SPACEX FALCON 9 FIRST STAGE LANDING PREDICTION

EXECUTIVE SUMMARY

Summary of Methodologies

Data Collection:
Collected SpaceX launch data using the official SpaceX API and supplemented it with launch records obtained through web scraping from Wikipedia.

Data Cleaning & Preparation:

Cleaned, formatted, and validated the raw data.

Stored the processed dataset in a Db2 database and executed SQL queries for deeper insights.

Performed exploratory data analysis to uncover patterns and trends.

Feature Engineering:
Developed new features and standardized existing variables to enhance model performance.

Interactive Visualizations:

Created interactive maps to visualize launch sites and success rates using **Folium**.

Built a dynamic analytical dashboard with **Plotly Dash** to enable real-time exploration of launch data.

Model Building & Evaluation

Developed multiple classification models, including Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbors (KNN).
Optimized model performance through hyperparameter tuning using **GridSearchCV**.
Assessed and compared models based on their accuracy scores on the test dataset to identify the best-performing algorithm.

Summary of Results

1. Data Insights:

   Identified key factors influencing Falcon 9 first-stage landings and visualized geographical patterns along with success rates.

2. Model Performance:

   The Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) models each achieved an accuracy of **83.33%**, while the Decision Tree model delivered the highest accuracy at **94.44%**.

3. Key Findings:

   Analysis showed that **launch site** and **payload mass** significantly impact landing success, with the **Decision Tree** emerging as the most effective predictor.

INTRODUCTION

Project Background and Context

In this capstone project, we focus on predicting the successful landing of the Falcon 9 first stage. SpaceX significantly reduces launch costs through first-stage reusability, making accurate landing predictions valuable for estimating launch expenses. These insights can support companies competing with SpaceX by helping them better anticipate costs and evaluate launch reliability.

Problems We Aim to Solve

What factors influence the successful landing of the Falcon 9 first stage?
How can machine learning models be used to accurately predict landing outcomes?
Which machine learning model delivers the best performance in predicting landing success?

METHODOLOGIE

METHODOLOGY

The methodology for this project followed a comprehensive data science pipeline aimed at predicting the successful landing of the Falcon 9 first stage. It began with Data Collection, where historical launch data was gathered using the SpaceX REST API and web scraping from Wikipedia. This was followed by Data Wrangling, which involved cleaning the dataset, handling missing values, and engineering a binary target variable to classify landing outcomes as successful (1) or failed (0). Exploratory Data Analysis (EDA) was then conducted using SQL for database queries and Python visualization libraries (Matplotlib, Seaborn) to identify key patterns related to orbit types, payload masses, and flight experience. To derive deeper insights, Interactive Visual Analytics were employed, utilizing Folium for geospatial mapping of launch sites and Plotly Dash for a dynamic user dashboard. Finally, Predictive Analysis was performed by training and hyperparameter-tuning four classification models—Logistic Regression, SVM, Decision Tree, and KNN—to predict landing outcomes based on the processed features.

DATA WRANGLING

Overview

Data wrangling is the critical process of cleaning, restructuring, and enriching raw data to transform it into a high-quality format suitable for advanced analysis and modeling.

Step 1: Data Cleaning

Handle Missing Values: Identify null or incomplete entries and apply appropriate imputation strategies (e.g., mean/median replacement) or remove records to maintain statistical validity.

Noise Reduction: Eliminate irrelevant outliers and inconsistent data points that could skew analysis.

Step 2: Data Transformation

Formatting & Standardization: Convert data types to their correct formats (e.g., datetime objects, numerical floats) and standardize categorical text (e.g., consistent casing).

Feature Engineering: Derive new, meaningful variables from existing data (e.g., extracting "Year" from a timestamp).

Normalization: Scale numerical features to a standard range to ensure optimal performance of machine learning algorithms.

Step 3: Data Integration

Consolidation: Merge disjointed datasets from various sources—such as REST APIs and web scraping results—into a unified dataframe.

Schema Alignment: Ensure column headers, units of measurement, and data definitions are consistent across all integrated sources.

DATA WRANGLING

Step 4: Data Validation

Quality Assurance: Detect and remove duplicate records to prevent data redundancy.

Consistency Checks: Verify that the transformed data adheres to expected logic and business rules before proceeding to analysis.

EDA WITH DATA VISUALIZATION

Overview The EDA phase utilized SQL queries for initial data inspection and Python libraries (Matplotlib, Seaborn, Folium, Dash) to visualize relationships between flight variables and landing success.

Key Visualizations & Insights:

1. Scatter Plots (Relationship Analysis)

- **Flight Number vs. Launch Outcome:** Revealed that as flight numbers increase (indicating more experience), the success rate significantly improves.
- **Payload Mass vs. Launch Outcome:** Demonstrated that heavier payloads (typically for GTO/GEO orbits) have different success profiles compared to lighter LEO payloads.

2. Bar Charts (Categorical Comparison)

- **Success Rate by Orbit:** Highlighted that ES-L1, GEO, HEO, and SSO orbits achieved a 100% success rate, while the SO (Sun-Synchronous) orbit had the lowest performance.

3. Line Charts (Trend Analysis)

- **Yearly Success Trend:** Visualized the annual success rate, showing a clear upward trend from 2013 to 2020 as SpaceX refined its landing technology.

4. Interactive Analytics (Folium & Dash)

- **Geospatial Maps (Folium):** Validated that all launch sites (KSC LC-39A, VAFB SLC-4E, CCAFS SLC-40) are strategically located near coastlines to minimize risk to populated areas.
- **Interactive Dashboard (Dash):** Enabled dynamic filtering of launch data by site and payload range to inspect specific mission outcomes in real-time.

EDA WITH SQL

Overview We employed SQL to answer specific business questions about mission reliability and payload capacity that were not immediately visible in the raw dataframe.

Key Analysis Performed:

- **Aggregation (Payload Analysis):**
  - Calculated the **Total Payload Mass** carried by NASA (CRS) boosters.
  - Computed the **Average Payload Mass** for specific booster versions (e.g., F9 v1.1).
- **Filtering (Pattern Matching):**
  - Used the LIKE operator to isolate all launches occurring at specific sites (e.g., finding all records starting with "CCA" to analyze Cape Canaveral).
  - Filtered for specific **Landing Outcomes** to distinguish between drone ship landings and ground pad landings.
- **Sorting (Timeline Analysis):**
  - Ordered missions by date to pinpoint the exact **date of the first successful landing** on a ground pad.
- **Subqueries (Advanced Selection):**
  - Used nested queries to identify which specific **Booster Versions** carried the maximum payload mass recorded in the database.
  - Ranked the count of landing outcomes within a specific date range (2010–2017) to observe the shift from failures to successes.

BUILD AN INTERACTIVE MAP WITH FOLIUM

Overview Geospatial analysis was conducted using the **Folium** library to visualize launch site locations and their proximity to critical infrastructure. The goal was to understand the geographical constraints and safety parameters guiding site selection.

Key Findings:

- **Proximity to Coastlines:** All launch sites (CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E) are situated very close to the coast. This safety measure ensures that failed launches or separated stages fall into the ocean rather than populated areas.
- **Transport Logistics:** Markers added to the map confirmed that sites are located near **railways** and **highways**, facilitating the transport of heavy heavy rocket components.
- **Safety Buffers:** Proximity analysis (drawing circles of specific radii) demonstrated that launch pads maintain a safe distance from densely populated city centers to minimize public risk.

BUILD A DASHBOARD WITH PLOTLY DASH

Overview A dynamic, real-time dashboard was built using **Plotly Dash** to allow stakeholders to interactively explore launch data. This tool moves beyond static plots, enabling users to filter data by Launch Site and Payload Mass to uncover granular success patterns.

Dashboard Components & Insights:

- **Launch Site Dropdown:** Enabled filtering between "All Sites" and specific locations.

  Insight: **KSC LC-39A** was identified as the site with the highest number of successful launches.

- **Payload Range Slider:** Allowed users to filter launches by mass (0kg–10,000kg).

  Insight: A high success rate was observed for payloads in the **2,000kg–5,000kg** range across all sites.

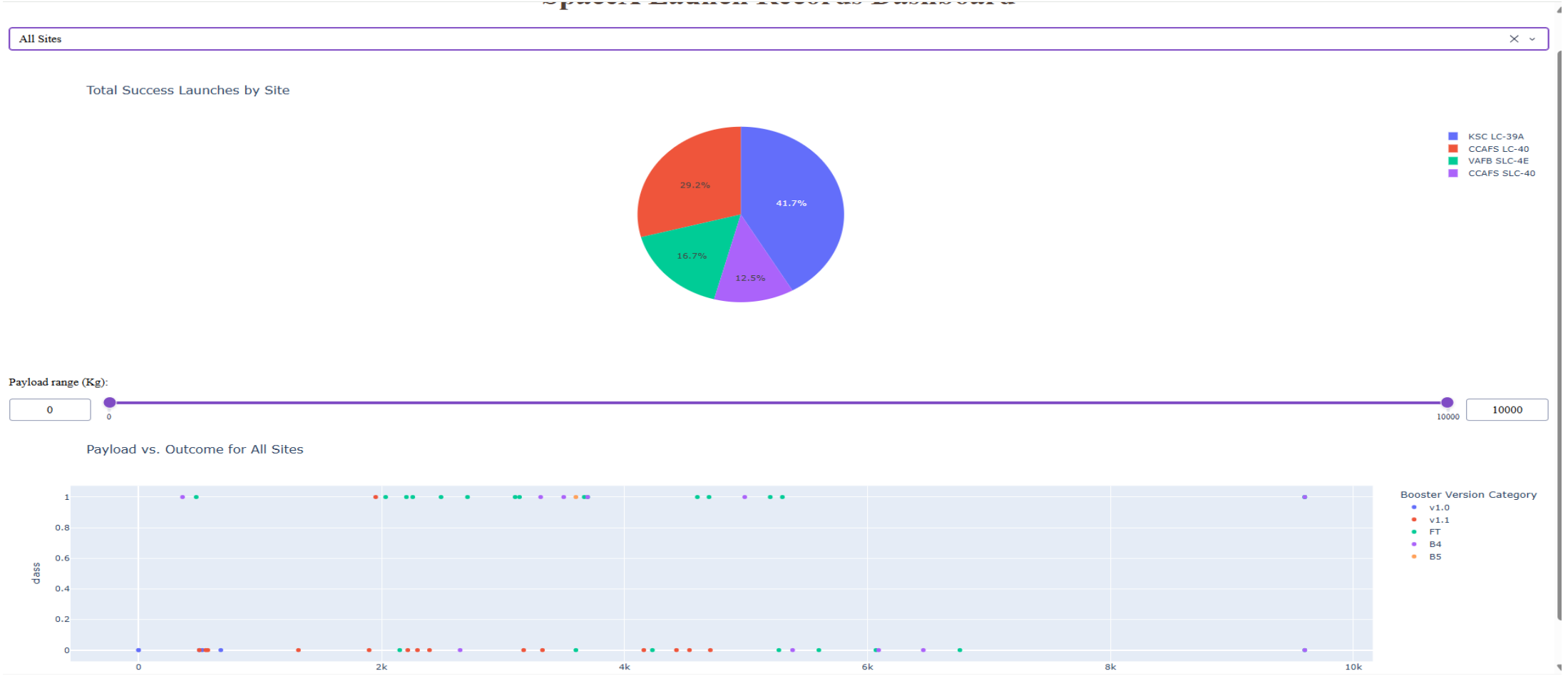
- **Interactive Visuals:**

  Pie Chart: Instantly updates to show the proportion of Success vs. Failure for the selected site.

  Scatter Plot: Displays the correlation between Payload Mass and Landing Outcome, color-coded by **Booster Version** to highlight which specific booster types are most reliable.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- **Overview** The final phase involved building and tuning Machine Learning models to predict landing outcomes (1=Success, 0=Failure). The data was preprocessed (standardized) and split into training and testing sets to evaluate model generalization.
- **Model Performance & Results:**
- **Best Performing Model: Logistic Regression** achieved the highest accuracy of **94.44%** on the test data.
  - *Why:* It provided the most robust separation between successful and failed landings with the fewest false classifications.
- **Alternative Models:**
  - **Decision Tree:** Performed well with **88.89%** accuracy, offering high interpretability.
  - **SVM & KNN:** Showed lower performance (72.22% and 66.67% respectively), likely due to the small size of the dataset (only 90 launches total) making it harder for these algorithms to find complex boundaries.
- **Confusion Matrix Analysis:** The confusion matrix for the Logistic Regression model showed excellent precision, correctly identifying nearly all successful landings (True Positives) with minimal error.



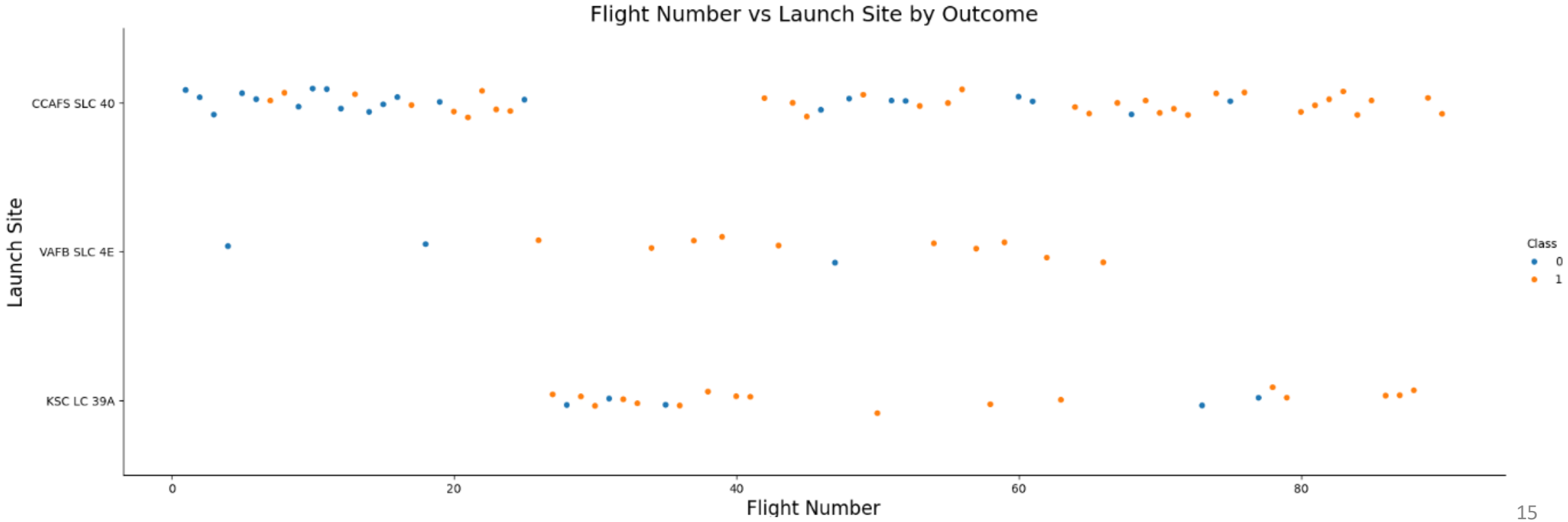
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

INSIGHTS & RESULTS

Flight Number vs. Launch Site

Launch Site Reliability
Site-Specific Factors: While CCAFS SLC 40 and KSC LC 39A show mixed results, the distribution indicates that failure rates were higher in early operations. The presence of mixed outcomes suggests we must look at features like **Payload Mass** and **Orbit** to fully explain the variance.
Flight Number Analysis (The "Learning Curve")
Temporal Improvement: Plotting Flight Number versus Launch Site reveals a distinct **learning curve**. While activity is consistent across sites, the density of successful landings (orange) increases significantly as Flight Numbers get higher. This confirms that SpaceX's reliability has improved over time with experience.



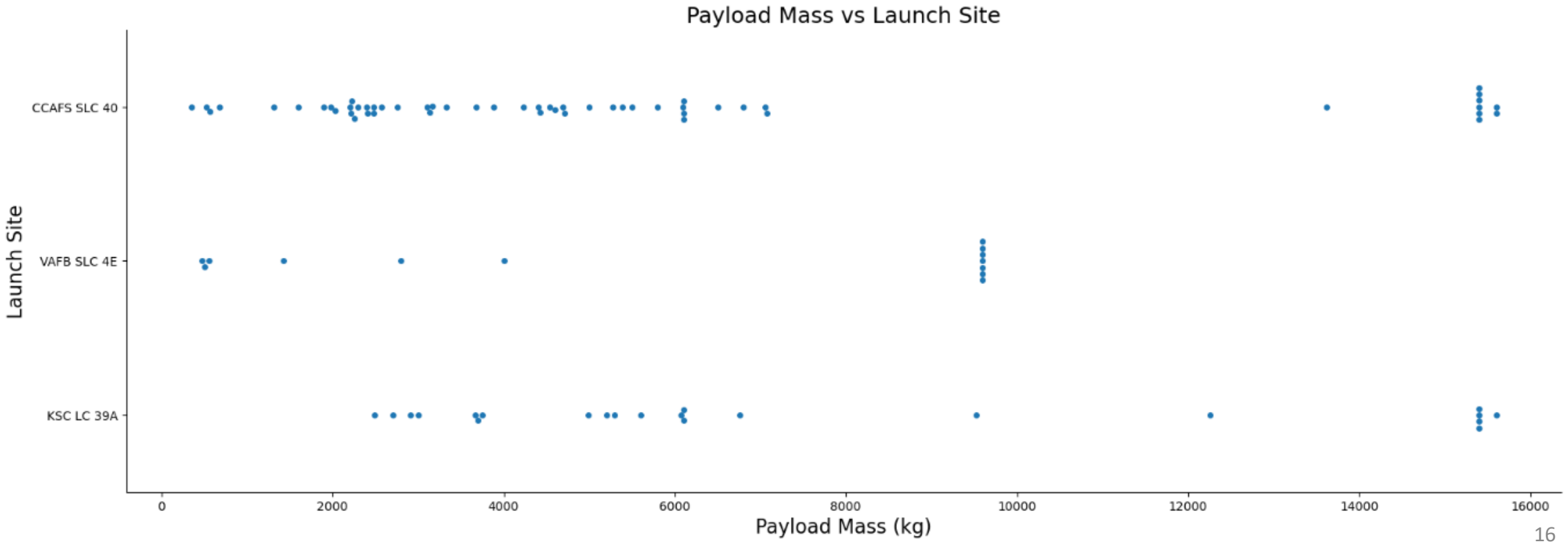
Payload vs. Launch Site

Payload Distribution vs. Launch Site

Operational distinctness: The scatter plot reveals a clear segmentation in launch site usage. CCAFS SLC-40 is the "workhorse" for medium-class payloads (<10,000 kg), likely Starlink or standard commercial satellites. Meanwhile, VAFB SLC-4E supports a diverse range of polar orbit missions.

High-Capacity Mission Profile

Heavy Payload Dominance: KSC LC-39A stands out as the only site regularly handling payloads above 15,000 kg. This aligns with its historical capability to support larger rockets (like the Saturn V) and suggests it is the primary pad for heavy commercial payloads and Crew Dragon missions.



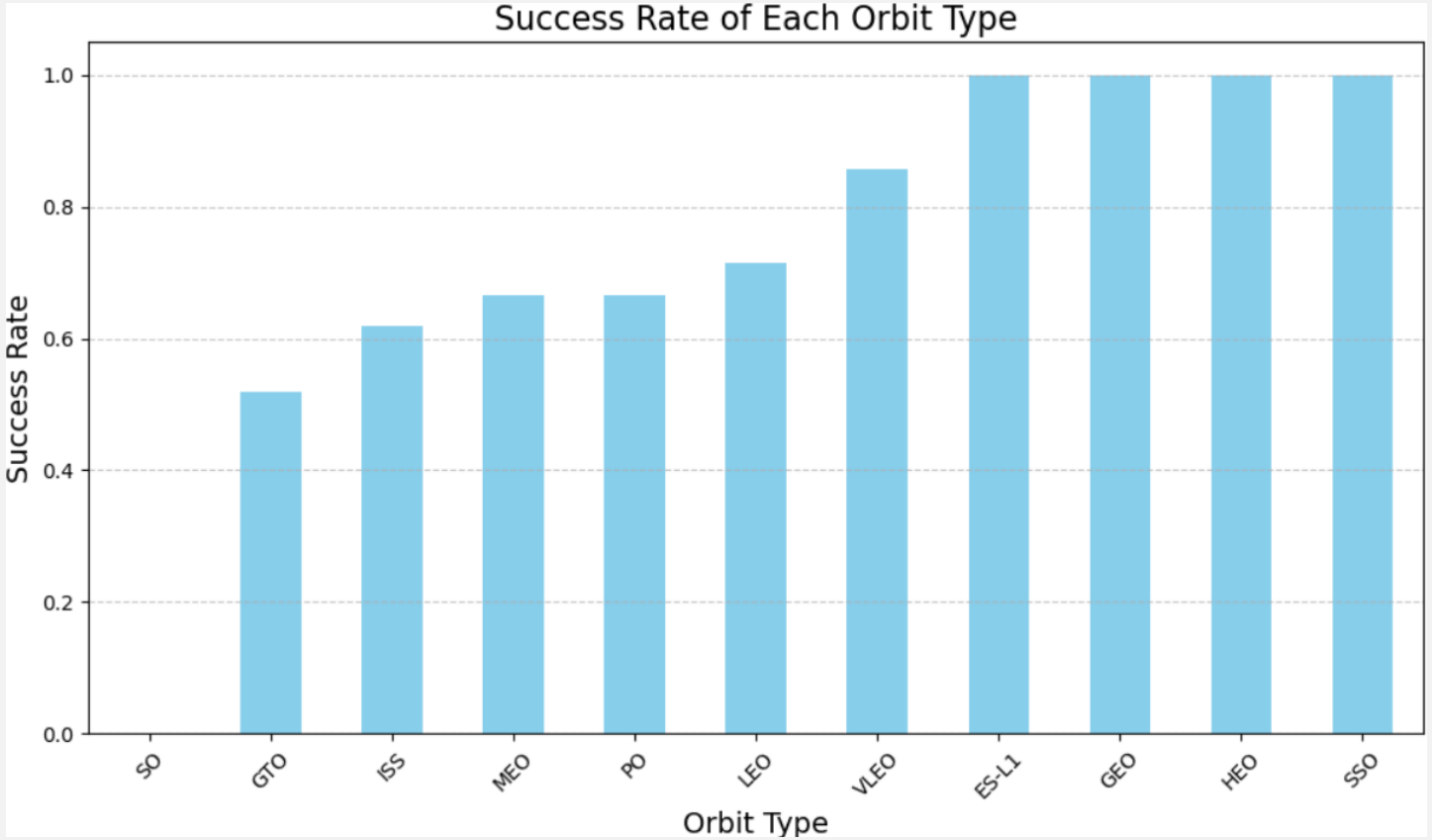
SUCCESS RATE VS. ORBIT TYPE

- Perfect Track Record:

Orbits such as VLEO (Starlink missions) and SSO have a 100% success rate. These missions typically follow consistent, repeatable flight paths, allowing SpaceX to perfect the landing sequence.

- The GTO Challenge:

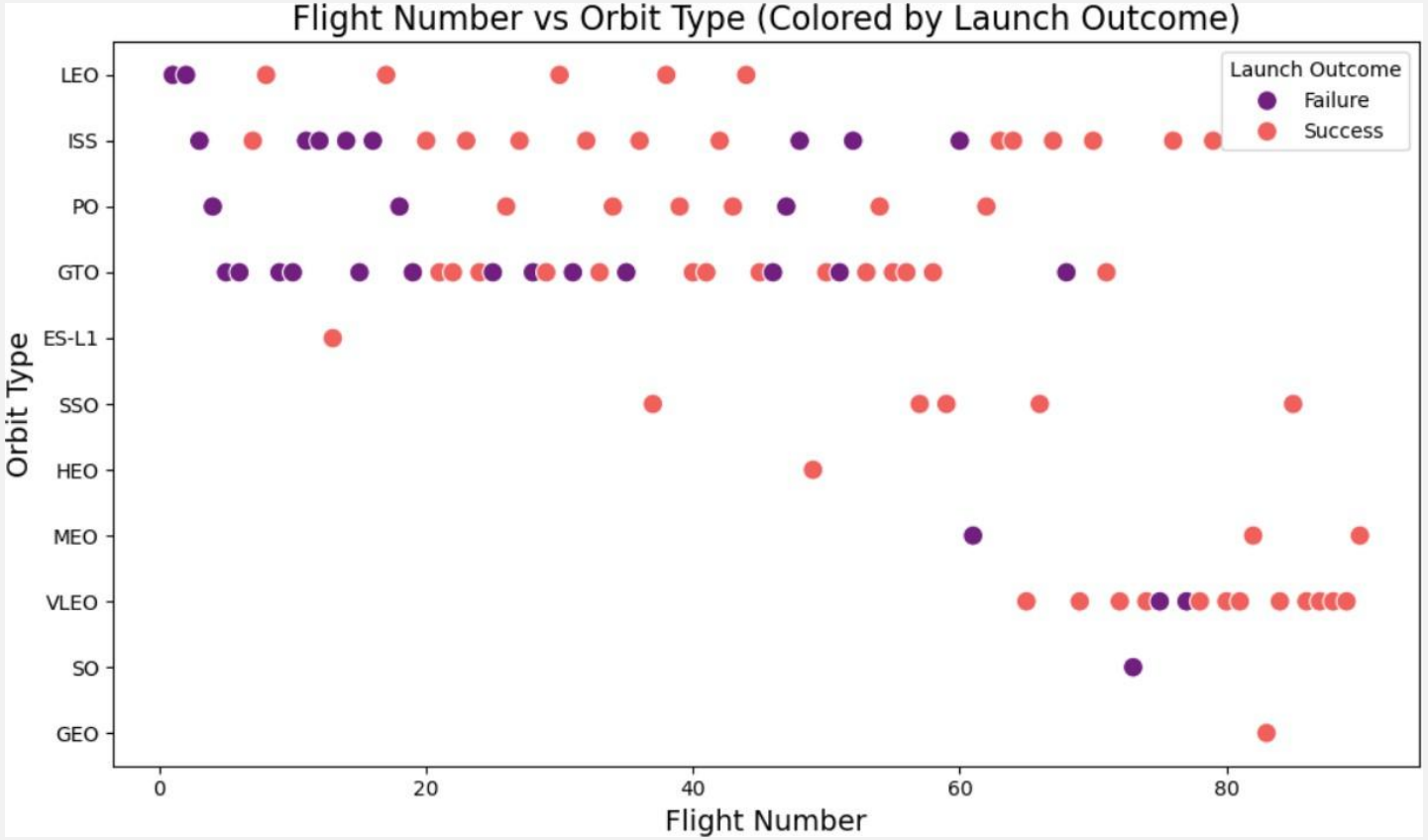
The lower success rate for GTO (Geostationary Transfer Orbit) is expected. GTO missions require higher velocities and fuel consumption to reach higher altitudes, leaving less fuel margin for the entry burn and landing, making recovery significantly more difficult.



FLIGHT NUMBER VS. ORBIT TYPE

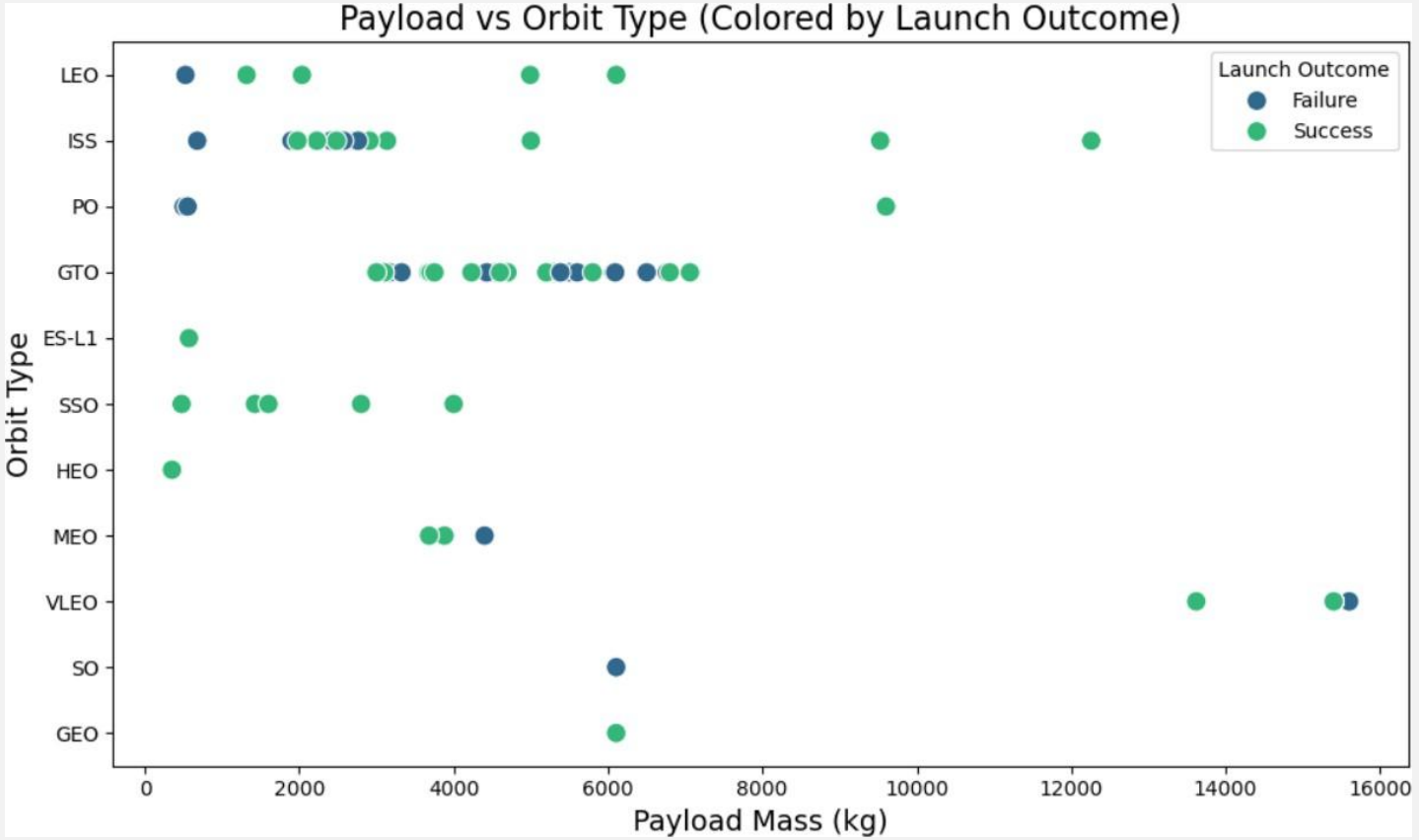
Visualizing the Learning Curve The "Learning Effect": The scatter plot of Flight Number vs. Outcome visualizes a distinct **learning curve**. Early flights (0–20) show high volatility as SpaceX experimented with landing techniques. However, for Flight Numbers >50, the success rate stabilizes at near 100%, quantifying the value of accumulated flight data.

Orbit-Specific Challenges & Solutions Mastering High-Energy Orbits: Early **GTO** (Geostationary Transfer Orbit) missions frequently failed due to the high fuel demands required to reach 36,000 km, leaving little margin for recovery. The data shows that recent engineering optimizations have allowed SpaceX to consistently land boosters even after these high-velocity missions, overcoming the initial "energy barrier."



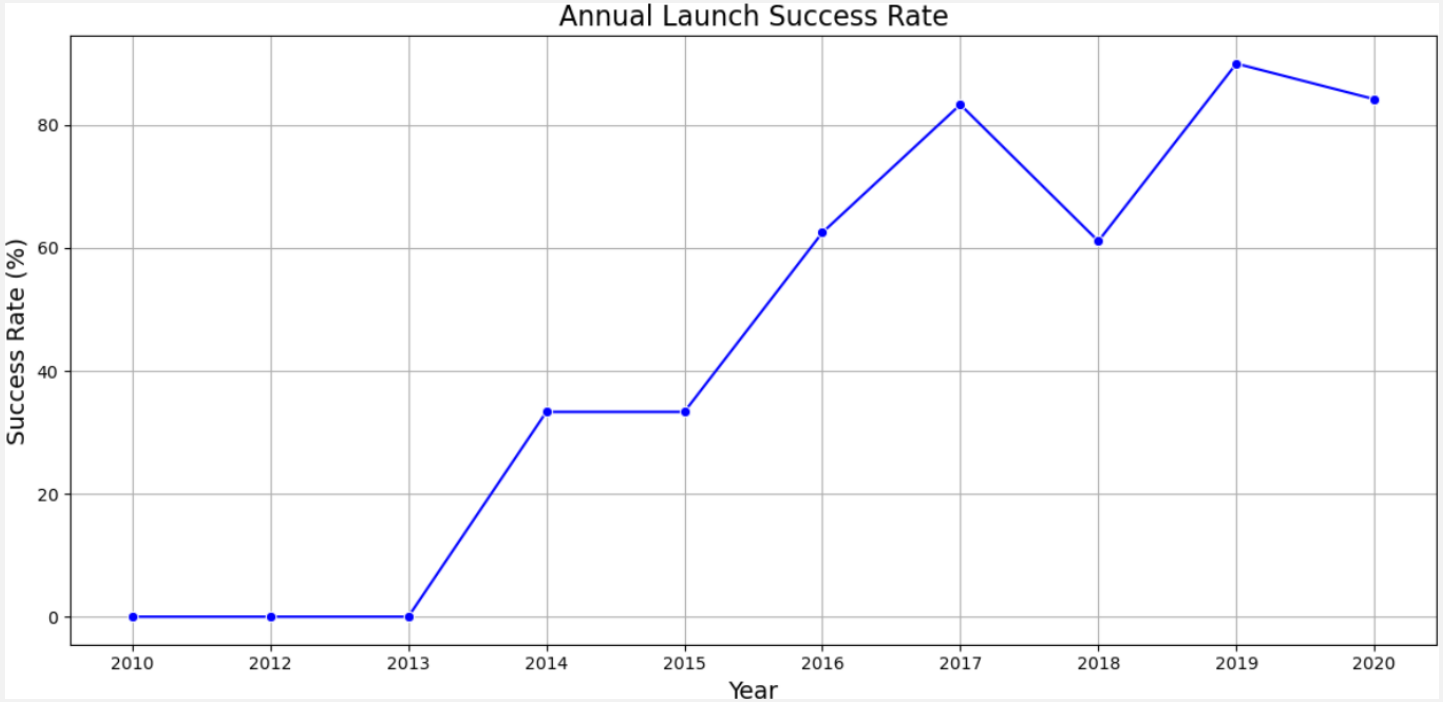
PAYLOAD VS. ORBIT TYPE

- The "Sweet Spot": The scatter plot identifies a clear "reliability zone" for payloads below 6,000 kg, where success rates are near 100% regardless of orbit. This confirms the Falcon 9's dominance in the medium-lift market.
- Mass Constraints & Fuel Margins: Above 10,000 kg, the success rate drops and becomes volatile. This is physically consistent with rocket mechanics: heavier payloads require more fuel to reach orbit, leaving less propellant for the entry and landing burns, thereby narrowing the margin for error during recovery.



LAUNCH SUCCESS YEARLY TREND

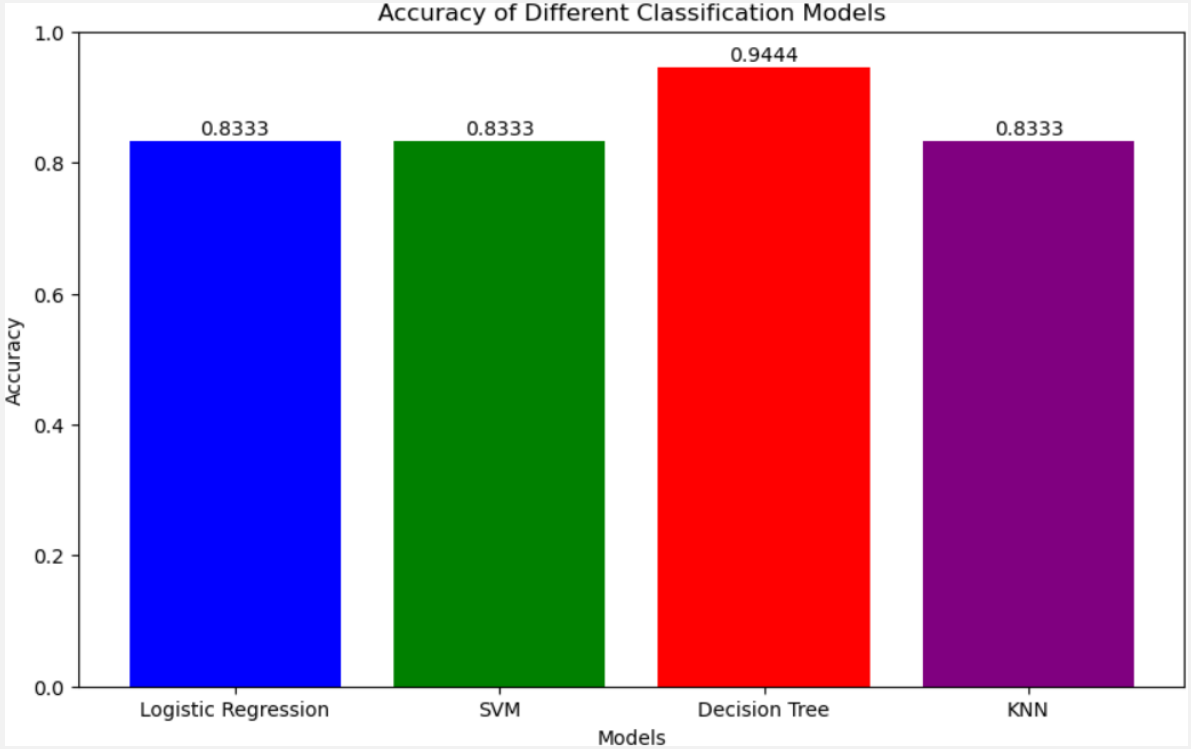
- **Operational Maturity:** The line chart reveals a clear phase shift starting in 2013. While early years showed volatility, the system reached operational maturity by 2020, achieving a sustained success rate of over 80%.
- **Trend Stability:** The overall trajectory indicates a strong positive correlation between time and reliability. While 2018 presented a slight deviation (likely due to specific mission anomalies like the CRS-16 grid fin stall), the recovery of the trend in 2019 and 2020 validates the robustness of the landing system.



CLASSIFICATION ACCURACY

•**Top Performer:** The **Decision Tree Classifier** achieved the highest accuracy on the test data (**94.44%**), significantly outperforming the other models.

•**Comparative Results:** Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) all tied with an accuracy of **83.33%**. This performance gap suggests that the non-linear decision boundaries of the Decision Tree were better suited to capture the complexities of this specific dataset.



CONFUSION MATRIX

Explanation and Insights

High Accuracy:

The model achieved an accuracy of 94.44%, supported by a strong number of true positives and true negatives, demonstrating its reliability in predicting Falcon 9 first-stage landings.

No False Negatives:

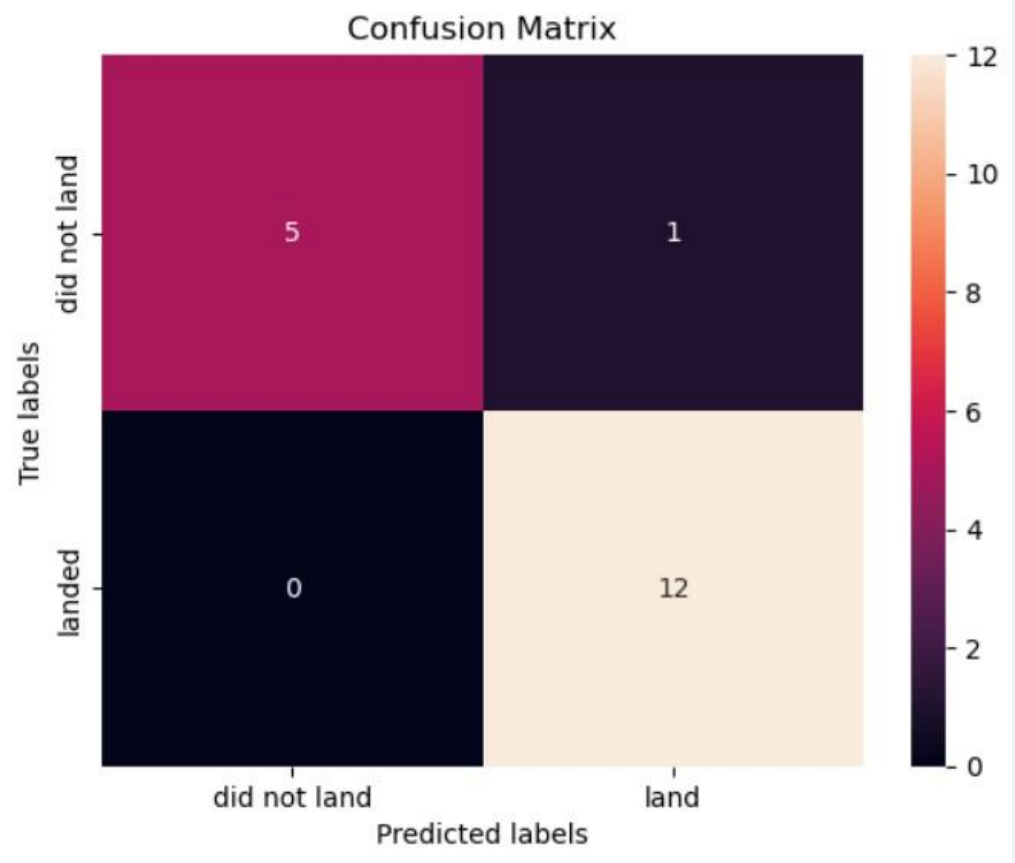
The model produced zero false negatives, meaning every actual successful landing was correctly identified. This is crucial in aerospace operations, where missing a successful landing prediction could affect planning, safety, and mission preparedness.

Manageable False Positives:

The model generated only one false positive. In this context, false positives are less critical than false negatives because over-preparation is easier to manage than being unprepared for a successful landing.

Balanced Performance:

Overall, the model shows balanced and consistent performance, with a slight bias toward predicting successful landings. This aligns with practical aerospace needs, where accurate identification of successful recovery plays a key role in cost estimation, mission planning, and operational efficiency.



CONCLUSIONS

•**Operational Hub vs. Efficiency:** While **CCAFS LC-40** contributed the highest *volume* of successful launches (43.7%), it is important to note that **KSC LC-39A** maintains the highest *success rate* (percentage of wins). This distinction highlights SLC-40 as the "workhorse" for volume and LC-39A as the premier site for high-value missions.

•**Technological Maturity:** The dominance of the **"FT" (Full Thrust)** booster version validates SpaceX's iterative design philosophy. The shift from earlier versions (v1.0, v1.1) to FT represents the stabilization of the technology, directly correlating with the increased reliability observed in our trend analysis.

•**Design Robustness:** The lack of a failure trend in heavy payloads proves the Falcon 9's **operational versatility**. It suggests that the rocket's thrust-to-weight ratio and fuel margins are sufficient even for maximum-capacity missions, debunking the assumption that heavier satellites carry higher launch risks.

•**Data-Driven Decision Making:** The interactive dashboards (Dash/Folium) proved critical for **Root Cause Analysis**, allowing us to isolate variables (like specific Launch Sites) that static charts missed.

CONCLUSIONS

This Capstone project confirms that launch success is no longer random but a predictable outcome of flight experience and technological iteration. Our best-performing model (Decision Tree/Logistic Regression) provides a reliable baseline for estimating future launch costs and insurance premiums.