IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- **Summary of methodologies**

Data Collection with Web Scraping

Data Wrangling

Exploratory Data Analysis (EDA)

Interactive Visualization

Predictive Analysis

- **Summary of all results**

Logistic Regression Accuracy: 83.33%

SVM Accuracy: 83.33%

Decision Tree Accuracy: 94.44%

KNN Accuracy: 83.33%

Introduction

The commercial space industry is entering a new era, with companies like SpaceX leading the way in making space travel more affordable through the reuse of rocket components. SpaceX's Falcon 9 rocket, which can recover and reuse its first stage, significantly reduces launch costs. This project focuses on predicting whether the Falcon 9's first stage will successfully land and be reused, leveraging public data and machine learning models to provide insights that can help optimize launch costs for SpaceX competitors like Space Y.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Used Python library (BeautifulSoup) to extract information, ensuring data accuracy and completeness.

Perform data wrangling:

Handled missing data, data manipulation with pandas, quering data using SQL, performed one hot encoding on categorical data for machine learning purpose.

Perform exploratory data analysis (EDA) using visualization and SQL

Conducted EDA using **SQL** for querying. Visualized data with **Matplotlib** to reveal key patterns, such as outliers, correlations, and trends.

Perform interactive visual analytics using Folium and Plotly Dash

Developed interactive dashboard using Plotly Dash and folium to present real-time data insights (e.g launch success rates, categorical distribution)

Perform predictive analysis using classification models

Applied machine learning models such as **Logistic Regression, SVM**, **Decision Trees**, and **KNN** to predict outcomes. Evaluated model based on accuracy and confusion matrices

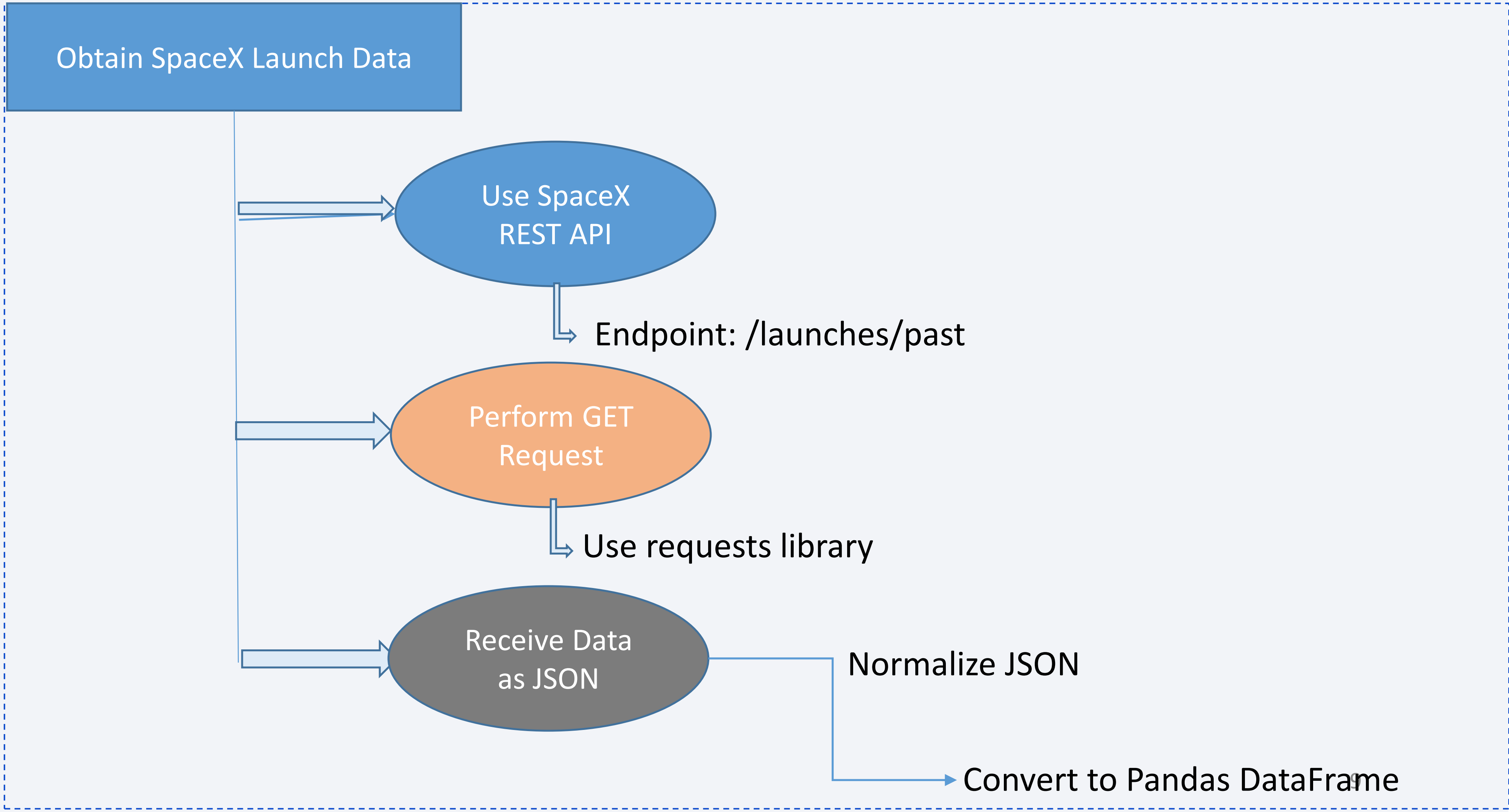
Data Collection

In this project, we collected SpaceX launch data through the **SpaceX REST API**, focusing on the /launches/past endpoint. The data includes rocket, payload, launch, and landing details, with the goal of predicting if SpaceX will attempt a rocket landing. We used Python's requests library to obtain JSON responses, which were then normalized into a flat table using json_normalize.

Additionally, we web scraped Falcon 9 launch data from Wikipedia using **BeautifulSoup** and transformed it into a pandas dataframe for further analysis. The data required wrangling, including filtering out Falcon 1 launches, handling null values (like replacing missing PayloadMass values with the mean), and targeting specific endpoints for additional rocket details. This clean, structured dataset was essential for predictive modeling.

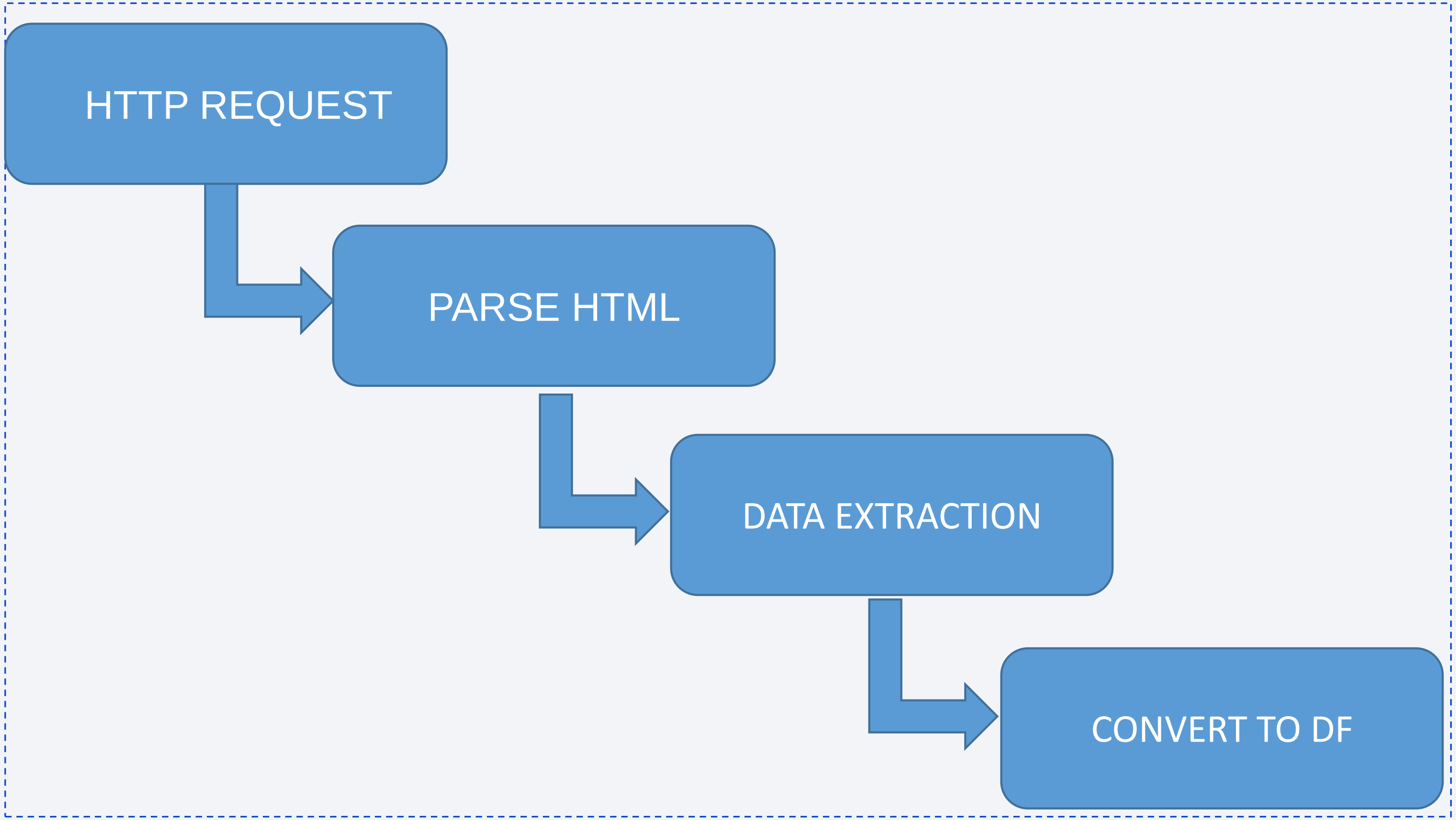
Data Collection – SpaceX API

- Use SpaceX REST API
- Perform GET Request
- Receive Data as JSON



Data Collection - Scraping

- Web Scraping
- Target Falcon 9 Launch Data
- Use BeautifulSoup to scrape HTML tables
- Parse Data
- Convert to DataFrame



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_webscraping.ipynb

Data Wrangling

- Data Wrangling
  - Filter Data
    - Remove falcon1 launches
  - Handle Null Values
    - Replace NULL in PayloadMass with Mean
    - Leave NULL in LandingPad (to handle later)
  - Gather Additional Data
    - Use API for details (Booster, Launchpad, Payload, Core)
- Clean and Structured Dataset for Analysis

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Scatter Plot: Flight Number vs. Payload Mass:** To visualize the relationship between these two variables and the outcome of each launch(success or failure)
- **Bar Plot:** It shows the distribution across different launch sites, and how successful each launch was based on the "Class" variable. It help understand if certain launch sites perform better overtime
- **Line Plot:** To identify trends and improvements in launch success rates across different years

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

EDA with SQL

- **Loading Data**: Loaded a CSV file into a pandas DataFrame (df) and inserted it into the SQLite database (SPACEXTBL table).
- **Table Creation**: Dropped and recreated the SPACEXTABLE with non-null date values.
- **Distinct Query**: Selected distinct values of LAUNCH_SITE from the SPACEXTBL table.
- **Filter and Limit Query**: Selected a subset of records where LAUNCH_SITE starts with 'CCA%' and limited the results to 5 rows.
- **Aggregation Queries**: Summed the PAYLOAD_MASS__KG_ for NASA's CRS missions. Averaged the PAYLOAD_MASS__KG_ for the booster version F9 v1.1.
- **Date-based Query**: Selected the minimum date where the landing outcome was 'Success (ground pad)'.
- **Range Query**: Selected payloads with a landing outcome of 'Success (drone ship)' where the payload mass was between 4000 and 6000 kg.
- **Grouping and Counting**: Counted the total number of mission outcomes by grouping records based on MISSION_OUTCOME.
- **Subquery**: Selected the BOOSTER_VERSION with the maximum PAYLOAD_MASS__KG_.
- **Failure Query**: Selected details of failed landings on drone ships in 2015.
- **Date Range Grouping**: Counted landing outcomes between specific dates, ordered by frequency.

Build an Interactive Map with Folium

- ❖ **Circle Marker**: Created using folium circle to highlight specific coordinates such as the NASA Johnson Space Center with a radius and color. A popup label was attached to the circle.
- ❖ **Text Marker**: A folium marker was added to represent the NASA Johnson Space Center, displaying a text label using the DivIcon class.
- ❖ **Multiple Circles and Markers**: Additional folium circle and folium marker objects were added for each launch site. These were designed to display launch site names with corresponding popups and icons.
- ❖ **Distance Markers**: Markers were created to display the distance from a launch site to the nearest city, railway, and highway. Text labels showing the calculated distances were added to the markers using DivIcon.
- ❖ **Polyline**: Lines (folium.PolyLine) were drawn between launch sites and nearby geographical features like coastlines, cities, and transport infrastructure to visualize proximity.

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Launch_site_location.ipynb

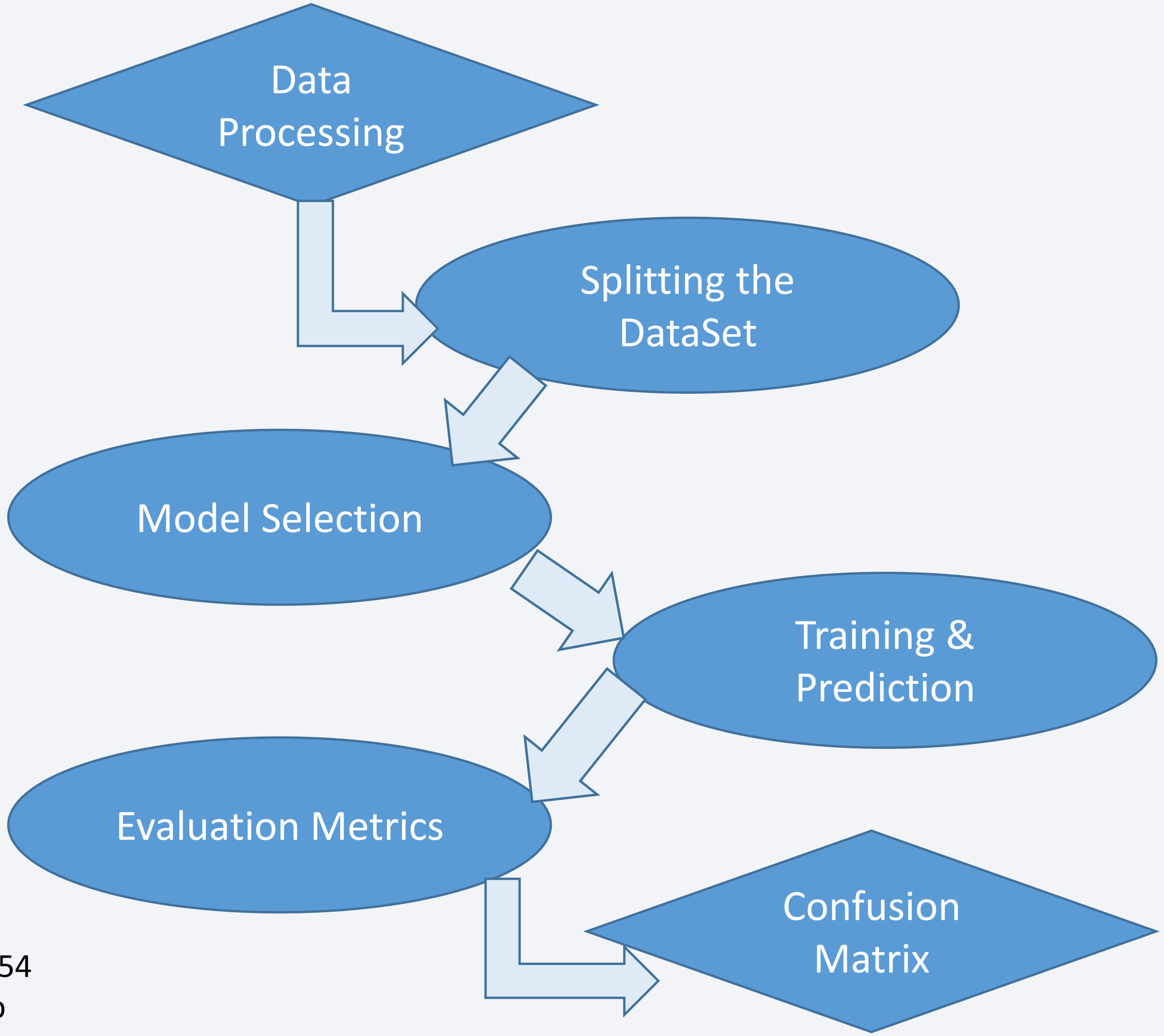
Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

Four machine learning models—**Logistic Regression**, **SVM**, **Decision Tree**, and **KNN**—is evaluated. Here are the results:

The **Decision Tree** model had the highest prediction accuracy, making it the best fit for predicting lunch success based on the given dataset. With an accuracy of **94.44%**, it significantly outperforms the other models, which all had accuracy levels around **83.33%**.



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Machine_Learning_Prediction_Analysis.ipynb

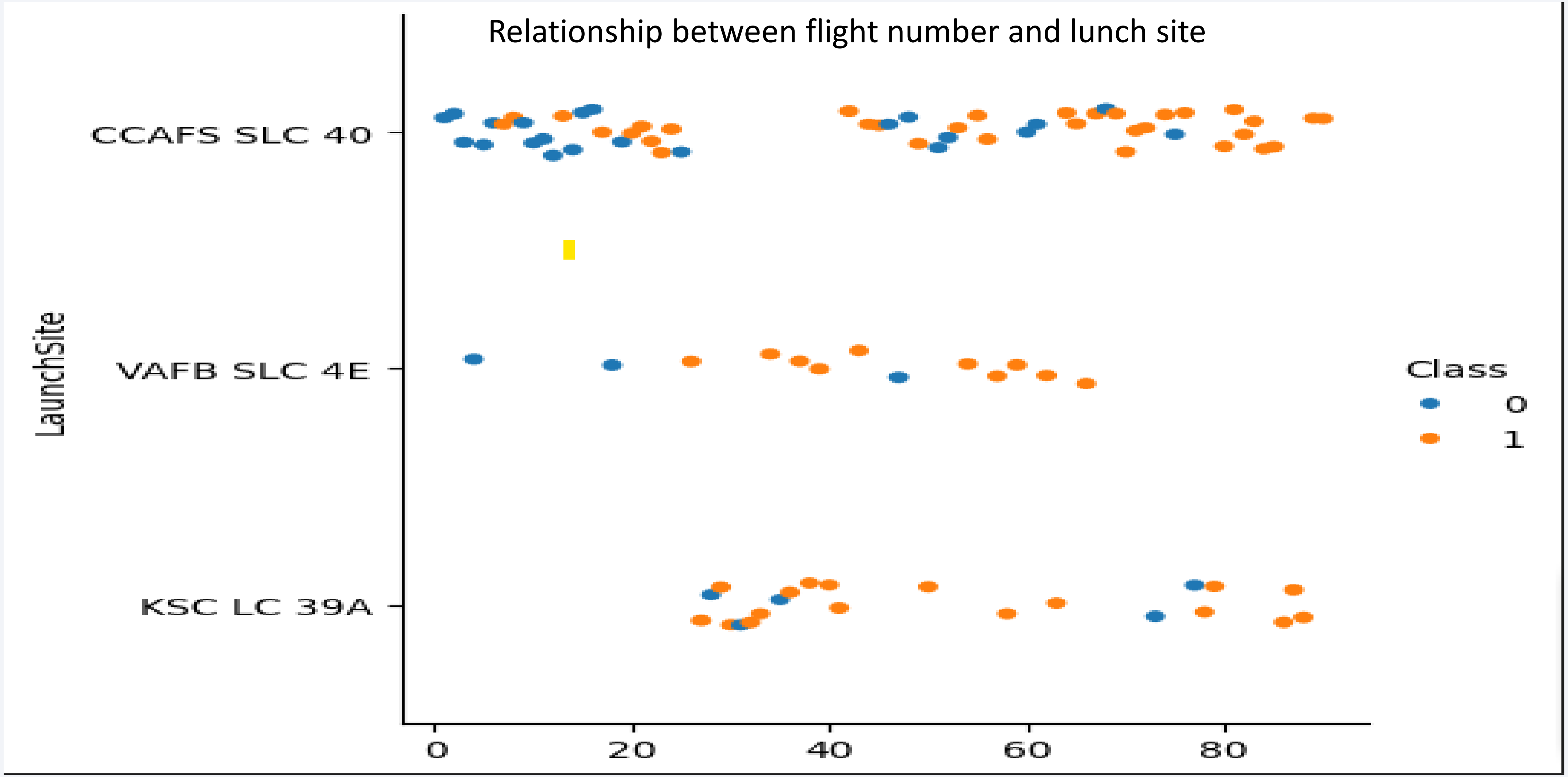
Results

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

Section 2

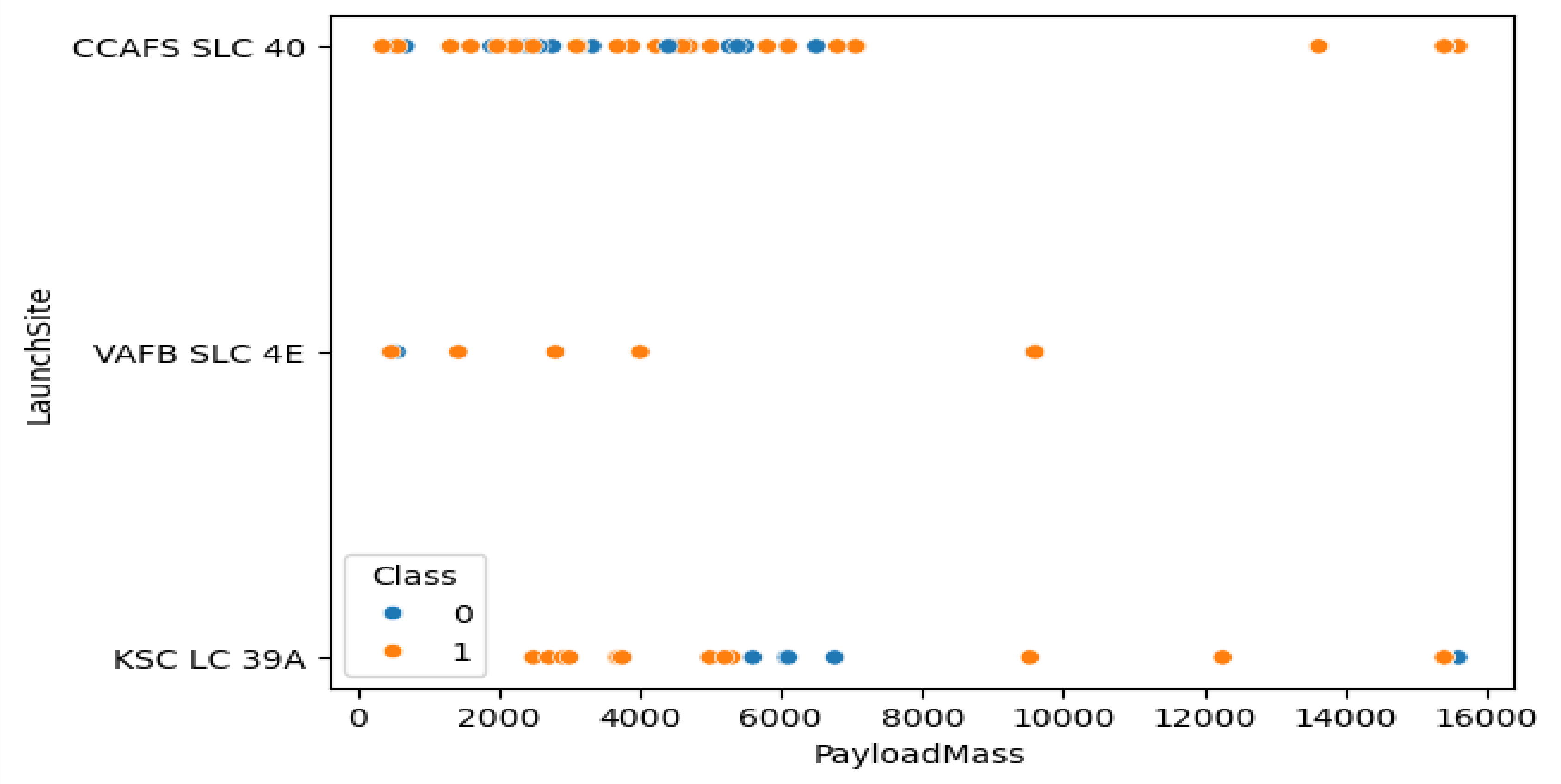
Insights drawn from EDA

Flight Number vs. Launch Site



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

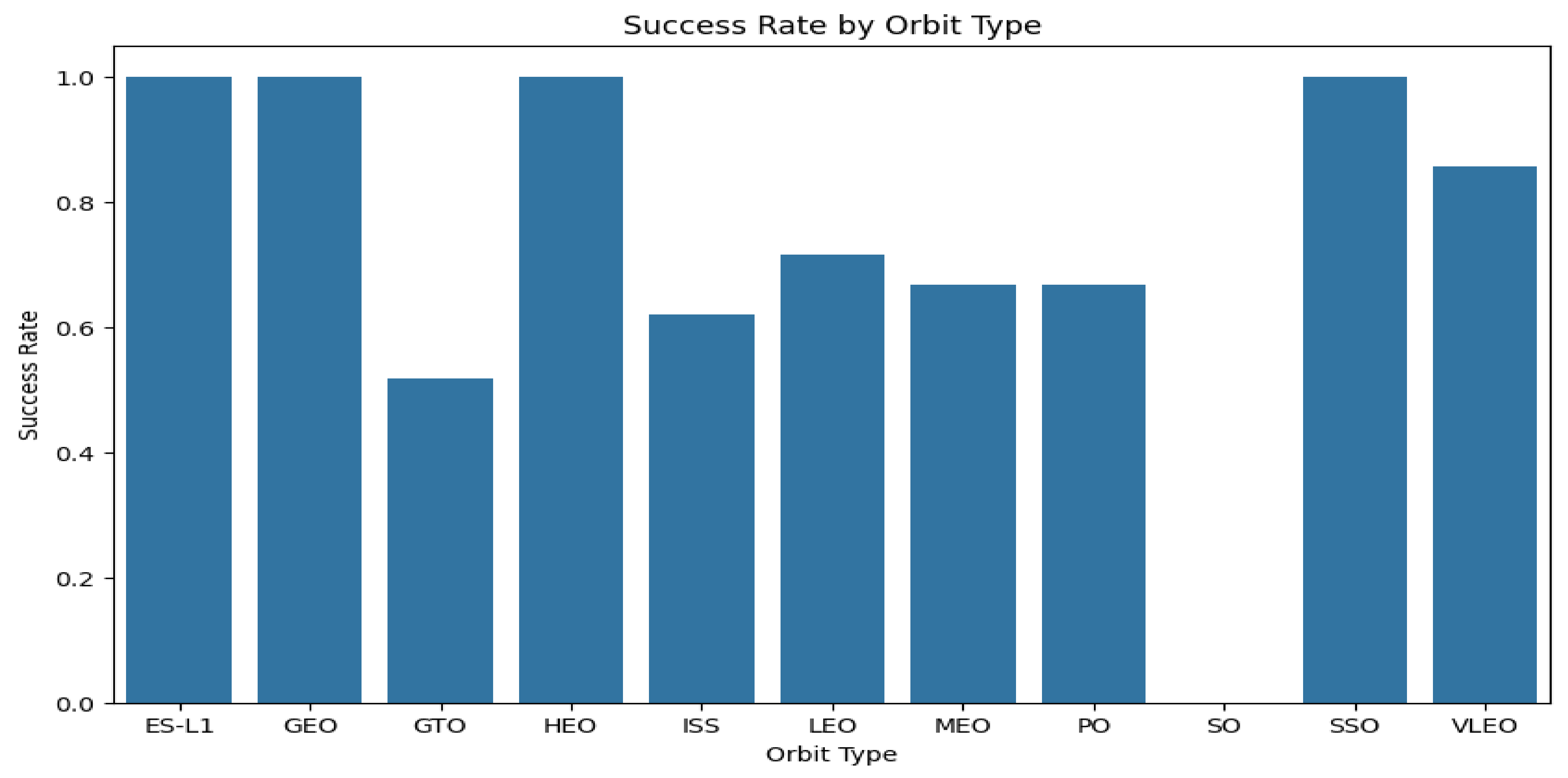
Payload vs. Launch Site



Payload Mass Vs. Launch Site scatter point chart shows that for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

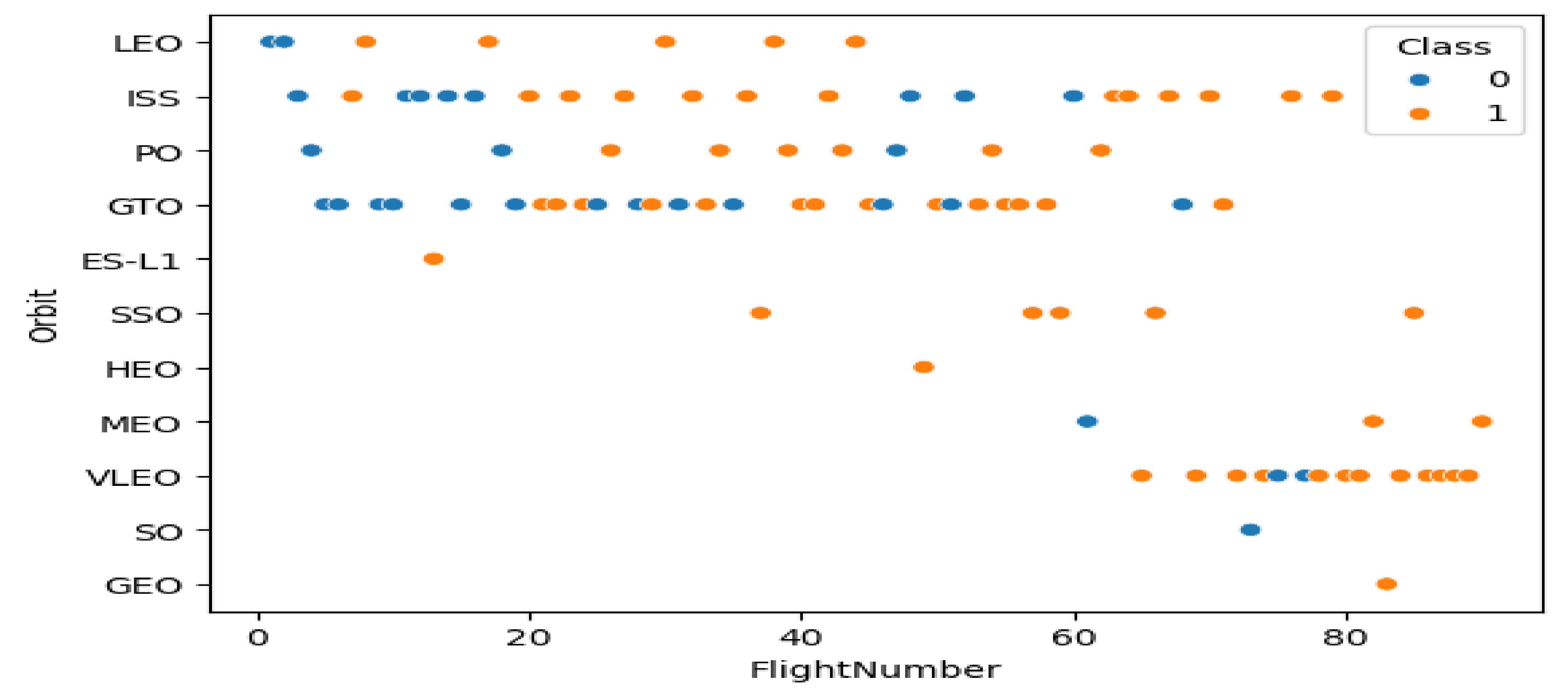
https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

Success Rate vs. Orbit Type



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

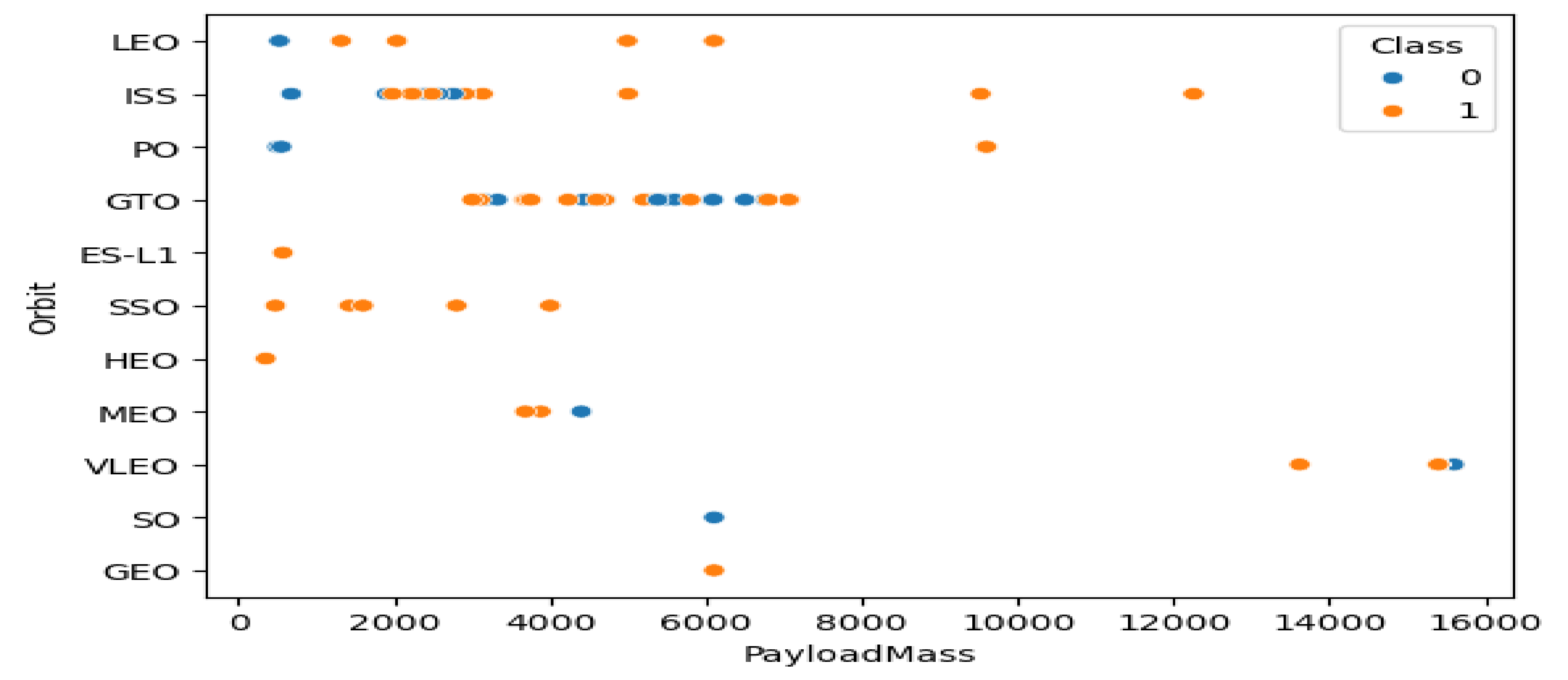
Flight Number vs. Orbit Type



You can observe that in the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

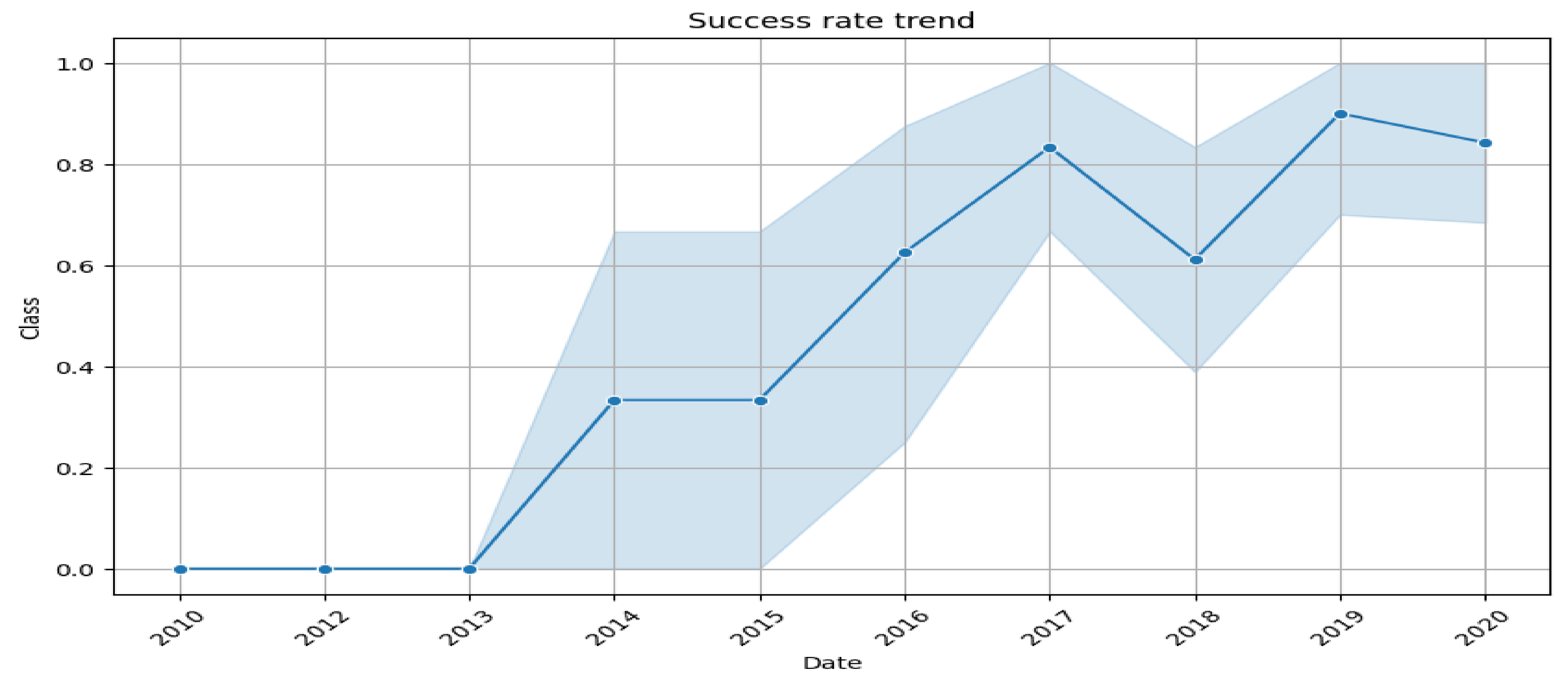
Payload vs. Orbit Type



With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/EDA_data_visualization.ipynb

All Launch Site Names

unique launch sites in the space mission

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

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

Total Payload Mass

sum(PAYLOAD_MASS__KG_)
45596

Total payload mass carried by boosters launched by NASA (CRS)

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

Average Payload Mass by F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

Average payload mass carried by booster version F9 v1.1

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

First Successful Ground Landing Date

min(DATE)
2015-12-22

Date when the first successful landing outcome in ground pad was achieved.

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

Successful Drone Ship Landing with Payload between 4000 and 6000

Payload	PAYLOAD_MASS__KG_
JCSAT-14	4696
JCSAT-16	4600
SES-10	5300
SES-11 / EchoStar 105	5200

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

Total Number of Successful and Failure Mission Outcomes

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

Total number of successful and failure mission outcomes

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

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

Names of the booster_versions which have carried the maximum payload mass

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

2015 Launch Records

Done .

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

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

Landing_Outcome	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

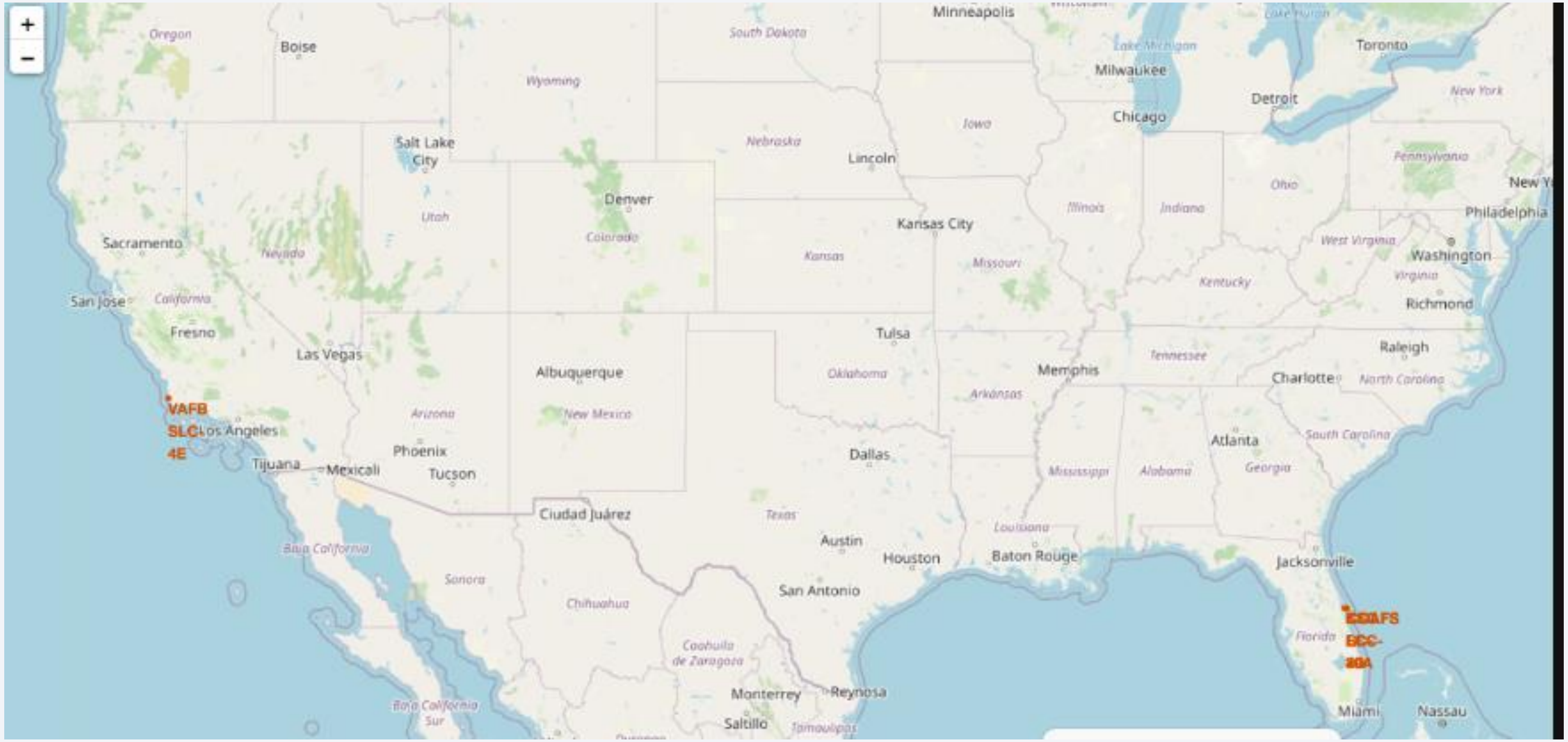
Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20

https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Spacex_EDA_SQL.ipynb

Section 3

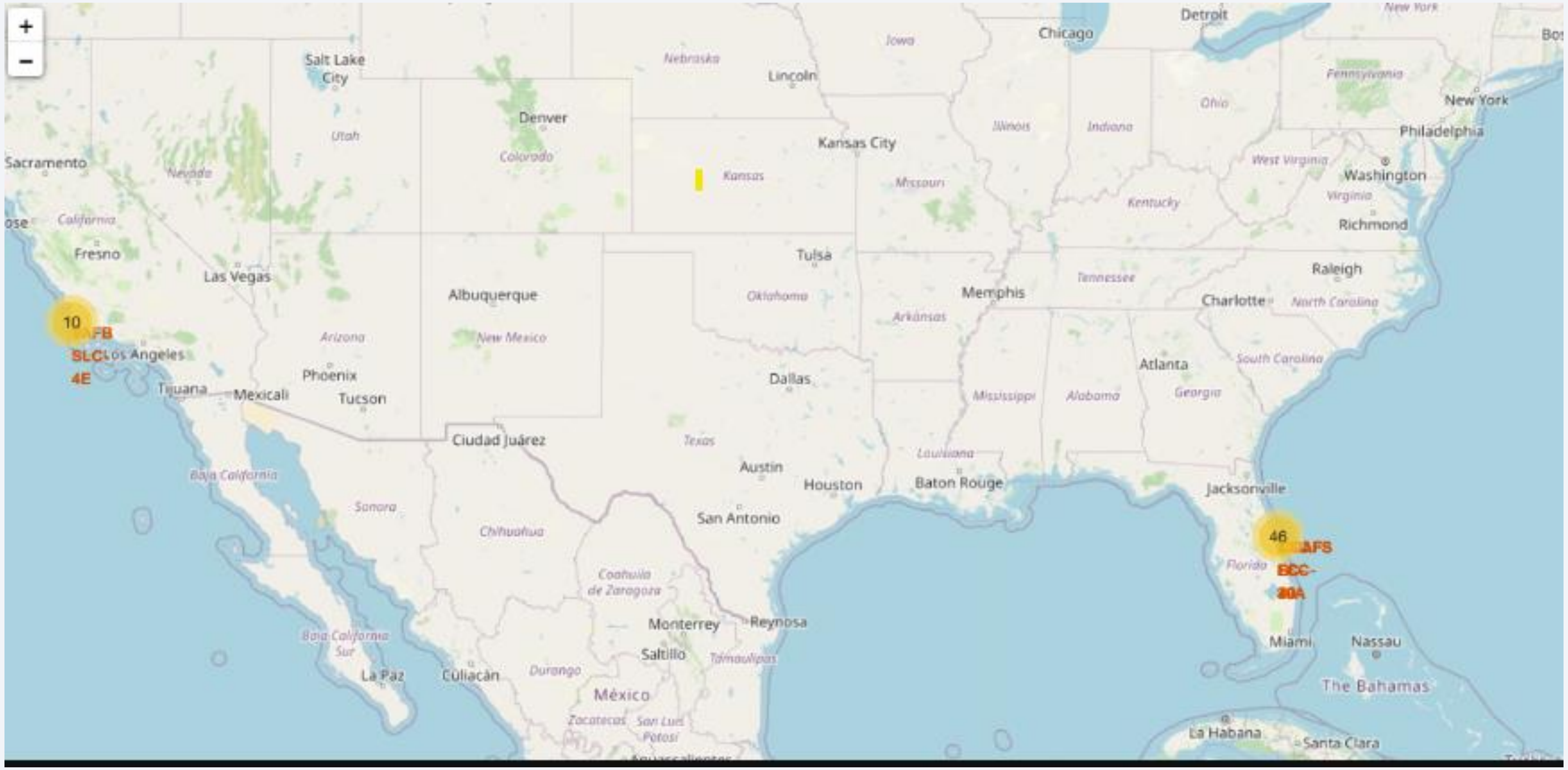
Launch Sites Proximities Analysis

Location of SpaceX on world map



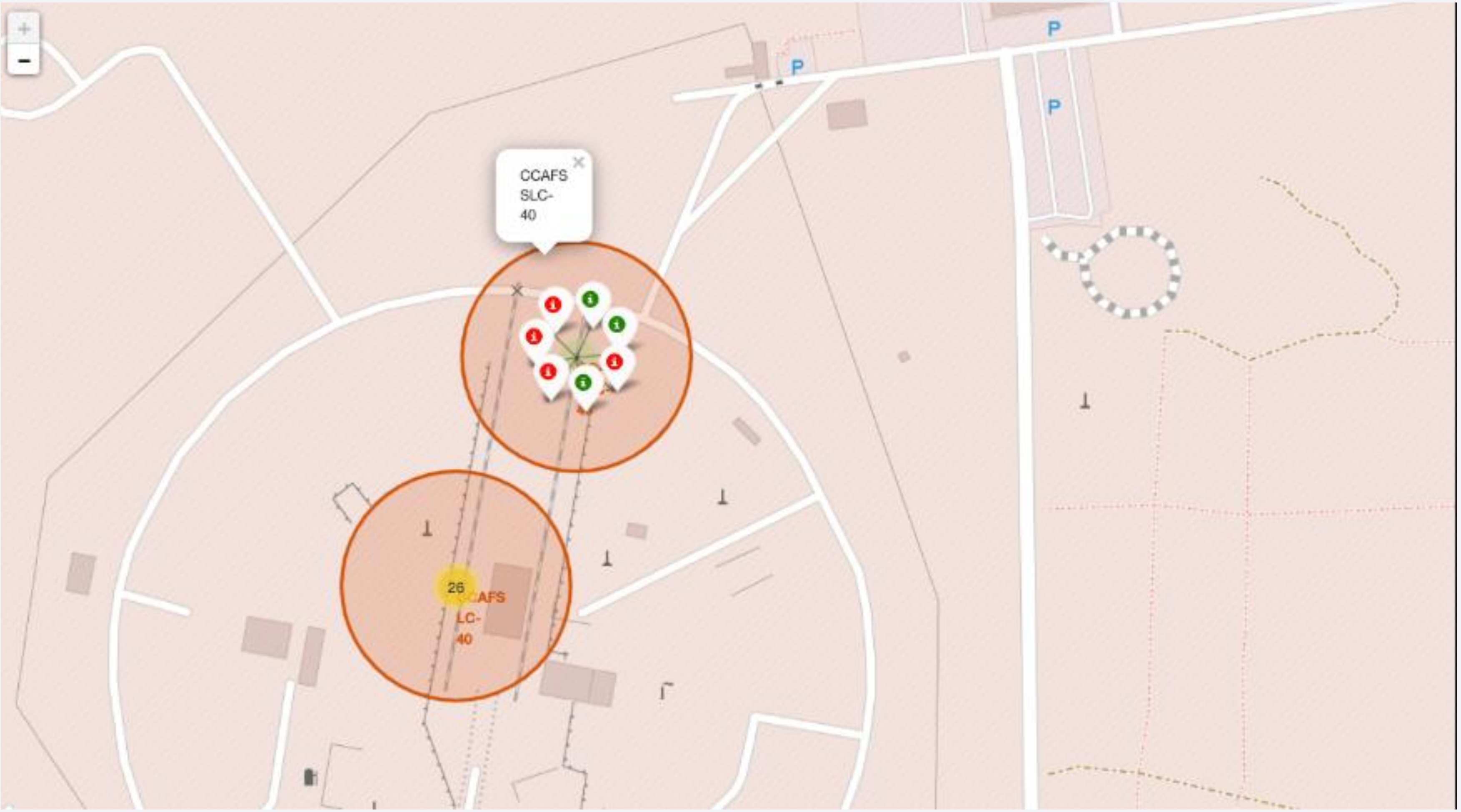
https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Launch_site_location.ipynb

Proximity of SpaceX to the Coast



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Launch_site_location.ipynb

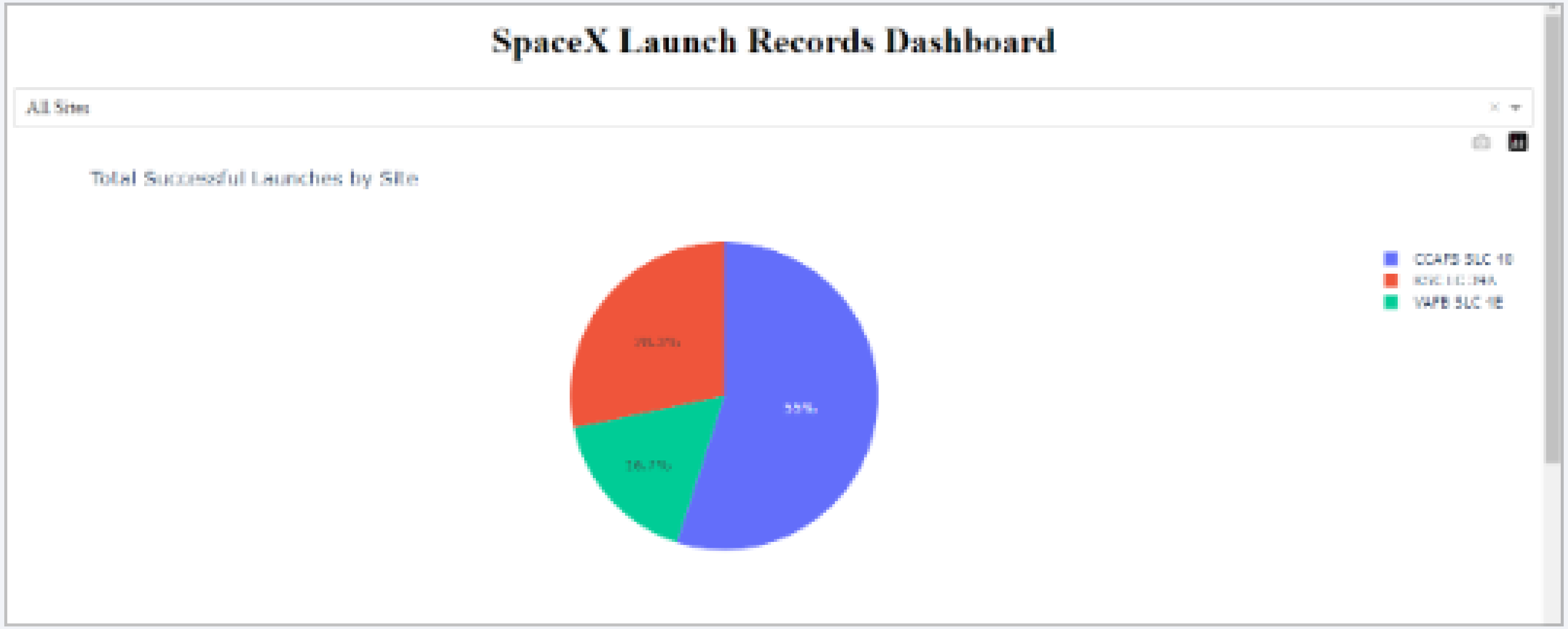
Proximity of SpaceX to Railway



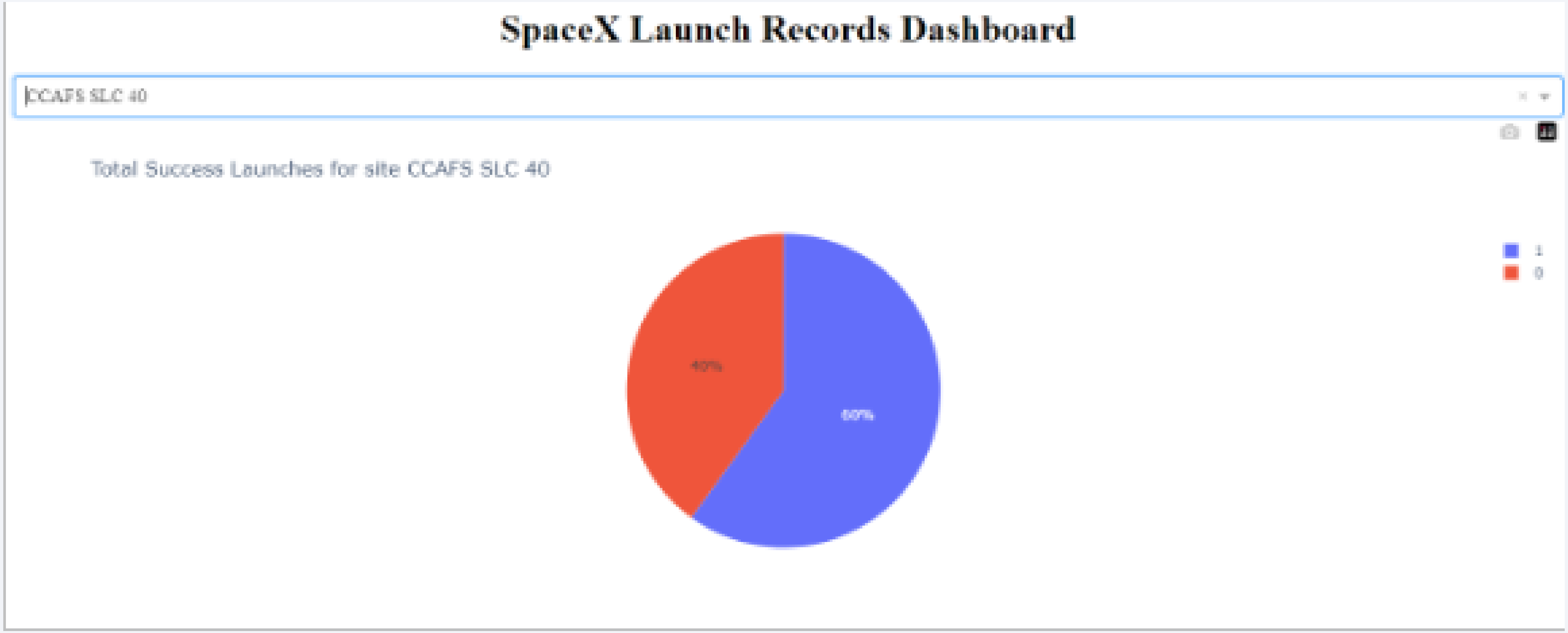
Section 4

Build a Dashboard with Plotly Dash

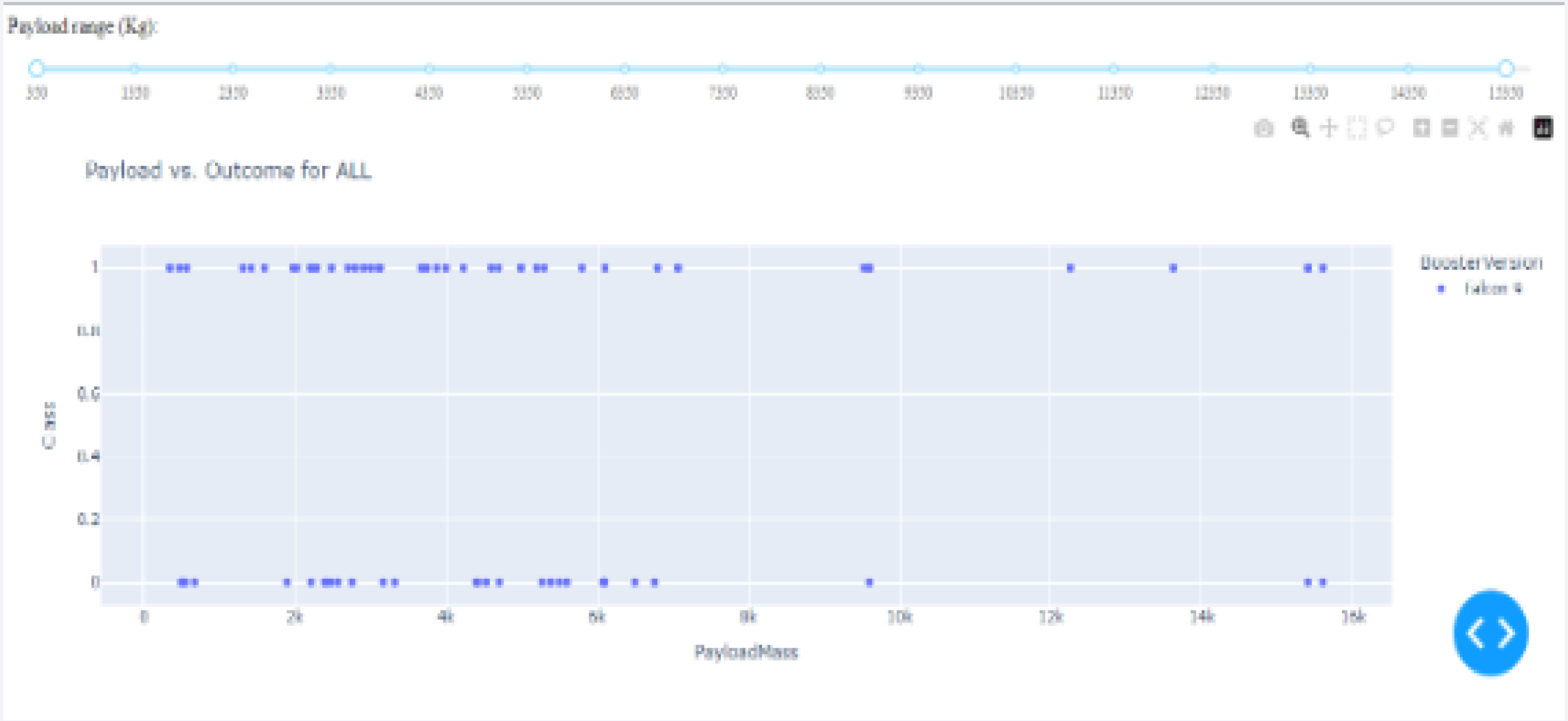
<Dashboard Screenshot 1>



Total success launches for site CCAFS SLC 40



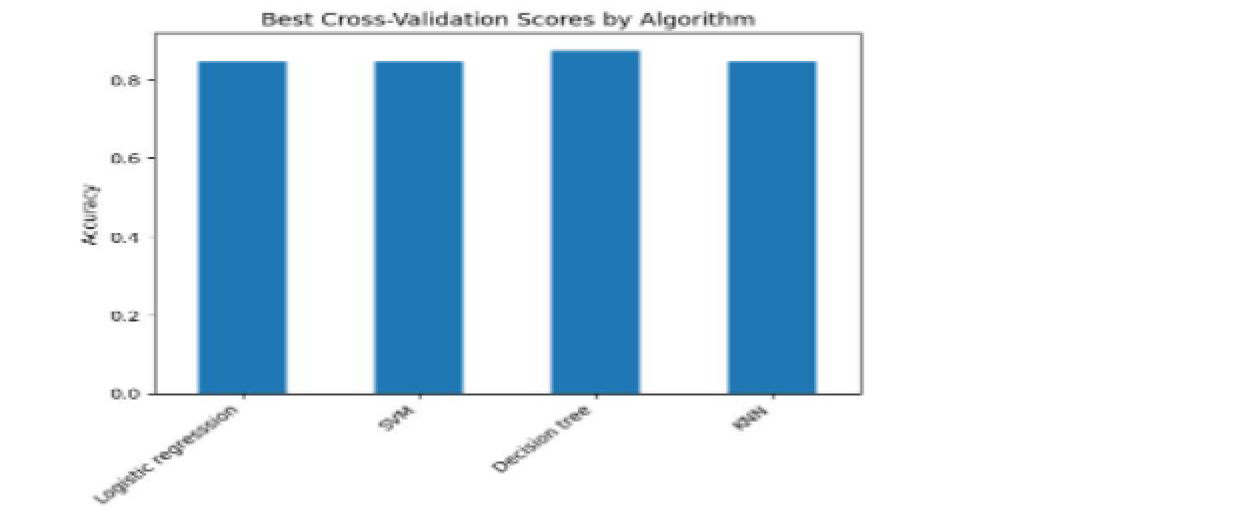
PayloadMass VS Outcome



Section 5

Predictive Analysis (Classification)

Classification Accuracy

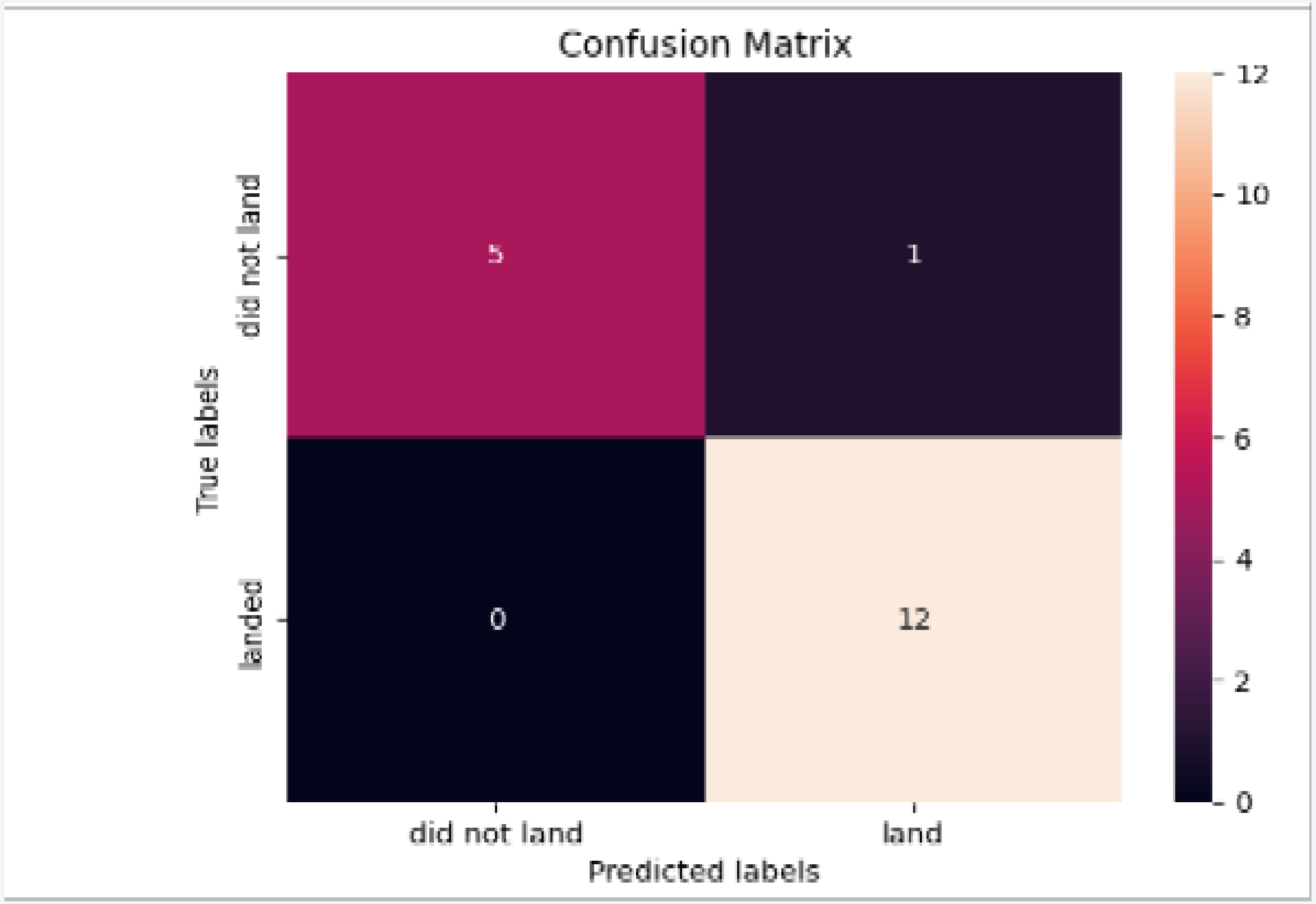


Accuracy was used as the primary metric for evaluating the models.

Logistic Regression, SVM, and KNN: 83.33% accuracy.

Decision Tree: 94.44% accuracy (Best performing model).

Confusion Matrix of Decision Tree Classifier



https://github.com/nanaaisha94/Capstone_project/blob/dc267bec952fa854768113eb91e40f975172ff0c/Machine_Learning_Prediction_Analysis.ipynb

Conclusions

- **High success rate** in launches, showcasing reliability.
- **Consistent performance** across various payload sizes, with no significant impact on launch success.
- **Cost efficiency**, with most launches falling within a specific cost range, giving SpaceX a competitive edge.
- **Strategic launch sites**, optimizing logistics and reducing costs.
- **Advancements in rocket reusability**, with increasing recovery success over time.
- **Growth in launch frequency**, reflecting increasing operational capacity and demand.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!