



IBM DATA SCIENCE

CAPSTONE PROJECT

OUTLINE

Summary

Introduction

Methodology

Results

Conclusion

SUMMARY OF METHODOLOGIES

. Data Collection

- Historical launch data was gathered from SpaceX's official website, public datasets, and relevant spaceflight databases. The data includes information on launch dates, payload mass, launch sites, orbit types, rocket versions, and landing outcomes.

Data Wrangling

- Raw data was cleaned and transformed to ensure consistency and usability. This included handling missing values, converting categorical variables into numerical format, normalizing numerical features, and engineering new variables to enhance model input quality.

INTRODUCTION

Project Background and Context

SpaceX has revolutionized space travel by dramatically lowering launch costs, with a Falcon 9 launch priced at approximately \$69.75 million as of 2024—significantly more affordable than traditional alternatives. A critical driver of this cost efficiency is SpaceX's innovative use of reusable rocket technology. Since 2017, the company has routinely landed and reused the first stage of the Falcon 9 rocket with a high success rate, showcasing the practical and economic benefits of reusability in orbital launches.

Exploratory Data Analysis (EDA)


- EDA was performed to understand the structure and distribution of the data. Statistical summaries and visualizations (e.g., histograms, box plots, and correlation matrices) were used to identify patterns, trends, and potential relationships between features and landing success.
-

. Interactive Visual Analysis

- Interactive dashboards and plots were created using tools such as Plotly and Dash. These visualizations allowed for dynamic exploration of the data, enabling users to filter and examine how different variables—like payload mass or orbit type—impact landing outcomes.

Predictive Analysis (Classification)

- Multiple machine learning algorithms were applied to model and predict the binary outcome of landing success (success vs. failure). Models evaluated included Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting.

- 
-
- Understanding the factors that contribute to a successful first-stage landing is essential for accurately estimating launch costs and optimizing mission planning. This knowledge not only supports more informed decision-making for customers and stakeholders but also provides valuable insight into the operational reliability of reusable space systems.

TASKS PERFORMED

- Data collection methodology
- Perform data wrangling (handling missing values, filter data)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models (Use Scikit-Learn for Preprocessing data)

DATA COLLECTION SCRAPPING

- Send API Request (SpaceX endpoint: /launches)
- Get JSON Response & Parse JSON Data (Extract nested fields like rocket, payloads, cores)
- Normalize & Flatten Data (Using pandas.json_normalize or loops)
- Select Required Columns (e.g., FlightNumber, Date, PayloadMass, Orbit)
- Handle Missing Values (in Payload Mass)
- Store in DataFrame (pandas DataFrame for analysis) and save
- <https://github.com/SurajKumar1411/data-science-1/blob/main/spacex-data-collection-api.ipynb>

DATA WRANGLING

- In the dataset, booster landings are categorized based on their success or failure and the type of landing attempted. For example, "True Ocean" indicates a successful landing in the ocean, whereas "False Ocean" means the attempt to land in the ocean was unsuccessful. Similarly, "True RTLS" refers to a successful return-to-launch-site ground landing, while "False RTLS" means the booster failed to land on the ground pad. "True ASDS" shows a successful drone ship landing, and "False ASDS" indicates an unsuccessful one. For training purposes, we simplify these outcomes into binary labels: "1" represents a successful landing, and "0" indicates a failed landing attempt.
- <https://github.com/SurajKumar1411/data-science-1/blob/main/spacex-Data%20wrangling.ipynb>

EDA WITH SQL

- Select All Launch Records Queried all columns from the SPACEXTBL table to preview the dataset. •Filter by Booster Version Retrieved specific launches using conditions like `BoosterVersion = 'F9 v1.1'`.
- Count Successful Lanches Counted records where `LandingOutcome` indicated a successful mission.
- Group by Launch Site Counted number of launches per `LaunchSite`. EDA with SQL [GitHub File Link](#)
- Filter by Launch Outcome Extracted launches based on successful or failed outcomes using `WHERE` conditions.

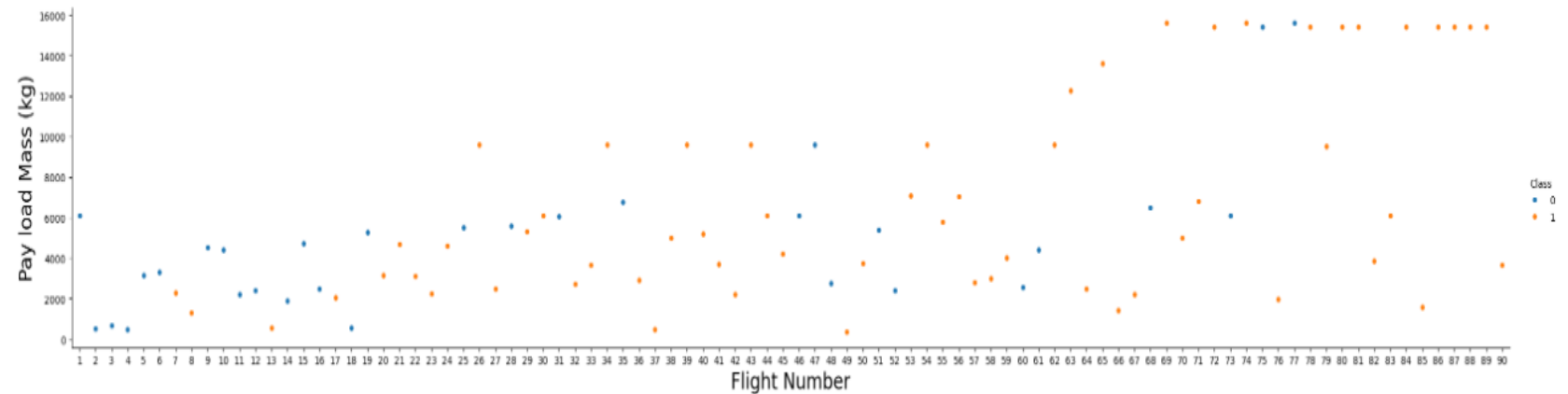
- Group by Orbit Type Summarized data by different Orbit categories to analyze usage frequency.
 - Calculate Average Payload Mass Used `AVG(PayloadMass)` to find average payload weights per orbit.
-
- Filter by Date Range Queried launches within specific dates using `WHERE Date BETWEEN`
 - Sort Launches Ordered launches by `PayloadMass` or `Date` using `ORDER BY`. Limit Records Restricted query results using `LIMIT` to show top N records
 - https://github.com/SurajKumar1411/data-science-1/blob/main/jupyter-eda-sql_sqlite.ipynb

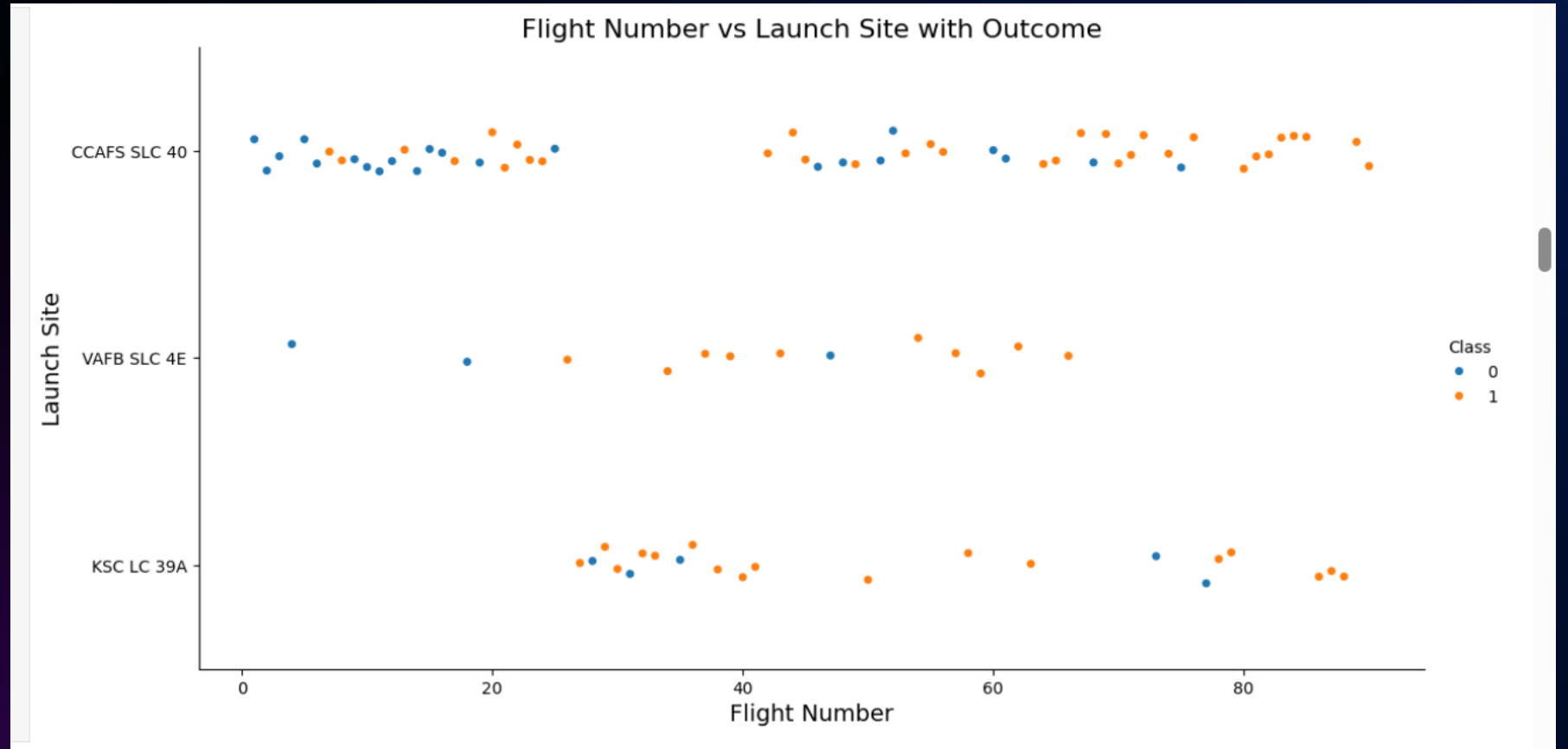
EDA WITH DATA VISUALIZATION

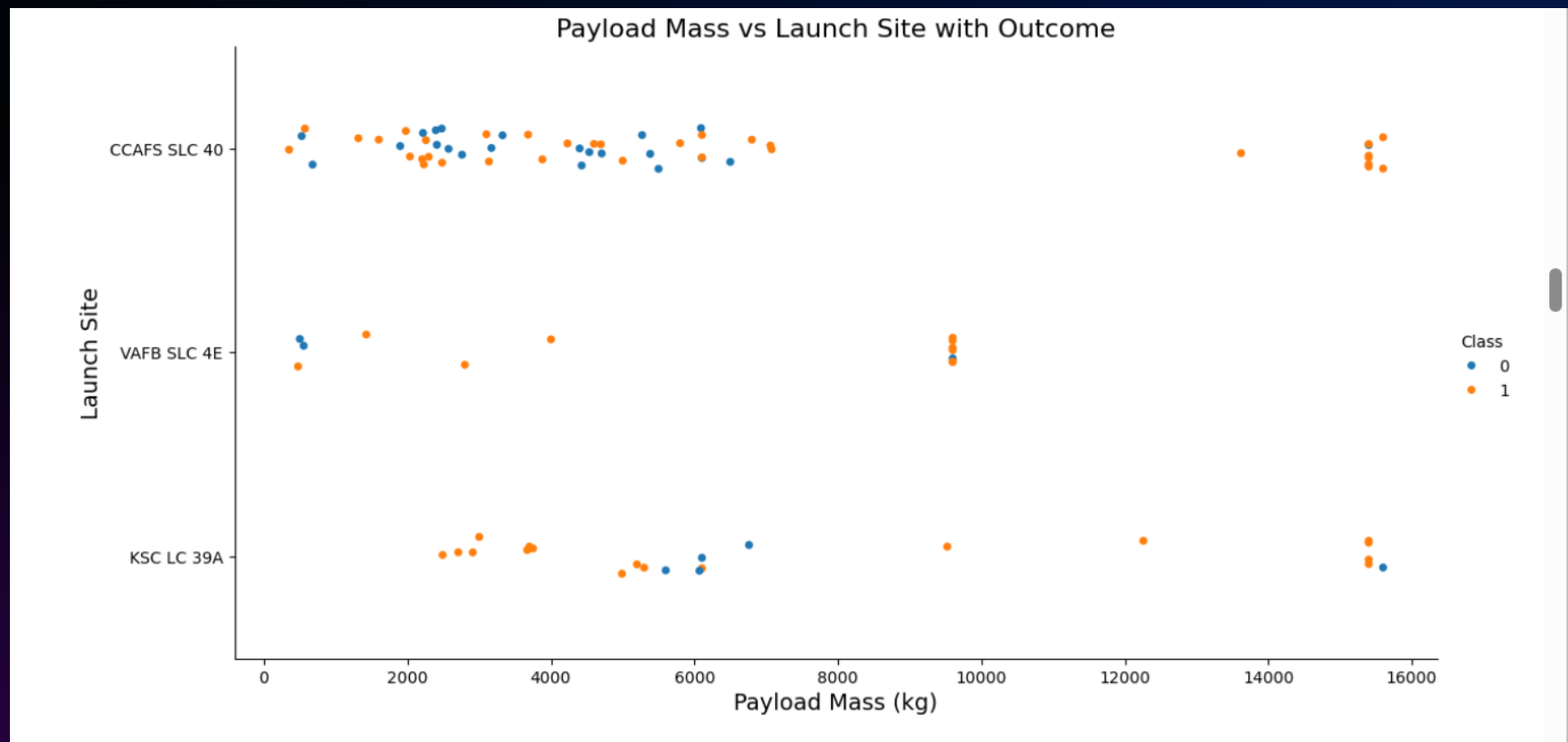
- Bar Chart (Launch Sites) Shows number of launches per site – highlights most used sites. Pie Chart (Outcomes) Displays success vs. failure rate – quick success overview. EDA with Data Visualization GitHub File Link .Scatter Plot (Payload vs Outcome) Examines impact of payload mass on launch success. Box Plot (Payload by Orbit) Compares payload mass ranges across orbit types. Line Plot (Launches Over Time) Tracks launch frequency trends. Bar Chart (Orbits) Identifies most frequently targeted orbits.
- <https://labs.cognitiveclass.ai/v2/tools/jupyterlite?ulid=ulid-1d4dead7d6a986c25d9bfef7047747f0b0e671e0>

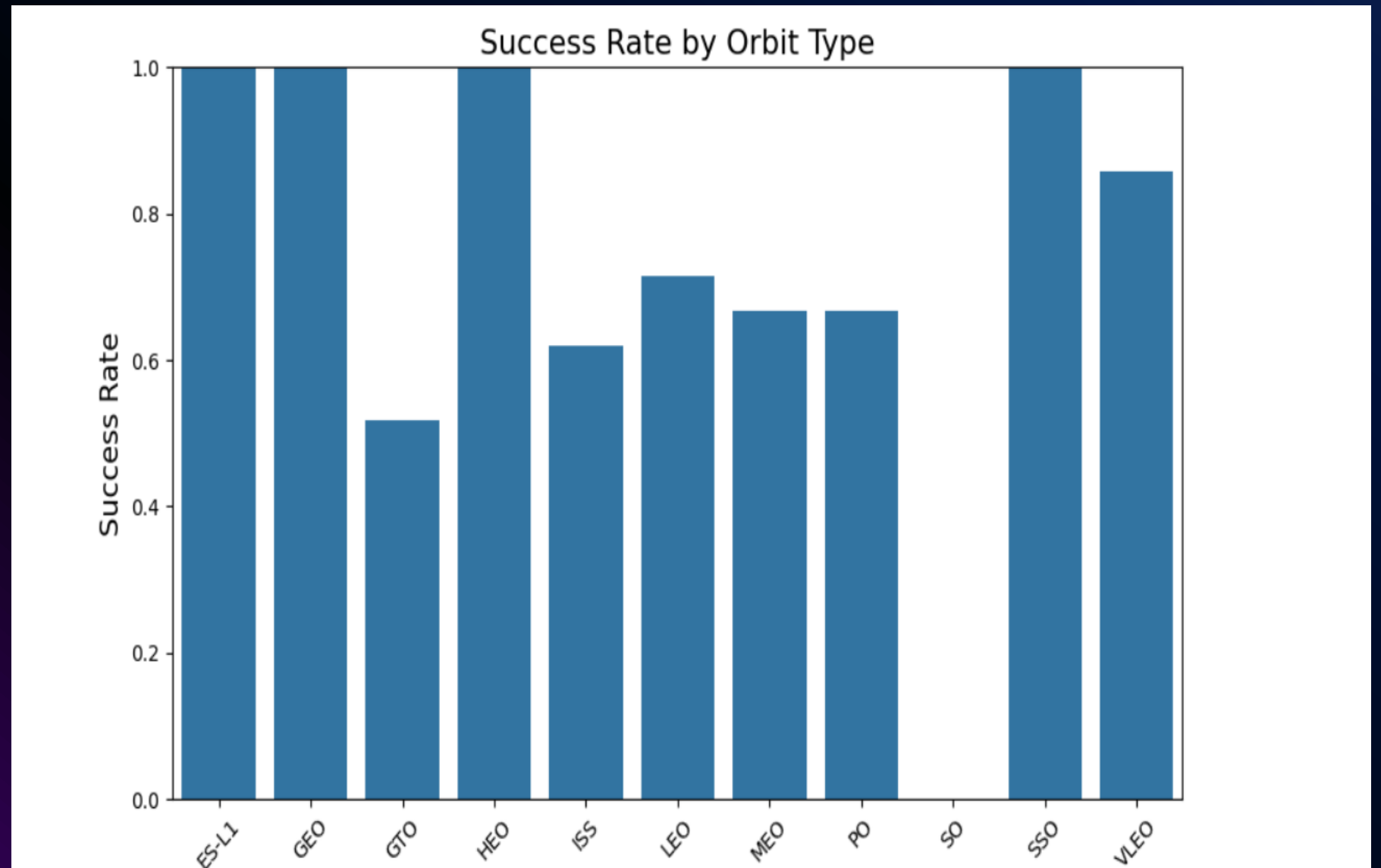
We can plot out the `FlightNumber` vs. `PayloadMass` and overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.

```
[5]: sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



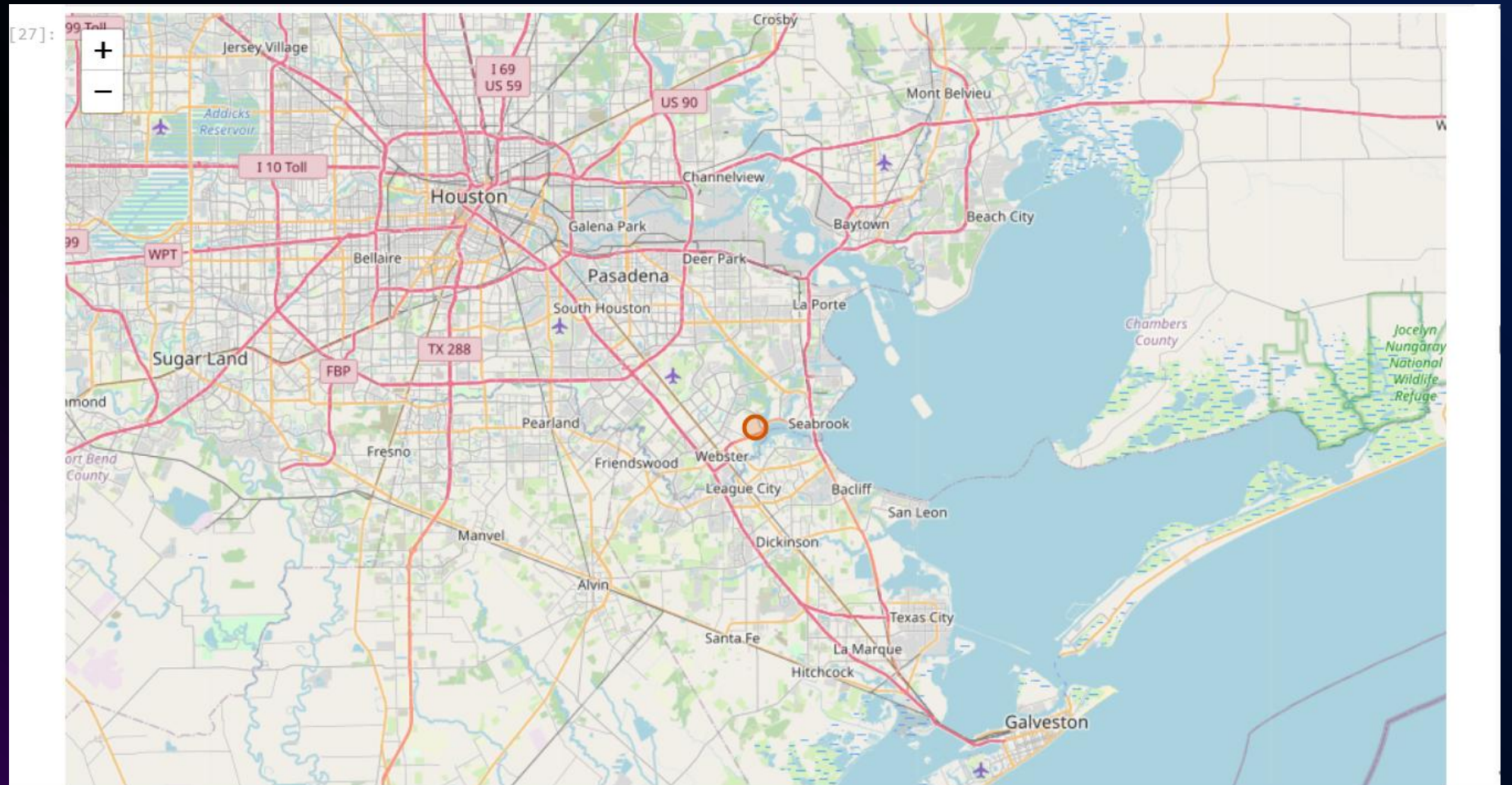




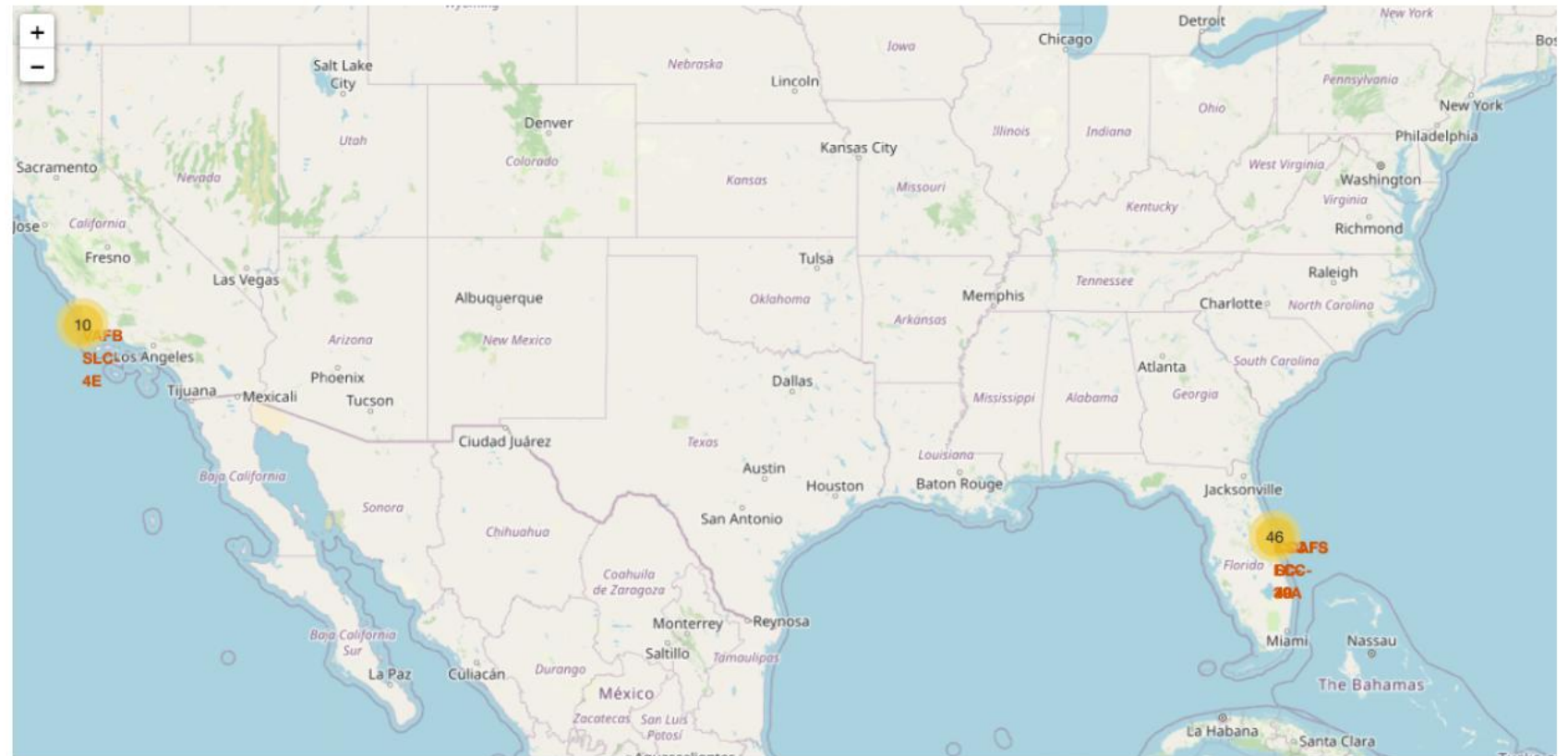


BUILD INTERACTIVE MAP WITH FOLIUM

- Folium Map Objects and Purpose Build an Interactive Map with Folium GitHub File Link
- Marker To identify and label launch site locations on the map.
- Circle To visually represent proximity range and analyze nearby facilities.
- Marker with Popup To highlight infrastructure relevant to launch operations. Line (PolyLine) To visualize distance and spatial relationship between launch sites and key infrastructure.
- <https://labs.cognitiveclass.ai/v2/tools/jupyterlite?ulid=ulid-f6ab70283e72c16d2ae5741ccd08bc625cf53bda>



Your updated map may look like the following screenshots:





From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates.

BUILD DASHBOARD WITH PLOTLY DASH

- The Plotly Dash dashboard offers interactive visualizations of space mission data
Pie Chart: Displays the total successful launches per site, with a dropdown to filter success/failure by site. Scatter Plot: Shows the correlation between launch outcomes and payload mass, with filters for payload range and booster version. Launch Site Dropdown: Allows selection of a site, adjusting the pie chart to show success/failure counts. Payload Mass Slider: Enables filtering of payload mass range for detailed analysis. These features provide a dynamic, user-friendly way to explore the data and analyze launch outcomes.

Total Success Launches by Site



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches.

Launch Success Count for All Sites

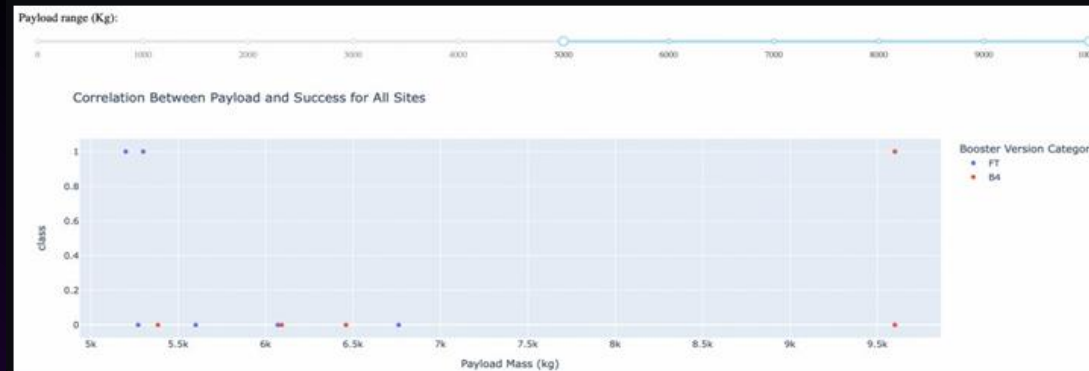
Total Success Launches for Site KSC LC-39A



The launch site KSC LC-39 A had the highest rate of successful launches, with a 76.9% success rate.

LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

Payload Mass vs. Launch Outcome for All Sites

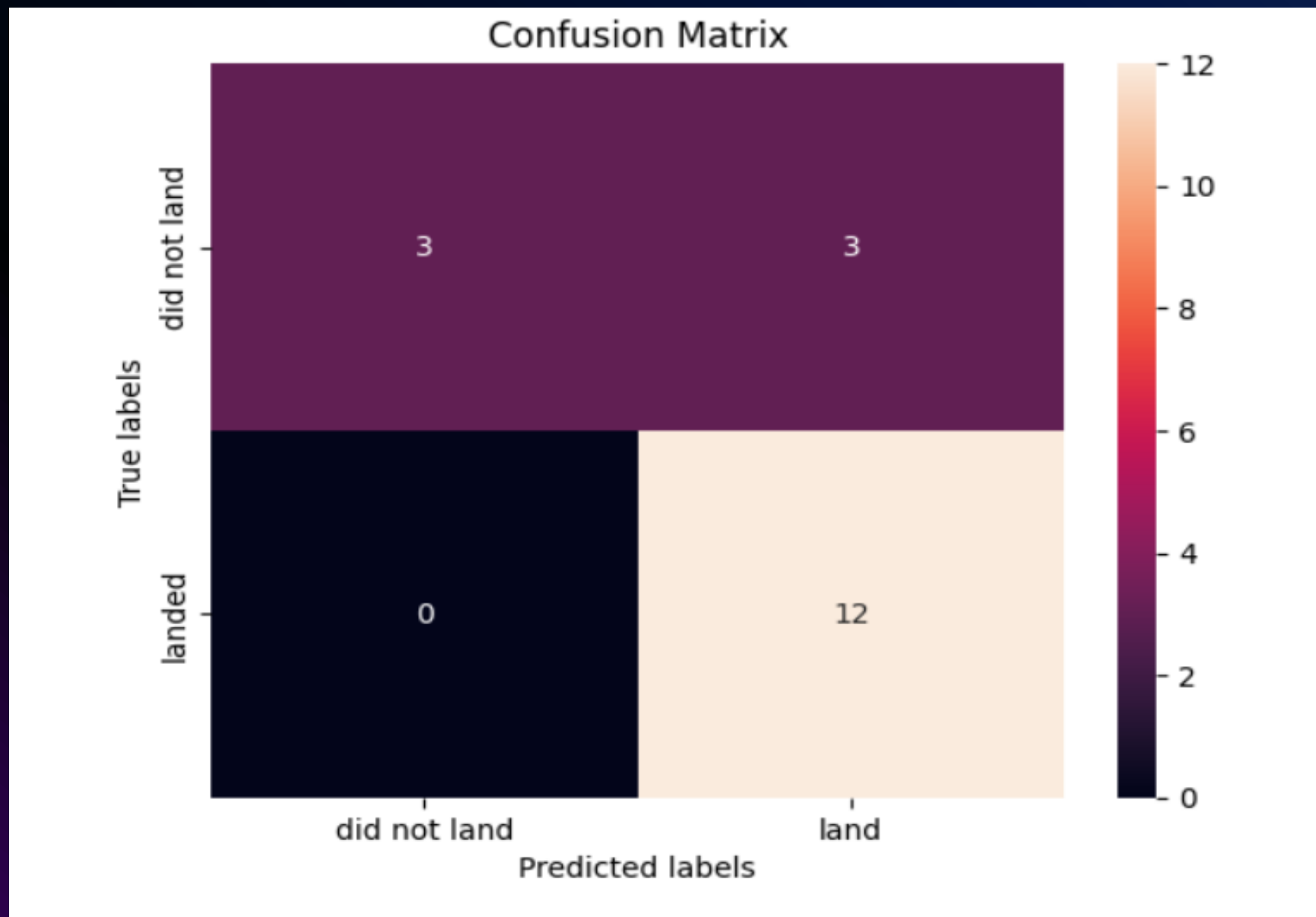


From these 2 plots, it can be shown that the success for massive payloads is lower than that for low payloads.



CONFUSION MATRIX

- As shown previously, best performing classification model is the Support Vector Machine Model , with an accuracy of 83%.
- This is explained by the confusion matrix, which shows only 3 out of 18 total results classified incorrectly (false positives, shown in the top-right corner).
- The other 15 results are correctly classified (3 did not land, 12 did land)



CONCLUSION

As the number of launches increased over time, so did the success rate at each launch site—early missions were more prone to failure, highlighting how experience improves reliability.

- From 2010 to 2013, there were no successful landings, resulting in a 0% success rate during that period. However, post-2013, the success rate showed a steady rise, with minor setbacks occurring in 2018 and 2020. Since 2016, the likelihood of a successful landing has consistently stayed above 50%.
- Certain orbit types—ES-L1, GEO, HEO, and SSO—demonstrated a 100% success rate. While GEO, HEO, and ES-L1 each had just one launch (limiting the strength of that conclusion), SSO stood out with five successful missions, making its performance more meaningful.

- Orbit categories such as PO, ISS, and LEO tended to show higher success rates with heavier payloads. Launches to VLEO were also associated with large payloads, which aligns with expectations for that orbit type.

KSC LC-39A proved to be the most successful launch site, contributing 41.7% of all successful missions, and achieving a 76.9% success rate overall.

In contrast, missions involving very large payloads (above 4000 kg) were generally less successful compared to those carrying lighter loads.

Among the classification models tested, the Support Vector Machine (SVM) model performed best, reaching an accuracy of 83%

THANK YOU