SPACEX LAUNCH ANALYSIS & PREDICTIVE MODELING

IBM Data Science Capstone Project

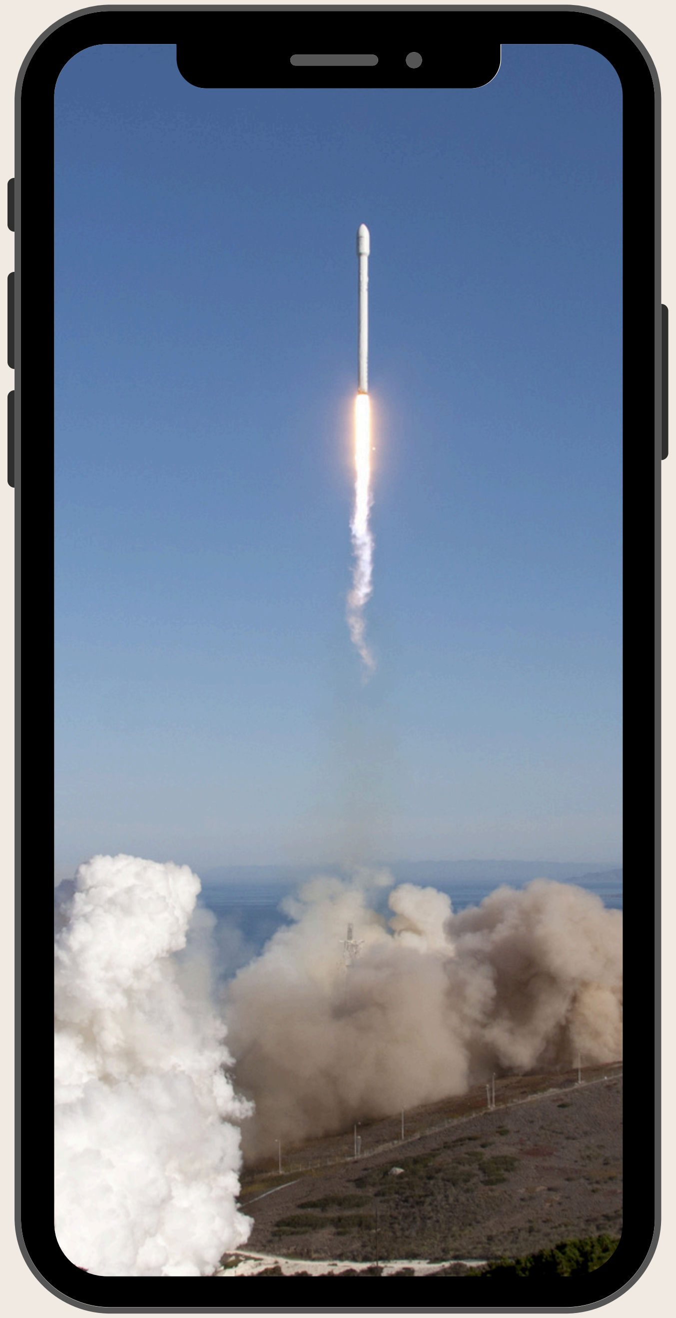
Autor: Alberto Martín Sánchez

En GitHub: Aiberto-Floppy

TABLE OF CONTENTS

1. **Executive Summary**
2. **Introduction**
3. **Data Collection**
4. **Data Wrangling**
5. **EDA & Visualization**
6. **EDA with SQL**
7. **Interactive Map (Folium)**
8. **Plotly Dash Dashboard**
9. **Predictive Analysis**
10. **Conclusion**

EXECUTIVE SUMMARY



- **Goal**: Predict SpaceX Falcon 9 first-stage landing success.
- **Data Sources**: SpaceX API, Wikipedia, and CSV datasets.
- **Methods**: Data wrangling, EDA (visual & SQL), interactive mapping, and machine learning models.
- **Key Finding**: Decision Tree achieved the highest test accuracy (94%).
- **Impact**: Supports cost estimation and competitive bidding against SpaceX.

INTRODUCTION

- **Context:** SpaceX reuses first-stage boosters, reducing launch costs.
- **Objective:** Analyze launch data to predict landing success.
- **Approach:** Combine data collection, EDA, visualization, and predictive modeling.
- **Benefit:** Insights to help competitors estimate costs and plan bids.

DATA COLLECTION & WRANGLING

Data Sources

- SpaceX REST API (launch details: rockets, payloads, cores).
- Wikipedia HTML tables (Falcon 9 & Falcon Heavy launches).
- Preprocessed CSV datasets for consistent analysis.

Process

- Collected launch details (site, payload, booster, outcomes).
- Cleaned missing values (e.g., payload mass).
- Filtered Falcon 9 launches only.

Output

- Clean dataset with standardized columns ready for Exploratory Data Analysis (EDA).

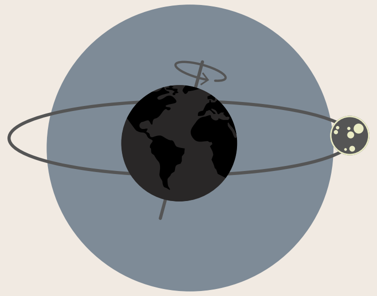
EDA & VISUALIZATION (PARTE 1)

Exploratory Data Analysis – Overview

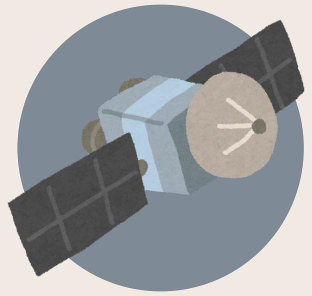
- Analyzed launch frequency across sites and payload mass distribution.
- Observed trends: higher success rates with increasing flight numbers.
- Visualized payload mass vs. success rate using scatterplots.

Success Rates by Orbit Type

EDA & VISUALIZATION (PARTE 2)



Calculated success rates for each orbit (GTO, LEO, ISS, etc.)



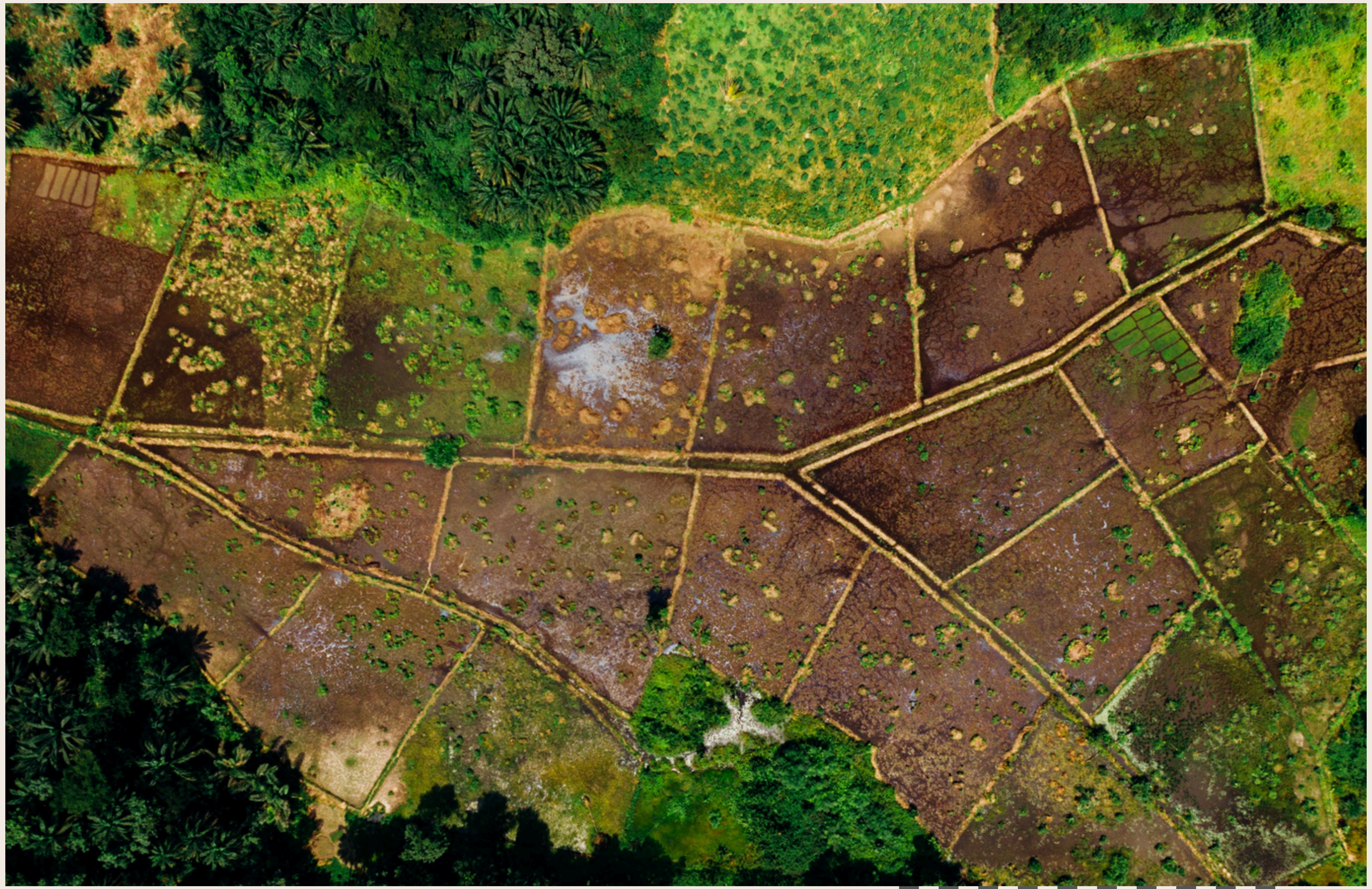
LEO and ISS showed higher success with heavy payloads.

GTO success rates were more variable and not tied to payload mass.

Flight Number vs. Launch Success

EDA & VISUALIZATION (PARTE 3)

- Early flights had lower success rates.
- Success rate improved significantly as flight number increased.
- Suggests learning curve in SpaceX launch operations.



EDA & VISUALIZATION (PARTE 4)

Payload Mass vs. Launch Success

- Heavier payloads often succeed in LEO, Polar, and ISS orbits.
- For GTO missions, success is mixed — heavy payloads are more challenging.
- Highlights the role of mission type in success probability.

EDA WITH SQL: QUERYING LAUNCH DATA

- **Purpose**: Use SQL queries to explore SpaceX launch data.
- **Focus**: Launch sites, payload mass, and mission outcomes.
- **Benefit**: Quick insights directly from the database before deeper visualization.

EDA WITH SQL: LAUNCH SITES DISTRIBUTION (SQL QUERY)

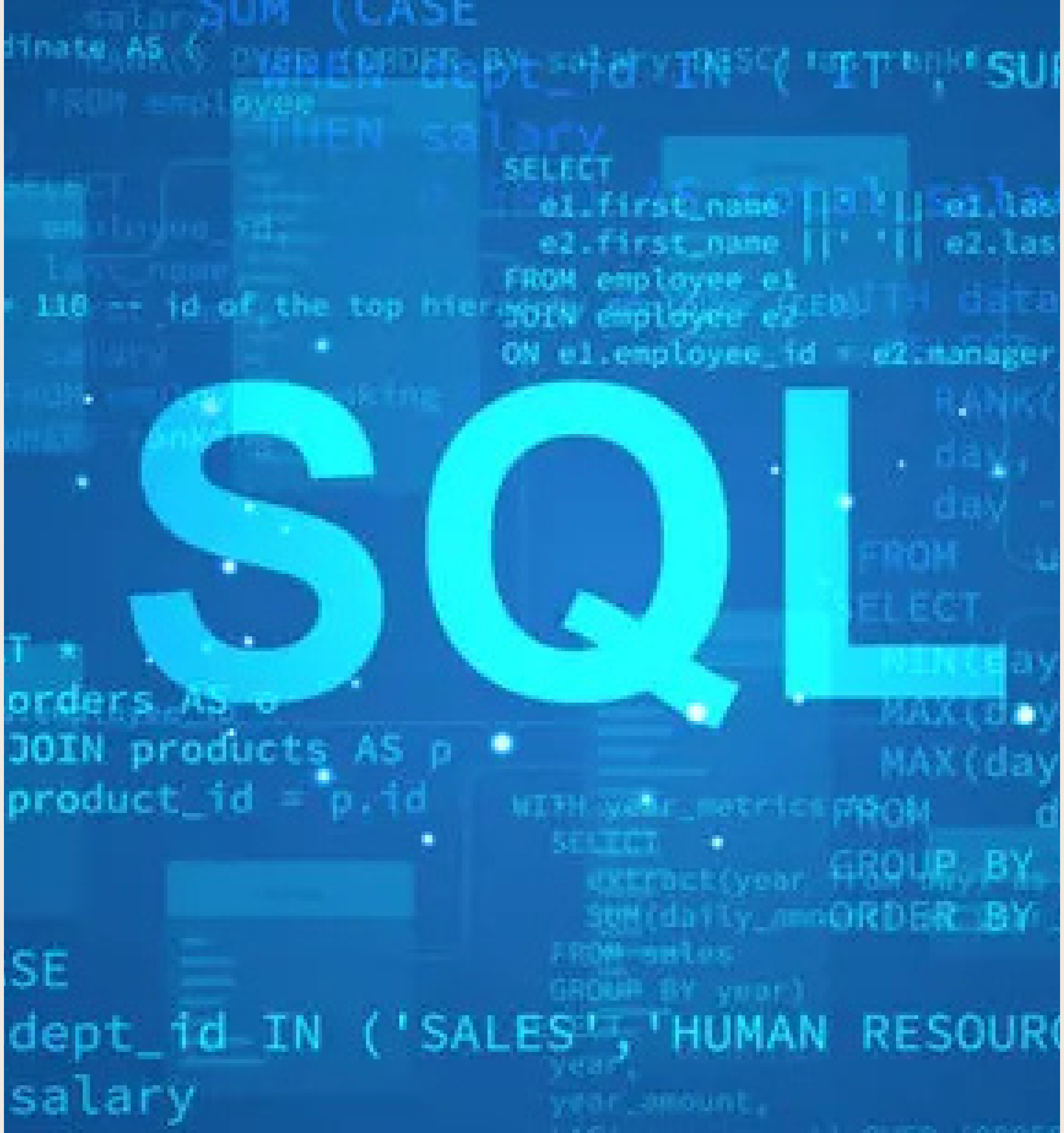
- **Query:** Count launches per site.
- **Key finding:** CCAFS SLC-40 has the highest number of launches.
- **Sites analyzed:** 4 main launch sites.

EDA WITH SQL: PAYLOAD MASS VS ORBIT (SQL QUERY)

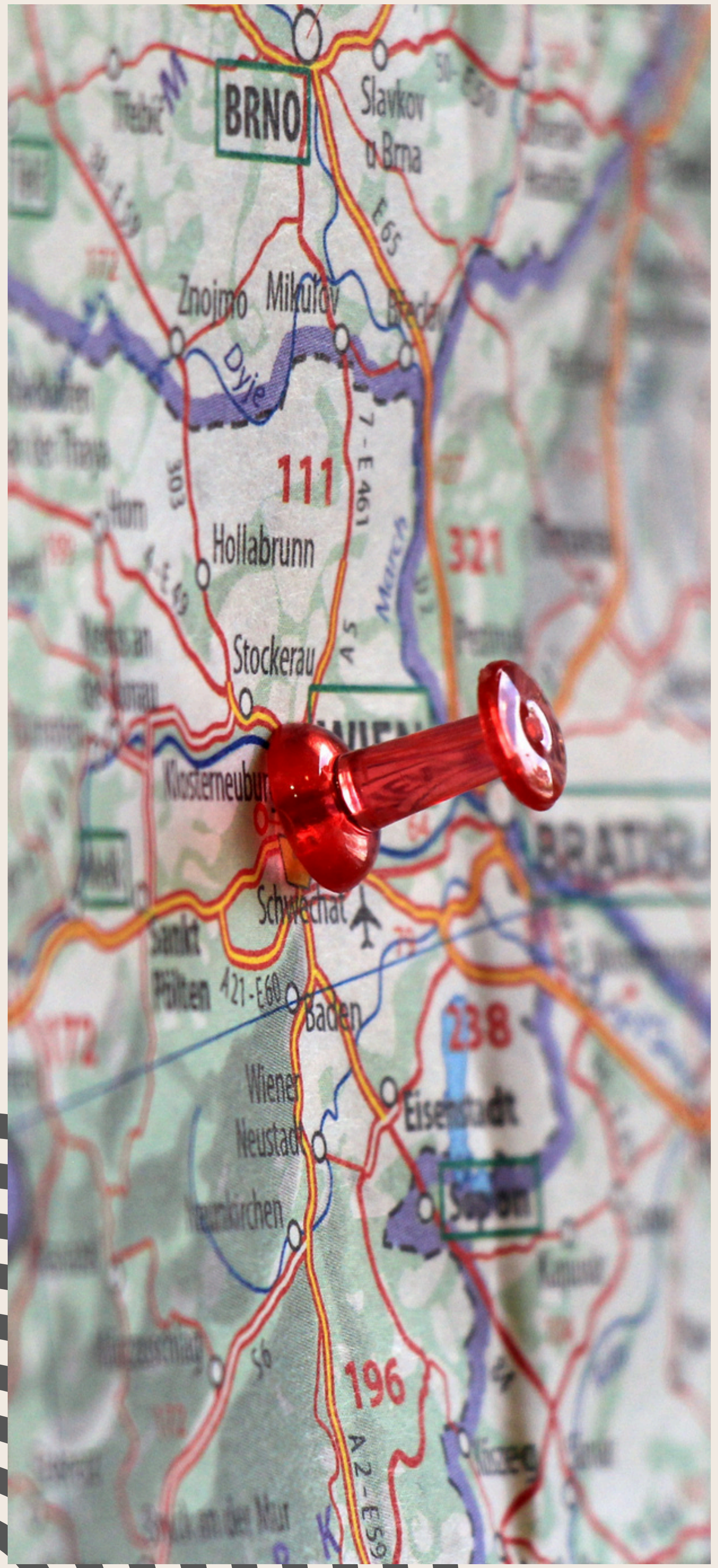
- **Query:** Calculate payload distribution by orbit type.
- **Key insight:** GTO and ISS dominate heavy payloads.
- **Trend:** LEO and VLEO carry lighter payloads.

EDA WITH SQL: QUERIES

- Analyzed launch data using SQL queries.
- Found CCAFS SLC-40 had the most launches (55).
- Highest success rates observed for ISS and LEO orbits.
- Results validated trends found in Python visualizations.



Interactive Map with Folium



- Mapped all launch sites with success/failure markers.
- Visualized proximity to coastlines, railways, and highways.
- Identified strategic locations near the coast for safe landings.

INTERACTIVE DASHBOARD WITH PLOTLY DASH

- Built real-time interactive dashboard for launch analysis.
- Features: dropdowns, sliders, pie charts, scatter plots.
- Explored success rate by site and payload range dynamically.

PREDICTIVE ANALYSIS METHODOLOGY

Methods Used

Applied logistic regression, SVM, decision trees, and KNN. Hyperparameters tuned via GridSearchCV.

Data Sources

Combined data from SpaceX API, Wikipedia tables, and CSV datasets provided for analysis.

Timeline

Data collection → EDA → Interactive visuals (Folium, Dash) → Model training and evaluation.

Data Nature

Mixed numerical (payload, flight number) and categorical (launch site, orbit) features standardized for models.

PREDICTIVE ANALYSIS RESULTS

Model 1

Logistic Regression: 83.3% accuracy

Model 2

SVM: 83.3% accuracy

Model 3

Decision Tree: 94.4% accuracy (best)

Model 4

KNN: 83.3% accuracy

CONCLUSION & KEY TAKEAWAYS

- SpaceX launch success strongly tied to experience (flight number) and orbit type.
- ISS and LEO orbits show the highest success rates.
- Decision Tree model performed best (94% accuracy on test data).
- Analysis supports cost predictions and site optimization for future launches.

MUCHAS GRACIAS

Julio de 2025