Data Science Capstone project

Shubhadeep Sarkar

13th September, 2021

Outline

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

Executive Summary

- Methodologies:-
  1. Web Scrapping and Data Wrangling
  2. EDA with SQL and visualization with Python
  3. Dashboard with plotly
  4. Predictive Analysis with ML

- Summary of all results

Introduction

There’s no company in the world which haven’t heard of SpaceX and the feats it achieved. The most notable achievement of SpaceX is that they are the first ones to have successfully landed a launched rocket back on the ground and managed to reuse the same old part for the next venture. So if any alternate company that wants to compete them needs recommendation data as on how successful they might be, cause after all no one wants to huge amount of money on failures.

So we will predict if the Falcon 9, the rocket which SpaceX managed to land back successfully, in it’s 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.

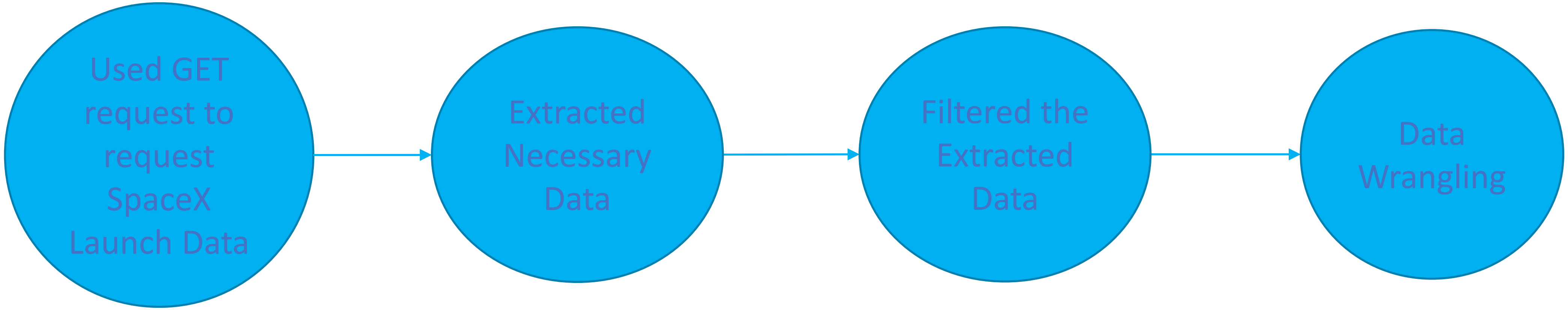
Methodology

- Data collection methodology:
  - After making a GET request to the SpaceX API and cleaned the data. Then requested Falcon 9 wiki Page and used BeautifulSoup to extract necessary Data and tables from the webpage.
- Perform data wrangling
  - Once the Data was collected, Data Frames were made from the collected data, the data was reformatted to ensure homogeneity in the same Columns and the missing vales were dealt with appropriate methods.
- 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

Methodology

Data collection

- The Data was collected from do different Sources. In the first one a request was made to the SpaceX API, and then the Data for the Data Frames was collected. As for the second one, a request was made for the Falcon9 Launch HTML page as an HTTP response, from there the data and tables were extracted using BeautifulSoup. After the extraction of data, certain operations known as Data Wrangling were made to filter the data and find certain fields.

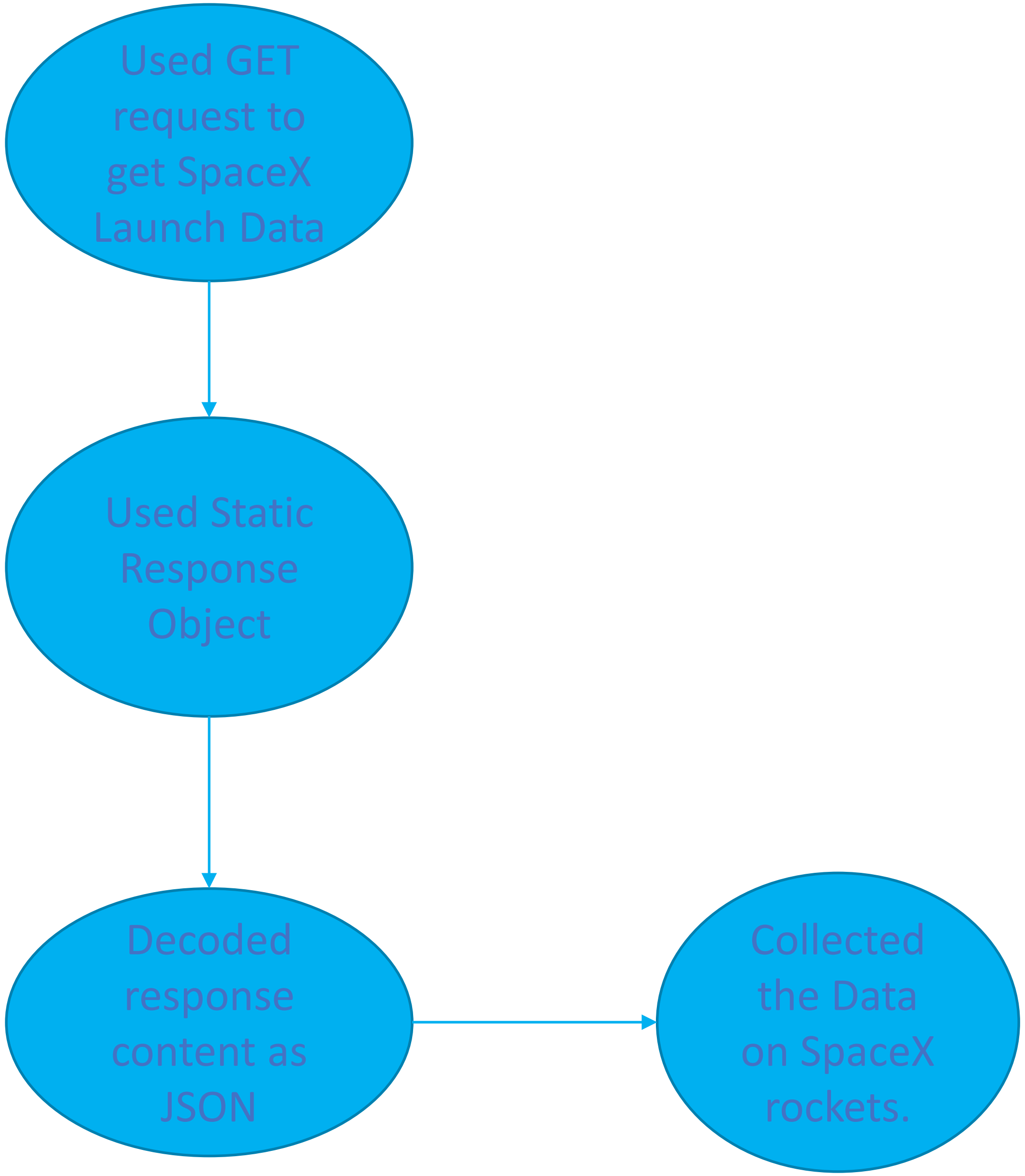

Used GET request to request SpaceX Launch Data
Extracted Necessary Data
Filtered the Extracted Data
Data Wrangling

Data collection – SpaceX API

- Made a get request to the SpaceX API.
- Made static response object to make the requested JSON result more consistent.
- Data required (like rocket data, launchpad data, payload mass data, etc.) were extracted.
- After extracting the data, data frames was created using Pandas.

Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_Capstone_Project_VzjrkHTzY.ipynb



Used GET request to get SpaceX Launch Data

Used Static Response Object

Decoded response content as JSON

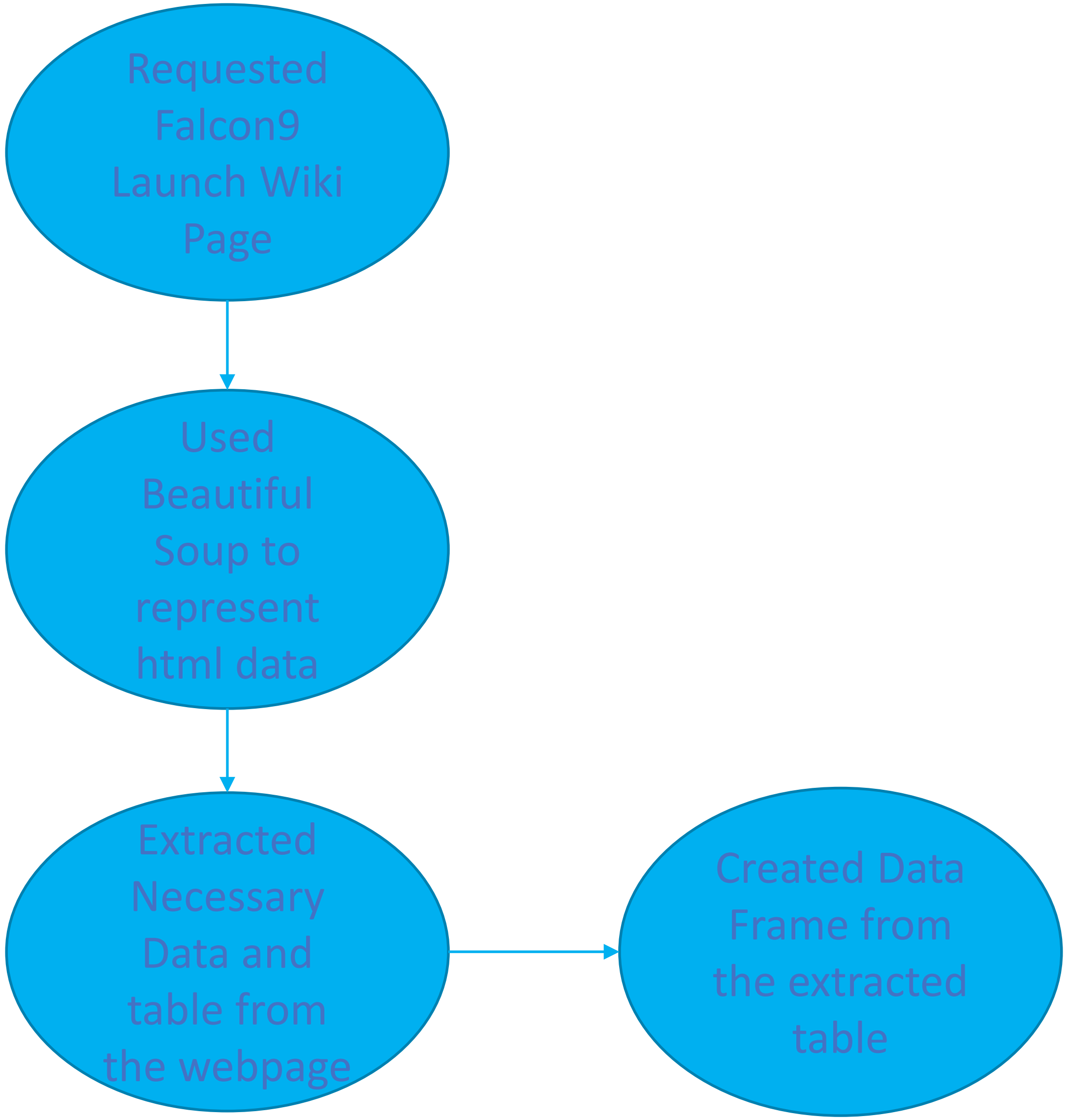
Collected the Data on SpaceX rockets.

Data collection – Web scraping

- HTTP GET method was used to request the Falcon9 Launch HTML page as an HTTP response.
- A BeautifulSoup object was created from the HTML response.
- Columns required (like Flight No., Launch Site, Payload, Payload Mass, etc.) were extracted from the tables present on the HTML response.
- Data Frame was created from the data extracted from the tables using pandas.

Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_Applied_Data_Science_Capstone_ktoVfTUAx.ipynb



Requested Falcon9 Launch Wiki Page

Used Beautiful Soup to represent html data

Extracted Necessary Data and table from the webpage

Created Data Frame from the extracted table

Data wrangling

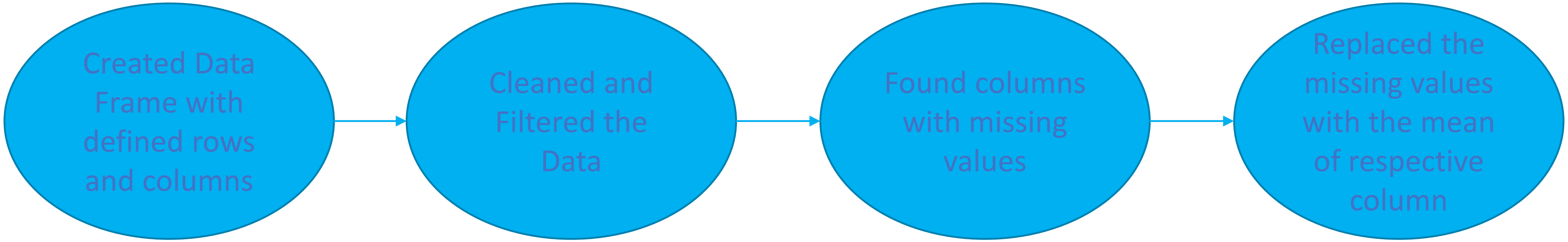
Once the collection of Data was finished, the Data was then filtered. The extracted data contained records of Falcon9 and other rockets. So the record of other rockets were filtered out since we will be dealing with only the launch data of Falcon9. After filtering out the launch data of Falcon9, columns were identified which had missing values, which came out to be Payload mass and Landing pad. But the Landing pad will retain the none values (default value for missing) as it denoted that the launch pad was not used. So only the Payload mass column was dealt with replacing the missing values with mean Payload Mass.

After the missing values in the Data Frames have been dealt, the following were calculated-

1. Number of Launches on each Site
2. Number and Occurrence of each orbit
3. Number and Occurrence of mission outcome per Orbit Type
4. Finally created landing outcome label called ‘Class’ from outcome column.

- Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_Capstone_Bk77PTKe1.ipynb



Created Data Frame with defined rows and columns

Cleaned and Filtered the Data

Found columns with missing values

Replaced the missing values with the mean of respective column

EDA with data visualization

- Summarize what charts were plotted and why used those charts

For data visualization, 6 scatterplot, 1 bar graph and 1 line graph were plotted using the collected and filtered data. These graphs help in the visualization of how each criteria is related to the launch outcome and are easy to understand. The last graph, the line graph allows us to compare the success rate of each year.

- Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_EDA_part-2_aAf7UjQZm.ipynb

EDA with SQL

- Summarize performed SQL queries using bullet points
- The data frames were first uploaded to IBM Database, from where we accessed them using Jupyter Notebook.
- The list of queries performed using SQL are:-

1. Displayed the names of unique launch sites in the space mission
2. Displayed 5 launch sites where their names begin with ‘CCA’
3. Displayed total payload mass carried by boosters launched by NASA (CRS)
4. Displayed average payload mass carried by booster version F9 v1.1
5. Listed the dates when successful landing outcome in ground pad was achieved
6. Listed the names of the boosters with success in drone ship and have payload mass >4000 and <6000
7. Listed total number of successful and failure mission outcomes
8. Listed the names of the booster versions which have carried the maximum payload mass
9. Listed the records which display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015
10. Ranked count of successful landing outcomes between 4th June, 2010 and 20th March, 2017 in descending order

- Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_EDA_with_sql_9HBNuQQp2.ipynb

Build an interactive map with Folium

- First we marked all the sites with a red color circle which look like a red dot when zoomed out sufficiently along with their names.
- After all the sites were marked, we marked all the successful and failed launches for each site on the map.
- Each successful launch is indicated by a green marker while a failed launch is indicated by red marker.
- Since many of these markers coincided at the same coordinates, marker clusters was created. It easily simplified the map.
- Then finally distance between launch sites and their proximities were calculated. The shortest distance is represented by a green line denoting the proximity coordinates with a red marker. This showed which proximities were closest to the space stations.
- Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_Dashboard_9CPtSq2Cx.ipynb

Build a Dashboard with Plotly Dash

- A dashboard is also included which shows the success and failure pie chart for each and all launch site, which can be selected from the dropdown menu. A scatter plot was also included with slider to plot the chart between a range of values.
- The pie chart shows the success rate of the missions carried out between all the launch sites and for each individual site, it shows the success and failure rate. This allows us to easily identify which site has most successful launch rate compared to the others and also tells us about their failure rates.
- The scatter plot tells us about the payload mass (in kgs) of all the successful launches among all the sites, and if individually selected, it tells us about the payload mass (in kgs) of each successful and failed launch.

- Link:-

https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/dashboard_Python/spacex_dash_app_code.py

Predictive analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- After importing all the important libraries required, first a function was created which plots our confusion matrix. Once the data is loaded, the class column is converted to NumPy array and assigned to a variable Y. The data in X is standardized and reassigned to itself. Now we split the data in X and Y into training and testing data, after which we find the best Hyperparameter for SVM, Classification Trees and Logistic Regression. At the end , we find the best method by using the test data.
- Link:-

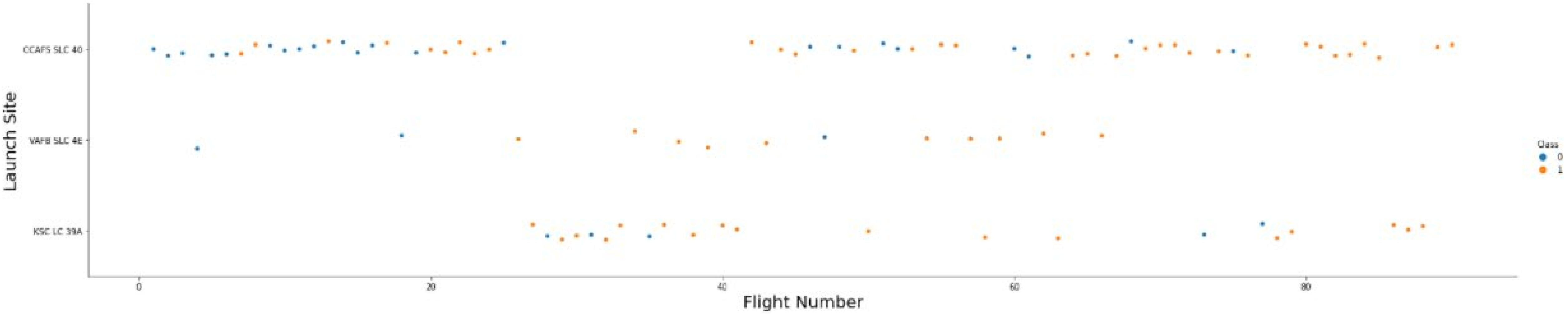
https://github.com/Osir1s-spec/coursera_capstone_project/blob/main/assets/notebook/notebook_ML_4zdma4gG3.ipynb

Results

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

EDA with Visualization

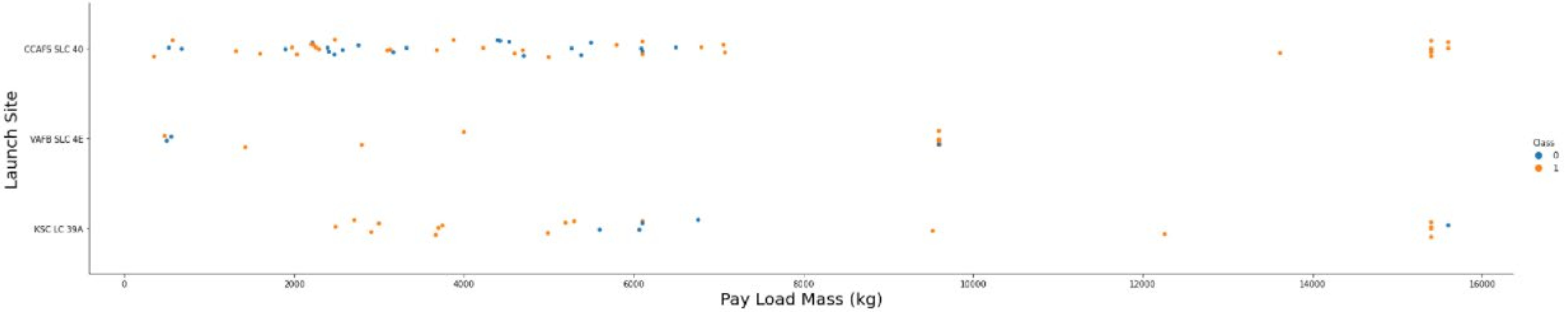
Flight Number vs. Launch Site



The scatter plot above shows all the flight number at each launch site where blue color is for class 0 and orange color is for class1.

From the plot we can see that there are 3 launch site in total, with Launch site CCAFS SLC 40 being used the most number of time.

Payload vs. Launch Site



The scatter plot above show us the payload mass in kg carried in each mission conducted at each site with color blue showing class 0 and color orange showing class 1.

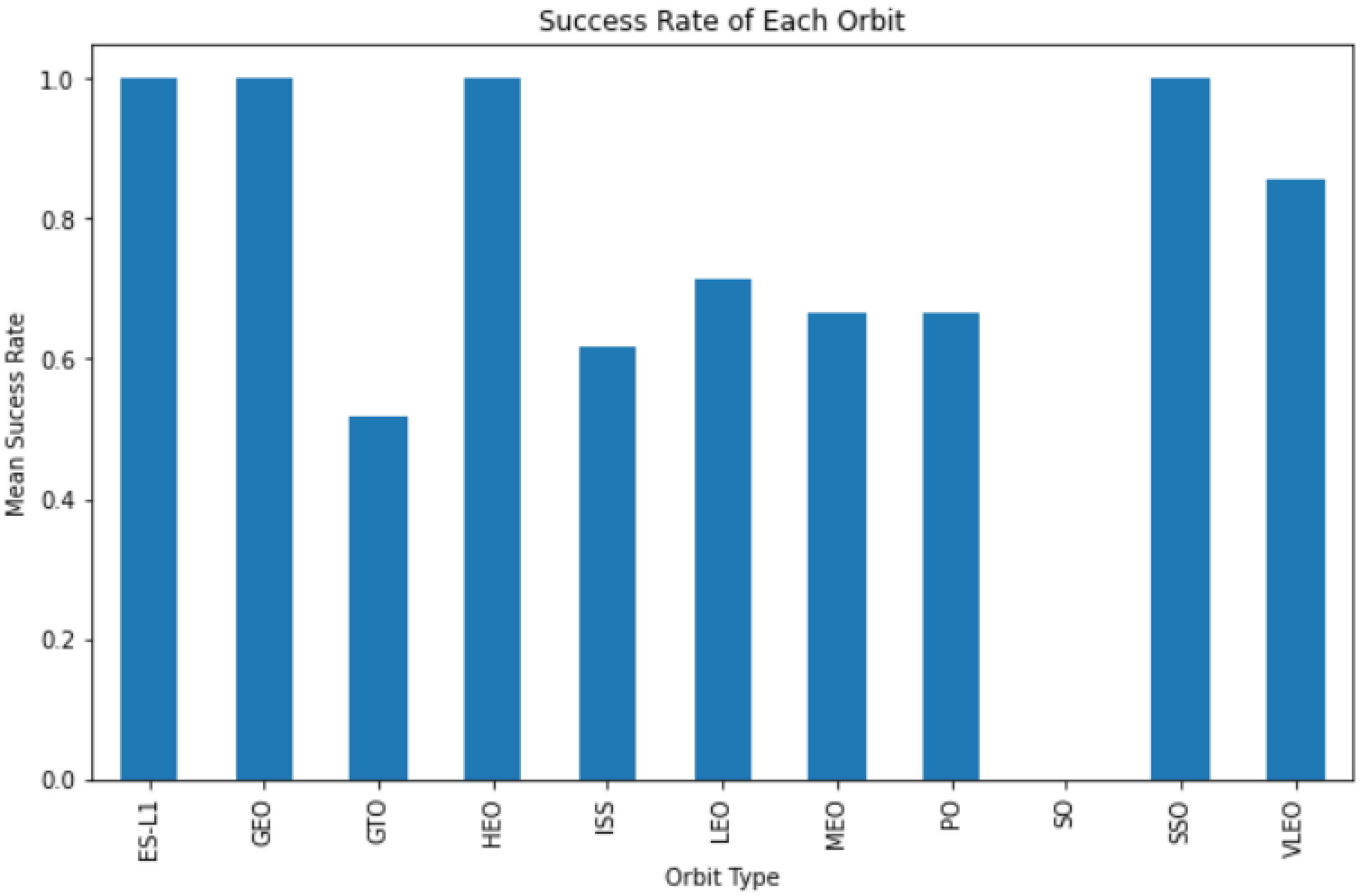
From the above plot it can be seen that be for payload with mass between 2000 kg to 8000 kg, launch site CCAFS SLC 40 has been used the most and for payload above 14000 kg, launch site CCAFS SLC 40 and launch site KSC LC 39A both has been used. Whereas for payload between 8000 kg and 10000 kg, launch site VAFB SLC 4E has been used.

Success rate vs. Orbit type

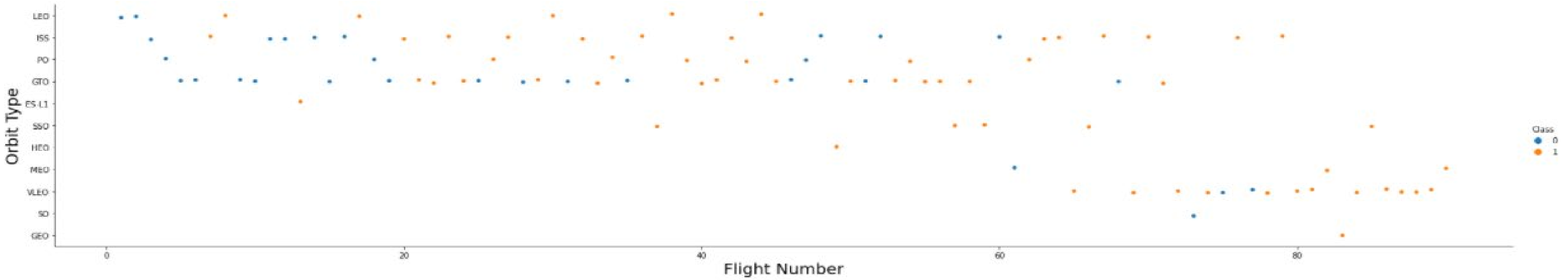
The bar chart shows the mean success rate of all the missions carried out for each orbit type.

From the bar we can find that orbits ES-L1, GEO, HEO and SSO have the highest mean success rate.

Orbit SO has nearly 0 mean success rate, which can be either because no mission for that orbit were carried out or all the missions carried out for this orbit is zero.



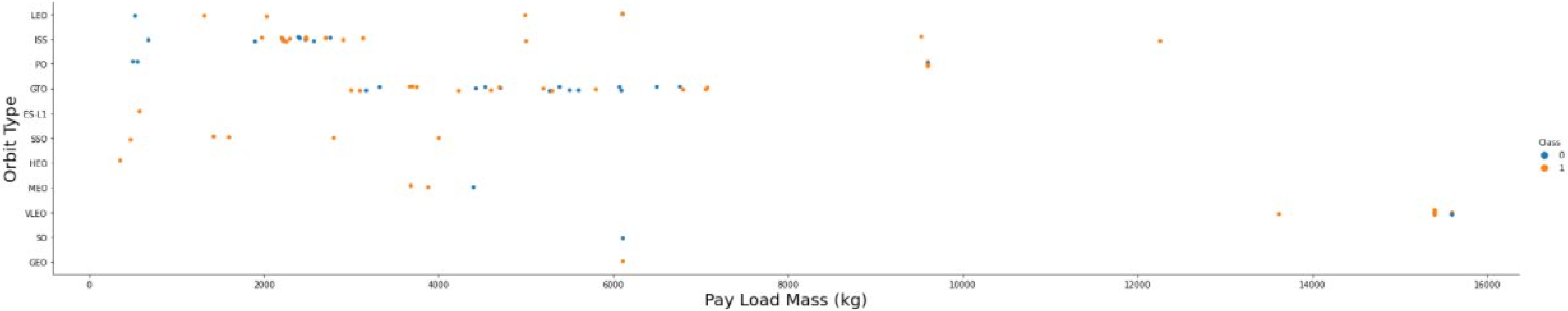
Flight Number vs. Orbit type



The scatter plot above shows the flight number of each orbit type mission carried out, where blue depicts class 0 and orange depicts class 1.

From the scatter plot we find the most of the initial flights were meant for orbit LEO, ISS, PO and GTO, where as the most recent flights were meant for GVLEO orbit.

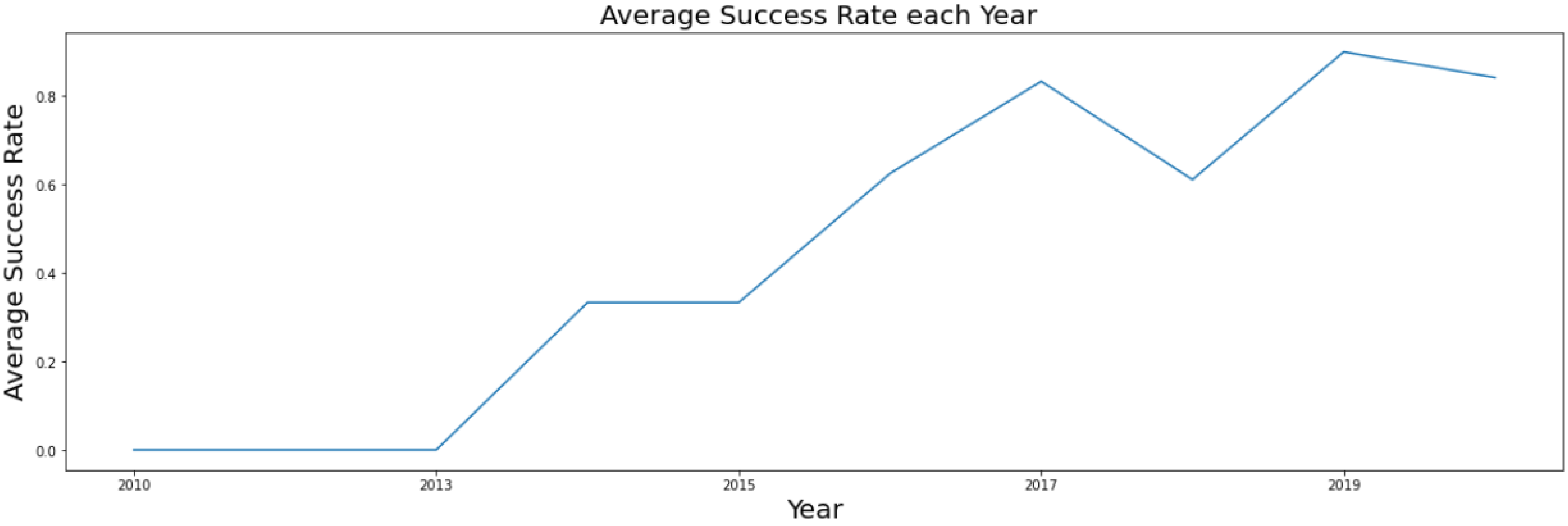
Payload vs. Orbit type



The above scatter plot shows the payload mass in kg delivered for each orbit type, where blue color depicts class 0 and orange color depicts class 1.

From the above scatter plot we can find that most of lighter payload (between 0 and 8000 kg) were carried for ISS and GTO orbits where as the heaviest payload (above 14000 kg) were carried out for VLEO orbit.

Launch success yearly trend

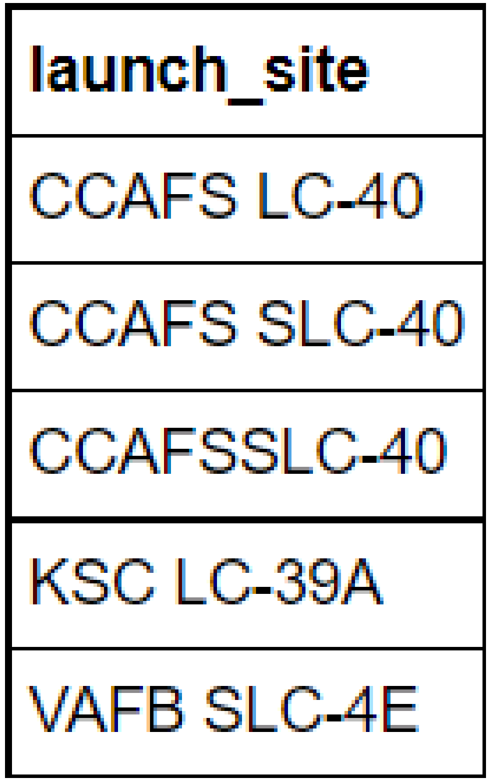


The above line graph shows the success rate during each year.

From the line graph it is seen that success rate gradually increased from 2013 to 2017, going from 0 to 0.8, and then dropped in 2018 to 0.6, and finally reached success rate above 0.8 in 2019.

EDA with SQL

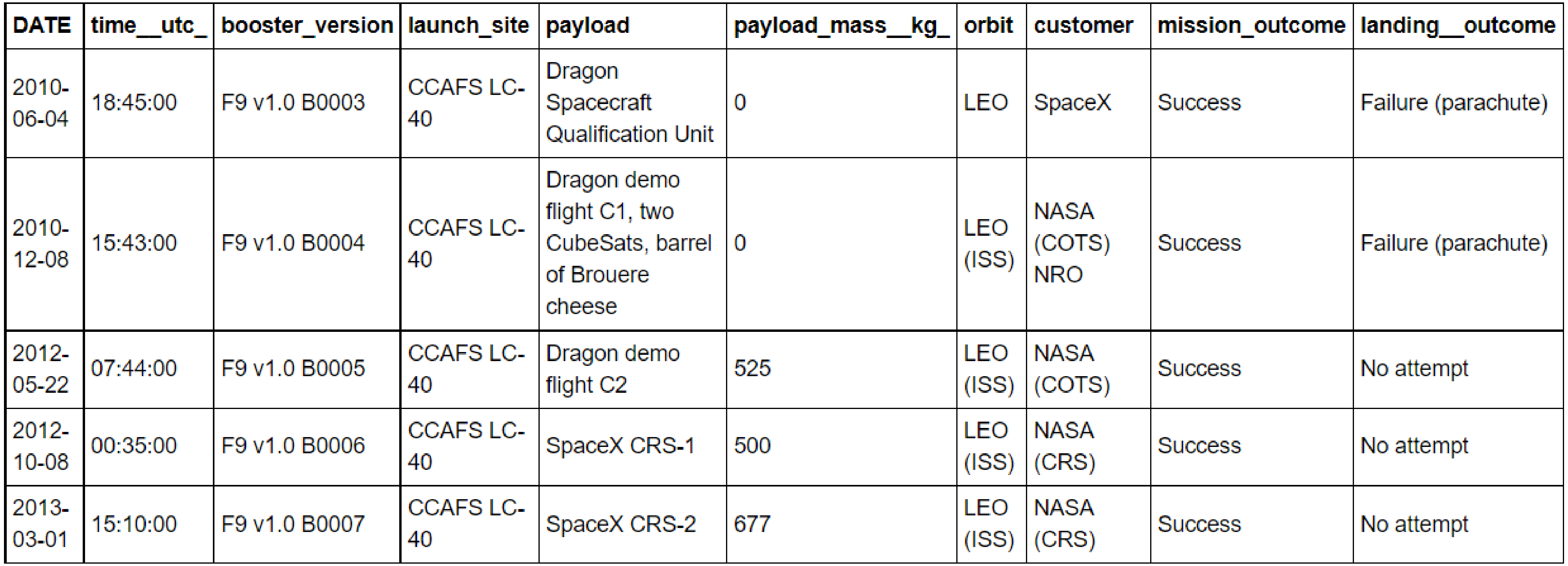
All launch site names



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

- This is a list of all the launch sites used by SpaceX.

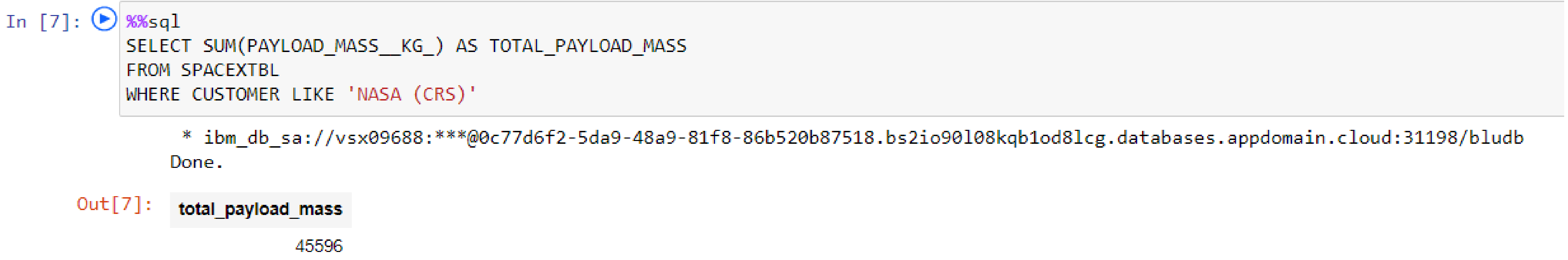
Launch site names begin with `CCA`



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

- The above table shows data on all the launch sites used by SpaceX whose name start with ’CCA’.

Total payload mass



In [7]:

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTBL
WHERE CUSTOMER LIKE 'NASA (CRS)'
```

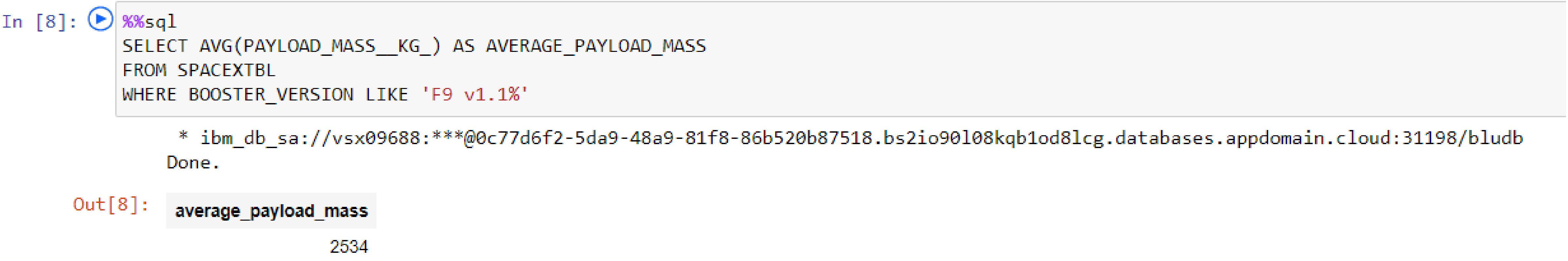
```
 * ibm_db_sa://vsx09688:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

total_payload_mass
45596

- The above query shows the total payload mass delivered by NASA, which is 45596 kg.

Average payload mass by F9 v1.1



```
In [8]: %%sql
        SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS
        FROM SPACEXTBL
        WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```

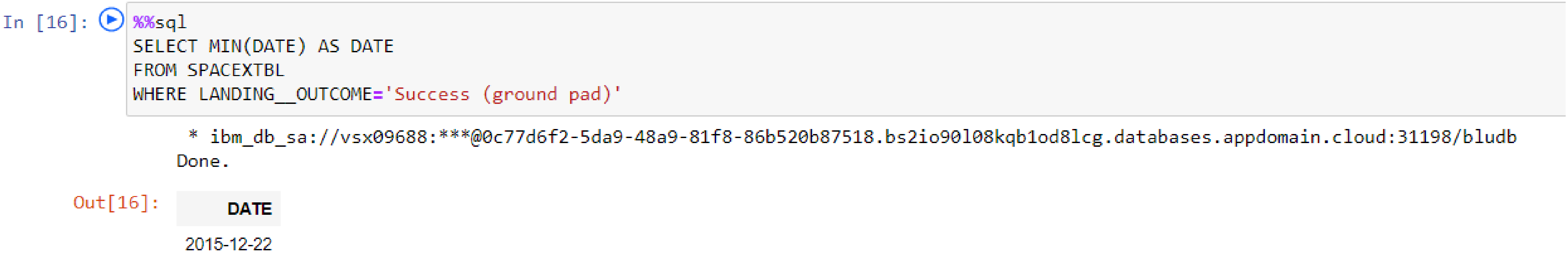
```
 * ibm_db_sa://vsx09688:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

average_payload_mass
2534

- The above query shows the average payload mass carried by booster version F9 v1.1, which is 2,534 kg.

First successful ground landing date



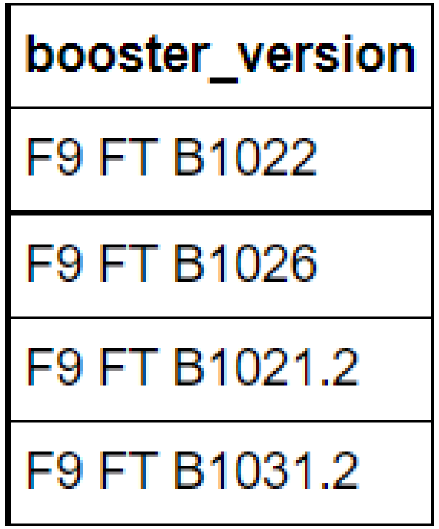
```
In [16]: %%sql
SELECT MIN(DATE) AS DATE
FROM SPACEXTBL
WHERE LANDING__OUTCOME='Success (ground pad)'

 * ibm_db_sa://vsx09688:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[16]:

DATE
2015-12-22

- The above query shows the date of first successful landing outcome in ground pad, which is 22nd December, 2012.

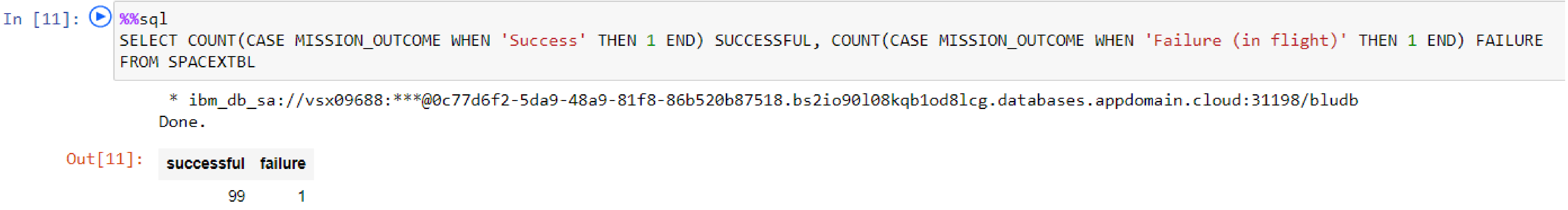
Successful drone ship landing with payload between 4000 and 6000



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

- This is a list of boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 600 kg.

Total number of successful and failure mission outcomes



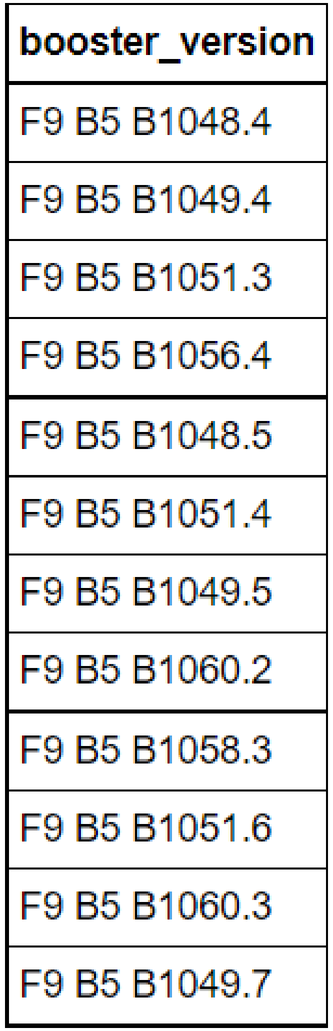
```
In [11]: %%sql
SELECT COUNT(CASE MISSION_OUTCOME WHEN 'Success' THEN 1 END) SUCCESSFUL, COUNT(CASE MISSION_OUTCOME WHEN 'Failure (in flight)' THEN 1 END) FAILURE
FROM SPACEXTBL
```

```
 * ibm_db_sa://vsx09688:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

```
Out[11]:
successful  failure
        99        1
```

- The above query shows the total number successful and failure mission outcome.
- From the query we find that a total of 100 missions were carried out by SpaceX whose success and failures were determined.

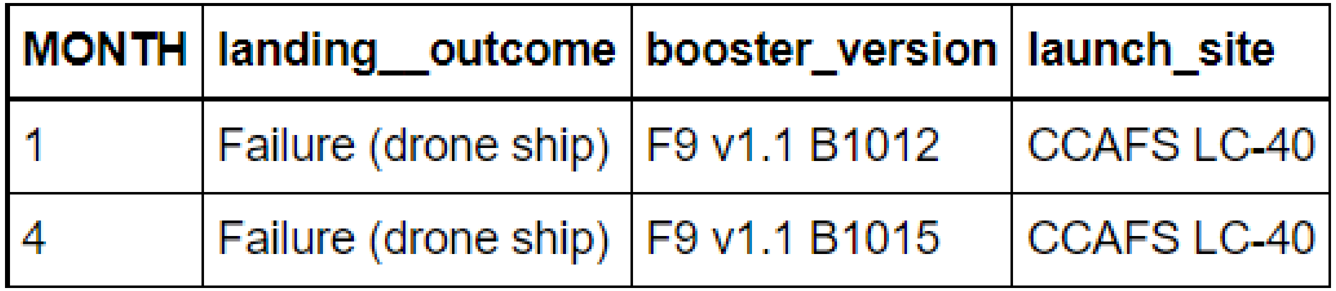
Boosters carried maximum **payload**



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

- This is a list of all booster versions which have carried maximum payload.

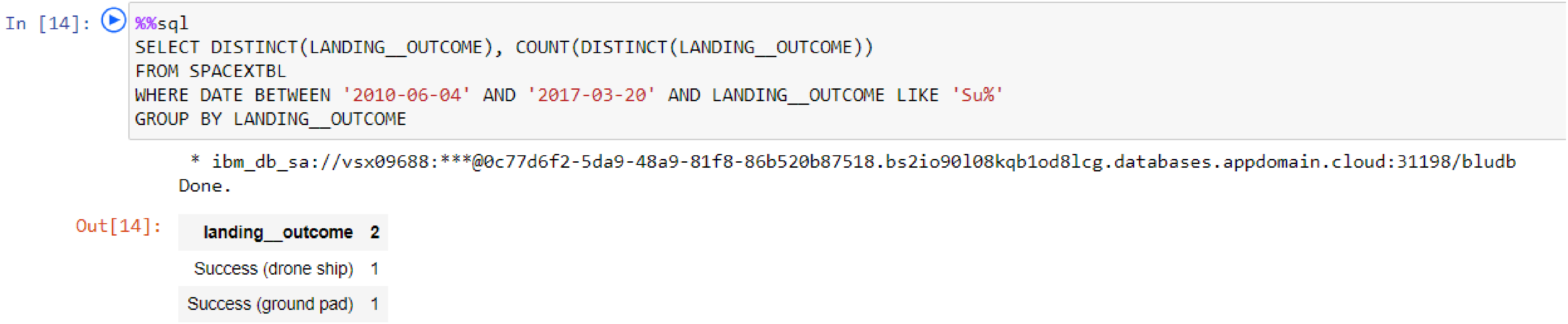
2015 launch records



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

- This is the launch record of boosters which had failure landing outcome on drone ship in the year 2015.
- From the list we can see that the failed outcomes were carried out in the month of January and April and both of them were launched from the same site.

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



In [14]:

```
%%sql
SELECT DISTINCT(LANDING__OUTCOME), COUNT(DISTINCT(LANDING__OUTCOME))
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__OUTCOME LIKE 'Su%'
GROUP BY LANDING__OUTCOME
```

```
 * ibm_db_sa://vsx09688:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

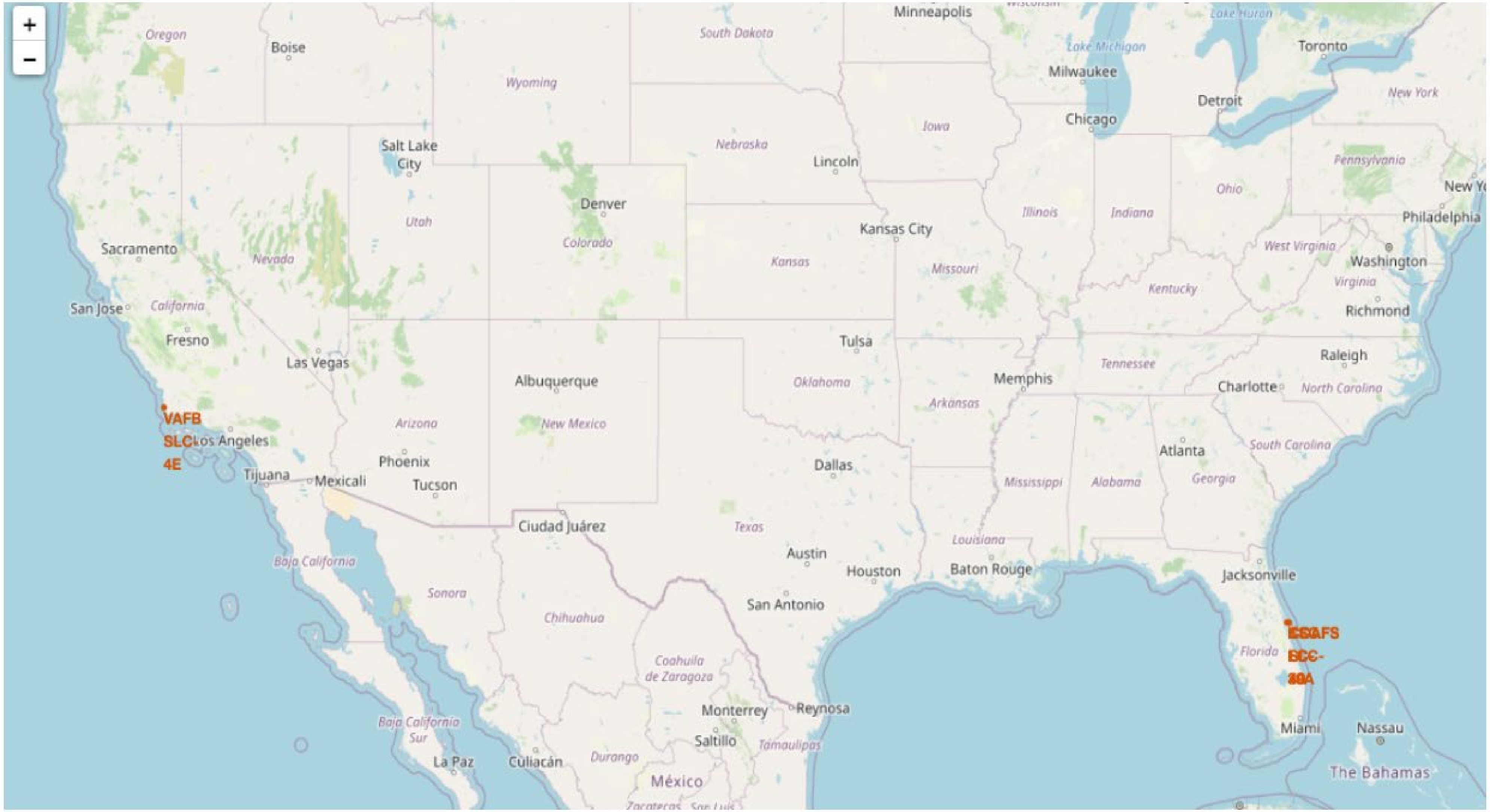
Out[14]:

landing__outcome	2
Success (drone ship)	1
Success (ground pad)	1

- The above query shows the rank of count of successful landing outcomes between the date 4th June, 2010 and 20th March, 2017.

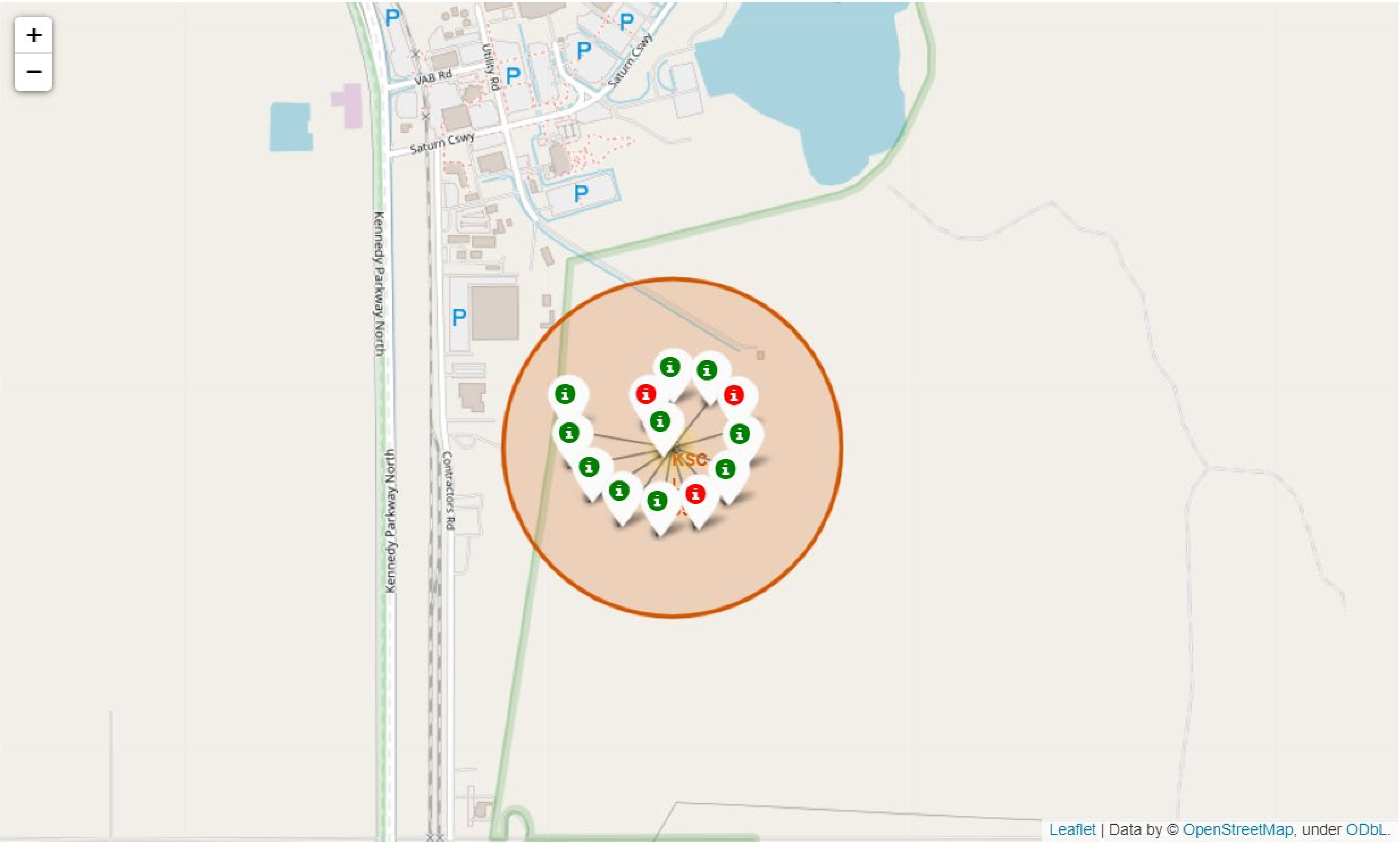
Interactive map with Folium

<Folium map screenshot 1> replace



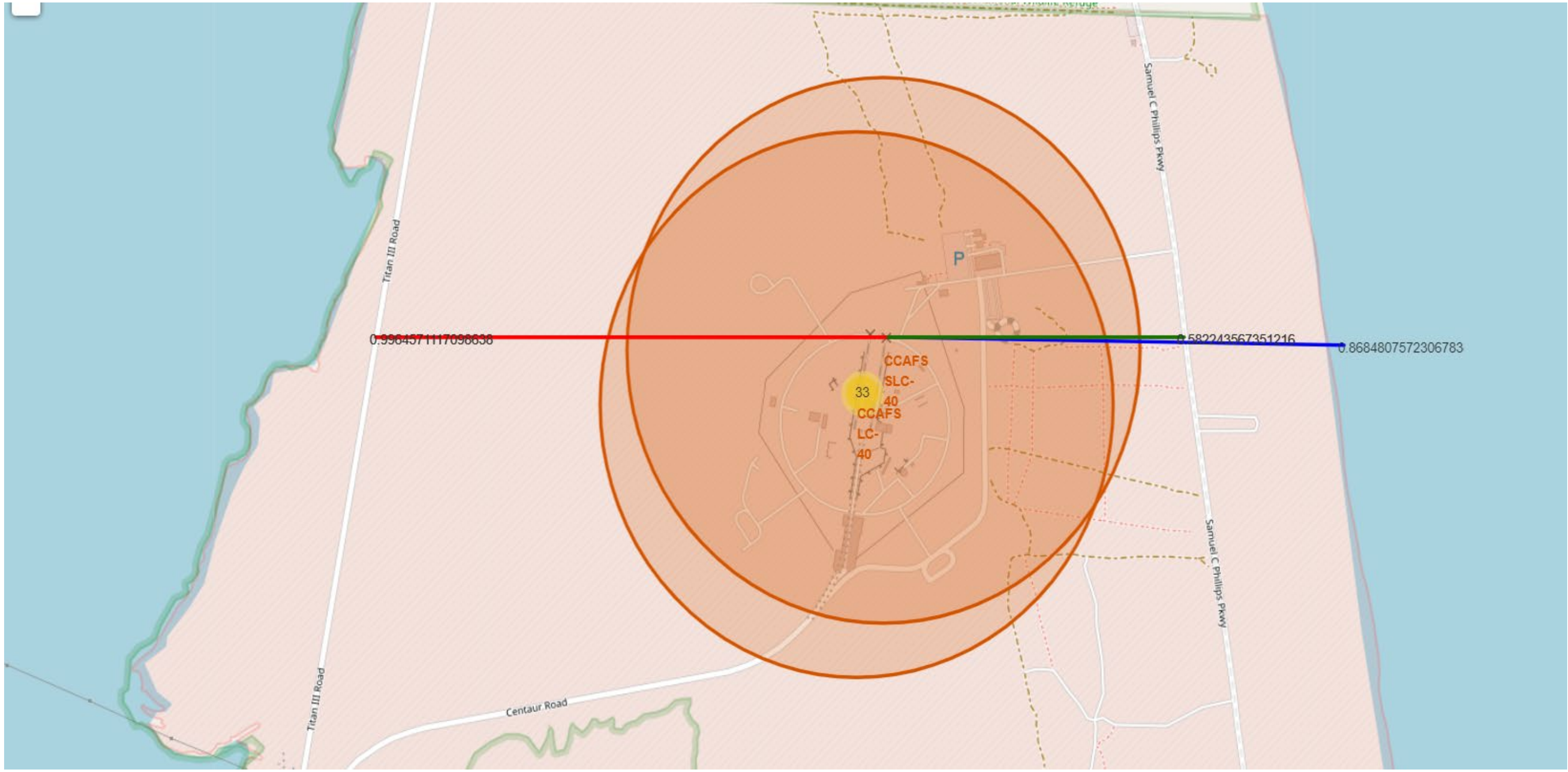
This map shows all the launch sites used by SpaceX, denoted by red color.

<Folium map screenshot 2> replace



- Upon clicking each individual map, it will show markings shown in the picture. Here we show the markings of launch site KSC LC-39A.
- These marking show the class of each mission where green color denote class 1 and red color denote class 0.

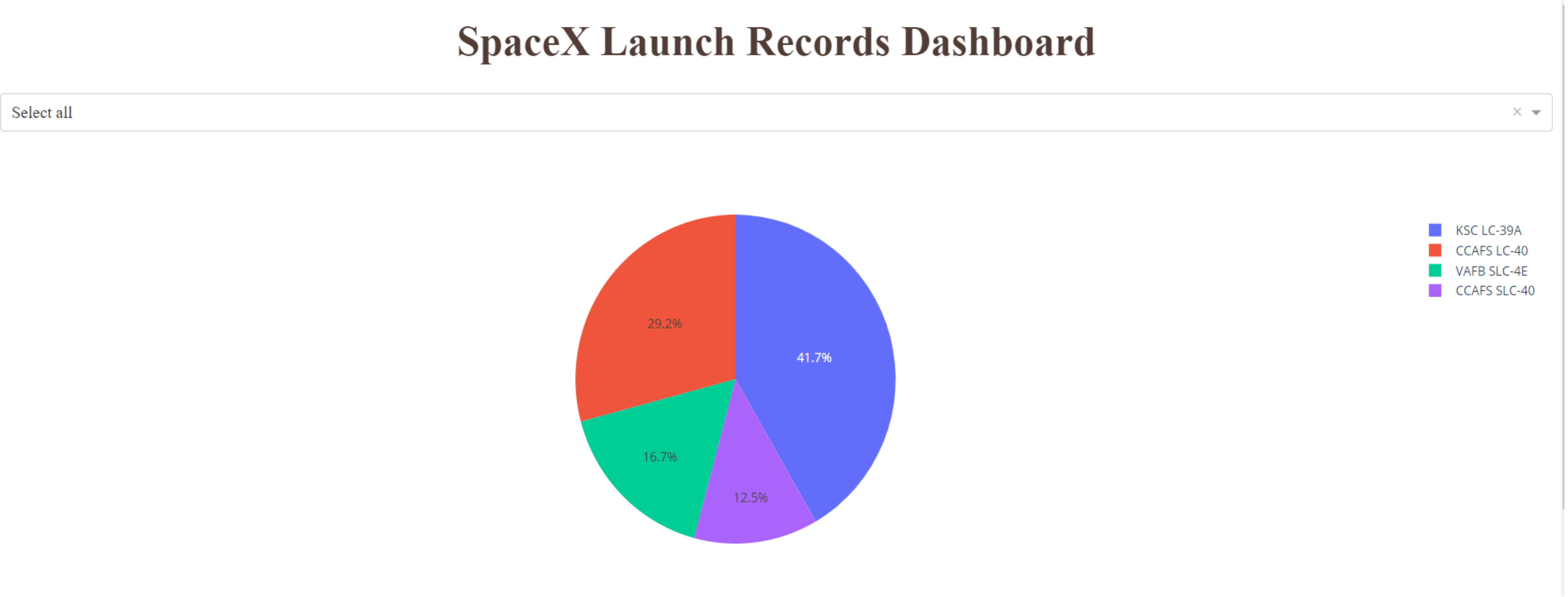
<Folium map screenshot 3> replace



- As shown in the picture, proximities of launch site CCAFS SLC 40 were found.
- The red line show the closes highway, blue line shows the closest coastline and the green line shows the closest railway line.

Build a Dashboard with Plotly Dash

<Dashboard screenshot 1> replace

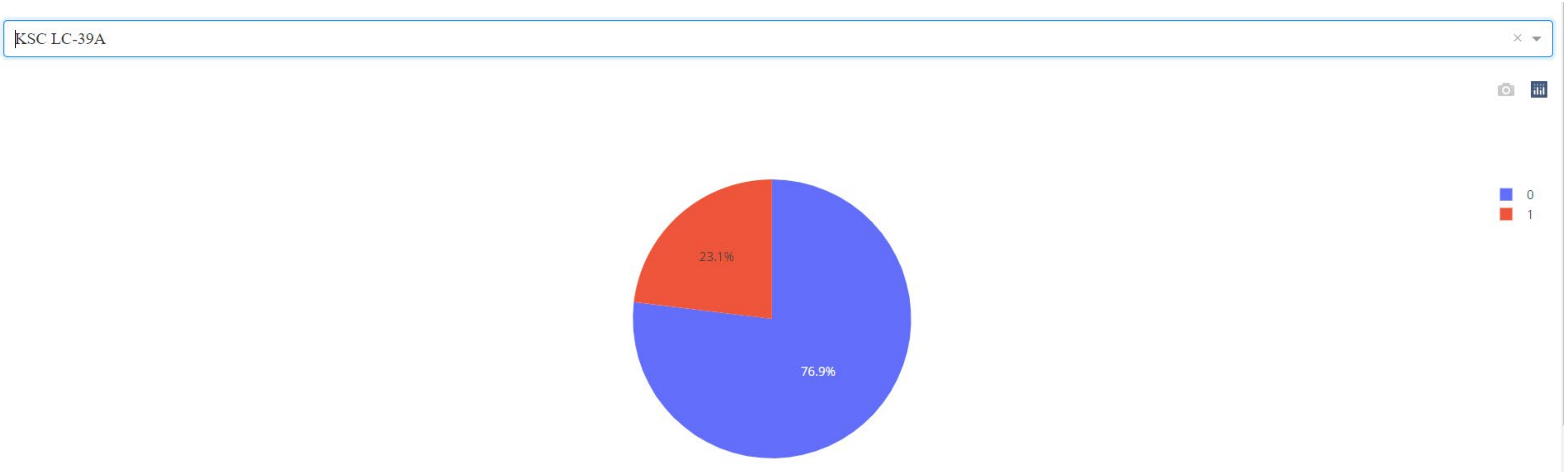


SpaceX Launch Records Dashboard

Select all

- The pie chart above shows the success rate for all the launch sites used by SpaceX.
- From the pie chart, it can concluded that launch site KSC LS-39A has the highest success rate compared to others.

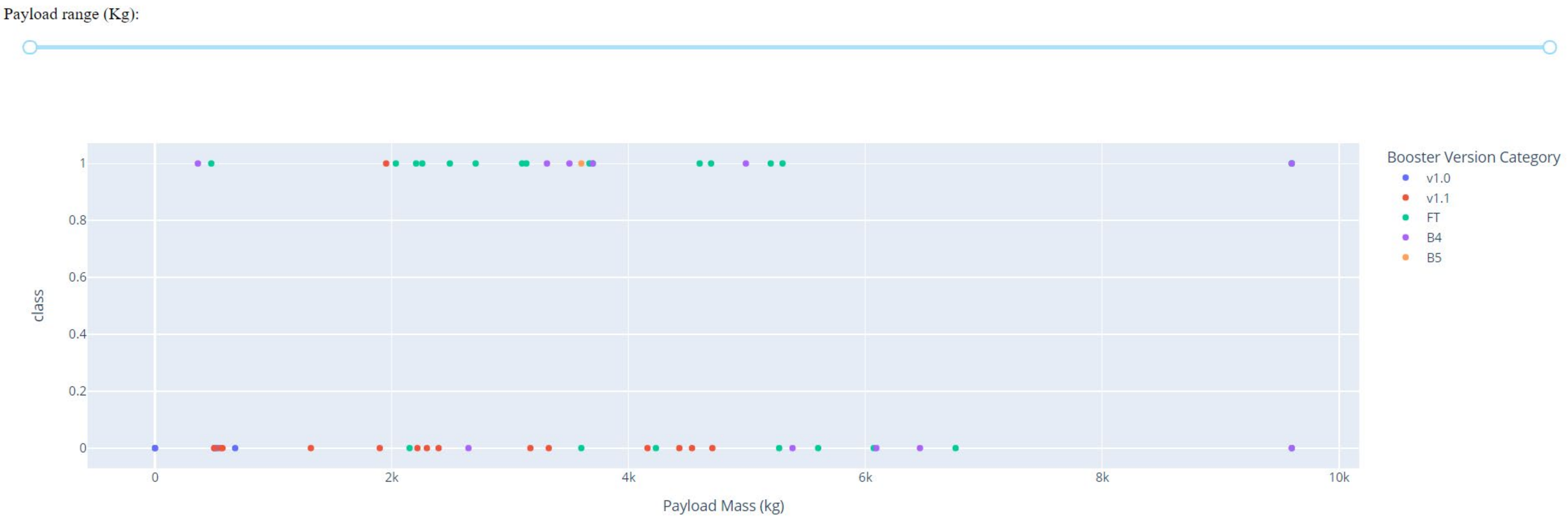
<Dashboard screenshot 2> replace



KSC LC-39A

- The above pie chart shows the success rate of launch site KSC LC-39A, which is 76.9% and is highest among all the launch sites used by SpaceX.

<Dashboard screenshot 3>



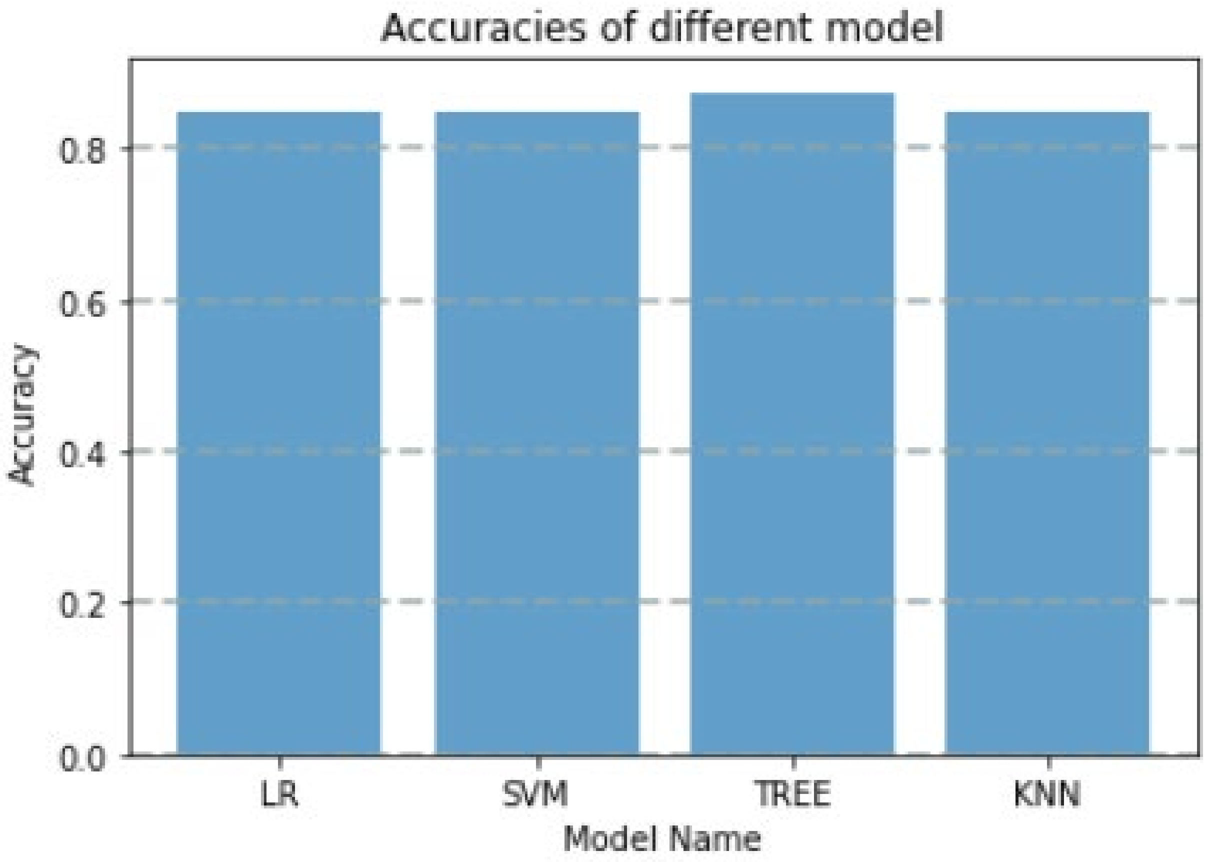
Payload range (Kg):

- The above scatter plot shows the payload carried during each successful and failed missions, where each color denoted a booster version.
- The range of payload mass can be changed by using the slider provided above.

Predictive analysis (Classification)

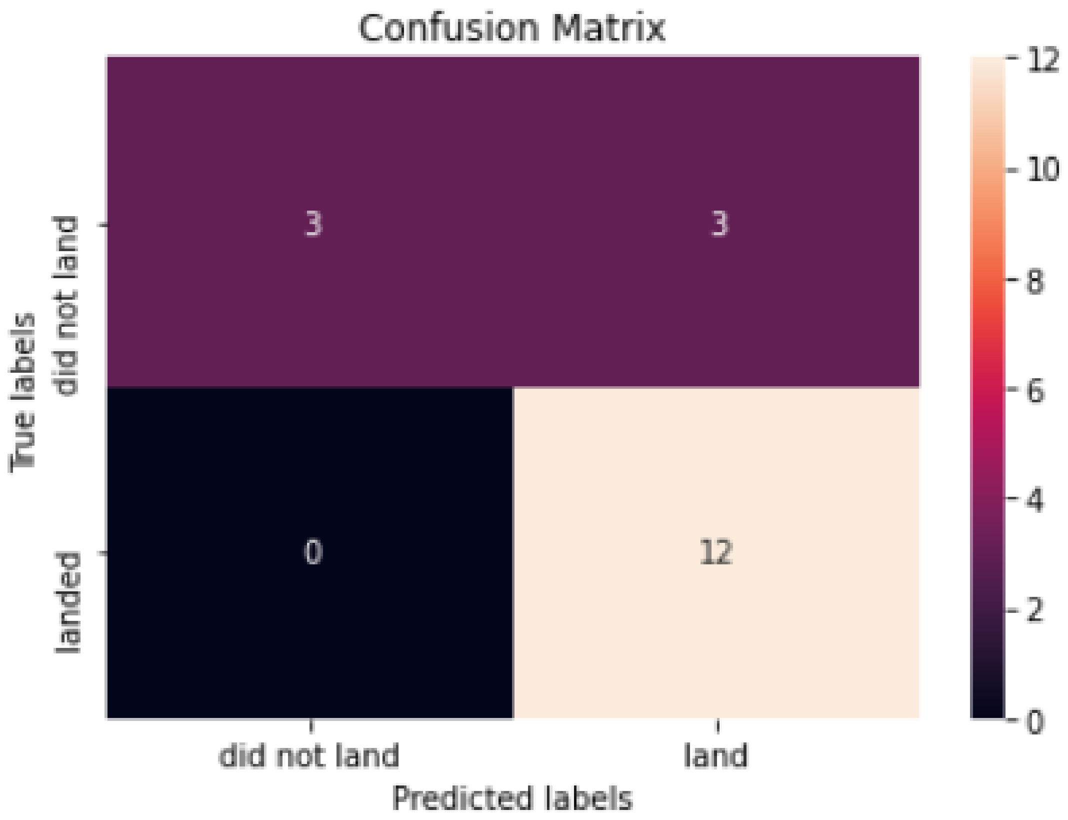
Classification Accuracy

- The bar chart here shows the accuracy comparison of different models used.
- From the bar chart, we find that TREE has the best accuracy among all the models used.



Confusion Matrix

- This is the confusion matrix for the TREE model as this model had the highest accuracy compared to others.
- From the confusion matrix we find that the TREE model had nearly predicted accurately for most of the cases except for 3 cases, where the outcome was supposed to land but actually it did not land.



CONCLUSION

- One can conclude that SpaceX has used launch site CCAFS SLC 40, but they achieved most of their success on launch site KSC LC-39A, which is a massive 76.9% success rate.
- Also it can noted that launch site KSC LC-39A has achieved a lot of success on various range of payload compared to it's counterpart CCAFS SLC 40.
- Most of lighter payload (between 0 and 8000 kg) were carried for ISS and GTO orbits where as the heaviest payload (above 14000 kg) were carried out for VLEO orbit. But SSO orbit had all it's missions successful.
- Booster Versions falling under FT category has enjoyed most success with payload in between 2000 kg to 6000 kg, but the most popular booster version.
- For prediction of future outcomes of missions, it is seen that nearly all model had the same accuracy but TREE model had higher accuracy than others, so it is recommended that TREE model be used for future prediction. In case to be double sure if the predictions are accurate, then it is recommended to use KNN or SVM model, either of them id fine, after TREE model.

APPENDIX

The following links were used while creating the dashboard:-

- https://dash.plotly.com/dash-core-components/dropdown?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
- https://dash.plotly.com/dash-core-components/rangeslider?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
- https://plotly.com/python/pie-charts/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01
- https://plotly.com/python/line-and-scatter/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDS0321ENSkillsNetwork26802033-2021-01-01