Winning Space Race with Data Science

Nhi Nguyen
17-Feb-2025

Outline

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

Executive Summary

- Summary of methodologies:

-Data Collection

-Data Wraggling

-Perform exploratory data analysis (EDA) using visualization and SQL

-Perform interactive visual analytics using Folium and Plotly Dash

-Perform predictive analysis using classification models

- Summary of all results

Introduction

- Project background and context:

Take a role of Data Scientist at a startup aiming to compete with SpaceX in the commercial space industry. The primary goal is to develop predictive models to determine the likelihood of the SpaceX Falcon 9 rocket's first stage successfully landing. This capability is crucial because SpaceX's cost advantage—offering launches at 62 million dollars compared to competitors' 165 million dollars—is largely due to their ability to reuse the first stage of their rockets. Accurately predicting landing success can enable your company to make informed bids and offer competitive pricing for satellite launch services

- Problems you want to find answers:

-Predict the First Stage Landing Success

-Identify key factors such as payload, launch site, weather conditions

-Cost-Benefit Analysis

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Use SpaceX Rest API
- Perform data wrangling
  - Use pd.json_normalize function to normalize a structured JSON data into a flat table
  - Deal with nulls
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building and evaluating classification models to ensure the best results

Data Collection

- Dataset: SpaceX REST API

-Use get request to get the launch data. Use .json() method to review the result

-Use json_normalize to convert structured json data into a flat table

-Use Python BeautifulSoup package to web scrape HTML tables -> parse data and convert into Pandas dataframe

-Transform into a clean dataset which provides meaningful data: Wrangling Data using an API, Sampling Data, and Dealing with Nulls

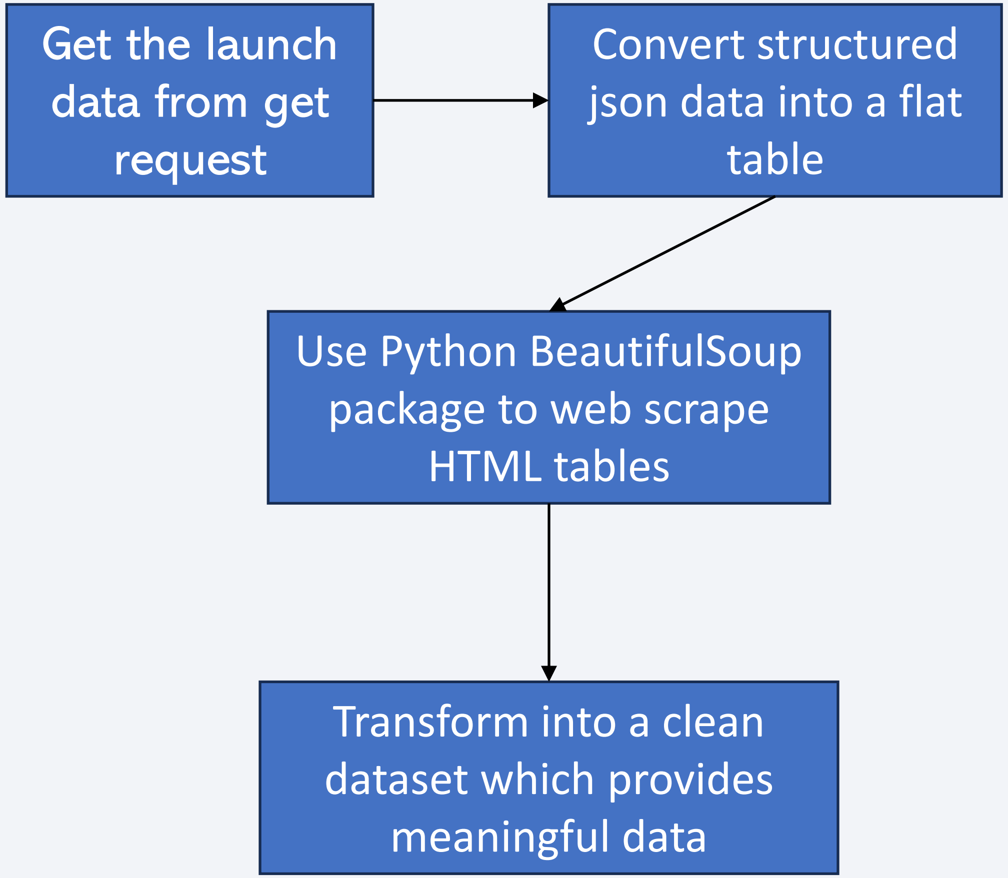
Data Collection – SpaceX API

-Use get request to get the launch data. Use .json() method to review the result

-Use json_normalize to convert structured json data into a flat table

-Use Python BeautifulSoup package to web scrape HTML tables -> parse data and convert into Pandas dataframe

-Transform into a clean dataset which provides meaningful data: Wrangling Data using an API, Sampling Data, and Dealing with Nulls



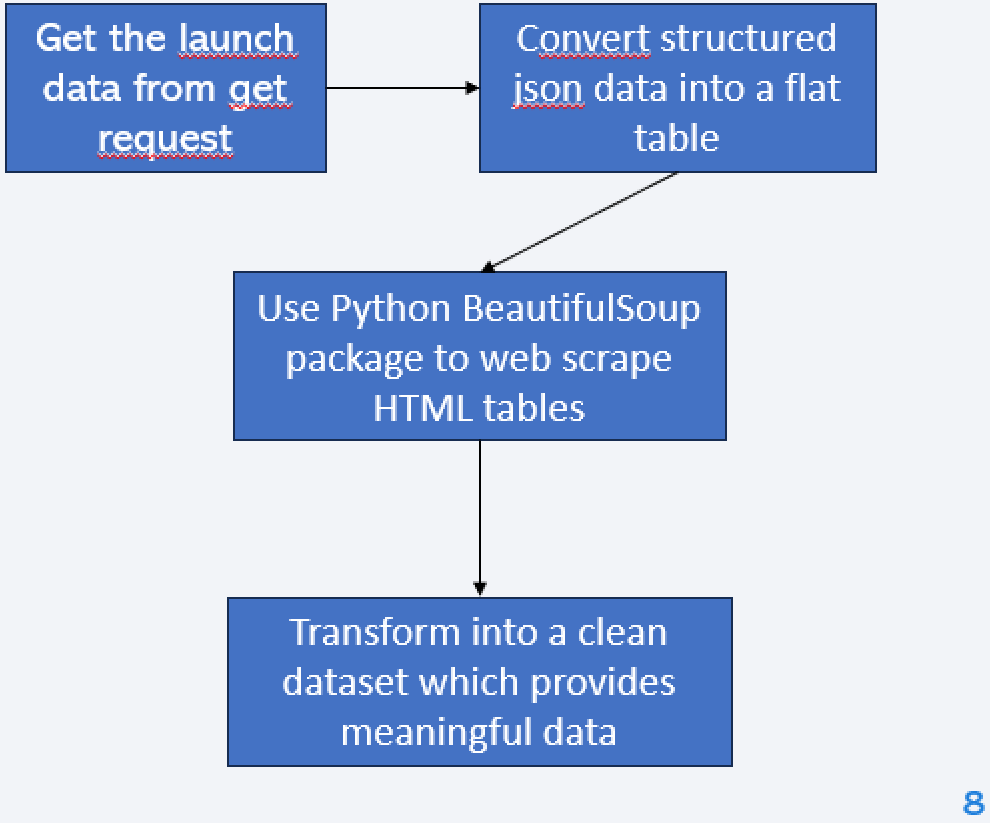
Data Collection - Scraping

-Use get request to get the launch data. Use .json() method to review the result

-Use json_normalize to convert structured json data into a flat table

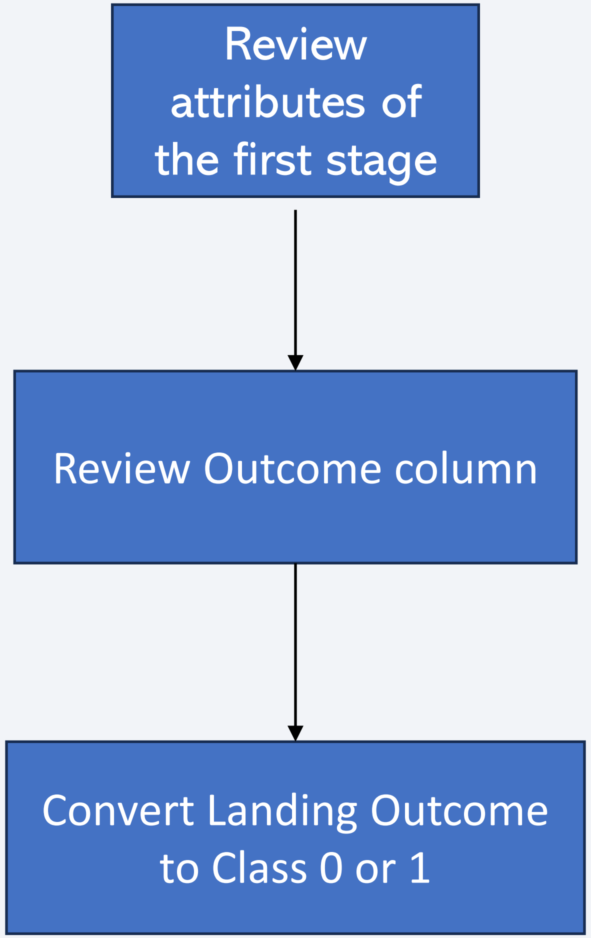
-Use Python BeautifulSoup package to web scrape HTML tables -> parse data and convert into Pandas dataframe

-Transform into a clean dataset which provides meaningful data: Wrangling Data using an API, Sampling Data, and Dealing with Nulls



Data Wrangling

- Review attributes of the first stage: Flight Number, Date, Booster version, Payload mass Orbit, Launch Site
- Review Outcome column
- Convert Landing Outcome to Classes y: 0 or 1. 0 means the booster did not land and 1 means the booster did land.



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Bar Charts – To visualize the frequency of successful and failed landings across different launch sites, helping identify which locations are more favorable for landings.
- Scatter Plots – Used to explore relationships between numerical variables, such as payload mass and landing success, identifying potential patterns.
- Line Charts – Showing trends in launches over time, helping track changes in success rates and other key metrics.
- Pie Charts – To display categorical distributions, such as the proportion of successful vs. failed landings.
- Box Plots – Used to examine the spread and outliers in numerical variables, such as payload mass, providing insights into the impact of various factors on landing outcomes.

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

-Data Retrieval: Extracted relevant data from the database for analysis.

-Data Filtering: Applied conditions to focus on specific subsets of data.

-Aggregation: Calculated summary statistics such as averages and counts. \

-Joins: Combined data from multiple tables to enrich the analysis.

-Sorting: Ordered data based on specific columns to identify trends.

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Markers: Placed at SpaceX launch sites to indicate their exact locations. These help visualize where the launches take place.
- Circle Markers: Added around launch sites to represent areas of interest and highlight surrounding regions. The radius of the circles provides spatial context.
- Lines (Polylines): Used to connect launch sites to other points (e.g., nearest city, ocean landing zones). These lines illustrate distances and trajectories relevant to rocket launches.
- These objects were added to enhance geographical understanding of launch site locations, analyze proximity to key areas, and visualize potential landing zones for Falcon 9 rockets.

Build a Dashboard with Plotly Dash

- Success Pie Chart: Displays the proportion of successful versus failed launches for selected launch sites -> Provides a quick visual summary of success rates for each launch site, aiding in identifying sites with higher performance.
- Success-Payload Scatter Plot: Illustrates the correlation between payload mass and launch success, highlighting trends across different payload ranges -> Helps in understanding how payload mass influences launch outcomes, which is crucial for mission planning and risk assessment

Predictive Analysis (Classification)

- Data Preprocessing: Preparing the dataset by handling missing values, encoding categorical variables, and scaling numerical features to ensure compatibility with machine learning algorithms.
- Feature Selection: Identifying and selecting the most relevant features that influence the landing outcome, such as payload mass, launch site, orbit type, and booster version.
- Model Selection: Choosing appropriate classification algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), to evaluate their performance on the dataset.
- Model Training and Evaluation: Splitting the data into training and testing sets, training each model, and assessing their performance using metrics like accuracy, precision, recall, and F1-score.
- Hyperparameter Tuning: Optimizing model parameters through techniques such as Grid Search or Random Search to enhance predictive accuracy. Model Comparison and Selection: Comparing the performance of all trained models and selecting the one with the highest evaluation metrics for deployment.

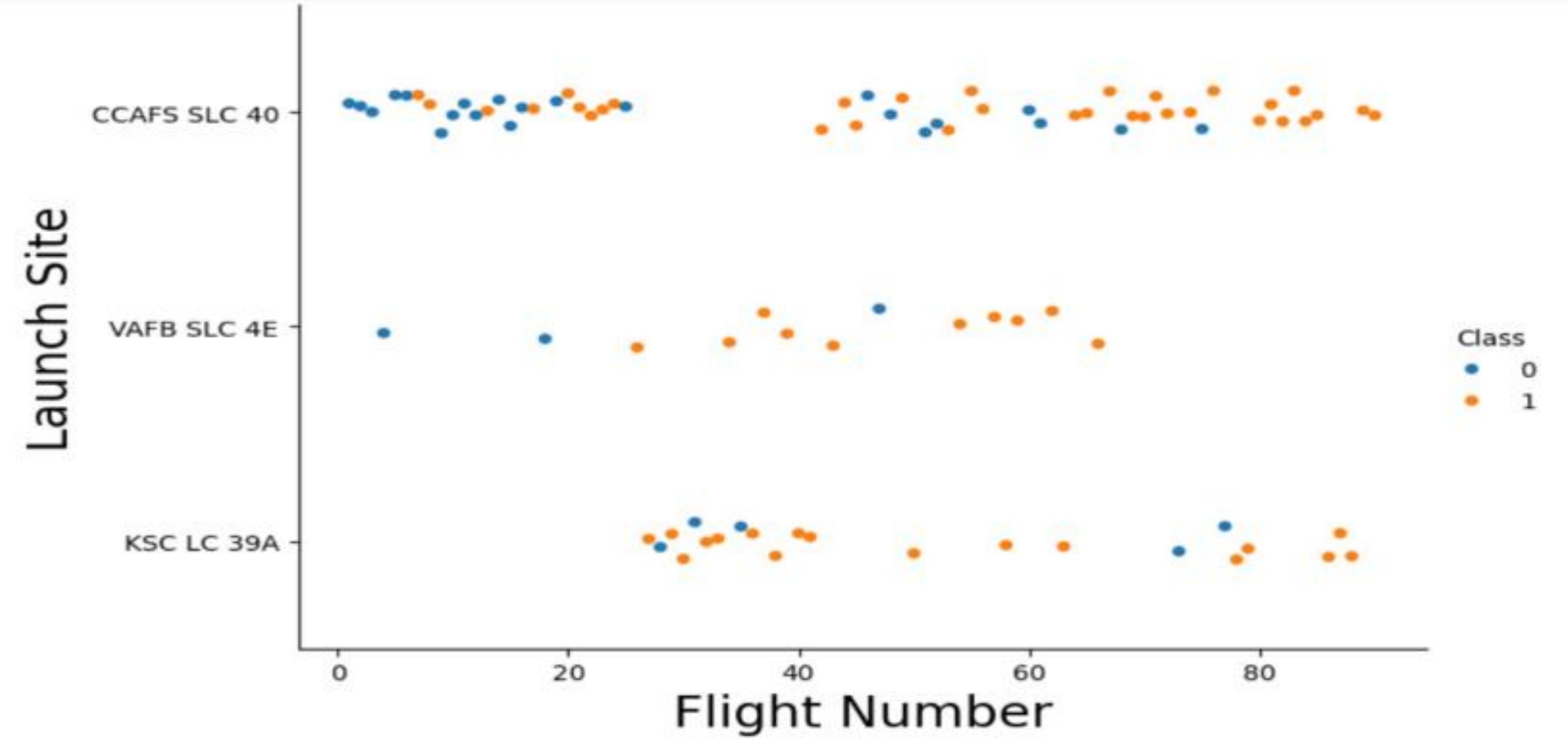
Results

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

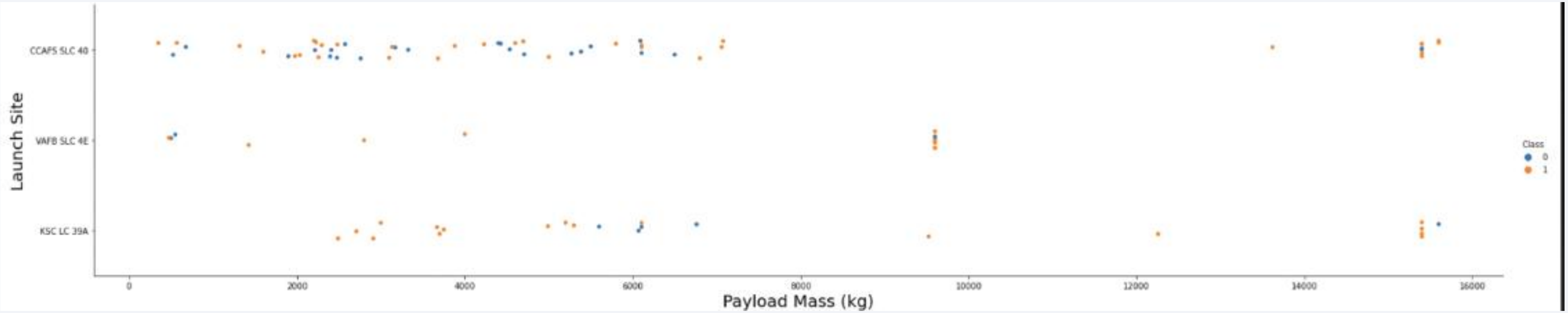
Section 2

Insights drawn from EDA

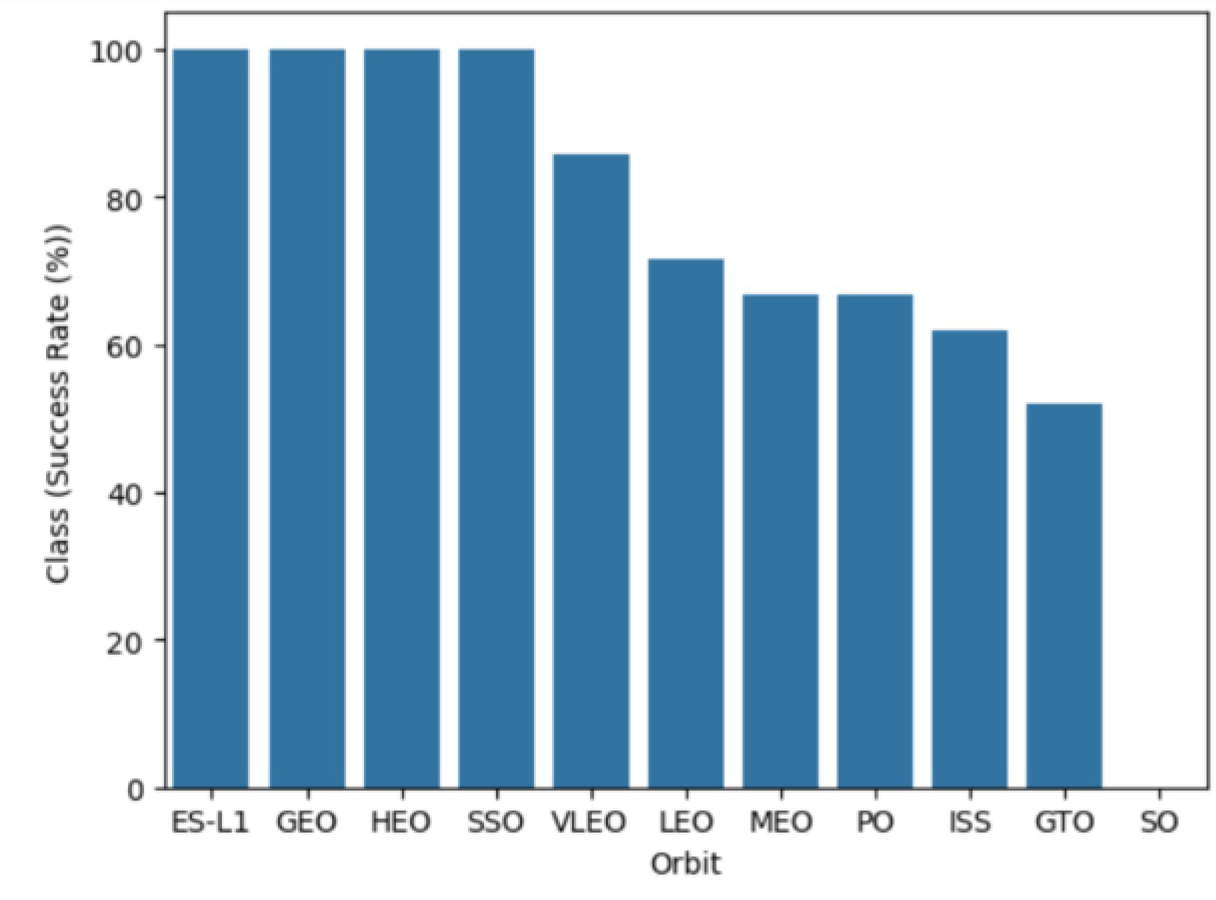
Flight Number vs. Launch Site



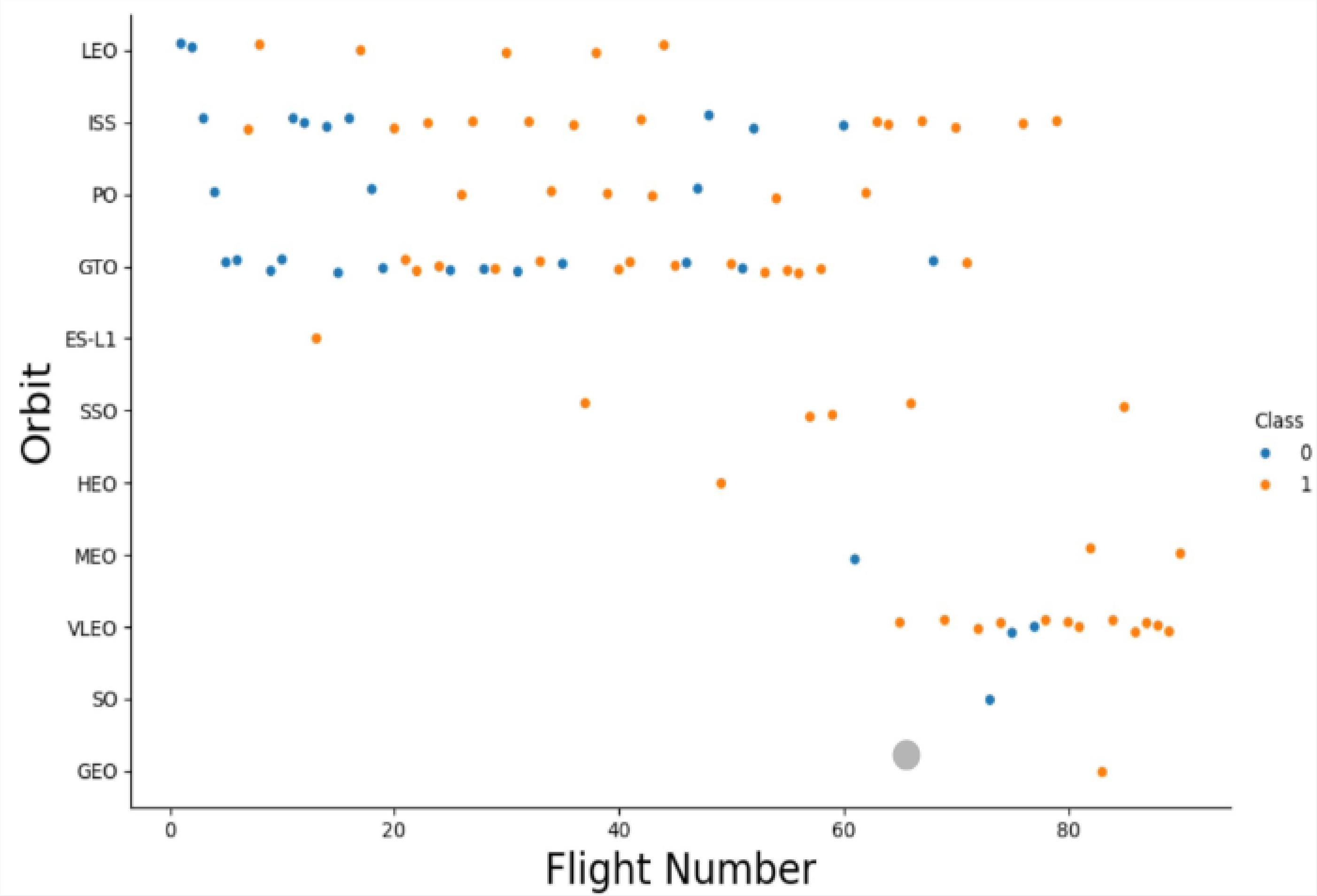
Payload vs. Launch Site



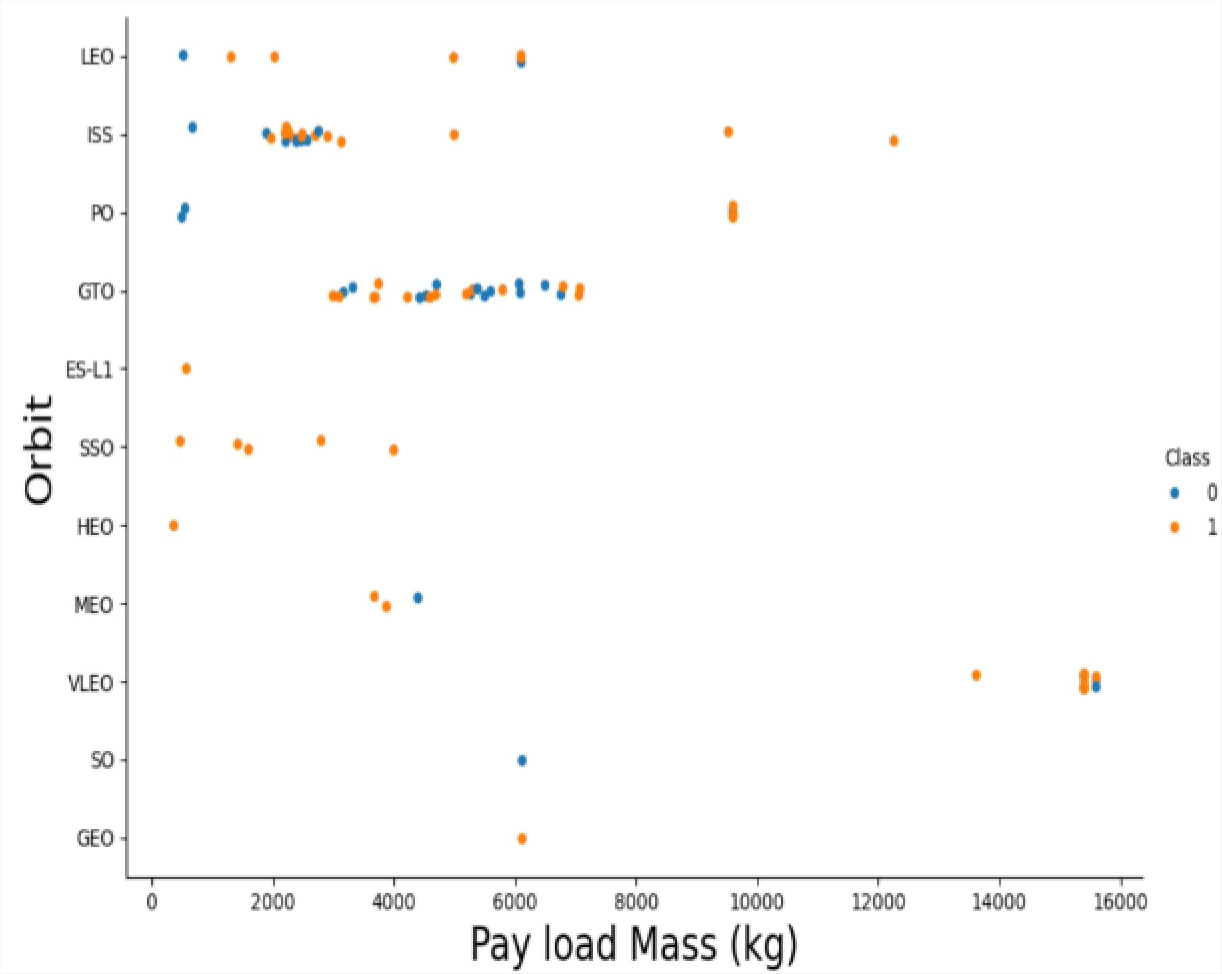
Success Rate vs. Orbit Type



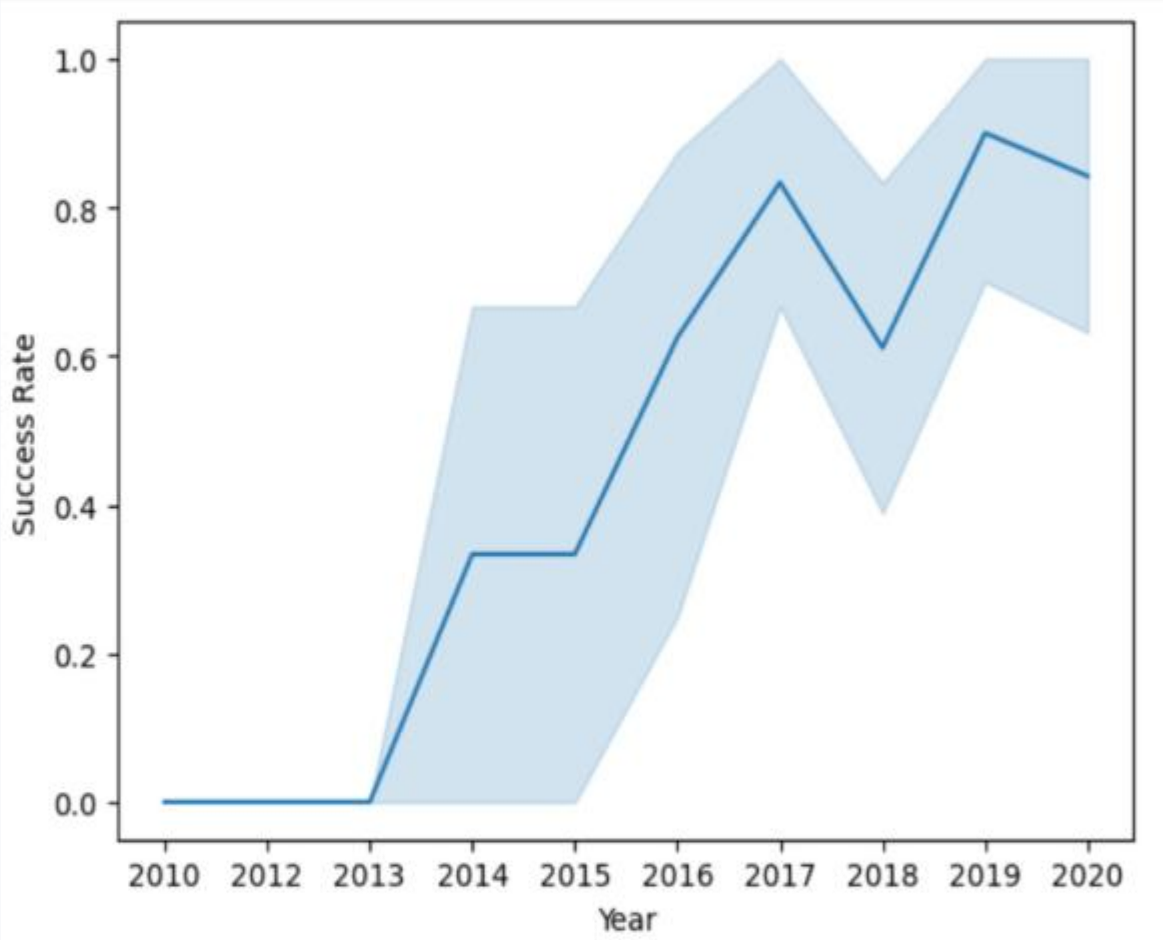
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

Display the names of the unique launch sites in the space mission

In [31]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[31]:

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM 'SPACEXTBL' WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landin _Outcom
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failu (parachut
08-12-2010	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	Failu (parachut
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attem
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attem
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attem

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

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

- Present your query result with a short explanation here

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

- Replace <Folium map screenshot 1> title with an appropriate title
- Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title

- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map

- Explain the important elements and findings on the screenshot

<Folium Map Screenshot 3>

- Replace <Folium map screenshot 3> title with an appropriate title

- Explore the generated folium map and show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed

- Explain the important elements and findings on the screenshot

Section 4

Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>

- Replace <Dashboard screenshot 1> title with an appropriate title

- Show the screenshot of launch success count for all sites, in a piechart

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title

- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

- Show the confusion matrix of the best performing model with an explanation

Conclusions

- Point 1
- Point 2
- Point 3
- Point 4
- …

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

Thank you!