IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Silva Fernandes
16/08/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Analysis
- Predictive Analysis (Classification)
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection API
  - Data Collection with Web Scrapping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - EDA with Visualization
  - Interactive Visual Analytics with Folium
  - Building a Dashboard with Plotly Dash
  - Machine Learning Prediction

- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Introduction

- Project background and context

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is **SpaceX.** SpaceX's accomplishments include: Sending spacecraft to the International Space Station; Starlink, a satellite internet constellation providing satellite Internet access; Sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers
  - 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?

Section 1

Methodology

Methodology

- Data collection methodology:
  - Using SpaceX Rest API and Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering the data, dealing with missing values and using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune and evaluate classification models to ensure the best results

Data Collection

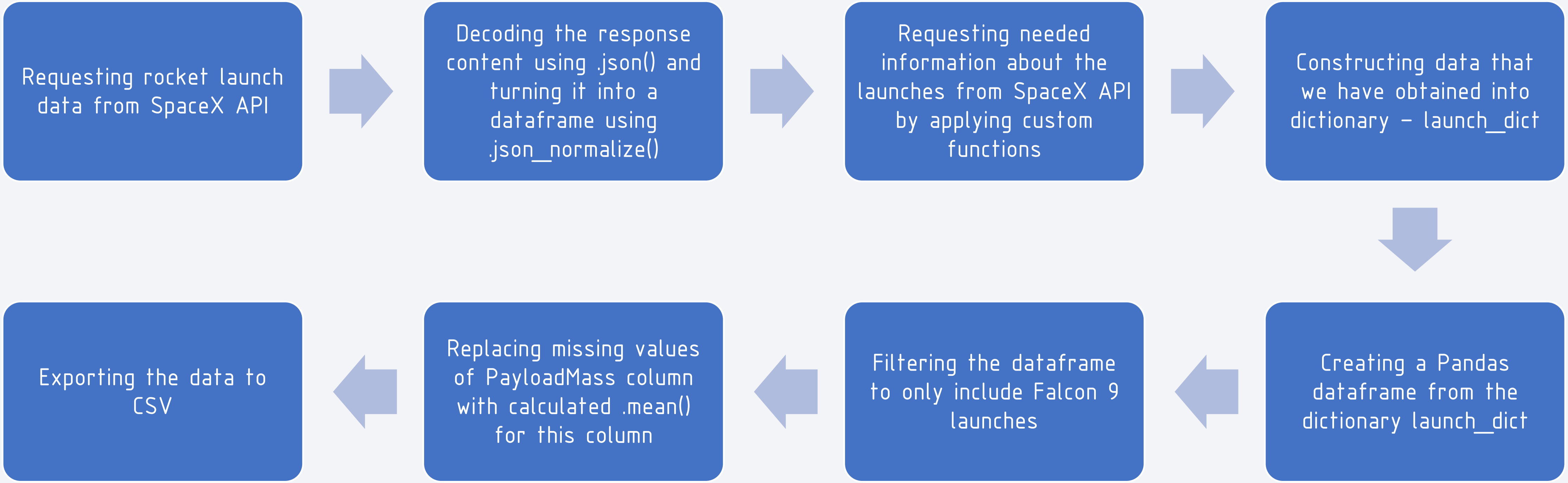
Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

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

Data Collection – SpaceX API



Requesting rocket launch data from SpaceX API

Decoding the response content using .json() and turning it into a dataframe using .json_normalize()

Requesting needed information about the launches from SpaceX API by applying custom functions

Constructing data that we have obtained into dictionary – launch_dict

Creating a Pandas dataframe from the dictionary launch_dict

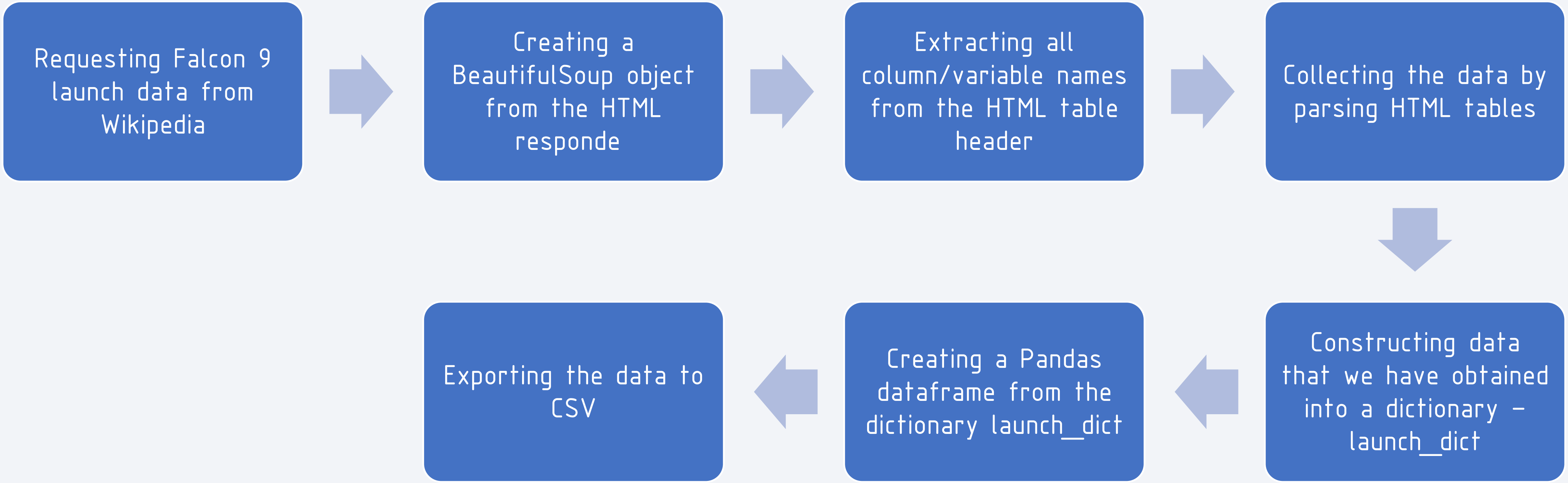
Filtering the dataframe to only include Falcon 9 launches

Replacing missing values of PayloadMass column with calculated .mean() for this column

Exporting the data to CSV

Data Collection API (GitHub URL)

Data Collection – Web Scraping



Requesting Falcon 9 launch data from Wikipedia

Creating a BeautifulSoup object from the HTML responde

Extracting all column/variable names from the HTML table header

Collecting the data by parsing HTML tables

Constructing data that we have obtained into a dictionary – launch_dict

Creating a Pandas dataframe from the dictionary launch_dict

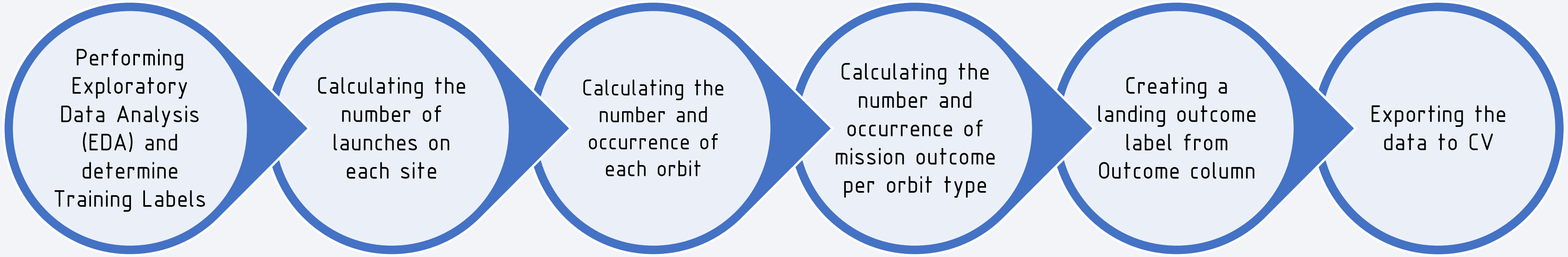
Exporting the data to CSV

Data Collection with Web Scraping (GitHub URL)

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, *True Ocean* means the mission outcome was successfully landed to a specific region of the ocean while *False Ocean* means the mission outcome was unsuccessfully landed to a specific region of the ocean. *True RTLS* means the mission outcome was successfully landed to a ground pad while *False RTLS* means the mission outcome was unsuccessfully landed to a ground pad. *True ASDS* means the mission outcome was successfully landed on a drone ship while *False ASDS* means the mission outcome was unsuccessfully landed on a drone ship.

In this chapter we mainly converted those outcomes into Training Labels with "*1*" means the booster successfully landed and "*0*" means it was unsuccessful.


Performing Exploratory Data Analysis (EDA) and determine Training Labels
Calculating the number of launches on each site
Calculating the number and occurrence of each orbit
Calculating the number and occurrence of mission outcome per orbit type
Creating a landing outcome label from Outcome column
Exporting the data to CV

Data Wrangling (GitHub URL)

EDA with SQL

- SQL queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the records which will display the failure landing outcomes in drone ship, booster versions and launch site for the months in year 2015
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

EDA with SQL (GitHub URL)

EDA with Data Visualization

- Charts that 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

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).

EDA with Visualization (GitHub URL)

Build an Interactive Map with Folium

- Mark all launch sites on a map
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts
- Mark the success/failed launches for each site on the map
  - Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates
- Calculate the distances between a launch site to its proximities
  - Added coloured Lines to show distances between the Launch Site VAFB SLC-4E (as an example) and its proximities like Railway, Highway, Coastline and Closest City

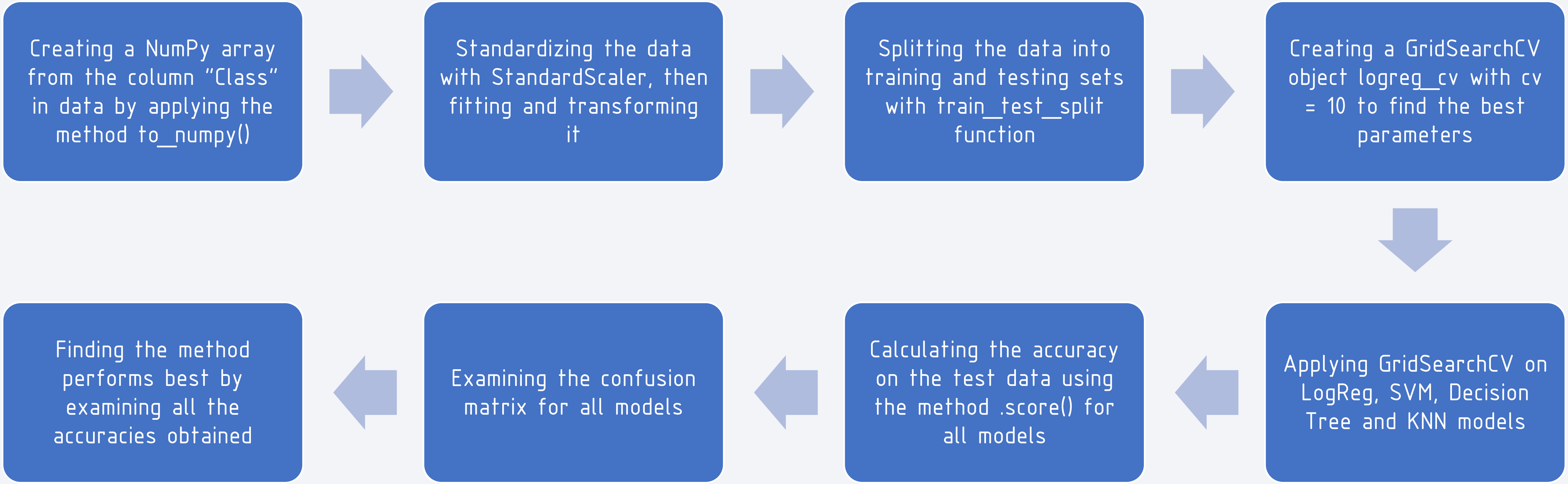
Interactive Visual Analytics with Folium (GitHub URL)

Build a Dashboard with Plotly Dash

- Launch Sites Dropdown List
  - Added a dropdown list to enable Launch Site selection
- Pie Chart showing Success Launches (All Sites/Certain Site)
  - Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected
- Slider of Payload Mass Range
  - Added a slider to select Payload range
- Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions
  - Added a scatter chart to show the correlation between Payload and Launch Success

SpaceX Dash App (GitHub URL)

Predictive Analysis (Classification)



Creating a NumPy array from the column "Class" in data by applying the method to_numpy()

Standardizing the data with StandardScaler, then fitting and transforming it

Splitting the data into training and testing sets with train_test_split function

Creating a GridSearchCV object logreg_cv with cv = 10 to find the best parameters

Applying GridSearchCV on LogReg, SVM, Decision Tree and KNN models

Calculating the accuracy on the test data using the method .score() for all models

Examining the confusion matrix for all models

Finding the method performs best by examining all the accuracies obtained

Machine Learning Prediction (GitHub URL)

Results

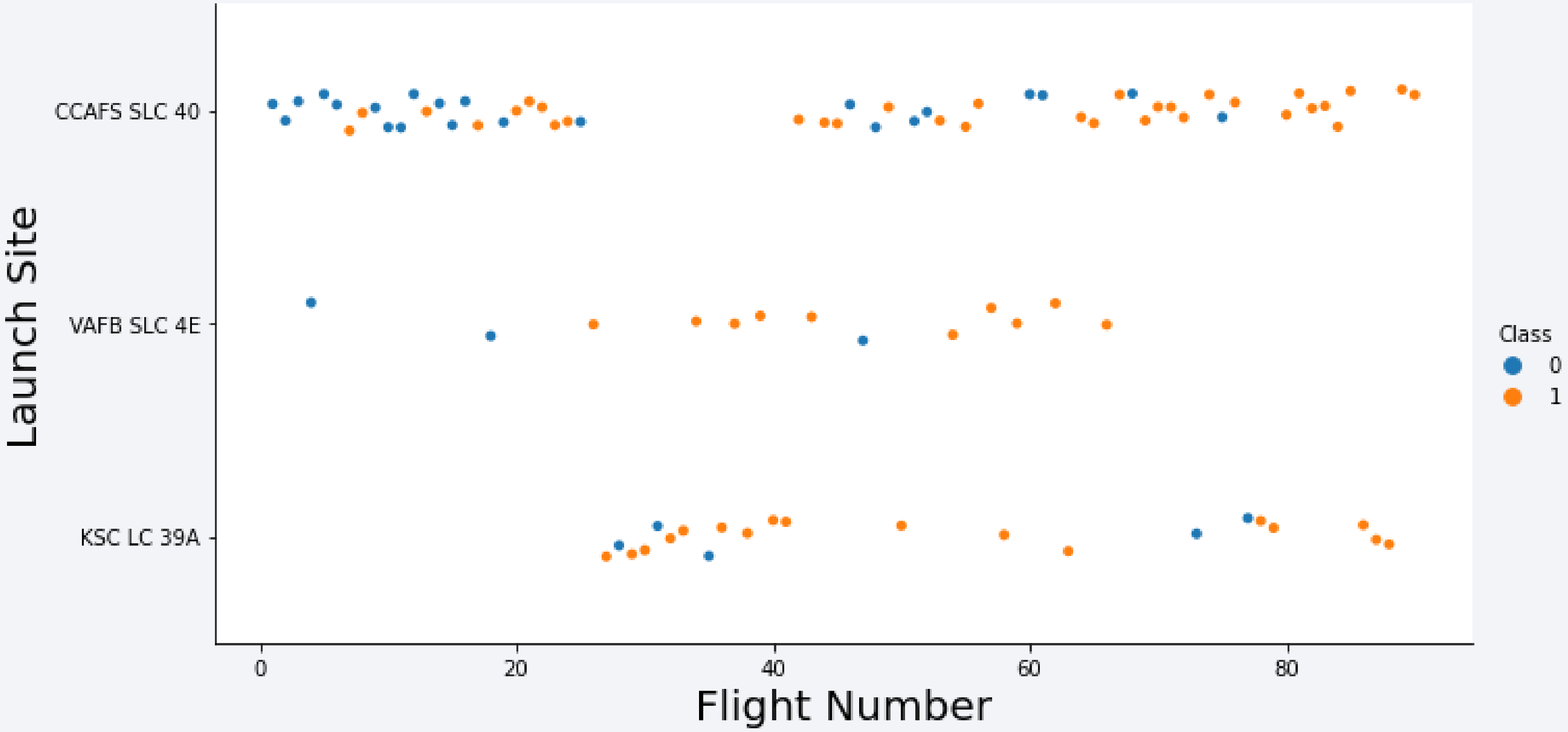
On the next slides, we will be able to see and check all the results obtained by:

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

Section 2

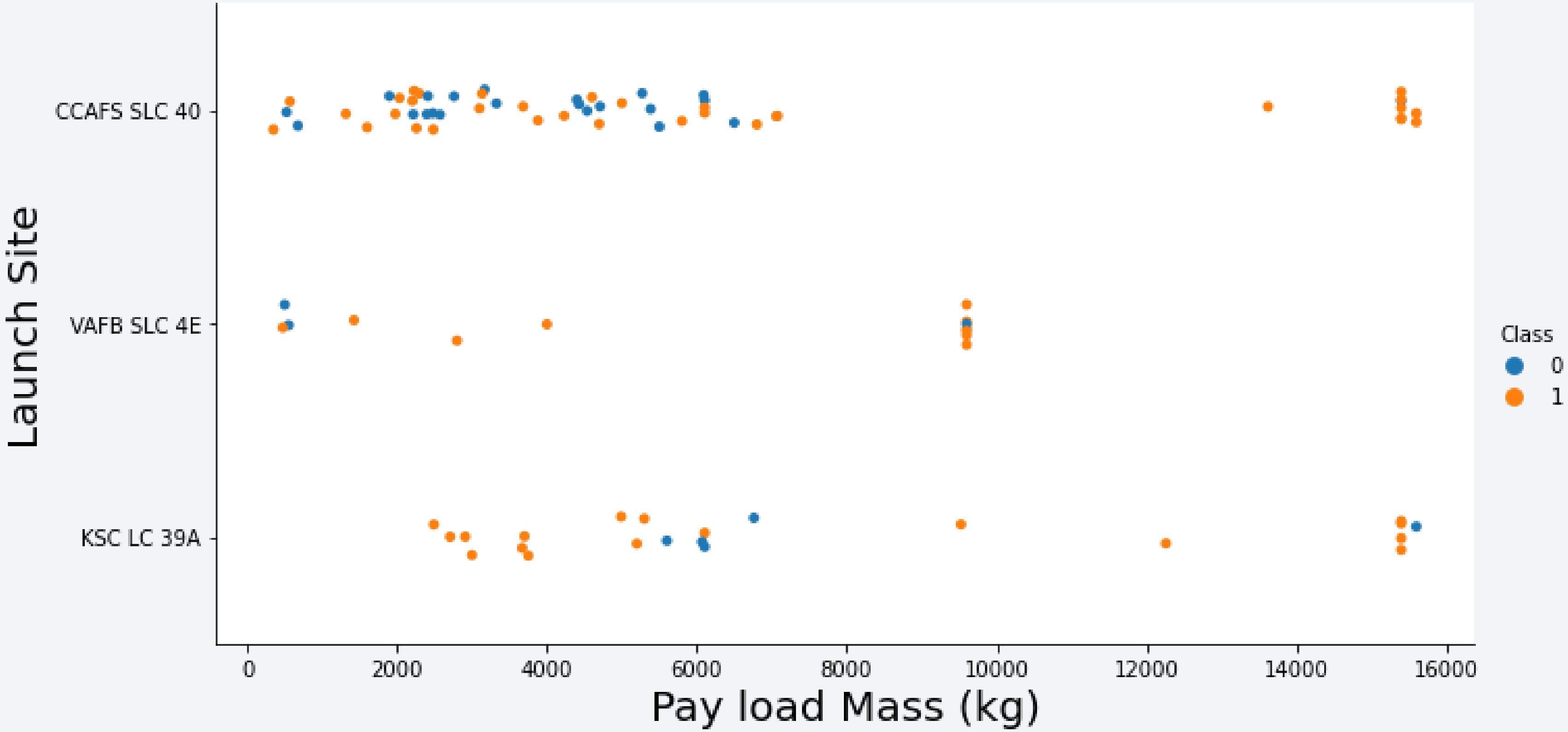
Insights drawn from EDA

Flight Number vs. Launch Site



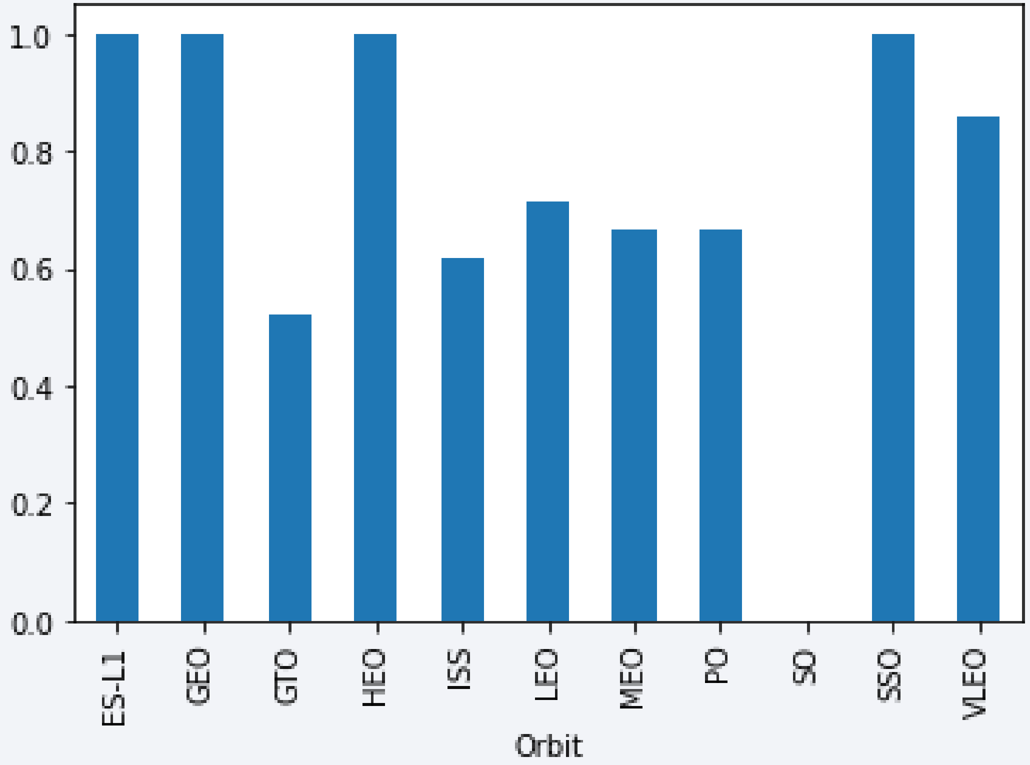
- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates than CCAFS SLC 40.
- We can see that each new launch has a higher rate of success.

Payload vs. Launch Site



- For VAFB-SLC 4E launch site there are no rockets launched for heavy payload mass (>10 000 kg)
- 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.

Success Rate vs. Orbit Type

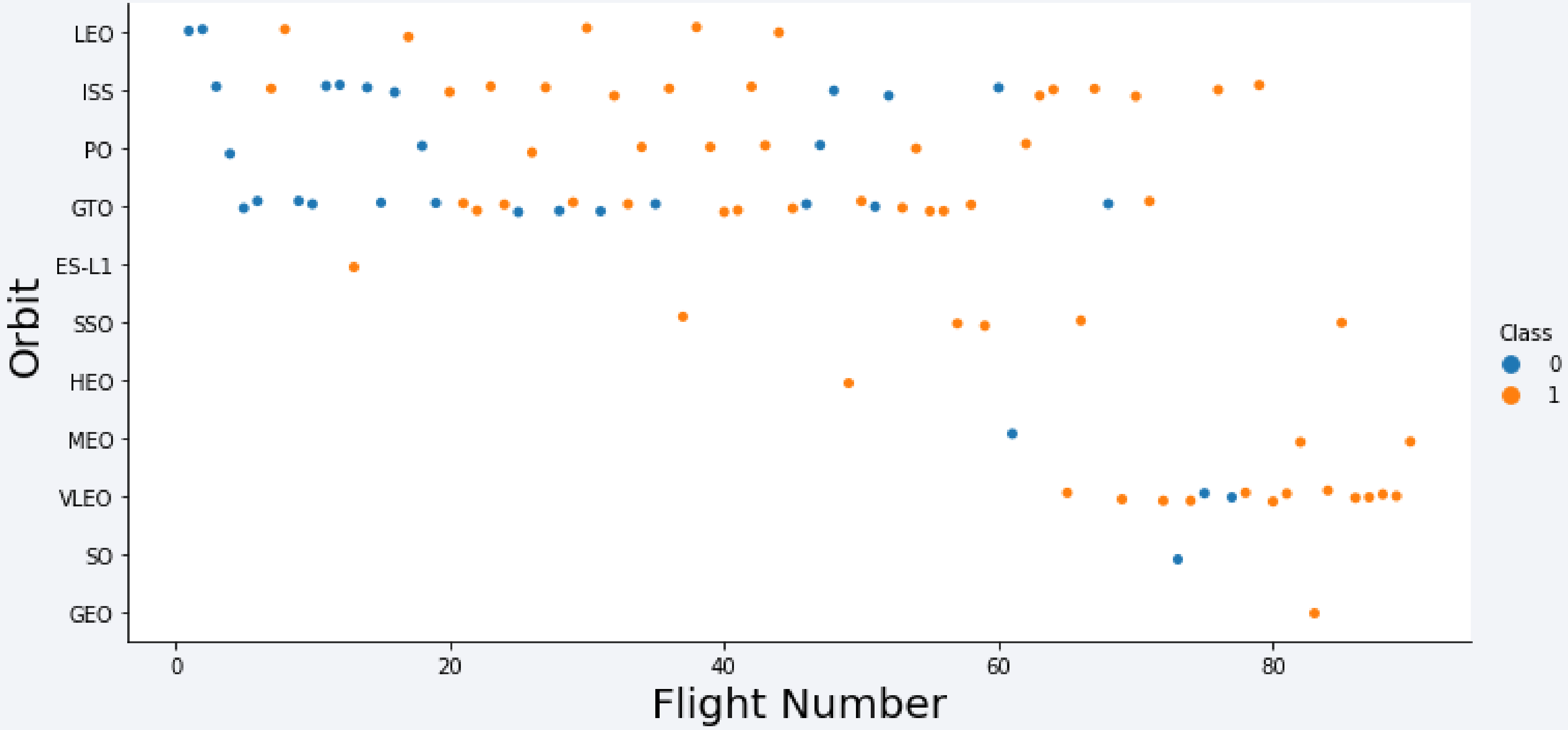


- Orbits with 100% success rate:
  - ES-L1, GEO, HEO and SSO

- Orbits with 0% success rate:
  - SO

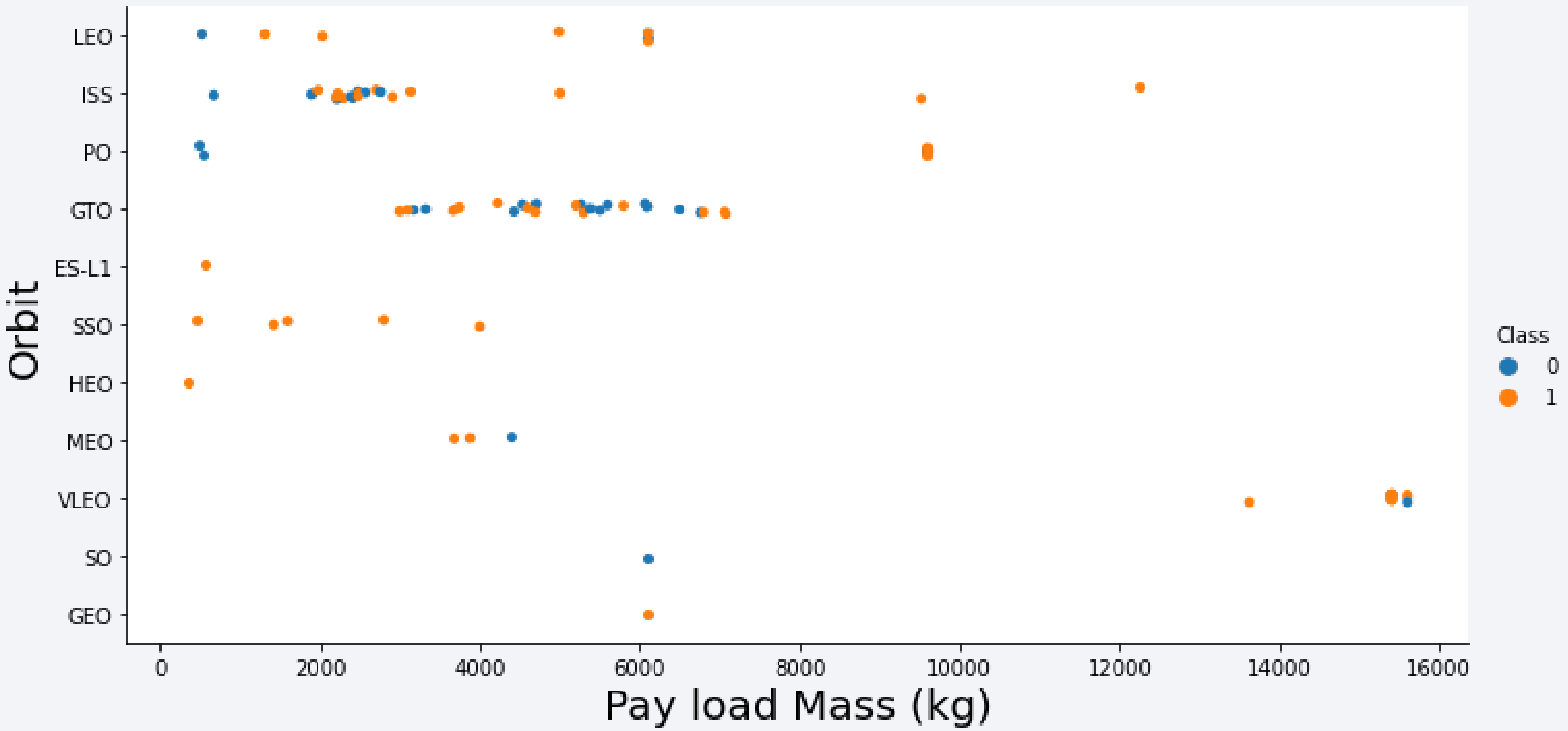
- Orbits with success rate between 50% and 85%:
  - GTO, ISS, LEO, MEO and PO

Flight Number vs. Orbit Type



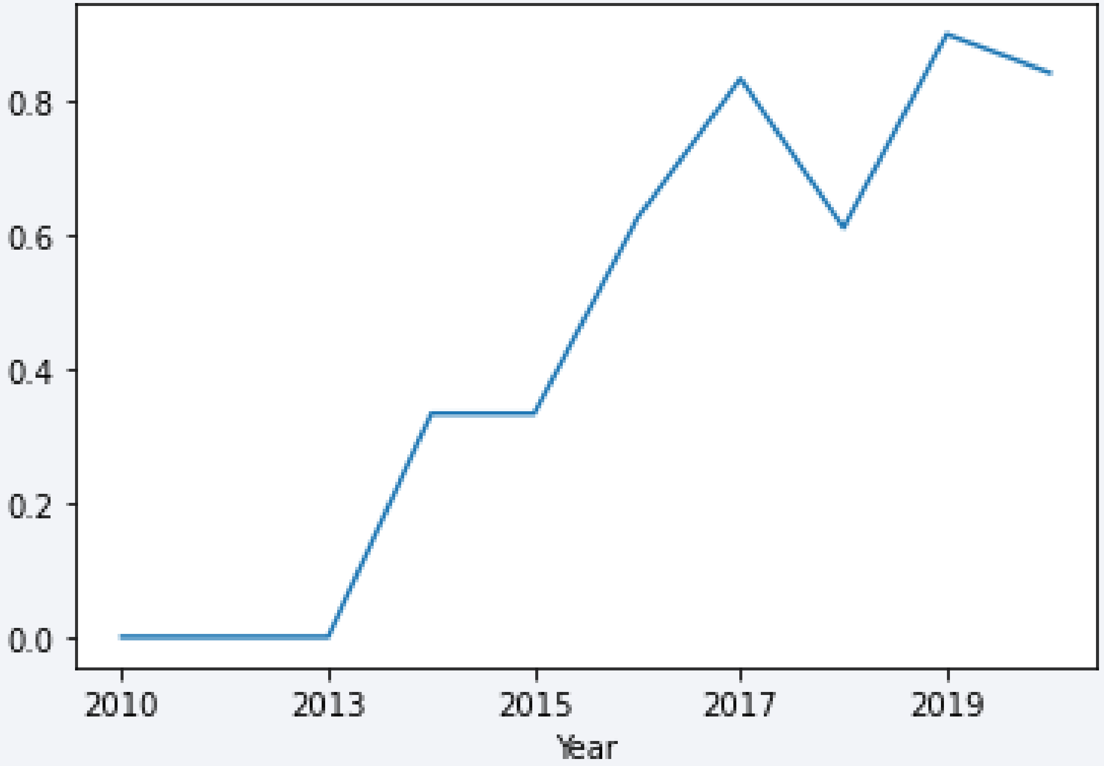
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



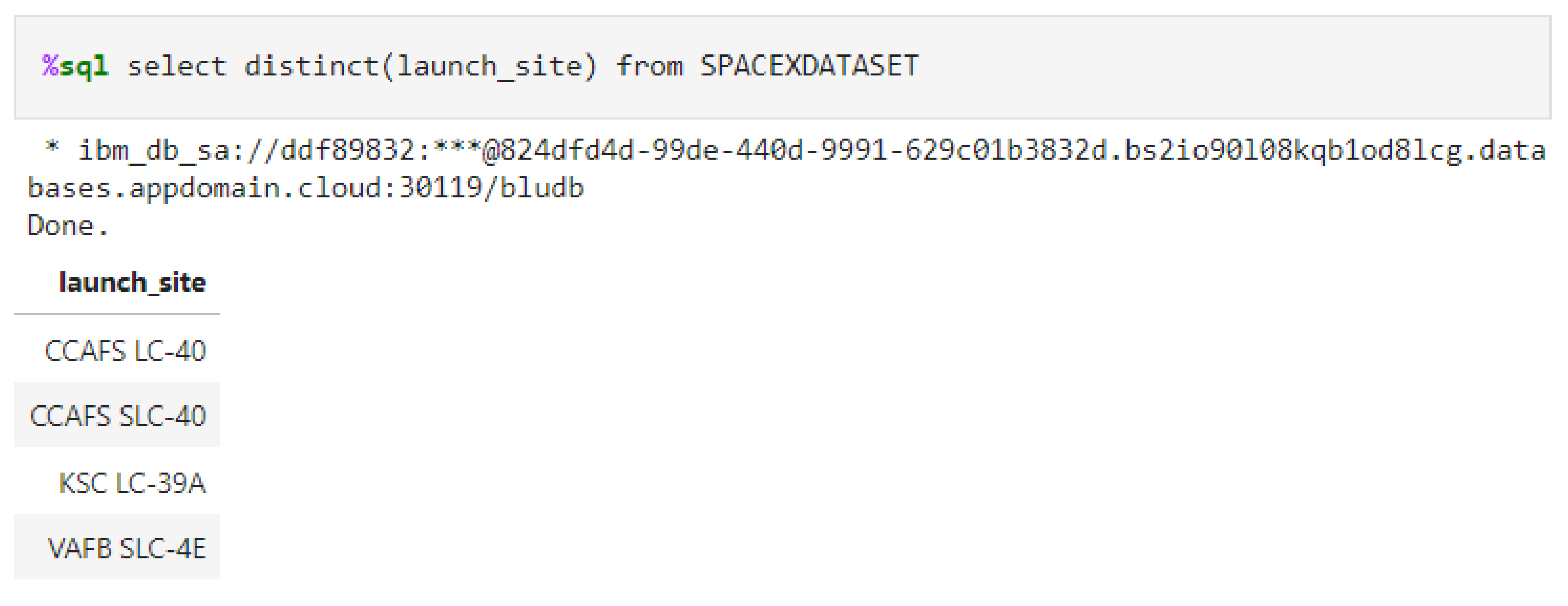
- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here.

Launch Success Yearly Trend



The success rate since 2013 kept increasing until 2020, however between 2017 and 2018, the success rate dropped about 20%.

All Launch Site Names



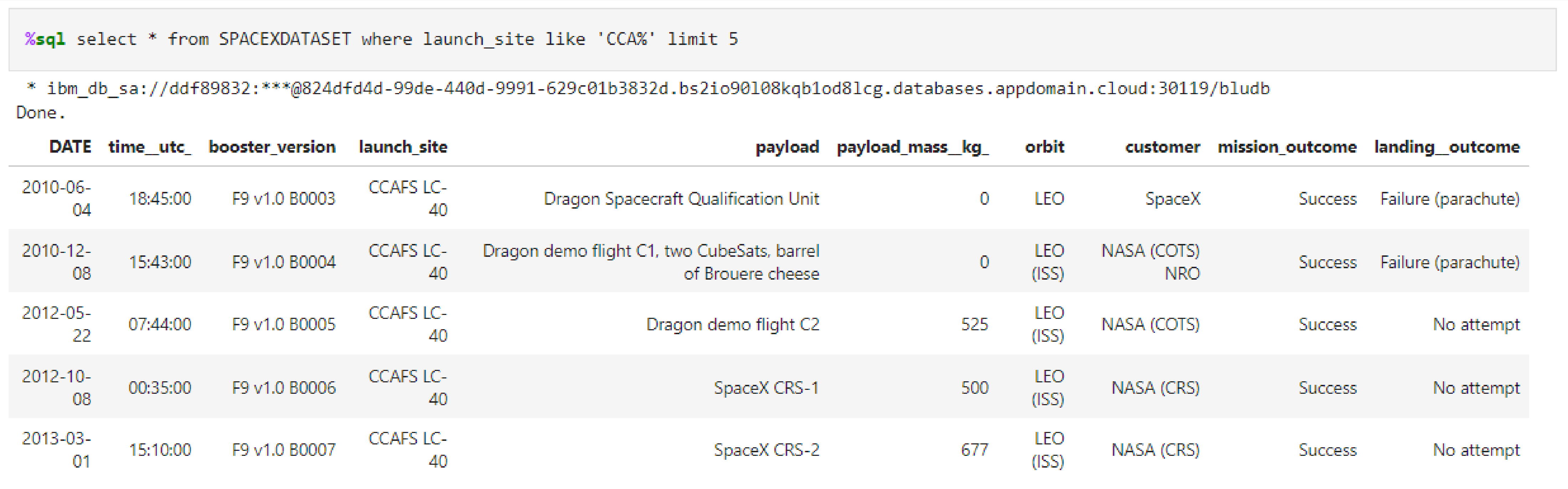
```
%sql select distinct(launch_site) from SPACEXDATASET
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.data
bases.appdomain.cloud:30119/bludb
Done.
```

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

Display the names of the unique launch sites in the space mission. The result were 4 different launch sites.

Launch Site Names Begin with 'CCA'



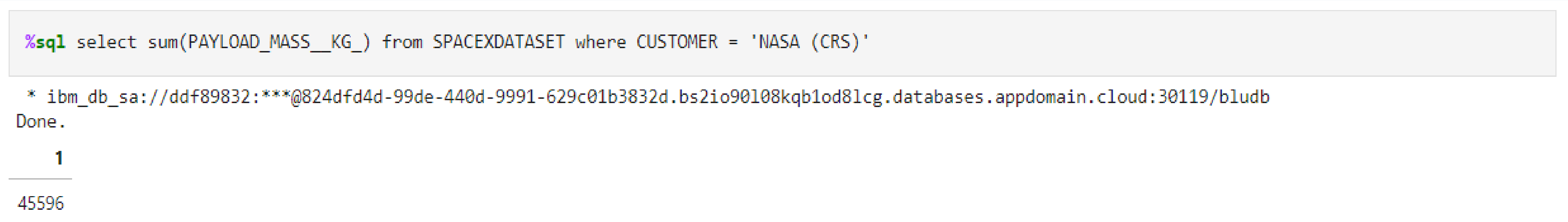
```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/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

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

Total Payload Mass

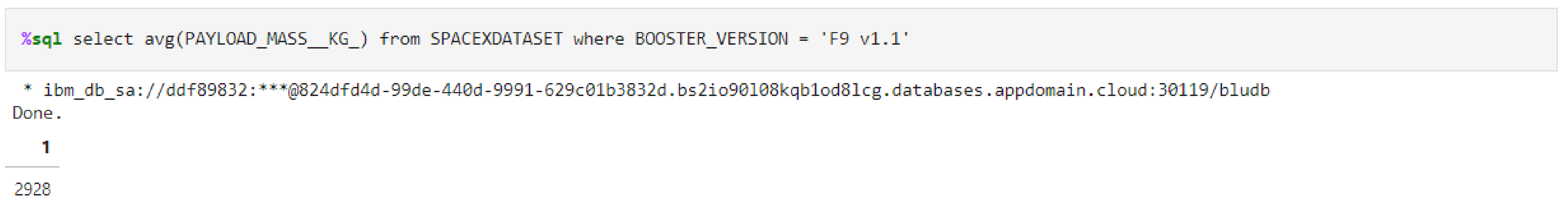


```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
    1
45596
```

Display the total payload mass carried by boosters launched by NASA (CRS). The result was 45 596 kg.

Average Payload Mass by F9 v1.1



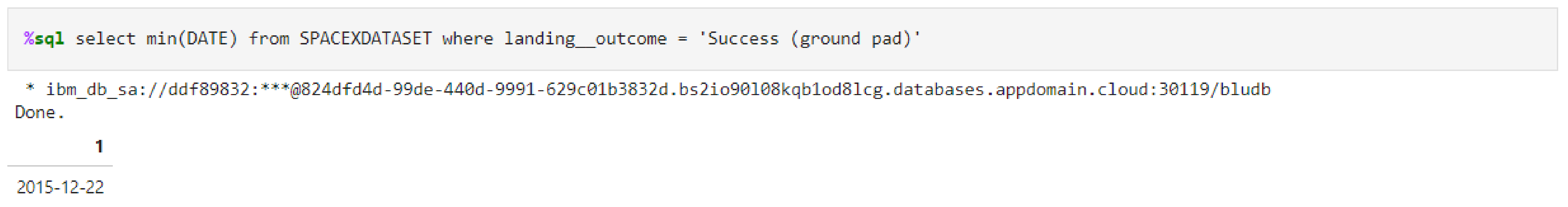
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2928

The average payload mass carried by booster version F9 v1.1 was 2928 kg.

First Successful Ground Landing Date



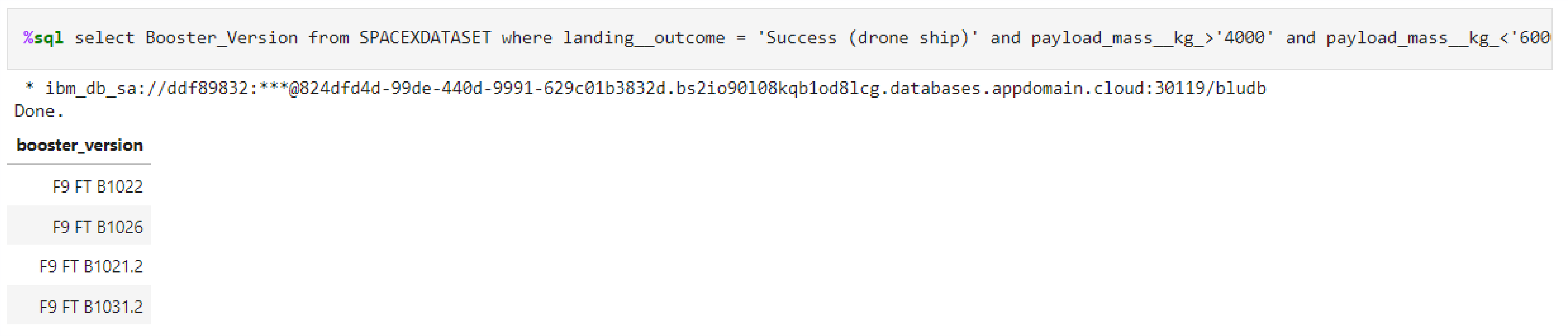
```
%sql select min(DATE) from SPACEXDATASET where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
2015-12-22

The date when the first successful landing outcome in ground pad was achieved was on December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000



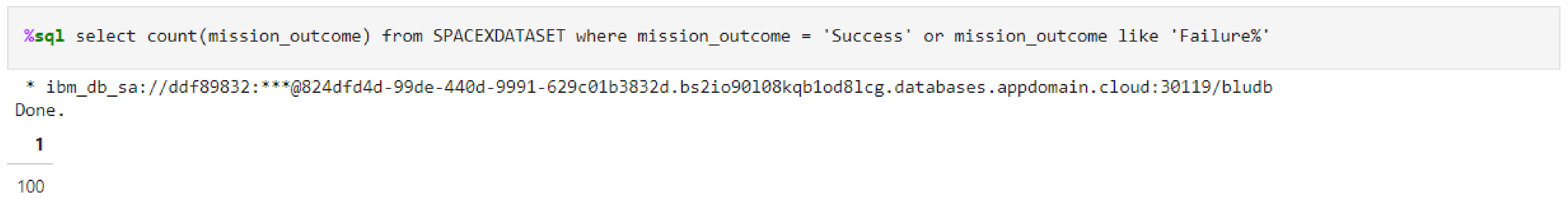
```
%sql select Booster_Version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_>'4000' and payload_mass__kg_<'600
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are **F9 FT B1022**, **F9 FT B1026**, **F9 FT B1021.2** and **F9 FT B1031.2**.

Total Number of Successful and Failure Mission Outcomes



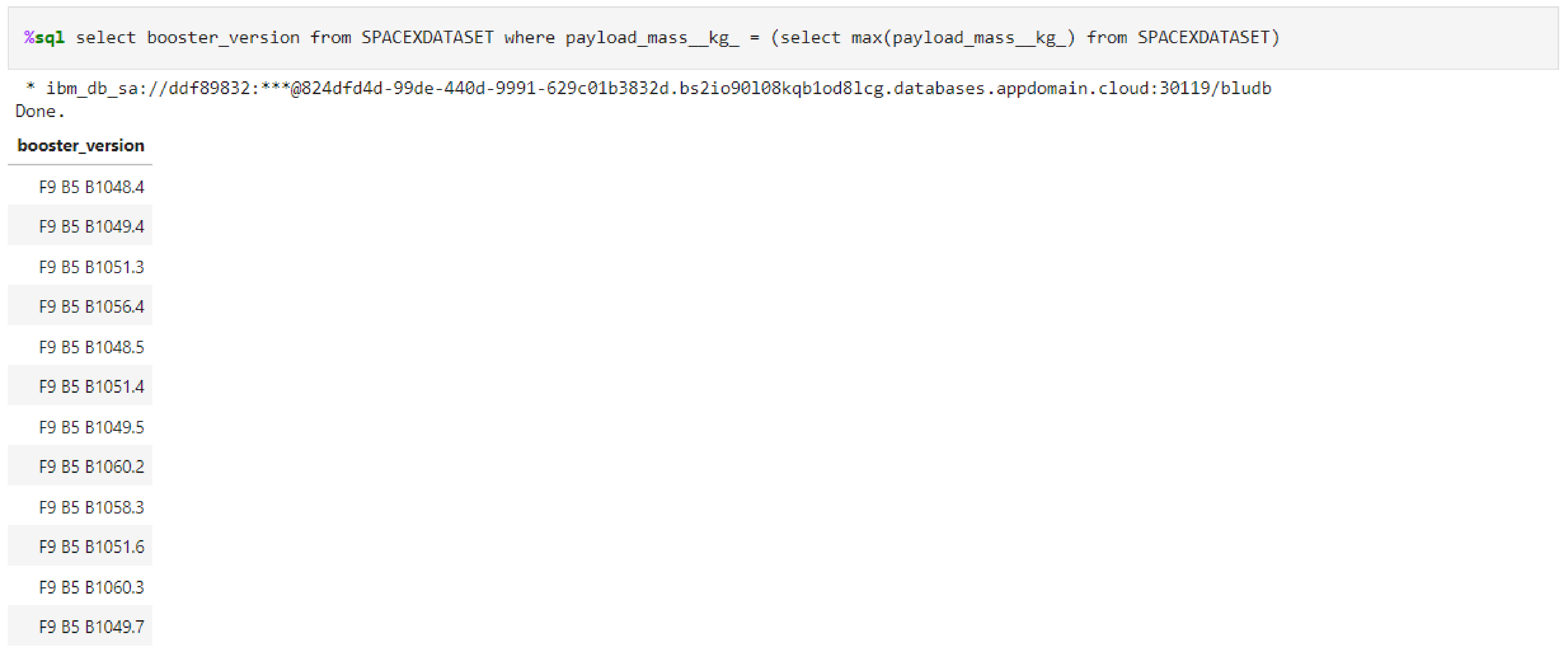
```
%sql select count(mission_outcome) from SPACEXDATASET where mission_outcome = 'Success' or mission_outcome like 'Failure%'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

1
100

There were 100 successful and failure mission outcomes.

Boosters Carried Maximum Payload



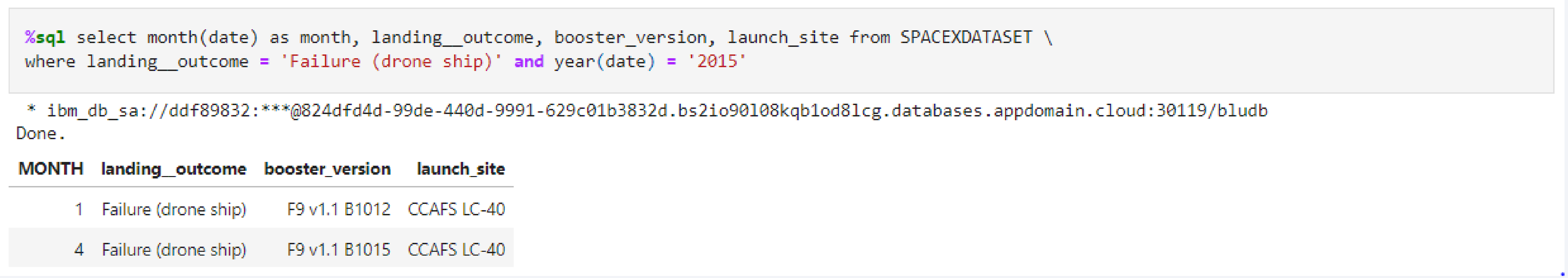
```
%sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/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

Using a subquery, there were found 12 different booster versions which have carried the maximum payload mass.

2015 Launch Records



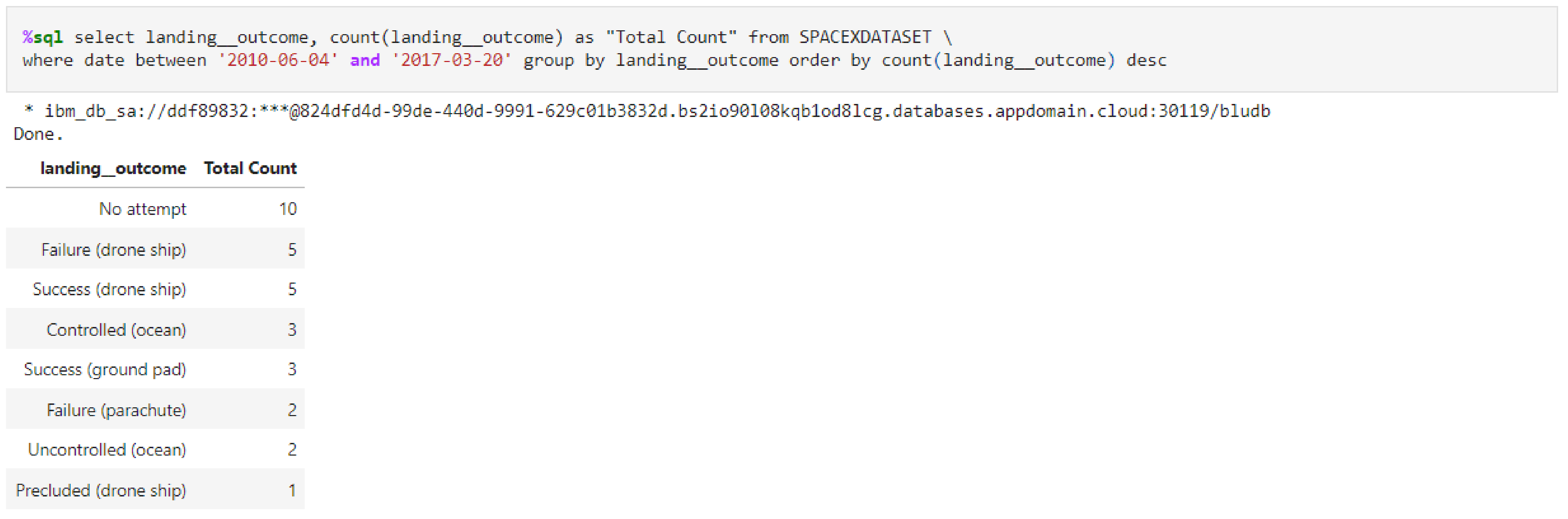
```
%sql select month(date) as month, landing__outcome, booster_version, launch_site from SPACEXDATASET \
where landing__outcome = 'Failure (drone ship)' and year(date) = '2015'
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015. The results presents one failed landing outcome in January and another in April, on the same launch site (CCAFS LC-40).

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



```
%sql select landing__outcome, count(landing__outcome) as "Total Count" from SPACEXDATASET \
where date between '2010-06-04' and '2017-03-20' group by landing__outcome order by count(landing__outcome) desc
```

```
 * ibm_db_sa://ddf89832:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

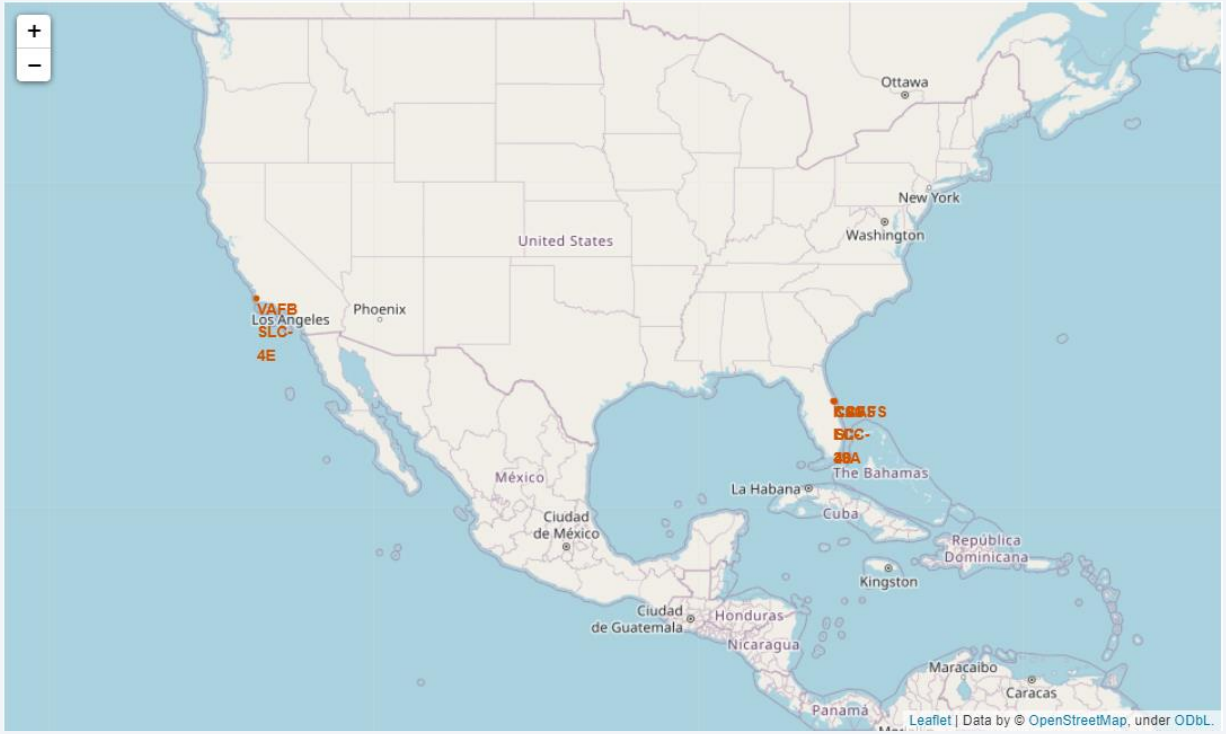
landing__outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Between 2010-06-04 and 2017-03-20, we can count 31 launches.

Section 3

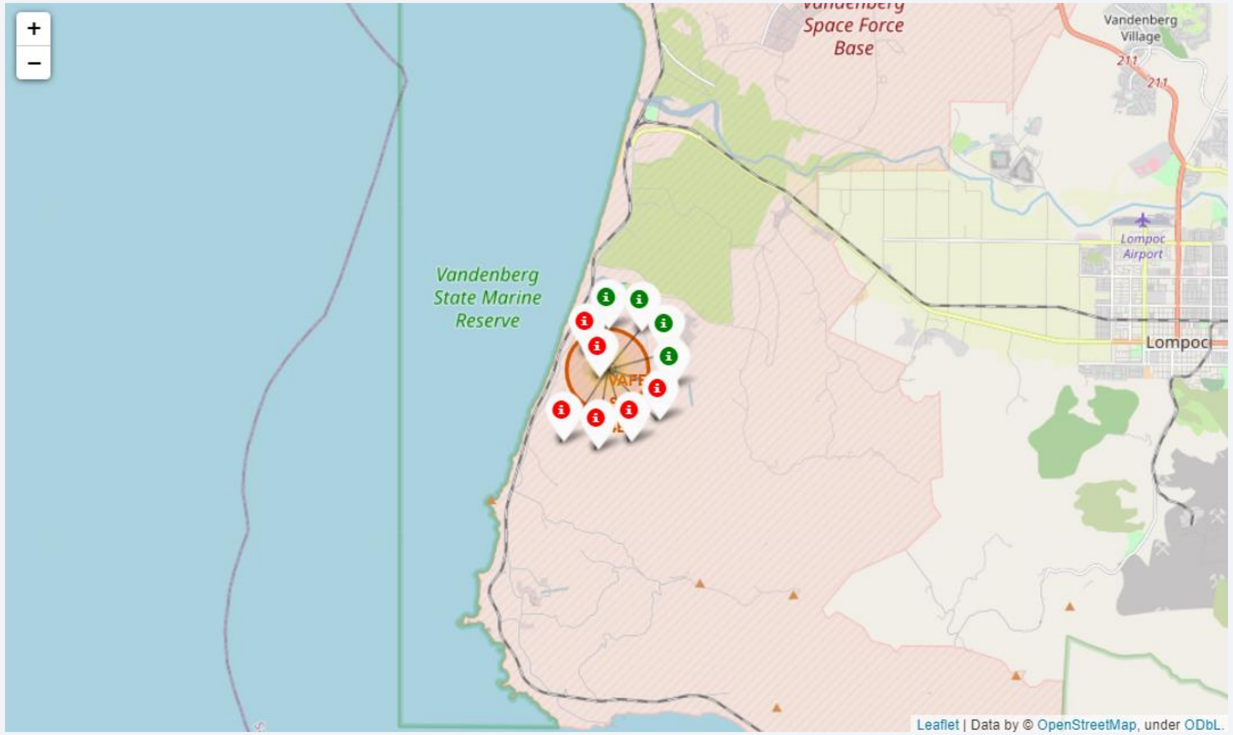
Launch Sites Proximities Analysis

Launch sites locations markers on a global map



- Are all launch sites in proximity to the Equator line?
  - Yes, most of them, because anything on the surface of the Earth at the equator is already moving at 1670 kilometers per hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- Are all launch sites in very close proximity to the coast?
  - Yes, they are, because launching a rocket from the east coast gives an additional boost to the rocket, due to the rotational speed of Earth. Also, these rockets travel eastward, so if anything goes wrong during their ascent, the debris would essentially fall into an ocean's waters, far away from densely populated areas.

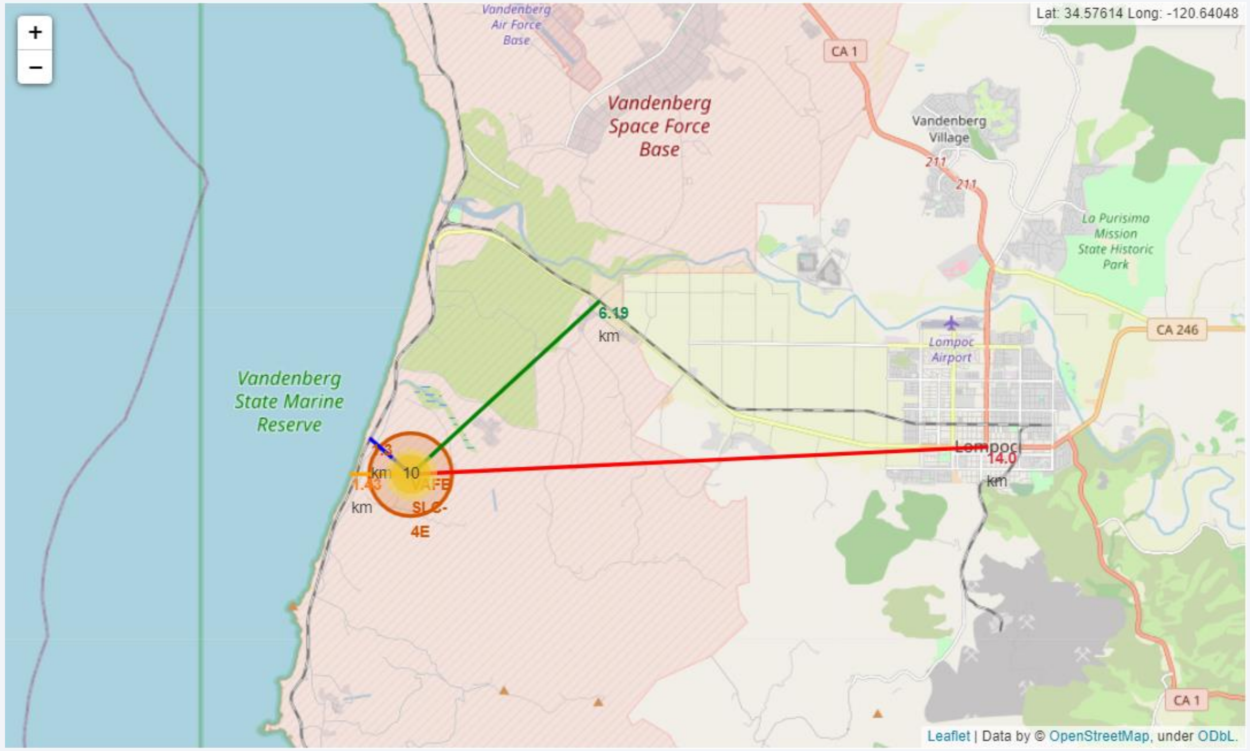
Success and failed launches for each site on the map


+
−
Vandenberg Space Force Base
Vandenberg Village
211
Lompoc Airport
Vandenberg State Marine Reserve
VAFB
Lompoc
Leaflet | Data by © OpenStreetMap, under ODbL.

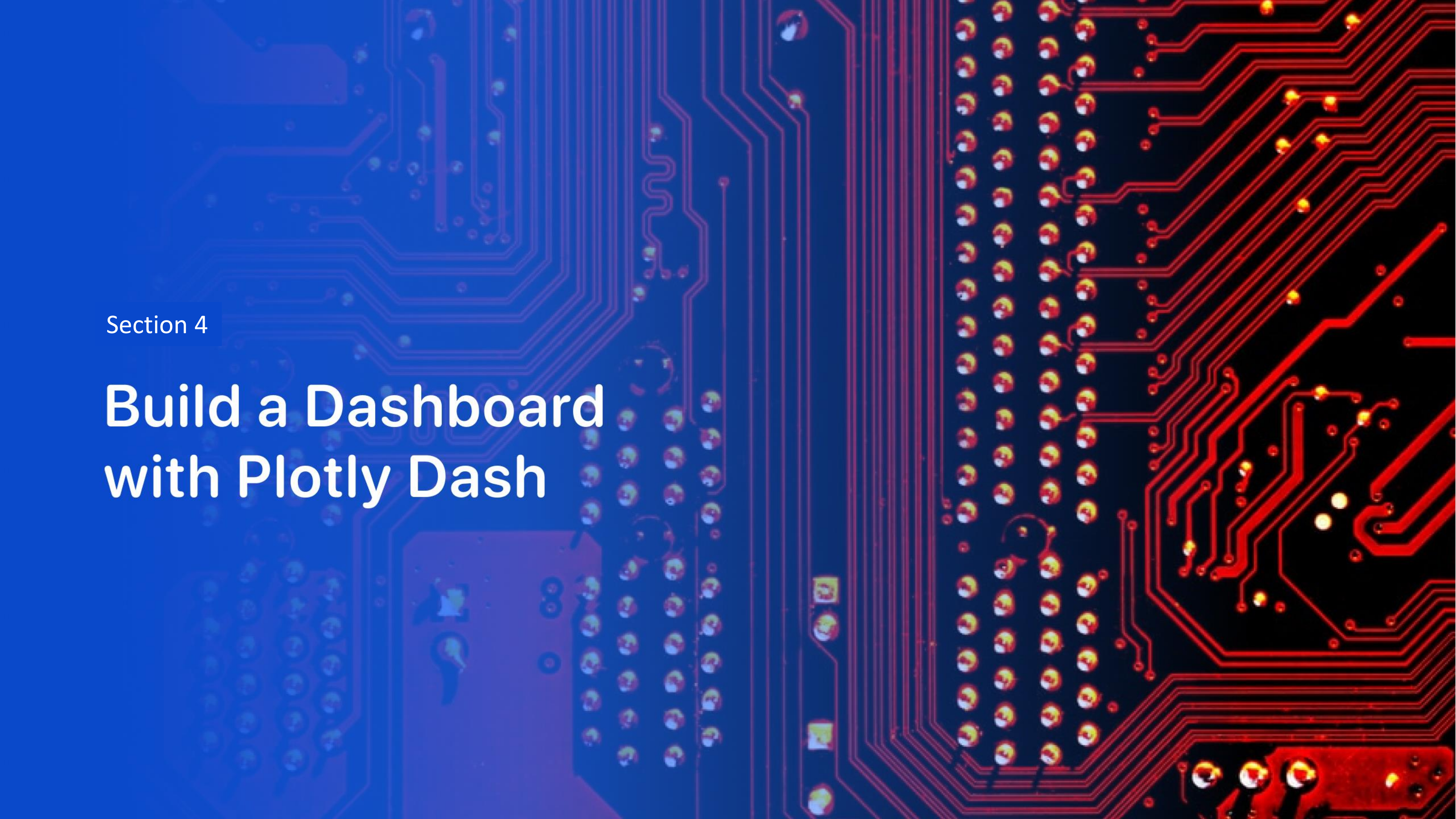
- If a launch was successful (class = 1), then we use a green marker and if a launch was failed, we use a red marker (class = 0).

- We can see that in VAFB SLC-4E launch site, in 10 launches, 4 of them were successful.

Distance from VAFB SLC-4E launch site to its proximities


+
−
Lat: 34.57614 Long: -120.64048
Vandenberg Air Force Base
CA 1
Vandenberg Space Force Base
Vandenberg Village
211
La Purisima Mission State Historic Park
6.19 km
CA 246
Lompoc Airport
Vandenberg State Marine Reserve
Lompoc
14.0 km
1.43 km
10
VAFB SLC-4E
CA 1
Leaflet | Data by © OpenStreetMap, under ODbL.

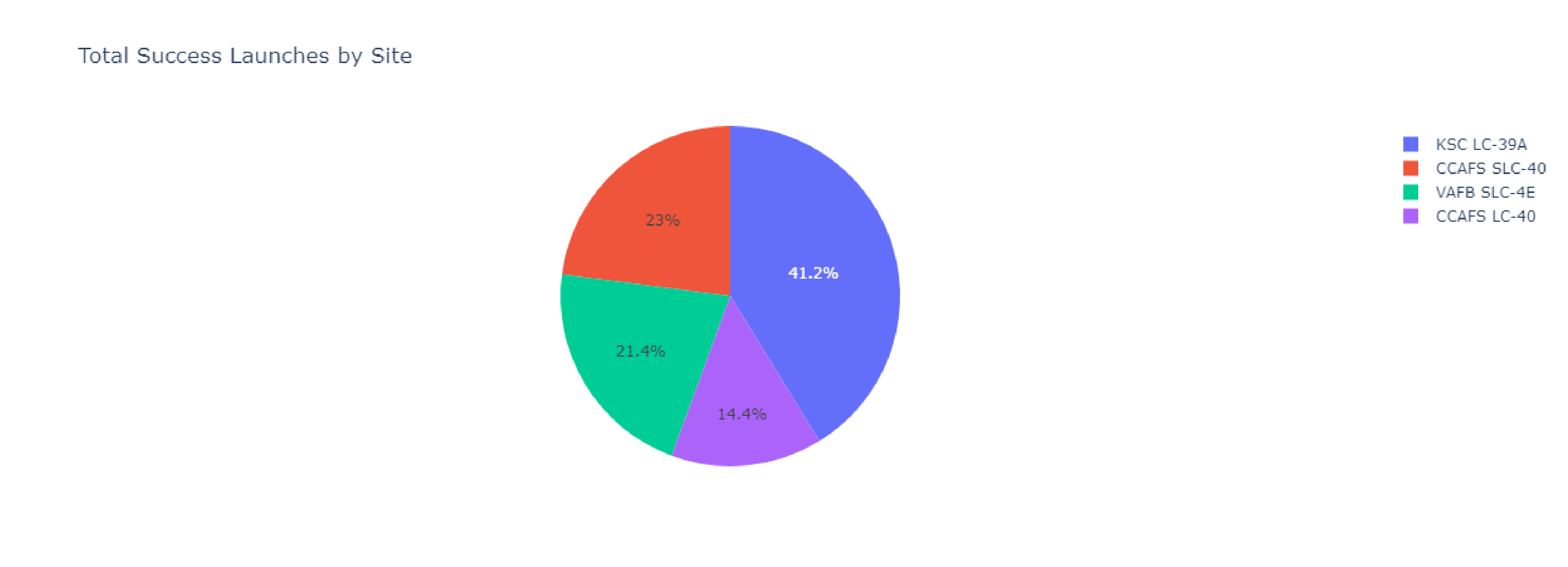
- From the visual analysis of the launch site VAFB SLC-4E we can clearly see that it is:
  - very close to railway (1.3 km)
  - close to highway (6.19 km)
  - very close to coastline (1.43 km)
- Also the launch site VAFB SLC-4E is relative close to its closest city Lompoc (14 km).



Section 4

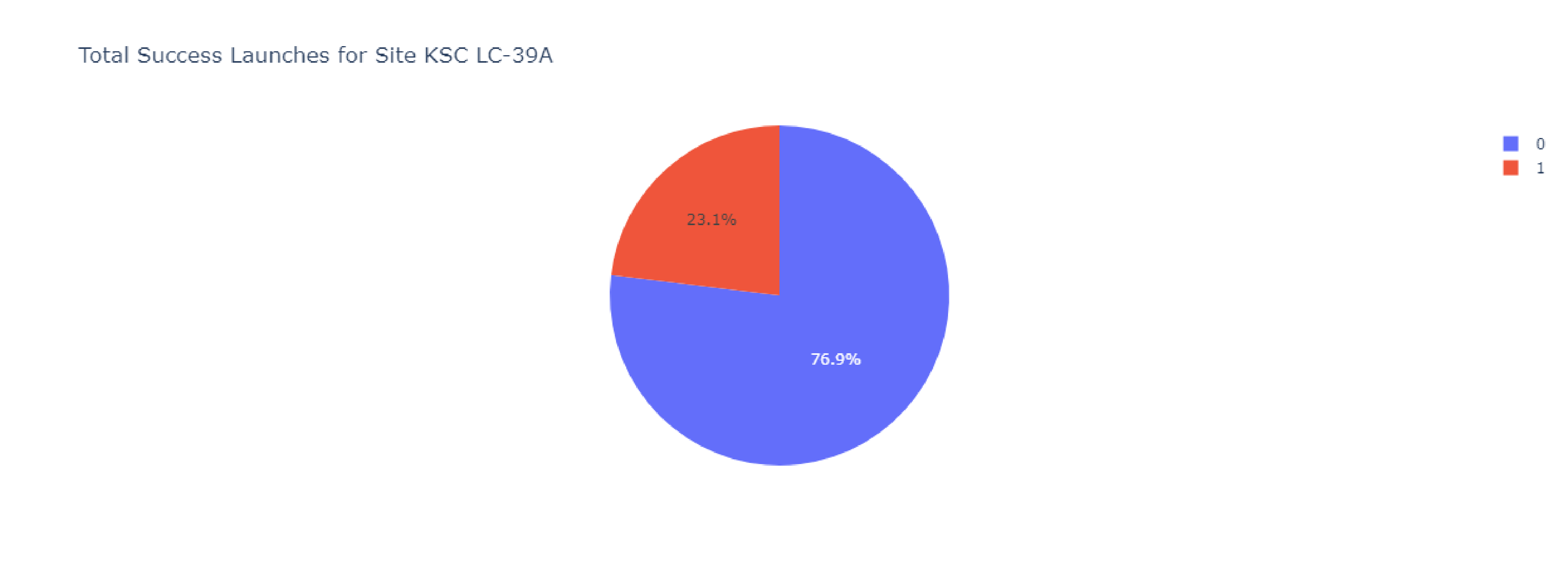
Build a Dashboard with Plotly Dash

Total Success Launches by Site



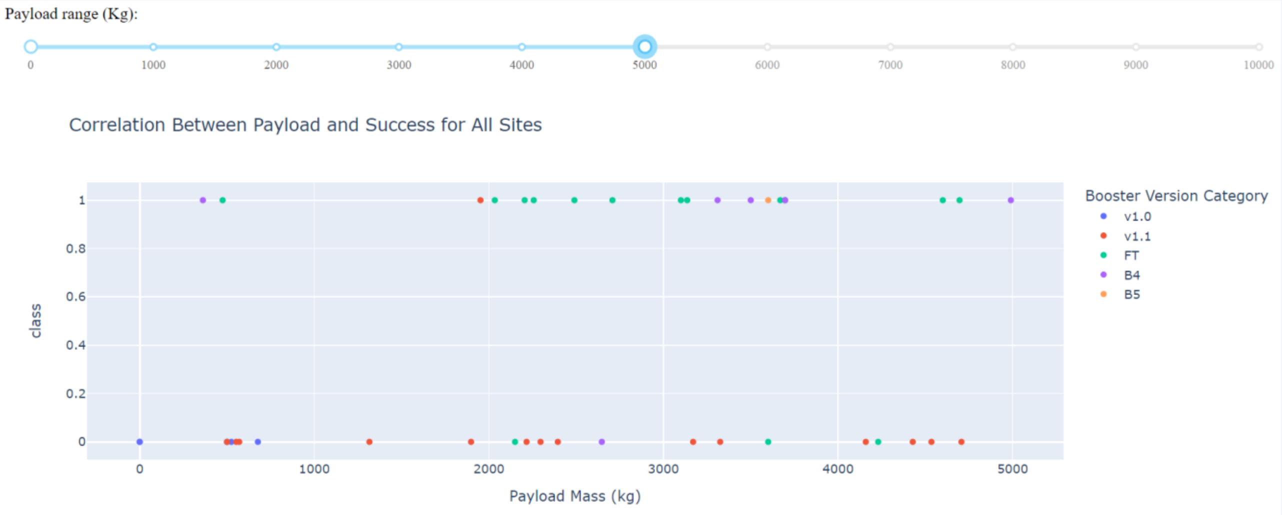
- From all sites, we can clearly see that KSC LC-39A is the launch site with higher success launches.
- CCAFS LC-40 is the launch site with the lowest success launches.

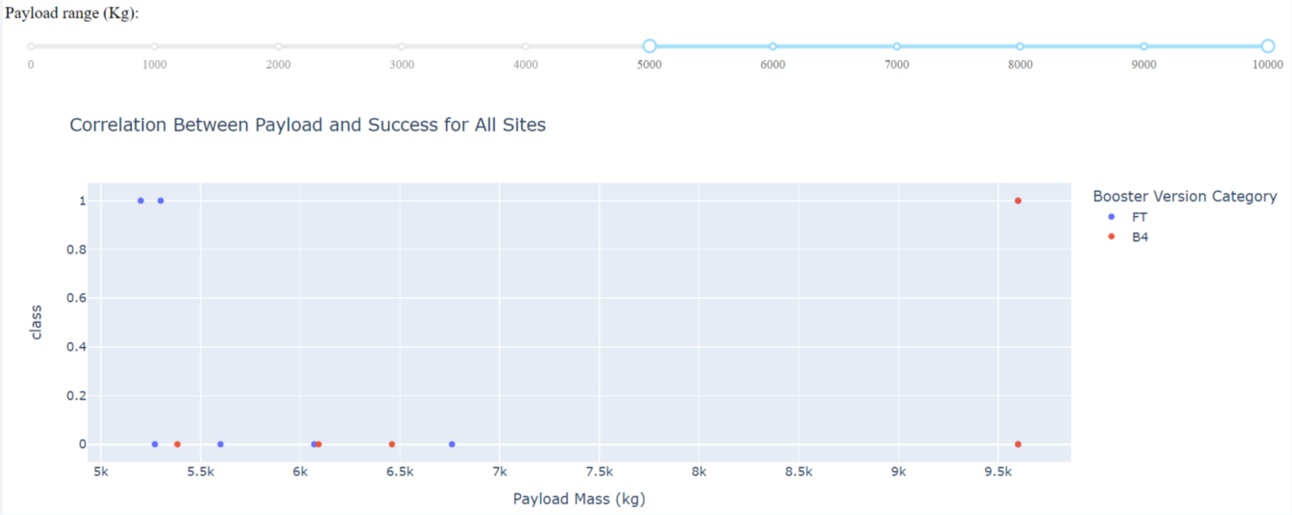
Launch Site with highest launch success ratio



- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landings.

Payload Mass vs Launch Outcome for all sites



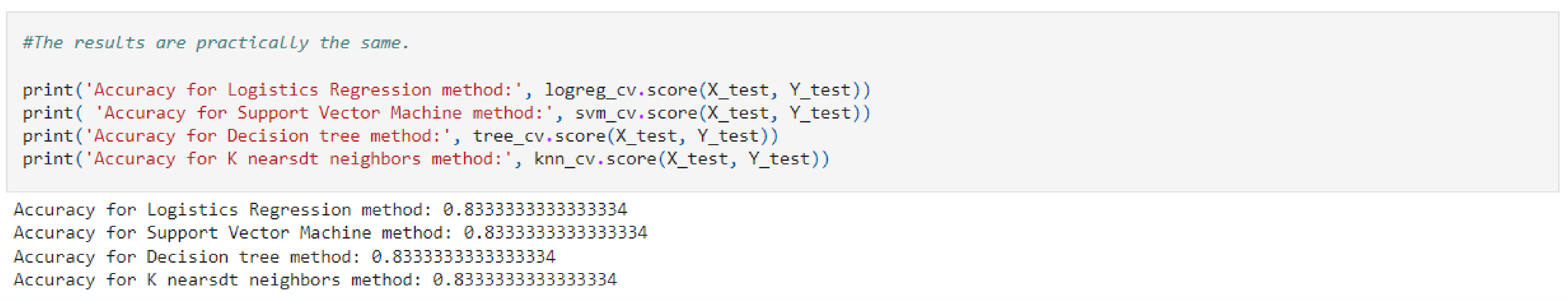


- We can see from the charts that payloads between 2000 and 5500 kg have the highest success rate.
- Booster version FT has the highest success rate between that payloads range.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



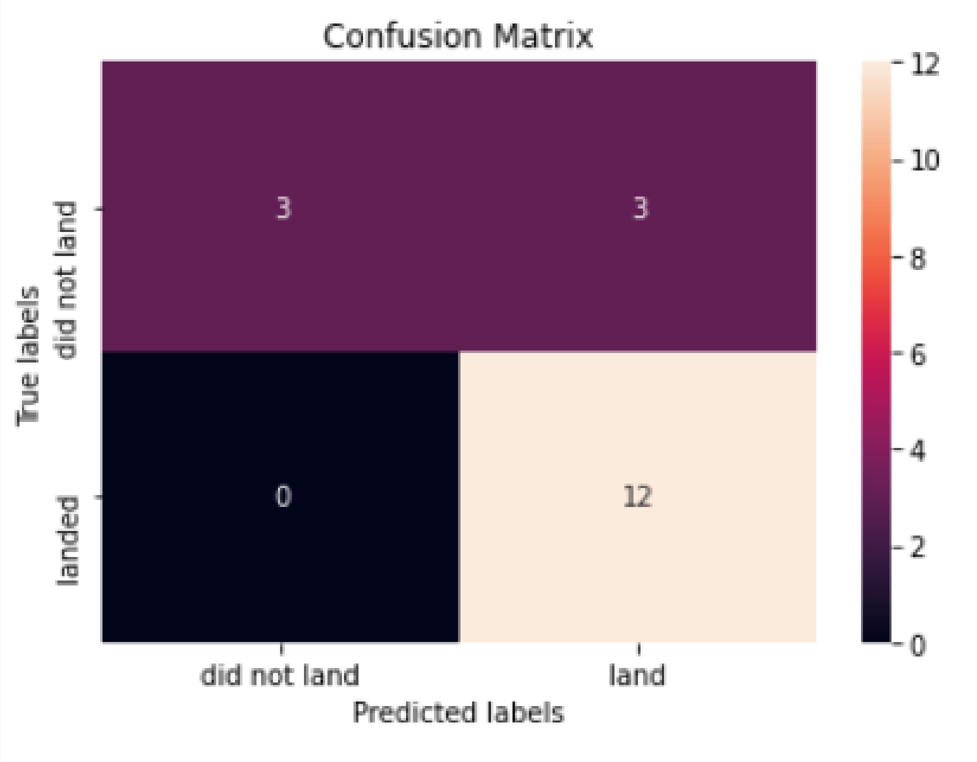
```
#The results are practically the same.

print('Accuracy for Logistics Regression method:', logreg_cv.score(X_test, Y_test))
print( 'Accuracy for Support Vector Machine method:', svm_cv.score(X_test, Y_test))
print('Accuracy for Decision tree method:', tree_cv.score(X_test, Y_test))
print('Accuracy for K nearsdt neighbors method:', knn_cv.score(X_test, Y_test))
```

```
Accuracy for Logistics Regression method: 0.8333333333333334
Accuracy for Support Vector Machine method: 0.8333333333333334
Accuracy for Decision tree method: 0.8333333333333334
Accuracy for K nearsdt neighbors method: 0.8333333333333334
```

- Based on the scores of the test set, we can not confirm which method performs best because the accuracy results are practically the same.
- This is because the dataset is small and has lesser values (18 samples).

Confusion Matrix



Since the results are practically the same, the confusion matrix is similar for all methods.

Conclusions

- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Launches with a low payload mass show better results than launches with a larger payload mass.
- The success rate of launches increases over the years.
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- The accuracy results are practically the same, therefore we can not conclude which method (Logistics Regression, Support Vector Machine, Decision tree or KNN) performs best.

Thank you!