IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Naomi Kamau
22/04/2025

Outline

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

Executive Summary

In this project, I analyzed historical data from SpaceX missions to identify key factors influencing rocket landing success rates and predict the success of first stage landings. This is important to know since reusability saves millions of dollars per launch. I explored data collection using web scraping and APIs, performed EDA with SQL and visualizations, built an interactive dashboard with Dash and Folium maps, and created predictive models using machine learning to predict landing success. My findings revealed trends across launch sites, payload mass and booster versions, allowing me to better understand what constitutes a successful mission.

Introduction

The commercial space industry is rapidly evolving, and SpaceX is a key player in transforming space transportation. This project aims to extract valuable insights from SpaceX mission data, with a particular emphasis on landing success. Understanding these factors helps to reduce costs and maximize reusability.

Business Question: Can we identify patterns to predict whether a Falcon 9 rocket will land successfully?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology: Data was collected from multiple sources ie APIs(SpaceX, OpenNotify), Web Scrapping(Wikipedia) and CSV datasets.
- Perform data wrangling: Combined all datasets into a unified dataframe, cleaned missing values, extracted new features, standardized and encoded features for machine learning.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models: Built classification models using Logistic Regression, SVM, Decision Tree, KNN to predict landing success. Used GridSearchCV and cross-validation to optimize hyperparameters.

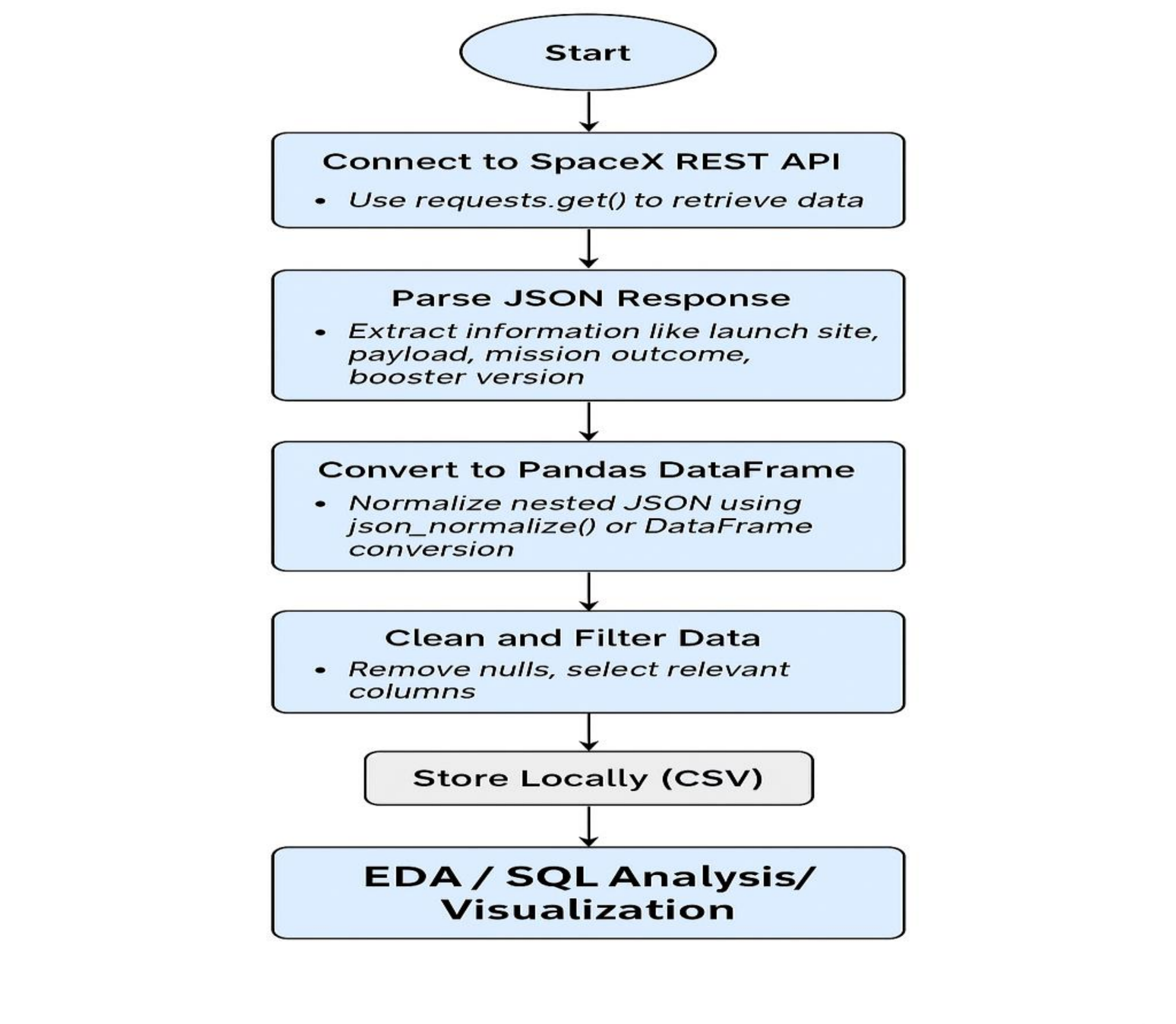
Data Collection

I used two data sources:

- **API Calls** to SpaceX launch data via OpenNotify and Launch Library APIs.
- **Web Scraping** from SpaceX's Wikipedia pages using BeautifulSoup.

Data Collection – SpaceX API

- **API endpoint used:** https://api.spacexdata.com/v4/launches
- **Tools used:** Python requests, json, pandas
- **Purpose:** To collect real-world data on SpaceX launches for analysis
- **Output:** Structured dataset ready for cleaning and exploration
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Used **BeautifulSoup** and **requests** libraries to scrape launch data from the **SpaceX website**.
- Parsed HTML tables to extract mission details like date, site, payload, and landing outcomes.
- Cleaned and structured the data using **pandas** for consistency and further analysis.
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Data wrangling involved:

- Combining all datasets into a unified dataframe.
- Cleaned missing values in fields like PayloadMass and LandingPad.
- Extracted new features like Booster Type, Launch Year, and Success Class.
- Standardized and encoded features for machine learning.
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Objective: Understand patterns and relationships between features like launch sites, payload mass, orbit types, and mission outcomes.**
- **Charts and Why They Were Used: Flight Number vs Launch Site (Catplot)-**To check how launch frequency varied across sites and whether success rates improved over time.
- **Payload Mass vs Launch Site (Catplot)-**To explore how payload mass differs by site and its potential impact on success.
- **Orbit Type vs Success Rate (Bar Chart)-**To visualize which orbits have higher landing success rates.
- **Flight Number vs Orbit (Scatter Plot)-**To identify whether certain orbits are more likely to succeed as SpaceX gained experience.
- **Payload Mass vs Orbit (Scatter Plot)-**To analyze the effect of payload size on launch outcomes for different orbit types.
- **Success Rate Over Years (Line Chart)-**To show how SpaceX's mission success rate improved annually since 2013.
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/edadataviz.ipynb

EDA with SQL

- **Launch Sites:** Retrieved all unique launch site names.
- **Payload Analysis:**
  - Total payload mass by NASA (CRS) launches.
  - Average payload mass for F9 v1.1 booster version.
- **Landing Outcomes:**
  - Counted successful and failed missions.
  - Ranked landing outcomes by frequency between 2010–2017.
- **Filtering Records:**
  - Selected launches from CCAFS sites.
  - Filtered launches with max payload mass.
  - Listed 2015 failed drone ship landings with month and booster details.
- **Date-Based Queries:** Found first successful ground pad landing date.
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

- **Folium Map: Objects and Purpose**
- **Markers:** Plotted each launch site and labeled them with site names.
  → *To visually identify SpaceX launch locations on the map.*
- **Colored Circles:** Highlighted launch site areas.
  → *To emphasize launch zones and make them easy to spot.*
- **Clustered Markers (Green/Red):** Represented individual launch outcomes (success/failure).
  → *To show launch frequency and performance visually.*
- **Lines (Polylines):** Connected launch sites to nearby points like cities, coastlines, highways.
  → *To measure and visualize proximity for site analysis.*
- **MousePosition Plugin:** Displayed coordinates on hover.
  → *To assist in capturing exact nearby location coordinates.*
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

I used the following visualizations and interactions:

- **Pie Chart (Success Rates):**
  → *Displays overall and site-specific launch success/failure counts.*
- **Scatter Plot (Payload vs Success):**
  → *Shows relationship between payload mass and launch success.*
  → *Color-coded by Booster Version for deeper insights.*
- **Dropdown Menu:**
  → *Allows selection of a specific launch site or all sites.*
- **Payload Range Slider:**
  → *Enables filtering launches based on payload mass range.*
- These interactions allow dynamic, real-time exploration of SpaceX launch data for better pattern discovery.
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/spacex-dash-app.py.1

Predictive Analysis (Classification)

Key Phrases:

- **Data Preprocessing:** Standardized features using StandardScaler()
- **Train-Test Split:** Used train_test_split() (80% train, 20% test)
- **Model Training:** Built and tuned models using GridSearchCV() with cv=10
- **Evaluation:** Compared accuracy and confusion matrix for each model
- **Best Model: Decision Tree** (Highest validation accuracy: **88.9%**)
- **Classification Models Tested:** Logistic Regression, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbors (KNN)
- https://github.com/N-Shelmith/capstoneprojectrepo/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

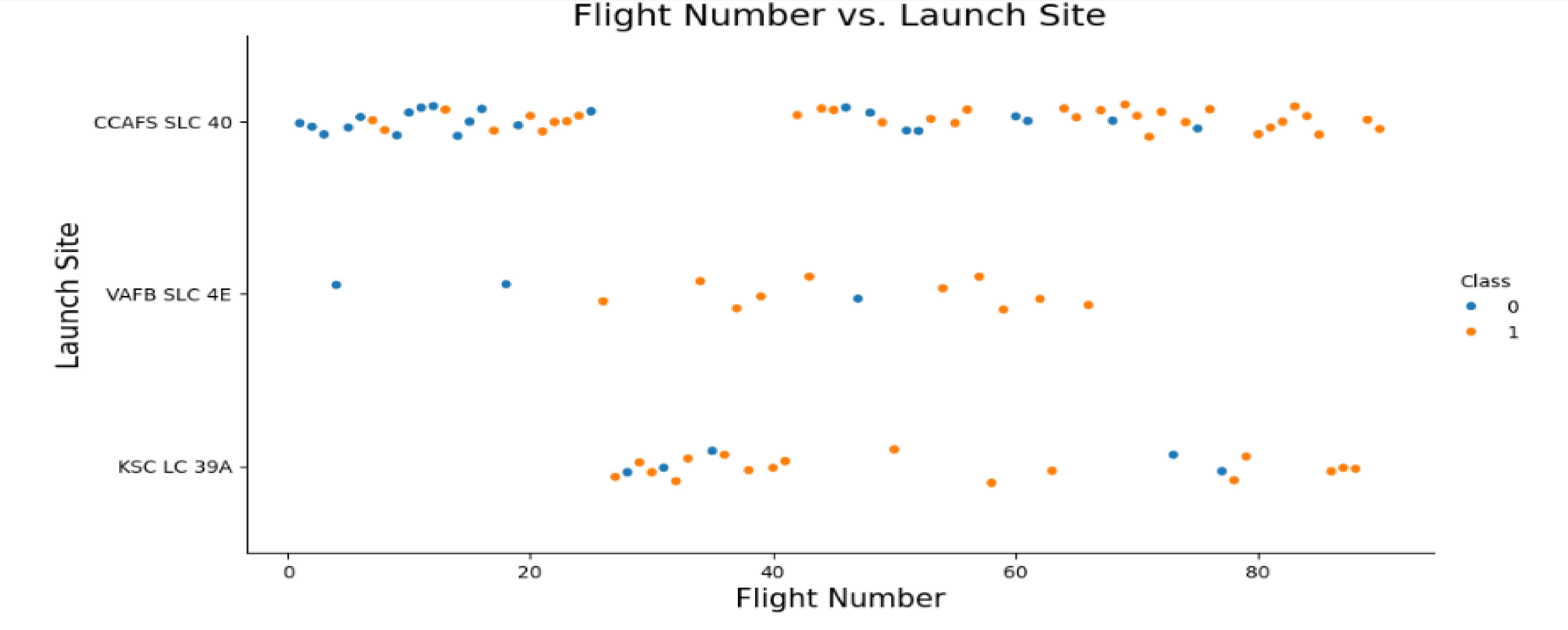
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

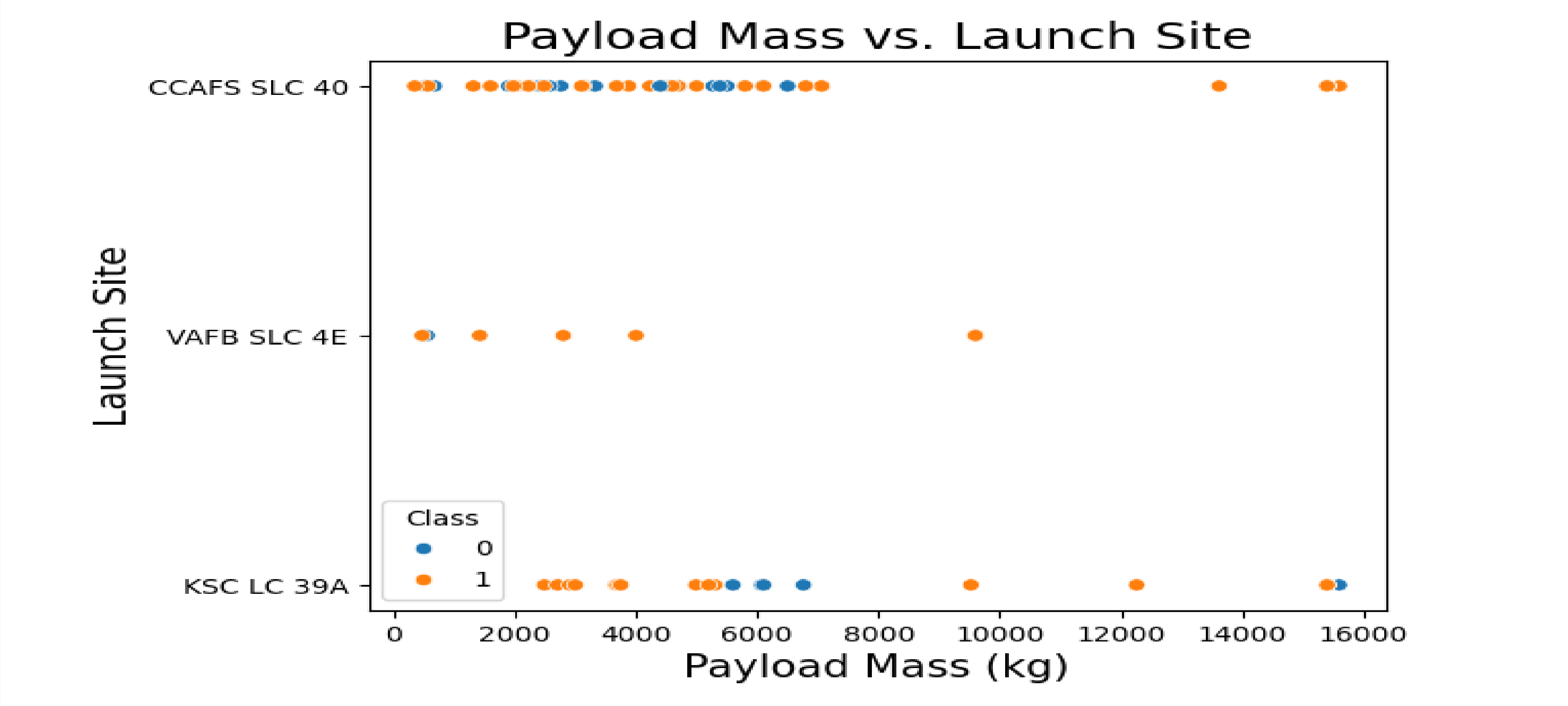
Insights drawn from EDA

Flight Number vs. Launch Site

- This scatterplot helped identify how frequently each launch site was used over time. **CCAFS SLC 40** was used more frequently.

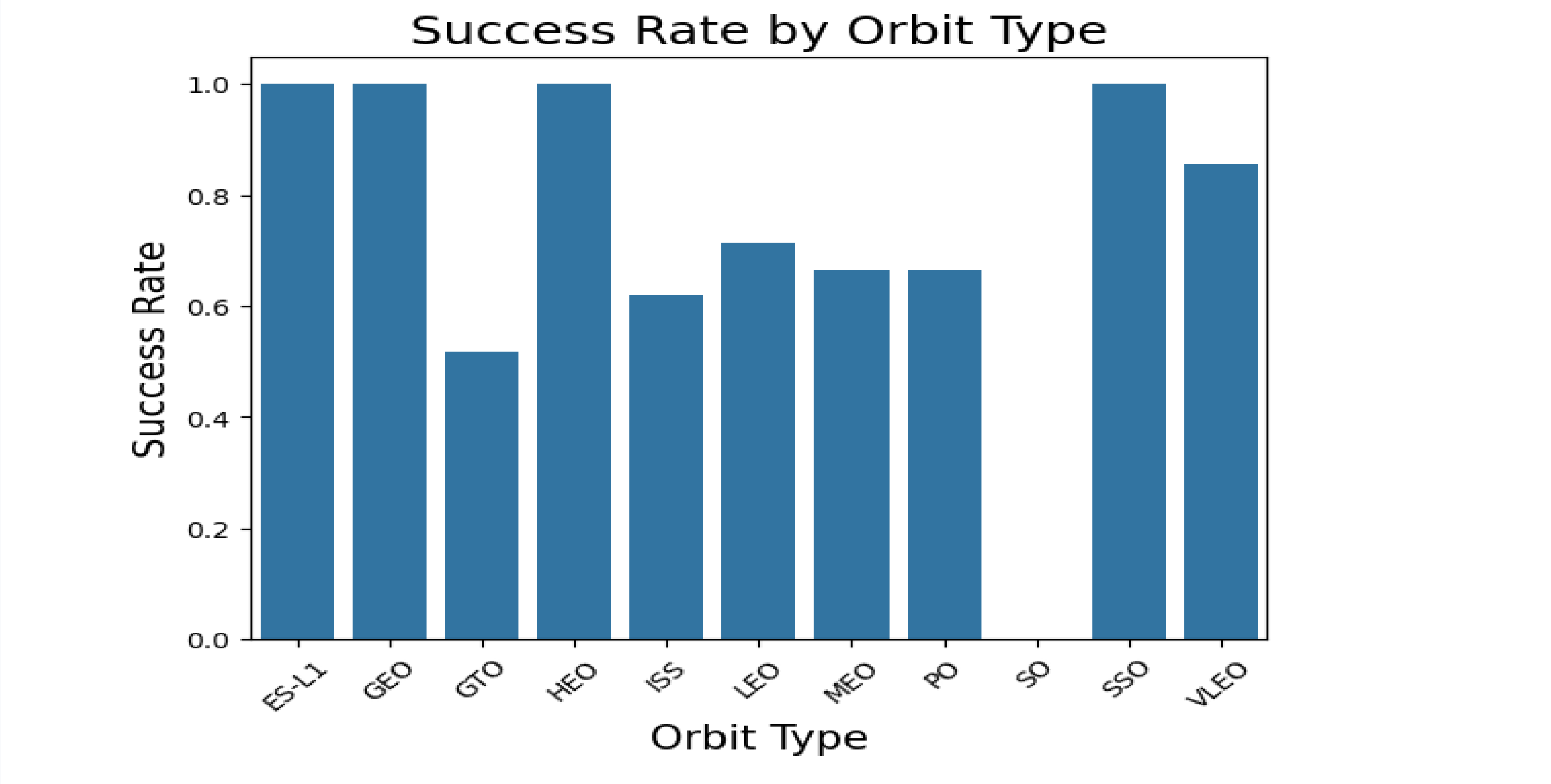


Payload vs. Launch Site



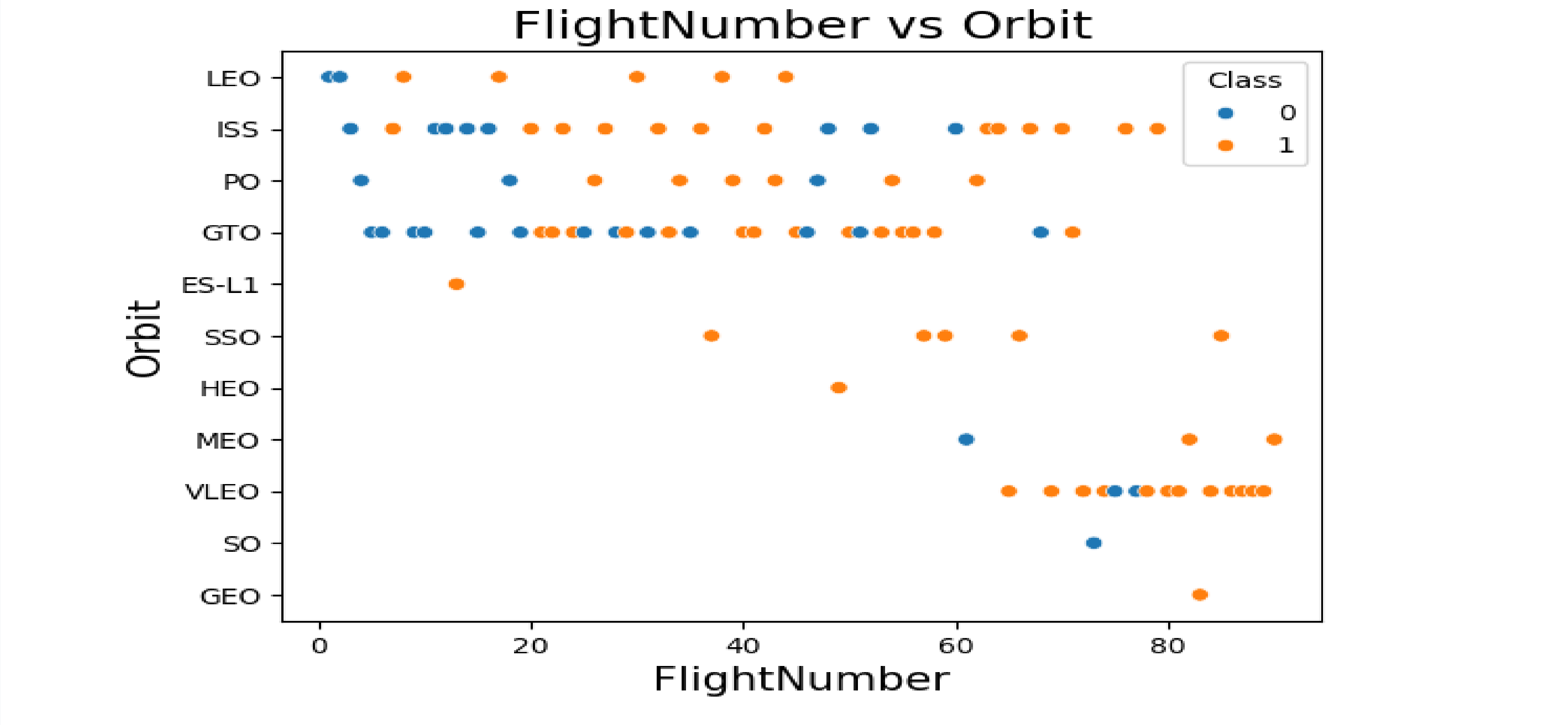
- Showed that **KSC LC-39A** handled more **heavy payload launches**, while **VAFB SLC-4E** had smaller payloads.

Success Rate vs. Orbit Type



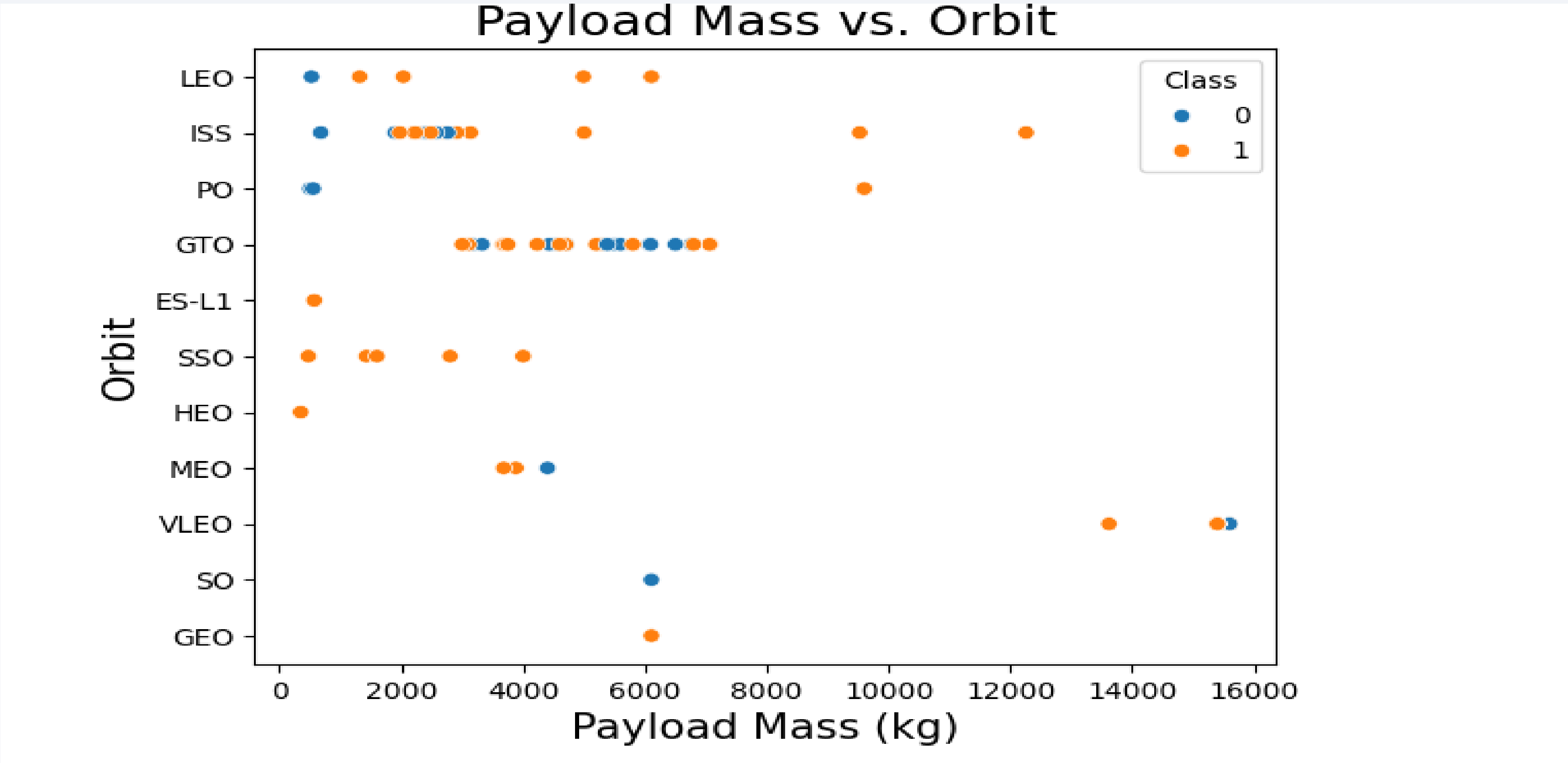
- **ES-L1,GEO,HEO and SSO orbits** had the **highest success rates**, showing they are more reliable for landings.

Flight Number vs. Orbit Type



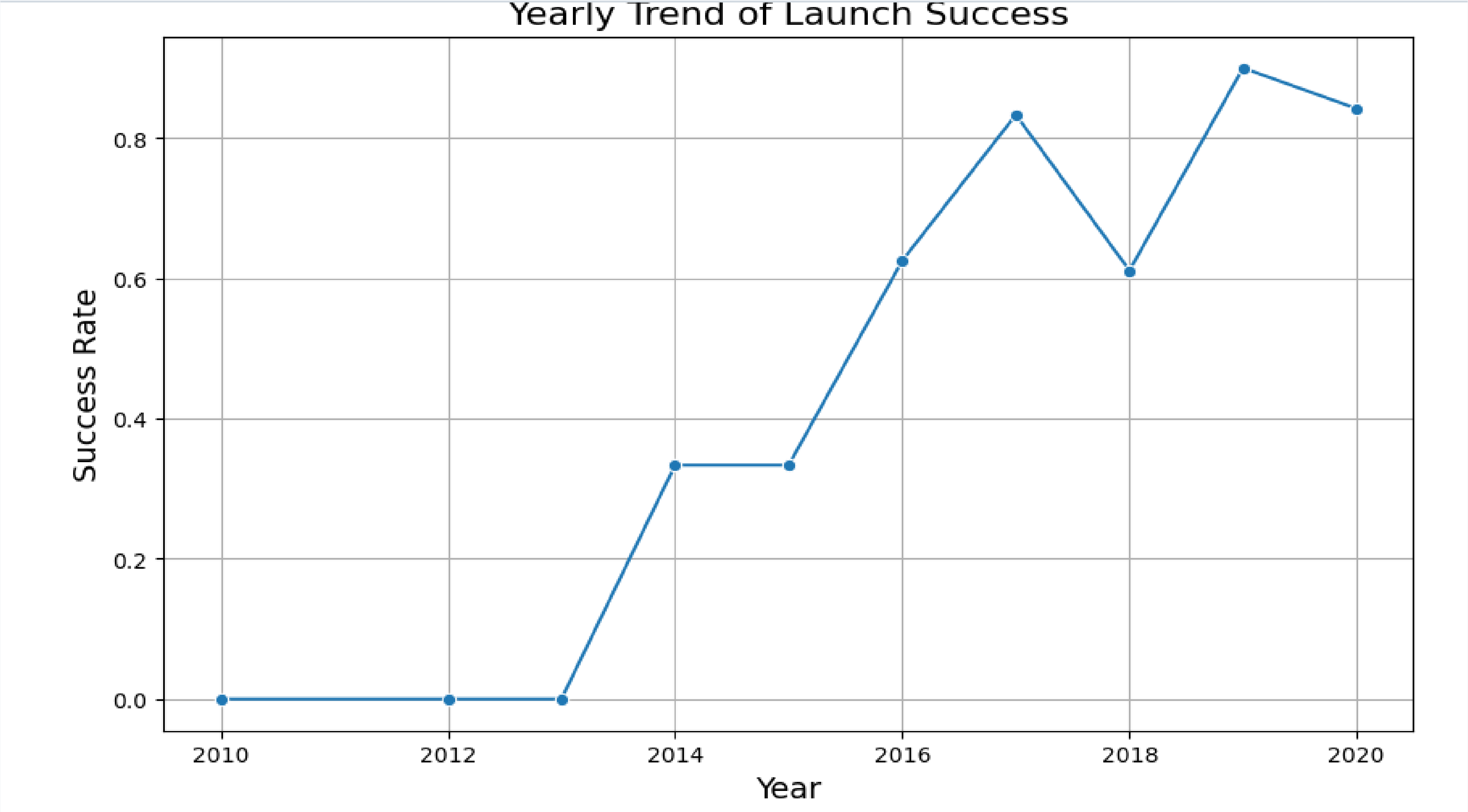
- In the **LEO** orbit, success seems to be related to the number of flights. Conversely, in the **GTO** orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- Revealed a **consistent improvement in mission success** over the years, especially after 2013.

All Launch Site Names

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

- Listed all four distinct launch sites used by SpaceX.

Launch Site Names Begin with 'CCA'

```
%sql SELECT *FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%'LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parach
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 (parach
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

Total Payload Mass

- Calculated **total payload mass launched** for NASA's CRS missions which was **48213.**

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") AS Total_Payload FROM SPACEXTBL WHERE "Customer" LIKE '%NASA (CRS)%';
```

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

Total_Payload
48213

Average Payload Mass by F9 v1.1

- Calculated the average payload mass carried by booster version F9 v1.1 which was **2534.65**

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") AS Avg_Payload FROM SPACEXTBL WHERE "Booster_Version" LIKE 'F9 v1.1%';
```

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

Avg_Payload
2534.6666666666665

First Successful Ground Landing Date

- Identified the **earliest mission** with a successful landing on a ground pad which was **2015-12-22**

```
%sql SELECT MIN("Date") AS First_Success_Ground_Pad FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

First_Success_Ground_Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Listed the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE "Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MAS
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes: Success was 100 and Failure 1.

```
%sql SELECT "Mission_Outcome", COUNT(Mission_Outcome) AS Count FROM SPACEXTBL GROUP BY "Mission_Outcome";
```

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

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass of **15,600**.

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

- Highlighted months with **drone ship landing failures** in 2015.

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

Month	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

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

- 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.

```
%%sql
SELECT "Landing_Outcome", COUNT(*) AS Outcome_Count
FROM SPACEXTBL
WHERE
    CAST(substr("Date", -4) AS INTEGER) BETWEEN 2010 AND 2017
GROUP BY "Landing_Outcome"
ORDER BY Outcome_Count DESC;
```

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

Landing_Outcome	Outcome_Count

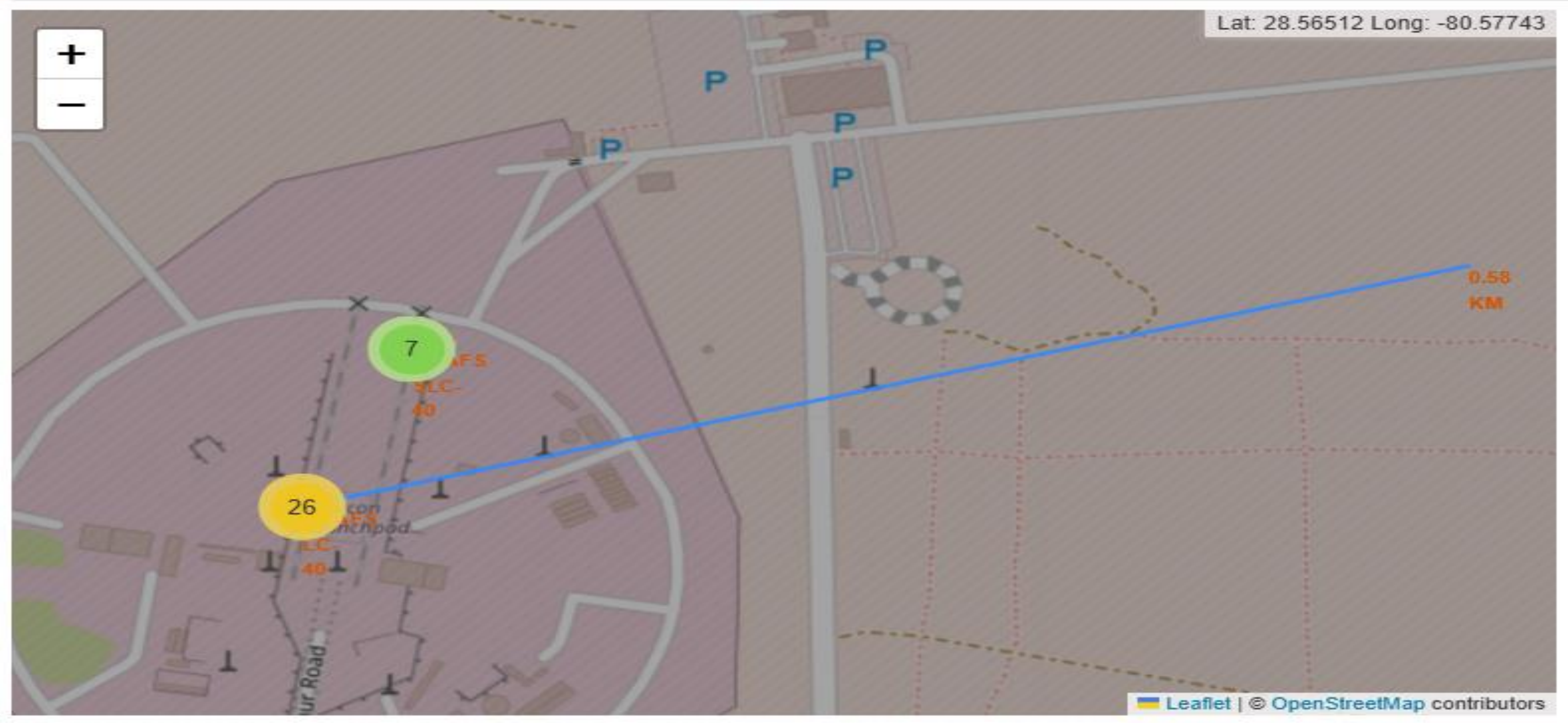
Section 3

Launch Sites Proximities Analysis

Folium Circle and Marker on Launch Sites

Success/Failed launches for each site

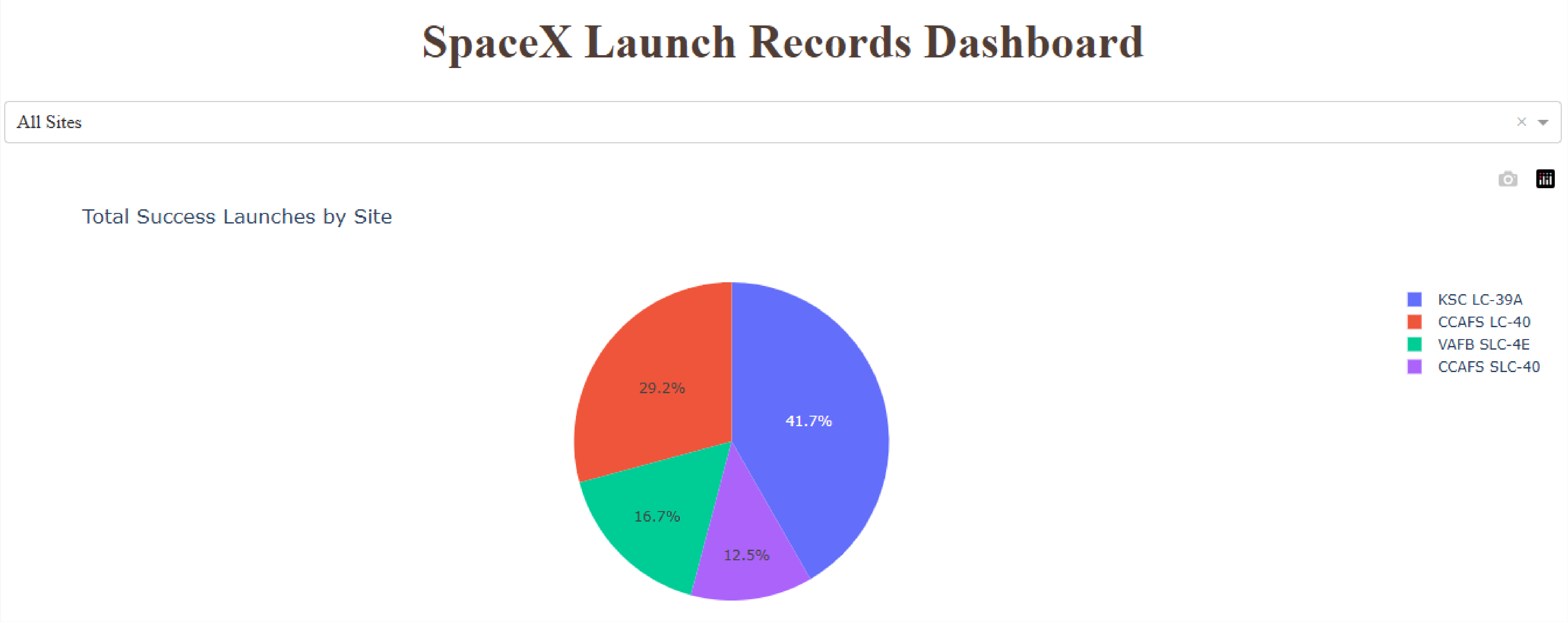
Distance between Launch Sites & Proximities



Section 4

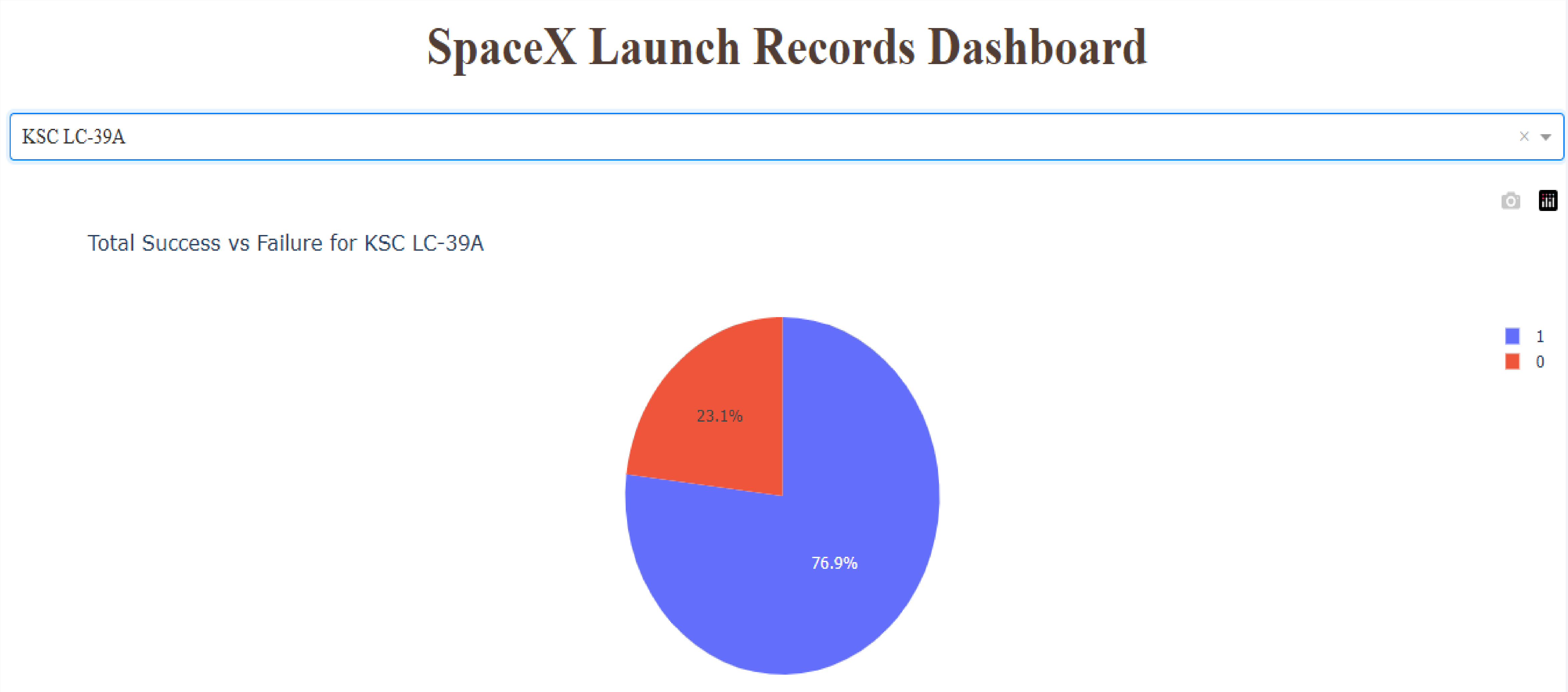
Build a Dashboard with Plotly Dash

Pie Chart for All sites launch success



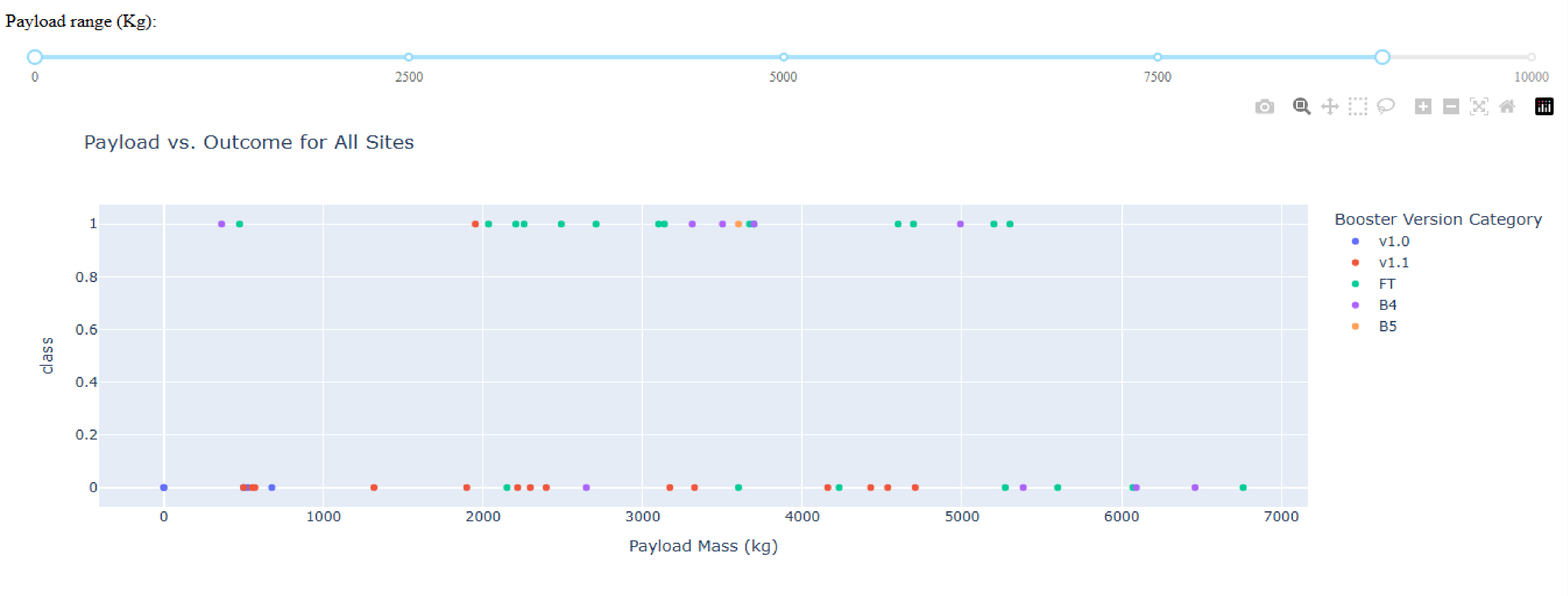
- The pie chart reveals that KSC LC-39A has the highest success of 41.7% and CCAFS SLC 40 has the lowest success rate of 12.5%.

Pie Chart for KSC LC-39A



- This site has a success rate of 76.9% and Failure of 23.1%

Payload vs. Launch Outcome scatter plot for all sites

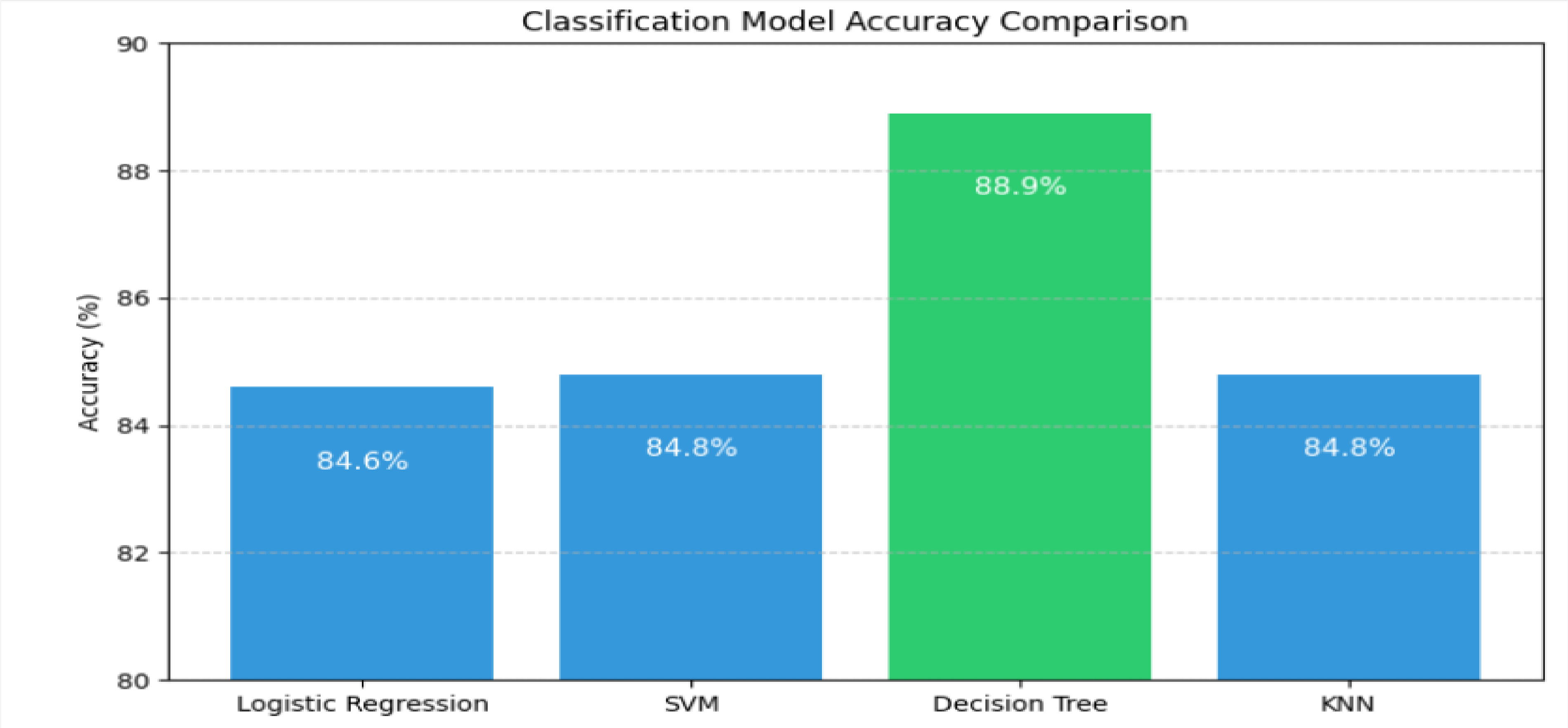


- With a payload mass range of 2000-5500kg booster version FT has a higher success rate.

Section 5

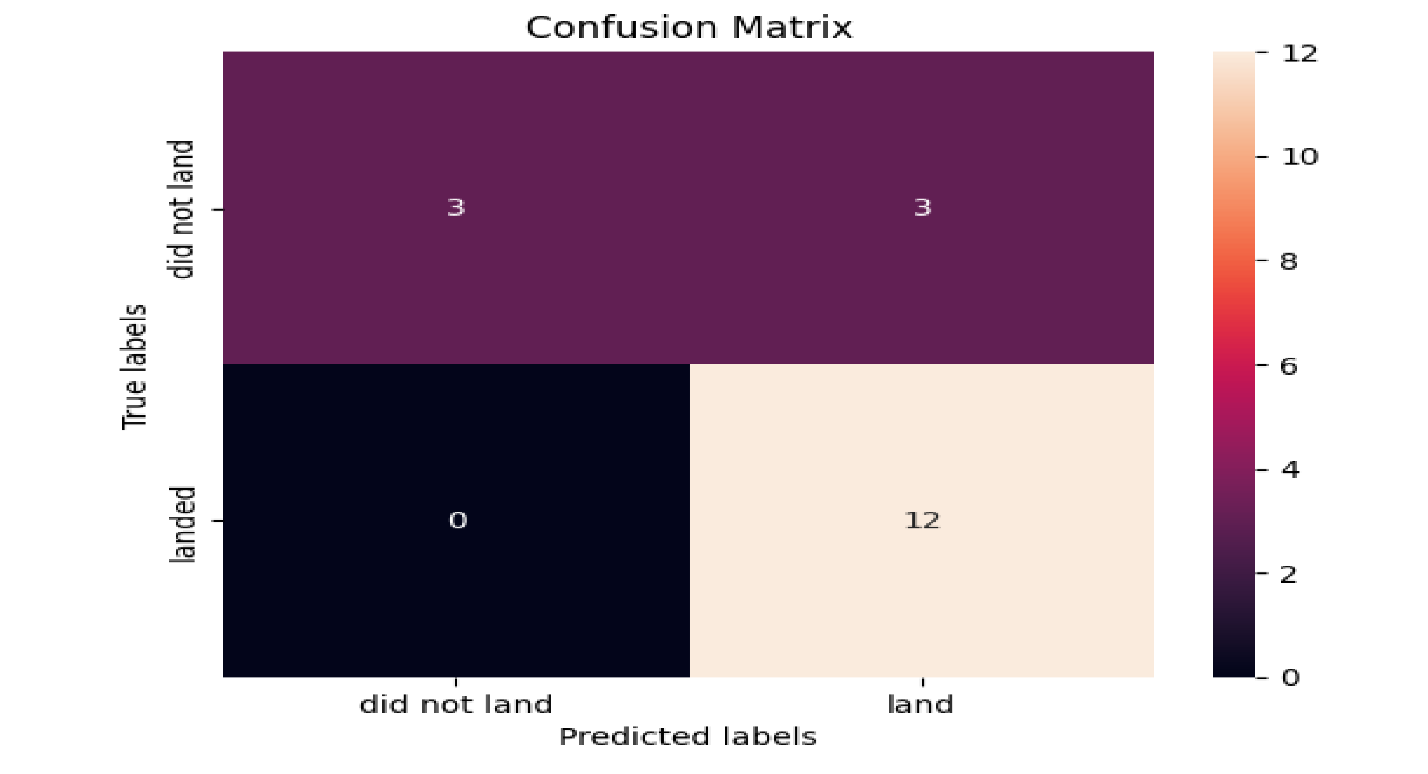
Predictive Analysis (Classification)

Classification Accuracy



- The model with the highest classification accuracy is **Decision Tree** of **88.9%**.

Confusion Matrix



- **True Positives (e.g., 12)**: Model correctly predicted successful landings.
- **False Positives (e.g., 3)**: Model incorrectly predicted a success when the actual result was failure.
- **True Negatives**: Correctly identified failed landings.
- **False Negatives**: Fewer in number, indicating good sensitivity.

Conclusions

- **I successfully analyzed SpaceX launch data** using Python, SQL, Folium, Plotly Dash, and Machine Learning.
- **Exploratory analysis** revealed that **KSC LC-39A** had the highest success rate and payloads between **4000–6000 kg** often led to successful landings.
- **Decision Tree Classifier** achieved the highest model accuracy (88.9%) and can be leveraged for future launch outcome predictions.
- Mapping tools showed launch sites were **strategically close to coastlines and transport infrastructure**.
- **Recommendation**: SpaceX and similar companies could optimize booster designs and payload planning based on model predictions to reduce failure rates and cost.

Recommendations

- SpaceX could **enhance prediction models by including weather, wind speed, and temperature data**.
- Use ML classification results in **real-time mission control decision support systems**.
- Explore **deep learning or ensemble methods** to further improve classification performance.
- Regularly retrain models with **updated data to keep predictions accurate** over time.
- Predicting successful landings supports **SpaceX's goal of reusability**.
- Reducing failure risks saves **millions per launch ($60M+ per rocket)**.

Appendix

Python Code Snippets

- Data collection via REST API and web scraping using requests, BeautifulSoup, and pandas.
- Interactive map plotting using Folium, DivIcon, and MarkerCluster.
- Model building using LogisticRegression, SVM, DecisionTreeClassifier, and KNeighborsClassifier.

SQL Queries

- Total and average payload mass by launch site and customer.
- Filtering launch success/failure records by conditions (date, location, outcome).
- Ranking landing outcomes between selected date ranges.

Charts & Outputs

- EDA plots: bar charts, pie charts, line graphs.
- Dashboard: Dropdowns, sliders, and scatter plots for interaction.
- Confusion matrix heatmaps for model evaluation.
- Accuracy comparison bar chart for model performance.

Datasets Used

- spacex_launch_dash.csv, dataset_part_2.csv, dataset_part_3.csv, and augmented SpaceX API/web scraped data.

Thank you!