IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jonathan Spencer
June 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Collect data using SpaceX REST API and web scraping techniques
  - Wrangle data to create success/fail outcome variable
  - Explore data with data visualisation techniques, considering the following factors: payload, launch site, flight number and yearly trend
  - Analyse the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
  - Explore launch site success rates and proximity to geographical markers
  - Visualise the launch sites with the most success and successful payload ranges
  - Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K - nearest neighbor (KNN)

- Summary of all results

  Exploratory Data Analysis:
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES -L1, GEO, HEO, and SSO have a 100% success rate

  Visualisation/Analytics:
  - Most launch sites are near the equator, and all are close to the coast

  Predictive Analytics:
  - All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station (ISS), launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Steps

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data – by filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Explore data via EDA with SQL and data visualisation techniques
- Visualise the data using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

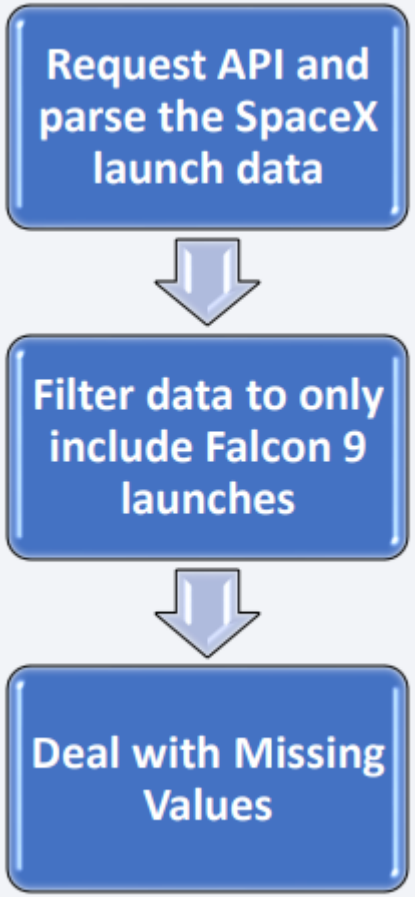
Data Collection

Steps

- Request data from SpaceX API (rocket launch data)
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with calculated .mean()
- Export data to csv file

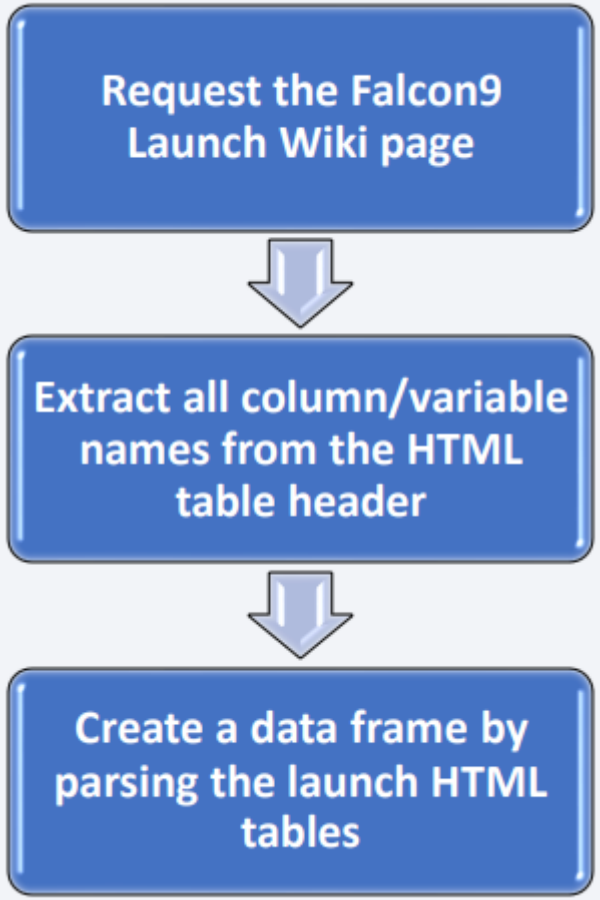
Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and perform some basic data wrangling.
- https://github.com/JonathanS-cmd/Jonathan/blob/main/01_SpaceX_Data_Collection_API.ipynb



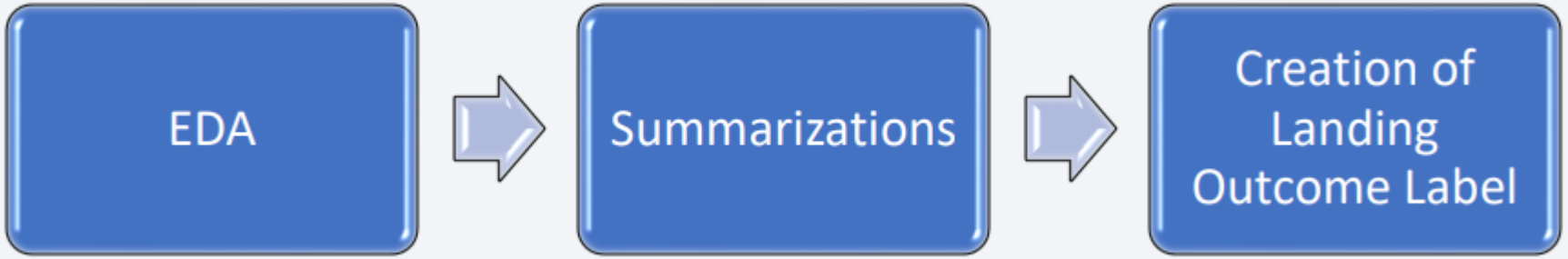
Data Collection - Scraping

- We applied web scrapping to webscrape Falcon 9 launch records with BeautifulSoup
- https://github.com/JonathanS-cmd/Jonathan/blob/main/02_SpaceX_Web_Scraping.ipynb



Data Wrangling

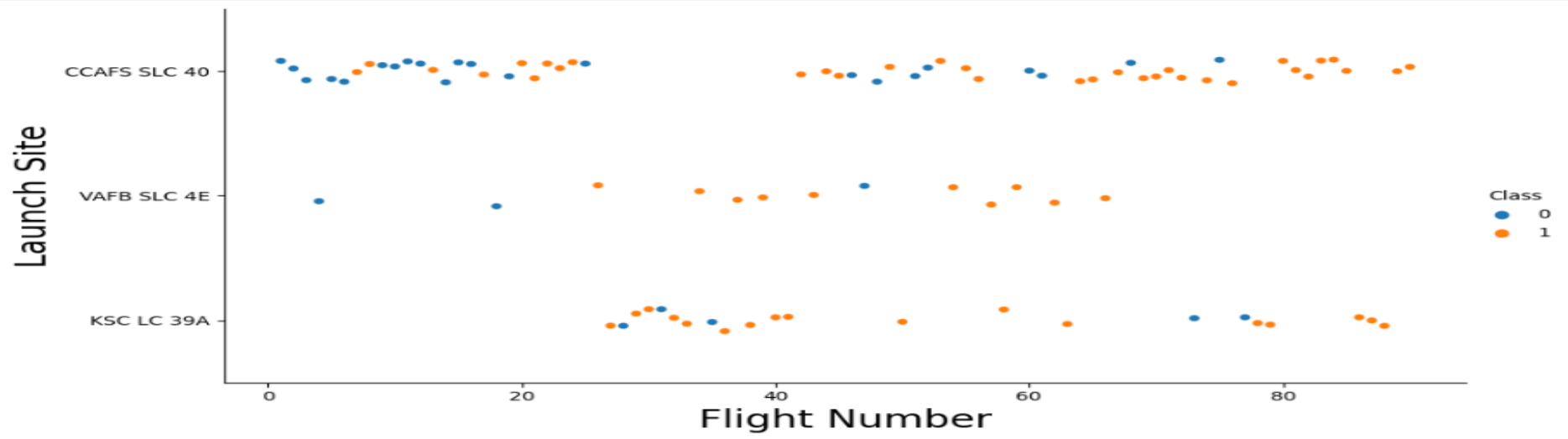
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset.
- Then the summarised launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally, the landing outcome label was created from Outcome column.



- https://github.com/JonathanS-cmd/Jonathan/blob/main/03_SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barplots were used to visualise 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



- https://github.com/JonathanS-cmd/Jonathan/blob/main/05_SpaceX_EDA_Data_Visualization.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the Jupyter notebook.
- We applied EDA with SQL to get insight from the data like first successful landing date and Total Payload Mass.
- https://github.com/JonathanS-cmd/Jonathan/blob/main/04_SpaceX_EDA_SQL.ipynb

Build an Interactive Map with Folium

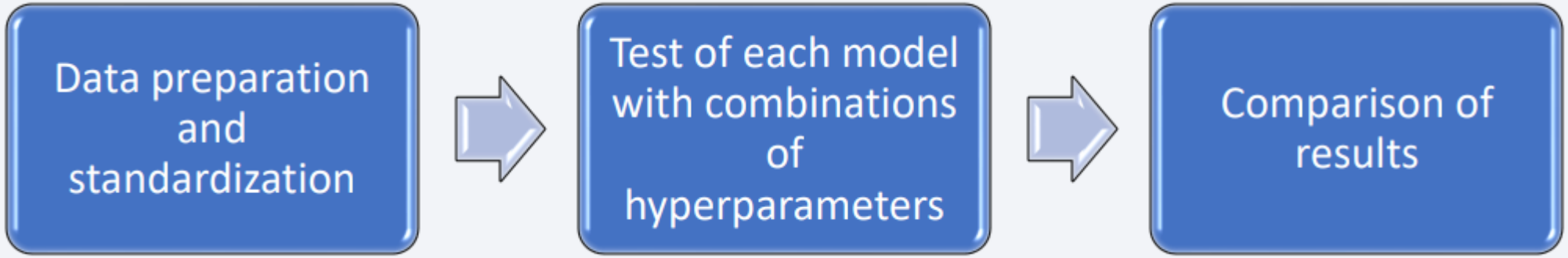
- We utilised a folium map to visualise launch sites and their corresponding success or failure outcomes. By assigning a value of 0 for failure and 1 for success, we marked the launch sites with markers, circles, and lines on the map to indicate their respective outcomes. We employed color-labeled marker clusters to identify launch sites with higher success rates. Additionally, we measured the distances between each launch site and its surrounding areas, exploring factors such as proximity to railways, highways, coastlines and cities
- https://github.com/JonathanS-cmd/Jonathan/blob/main/06_SpaceX_Interactive_Visual_Analytics_Folium.ipynb

Build a Dashboard with Plotly Dash

- We developed an interactive dashboard using Plotly Dash. Within the dashboard, we included pie charts that visualise the total number of launches from specific sites. Additionally, we created scatter graphs to examine the relationship between the launch outcome and the payload mass (in kilograms) for various booster versions.
- https://github.com/JonathanS-cmd/Jonathan/blob/main/07_SpaceX_Interactive_Visual_Analytics_Plotly.py

Predictive Analysis (Classification)

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



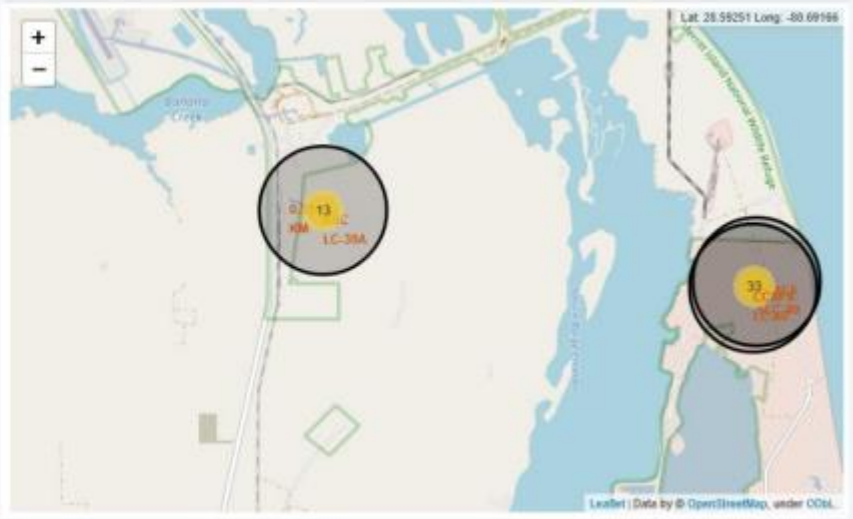
- https://github.com/JonathanS-cmd/Jonathan/blob/main/09_SpaceX_Machine_Learning_Prediction.ipynb

Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites
  - The first launches were done to Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2,928 kg
  - The first success landing outcome happened in 2015 five years after the first launch
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average
  - Almost 100% of mission outcomes were successful
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - The number of landing outcomes became as better as years passed

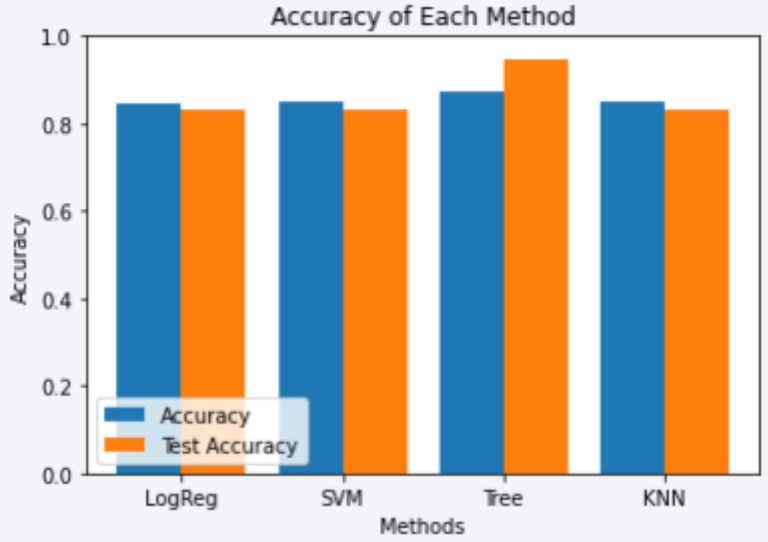
Results

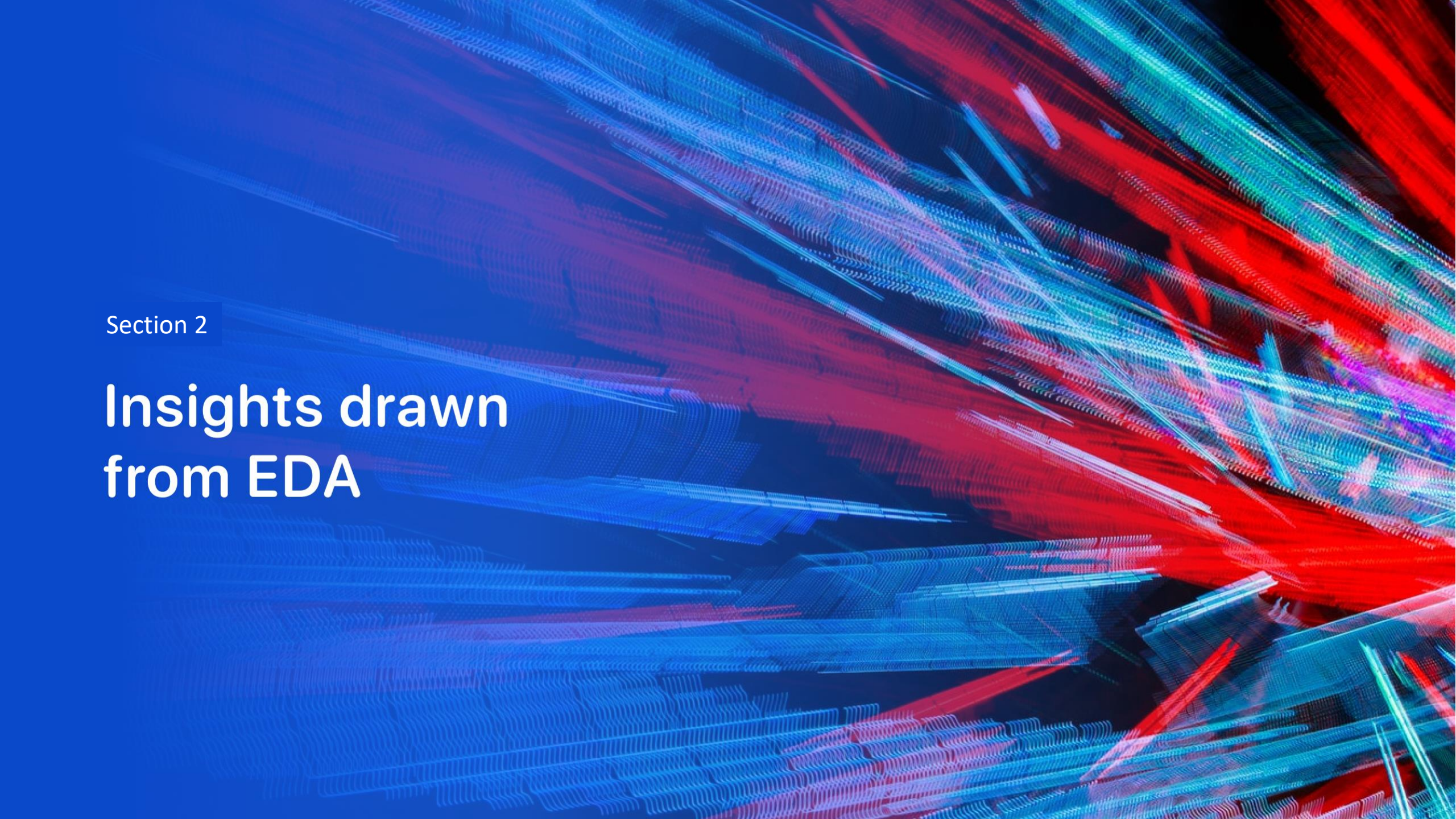
- Using interactive analytics, it was possible to identify launch sites are in safe places, for example, near the ocean with good surrounding infrastructure nearby
- Most launches happen at east cost USA sites



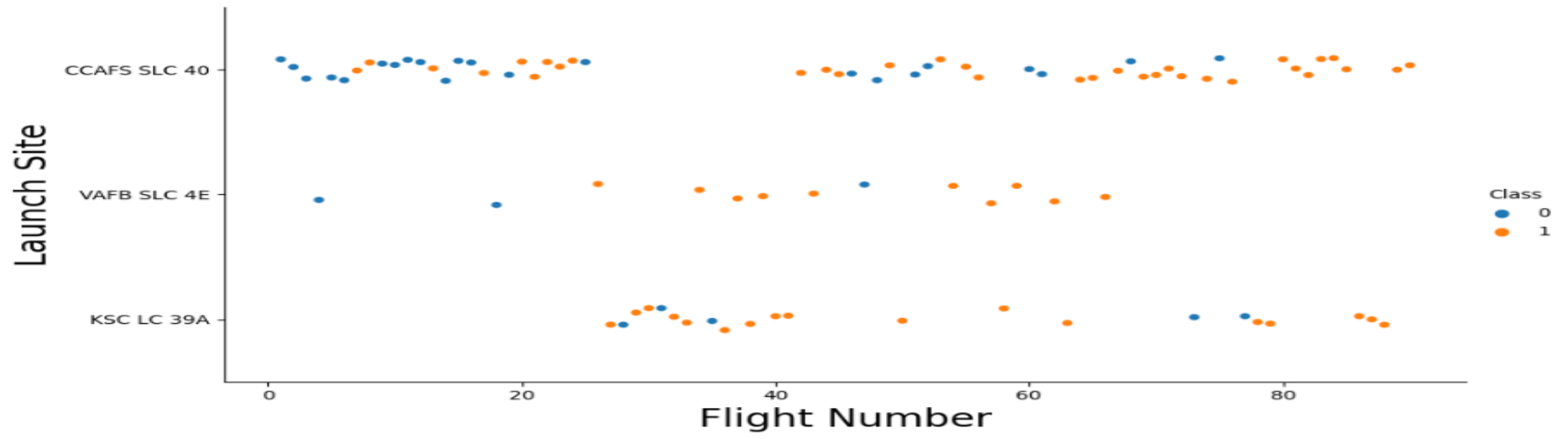
Results

- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having an accuracy of over 87% and accuracy for test data greater than 94%.




Section 2
Insights drawn from EDA

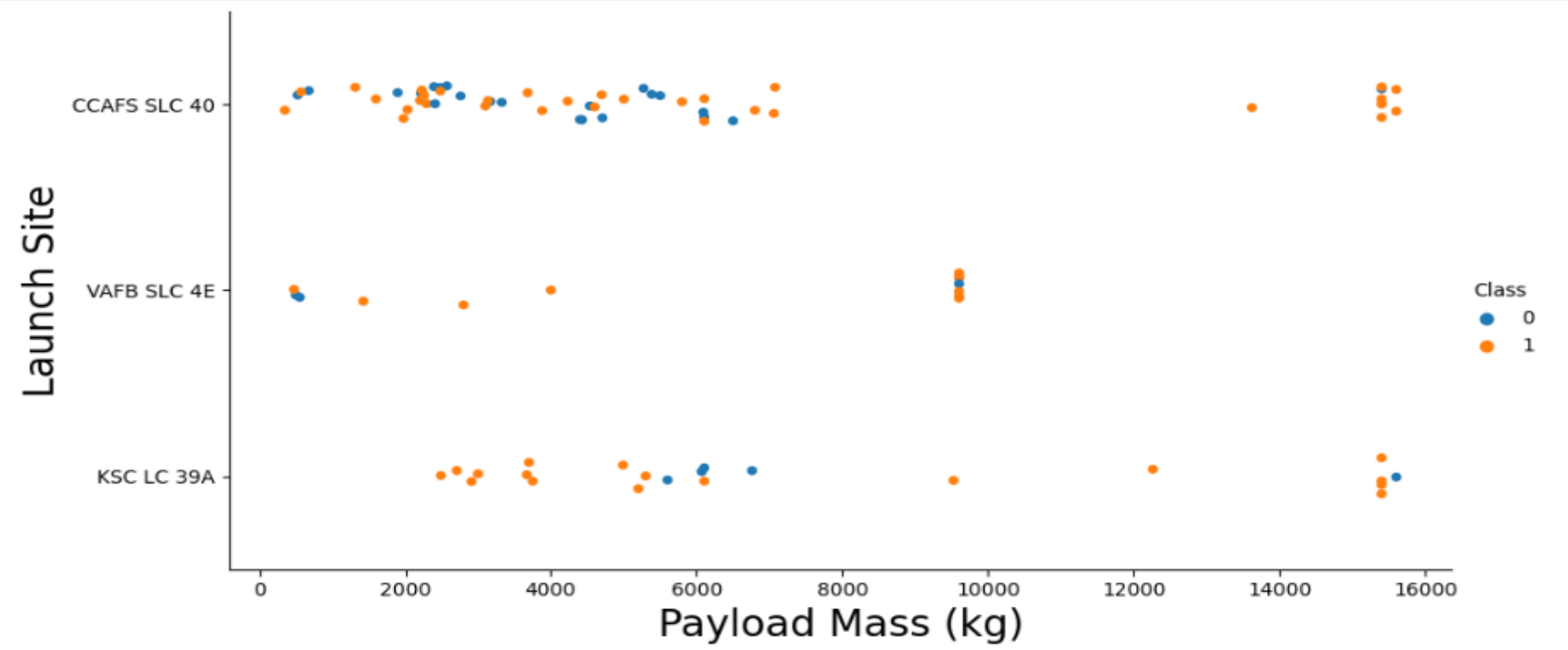
Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 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.

Payload vs. Launch Site

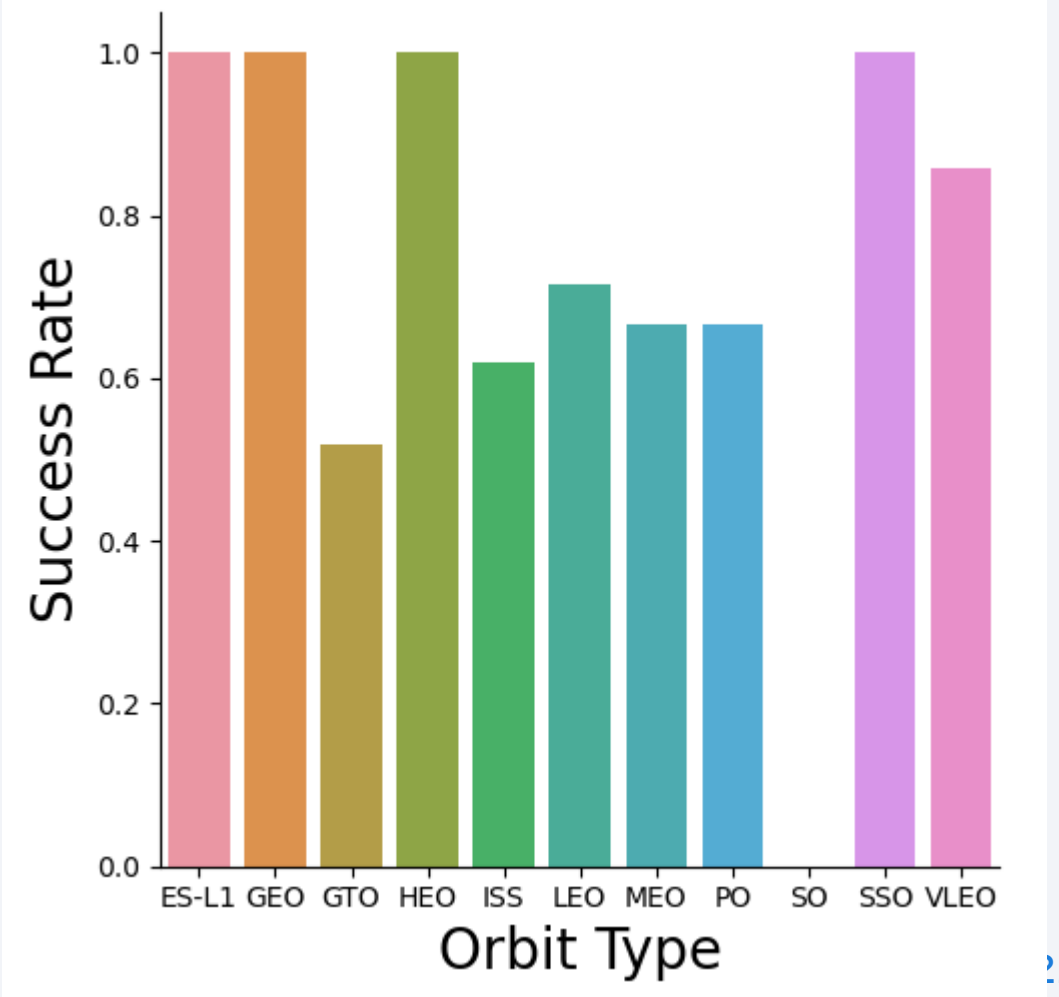
- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



Success Rate vs. Orbit Type

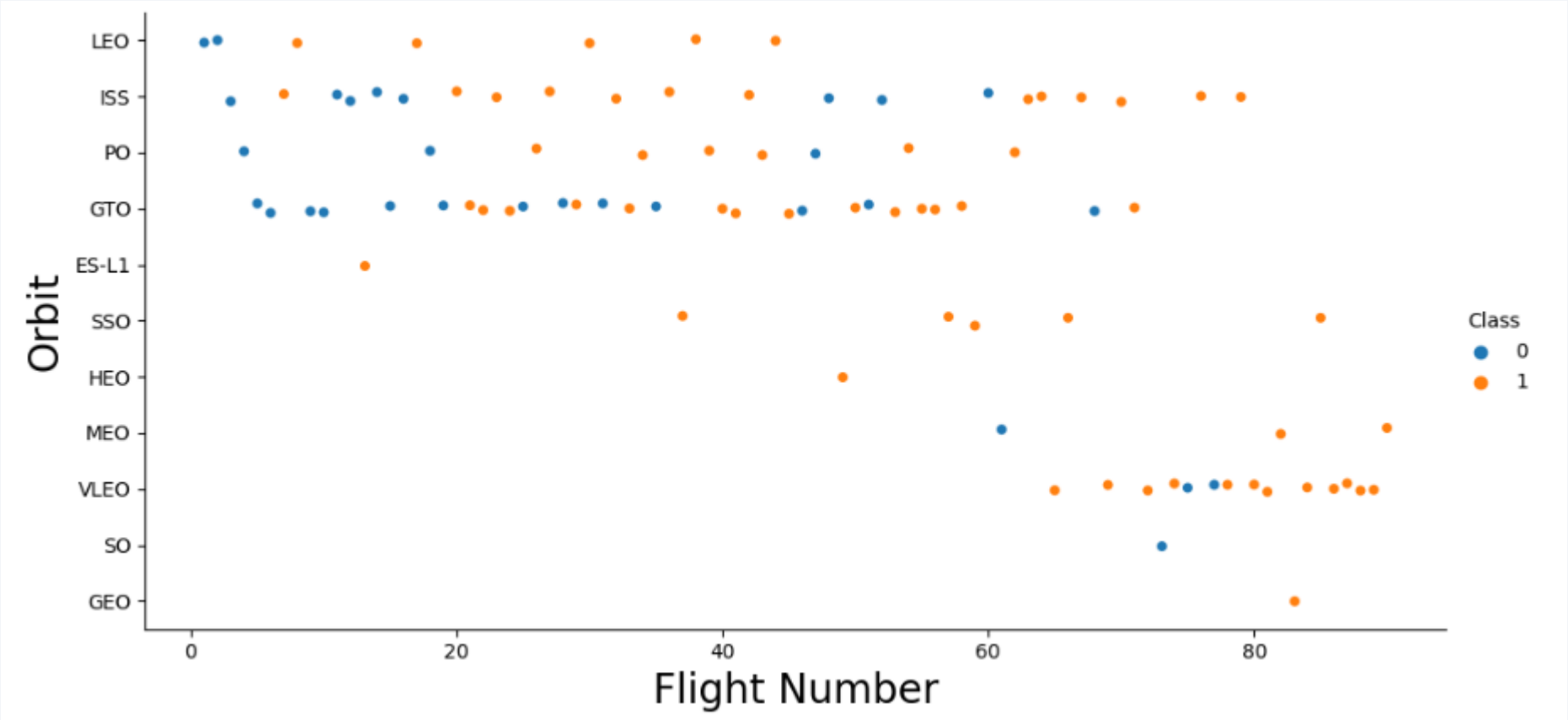
The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO



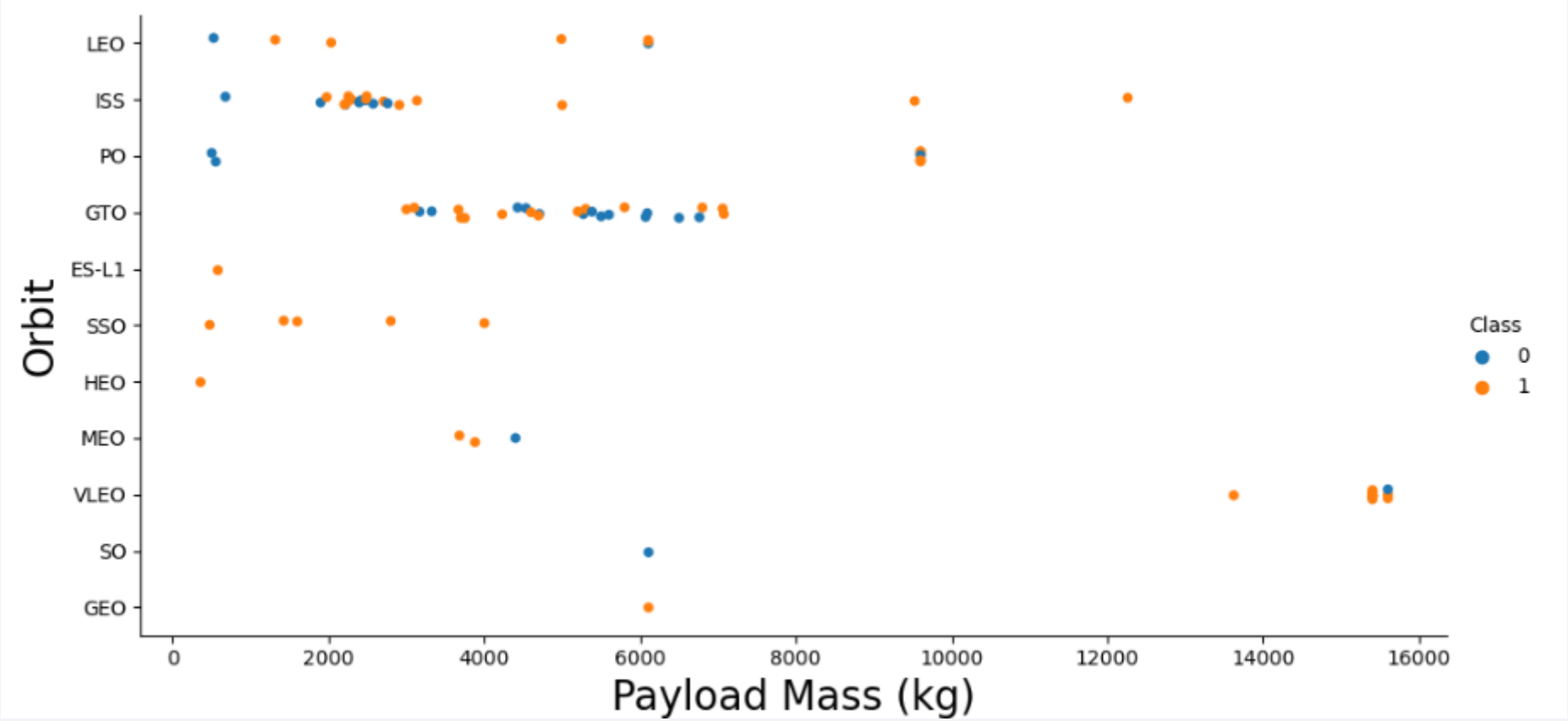
Flight Number vs. Orbit Type

- Apparently, success rate improved over time to all orbits
- VLEO orbit seems a new business opportunity, due to recent increase of its frequency



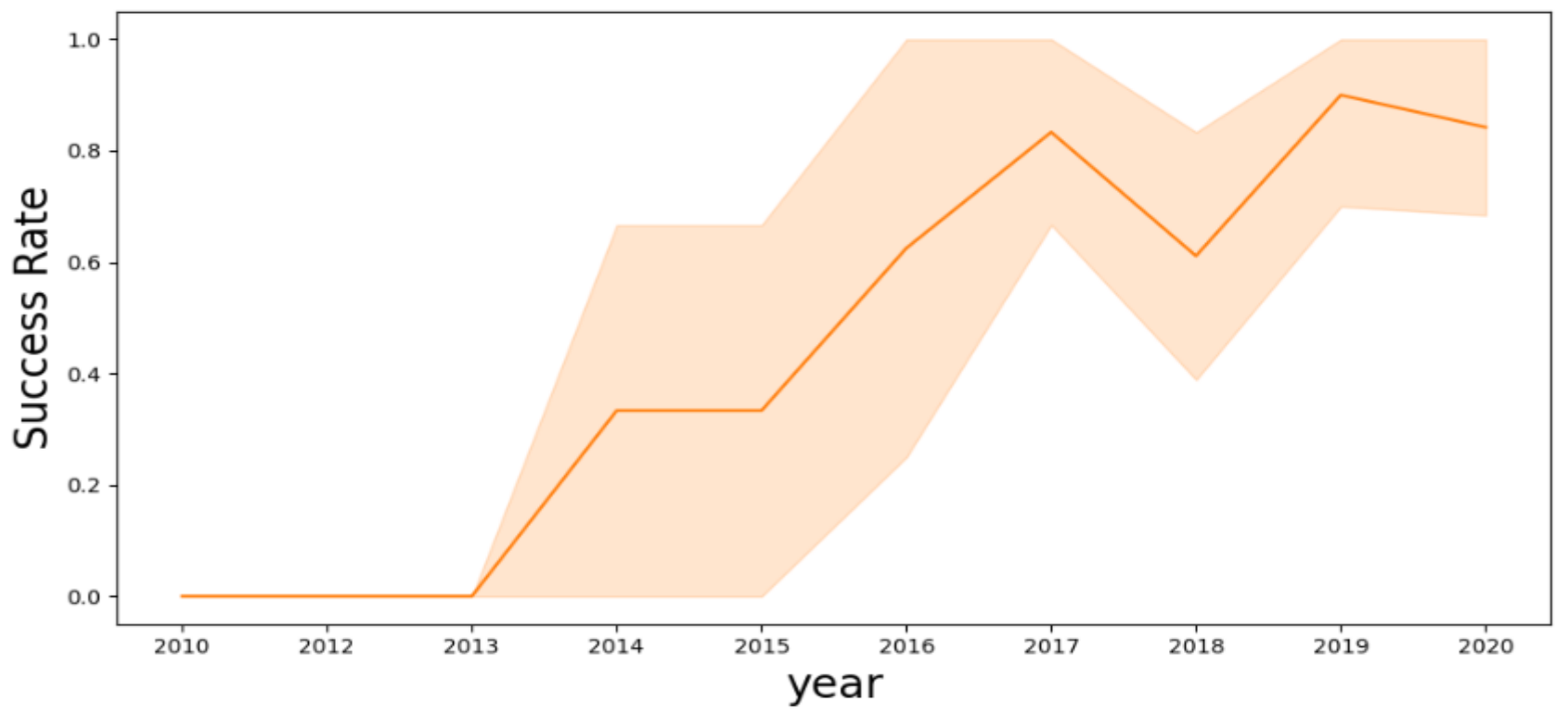
Payload vs. Orbit Type

- Apparently, there is no relation between payload and success rate to orbit GTO
- ISS orbit has the widest range of payload and a good rate of success
- There are few launches to the orbits SO and GEO.



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020
- It seems that the first three years were a period of adjustments and improvements of technology.



All Launch Site Names

- According to data, there are four launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A and VAFB SLC-4E
- They are obtained by selecting unique occurrences of "*launch_site*" values from the dataset.

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

Date	Time UTC	Booster Version	Launch Site	Payload	Payload Mass kg	Orbit	Customer	Mission Outcome	Landing Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	**CCA**FS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	**CCA**FS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	**CCA**FS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	**CCA**FS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	**CCA**FS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Total payload carried by boosters from NASA: 111.268 Kg
- Total payload calculated above, by summing all payloads whose codes contain ‘CRS’, which corresponds to NASA.

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2,928 Kg

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 Kg.

First Successful Ground Landing Date

- First successful landing outcome on ground pad = 22 Dec 2015
- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster Version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1022
F9 FT B1026

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

Mission Outcome	Occurrences
Success	99
Success (payload status unclear)	1
Failure (in flight)	1

- Grouping mission outcomes and counting records for each group led us to the summary above

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass:

Booster Version (...)
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3

Booster Version
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- These are the boosters which have carried the maximum payload mass registered in the dataset

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Booster Version	Launch Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- The list above has only two occurrences

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

- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order:

Landing Outcome	Occurrences
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

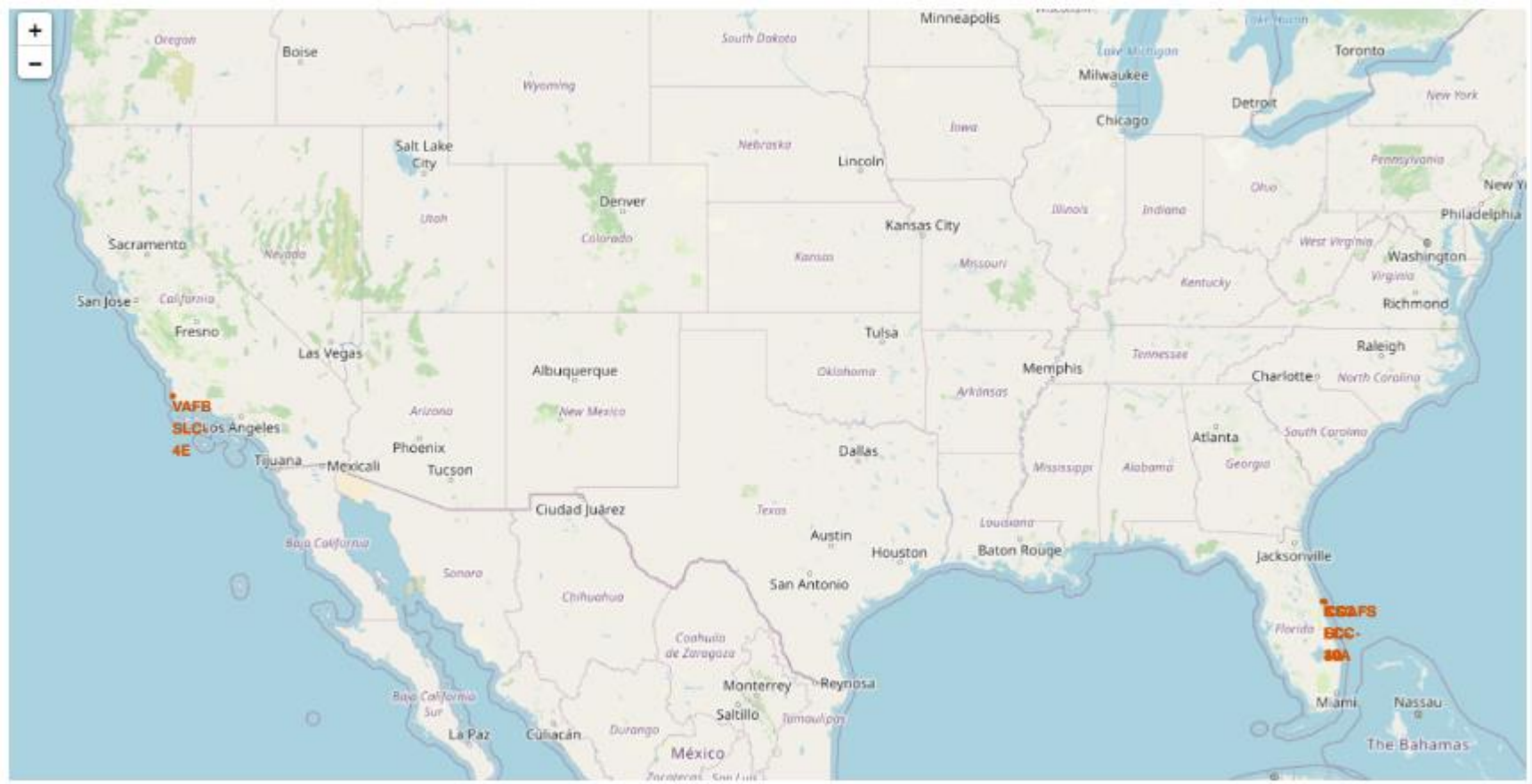
- This view of data alerts us that “No attempt” must be taken in account

Section 3

Launch Sites Proximities Analysis

Launch Sites

- Sites are all near the ocean, but also good infrastructure e.g. roads, railways etc.



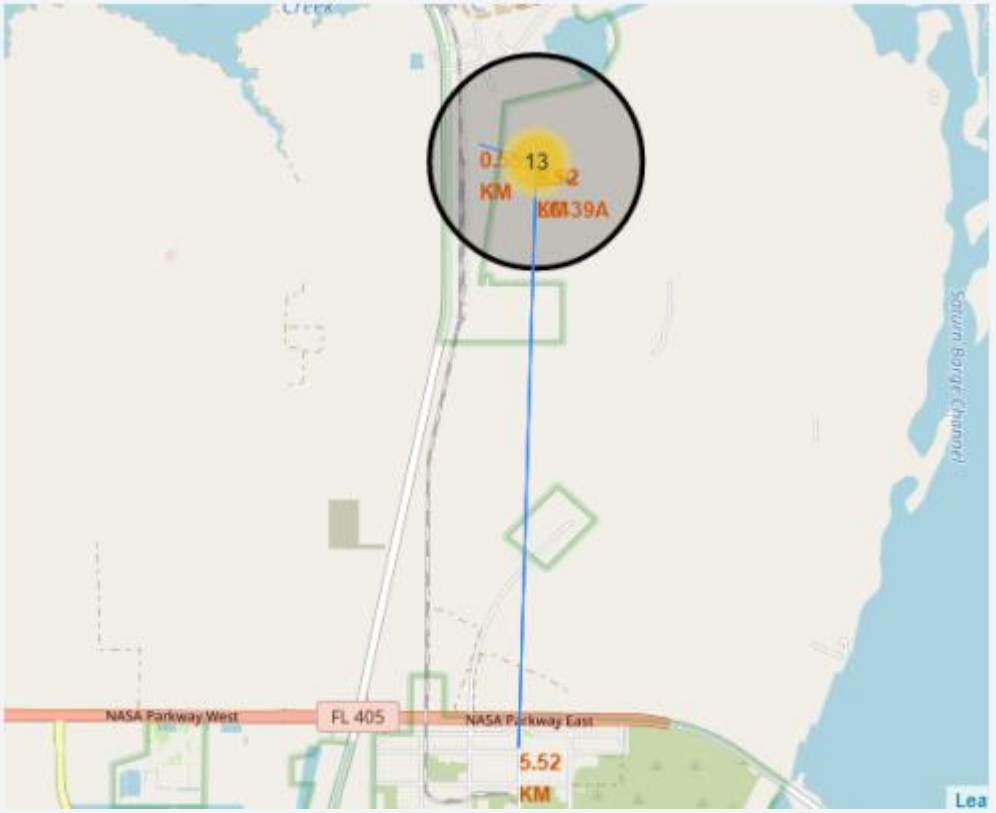
Launch Outcomes by Site

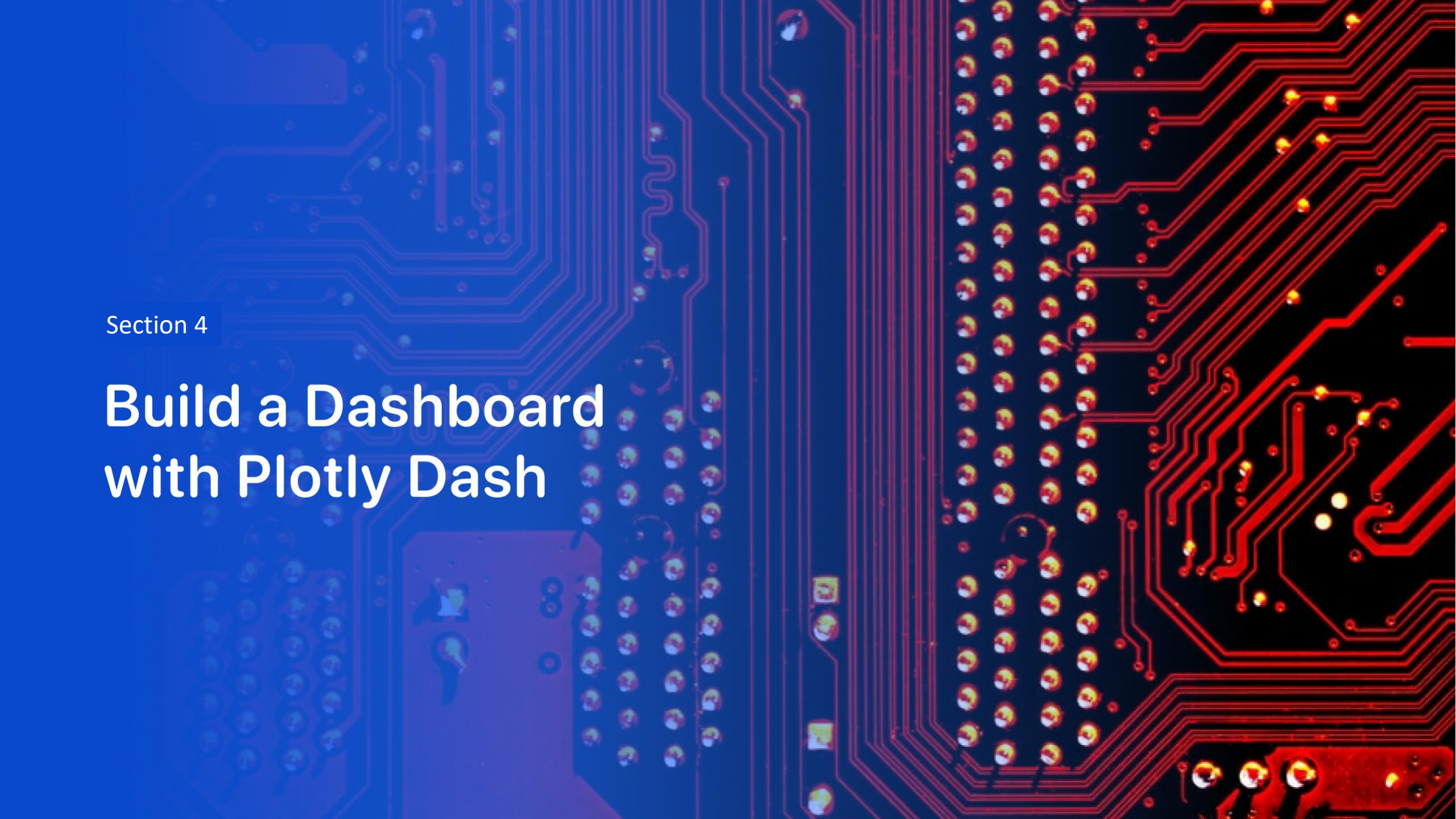
- Example of KSC LC-39A launch site launch outcomes

- Green markers indicate successful and red ones indicate failure

Launch Site Transport Proximity

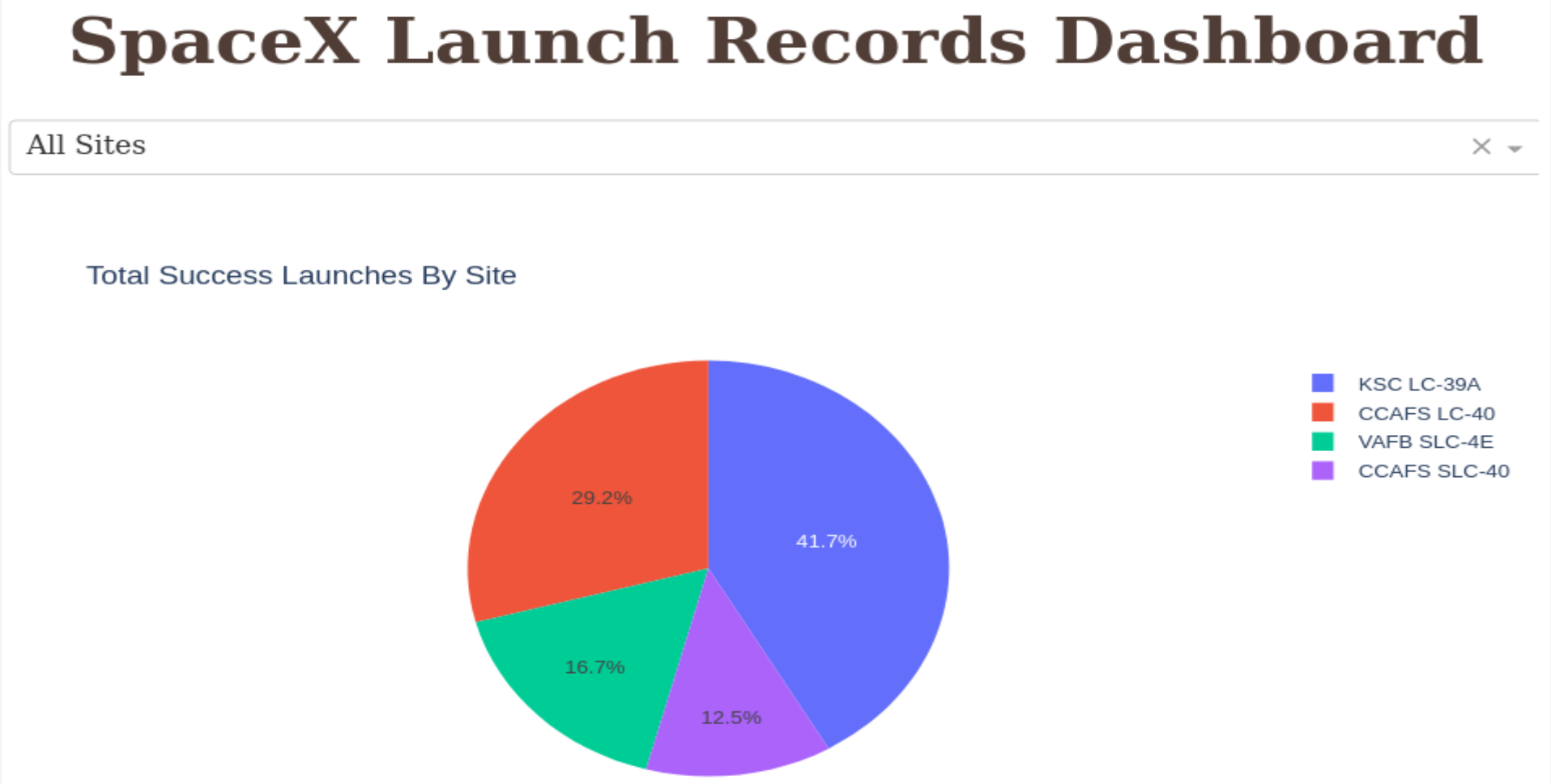
- Launch site KSC LC-39A has good surrounding logistics being near railroad and road and relatively far from inhabited areas




Section 4
Build a Dashboard with Plotly Dash

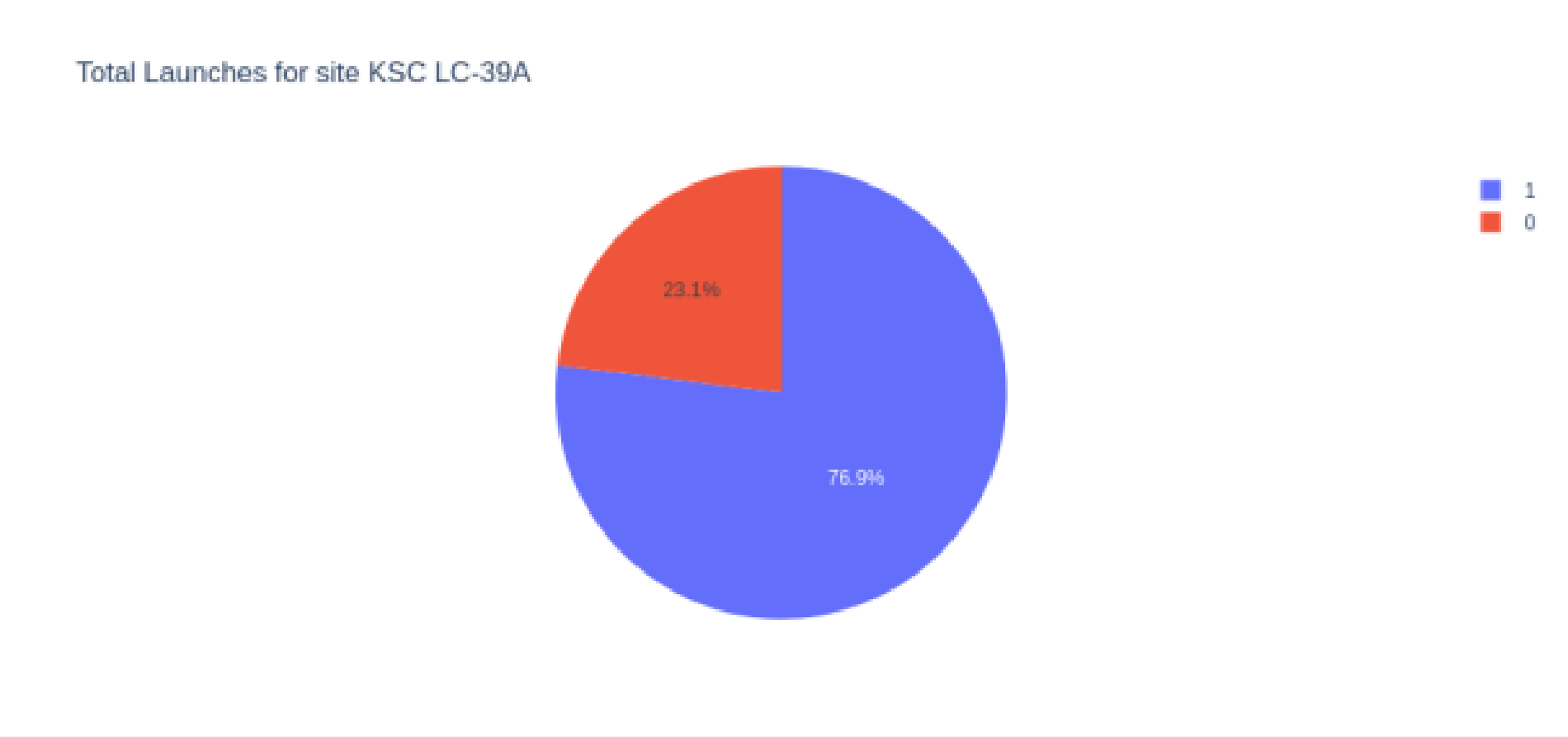
Successful Launches by Site

- The place from where launches are performed seems to be a very important factor in mission success



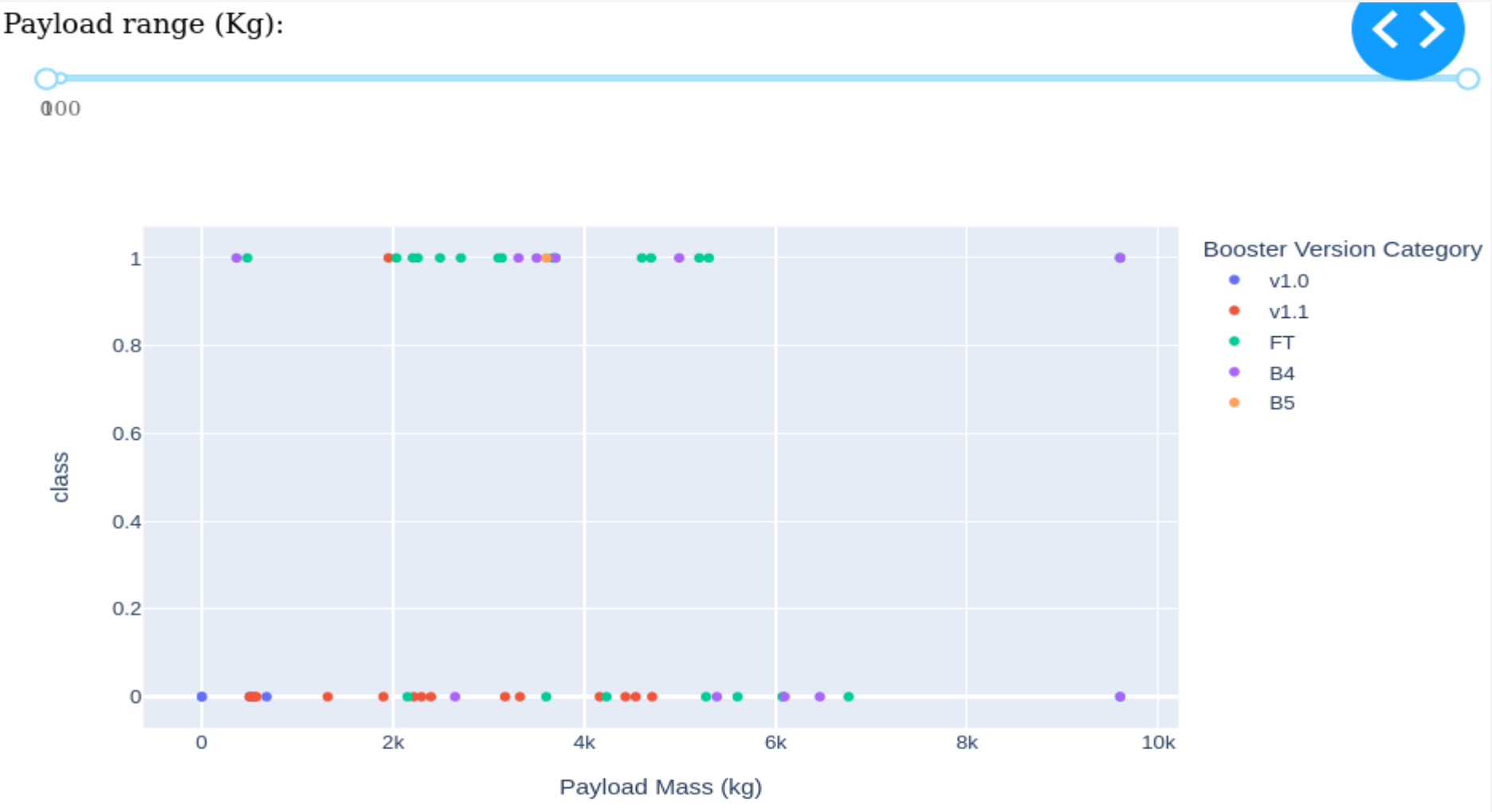
Launch Success Ratio for KSC LC-39A

- 76.9% of launches are successful in this site (1 vs 0)



Payload vs. Launch Outcome

- Payloads under 6,000 Kg and FT boosters are the most successful combination

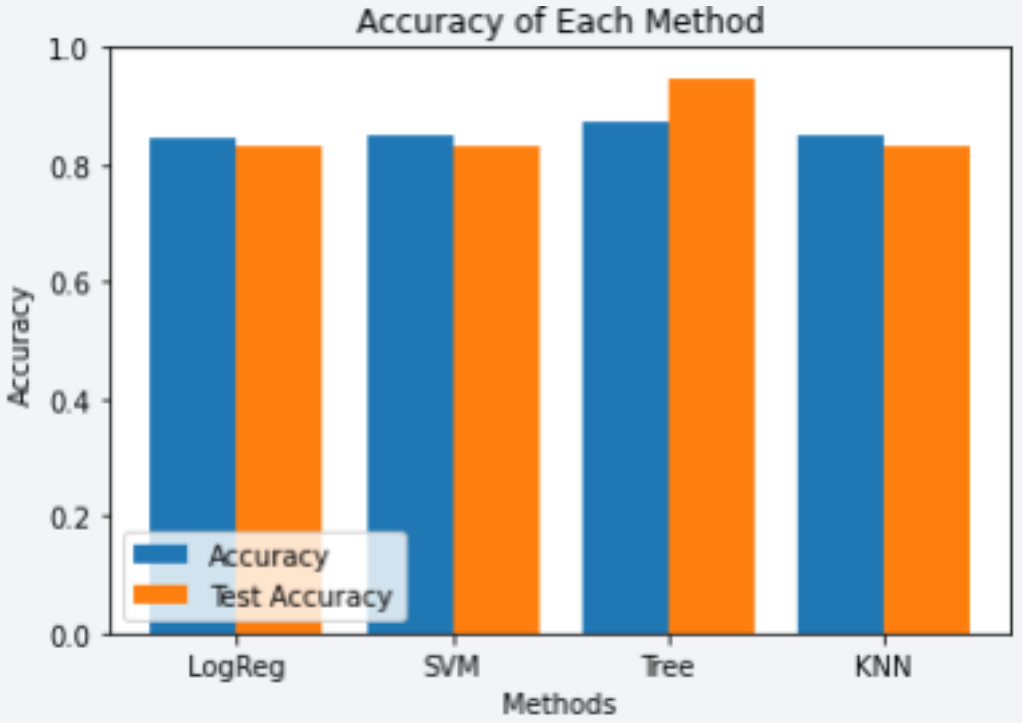


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix of Decision Tree Classifier

- The Confusion Matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones

Conclusions

- Different data sources were analysed throughout the process
- The best launch site is KSC LC-39A
- Launches above 7,000kg are less risky than others
- Although most mission outcomes were successful, landing outcomes seem to improve over time, which suggests an evolution of processes and rockets
- The Decision Tree Classifier can be used to predict successful landings and increase profitability

Appendix

- Underlying data available on https://github.com/JonathanS-cmd/Jonathan

Thank you!