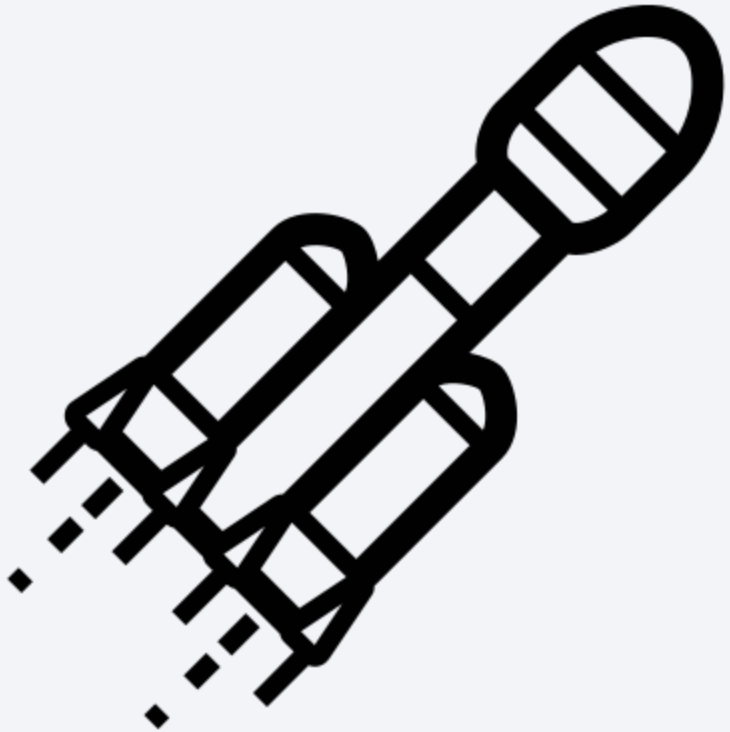
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dipankar Purkayastha
27th September, 2025

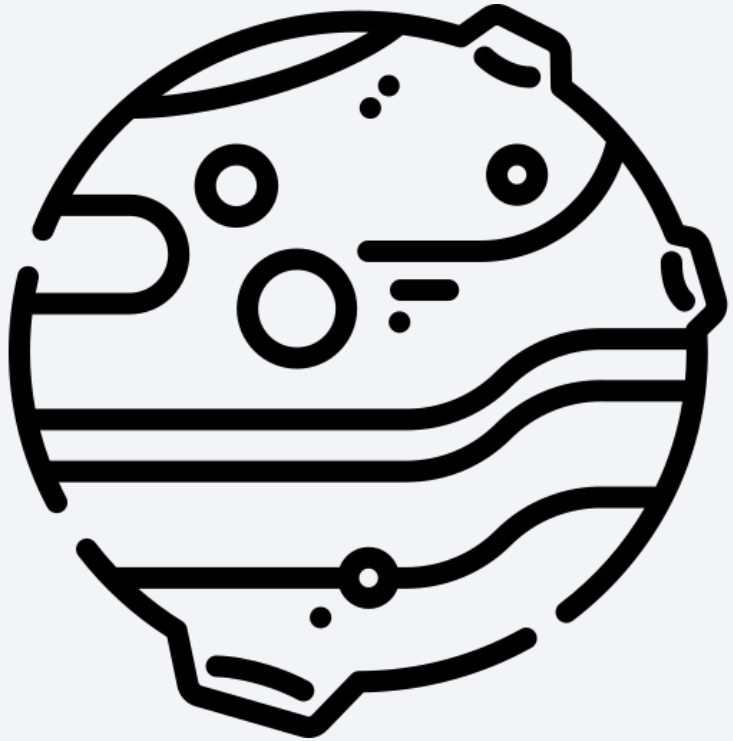
Outline

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



Executive Summary

- **Summary of methodologies**
    1. Data collection
    2. Data wrangling
    3. Exploratory Data Analysis with Data Visualization
    4. Exploratory Data Analysis with SQL
    5. Building an interactive map with Folium
    6. Building a Dashboard with Plotly Dash
    7. Predictive analysis (Classification)
- **Summary of all results**
    1. Exploratory Data Analysis Results
    2. Interactive Analysis demo in screenshots
    3. Predictive analysis results



Introduction

- **Project background and context**

  SpaceX is the most successful company of the commercial space age, making space travel affordable. The company advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- **Problems you want to find answers**

  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years? - What is the best algorithm that can be used for binary classification in this case?

Introduction

- **Project Objective**

  The project objective is to evaluate the viability of the new company SpaceY to make prediction on Space X by training and developing models.

- **Desirable answers**

  - The best way to estimate the total cost for launches, by predicting successful landing of the first stage of rockets
  - Where is the best place to make launches.

Section 1

Methodology

Methodology

- **Data collection methodology:**
  - Using SpaceX Reset API
  - Using Web Scrapping from Wikipedia
- **Perform data wrangling**
  - Filtering the data
  - Dealing with missing values
  - Using one hot encoding to prepare the data to a binary classification
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - How to build, tune, evaluate classification models

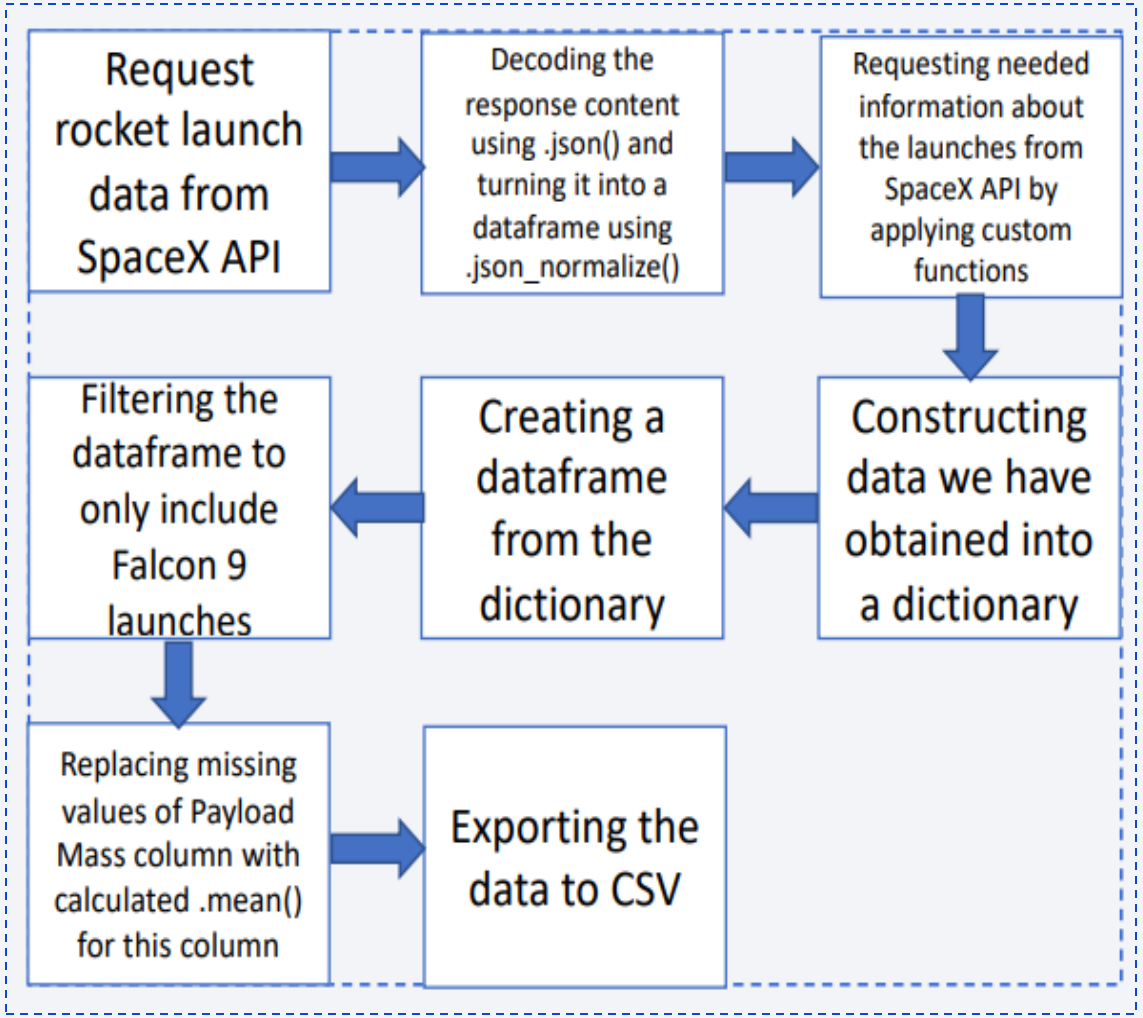
Data Collection

- Used **two methods**:
- **SpaceX REST API** → technical details (FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, etc.)
- **Wikipedia Web Scraping** → additional info (Flight No., Payload, Customer, Launch outcome, Booster landing, Date & Time, etc.)
- Combined both sources for **complete launch data** and detailed analysis.

Data Collection – SpaceX API

- Data collection using SpaceX REST calls using key phrases and flowcharts

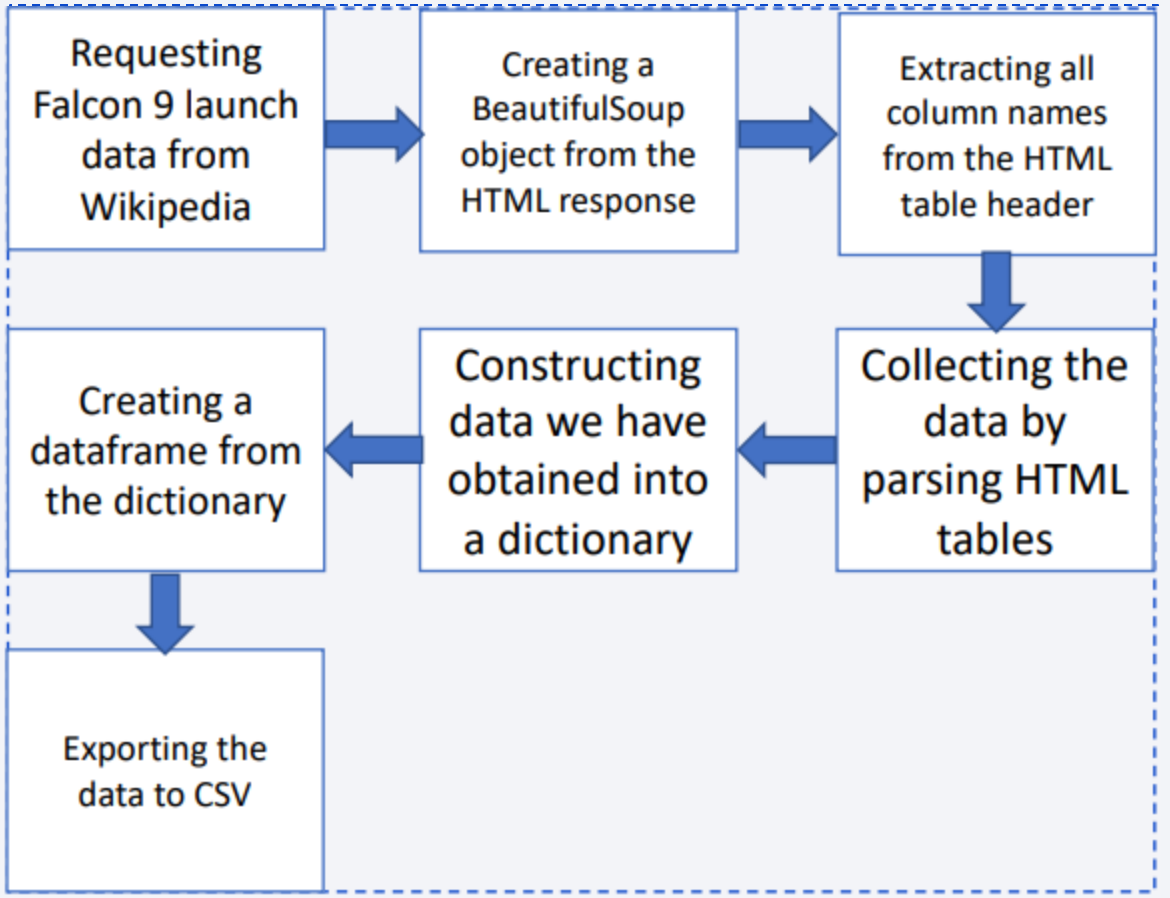
- GitHub URL of the completed SpaceX API: https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/blob/main/01%20Data%20Collection%20and%20Wrangling/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

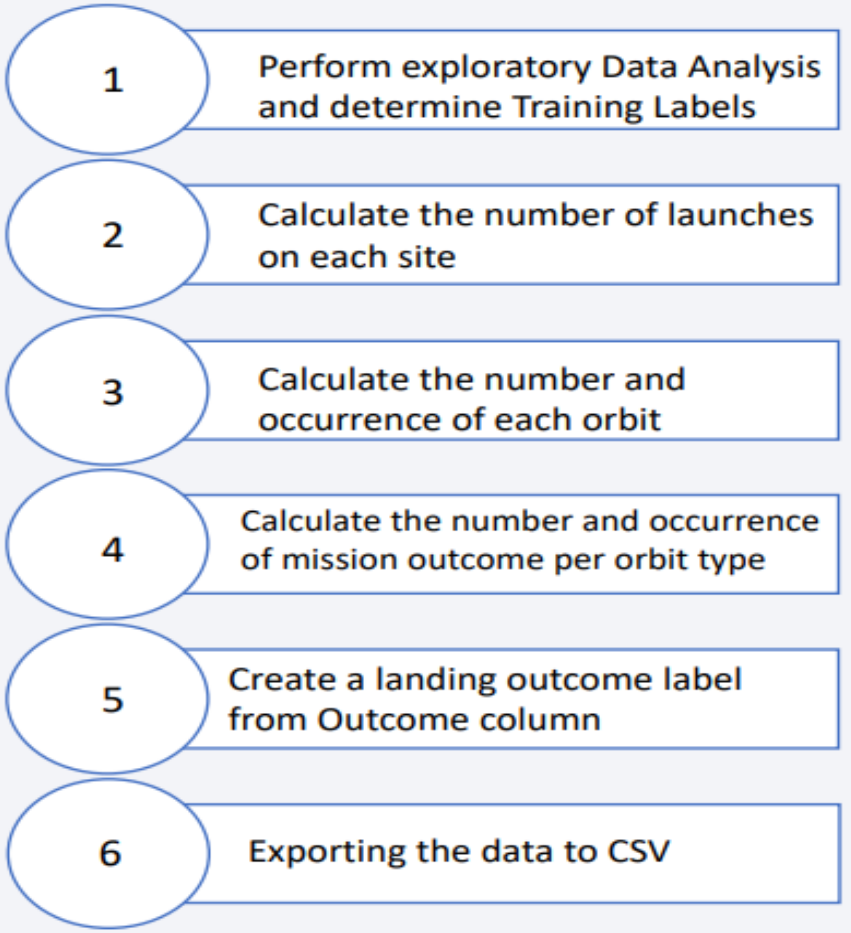
- Data collection was used web scraping process using key phrases and flowcharts.

- GitHub URL of the completed SpaceX API: https://github.com/aimless-coder/SpaceY_IBM_Applied_Data_Science_Project/blob/main/01%20Data%20Collection%20and%20Wrangling/jupyter-labs-webscraping.ipynb



Data Wrangling

- **Booster Landing Outcomes**
- **Ocean**: True → successful ocean landing, False → failed ocean landing
- **RTLS**: True → successful ground pad landing, False → failed ground pad landing
- **ASDS**: True → successful drone ship landing, False → failed drone ship landing
- **Training Labels**:
  - 1 → **Successful landing**
  - 0 → **Unsuccessful landing**
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/blob/main/01%20Data%20Collection%20and%20Wrangling/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- **Charts & Purpose**
  - **Flight Number vs Payload Mass (scatter/strip plot)** → To analyze how payload size and number of launches relate to success/failure.
  - **Flight Number vs Launch Site (scatter plot)**→ To see success rates at different launch sites across missions.
  - **Payload Mass vs Launch Site (scatter plot)** → To study how payload weight affects landing success depending on site.
  - **Success Rate by Orbit (bar chart)** → To compare booster landing success across orbit types.
  - **Flight Number vs Orbit (scatter plot)** → To check how mission orbits vary with experience and success.
  - **Payload Mass vs Orbit (scatter plot)** → To examine relationship between payload weight, orbit type, and success.
  - **Success Rate by Year (line chart)** → To track improvement in landing success over time.
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/blob/main/02%20EDA/edadataviz.ipynb

EDA with SQL

- **SQL Queries Performed**
  - Created cleaned dataset (removed missing dates).
  - Retrieved unique launch sites & filtered by site codes.
  - Calculated payload stats (total, average, max).
  - Identified first successful landing dates.
  - Analyzed booster versions & outcomes (success/failure by conditions).
  - Counted missions and landing outcomes over time.
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/blob/main/02%20EDA/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

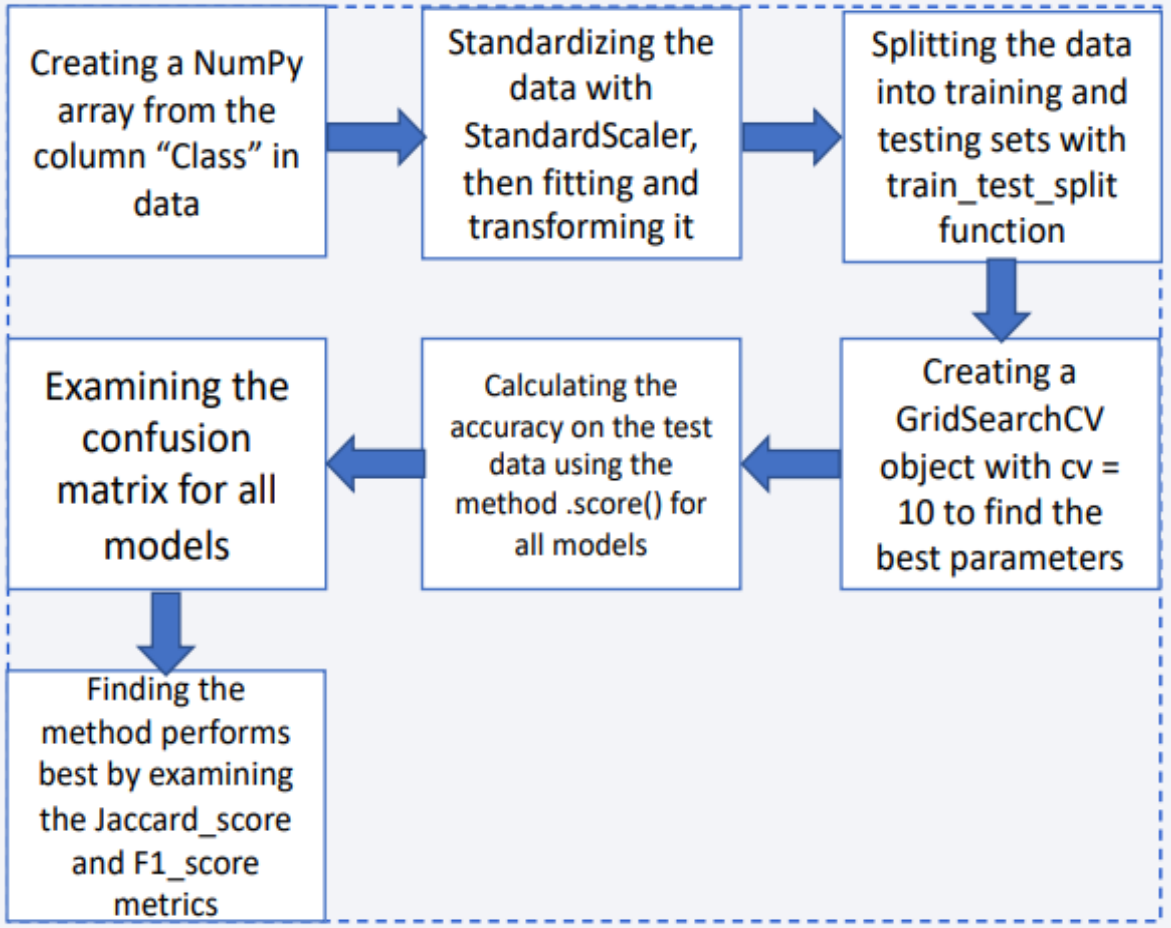
- Map Objects in Folium
  - **Markers** – show launch site locations
  - **Circles** – highlight surrounding areas/radius
  - **Lines** – connect points, show paths/distances
  - **Popups/Tooltips** – display site details
- **Purpose:** Clearly locate sites, show influence zones, illustrate connections, and provide quick info.
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/blob/main/03%20Visual%20Anlytics%20and%20Dashboard/DV0101EN-Exercise-Generating-Maps-in-Python.ipynb

Build a Dashboard with Plotly Dash

- **Plots**
  - Pie Chart → launch success by site
  - Scatter Plot → payload vs. success, colored by booster
- **Interactions**
  - Dropdown → select launch site
  - Range Slider → filter payload range
- **Why**
  - Compare sites & success rates
  - Analyze payload–outcome relation
  - Explore booster performance
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/tree/main/03%20Visual%20Anlytics%20and%20Dashboard/Plotly%20Dashboard

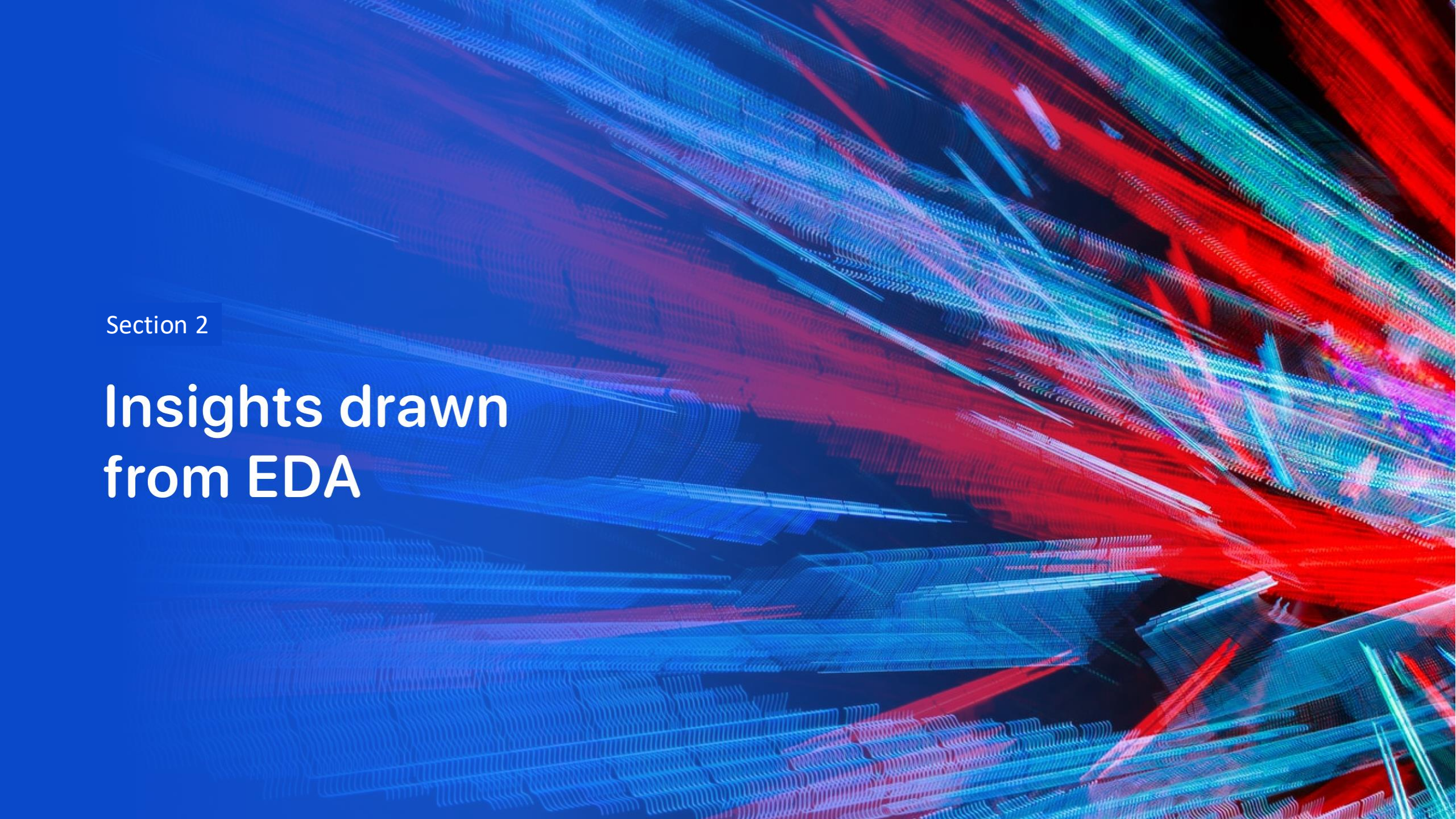
Predictive Analysis (Classification)

- **Steps Followed**
  - **Data Preparation** → Cleaned and encoded categorical variables, normalized features.
  - **Model Building** → Trained multiple classifiers (Logistic Regression, SVM, Decision Tree, KNN).
  - **Evaluation** → Compared models using accuracy, confusion matrix, and classification report.
  - **Hyperparameter Tuning** → Used GridSearchCV/parameter tuning to optimize models.
  - **Best Model Selection** → Picked the highest-performing model based on test accuracy.
- **Outcome -** Identified the best classification model for predicting SpaceX launch success.
- **GitHub URL:** https://github.com/aimless-coder/SpaceY_IBM_Applied_DataScience_Project/tree/main/03%20Visual%20Anlytics%20and%20Dashboard/Plotly%20Dashboard

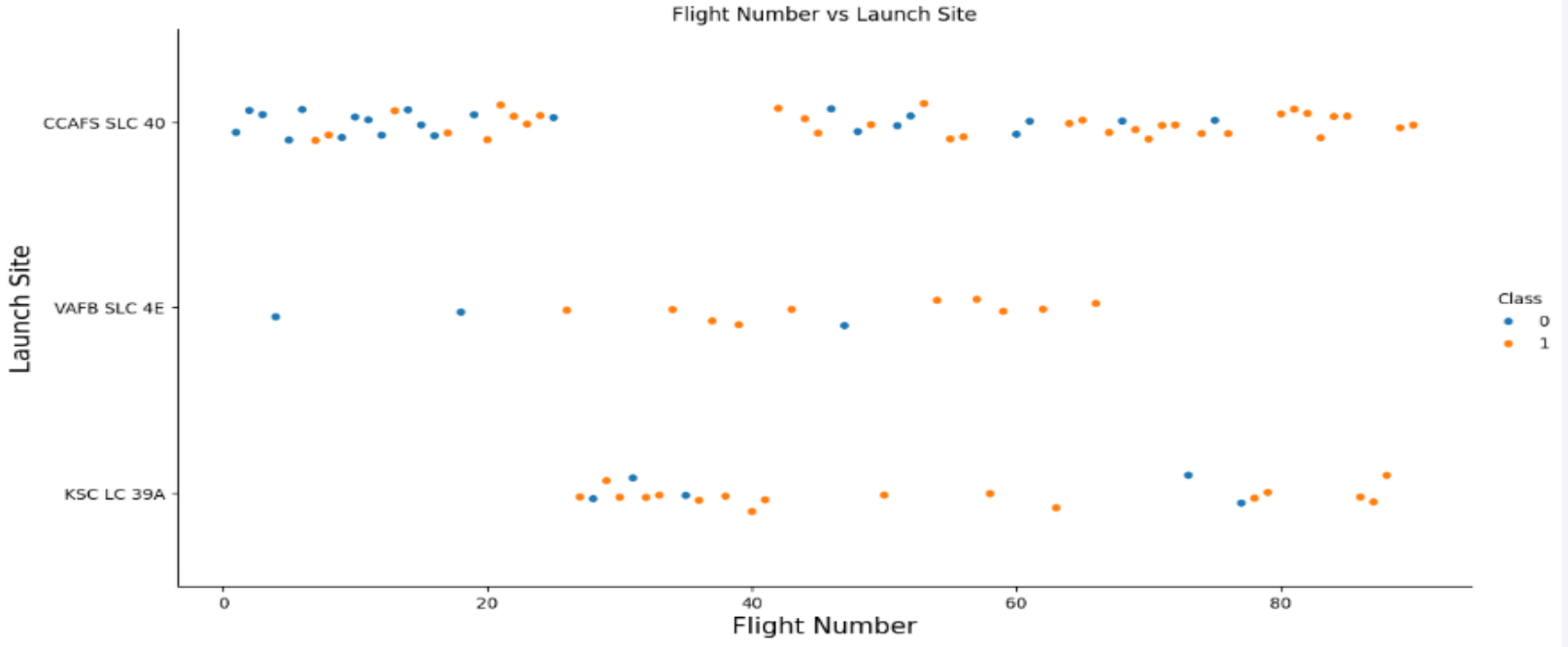


Results

- **EDA Findings**
  - Launch success rose from 0% (2010–13) → ~90% (2019).
  - Heavy payloads (>6,000 kg) show high success.
  - KSC LC-39A & CCAFS SLC-40 = 41.7% each of total successes.
  - GTO = most challenging orbit (~52% success).
- **Interactive Analytics**
  - Folium maps show strategic East & West Coast launch sites.
  - Plotly dashboards: site & payload filters highlight success patterns.
  - Spatial clusters identified high-performance zones.
- **Predictive Models**
  - Logistic Regression / SVM / KNN ≈ 83% accuracy.
  - **Decision Tree = Best (88.9% accuracy).**
  - Balanced performance confirmed via confusion matrix.

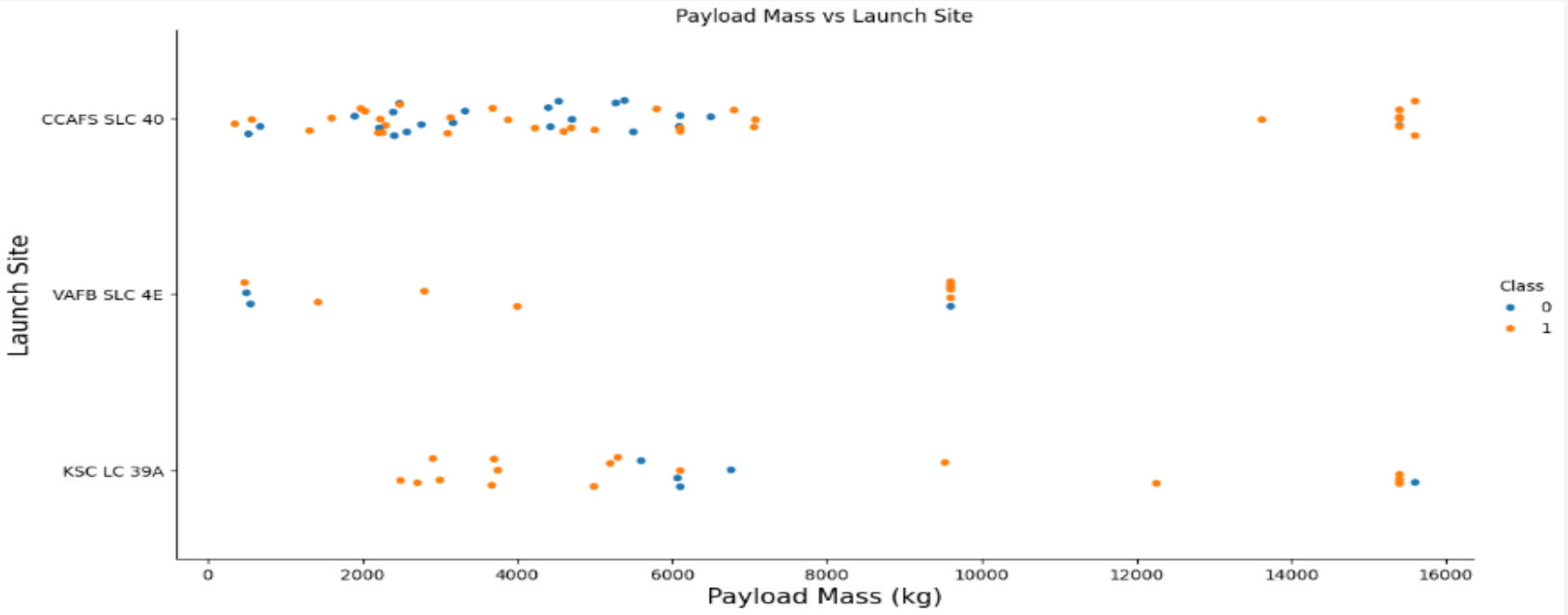

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- **CCAFS SLC 40 is the primary launch site**, hosting the greatest number of flights across the sequence, while **VAFB SLC 4E** is the least used.
- **Launch success (Class 1) is unevenly distributed:** Most launch sites show a mix of success and failure, but success becomes generally more frequent for flights at higher **Flight Numbers** (later in the sequence) at sites like **CCAFS SLC 40** and **KSC LC 39A**.

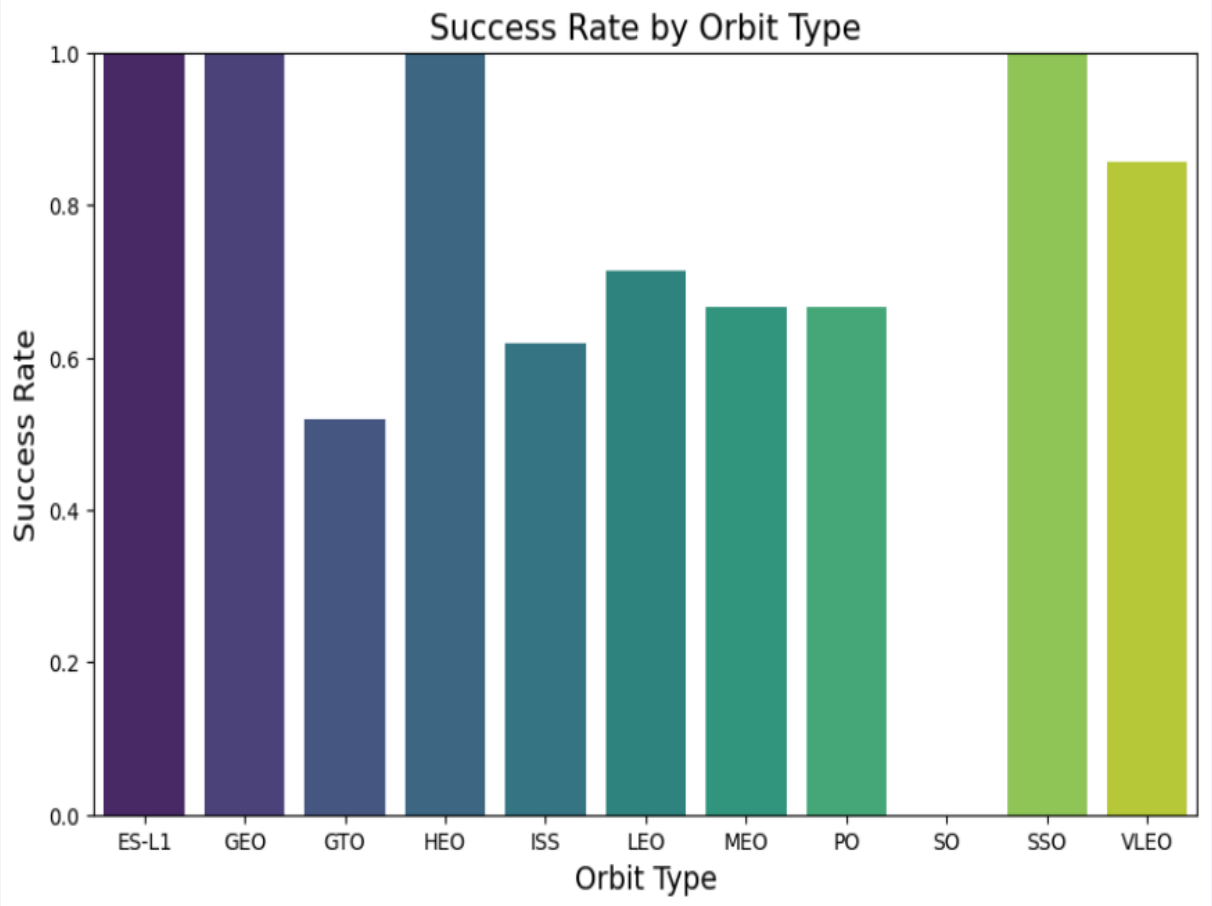
Payload vs. Launch Site



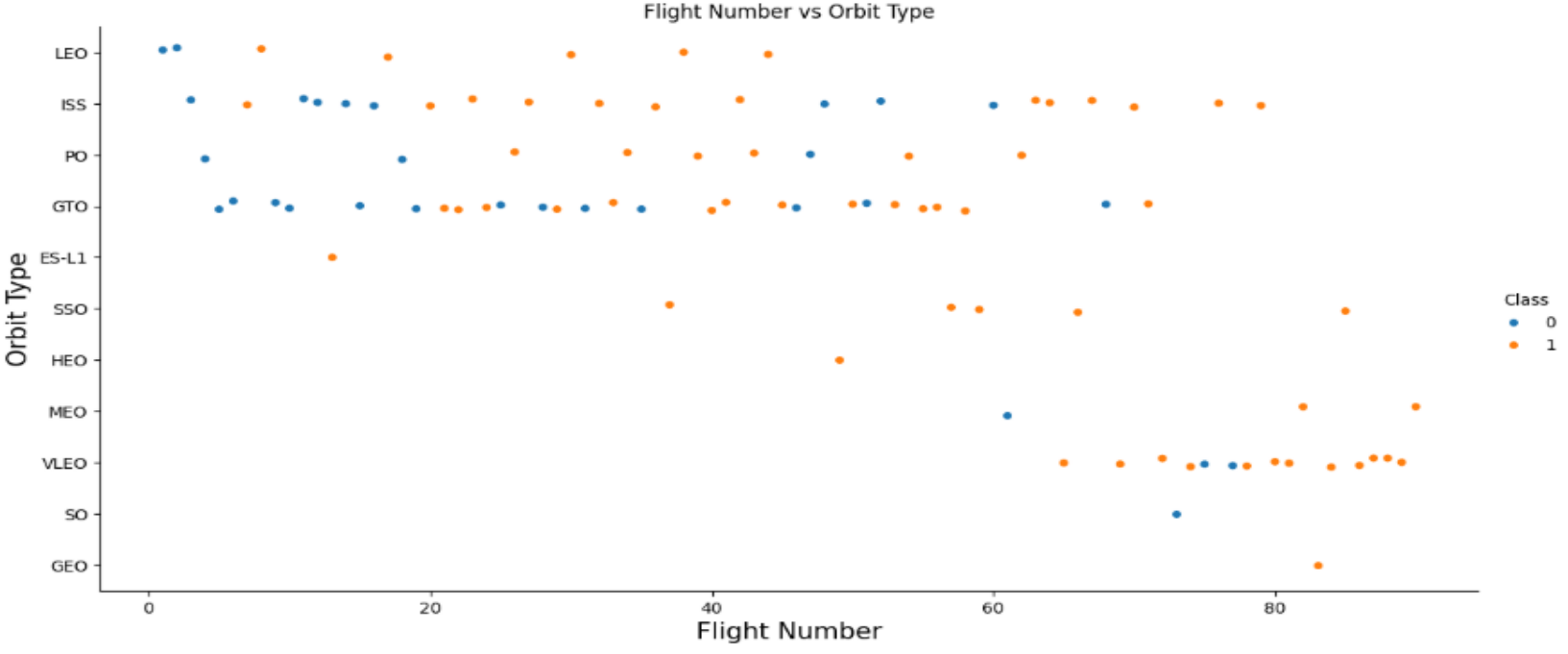
- **CCAFS SLC 40 is the High-Volume Site:** It launches the **widest and largest number of payloads**, dominating the low to mid-mass range (under 8,000 kg).
- **Specialization in Heavy Payloads:** The heaviest payloads (around 15,000 kg) are primarily launched from **CCAFS SLC 40** and **KSC LC 39A**, with the latter also handling unique very-heavy flights. **VAFB SLC 4E** focuses on specific medium-to-heavy missions (around 9,500 kg).

Success Rate vs. Orbit Type

- **Orbits with 100% Success Rate:** Several orbit types—**ES-L1**, **GEO**, **HEO**, and **SSO**—have a perfect success rate of 1.0 (or 100%). This suggests that the rocket system is highly reliable for these specific mission profiles.
- **GTO has the lowest success rate, while LEO, MEO, and PO are clustered in the mid-range: GTO (Geostationary Transfer Orbit)** has the lowest success rate at approximately 0.52 (or 52%), indicating it is the most challenging orbit to achieve successfully. Conversely, common orbits like **LEO (Low Earth Orbit)**, **MEO (Medium Earth Orbit)**, and **PO (Polar Orbit)** all have success rates clustered between approximately 0.67 and 0.72, suggesting moderate but reliable performance for these frequent missions.

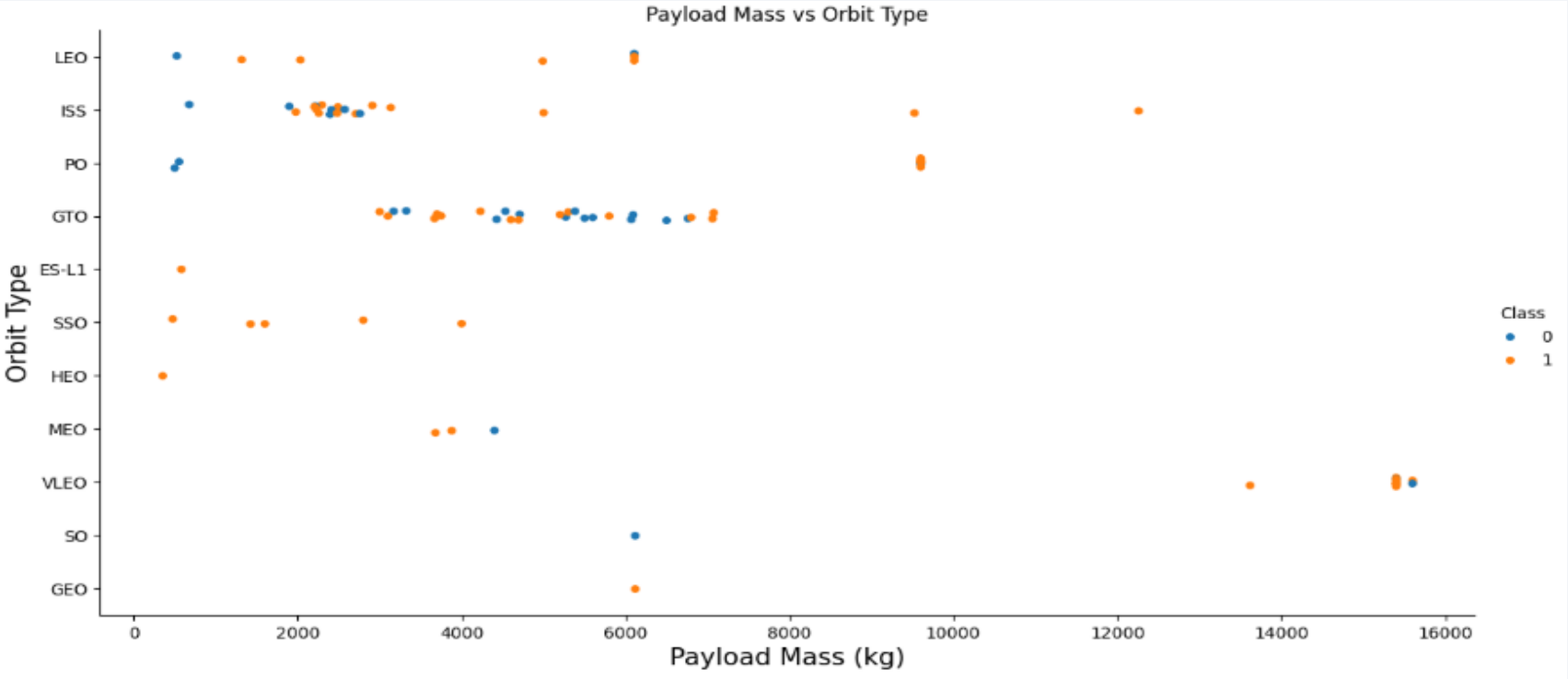


Flight Number vs. Orbit Type



- **GTO is the Primary Mission:** It is the **most frequent orbit** and has been flown consistently from the first to the last flight number.
- **VLEO is a Newer, Highly Successful Orbit:** It began around **Flight 60** and shows an almost perfect success rate (**Class 1**). Early flights focused heavily on **LEO** and **ISS**.

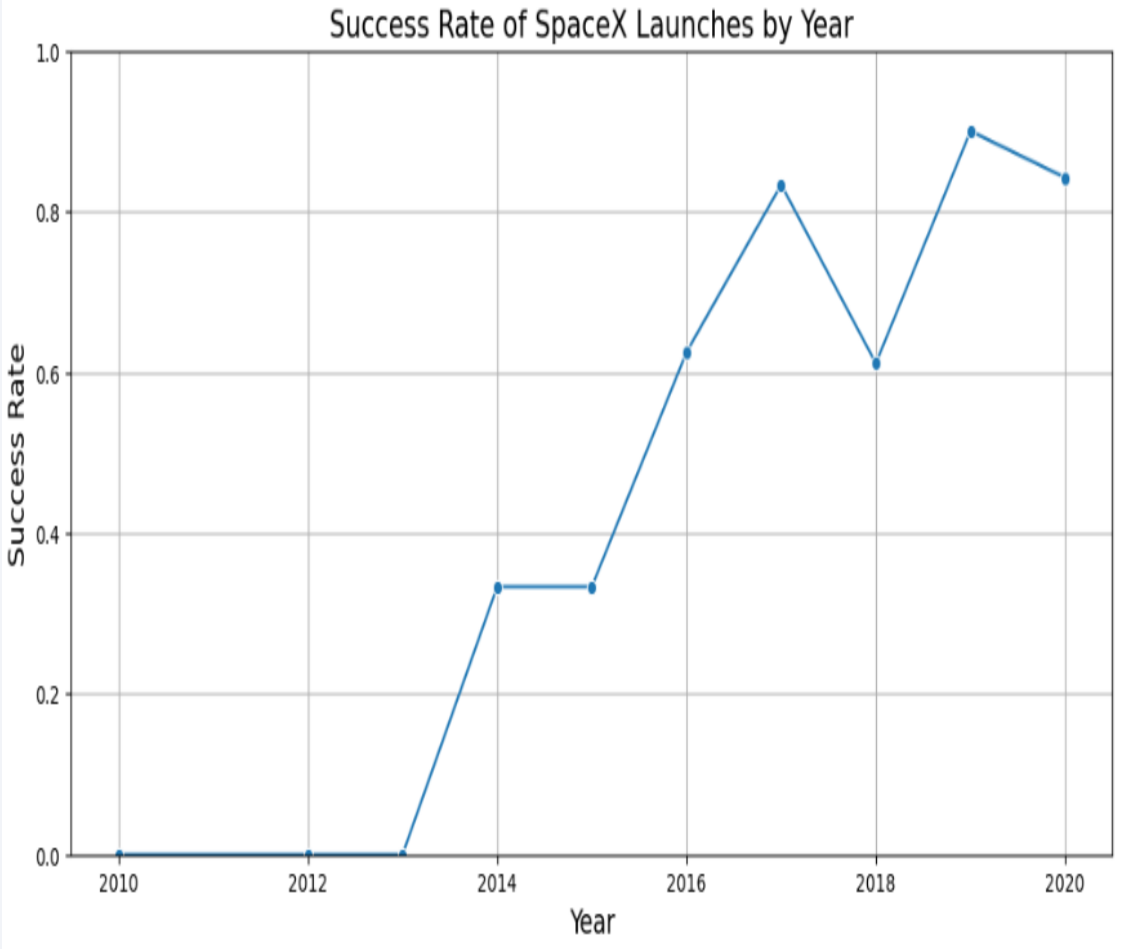
Payload vs. Orbit Type



- **GTO Dominates Mid-Mass Payloads: GTO** is the most frequent orbit, accommodating the largest number of missions in the 1,500 kg to 7,000 kg range, with a high success rate.
- **VLEO Handles the Heaviest Payloads Successfully:** Payloads over 14,000 kg are primarily launched to **VLEO** and have a near-perfect **Class 1** success rate.

Launch Success Yearly Trend

- **Massive Reliability Improvement:** Success rate soared from **0% (2010–2013) to ~90% (2019)**, showing strong system maturation.
- **Recent Success Volatility:** Despite the overall trend, annual success rates have fluctuated significantly since **2016** (e.g., 83% in 2017, 61% in 2018, 90% in 2019).

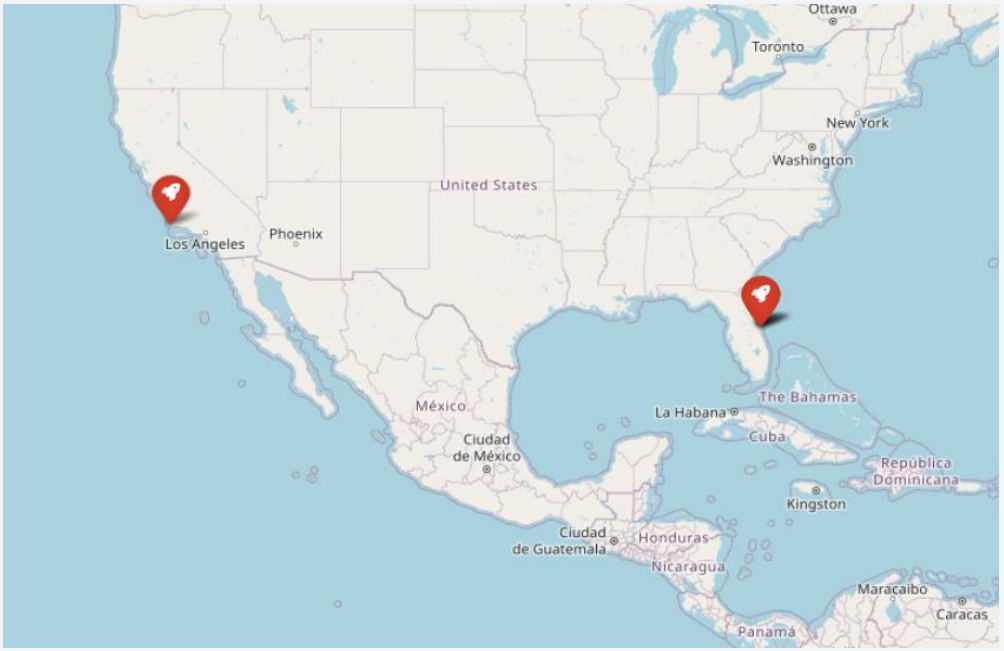


Section 3

Launch Sites Proximities Analysis

Geographic Distribution of Launch Sites

- The launch sites are strategically located on **both the East and West Coasts of the continental United States**.
- The two primary locations are near **Cape Canaveral, Florida** (East Coast, for easterly orbits) and near **Los Angeles, California** (West Coast, often used for polar or southerly orbits).

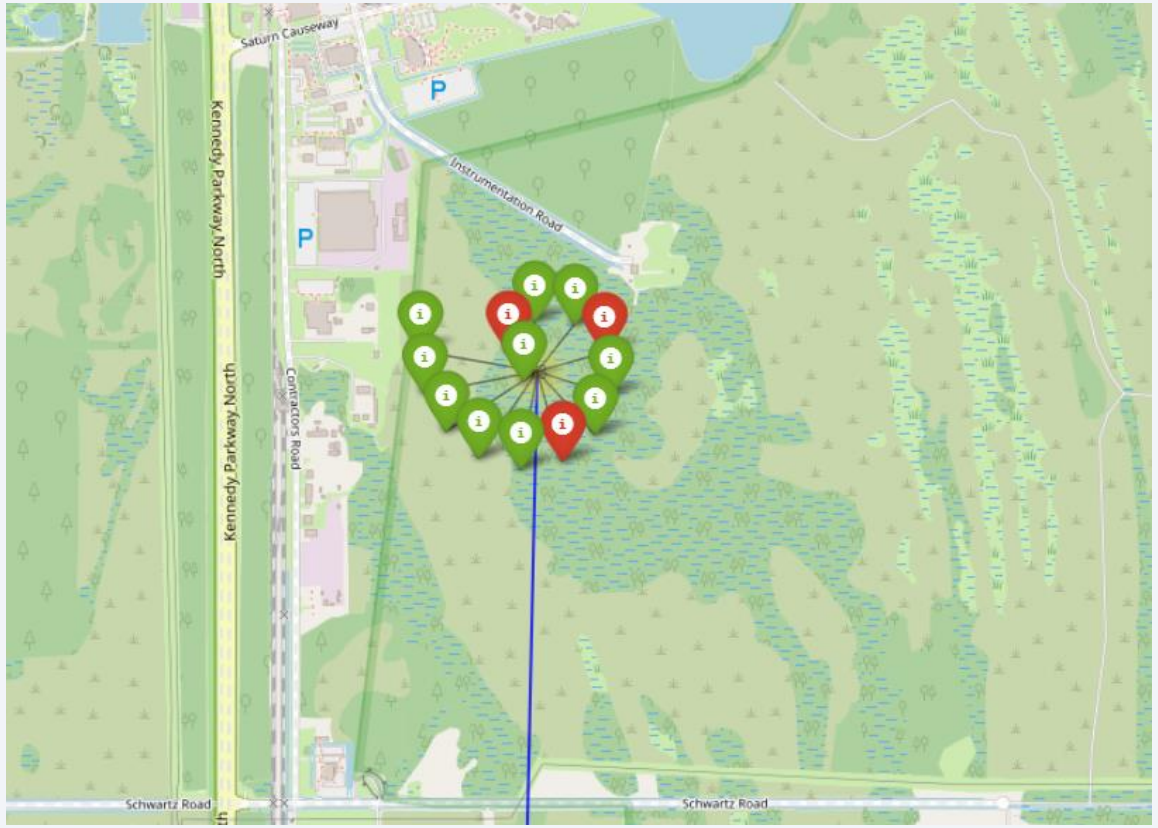


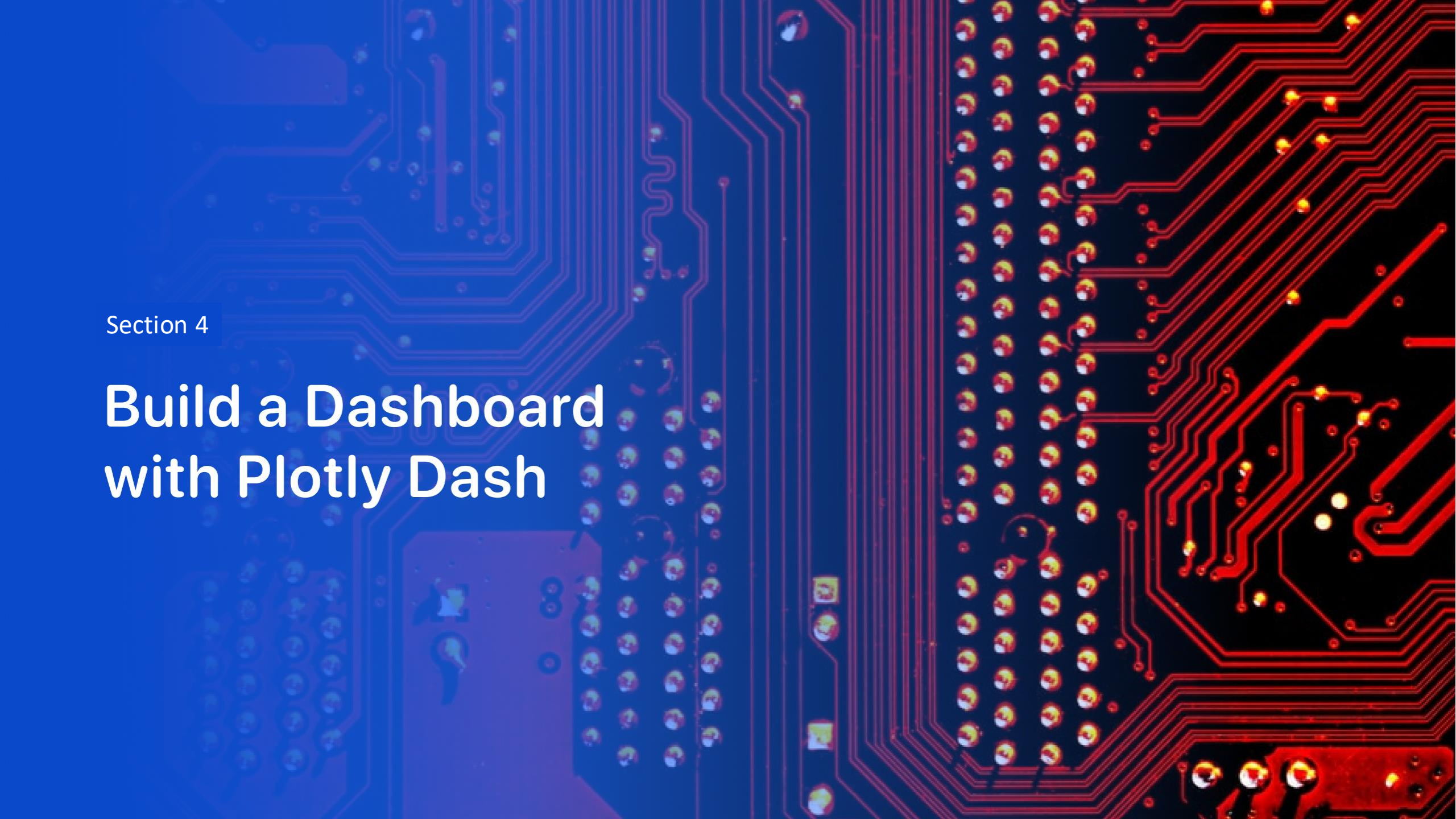
Launch Outcomes

- The map displays data points arranged in a **distinct spiral pattern** around a central location (labeled '26'), suggesting a systematic outward survey or distribution from that core point.
- The points are categorized into **two outcomes** represented by the **red** (e.g., failure, hazard) and **green** (e.g., success, safe) pins, showing the classified result at each location.

Localized Geospatial Survey near Kennedy Space Center

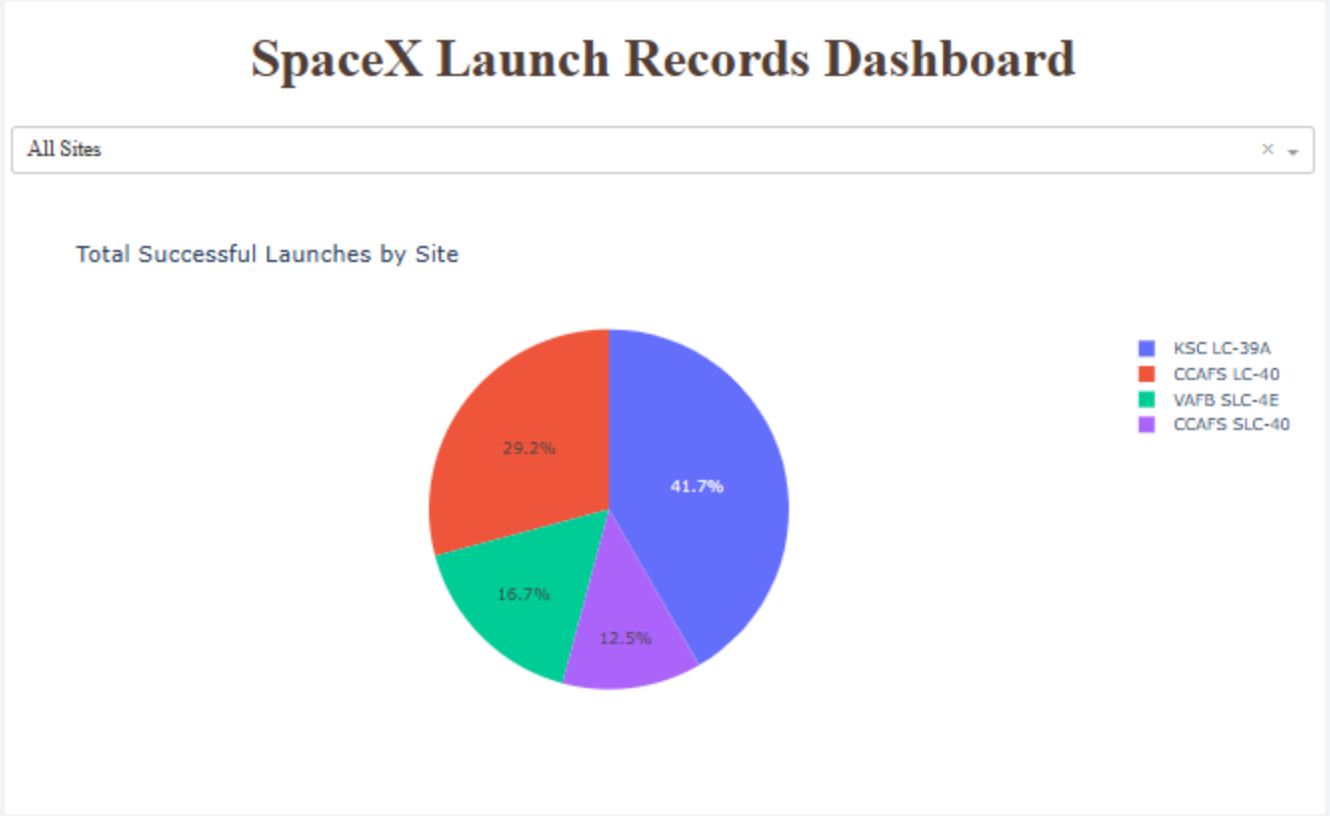
The image displays a localized cluster of data points (green and red pins) within a marshland area, bordered by major roads like **Kennedy Parkway North** and **Instrumentation Road**, suggesting a specific test or survey zone within the Cape Canaveral/KSC complex.




Section 4
Build a Dashboard with Plotly Dash

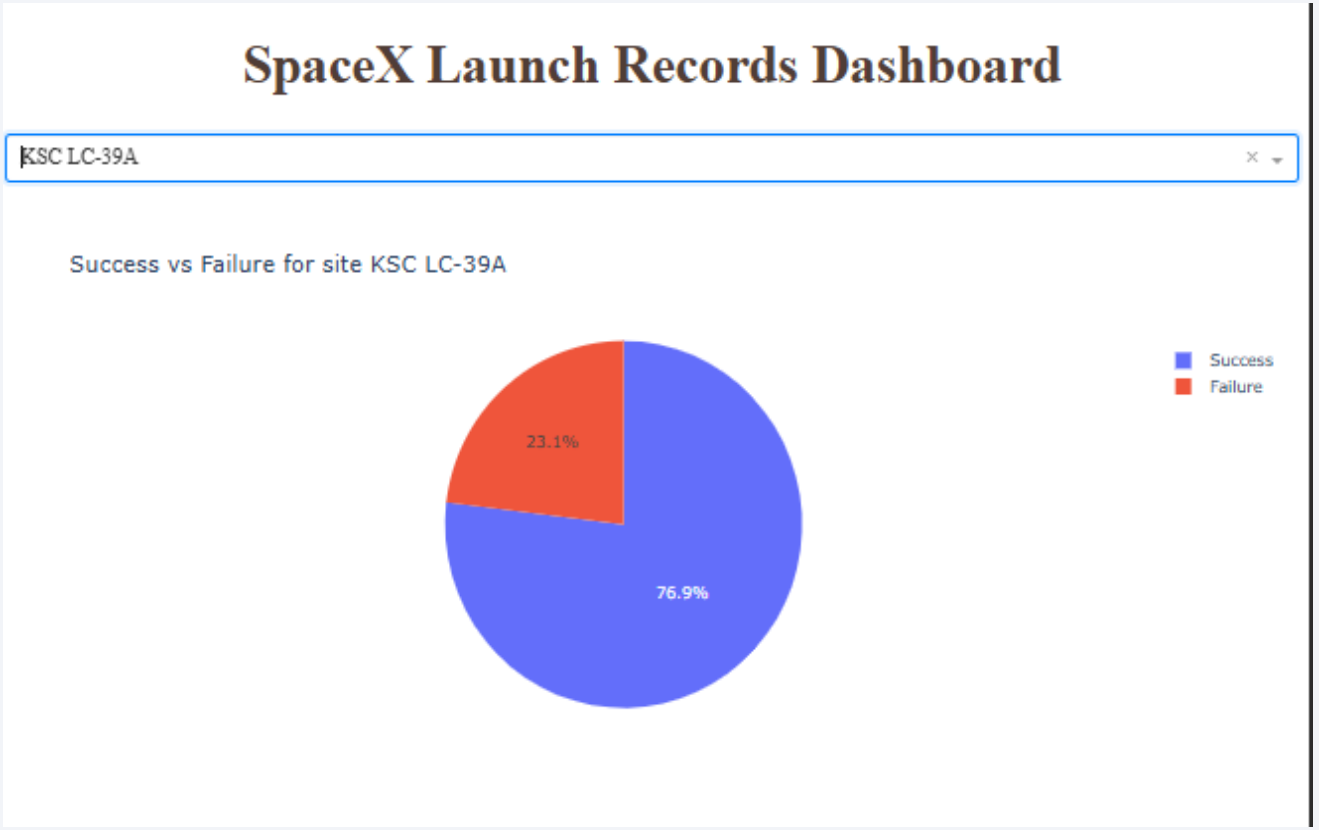
Successful Launch Contribution by Site

- **KSC LC-39A is the top launch site**, contributing the single largest share of successful missions at **41.7%**.
- **CCAFS SLC-40 (combined) is equally successful**, with its two segments adding up to a total of **41.7%** of all successful launches.



KSC LC-39A Performance Metrics

- **High Success Rate:** The site demonstrates a strong performance with **Success (blue) at 76.9%**.
- **Success Dominates:** Launches from **KSC LC-39A** are overwhelmingly successful, outnumbering failures (red, 23.1%) by more than a 3-to-1 margin.



Payload vs. Outcome

- **Failures span the light-to-mid range** (up to ~6,000 kg), occurring with various booster versions.
- **Heavy payloads (over ~6,000 kg) are uniformly successful (Class 1)**, demonstrating high reliability for high-mass missions.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

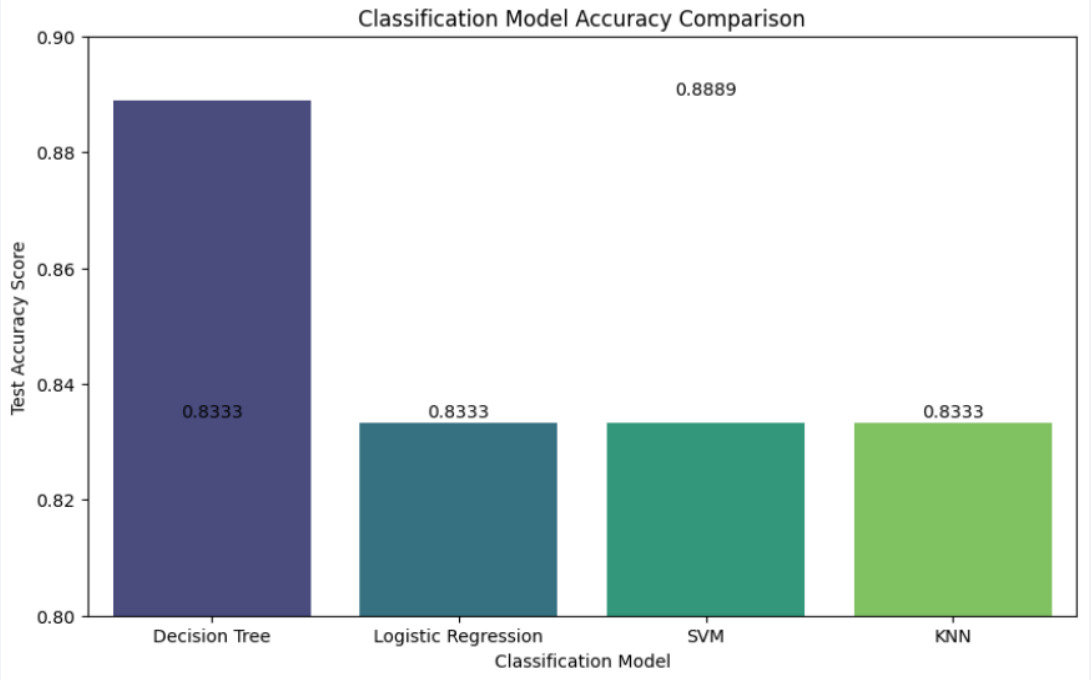
Logistic Regression test accuracy: 0.8333333333333334

SVM test accuracy: 0.8333333333333334

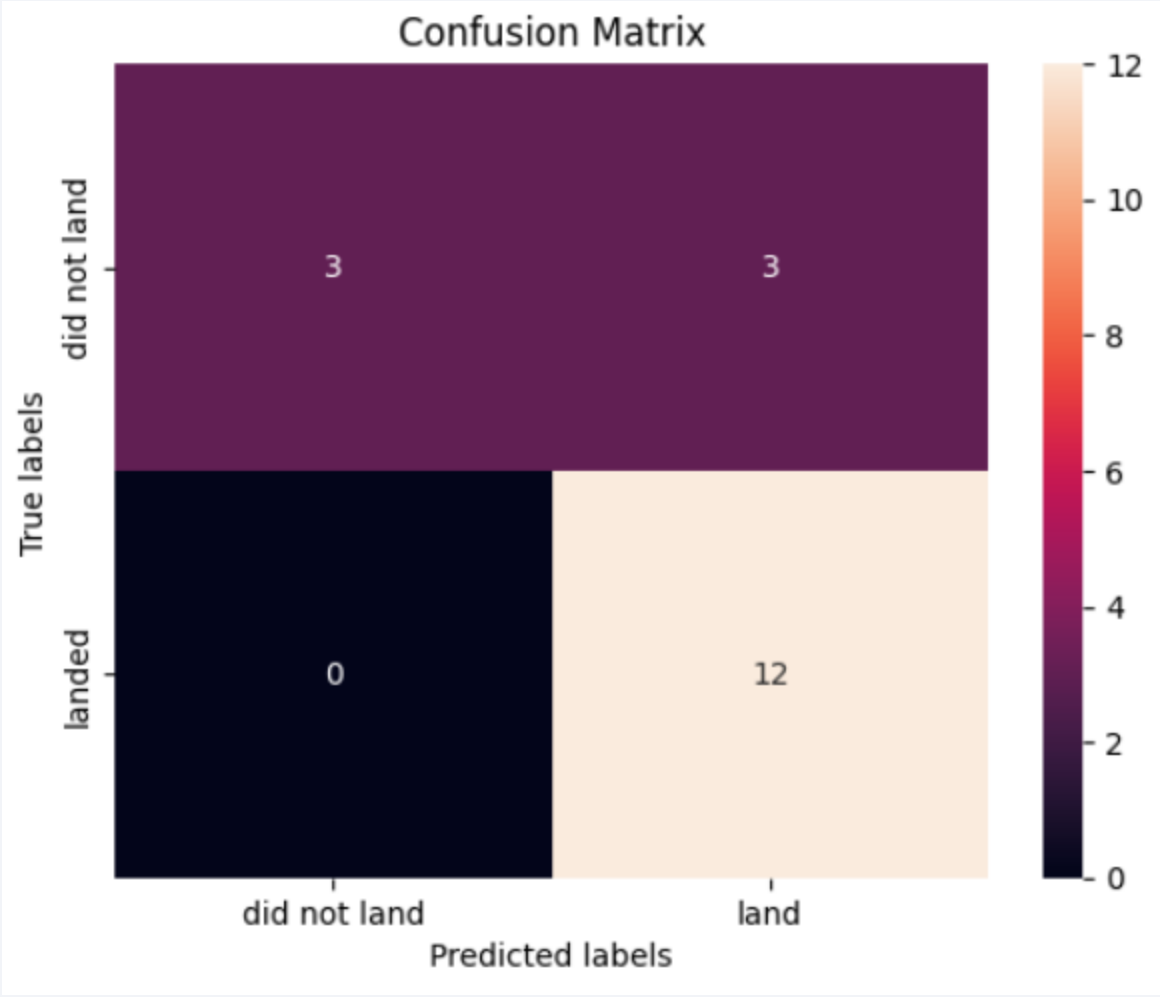
Decision Tree test accuracy: 0.8888888888888888

KNN test accuracy: 0.8333333333333334

The model that performs best on the test data is: Decision Tree with accuracy: 0.8888888888888888



Confusion Matrix



Conclusions

- SpaceX landing success improved dramatically over time.
- Launch site & payload mass are key drivers of outcomes.
- Orbit type matters: GTO toughest, VLEO/LEO/SSO highly reliable.
- Decision Tree model is most effective for predicting landings.
- EDA + interactive dashboards + ML → strong framework for cost & mission planning.

Thank you!