IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Carlos Rey
21/11/2021

Outline

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

Executive Summary

- Methodologies:
  - Web Scrapping with BeautifulSoup
  - Data wrangling with pandas and numpy
  - First EDA with SQL
  - EDA and Preparing Data Feature Engineering
  - Location Analysis with Folium
  - Machine Learning Prediction with multiple models
- Results:

Introduction

- Rocket launches usually cost upward of 165 million dollars each.
- SpaceX advertises Falcon 9 rocket launches will cost 62 million dollars this could be possible because of reusing first stage.
- These savings could make possible easier and more reachable spatial research.
- Can we determine which variables influence a successful first stage land?
- Can we predict if a first stage will land?
- Thus, can we ensure first stage savings?

Section 1

Methodology

Methodology

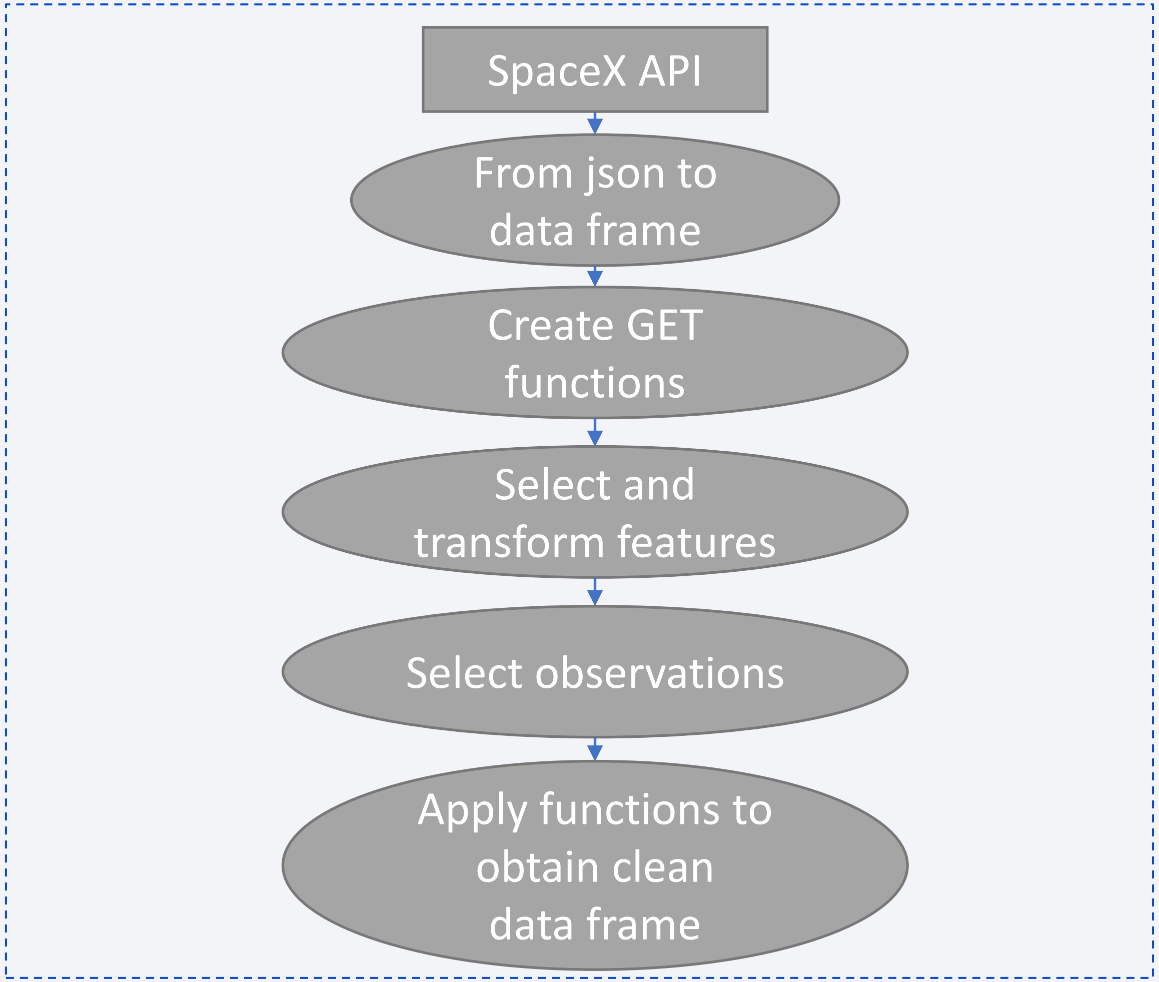
Executive Summary

- Data collection methodology:
  - Web scraping from Wikipedia Falcon 9 historical launch records.
- Perform data wrangling
  - Preprocess data and add a new column ‘class’
- 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, evaluate classification models

Data Collection – SpaceX API

- Data collection from SpaceX API
    - Extract data
    - Data wrangling
    - Filter data frame
    - Deal with missing values

- GitHub Repository



SpaceX API

From json to data frame

Create GET functions

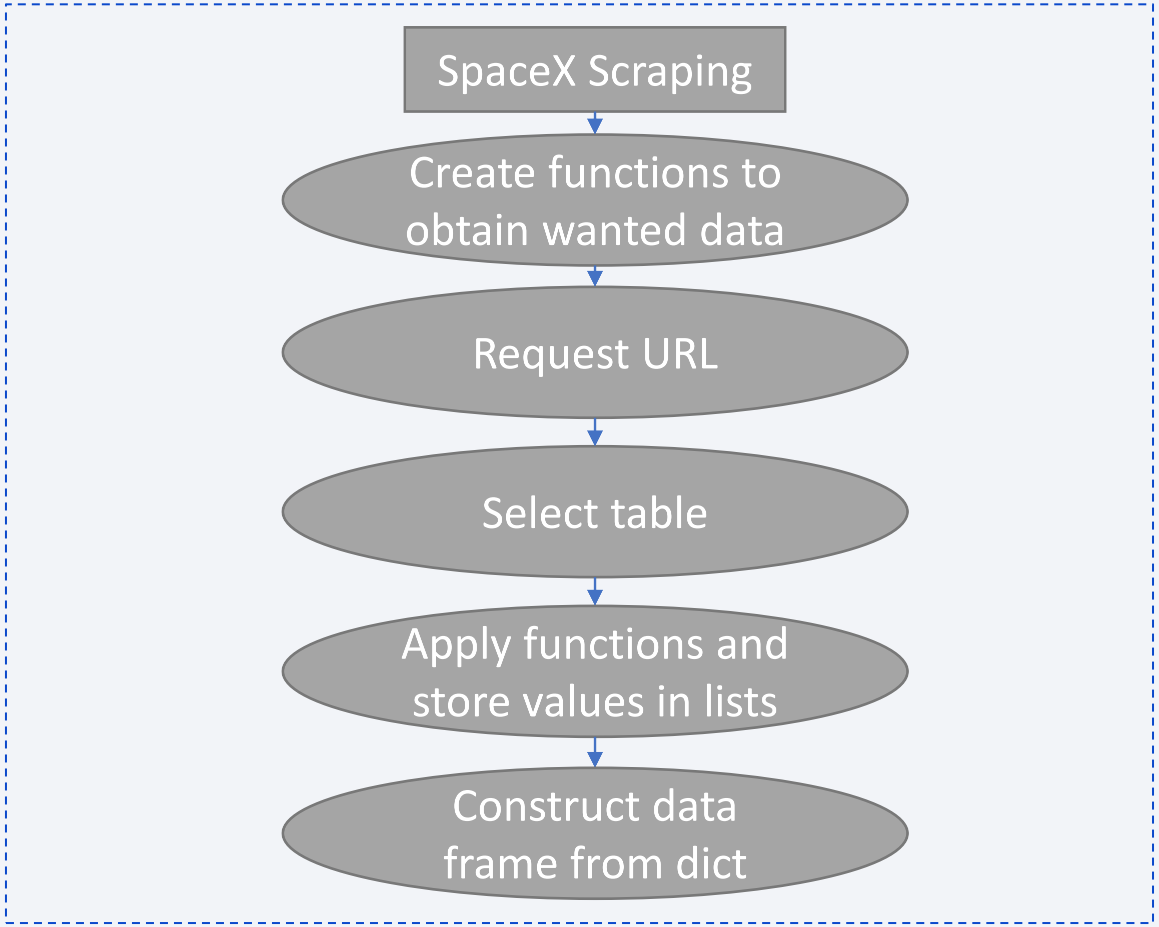
Select and transform features

Select observations

Apply functions to obtain clean data frame

Data Collection - Scraping

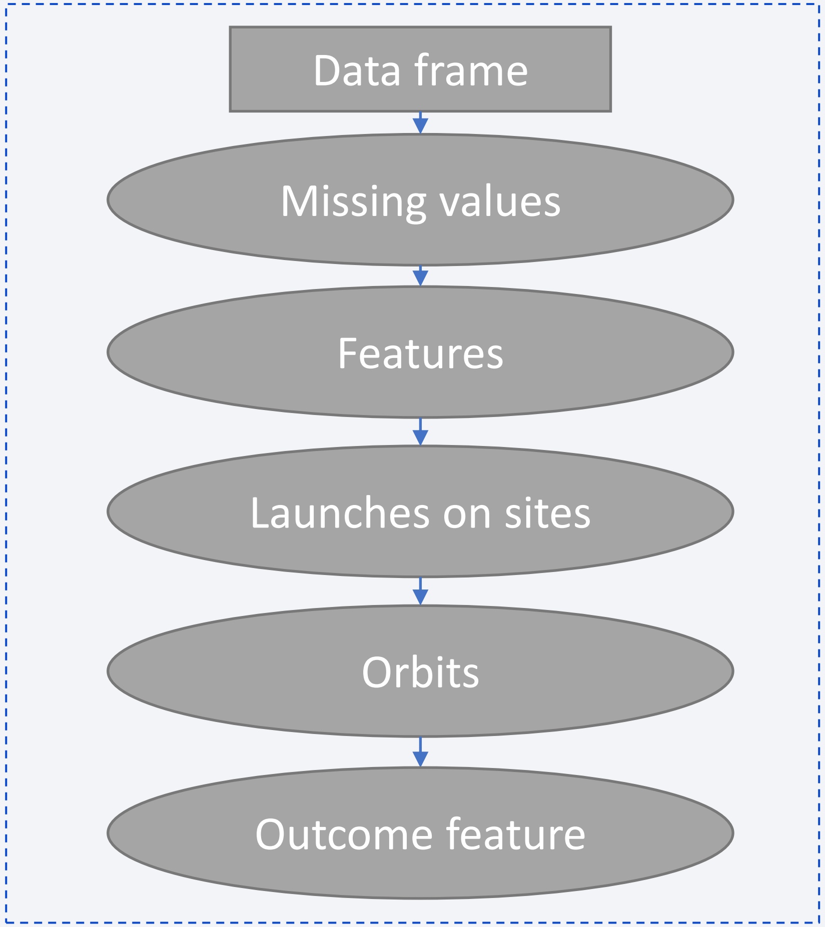
- Extract Falcon 9 launch records HTML table from Wikipedia
- Using BeautifulSoup
- Parse HTML tables to create data frame
- GitHub


SpaceX Scraping
Create functions to obtain wanted data
Request URL
Select table
Apply functions and store values in lists
Construct data frame from dict

Data Wrangling

- Identification and sum of missing values per feature.
- Identification of feature types.
- Analysis of launches on each site.
- Analysis of number and occurrence of each orbit.
- Analysis of mission outcome number and occurrence per orbit type
- Creation of new column which determine outcome:
  - Successful landing
  - Bad landing

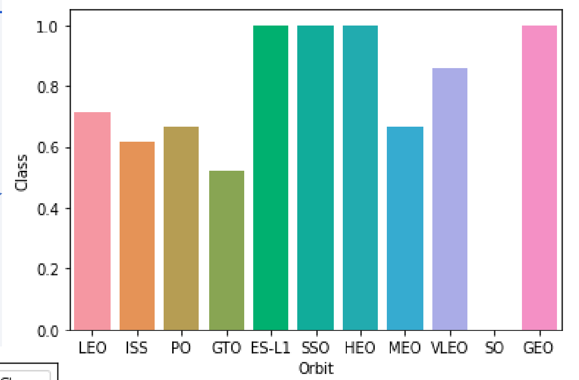
- GitHub

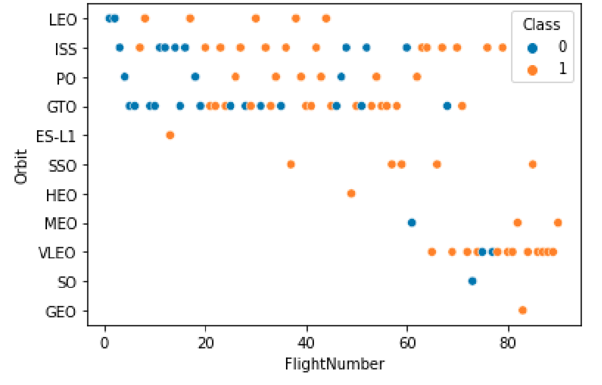


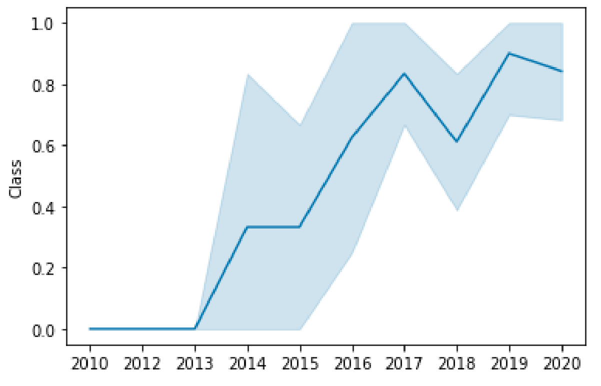
EDA with Data Visualization

- Relation between Flight Number and Launch Site.
- Relationship between Payload and Launch Site.
- Relationship between success rate of each orbit type.
- Relationship between Flight Number and Orbit type
- Relationship between Payload and Orbit type
- Launch success yearly trend

- GitHub







EDA with SQL

- Display names of unique launch sites.
- 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.
- 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 between 4000 and 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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20
- GitHub

Build an Interactive Map with Folium

- I created and added:
  - Red circles and markers, to point cities.
  - Green markers, to point successful launch outcomes.
  - Red markers, to point unsuccessful launch outcomes.
  - Lines to calculate distance between near coastline and launch site.
  - Lines to calculate distance between near city and launch site.

- GitHub

Build a Dashboard with Plotly Dash

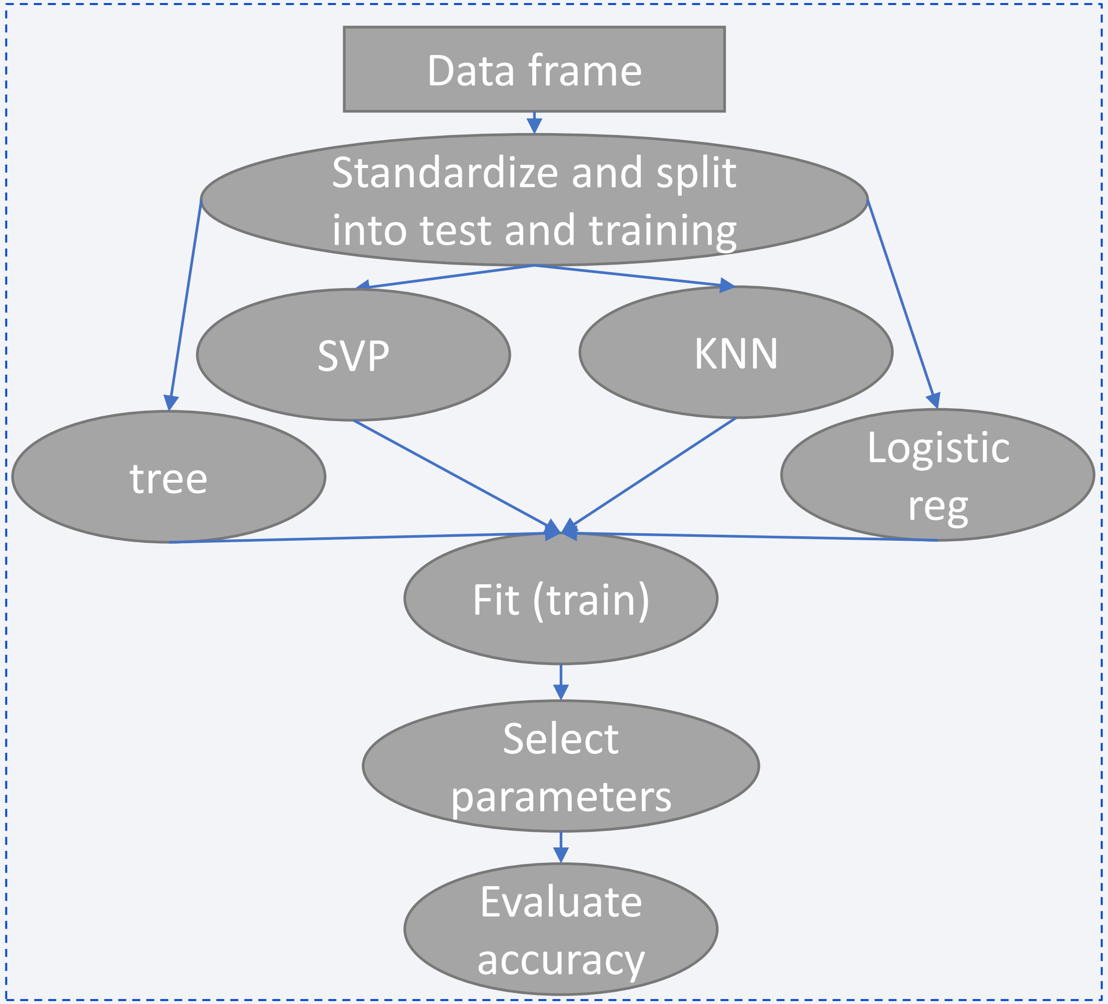
- I created:
  - A pie chart of total successful launches count for all sites and for every launch site.
  - Pie charts for each launch site showing success vs failed counts.
  - A scatter plot to show correlation between payload and launch success within a range (selected with a slider).
- What can you visualize with these plots?
  - Ratio of successful launches for all launch sites and for each one.
  - Relationship between payload and launch success in selected payload range.

- GitHub

Predictive Analysis (Classification)

- Standardize data and split into training and test data
- For logistic regression, support vector machine, tree classifier and k nearest neighbors:
  - I trained training data.
  - Select best parameters and best accuracy.
  - Plot confusion matrix and the accuracy of the model.

- [GitHub]()

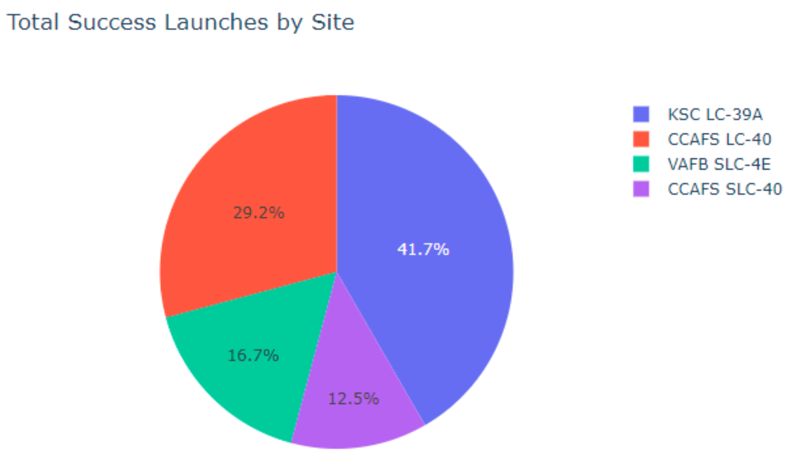


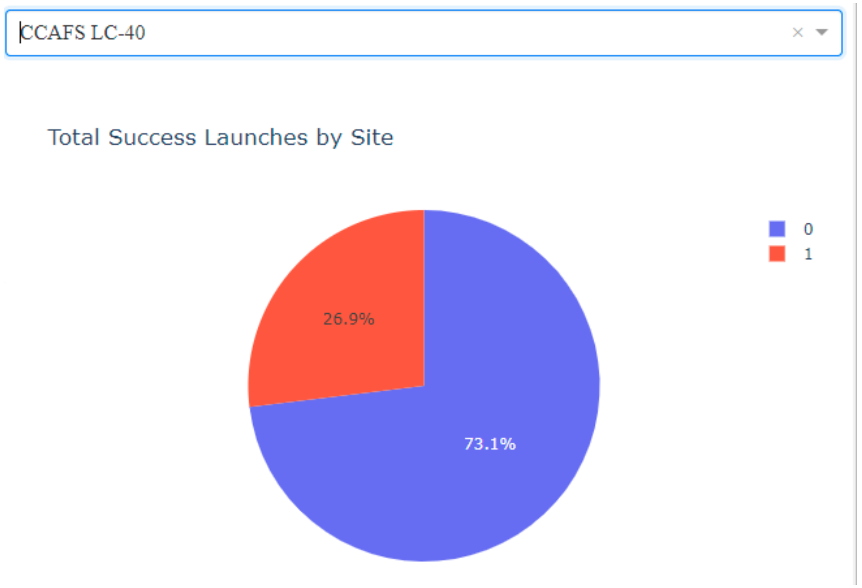
Results

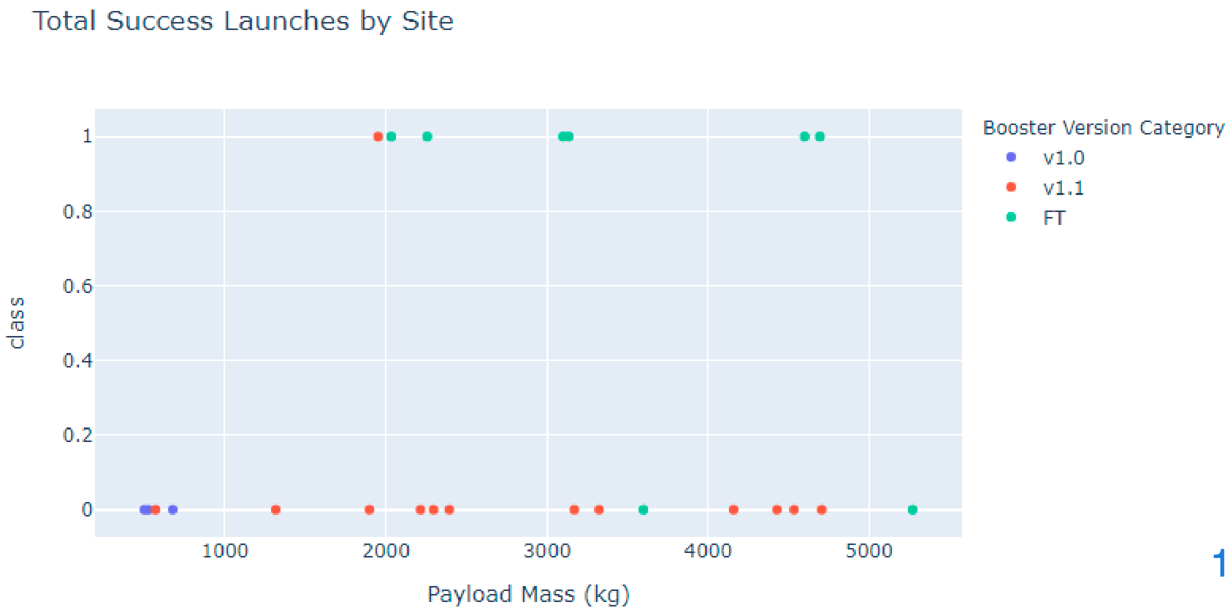
- Exploratory data analysis results:
  - 2928kg average payload mass carried by booster version F9 v1.1
  - First successful landing in ground pad was 2015-12-22
  - The success landing ratio is greater than the failure ratio but must of the time no landing was tried.
  - ES-L1, SSO, HEO and GEO orbits have a 100% success rate.
  - These orbits also have fewer Flight Numbers (they are new ones)
  - Success rate is increasing yearly.
- Interactive analytics demo in screenshots
- Predictive analysis results

Results

- Interactive analytics

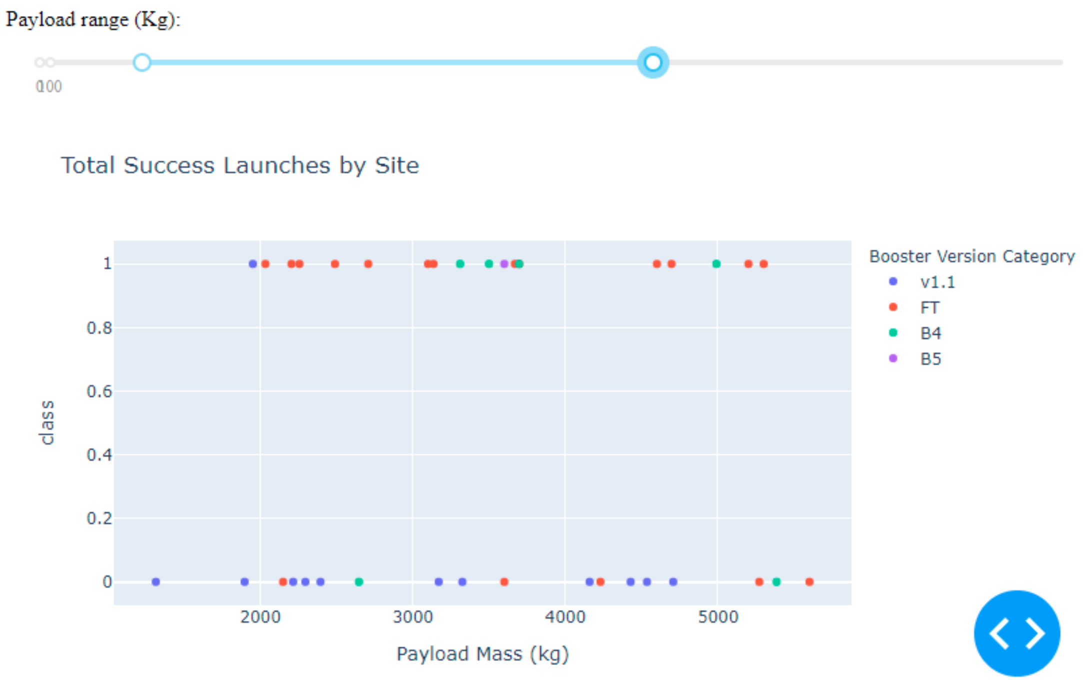



CCAFS LC-40

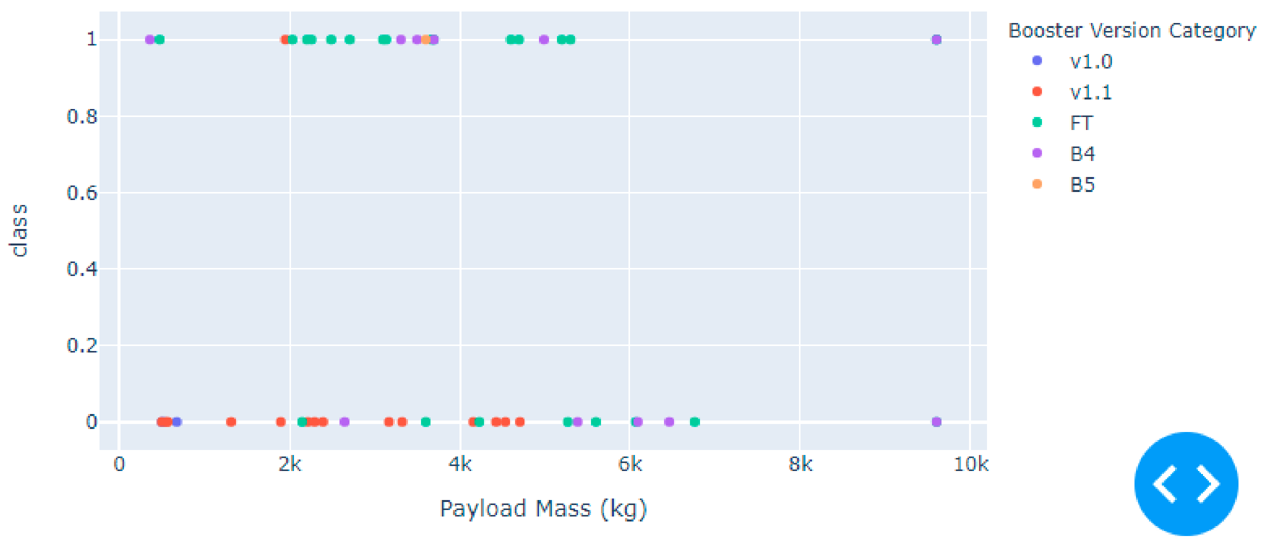


Results

- Interactive analytics


Payload range (Kg):

Total Success Launches by Site


Total Success Launches by Site

Results

- Predictive analysis results:
  - Logistic regression:
    - Parameters:
      - C: 0.01
      - penalty : 12
      - solver: 'lbfgs'
    - Accuracy: 0.834
  - Support Vector Machine:
    - Parameters:
      - C: 0.01
      - gamma : 0.03162
      - kernel: sigmoid
    - Accuracy: 0.834

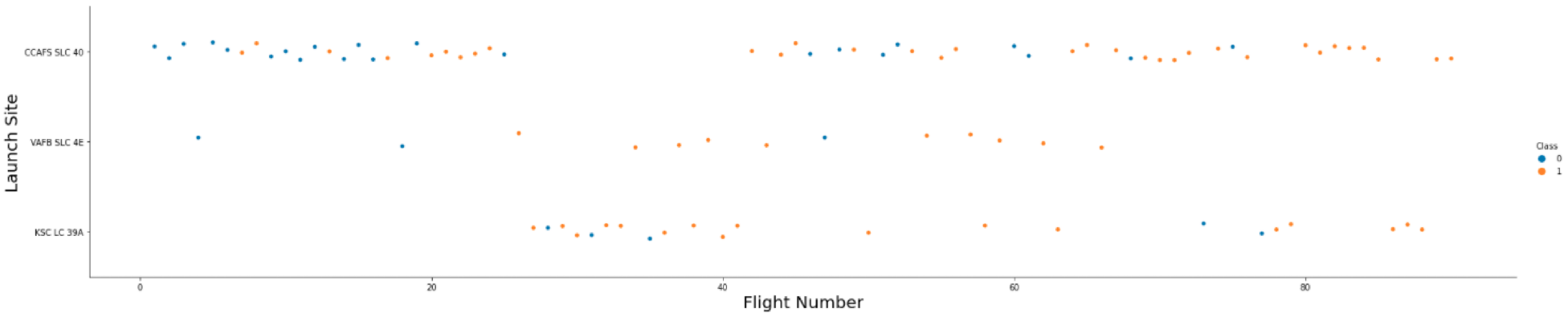
- Decision tree
  - Parameters:
    - 'criterion': 'entropy’
    - 'max_depth': 4
    - 'max_features': 'sqrt’
    - 'min_samples_leaf': 4
    - 'min_samples_split’: 2
    - 'splitter': 'random'}
  - Accuracy: 0.834
- K nearest neighbors
  - Parameters:
    - 'algorithm': 'auto’
    - 'n_neighbors': 10,
    - p': 1
  - Accuracy: 0.834

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Flight Number vs. Launch Site



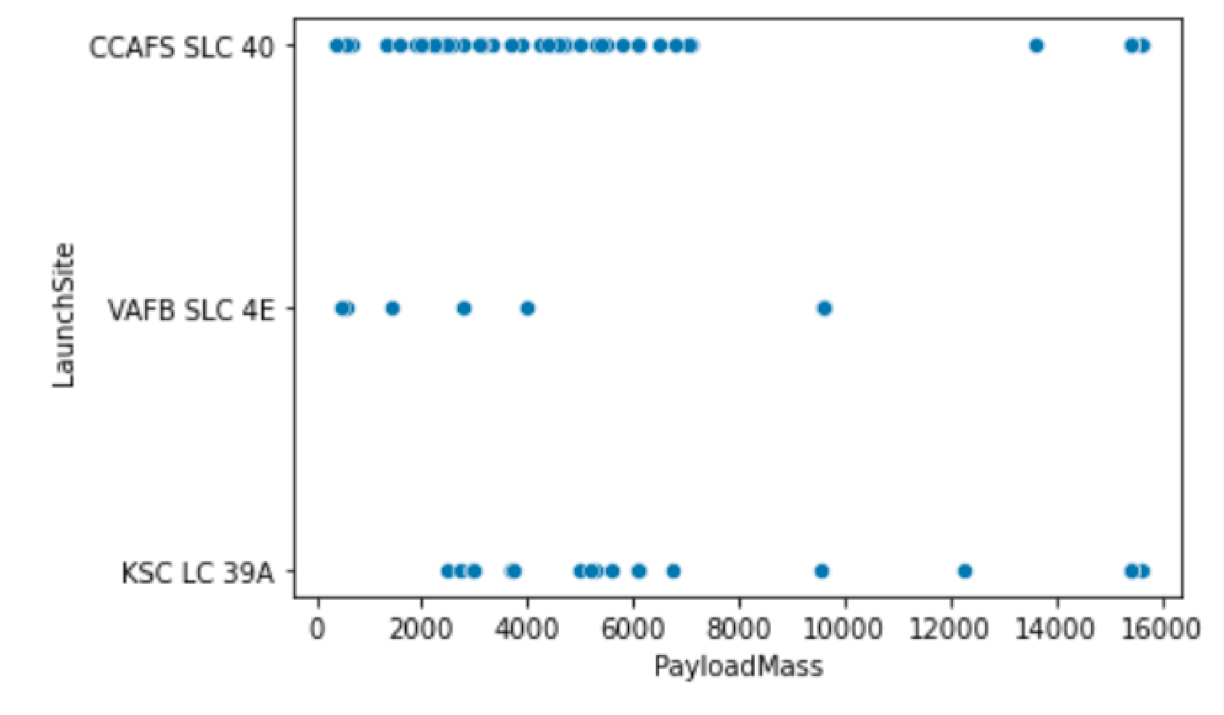
Each point represents a launch. If it is blue landing failed, if it is yellow landing was successful.

CCAFS SLC 40 is the older launch site, and which has more flights.

VAFB SLC 4E has fewer flights than the others, and last flights didn’t take place there.

Payload vs. Launch Site

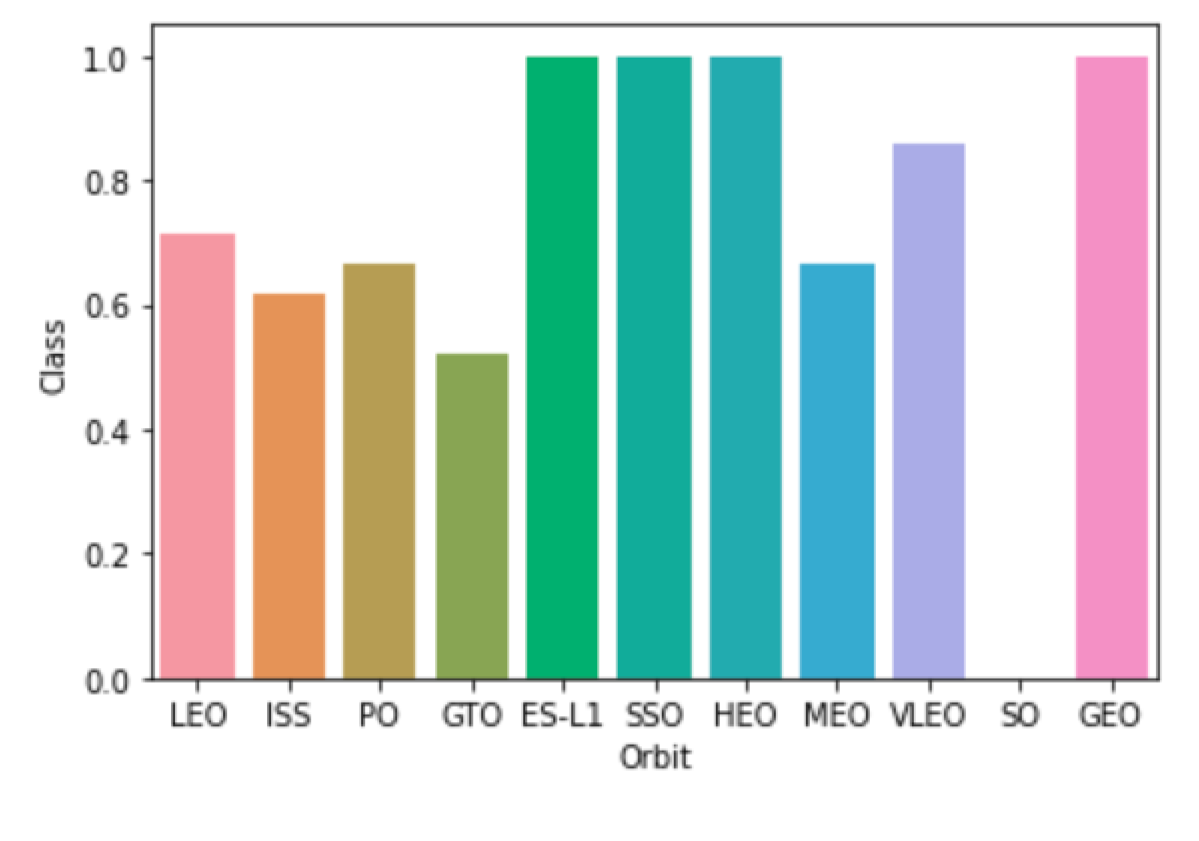
- Payload vs. Launch Site



- Each point represents a launch.
- We can see the relationship between launch site and payload mass.
- CCAFS SLC 40 and VAFB SLC 4E usually have launches with fewer payload mass.

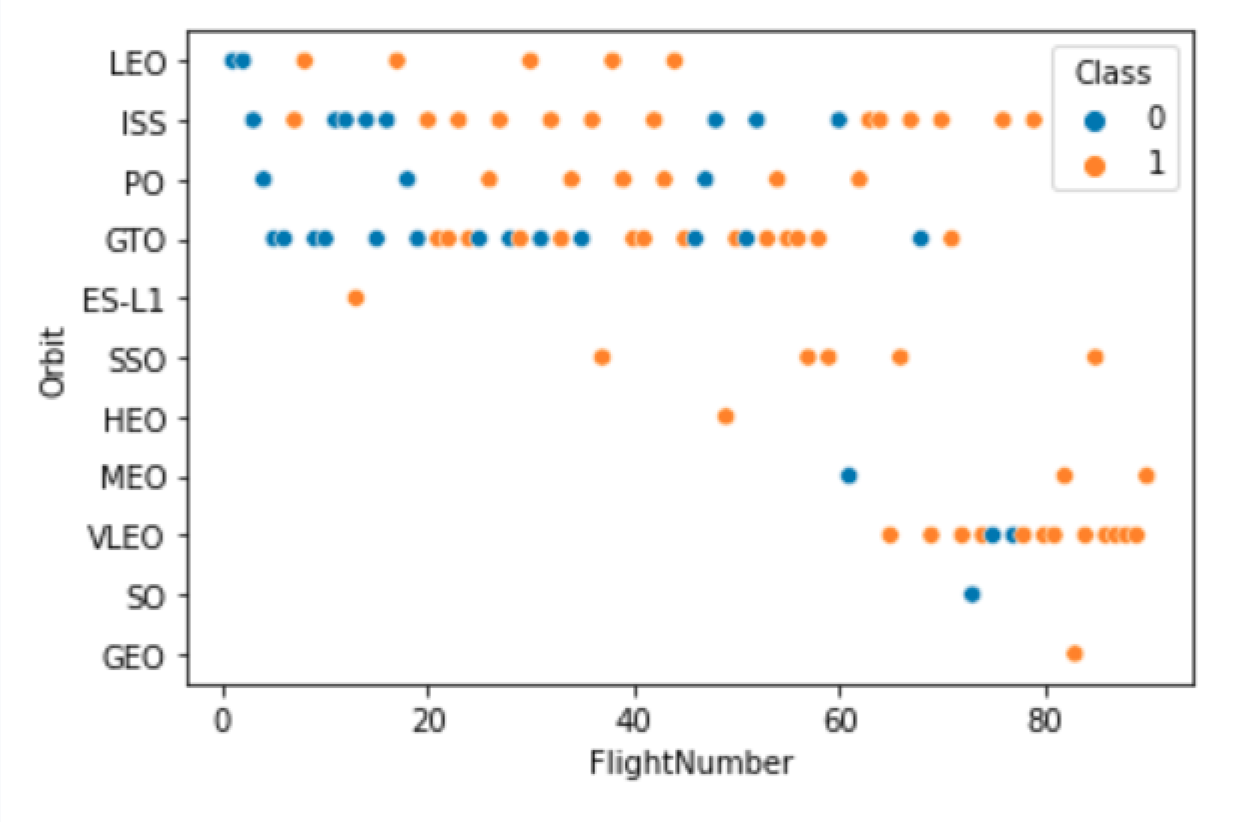
Success Rate vs. Orbit Type

- Each bar represents an orbit.
- Each length bar represents success rate.
- SO orbit has 0 success rate.
- ES-L1, SSO, HEO and GEO have 1 success rate



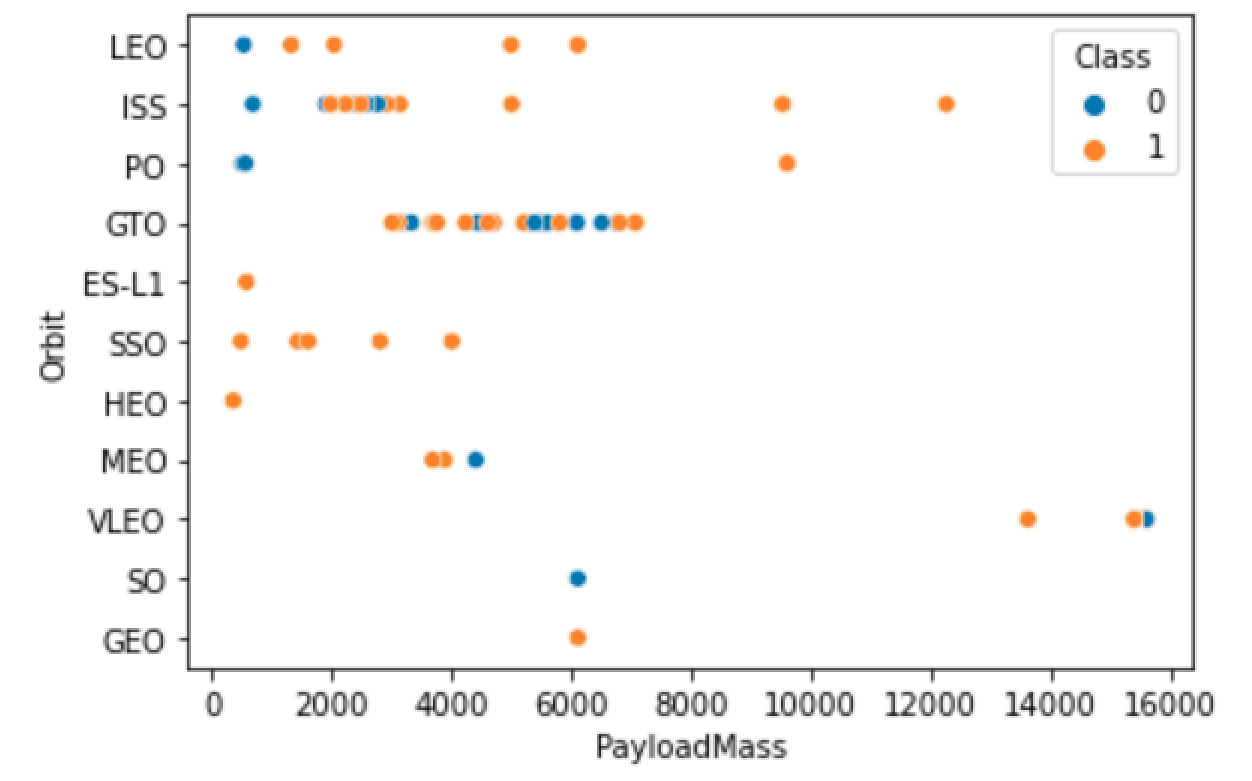
Flight Number vs. Orbit Type

- Each point represents a launch.
- If point is blue landing failed
- If point is yellow landing was successful.
- Plot represents relationship between flight number and orbit.
- We can see:
  - When flight number increases number of flights decreases.
  - Some orbits has just a few flight numbers.



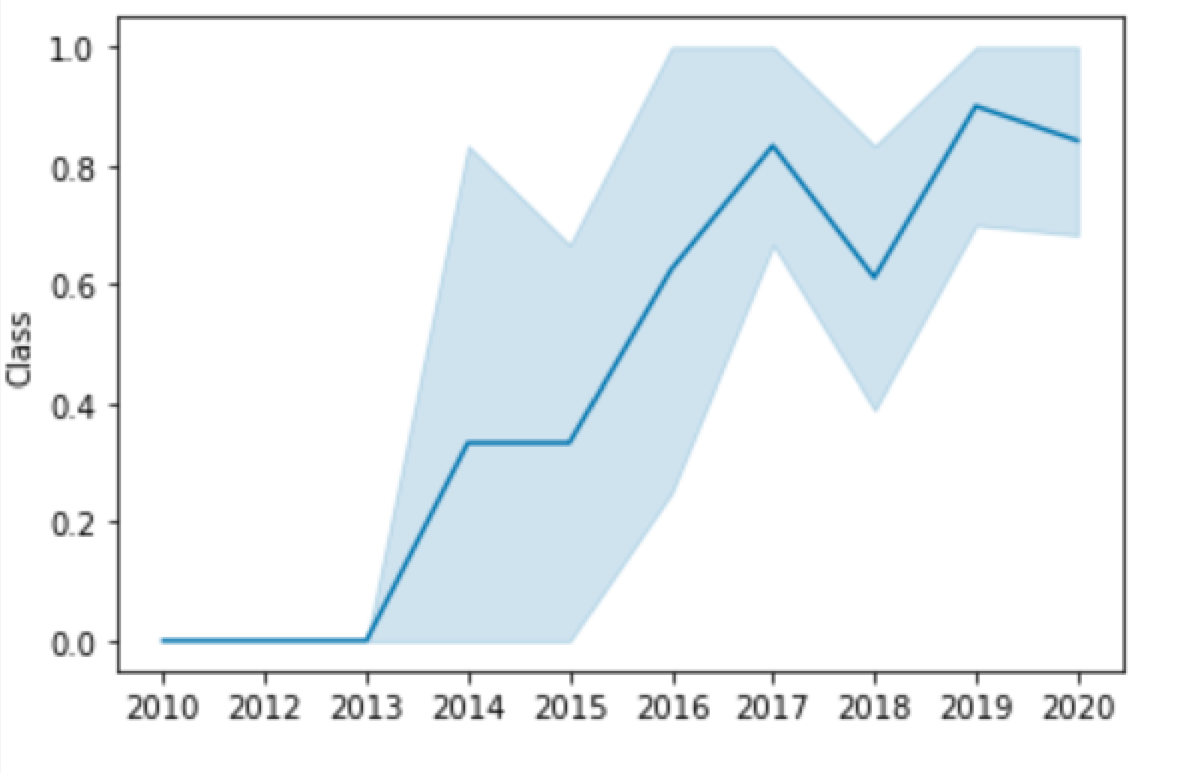
Payload vs. Orbit Type

- Each point represents a launch.
- If point is blue landing failed
- If point is yellow landing was successful.
- Plot represents relationship between Payload mass and orbit type.
- We can see:
  - When payload mass increases number of flights decreases.
  - Most orbits have less than 6000kg payload mass on average.

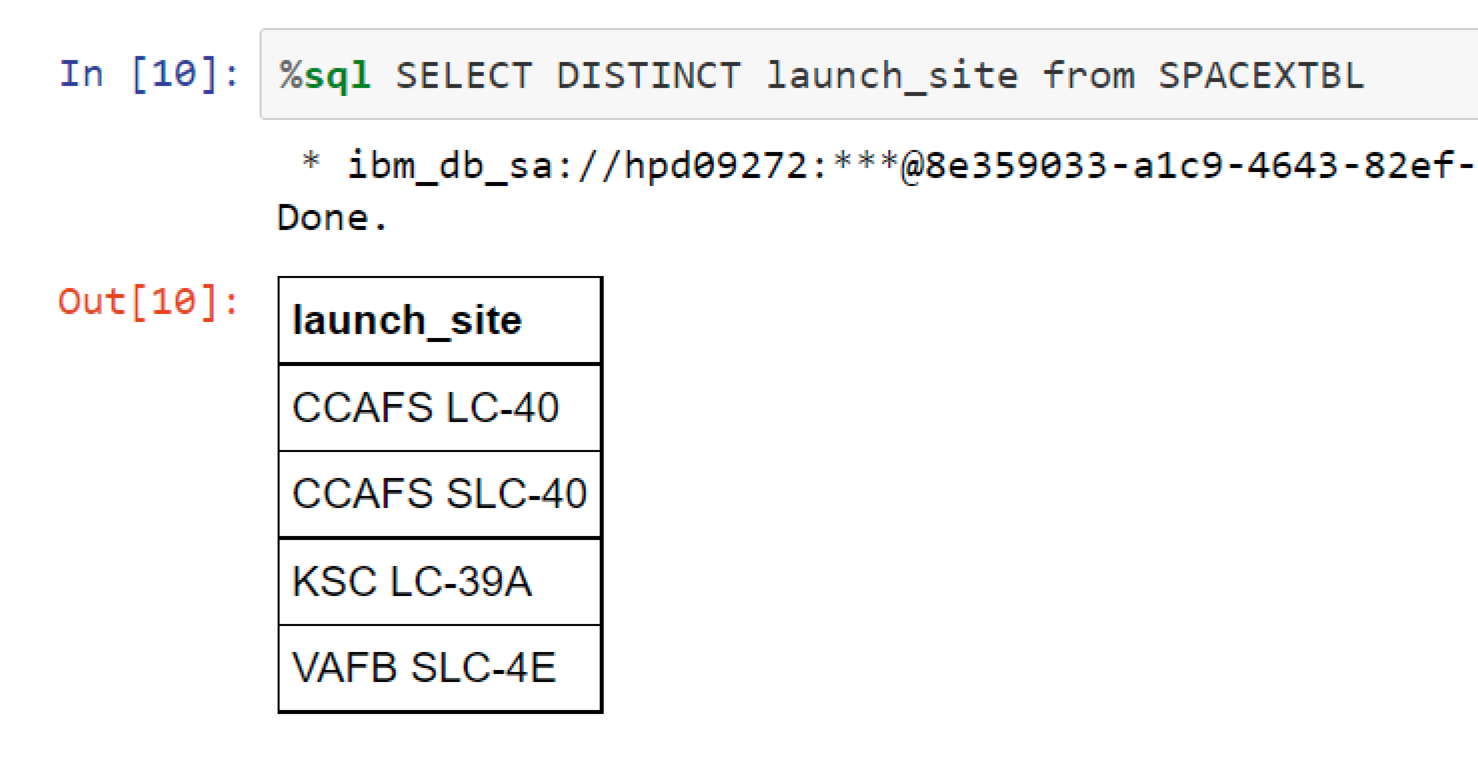


Launch Success Yearly Trend

- Line represents the ratio of successful landing yearly.
- As we can see, success rate is increasing every year.



All Launch Site Names



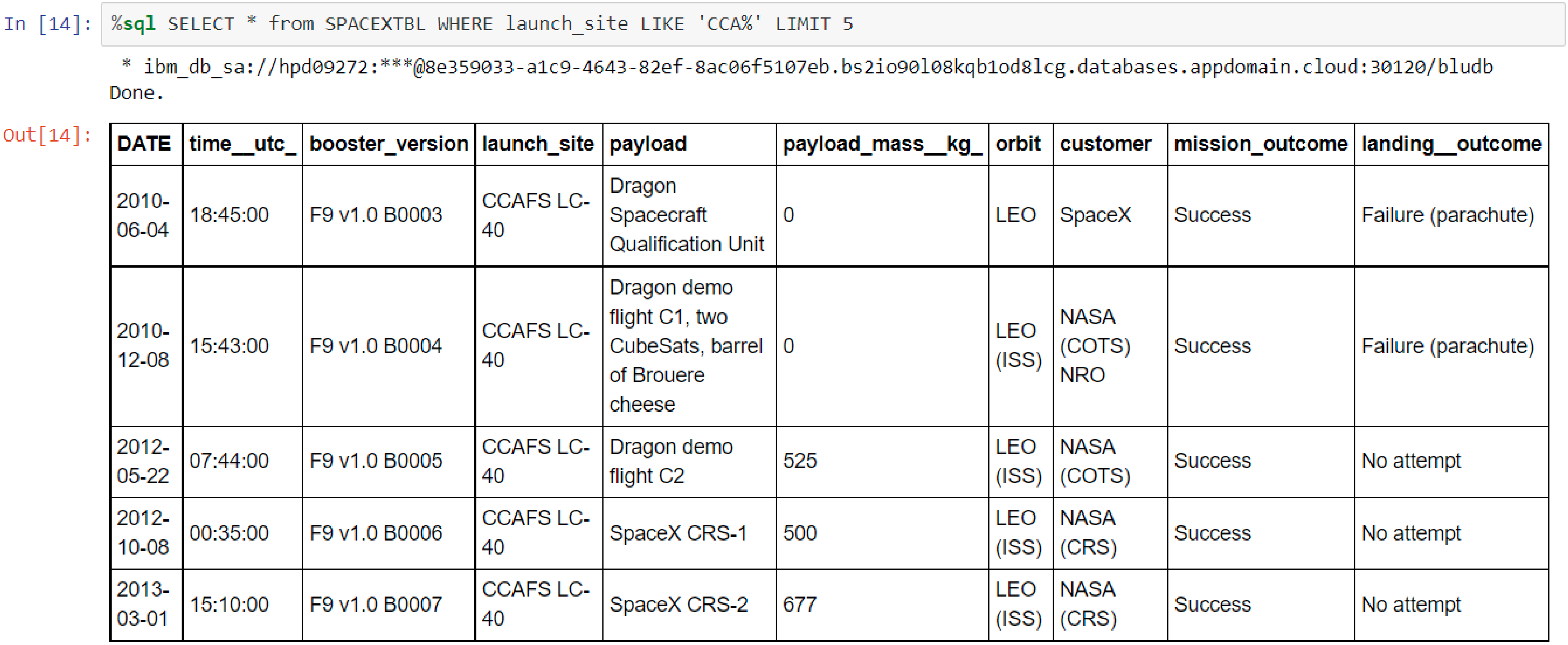
```
In [10]: %sql SELECT DISTINCT launch_site from SPACEXTBL

 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-
Done.
```

Out[10]:

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

- I selected every unique launch site.

Launch Site Names Begin with 'CCA'



```
In [14]: %sql SELECT * from SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5
```

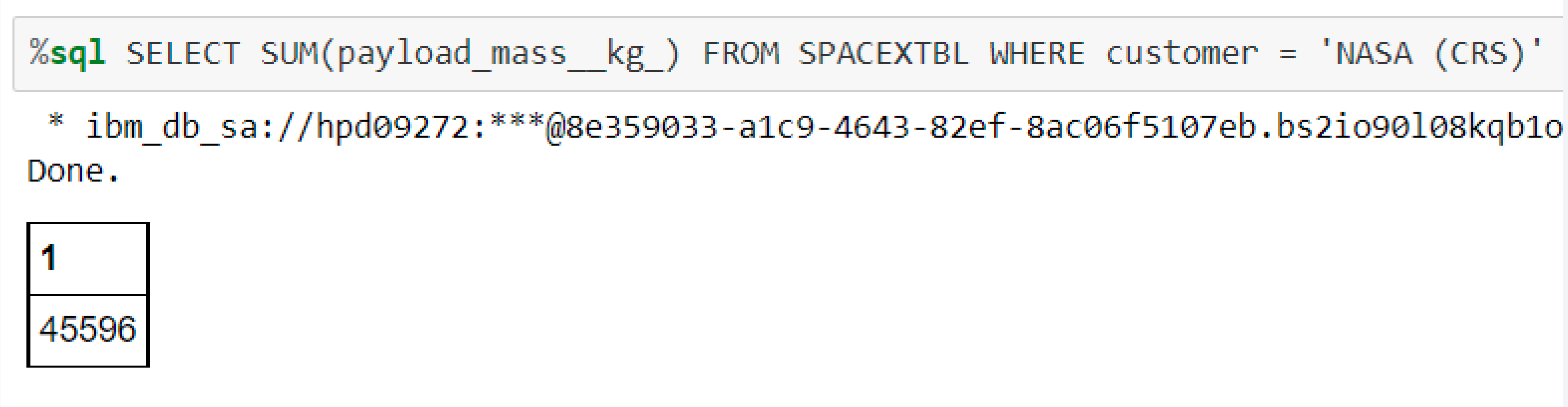
```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[14]:

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

- I showed the first 5 launch sites starting with CCA

Total Payload Mass



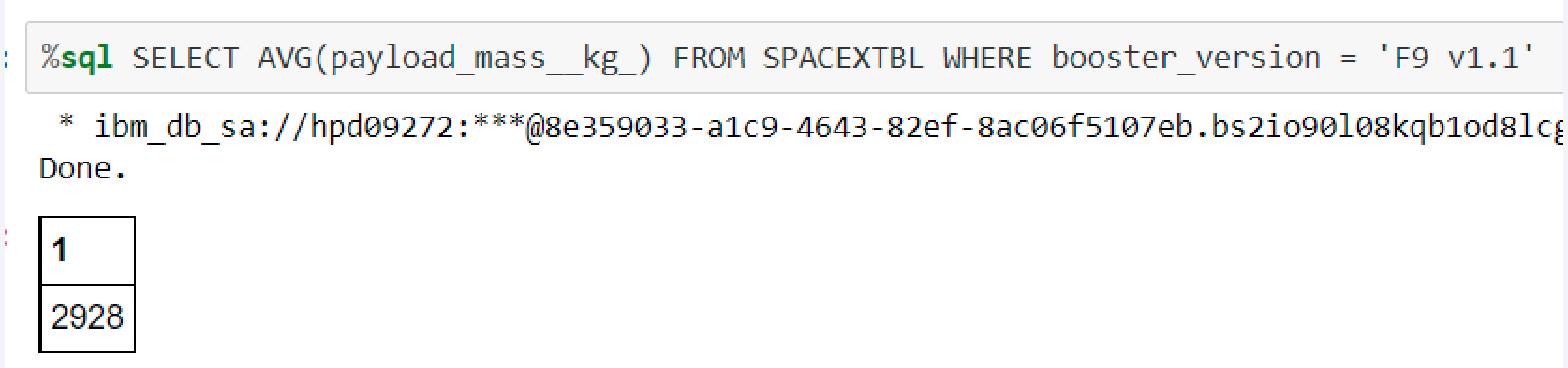
```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1o
Done.
```

1
45596

- I calculated the sum of payload carried by boosters from NASA

Average Payload Mass by F9 v1.1



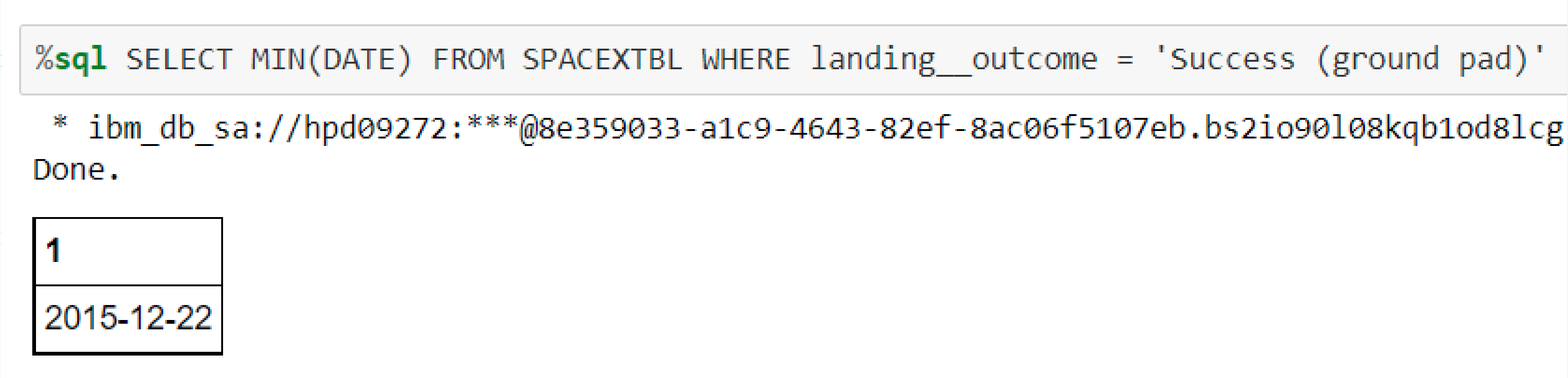
```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg
Done.
```

1
2928

- I showed the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date



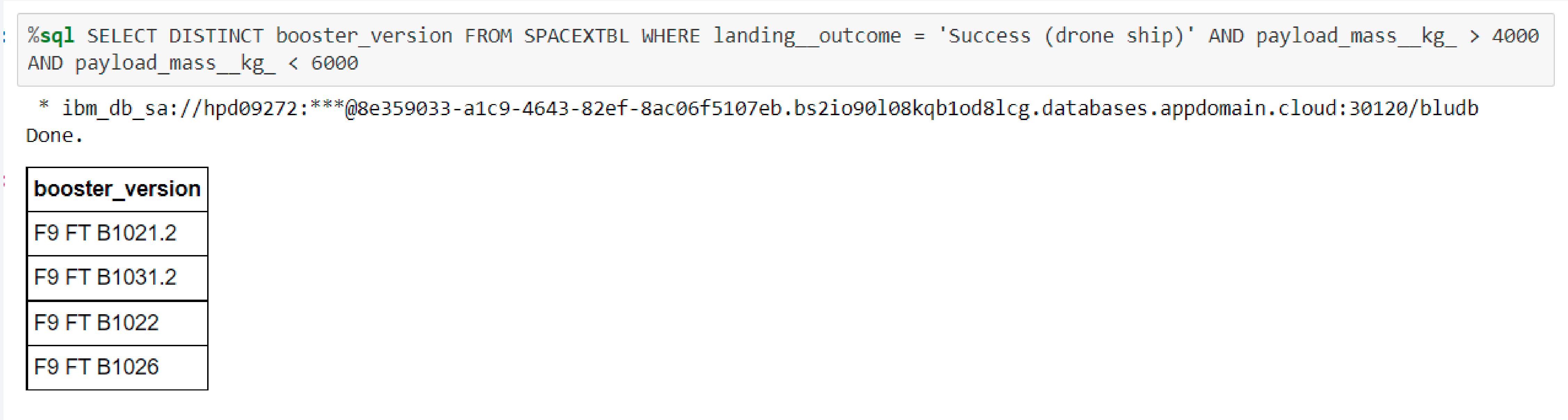
```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg
Done.
```

1
2015-12-22

- I showed the date of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000



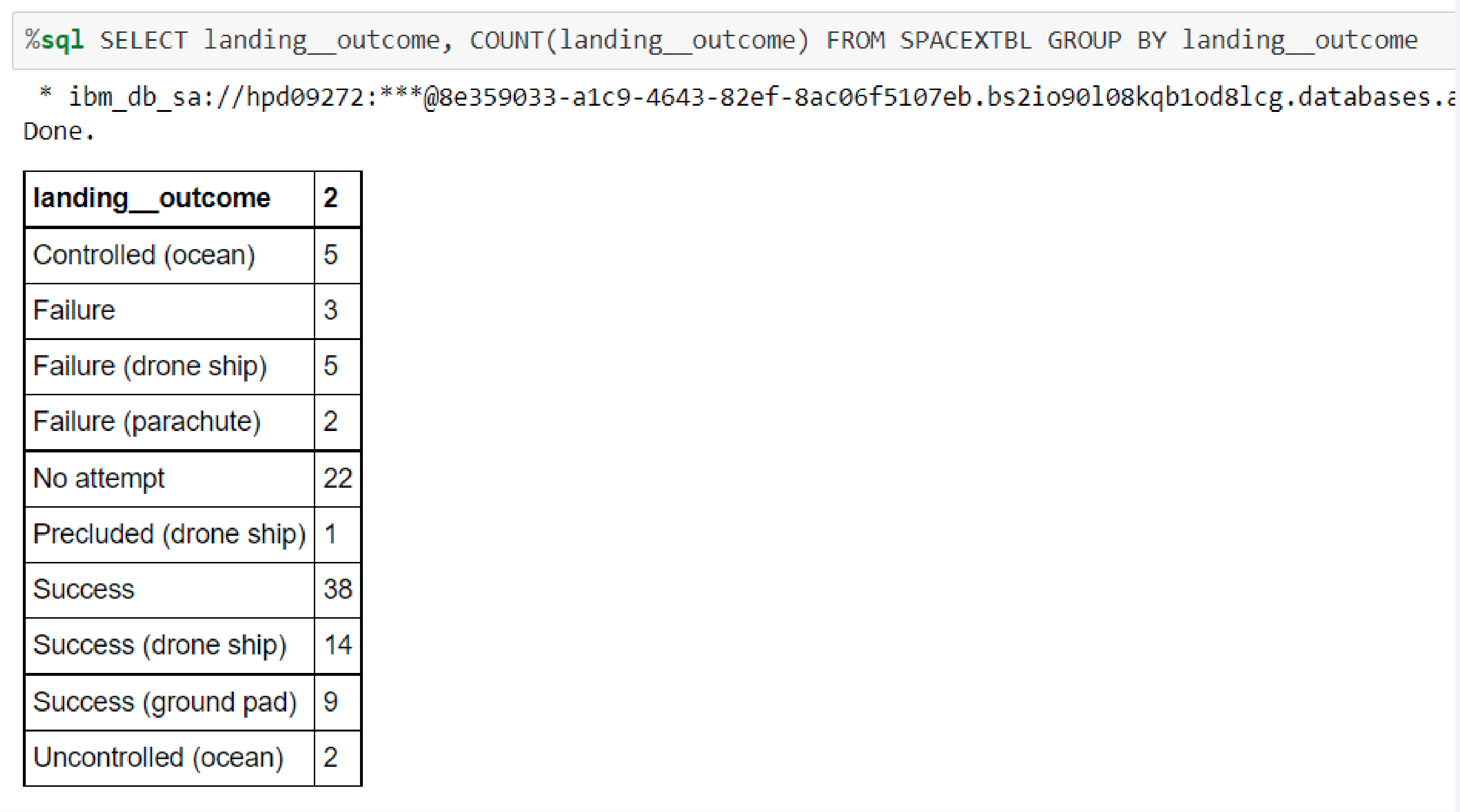
```
%sql SELECT DISTINCT booster_version FROM SPACEXTBL WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000
AND payload_mass__kg_ < 6000
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

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

- I showed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes



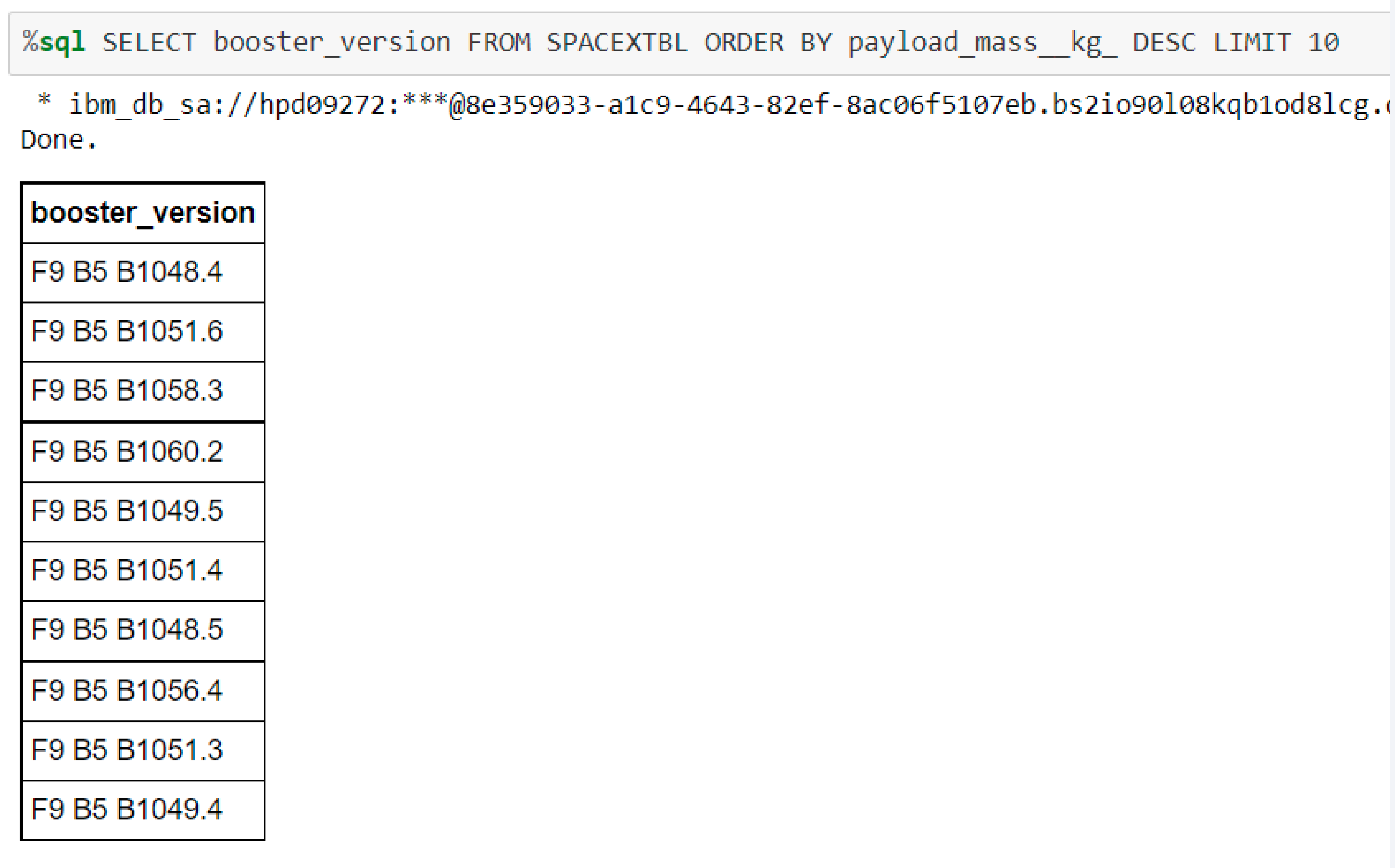
```
%sql SELECT landing__outcome, COUNT(landing__outcome) FROM SPACEXTBL GROUP BY landing__outcome
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.a
Done.
```

landing__outcome	2
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

- I showed the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload



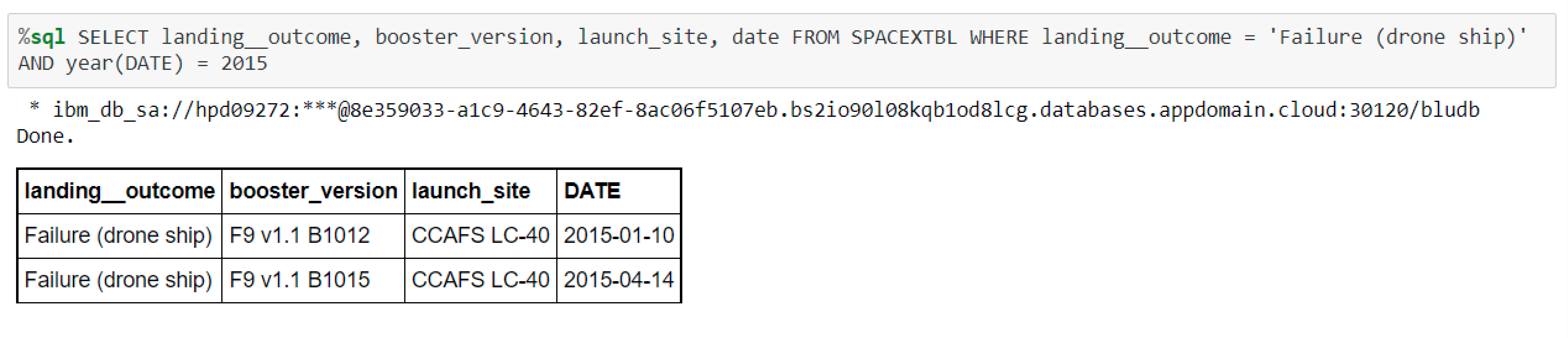
```
%sql SELECT booster_version FROM SPACEXTBL ORDER BY payload_mass__kg_ DESC LIMIT 10
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.
Done.
```

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

- Names of the booster which have carried the maximum payload mass

2015 Launch Records



```
%sql SELECT landing__outcome, booster_version, launch_site, date FROM SPACEXTBL WHERE landing__outcome = 'Failure (drone ship)'
AND year(DATE) = 2015
```

```
 * ibm_db_sa://hpd09272:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

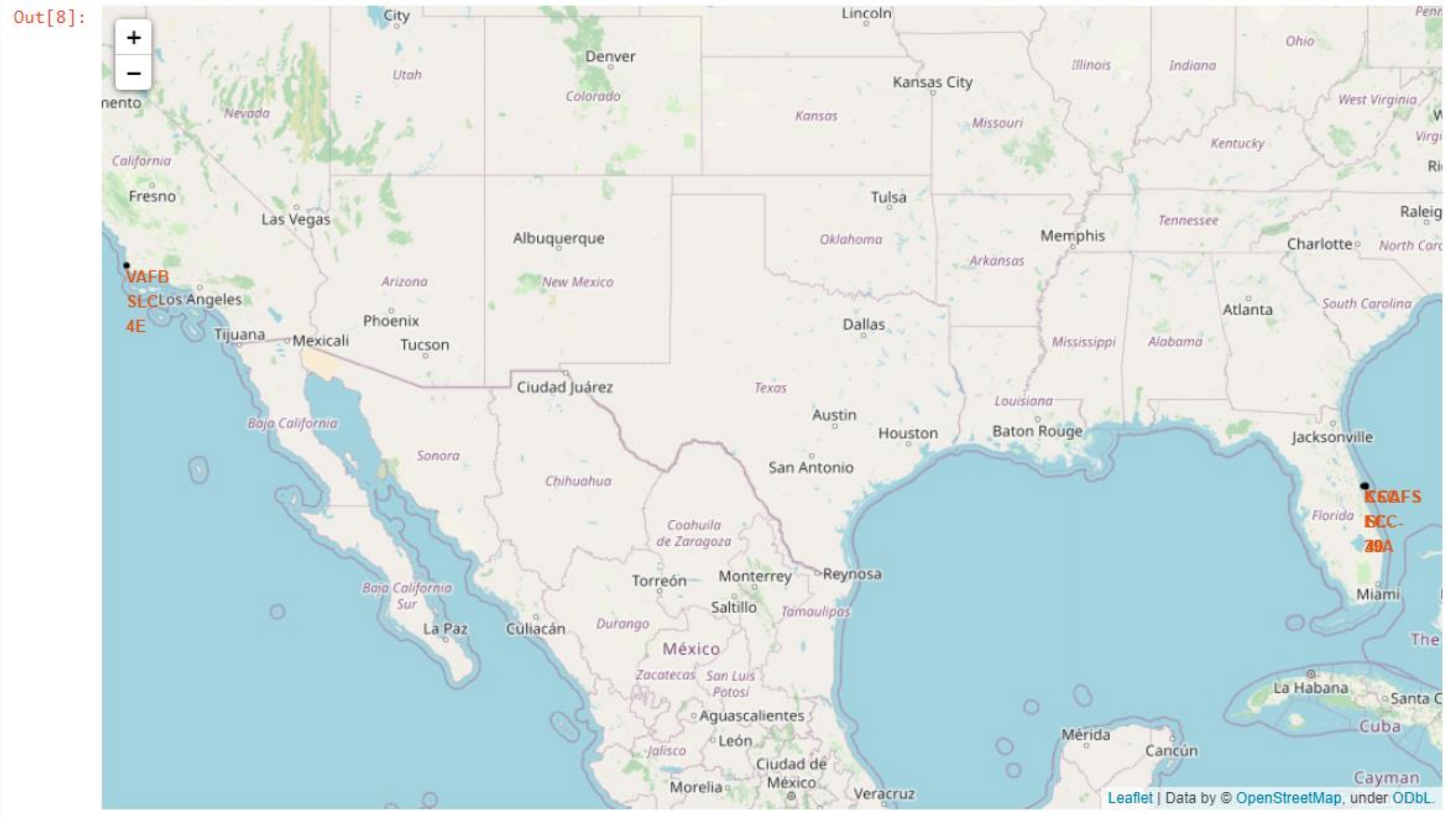
landing__outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

Section 4

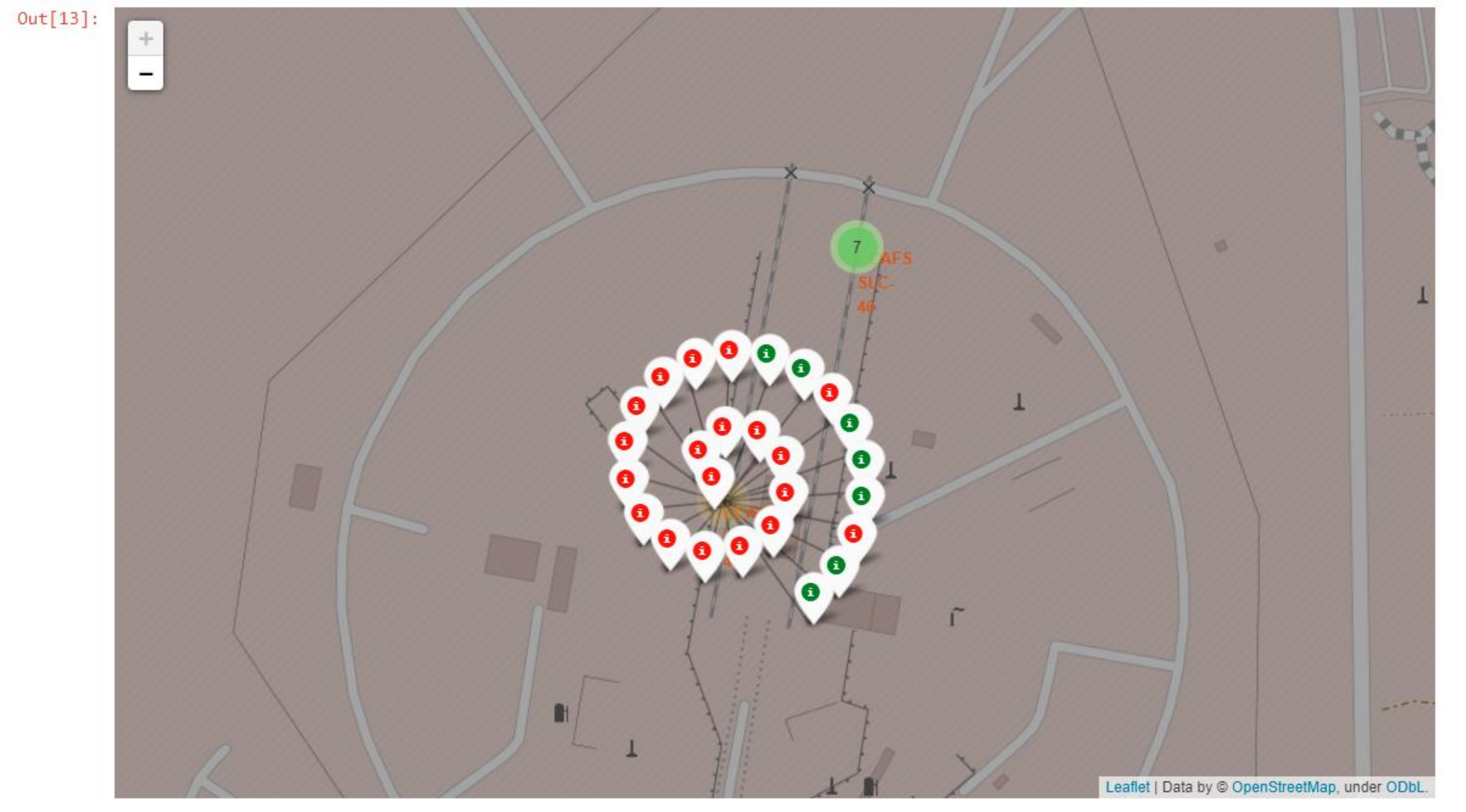
Launch Sites Proximities Analysis

All launch sites on a map



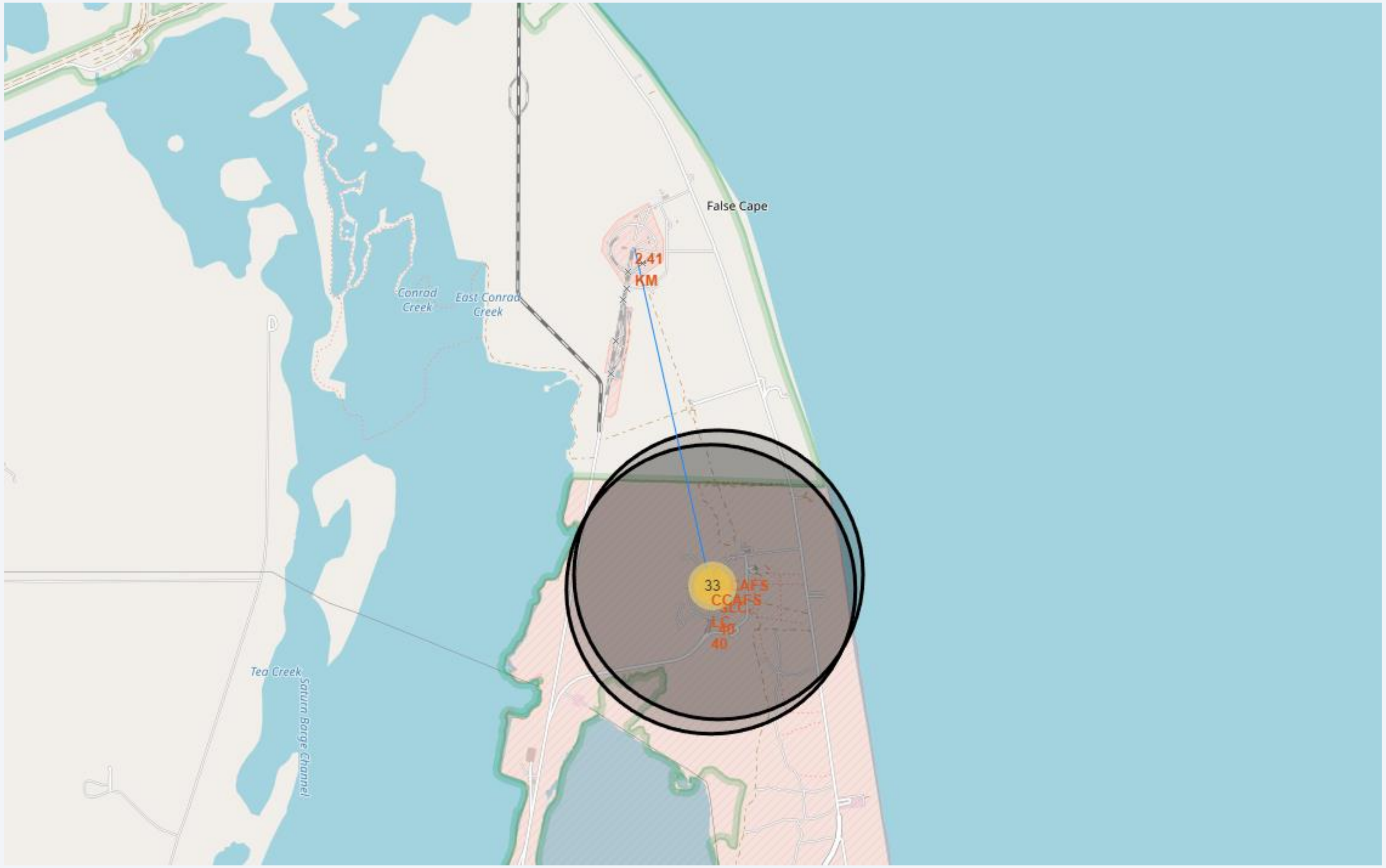
- All launches sites marked with a black circles and red letters.

Success/failed on map

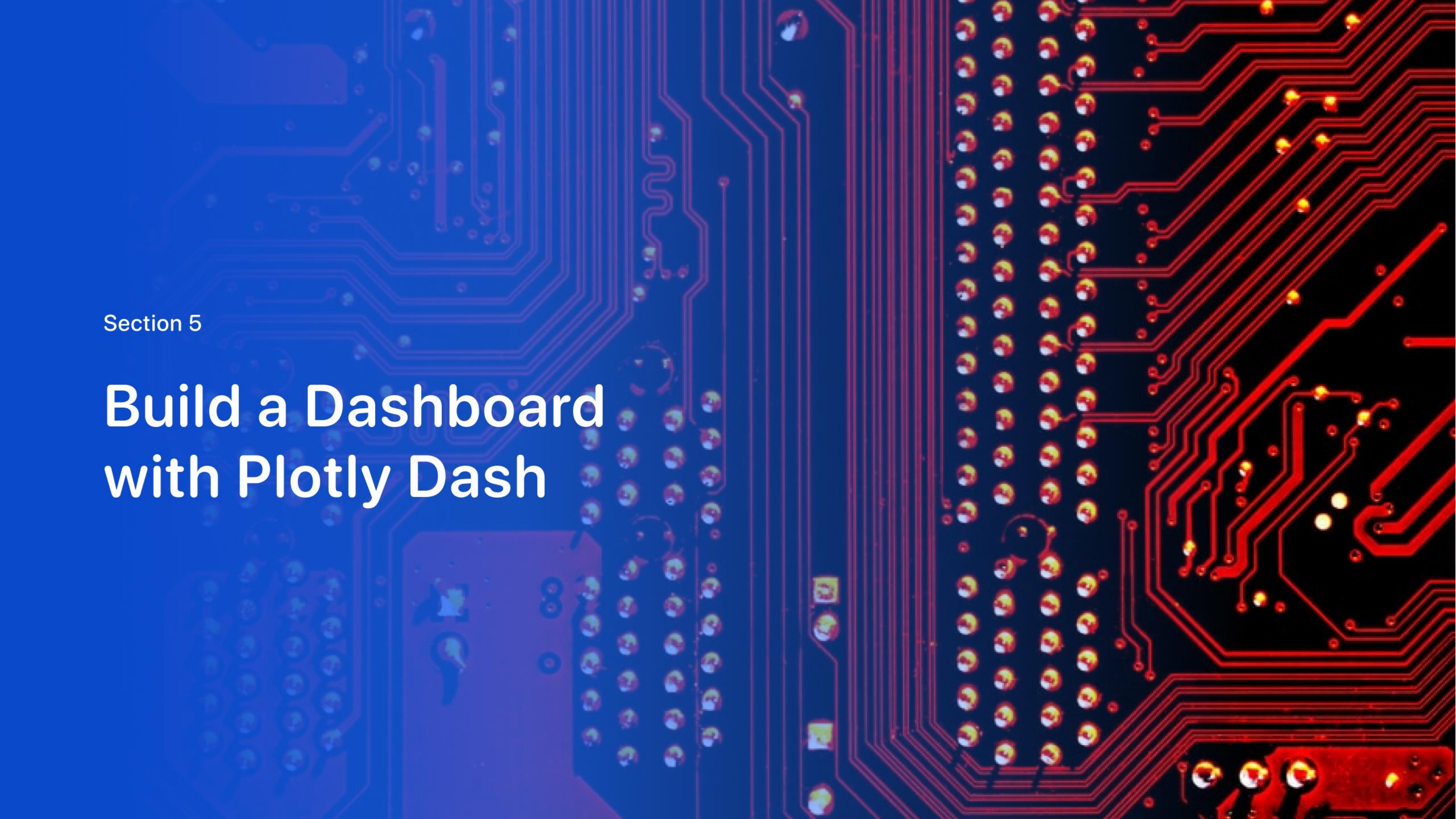


All launches marked of CCAFS SLC-40. Green are successful launches. Red are failed launches.

Launch site to highway



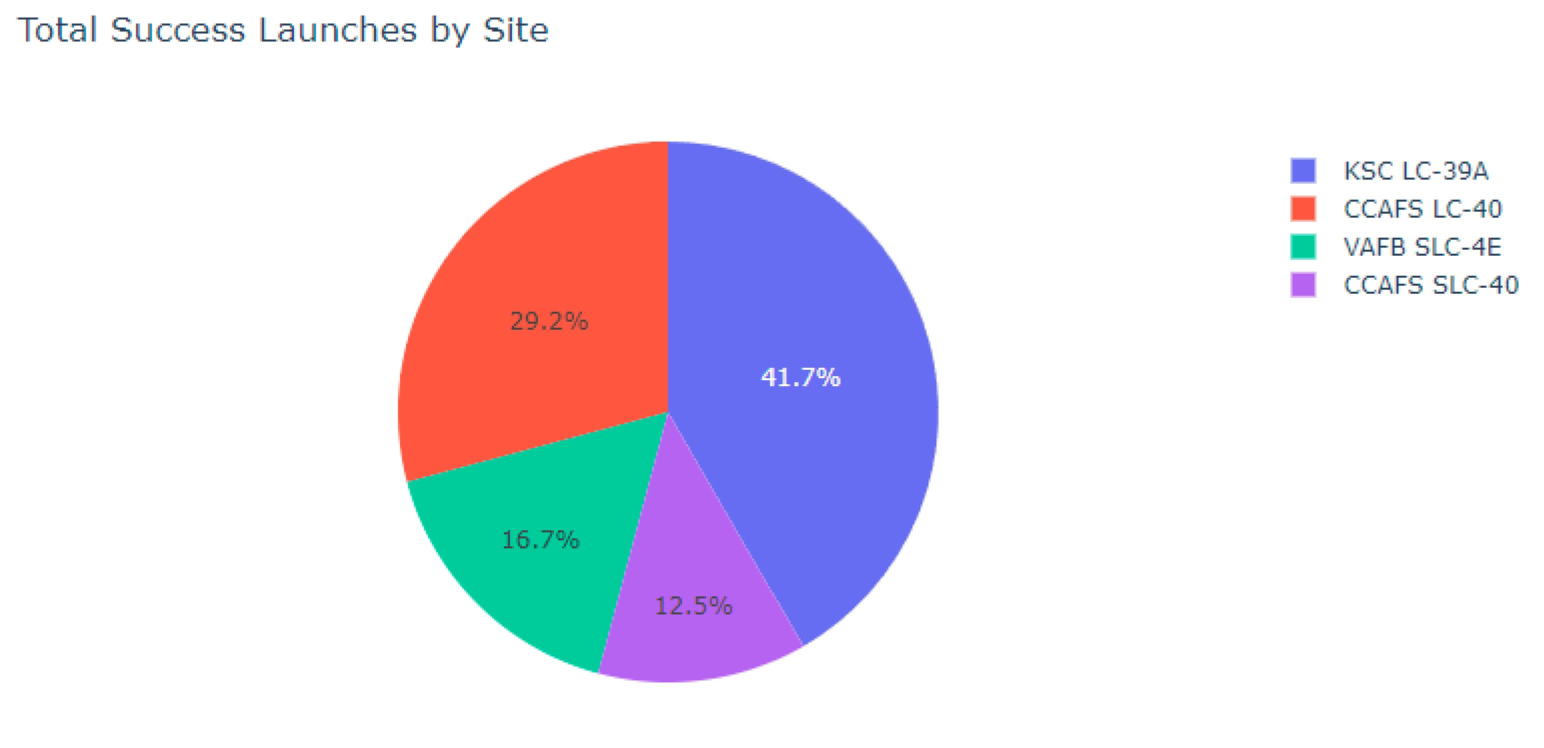
Distance between CCAFS SLC-40 launch site and the nearest highway.



Section 5

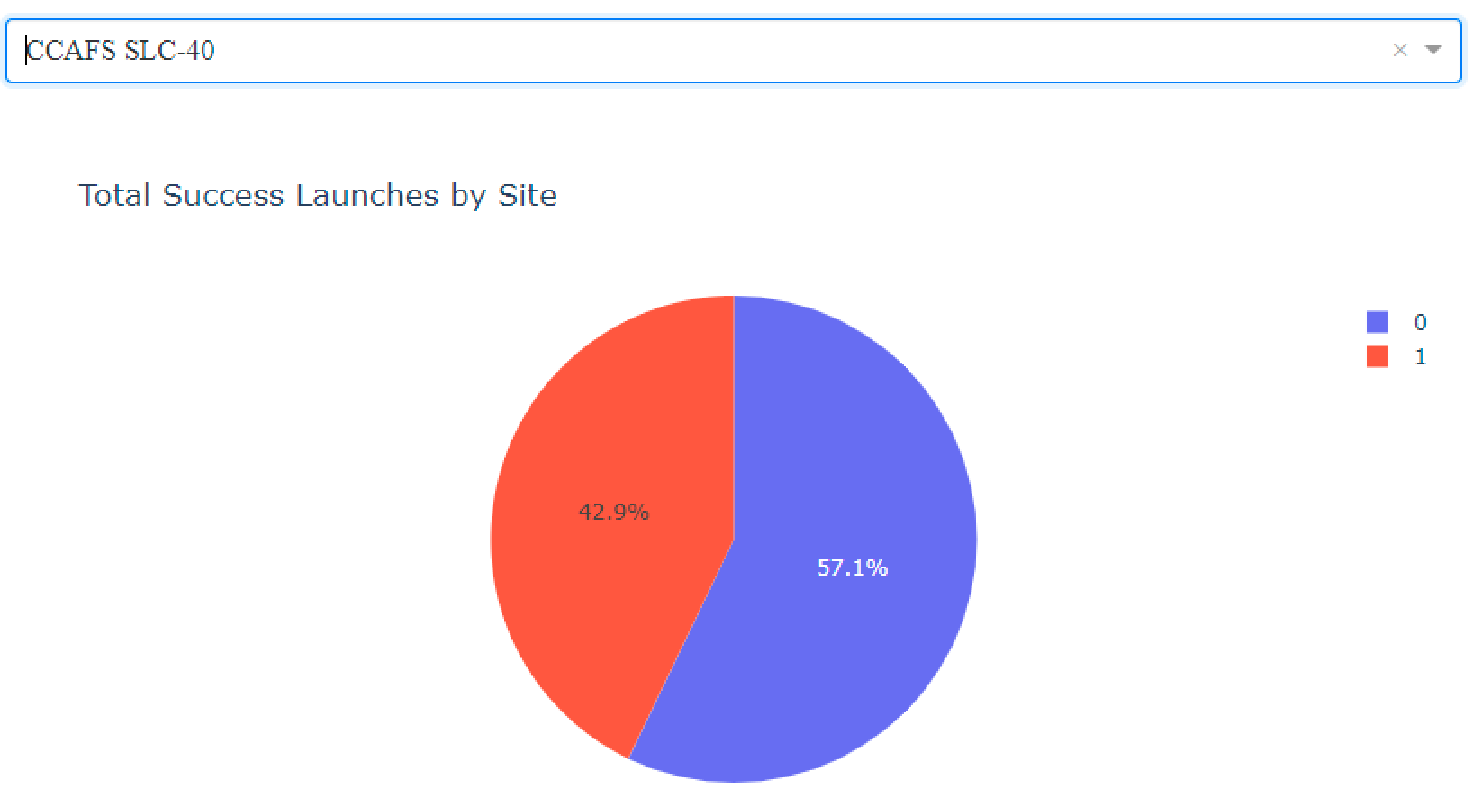
Build a Dashboard with Plotly Dash

Launch success for all launch sites



Each success percentage for each launch site.

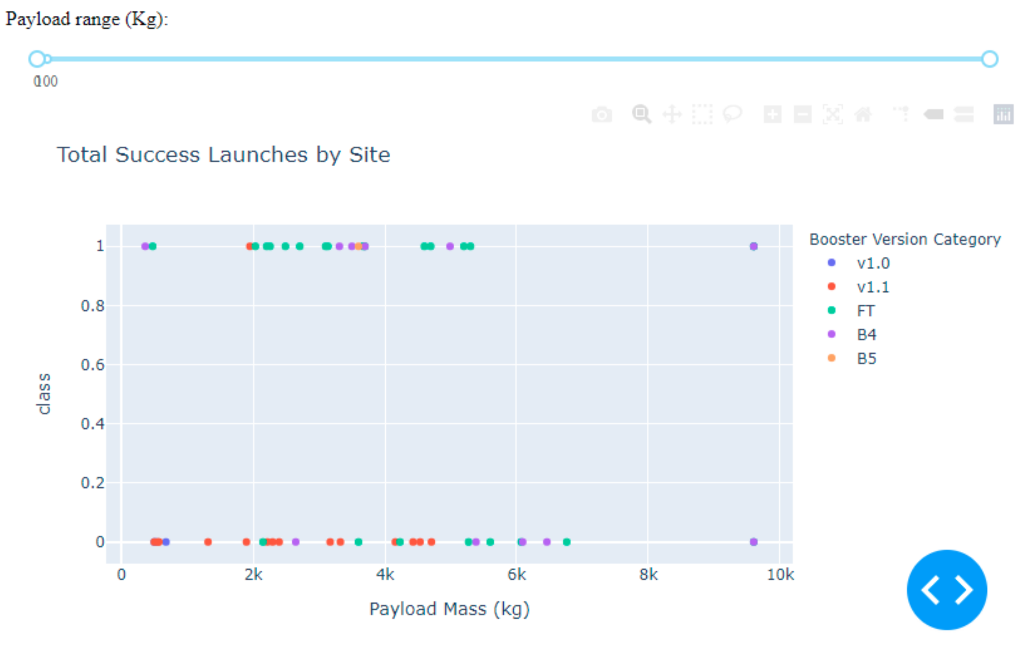
Success ratio for CCAFS SLC-40



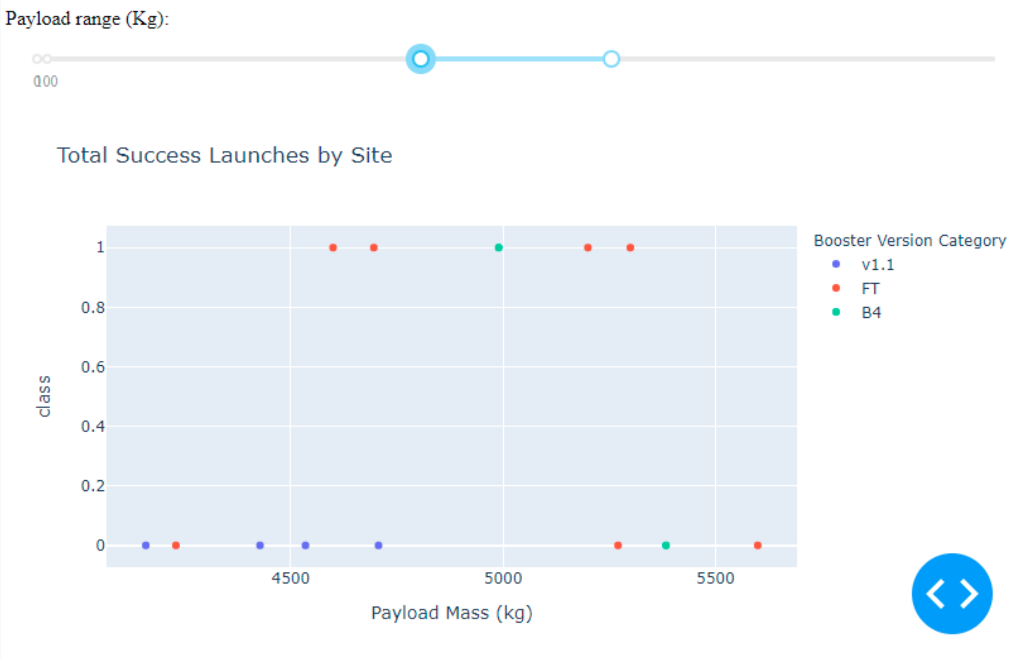
CCAFS SLC-40

Success percentage for CCAFS SLC-40, which was the launch site with higher success ratio (42.9%)

Payload vs Launch Outcome



Scatter plot showing relationship between payload and launch outcome for every launch site and all range of payload (kg)



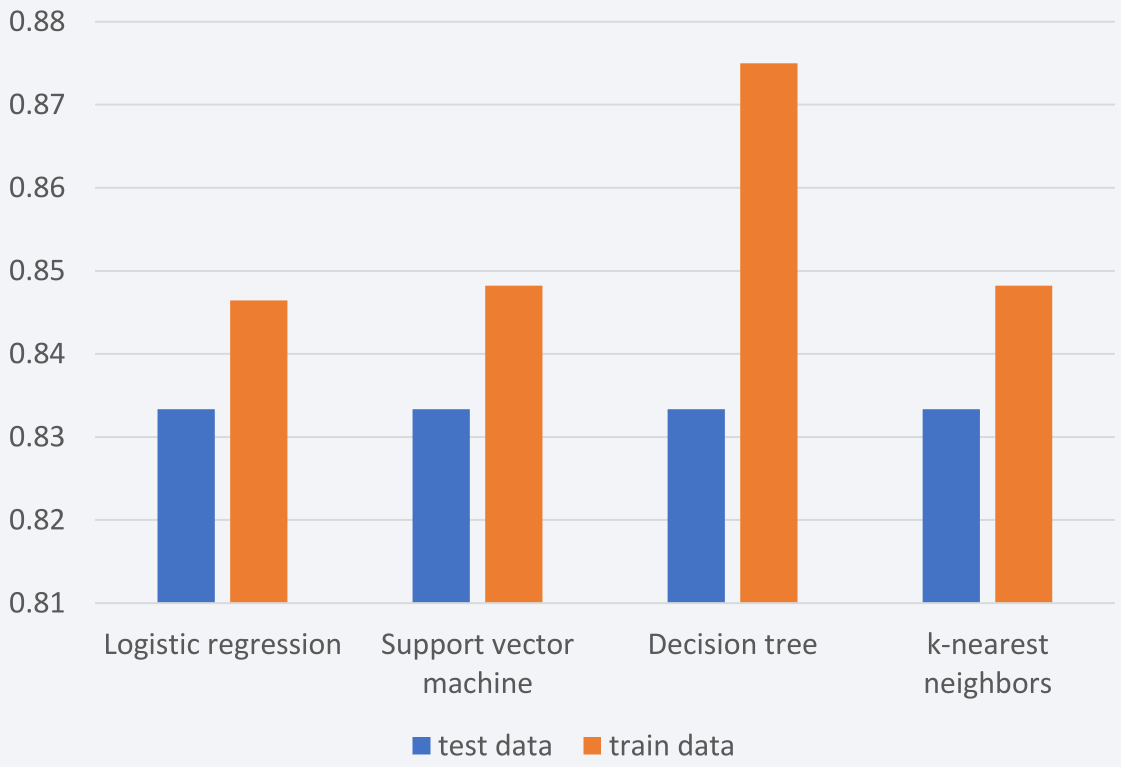
Same plot but constraint the range of payload (kg) between 4000kg and 6000kg. Different colors for different Booster versions.

Section 6

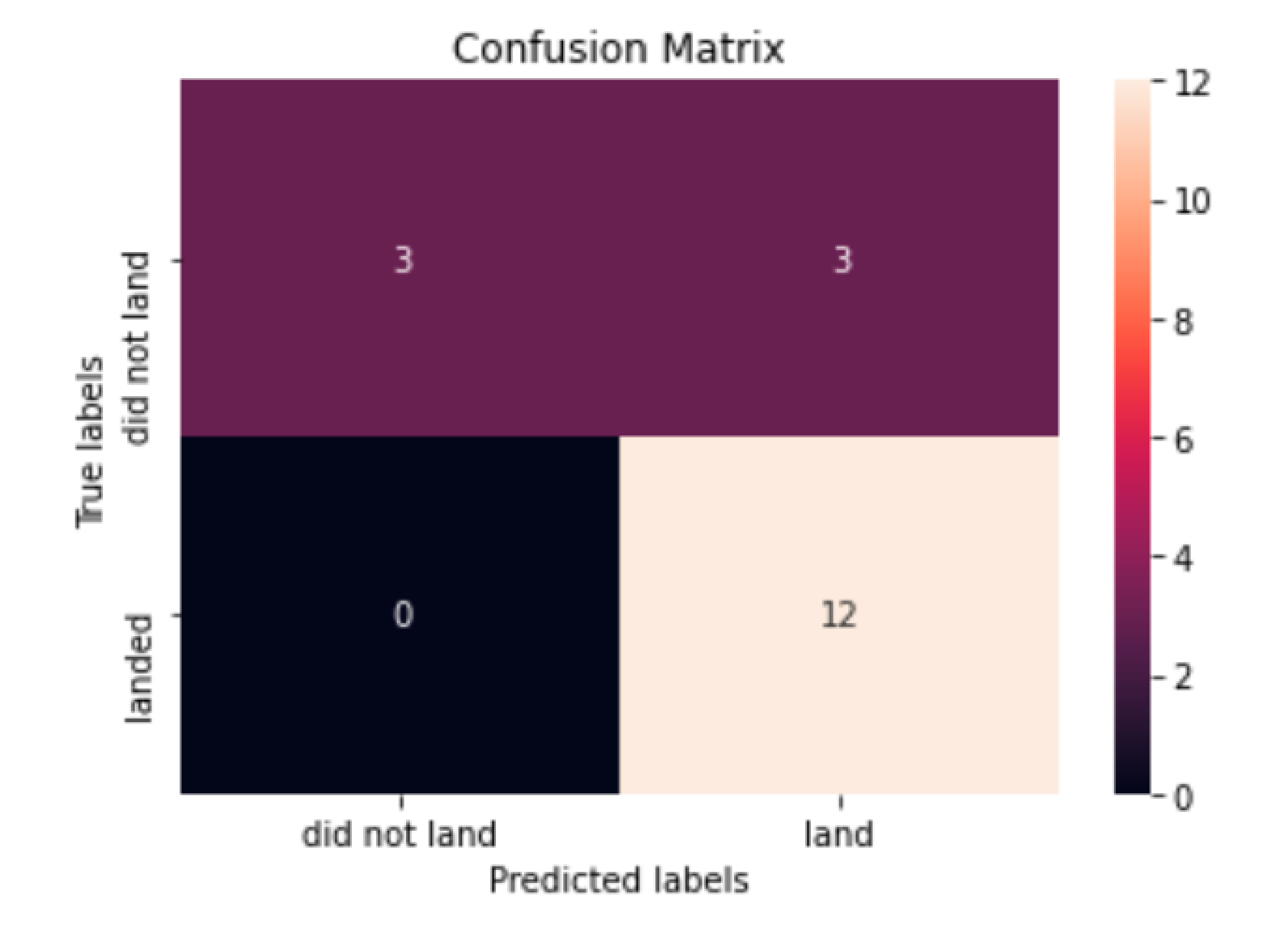
Predictive Analysis (Classification)

Classification Accuracy

- All models have the same accuracy with test data = 0.834.
- Decision tree has the higher accuracy with train data.



Confusion Matrix



Conclusions

- We have trained and obtained a model to predict if a launch would be successful or not depending in other variables.
- Now, we can optimize our launches to increase savings and make progress.
- With the work presented will be easier analyze future data and present it with interactive plots, maps and other methods here used.
- Use these data to improve our launches is the next step.

Thank you!