IBM Data Science Capstone

Executive Summary

Summary of Methodology

- Data Collection & Processing:
  - **API and Web Scraping:** Collected historical launch data from SpaceX API and Wikipedia.
  - **Data Normalization:** Transformed data into a structured format.
  - **Filtering & Sampling:** Selected relevant Falcon 9 launches for analysis.
  - **Missing Value Handling:** Filled missing payload mass values with the mean.
  - **Data Cleaning:** Standardized data for modeling.
  - **One-Hot Encoding:** Converted categorical variables into numerical format.
  - **Feature Selection:** Identified key features for predicting landing success.

- Model Development & Evaluation:
  - **Machine Learning Models:** Trained multiple models (Logistic Regression, SVM, Decision Tree, KNN) to classify landing success.
  - **Hyperparameter Optimization:** Tuned model parameters for better accuracy.
  - **Model Comparison:** Compared model performance and identified areas for improvement.

Introduction

- Project Background and Contest
  - The commercial space industry is growing, with companies focusing on reusable rocket technology to make space travel more affordable.
  - SpaceX's Falcon 9 rocket is cost-effective due to its reusable first stage.
  - Not all Falcon 9 missions allow for first-stage recovery.
  - The goal is to build a machine learning model that predicts Falcon 9 first-stage landing success to inform launch cost and planning decisions for Space Y, a fictional competitor.

- Research Questions:
  - Predicting Falcon 9 first-stage landing success: Can we accurately predict the outcome of a Falcon 9 landing? This is significant for cost reduction and strategic planning.
  - Identifying key factors: What are the main factors influencing landing success (e.g., payload weight, weather conditions, launch site)?
  - Model accuracy: How accurate are machine learning models in predicting landing outcomes?
  - Impact of data quality: How does data quality and completeness affect model accuracy?

Methodology

- Data Collection Methodology:
  - Used SpaceX API and web scraping to collect historical launch data.
  - Used additional sources to supplement the data.
- Data Wrangling:
  - Selected relevant features for predicting Falcon 9 first-stage landing success.
- Exploratory Data Analysis (EDA):
  - Used visualization and SQL to understand the data.
- Interactive Visual Analytics:
  - Used Folium and Plotly Dash for interactive visualizations.
- Predictive Analysis:
  - Used classification models to predict landing success.
  - Used GridSearch to optimize hyperparameters.

Data Collection

- Data collection involves gathering and measuring information from specific variables within a defined system. This process allows for answering key questions and evaluating outcomes effectively. In this case, the dataset was obtained through two methods: REST API and web scraping from Wikipedia.
- REST API Method
  - The process began by sending a GET request to the API. The response content was then decoded into JSON format. Using the json_normalize() function, we converted this JSON data into a structured pandas DataFrame. Afterward, we performed data cleaning, checking for any missing values and filling them as necessary to ensure consistency.
- Web Scraping Method
  - For web scraping, we utilized BeautifulSoup to extract launch records presented as an HTML table. The table was parsed and then converted into a pandas DataFrame, enabling us to analyze the data further.

Data Collection

SpaceX API

- Access Data: Retrieve SpaceX launch data through their API in JSON format.
- Parse Data: Extract key details like launch site, booster version, and payload.
- Normalize Data: Flatten hierarchical JSON using json_normalize() for easy analysis.
- Select Key Features: Focus on attributes like launch date, mission name, and landing outcomes.
- Convert to DataFrame: Organize data into a Pandas DataFrame for easy manipulation.
- Filter for Falcon 9: Limit data to Falcon 9 launches, excluding others.
- Handle Missing Values: Replace missing values (e.g., PayloadMass) with the column mean for completeness.

Data Collection

Scraping

- Request Wikipedia HTML: Retrieve the HTML content of the relevant SpaceX launch data page.
- Parse with BeautifulSoup: Use BeautifulSoup and the html5lib parser to process the HTML content for easy data extraction.
- Locate Launch Table: Identify the HTML table containing the SpaceX launch data.
- Initialize Dictionary: Create an empty dictionary to store the extracted data with appropriate keys.
- Extract Data to Dictionary: Loop through table cells and populate the dictionary with relevant data.
- Convert to DataFrame: Convert the dictionary into a Pandas DataFrame for easier analysis and manipulation.

Data Wrangling

- Data Cleaning and Unification: Standardized and cleaned the dataset, ensuring consistency and data integrity.
- Exploratory Data Analysis (EDA):
  - Launch Site Analysis: Calculated the number of launches per site to identify trends.
  - Mission Outcome Analysis: Analyzed mission outcomes across orbit types.
  - Landing Outcome Labeling: Created a new "landing outcome" variable from the "outcome" column for easier analysis.
- Export: Saved the final cleaned dataset to a CSV for further exploration and modeling.

EDA with Data Visualitzation

- Visualization:
  - Scatter Plots: Used to explore the relationships between attributes and identify factors affecting landing outcomes.
  - Bar Charts: Visualized the distribution of mission outcomes across orbit types to identify orbits with higher success rates.
  - Line Plots: Tracked trends in launch success over time to spot potential patterns or anomalies.
- Feature Engineering:
  - Categorical Variable Encoding: Applied one-hot encoding to convert categorical variables (launch site, orbit type, landing outcome) into numerical format for machine learning compatibility.

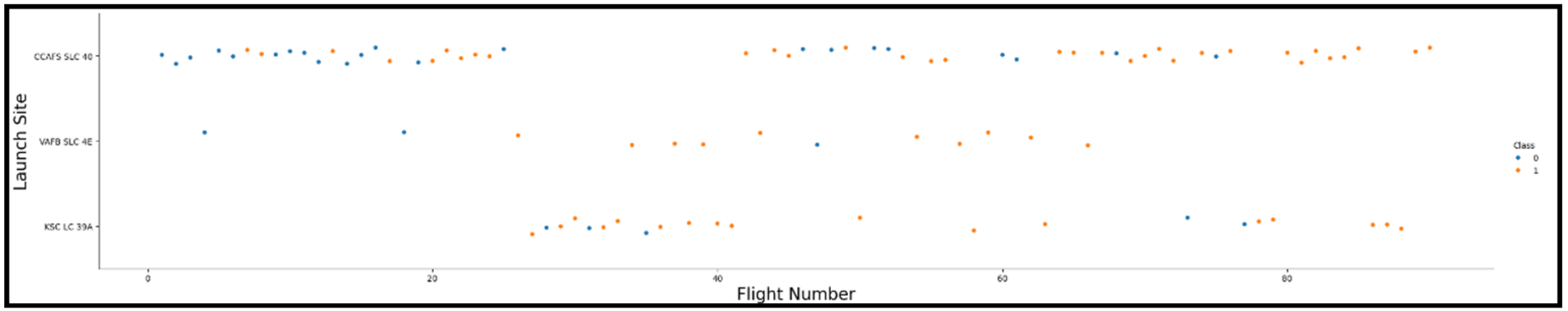
EDA with Data Visualitzation

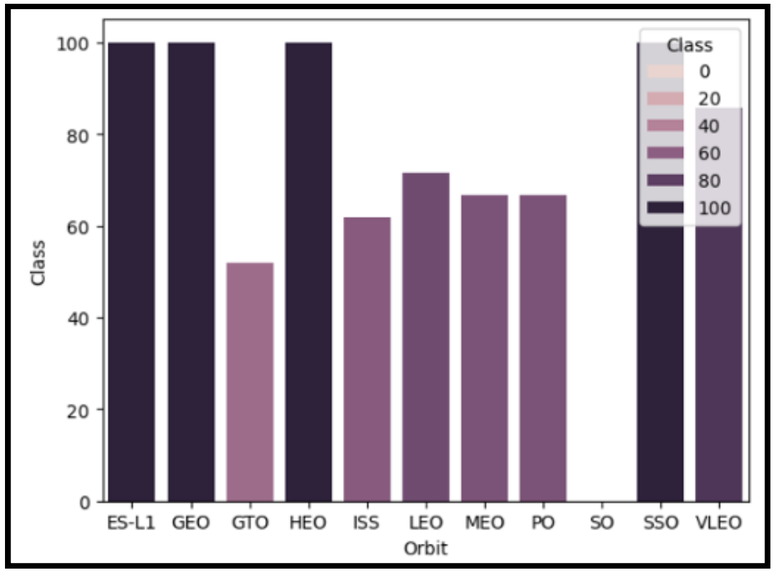
- Visualization:
  - Scatter Plots: Used to explore the relationships between attributes and identify factors affecting landing outcomes.
  - Bar Charts: Visualized the distribution of mission outcomes across orbit types to identify orbits with higher success rates.
  - Line Plots: Tracked trends in launch success over time to spot potential patterns or anomalies.
- Feature Engineering:
  - Categorical Variable Encoding: Applied one-hot encoding to convert categorical variables (launch site, orbit type, landing outcome) into numerical format for machine learning compatibility.

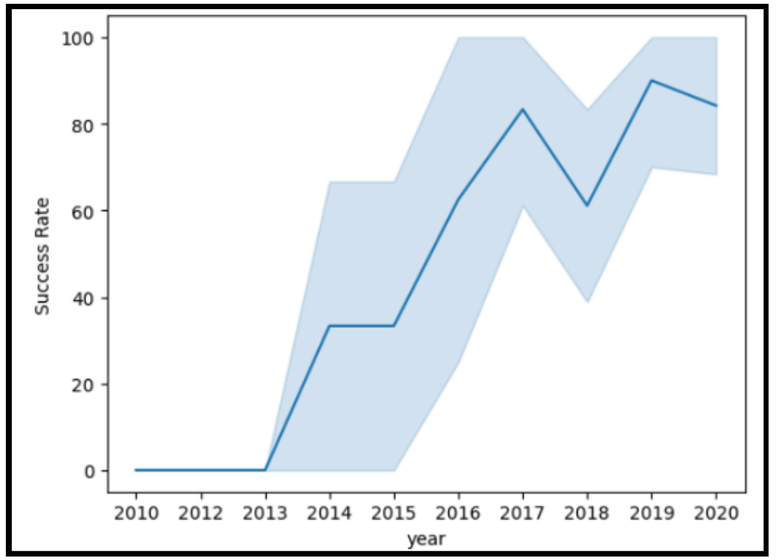
EDA with Data Visualitzation

- Flight Number vs. Launch Site (Scatter Plot): Analyzed the distribution of launches across different sites over time.
- Payload Mass vs. Launch Site (Scatter Plot): Investigated the relationship between payload mass and launch site.
- Success Rate vs. Orbit Type (Bar Chart): Studied the success rate across different orbit types.
- Payload Mass vs. Flight Number (Scatter Plot): Explored the correlation between payload mass and mission sequence.
- Flight Number vs. Orbit Type (Scatter Plot): Examined the evolution of orbit types across missions.
- Payload vs. Orbit Type (Scatter Plot): Studied the relationship between payload mass and target orbit.
- Launch Success Yearly Trend: Analyzed launch success trends over time.

EDA with Data Visualitzation







EDA with SQL

- Data Integration and Analysis:
  - Database Integration: Integrated the dataset into an IBM DB2 database for centralized storage and easy querying.
  - Launch Site Analysis: Queried unique launch sites and their corresponding mission outcomes.
  - Payload Analysis: Analyzed the distribution of payload sizes across different customers and missions.
  - Booster Analysis: Investigated the various booster versions used in the launches.
  - Landing Outcome Analysis: Examined the frequency and types of landing outcomes.
  - Using SQL queries, we efficiently extracted insights from the dataset, enabling deeper analysis.

EDA with SQL

- Launch Site Exploration:
  - Identified unique launch site names.
  - Filtered for sites starting with "CCA".
- Payload Analysis:
  - Calculated the total payload mass for NASA CRS missions.
  - Determined the average payload mass for F9 v1.1 boosters.
- Landing Outcome Analysis:
  - Identified the date of the first successful ground pad landing.
  - Filtered for boosters with successful drone ship landings and specific payload mass ranges.
  - Analyzed the total number of successful and failed missions.
  - Identified booster versions with the maximum payload mass.
  - Analyzed failed drone ship landings in 2015 by booster version and launch site.
  - Ranked landing outcomes between 2010-06-04 and 2017-03-20 in descending order.

- These SQL queries laid the groundwork for data visualization and modeling.

Build an Interactive Map with Folium

- Interactive Map Visualization:
  - Geocoding Launch Sites: Plotted launch sites on an interactive map using latitude and longitude coordinates with circle markers.
  - Visualizing Launch Outcomes: Mapped launch outcomes (success = 1, failure = 0) to different marker colors (green for success, red for failure). Used MarkerCluster() for efficient visualization of multiple markers.

- Spatial Analysis:
  - Distance Calculations: Used the Haversine formula to calculate distances between launch sites and key landmarks:
  - Railways
  - Highways
  - Coastlines
- Nearby cities
- Key Questions:
  - How close are launch sites to transportation infrastructure?
  - What is the spatial relationship between launch sites and urban areas?

Build a Dashboard with Plotly Dash

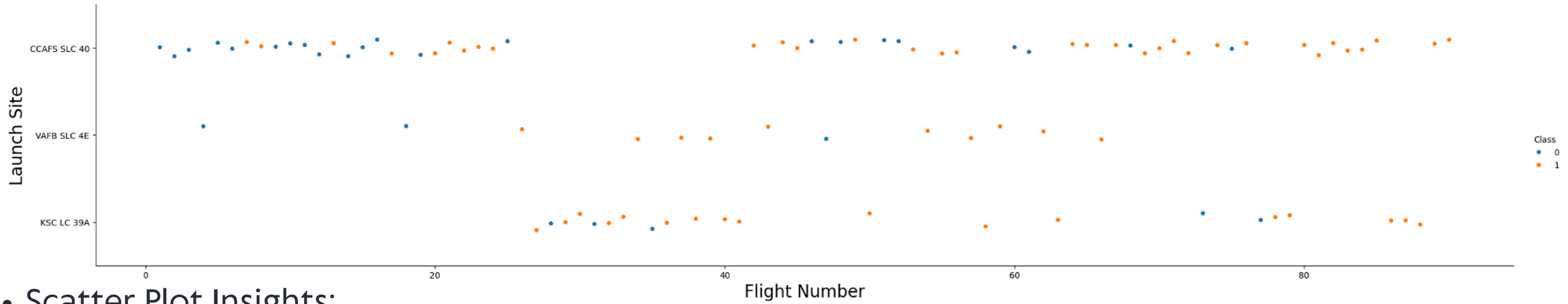
- Interactive Dashboard:
  - Platform: Developed using Plotly Dash for dynamic and user-friendly data exploration.
  - Key Visualizations:
    - Pie Charts: Showed the distribution of launches across different launch sites for a clear site activity overview.
    - Scatter Plots: Examined the relationship between mission outcome and payload mass across booster versions, highlighting trends and correlations.
  - The dashboard enables users to interact with the data, explore trends, and uncover insights about space launch patterns.

Predictive Analysis for Classification

- Data Preparation and Model Selection:
  - Data Loading: Imported the dataset into NumPy and Pandas for efficient manipulation.
  - Data Transformation and Splitting: Preprocessed the data and split it into training and test sets.
  - Algorithm Selection: Chose appropriate machine learning algorithms based on the problem and dataset characteristics.
  - Hyperparameter Tuning: Used GridSearchCV to optimize hyperparameters for each algorithm.
- Model Evaluation:
  - Performance Metrics: Assessed models using accuracy, precision, recall, and F1-score.
  - Hyperparameter Insights: Analyzed tuned hyperparameters to determine optimal configurations.
  - Confusion Matrix Analysis: Visualized confusion matrices for detailed classification performance.
- Model Improvement:
  - Feature Engineering: Created new features to improve model performance.
  - Algorithm Tuning: Refined hyperparameters and tested alternative algorithms to enhance accuracy.
- Model Selection:
  - Chose the model with the highest accuracy and best performance across metrics for final predictions on unseen data.

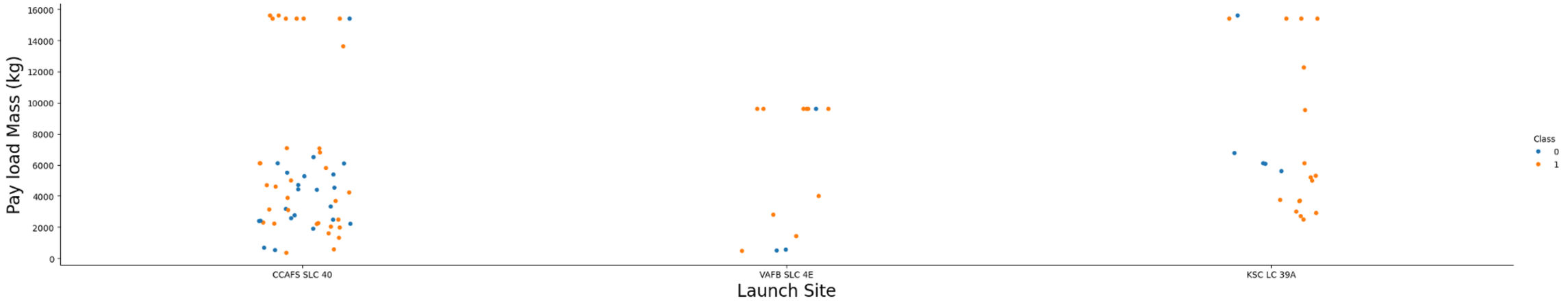
- Expected Outcomes:
  - Exploratory Data Analysis (EDA) Results: A report summarizing key findings, data quality assessments, statistical summaries, and visualizations.
  - Interactive Analytics Demo: Screenshots of an interactive dashboard (using Plotly Dash or Tableau) to allow dynamic data exploration and insights.
  - Predictive Analysis Results: A detailed report on predictive modeling, including model selection, feature engineering, training, evaluation, and performance metrics, with insights into predictive capabilities.

Flight Number vs Launch Site



- Scatter Plot Insights:
  - Positive Correlation Between Flight Number and Success Rate: As the number of flights increases at a specific launch site, the success rate improves, suggesting that experience and technological advancements contribute to mission success.
  - Significant Breakthrough Around Flight 20: A noticeable shift in success rates around flight number 20, likely due to a major technological breakthrough or improved procedures.
  - CCAFS SLC40 – Dominant Launch Site: CCAFS SLC40 has the highest launch volume but shows a weaker correlation between flight number and success rate, indicating other influencing factors beyond experience.
- Further analysis is needed to explore the specific factors affecting success rates across different sites and flight numbers.

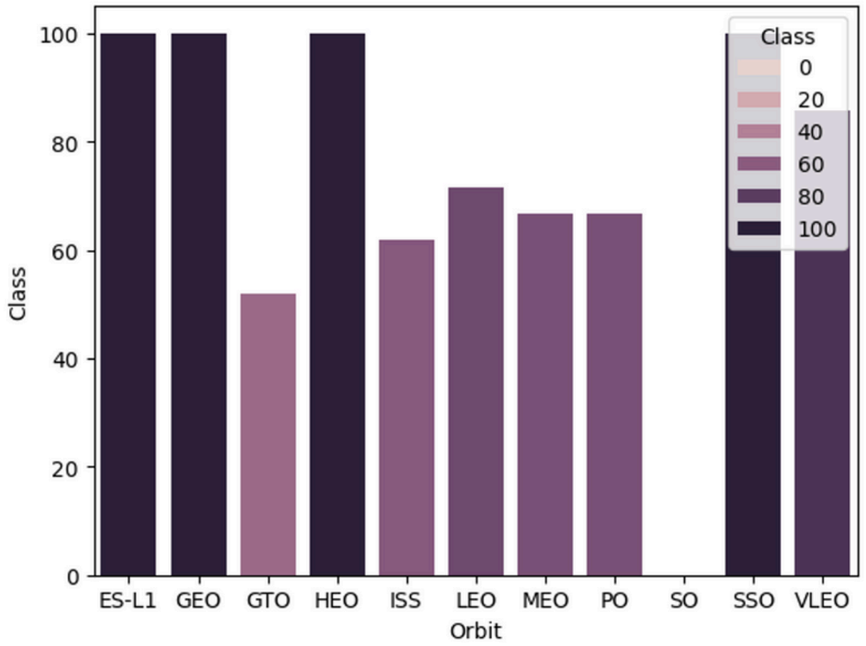
Payload vs Launch Site



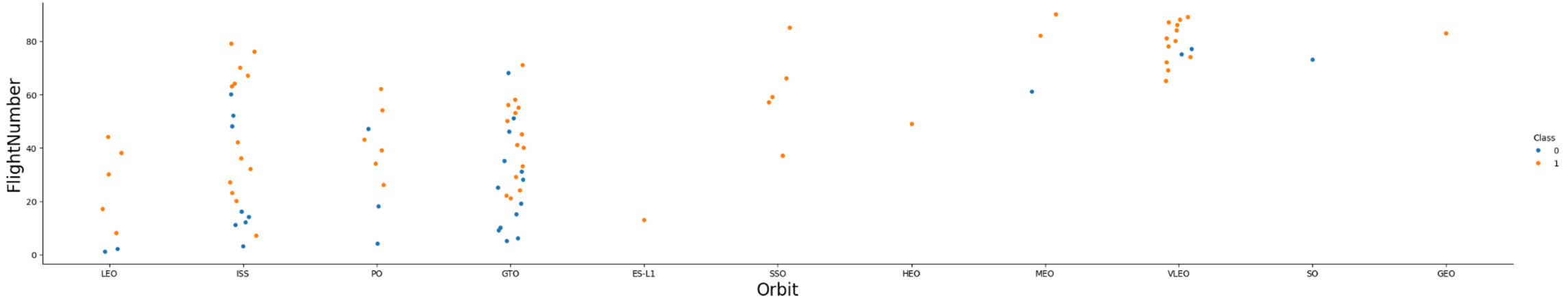
- Scatter Plot Insights on Payload Mass and Mission Success:
  - Payload Mass Threshold: Missions with payloads over 7000 kg show a higher success probability, indicating that advancements in technology and operations enable successful launches of heavier payloads.
  - Launch Site-Specific Payload Preferences: Launch sites cater to specific payload mass ranges, likely influenced by vehicle capabilities, infrastructure, and mission objectives.
  - No Clear Site-Payload-Success Correlation: While payload mass is critical for success, no clear relationship between launch site and payload mass in terms of success rate is observed. Other factors, such as vehicle performance, weather, and mission complexity, likely affect the outcome.
- Further analysis, like statistical tests or machine learning models, is needed to uncover deeper relationships between these variables and mission success.

Success Rate vs Orbit Type

- The figure suggests a potential correlation between orbit type and landing outcome:
  - ·High Success Rate Orbits: Orbits such as SSO, HEO, GEO, and ES-L1 appear to have a 100% success rate in landing.
  - ·Low Success Rate Orbit: The SO orbit, on the other hand, seems to have a 0% success rate.
  - ·Caveats and Further Analysis:
- ·It's important to note that the limited sample size for certain orbits, particularly GEO, SO, HEO, and ES-L1, may skew the results. With only one occurrence for each of these orbits, it's difficult to draw definitive conclusions about their impact on landing outcomes.
- To gain a more accurate understanding of the relationship between orbit type and landing success, a larger dataset with more diverse and representative samples is necessary. This will enable us to conduct more robust statistical analysis and identify any underlying patterns or trends.

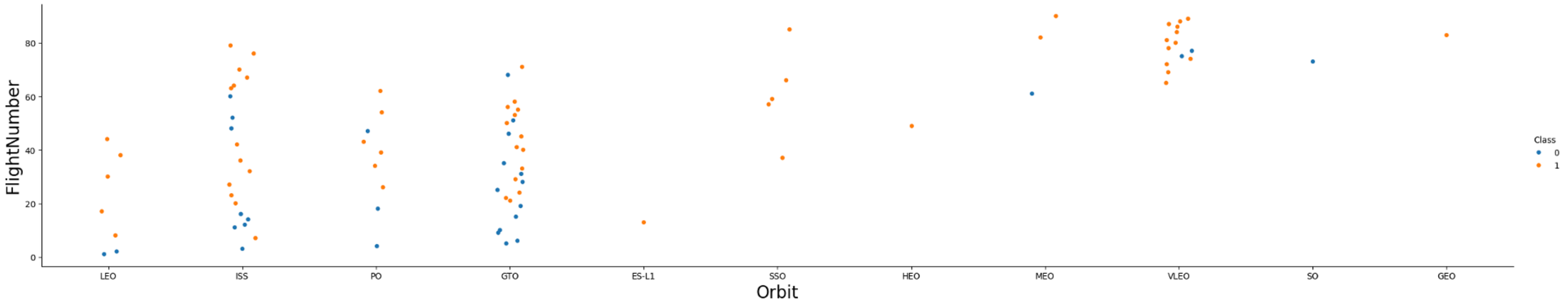


Flight Number vs Orbit Type



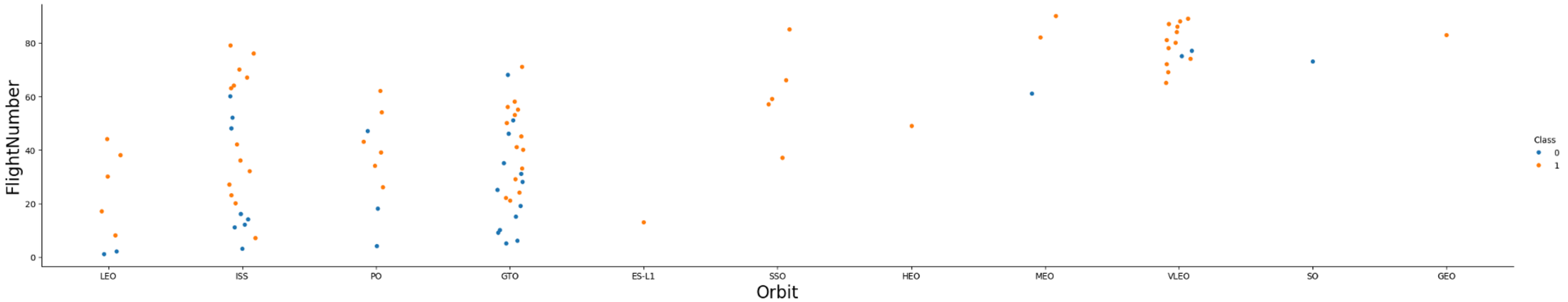
- Scatter Plot Insights on Flight Number and Success Rate for Orbits:
  - Positive Correlation for LEO Orbits: For LEO orbits, as the number of launches increases, the success rate improves, indicating that SpaceX has gained significant experience and enhanced capabilities for LEO missions.
  - GTO Orbit Variability: GTO orbits show more variability with no clear correlation between flight number and success rate, likely due to the complexity of GTO missions, such as higher energy requirements and intricate trajectory maneuvers.
  - SpaceX's Focus on Lower Orbits: SpaceX's focus on LEO and recent shift towards VLEO align with success trends, as these orbits appear to be their areas of expertise with consistent success.
  - Potential for Sun-Synchronous Orbits: The scatter plot suggests the potential for high success rates in Sun-Synchronous orbits, though further analysis with a larger dataset is needed to confirm this trend.

Flight Number vs Orbit Type



- Key Insights from Scatter Plot: Flight Number vs. Success Rate by Orbit:
  - LEO Orbits: Success rate improves as launches increase, indicating SpaceX's growing expertise in LEO missions.
  - GTO Orbits: No clear correlation, likely due to the complexities of GTO missions.
  - Focus on Lower Orbits: SpaceX’s consistent success in LEO and VLEO missions aligns with their strategic focus on these orbits.
  - Sun-Synchronous Orbits: Potential for high success, but more data is needed to confirm.

Payload vs Orbit Type



- Key Findings:
  - Payload Mass Impact:
  - Positive: Heavier payloads correlate with higher success in LEO, ISS, and P0 orbits.
  - Negative: Heavier payloads reduce success in MEO and VLEO orbits.
  - No Correlation: GTO orbits show no clear trend.
- Orbit-Specific Trends:
  - LEO/SSO: Lighter payloads dominate.
  - VLEO: Heavier payloads are successful.
- Data Limitations:
  - Limited data for SO, GEO, and HEO orbits.
  - Further analysis with more data is needed.
  -