IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paulo Henrique Pereira da Cunha
April 2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Insights drawn from EDA
- Launch Sites Proximities Analysis
- Build a Dashboard with Plotly Dash
- Predictive Analysis (Classification)
- Conclusions
- Appendix

Executive Summary

- This project explores whether it is possible to accurately predict the landing success of SpaceX's Falcon 9 first stage using historical mission data. Successful landings significantly reduce launch costs by enabling rocket reuse, which is a critical factor in modern aerospace economics.
- The analysis covers the full data science workflow, from data acquisition to machine learning, uncovering patterns and developing predictive insights. Visualizations and interactive dashboards offer both technical and intuitive understanding of the factors affecting landing outcomes.
- The final model, a Decision Tree Classifier, achieved a test accuracy of 90,4%, offering the best balance between performance and interpretability. Key drivers of success include payload mass, launch site, orbit type, and hardware reuse history.
- These findings not only validate the potential of machine learning in aerospace operations, but also provide a framework that can be adapted to other launch systems or mission planning scenarios.

Introduction

Project Goal:

To predict Falcon 9 first stage landing outcomes using historical SpaceX data.

Research Questions:

- What are the key factors influencing landing success?
- How accurate can a machine learning model be in predicting landings?
- Can interactive analytics improve understanding of launch patterns?

Context:

Successful landings reduce costs drastically. Understanding what drives them is of major business interest.

Section 1

Methodology

Methodology

Data Sources:

SpaceX REST API (launches, rockets, payloads, launchpads) and Wikipedia (web scraping)

Data Preparation:

Cleaning, merging, feature engineering (block version, reuse count), label creation

Exploratory Analysis:

Visualizations (Seaborn, Matplotlib) and SQL queries to explore launch outcomes

Interactive Analytics:

Folium maps (site markers, outcome colors, proximities) and Plotly Dash dashboard

Predictive Modeling:

Classification with Logistic Regression, Decision Tree, SVM, KNN; hyperparameter tuning with GridSearchCV; evaluation with accuracy and confusion matrix

Data Collection

API (SpaceX)

- Accessed endpoints for launches, rockets, payloads, cores, and launchpads using requests.get(); converted JSON responses into DataFrames with json_normalize().

Web Scraping (Wikipedia)

- Retrieved Falcon 9 launch history table using BeautifulSoup; parsed HTML elements, removed annotations and special characters, and structured data with Pandas.

Data Parsing

- Extracted nested attributes like payload mass, orbit type, landing success, booster version, and reuse count using custom Python functions.

Data Integration

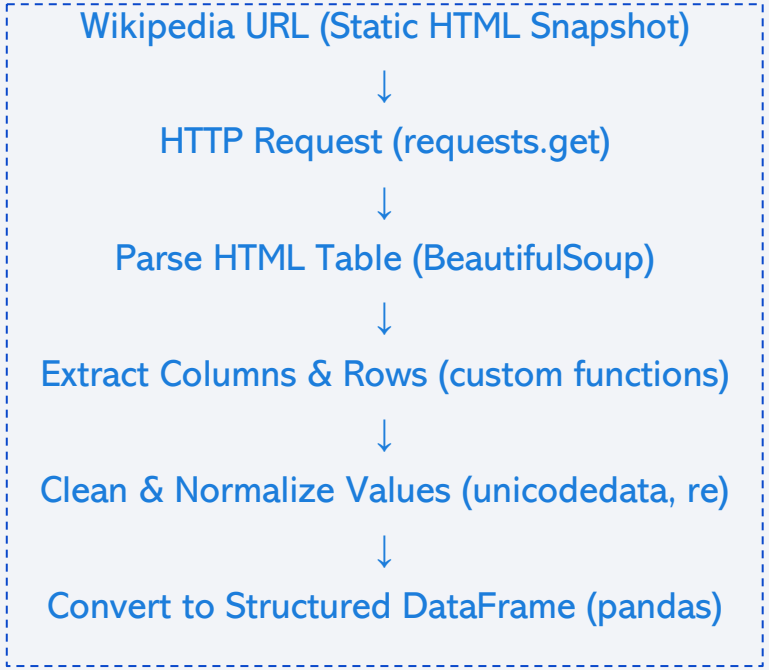
- Merged API and scraped datasets; standardized column names and data formats; removed duplicates and inconsistencies to create a clean master dataset.

Data Collection – SpaceX API

- Accessed historical launch data via SpaceX REST API
- Performed multiple GET requests using Python's requests library
- Used custom helper functions to retrieve and extract nested information (e.g., booster, payload, cores)
- Normalized JSON responses with pandas.json_normalize() to create structured DataFrames
- Created master dataset with key fields: Flight Number, Payload Mass, Orbit, Reuse, Block, Launch Site, Landing Outcome
- **GitHub URL:** https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

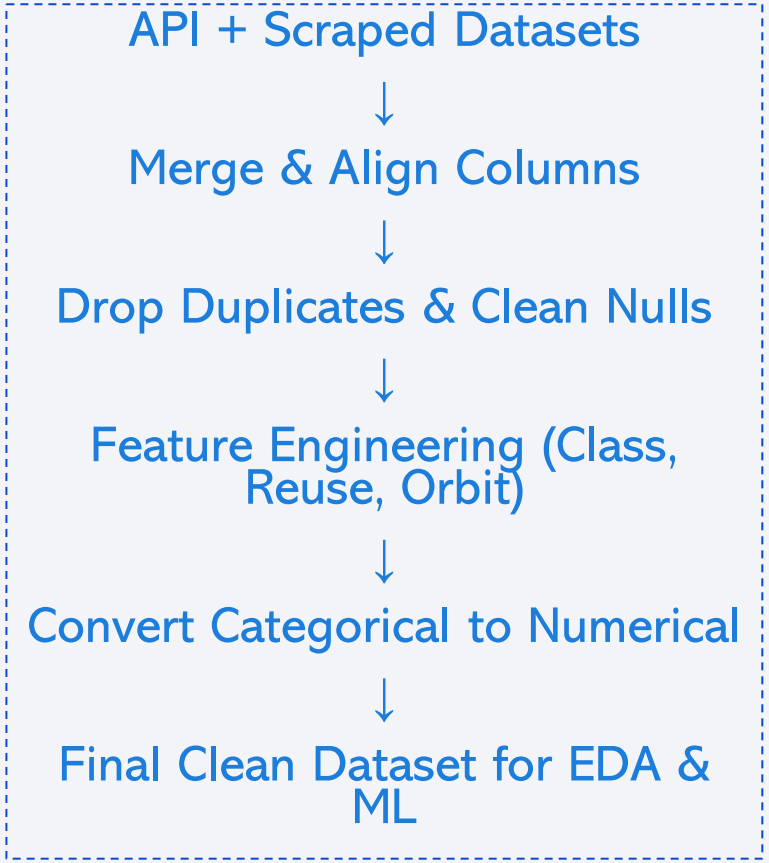
Data Collection - Scraping

- Extracted Falcon 9 and Falcon Heavy launch records from Wikipedia
- Targeted HTML tables using requests.get() and BeautifulSoup
- Parsed column headers and row values with custom functions (extract_column_from_header, booster_version, etc.)
- Cleaned and converted scraped HTML tables into structured Pandas DataFrame
- Used unicodedata and re to normalize values (e.g., payload mass in kg)
- **GitHub URL:** https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Merged datasets from API and web scraping to create a unified structure
- Removed irrelevant columns and duplicated entriesHandled missing values (e.g., LandingPad, PayloadMass, Outcome)
- Engineered new features:
  → Class (landing success: 1 or 0)
  → Block, ReusedCount, Orbit encoding
- Converted categorical values into numerical format for modeling
- Verified data types and cleaned inconsistencies across columns
- **GitHub URL:**
  https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- **Flight Number vs. Payload Mass (with Class)**
  → Scatter plot to observe patterns over time and relation with payload size.
  → Shows trend of higher success rates in later flights and lighter payloads.
- **Flight Number vs. Launch Site (with Class)**
  → Strip plot to assess performance by launch site over time.
  → Reveals which sites achieved more successful landings.
- **Payload Mass vs. Orbit (with Class)**
  → Box plot to compare landing success across orbit types and payload weights.
  → Highlights risky orbits and their relation to heavier loads.
- **Orbit vs. Outcome Count**
  → Bar plot to evaluate which orbit types are associated with success/failure.
  → Useful to identify operational challenges by mission type.

EDA with SQL

- **Count of Total Missions per Launch Site**
  → Identified the most frequently used sites
- **Average Payload Mass by Orbit Type**
  → Highlighted orbits with heavier vs. lighter payloads
- **Most Frequent Booster Version**
  → Revealed which hardware models were used most often
- **Earliest Successful Landing by Site**
  → Helped map timeline of reusable rocket success
- **Success vs. Failure Counts by Outcome Type**
  → Summarized overall landing performance
- **Sum of Payload Mass by Booster Version**
  → Quantified delivery capacity over time
- **GitHub URL:** https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **Launch Site Markers (folium.Marker)**
  → Marked geographic positions of each launch site
  → Added labels with site names for easy identification
- **Success/Failure Circles (folium.Circle)**
  → Represented individual launches by outcome
  → Color-coded for visual distinction (success vs. failure)
- **Proximity Lines (folium.PolyLine)**
  → Connected launch sites to nearby features (highways, coastlines, railways)
  → Helped evaluate logistical and geographical influences on launch outcomes
- **Map Layers and Zoom Control**
  → Enabled dynamic interaction with specific sites and regions
  → Provided intuitive spatial understanding of launch distribution
- **GitHub URL:**
  https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

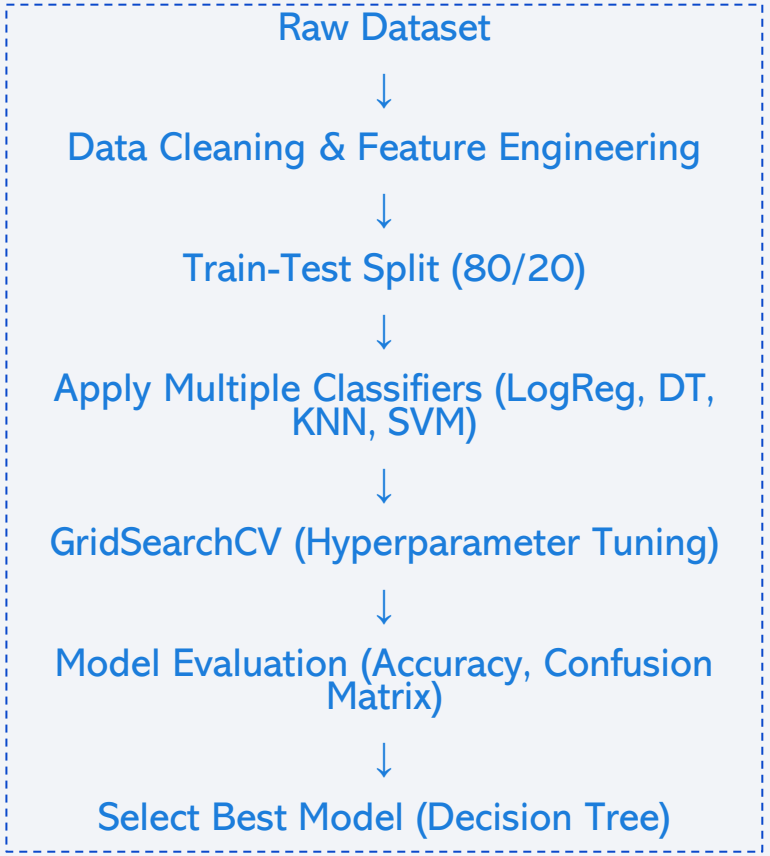
- **Pie Chart – Total Success Launches by Site**
  → Visualizes proportion of successful launches per site
  → Helps compare site performance at a glance
- **Scatter Plot – Payload Mass vs. Launch Success**
  → Displays correlation between payload size and mission outcome
  → Points colored by booster version category to enrich insight
- **Dropdown Filter – Launch Site Selector**
  → Enables selection of individual sites or all sites
  → Allows dynamic filtering of both plots
- **Payload Range Slider**
  → Interactive range selector for payload mass
  → Refines scatter plot to highlight mass-related trends
- **GitHub URL:**
  https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/spacex-dash-app.py

Why These Elements Were Added:

- To **explore key success factors interactively**
- To **compare sites**, payloads, and hardware versions visually
- To **enhance usability** for both analysts and stakeholders
- To **make insights accessible** without running any code

Predictive Analysis (Classification)

- Created a **classification pipeline** to predict Falcon 9 first stage landing success
- Selected features: **PayloadMass, Orbit, FlightNumber, Block, ReusedCount**, etc.
- Split data into **training and test sets** (80/20)
- **Applied and compared models**:→ Logistic Regression→ Decision Tree→ K-Nearest Neighbors→ Support Vector Machine (SVM)
- Used **GridSearchCV** for hyperparameter tuning
- Evaluated models using **accuracy score** and **confusion matrix**
- **Predictive Modeling – Best Model: Decision Tree (90.4%)**
- **GitHub URL:** https://github.com/paulohenriquecunha/FinalProjectDataScience/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



Results

EDA Highlights

- Success rate increased in later flights
- Lighter payloads → higher landing success
- Certain sites (e.g., KSC LC-39A) and orbits (e.g., LEO) had better outcomes

Interactive Analytics

- **Folium Map:** Launch sites with outcome markers; visualized proximity to coast, roads, rail
- **Dash Dashboard:** Pie chart (site success), scatter plot (payload vs. outcome), filters for site & mass

Predictive Modeling

- **Decision Tree:** Best model with 90.4% accuracy
- Key features: reuse count, orbit type, payload mass
- Confusion matrix showed strong prediction performance

Section 2

Insights drawn from EDA

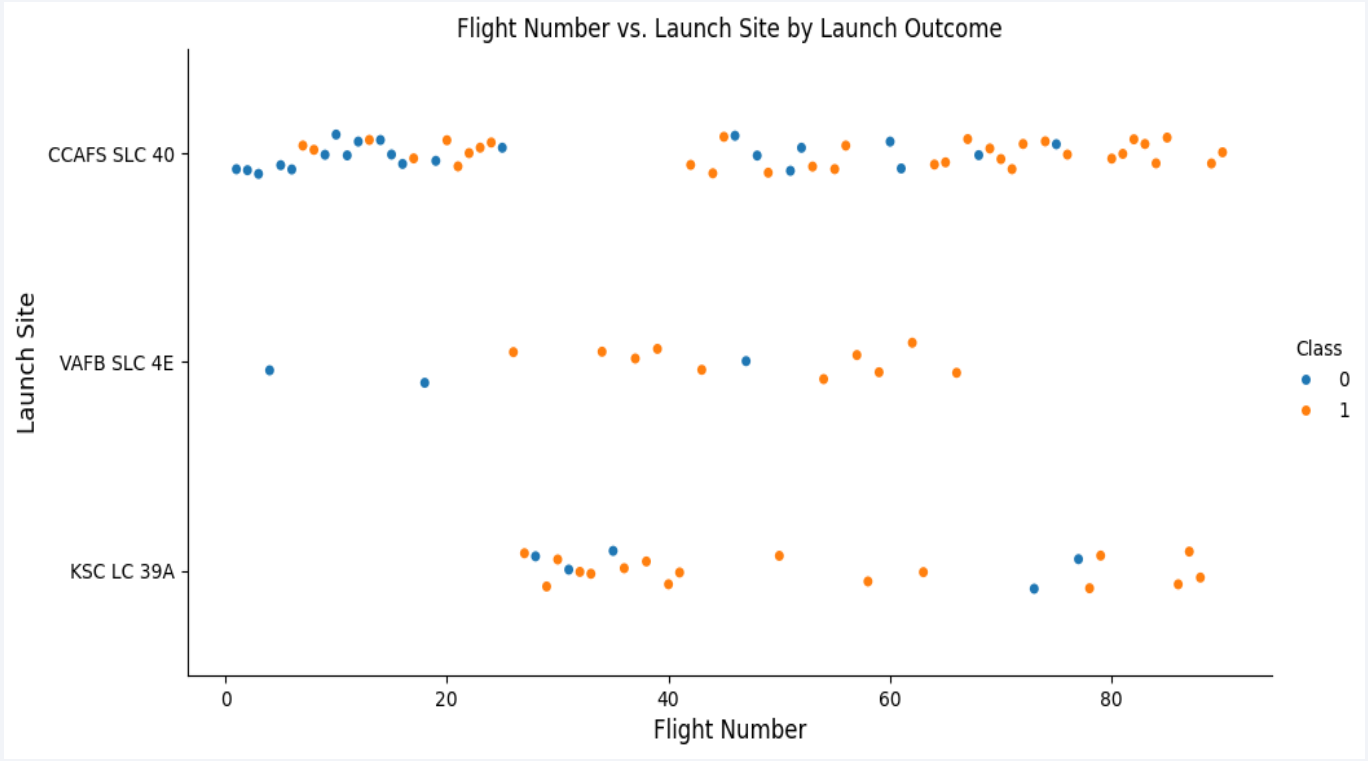
Flight Number vs. Launch Site

Plot:

- Scatter (strip) plot: Flight Number (x) vs. Launch Site (y)
- Color: Class (0 = failure, 1 = success)

Insight:

- The plot shows how launch success varies by site and flight number.
- Higher flight numbers (later missions) tend to have more successes, especially at CCAFS SLC-40 and VAFB SLC-4E.
- KSC LC-39A has a high success rate across all flights.



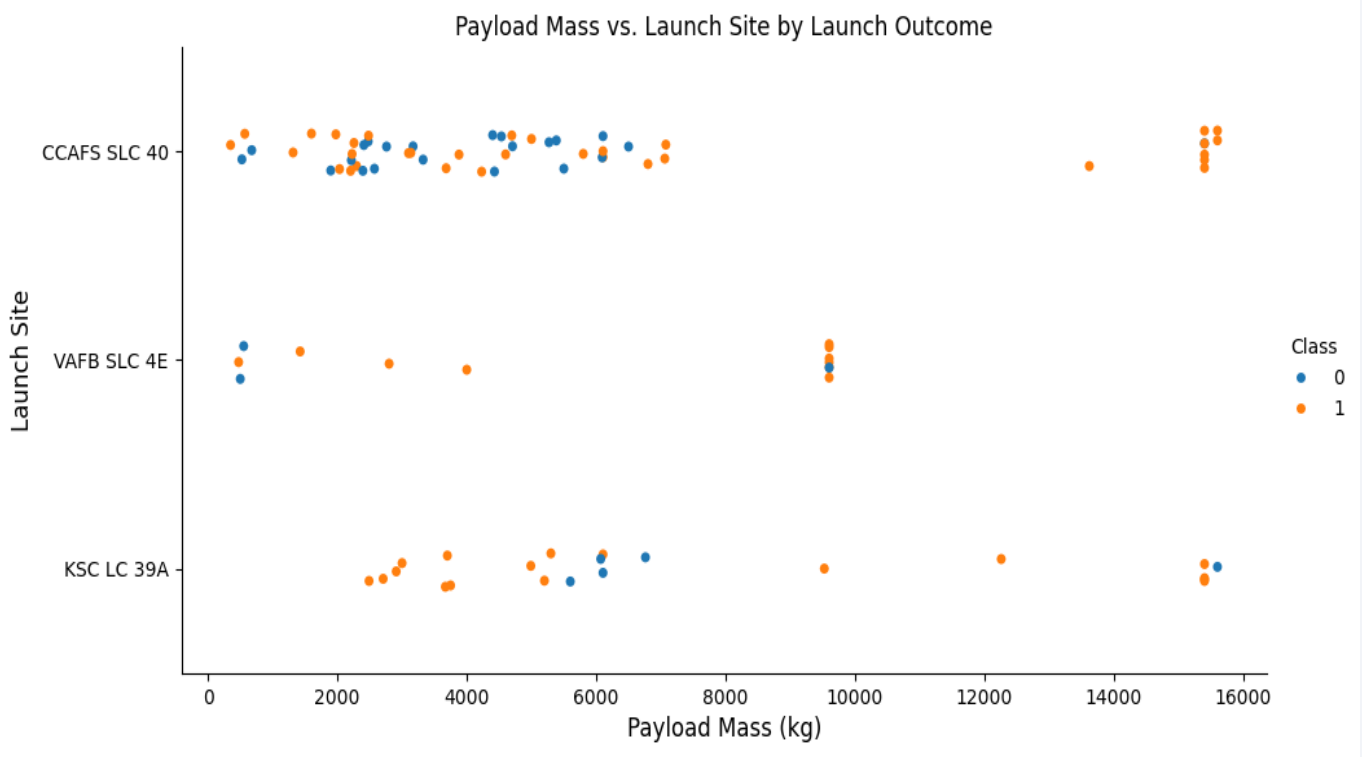
Payload vs. Launch Site

Plot:

- Scatter plot: Payload Mass (x) vs. Launch Site (y)
- Color: Class (landing success)

Insight:

- Heavier payloads (>10,000 kg) are launched from KSC LC-39A and CCAFS SLC-40.
- Success rate is high across all payload ranges, but failures cluster at lower payload masses.



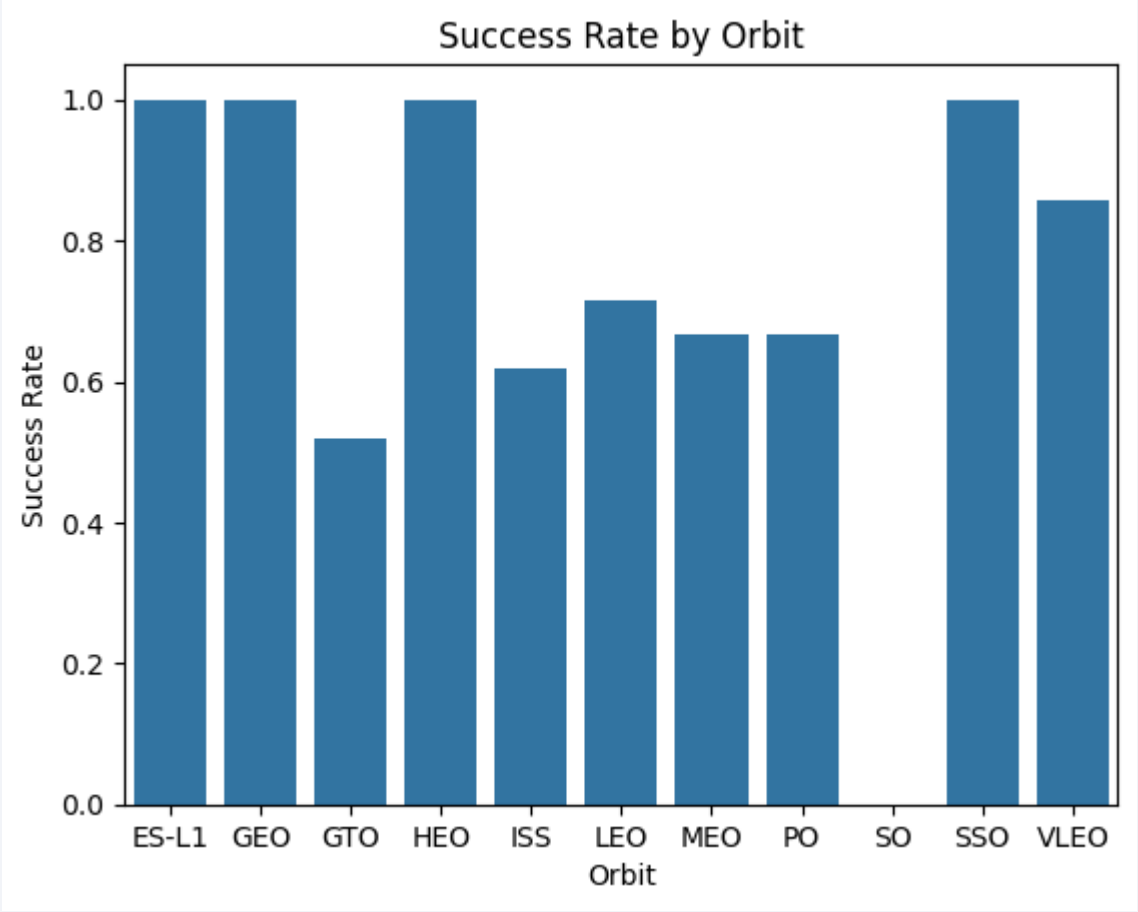
Success Rate vs. Orbit Type

Plot:

- Bar chart showing success count or success rate by orbit

Insight:

- Success rates vary significantly by orbit type.
- ES-L1, GEO, HEO, and SSO achieved 100% success, while LEO, ISS, MEO, VLEO, and PO showed consistent performance.
- GTO had the lowest success rate, indicating higher mission complexity.
- This suggests orbit type is a key factor in predicting landing outcomes.



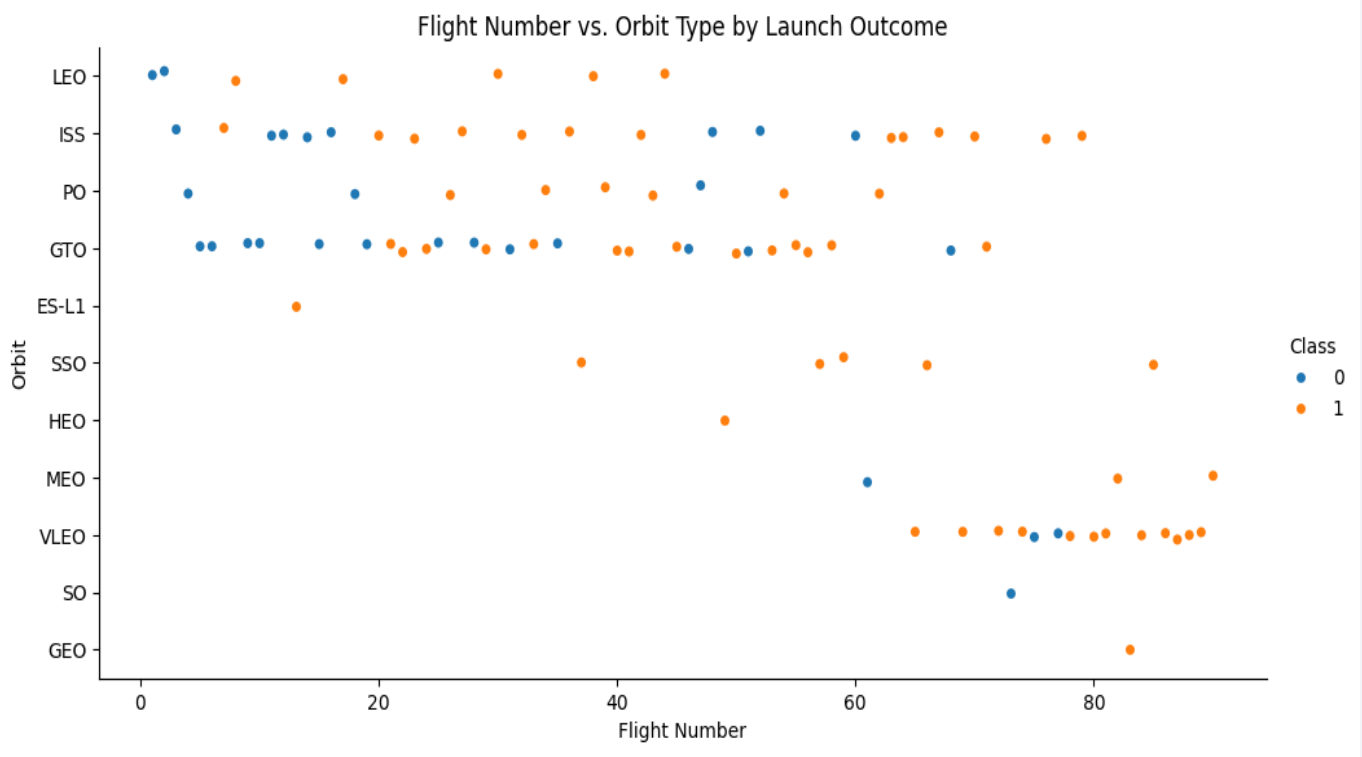
Flight Number vs. Orbit Type

Plot:

- Scatter plot: Flight Number (x) vs. Orbit (y)
- Color: Outcome (Class)

Insight:

- Distribution of orbit types changed over time.
- Some orbits became more frequent and successful in later flights.



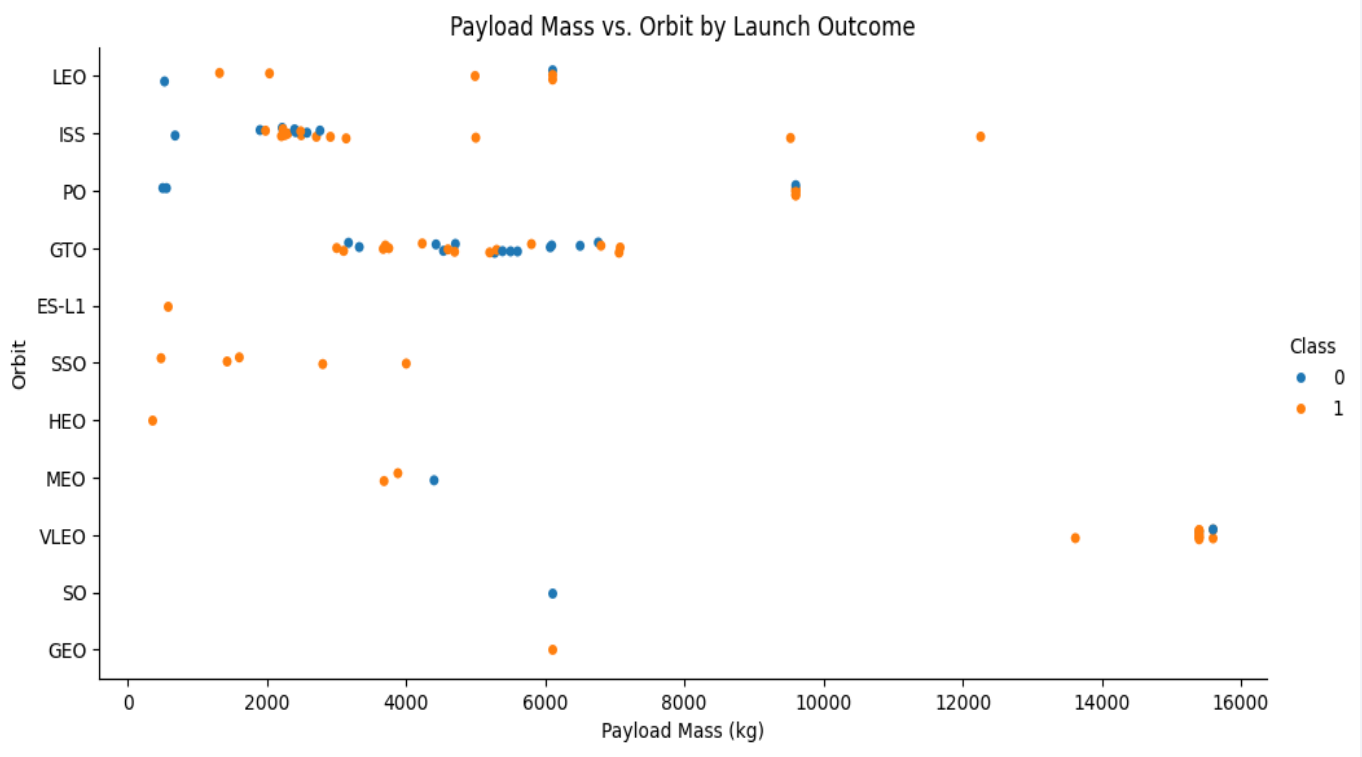
Payload vs. Orbit Type

Plot:

- Scatter plot: Payload Mass (x) vs. Orbit Type (y)
- Color: Success class

Insight:

- Different orbits handled different payload sizes.
- GTO (Geostationary Transfer Orbit) missions carry the heaviest payloads.
- Failures are rare but occur across all payload ranges.
- Some orbit-payload combinations were riskier.



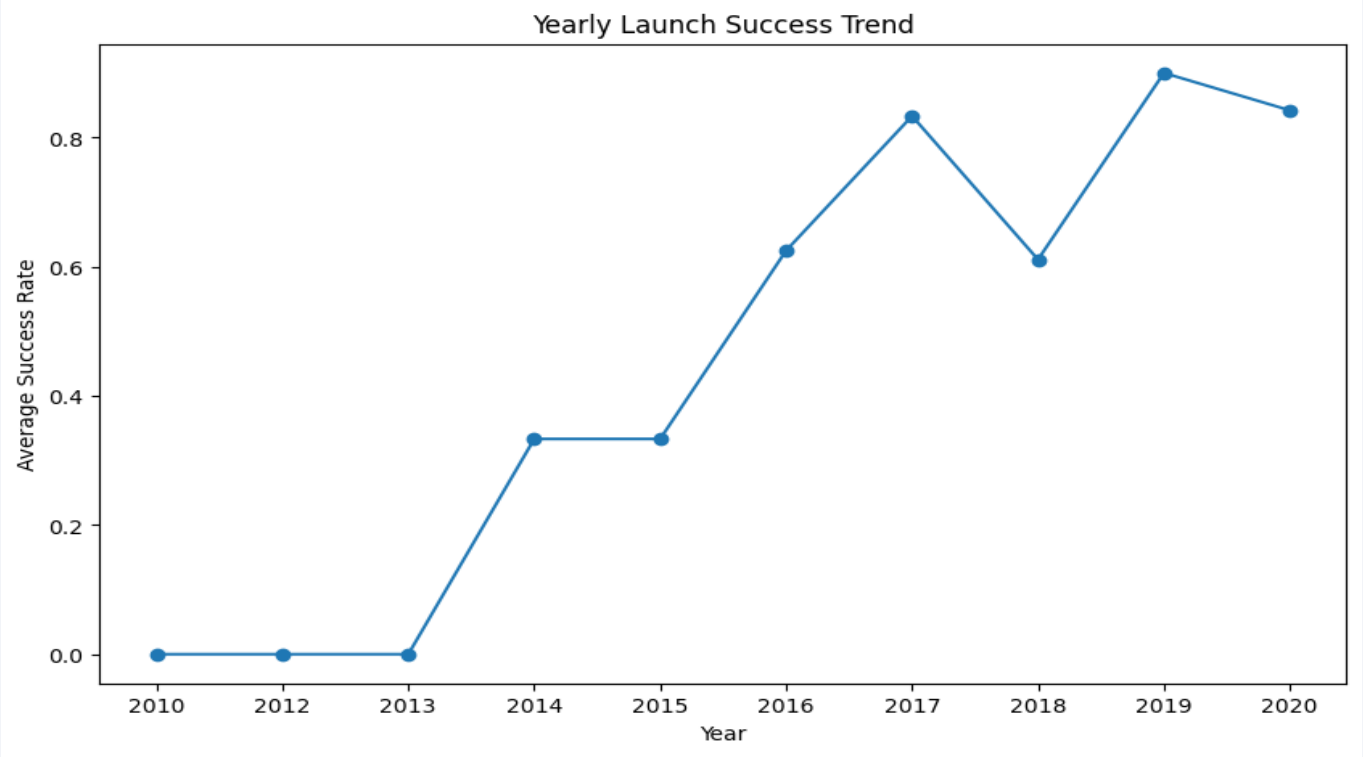
Launch Success Yearly Trend

Plot:

- Line chart: Year (x) vs. Average Landing Success (y)

Insight:

- Overall improvement in success rate year-over-year, peaking at ~90% after 2015.
- Clear trend of reliability increasing over time.



All Launch Site Names

SQL Query:

- SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

Unique Launch Sites Found:

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

Explanation:

- This query identifies all unique launch sites used by SpaceX. The results show 4 unique launch sites used by SpaceX, including 2 at Cape Canaveral (CCAFS) and one each at Vandenberg and Kennedy Space Center.

Launch Site Names Begin with 'CCA'

SQL Query:

- SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

Result:

- Displays 5 early missions from Cape Canaveral (CCAFS) launch sites.

Explanation:

- Filters for launch sites starting with 'CCA' (Cape Canaveral), showing the first 5 records including dates, payloads, and outcomes.

Total Payload Mass

SQL Query:

- SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%';

Total Payload:

- 48,213 kg

Explanation:

- Calculates the combined payload mass delivered for NASA's Commercial Resupply Services missions.

Average Payload Mass by F9 v1.1

SQL Query:

- SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1’;

Average Payload:

- ~2,928.4 kg

Explanation:

- Shows the typical payload capacity for SpaceX's Falcon 9 v1.1 booster variant.

First Successful Ground Landing Date

SQL Query:

- SELECT MIN(Date) AS First_Success FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)';

Date:

- 2015-12-22

Explanation:

- Marks the historic first time SpaceX successfully landed a booster vertically on land.

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query:

- SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)’ AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

Boosters:

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

Explanation:

- Lists boosters that achieved drone ship landings with medium-weight payloads (4-6 metric tons).

Total Number of Successful and Failure Mission Outcomes

SQL Query:

- SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome;

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Explanation:

- Demonstrates SpaceX's high success rate.

Boosters Carried Maximum Payload

SQL Query:

- SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

Result:

- 12 records of F9 B5 boosters carrying 15,600 kg

Explanation:

- Identifies the most powerful booster configuration and its payload capacity.

2015 Launch Records

- **SQL Query:**
  SELECT substr(Date,6,2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE
  WHERE substr(Date,0,5)='2015' AND Landing_Outcome = 'Failure (drone ship)';

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanation:

- Shows SpaceX's early challenges with drone ship landings in 2015 before perfecting the technology.

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

SQL Query:

- SELECT Landing_Outcome, COUNT(*) AS Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20’ GROUP BY Landing_Outcome ORDER BY Count DESC;

Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanation:

- Tracks the evolution of landing attempts during SpaceX's experimental phase.


Section 3
Launch Sites
Proximities Analysis

Global Distribution of SpaceX Launch Sites

Description:

This map displays the locations of all four SpaceX launch sites:

- CCAFS LC-40 and CCAFS SLC-40 in Florida (USA).
- KSC LC-39A in Florida (USA).
- VAFB SLC-4E in California (USA).

Key Findings:

- All launch sites are located near coastlines, facilitating safe launch trajectories over water.
- Sites are clustered in low-latitude regions (close to the Equator), which is optimal for achieving orbital velocity efficiently.

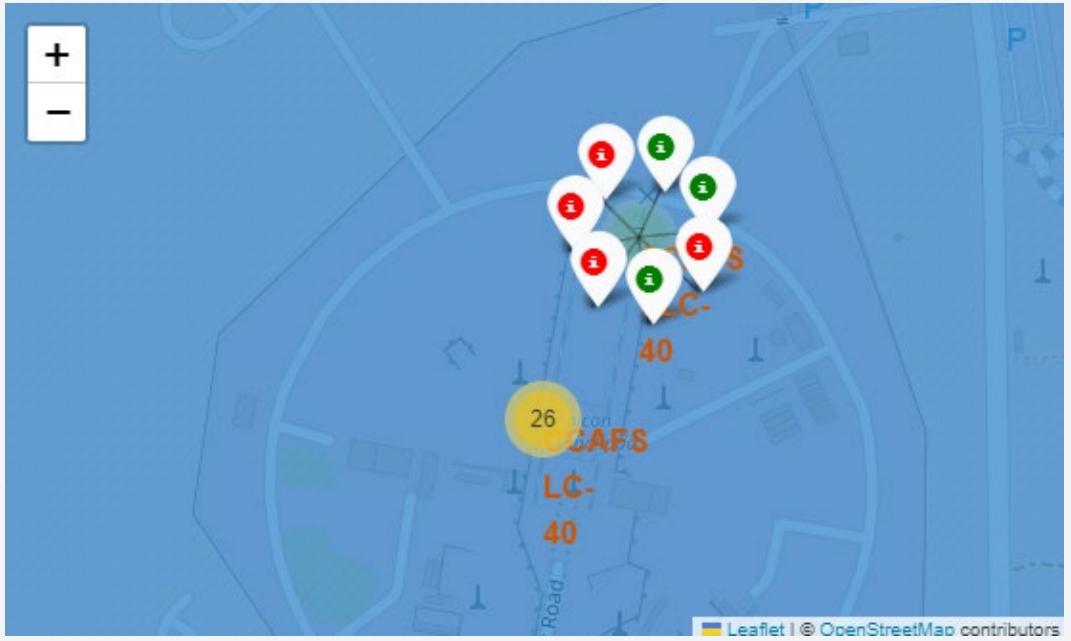
Launch Outcomes Visualized by Color Markers

Description:

- Green markers: Successful launches (class=1).
- Red markers: Failed launches (class=0).
- Marker clusters simplify visualization for high-density launch records (e.g., CCAFS SLC-40).

Key Findings:

- KSC LC-39A shows a high success rate (mostly green markers).
- CCAFS SLC-40 has a mix of outcomes, indicating variability in launch success.
- Proximity to support infrastructure (e.g., NASA facilities) may correlate with higher success rates.



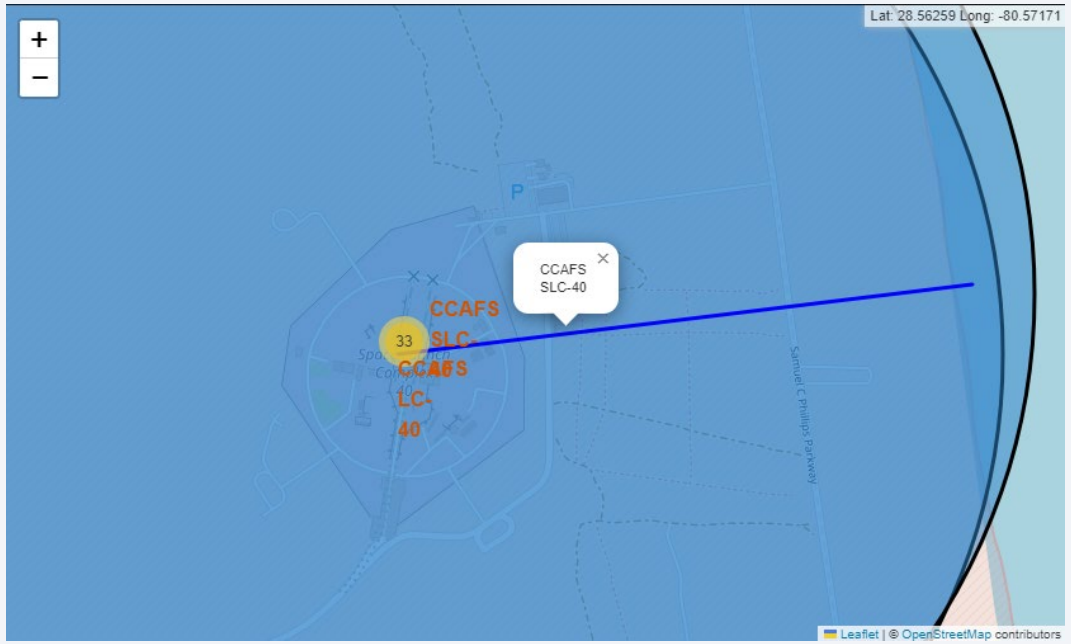
Launch Site Proximity to Coast

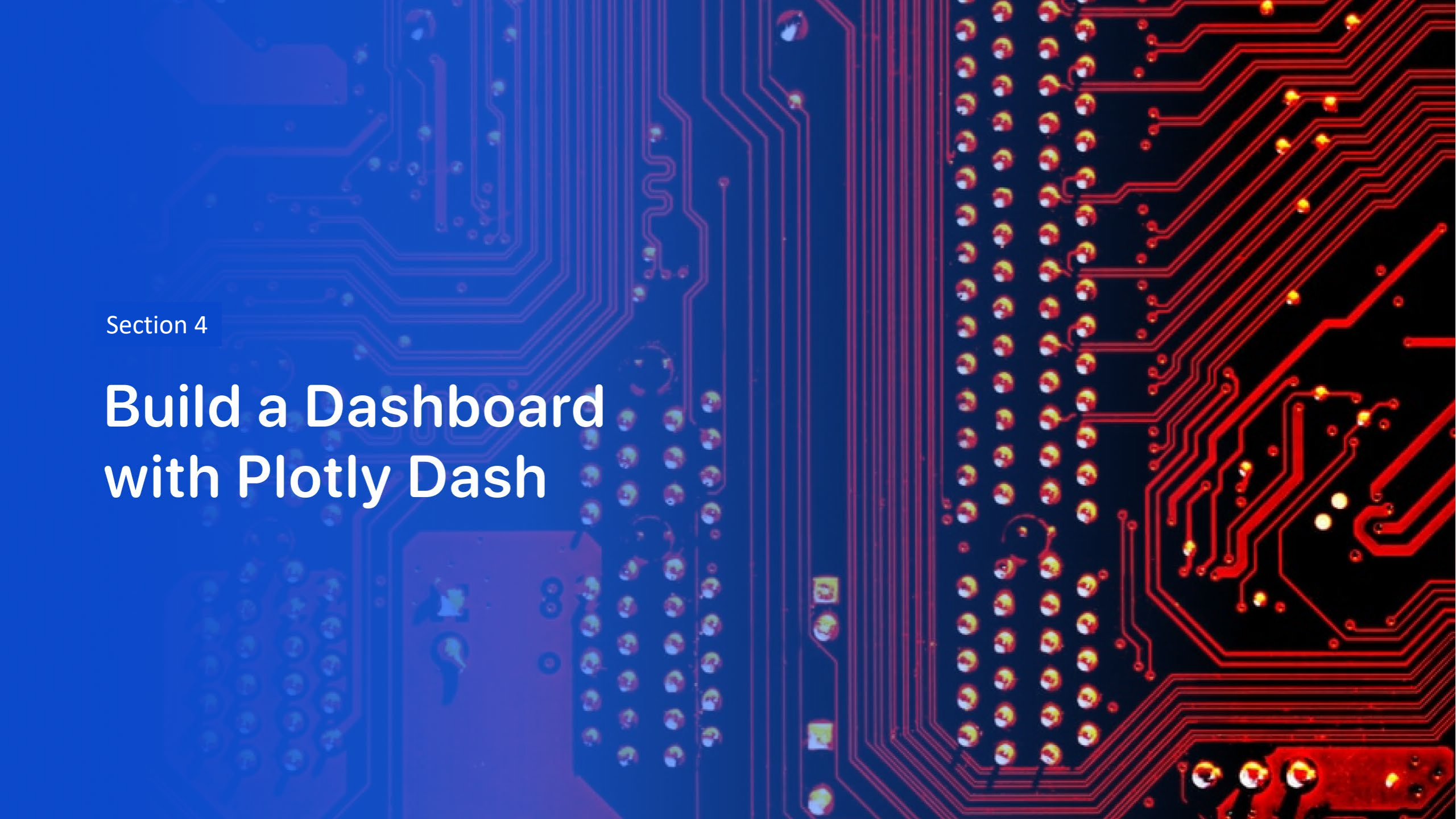
Description:

- This screenshot shows a launch site and its distance from coastline. Using polylines, we calculate and display distances to nearby locations, such as a highway and railway

Key Findings:

- Launch sites are strategically placed near logistical infrastructure while maintaining safe distances from populated areas.
- Coastal proximity minimizes risks to human settlements during failures.

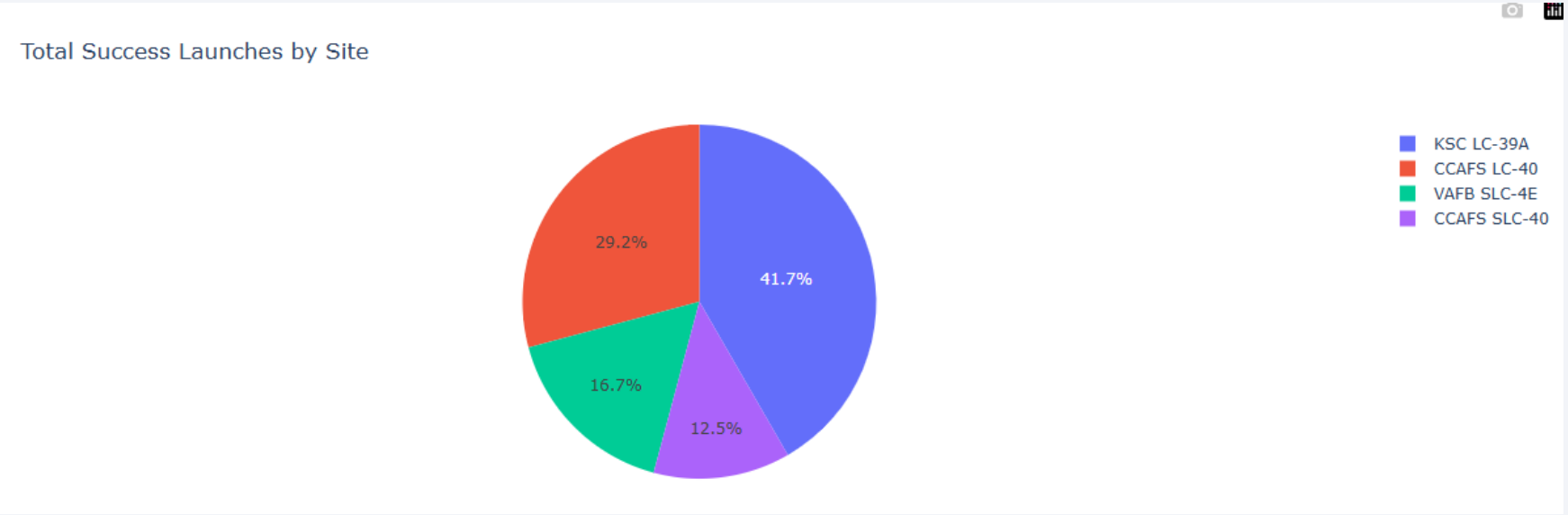



Section 4
Build a Dashboard with Plotly Dash

Launch Success Count by Site

Explanation:

- This pie chart displays the total number of successful launches across all SpaceX launch sites.



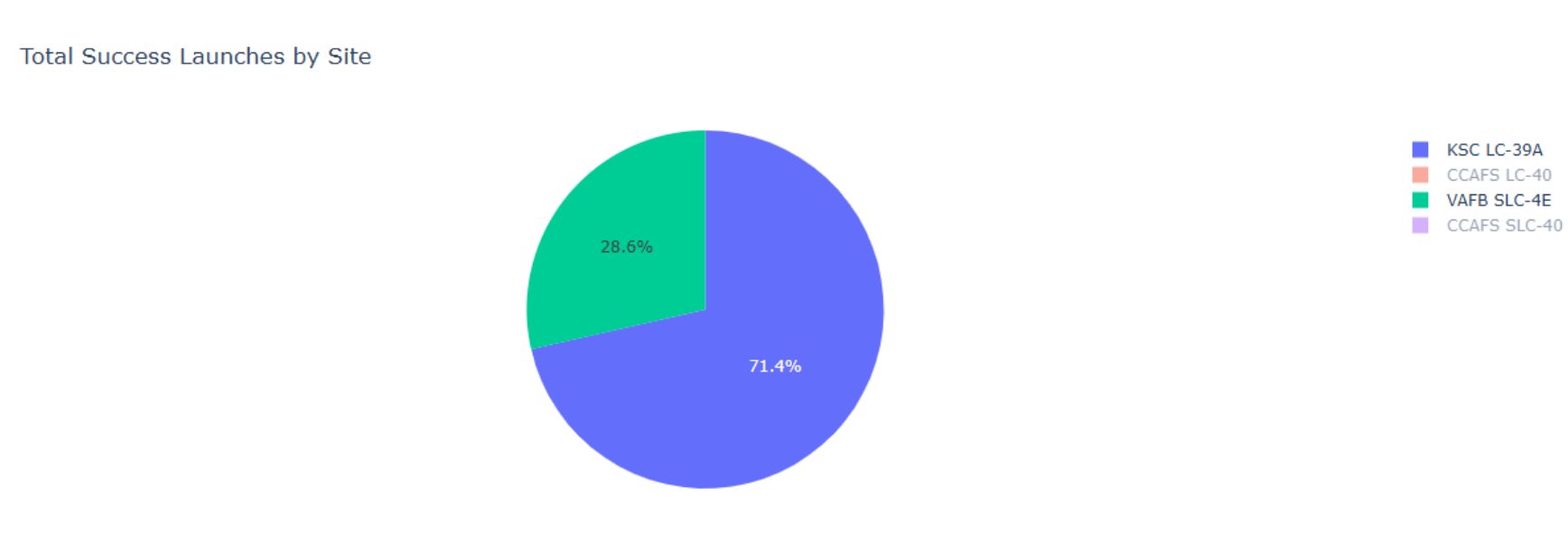
Key findings:

- KSC LC-39A has the highest number of successes (41.7%), followed by CCAFS LC-40.
- This highlights site-level performance and allows us to compare reliability among different launch locations.

Highest Launch Success Ratio Site

Explanation:

- This chart zooms in on the launch site with the highest success rate — KSC LC-39A, which has a 71.4% success rate, in contrast to VAFB SLC-4E.



Key findings:

- This site consistently demonstrates better launch outcomes.
- Ideal location to prioritize for future missions due to reliability.

Payload vs Launch Outcome Scatter Plot

Explanation:

- This scatter plot visualizes the correlation between payload mass and launch outcome across all sites.

Key findings:

- Successful launches (class = 1) are spread across payload ranges.
- Booster versions like B5 show higher success rates for heavy payloads (4000–6000 kg).
- The interactive slider allows users to isolate trends for specific mass ranges.

Section 5

Predictive Analysis (Classification)

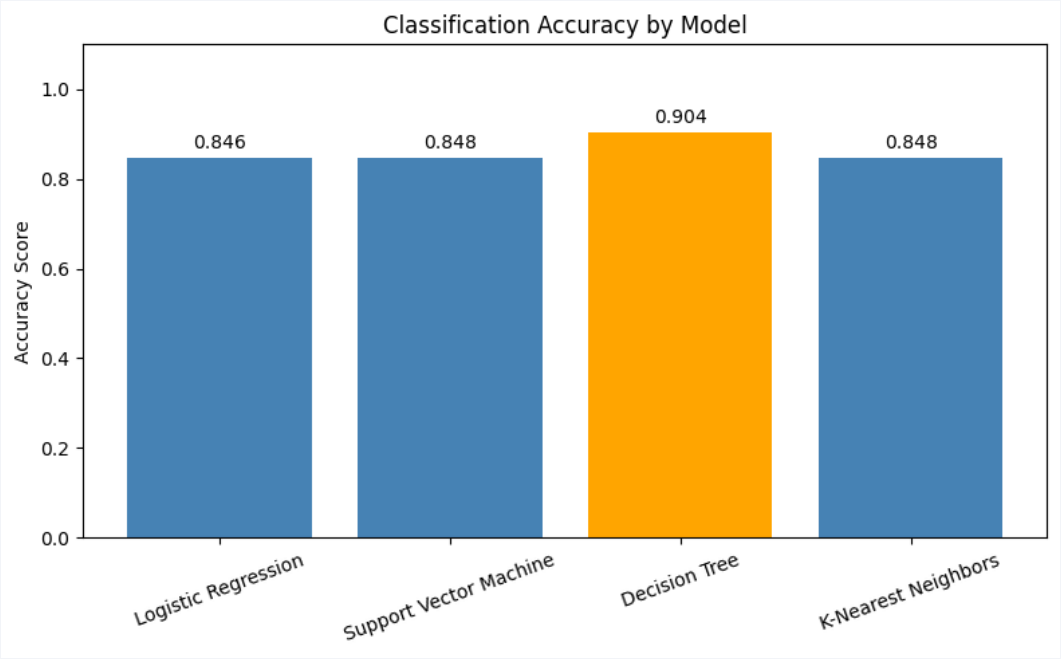
Classification Accuracy

Description:

- We compared Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbors Classifier using GridSearchCV.
- The best performing model was Decision Tree with 90.4% accuracy.

Insight:

- The Decision Tree model outperformed all others in accuracy, making it the most suitable for our binary classification task involving SpaceX launch outcome prediction.



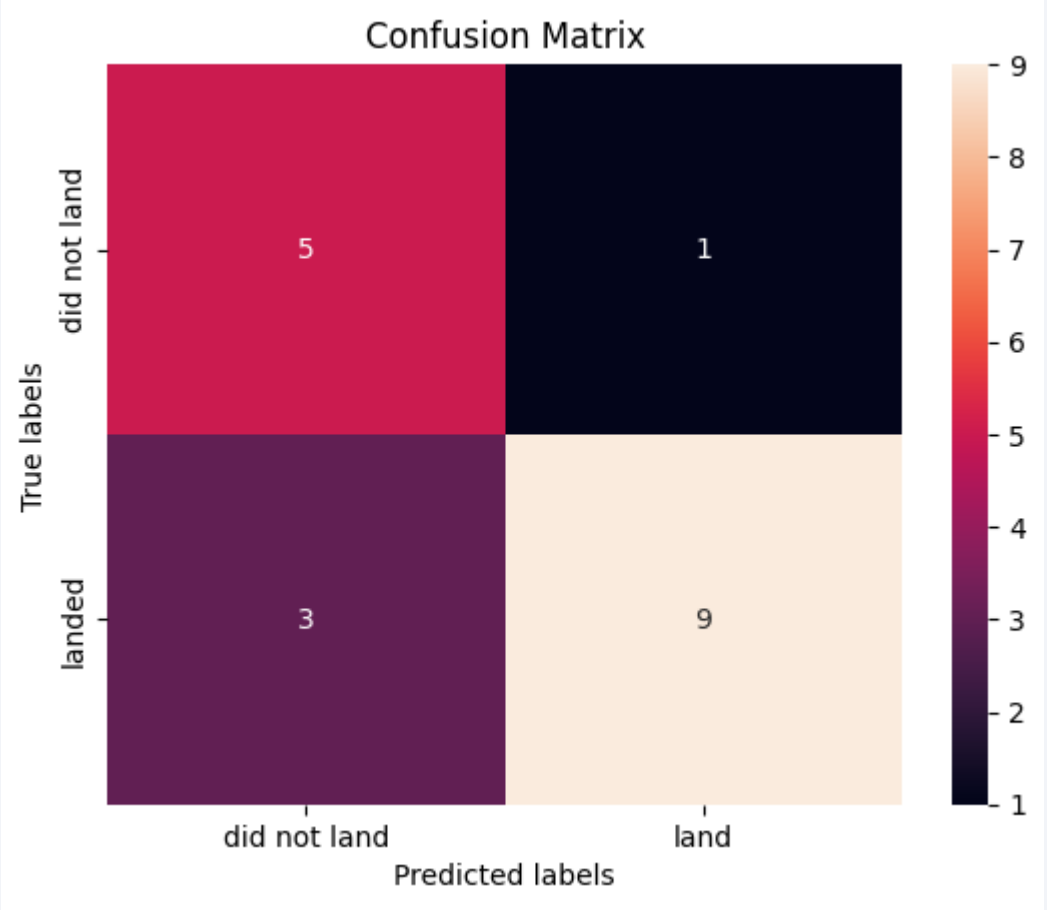
Confusion Matrix

Explanation:

- The model correctly classified:
  → 9 successful landings
  → 5 failed landings
  → 1 launch was incorrectly predicted as success
  → 3 successful landings were missed by the model

Insight:

- The model Decision Tree shows strong precision (90.4%) on predicting landings and a balanced performance across both classes, making it suitable for this binary classification task.



Conclusions

- This project successfully predicted Falcon 9 first stage landing outcomes with high accuracy, validating the potential of machine learning in aerospace applications.
- It demonstrated the full data science workflow — from API and web scraping to feature engineering, analysis, visualization, and model development.
- Key patterns in mission success were uncovered through SQL queries and exploratory visualizations, offering valuable operational insights.
- Interactive tools such as Folium maps and Plotly Dash dashboards made complex data more accessible and actionable.
- Overall, this project highlights how data-driven approaches can support strategic decision-making and innovation in space launch operations.

Appendix

GitHub Repository:

- https://github.com/paulohenriquecunha/FinalProjectDataScience
- Includes all notebooks, scripts, and visualizations

Datasets Used:

- SpaceX Public API
- Wikipedia (web-scraped Falcon 9 & Falcon Heavy launch data)

Python Libraries:

- Pandas, Seaborn, Matplotlib, Folium, Plotly Dash, Scikit-learn, SQLAlchemy

Tools & Platforms:

- JupyterLab, Google Colab, GitHub, Visual Studio Code

Thank you!