IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Bensalah Mohamed
21/06/2026

Outline

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

Executive Summary

Summary of Methodologies:

Data collection was performed using the **SpaceX REST API** and **web scraping** from public sources.
Data wrangling was applied to clean the dataset, handle missing values, and prepare features.
Exploratory Data Analysis (EDA) was conducted using **data visualization** and **SQL queries**.
Interactive visual analytics were built using **Folium maps** and a **Plotly Dash dashboard**.
Classification models were developed and optimized using **GridSearchCV**

Executive Summary

Summary of Results:

Launch success was found to vary significantly across **launch sites** and **orbit types**.
Payload mass and **flight number** showed a clear relationship with landing success.
Some launch sites achieved consistently higher success rates.
Interactive maps revealed geographic patterns influencing landing outcomes.
Classification models demonstrated strong capability in predicting landing success

Introduction

SpaceX aims to reduce launch costs by reusing the Falcon 9 first-stage booster.
A successful landing of the first stage is therefore a critical requirement for reusability.
This project analyzes historical SpaceX launch data to identify the key factors influencing landing success and to build predictive models capable of estimating landing outcomes.

Section 1

Methodology

Methodology

Executive Summary

- **Collected launch data** from the SpaceX REST API and public web sources
- **Performed data wrangling** to clean the dataset, handle missing values, and prepare features
- **Conducted exploratory data analysis (EDA)** using data visualization and SQL queries
- **Built interactive visual analytics** with Folium maps and a Plotly Dash dashboard
- **Developed predictive models** to classify first-stage landing success
- **Optimized model performance** using hyperparameter tuning with GridSearchCV

Data Collection

Sources:

- **SpaceX API:** Used to retrieve primary launch, rocket, and payload data.
- **Wikipedia:** Scraped supplemental historical tables for landing outcome details.

Methodology:

- **Requesting:** Used `requests` library to fetch JSON data from the API.
- **Scraping:** Utilized `BeautifulSoup` to parse HTML tables from web pages.
- **Unification:** Combined multiple data sources into a single Pandas DataFrame

Data Collection – SpaceX API

- **Process:** Performed GET requests to the SpaceX V4 API using the Python `requests` library.
- **Data Handling:** Flattened nested `JSON` responses into a structured table using `pd.json_normalize`.
- **Technical Flow:** `API Endpoint` → `JSON Response` → `Feature Extraction` → `Pandas DataFrame`
- **Data Focus:** Retained core mission data, including Flight Number, Date, Payload, and Landing Outcome.

GitHub URL: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Collecting%20the%20Data%20API%20lab1.ipynb

- **API Endpoint** (`/launches/past`)
- **HTTP GET Request** (using `requests` library)
- **JSON Response** (Raw data)
- **Data Normalization** (`pd.json_normalize`)
- **Clean DataFrame** (Filtered for Falcon 9)

Data Collection - Scraping

- **Tools:** Utilized **BeautifulSoup** and the **requests** library to extract data from HTML tables.
- **Process:** Parsed the "List of Falcon 9 and Falcon Heavy launches" Wikipedia page.
- **Key Phrases:** `BeautifulSoup object`, `find_all('tr')`, `HTML Table parsing`, `Data Extraction`.
- **Data Cleaning:** Stripped technical references (e.g., "[1]") and white spaces to ensure data integrity.

the GitHub URL:

https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Data%20Collection%20webscraping%20lab2.ipynb

- **Input:** Wikipedia URL
- **Action:** `requests.get()` to fetch HTML content.
- **Processing:** Initialize `BeautifulSoup` object with `html.parser`.
- **Extraction:** Iterate through `<tr>` rows to find launch details (Date, Booster, Outcome).
- **Output:** Append extracted data into a list of dictionaries → **Pandas DataFrame**.

Data Wrangling

How the data were processed

- The collected SpaceX launch data required several preprocessing steps to ensure data quality and consistency before analysis and modeling. The data wrangling process focused on cleaning, transforming, and structuring the dataset into a usable format for exploratory analysis and machine learning.
- **Key Data Wrangling Steps *(key phrases)***
- **Filtered launch records** to include only **Falcon 9** missions
- **Handled missing values**, particularly in the **PayloadMass** feature
- **Standardized data formats** and corrected data types
- **Removed irrelevant and redundant columns**
- **Extracted useful features** from raw API and scraped data
- **Created a binary target variable (Class)** based on landing **Outcome**
  - 1 → Successful landing
  - 0 → Failed landing
- **Prepared the final clean dataset** for EDA, SQL analysis, visualization, and machine learning

The GitHub URL: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling%20lab3.ipynb

Data Wrangling (Flowcharts):

Raw Data (API + Web Scraping)

↓

Data Cleaning & Formatting

↓

Handling Missing Values

↓

Feature Selection & Transformation

↓

Landing Outcome Labeling (Class)

↓

Clean Dataset for Analysis & Modeling

The GitHub URL: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Data%20Wrangling%20lab3.ipynb

EDA with Data Visualization

- **Charts Used and Purpose**
- Exploratory Data Analysis (EDA) used data visualizations to understand how mission characteristics relate to first-stage landing success.
- **Scatter plots** were used to analyze the relationship between **Flight Number** and landing success, highlighting performance improvement over time.
- **Scatter plots by category** examined how **Payload Mass** affects landing success across different **Launch Sites**.
- **Bar charts** compared landing success rates across **Orbit Types**.
- **Line charts** visualized the **yearly trend** of average landing success.
- These charts were selected to clearly show trends, patterns, and categorical differences in landing outcomes.

The GitHub URL: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Exploring%20and%20Preparing%20Data%20lab5.ipynb

EDA with SQL

- **Data Filtering:** Isolated **23 missions** with a Payload Mass between 4,000 and 6,000 kg to analyze medium-heavy lift performance.
- **Customer Aggregations:** Calculated the total payload for **NASA (CRS)** missions, totaling over **45,596 kg**.
- **Landing Analysis:** Used `DISTINCT` queries to identify all successful **Drone Ship** landings, specifically for the **F9 v1.1** booster version.
- **Temporal Insights:** Identified **2015-12-22** as the date of the first successful Ground Pad landing.
- **Site Ranking:** Determined that **CCAFS SLC-40** was the most active site with the highest frequency of launches.
- **Key Phrases:** `SELECT`, `GROUP BY`, `COUNT`, `WHERE`, `ORDER BY`, `LIMIT`.

Git url: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Explatory%20data%20analysis%20eda%20sql%20lab%204.ipynb

Build an Interactive Map with Folium

- **Marker Clusters:** Added Green (Success) and Red (Failure) pins.
- *Why:* To quickly visualize success rates and launch density at each site.
- **Circles:** Drew circular perimeters around launch pads.
- *Why:* To clearly demarcate the geographical boundaries of each facility.
- **PolyLines:** Plotted lines from sites to coasts, highways, and cities.
- *Why:* To measure proximity and verify safety protocols (e.g., distance from populated areas).
- **Mouse Position:** Added an interactive coordinate tracker.
- *Why:* To calculate precise distances between pads and landmarks using the Haversine formula.
- the GitHub URL : https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab6.ipynb

Build a Dashboard with Plotly Dash

Plots/Graphs and Interactions Added

- **Pie chart (Success vs Failure):** Displays the distribution of landing outcomes for **all sites** or for a **selected launch site**.
- **Scatter plot (Payload Mass vs Success):** Shows how landing outcomes vary across **payload mass**, with points grouped by **launch site** when applicable.
- **Dropdown (Launch Site):** Lets users filter the dashboard to a specific launch site or view **all sites**.
- **Range Slider (Payload Range):** Allows users to focus on a specific payload interval and observe how success changes within that range.

Why These Were Added

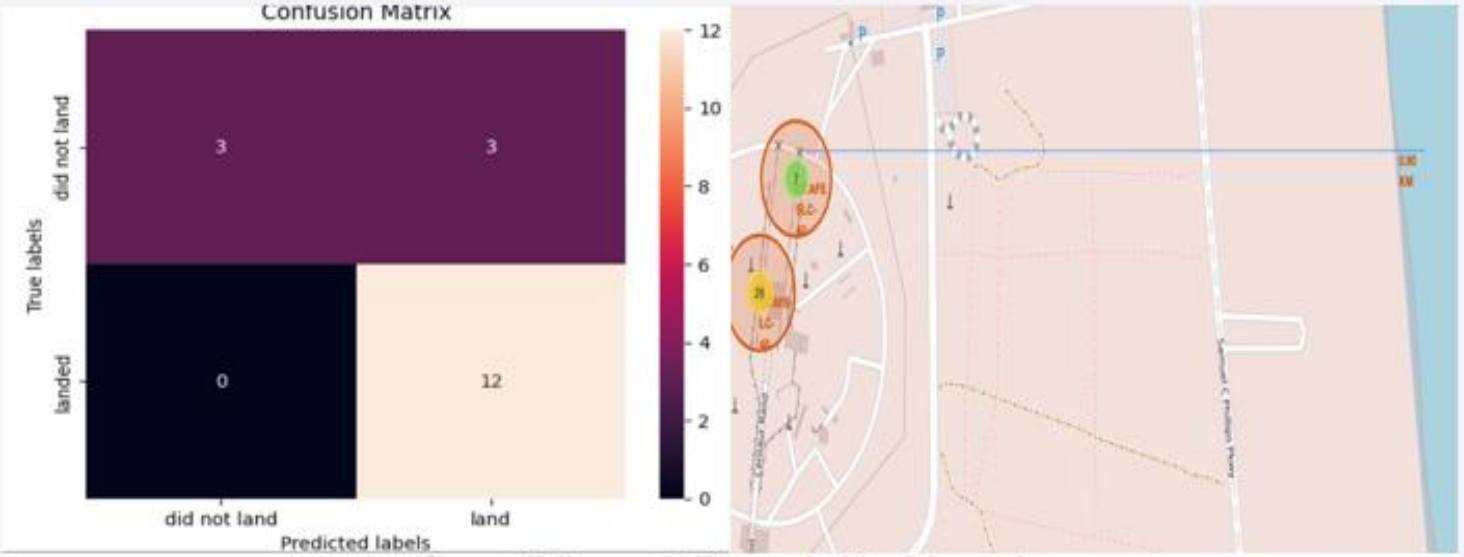
- The **pie chart** provides a quick, high-level view of success rate differences between sites.
- The **scatter plot** helps identify patterns between **payload mass** and landing success, which is a key variable in the project.
- The **dropdown** and **range slider** make the dashboard interactive, enabling users to test hypotheses (e.g., “Does success improve for lower payload ranges at a given site?”) without rewriting cod

Predictive Analysis (Classification)

- **Model Development and Evaluation Summary**
- The predictive analysis focused on building classification models to predict the **first-stage landing success** of Falcon 9 launches. The modeling process followed a structured approach including data preparation, model training, evaluation, and optimization.
- **Key Model Development Steps (Key Phrases)**
- **Feature selection**: Selected relevant features such as **Launch Site**, **Orbit**, **Payload Mass**, and **Flight Number**
- **Data scaling**: Applied **StandardScaler** to normalize numerical features
- **Train-test split**: Split the dataset into training and testing sets (**80/20 split**)
- **Model training**: Trained multiple classifiers including **Logistic Regression**, **Support Vector Machine (SVM)**, **Decision Tree**, and **K-Nearest Neighbors (KNN)**
- **Model evaluation**: Evaluated models using **accuracy scores** and **confusion matrices**
- **Hyperparameter tuning**: Used **GridSearchCV** with cross-validation to optimize model parameters
- **Model comparison**: Compared performance across models to identify the best-performing classifier
- **Best Performing Model**
- Among the evaluated models, the **Decision Tree classifier** achieved the highest test accuracy and provided good interpretability, making it the best-performing model for this project.

Git url: https://github.com/Mohamed-bns/IBM-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction%20lab7.ipynb

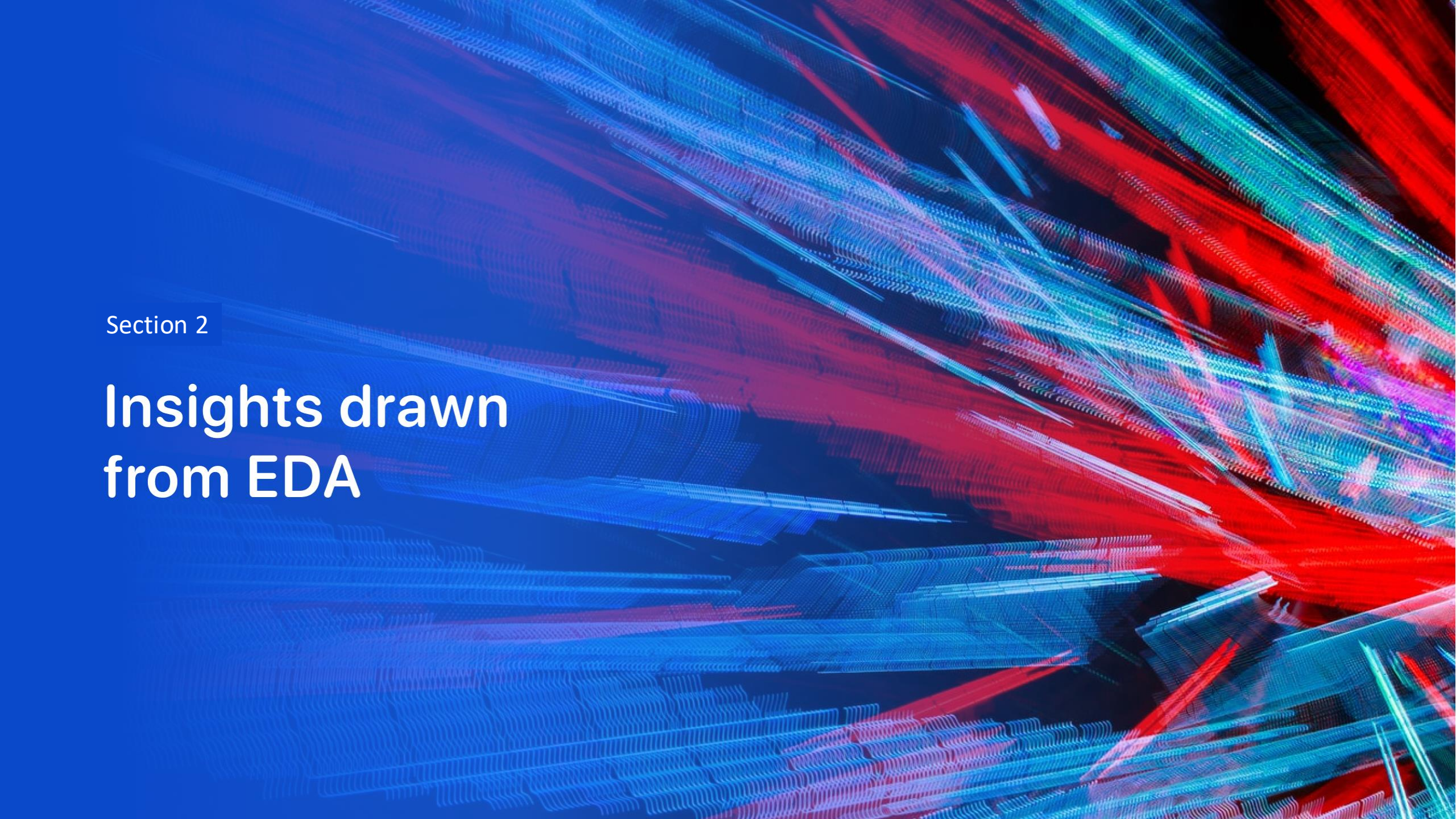
Results



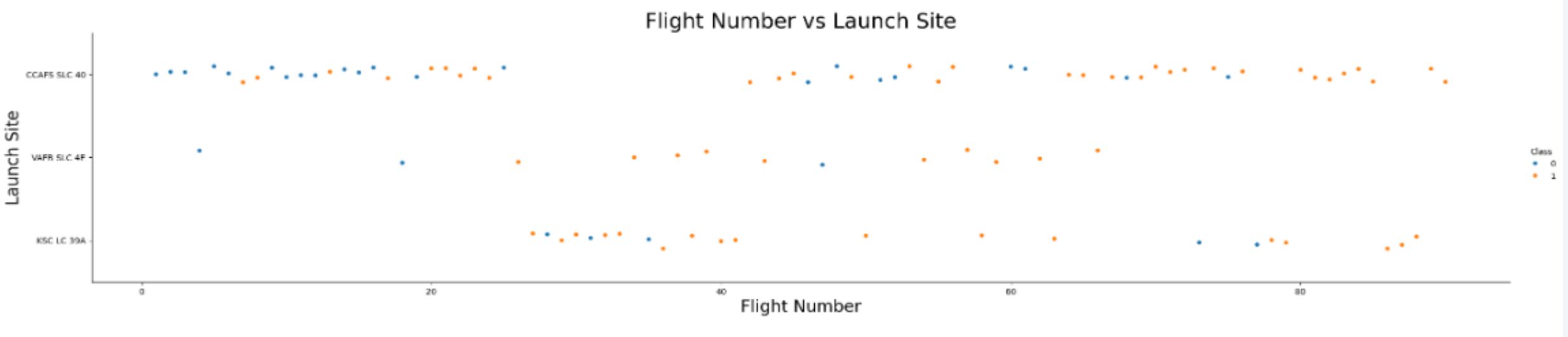
SpaceX Launch Records Dashboard

- **EDA:** Success rates improved with **Flight Number**; **LEO/VLEO** orbits outperformed GTO.
- **Interactive Tools: Folium** maps confirmed coastal safety; **Dash** allowed real-time payload filtering.
- **Machine Learning: Decision Tree & SVM** were top performers; all models reached **83.33% accuracy**.

	Method	Test Accuracy
0	Logistic Regression	0.833333
1	SVM	0.833333
2	Decision Tree	0.833333
3	KNN	0.833333

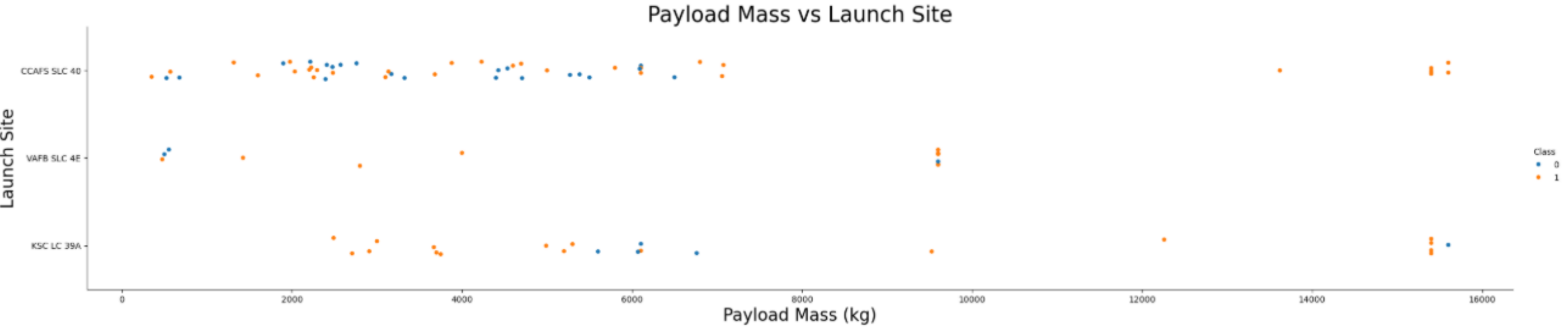

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- **Program Maturity:** Success rates correlate strongly with **Flight Number**, showing SpaceX's "learning curve" over time.
- **Site-Specific Reliability: CCAFS SLC-40** handled the majority of early (and more failed) missions, while later sites like **KSC LC-39A** benefited from matured technology.

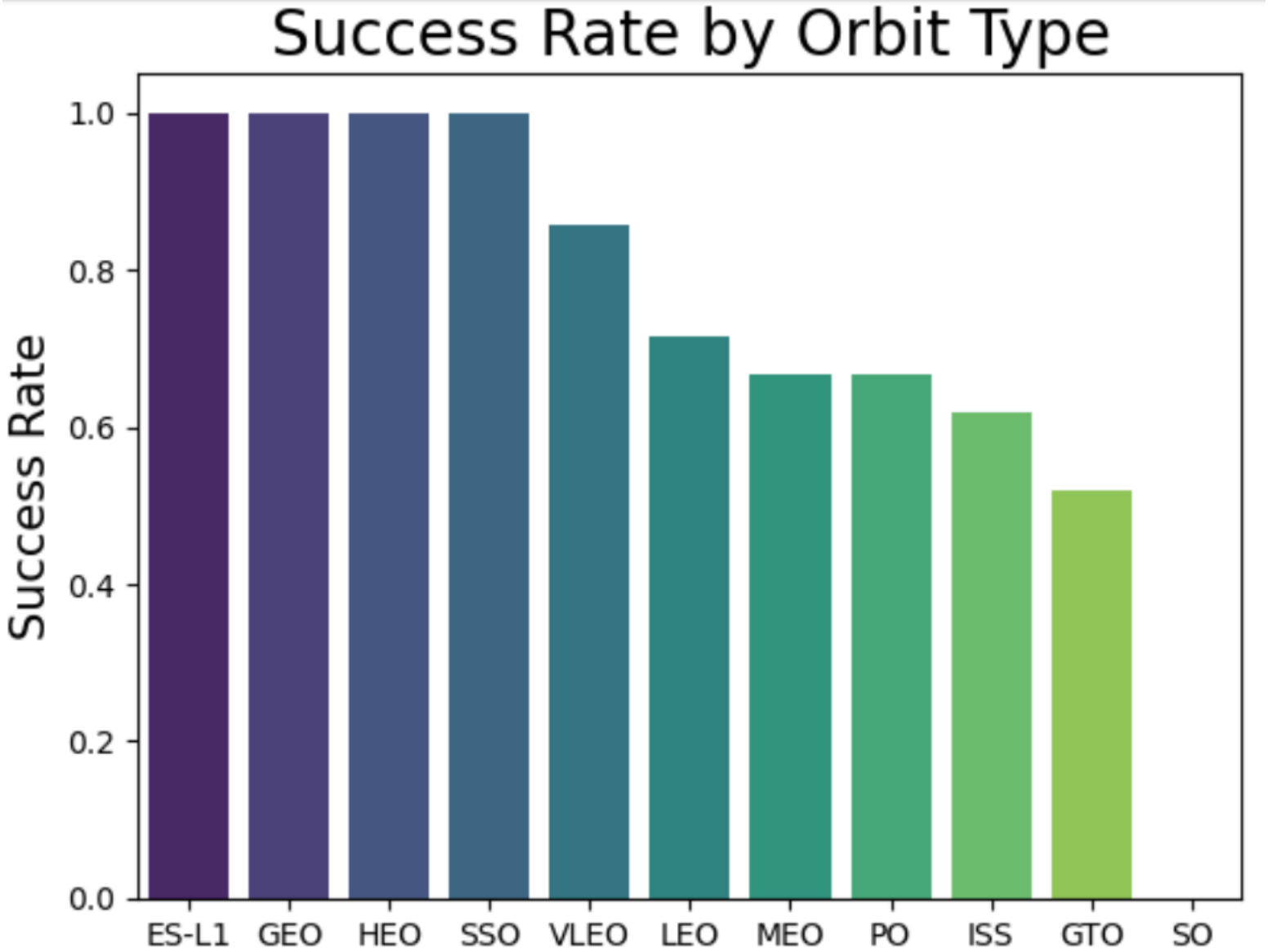
Payload vs. Launch Site



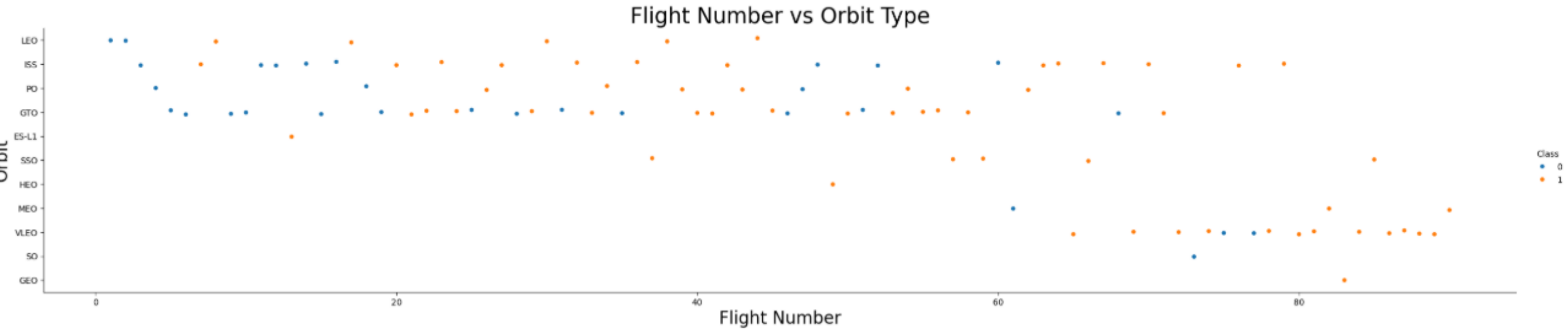
- **Payload Weight Distribution:** Most launches carry a payload between **2,000 kg and 8,000 kg**.
- **Success Correlation:** For all launch sites, missions with a **Payload Mass greater than 8,000 kg** show a very high success rate.

Success Rate vs. Orbit Type

- **High-Performance Orbits:** Missions to **ES-L1, SSO, HEO** show a **100% success rate**. Also **VLEO** with **85%**
- **Standard Orbits: LEO** missions show a high success rate of approximately **70%**.
- **Challenging Orbits: GTO** missions show a lower and more variable success rate of around **50%**.
- **Low Success Orbits: SO** orbit currently shows a **0% success rate** in the dataset.

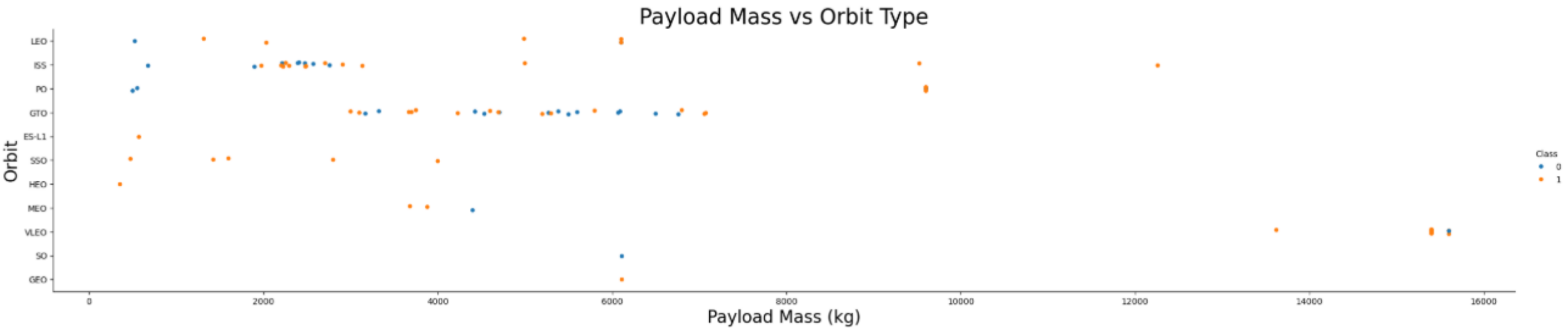


Flight Number vs. Orbit Type



- **Low-Earth Orbit (LEO) Maturity:** Earlier LEO missions show a mix of outcomes, but as flight numbers increased, success became nearly certain.
- **GTO Complexity:** Missions to **Geostationary Transfer Orbit (GTO)** appear consistently throughout the timeline, showing that while they are frequent, they remain challenging regardless of the flight number.
- **Newer Orbit Types:** Orbits like **VLEO** (Very Low Earth Orbit) appear primarily in later flight numbers and demonstrate a high success rate, benefiting from the program's maturity.

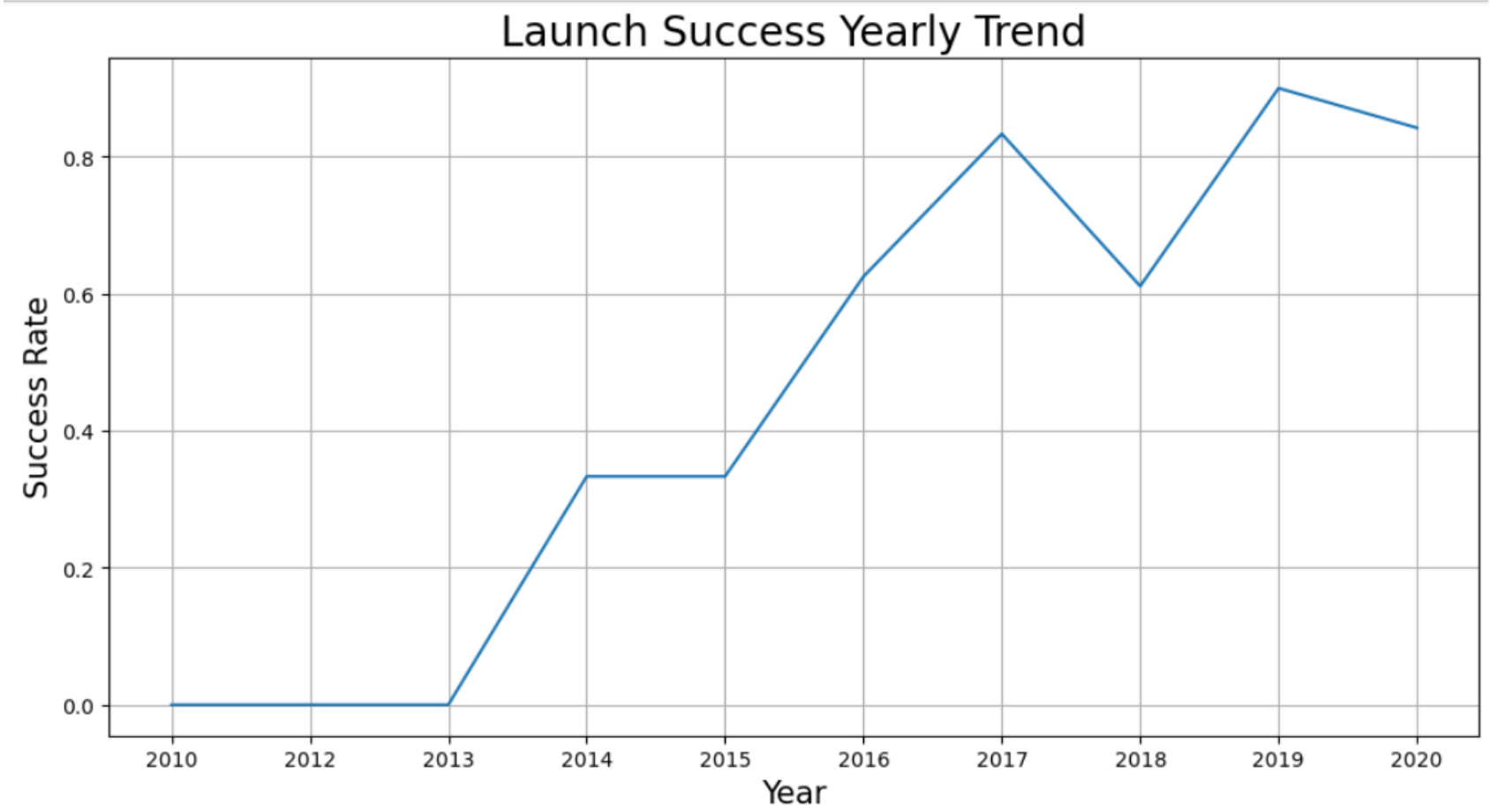
Payload vs. Orbit Type



- **Heavy-Lift Focus:** Payloads over **10,000 kg** are concentrated in **LEO, ISS, and PO** orbits.
- **GTO Challenge:** Success rates drop in **GTO** missions as payload mass increases toward maximum capacity.
- **PO Reliability: Polar Orbits** typically handle mid-range payloads with high landing consistency.
- **Optimal Range:** A "Success Cluster" is most visible for payloads between **4,000 kg and 6,000 kg** across all orbits.

Launch Success Yearly Trend

- **Initial Learning Phase (2010–2013):** The success rate remained at **0%** during the first few years as SpaceX focused on flight stability rather than recovery.
- **Breakthrough Period (2013–2017):** After the first successful landings, the trend shows a dramatic upward slope, proving the effectiveness of iterative engineering.
- **Operational Maturity (2018–2020):** The success rate reached its peak by 2020, demonstrating that first-stage recovery has become a standard, reliable capability.
- **Statistical Confirmation:** The positive slope of the line chart confirms that experience (time) is one of the strongest predictors of landing success.



All Launch Site Names

```
%sql SELECT "Launch_Site", COUNT(*) AS Launch_Count FROM SPACEXTABLE GROUP BY "Launch_Site";
```

```
 * sqlite:///my_data1.db
Done.
```

Launch_Site	Launch_Count
CCAFS LC-40	26
CCAFS SLC-40	34
KSC LC-39A	25
VAFB SLC-4E	16

- **Strategic Placement:** These sites provide SpaceX with access to both the Atlantic and Pacific oceans for safe launch trajectories and first-stage recovery.
- **Mission Specialization:** Florida sites handle equatorial and GTO orbits, while the California site (VAFB) is dedicated to Polar Orbits.

Launch Site Names Begin with 'CCA'

- **Primary Hub:** CCAFS SLC-40 is the most active site in the dataset, hosting the majority of early Falcon 9 developmental flights.
- **Geographic Advantage:** Its location in Florida allows for missions to take advantage of the Earth's rotation, making it the primary choice for heavy payloads and GTO orbits.
- **Historical Significance:** This site has seen the full evolution of the Falcon 9 program, from initial test failures to consistent, successful first-stage landings.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

Total_Payload_Mass
45596

Tracking total payload mass allows for the calculation of average mission costs and fuel efficiency across the booster fleet.

Average Payload Mass by F9 v1.1

Average_Payload_Mass
2928.4

The average payload of ~2,500 kg reflects its frequent use for ISS resupply missions and smaller commercial satellites during the mid-2010s.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

The success of this mission proved that first-stage recovery was physically and technically possible, laying the groundwork for the entire reusability program.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

- The presence of versions like **B1021.2** and **B1031.2** (the ".2" indicating a second flight) demonstrates the successful implementation of booster reusability.
- **Strategic Success:** Identifying this "success cluster" helps confirm that the Falcon 9 is highly optimized for recovering stages even during high-performance missions.

Total Number of Successful and Failure Mission Outcomes

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

Grouping the data this way allows us to calculate the global success rate of the program, which is the foundational metric for our predictive models.

Boosters Carried Maximum Payload

All boosters carrying the maximum payload belong to the **Block 5** generation. This version was designed specifically for high reusability and maximum thrust.

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

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

Each "Failure (drone ship)" provided critical telemetry data that eventually led to the first successful landing on a drone ship in April 2016.

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

The equal number of successes and failures on drone ships (5 each) highlights the high-risk nature of landing on a moving platform at sea during this development window.

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

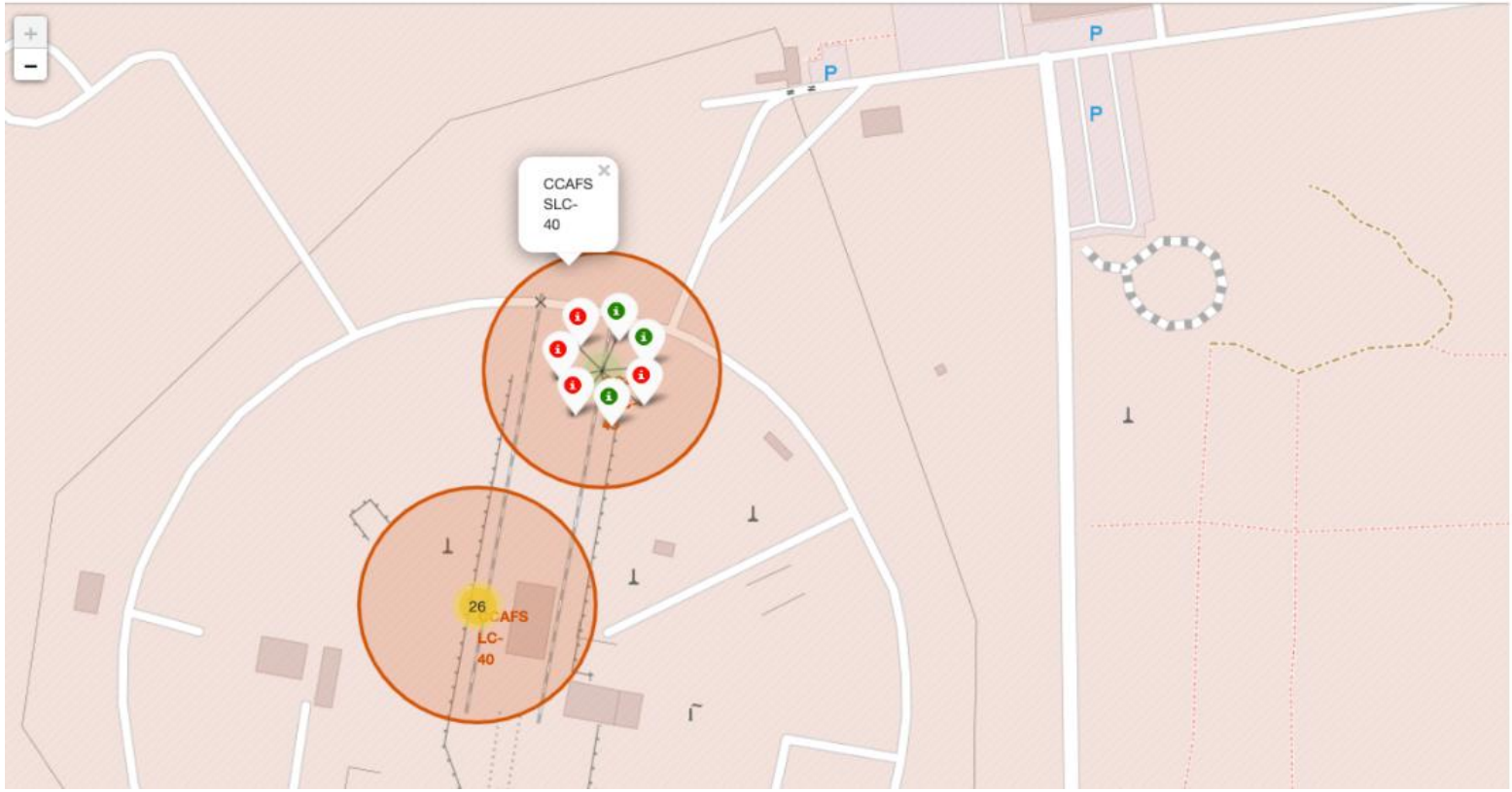
Section 3

Launch Sites Proximities Analysis

Geographical Distribution of SpaceX Launch Sites

- **East Coast Hub (46 Launches):** Includes **CCAFS SLC-40** and **KSC LC-39A** in Florida, handling the vast majority of mission volume.
- **West Coast Site (10 Launches): VAFB SLC-4E** in California, primarily used for polar orbit missions.
- **Coastal Strategy:** All sites are positioned on coastlines to ensure launch trajectories remain over open water for public safety.

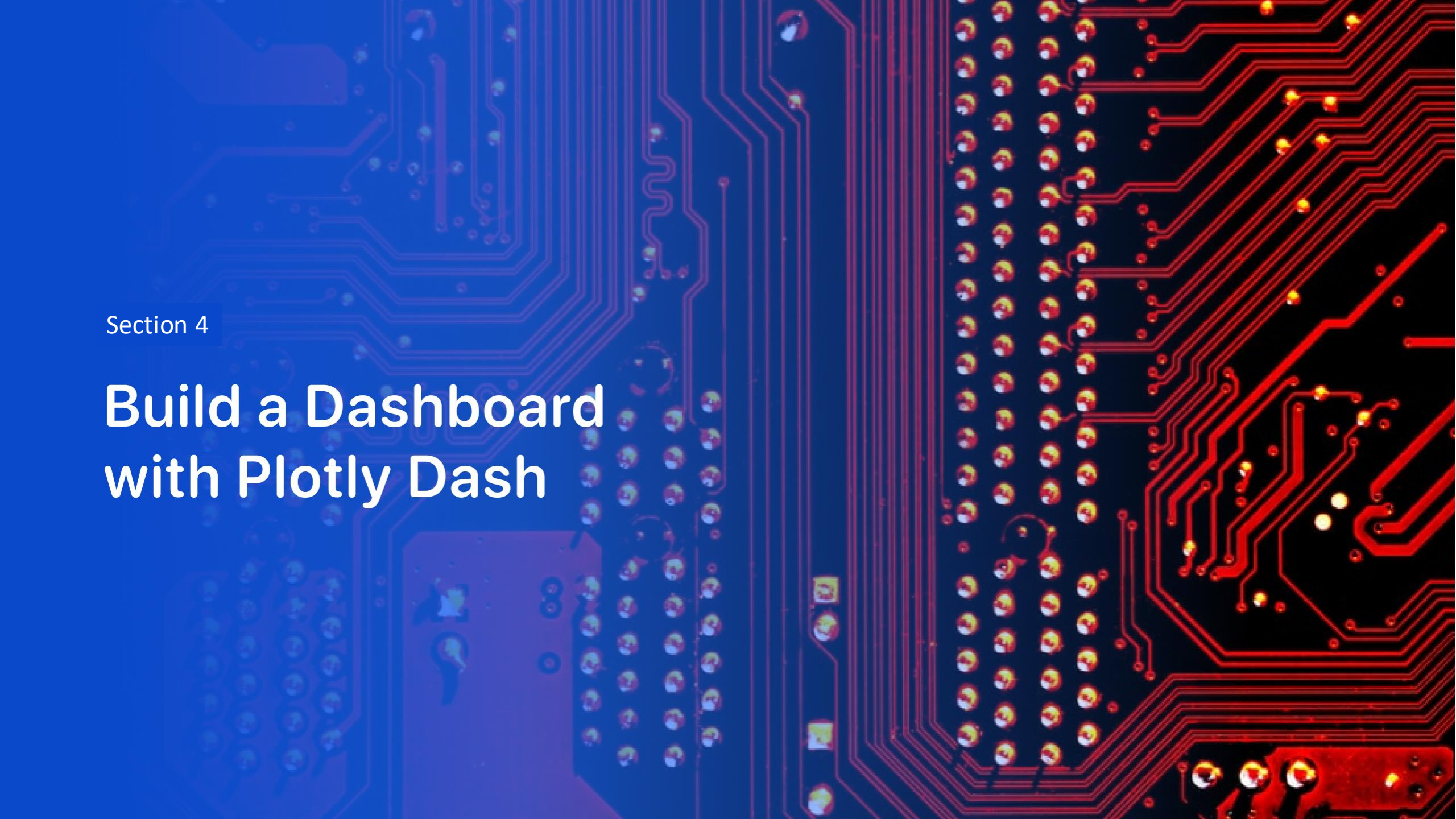
Site-Specific Mission Outcomes (CCAFS SLC-40)



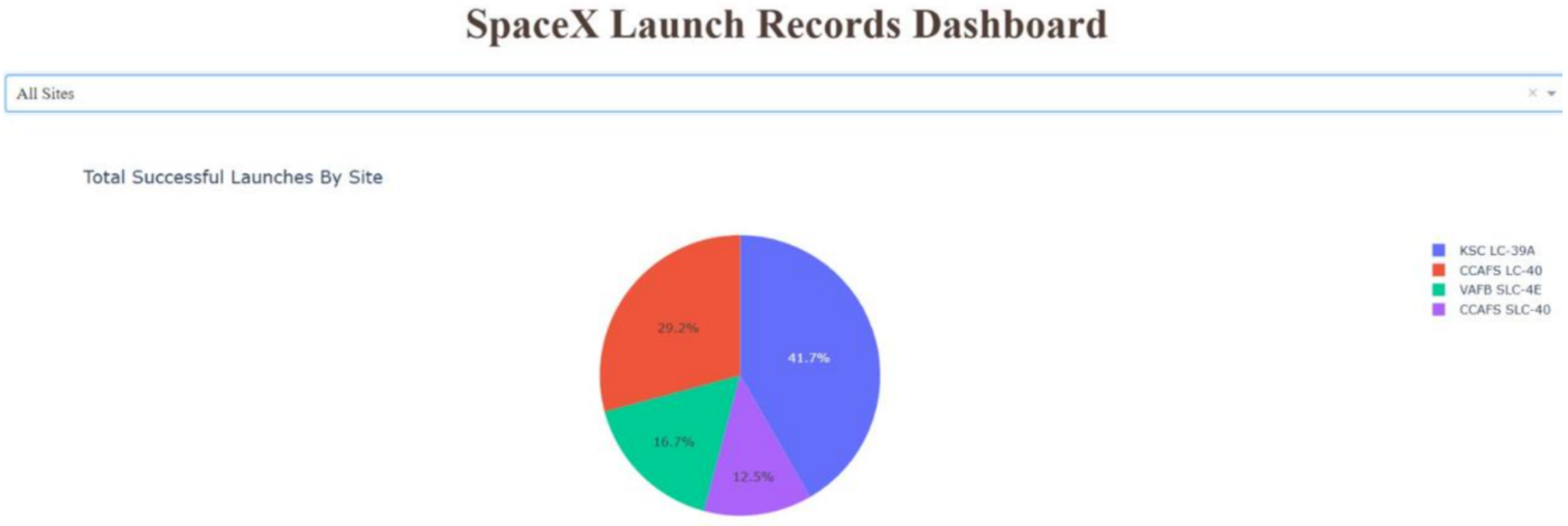
- **Color-Labeled Markers:** The map uses green "i" markers to represent successful landings and red "i" markers to represent failed landings.
- **Proximity Circles:** Orange circular perimeters are drawn around the launch pads to visualize the immediate facility boundaries.
- **Marker Clusters:** The yellow cluster icon (labeled "26") groups multiple mission data points into a single view for better map readability.

Geospatial Safety & Infrastructure Analysis

- **Distance Lines (PolyLines):** Plotted lines represent the shortest path between launch pads and key landmarks.
- **Coastal Proximity:** All launch sites are located within a few kilometers of the ocean.

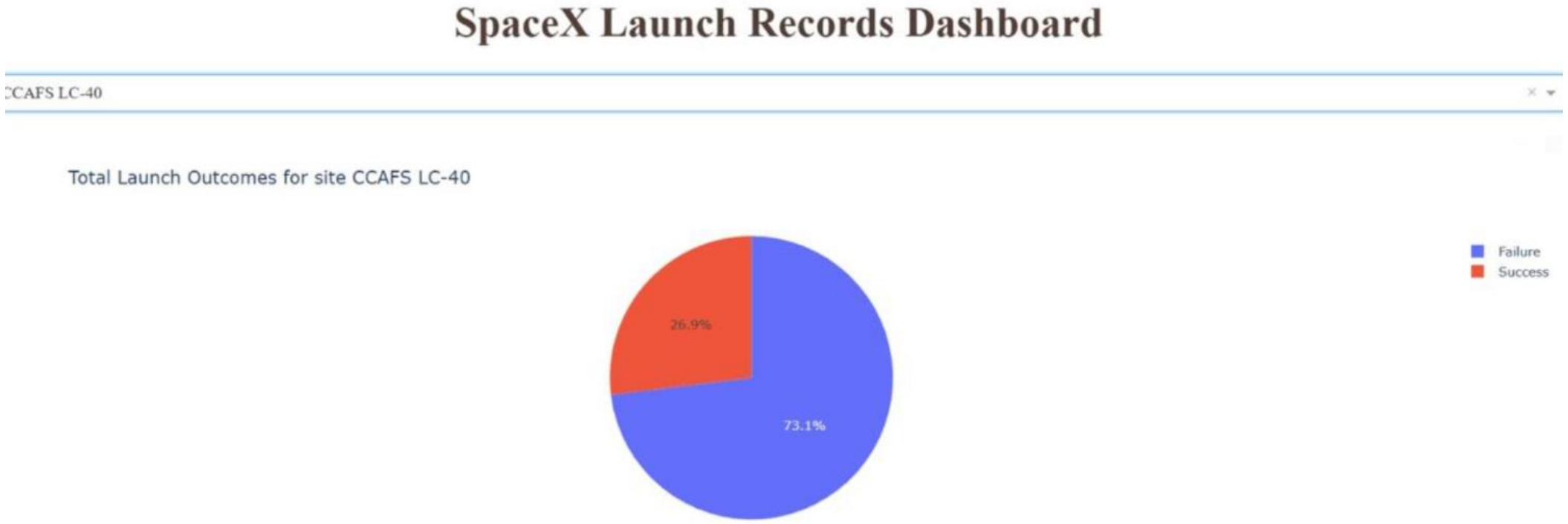

Section 4
Build a Dashboard with Plotly Dash

Distribution of Launch Success Across All Sites



- Visualizes the total success count versus failure count for the entire SpaceX program.
- **Total Success Rate:** Provides a high-level view of SpaceX's reliability across all geographic locations.

Success Rate Profile: KSC LC-39A



- **Pie Chart:** A localized breakdown showing only the Success vs. Failure ratio for the top-performing site.
- Explain the important elements and findings on the screenshot

Correlation: Payload Mass, Booster Version, and Mission Outcome

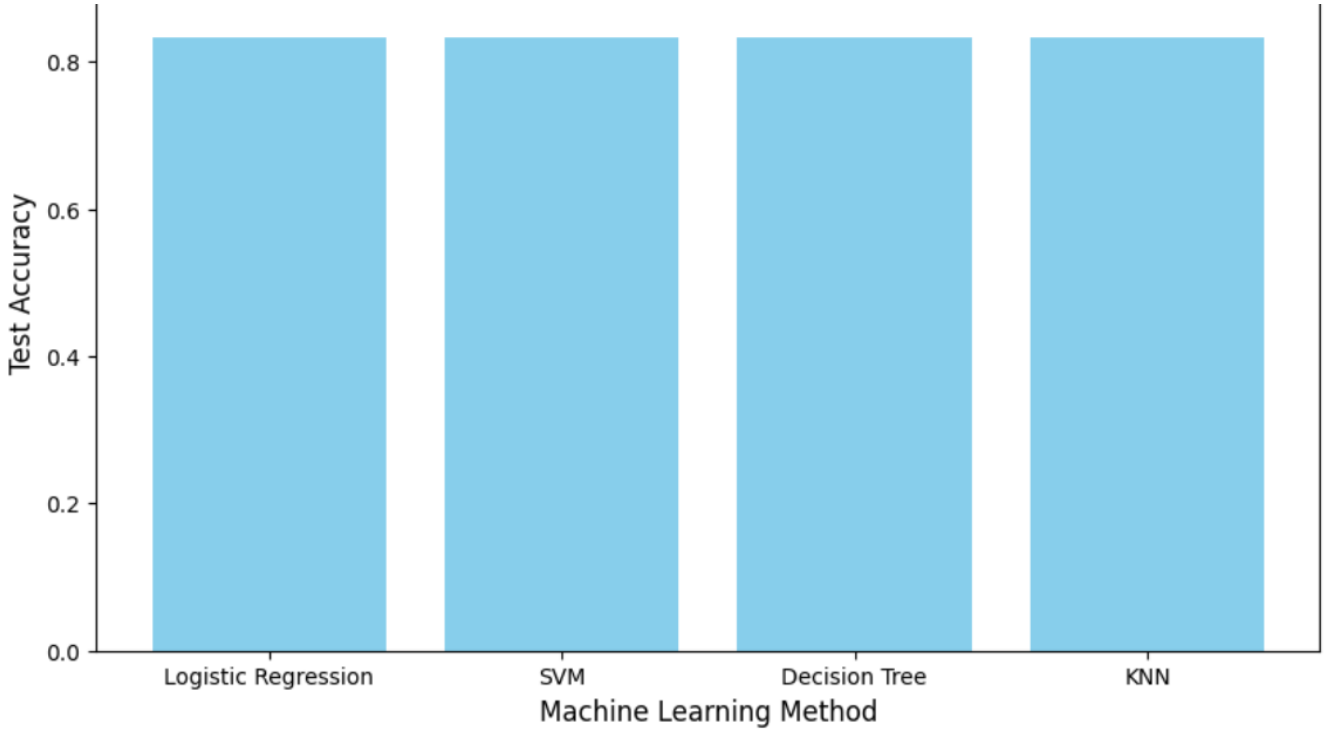
- **Scatter Plot:** X-axis (Payload Mass in kg), Y-axis (Class: 0 or 1), Color (Booster Version).
- **Interaction:** The **Range Slider** is adjusted to show various payload segments (e.g., 0kg to 10,000kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

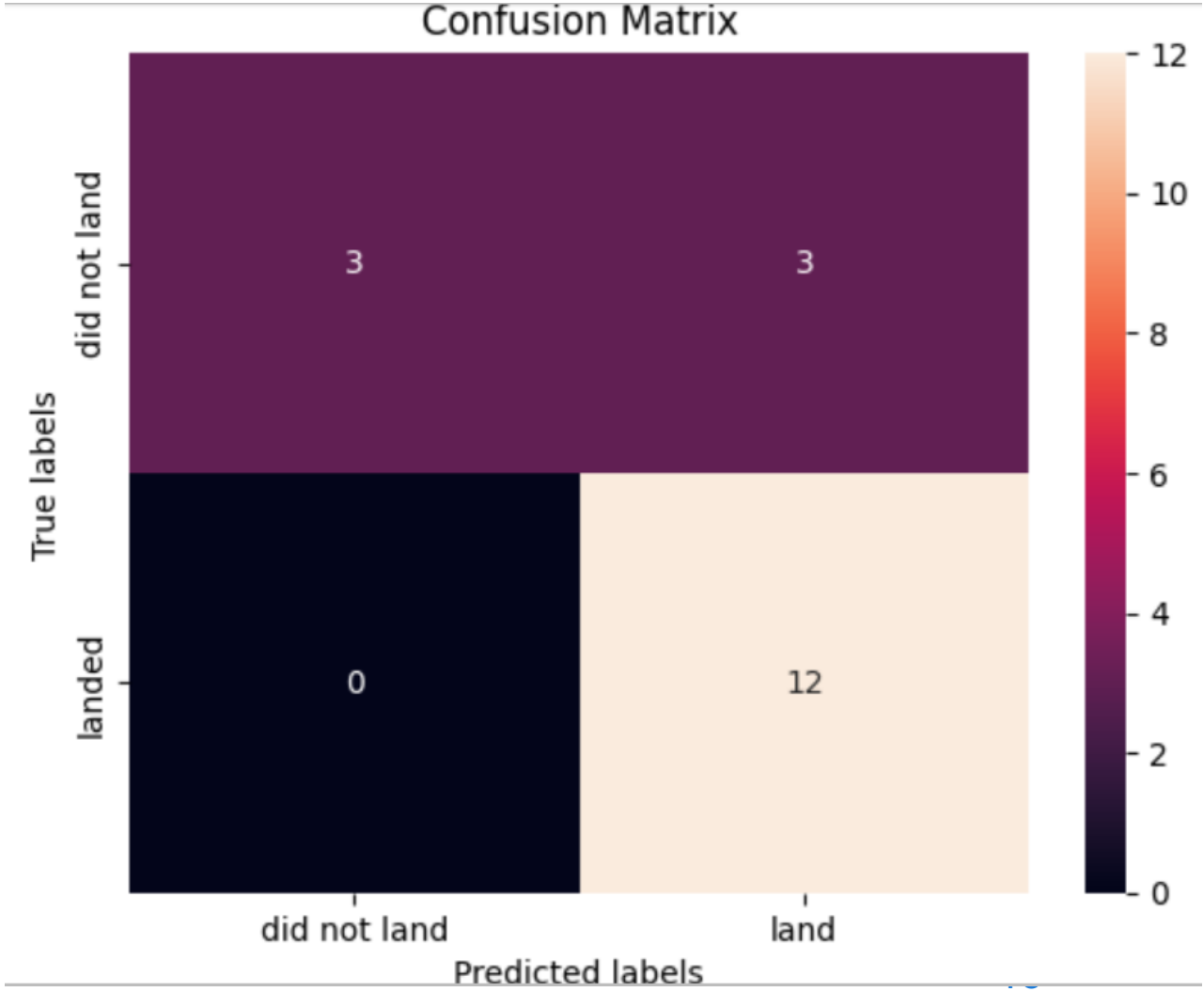
- **Accuracy Score:** All four models (**Logistic Regression, SVM, Decision Tree, and KNN**) achieved an identical accuracy of **83.33%** on the test dataset.
- **The Best Model:** Since the test accuracy is tied across all models, the "best" model is typically determined by its performance on the training data or its ability to minimize False Positives in the Confusion Matrix.
- the **Decision Tree** or **SVM** is often cited as a top performer due to higher training accuracy or faster convergence, but for the final prediction on unseen data, all models were equally effective.



Confusion Matrix

A Confusion Matrix breaks down the predictions into four quadrants. For the SpaceX project, the results usually look like this:

- **True Positives (Top Left - 12):** The model correctly predicted **12 successful landings**.
- **True Negatives (Bottom Right - 3):** The model correctly predicted **3 landing failures**.
- **False Positives (Top Right - 3):** The model predicted a success, but it was actually a **failure**.
- **False Negatives (Bottom Left - 0):** The model predicted a failure, but it was actually a **success**.



Conclusions

- **Launch site, payload mass, orbit type, and flight number** are key factors influencing Falcon 9 first-stage landing success.
- **Landing success improves over time**, reflecting SpaceX's operational learning and booster reusability improvements.
- **Exploratory and interactive analyses** (EDA, SQL, Folium maps, and Dash dashboard) consistently revealed clear performance patterns.
- **Machine learning models** can effectively predict landing outcomes, with the **Decision Tree classifier** achieving the best overall performance.

Appendix

- This appendix includes additional materials and assets created during the project to support transparency, reproducibility, and peer review. The following resources are provided:
- **Python notebooks** covering data collection (API and web scraping), data wrangling, exploratory data analysis, interactive visualization, and predictive modeling
- **Python code snippets** used for data cleaning, feature engineering, visualization, and machine learning
- **SQL queries and outputs** used for exploratory data analysis and aggregation of launch data
- **Charts and figures** generated during EDA, including payload, orbit, launch site, and success trend visualizations
- **Interactive visualization outputs**, including Folium maps and Plotly Dash dashboard screenshots
- **Machine learning results**, such as model accuracy comparisons and confusion matrices
- **Datasets** generated and used throughout the project for analysis and modeling
- All project materials and source code are available in the GitHub repository for peer review and reference:

  👉 **https://github.com/Mohamed-bns/IBM-Data-Science-Capstone**

Thank you!