IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arthur Chong
05.31.2025

Outline

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

Executive Summary

- Summary of methodologies
  - Data Preprocessing & Cleaning
  - Exploratory Data Analysis (EDA)
  - Feature Engineering
  - Model Building
  - Model Evaluation

- Summary of All Results
  - Best Launch Site: KSC LC-39A has the most successful landings.
  - Payload Insights:
    - Heavier payloads were more successful in Polar, LEO, and ISS.
  - Trend Over Time:
    - Success rates have increased since 2013.
  - Best Performing Model:
    - Decision Tree Classifier achieved the highest accuracy of 0.89.

Introduction

SpaceX advertises the Falcon 9 rocket with a launch cost of **$62 million**, significantly lower than other providers whose costs can exceed **$165 million**.

This cost advantage is due to SpaceX's ability to reuse the rocket's first stage, reducing manufacturing and operational expenses.

- In this capstone, we aim to predict whether the Falcon 9's first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scrapping from Wikipedia
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Overview
  - **Source**: Falcon 9 launch records from Wikipedia HTML table.
  - **Tool**: Scrapped using **Beautiful Soup**, a Python library for parsing HTML content.

- Steps in Pipeline:
  1. Web scrapping
  2. HTML parsing
  3. Data Extraction
  4. Data Frame Conversion

Data Collection – SpaceX API

- **REST API Calls with Python**
  - Used requests.get() to send HTTP get requests and retrieve responses.
- **Parsing JSON Data**
  - Loaded JSON into Python dictionaries using response.json()
- **Data Normalization**
  - Flattened nested fields using json-normalize or pd.json_normalize
- **GitHub**
  - https://github.com/kchong99/SpaceX_Falcon9/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Send GET Request via Python (requests)

Receive JSON Response

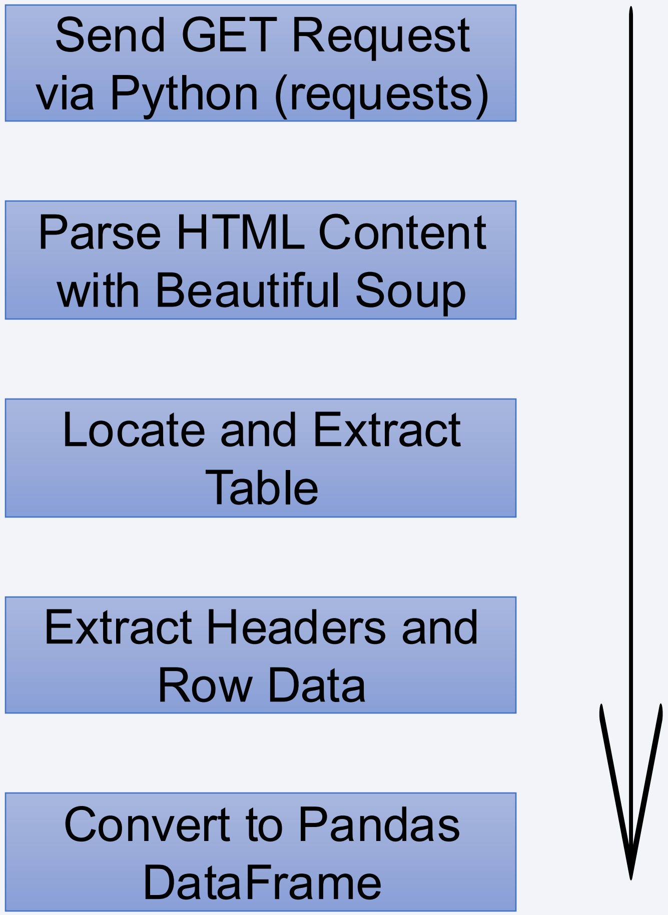
Extract & Normalize JSON Data

Convert to Pandas Data Frame

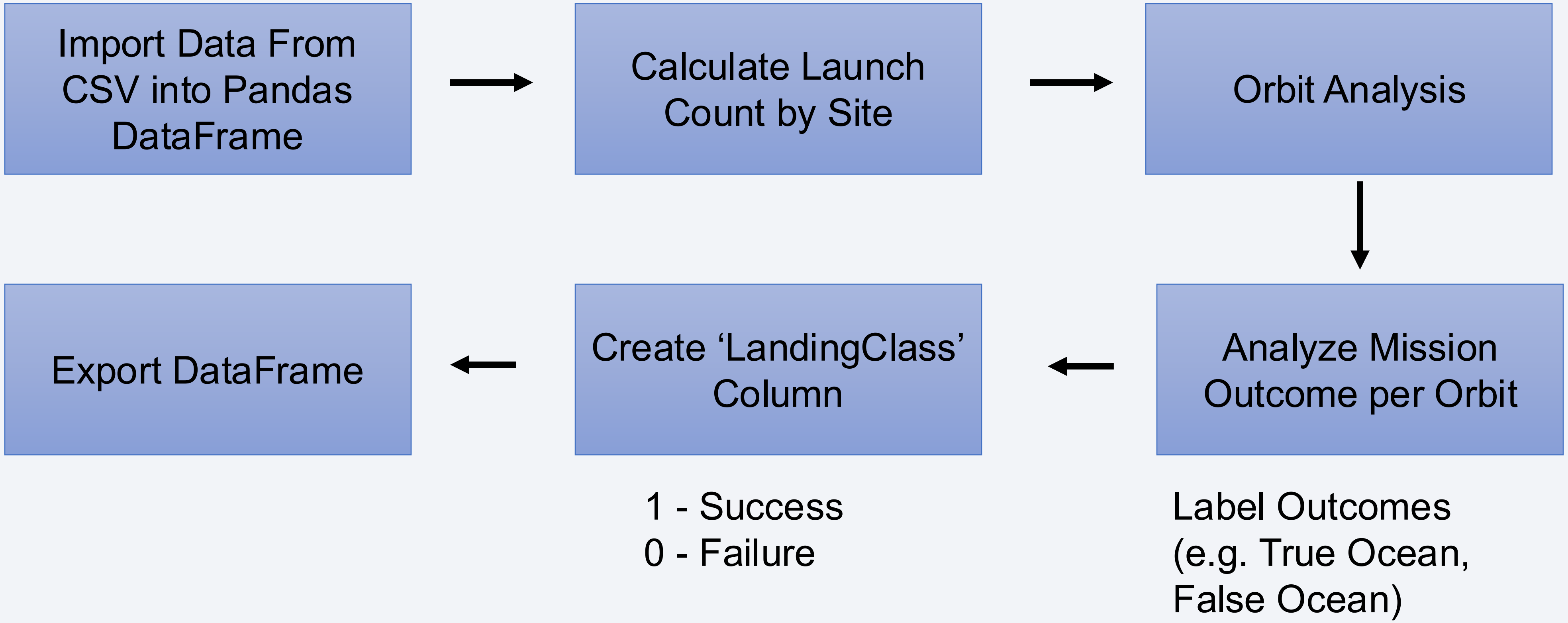
Use Data for Analysis

Data Collection - Scraping

- **Target URL Identification**
  - Located the Wikipedia page containing Falcon 9 launch records in table format.
- **HTML Parsing**
  - Employed BeautifulSoup to parse HTML and navigate through the page structure.
- **Table Extraction**
  - Located the relevant <table> tag containing the data using tag and class names.
- **Github**
  - https://github.com/kchong99/SpaceX_Falcon9/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling



GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- **Scatter Plots**
  Used to show relationships between two continuous or categorical variables.
  - Flight No. vs Launch Site
  - Flight No. vs Payload Mass

- **Bar Chart**
  For comparing categorical data
  - Success Rate by Orbit

- **Line Chart**
  Ideal for showing trends over time.
  - Yearly Launches by Site

GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/edadataviz.ipynb

EDA with SQL

- Selected **distinct launch sites** from the SpaceXTable
- Used **LIKE** to filter launch sites and compute total payload mass
- Calculated **average payload mass** for **Booster Version F9 v1.1** using AVG()
- Retrieved **date of first successful landing** using MIN() on landing success data
- Filtered **booster names** with:
  - Payload mass between **4000 and 6000 kg**
  - **Successful drone ship landings**
- Counted **total number of successful and failed landings** using COUNT()
- Used a **subquery** to find all **booster versions that carried the maximum payload mass**
- Applied SUBSTR(Date, ...) to extract **month names of failed landings**

GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- **folium.Circle**
  - Highlights a specific coordinate area with a visible radius
  - Used to mark the coverage area around launch sites

- **folium.Marker**
  - Marks specific launch site or location with a descriptive label
  - Helps pinpoint important coordinates on the map

- **MarkerCluster()**
  - Groups nearby markers to avoid clutter
  - Improves readability on maps with multiple markers at or near the same location

- **folium.PolyLine**
  - Draws lines between key coordinates, such as between a launch site and the coastline
  - Helps visualize spatial relationships and distance

GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- **Dropdown Menu**: Allows users to select a specific SpaceX launch site or view data for all sites.

- **Pie Chart**
  - Displays the **total number of successful launches** for each site when "All Sites" is selected.
  - Displays the **success vs. failure breakdown** for a selected individual site.

- **Range Slider**: Enables users to filter data by payload mass (in kilograms).

- **Scatter Plot**
  - Shows the **correlation between payload mass and launch outcome**.

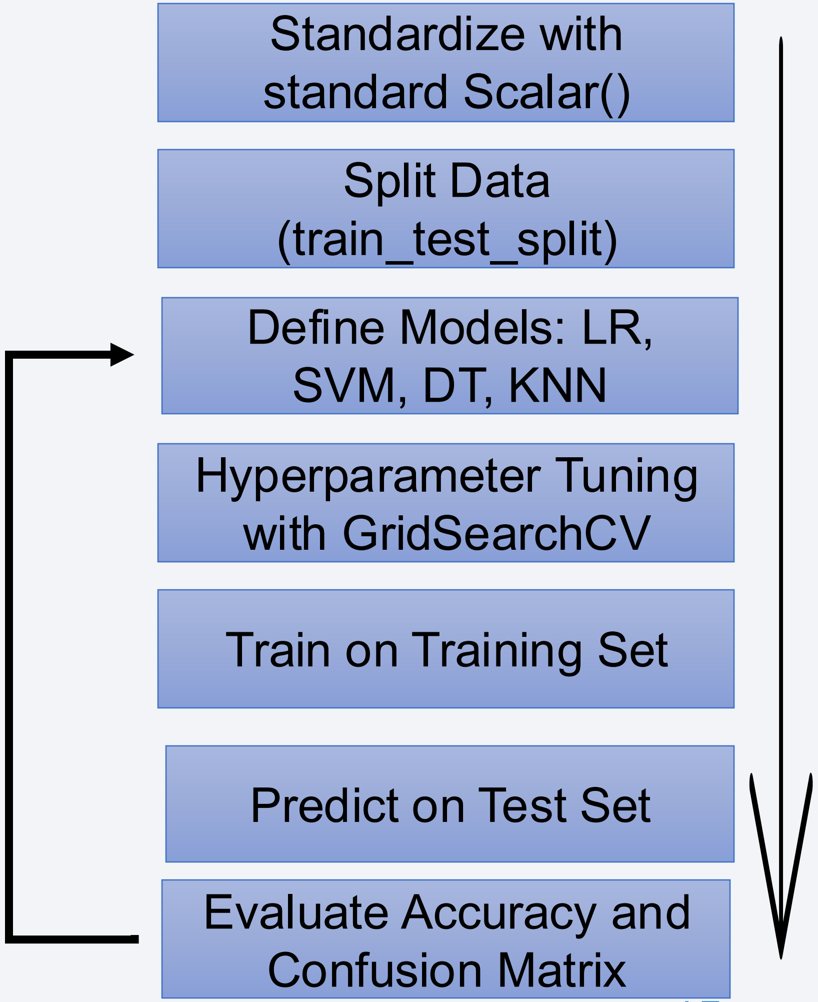
GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/spacex-dash-app.py

Predictive Analysis (Classification)

- Selected multiple classification models:
  - **Logistic Regression**
  - **Support Vector Machine (SVM)**
  - **Decision Tree Classifier**
  - **K-Nearest Neighbors (KNN)**

- Used **GridSearchCV** to find the **best hyperparameters** for each model.

- **Model Evaluation**
  - **Accuracy score**
  - **Confusion Matrix**



GitHub: https://github.com/kchong99/SpaceX_Falcon9/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Success Rate by Orbit Type
  - Orbit Type vs. Flight Number
  - Payload vs. Orbit Type
  - Yearly Success Rate

- Interactive analytics results
  - Pie Charts
  - Scatter Plot Dashboard
  - Dash App

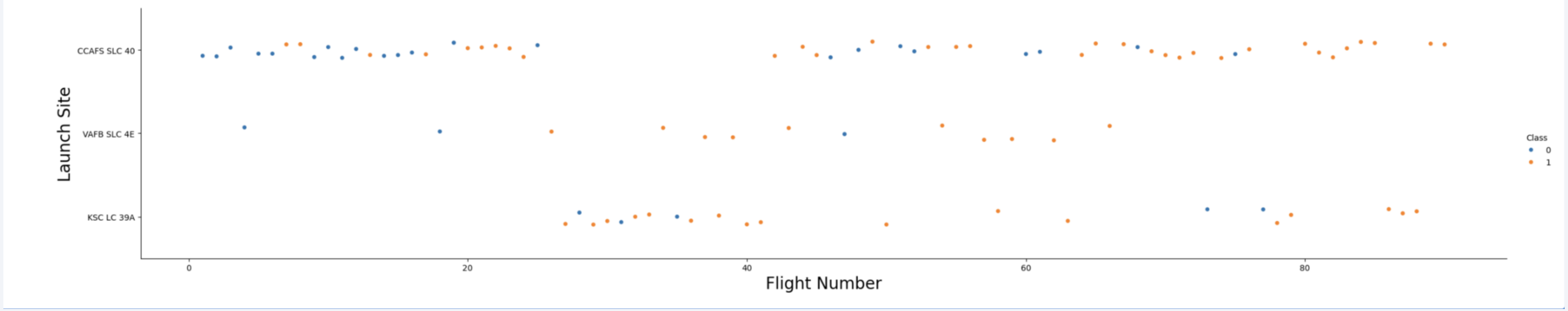
- Predictive analysis results
  - Models Trained
  - Training Process
  - Model Performance Comparison
  - Best Model Evaluation

Section 2

Insights drawn from EDA

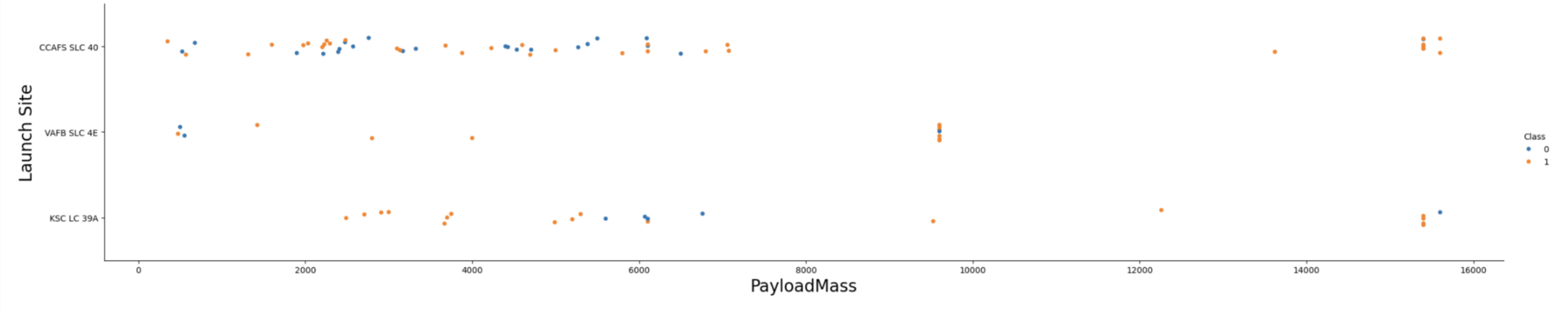
Flight Number vs. Launch Site

- **CCAFS SLC-40**:
  - Has the **highest number of flights** (most data points along x-axis).
  - Shows **more successful landings** (colored points with Class = 1).
- **KSC LC-39A**:
  - Has **fewer flight numbers**, indicating less activity from this site.



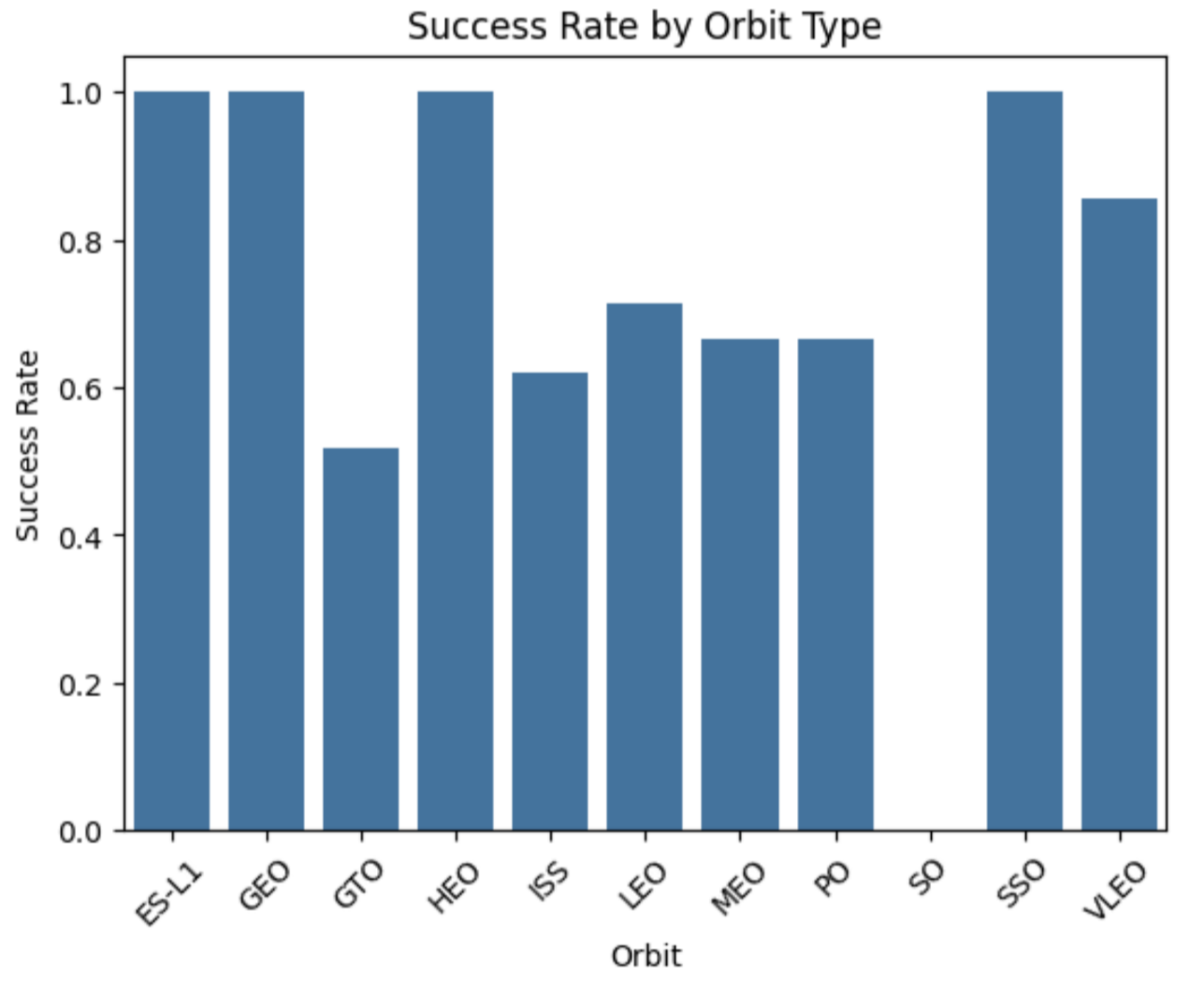
Payload vs. Launch Site

- **CCAFS SLC-40**:
  - Achieved **the most successful landings at high payloads**, even **beyond 14,000 kg**.
- **KSC LC-39A**:
  - Had **a failed landing** at payloads over **14,000 kg.**
- **VAFB SLC-4E**:
  - Its **maximum payload mass is < 10,000 kg.**



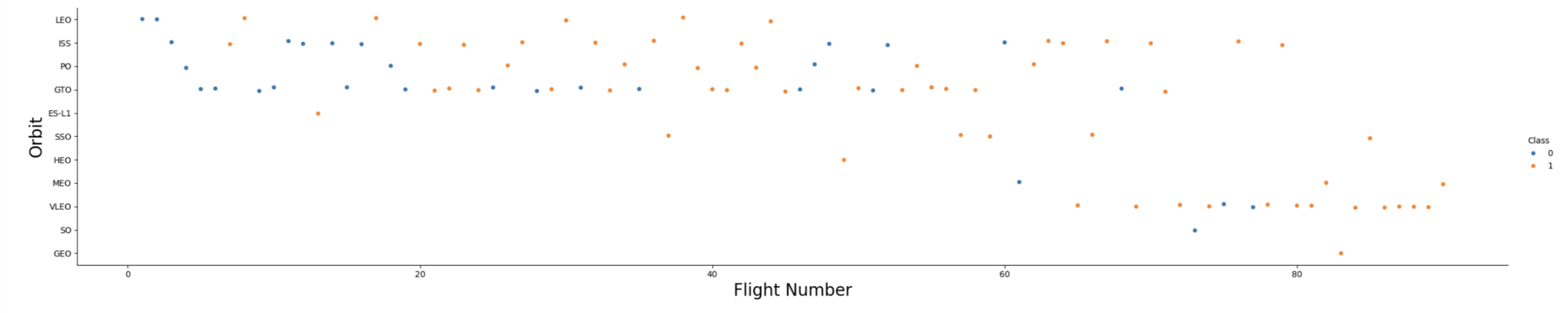
Success Rate vs. Orbit Type

- **Highest Success Rates**:
  - **ES-L1, GEO, HEO, and SSO** all have **perfect or near-perfect launch success rates (≈ 1.0)**.
- **Moderate Success Rate**:
  - **GTO (Geostationary Transfer Orbit)** has a **success rate around 0.5**
- **Lowest Success Rate**:
  - **SO (Sub-Orbital)** has a **success rate of 0**, indicating **failures or aborted missions** for this specific orbit.



Flight Number vs. Orbit Type

- **LEO (Low Earth Orbit)**:
  - Shows a **clear trend of increasing success rate** with higher flight numbers.
- **GTO (Geostationary Transfer Orbit)**:
  - No clear trend or pattern between experience and success.
- **GEO (Geostationary Orbit)**:
  - Only **1 successful launch** was observed.
- **SO (Sub-Orbital)**:
  - Shows a **failed landing**, indicating **no success yet** for this orbit type in the dataset.

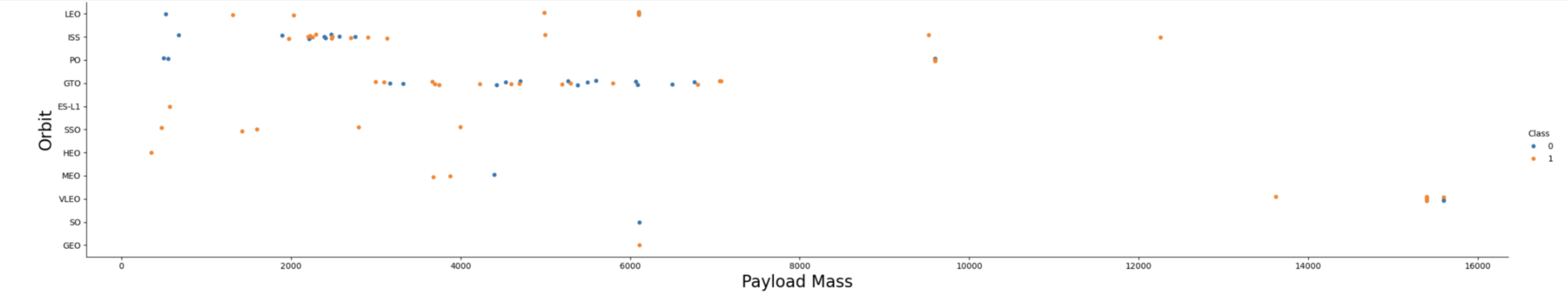


Payload vs. Orbit Type

- **Polar, LEO, and ISS Orbits**:
  - These orbits show **high success rates** even at **heavier payloads**.

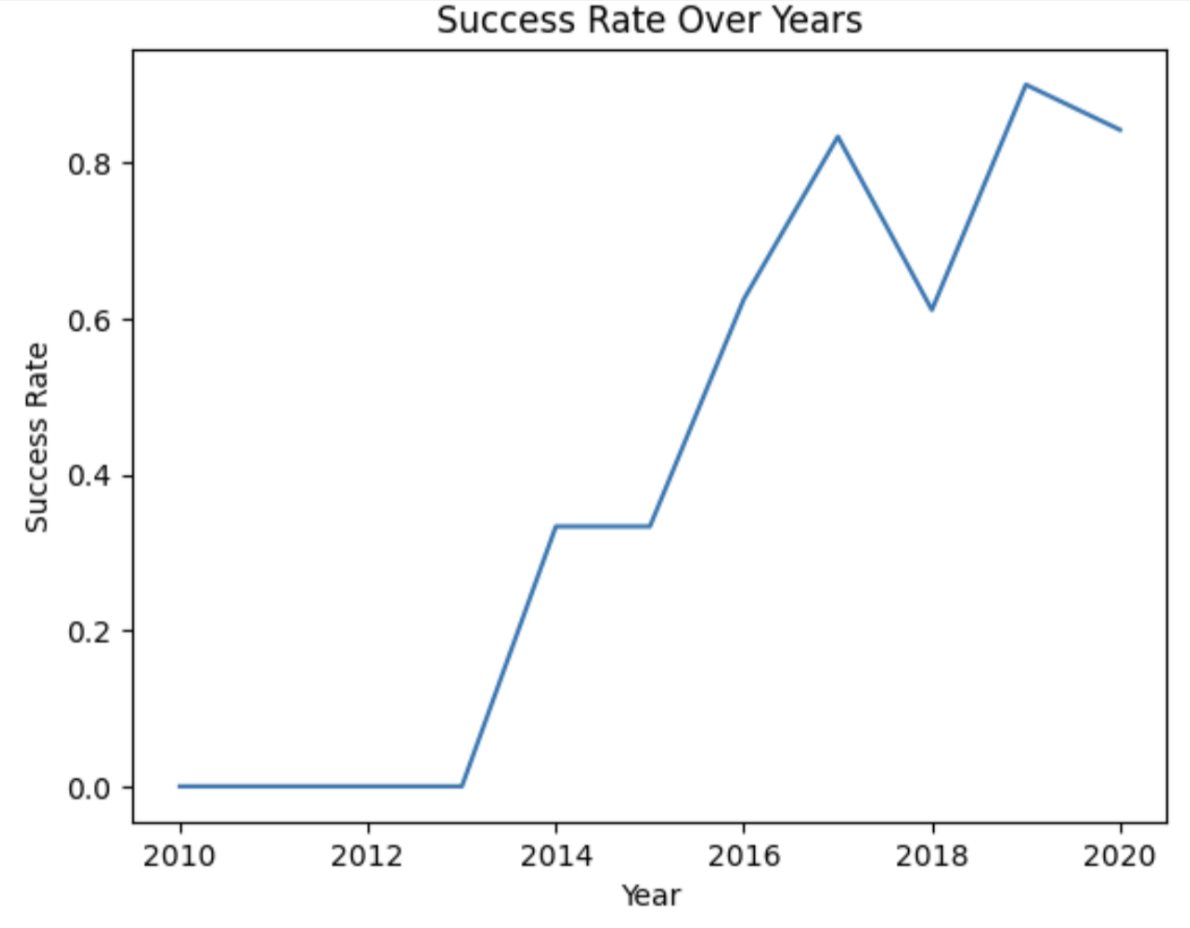
- **GTO (Geostationary Transfer Orbit)**:
  - **Both successful and unsuccessful landings** are observed across various payload masses.

- **SSO (Sun-Synchronous Orbit)**:
  - Shows **nearly 100% success rate**, regardless of payload mass.



Launch Success Yearly Trend

- **2010–2013**:
    - Success rate remains **at 0%**, indicating **no successful landings** during early missions.
- **2013–2020**:
    - General **upward trend in success**
- **2017–2018 and 2019–2020**:
    - Noticeable **dips in success rate**, possibly due to experimental missions or increased mission complexity.



All Launch Site Names

- SQL Query:
  - SELECT DISTINCT Launch_Site FROM SPACEXTABLE;

- This query retrieves the distinct launch site name from the SpaceX Table, eliminating duplicates to show the unique locations.

- Results

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

Launch Site Names Begin with 'CCA'

- The query selects the first 5 rows from the SpaceX Table where the Launch_Site name starts with “CCA”

- All 5 returned entries have the **launch site "CCAFS LC-40"**

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 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

Total Payload Mass

- The total payload mass carried by NASA (CRS) missions is **48,213 kg**.
- The query calculates the total payload mass (in kilograms) delivered by SpaceX booster for missions where the customer is NASA (CRS).

SUM_Payload_Mass
48213

Average Payload Mass by F9 v1.1

- The **average payload mass** carried by **F9 v1.1** boosters is **2,534.67 kg**.
- The query calculates the average payload mass (in kilograms) for launching using F9 v1.1 booster version.

AVG_Payload_Mass
2534.6666666666665

First Successful Ground Landing Date

- The first successful ground pad landing occurred on **December 22, 2015**.
- The query finds the earliest date when a SpaceX booster successfully landed on a ground pad.

Earliest_Successful_Landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The query lists all booster versions that successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000 kg.
  - F9 FT B1022
  - F9 FT B1026
  - F9 FT B1021.2
  - F9 FT B1031.2

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

Total Number of Successful and Failure Mission Outcomes

- The query counts the total number of missions with outcomes classified as **Success** or **Failure**.
- **Results**
  - **Successful missions:** 98
  - **Failure Mission**: 0

Mission_Outcome	Outcome_Count
Success	98

Boosters Carried Maximum Payload

- Name of the booster which have carried the maximum payload mass
  - 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 query identifies the booster versions that carried the **maximum payload mass** recorded in the dataset by comparing each payload mass to the overall maximum.

2015 Launch Records

- The retrieves data from the year **2015** where **drone ship landings failed**, including:
  - Month of launch
  - Booster version used
  - Launch site name

- This records show the two failed drone ship landing attempts in 2015:

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

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

- The query ranks the different landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

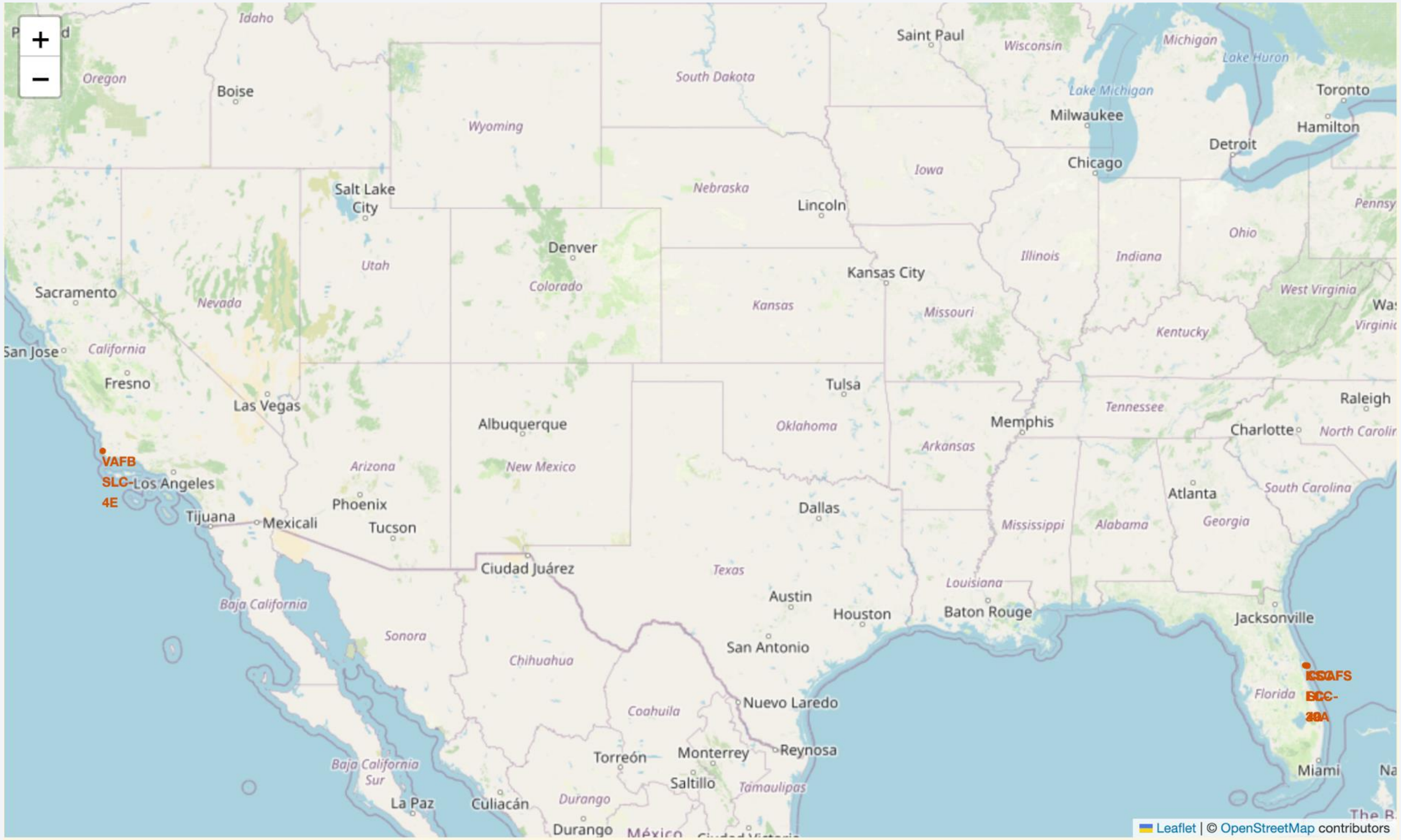
- The results display the most frequent landing outcomes during this period, helping assess **SpaceX's landing success rate trends** over time.

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

Global Distribution of SpaceX Launch Sites

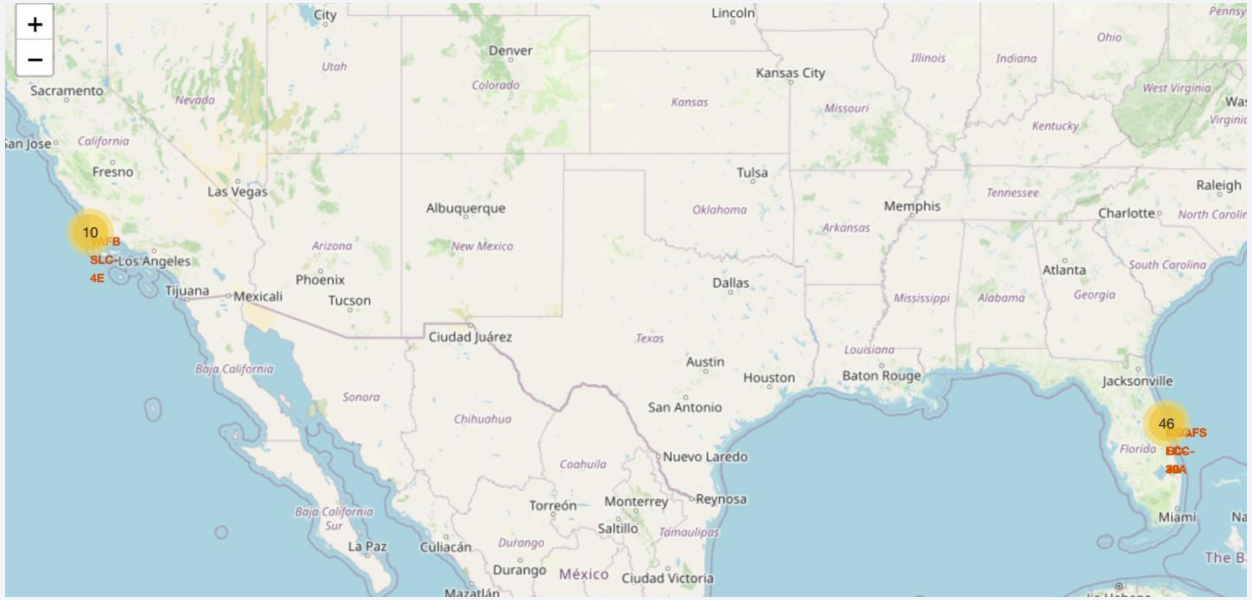


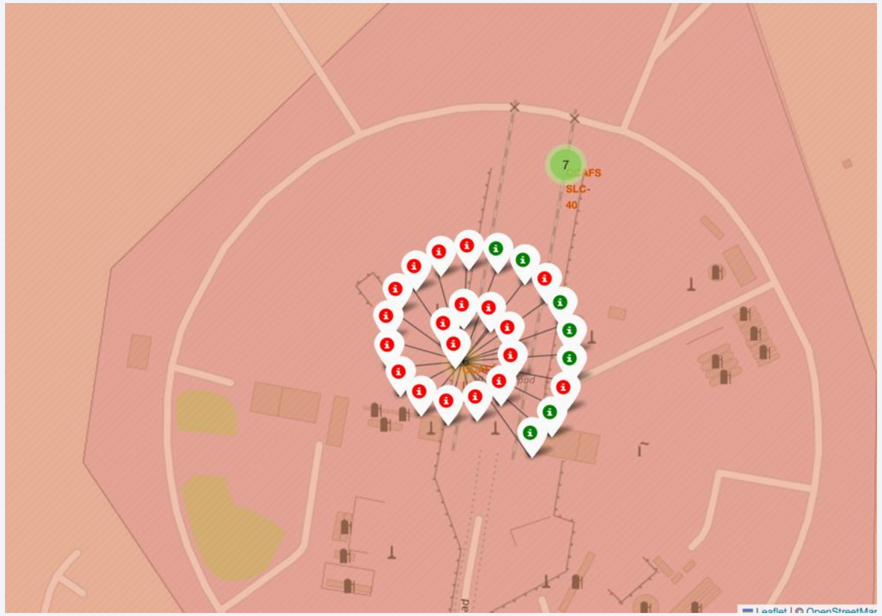
- **Highlighted Sites:**
  - **California:** 1 launch site
  - **Florida:** Multiple launch sites

Color-Coded Launch Outcomes at SpaceX Launch Sites

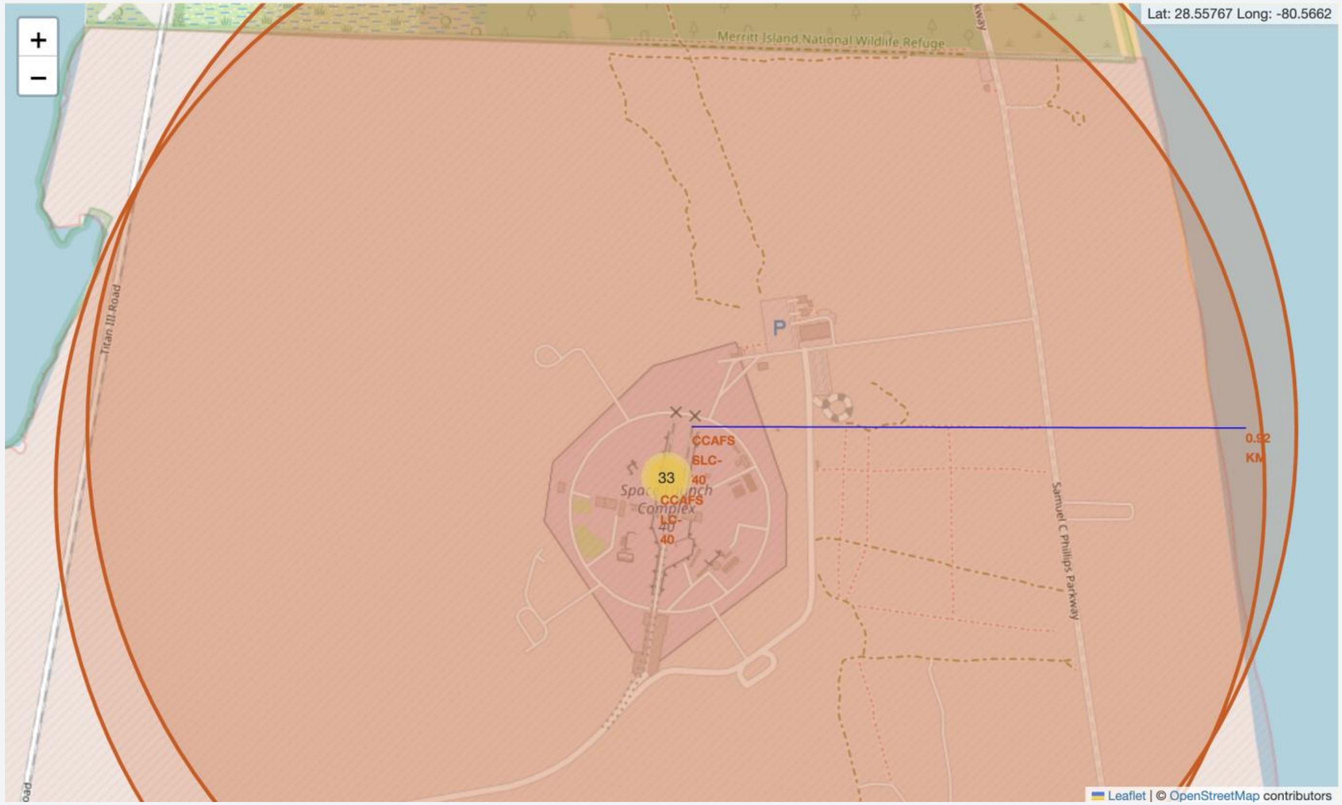
- **The first figure** gives an overview of landing outcomes across all launch sites.

- The **zoomed-in view** (e.g., in Florida) reveals detailed success/failure patterns at each site.
    - **Green markers** indicate **successful landings**
    - **Red markers** indicate **failed landings**





Distance from Launch Site to Nearest Coastline



- This **interactive Folium map** shows how close a SpaceX launch site is to nearby geographic features (e.g., **coastline**, **highway**, and **railway**).

- These visualizations help evaluate how well-positioned each site is with respect to:
  - **Transport access**
  - **Proximity to water**
  - **Rail access**

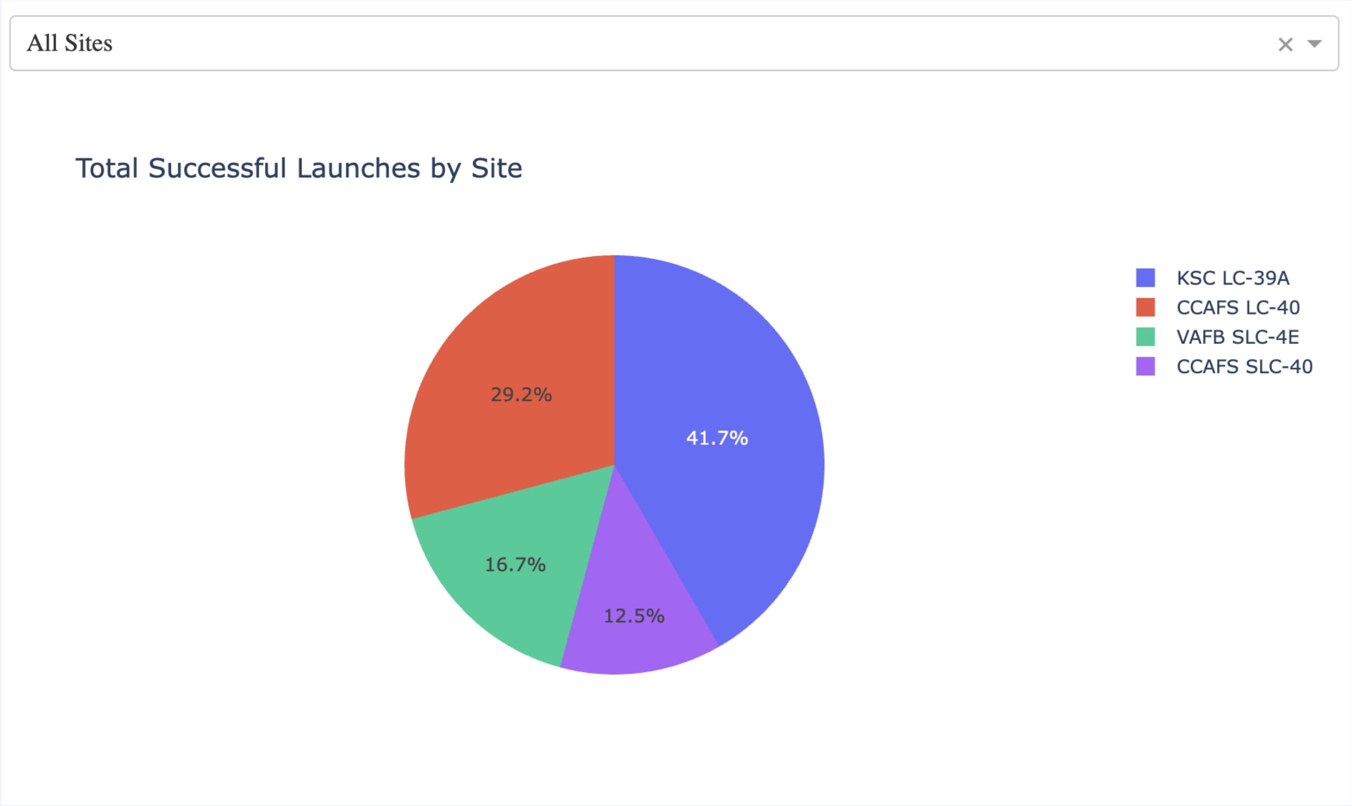
Section 4

Build a Dashboard with Plotly Dash

SpaceX Launch Success Count by Site

- **Pie Chart Overview**:
  - This visualization displays the **distribution of successful launches** across all SpaceX sites.
  - Each **color-coded segment** represents one **launch site**.

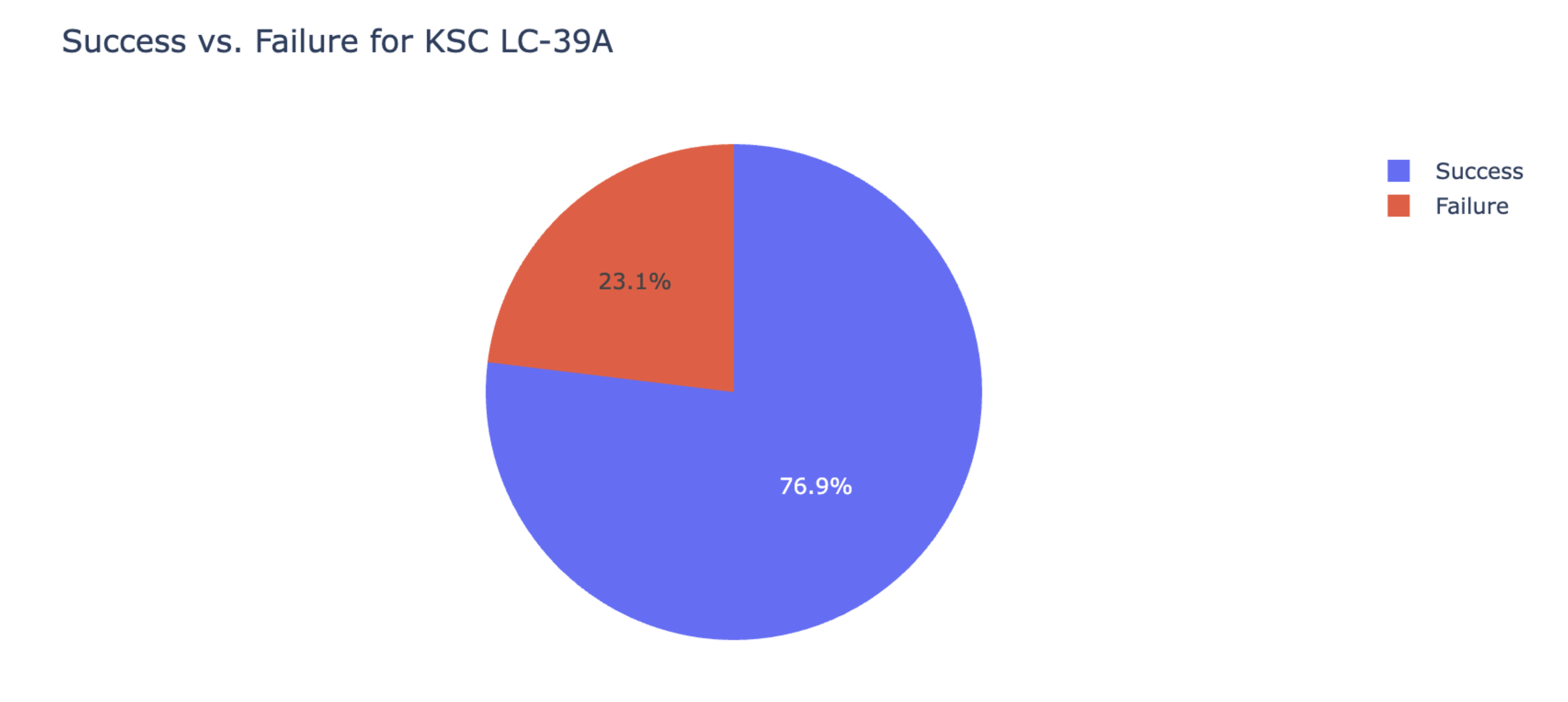
- **Findings**:
  - **KSC LC-39A** has the **highest number of successful launches**
  - **CCAFS SLC-40** has the **fewest successful launches**



KSC LC-39A Launch Success vs Failure Ratio

- **Two segments are shown:**
  - **Blue** for successful launches.
  - **Red** for failed launches.

- **Findings**:
  - **Success Rate**: 76.9% of the launches at KSC LC-39A were successful.
  - **Failure Rate**: 23.1% of the launches failed.



Payload vs. Launch Outcome by Booster Version

- **Scatter Plot Details**:
  - Each **color** corresponds to a different **Booster Version Category**.
  - The **range slider** below filters payload values dynamically, allowing users to explore trends across payload sizes.

- **Key Findings**:
  - **Booster Version FT** (Falcon 9 Full Thrust) displays a **high success rate**, with most of its data points aligned at the "1".

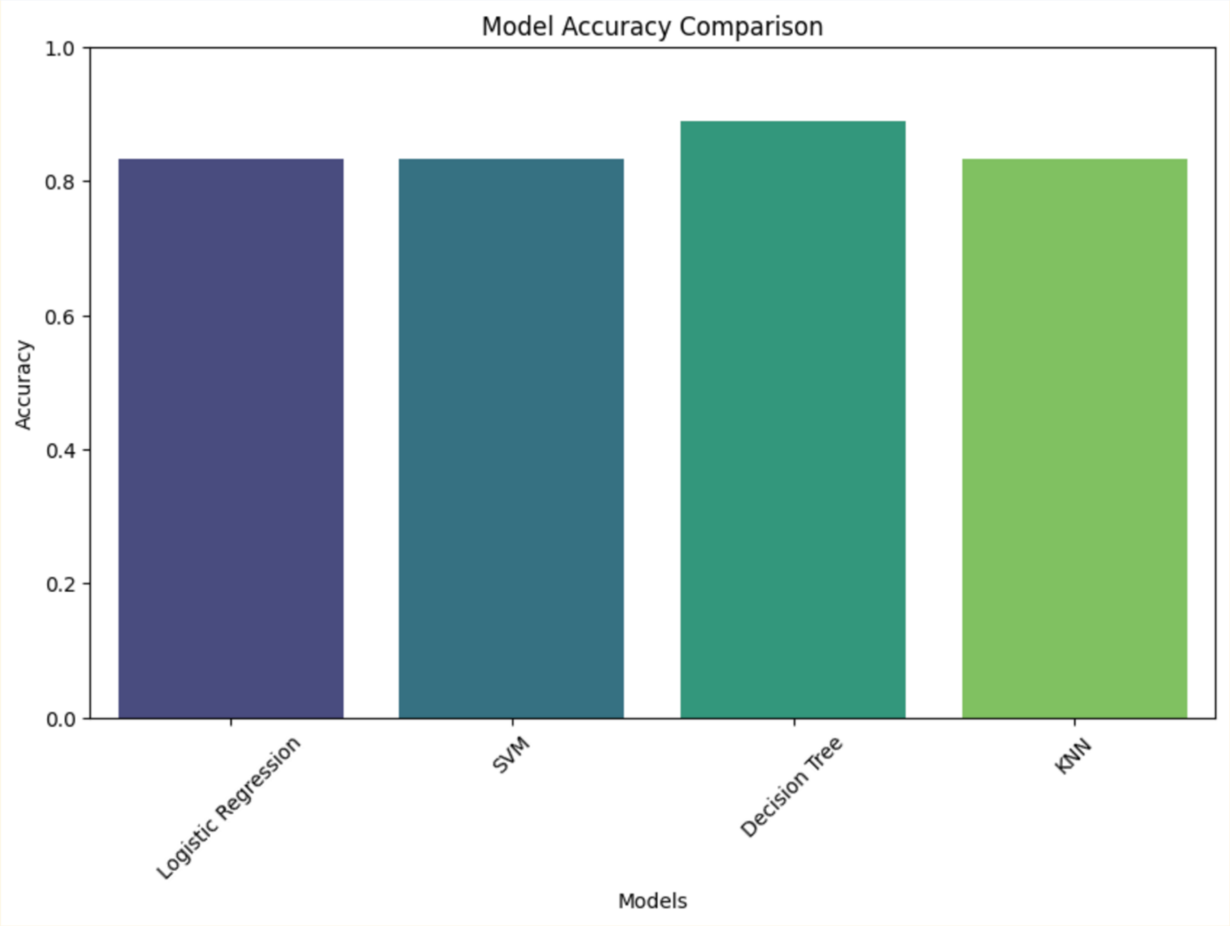
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- **Logistic Regression**: 84%
- **Support Vector Machine (SVM)**: 84%
- **Decision Tree**: **89%**
- **K-Nearest Neighbors (KNN)**: 84%

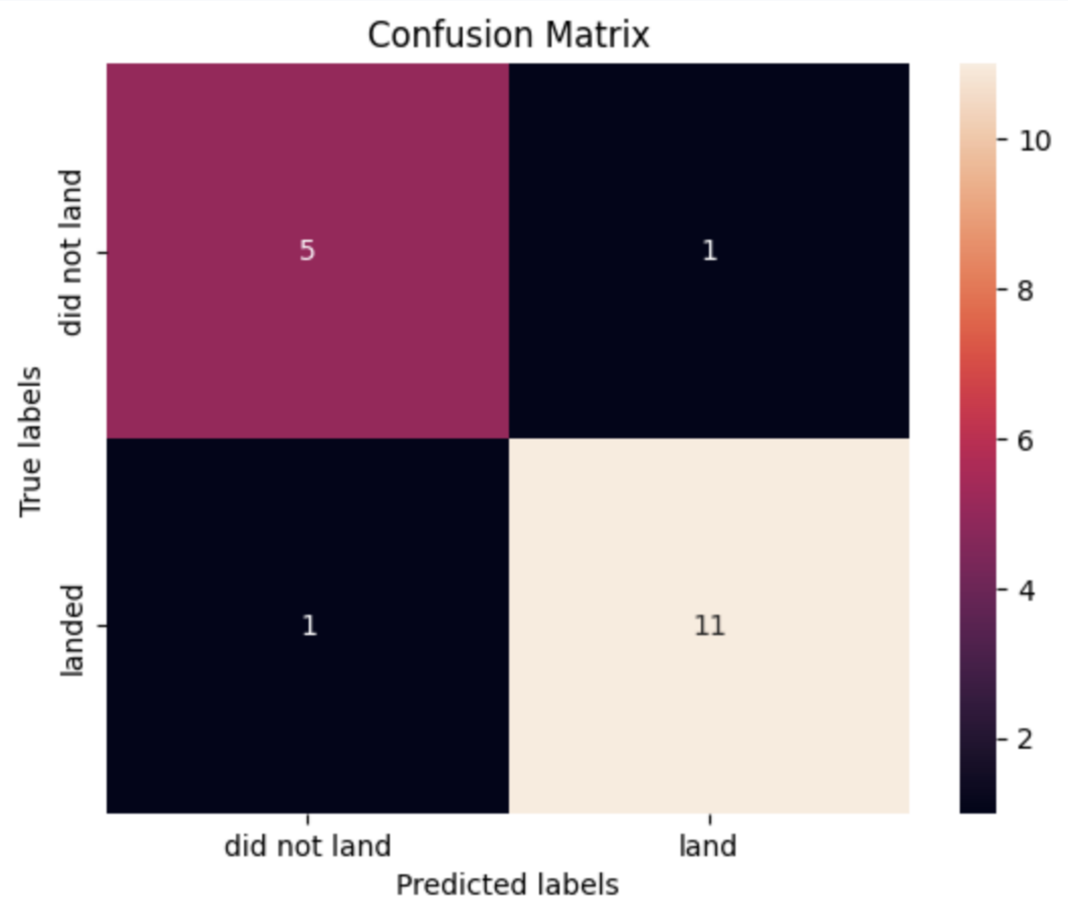
- **Decision Tree** achieved the **highest accuracy** at **89%**, making it the best-performing model among those evaluated.



Confusion Matrix

- **True Positives (11)**: Correctly predicted landings.
- **False Positives (1)**: Predicted landing, but it did not land.
- **True Negatives (5)**: Correctly predicted no landing.
- **False Negatives (1)**: Predicted no landing, but it actually landed.

- The model does a **great job identifying landings**, with only **2 errors out of 18**.



Conclusions

1. KSC LC-39A and CCAFS SLC-40 are the two most active launch sites, with KSC LC-39A having the highest number of successful landings

2. Payload mass influences landing success: Booster version FT and B4 are more successful with heavy payloads.

3. Orbit type matters: Orbits such as SSO, ES-L1, and GEO have the highest success rates, while GTO has a mixed outcome and SO has zero success.

4. Launch history shows performance growth: The yearly success rate has steadily increased from 2013 to 2020

5. Decision Tree Classifier outperformed other models with an accuracy of 0.89, making it the most effective for predicting landing outcomes based on mission parameters.

Thank you!

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project