IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gustavo Cardoso
2023-01-15

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using API and Web Scraping and Data Wrangling;
  - Exploratory Data Analysis (EDA) using SQL, Pandas and Matplotlib;
  - Data visualization with Interactive Visual Analytics and Dashboard;
  - Predictive Analysis with Machine Learning;
- Summary of all results
  - Successful data collection from API and web Scraping;
  - Analysis of most important features of dataset;
  - Several machine learning models trained for best prediction;

Introduction

- Project background and context

  SpaceX is one of the most successful rocket companies by managing to launch at an relatively low price. The aim of this project is to predict the cost of each launch of a new company, Space Y, using data of SpaceX launches.

- Problems you want to find answers
  - Determine the price of each launch;
  - Find the most important features to predict this price and the successful rate;
  - Find possible relations between factors;

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collected from Space X API: https://api.spacexdata.com/v4/launches/past and through web scrapping Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - It was created a column representing the landing outcome to improve the data;
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was divided in training and testing data and evaluated by different classifications models.

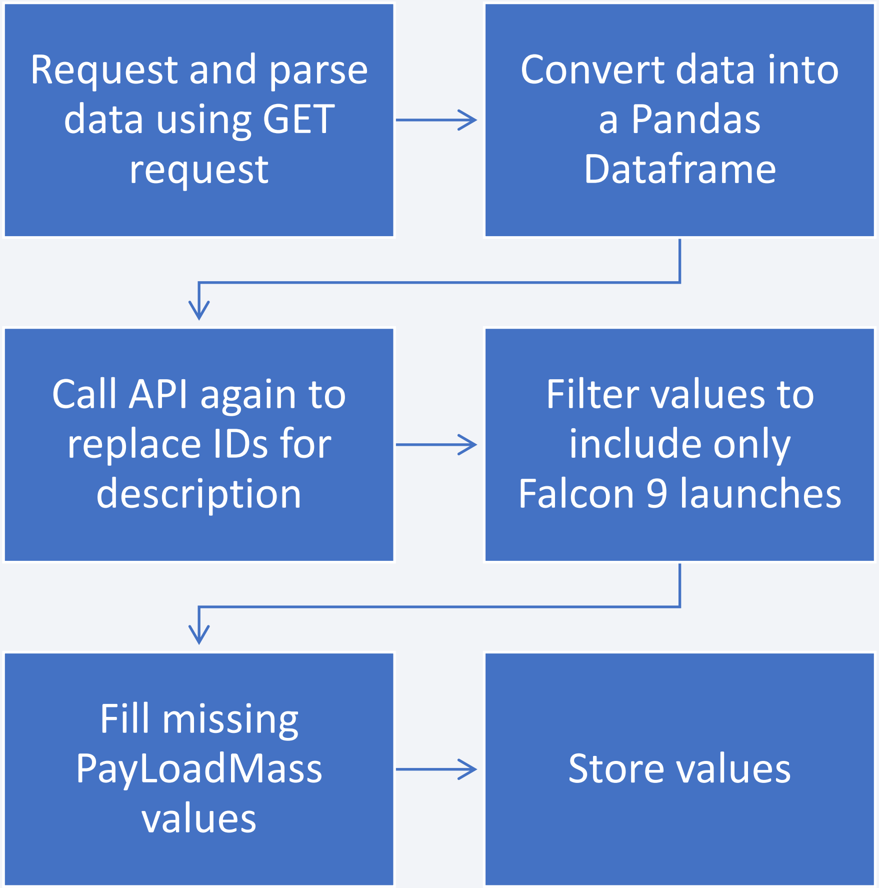
Data Collection

- Data collected from Space X API: https://api.spacexdata.com/v4/launches/past. It was also necessary call other endpoints of the SpaceX API to achieve a better understanding.
- In addition, to improve the dataset, it was applied a web scrapping process in the following Wikipedia page: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Data was collected, converted, cleaned and filtered as indicated in the flowchart beside.

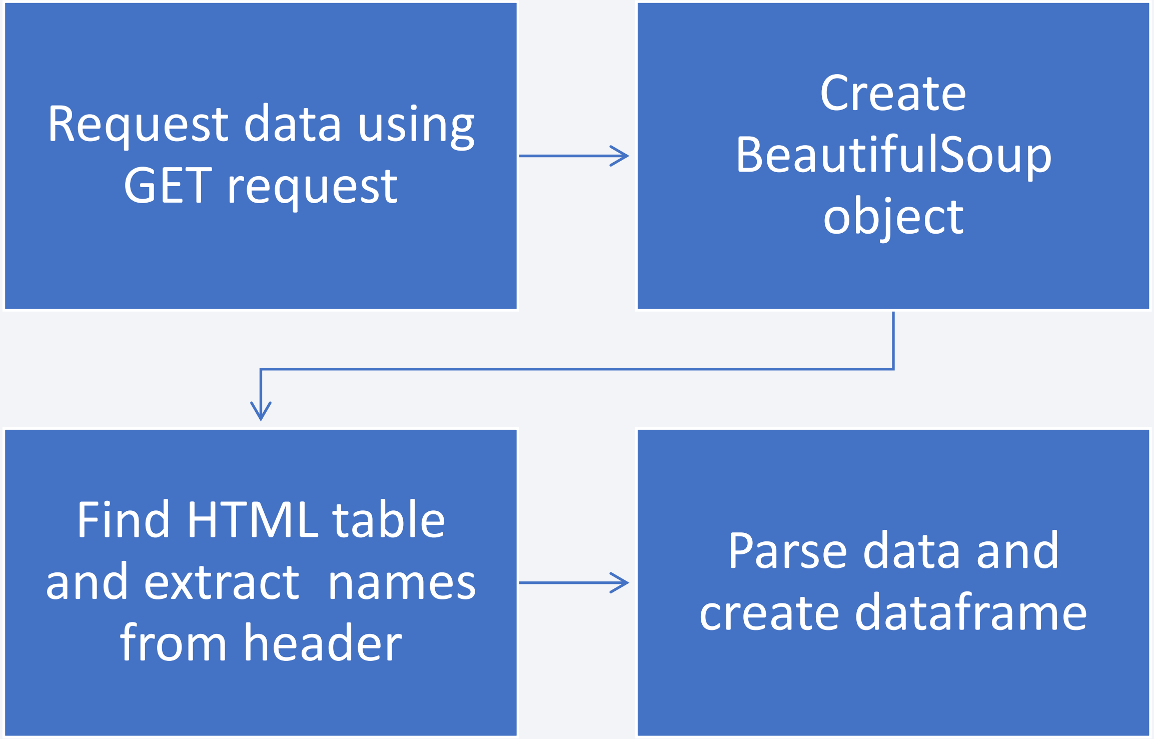
- GitHub URL: [Data Collection SpaceX API]


Request and parse data using GET request
Convert data into a Pandas Dataframe
Call API again to replace IDs for description
Filter values to include only Falcon 9 launches
Fill missing PayLoadMass values
Store values

Data Collection - Scraping

- The Wikipedia page of the Falcon 9 launches was used to obtain data through web scraping

- GitHub URL: [Data Collection with Web Scrapping]()



Request data using GET request

Create BeautifulSoup object

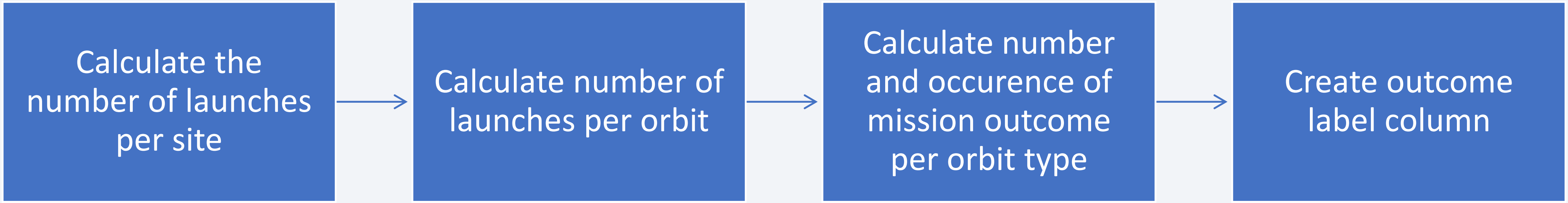
Find HTML table and extract names from header

Parse data and create dataframe

Data Wrangling

- In this stage, an initial exploratory data analysis (EDA) was performed. The main objective was summarize the data and convert the multiples possible outcomes into successful or unsuccessful landing.

- GitHub URL: Data Wrangling



Calculate the number of launches per site

Calculate number of launches per orbit

Calculate number and occurence of mission outcome per orbit type

Create outcome label column

EDA with Data Visualization

- Scatter, bar and line charts were plotted for better understanding of the relations between launch site, payload mass, success rate, orbit type and yearly trend.
- Scatter plots are better for understanding correlation between two variables, while line charts are better for time series and bar charts for comparison.

- GitHub URL: EDA with Data Visualization

EDA with SQL

- After loading the dataset into a DB2 database instance, the following queries were performed:
  - Names of unique launch sites;
  - Display 5 records where launch sites begin with ‘CCA’;
  - Total payload mass carried by boosters lunched by NASA (CRS);
  - Average payload mass carried by booster version F9 v1.1;
  - Date of the first successful landing outcome in ground pad;
  - Boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
  - Total number of successful and failure mission outcomes;
  - Booster versions which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, with booster versions, launch sites names in the year 2015;
  - Rank of the count of landing outcomes between 2010-06-04 and 2017-03-2;
- GitHub URL: EDA with SQL

Build an Interactive Map with Folium

- All launch sites were marked in a folium map with a marker and a circle around it;
- For each launch site, a marker cluster was created with a marker for each of their launch records, using colors to distinguish successful and unsuccessful ones;
- Lines were drawn to indicate the distance of the launch sites and its proximities, like closest city, railway or highway;

- GitHub URL: Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- Next, a interactive dashboard containing two charts were created using Plotly Dash:
    - A pie chart showing the total successful launches per launch site;
    - A scatter plot showing the relation between the outcome and the payload mass
- In addition, a dropdown containing all launch sites was created to filter data in both charts as well as a range slider to control the range of the payload mass of the scatter plot.

- GitHub URL: SpaceX Dash App.py

Predictive Analysis (Classification)

- To perform predictive analysis, the data was transformed and split in training and testing data sets;
- Four different models were used: Support Vector Machine, Classification Tree, Logistic Regression and K-Nearest Neighbor;
- For each of them were performed procedures to find the best hyperparameter.
- To evaluate each model, the chosen metrics were: F1-Score, Accuracy and Jaccard Score;

- GitHub URL: Machine Learning Prediction lab.ipynb

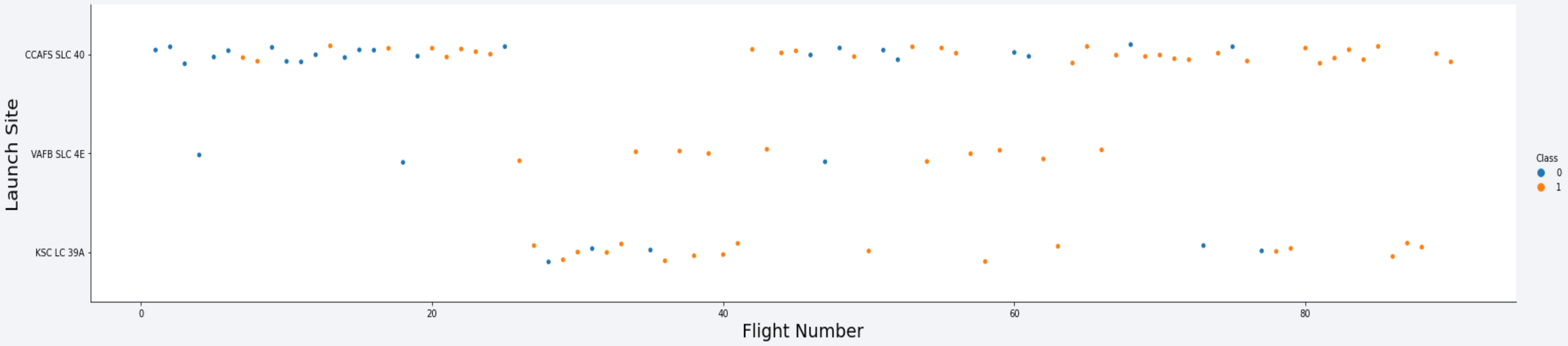
Results

- In the following sections, the following topics it will be discussed:
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

Section 2

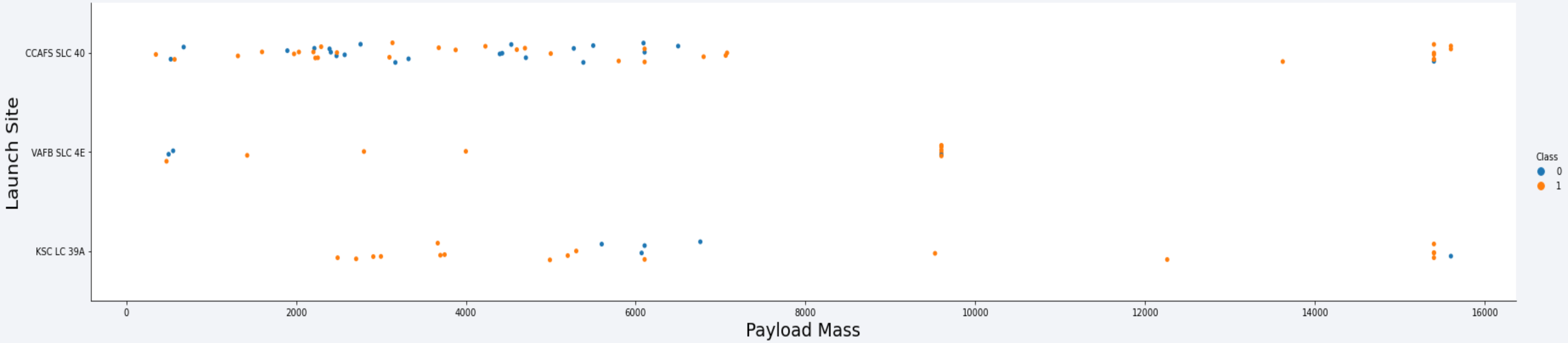
Insights drawn from EDA

Flight Number vs. Launch Site



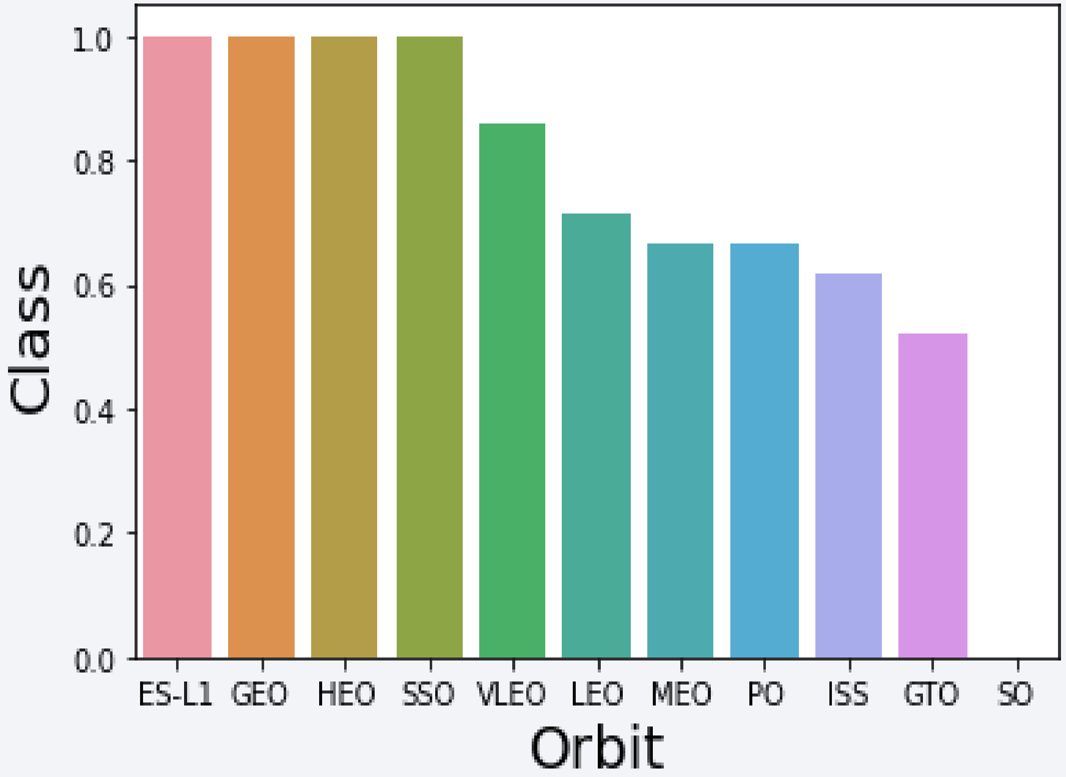
- The CCAFS SLC 40 site has most of the launches while VAFB SLC 4E has the fewest;
- The overall success rate greatly improved over the time;
- From the last 20 launches, only 3 were unsuccessful, meaning a 85% success rate. With 5 successful launches out of 7 in KSC LC 39A and 12 out of 13 in CCAFS SLC 40;

Payload vs. Launch Site



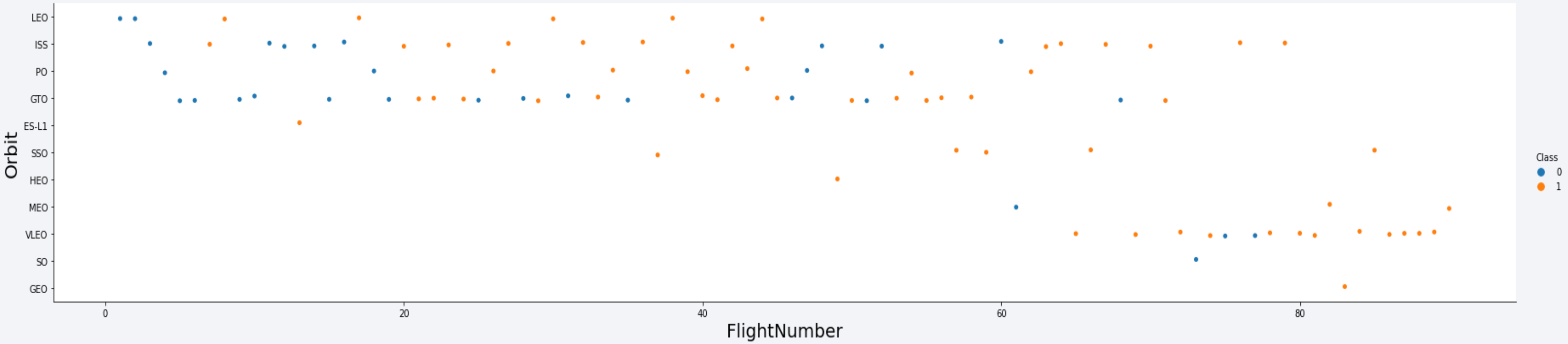
- In general, higher payload mass results in better success rate;
- VAFB SLC 4E had no launches with over 10000kg;
- KSC LC 39A had 100% success rate under 5000kg;

Success Rate vs. Orbit Type



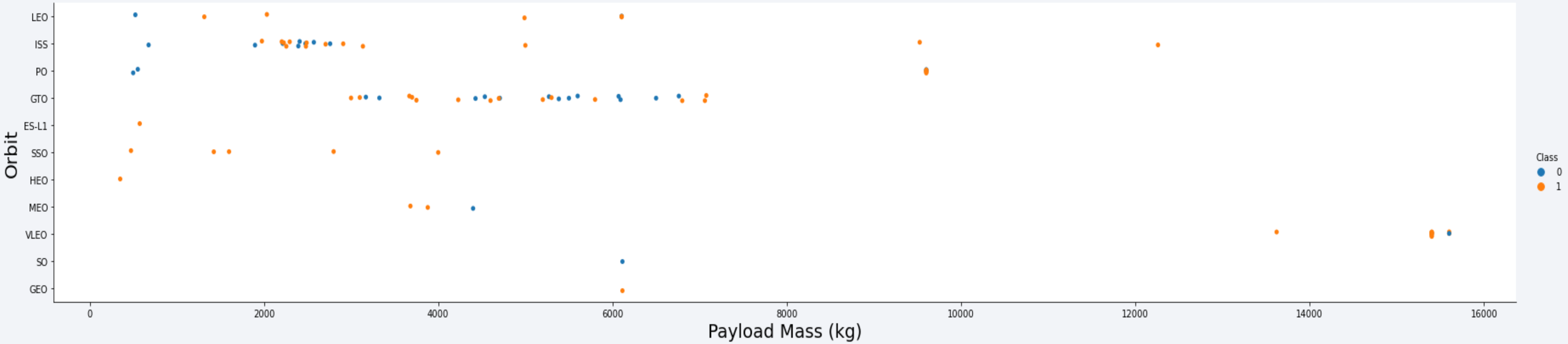
- ES-L1, GEO, HEO and SSO were the most successful with 100% success rate;
- SO had 0% success rate;

Flight Number vs. Orbit Type



- LEO orbit appears to have a relation with number of flights;
- ES-L1, HEO, SO and GEO had one launch each, so it is not posible to determine if there was a relatioship between between flight number and orbit in these cases;

Payload vs. Orbit Type

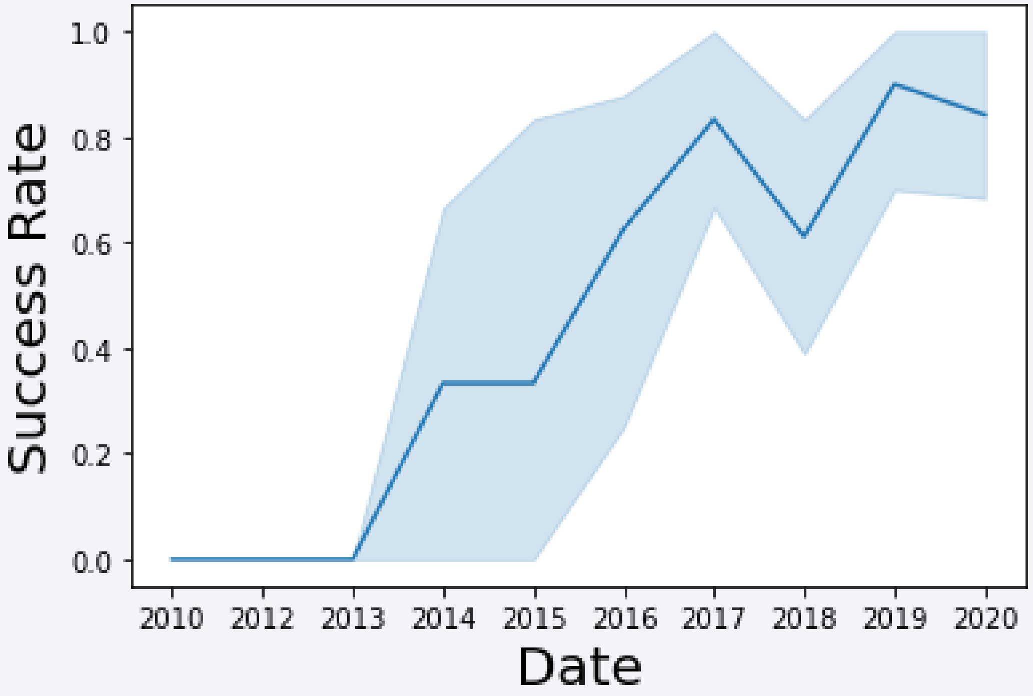


- Payload mass had a positive relation with LEO, PO and ISS orbits;
- On the other hand, payload mass had a negative relation with GTO orbit;

Launch Success Yearly Trend

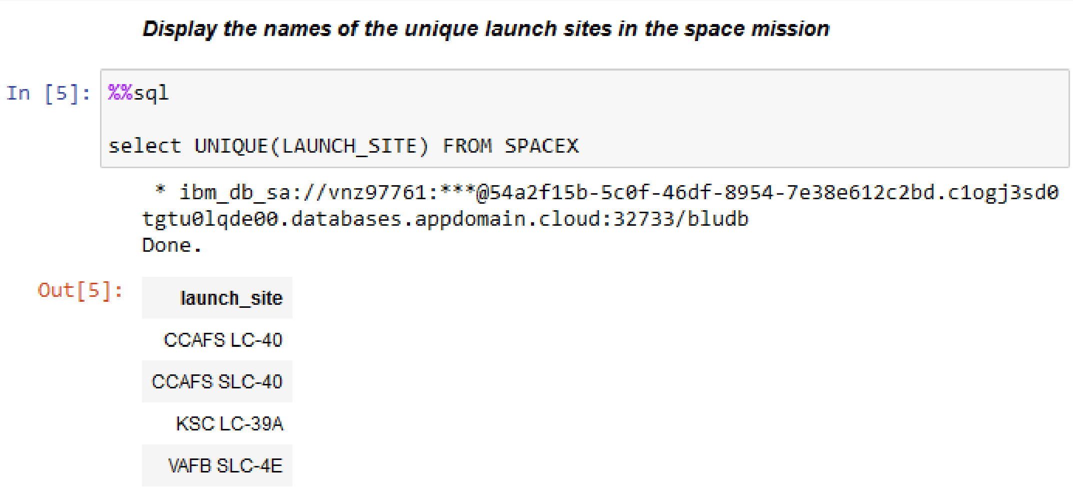
- The first three years had no improvement in the success rate;

- From 2013 to 2020 the success rate greatly improved;



All Launch Site Names

- It was possible to get the names of launch sites using the UNIQUE SQL constraint;



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

```
In [5]: %%sql

        select UNIQUE(LAUNCH_SITE) FROM SPACEX
```

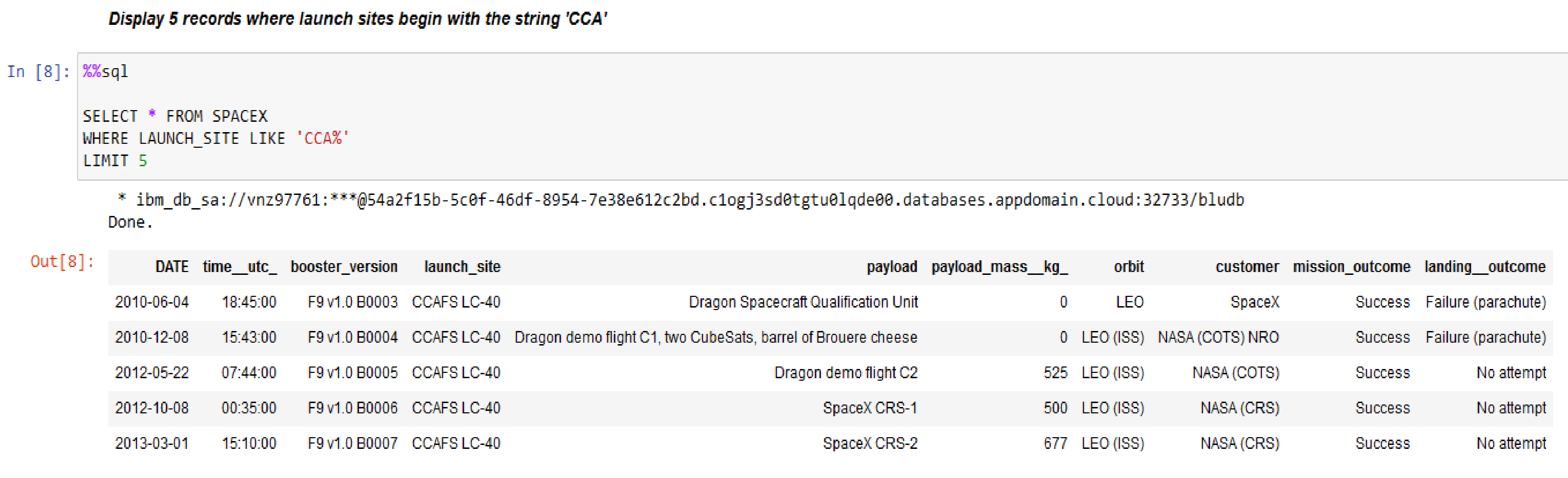
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0
tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[5]:

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

Launch Site Names Begin with 'CCA'

- The following query was used to find 5 records where launch sites begin with `CCA`, using the constraint LIMIT with the WHERE command:



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

```
In [8]: %%sql

        SELECT * FROM SPACEX
        WHERE LAUNCH_SITE LIKE 'CCA%'
        LIMIT 5
```

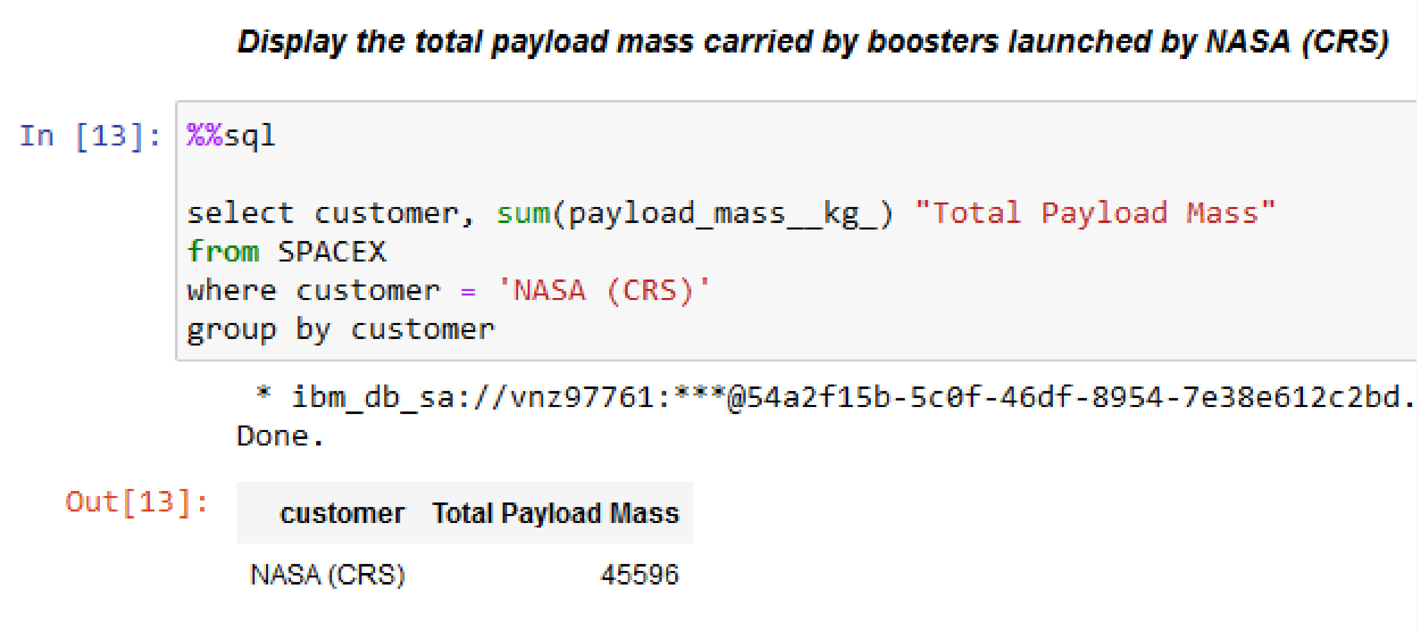
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[8]:

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

- To calculate the total payload mass carried by boosters launched by NASA, it was necessary to use the SUM function and GROUP BY operator:



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

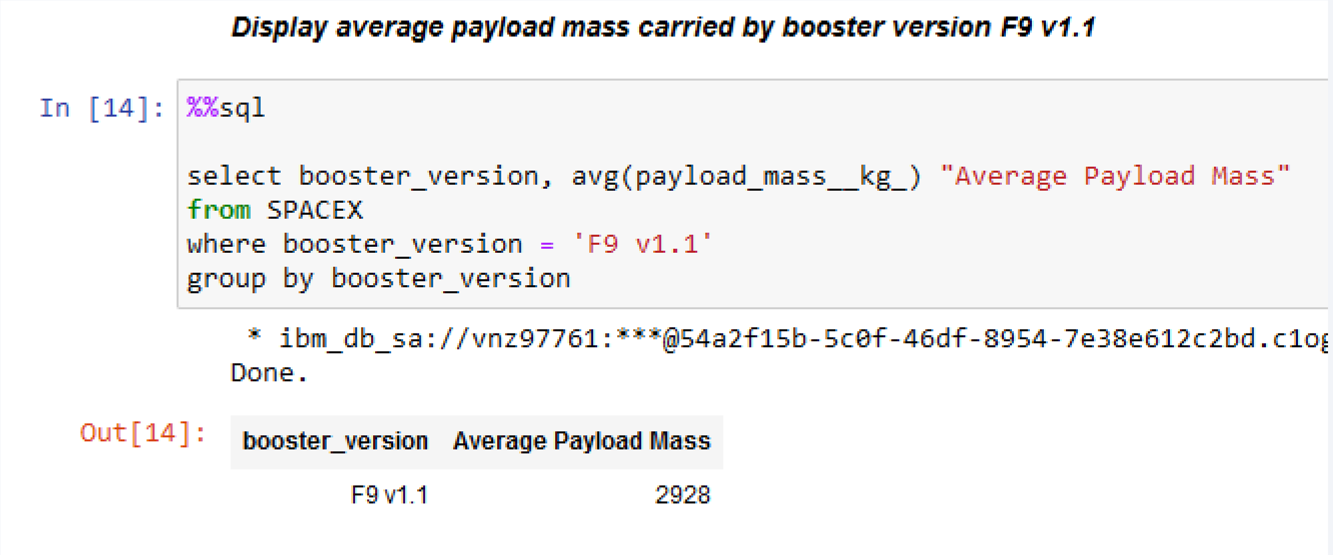
```
In [13]: %%sql

         select customer, sum(payload_mass__kg_) "Total Payload Mass"
         from SPACEX
         where customer = 'NASA (CRS)'
         group by customer

 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.
Done.
```

Out[13]:

customer	Total Payload Mass
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The average payload mass caried by booster version F9 v1.1 was calculated using the AVG function combined with WHERE clause



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

In [14]:
```
%%sql

select booster_version, avg(payload_mass__kg_) "Average Payload Mass"
from SPACEX
where booster_version = 'F9 v1.1'
group by booster_version
```

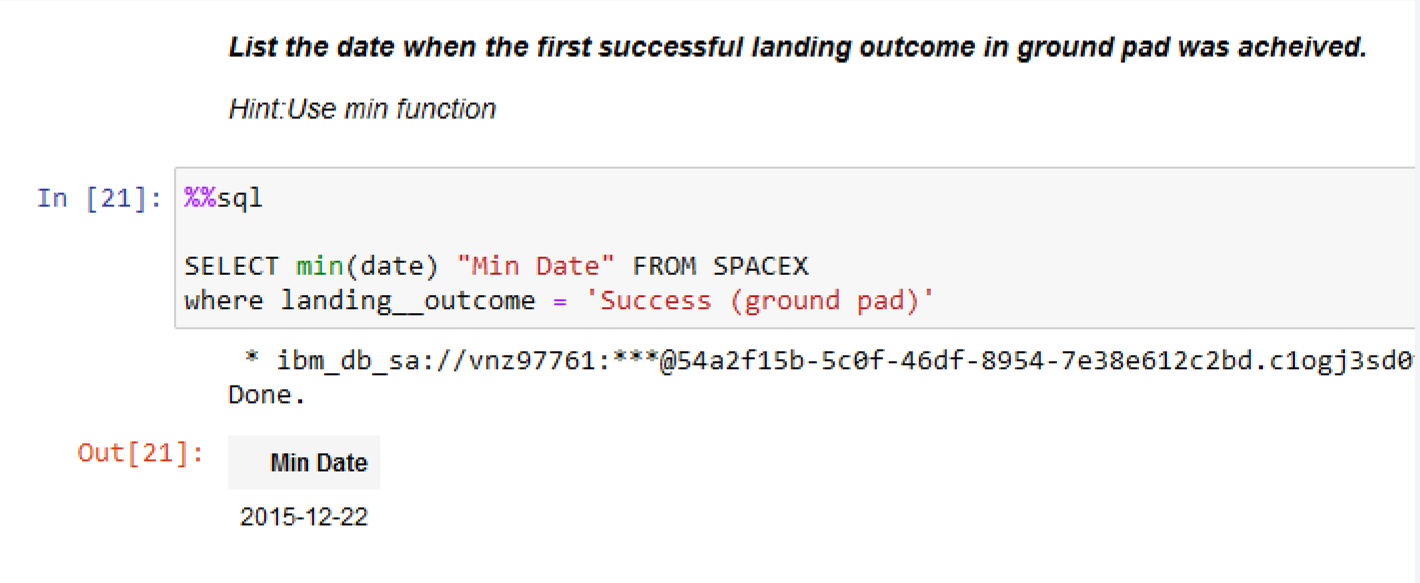
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1og
Done.
```

Out[14]:

booster_version	Average Payload Mass
F9 v1.1	2928

First Successful Ground Landing Date

- The first successful landing outcome in ground was achieved in 2015-12-22. To find this date, the MIN function was used:



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

Hint:Use min function

In [21]:

```
%%sql

SELECT min(date) "Min Date" FROM SPACEX
where landing__outcome = 'Success (ground pad)'
```

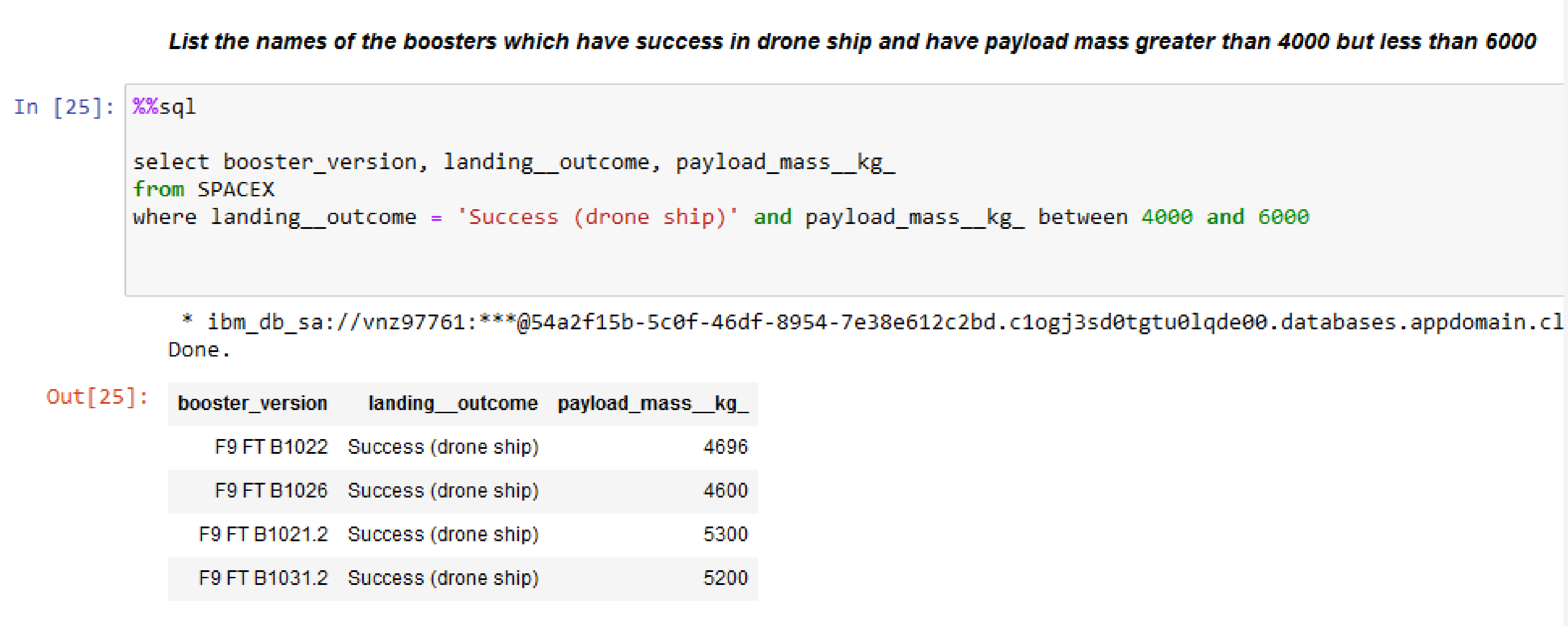
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0
Done.
```

Out[21]:

Min Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 the AND operator was used to combine two restrictions in the WHERE clause:



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

In [25]:

```
%%sql

select booster_version, landing__outcome, payload_mass__kg_
from SPACEX
where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

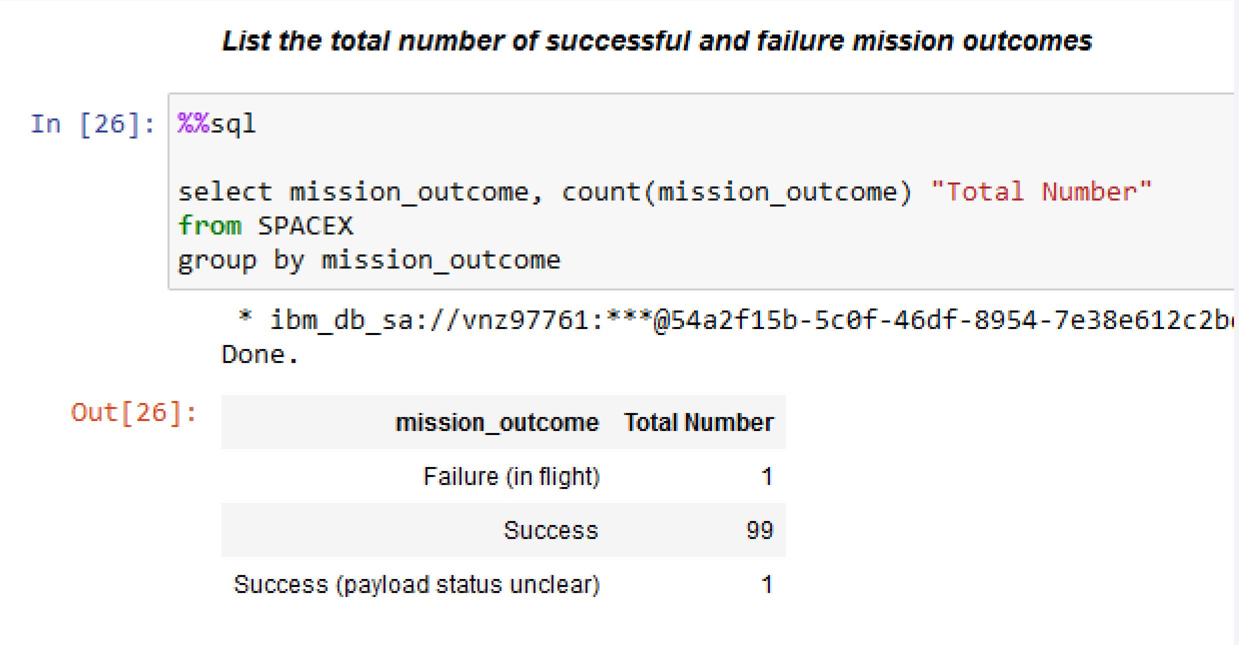
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cl
Done.
```

Out[25]:

booster_version	landing__outcome	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

- To calculate the total number of successful and failure mission outcomes, it was necessary to use the COUNT function with the GROUP BY operator:



List the total number of successful and failure mission outcomes

In [26]:

```
%%sql

select mission_outcome, count(mission_outcome) "Total Number"
from SPACEX
group by mission_outcome
```

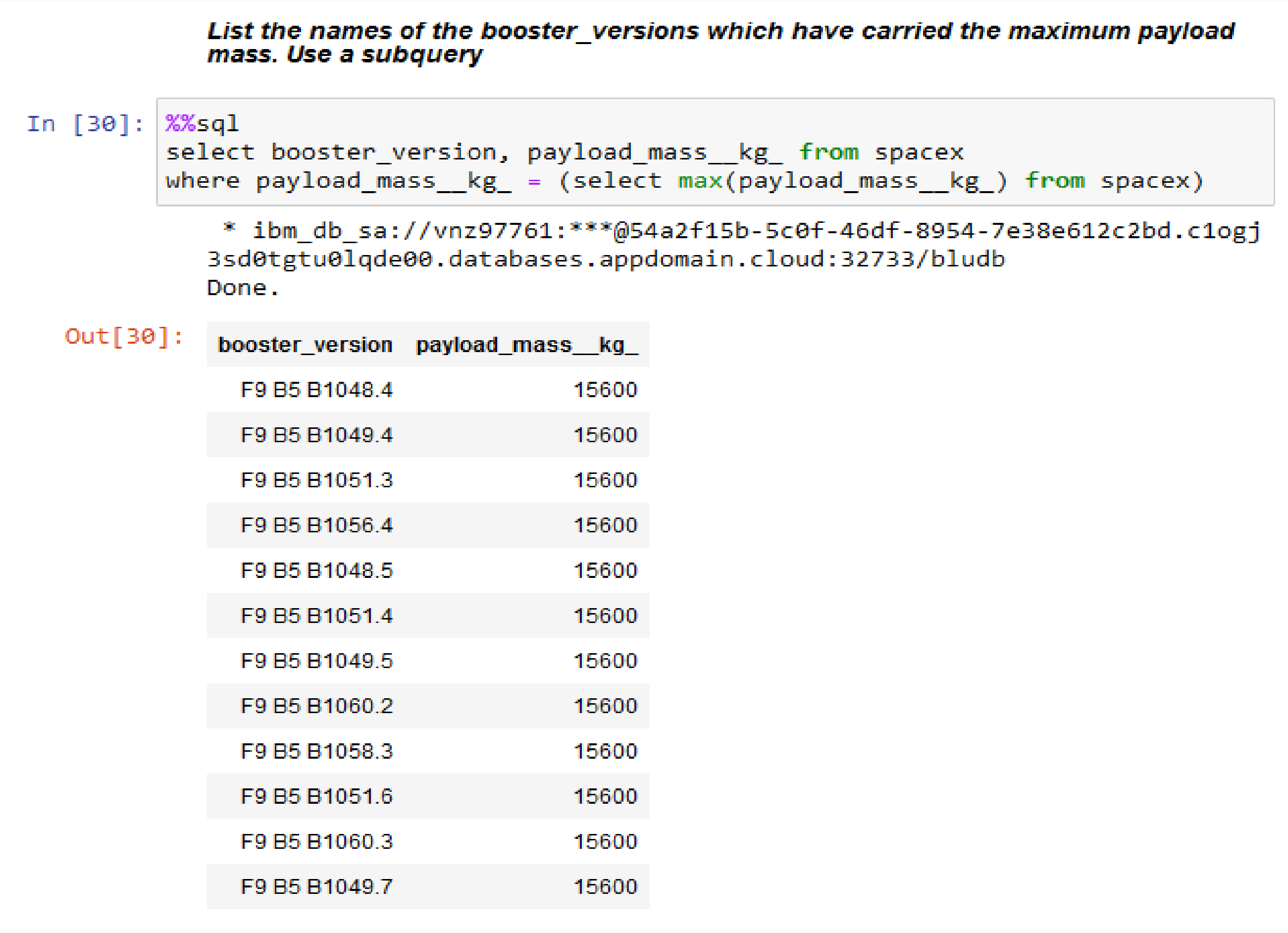
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2b
Done.
```

Out[26]:

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

Boosters Carried Maximum Payload

- To list the names of the booster which have carried the maximum payload mass, the solution consists of using a subquery in the WHERE clause;



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

```
In [30]: %%sql
         select booster_version, payload_mass__kg_ from spacex
         where payload_mass__kg_ = (select max(payload_mass__kg_) from spacex)
```

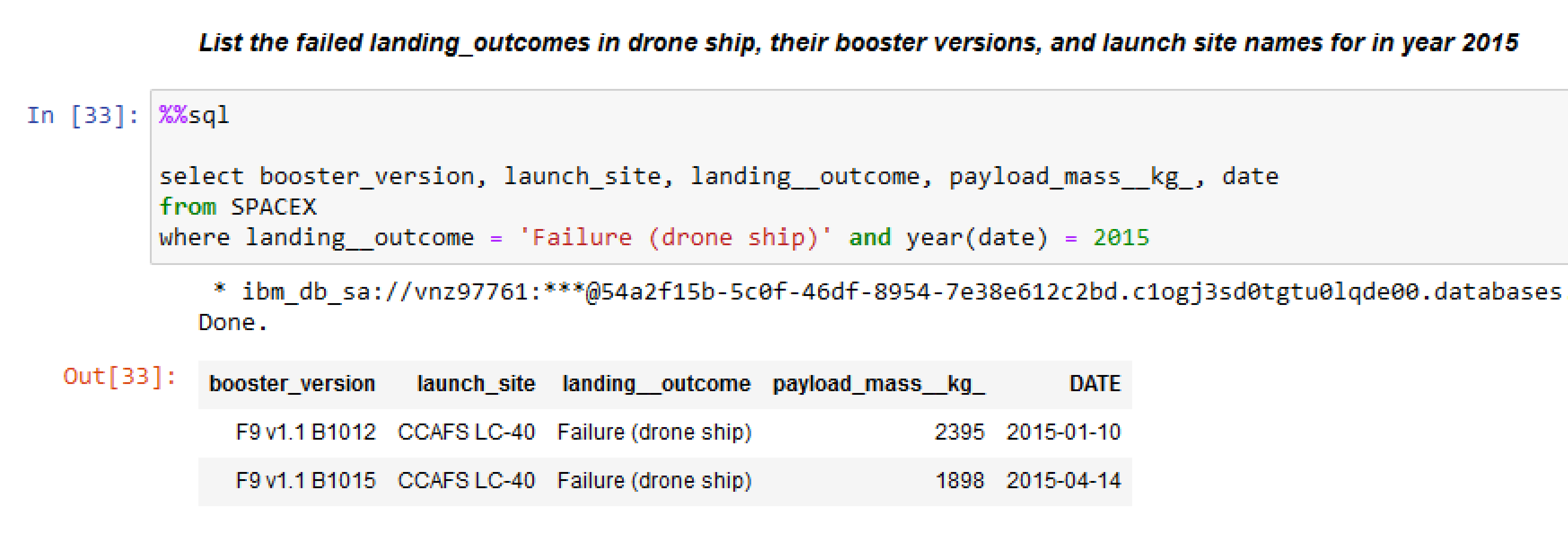
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj
3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[30]:

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

- To list the failed landing outcomes in drone ship, their booster versions, and launch site names in 2015 the YEAR function was combined with the AND operator in the WHERE clause:



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

In [33]:

```
%%sql

select booster_version, launch_site, landing__outcome, payload_mass__kg_, date
from SPACEX
where landing__outcome = 'Failure (drone ship)' and year(date) = 2015
```

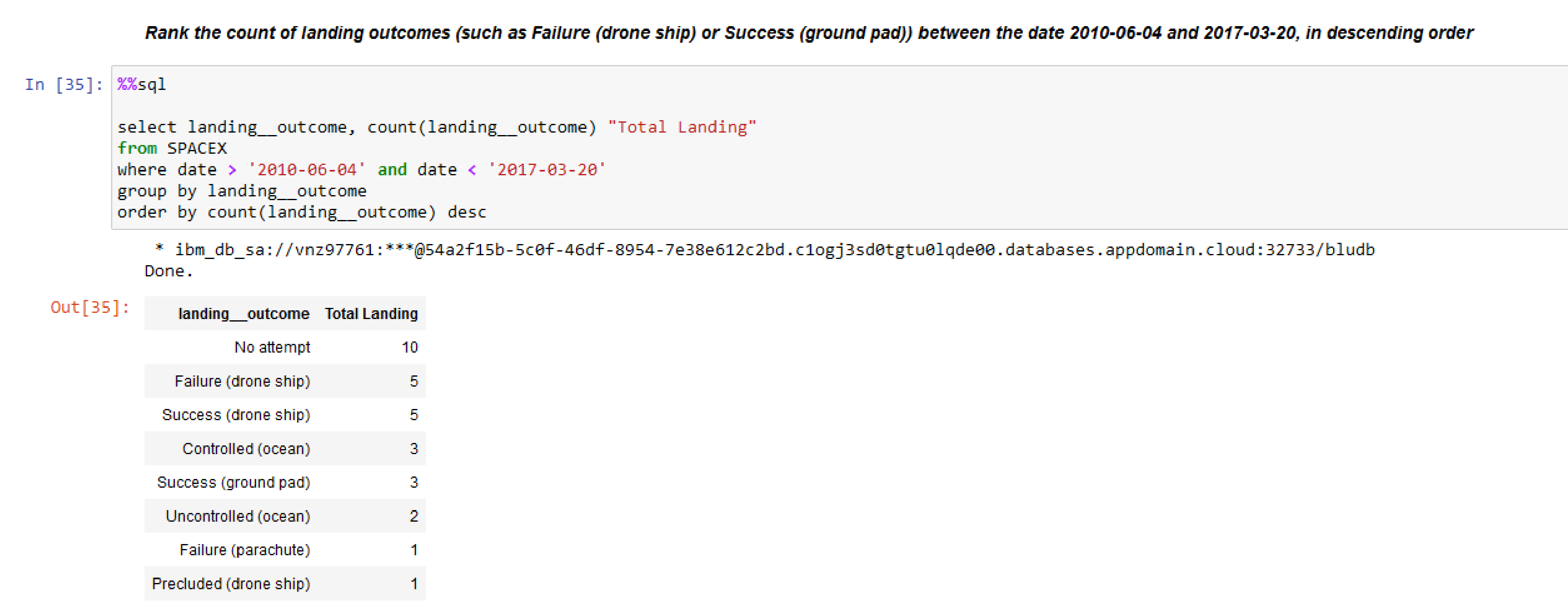
```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases
Done.
```

Out[33]:

booster_version	launch_site	landing__outcome	payload_mass__kg_	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2395	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	1898	2015-04-14

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

- To get a rank of the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order, it was used the COUNT function with the operators GROUP BY and ORDER BY:



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

In [35]:

```
%%sql

select landing__outcome, count(landing__outcome) "Total Landing"
from SPACEX
where date > '2010-06-04' and date < '2017-03-20'
group by landing__outcome
order by count(landing__outcome) desc
```

```
 * ibm_db_sa://vnz97761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

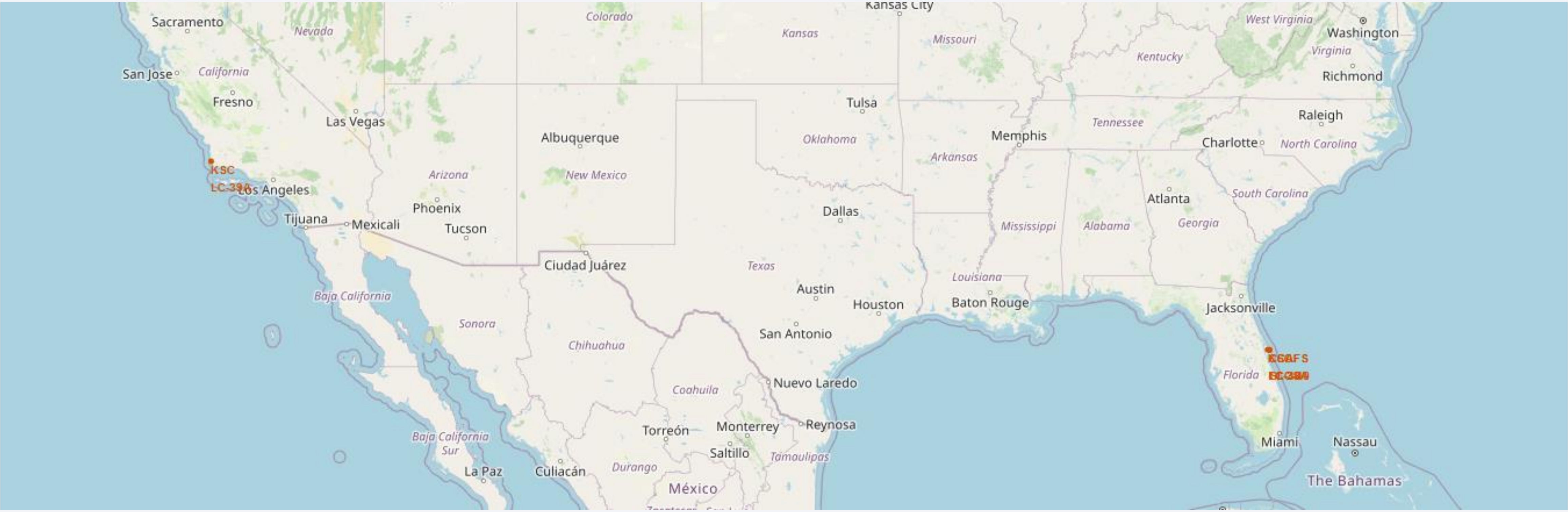
Out[35]:

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

Section 3

Launch Sites Proximities Analysis

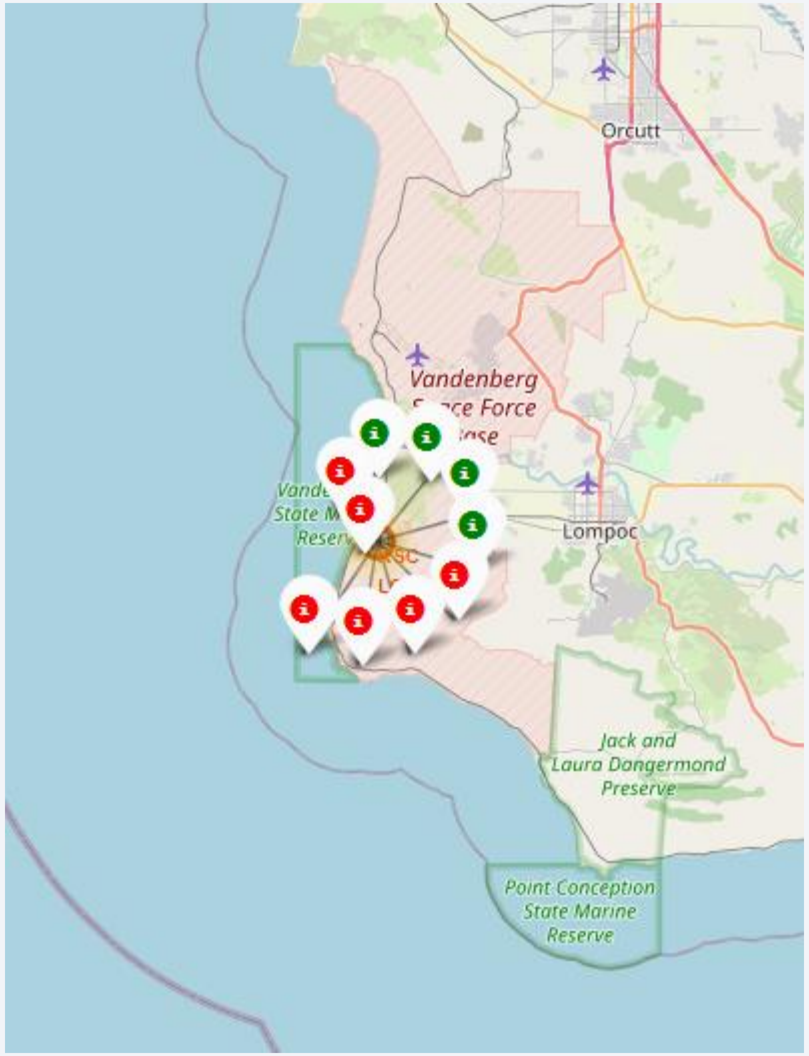
Folium map with all launch sites



- It is possible to see in the map that all launch sites are located in the United States coasts, one in California and two in Florida;
- The proximity with the cost can be explained by the fact that launches toward the sea minimizes the risk of damage from falling debris;

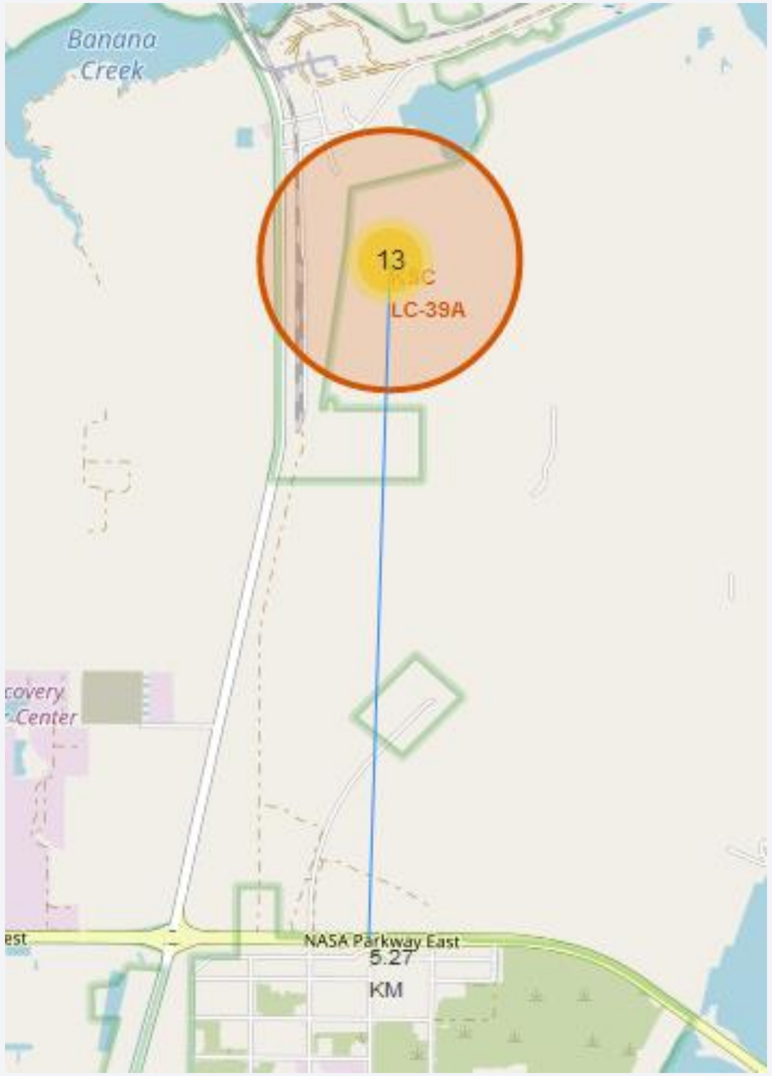
Map with outcome labeled markers

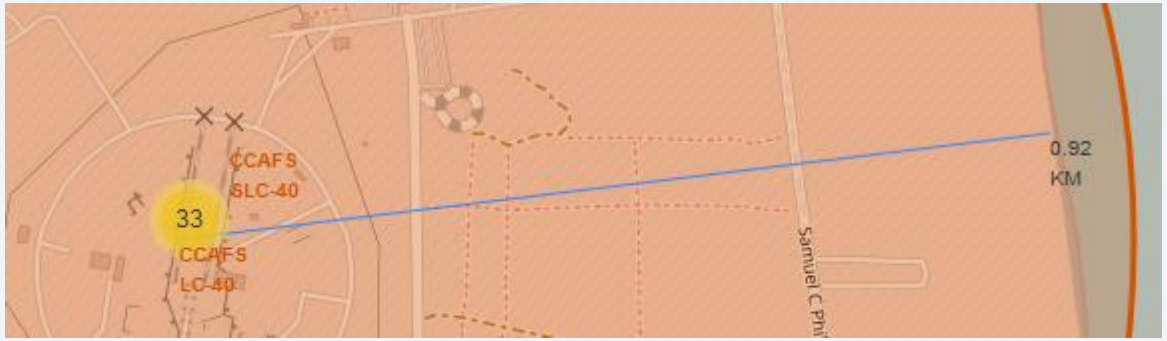
- To each launch it was placed a marker, with color to distinguish failure (red) from success (green)



Maps with distance to proximities

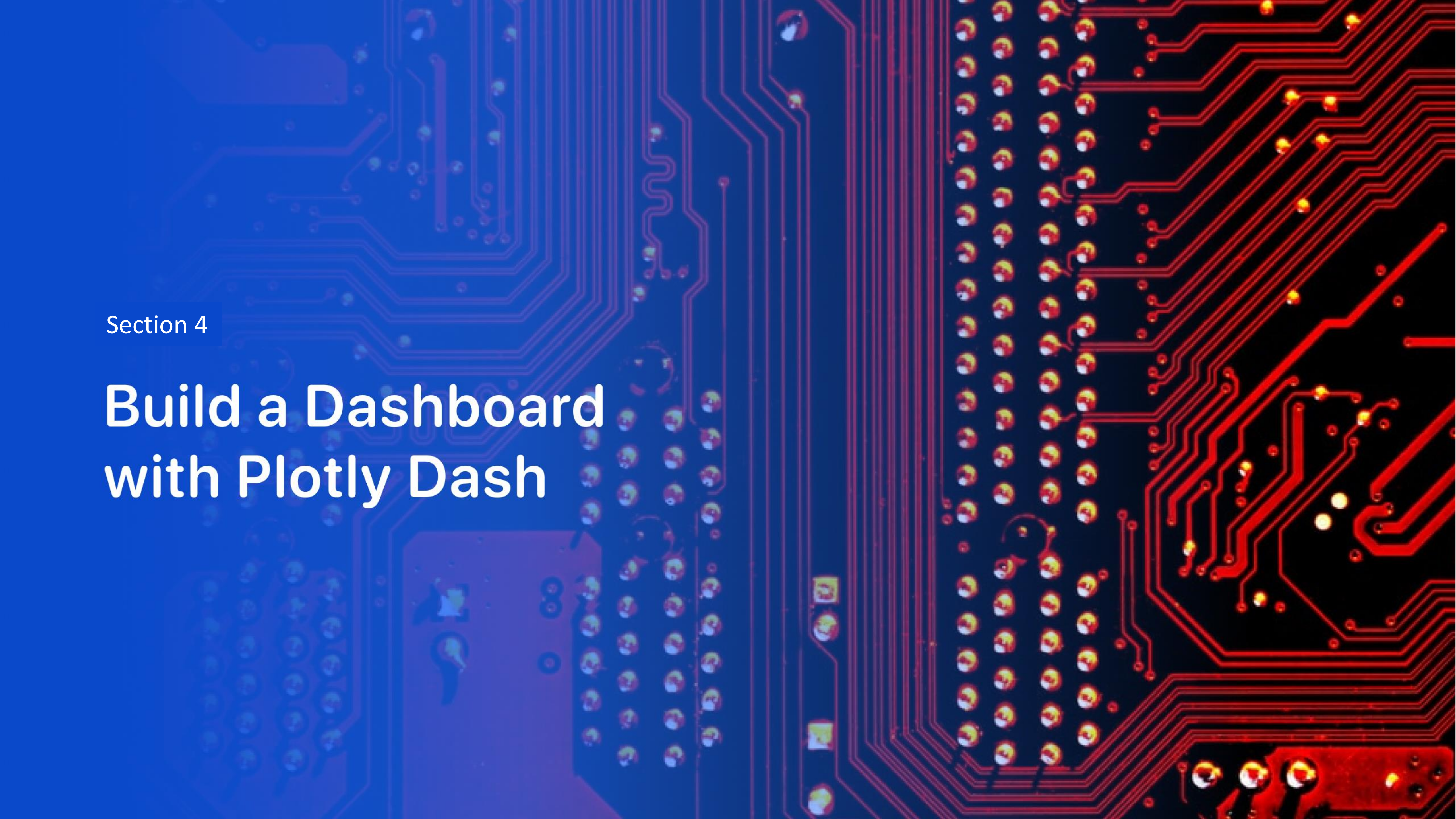
- The distance of various proximities (like closest city, coastline, railway, highway, etc.) was also measured;
- In the example beside, it was calculate the distance from KSC LC 39A to the nearest highway;
- Below, the distance from CCAF-SE LC-40 to coastline;





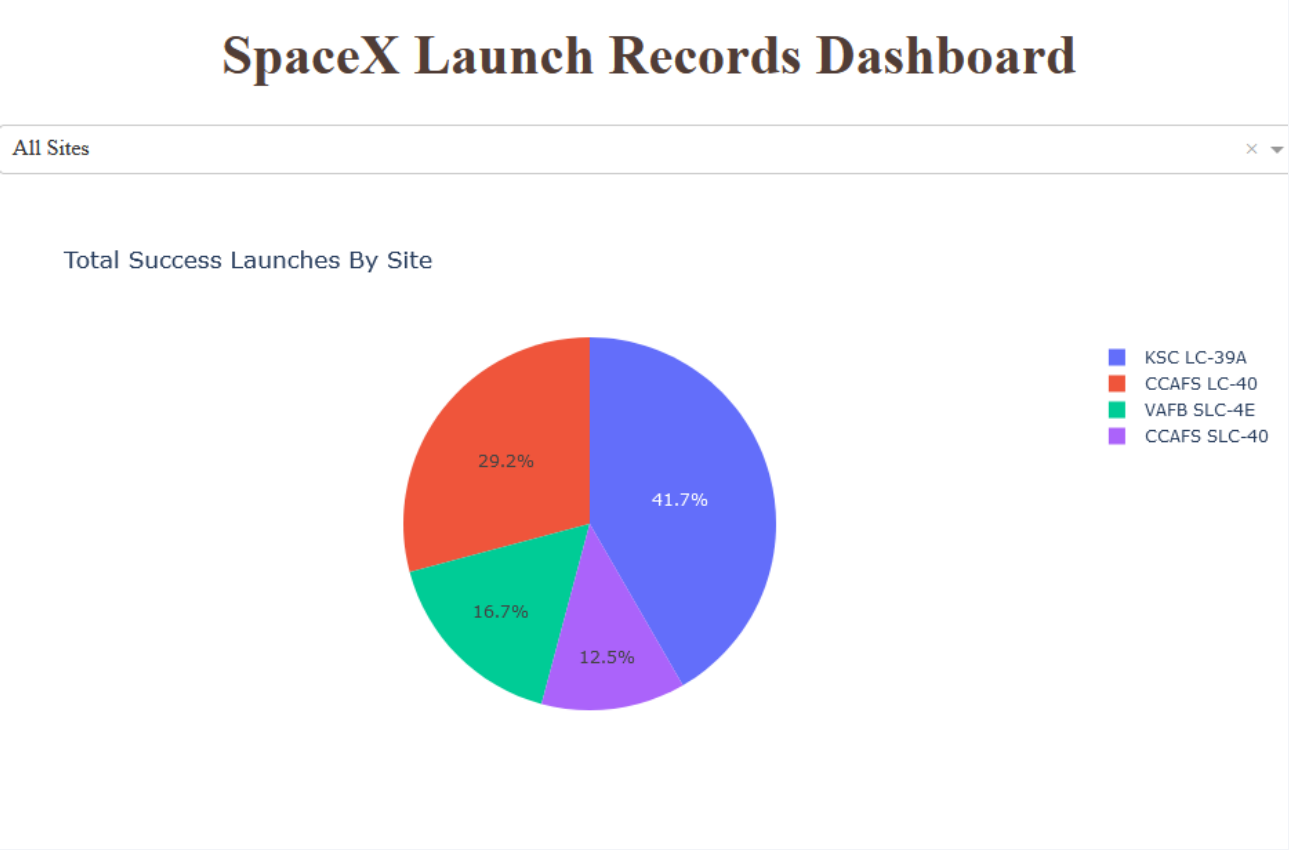
Section 4

Build a Dashboard with Plotly Dash



Dashboard – Successful Launches by Site

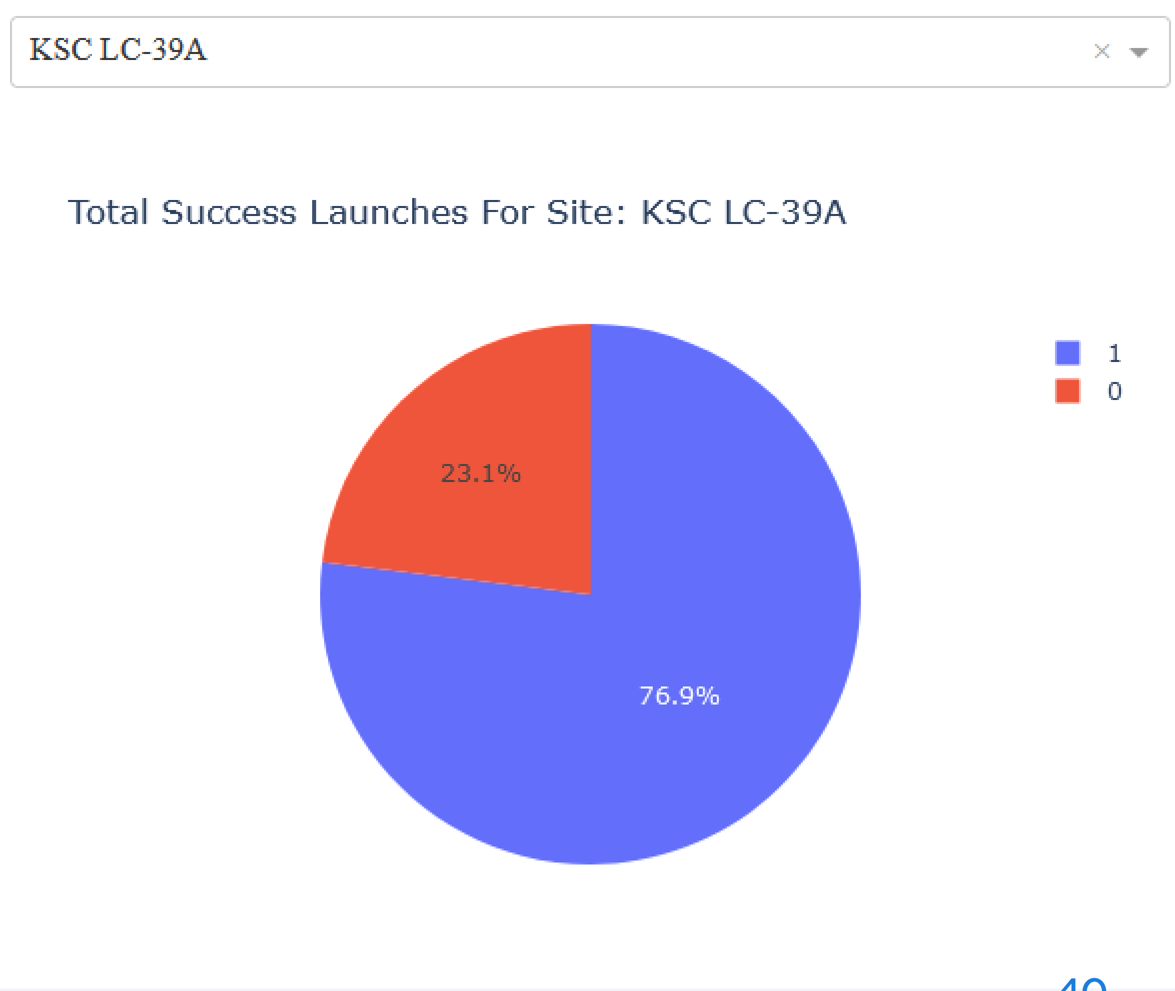
- Using Plotly Dash, a dashboard was created;
- The first graph is a piechart containing the count of successful launches by site;
- There is also a dropdown where is possible to chose a specific Site;
- From all sites, KSC LC-39A had the most successful launches;



SpaceX Launch Records Dashboard

Highest Success Rate Site

- Using the dashboard interactivity it was possible to conclude that KSC LC-39A has the highest success rate, with 76,9% of successful launches
- The success rate of others sites:
  - CCAFS LC-40: 73,1%
  - CCAFS SLC-40: 57,1%
  - VFAB SLC-4E: 60%

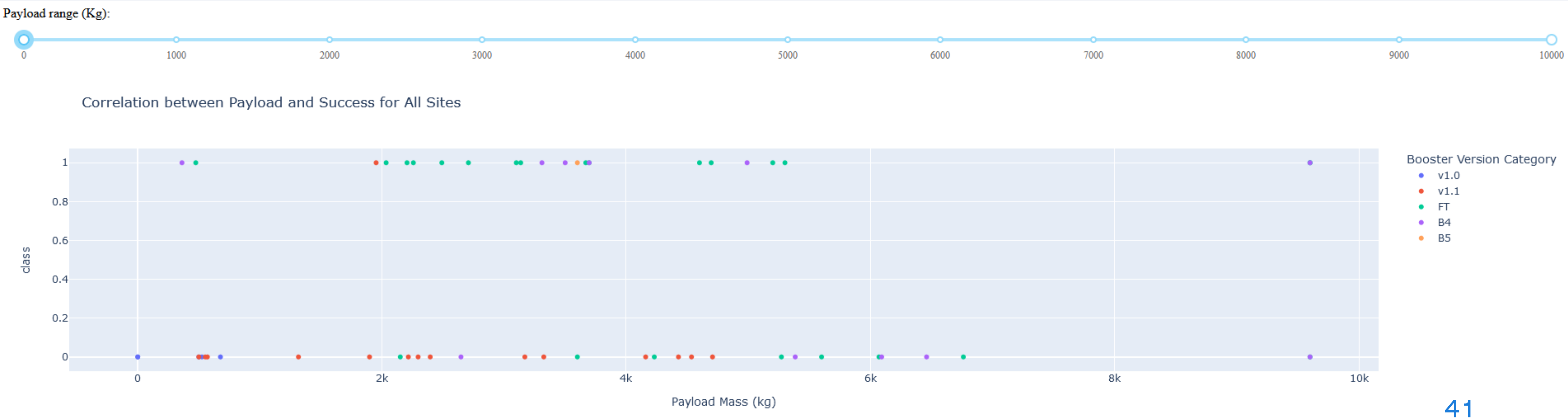


KSC LC-39A

Total Success Launches For Site: KSC LC-39A

Scatterplot Payload vs Launch Outcome

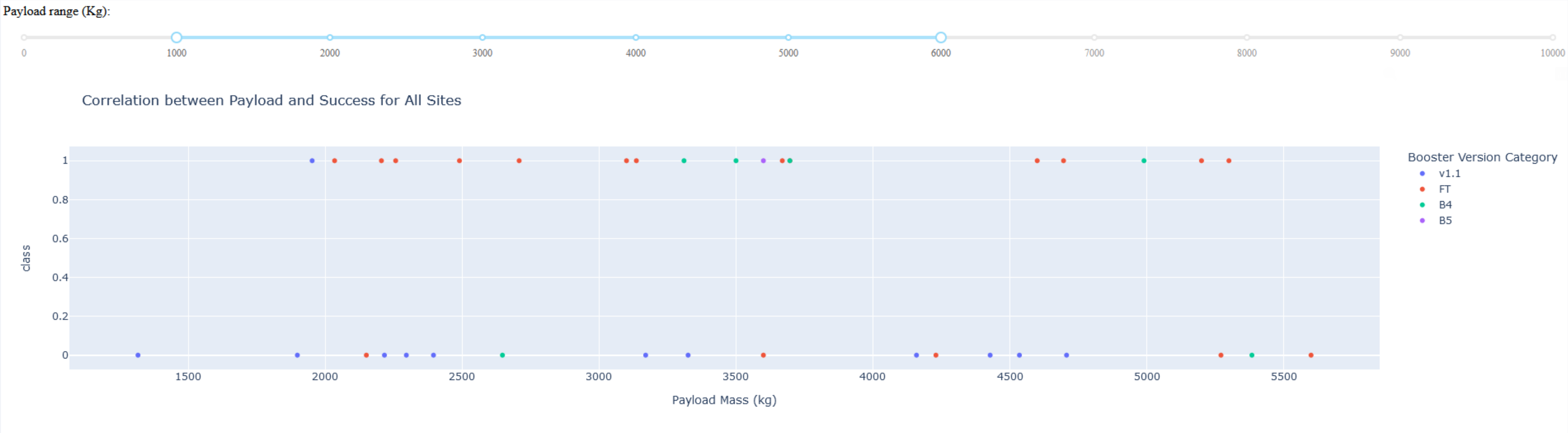
- The second chart of the dashboard is a scatterplot of the payload mass vs launch outcome, colored by booster version category;
- There is a range slider where is possible to change the payload range;



Payload range (Kg):

Scatterplot Payload vs Launch Outcome

- It is possible to conclude from the chart that the most successful payload mass range is from 1952 to 5300kg.



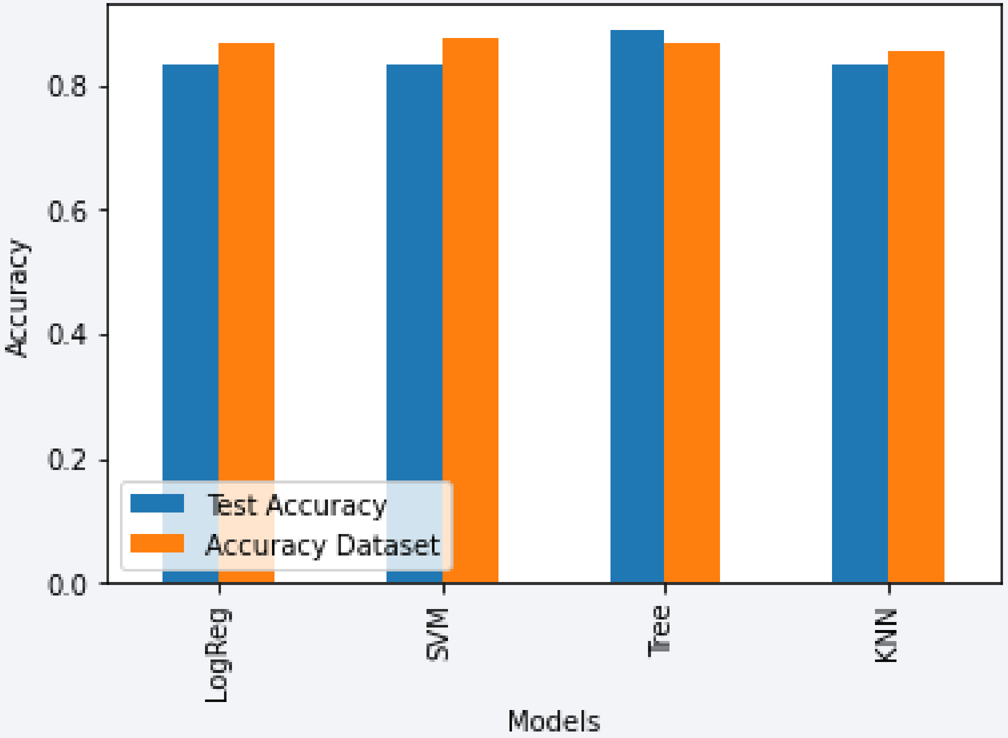
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four models were tested: Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree (Tree), K-Nearest Neighbor (KNN)

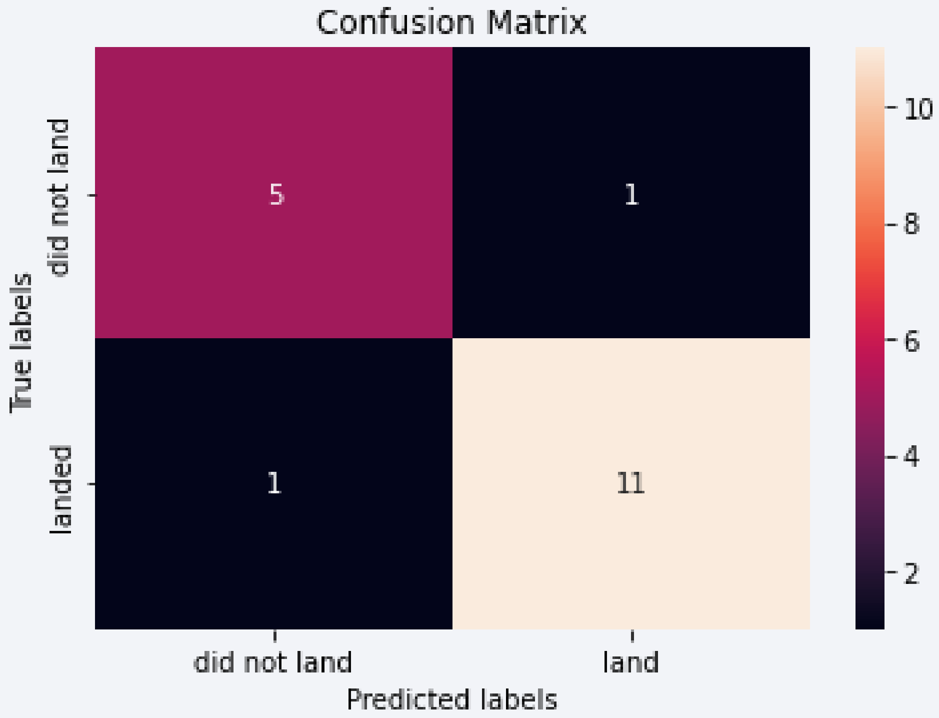
- Based on the models scores, the overall best model was the Decision Tree;



Confusion Matrix

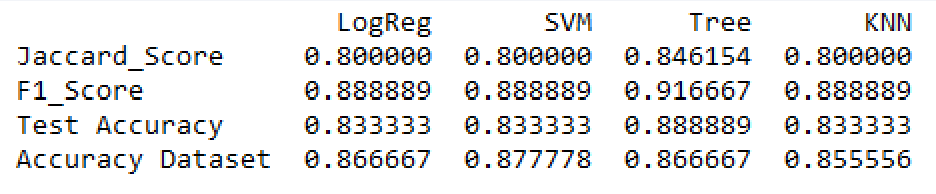
- The confusion matrix of the Decision Tree Model is shown beside;

- It can be concluded that this model has an equally low number of false positives and false negatives (one of each), indicating a good level o accuracy;



Classification Scores

- In addition, other evaluation methods were tested:
    - Jaccard Score;
    - F1 Score
- Theses methods also showed that Decision Tree is the best model for this situation;



```
                    LogReg       SVM      Tree       KNN
Jaccard_Score     0.800000  0.800000  0.846154  0.800000
F1_Score          0.888889  0.888889  0.916667  0.888889
Test Accuracy     0.833333  0.833333  0.888889  0.833333
Accuracy Dataset  0.866667  0.877778  0.866667  0.855556
```

Conclusions

- There are four launch sites, all of them in USA and near the coastline;
- CCAFS SLC 40 site had most of the launches, but KSC LC-39A was the most successful one while the VFAB SLC-4E was the least successful;
- ES-L1, GEO, HEO and SSO were the most successful orbits;
- Success rate have been improving over the years;
- The most successful payload mass range is from 1952 to 5300kg;
- Decision Tree is the best model, with low number of false positives and false negatives;

Appendix

- Higher resolutions screenshots of the Plotly dashboard were added to the github page;

- GitHub Repository: https://github.com/dc-gustavo/IBM-Data-Sience-Capstone-Project

Thank you!