IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sebastian Rada

December 2025

Outline

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

Executive Summary

Summary of methodologies Collected SpaceX launch data via API and web scraping, cleaned and prepared it, explored patterns using visualizations and SQL, built an interactive dashboard, and trained classification models (e.g., logistic regression, decision trees) to predict first-stage landing success.

Summary of all results Launch site, booster version, and payload mass were the strongest predictors. Models achieved roughly 80–90% accuracy, showing that newer boosters land more reliably and heavier payloads reduce success probability.

Introduction

Project background and context:

The project focuses on SpaceX’s Falcon 9 rockets, whose ability to land and be reused greatly reduces launch costs. As a data scientist for a competing company, your goal is to analyze SpaceX’s historical launches to understand what drives first-stage landing success.

Problems you want to find answers:

Which factors most influence landing success? How accurately can we predict whether a launch will land successfully? And how can these insights help estimate costs and improve competitive bidding on future launches?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification model

Data Collection

- Combined SpaceX API data + Wikipedia web-scraped tables
- Integrated multiple sources to build a complete launch dataset
- Ensured consistency through cleaning, merging, and formatting
- Output: final analytical dataset for EDA and ML modeling

SpaceX API —▶

\

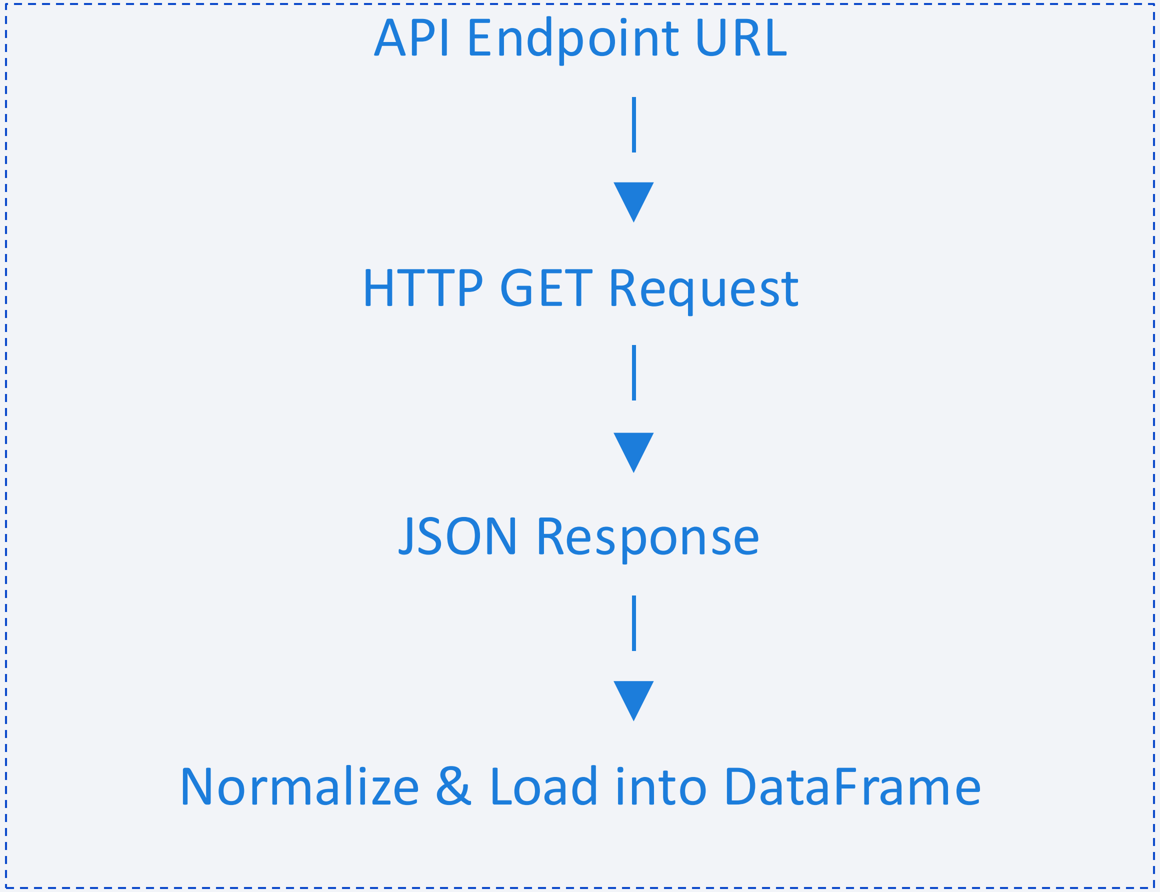
▶ Data Extraction –▶ Cleaning & Formatting–▶
Merged Dataset

/

Wikipedia ——▶

Data Collection – SpaceX API

- Used SpaceX REST API to retrieve historical Falcon 9 launch data
- Extracted: launch dates, booster versions, payload mass, launch sites, landing outcomes
- Queried endpoints using HTTP GET requests
- Converted JSON responses into pandas DataFrames
- Link



API Endpoint URL

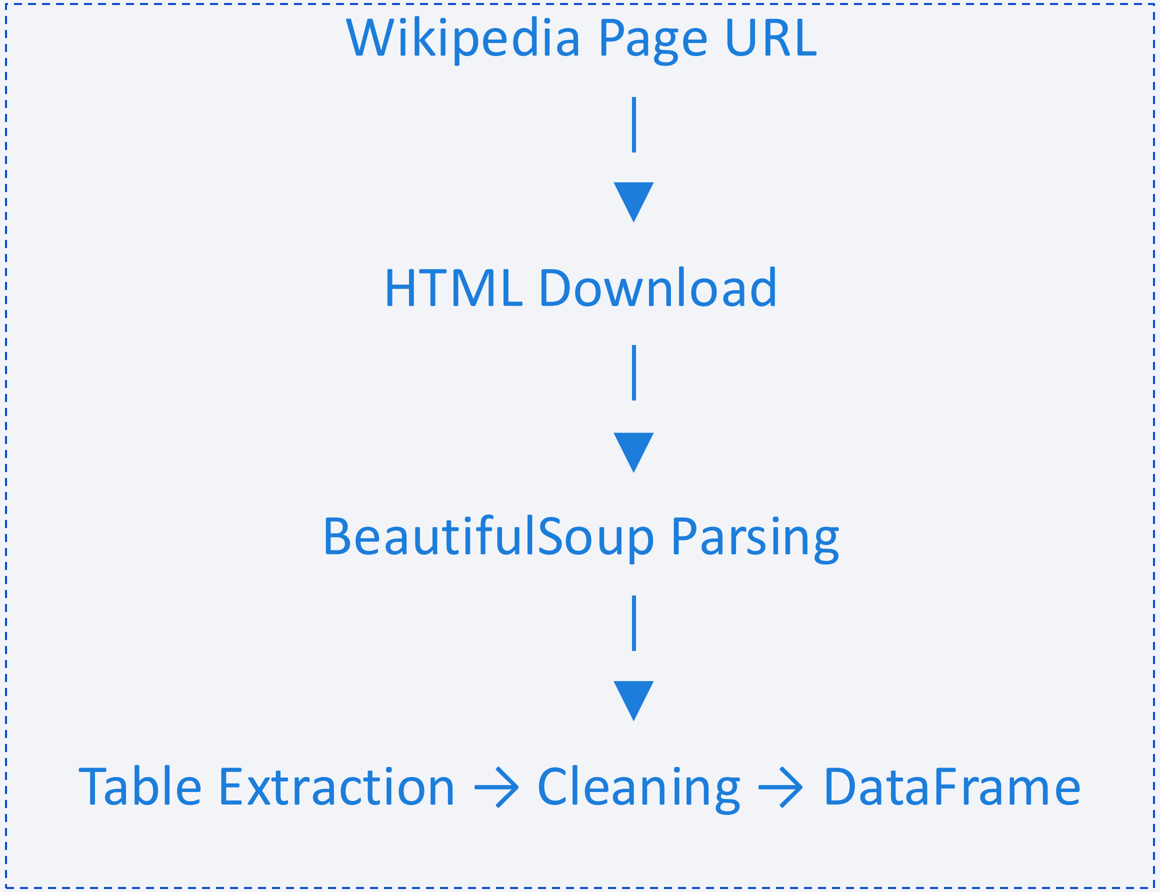
HTTP GET Request

JSON Response

Normalize & Load into DataFrame

Data Collection - Scraping

- Scraped Wikipedia tables containing additional Falcon 9 launch info
- Used BeautifulSoup to parse HTML and extract table rows
- Cleaned and standardized fields not available in the API
- Joined scraped data with API data to complete missing attributes
- Link



Wikipedia Page URL

HTML Download

BeautifulSoup Parsing

Table Extraction → Cleaning → DataFrame

Data Wrangling

- Cleaned and standardized raw data from both API and web scraping
- Converted data types and resolved missing or inconsistent values
- Engineered new features relevant to landing prediction
- Prepared a final, analysis-ready dataset for EDA and machine learning
- Link

EDA with SQL

- Inspected table structure
- Filtered launches by launch site
- Calculated total payload mass
- Computed the average payload mass
- Retrieved all distinct landing outcomes
- Found the earliest date
- Identified booster versions
- Counted missions by mission outcome
- Listed distinct mission outcomes
- Selected booster version(s)
- Extracted 2015 launches
- Link

EDA with Data Visualization

- Bar Charts: Compared landing success rates across launch sites and booster versions to identify which categories showed the highest performance
- Scatter Plots: Explored the relationship between payload mass and landing outcome, helping detect trends or thresholds
- Histograms: Examined distributions of payload mass and flight numbers to understand data spread and detect outliers
- Box Plots: Compared payload distributions across orbits and launch sites to evaluate variability and differences among groups
- Line Charts / Time Series: Visualized landing success over time to observe improvements in reliability across years
- Link

Build an Interactive Map with Folium

Map Objects Created

- Base **Folium map** centered on NASA JSC
- **Circle markers** for each launch site
- **Text labels** using DivIcon
- **Success/failure markers** (green/red)
- **MarkerCluster** to group overlapping points
- Link

Porpuse

- Highlight and label launch sites
- Visualize landing outcomes clearly
- Reduce clutter with clustering
- Analyze geographic factors (distance)

Build a Dashboard with Plotly Dash

Plots and Interactions Added

- **Pie Chart (All Sites):** Shows total number of successful launches across all SpaceX launch sites
- **Pie Chart (Selected Site):** Displays success vs. failure outcomes for a specific launch site chosen from the dropdown
- **Scatter Plot:** Plots payload mass vs. launch success, colored by booster version category
- **Dropdown Menu:** Allows users to select a specific launch site or view all sites
- **Range Slider:** Lets users filter launches by payload mass range
- Link

Predictive Analysis (Classification)

I built the classification model by first cleaning and standardizing the dataset, then splitting it into training and testing subsets. Using GridSearchCV, I trained four algorithms—Logistic Regression, SVM, Decision Tree, and KNN—while tuning their hyperparameters through cross-validation to identify the best settings. After evaluating each model on validation and test sets, I compared their accuracies and confusion matrices. The Decision Tree achieved the highest validation score, but Logistic Regression, SVM, and KNN performed more consistently on the test set, making them the most reliable models overall

Link

Results

EDA Results

- Success rates vary by launch site and booster version
- Payload has a moderate relationship with landing success
- Newer boosters show higher reliability

Interactive Dashboard

- Dropdown selects site; slider filters payload
- Pie charts show site performance; scatter shows payload vs. success
- All graphs update dynamically for exploration

Predictive Analysis

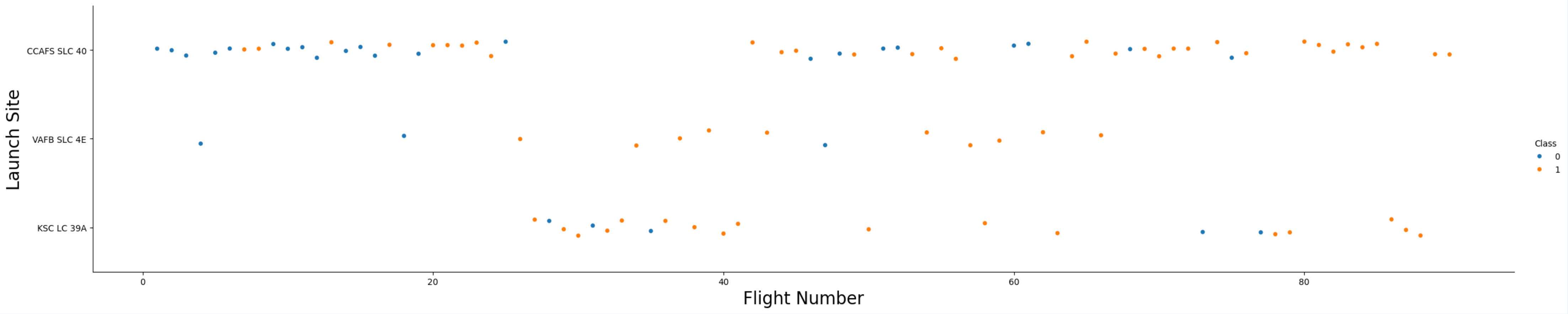
- Trained Logistic Regression, SVM, KNN, Decision Tree
- Tuned with GridSearchCV; compared accuracy
- SVM and Logistic Regression were the most reliable models

Section 2

Insights drawn from EDA

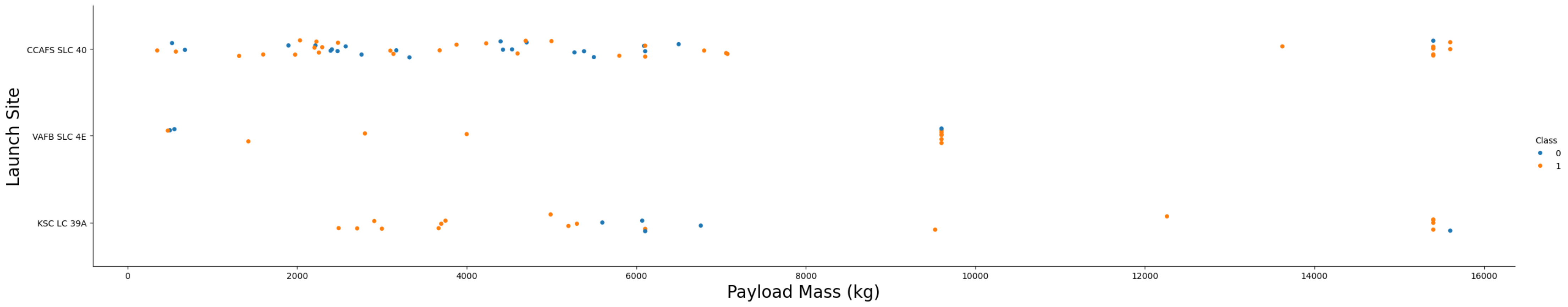
Flight Number vs. Launch Site

- This scatter plot shows how launch success (orange dots) and failure (blue dots) are distributed across flight numbers for each SpaceX launch site. Each row represents a different launch site, and each point represents one launch. The graph reveals that most later flights tend to be successful, especially at CCAFS SLC 40 and KSC LC 39A, suggesting improved reliability over time. The VAFB SLC 4E site has fewer launches but shows a similar pattern of increasing success



Payload vs. Launch Site

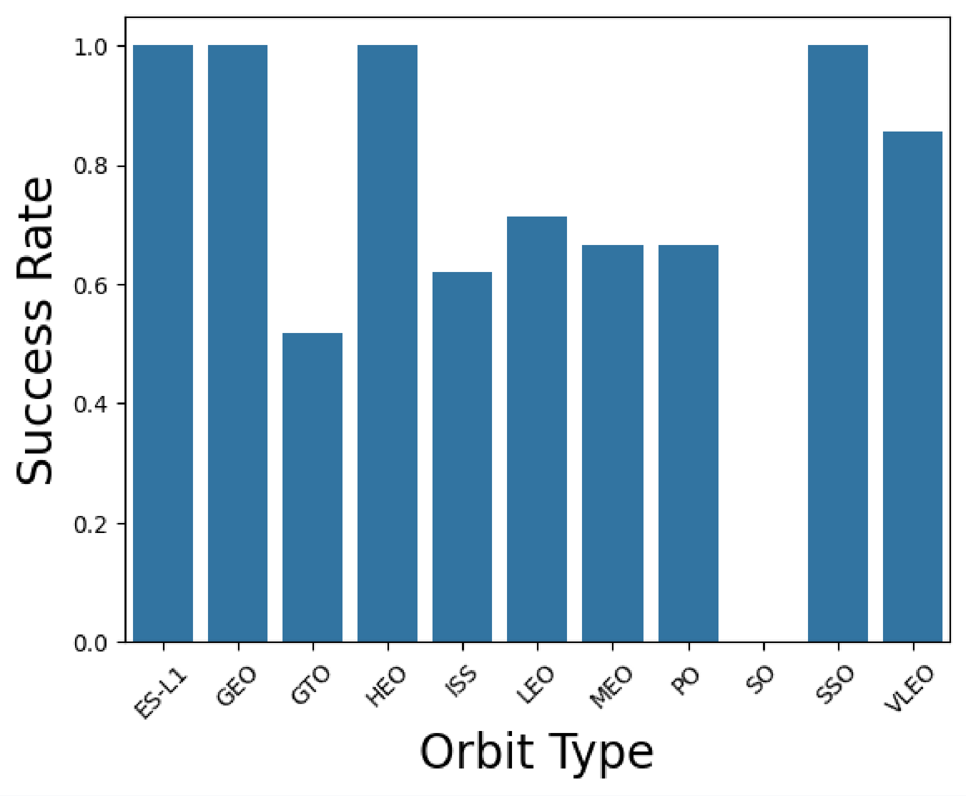
- This plot shows launch outcomes across different payload masses and sites. Orange points (success) appear across the full payload range, while failures (blue) cluster in fewer areas. Overall, success remains high even at heavier payloads, suggesting payload mass is not a major limiting factor



Success Rate vs. Orbit Type

This bar chart shows the launch success rate for each orbit type. Most orbit categories such as ES-L1, GEO, HEO, and SSO have very high success rates, while GTO stands out with a lower success rate

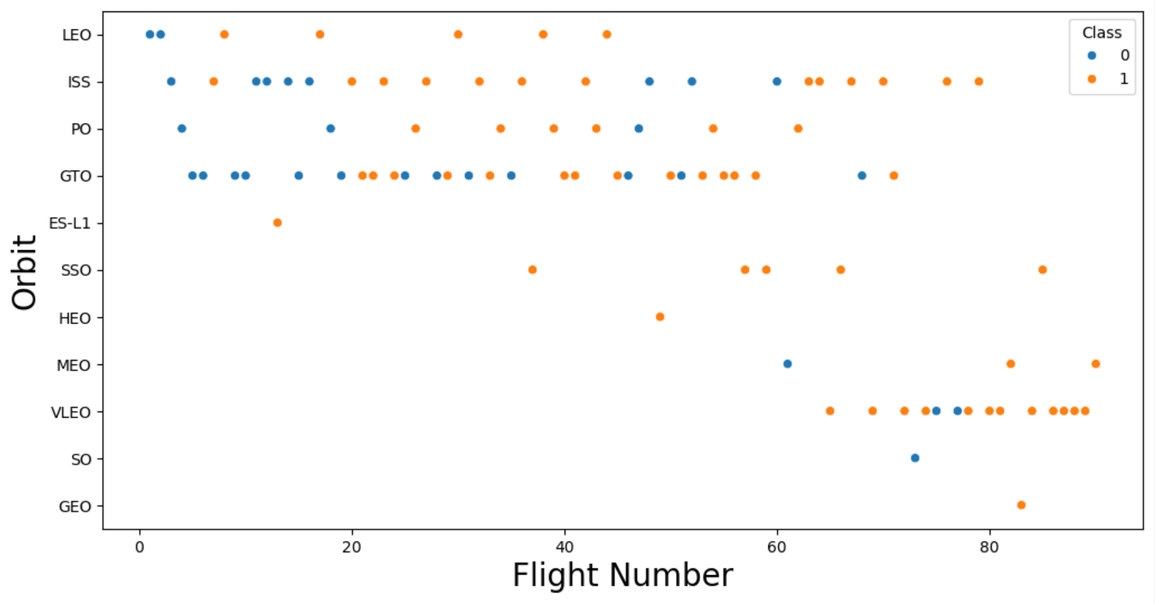
Overall, success remains strong across most orbits with some variation depending on mission type



Flight Number vs. Orbit Type

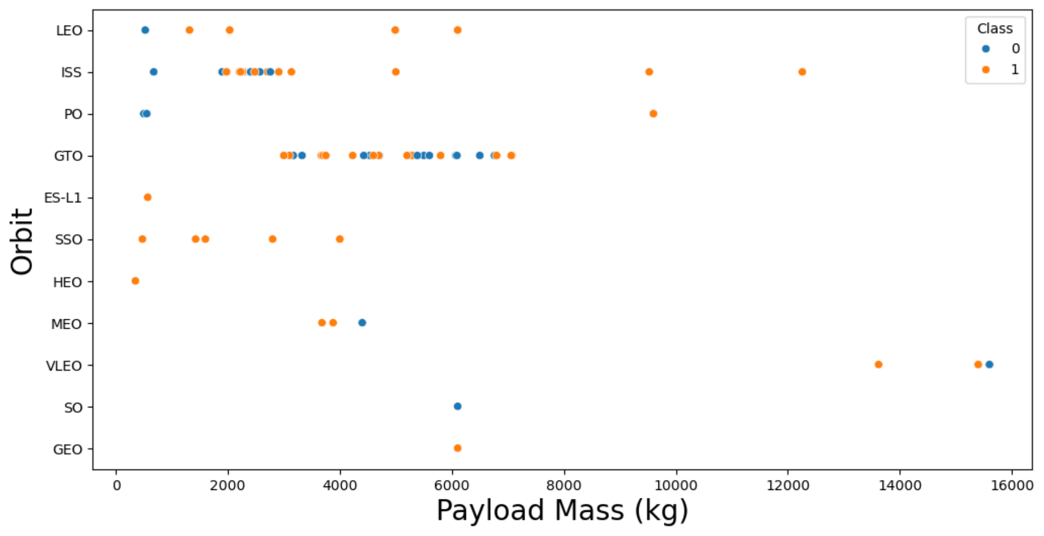
This plot shows launch outcomes across different orbit types and flight numbers. Orange points represent successful landings and blue points represent failures

Most orbits show a high concentration of successful launches, especially in later flights, which suggests performance improves over time



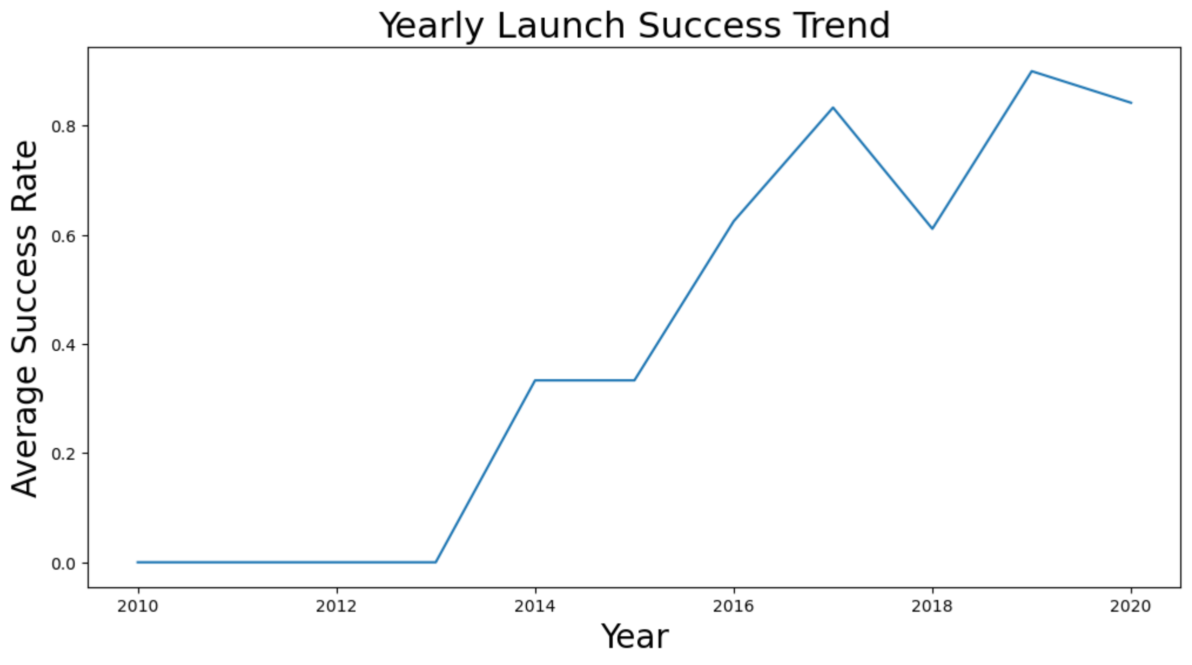
Payload vs. Orbit Type

- Shows **payload mass (x-axis)** vs. **orbit type (y-axis)**
- **Two colors** represent two launch classes (0 and 1)
- Lets you compare how each class is distributed across different orbits
- **Insight:** Class 1 tends to handle **heavier payloads**, especially above ~10,000 kg
- **Insight:** LEO and GTO have the **highest concentration** of launches from both classes



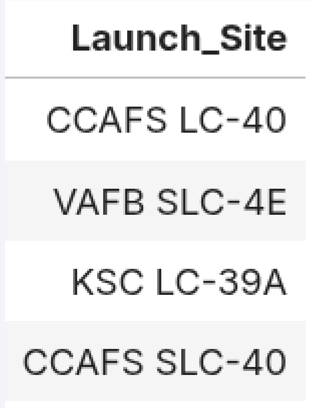
Launch Success Yearly Trend

- Line chart showing average launch success rate per year (2010–2020)
- Success rate starts at 0, then steadily increases after 2013
- Peaks around 2019, with a slight dip in 2020
- Insight: The success rate shows a clear upward trend, suggesting major reliability improvements over the decade



All Launch Site Names

SELECT DISTINCT "Launch_Site" removes duplicates, giving a clean list of all different launch locations used in the dataset



Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Columns:

Date, Time (UTC), Booster_Version, Launch_Site, Payload, PAYLOAD_MASS__KG_, Orbit, Customer, Mission_Outcome, Landing_Outcome

Explanation:

These columns appear in the query output because the SELECT * statement returns every field from rows where Launch_Site begins with "CCA", limited to the first 5 matching launches

Total Payload Mass

Query Result: Total payload mass carried by boosters for NASA (CRS) missions: 45,596 kg

Explanation: The query uses SUM(PAYLOAD_MASS__KG_) to add up all payload masses from rows where the customer is NASA (CRS), giving the total payload delivered for their resupply missions

Average Payload Mass by F9 v1.1

Query Result: Average payload mass for booster version F9 v1.1: 2928.4 kg

Explanation: The query calculates the mean payload mass by selecting only rows where the booster version is F9 v1.1 and applying AVG() to the payload mass column.

First Successful Ground Landing Date

- Query Result: 2015-12-22

- Explanation: Filtering the table to rows where Landing_Outcome = 'Success (ground pad)' and selecting the earliest date gives December 22, 2015, which is the first successful Falcon 9 ground-pad landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query Result:

- F9 FT B1022
- F9 FT B1026
- F9 FT B1021.2
- F9 FT B1031.2

Explanation: These boosters meet all three conditions:

- Successful drone-ship landing
- Payload mass between 4000 and 6000 kg.

The query filters for those criteria and returns the matching booster names.

Total Number of Successful and Failure Mission Outcomes

Query Result:

- Successful missions: 100
- Failed missions: 1

Explanation:nFrom the grouped results:

- Success appears in three forms (Success, Success , and Success (payload status unclear)), totaling 100 successful missions.
- Only 1 mission is marked as Failure (in flight) → counted as failure.

Boosters Carried Maximum Payload

Query Result:

- F9 B5 B1048.4
- F9 B5 B1049.4
- F9 B5 B1051.3
- F9 B5 B1056.4
- F9 B5 B1048.5
- F9 B5 B1051.4
- F9 B5 B1049.5
- F9 B5 B1060.2
- F9 B5 B1058.3
- F9 B5 B1051.6
- F9 B5 B1060.3
- F9 B5 B1049.7

Explanation: These boosters all carried the maximum payload mass recorded in the dataset. The inner query finds the highest payload value, and the outer query returns all boosters that match that maximum.

2015 Launch Records

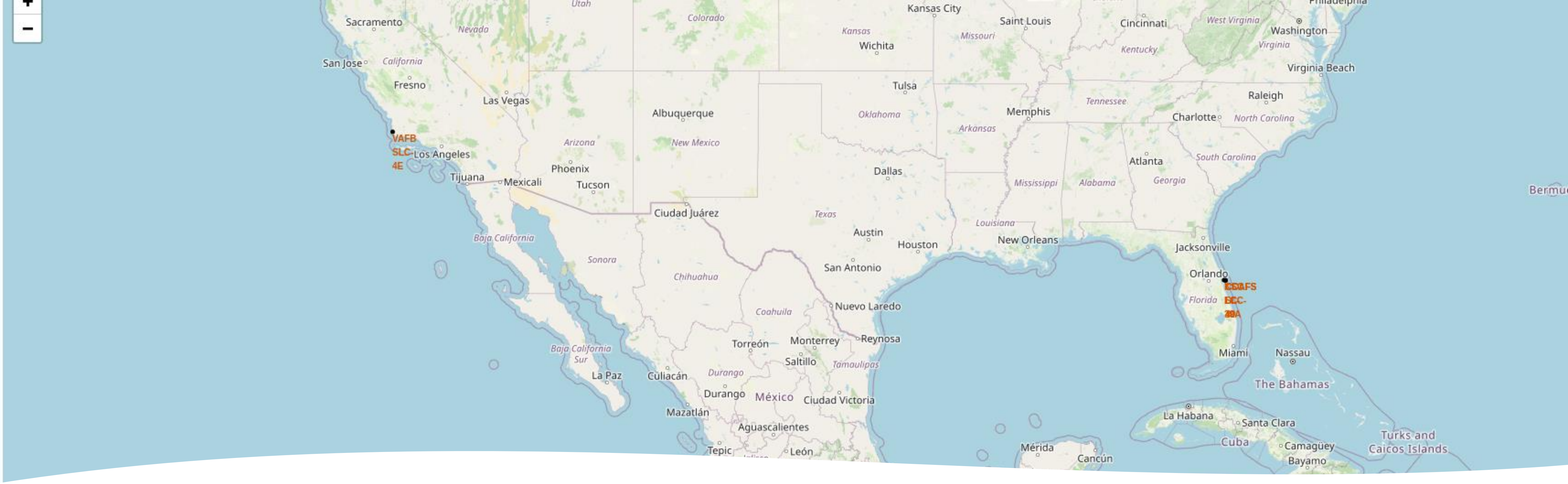
Query Result:

- January — F9 v1.1 B1012 — CCAFS LC-40 — Failure (drone ship)
- April — F9 v1.1 B1015 — CCAFS LC-40 — Failure (drone ship)

Explanation: The query filters for launches in 2015 where the landing outcome was "Failure (drone ship)", then extracts the month name, booster version, and launch site. Two such failures occurred that year.

Section 3

Launch Sites Proximities Analysis



Global Map of SpaceX Launch Site Locations

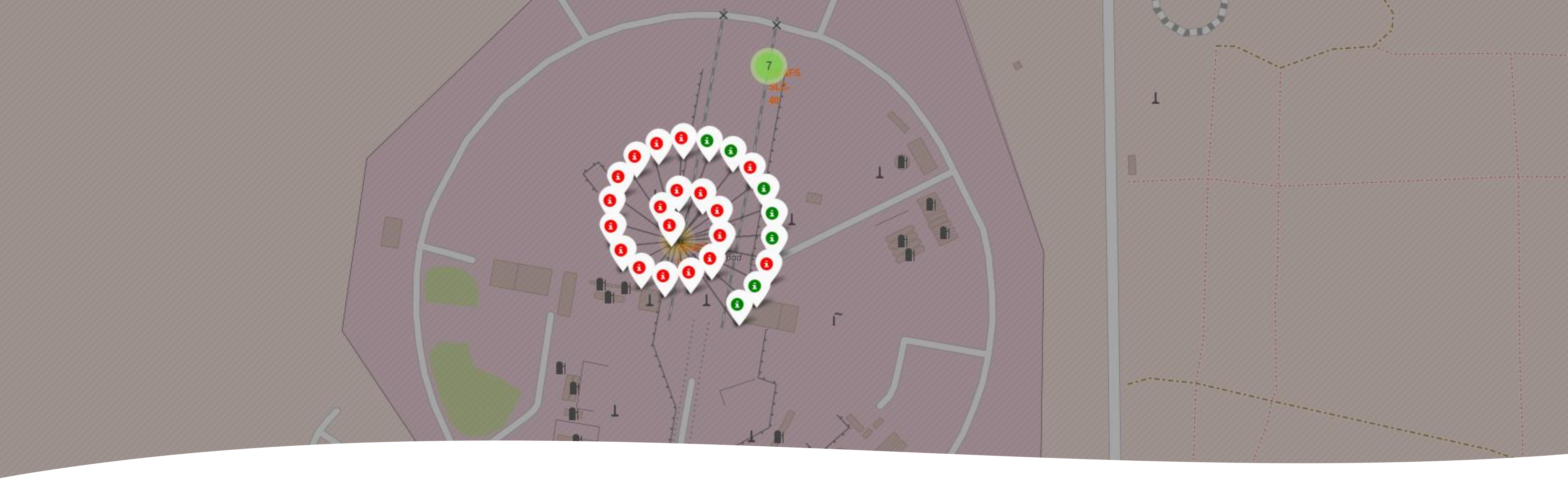
The map shows all SpaceX launch sites from the dataset. Three sites appear:

- VAFB SLC-4E (California)
- CCAFS LC-40 (Florida)
- KSC LC-39A (Florida)

All launch sites are located in the United States.

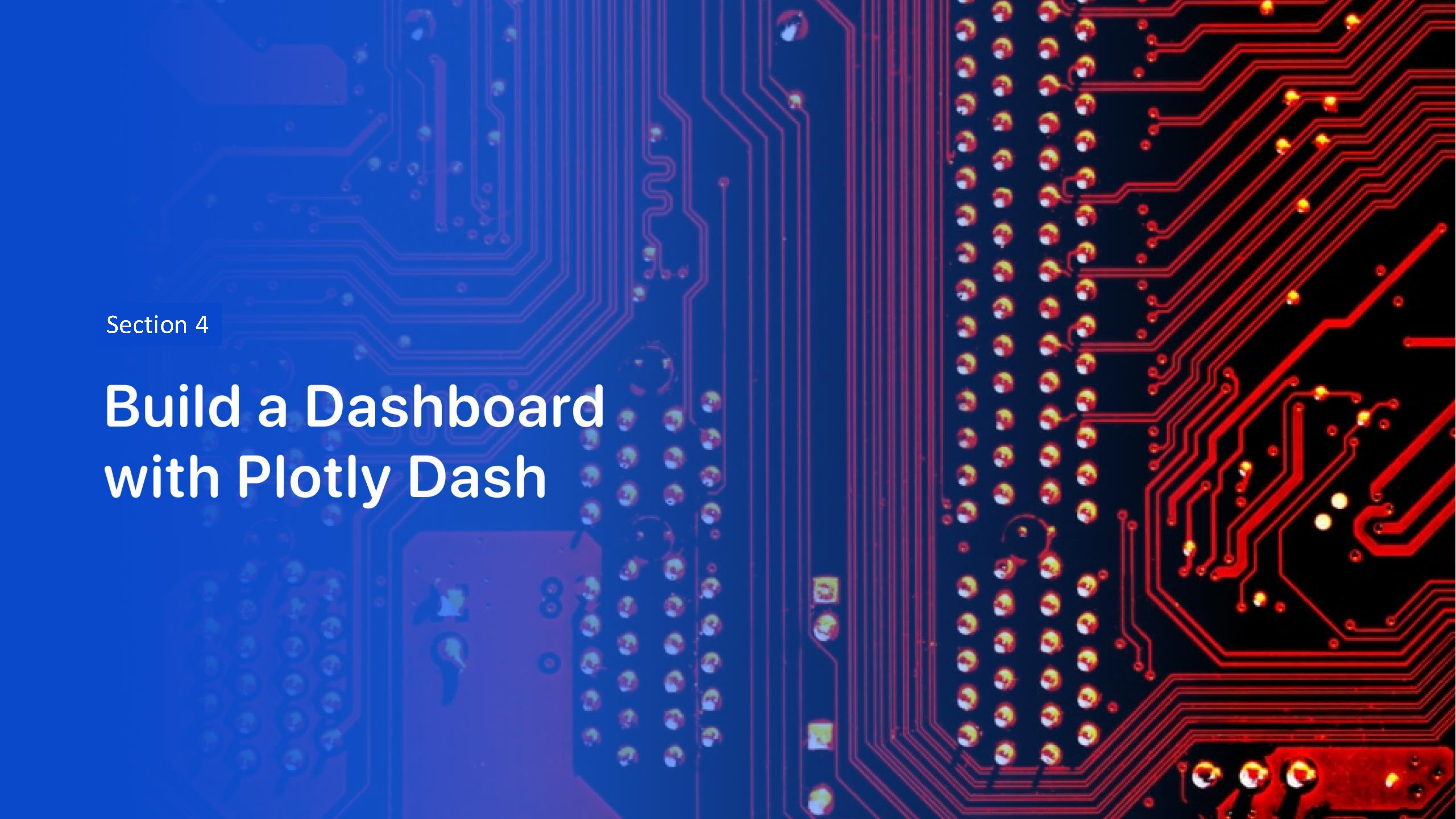
Florida sites appear close together because they are only a few kilometers apart.

West Coast is used mainly for polar orbits, while East Coast supports most other missions.



Color-Coded Launch Outcomes at CCAFS and KSC Launch Sites

- Each marker represents a Falcon 9 launch, color-coded by landing outcome:
- Green markers → Successful landing
- Red markers → Failed landing
- The MarkerCluster groups markers into a spiral-like cluster when zoomed in, making dense launch activity easy to visualize.
- Green markers grow in frequency over time, reflecting SpaceX’s increasing landing reliability.



Section 4

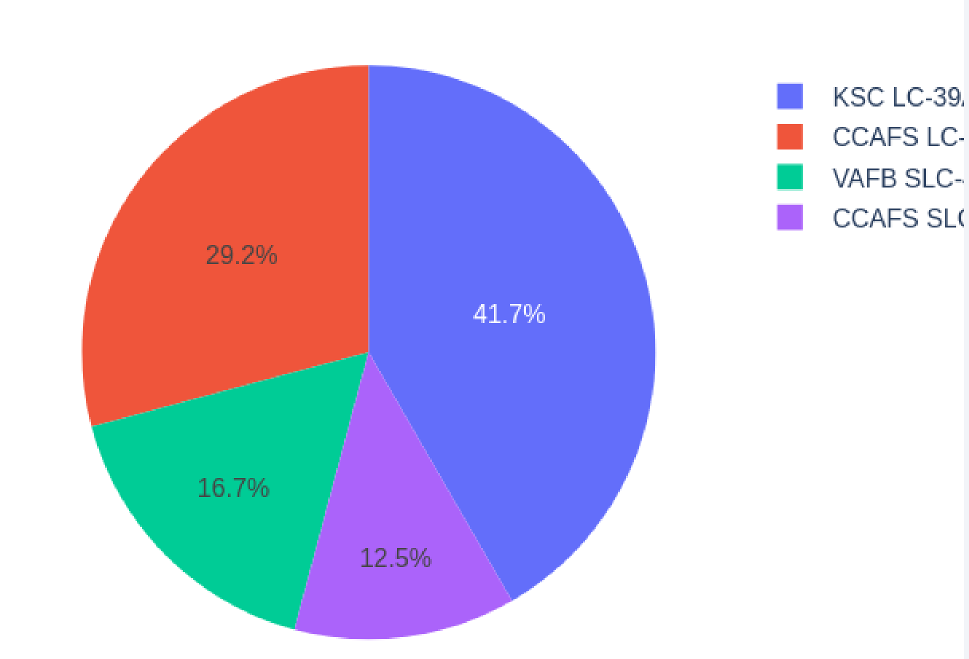
Build a Dashboard with Plotly Dash

Launch Success Distribution Across All SpaceX Sites

- KSC LC-39A has the highest number of successful launches.

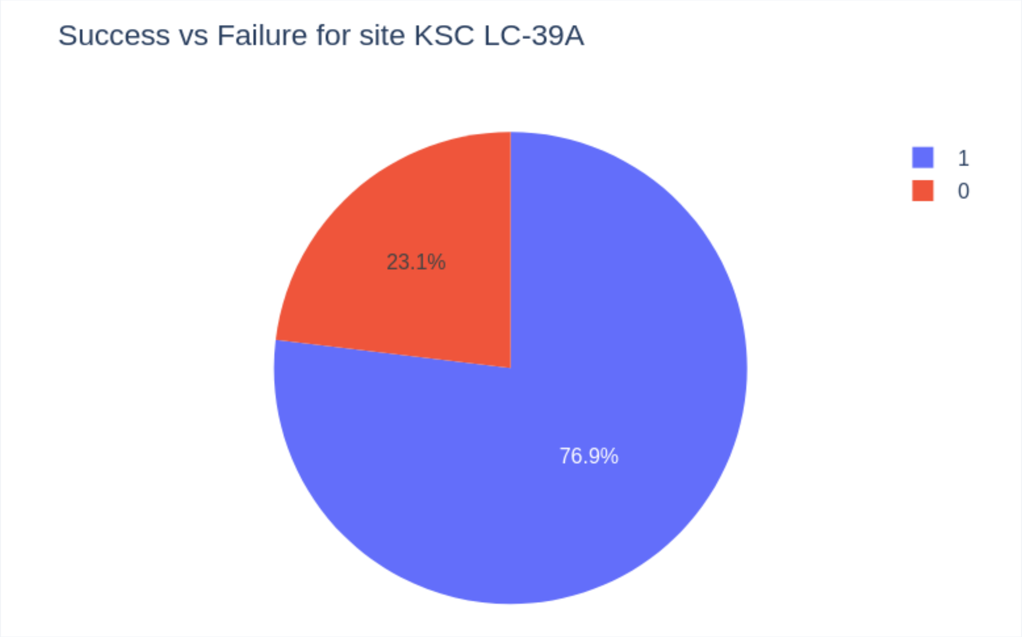
- CCAFS LC-40 also contributes a large share of successes.

- VAFB SLC-4E shows fewer successes due to fewer missions.



Success vs Failure Breakdown for KSC LC-39A

- The pie chart shows successful launches (1) vs failed launches (0) at KSC LC-39A.
- KSC LC-39A has a high success ratio (≈77%), making it the top-performing launch site.
- Failures are relatively few, representing only about 23% of launches.
- This confirms LC-39A as SpaceX’s most reliable site in the dataset.



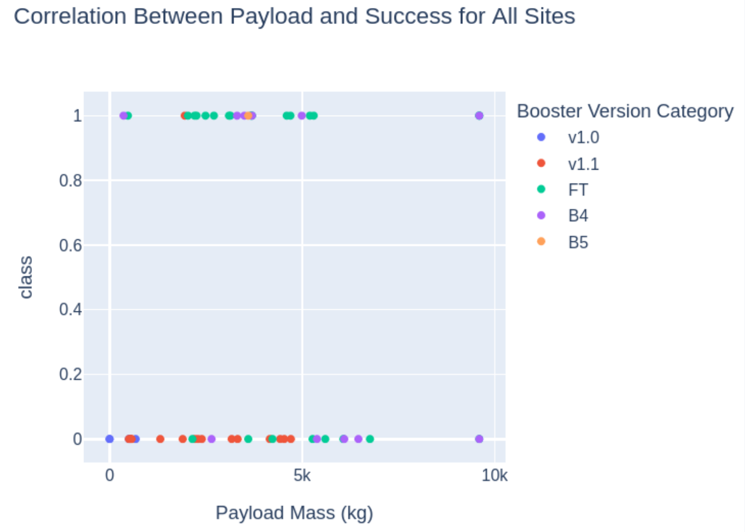
Payload vs Launch Outcome

- Successes occur across almost all payload levels, from very low to nearly 10,000 kg.

- FT (Full Thrust) and B4 boosters show many successful launches, indicating high reliability.

- Most failures happen in the lower to mid payload range (0–6000 kg) and with older boosters (v1.0, v1.1).

- There is no strong correlation between payload mass and success — booster version matters more than payload weight.

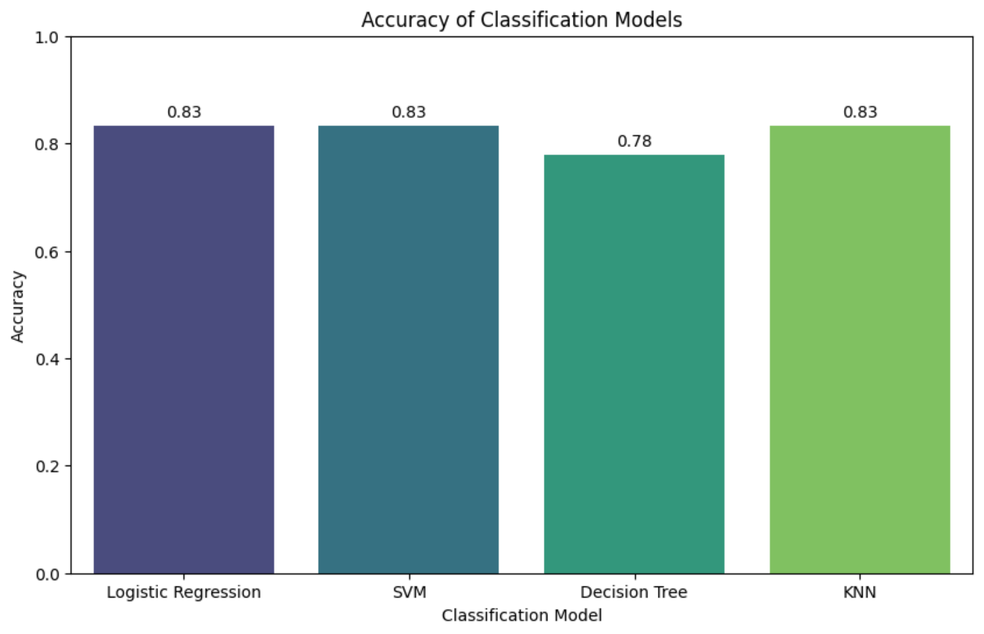


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistic Regression, SVM, and KNN all achieve 0.83 accuracy.
- Decision Tree performs slightly lower at 0.78.
- Overall, three models tie for the best performance.

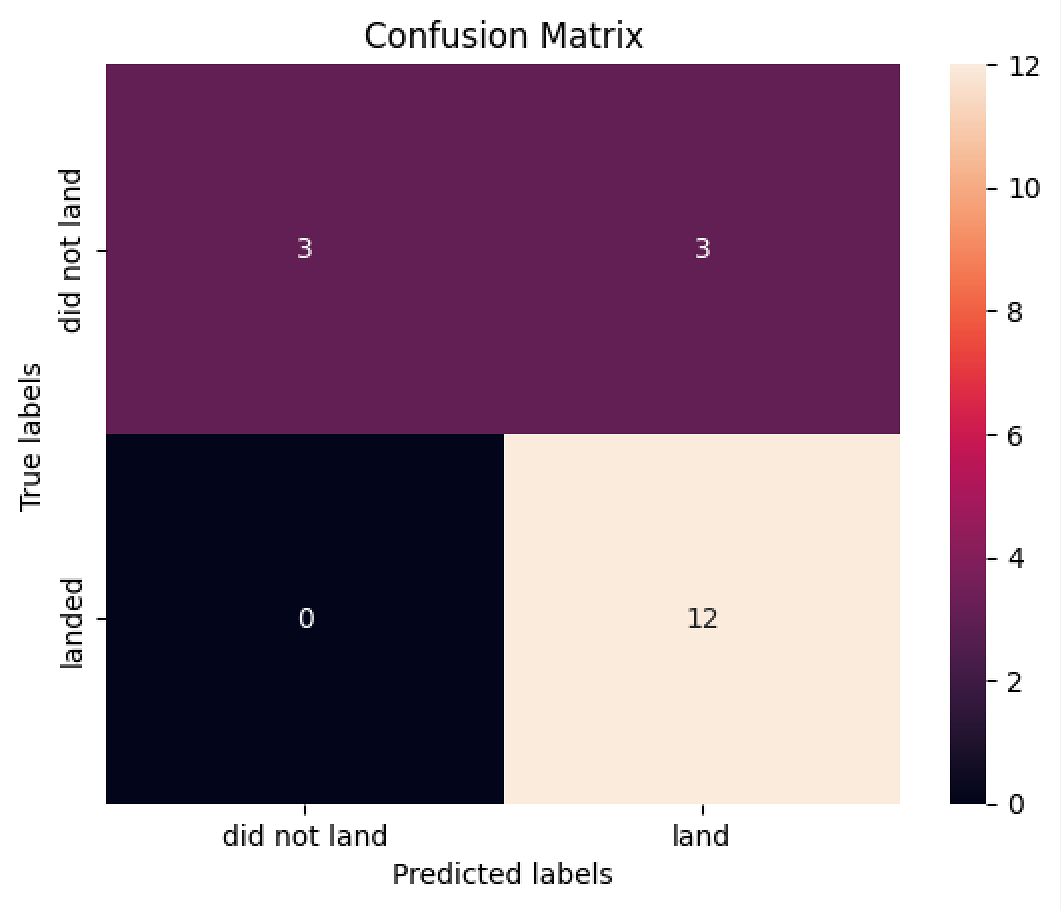


Confusion Matrix

- Correct predictions: 3 for “did not land” and 12 for “landed.”

- The model misclassified 3 cases as “landed” when they actually “did not land.”

- No false negatives: the model never predicted “did not land” for a true landing.



Conclusions

- This capstone project demonstrated the full data science workflow — from data collection and cleaning to exploratory analysis, visualization, modeling, and evaluation.
- Using real SpaceX launch data, we explored key factors that influence Falcon 9 landing success, including launch site, payload mass, and booster version.
- Visual analyses showed that some launch sites, such as KSC LC-39A, consistently achieve higher success rates, and that newer booster versions tend to perform better.
- Machine learning models (Logistic Regression, SVM, Decision Tree, and KNN) were built to predict landing outcomes.
- Most models achieved strong performance, with Logistic Regression, SVM, and KNN reaching 83% accuracy.
- The confusion matrix revealed that the selected model is particularly strong at predicting successful landings, with very few misclassifications.

Thank you!