Data Science Capstone Project

Serhii Ozhereliev

13.11.2022

OUTLINE

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

EXECUTIVE SUMMARY

- Data collection via Web Scraping and API
- EDA with SQL and Pandas
- EDA via visualization
- Folium interactive map
- Plotly Dash dashboard
- ML classification

INTRODUCTION

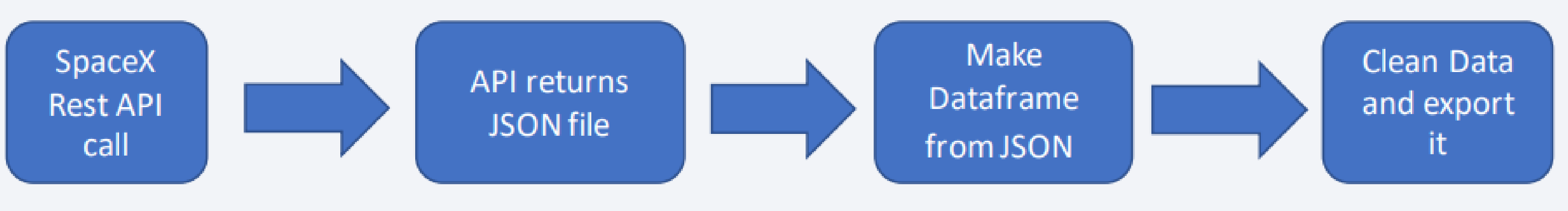
- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against Space X for a rocket launch.
- Questions to answer
  - How do different variables such as payload mass, launch site and orbit affect the success of the first stage landing?
  - How the rate of successful landings change over the years?
  - How to predict if the first stage will land?

METHODOLOGY

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Data wrangling:
  - Data filtering
  - Dropping unnecessary data
  - Dealing with missing values
  - One Hot encoding
- EDA with SQL, Pandas and Data Visualization
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Choosing the best classification model for predicting successful landing

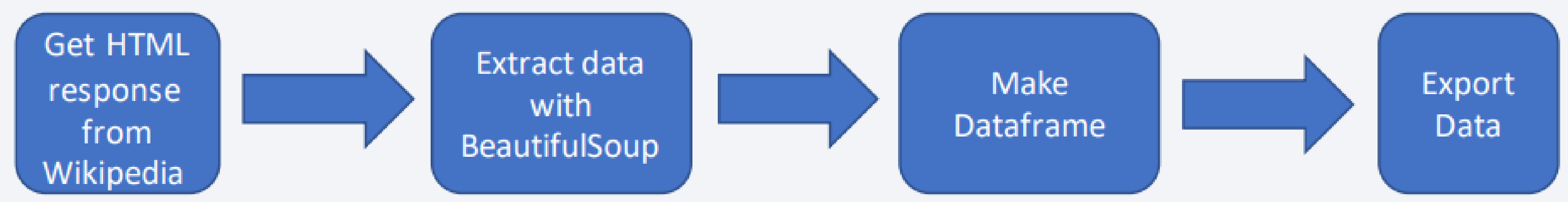
DATA COLLECTION

- 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 web scraping 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.
- During data wrangling we mainly converted those outcomes into Training Labels with `1` means the booster successfully landed `0` means it was unsuccessful.

EDA with Data Visualization

- Charts were plotted:
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
- Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.
- Line charts show trends in data over time (time series).

EDA with SQL

- We performed SQL queries to gather and understand data from dataset:
- Displaying the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA' .
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.
- Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 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

- 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.

Predictive Analysis (Classification)

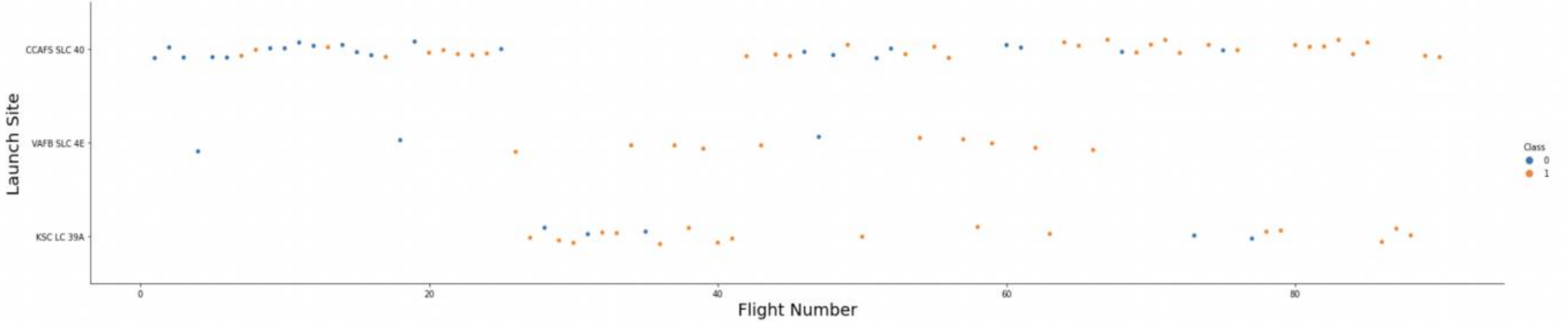
- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get 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)

Results

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

EDA with Visualization Results

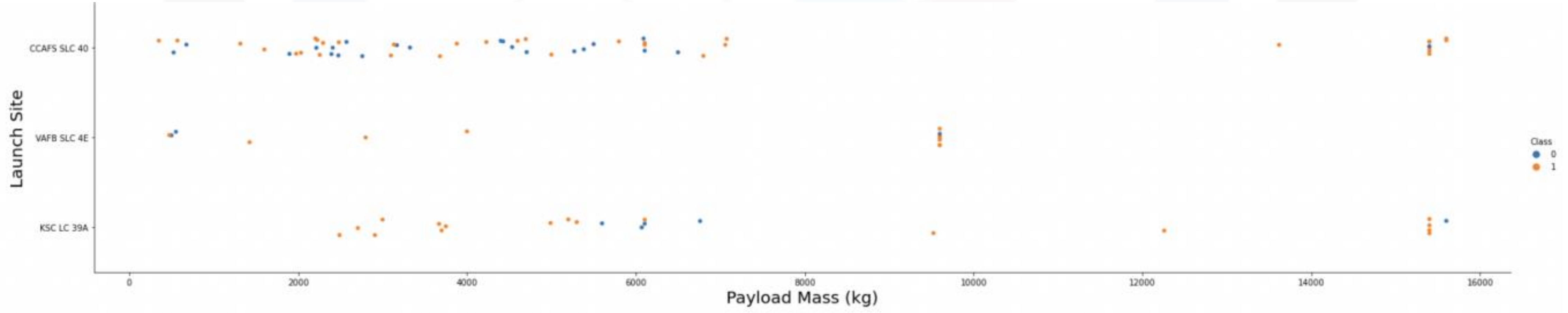
Flight Number vs. Launch Site



- 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.

EDA with Visualization Results

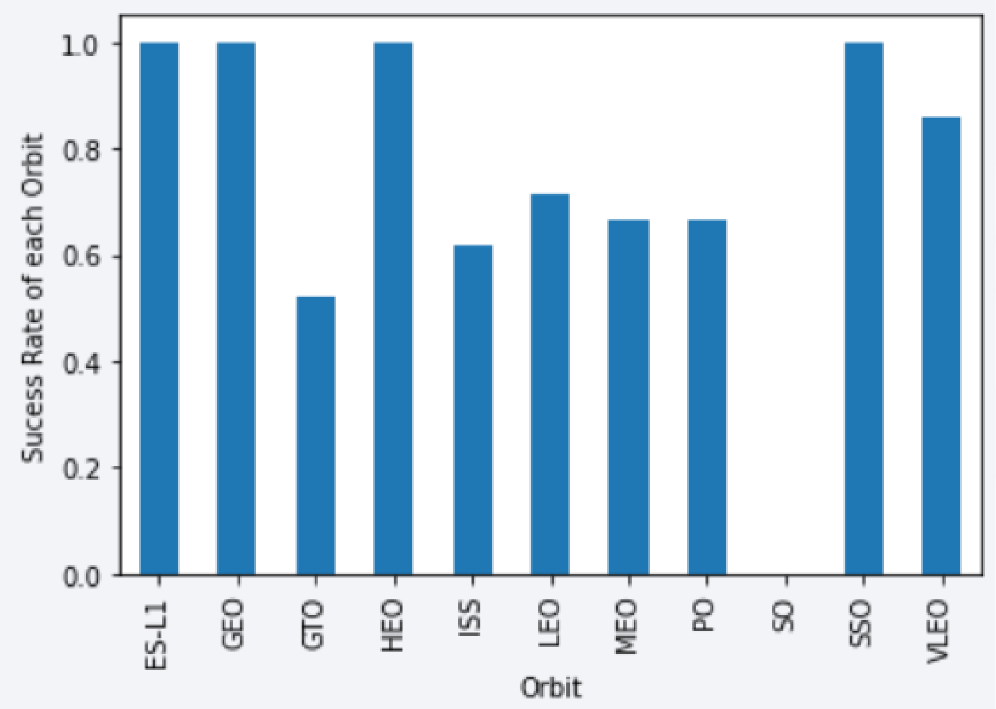
Payload vs. Launch Site



- 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.

EDA with Visualization Results

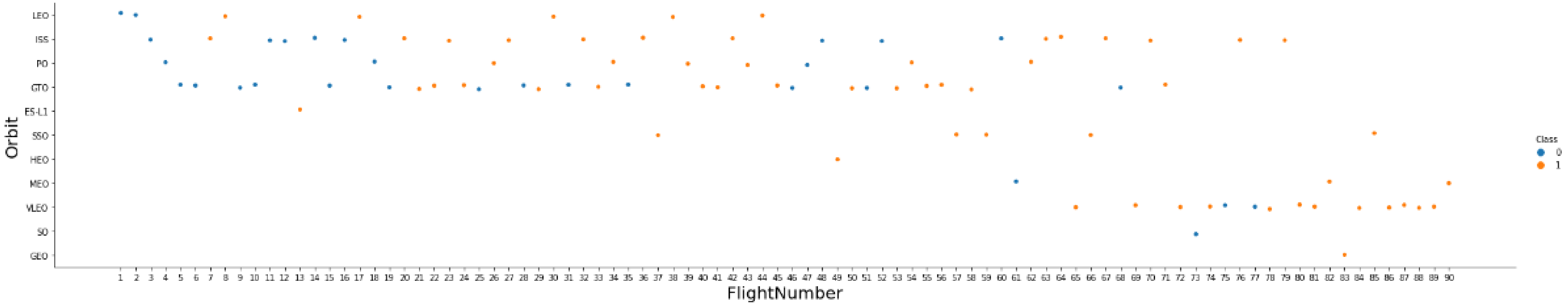
Success rate vs. Orbit type



- 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

EDA with Visualization Results

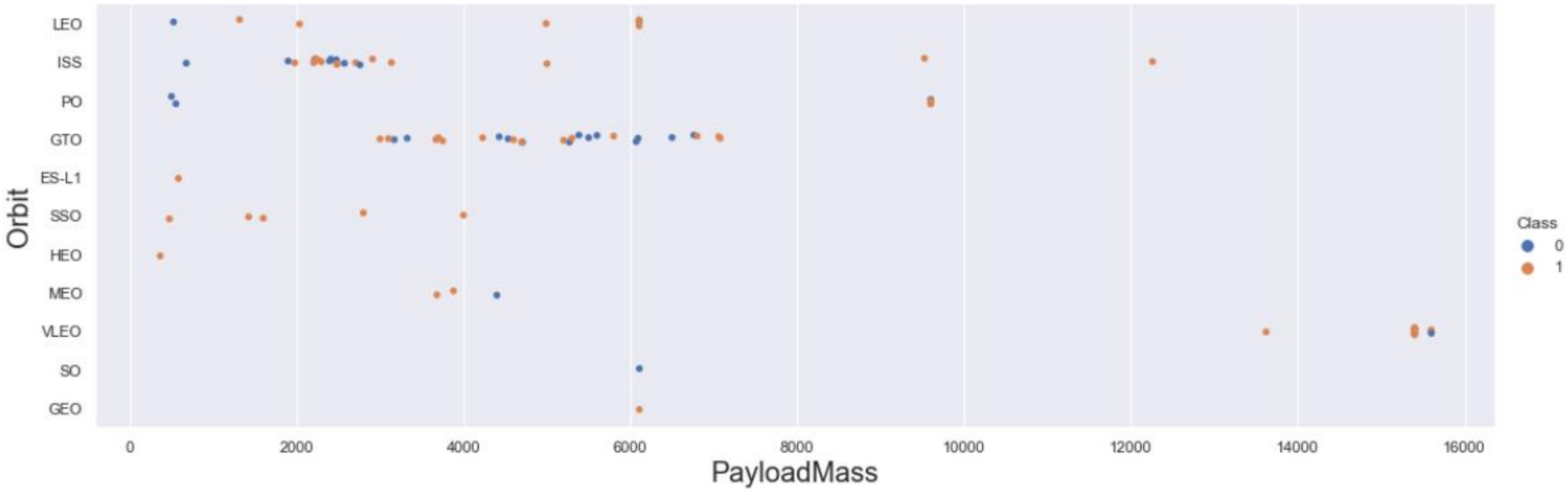
Flight Number vs. Orbit type



- 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.

EDA with Visualization Results

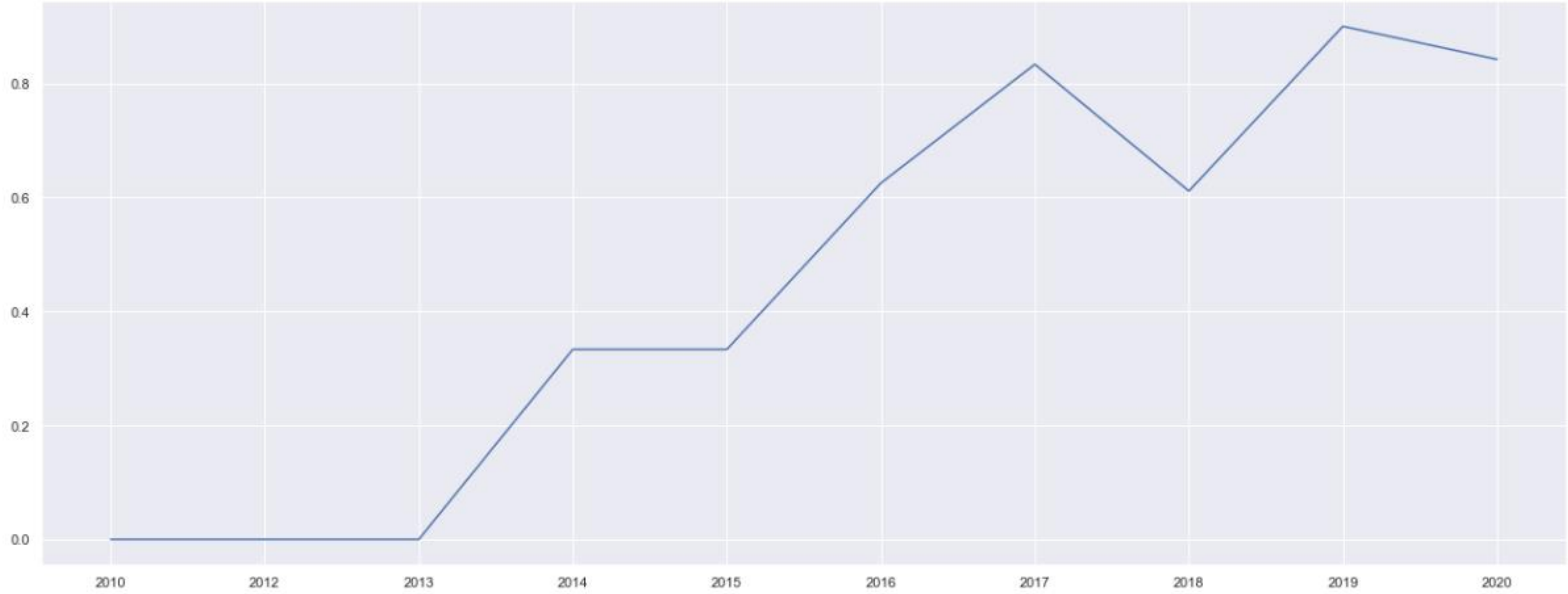
Payload Mass vs. Orbit type



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

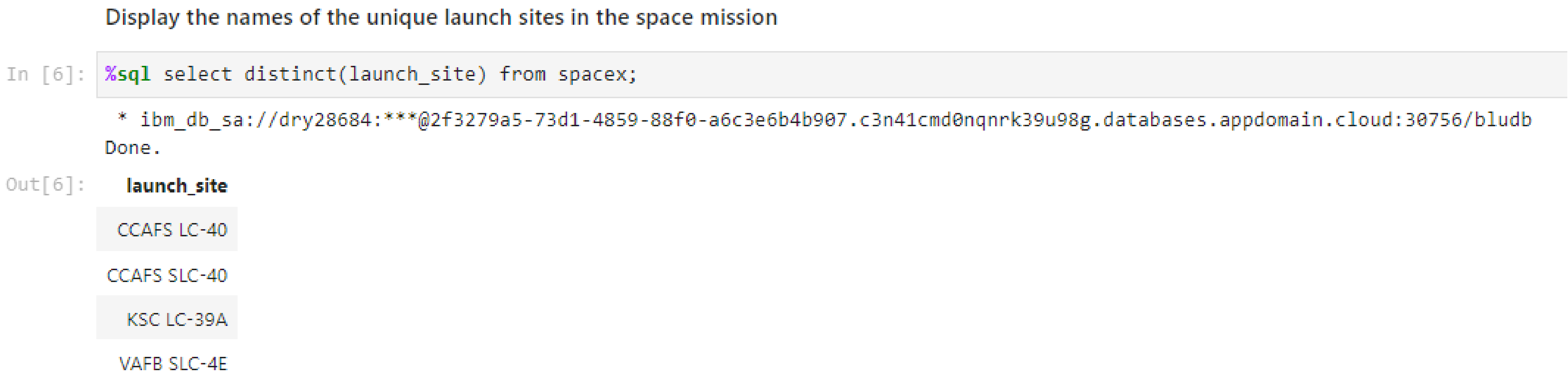
EDA with Visualization Results

Launch success yearly trend



- The success rate since 2013 kept increasing till 2020

EDA with SQL



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

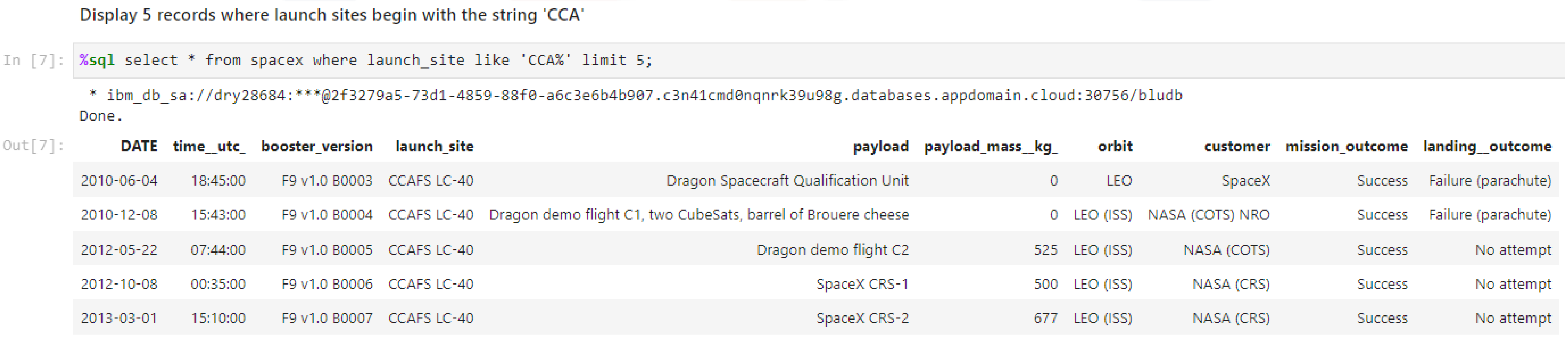
```
In [6]: %sql select distinct(launch_site) from spacex;
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[6]:

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

EDA with SQL



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

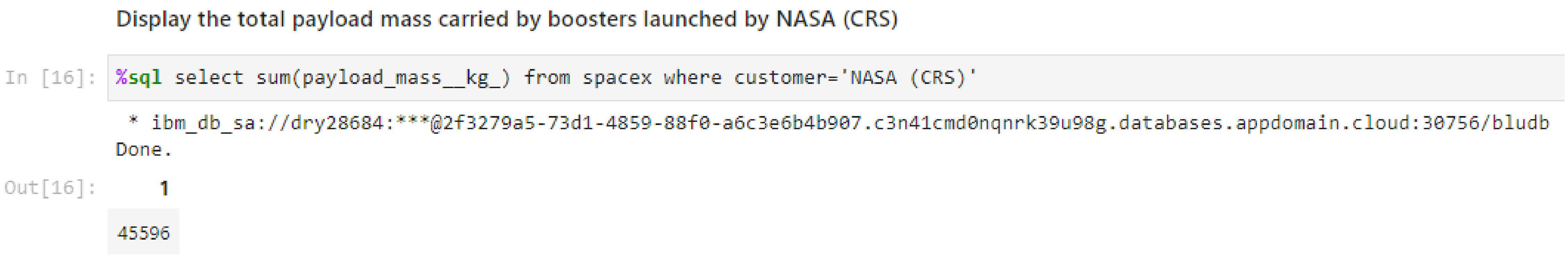
```
In [7]: %sql select * from spacex where launch_site like 'CCA%' limit 5;
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[7]:

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

EDA with SQL



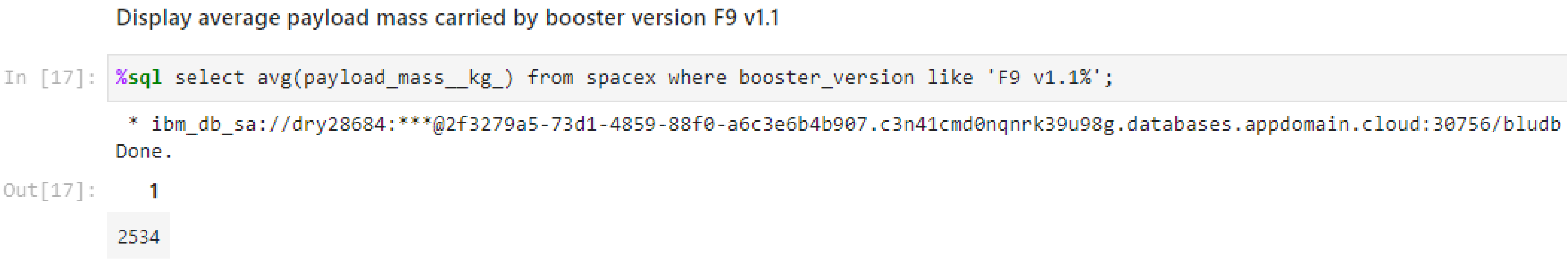
Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [16]: %sql select sum(payload_mass__kg_) from spacex where customer='NASA (CRS)'

 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.

Out[16]:     1
         45596
```

EDA with SQL



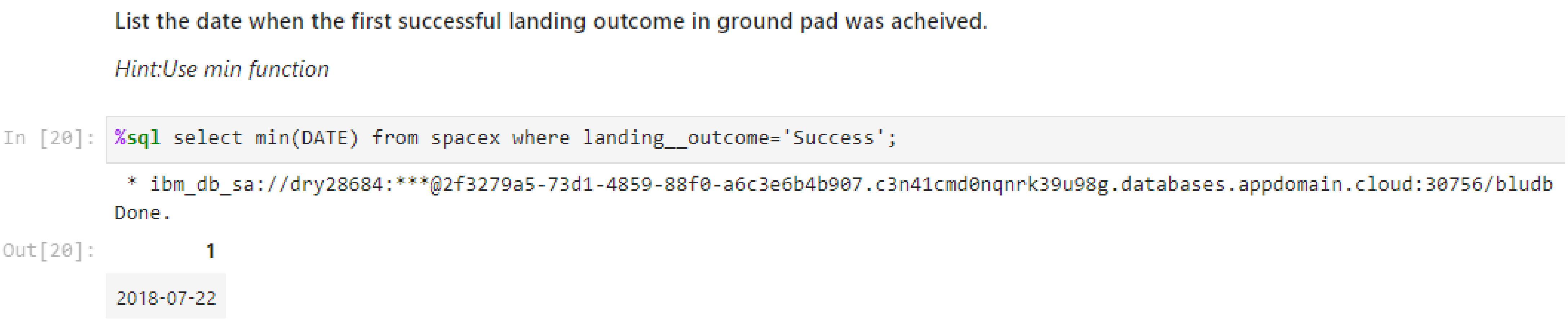
Display average payload mass carried by booster version F9 v1.1

```
In [17]: %sql select avg(payload_mass__kg_) from spacex where booster_version like 'F9 v1.1%';
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

```
Out[17]:  1
          2534
```

EDA with SQL



List the date when the first successful landing outcome in ground pad was acheived.

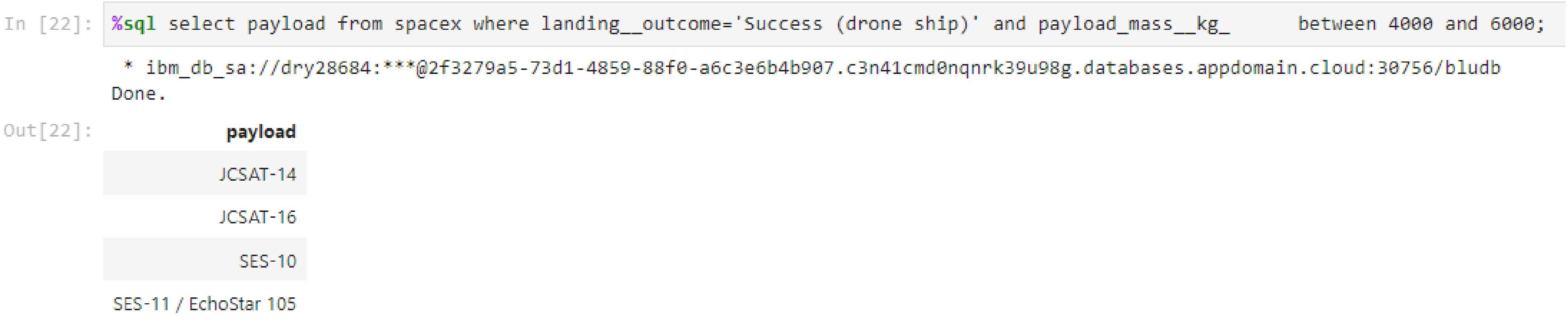
Hint:Use min function

In [20]:
```
%sql select min(DATE) from spacex where landing__outcome='Success';
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[20]:
```
         1
2018-07-22
```

EDA with SQL



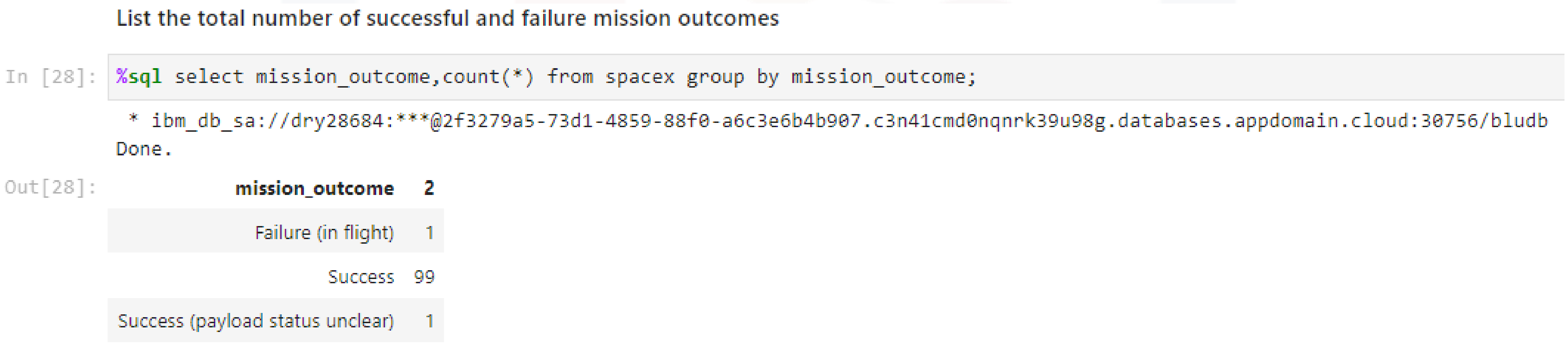
```
In [22]: %sql select payload from spacex where landing__outcome='Success (drone ship)' and payload_mass__kg_      between 4000 and 6000;

 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[22]:

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

EDA with SQL



List the total number of successful and failure mission outcomes

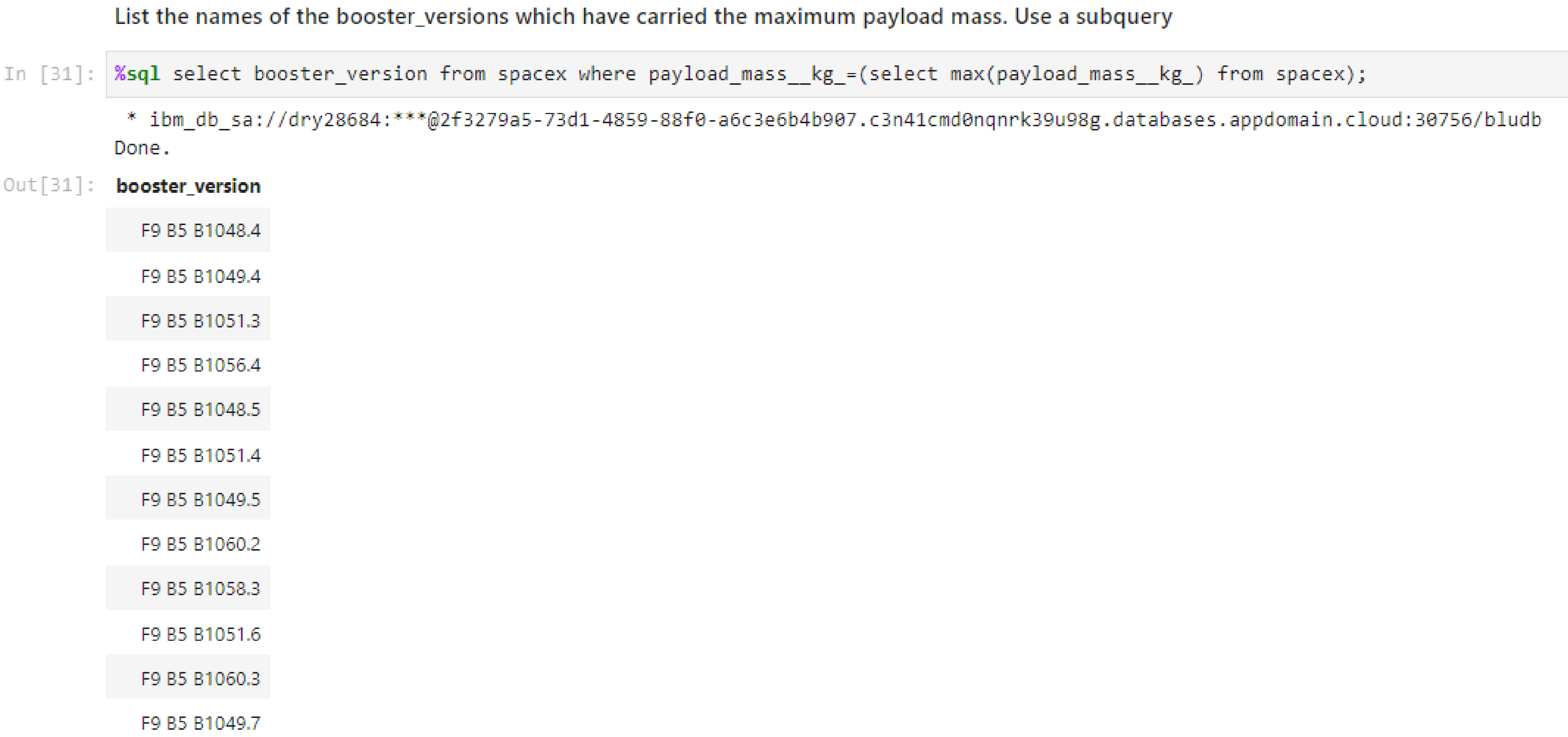
```
In [28]: %sql select mission_outcome,count(*) from spacex group by mission_outcome;
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[28]:

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

EDA with SQL



List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

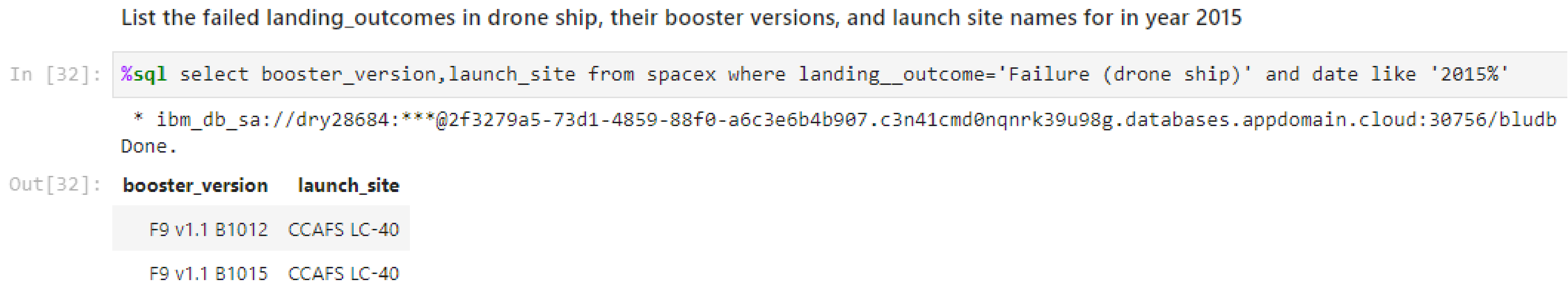
In [31]:
```
%sql select booster_version from spacex where payload_mass__kg_=(select max(payload_mass__kg_) from spacex);
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[31]:

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

EDA with SQL



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

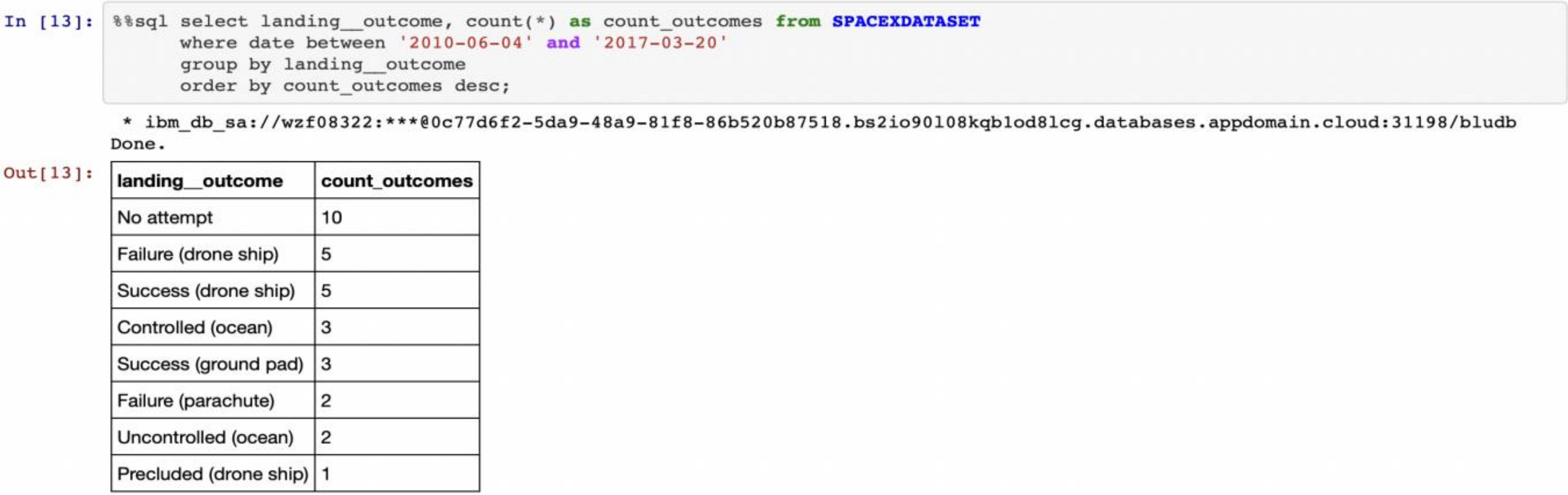
```
In [32]: %sql select booster_version,launch_site from spacex where landing__outcome='Failure (drone ship)' and date like '2015%'
```

```
 * ibm_db_sa://dry28684:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/bludb
Done.
```

Out[32]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

EDA with SQL



```
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;
```

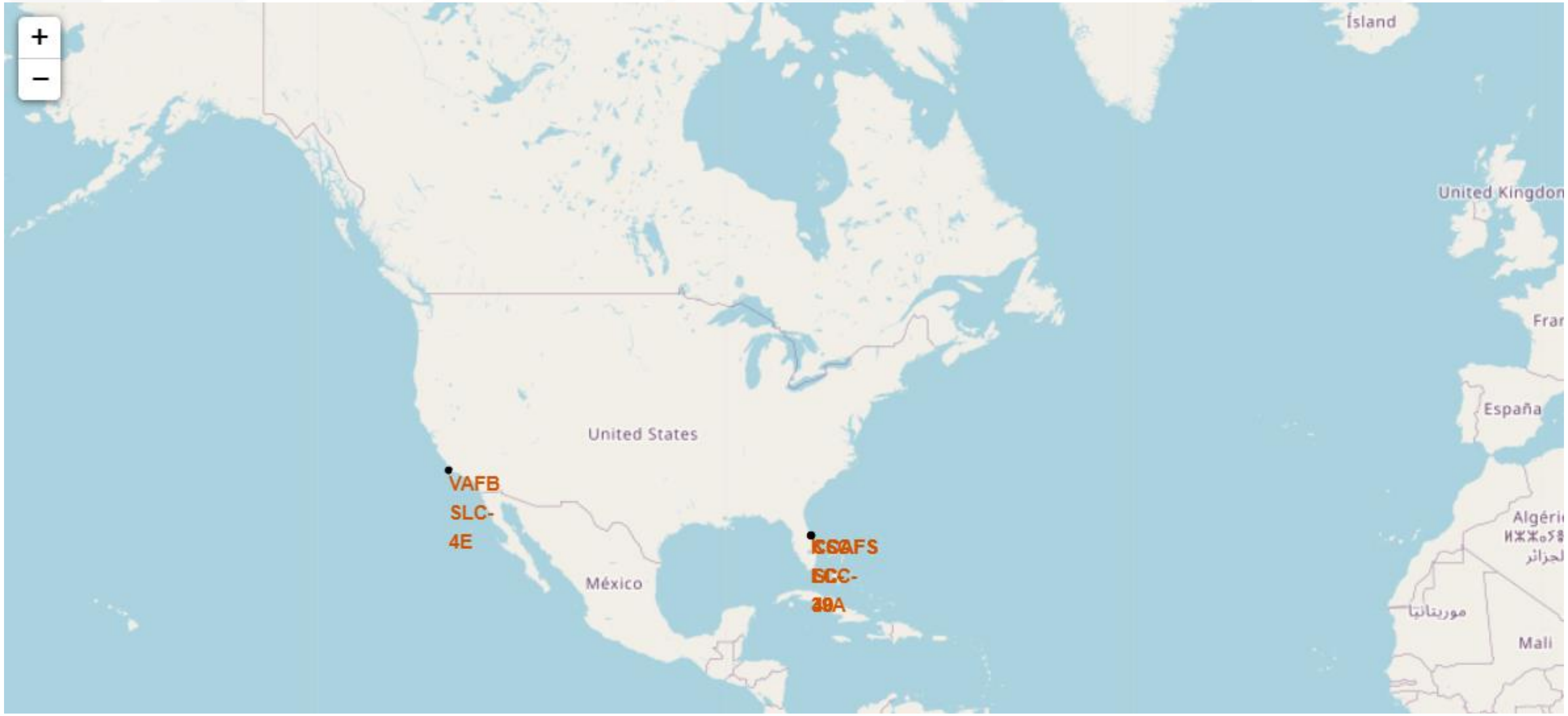
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

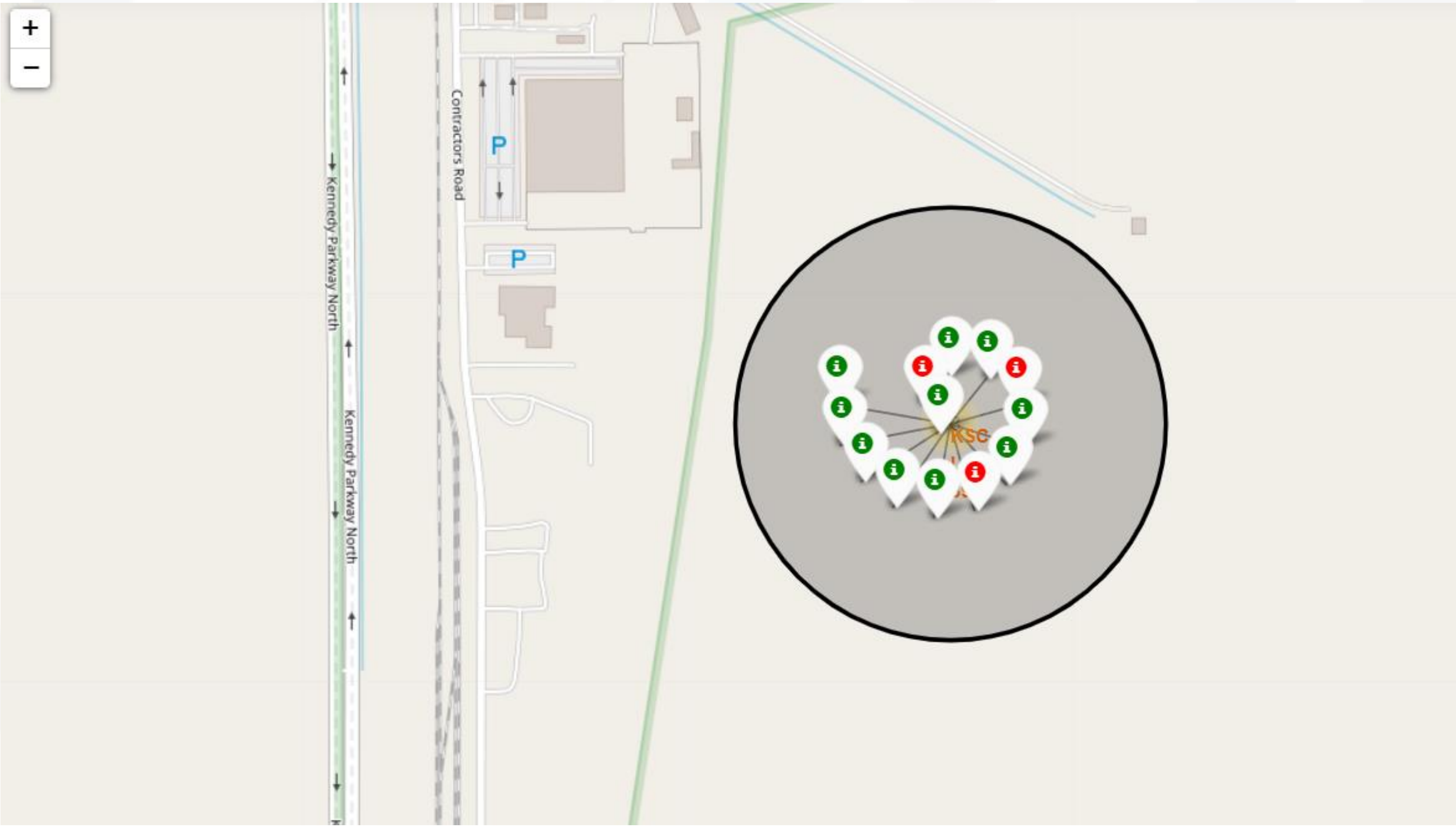
Interactive map with Folium

All launch sites' location markers on a global map

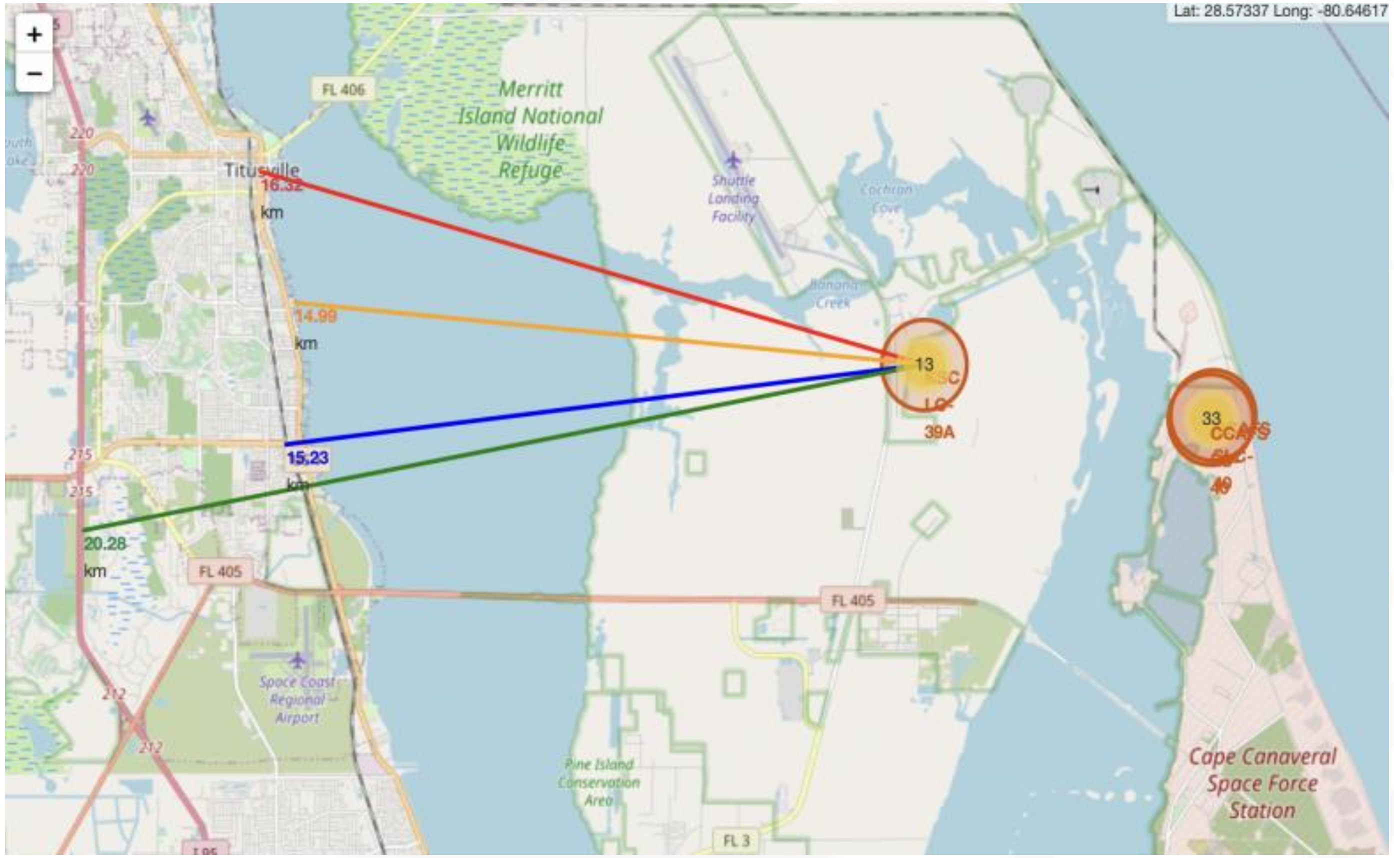


Interactive map with Folium

Color-labeled launch records on the map (green – successful, red – failed)



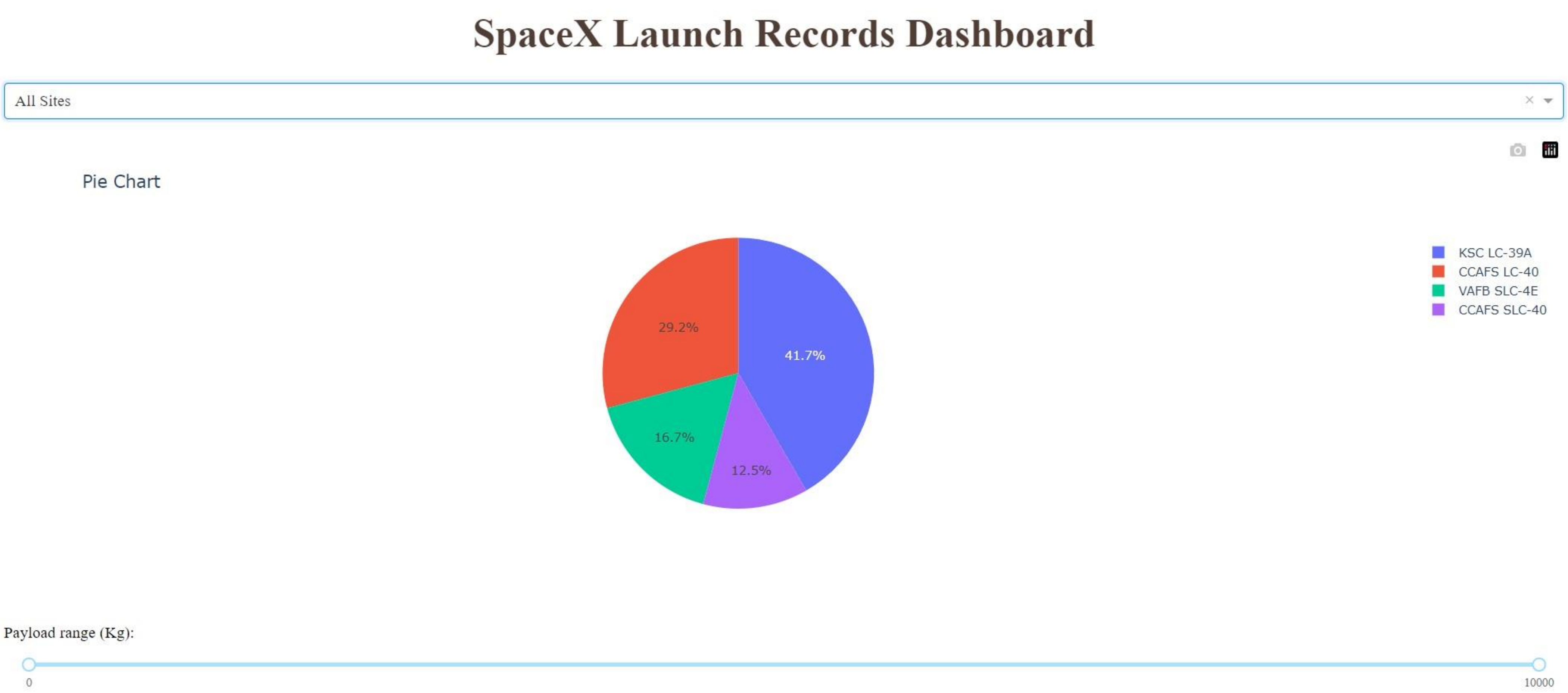
Interactive map with Folium



+
−
Lat: 28.57337 Long: -80.64617
FL 406
Merritt Island National Wildlife Refuge
Titusville
16.32 km
14.99 km
15.23 km
20.28 km
Shuttle Landing Facility
Cochran Cove
Banana Creek
13
KSC LC-39A
33
CCAFS SLC-40
FL 405
Space Coast Regional Airport
Pine Island Conservation Area
FL 3
Cape Canaveral Space Force Station

- From the visual analysis of the launch site KSC LC-39A we can clearly see that it is:
  - relative close to railway (15.23 km)
  - relative close to highway (20.28 km)
  - relative close to coastline (14.99 km)
- Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).
- Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.

Build a Dashboard with Plotly Dash



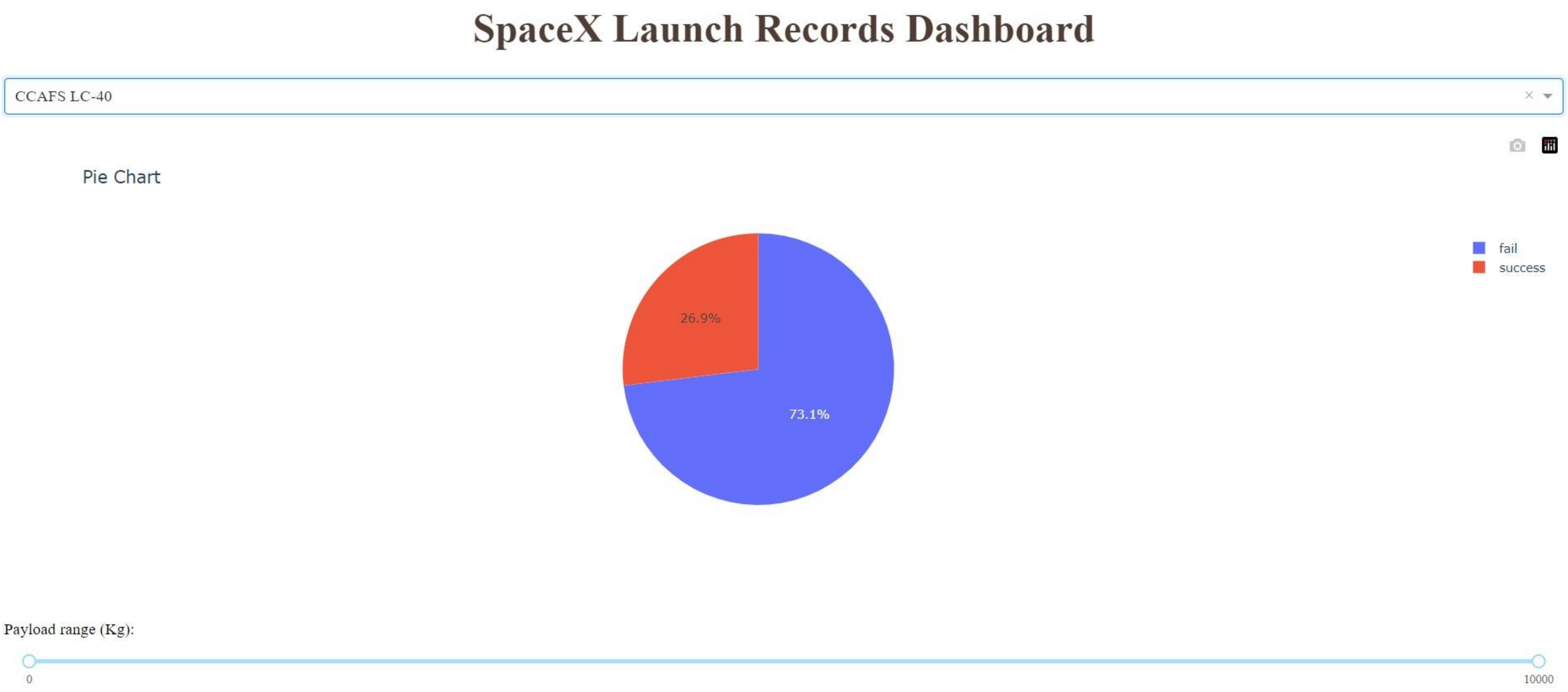
SpaceX Launch Records Dashboard

All Sites

Payload range (Kg):

0 – 10000

Build a Dashboard with Plotly Dash



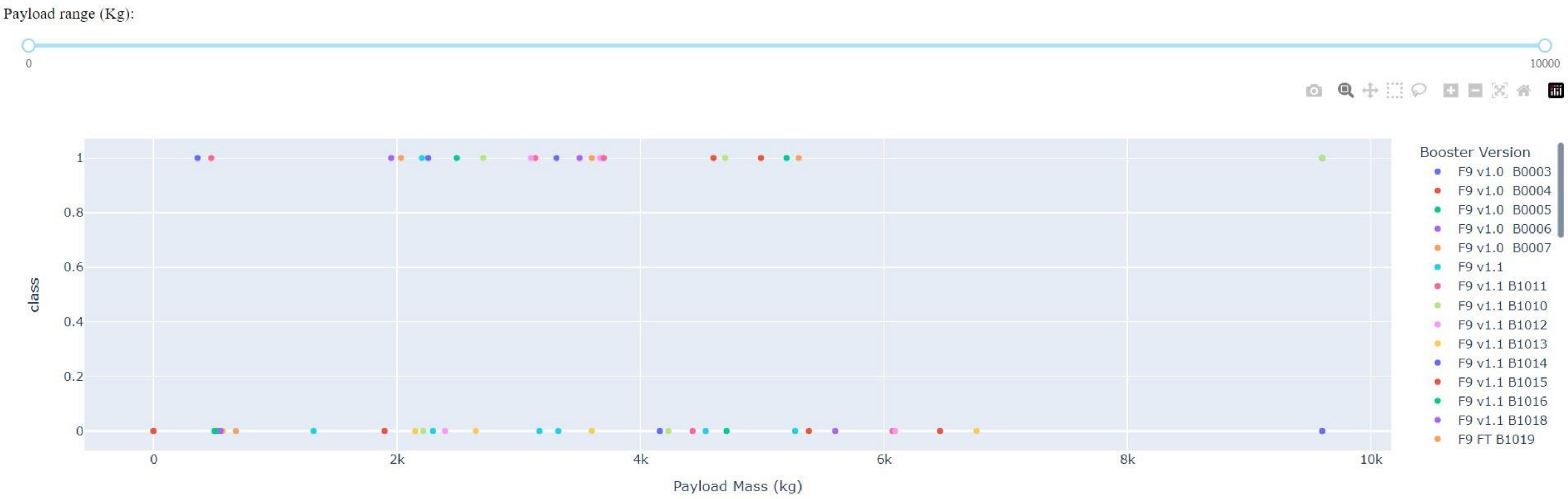
SpaceX Launch Records Dashboard

CCAFS LC-40

Payload range (Kg):

Build a Dashboard with Plotly Dash

Payload Mass vs. Class for all sites

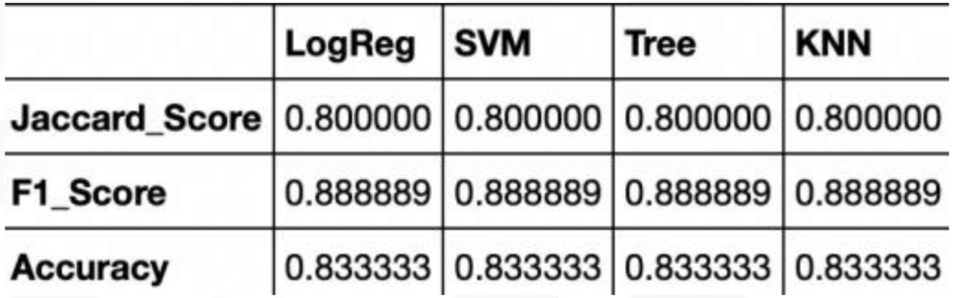


Predictive analysis (Classification)

Classification Accuracy

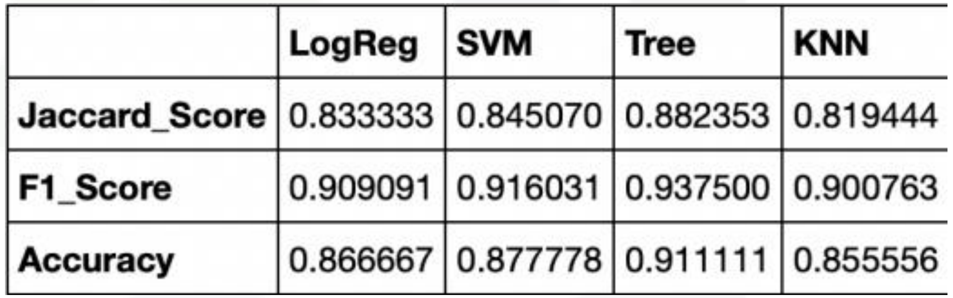
- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

Scores and Accuracy of the Test Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

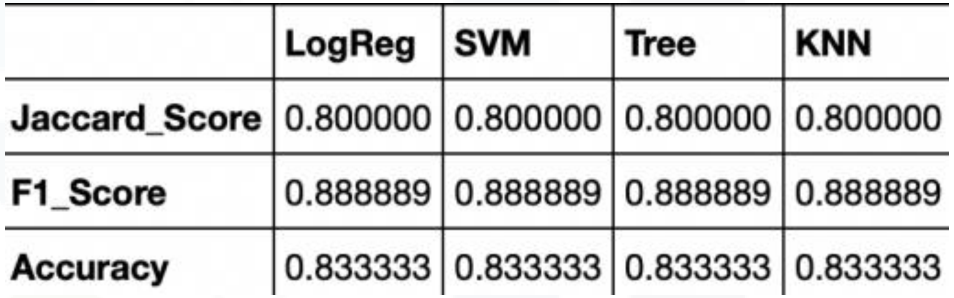
Scores and Accuracy of the Entire Data Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

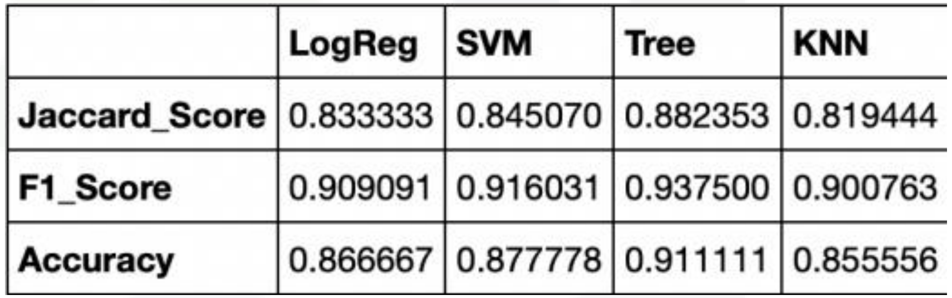
Predictive analysis (Classification)

Scores and Accuracy of the Test Set



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

Scores and Accuracy of the Entire Data Set

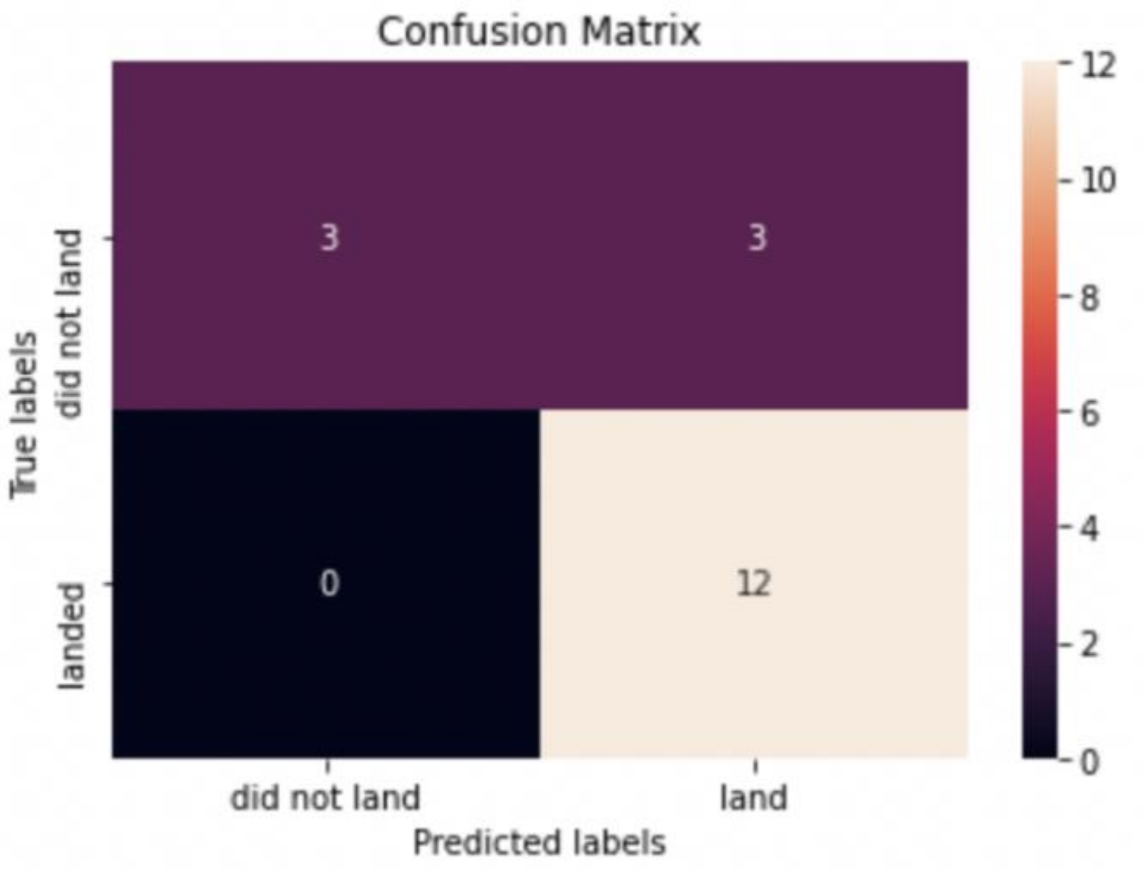


	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Predictive analysis (Classification)

Confusion Matrix

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



CONCLUSION

- 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.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.