IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexander Brakas
November 27, 2023

Outline

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

Executive Summary

Summary of methodologies:

- Data Collected from web scraping SpaceX Wiki and SpaceX REST API
- Explore data after cleaning considering key factors like yearly trend and launch locations
- Analyzed dataset by calculating outcomes of the failed launch attempts and successful launch attempts.
- Create visualizations for comparing trends and what causes successful launches
- Build Models to predict success landings, multiple models will by tested to find the most accurate.

Summary of results:

- The 4 launch sites, ES-L1, GEO, HEO, and SSO, were 100% success rates for launches
- With more launches the greater rate of success the launches had
- Every model that was tested performed the same on the given test set
- Most launch sites are close to the equator to increase chance for successful launches

Introduction

Background

SpaceX is a leader in the space industry, revolutionizing space travel and the way we think about making rockets. Our main objective is to investigate why SpaceX can launch rockets for so much cheaper and why is the success rate of cheaper rocket launches. The key aspect that would allow SpaceY to complete with SpaceX is to find a way to reuse the first stage of the rocket at a more consistent rate to allow for cheaper rocket launches.

Investigate

- What affects the first-stage landing success rate
- What predictive model with most accurately predict if the landing will be a success

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Web scrapping SpaceX wiki page and SpaceX REST API
- Perform data wrangling
  - Filling in missing values and creating categorical data values with one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Cross validation with grid search to get optimal score for each model

Data Collection – SpaceX API

- Request rocket launch data from API
- Convert data to a dataframe
- Filter the data so we only have Falcon 9 launch data
- Replace missing values
- Export data

Data Collection - Scraping

- Gathered HTML page from Wikipedia
- Created Beautiful soup object
- Parsed HTML for tables
- Pulled data from the tables
- Loaded data into a dataframe
- Exported the data

Data Wrangling

- Base level exploration of data
- Count number of launches at each site and in each orbit
- Explore landing outcome labels
- Create classes for key outcomes
- Find the average rate of success

EDA with Data Visualization

- Created charts comparing the following:
  - Payload Mass and Launch Site
  - Flight Number and Launch site
  - Payload Mass and Orbit
  - Flight Number and Payload
- The charts changed to appropriately represent the trends and similarities for each comparison

EDA with SQL

- Named the unique launch sites
- Calculated the total payload mass launched by NASA
- Get the first date of a successful recovery
- Get total successes and failures
- Tally landing outcomes between 2010 July 4 till 2017 March 20
- Find booster that carried the biggest payload

Build an Interactive Map with Folium

- Added markers to each launch site
- Added tags with markers to indicate the name of the launch site
- Created a cluster for successful launches and marked them on the map with green
- Added lines to show between distance between launch sites

Build a Dashboard with Plotly Dash

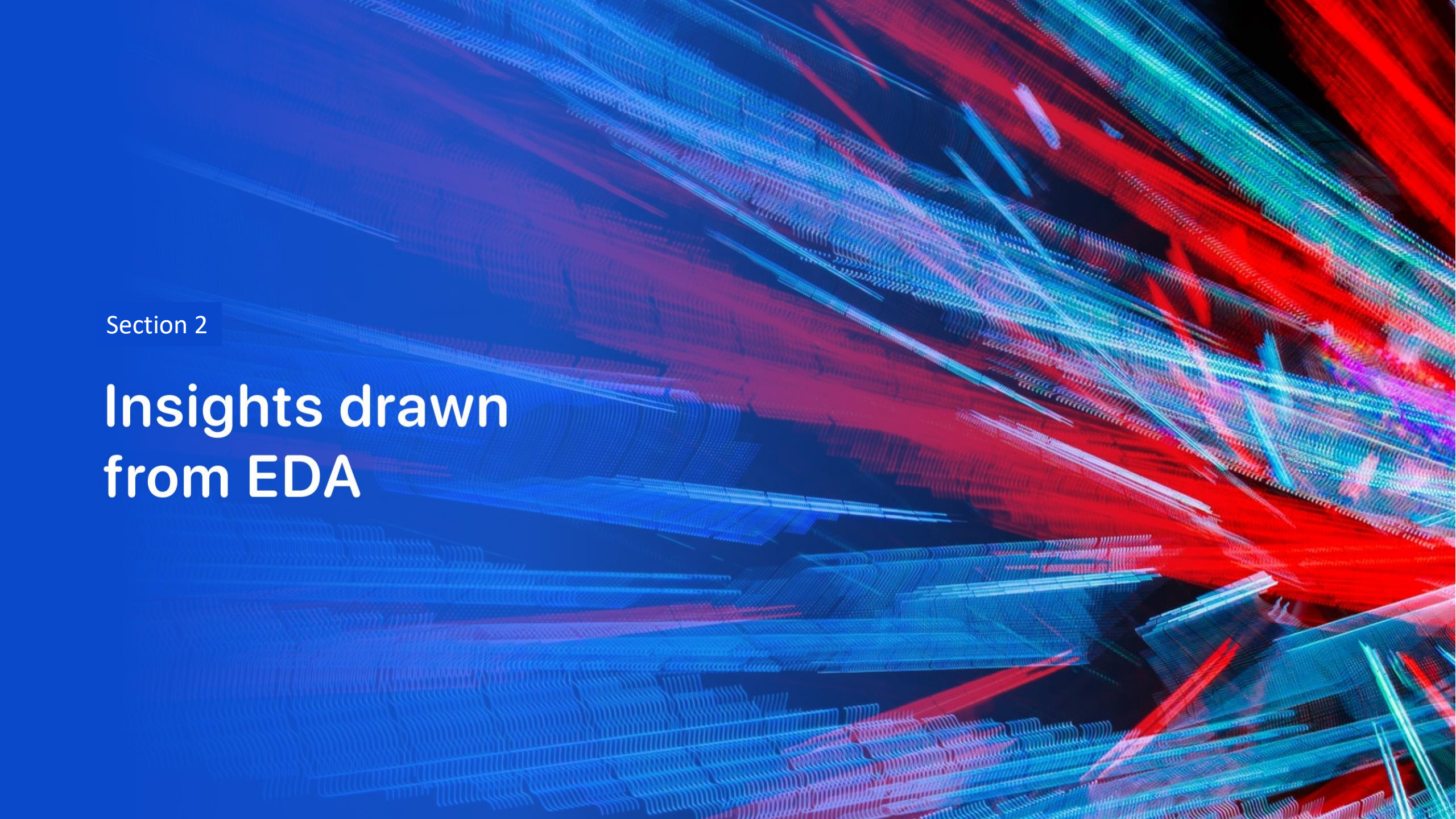
- Create a dropdown to allow user to pick the launch site
- Slider to limit payload mass to given range
- Create scatter plots to visualize to see payload mass compared to
- Pie chart to show the launch successes

Predictive Analysis (Classification)

- Standardize the data
- Split data into training and test sets
- Created decision tree, support vector machine, K-nearest neighbors, and logistic regression models using gridsearch with 10 cross validations
- Determine accuracy of each model
- Represent the results as a confusion matrix

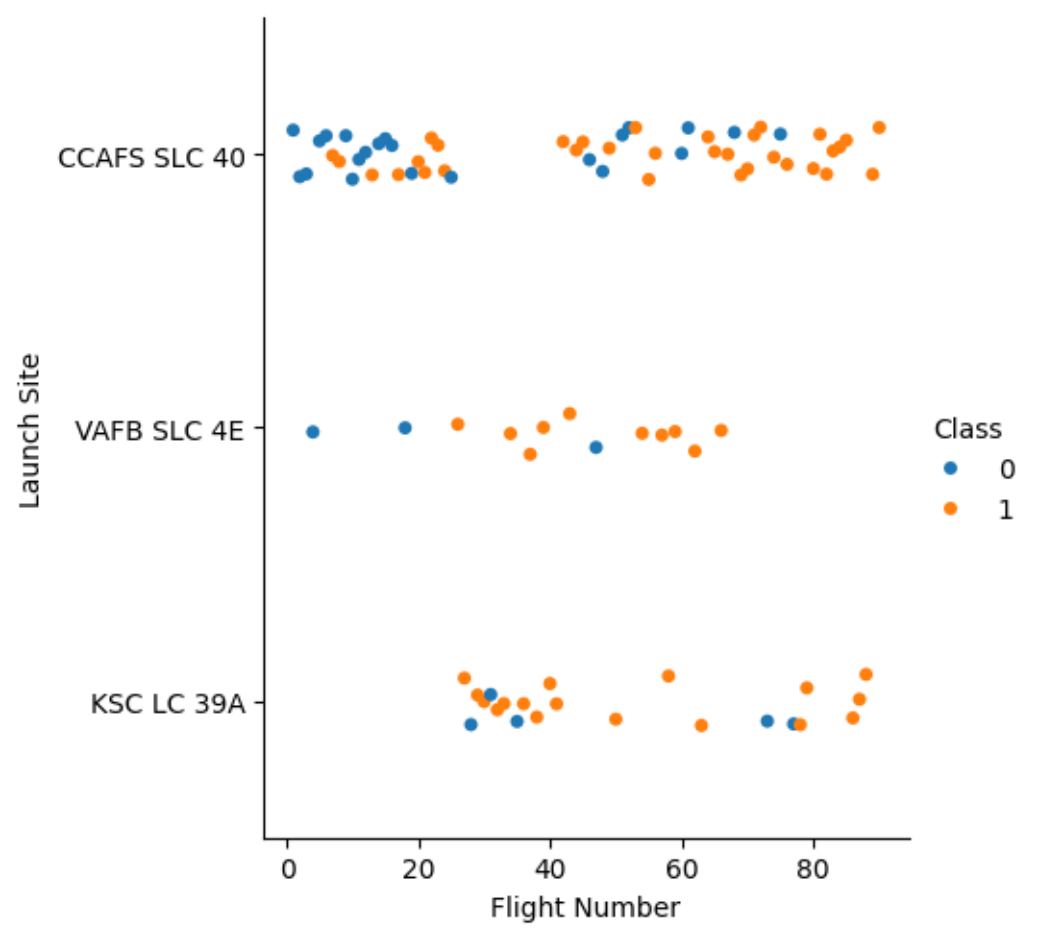
Results

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


Section 2
Insights drawn from EDA

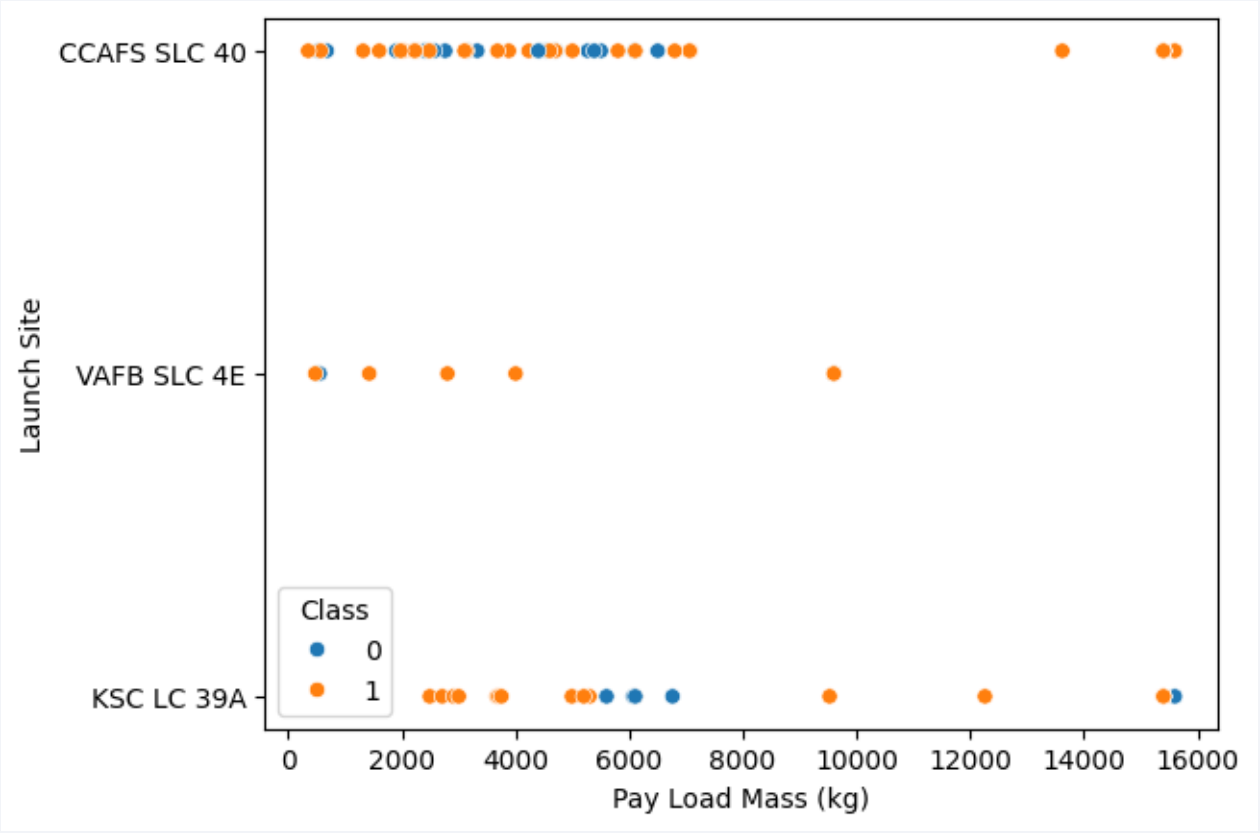
Flight Number vs. Launch Site

- CCAFS SLC 40 had almost all the first 20 flights
- Lots of fails (0 class) in the lower flight number range
- VAFB SLC 4E least launches



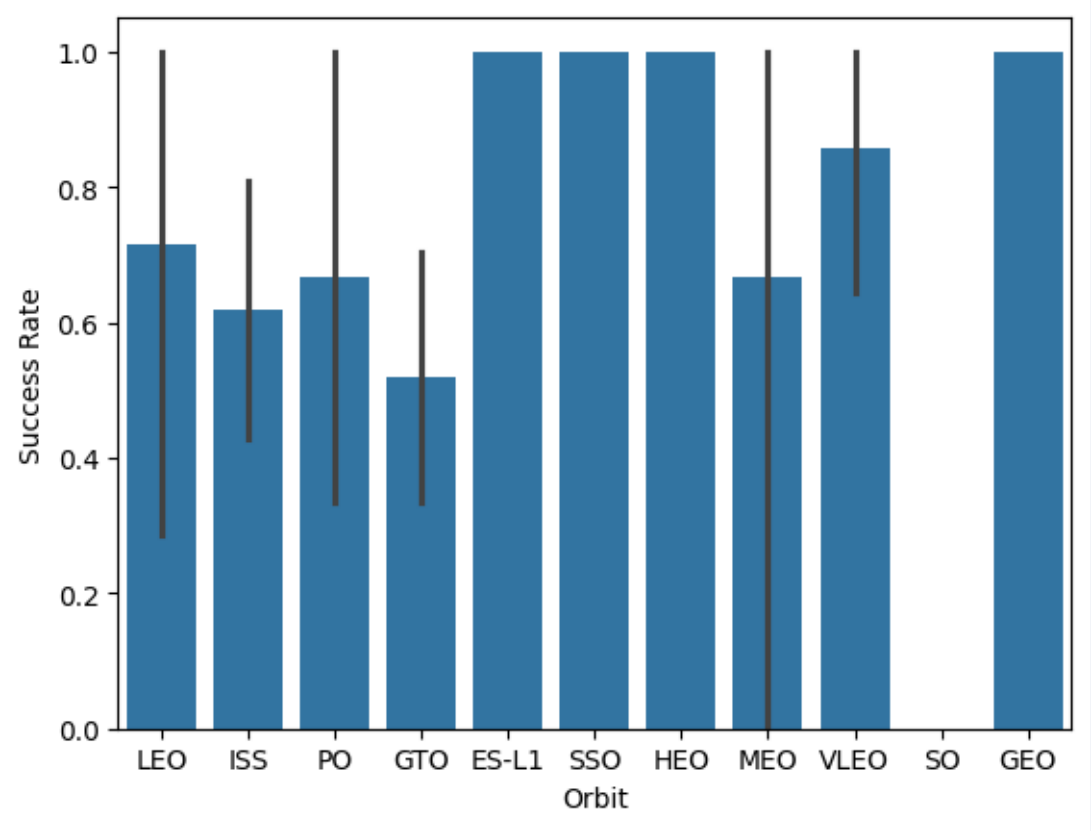
Payload vs. Launch Site

- As the payload mass increases, the number of failures trends up slightly (class 0)
- Seems to have a bubble of class 0 around 6000 kg



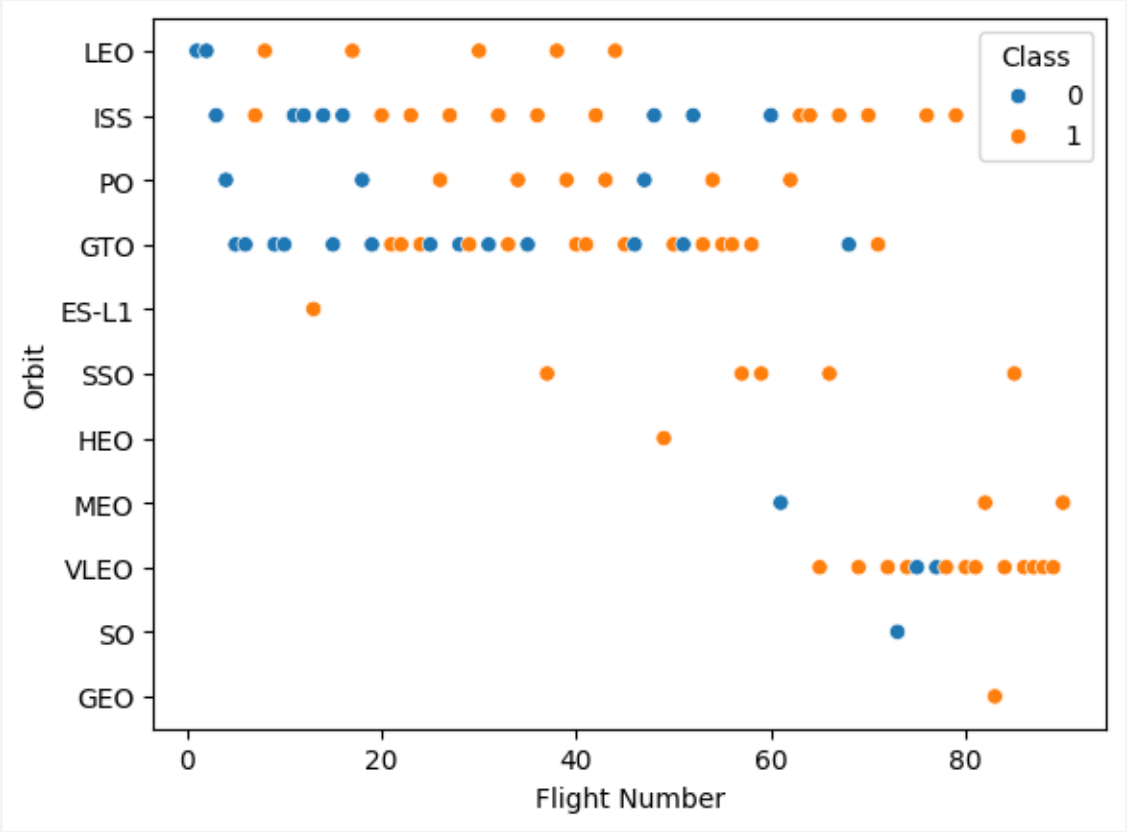
Success Rate vs. Orbit Type

- Four orbits with 100% success rate: ES-L1, SSO, HEO, and GEO
- SO orbit has 100% failure rate
- Rest of orbits have a lot of deviation on successes



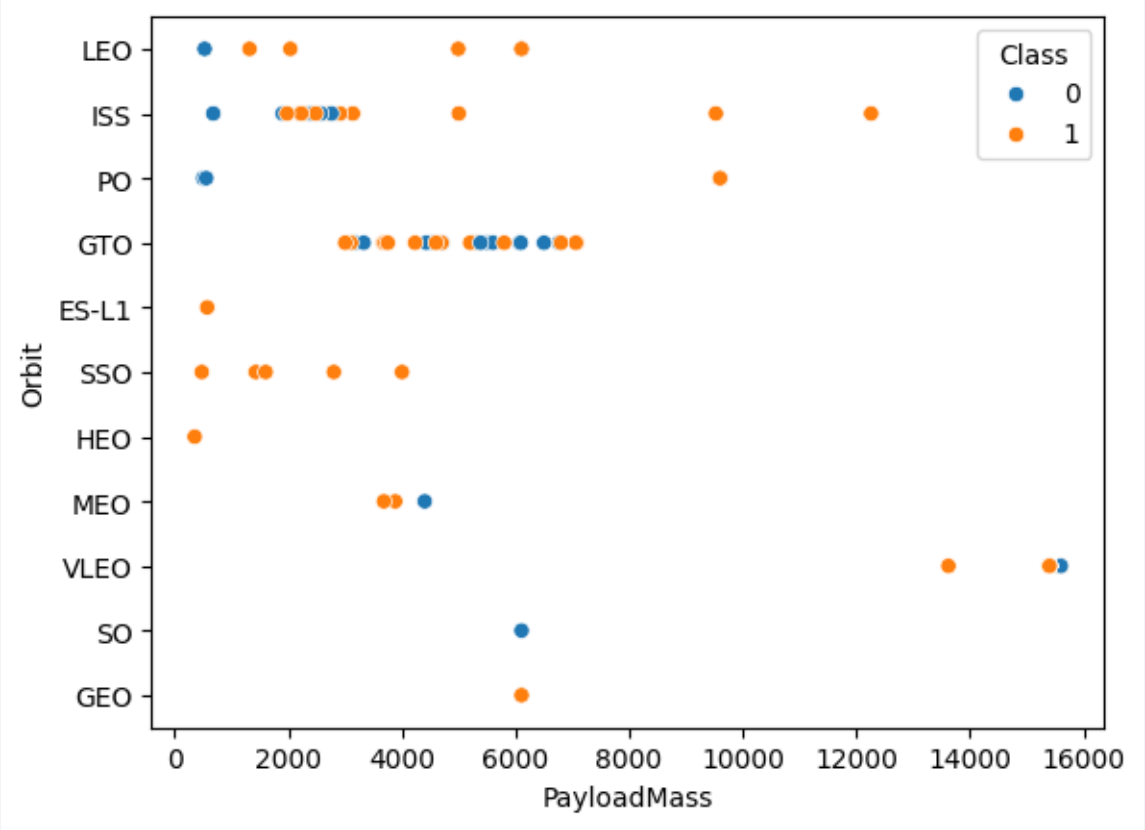
Flight Number vs. Orbit Type

- Important note, first flight numbers used LEO, ISS, PO, and GTO mainly
- Shows discrepancy in orbit and success rate
- SO and GEO only had 1 launch



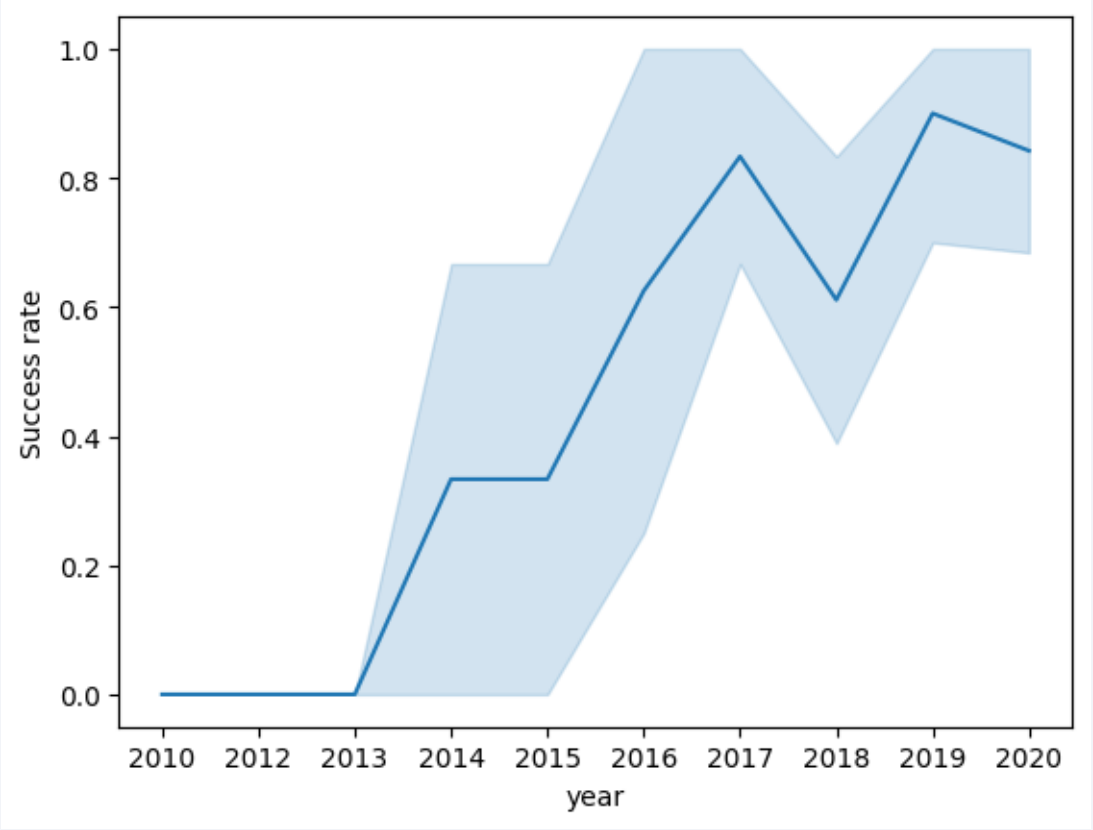
Payload vs. Orbit Type

- LEO, ISS, PO have better success rates with greater payload mass
- ES-L1, SSO, HEO have only had lower payload mass than then bubble of fails (<6000)



Launch Success Yearly Trend

- Success rate went up over the years
- Lowest deviation was raised over time, so the chance of a failure significantly went down

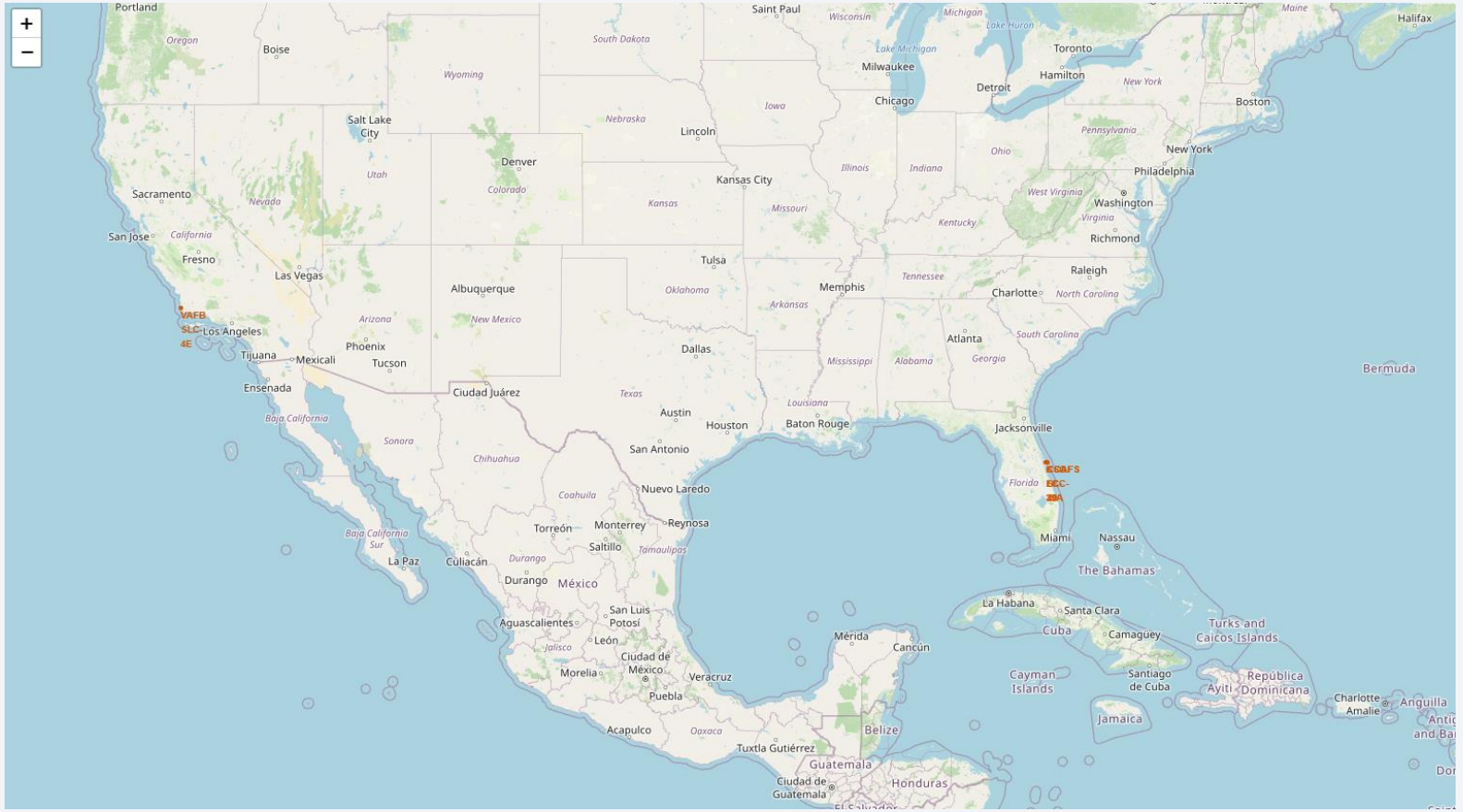


Section 3

Launch Sites Proximities Analysis

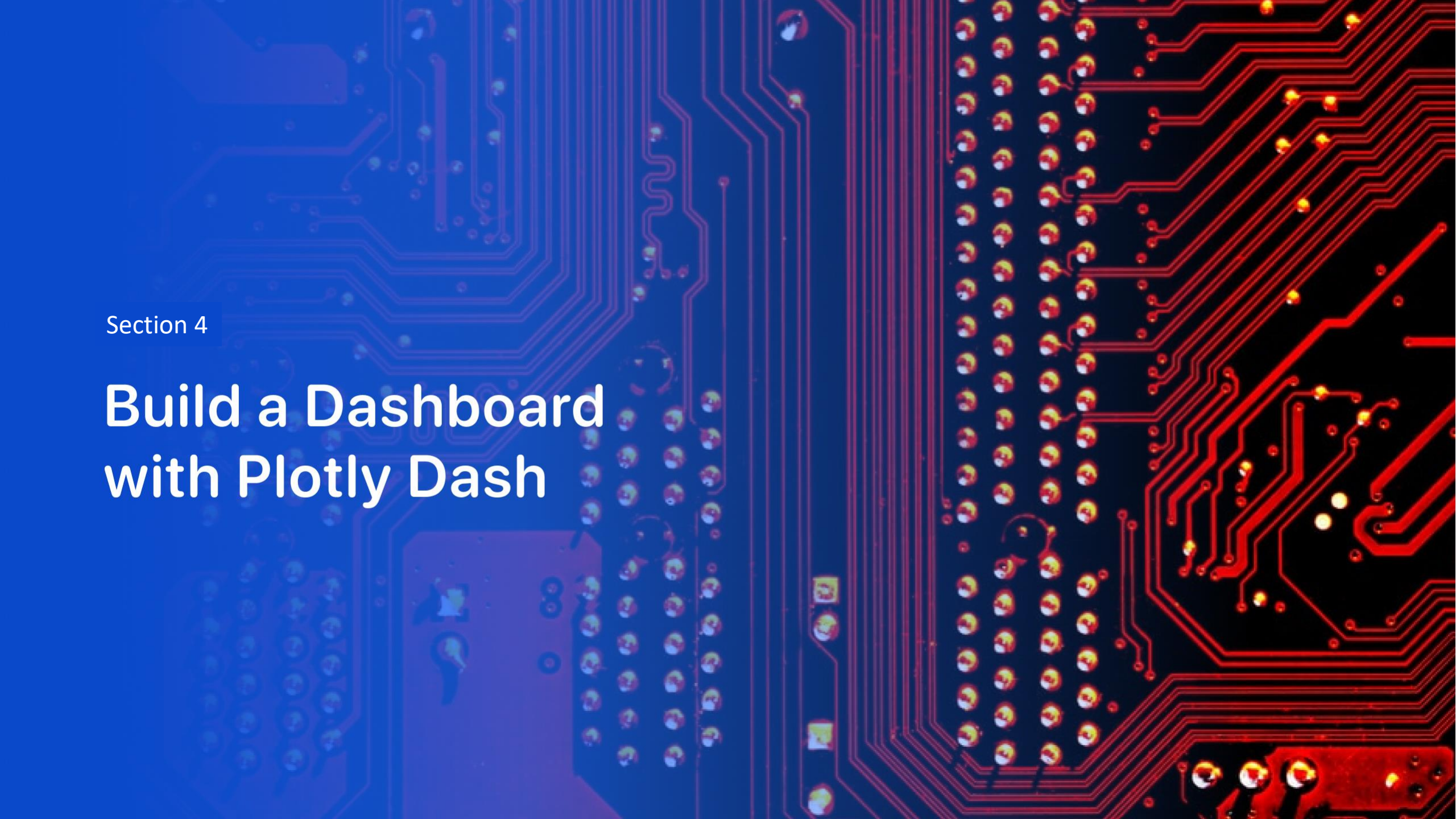
All launch sites

- Launch sites are only located close to the coast
- Launch sites close to equator



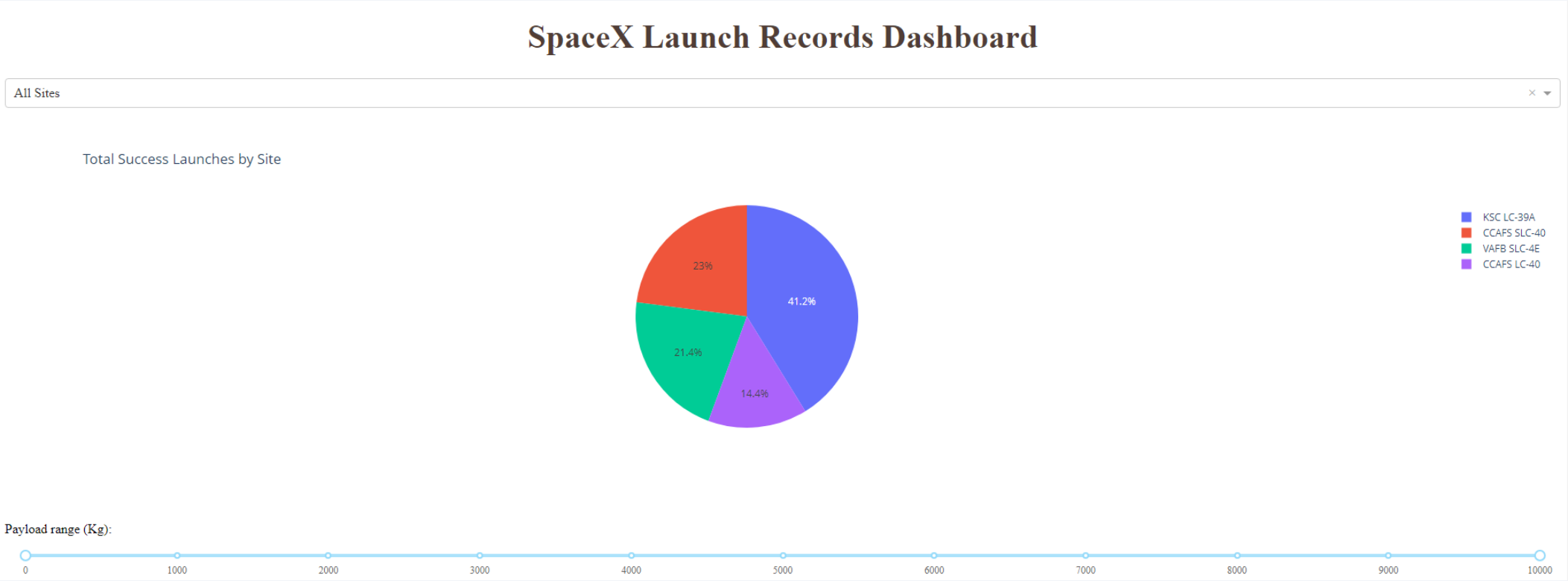
Successful launches per site

- More success towards East coast compared to West coast


Section 4
Build a Dashboard with Plotly Dash

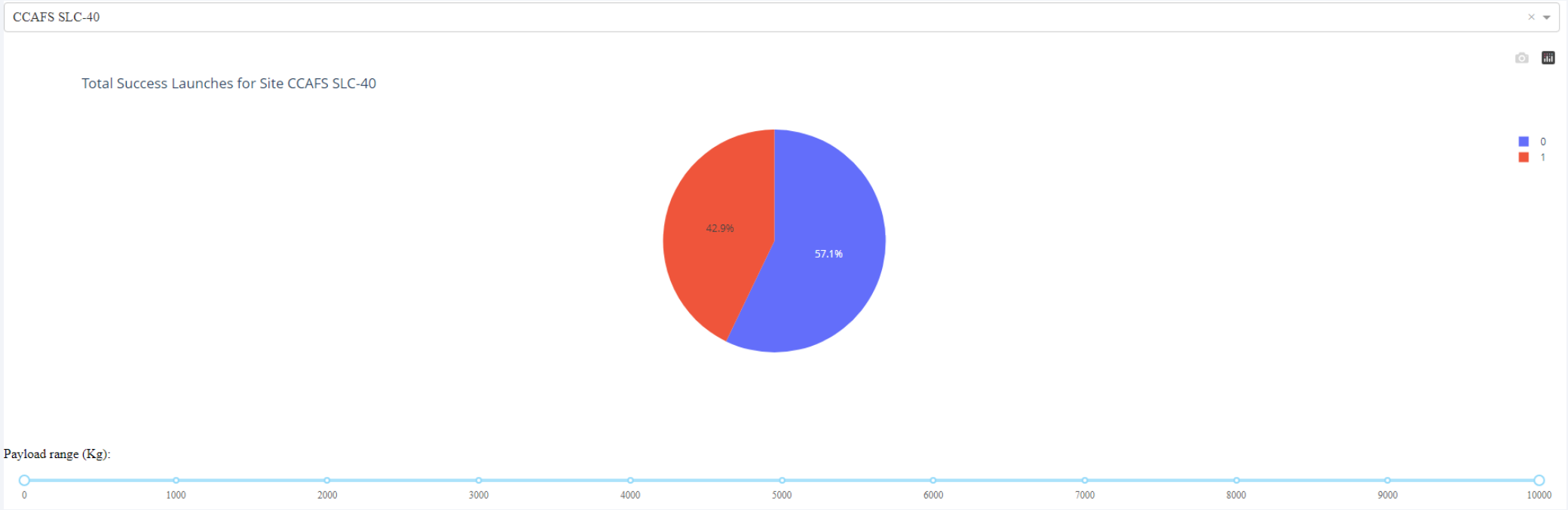
Dashboard Setup

- KSC LC-29A had the most total successful launches across all sites with all payload as the range



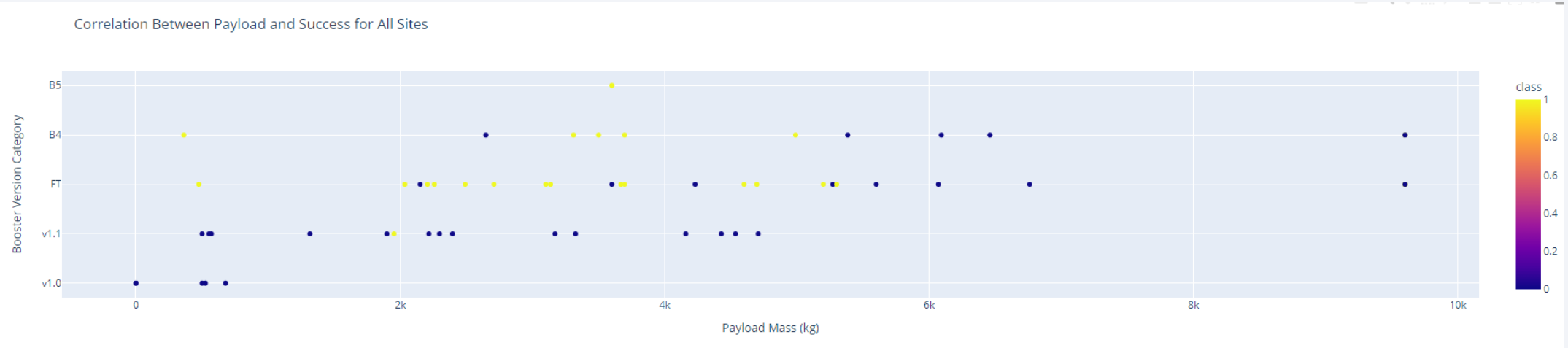
Highest launch success ratio

- CCAFS SLC-40 had the best success to fail ratio with 42.9% class 1 outcome
- Even though KSC LC-29A had the most successful launches they had more class 0 as well



Outcome scatter plot

- Later boosters had heavier payloads
- Early boosters had an extremely high failure rate
- All boosters with payloads above 6,000 kg failed

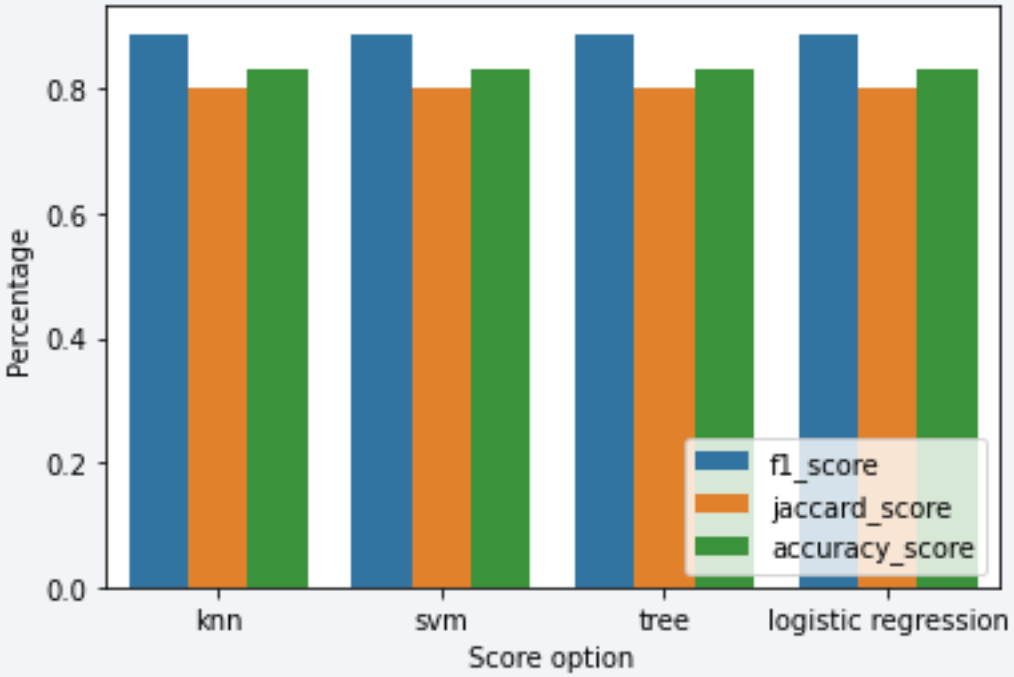


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Every model had the same score for all categories
- f1 score is the highest score
- Jaccard score is the lowest score



Confusion Matrix

- Matrix contains 3 type I errors (false positives)
- 12 true positives and 3 true negatives
- Recall: 12/(12+0) = 100%
- Precision: 12/(12+3) = 80%
- F1 score: 2*(1*0.8)/(1+0.8) = 88.9%
- Accuracy: (12+3)/(12+3+3+0) = 83.3%
- Jaccard: Intersect/Union = 12/15 = 80%

*NOTE: Jaccard score is with binary case so only class of 1 is considered for intersect

Conclusions

- All models performed the same
- Most launch sites were on the coast and as close to the equator as possible to minimize launch distance
- The launches became more successful over time
- Payload mass has a negative impact on the success rate

Appendix

- Original dataframe for score graph

	knn	svm	tree	logistic regression
f1_score	0.888889	0.888889	0.888889	0.888889
jaccard_score	0.800000	0.800000	0.800000	0.800000
accuracy_score	0.833333	0.833333	0.833333	0.833333

Thank you!