IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mandy Priscilla
22 November 2022

Outline

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

Executive Summary

Summary of methodologies

- The data is collected using API and webscraping. In the data wrangling part, the landing outcome is converted to 1 for success and 0 for failed. The exploratory data analysis is done using SQL and Visualization (Python). Map visualization is done using Folium. Interactive visualization is done using Plotly Dash. Machine Learning is used for predictive analysis and determine which model has the best accuracy.

Summary of all results

- Most of the launches are successful.

Introduction

Project background and context

- SpaceX is able to send manned missions to Space because the rocket launches are relatively inexpensive. Much of the SpaceX savings is because it can reuse its first stage. Therefore, if the first stage can be determined whether it will land, the cost of a launch can be determined. Spaces X's Falcon 9 launch like regular rockets. A new company, Space Y, would like to compete with SpaceX.

Problems you want to find answers

- This project aimed to determine the price of each launch to help Space Y in competing with SpaceX. This project will also determine if SpaceX will reuse the first stage, train a machine learning model, and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

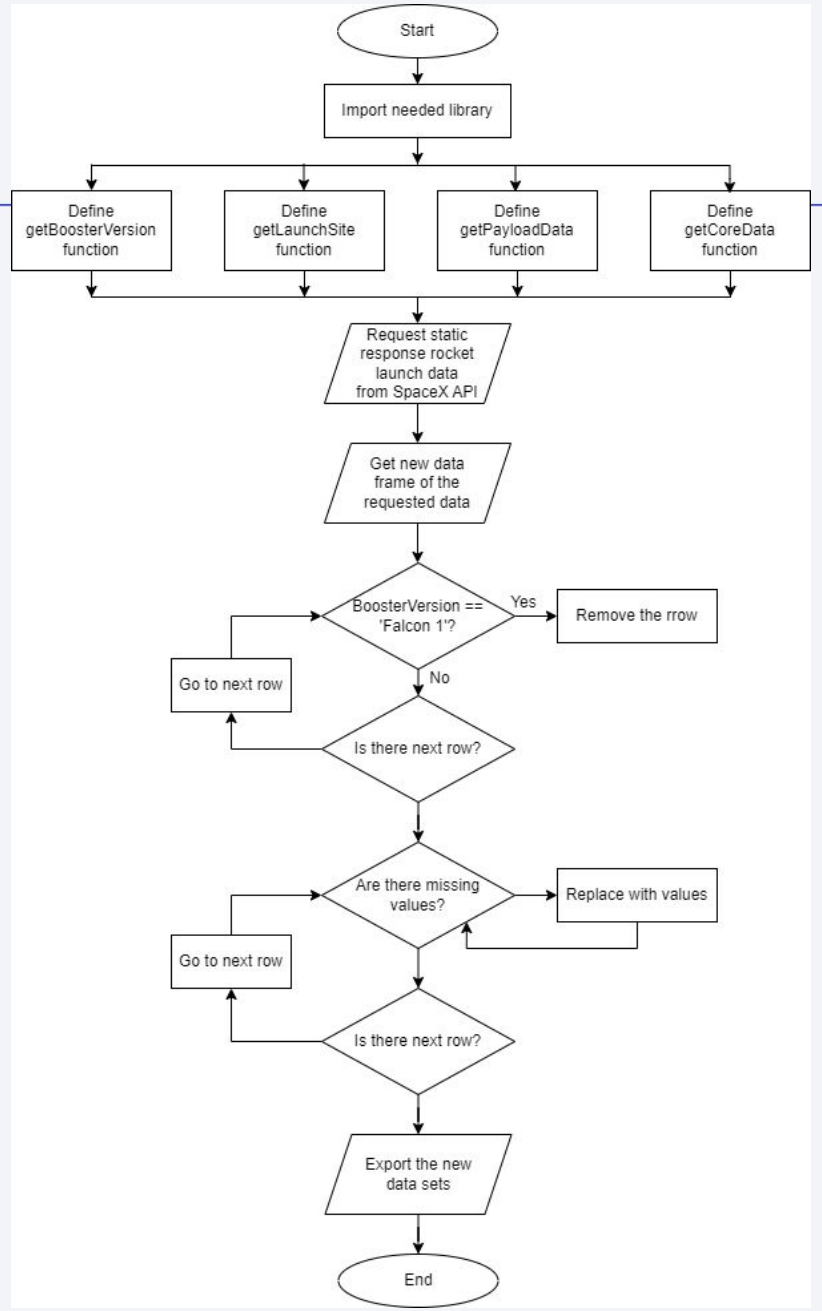
- Data collection methodology:
  - The data was collected using API and webscraping.
- Perform data wrangling
  - The data was processed to create landing outcome label.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- The data sets were collected from the Space X Data API. Some helper functions used to extract the rocket, launchpad, payload, and cores data.
- SpaceX launch data requested and parsed using the GET request. A new data frame constructed to store the data obtained. The data filtered to only contains Falcon 9 launches.
- The data must be clean before processed. Therefore, the missing values were handled by replacing null values in Landing Pad to None, and NaN values in Payload Mass with the average Payload Mass.
- The data sets then exported as a csv file to be used next.

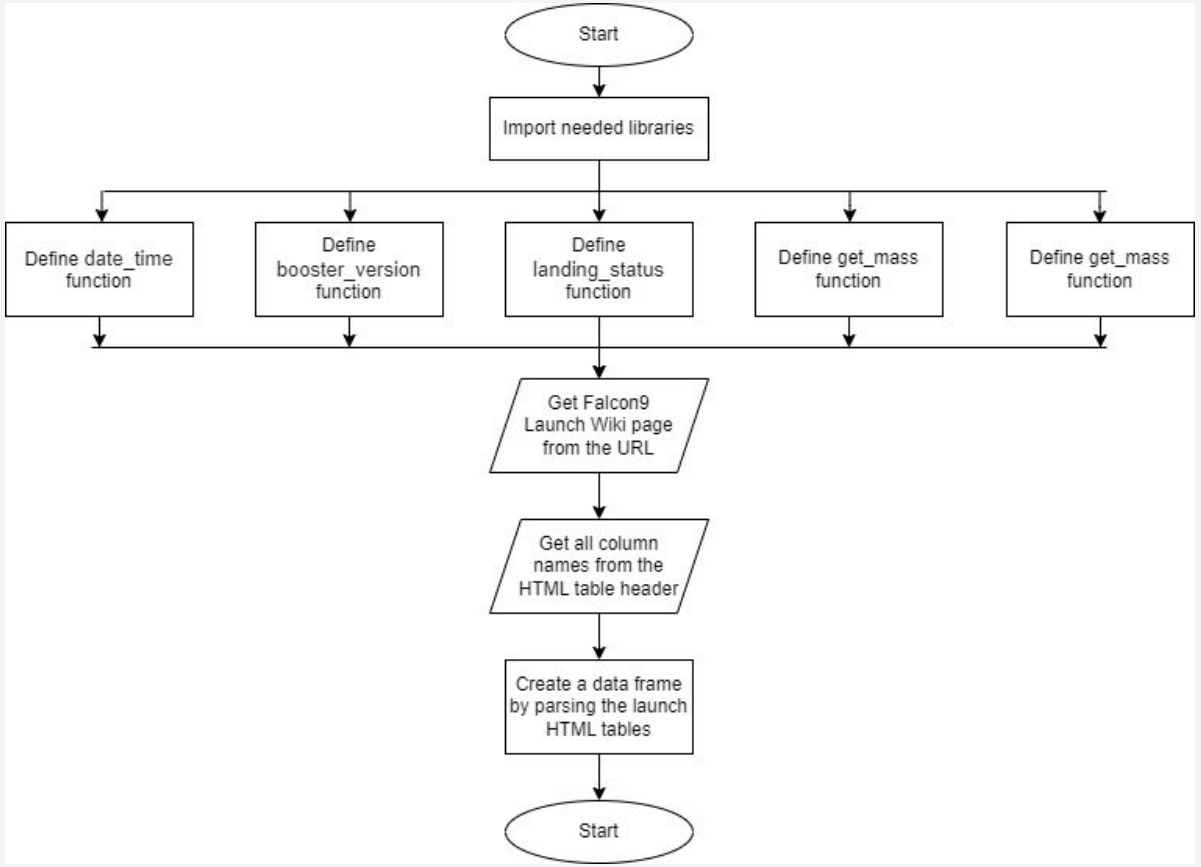
Data Collection – SpaceX API

- The process begin with importing the needed libraries and defining the needed functions. Then, the data requested from the API. The new data frame is created.
- For all rows, if the BoosterVersion is Falcon 1, then the row will be removed. After fitered, the missing values will be replaced and the new data set is obtained.
- Data Collection on GitHub



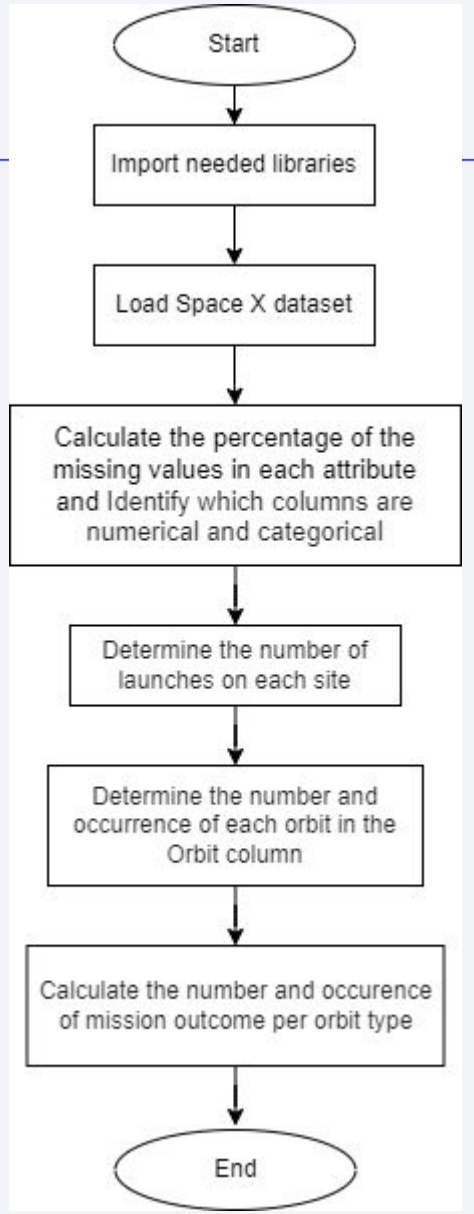
Data Collection - Scraping

- The process begin with importing the needed libraries and defining the needed functions. Then, the Falcon9 Launch HTML page requested as an HTTP response, and collect all relevant column names from the HTML table header. Lastly, a new Pandas dataframe formed from column names in previous task.
- Data Collection with Web Scraping on GitHub



Data Wrangling

- After the data loaded from the url, the percentage of null values of the data at first calculated. Then, each columns identified whether it contains numerical or categorical values.
- The next steps is determining the number of launches on each site, the number and occurrence of each orbit in the Orbit column, the number of landing_outcomes, and lastly creating a landing outcome label from Outcome column.
- Data Wrangling on GitHub



EDA with Data Visualization

- The visualization used scatter plot to show the correlation of two variables, bar chart to show and compare values of different groups, and line chart to show changes of values over time.
- The correlations that shown by scatter plot diagrams are Flight Number VS Launch Site, Payload Mass VS Launch Site, Flight Number VS Orbit Type, and Payload Mass VS Orbit Type.
- The success rate of each orbit types presented by a bar chart.
- The changes of success rate over year shown by a line chart.
- EDA with Data Visualization on GitHub

EDA with SQL

The SQL queries performed :

- SELECT
- DISTINCT
- FROM
- WHERE
- LIKE
- LIMIT
- AS

EDA with SQL on GitHub

Build an Interactive Map with Folium

- Markers used : icon=DivIcon, icon_size=(20,20), icon_anchor=(0,0), added to identify which launch sites have relatively high success rates.
- Circles used : radius=1000, color='#000000', added to highlight circle area with a text label on a specific coordinate.
- Lines used : folium.PolyLine(locations=coordinates, weight=1), added to measure the distance of specific point to the nearest railway, highway, and city in the map.
- Data Visualization with Folium

Build a Dashboard with Plotly Dash

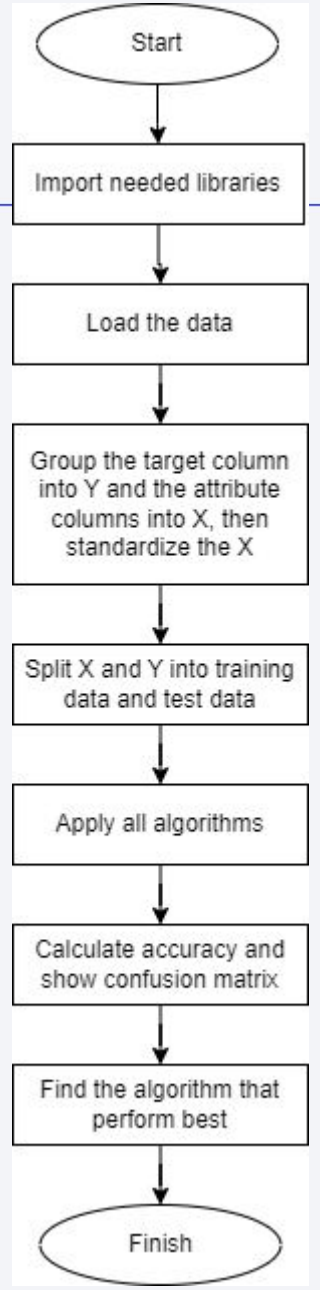
- Plot/graphs and interactions added to the dashboard :
- dropdown to select booster version
- a pie chart to show the total successful launches count for the selected sites from dropdown
- slider to select payload range of the selected site from dropdown to only show the result of certain range in order to be clearer
- scatter chart to show the correlation between payload and launch success of the selected site from dropdown

- Interactive Dashboard with Plotly Dash

Predictive Analysis (Classification)

- The classification model was built by splitting the data into 80% training data and 20% test data. The algorithms used and the accuracy is to be compared are Logistic Regression, SVM, Decision Tree, and KNN. The algorithm with the best accuracy turned out to be Decision Tree, with 88,3% accuracy and 83,3% test accuracy.

- Machine Learning Prediction on GitHub



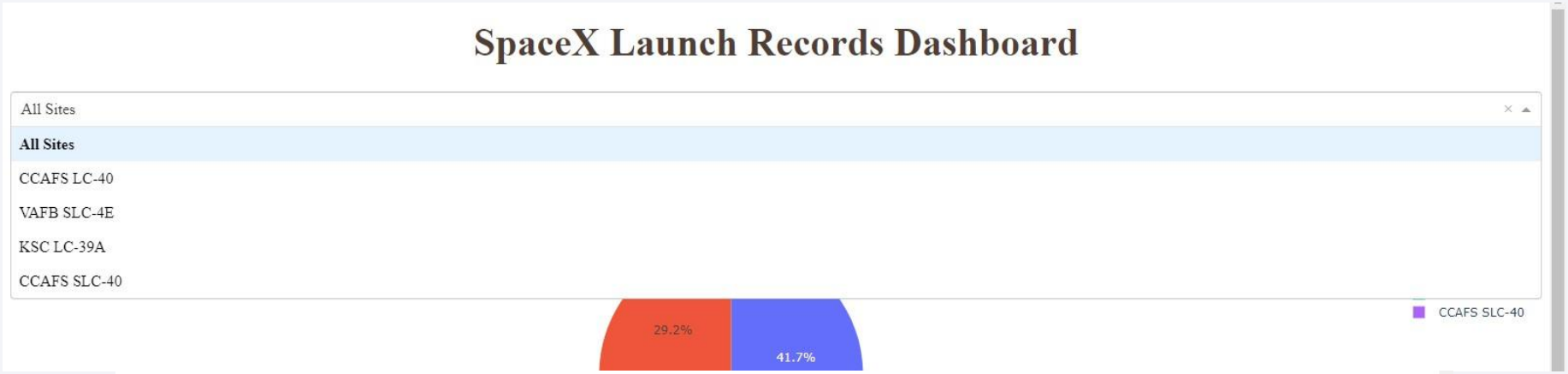
Results

Exploratory data analysis results

- There are 4 distinct launch sites : CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, dan CCAFS SLC-40.
- The total payload mass carried by boosters launched by NASA (CRS) is 111,268 kg.
- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.
- The first successful landing outcome in ground pad was achieved on 01-05-2017.
- Boosters that have success in drone ship and have payload mass between 4000-6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.
- Orbits that have the highest success rate are ES-L1, GEO, HEO, and SSO.
- The yearly trend of launch success is increasing until 2017, but then decreasing until 2020.

Results

Interactive analytics demo in screenshots



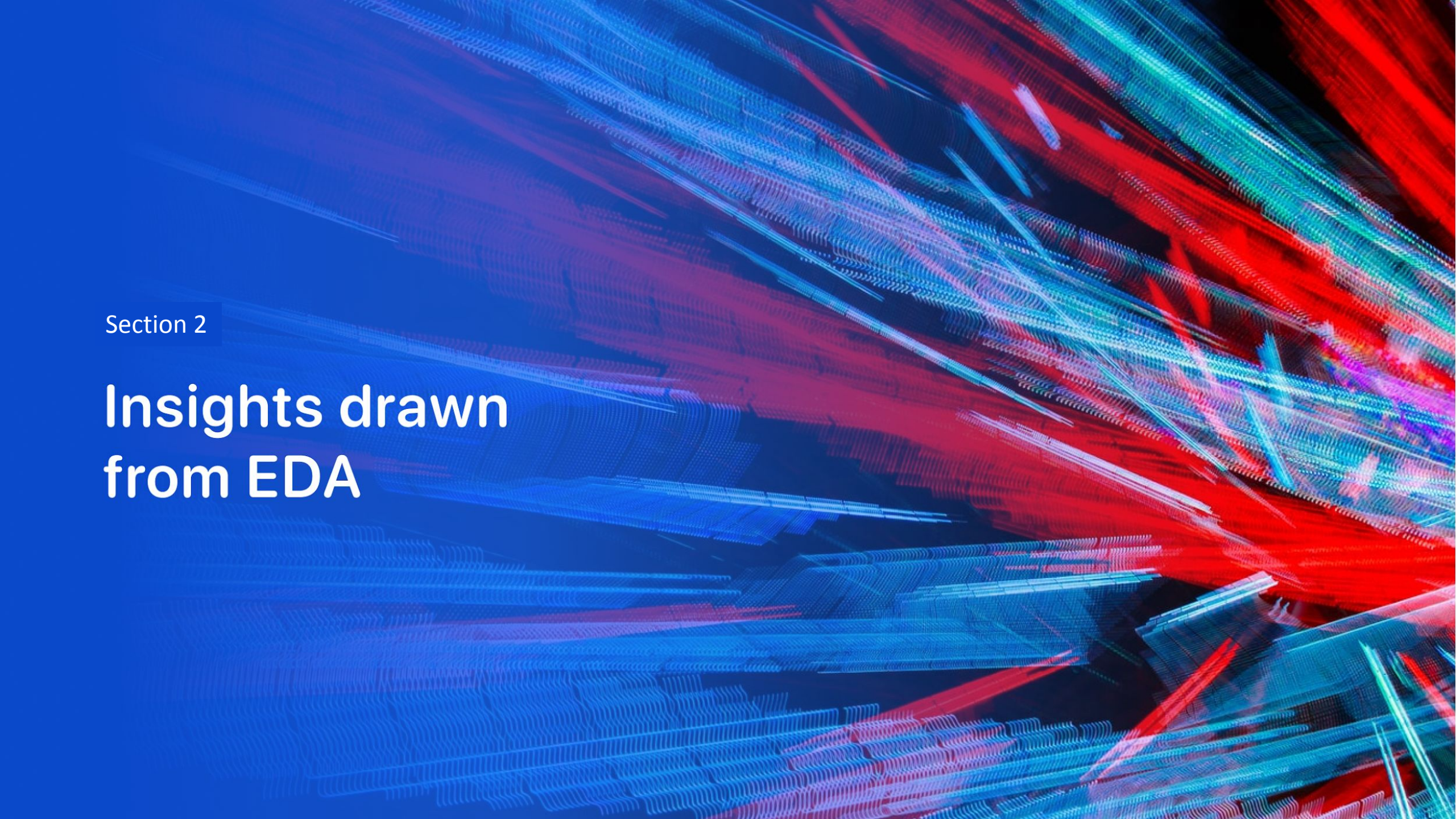
Results

Predictive analysis results

- The Decision Tree Classification Model has the highest accuracy and test accuracy.
- The accuracy of Decision Tree Confusion Matrix is 66.67%.

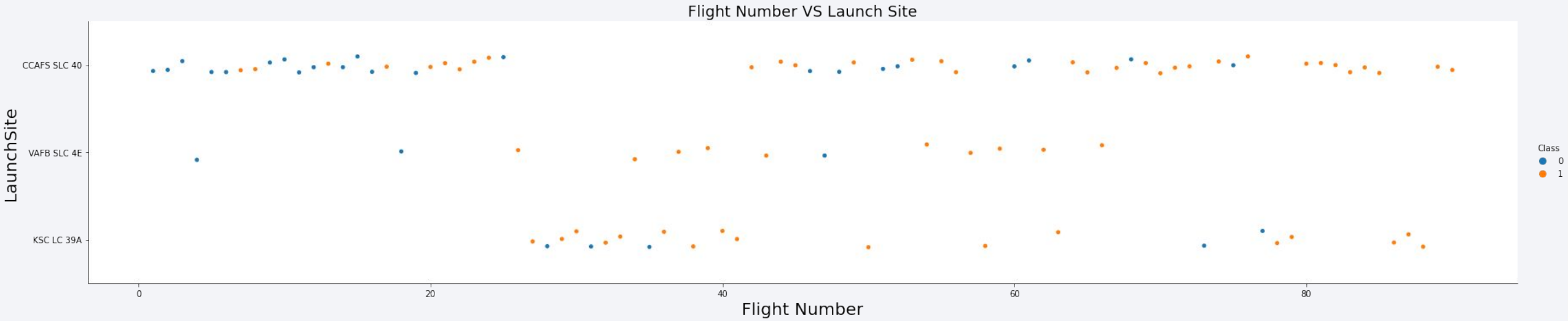
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

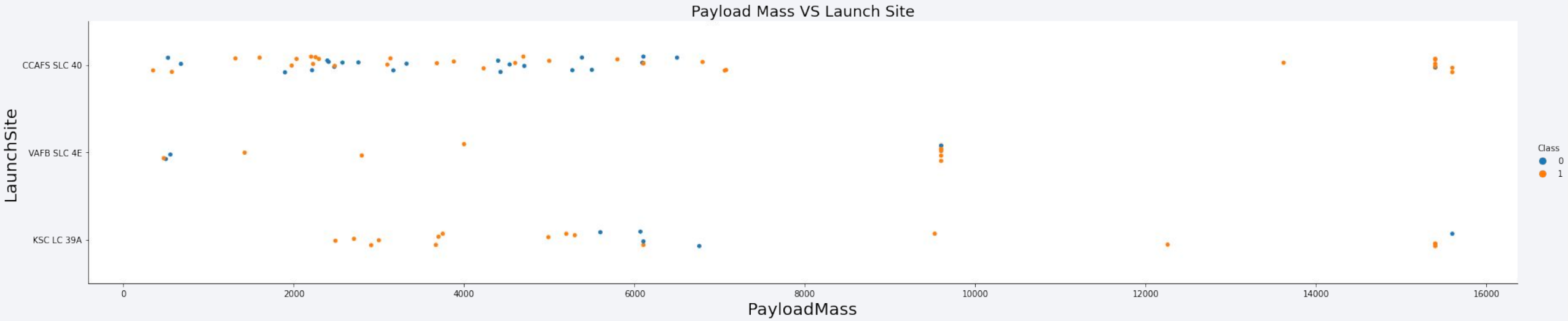
- Scatter plot of launch site of each flight number



- VAFB SLC 4E has the highest rate of success of 76.9%.
- Successful launches of VAFB SLC 4E ranged in flight number 25-45 and 50-70.

Payload vs. Launch Site

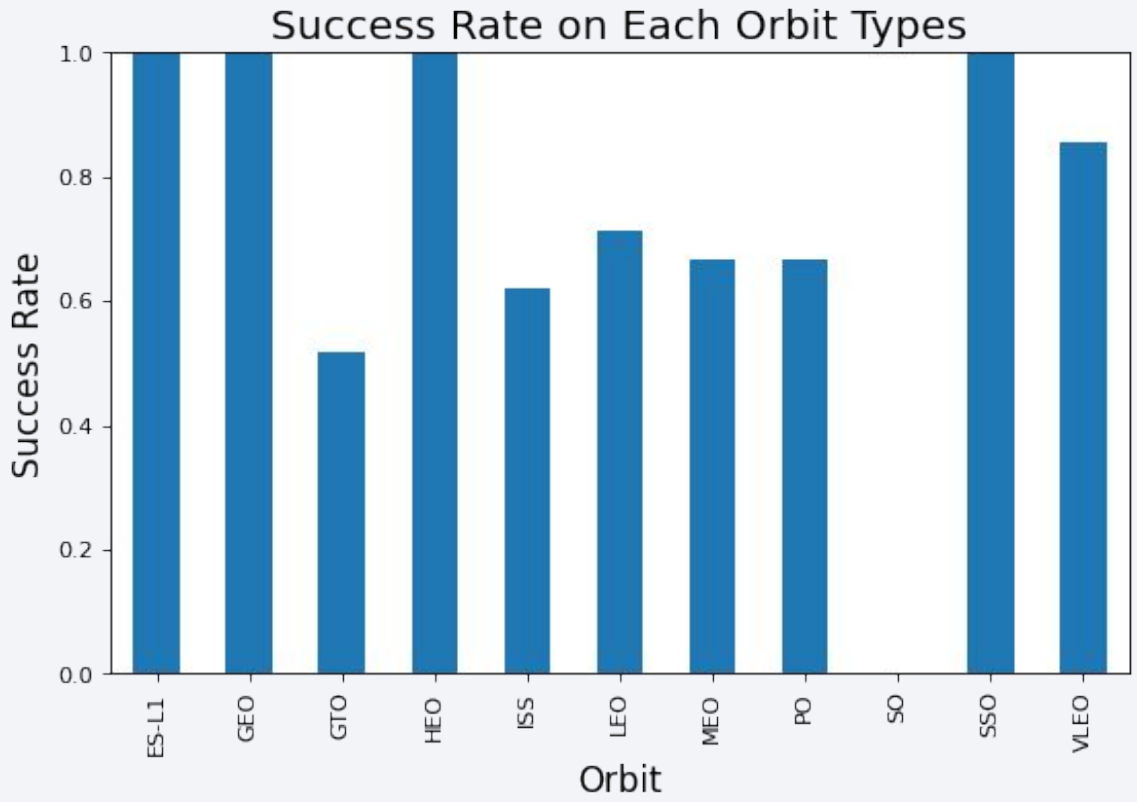
- Scatter plot of launch site of each payload mass



- VAFB SLC 4E has the highest rate of success of 76.9%.
- Successful launches of VAFB SLC 4E have payload mass between 0-8.000 kg.
- Most of the launches are from CCAFS SLC 40.

Success Rate vs. Orbit Type

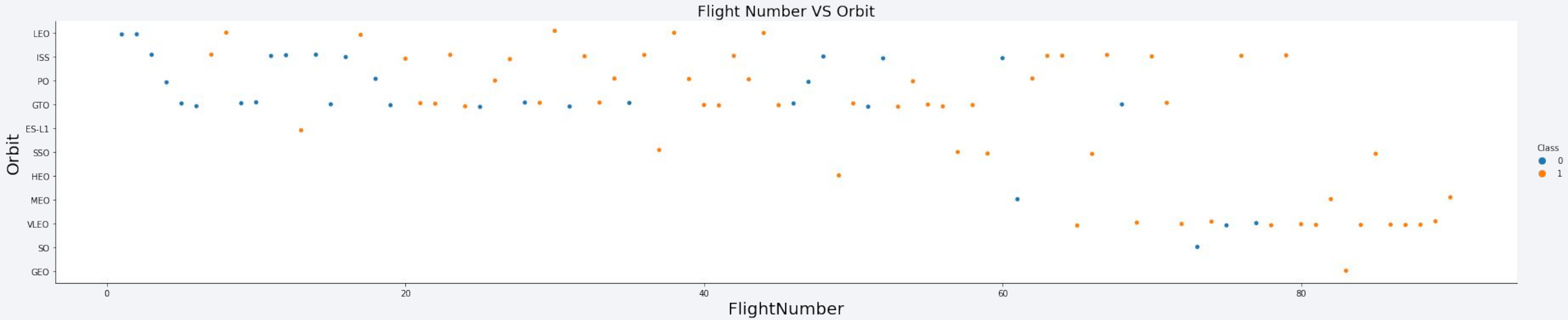
- Bar chart for the success rate of each orbit type



- The orbits that have the highest success rate is ES-L1, GEO, HEO, and SSO with 100% success rate.

Flight Number vs. Orbit Type

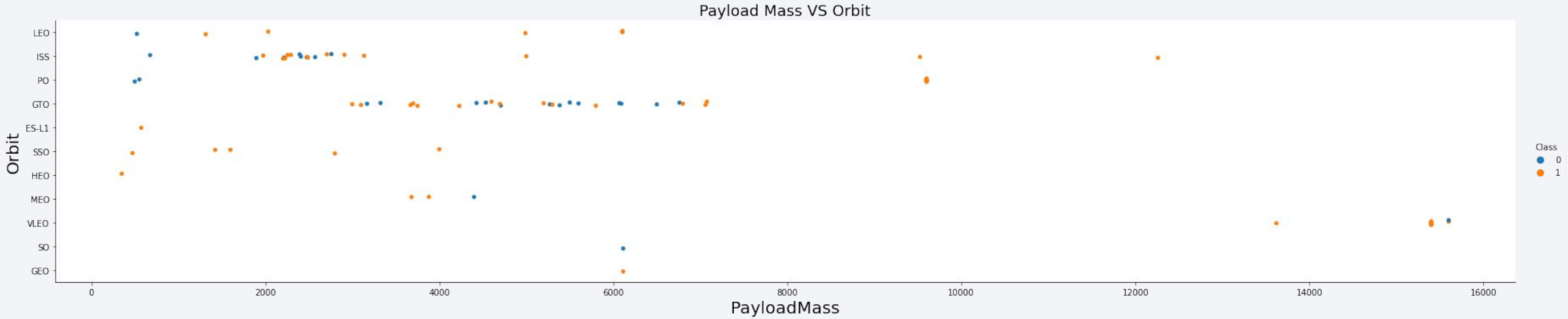
- Scatter plot of orbit of each flight number



- Most launches are in LEO, ISS, PO, and GTO orbits.
- Most successful launches' flight number are between 20-45.

Payload vs. Orbit Type

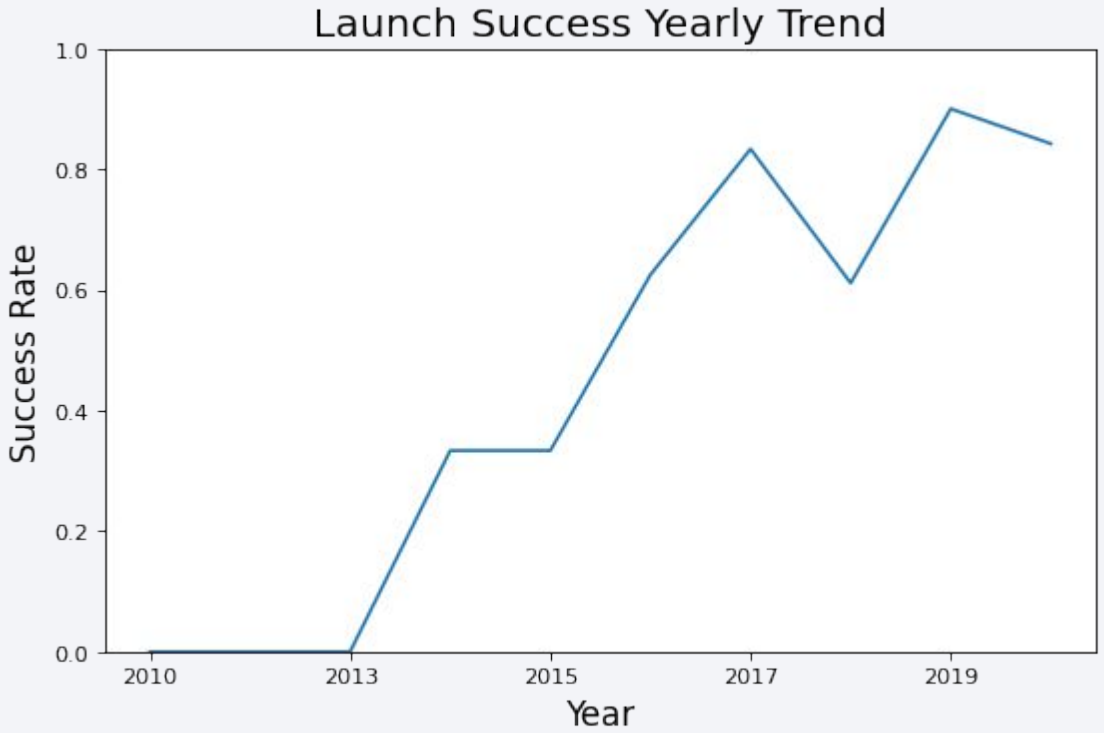
- Scatter plot of orbit of each flight number



- Most launches have the payload mass between 0-8.000 kg.
- Most launches are in GTO orbit.

Launch Success Yearly Trend

- Line chart of yearly average success rate



- The yearly trend of successful launch is increasing until 2017 and decreasing until 2020.

All Launch Site Names

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

These are the distinct values of Launch Site column.

Launch Site Names Begin with 'CCA'

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

- These are the first 5 rows of launch site CCAFS LC-40.

Total Payload Mass

TOTAL_PAYLOAD_MASS
111268

- The total payload carried by boosters from NASA is 111,268 kg.

Average Payload Mass by F9 v1.1

AVG_PAYLOAD
2928.4

- The average payload mass carried by booster version F9 v1.1 is 292,8.4 kg.

First Successful Ground Landing Date

FIRST_SUCCESS_GP
01-05-2017

- The first successful landing outcome on ground pad is happening on 1 May 2017.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are F9 FT B1022, F0 FT B1026, F9 FT B1021.2, and F9 FT B1031.2.

Total Number of Successful and Failure Mission Outcomes

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- There are 1 failure and 100 successful mission outcomes.

Boosters Carried Maximum Payload

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
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 names of the booster which have carried the maximum payload mass.

2015 Launch Records

MONTH	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- These are the months, booster versions, and launch site names of failed landing_outcomes in drone ship in year 2015.

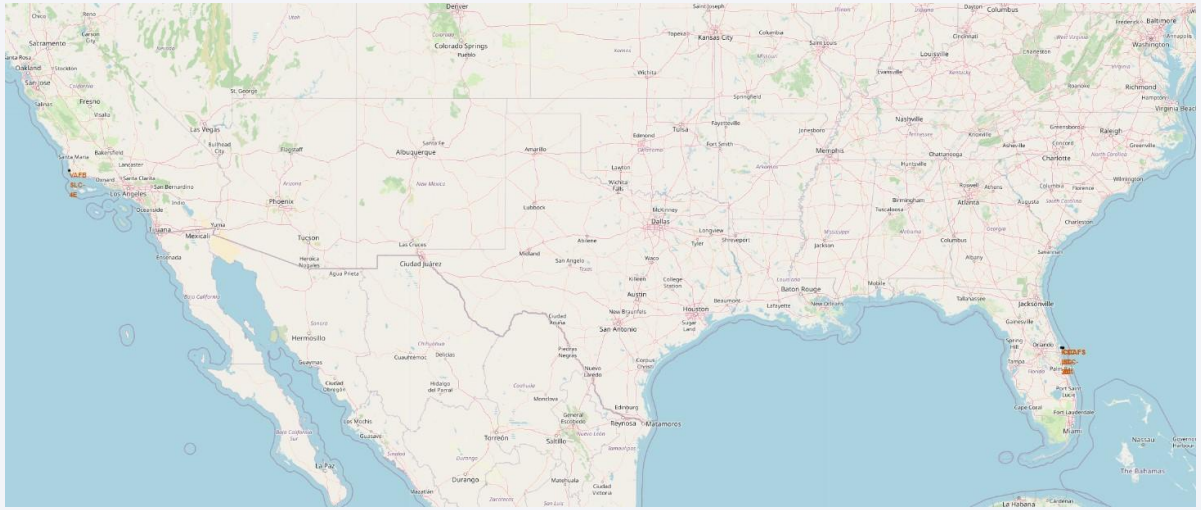
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

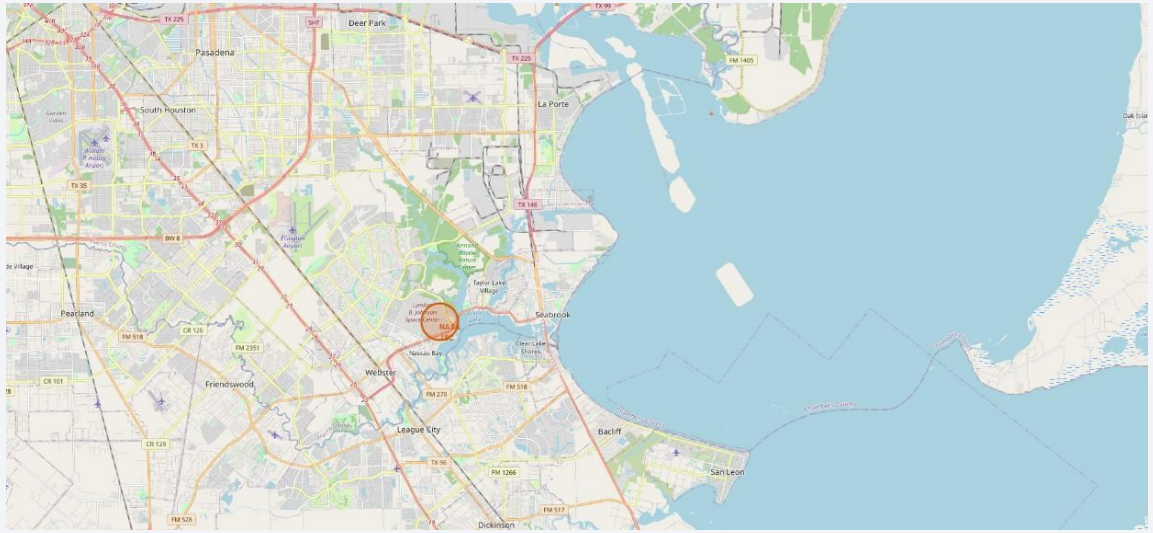
Landing _Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Controlled (ocean)	3
Failure	3
Failure (parachute)	2
No attempt	1

- These are the counts 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.

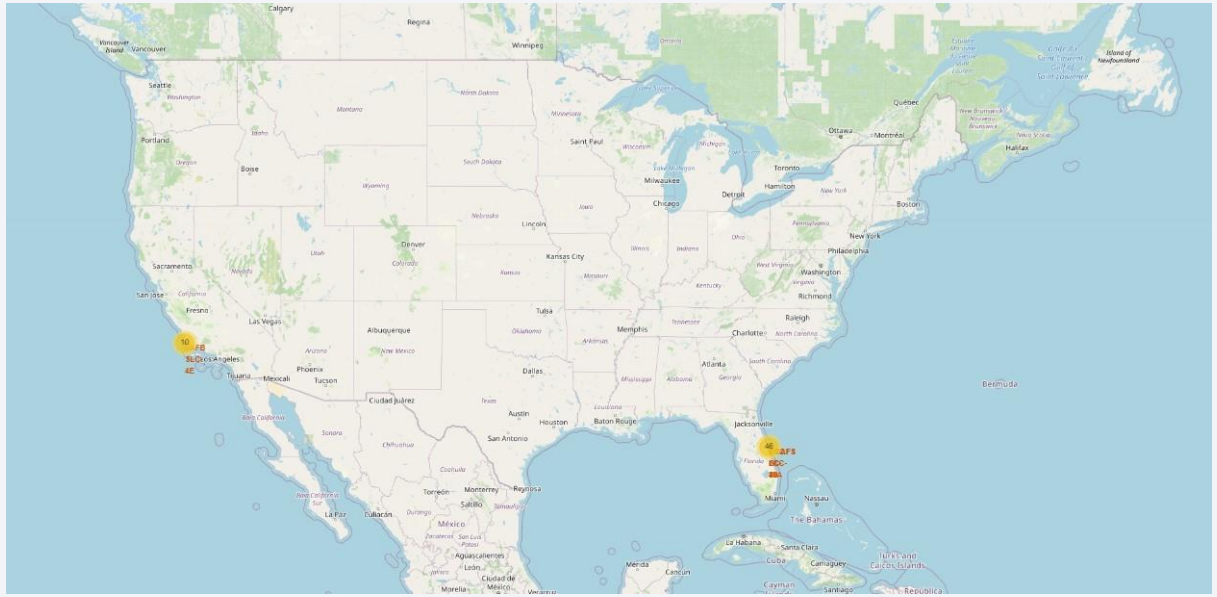

Section 3
Launch Sites
Proximities Analysis

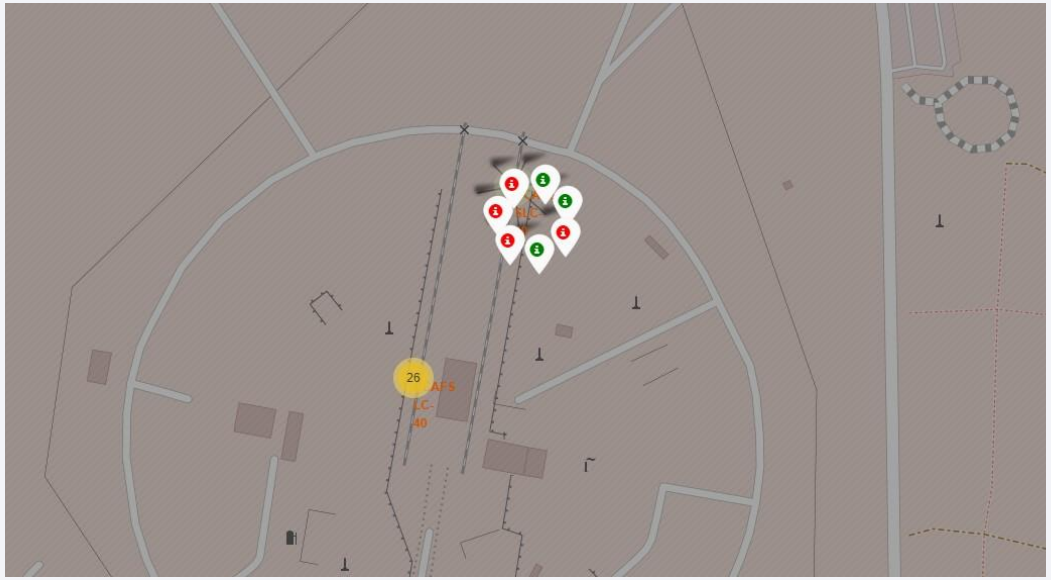
Launch Sites Locations



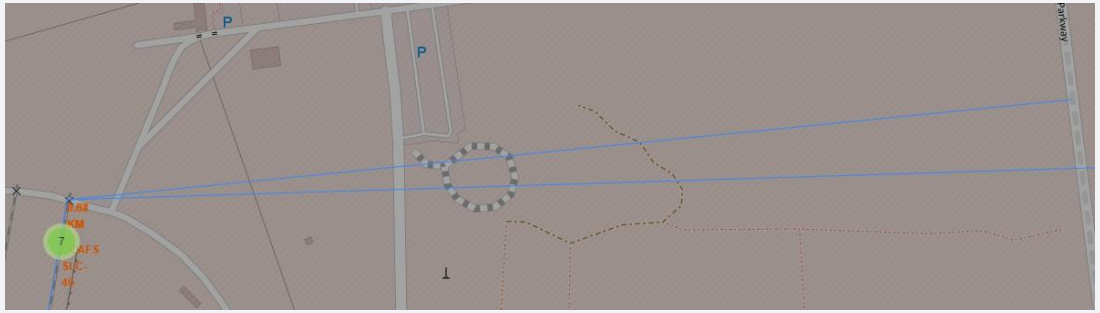


Sites Launch Outcomes

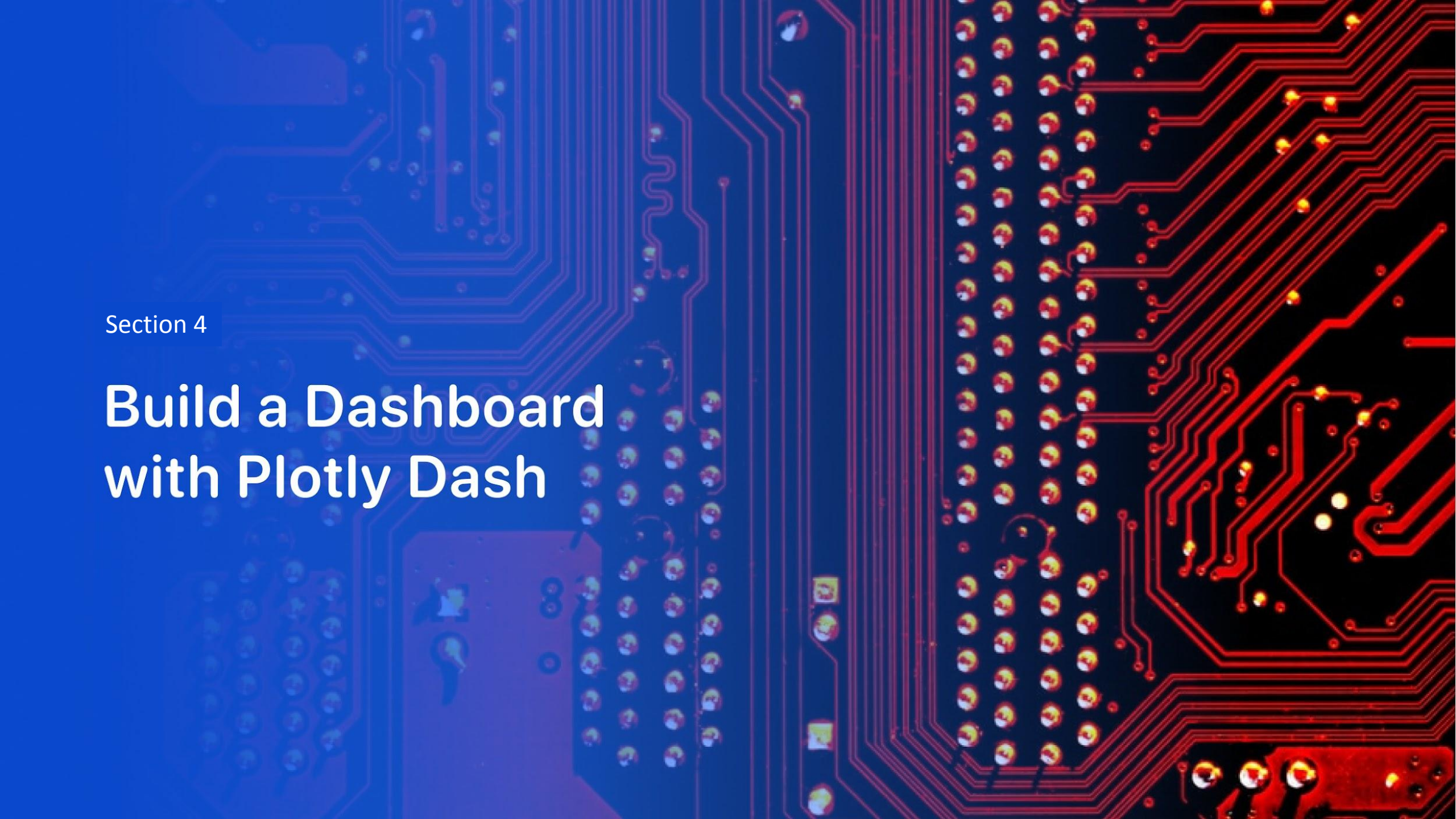




Distance of a Launch Site to Its Proximities

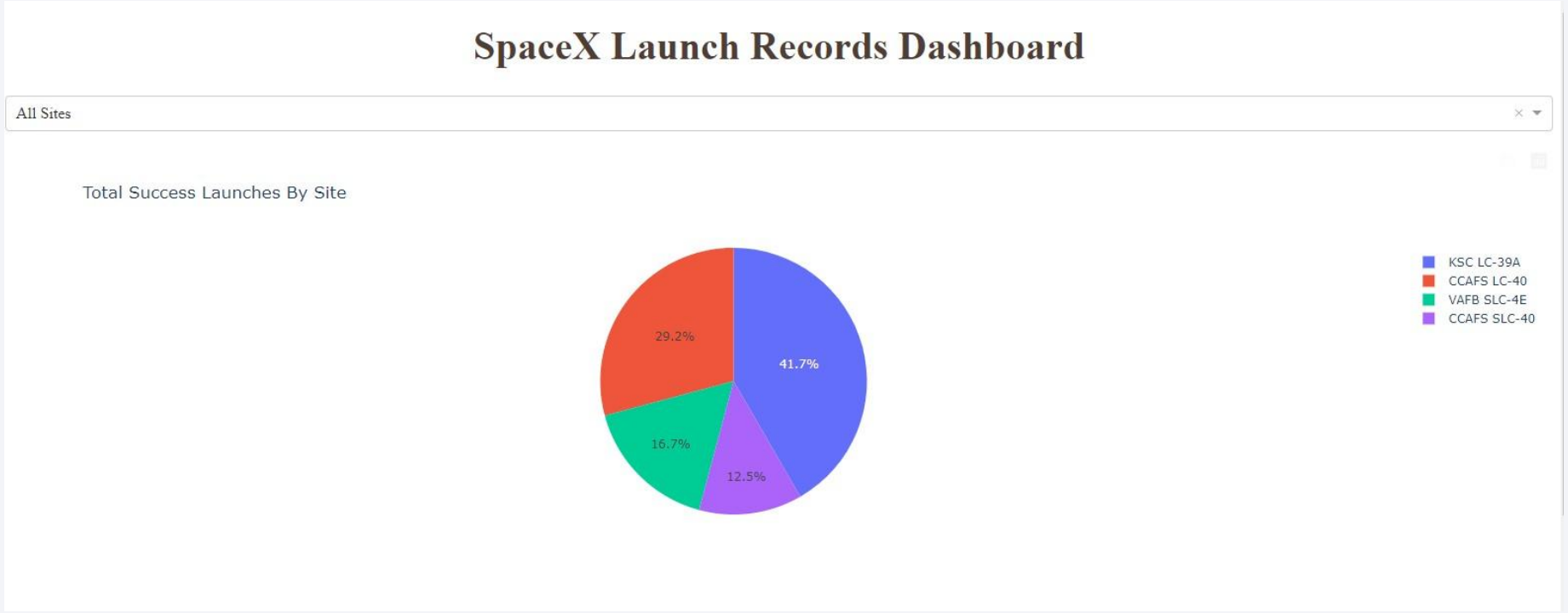


- The generated folium map show a selected launch site to its nearest coastline, railway, and highway with the distance of 0.86 km, 0.08 km, and 0.67 km respectively.


Section 4
Build a Dashboard with Plotly Dash

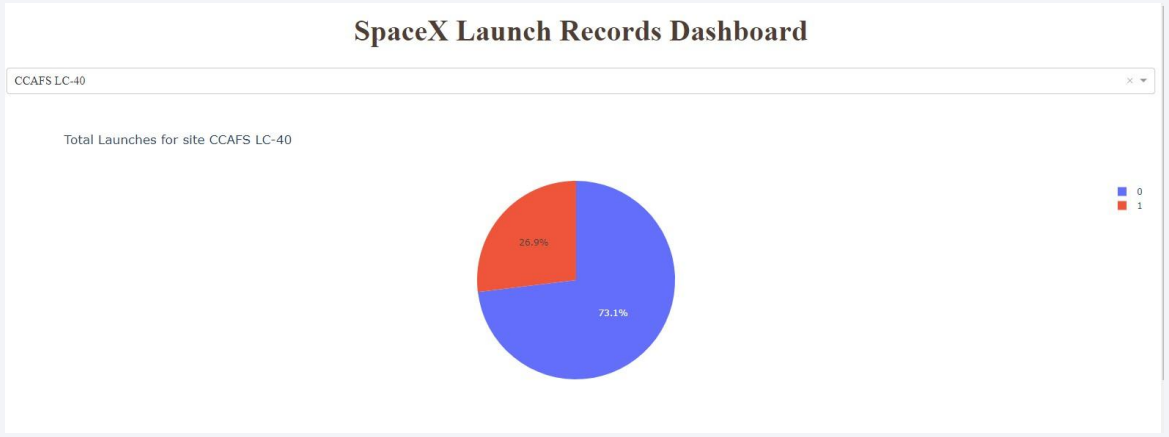
Total Success Launches by Site

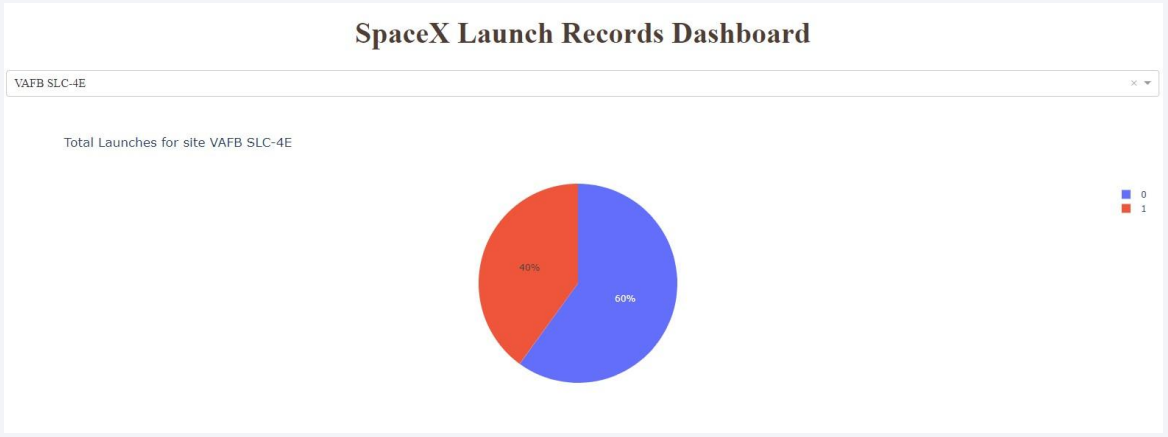
Launch success count for all sites

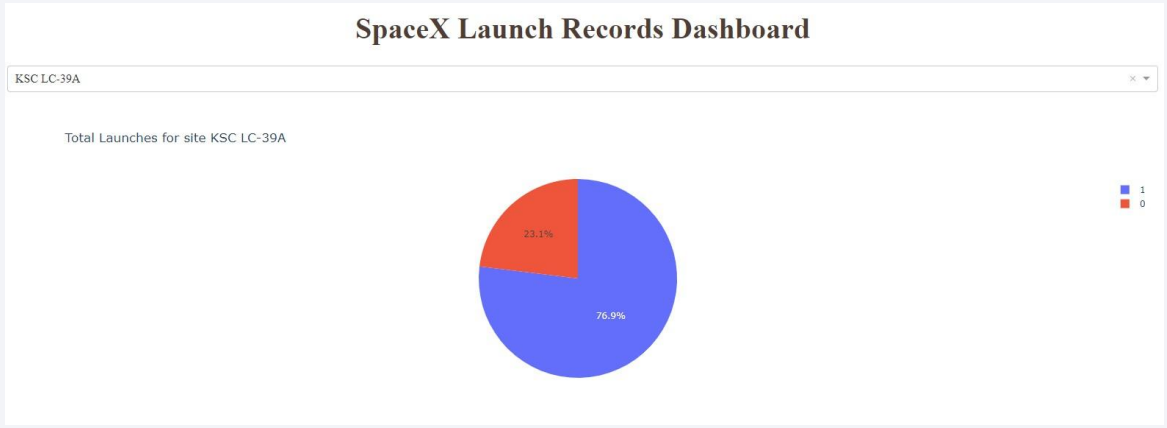


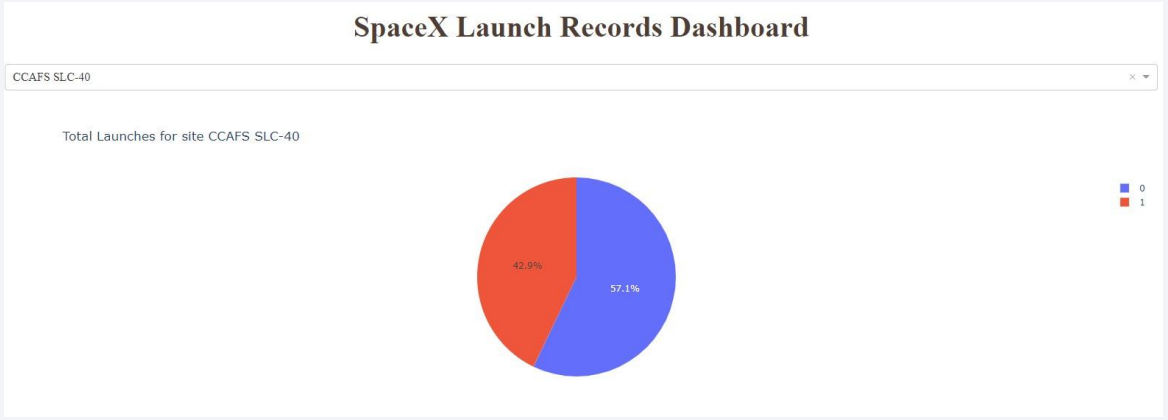
- The KSC-LC-39A site has the most success launches compared to the other sites, with 41.7% of the number of all sites.

Launch Success Ratio of All Sites



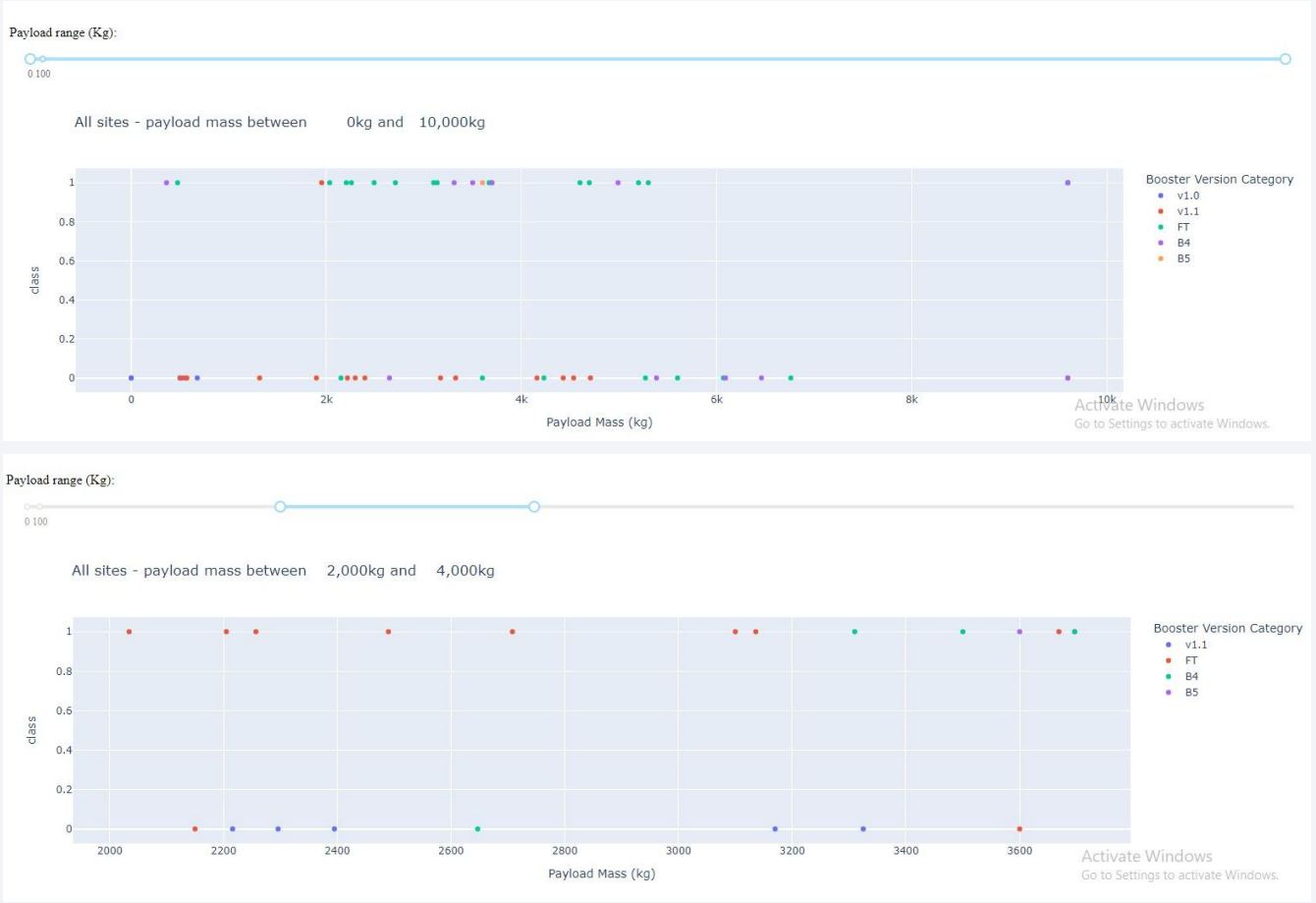






- The KSC-LC-39A site has the highest success ratio compared to the other sites, with 76.9% success percentage.

Payload vs. Launch Outcome of All Sites



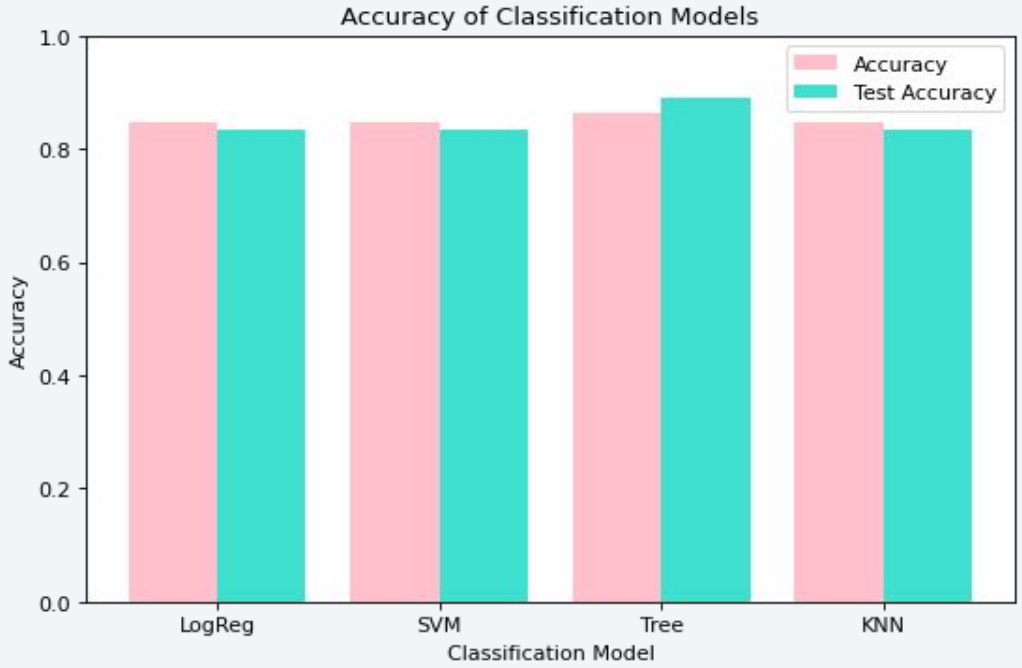
The booster version that has the largest success rate is FT, with payload range 2,000 – 4,000 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

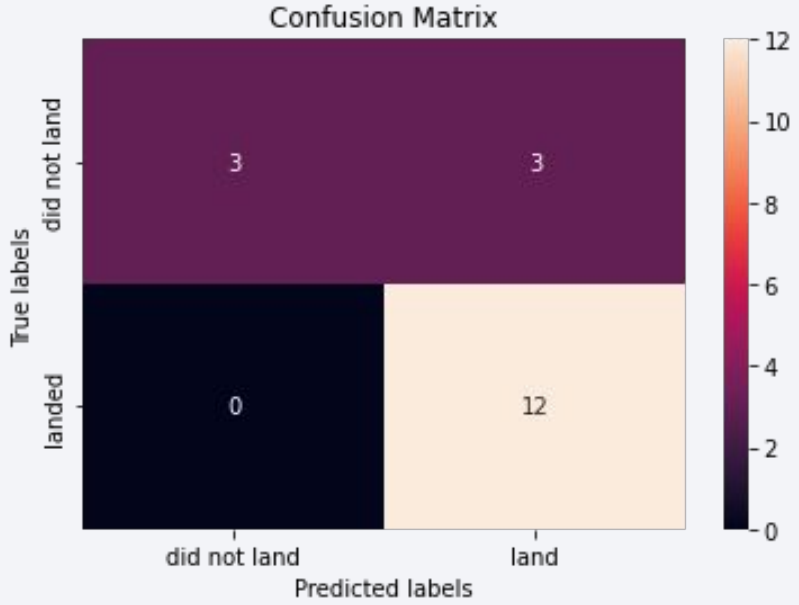
- Bar chart of the built model accuracy for all built classification models.



- Decision Tree has the highest classification accuracy.

Confusion Matrix

- Confusion matrix of Decision Tree model



The confusion matrix above shows that true positive (TP) = 12, true negative (TN) = 0, false positive (FP) = 3, and false negative (FN) = 3. The accuracy is (TP+TN)/(TP+TN+FP+FN) = (12+0)/(12+0+3+3) = 12/18 = 66.67%.

Conclusions

- The Decision Tree is the best classification model.
- Most launches are successful.

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!