IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kate Nowak
August 2022

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
  - Unlike other rocket providers, SpaceX's Falcon 9 can recover the first stage. Sometimes the first stage does not land. Sometimes it will crash. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
  - SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- Questions we seek to answer with data
  - How do variables such as payload, launch site, number of flights, and orbits affect the success of the first stage landing?
  - With what accuracy can we predict whether a first stage will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Webscraping from Wikipedia with BeautifulSoup
- Perform data wrangling
  - Filtered the data
  - Dealt with missing values
  - Used one-hot encoding to prepare the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium
- Perform predictive analysis using classification models
  - Used GridSearchCV to optimize and compare logistic regression, SVM, and classification trees

Data Collection

- Method 1: made a get request to the SpaceX API to get IDs and information about each launch.
- Method 2: performed web scraping with BeautifulSoup to collect Falcon 9 historical launch records from Wikipedia.

We used both of these methods in order to get more complete information.

Data Collection – SpaceX API

Features collected:

- Flight Number
- Date
- BoosterVersion
- PayloadMass
- Orbit
- LaunchSite
- Outcome
- Flights
- GridFins
- Reused
- Legs
- LandingPad
- Block
- ReusedCount
- Serial
- Longitude
- Latitude

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/44ab4ac2313aaeba9f6a36b6d005a7601528802f/week1-1%20Data%20Collection%20API%20Lab.ipynb

Steps:

1. Request and parse SpaceX launch data using GET request
2. Filter the dataframe to only include Falcon 9 launches
3. Replaced missing PayloadMass values with mean

Data Collection - Scraping

Columns collected:

- Flight No.
- Date and time
- Launch site
- Payload
- Payload mass
- Orbit
- Customer
- Launch outcome

Notebook URL:

https://github.com/k8nowak/DS-Capstone/blob/985efcf0785efa7283fa5b3f661607e87eb70765/week%201-2%20Data%20Collection%20with%20Web%20Scraping%20Lab.ipynb

Steps:

1. Request Falcon 9 launch data from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
2. Create a BeautifulSoup object
3. Extract column names from HTML table headers
4. Create a dictionary launch_dict and then a dataframe df.

Data Wrangling

- We computed the number of launches on each site, and the number of flights to each orbit using value_counts() method.
- However the main goal of this process was to convert detailed outcome data in the dataset to a “Class” feature where “1” means the booster successfully landed and “0” means it was unsuccessful.
- This was accomplished by
  - enumerating the bad outcomes
  - generating a list with “0” if the flight contained a bad outcome and “1” otherwise
  - appending this list to the dataframe in a “Class” column.
- The success rate for all launches is ~67%.

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/d2de21fd8896eaf920120a6fc73dd2bfdb33f5ff/Week%201-3%20Data%20Wrangling%201.ipynb

EDA with Data Visualization - Feature Selection

We select the features that appear to have an affect on success rate:

- FlightNumber
- PayloadMass
- Orbit
- LaunchSite
- Flights
- GridFins
- Reused
- Legs
- LandingPad
- Block
- ReusedCount
- Serial

We create dummy variables to categorical columns with one-hot encoding.
Then, cast all numeric columns to float64.

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/0326996ad18b67bc29302ebfff7a9014e3f0a99b/week%202-2%20EDA%20with%20Visualization.ipynb

EDA with SQL

Note: I had trouble accessing the “Landing _Outcomes” column, so I renamed it to “LO”.

SQL queries performed:

1. Select names of unique launch sites.
2. Select 5 records where launch sites begin with ‘CCA’
3. Select total payload mass carried by boosters launched by NASA (CRS)
4. Select average payload mass carried by booster version F9 v1.1
5. Find the date when the first successful landing outcome in ground pad was achieved. (2015-12-22)
6. Select 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. Using a subquery, 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

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/a9357ba04228331496a37ded675500edc8f0d218/Week%202-1%20EDA%20with%20SQL%20Lab.ipynb

Build an Interactive Map with Folium

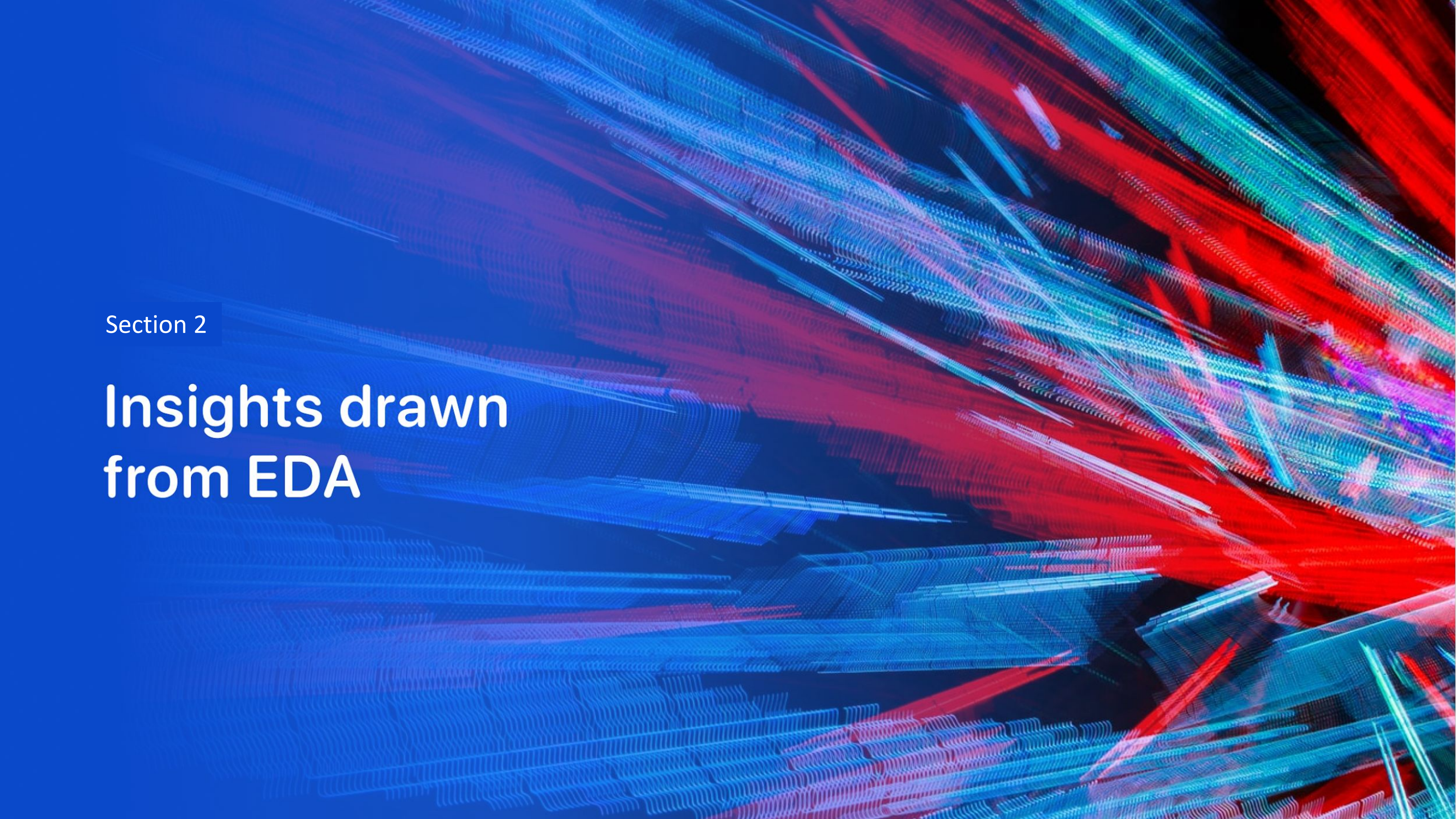
1. Marked all the launch sites on the map with a circle.
2. Marked the success/failed launches for each site on the map using a green or red marker in a marker_cluster.
3. Calculated the distance between a launch site to the coastline, and added a line showing this distance to the map.

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/3d00f2a6b17b6f97be5c5199dc817e751a0f5be4/Week%203-1%20Folium.ipynb

Predictive Analysis (Classification)

- Used previously-described features (standardized) as X, and [‘Class’] as Y.
- Created train and test sets with test_size of 20%.
- Used GridSearchCV to try a variety of parameters for logistic regression, SVM, Decision Tree Classifier, and K Nearest Neighbors.
- Scored the test data and plotted a confusion matrix for each.
- Each of the models performed the same on the test data, with a score of ~83% and an identical confusion matrix.

Notebook URL:
https://github.com/k8nowak/DS-Capstone/blob/694558b9ccd549d645c24a8fe7fa487193ed39a6/Week%204-1%20Machine%20Learning.ipynb

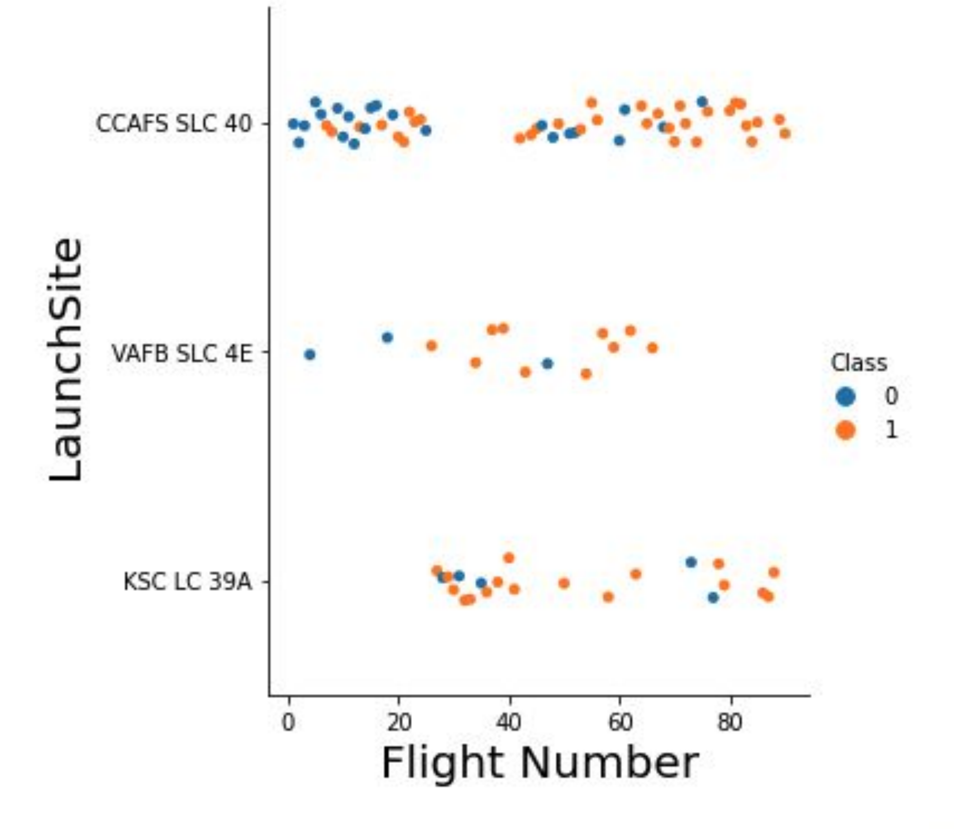


Section 2

Insights drawn from EDA

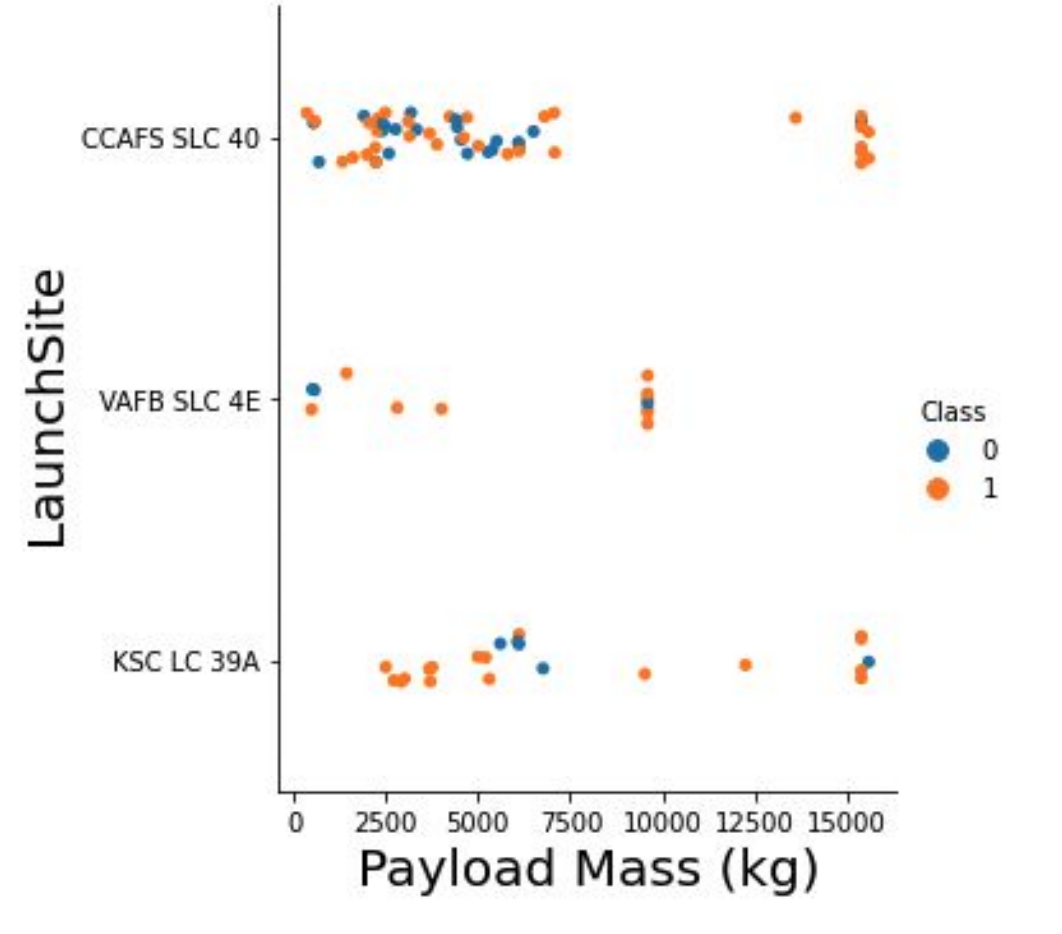
Flight Number vs. Launch Site

A visualization of flight number vs. launch site indicates that later flights were more successful than earlier flights, and the sites VAFB and KSC had greater success rates than CCAFS.



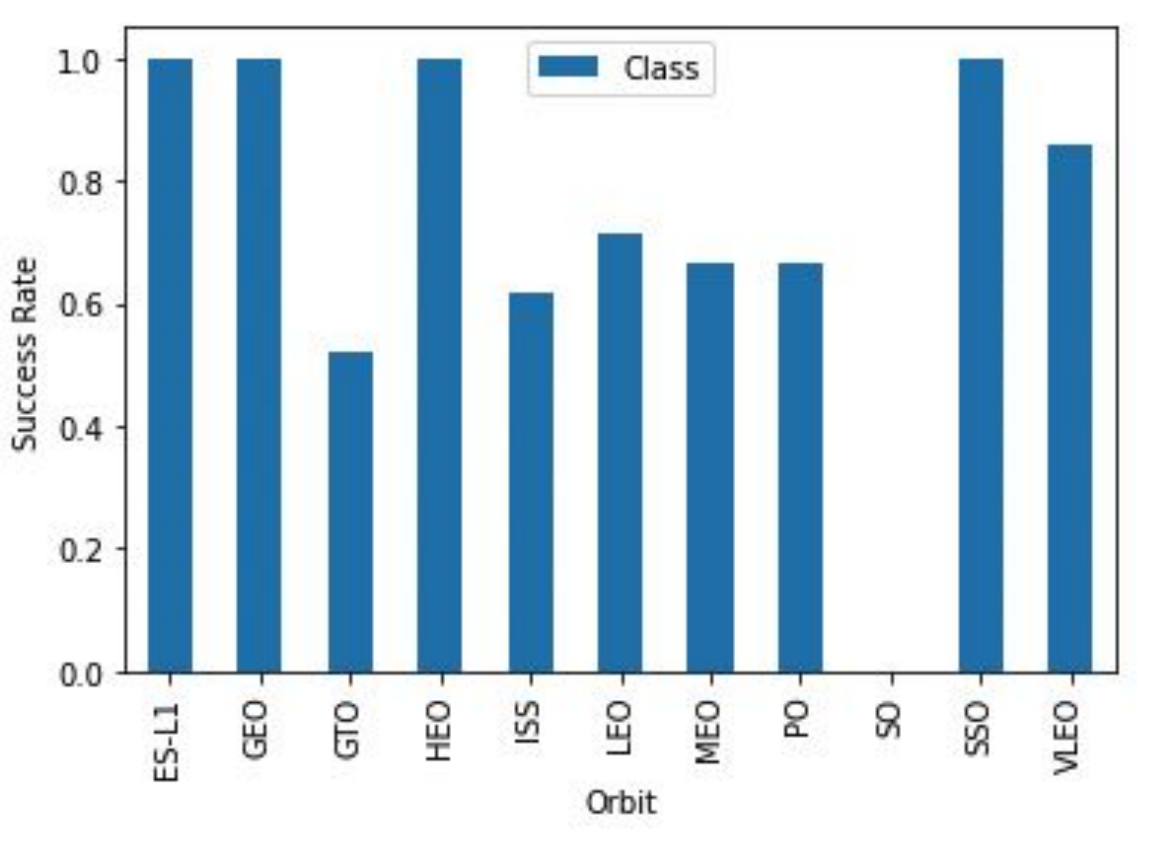
Payload vs. Launch Site

A visualization of payload mass vs. launch site indicates that VAFB was not used to launch payload more massive than 10,000 kg. Overall there were more launches of lighter payloads than heavier.



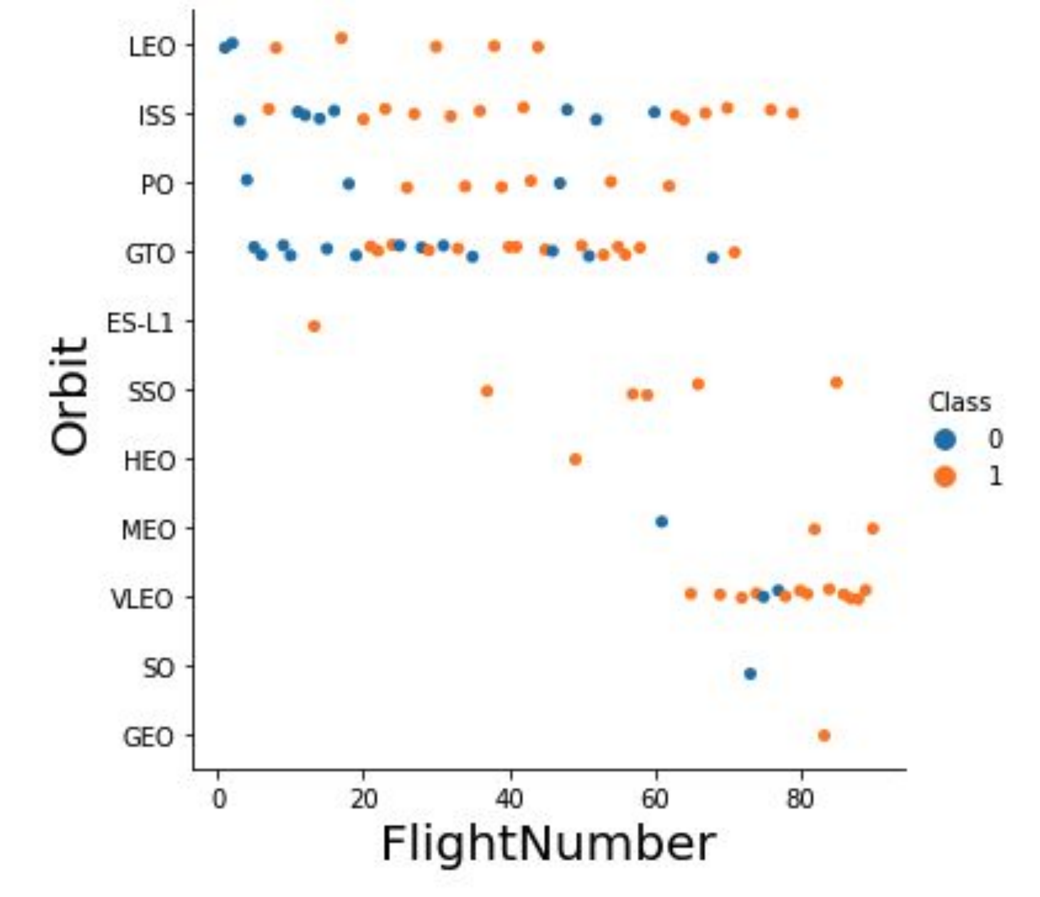
Success Rate vs. Orbit Type

The orbits with a 100% success rate were ES-L1, GEO, HEO, and SSO. The orbit with the lowest success rate of ~50% was GTO. Success rates for other orbits are shown.



Flight Number vs. Orbit Type

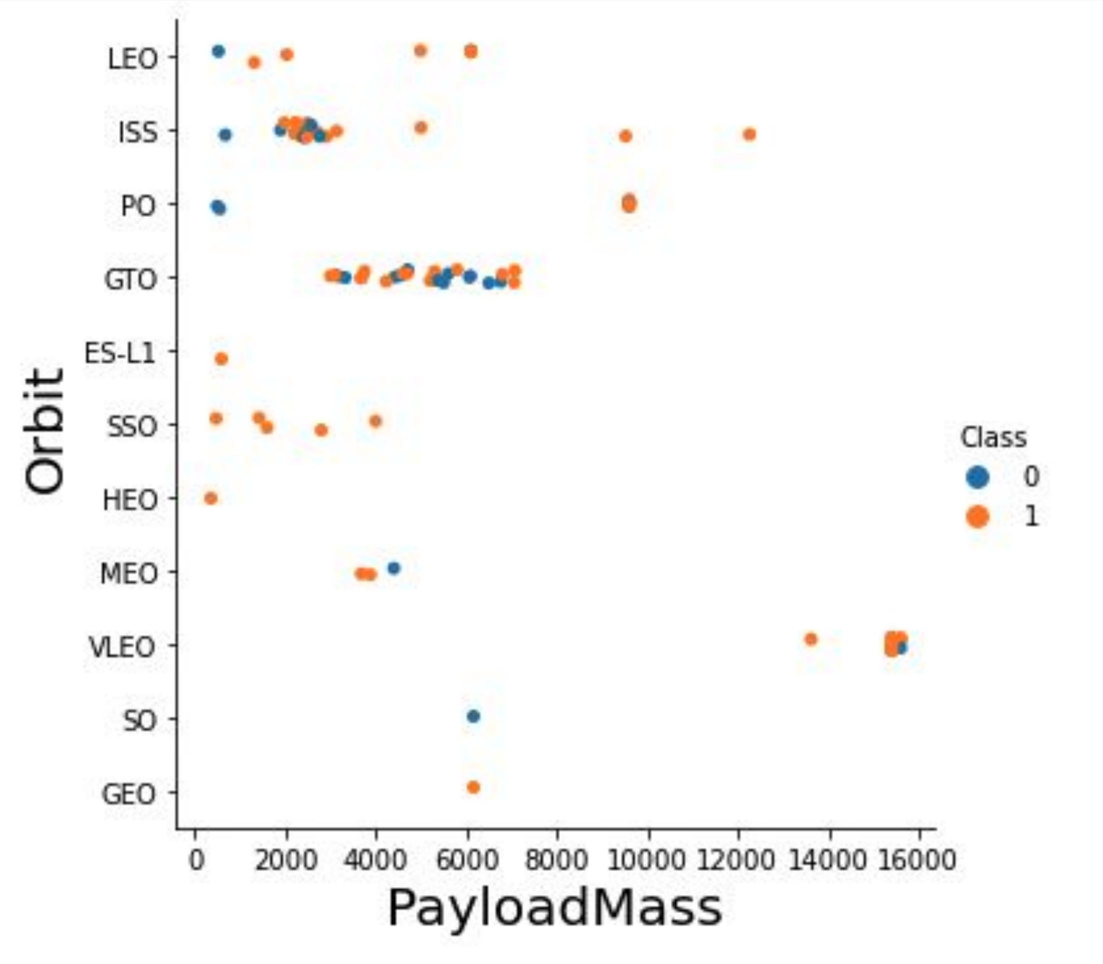
Flight number vs orbit reveals patterns. Some orbits were not attempted until later flight numbers. In the LEO orbit success appears related to the number of flight, but there seems to be no relationship between flight number and success for GTO orbit.



Payload vs. Orbit Type

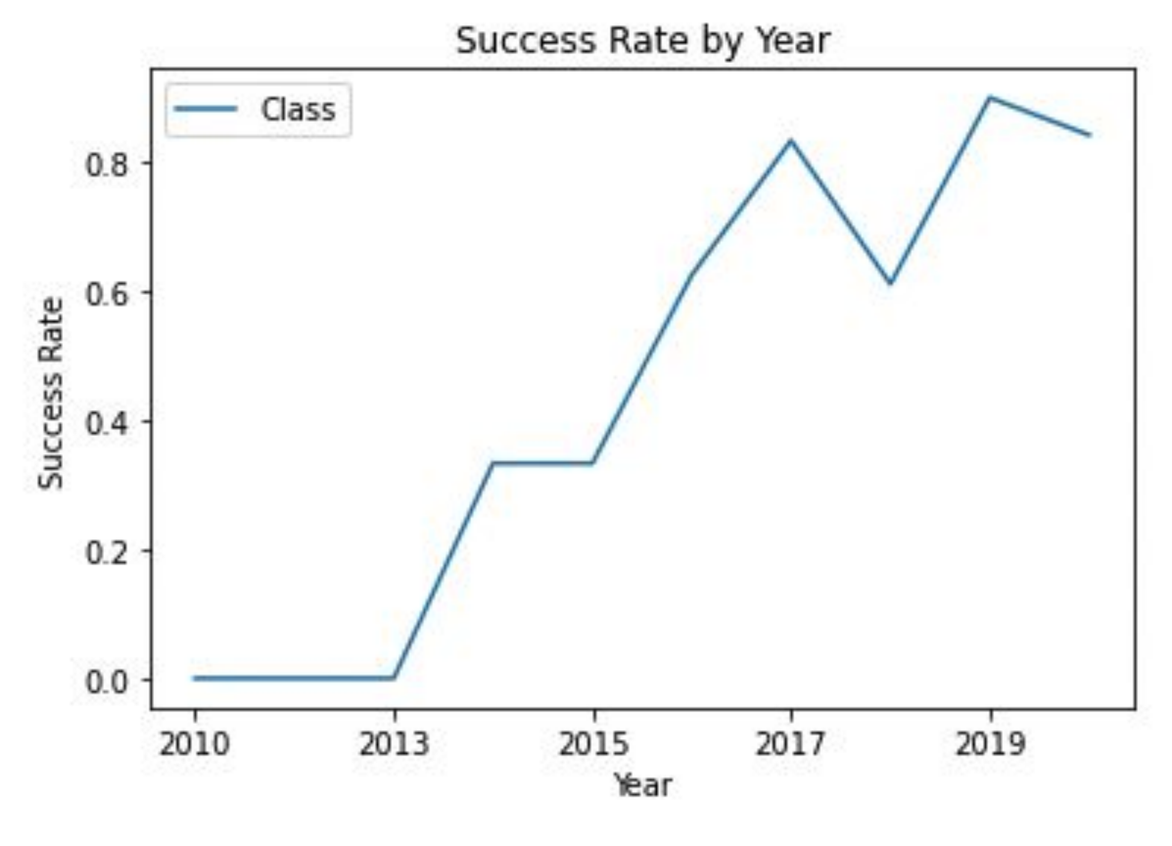
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO, and ISS.

However for GTO we cannot distinguish this well as both successful and unsuccessful missions are seen.

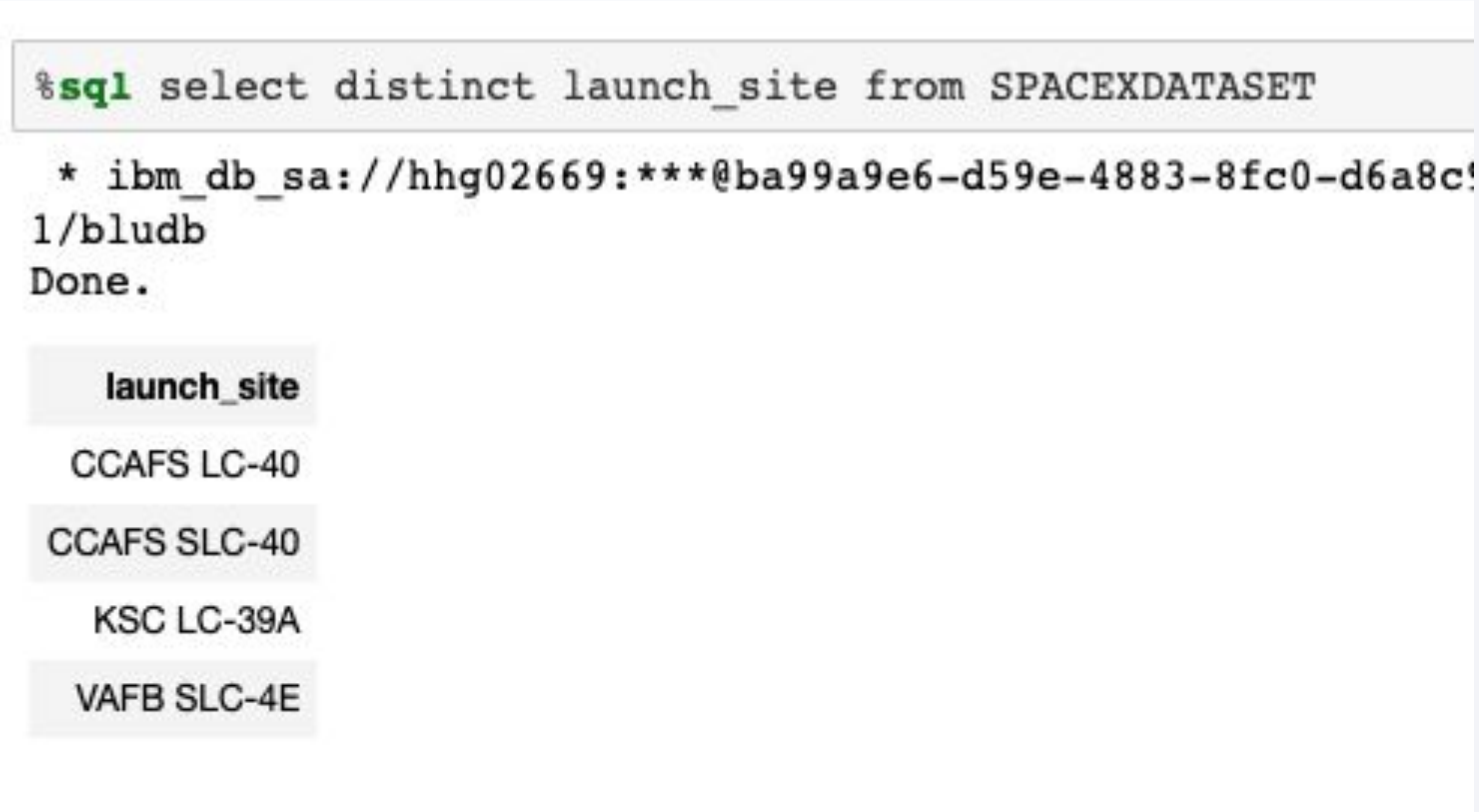


Launch Success Yearly Trend

Year over year, the success rate has tended to increase.



All Launch Site Names

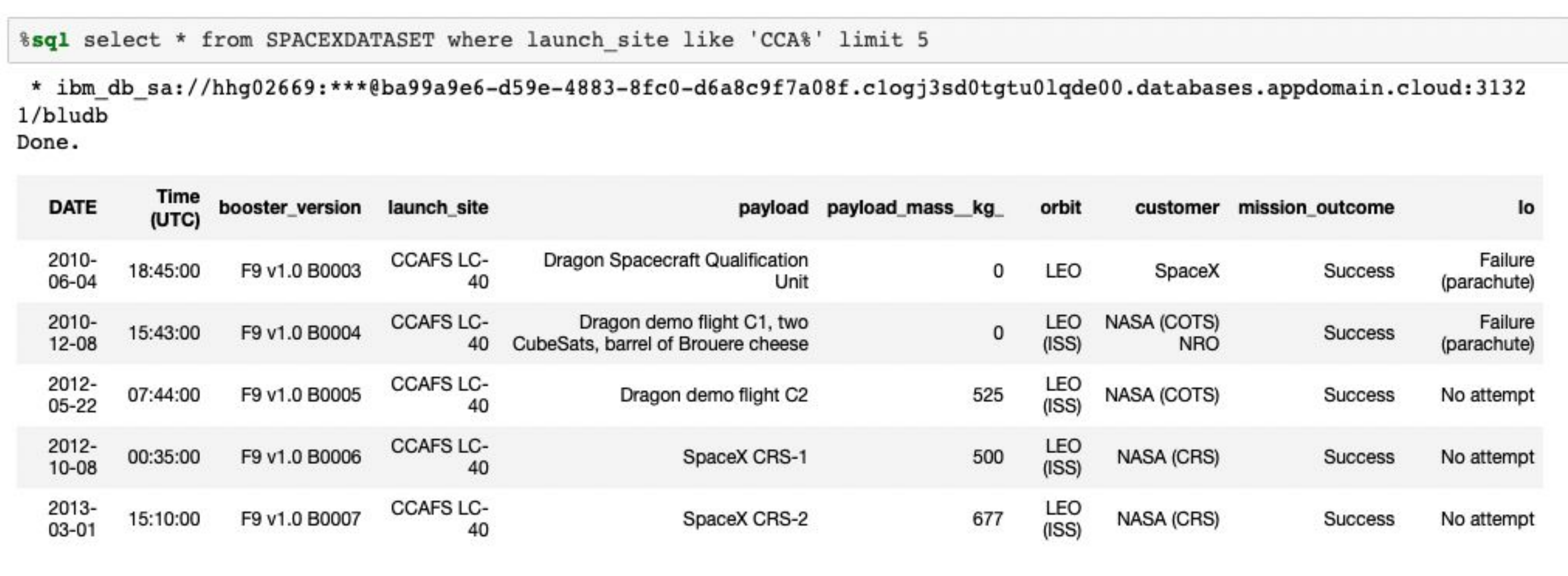


```
%sql select distinct launch_site from SPACEXDATASET
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c
1/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

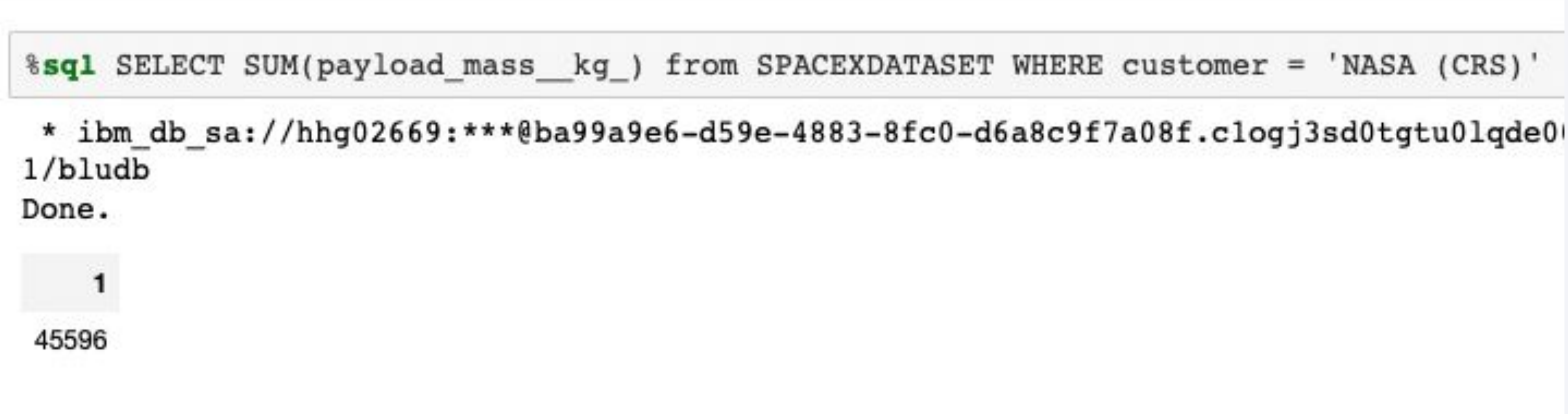


```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:3132
1/bludb
Done.
```

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	lo
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

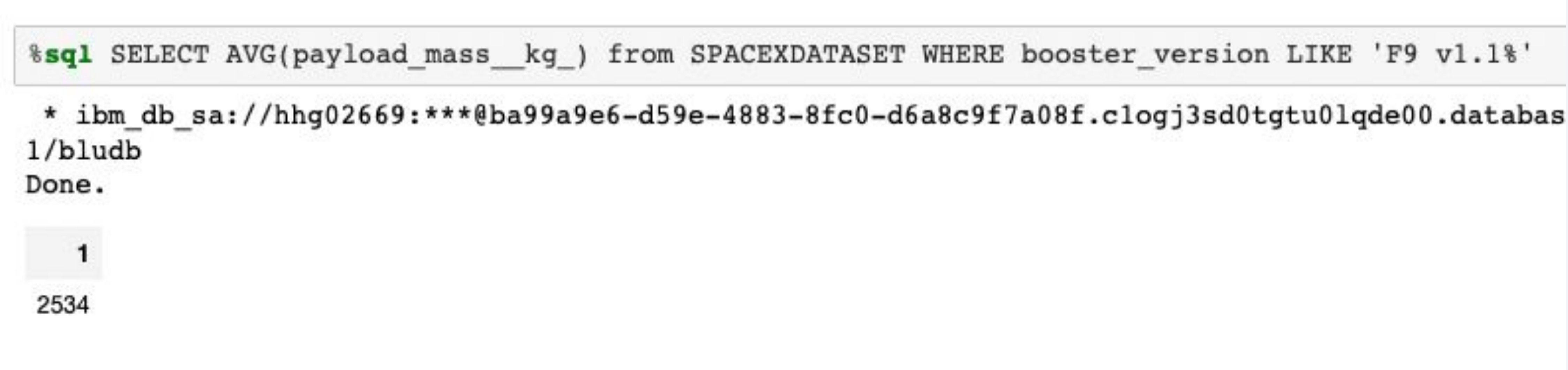


```
%sql SELECT SUM(payload_mass__kg_) from SPACEXDATASET WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde0
1/bludb
Done.

    1
45596
```

Average Payload Mass by F9 v1.1

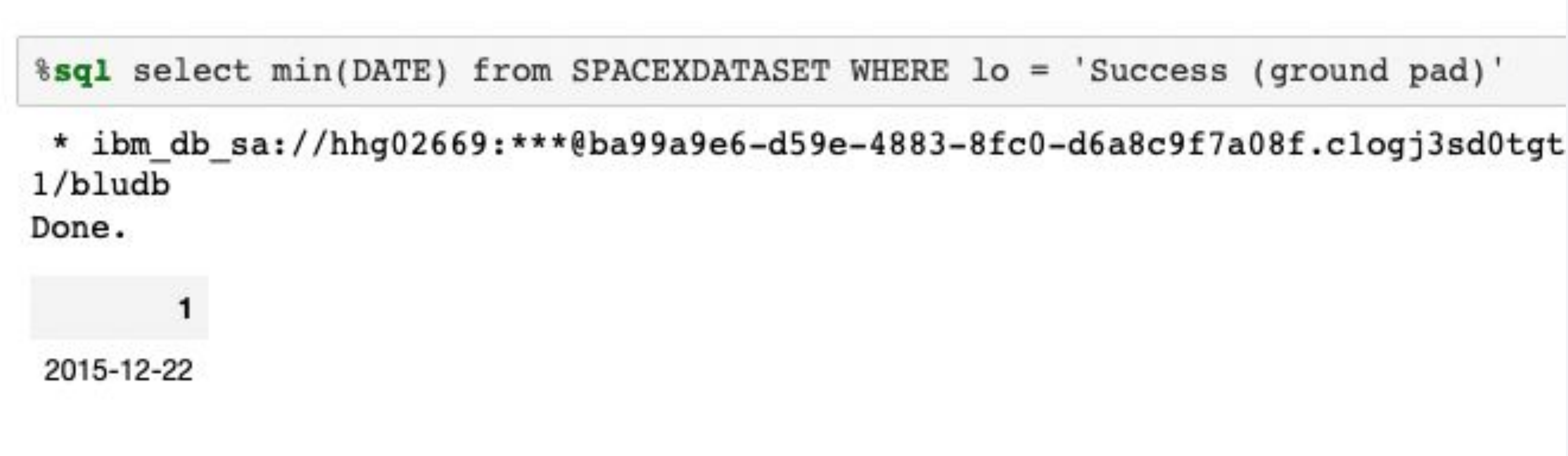


```
%sql SELECT AVG(payload_mass__kg_) from SPACEXDATASET WHERE booster_version LIKE 'F9 v1.1%'
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databas
1/bludb
Done.
```

1
2534

First Successful Ground Landing Date

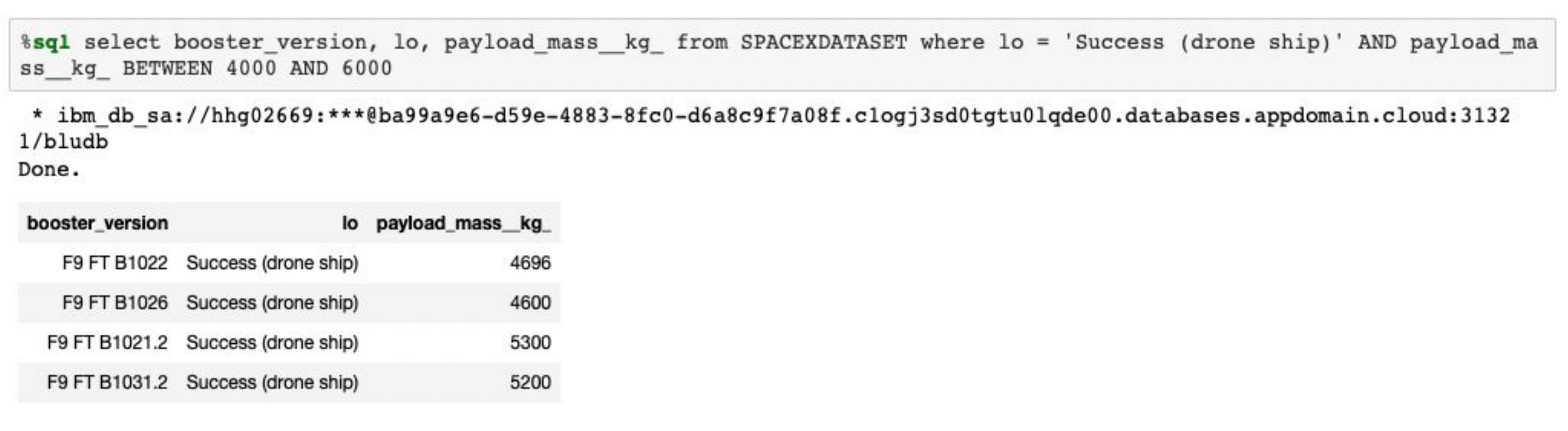


```
%sql select min(DATE) from SPACEXDATASET WHERE lo = 'Success (ground pad)'
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgt
1/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

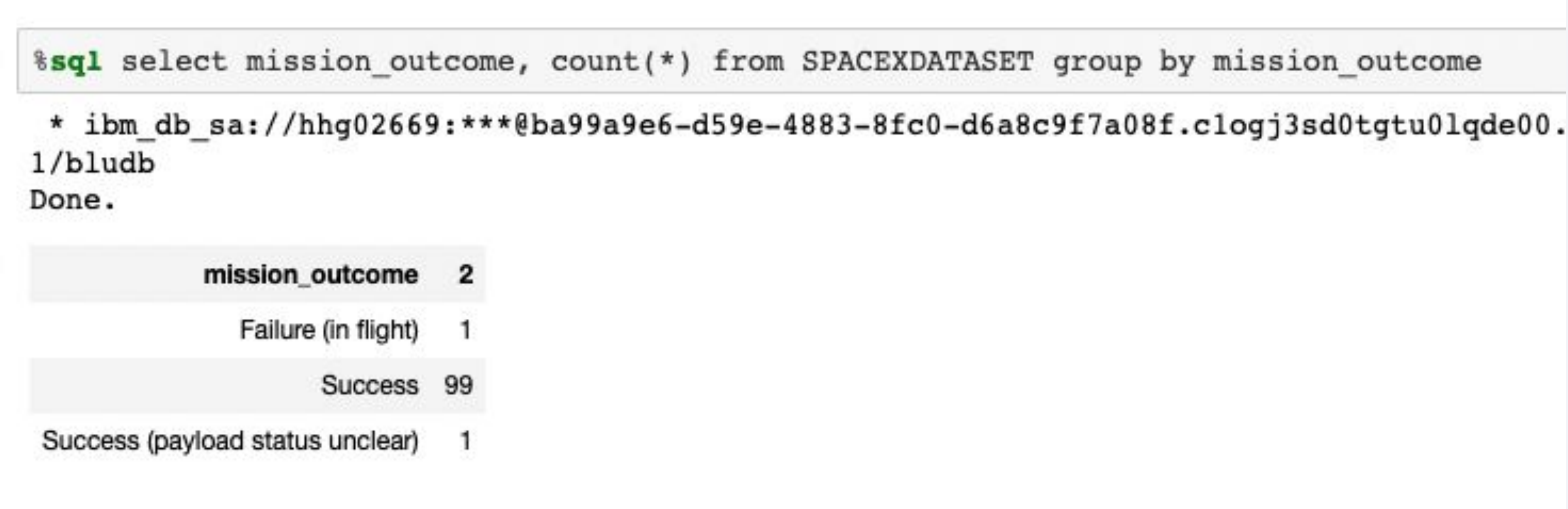


```
%sql select booster_version, lo, payload_mass__kg_ from SPACEXDATASET where lo = 'Success (drone ship)' AND payload_ma
ss__kg_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:3132
1/bludb
Done.
```

booster_version	lo	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

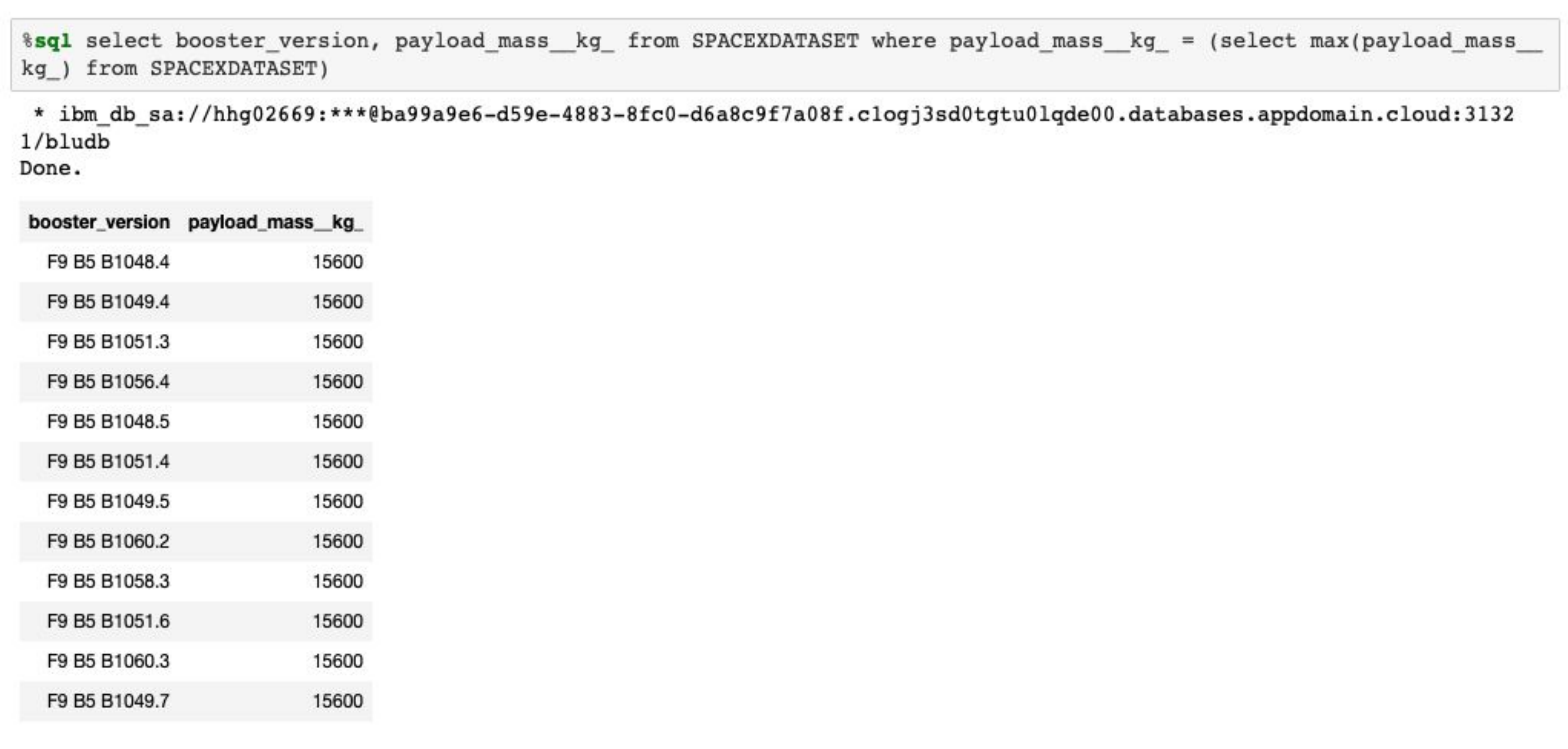


```
%sql select mission_outcome, count(*) from SPACEXDATASET group by mission_outcome
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.
1/bludb
Done.
```

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

Boosters Carried Maximum Payload

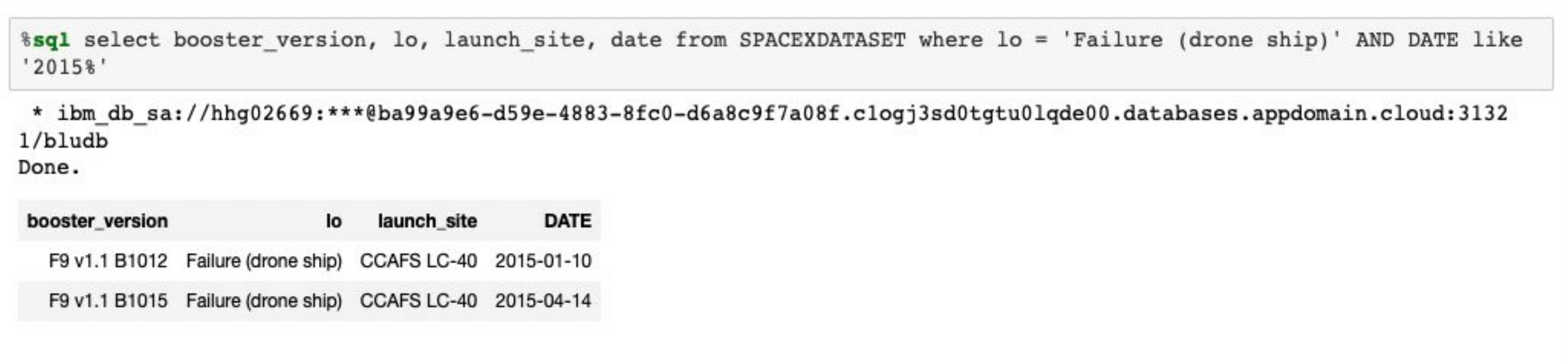


```
%sql select booster_version, payload_mass__kg_ from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__
kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:3132
1/bludb
Done.
```

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

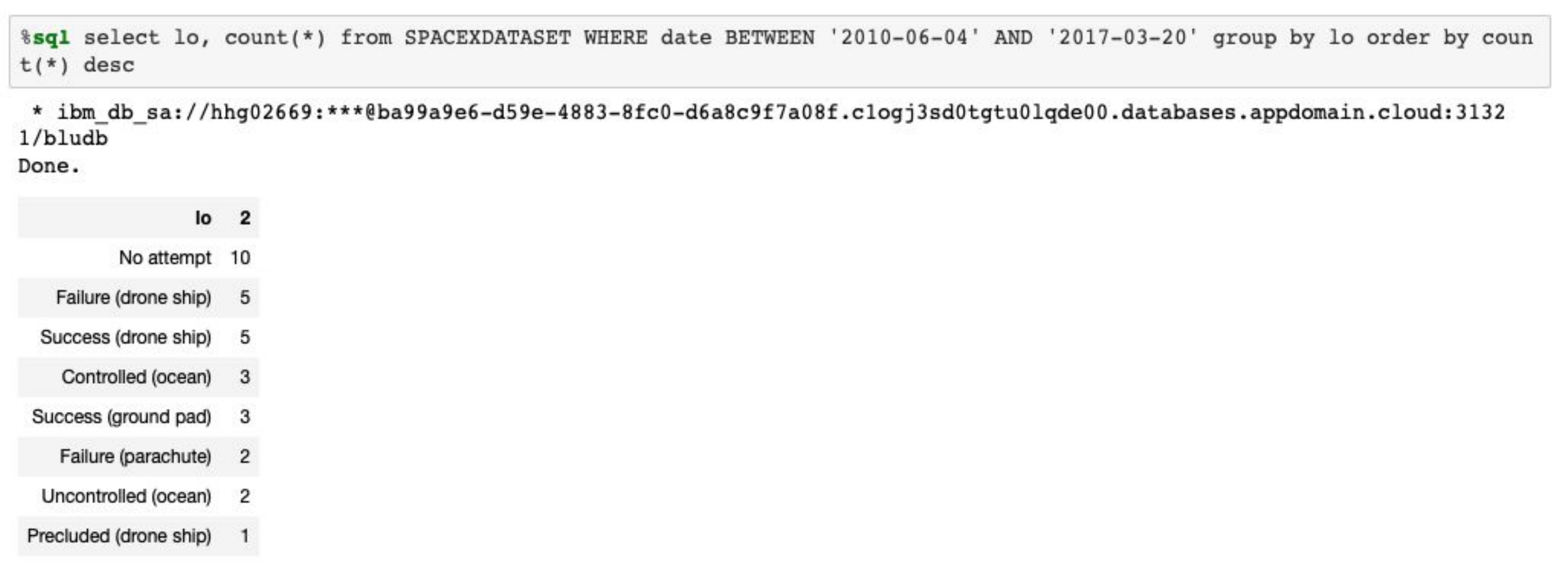


```
%sql select booster_version, lo, launch_site, date from SPACEXDATASET where lo = 'Failure (drone ship)' AND DATE like
'2015%'
```

```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:3132
1/bludb
Done.
```

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

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



```
%sql select lo, count(*) from SPACEXDATASET WHERE date BETWEEN '2010-06-04' AND '2017-03-20' group by lo order by coun
t(*) desc
```

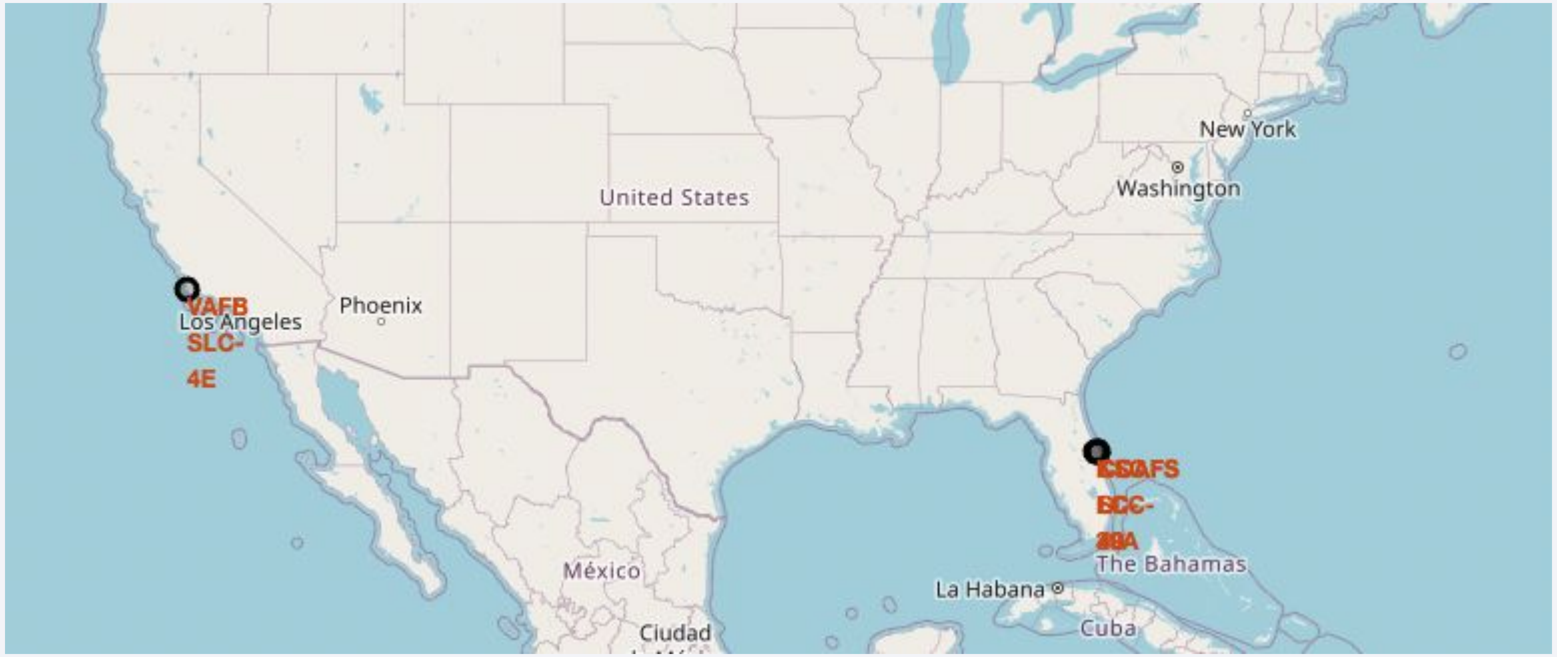
```
 * ibm_db_sa://hhg02669:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:3132
1/bludb
Done.
```

lo	2
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

Section 3

Launch Sites Proximities Analysis

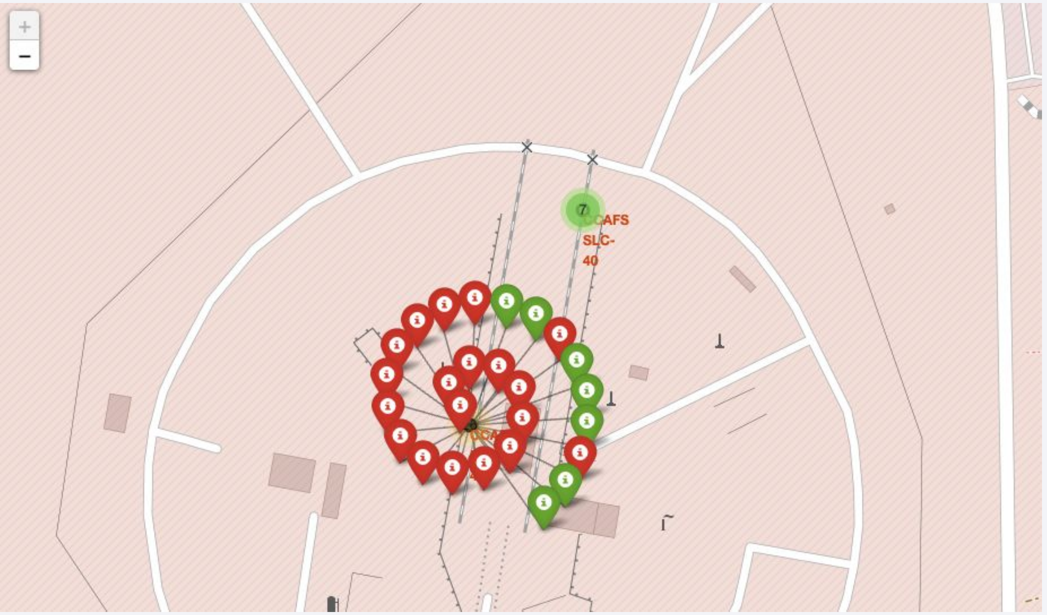
Launch Sites



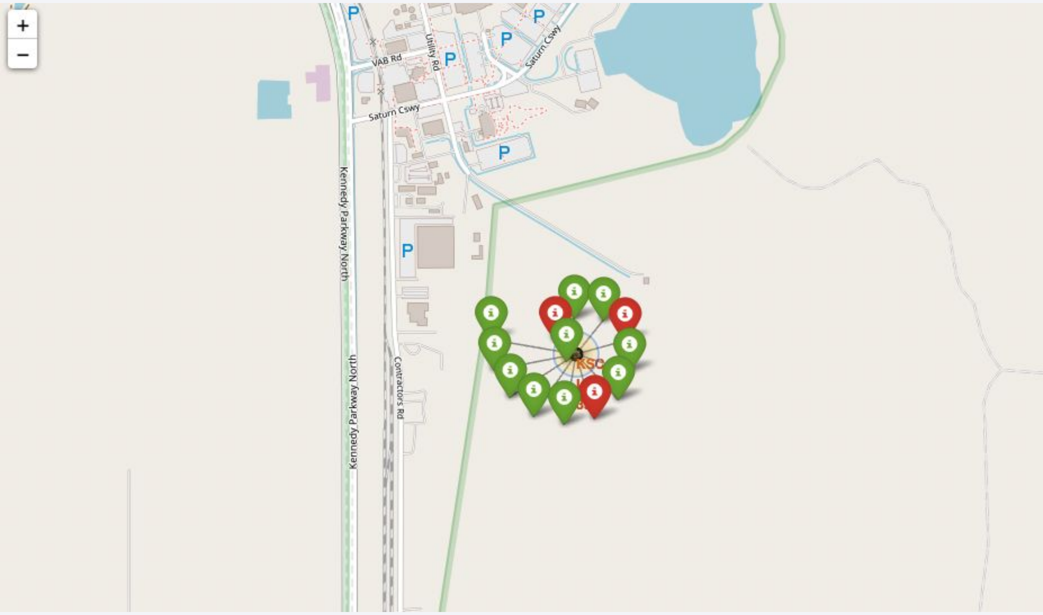
Launch sites are located at two coastal areas in the United States.

<Folium Map Screenshot 2>

Each launch site can be expanded to show number of successful and failed launches.
Two examples are screenshotted.

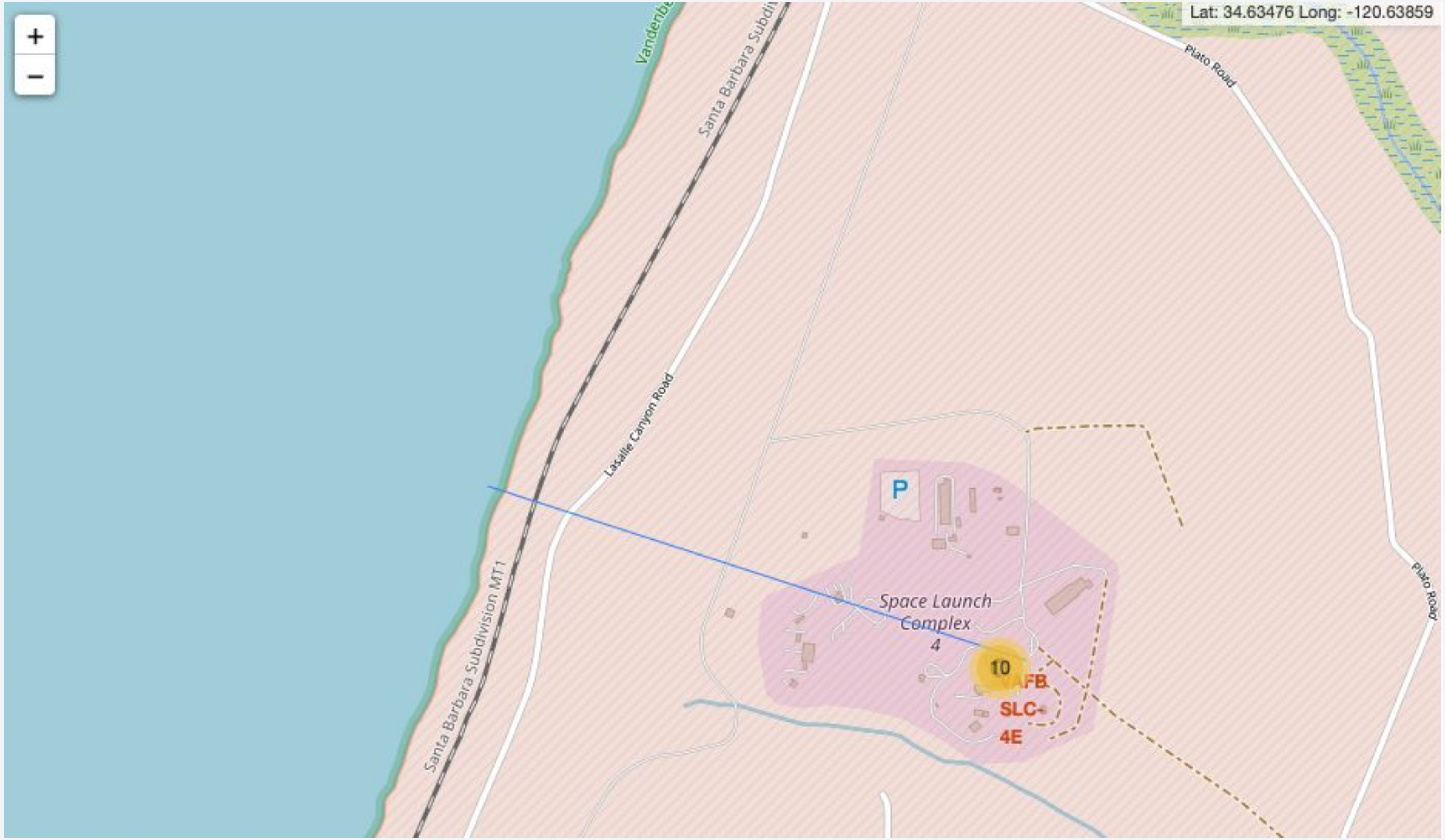


CAFS LC-40



KSC LC-39A

Distance from launch site to coast is plotted



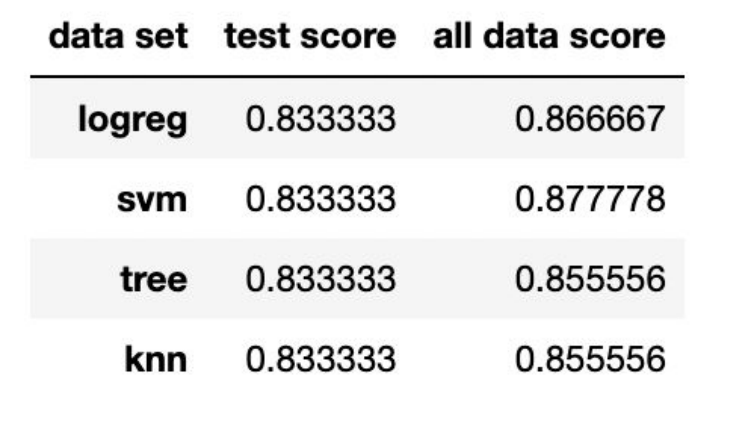
Section 5

Predictive Analysis (Classification)

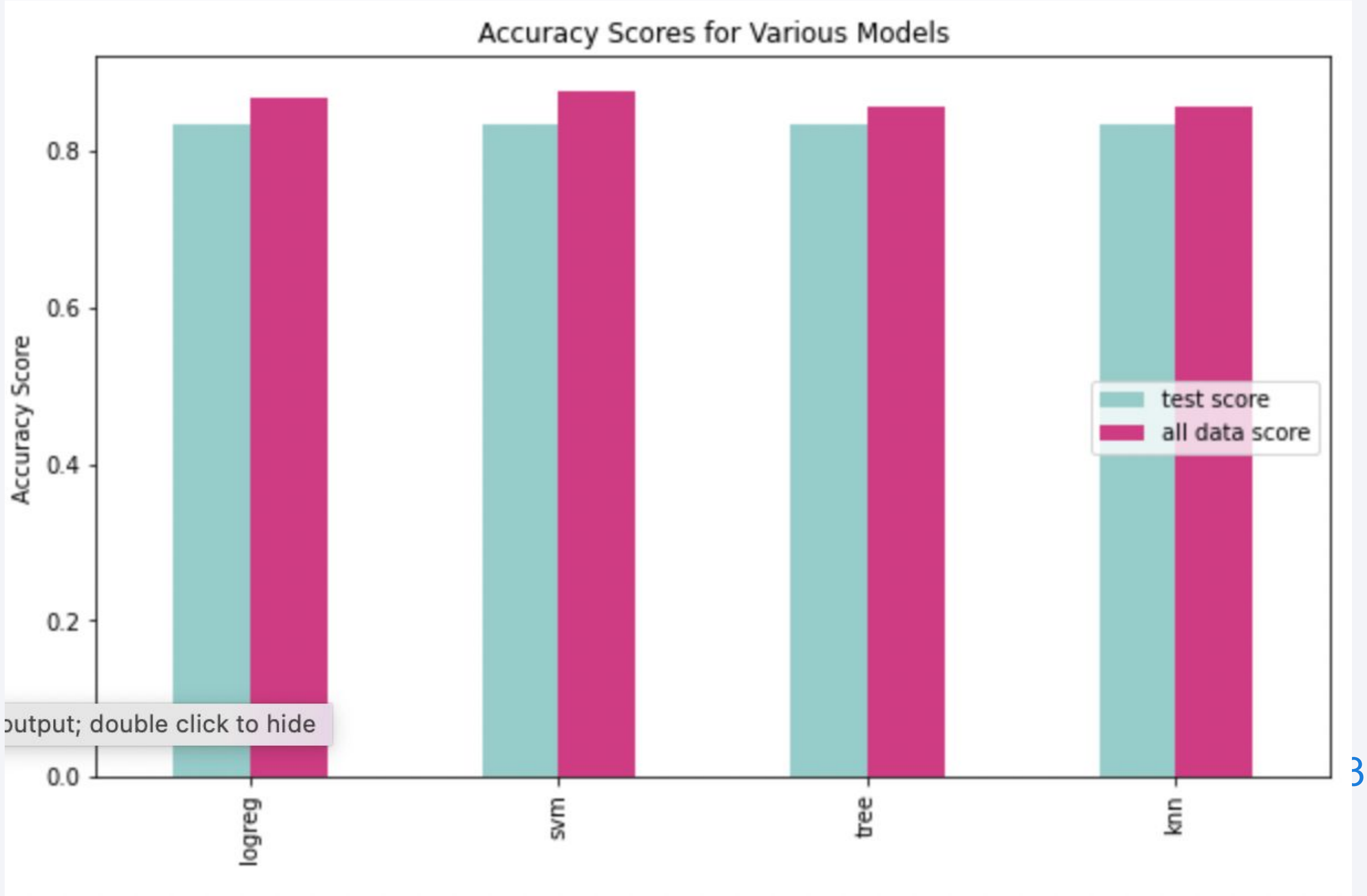
Classification Accuracy

- Every model that was tried performed the same on the test data, with slight differences when scoring all data.

- This lack of differentiation may be due to the relatively small size of the data set used to train the model.

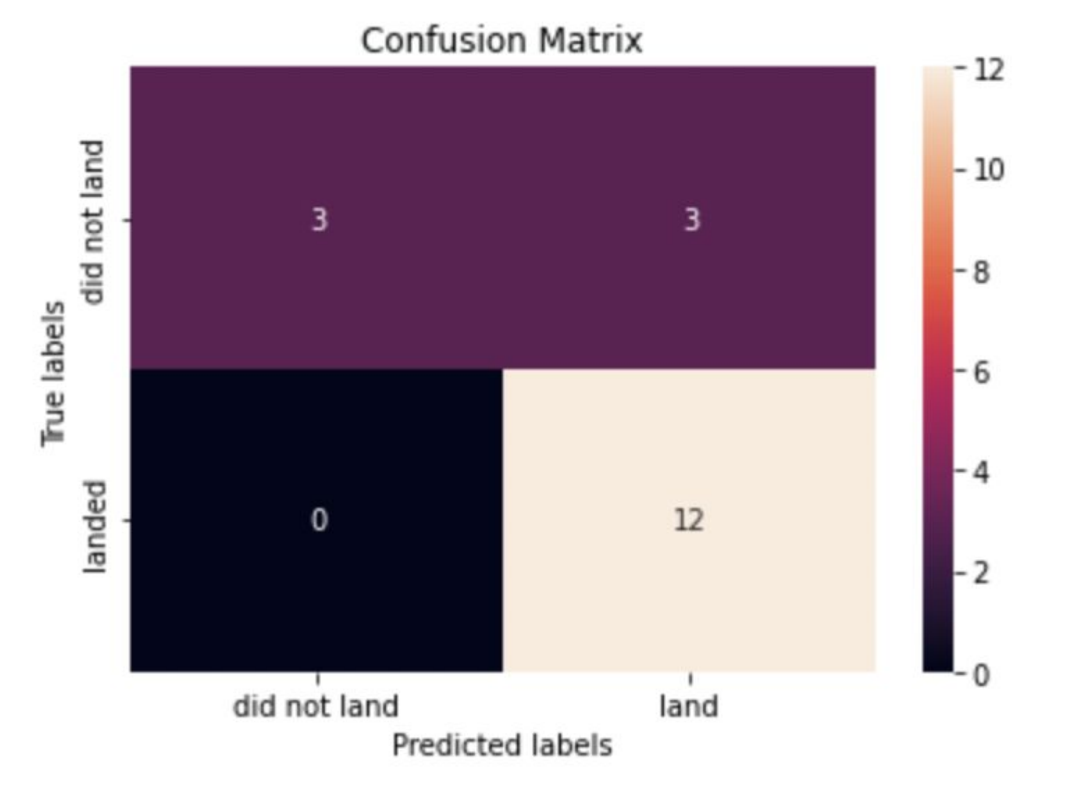


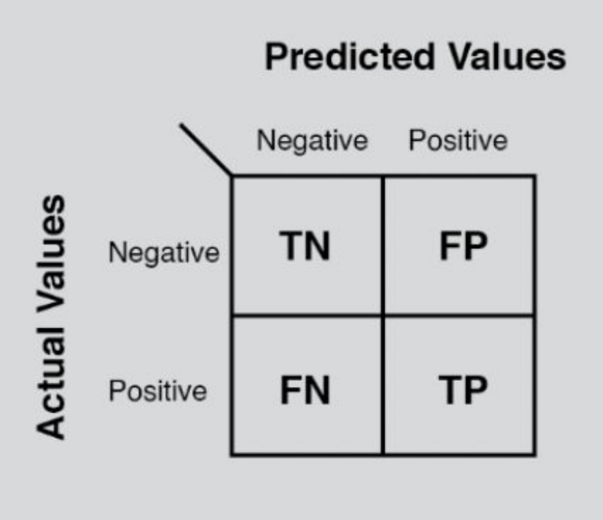
data set	test score	all data score
logreg	0.833333	0.866667
svm	0.833333	0.877778
tree	0.833333	0.855556
knn	0.833333	0.855556



Confusion Matrix

- The confusion matrix was identical for each model.
- We can distinguish between different classes, and the major concern is false positives.





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

Conclusions

- There was no clear best model for this dataset. A good next step would be to analyze a larger data set.
- Launches with lower payload mass have a higher success rate than launches with a higher payload mass.
- All of the launch sites are in close proximity to a coastline, and located in lower latitudes.
- The success rate of launches increases year over lear.
- KSC LC-39A has the highest success rate of all the launch sites.
- Orbits GEO, HEO, SSO, and ES-L1 have 100% success rates (but for a small number of launches).

Thank you!