IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dragan Ignatović
25.03.2023.

Outline

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

Executive Summary

- Summary of methodologies
    - The project followed the Data Science methodology involving data collection, data wrangling, exploratory data analysis, data visualization, model development, model evaluation, and reporting of the results.
    - The project was done using Jupyter Notebook, Rest API, IBM DB2, Python and SQL.
- Summary of all results
    - Exploratory data analysis results
    - Interactive analytics
    - Predictive analysis

Introduction

- Project background and context
  - The project is a part of the IBM Data Science Professional Certificate and Applied Data Science with Python Specialization. It involves assuming the role of a Data Scientist working for a startup that intends to compete with SpaceX.
  - The project involves data collection, data wrangling, exploratory data analysis, model development, and model evaluation to provide accurate predictions.
- Problems you want to find answers
  - The task is to predict the success of the landing of the first stage of the SpaceX Falcon 9 rocket, which can help the competing startup in making more informed bids against SpaceX for a rocket launch.

Section 1

Methodology

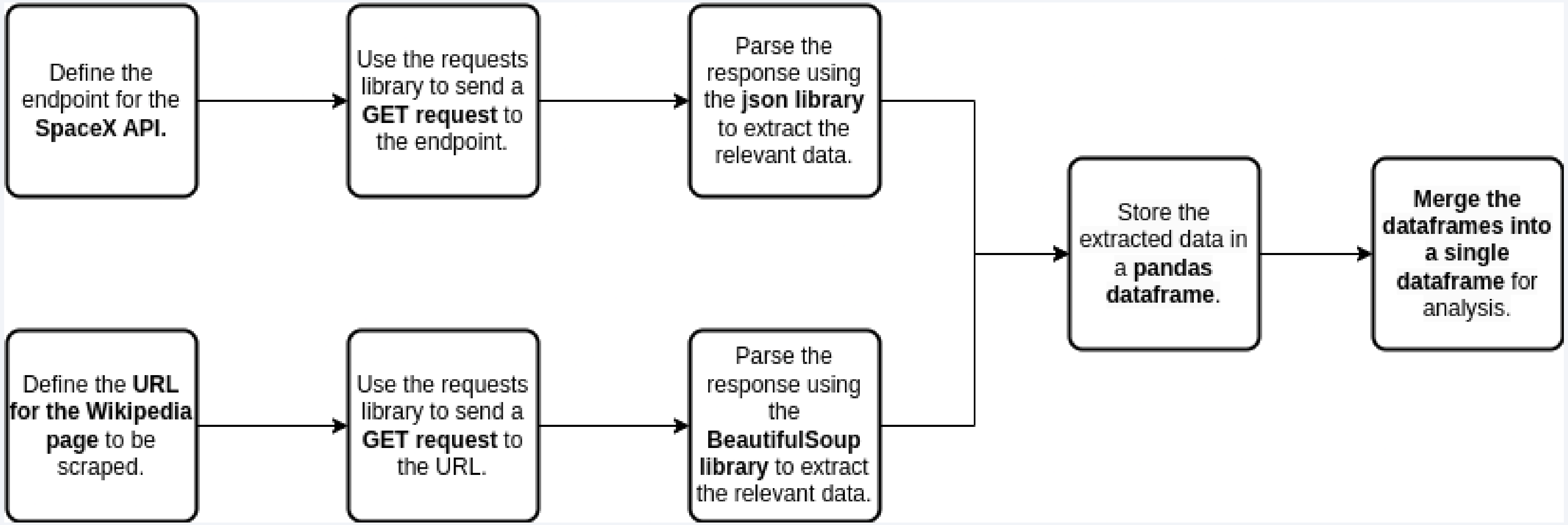
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - The data was processed using the pandas library and numpy library in Jupyter Notebook.
  - The data was explored using various methods such as checking for missing values, calculating the number of launches on each launch site, the number and occurrence of each orbit and the number and occurrence of mission outcomes per orbit type.
  - Finally, a landing outcome label was created from the Outcome column and the data was saved as a new CSV file.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Linar Regression(LR), K Nearest Neigbour(KNN), Support Vector Machine(SVM) and Decision Three(DT) models have been built using corresponding libraries in python and then evaluated for the the best classifier

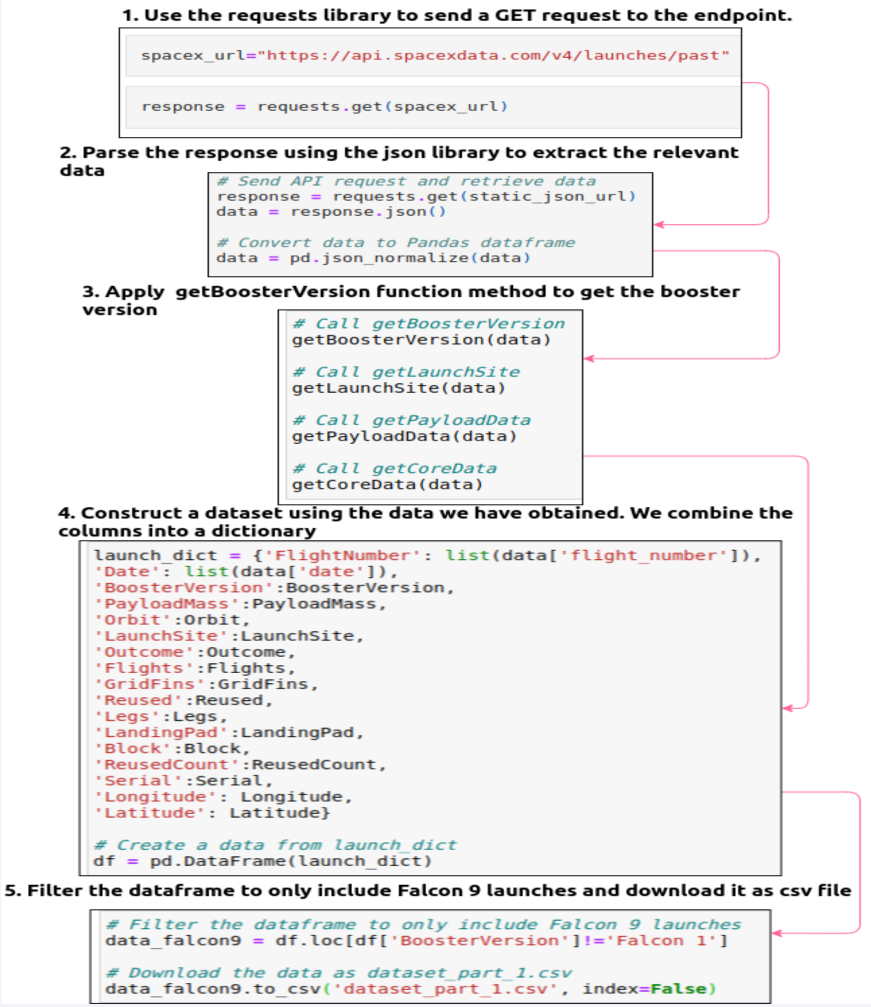
Data Collection

- For the SpaceX Rest API, the data was collected by making a GET request to the SpaceX API using Python's **requests** library. The API endpoint used was https://api.spacexdata.com/v4/launches. The response from the API was in JSON format, which was then parsed and converted into a Pandas data frame.
- For web scraping Falcon 9 launch records, the data was collected by using Python's **requests** library and BeautifulSoup to extract a Falcon 9 launch records HTML table from Wikipedia. The table was then parsed and converted into a Pandas data frame.



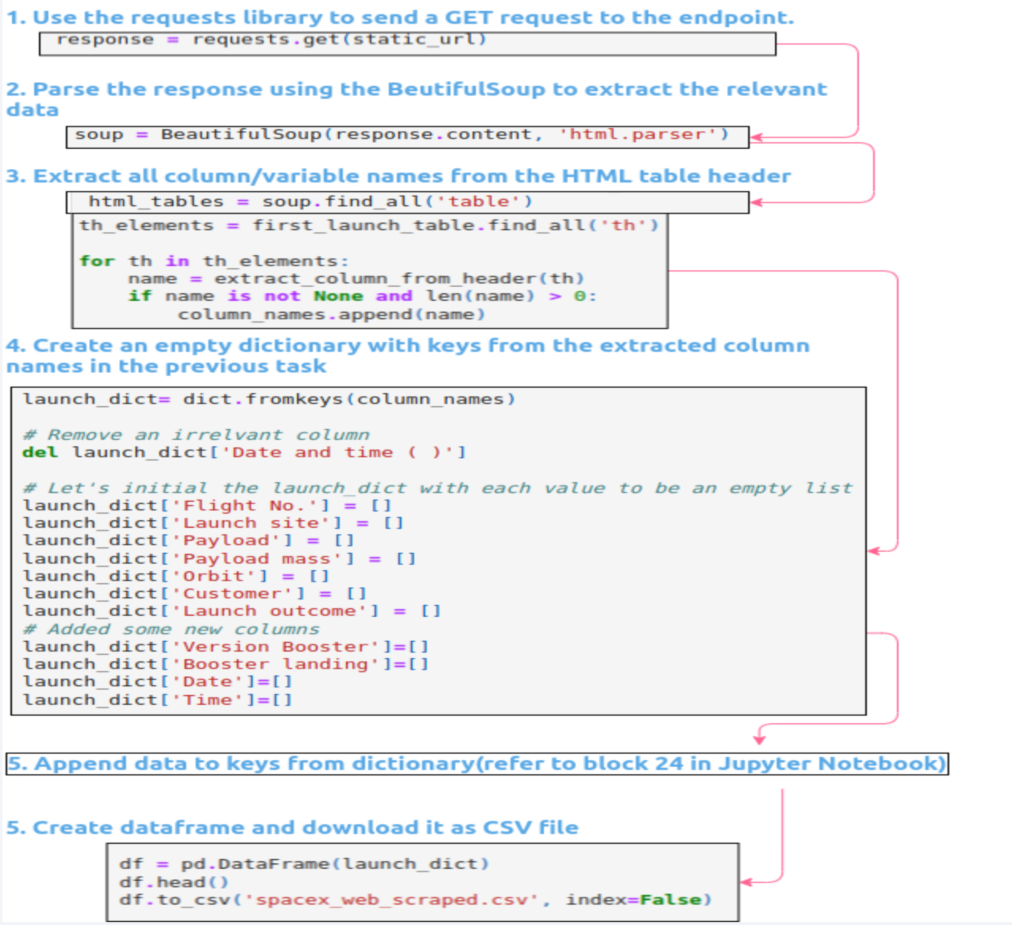
Data Collection – SpaceX API

GitHub URL of the completed SpaceX API calls notebook: https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_%20Data%20Collection_API.ipynb



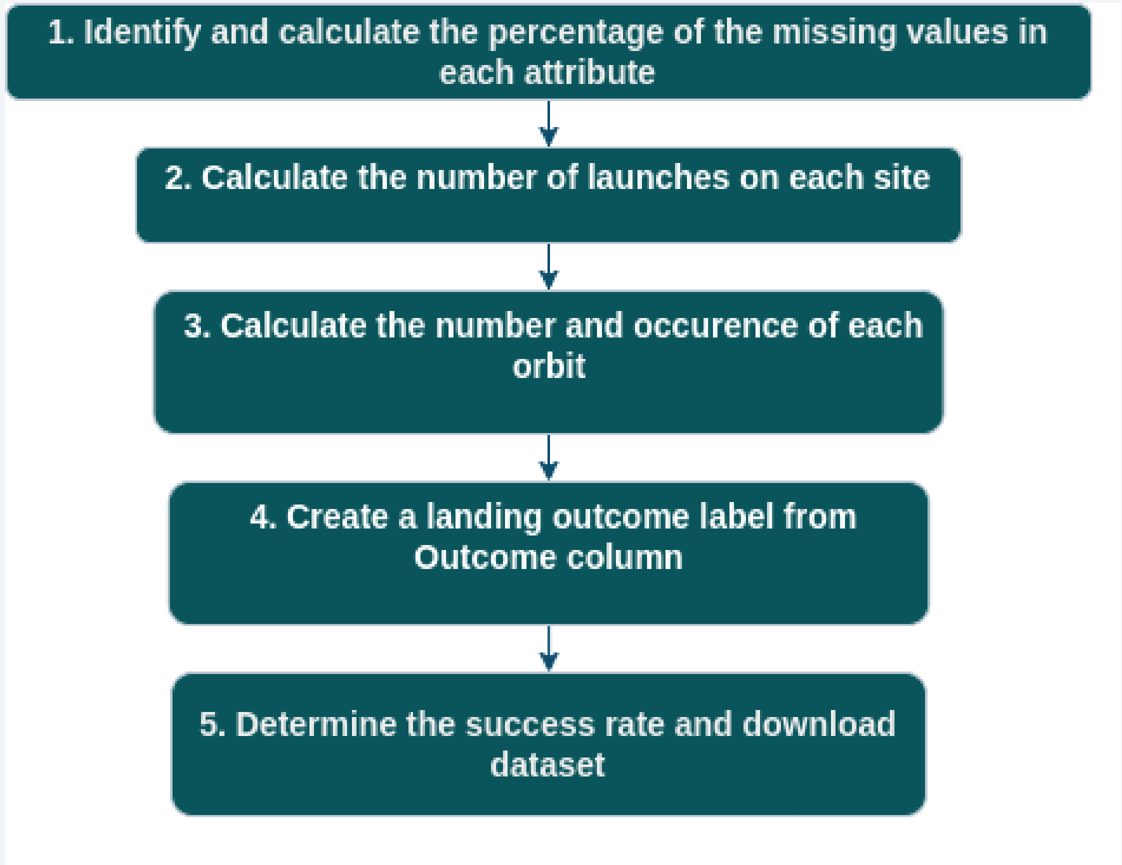
Data Collection - Scraping

GitHub URL of the completed web scraping notebook: https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_%20Data%20Collection%20with%20WebScraping.ipynb



Data Wrangling

GitHub URL of completed data wrangling notebook: https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatterplot between Flight Number and Launch Site
  - Lunch sites have highest success rate within higher flight numbers(40-80)
- Scatterplot between Payload and Launch Site
  - For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)
- Barchart between Success Rate of each Orbit Type
  - ES-L1, GEO, HEO and SSO have the same success rate and it's the highest
  - GTO has lowest success rate
- Scatterplot between FlightNumber and Orbit type
  - In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- Scatterplot between Payload and Orbit type
  - With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
  - However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Lince chart for launch success yearly trend
  - The chart has upward trend and highest success rate was in 2019

- GitHub URL of completed EDA with data visualization notebook: https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_%20EDA%20with%20Visualization%20-%20Not%20Watson.ipynb

EDA with SQL

- *Names of the unique launch sites in the space mission*
- *5 records where launch sites begin with the string 'CCA'*
- *The total payload mass carried by boosters launched by NASA (CRS)*
- *Average payload mass carried by booster version F9 v1.1*
- *The date when the first successful landing outcome in ground pad was acheived*
- *The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000*
- *The total number of successful and failure mission outcomes*
- *The names of the booster_versions which have carried the maximum payload mass*
- *The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015*
- *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*

GitHub URL of completed EDA with SQL notebook:
https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- **folium.Marker()** was used to mark launch sites on the map
- **folim.Circle()** was used to create circles around the launch sites on the map
- **folium.Icon()** was used to create icons on the map
- **folium.PolyLine()** was used to create polynomal line between the points
- **markerCluster()** was used to simplify the maps which had several markers with similar coordination

GitHub URL of completed interactive map with Folium map:
https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_Interactive%20Visaualization.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown menu:** Allows the user to select a specific launch site or view data for all launch sites.
- **Pie chart:** Shows the total number of successful launches for all sites or for a specific site.
- **Scatter plot:** Shows the relationship between Payload Mass and Booster Version Category for all sites or for a specific site.
- **Range slider:** Allows the user to select a range of Payload Mass (kg).

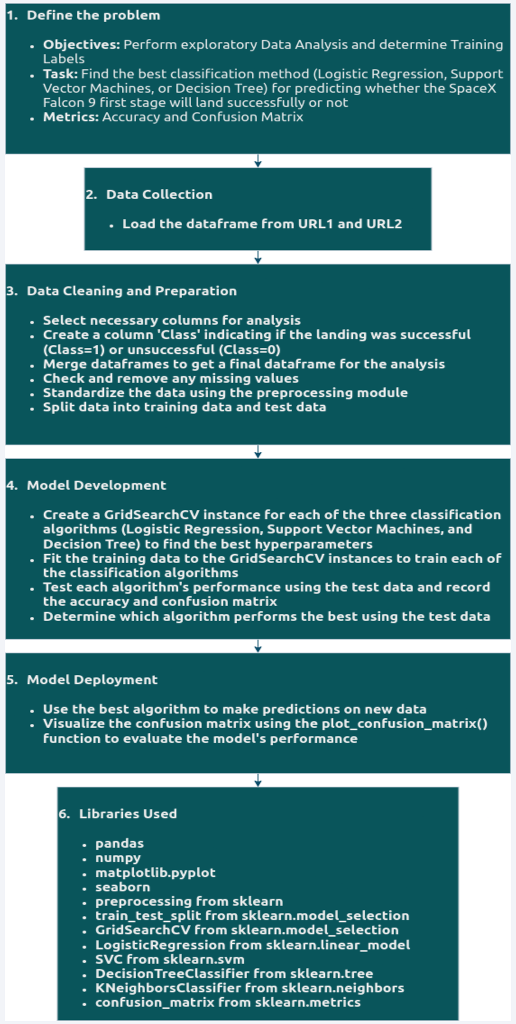
- These visualizations and interactions allow the user to explore and analyze the SpaceX launch data in different ways.
- The dropdown menu allows the user to select a specific launch site and view the data for that site.
- The pie chart shows the total number of successful launches, allowing the user to quickly compare different sites or see the overall success rate.
- The scatter plot shows the relationship between Payload Mass and Booster Version Category, allowing the user to identify any correlations between these variables.
- Finally, the range slider allows the user to filter the data by Payload Mass, making it easier to analyze specific subsets of the data.

GitHub URL of completed Plotly Dash lab:
https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/Dashboard/spacex_dashboard_working.py

Predictive Analysis (Classification)

- The first steps include performing exploratory data analysis to determine training labels, creating a column for the class, standardizing the data, and splitting the data into training and test sets.
- The next step is to find the best hyperparameters for SVM, Classification Trees, and Logistic Regression models, followed by selecting the best-performing method using test data.
- The necessary libraries for the project include pandas, numpy, matplotlib,seaborn, sklearn.preprocessing, sklearn.model_selection, sklearn.linear_model, sklearn.svm, sklearn.tree, and sklearn.neighbors.
- The project also includes defining auxiliary functions, such as a function to plot the confusion matrix. The project loads two data frames from URLs and assigns them to variables, data and X.

GitHub URL of completed predictive analysis lab: https://github.com/Gareleon/IBM_DS_Capstone/blob/b1d69f6a83b0fc3808342808a94719438c61391e/Final%20Capstone/CP_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



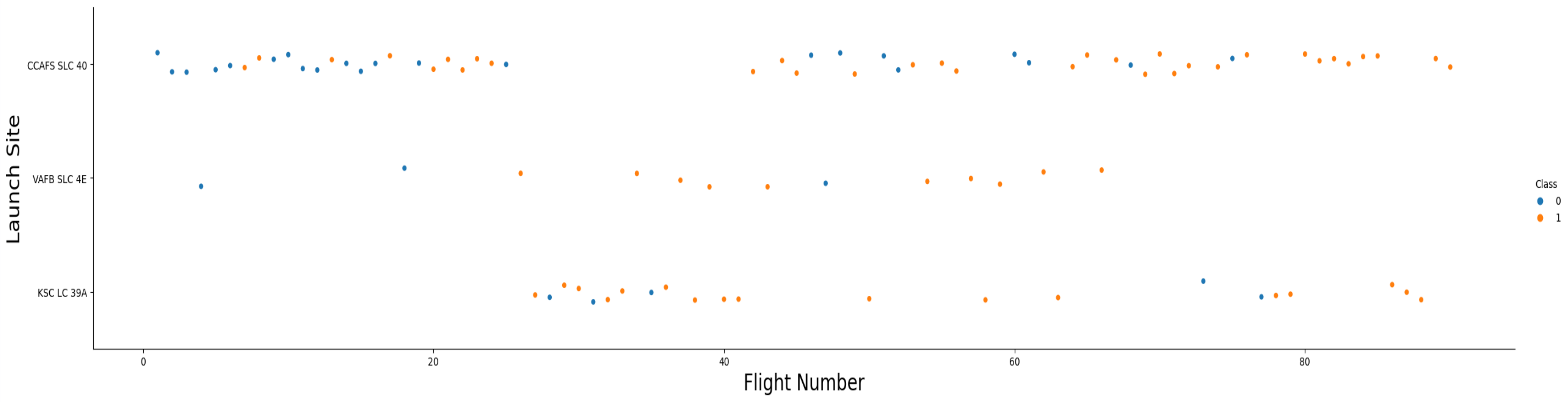
Results

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

Section 2

Insights drawn from EDA

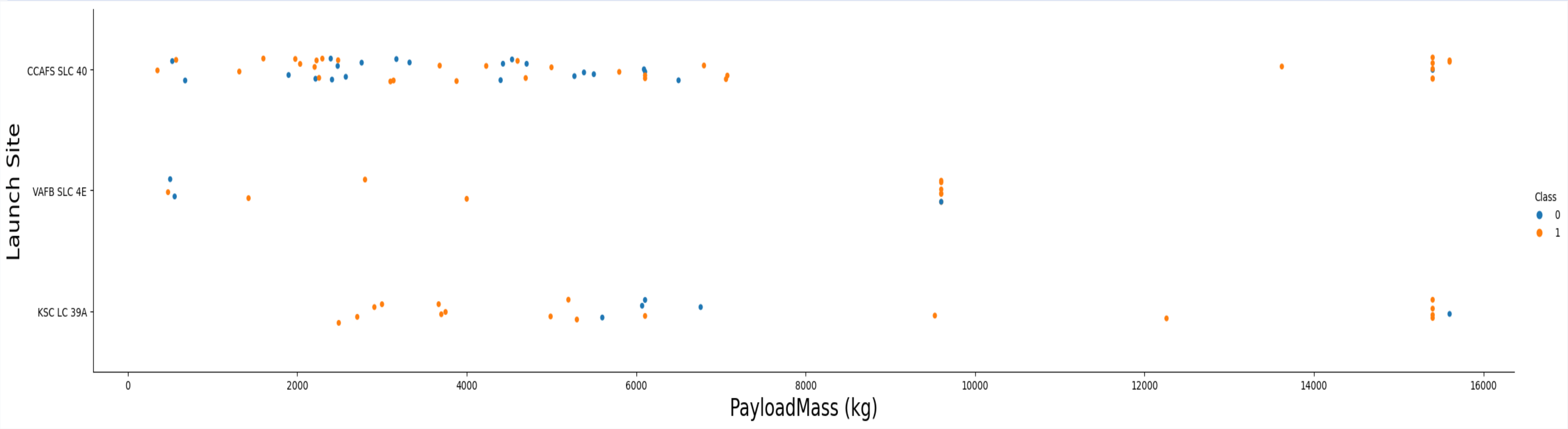
Flight Number vs. Launch Site



Lunch sites have highest success rate within higher flight numbers(40-80)

Highest number of launches is from CCAFS SLC 40

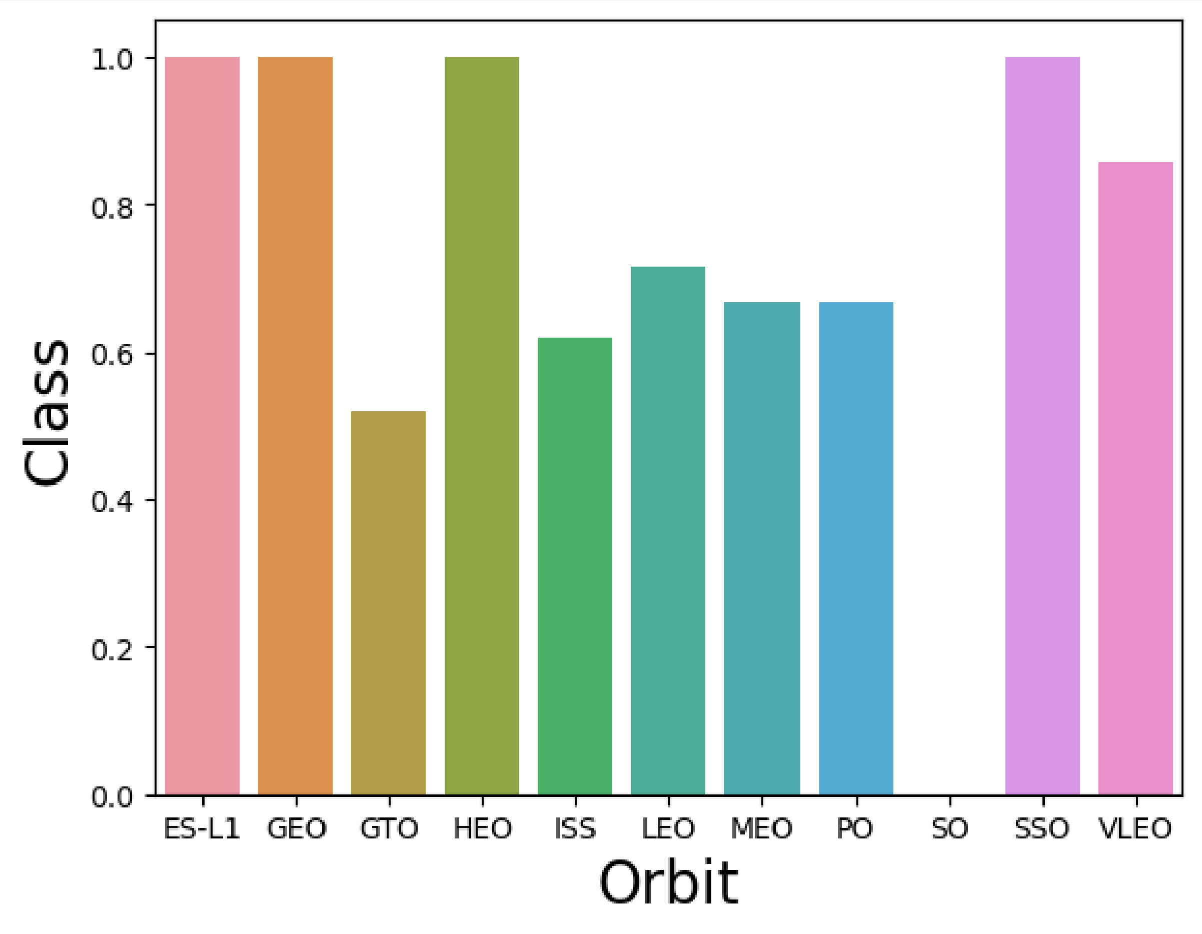
Payload vs. Launch Site



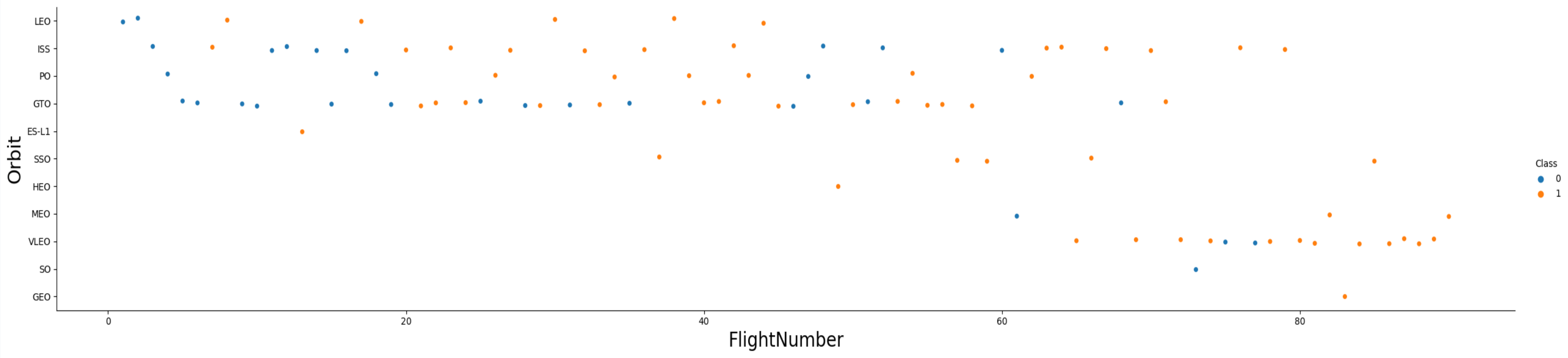
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have highest and the same success rate results
- GTO has lowest success rate

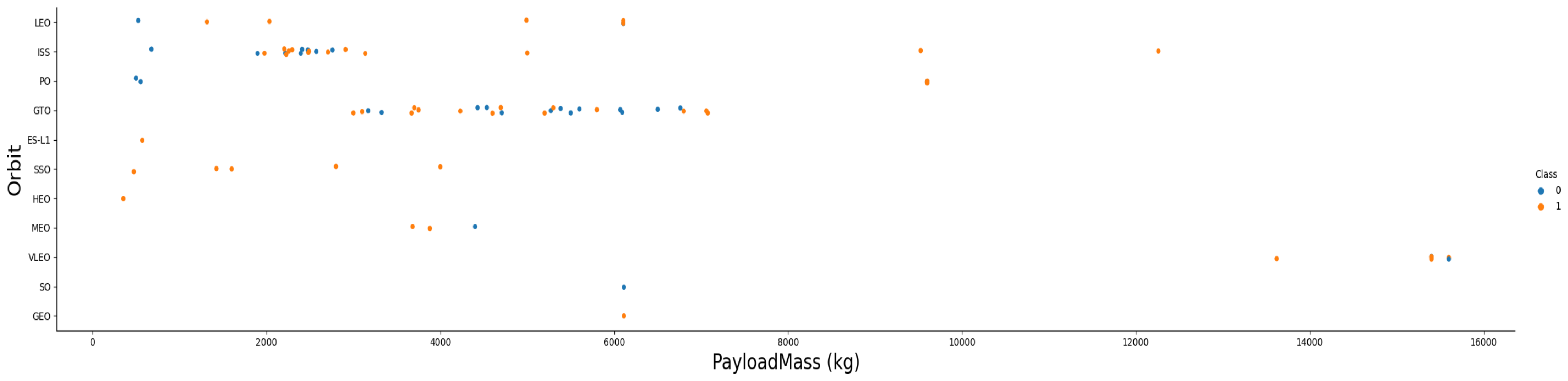


Flight Number vs. Orbit Type



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

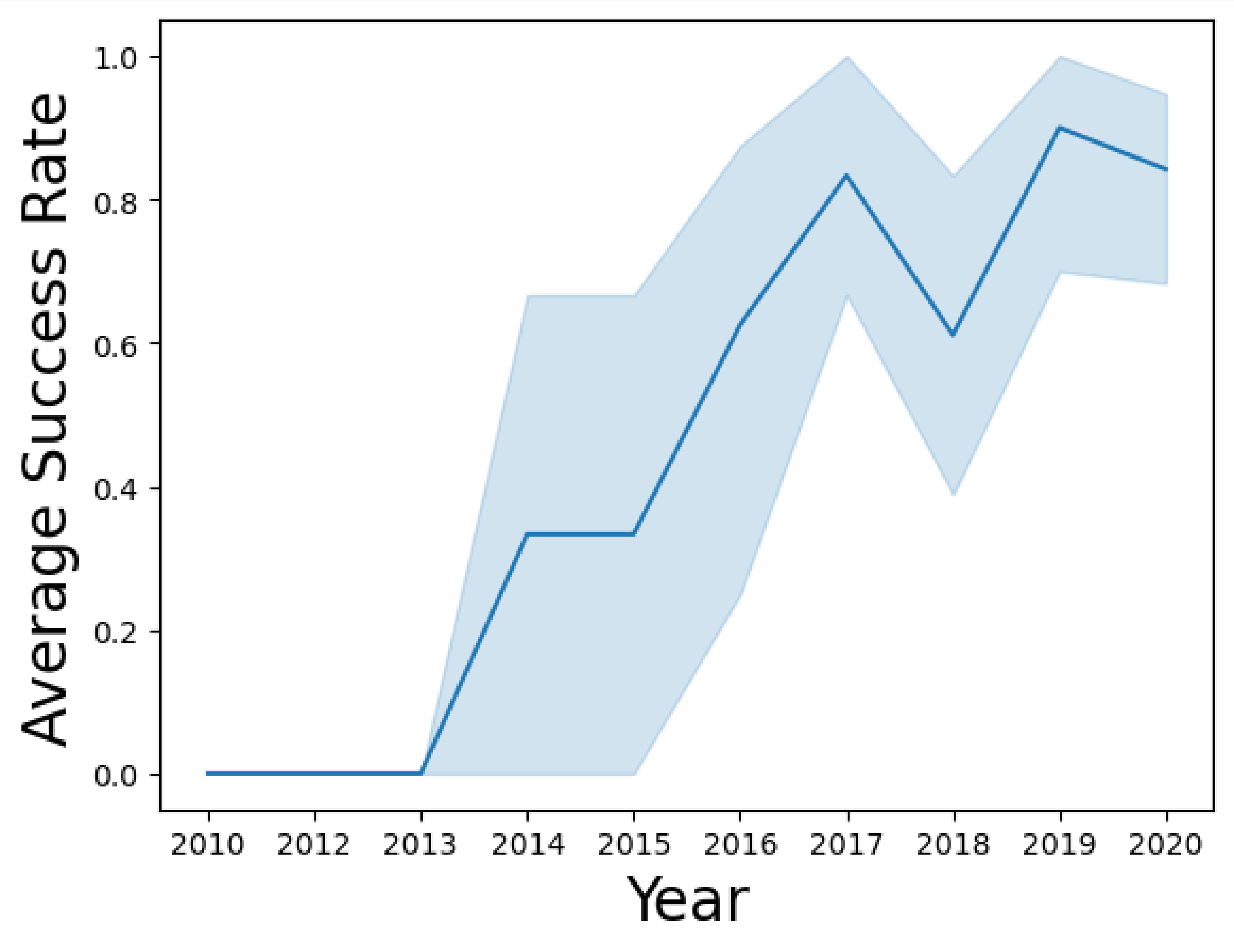
Payload vs. Orbit Type



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

Launch Success Yearly Trend

- The line chart has upward trend and highest success rate was in 2019



All Launch Site Names

```
%%sql
SELECT Distinct(launch_site) FROM spacex;
```

Out[12]:

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

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT launch_site
FROM spacex
WHERE launch_site LIKE 'CCA%'
LIMIT 5
;
```

Out[16]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%%sql
SELECT SUM(payload_mass_kg_)
FROM spacex
WHERE customer = 'NASA (CRS)'
;
```

Out[20]:

1
45596

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(payload_mass_kg_)
FROM spacex
WHERE booster_version = 'F9 v1.1'
;
```

Out[21]:

1
2928

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE)
FROM spacex
WHERE landing_outcome = 'Success (ground pad)'
;
```

Out[23]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%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
;
```

Out[24]:

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

```
%%sql
SELECT mission_outcome,
   COUNT(mission_outcome)  as
   total_number
FROM spacex
GROUP BY mission_outcome
;
```

Out[30]:

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

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, payload_mass_kg_
FROM spacex
WHERE
payload_mass_kg_ = (SELECT
   MAX(payload_mass_kg_) FROM spacex)
;
```

Out[31]:

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

```
%%sql
SELECT landing_outcome, booster_version,
   launch_site, DATE
FROM spacex
WHERE
landing_outcome = 'Failure (drone ship)'
AND
year(DATE) = '2015'
;
```

Out[67]:

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

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

```
%%sql
SELECT COUNT(landing_outcome) as COUNT
FROM spacex
WHERE DATE(DATE) BETWEEN DATE('2010-06-04') AND DATE('2017-03-20')
GROUP BY landing_outcome
ORDER BY COUNT (landing_outcome) DESC
;
```

Out[62]:

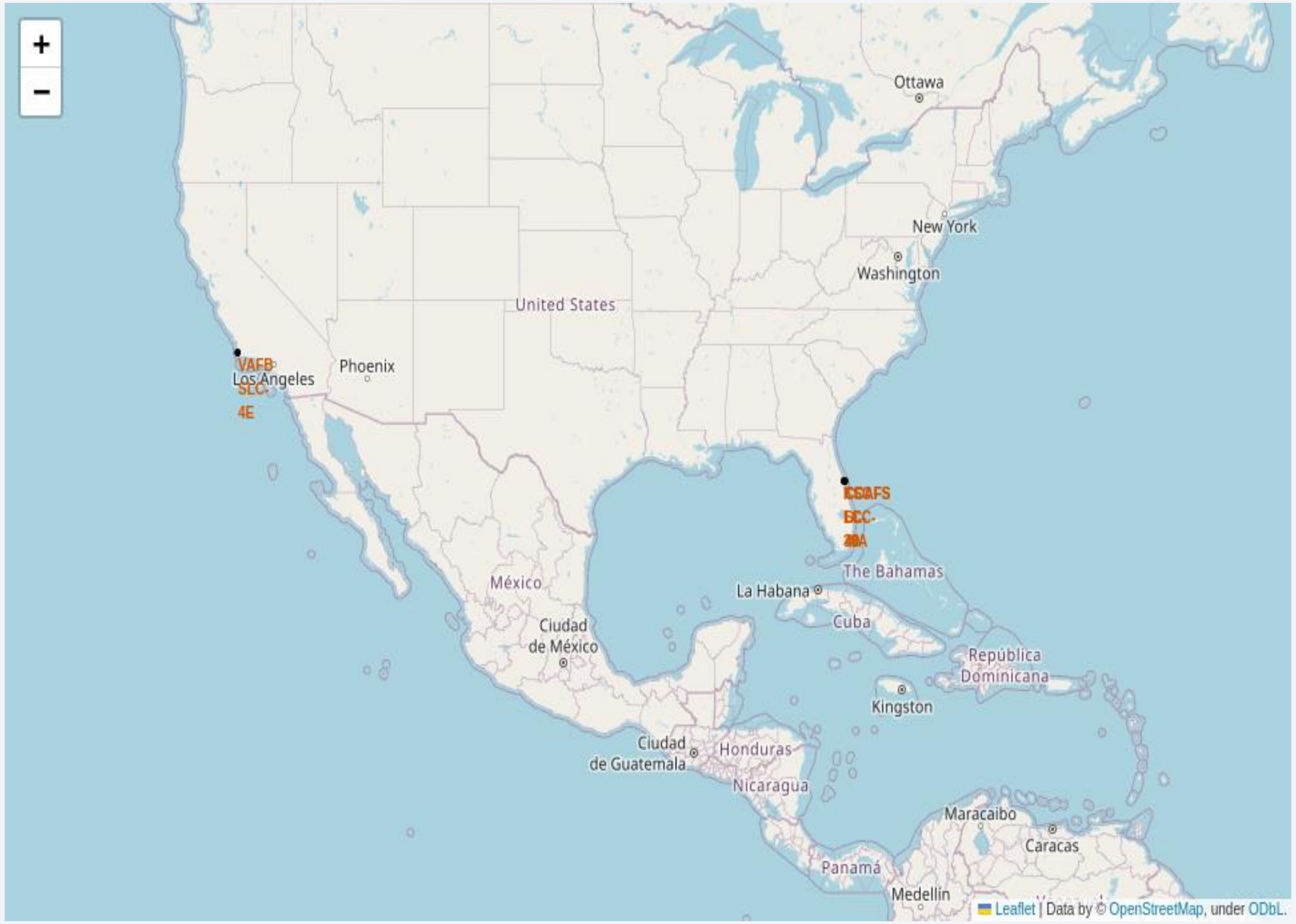
COUNT
10
5
5
3
3
2
2
1

Section 3

Launch Sites Proximities Analysis

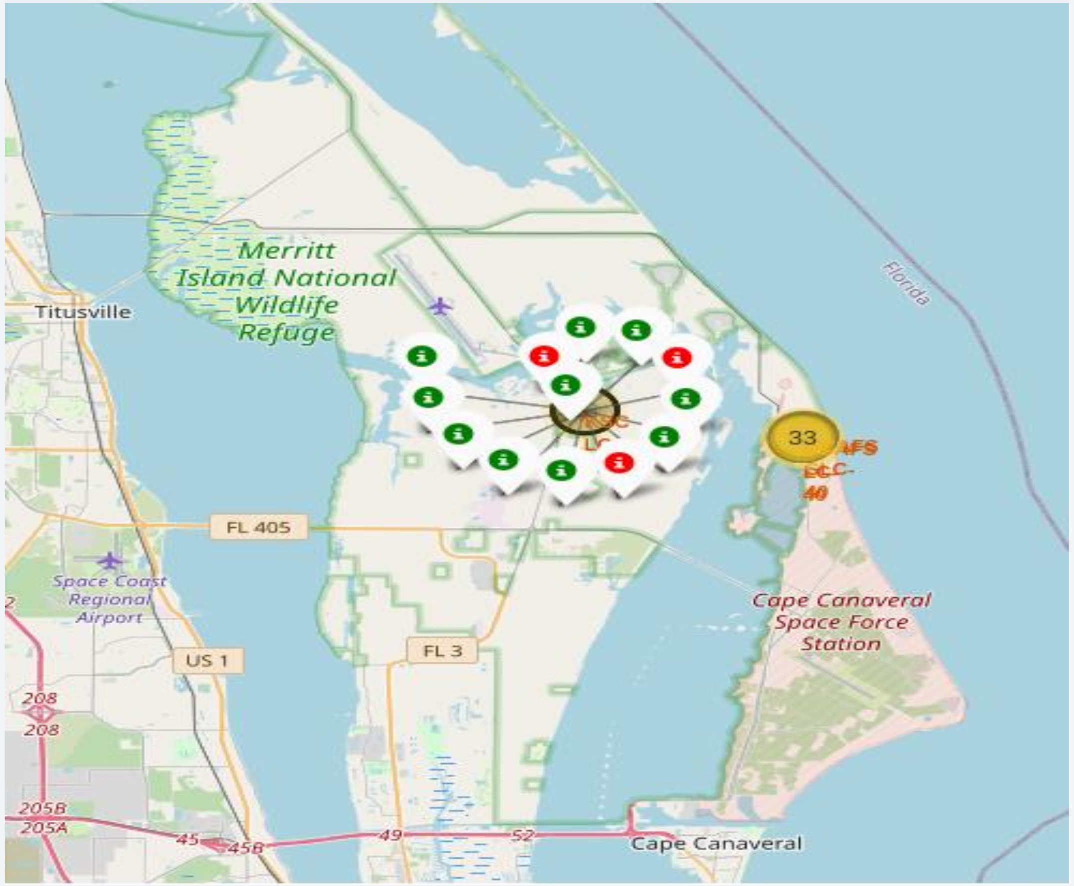
All launch sites on a map

- All launch sites are in proximity to the Equator line.
- Are all launch sites are very close proximity to the coast.



Launch outcomes on the map

- Green color = successful outcome
- Red color = unsuccessful outcome



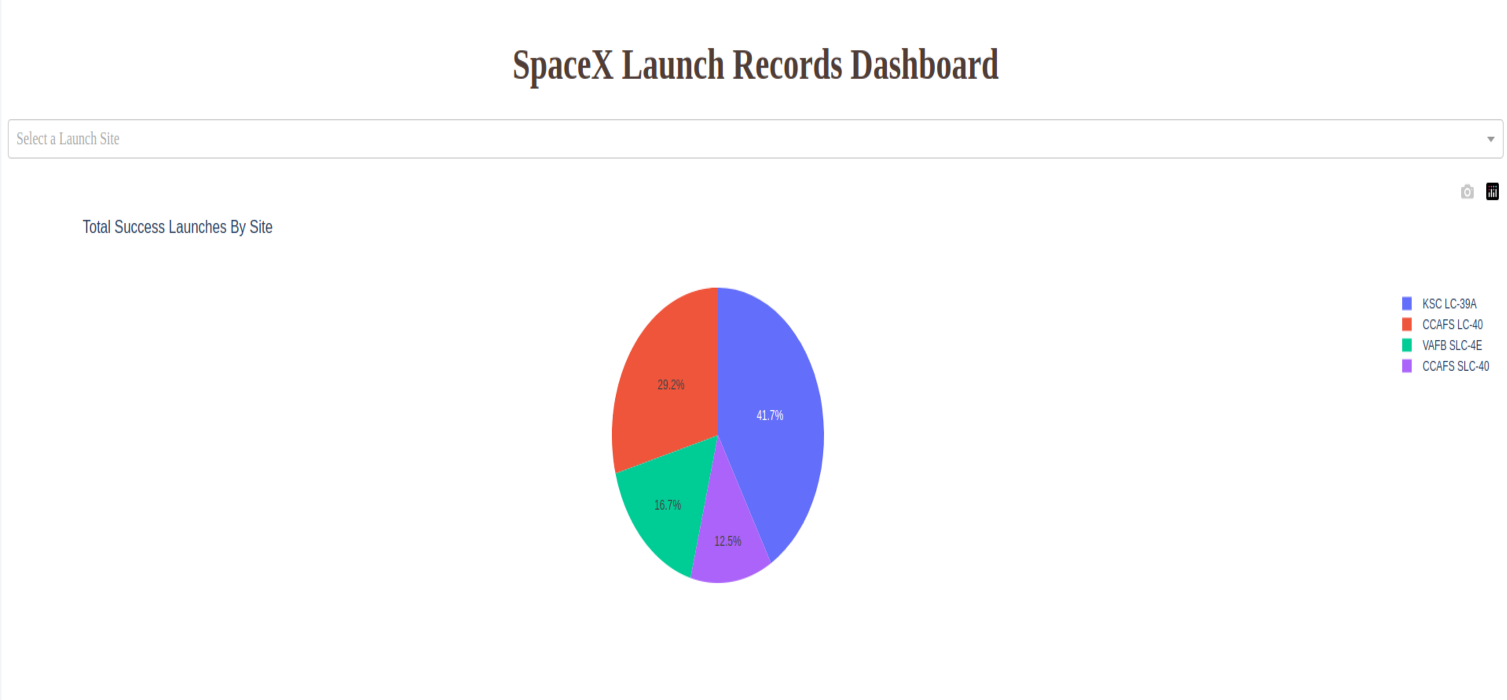
Launch site to it's proximites(railway, highway, coastline)

Section 4

Build a Dashboard with Plotly Dash

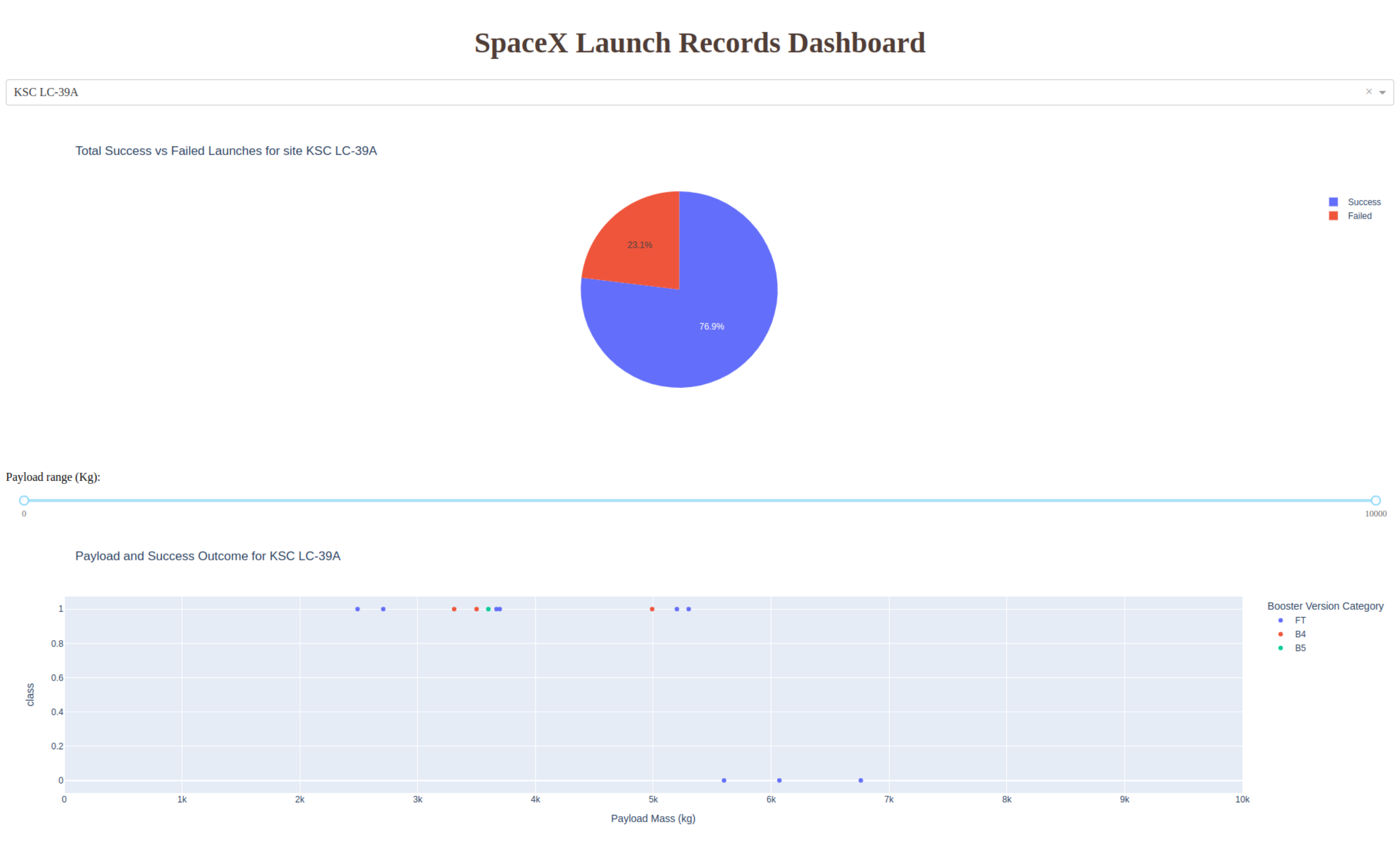
Total Success Launches by Site

- KSC LC-39A has the highest success score with 41.7%

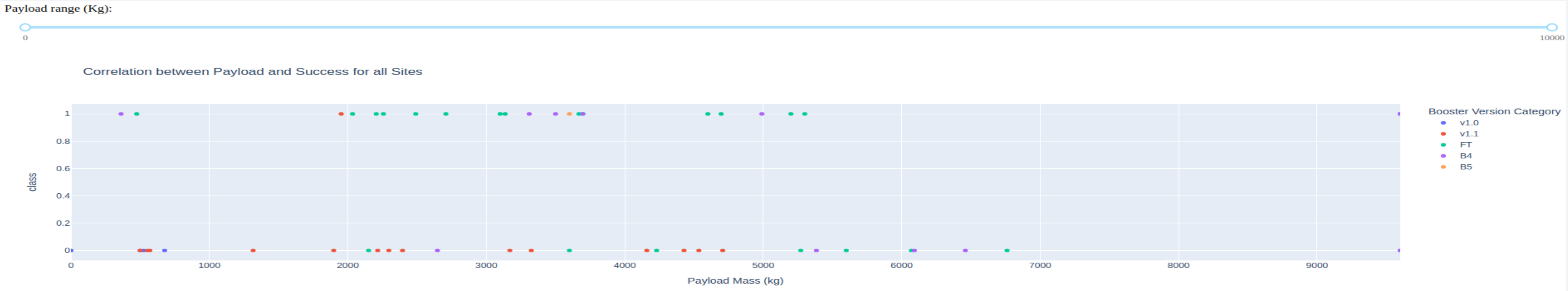


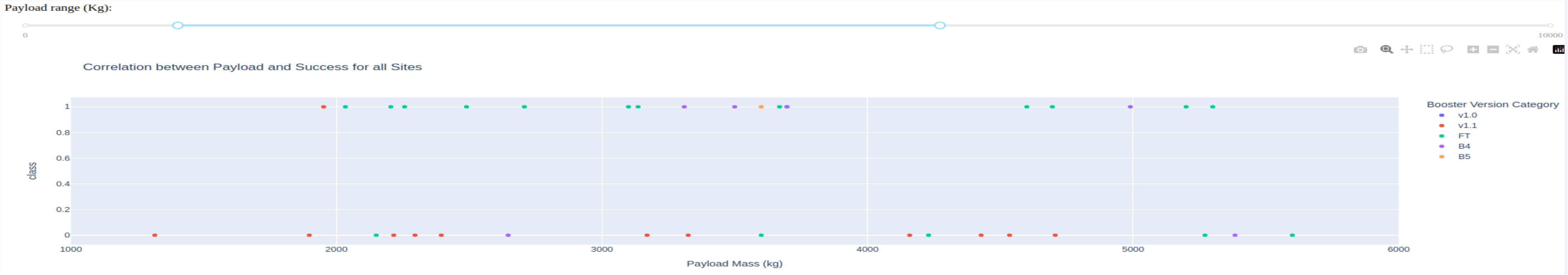
Total Success vs Failed Launches for site KSC LC-39A

- KSC LC-39A has success rate is 76.9% with payload range of 2000 kg – 10000 kg
- FT booster version has the most success



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





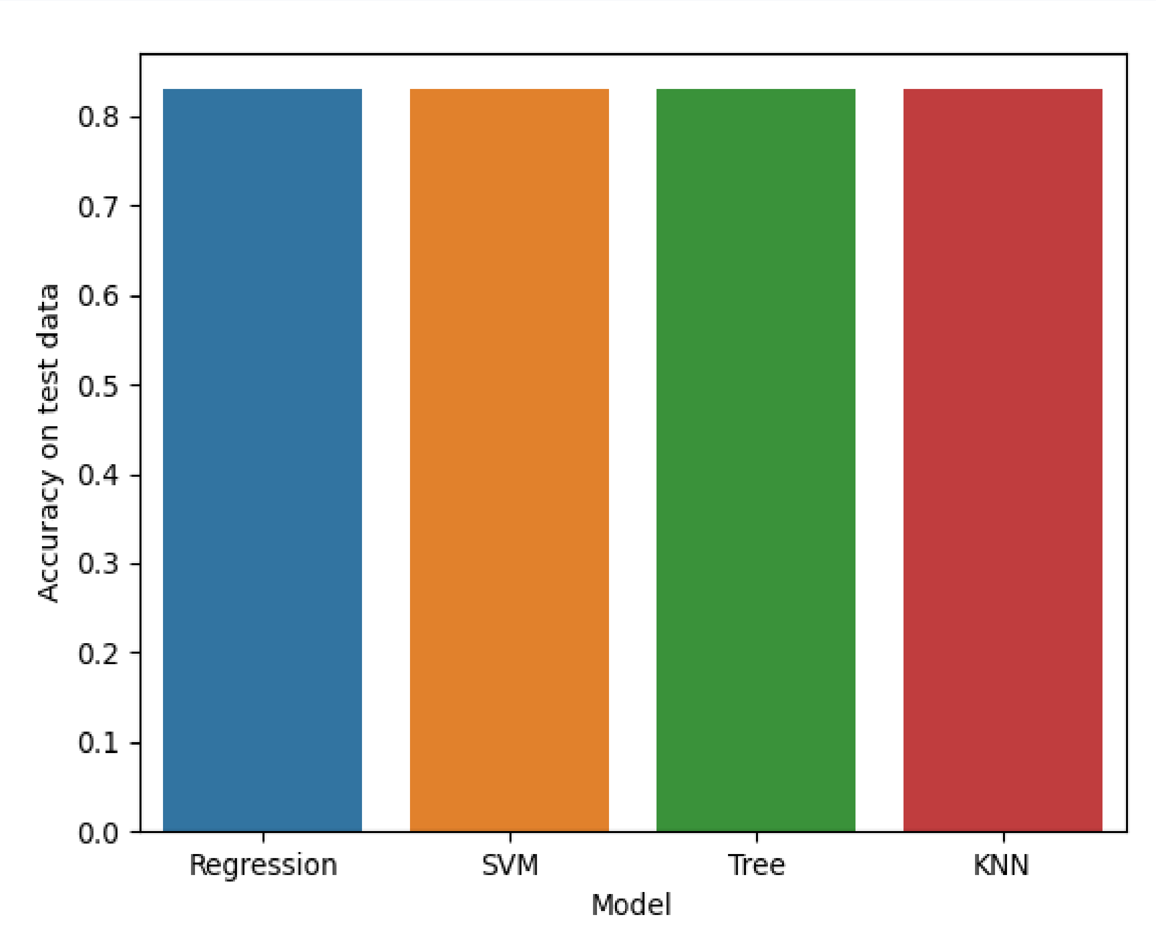
- Payload Range from 1000kg to 6000kg has the higher success range
- Booster version FT is the most frequent

Section 5

Predictive Analysis (Classification)

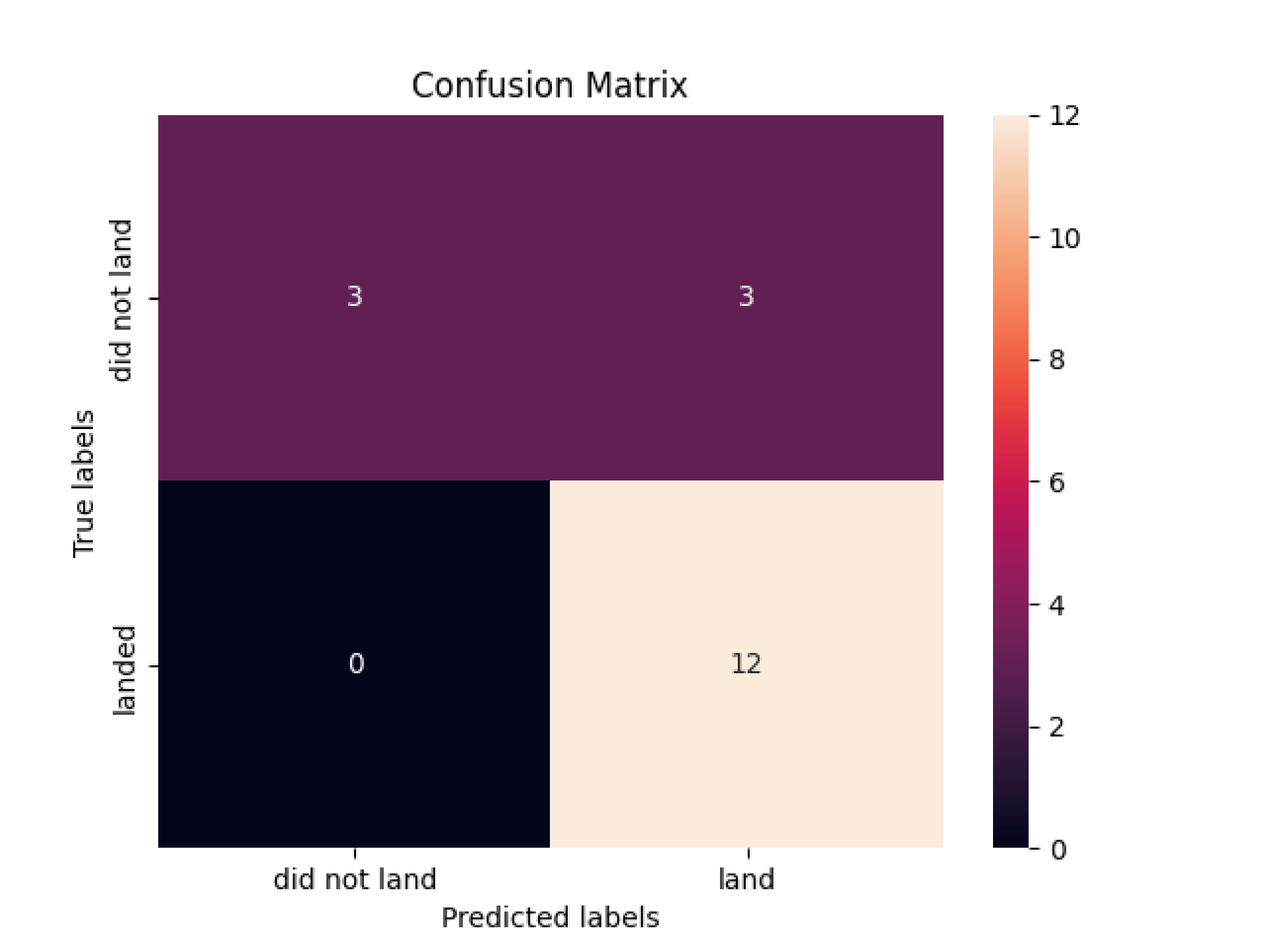
Classification Accuracy

From these results, we can see that all models perform the same.



Confusion Matrix

As all models preform the same, they have identical confusion matrix plots



Conclusions

- All models we tried perform the same.
- KSC LC-39A has the highest success score with 41.7%
- FT booster version has the most success
- Payload Range from 1000kg to 6000kg has the higher success range
- There is upward yearly Launch Success Trend, the highest success rate for now was in 2019

Thank you!