IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ka Chun Leung
15th February 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies:
  - Data is gathered via API and web scraping and then processed into desired dataset
  - Exploratory data analysis (EDA) is performed via visualization and SQL to acquire desired plots needed to explore the relationship between launch success and various factors
  - Interactive visual analytics is performed via Folium and Plotly Dash to produce interactive labeled map and dashboard
  - Predictive analysis is performed via building, tuning and evaluating different classification models (Logistic regression, Decision tree, SVM, KNN) in order to predict the outcome of the launch
- Summary of all results:
  - Payload, number of flights and orbit types are all contributing factors to a successful launch
  - The average launch success rate of SpaceX Falcon 9 first stage kept increasing from 2013 to 2020
  - Launch site KSC LC-39A has the highest launch success rate
  - Payload range 2000-4000 kg and booster version FT have the highest launch success rate
  - All classification models built have 83% accuracy to predict the outcome of the launch

Introduction

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage

- Therefore, if we can determine if the first stage will land, we can determine the cost of a launch

- This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

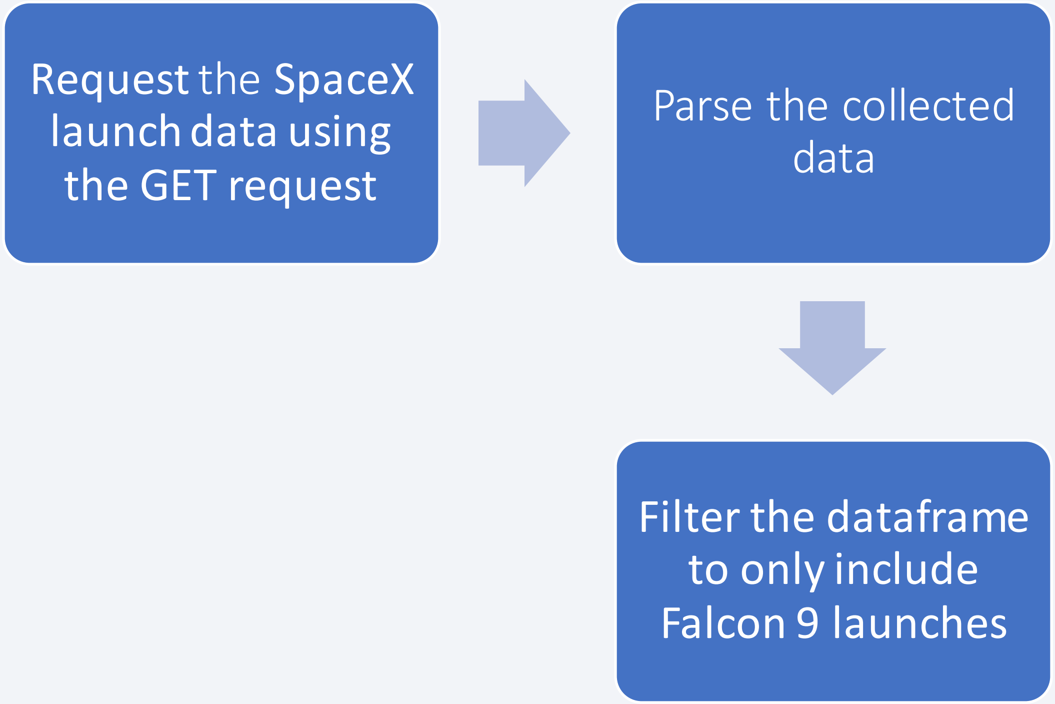
Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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

The completed SpaceX API calls notebook: https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb

Flowchart of SpaceX API calls



Request the SpaceX launch data using the GET request

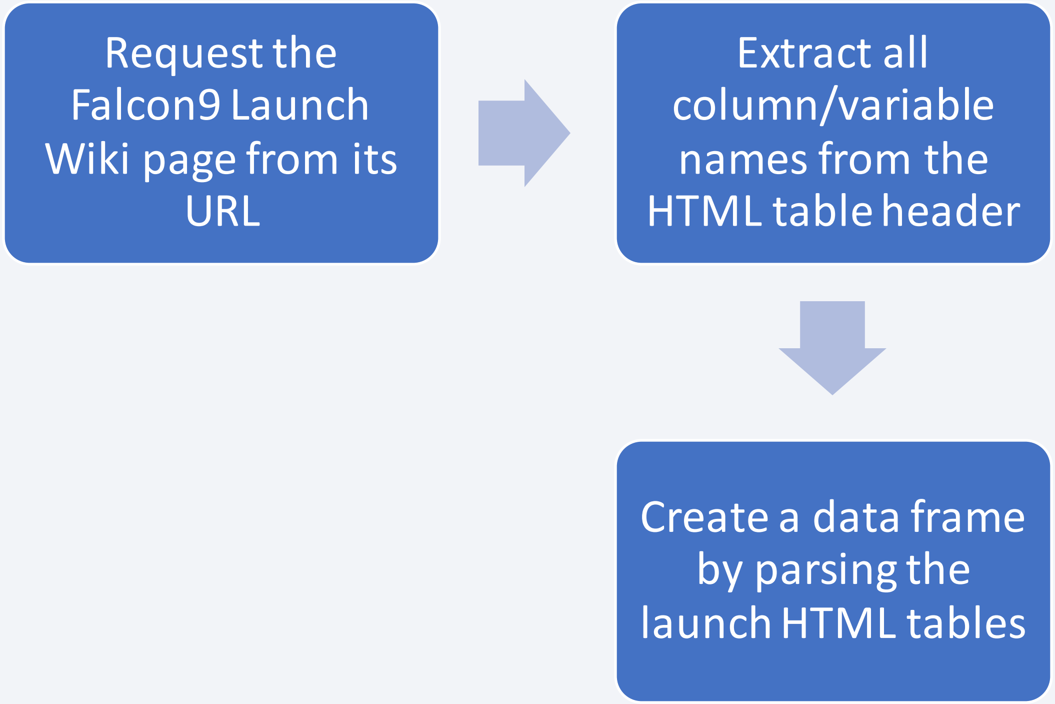
Parse the collected data

Filter the dataframe to only include Falcon 9 launches

Data Collection - Scraping

The completed web scraping notebook: https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/fe1727b8b7a106d53411b8e37ee07f4153218323/Data%20Collection%20with%20Web%20Scraping.ipynb

Flowchart of Web Scraping



Request the Falcon9 Launch Wiki page from its URL

Extract all column/variable names from the HTML table header

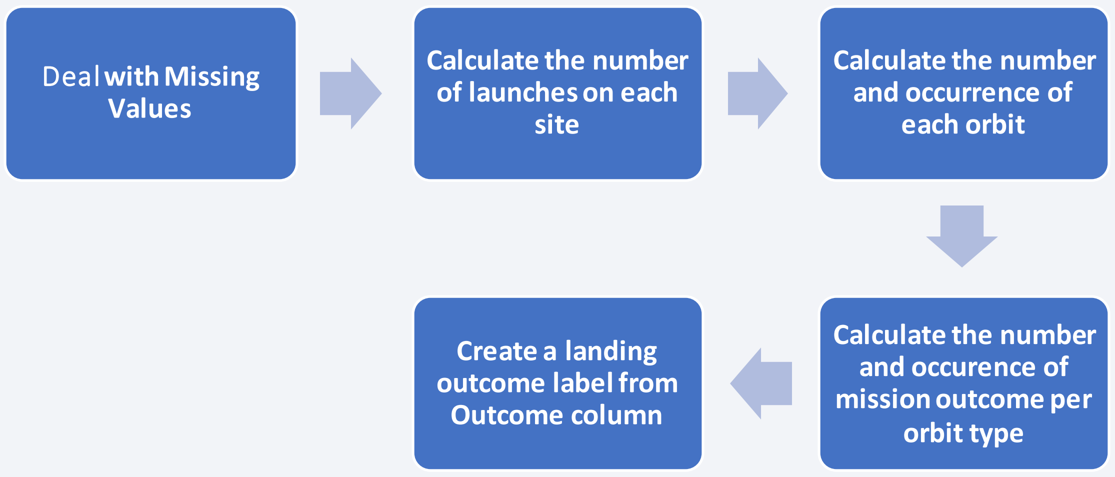
Create a data frame by parsing the launch HTML tables

Data Wrangling

The completed data wrangling notebooks:

- https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/master/Data%20Collection%20API.ipynb
- https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/1517a1ae1c5efc699c08629be098f40539954eb5/EDA.ipynb

Flowchart of Data Wrangling


Deal with Missing Values
Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate the number and occurence of mission outcome per orbit type
Create a landing outcome label from Outcome column

EDA with Data Visualization

- The following charts were plotted:
  - Flight Number vs. Payload Mass scatter point chart and overlay the outcome of the launch
    - To see how the Flight Number and Payload variables would affect the launch outcome
  - Flight Number vs. Launch Site scatter point chart
    - to find patterns in the Flight Number vs. Launch Site scatter point plots
  - Payload Vs. Launch Site scatter point chart
    - to observe if there is any relationship between launch sites and their payload mass
  - Bar chart for the sucess rate of each orbit
    - to Analyze the plotted bar chart and find which orbits have high sucess rate
  - Flight Number Vs. Orbit type scatter point chart
    - to see if there is any relationship between Flight Number and Orbit type
  - Payload vs. Orbit scatter point charts
    - to reveal the relationship between Payload and Orbit type
  - a line chart with x axis to be Year and y axis to be average success rate
    - to get the average launch success trend

- **The completed EDA with data visualization notebook:** https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/1517a1ae1c5efc699c08629be098f40539954eb5/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- The summary the 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 using a subquery
  - 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 (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- **The completed EDA with SQL notebook:** https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/1517a1ae1c5efc699c08629be098f40539954eb5/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Summary of what map objects created and added to a folium map:
    - A Circle and a Marker for each launch site
        - To mark all launch sites on a map with circles and icons showing their names
    - A Marker Cluster , a Marker object for each launch result in spacex_df data frame
        - To mark the success/failed launches for each site on the map so that which launch sites have relatively high success rates can be easily identified
    - A Mouse Position
        - To find the coordinates of any points of interests easily
    - A Marker on the selected closest coastline point
        - To display the distance between coastline point and launch site
    - A Poly Line using the coastline coordinates and launch site coordinate
        - To draw a Poly Line between a launch site to the selected coastline point
    - A Marke with distance to a closest city, railway, highway and a line between the marker to the launch site
        - To calculate the distances between a launch site to its proximities
- **The completed interactive map with Folium map:** https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/1517a1ae1c5efc699c08629be098f40539954eb5/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

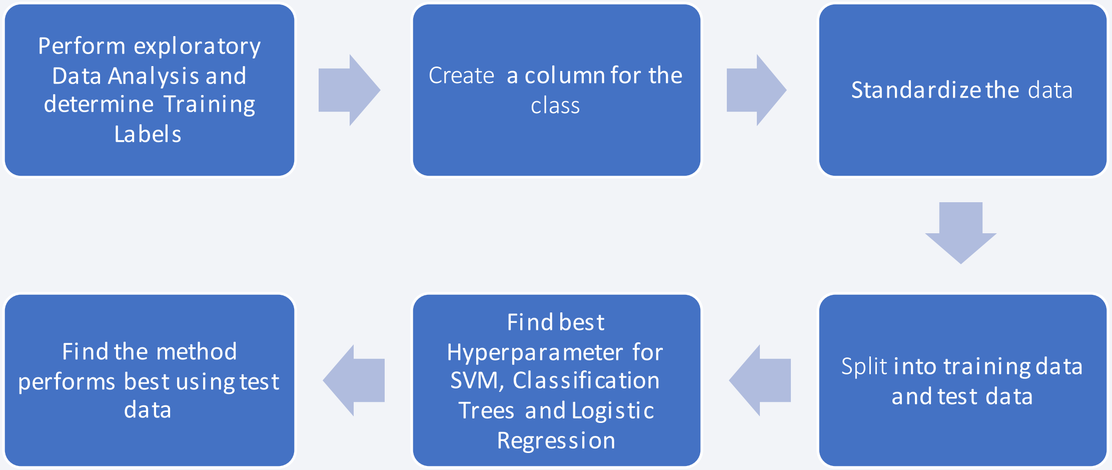
Build a Dashboard with Plotly Dash

- Summary of what plots/graphs and interactions have been added to a dashboard:
  - A Launch Site Dropdown menu and a Pie Chart
    - To get the selected launch site from dropdown menu and render a pie chart visualizing launch success counts
  - A Range Slider for selecting various Payload Range and a Scatter Chart
    - To visually observe how payload may be correlated with mission outcomes for selected site(s)
- The GitHub URL of the completed Plotly Dash lab: https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/92192371a416e4b012ae1f1ea07420cb6656f795/spacex_dash_app.py

Predictive Analysis (Classification)

The GitHub URL of the completed predictive analysis lab: https://github.com/Snakeybob/Applied-Data-Science-Capstone/blob/92192371a416e4b012ae1f1ea07420cb6656f795/Machine%20Learning%20Prediction.ipynb

Model development process



Perform exploratory Data Analysis and determine Training Labels

Create a column for the class

Standardize the data

Split into training data and test data

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

Find the method performs best using test data

Results

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

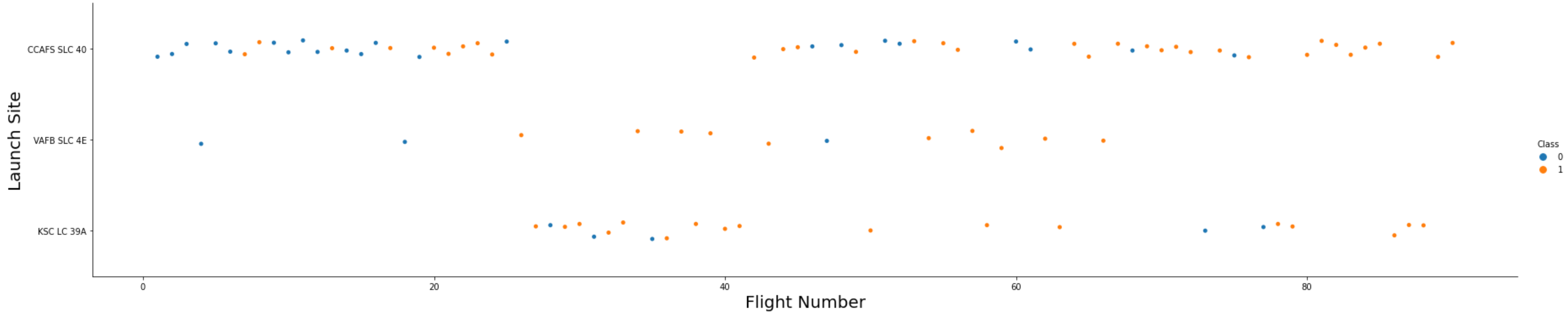
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Observation:

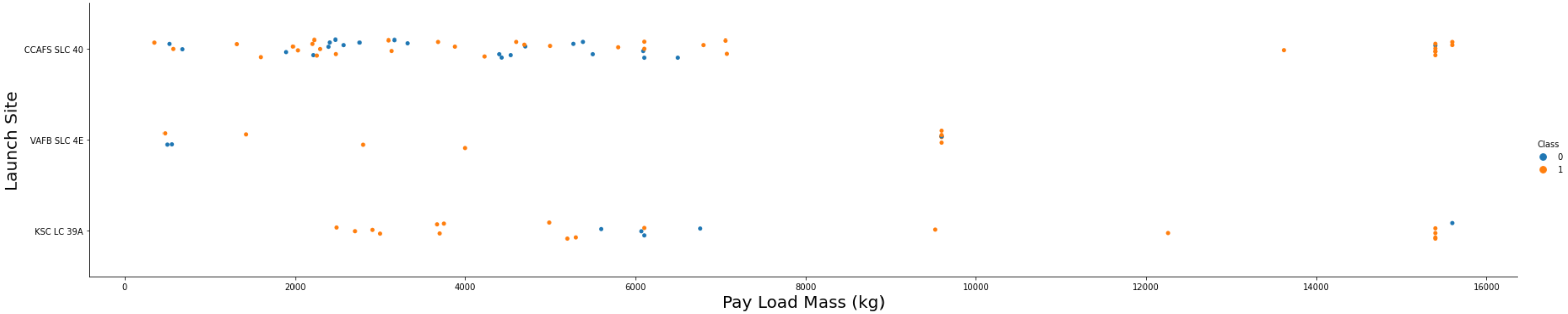
- The success rate for launch site **CCAFS SLC-40** seems to increase as the flight number increases



Payload vs. Launch Site

Observation:

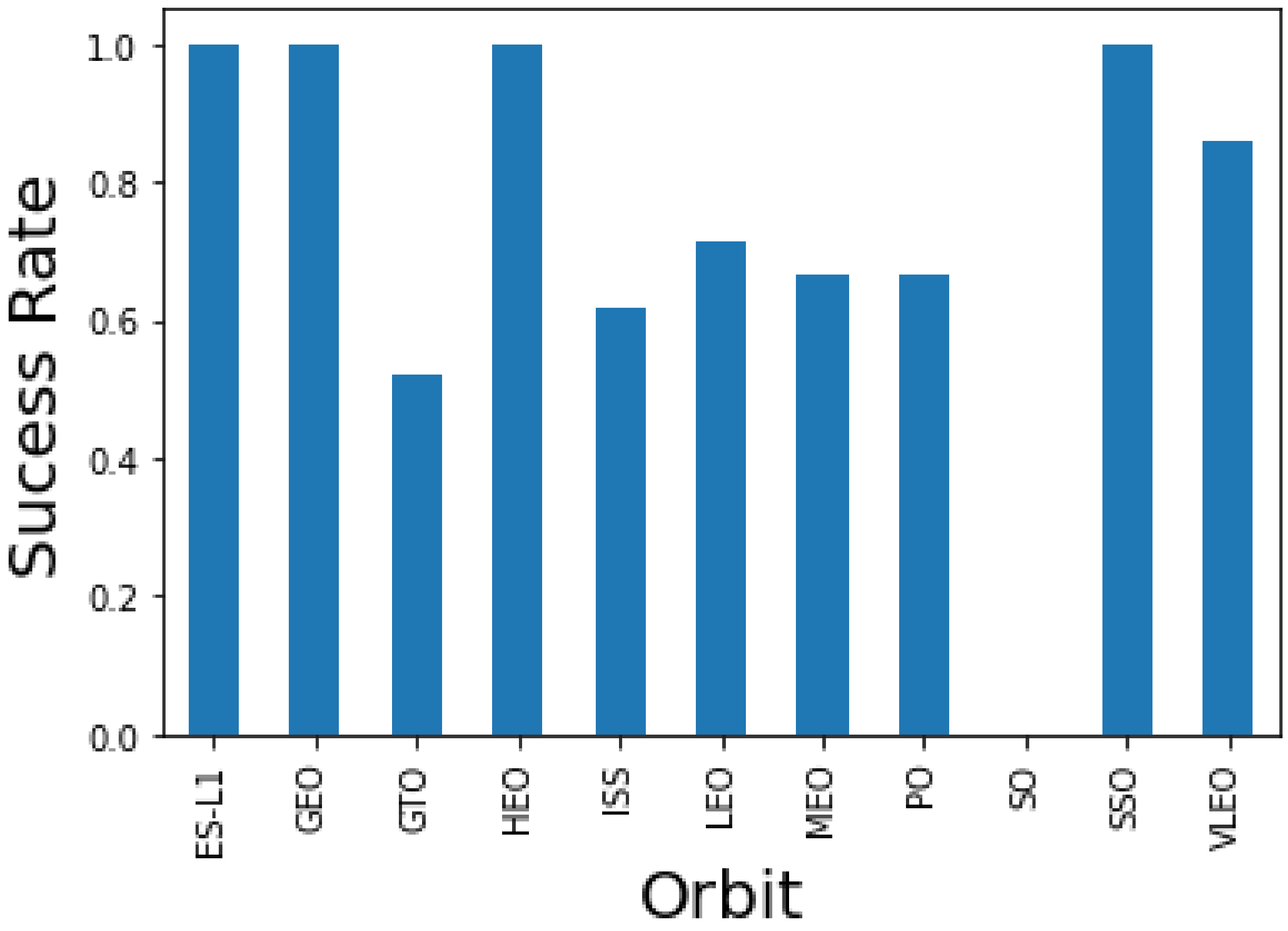
- There are no rockets launched for heavy payload mass(greater than 10000kg) for the **VAFB-SLC** launch site
- The success rate appears to be higher for heavy payload mass(greater than 10000kg) for the **CCAFS SLC-40** launch site



Success Rate vs. Orbit Type

Observation:

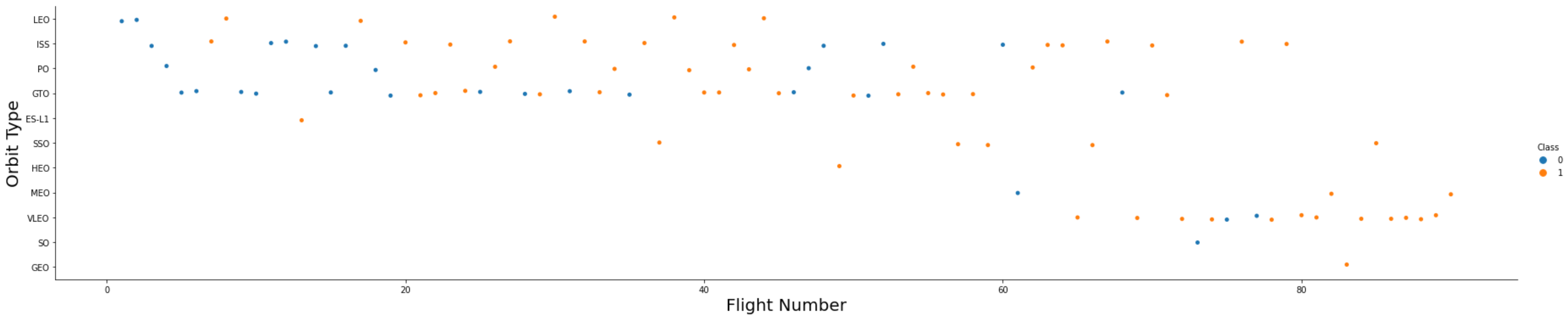
- The success rate are the highest for the following orbits:
    - **ES-L1**
    - **GEO**
    - **HEO**
    - **SSO**
- Orbit **GTO** has the lowest success rate



Flight Number vs. Orbit Type

Observation:

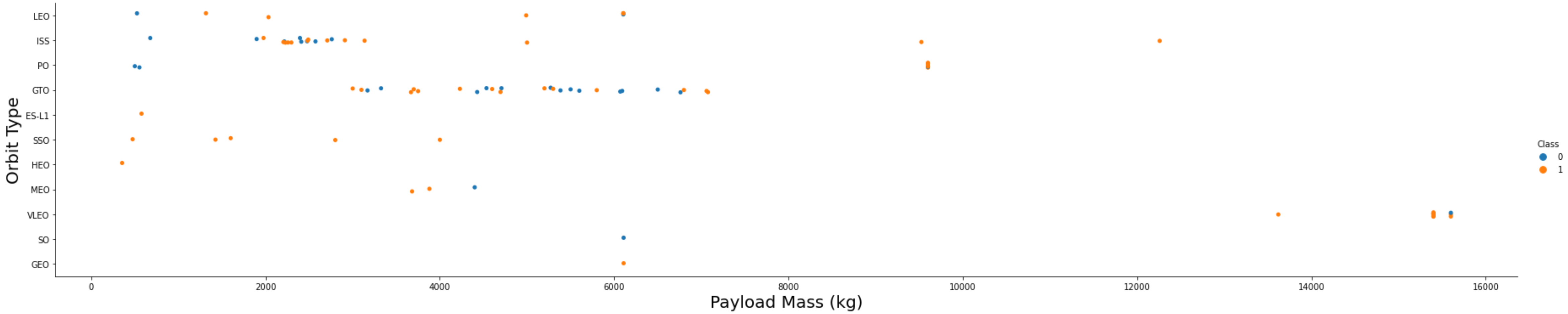
- The Success appears related to the number of flights in the **LEO** orbit



Payload vs. Orbit Type

Observation:

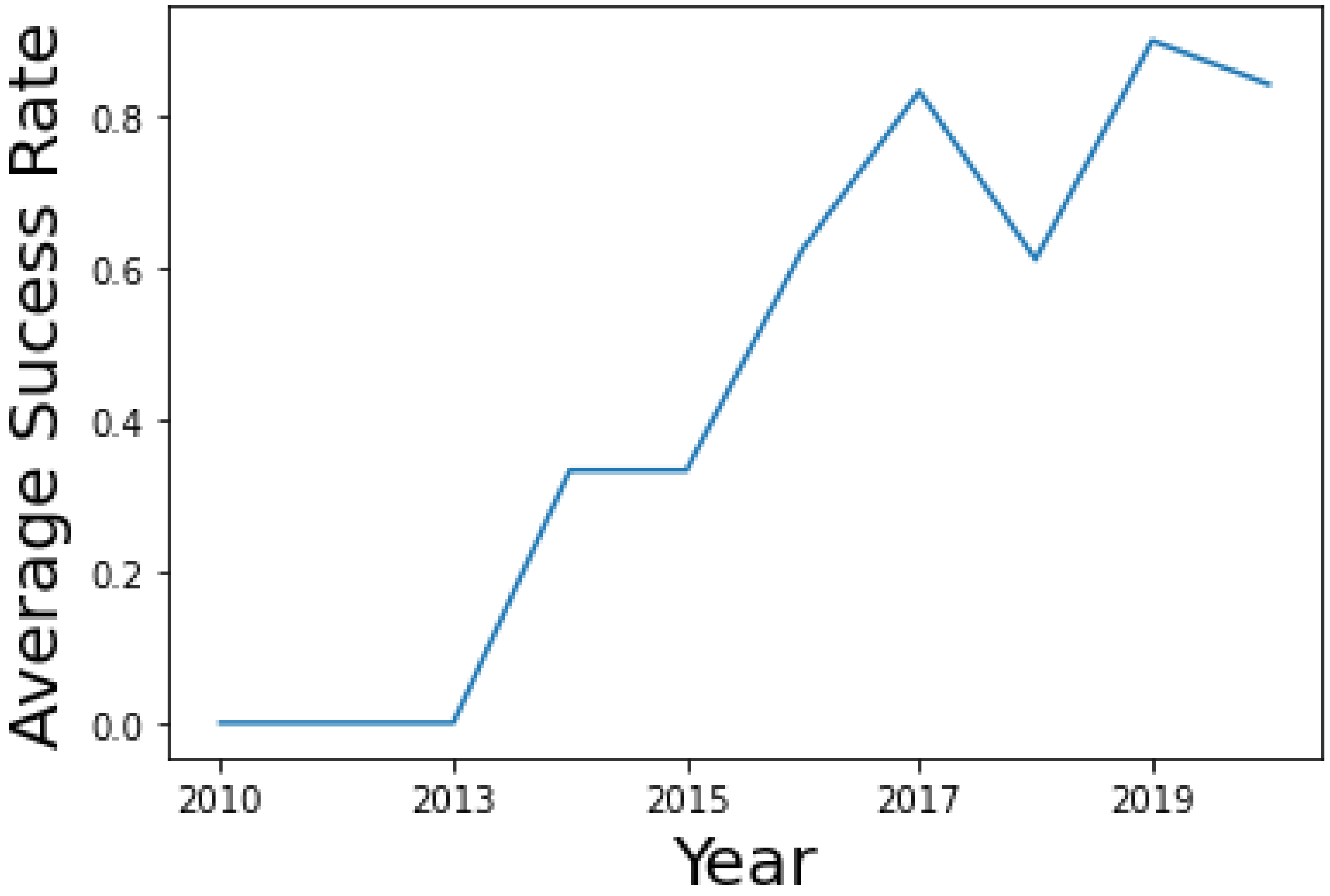
- The successful landing or positive landing rate are higher for **Polar**, **LEO** and **ISS** with heavy payloads



Launch Success Yearly Trend

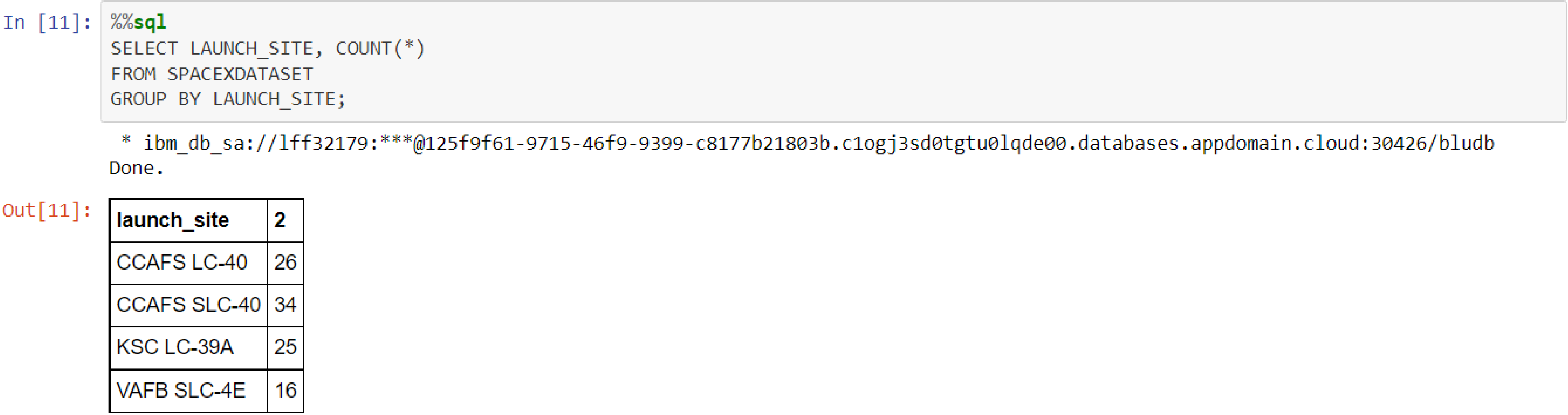
Observation:

- The sucess rate since 2013 kept increasing till 2020



All Launch Site Names

- The names of the unique launch sites are found using the following query



```
In [11]: %%sql
         SELECT LAUNCH_SITE, COUNT(*)
         FROM SPACEXDATASET
         GROUP BY LAUNCH_SITE;
```

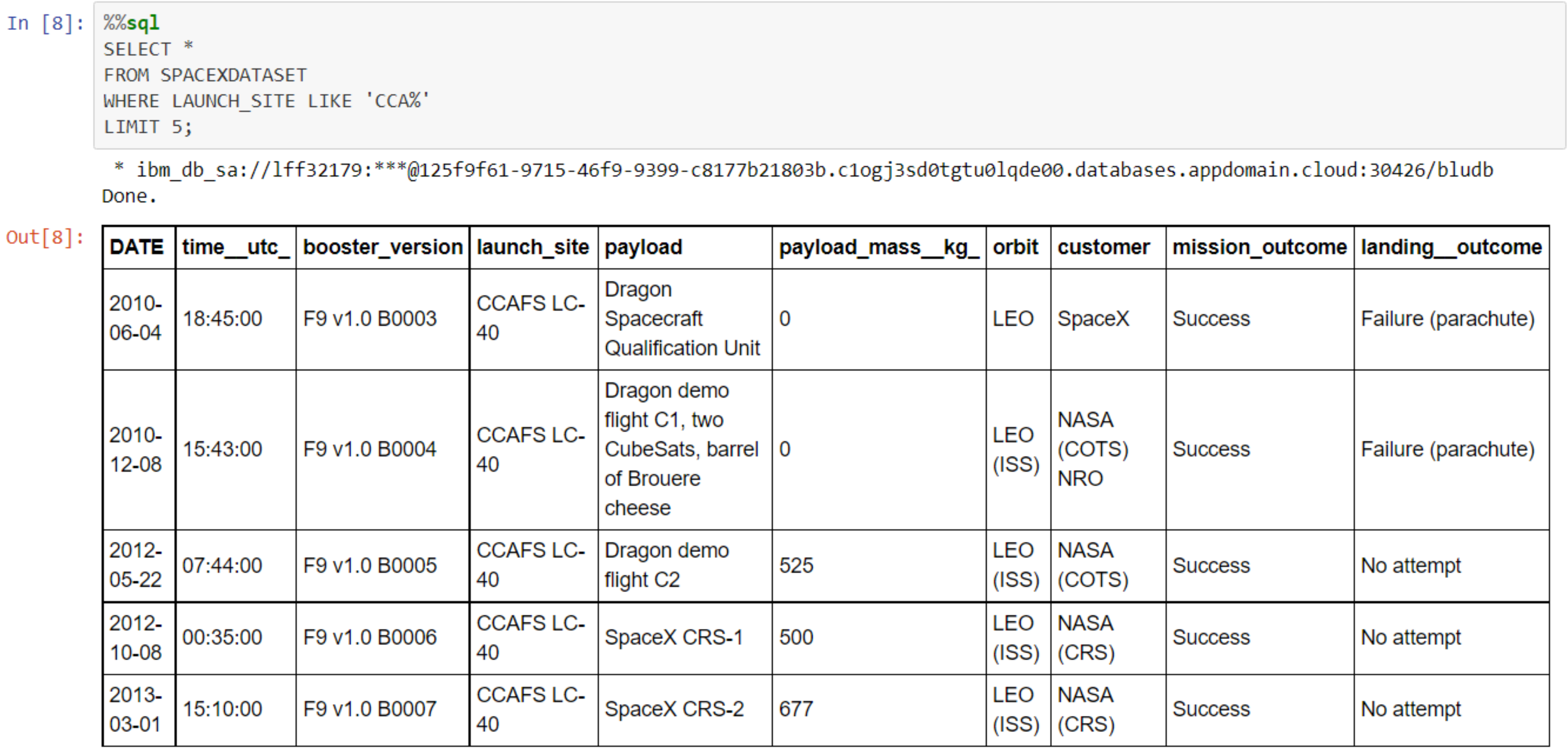
```
 * ibm_db_sa://lff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[11]:

launch_site	2
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

Launch Site Names Begin with 'CCA'

- The first 5 records where launch sites begin with `CCA` are found using the following query



In [8]:

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

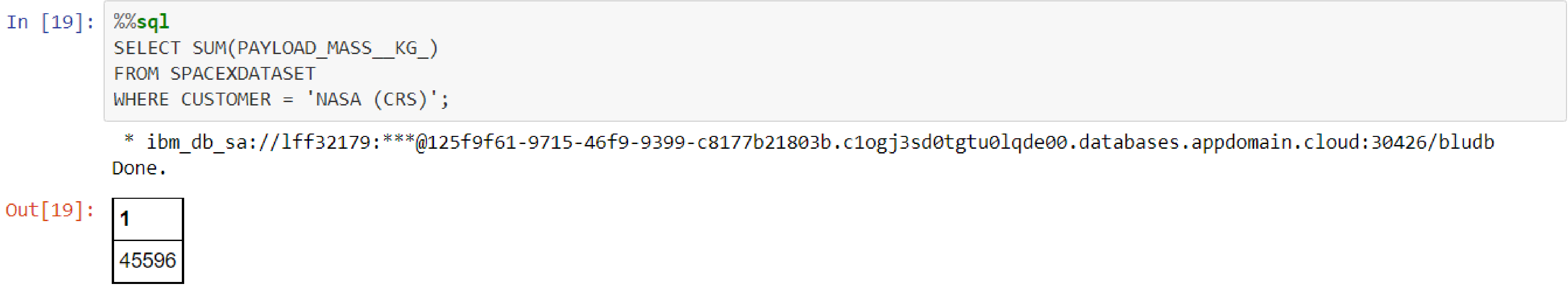
```
 * ibm_db_sa://1ff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

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

- The total payload carried by boosters from NASA is calculated using the following query



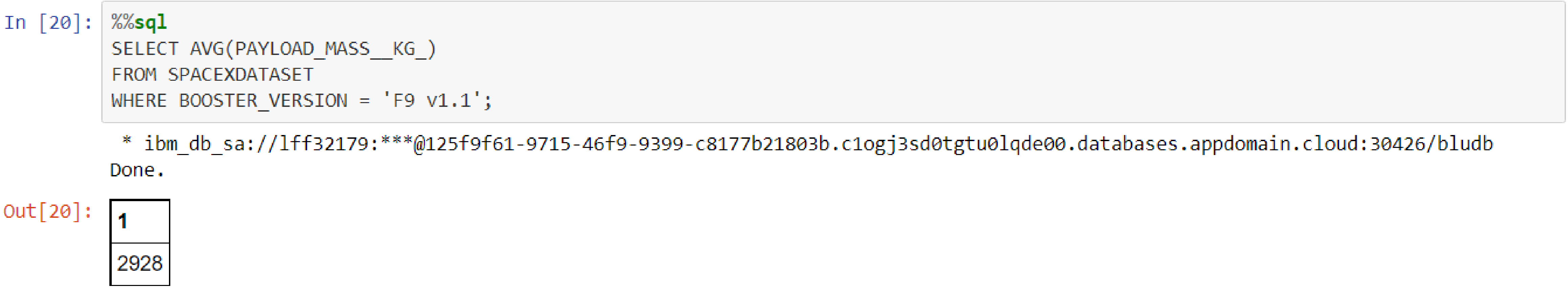
```
In [19]: %%sql
         SELECT SUM(PAYLOAD_MASS__KG_)
         FROM SPACEXDATASET
         WHERE CUSTOMER = 'NASA (CRS)';

 * ibm_db_sa://lff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[19]:

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is calculated using the following query



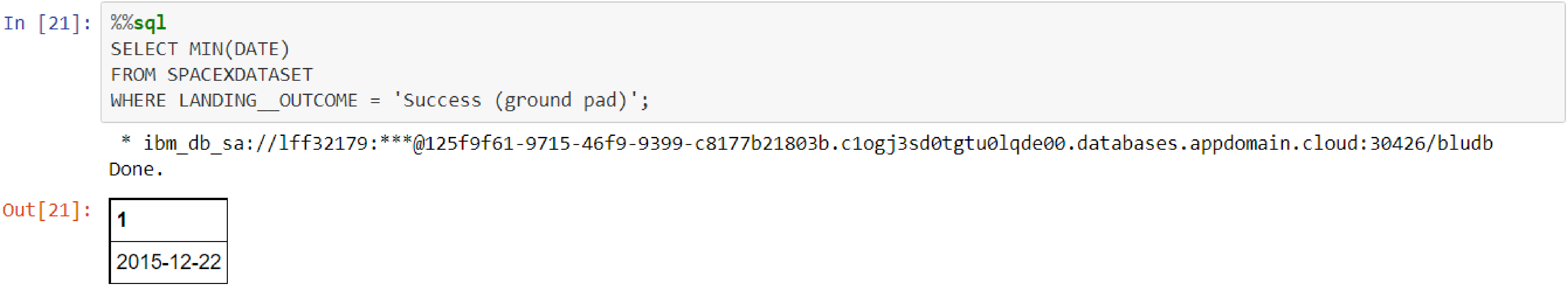
```
In [20]: %%sql
         SELECT AVG(PAYLOAD_MASS__KG_)
         FROM SPACEXDATASET
         WHERE BOOSTER_VERSION = 'F9 v1.1';

          * ibm_db_sa://lff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
         Done.
```

Out[20]:

1
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is found using the following query

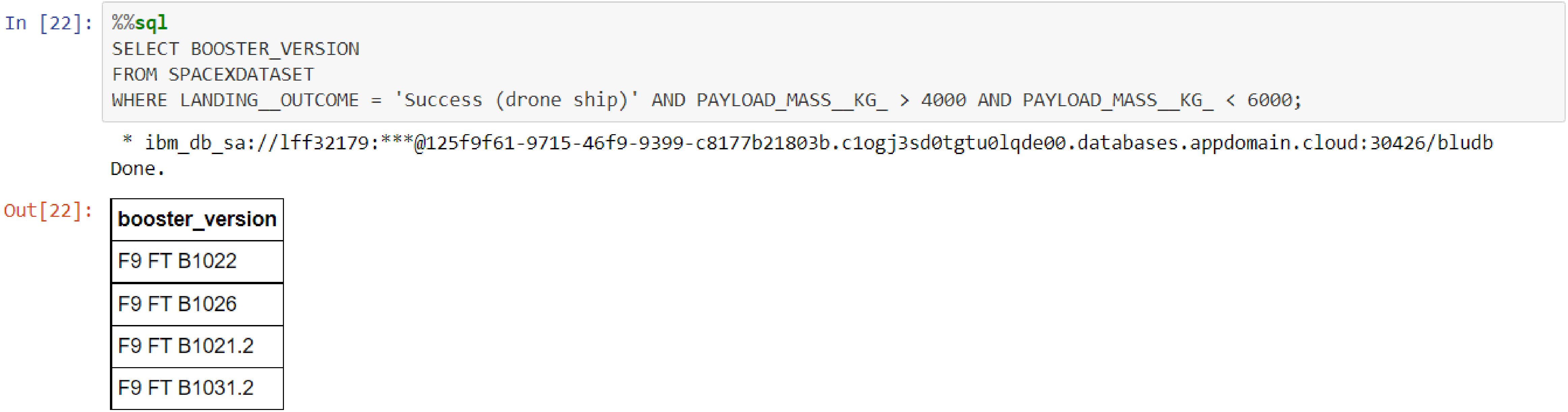


```
In [21]: %%sql
         SELECT MIN(DATE)
         FROM SPACEXDATASET
         WHERE LANDING__OUTCOME = 'Success (ground pad)';

          * ibm_db_sa://lff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
         Done.

Out[21]:
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed using the following query



In [22]:

```
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

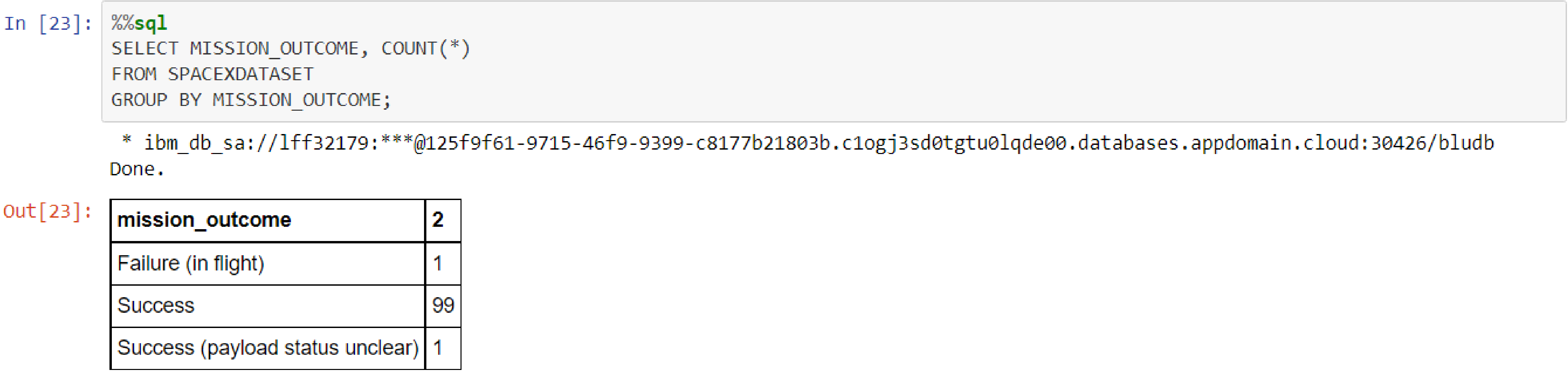
```
 * ibm_db_sa://1ff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[22]:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are calculated using the following query



In [23]:

```
%%sql
SELECT MISSION_OUTCOME, COUNT(*)
FROM SPACEXDATASET
GROUP BY MISSION_OUTCOME;
```

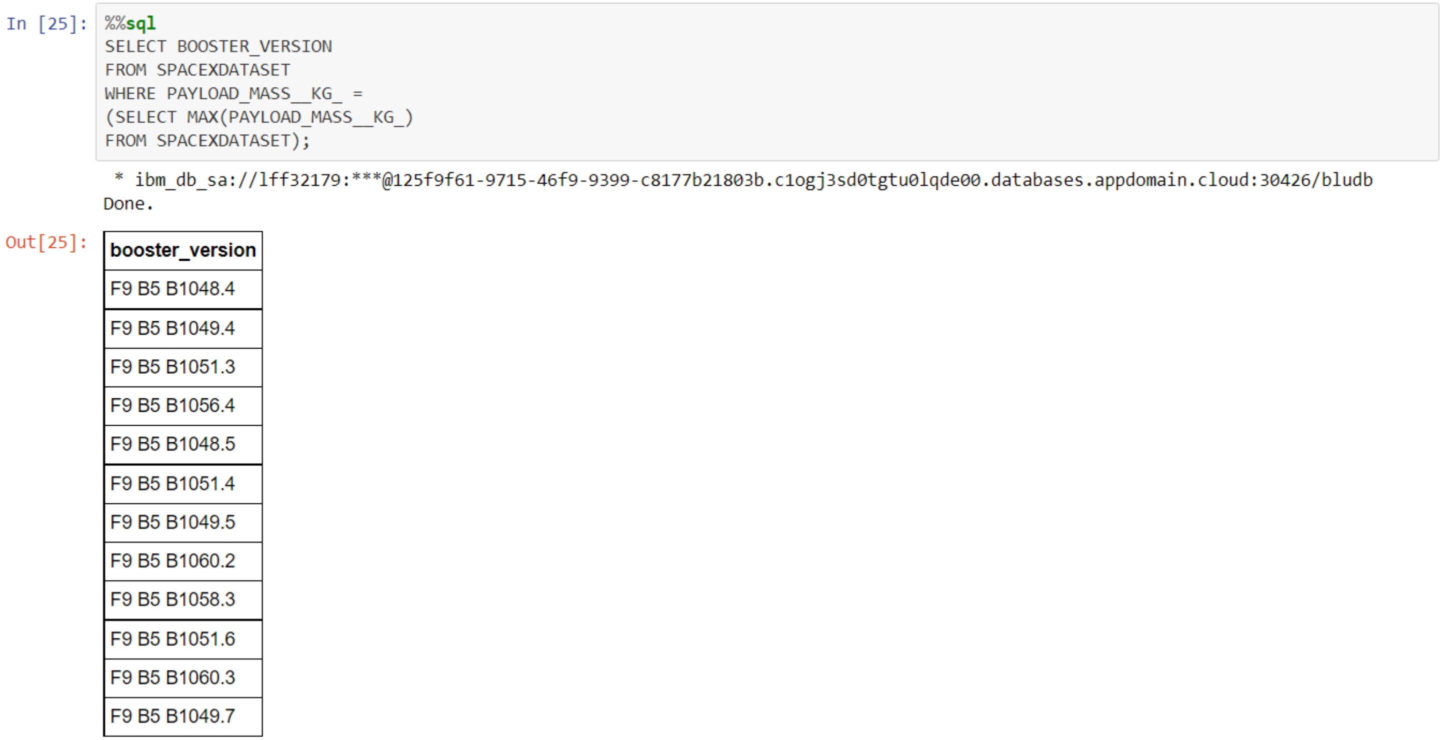
```
 * ibm_db_sa://1ff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[23]:

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

Boosters Carried Maximum Paylo ad

- The names of the booster which have carried the maximum payload mass are listed using the following query

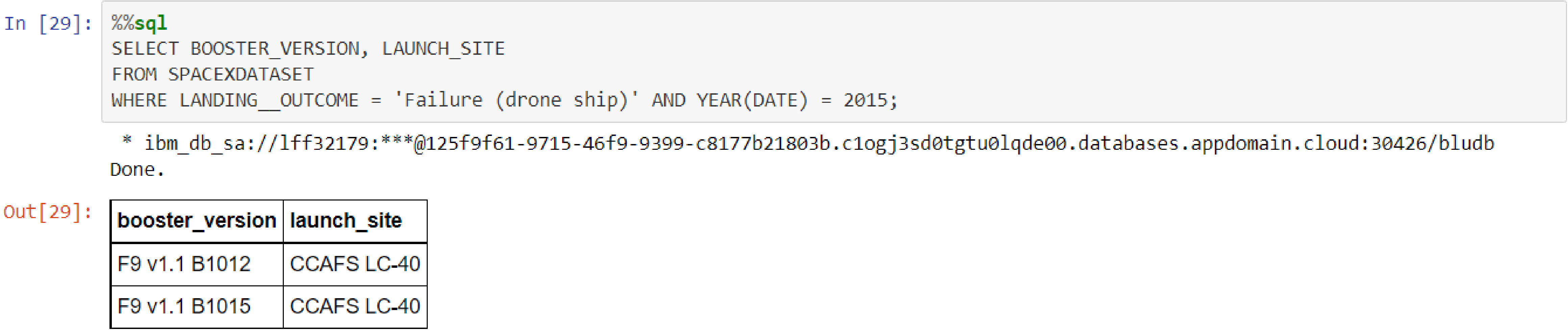


```
In [25]: %%sql
         SELECT BOOSTER_VERSION
         FROM SPACEXDATASET
         WHERE PAYLOAD_MASS__KG_ =
         (SELECT MAX(PAYLOAD_MASS__KG_)
         FROM SPACEXDATASET);

          * ibm_db_sa://1ff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
         Done.

Out[25]:
```

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are listed using the following query



```
In [29]: %%sql
SELECT BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXDATASET
WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

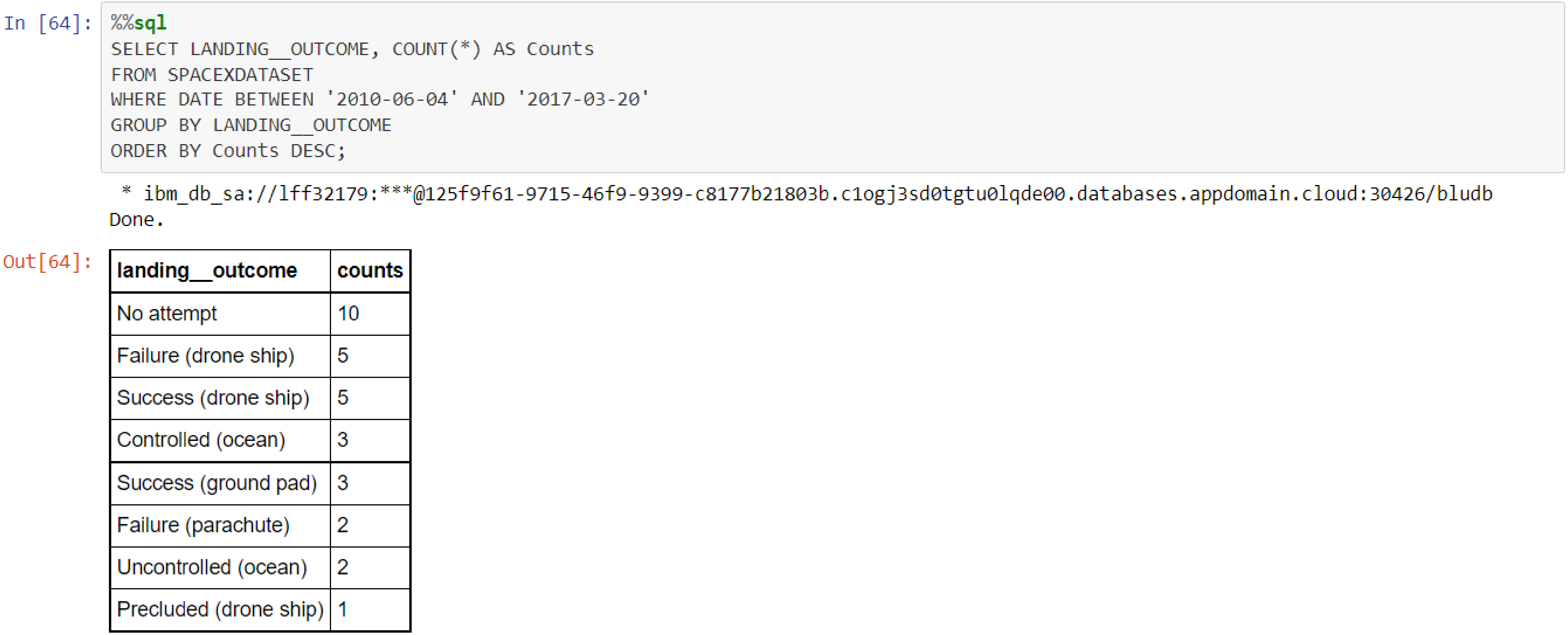
```
 * ibm_db_sa://lff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[29]:

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, are ranked in descending order using the following query



In [64]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(*) AS Counts
FROM SPACEXDATASET
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY Counts DESC;
```

```
 * ibm_db_sa://1ff32179:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/bludb
Done.
```

Out[64]:

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

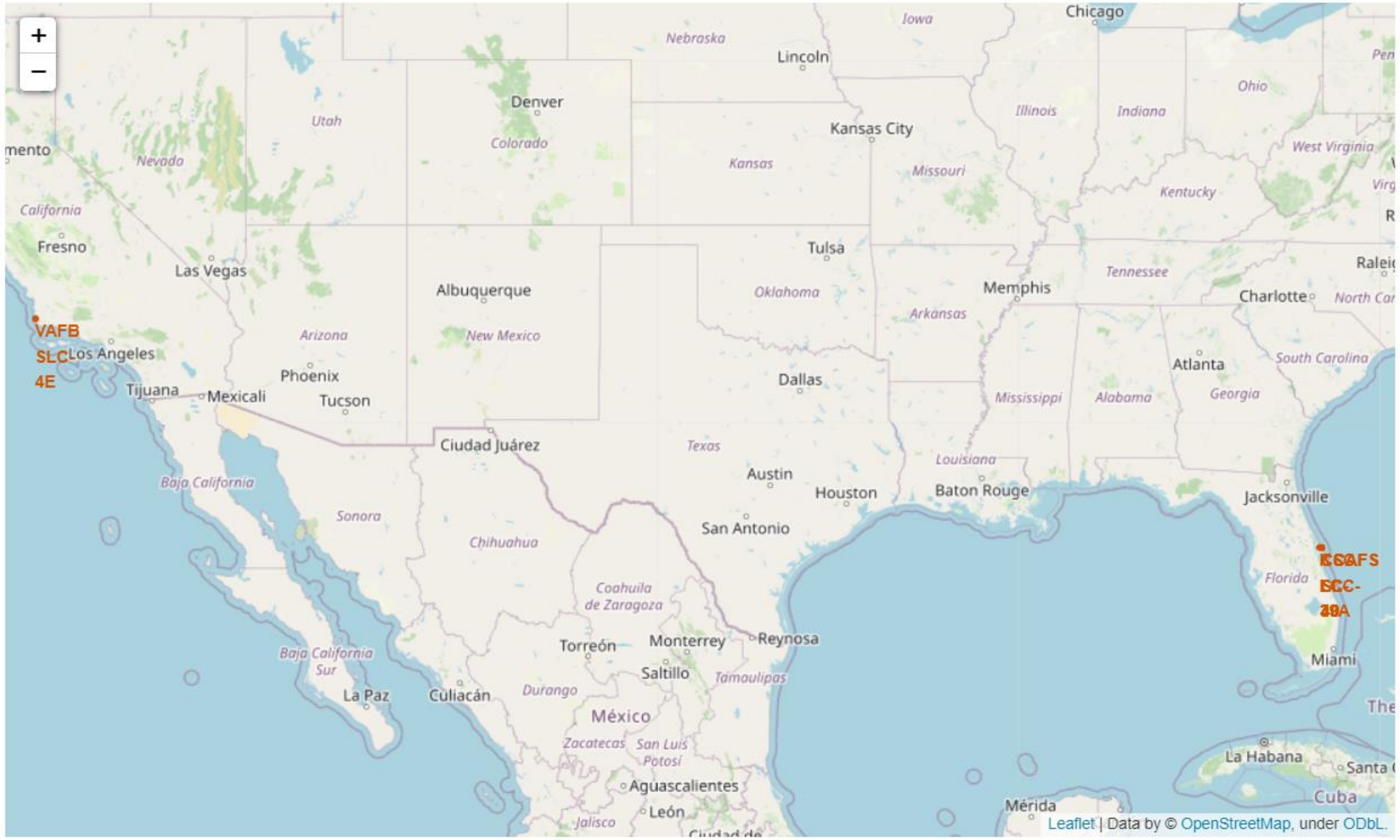
Section 3

Launch Sites Proximities Analysis

All launch sites' location markers on the map

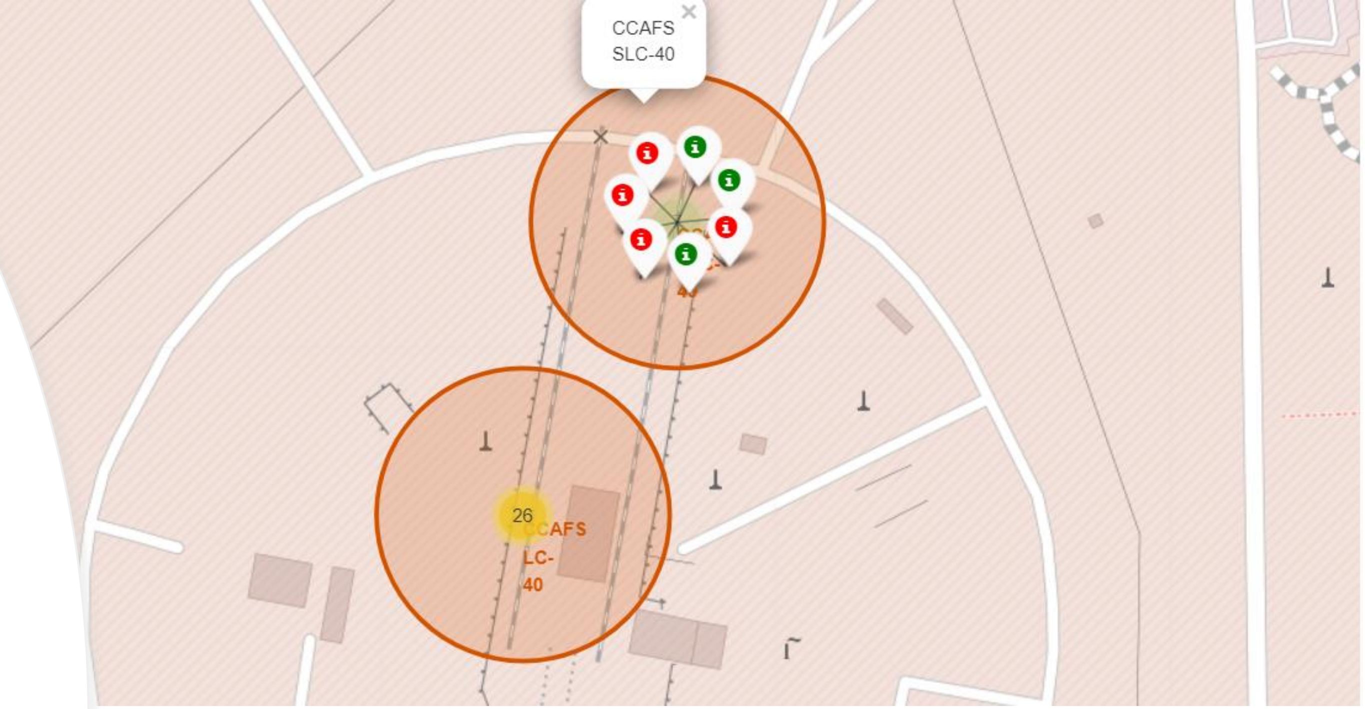
Observation:

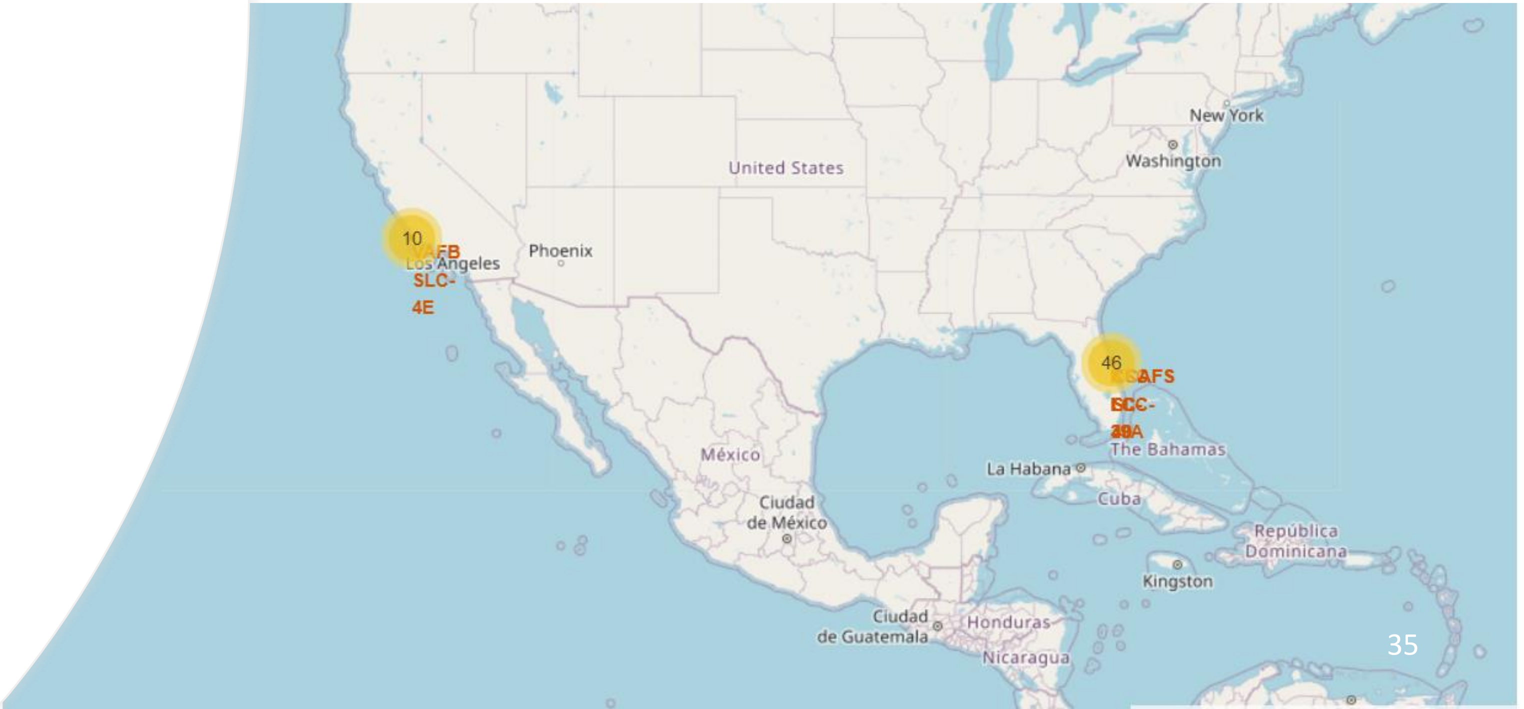
- All launch sites are in proximity to the Equator line



Color-labeled launch outcomes on the map

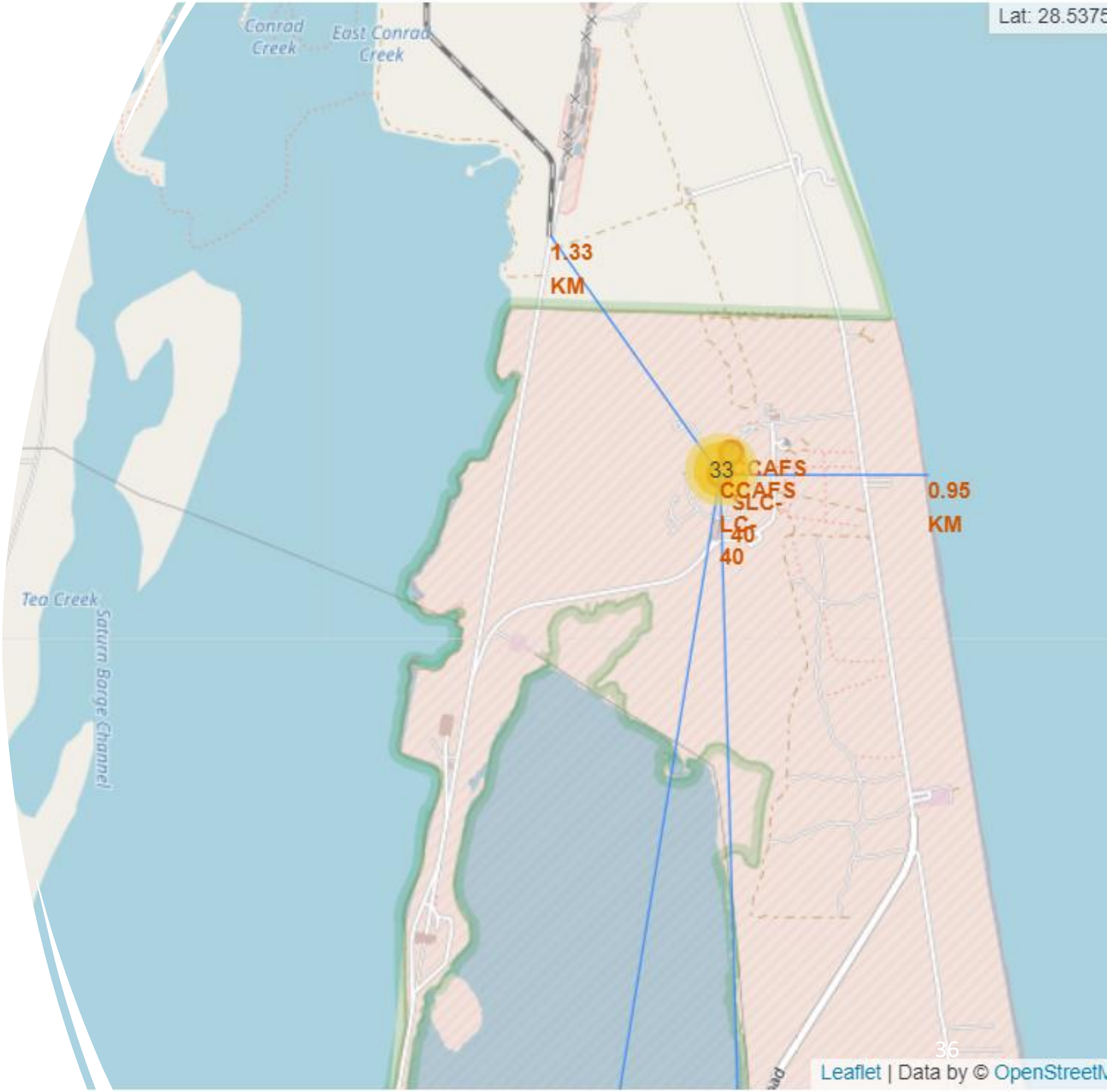
- The color-labeled markers in marker clusters makes it easy to identify which launch sites have relatively high success rates from
- For example, site **CCAFS SLC-40** has 3 success launches out of 7 launches in total

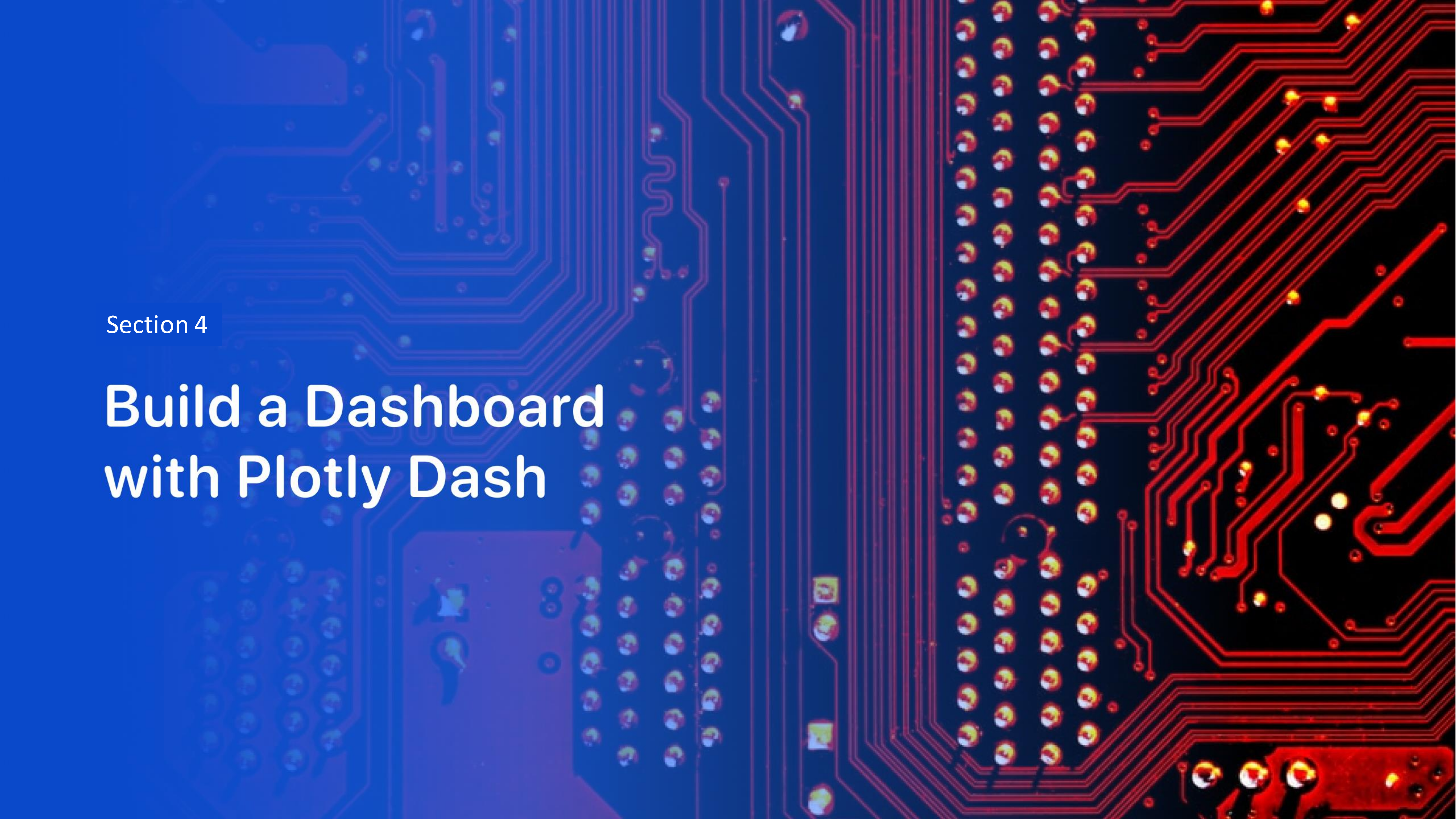




A selected launch site to its proximities on the map

- Launch site **CCAFS LC-40** to its proximities such as railway, highway, coastline, with distance calculated and displayed
- While it is close to the coastline and railways, the launch site keeps certain distance to highways and cities due to safety concern





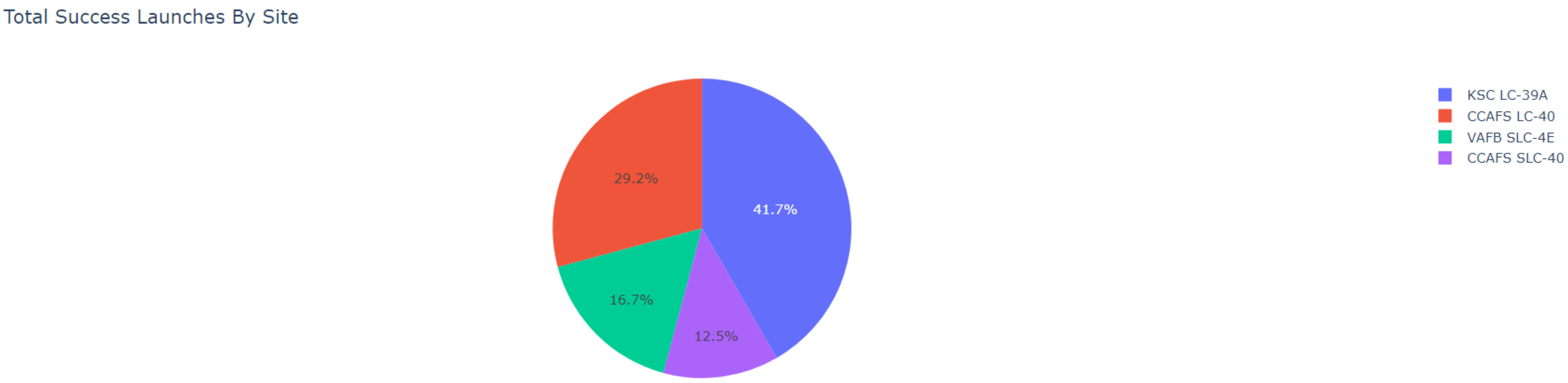
Section 4

Build a Dashboard with Plotly Dash

A pie chart of total success launches by site

Observation:

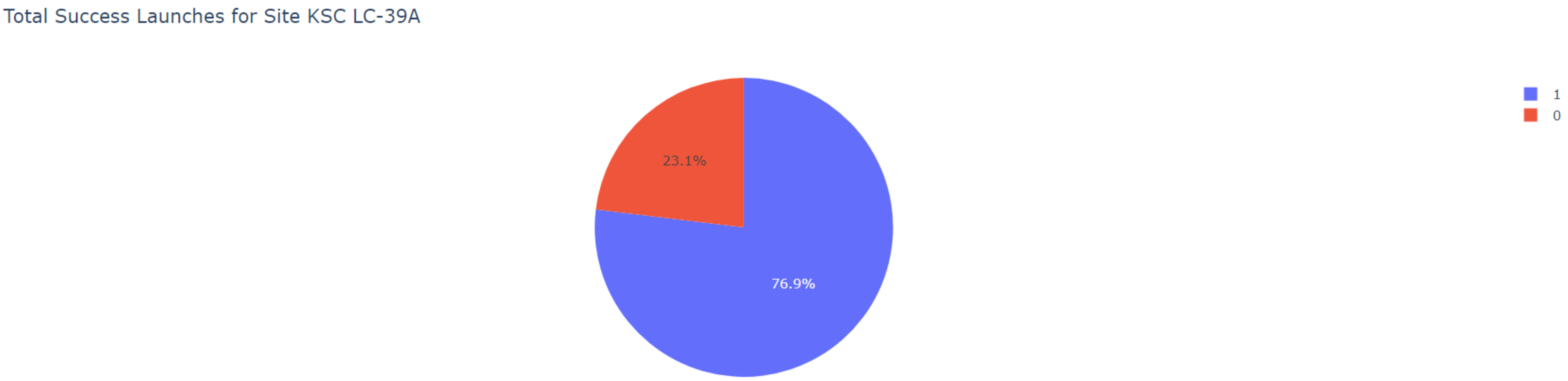
- The launch site that has the largest launch success rate is **KSC LC-39A**
- The launch site that has the lowest launch success rate is **CCAFS SLC-40**



A pie chart for the launch site KSC LC-39A (highest success launch rate)

Observation

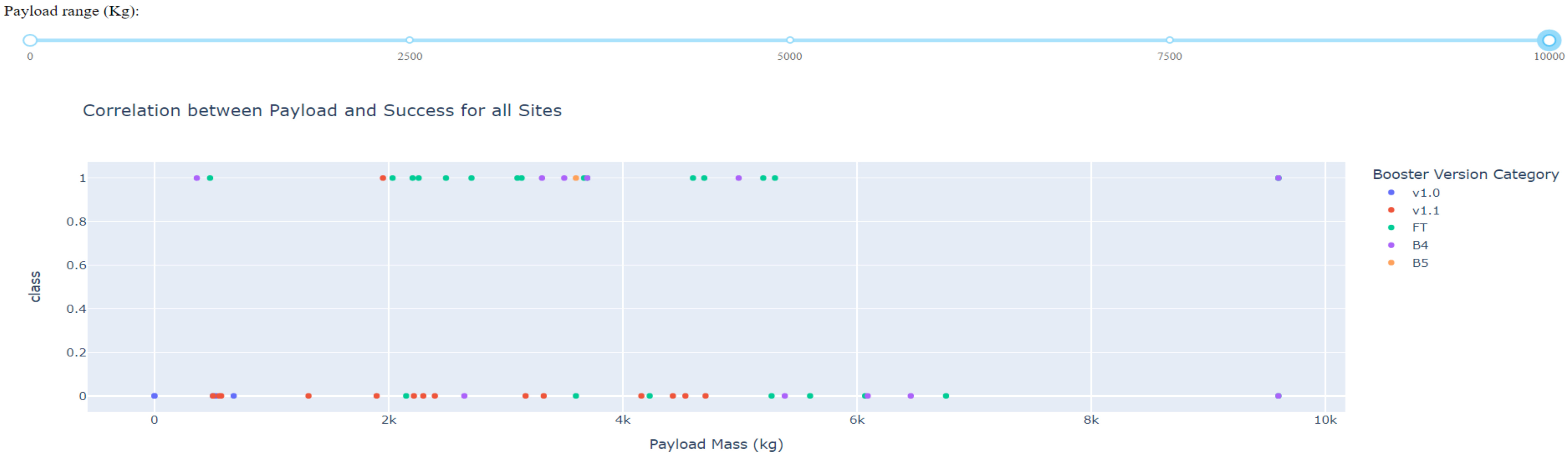
- The launch success rate for site KSC LKC-39A is 76.9% as shown from the pie chart



Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

Observation

- Payload range between **2000-4000kg** appears to have the highest launch success rate while **0-2000kg** has the lowest
- Booster version **FT** has the highest launch success rate

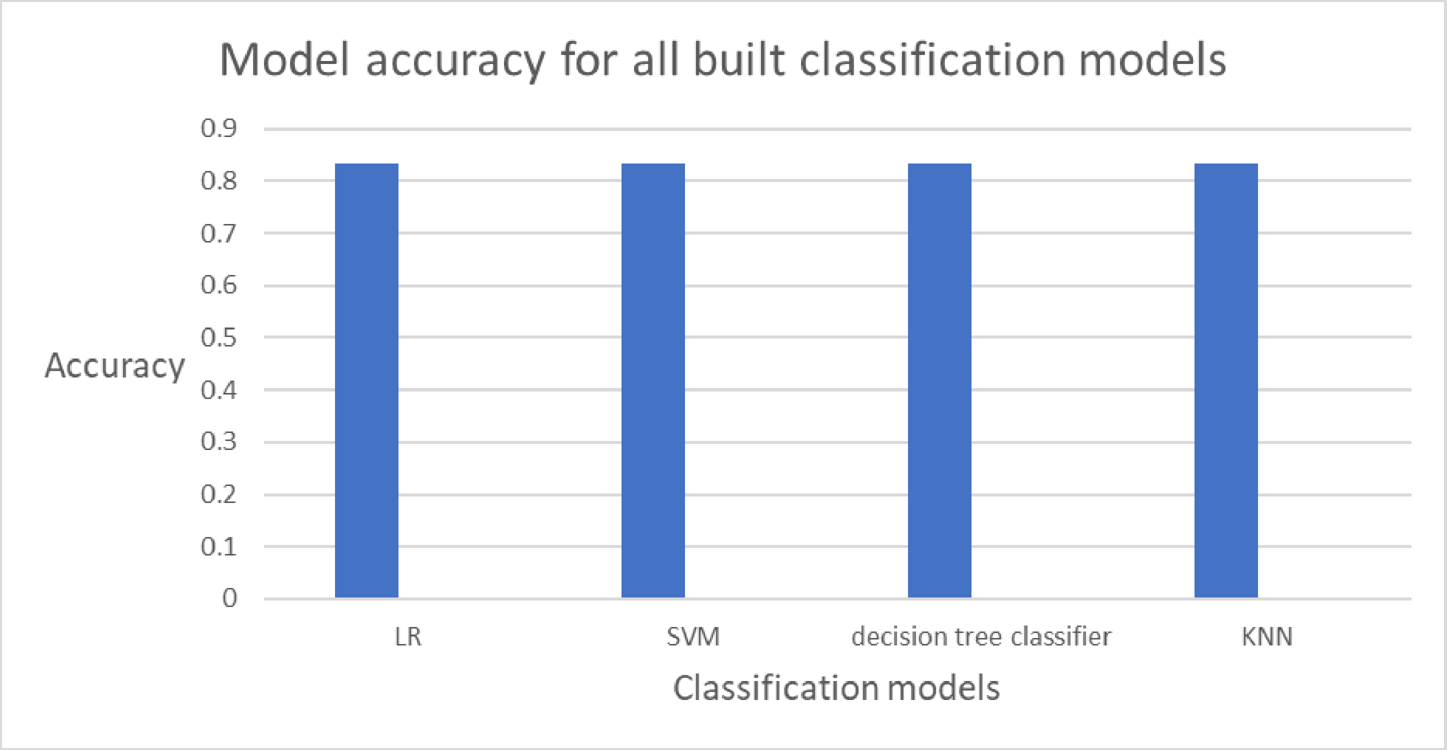


Section 5

Predictive Analysis (Classification)

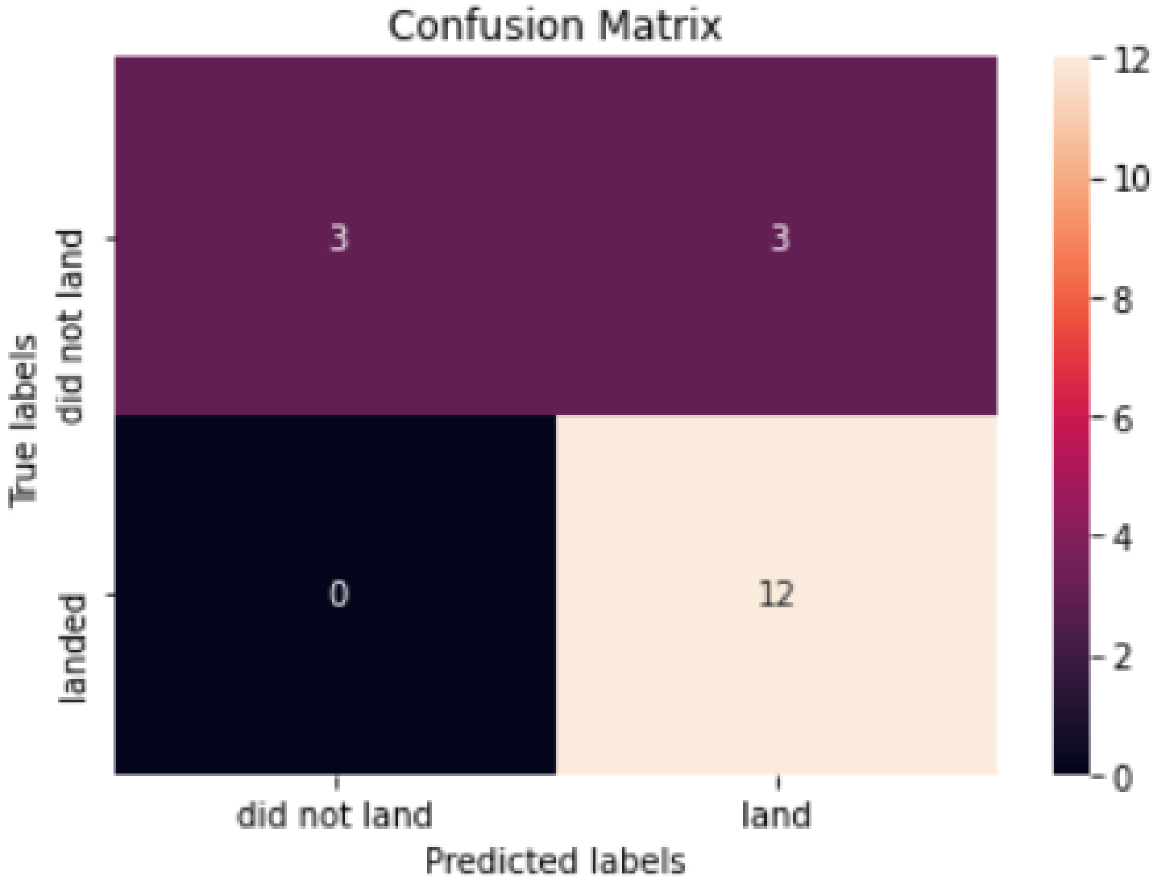
Classification Accuracy

- **All models** have the **same** accuracy which is 83.33%



Confusion Matrix

- The results are the **same** for **all the methods**
- The major problem are the false positives



Conclusions

- The number of flights and payload may be a factor in launch success
- Launch success rate differs in different orbits; Highest in orbits **ES-L1, GEO, HEO, SSO;** Lowest in orbits **GTO**
- The average launch success rate for SpaceX has been improving since 2013
- **KSC LC-39A** is the launch site with the highest launch success rate so further research could be conducted to find out the reason of it
- Payload range between **2000-4000kg** and booster version **FT** have the highest success rate so further research could be conducted to find out the reason of it and use as a reference
- All built classification models have an accuracy of 83% and there is room for improvement by dealing with the problem of false positives

Thank you!