IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Leona Garabedian
August 5, 2025

# Outline

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

# Executive Summary

- Summary: Methodologies
  - Data Collection & Preparation
    - Retrieved Falcon 9 launch data via SpaceX REST API and Wikipedia web scraping.
    - Data wrangling component of cleaning data by keeping only Falcon 9 launches, filling missing values, and linking IDs to clear info.
    - Created training labels showing if the first stage landed successfully (yes = 1, no = 0).
- Exploratory Data Analysis & Visualization:
    - Created scatter plots, line plots, and bar charts with Python and Pandas to explore data distributions and patterns.
    - Built interactive dashboards with Plotly Dash to show launch results and key stats.
    - Created maps with Folium to study launch site locations and distances.
- Machine Learning Modeling:
    - Split data into training and test sets.
    - Trained models: SVM, Decision Trees, Logistic Regression.
    - Tuned model settings with grid search to improve accuracy.
    - Tested models and picked the best one to predict first stage landings.

# Introduction

- Project Background
  - I am a data scientist at Space Y, a new rocket company aiming to compete with SpaceX.
  - SpaceX offers Falcon 9 launches with reusable first stages, drastically reducing launch costs.
  - My goal is to determine launch prices for Space Y by analyzing SpaceX's public launch data.
  - Instead of rocket science, I will use machine learning to predict if SpaceX will reuse the Falcon 9 first stage.
  - This prediction directly impacts launch costs and competitive pricing strategies.
- Problems to Solve
  - How can I accurately predict whether SpaceX will reuse the Falcon 9 first stage using public data?
  - Which features best indicate the likelihood of first stage reuse?
  - How can Space Y leverage these predictions to set competitive launch prices?
  - How can I build dashboards that help the team make data-driven decisions?

Section 1

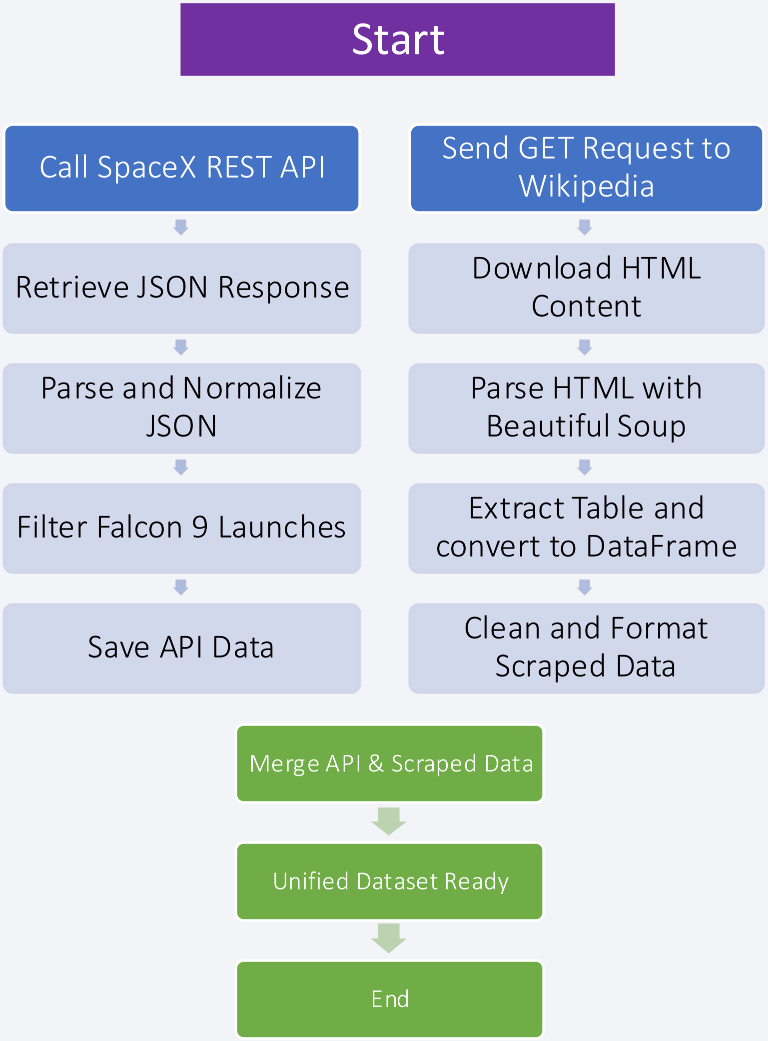
# Methodology

# Methodology

I. Data collection methodology
   A. Used SpaceX REST API to retrieve launch data.
   B. Scraped Wikipedia for supplemental info.
II. Perform data wrangling
   A. Filtered for Falcon 9 launches and cleaned missing values.
   B. Standardized dates, merged datasets, encoded categorical variables.
III. Perform exploratory data analysis (EDA) using visualization and SQL
IV. Perform interactive visual analytics using Folium and Plotly Dash
V. Perform predictive analysis using classification models
   A. Split data into train/test sets, used grid search for hyperparameter tuning, and validated performance with confusion matrices and test metrics.

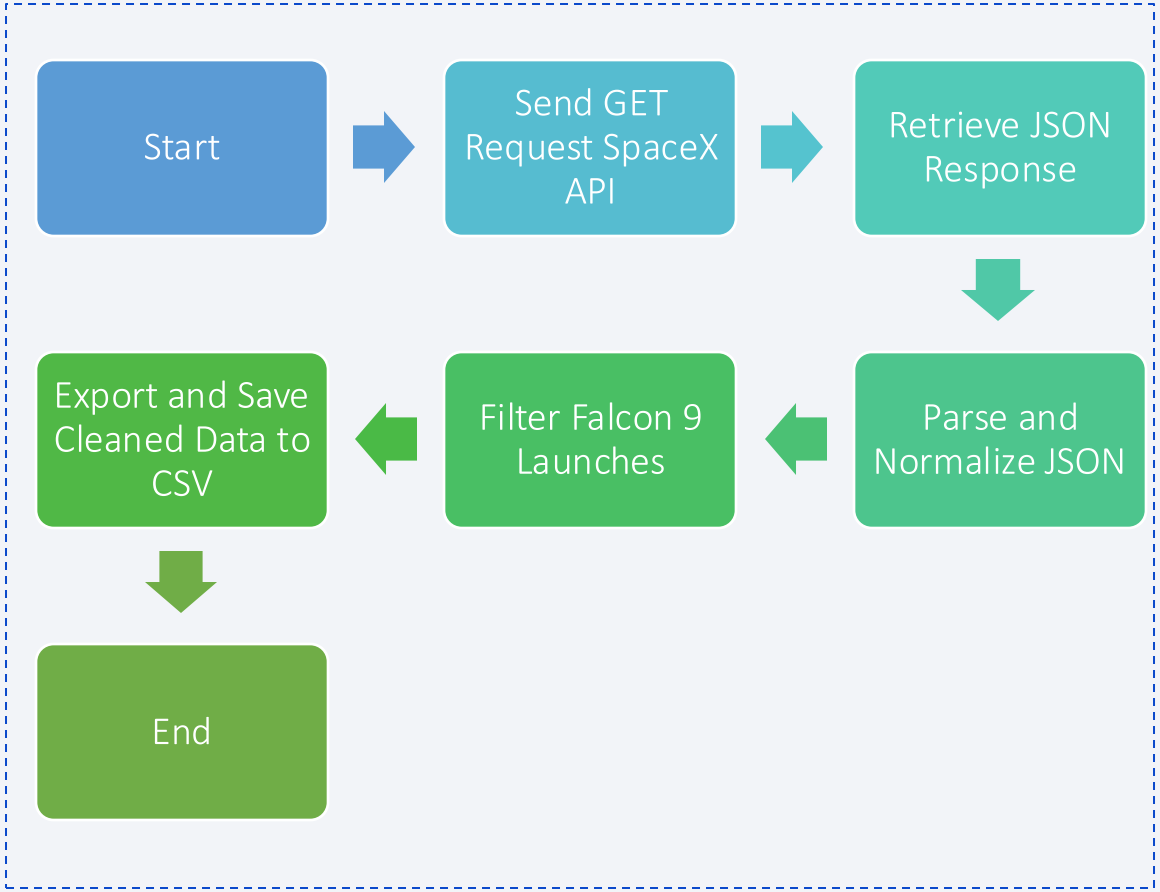
# Data Collection

1. **Access SpaceX REST API** to retrieve launch data in JSON format.
2. **Parse and normalize JSON** into pandas Data Frames.
3. **Filter Falcon 9 launches** for focused analysis.
4. **Scrape Wikipedia pages** for supplemental launch and landing data.
5. **Parse HTML tables** using BeautifulSoup to extract additional datasets.
6. **Merge API and scraped data** into a unified dataset.
7. **Export cleaned dataset** for further analysis.



# Data Collection – SpaceX API

1. Access SpaceX REST API: Send HTTP GET request.
2. Retrieve JSON Data: Obtain list of launch records in JSON format.
3. Parse and Normalize JSON: Use pd.json_normalize() to flatten JSON into tabular DataFrame.
4. Extract Relevant Fields: Flight number, mission name, launch year, rocket details, payload, orbit, launch site.
5. Filter for Falcon 9 Launches: Retain only records where Falcon 9 launches.
6. Save Cleaned Dataset: Export to CSV for downstream analysis.

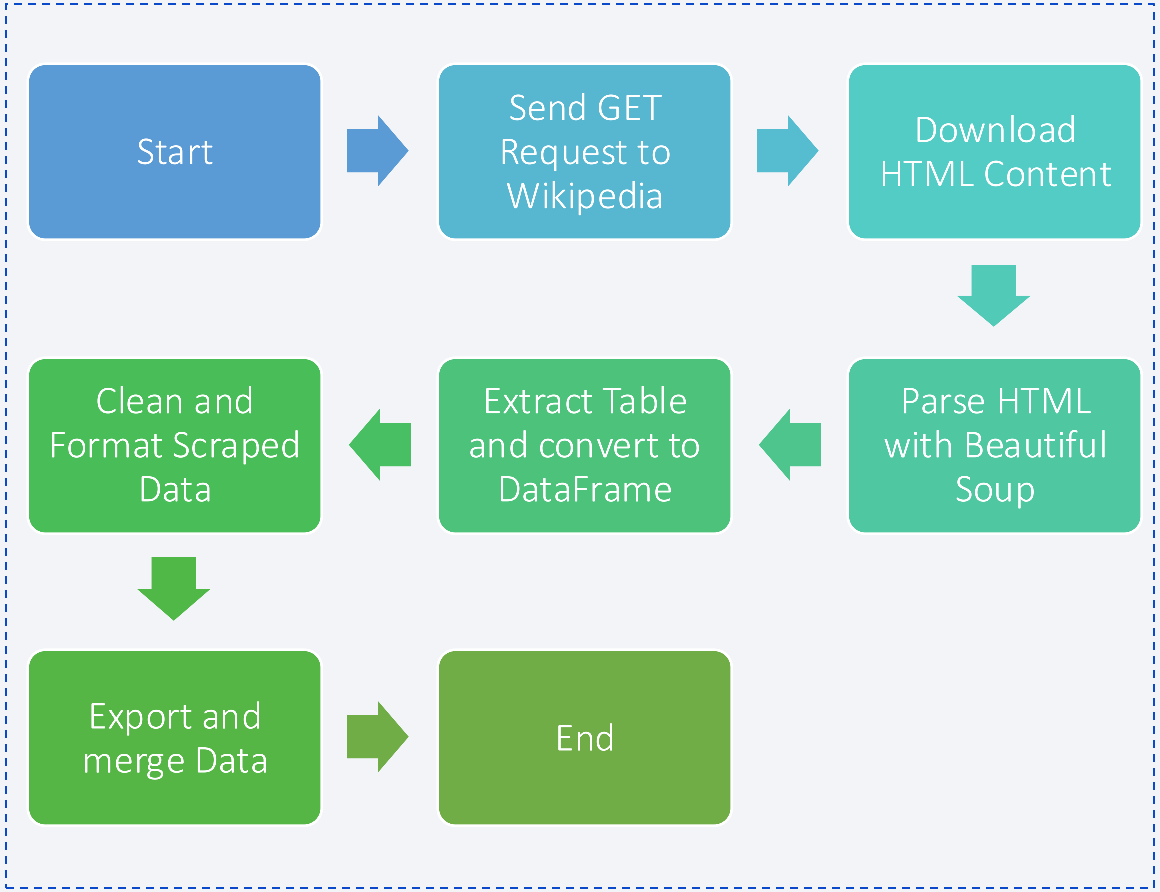



https://github.com/Leonagarabedian/Collecting-SpaceX-Data

# Data Collection - Scraping

1. Send HTTP GET request to SpaceX Wikipedia pages.
2. Download and parse raw HTML content.
3. Use BeautifulSoup to locate launch-related data tables.
4. Extract tables and convert to pandas DataFrames.
5. Clean and format scraped data (fix headers, remove empty rows).
6. Export or merge scraped data for analysis.

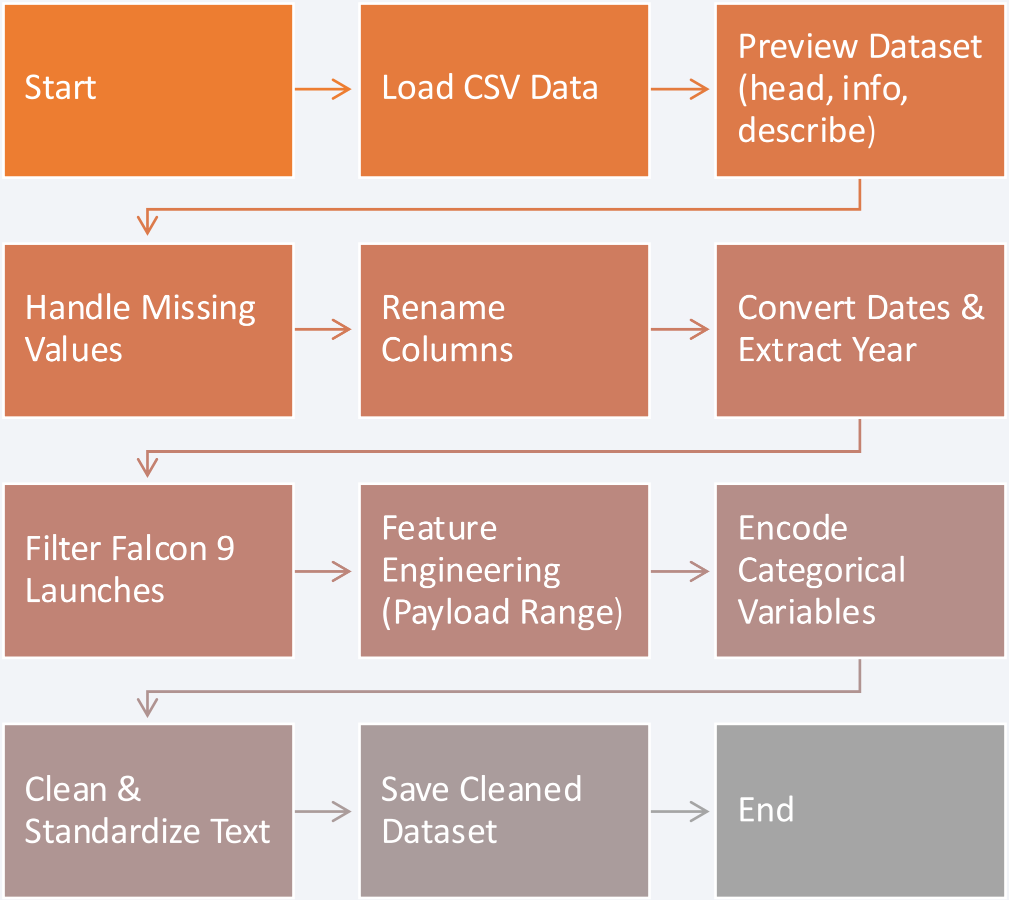
https://github.com/Leonagarabedian/Web-Scrapping-From-Wikipedia



# Data Wrangling

1. Load SpaceX launch CSV data into pandas DataFrame.
2. Preview dataset structure and summary statistics.
3. Identify and handle missing values by dropping incomplete rows.
4. Rename columns for clarity and consistency.
5. Convert date columns to datetime format and extract launch year.
6. Filter dataset to retain only Falcon 9 launches.
7. Engineer new features such as payload mass ranges.
8. Encode categorical variables (e.g., orbit, launch site) using one-hot encoding.
9. Clean and standardize textual data.
10. Save the cleaned and processed dataset as CSV.

https://github.com/Leonagarabedian/Data-Wrangling-SpaceX-Falcon9-Landing-Prediction



| Data Visualizations | Insights |
|---|---|
| Flight Number vs. Payload Mass (Success Overlay) | *Scatter plot with class (success/failure) overlay*<br>Analyze combined effect of launch order and payload on mission outcomes. |
| Flight Number vs. Launch Site (Success Overlay) | *Scatter plot with class overlay*<br>Examine relationship between launch sequence, site, and success rates. |
| Payload Mass vs. Launch Site (Success Overlay) | *Scatter plot with class overlay*<br>Examine relationship between launch sequence, site, and success rates. |
| Flight Number vs. Orbit Type | *Scatter plot*<br>Visualize how launch sequence correlates with mission orbit targets. |
| Payload Mass vs. Orbit Type | *Scatter plot*<br>Explore payload distribution across different orbit categories. |
| Success Rate by Orbit Type | *Bar chart*<br>Assess how mission success varies across orbit types. |
| Launch Success Trend Over Time | *Line chart*<br>Track annual improvements or changes in mission success rates. |

# EDA with Data Visualization

https://github.com/Leonagarabedian/Falcon9-Landing-EDA-Visuals

# EDA with SQL

**SQL queries collectively aimed to extract, filter, and summarize critical launch and landing data**

- Selected unique launch sites to identify launch locations.
- Filtered and displayed launches where launch sites begin with specific prefixes (e.g., “CCA”).
- Counted the number of successful and failed missions overall and by launch site.
- Queried boosters with payload mass between specified ranges and successful drone ship landings.
- Retrieved booster versions that carried the maximum payload mass using aggregate functions and subqueries.
- Ranked landing outcomes by count for a specific date range.
- Extracted month and year from launch dates to analyze monthly and yearly trends.
- Grouped launches by orbit type to summarize mission counts per orbit.
- Joined launch and landing outcome data for detailed mission success analysis.

https://github.com/Leonagarabedian/SpaceX-SQL-Analysis

# Build an Interactive Map with Folium

Folium was used to explore the geographic and spatial factors that may influence SpaceX rocket launches.

Markers
- Placed at exact coordinates of launch sites (CCAFS SLC 40, VAFB SLC 4E, KSC LC 39A).
  - Identified and mapped all SpaceX launch sites on a geographic map.

Circles
- Drawn around launch sites with ~1000m radius, colored and semi-transparent.
  - Visually emphasize spatial area and buffer zone around each launch facility.

Popup Labels
- Attached to markers/circles showing launch site names on click.
  - Provide interactive, on-demand information without cluttering the map.

Lines (Polylines)
- Connected launch sites to nearby landmarks (coastlines, highways, trainline, and city ) using lines. Visualize spatial relationships and proximity between sites and features.

https://github.com/Leonagarabedian/Interactive-Visual-Analytics-with-Folium-SpaceX-Launch-Sites

# Build a Dashboard with Plotly Dash

1. Launch Success Counts Pie Chart : Visualizes **total** successful launches by each SpaceX launch site for an easy performance overview.

2. Launch Success Ratio Pie Chart (Per Site): Shows success vs failure percentages for the selected launch site to assess reliability.

3. Payload Mass vs. Launch Outcome Scatter Plot: Displays the impact of payload mass on launch success, segmented by booster version, revealing performance trends.

4. Launch Site Dropdown Menu: filters all plots to focus analysis on a selected launch site.

5. Payload Mass Range Slider: Dynamically filters scatter plot data by payload mass to explore success rates within specific payload ranges.

https://github.com/Leonagarabedian/SpaceX-Visuals-Application

# Predictive Analysis (Classification)

Model Development & Evaluation Process

- <u>Data Collection & Cleaning:</u> Gathered Falcon 9 launch data and prepared it for modeling.
- <u>Train-Test Split:</u> Separated data to train models and test their performance.
- <u>Model Training:</u> Built several classifiers: Logistic Regression, SVM, Decision Tree, KNN.
- <u>Hyperparameter Tuning:</u> Used grid search to find the best settings for each model.
- <u>Model Evaluation:</u> Compared models using accuracy, precision, recall, and F1-score.
- <u>Best Model Selection:</u> Chose Decision Tree as best based on accuracy and balanced performance.


https://github.com/Leonagarabedian/Machine-Learning-Space-X-Falcon-9-First-Stage-Landing-Prediction/tree/main

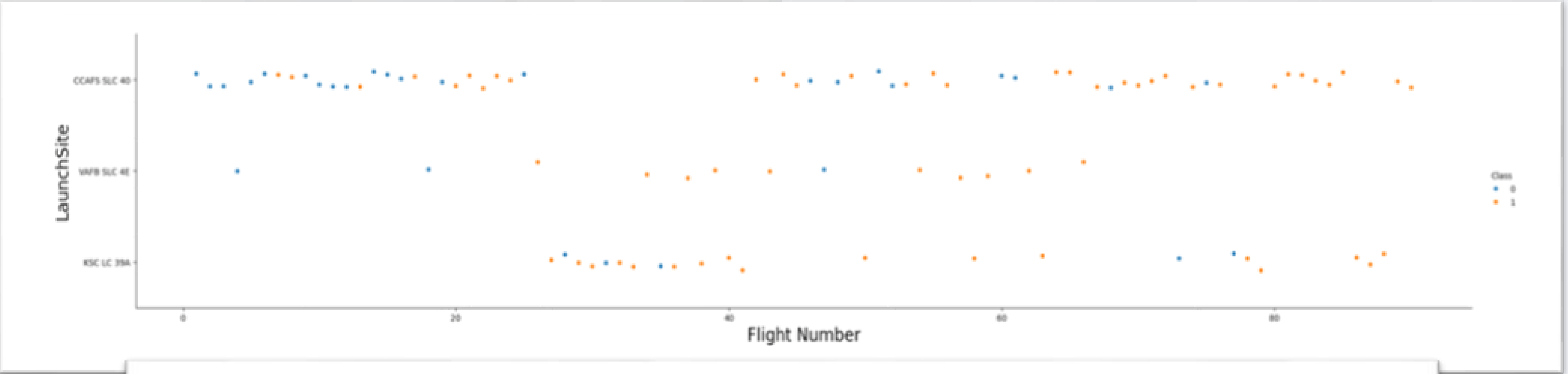
# Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

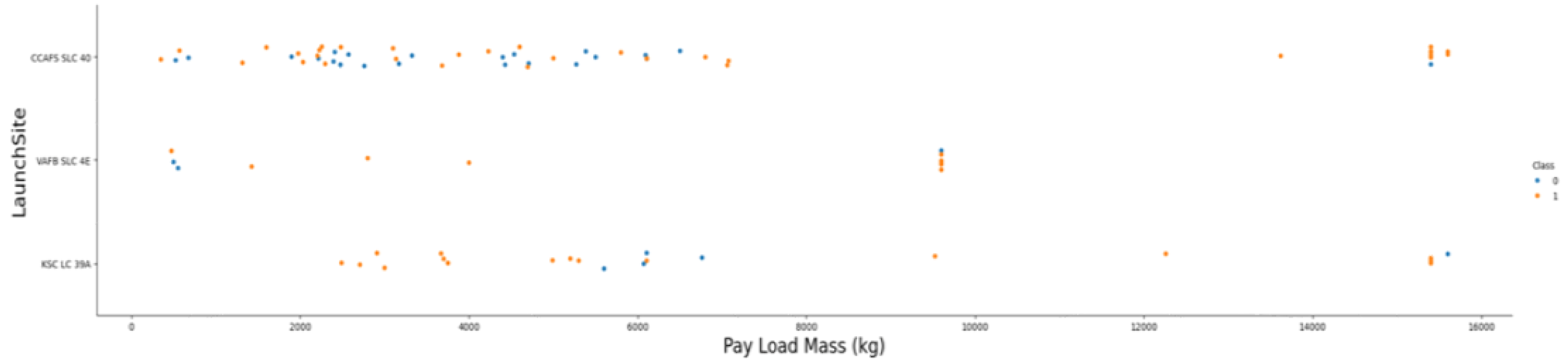
# Insights drawn from EDA

# Flight Number vs. Launch Site



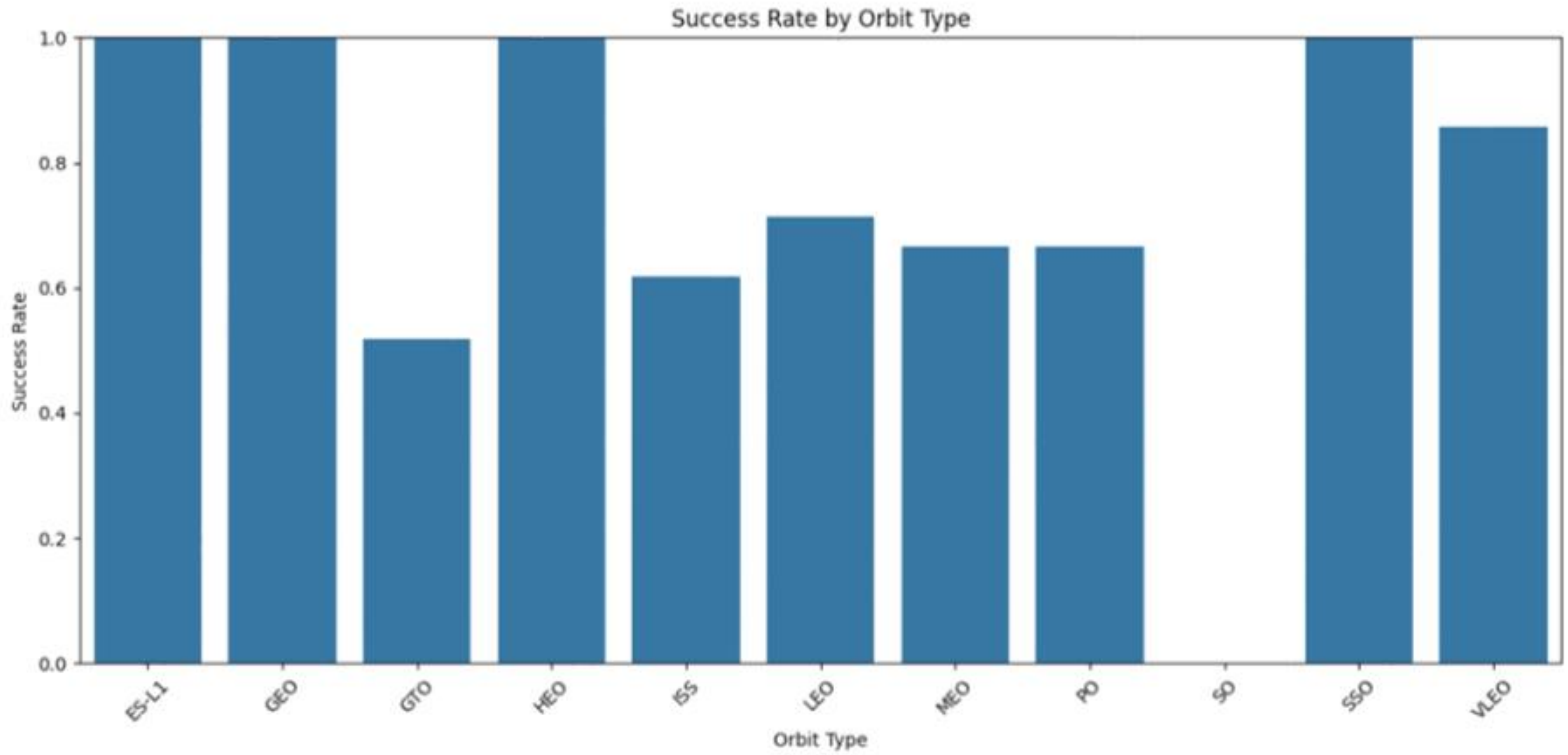

plots. CCAFS SLC 40 has the highest number of flights across the range of flight numbers. VAFB SLC 4E has fewer flights and they are more spread out. KSC LC 39A has a moderate number of flights, primarily concentrated in the middle flight numbers. Flight numbers increase along the x-axis, showing a timeline or sequence of launches. Launch site points labeled on the y-axis, indicating that launches from these sites occurred throughout the mission history. The Class color indicates the success (1) or failure (0) of the booster landing. At CCAFS SLC 40, there is a balanced mix of successful and unsuccessful landings over most flight numbers, but with some clusters of successes towards higher flight numbers. At VAFB SLC 4E, most points are orange (Class 1), suggesting a higher success rate at this site, though there are less launches. At KSC LC 39A, the data shows more orange points as well.

# Payload vs. Launch Site



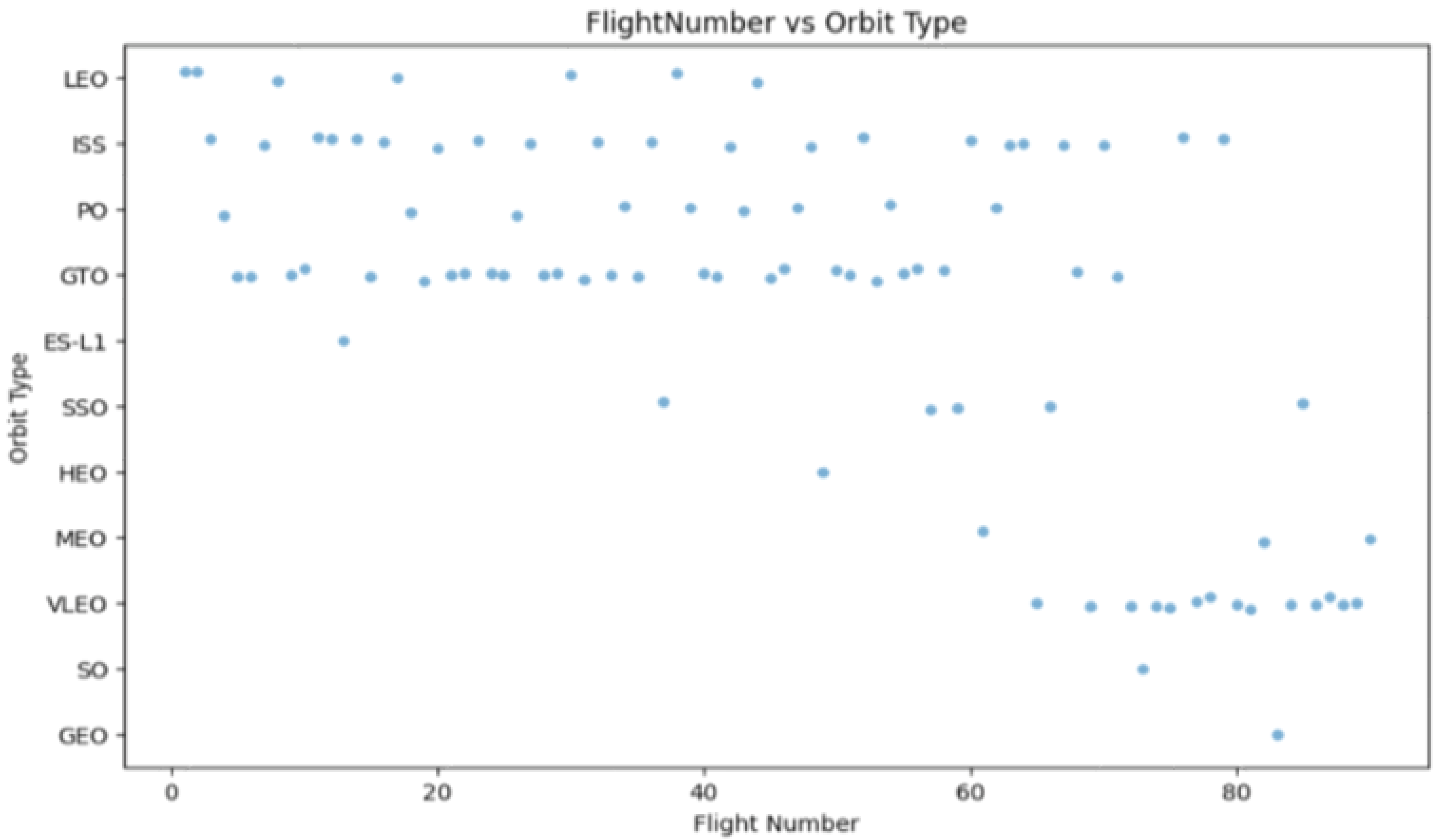

Now if you observe Payload Mass Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

# Success Rate vs. Orbit Type



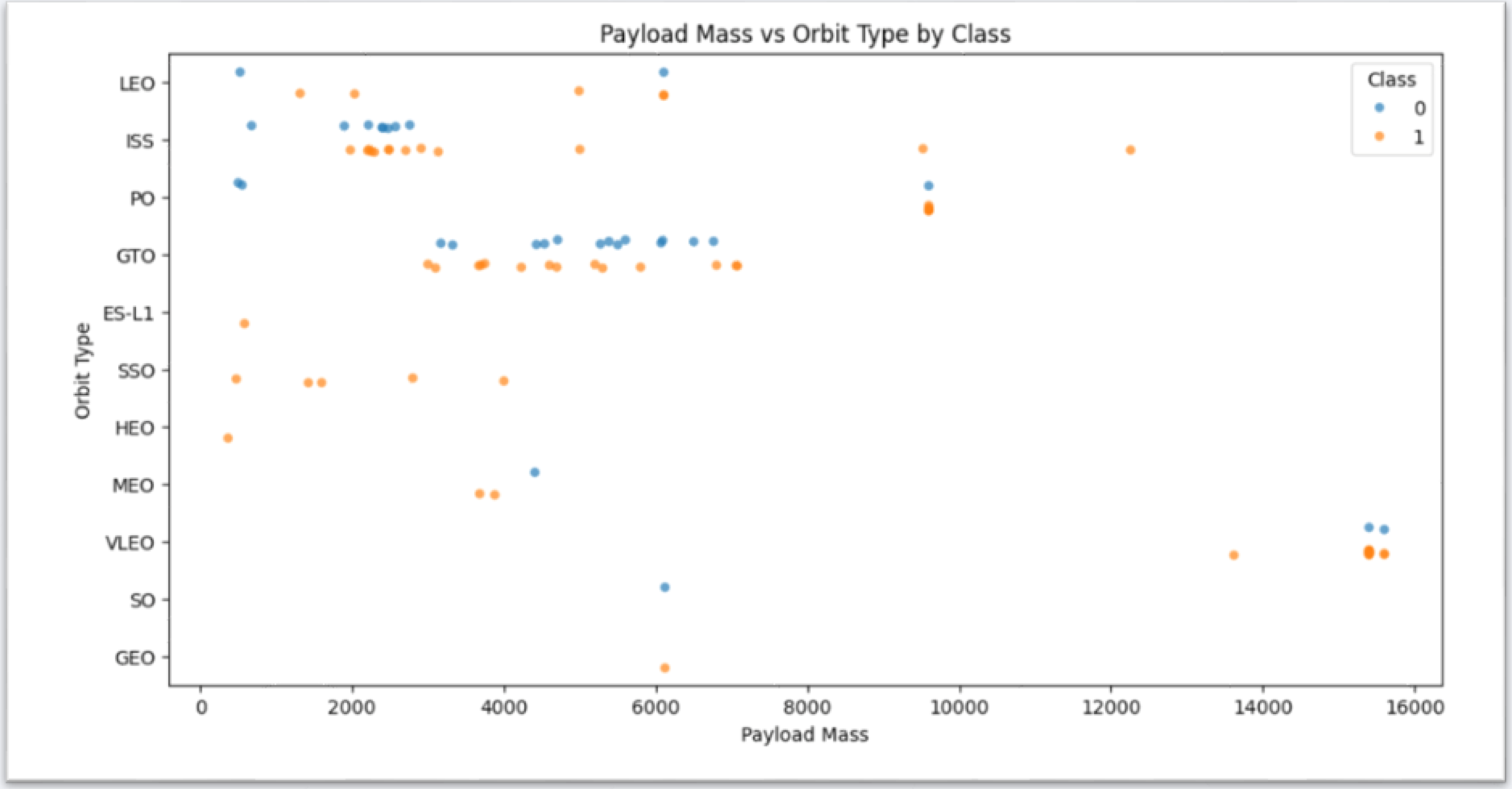

Analyze the plotted bar chart to identify which orbits have the highest success rates. Orbits with the highest success rates are ES-L1, GEO, HEO, and SSO.

# Flight Number vs. Orbit Type




You can observe that 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.

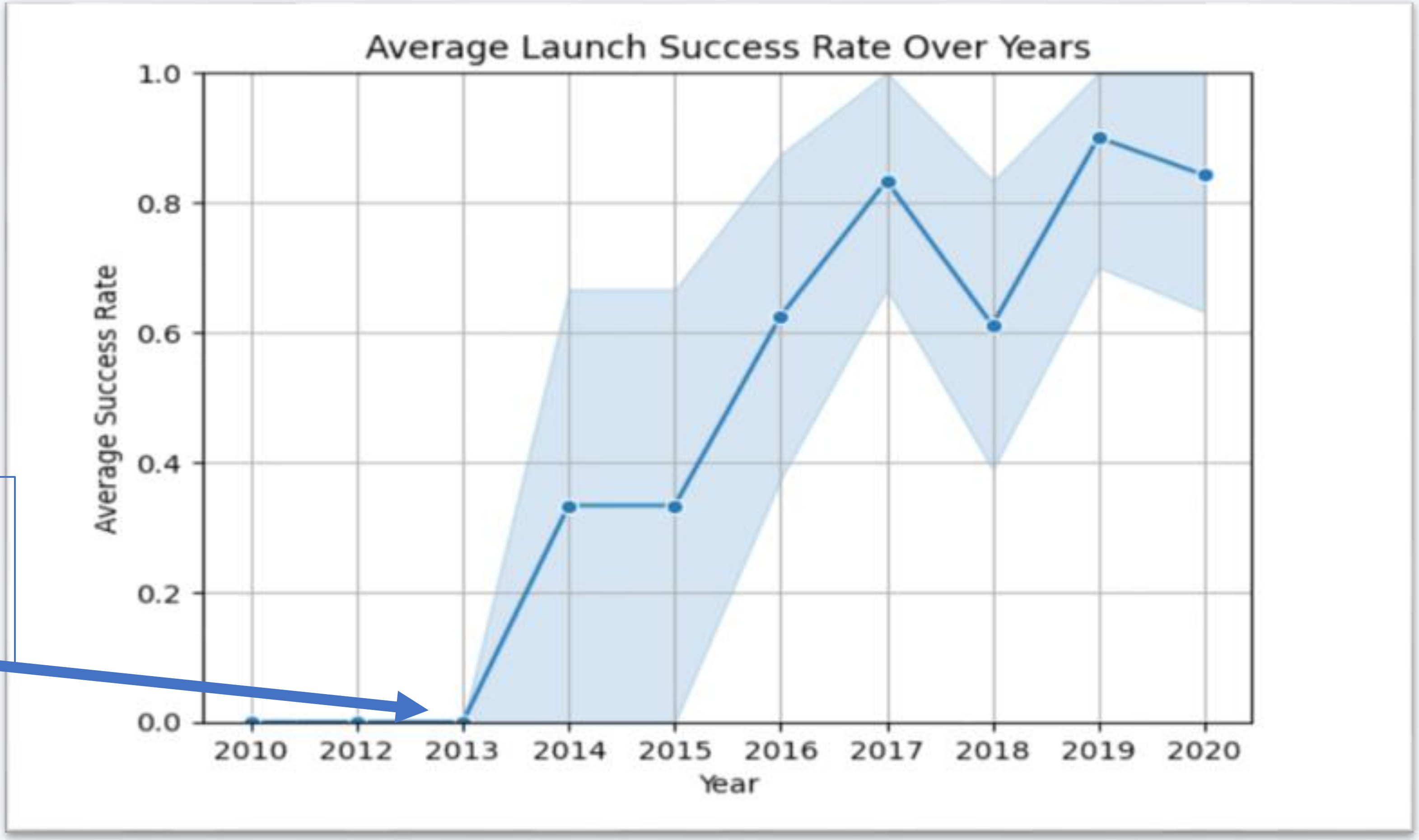
# Payload vs. Orbit Type




<u>Payload masses vary widely across different orbit types for both classes.</u>

- LEO, ISS, and GTO have launches with mixed classes and varying payloads.
- Higher payload masses (above 10,000 kg) mainly occur in VLEO.
- Some orbits (SSO, ES-L1, HEO) show more Class 1 points at lower payloads.

# Launch Success Yearly Trend




Success Rate since 2013 increase until 2020

# All Launch Site Names

## Understanding SpaceX Launch Sites

By analyzing the SpaceX dataset and executing SQL queries, I identified four unique launch sites where missions are conducted.

Table 1. SpaceX Launch Sites

| **Launch Site** | CCAFS LC-40 | VAFB SLC-4E | KSC LC-39A | CCAFS SLC-40 |
|---|---|---|---|---|

**CCAFS LC-40** and **CCAFS SLC-40** are at Cape Canaveral Air Force Station, Florida.

**VAFB SLC-4E** is at Vandenberg Air Force Base, California.

**KSC LC-39A** is the historic Kennedy Space Center launch complex.

# Launch Site Names Begin with 'CCA'

Table 2. Records from Launch Sites Starting 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 | 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 |

*These records highlight early SpaceX missions launched from Cape Canaveral Air Force Station (CCAFS LC-40). Payloads mainly served NASA's ISS cargo needs. Early recovery efforts showed booster landing failures or no landing attempts, reflecting SpaceX's development phase.*

# Total Payload Mass

Query Result

| Total Payload (kg) |
| --- |
| 45,596 |

The total payload mass delivered by SpaceX boosters for NASA missions is 45,596 kg. This sum highlights the significant volume of cargo transported to support NASA's programs, including resupply missions to the International Space Station.

# Average Payload Mass by F9 v1.1

Query Result

| Average Payload (Kg) |
| --- |
| 2,928.4 |

The average payload mass carried by the SpaceX booster version **F9 v1.1** is approximately **2,928.4 kg**. This reflects the typical cargo capacity for missions using this booster version, showing its role in medium payload deliveries.

# First Successful Ground Landing Date

Query Result

| First Successful Landing Date |
|---|
| 2015-12-22 |

The earliest recorded date for a successful booster landing on a ground pad was December 22, 2015. This milestone marked a key achievement in SpaceX's efforts toward booster reusability and cost reduction in space launches.

| Booster Version |
| --- |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

## Successful Drone Ship Landing with Payload between 4000 and 6000

These booster versions have successfully landed on drone ships after missions carrying payloads between 4,000 kg and 6,000 kg. Successful drone ship landings are vital for recovering boosters from missions with higher velocity requirements, enabling SpaceX's goal of rocket reusability.

# Total Number of Successful and Failure Mission Outcomes

| Total Success | Total Failure |
|---|---|
| 100 | 1 |

Out of all recorded missions, 100 were successful while 1 mission failed. This demonstrates SpaceX's high success rate in their launch operations.

# Boosters Carried Maximum Payload

These booster versions each carried the maximum recorded payload mass of 15,600 kg. This represents the peak cargo capacity achieved by SpaceX boosters in the dataset, demonstrating their capability to transport heavy payloads.

| 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

These records show the months in 2015 when SpaceX experienced failed booster landings on drone ships. Both failures occurred early in the year from launches at Cape Canaveral Air Force Station (CCAFS LC-40), highlighting challenges in booster recovery during that period.

| 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

| 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 |

This ranking shows the frequency of various landing outcomes over the specified period. The majority were "No attempt," followed by both successful and failed landings on drone ships, indicating early-stage testing and recovery efforts by SpaceX.

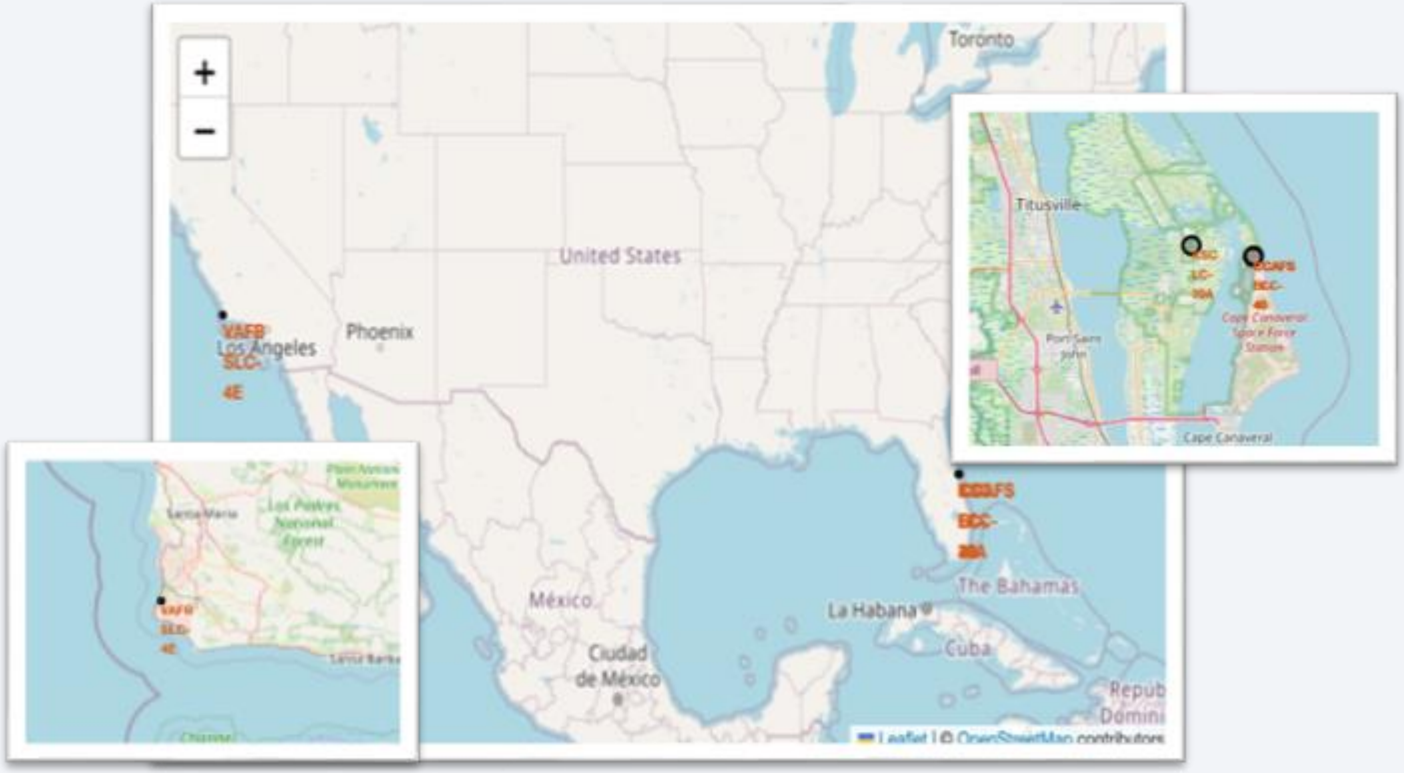
Section 3

# Launch Sites Proximities Analysis

# Global Locations of SpaceX Launch Sites

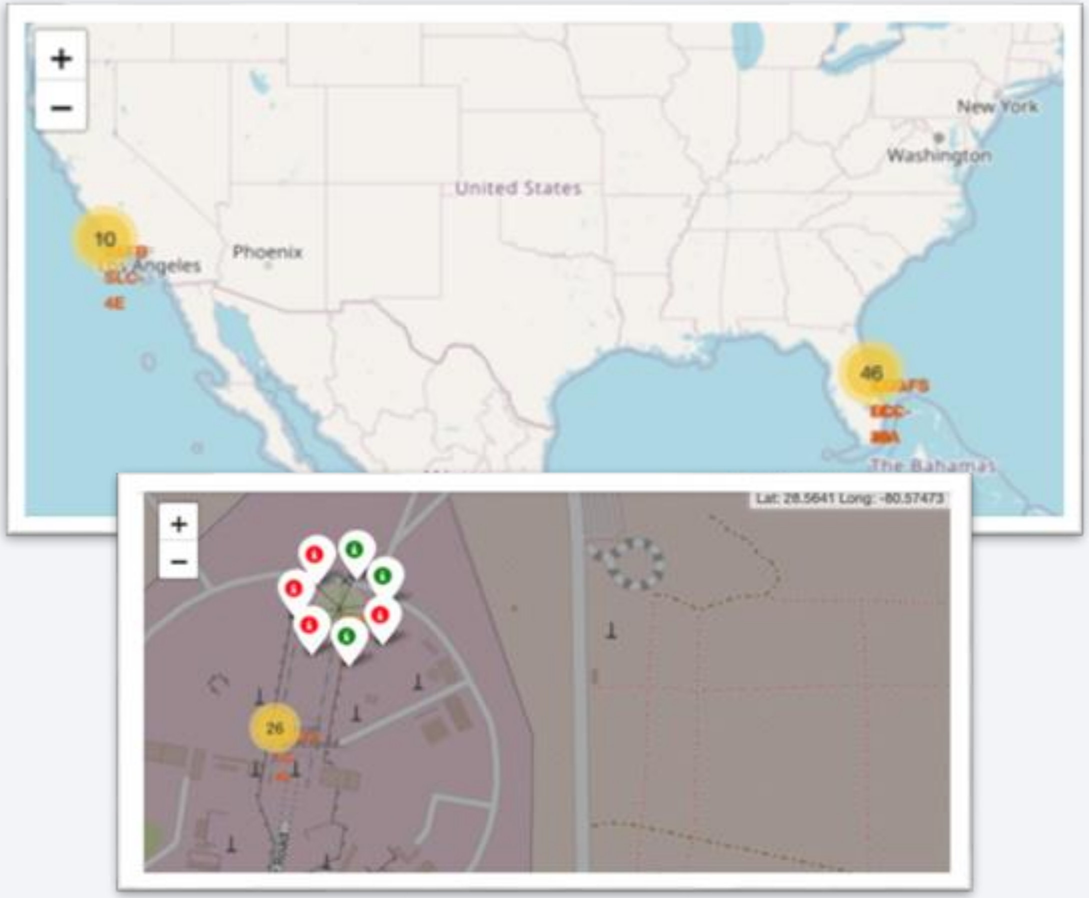
SpaceX launch sites are marked with black dots and labels:

- **CCAFS LC-40** and **CCAFS SLC-40** (Cape Canaveral, Florida).
- **KSC LC-39A** (Kennedy Space Center, Florida).
- **VAFB SLC-4E** (Vandenberg Air Force Base, California).
- Launch sites are not in proximity of the equator line.
- All are very close to the coast.
- Map includes zoom controls and shows geographic context, aiding spatial understanding of site locations relative to major cities.



# Launch Outcomes Across SpaceX Launch Sites

- Color-coded Markers:
  The map uses colored markers to represent different launch outcomes:
  - Green markers indicate successful launches.
  - Red markers indicate failed launches or issues.
- Yellow clusters show aggregated numbers of launches in close proximity.
- Launch Site Clusters:
  The clusters with numbers (e.g., 46 and 10) show the total launches at that site, providing a quick sense of launch frequency and success at each location.

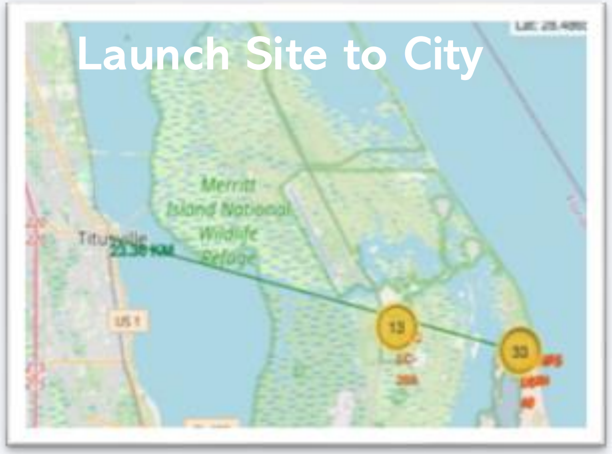


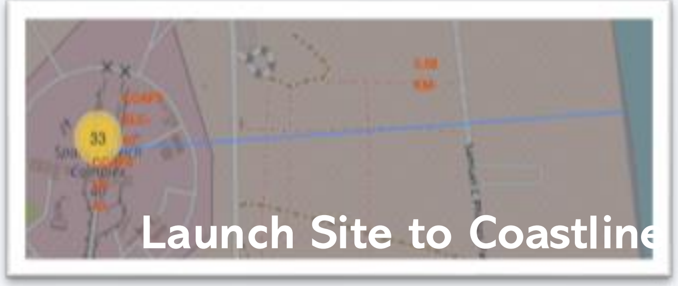
# Launch Site Proximity to Key Landmarks

The map displays the selected launch site with lines connecting it to nearby landmarks: railway, highway, and coastline, with distances labeled (e.g., 0.66 KM, 0.94 KM, 23.36 KM, 0.58 KM).

- **Close Proximity to Railways:**
  Facilitating transport and logistics.
- **Close Proximity to Highways:**
  Providing easy road access for personnel and cargo.
- **Close Proximity to Coastline:**
  Essential for launch trajectory safety and maritime recovery operations.
- **Distance from Cities:**
  maintains a more significant distance from populated urban areas, reducing risk and ensuring safety zones.





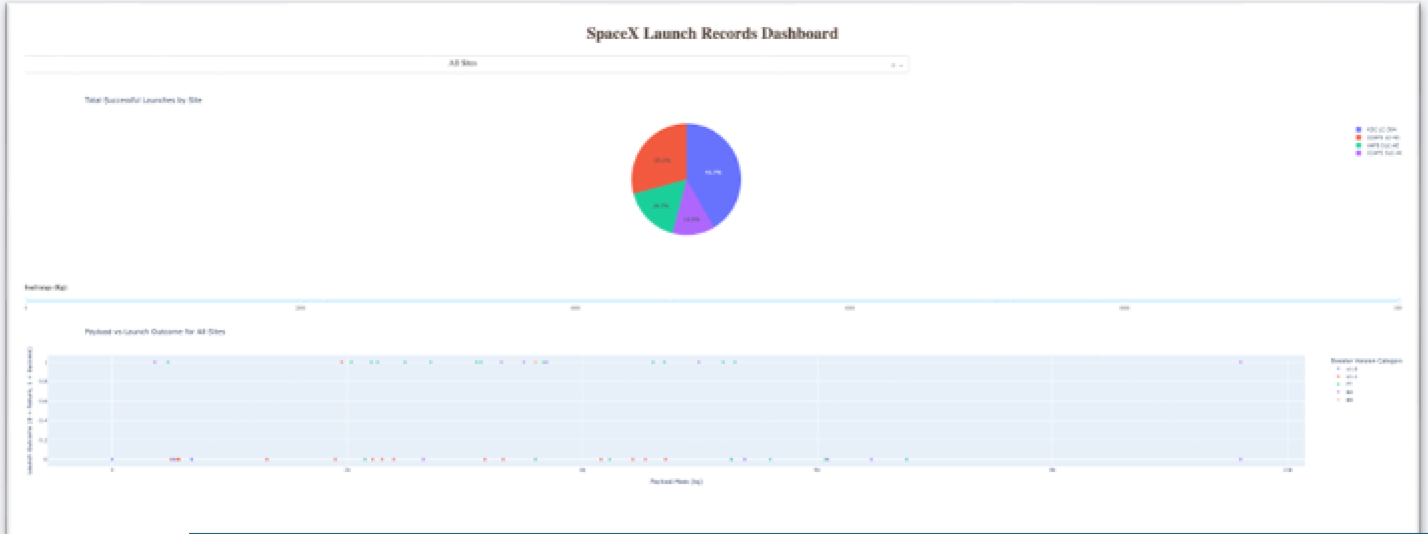




**Summary:**
These proximity measurements help assess logistical convenience, safety considerations, and environmental impact around the launch site.

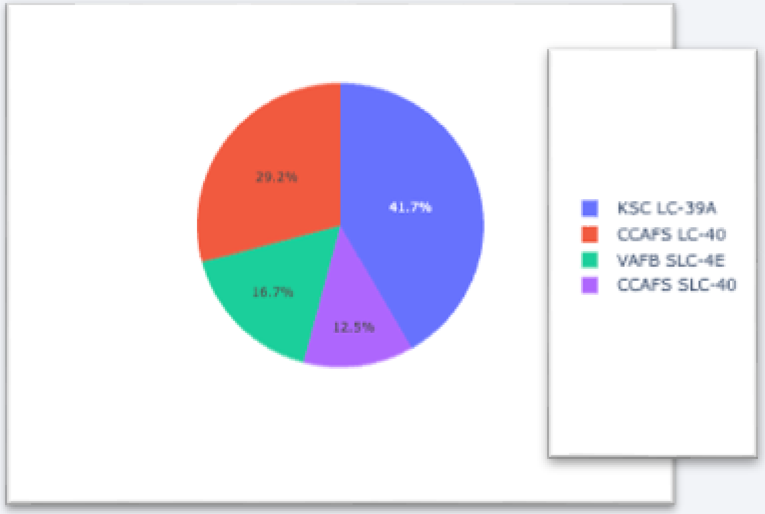
Section 4

# Build a Dashboard with Plotly Dash

# Launch Success Counts Across SpaceX Launch Sites




Plotly dashboard allowing users to interactively analyze SpaceX records

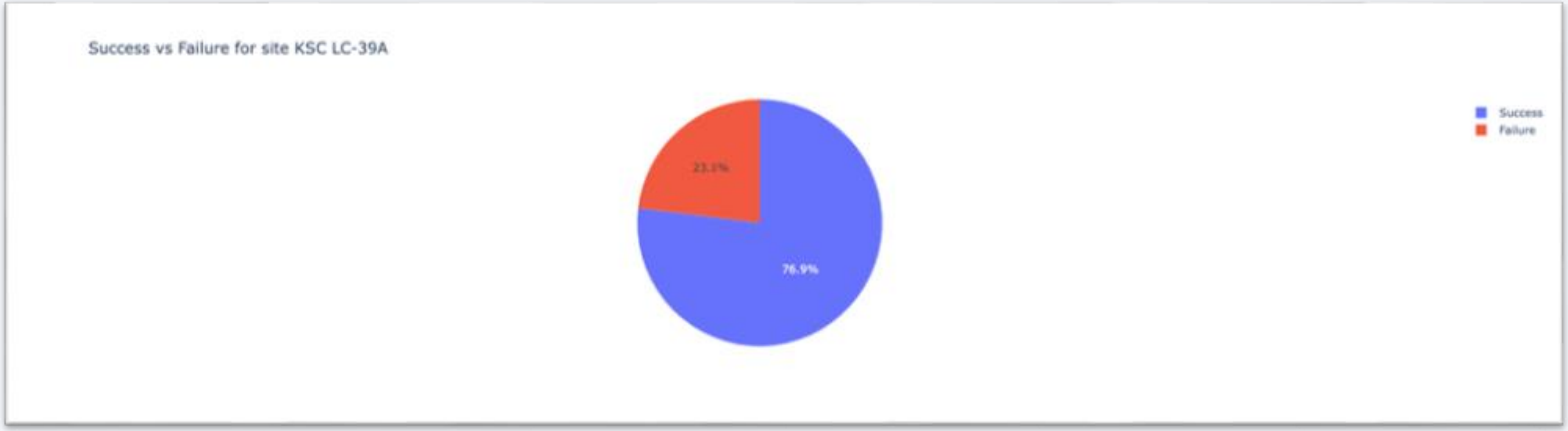



Launch Success:
Kennedy Space Center (KSC LC-39A) leads in total successful launches (about 42%), followed by CCAFS LC-40 (29%).

Features include:

1. Pie chart showing total successful launches by site.
2. Scatter plot of payload mass vs. launch outcome, categorized by booster version.
3. Payload mass slider for filtering launches by payload range.
4. Dropdown menu for selecting specific launch sites.

# Launch Success vs Failure Ratio for KSC LC-39A




**Success Dominates:**
The pie chart shows that **76.9%** of launches from KSC LC-39A were successful, indicating a strong track record at this launch site.
**Failure Rate:**
The failure rate of **23.1%** reflects the challenges and risks inherent in space launches but remains relatively low compared to success.
**Operational Insight:**
KSC LC-39A is a reliable and frequently used launch site for SpaceX, contributing significantly to overall mission success.

# Payload vs. Launch Outcome Scatter Plot Across All Sites


<u>Payload Range with Highest Success</u> ➡ payloads between 362 kg and 5300 kg across all booster versions.

<u>Payload Range with Lowest Success</u> ➡ Payloads in the range 0 kg to 6761 kg reflecting challenges with very light or very heavy payloads.

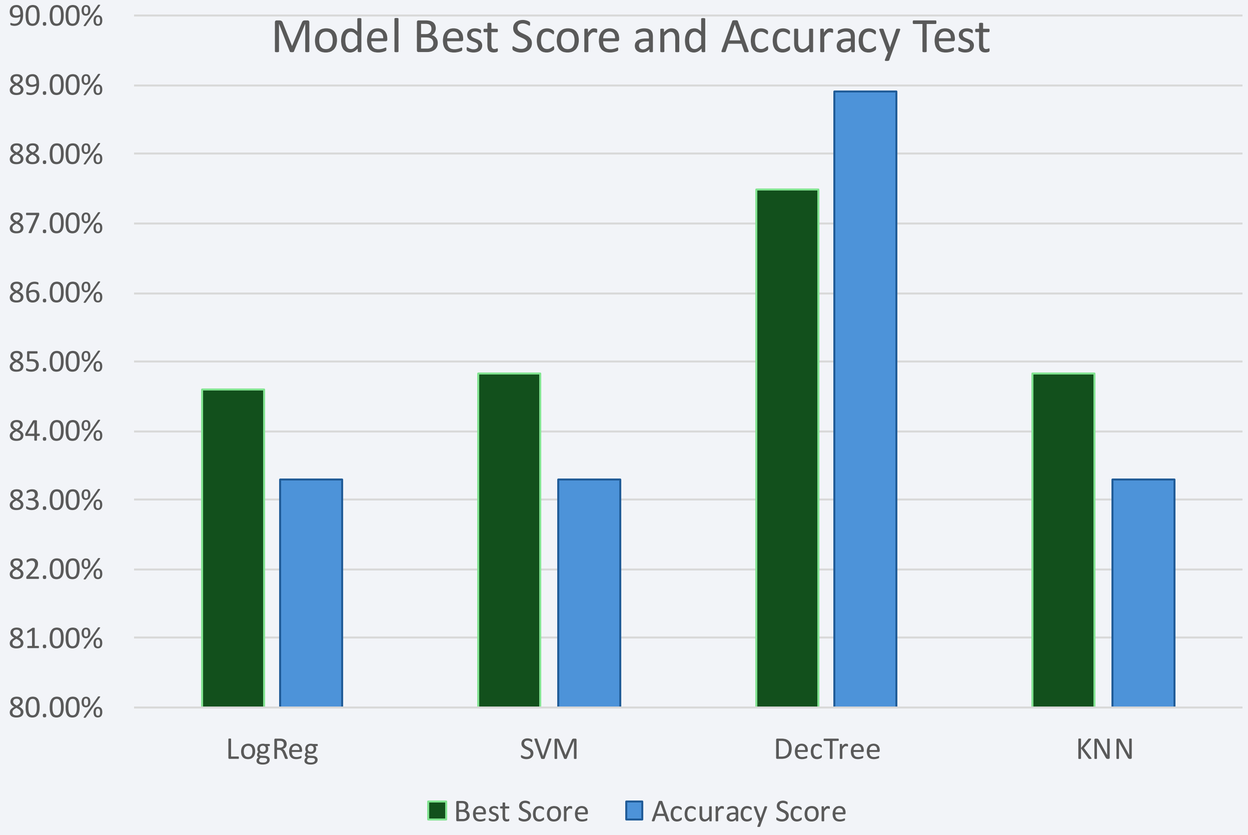
<u>Booster Version Performance:</u> The FT (Falcon 9 Full Thrust) booster version demonstrates the highest launch success rates, indicating improved reliability and performance in recent missions.

<u>Scatter Plot Insights:</u> Successful launches (indicated by 1 on the y-axis) cluster in specific payload ranges and booster versions, while failures (0) are scattered, highlighting operational risk factors.

Section 5

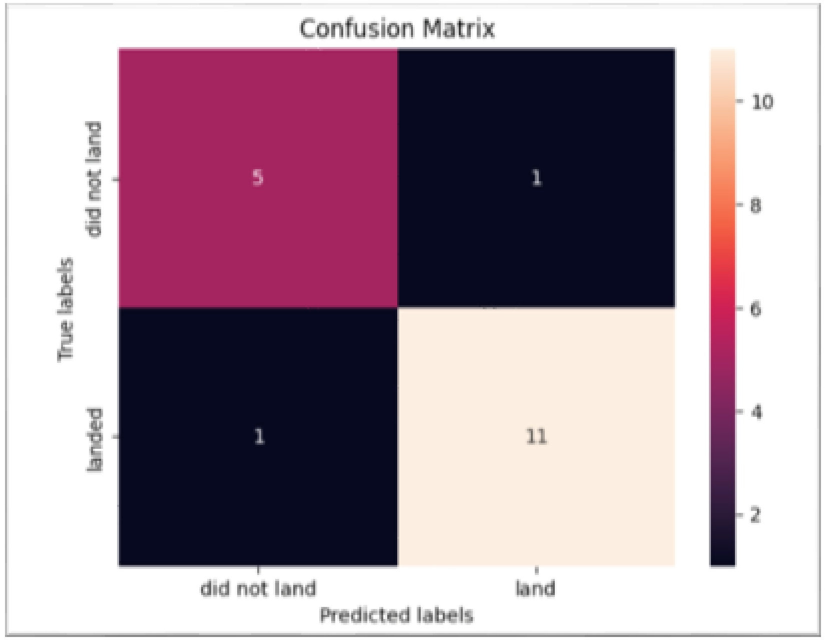
# Predictive Analysis (Classification)

# Classification Accuracy




| Model | Best CV Accuracy | Test Accuracy |
|---|---|---|
| Logistic Regression | 0.8464 | 0.8333 |
| SVM | 0.8482 | 0.8333 |
| Decision Tree | 0.8750 | 0.8889 |
| KNN | 0.8482 | 0.8333 |

The tested models produced the above accuracy results.
The Decision Tree Model had the highest classification accuracy.




# Confusion Matrix

- The Decision Tree model
  - achieved a high classification accuracy of 88.89%.

The confusion matrix reveals that out of 18 total predictions, 16 were correctly classified.

- There were 2 misclassifications
  - 1 false positive (predicted "land" when it actually "did not land").
  - 1 false negative (predicted "did not land" when it actually "landed").

This highlights the model's effectiveness while also showing some minor errors primarily in distinguishing between the two classes.

# Conclusions

Solved Problems:

1. How can I accurately predict whether SpaceX will reuse the Falcon 9 first stage using public data?

   By using the Decision Tree model which achieved 88.9% accuracy in predicting first stage reuse by using publicly available launch data (e.g., flight number, orbit type, payload mass, launch site).

   The confusion matrix shows the model's reliable performance with very few misclassifications, confirming that public data contains strong predictive signals.

# Conclusions

Solved Problems:

2. Which features best indicate the likelihood of first stage reuse?

- Flight Number: Increasing flight numbers correlate with higher first stage reuse success, showing booster reliability improves over time.
- Orbit Type: Missions to orbits like ES-L1, GEO, HEO, SSO have more successful landings; GTO orbits show less predictable reuse patterns.
- Payload Mass: Heavier payloads generally reduce reuse likelihood but with variations depending on orbit and launch site.
- Launch Site: Sites like KSC LC-39A and CCAFS LC-40 show higher reuse success rates, indicating location impacts outcomes.
- Booster Version: The FT booster version shows the highest success rate, reflecting design improvements.

# Conclusions

Solved Problems:

3. How can Space Y leverage these predictions to set competitive launch prices?

By accurately forecasting first stage reuse probability, Space Y can estimate cost savings from reuse vs. new builds.

Knowing the key factors that affect reuse success allows Space Y to price launches more competitively for certain payload sizes, orbits, and launch sites.

This predictive insight supports optimized bidding and operational planning aligned with reliability trends.

# Conclusions

**Solved Problems:**

3. How can Space Y leverage these predictions to set competitive launch prices?

By accurately forecasting first stage reuse probability, Space Y can estimate cost savings from reuse vs. new builds.

Knowing the key factors that affect reuse success allows Space Y to price launches more competitively for certain payload sizes, orbits, and launch sites.

This predictive insight supports optimized bidding and operational planning aligned with reliability trends.

4. How can I build dashboards that help the team make data-driven decisions?

Interactive dashboards(Plotly) visualize key metrics like flight success trends, payload distributions, and reuse predictions.

# Appendix

Link to any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets created during this project.

https://github.com/Leonagarabedian

Thank you!