IBM Data Science Capstone Project

Prathik Ramachandran
18/01/2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Summary of methodologies
    - Data Collection via API, Web Scraping
    - Exploratory Data Analysis (EDA) with Data Visualization
    - EDA with SQL
    - Interactive Map with Folium
    - Dashboards with Plotly Dash
    - Predictive Analysis
- Summary of all results
    - Exploratory Data Analysis results
    - Interactive maps and dashboard
    - Predictive results

INTRODUCTION

Background : SpaceX a rocket company launches satellites at low price like 70% less than their competitor since they land their satellites for reusing them to launch . The aim of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The price difference is explained by the fact that SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch.

- Problems you want to find answers
  - What are the main characteristics of a successful or failed landing ?
  - What are the effects of each relationship of the rocket variables on the success or failure of a landing ?
  - What are the conditions which will allow SpaceX to achieve the best landing success rate ?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

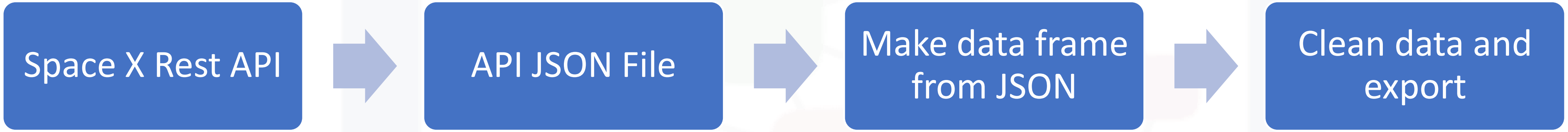
METHODOLOGY

- Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization andSQL
- Perform interactive visual analytics using Folium andPlotlyDash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

DATA COLLECTION

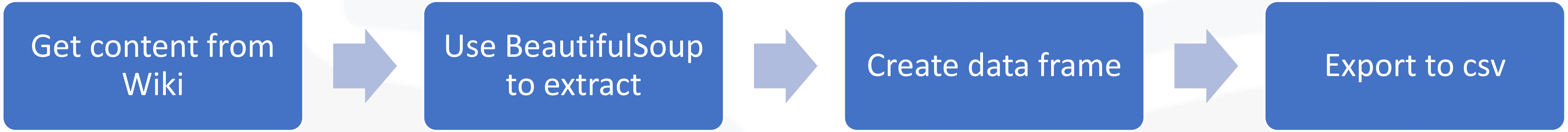
Datasets are collected from Rest SpaceX API and web scrapping Wikipedia



Space X Rest API → API JSON File → Make data frame from JSON → Clean data and export

Link to Data Collection

The information obtained by the web scrapping of Wikipedia are launches, landing, payload information.
URL:
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Get content from Wiki → Use BeautifulSoup to extract → Create data frame → Export to csv

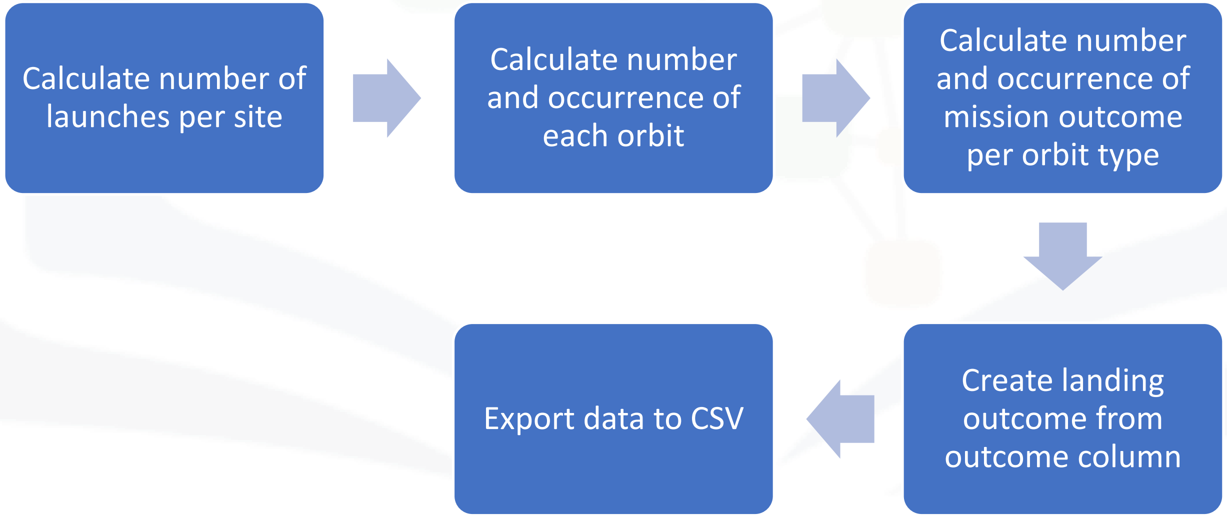
Link to Web Scrapping

DATA WRANGLING

- In the dataset, there are several cases where the booster did not land successfully.
  - •True Ocean, True RTLS, True ASDS means the mission has been successful.
  - •False Ocean, False RTLS, False ASDS means the mission was a failure.

•We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

Link to data Wrangling


Calculate number of launches per site
Calculate number and occurrence of each orbit
Calculate number and occurrence of mission outcome per orbit type
Create landing outcome from outcome column
Export data to CSV

EDA WITH DATA VISUALIZATION

Scatter plot

- Flight number & Launch Sites-Visualizing the launch from every site .
- Payload & Launch Sites-Payload launch from sites
- Success rate & Orbit type-Success rate compared to the orbit type
- Flight number & Orbit Type -Type of orbit for each launch
- Payload & Orbit type -Payload and the orbit .
- Trend of success rate-Trend of the success rate over the years .

Bar Graph
Success rate and Orbit

Line Graph
Success rate and Year

Link for EDA Data visualization

EDA WITH SQL

Performed the following SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 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 date 2010-06-04 and 2017-03-20 in descending order

[Link to EDA With SQL]()

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.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured 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 coloured 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 Interactive map to Folium

Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

- Link to Dashboard with Plotly

Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

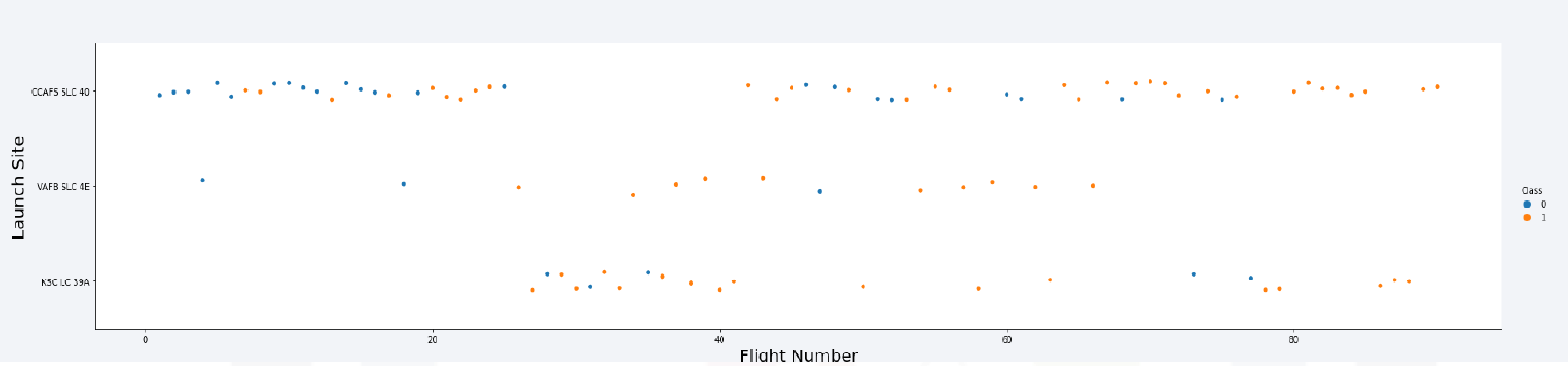
Link to Machine Learning Prediction

RESULTS

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

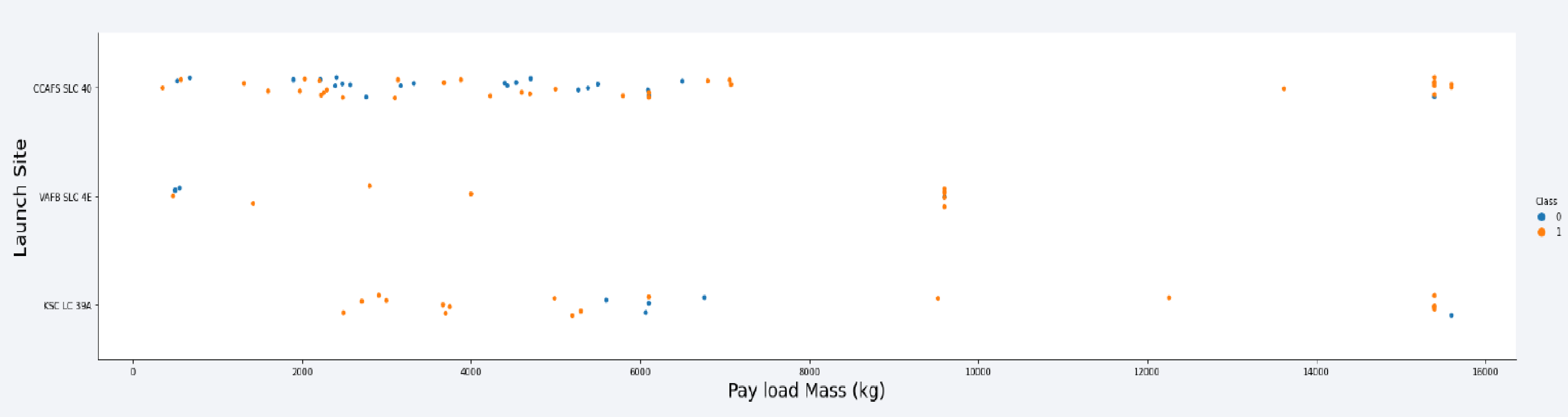
EDA With Visualization

Flight Number vs Launch Site



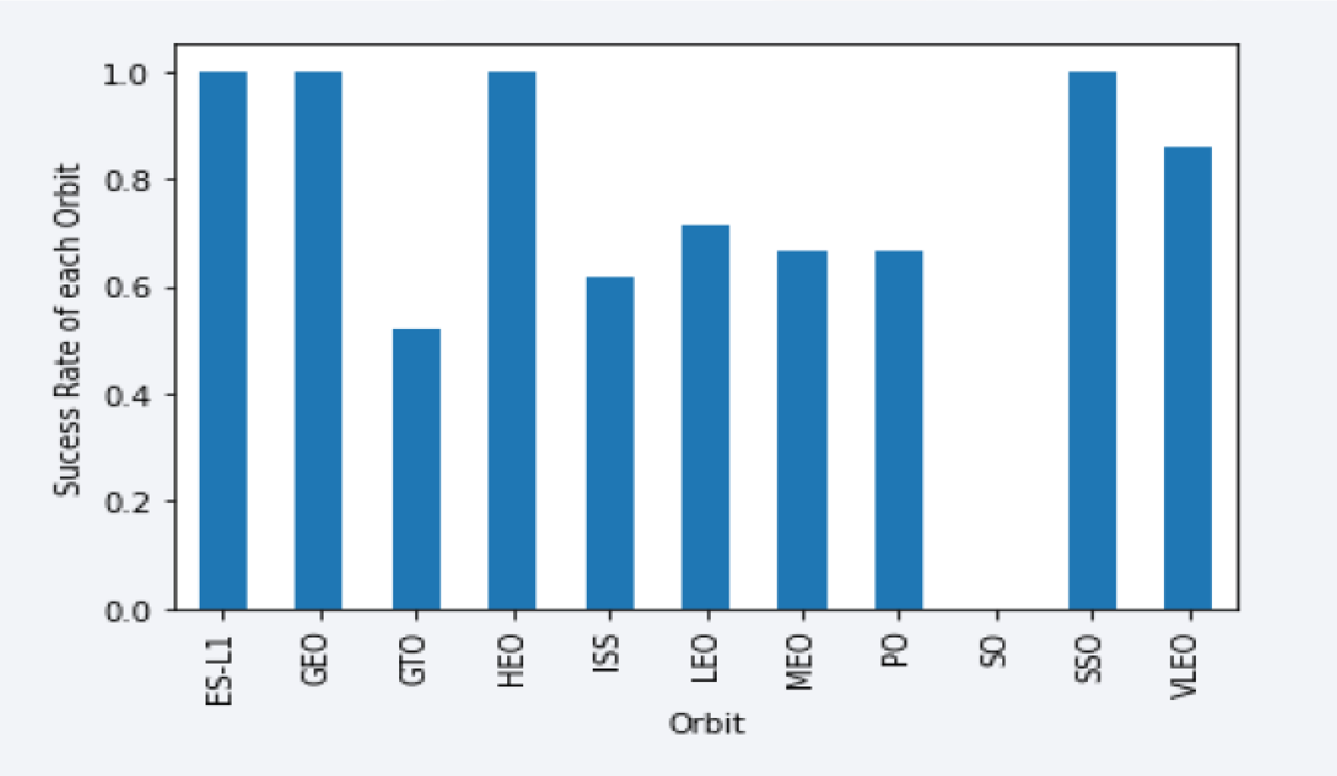
Per the graph above, we observe that for each site, the success rate is increasing.

Payload vs Launch Site



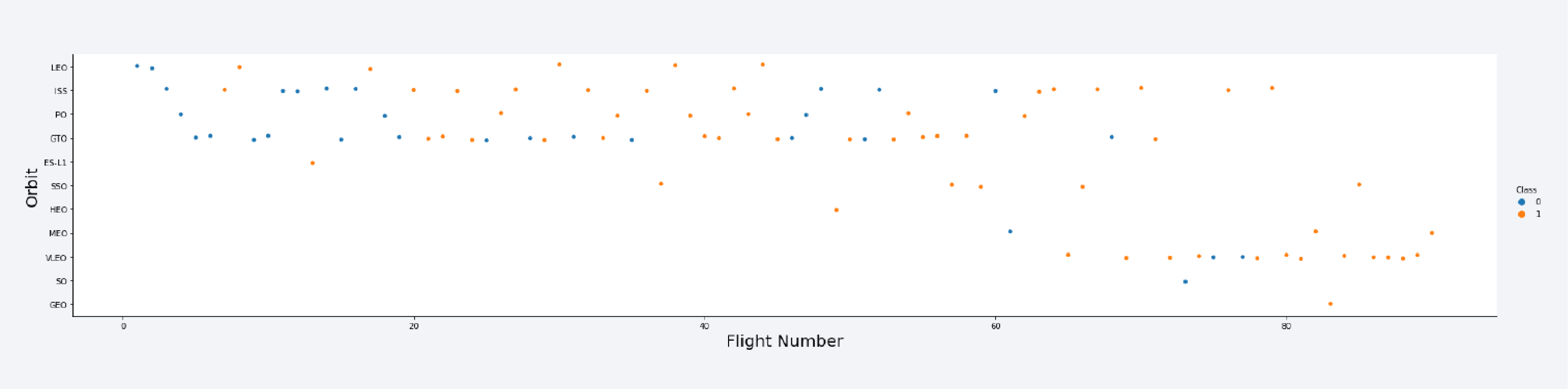
Per the graph above, we observe that a heavier payload may be a consideration for a successful landing. On the other hand, a too heavy payload can make a landing fail.

Success Rate vs Orbit Type



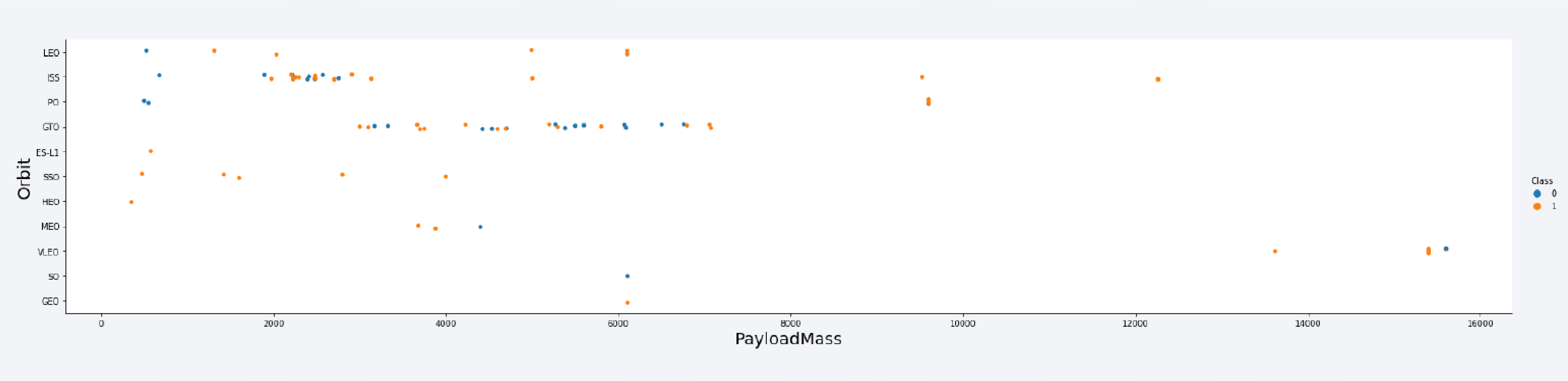
- Orbits with 100% success rate: - ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: - SO
- Orbits with success rate between 50% and 85%: - GTO, ISS, LEO, MEO, PO

Flight Number vs Orbit Type



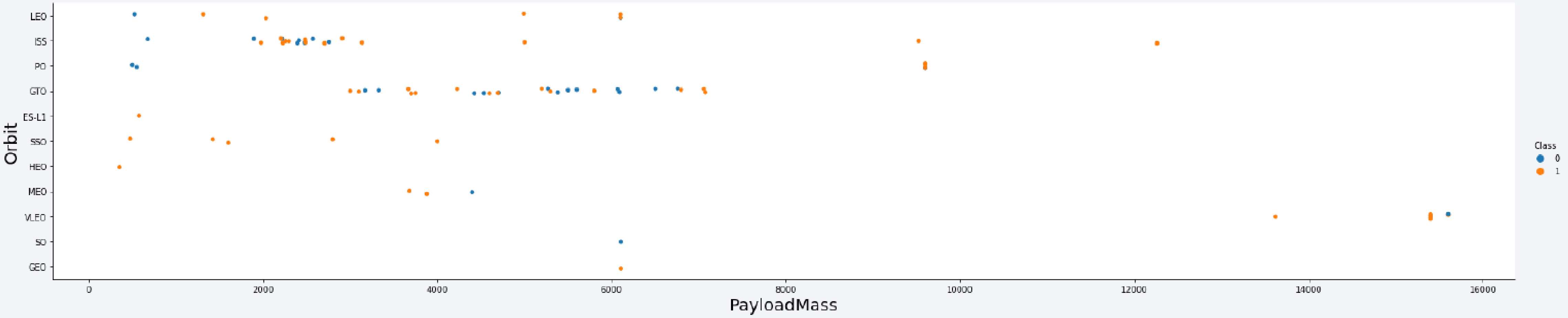
Per the graph above, we observe that for the success rate increases with the number of flights for the LEO orbit. For some orbits like GTO, there is no relation between the success rate and the number of flights. But we can suppose that the high success rate of some orbits like SSO or HEO is due to the knowledge learned during former launches for other orbits.

Payload vs Orbit Type



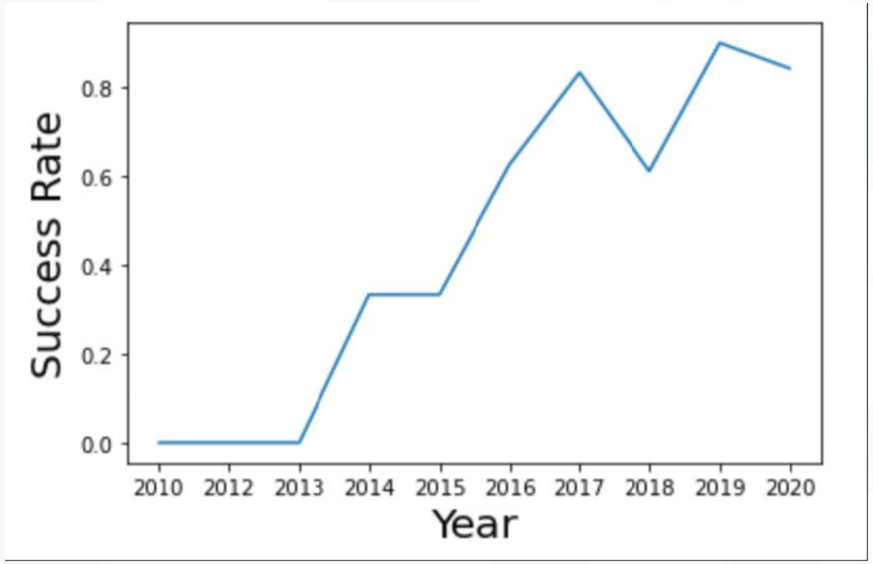
Per the graph above, we observe that heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Payload vs Orbit Type



Per the graph above, we observe that heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

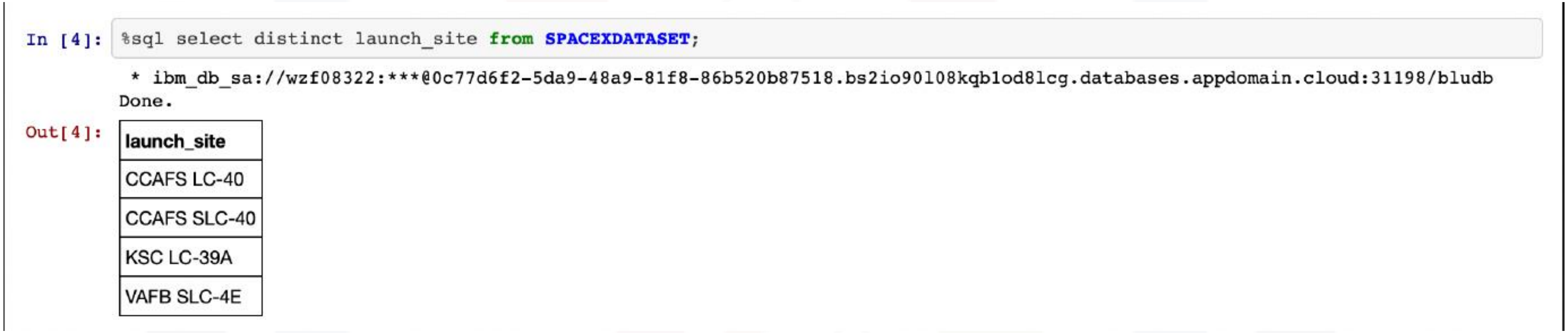
Launch Success Yearly Trend



Per the graph above, we observe that from 2013, the success rate has been increasing till 2020

EDA With SQL

All Launch Site Names



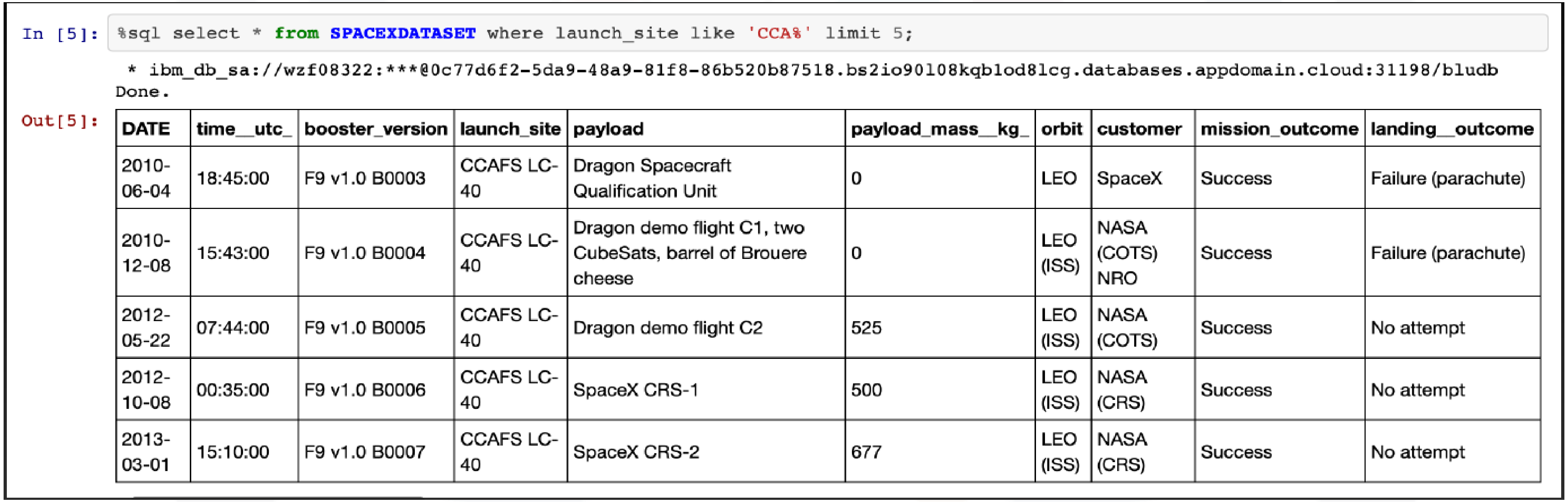
```
In [4]: %sql select distinct launch_site from SPACEXDATASET;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Explanation: Displaying the names of the unique launch sites in the space mission.

Launch Site Names begin with CCA



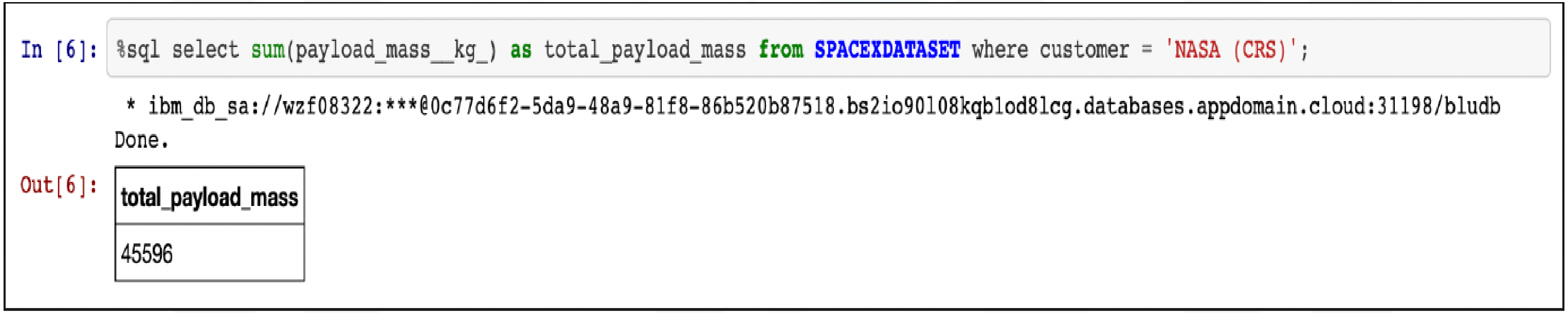
```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

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

Total Payload mass



```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

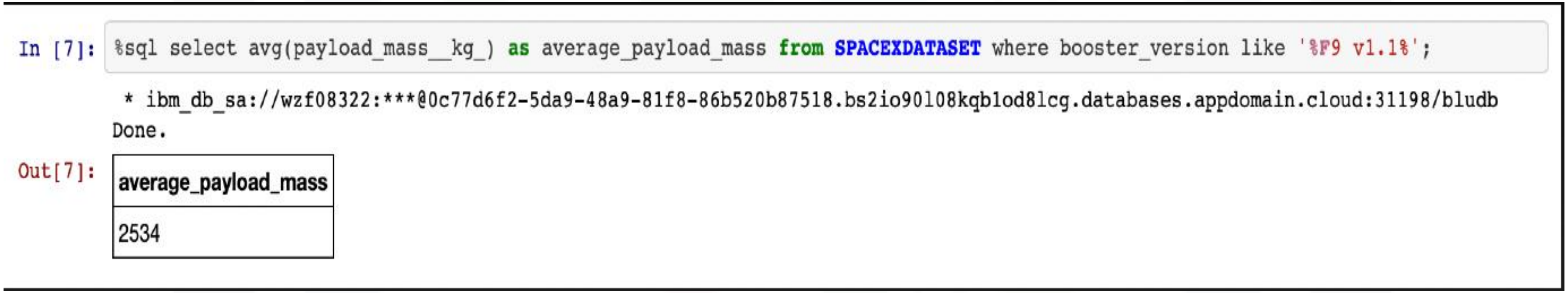
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

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

Average Payload Mass by F9 v1.1

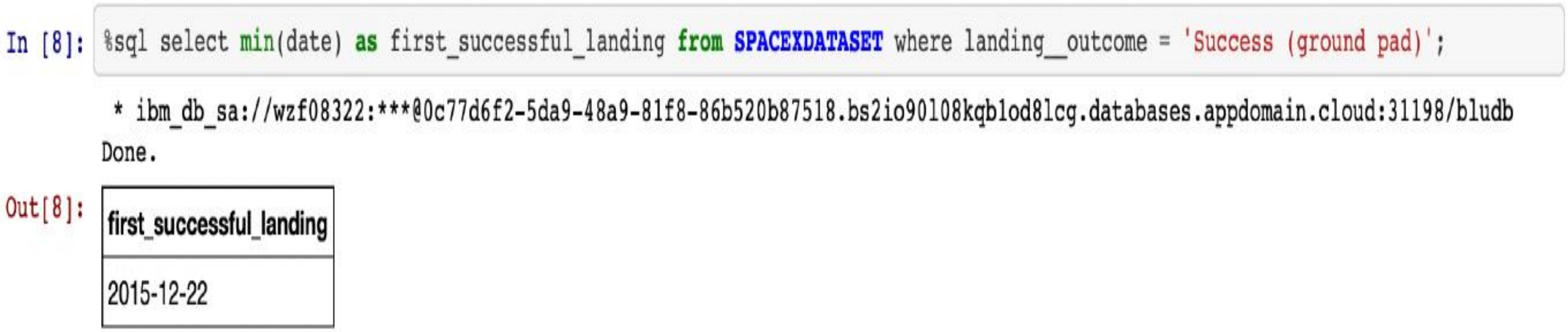


```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[7]:
```

average_payload_mass
2534

Explanation: This query returns the average of all payload masses where the booster version contains the substring F9 v1.1.

First Successful Ground Landing Date



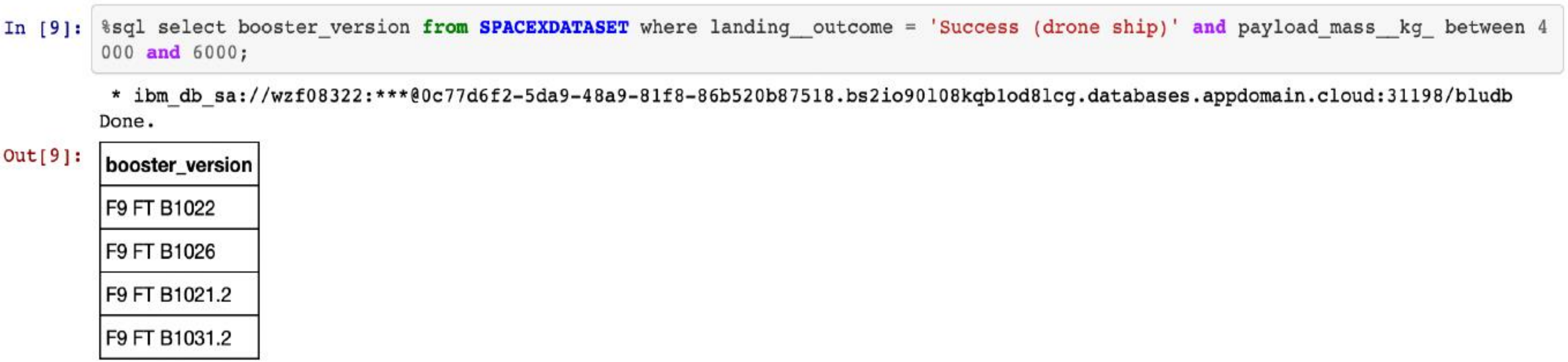
```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Explanation: Listing the date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

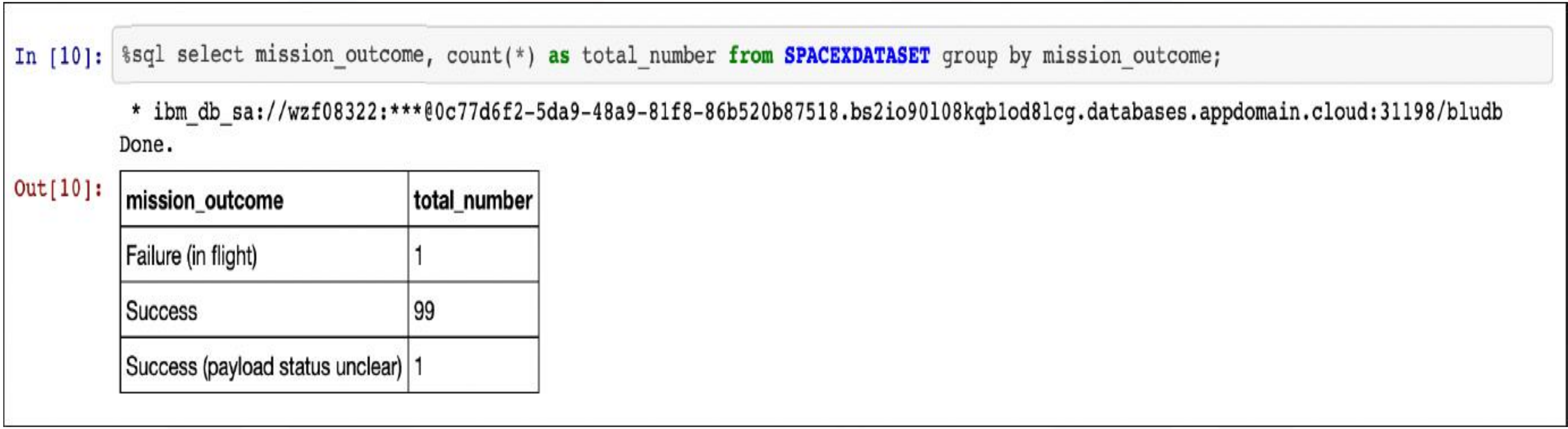


```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[9]:
```

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

Explanation: Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total number of successful and failure mission outcomes

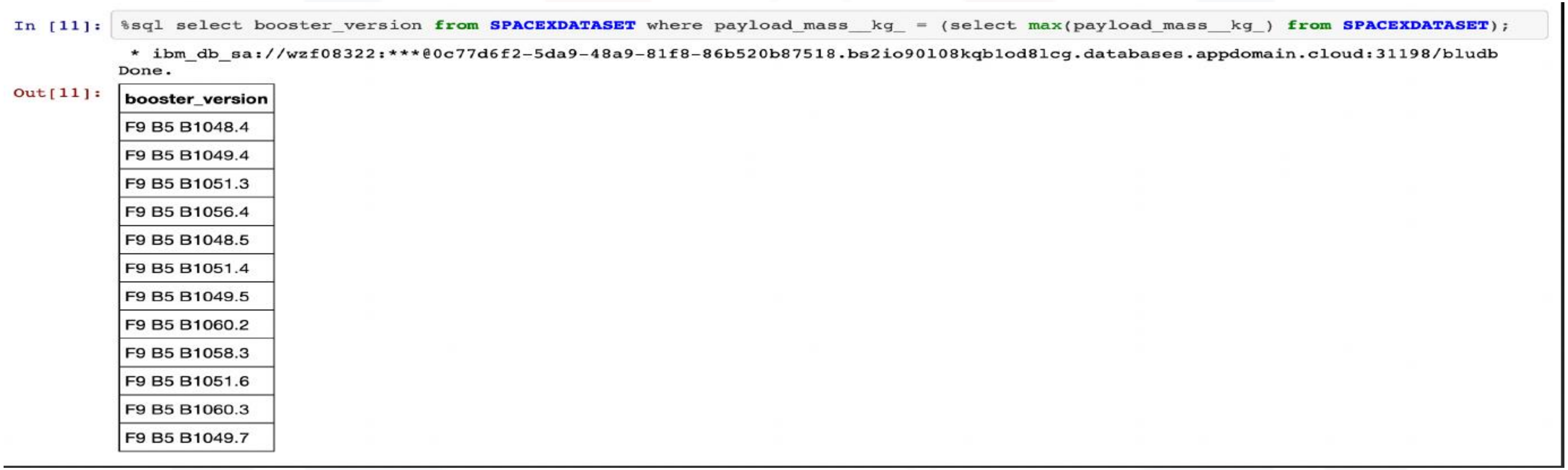


```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[10]:
```

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

Explanation:
With the first SELECT, we show the subqueries that return results. The first subquery counts the successful mission. T
The WHERE clause followed by LIKE clause filters mission outcome.
The COUNT function counts records filtered.

Boosters carried maximum payload

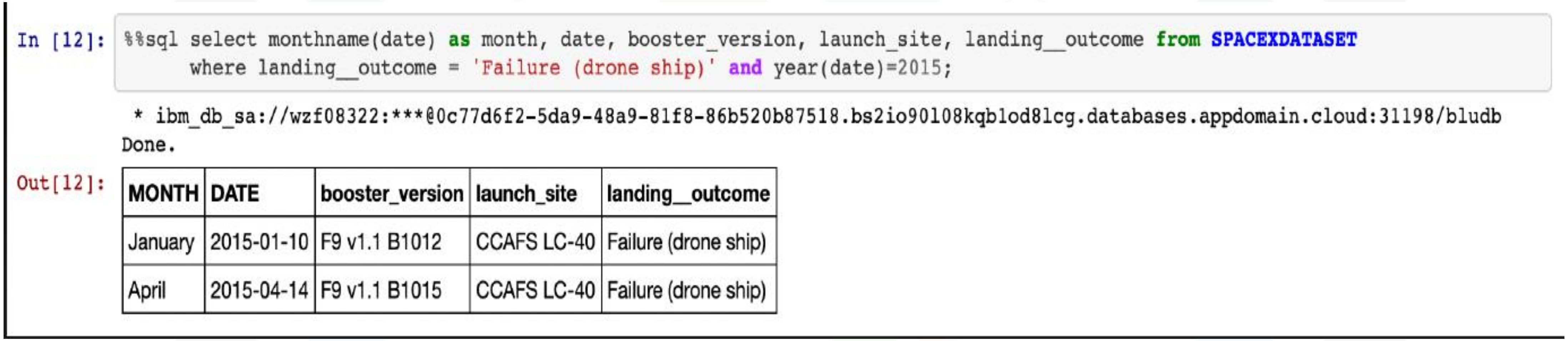


```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[11]:
```

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

Explanation:
Listing the names of the booster versions which have carried the maximum payload mass.

2015 launch records



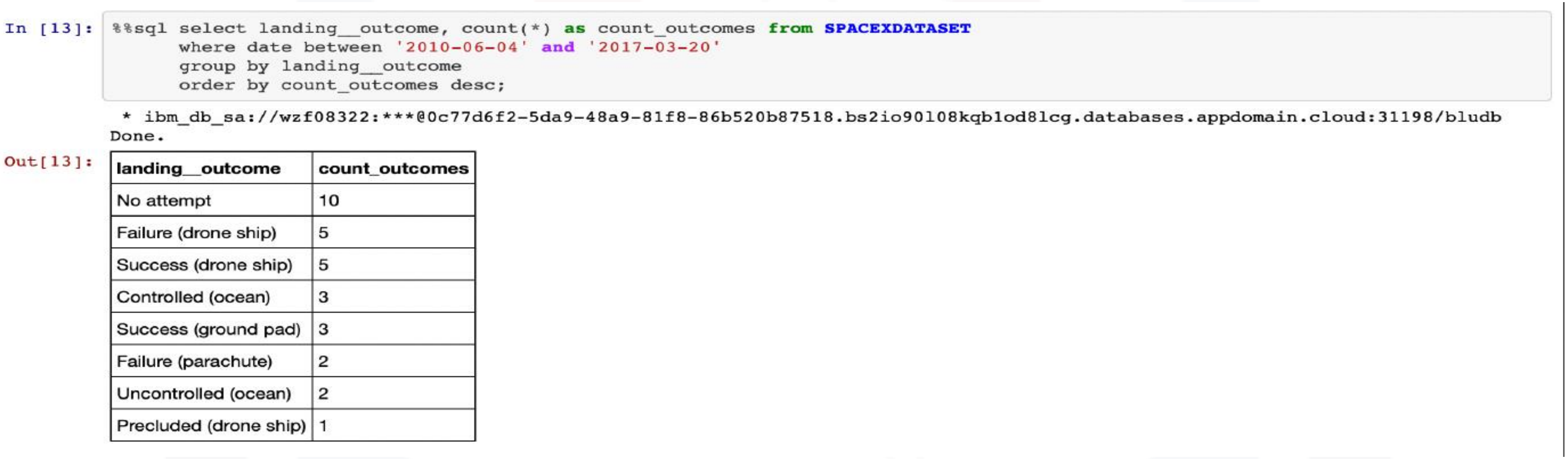
```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
               where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqblod8lcg.databases.appdomain.cloud:31198/bludb
         Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Explanation:
Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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



```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
                where date between '2010-06-04' and '2017-03-20'
                group by landing__outcome
                order by count_outcomes desc;

          * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
         Done.
Out[13]:
```

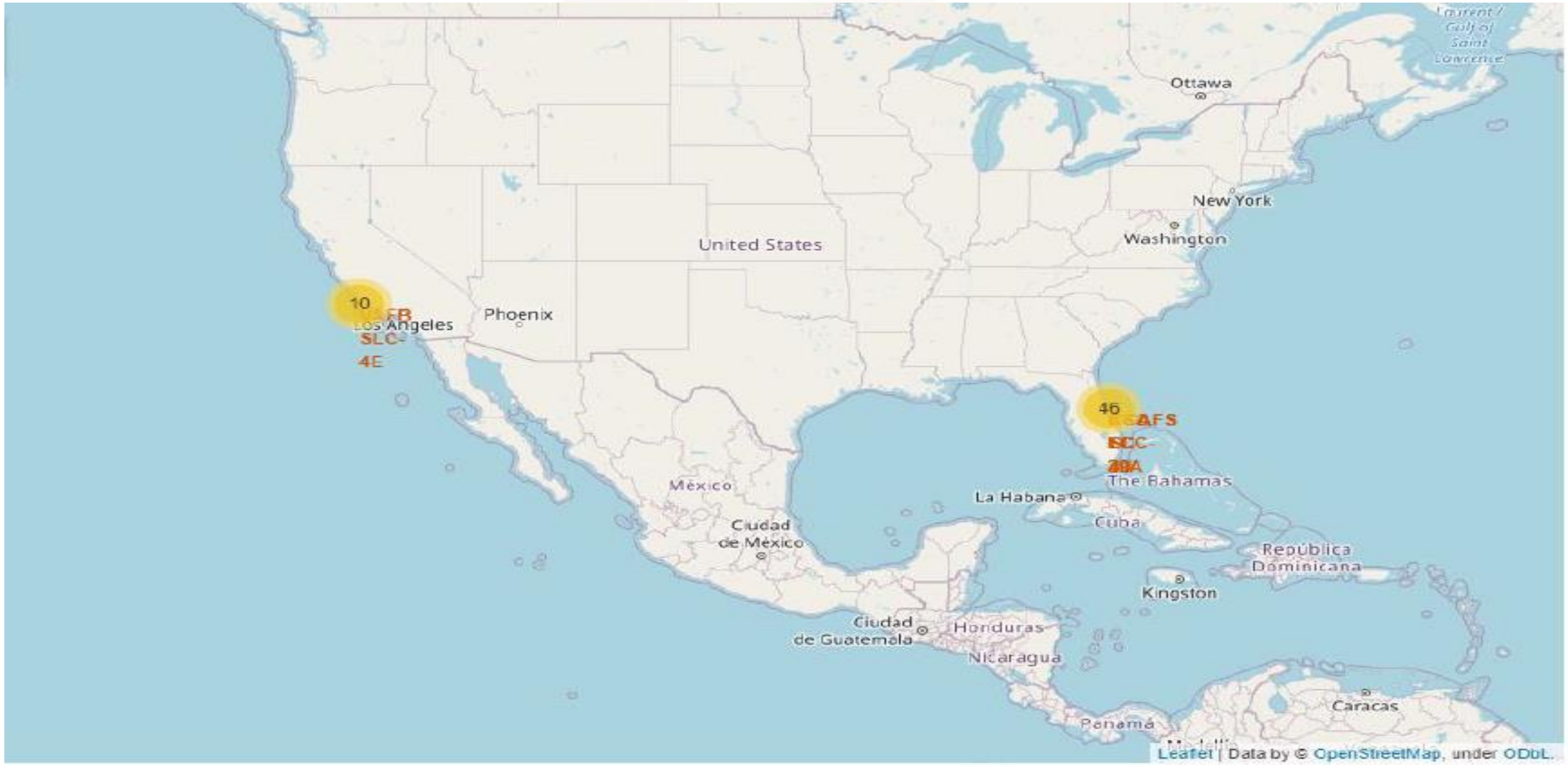
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:

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.

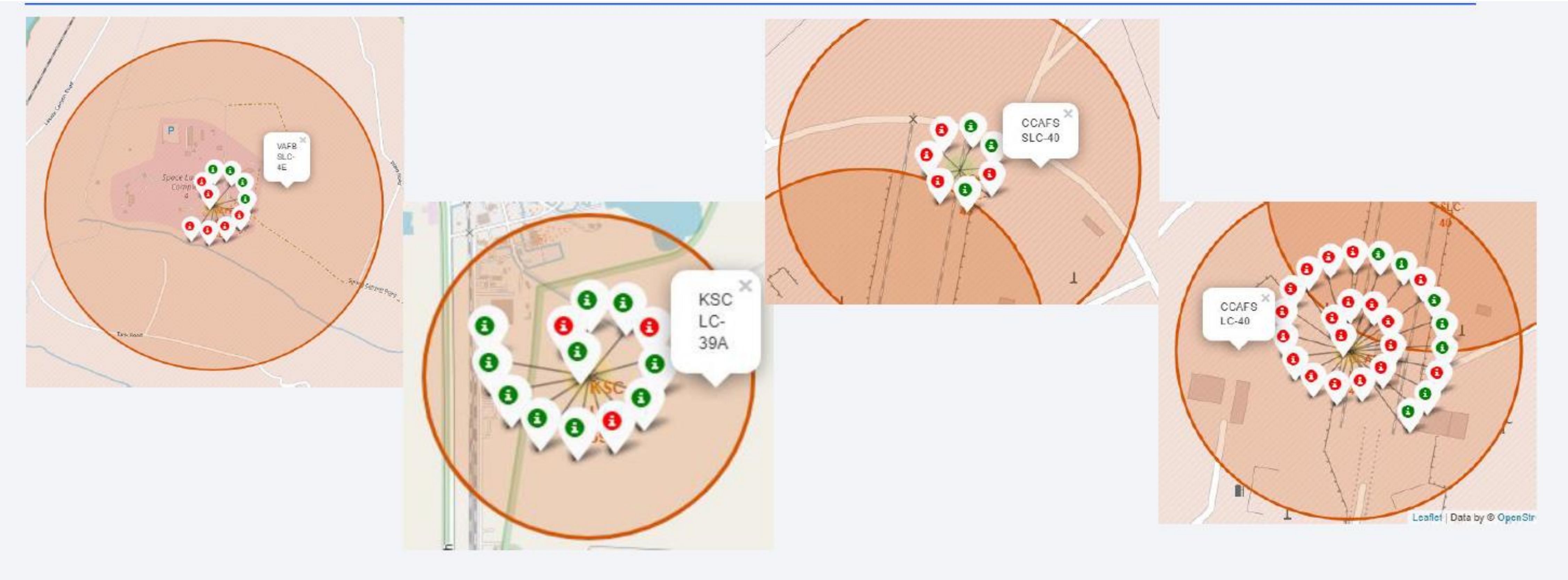
Interactive Map with Folium

All launch sites' location markers on a global map



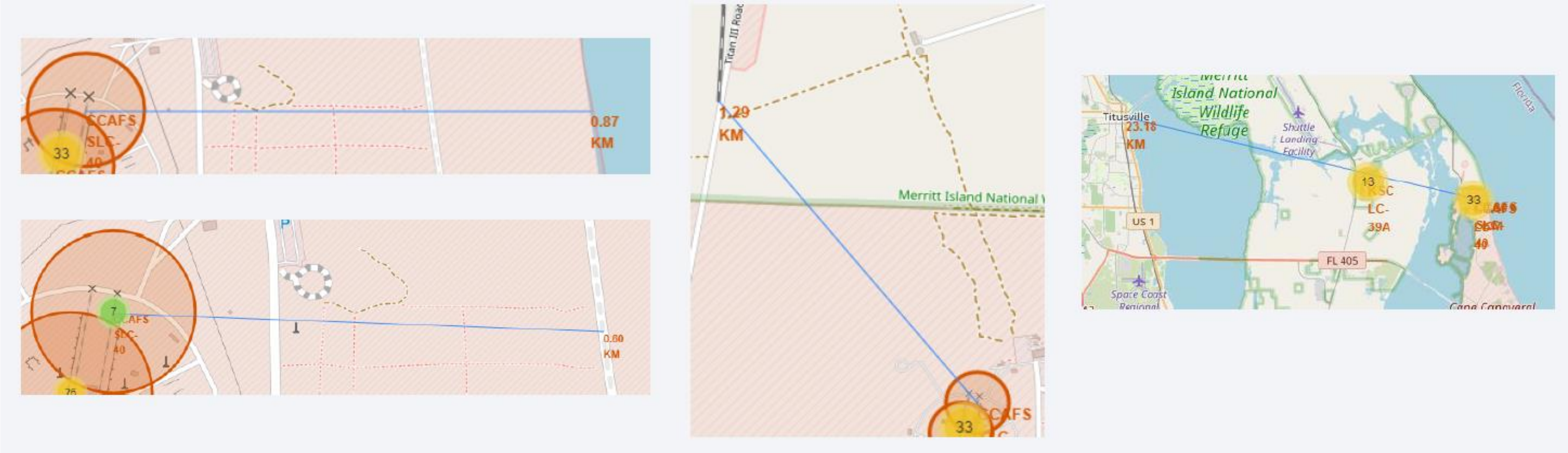
We see that Space X launch sites are located on the coast of the United States

Folium map –Color Labeled Markers



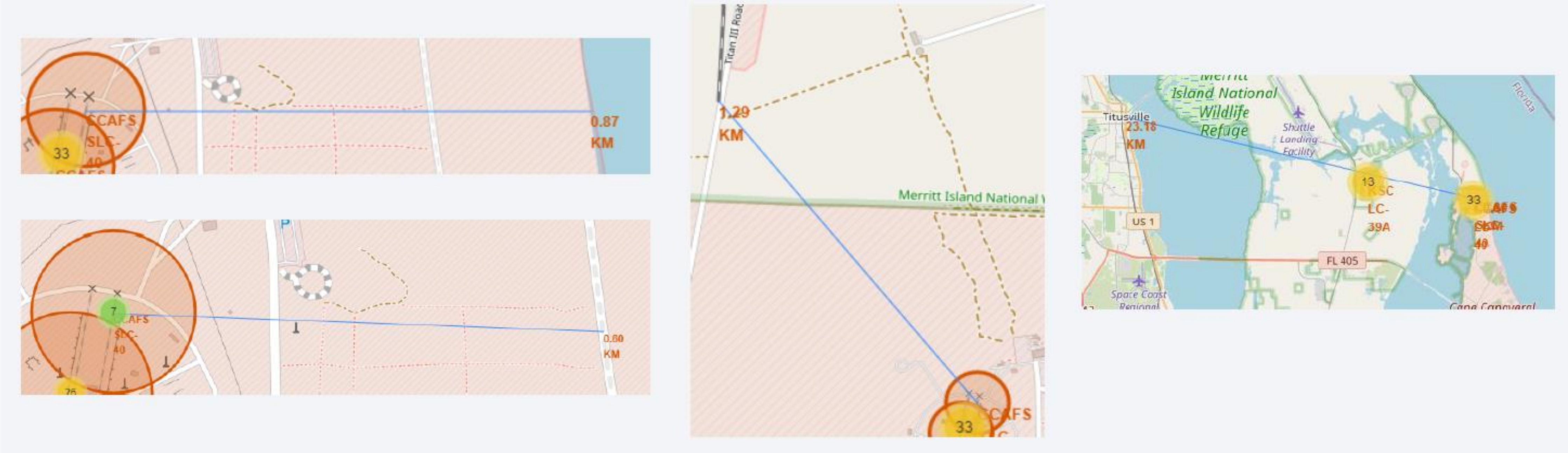
Greenmarker represents successful launches. Redmarker represents unsuccessful launches. We note that KSC LC-39A has a higher launch success rate.

Folium Map –Distances between CCAFS SLC-40 and its proximities



Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40in close proximity to highways ? Yes
Is CCAFS SLC-40in close proximity to coastline ? Yes
DoCCAFS SLC-40keeps certain distance away from cities ? No

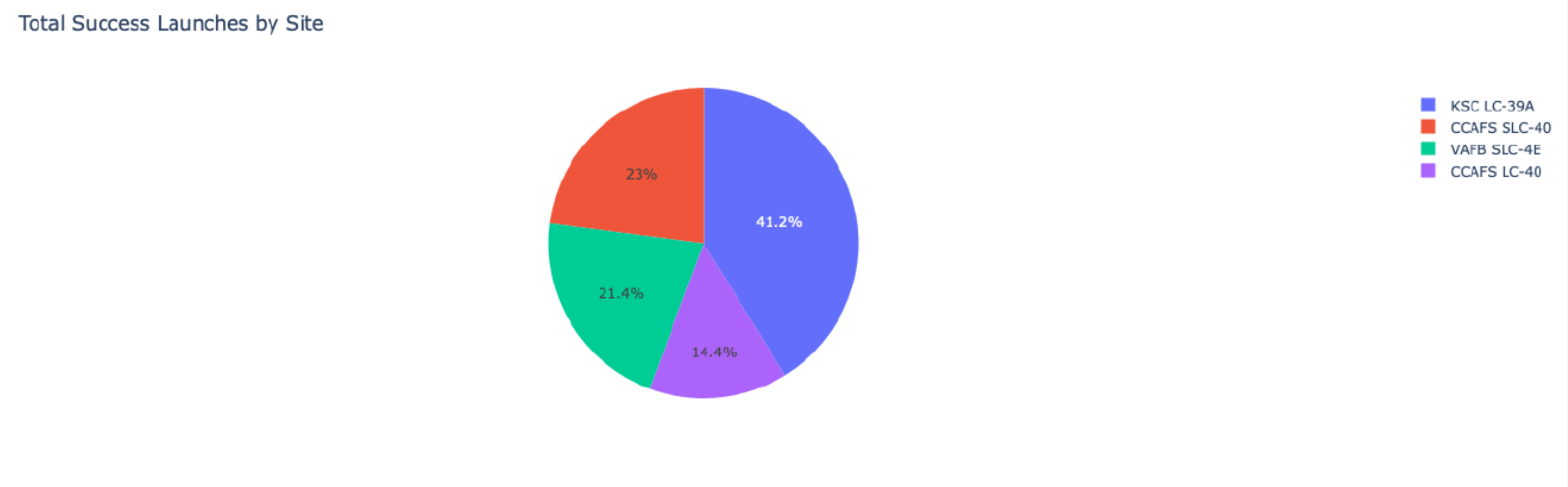
Folium Map –Distances between CCAFS SLC-40 and its proximities



Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40in close proximity to highways ? Yes
Is CCAFS SLC-40in close proximity to coastline ? Yes
DoCCAFS SLC-40keeps certain distance away from cities ? No

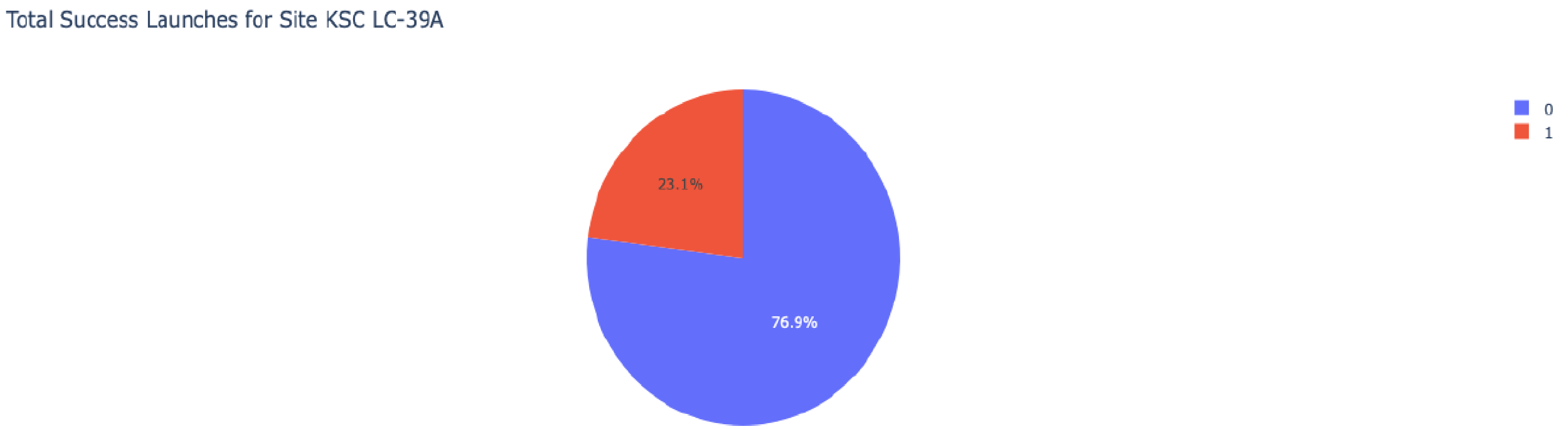
Build a Dashboard with Plotly Dash

Dashboard –Total success by Site



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

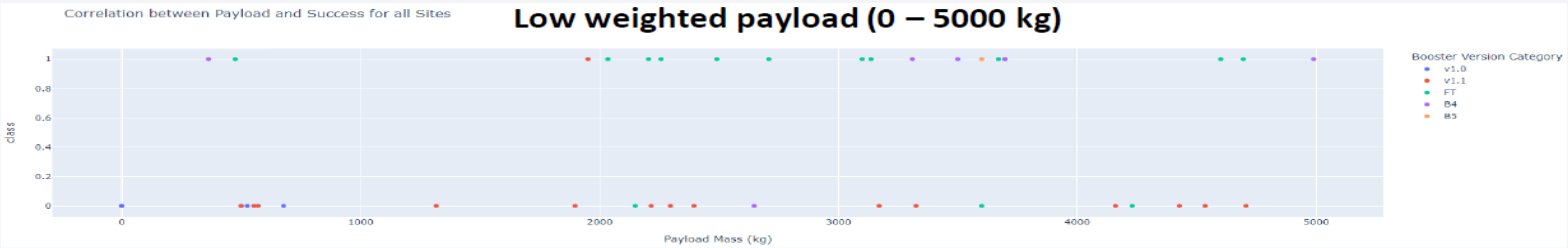
Launch site with highest launch success ratio



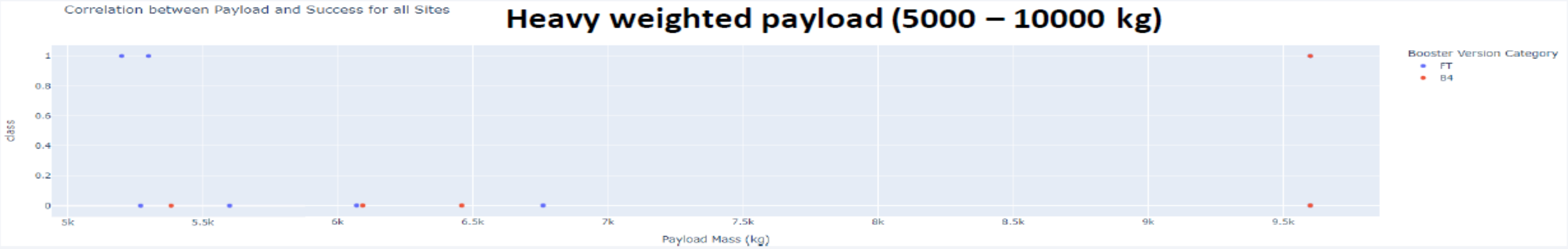
Total Success Launches for Site KSC LC-39A

KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Dashboard –Payload mass vs Outcome for all sites with different payload mass selected



Low weighted payload (0 – 5000 kg)

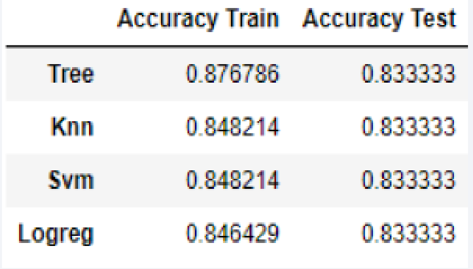


Heavy weighted payload (5000 – 10000 kg)

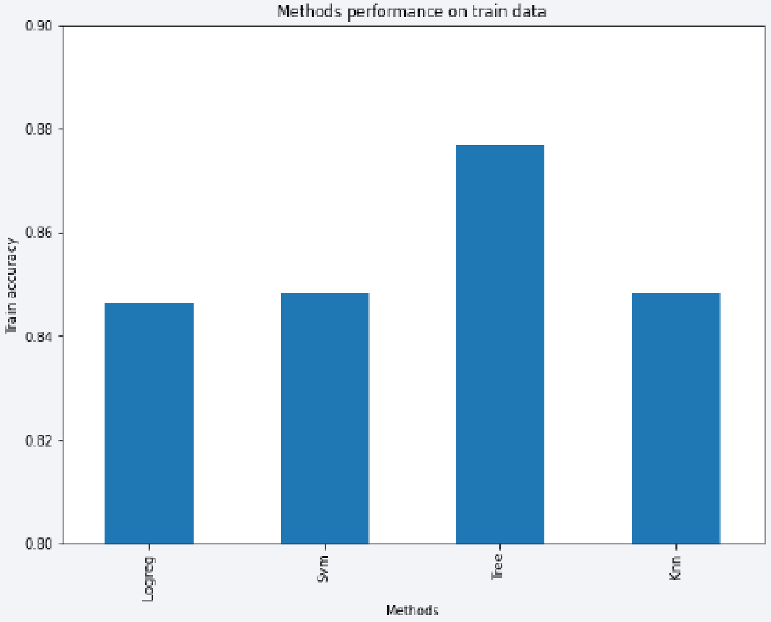
Low weighted payloads have a better success rate than the heavy weighted payloads.

Predictive Analysis (Classification)

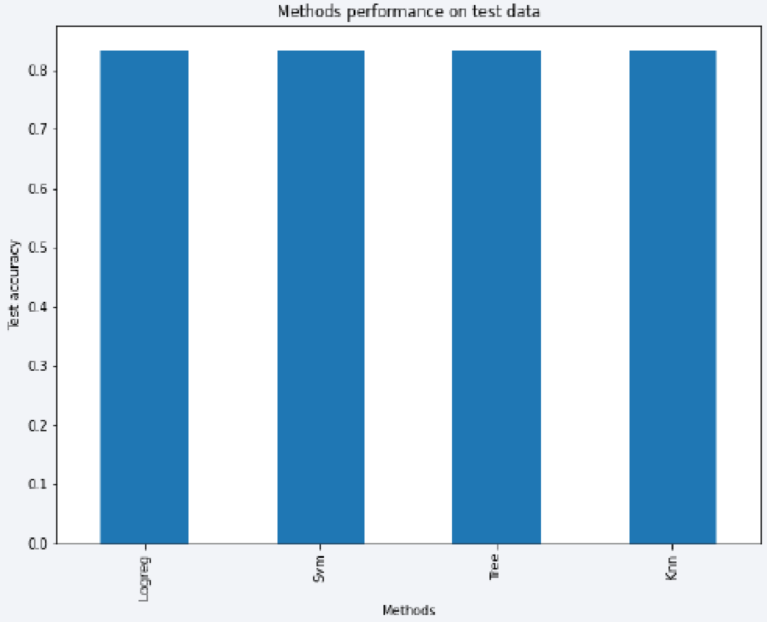
Classification Accuracy



	Accuracy Train	Accuracy Test
Tree	0.876786	0.833333
Knn	0.848214	0.833333
Svm	0.848214	0.833333
Logreg	0.846429	0.833333



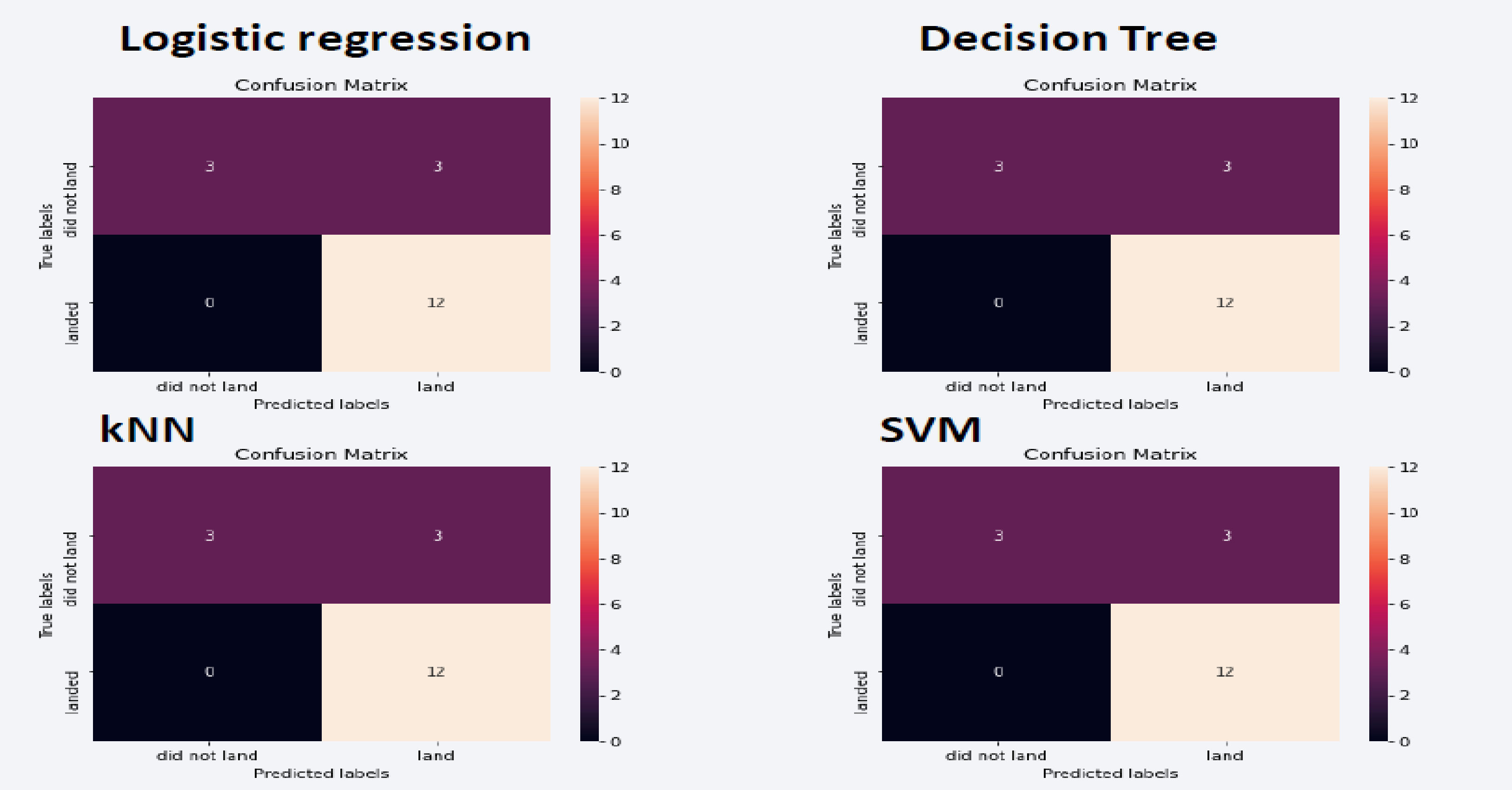
Methods performance on train data

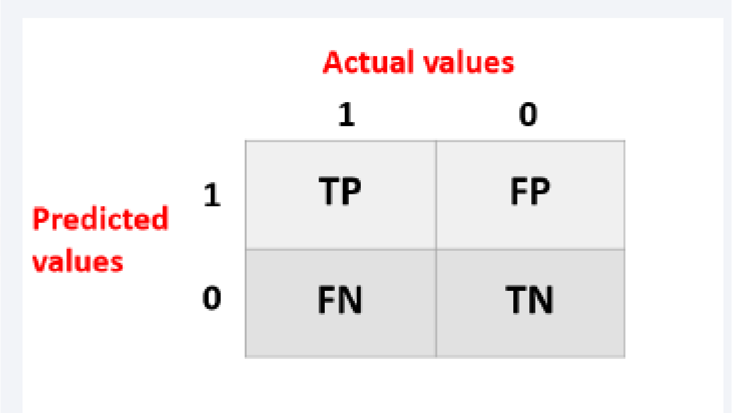


Methods performance on test data

For accuracy test, all methods performed similar. We could get more test data to decide between them. But if we really need to choose one right now, we would take the decision tree.

Confusion Matrix





		Actual values	
		1	0
Predicted values	1	TP	FP
	0	FN	TN

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

Conclusion

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.

- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.

- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.

- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.

- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.

THANK YOU