Winning Space Race with Data Science

Rahul Arya

11-12-2024

Outline

Executive Summary

Methodology Overview

To address the business problem, I employed the following methods:

1. **Data Collection**:
   - **Web Scraping**: Extracted historical SpaceX launch data from Wikipedia.
   - **API Calls**: Gathered detailed, structured data from SpaceX's REST API.

2. **Exploratory Data Analysis (EDA)**:
   - Used Python libraries (pandas, matplotlib, seaborn) to clean, visualize, and explore data for meaningful patterns.

Executive Summary

3. Interactive Analysis:

- Developed an interactive dashboard with **Plotly Dash** to analyze launch data.
- Created geographic visualizations with **Folium** to study launch site proximity.

4. Predictive Modeling:

- Applied machine learning models (SVM, Decision Trees, Logistic Regression) to predict landing success.
- Performed hyperparameter tuning to find the best model for accurate predictions.

Executive Summary

Key Result

The final model accurately predicts the success of Falcon 9 first-stage landings, providing valuable insights for cost reduction and competitive bidding.

Introduction

Background

- SpaceX has revolutionized the space industry by making rocket launches more affordable through reusable rocket technology.
- The Falcon 9 rocket's first stage is designed to land and be reused, significantly reducing launch costs from $165 million to $62 million.

Business Problem

- Predicting the success of Falcon 9 first-stage landings is critical to evaluating cost efficiency.
- Competitors can use this prediction model to make informed decisions when bidding against SpaceX for launch contracts.

Introduction

Objective

- Develop a predictive model to determine if the Falcon 9 first stage will successfully land, leveraging historical and technical data.
- Provide insights into the factors influencing landing success to aid decision-making and improve cost optimization strategies.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - **Web Scraping**: Used BeautifulSoup to extract historical launch data from Wikipedia, focusing on key parameters like time & dates, booster versions, and outcomes.
  - **API Calls**: Accessed SpaceX's REST API to retrieve detailed and real-time information, such as rocket specifications, payload details, and launchpad coordinates.

Methodology

Executive Summary

- Perform data wrangling:
  - Handled missing and inconsistent values in the dataset.
  - Standardized data formats for easier analysis (e.g., date).
  - Created a 'class' column to label outcomes, categorizing them based on success or failure for better analysis and modeling.
- Perform exploratory data analysis (EDA) using visualization and SQL:
  - Conducted using Python libraries like pandas, matplotlib, and seaborn.
  - Visualized key trends and patterns, such as:Payload mass vs. landing success.
    - Success rates by launch site and orbit type.
    - Leveraged SQL queries for in-depth data exploration and filtering.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash:
  - Built a **Plotly Dash** dashboard to analyze launch records interactively, including pie charts and scatter plots.
  - Created maps using **Folium** to study launch site proximity and surrounding infrastructure.
- Perform predictive analysis using classification models:
  - Employed machine learning models:
    - Logistic Regression, Decision Trees, and SVM.
  - Optimized model performance through hyperparameter tuning.
  - Evaluated model success with metrics like accuracy, precision, and recall.

Data Collection

1. Web Scraping:

- **Source**: Wikipedia's SpaceX launch page.
- **Tools Used**:
  - **BeautifulSoup**: For parsing HTML and extracting relevant data from tables.
  - **Requests**: To fetch the webpage content.
- **Data Extracted**:Launch dates, payload mass, booster version, landing status, and mission details.

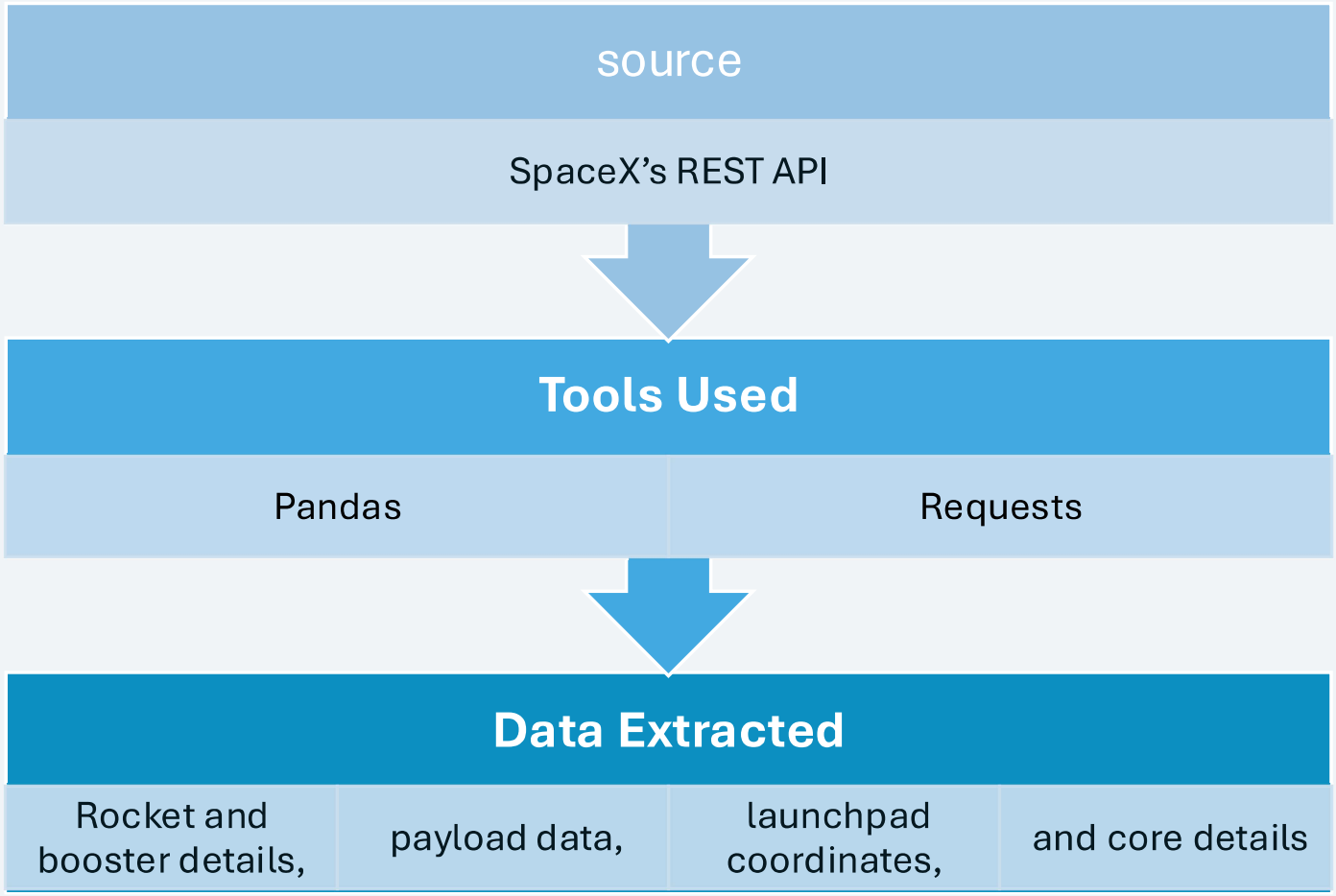
Data Collection

2. API Calls:

- **Source**: SpaceX's REST API.
- **Tools Used**:
  - **Requests**: To interact with the API and fetch JSON data.
  - **Pandas**: To convert JSON into structured DataFrames for analysis.
- **Data Extracted**:Rocket and booster details, payload data, launchpad coordinates, and core details.

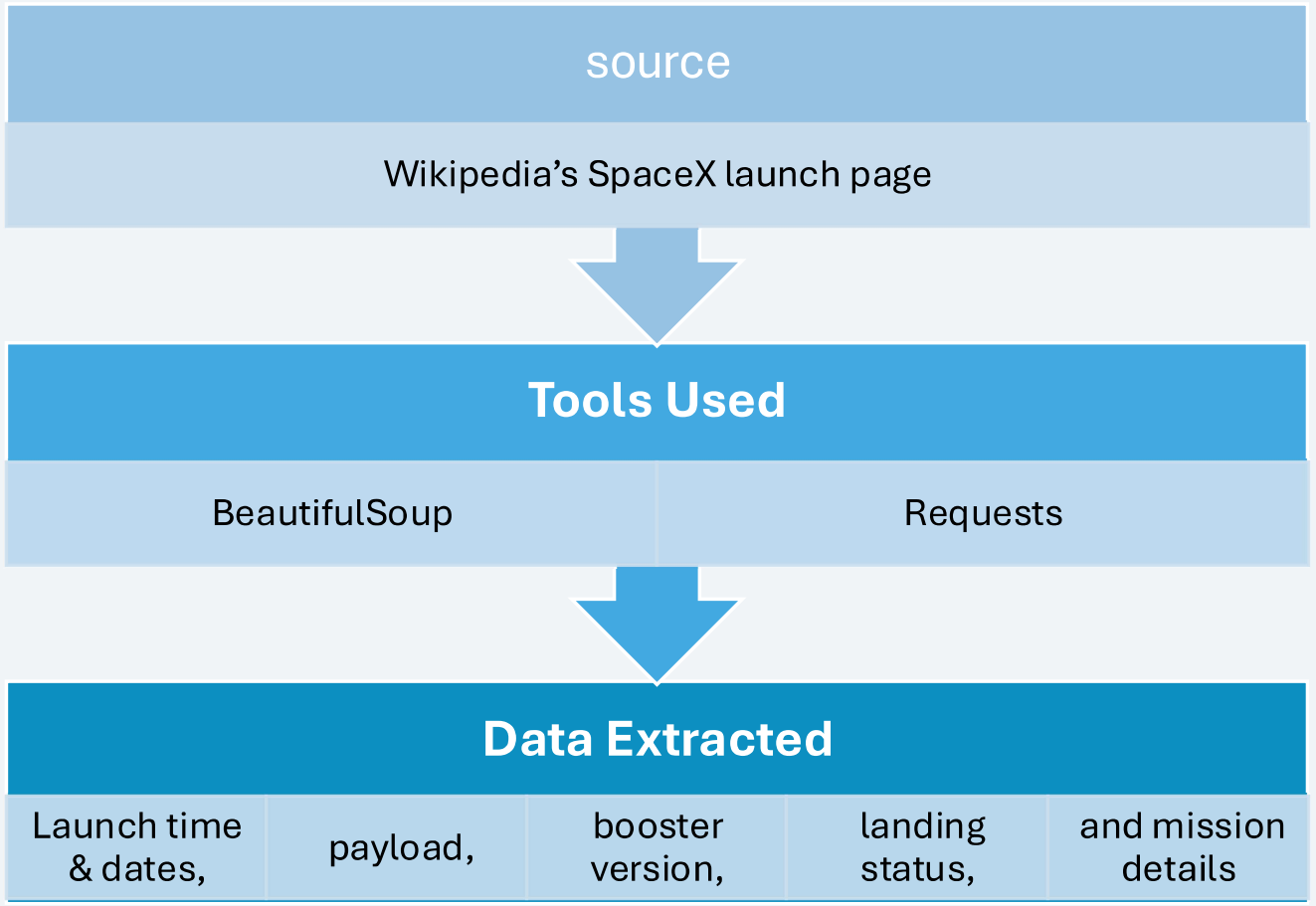
Data Collection – SpaceX API

- GitHub URL : capstone/1.jupyter-labs-spacex-data-collection-api.ipynb at main · lonewolf1810/capstone



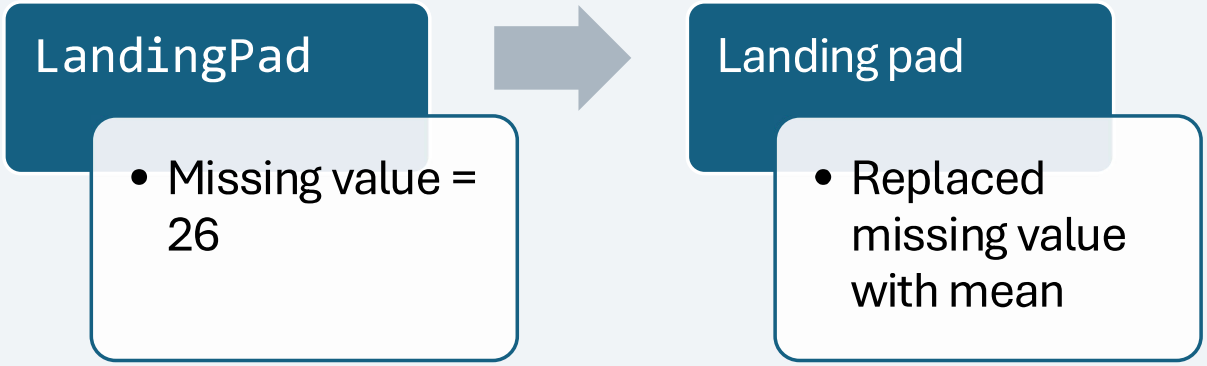
Data Collection - Scraping

- GitHub URL
  capstone/2.jupyter-labs-webscraping.ipynb at main · lonewolf1810/capstone



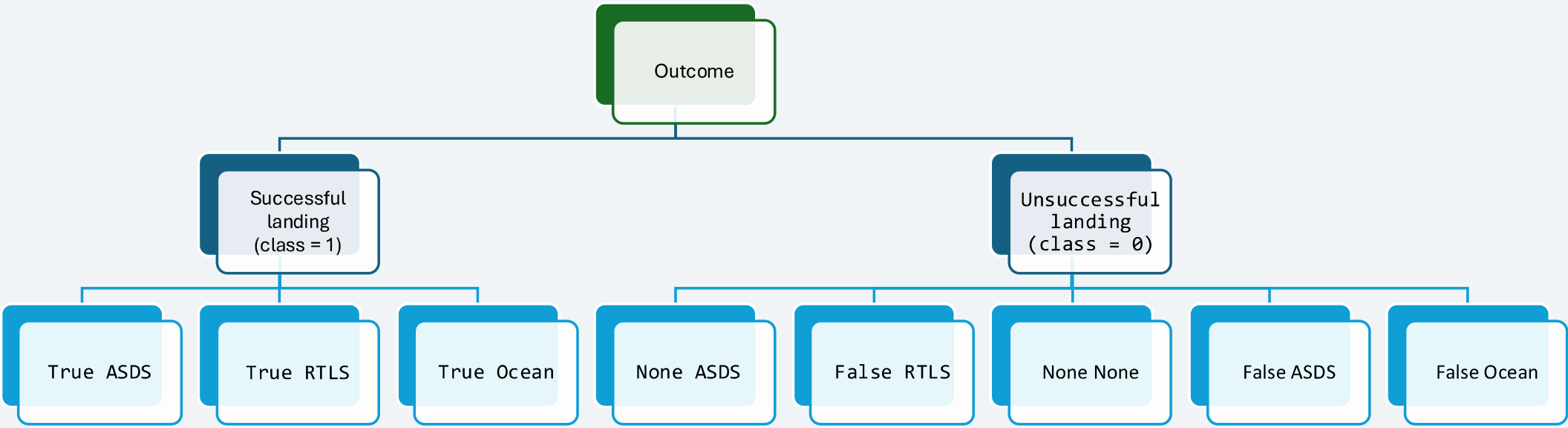
Data Wrangling

- **Handling Missing Data**
  - **Landing Pad**: Returned None for missions without landing pads. This is expected and does not require further handling.
  - **Payload Mass**: Replaced missing values with the mean of the column for consistent analysis.



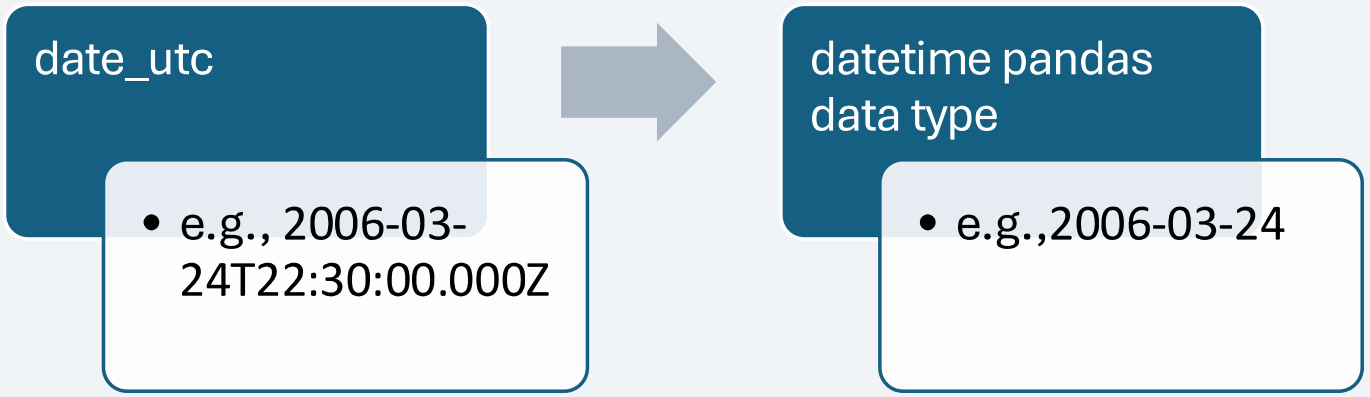
Data Wrangling

- **Feature Engineering**
  - Created a `class` column to label outcomes:
    - `1`: Successful landing. ;
    - `0`: Unsuccessful landing.



Data Wrangling

- **Data Formatting**
  - Standardized date and time formats.



- GitHub URL capstone/3.labs-jupyter-spacex-Data wrangling.ipynb at main · lonewolf1810/capstone

EDA with Data Visualization

- **Flight Number vs. Launch Site (Scatter Plot)**
  - *Purpose*: To analyze the consistency and frequency of launches at different sites.
- **Payload Mass and Launch Site (Scatter Plot)**
  - *Purpose*: To explore payload capacities handled by specific sites.
- **Success Rate by Orbit Type (Bar Chart)**
  - *Purpose*: To identify which orbit types achieve higher success rates.
- **Flight Number vs. Orbit Type (Scatter Plot)**
  - *Purpose*: To observe the variety of orbit types as the number of missions increases.
- **Payload Mass vs. Orbit Type (Scatter Plot)**
  - *Purpose*: To investigate compatibility between payload masses and orbit types.

EDA with Data Visualization

- **Yearly Launch Success Trend (Line Chart)**
  - *Purpose*: To visualize improvements in launch success rates over time.

- GitHub URL capstone/5.edadataviz.ipynb at main · lonewolf1810/capstone

EDA with SQL

- **Unique Launch Sites**: Identified all distinct launch sites used by SpaceX.
- **Filtered Launch Sites**: Retrieved records where launch sites begin with "CCA".
- **Payload Analysis**: Calculated total payload mass for NASA CRS missions and average payload mass for booster version F9 v1.1.
- **Landing Success Dates**: Found the date of the first successful ground pad landing.
- **Drone Ship Landings**: Listed boosters successful on drone ships with payloads between 4000 and 6000 kg.
- **Year-Specific Failures**: Filtered records by landing outcomes, booster versions, and launch sites for failures in 2015.

EDA with SQL

- **Mission Outcomes**: Counted successful and failed mission outcomes.
- **Maximum Payload Carriers**: Identified boosters carrying the maximum payload mass using subqueries.

- GitHub URL capstone/4.capstone_sql.ipynb at main · lonewolf1810/capstone

Build an Interactive Map with Folium

- **Launch Sites Visualization**: Displayed all SpaceX launch sites as markers with circles and pop-ups for identification.
- **Class Markers**: Added markers for each launch event, colored based on success (green) or failure (red).
- **Proximity Analysis**: Calculated distances from launch sites to nearby infrastructure (e.g., coastlines, railways, highways, and cities) and visualized them with markers and lines.
- **Mouse Position Plugin**: Enabled live display of latitude and longitude when hovering over the map.
- **Purpose :**
  - To analyze launch site locations, assess proximity to critical infrastructure, and visualize mission outcomes interactively.
- nbviewer GitHub URL : Jupyter Notebook Viewer (Note: Ctrl + Click to open link)

Build a Dashboard with Plotly Dash

- **Key Features :**
  - **Launch Site Dropdown**: Allows users to filter data by specific launch sites or view all sites.
  - **Payload Range Slider**: Enables selection of payload mass ranges to customize the analysis.
  - **Success Pie Chart**: Visualizes the success rate of launches for selected sites.
  - **Scatter Plot**: Shows the relationship between payload mass and launch outcomes, categorized by booster versions.
- **Purpose :** The dashboard enables interactive analysis of launch data, making it easy to identify trends, success rates, and performance based on payload and launch site.
- GitHub URL capstone/7.capstone_dash.py at main · lonewolf1810/capstone

Predictive Analysis (Classification)

- **Overview:** Built a machine learning pipeline to predict whether the Falcon 9 first stage will land successfully, utilizing **Logistic Regression**, **SVM**, **Decision Trees**, and **KNN**.
- **Objective**: Predict landing success to reduce launch costs.
- **Steps**
  - **Data Preparation**: Standardized data, created a class column, split into training and test data.
  - **Model Training**: Tuned hyperparameters using GridSearchCV for optimal performance.
- **Results**: **Best Model**: Decision Tree Classifier with accuracy **0.9444** on the test data. (see appendix A for features)
  - All models showed **0.8333** accuracy on test data.
- GitHub URL: capstone/8.capstone_predictive_analysis.ipynb at main · lonewolf1810/capstone

Predictive Analysis (Classification)

Prepare data

- Standardized data
- Create a class column
- split into training and test data

Train model

- **Models**: Logistic Regression, SVM, Decision Trees, KNN
- **Hyperpara. Tuning**: **GridSearchCV**

Evaluate model

- Assess models using accuracy and confusion matrix on the test set

Best model

- Decision Tree showed the best performance, achieving **0.9444 accuracy** on test data.

Results

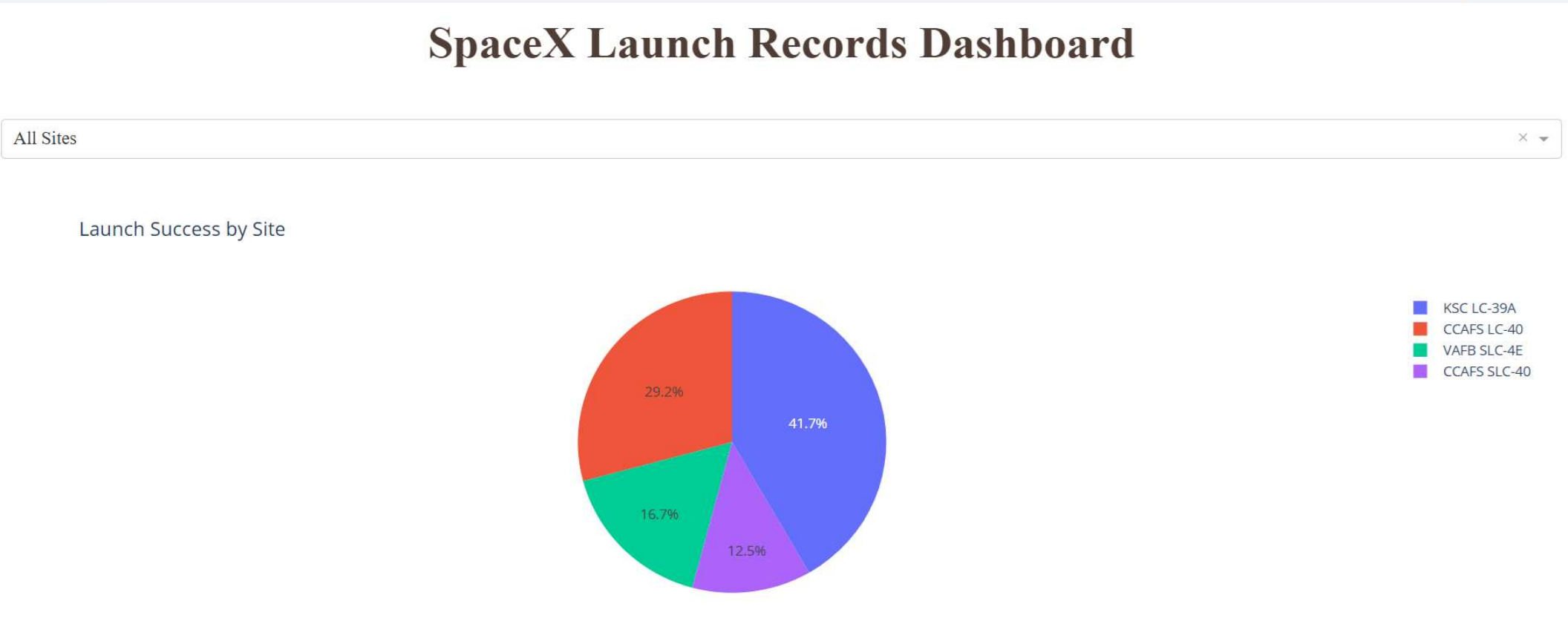
- Success rate of landing successfully : 66.66%
- **SQL Analysis**:
  - Identified key launch sites and summarized payload statistics.
  - Most missions were successful, with **38 successful landings** and **99 successful mission outcomes overall**.
  - High payload mass was associated with NASA CRS missions and certain booster versions.
- **Pandas and Matplotlib Analysis**:
  - Success rates increased over time, with steady improvement from **2013 to 2020**.

Results

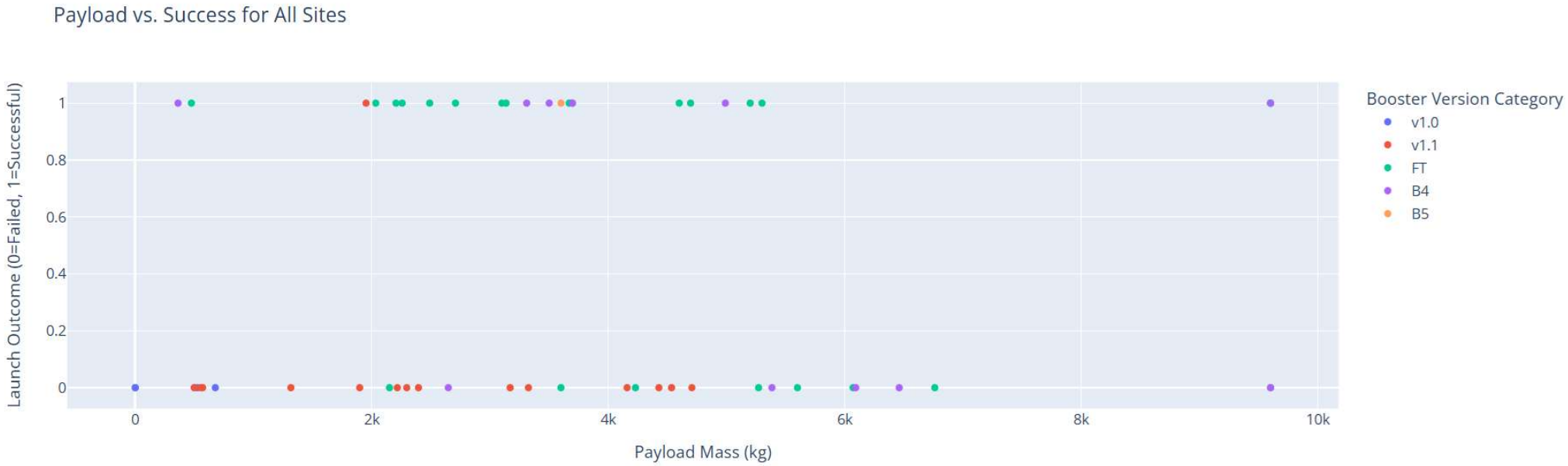
- Heavy payloads were more successful in Polar, LEO, and ISS orbits, while GTO showed mixed outcomes.
- Success ratio improved with flight numbers, particularly at CCAFS SLC-40 and KSC LC-39A.

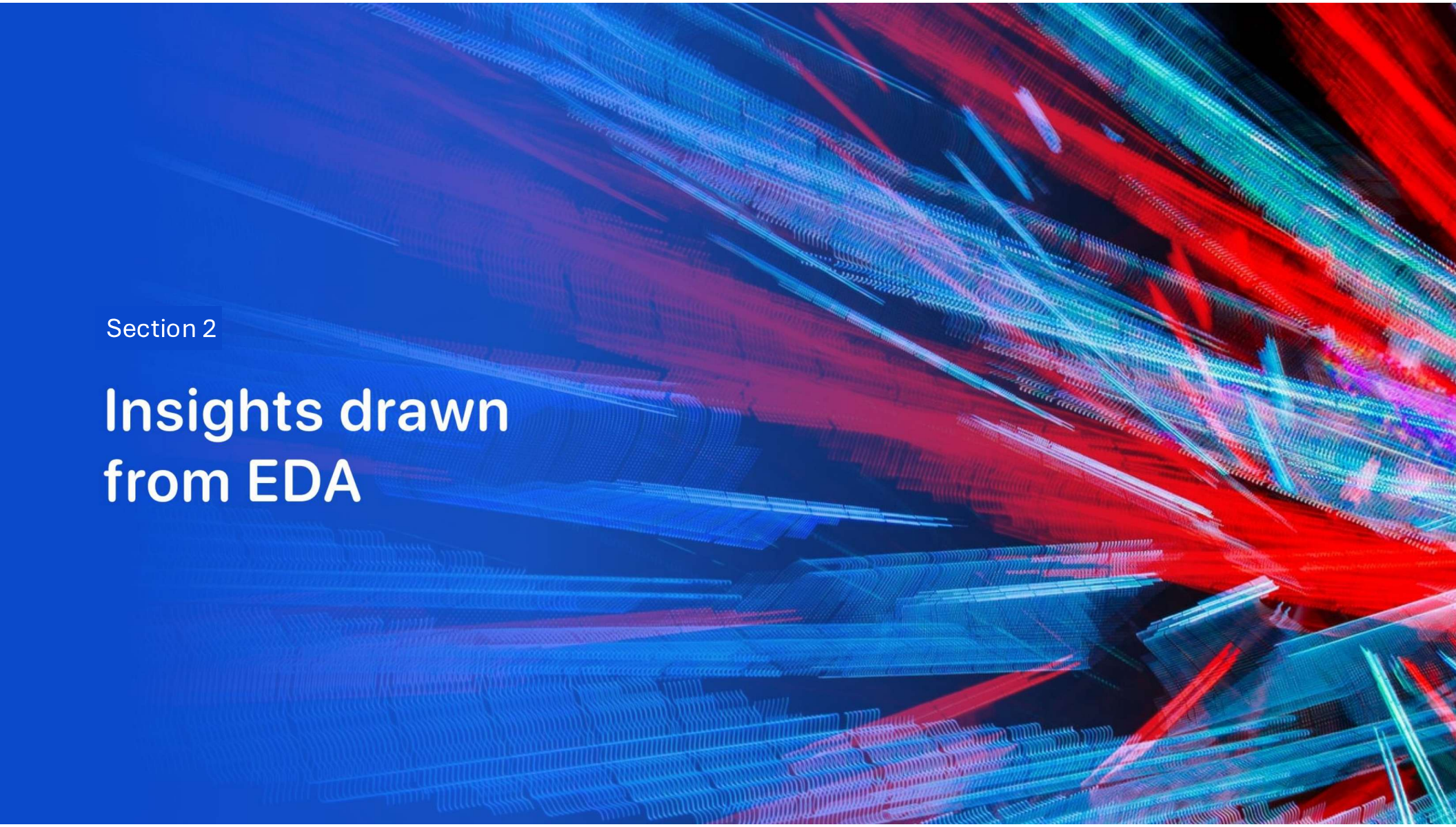
- **Predictive analysis results:**
  - **Best Model**: Decision Tree Classifier with accuracy **0.9444** on the test data. (see appendix A for features)
  - All models showed **0.8333** accuracy, but Decision Tree performed best with tuned hyperparameters.

Results Interactive analytics demo in screenshots



Results Interactive analytics demo in screenshots

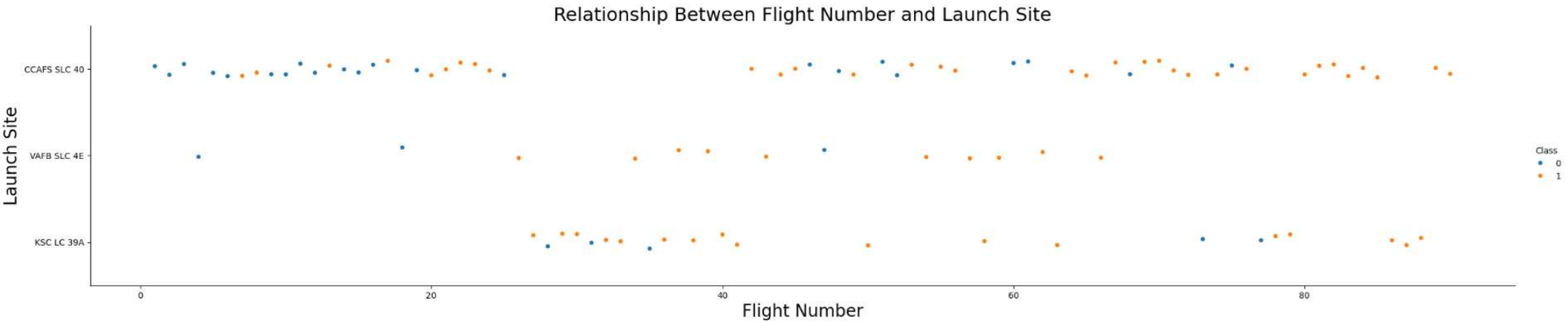




Section 2

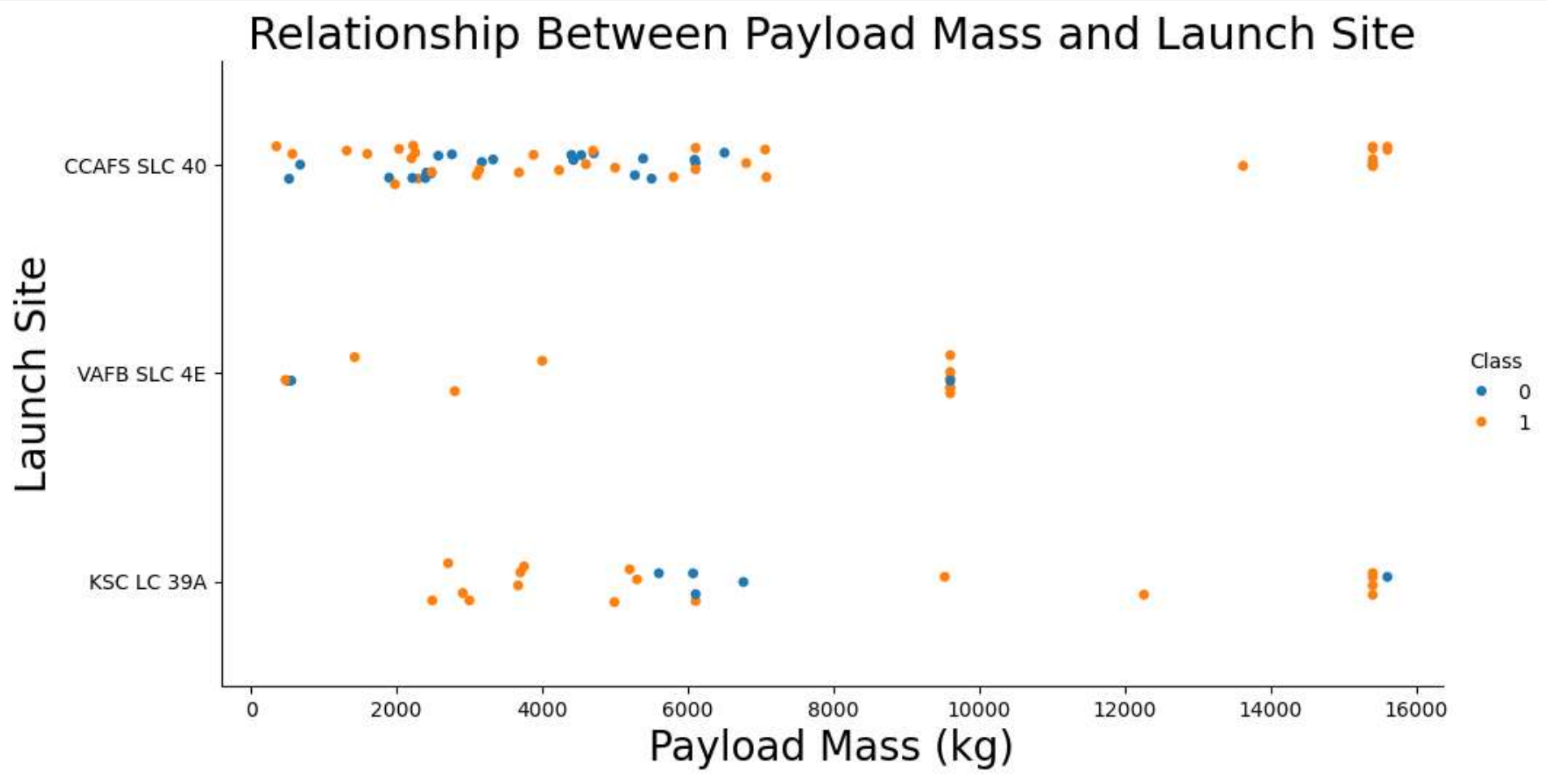
Insights drawn from EDA

Flight Number vs. Launch Site



- As the number of flights increases, the success rate at each launch site also improves.

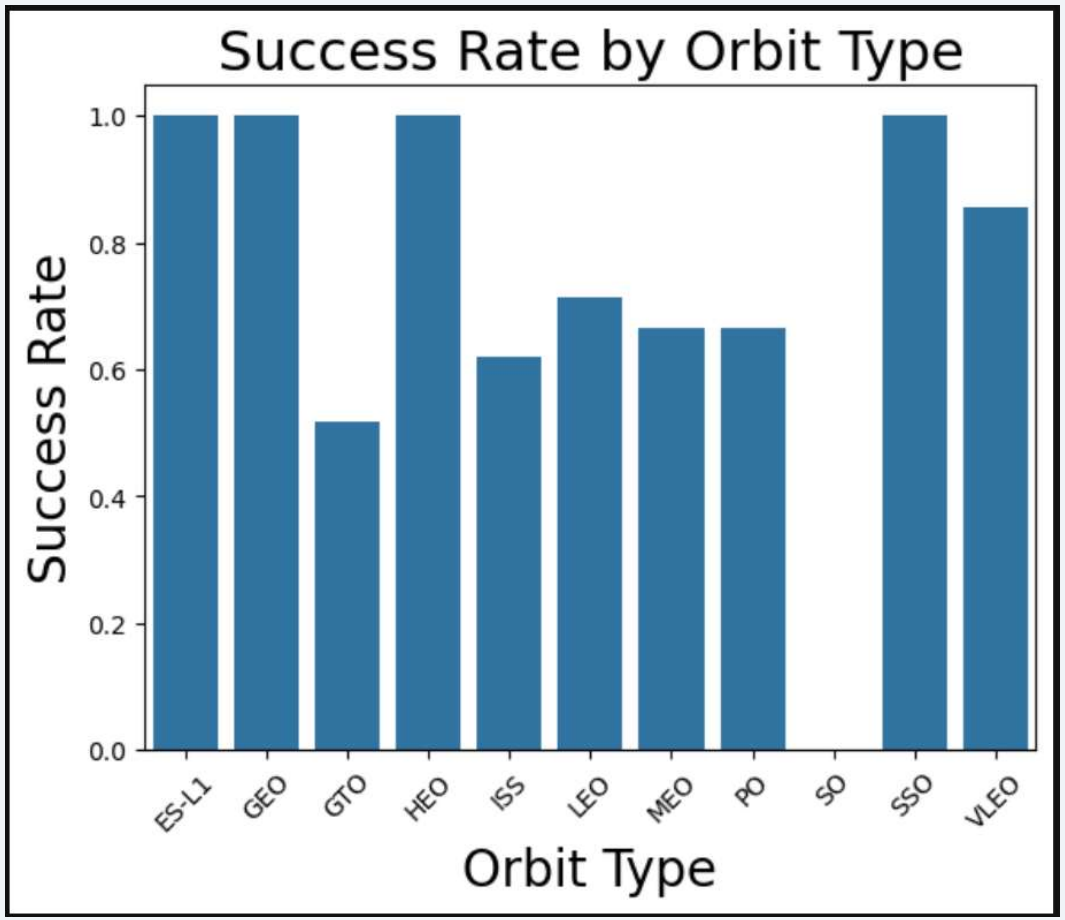
Payload vs. Launch Site



Payload vs. Launch Site

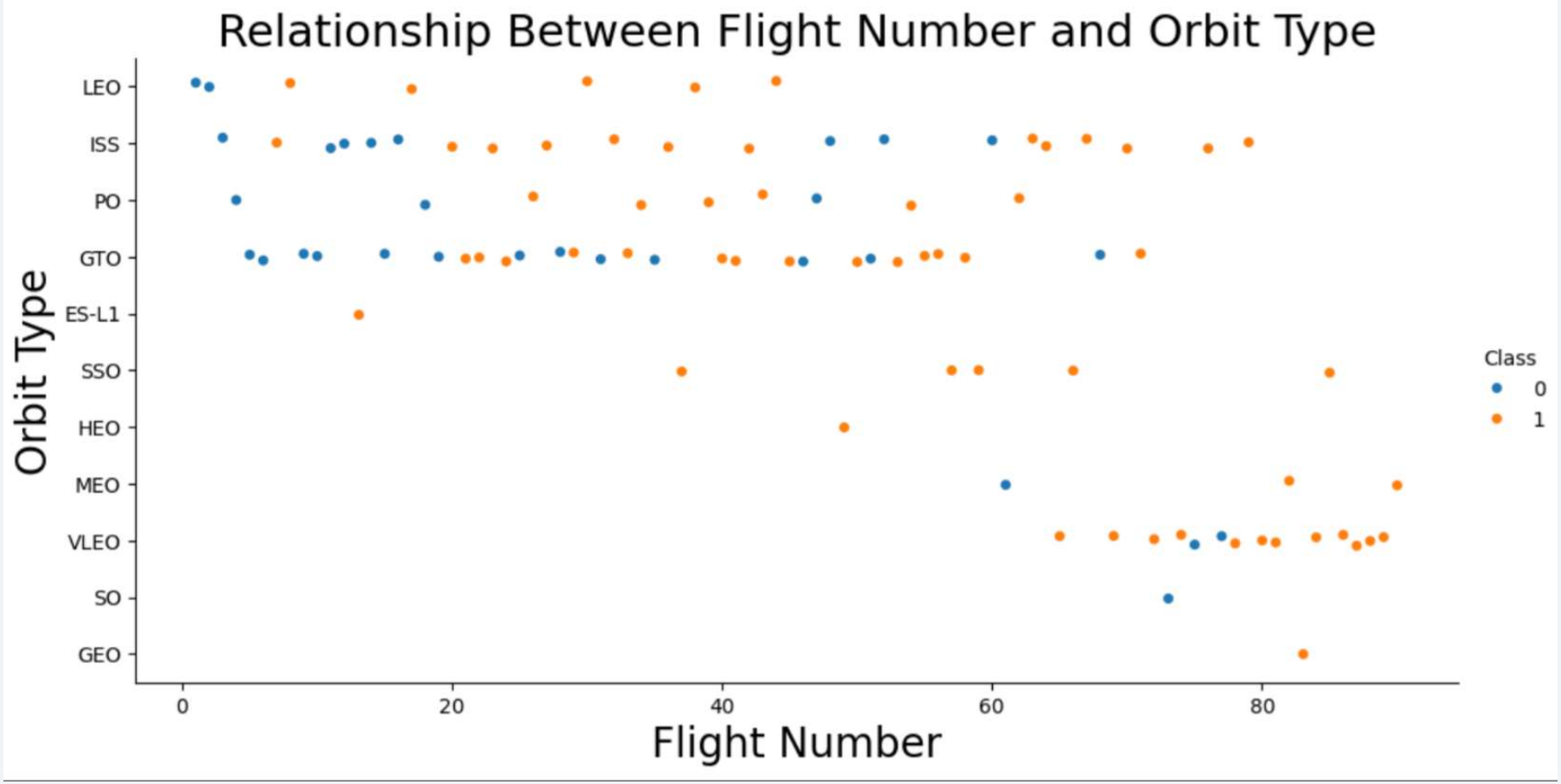
- VAFB SLC-4E launched no heavy payloads (>10,000 kg), while success rates increase for heavy payloads at CCAFS SLC-40 and KSC LC-39A.

Success Rate vs. Orbit Type



- High success rates for ES-L1, GEO, HEO, and SSO, though these orbits are rare.

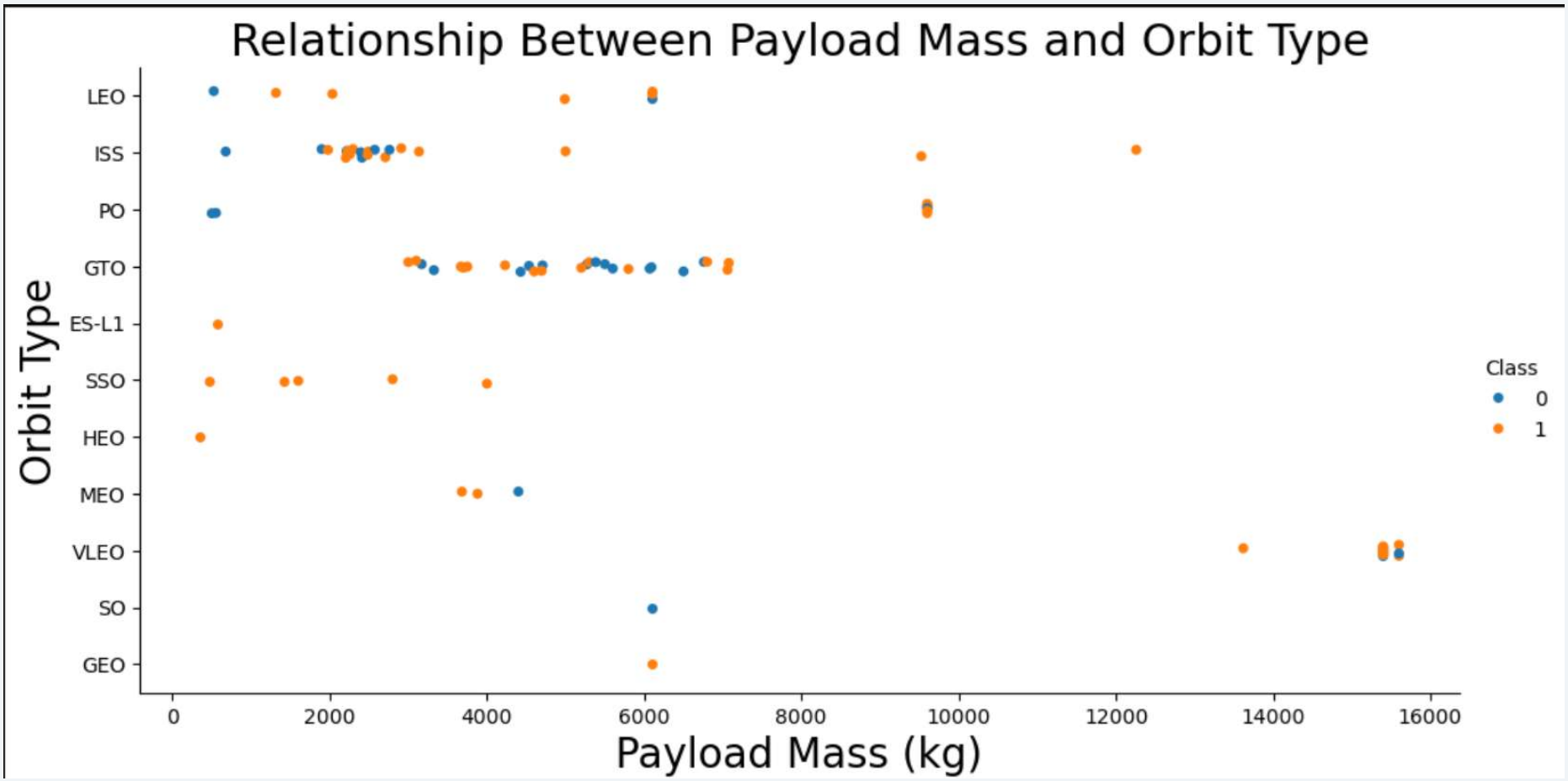
Flight Number vs. Orbit Type



Flight Number vs. Orbit Type

- In the LEO orbit, success seems to be related to the number of flights. Conversely, in the GTO orbit, there appears to be no relationship between flight number and success.

Payload vs. Orbit Type

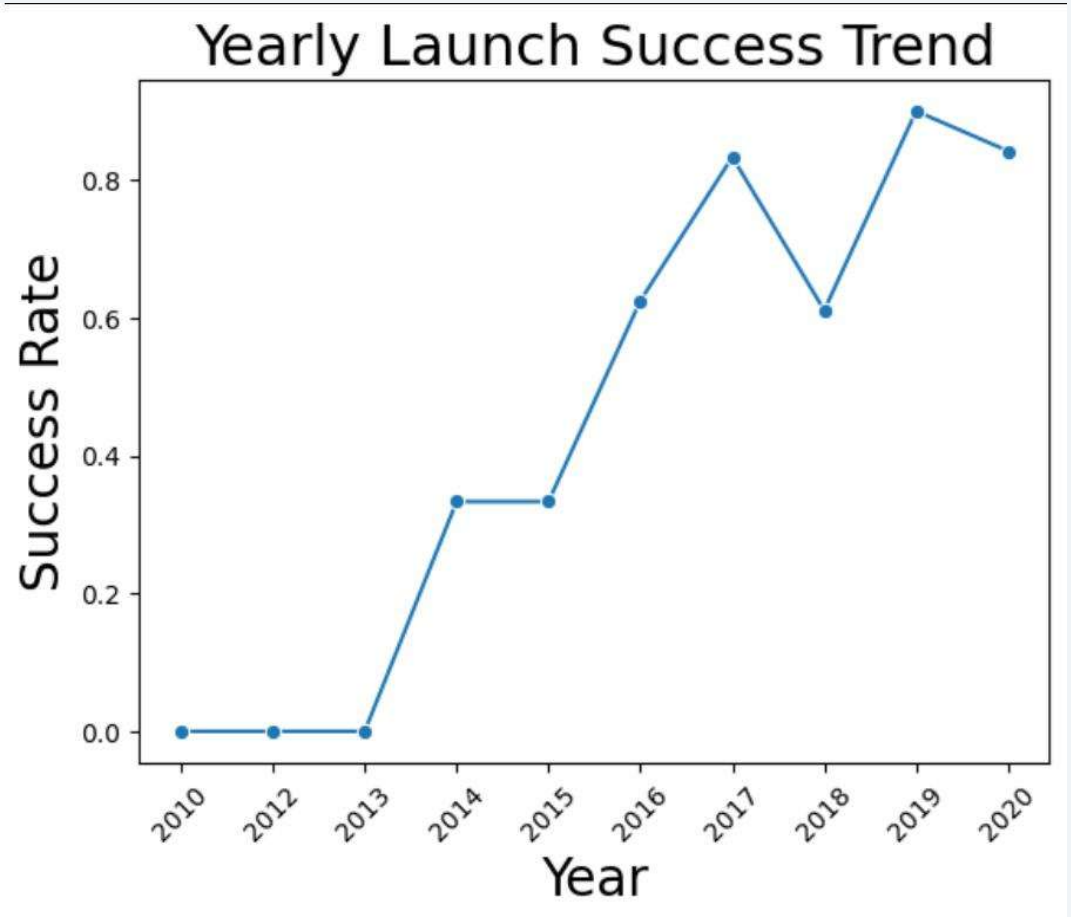


Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

- However, for GTO, it's difficult to distinguish between successful and unsuccessful landings as both outcomes are present.

Launch Success Yearly Trend



- Success rates steadily improved from 2013 to 2020.

All Launch Site Names

Launch Site
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40

- The query reveals the unique SpaceX launch sites present in the dataset: **CCAFS LC-40**, **VAFB SLC-4E** and **CCAFS SLC-40**.
- **Insight**: SpaceX operates multiple launch sites, strategically distributed to support various mission requirements based on geography and payload specifications.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	Fa9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Launch Site Names Begin with 'CCA'

- The data shows five SpaceX launches from **2010 to 2013** at **CCAFS LC-40**, with the following insights:
  - **Mission Outcomes**: All missions were successful despite challenges in landing technology.
  - **Landing Outcomes**: Early missions had either parachute failures or no landing attempts.
  - **Payload Trends**: Initial payloads were either light or test payloads, reflecting early development stages.
  - **Orbit**: Most missions targeted Low Earth Orbit (LEO) for research and resupply purposes.

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
45596

- **Query Result**: The total payload mass carried by boosters for NASA (CRS) missions is **45,596 kg**.
- This query highlights SpaceX's significant contribution to NASA's Commercial Resupply Services (CRS), reflecting the boosters' ability to handle substantial payloads for resupplying the International Space Station and conducting experiments in Low Earth Orbit (LEO).

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2928.4

- **Query Result**: The average payload mass carried by the **F9 v1.1** booster version is **2,928.4 kg**.
- This result showcases the payload handling capacity of the Falcon 9 v1.1 booster. It reflects its reliability in medium payload missions, likely designed to optimize costs while maintaining efficiency for Low Earth Orbit (LEO) and other target orbits.

First Successful Ground Landing Date

MIN(Date)
2015-12-22

- **Query Result**: The first successful ground landing occurred on **2015-12-22**.
- This marks a significant milestone for SpaceX as it demonstrates the company's ability to successfully recover and reuse Falcon 9 boosters by achieving a precise landing on a ground pad, reducing launch costs and advancing spaceflight reusability.

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

- These results highlight the capability of Falcon 9 boosters to handle medium-heavy payloads while achieving successful drone ship landings, showcasing SpaceX's progress in reusability and precision landings under challenging conditions.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	Total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- The vast majority of SpaceX missions have been successful, demonstrating the company's operational reliability. The single unclear payload status indicates an anomaly, while only one mission failure in flight highlights SpaceX's commitment to mission success and safety improvements over time.

Boosters Carried Maximum Payload

Booster_Version
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

- The Falcon 9 Block 5 (F9 B5) series boosters are capable of handling the heaviest payloads in SpaceX's missions. This highlights their superior engineering and efficiency in supporting high-mass missions, further advancing SpaceX's goal of reusability and cost reduction.

2015 Launch Records

Month_Name	Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- These results indicate two failed drone ship landings in 2015. Both occurred at **CCAFS LC-40** using the Falcon 9 v1.1 booster version, highlighting early challenges in achieving successful drone ship landings before SpaceX perfected its landing technology.

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

Landing_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

- The most frequent landing outcome was **No attempt** (10), followed by **Success (drone ship)** and **Failure (drone ship)**, each with 5 occurrences. This highlights SpaceX's increasing attempts at drone ship landings, with continued challenges in achieving successful recoveries.

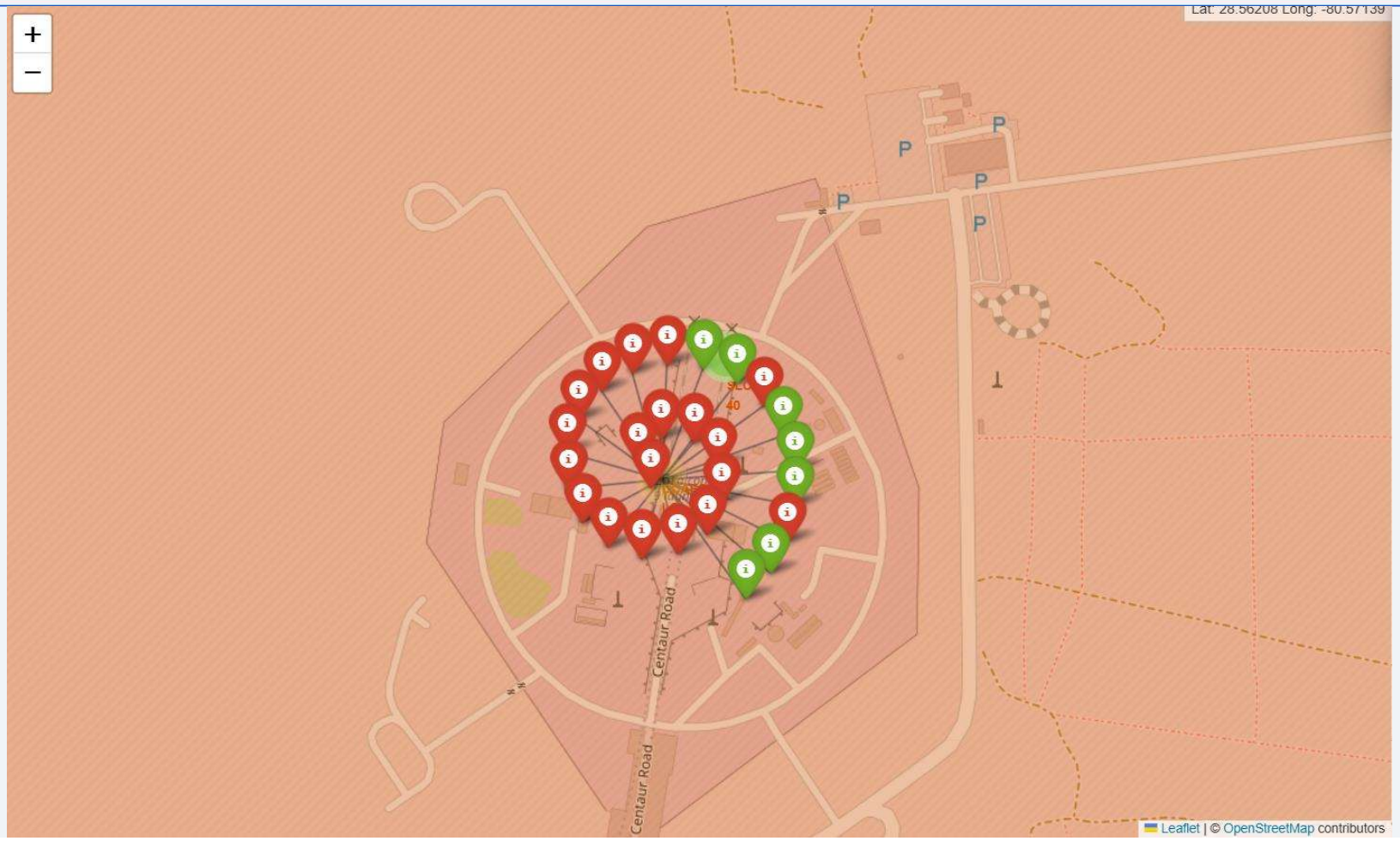

Section 3
Launch Sites
Proximities Analysis

Global Map of SpaceX Launch Sites

Global Map of SpaceX Launch Sites

- **Screenshot Explanation**:
  - The **map** shows the locations of all **SpaceX launch sites** globally.
  - Each site is represented by a **marker**, and the color coding or icons can indicate various **landing outcomes** or mission statuses.
- **Key Findings**:
  - Launch sites are spread across multiple geographies, including the U.S. East and West coasts, reflecting SpaceX's global operational reach.
  - The map helps visualize proximity to other critical infrastructure (e.g., coastlines).

SpaceX Launch Outcomes by Location



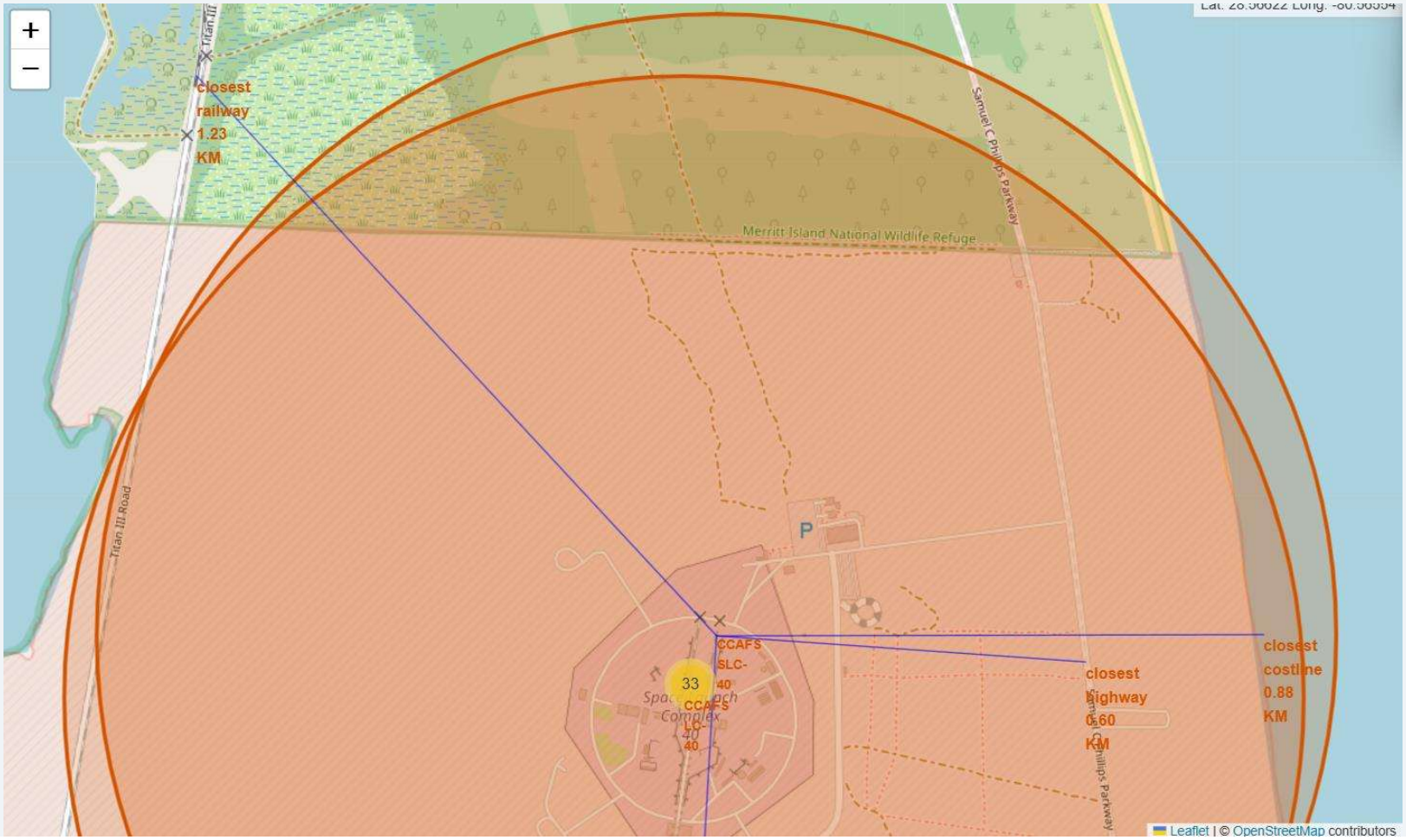
SpaceX Launch Outcomes by Location

- The **map** displays **SpaceX launch sites** with markers color-coded based on **landing outcomes**.

- **Color Legend**: Green for success and red for failure

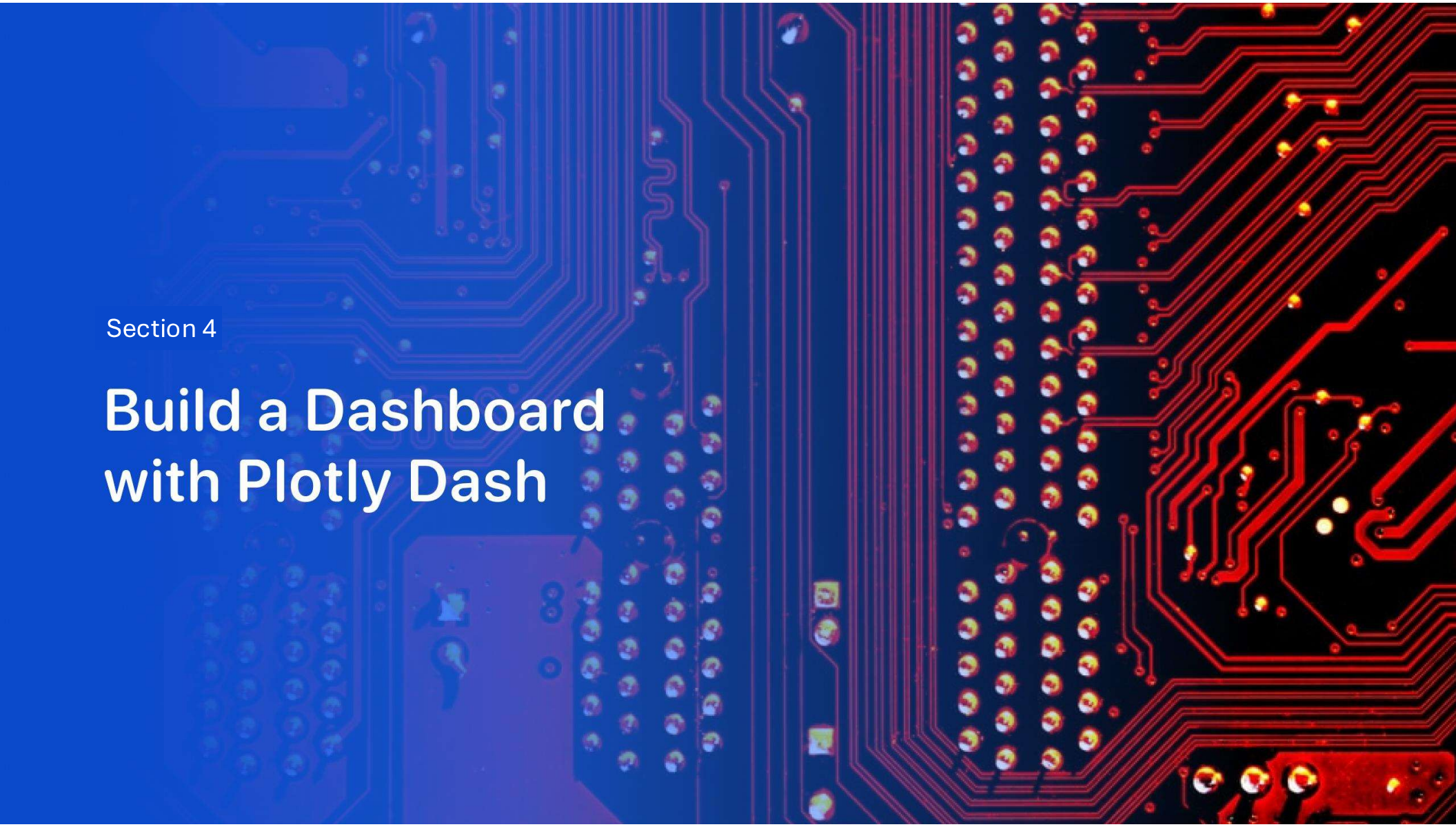
- **Key Findings**:
  - Successful landings are concentrated at specific sites.
  - This visualization provides insights into the **performance trends** across launch sites globally.

Proximity Analysis of SpaceX Launch Site



Proximity Analysis of SpaceX Launch Site

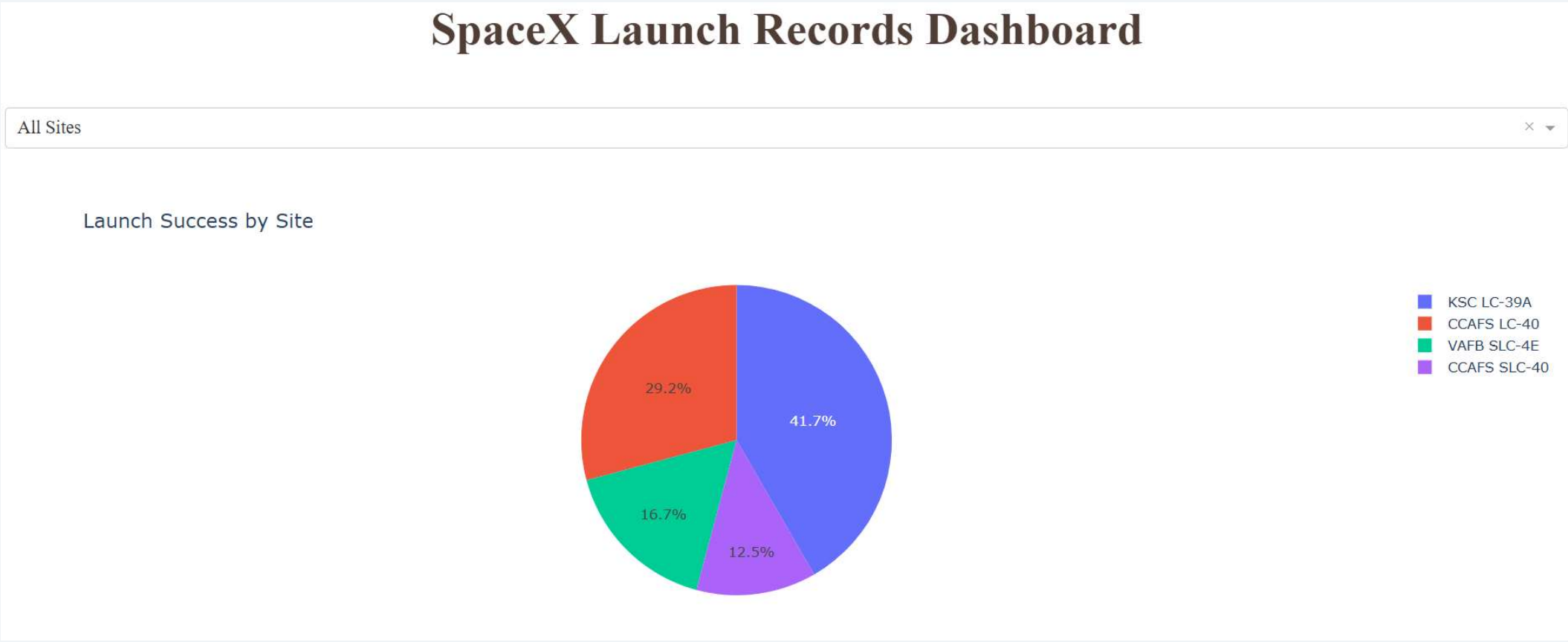
- **Screenshot Explanation**:
  - The map shows the **selected launch site** along with proximity markers for **railways**, **highways**, and **coastlines**.
  - **Distance Calculations**: The distances from the launch site to these nearby infrastructure elements are displayed for context.
- **Key Findings**:
  - The map illustrates how **geographical factors** like nearby transportation routes and coastlines influence launch site accessibility and safety.
  - This proximity data helps assess potential risks and logistical support for launches.



Section 4

Build a Dashboard with Plotly Dash

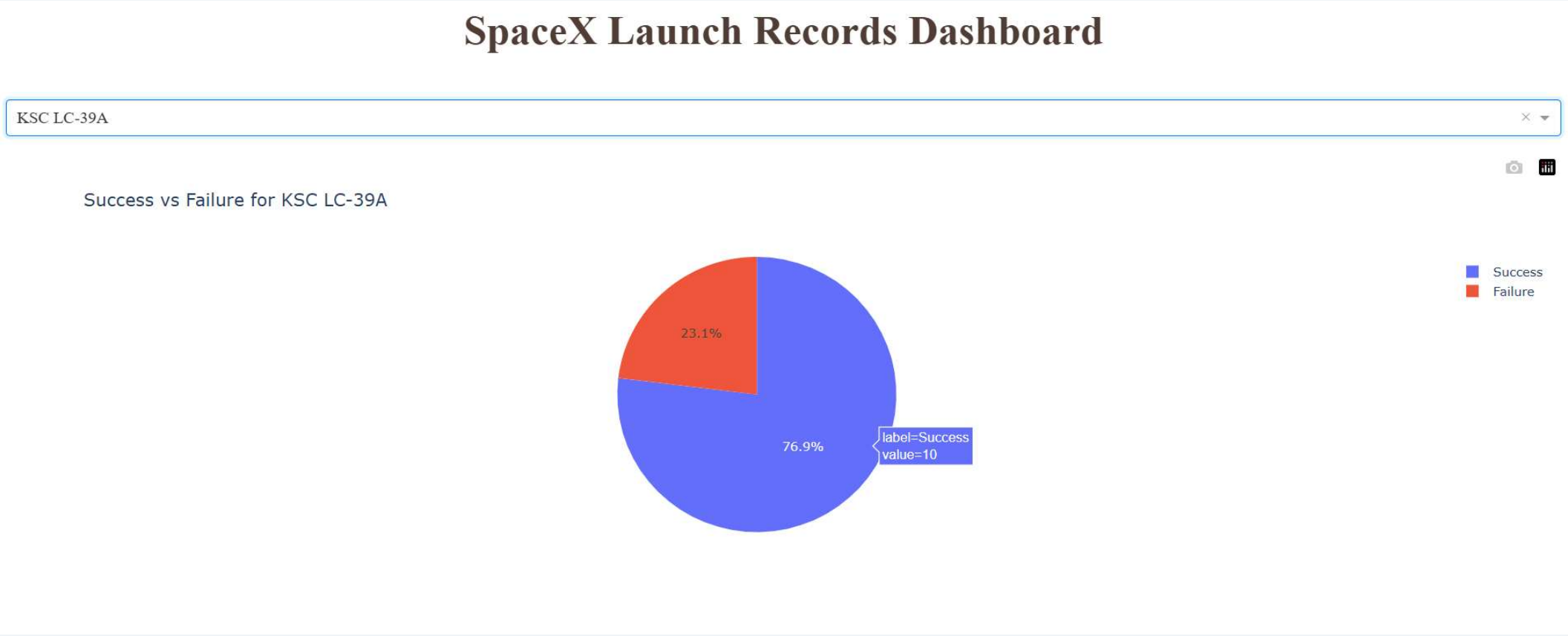
SpaceX Launch Success Distribution by Site (Pie Chart)



SpaceX Launch Success Distribution by Site (Pie Chart)

- **Screenshot Explanation**:
  - The **pie chart** shows the **success** distribution of launches across all SpaceX sites.
- **Key Elements**:
  - The **majority of launches** are successful, illustrating SpaceX’s operational efficiency across different locations.
  - This chart offers a clear visual representation of SpaceX's high success rate, further reinforcing their reputation for reliable rocket launches.

KSC LC-39A Launch Success Ratio



KSC LC-39A Launch Success Ratio

- The **pie chart** showcases the **launch success ratio** specifically for the **KSC LC-39A launch site**, highlighting the percentage of successful launches compared to failures.

- **Key Findings**:
  - **KSC LC-39A** has a **very high success rate**, indicating the site's effectiveness in supporting missions.
  - This chart emphasizes KSC LC-39A's pivotal role in SpaceX's operations and its critical contribution to the overall success of SpaceX's launch programs.

Payload vs. Launch Outcome for All Sites

Payload vs. Launch Outcome for All Sites

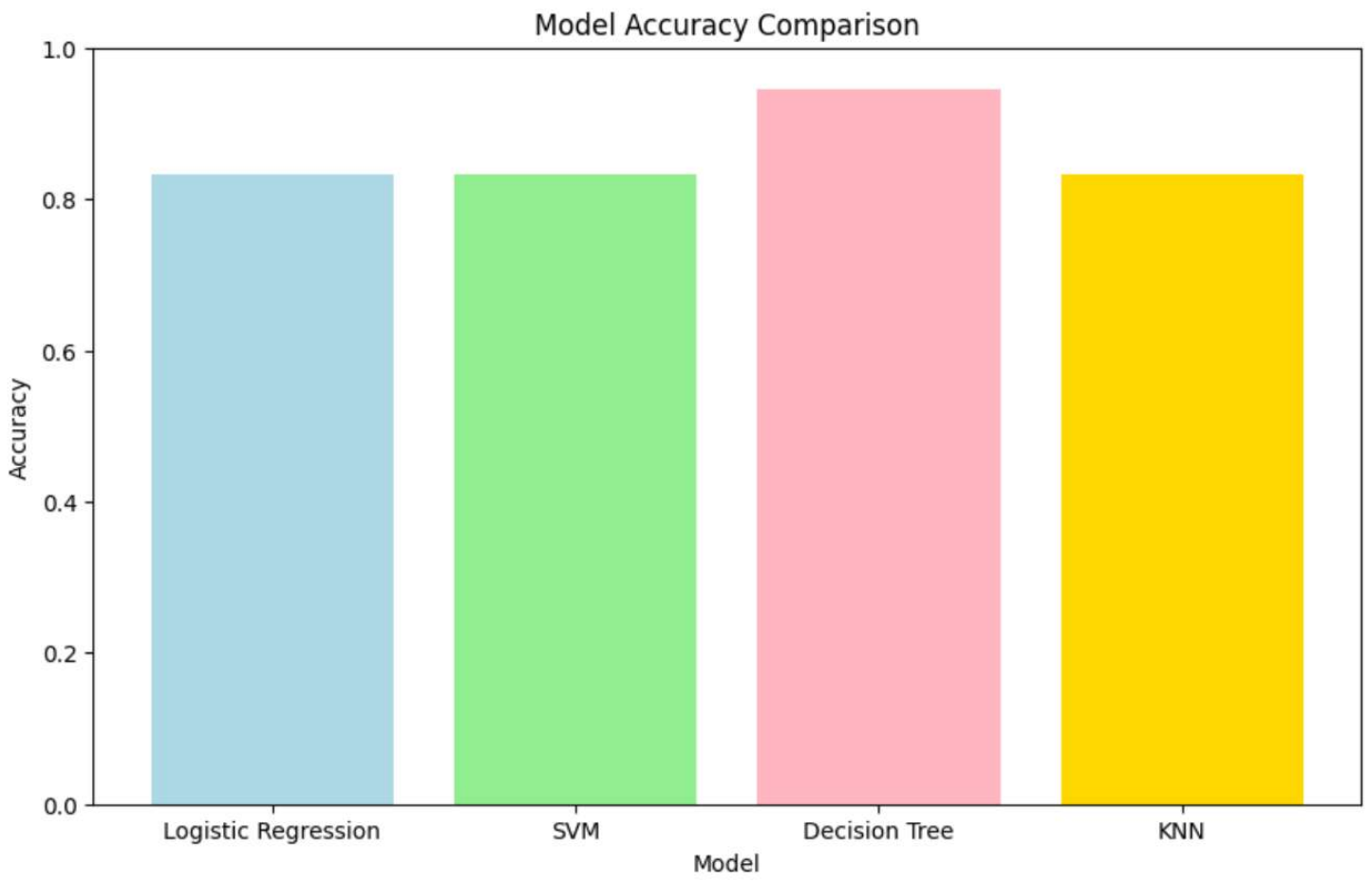
- The **scatter plot** visualizes the relationship between **payload mass** and **launch outcomes** across all SpaceX launch sites, with varying payload ranges selected using the **range slider**.

- **Key Findings**:
  - **F9 v1.1 boosters** tend to have a higher failure rate.
  - **FT boosters** show high success rates for payloads between **2000-6000 kg**.
  - **B4 boosters** perform best with payloads between **3000-5000 kg**, showing consistent success.
  - This highlights the correlation between booster versions, payload sizes, and success rates.

Section 5

Predictive Analysis (Classification)

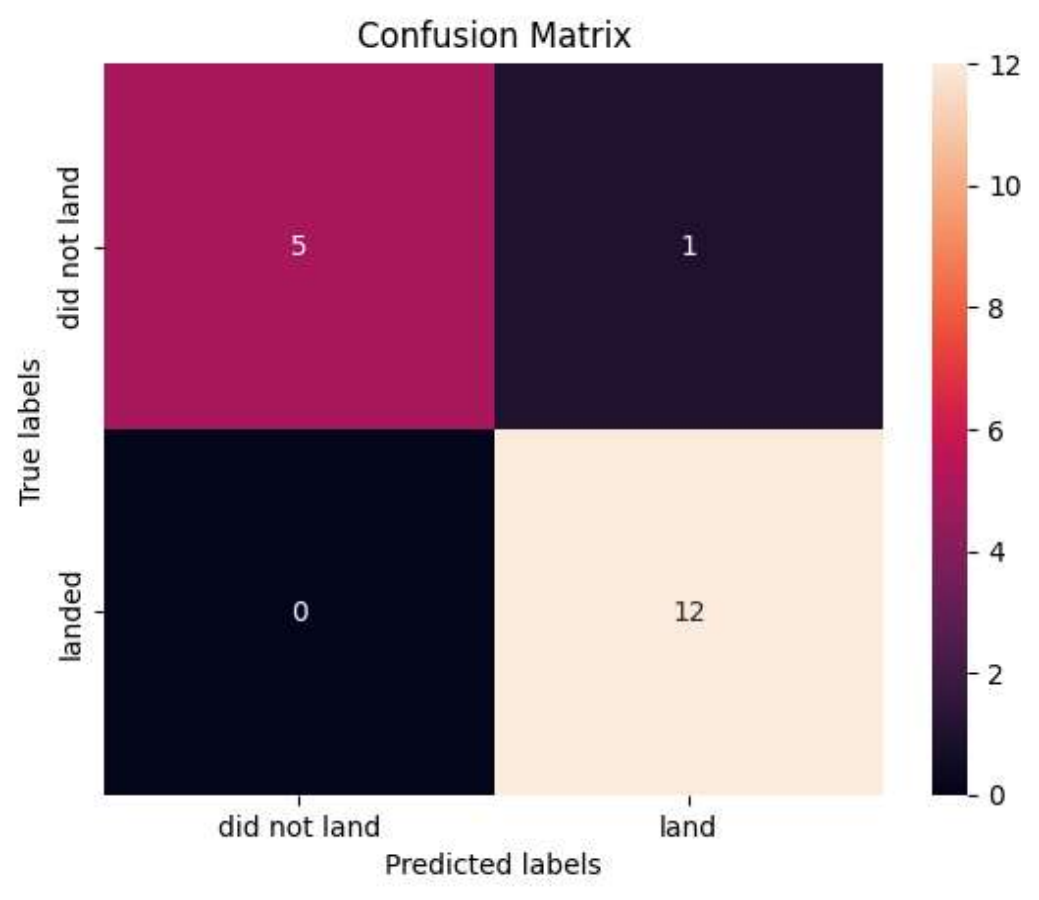
Classification Accuracy



- **Best Model**: Decision Tree Classifier with accuracy **0.9444** on the test data.

(see appendix A for features)

Confusion Matrix



- The confusion matrix shows the count of actual vs. predicted values for each class: **"Did not land" (0)** and **"Landed" (1)**.
- **True Positives (TP):** Correct predictions of "Landed".
- **True Negatives (TN):** Correct predictions of "Did not land".
- **False Positives (FP):** Misclassified "Did not land" as "Landed".
- **False Negatives (FN):** Misclassified "Landed" as "Did not land".

Conclusions

- The project highlights that **payload mass**, **booster version**, and **orbit type** are highly correlated with rocket landing outcomes.

- The **KSC LC-39A** site shows the highest success rate.

- Orbit types like **LEO** (Low Earth Orbit) and **ISS** (International Space Station) correlate with higher success rates because they are typically closer to Earth, requiring less energy and offering more control over the rocket's trajectory, making landings more predictable. **GTO** (Geostationary Transfer Orbit) has more variability in landing outcomes because it involves higher altitudes and more complex trajectories, requiring more energy and precision, which introduces greater challenges for landing.

Conclusions

- Successful landings are more likely for **payloads between 2000-6000 kg** and for **F9 FT** boosters.

- Launch sites near **coastlines (**reducing risks to populated areas), with access to **railways** and **highways** (efficient transportation of equipment and personnel), are ideal for success.

Appendix

A

Features for tree classifier :

- **{'criterion': 'entropy', 'max_depth': 8,'max_features': 'sqrt','min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'random'}**

Thank you!