IBM Developer
SKILLS NETWORK

Capstone project:
SpaceX Rocket launch analysis

Tanad Lerdbussarakam

04/11/2021

Outline

- Executive summary
- Introduction
- Methodology
- Result
- Conclusion
- Appendix

Executive summary

This project is aim to extracted insight from the SpaceX launch data and analysis to identify the influence factors and launch site for potential competitive commercial space programs. The data was collected through SpaceX API and Web scraping, where data wrangling was done to normalize and pre-process data to the dataframe. The dataset will be explored through exploratory data analysis and interactive map to identify the influence factors and launch site proximity. The classification models were made for further prediction analysis

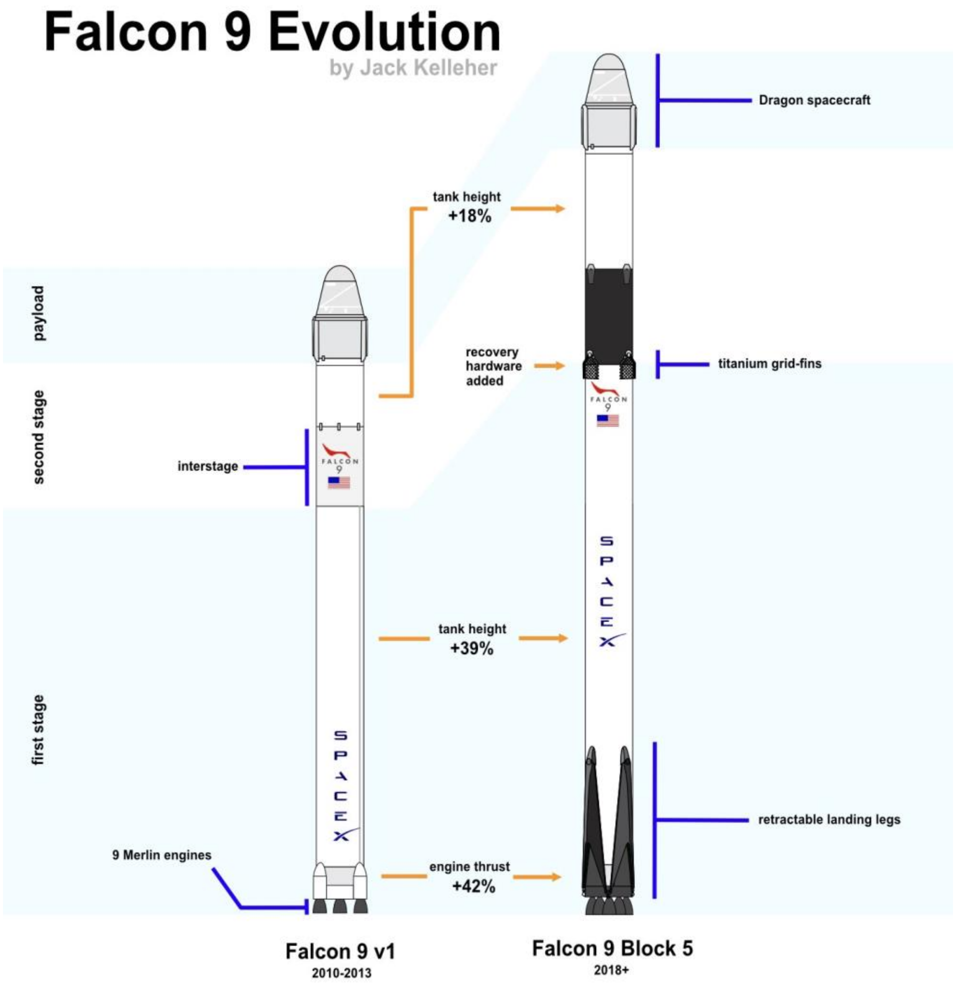
The project outcome is able to provide an insight of suitable Payload mass, Rocket version and orbit that would likely to yield successful. The best prediction model was also identified.

Introduction

Project Background

The current advance of space technology bring the possibility of space traveling become close to realization. SpaceX is currently the most successful in making commercial space traveling more achievable with relatively affordable launching cost of the model Falcon 9 due to the design where the first stage of the rocket could land and be reused.

Therefore, this project will aim to determine the cost of launch based on the first stage landing using the retrievable dataset from SpaceX using machine learning prediction model, which could be used for competitive project bid in commercial space program.



https://medium.com/@JackKelleher

Introduction

Objective

- Determine the major factors which influence the successful chance of the first stage landing such as;
  - Version
  - Orbit
  - Payload mass
- Determine the suitable launch site based successful rate and proximity area
- Determine the best classification model for further prediction analysis

Methodology

Methodology

- Data collection methodology
  - Collecting SpaceX data through Rest API and Web Scraping using BeautifulSoup
- Perform data wrangling
  - Filtering, Clearing void/null, One hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium map
- Perform predictive analysis using classification models
  - Build and tuning between Logistic Regression/ SVM/ Tree/ KNN
  - Evaluate each model through Accuracy, Jaccard Score and F1 score for model selection

GitHub link

Data collection

- From API: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/API%20Lab.ipynb
- From web scraping: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/Web%20Scraping.ipynb

Data wrangling: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/Data%20wrangling.ipynb

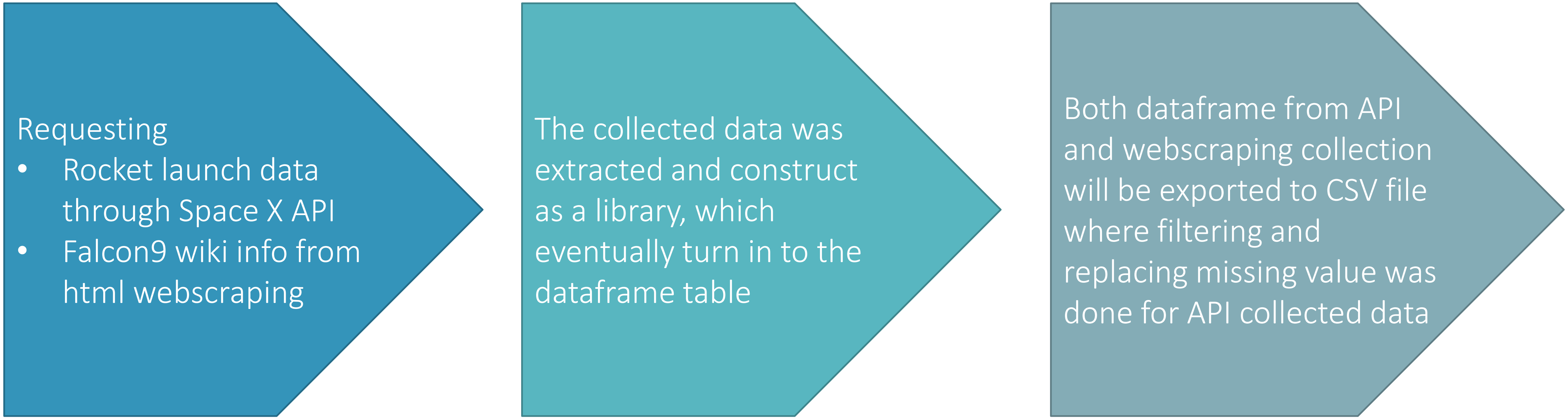
EDA

- Data Visualization: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/EDA%20with%20visual%20lab.ipynb
- SQL: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/EDA%20with%20SQL.ipynb

Folium map: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/Data%20Visualization%20with%20Folium.ipynb

ML classifier: https://github.com/MmIiRrUu/IBM-DS-Capstone-project/blob/b28579cb0bb6173476ea5f22d4d3a3b3003627d7/Machine%20Learning%20Prediction.ipynb

Data collection



Requesting

- Rocket launch data through Space X API
- Falcon9 wiki info from html webscraping

The collected data was extracted and construct as a library, which eventually turn in to the dataframe table

Both dataframe from API and webscraping collection will be exported to CSV file where filtering and replacing missing value was done for API collected data

Data wrangling

- **Exploratory Data analysis (EDA)** was done on the dataset to find the pattern and label if the booster landing is successful or not.
- The current label in the dataset was varied through different scenarios of success and failure such as;
  - TRUE/FALSE Ocean for specific ocean landing
  - TRUE/FALSE RTLS for ground pad landing
  - TRUE/FALSE ASDS for landing on a drone ship
- The main objective of data wrangling is to label simply as binary (1 or 0) whether if the landing is successful

Data wrangling

Data analysis tasks done:

- 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

EDA with Data Visualization

- **Visualize several relationship and trend from the dataset**
  - Flight Number and Launch Site (Scatter plot)
  - Payload and Launch Site (Scatter plot)
  - Success rate of each orbit type (Bar plot)
  - Flight Number and Orbit type (Scatter plot)
  - Payload and Orbit type (Scatter plot)
  - The launch success yearly trend (Line plot)
- **Select the features that will be used in success prediction through features engineering**
  - Use the function get_dummies and features dataframe to apply OneHotEncoder
  - Cast the entire dataframe to variable type float64

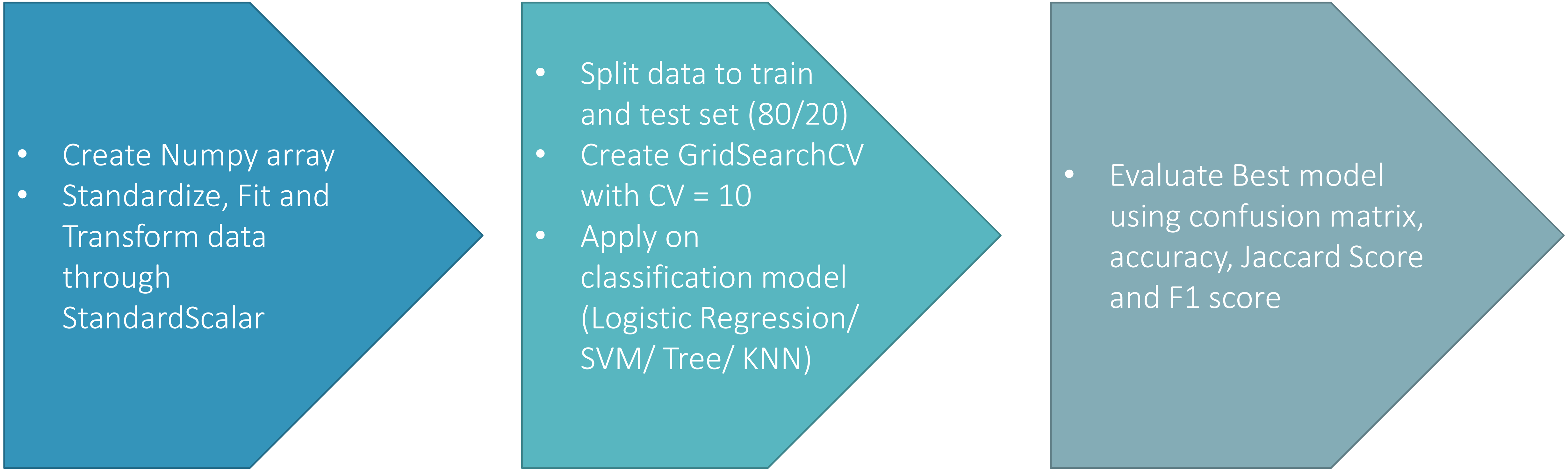
EDA with SQL

- Using Database **IBM DB2** for SQL Queries
- List of information extracted using SQL queries
  - The names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA’
  - The total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date when the first successful landing outcome in ground pad was achieved
  - The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - The total number of successful and failure mission outcomes
  - The names of the booster versions which have carried the maximum payload mass.
  - The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

Interactive map with Folium

- Mark all launch sites on a map
  - Create a folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
  - Add a circle for each launch site in data frame
- Mark the success/failed launches for each site on the map
  - Adding the launch outcomes for each site, and see which sites have high success rates
  - Create markers for all launch records through MarkerCluster
- Calculate the distances between a launch site to its proximities
  - Get coordinate for a mouse over a point on the map
  - Calculate the distance between two points on the map based on their Latitude and Longitude values
  - Create a folium.Marker to show the distance and draw a PolyLine between a launch site to the selected point(Coastline, City, Railway, Highway)

Predictive analysis (Classification)



- Create Numpy array
- Standardize, Fit and Transform data through StandardScalar

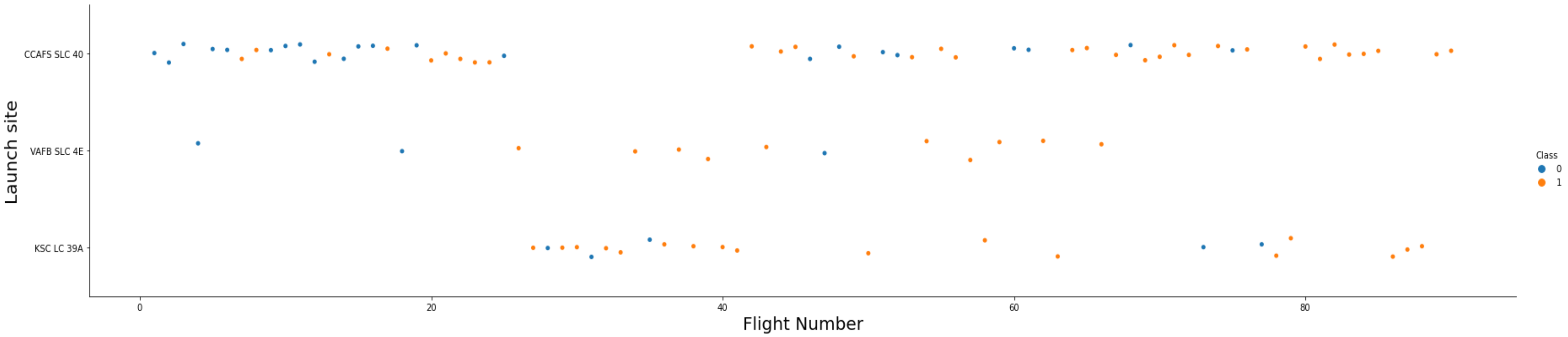
- Split data to train and test set (80/20)
- Create GridSearchCV with CV = 10
- Apply on classification model (Logistic Regression/ SVM/ Tree/ KNN)

- Evaluate Best model using confusion matrix, accuracy, Jaccard Score and F1 score

IBM Developer
SKILLS NETWORK

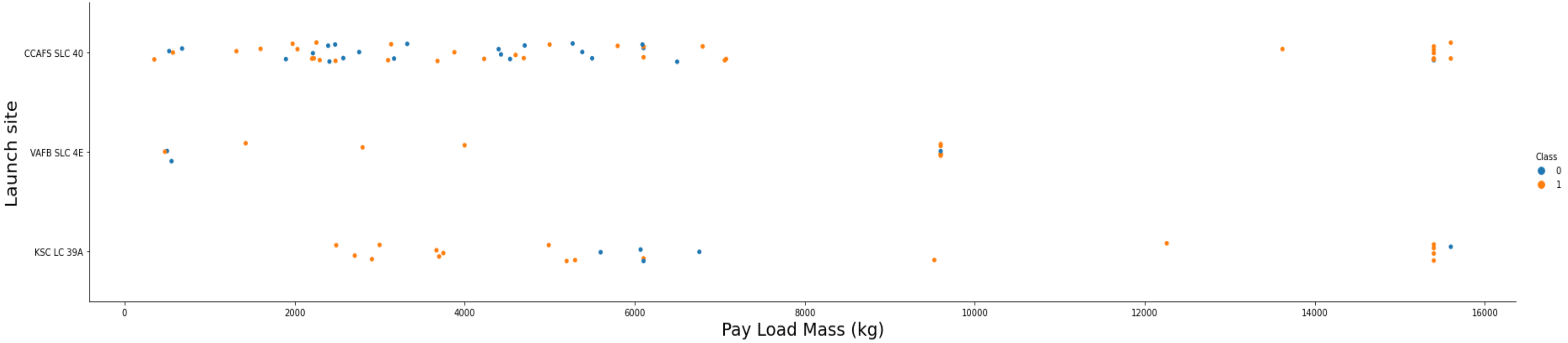
EDA on dataset through visualization tools: Result & Discussion

Launch site vs Flight number



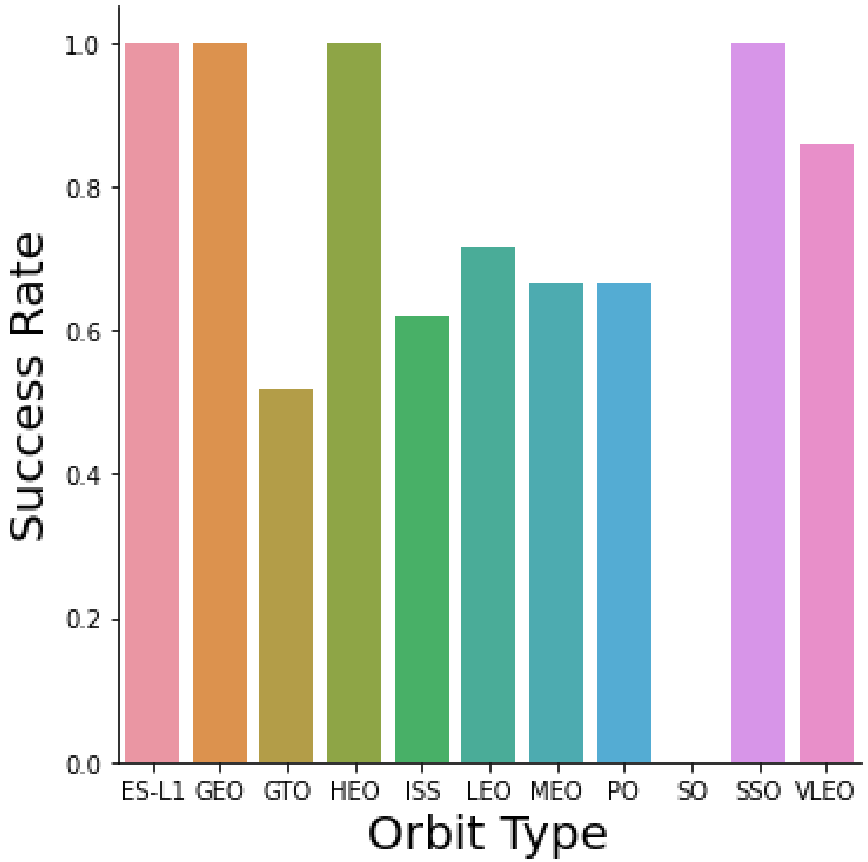
- The earlier launch had less successful rate while there are relatively more successful launch in the later flight
- CCAFS SLC 40 have been majorly used as a launch site but other launch site have more successful rate

Launch site vs Payload Mass



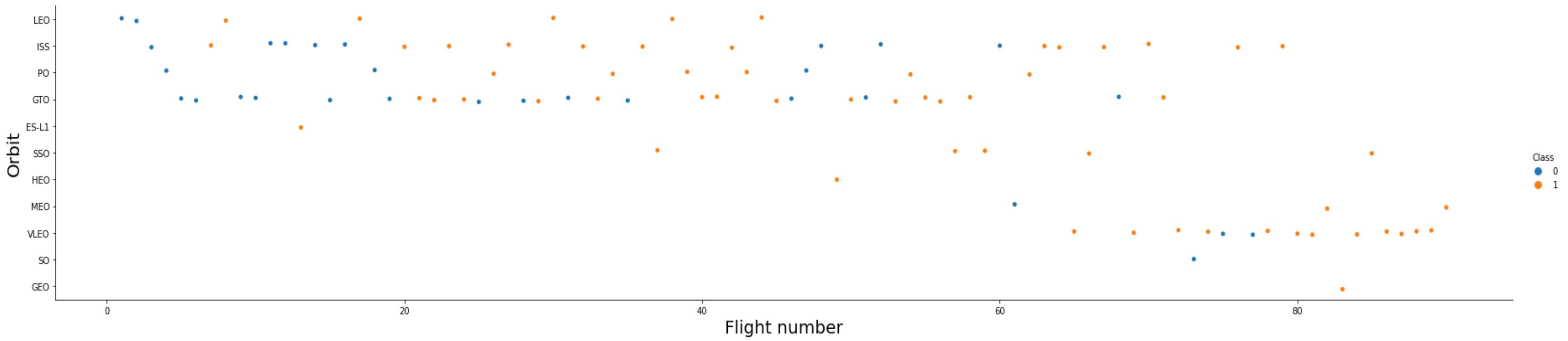
- Most launch have been done at Pay Load Mass below 8000 kg
- Most of successful launce use Pay Load Mass more than 7000 kg

Success rate for each orbit type



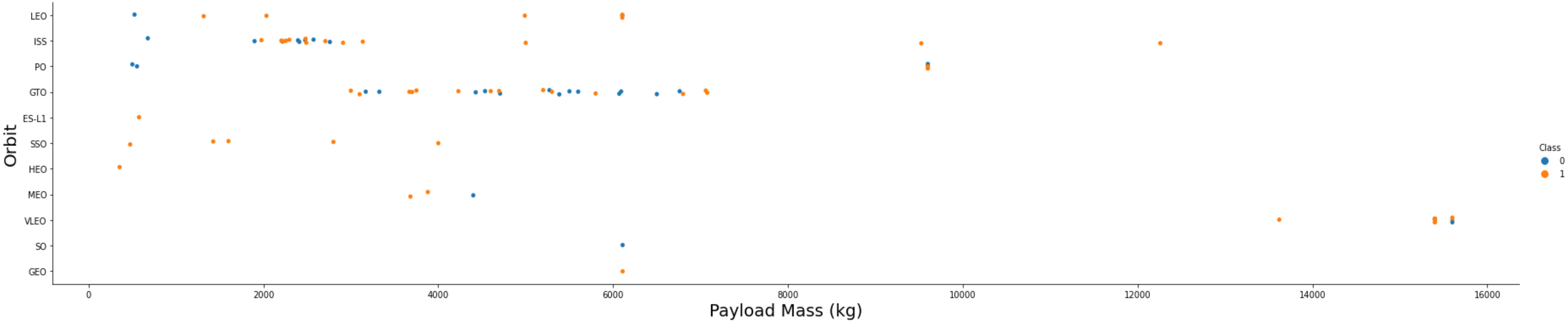
- ES-L1, GEO, HEO, SSO have the highest success rate among every orbit.
- No successful launch shows for orbit SO

Orbit vs Flight number



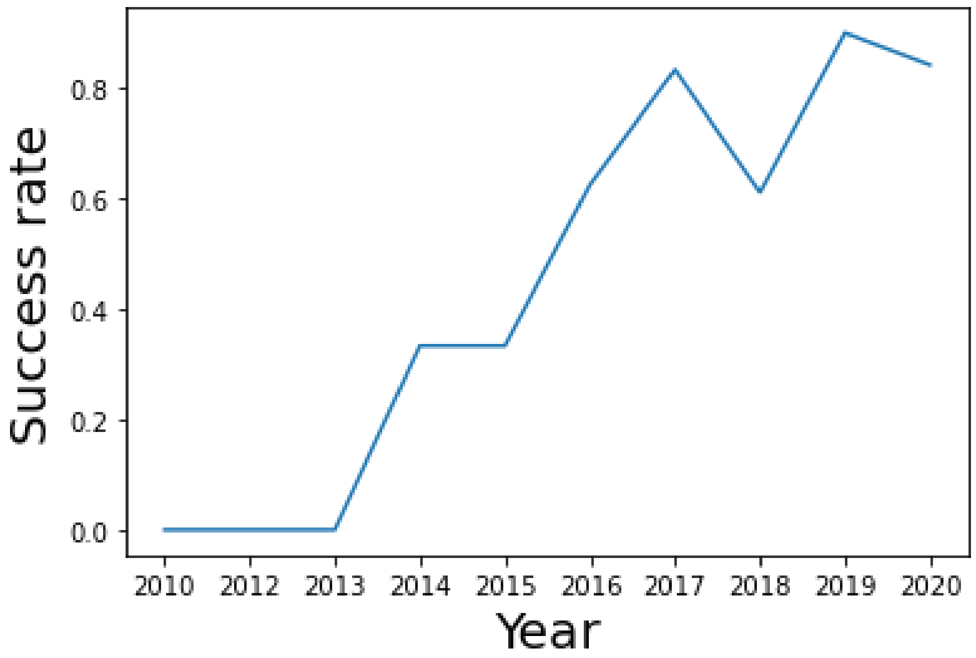
- LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- ISS, VLEO orbit have the most successful launch since the flight number 20, which indicate favorable orbit for rocket launch especially VLEO which often in the latest launch.

Orbit vs Payload mass



- With Payload Mass under 3000, the successful landing or positive landing rate are more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there

Success rate from 2010 - 2020

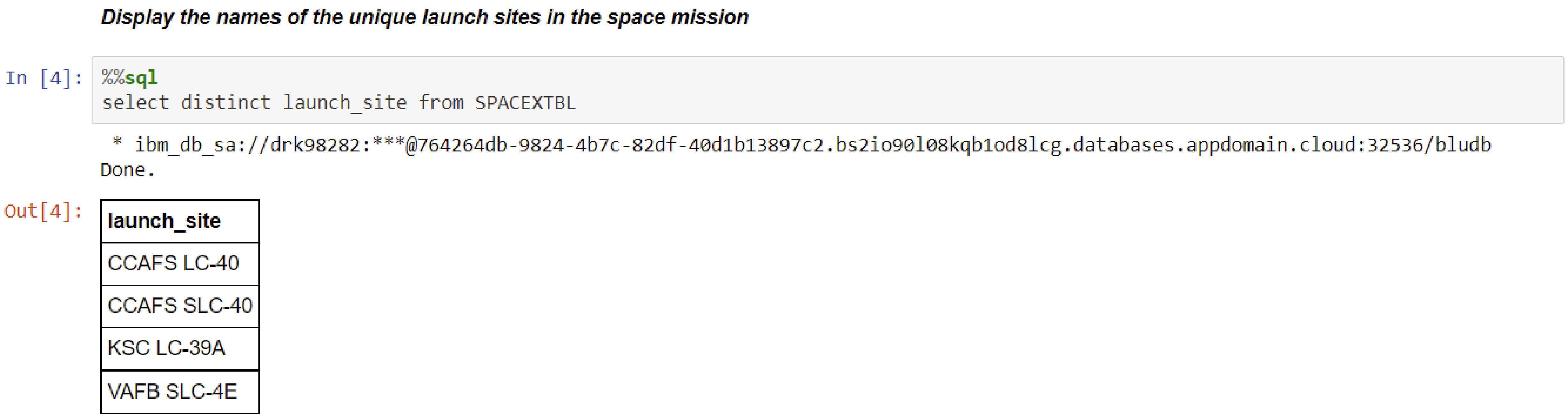


- The success rate since 2013 kept increasing till 2017
- Sharp decline found between 2017-2018 and keep increasing afterward

IBM Developer
SKILLS NETWORK

EDA on dataset through SQL: Result & Discussion

Exploratory insight from SQL queries



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

In [4]:

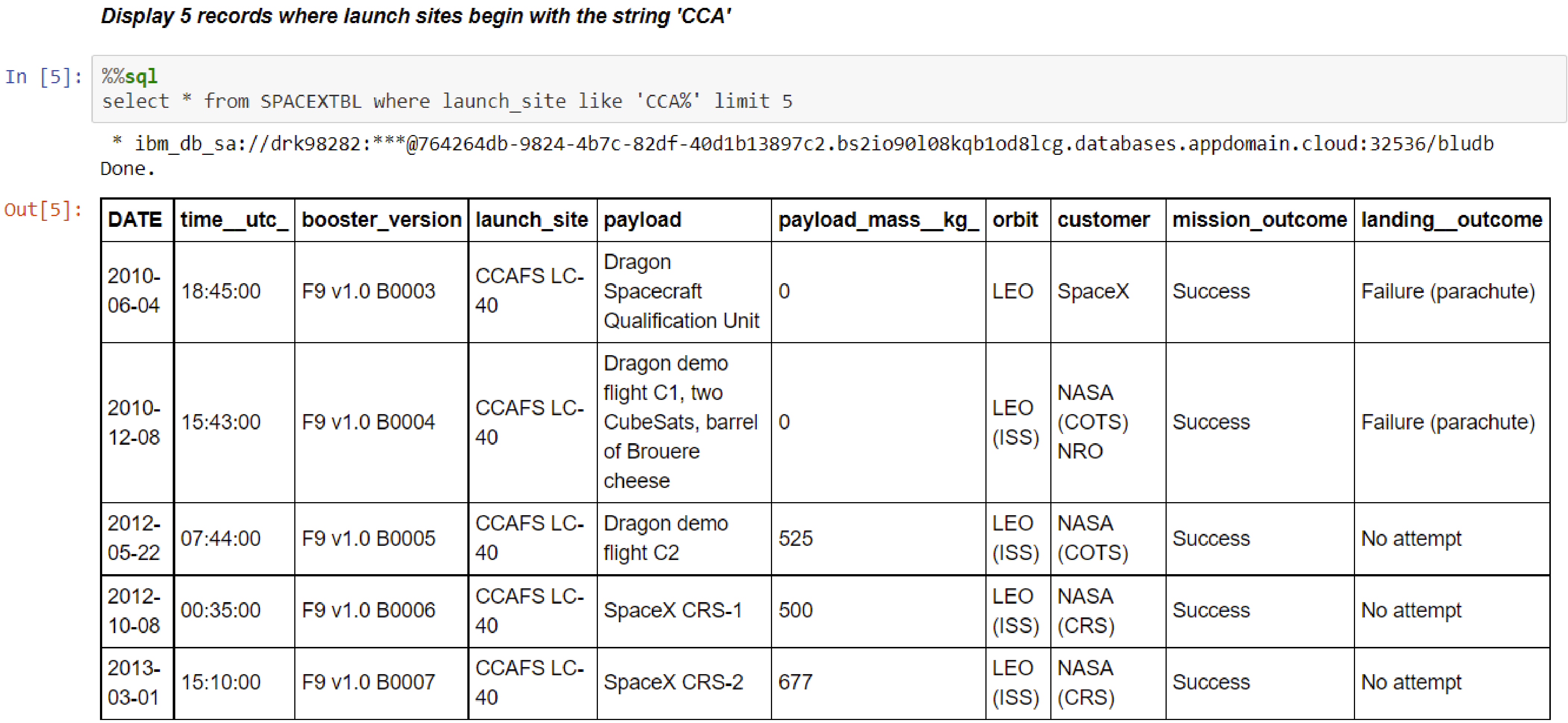
```
%%sql
select distinct launch_site from SPACEXTBL
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[4]:

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

Exploratory insight from SQL queries



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

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

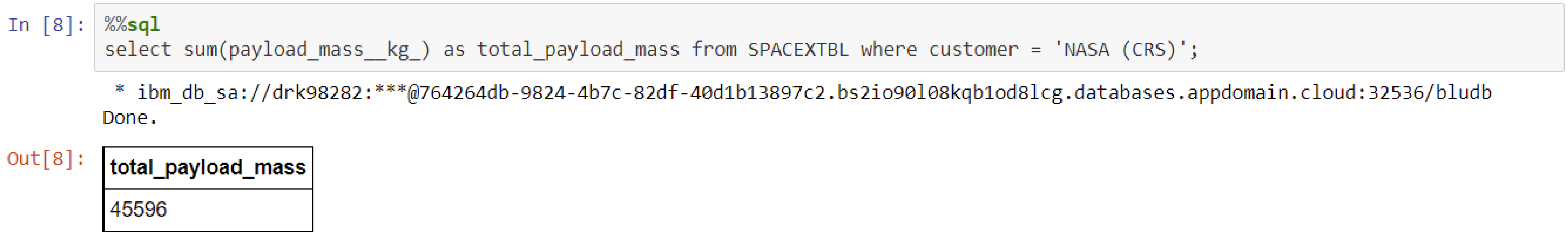
```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

Exploratory insight from SQL queries

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



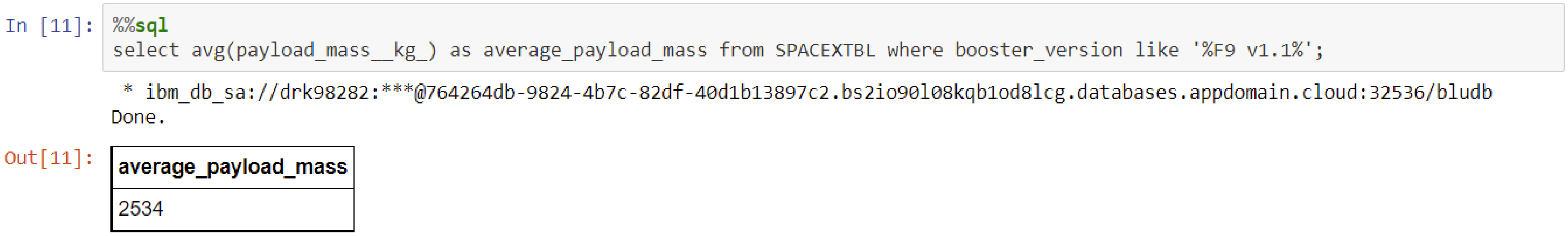
```
In [8]: %%sql
        select sum(payload_mass__kg_) as total_payload_mass from SPACEXTBL where customer = 'NASA (CRS)';

 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[8]:

total_payload_mass
45596

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



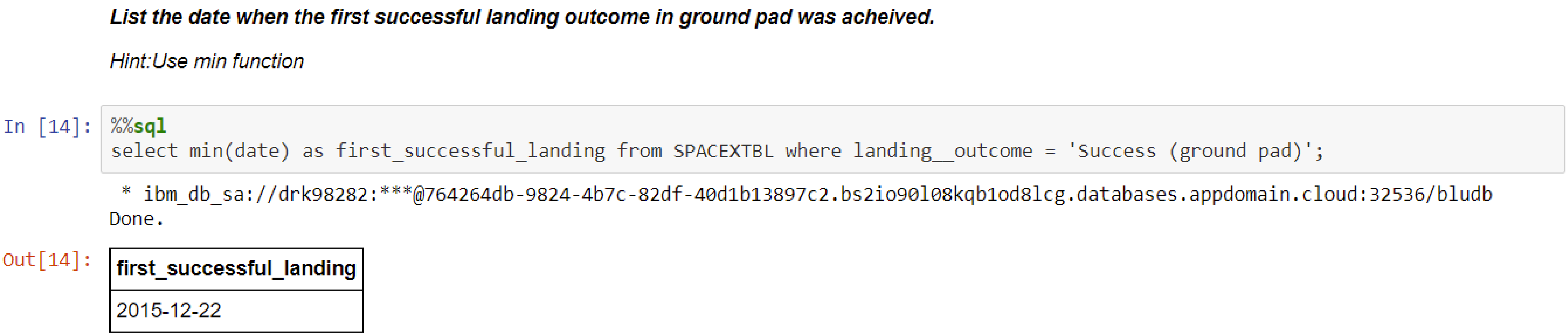
```
In [11]: %%sql
         select avg(payload_mass__kg_) as average_payload_mass from SPACEXTBL where booster_version like '%F9 v1.1%';

 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[11]:

average_payload_mass
2534

Exploratory insight from SQL queries



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

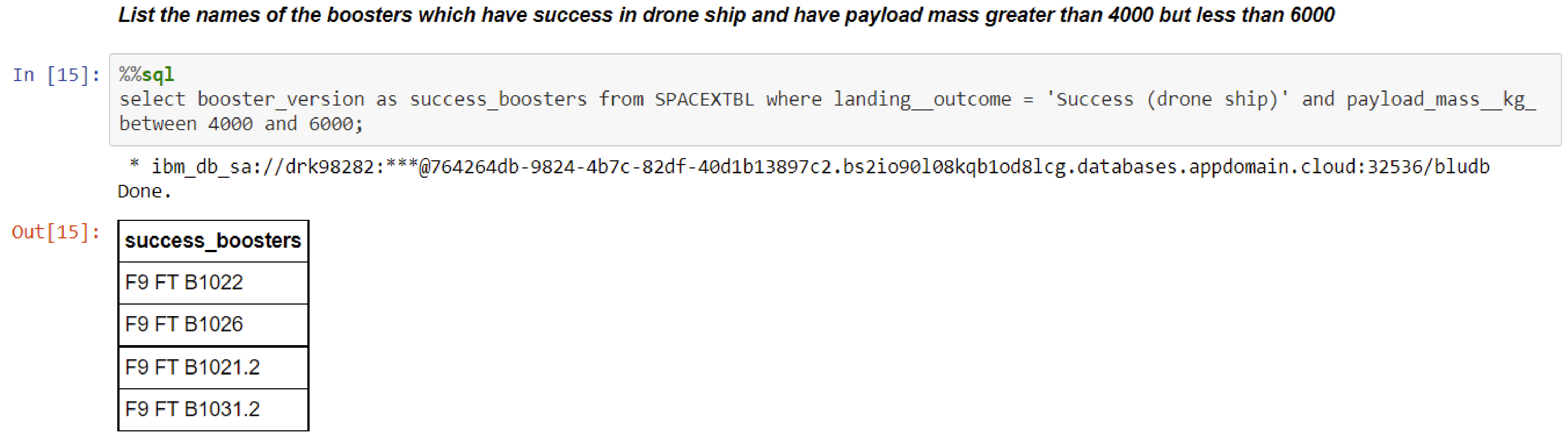
Hint:Use min function

```
In [14]: %%sql
select min(date) as first_successful_landing from SPACEXTBL where landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[14]:

first_successful_landing
2015-12-22



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

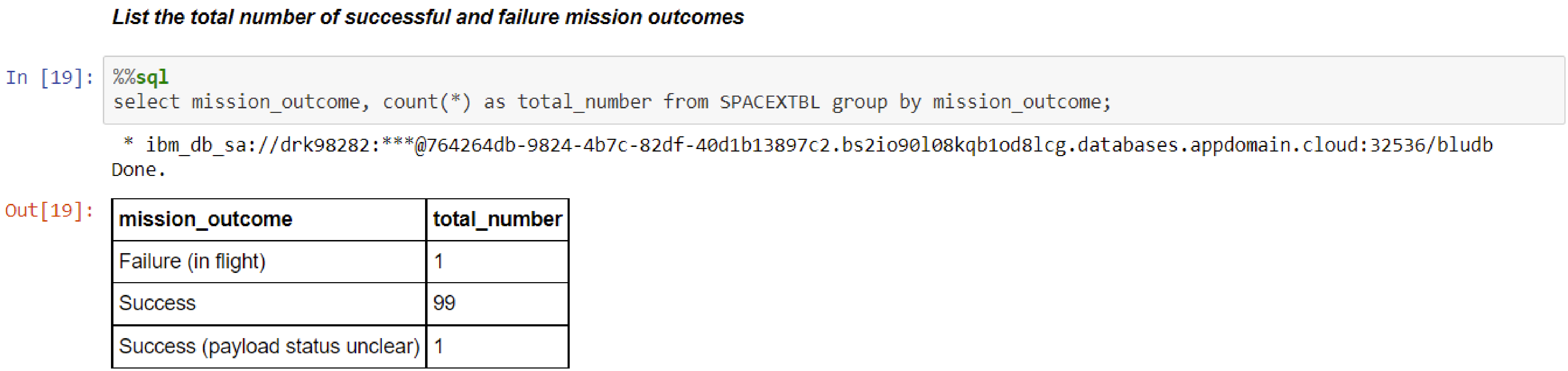
```
In [15]: %%sql
select booster_version as success_boosters from SPACEXTBL where landing__outcome = 'Success (drone ship)' and payload_mass__kg_
between 4000 and 6000;
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[15]:

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

Exploratory insight from SQL queries



List the total number of successful and failure mission outcomes

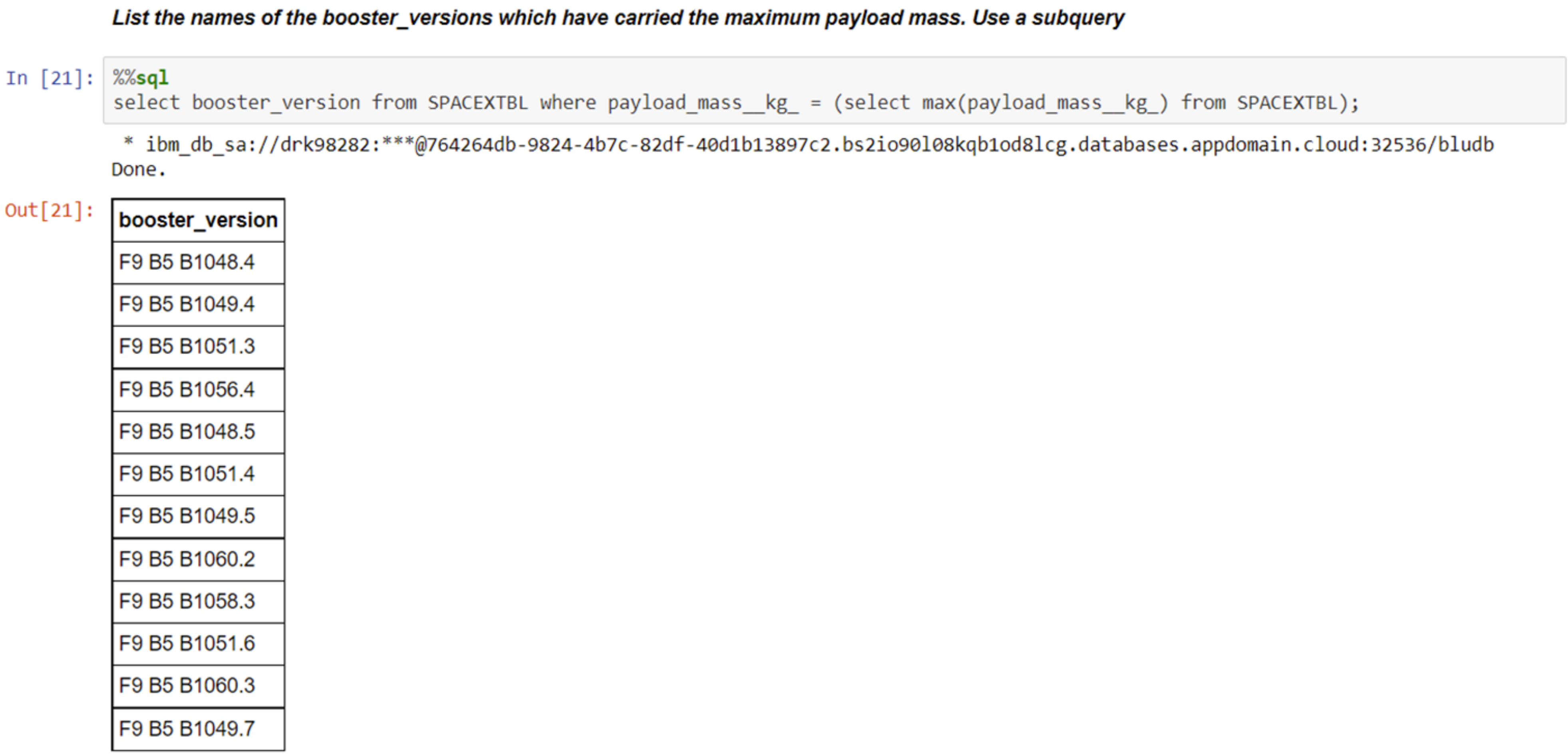
```
In [19]: %%sql
         select mission_outcome, count(*) as total_number from SPACEXTBL group by mission_outcome;
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[19]:

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

Exploratory insight from SQL queries



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

In [21]:

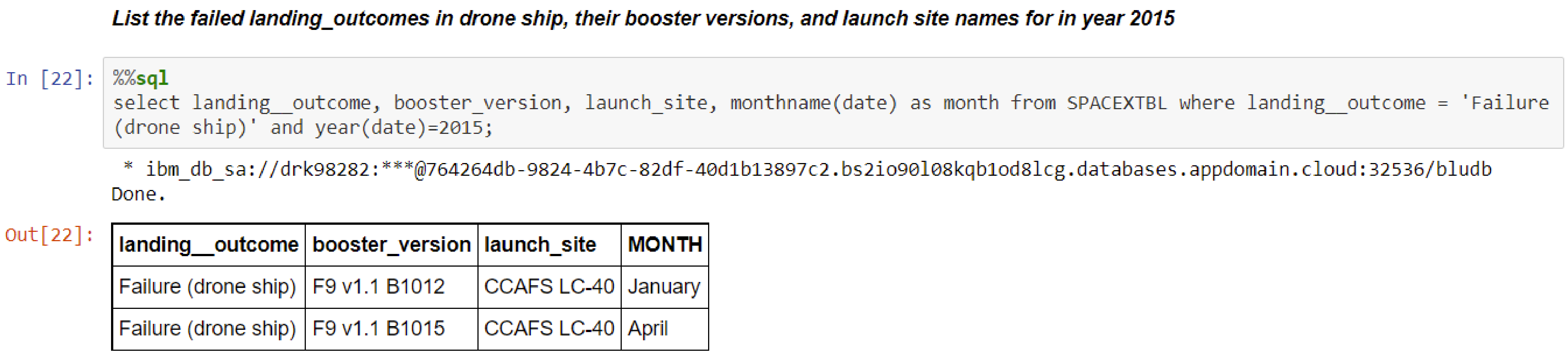
```
%%sql
select booster_version from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[21]:

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

Exploratory insight from SQL queries



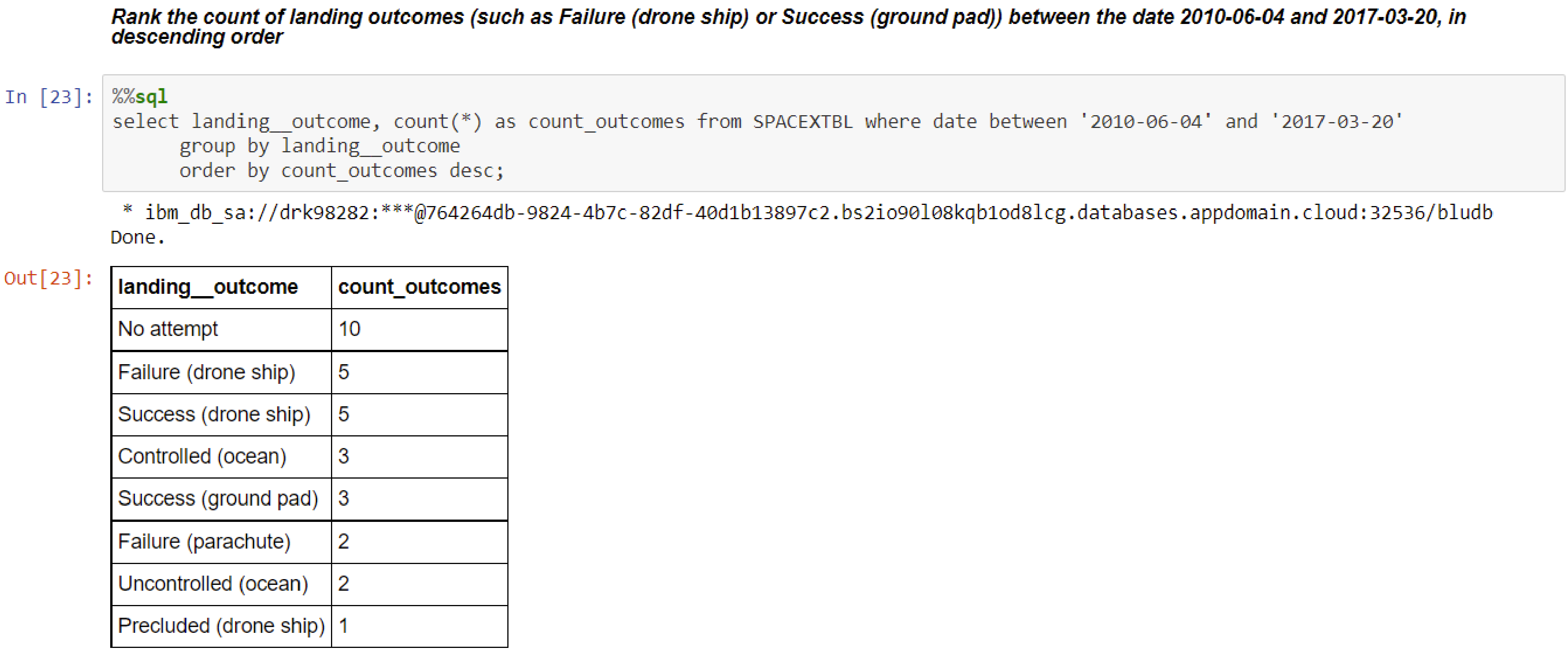
List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%%sql
select landing__outcome, booster_version, launch_site, monthname(date) as month from SPACEXTBL where landing__outcome = 'Failure
(drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

Exploratory insight from SQL queries



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

In [23]:

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

```
 * ibm_db_sa://drk98282:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

Out[23]:

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

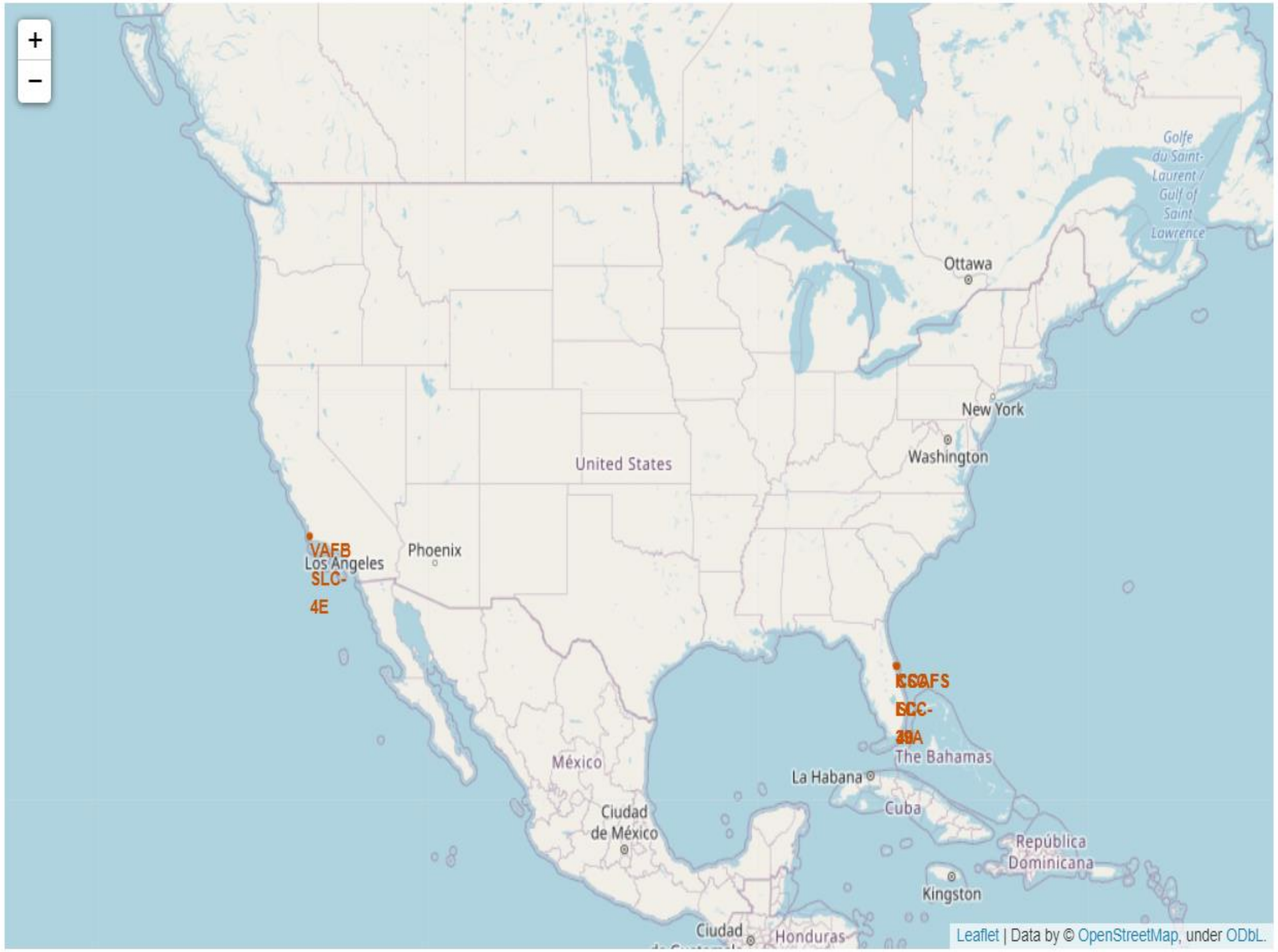
Exploratory insight from SQL queries

- Overall the query result is represent as per given tasks in methodology session, which indicate that capability to explore information from the dataset
- Based on mission outcome data, almost every missions have been successful, contrary to the success rate in the first stage landing
- There are 4 distinct launching site and 4 distinct successful booster version
- First success landing was from 2015

IBM Developer
SKILLS NETWORK

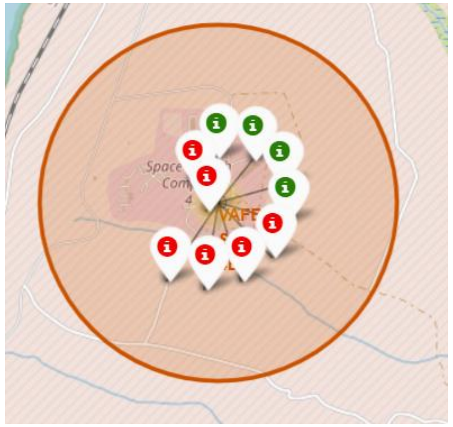
Insight from Interactive map with Folium: Result & Discussion

Interactive map – Launching site

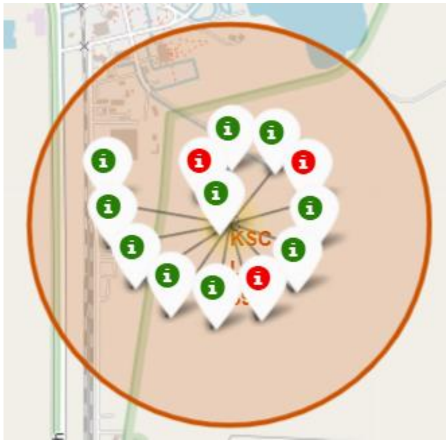


- Three of the launch is near the coastline of Florida and one in Los Angelis, which all located near the equator line as it is the location which Earth rotate the fastest. The inertia create from Earth rotation speed would help propelling the rocket
- The finding suggest that the launch site proximity is likely to close to the coastline due risk of falling debris from launching

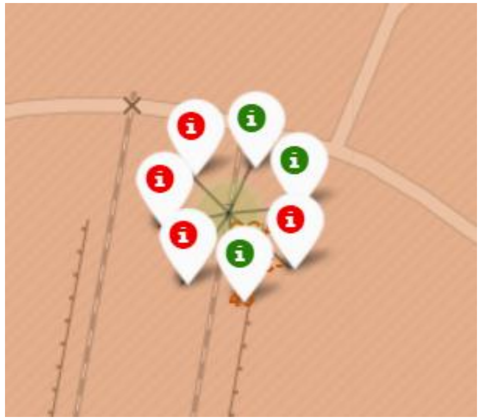
Interactive map – Successful rate at the launch site



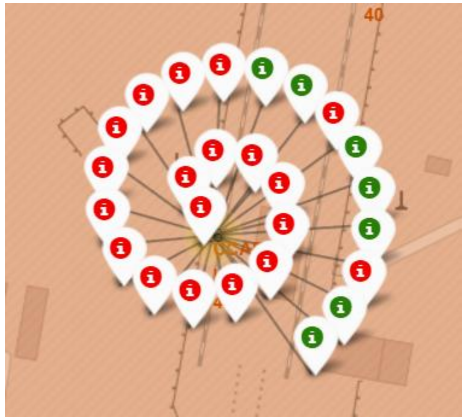
VAFB SLC 4E (LA)



KSC LC 39A (FL)



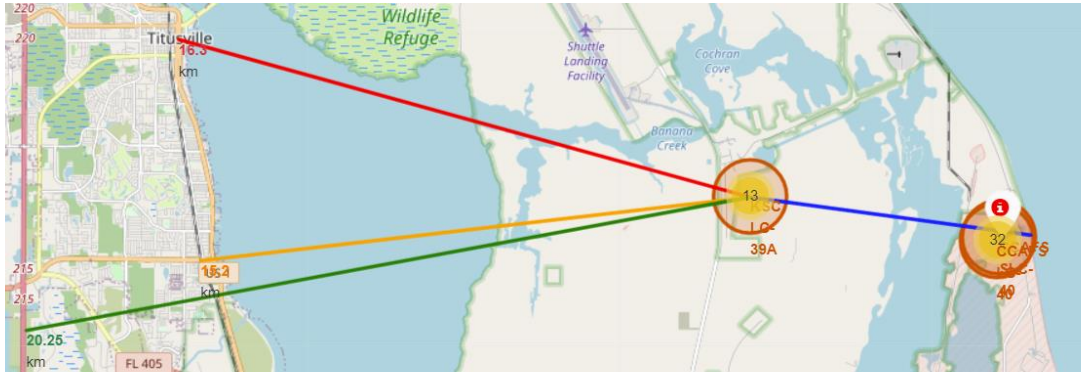
CCAFS SLC 40 (FL)



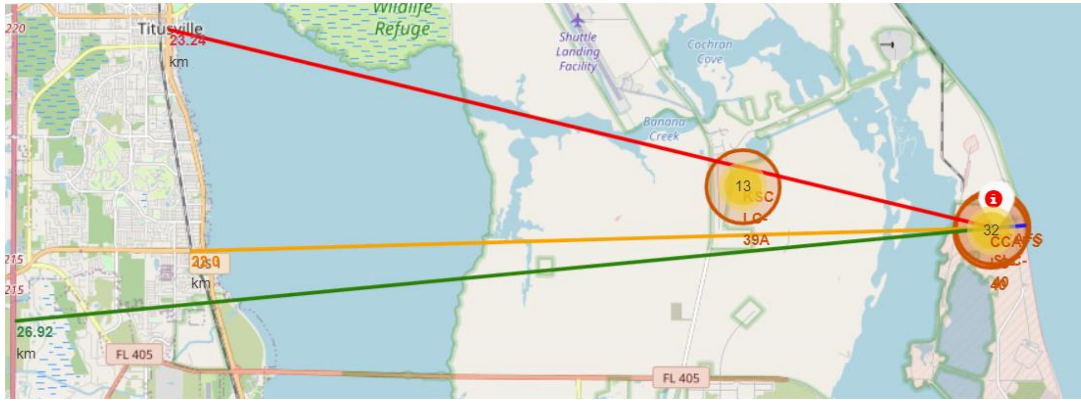
CCAFS LC 40 (FL)

- The finding shows that the launch site KSC LC 39A have the highest successful launch relatively to other launch sites, which indicate a favorable launch site in case of the next space rocket launch.
- CCAFS LC 40 have the most launch attempt compare to other launch sites, which could explore more to which factor could influence more activity in this site

Interactive map – The launch site's proximity



KSC LC 39A (FL)

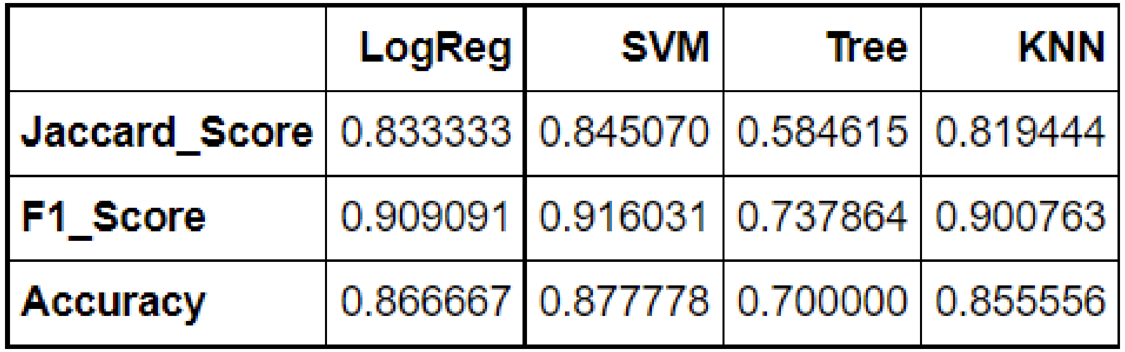


CCAFS LC 40 (FL)

- By checking the proximity between launch site and surrounding (Coastline, City, Railway, Highway), we found that KSC LC 39A is located relatively closer to the city but further to the coastline. While CCAFS LC 40 located closer to the coastline and further to the city
- This would explain the preference on using CCAFS LC 40 as the launch site over KSC LC 39A despite the lower successful rate.

Predictive analysis: Result & Discussion

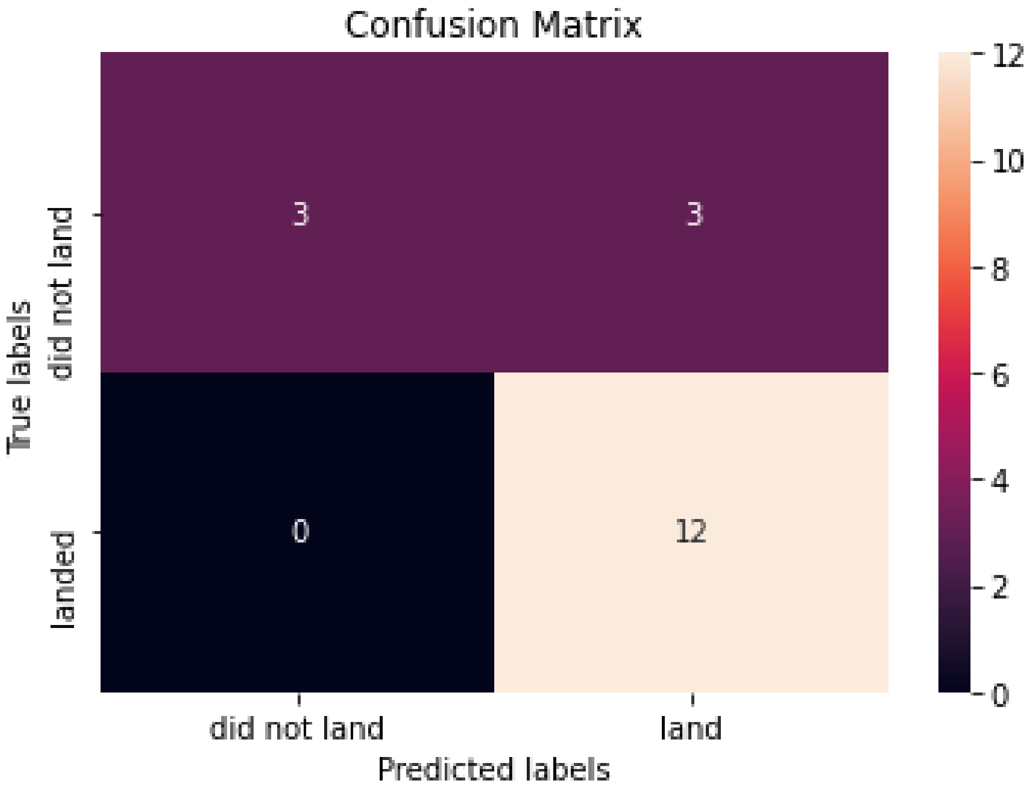
Evaluation score



	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.584615	0.819444
F1_Score	0.909091	0.916031	0.737864	0.900763
Accuracy	0.866667	0.877778	0.700000	0.855556

- From the evaluation in F1 score, Jaccard score and accuracy of the model, SVM have the highest score among all the classification model while Logistic Regression (LogReg) and KNN Model are very close in all accuracy compare to SVM
- Tree classifier have scored significantly poorer compare to all the model.

Confusion Matrix



- Confusion matrix for all models is quite similar with 15 out of 18 samples have been predicted correctly (True positive, True Negative)
- There are 3 out of 18 samples have been false positive, which could be the defection on all the current model

IBM Developer
SKILLS NETWORK

Conclusion

Conclusion

- Some influence factors for successful rocket launch mission was identified
  - The latest launch have more successful rate compare to the earlier launch, indicate trend successful over the time
  - Payload Mass at the range 6000 – 7000 kg or over would likely to be more successful
  - VLEO, LEO shown to be successful with the flight number, especially VLEO for the recent successful launch
- The launch site CCAFS LC 40 is more suitable choice due to the proximity while higher successful KSC LC 39A could be used for most of the low risk test or used further when successful rate at CCAFS LC 40 is increased.
- SVM model is the current best model that could be used for further prediction analysis

IBM Developer
SKILLS NETWORK

Thank you