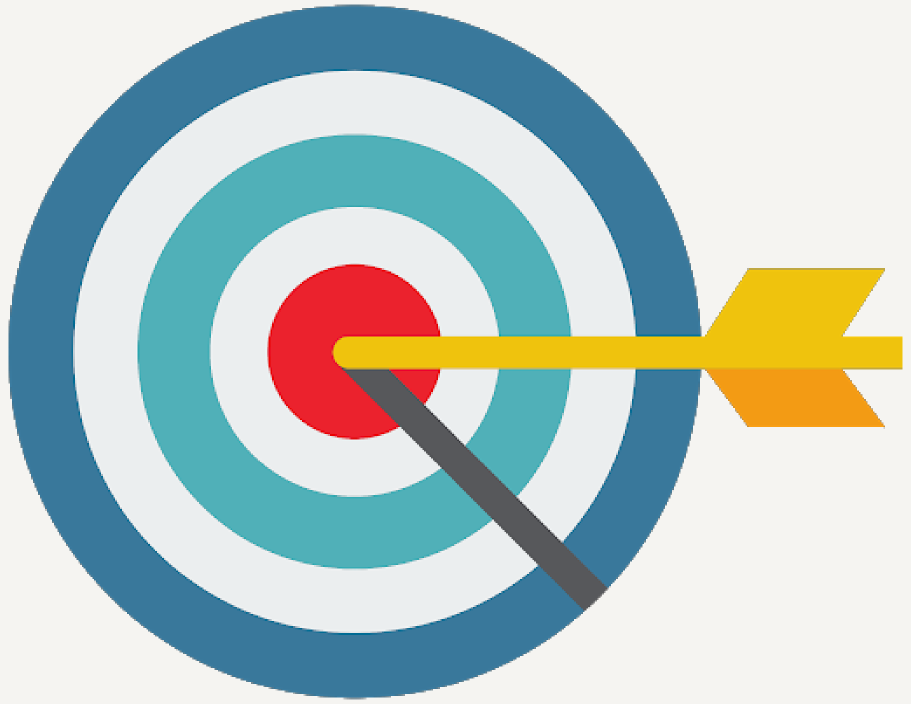
SPACEX

Data Science Capstone Project

Kareem Jamal
31.08.2024

Outline

- 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

Background

- SpaceX has become the leading company in the commercial space industry, transforming space travel by making it much more cost-effective. This success is largely attributed to the Falcon 9 rocket, which is listed on the company’s website with a launch price of $62 million. In comparison, other space providers charge over $165 million per launch. The significant cost reduction is primarily due to SpaceX's capability to reuse the rocket's first stage. By predicting whether the first stage can be successfully recovered and reused, we can accurately estimate the cost of a SpaceX launch. Using publicly available data and sophisticated machine learning models, we aim to predict the likelihood of first-stage reuse for upcoming missions.
- 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?

Introduction

Methodology

Methodology

Data Collection Methodology:

- **Utilized SpaceX REST API**
- **Employed Web Scraping from Wikipedia**

Data Wrangling:

- **Filtered and cleaned the data**
- **Addressed missing values**
- **Applied One-Hot Encoding to prepare the data for binary classification**

Exploratory Data Analysis (EDA):

- **Conducted EDA using visualizations and SQL**
- **Implemented interactive visual analytics with Folium and Plotly Dash**

Predictive Analysis:

- **Developed and fine-tuned classification models**
- **Evaluated models to ensure optimal performance**

Data wrangling

IN THE DATASET, THERE ARE VARIOUS INSTANCES WHERE THE BOOSTER DID NOT LAND SUCCESSFULLY. FOR EXAMPLE, A SUCCESSFUL OCEAN LANDING IS LABELED AS "TRUE OCEAN," WHILE AN UNSUCCESSFUL ATTEMPT IS LABELED AS "FALSE OCEAN." SIMILARLY, "TRUE RTLS" INDICATES A SUCCESSFUL LANDING ON A GROUND PAD, AND "FALSE RTLS" SIGNIFIES AN UNSUCCESSFUL ATTEMPT. LIKEWISE, "TRUE ASDS" DENOTES A SUCCESSFUL LANDING ON A DRONE SHIP, WHEREAS "FALSE ASDS" INDICATES A FAILED LANDING.

FOR OUR ANALYSIS, WE CONVERT THESE OUTCOMES INTO TRAINING LABELS, WHERE "1" REPRESENTS A SUCCESSFUL BOOSTER LANDING, AND "0" INDICATES AN UNSUCCESSFUL LANDING.

Data collection

The data collection process utilized a blend of API requests from the SpaceX REST API and web scraping techniques to extract data from a table in SpaceX's Wikipedia entr

- 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

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).

Build an interactive map with Folium

-Marker for NASA Johnson Space Center: Added a marker with a circle, popup label, and text label for NASA Johnson Space Center using its latitude and longitude coordinates as the starting location.

-Markers for All Launch Sites: Added markers with circles, popup labels, and text labels for all launch sites using their latitude and longitude coordinates to display their geographical locations and proximity to the Equator and coasts.

- Colored Markers for Launch Outcomes: Added colored markers for launch outcomes at each launch site. Green markers indicate successful launches, and red markers indicate failed launches. Marker clusters are used to identify launch sites with relatively high success rates.

Build a Dashboard with Plotly Dash

Dropdown List:

-Here are the revised descriptions:

•**Dropdown List for Launch Site Selection**: Added a dropdown list to enable the selection of a specific launch site.

•**Pie Chart Showing Successful Launches (All Sites/Certain Site)**: Added a pie chart to display the total count of successful launches for all sites, and the success vs. failure counts for a specific launch site if selected.

•**Slider for Payload Mass Range**: Added a slider to allow the selection of a payload mass range.

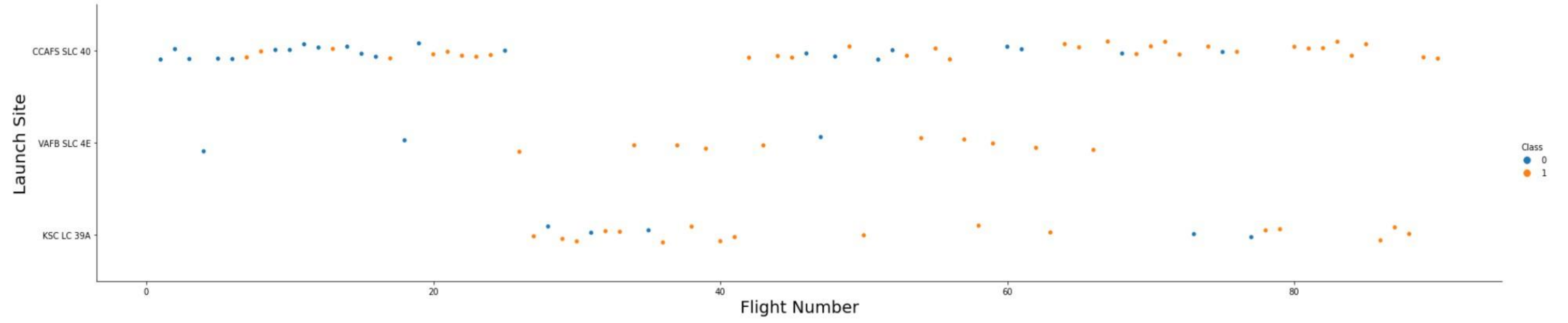
•**Scatter Chart of Payload Mass vs. Success Rate for Different Booster Versions**: Added a scatter chart to illustrate the correlation between payload mass and launch success for different booster versions.

Results

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

EDA with Visualization

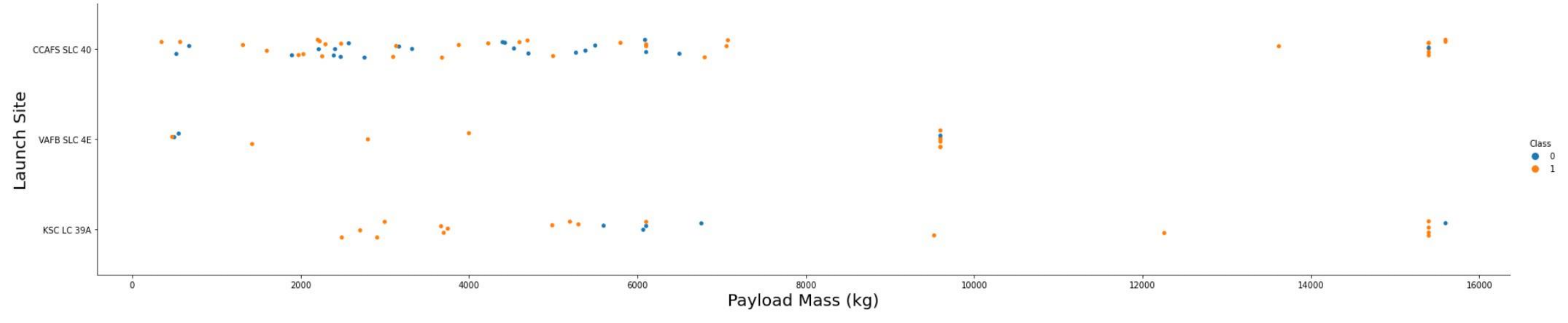
Flight Number vs. Launch Site



Explanation:

- 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



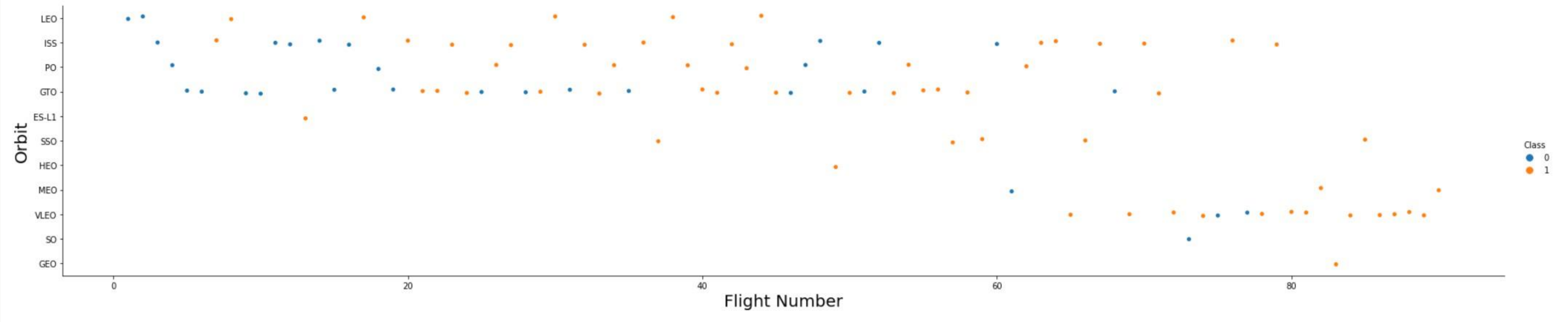
Explanation:

- 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.

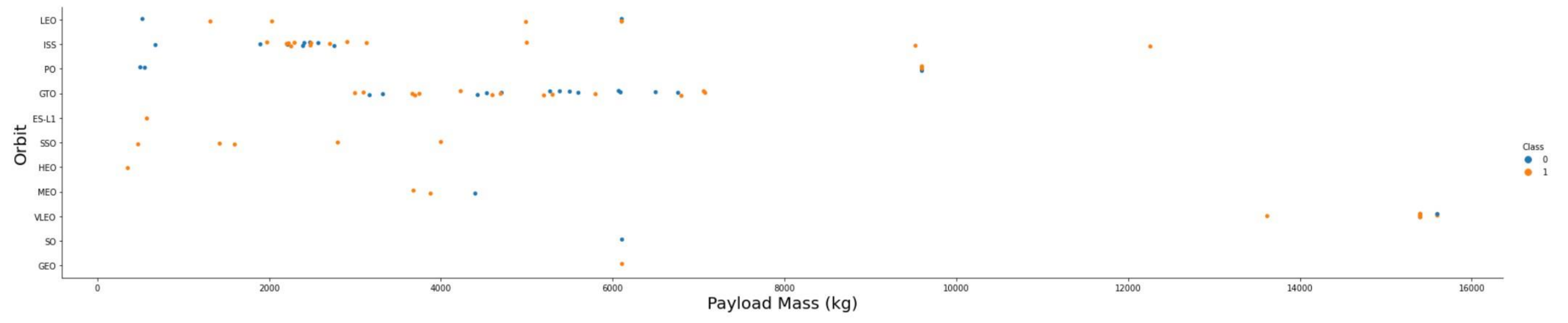
Flight Number vs. Orbit type

Explanation:

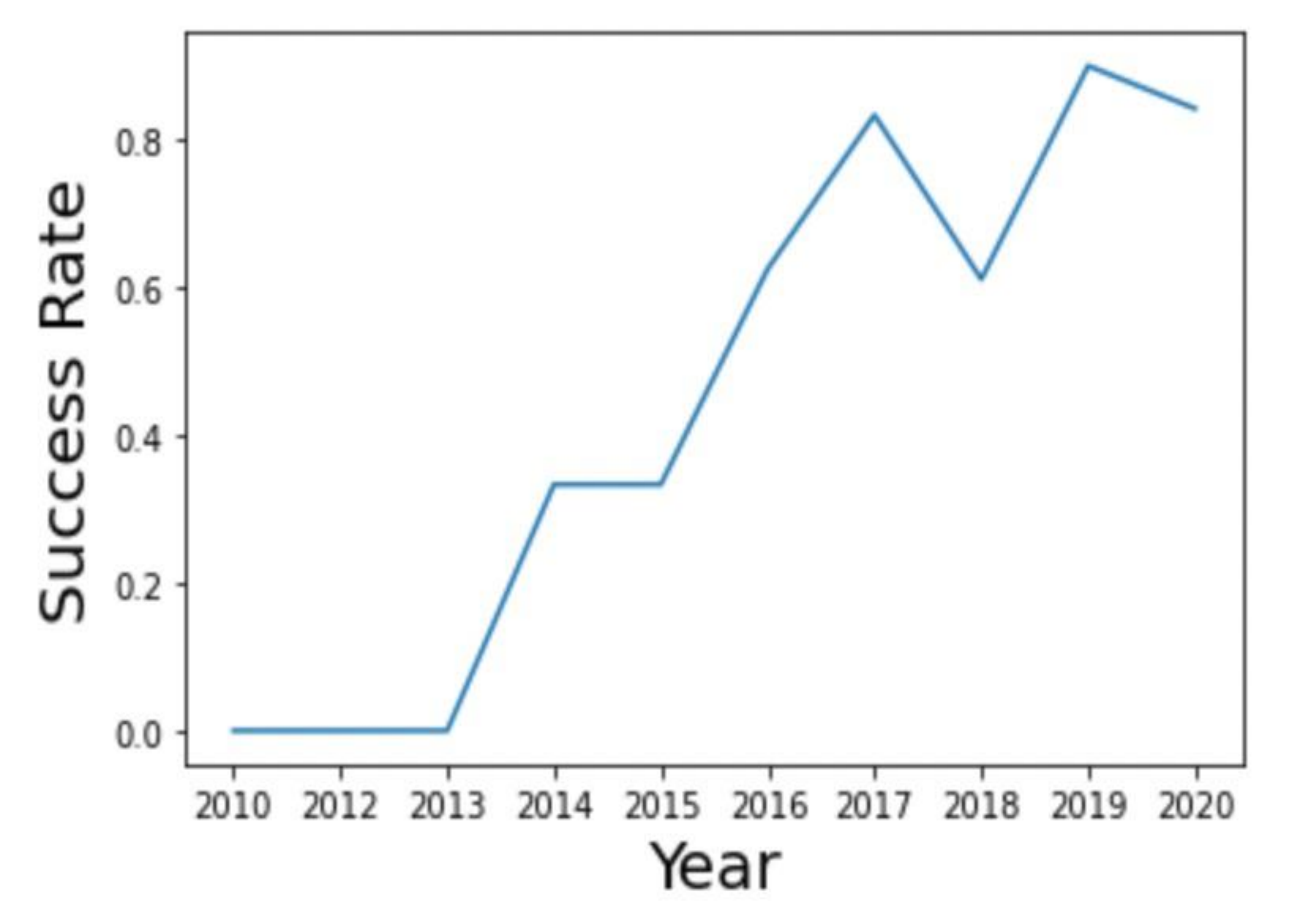
- In the LEO orbit, success seems to be correlated with the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between the number of flights and success.



Payload Mass vs. Orbit type



Launch success yearly trend

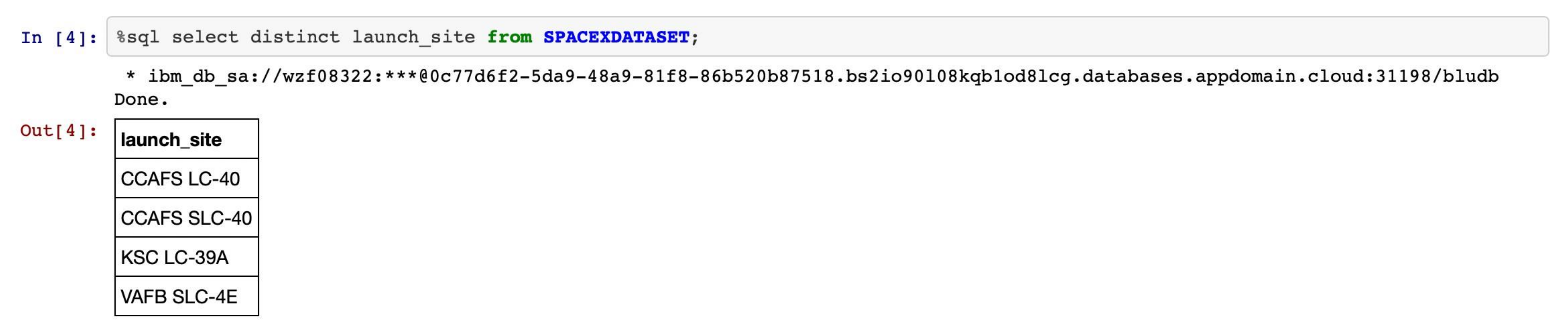


Explanation:

- The success rate increased till 2020.

EDA with SQL

All launch site names



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

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

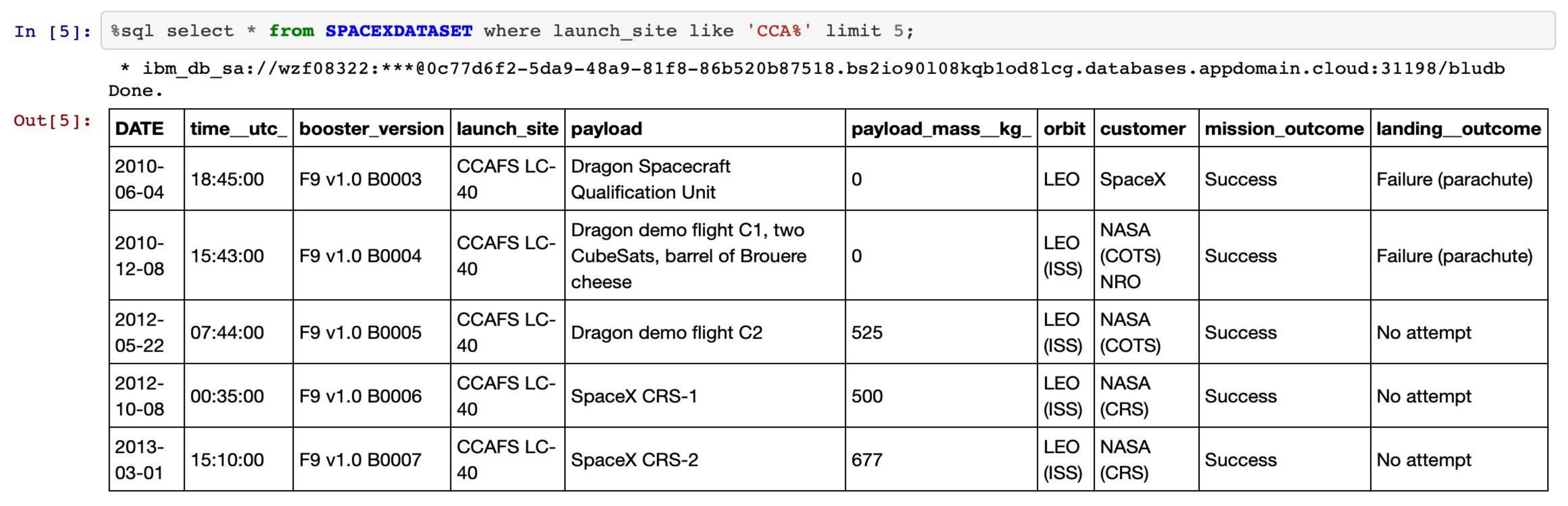
Out[4]:

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

Explanation:

- Displaying the names of the unique launch sites in the space mission.

Launch site names begin with `CCA`



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

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

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

Explanation:

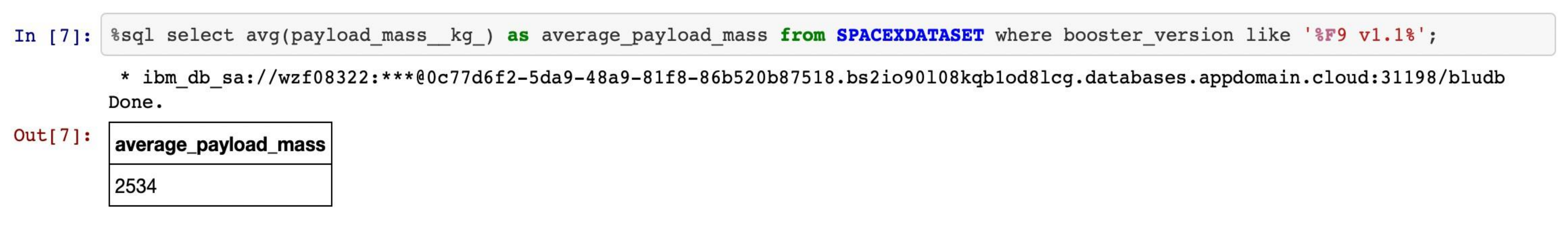
- Displaying 5 records where launch sites begin with the string 'CCA'.

EDA with SQL

revised SQL queries:

I. Displaying the names of the unique launch sites in the space mission.
II. Displaying 5 records where launch sites begin with the string ‘CCA’.
III. Displaying the total payload mass carried by boosters launched by NASA (CRS).
IV. Displaying the average payload mass carried by booster version F9 v1.1.
V. Listing the total number of successful and failed mission outcomes.
VI. Listing the names of the booster versions that have carried the maximum payload mass.

Average payload mass by F9 v1.1



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

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

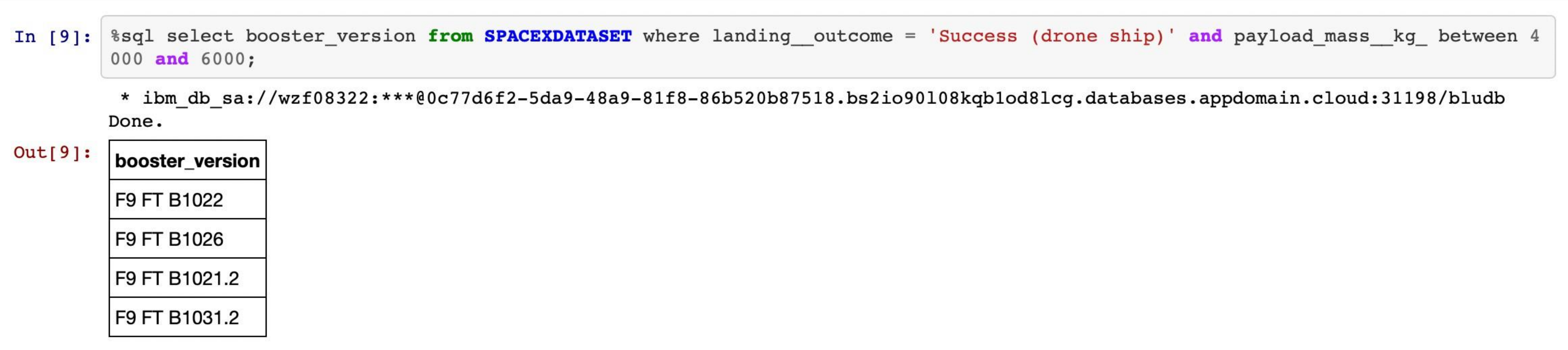
Out[7]:

average_payload_mass
2534

Explanation:

- Displaying average payload mass carried by booster version F9 v1.1.

Successful drone ship landing with payload between 4000 and 6000



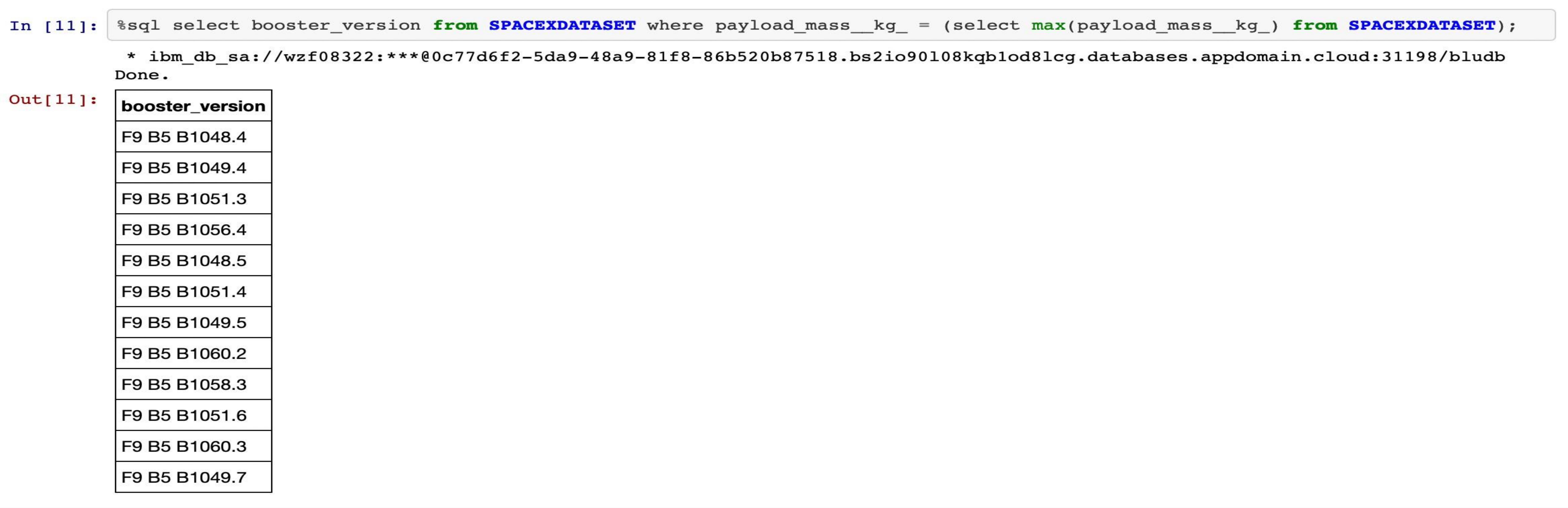
```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * 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

Explanation:

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

Boosters carried maximum payload



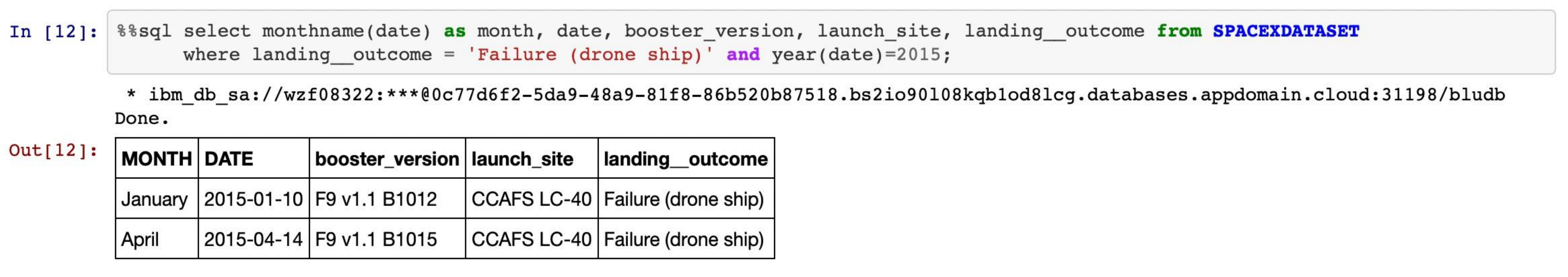
```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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

Explanation:

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

2015 launch records



```
%%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;
```

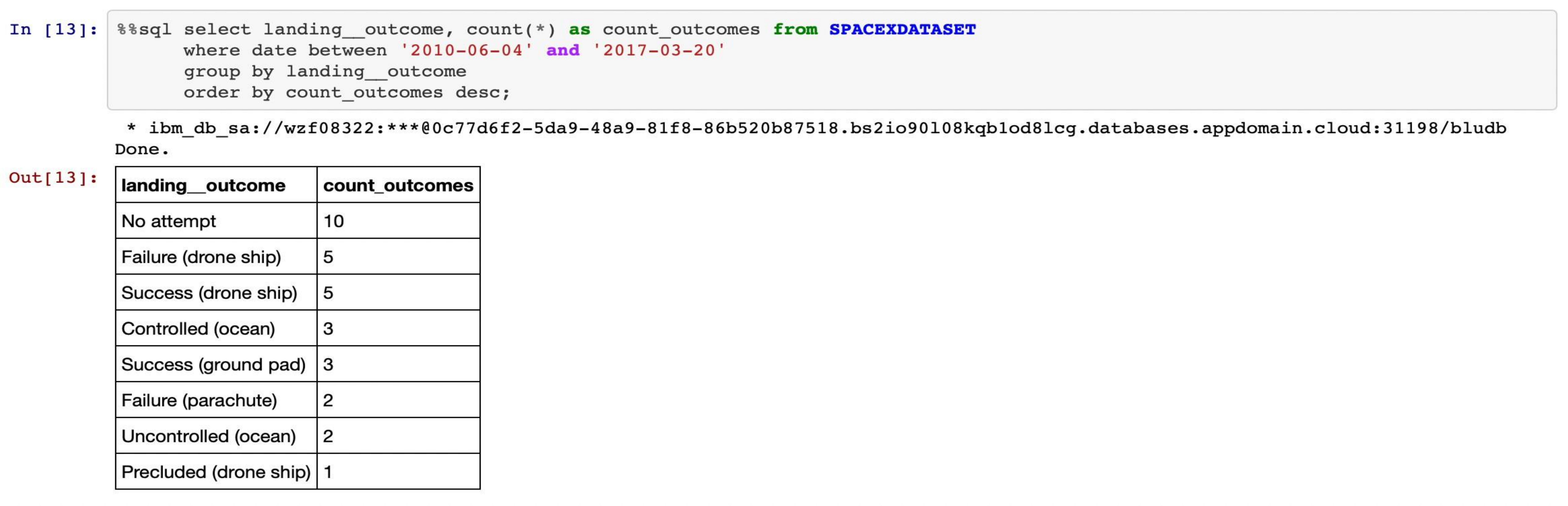
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

Explanation:

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

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



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

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

Out[13]:

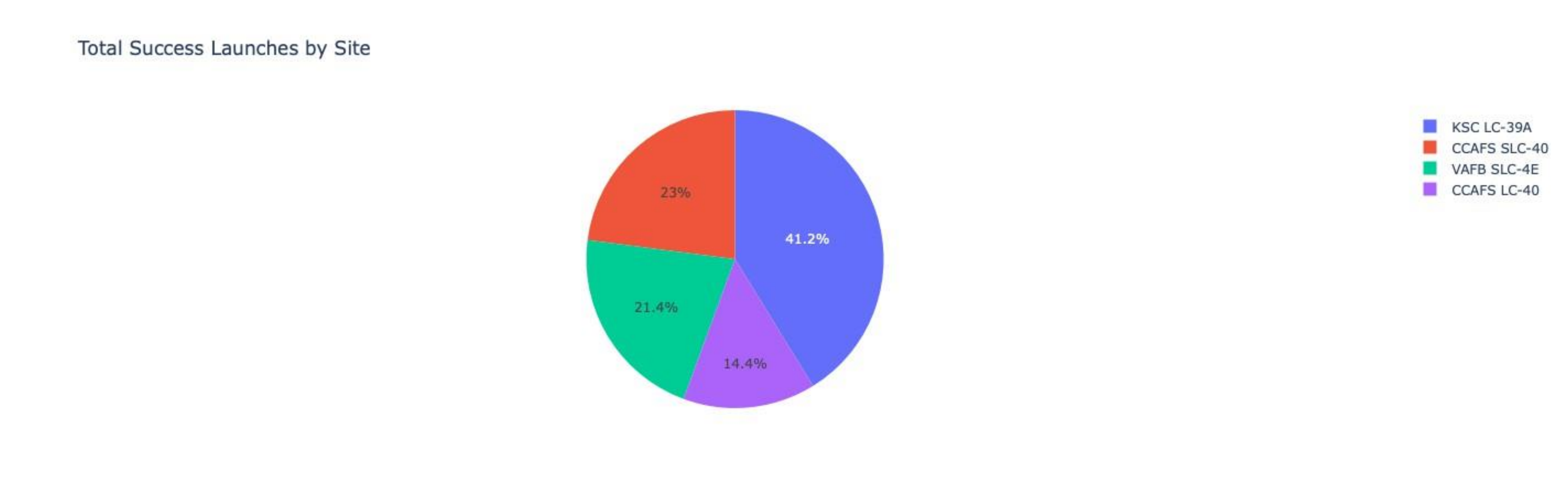
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Explanation:

- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

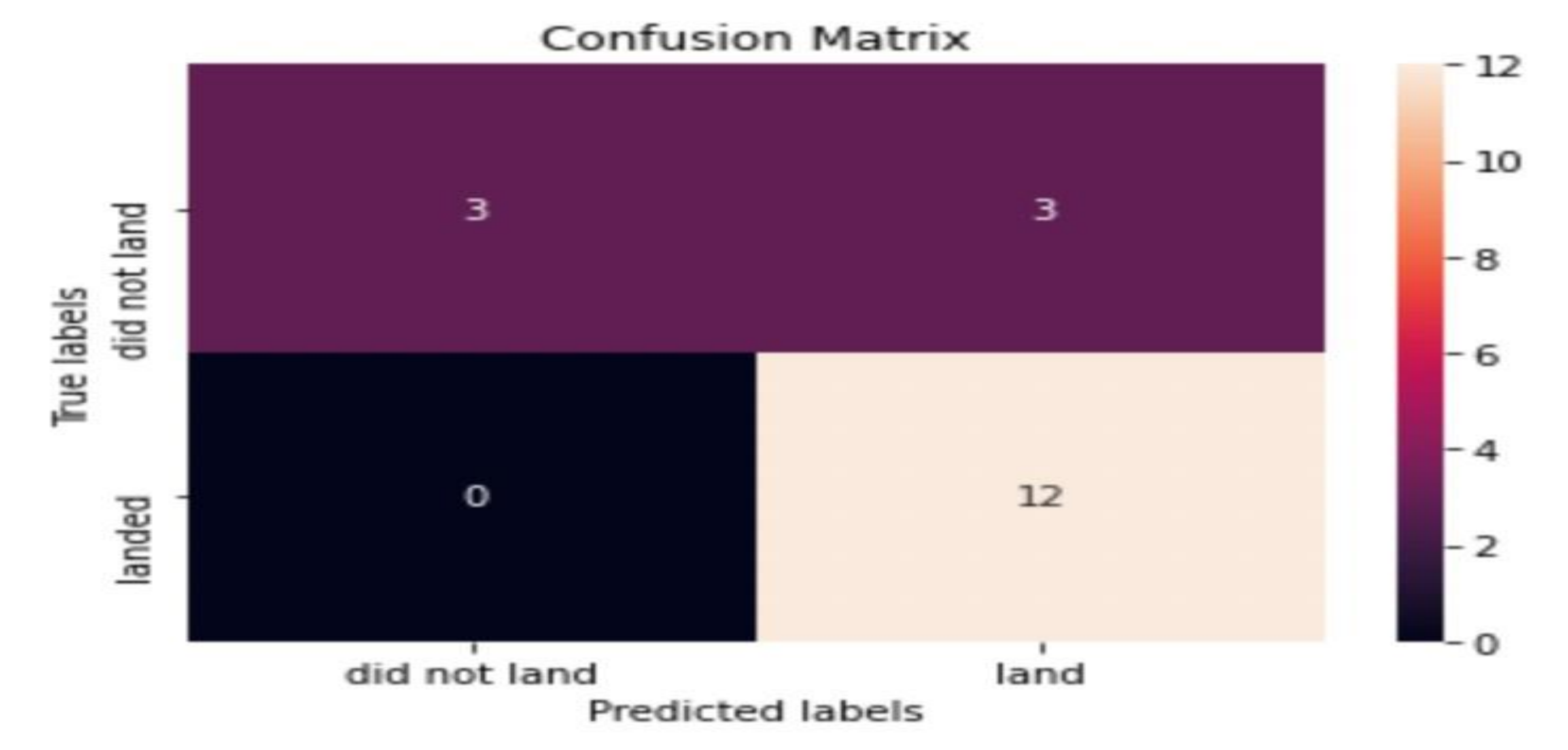
Build a Dashboard with Plotly Dash

Launch success count for all sites



Predictive analysis (Classification)

Confusion Matrix



Conclusion

- The Decision Tree Model is the most suitable algorithm for this dataset.
- Launches with a lower payload mass tend to show better results compared to those with a larger payload mass.
- Most launch sites are located near the Equator and are in close proximity to the coast.
- The success rate of launches has increased over the years.
- KSC LC-39A boasts the highest success rate among all the launch sites.
- Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate.