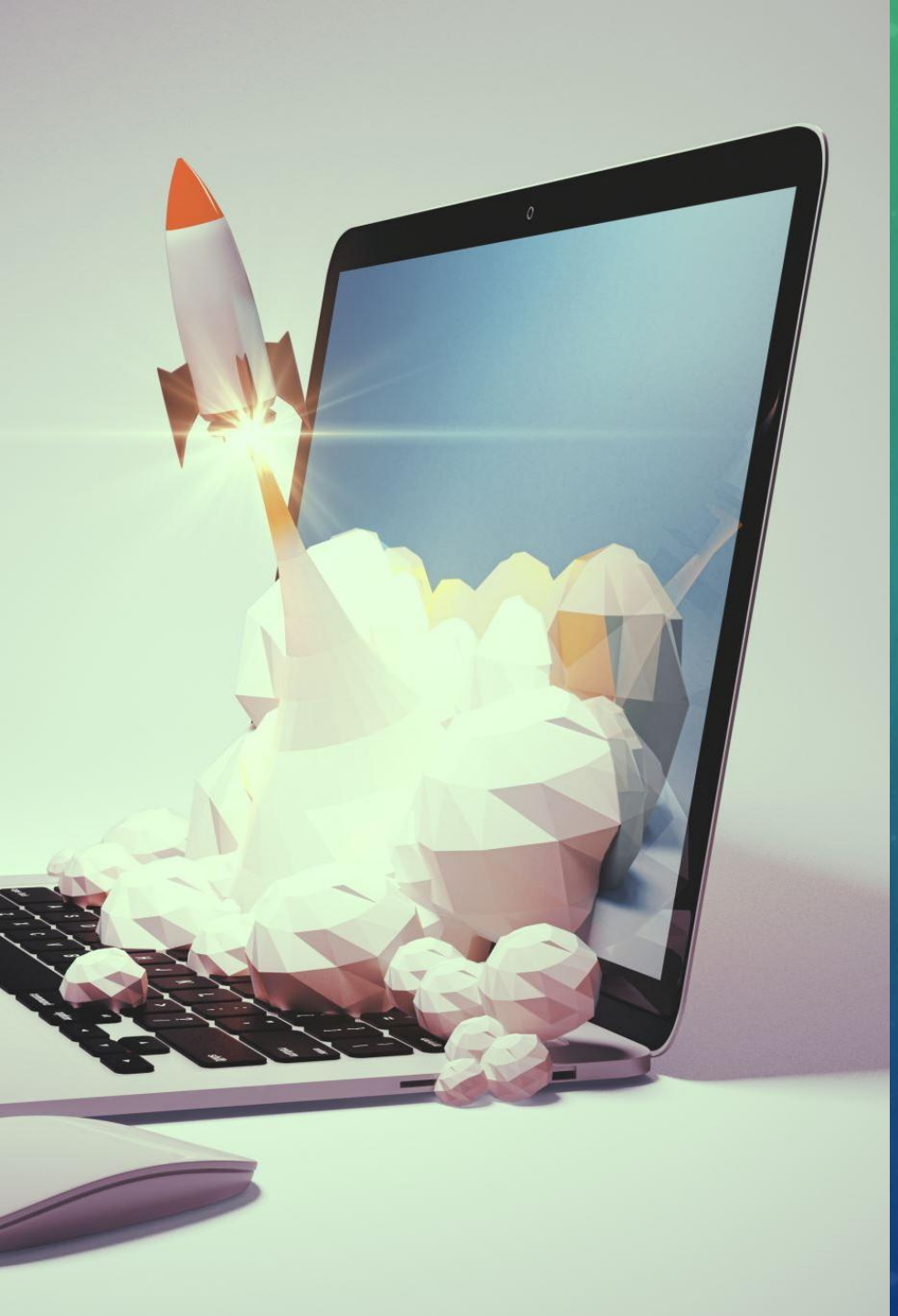
Space Y – Capstone Project

Franco Fernando
Padilla Chávez



EXECUTIVE SUMMARY

•This research aims to identify key factors contributing to a successful rocket landing. To achieve this, the following methodologies were employed:

- Data was collected using the SpaceX REST API and web scraping techniques.
- The data was cleaned and transformed to generate a binary success/failure outcome variable.
- Exploratory data analysis was performed using data visualization tools, focusing on variables such as payload, launch site, flight number, and annual trends.
- SQL was used to analyze the data, including calculations of total payload, payload ranges for successful launches, and the number of successful versus failed outcomes.
- Success rates at different launch sites were evaluated, along with their proximity to geographical landmarks.
- Predictive models—including logistic regression, support vector machines (SVM), decision trees, and K-nearest neighbors (KNN)—were built to forecast landing outcomes.

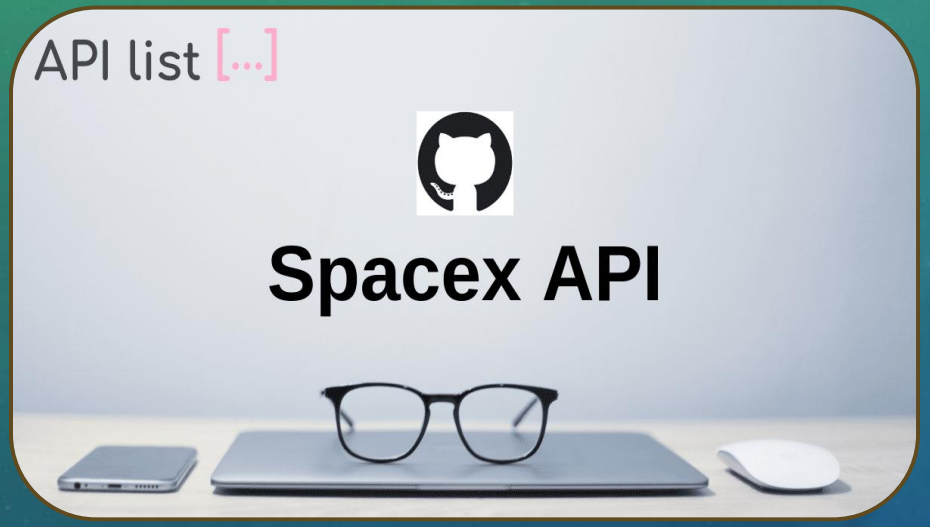
INTRODUCTION

•SpaceY, a pioneering company in the aerospace sector, aims to make space travel more accessible and affordable. Its major achievements include transporting spacecraft to the International Space Station, launching a satellite network to deliver global internet access, and conducting manned space missions. One of the key reasons SpaceY can keep launch costs relatively low—around $62 million per launch—is its innovative ability to reuse the first stage of its Falcon 9 rocket. In contrast, other providers that lack reusable technology face launch costs exceeding $165 million. By predicting whether the first stage will successfully land, we can estimate the potential cost of a launch. This can be accomplished using publicly available data combined with machine learning models to forecast stage recovery for SpaceY or its competitors.

METHODOLOGY

- Collect data through the SpaceX REST API and web scraping methods.
- Wrangle the data by filtering, handling missing values, and applying one-hot encoding to prepare it for analysis and modeling.
- Explore the dataset using Exploratory Data Analysis (EDA), SQL queries, and data visualization techniques.
- Visualize insights with tools such as Folium for mapping and Plotly Dash for interactive dashboards.
- Build classification models to predict landing outcomes, then tune and evaluate them to identify the most accurate model and optimal parameters.

DATA COLLECTION



We gathered data from two main sources:

- **SpaceX API** – An open-source REST API offering structured data on launches, rockets, cores, capsules, Starlink missions, launchpads, and landing sites.
- **Wikipedia** – A publicly accessible online encyclopedia, used to extract additional information through web scraping.

DATA WRANGLING

In this stage, we utilized Pandas and NumPy to load and prepare the collected data for analysis. The main objective was to clean the dataset and identify relevant features suitable for training machine learning models. This process involved addressing missing values, resolving formatting inconsistencies, and ensuring the data was properly structured for effective exploratory analysis and modeling.

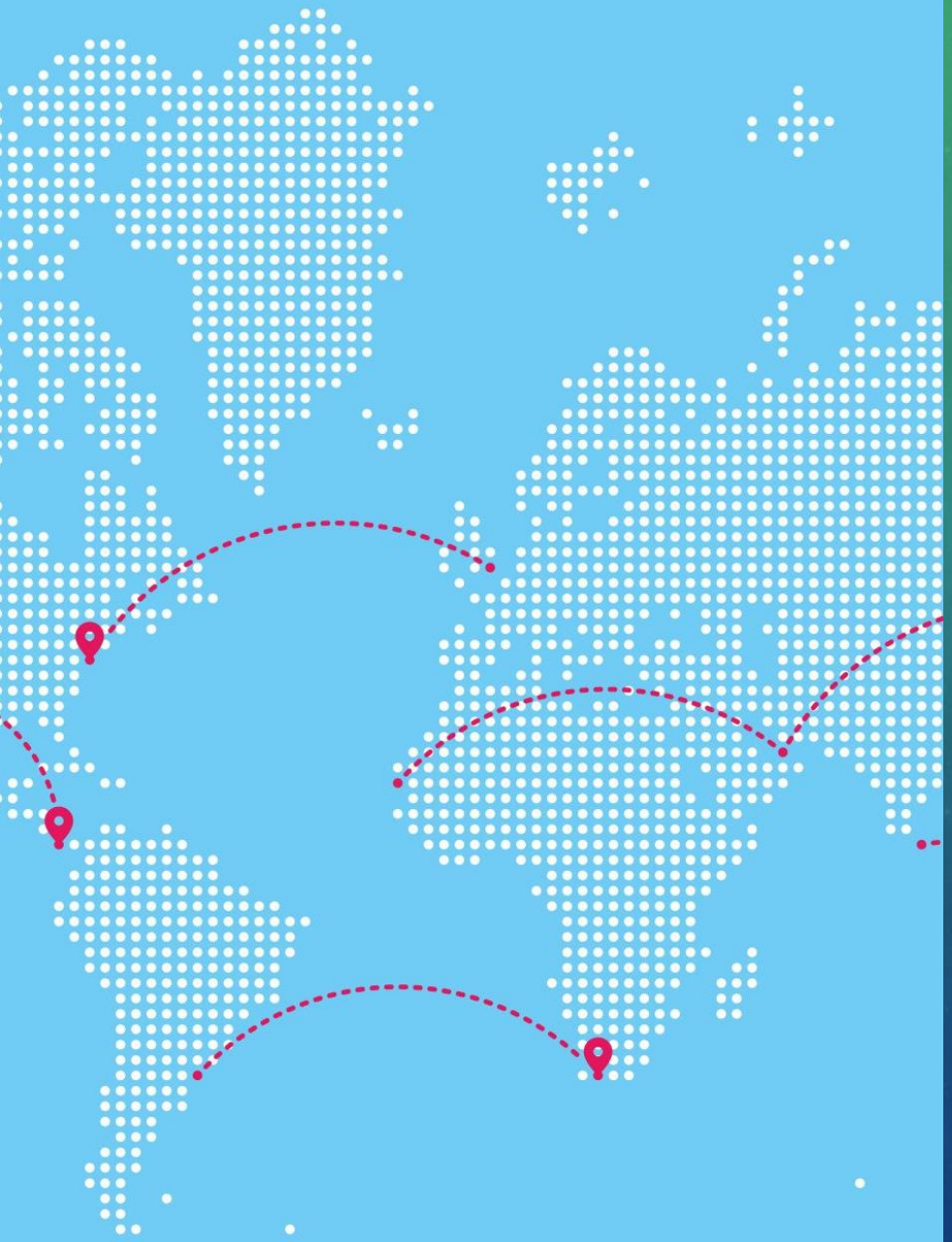
EDA WITH VISUALIZATION

At this stage, we conducted exploratory data analysis (EDA) to examine the relationships between various features and the target variable. Leveraging Seaborn and Matplotlib, we visualized correlations and uncovered important patterns within the dataset. We also applied feature engineering techniques, including the conversion of categorical variables into dummy variables, to ensure the data was properly structured for model development.

EDA WITH SQL

In this phase, we utilized SQL queries—examples of which are shown on the right—to support our exploratory data analysis (EDA) on the collected dataset. These queries allowed us to extract specific insights, including:

- Retrieving all unique launch site names used in space missions.
- Displaying five records of launch sites with names starting with the prefix 'CCA'.
- Calculating the total payload mass carried by boosters in NASA (CRS) missions.
- Identifying the date of the first successful landing on a ground pad.
- Listing the names of boosters that successfully landed on a drone ship and carried payloads between 4000 kg and 6000 kg.



MAP WITH FOLIUM

•We used the **Folium** library to visualize geospatial data through interactive maps. Specifically, we mapped four SpaceX launch sites using circle markers to indicate their geographic coordinates:

- **CCAFS LC 40**: (28.562302, -80.577356)
- **CCAFS SLC 40**: (28.563197, -80.576820)
- **KSC LC 39A**: (28.573255, -80.646895)
- **VAFB SLC 4E**: (34.632834, -120.610746)

•Circle markers were added to represent Falcon 9 rocket launch locations, and custom markers were used to display the outcomes of first-stage landings (success or failure).
We also calculated and visualized the distances from **CCAFS LC 40** to the nearest city, coastline, and highway, using **PolyLine** objects to draw these connections on the map.

DASHBOARD WITH PLOTLY DASH

Interactive Dashboard Features

- **Dropdown Menu** – Enabled users to select a specific launch site from the following options: All Sites, CCAFS LC 40, CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A.
- **Pie Chart** – Displayed the total number of successful launches for each selected launch site.
- **Payload Slider** – Added a slider to filter launches based on payload mass, with a range from 0 to 10,000 kg.
- **Scatter Plot** – Visualized the relationship between payload mass and launch success to highlight possible correlations.

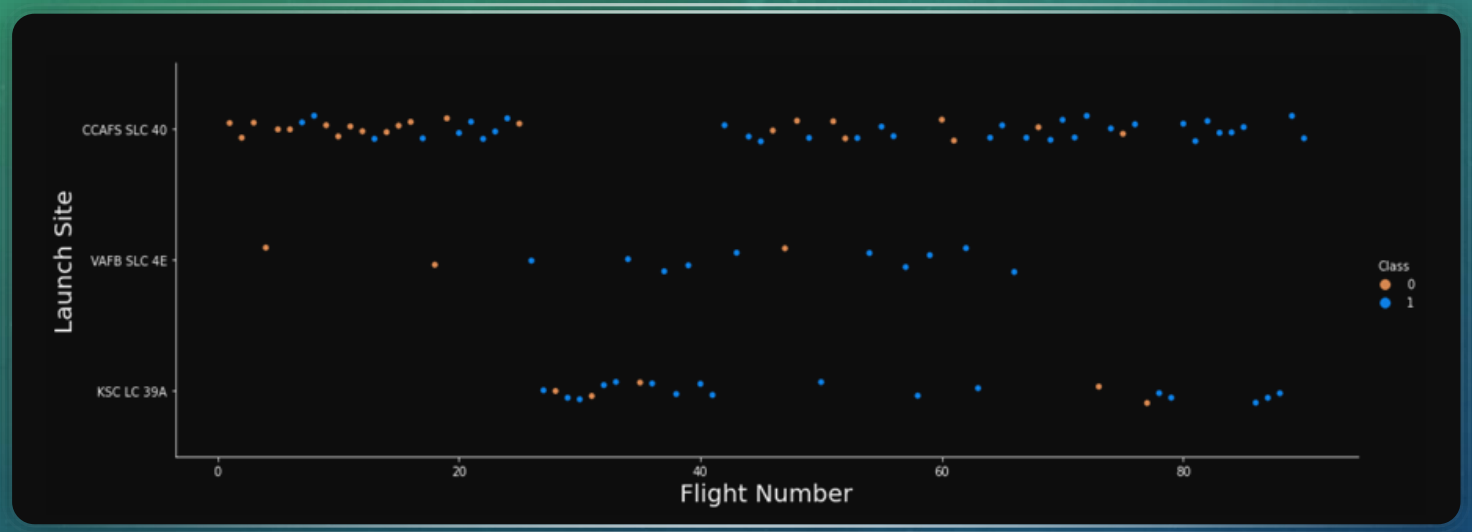
PREDICTIVE ANALYTICS

- Machine Learning Pipeline
- Imported all necessary libraries required for model training and evaluation.
- Loaded the cleaned and preprocessed dataset.
- Standardized feature values to ensure consistent scaling and prevent model bias.
- Split the data into training (80%) and testing (20%) sets.
- Initialized four classification algorithms:
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree (DT)
  - K-Nearest Neighbors (KNN)
- Applied Grid Search for hyperparameter tuning to identify the optimal settings for each model.
- Evaluated model performance using the following metrics:
  - Confusion Matrix
  - F1 Score
  - Jaccard Index

These evaluation metrics helped determine the most effective model for potential deployment.

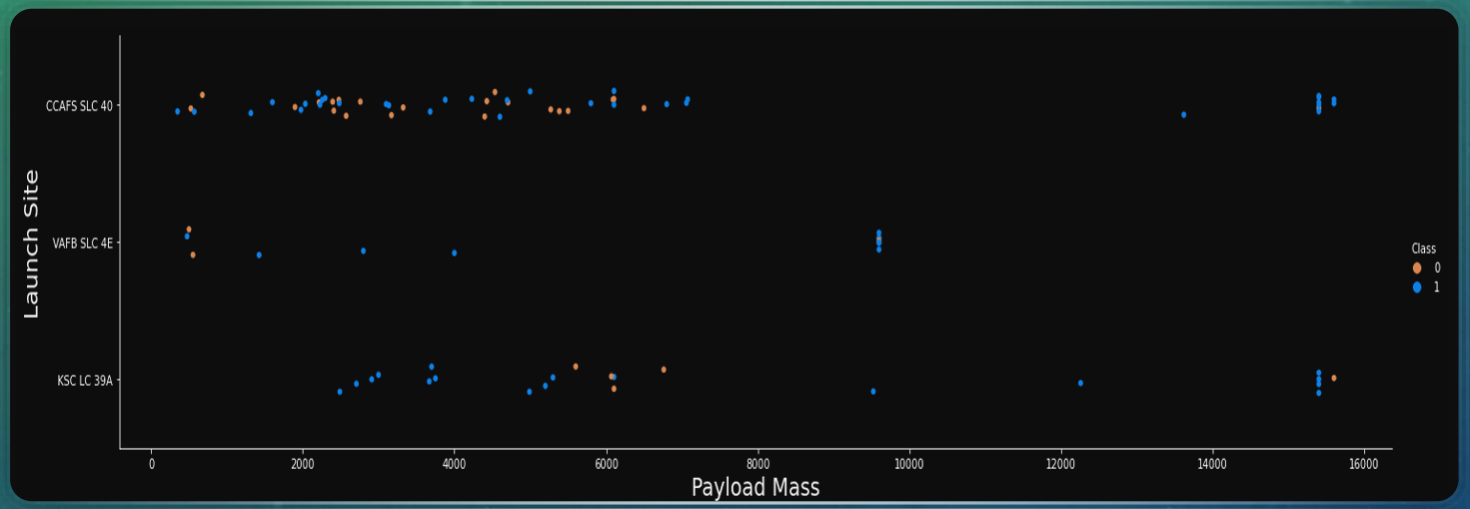
RESULTS

FLIGHT NUMBER VS. LAUNCH SITE

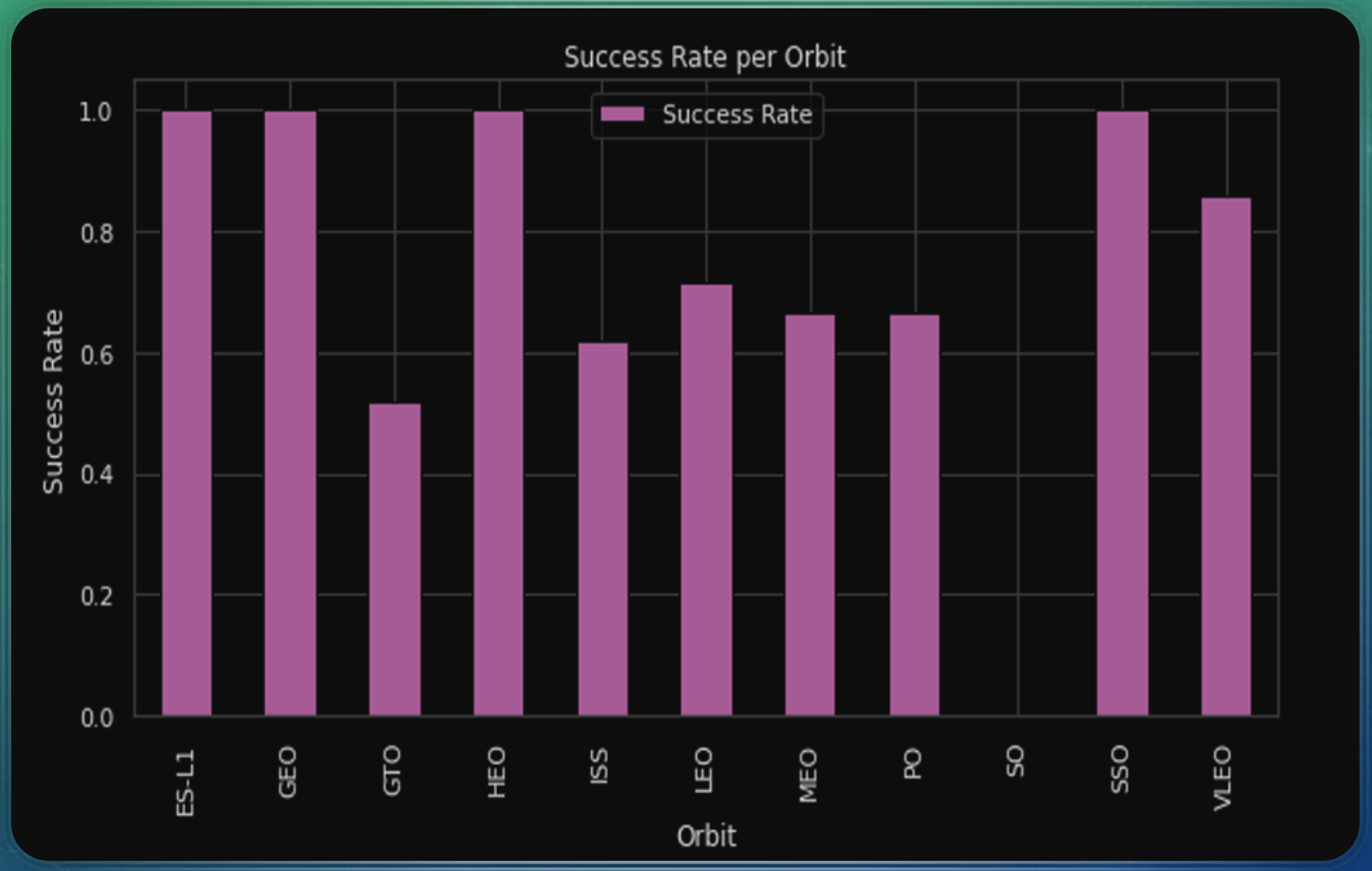


- Early flights showed a lower success rate, with more failures (orange markers).
- More recent flights exhibited a higher success rate, indicated by more successes (blue markers).
- Approximately half of all launches took place at the CCAFS SLC 40 launch site.
- Launches from VAFB SLC 4E and KSC LC 39A demonstrated higher success rates compared to other sites.Overall, the data suggests that newer launches are more likely to succeed.

PAYLOAD VS. LAUNCH SITE

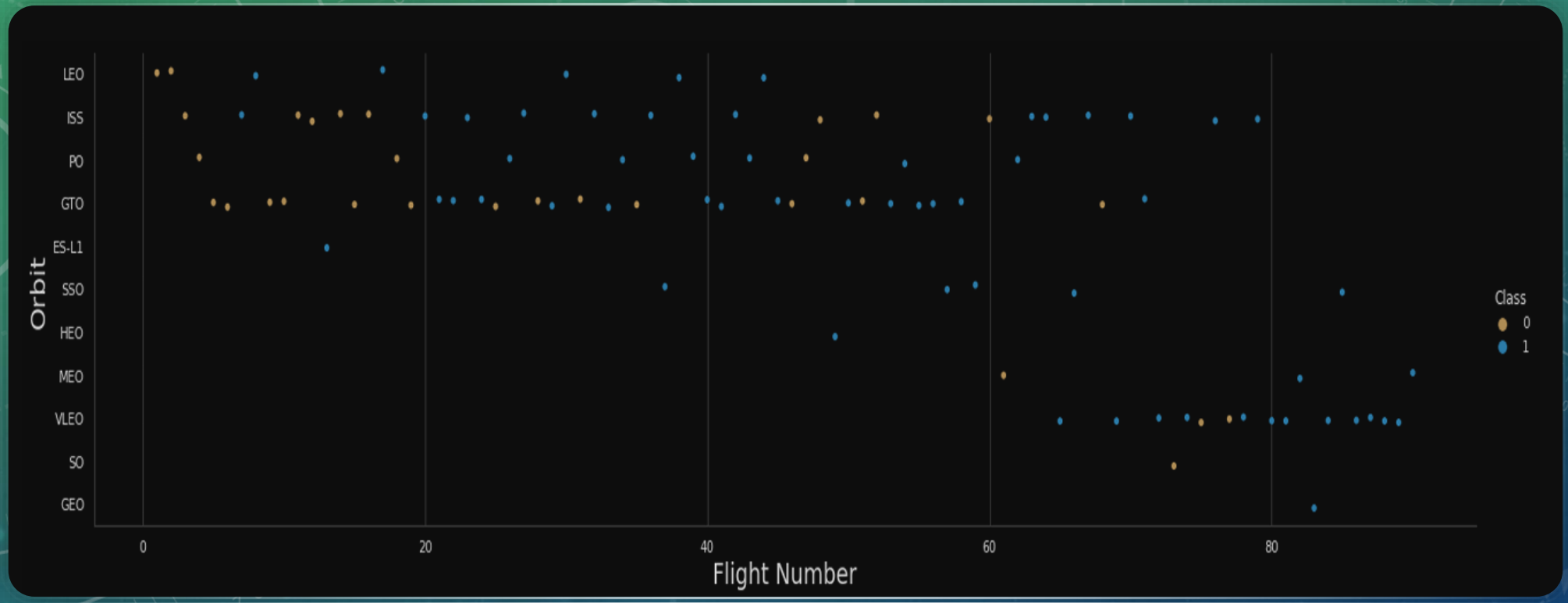


- In general, higher payload masses (kg) are associated with a higher success rate.
- Most launches exceeding 7,000 kg in payload were successful.
- The KSC LC 39A launch site showed a 100% success rate for payloads under 5,500 kg.
- The VAFB SLC 4E site has not conducted any launches with payloads above approximately 10,000 kg.

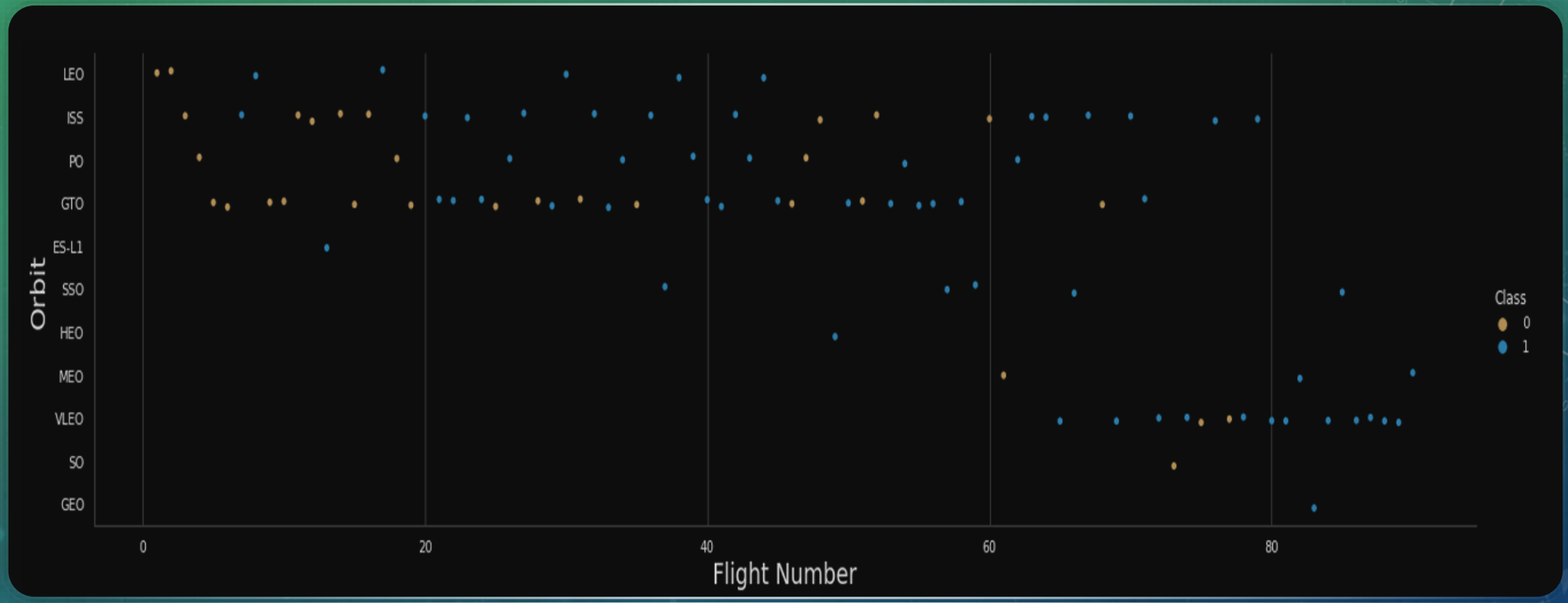


SUCCESS RATE BY ORBIT

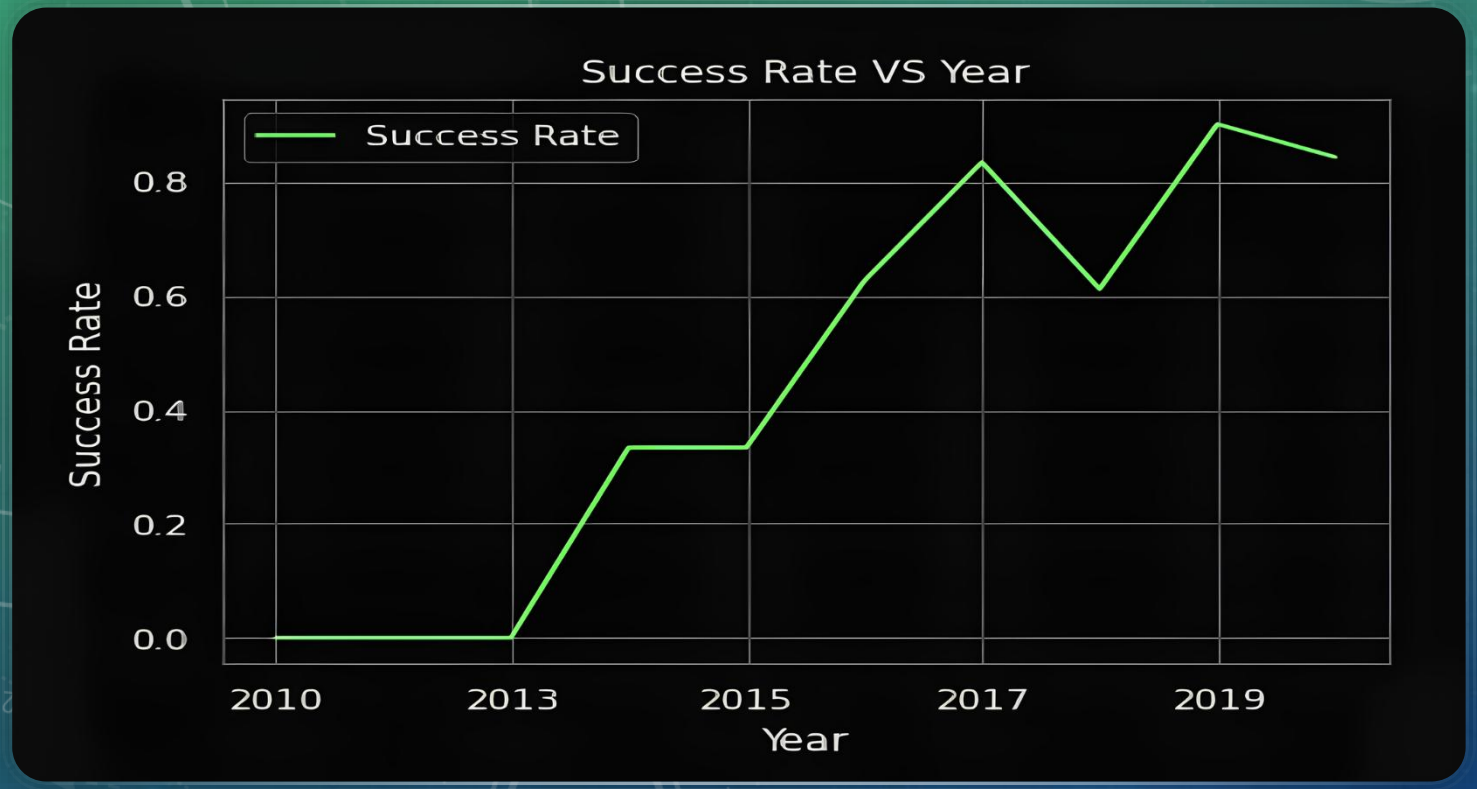
- **100% Success Rate**:
  - ES-L1 (Earth–Sun Lagrange Point 1)
  - GEO (Geostationary Orbit)
  - HEO (Highly Elliptical Orbit)
  - SSO (Sun-Synchronous Orbit)
- **Moderate Success Rate (50%–80%)**:
  - GTO (Geostationary Transfer Orbit)
  - ISS (International Space Station)
  - LEO (Low Earth Orbit)
  - MEO (Medium Earth Orbit)
  - PO (Polar Orbit)
- **0% Success Rate**:
  - SO (Solar Orbit)



FLIGHT NUMBER VS. ORBIT



PAYLOAD VS. ORBIT TYPE



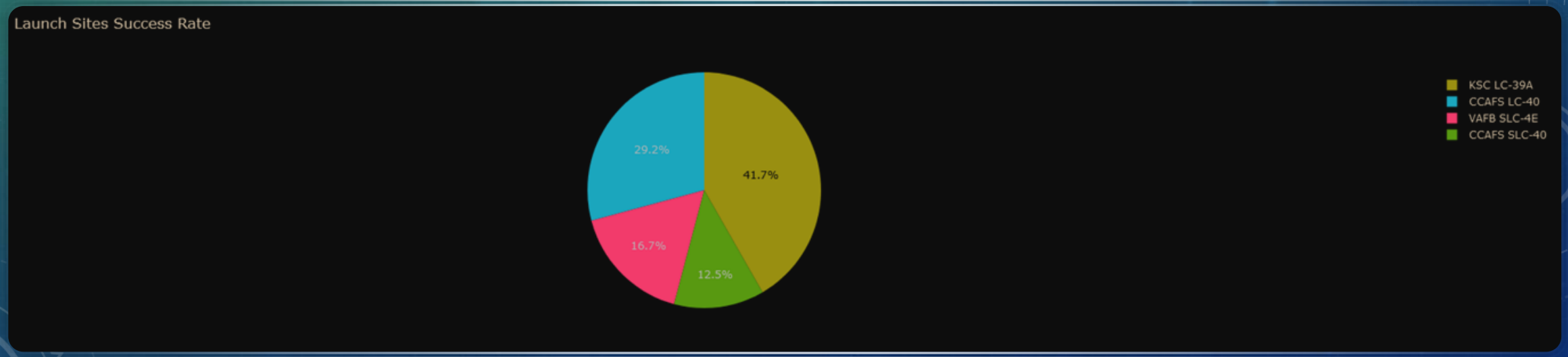
LAUNCH SUCCESS OVER TIME


DASHBOARD WITH PLOTLY

DASHBOARD WITH PLOTLY

•Success as a Percentage of Total Launches

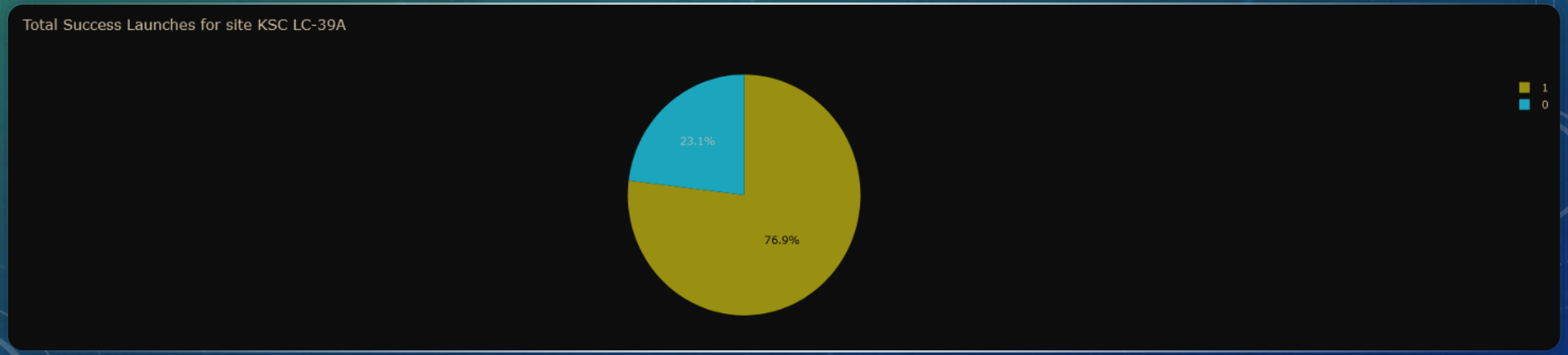
- KSC LC-39A accounts for the highest proportion of successful launches among all launch sites, contributing 41.2% of total successes.



LAUNCH SUCCESS (KSC LC-29A)

- **Success as Percent of Total**

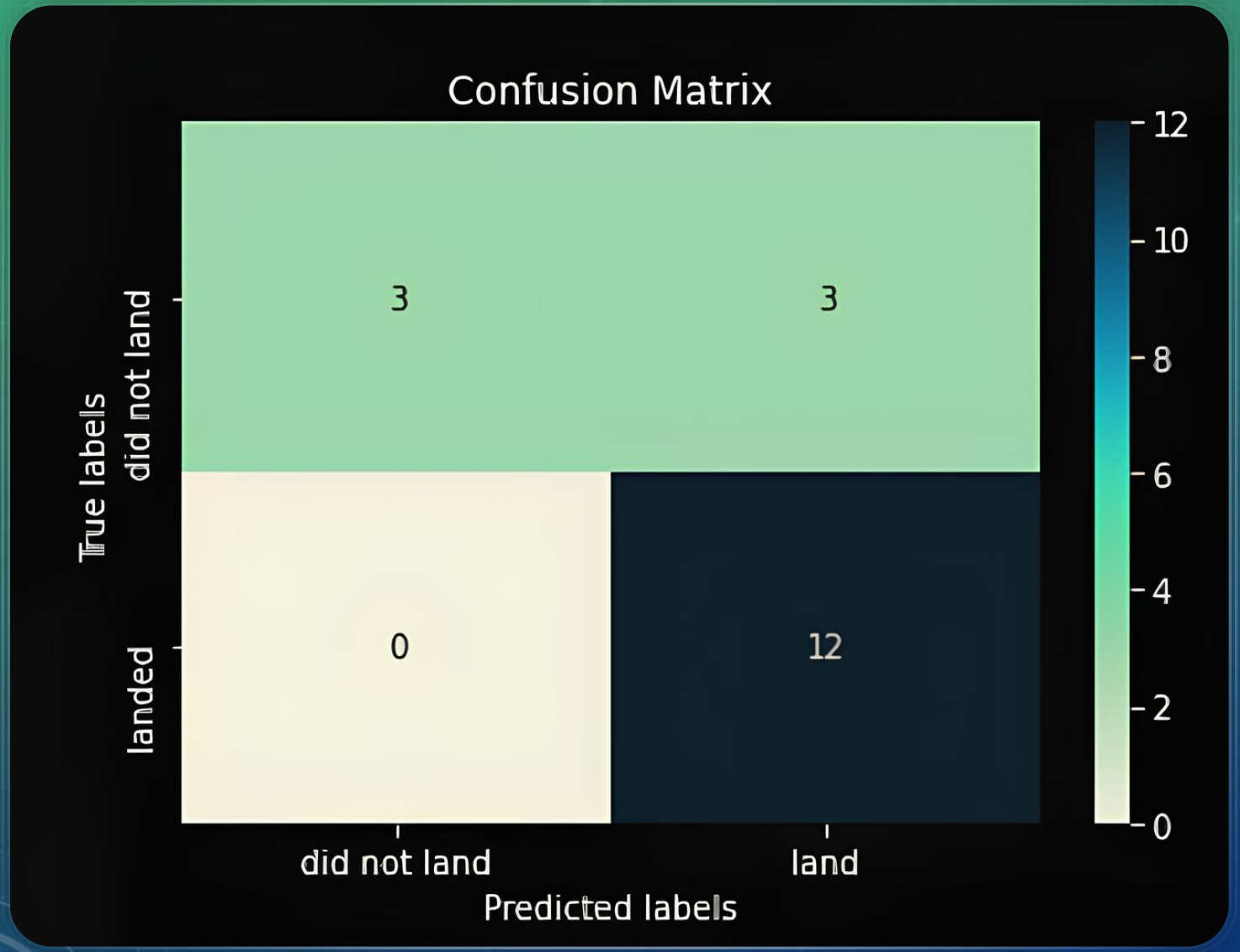
- At **KSC LC-39A**, **76.9%** of missions were **successful**, while **23.1%** resulted in **failure**.



DASHBOARD

•This interactive scatter plot illustrates the relationship between payload mass and launch outcomes. The visualization suggests that boosters carrying payloads under 4,000 kg tend to have a higher success rate. This indicates that lighter payloads may be associated with more reliable launch outcomes, regardless of the booster version used.

PREDICTIVE ANALYSIS



CLASSIFICATION

•A confusion matrix provides a summary of a classification model's performance by showing the number of correct and incorrect predictions. In this case, all models produced identical confusion matrices, which revealed the following results:

- 12 True Positives
- 3 True Negatives
- 3 False Positives (Type I Errors)
- 0 False Negatives

•The presence of false positives indicates that some launches were incorrectly predicted as successful, which is a concern when accuracy is critical—especially in high-stakes applications like rocket landings.

CONCLUSION

- **KSC LC-39A** is the most reliable launch site, with the **highest success rate**, including **100% success** for payloads under 5,500 kg.
- **Launch success has improved over time**, with newer missions showing higher reliability.
- **Higher payload mass** is generally correlated with **increased success rates**, especially for payloads above 4,000 kg.
- Launch sites are **strategically located near the equator and coastlines** to maximize fuel efficiency and safety.
- Orbits like **ES-L1, GEO, HEO, and SSO** have **100% success**, while **GTO and SO** show lower reliability.
- **Interactive visualizations** (Folium, Plotly Dash) enhanced exploration of spatial and payload-related trends.
- Among all models, the **decision tree slightly outperformed others**, though all showed similar results with some **false positives**.
- A **data-driven approach** combining EDA, geospatial mapping, and machine learning provides **strategic insights** for improving launch outcomes and supporting **competitive growth** in the commercial space industry.


THANK
YOU