Winning Space Race with Data Science

Aditya Kumar
23 JAN 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with Data Visualization
  - EDA 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

Project background and context

- The era of commercial space has arrived
- Several companies are making space travel affordable for everyone
- SpaceX is the most successful of them

- their rocket launch is relatively inexpensive
- SpaceX advertises Falcon 9 rocket launches, with a cost of 62 million dollars.
- It's because SpaceX can reuse the first stage
- Therefore, we will predict if the Falcon 9 first stage will land successfully

Problems you want to find answers

- Correlations between each rocket variables and successful landing rate
- Conditions to get the best results and ensure the best successful landing rate

Methodology

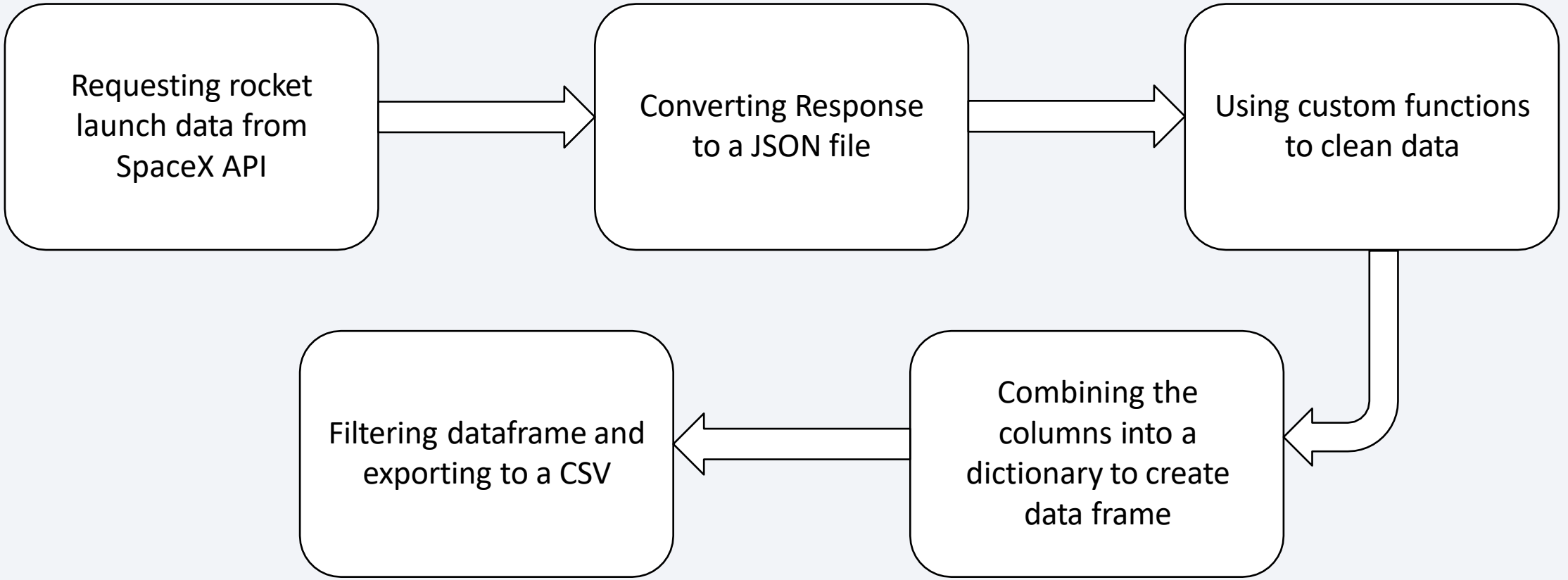
Executive Summary

- Data collection methodology:
  - SpaceX API & Web Scraping SpaceX Wikipedia Page
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

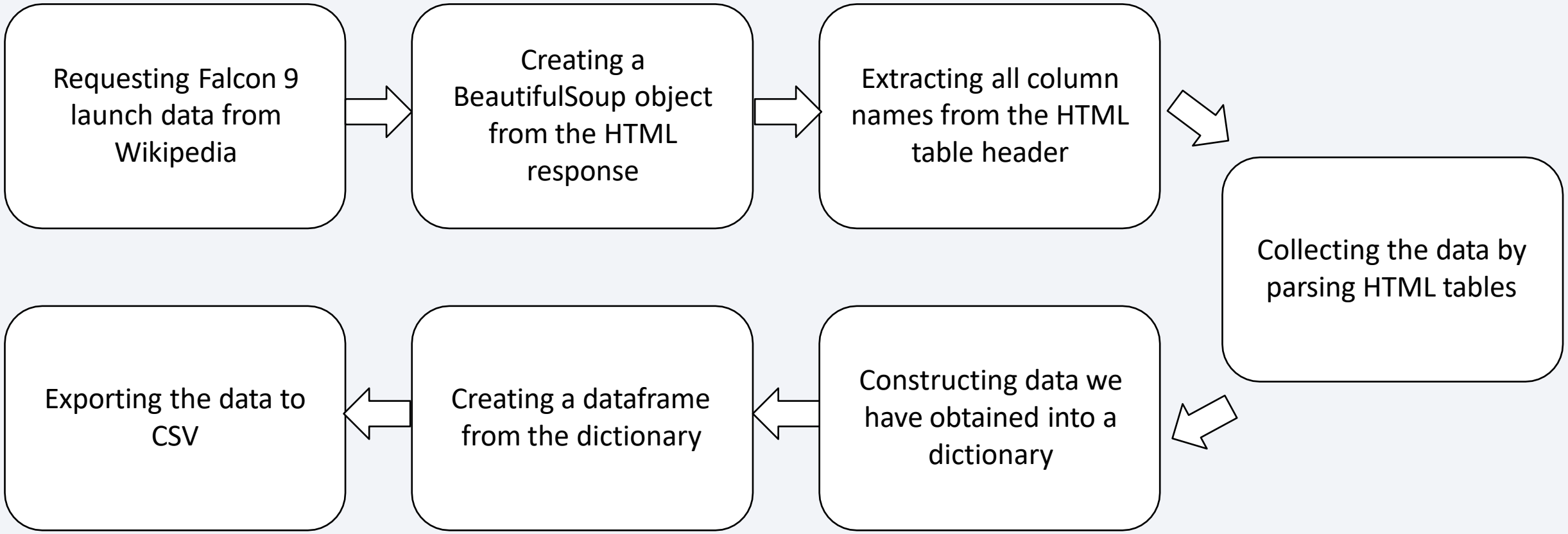
Data Collection

- Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia.
- Data Columns are obtained by using SpaceX REST API:
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Data Columns are obtained by using Wikipedia Web Scraping:
  Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection - SpaceX API

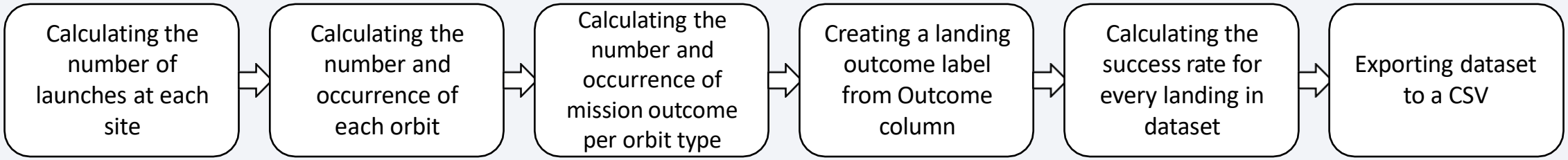


Data Collection - Scraping



Data Wrangling

- There are several cases in which the booster failed to successfully land on the dataset, and sometimes it attempted to land but failed because of accident.
- Converting the results into training labels:
  - 1 = successful / 0 = failure



EDA with Data Visualization

Scatter chart:

- Flight Number vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type, Payload vs. Orbit Type.
- A scatter plot shows how much one variable is affected by another. The relationship between two variables is called a correlation. This plot is generally composed of large data bodies.

Bar chart:

- Orbit Type vs. Success Rate.
- A Bar chart makes it easy to compare datasets between multiple groups at a glance. One axis represents a category and the other axis represents a discrete value. The purpose of this chart is to indicate the relationship between the two axes

Line chart:

- Year vs. Success Rate
- A Line chart shows data variables and trends very clearly and helps predict the results of data that has not yet been recorded

EDA with SQL

- Performed SQL queries:
  - Displaying the names of the unique launch sites in the space mission
  - Displaying 5 records where launch sites begin with the string ‘CCA’
  - Displaying the total payload mass carried by boosters launched by NASA (CRS)
  - Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date when the first successful landing outcome in ground pad was achieved
  - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster versions which have carried the maximum payload mass

Build an Interactive Map with Folium

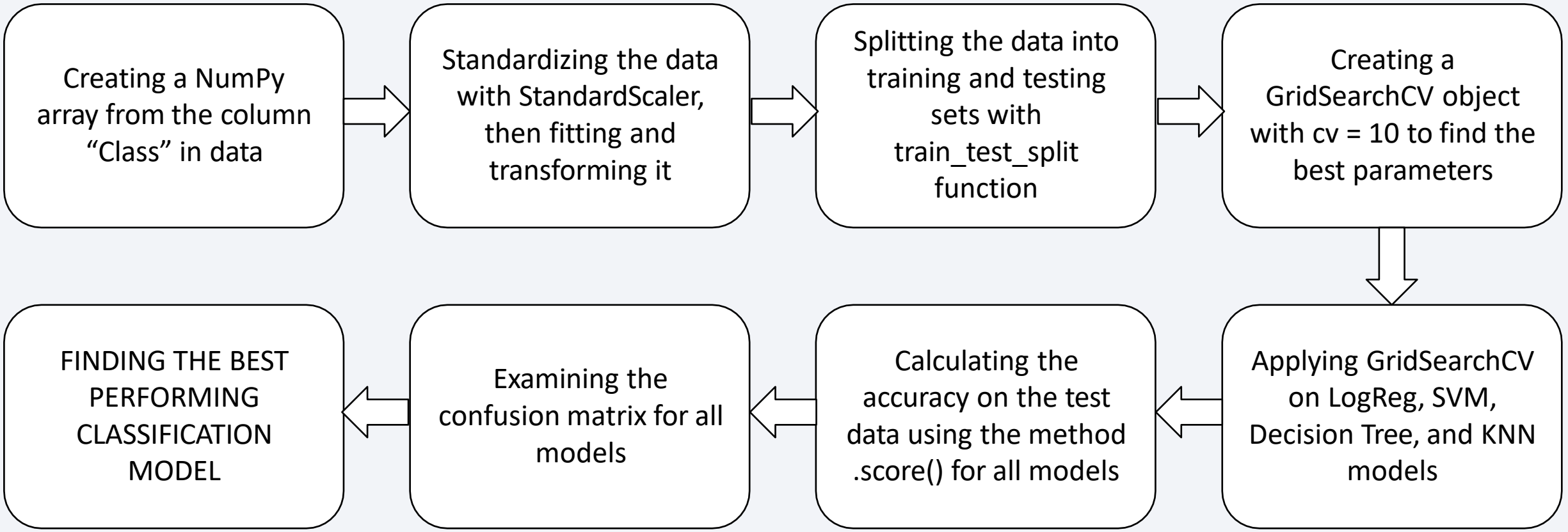
- Markers of all Launch Sites:
  - Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.
  - Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.
- Coloured Markers of the launch outcomes for each Launch Site:
  - Added coloured 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 Cit

Build a Dashboard with Plotly Dash

The dashboard application contains a pie chart and a scatter point chart.

- Pie chart
  - For be selected to indicate a successful landing distribution across all launch sites or to indshowing total success launches by sites
  - This chart can icate the success rate of individual launch sites.
- Scatter chart
  - For showing the relationship between Outcomes and Payload mass(Kg) by different boosters
  - Has 2 inputs: All sites/individual site & Payload mass on a slider between 0 and 10000 kg
  - This chart helps determine how success depends on the launch point, payload mass, and booster version categories

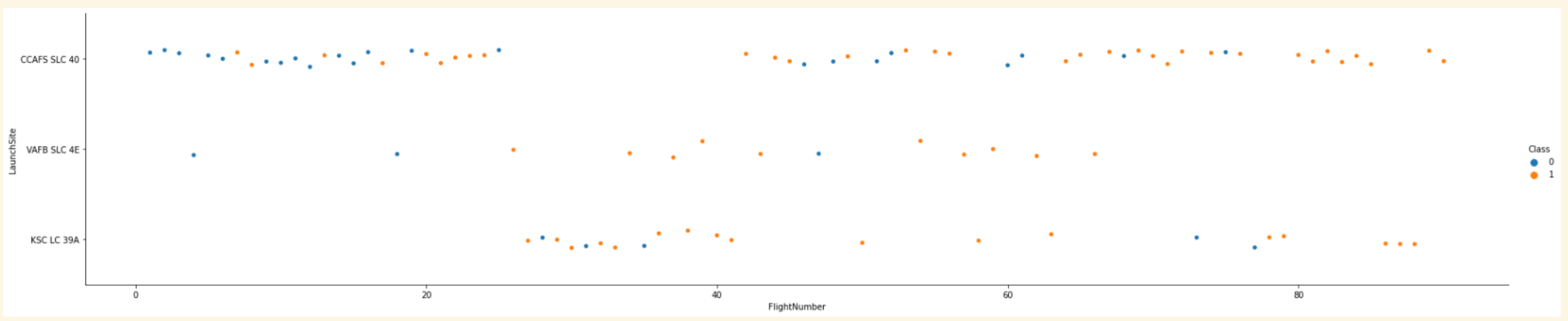
Predictive Analysis (Classification)



Results

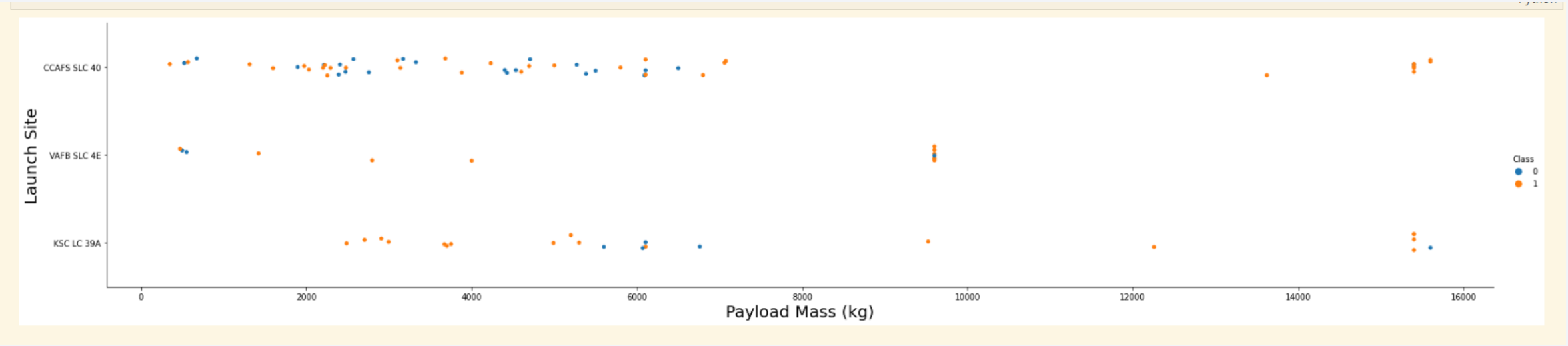
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Flight Number vs. Launch Site



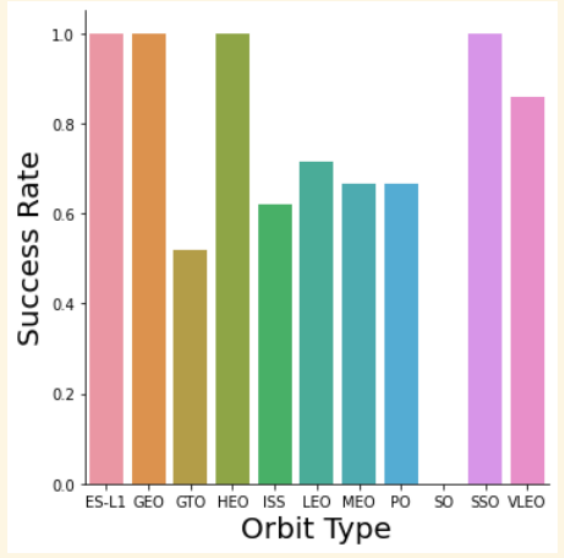
Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



For every launch site the higher the payload mass, the higher the success rate. Most of the launches with payload mass over 7000 kg were successful. KSCLC39A has a 100% success rate for payload mass under 5500 kg too

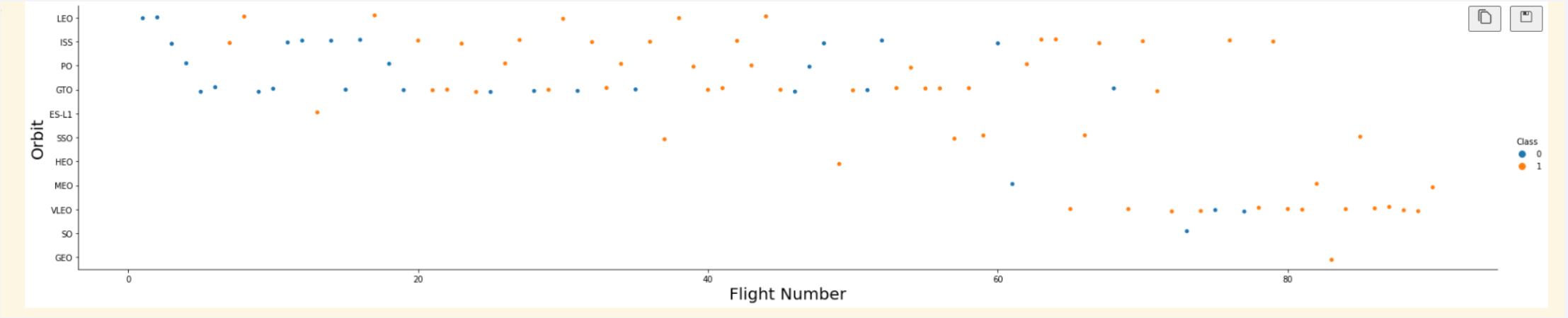
Success Rate vs. Orbit Type



Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%).

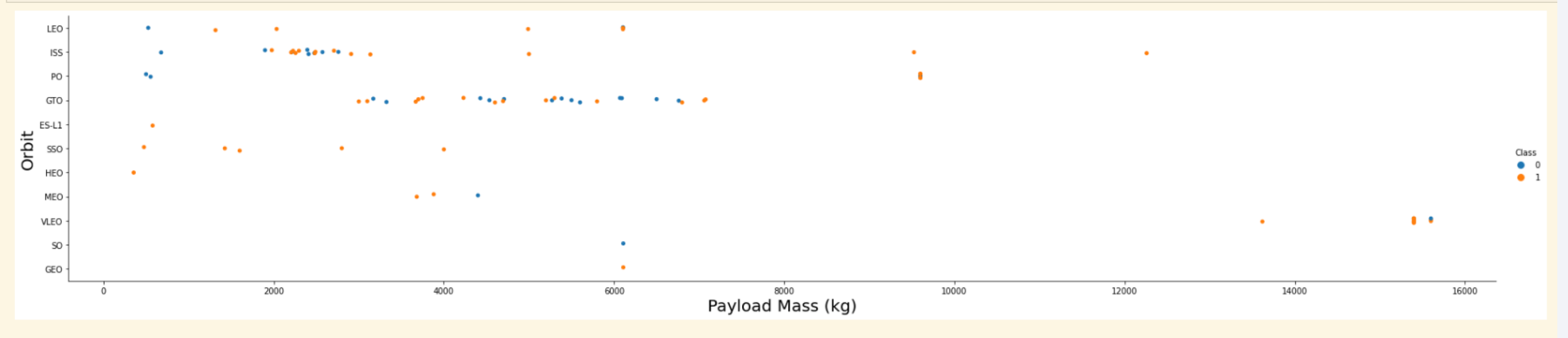
On the other hand, the success rate of orbit type GTO is only 50%, and it is the lowest except for type SO, which recorded failure in a single attempt

Flight Number vs. Orbit Type



Launch Orbit preferences changed over Flight Number.
Launch Outcome seems to correlate with this preference.
SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type

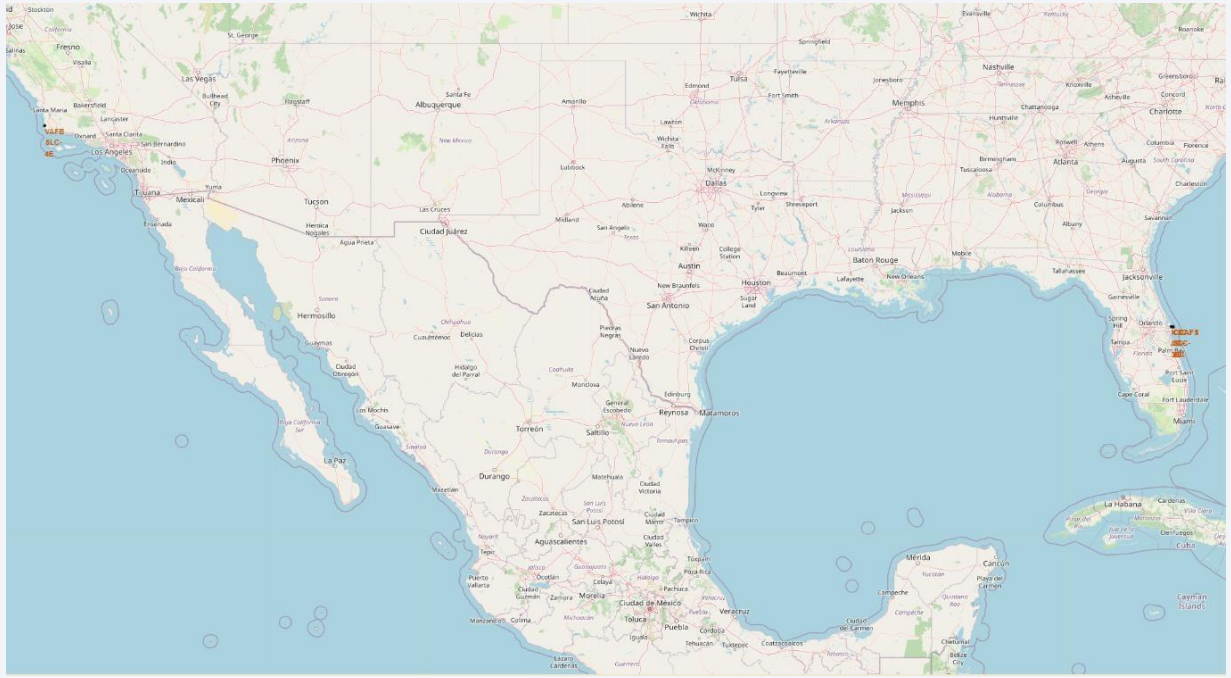


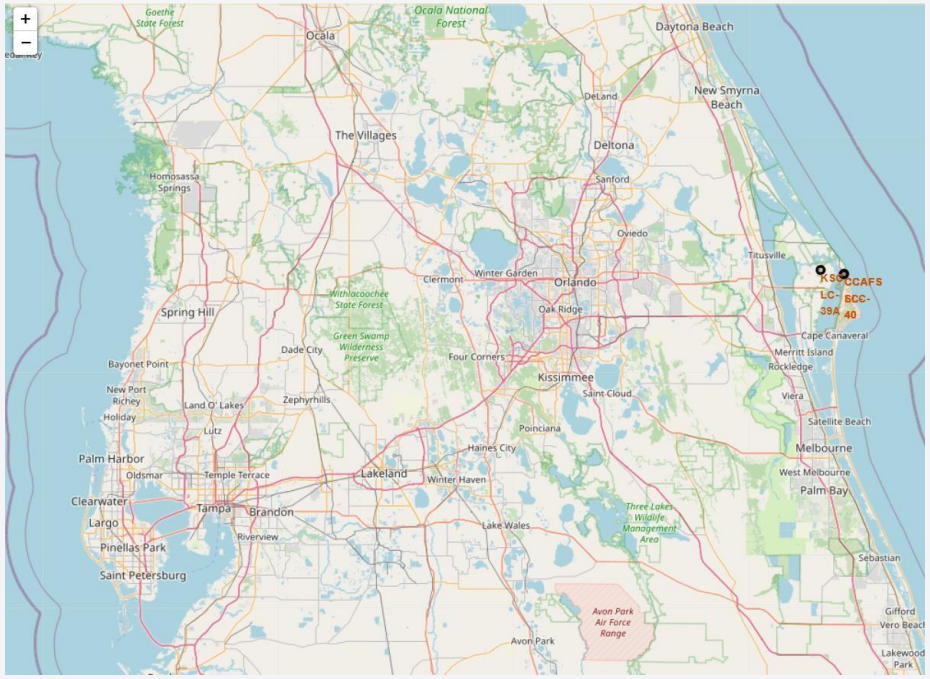
Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

The other most successful orbit VLEO only has payload mass values in the higher end of the range

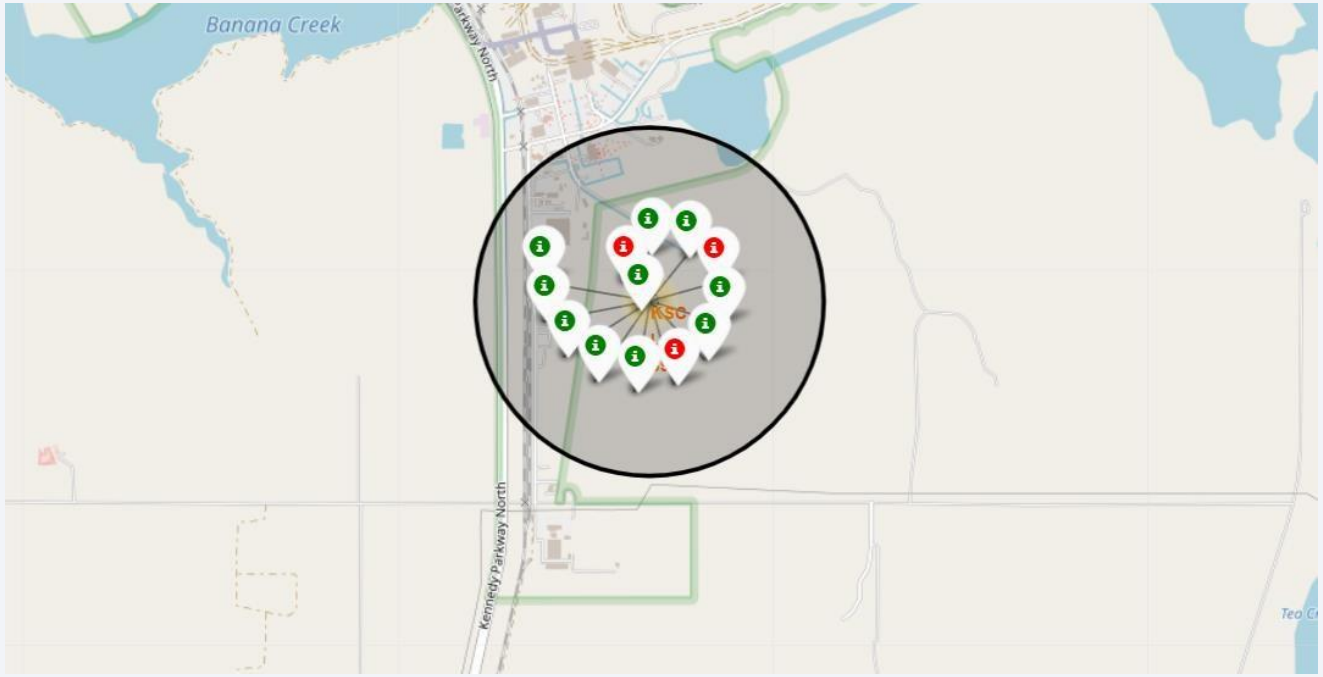
All launch sites' location markers on a global map





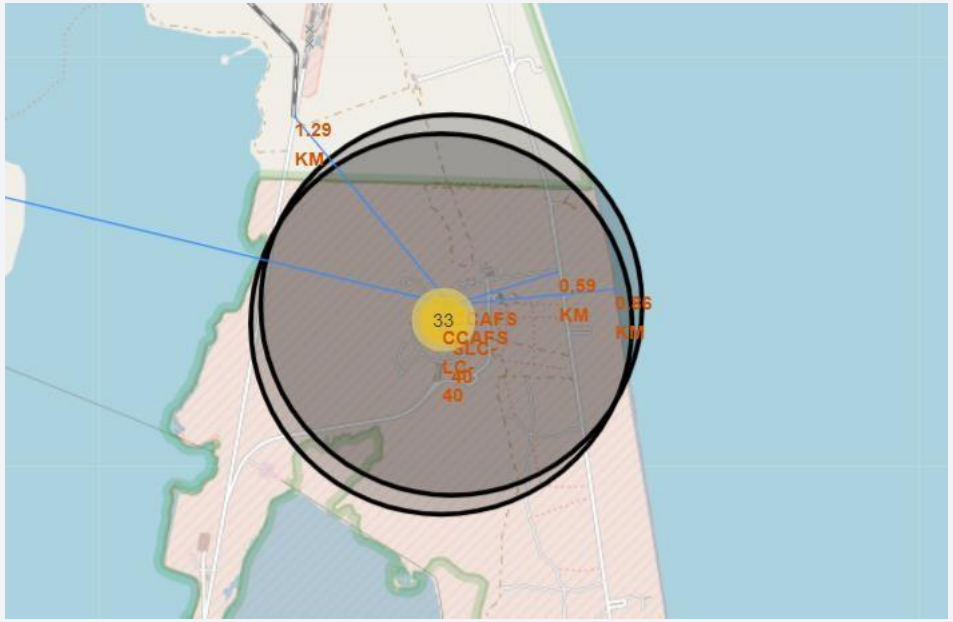
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean

Colour-labeled launch records on the map



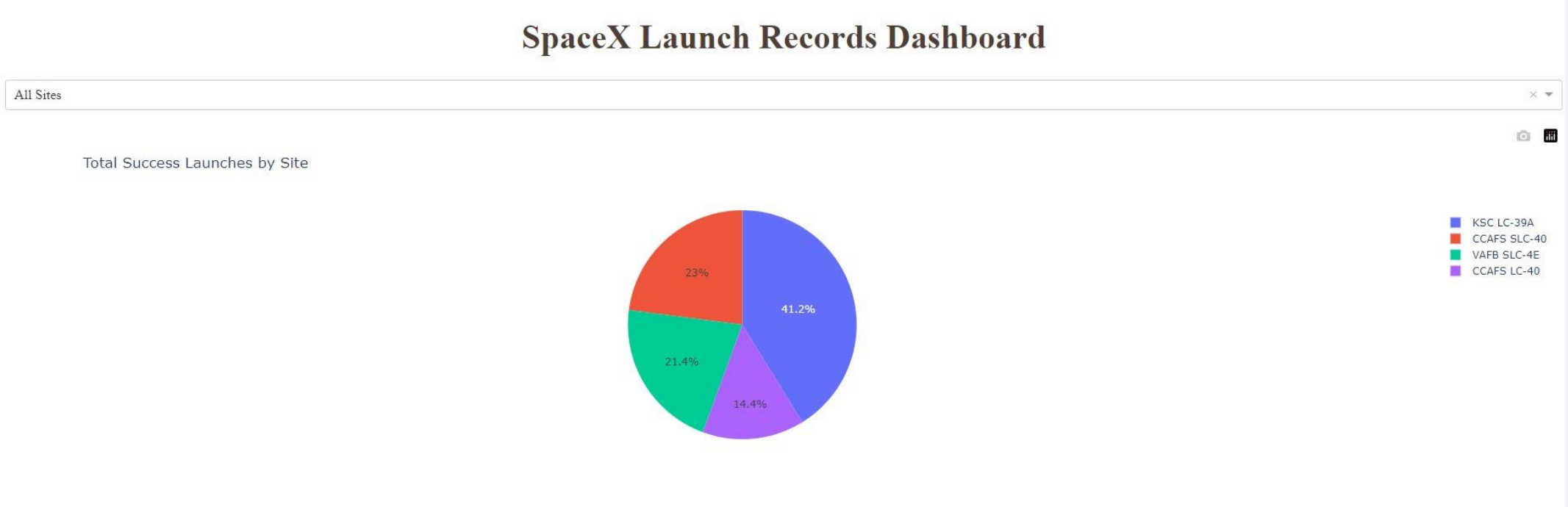
Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

Proximities of Launch Sites



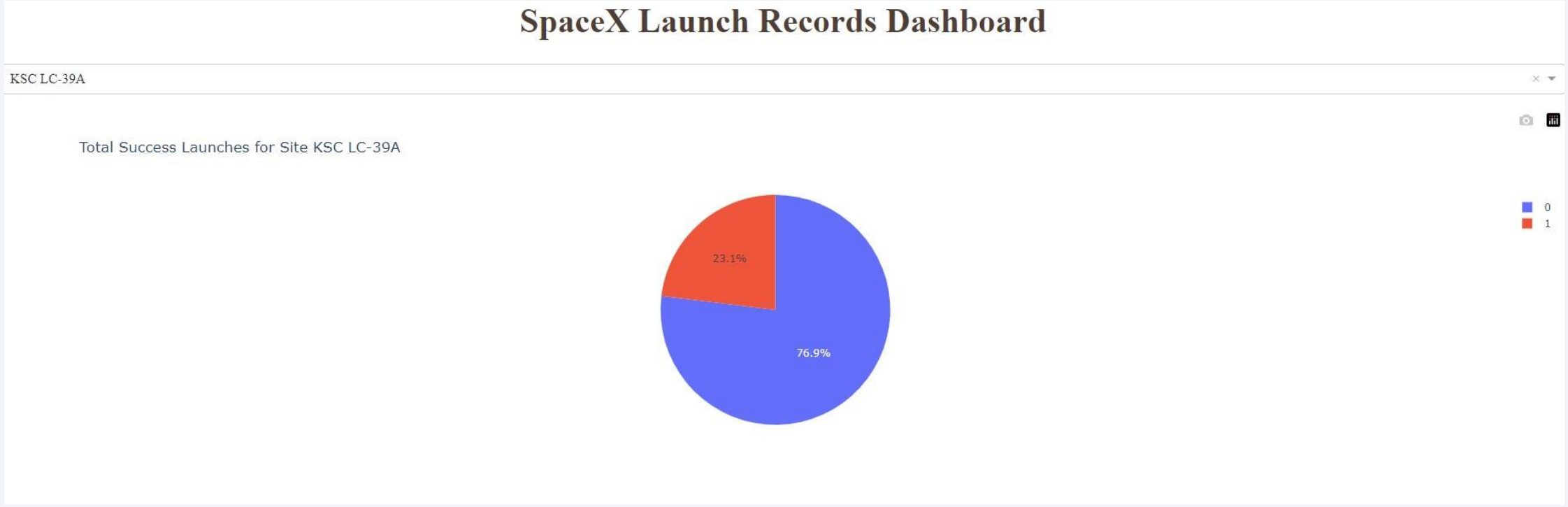
It can be found that the launch site is close to railways and highways for transportation of equipment or personnel, and is also close to coastline and relatively far from the cities so that launch failure does not pose a threat

Launch success count for all sites



The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

Launch site with highest launch success ratio

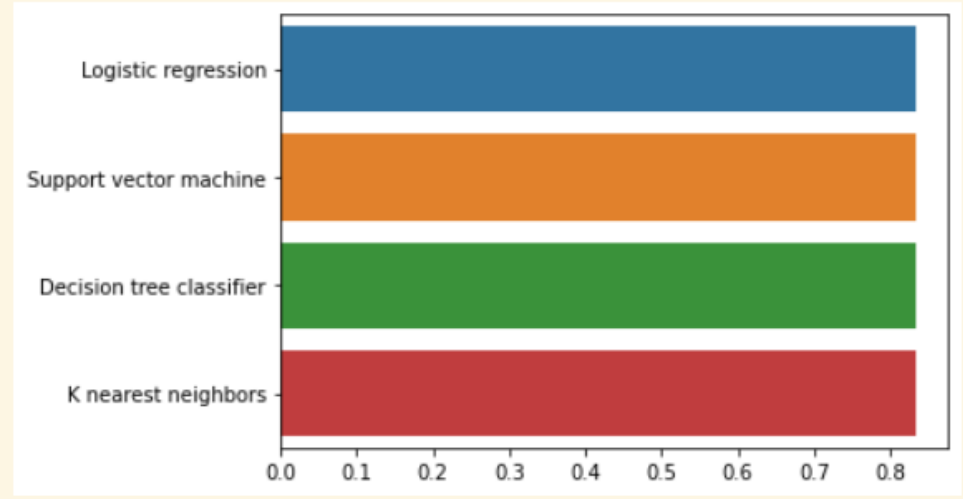


KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3 landing failures (23.1%)

<Dashboard Screenshot 3>

These figures show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is higher than that of heavy weighted payloads(5000-10000 kg)

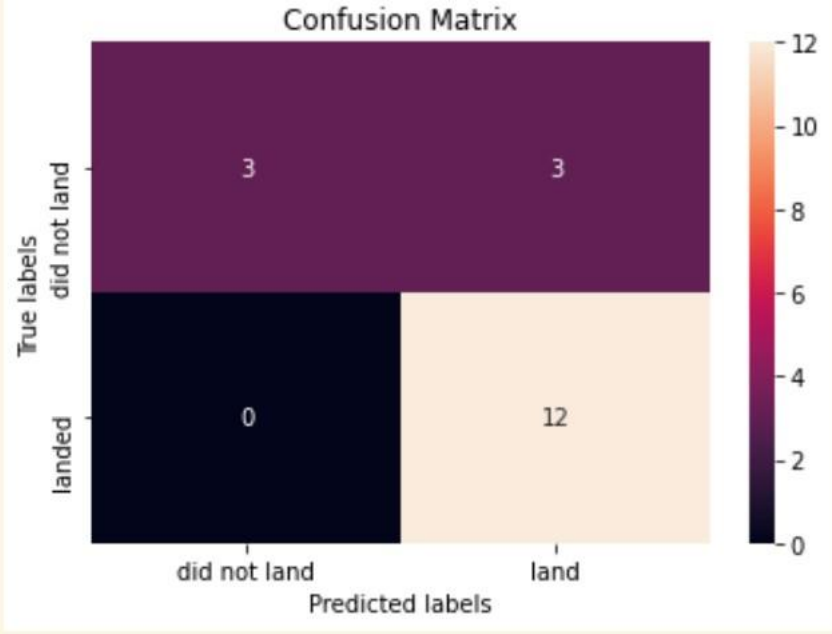
Classification Accuracy



	Method	Accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.833333
3	K nearest neighbors	0.833333

- In the test set, the accuracy of all models was virtually the same at 83.33%.
- It should be noted that the test size was small at 18.
- Therefore, more data is needed to determine the optimal model

Confusion Matrix



- The confusion matrix is the same for all models because all models performed the same for the test set.
- The models predicted :
  - 12 successful landings when the true label was successful
  - 3 failed landings when the true label was failure.
  - 3 predictions that said successful landings when the true label was failure (false positive).
- Overall, these models predict successful landings

Conclusions

- As the number of flights increased, the success rate increased, and recently it has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- The launch site is close to railways, highways, and coastline, but far from cities.
- KSLC-39A has the highest number of launch successes and the highest success rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heavy weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems that more data is needed to determine the optimal model due to the small data size.

Appendix

- GitHub URL

Thank you!