WINNING SPACE RACE WITH DATA SCIENCE

DEEPIKA UTTAM SAMBREKAR

OUTLINE

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

EXECUTIVE SUMMARY

- Acquiring data through web scraping and the SpaceX API.
- Conducting Exploratory Data Analysis (EDA), which encompassed tasks such as data wrangling, data visualization, and interactive visual analytics.
- Utilizing Machine Learning for predictive modeling.
- Summarizing all findings.
- Valuable data was successfully gathered from publicly available sources.
- EDA facilitated the identification of key features for predicting the success of launches.
- The Machine Learning Prediction phase determined the optimal model for understanding the crucial characteristics that contribute to a successful launch, leveraging the entire dataset.

INTRODUCTION

- The objective is to evaluate Space Y's possible competitiveness against Space X, with an emphasis on gaining the following important insights

METHODOLOGY

Executive Summary

- Data collection methodology
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

METHODOLOGY

DATA COLLECTION

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/ and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

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: https://github.com/tflores/applied-data science capstone/blob/master/Data%20Collection%20API.ipynb

DATA COLLECTION - SCRAPING

- Information regarding SpaceX launches is additionally accessible through Wikipedia.
- Data is retrieved from Wikipedia following the outlined flowchart and subsequently stored for future reference.

DATA WRANGLING

- Initially, exploratory data analysis (EDA) was conducted on the dataset.
- Following that, summaries were generated for launches at each site, occurrences of each orbit, and occurrences of mission outcomes based on orbit types.
- Lastly, a landing outcome label was created from the information in the Outcome column.

EDA WITH DATA VISUALIZATION

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:
- Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

EDA WITH DATA VISUALIZATION

- To explore data, scatterplots and barplots were used to visualize the relationship between pair of features
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

EDA WITH SQL

- Retrieving the names of unique launch sites in space missions.
- Identifying the top 5 launch sites with names starting with 'CCA.'
- Calculating the total payload mass carried by boosters launched by NASA for CRS missions.
- Determining the average payload mass carried by booster version F9 v1.1.
- Finding the date of the first successful landing outcome on a ground pad.
- Listing the names of boosters with successful drone ship landings and payload mass between 4000 and 6000 kg.
- Summing up the total number of successful and failed mission outcomes.
- Identifying the names of booster versions that have carried the maximum payload mass.
- Retrieving information about failed landing outcomes on drone ships, including booster versions and launch site names, specifically for the year 2015.
- Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between June 4, 2010, and March 20, 2017.

BUILD AN INTERACTIVE MAP WITH FOLIUM

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site; and
  - Lines are used to indicate distances between two coordinates.

BUILD A DASHBOARD WITH PLOTLY DASH

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

PREDICTIVE ANALYSIS (CLASSIFICATION)

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

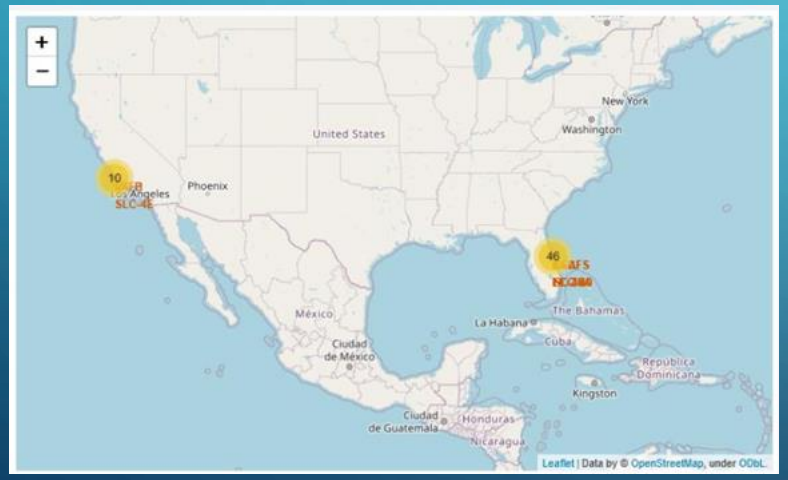
RESULTS

Findings from the exploratory data analysis include:

- Space X utilizes four distinct launch sites.
- Initial launches were directed towards Space X itself and NASA.
- The average payload of the F9 v1.1 booster is 2,928 kg.
- The first successful landing outcome occurred in 2015, five years after the initial launch.
- Numerous Falcon 9 booster versions demonstrated successful landings on drone ships with payloads surpassing the average.
- Nearly 100% of mission outcomes were successful.
- In 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, experienced failed landings on drone ships.
- The success rate of landing outcomes improved over the years.

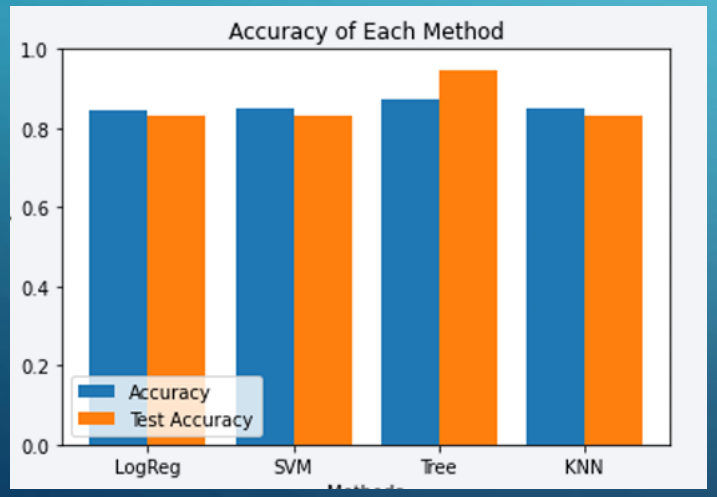
RESULTS

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.



RESULTS

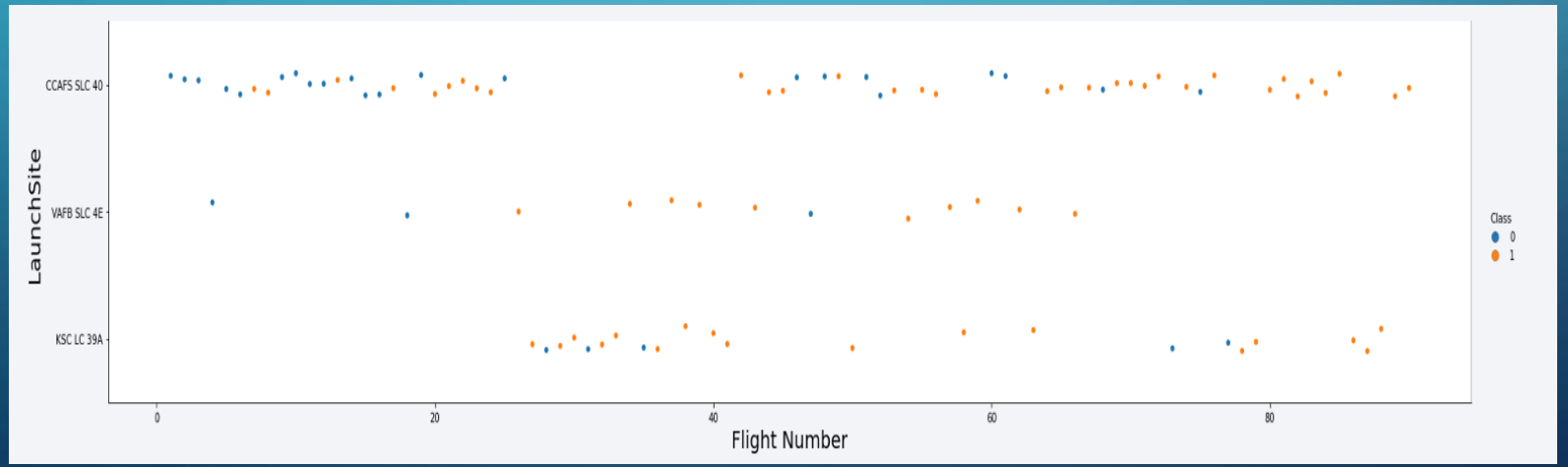
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 84% and accuracy for test data over 87%.



INSIGHTS DRAWN FROM EDA

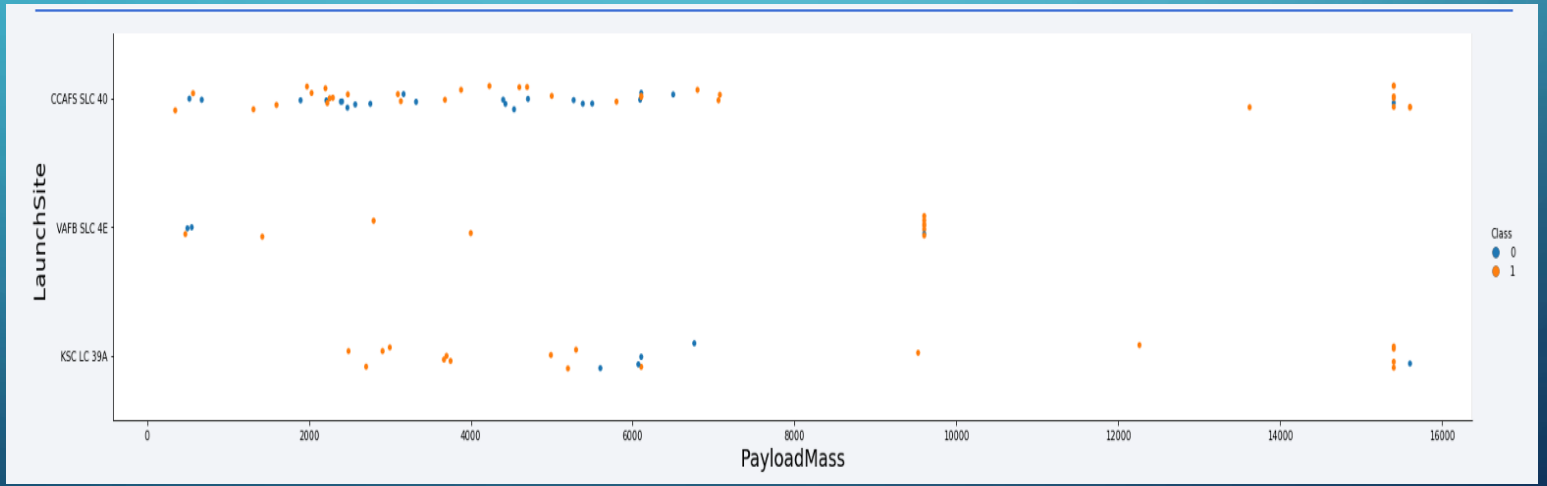
FLIGHT NUMBER VS. LAUNCH SITE

- Based on the depicted plot, it is evident that the most favorable launch site currently is CCAF5 SLC 40, with a majority of recent launches achieving success. Following closely in second place is VAFB SLC 4E, and in third place is KSC LC 39A. Additionally, the overall success rate has visibly improved over time.

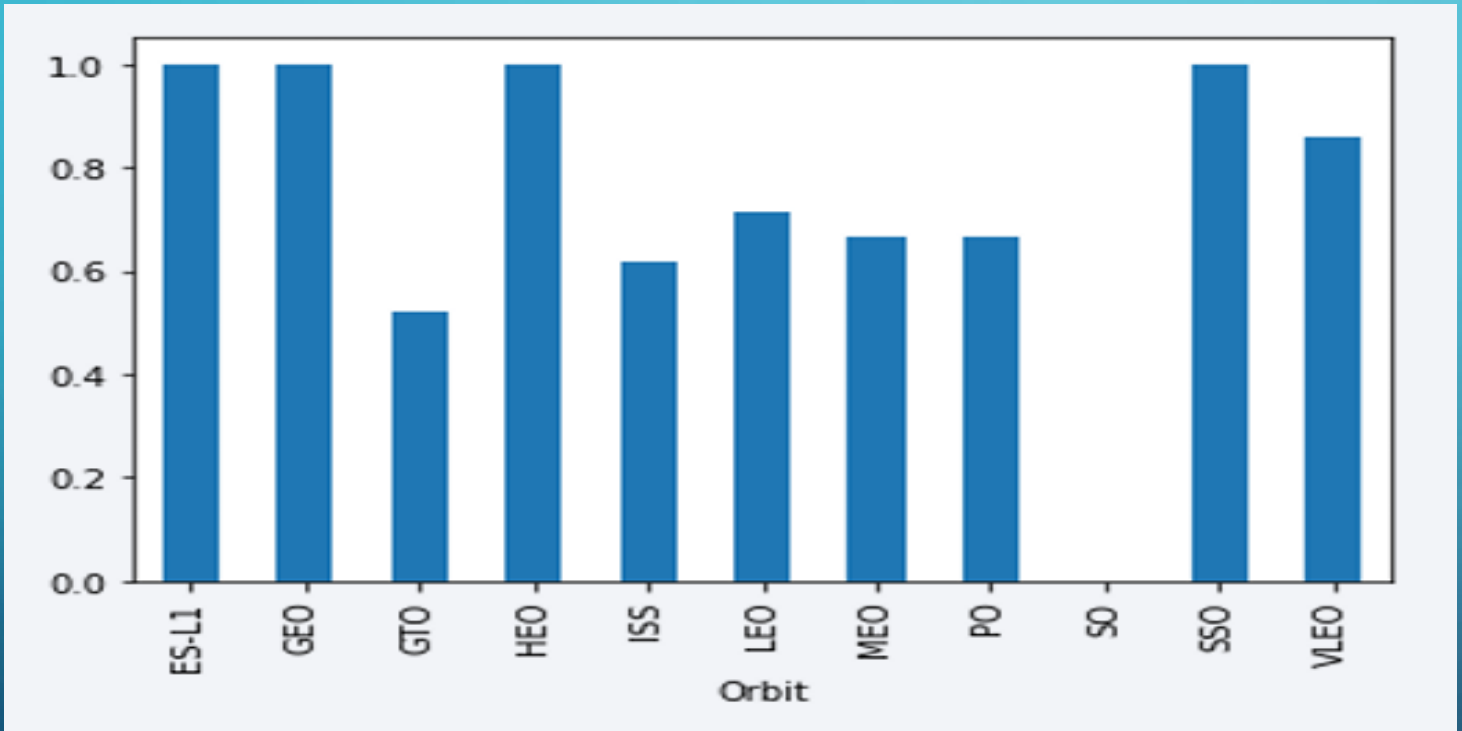


AYLOAD VS. LAUNCH SITE

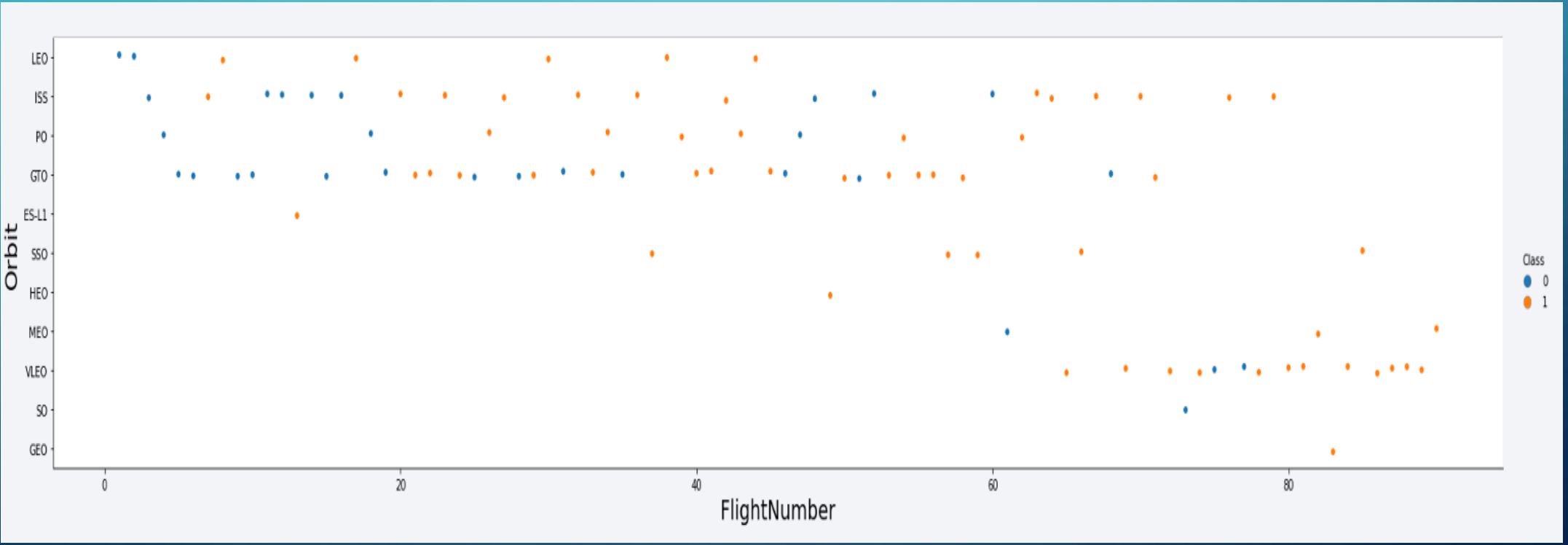
- Payloads exceeding 9,000kg, approximately the weight of a school bus, demonstrate an excellent success rate. Furthermore, payloads surpassing 12,000kg appear feasible primarily on the CCAFS SLC 40 and KSC LC 39A launch sites.



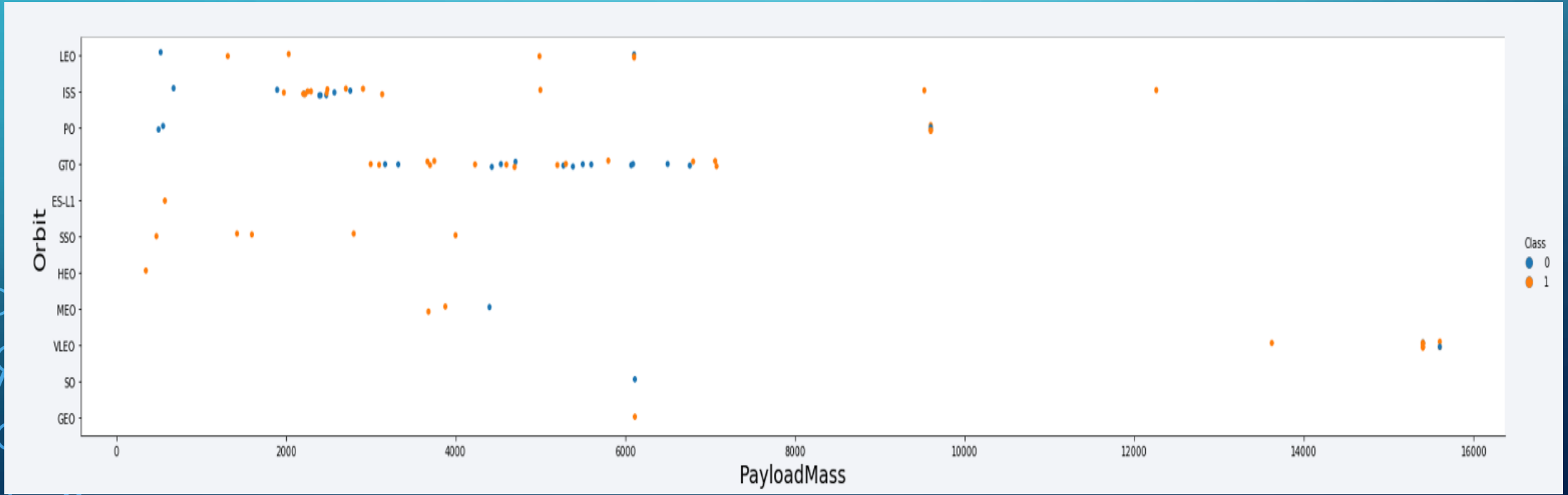
SUCCESS RATE VS. ORBIT TYPE



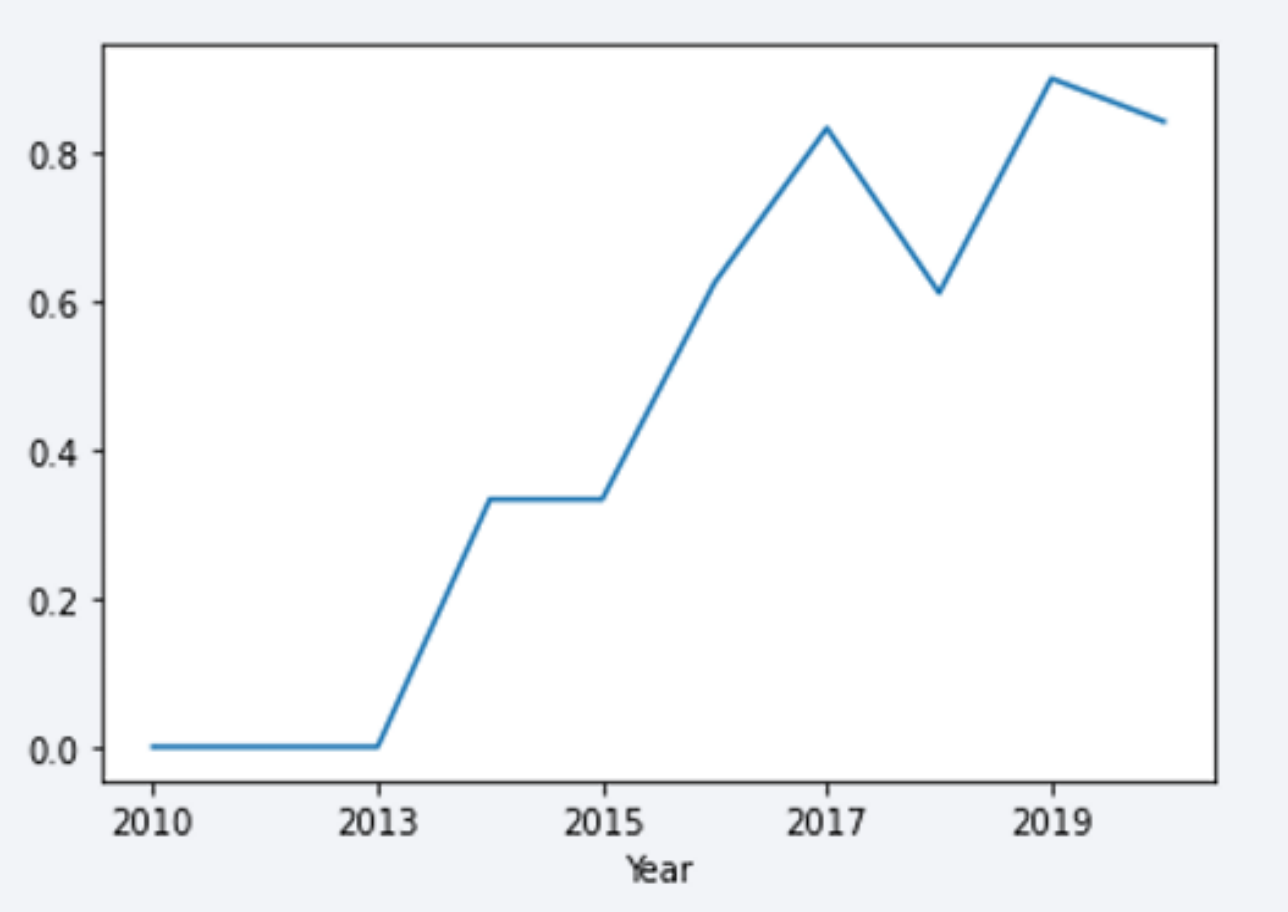
FLIGHT NUMBER VS. ORBIT TYPE



PAYLOAD VS. ORBIT TYPE



LAUNCH SUCCESS YEARLY TREND

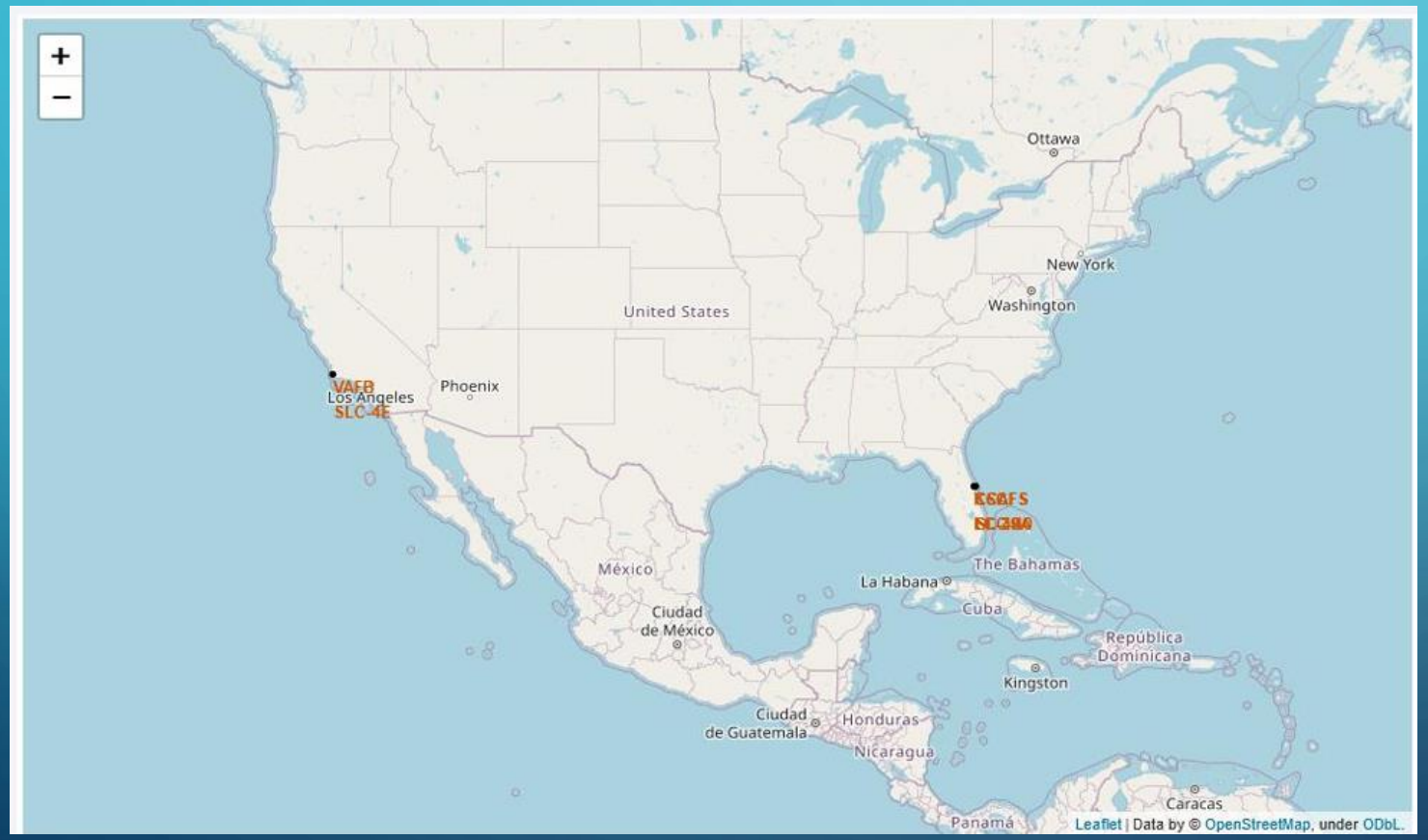


ALL LAUNCH SITE NAMES

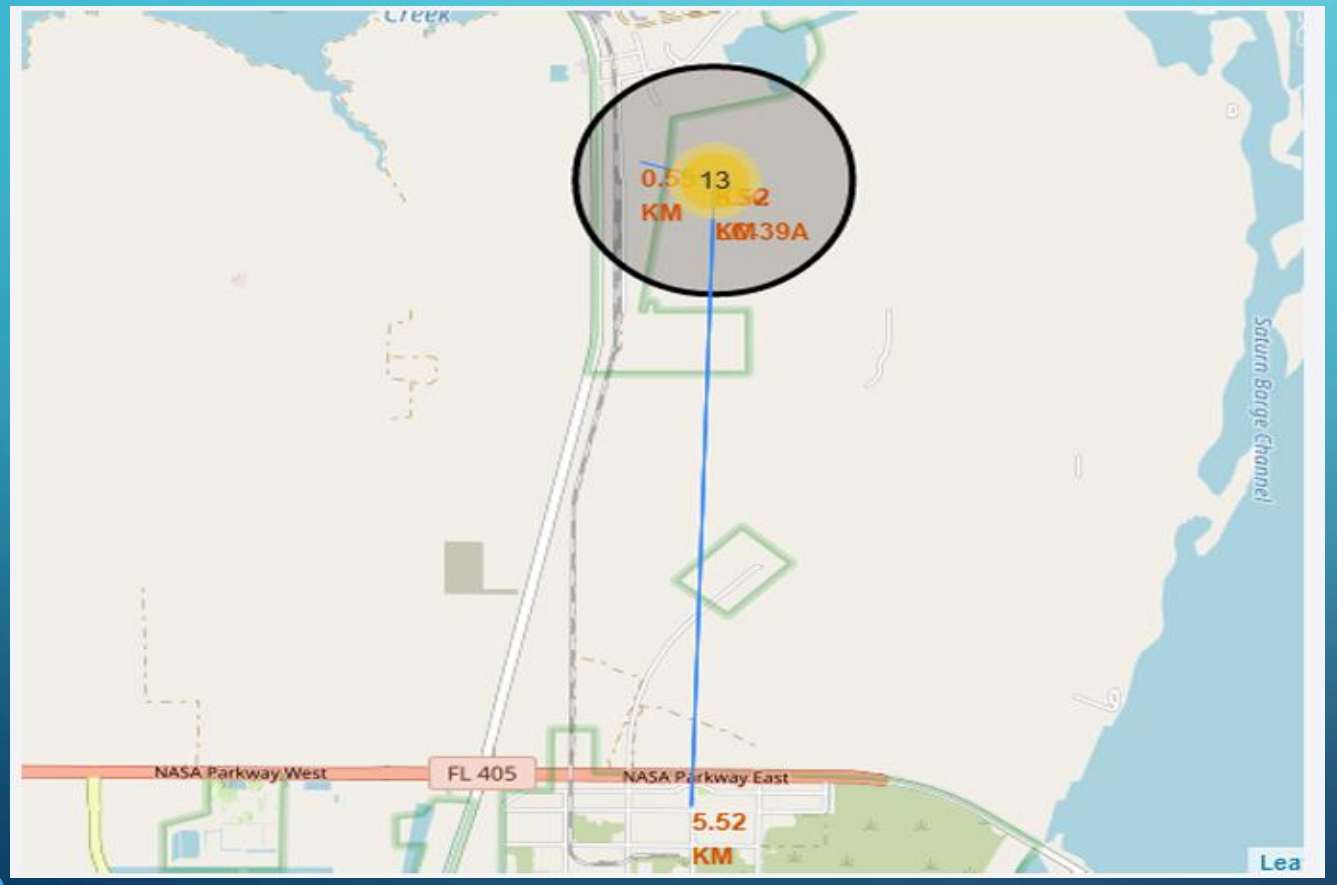
Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

LAUNCHE SITES PROXIMITIES ANALYSIS

ALL LAUNCH SITES

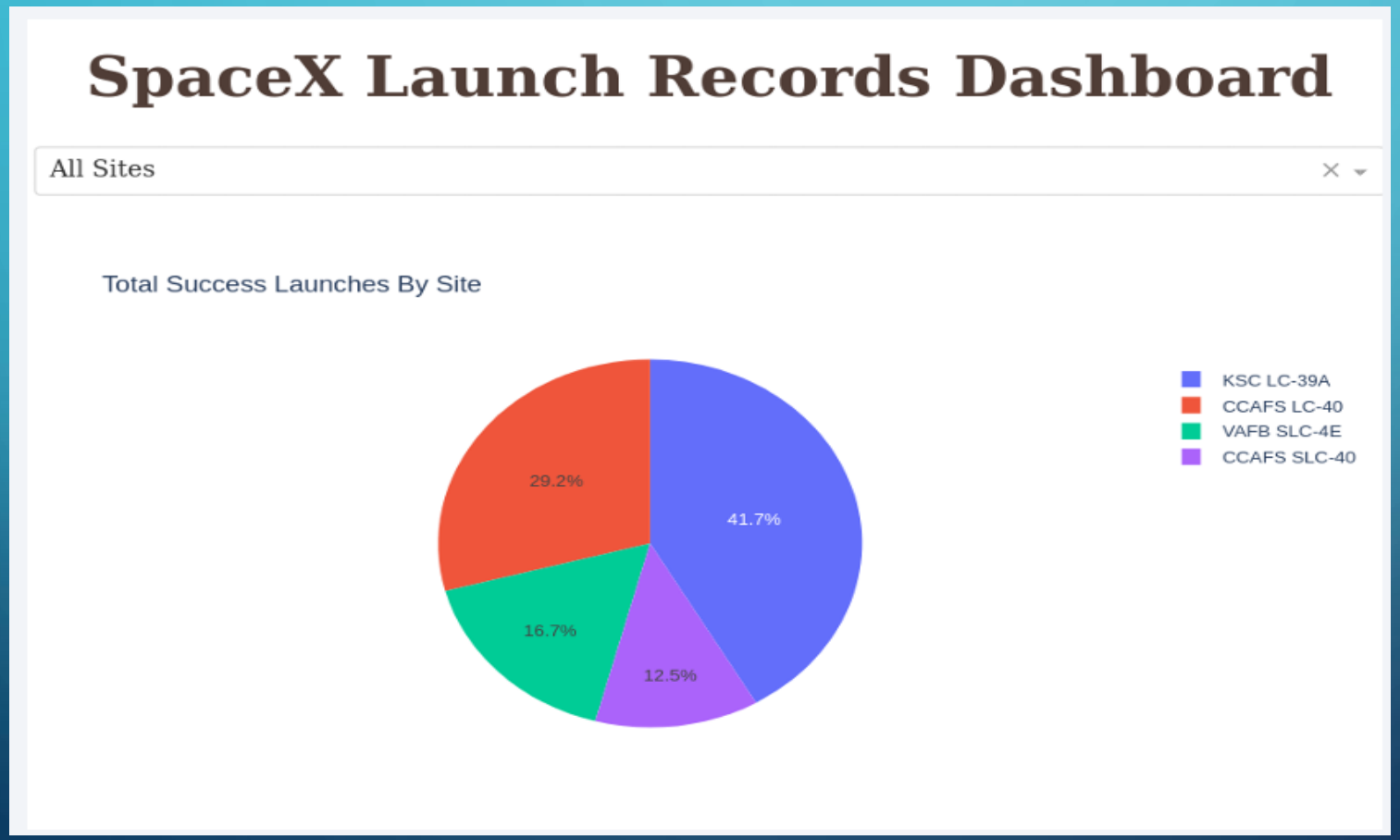


LOGISTICS AND SAFETY

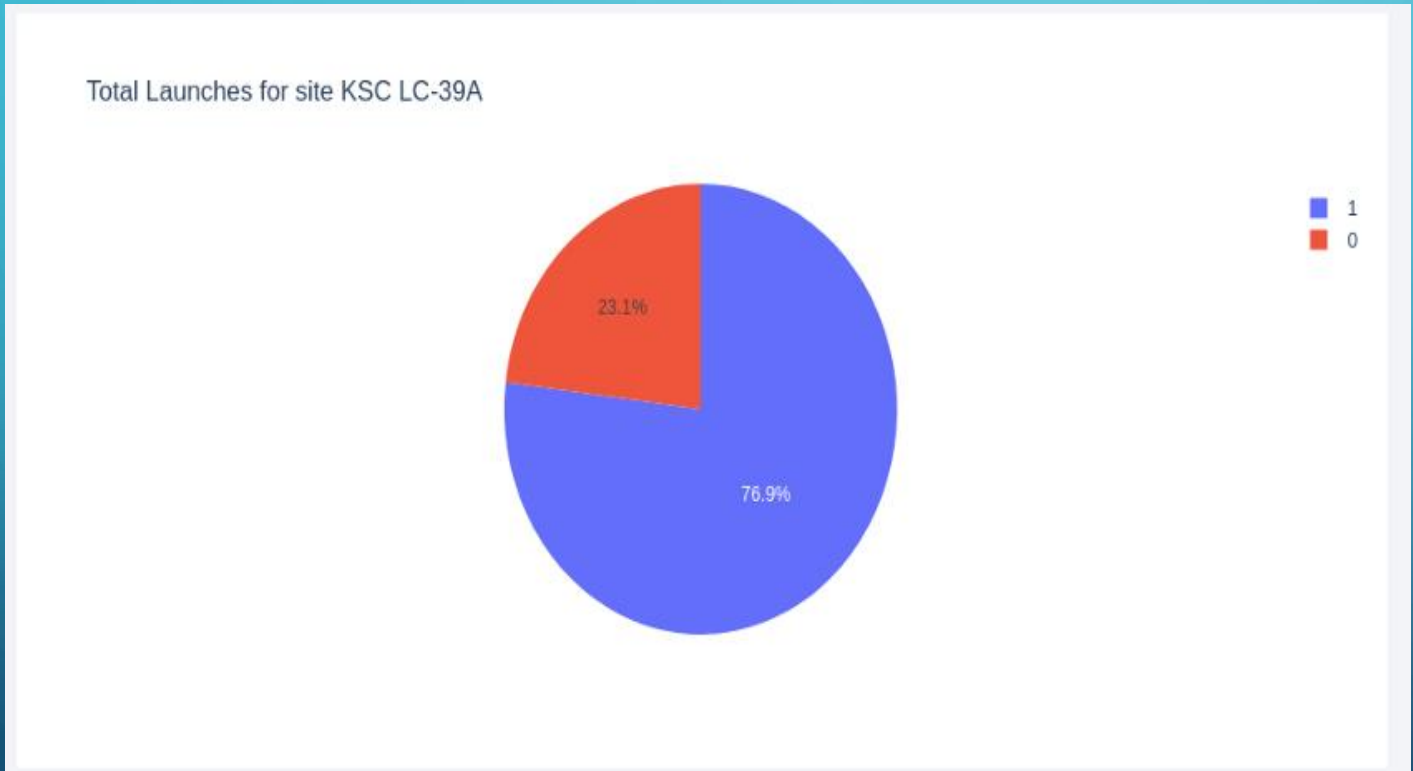


BUILD A DASHBOARD WITH PLOTLY DASH

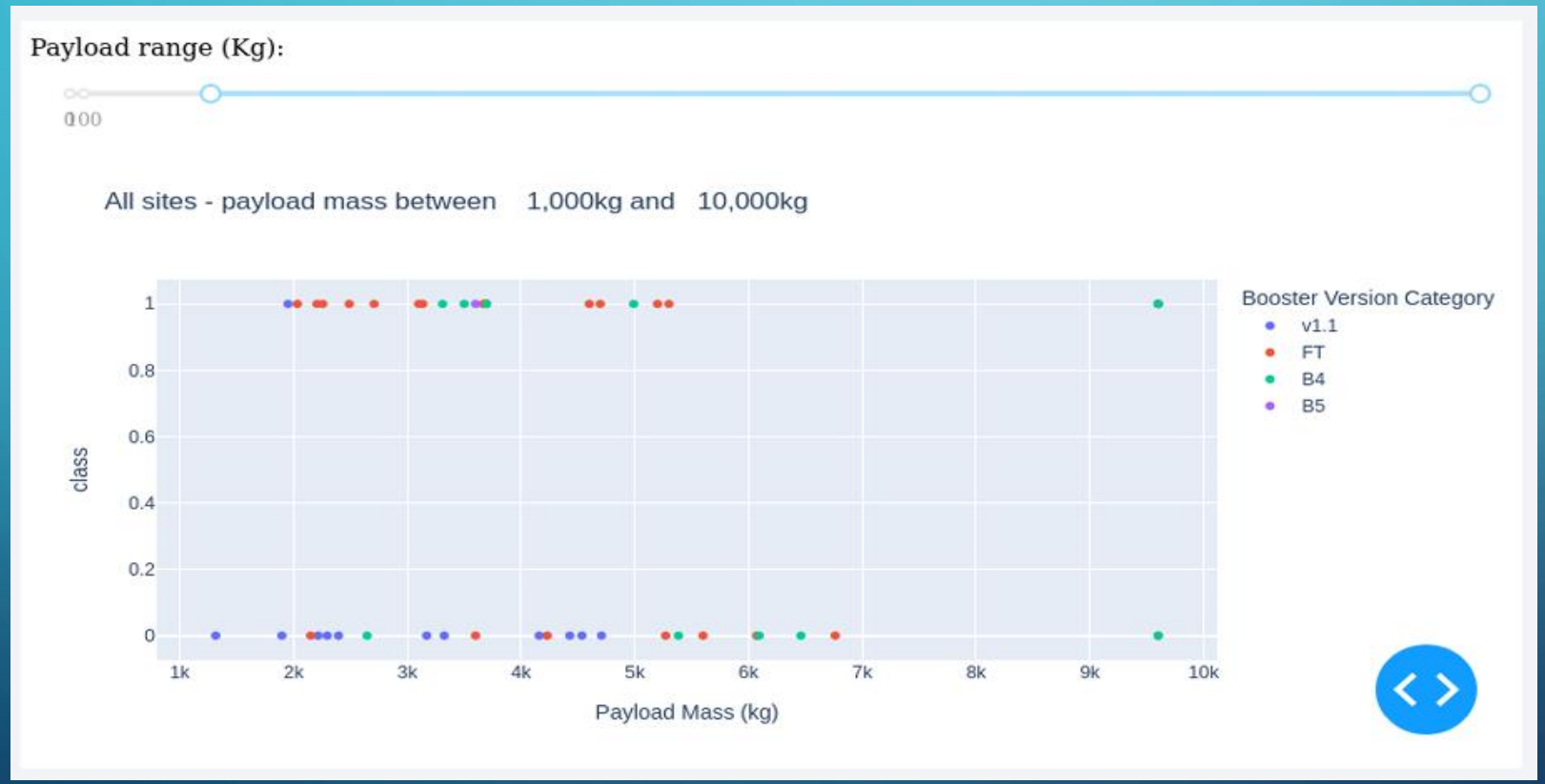
SUCCESSFUL LAUNCHES BY SITE



LAUNCH SUCCESS RATIO FOR KSC LC-39A



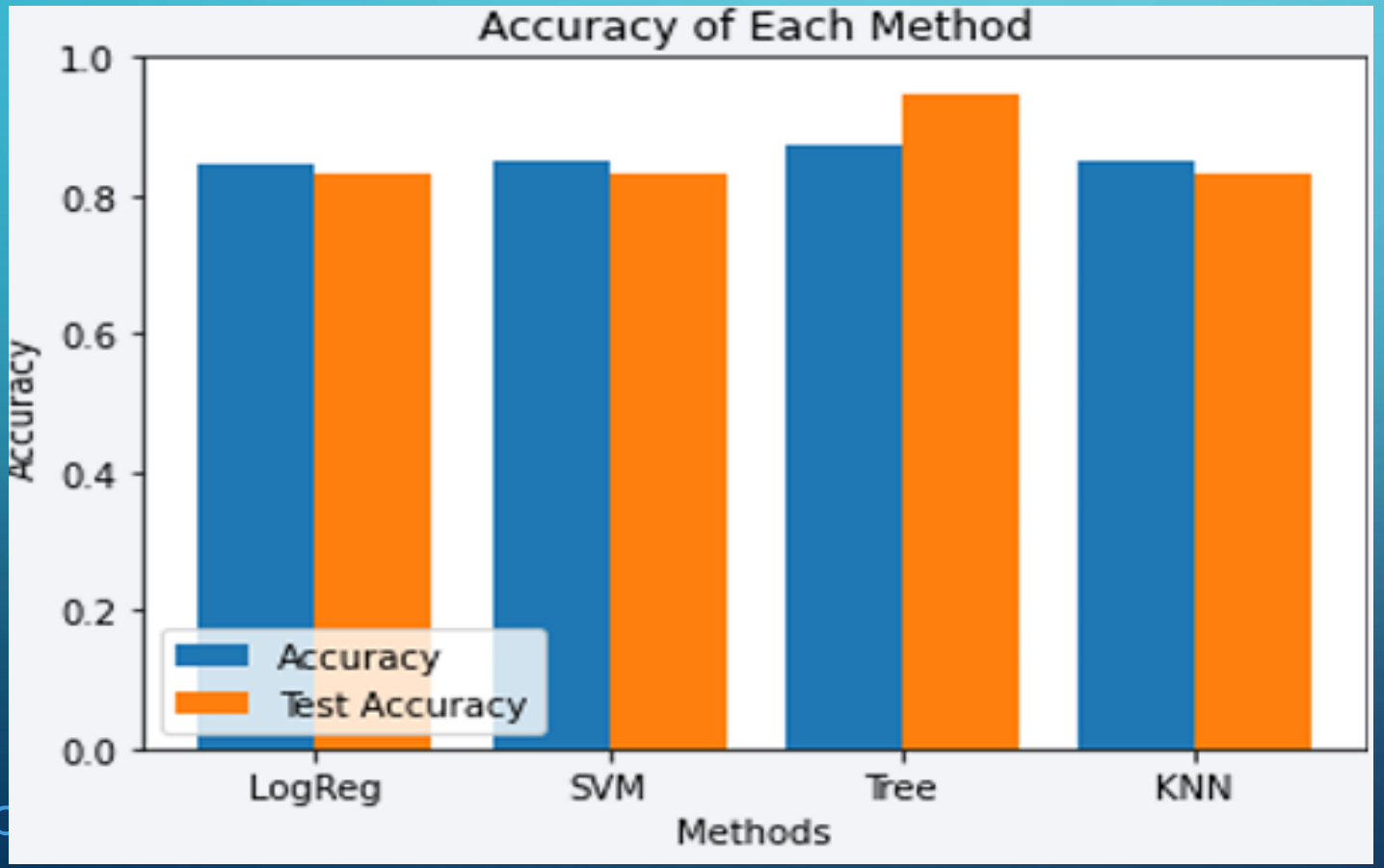
PAYLOAD VS. LAUNCH OUTCOME



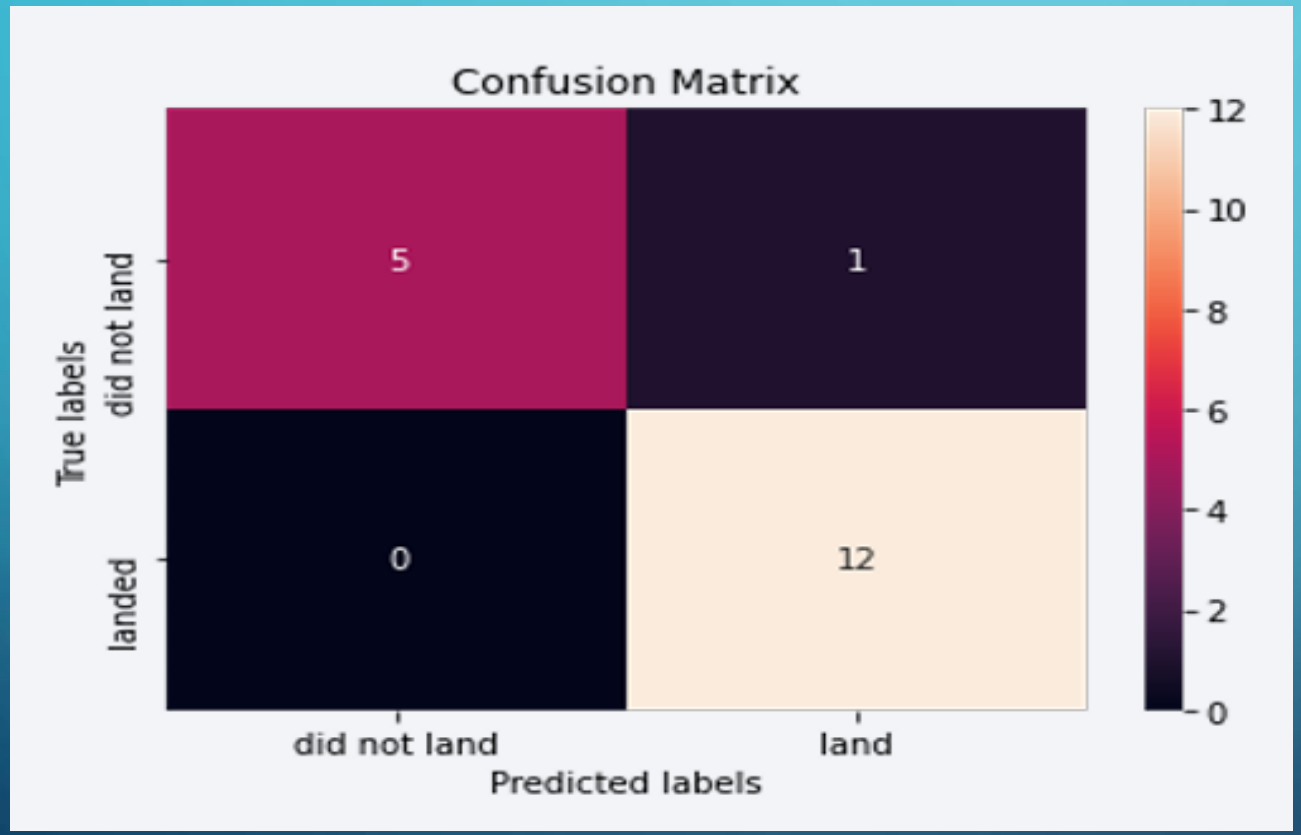
PAYLOAD VS. LAUNCH OUTCOME

PREDICTIVE ANALYSIS

CLASSIFICATION ACCURACY



CONFUSION MATRIX OF DECISION TREE CLASSIFIER



CONCLUSIONS

- Various data sources were scrutinized, leading to refined conclusions throughout the analysis.
- KSC LC-39A is identified as the optimal launch site.
- Launches with payloads exceeding 7,000kg entail lower risks.
- While a majority of mission outcomes are successful, the success of landing outcomes appears to improve over time, reflecting advancements in processes and rocket technologies.
- The utilization of a Decision Tree Classifier holds potential for predicting successful landings and subsequently enhancing profits.

APPENDIX

- To enhance model tests, it is crucial to assign a value to the np.random.seed variable.
- Due to Folium not displaying maps on GitHub, screenshots were taken as an alternative.