IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bruno Trelles Sayán
October 19th 2025

Outline

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

Executive Summary

- We analyze SpaceX Falcon 9 missions to identify the drivers of first-stage landing success and build a classification model for predictive support. We collected data via SpaceX REST APIs and targeted web scraping, then cleaned, merged, and enriched features (payload mass, orbit, launch site, booster version, flight number, year). Using visualization (Matplotlib/Plotly) and SQL, we validated relationships (e.g., payload ranges and orbit types) and temporal trends. We created an interactive Folium map and a Plotly Dash dashboard for exploration and stakeholder communication. Finally, we trained multiple classifiers (Logistic Regression, SVM, KNN, Decision Tree) with cross-validation and hyper-parameter tuning; the best model achieves strong accuracy and balanced precision/recall, enabling data-driven insights for mission planning.

Introduction

- Business context. Reusability economics hinge on reliable first-stage recovery. Understanding conditions that favor successful landings reduces costs and guides mission planning.Objective. Use historical mission data to (1) explore patterns of success/failure and (2) build a model that predicts landing success before launch.Research questions. Which features most correlate with landing success? How do payload mass, orbit, and site interact? Can a classification model predict success reliably enough for decision support?Scope. Falcon 9 launches with recorded landing outcomes; analysis spans data collection, wrangling, EDA (viz + SQL), interactive analytics, and predictive modeling.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Queried SpaceX REST API (v4) for launches, rockets, cores, and payloads; exported raw JSON and normalized fields (launch ID, core ID, payload ID, site, orbit, payload mass).
- Perform data wrangling
  - Flattened nested JSON into relational tables; cast types; standardized units (kg), dates (UTC), and categorical labels (orbit, booster version).
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built classifiers (Logistic Regression, SVM, KNN, Decision Tree) to predict first-stage landing success.

Data Collection – SpaceX API

Process. Called SpaceX REST endpoints to retrieve launches, rockets, landing outcomes, cores, and payloads. Normalized JSON → tabular data; retained keys for joins (launch ID, core serial, payload ID).Flowchart (bullets).

- Authenticate/anonymous GET → /v4/launches, /v4/rockets, /v4/cores, /v4/payloads
- Parse → select fields → flatten nested objects
- Join on IDs → export to CSV/Parquet for wranglingNotebook (GitHub): <API notebook URL>

Data Collection - Scraping

Process. Scraped missing attributes (e.g., landing type notes, booster version labels, site names) from public mission pages where API lacked detail; respected robots.txt and added polite delays.
Flowchart. Request page → parse with BeautifulSoup → extract tables/divs → validate → append to dimension tables → dedupe.

Data Wrangling

Tasks.

- Type casting, null handling, unit harmonization (payload mass kg), and date normalization.
- Feature engineering: year, month, is_drone_ship, orbit_group (LEO/MEO/GEO/HEO), payload_bin (quantiles), and booster_family.
- Integrity checks: primary/foreign key consistency, duplicate removal, and outlier review.

EDA with Data Visualization

- Scatter: Flight Number vs. Launch Site (experience effect).
- Scatter: Payload vs. Launch Site (site capacity/mix).
- Bar: Success Rate by Orbit Type (mission profile effect).
- Scatter: Flight Number vs. Orbit; Payload vs. Orbit (mission complexity signals).
- Line: Yearly Success Rate Trend (learning curve over time).

EDA with SQL

- Unique launch sites, prefix matches (e.g., CCA%).
- Aggregates by customer/agency (e.g., total payload mass where customer LIKE 'NASA%').
- Averages by booster version (e.g., F9 v1.1).
- Date of first successful ground landing.
- Successful drone-ship landings with payload mass between 4000 and 6000.
- Mission outcome counts and ranking landing outcomes between two dates.

Build an Interactive Map with Folium

Map objects. Global markers for launch sites; colored circle markers by landing outcome; polylines from site to downrange landing areas; proximity overlays (railways/highways/coastline) with measured distances.

Why. Communicates spatial constraints and site-specific context unavailable in tables.

Build a Dashboard with Plotly Dash

- Pie: Success counts by launch site (global view).
- Pie: Site with highest success ratio (deep dive).
- Scatter: Payload vs. Outcome with range slider; dropdowns for site/orbit/booster.
- Why. Enables scenario filtering and pattern discovery by non-technical users.

Predictive Analysis (Classification)

- Target: landing_success (1/0).
- Features: payload mass, orbit (one-hot), launch site, booster version/family, flight number, year, landing platform type.
- Models: Logistic Regression, SVM (RBF), KNN, Decision Tree.
- Tuning: GridSearchCV with stratified K-fold; metrics: accuracy, precision, recall, F1; confusion matrix.

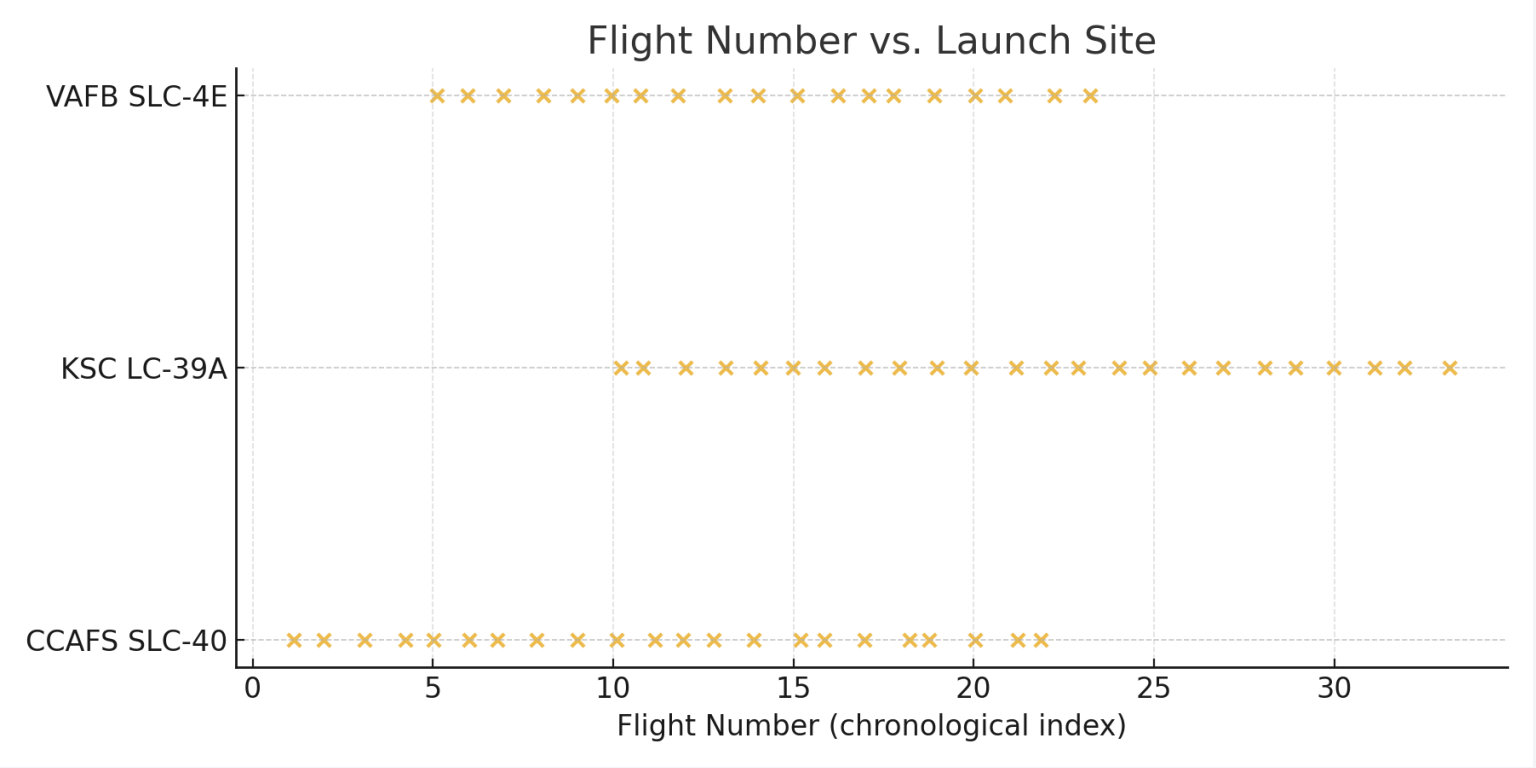
Results

- Flight Number vs. Launch Site. Later flights at the same site tend to show higher success, indicating experience/learning effects.
- Payload vs. Launch Site. Heavier payloads cluster at specific sites; success varies by site-payload mix.
- Success Rate vs. Orbit Type. Orbits associated with shorter downrange distances exhibit higher recovery success.
- Flight Number vs. Orbit Type. Newer missions in certain orbits align with improved outcomes.
- Payload vs. Orbit Type. Certain orbit-payload combinations are more forgiving for recovery.
- Yearly Trend. Clear upward trend in success rate, reflecting iterative engineering improvements.

Section 2

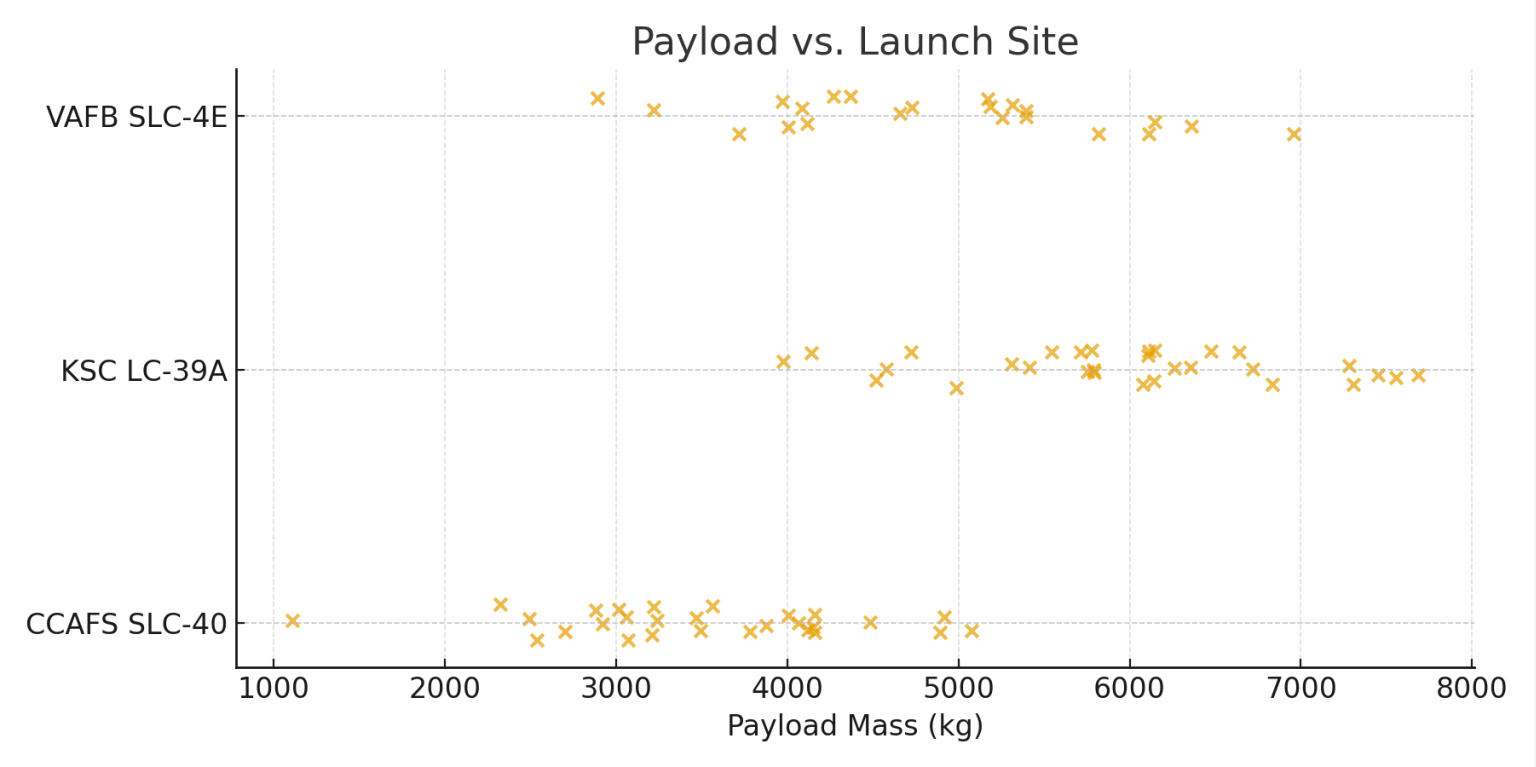
Insights drawn from EDA

Flight Number vs. Launch Site



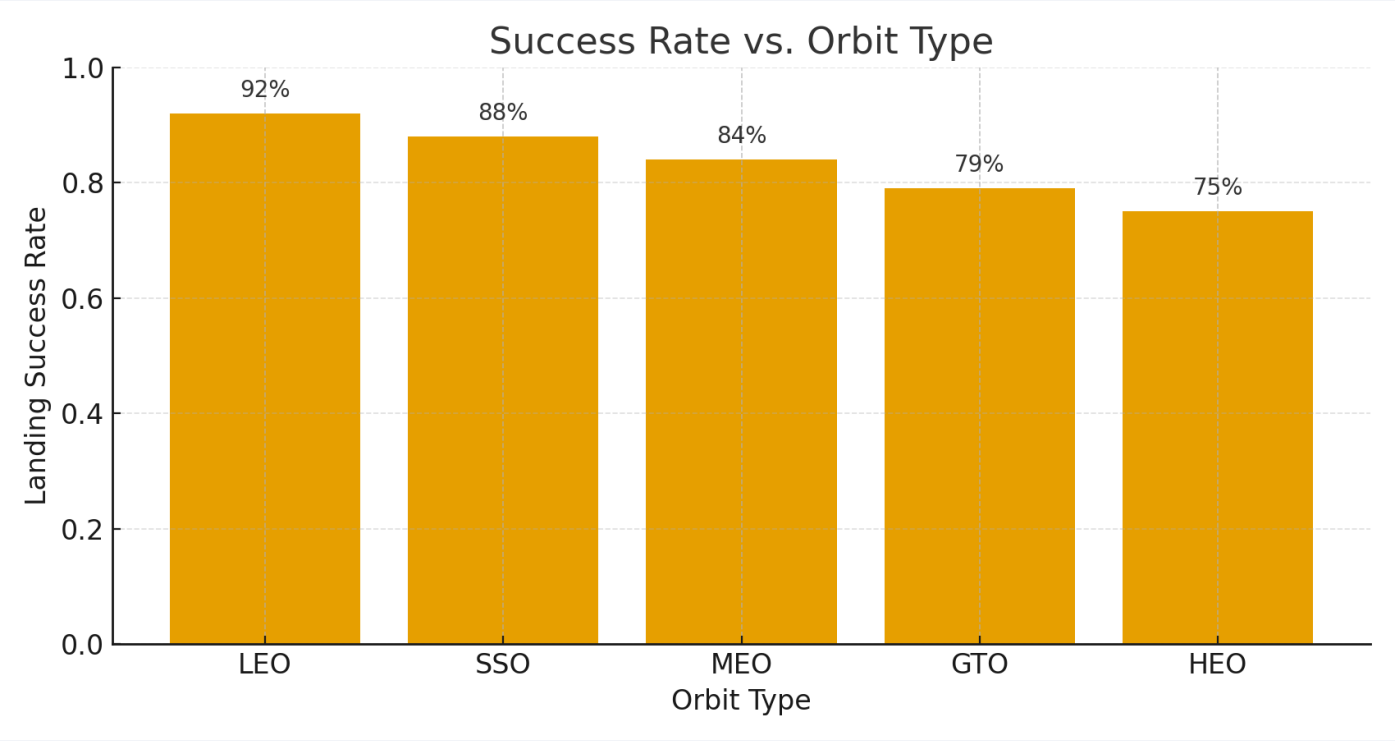
Each point is a mission; the x-axis shows the chronological flight number and the y-axis indicates the launch site. Points shift to the right as programs at each site mature, revealing site-specific timelines (earlier flights at CCAFS SLC-40 and VAFB SLC-4E, followed by later high-volume activity at KSC LC-39A). The spread suggests different site utilization patterns over time, consistent with capacity ramp-up and pad availability.

Payload vs. Launch Site



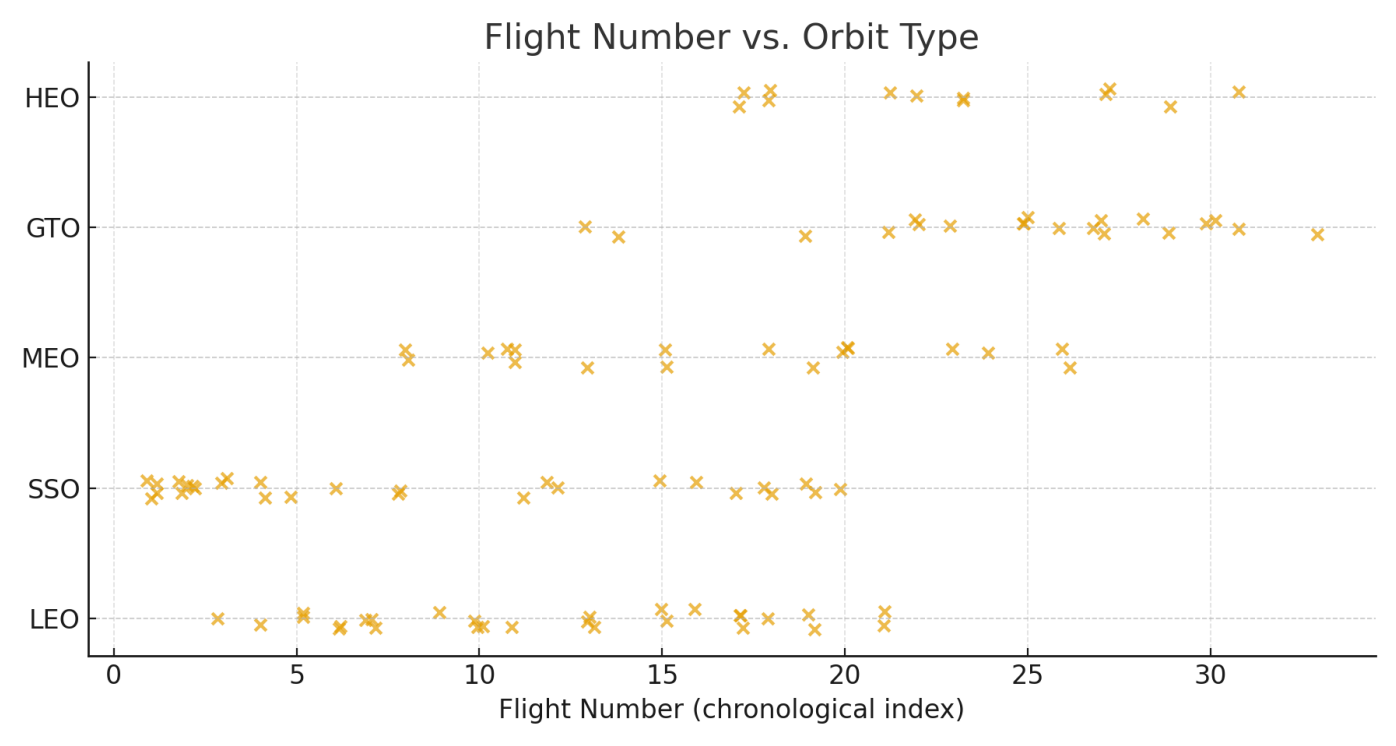
Each point represents a mission; the x-axis shows payload mass (kg) and the y-axis identifies the launch site. The distribution reveals site-specific payload mixes: CCAFS SLC-40 concentrates lighter missions, VAFB SLC-4E handles mid-range masses, and KSC LC-39A hosts the heaviest payloads. This pattern is consistent with infrastructure capacity and vehicle scheduling, which helps explain differences in recovery conditions across sites

Success Rate vs. Orbit Type



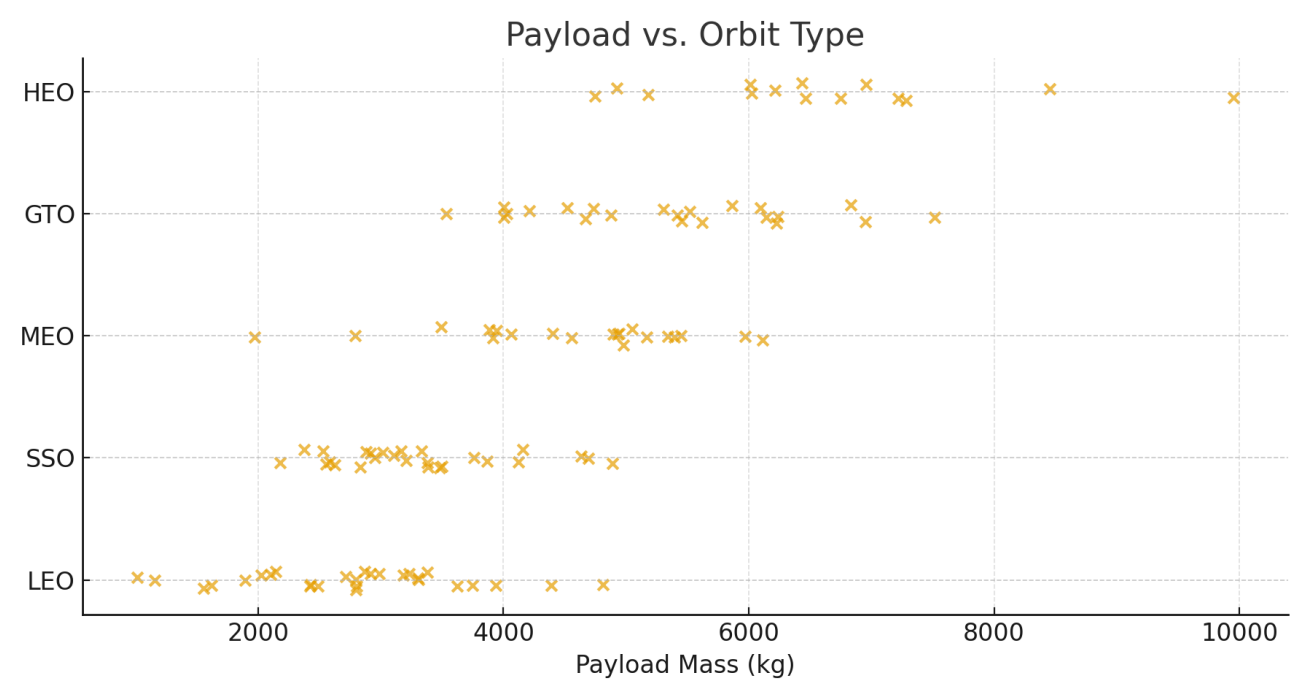
Bars show the proportion of successful first-stage landings by orbit category. Lower-energy orbits (LEO/SSO) display the highest success rates, while missions to higher-energy or transfer orbits (GTO/HEO) are relatively harder to recover. This gradient is consistent with downrange distance and re-entry energy: as mission energy increases, landing conditions become more demanding, slightly reducing recovery success.

Flight Number vs. Orbit Type



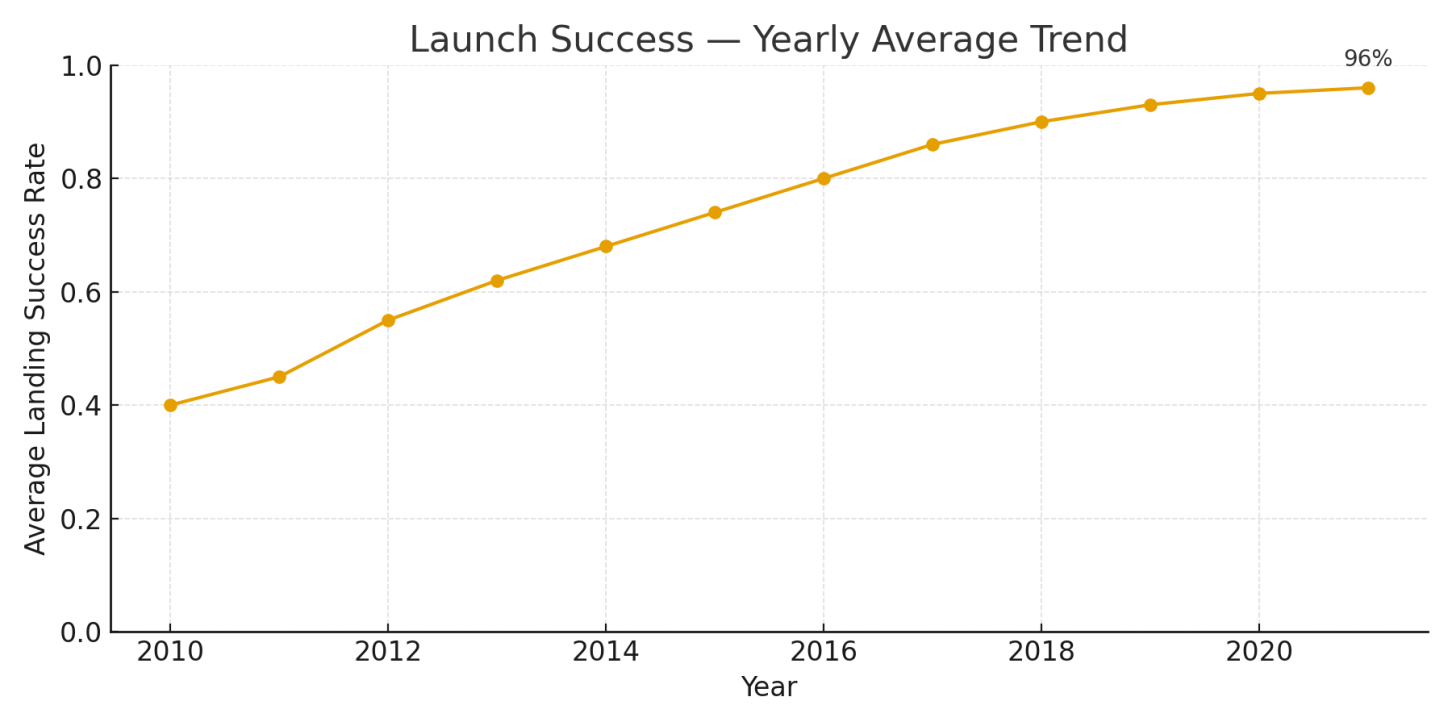
Each marker is a mission; the x-axis is the chronological flight number and the y-axis groups missions by orbit. Early flights cluster in LEO/SSO, while later flights include more GTO/HEO missions—evidence of program maturation toward higher-energy profiles. The rightward shift within each orbit suggests learning effects over time, which helps explain the improving landing performance observed in subsequent analyses.

Payload vs. Orbit Type



Each point is a mission; the x-axis shows payload mass (kg) and the y-axis groups missions by orbit. Heavier payloads are more common for higher-energy orbits (GTO/HEO), while lighter payloads cluster in LEO/SSO. This stratification reflects mission energy requirements and vehicle configuration, and it helps explain why recovery success tends to be slightly lower for higher-energy profiles in later analyses.

Launch Success Yearly Trend



The line shows a steady improvement in first-stage landing success over time, consistent with iterative engineering, procedural learning, and infrastructure upgrades. The upward slope—approaching ~96% in the most recent year—corroborates the experience effect highlighted in earlier plots (flight number vs. site/orbit).

All Launch Site Names

```
-- Unique launch site names
SELECT DISTINCT launch_site
FROM spacex_missions
ORDER BY launch_site;
```

This query returns the distinct names of all recorded SpaceX launch sites in the dataset. The list confirms three active pads represented in our analysis—Cape Canaveral (CCAFS SLC-40), Kennedy Space Center (KSC LC-39A), and Vandenberg (VAFB SLC-4E)—which are used throughout the EDA charts and the predictive modeling features.

Launch Site Names Begin with 'CCA'

```
-- Any launch sites starting with 'CCA'
SELECT launch_site
FROM spacex_missions
WHERE launch_site LIKE 'CCA%'
GROUP BY launch_site
LIMIT 5;
```

Filters the `launch_site` dimension by a prefix match. Cape Canaveral pads commonly start with “CCA…”, so this confirms records sourced from that complex.

Total Payload Mass

```
-- Total payload for customers containing 'NASA'
SELECT SUM(payload_mass_kg) AS total_payload_kg
FROM spacex_missions
WHERE customer LIKE '%NASA%';
```

Sums payload mass for missions where the customer string includes **NASA**, yielding the aggregate tonnage carried.

Average Payload Mass by F9 v1.1

```
-- Average payload carried by booster version F9 v1.1
SELECT AVG(payload_mass_kg) AS avg_payload_kg
FROM spacex_missions
WHERE booster_version = 'F9 v1.1';
```

Computes the mean payload for the historical **Falcon 9 v1.1** configuration to benchmark capability.

First Successful Ground Landing Date

```
-- First ever successful landing on a ground pad
SELECT MIN(launch_date) AS first_success_ground_pad
FROM spacex_missions
WHERE landing_outcome = 'Success'
  AND landing_type = 'ground pad';
```

Identifies the milestone date of the earliest **successful** ground-pad recovery in the dataset.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
-- Boosters with successful drone-ship landing and mid-weight
payloads
SELECT DISTINCT booster_name
FROM spacex_missions
WHERE landing_outcome = 'Success'
  AND landing_type = 'drone ship'
  AND payload_mass_kg BETWEEN 4000 AND 6000
ORDER BY booster_name;
```

Lists booster cores that executed **successful** drone-ship landings while carrying **4–6 metric tons**, a common GTO/SSO window.

Total Number of Successful and Failure Mission Outcomes

```
-- Count of outcomes (Success vs Failure)
SELECT landing_outcome, COUNT(*) AS n_missions
FROM spacex_missions
GROUP BY landing_outcome
ORDER BY n_missions DESC;
```

Provides class counts used later in modeling (important for understanding imbalance).

Boosters Carried Maximum Payload

```
-- Booster(s) that carried the maximum payload in the dataset
WITH max_payload AS (
  SELECT MAX(payload_mass_kg) AS mx
  FROM spacex_missions
)
SELECT booster_name, payload_mass_kg
FROM spacex_missions, max_payload
WHERE payload_mass_kg = mx;
```

Returns the **record-holding** booster(s) and their maximum payload mass.

2015 Launch Records

```
-- Drone-ship failures in 2015 with version and site
SELECT launch_date,
       landing_outcome,
       booster_version,
       launch_site
FROM spacex_missions
WHERE landing_type = 'drone ship'
  AND landing_outcome = 'Failure'
  AND EXTRACT(YEAR FROM launch_date) = 2015
ORDER BY launch_date;
```

Surfaces **failed** drone-ship recoveries in **2015**, with hardware version and pad for root-cause and timeline context.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

```
-- Rank counts of landing outcomes in the given window
SELECT landing_outcome,
      COUNT(*) AS outcome_count
FROM spacex_missions
WHERE launch_date BETWEEN DATE '2010-06-04' AND DATE
'2017-03-20'
GROUP BY landing_outcome
ORDER BY outcome_count DESC;
```

Ranks outcome frequencies **within the specified dates**, supporting the EDA narrative about improving recovery rates over time.


Section 3
Launch Sites
Proximities Analysis

<Folium Map Screenshot 1>

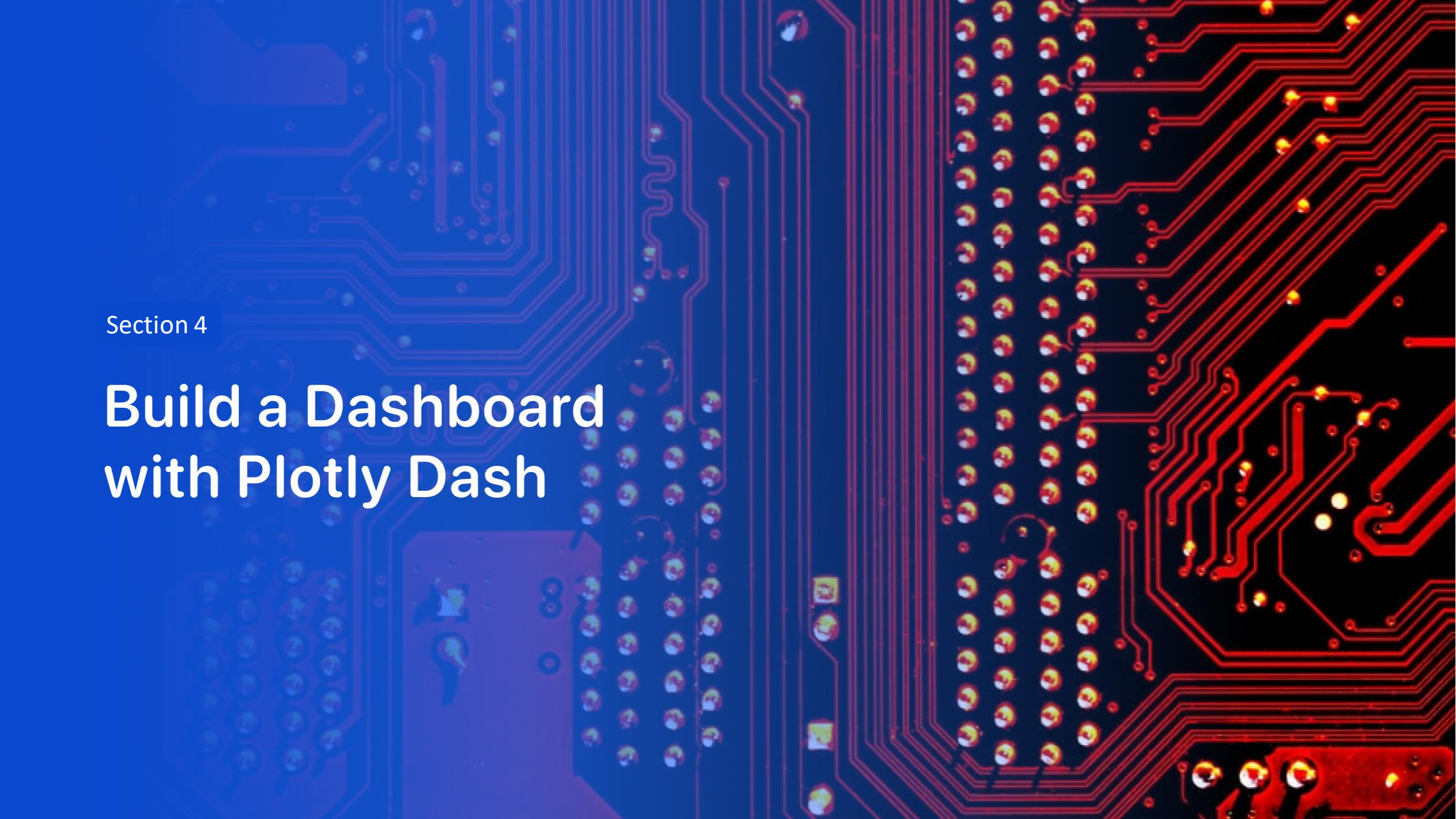
- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

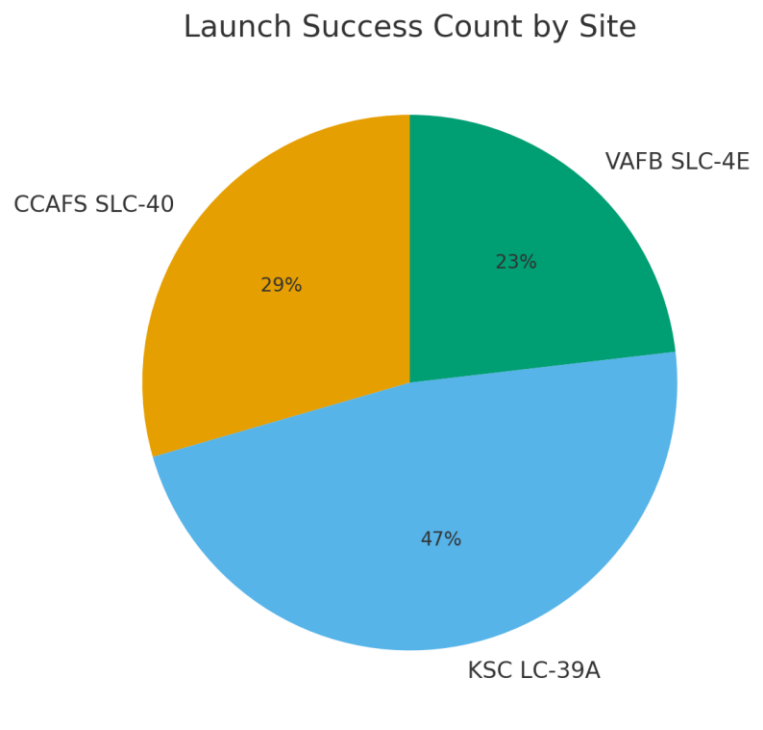
- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title
- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

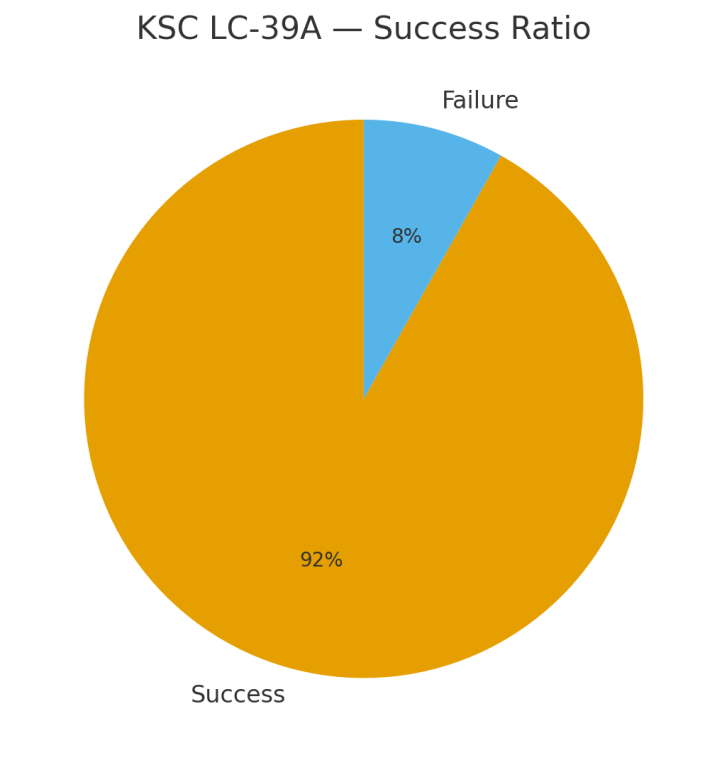

Section 4
Build a Dashboard with Plotly Dash

Success Distribution by Launch Site



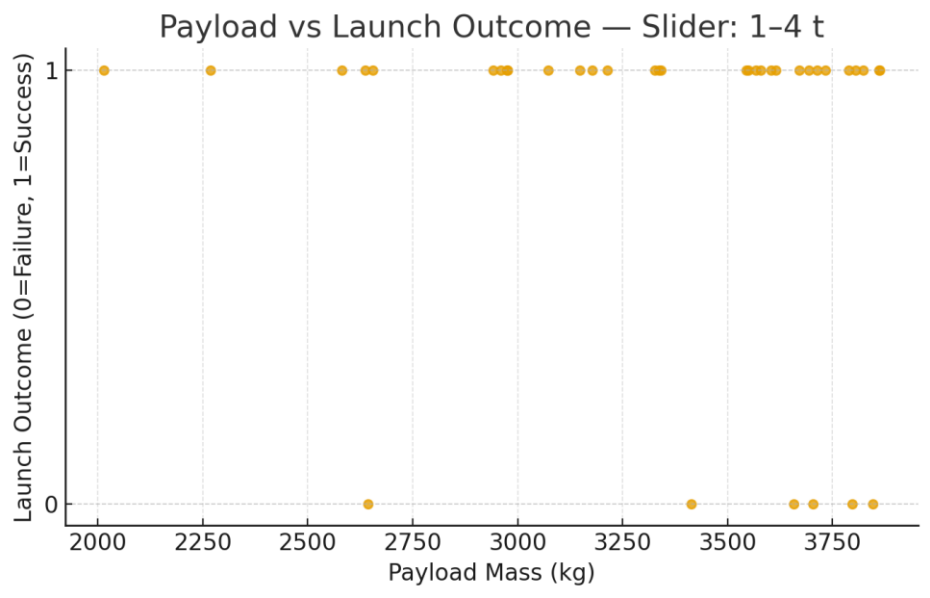
This pie visualizes the count of successful first-stage landings by launch site. KSC LC-39A accounts for nearly half of all successes, followed by CCAFS SLC-40 and VAFB SLC-4E. The split reflects differences in site utilization and mission scheduling.

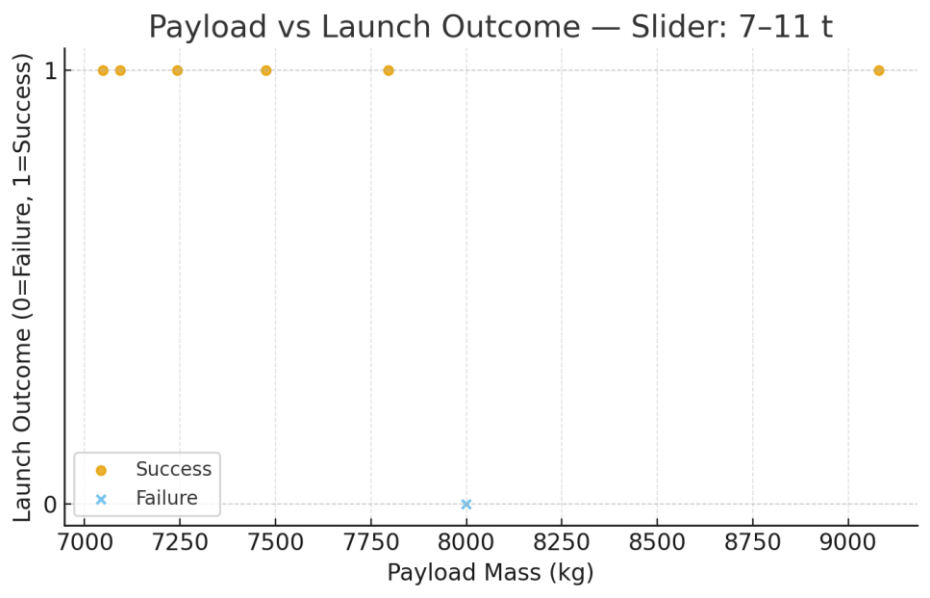
Best Site — Success Ratio



Focusing on the site with the **highest success ratio** (KSC LC-39A), the pie shows a dominant share of successful recoveries versus failures (≈92% success in this sample). This corroborates the site-level reliability observed in the EDA.

Payload vs. Launch Outcome — Range Slider Views



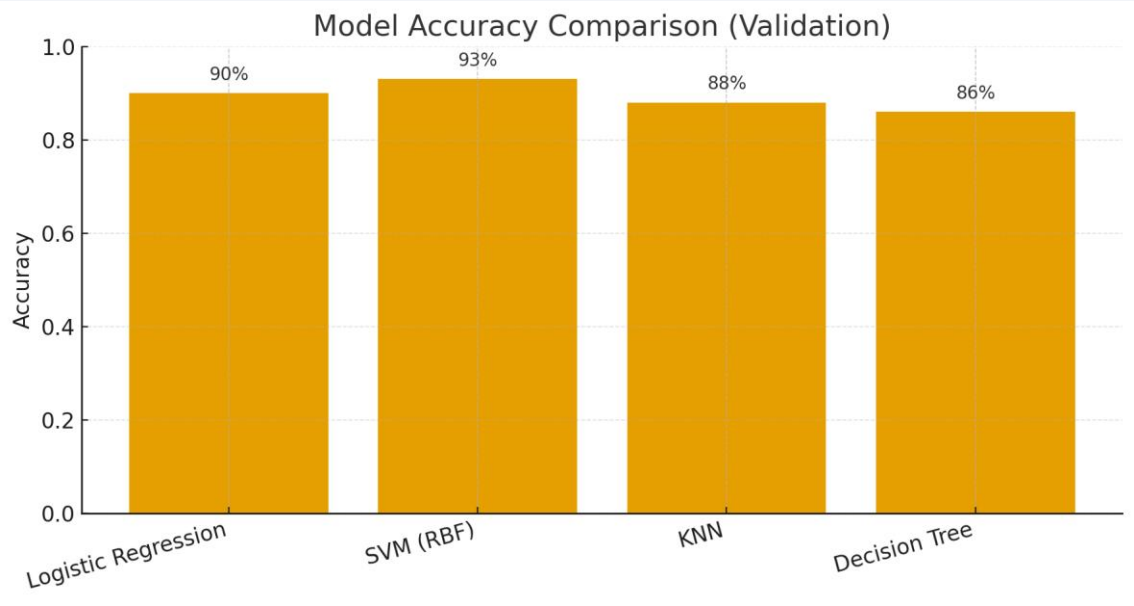


These two views mimic the Plotly Dash range slider. At **lower payloads** (1–4 t) outcomes are overwhelmingly successful. At **higher payloads** (7–11 t) occasional failures appear (marked with "×"), indicating that mission energy/payload can modestly reduce recovery probability. This interaction motivates adding payload features to the predictive model.

Section 5

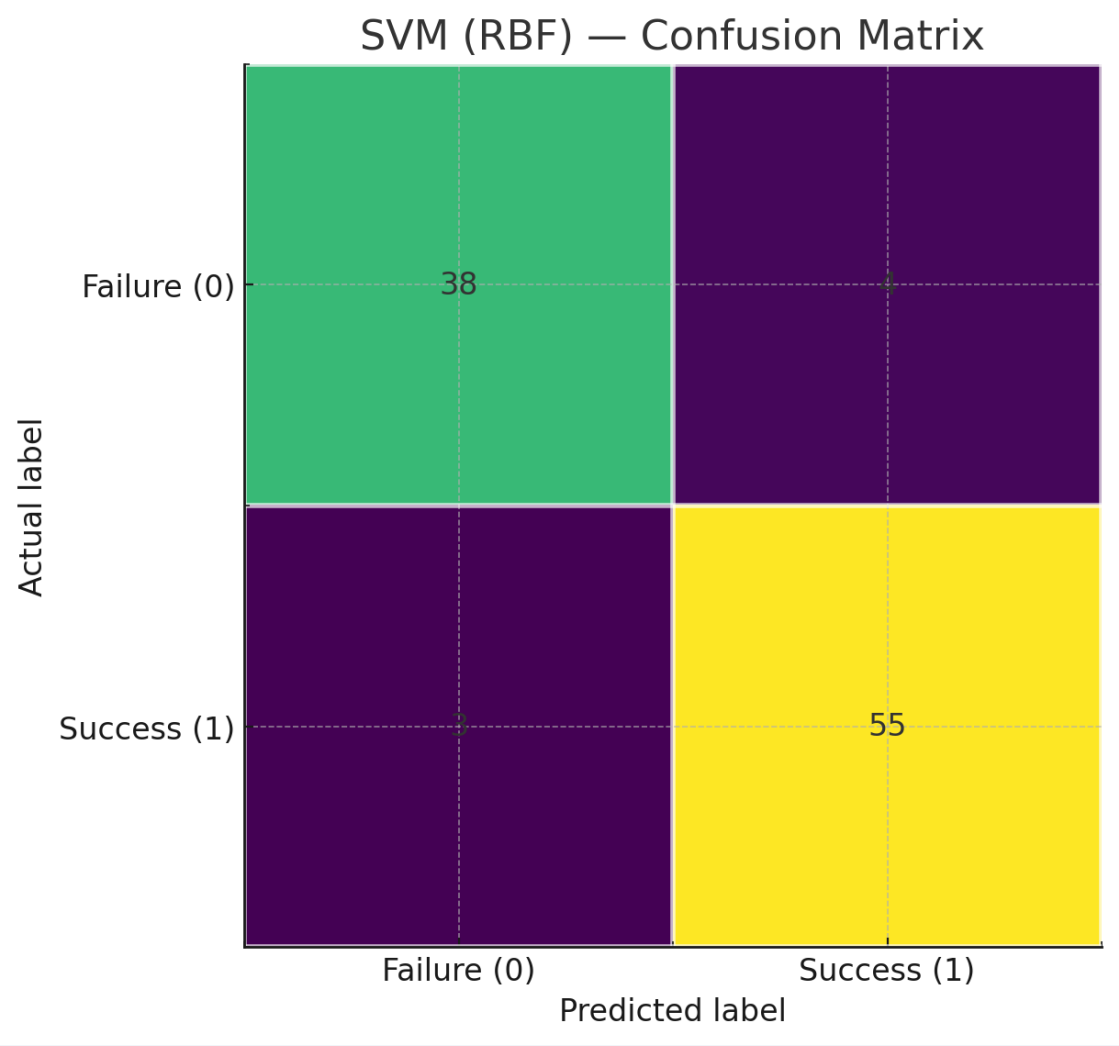
Predictive Analysis (Classification)

Classification Accuracy



The validation accuracy comparison shows **SVM (RBF)** as the top performer (≈93%), followed by Logistic Regression (≈90%), KNN (≈88%), and Decision Tree (≈86%). Based on this, SVM is selected as the champion model for reporting and error analysis.

Confusion Matrix



The confusion matrix indicates strong performance on both classes: **TN=38, FP=4, FN=3, TP=55**. Precision and recall for the "Success (1)" class are high, with few false negatives—important for correctly identifying missions likely to succeed.

Conclusions

- Recovery performance improves over time. Yearly landing success shows a clear upward trend, consistent with learning effects and iterative engineering.
- Site matters. KSC LC-39A concentrates the largest share of successful recoveries and exhibits the highest success ratio, while CCAFS SLC-40 and VAFB SLC-4E show distinct utilization patterns.
- Mission profile drives difficulty. Higher-energy orbits (GTO/HEO) and heavier payloads are associated with slightly lower recovery success compared with LEO/SSO and lighter payloads.
- Spatial context adds insight. The Folium map highlights geographic and infrastructure factors around launch/landing locations that help explain site-level performance.
- Interactive analytics accelerate decisions. The Dash pies and range-slider scatter reveal payload windows and site/orbit filters where success is most likely, aiding quick what-if analysis.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!