IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Bernard Joshua Raja Rajan
12th October 2024

# Outline

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

# Executive Summary

This project focused on analyzing SpaceX launch data to identify factors that affect launch costs and predict rocket reusability. The key methodologies and results are summarized below:

**Methodologies:**

- **Data Collection:** Launch data was gathered through the SpaceX API and web scraping from relevant sources.
- **Data Wrangling:** The collected data was cleaned and transformed for analysis.
- **Exploratory Data Analysis (EDA):** Visualized key trends using charts and performed SQL queries to understand launch patterns.
- **Predictive Modeling:** Logistic Regression, SVM, and Decision Tree models were tested to predict the success of rocket landings.

**Key Results:**

- Successful landings were more likely with lighter payloads and specific orbits, such as Low Earth Orbit (LEO).
- KSC LC-39A was identified as the most successful launch site in terms of landing outcomes.
- Logistic Regression and SVM models showed the highest accuracy in predicting successful landings.

# Introduction

The aerospace industry is rapidly changing, and companies need to make smart decisions to stay competitive. Space Y, a new rocket company, wants to optimize its operations by using data to help predict launch costs and improve rocket reusability. In this capstone project for the "Applied Data Science" course, I will act as a data scientist, analyzing public data related to SpaceX launches.

The main goals of this project are to answer two important questions:

- What factors affect the price of launching a rocket?
- Can we predict if SpaceX will reuse the first stage of its rockets?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

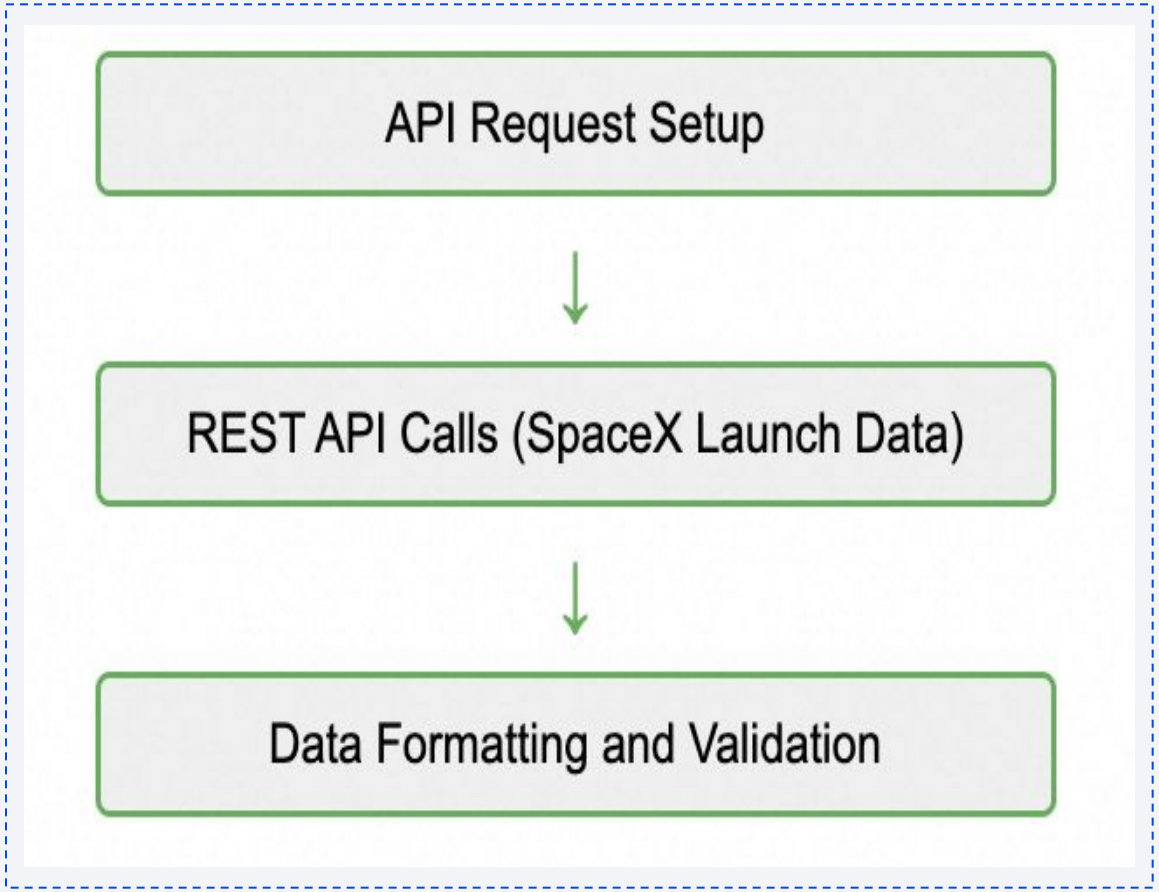
# Data Collection – SpaceX API

**Objective**: The goal was to collect data related to SpaceX Falcon 9 launches and predictions for first-stage landings.

**Collection Process**:

1. **API Connection Setup**: The SpaceX REST API was accessed to fetch launch data.
2. **Data Retrieval**: Multiple REST API calls were made to gather information such as launch dates, payloads, landing outcomes, and rocket configurations.
3. **Data Formatting**: The retrieved data was processed and formatted into structured datasets for analysis.
4. **Data Validation**: Collected data was verified for completeness and accuracy.

https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/00-SpaceX-Data-Collection-API.ipynb

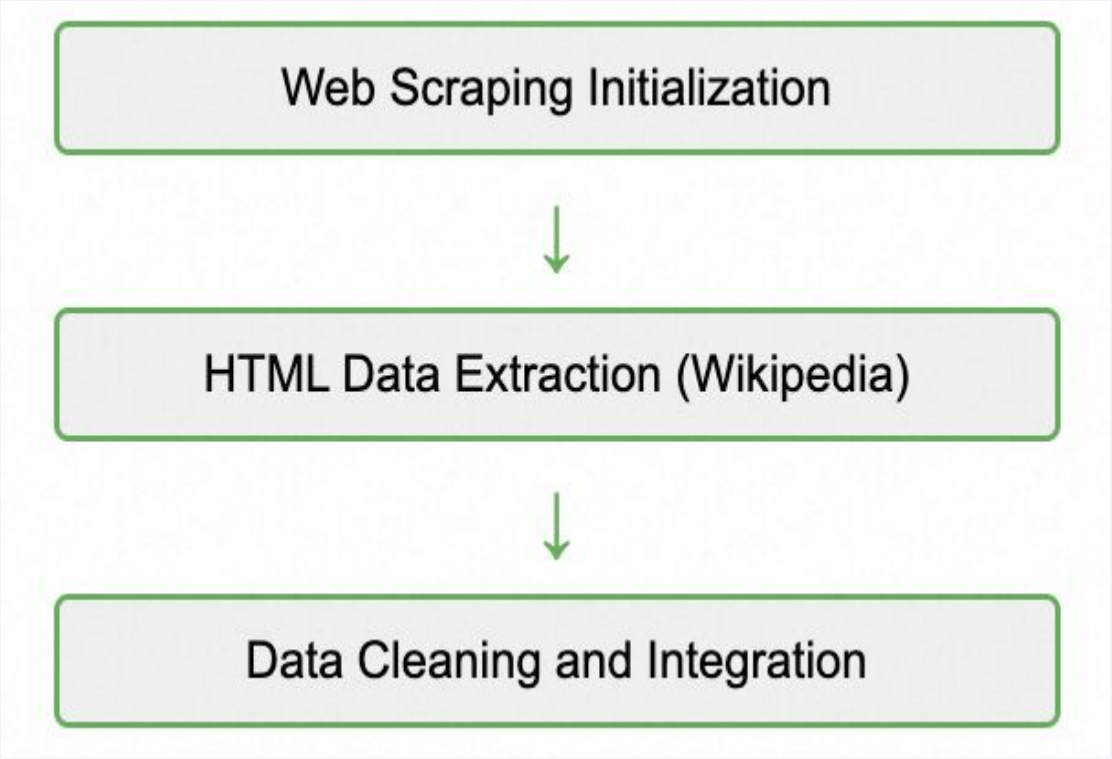


# Data Collection - Scraping

**Target URL**: Data was scraped from the Wikipedia page for "List of Falcon 9 and Falcon Heavy launches."

**Process Steps**:

1. **Accessing the Webpage**: The webpage was accessed using a web scraping library (e.g., BeautifulSoup).
2. **Extracting Data**: Relevant data such as launch dates, rocket names, and mission outcomes was extracted from the HTML table.
3. **Cleaning the Data**: The extracted information was cleaned and converted into a structured format like a DataFrame for further use.
4. **Data Integration**: The cleaned data was integrated with other datasets to create a comprehensive dataset for analysis.




https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/01-Webscraping-SpaceX.ipynb

# Data Wrangling

**Data Loading**:

- The dataset is loaded from a CSV file using `pd.read_csv()`. It contains various details, including numerical and categorical data.

**Data Exploration**:

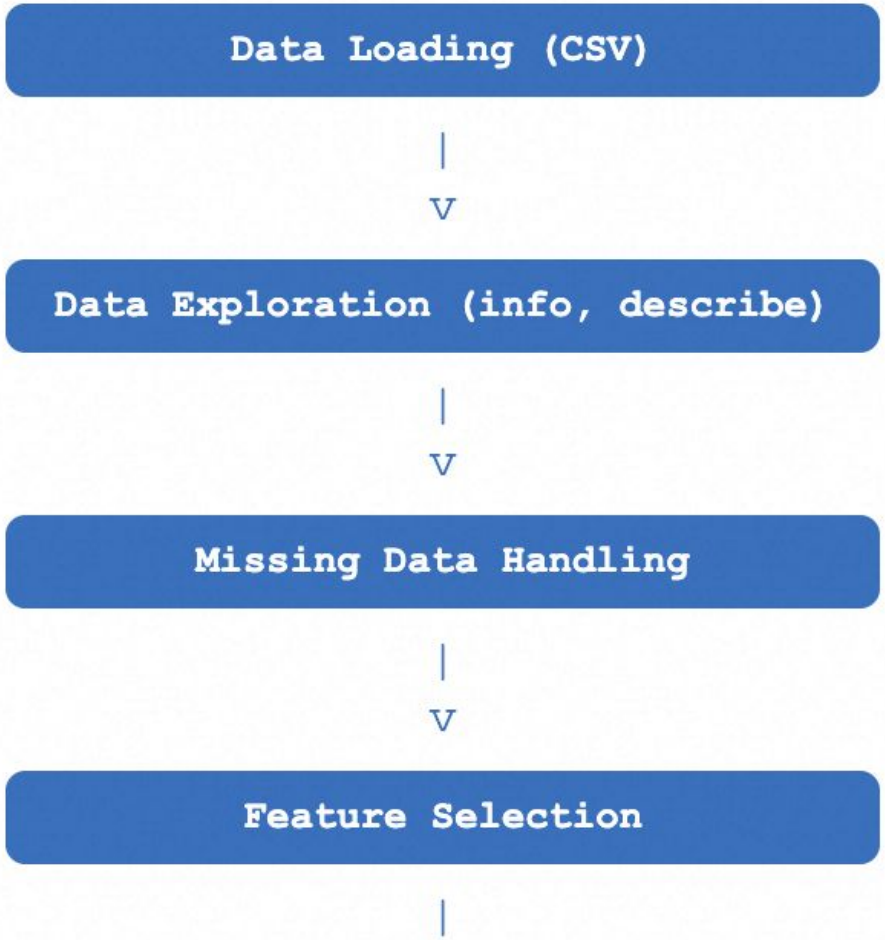
- The data is inspected with `data.info()` and `data.describe()` to understand the structure, data types, and basic statistics.

**Missing Data Handling**:

- Identify missing values using `data.isnull().sum()`.
- Fill missing numerical values with the median of the respective columns using `data.fillna()` or drop rows with missing values if appropriate.

**Feature Selection**:

- Only the necessary columns are selected for further analysis (e.g., filtering columns with `data[['column1', 'column2']]`).



# Data Wrangling

**Data Transformation**:

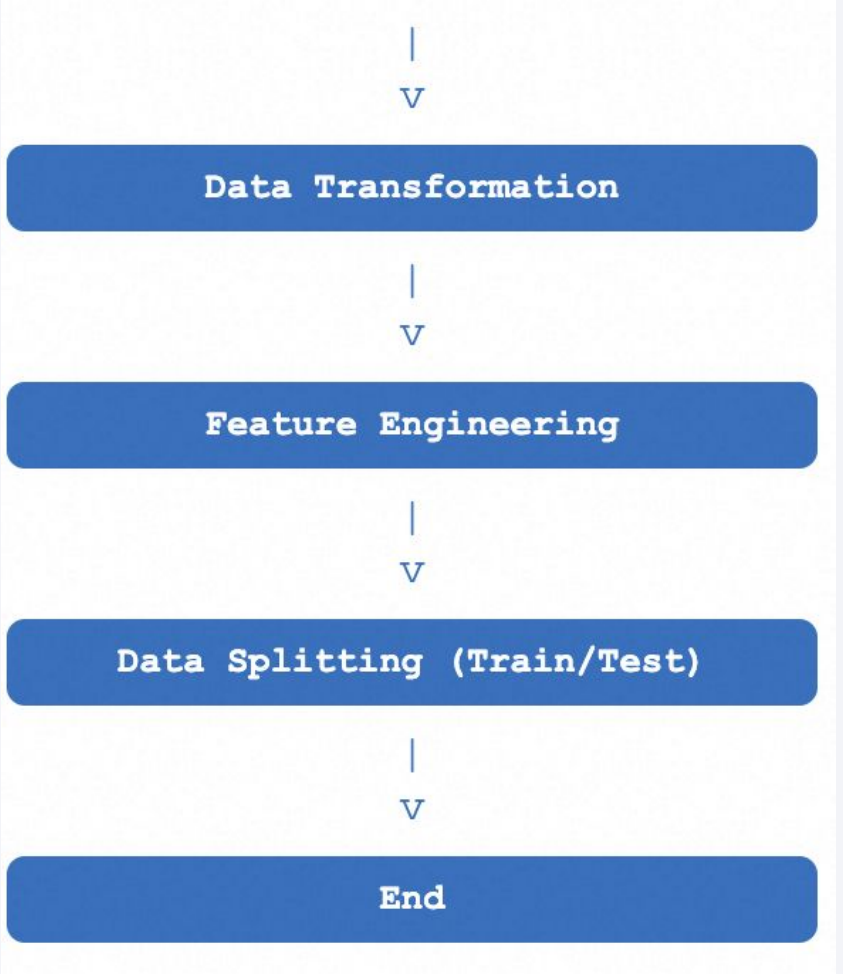
- Categorical variables are encoded using one-hot encoding or label encoding to convert them into numerical formats.
- Numerical features may be normalized or scaled using methods such as Min-Max Scaling or Standardization to ensure they are within a similar range.

**Feature Engineering**:

- New features are created from the existing ones, if needed, for better predictive performance (e.g., creating polynomial features).

**Data Splitting**:

- Split the data into training and test sets using a function like `train_test_split()` to prepare for model training.




https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/02-SpaceX-Data-Wrangling.ipynb

# EDA with Data Visualization

**Flight Number vs. Payload Mass (with Launch Outcome):**

- **Chart Type:** Scatter Plot (Categorical)
- **Purpose:**
  This chart helps identify if an increase in flight attempts improves the success rate. It also examines the impact of payload mass on landing success. As flight numbers increase, successful landings become more frequent, even for heavier payloads.

**Flight Number vs. Launch Site:**

- **Chart Type:** Scatter Plot (Categorical)
- **Purpose:**
  This chart explores the distribution of launches across different sites, showing how frequently each site is used over time. It reveals trends in success rates by site, helping understand which locations may influence launch outcomes.

**Payload Mass vs. Launch Site:**

- **Chart Type:** Scatter Plot (Categorical)
- **Purpose:**
  This chart assesses if payload mass varies across launch sites. It highlights that certain sites (e.g., VAFB-SLC) do not handle heavy payloads, which could influence their respective success rates.

https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/04-EDA-Visualization.ipynb

# EDA with Data Visualization

**Success Rate by Orbit Type:**

- **Chart Type:** Bar Plot
- **Purpose:**
  This chart investigates how different orbits impact launch success. It identifies orbits like LEO with high success rates and others like GTO with mixed outcomes, indicating the challenges associated with certain orbit types.

**Flight Number vs. Orbit Type:**

- **Chart Type:** Scatter Plot (Categorical)
- **Purpose:**
  This chart analyzes whether repeated flight attempts affect success rates across various orbits. It shows a positive trend for LEO with higher success rates as flight numbers increase, while no clear pattern is observed for GTO.

**Payload Mass vs. Orbit Type:**

- **Chart Type:** Scatter Plot (Categorical)
- **Purpose:**
  This chart examines whether payload mass affects the success rate in different orbits. It reveals that heavy payloads have better outcomes in certain orbits like Polar and ISS, but for GTO, outcomes are more unpredictable.

**Launch Success Trend Over the Years:**

- **Chart Type:** Line Plot
- **Purpose:**
  This chart visualizes the yearly trend in success rates, showing how SpaceX's success rate improved significantly over time, especially from 2013 onward. It indicates growing reliability with more launches.

# EDA with SQL

- **SQL Queries Executed**:
  - Loaded SpaceX launch data into an SQLite database.
  - Queried the database to calculate launch success rates by year.
  - Identified correlations between launch cost and landing success.
  - Grouped launch data by rocket type and launch site to analyze trends.
  - Performed aggregate functions (e.g., `COUNT`, `SUM`) to evaluate total launches, successful landings, and reusability factors.

# Build an Interactive Map with Folium

**CircleMarker for Launch Sites**

- Used `folium.CircleMarker` to mark each SpaceX launch site.
- Added `Popup` and `Tooltip` for detailed site information.

**Color-Coded Markers for Outcomes**

- `Icon` colors: Green for successful launches, red for failures.
- Purpose: Differentiate launch outcomes at various locations.

**MarkerCluster for Grouping**

- Applied `folium.plugins.MarkerCluster` to reduce clutter.
- Enhanced map clarity by clustering closely located markers.

https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/05-Geolocation-Visualization.ipynb

# Build a Dashboard with Plotly Dash

1. **Dropdown for Launch Site Selection**
   - A `dcc.Dropdown` component was added to allow users to select a specific launch site or view data from all launch sites.
   - **Reason:** This provides the flexibility to analyze overall trends or focus on a particular site, helping users understand the performance differences across different locations.
2. **Pie Chart for Launch Success Counts**
   - A pie chart was created using `dcc.Graph` and `plotly.express` to display the proportion of successful vs. failed launches. The chart updates dynamically based on the selected launch site.
   - **Reason:** The pie chart gives a clear, visual summary of the launch outcomes for the chosen site, making it easy to see the success rate and compare it across different sites.
3. **Payload Range Slider**
   - A `dcc.RangeSlider` component was added to enable users to filter the data based on the payload mass range.
   - **Reason:** Allows users to focus on specific payload ranges and understand how payload mass influences launch success, helping identify trends or correlations.
4. **Scatter Chart for Payload vs. Launch Success**
   - A scatter plot was implemented to show the relationship between payload mass and launch success, with color coding for successful and failed launches. The chart updates based on both the selected launch site and payload range.
   - **Reason:** The scatter plot provides insight into how payload mass affects the probability of launch success, allowing for a more detailed analysis when compared to just looking at overall success rates.

### Interactivity Explanation:

- **Dropdown Selection (Launch Site) and Payload Range Filtering:**
  - Both the pie chart and scatter plot dynamically update based on the user's selections, allowing for tailored data exploration.
  - **Reason for Interactivity:** Enables a customizable analysis experience, making the dashboard useful for different types of users, from casual viewers to data analysts interested in specific subsets of the data.

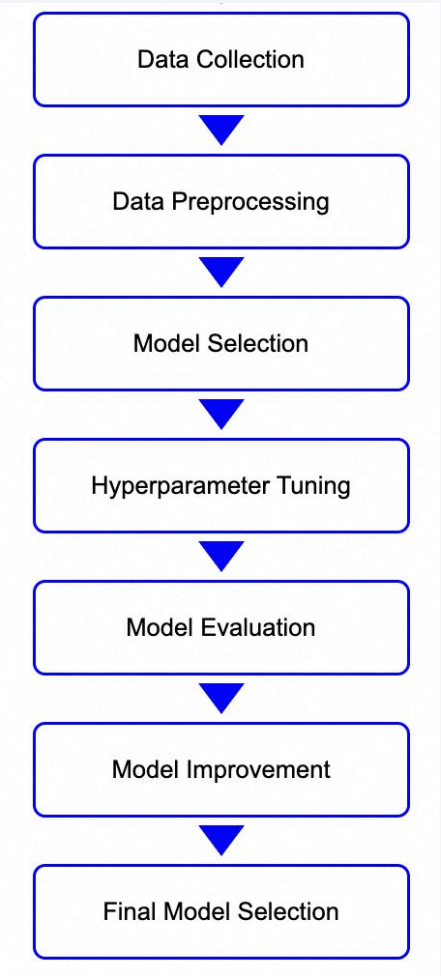


https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/06-SpaceX-Dash-App.py

# Predictive Analysis (Classification)

1. **Data Preparation:**
   - Load data and explore.
   - Standardize features for consistency.
   - Split data into training and testing sets (80% train, 20% test).
2. **Model Selection:**
   - Compare multiple models: Logistic Regression, SVM, Decision Tree.
   - Use `GridSearchCV` for hyperparameter tuning (cross-validation set to 10).
3. **Model Evaluation:**
   - Evaluate performance using accuracy score.
   - Use confusion matrix to identify true positives, false positives, etc.
   - Focus on minimizing false positives.
4. **Improvement:**
   - Fine-tune hyperparameters to improve accuracy.
   - Use best hyperparameters found from `GridSearchCV`.
5. **Best Model Selection:**
   - Choose the model with the highest test accuracy.
   - Use metrics (e.g., accuracy, confusion matrix) to confirm selection.

https://github.com/BernardJoshua/IBM-DS-Capstone-SpaceX/blob/main/07-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

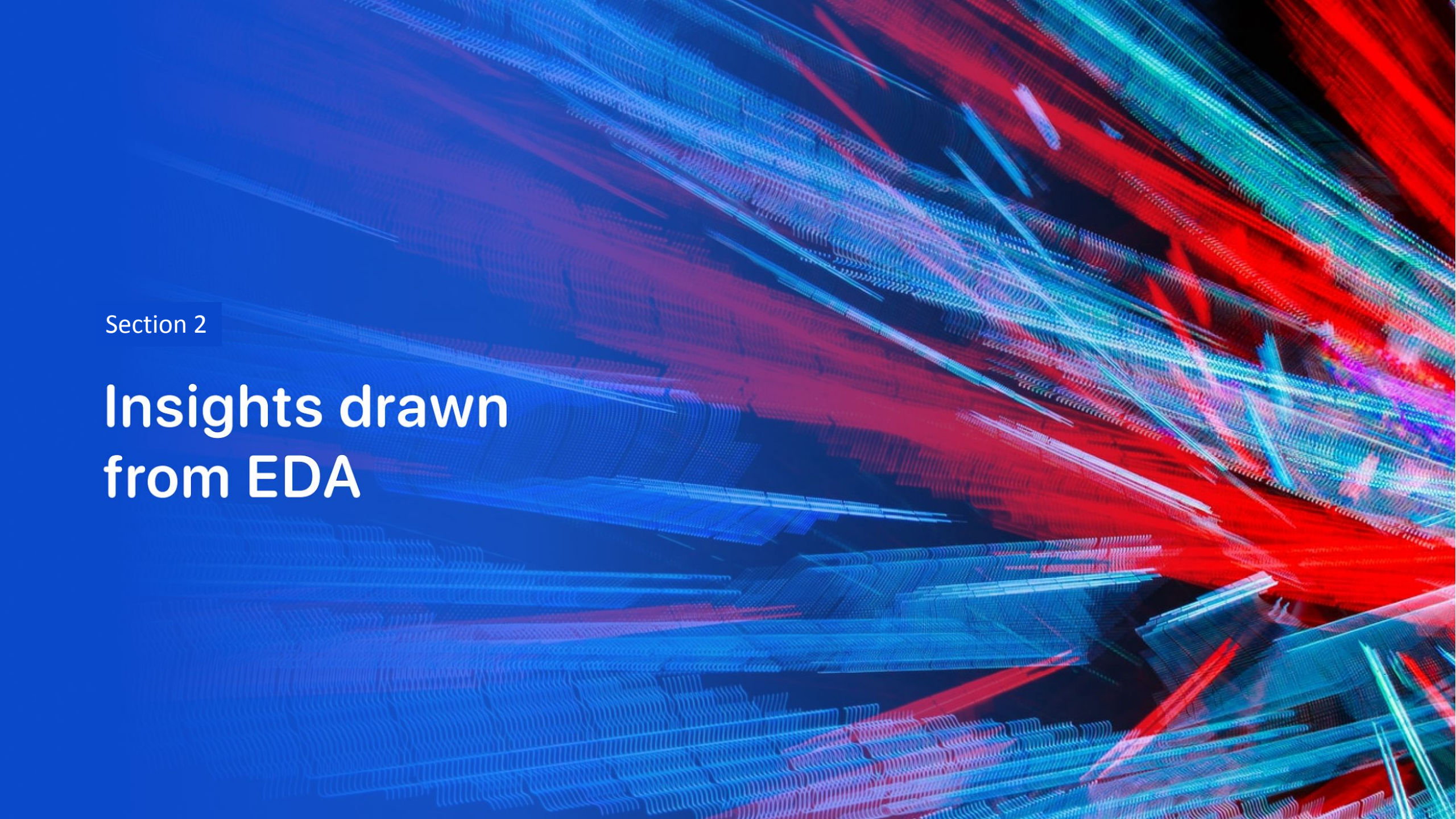


# Results

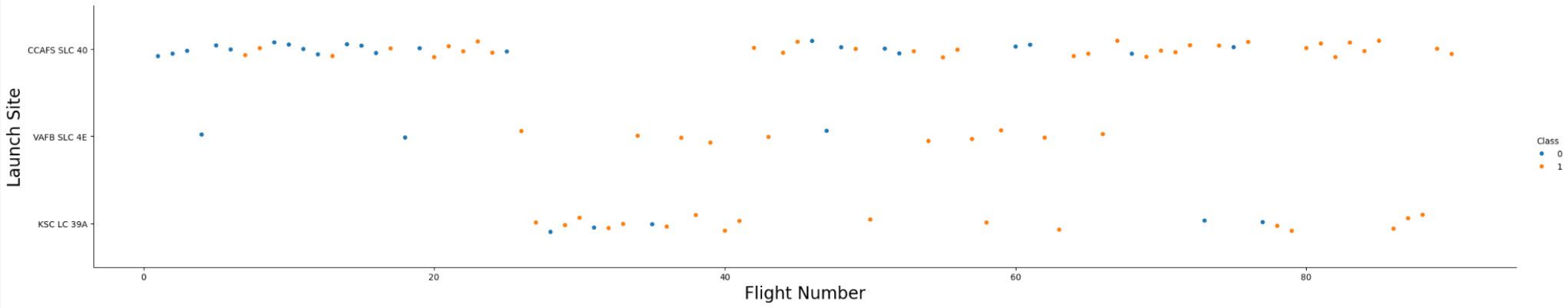
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

# Insights drawn from EDA

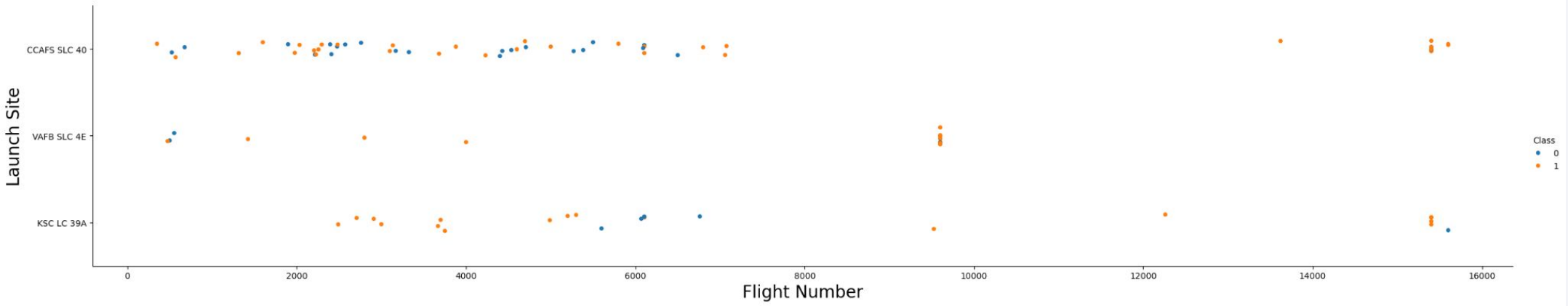


# Flight Number vs. Launch Site




- Ask we can see CCAFS SLC-40 has had the most number of flights , followed by KSC LC-39A and then VAFB SLC-4E.
- Despite CCAFS SLC-40 have more flights the ratio of successful flights to landings is higher for KSC LC-39A. VAFB SLC-4E has the highest ratio but overall the least amount of flights.

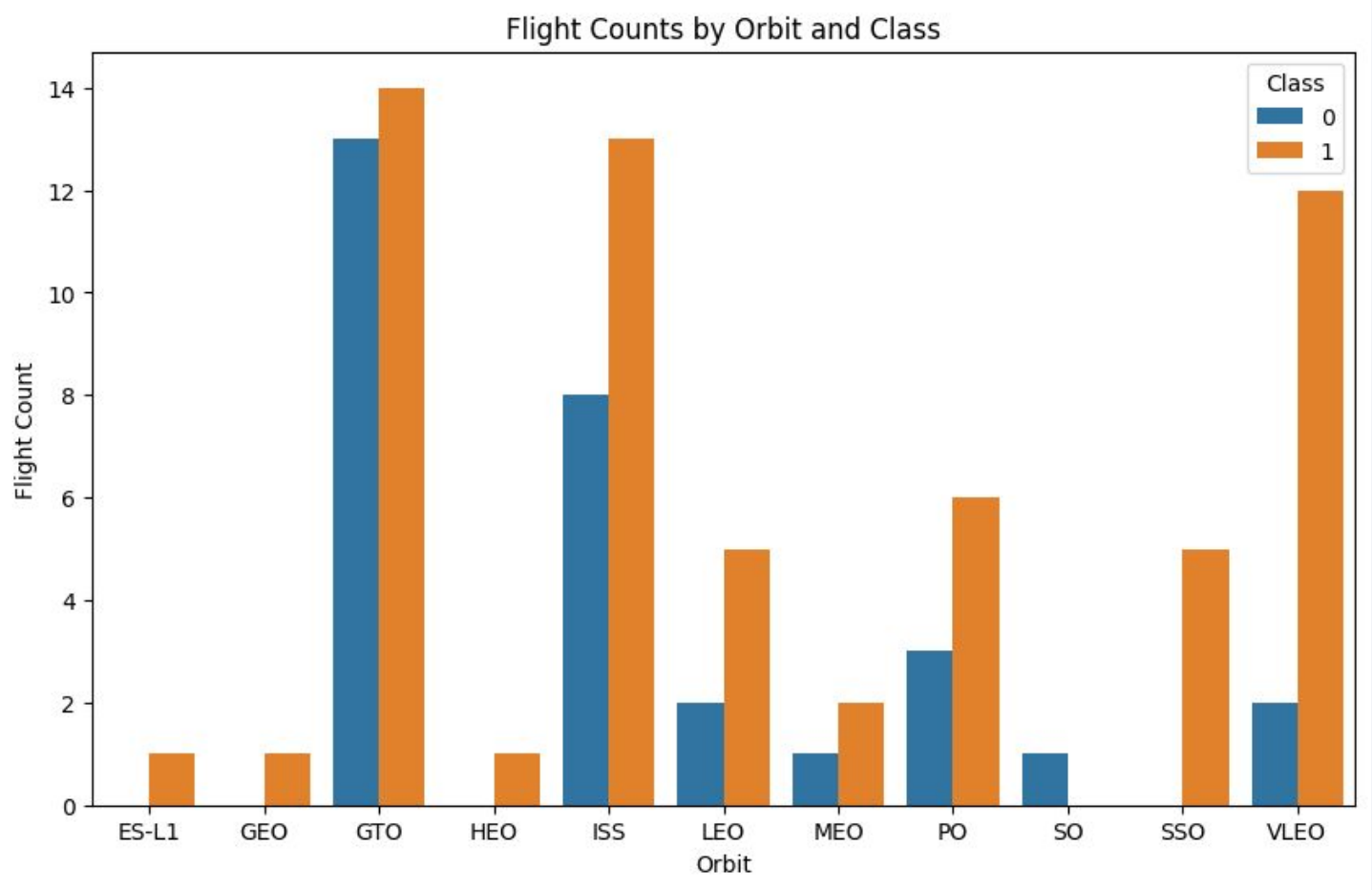
# Payload vs. Launch Site



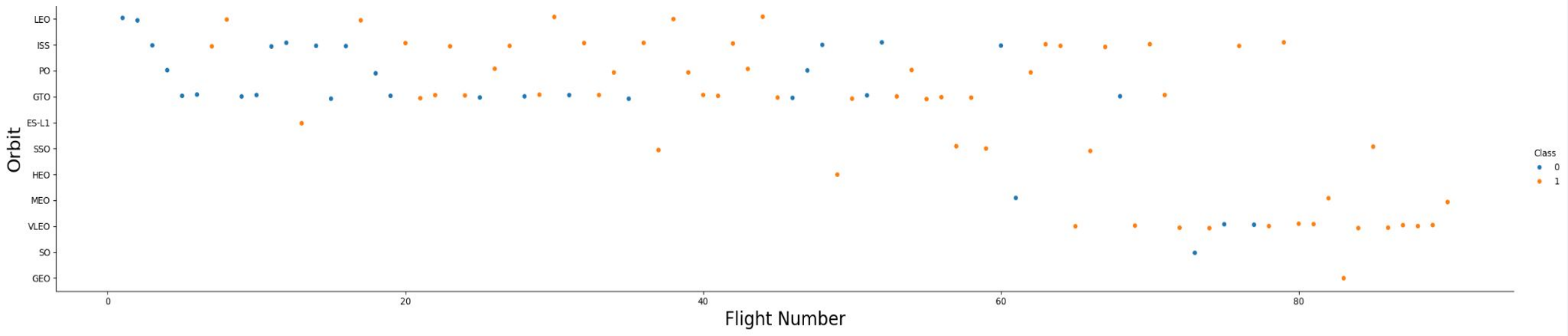

- CCAFS SLC-40 has carried a range of payloads ranging from < 3000 kg all the way up to +/- 17000 kg.
- KSC LC-39A also has had flights carrying a range of payloads but the lightest starts at > 3000kg
- VAFB SLC-4E has carried a smaller range of payloads capped at +/- 9000 kg.
- CCAFS SLC-40 has had a successful flight land carrying a higher payload +/- 17000 kg while a flight with a similar payload did not have a successful landing for KSC LC-39A.

# Success Rate vs. Orbit Type

- Flights that took off to SSO has the highest successful landing ratio, with no failed landings , followed by VLEO.
- Since SSO has less flights than VLEO we can conclude that VLEO flights land successfully the most.

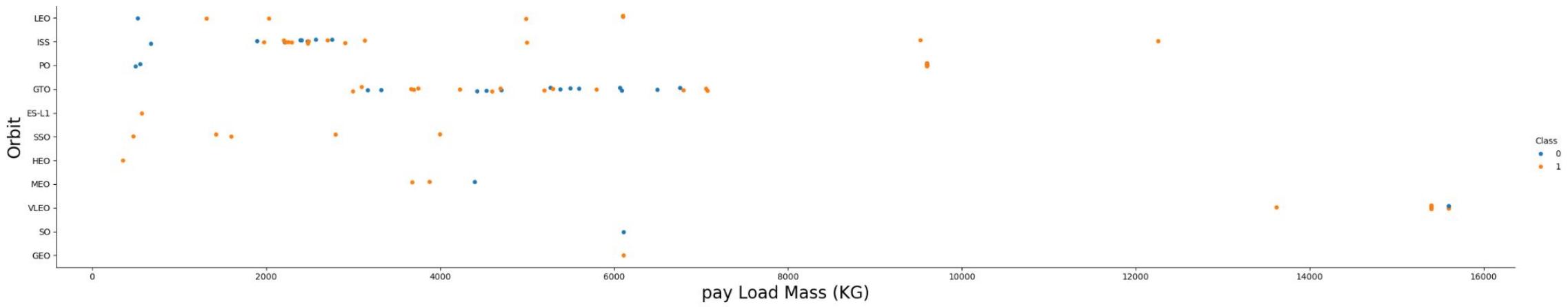


# Flight Number vs. Orbit Type




- 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.

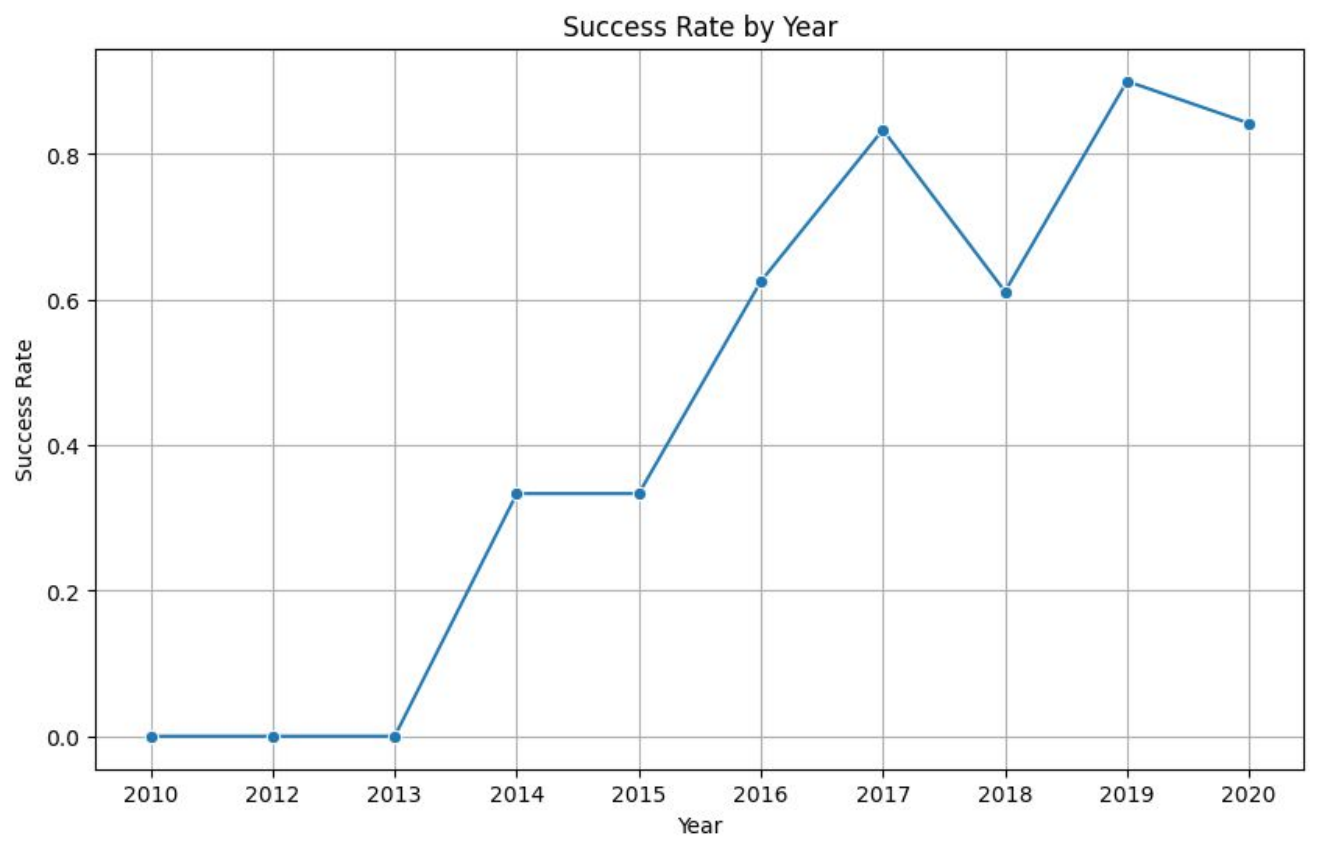
# Payload vs. Orbit Type




- 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




- The success rate since 2013 kept increasing till 2020.

# All Launch Site Names

- There are 4 Launch Sites that SpaceX utilizes:
  - CCAFS LC-40
  - VAFB SLC-4E
  - KSC LC-39A
  - CCAFS SLC-40

# Launch Site Names Begin with 'CCA'

| Date | Time (UTC) | Booster Version | Launch Sites | 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 |

- Early Falcon 9 launches (2010-2013) focused on test and resupply missions.
- Missions targeted Low Earth Orbit, including the ISS.
- Payloads included demo units and NASA cargo.
- All missions were successful, achieving orbit.
- Booster recovery faced challenges, with some failures and no attempts.

# Total Payload Mass

- The total payload mass of 45,596 kg represents the combined weight of cargo delivered across all SpaceX Falcon 9 launches to date.

# Average Payload Mass by F9 v1.1

- The average payload mass of 38,020 kg for the Falcon 9 v1.1 is calculated based on the total payload mass carried across all booster versions. This figure reflects the typical load capacity for missions when using the Falcon 9 v1.1, demonstrating its role in handling a substantial portion of SpaceX's overall launch payloads.

# First Successful Ground Landing Date

- The date of the first successful landing outcome on a ground pad for SpaceX was **December 22, 2015**. This historic achievement occurred during the Falcon 9 Flight 20 mission, marking the first time a booster was recovered intact after launching a payload into orbit, paving the way for reusable rocket technology.

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The following boosters successfully landed on a drone ship and carried payloads weighing more than 4,000 kg but less than 6,000 kg:
  - **F9 FT B1022**
  - **F9 FT B1026**
  - **F9 FT B1021.2**
  - **F9 FT B1031.2**
- These missions show that SpaceX was able to recover its rockets after delivering heavy cargo to space, highlighting the Falcon 9's ability to handle medium-sized payloads while promoting reusable rocket technology.

# Total Number of Successful and Failure Mission Outcomes

- **Total Successful Missions:** 100
- **Total Failed Missions:** 1
- This data shows that out of all the SpaceX missions, 100 were successful, while only 1 resulted in failure. This high success rate demonstrates SpaceX's reliability in launching payloads into space.

# Boosters Carried Maximum Payload

- The booster that has carried the maximum payload mass is:
  - **F9 B5 B1060.3** with a payload mass of **15,600 kg**.
- This booster achieved the highest payload capacity in a single mission, showcasing the Falcon 9's ability to deliver large cargo loads effectively into orbit.

# 2015 Launch Records

The following are the failed landing outcomes on a drone ship in 2015, along with their booster versions and launch site names:

- **January**:
  - **Booster Version**: F9 v1.1 B1012
  - **Launch Site**: CCAFS LC-40
- **April**:
  - **Booster Version**: F9 v1.1 B1015
  - **Launch Site**: CCAFS LC-40

These failures highlight the challenges SpaceX faced while developing its rocket recovery technology during that year.

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

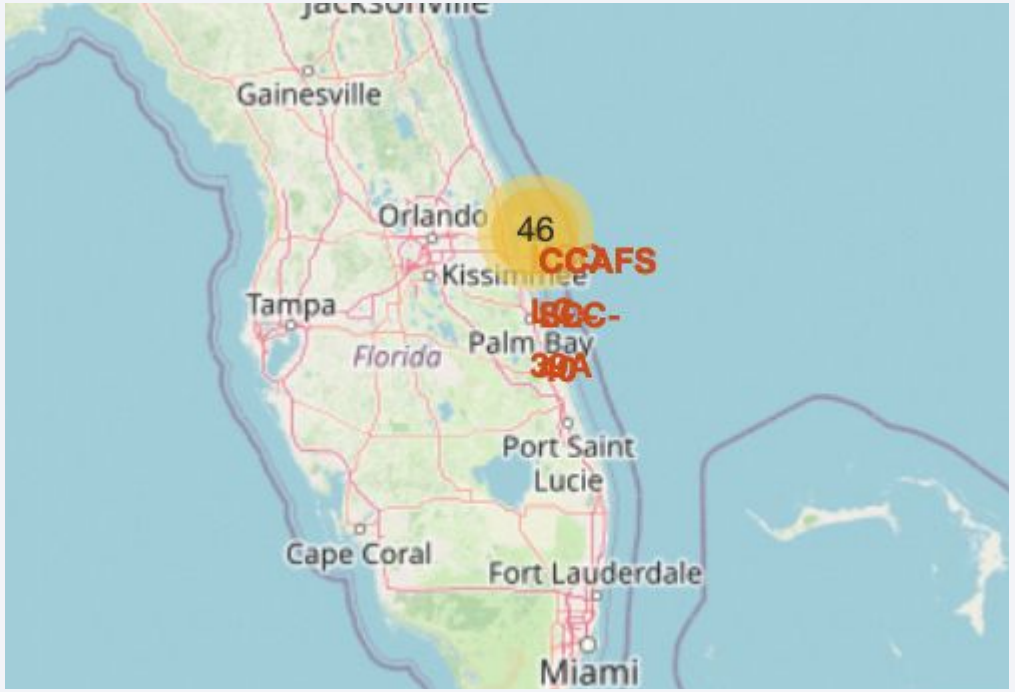
The landing outcomes between June 4, 2010, and March 20, 2017:

- **No attempt**: 10
- **Failure (drone ship)**: 5
- **Success (drone ship)**: 5
- **Controlled (ocean)**: 3
- **Success (ground pad)**: 3
- **Failure (parachute)**: 2
- **Uncontrolled (ocean)**: 2
- **Precluded (drone ship)**: 1

This ranking shows that the most common outcome during this period was "No attempt," indicating several missions did not attempt a landing. The successes and failures on the drone ship are equal, reflecting the mixed results in recovery attempts.

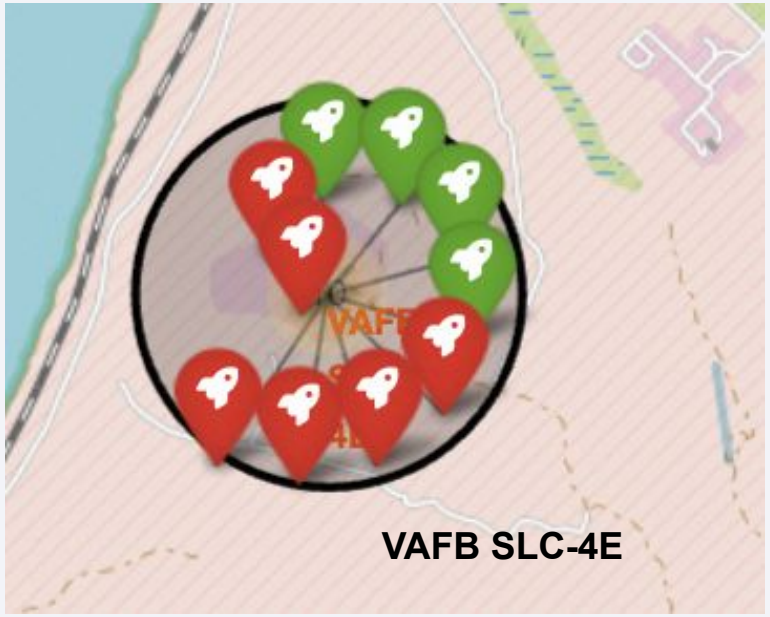

Section 3
Launch Sites
Proximities Analysis

# Locations of Launch Sites

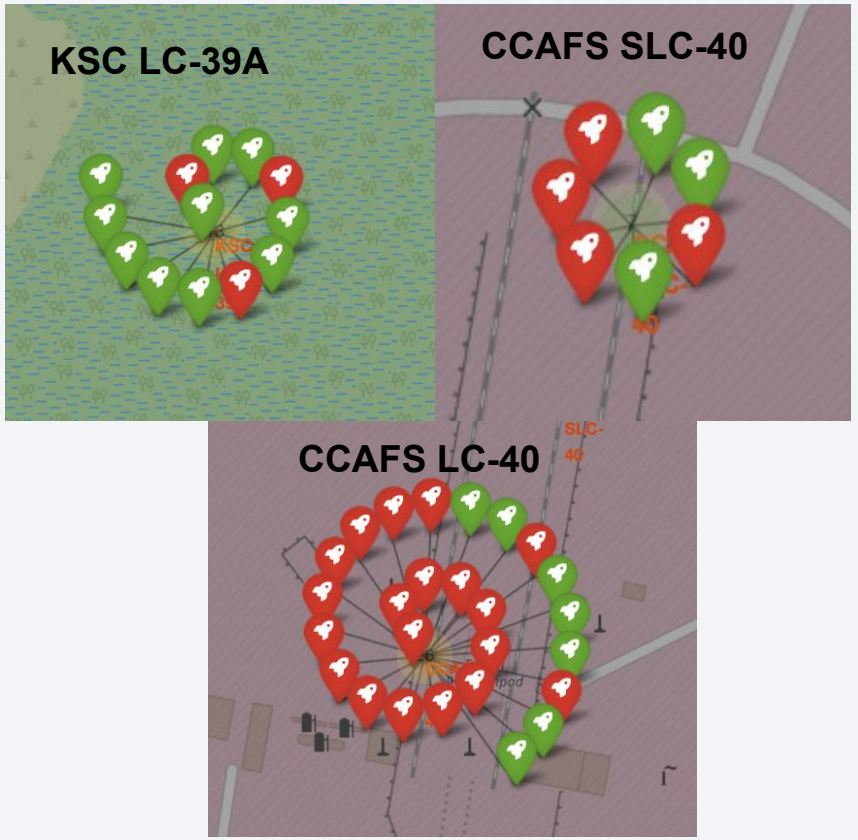




- All launch sites are located in the United States.
- VAFB SLC-4E is located in West Coast in California while the rest are located in the South-Eastern Coast of America in Florida.

# Successful Launches By Locations




West-Coast



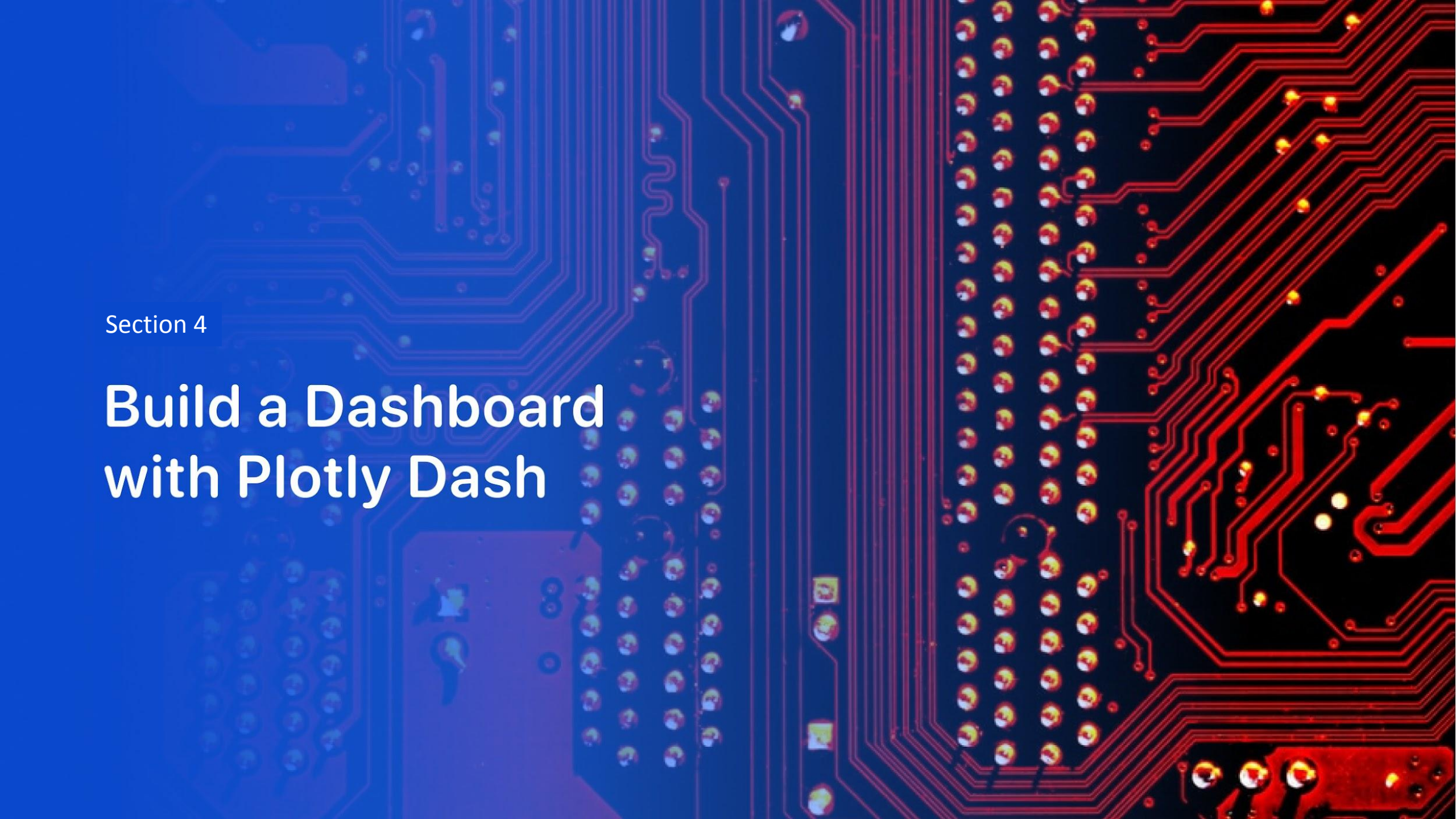

South-Eastern Coast

- Seems like launches in the South-Eastern coast have more successful landings then the ones on the West Coast.

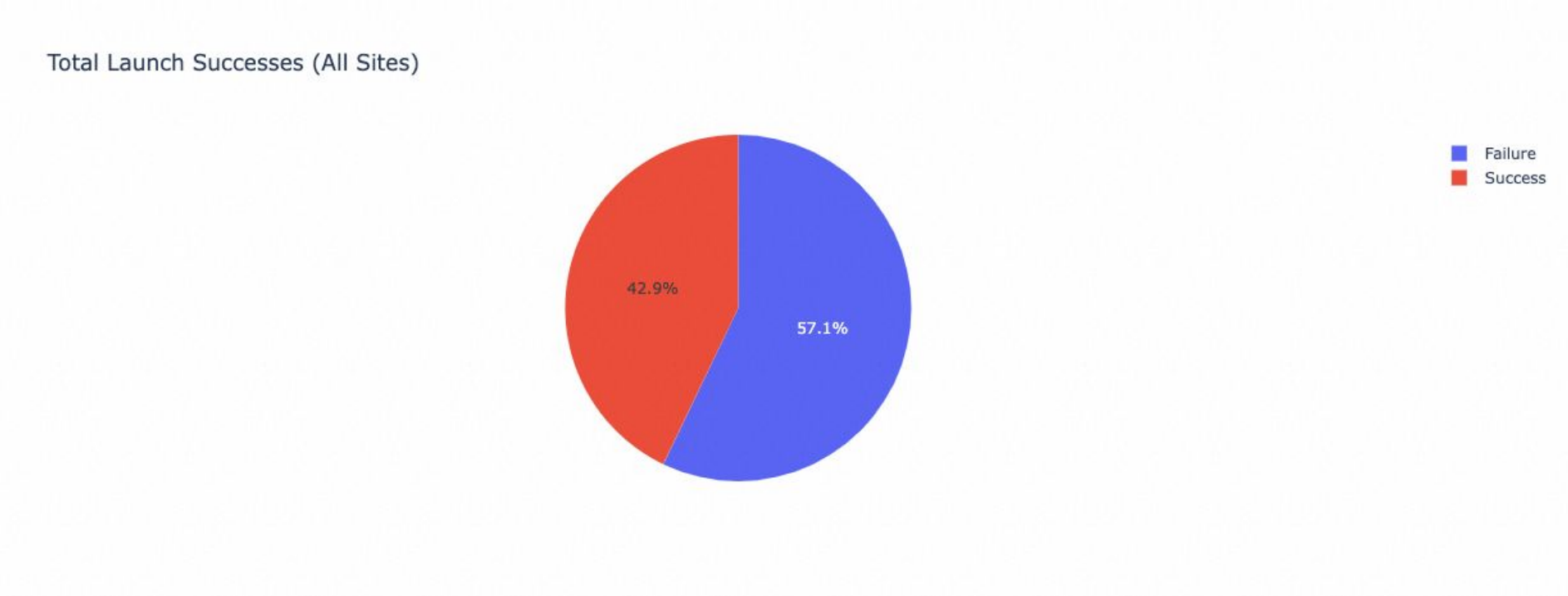
# Proximities and Climate


- The two launch sites CCAFS SLC-40 , CCAFS LC-40, VAFB SLC-4E both are located closer to the coast line. Which could explain why they have more flights and less successful landings (CCAFS SLC-40) and less flights overall (CCAFS LC-40).

- Being closer to the coastline could affect the flights as the weather conditions at these locations may not be as optimal as KSC LC-39A which is more sheltered to the elements.

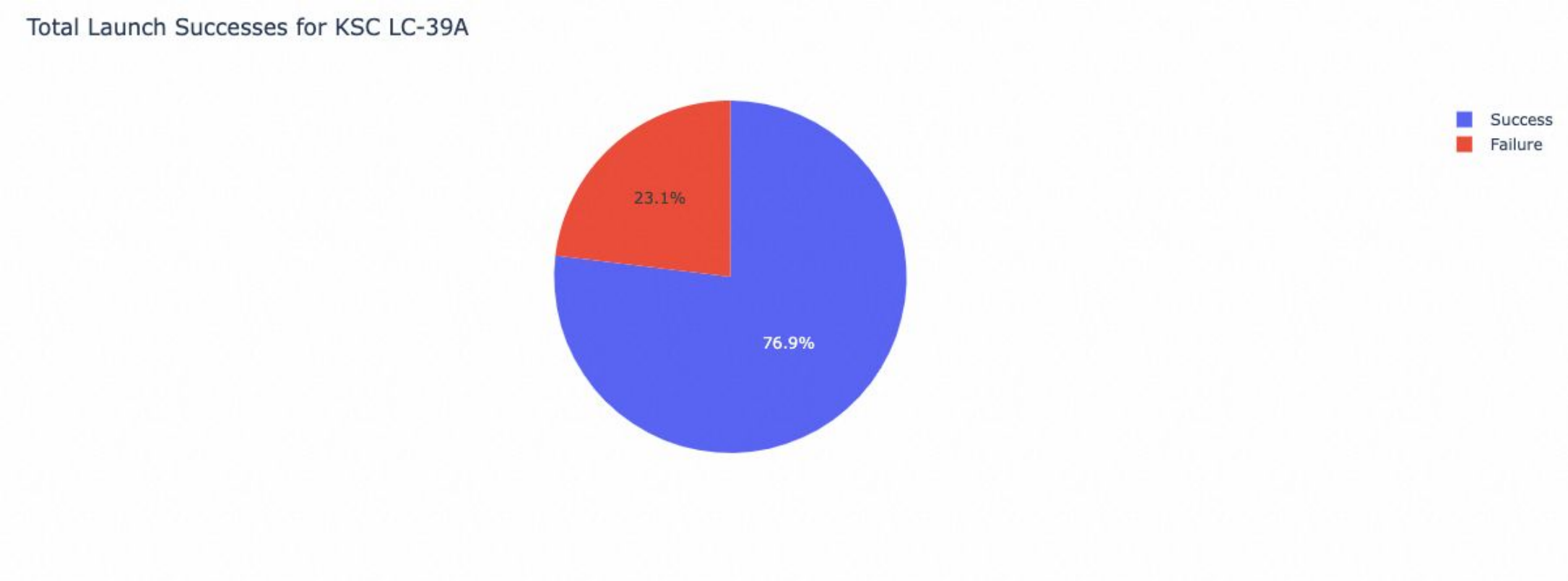

Section 4
Build a Dashboard with Plotly Dash

# Total Lauch Successes (All Sites)



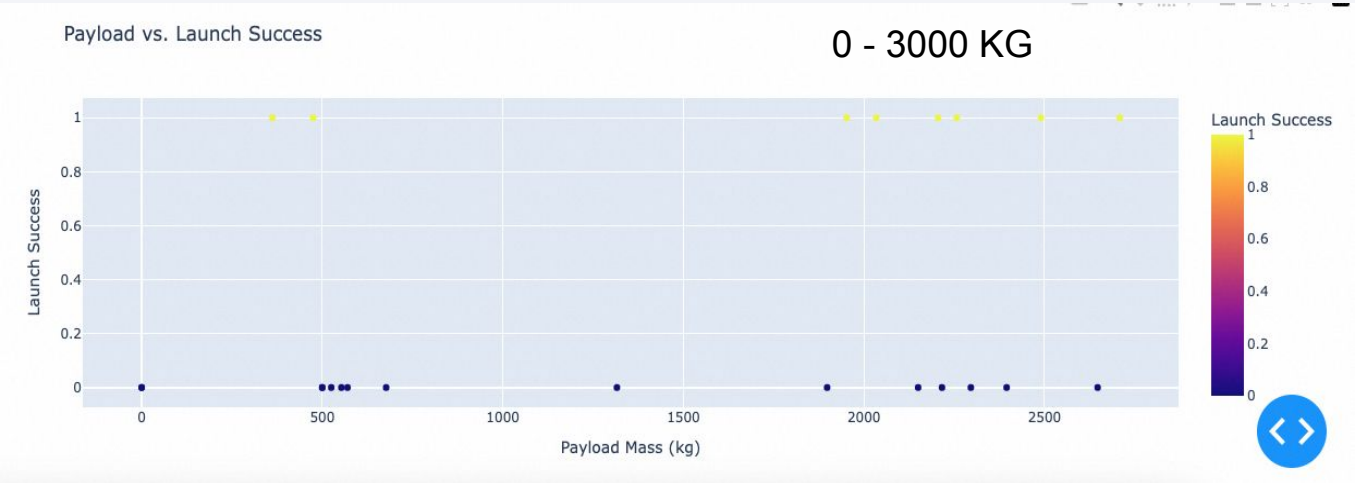

- The overall launch success is around 42.9% which means that more than half of all launches fail.

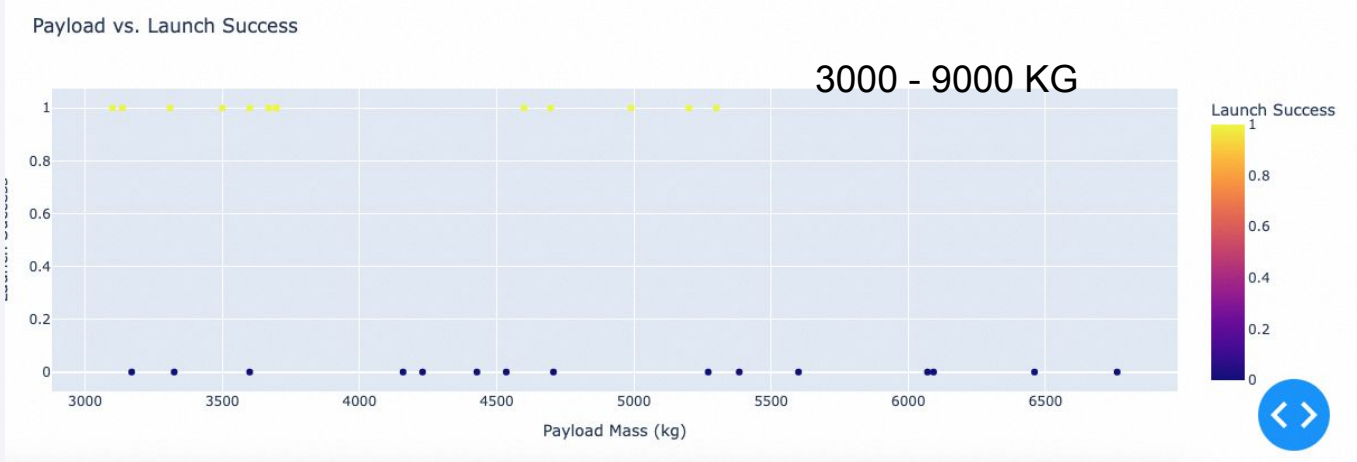
# Most Successful Launch Site.




- KSC LC-39A has the highest success ratio for launches with 10 out of 13 launches being successful.

# Launch Success By Payload Mass



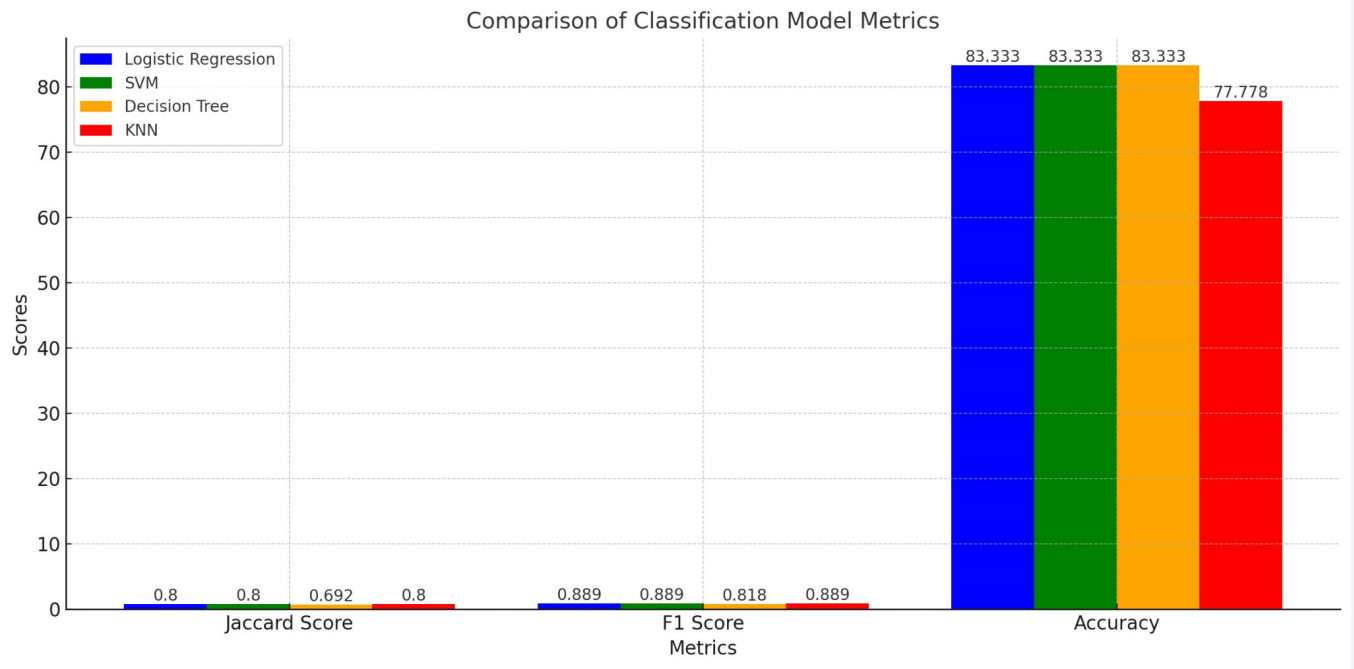

- The heavier the payload the higher the likelihood of launch failure.

Section 5

# Predictive Analysis (Classification)

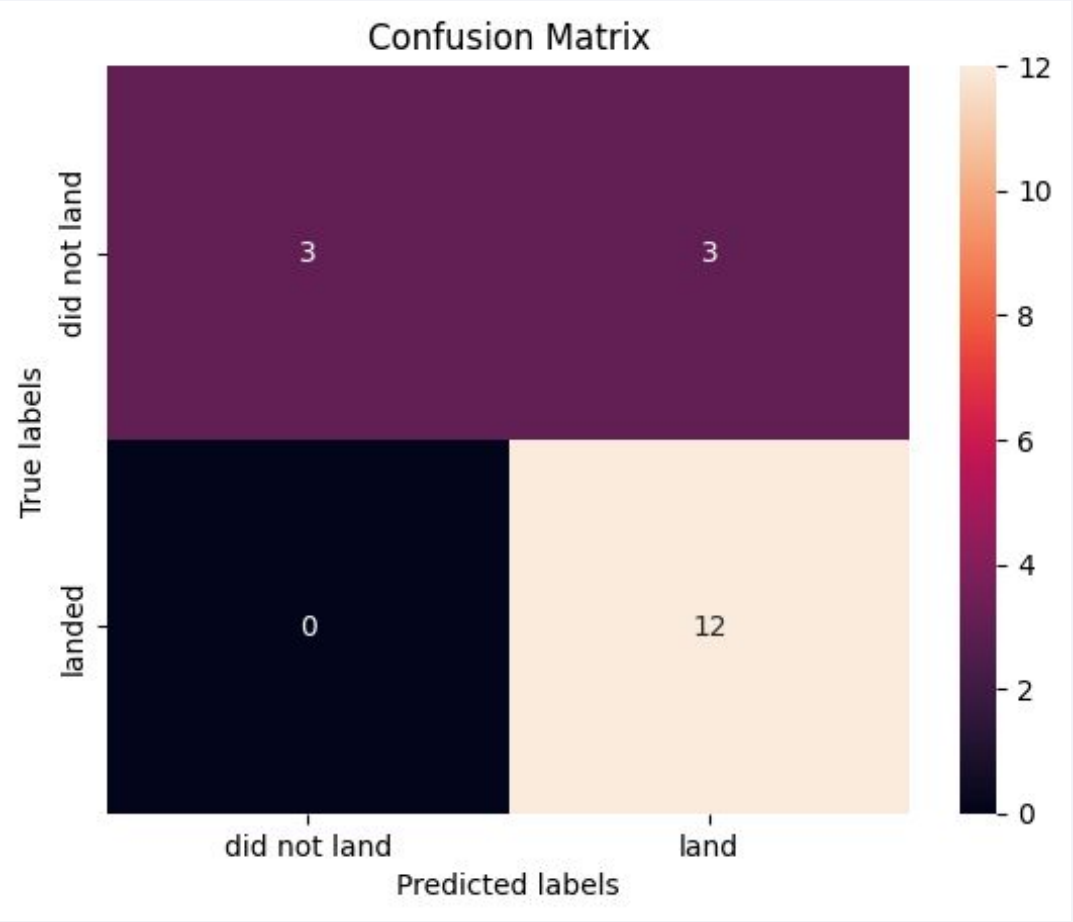
# Classification Accuracy

- Logistic Regression and SVM perform equally well in all three metrics.
- Decision Tree has slightly lower Jaccard and F1 scores but matches in accuracy.
- KNN has good Jaccard and F1 scores but lower accuracy compared to the others.
- The model with the best performing accuracy and metrics are either Logistic Regression or SVM.
- We will be using Logistic Regression.



# Confusion Matrix

- The Logistic Regression Model has managed  to label the did not land outcome correctly (100%), however missed the land outcome by 3 times (25%).
- Other than that the model seems to be performing well.



# Conclusions

- Rocket reuse and launch success depend on factors like payload weight, orbit, and launch site.
- Heavier payloads are more successful in certain orbits (like Low Earth Orbit).
- Some launch sites, like KSC LC-39A, have a higher success rate for landings.
- Logistic Regression and SVM models worked best to predict rocket reuse, helping SpaceX plan future missions and save costs.

# Thank you!