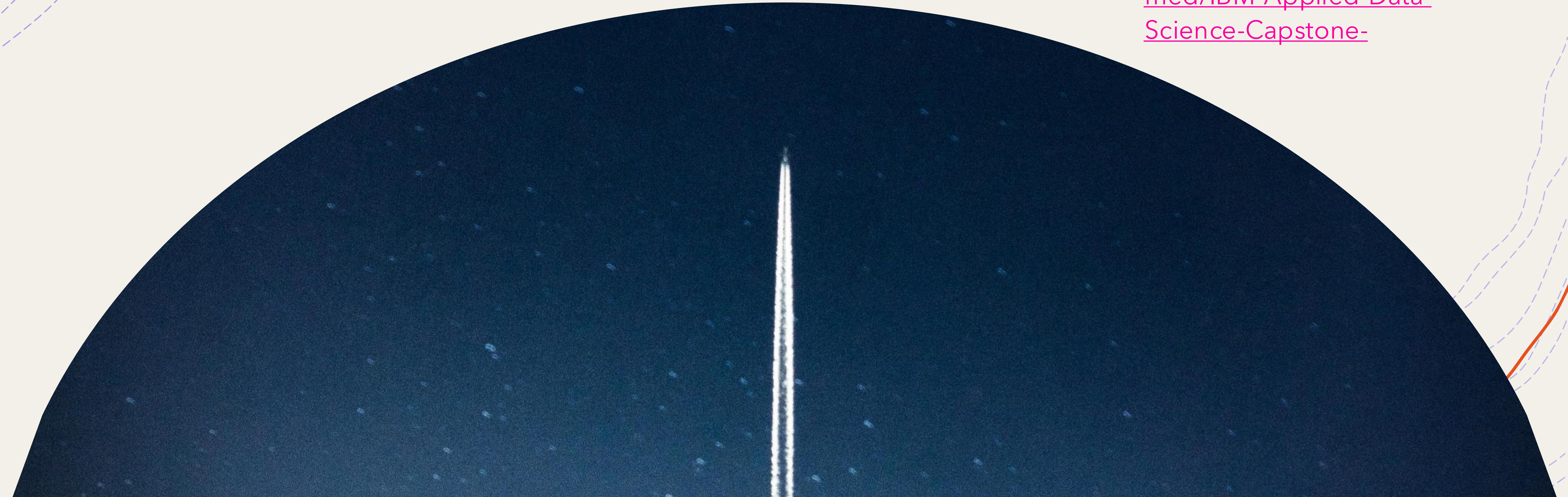
Applied Data Science Capstone Project

Done By:

Amruha Ahmed

GitHub Link:

https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-



Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

1. Executive Summary

The following methods were applied to gain a comprehensive picture of the data available:

- Data Collection
- Data Wrangling
- EDA
- Interactive Analytical Dashboard
- Predictive Analysis

The following points were inferred:

- EDA and visualizations using dashboards helped in understanding how each parameter is affected by other and gain valuable insights
- Accuracy of predicting whether the first stage of Falcon 9 rocket will land or not is 83.33%

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

2. Introduction

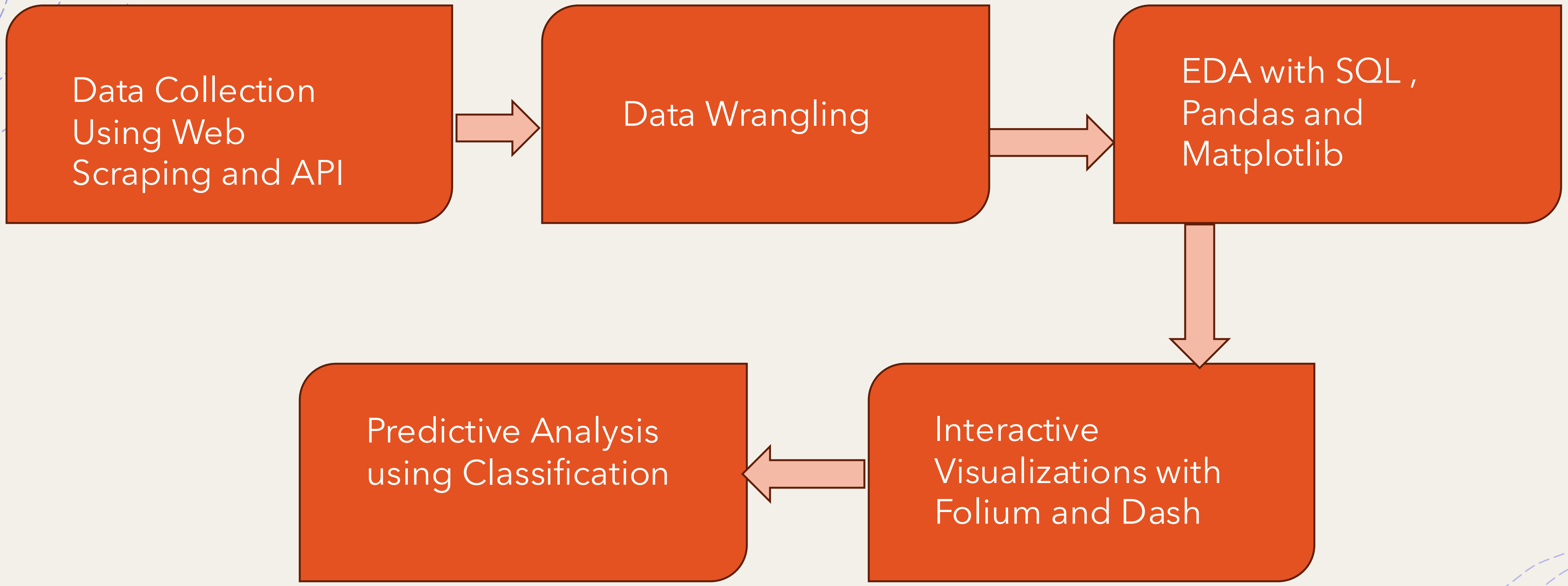
Space X advertises Falcon 9 rocket launches on its website with a cost of **62 million dollars**; other providers cost upward of **165 million dollars** each, much of the savings is because Space X can reuse the first stage.

The goal is to

- determine the price of each launch for **Space Y , that is competing with SpaceX.**
- Gather information about SpaceX
- Creating dashboards for better insights
- Whether SpaceX will reuse the first stage of Falcon 9 or not using machine learning

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

3. Methodology



Data Collection Using Web Scraping and API

Data Wrangling

EDA with SQL , Pandas and Matplotlib

Interactive Visualizations with Folium and Dash

Predictive Analysis using Classification

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Data Collection

GitHub Link for Data Collection using API:https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20collection%20using%20api.ipynb

GitHub Link for Data Collection using Web Scraping: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20collection%20using%20web%20scraping.ipynb

3.1 Data Collection

Data Collection using API

- URL is used to target a specific endpoint of the API to get past launch data.
- perform a get request using the requests library to obtain the launch data, which we will use to get the data from the API.
- Result is views using .json()
- To convert JSON into pandas dataframe, json_normalize() is used

Data Collection using Web Scraping

- Python BeautifulSoup package is used to web scrape some HTML tables that contain valuable Falcon 9 launch records
- Data is parsed from those tables and convert them into a Pandas data frame for further visualization and analysis.
- Instances in the data containing Falcon 1 have to be removed
- Null values of Payload Mass are filled using the mean of Payload Mass

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Data Wrangling

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/data%20wrangling.ipynb

3.2 Data Wrangling

Data Wrangling of the dataset involved the following steps

- Calculating the number of launches on each site
- Calculating the number and instance of each orbit
- Calculating the number and ocurrence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- Exporting the resultant dataframe into dataset_part_2.csv

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

EDA with SQL

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/eda%20with%20sql.ipynb

3.3.1 EDA with SQL

Exploratory Data Analysis of the dataset using SQL involved the following tasks :

- Installing SQL alchemy
- Connecting to a database
- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first succesful landing outcome in ground pad was acheived.

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

3.3.1 EDA with SQL Continued

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Ranking 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.

EDA with Pandas and Matplotlib + Feature Engineering

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/eda%20with%20pandas%20and%20matplotlib.ipynb

3.3.2 EDA with Pandas and Matplotlib

Exploratory Data Analysis of the dataset using Pandas and Matplotlib libraries in Python involved the following tasks :

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload Mass and Launch Site
- Visualize the relationship between success rate of each orbit type+
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload Mass and Orbit type
- Visualize the launch success yearly trend

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

3.3.2 Feature Engineering

Feature Enginering of the dataset using Python involved the following tasks :

- Create dummy variables to categorical columns of Orbits, Launch Site, Landing Pad, Serial using One Hot Encoding
- Cast all numeric columns to `float64`

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Interactive Visualizations with Folium

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/interactive%20visualizations%20using%20folium.ipynb

3.4.1 Interactive Visualizations with Folium

Interactive Launch Sites Locations Analysis with Folium is done through the following tasks:

- create a folium `Map` object,
- Create and add `folium.Circle` and `folium.Marker` for each launch site on the site map
- Create a new column in `spacex_df` dataframe called `marker_color` to store the marker colors based on the `class` value
- Mark the success/failed launches for each site on the map
- For each launch result in `spacex_df` data frame, add a `folium.Marker` to `marker_cluster`
- Calculate the distances between a launch site to its proximities

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Interactive Dashboard with Dash

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/interactive%20dashboard%20using%20dash.py

3.4.2 Interactive Dashboard with Dash

The following steps were involved to create an interactive dashboard using Plotly Dash:

- Reading the airline data into pandas dataframe
- Creating a dash application
- Creating an app layout
- Adding a dropdown list to enable Launch Site selection
- Adding a callback function for `site-dropdown` as input, `success-pie-chart` as output. And a function decorator to specify function input and output

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

3.4.2 Interactive Dashboard with Dash Continued

- Adding a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, showing the Success vs. Failed counts for the site
- Adding a slider to select payload range
- Adding a callback function for `site-dropdown` and `payload-slider` as inputs, `success-payload-scatter-chart` as output
- Adding a scatter chart to show the correlation between payload and launch success
- Running the app

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Predictive Analysis using Classification

GitHub Link: https://github.com/AmruhaAhmed/IBM-Applied-Data-Science-Capstone-/blob/main/predictive%20analysis(classification).ipynb

3.5 Predictive Analysis using Classification

- In order to create a machine learning pipeline to predict if the first stage will land, the following steps are used:
- Loading the dataframe
- Creating a NumPy array from the column Class(target column)
- Standardizing the data in independent columns
- Using the function train_test_split to split the data X and Y into training and test data and the test size=20 % and training size=80% of the total data
- Creating a logistic regression object then create a GridSearchCV object cross validation = 10. Fitting the object to find the best parameters from the dictionary parameters

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

3.5 Predictive Analysis using Classification

-

  Creating a support vector machine object then create a GridSearchCV object with cross validation = 10. Fitting the object to find the best parameters from the dictionary parameters.
- Creating a decision tree classifier object then create a GridSearchCV object with cross validation = 10. Fitting the object to find the best parameters from the dictionary parameters.
- Creating a k nearest neighbors object then create a GridSearchCV object with cv = 10. Fitting the object to find the best parameters from the dictionary parameters.
- Making confusion matrix and evaluating the models using accuracy, precision, recall and f1 score

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

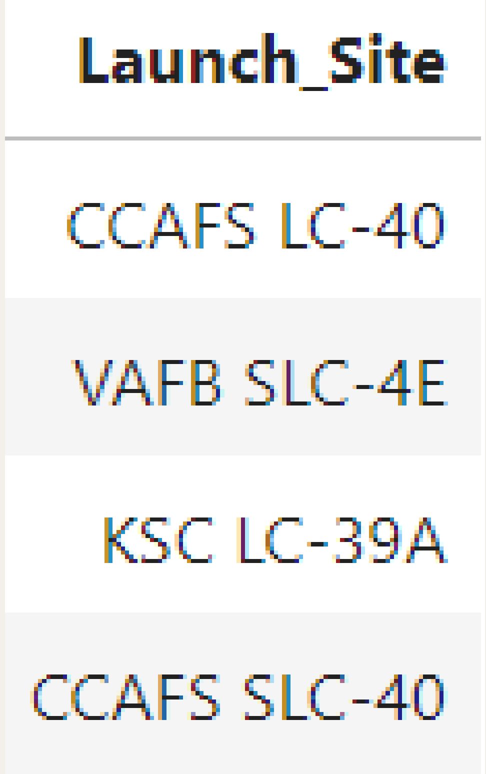
4. Results

The results are categorized into 5 critical stages

- EDA with SQL
- EDA with Matplotlib and Pandas
- Interactive Visualizations with Folium
- Interactive Dashboard using Dash
- Predictive Analysis using Classification

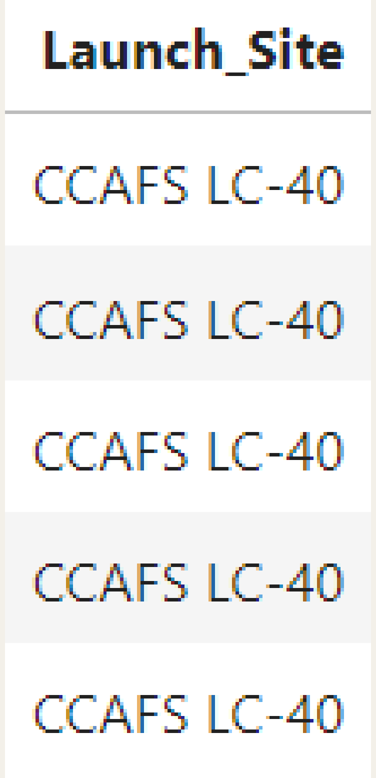
4.1 EDA with SQL

Results of Task 1:Display the names of the unique launch sites in the space mission



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

Results of Task 2:Display 5 records where launch sites begin with the string 'CCA'

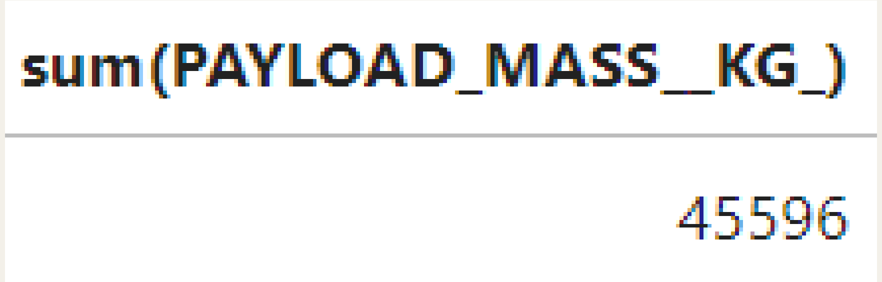


Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

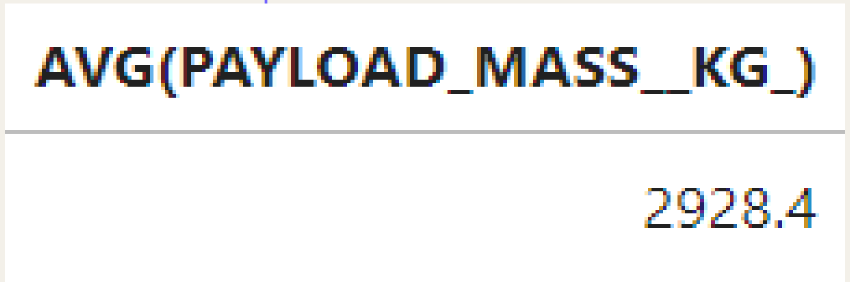
4.1 EDA with SQL

Results of Task 3:Display the total payload mass carried by boosters launched by NASA (CRS)



sum(PAYLOAD_MASS__KG_)
45596

Results of Task 4:Display average payload mass carried by booster version F9 v1.1

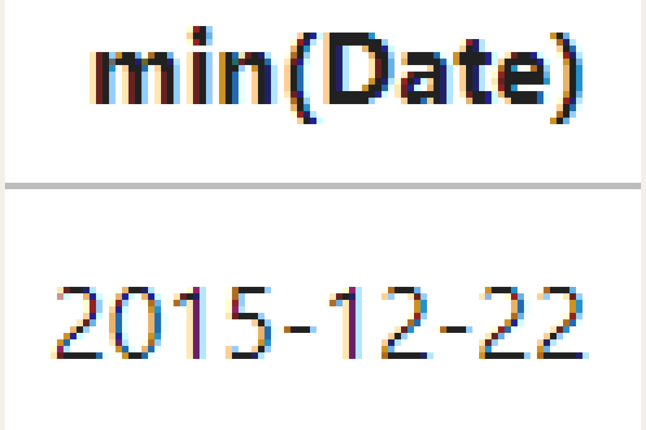


AVG(PAYLOAD_MASS__KG_)
2928.4

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

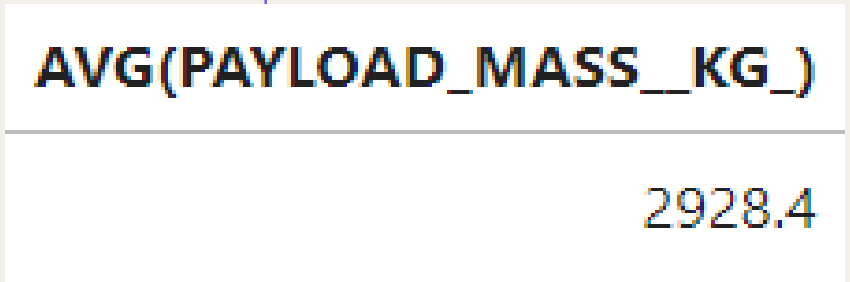
4.1 EDA with SQL

Results of Task 5:List the date when the first succesful landing outcome in ground pad was acheived.



min(Date)
2015-12-22

Results of Task 6:List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

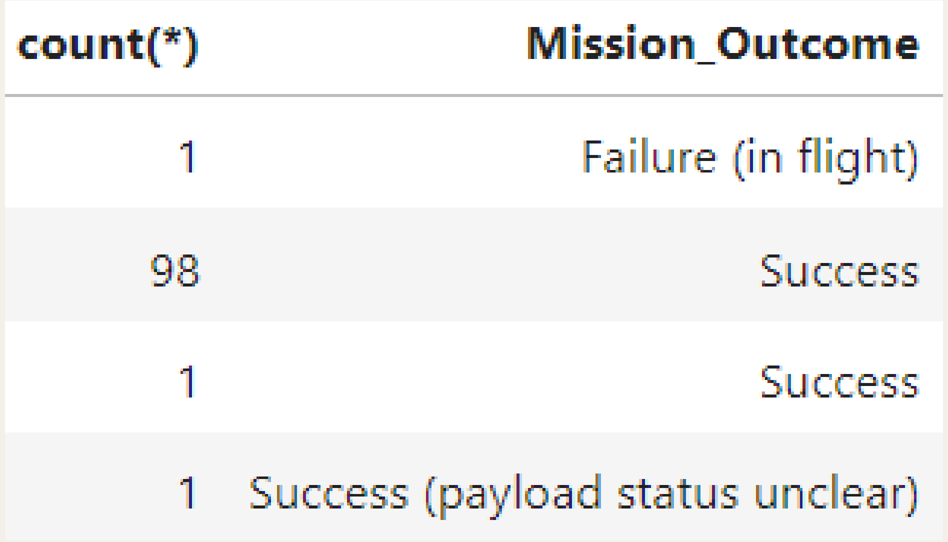


AVG(PAYLOAD_MASS__KG_)
2928.4

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

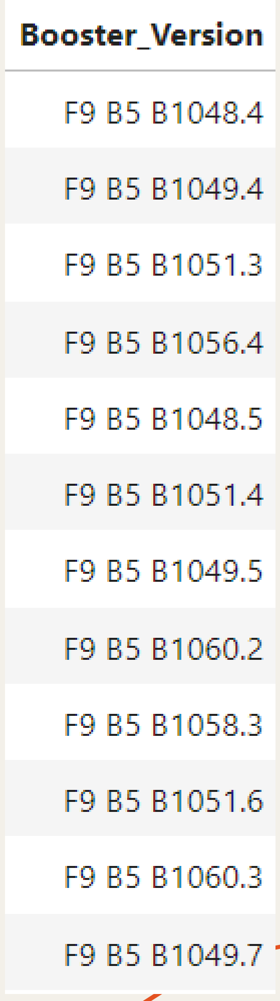
4.1 EDA with SQL

Results of Task 7: List the total number of successful and failure mission outcomes



count(*)	Mission_Outcome
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Results of Task 8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

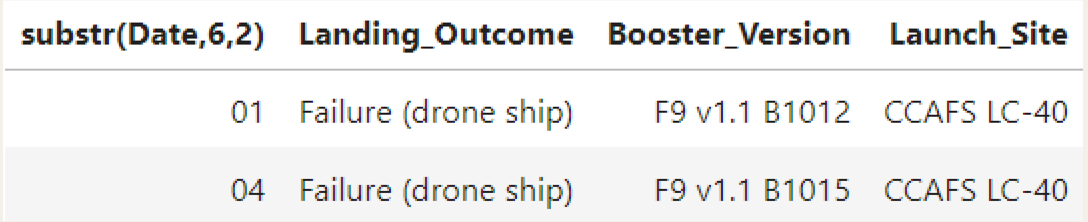


Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

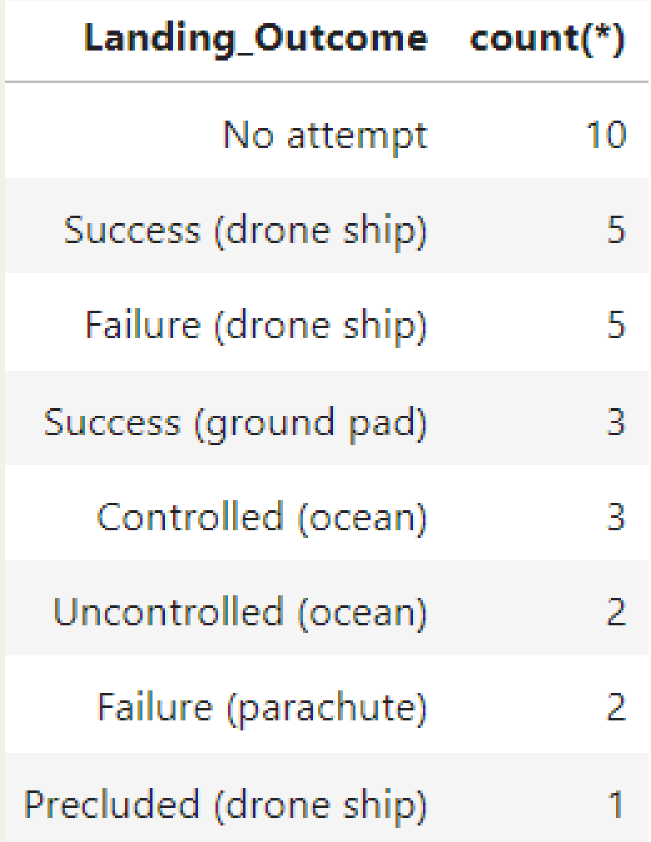
4.1 EDA with SQL

Results of Task 9: List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.



substr(Date,6,2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Results of Task 10: 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



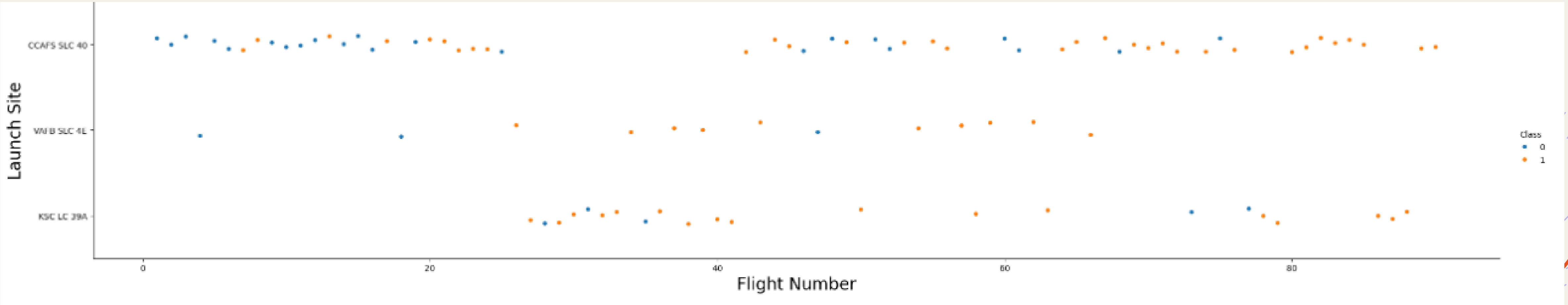
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.2 EDA using Pandas and Matplotlib

Results of Task 1: Visualize the relationship between Flight Number and Launch Site

Insights Gathered : Flight Numbers are higher in CCAFS SLC 40 , with most of them being successful

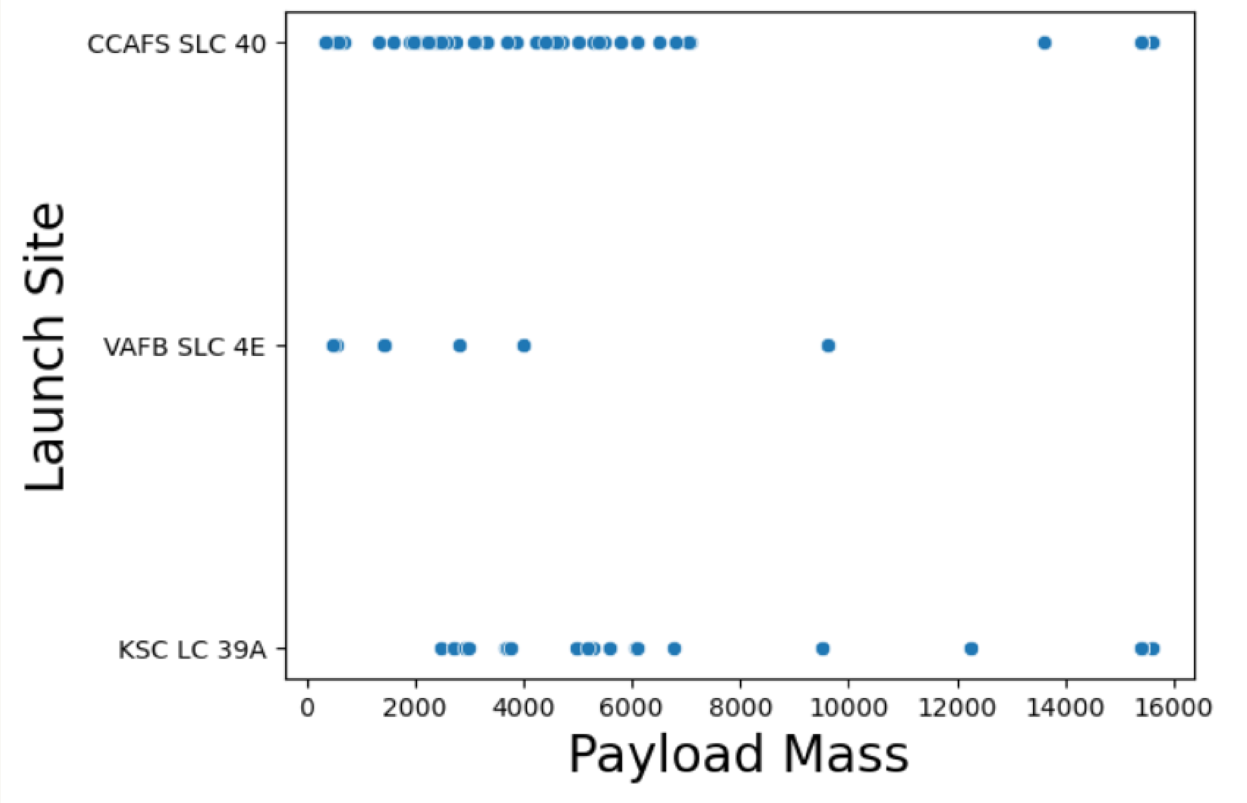


https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone

4.2 EDA using Pandas and Matplotlib

Results of Task 2: Visualize the relationship between Payload Mass and Launch Site

Insights Gathered : CCFAS SLC 40 AND KSC LC 39 A have the highest Payload Masses recorded



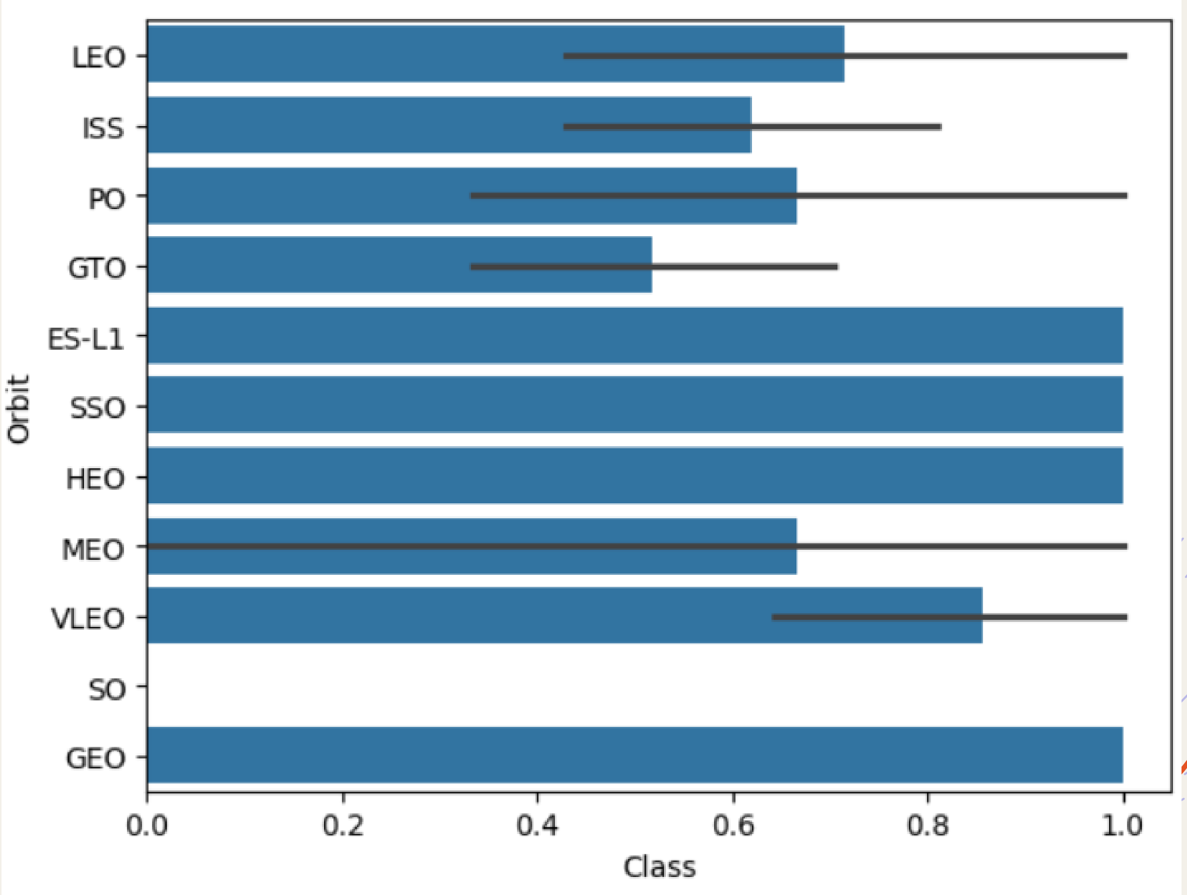
https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.2 EDA using Pandas and Matplotlib

Results of Task 3: Visualize the relationship between success rate of each orbit type

Insights Gathered : ES –L1 , SSO,HEO and GEO have the highest success rate whereas SO has the lowest success rate

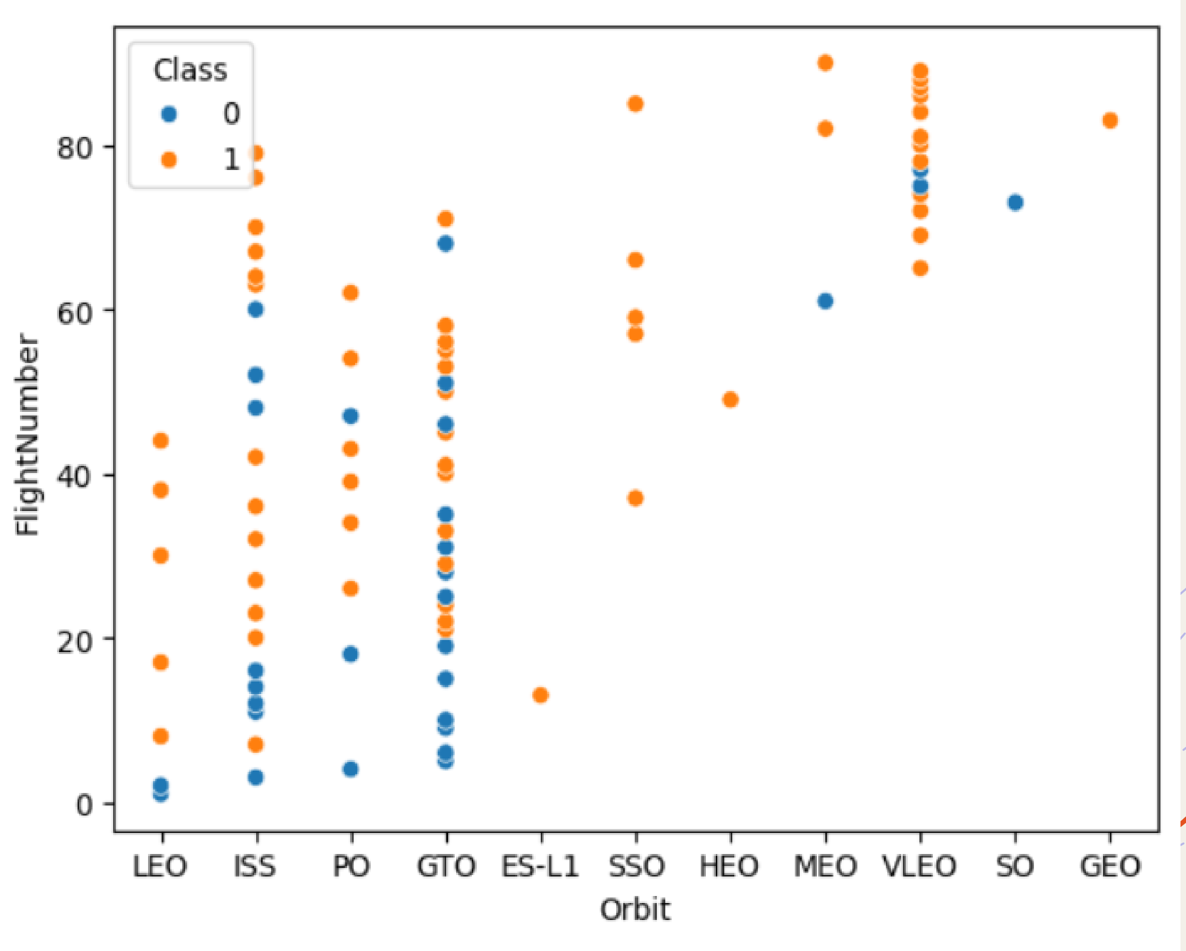
https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-



4.2 EDA using Pandas and Matplotlib

Results of Task 4 : Visualize the relationship between FlightNumber and Orbit type

Insights Gathered : In LEO orbit, higher number of flights has higher success rate. IN SSO, each flight has high success rate but there is no clear pattern fr the rest of the orbits

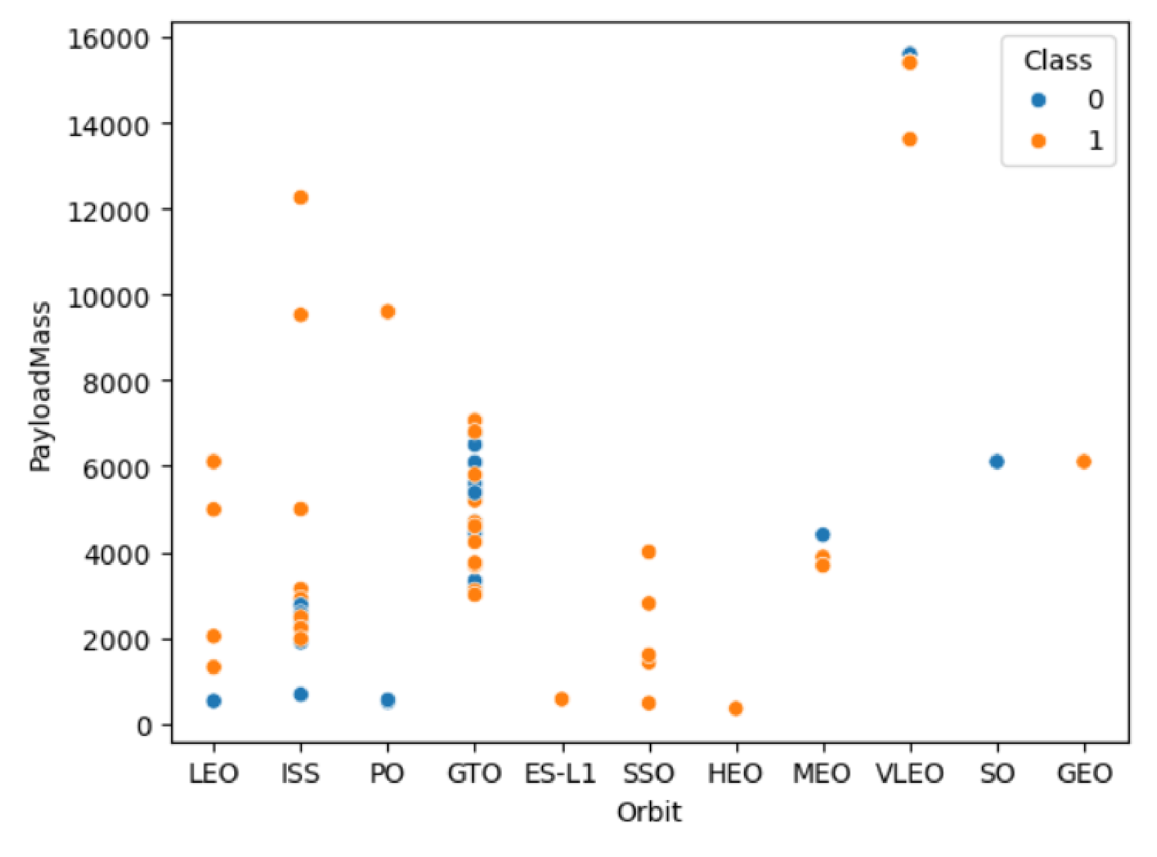


https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.2 EDA using Pandas and Matplotlib

Results of Task 5: Visualize the relationship between Payload Mass and Orbit type

Insights Gathered : for orbit types LEO, SSO, higher payload mass guarentees success rate

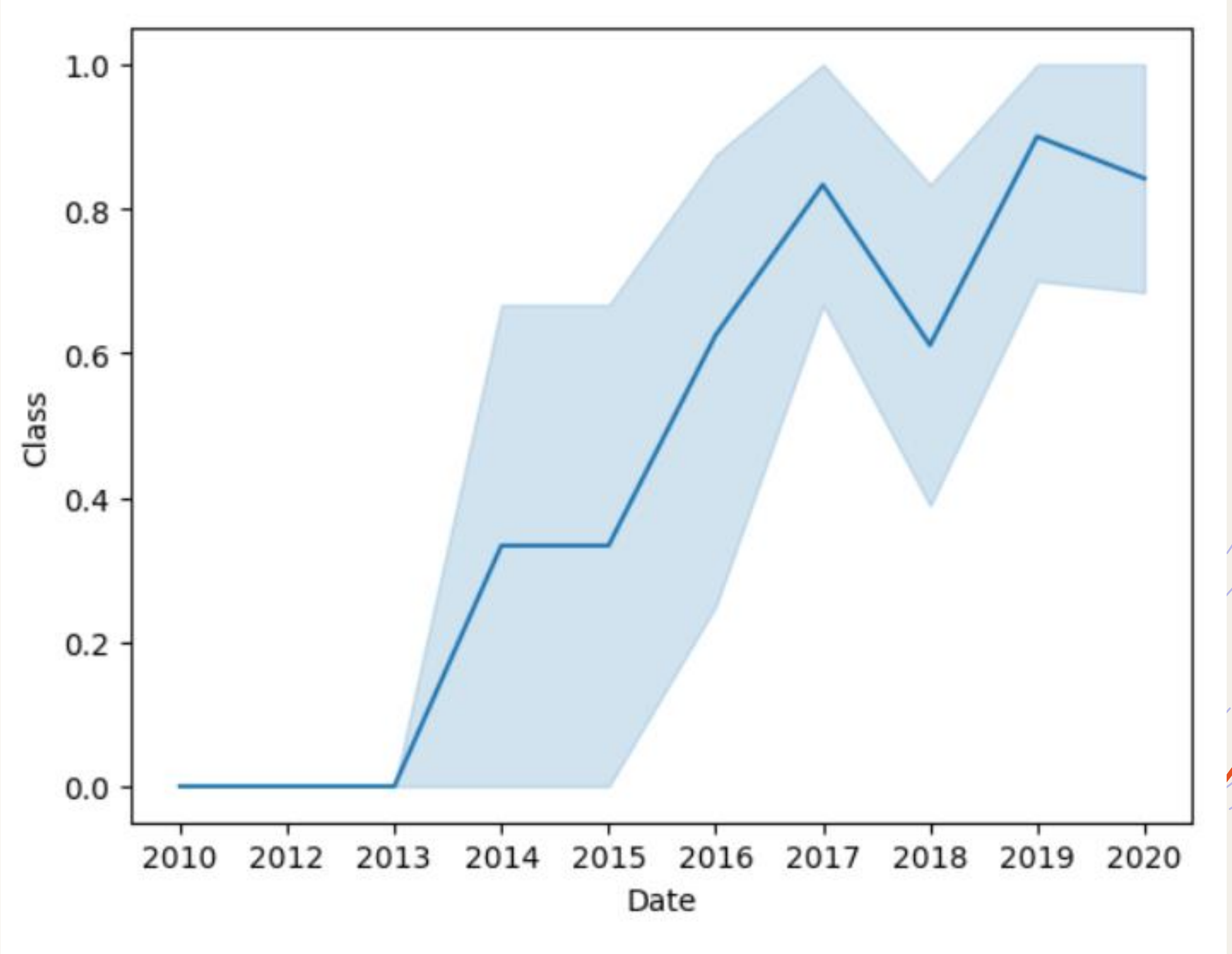


https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.2 EDA using Pandas and Matplotlib

Results of Task 6: Visualize the launch success yearly trend

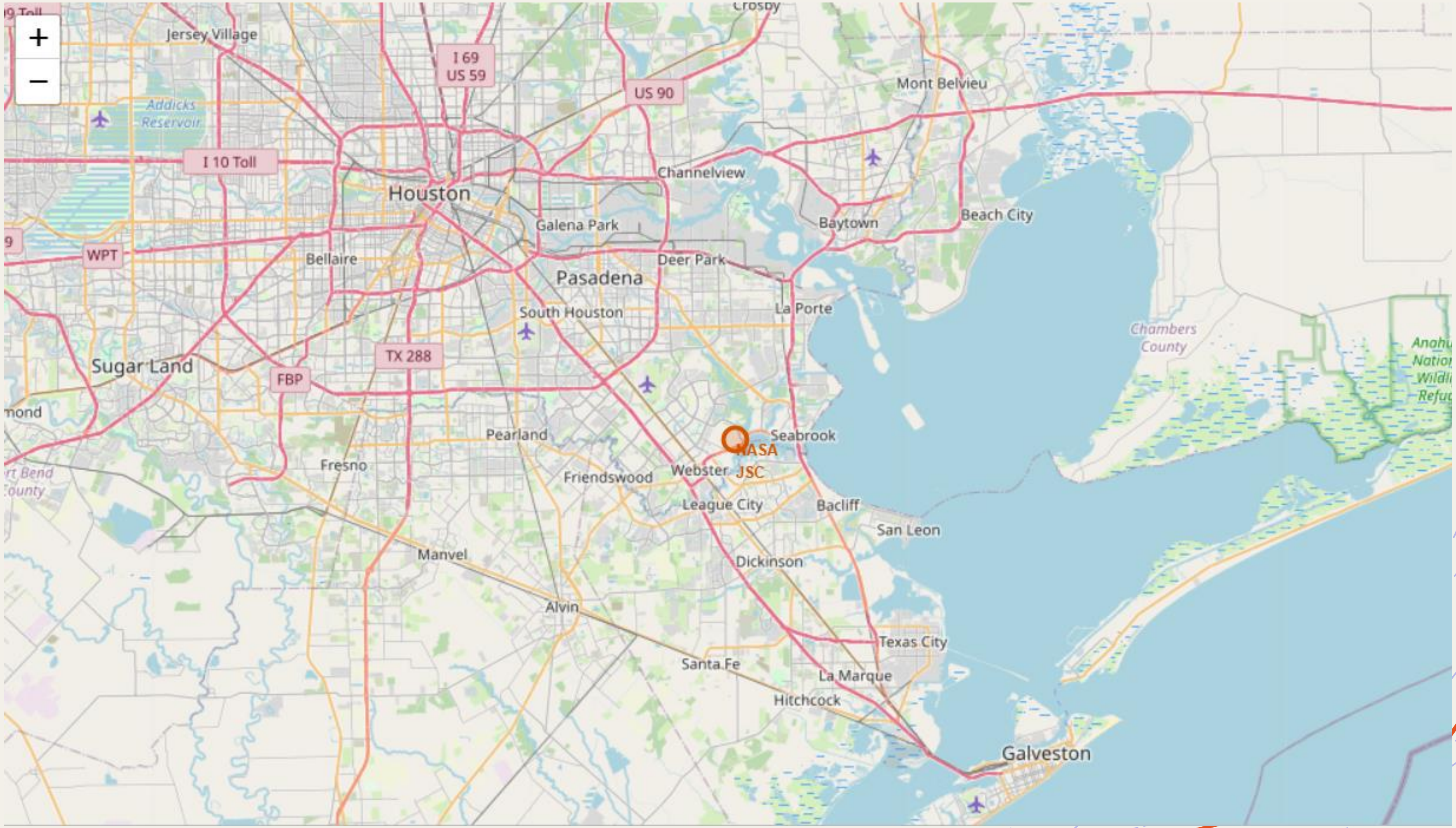
Insights Gathered : there is a steady increase in success rate form 2010 to 2020 , with a slight dip in 2018



https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.3 Interactive Visualizations with Folium

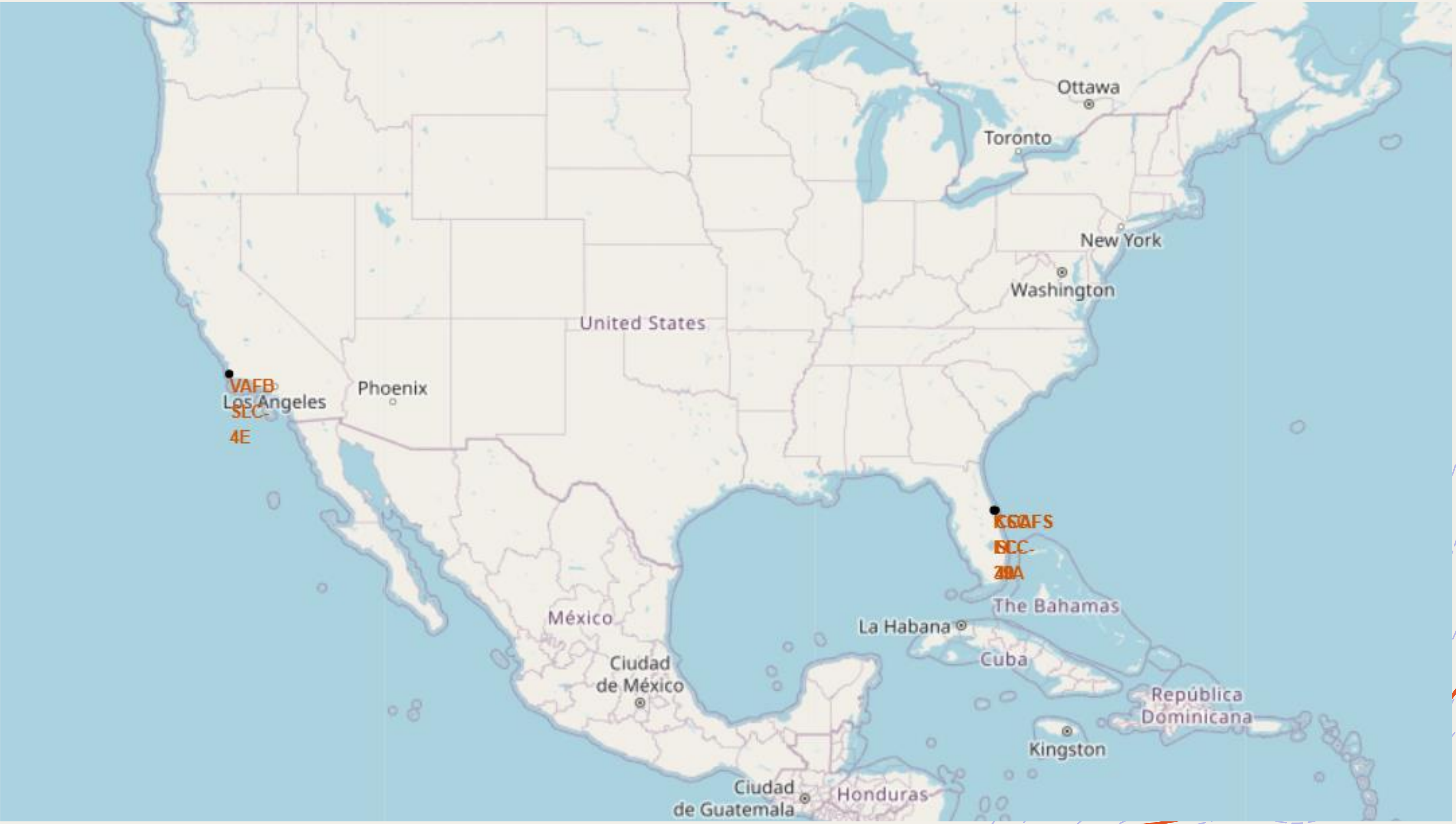
folium Map object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.



https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.3 Interactive Visualizations with Folium

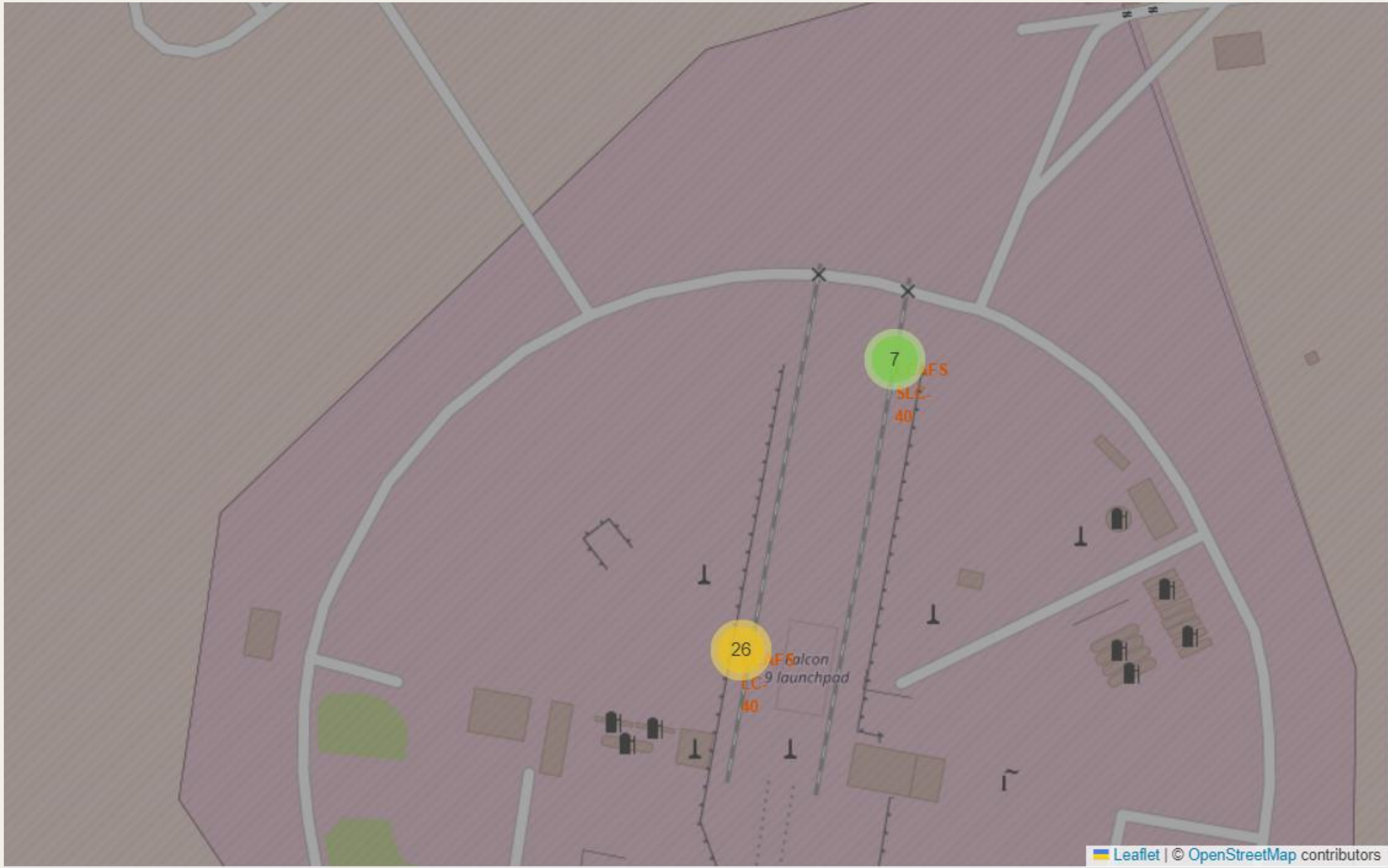
Marking Launch Sites on the Map



https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.3 Interactive Visualizations with Folium

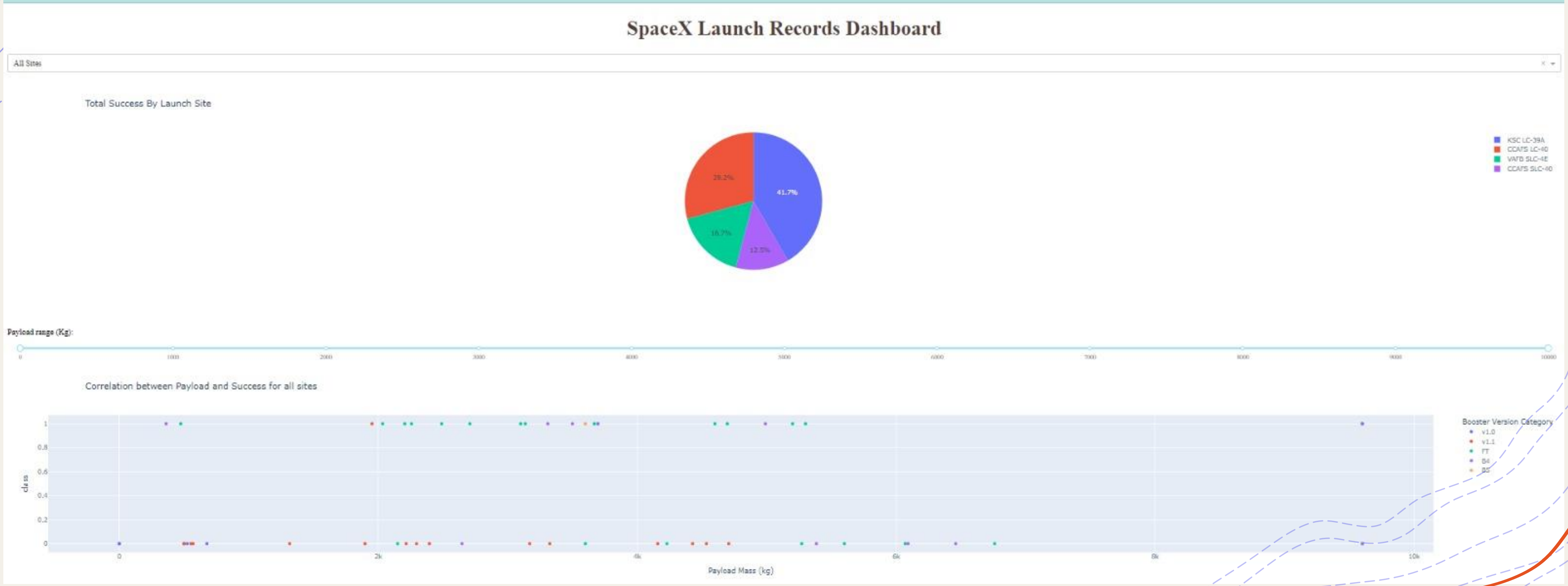
Creating markers for all launch records



https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.4 Interactive Dashboard using Dash

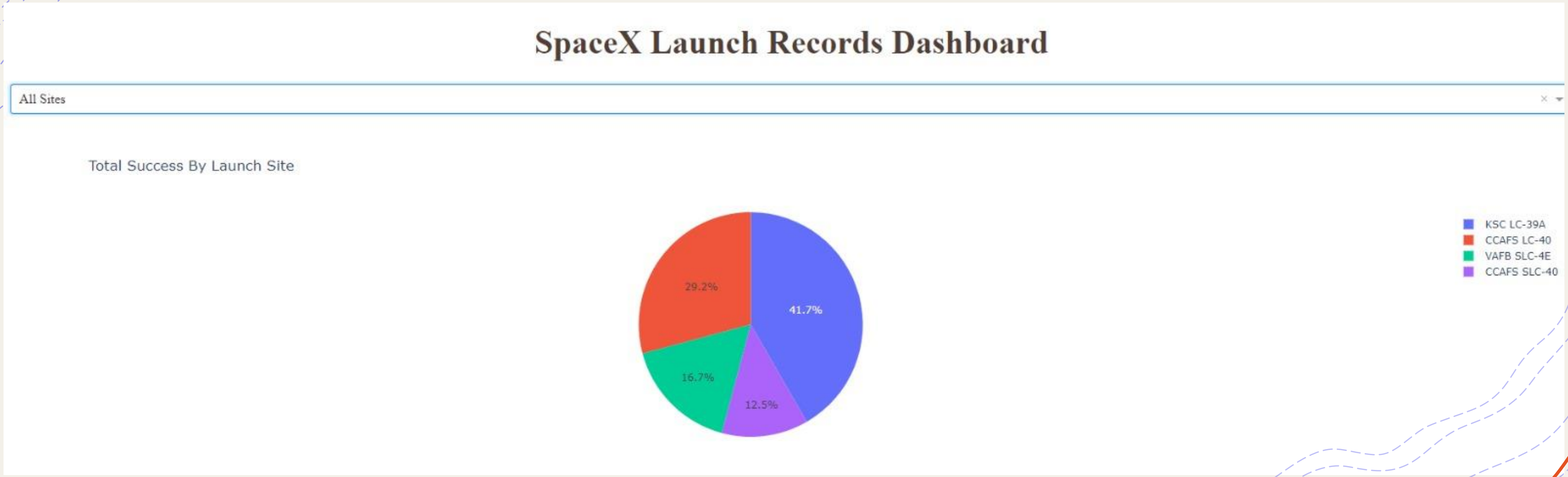
- View of the entire dashboard with dropdown and slider



https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.4 Interactive Dashboard using Dash

- KSC LC 39 A has the highest total success by launch site
- CCAFS SLC 40 has the least success by launch site



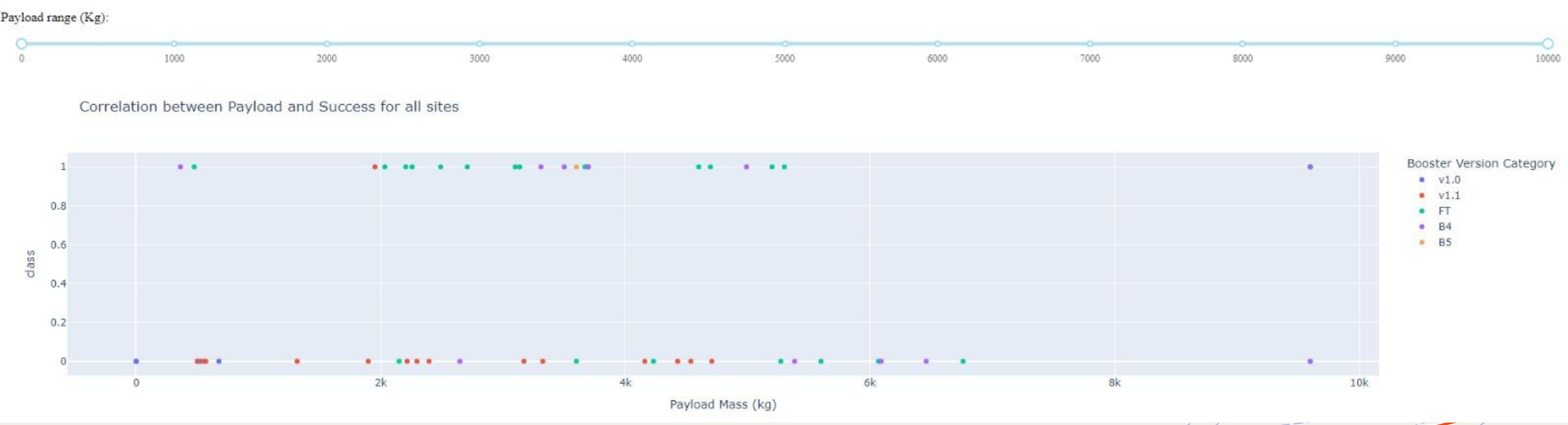
SpaceX Launch Records Dashboard

All Sites

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.4 Interactive Dashboard using Dash

- Graph for Correlation between Payload and Success for all sites
- Made interactive using the filters of payload range and Launch sites drop down
- 2k to 6k payload range has the highest amount of success rate
- FT Booster Version has the highest amount of success rate



Payload range (Kg):

Correlation between Payload and Success for all sites

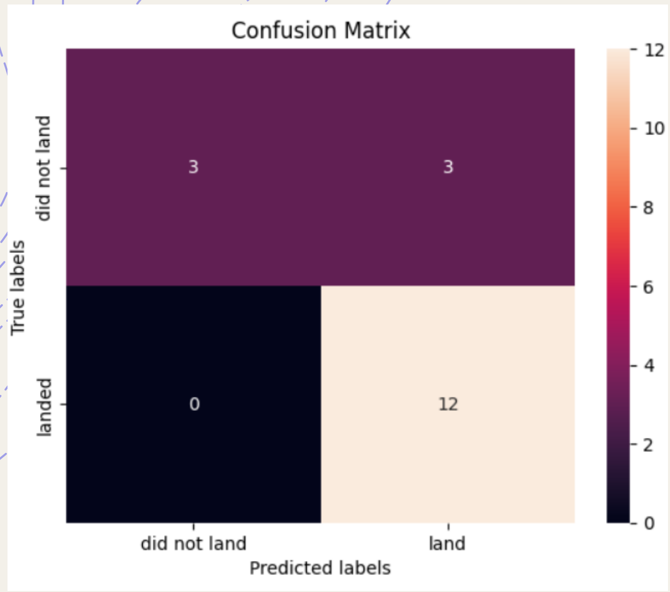
https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

4.5 Predictive Analysis using Classification

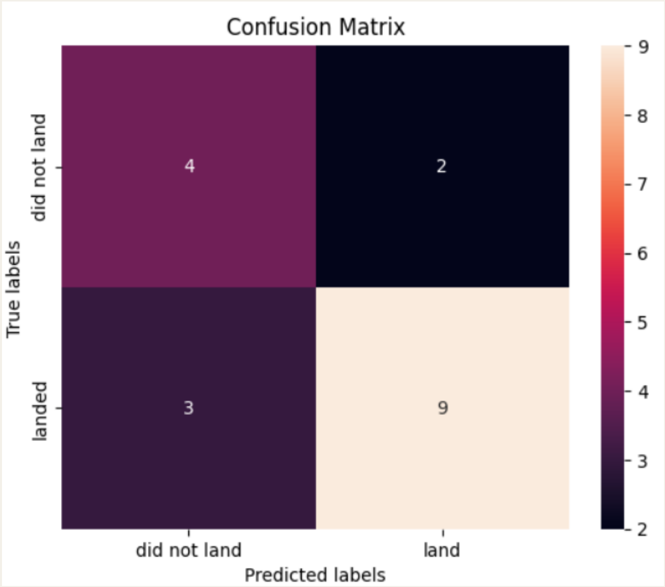
Model	Best Parameters Chosen
Logistic Regression	'C'= 0.01, 'penalty'='l2', 'solver'= 'lbfgs'
Support Vector Machine	'C'= 1.0, 'gamma'= 0.03162277660168379, 'kernel'= 'sigmoid'
Decision Tree	'criterion'='gini', 'max_depth'= 2, 'max_features'= 'sqrt', 'min_samples_leaf'= 1, 'min_samples_split'= 5, 'splitter'= 'random'
K Nearest Neighbors	'algorithm'= 'auto', 'n_neighbors'= 10, 'p'= 1

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

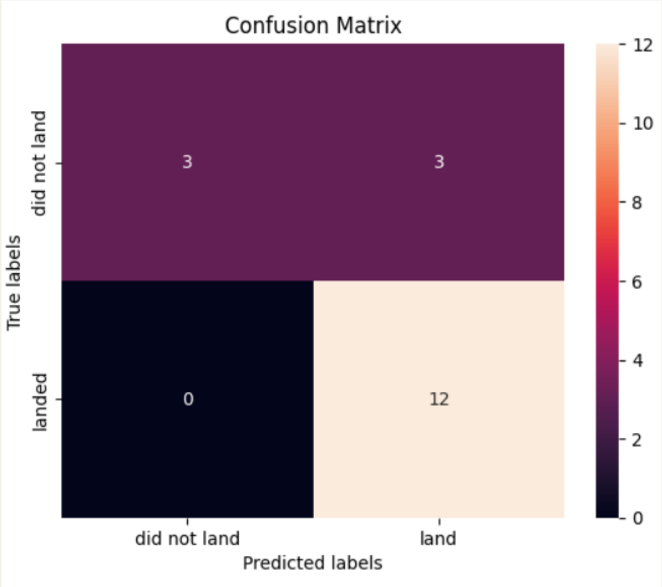
4.5 Predictive Analysis using Classification



Confusion Matrix for Logistic Regression

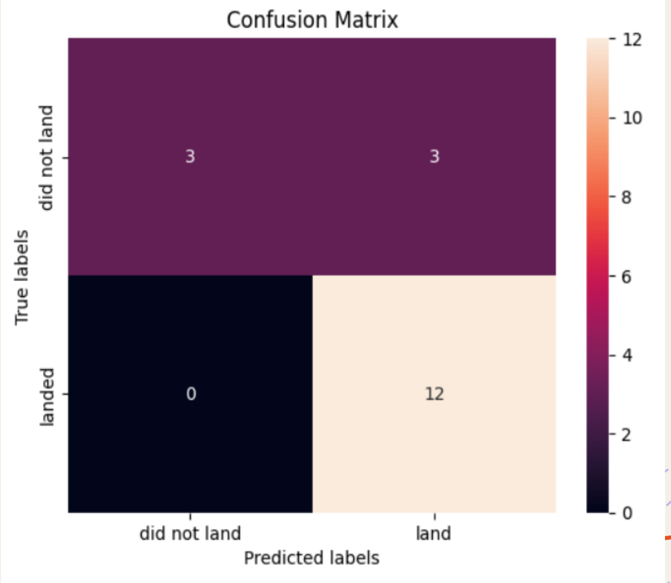


Confusion Matrix for Decision Tree



Confusion Matrix for Support Vector Machine

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-



Confusion Matrix for K Nearest Neighbors

4.5 Predictive Analysis using Classification

Model	accuracy	precision	recall	F1 score
Logistic Regression	0.8333333333333	0.8	1	0.88888888
Support Vector Machine	0.83333333333333	0.8	1	0.88888888
Decision Tree	0.72222222222222	0.81818	0.75	0.7826086
K Nearest Neighbors	0.833333333333	0.8	1	0.88888888

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

5. Conclusion

- KNN, SVM and Logistic Regression are the best performing models
- Flight Numbers are higher in CCAFS SLC 40 , with most of them being successful
- CCFAS SLC 40 AND KSC LC 39 A have the highest Payload Masses recorded
- KSC LC 39 A has the highest total success by launch site
- CCAFS SLC 40 has the least success by launch site
- 2k to 6k payload range has the highest amount of success rate
- FT Booster Version has the highest amount of success rate

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-

Thank You

https://github.com/AmruhaAhmed/IBM-Applied-Daa-Science-Capstone-