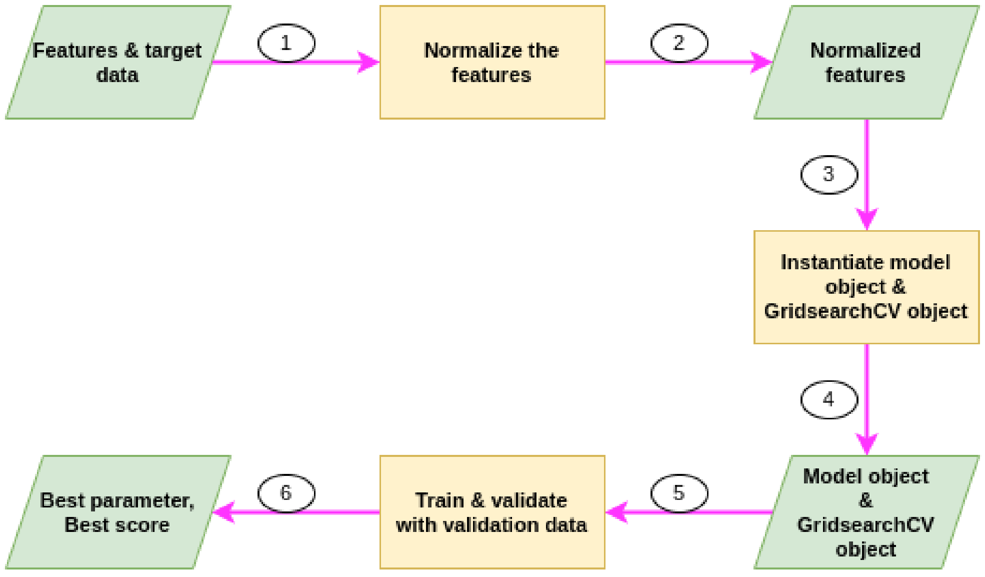

Features & target data
1
Normalize the features
2
Normalized features
3
Instantiate model object & GridsearchCV object
4
Model object & GridsearchCV object
5
Train & validate with validation data
6
Best parameter, Best score

Fig: Predictive Analysis Flowchart

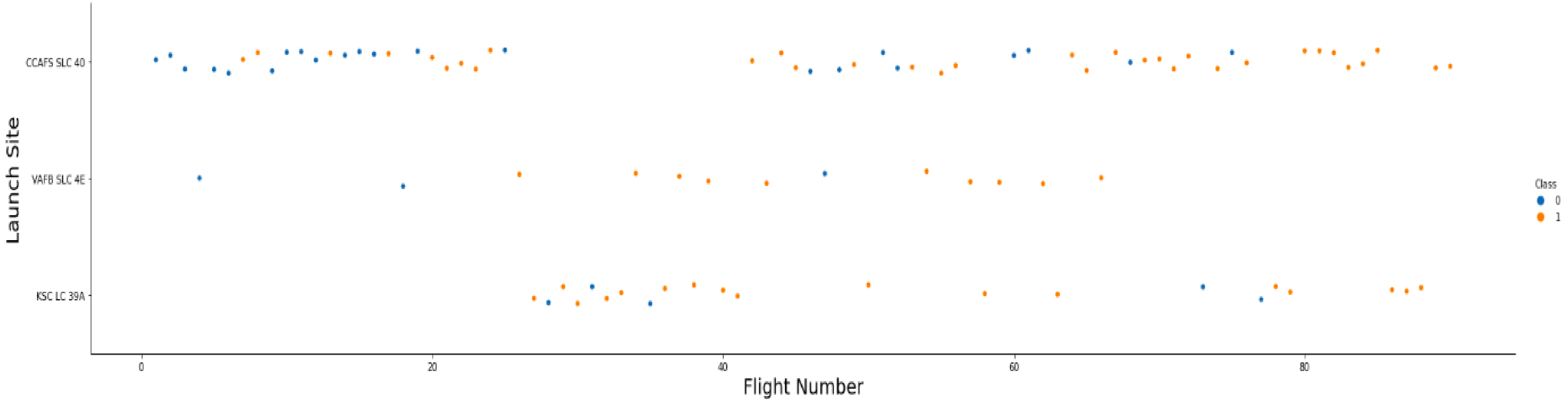
Predictive analysis notebook link

Results

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

EDA with Visualization

Flight Number vs. Launch Site

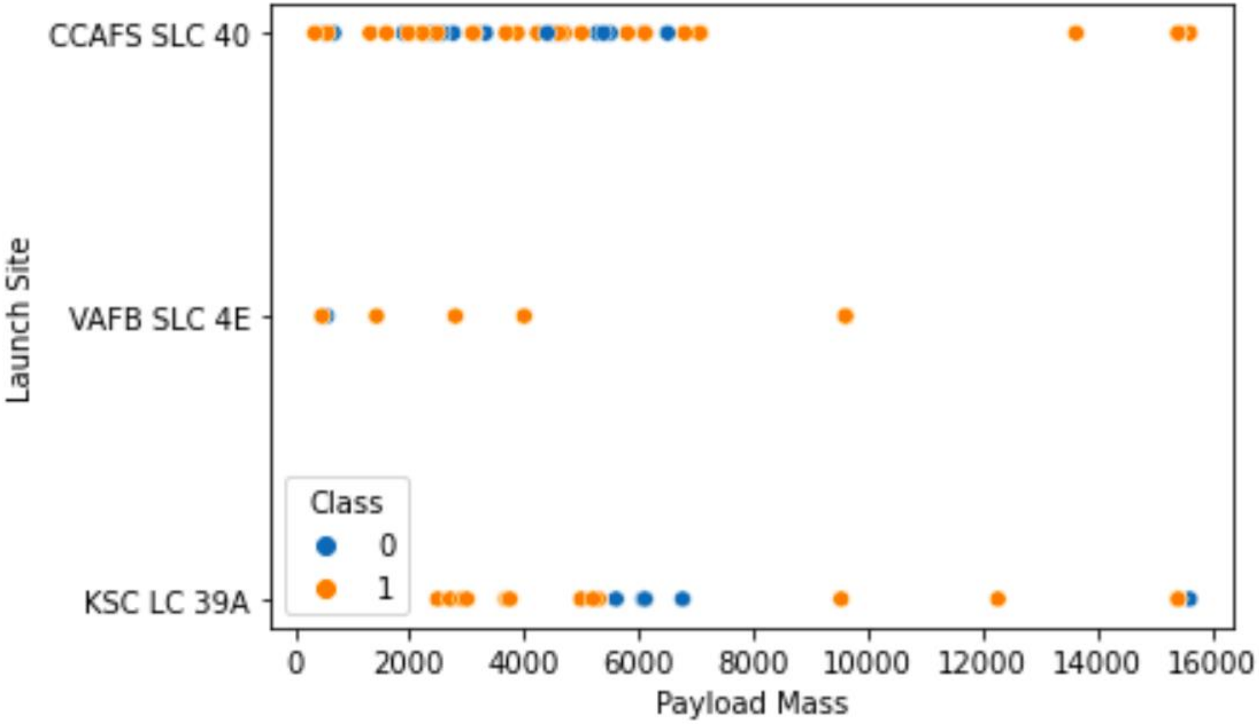


From the chart it is clear that, launch site CCAFS SLC 40 has the maximum number of launches, also most failure launches

Payload vs. Launch Site

Observations

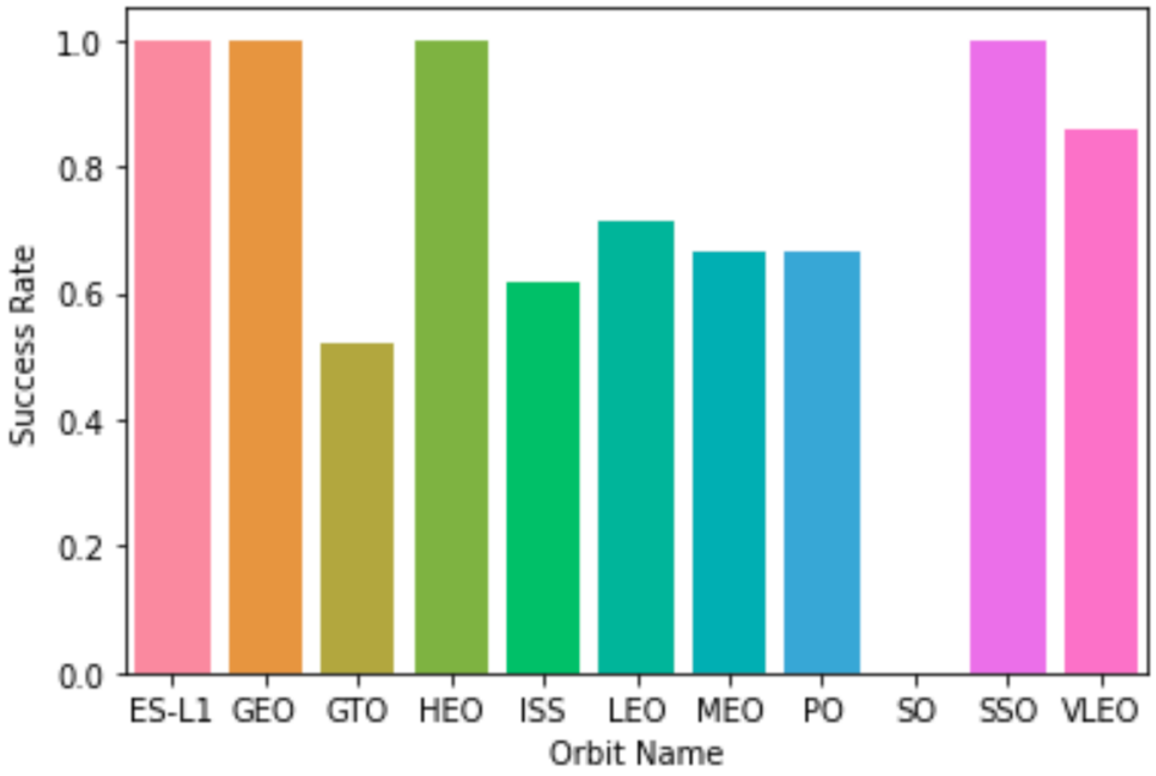
1) Launch site CCAFS SLC 40 has more success rate when the payload mass is larger.
2) Launch site VAFB SLC 4E seems to be has not impact on payload mass



Success rate vs. Orbit type

Observations

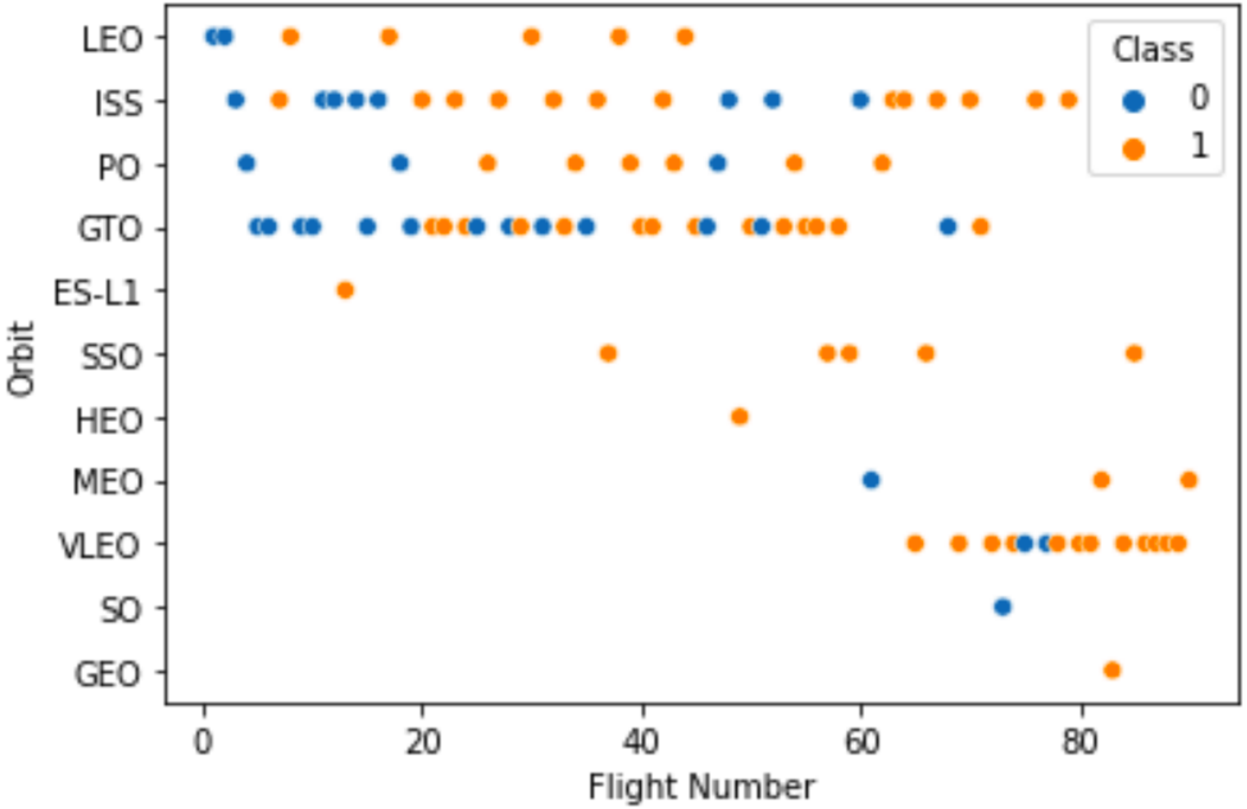
1) Orbit ES-L1, GEO, SSO & HEO seems to have more success rate than other orbits.



Flight Number vs. Orbit type

Observations

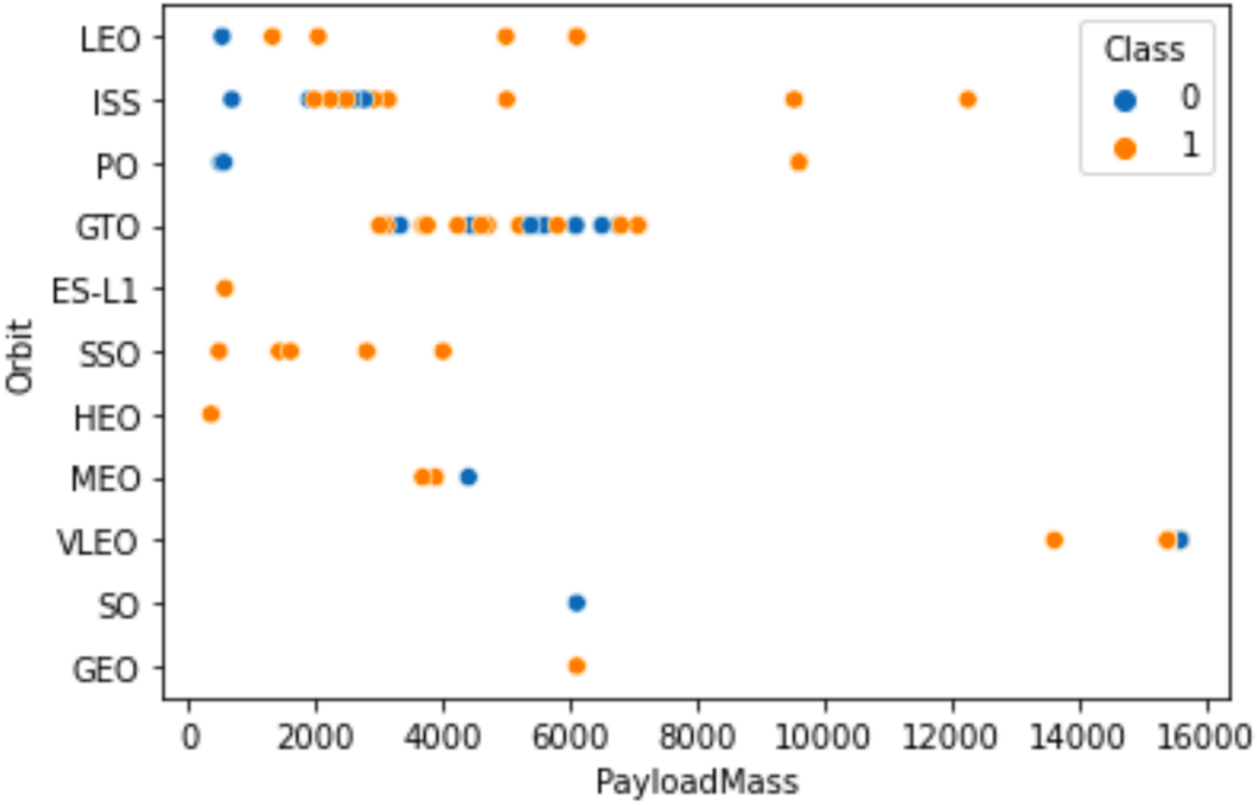
1) With increase of flight no. Each orbit seems to have more success rate in launches.



Payload vs. Orbit type

Observations

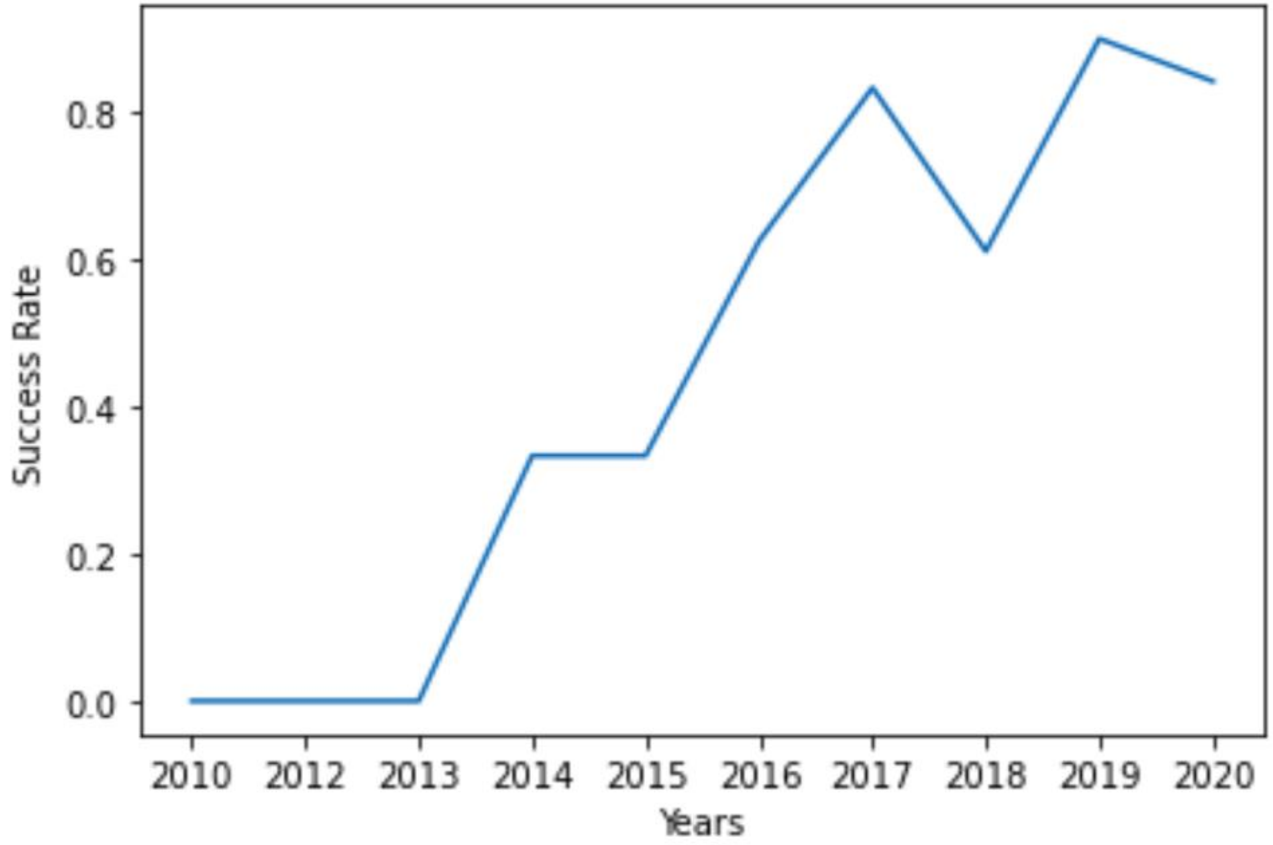
1) LEO orbit seems to have more success rate with more payload mass.



Launch success yearly trend

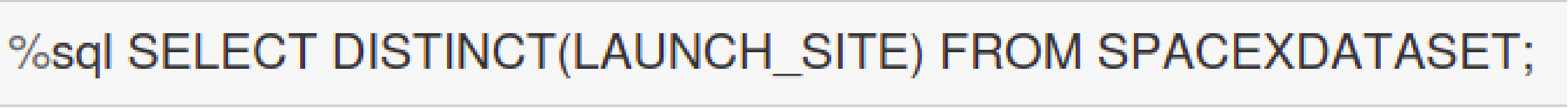
Observations

1) From 2013 until 2020 success of launch & reuse of first stage is increasing consistently, though there is a slight fall in 2018.



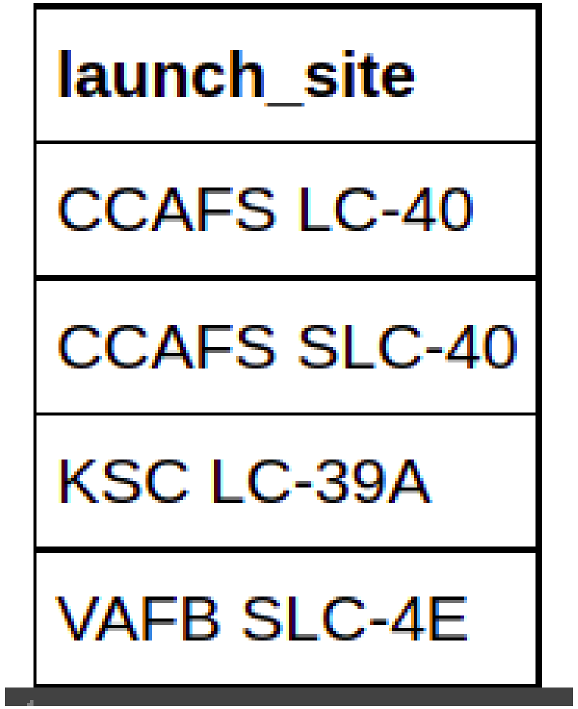
EDA with SQL

All launch site names



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

- DISTINCT clause has been used for finding unique launch sites from the SPACEXDATASET table.

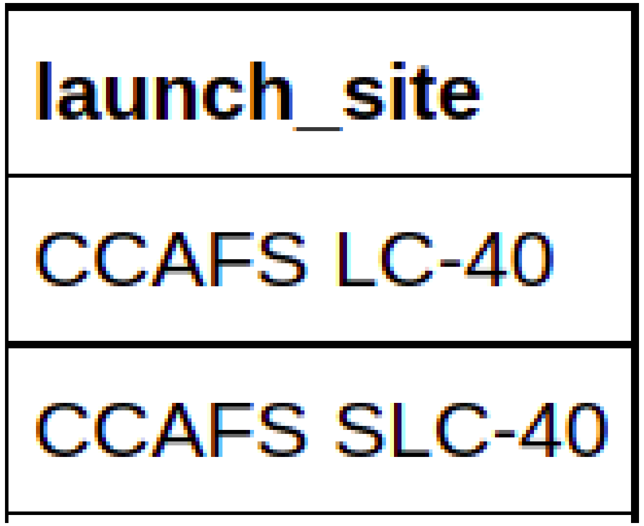


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch site names begin with `CCA`

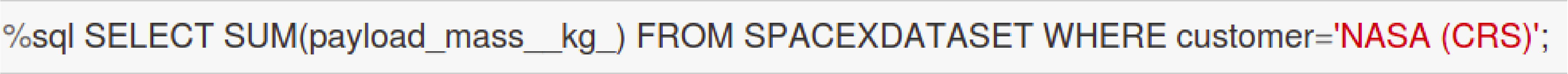
```
%sql SELECT launch_site FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%';
```

Used WHERE clause to filter the launch site names, & the LIKE clause to input the pattern



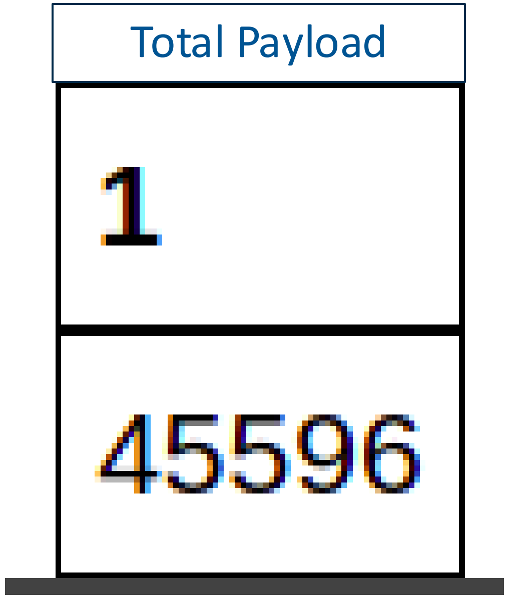
launch_site
CCAFS LC-40
CCAFS SLC-40

Total payload mass



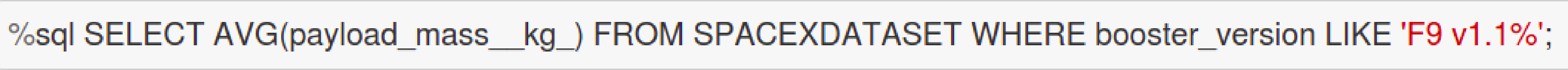
```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer='NASA (CRS)';
```

Used the SUM built in function to calculate the total payload mass & also used the WHERE clause to filter by the customer.



Total Payload
1
45596

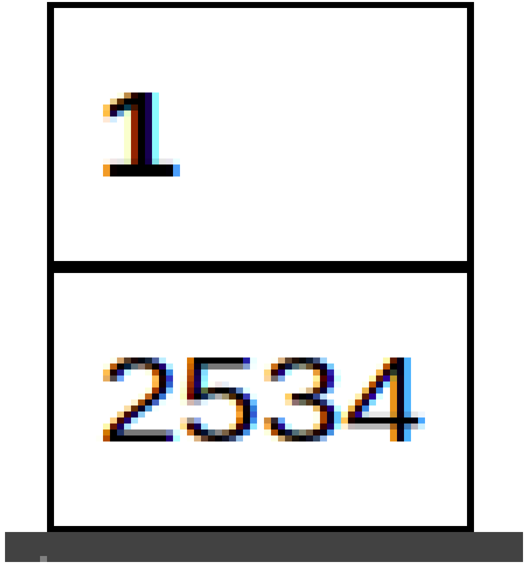
Average payload mass by F9 v1.1



```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1%';
```

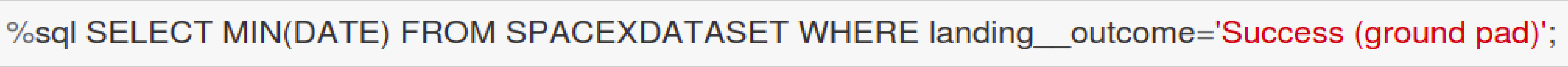
Used the AVG built in function for calculating the average payload mass & also used the WHERE clause for filtering the records by booster_version F9 v1.1

Avg. Payload F9



1
2534

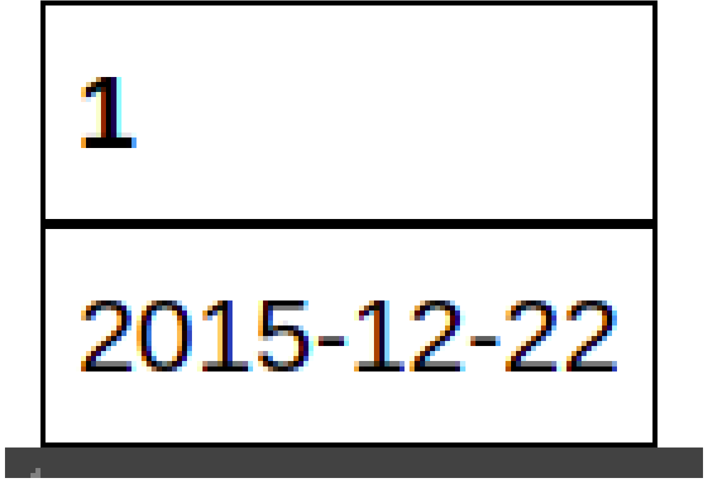
First successful ground landing date



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

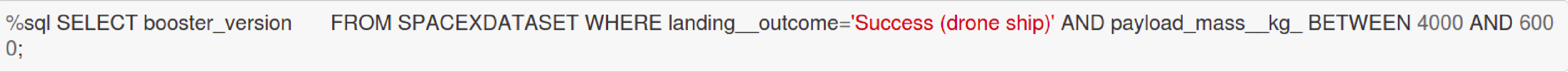
Used the MIN builtin function for calculating the first date & also used the WHERE clause to filter the record with landing_outcome.

First successful landing date



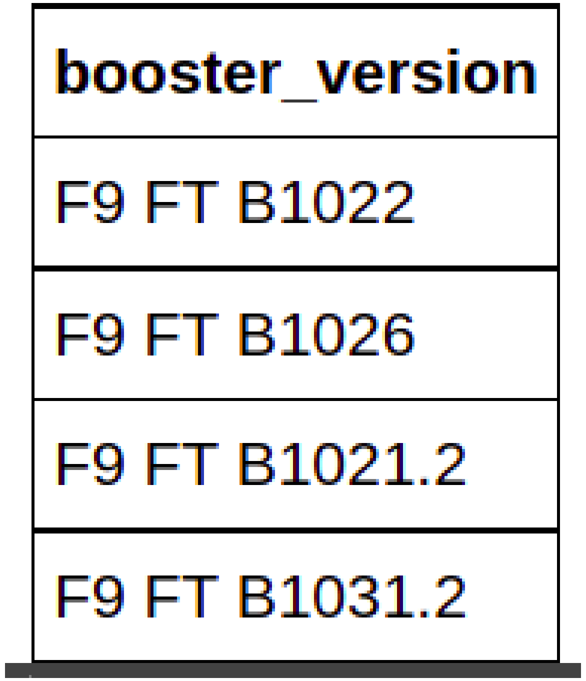
1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000



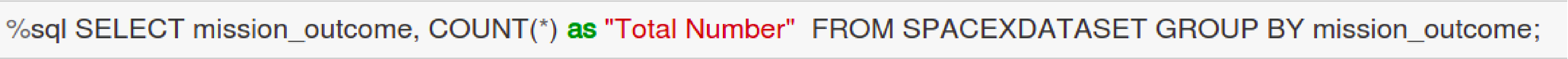
```
%sql SELECT booster_version      FROM SPACEXDATASET WHERE landing__outcome='Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND 600
0;
```

Used SELECT query to select the booster versions & also used the WHERE clause for filtering with landing_outcome & payload_mass__kg_



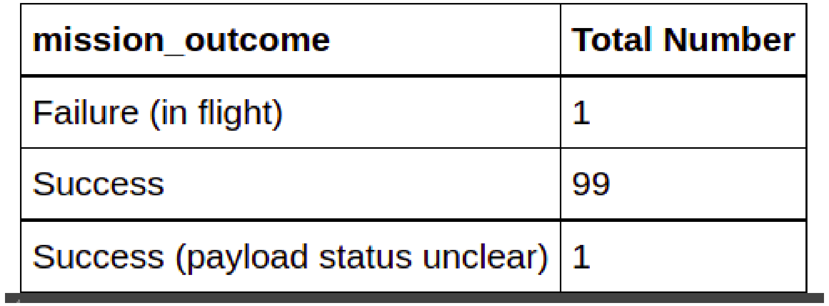
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total number of successful and failure mission outcomes



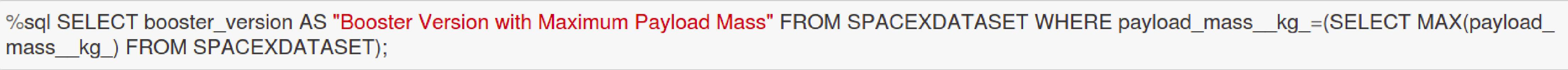
```
%sql SELECT mission_outcome, COUNT(*) as "Total Number"  FROM SPACEXDATASET GROUP BY mission_outcome;
```

Used COUNT built in function for getting the total number of occurences as success & failure.



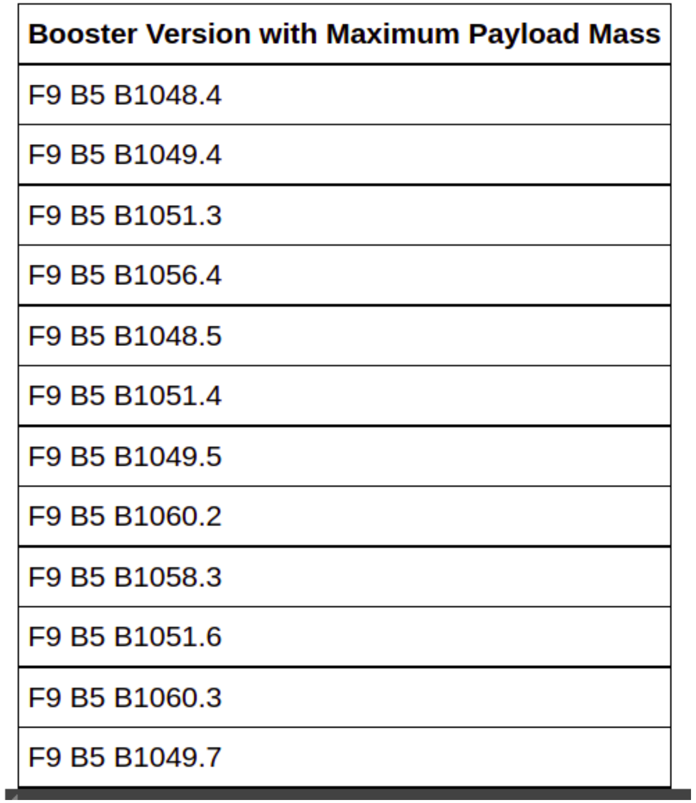
mission_outcome	Total Number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum payload



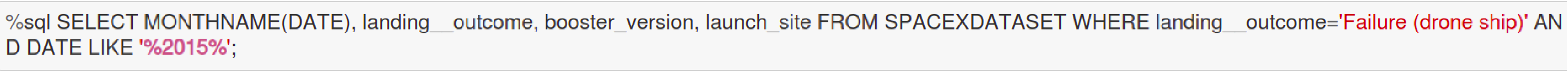
```
%sql SELECT booster_version AS "Booster Version with Maximum Payload Mass" FROM SPACEXDATASET WHERE payload_mass__kg_=(SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET);
```

Used sub query for finding the maximum payload, then compared that maximum amount for finding the booster version that satisfies the condition.



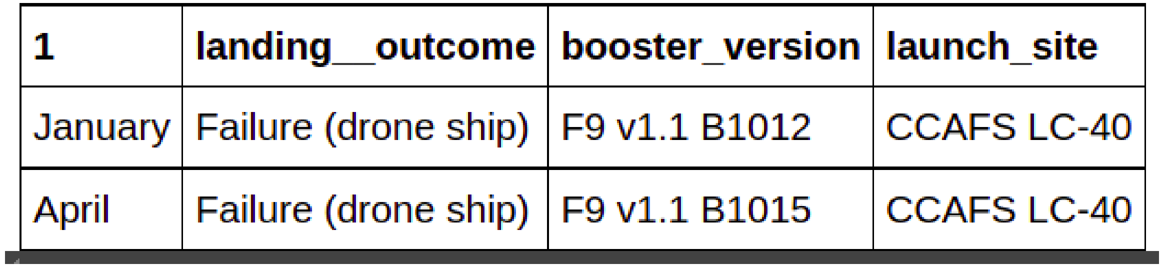
Booster Version with Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 launch records



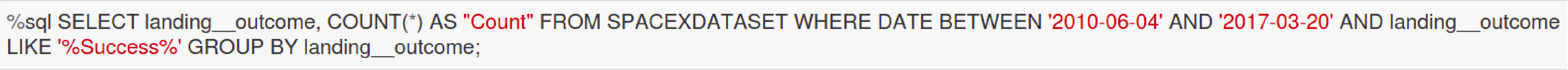
```
%sql SELECT MONTHNAME(DATE), landing__outcome, booster_version, launch_site FROM SPACEXDATASET WHERE landing__outcome='Failure (drone ship)' AN
D DATE LIKE '%2015%';
```

Used MONTHNAME function for getting the monthname from DATE, also used the LIKE clause for finding records relevant to 2015



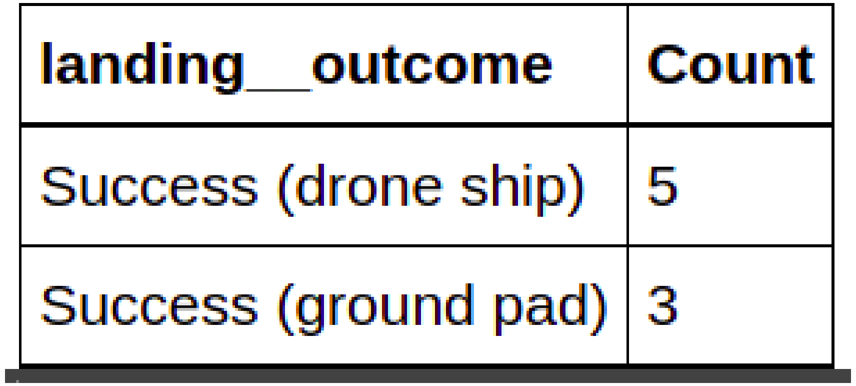
1	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20



```
%sql SELECT landing__outcome, COUNT(*) AS "Count" FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome
LIKE '%Success%' GROUP BY landing__outcome;
```

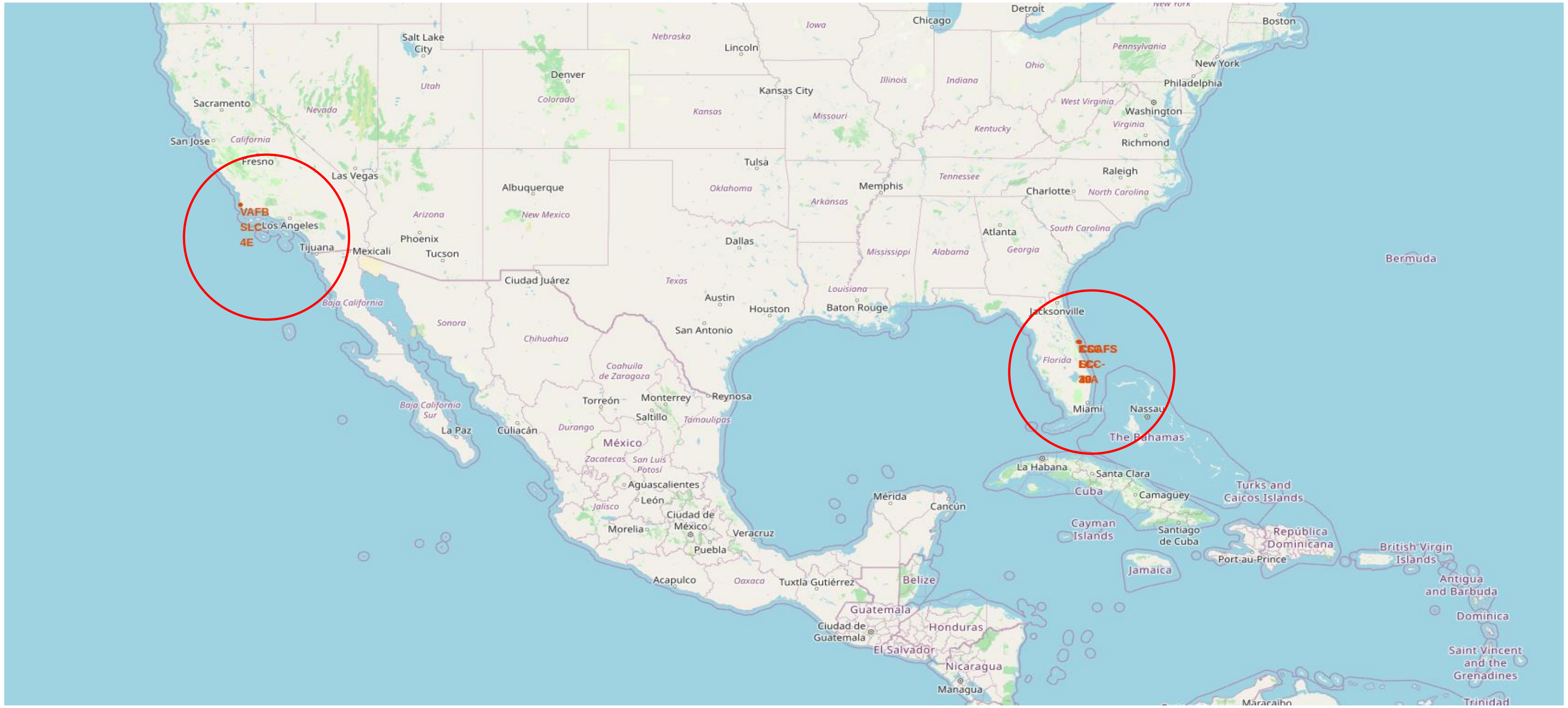
Used COUNT built in function for getting the total number of success launch both for dron ship & ground pad



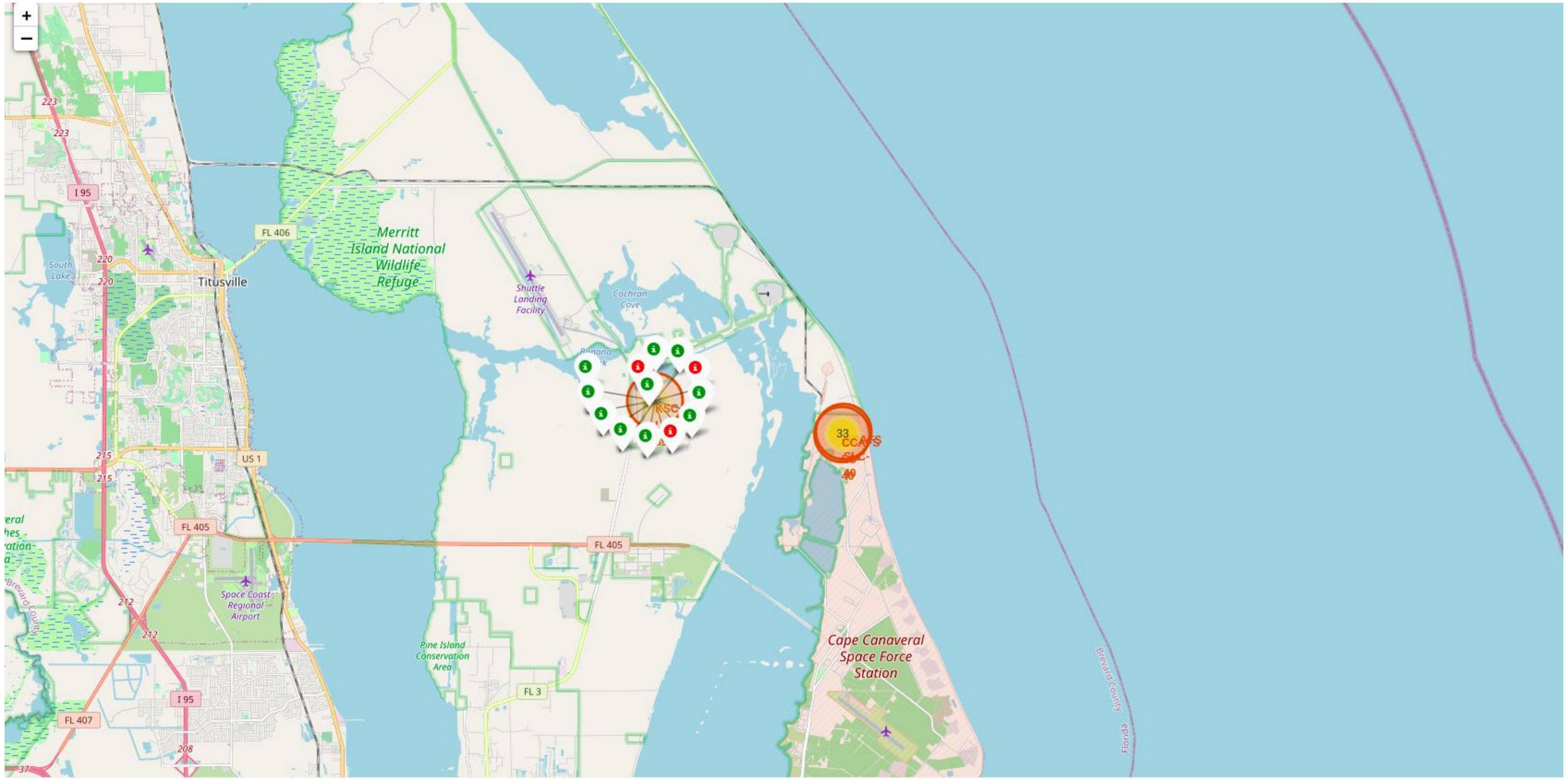
landing__outcome	Count
Success (drone ship)	5
Success (ground pad)	3

Interactive map with Folium

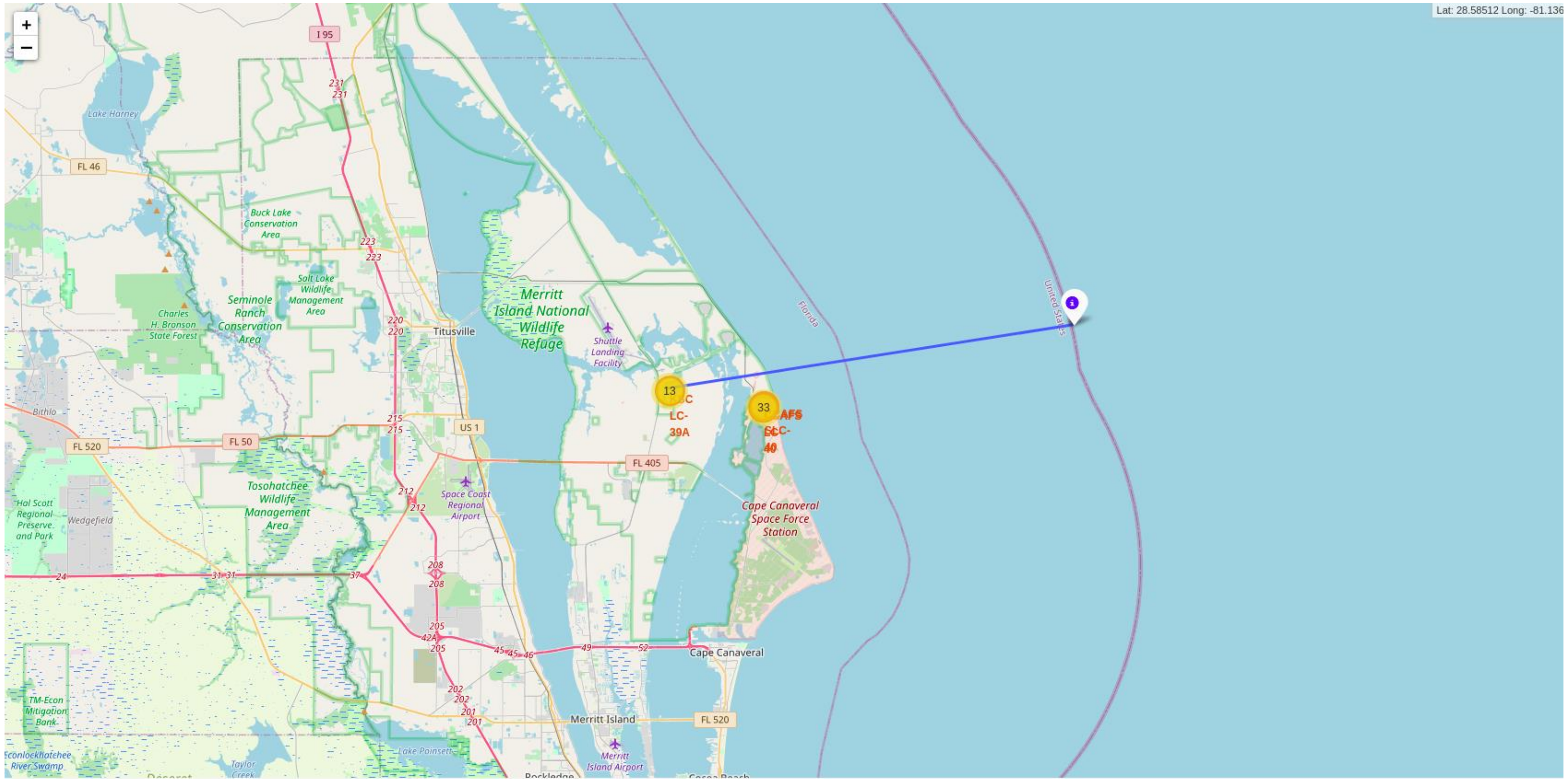
All launch sites marker in map



Color labeled launch records map

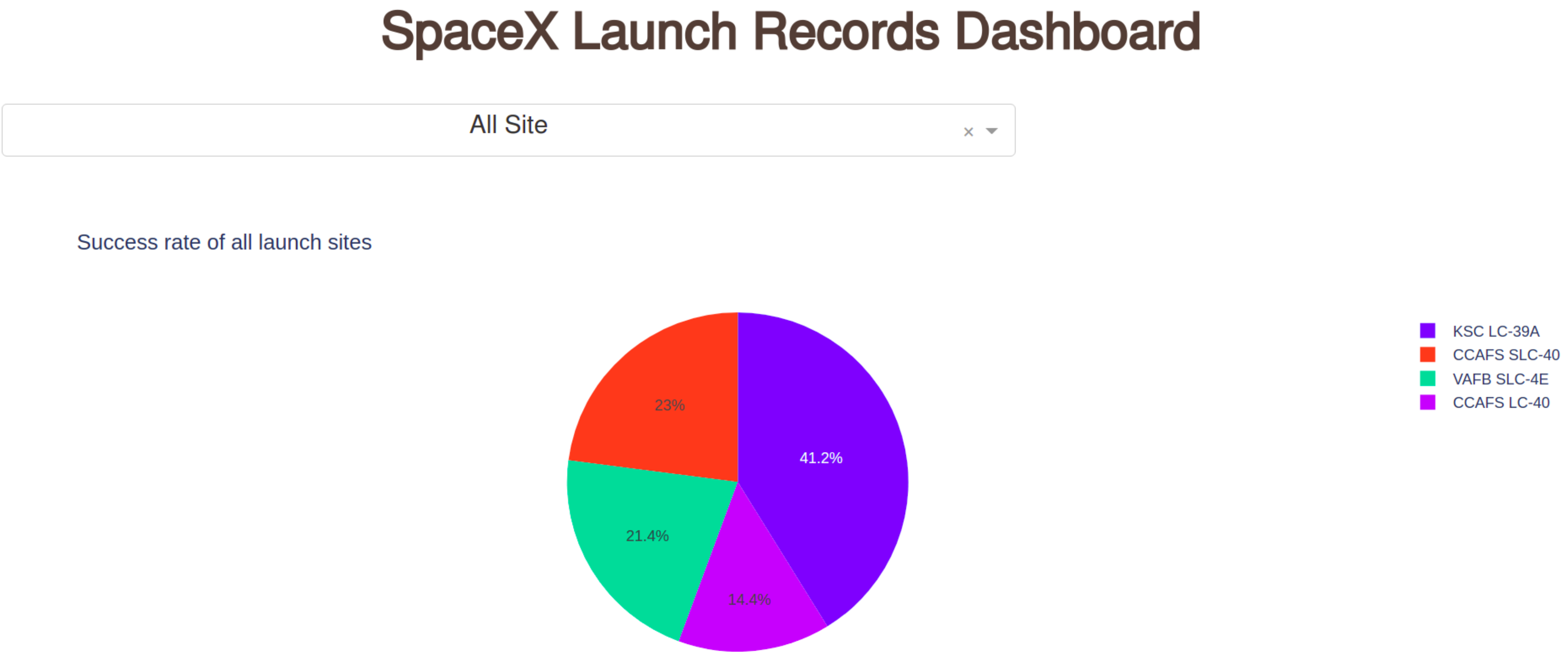


Proximity to other GEO location


Lat: 28.58512 Long: -81.136
+
−
13
LC-
39A
33
AFS
LC-
40
Florida
United States
Merritt Island National Wildlife Refuge
Shuttle Landing Facility
Cape Canaveral Space Force Station
Titusville
Cape Canaveral
Merritt Island
Merritt Island Airport
Space Coast Regional Airport
Tosohatchee Wildlife Management Area
Seminole Ranch Conservation Area
Buck Lake Conservation Area
Salt Lake Wildlife Management Area
Charles H. Bronson State Forest
Lake Harney
Bithlo
Wedgefield
Hal Scott Regional Preserve and Park
TM-Econ Mitigation Bank
Econlockhatchee River Swamp
Lake Poinsett
Rockledge
Cocoa Beach
I 95
FL 46
FL 50
FL 520
FL 405
US 1

Build a Dashboard with Plotly Dash

All sites launch success counts

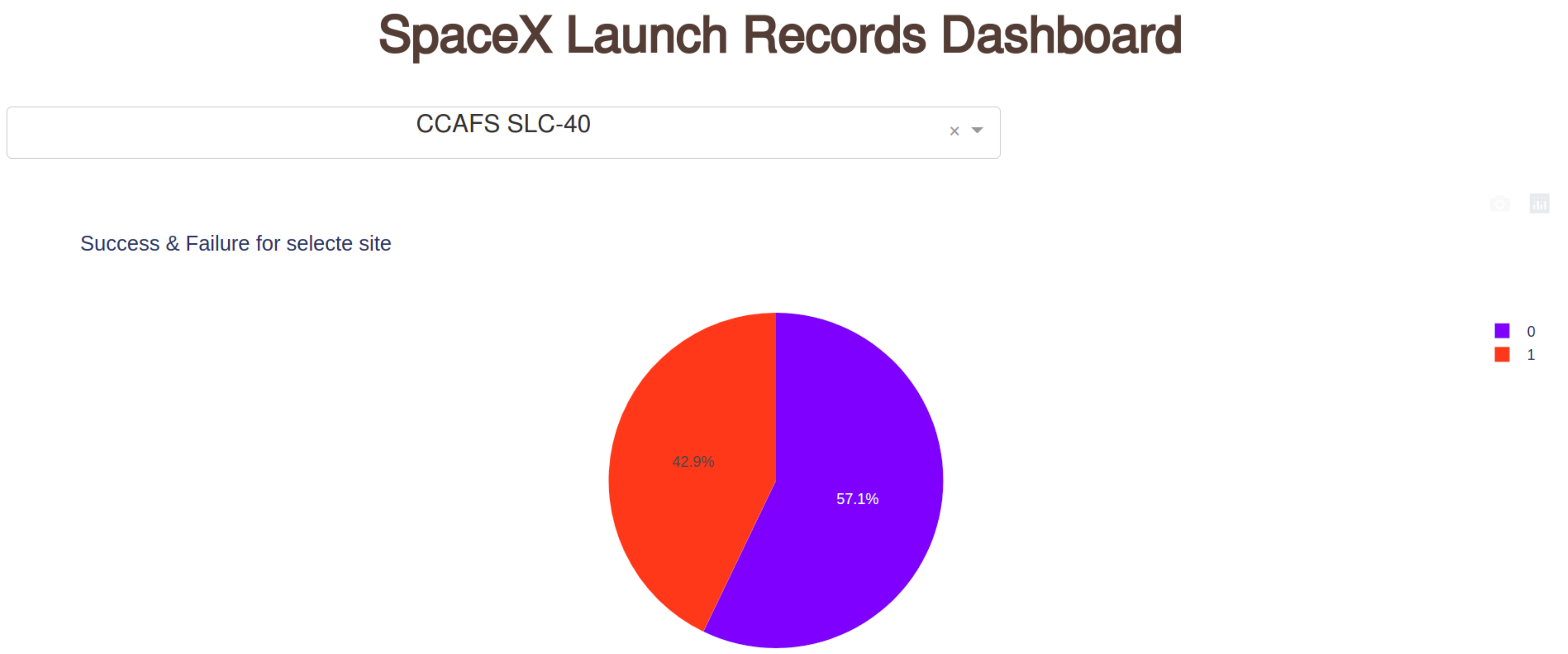


SpaceX Launch Records Dashboard

All Site × ▾

Success rate of all launch sites

Launch site with highest launch success ratio

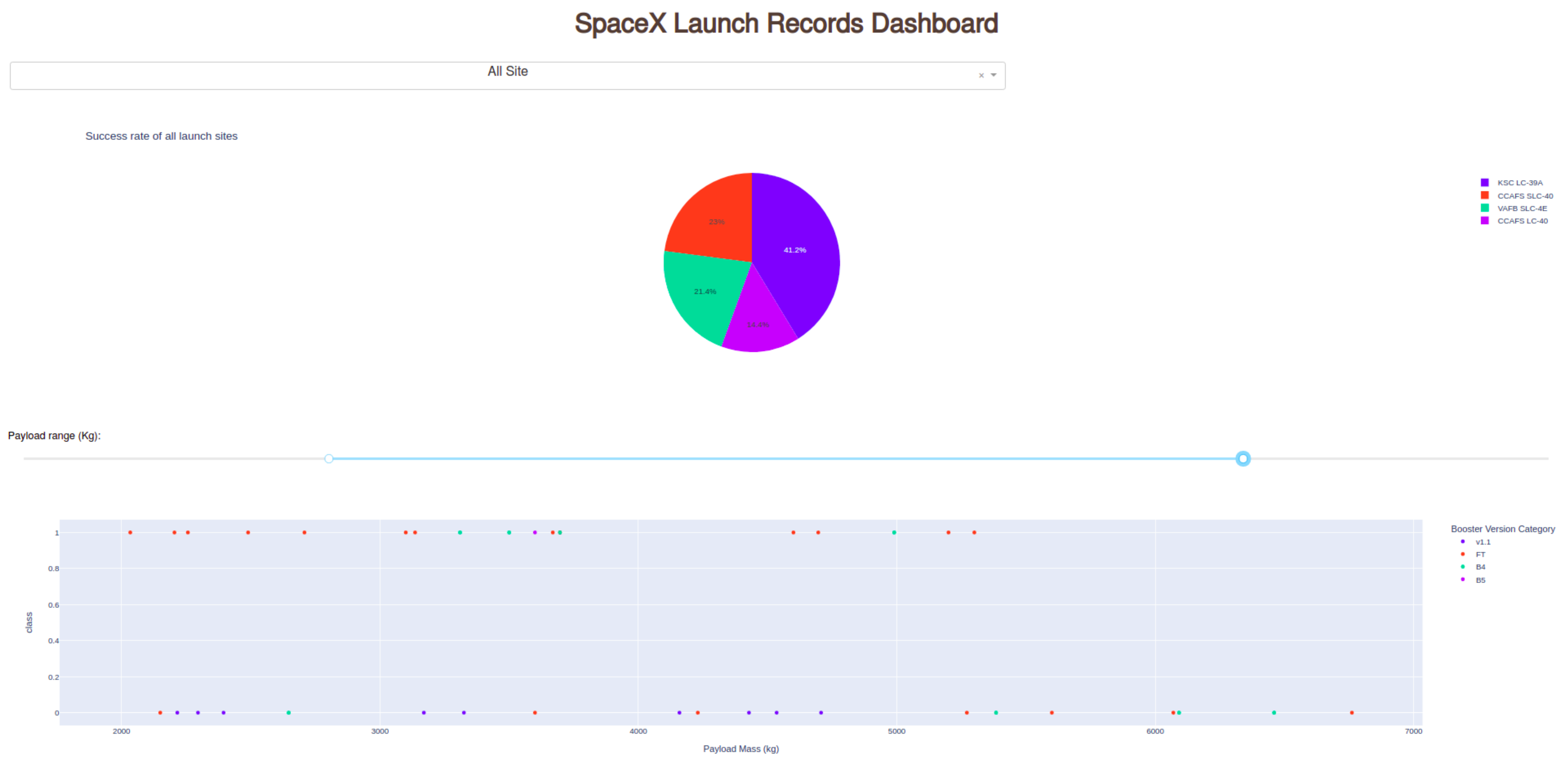


SpaceX Launch Records Dashboard

CCAFS SLC-40 × ▾

Success & Failure for selecte site

Payload vs. Launch outcome scatter plot



SpaceX Launch Records Dashboard

All Site

Success rate of all launch sites

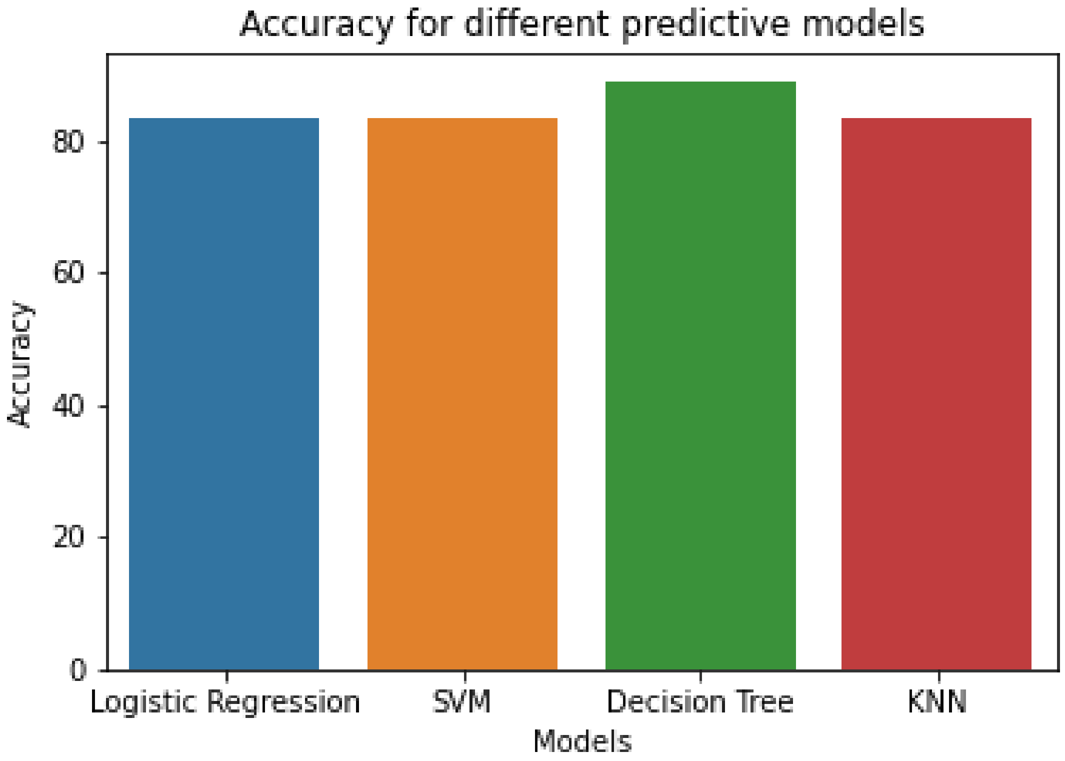
Payload range (Kg):

Predictive analysis (Classification)

Classification Accuracy

Decision tree model has the highest accuracy on test data as:

88.89%

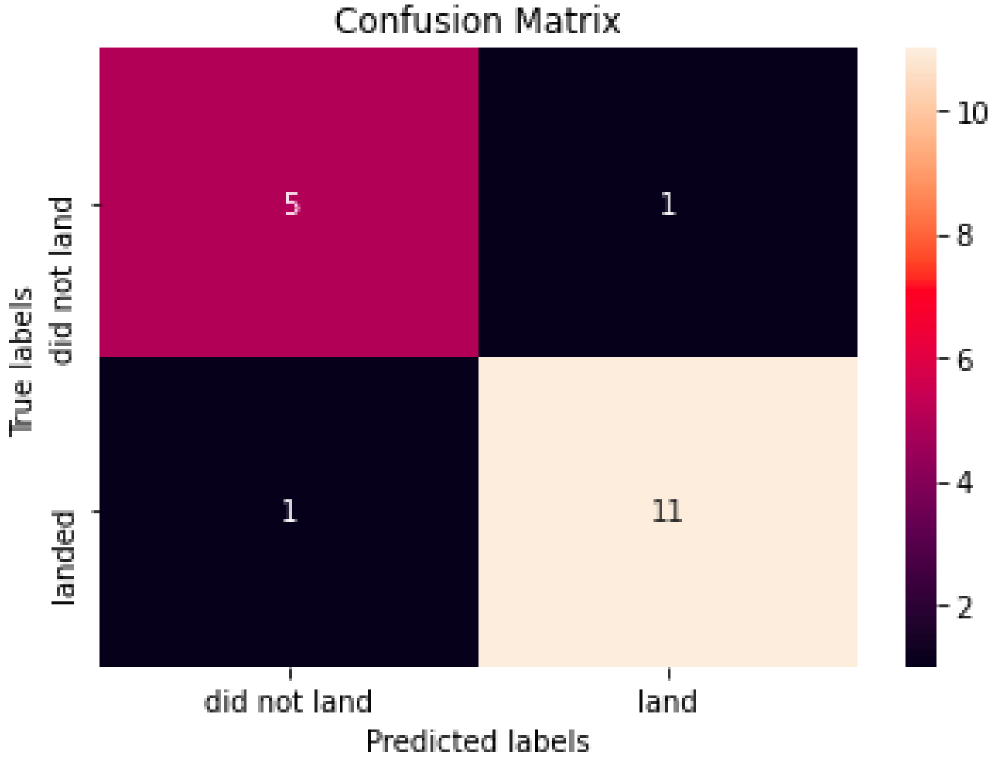


Confusion Matrix

As shown in the confusion matrix, the number in different blocks signifies-

True positive = 5 + 11
False positive = 1
False negative = 1

The number of false positive is less than other models.



CONCLUSION

- Few parameters have impact on the success of launch of first stage of falcon 9. E.g. Payload, launch site etc.
- With this current historical data the best predictive model can predict success or failure of launch with **88.89%** accuracy.

APPENDIX

- Plotting interactive charts with Plotly