Data Science Capstone project

<Bui Nguyen Hoang Trieu>

<25/08/2021>

Outline

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

Executive Summary

- Summary of methodologies
  - Python libraries: Pandas, NumPy, Matplotlib, Pyplot, Dash, scikit-learn, scipy, BeautifulSoup
  - Pandas, NumPy, Matplotlib, Pyplot: Data Analyzing & Visualization
  - scikit-learn, scipy: Machine Learning models
  - BeautifulSoup: Web scraping
- Summary of all results
  - By using various techniques and analysis, the optimum models are found.
  - However, there are considerations that must be taken along with the outcome from the model.
  - Other techniques to assess the success rate of launch is necessary to save costs and risk management.

Introduction

- Project background and context
  - SpaceY, a new rocket company wants to compete with SpaceX

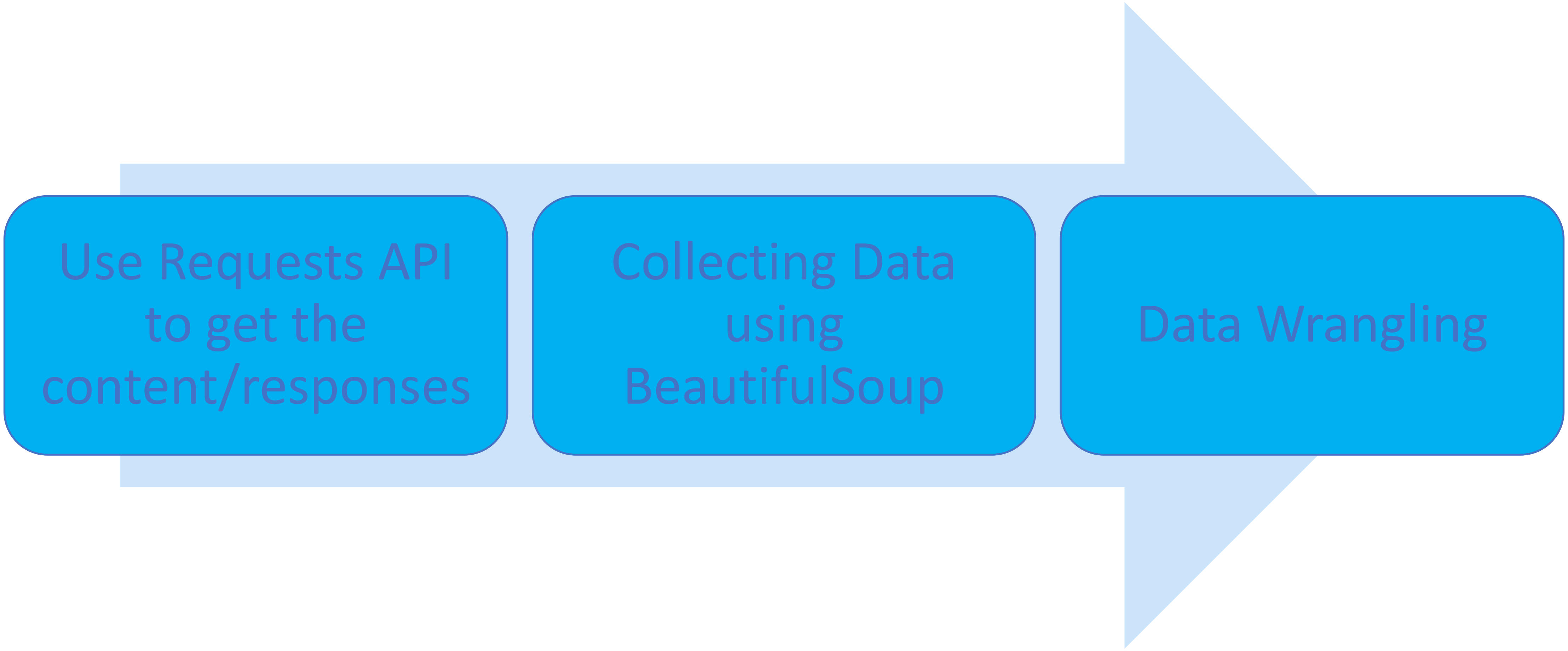
- Problems you want to find answers
  - Determining the price of each launch
  - Using Machine Learning models to determine if the land will success or not
  - Predict if we can reuse the first stage

Methodology

- Data collection methodology:
  - Use web scraping from SpaceX public data or Wikipedia
- Perform data wrangling
  - Classify landing outcomes to Class 0 or 1, which 0 means a bad outcome and 1 is a good outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train the model and perform Grid Search to find best hyperparameters for Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors models.

Methodology

Data collection



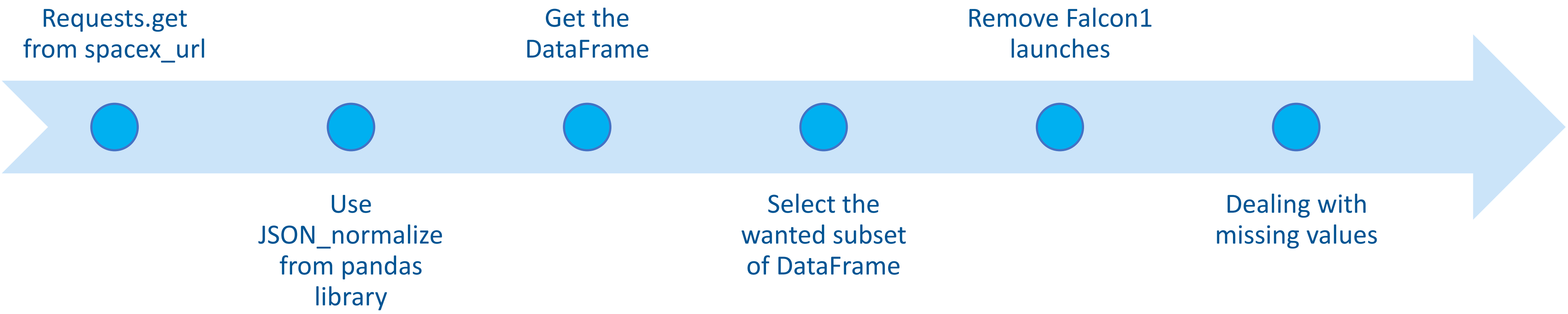
Use Requests API to get the content/responses

Collecting Data using BeautifulSoup

Data Wrangling

Data collection – GitHub
SpaceX API

Flowchart of SpaceX API calls here



Requests.get from spacex_url

Use JSON_normalize from pandas library

Get the DataFrame

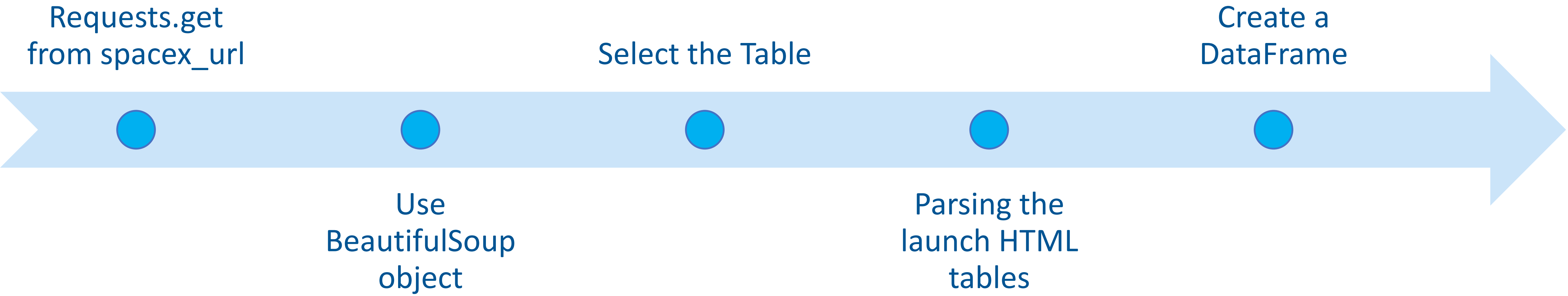
Select the wanted subset of DataFrame

Remove Falcon1 launches

Dealing with missing values

Data collection – GitHub
Web Scraping

Flowchart of Web Scraping here



Requests.get from spacex_url

Use BeautifulSoup object

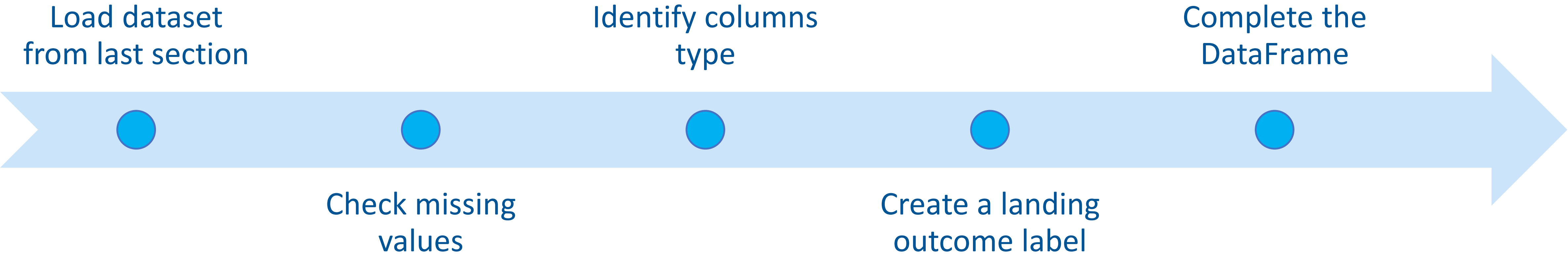
Select the Table

Parsing the launch HTML tables

Create a DataFrame

Data Wrangling GitHub

Flowchart of Data Wrangling here



Load dataset from last section

Check missing values

Identify columns type

Create a landing outcome label

Complete the DataFrame

EDA with data visualization

GitHub URL: GitHub

Charts for exploration:

- FlightNumber vs PayloadMass (Scatter)
- FlightNumber vs LaunchSite (Scatter)
- PayloadMass vs LaunchSite (Scatter)
- Orbit vs FlightNumber (Scatter)
- PayloadMass vs Orbit (Scatter)
- Success rate of each Orbit (Bar)
- Yearly trend of success rate (Line)

Scatter	Determining the relationship between various parameters and their combined influence on the success rate
Bar	Determining success rate of each orbit and comparison
Line	Determining the trend of success rate over time

EDA with SQL

GitHub URL: GitHub

Summarize performed SQL queries using bullet points

- SELECT DISTINCT (LAUNCH_SITE) FROM SPACEXTBL
- SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS_CARRIED FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)’
- SELECT avg(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_CARRIED FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1’
- select min(DATE) from spacextbl where landing__outcome like '%Success (ground pad)%’
- select unique(booster_version) from spacextbl where where landing__outcome = 'Success (drone ship)' andpayload_mass__kg_ between 4000 and 6000;
- select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome;
- select unique(booster_version) from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
- select landing__outcome,booster_version,launch_site from spacextbl where landing__outcome='Failure (drone ship)' and date like '2015%’
- select landing__outcome, count(landing__outcome) as Total from spacextbl where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome) desc

Build an interactive map with Folium

GitHubURL: GitHub

Added:

- A blue circle at NASA Johnson Space Center's coordinate with a icon showing its name
- For each launch site, add a Circle object based on its coordinate (Lat, Long) values. In addition, add Launch site name as a popup label
- Marker clusters were added to group the large number of launch sites together on the map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
- A polyline between a selected launch site and coastline indicating the distance

Build a Dashboard with Plotly Dash

GitHubURL: GitHub

Added:

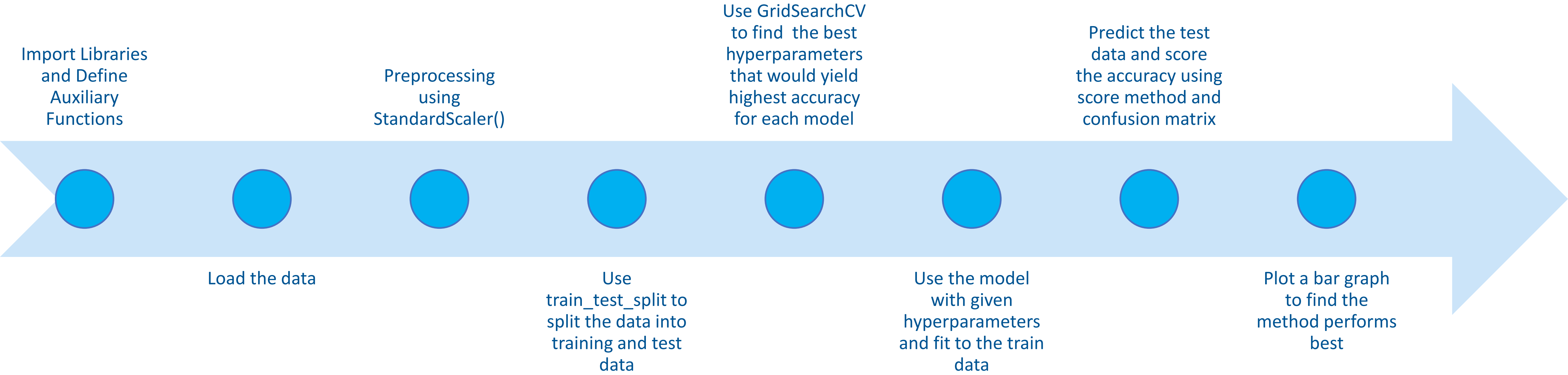
- A Launch Site Drop-down Input Component
- A callback function to render Pie Chart of success rate based on selected site dropdown
- A Range Slider to Select Payload
- A callback function to render Scatter Plot

Pie Chart is connected to the Dropdown

Scatter Plot is connected to both Dropdown and Range Slider

Predictive analysis (Classification)

GitHubURL: GitHub



Import Libraries and Define Auxiliary Functions

Load the data

Preprocessing using StandardScaler()

Use train_test_split to split the data into training and test data

Use GridSearchCV to find the best hyperparameters that would yield highest accuracy for each model

Use the model with given hyperparameters and fit to the train data

Predict the test data and score the accuracy using score method and confusion matrix

Plot a bar graph to find the method performs best

Results

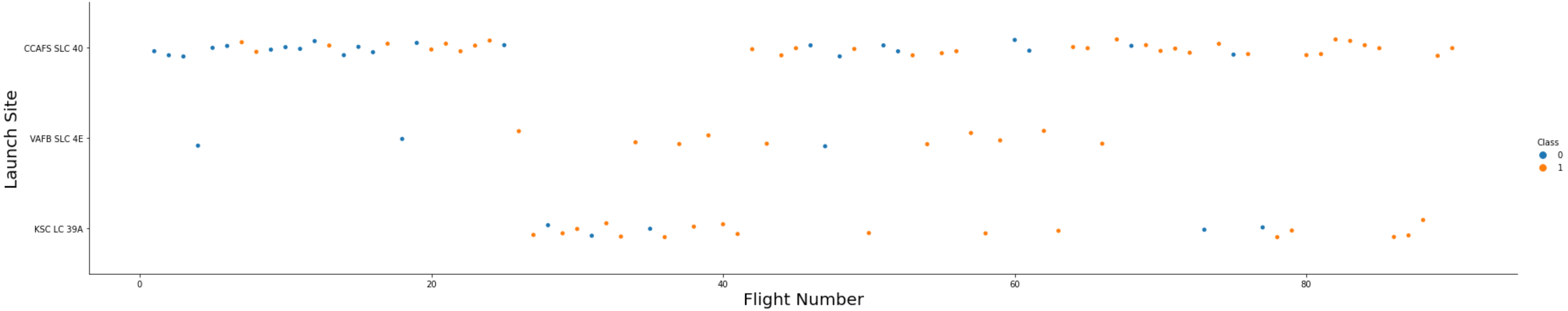
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA with Visualization

Flight Number vs. Launch Site

The success rate increase along with the flight number.

CCAFS LC-40 have lower success rate compared to the remaining launch site.



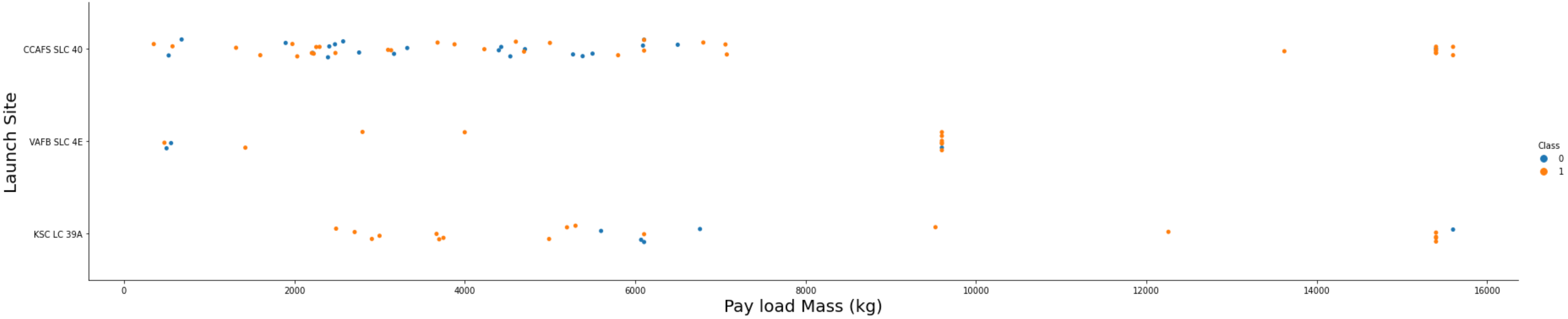
Payload vs. Launch Site

CCAFS SLC 40 has no clear pattern between Payload Mass and Launch Site.

The remain two has higher success rate.

VAFB SLC 4E did not test Payload Mass over 10,000kg

At Payload Mass over 14,000 kg, CCAFS SLC 40 has higher success rate compared to KSC LC 39A

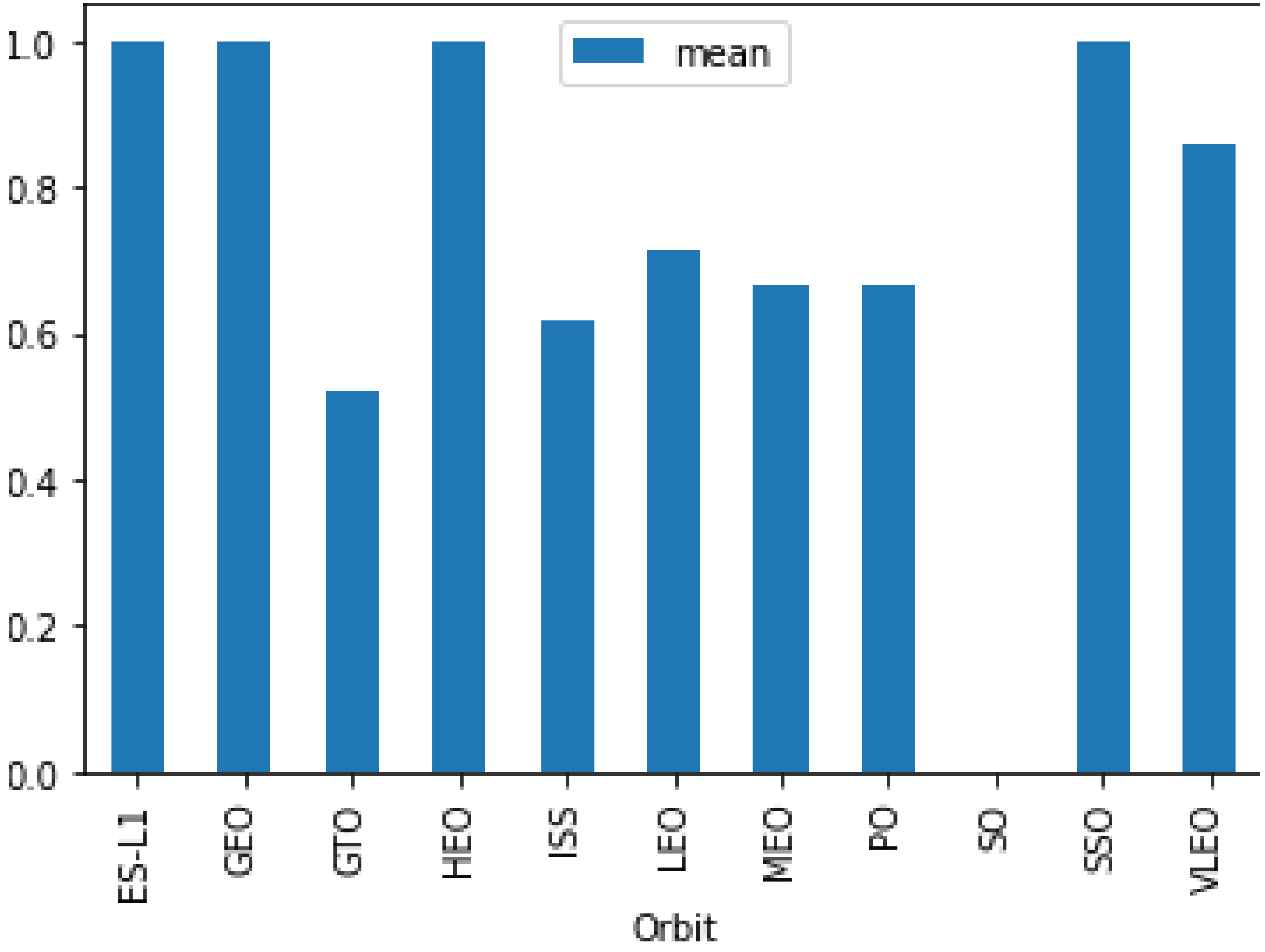


Success rate vs. Orbit type

SO has no successful landing.

ES-L1, GEO, HEO, SSO has 100% success rate.

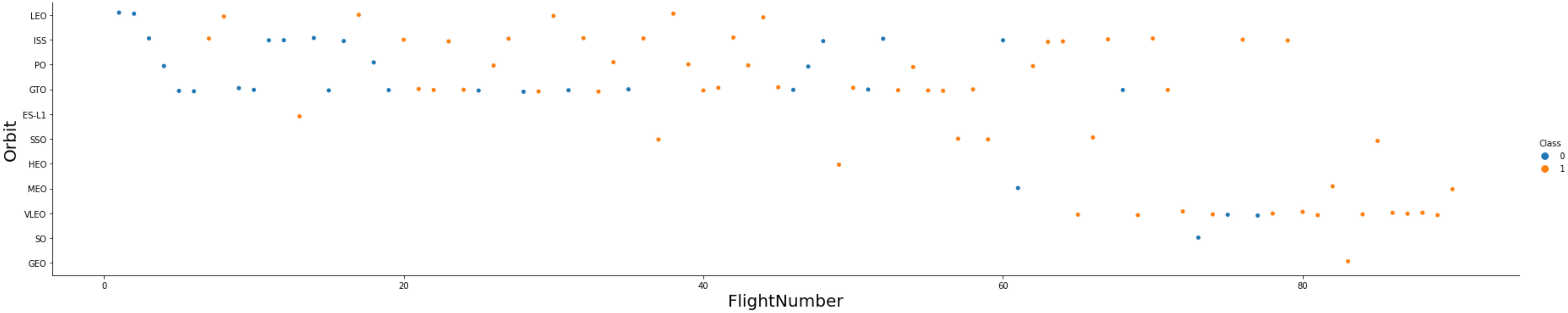
Other than SO, GTO has the lowest success rate which is only about 50%.



Flight Number vs. Orbit type

It can be seen that for LEO and VLEO orbits 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.

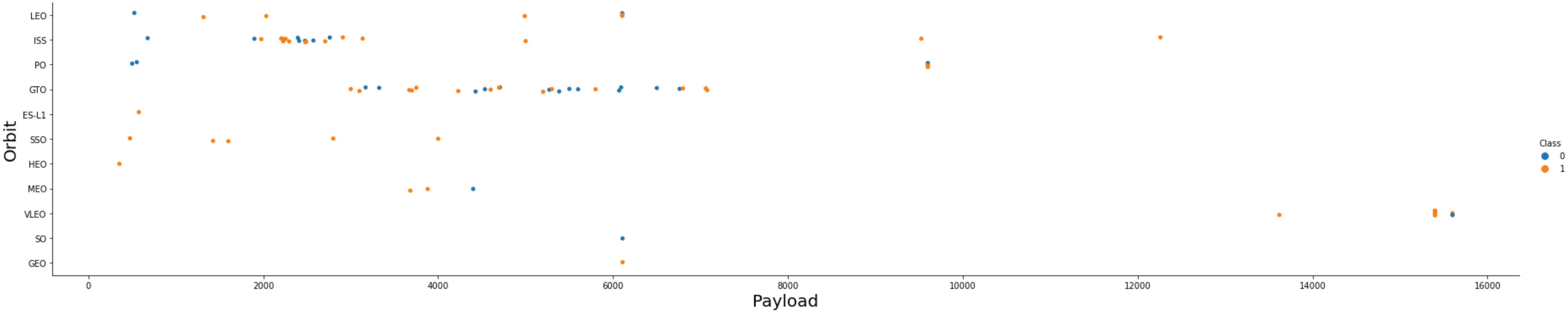
For success rate of each orbit, please refer to previous slide.



Payload vs. Orbit type

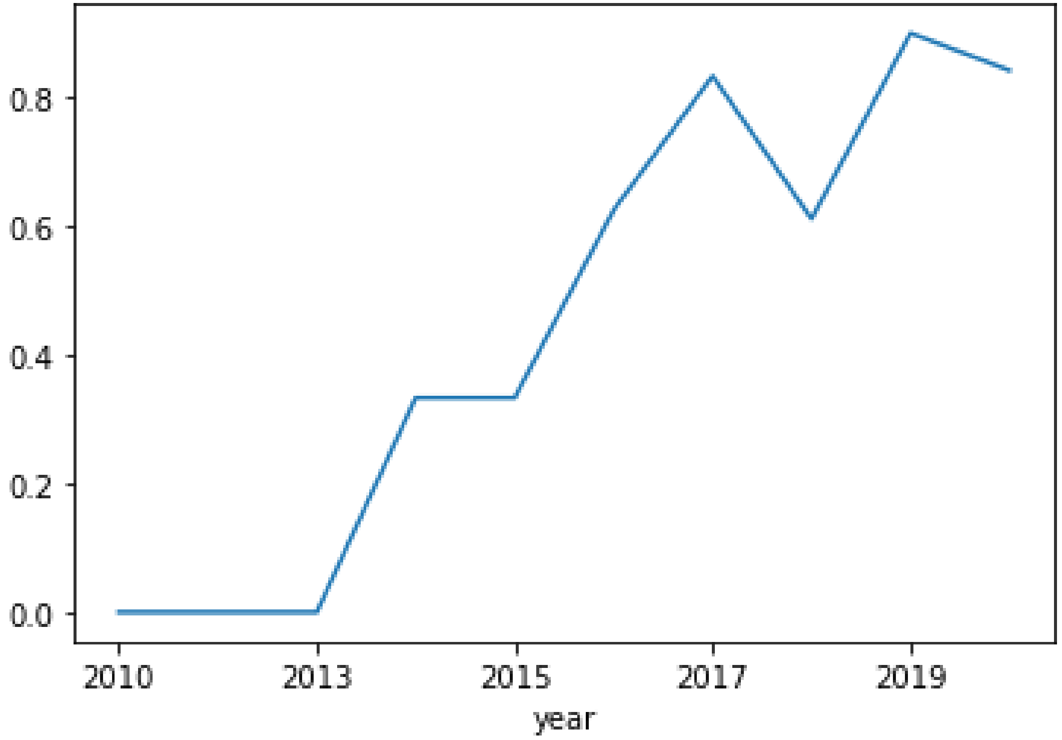
Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

The capability of payload over 14,000 kg at VLEO orbit seems to reach its maximum at around 16,000 kg.



Launch success yearly trend

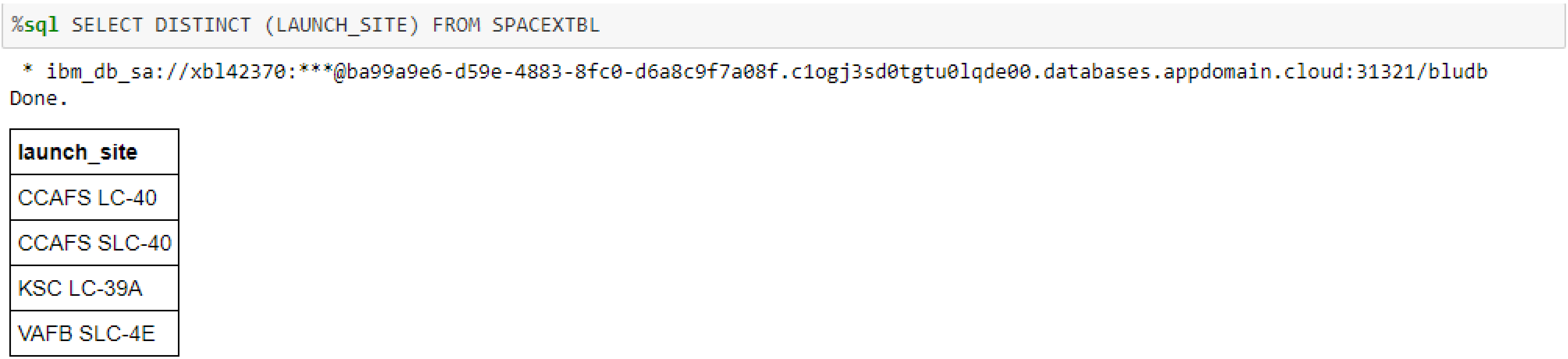
- The success rate since 2013 kept increasing till 2020.



EDA with SQL

All launch site names

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



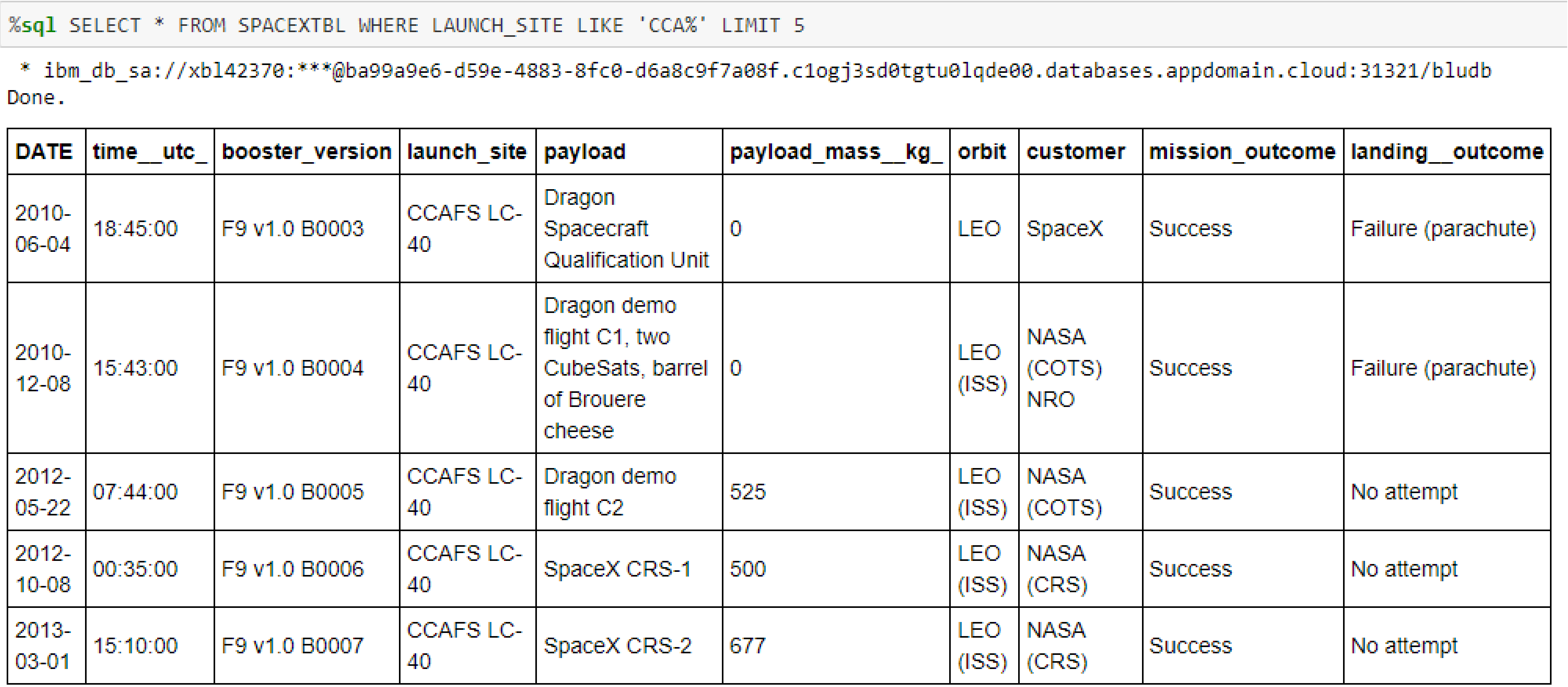
```
%sql SELECT DISTINCT (LAUNCH_SITE) FROM SPACEXTBL
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

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

Launch site names begin with `CCA`

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



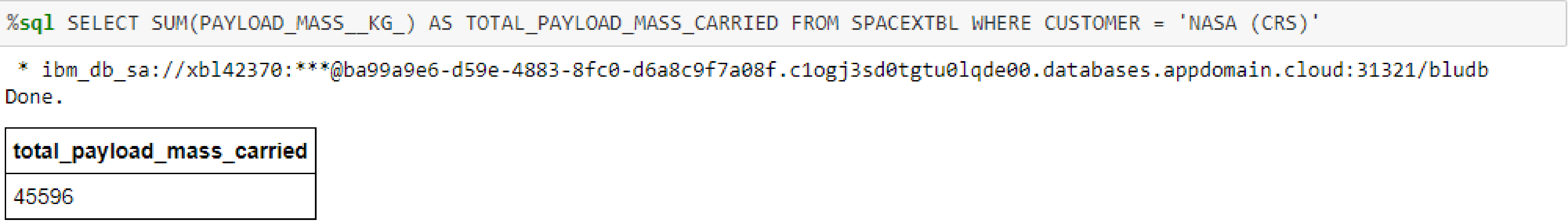
```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

The total payload mass carried by boosters launched by NASA (CRS)



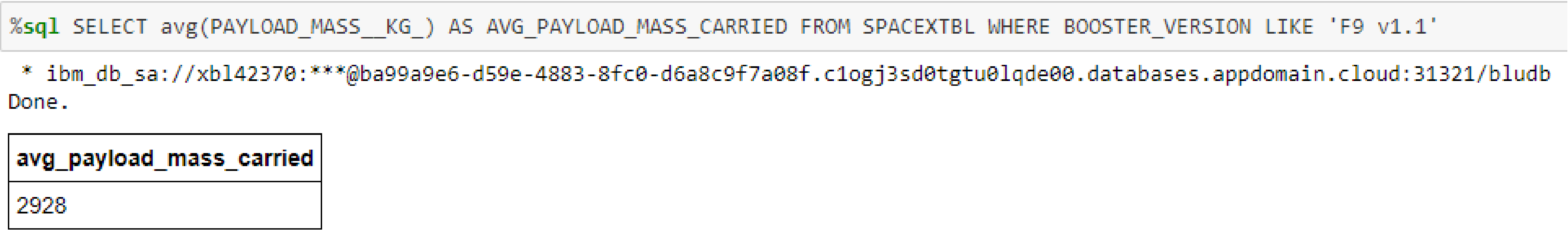
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS_CARRIED FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

total_payload_mass_carried
45596

Average payload mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1



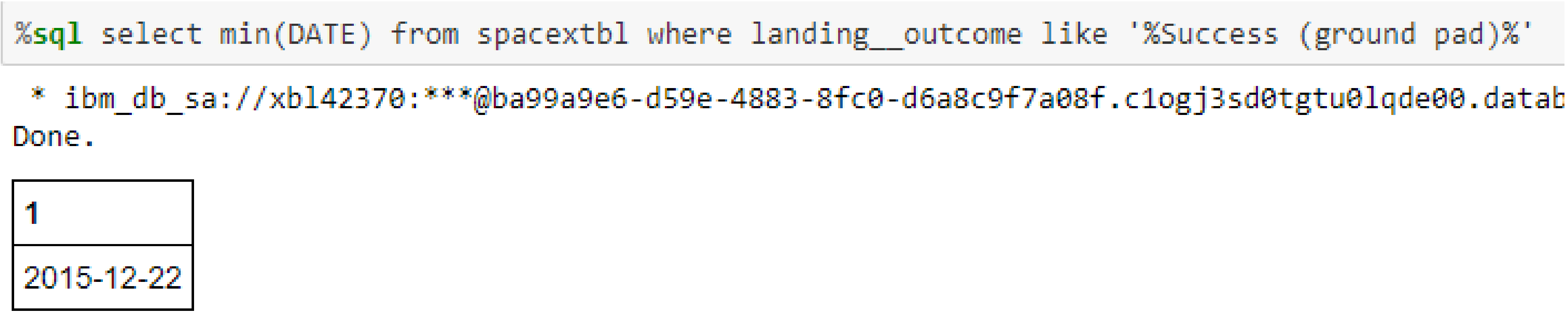
```
%sql SELECT avg(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_CARRIED FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

avg_payload_mass_carried
2928

First successful ground landing date

The date when the first succesful landing outcome in ground pad was achieved



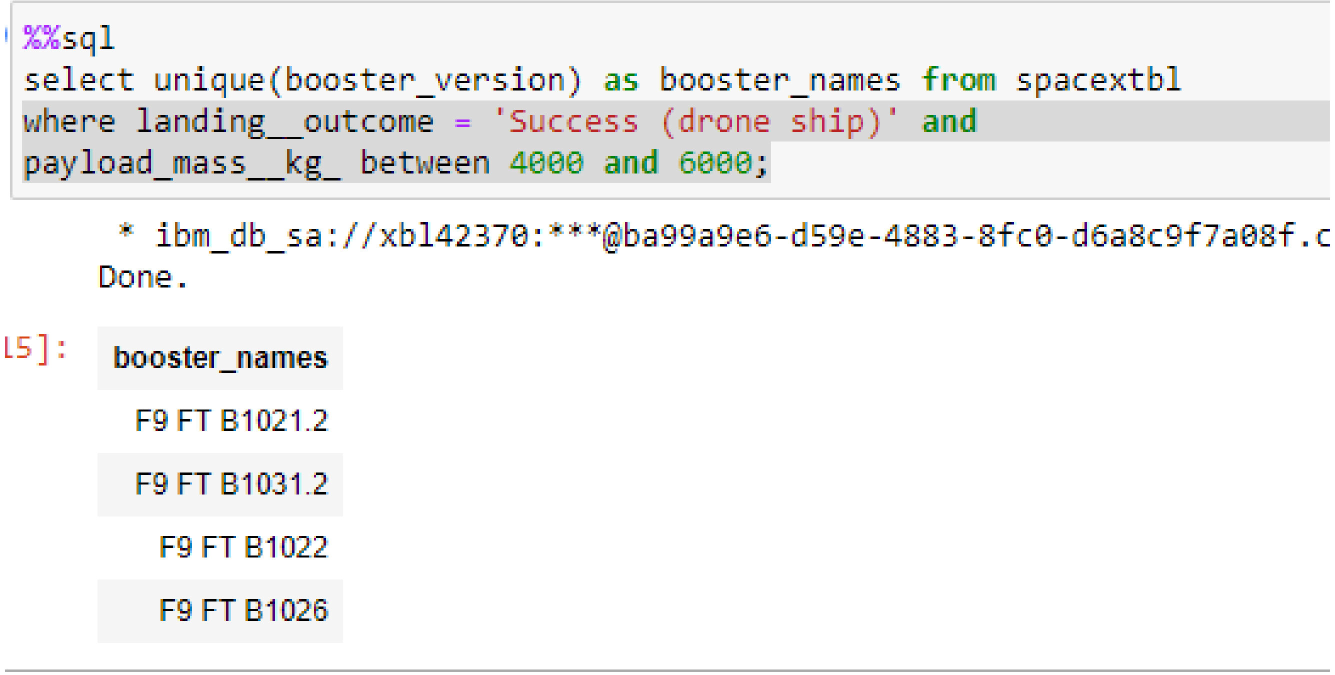
```
%sql select min(DATE) from spacextbl where landing__outcome like '%Success (ground pad)%'
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.datab
Done.
```

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

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



```
%%sql
select unique(booster_version) as booster_names from spacextbl
where landing__outcome = 'Success (drone ship)' and
payload_mass__kg_ between 4000 and 6000;
```

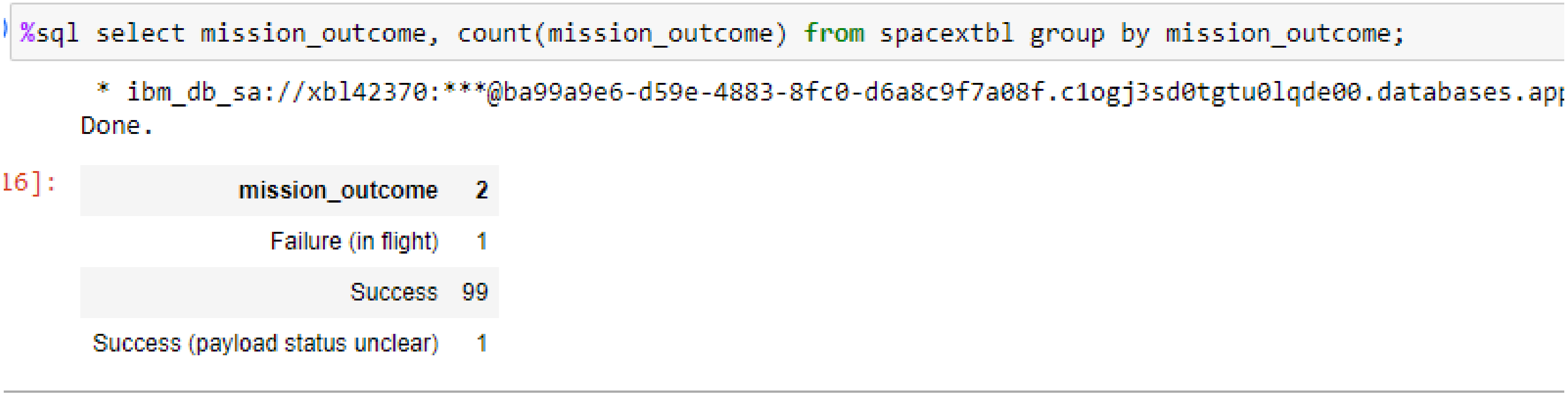
```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c
Done.
```

l5]:

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

Total number of successful and failure mission outcomes

The total number of successful and failure mission outcomes



```
%sql select mission_outcome, count(mission_outcome) from spacextbl group by mission_outcome;
```

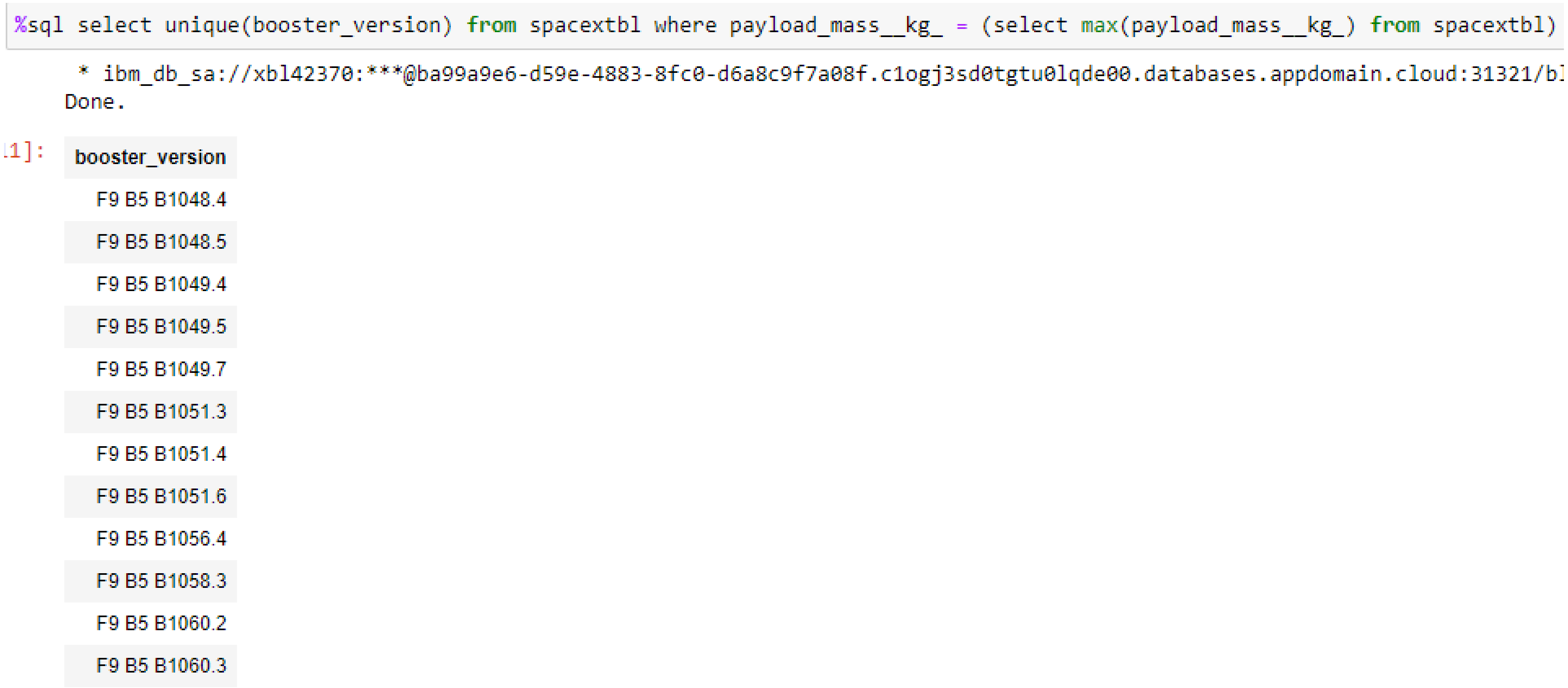
```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.apl
Done.
```

16]:

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum payload

The names of the booster_versions which have carried the maximum payload mass



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

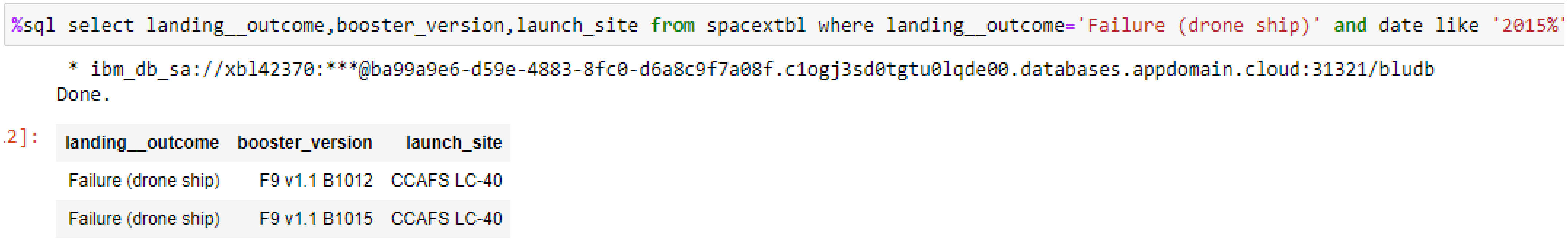
```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bl
Done.
```

1]:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 launch records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015



```
%sql select landing__outcome,booster_version,launch_site from spacextbl where landing__outcome='Failure (drone ship)' and date like '2015%'
```

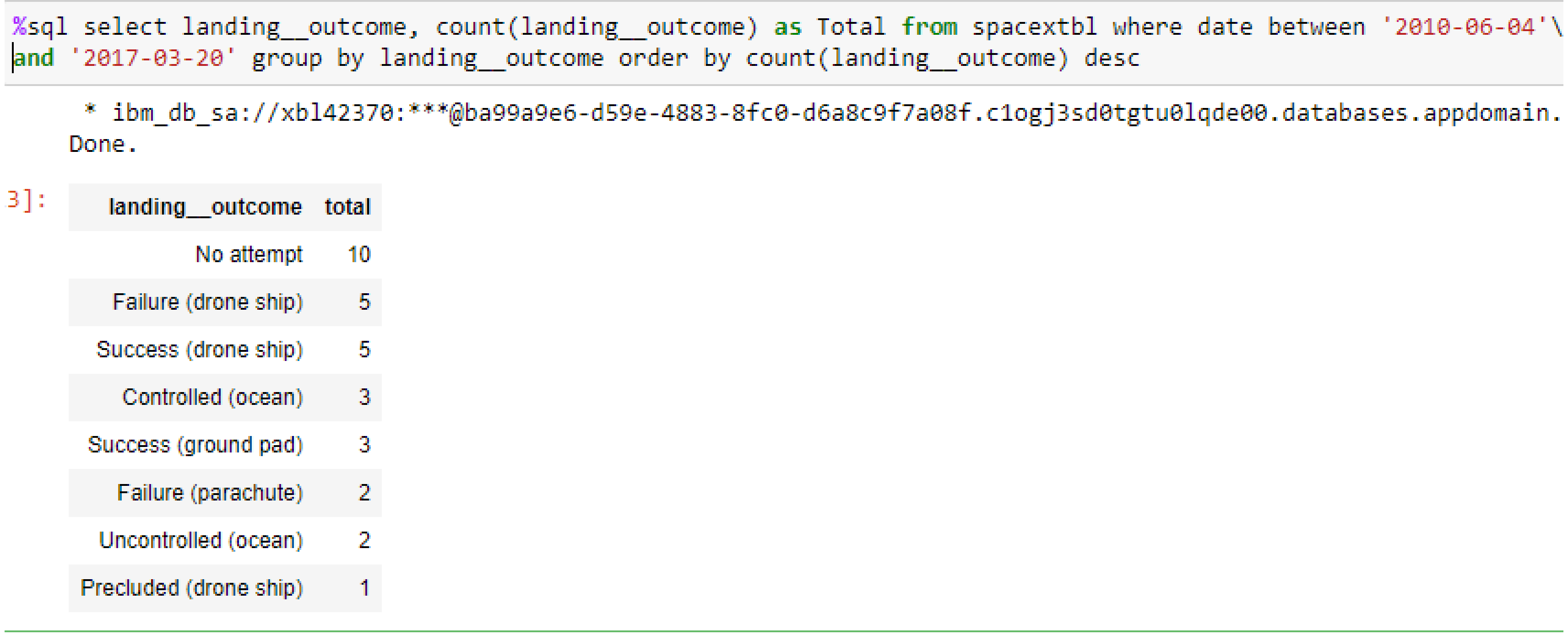
```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

.2]:

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

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

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



```
%sql select landing__outcome, count(landing__outcome) as Total from spacextbl where date between '2010-06-04'\
and '2017-03-20' group by landing__outcome order by count(landing__outcome) desc
```

```
 * ibm_db_sa://xbl42370:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.
Done.
```

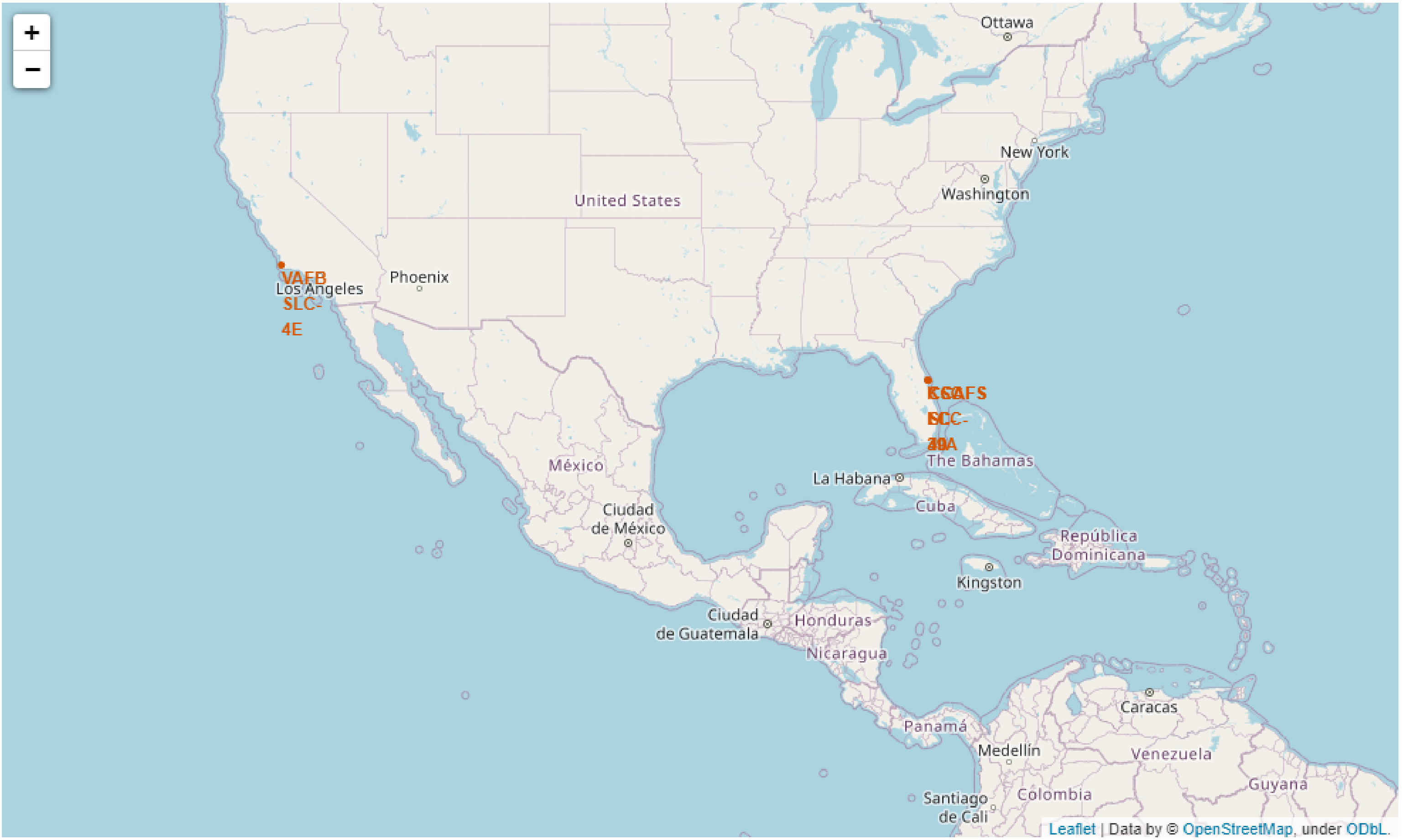
3]:

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

Interactive map with Folium

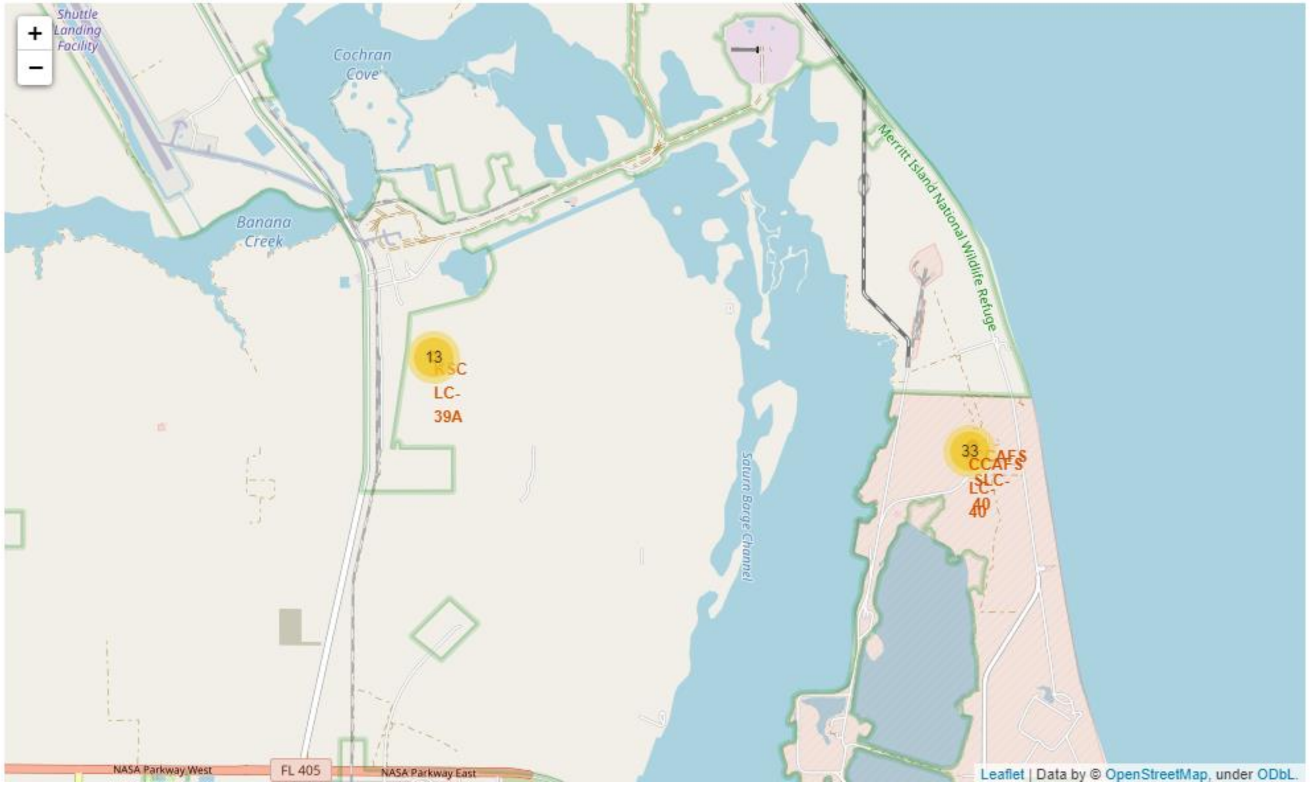
All Launch Sites Locations

All launch sites are in coastal areas



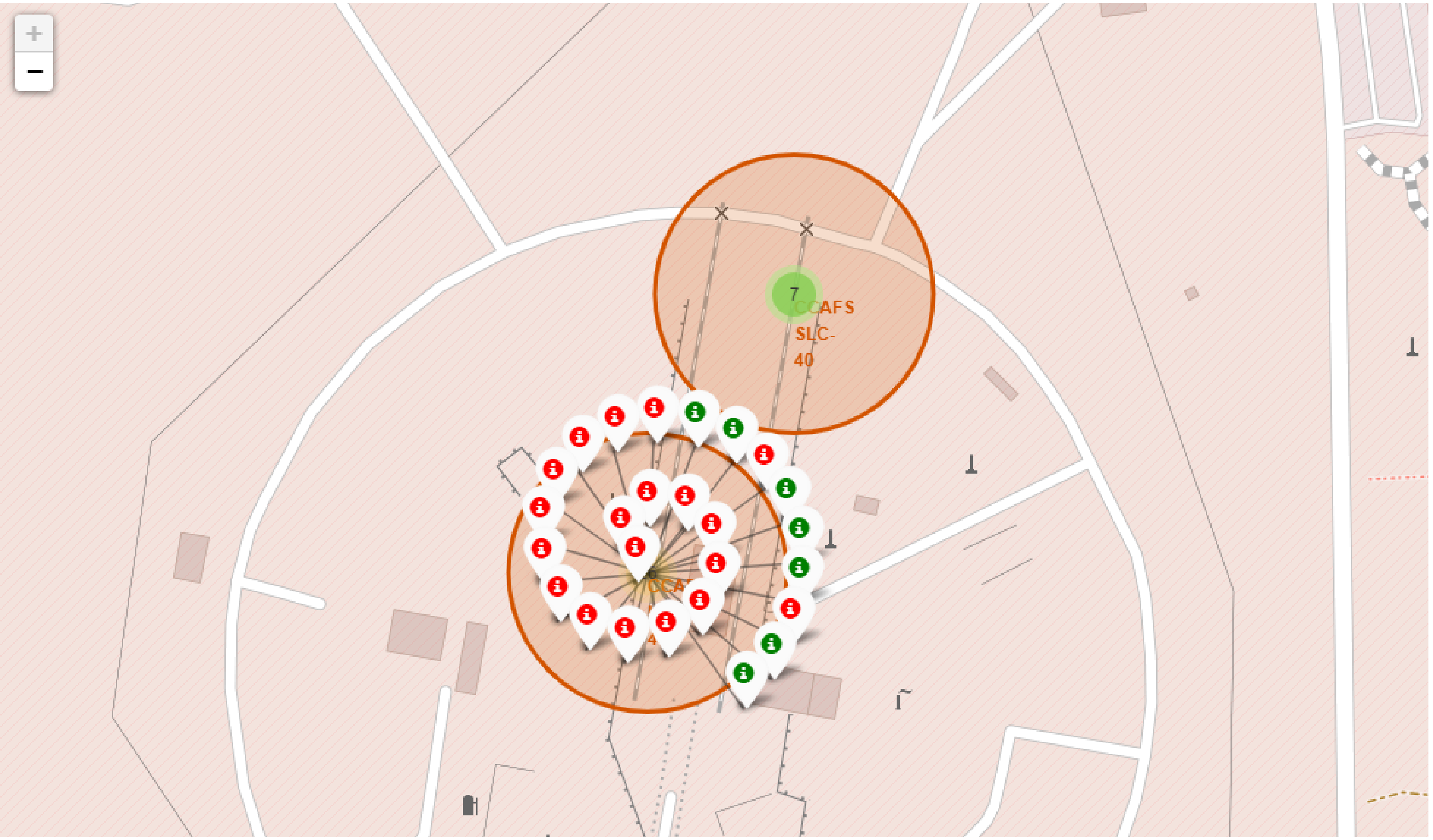
Marked outcome of launches for each site on the map

The outcome of each launch at each marker cluster are color coded with success being green and failures indicated by red.

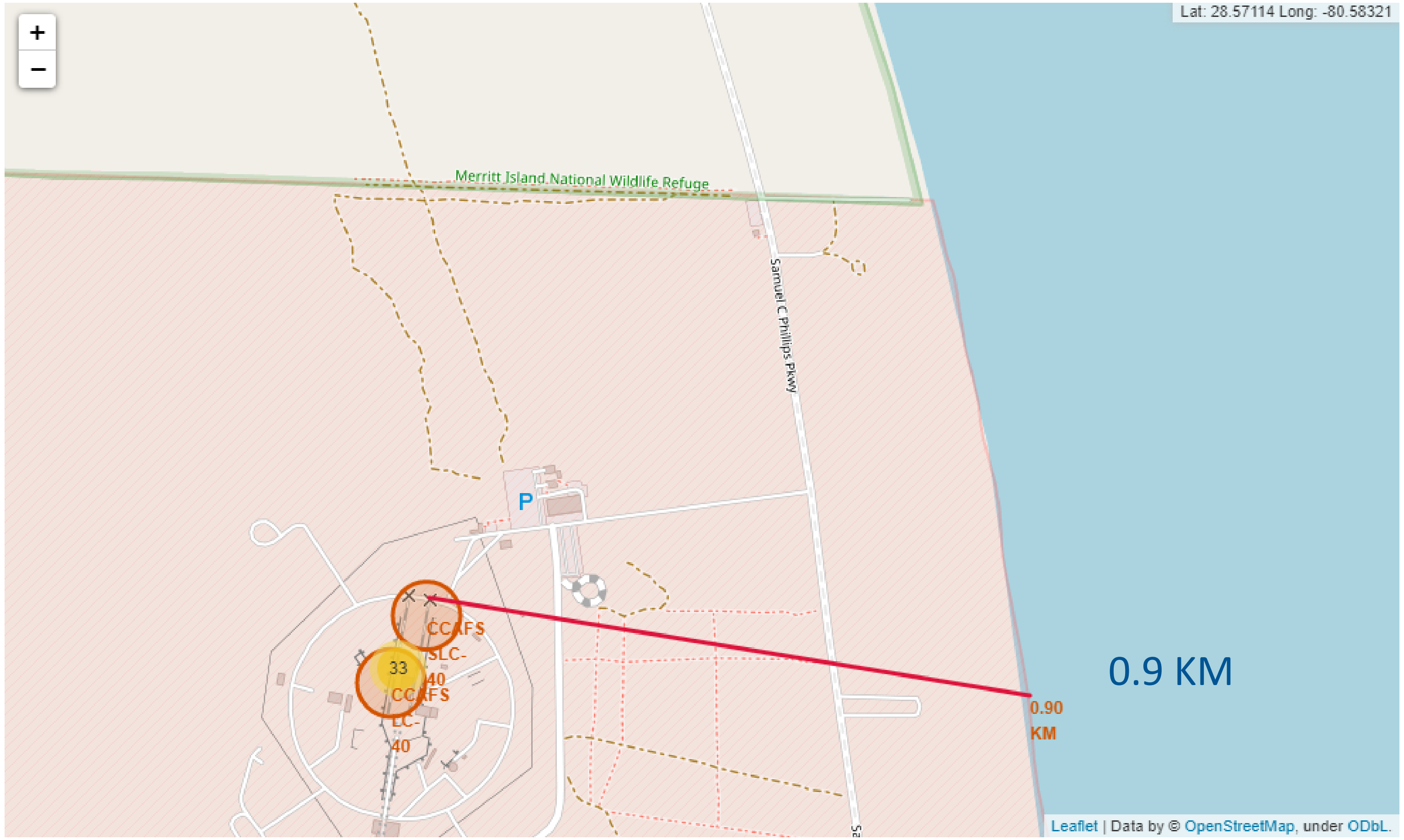


Marked outcome of launches for each site on the map

The outcome of each launch at each marker cluster are color coded with success being green and failures indicated by red.

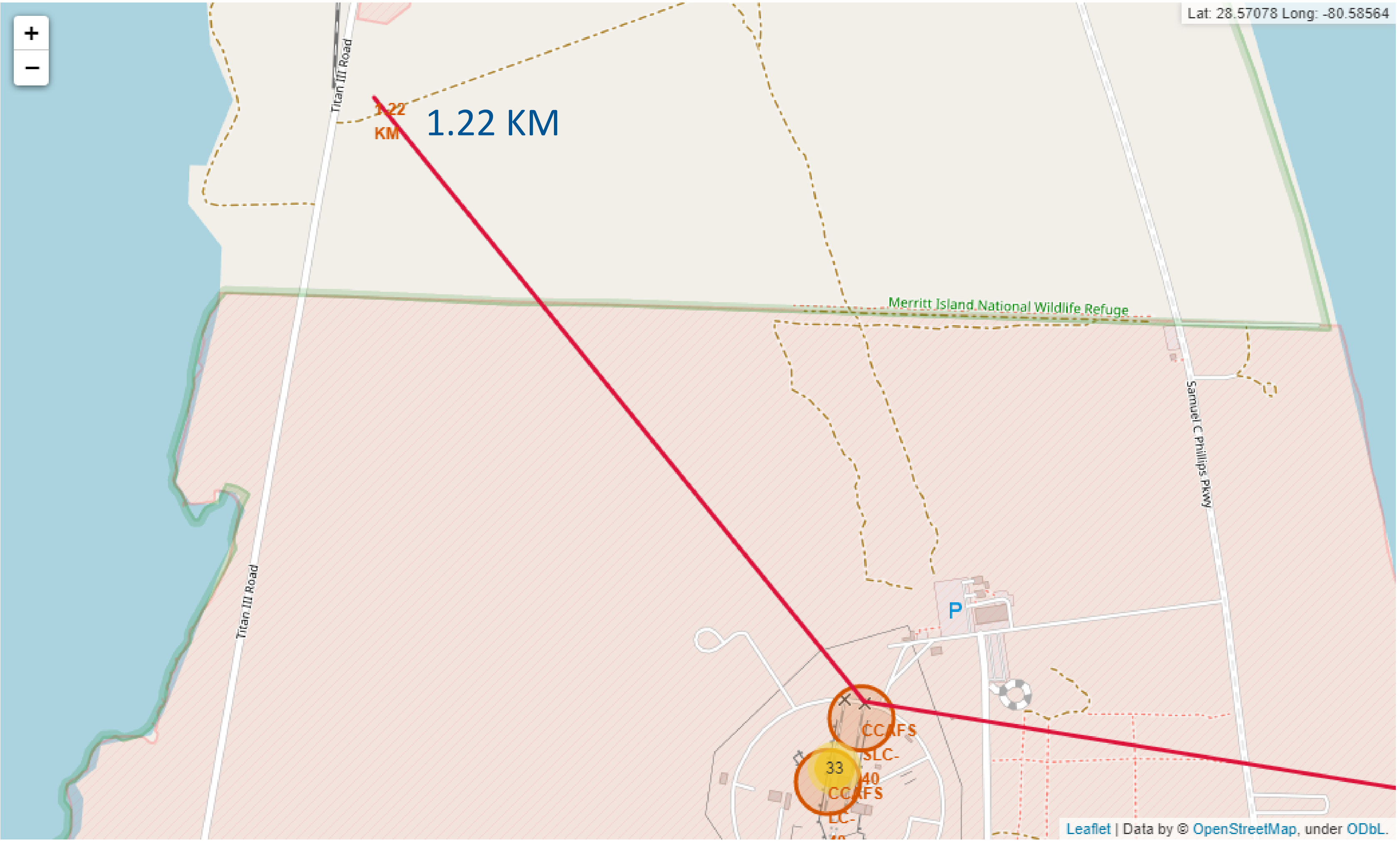


Distances between launch site and the coastline


+
−
Lat: 28.57114 Long: -80.58321
Merritt Island National Wildlife Refuge
Samuel C Phillips Pkwy
P
CCAFS
SLC-
40
33
CCAFS
LC-
40
0.90
KM
0.9 KM
Leaflet | Data by © OpenStreetMap, under ODbL.

Distances between launch site and the railway

```
1.2173502585644906
```



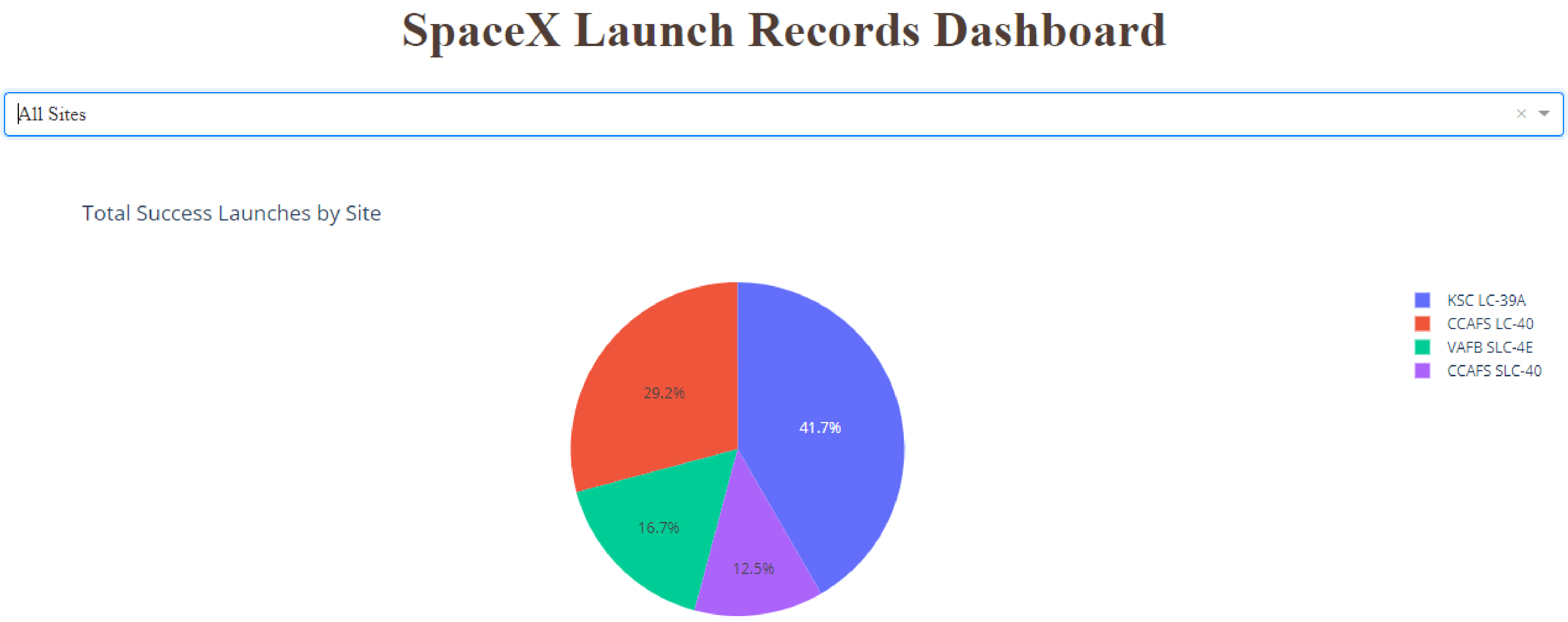
1.22 KM

Lat: 28.57078 Long: -80.58564
+
−
Titan III Road
1.22 KM
Merritt Island National Wildlife Refuge
Samuel C Phillips Pkwy
P
CCAFS SLC-40
33
CCAFS LC-40
Leaflet | Data by © OpenStreetMap, under ODbL.

Build a Dashboard with Plotly Dash

Success Rate of all launch sites

KSC LC-39A has the highest success rate while CCAFS SLC-40 has the lowest.

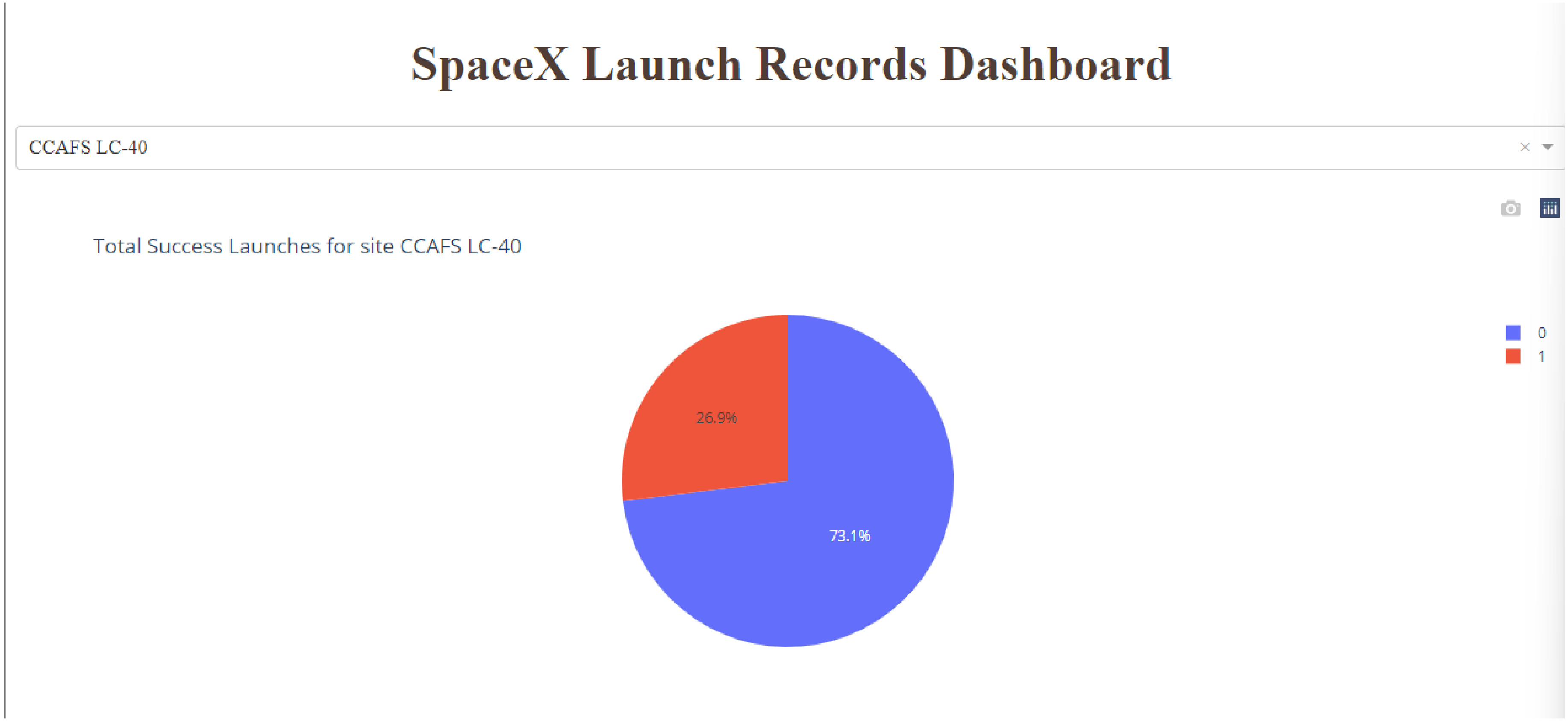


SpaceX Launch Records Dashboard

All Sites

Success Rate of CCAFS LC-40

Launches at this site has nearly 27% success rate.



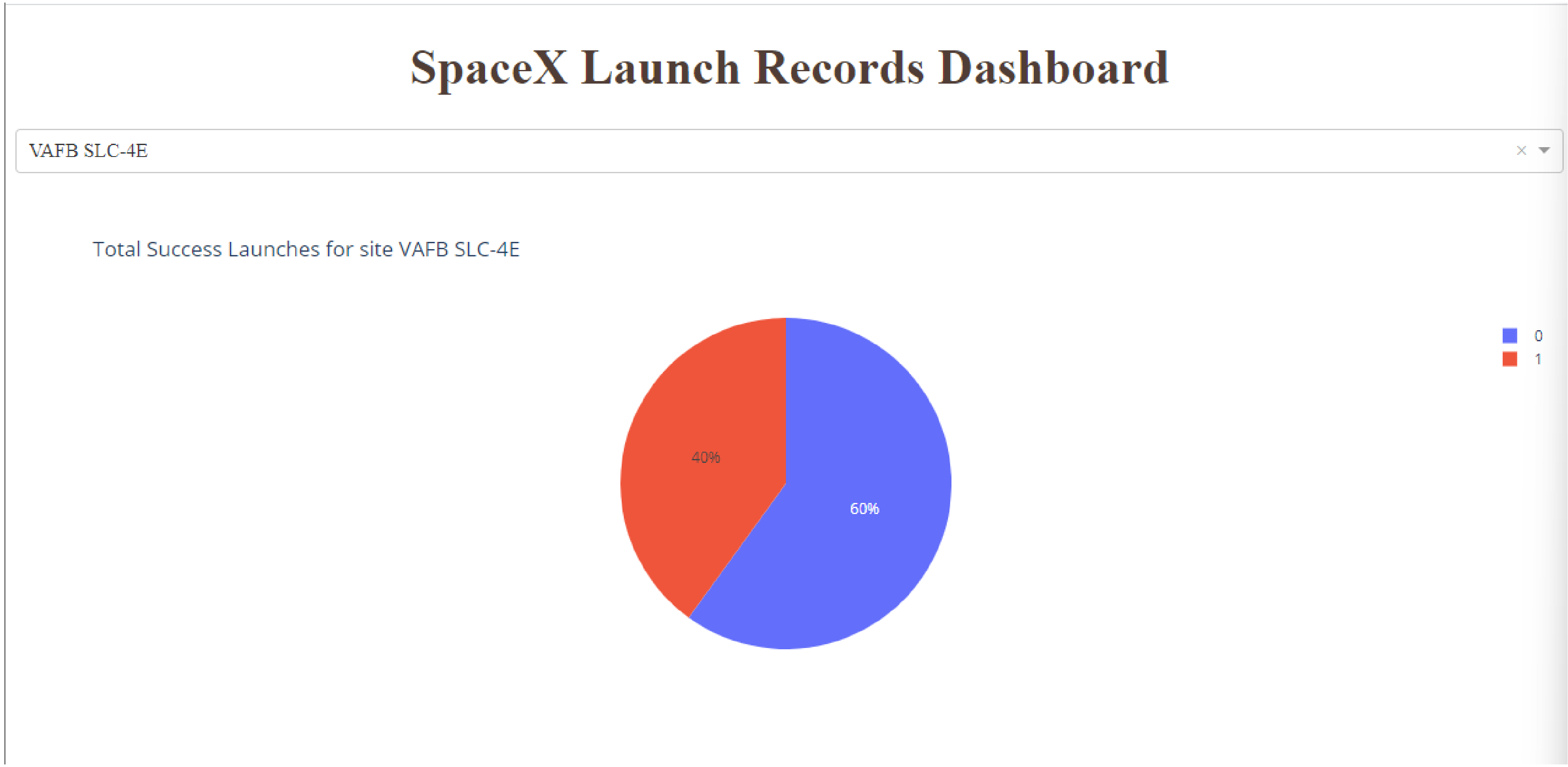
SpaceX Launch Records Dashboard

CCAFS LC-40

Total Success Launches for site CCAFS LC-40

Success Rate of VAFB SLC-4E

Launches at this site has 40% success rate.



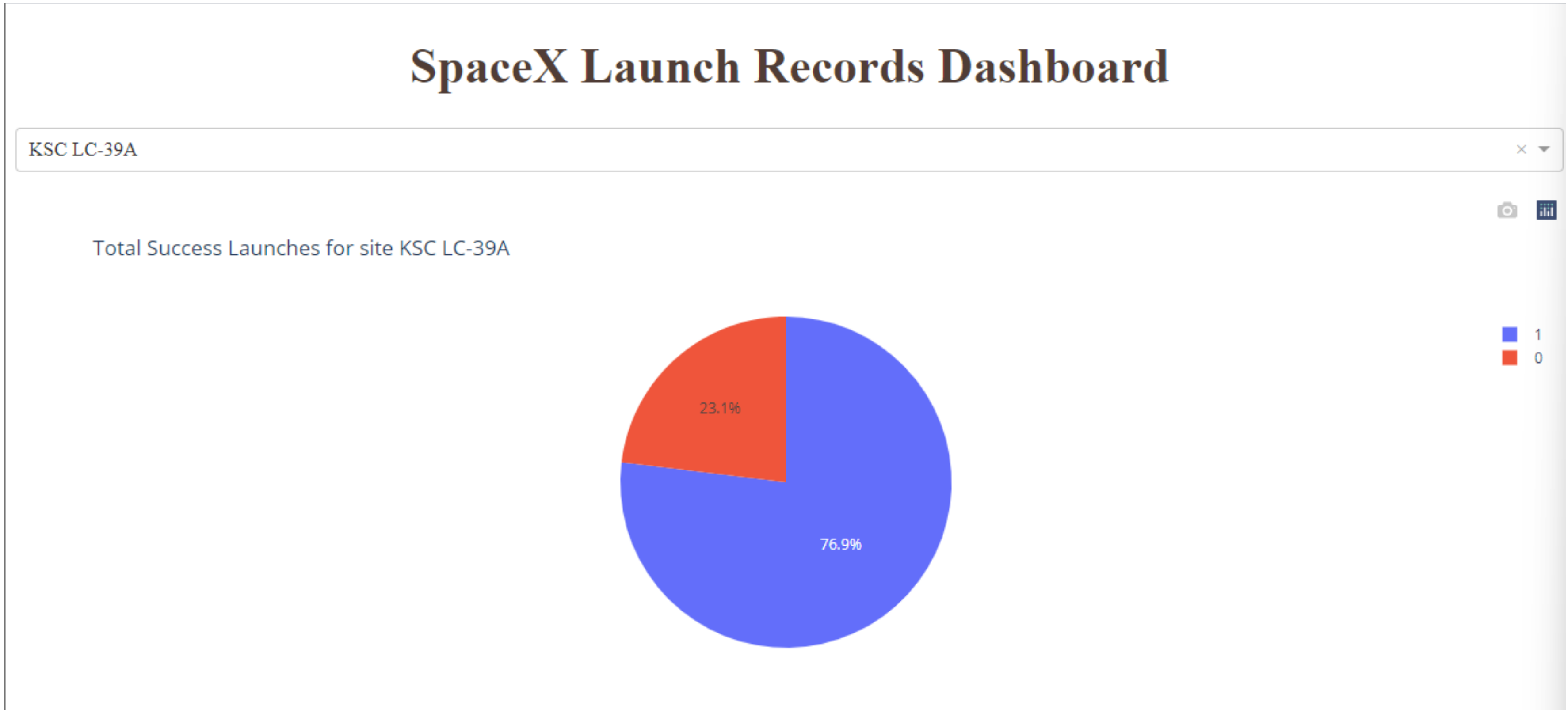
SpaceX Launch Records Dashboard

VAFB SLC-4E

Total Success Launches for site VAFB SLC-4E

Success Rate of KSC LC-39A

Launches at this site has over 23% success rate.



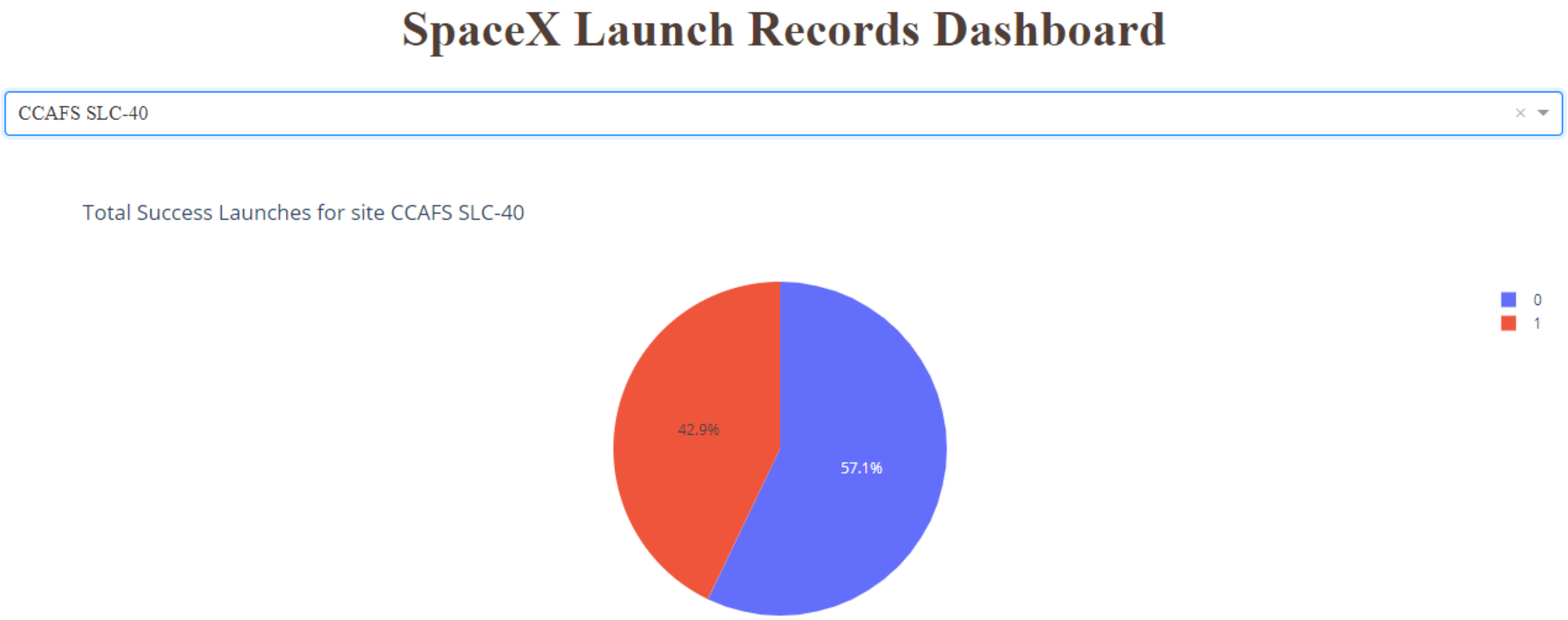
SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Success Rate of CCAFS SLC-40

Launches at this site has over 42% success rate.



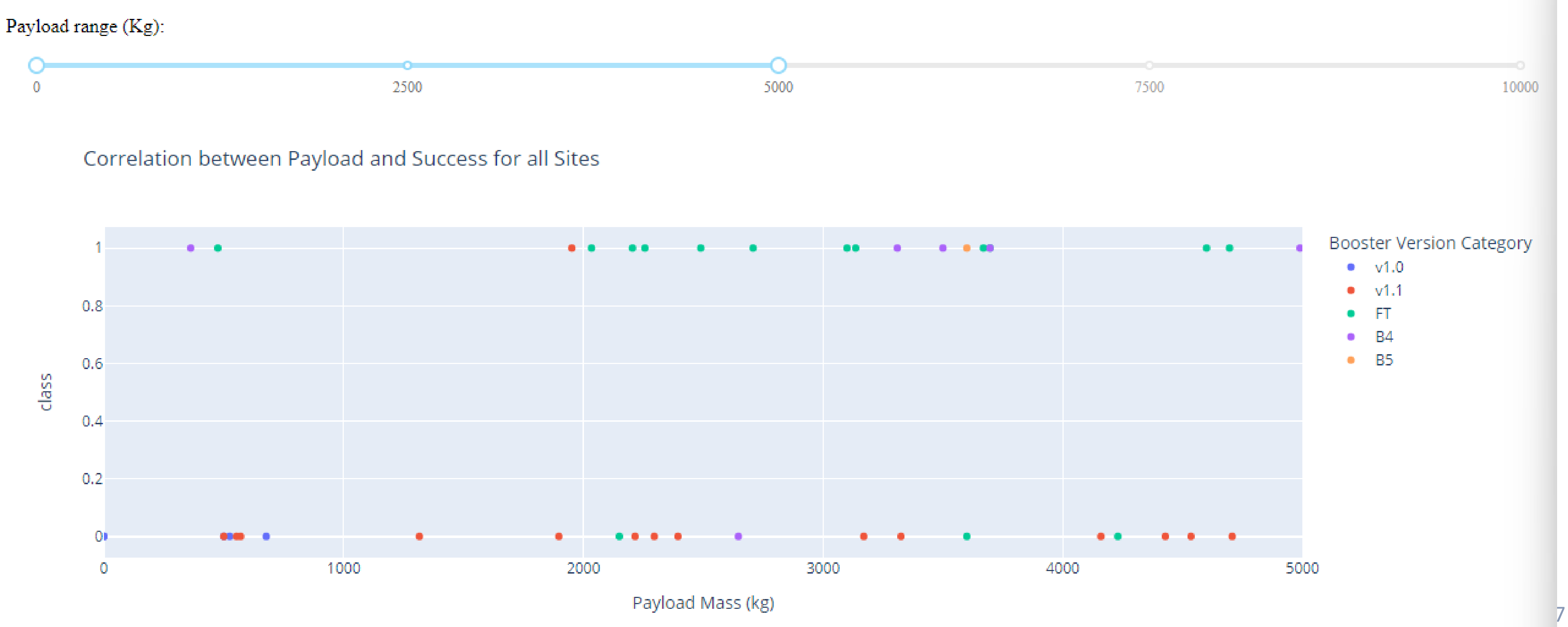
SpaceX Launch Records Dashboard

CCAFS SLC-40

Total Success Launches for site CCAFS SLC-40

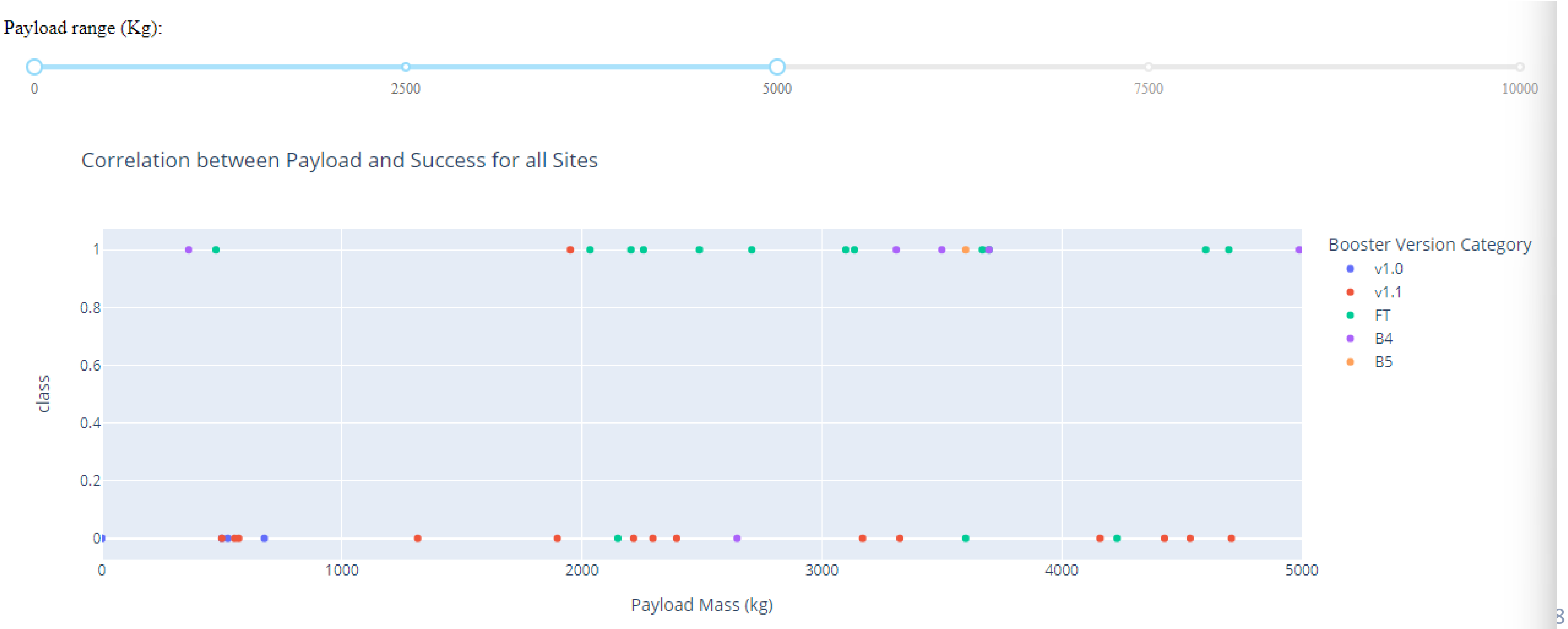
Payload vs Launch Outcome with Range Slider (1)

For payload range under 5000 kg, Booster v1.1 has significantly high unsuccessful rate while FT has the highest rate of success.



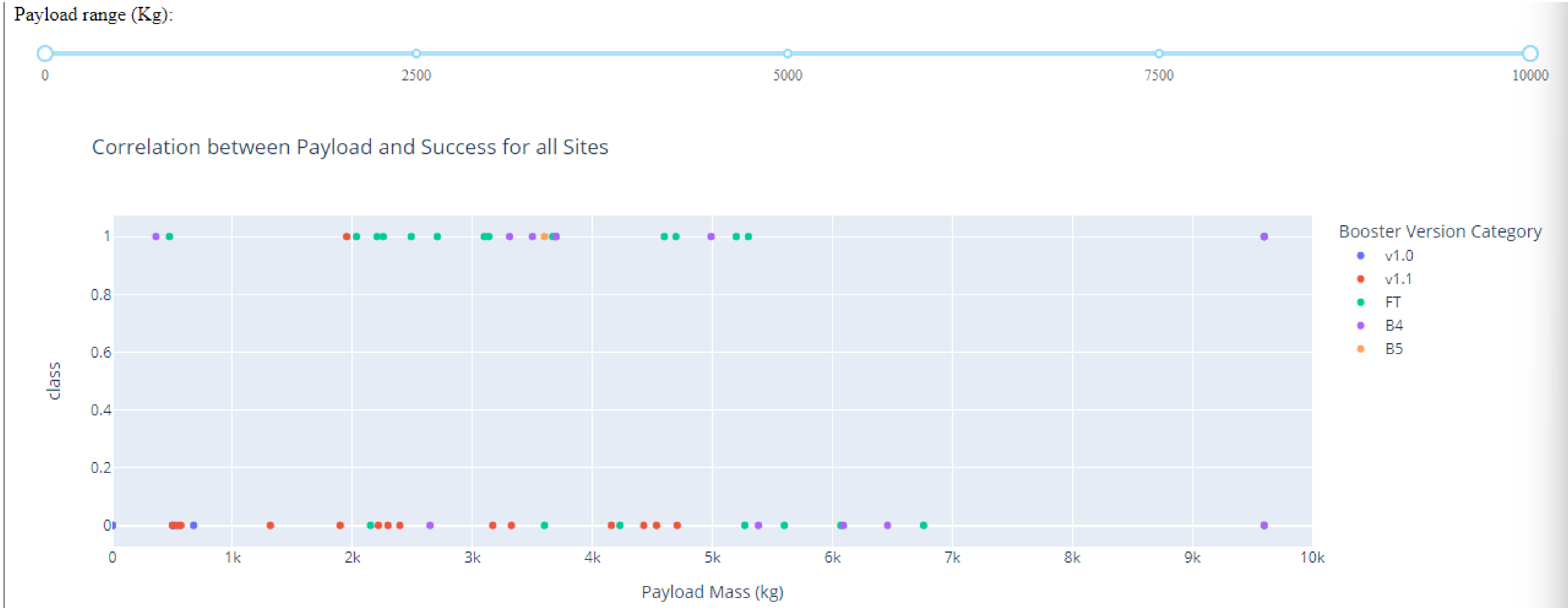
Payload vs Launch Outcome with Range Slider (2)

For payload range over 5000 kg, Booster FT has lower success rate and along with B4, these two boosters show that the higher payload lower the success rate.



Payload vs Launch Outcome with Range Slider (3)

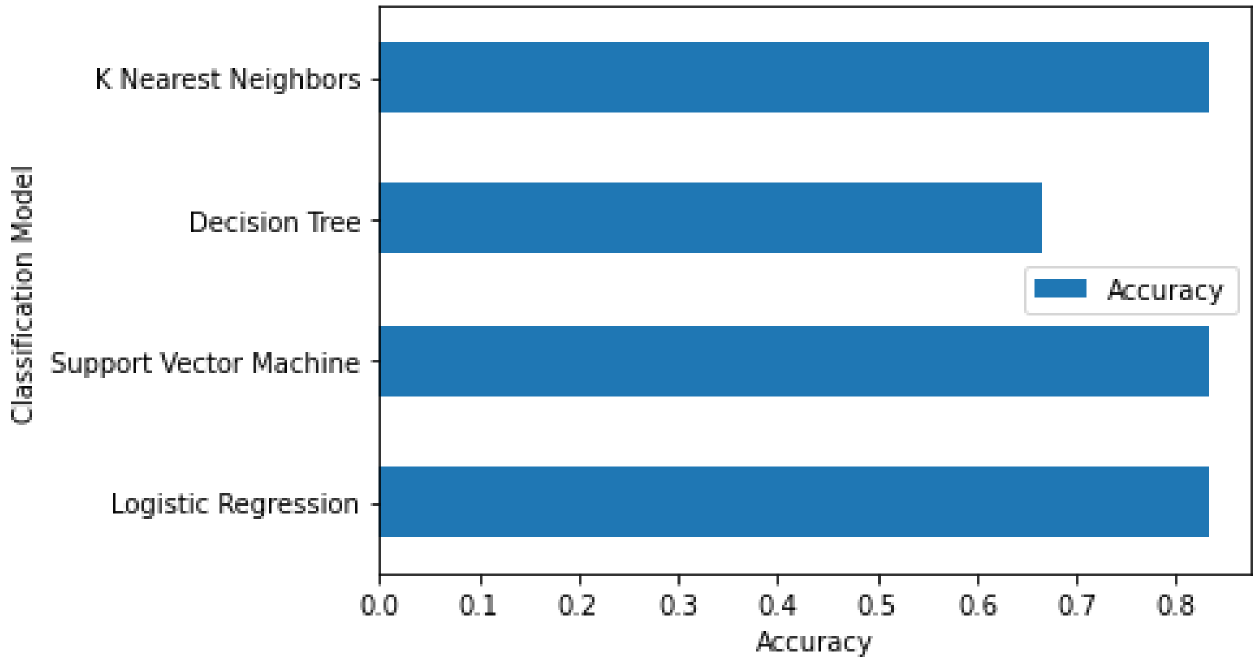
For all payload range, FT Booster has clear correlation between success rate and payload mass as it dominates the success rate of payload below 5000kg. B4 Booster show no clear correlation. V1.1 has the lowest success rate amongst all.



Predictive analysis (Classification)

Classification Accuracy

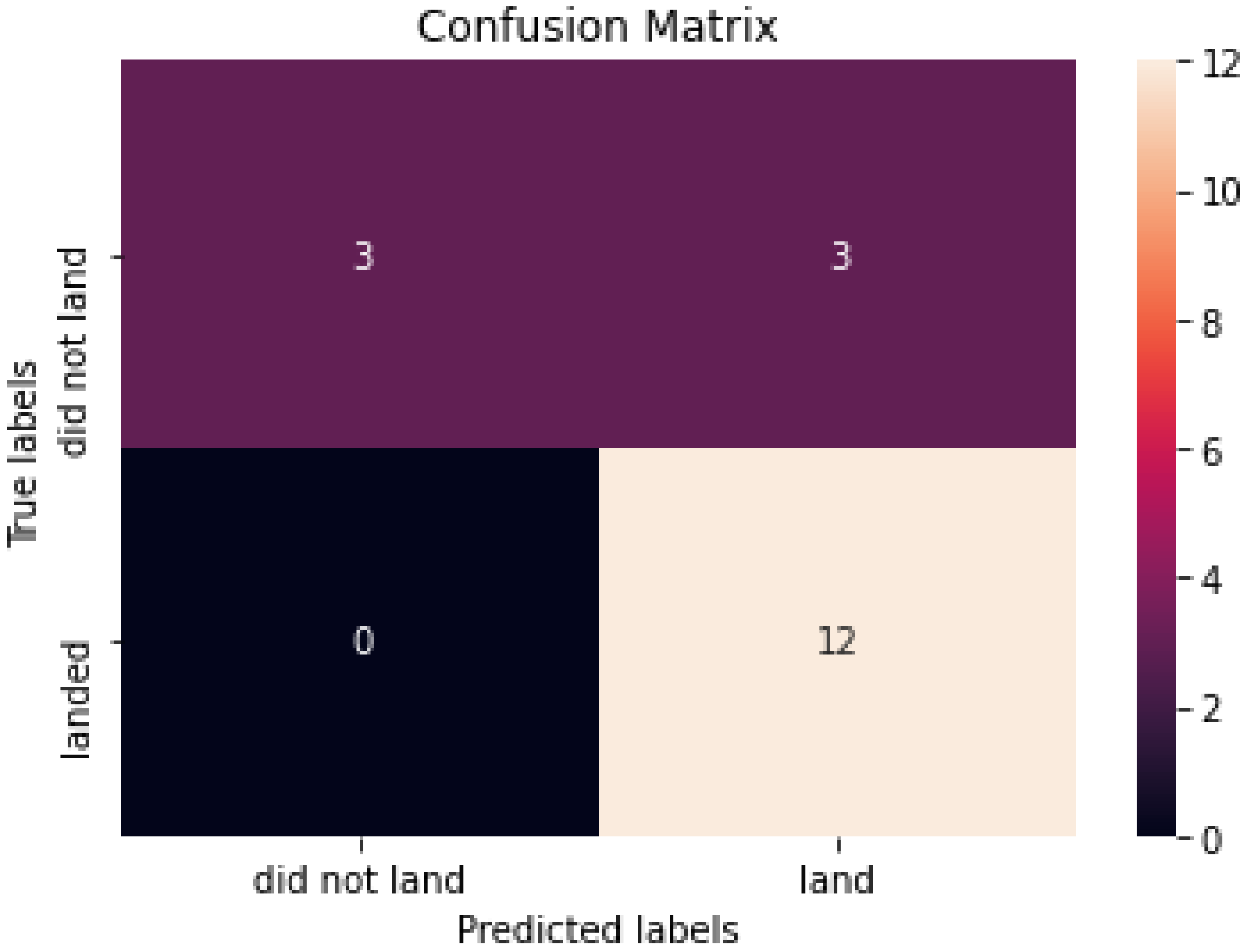
All model except for Decision tree has around 83% accuracy, which is a high accuracy.



Confusion Matrix

All models other than Decision Tree has the same confusion matrix as follow.

The consideration point here is that there are 17% of false positive and 0% of false negative, therefore the outcome of this predictive model should be carefully assessed using other methods other than machine learning only as this percentage of false positive may cost 100 million per launch.



CONCLUSION

- The best models for this dataset are: kNN Model, Logistics Regression Model and SVM with sigmoid engine.
- The best accuracy is 83% equally amongst above mentioned models.
- Techniques other than machine learning must be taken into consideration since the model has relatively high rate of false positive which will cost a bunch of money.

APPENDIX: References

GitHub URL:

GitHub Repo

SpaceX data

Wikipedia

IBM Data Scientist Professional Certificate powered by Coursera

THANK YOU!