Winning the Space Race with Data Science

SpaceX Falcon 9 Landing Prediction – End-to-End Data Science Project

Author: Chetan Sai Abhishek Injavarapu

Institution: IBM Applied Data Science Capstone (Coursera)

GitHub Repository: https://github.com/chetan-957/IBM-Capstone

# 1. Executive Summary

This project implements a complete end-to-end data science workflow to predict the successful landing of SpaceX Falcon 9 first-stage boosters. Reusable rocket technology has significantly reduced launch costs, but landing outcomes depend on multiple operational and technical factors.

The workflow included data collection via the SpaceX REST API, web scraping of landing outcomes, data wrangling, SQL-based exploratory data analysis, interactive visual analytics using Folium and Plotly Dash, and predictive modeling using classification algorithms.

All four classification models achieved approximately 83% test accuracy. The Decision Tree model demonstrated the highest cross-validation accuracy (~88.9%) and achieved perfect recall for successful landings.

# 2. Introduction

SpaceX revolutionized aerospace engineering through reusable Falcon 9 rockets. Successful recovery of first-stage boosters reduces costs and improves mission efficiency.

This report investigates key determinants of landing success, including payload mass, orbit type, launch site, booster version, and operational maturity.

**Research Questions:**

• What factors most influence landing success?

• Does payload mass impact landing probability?

• Do launch sites and orbit types affect outcomes?

• Has landing reliability improved over time?

• Can machine learning models accurately predict landing success?

# 3. Methodology

## 3.1 Data Collection

Launch data was collected using the SpaceX REST API, extracting flight number, payload mass, orbit type, launch site, booster version, and landing outcome. Landing history data was scraped from Wikipedia using BeautifulSoup and merged with API results into a structured dataset.

## 3.2 Data Wrangling

Data preprocessing included filtering Falcon 9 missions, removing irrelevant columns, handling missing values, encoding categorical variables, and creating a binary target variable representing landing success (1) or failure (0).

## 3.3 Exploratory Data Analysis

EDA was conducted using SQL and Python visualizations. Bar charts, scatter plots, and line charts were used to analyze launch site performance, payload trends, orbit complexity, and yearly success rates.

## 3.4 Interactive Analytics

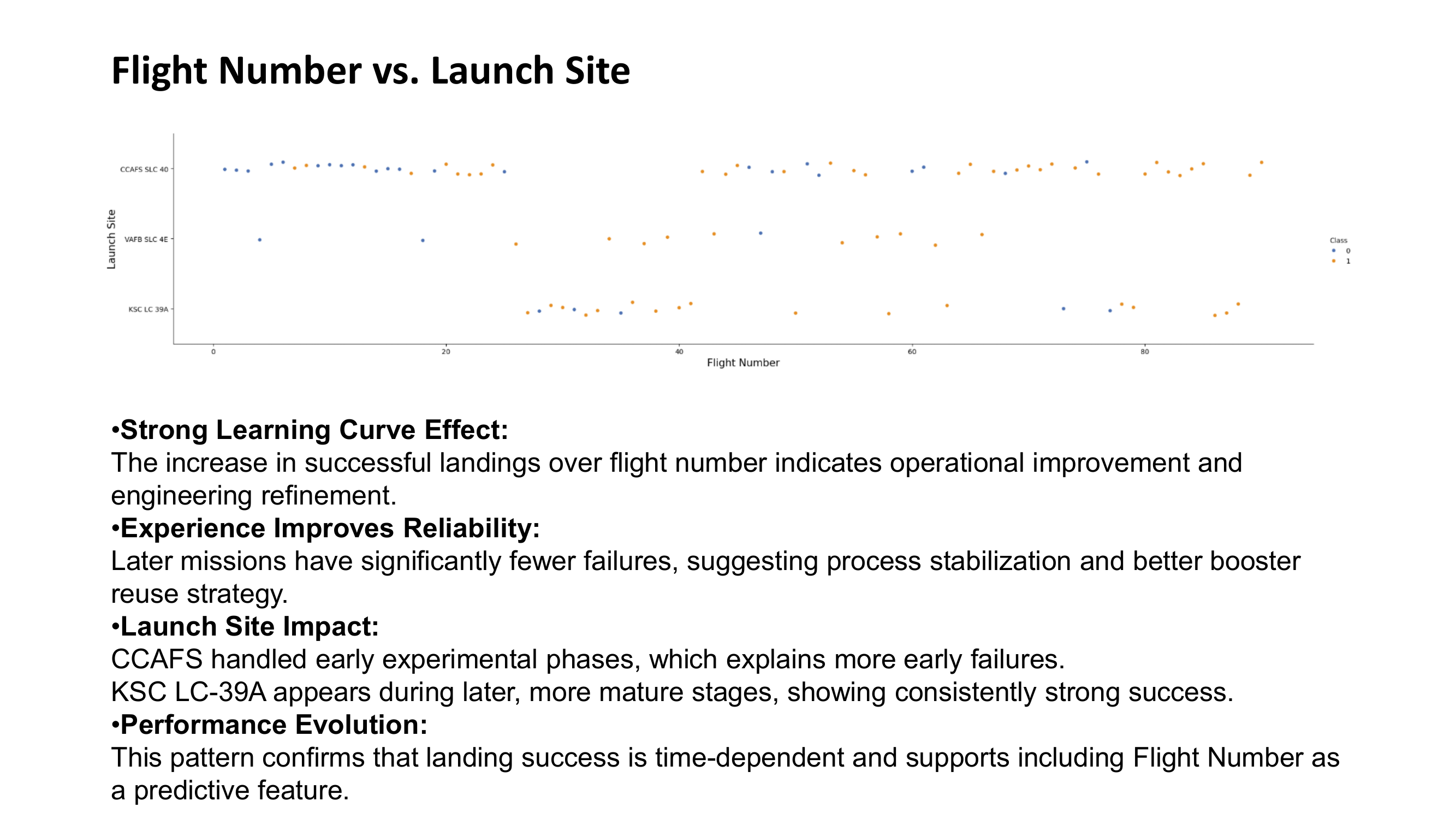
Interactive maps were built using Folium to visualize geographic distribution and launch site proximity analysis. A Plotly Dash dashboard enabled dynamic filtering by launch site and payload range.

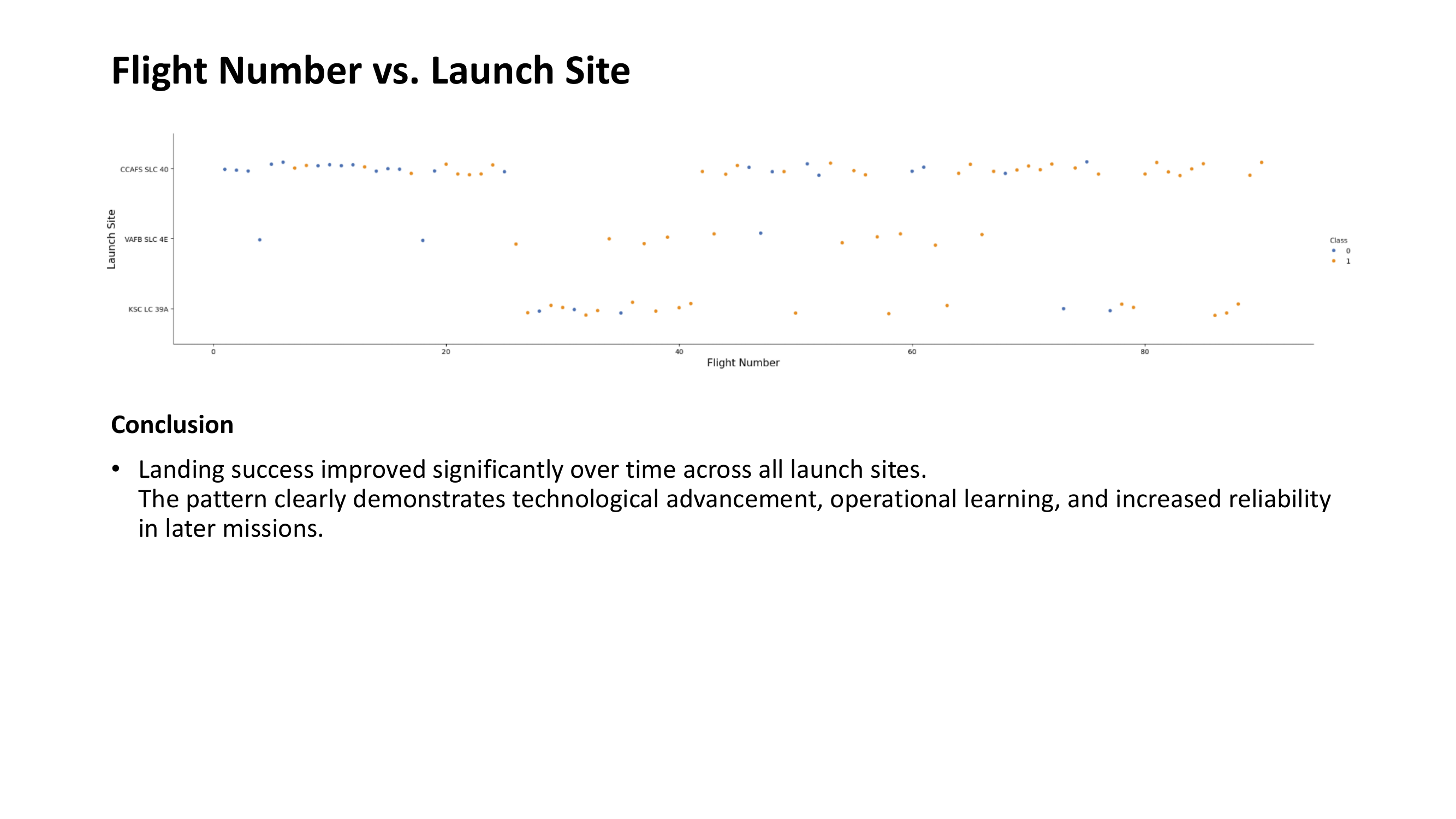
## 3.5 Predictive Modeling

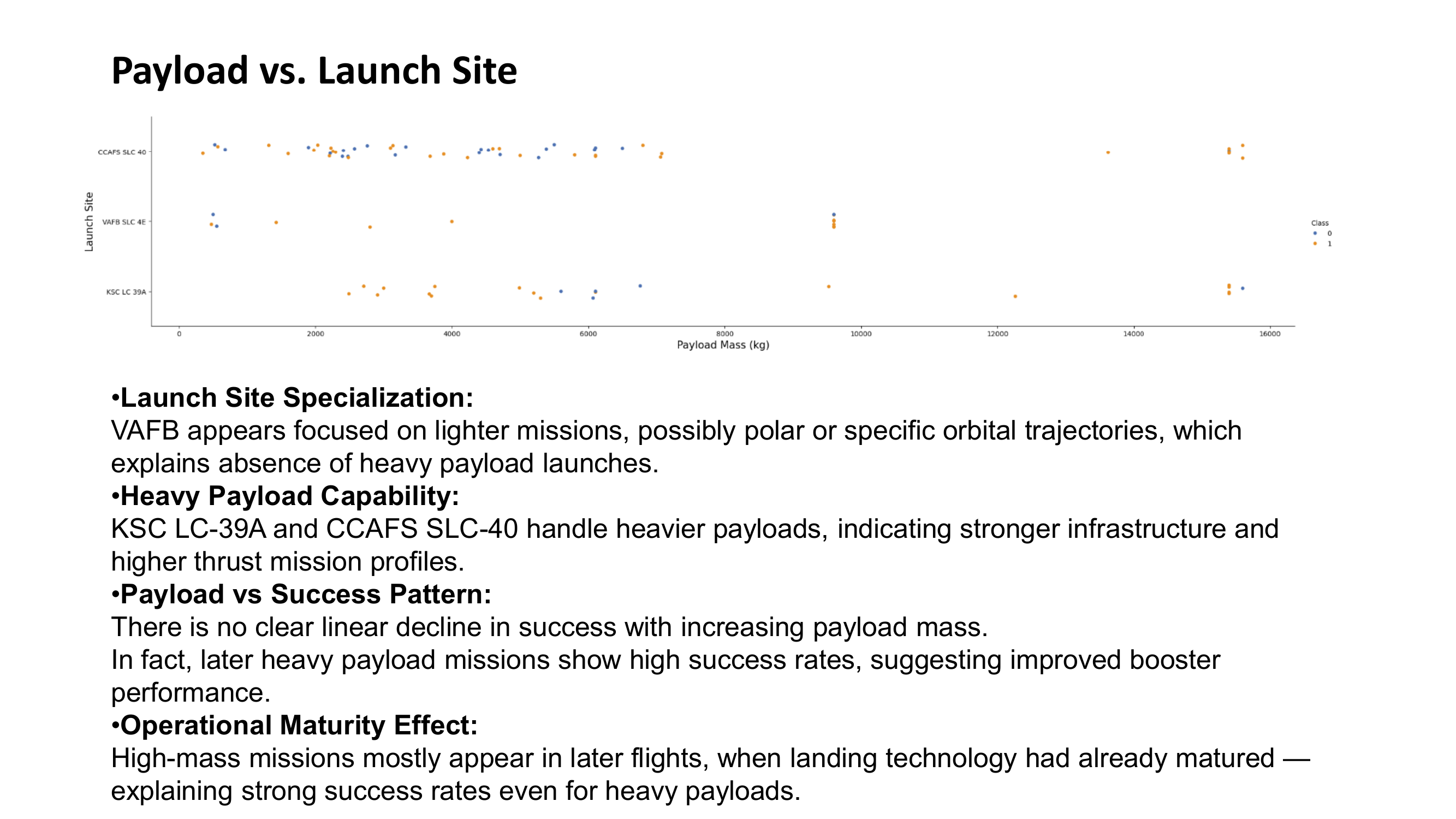
Four classification models were implemented: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors. Data was split 80/20 for training and testing, and 10-fold cross-validation was used for hyperparameter tuning.

# Analytical Results & Visual Insights

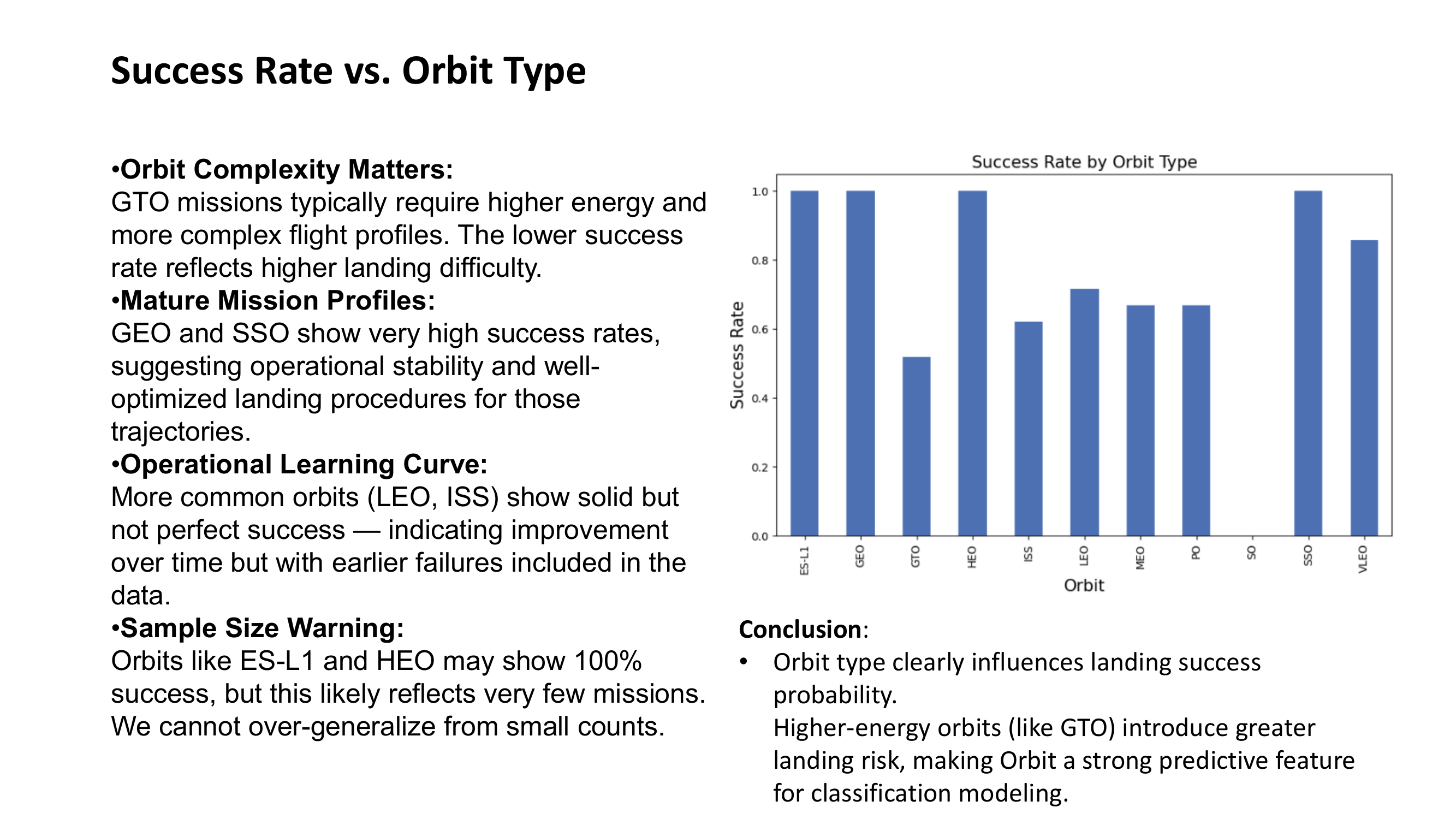
The following figures summarize key analytical insights derived from EDA, SQL analysis, geospatial modeling, interactive dashboards, and predictive classification.

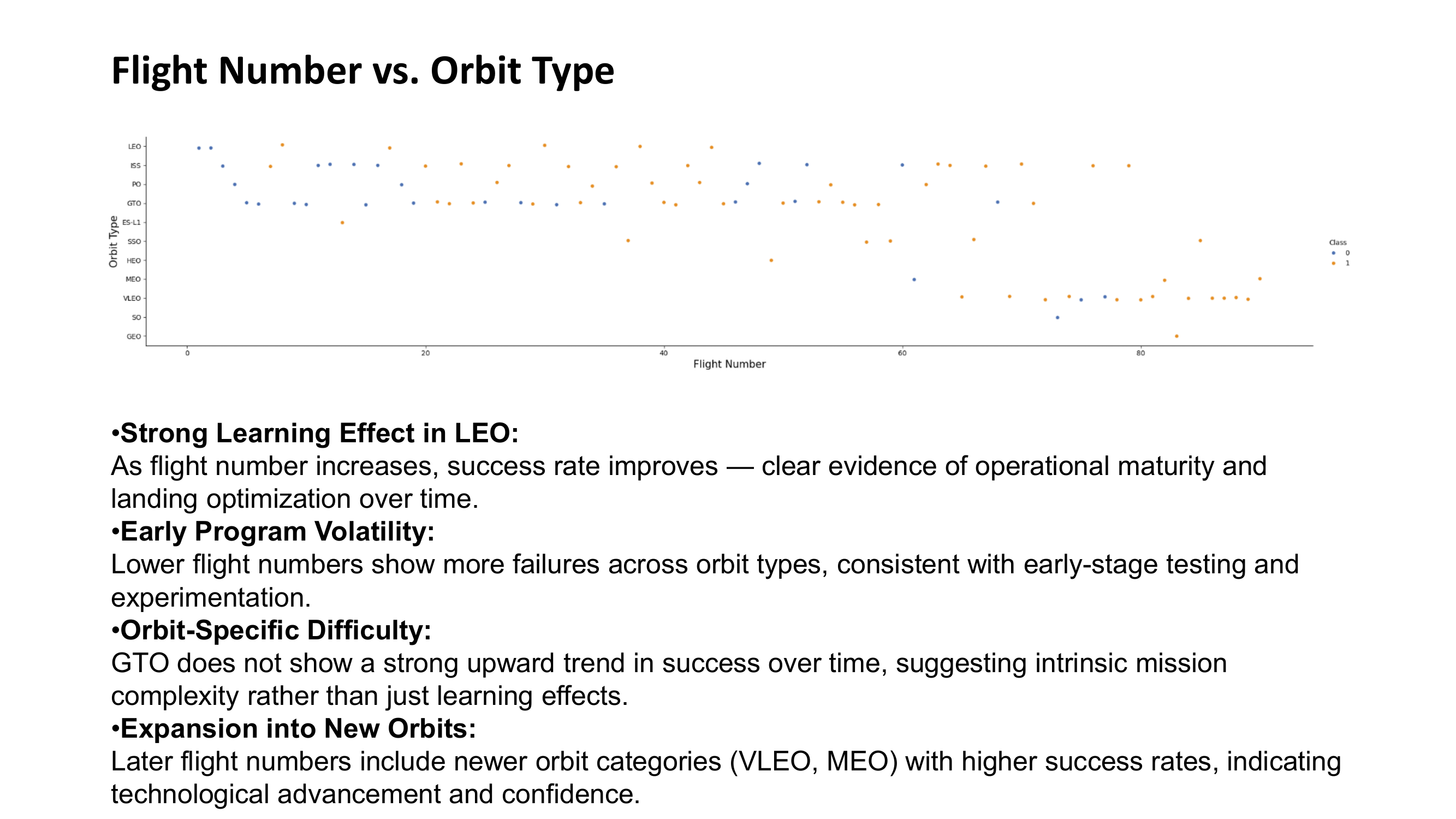


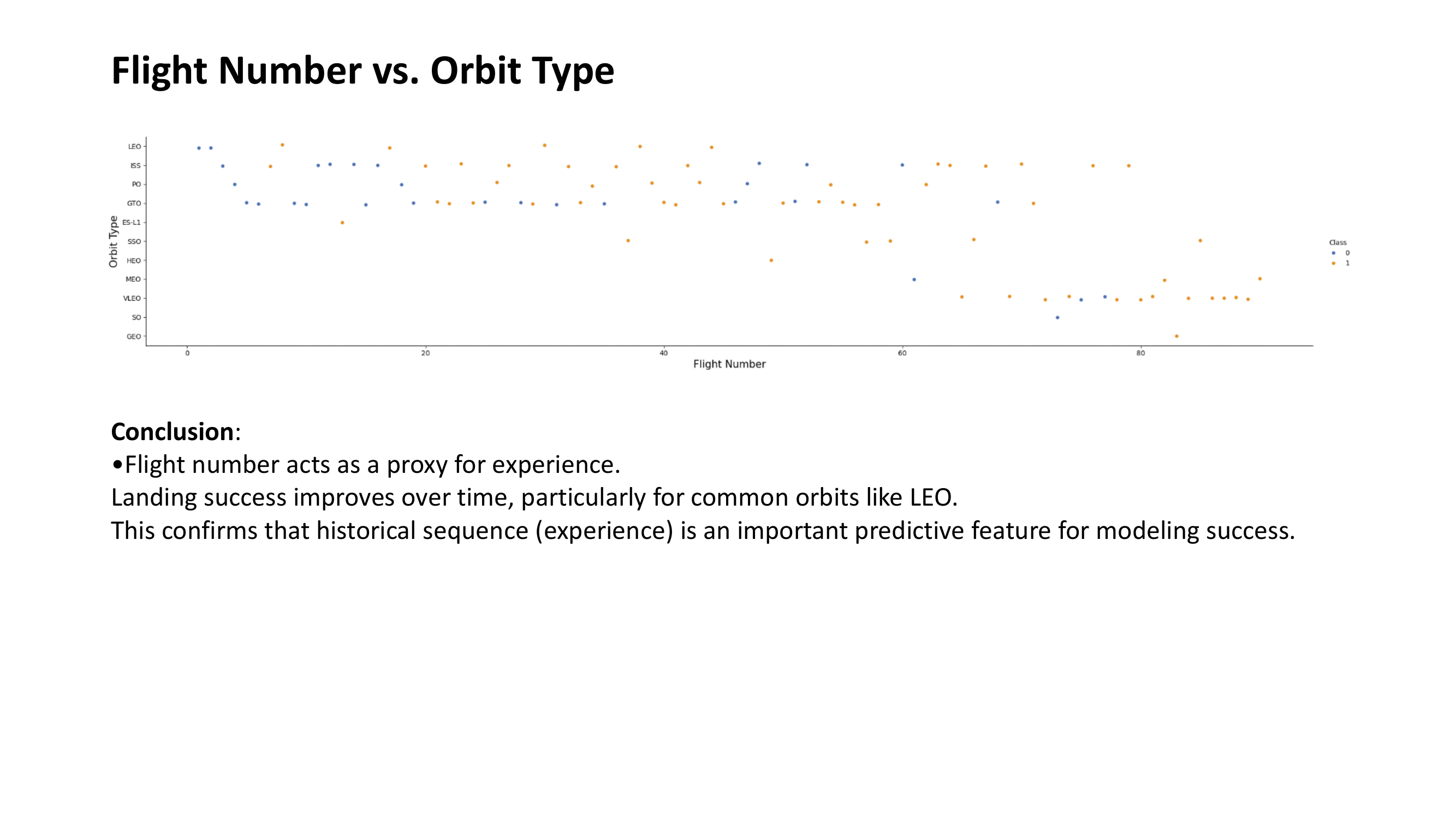


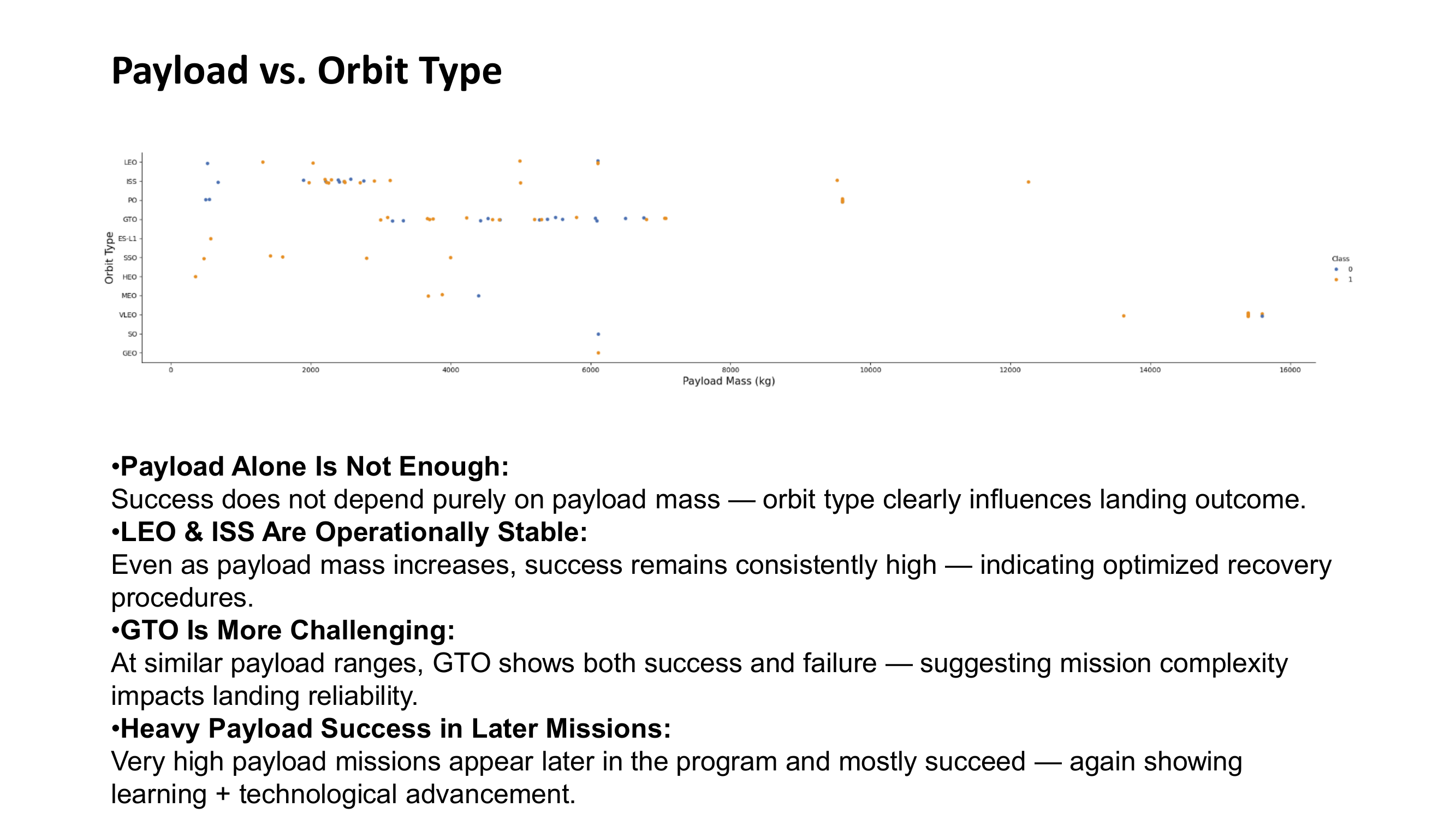


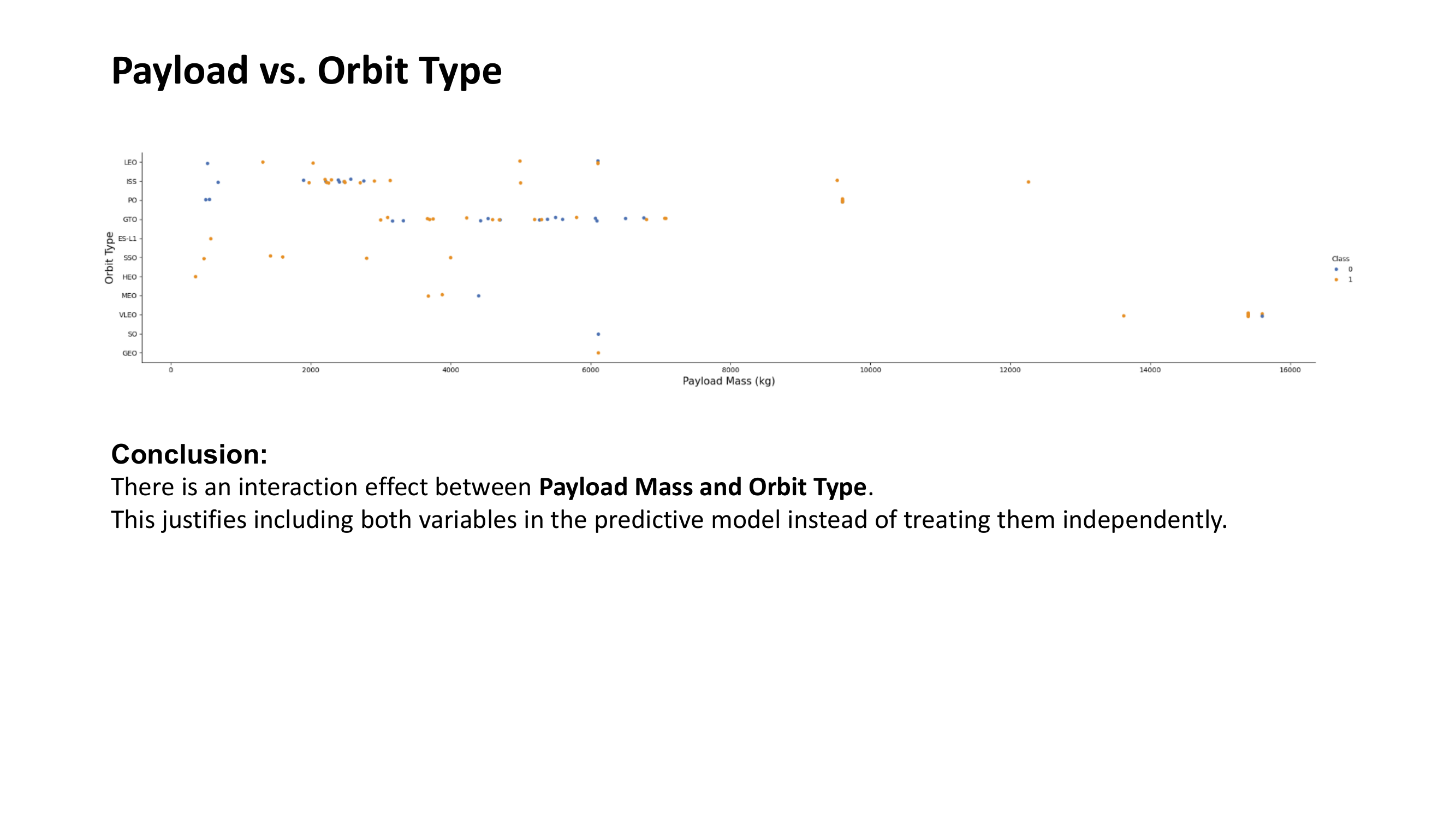


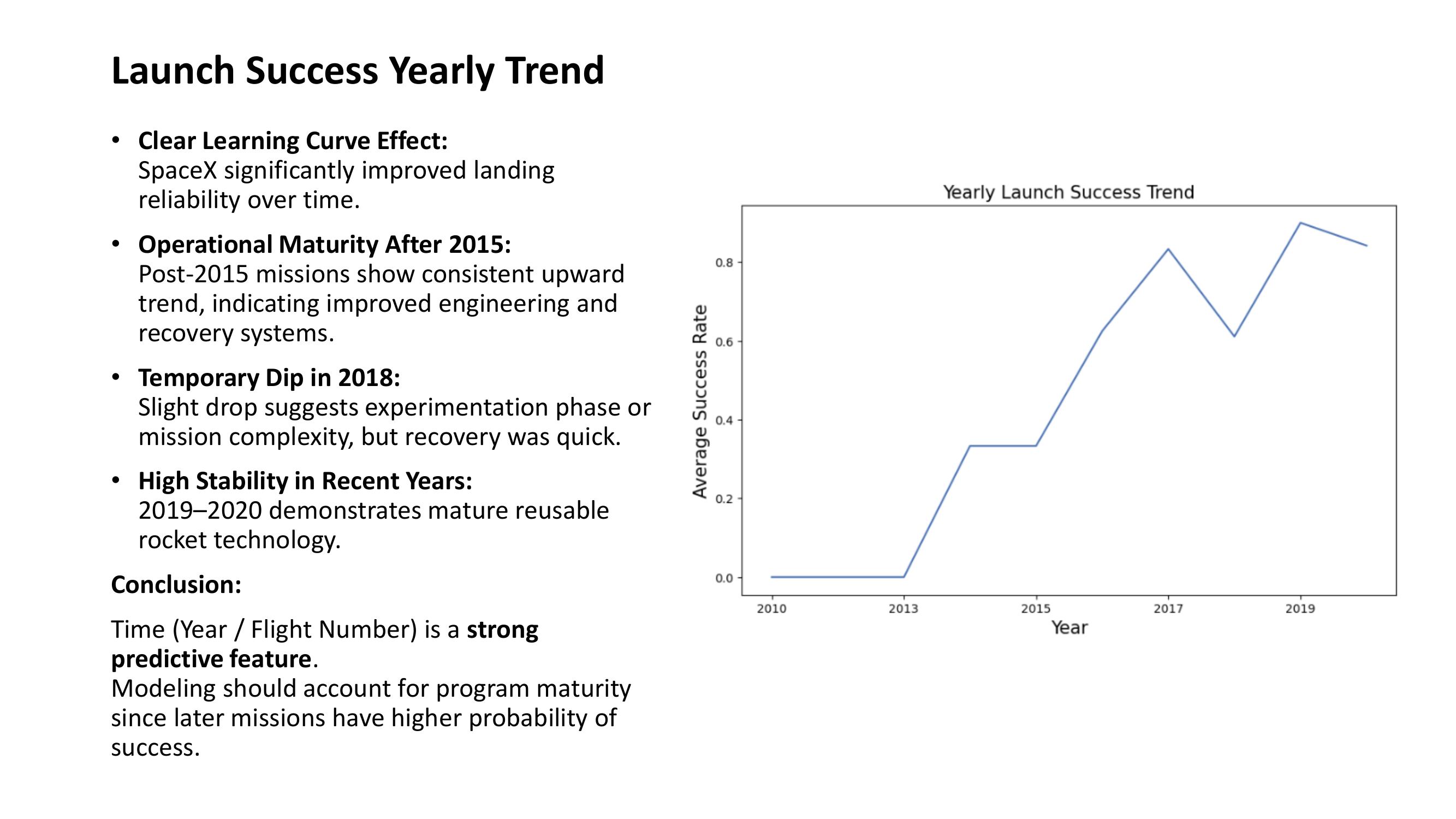


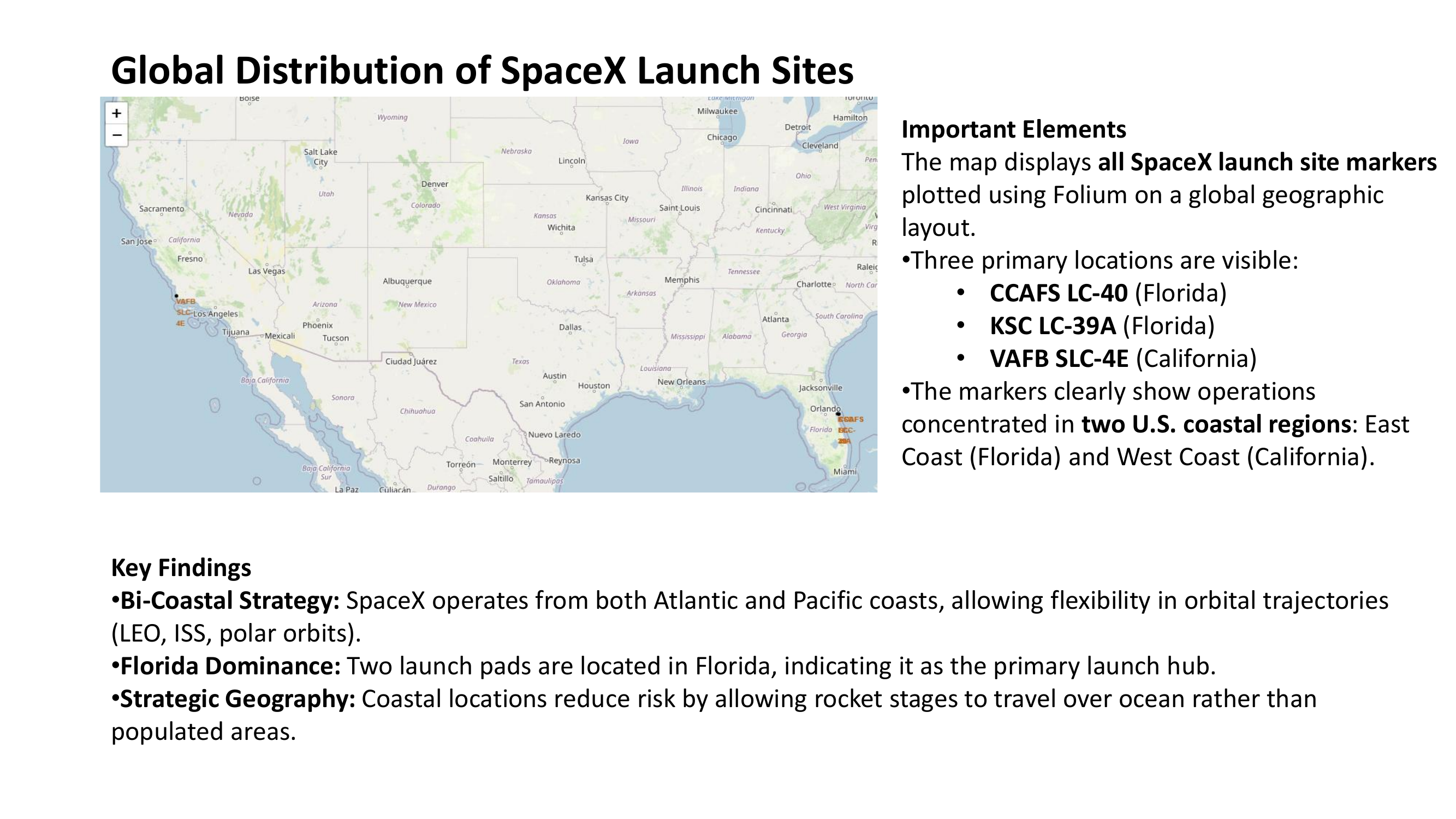


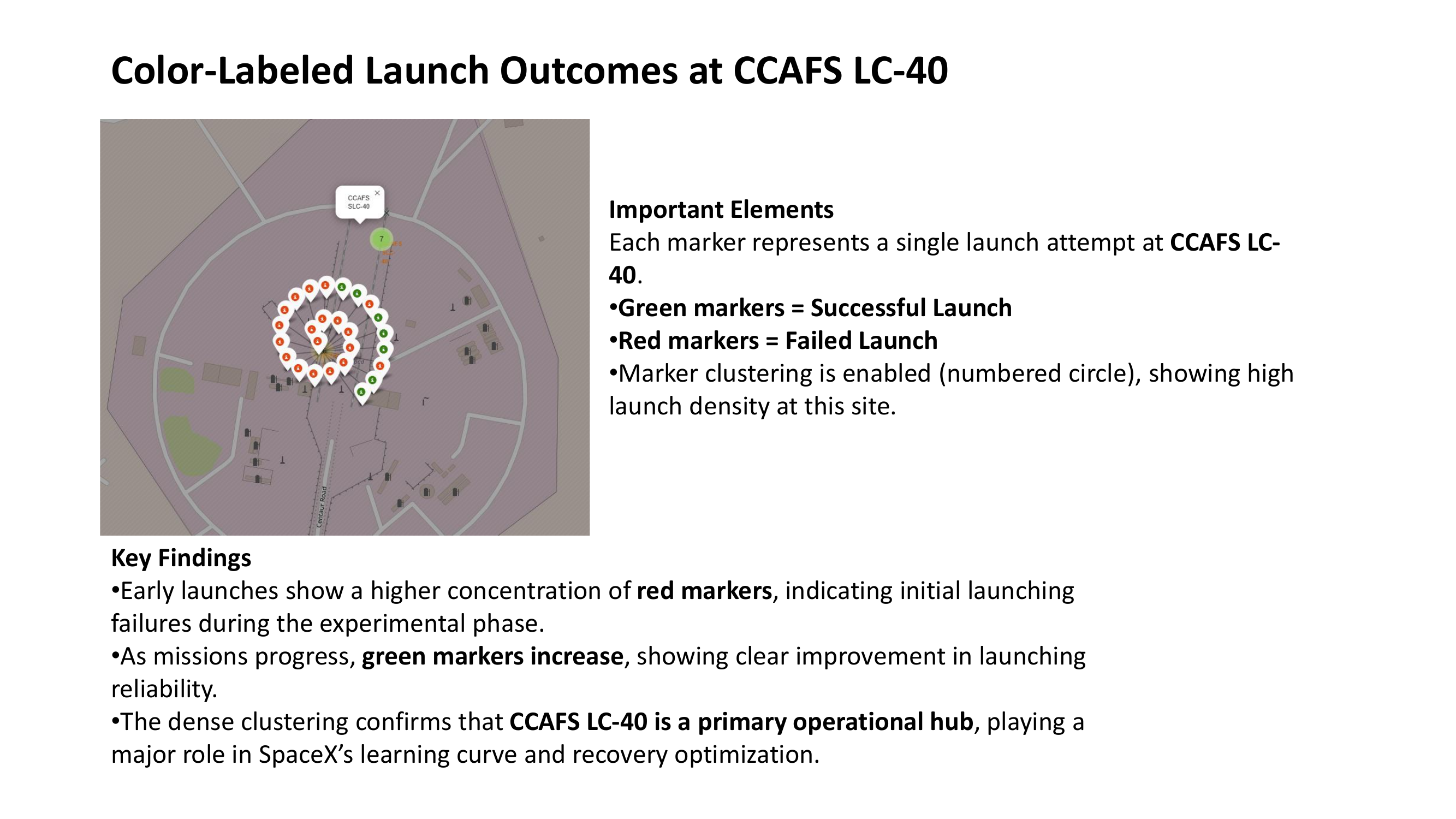


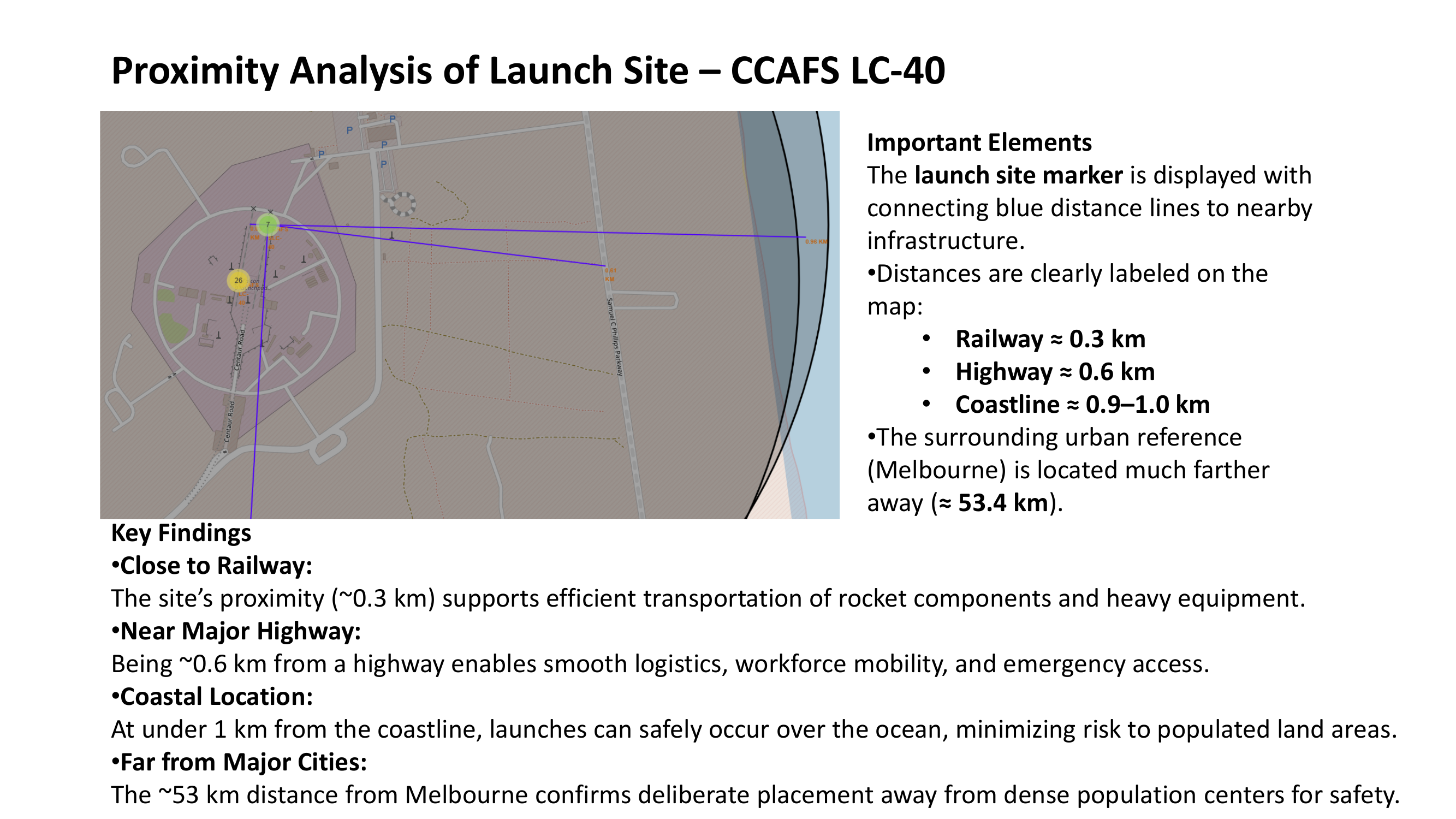


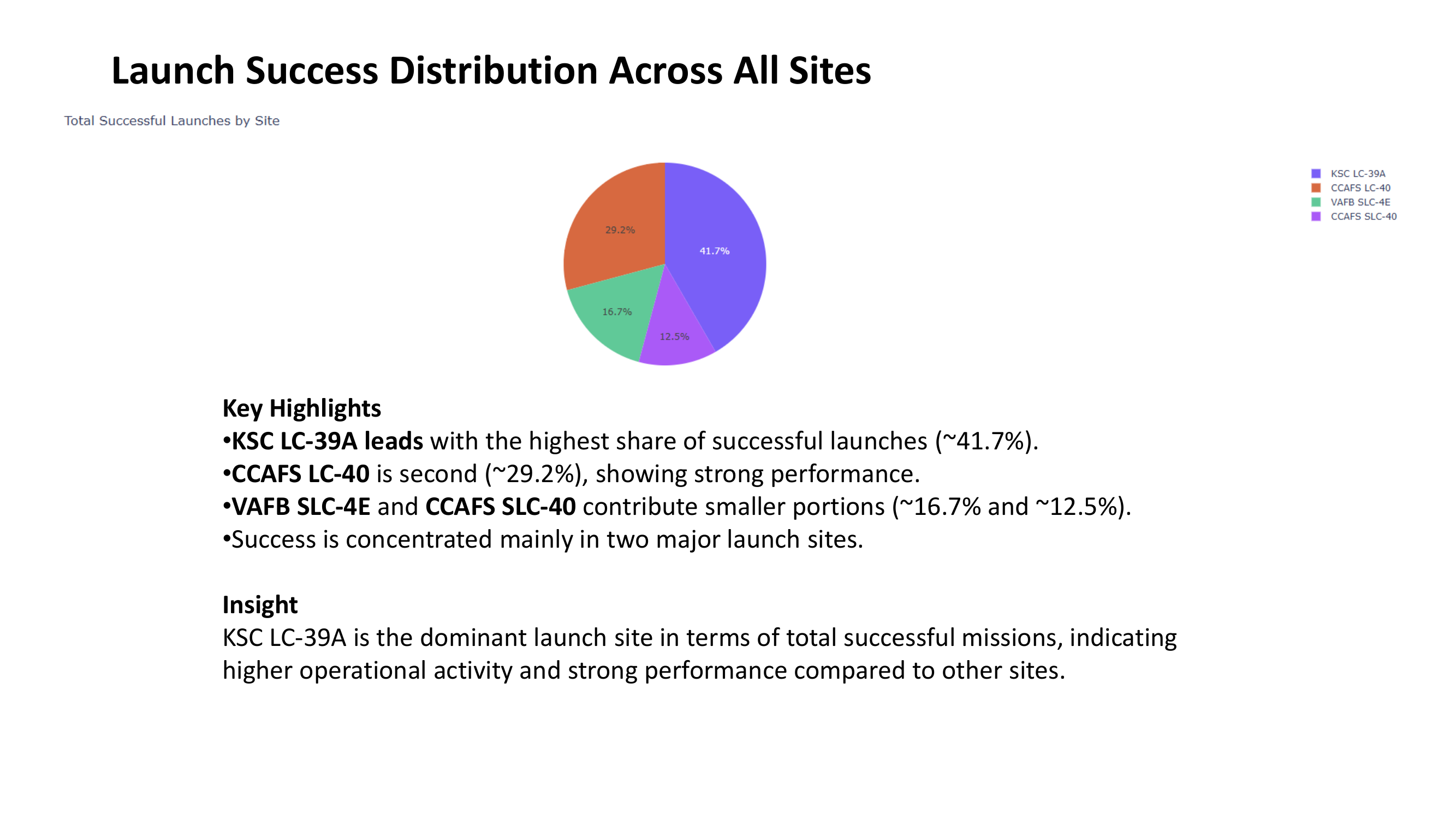


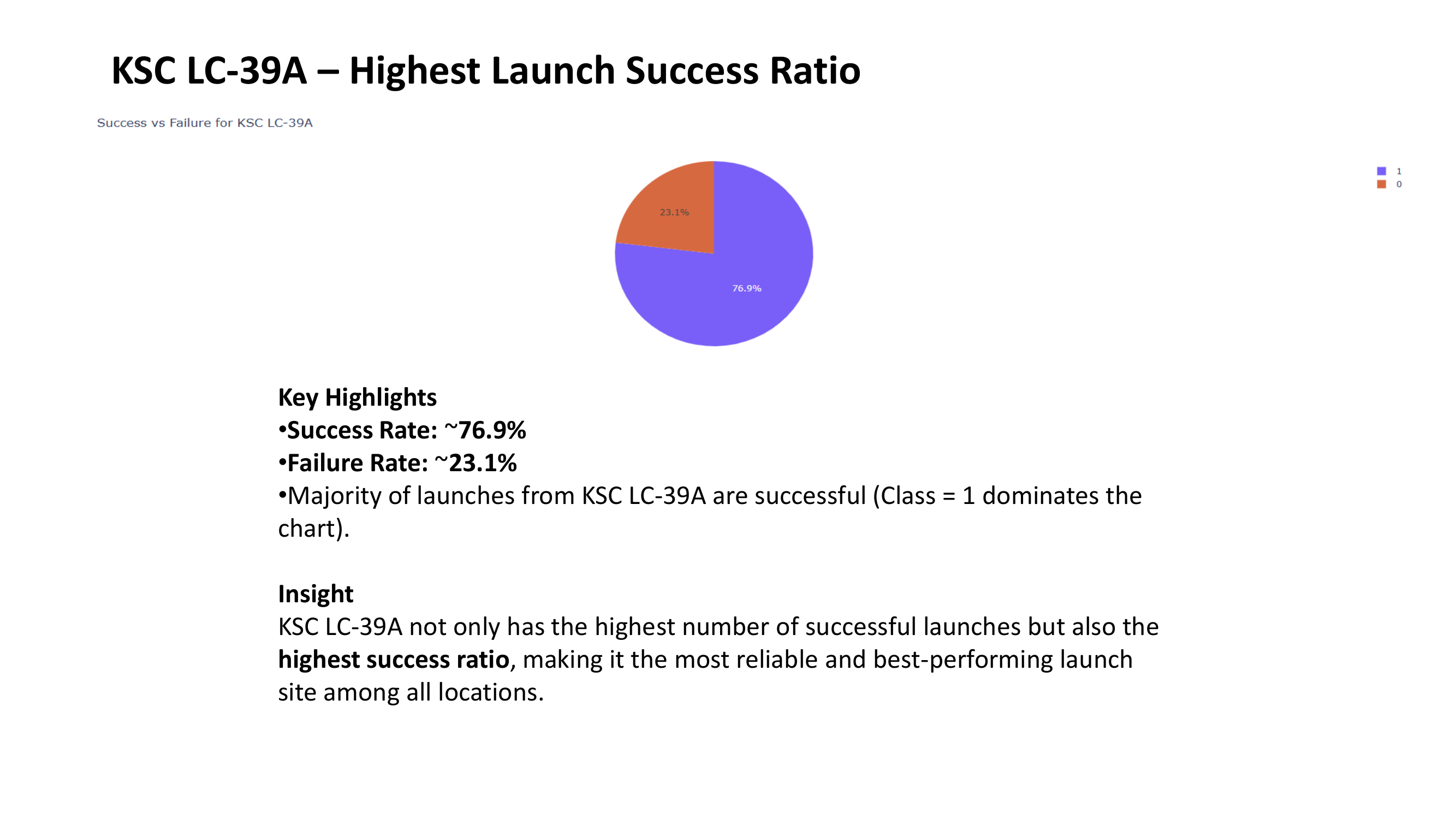


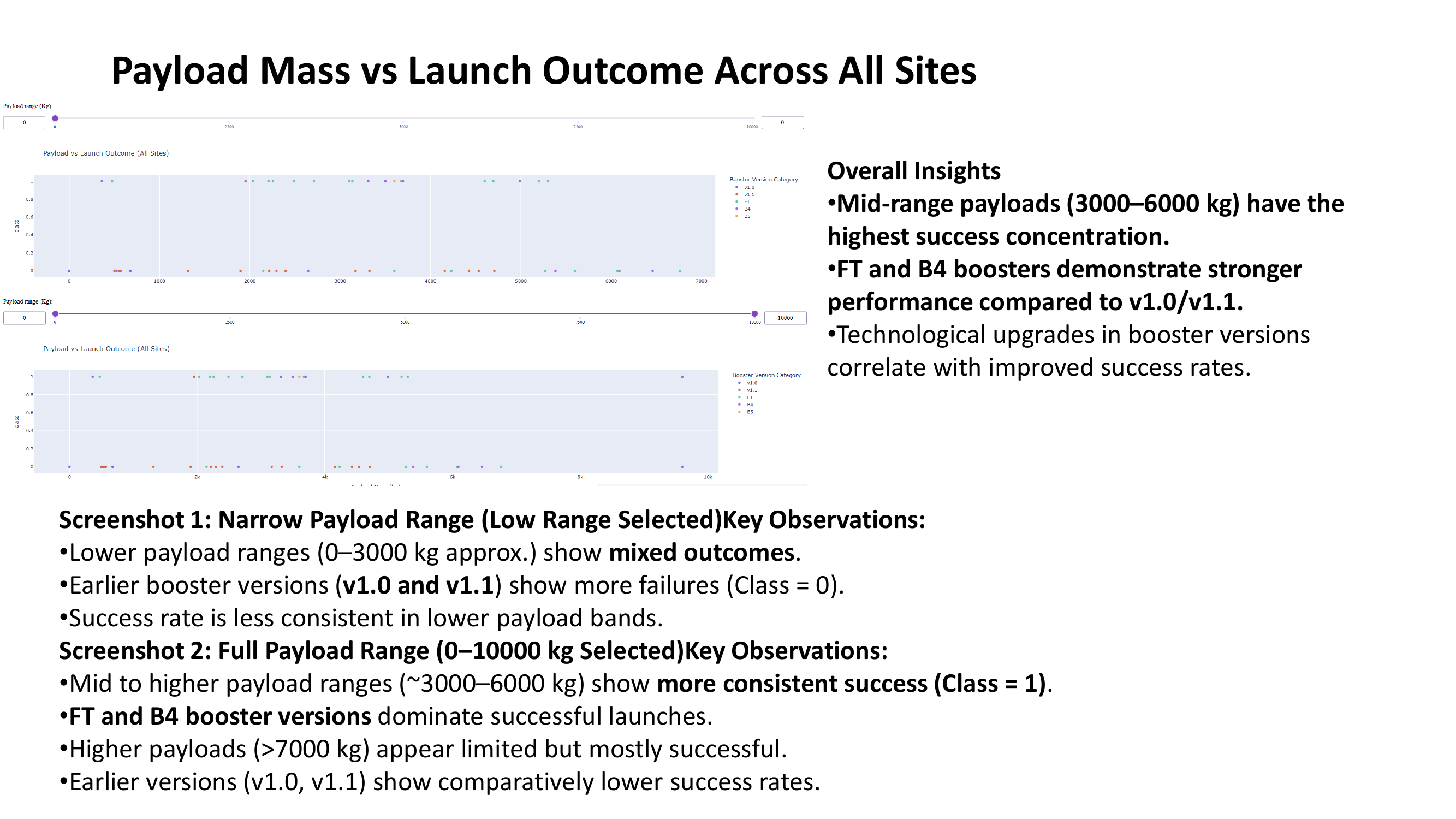




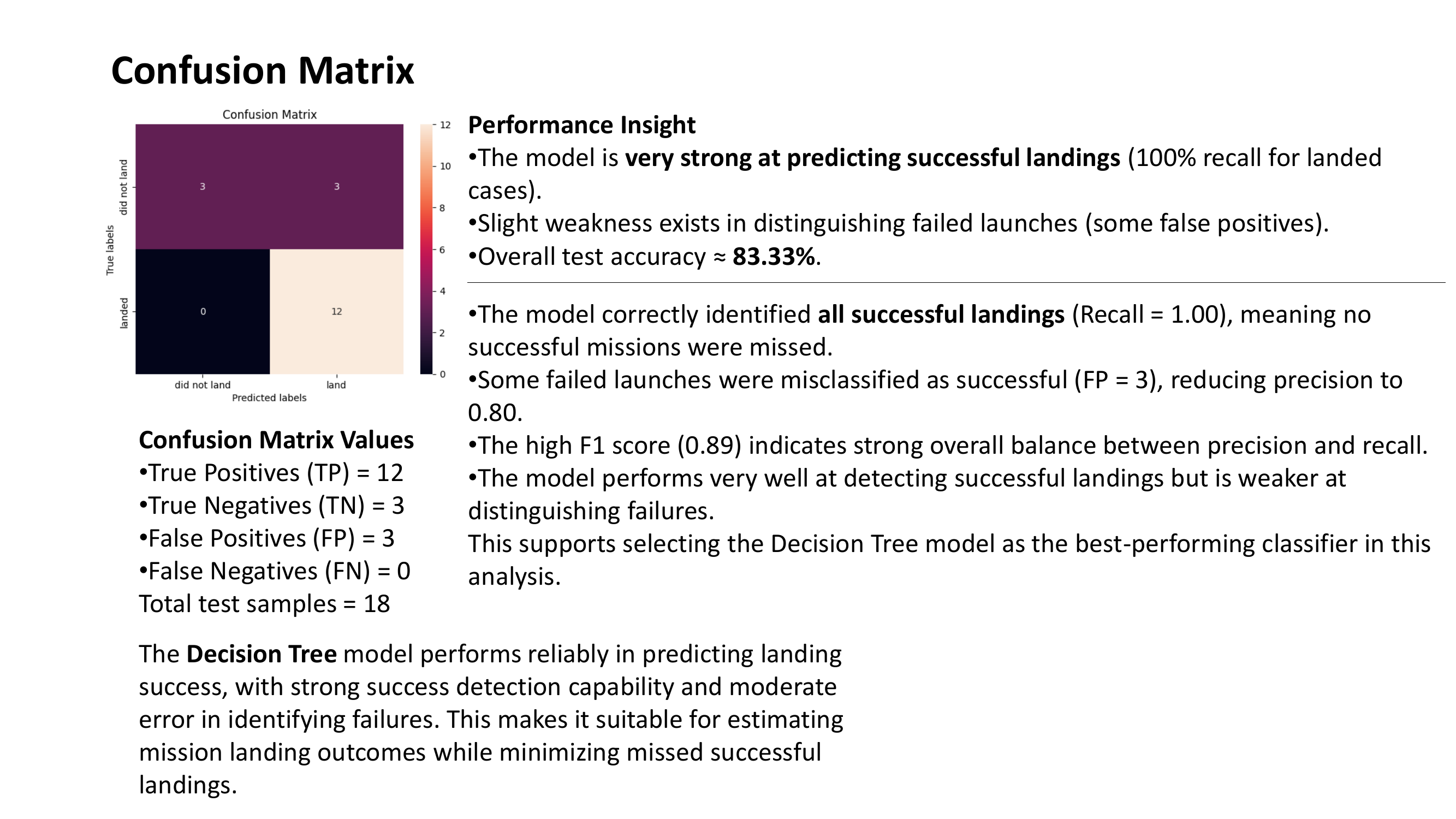












# 4. Results

Landing success improved significantly over time, demonstrating a strong learning curve effect.

KSC LC-39A emerged as the most reliable launch site, while GTO missions exhibited lower landing success due to higher mission complexity.

**Model Performance:**

• Test Accuracy (All Models): ~83.33%

• Best Model: Decision Tree

• Cross-Validation Accuracy: ~88.9%

• Confusion Matrix: TP=12, TN=3, FP=3, FN=0

• Recall (Success Detection): 1.00

• F1 Score: ~0.89

# 5. Discussion

The analysis confirms that booster version evolution, orbit complexity, payload interaction, and operational maturity significantly influence landing success.

# Predictive Modeling Performance

All four models achieved comparable test accuracy (~83.33%). Decision Tree demonstrated the strongest cross-validation performance (~88.9%) and perfect recall (1.00) for successful landings.

Confusion Matrix Summary:  
TP = 12  
TN = 3  
FP = 3  
FN = 0  
  
Model shows strong success detection capability with moderate false positives.

# 6. Conclusion

This project demonstrates how data science and machine learning can support aerospace decision-making and performance optimization. Predictive models can assist in mission risk assessment, cost estimation, and operational planning.

The findings validate that technological evolution and accumulated operational experience have significantly improved Falcon 9 landing reliability.

# 7. References

• SpaceX REST API Documentation

• Wikipedia Falcon 9 Launch Records

• Scikit-learn Documentation

• IBM Applied Data Science Capstone Labs