IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

SHEN Pingya
2024.2.22

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
- Problems you want to find answers

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - use library dataframe and js in jupyter notebook to fecth data from a url
- Perform data wrangling
  - data cleaning and data transformation
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - choose and train a Model with training set, then hyperparameter tuning use cross-validation, finally model evaluation with test set

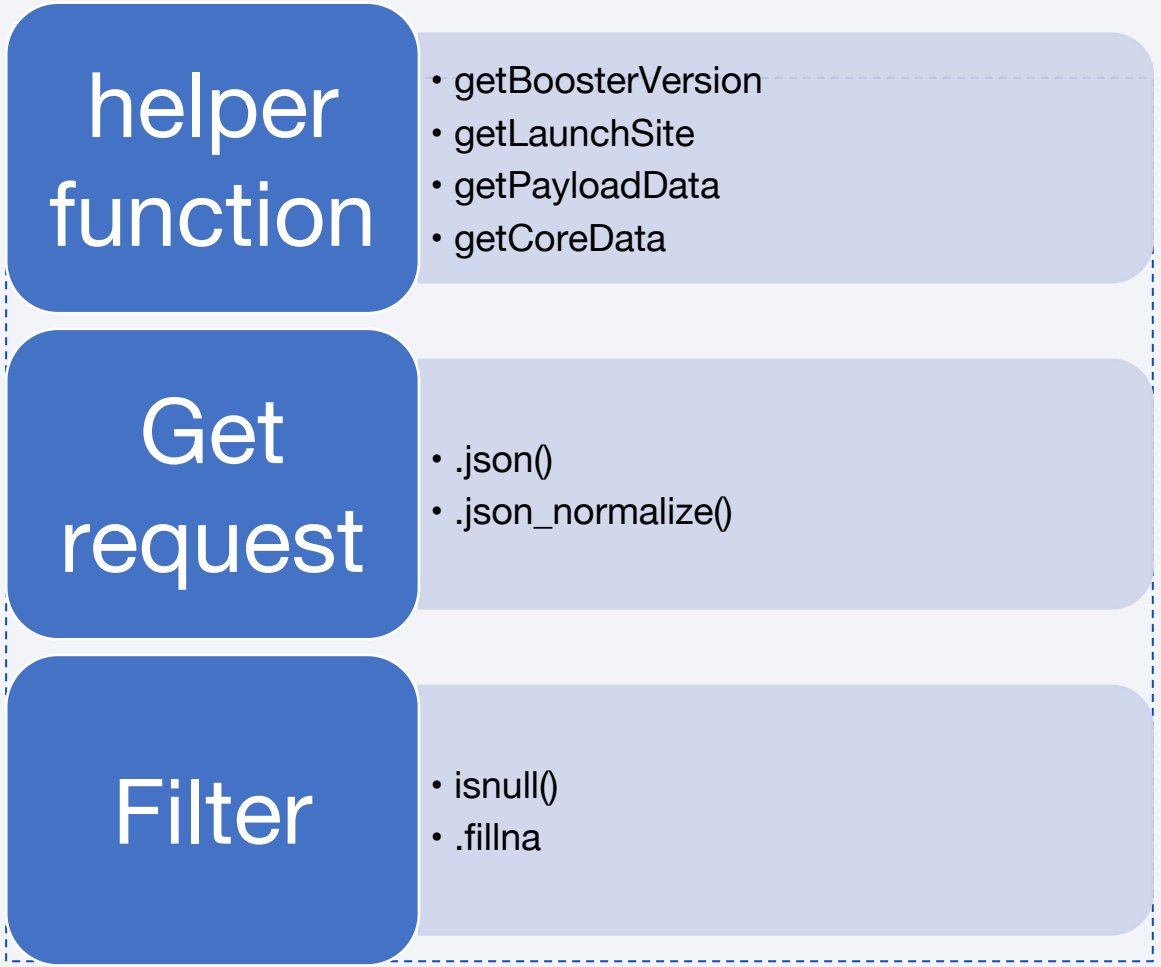
Data Collection

- I define a series of helper functions that help me use the API to extract information using identification numbers in the launch data.
- To make the requested JSON results more consistent, iuse the static response object for this project
- Filter the data dataframe and the data wrangling

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose



Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL of the completed web scraping notebook, as an external reference and peer–review purpose

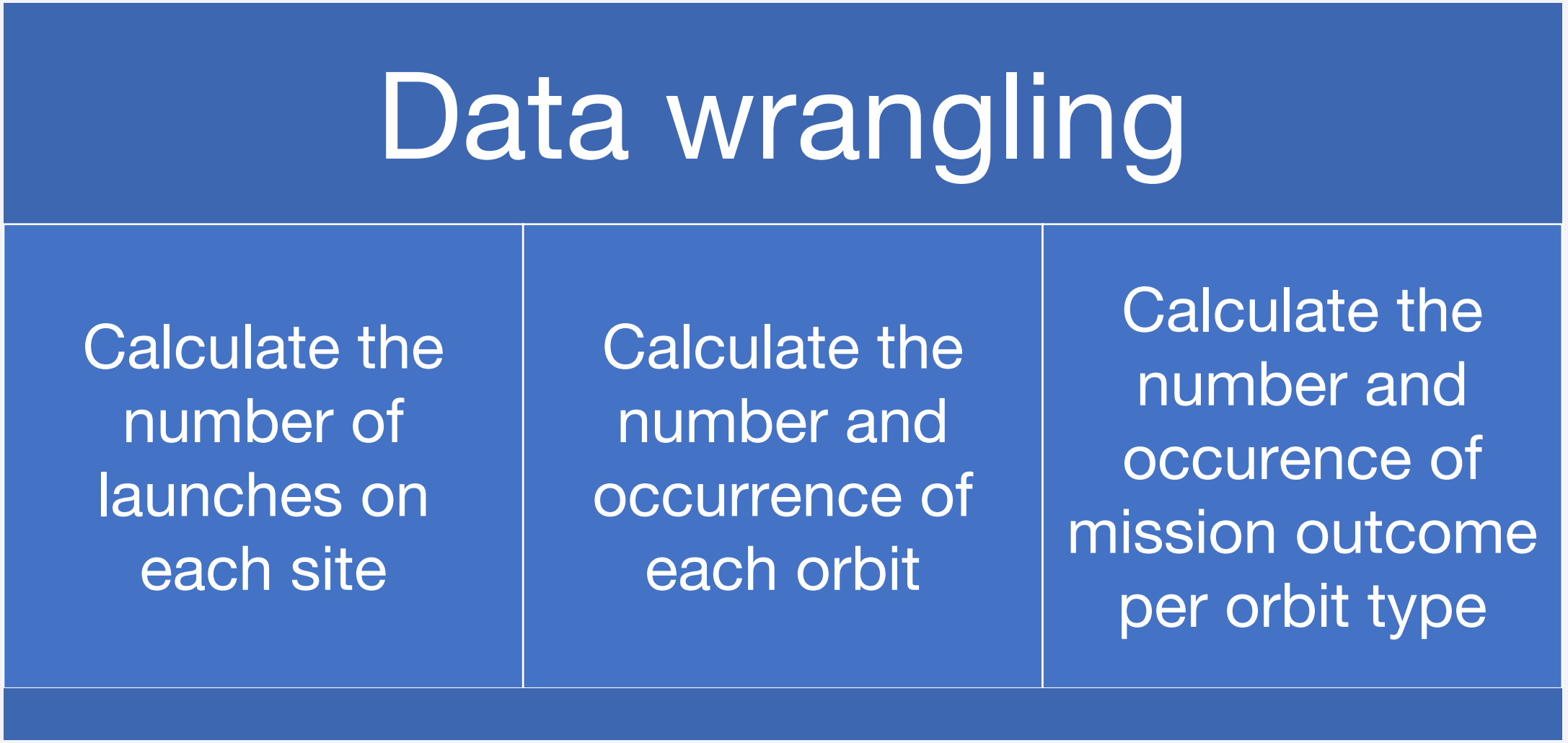
spacex_url="..."

- response = requests.get(spacex_url)

static_json_url=”...”

- response.status_code

Data Wrangling



EDA with Data Visualization

lineplot

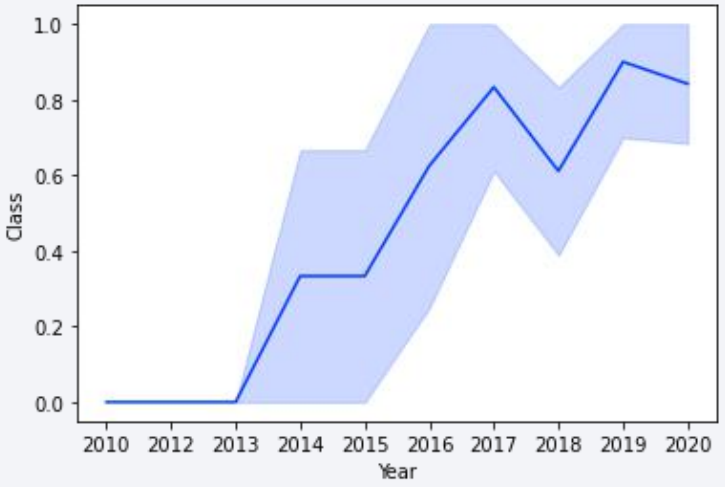
- Plot a line chart with x axis to be Year and y axis to be average success rate, to get the average launch success trend.

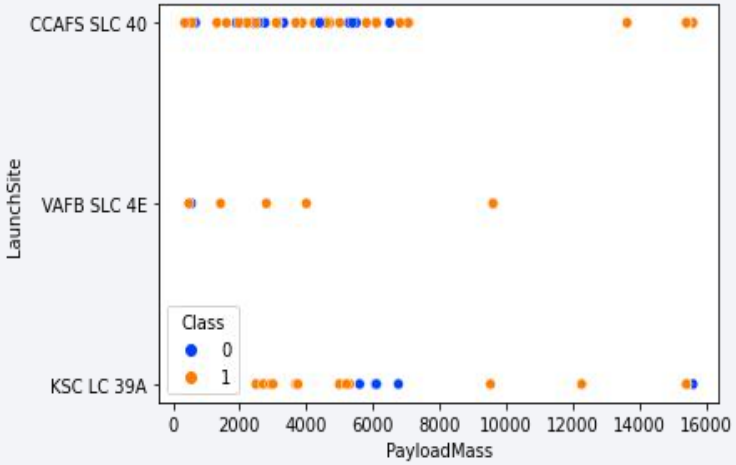
scatterplot

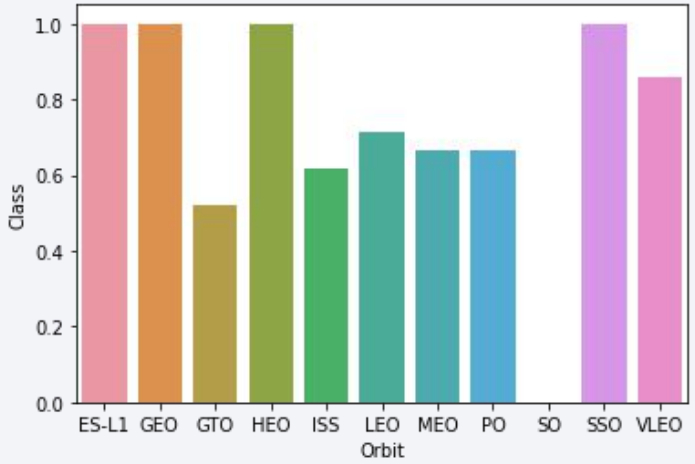
- We also want to observe if there is any relationship between launch sites and their payload mass.

barplot

- To compare different Orbit in class







EDA with SQL

- %sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
- %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5
- %sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
- %sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
- %sql select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'
- %sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
- AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')

Build an Interactive Map with Folium

- I first need to create a folium `Map` object, with an initial center location to be NASA Johnson Space Center at Houston, Texas.
- I use `folium.Circle` to add a highlighted circle area with a text label on a specific coordinate.
- I enhance the map by adding the launch outcomes for each site, and see which sites have high success rates.
- I add a `MousePosition` on the map to get coordinate for a mouse over a point on the map.

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer–review purpose

Predictive Analysis (Classification)

In this dashboard application, I added a line plot that shows the average flight delay time across different months for a selected year. Here's a summary of the plots and interactions added:

Line Plot

- The line plot displays the average flight delay time (in minutes) on the y-axis and the months of the selected year on the x-axis.
- The plot provides insights into how flight delay times vary throughout the year.

Input Year Interaction

- Users can input a specific year using a numeric input field.
- This interaction allows users to explore the flight delay patterns for different years.

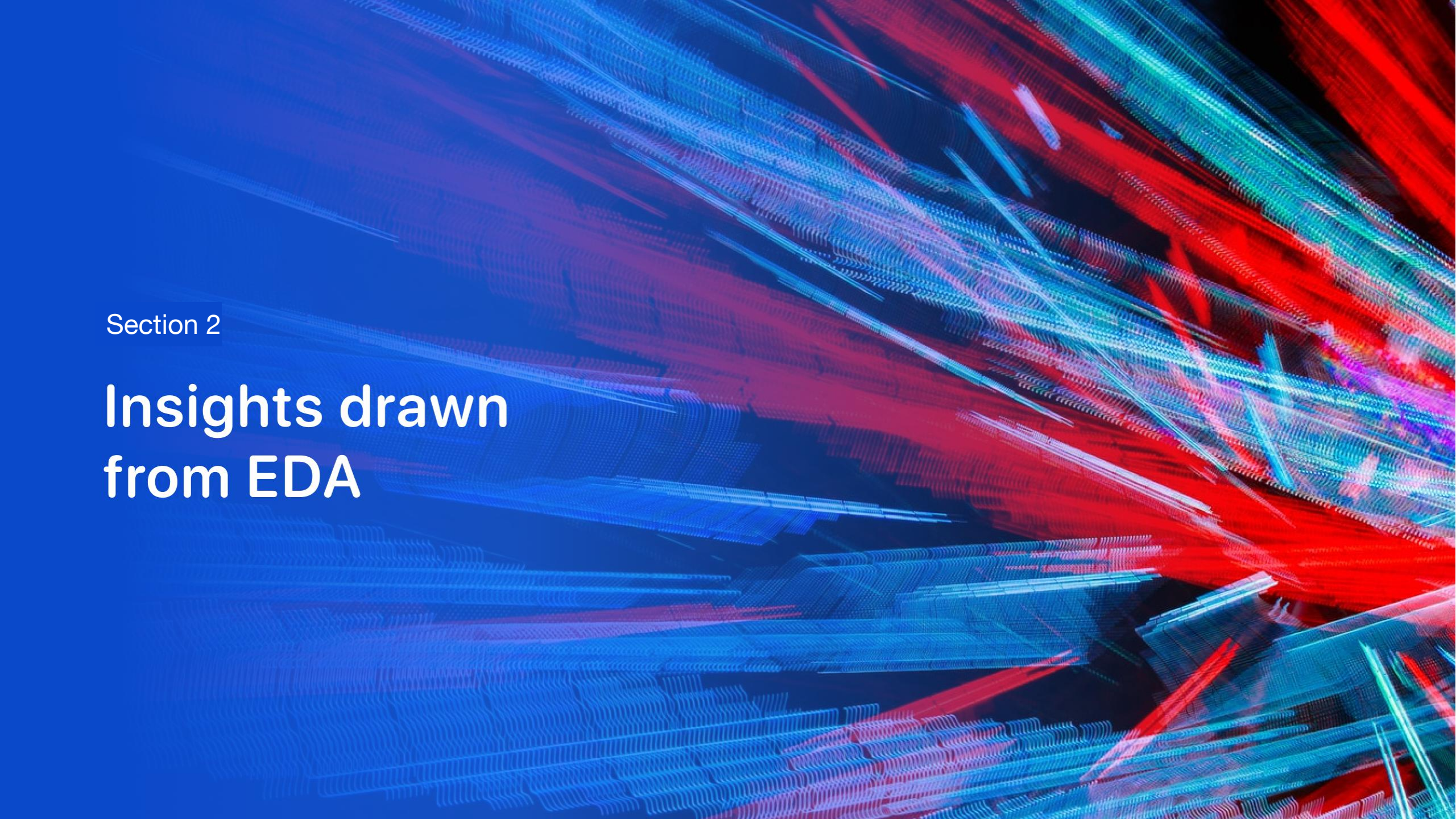
Results

Visualization of Flight Delay Trends:

- The line plot effectively visualizes the trend of average flight delay times over different months of the year. This allows users to identify patterns or seasonal variations in flight delays.

Dynamic Updating Based on User Input:

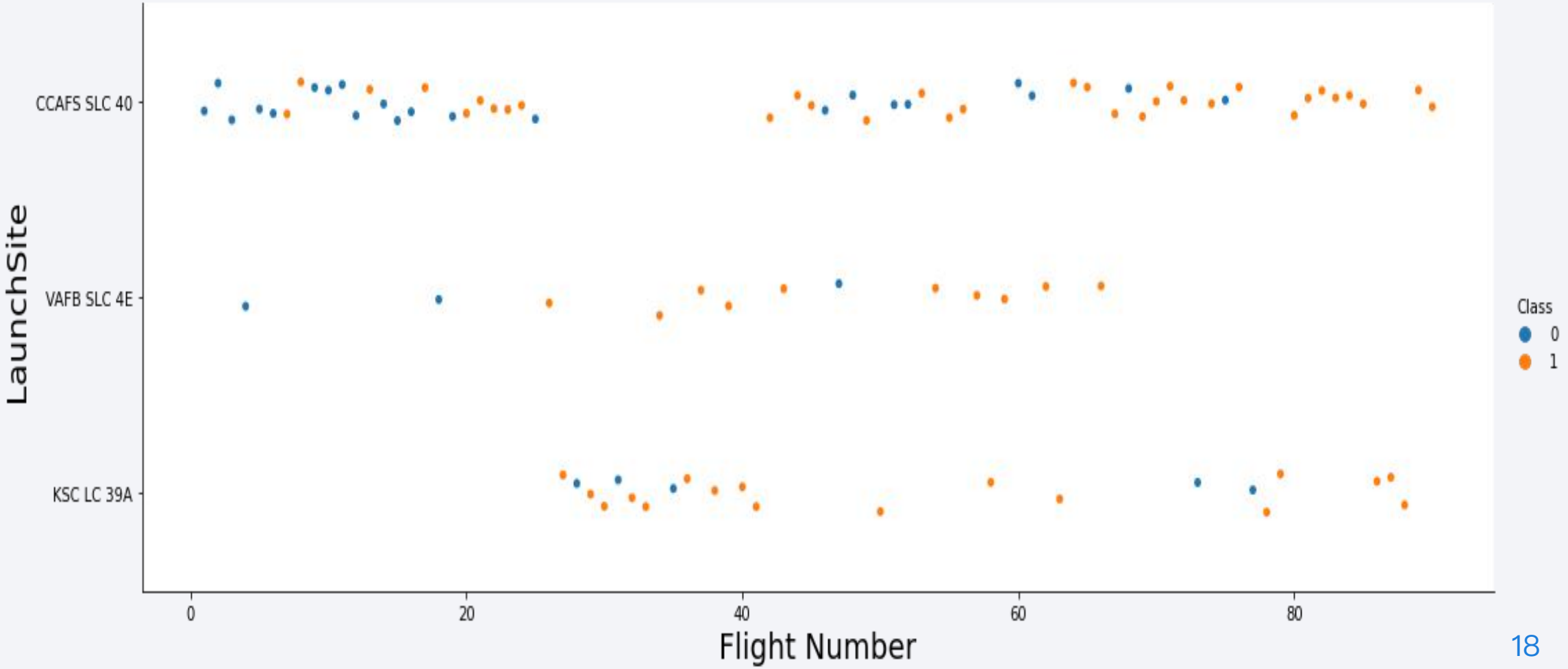
- The input field for selecting the year allows users to interactively explore flight delay data for different years.
- Dynamic updating of the line plot ensures that users can instantly see the impact of their selection on the visualization.



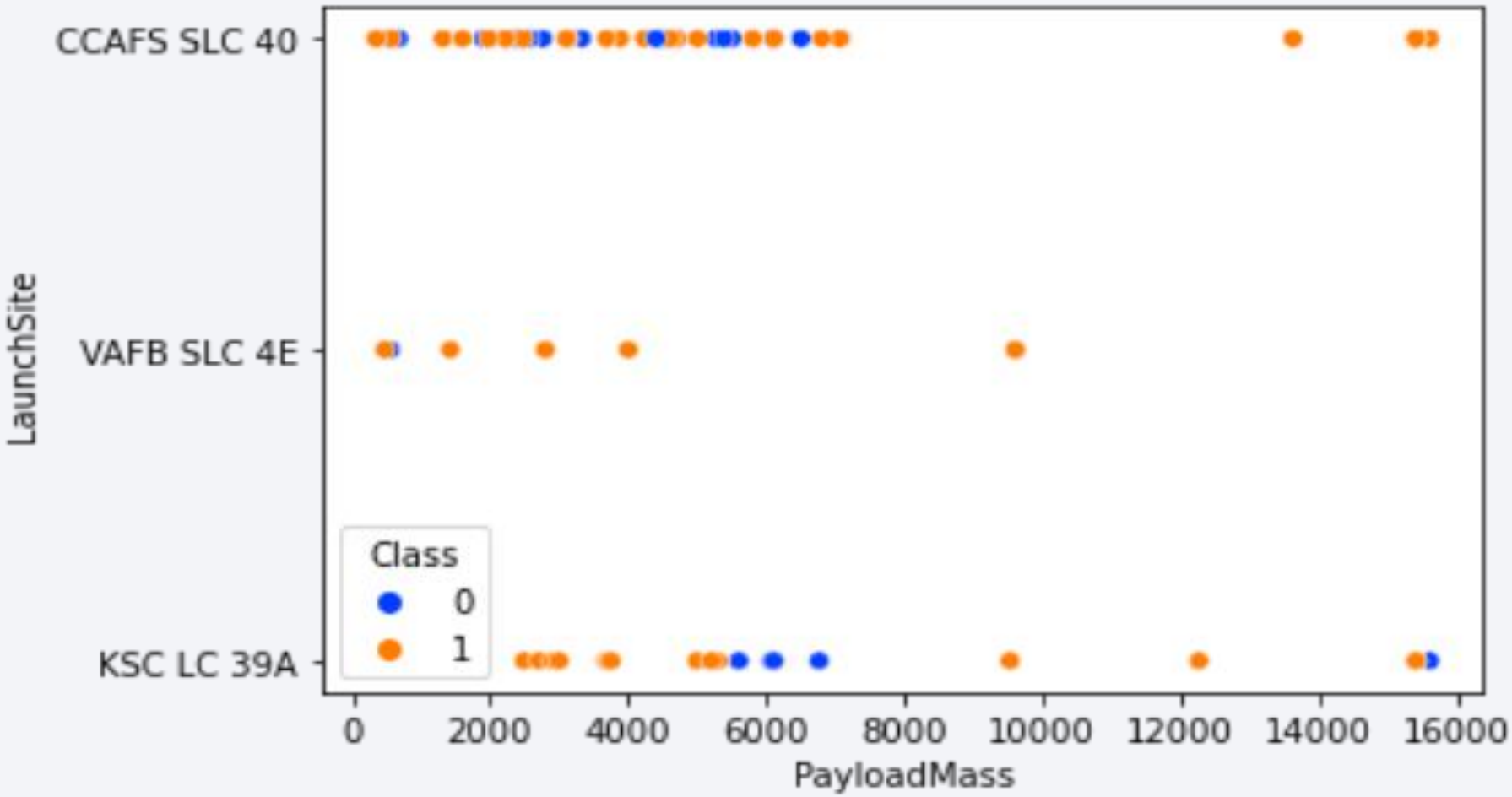
Section 2

Insights drawn from EDA

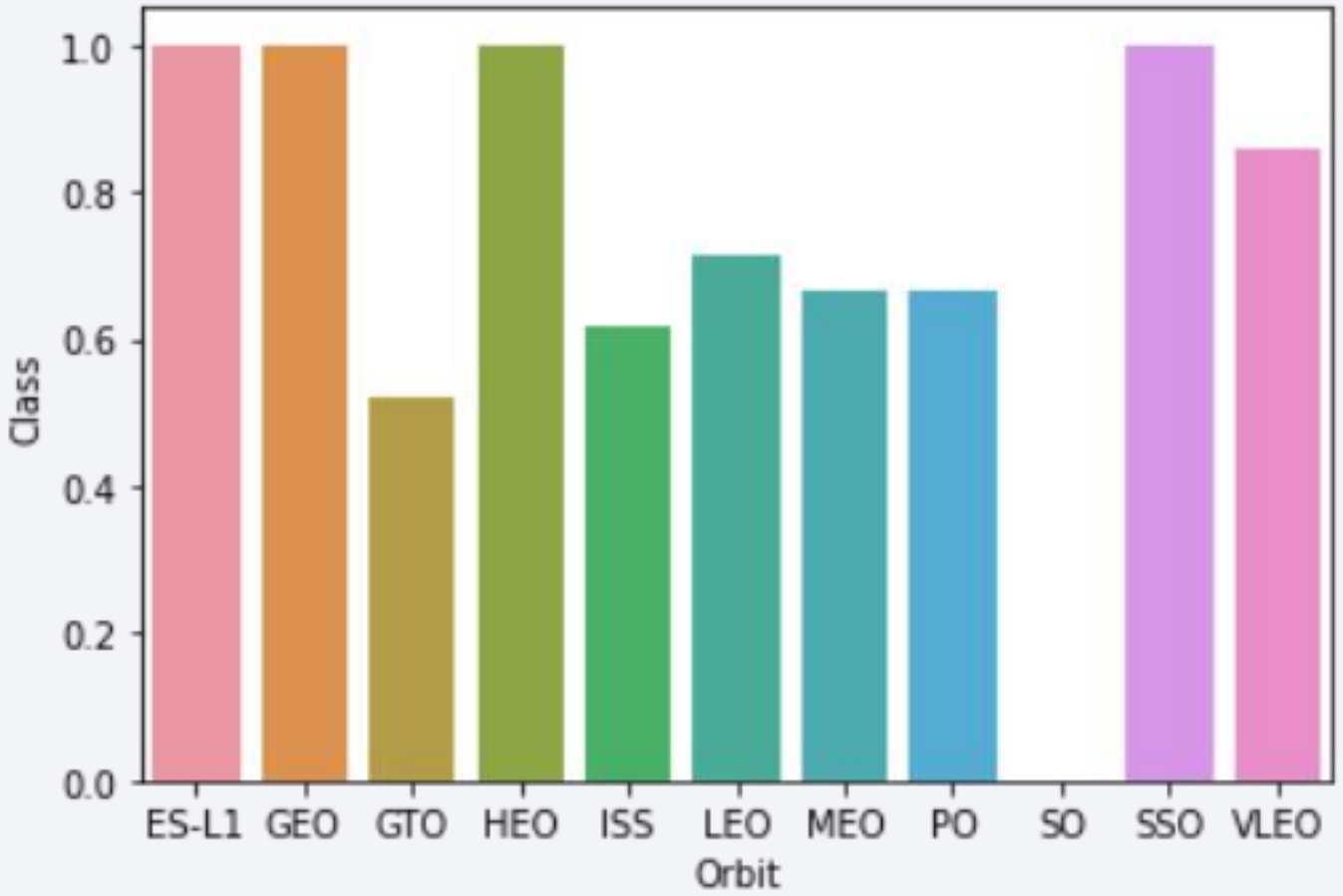
Flight Number vs. Launch Site



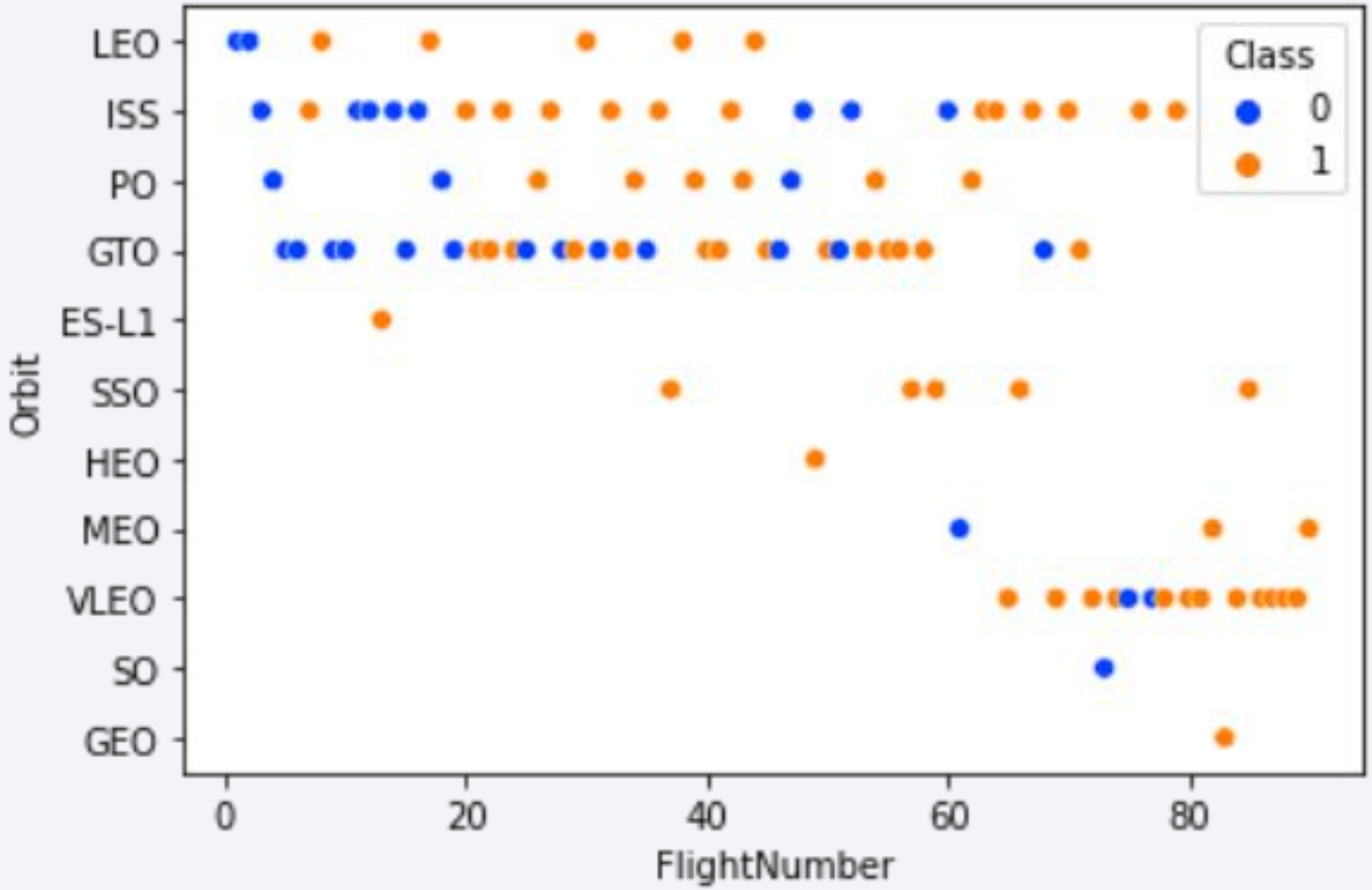
Payload vs. Launch Site



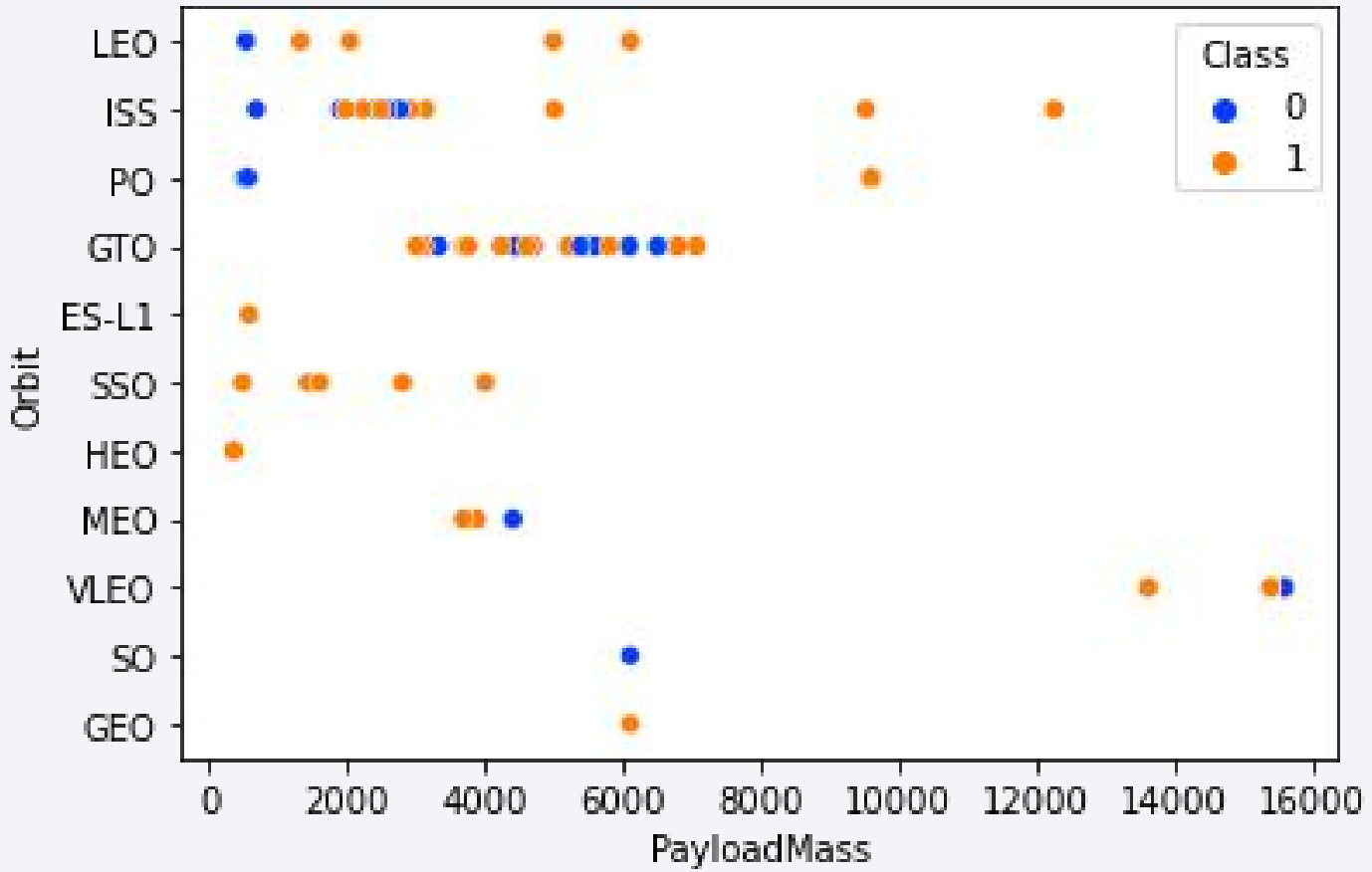
Success Rate vs. Orbit Type



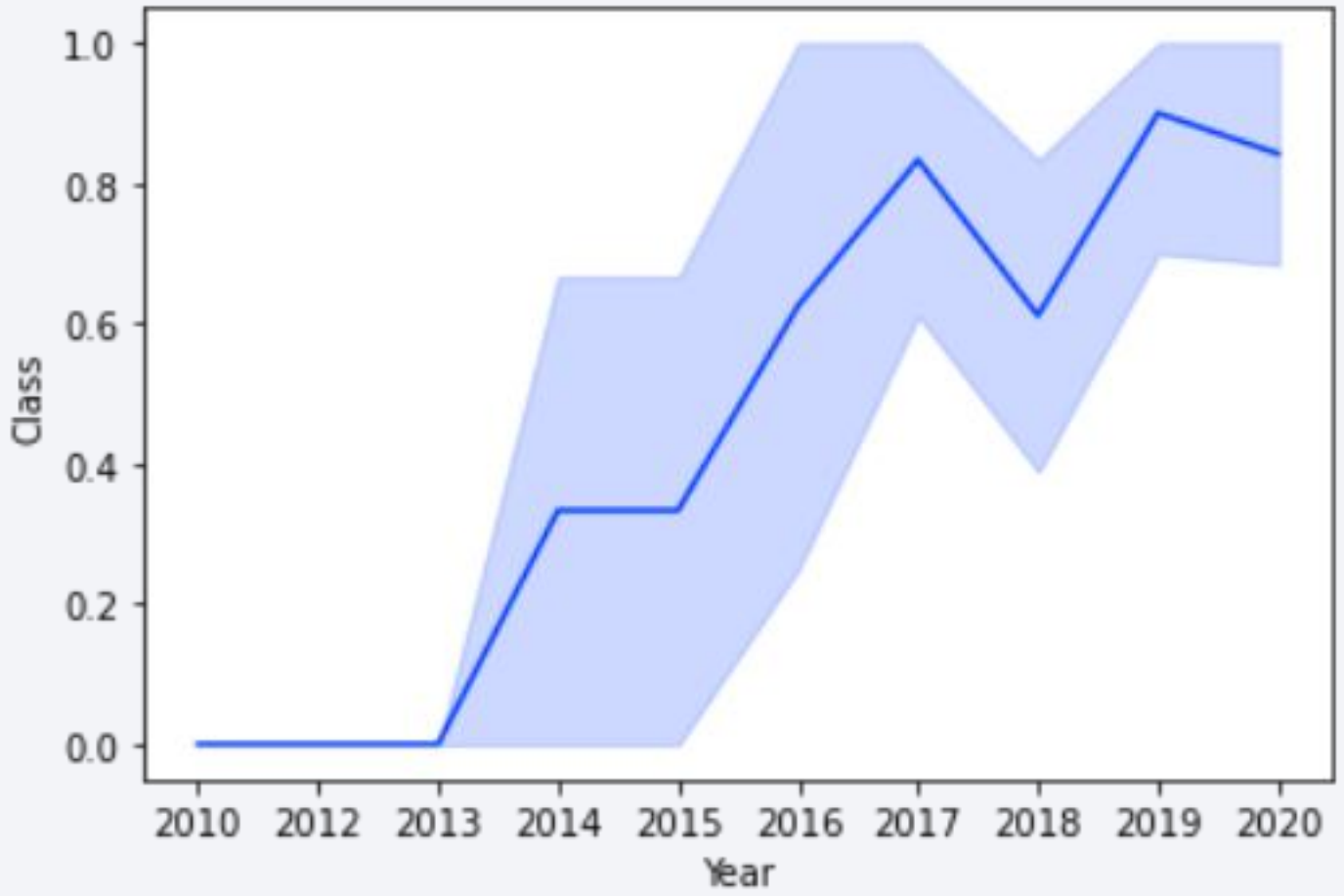
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

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

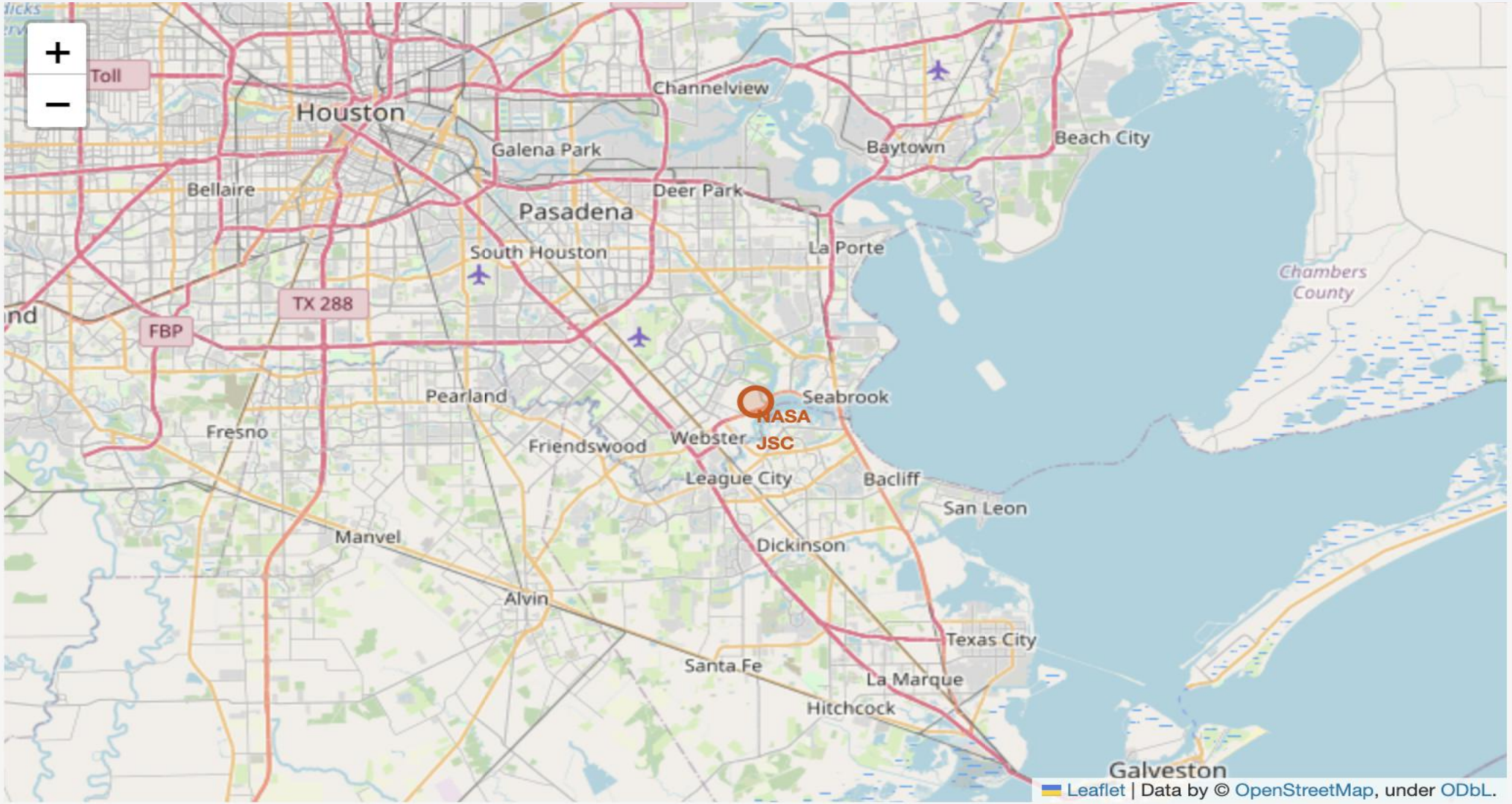
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

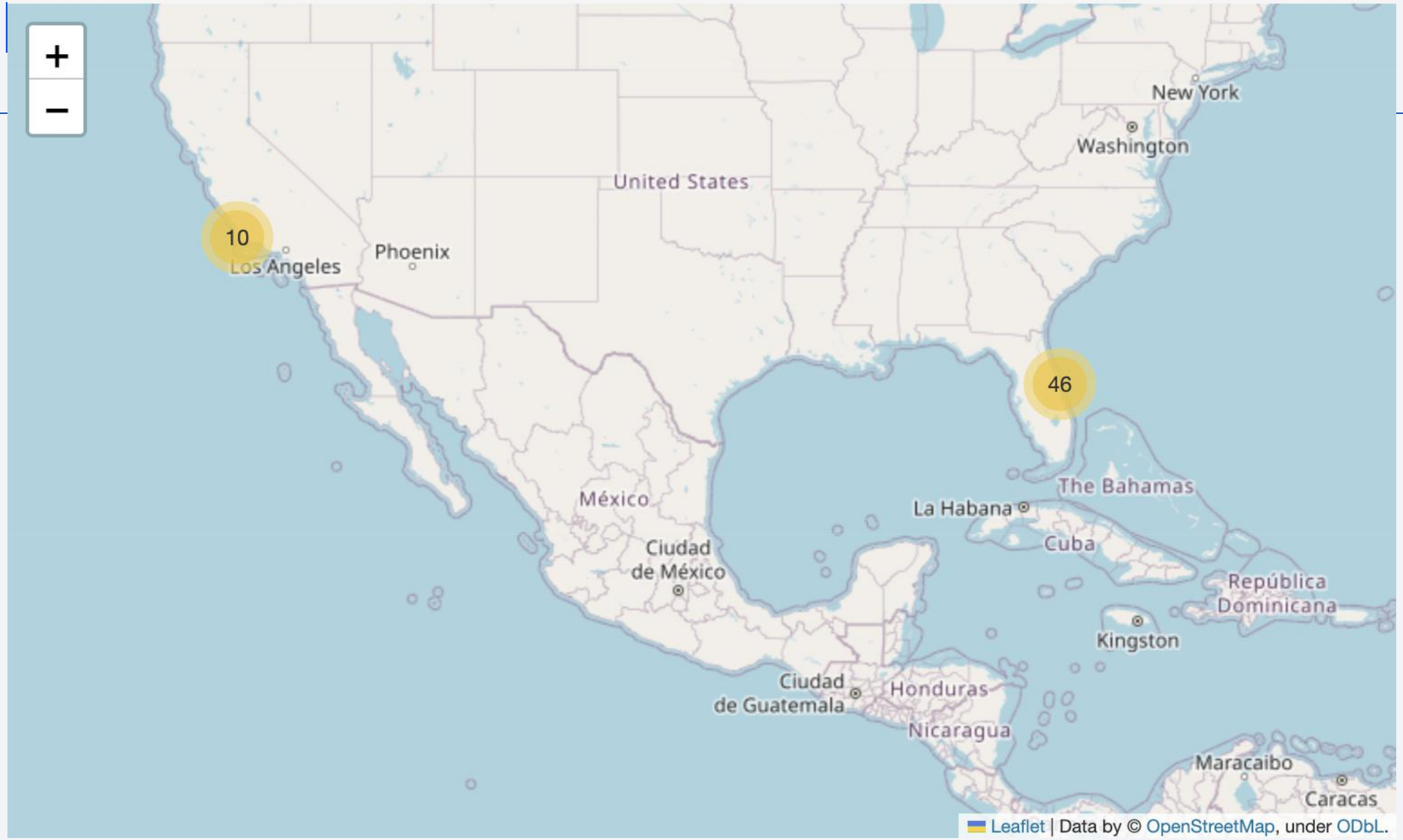
- Present your query result with a short explanation here

Section 3

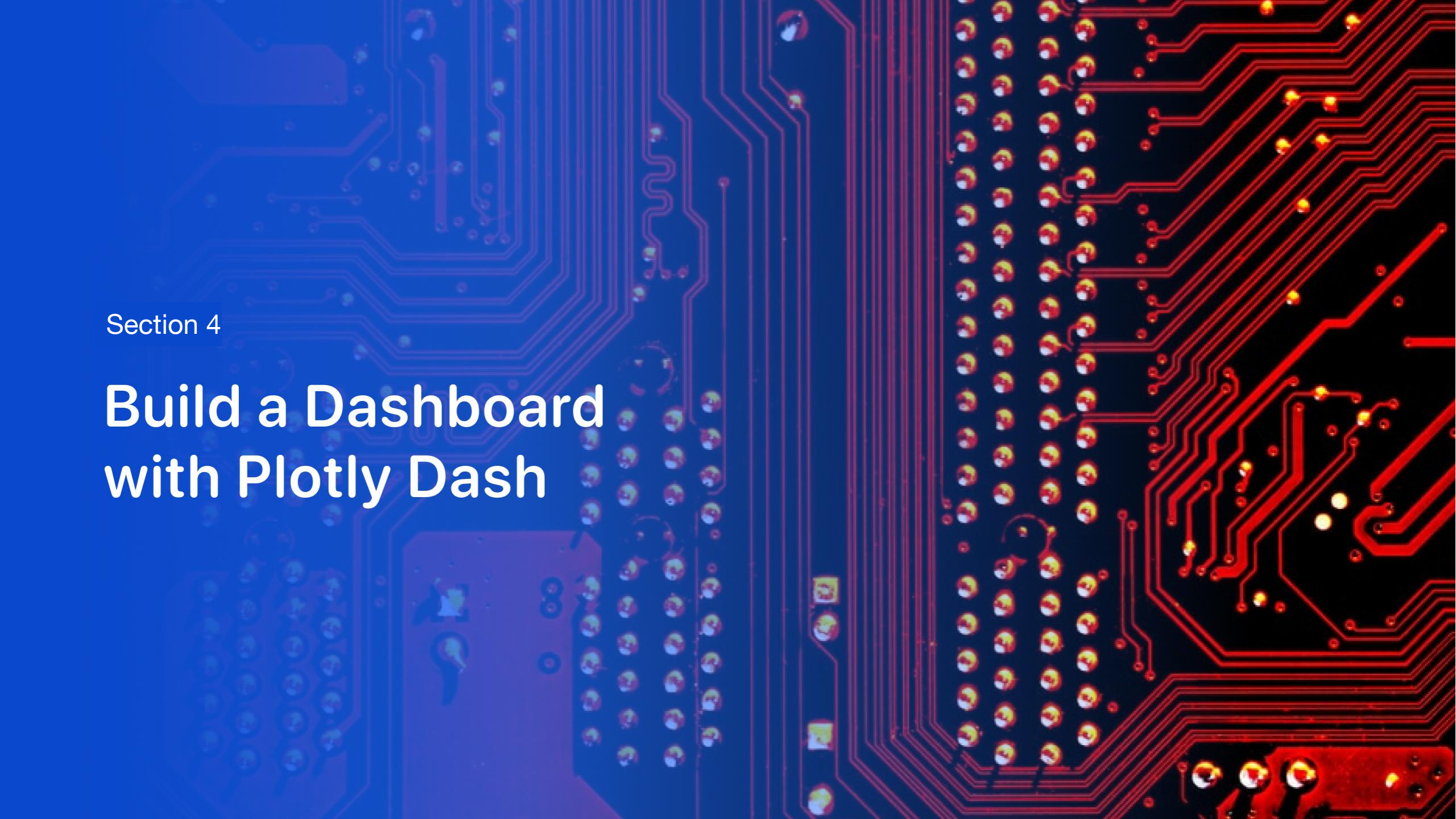
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>




+
−
New York
Washington
United States
10
Los Angeles
Phoenix
46
México
Ciudad
de México
La Habana
The Bahamas
Cuba
República
Dominicana
Kingston
Ciudad
de Guatemala
Honduras
Nicaragua
Maracaibo
Caracas
Leaflet | Data by © OpenStreetMap, under ODbL.

<Folium Map Screenshot 3>


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

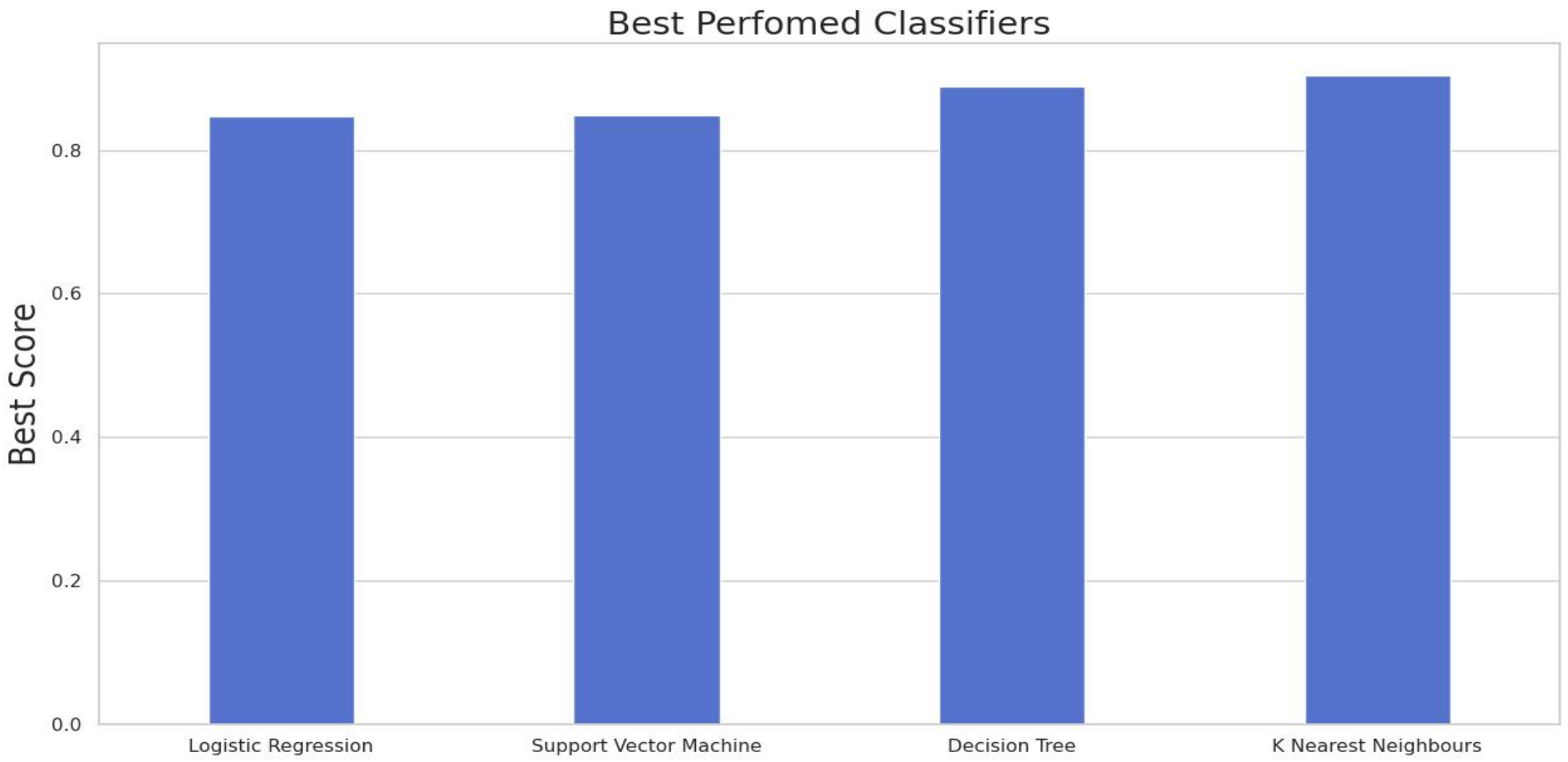
<Dashboard Screenshot 3>

- 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.

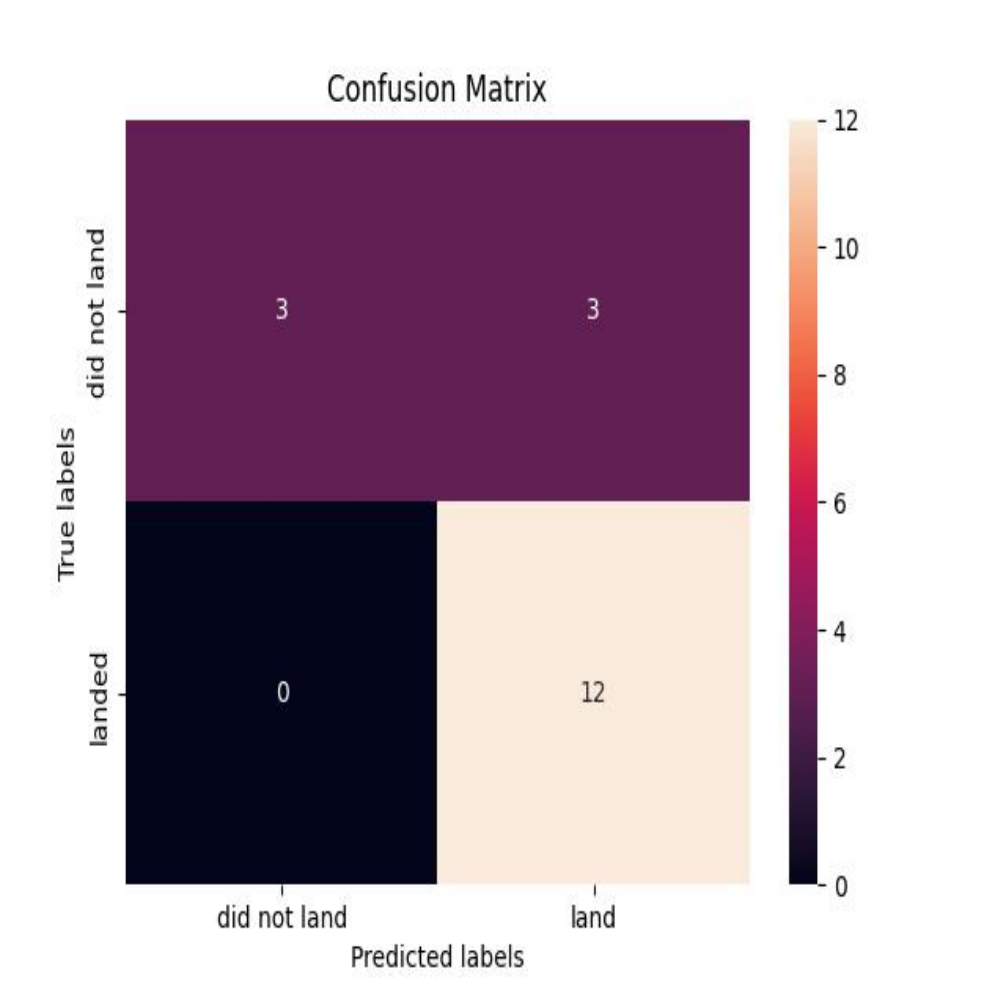
Section 5

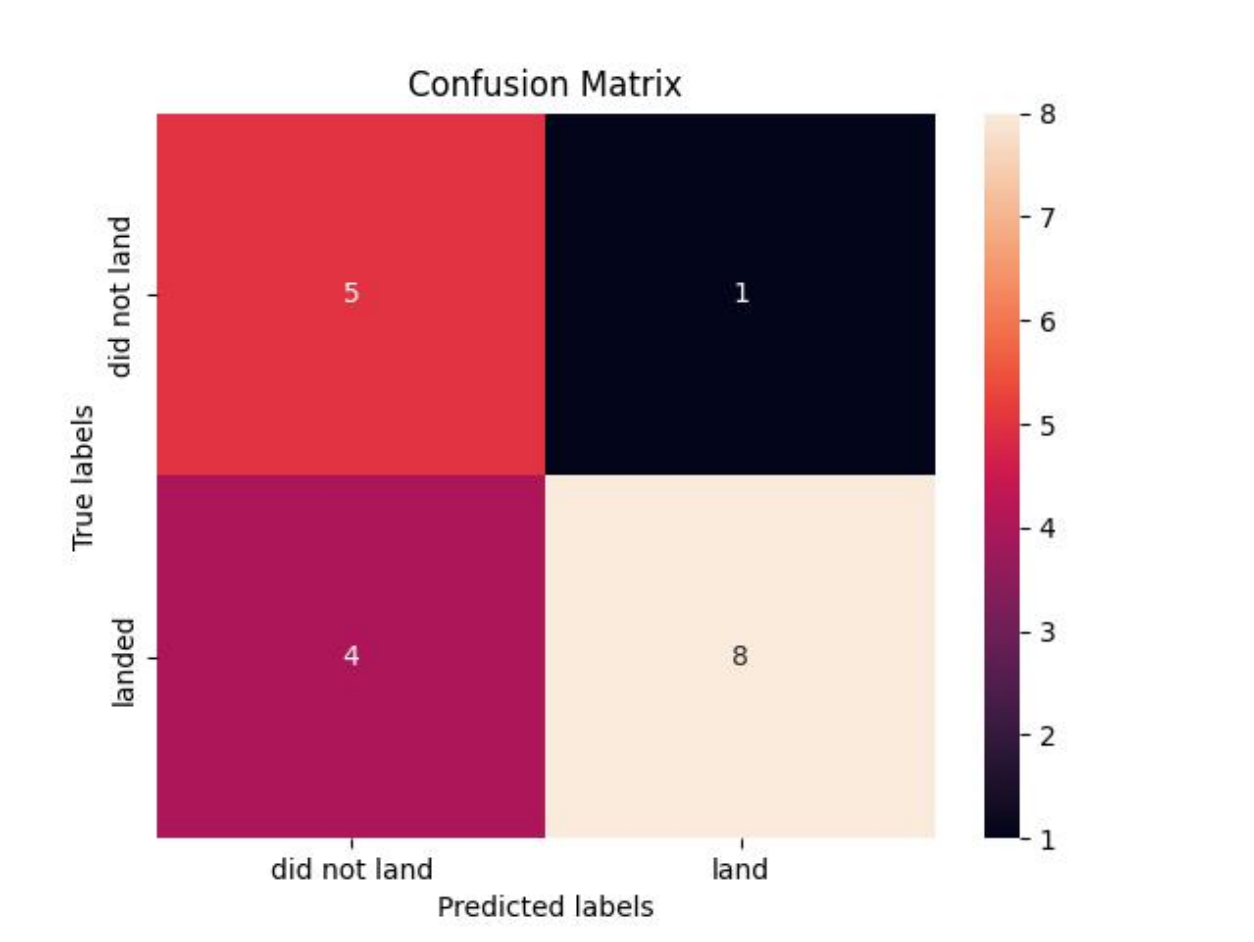
Predictive Analysis (Classification)

Classification Accuracy



Confusion Matrix





Conclusions

- I found that the scores predicted by these models are generally similar, but they are different in detail.

	Classifiers	Accuracy Score	Best Score
0	Logistic Regression	0.833333	0.846429
1	Support Vector Machine	0.833333	0.848214
2	Decision Tree	0.777778	0.889286
3	K Nearest Neighbours	0.722222	0.903571

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!