Data Science Capstone project

Dovber Kaplan

25 August 2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results (Visual charts, dashboard)
- Conclusion
- Appendix

Executive Summary

- The project done aims to predict if the Falcon 9 SpaceX rocket first stage will land successfully. The project will provide EDA regarding SpaceX rocket launches and first stage landings. All data was gathered by web-scraping as well as requesting data from the SpaceX open API. It was then processed using data manipulation tools in Python such as Pandas. EDA was provided using tools and SQL queries. Logistic regression, SVM Decision tree and KNN was trained and optimized to predict the success rate of the first stage landing. The best performing model was then chosen by the accuracy score.
- EDA shows a list of different factors that affect the success in question. 80%+ accuracy was achieved for the predictions. The false positives rate is the point for further improvements of predicting model based matrices.

Introduction

- SpaceX is a company that provides Falcon 9 rocket launches. The cost of a launch is $62 mil, vs other companies which offer similar services, at around $165 mil. This is since first stage can be reused with SpaceX. If the success rate of the first stage can be determined, the cost of launch can be predicted.

- We want to find out what the success rate of Falcon 9's first stage actually is to determine what the cost of launch will be.

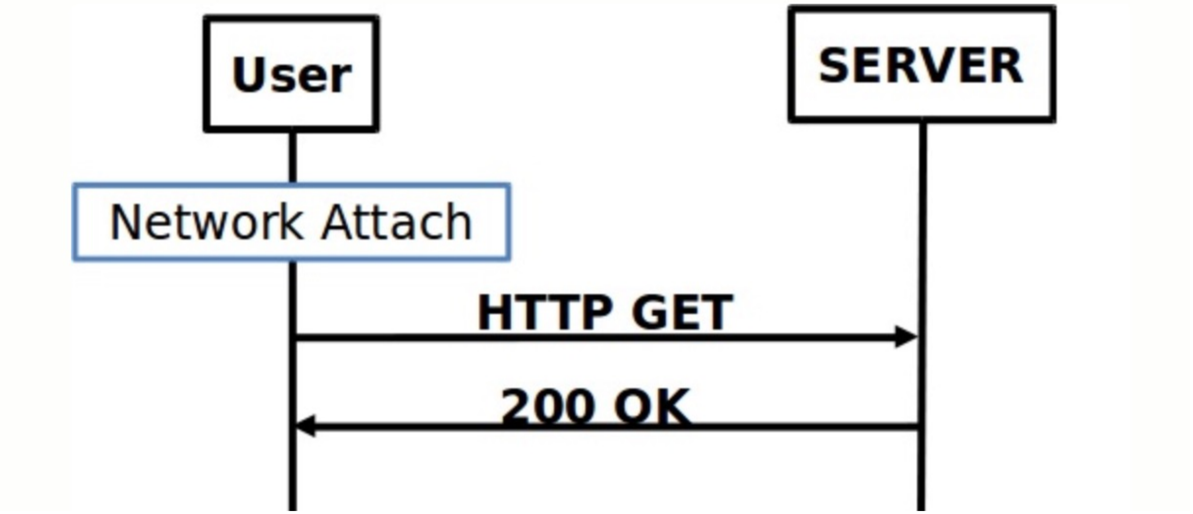
Methodology

- Data collection methodology:
  - Data was web-scraped from the related Wikipedia article.
  - Data was also retrieved from the official SpaceX open REST API.
- Perform data wrangling
  - Irrelevant records were first removed. We then replaced missing data with mean values. We examined the data-types in the data sets.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - The data and features of the dataset were visualized with plots and charts.
  - SQL queries were used to gain additional insights into the data.
  - Models for prediction were prepared based on visual analysis results.
- Perform interactive visual analytics using Folium and Plotly Dash
  - Plotly was used to visualize spaceX data related dashboard as well as Folium maps, to more easily analyze our data.
- Perform predictive analysis using classification models
  - Features were standardized and logistic regressions, SVM, decision trees and KNN methods were used for predictive models.

Data collection – SpaceX API

https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/api.ipynb

The data was pulled from the official SpaceX API, which provides documentation as my work shows in the above file.

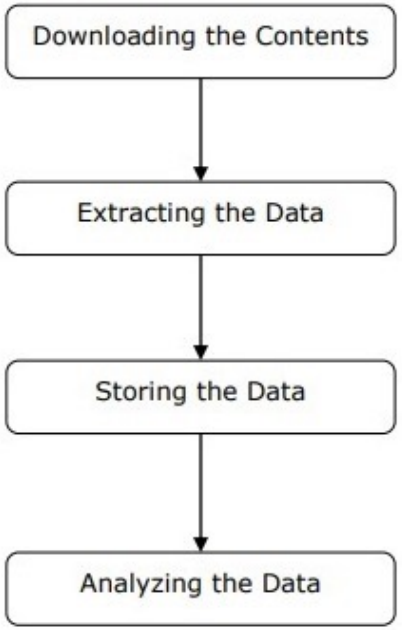


Get requests were used to retrieve the required data from the API.

Data collection – Web scraping

https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/webscraping.ipynb

The data was pulled from the the relavent Wikipedia article. The HTML was then processed to retrieve the data we needed.



Downloading the Contents

Extracting the Data

Storing the Data

Analyzing the Data

A Python library, BeautifulSoup was used to process the HTML we retrieved.

Data wrangling

- Irrelevant records were first removed.
- We then replaced missing data with mean values.
- We examined the data-types in the data sets to check they were all easy to work with.
- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/data-wrangling.ipynb

EDA with data visualization

- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/visualization.ipynb
- A scatter plot was used for the following
  - Landing outcome for flight number vs payload mass (This was used to check if landing success rate increased in later flights and to see if success is higher for higher payload masses.)
  - Landing outcome for flight number vs launch site (This was used to check distribution of launches between the launch sites. We checked if the rate of success changed in the later launches)
  - Landing outcome for payload mass vs launch site (This was used to check the distribution of launches with different payload masses, and to see if this affects success rates.)
  - Landing outcome for payload mass vs orbit type (This was used to check the influence on payload mass on success rates for different orbit types.)
  - Landing outcome for flight number vs orbit type (This was used to check older vs later flight successes in different orbit types.)

EDA with data visualization

- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/visualization.ipynb
- A bar plot was used for the following
  - Success rate for orbit type. (This was used to check if different orbits have different success rates.)
- A line chart was used for launch successes per year. (This was used to check the success change from older (~2013) to later (~2020) flights.

EDA with SQL

- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/eda.ipynb
- SQL queries:
  - Select names of unique launch sites
  - Select total payload mass by bossters launched by NASA (CRS)
  - Select average payload mass carried by VF9 v1.1
  - Select dates of the first successful landing on ground-pad was achieved.
  - Select the boosters which have succeeded and have payload mass between 4000 and 6000.
  - Select successes total vs failures total.
  - Select the names of the booster versions which carried the max payload mass.
  - Select the successful landings between 2010/06/4 and 2017/3/20

Build an interactive map with Folium

- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/analytics-and-dashboard.ipynb
- Circle, popup labels and text labels were added for:
  - NASA Johnsohn Space Center locations, to show the NASA commands centers.
  - Launch sites to show all the launch sites
- Color labeled markers were added for the success or failures per launch site so we can easily were successes or failures were per launch site.
- Polyline and text labels were added for the nearest railways, cities, coasts and highways, to measure the distances.

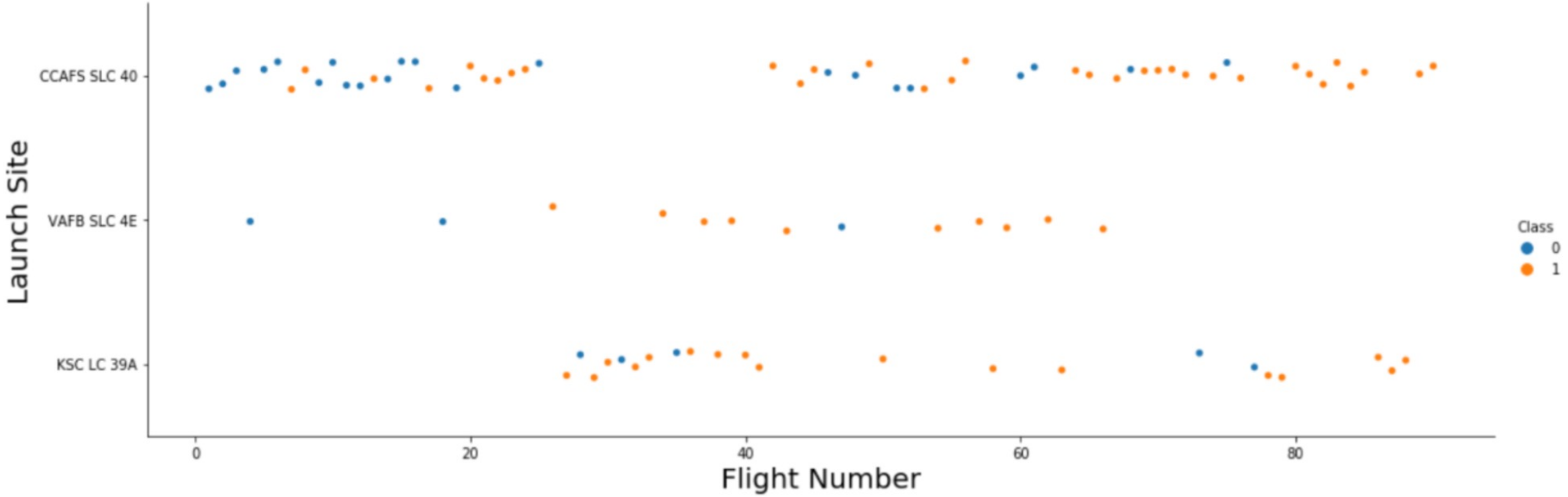
Build a Dashboard with Plotly Dash

- https://github.com/Droidking18/applied-data-science-capstone-IBM/blob/master/spacex-dash.py
- Launch site dropdowns were added to enable interactive launch site selection.
- Pie chart of successful launches was added to show success vs failure per launch site.
- A slider for payload mass range was added, in order to examine the success rate per payload mass.
- A scatter chart for booster version for success rate vs payload mass to show the correlation between payload mass and launch success for each booster.

EDA with Visualization

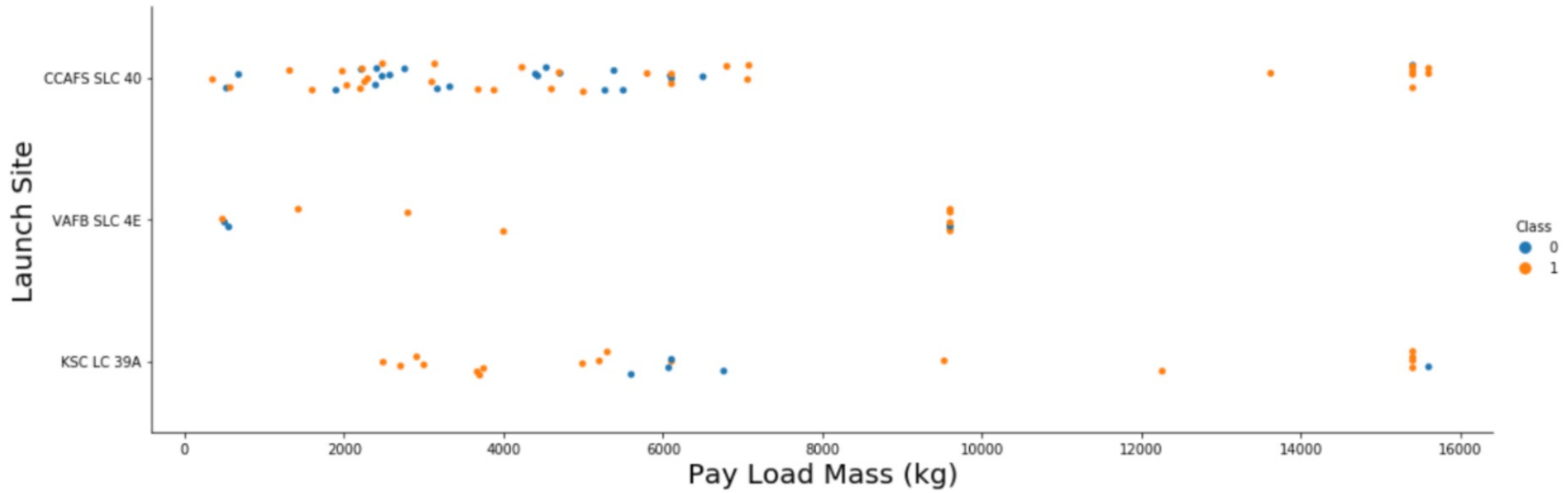
Flight Number vs. Launch Site

- CCAFS SLC 40 is used the most.
- It also has the highest number of failed launches.
- After flight 78, all launches were successful.



Payload vs. Launch Site

- The higher the payload mass is, the higher the success.
- KSC LC 39A has the highest success rate, but payload between 5000 kg and 7000 kg had issues.
- Most failure was payload below 7000kg



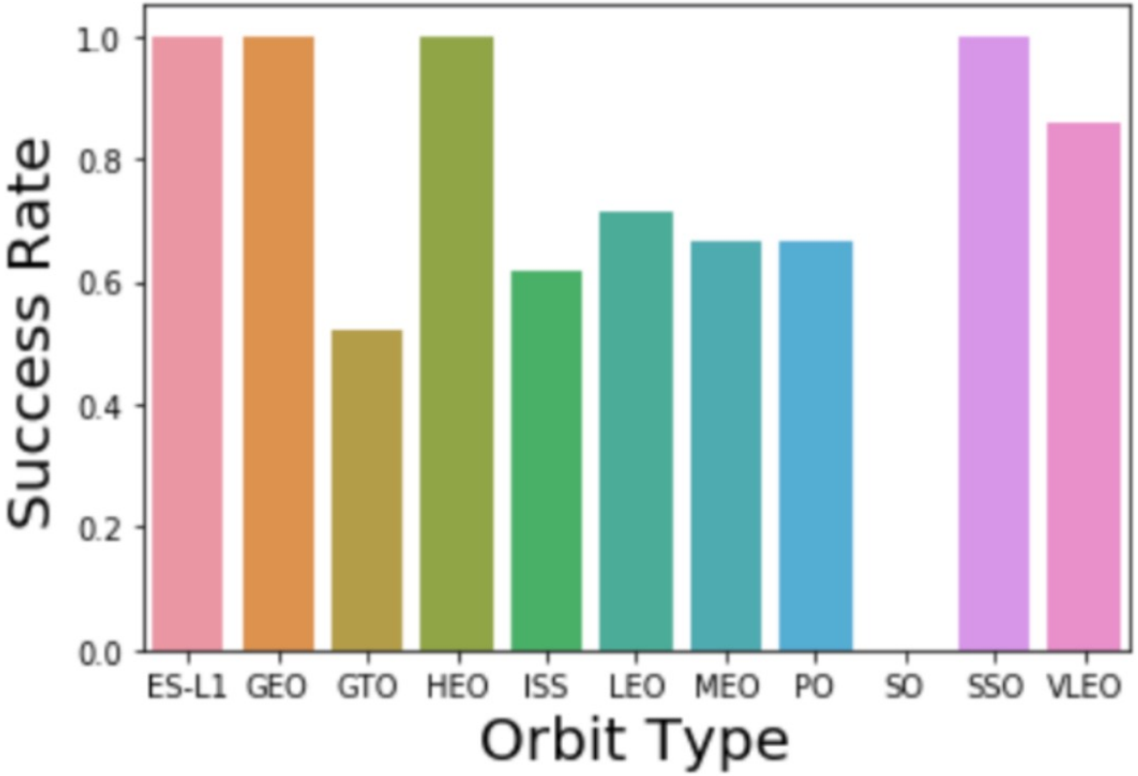
Success rate vs. Orbit type

ES L1, GEO, HEO and SSO have 100% success rate.

SO has 0% success rate.

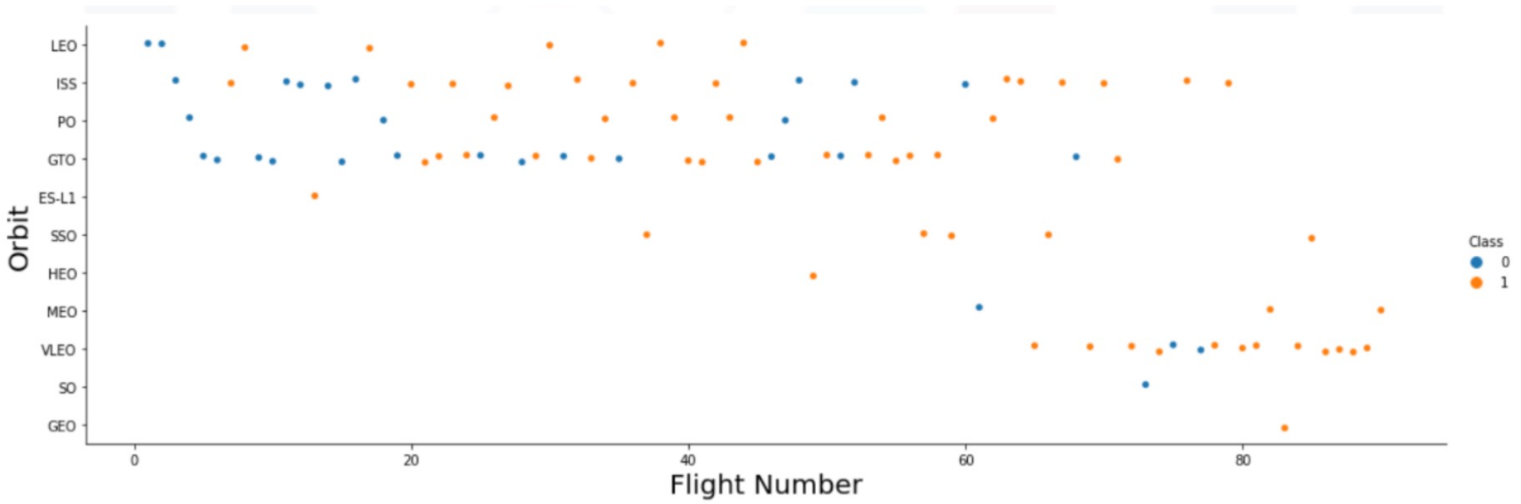
VLEO is above 80% success.

The remaining orbit types had success rates ranging from 50% to 80%.



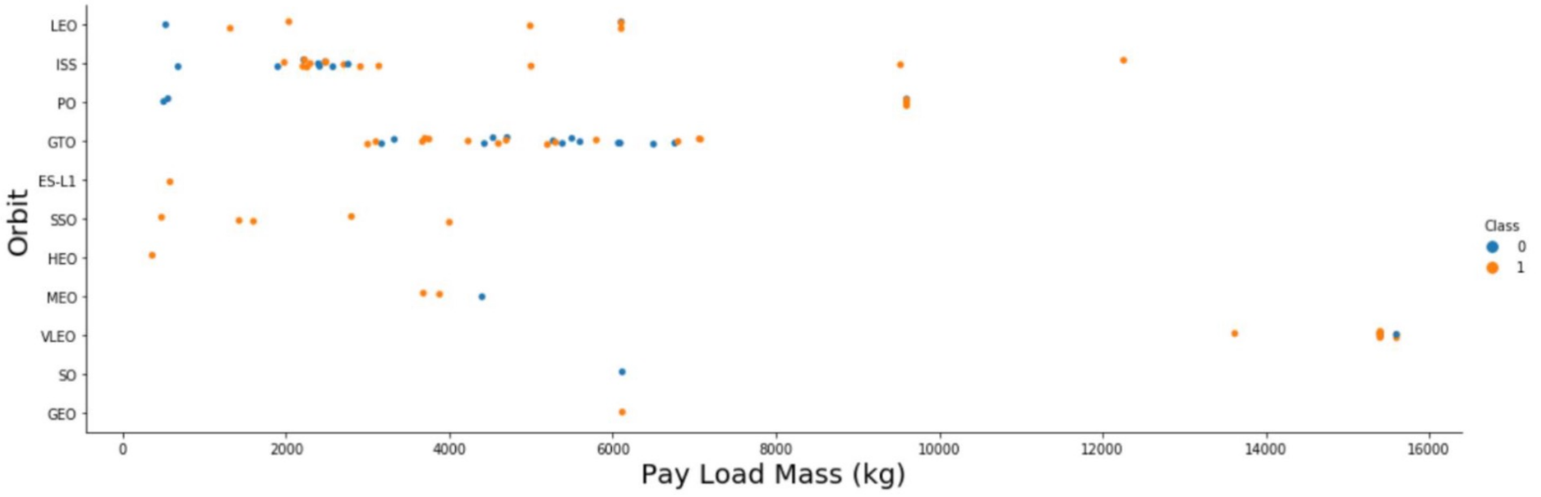
Flight Number vs. Orbit type

- LEOs success is related to the number of flights.
- The wasn't a visible relationship between GTO and flight number.
- Launches to VLEO were made later.



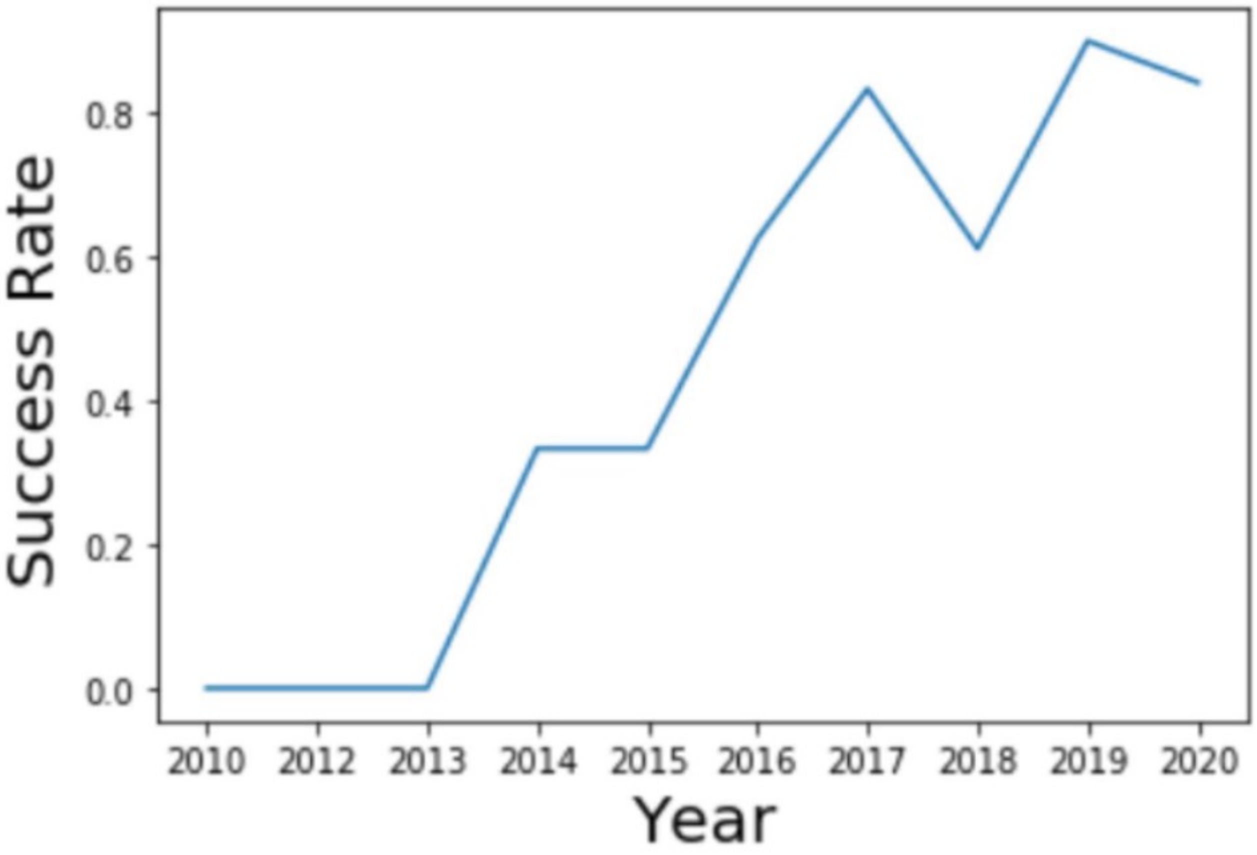
Payload vs. Orbit type

- Heavy payloads have negative influence on GTO orbits.
- Heavy payloads have positive influence on LEO and ISS orbits.



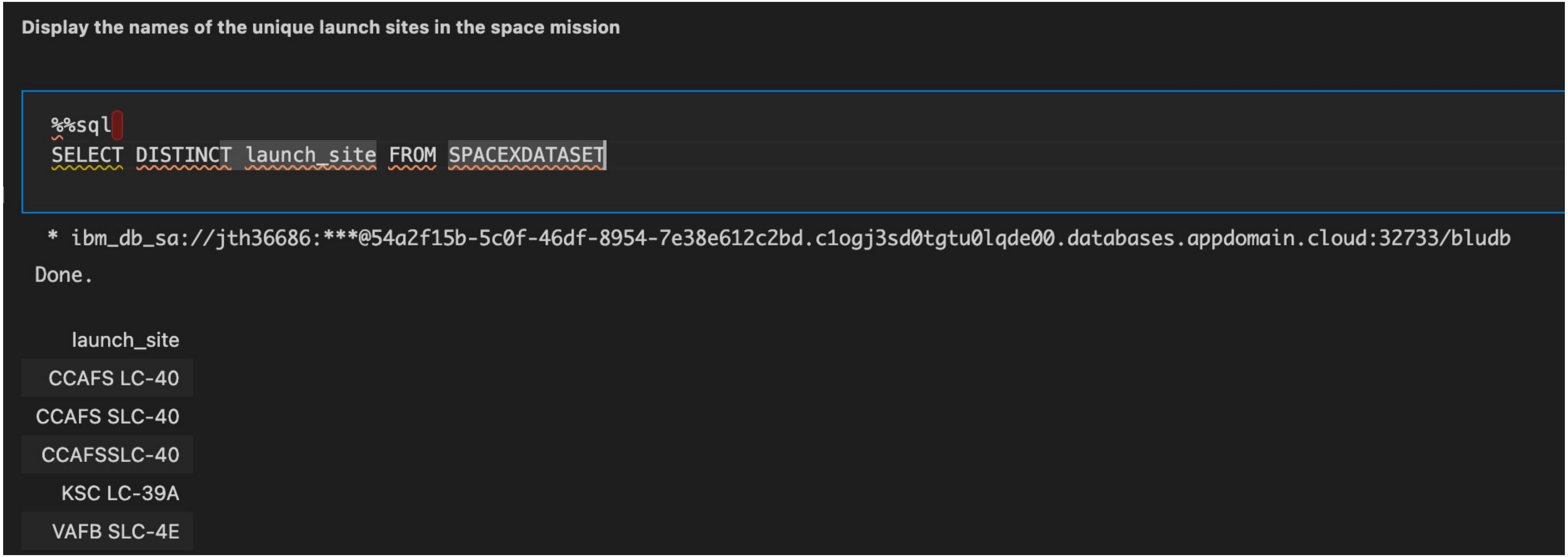
Launch success yearly trend

Success rates kept increasing from 2013 to 2020 with minor drops in 2017 and 2019.



EDA with SQL

All launch site names



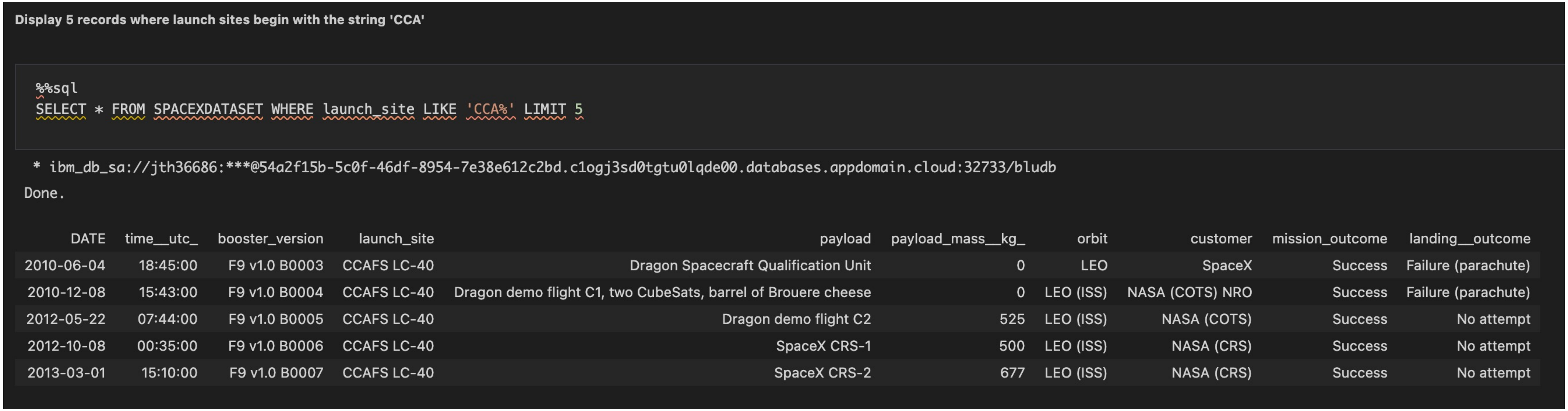
Display the names of the unique launch sites in the space mission

```
%%sql
SELECT DISTINCT launch_site FROM SPACEXDATASET
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Launch site names begin with `CCA`



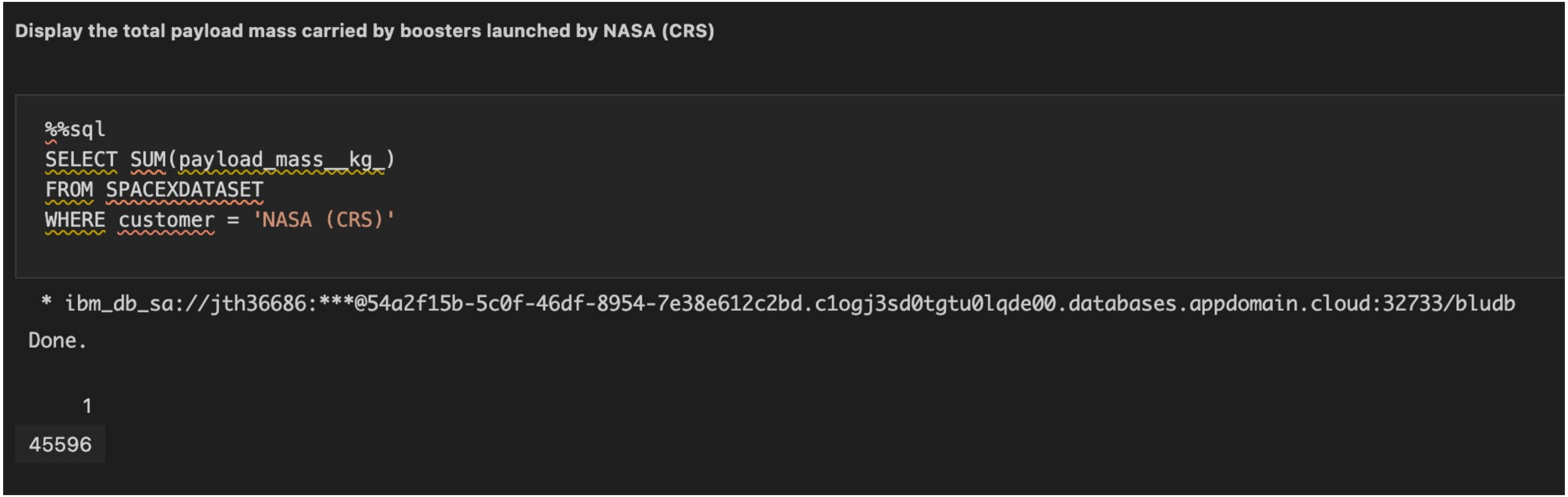
Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
SELECT * FROM SPACEXDATASET WHERE launch_site LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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



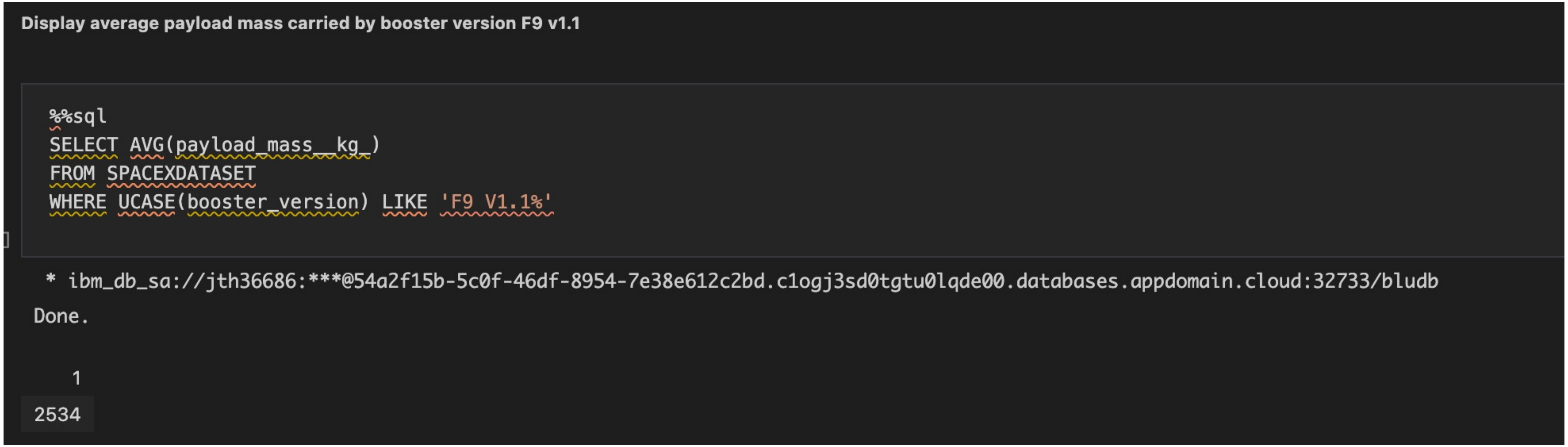
Display the total payload mass carried by boosters launched by NASA (CRS)

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

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

    1
45596
```

Average payload mass by F9 v1.1



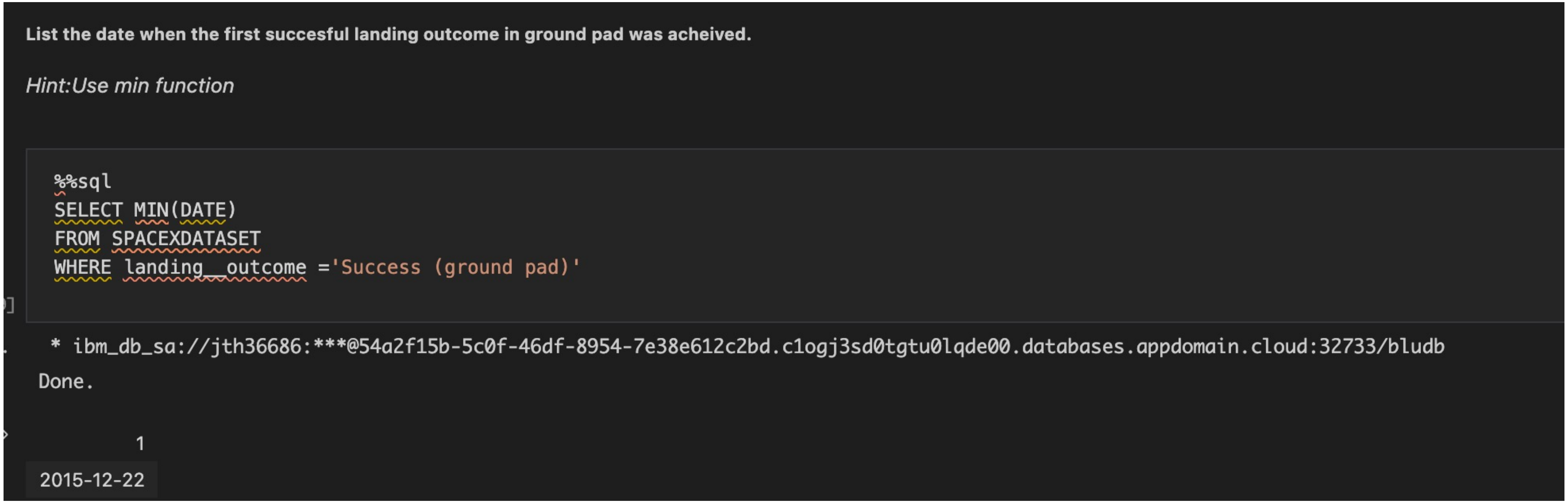
Display average payload mass carried by booster version F9 v1.1

```
%%sql
SELECT AVG(payload_mass__kg_)
FROM SPACEXDATASET
WHERE UCASE(booster_version) LIKE 'F9 V1.1%'
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

   1
2534
```

First successful ground landing date



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

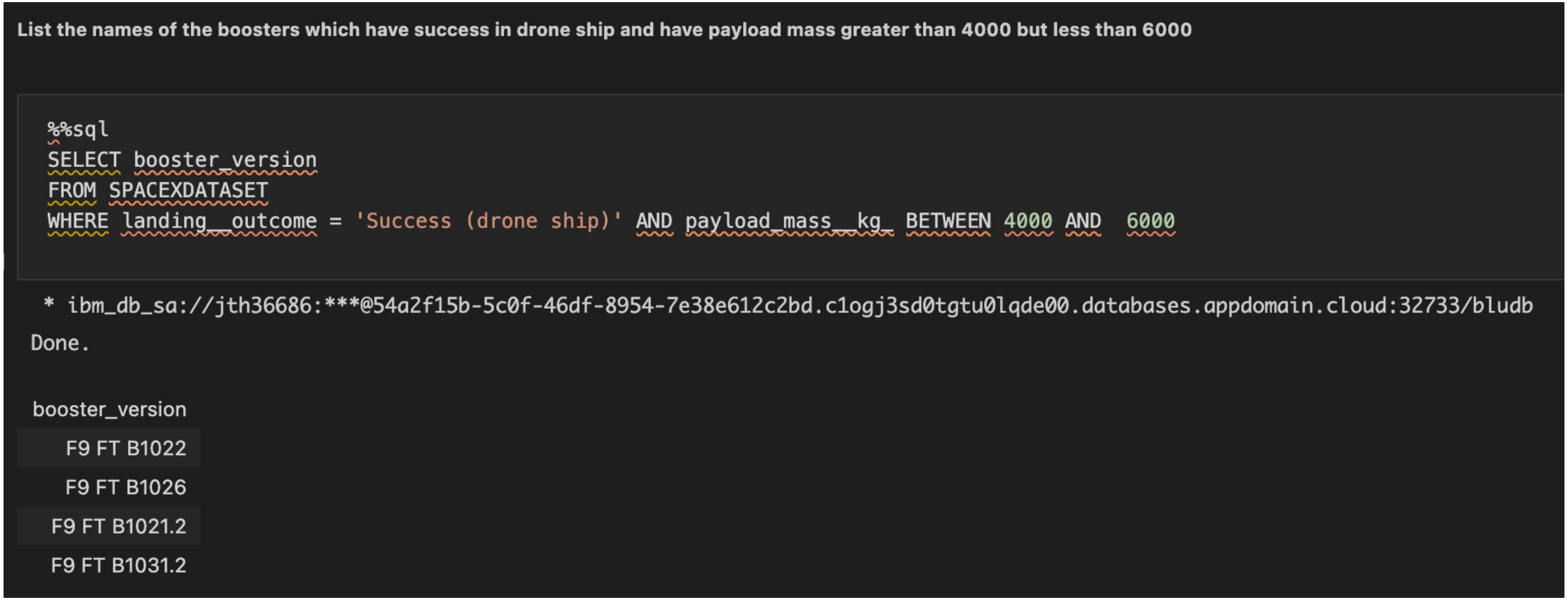
Hint:Use min function

```
%%sql
SELECT MIN(DATE)
FROM SPACEXDATASET
WHERE landing__outcome ='Success (ground pad)'
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000



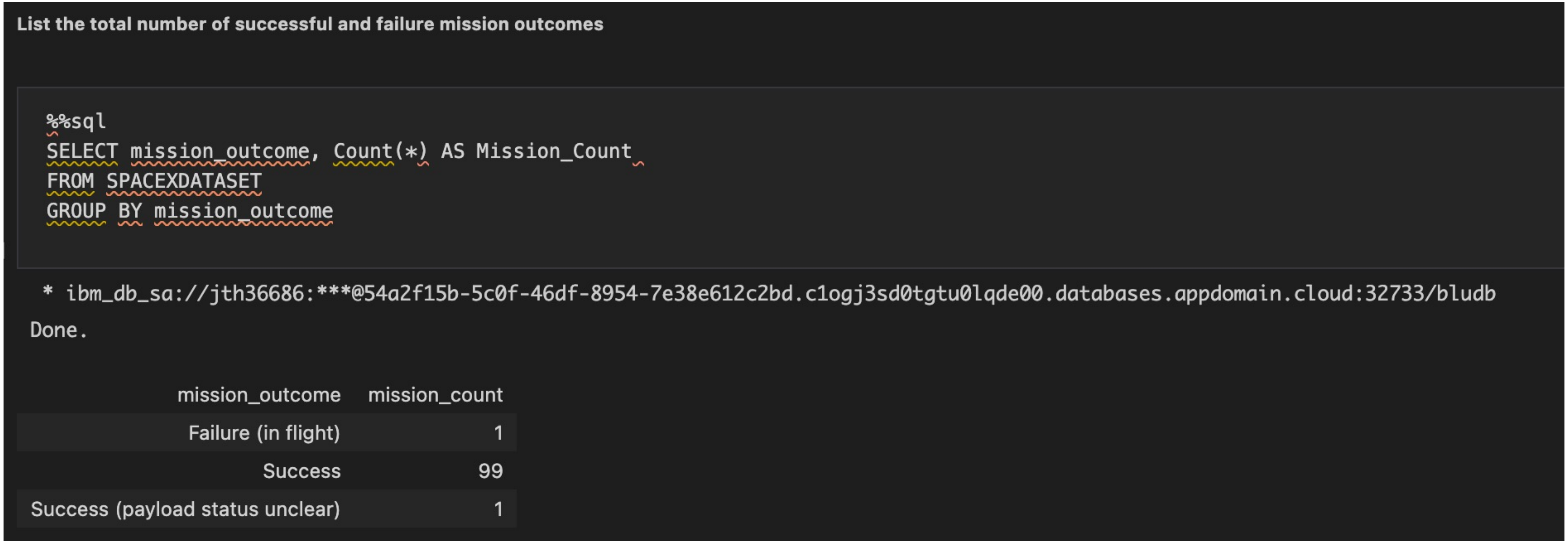
List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4000 AND  6000
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Total number of successful and failure mission outcomes



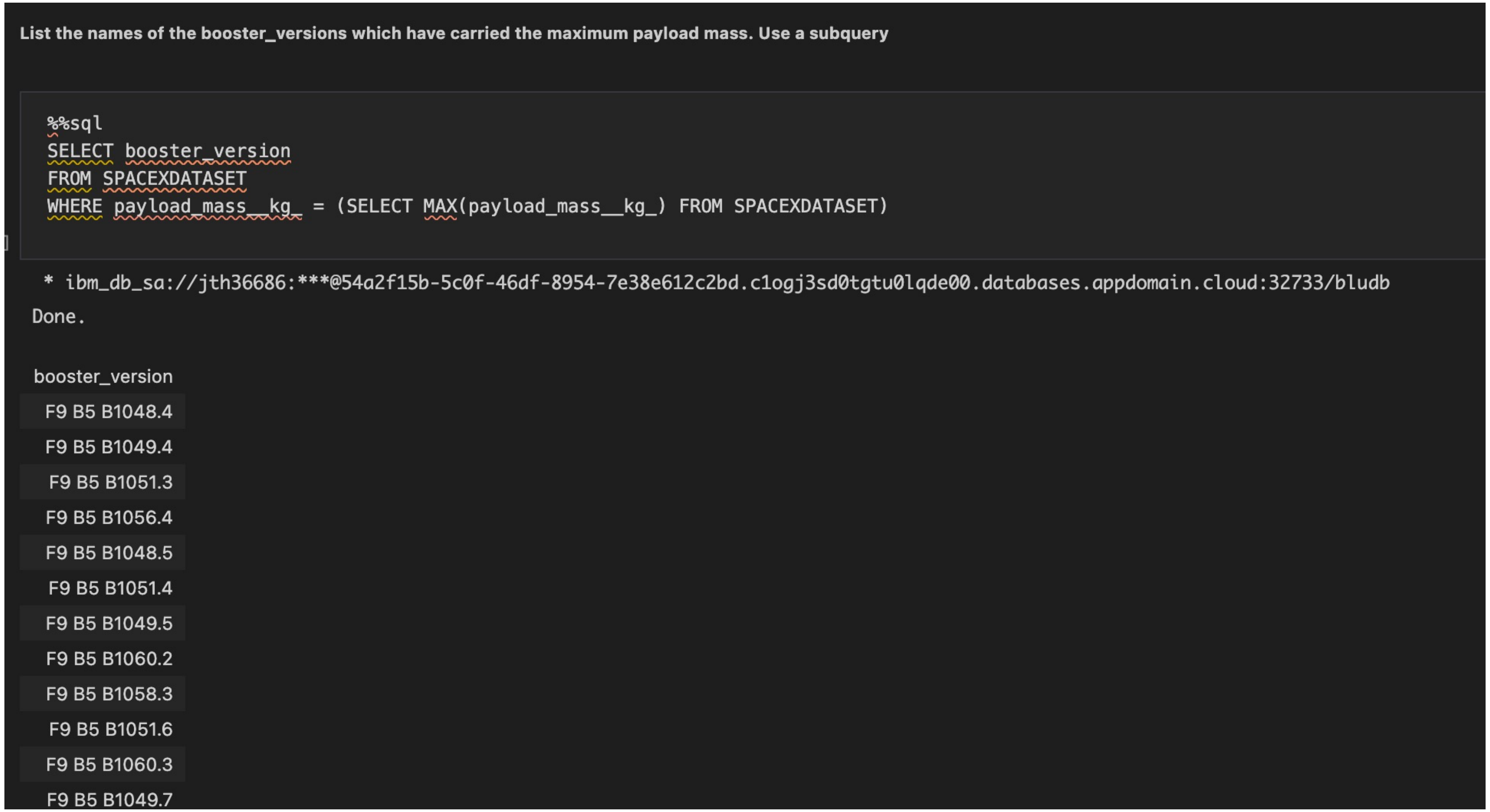
List the total number of successful and failure mission outcomes

```
%%sql
SELECT mission_outcome, Count(*) AS Mission_Count
FROM SPACEXDATASET
GROUP BY mission_outcome
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

Boosters carried maximum **payload**



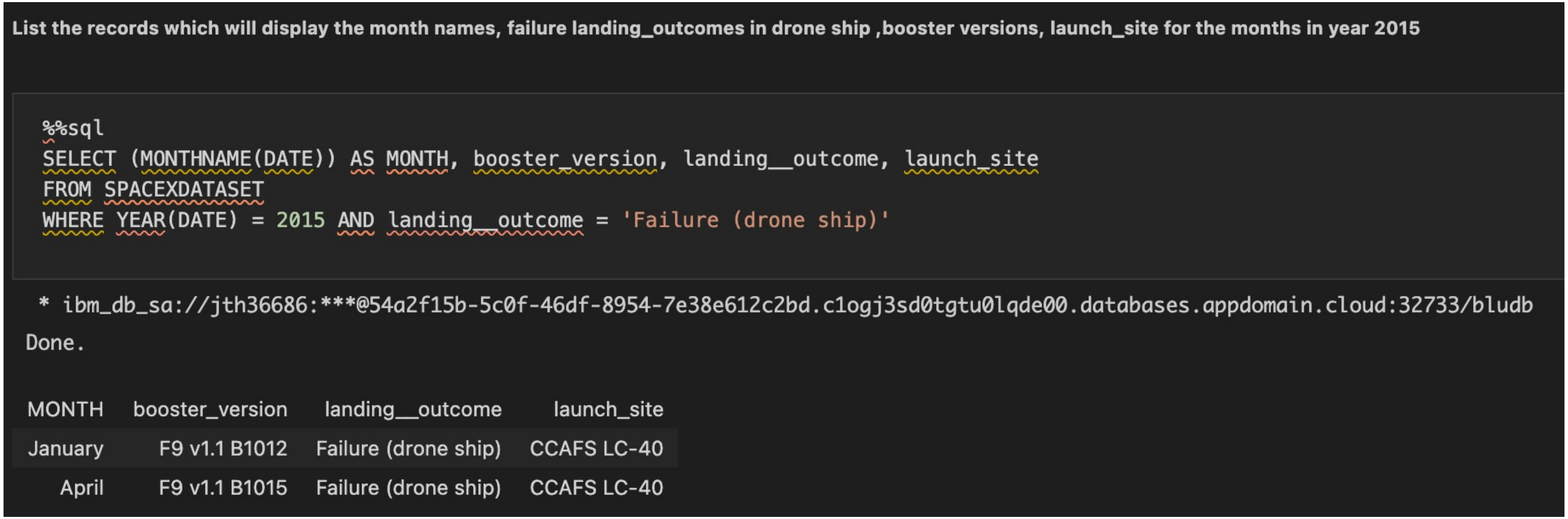
List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXDATASET)
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

2015 launch records



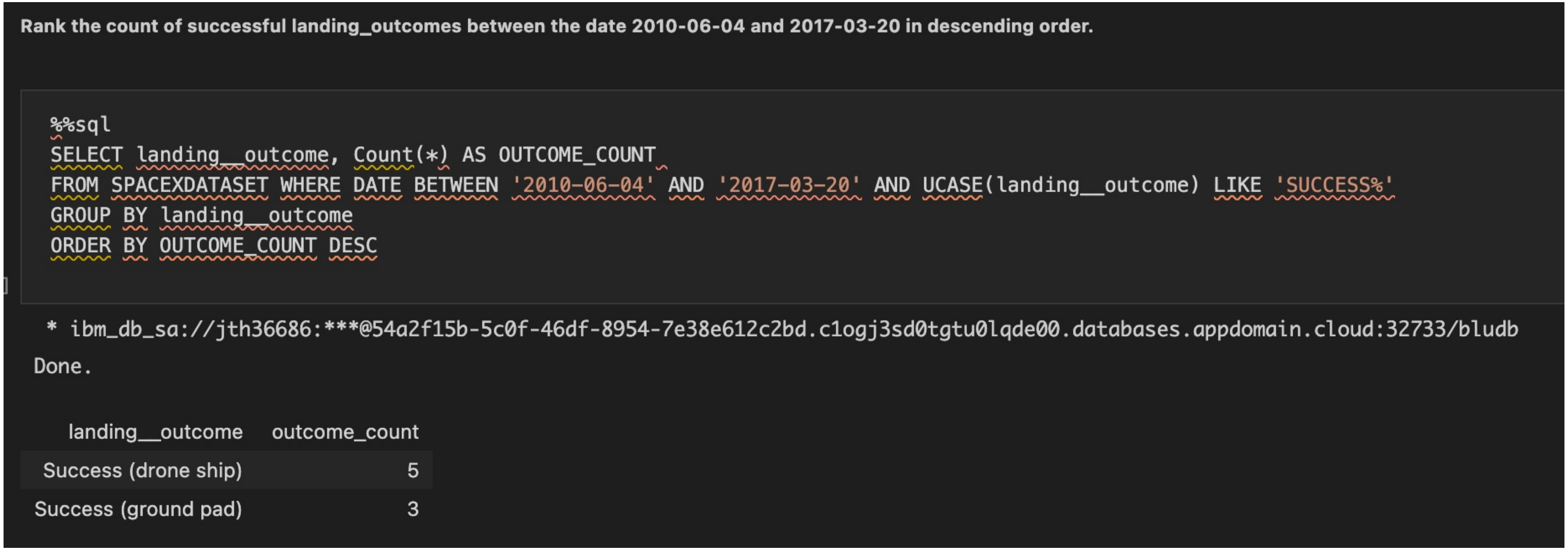
List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

```
%%sql
SELECT (MONTHNAME(DATE)) AS MONTH, booster_version, landing__outcome, launch_site
FROM SPACEXDATASET
WHERE YEAR(DATE) = 2015 AND landing__outcome = 'Failure (drone ship)'
```

```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

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

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



Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

```
%%sql
SELECT landing__outcome, Count(*) AS OUTCOME_COUNT
FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND UCASE(landing__outcome) LIKE 'SUCCESS%'
GROUP BY landing__outcome
ORDER BY OUTCOME_COUNT DESC
```

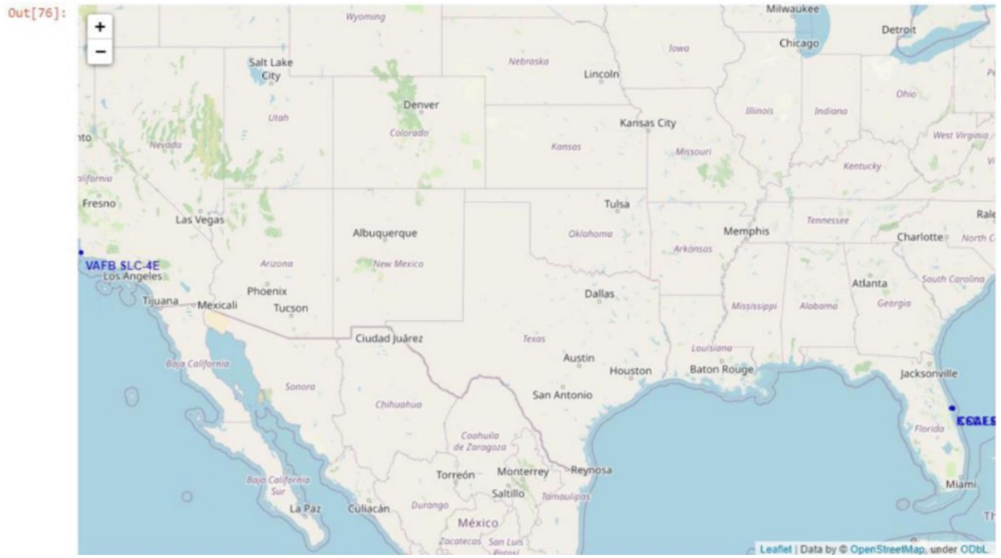
```
 * ibm_db_sa://jth36686:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.

    landing__outcome   outcome_count
 Success (drone ship)              5
Success (ground pad)              3
```

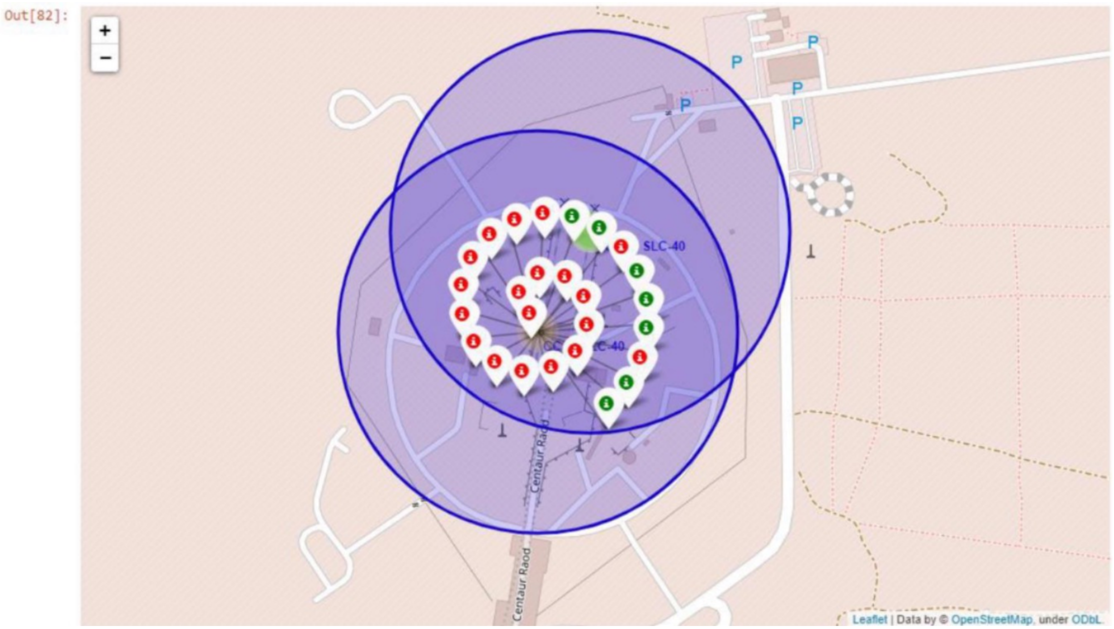
Interactive map with Folium

Launch Sites map

- Most launch sites are in close proximity to the equator where earth moves faster (±1670 km/hr)
- All sites are in close proximity to the coast, which can minimize rists of explosions or debris falling near people or buildings.

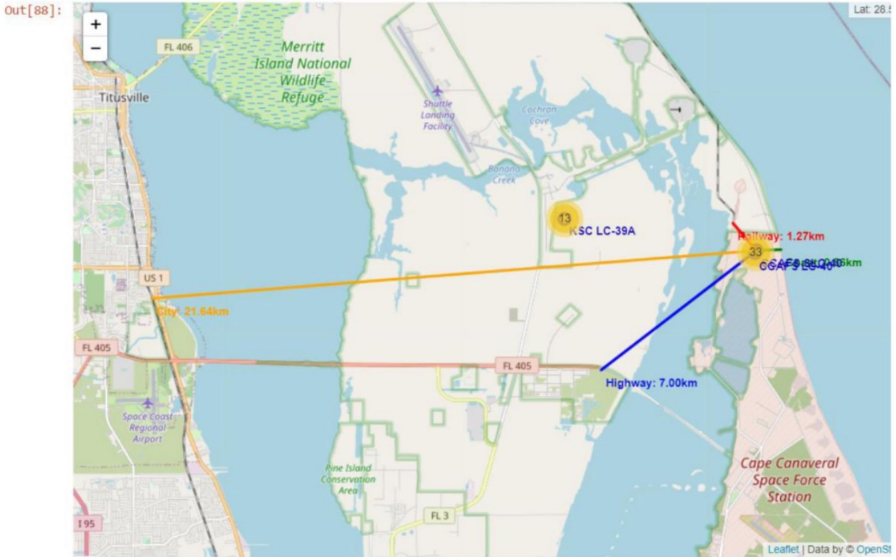


Success rate for CCAFS LC-40



This launch site did not have a very high success rate as the map shows.

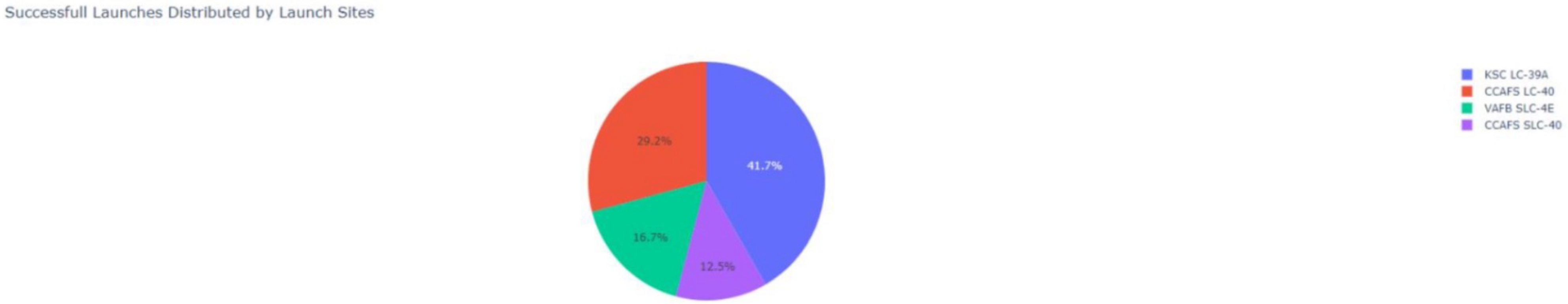
Locations of CCAFS LC-40



- It is in close proximity to water, coast, which minimizes risk of destruction.
- It is close to railway, to enable cheaper transport of material.
- It is far from city centers to reduce risk of destructions.

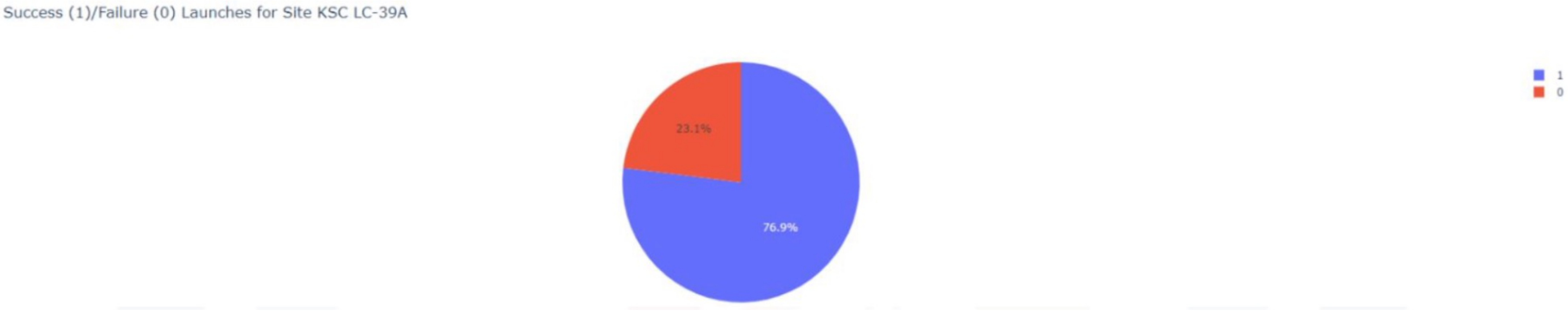
Build a Dashboard with Plotly Dash

Launch success rate for all sites



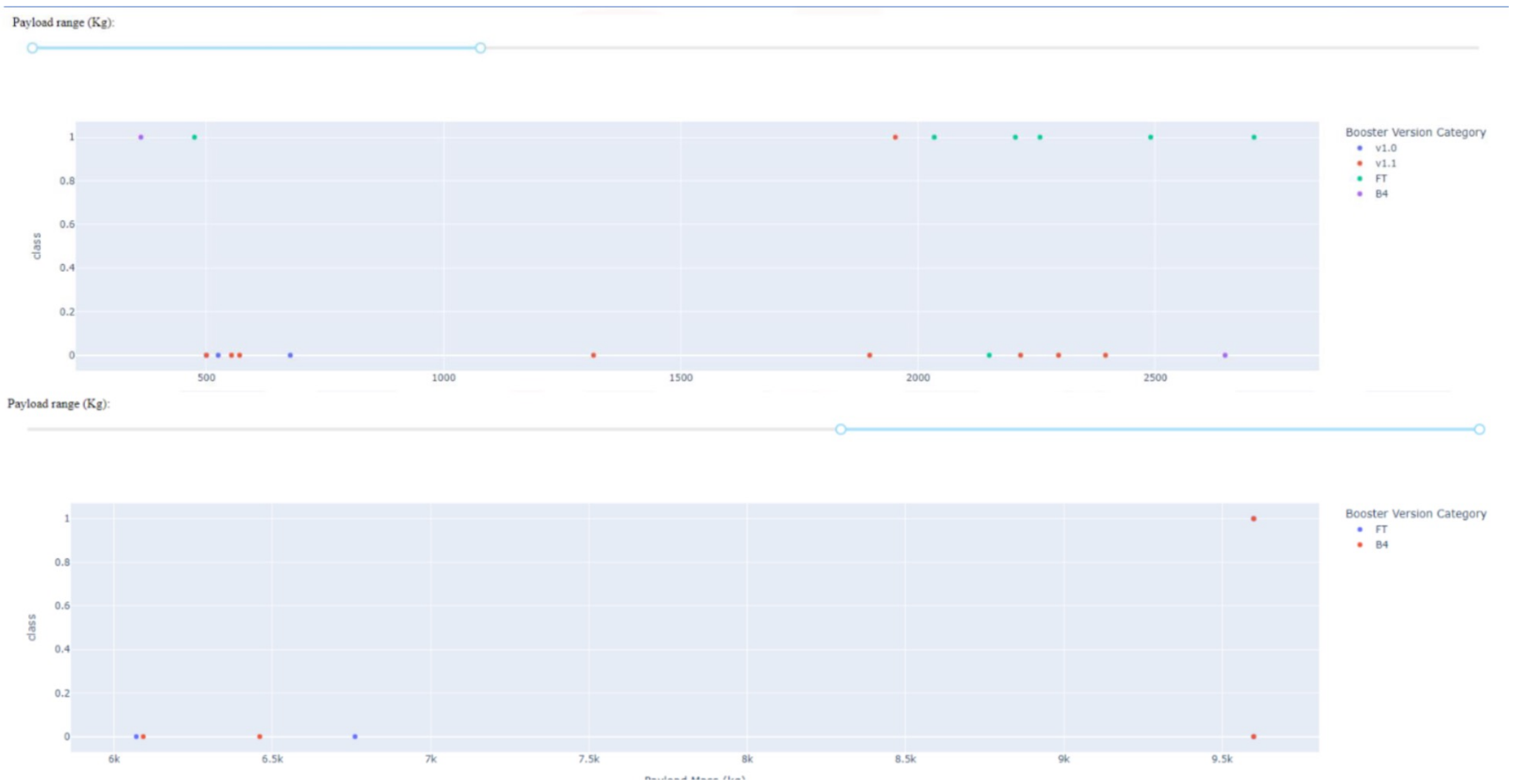
Most success was on the KSC LC-39A site, 41.7%

Success rate for KSC LC-39A



The launch site in question has the highest success rate.

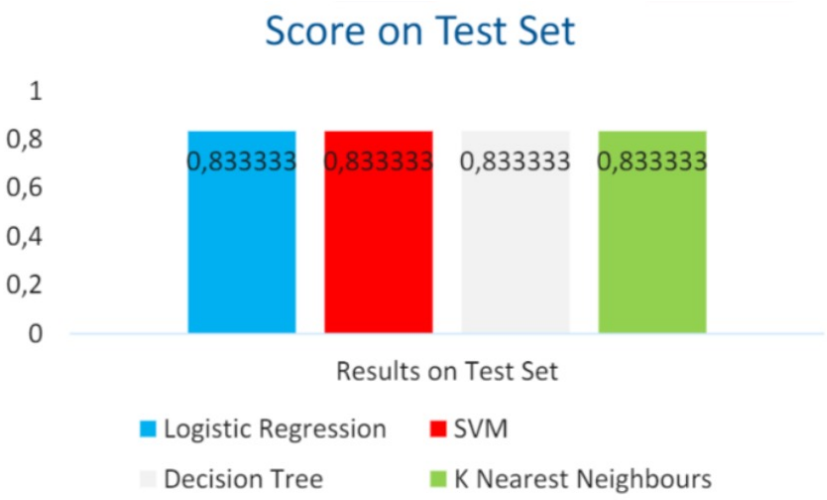
Payload vs success rate.

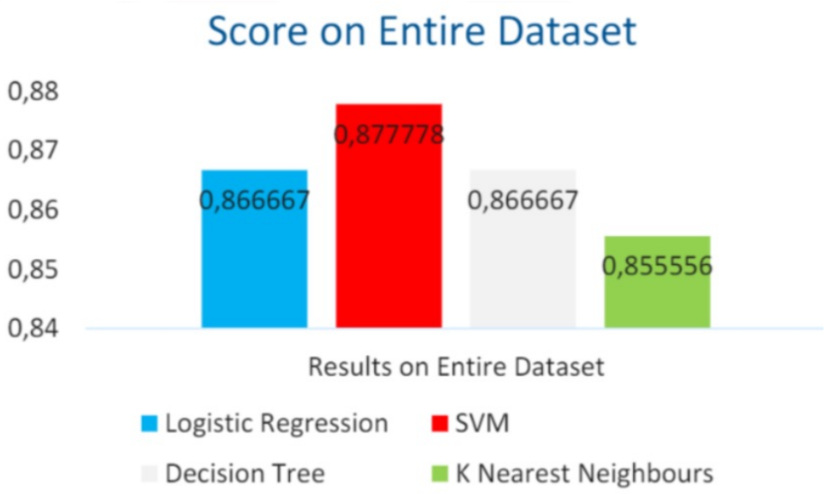


Predictive analysis (Classification)

Classification Accuracy

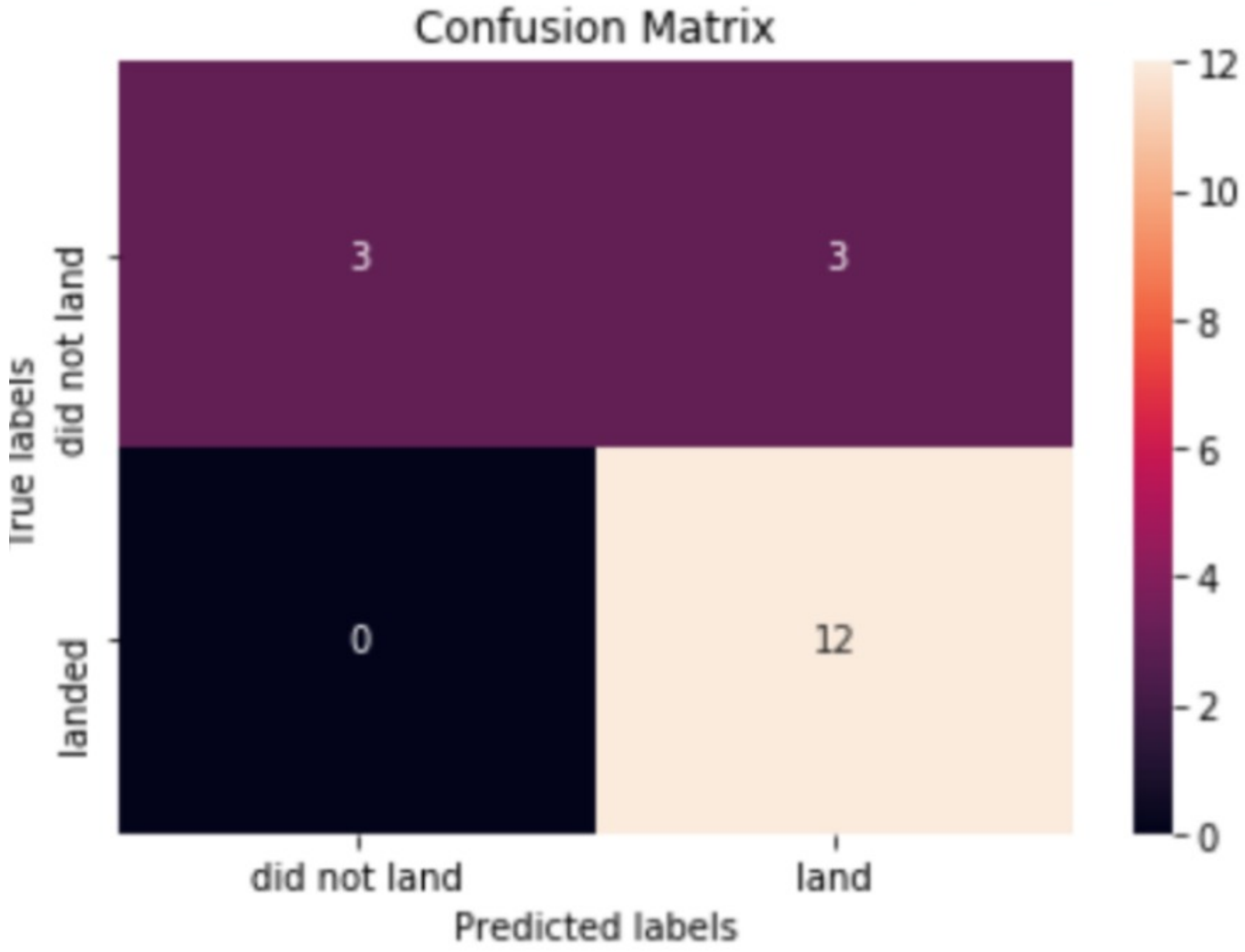
- All models showed an accuracy rate of over 80%.
- All models showed an accuracy of 83% on the test set.
- SVM performed best on the entire set, at 88%.


Score on Test Set


Score on Entire Dataset

Confusion Matrix

- The major problem to note is high false positivity rate.
- This is the point of improvement.



CONCLUSION

- Gathered datasets provided a good data predicting if Falcon9 first stage with land successfully.
- Various EDA techniques showed a list of factors affect the success of the first stage landing.
- SVM method provided the highest accuracy on the dataset.
- False positives is the point of improvement of the confusion matric analysis.