IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohsen Goodarzi

January 2024

Outline

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

Introduction

- Project background and context:
  - Commercial Space Boom:
    - Companies like SpaceX, Virgin Galactic, Rocket Lab, and Blue Origin making space more accessible.
  - SpaceX's Success Story:
    - Manned missions, Starlink satellite constellation, and cost-efficiency due to reusable rockets, especially the Falcon 9.
  - Falcon 9 and Its Significance:
    - Falcon 9 rocket's importance in SpaceX's success, particularly its first stage, which is crucial for cost savings.
  - Visualizing Falcon 9:
    - Forest Katsch's diagrams for a better understanding of Falcon 9's scale, aiding in cost estimation.

Introduction

- Problems you want to find answers
  - Launch Cost Determination:
    - Core problem: understanding the factors influencing the cost of a SpaceX launch, crucial for a new player like Space Y aiming to compete.
  - First Stage Reusability Prediction using Machine Learning :
    - Objective of predicting if SpaceX will reuse the first stage, shifting from traditional rocket science to a machine learning approach using public information.
  - Space Y's Competitive Edge:
    - Empowering Space Y to competitively price launches against SpaceX through precise cost estimation and strategic planning.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX REST API:
    - Use SpaceX REST API for launch data.
    - Gather rocket, payload, launch, and landing information.
    - Aim to predict SpaceX's rocket landing attempts.
  - Web Scraping with BeautifulSoup:
    - Supplement data from Wiki pages using BeautifulSoup.
    - Store scraped data in lists for dataset creation.
    - Merge API and web-scraped data for a comprehensive dataset.

Methodology

- Perform data wrangling
  - API-based Data Enhancement:
    - Use SpaceX API to gather specific data, replacing identification numbers with actual data (Booster, Launchpad, Payload, Core).
  - Filtering Falcon 1 Launches:
    - Exclude Falcon 1 launches from the dataset to focus exclusively on Falcon 9.
  - Handling NULL Values – Payload Mass:
    - Calculate the mean of Payload Mass data to replace NULL values, ensuring dataset integrity.
  - NULL Values in Landing Pad:
    - Leave NULL values in the Landing Pad column intact, representing scenarios without landing pad usage.

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL
  - Exploratory Data Analysis (EDA):
    - Utilize EDA as the initial step in the data science project.
    - Employ visualization techniques to observe trends, such as the improving success rate since 2013.
  - Launch Site Success Rates:
    - Identify variations in success rates across different launch sites.
    - Leverage SQL to analyze and extract insights from the dataset, preparing for machine learning.
  - Attribute Correlation:
    - Determine attributes correlated with successful landings through EDA.
    - Utilize one-hot encoding in SQL for categorical variables, optimizing the dataset for machine learning predictions.

Methodology

- Perform interactive visual analytics using Folium and Plotly Dash:
- Exploring Data in Real-Time:
  - Interactive visual analytics allows users to explore and manipulate data dynamically.
  - Common interactions like zooming, panning, filtering, and searching enhance the exploration process.
- Effective Pattern Discovery:
  - Users can find visual patterns more efficiently with interactive visual analytics.
  - Dashboards and interactive data visualizations provide a compelling and dynamic way to present findings.
- Using Folium for Launch Site Analysis:
  - Folium is employed to create an interactive map displaying launch site locations and their proximities.
  - Exploration of the map helps discover patterns and aids in choosing optimal launch sites based on geo and proximity data.
- Plotly Dash Dashboard Application:
  - Utilizing Plotly Dash to build a dashboard with input components like dropdown lists and range sliders.
  - The dashboard facilitates interaction with pie charts and scatter point charts, offering a user-friendly way to gain insights from the SpaceX dataset.

Methodology

- Perform predictive analysis using classification models

- Building a Machine Learning Pipeline:
  - Construct a pipeline to predict the successful landing of the Falcon 9 first stage.
  - Include preprocessing steps to standardize data and train-test split to divide data for training and testing.
- Building, Tuning, and Evaluating Classification Models:
  - Train various classification models, including Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors.
  - Utilize Grid Search to find optimal hyperparameters for each model, enhancing their performance.
- Evaluating Model Performance:
  - Determine the model with the highest accuracy using training data.
  - Output confusion matrix to assess model performance in predicting successful first stage landings.

Data Collection

Requesting and Parsing SpaceX Launch Data

- Utilized GET request to retrieve SpaceX launch data.
- Parsed the data to extract relevant information using Python libraries.

Filtering Data to Include Only Falcon 9 Launches

- Filtered the dataframe to include only Falcon 9 launches.
- Ensured focus on specific rocket type for analysis.

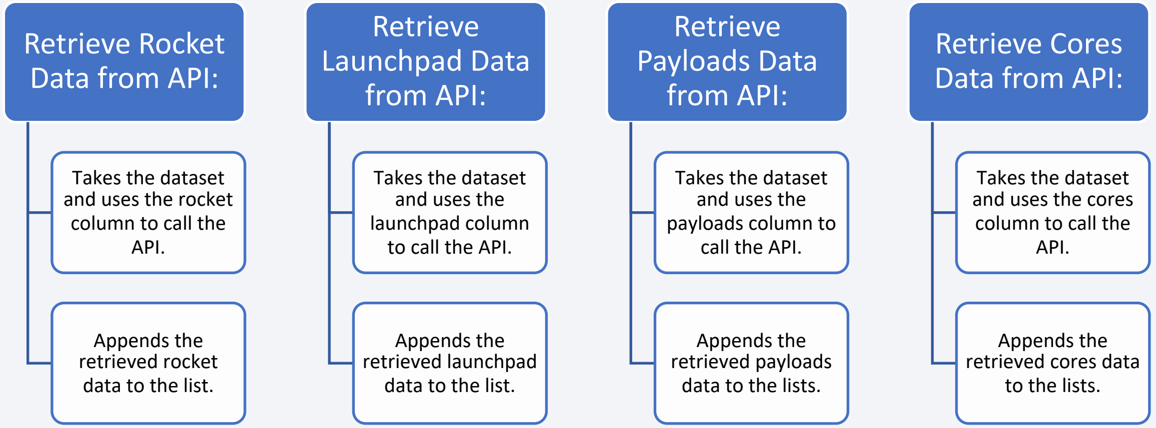
Data Wrangling: Handling Missing Values

- Calculated mean for PayloadMass using .mean() function.
- Replaced missing values (np.nan) in the dataset with the calculated mean using .replace() function.

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/jupyter-labs-webscraping.ipynb

Request Falcon 9 Launch Wiki Page:

- Initiates a request to the Falcon 9 Launch Wiki page from its URL.

Extract Column/Variable Names from Table Header:

- Extracts all column/variable names from the HTML table header of the Falcon 9 Launch Wiki page.

Create DataFrame by Parsing Launch HTML Tables:

- Parses the HTML tables on the Falcon 9 Launch Wiki page to create a data frame.
- Utilizes the extracted column/variable names for DataFrame creation.

Data Wrangling

- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/labs-jupyter-spacex-Data wrangling.ipynb

Calculate Number of Launches on Each Site

- Counts the number of launches on each launch site.
- Utilizes the launch site column in the dataset to calculate the frequency of launches.

Calculate Number and Occurrence of Each Orbit:

- Determines the number and occurrence of each orbit.
- Analyzes the orbit column in the dataset to calculate the frequency of different orbits.

Calculate Number and Occurrence of Mission Outcomes by Orbit:

- Determines the number and occurrence of mission outcomes for each orbit.
- Utilizes the orbit and outcome columns in the dataset to calculate the frequency of mission outcomes for different orbits.

Create Landing Outcome Label from Outcome Column:

- Creates a landing outcome label from the outcome column.
- Assigns a specific label based on the outcome column values, indicating whether the landing was successful, unsuccessful, or other.

EDA with Data Visualization

- Visualization:
  - Scatter plots:
    - Flight Number vs. Payload
    - Flight Number vs. Launch Site
    - Payload Mass (kg) vs. Launch Site
    - Payload Mass (kg) vs. Orbit type
  - Other plots:
    - Success rate vs Orbit
    - Success rate vs Year

- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- Queries:
  - Display:
    - Unique launch site names
    - Five instances where the launch site begins with 'CCA'
    - Cumulative payload mass transported by boosters launched by NASA under CRS missions
    - Mean payload mass transported by booster version F9 v1.1.

EDA with SQL

- Queries:
  - List:
    - Date of initial successful ground pad landing
    - Boosters names that achieved successful landings on a drone ship with payload masses between 4,000 and 6,000
    - Overall count of successful and unsuccessful missions
    - Booster versions that transported the maximum payload
    - Unsuccessful drone ship landings, including their booster version and launch site, specifically within the year 2015
    - Tally of landing outcomes from June 4, 2010, to March 20, 2017, in descending order.

- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_1.ipynb

Build an Interactive Map with Folium

- To explore the correlation between the success rate and various properties of launch sites, such as geographical location and proximity to transportation infrastructure and urban areas, the following map visualizations have been developed:
  - Launch Site Markers:
    - Blue circle marker placed at NASA Johnson Space Center's coordinates with a popup label displaying its name.
    - Red circle markers added at all launch site coordinates with popup labels showing their names.
  - Launch Outcome Markers:
    - Markers colored green for successful launches and red for unsuccessful launches, depicting launch success rates at each site.
  - Launch Site Proximity Distances:
    - Colored lines representing distances between launch site CCAFS SLC-40 and the nearest coastline, railway, highway, and city.
- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite1.ipynb

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites:
  - This feature provides users with a list in the form of a dropdown menu, allowing them to select either all available launch sites or a specific launch site of their choice. This functionality enhances user interaction and customization, enabling them to focus on specific launch sites as needed.

- Dashboard with Plotly Dash:
  - The dashboard interface is built using Plotly Dash, a powerful visualization library. It incorporates various interactive elements such as sliders, charts, and dropdown menus to present data in an intuitive and dynamic manner. This dashboard offers users a comprehensive view of the data and facilitates easy exploration and analysis.

- Slider of Payload Mass Range:
  - A slider component is included in the dashboard interface, enabling users to select a specific range of payload masses. This interactive feature allows users to filter and analyze data based on their payload mass criteria, enhancing the granularity of analysis and providing insights into the relationship between payload mass and other variables.

Build a Dashboard with Plotly Dash

- Pie Chart Showing Successful Launches:
  - This visualization component presents data on successful and unsuccessful launches in the form of a pie chart. By representing the data as proportions of a whole, users can quickly grasp the distribution of successful launches relative to total launches. This visual representation aids in understanding the success rate and identifying trends or patterns in launch outcomes.

- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version:
  - The scatter chart visualizes the relationship between payload mass and launch success rate across different booster versions. By plotting payload mass on one axis and success rate on the other, users can examine how changes in payload mass impact the likelihood of launch success for different booster versions. This visualization enables users to identify correlations and make informed decisions based on payload considerations.

- Explain why you added those plots and interactions

- https://github.com/Mohsen-DS/IBM_Capstone/commit/e5c1b7633434c132bfcd2a702b7d761f8d806745

Predictive Analysis (Classification)

- Data Preparation and Standardization:
  - Convert the 'Class' column into a NumPy array and standardize the data using StandardScaler to ensure uniform scaling across features.
- Model Training and Evaluation:
  - Split the standardized data into training and testing sets using train_test_split for model evaluation.
  - Implement GridSearchCV with cv=10 to optimize model parameters, enhancing model performance.
- Model Selection and Performance Assessment:
  - Apply GridSearchCV on various classification algorithms, including logistic regression, support vector machine, decision tree, and K-nearest neighbor.
  - Calculate accuracy scores on the test data for all models and assess confusion matrices to evaluate their predictive capabilities.
- Best Model Identification:
  - Determine the best-performing model based on evaluation metrics such as Jaccard Score, F1 Score, and overall accuracy, providing insights into the most effective classification algorithm for the given dataset.
- https://github.com/Mohsen-DS/IBM_Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

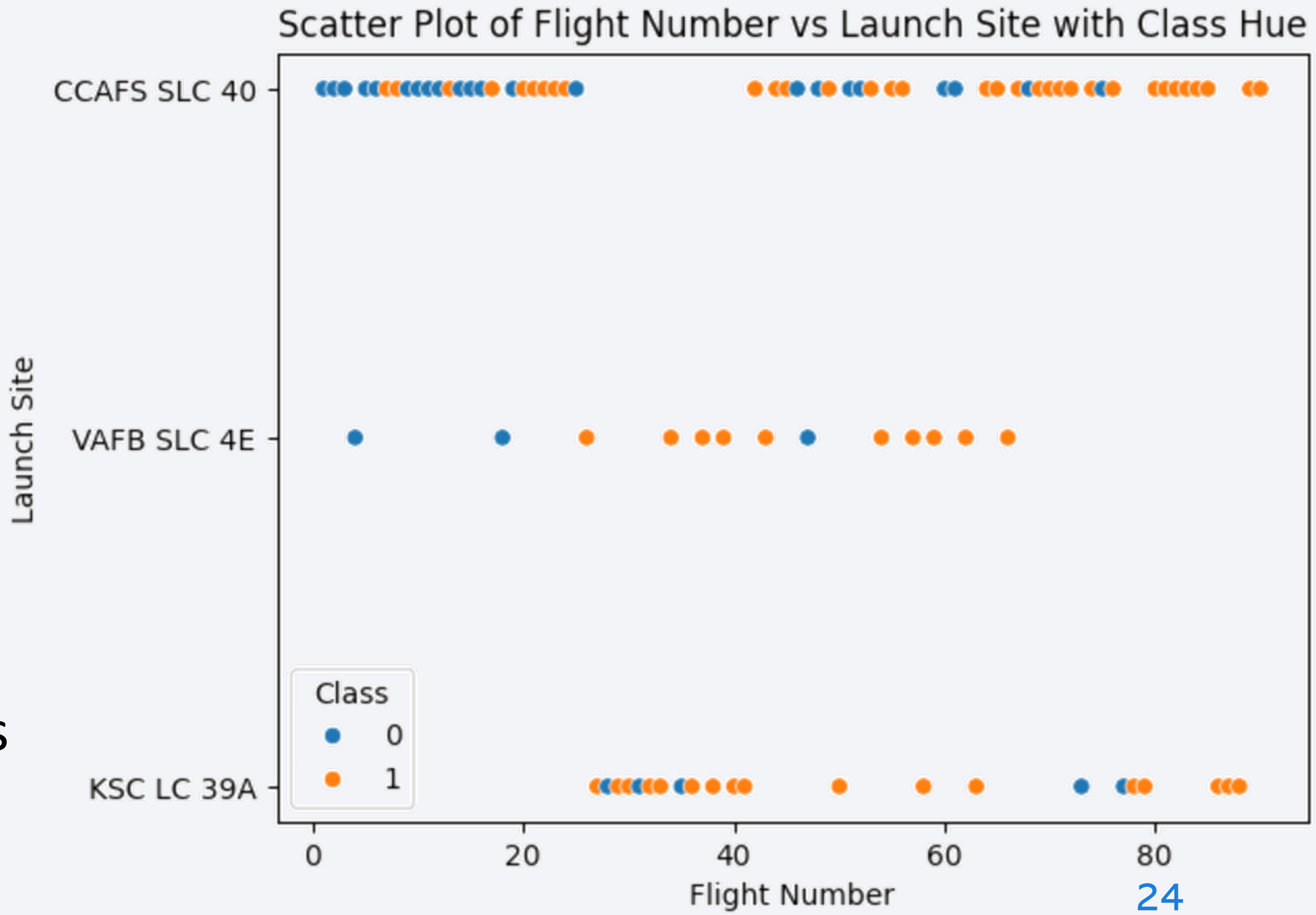
- Exploratory data analysis results
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- Interactive analytics demo in screenshots
  - Most launch sites are near the equator, and all are close to the coast
  - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities
- Predictive analysis results
  - Decision Tree model is the best predictive model for the dataset

Section 2

Insights drawn from EDA

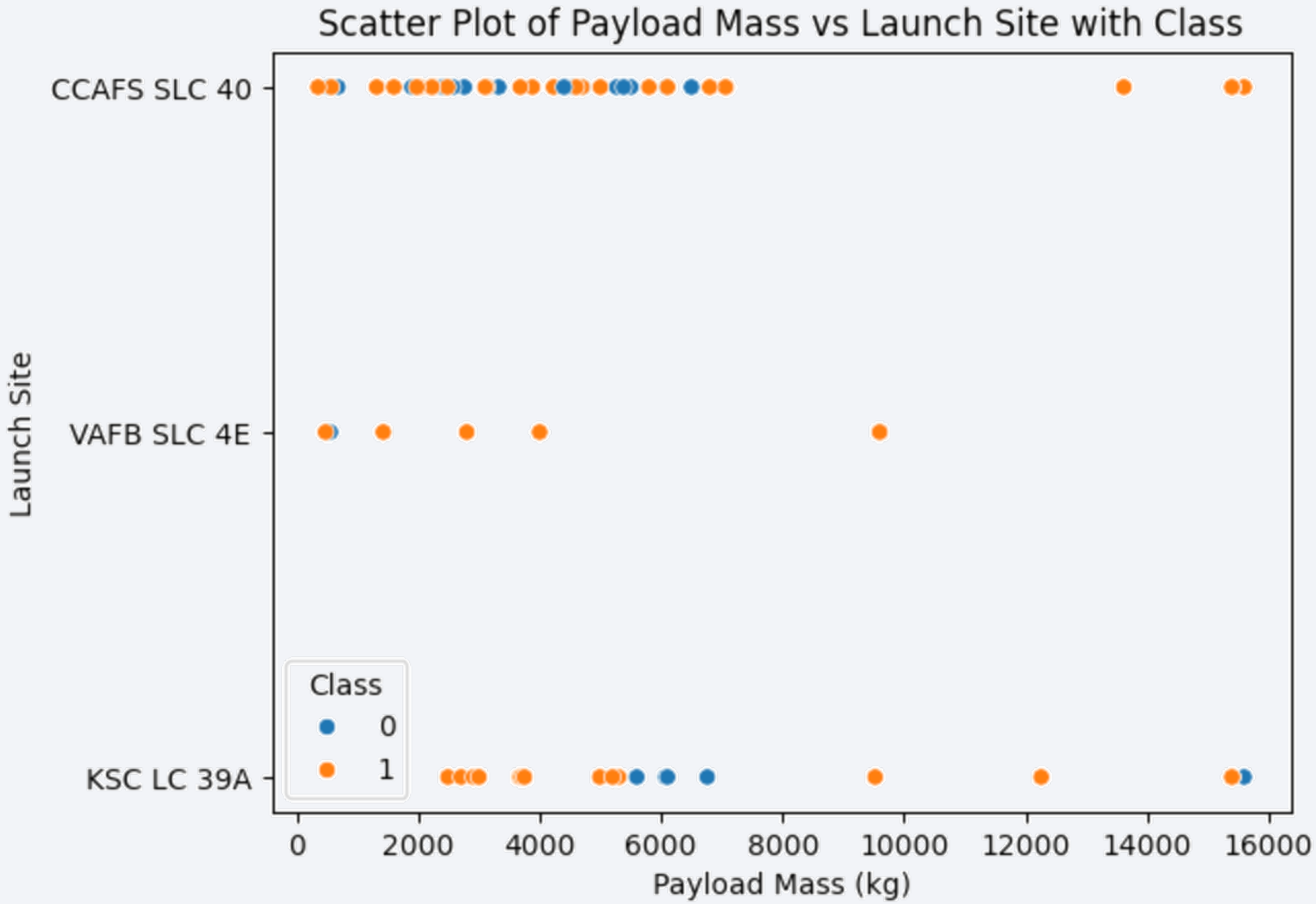
Flight Number vs. Launch Site

- Exploratory Data Analysis
  - Earlier flights had a lower success rate (blue = fail)
  - Later flights had a higher success rate (orange = success)
  - Around half of launches were from CCAFS SLC 40 launch site
  - VAFB SLC 4E and KSC LC 39A have higher success rates
  - We can infer that new launches have a higher success rate



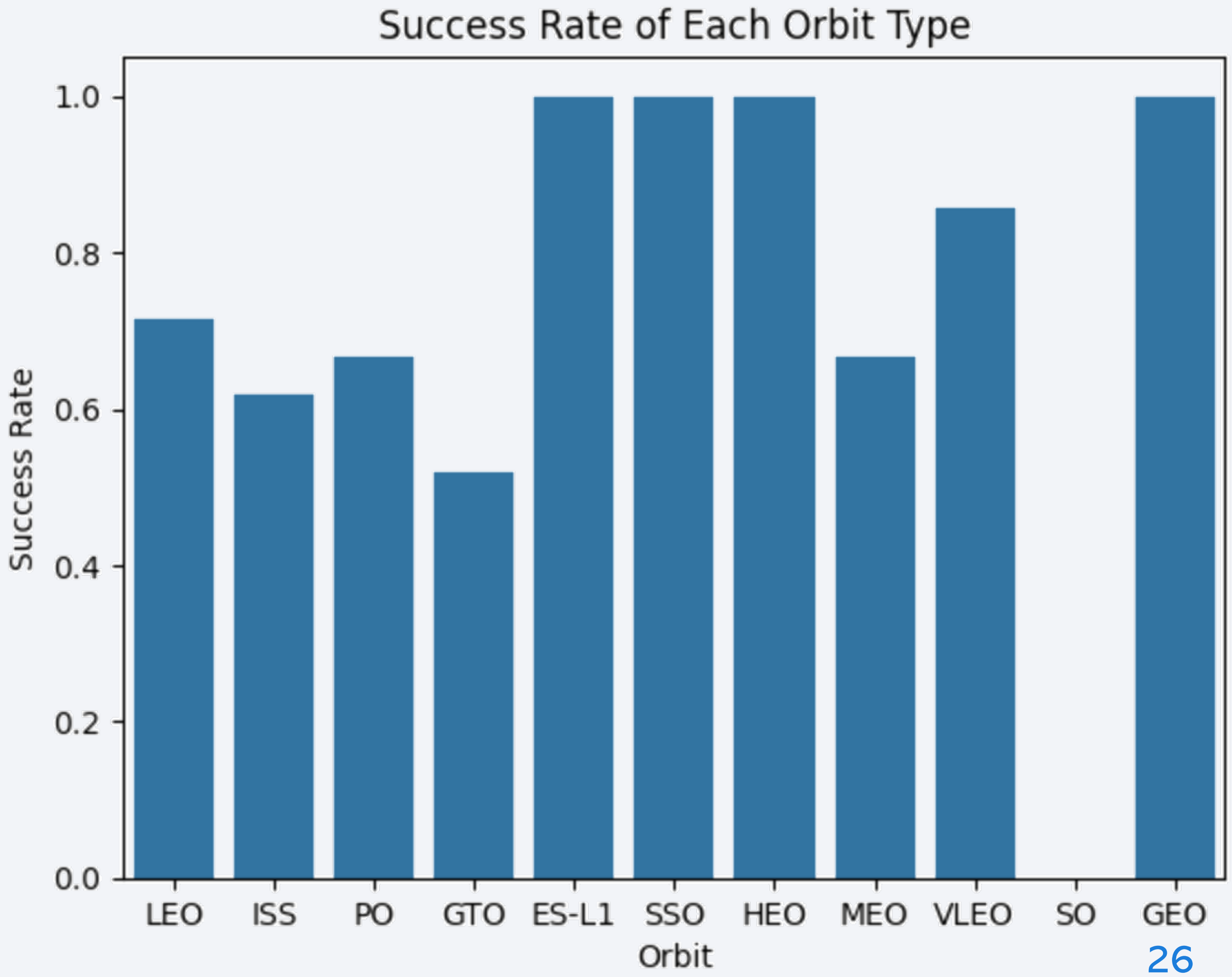
Payload vs. Launch Site

- Exploratory Data Analysis
  - Typically, the higher the payload mass (kg), the higher the success rate
  - Most launces with a payload greater than 7,000 kg were successful
  - KSC LC 39A has a 100% success rate for launches less than 5,500 kg
  - VAFB SKC 4E has not launched anything greater than ~10,000 kg



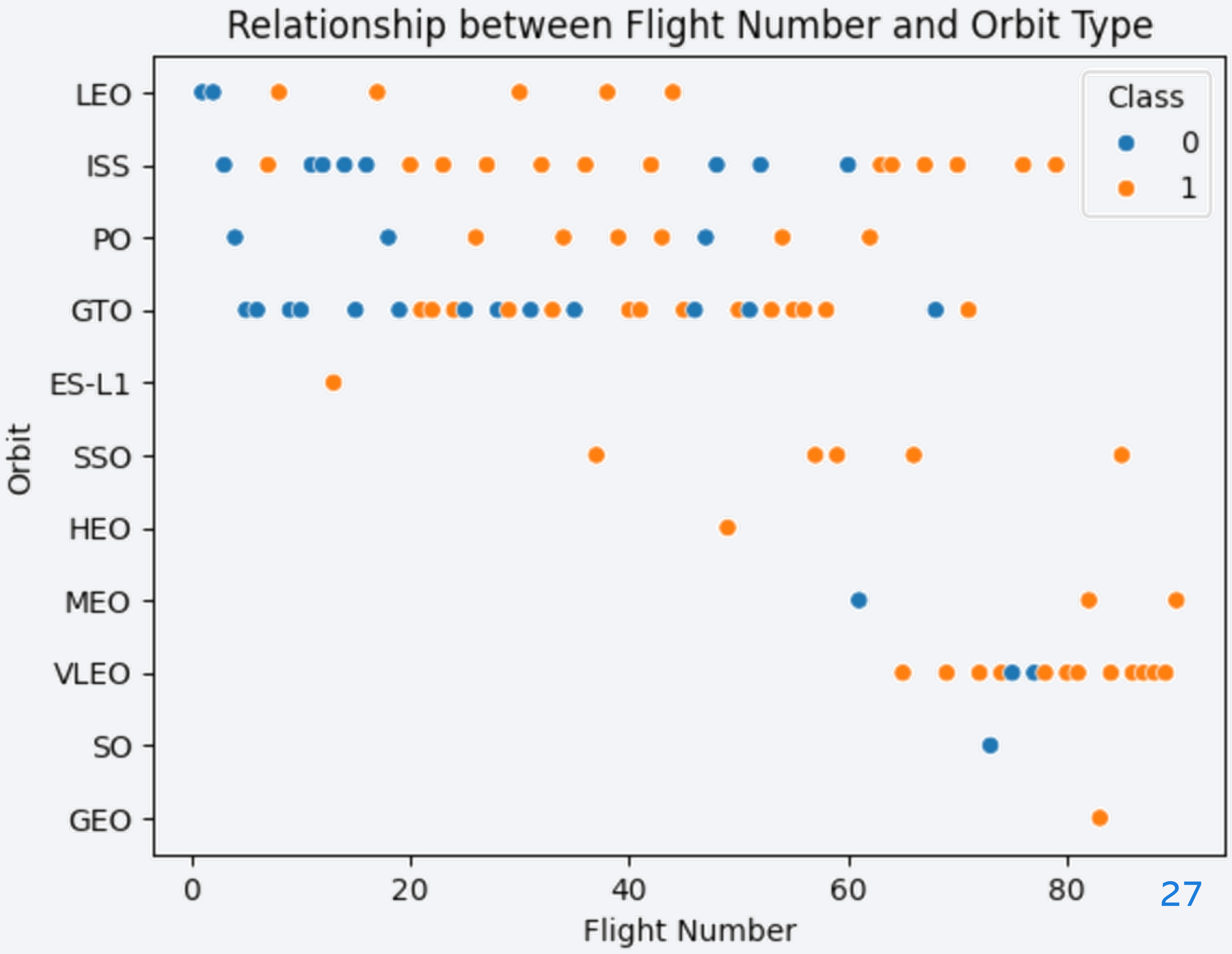
Success Rate vs. Orbit Type

- Exploratory Data Analysis
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



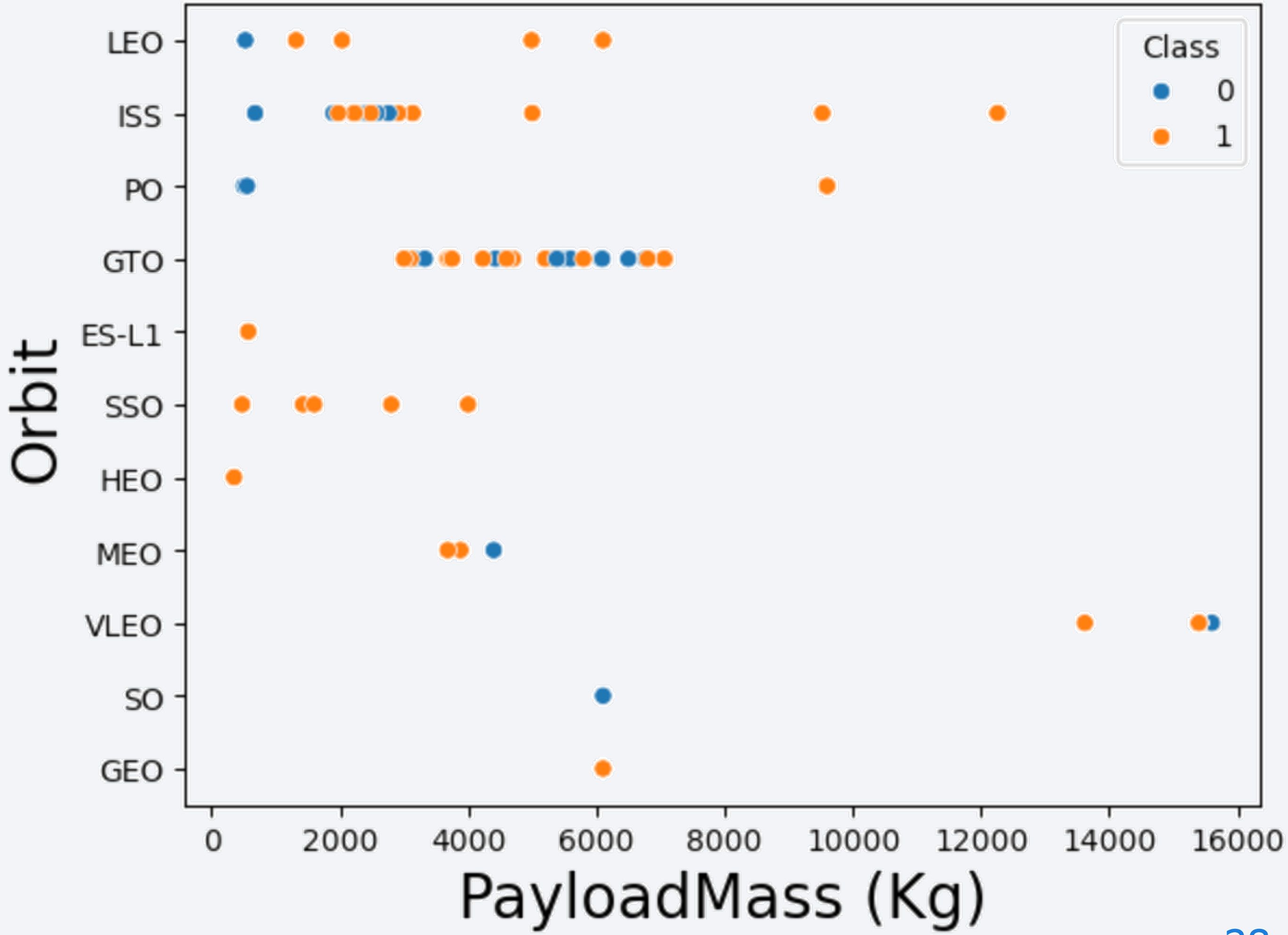
Flight Number vs. Orbit Type

- Exploratory Data Analysis
  - The success rate typically increases with the flight number for each orbit
  - VLEO includes 2 fails despite having promising initial flights.



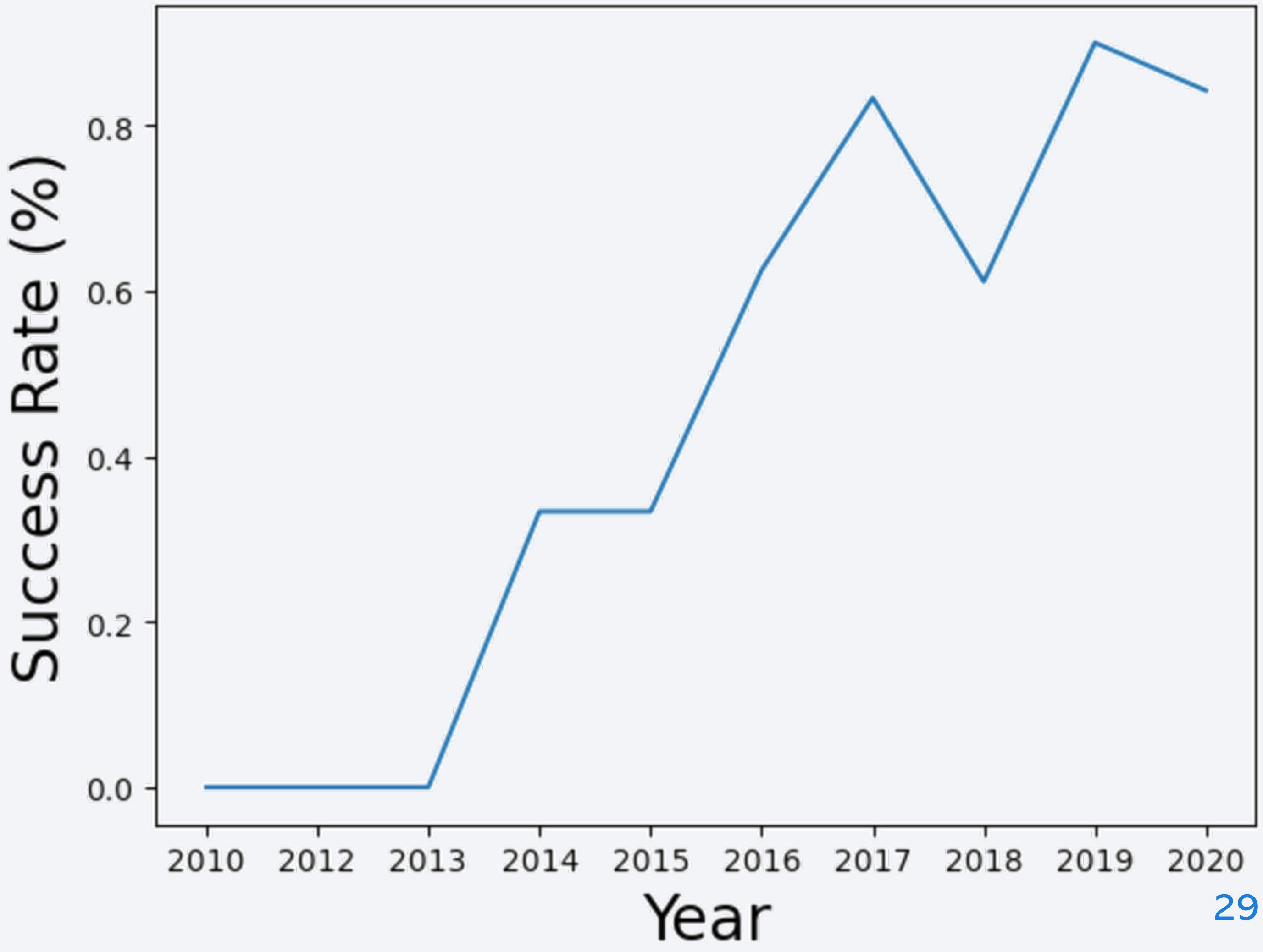
Payload vs. Orbit Type

- Exploratory Data Analysis
  - Heavy payloads are better with LEO, ISS and PO orbits
  - The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- Exploratory Data Analysis
  - Overall, the success rate has improved since 2013
  - The success rate improved from 2013-2017 and 2019
  - The success rate decreased in 2018 and 2020



All Launch Site Names

- Launch Site
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

- I have used GROUP BY to get the unique sites name

```
%sql SELECT Launch_Site FROM SPACEXTABLE GROUP BY Launch_Site
```

```
 * sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

- Here I have used LIKE ‘CCA%’ to find the site names start with CCA and used LIMIT 5 to only display 5 records

Display 5 records where launch sites begin with the string 'CCA'

[11]:
```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

[11]:

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

Total Payload Mass

- Using SUM and AVG, total payload mass carried by boosters launched by NASA (CRS) and average payload mass carried by booster version F9 v1.1 are displayed respectively.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_load_By_NASA FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Total_load_By_NASA
45596

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_load_By_F9_v1 FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

AVG_load_By_F9_v1
2928.4

First Successful Ground Landing Date

- I have used LIKE to find the all successful landing and ORDER BY date and LIMIT 1 to display the first one.

```
%sql SELECT * FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success%' ORDER BY Date LIMIT 1
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- I have used the logical combination of successful outcome and the playmass load of 4000<6000kg to display four launches.

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ <6000
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT Mission_Outcome, COUNT(Mission_Outcome) FROM SPACEXTABLE GROUP BY Mission_Outcome
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

```
%sql SELECT  Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ =\
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE )
```

```
 * sqlite:///my_data1.db
Done.
```

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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql SELECT Date, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date,0,5) = '2015'\
AND Landing_Outcome LIKE '%Failure (drone ship)%';
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Landing_Outcome	Booster_Version	Launch_Site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Ranking of 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

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count\
FROM SPACEXTABLE\
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'\
GROUP BY Landing_Outcome\
ORDER BY Outcome_Count DESC;
```

```
 * sqlite:///my_data1.db
Done.
```

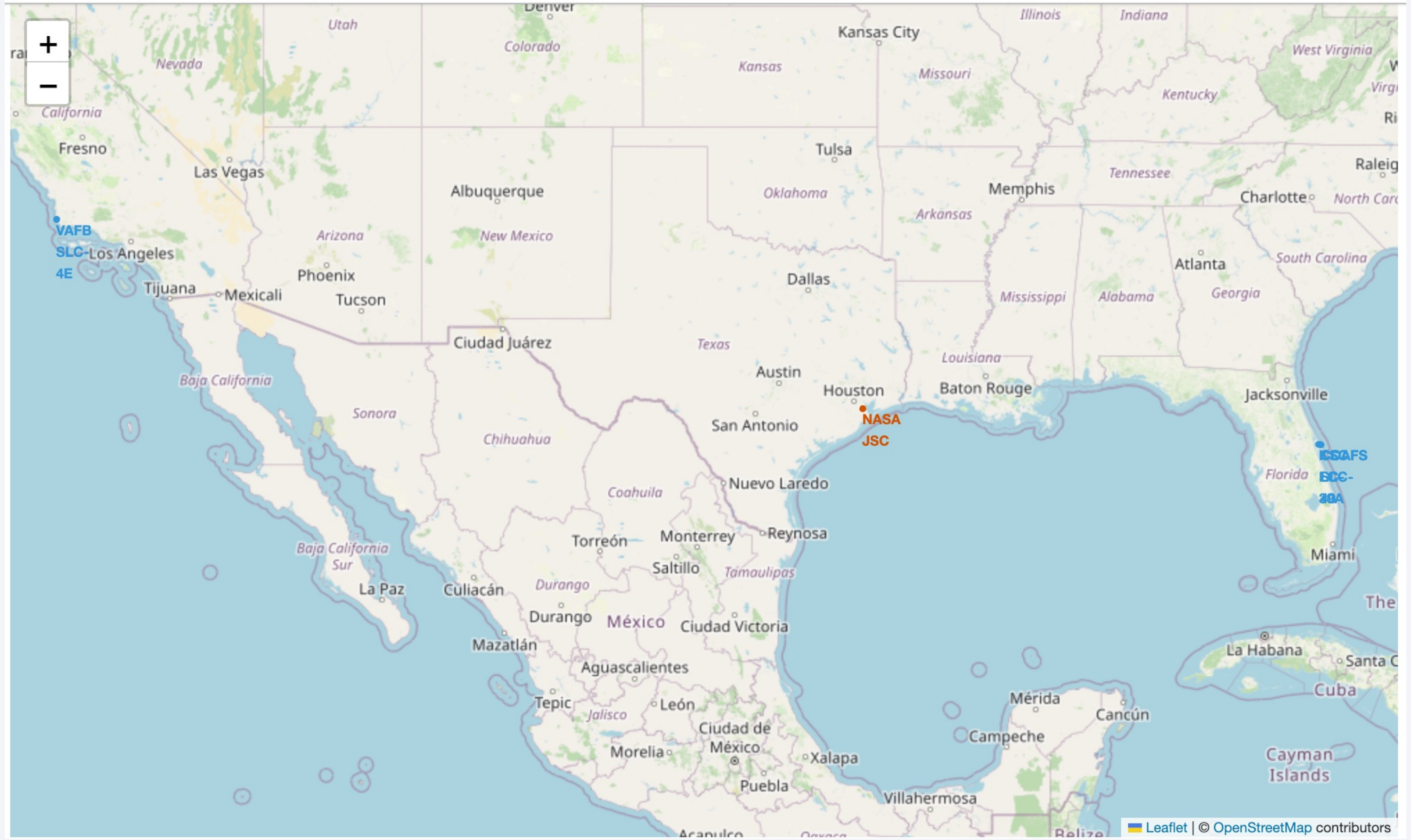
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

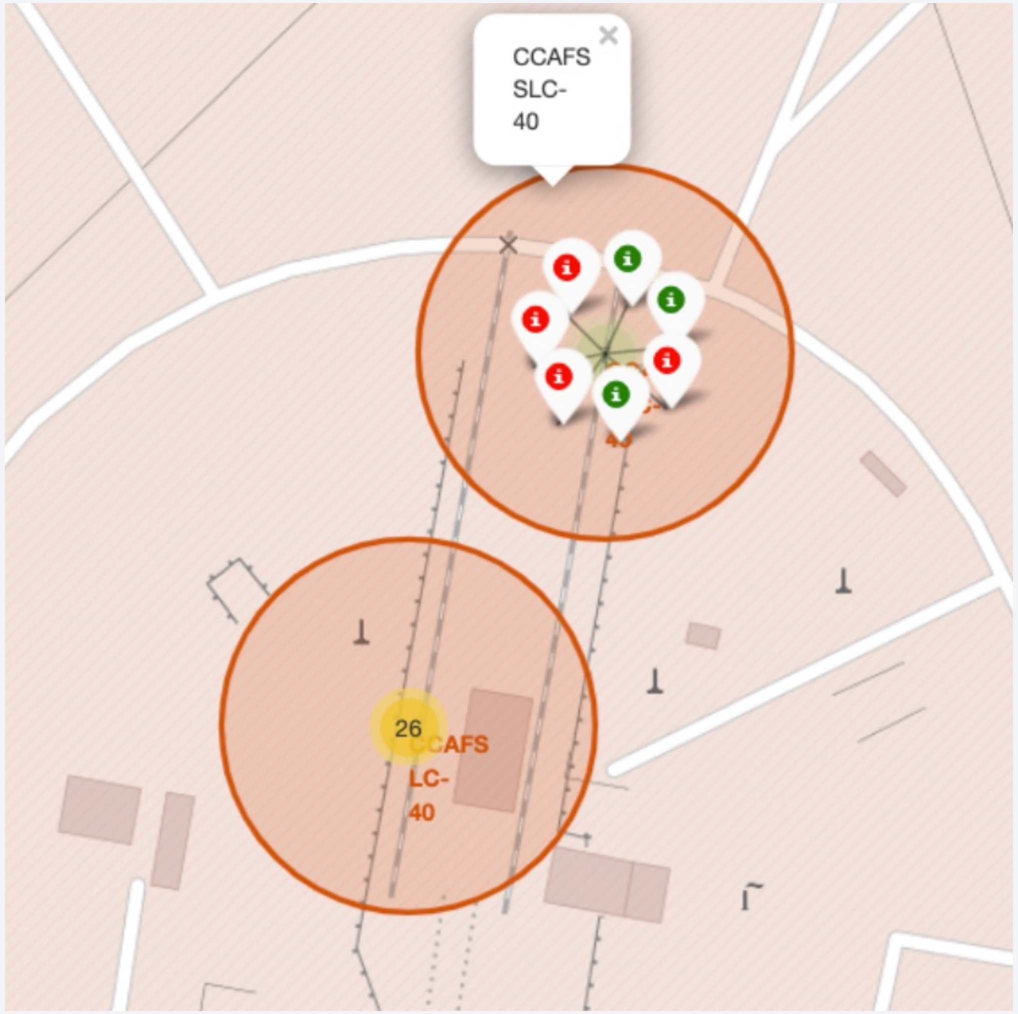
Launch Sites

Proximity to the Equator: Launch sites located closer to the equator benefit from easier access to equatorial orbits, leveraging the Earth's rotation for a prograde trajectory. Rockets launched from such sites receive a natural boost from the Earth's rotational speed, resulting in potential cost savings by reducing the need for additional fuel and boosters.



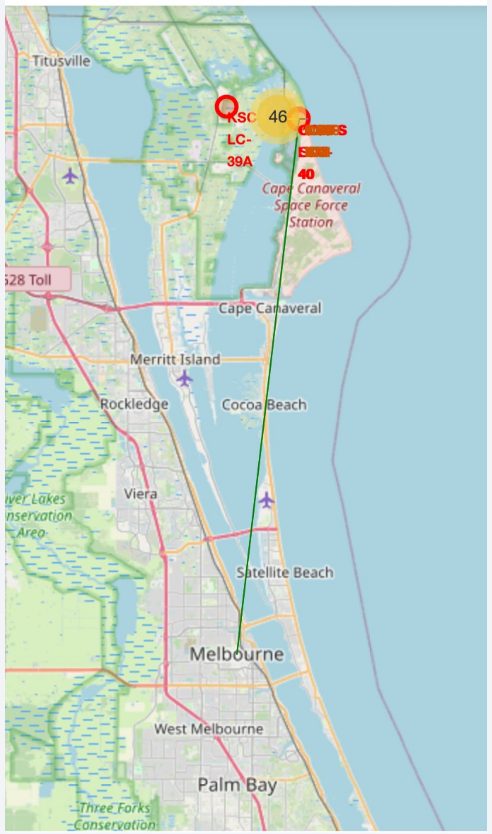
Launch Outcomes

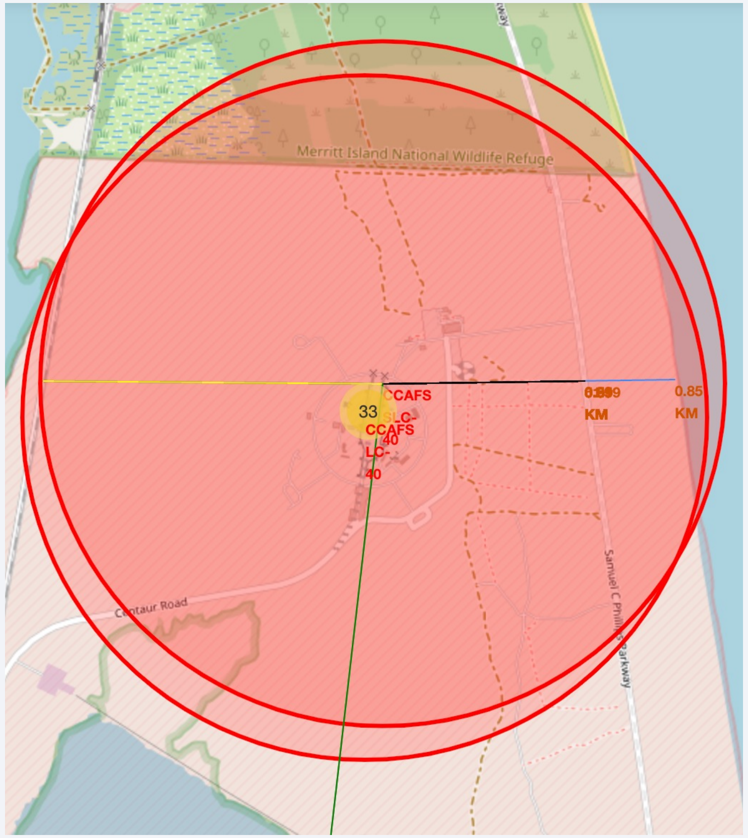
- At Each Launch Site:
  - Launch Outcomes:
    - Successful launches are denoted by green markers, while unsuccessful launches are represented by red markers.
    - For instance, launch site CCAFS SLC-40 exhibits a success rate of 3 out of 7 launches, equivalent to 42.9%.



Launch Site Proximity Distances

- CCAFS SLC-40:
- Coastal Location: Coastal positioning aids in ensuring that discarded rocket stages or unsuccessful launches do not endanger individuals or property along the launch trajectory.
- Safety and Security: Implementation of exclusion zones around the launch site is essential to prevent unauthorized access and ensure the safety of personnel.
- Transportation and Infrastructure, and Urban Proximity: Optimal location necessitates a balance between distance from potential damage caused by failed launches and proximity to transportation networks such as roads, railways, and docks, facilitating the transportation of personnel and materials to and from the site to support launch operations.



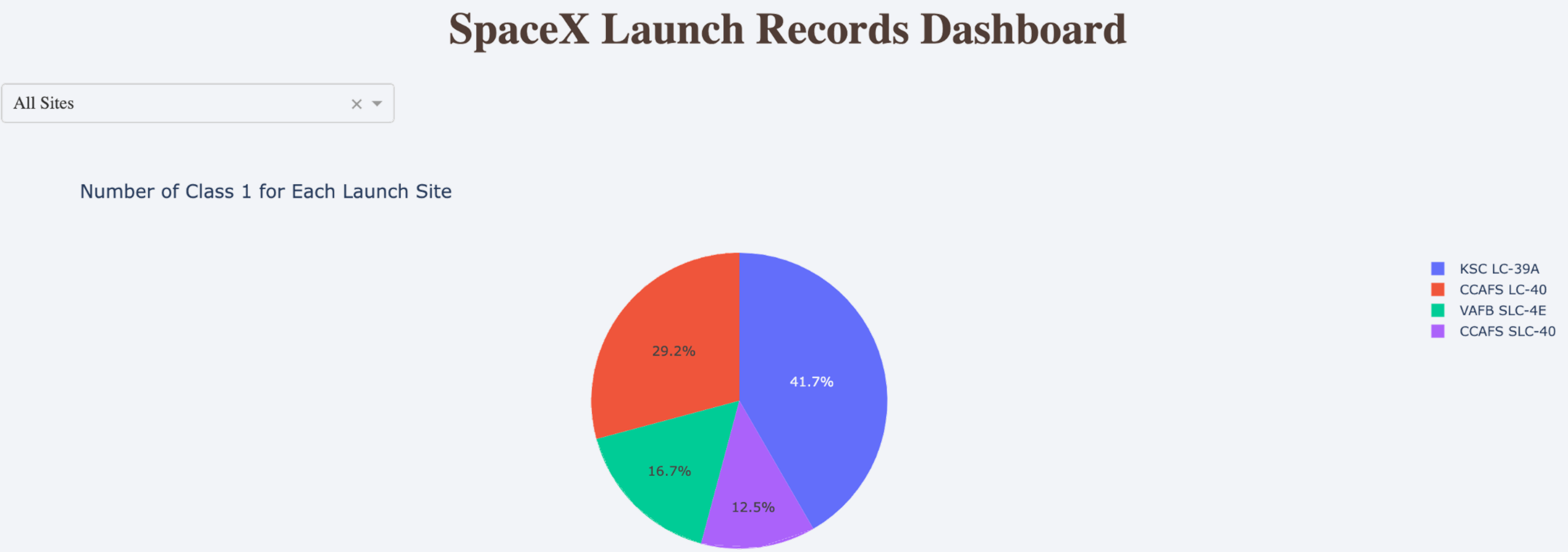


Section 4

Build a Dashboard with Plotly Dash

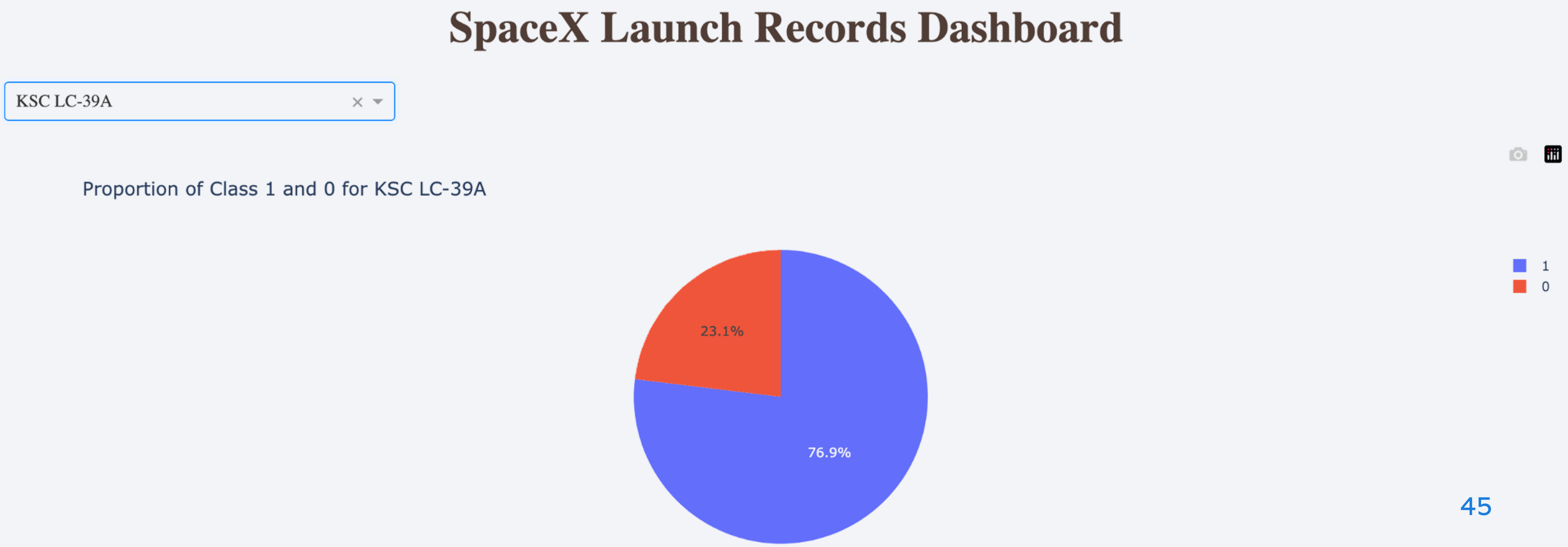
Sites Launch Success Counts

- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



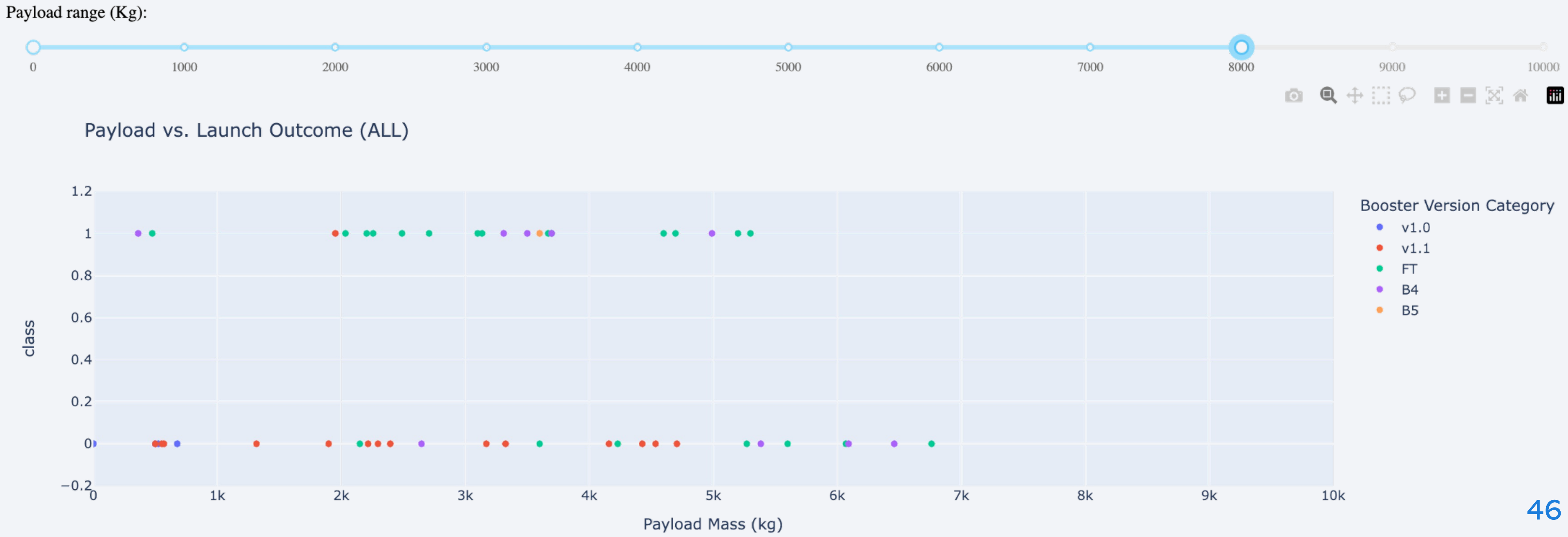
Launch Sites Success Rate

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches vs 3 failed launches



Payload Mass vs Success

- Payloads between 2,000 kg and 5,000 kg have the highest success rate by Booster Version

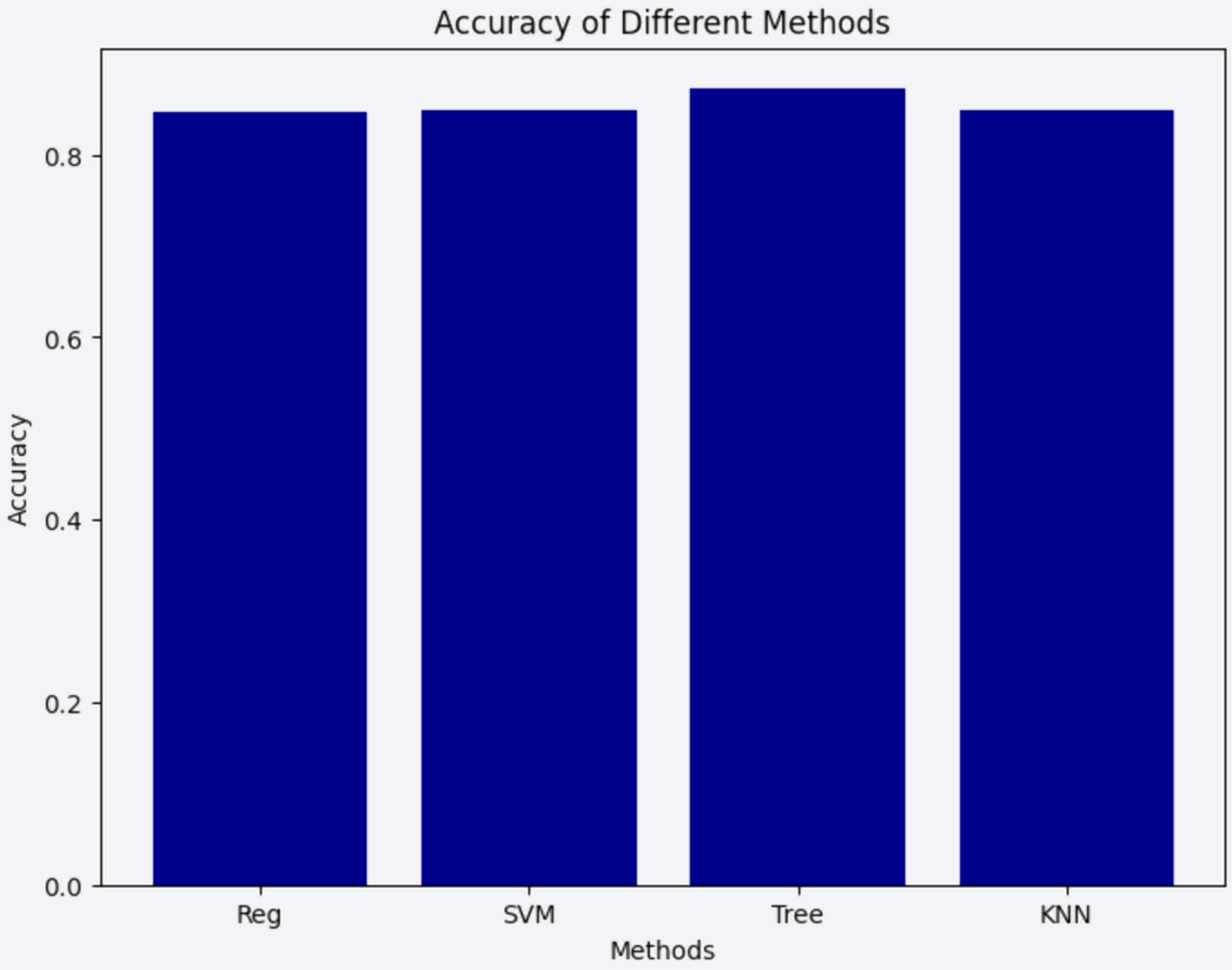


Section 5

Predictive Analysis (Classification)

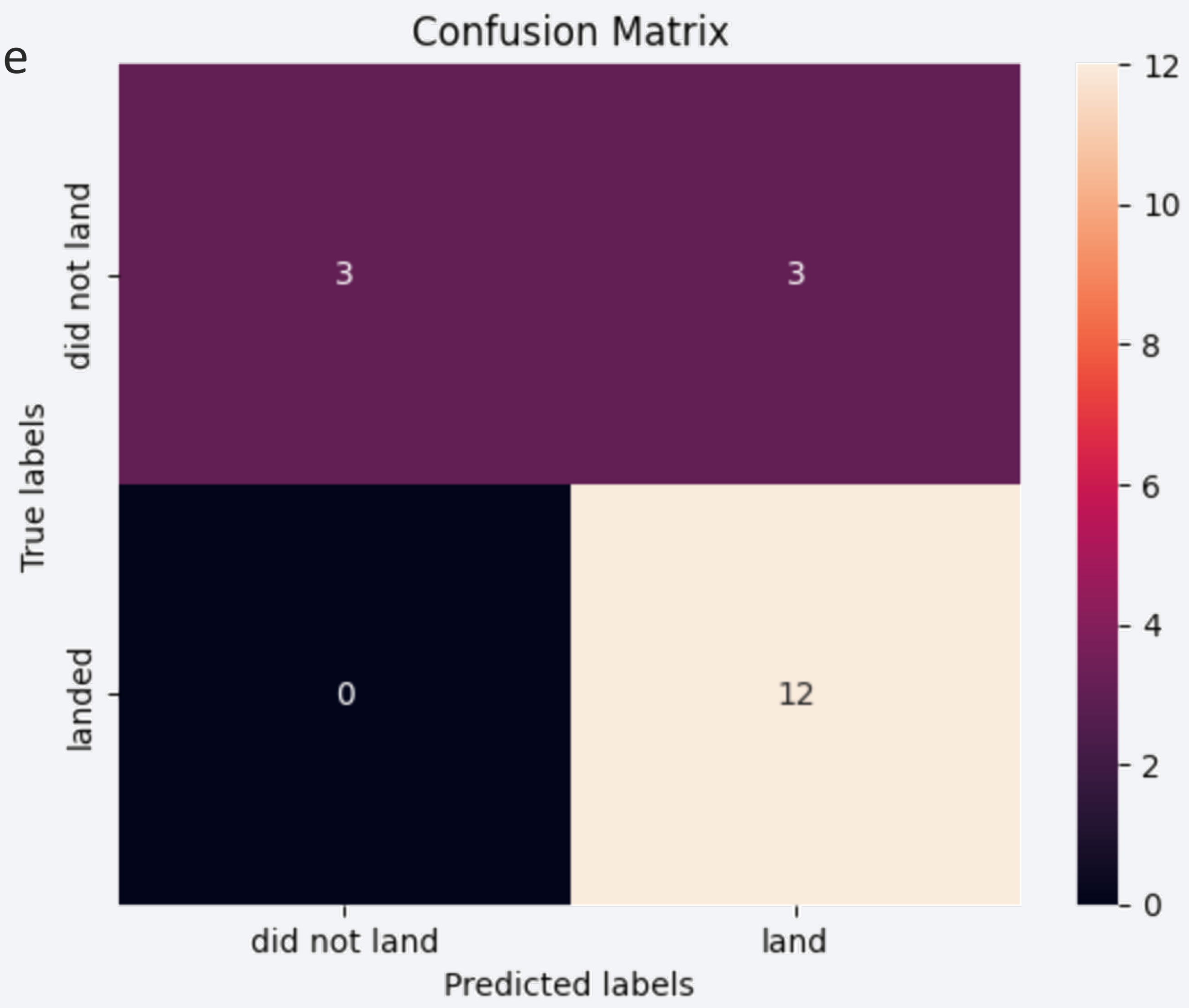
Classification Accuracy

- Decision Tree exhibits the best accuracy



Confusion Matrix

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical
  - Confusion Matrix Outputs:
    - 12 True positive
    - 3 True negative
    - 3 False positive
    - 0 False Negative



Conclusions

Key Insights:

- **Model Performance:** The models performed comparably on the test set, with a slight edge observed for the decision tree model.
- **Equatorial Advantage:** Most launch sites strategically located near the equator benefit from Earth's rotational speed, reducing fuel and booster requirements.
- **Coastal Proximity:** All launch sites are strategically situated near coastlines, ensuring safety and facilitating logistical operations.
- **Success Trends:** Over time, there is a notable upward trend in launch success rates, indicative of technological advancements and operational improvements.
- **KSC LC-39A Success:** KSC LC-39A demonstrates the highest success rate among launch sites, particularly excelling in launches under 5,500 kg.
- **Orbit Success:** Certain orbits consistently achieve a 100% success rate, including ES-L1, GEO, HEO, and SSO.
- **Payload Mass Impact:** Across all launch sites, higher payload masses correlate positively with success rates.

Conclusions

- **Future Research Considerations:**
  - **Dataset Expansion:** Augmenting the dataset size can enhance predictive analytics outcomes, facilitating broader generalization of findings.
  - **Feature Analysis and PCA:** Conducting further analysis through feature exploration or principal component analysis (PCA) may enhance model accuracy by identifying critical features.
  - **Exploring XGBoost:** Investigating the potential of XGBoost, a powerful yet unexplored model in this study, could provide valuable insights into its comparative performance with other classification models.

Thank you!