


# DATA SCIENCE CAPSTONE PROJECT

**Owner : Nguyễn Đăng Huỳnh Châu**

**04/06/2024**



# OVERVIEW

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

# EXECUTIVE SUMMARY

## Summary of methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

## Summary of all results

- Exploratory Data Analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

# INTRODUCTION

- **The objective:**

- Evaluating the viability of the new company Space Y to compete with Space X.

- **Questions are:**

- How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
- Does the rate of successful landings increase over the years?
- What is the best algorithm that can be used for binary classification in this case?



# METHODOLOGY

## 1. Data collection methodology:

- Using SpaceX Rest API
- Using Web Scrapping from Wikipedia

## 2. Performed data wrangling

- Filtering the data
- Dealing with missing values
- Using One Hot Encoding to prepare the data to a binary classification

## 3. Performed exploratory data analysis (EDA) using visualization and SQL

## 4. Performed interactive visual analytics using Folium and Plotly Dash

## 5. Performed predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best results

# DATA COLLECTION

Data collection process involved a combination of API requests from SpaceX REST

API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

## SpaceX REST API

Data Columns are obtained by using SpaceX REST API

## Wikipedia Web Scraping

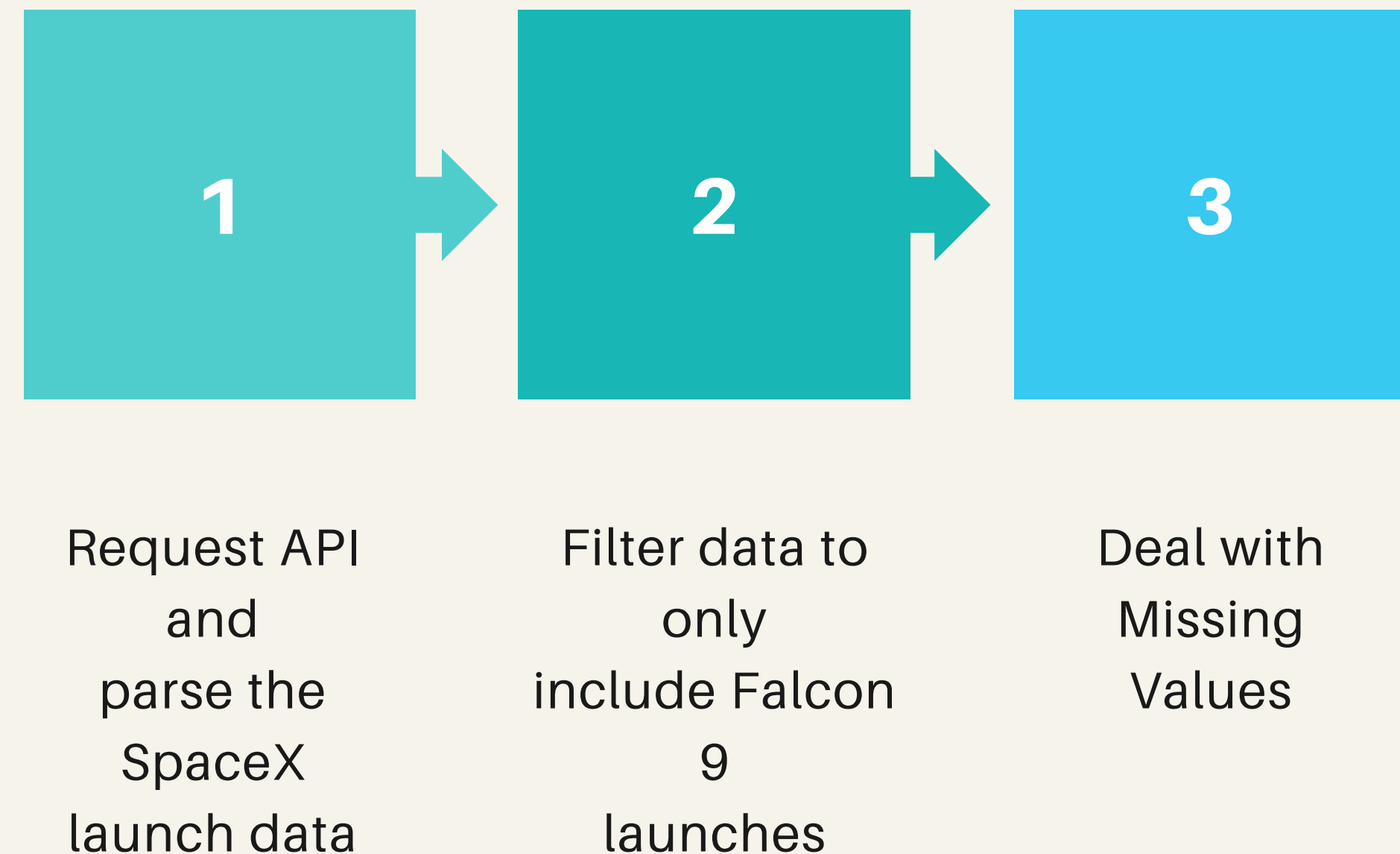
Data Columns are obtained by using Wikipedia Web Scraping

# DATA COLLECTION : SPACEX API

- SpaceX offers a public API from where data can be obtained and then used:

- This API was used according to the flowchart beside and then data is persisted.

[source code: data collection by SpaceX API](#)

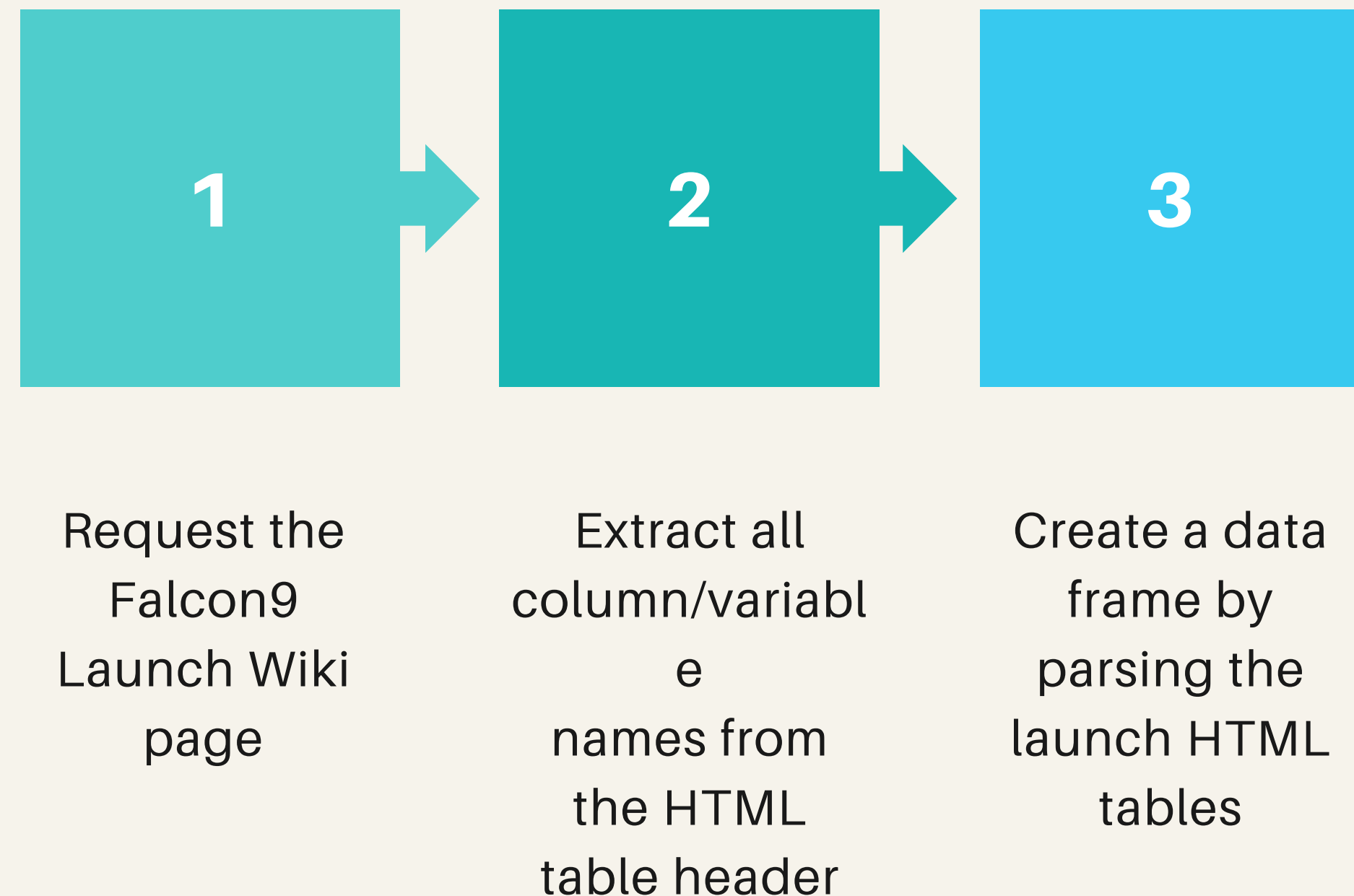




# DATA COLLECTION : SCRAPING

- Data from SpaceX launches can also be obtained from Wikipedia
- Data are downloaded from Wikipedia according to the flowchart and then persisted.

[source code: data collection web scraping](#)





# DATA WRANGLING

1

Initially some Exploratory Data Analysis (EDA) was performed on the dataset.

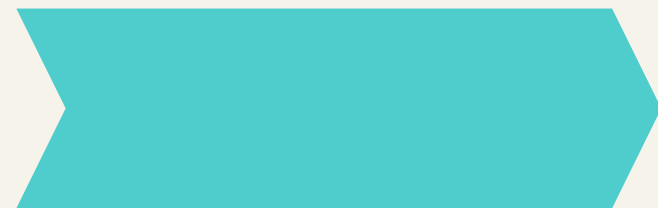
2

The summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.

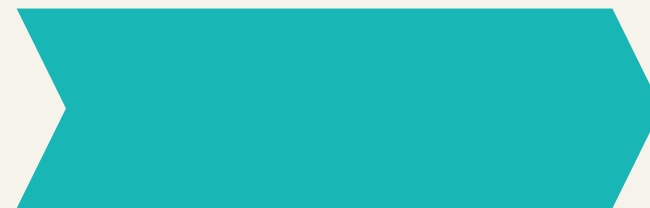
3

Finally, the landing outcome label was created from Outcome column.

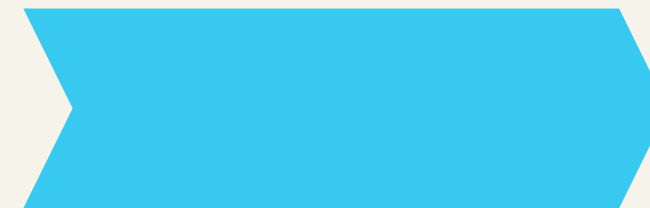
[source code: data wrangling](#)



EDA



Summarizations



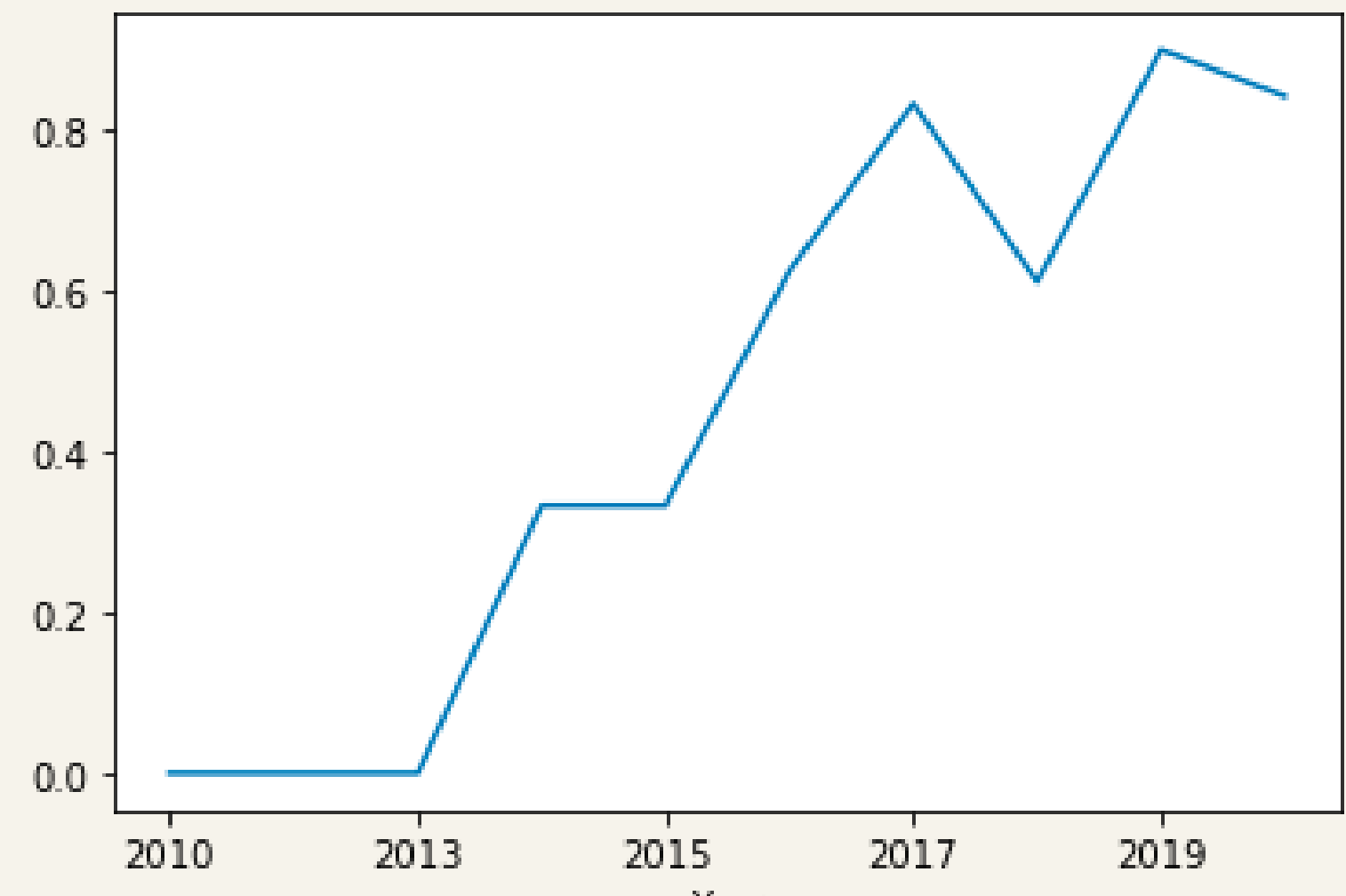
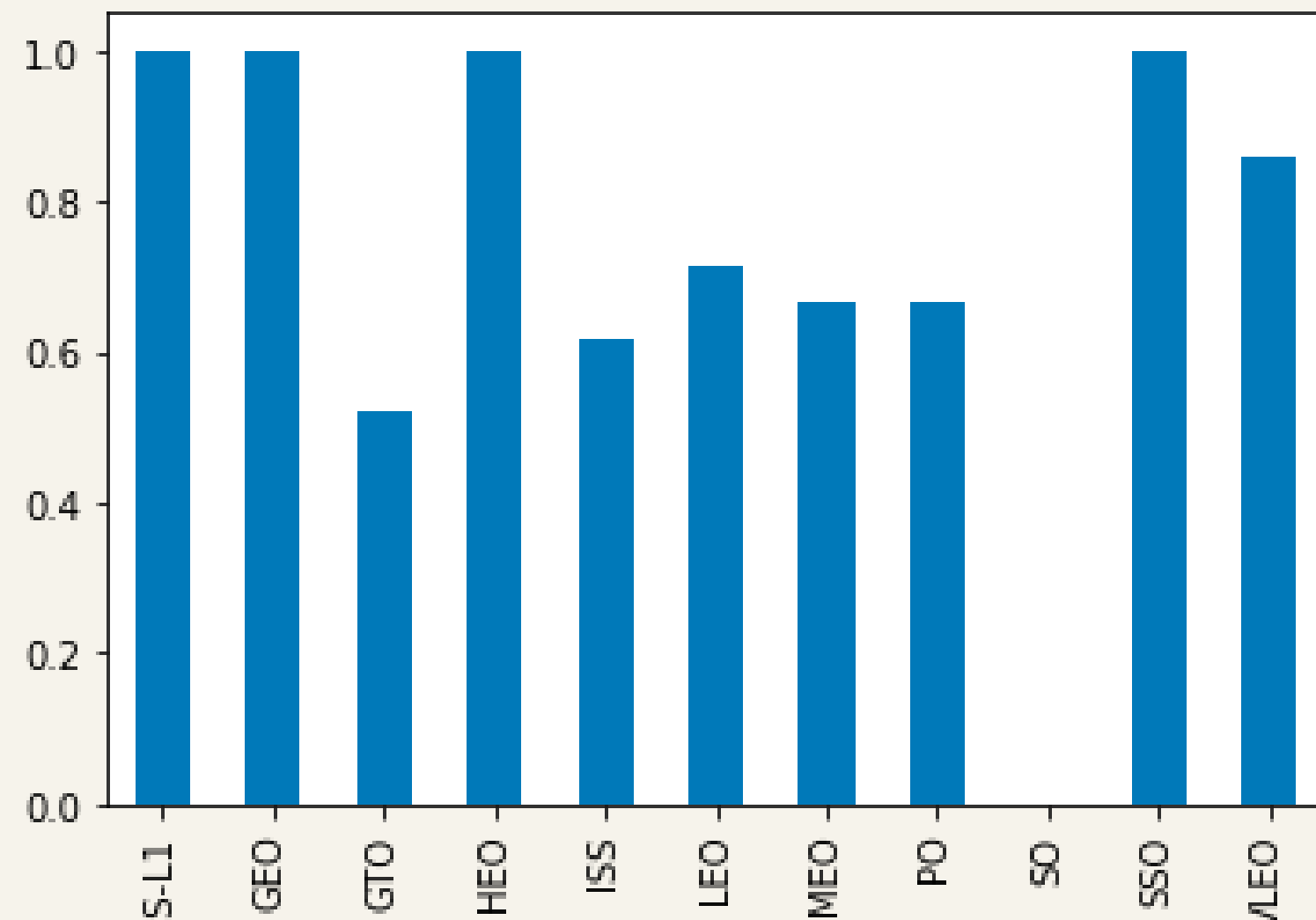
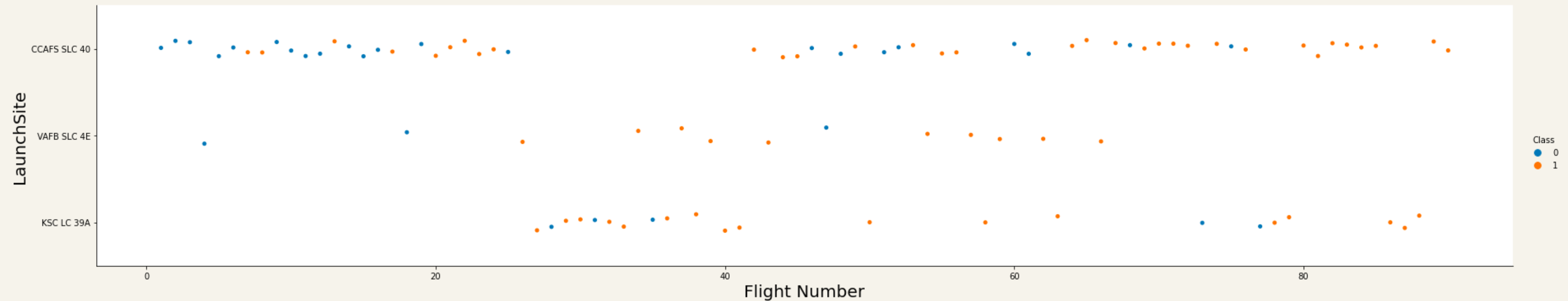
Creation of  
Landing  
Outcome Label

# DATA VISUALIZATION

- **Charts were plotted:**
  - Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- **Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.**
- **Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.**
- **Line charts show trends in data over time (time series).**

[source code: data visualisation](#)

# DATA VISUALIZATION



# EDA WITH SQL

The following SQL queries were performed:

[source code: data visualisation with SQL](#)

- Names of the unique launch sites in the space mission;
- Top 5 launch sites whose name begin with the string 'CCA';
- Total payload mass carried by boosters launched by NASA (CRS);
- Average payload mass carried by booster version F9 v1.1;
- Date when the first successful landing outcome in ground pad was achieved;
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg;
- Total number of successful and failure mission outcomes;
- Names of the booster versions which have carried the maximum payload mass;
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015;
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

# FOLIUM INTERACTIVE MAP

[source code: Folium interactive map](#)

- **Markers of all Launch Sites:**

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

- **Coloured Markers of the launch outcomes for each Launch Site:**

- Added color Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

- **Distances between a Launch Site to its proximities:**

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

# DASHBOARD WITH PLOTLY DASH

- **Launch Sites Dropdown List:**

- Added a dropdown list to enable Launch Site selection.

[source code: Dashboard Plotly Dash](#)

- **Pie Chart showing Success Launches (All Sites/Certain Site):**

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

- **Slider of Payload Mass Range:**

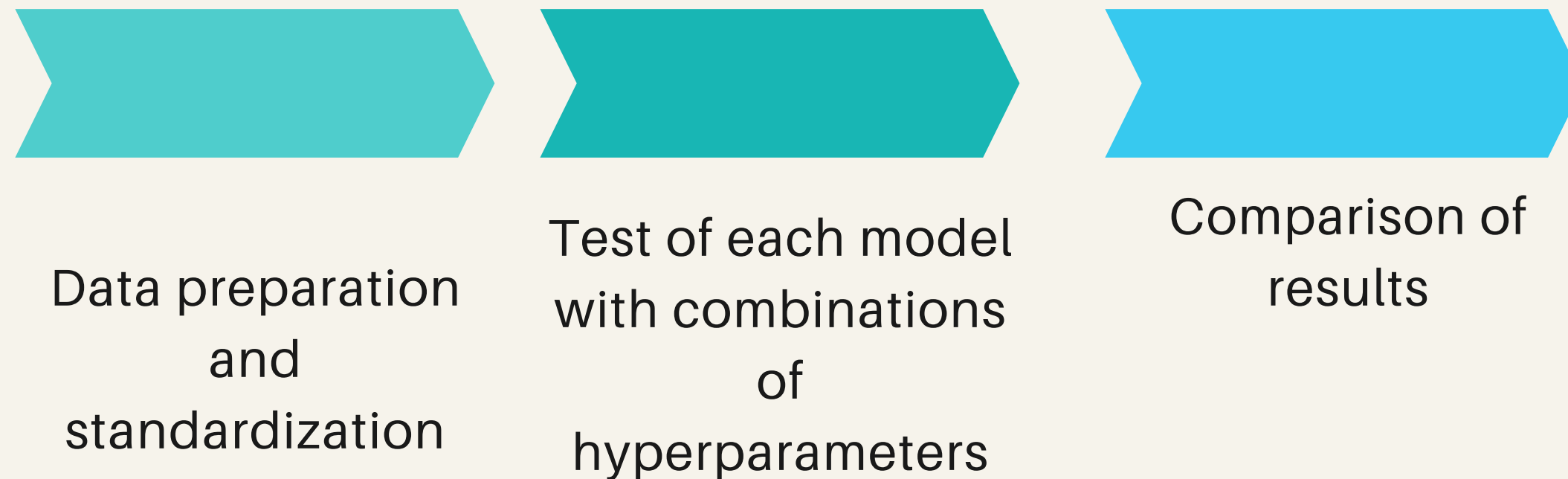
- Added a slider to select Payload range.

- **Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:**

- Added a scatter chart to show the correlation between Payload and Launch Success.

# PREDICTIVE ANALYSIS (CLASSIFICATION)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.

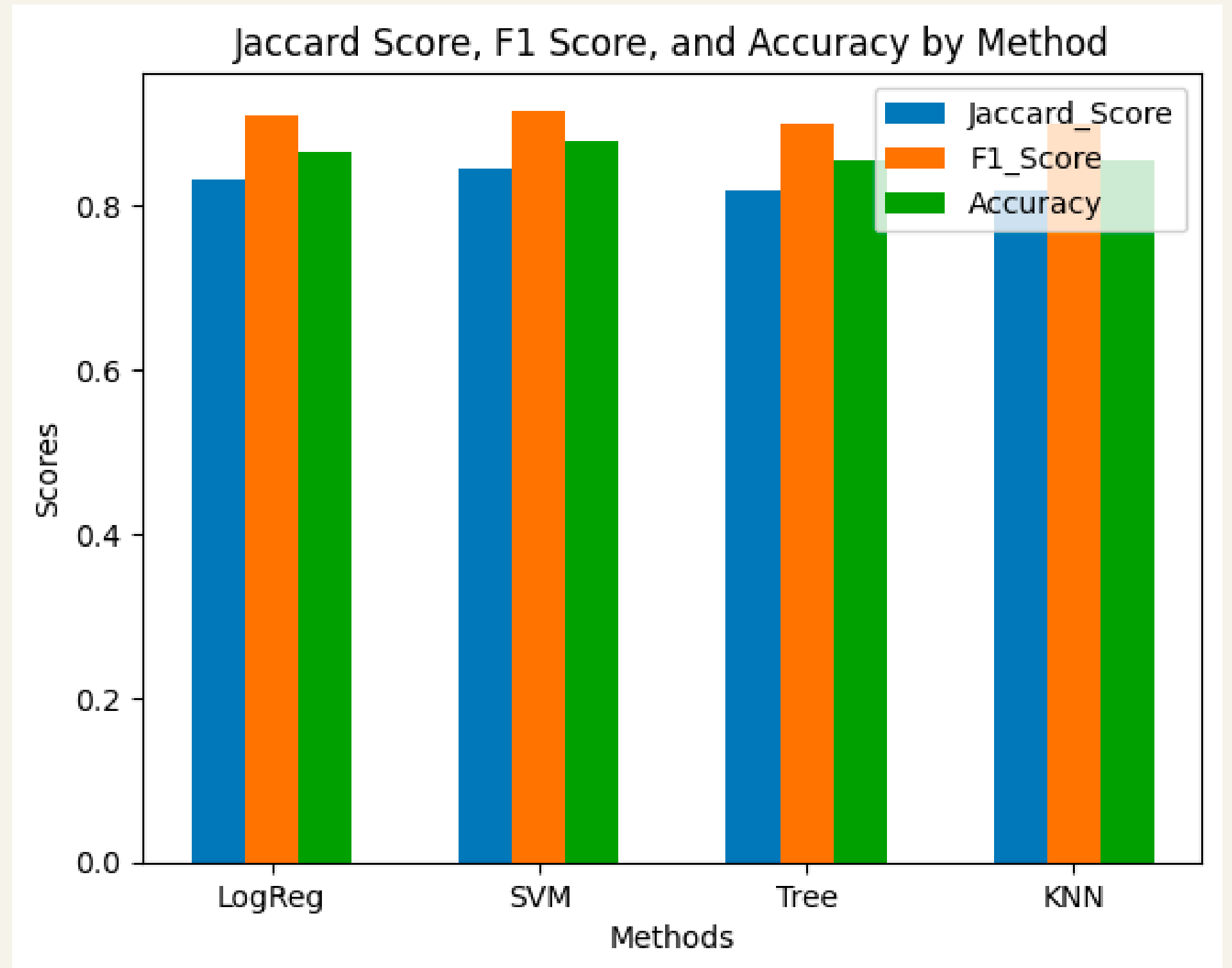


[source code: predictive analysis](#)

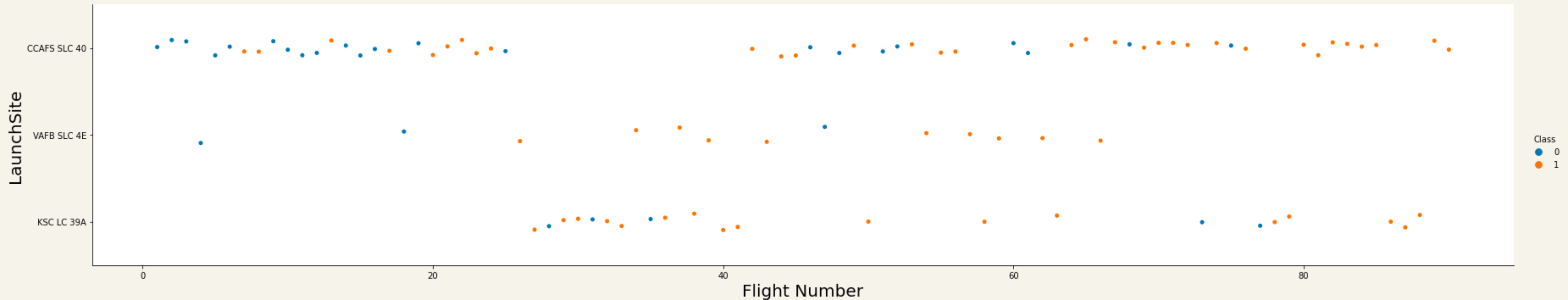


# RESULT

- **Predictive Analysis showed that SVM is the best model to predict successful landings, having Jaccard Score over 84%, F1 Score over 91% and accuracy over 87%.**

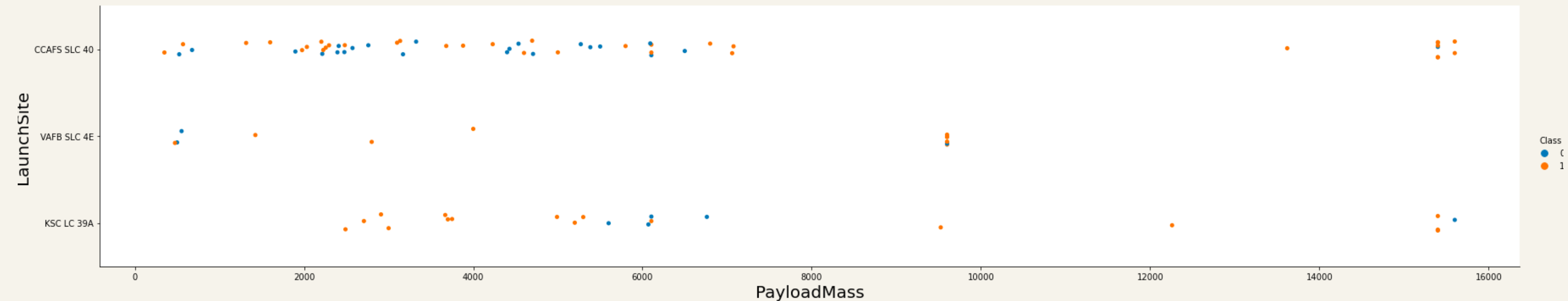


# RESULT



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAFS SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

# RESULT



- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

# CONCLUSION

- Different data sources were analyzed, refining conclusions along the process.
- The best launch site is KSC LC-39A;
- Launches above 7,000kg are less risky;
- SVM can be used to predict successful landings and increase profits.



**Larana University | 2024**

**THANK YOU**

**Presented By : Adeline Palmerston**