IBM Developer
SKILLS NETWORK

CAPSTONE PROJECT

OUTLINES:-

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

EXECUTIVE SUMMARY

Summary of methodologies:

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

INTRODUCTION

Project background and context:

SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

INTRODUCTION

Questions to be answered

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

- 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

1- DATA COLLECTION METHODOLOGY:
- USING SPACEX REST API
- USING WEB SCRAPPING FROM WIKIPEDIA

2- PERFORMED DATA WRANGLING
- FILTERING THE DATA
- DEALING WITH MISSING VALUES
- USING ONE HOT ENCODING TO PREPARE THE DATA TO A BINARY CLASSIFICATION

3- PERFORMED EXPLORATORY DATA ANALYSIS (EDA) USING VISUALIZATION AND SQL PERFORMED
4- INTERACTIVE VISUAL ANALYTICS USING FOLIUM AND PLOTLY DASH PERFORMED PREDICTIVE
5- ANALYSIS USING CLASSIFICATION MODELS

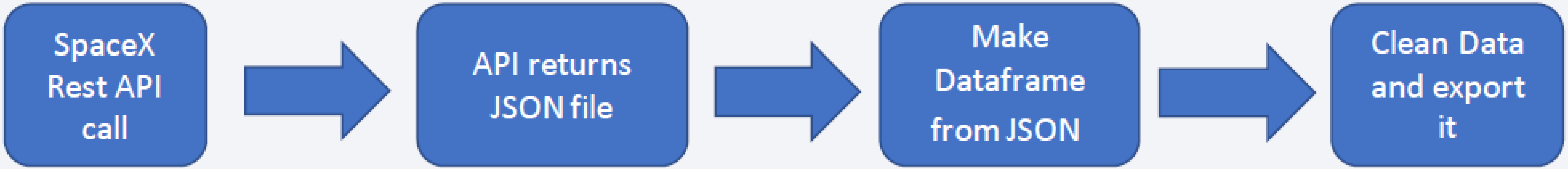
- BUILDING, TUNING AND EVALUATION OF CLASSIFICATION MODELS TO ENSURE THE BEST RESULTS

DATA COLLECTION

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX’s Wikipedia entry. We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis. Data Columns are obtained by using SpaceX REST API: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude Data Columns are obtained by using Wikipedia Web Scraping: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection

- Datasets are collected from Rest SpaceX API and webscrapping Wikipedia
  - The information obtained by the API are rocket, launches, payload information.
    - The Space X REST API URL is api.spacexdata.com/v4/



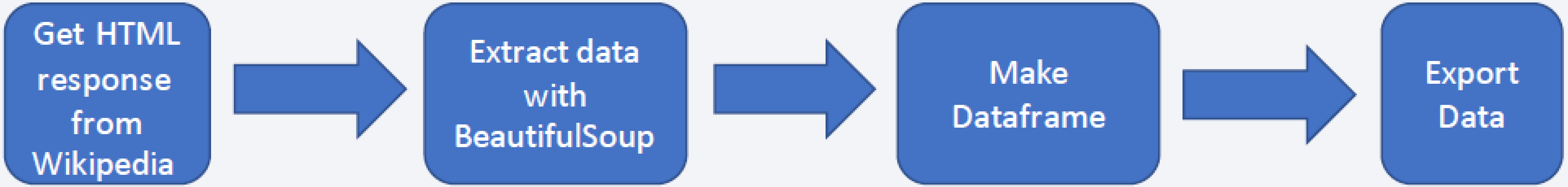
SpaceX Rest API call

API returns JSON file

Make Dataframe from JSON

Clean Data and export it

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



Get HTML response from Wikipedia

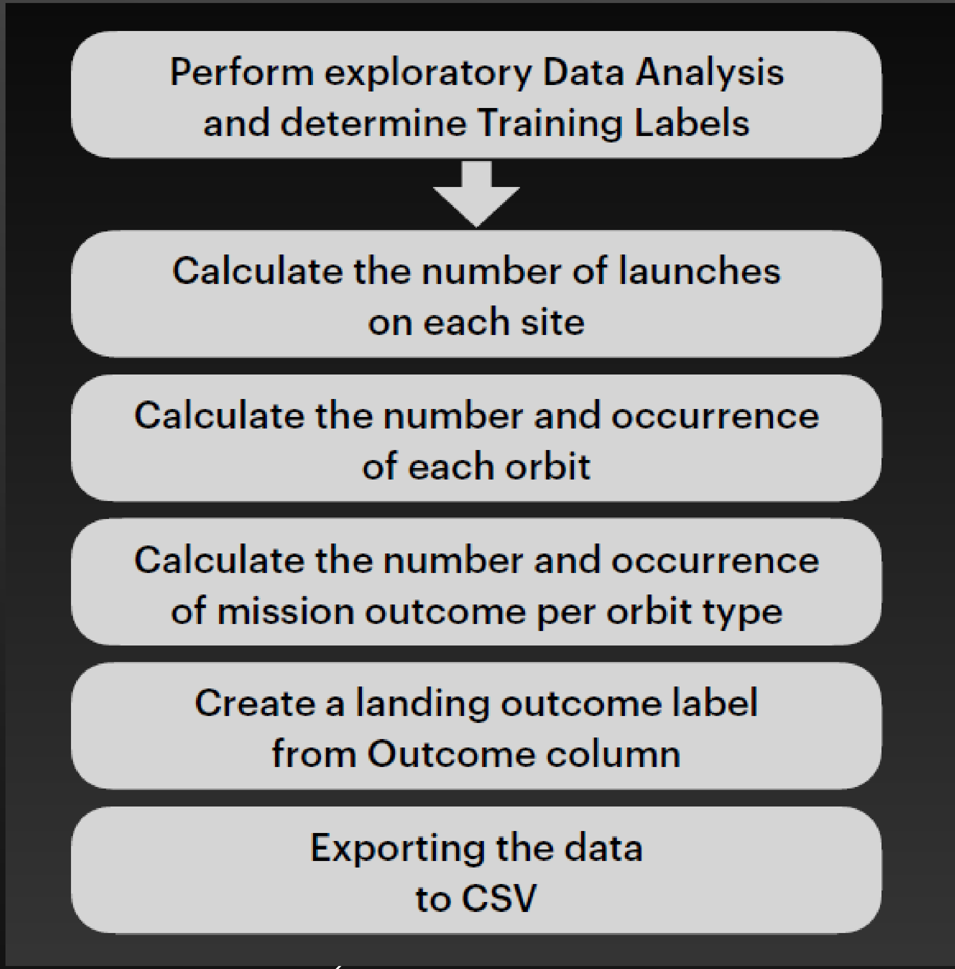
Extract data with BeautifulSoup

Make Dataframe

Export Data

DATA WRANGLING

IN THE DATA SET, THERE ARE SEVERAL DIFFERENT CASES WHERE THE BOOSTER DID NOT LAND SUCCESSFULLY. SOMETIMES A LANDING WAS ATTEMPTED BUT FAILED DUE TO AN ACCIDENT; FOR EXAMPLE, TRUE OCEAN MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED TO A SPECIFIC REGION OF THE OCEAN WHILE FALSE OCEAN MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A SPECIFIC REGION OF THE OCEAN. TRUE RTLS MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED TO A GROUND PAD FALSE RTLS MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED TO A GROUND PAD. TRUE ASDS MEANS THE MISSION OUTCOME WAS SUCCESSFULLY LANDED ON A DRONE SHIP FALSE ASDS MEANS THE MISSION OUTCOME WAS UNSUCCESSFULLY LANDED ON A DRONE SHIP. WE MAINLY CONVERT THOSE OUTCOMES INTO TRAINING LABELS WITH “1” MEANS THE BOOSTER SUCCESSFULLY LANDED, “0” MEANS IT WAS UNSUCCESSFUL.


Perform exploratory Data Analysis and determine Training Labels
Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate the number and occurrence of mission outcome per orbit type
Create a landing outcome label from Outcome column
Exporting the data to CSV

EDA WITH DATA VISUALIZATION

Scatter Graphs

- 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

This shows Correlation

Bar Graph

- Success rate vs. Orbit

This kind of graph shows the relationship between numeric and categoric variables.

Line Graph

- Success rate vs. Year

This kind of graphs shows data variables and their trends. Also, Line graphs can help to show global behavior and make prediction for unseen data.

EDA WITH SQL

PERFORMED 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

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.

BUILD A DASHBOARD WITH PLOTLY DASH

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

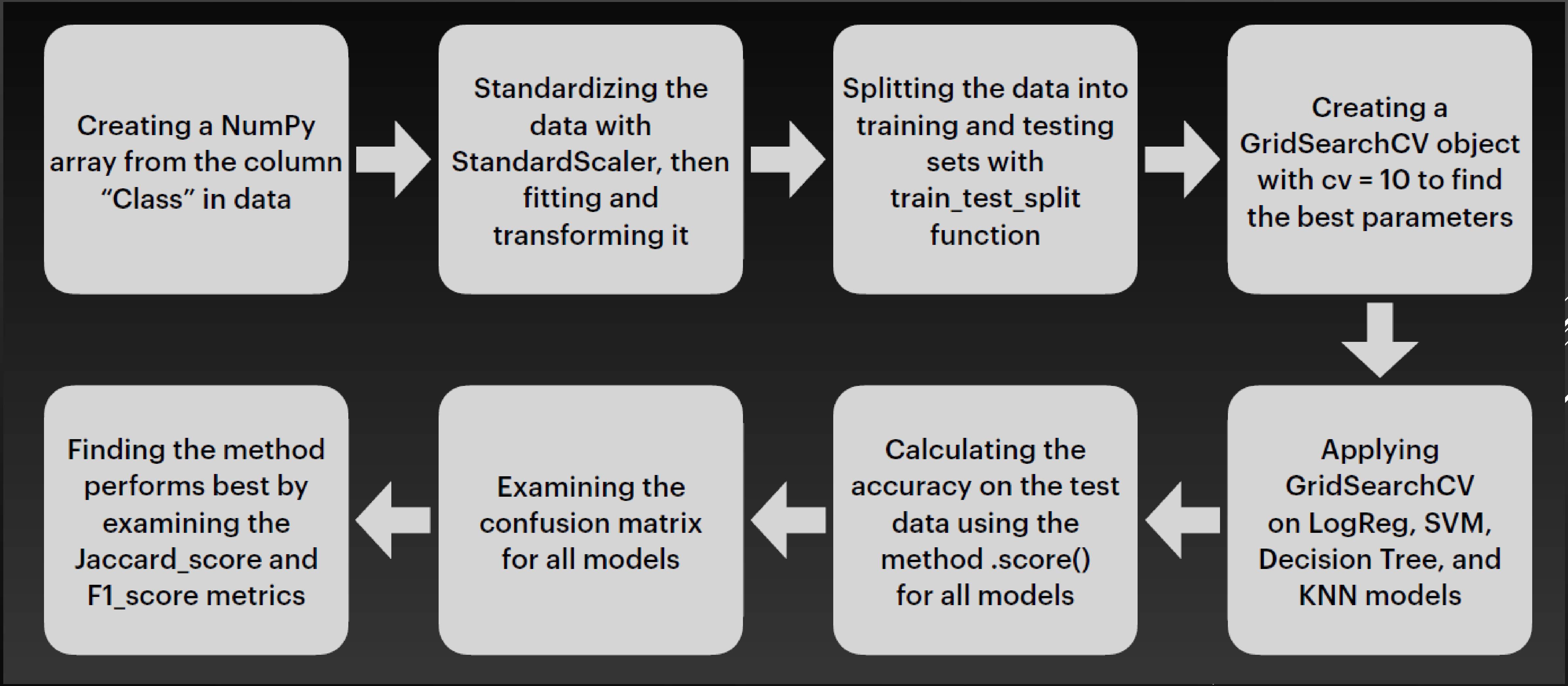
Dropdown allows a user to choose the launch site or all launch sites

Pie chart shows the total success and the total failure for the launch site chosen with the dropdown component.

Range slider allows a user to select a payload mass in a fixed range

Scatter chart shows the relationship between two variables, in particular Success vs Payload Mass

PREDICTIVE ANALYSIS (CLASSIFICATION)



Creating a NumPy array from the column "Class" in data

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Calculating the accuracy on the test data using the method .score() for all models

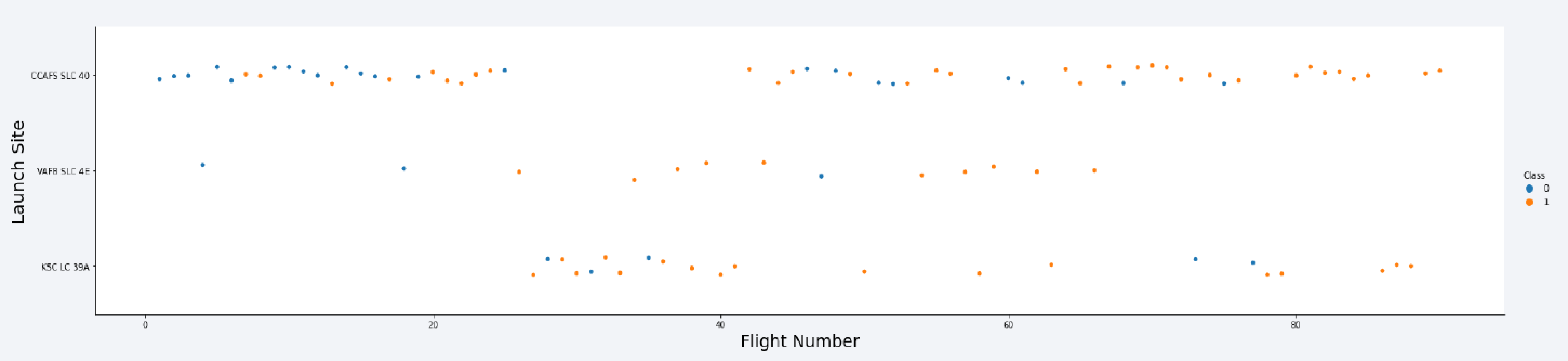
Examining the confusion matrix for all models

Finding the method performs best by examining the Jaccard_score and F1_score metrics

RESULTS

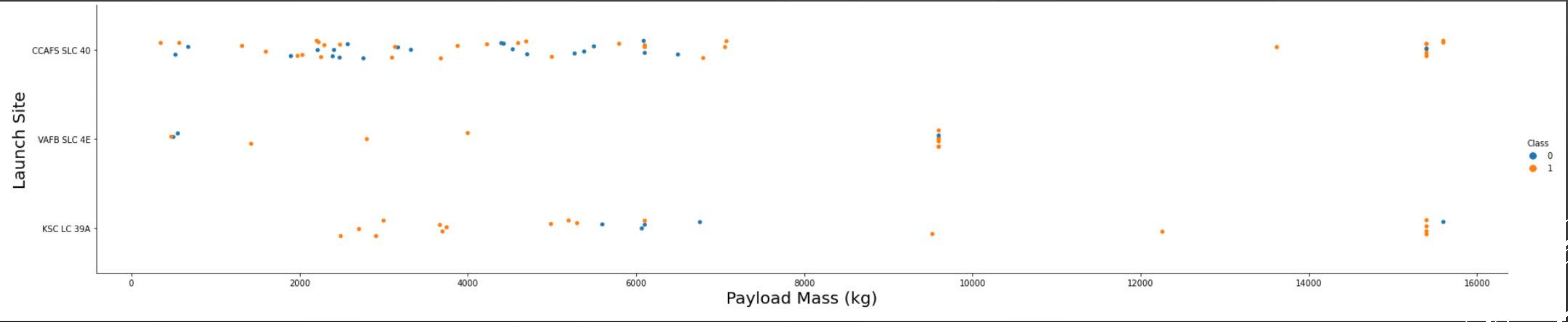
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

FLIGHT NUMBER VS. LAUNCH SITE



- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

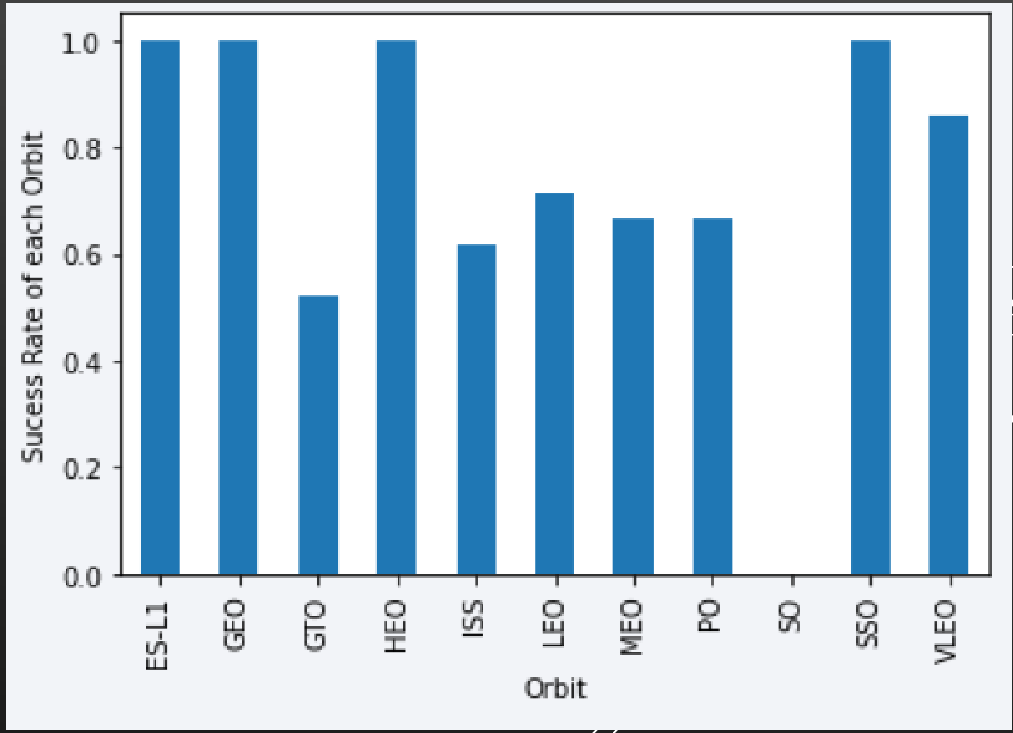
PAYLOAD VS. LAUNCH SITE



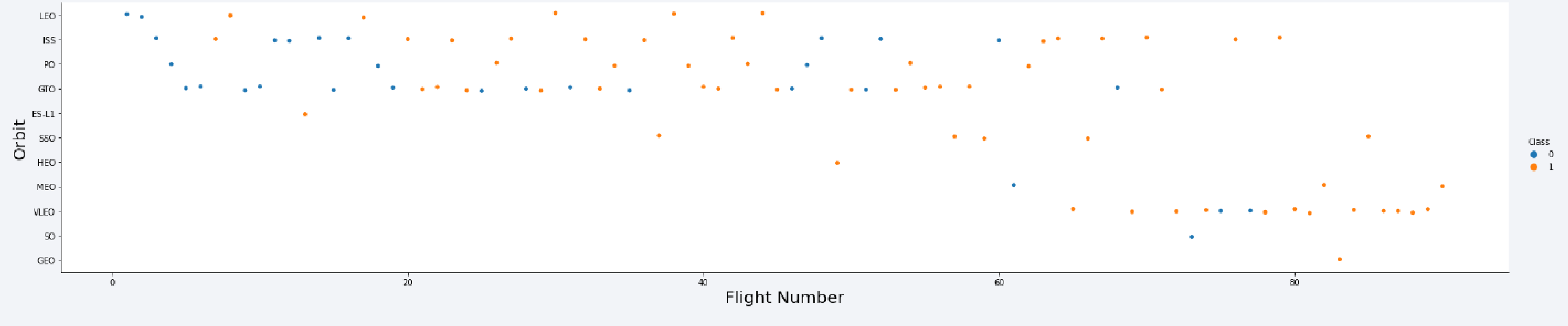
- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

SUCCESS RATE VS. ORBIT TYPE

- 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

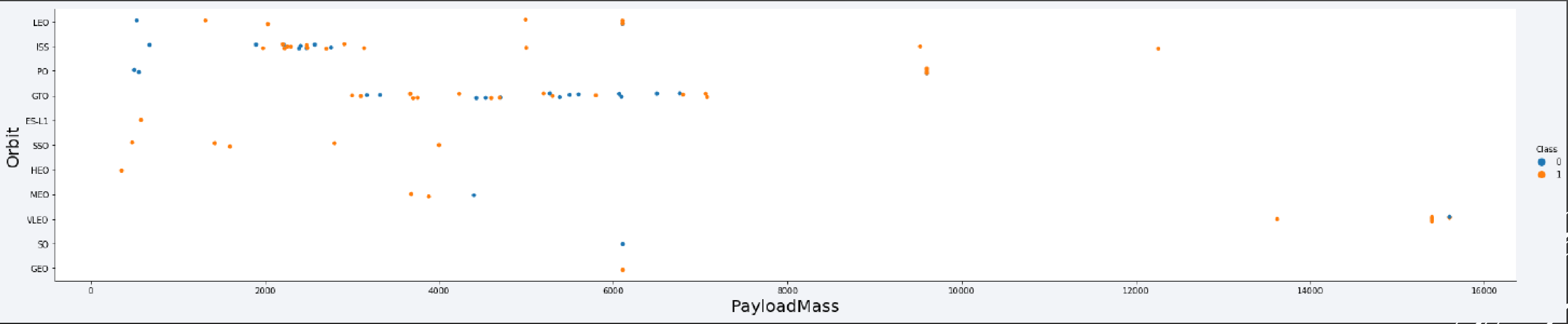


FLIGHT NUMBER VS. ORBIT TYPE



In the LEO orbit 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.

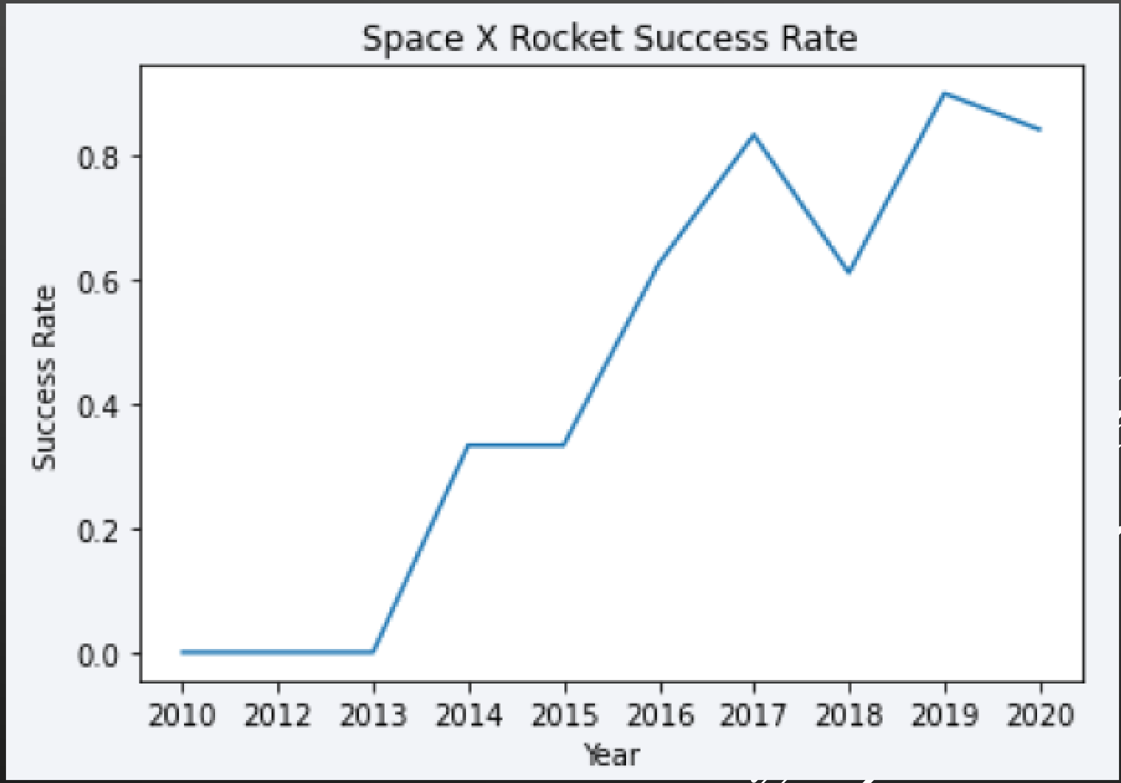
PAYLOAD MASS VS. ORBIT TYPE



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

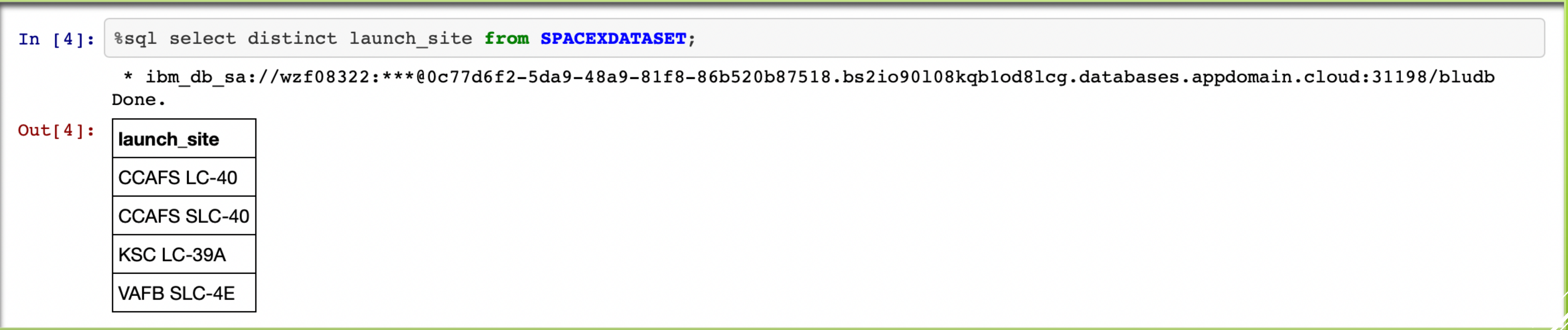
SUCCESS RATE VS. ORBIT TYPE

The success rate since 2013 kept 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.
```

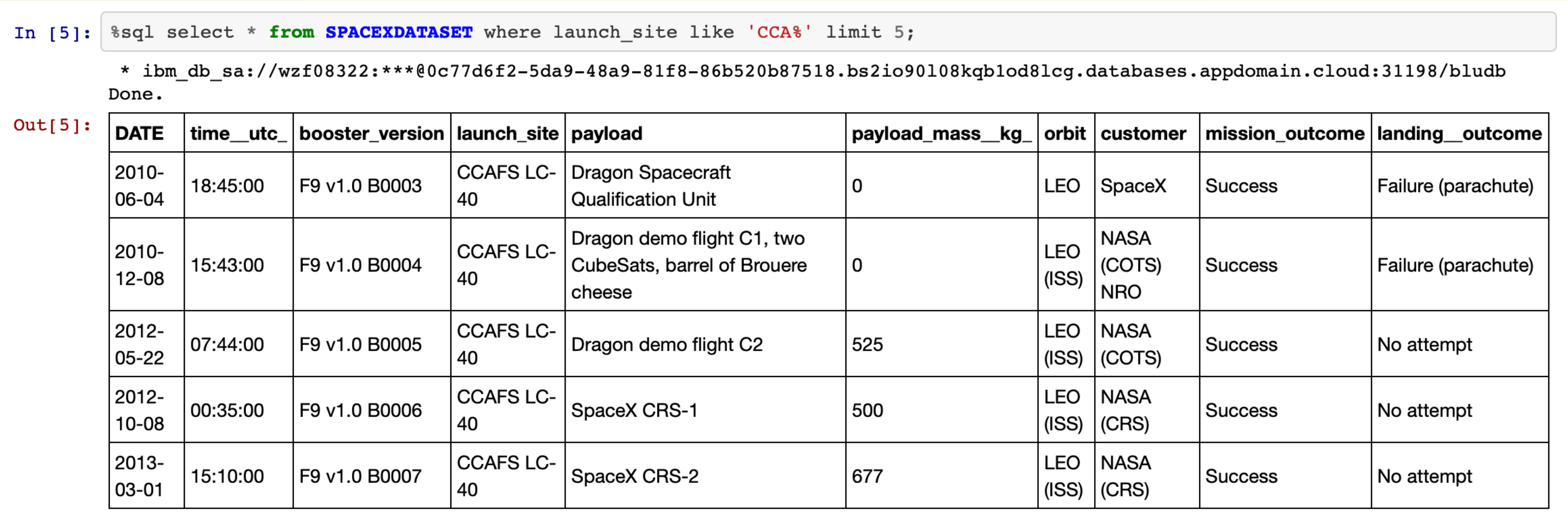
Out[4]:

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

NAMES OF THE UNIQUE LAUNCH SITES

IN THE SPACE MISSION

Launch site names begin with `CCA`

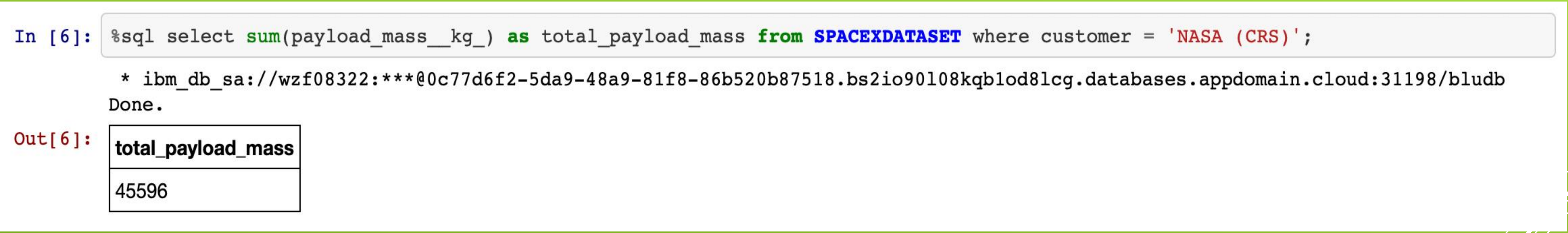


```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;

         * 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

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

Total payload mass

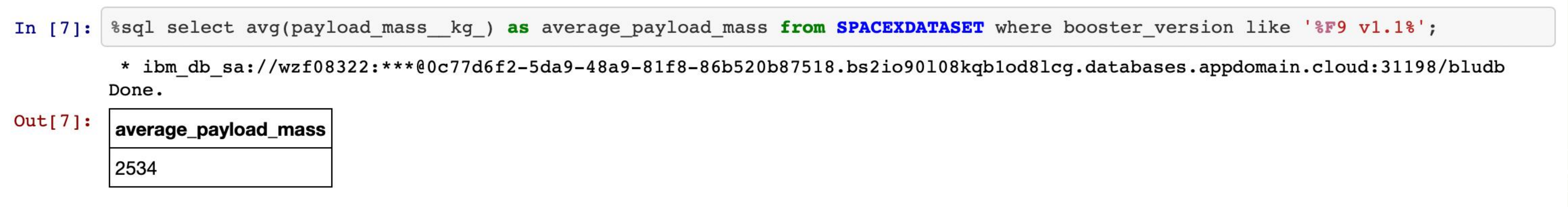


```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.
Out[6]:
```

total_payload_mass
45596

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

Average payload mass by F9 v1.1



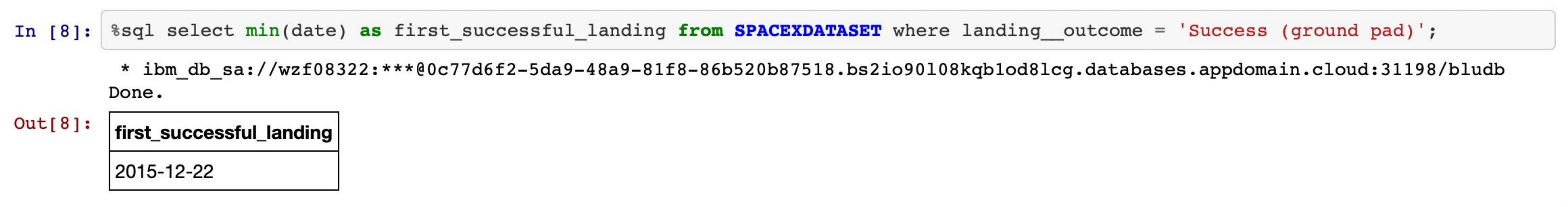
```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

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

Out[7]:

average_payload_mass
2534

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

First successful ground landing date



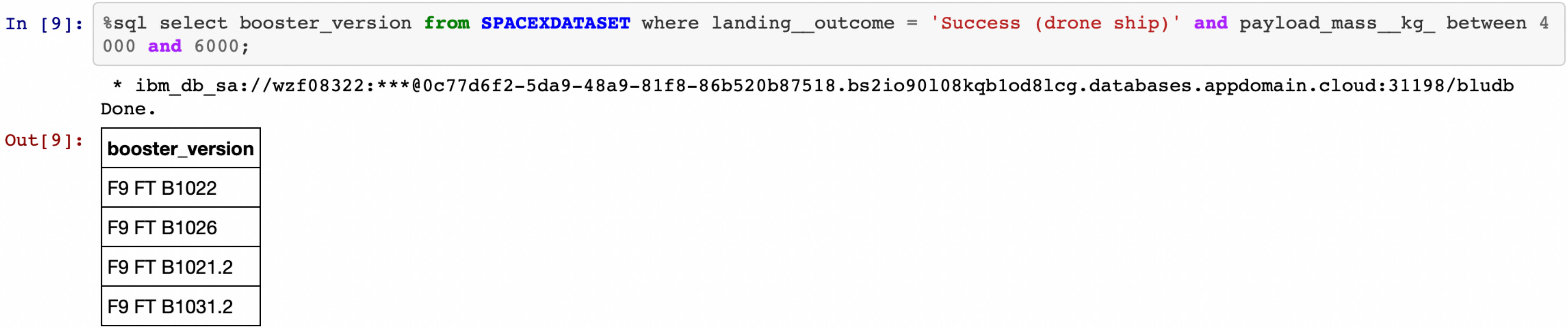
```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

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

Out[8]:

first_successful_landing
2015-12-22

Listing the date when the first successful landing outcome in groundpad 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;
```

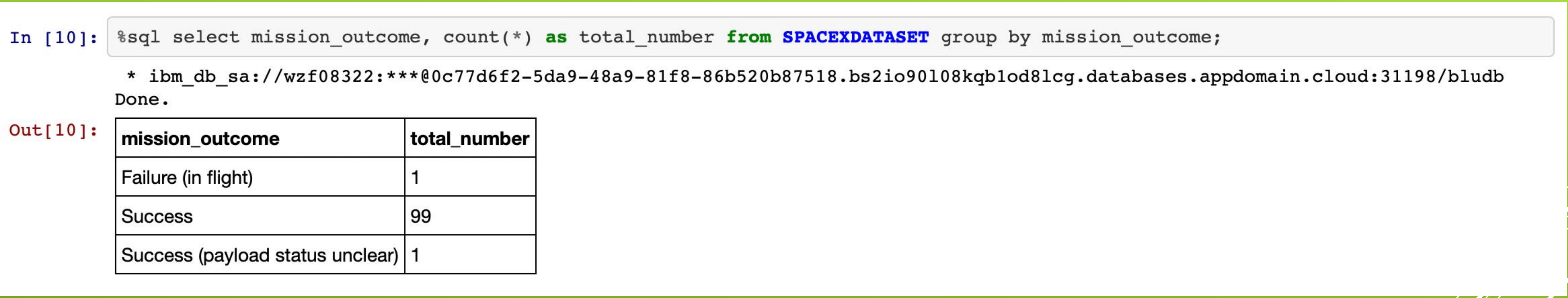
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes



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

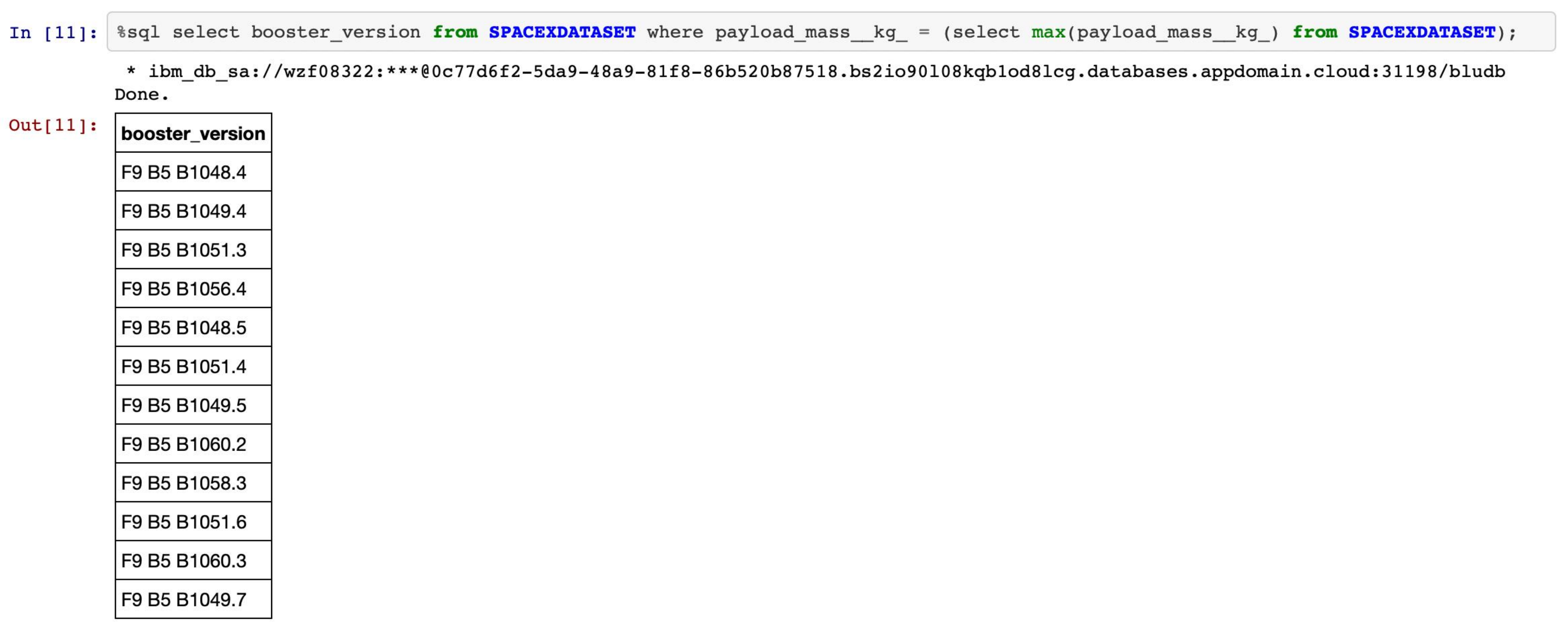
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

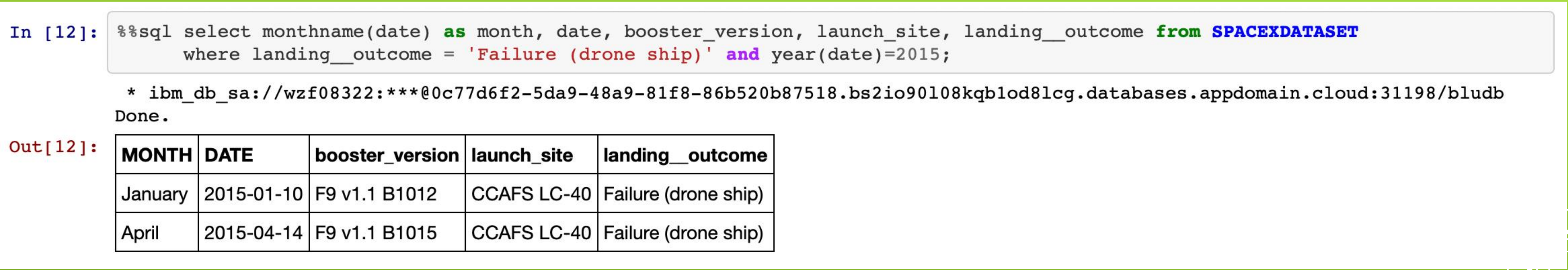


```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

          * 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

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;
```

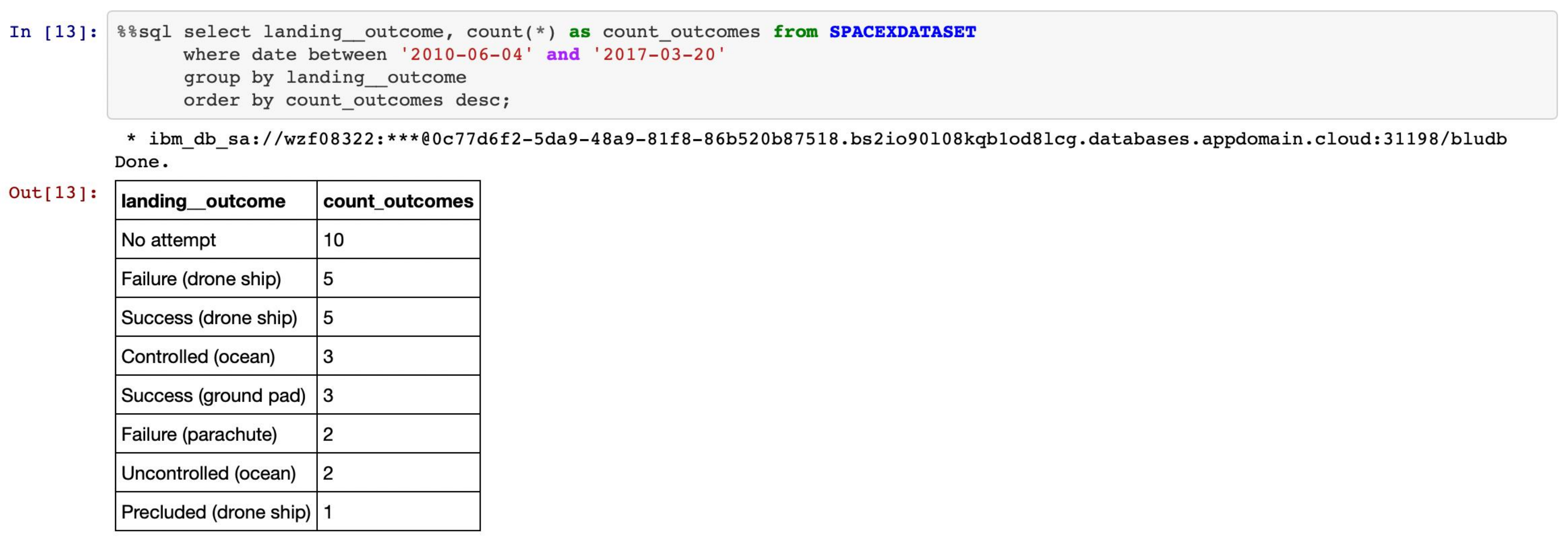
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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)

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



```
%%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;
```

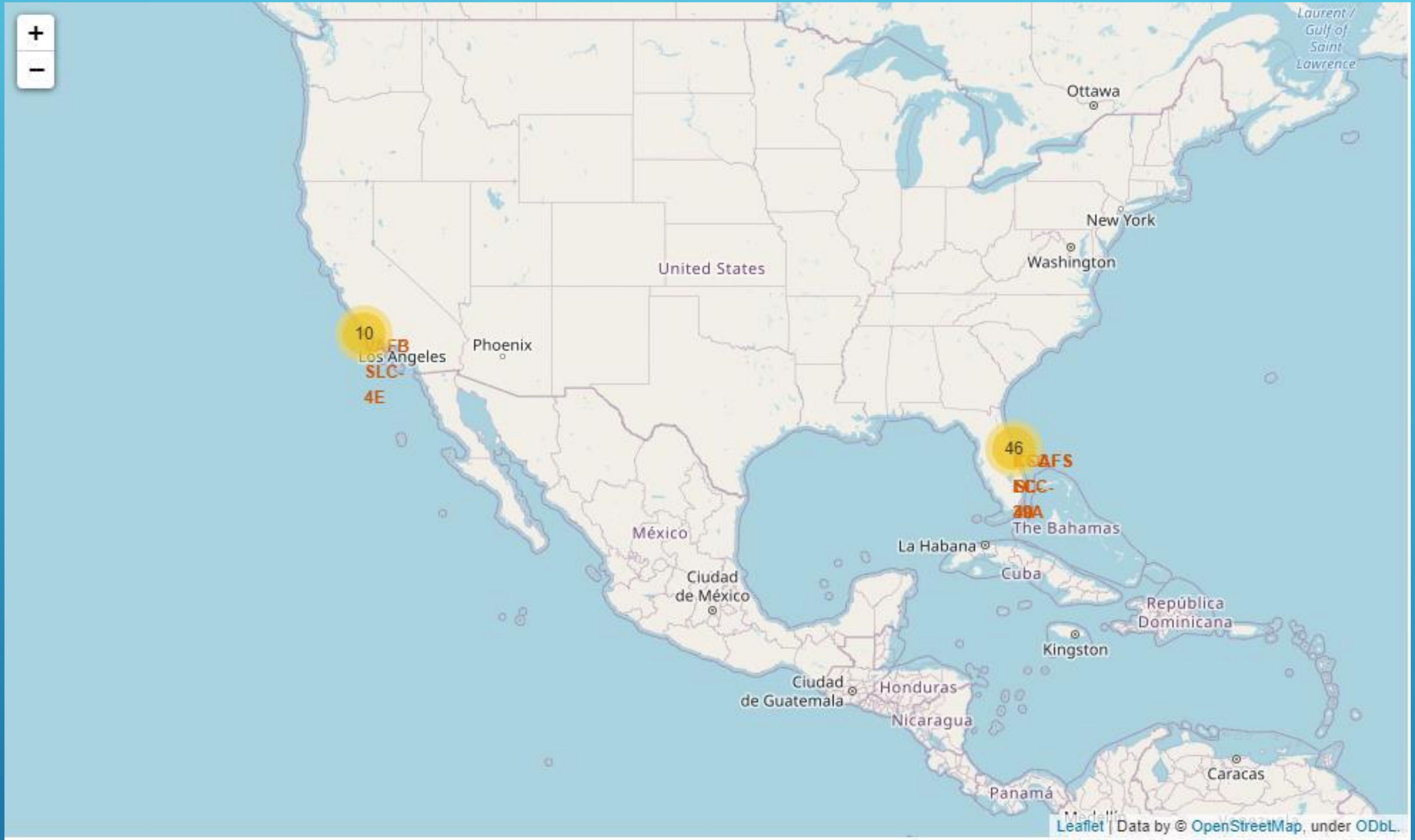
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

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

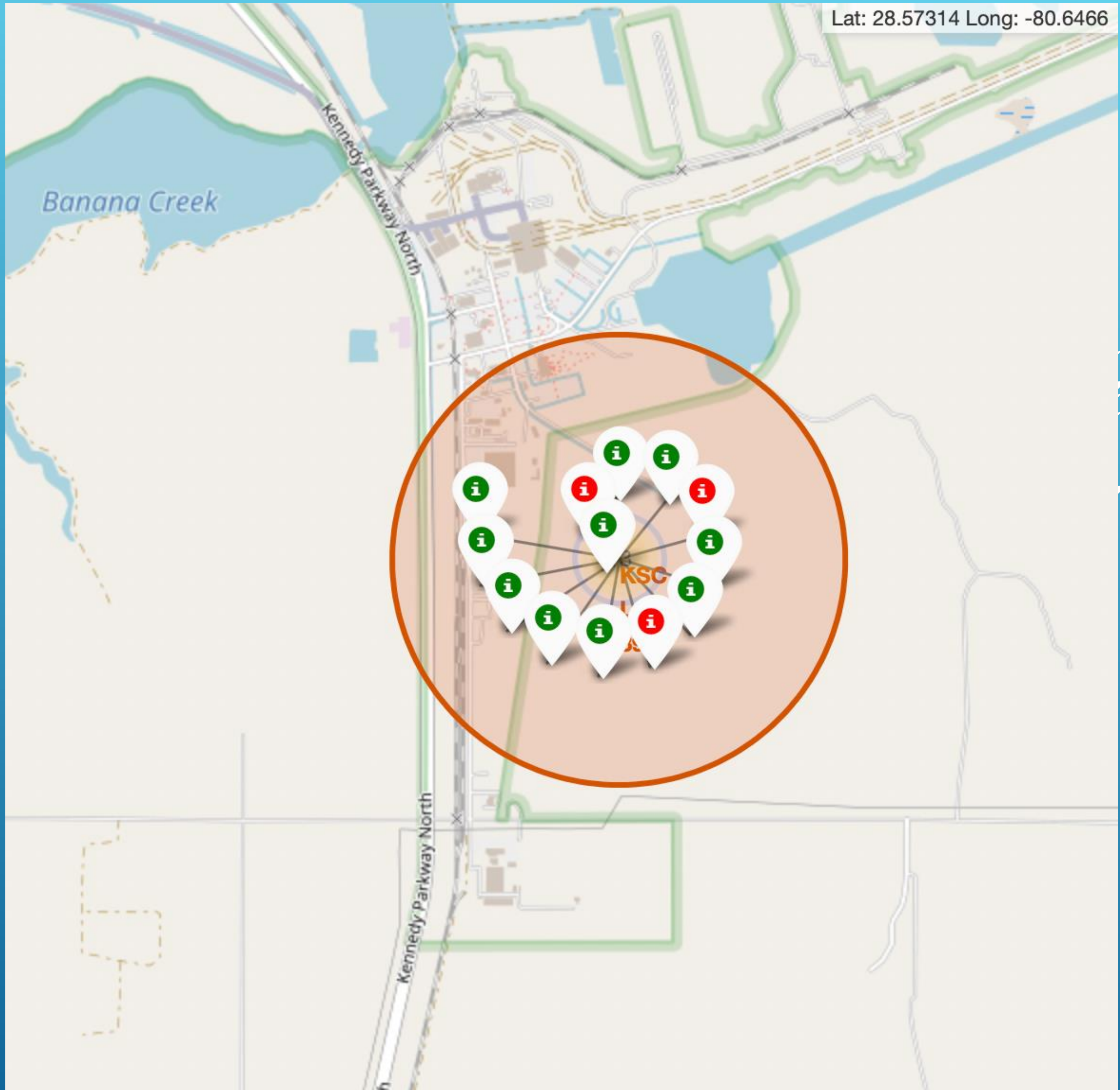
FOLIUM MAP - GROUND STATIONS



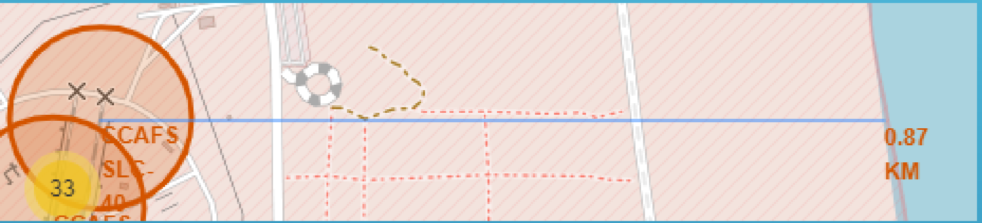
We see that Space X launch sites are located on the coast of the United States

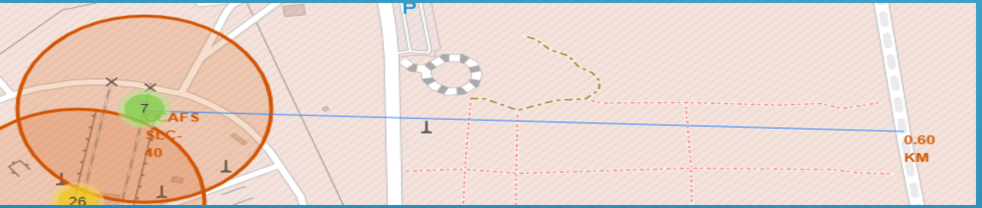
FOLIUM MAP COLOR LABELED MARKERS

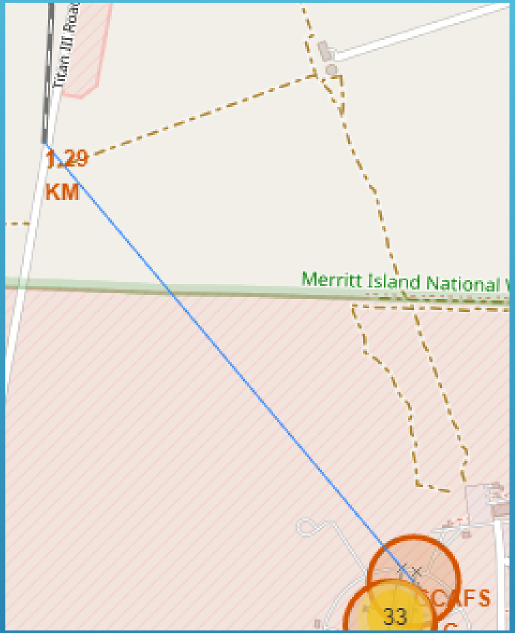
From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- Green Marker = Successful Launch
- Red Marker = Failed Launch
• Launch Site KSC LC-39A has a very high Success Rate

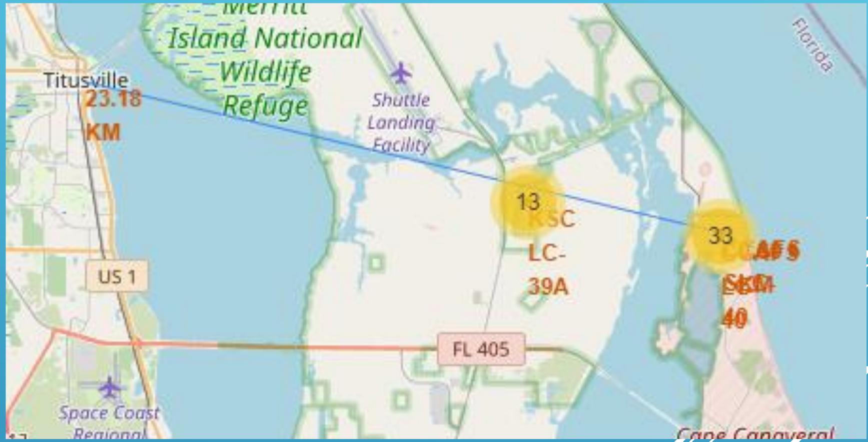


FOLIUM MAP DISTANCES BETWEEN CCAFS SLC 40 AND ITS PROXIMITIES







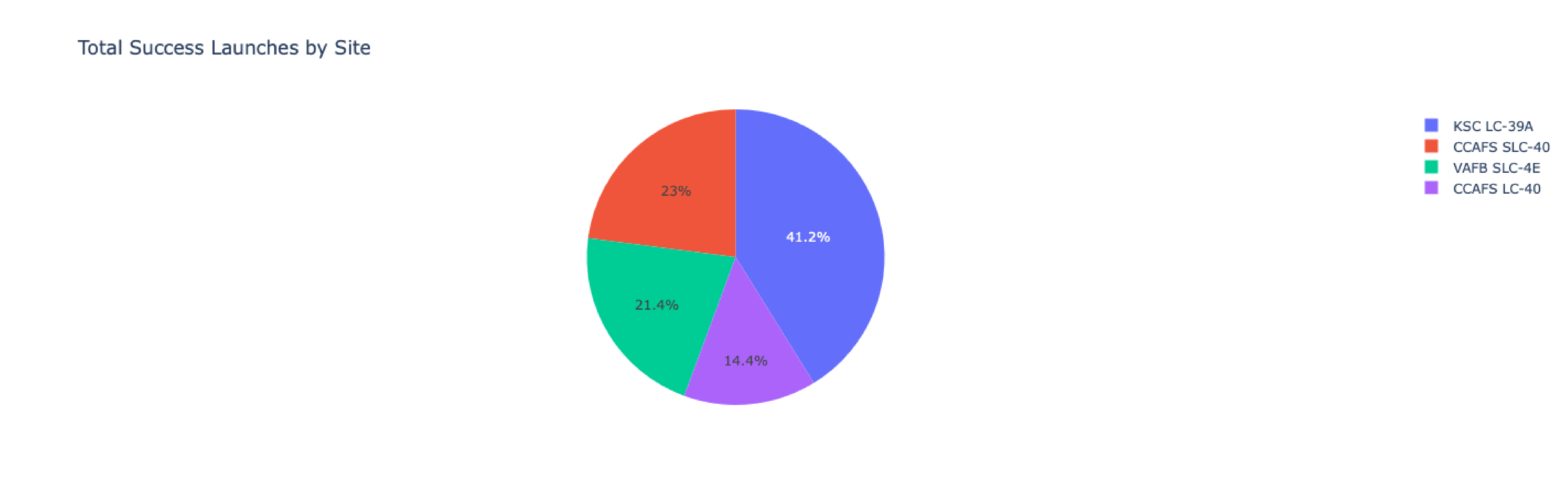


We see that:

CCAFS SLC-40 in close proximity to railways
CCAFS SLC-40in close proximity to highways
CCAFS SLC-40in close proximity to coastline

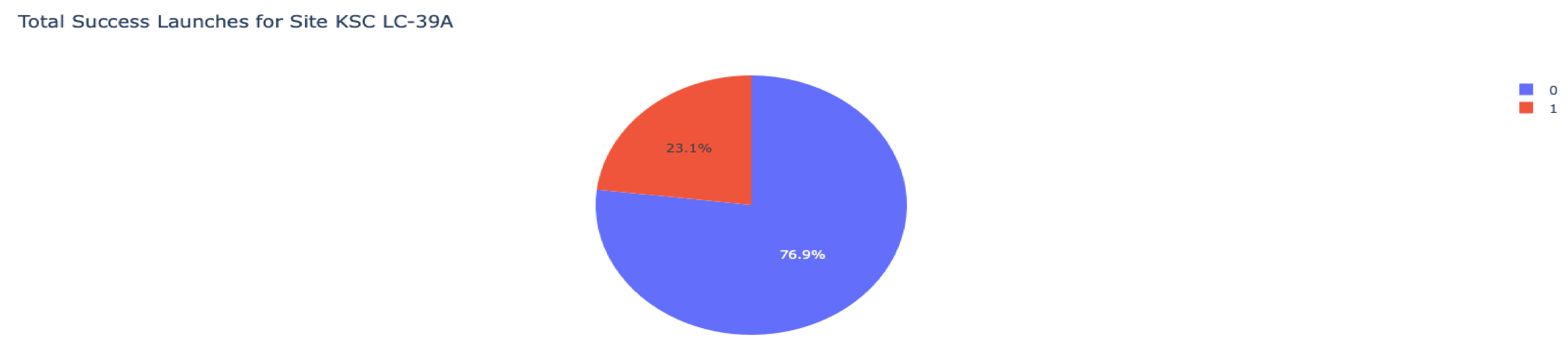
BUILD A DASHBOARD WITH PLOTLY DASH

LAUNCH SUCCESS COUNT FOR ALL SITES



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

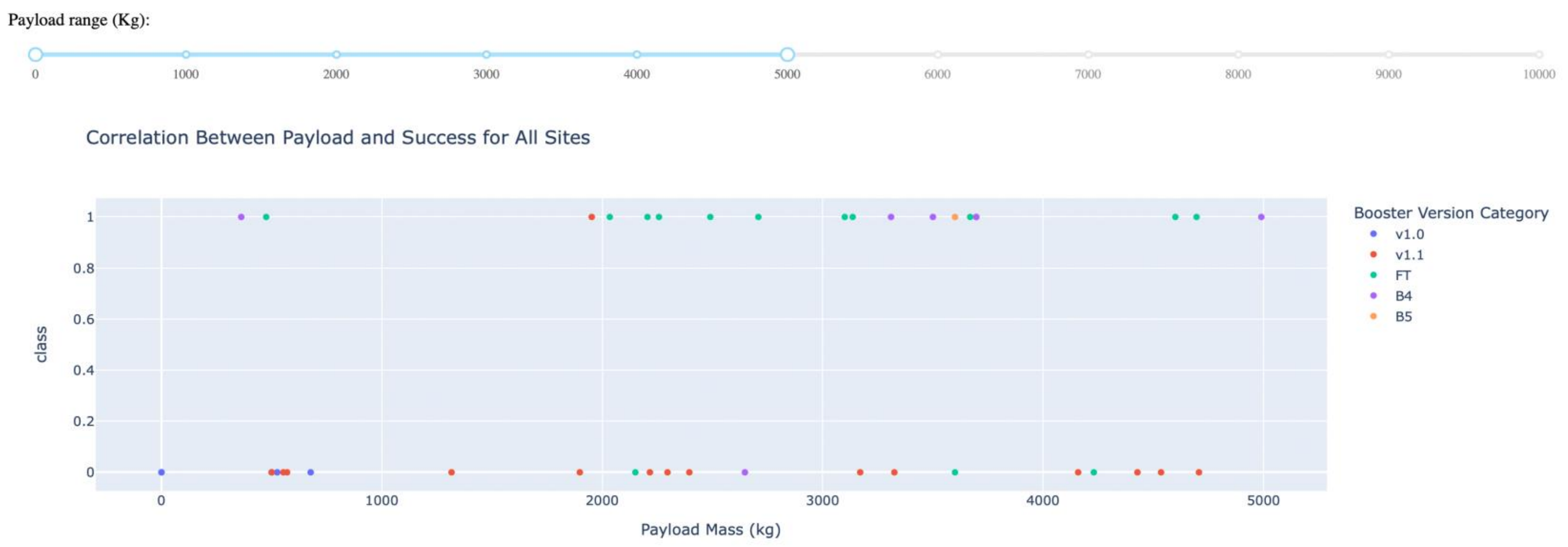
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

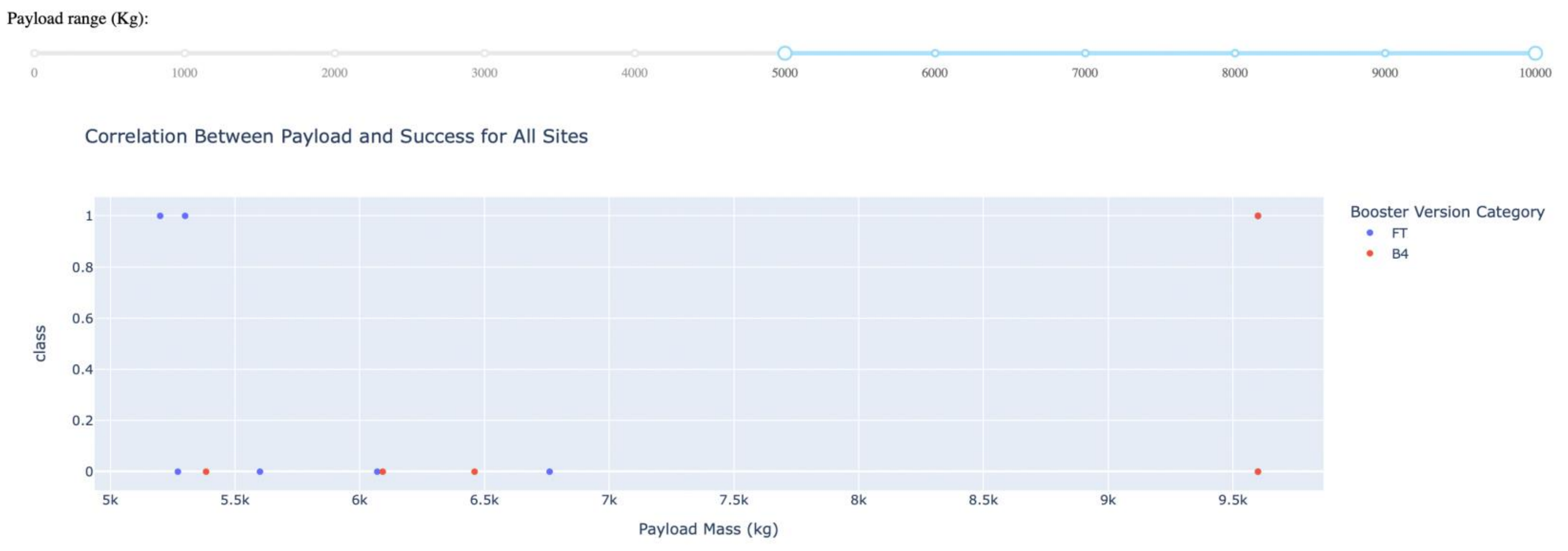


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES

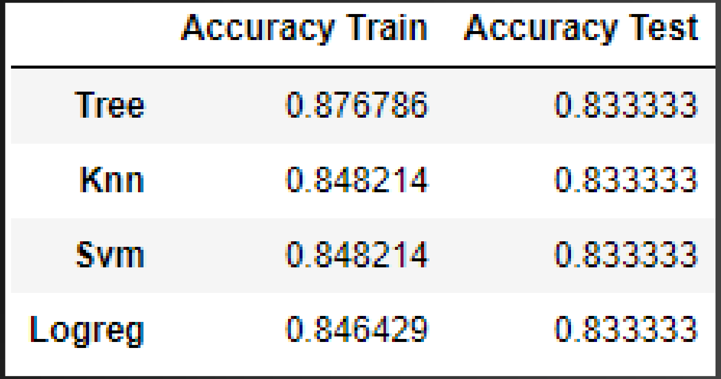
The charts show that payloads between 2000 and 5500 kg have the highest success rate.


Payload range (Kg):

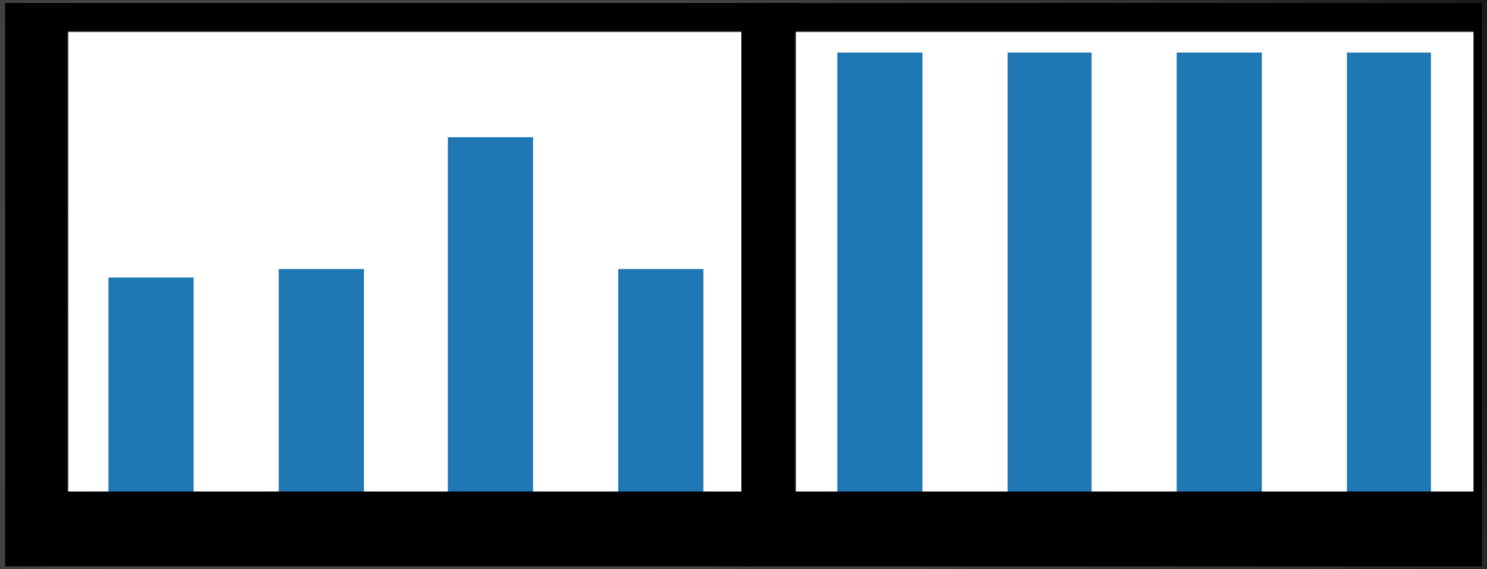

Payload range (Kg):

PREDICTIVE ANALYSIS (CLASSIFICATION)

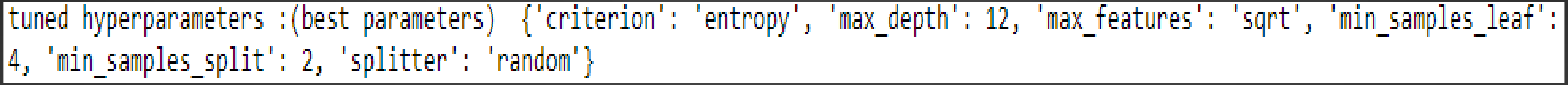
CLASSIFICATION ACCURACY



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



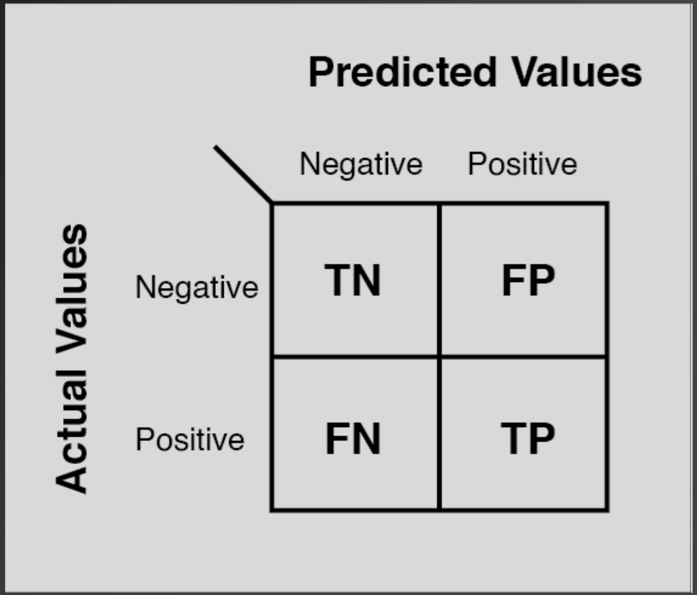
For accuracy test, all methods performed similar. We could get more test data to decide between them. But if we really need to choose one right now, we would take the decision tree:-



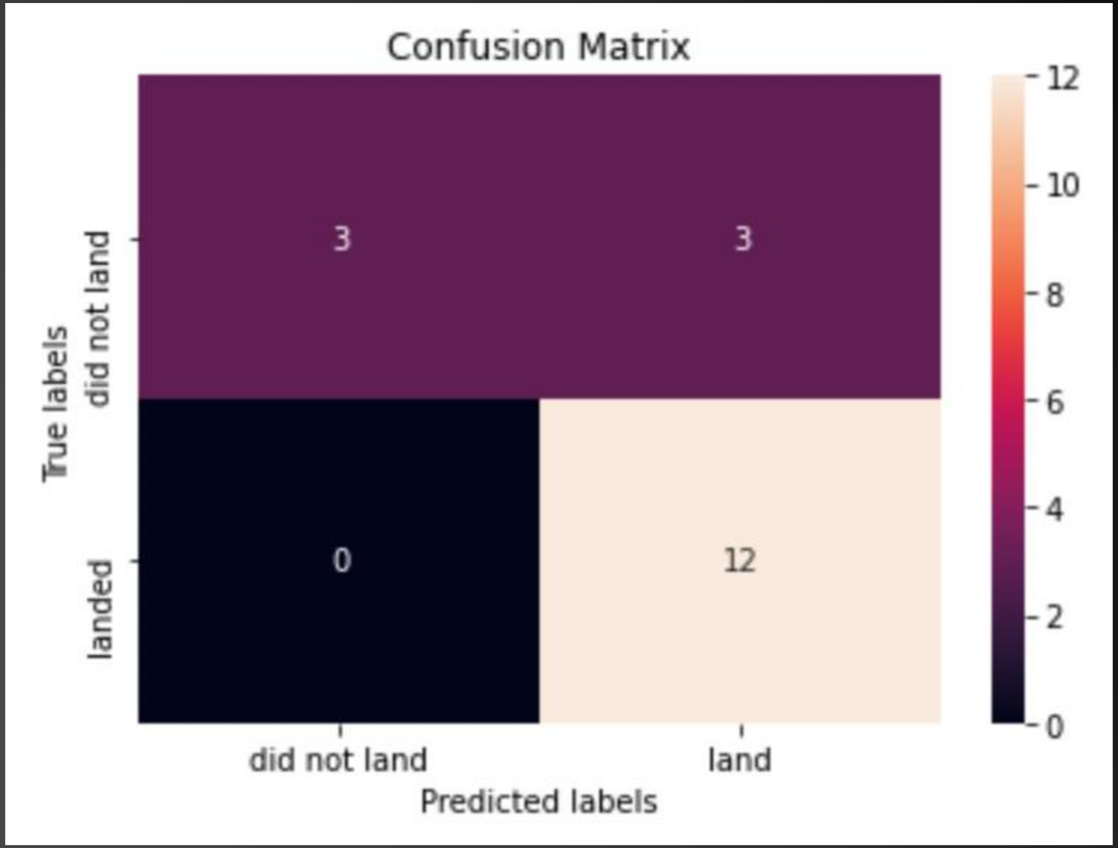
```
tuned hyperparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf':
4, 'min_samples_split': 2, 'splitter': 'random'}
```

Illustration:

Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP



CONFUSION MATRIX

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.
- 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.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass.

APPENDIX

Special thanks to:

- IBM
- Coursera
- Instructors

I wish you have enjoyed

Made by/ Mohamed Mohsen