IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yash Shah
19 July, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with SQL and Data Visualization
  - Building an Interactive Map with Folium
  - Building Dashboard with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - EDA Results
  - Interactive Analytics
  - Predictive Analysis

Introduction

- Project background and context
  - Our objective is to make a prediction regarding the successful landing of the Falcon 9 first stage. On SpaceX's website, they promote Falcon 9 rocket launches at a price of $62 million, while other providers charge over $165 million per launch. The significant cost reduction is primarily attributed to SpaceX's ability to reuse the first stage. Consequently, if we can ascertain the first stage's landing outcome, we can also determine the overall launch cost.
- Problems you want to find answers
  - Identifying key factors that contribute to successful landings.
  - Analyzing the relationship between launch site, payload mass, and mission outcomes.
  - Providing insights and recommendations to optimize future launches.
  - Train a machine learning model and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

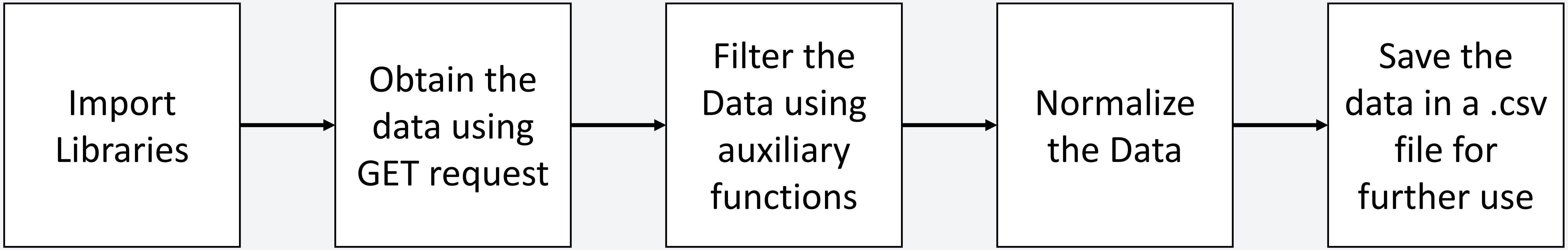
Methodology

Executive Summary

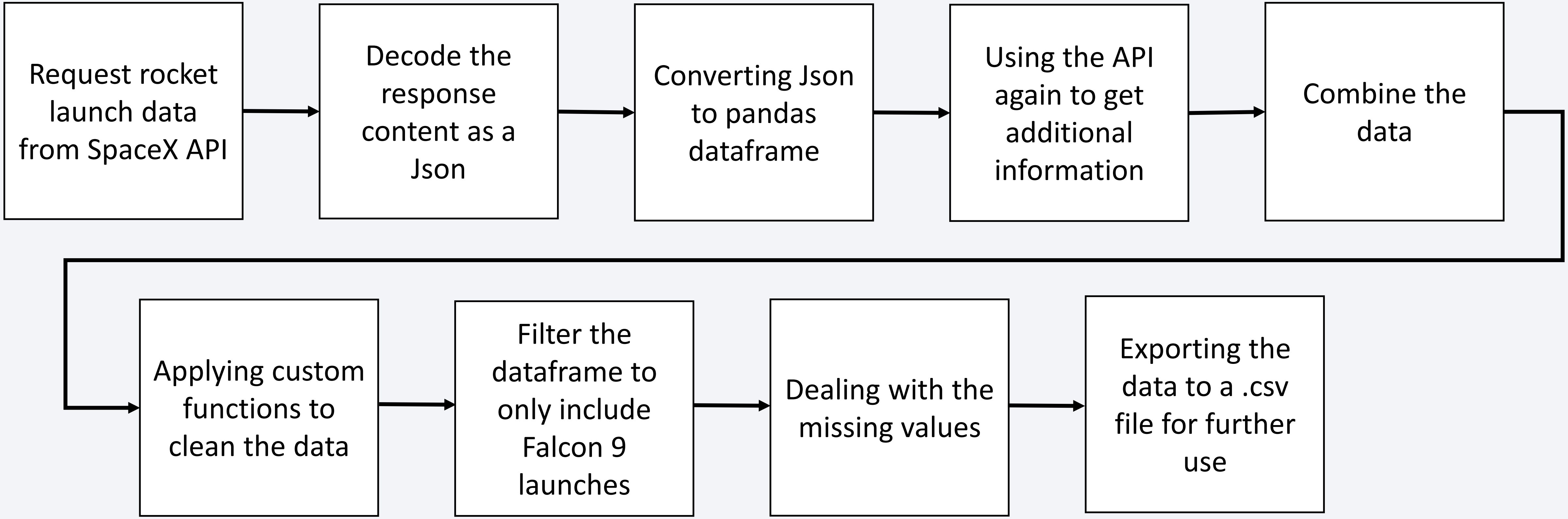
- Data collection methodology:
  - SpaceX API and Web Scrapping from Wikipedia.
- Perform data wrangling:
  - Data was cleansed and One Hot Encoding was performed.
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Python libraries, including Matplotlib and Seaborn, facilitated exploratory data analysis (EDA), while SQL was employed for data manipulation and querying.
- Perform interactive visual analytics using Folium and Plotly Dash:
  - Employ the Folium and Plotly Dash to create visual representations of previously observed correlations.
- Perform predictive analysis using classification models:
  - A machine learning pipeline was created to predict if the first stage will land or not.

Data Collection

- Data was collected using SpaceX API and Web Scrapping of a Wikipedia Page.

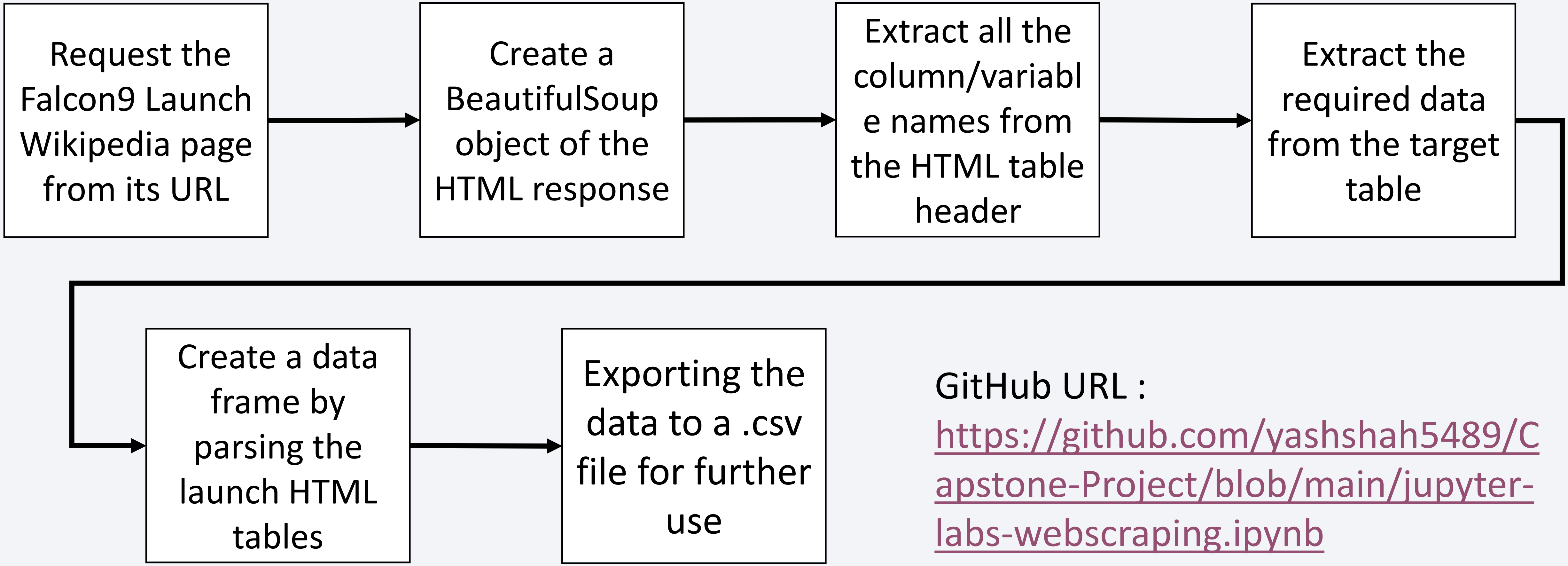


Data Collection – SpaceX API



GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

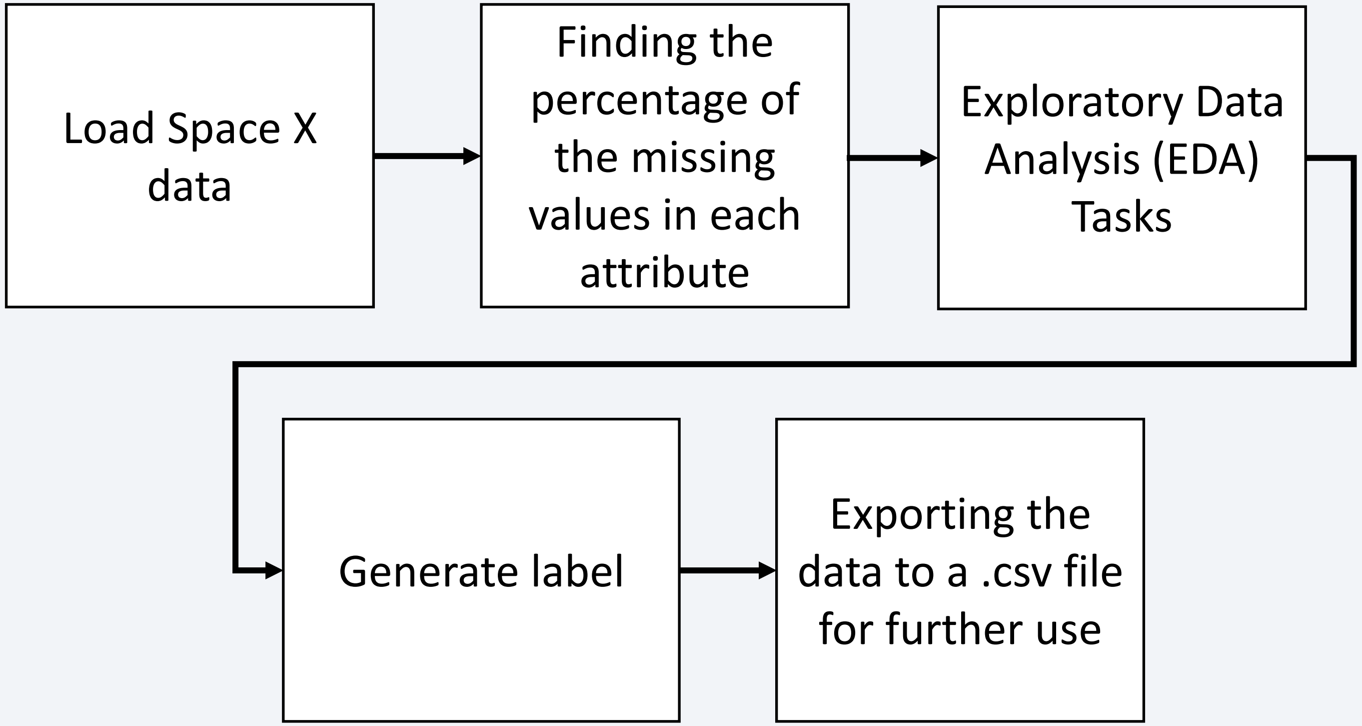
Data Collection - Scraping



GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- In this section, some Exploratory Data Analysis (EDA) is performed to find some patterns in the data and determine what would be the label for training supervised models.



GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- The following charts were plotted:
  - Categorical Plot
    1. Flight Number vs. Payload
    2. Flight Number vs. Launch site
    3. Payload vs. Launch Site
    4. Flight Number vs. Orbit
    5. Payload vs. Orbit
  - Bar Chart
    1. Orbit vs. Class
  - Line Plot
    1. Date vs. Class

- The aforementioned plots have been chosen due to their ability to prominently showcase and underscore the relationship that exists among the variables under consideration.

GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

➢ The following SQL queries were 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. Use a subquery
- 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.
- 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.

GitHub URL: https://github.com/yashshah5489/Capstone-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(3)%20(1).ipynb

Build an Interactive Map with Folium

- Markers: Used to represent launch sites, indicating their locations on the map.
- Circles: Added to highlight proximity areas, such as coastlines or cities, relative to the launch sites.
- Marker Clusters: Employed to group markers with the same coordinates for improved visualization.
- Lines: Drawn to depict distances between launch sites and their proximities.

➢ These objects were added to enhance the visual representation and facilitate exploration of the SpaceX launch sites and their relationships with various factors. They provide a clear view of site locations, proximity ranges, and distances, aiding in the analysis of geographical patterns and correlations.

GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/lab_jupyter_launch_site_location.jupyterlite%20(1).ipynb

Build a Dashboard with Plotly Dash

- Dropdown List: Allows selection of launch sites to filter the data.
- Pie Chart: Displays success counts for all sites or a selected site.
- Payload Range Slider: Enables filtering of data based on payload mass range.
- Scatter Chart: Depicts the correlation between payload mass and launch success.

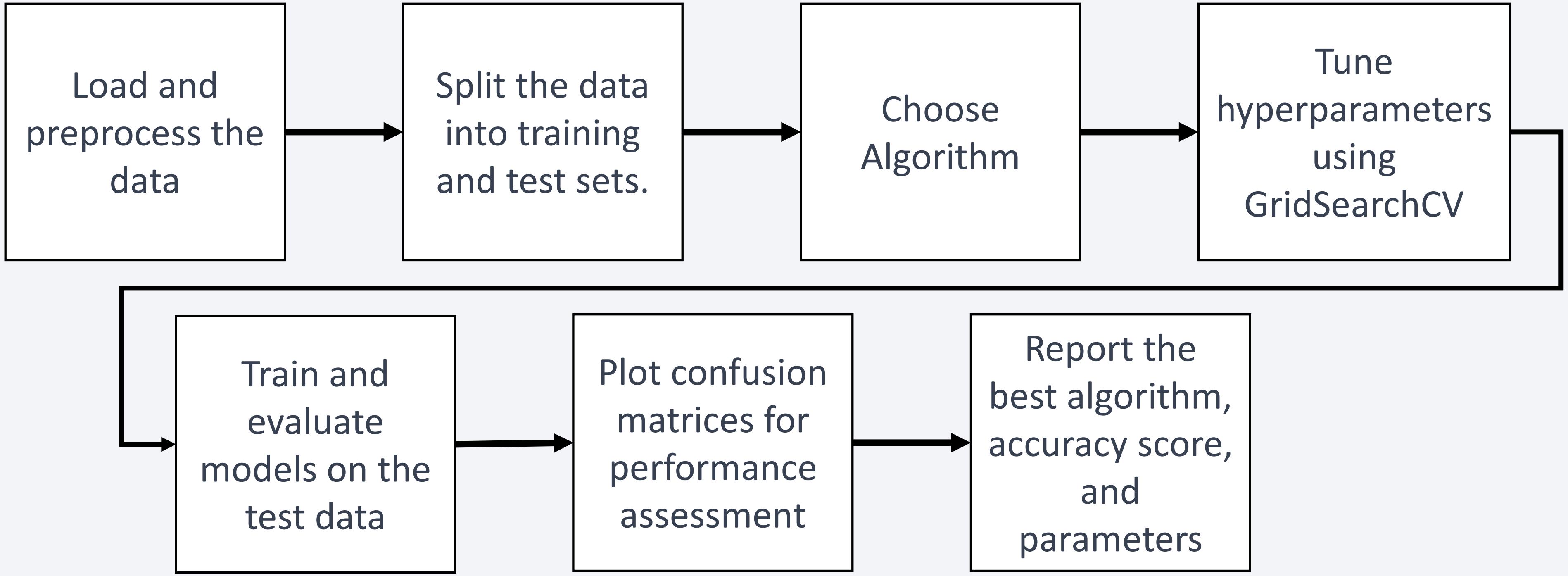
➢ These plots and interactions were added to provide a comprehensive and interactive dashboard experience. The dropdown, pie chart, range slider, and scatter chart allow users to explore different aspects of the SpaceX launch records, including launch site comparisons, success rates, payload analysis, and correlations. The interactive nature of the dashboard enables users to dynamically adjust the parameters and visually explore the data, facilitating a more intuitive understanding of the launch records.

GitHub URL :
https://github.com/yashshah5489/Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

The classification model was built by loading and preprocessing the dataset, splitting it into training and test sets. Multiple algorithms (Logistic Regression, SVM, Decision Tree, KNN) were considered. GridSearchCV was used to find the best hyperparameters for each algorithm through cross-validation. The models were trained and tuned using the GridSearchCV objects. The accuracy scores were evaluated on the test data, and the model with the highest score was selected as the best performing model.

GitHub URL : https://github.com/yashshah5489/Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



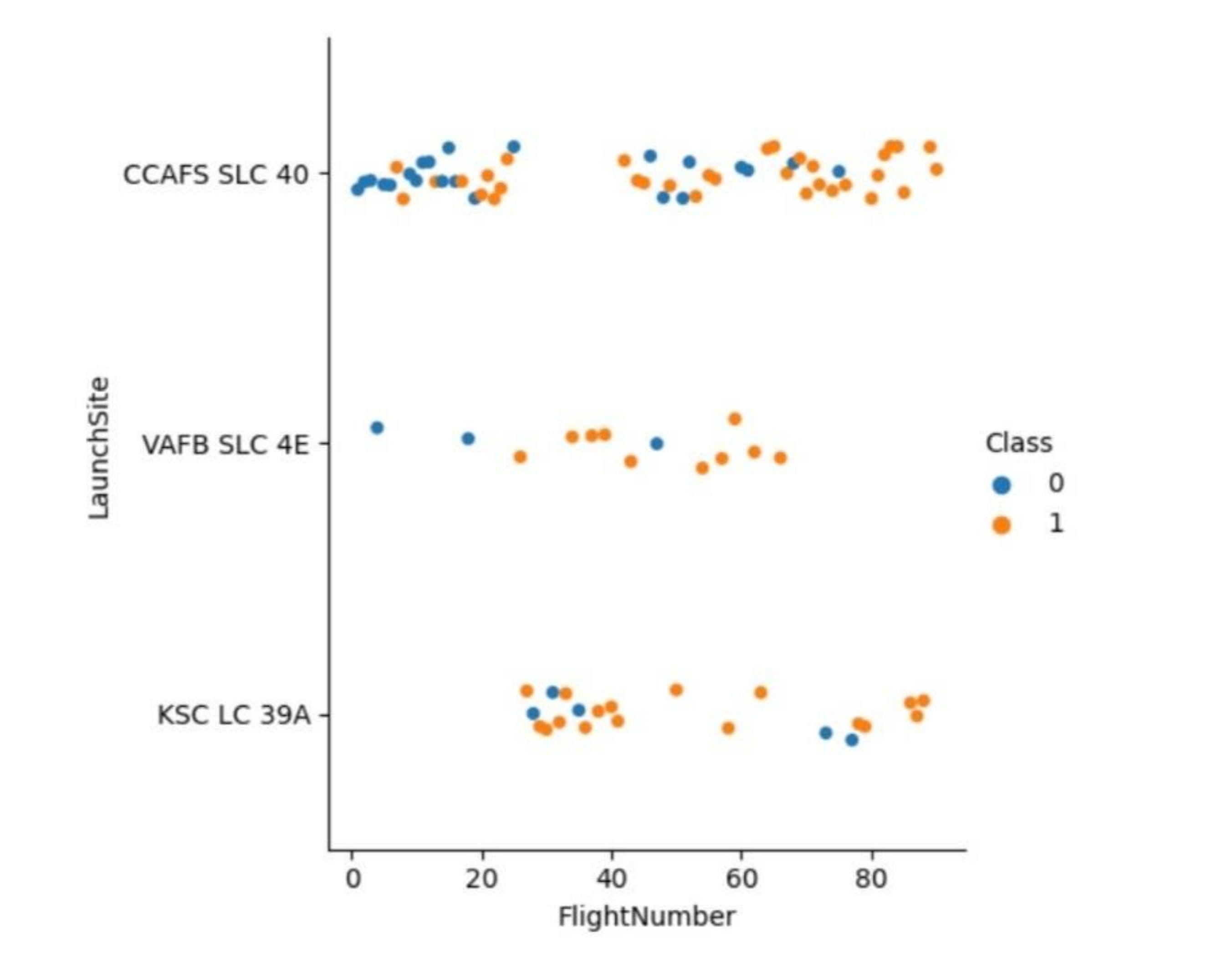
Results

- Exploratory data analysis results
  - Explored dataset of SpaceX Falcon 9 rocket launches.
  - Analyzed features like Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, etc.
  - Used descriptive statistics and visualizations.
  - Gained insights on variable distributions and correlations.
  - Identified patterns and relationships for predictive analysis.
- Interactive analytics demo in screenshots
  - Interactive demo with charts, graphs, and user interactions.
  - Filters and input components to explore specific data.
  - Enabled users to gain insights and patterns interactively.
- Predictive analysis results
  - Trained and evaluated classification algorithms (Logistic Regression, SVM, Decision Tree, KNN).
  - Tuned hyperparameters using GridSearchCV.
  - Calculated and compared accuracy scores.
  - Used confusion matrix to assess model performance.
  - Reported best algorithm, accuracy score, and optimal hyperparameters.
  - Predicted success or failure of Falcon 9 first stage landing.

Section 2

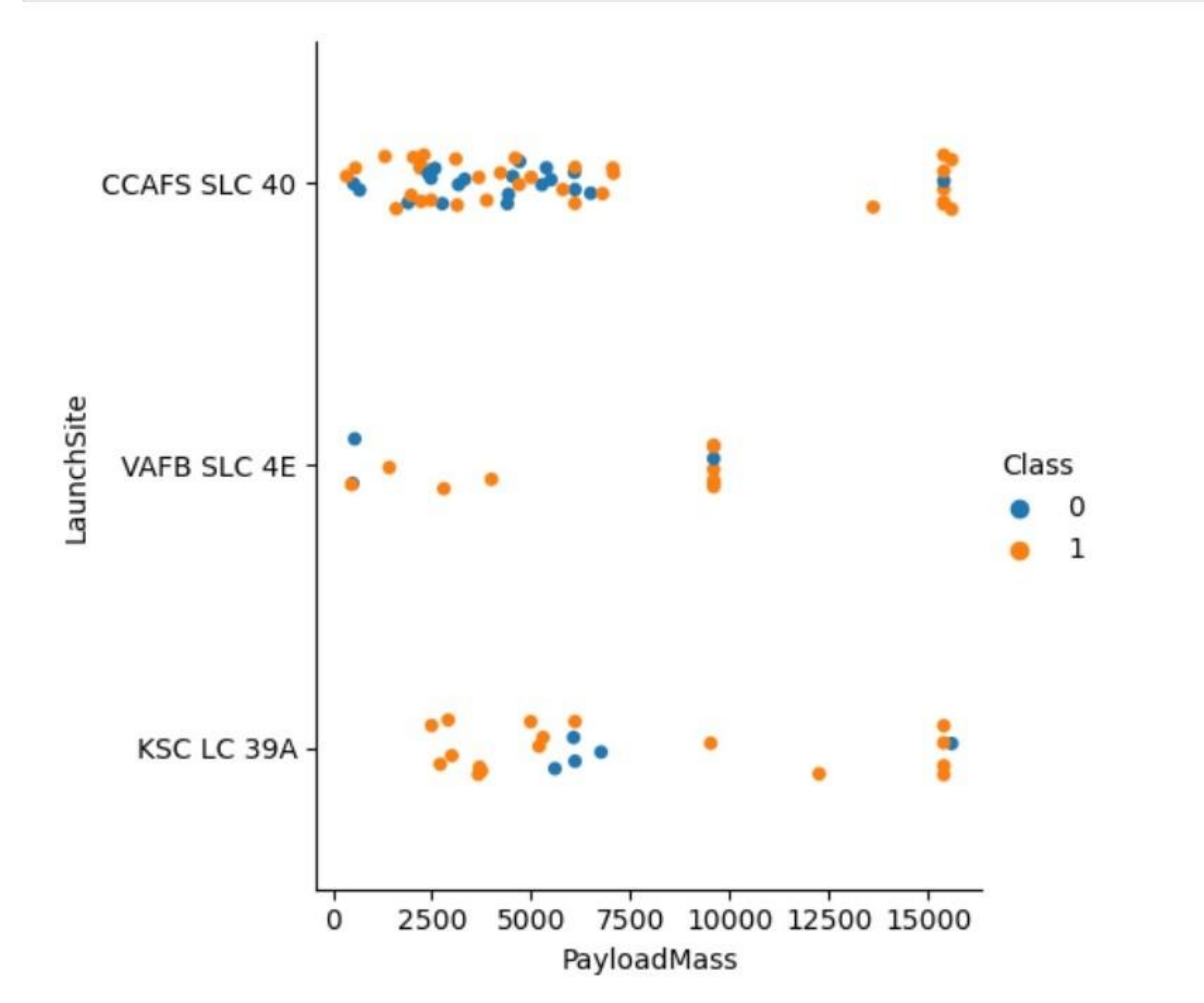
Insights drawn from EDA

Flight Number vs. Launch Site



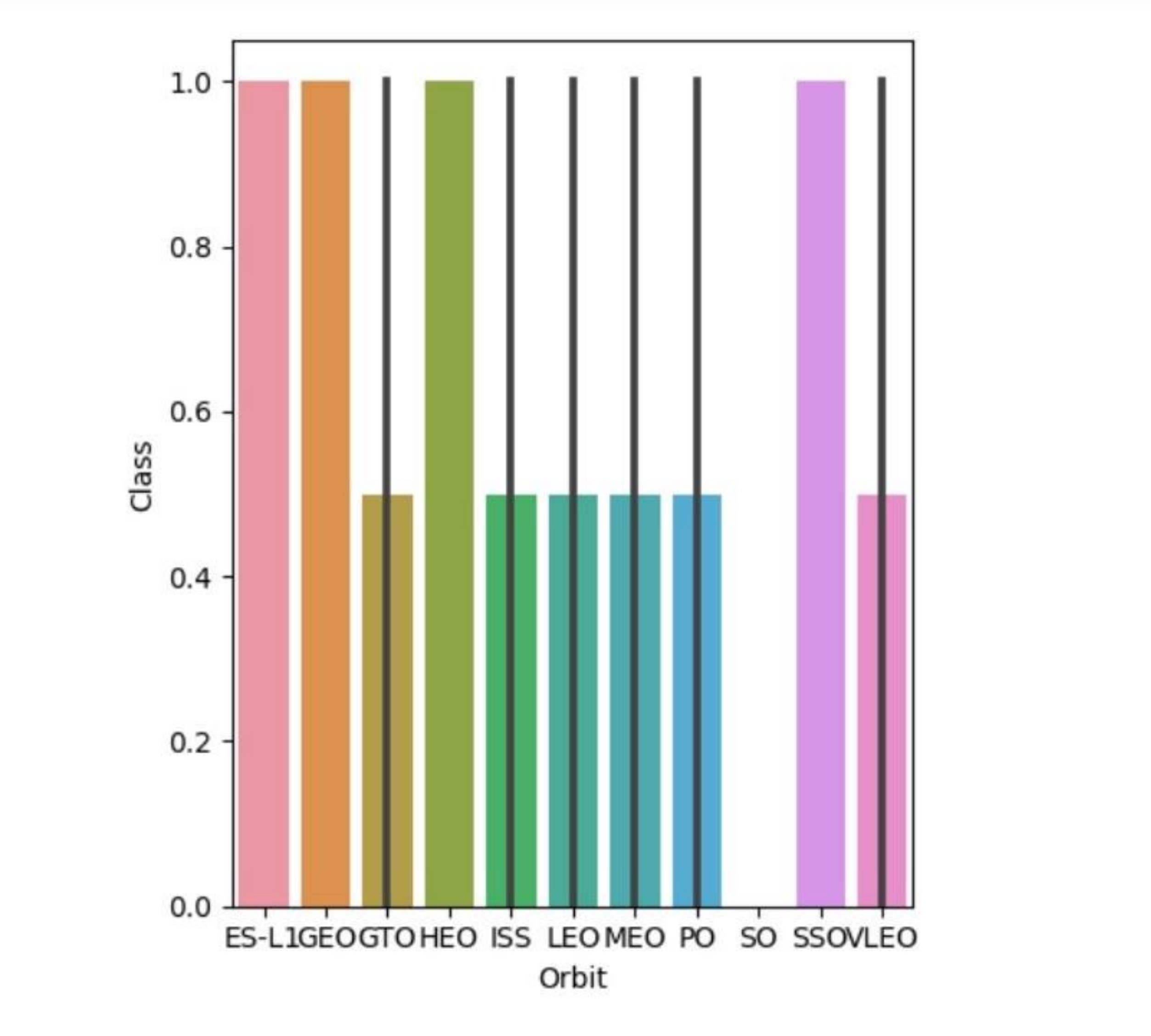
- The number of launches conducted from CCAFS SLC 40 is noticeably higher compared to the number of launches carried out from other sites in the dataset.

Payload vs. Launch Site



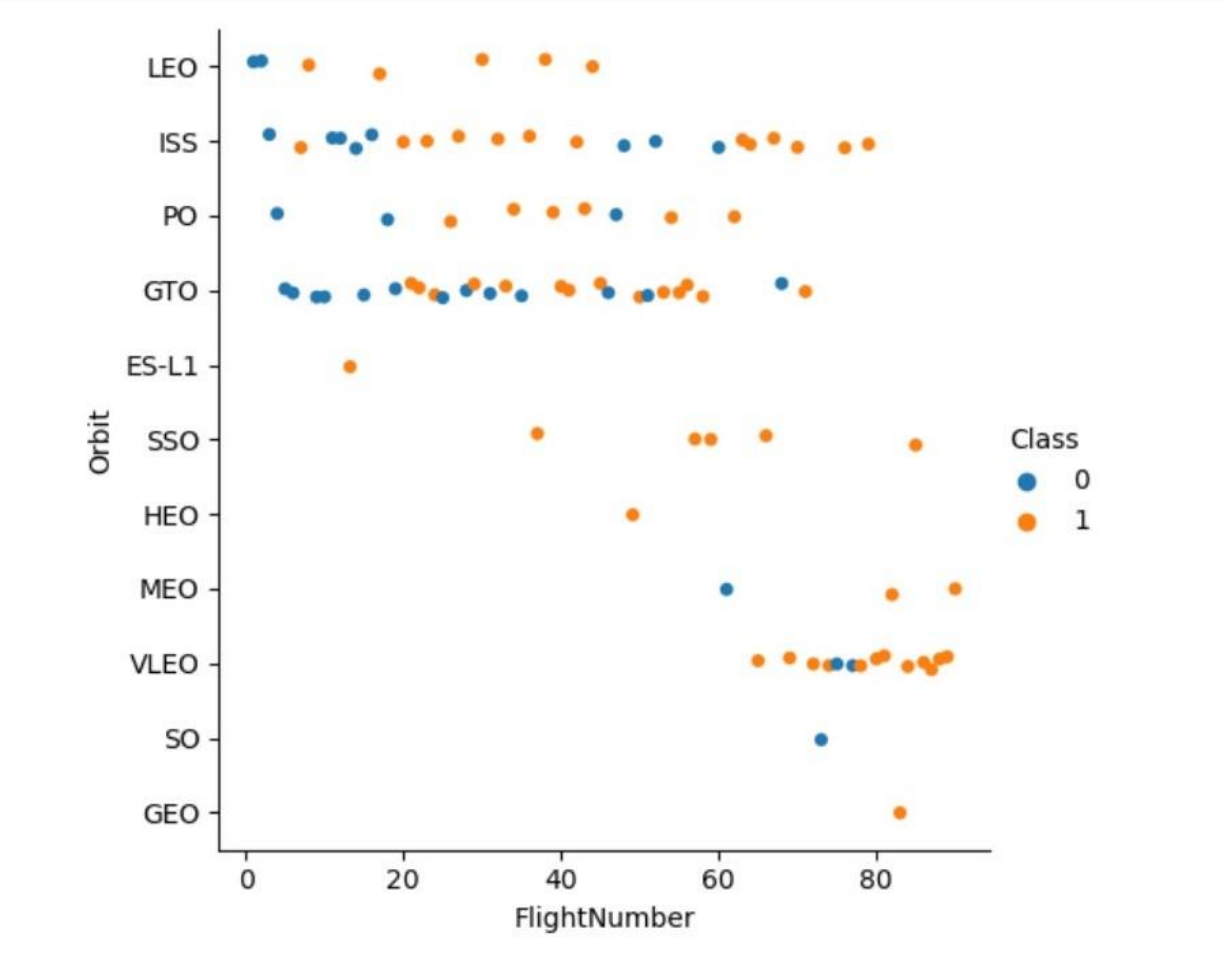
- The success rate of launches with a payload mass greater than 7500 kg is remarkably high.
- The launch site CCAFS SLC 40 is relatively more utilized than other launch sites for launching payloads with lower masses.

Success Rate vs. Orbit Type



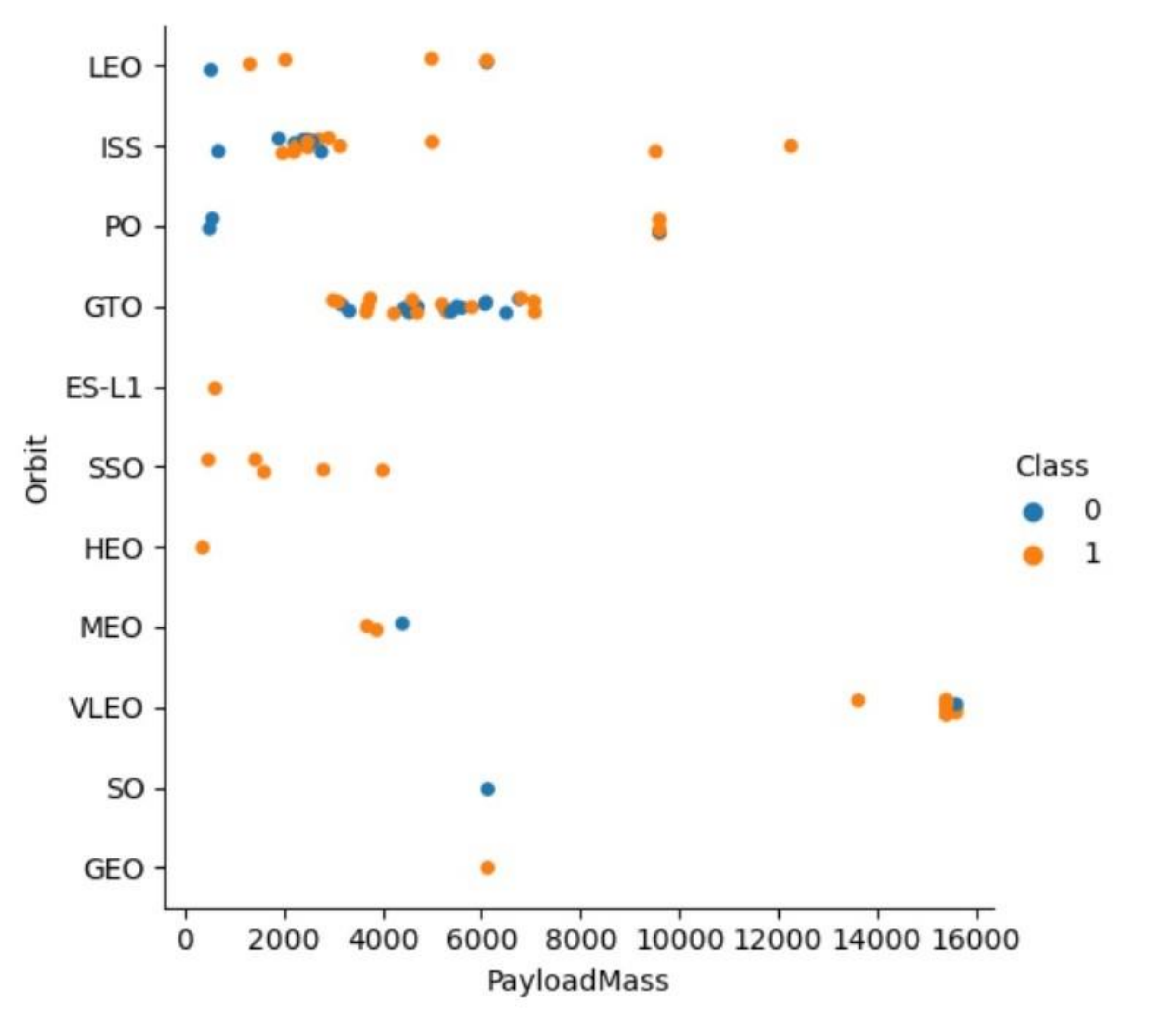
- ES-L1, GEO, HEO and SSO have the highest success rate
- SO orbit has the lowest success rate

Flight Number vs. Orbit Type



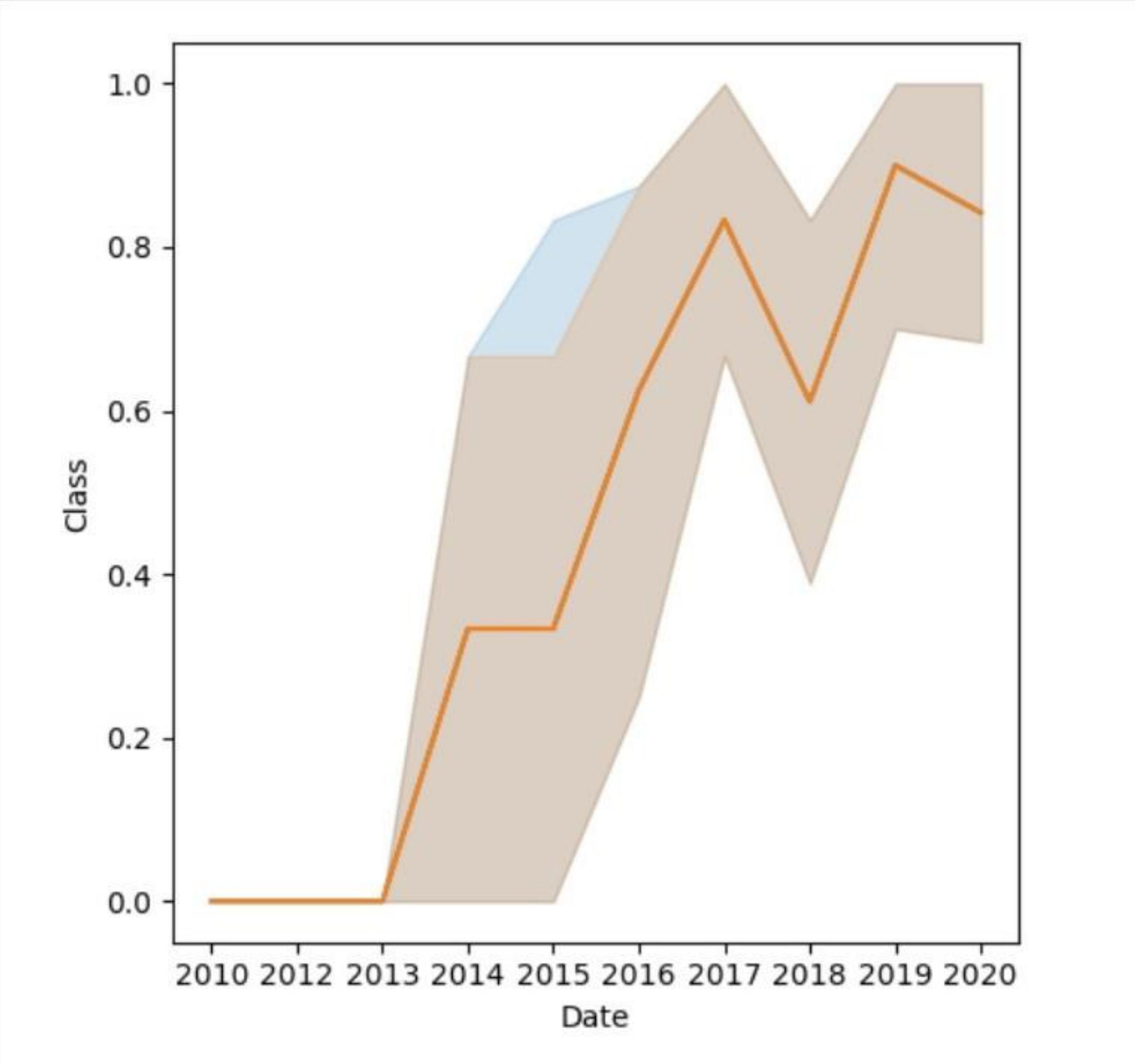
- For lower flight numbers (1-30), there is a mix of different orbit types including LEO, ISS, PO, and GTO.
- As the flight number increases, specifically from around 50 onwards, there is a noticeable increase in launches to VLEO and SSO orbits.
- VLEO and SSO orbits become more dominant in the later stages of the flight numbers.

Payload vs. Orbit Type



- There is no clear relationship between payload mass and orbit type. The payload mass varies for each orbit type, indicating that the required payload mass can differ based on the specific mission objectives and the destination orbit.

Launch Success Yearly Trend



- From 2012 to 2020, there is an upward trend in the success rate, indicating improvements in launch reliability and operational efficiency.
- During the period between 2017 and 2018, there was a brief decline in the success rate. However, it is important to note that this decrease was temporary, as the success rate promptly returned to its upward trend.
- The highest success rate is achieved in the later years, indicating the growing maturity and reliability of SpaceX's Falcon 9 rockets.

All Launch Site Names

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

```
%sql SELECT DISTINCT(LAUNCH_SITE)
FROM SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

%sql SELECT * FROM SPACEXTBL WHERE (LAUNCH_SITE) LIKE ‘CCA%’ LIMIT 5;

Total Payload Mass

SUM (PAYLOAD_MASS__kg_)
45596.0

%sql SELECT
SUM(PAYLOAD_MASS_KG)
FROM SPACEXTBL WHERE
CUSTOMER = ‘NASA(CRS)’;

Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

%sql SELECT AVG(PAYLOAD_MASS_KG_) FROM SPACEXTBL
WHERE BOOSTER_VERSION = ‘F9 v1.1’;

First Successful Ground Landing Date

First_Successful_Landing_Date
22/12/2015

```
%sql SELECT MIN(DATE) AS
First_Successful_Landing_Date FROM SPACEXTBL
WHERE Landing_Outcome = ‘Success (ground pad)’;
```

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

%sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome==‘Success (drone ship)’ AND PAYLOAD_MASS_KG_>4000 AND PAYLOAD_MASS_KG_<6000;

Total Number of Successful and Failure Mission Outcomes

successful mission	failure mission
100	1

```
%sql SELECT (SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTBL
WHERE MISSION_OUTCOME LIKE '%Success%') AS "successful mission",
(SELECT COUNT(MISSION_OUTCOME) FROM SPACEXTBL WHERE
MISSION_OUTCOME LIKE '%Fail%') AS "failure mission";
```

Boosters Carried Maximum Payload

boosterversion
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

```
%sql SELECT BOOSTER_VERSION AS boosterversion from SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

2015 Launch Records

Month	Failure_Landing_Outcomes	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

```
%sql SELECT substr(Date, 4, 2) AS Month, Landing_Outcome AS
Failure_Landing_Outcomes,  Booster_Version, Launch_Site FROM SPACEXTBL
WHERE substr(Date, 7, 4) = '2015' AND Landing_Outcome LIKE 'Failure (drone
ship)'
```

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

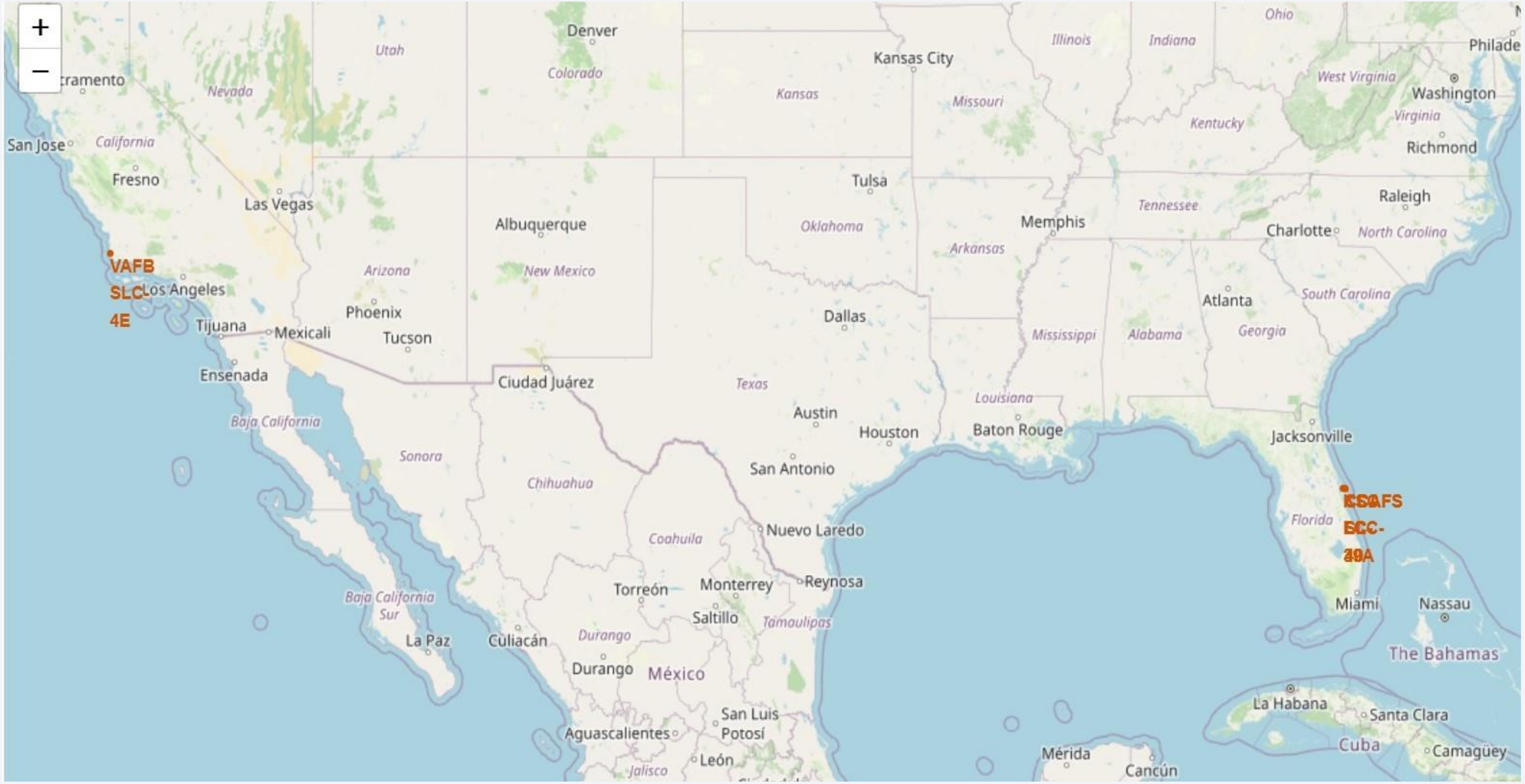
Landing_Outcome	Count_Landing_Outcomes
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

%sql SELECT Landing_Outcome, COUNT(*) AS Count_Landing_Outcomes FROM SPACEXTBL WHERE substr(Date, 7, 4) OR substr(Date, 4, 2) BETWEEN '2010-06' AND '2017-03' GROUP BY Landing_Outcome ORDER BY Count_Landing_Outcomes DESC;

Section 3

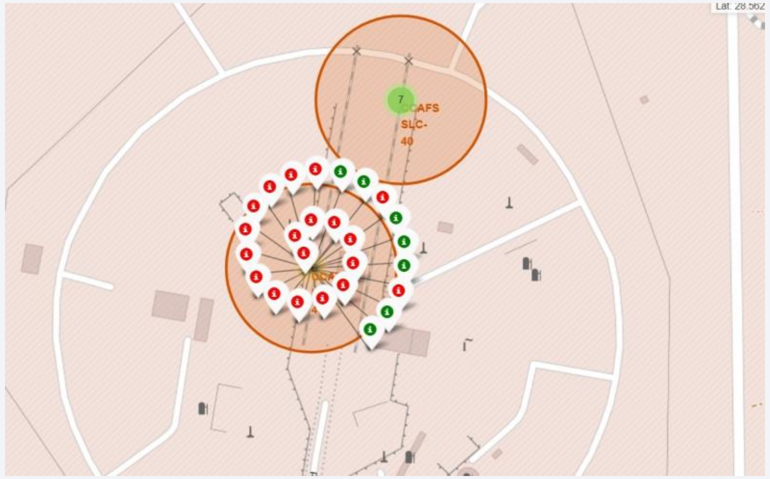
Launch Sites Proximities Analysis

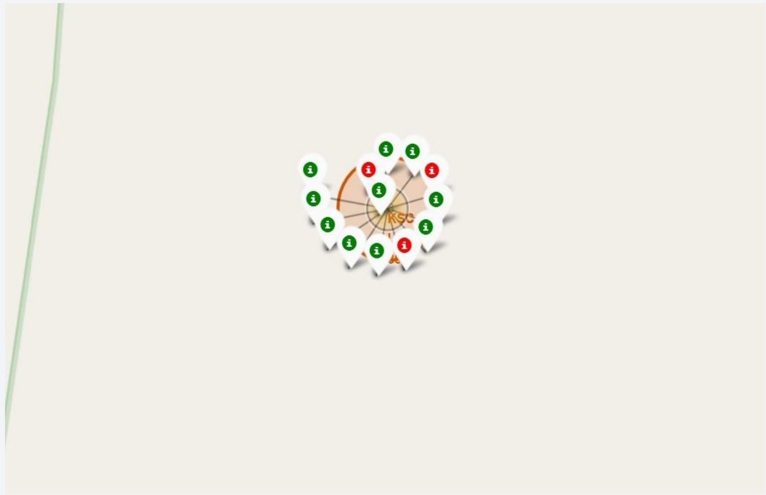
<Folium Map Screenshot 1>



All launch sites in the U.S.A. are marked on the map.

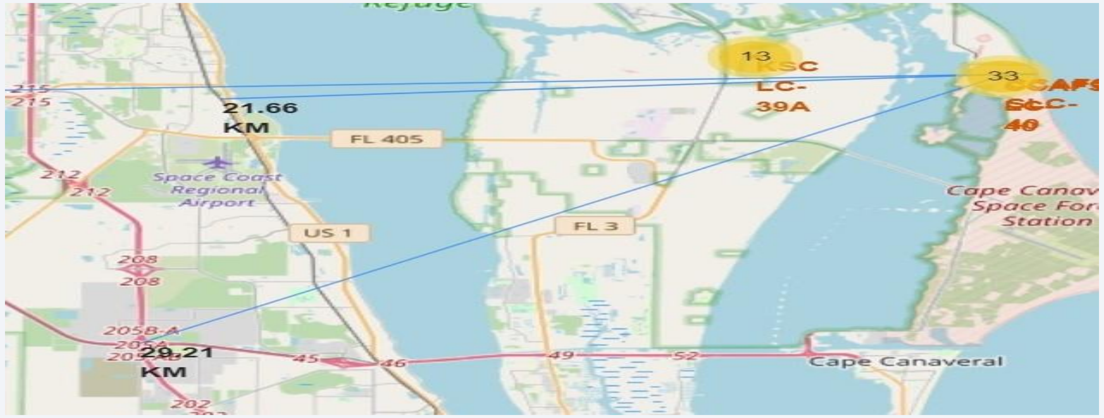
<Folium Map Screenshot 2>

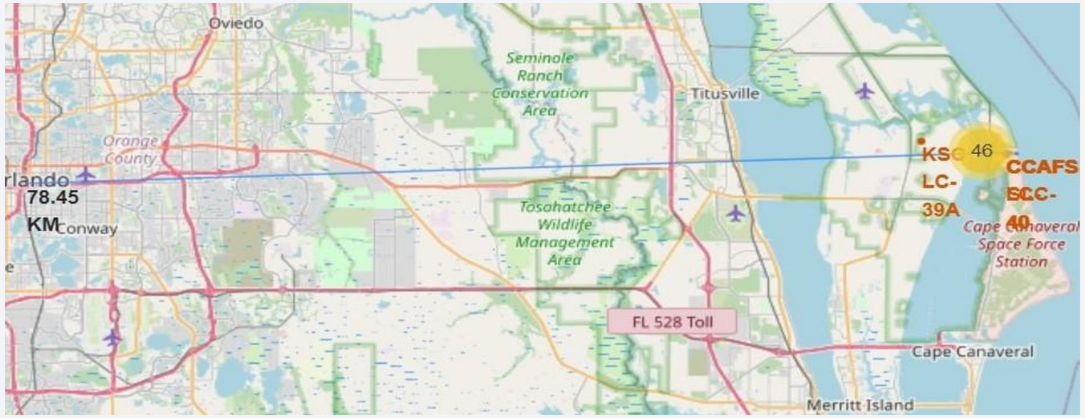




If marker colour is green then it means outcome was successful and if it is red then it means that outcome was failure.

<Folium Map Screenshot 3>



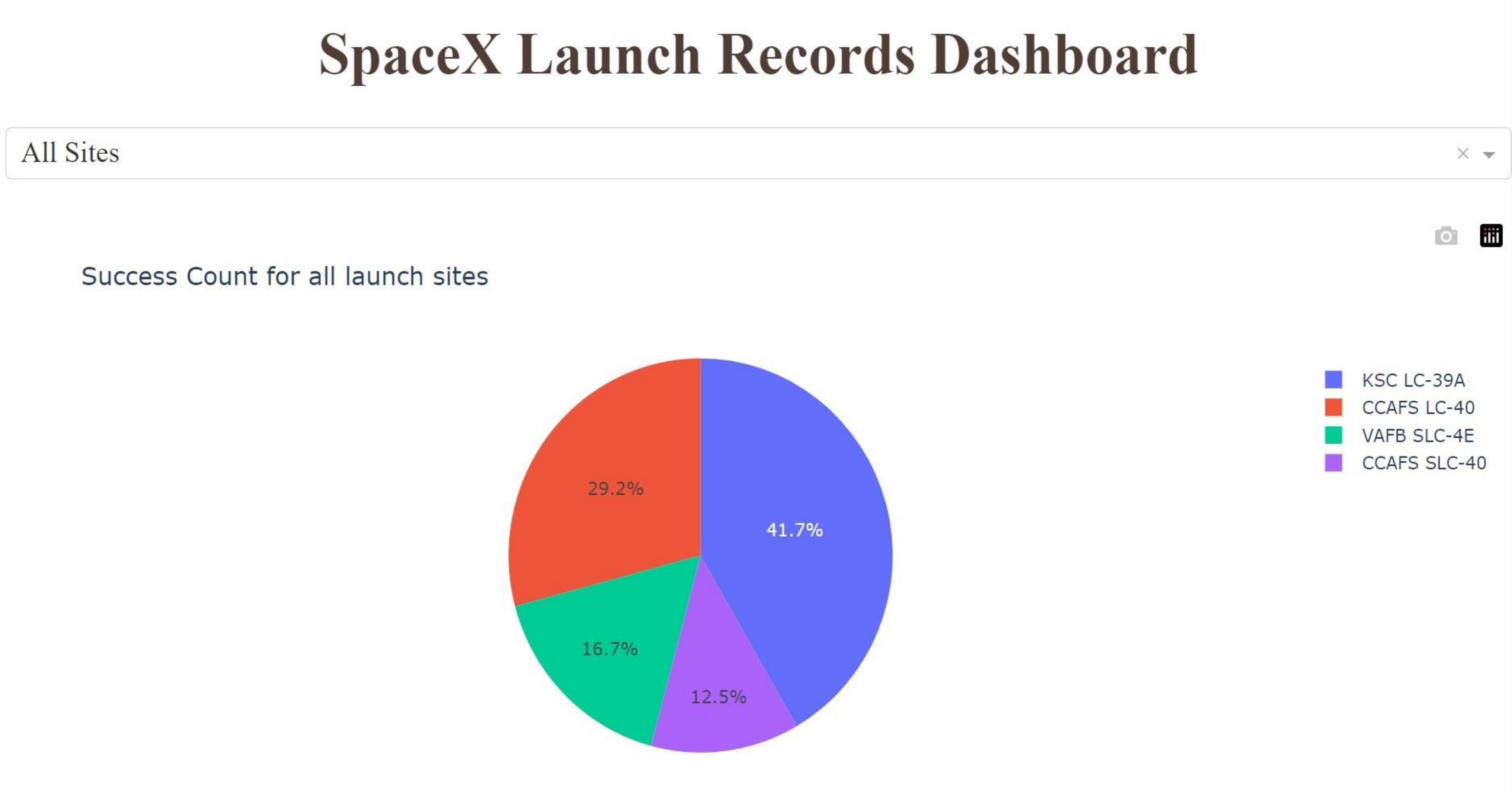


- The railway, highway, and coastline are represented by lines connecting the launch site to these points of interest.
- Each line is accompanied by a marker that displays the calculated distance from the launch site to that specific point of interest.

Section 4

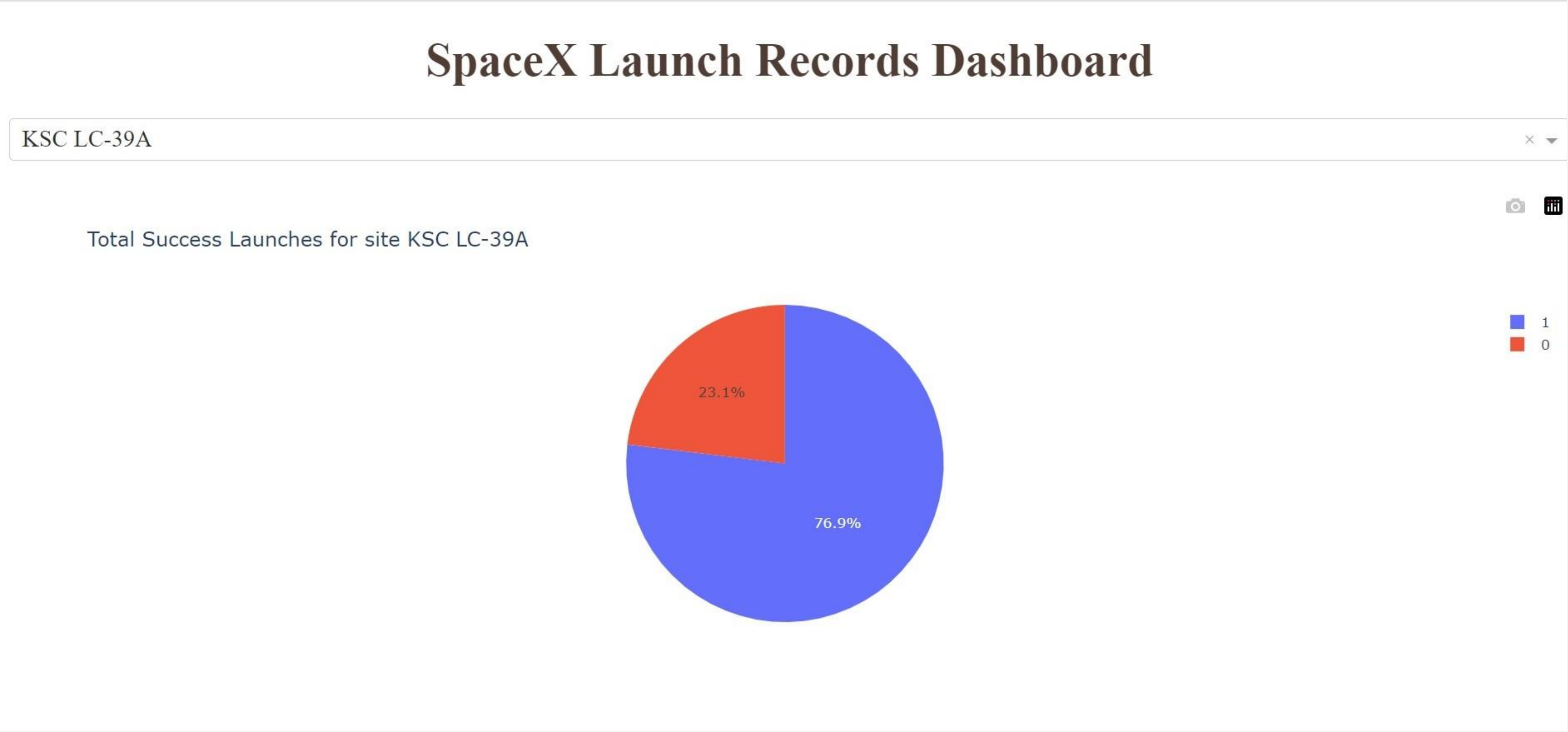
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



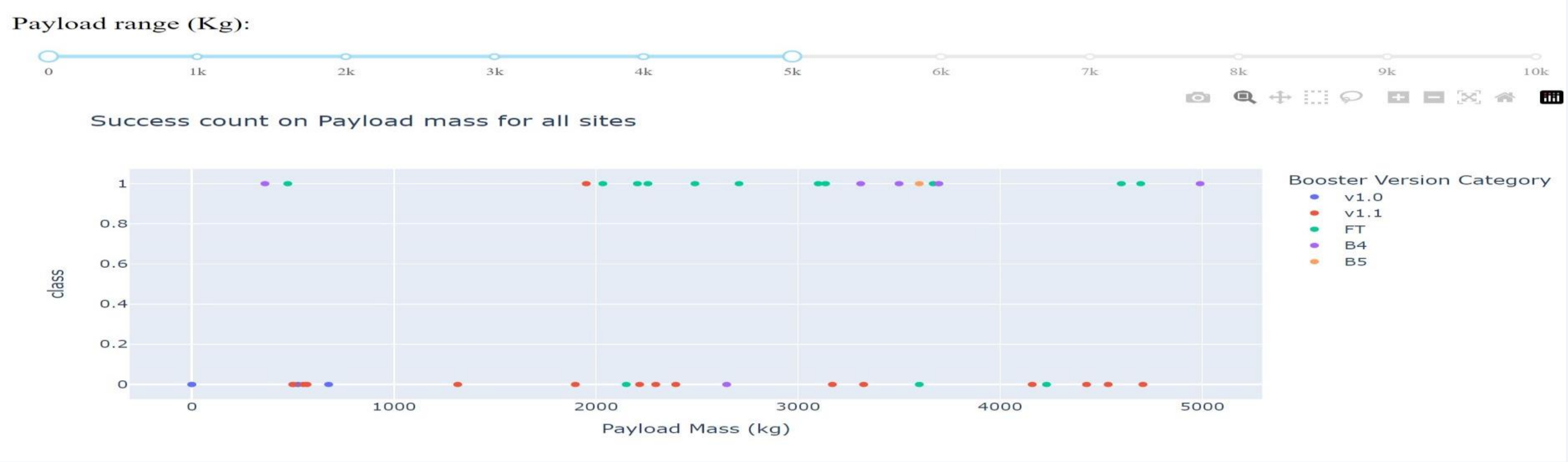
- The pie chart reveals that KSC LC-39A boasts the highest number of successful launches among all the launch sites, accounting for 41.7% of the total. It is closely followed by CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40, which represent 29.2%, 16.7%, and 12.5% of the successful launches, respectively.

<Dashboard Screenshot 2>



- The data depicted in the flowchart reveals that launches conducted from KSC LC-39A exhibit a success rate of 76.9% and a failure rate of 23.1%.

<Dashboard Screenshot 3>

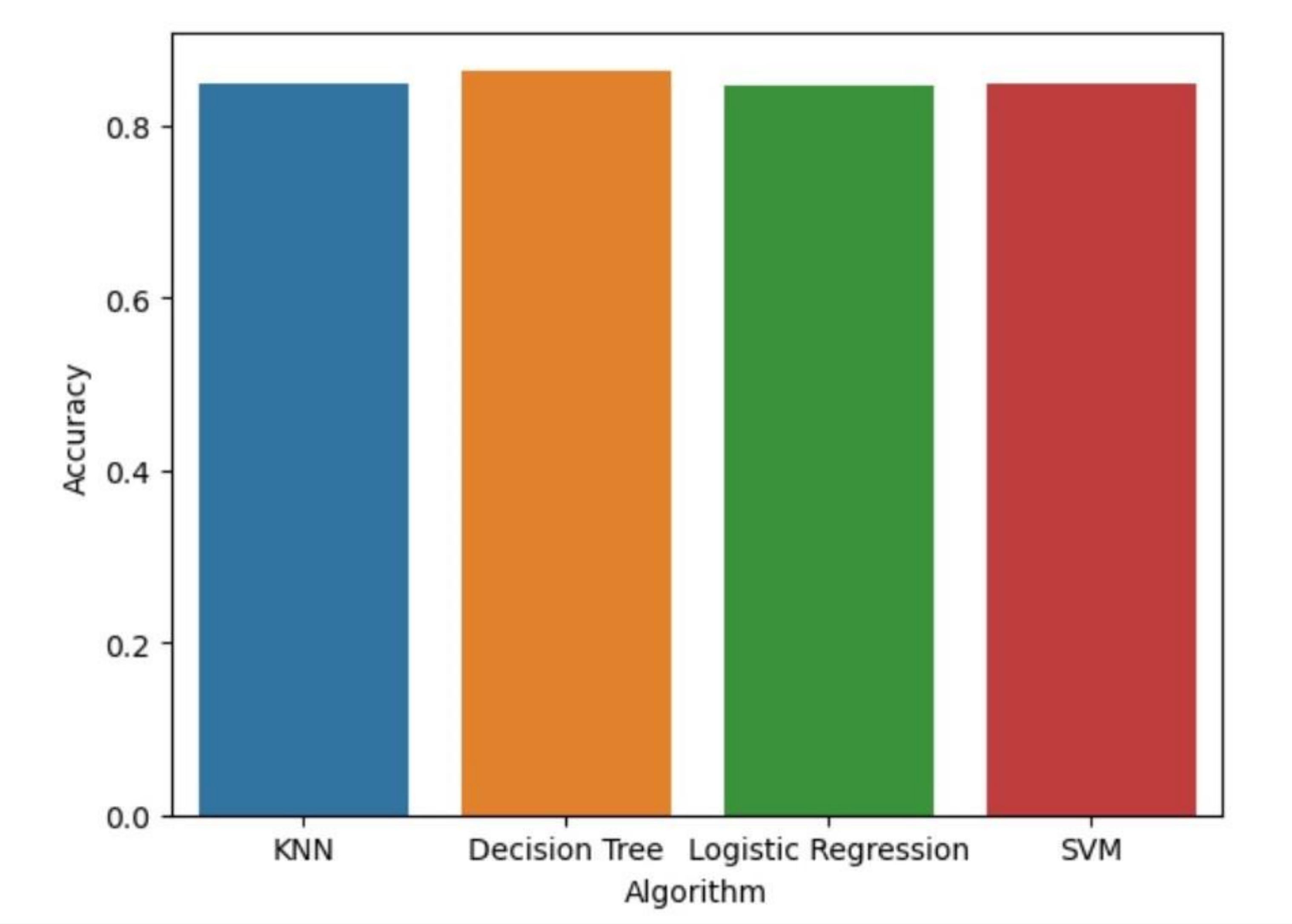


- The slider reveals that payloads in the lower weight range (0-5000 kg) exhibit a higher success rate compared to payloads in the heavier weight range (5000-10000 kg).

Section 5

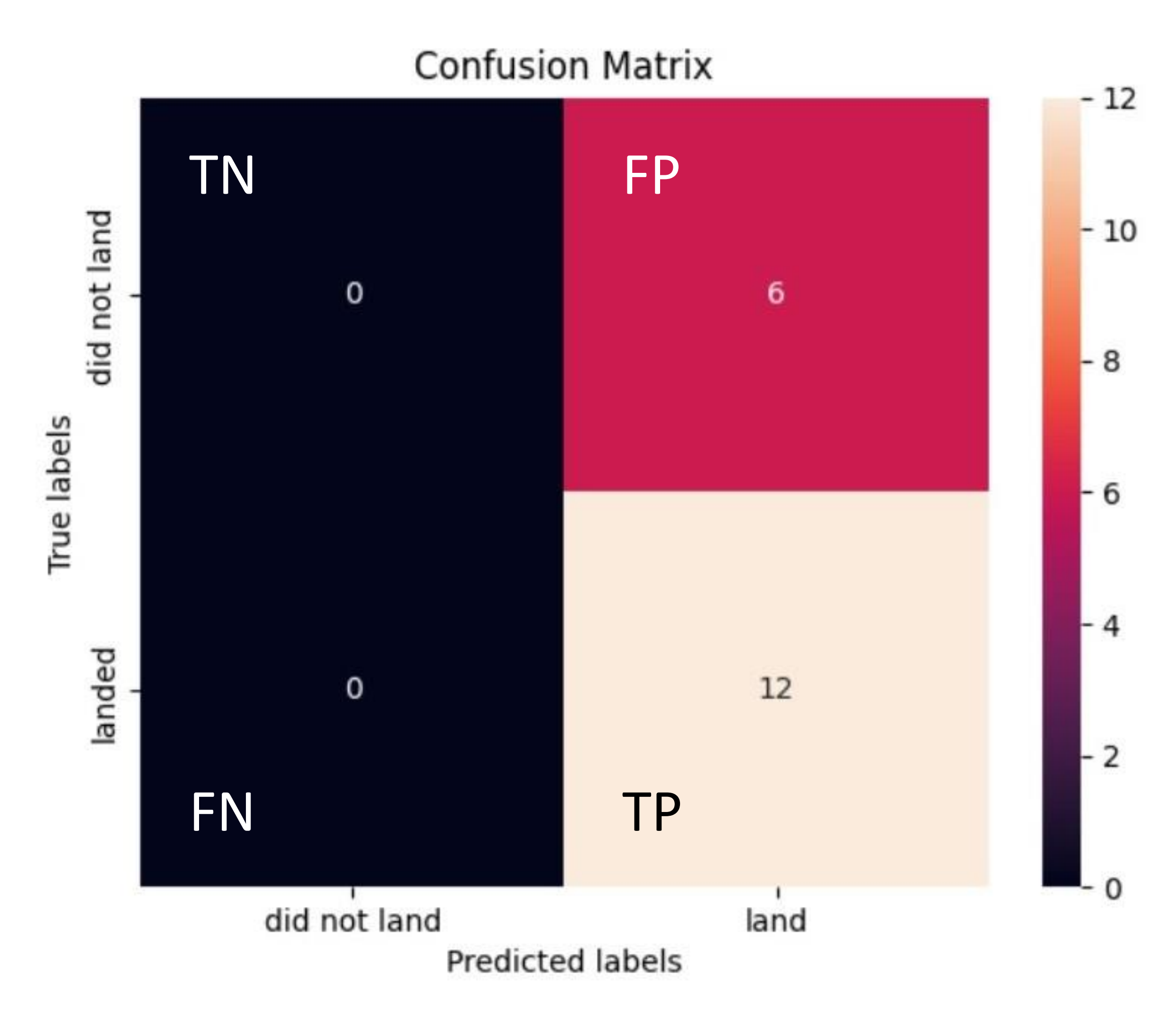
Predictive Analysis (Classification)

Classification Accuracy



- The decision tree algorithm outperforms other models with an accuracy score of approximately 0.87, making it the most accurate predictor for successful Falcon 9 rocket first stage landings.
- Its hierarchical structure effectively captures patterns and features, enabling precise predictions.

Confusion Matrix



- True Negatives (TN): These are the instances correctly classified as "did not land." The model accurately predicted that these instances would not result in a successful landing.
- False Positives (FP): These are the instances incorrectly classified as "landed" when they actually did not land. The model falsely predicted a successful landing for these instances.
- False Negatives (FN): These are the instances incorrectly classified as "did not land" when they actually landed. The model failed to predict a successful landing for these instances.
- True Positives (TP): These are the instances correctly classified as "landed." The model accurately predicted a successful landing for these instances.

Conclusions

- The quantity of launches performed at CCAFS SLC 40 is notably greater when compared to launches from other locations.
- The success rate of launches carrying a payload mass exceeding 7500 kg is notably high.
- The launch site CCAFS SLC 40 sees higher utilization compared to other launch sites for launching payloads with lower masses.
- Missions conducted in ES-L1, GEO, HEO, and SSO orbits exhibit a higher rate of success.
- Between 2012 and 2020, there was a consistent upward trend in the success rate, suggesting advancements in launch reliability and operational efficiency.
- The later years exhibit the highest success rate, indicating the growing development and reliability of SpaceX's Falcon 9 rockets.
- KSC LC-39A stands out with the highest number of successful launches among all the launch sites, representing 41.7% of the total.
- Payloads within the lower weight range (0-5000 kg) demonstrate a higher success rate in comparison to payloads within the heavier weight range (5000-10000 kg).
- With an accuracy score of approximately 0.87, the decision tree algorithm outperforms other models, firmly establishing itself as the preeminent predictor for successful Falcon 9 rocket first stage landings.

Appendix

- The GitHub Repository link for the whole analysis is as follow:

https://github.com/yashshah5489/Capstone-Project

Thank you!