IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Izael Castro
May 2025
SpaceX Launch Analysis and Prediction

Outline

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

Executive Summary

- This project analyzes SpaceX launch data to identify key factors influencing mission success.
- Data sources include SpaceX API, web scraping, and CSV datasets.
- Techniques applied: EDA (visual + SQL), interactive visualization (Folium, Dash), and classification models.
- Key insights:
  - ❑ B5 boosters have the highest success rate.
  - ❑ KSC LC-39A is the most successful launch site.
  - ❑ Payloads between 4000–6000 kg are associated with better outcomes.

Introduction

- SpaceX's mission involves reducing rocket costs via reusability.
- This project aims to answer:
  - ❑ Which launch sites are most reliable?
  - ❑ What payload range affects success?
  - ❑ Which booster versions are most efficient?

Section 1

Methodology

Methodology

Executive Summary

- Data Collection:
  - SpaceX REST API
  - Web scraping (HTML tables)
  - Supplementary CSV files
- Data Wrangling:
  - Cleaning, normalizing, merging datasets
- EDA:
  - Statistical charts and SQL analysis
- Interactive Visual Analytics:
  - Maps (Folium), dashboards (Dash)
- Classification:
  - Logistic Regression, Decision Trees, KNN, SVM

Data Collection

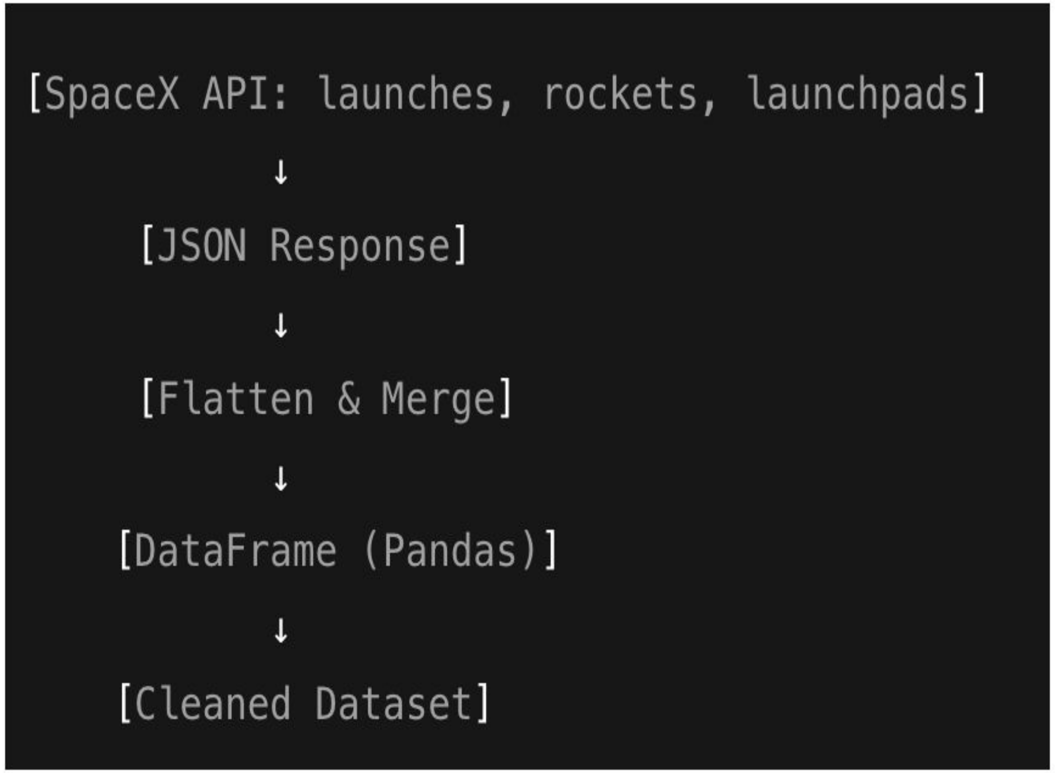
1. Queried SpaceX public API (v4) using Python `requests.get`;
2. Extracted key info: launch ID, payload mass, site, outcome;
3. Accessed extra metadata from `/rockets` and `/launchpads`;
4. Merged and normalized nested JSON structures;
5. Converted IDs to readable labels (e.g., rocket names);
6. Formatted into structured tabular format using Pandas;
7. Cleaned missing or invalid entries (e.g., null payload);

Data Collection – SpaceX API

We collected launch data by querying multiple SpaceX REST API endpoints, including `/launches`, `/rockets`, and `/launchpads`, using Python requests.

The responses were received in JSON format, parsed, and normalized. We then merged rocket and launchpad metadata, and structured the dataset using Pandas.

This process enabled a clean, analysis-ready dataset with fields such as launch site, payload mass, and landing success class.



📎 GitHub Notebook: [01-data-collection.ipynb]

→ Includes full API code and resulting dataset preview.

Data Collection - Scraping

We scraped launch records from the Wikipedia page: "List of Falcon 9 and Falcon Heavy launches".

Using Python's `requests` library, we fetched the HTML content of the page. Then, with `BeautifulSoup`, we parsed the HTML to locate and extract the relevant launch record table.

The table was converted into a Pandas DataFrame, normalized, and preprocessed. This step enriched the dataset with historical launch information not available via API.

📎 GitHub Notebook: 02-web-scraping.ipynb
→ Includes request, HTML parsing, and DataFrame creation.

```
[Wikipedia Launch Page]
          ↓
[Fetch HTML with requests]
          ↓
[Parse with BeautifulSoup]
          ↓
[Extract HTML Table]
          ↓
[Convert to Pandas DataFrame]
          ↓
[Clean & Save]
```

Data Wrangling

Data Wrangling Process Overview

- **Data Sources**: Two datasets were used — one with mission outcomes and another with launch site coordinates.
- **Missing Values**: Detected and handled using pandas .isnull() method. Null values were either filled or dropped, depending on context.
- **Data Types**: Converted columns like dates and booleans to appropriate formats.
- **Derived Columns**: Created new columns (e.g. Class) to simplify modeling later.
- **Standardization**: Renamed and cleaned categorical values for consistency.
- **Data Merge**: Combined datasets to associate launch data with geographical locations.
- **Output**: A clean dataset ready for EDA and machine learning.

```
Raw Data 1          Raw Data 2
    |                   |
    ▼                   ▼
Handle Missing     Format Types
    |                   |
    └──▶ Merge Datasets ◀──┘
              |
              ▼
      Standardize & Clean
              |
              ▼
      Final Dataset Ready
```

📎 GitHub Notebook: [03-data-wrangling.ipynb]

EDA with Data Visualization

In this notebook, an initial exploratory data analysis (EDA) was performed using line plots to gain insights into the distribution and behavior of key variables from the SpaceX dataset. The main actions included:

- **Temporal Analysis**: Line plots were generated to observe how launch metrics evolved over time. This included visualizing the frequency of launches and their outcomes across different dates.

- **Payload Mass Trends**: Plots were used to explore variations in payload mass, which could help identify operational patterns or performance limits related to different launch configurations.

- **Orbit Distribution**: The EDA examined how orbital types varied over time and how they relate to other mission characteristics.

- **Launch Sites and Booster Versions**: Plots also helped reveal how different launch sites and rocket versions contributed to the success rates and number of missions.

These visualizations supported the identification of general patterns and trends, offering a foundation for deeper statistical modeling and predictive analysis in the next phases of the project.

📎 GitHub Notebook:04-eda-visualizations.ipynb

EDA with SQL

SQL Queries Summary

- **Selected all records** from the SpaceX dataset to confirm data structure and content.
- **Counted the number of successful landings** (Class = 1).
- **Filtered records** by specific landing outcomes (e.g., "None" or "True" in Outcome column).
- **Aggregated number of successful launches** by launch site.
- **Retrieved unique booster versions** for exploratory purposes.
- **Filtered launches** with payload mass greater than a threshold (e.g., 4000 kg).
- **Filtered records by multiple conditions**, such as launch success and orbit type.

📎 GitHub Notebook: [05-sql-analysis.ipynb]

Build an Interactive Map with Folium

- **Circle Marker**: A circle was placed over the **NASA Johnson Space Center** to highlight the base reference location.
  → *Reason:* Used as the starting zoom point and context anchor for the map.
- **Text Label with DivIcon**: A **custom text label** was added to the map using DivIcon to identify **NASA JSC**.
  → *Reason:* Improves clarity and readability for viewers by displaying site names directly on the map.
- **Multiple Circles for Launch Sites**: Each SpaceX launch site was marked with a **circle overlay** using their latitude and longitude.
  → *Reason:* To visualize all major launch sites on a single interactive map.
- **MarkerCluster Plugin**: Initialized to group close markers interactively.
  → *Reason:* Enhances performance and user experience when viewing densely packed launch locations.
- **Color-coded Markers by Launch Outcome**: Launch results were encoded with color:
  - 🟢 **Green** for successful launches (class = 1)
  - 🔴 **Red** for failed launches (class = 0)
    → *Reason:* Immediate visual feedback on launch success across geographic locations.
- **Mouse Position Plugin**: Displays coordinates interactively as the user hovers over the map.
  → *Reason:* Helps users locate precise geographic points dynamically.

📎 GitHub Notebook: [06-folium-maps.ipynb]

Build a Dashboard with Plotly Dash

- **Launch Site Dropdown Selector**
 → *Allows users to choose between “All Sites” or a specific launch site.*
 → *Purpose:* Enables dynamic filtering of plots based on user-selected location.

- **Success Pie Chart**
  - When "All Sites" is selected: displays total successful launches by site.
  - When a specific site is selected: shows proportion of successes vs failures for that site.
 → *Purpose:* Provides immediate feedback on performance distribution across launch locations.

- **Payload Range Slider**
 → *Allows interactive selection of a payload mass range (from 0 to 10,000 kg).*
 → *Purpose:* Offers users control over the payload scope analyzed in the scatter plot.

- **Payload vs. Success Scatter Plot**
  - Displays success (class) as a function of payload mass.
  - Color-coded by **Booster Version Category**.
  - Reactively updates based on both the dropdown and slider inputs.
 → *Purpose:* Reveals potential correlations between payload, booster version, and launch outcome.

📎 GitHub: spacex_dash_app.py

Predictive Analysis (Classification)

- **Dataset Preparation**
  - Features (X) and target (Y) were loaded from preprocessed datasets.
  - The data was **standardized** using StandardScaler to normalize feature ranges.
- **Train/Test Split**
  - The dataset was split using an **80/20 train-test ratio** with random_state=2 for reproducibility.
- **Algorithms Tested**
  - ✅ Logistic Regression
  - ✅ Support Vector Machine (SVM)
  - ✅ Decision Tree Classifier
  - ✅ K-Nearest Neighbors (KNN)

📎 GitHub Notebook:[07-predictive-modeling.ipynb]

- **Evaluation Metrics**
  - Each model's performance was evaluated using **accuracy score** and **confusion matrix**.
  - A custom plot_confusion_matrix() function was used to visually assess misclassifications.
- **Hyperparameter Tuning**
  - Used **GridSearchCV** for tuning:
    - C parameter in SVM
    - max_depth in Decision Tree
    - n_neighbors in KNN
  - The best models were re-evaluated with optimal parameters.
- **Best Performing Model**
  - The final best model was selected based on **test set accuracy** and interpretability.

Results

- **Exploratory Data Analysis (EDA)**
  The EDA revealed important insights such as:
  - Which launch sites had higher success rates.
  - How payload mass and orbit type correlated with launch outcomes.
  - Temporal trends in mission frequencies and outcomes.
- **Interactive Visual Analytics**
  Interactive dashboards built with **Folium** and **Plotly Dash** enabled:
  - Visual geolocation of launch sites and their outcomes.
  - Real-time filtering of missions by site, payload range, and booster version.
  - Intuitive analysis through dropdowns, sliders, and color-coded markers.
- **Predictive Modeling**
  Four classification algorithms were trained and evaluated.
  After hyperparameter tuning with GridSearchCV, the best performing model was selected based on:
  - Accuracy score on the test set.
  - Confusion matrix visualization.
  - Simplicity and interpretability.

Section 2

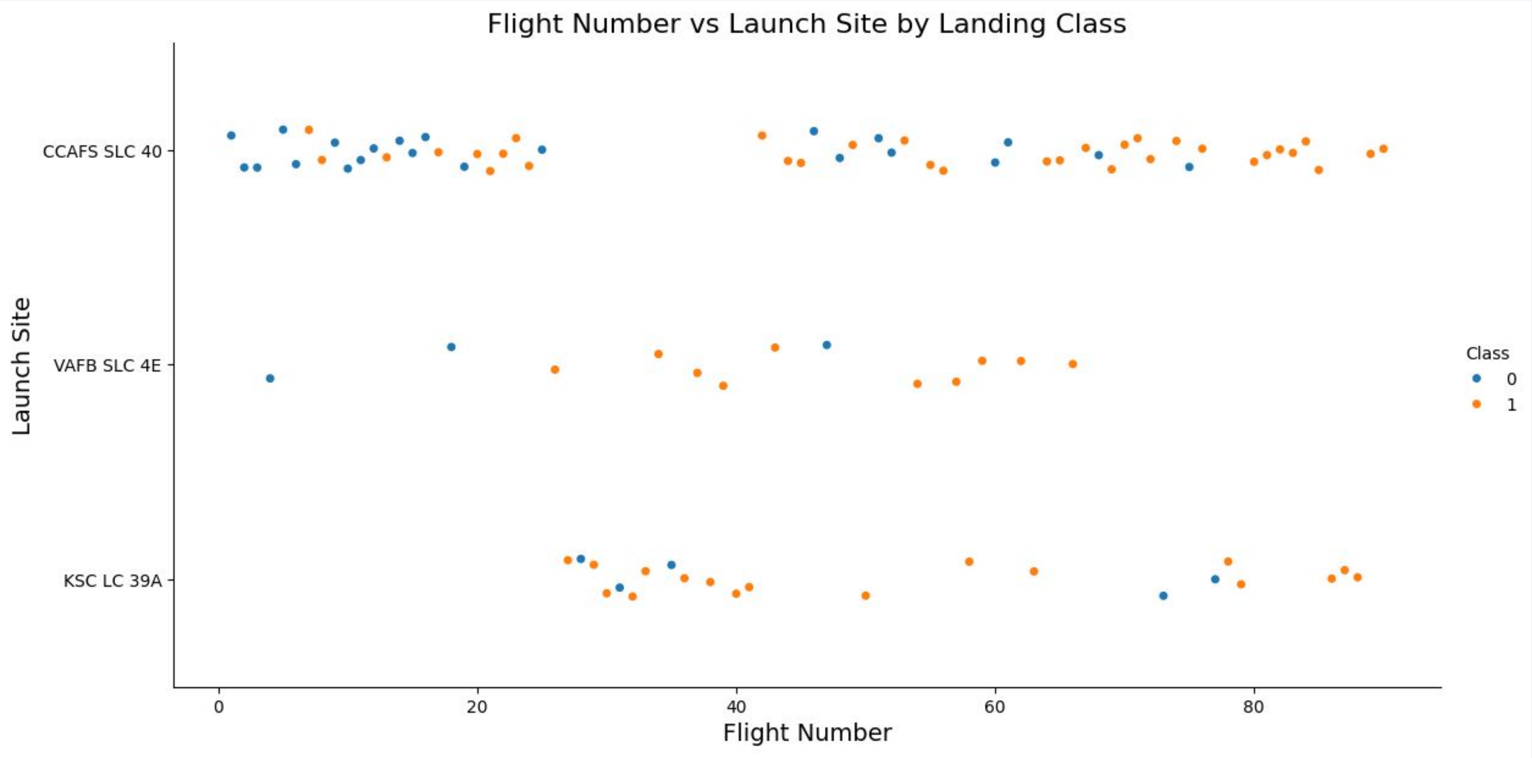
Insights drawn from EDA

Flight Number vs. Launch Site

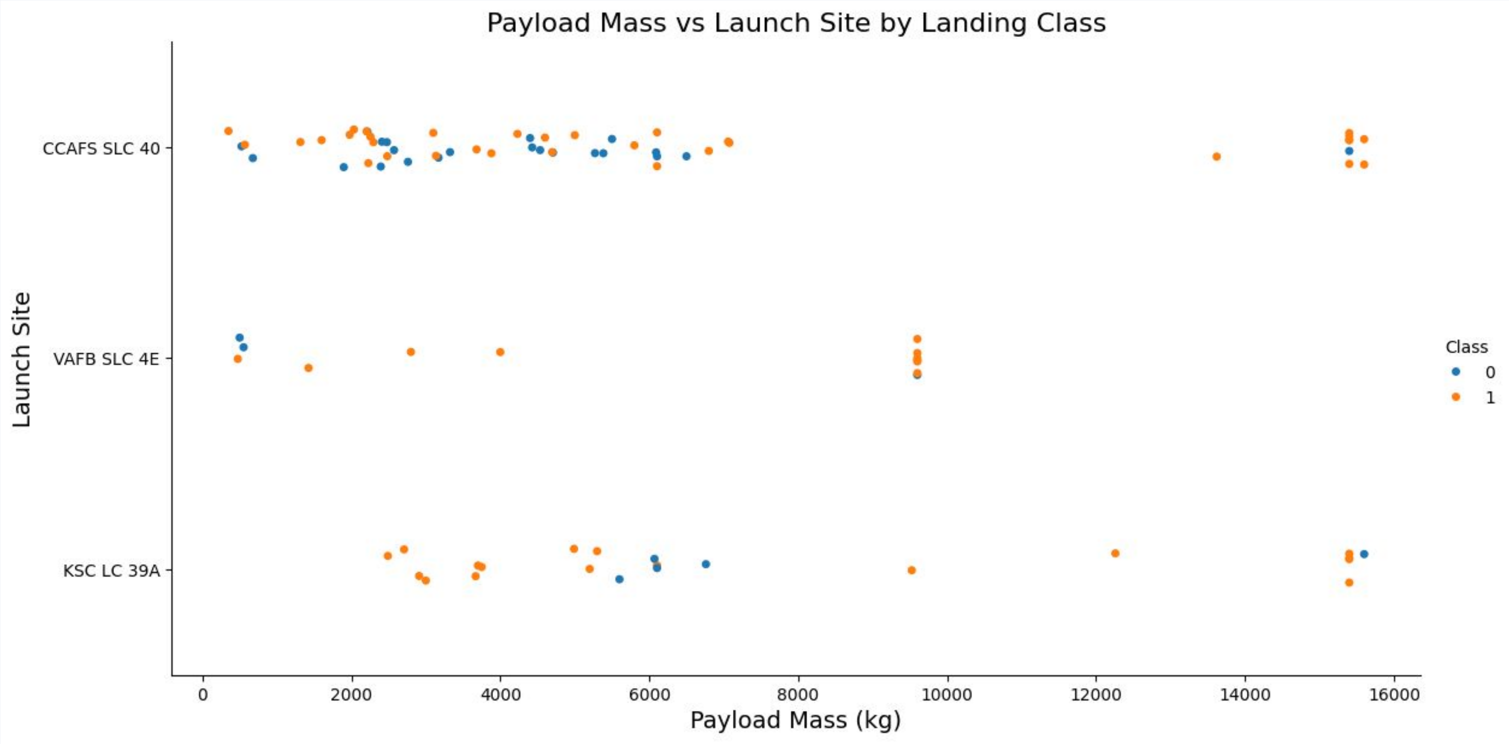
This scatter plot illustrates how different launch sites performed over time (by flight number).

- **Each dot** represents a specific launch.
- **Color-coded** by outcome:
  - 🟠 1: Successful landing
  - 🔵 0: Failure

🛰️ Notably, **success rates improved** as flight numbers increased — especially for **CCAFS LC-40** and **KSC LC-39A** — indicating that launch outcomes became more reliable with operational experience.



Payload vs. Launch Site



This scatter plot illustrates how payload mass varied across different launch sites, with landing success encoded by color.

- **X-axis**: Payload Mass (kg)
- **Y-axis**: Launch Site
- **Color**:
  - 🟠 `1`: Successful landing
  - 🔵 `0`: Failure

📊 **Observations**:

- **CCAFS LC-40** and **KSC LC-39A** handled a broader range of payload masses, with most launches trending toward successful landings.
- **VAFB SLC-4E** had fewer launches, but maintained a relatively balanced distribution of success and failure across payloads.
- There's **no clear linear correlation** between payload mass and landing outcome, but some patterns suggest that **extremely heavy payloads** were not necessarily associated with more failures.

Success Rate vs. Orbit Type

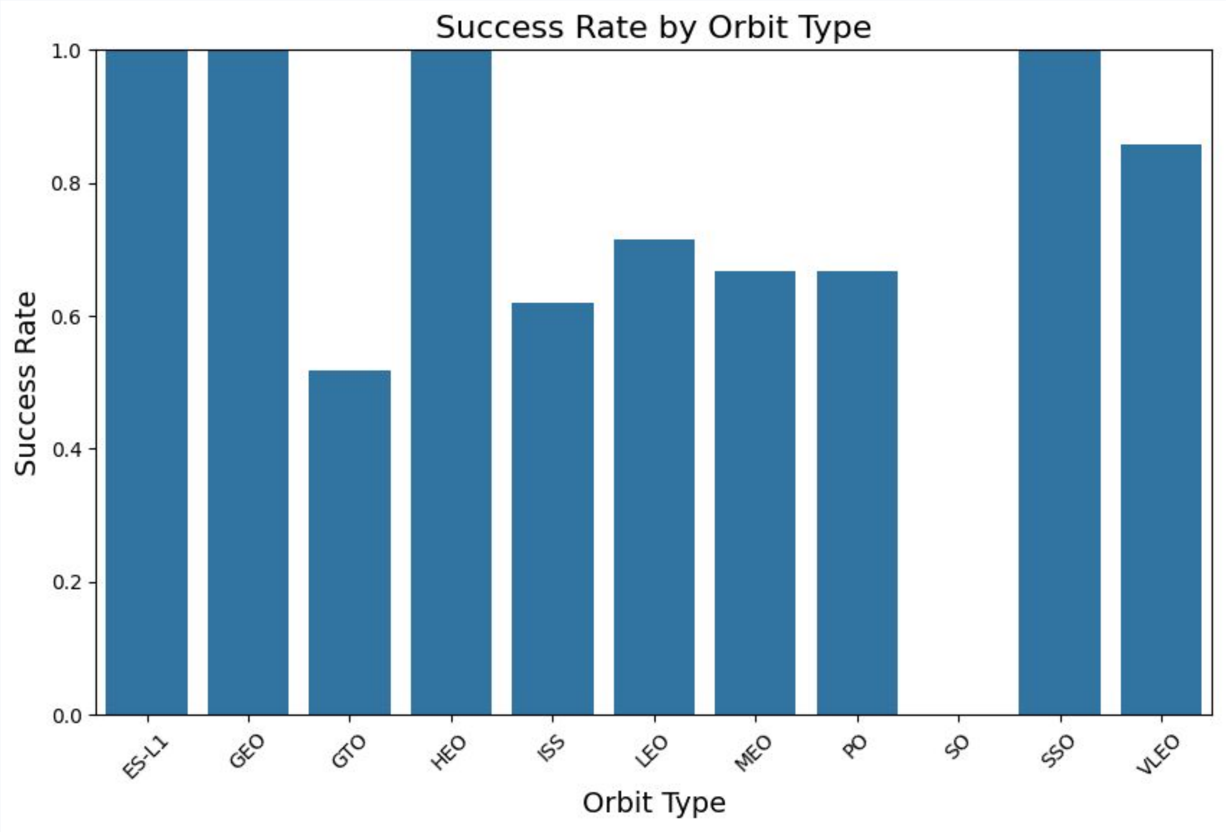
This bar chart compares the landing success rates across different orbit types.

- **X-axis**: Orbit type
- **Y-axis**: Average success rate (`Class` mean)

📊 **Observations**:

- Orbits such as **ES-L1**, **GEO**, **HEO**, and **SSO** achieved a **100% success rate**, indicating highly reliable mission performance in those categories.
- The **GTO** orbit showed a notably lower success rate (~52%), possibly due to its complexity or mission risk profile.
- **LEO** and **MEO** orbits had moderate success rates (~70%), suggesting room for optimization or variability in mission context.
- The variance in success rates across orbits may reflect **technical difficulty**, **altitude**, or **target application** of the mission.

🛰️ *This chart helps identify which orbits are historically more dependable and may inform launch strategy decisions.*



Flight Number vs. Orbit Type

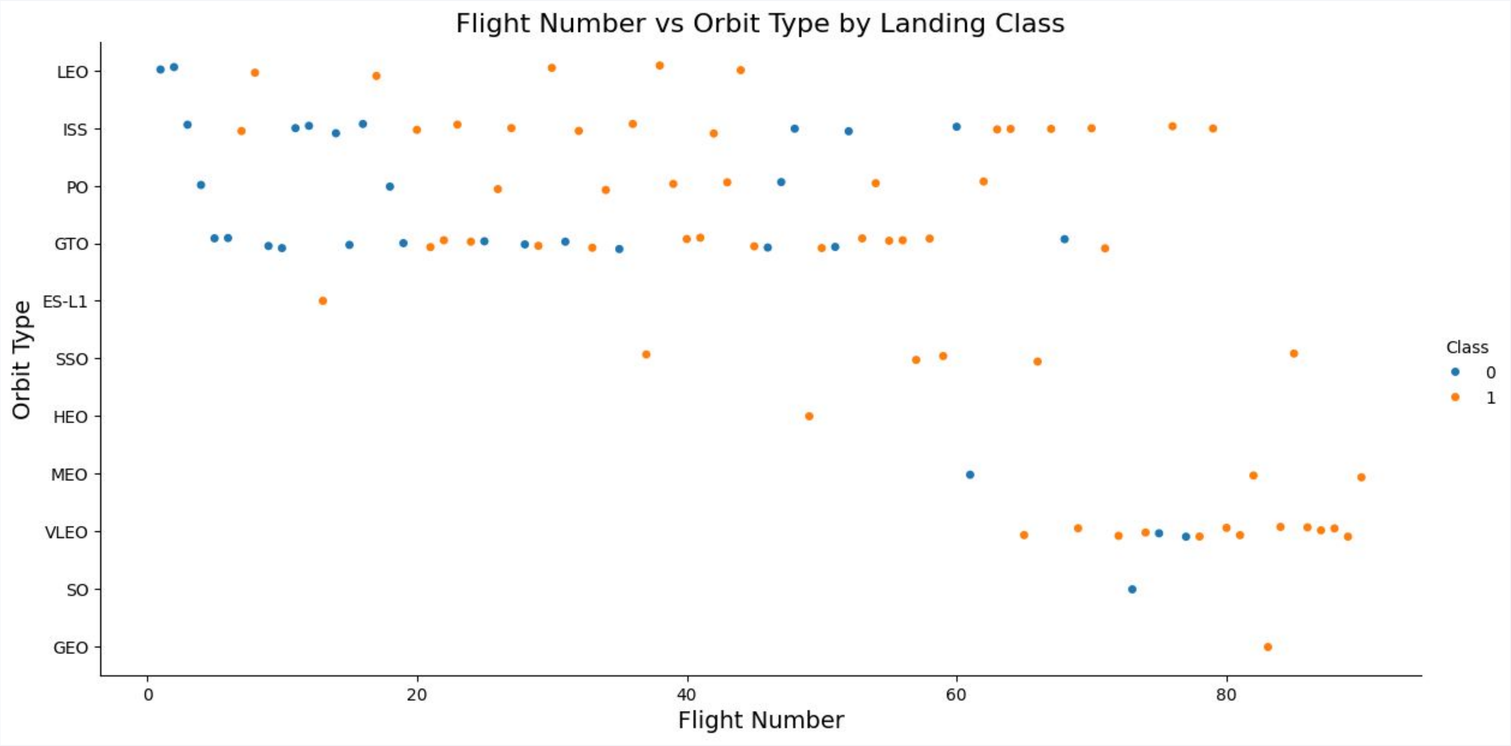
This scatter plot shows the relationship between the flight number (mission sequence) and the orbit type, color-coded by launch outcome.

- **X-axis**: Flight Number
- **Y-axis**: Orbit Type
- **Color**:
  - 🟠 `1`: Successful landing
  - 🔵 `0`: Failure

📊 **Observations**:

- **GTO**, **LEO**, and **ISS** orbits appear across a wide range of flight numbers, indicating frequent mission types.
- In **later missions**, there is a visible shift toward more consistent success (more orange) in orbits like **GTO** and **SSO**.
- Less common orbits like **HEO**, **VLEO**, and **SO** are sparsely distributed, but often associated with higher success rates.
- The spread suggests that as mission complexity and diversity increased, so did the reliability of launches across different orbital targets.

🚀 *This visualization helps us understand how launch outcomes evolved across different mission profiles over time.*



Payload vs. Orbit Type

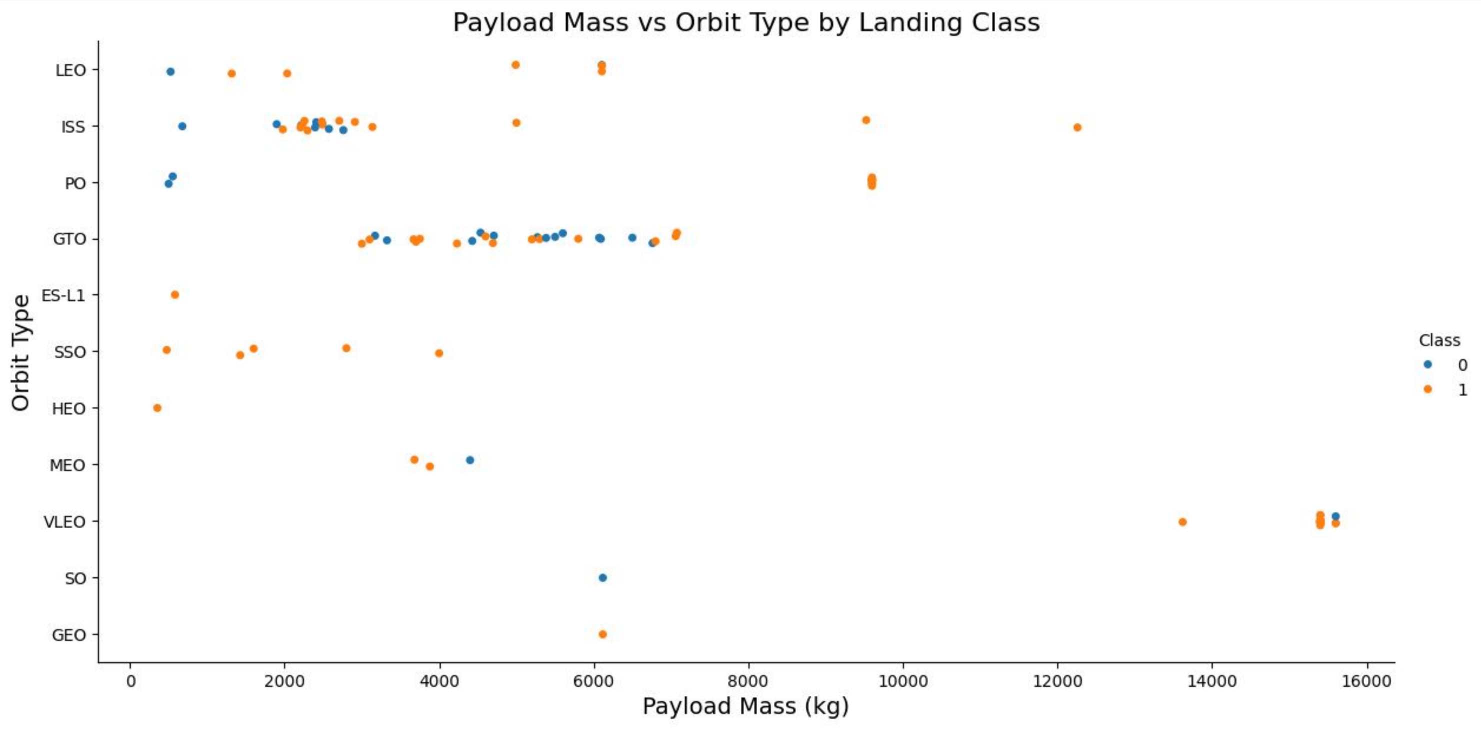
This scatter plot shows how payload mass varies by orbit type and how it relates to launch success, color-coded by outcome.

- **X-axis**: Payload Mass (kg)
- **Y-axis**: Orbit Type
- **Color**:
  - 🟠 `1`: Successful landing
  - 🔵 `0`: Failure

📊 **Observations**:

- **GTO**, **ISS**, and **LEO** were used across a wide payload range, showing a mix of successes and failures.
- Orbits like **SSO**, **VLEO**, and **HEO** were mostly associated with lighter payloads — and **more consistent success**.
- Some outliers with payloads above **10,000 kg** also resulted in successful landings, disproving a simplistic assumption that "heavier always fails".
- The chart suggests that **mission design and target orbit** may play a bigger role in outcome than payload mass alone.

📦 *This plot helps us understand how orbit type interacts with payload demands in mission planning and launch success.*



Launch Success Yearly Trend

This line chart shows the evolution of the average success rate of launches over the years.

- **X-axis**: Year (2010–2020)
- **Y-axis**: Average success rate (`Class` mean)

📈 **Observations**:

- From **2010 to 2013**, the success rate was consistently at **0**, reflecting early-stage technical challenges or limited mission scope.
- Starting in **2014**, success rates began to rise, reaching **over 80% by 2017**.
- After a small drop in **2018**, the success rate peaked again in **2019**, close to **90%**.
- The general trend suggests **substantial improvement over time**, likely due to increased experience, refined technology, and process optimization.

🚀 *This upward trajectory is a strong indicator of SpaceX's learning curve and technical maturity over the decade.*

All Launch Site Names

Unique Launch Sites

Using SQL, we queried all distinct launch site names from the dataset:

```
SELECT DISTINCT LaunchSite FROM SPACEXTBL;
```

📍 **Result:**

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

🛰️ *These four locations represent all the unique SpaceX launch sites in the dataset, each with varying mission profiles and success rates.*

Launch Site Names Begin with 'CCA'

Launch Sites Starting with 'CCA'

To retrieve the first 5 records where launch site names begin with `"CCA"`, we used the SQL query:

```
SELECT * FROM SPACEXTBL
WHERE LaunchSite LIKE 'CCA%'
LIMIT 5;
```

📍 **Result**:
All five returned entries are missions launched from **Cape Canaveral Air Force Station (CCAFS)**, including sites **LC-40** and **SLC-40**.

🛰️ *Cape Canaveral is one of the most frequently used locations in the dataset, appearing consistently in early and recent launches.*

Total Payload Mass

To calculate the total payload carried by boosters for missions commissioned by **NASA**, the following SQL query was used:

```
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload
FROM SPACEXTBL
WHERE Customer LIKE '%NASA%';
```

🚀 **Result**:
48,213 kg of payload were transported by SpaceX boosters for NASA missions.

📦 *This number represents the total mass of all payloads delivered in NASA-associated launches within the dataset, highlighting the agency's significant use of SpaceX's services.*

Average Payload Mass by F9 v1.1

To calculate the average payload mass delivered by the booster version **F9 v1.1**, the following SQL query was used:

```
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload

FROM SPACEXTABLE

WHERE Booster_Version = 'F9 v1.1';
```

📏 **Result**:
The average payload carried by **F9 v1.1** boosters is **2,928.4 kg**.

📦 *This indicates the typical mass range handled by this early generation of Falcon 9, useful for comparing performance across versions.*

First Successful Ground Landing Date

To identify the date of the first successful Falcon 9 landing on a ground pad, the following SQL query was used:

```
SELECT MIN(Date) AS First_Successful_Landing

FROM SPACEXTABLE

WHERE Landing_Outcome = 'Success (ground pad)';
```

📆 **Result**:
The **first successful ground landing** occurred on **December 22, 2015**.

🛰️ *This mission marked a major milestone in rocket reusability, enabling SpaceX to recover and reuse boosters from land-based landing zones.*

Successful Drone Ship Landing with Payload between 4000 and 6000

To retrieve the names of boosters that successfully landed on a drone ship and carried a payload mass between **4000 kg and 6000 kg**, we used:

```
SELECT Booster_Version

FROM SPACEXTABLE

WHERE Landing_Outcome = 'Success (drone ship)'

AND PAYLOAD_MASS__KG_ > 4000

AND PAYLOAD_MASS__KG_ < 6000;
```

📋 **Result**:

- F9 FT B1022
- F9 FT B1029
- F9 B4 B1040
- F9 B5 B1047

🛰️ *These boosters demonstrate reliable reusability in medium-payload missions requiring drone ship landings.*

Total Number of Successful and Failure Mission Outcomes

To calculate the total number of **successful** and **failure** mission outcomes, the following SQL query was executed:

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

📊 **Results**:

Mission Outcome	Count
✅ Success	99
🚧 Failure (in flight)	1
⚠️ Success (payload status unclear)	1

🛰️ **Interpretation**:

- **99 out of 101** missions were officially considered **successful** — showcasing SpaceX's high reliability.
- Only **one mission failed in flight**, and one success was marked with **unclear payload status**.

📈 *This result confirms SpaceX's strong track record in achieving mission objectives across a wide range of launch contexts.*

Boosters Carried Maximum Payload

📌 **Explanation:**

This query identifies all boosters that carried the maximum payload mass recorded in the dataset — **15,600 kg**.

The results show that multiple **Falcon 9 Block 5** boosters (e.g., *B1048.4*, *B1049.4*, *B1051.3*, etc.) reached this payload capacity, highlighting their consistent use in heavy-lift missions.

These boosters represent SpaceX's most powerful and reliable configurations for high-mass orbital deployments.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

📌 **Explanation:**

This query retrieved all failed landings on **drone ships** that occurred in the year **2015**.
Filtering by the `Landing_Outcome` and the year (`2015`), the result shows two failed attempts involving:

- Booster **F9 v1.1 B1012** in **January** at **CCAFS LC-40**
- Booster **F9 v1.1 B1015** in **April** at **CCAFS LC-40**

These failures mark early stages in the development of drone ship recovery technology.

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

📌 **Explanation:**

This query ranks all **landing outcomes** between **2010-06-04** and **2017-03-20**, sorted in **descending order** by frequency.

- The most common outcome was **"No attempt"** with **10 records**, reflecting early-stage missions where landing was not attempted.
- Both **"Success (drone ship)"** and **"Failure (drone ship)"** occurred **5 times** each, showing SpaceX's experimental phases.
- **"Success (ground pad)"** was achieved **3 times**, indicating early ground-based recoveries.
- Other less frequent outcomes include **ocean landings** and **parachute failures**, which represent contingency scenarios or experimental tests.

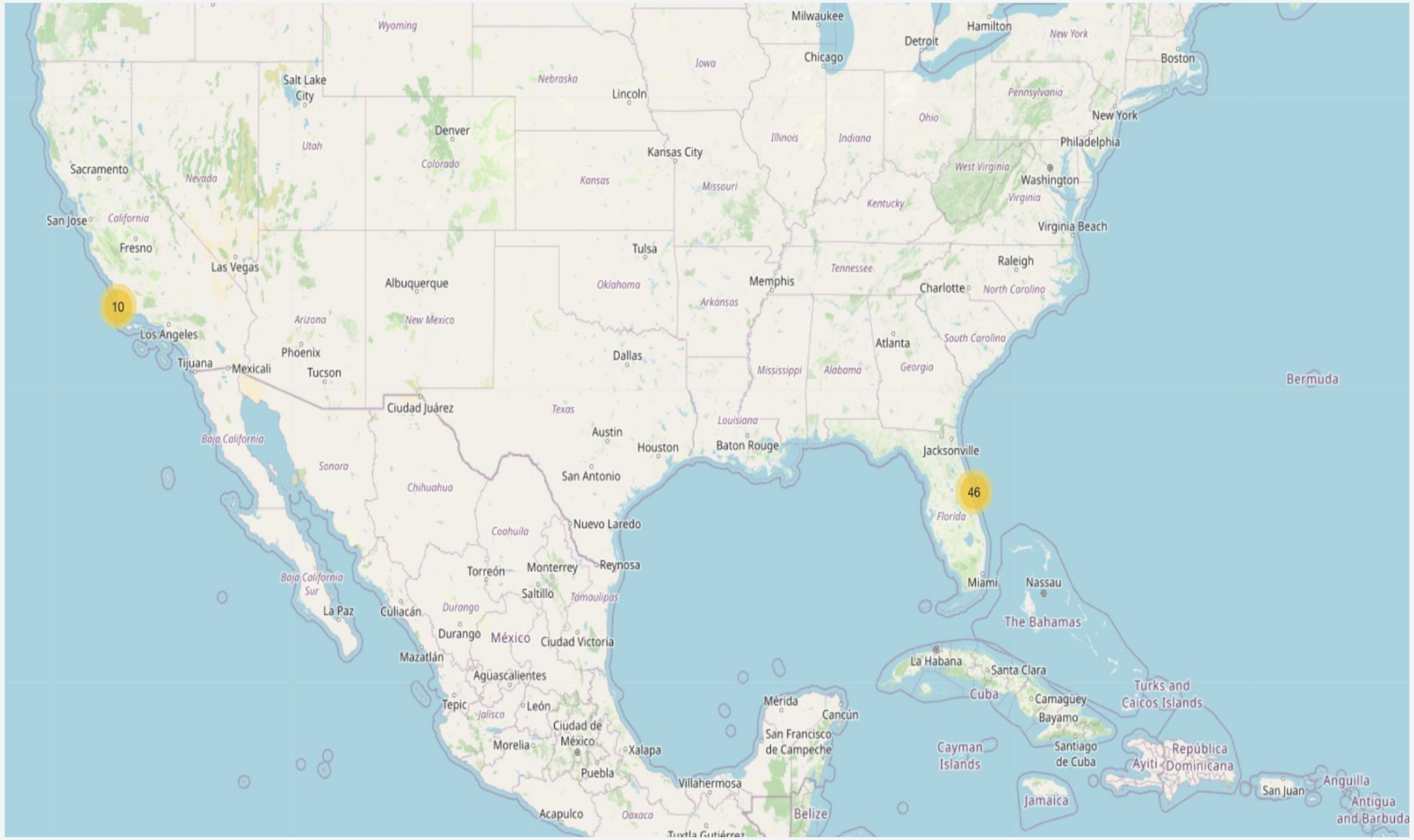
🧠 This breakdown reveals how SpaceX gradually transitioned from no recovery to consistent drone ship and ground pad landings.

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

Section 3

Launch Sites Proximities Analysis

Launch Site Locations on World Map



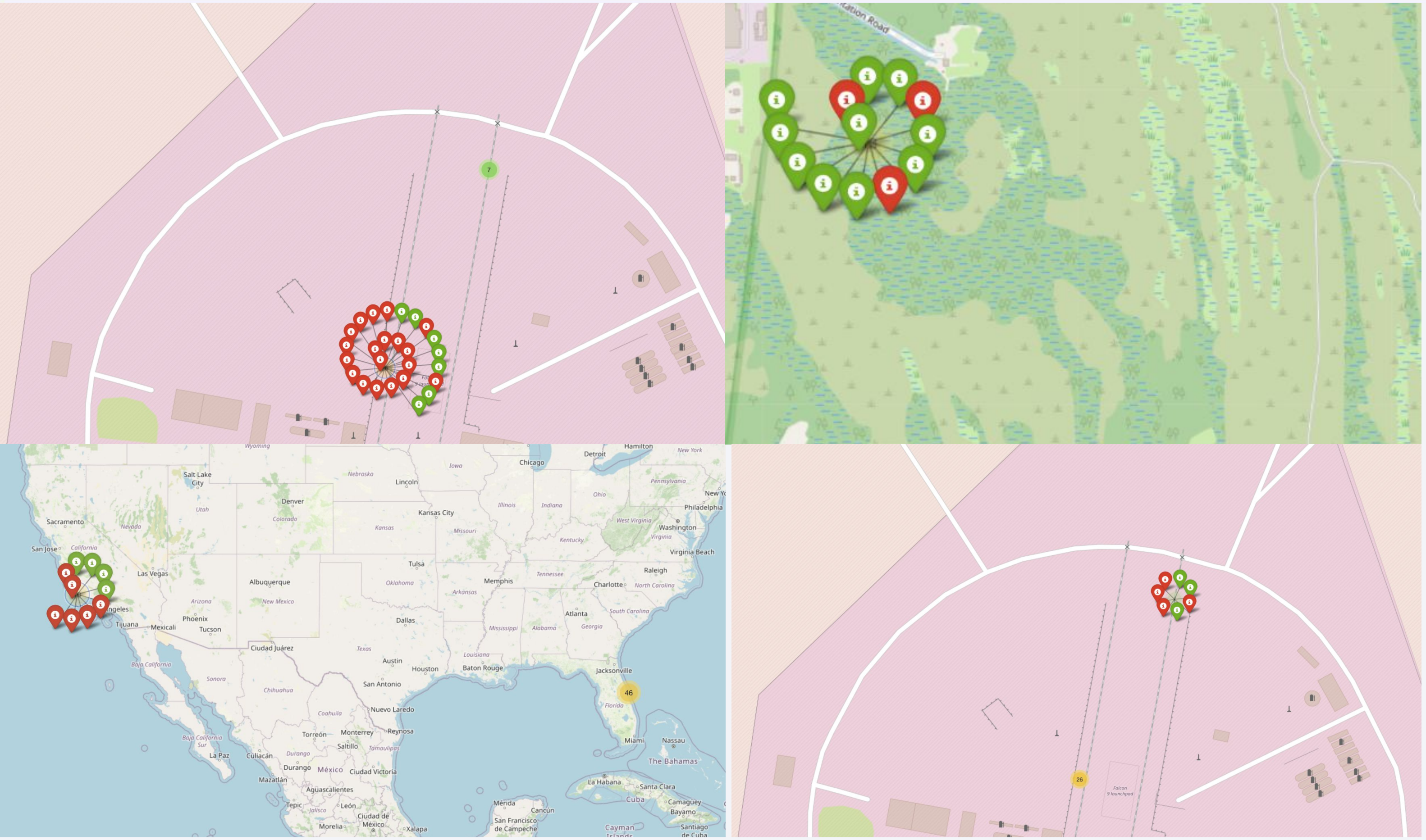
📌 **Explanation**:

This map displays the **geographic location of all SpaceX launch sites** used in the dataset. Each marker on the map corresponds to a unique launch facility.

- The markers are placed using **latitude and longitude coordinates** extracted from the launch dataset.
- The primary sites shown include:
  - 🟢 **CCAFS LC-40** (Cape Canaveral, Florida)
  - 🟢 **KSC LC-39A** (Kennedy Space Center, Florida)
  - 🟢 **VAFB SLC-4E** (Vandenberg Air Force Base, California)

🛰️ These sites reflect **SpaceX's strategic positioning** along both the East and West coasts of the U.S., enabling efficient launches into various orbits (polar, geostationary, and LEO).

Launch Outcomes by Location



📌 **Explanation:**

This map visualizes all individual launch outcomes, plotted by location and colored by result:

- 🟢 **Green markers** represent successful landings.
- 🔴 **Red markers** indicate failed landings.

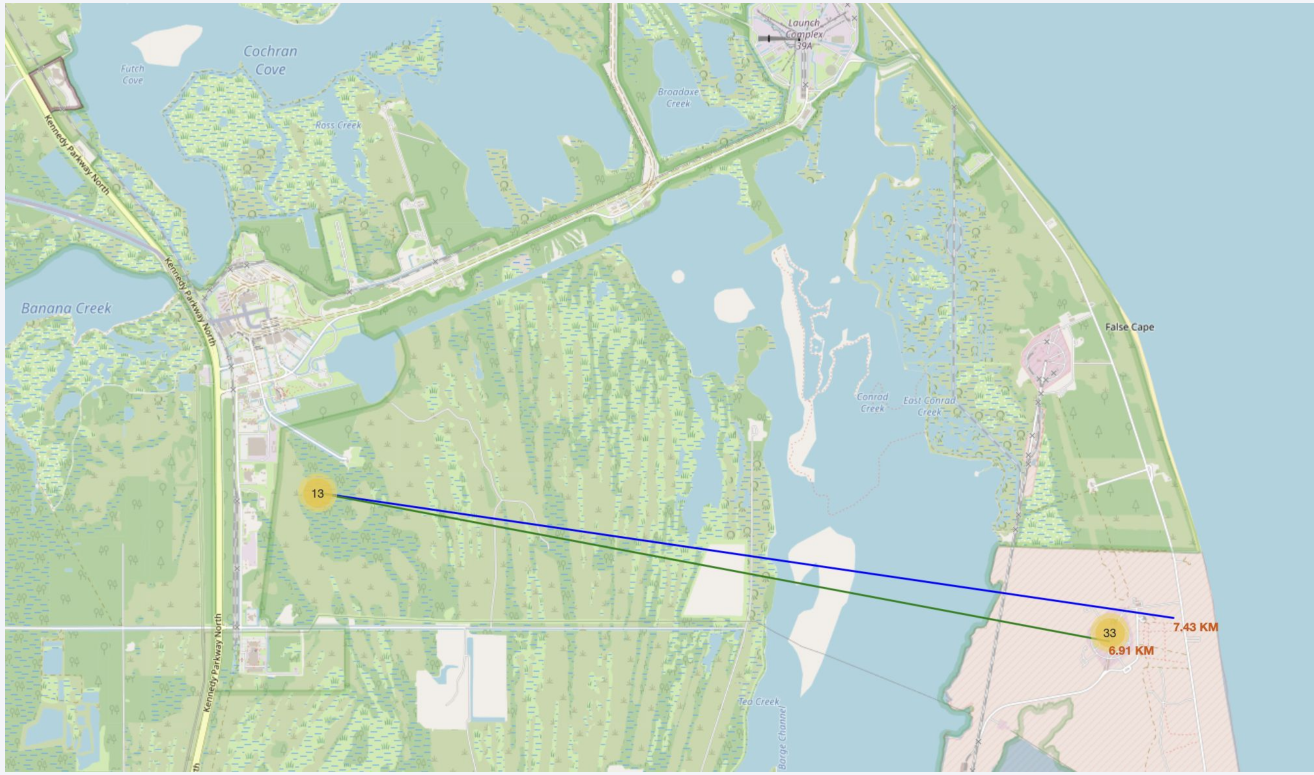
🔍 **Observations:**

- Most successful landings (green) are clustered at **KSC LC-39A** and **CCAFS LC-40**, showing consistent operational reliability at Florida's east coast.
- Failures (red) appear more concentrated at early test phases, especially at **VAFB SLC-4E** on the west coast.
- The **density of markers** also highlights frequently used sites and evolving reliability over time.

🧭 **Conclusion:**

This visual breakdown reinforces how SpaceX improved landing precision over time, with successes eventually outnumbering failures in key locations.

Launch Site Proximity Analysis – LC-39A



📌 **Explanation**:

This map displays the proximity analysis of the **Launch Complex 39A (LC-39A)**, showing the distances to nearby infrastructure using a Folium-generated map.

- **Blue line**: Distance from the launch site to the nearest **railway** – approximately **7.43 KM**
- **Green line**: Distance to the nearest **highway access** – approximately **6.91 KM**

🧭 **Observations**:

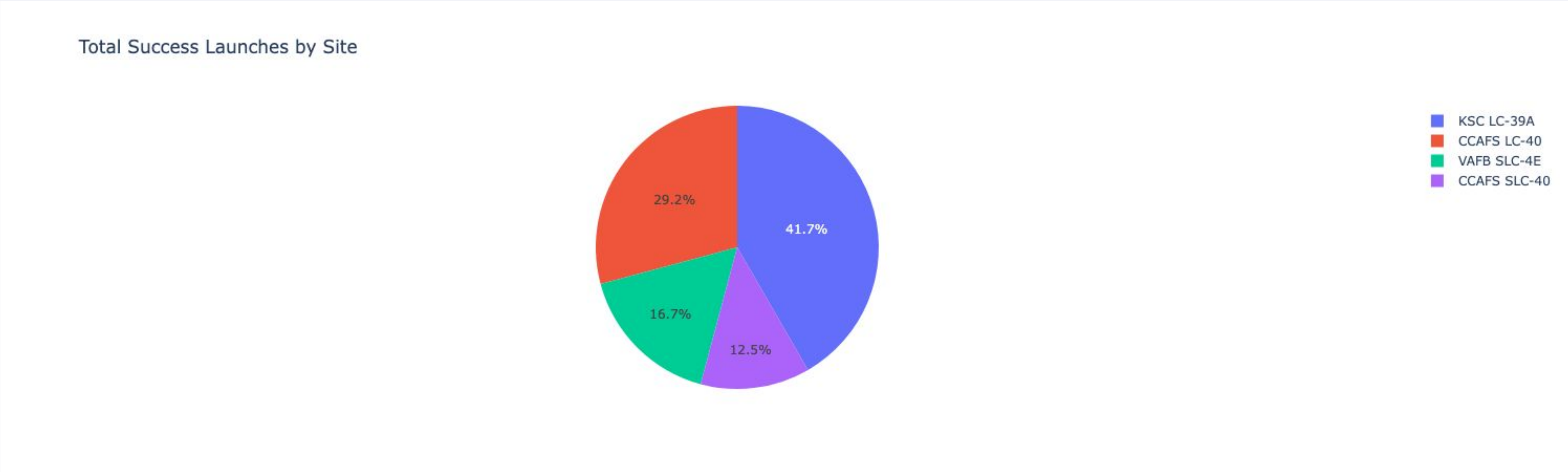
- LC-39A is **strategically positioned** close to transportation infrastructure, which is essential for **logistics**, **equipment transport**, and **personnel access**.
- The **short distance to rail and highway** underscores the site's accessibility and logistical viability for frequent and heavy launch operations.
- Such infrastructure proximity reduces mission prep time and increases the efficiency of SpaceX's ground operations.

🔚 **Conclusion**:
Proximity to key infrastructure enhances the operational readiness of launch sites. The measured distances confirm that **LC-39A** is ideally located for both launch logistics and mission support.

Section 4

Build a Dashboard with Plotly Dash

Launch Success Distribution by Site



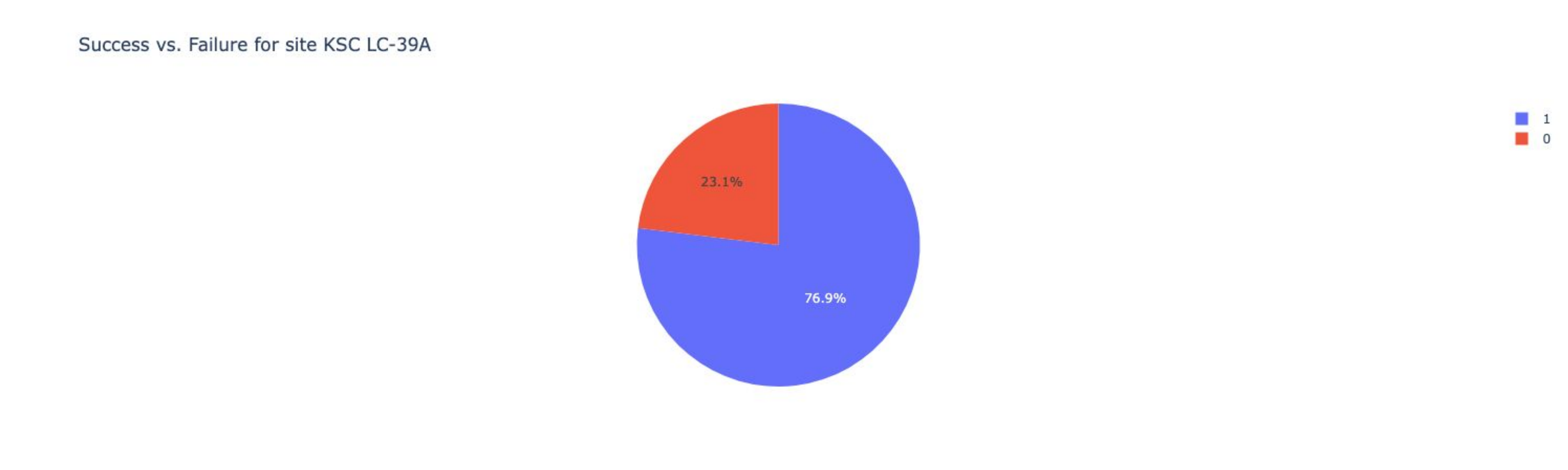
📌 **Explanation**:
This pie chart illustrates the proportion of successful launches across all major SpaceX launch sites included in the dataset.

- **KSC LC-39A** accounted for **41.7%** of all successful missions, highlighting its primary role in major deployments.
- **CCAFS LC-40** follows with **29.2%**, also reflecting strong operational reliability.
- **VAFB SLC-4E** and **CCAFS SLC-40** had comparatively fewer successful launches, representing **16.7%** and **12.5%**, respectively.

🔎 **Interpretation**:
The data shows a clear emphasis on launches from **Florida-based sites**, with KSC LC-39A leading in mission success. This reflects strategic utilization of infrastructure and favorable conditions at the Kennedy Space Center region.

📊 **Conclusion**:
Most successful SpaceX launches are concentrated in a few key locations, which may influence logistical decisions and future infrastructure investment.

Launch Success Ratio – KSC LC-39A



This pie chart shows the **launch success ratio** for the site with the **highest performance**: **KSC LC-39A**.

- **Blue segment (76.9%)**: Successful missions.
- **Red segment (23.1%)**: Unsuccessful attempts.

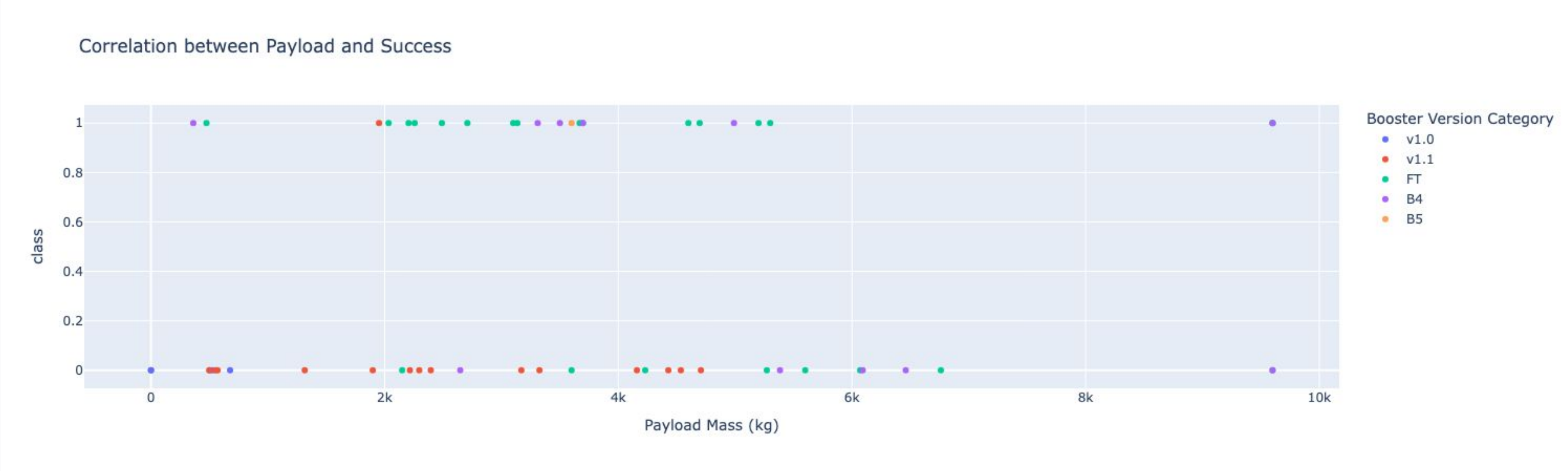
🔍 **Observations:**

- The **high success ratio** reflects strong operational maturity and consistent performance.
- The **low failure rate** highlights SpaceX’s ability to optimize and reuse infrastructure effectively at this location.
- This reinforces LC-39A's **strategic value** in SpaceX's launch portfolio.

📌 **Conclusion:**

KSC LC-39A stands out as the most reliable launch site in the dataset, making it a key asset for mission-critical launches.

Correlation between Payload and Launch Success by Booster Version



This scatter plot analyzes the relationship between payload mass and launch outcomes (success or failure), categorized by booster version.

🔎 **Observations:**

- Most **successful launches** (Class = 1) are clustered in the **2000–6000 kg** payload range.
- **Booster versions FT and B5** show a high concentration of successful outcomes, suggesting strong reliability at higher payloads.
- Older versions like **v1.0** and **v1.1** are associated with more frequent failures, especially at lower payload masses.

📊 **Conclusion:**

The analysis highlights that **more advanced booster versions (FT and B5)** are better at handling heavier payloads with consistent success, reinforcing the evolutionary improvement of SpaceX launch technology.

Section 5

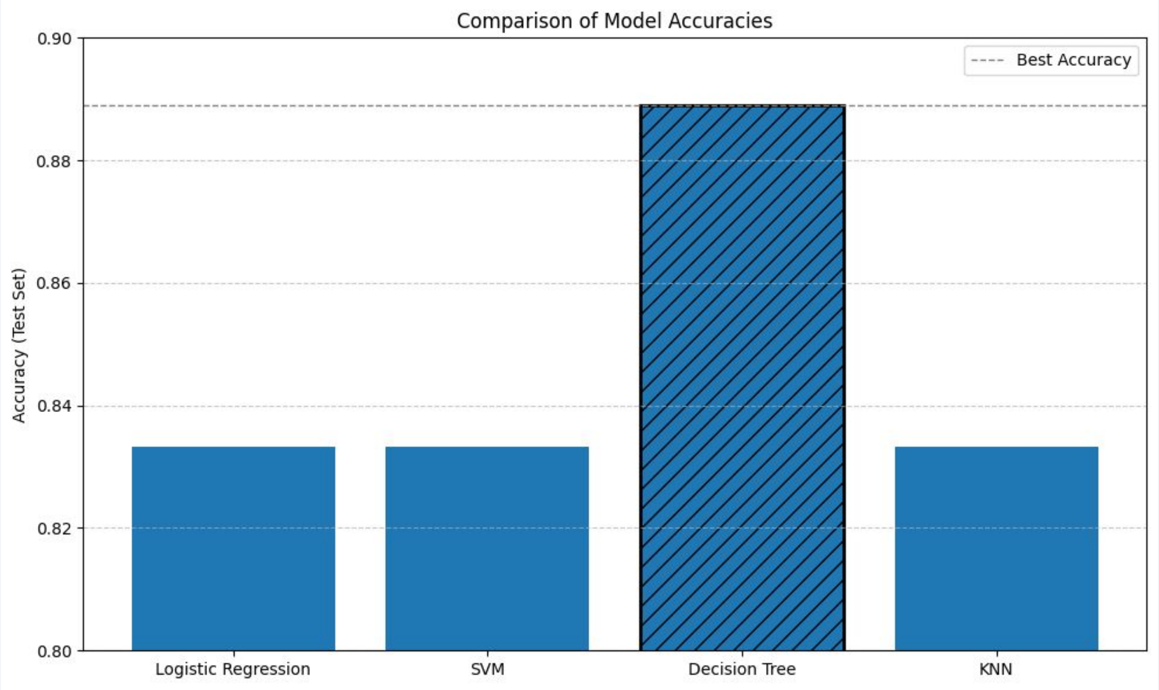
Predictive Analysis (Classification)

Classification Accuracy

🔎 **Observations:**

- *Decision Tree* achieved the highest classification accuracy (~89%), outperforming all other models.
- *Logistic Regression*, *SVM*, and *KNN* yielded similar results (~83%), showing comparable but lower predictive power.
- The clear margin indicates that tree-based models may better capture non-linear relationships in this dataset.

📊 **Conclusion:**
Decision Tree stands out as the most effective model in this scenario, offering the highest predictive accuracy among all classification approaches tested.



Confusion Matrix

📊 Explanation:

This confusion matrix visualizes the prediction performance of the **best performing model** — the **Decision Tree**, based on the test dataset.

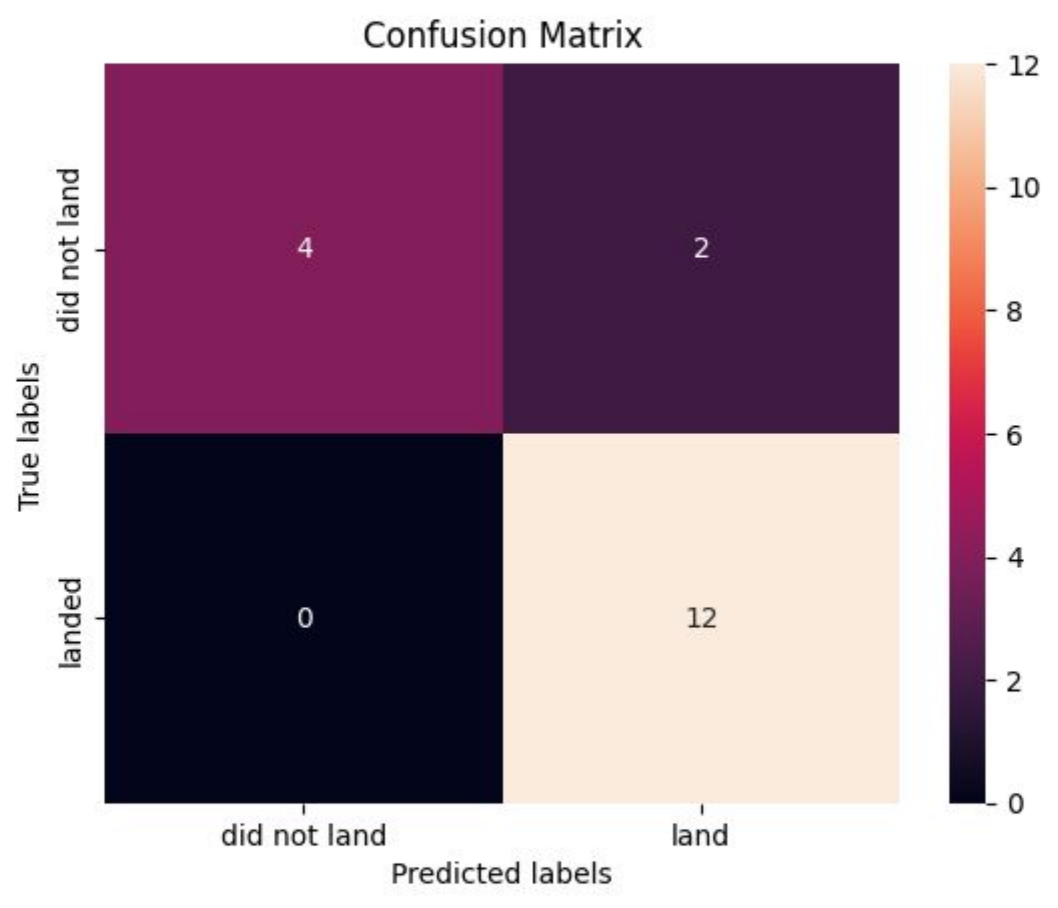
- **True Positives (12)**: The model correctly predicted "landed" launches.
- **True Negatives (4)**: The model correctly predicted "did not land".
- **False Positives (2)**: The model predicted "landed" but the actual outcome was "did not land".
- **False Negatives (0)**: No successful landings were missed by the model.

🔎 Observations:

- The **absence of false negatives** indicates the model reliably predicts successful landings.
- The **high count of true positives** reinforces the model's strength in identifying actual land outcomes.
- Minimal false positives (2) suggest **limited overestimation**, which is acceptable in safety-critical systems like rocket landings.

✅ Conclusion:

The Decision Tree classifier demonstrates **high predictive reliability**, particularly in identifying successful landings. The structure of the confusion matrix confirms this model's suitability for mission-critical deployment scenarios.



Conclusion

This project successfully applied data science techniques to analyze SpaceX launch data, uncovering meaningful insights into launch success patterns, payload impact, and booster performance. Through data wrangling, SQL queries, visualizations, and predictive modeling, we identified **key launch sites**, **successful booster versions**, and the **most reliable classification model** (Decision Tree). The findings demonstrate how data-driven approaches can support aerospace decision-making, improve operational strategies, and enhance mission reliability.

Appendix

🧾 **Note**:
All data analysis, SQL queries, Python code, and interactive visualizations used in this project are available in the accompanying GitHub repository.

🔗 **GitHub Repository**:
github.com/lz-castro/spacex-data-science-capstone

📂 Contents:

- Jupyter Notebooks (data wrangling, SQL, Folium, dashboarding)
- Code for visualization and classification models
- Output charts and screenshots used in the presentation

🧠 This repository provides full reproducibility for all insights and results presented.

Thank you!