IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Justin Chan>
<23 May 2022>

Outline

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

Executive Summary

- Target: To determine if a first stage rocket could be landed with reference to SpaceX data
- Summary of methodologies
    - Data is collected from SpaceX API and Web Scrapping from Wikipedia
    - Data wrangling and cleaning was performed
    - Exploratory Data Analysis was performed, in the form of SQL and data visualization
    - Visual analytic with Folium and dash was performed
    - Machine learning was deployed to predict if the first stage will land, comparison between models were made

Executive Summary

- Summary of all results
  - First stage landing success rate has improved in recent years, reaching ~80% in recent years
  - Higher pay load mass often results in a higher success rate in landing missions
  - Launch site KSC LC-39A has the highest success rate in launching
  - Various classification models demonstrated similar accuracy, this may due to the lack of test sample (total sample size was 90, with 18 used as test sample)
    - The accuracy might be improved by adopting k-fold cross validation to the sample set
  - Company may consider launching at KSC LC-39A, with a payload mass below 5000kg as a start.

Introduction

- Project background and context
  - SpaceX has a much lower cost in launching rockets compared to other companies, and the savings is because SpaceX can reuse the first stage rockets
  - In order to determine the cost of each launch, data were gathered, including rocket types, locations, orbit type etc., to determine if the first stage will land
- Problems you want to find answers
  - What are the factors affecting the successful rate of first stage landing?
  - Which launch site has the best successful rate?
  - Could a landing outcome be predicted with the help of machine learning?
  - How could a company optimize its successful rate in landing missions?

Section 1

Methodology

Methodology

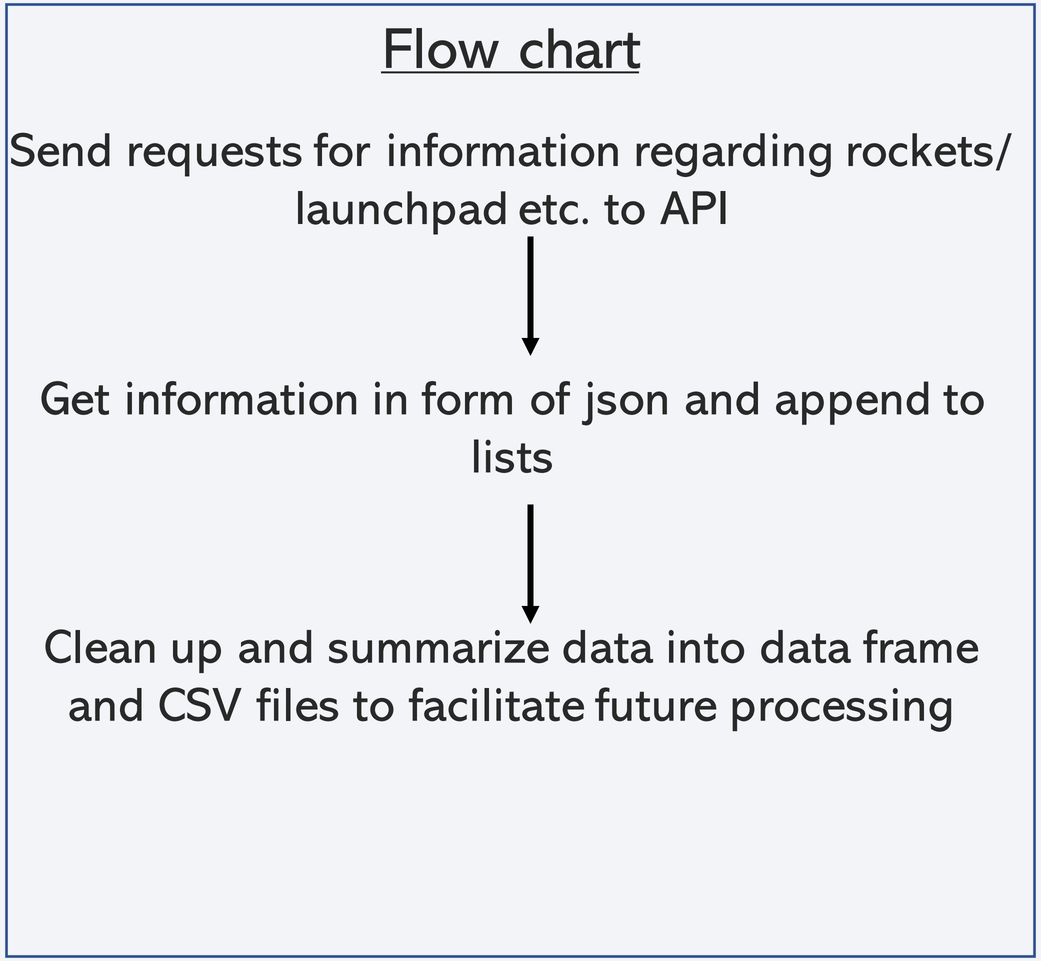
Executive Summary

- Data collection methodology:
  - Data was collected from SpaceX API and through web scraping on Wikipedia
- Perform data wrangling
  - Data cleaning was performed on raw data, and mission outcome is classified as either successful or fail
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After standardizing the data, the landing results are separated into train/test group and being fitted by various classification models for predictive analysis

Data Collection – SpaceX API

- Data were collected with SpaceX REST calls, and the correponding json files of interest were downloaded and processed, flow chat on the right

- The link to Github notebook could be found here.



Flow chart

Send requests for information regarding rockets/ launchpad etc. to API

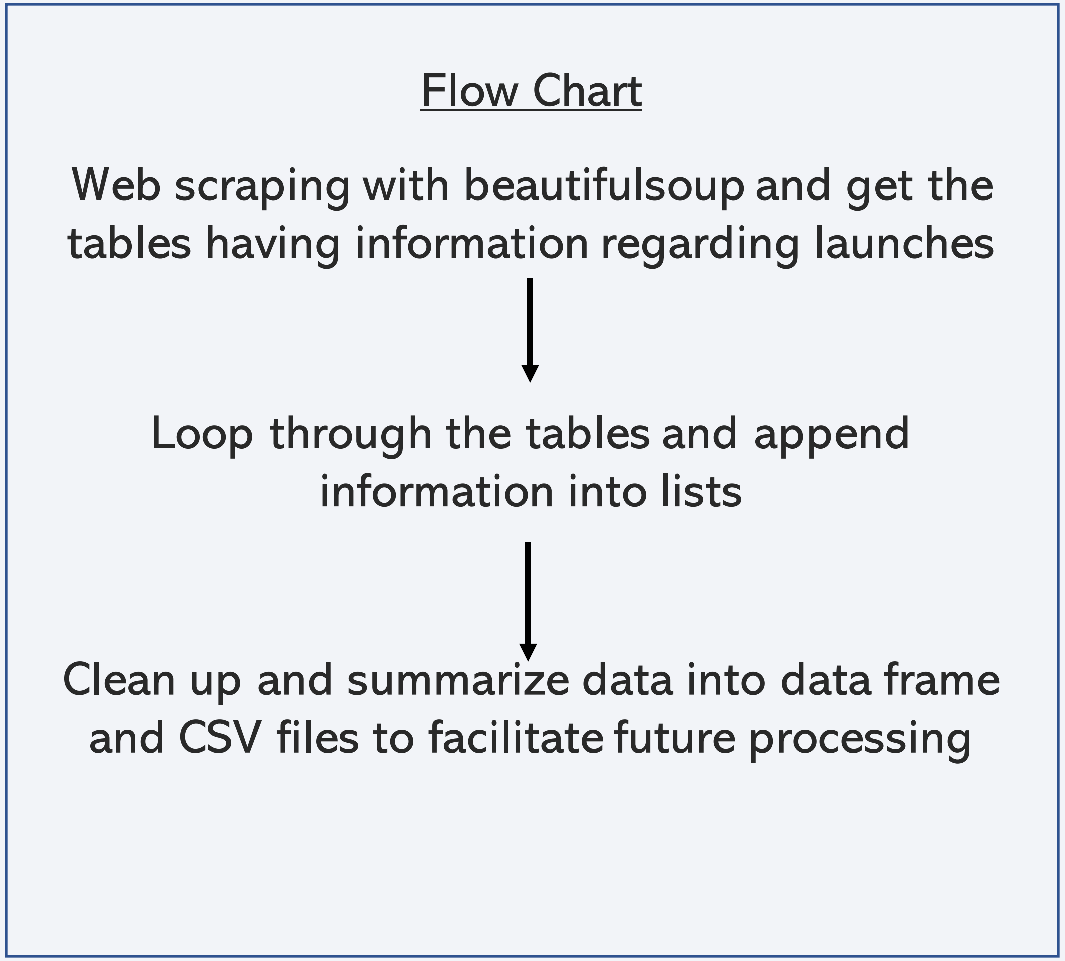
Get information in form of json and append to lists

Clean up and summarize data into data frame and CSV files to facilitate future processing

Data Collection - Scraping

- Web scrapping from Wikipedia with beautifulsoup was processed, flow chart on the right

- The link to Github notebook could be found here.



Flow Chart

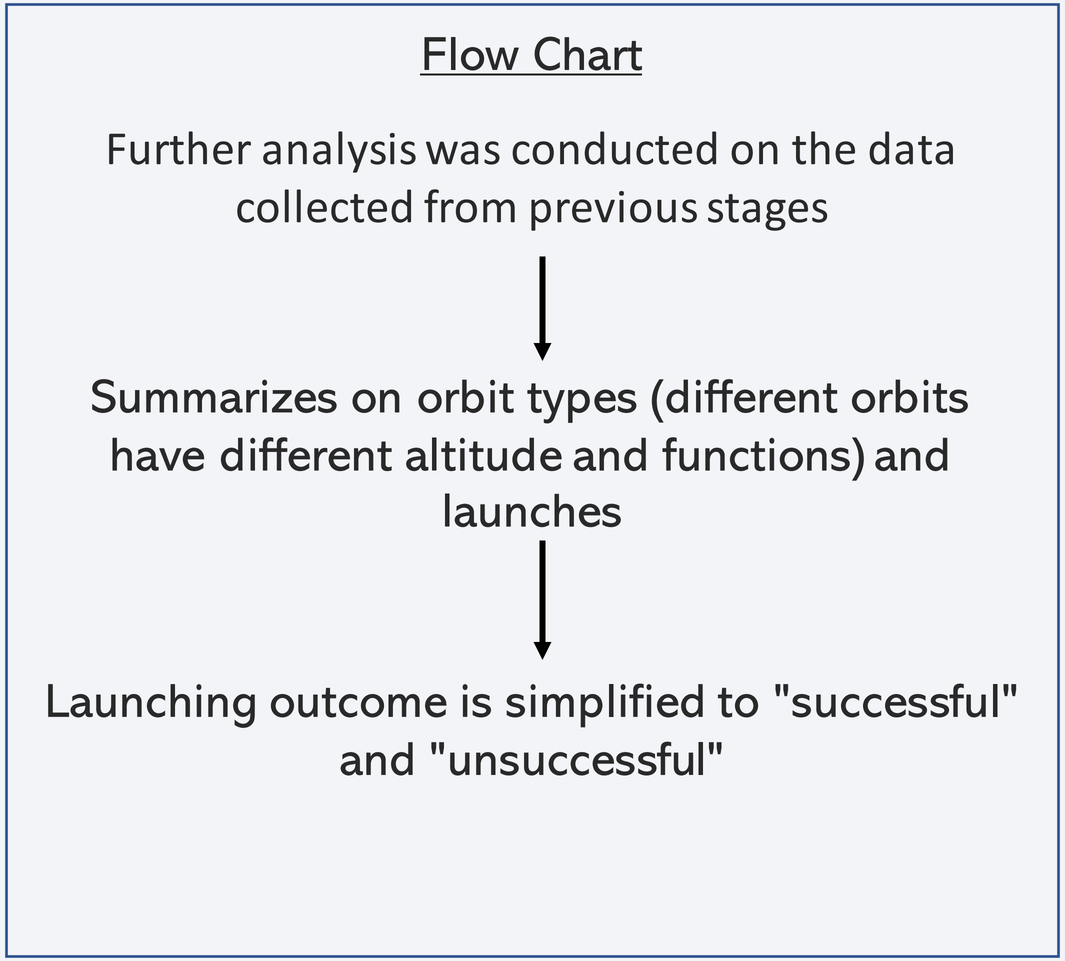
Web scraping with beautifulsoup and get the tables having information regarding launches

Loop through the tables and append information into lists

Clean up and summarize data into data frame and CSV files to facilitate future processing

Data Wrangling

- Further analysis was conducted on the data collected
- Mission outcomes was simplified from categorical to binary
- The link to Github notebook could be found here.



Flow Chart

Further analysis was conducted on the data collected from previous stages

Summarizes on orbit types (different orbits have different altitude and functions) and launches

Launching outcome is simplified to "successful" and "unsuccessful"

EDA with Data Visualization

- To better understand the relation between factors, especially on orbits, flight number, launch site, payload and time, following charts were plotted
  - Relationship between Flight Number and Launch Site
  - Relationship between Payload and Launch Site¶
  - Relationship between success rate of each orbit type
  - Relationship between FlightNumber and Orbit type
  - Relationship between Payload and Orbit type
  - Launch success rate yearly trend

The code and visualization could be found in Github notebook here.

EDA with SQL

The following SQL were performed

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. 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

The code and details could be found in the GitHub link here.

Build an Interactive Map with Folium

- Launch site locations, with object circles and markers were added to the map to visualize the locations of the launch sites
- Launch outcome were also added as marker cluster to see which sites have high success rates
- Mouse position object was added to identify coordinates of surrounding facilities, like railway, highway and coast
- Distance between the launch sites and surrounding facilities were added as line object, with marker object indicating the distance. This is to check if these facilities are in close proximity to the launch site for supply.

The code and details could be found in Github link here.

Build a Dashboard with Plotly Dash

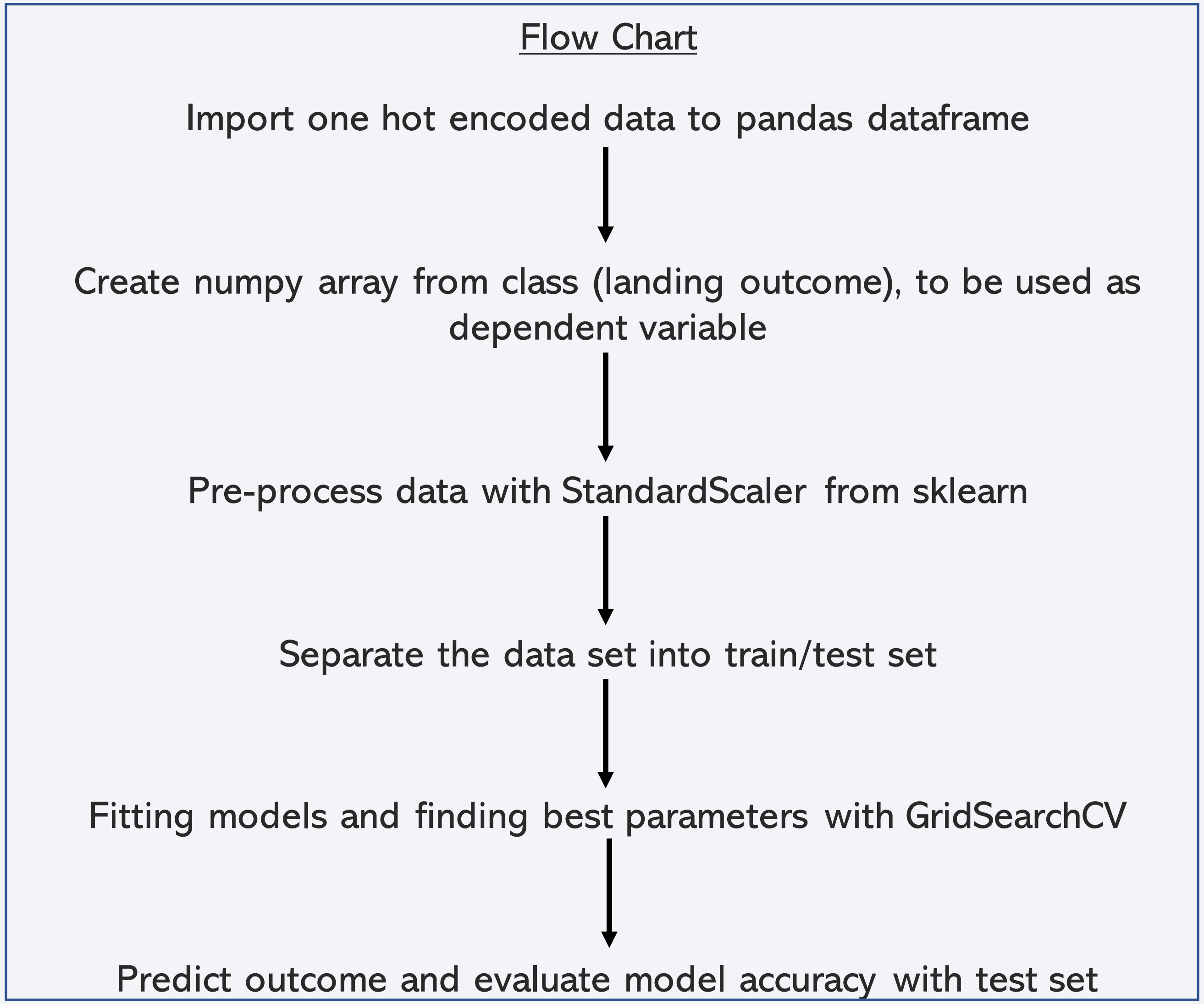
- Pie charts showing statistic of successful/failed launch in difference locations were plotted, to visualize the selected launch site(s) successful rate
- Scattered chart with slider adjusting the pay load mass was made, to visualize how payload may be correlated with mission outcomes for selected site(s)

The code and details could be found in Github link here.

Predictive Analysis (Classification)

- Predictive models were built, primarily with library pandas and sklearn
- Best parameters from a list were decided with the aid of GridSearchCV
- Flow chart shown on the right

The code and details could be found in the Github link here.



Flow Chart

Import one hot encoded data to pandas dataframe

Create numpy array from class (landing outcome), to be used as dependent variable

Pre-process data with StandardScaler from sklearn

Separate the data set into train/test set

Fitting models and finding best parameters with GridSearchCV

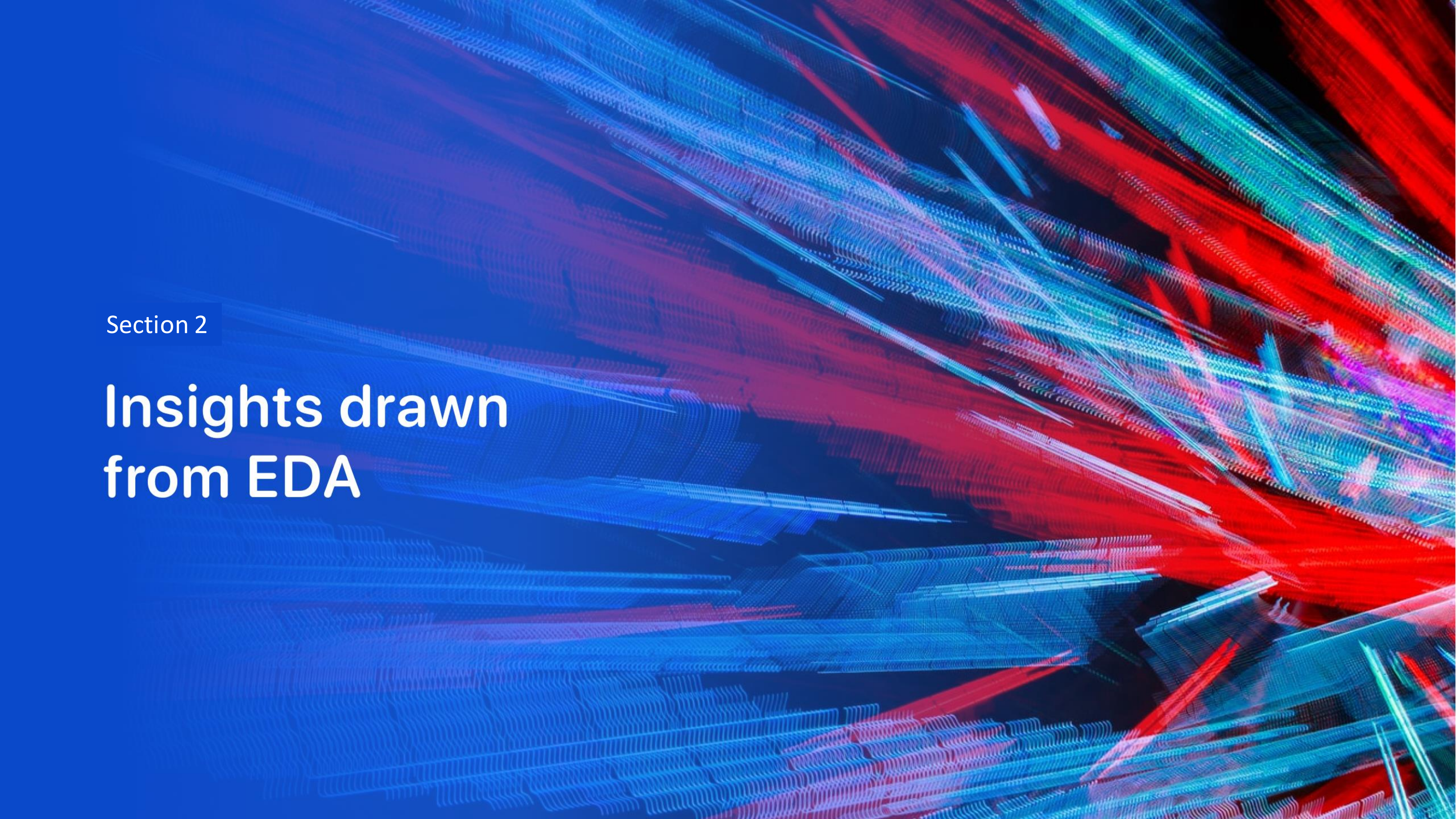
Predict outcome and evaluate model accuracy with test set

Results

- Following results are presented in details in upcoming session
    - Exploratory data analysis results
    - Interactive analytics demo in screenshots
    - Predictive analysis results

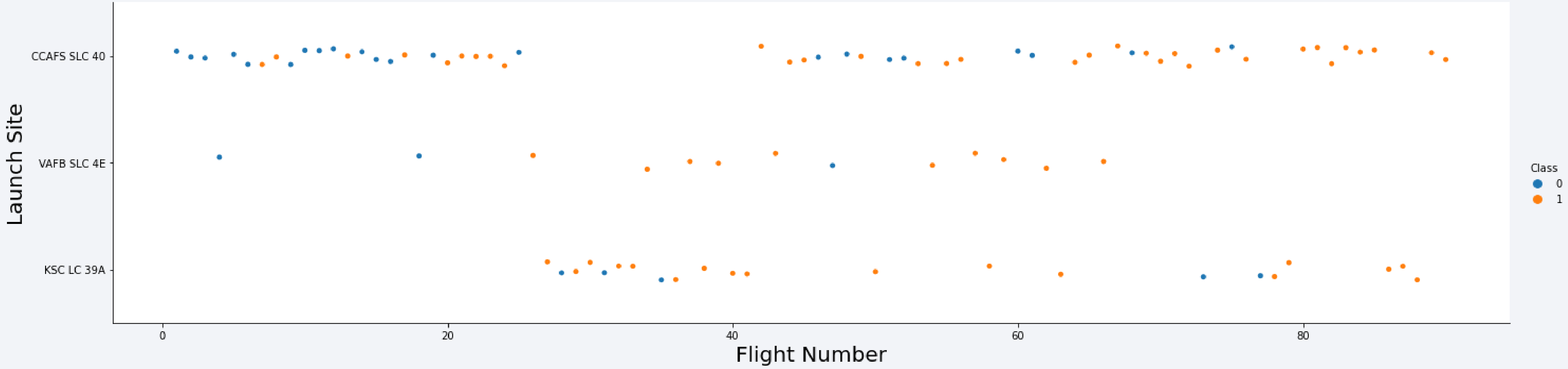
Section 2

Insights drawn from EDA



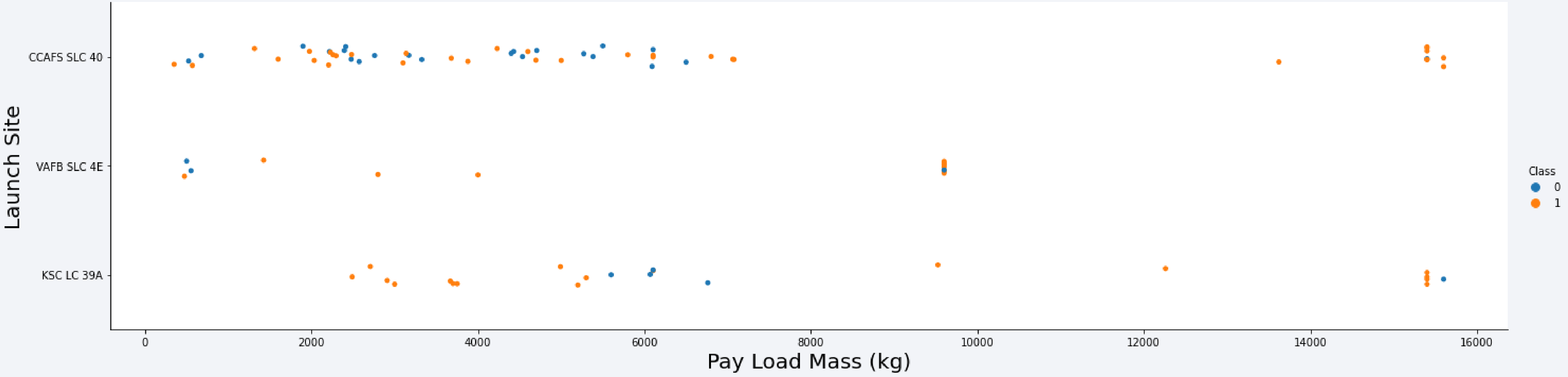
Flight Number vs. Launch Site

- From the figure it is noted that site CCAFS SLC 40 was used the most, with 9 consecutive successful launches in latest 9 flights. Site VAFB SLC 4E was used the least. Site KSC LC 39A was used occasionally, with 5 consecutive successful launches in latest 9 flights.



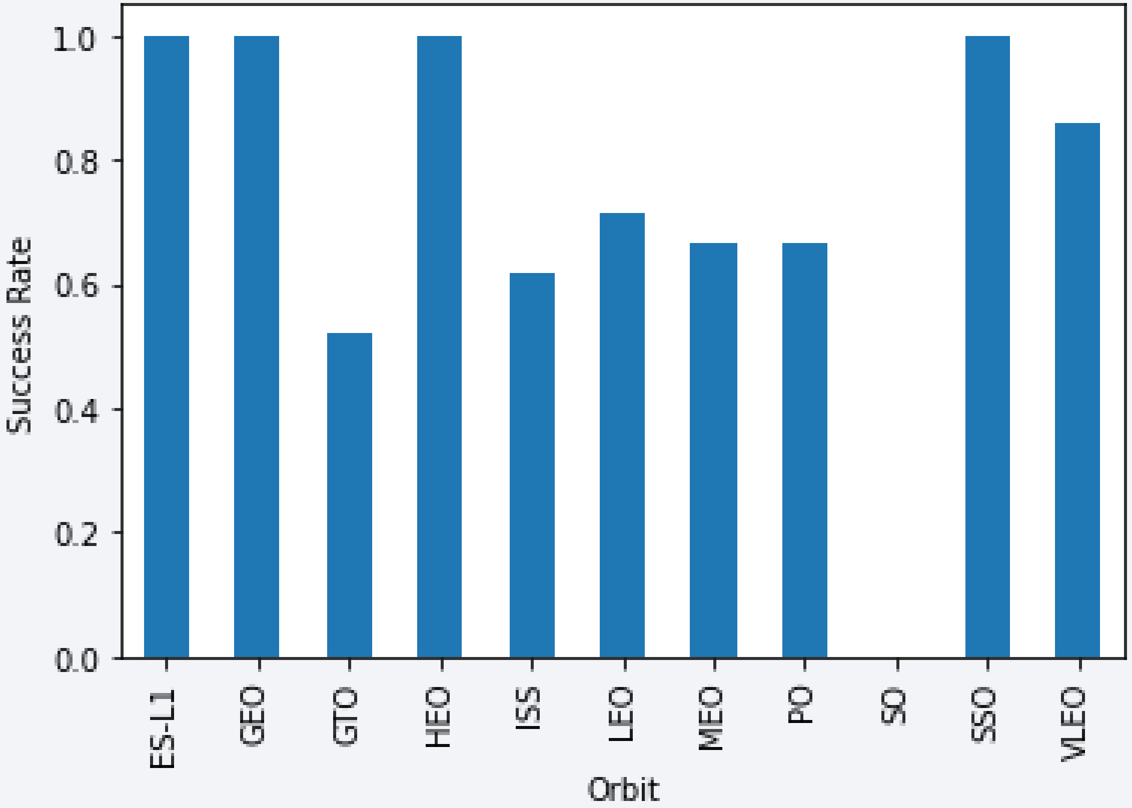
Payload vs. Launch Site

- From the figure it is noted that site VAFB SLC 4E has no launches above 10,000kg pay load mass
- Higher payload mass in general has a higher success rate



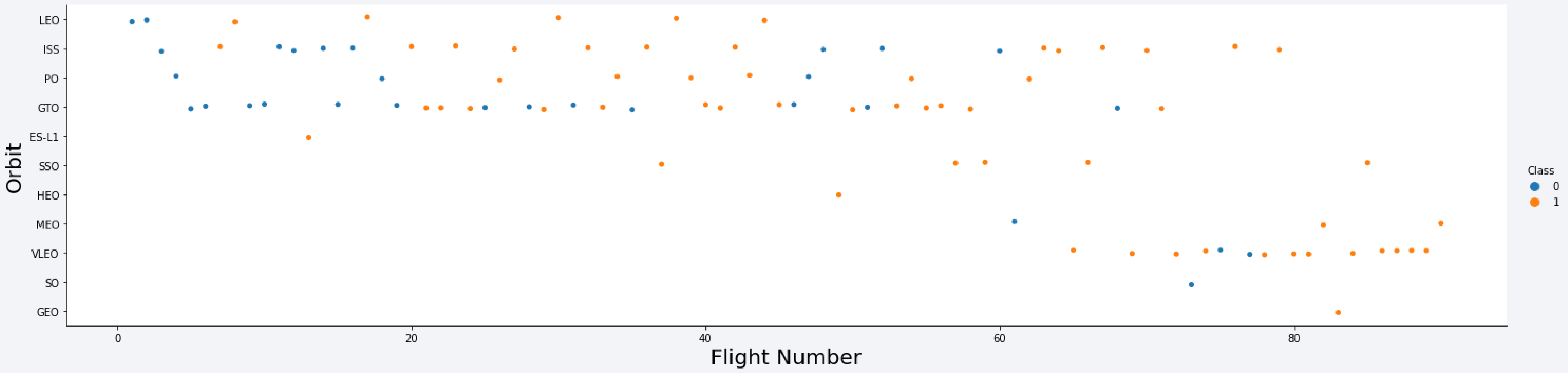
Success Rate vs. Orbit Type

- For the figure it is noted that Orbits ES-L1, GEO, GEO, SSO have the highest success rate of 1



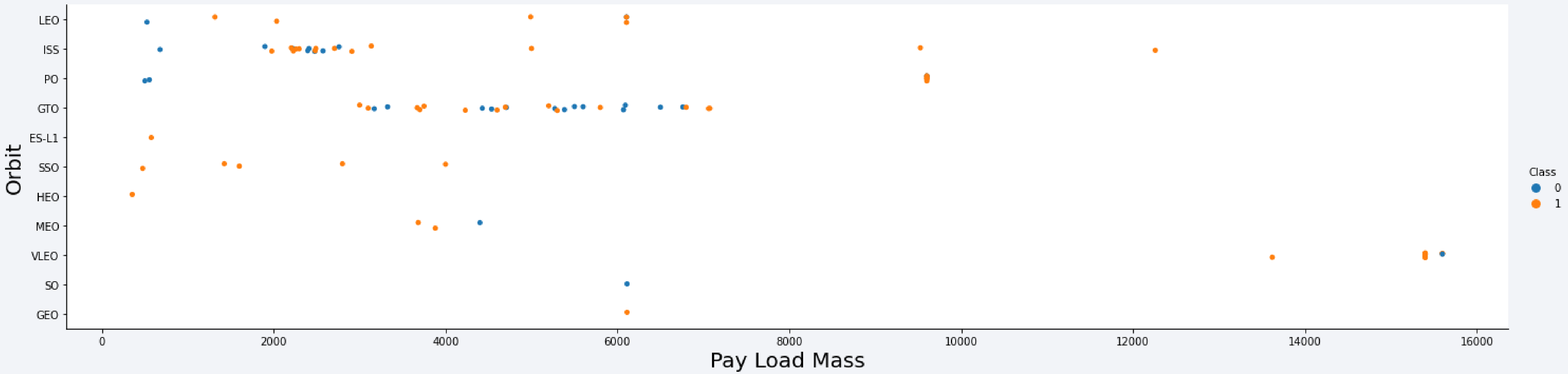
Flight Number vs. Orbit Type

- It is noted that in LEO, VLEO 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.



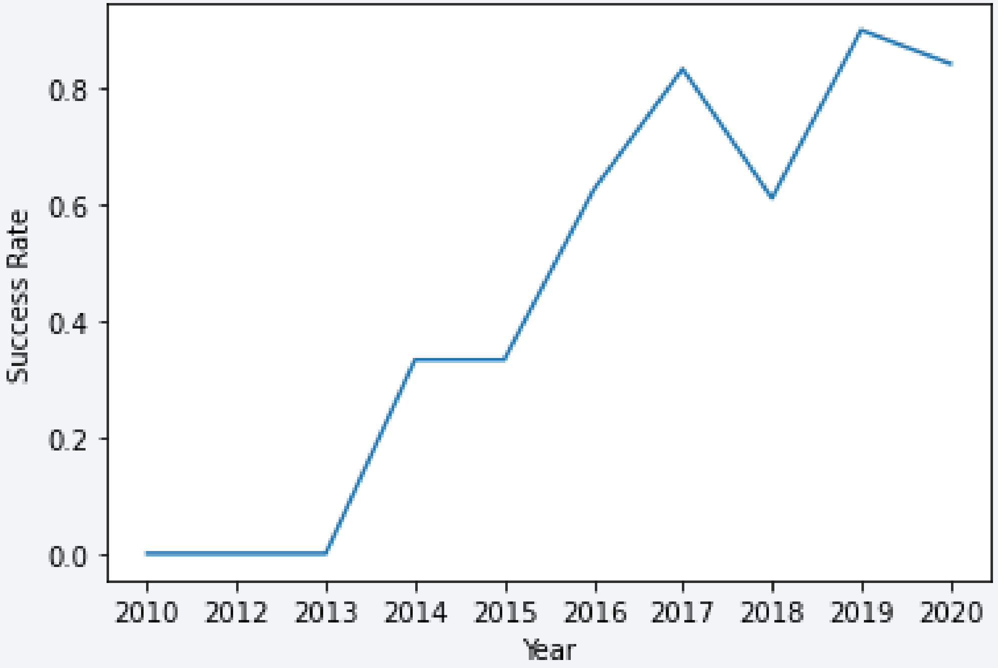
Payload vs. Orbit Type

- It is noted that in with heavy payloads the successful landing or positive landing rate are more for VLEO and ISS; whereas there is no trend for GTO orbit



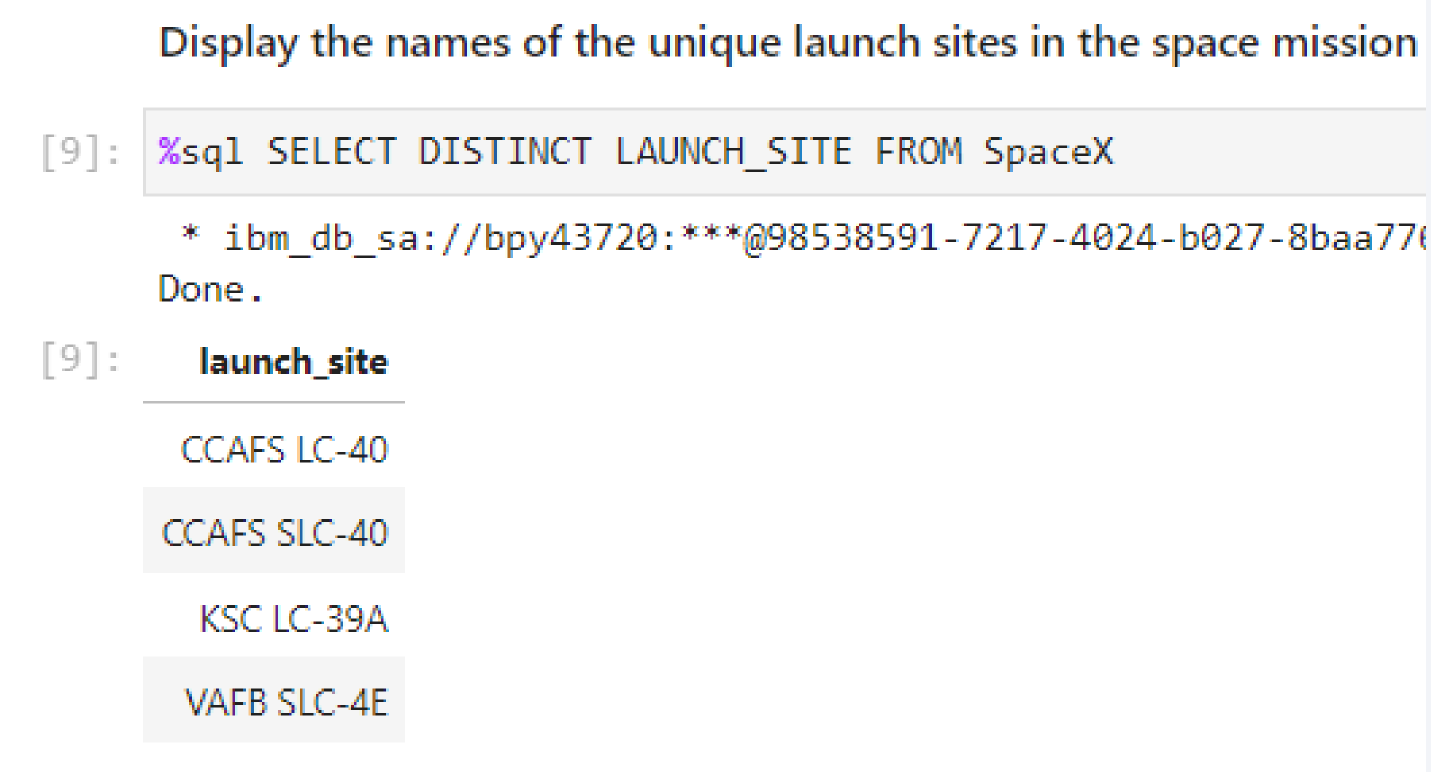
Launch Success Yearly Trend

- It is noted that the success rate has an increasing trend throughout the year, with close to 80% in operations since 2019



All Launch Site Names

- Finding the names of the unique launch sites with the query below, using the 'DISTINCT' function



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

```
[9]: %sql SELECT DISTINCT LAUNCH_SITE FROM SpaceX
```

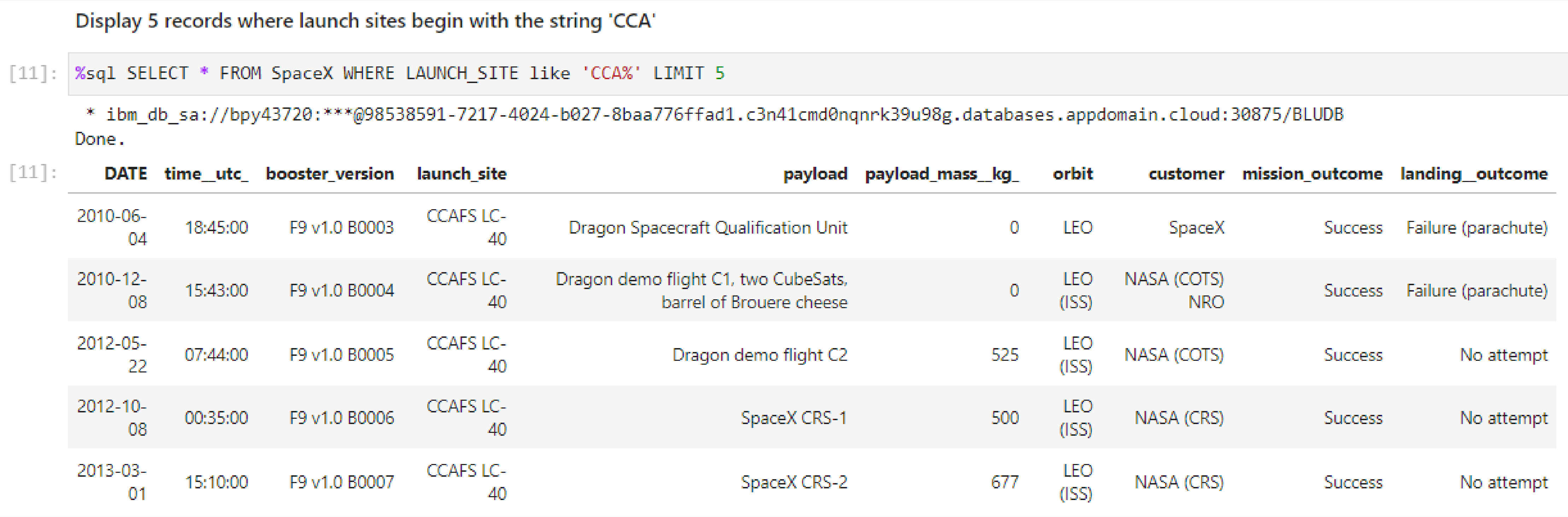
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa77
Done.
```

[9]:

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

Launch Site Names Begin with 'CCA'

- Displaying 5 records where launch sites begin with 'CCA' with 'LIMIT' function in query below



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

```
[11]: %sql SELECT * FROM SpaceX WHERE LAUNCH_SITE like 'CCA%' LIMIT 5
```

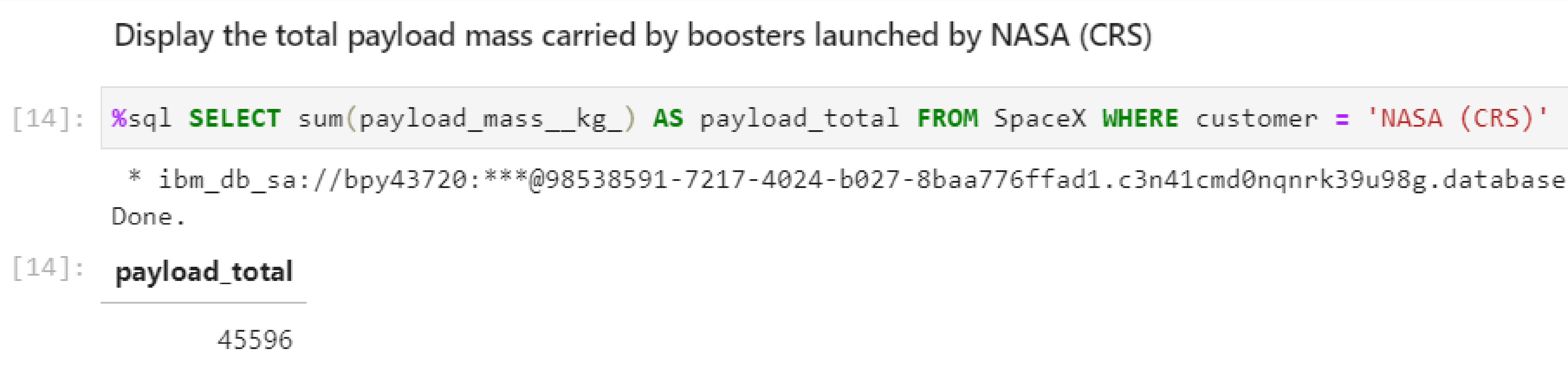
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/BLUDB
Done.
```

[11]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Calculating the total payload carried by boosters from NASA(CRS) with 'SUM' function in the query below



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

```
[14]: %sql SELECT sum(payload_mass__kg_) AS payload_total FROM SpaceX WHERE customer = 'NASA (CRS)'
```

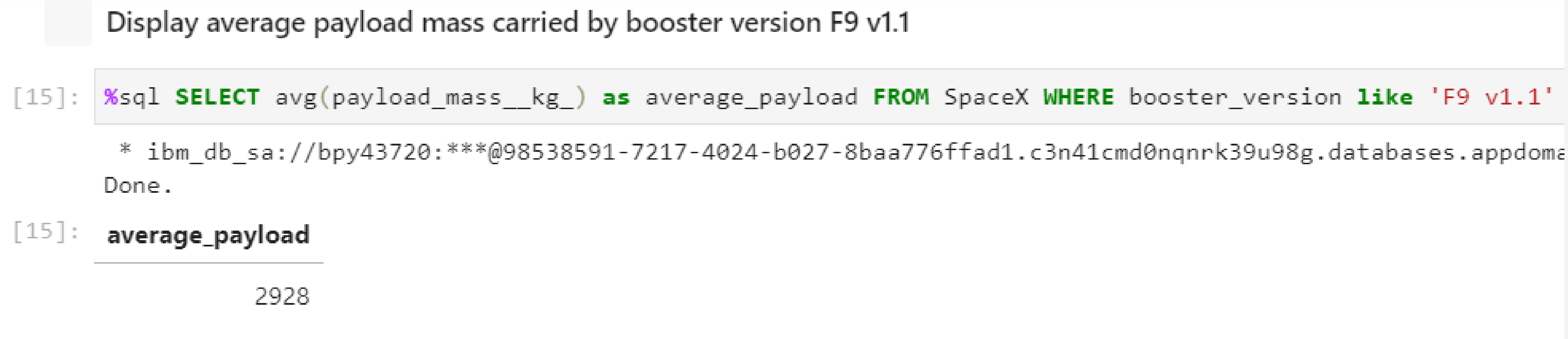
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.database
Done.
```

[14]:

payload_total
45596

Average Payload Mass by F9 v1.1

- Calculating the average payload mass carried by booster version F9 v1.1 with 'AVG' and 'LIKE' function in query below



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

```
[15]: %sql SELECT avg(payload_mass__kg_) as average_payload FROM SpaceX WHERE booster_version like 'F9 v1.1'
```

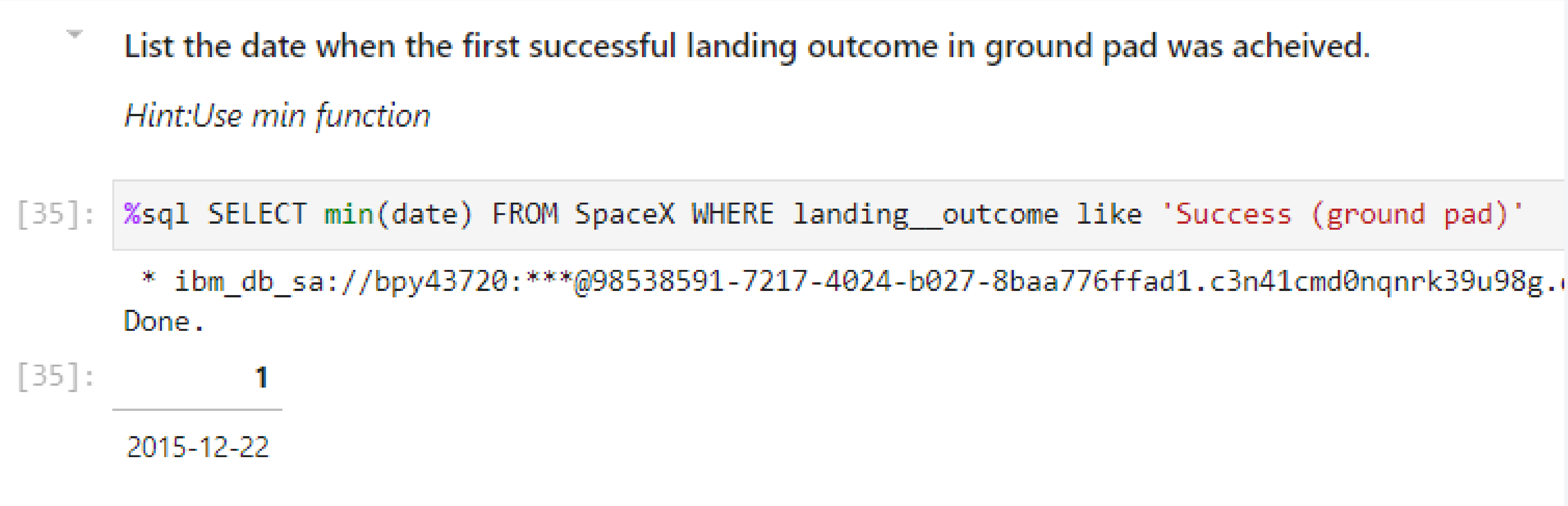
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdoma
Done.
```

[15]:

average_payload
2928

First Successful Ground Landing Date

- Finding the dates of the first successful landing outcome on ground pad, using 'MIN(date)' function in following query



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

Hint:Use min function

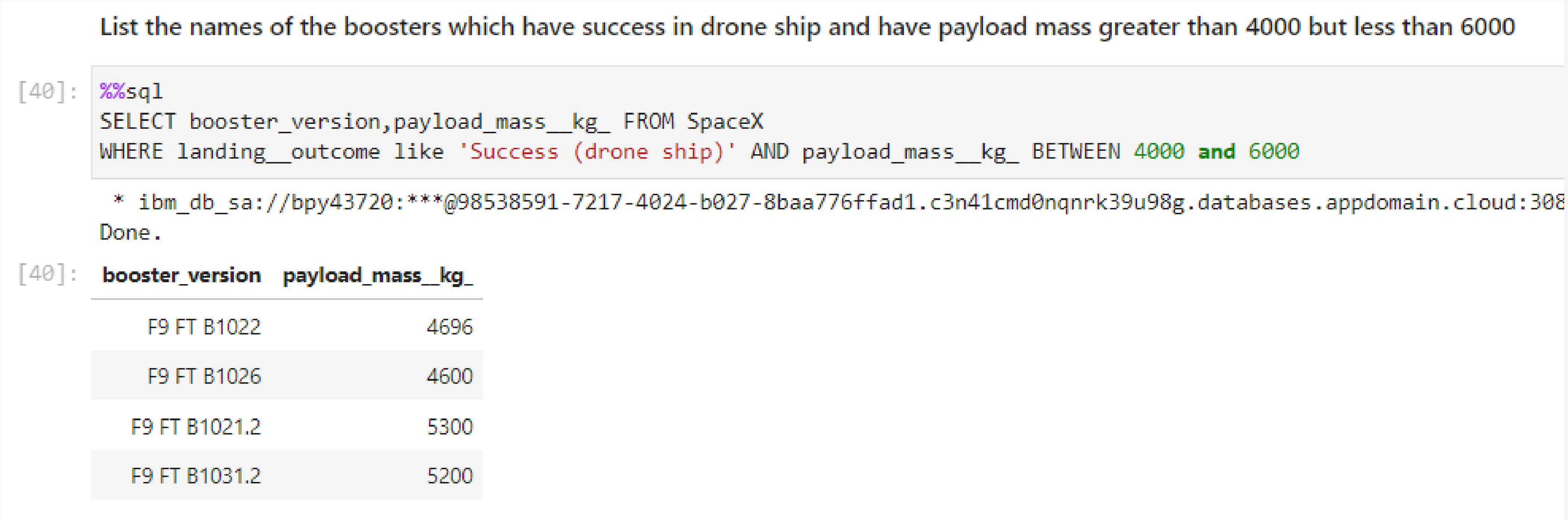
```
[35]: %sql SELECT min(date) FROM SpaceX WHERE landing__outcome like 'Success (ground pad)'

 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.
Done.
```

[35]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listing the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000, with 'BETWEEN' function shown in query below



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

[40]:
```
%%sql
SELECT booster_version,payload_mass__kg_ FROM SpaceX
WHERE landing__outcome like 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 and 6000
```

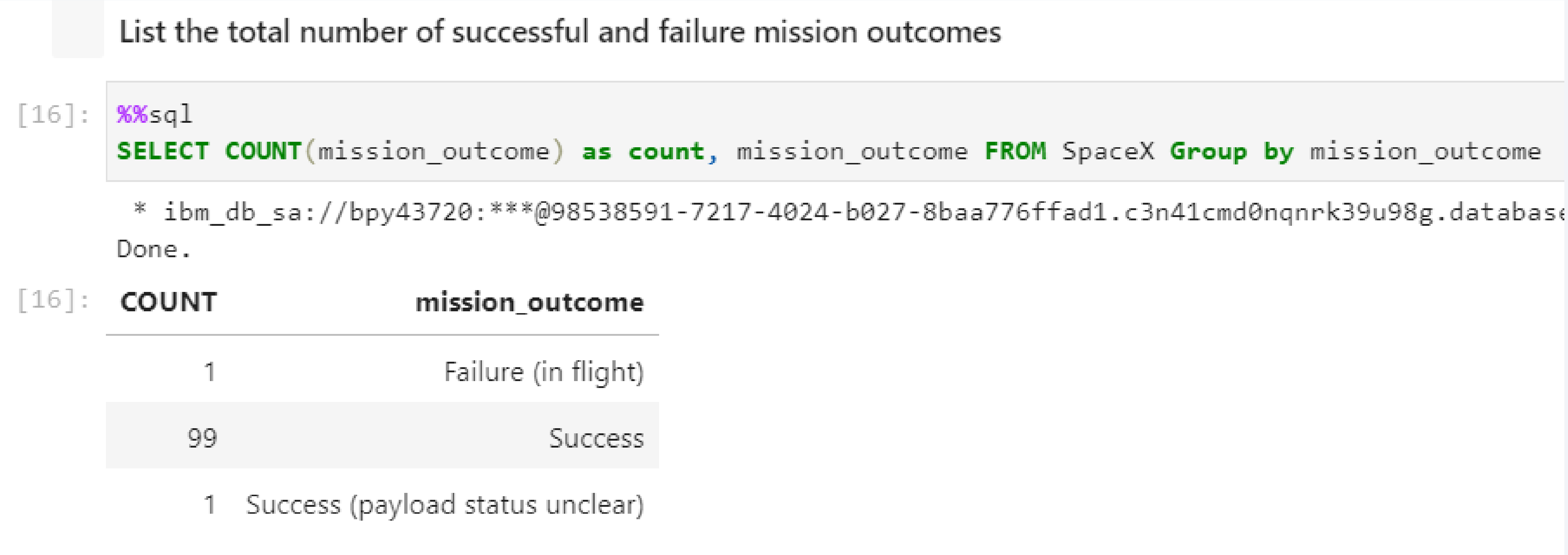
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:308
Done.
```

[40]:

booster_version	payload_mass__kg_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

- Calculating the total number of successful and failure mission outcomes with 'GROUP BY' function as shown in query below



List the total number of successful and failure mission outcomes

[16]:
```
%%sql
SELECT COUNT(mission_outcome) as count, mission_outcome FROM SpaceX Group by mission_outcome
```

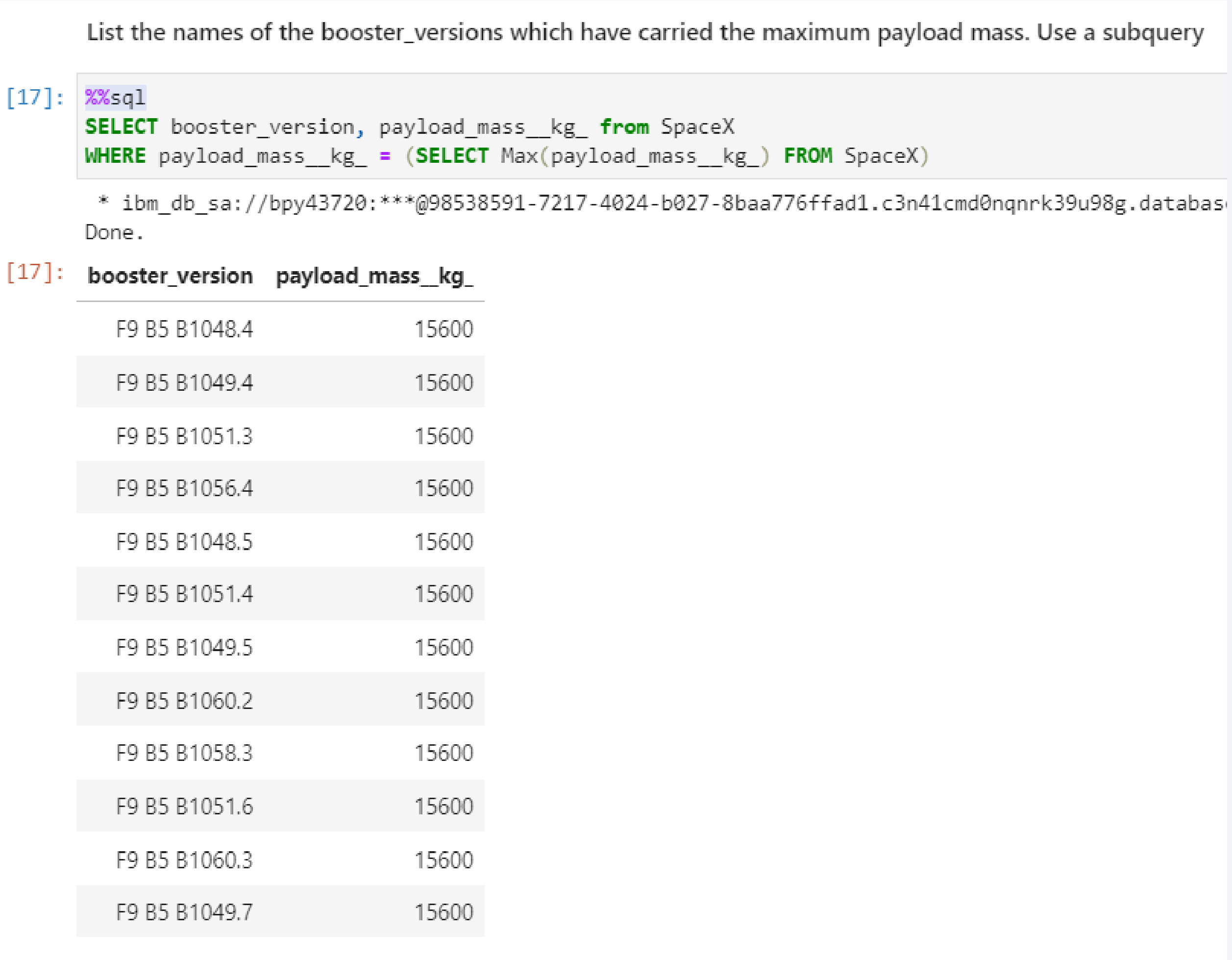
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databas
Done.
```

[16]:

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

Boosters Carried Maximum Payload

- Listing the names of the booster which have carried the maximum payload mass, using 'Max(payload_mass__kg_)' as subquery as shown below



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

```
[17]: %%sql
      SELECT booster_version, payload_mass__kg_ from SpaceX
      WHERE payload_mass__kg_ = (SELECT Max(payload_mass__kg_) FROM SpaceX)
```

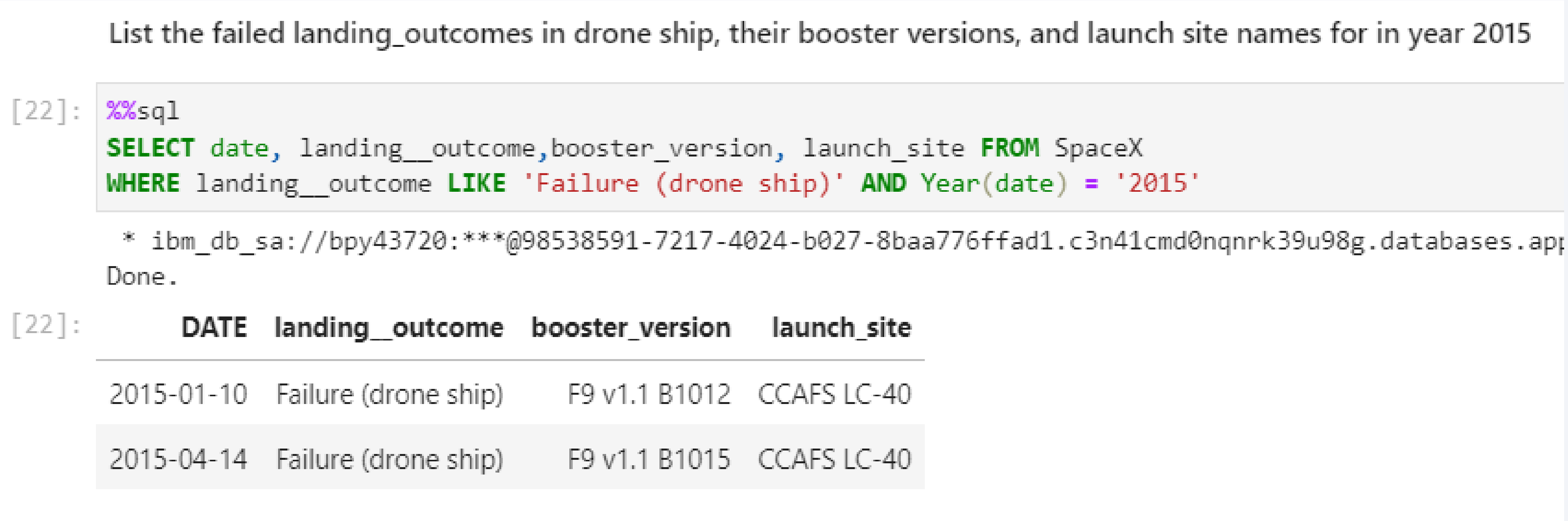
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databas
Done.
```

[17]:

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Listing the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015, with 'YEAR' function as shown in query below



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

[22]:
```
%%sql
SELECT date, landing__outcome,booster_version, launch_site FROM SpaceX
WHERE landing__outcome LIKE 'Failure (drone ship)' AND Year(date) = '2015'
```

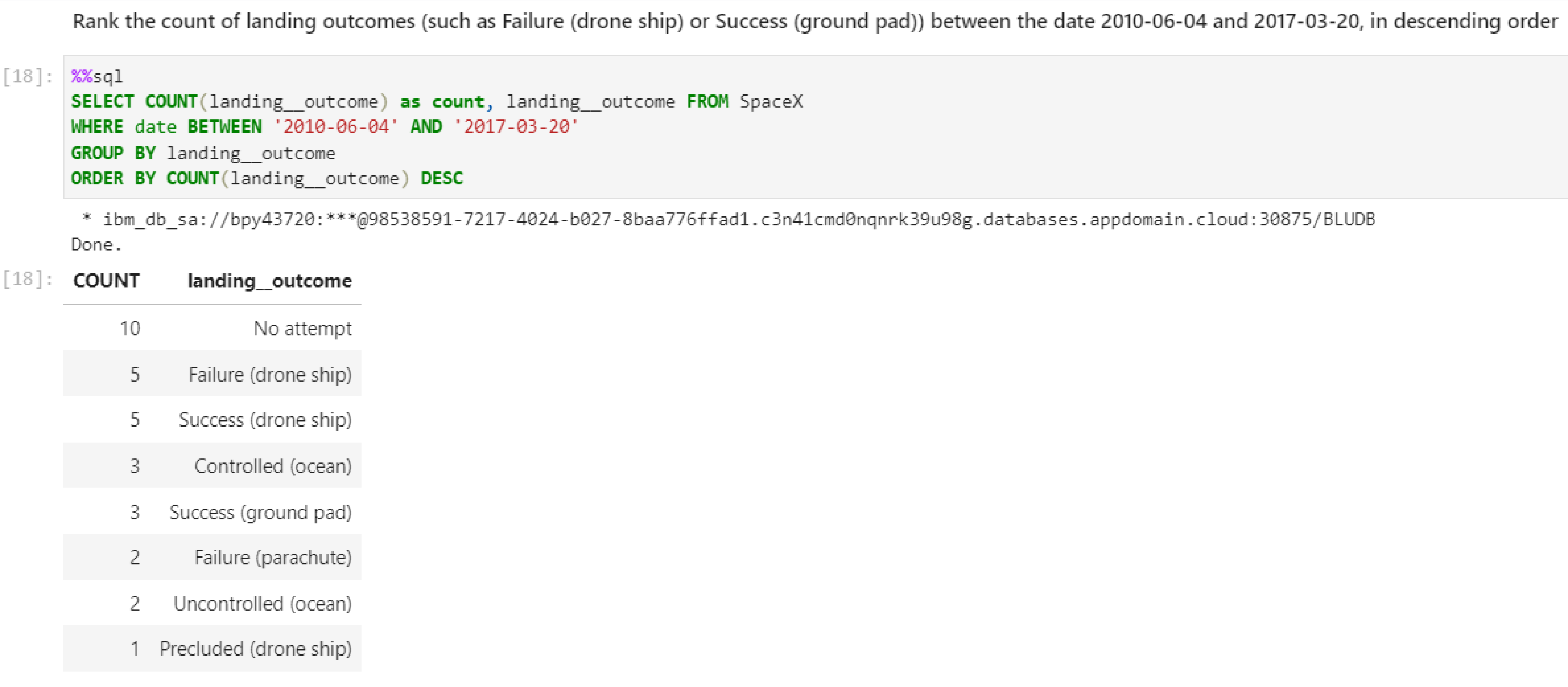
```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.app
Done.
```

[22]:

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

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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 as shown in query below, with functions 'GROUP BY' and 'ORDER BY'



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

[18]:
```
%%sql
SELECT COUNT(landing__outcome) as count, landing__outcome FROM SpaceX
WHERE date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY COUNT(landing__outcome) DESC
```

```
 * ibm_db_sa://bpy43720:***@98538591-7217-4024-b027-8baa776ffad1.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30875/BLUDB
Done.
```

[18]:

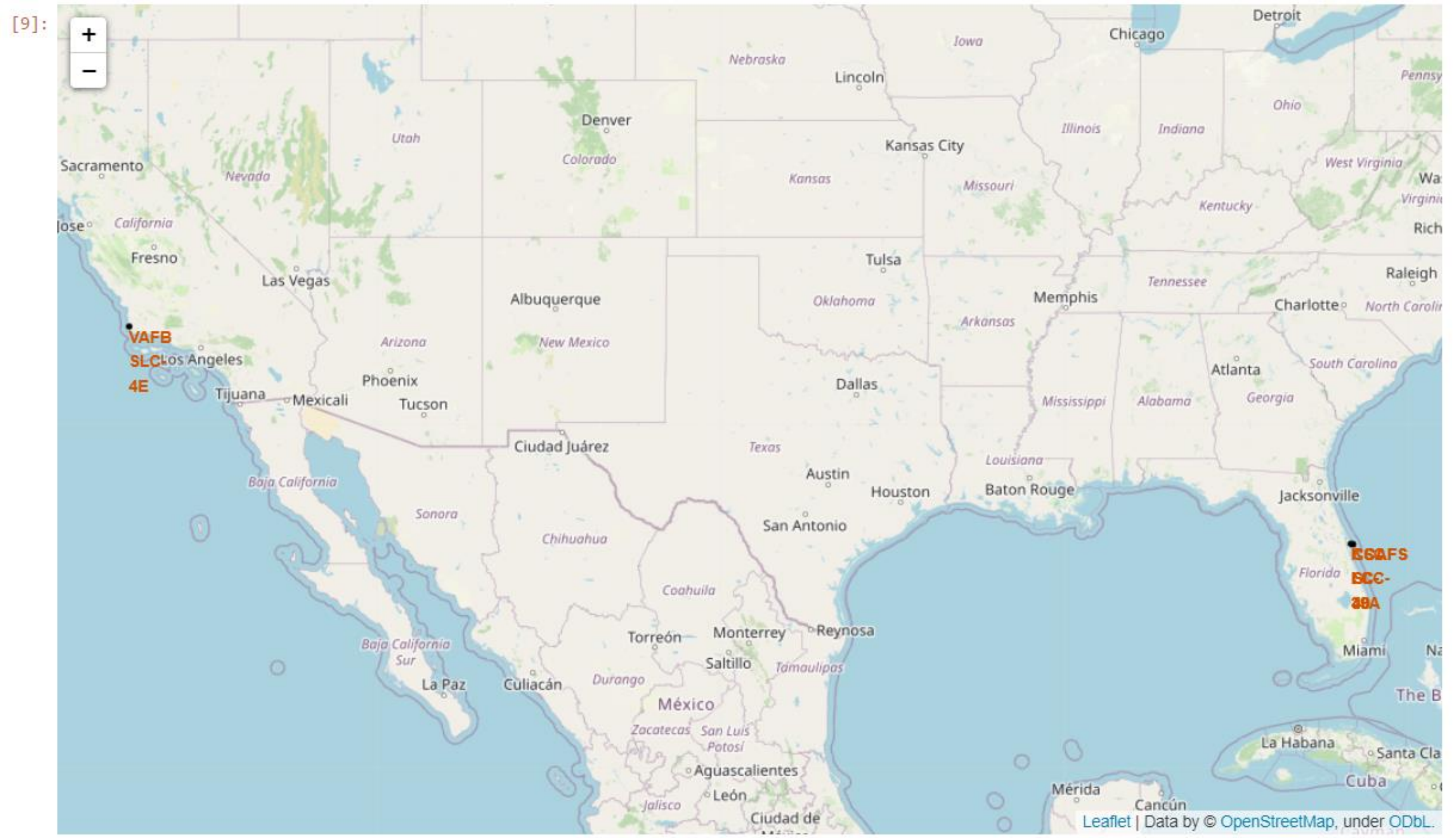
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 3

Launch Sites Proximities Analysis

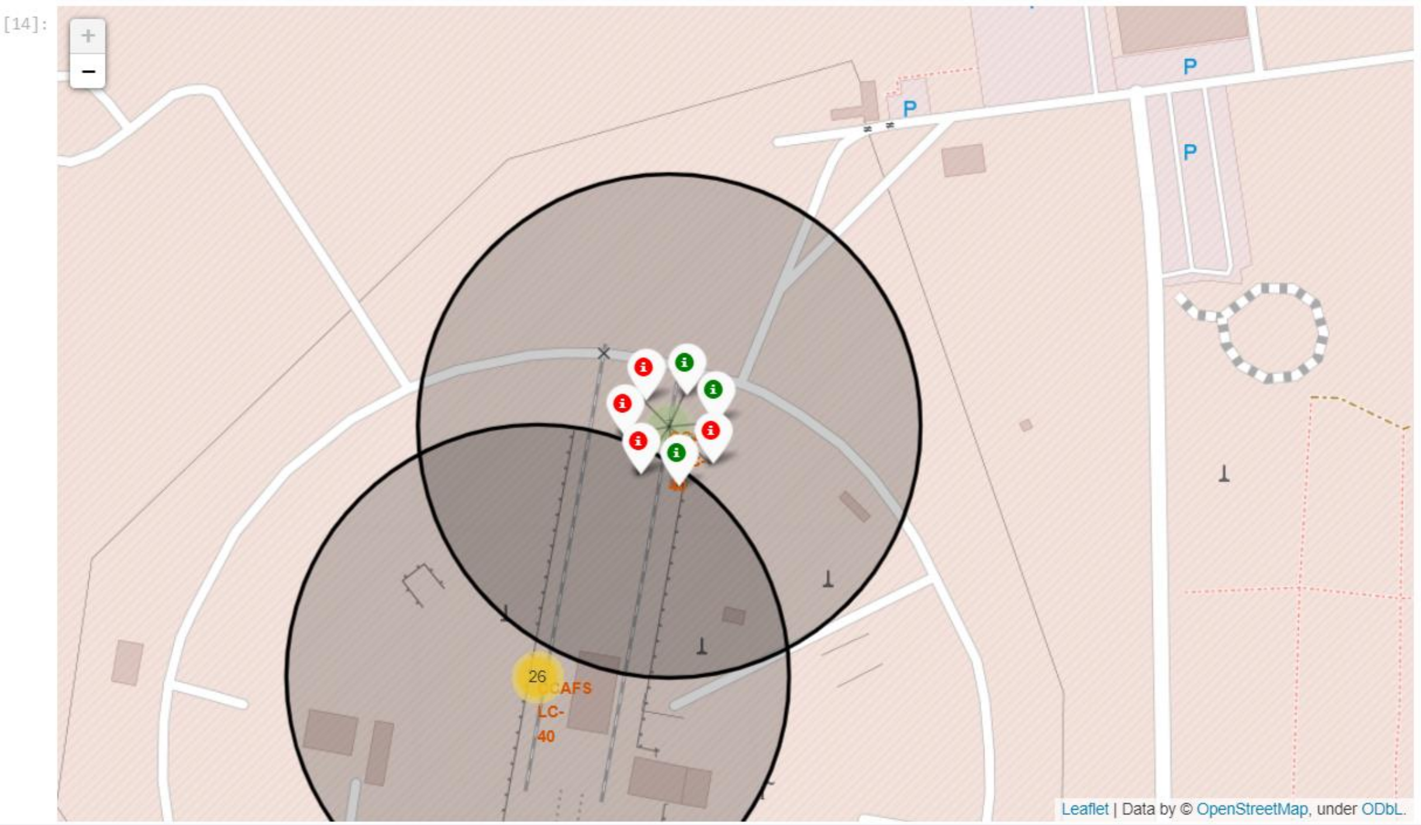
Launch Site Locations

- The launch site coordinates were extracted, circle and marker objects were added to the map to show their locations.
- Their locations in the global map shows all launch site are in Southern United States and are close to Equator, and are very close to coast



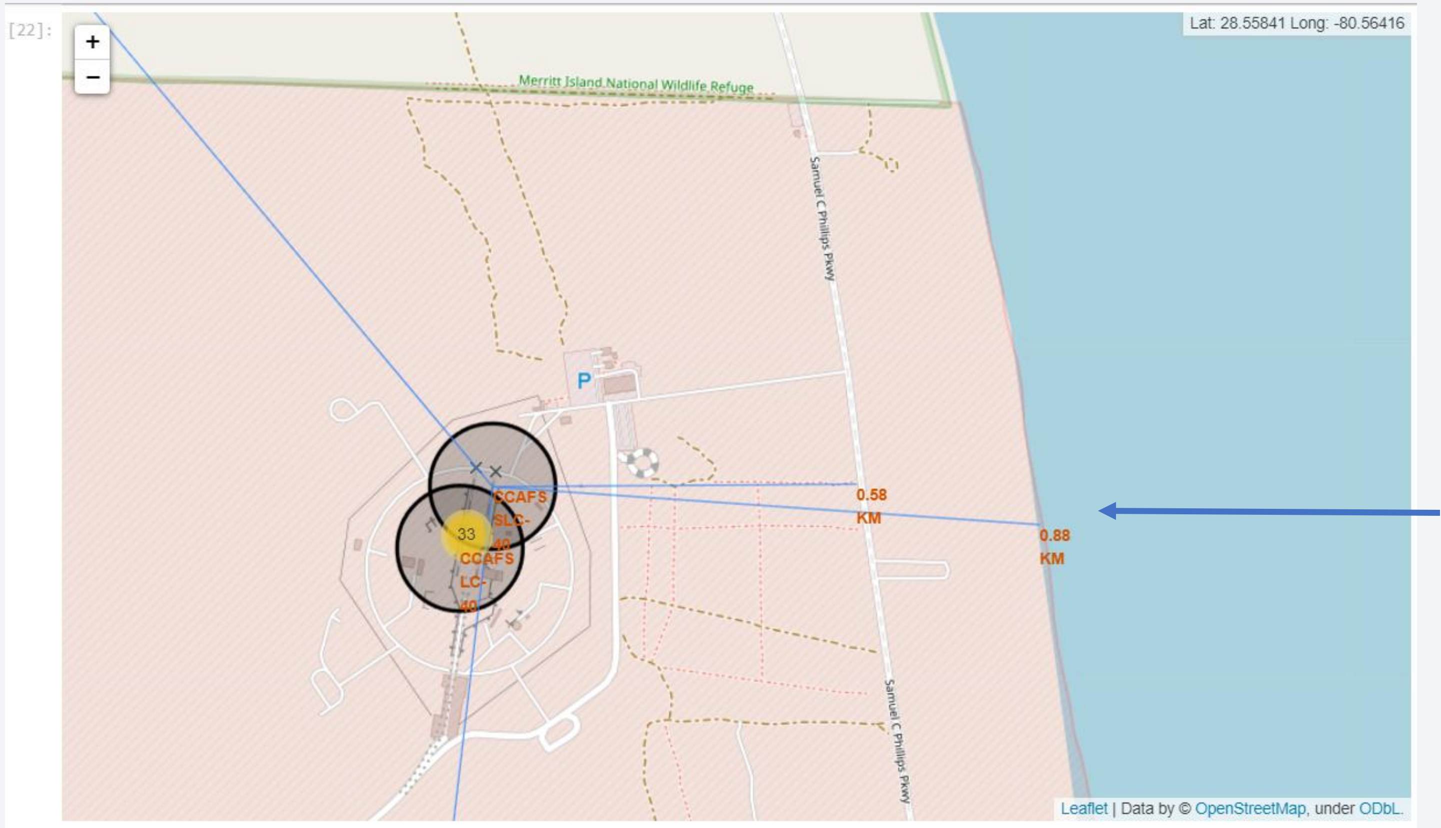
Success/failed launches for each site on the map

- Launch outcomes were added as marker cluster to the map
- Green ones are successful; red ones are failed
- This map can easily visualize the launching outcomes for a launching site

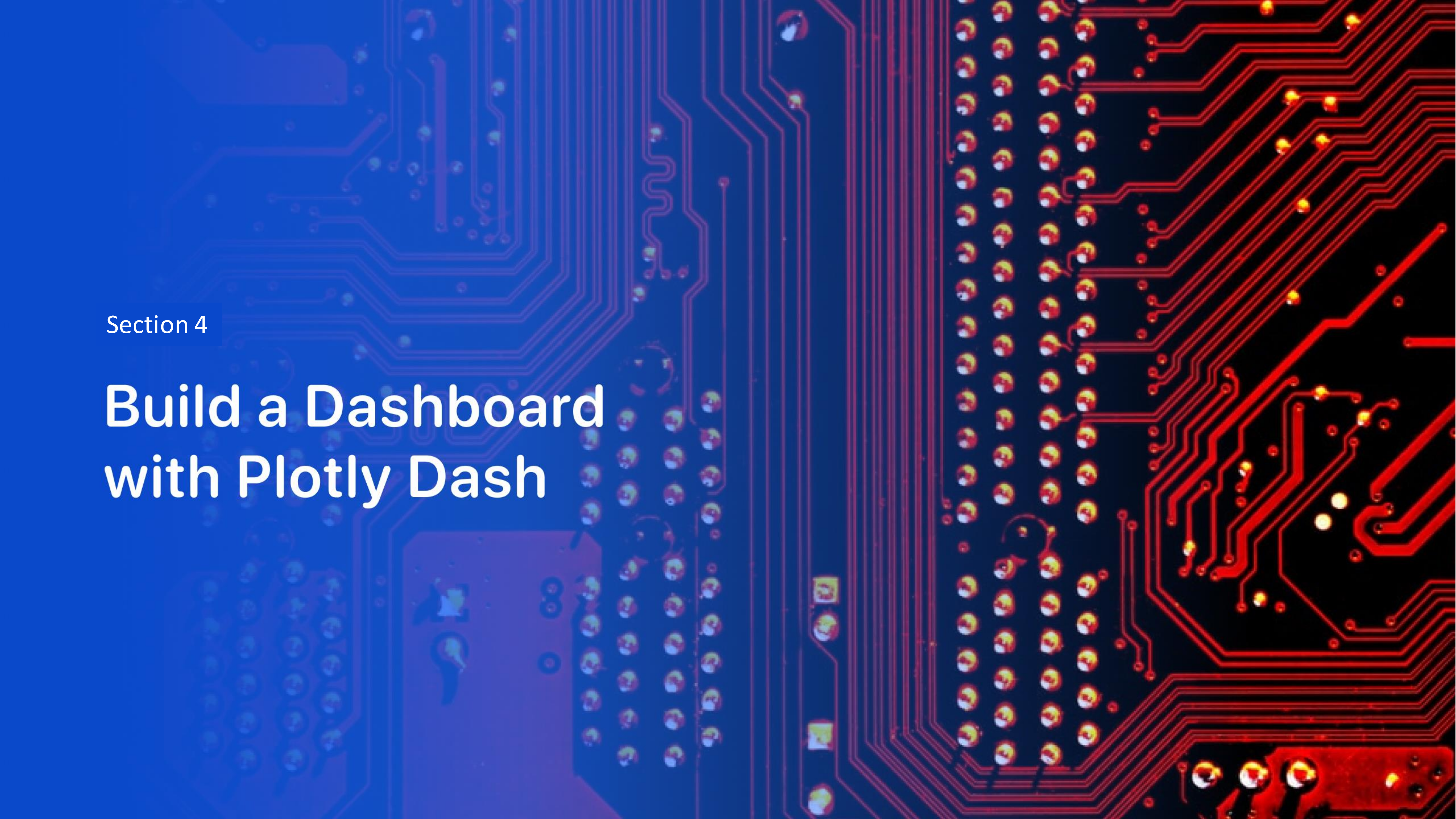


Launch site distance from surrounding features

- Distance between the launch site with surrounding features including highway, railway, coastline and city were added, by mark object and line object displaying the distance
- It is noted that the launch site was near the railway, highway and coastline to allow better supply
- The launch site is also located far away from city, potentially to minimize disturbance to public


[22]:
+
−
Lat: 28.55841 Long: -80.56416
Merritt Island National Wildlife Refuge
Samuel C Phillips Pkwy
P
CCAFS SLC-40
33
CCAFS LC-40
0.58 KM
0.88 KM
Samuel C Phillips Pkwy
Leaflet | Data by © OpenStreetMap, under ODbL.

Launch site is 0.58km from highway and 0.88km from coast

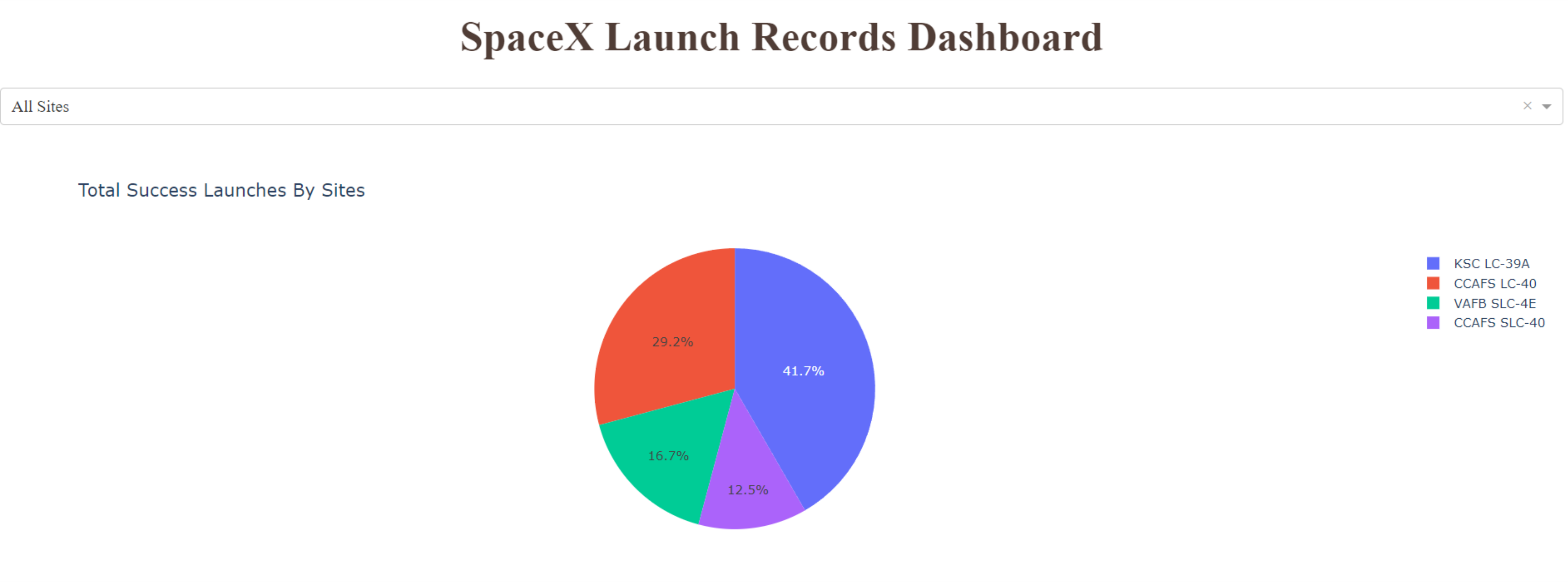


Section 4

Build a Dashboard with Plotly Dash

Total Success Launches by Sites

- Total number of success launches were plotted in a pie chart
- It is noted that launch site KSC LC-39A has the highest success count, followed by launch site CCAFS LC-40

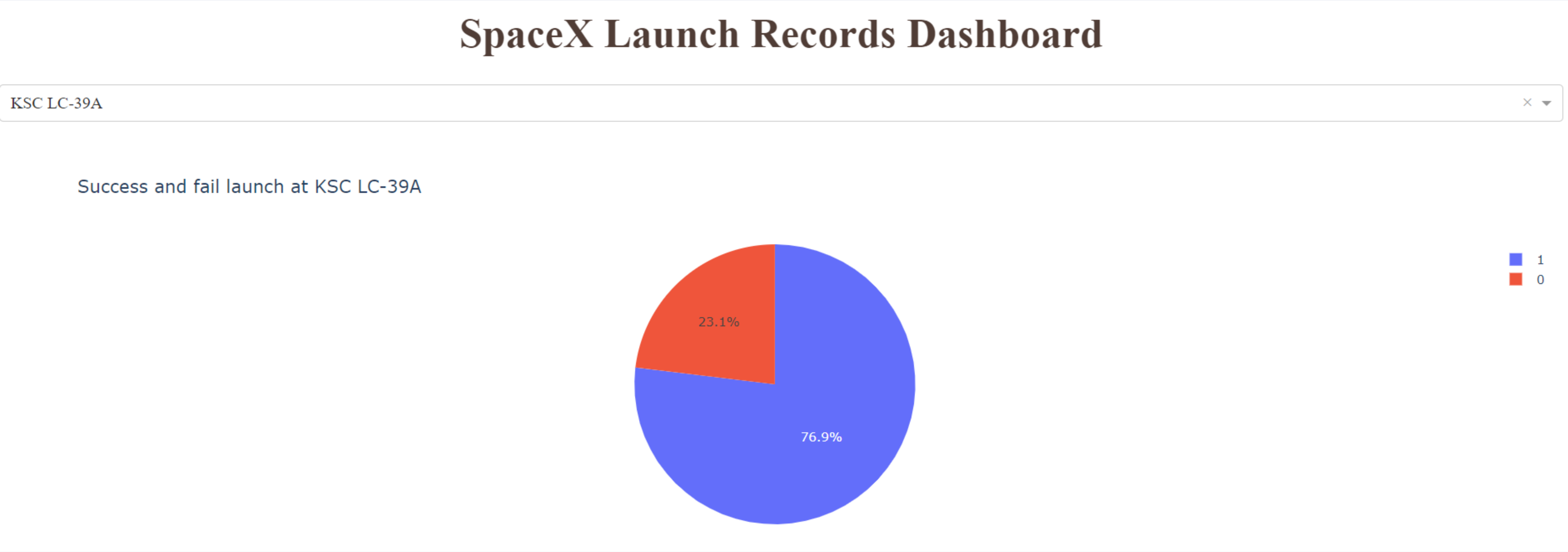


SpaceX Launch Records Dashboard

All Sites

Success rate of individual site

- Success and fail launch is plotted as pie chart (with 1 as success; 0 as fail)
- Different sites were compared and it was founded that site KSC LC-39A, with a success rate of 76.9% is highest among the four locations

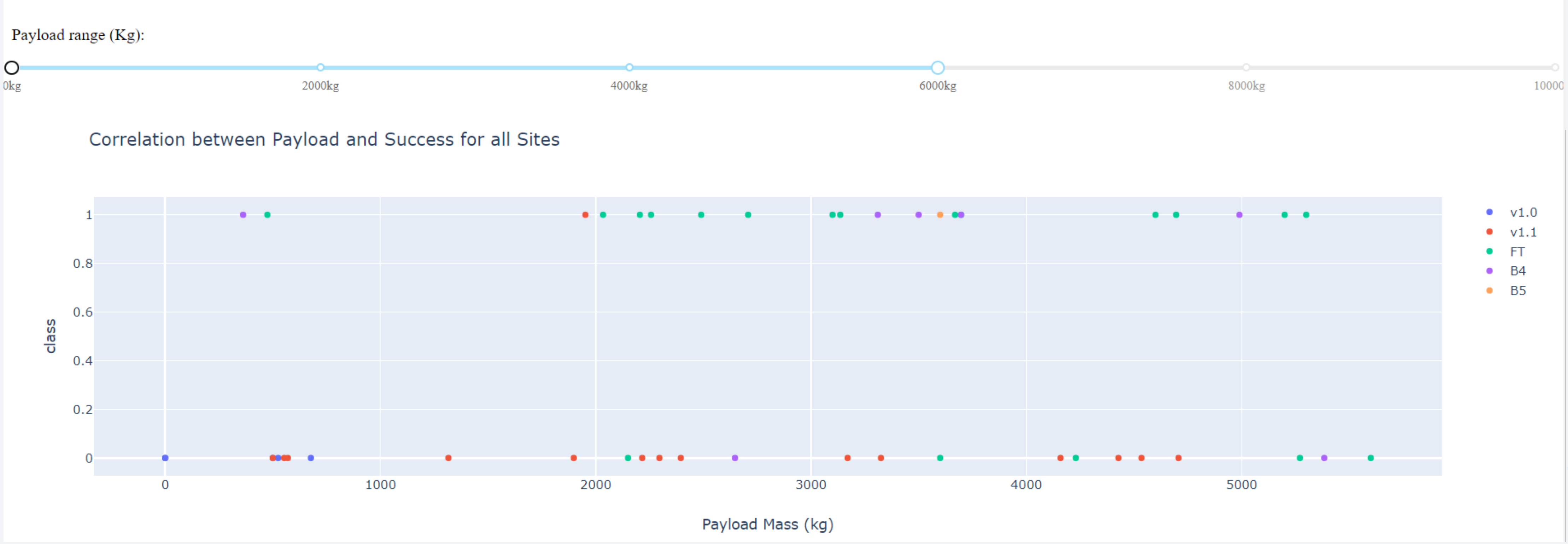


SpaceX Launch Records Dashboard

KSC LC-39A

Payload vs Launch Outcome

- Scattered plot for payload vs launch outcome was made, along with range slide for selecting the payload mass of interest
- Booster version FT has a higher successful rate.
- Payload mass between 3,000-4,000 kg has the highest successful rate



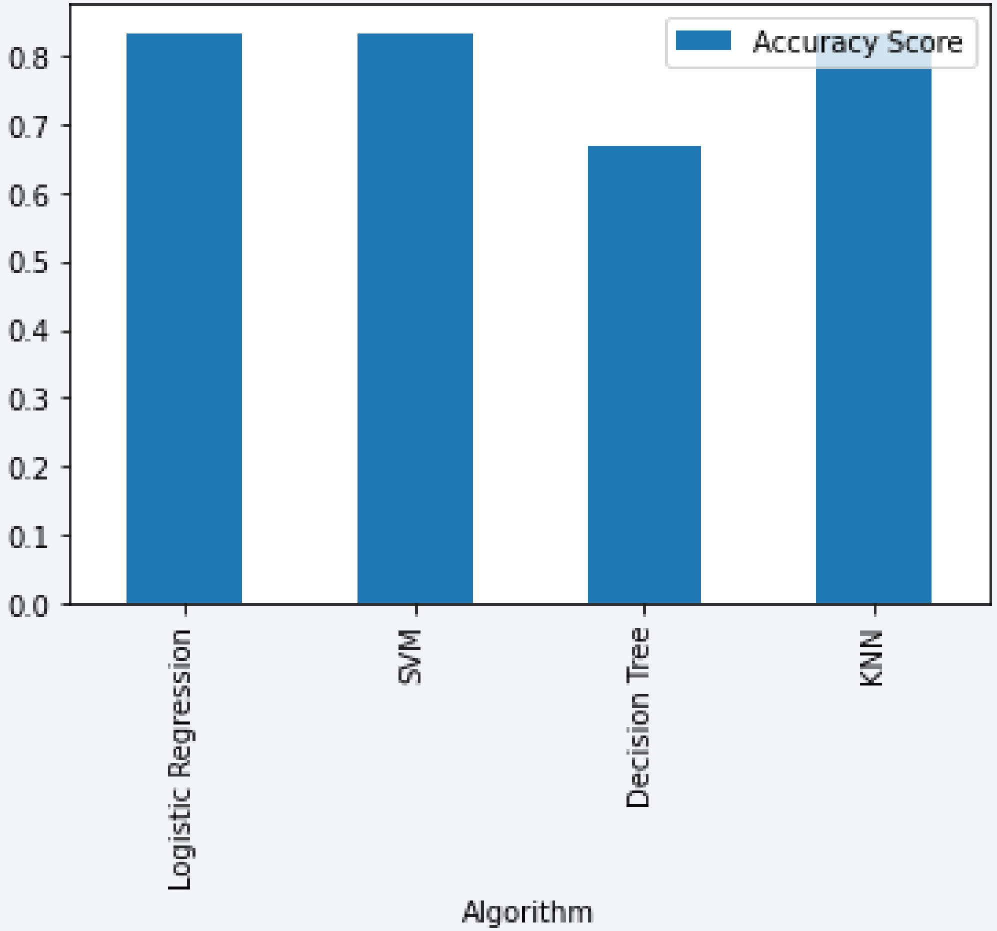
Section 5

Predictive Analysis (Classification)

Classification Accuracy

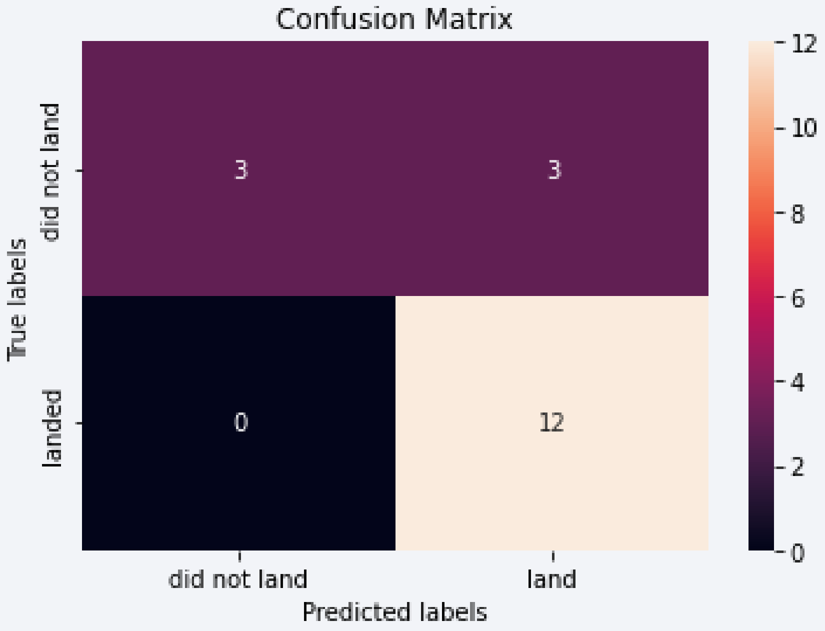
- Four models were built, including
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree
  - K Nearest Neighbour(KNN)

- LR, SVM and KNN demonstrated similar accuracy in the train/test set, with decision tree being the least accurate



Confusion Matrix

- The confusion matrix of all models are the same, primarily due to the lack of test data (sample size of 18)
- The models can distinguish between different classes
- The major problem is false positives (3 of the samples are determined 'land' by model, but the test indicates 'did not land'



Conclusions

- First stage landing success rate has improved in recent years, reaching ~80% in recent years
- Higher pay load mass often results in a higher success rate in landing missions
- Launch site KSC LC-39A has the highest success rate in launching
- Various classification models demonstrated similar accuracy, this may due to the lack of test sample (total sample size was 90, with 18 used as test sample)
  - The accuracy might be improved by adopting k-fold cross validation to the sample set
- Company may consider launching at KSC LC-39A, with a payload mass below 5000kg as a start.

Appendix and Note

- Slight difference might be observed from files generated by the notebook (output and csv files)
- This is due to there are slight change from SpaceX API with course material, as well as other discrepancies in course materials with web data
- To make the output comparable to course material, most of the input files (CSV files) are the downloaded version from IBM course

Thank you!