IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Siwarak Sawongnam
24 Novembem 2023

Outline

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

Executive Summary

- Data Analysis techniques :
  - Data was collected by using web scraping technique and Space X API
  - Explanatory Data Analysis (EDA) is utilized to perform data analytic.
  - Machine Learning technique is applied to make a prediction.
- Summary of all results
  - The data was collected from public sources.
  - EDA must be a first step to perform Machine Learning that can act as a feature engineering to find feature that can yield the most accuracy for success of launchings.
  - Machine Learning's result is pointed out that the important of data can drive decision.

Introduction

- The goal of this project is to evaluate the performance of the new company Space Y to beat with the Space X.

- The goal's answer :
  - The solution for estimating the total cost for launches by predicting successful of the landing of the rockets and where the best location to launch the rocket.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was collected from 2 sources : Space X API and WebScraping
- Perform data wrangling
  - The library pandas was very useful to perform wrangling.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics and patterns within a dataset. By combining visualization techniques and SQL queries, you can gain deeper insights into your data. Here's a summary of performing EDA using visualization and SQL.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
  - Performing interactive visual analytics using Folium and Plotly Dash involves creating interactive maps with Folium and building interactive web applications with Plotly Dash.
- Perform predictive analysis using classification models
  - Build accurate classification models by preparing and preprocessing data, selecting an appropriate algorithm, and optimizing hyperparameters. Evaluate models using train-test split, cross-validation, and metrics like accuracy and ROC-AUC. Document the process, interpret results, and deploy the model for production if needed.

Data Collection

- Data sets were collected ;
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

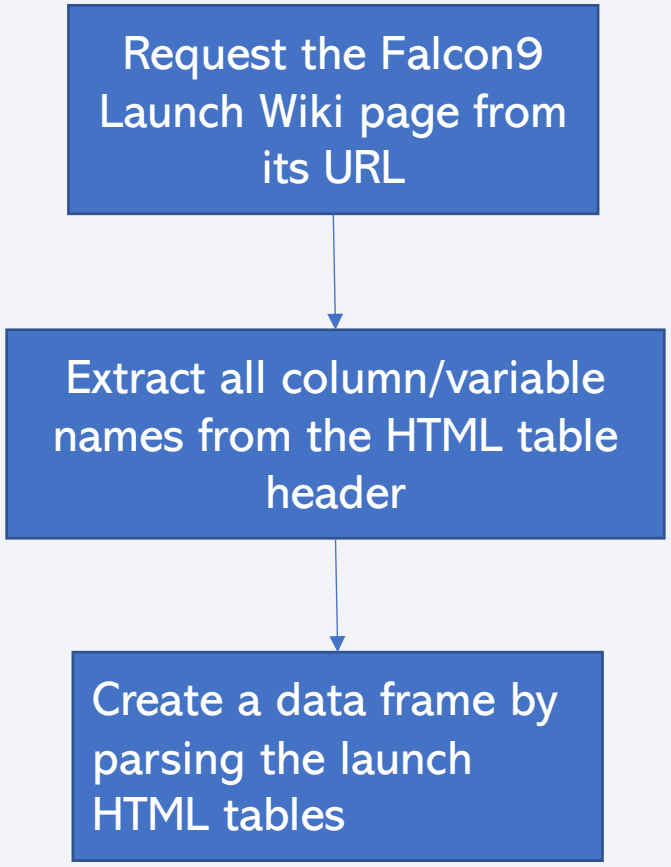
Data Collection – SpaceX API

- SpaceX provides a public API via which data may be accessed and utilized;
- The API was used according to the given flowchart and the data is collected.

Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/Data_Collection_API.ipynb

Data Collection - Scraping

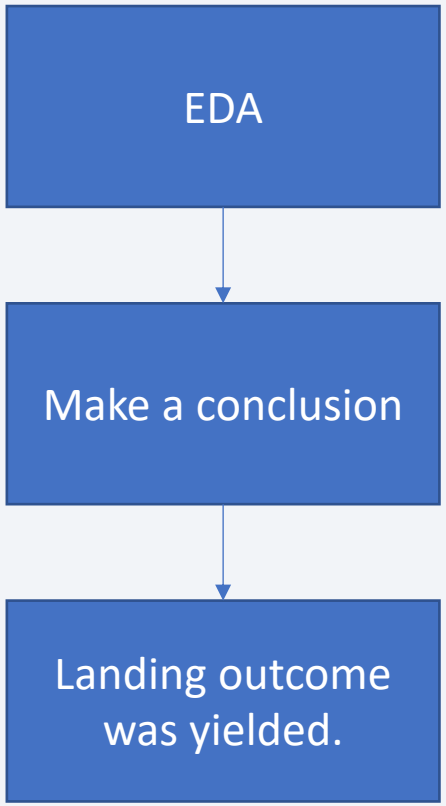
- Space X data was collected from Wikipedia as procedure given in the flowchart.



Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/Data_Collection_with_Web_Scraping.ipynb

Data Wrangling

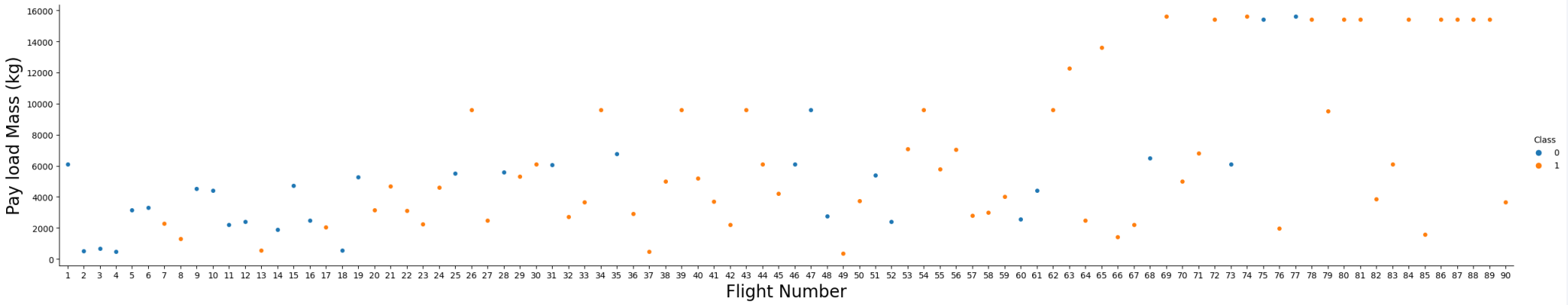
- First, we have explored data through EDA using SQL.
- Then we can summarize launches per site, occurrence of each of orbit and mission outcome per orbit type and, final we obtained the landing outcome as outcome label.



Source code : https://colab.research.google.com/github/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/Data_Wrangling.ipynb

EDA with Data Visualization

- The plot of the FlightNumber vs. PayloadMassand overlay the outcome of the launch. We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.



Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/EDA_with_Data_Visualization.ipynb

EDA with SQL

- The task of SQL queries were as the following:
  - Names of the distinct launch locations for the space mission should be displayed.
  - Show five entries where the string "CCA" appears at the start of launch sites.
  - Show the total tonnage of payload carried by all of NASA's booster launches (CRS)
  - Show the average mass of payload that the booster version F9 v1.1 can carry.
  - Give the date of the first successful landing that took place on the ground pad.
  - Enumerate the launchers' names that have been successful in drone ship applications and whose payload mass is more than 4,000 but less than 6,000.
  - Total the number of mission outcomes that were successful and unsuccessful.
  - Put the names of the booster models that have transported the most payload mass in order.
  - List the data for the months of 2015 that will show the launch location, booster versions, month names, and failed landing results in the drone ship.
  - Sort the number of landing outcomes (e.g., ground pad success or drone ship failure) between 2010-06-04 and 2017-03-20 in decreasing order.

Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/EDAwithSQL.ipynb

Build an Interactive Map with Folium

- Map objects such as markers cluster, circles, lines added to a folium map ;
  - Mark all launch sites on a map to point out a launch sites.
  - Mark the success/failed launches for each site on the map to indicate which locations/positions are success or failed.
  - The distances between a launch site to its proximities to point the distance between two coordinates.

Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/Interactive_Visual_Analytics_with_Folium_lab.ipynb

Build a Dashboard with Plotly Dash

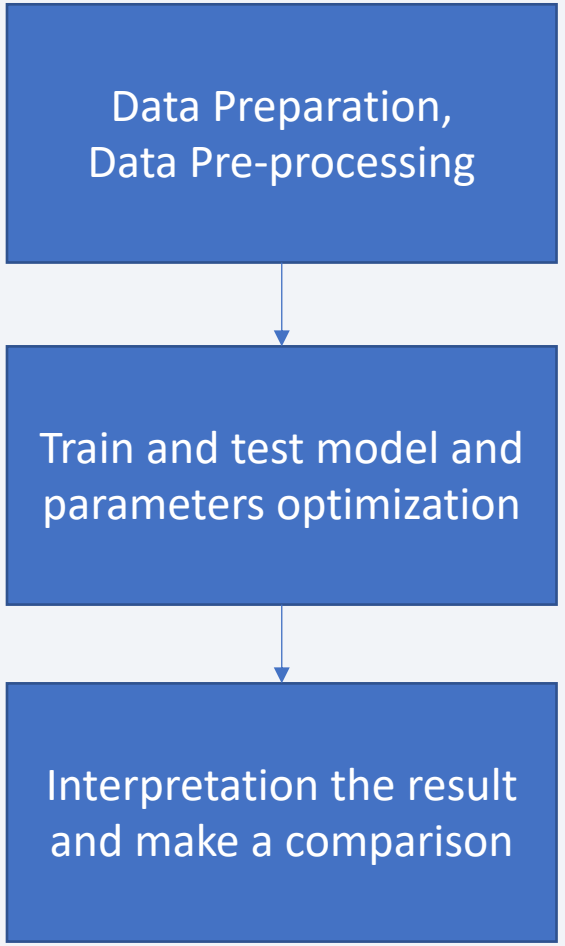
- We calculate the percentage of launches by site and payload range by using Plotly Dash.
- The combination of two graphs allowed the analytic task to perform easily and help identifying the best place to launch.

Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Four machine learning model that were used to classify were performed and compared the result ; logistic regression, support vector machine, decision tee and k nearest neighbors.

Source code : https://github.com/cianyu12/Applied-Data-Science-Capstone-IBM/blob/main/Machine_Learning_Prediction.ipynb



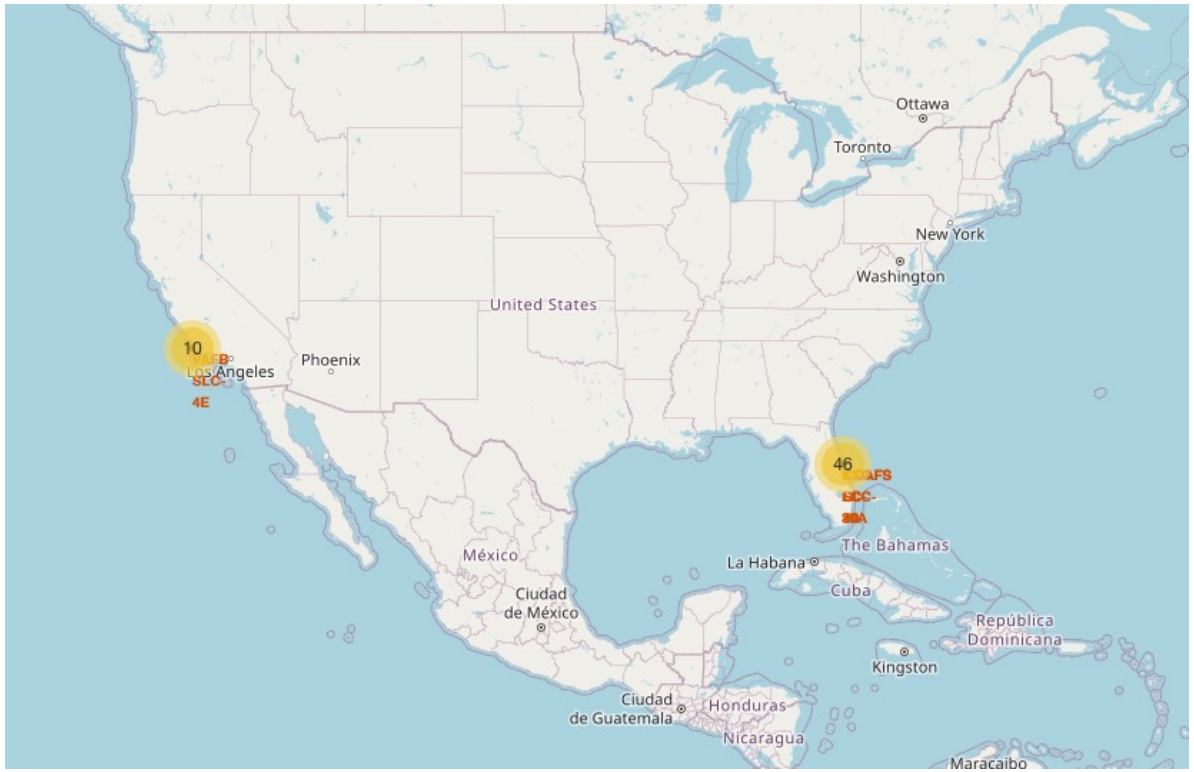
Results

- Results of the exploratory data analysis show that:
  - Space X uses four different launch sites.
  - NASA and Space X conducted the first launches.
  - The average payload of the F9 v1.1 booster is 2,928 kg.
  - The first successful landing occurred in 2015, five years after the first launch.
  - Almost all mission outcomes were successful
  - Two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, failed to land in drone ships in 2015
  - The number of landing outcomes improved over time.

Results

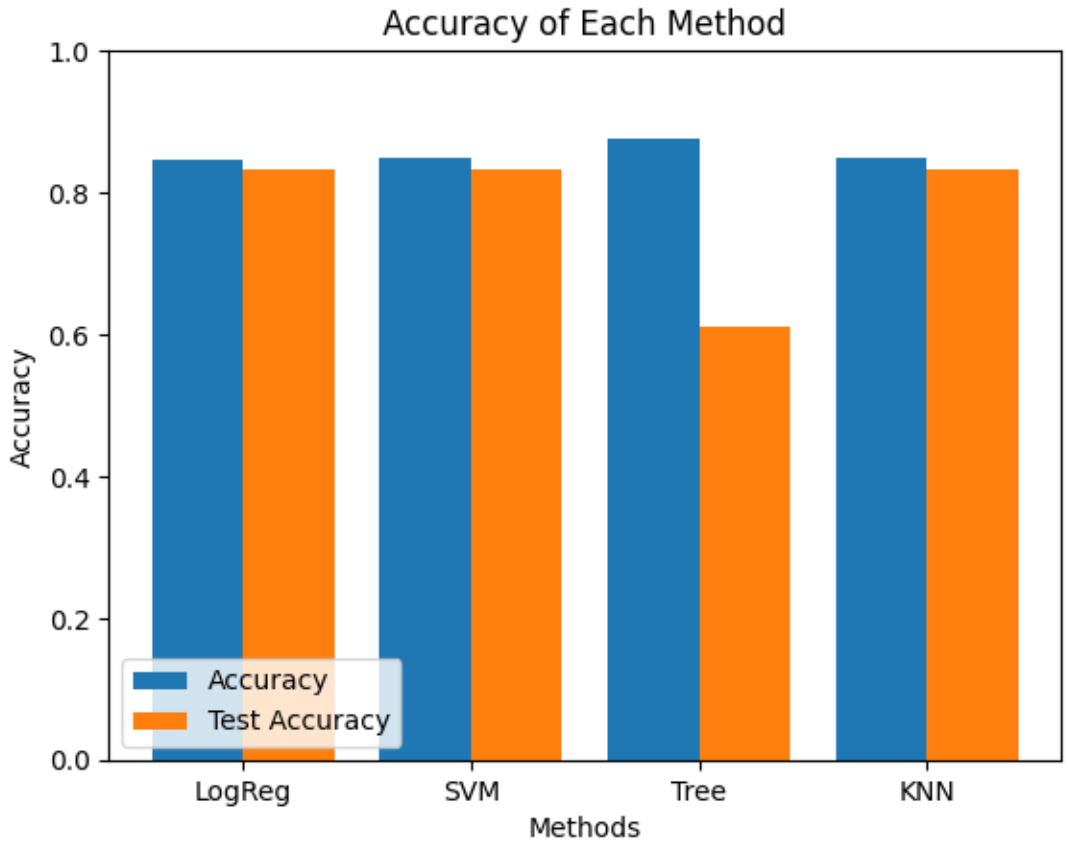
It was possible to determine through interactive analytics that launch sites historically had strong logistical support and were located in secure areas-near the sea, for example.

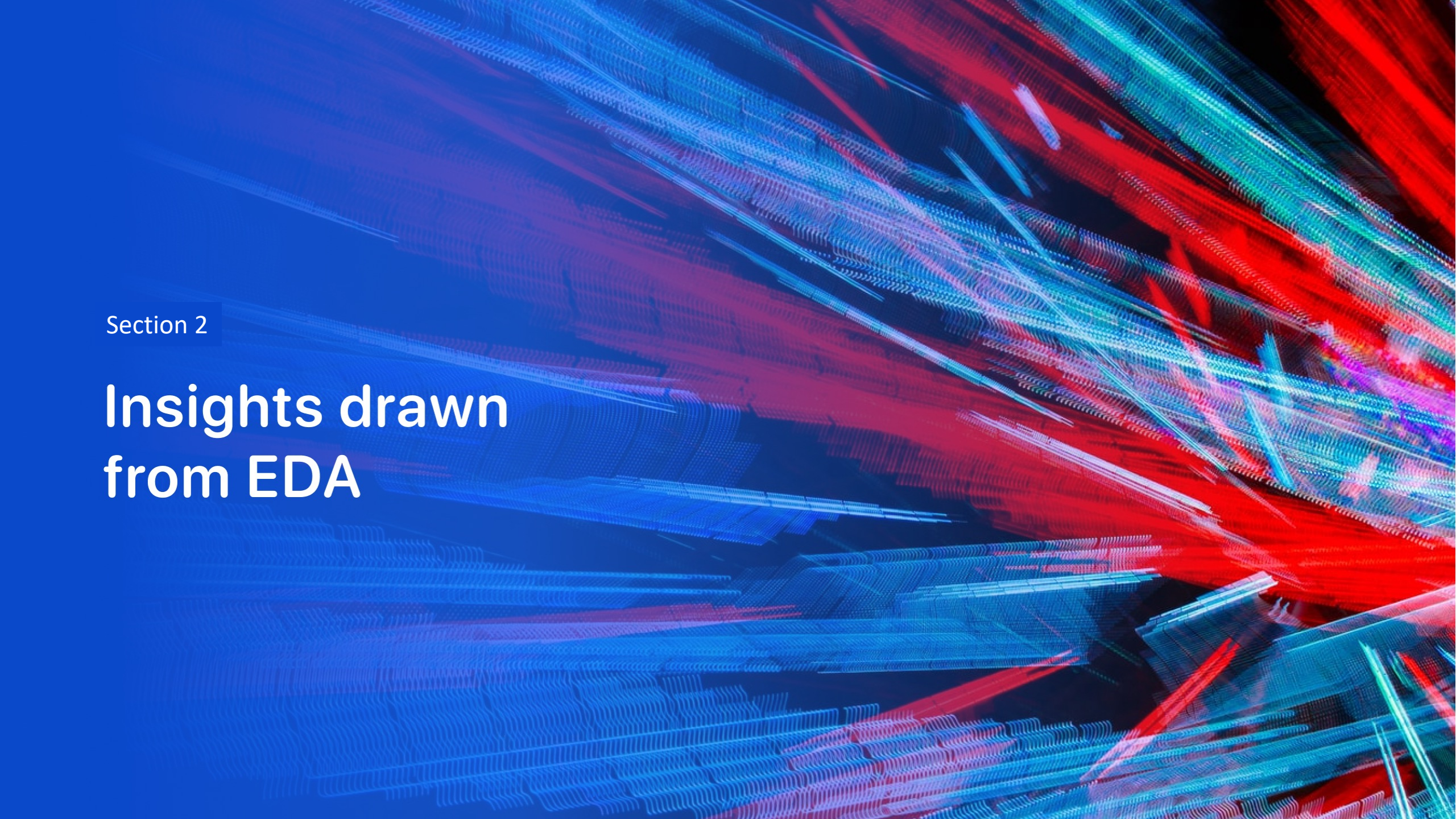
• East Coast launch sites host the majority of launches.



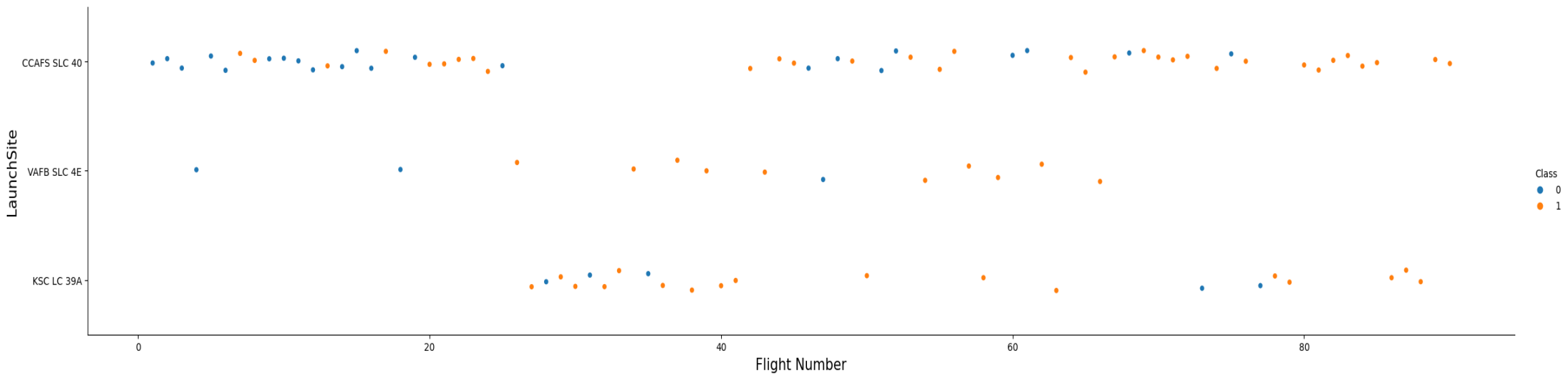
Results

The outcome of three machine learning models provided the similar result which are logistic regression, support vector machine and KNN having accuracy in training and test dataset around 83%. However, the accuracy in train dataset for Decision tree model overcome another models but the test accuracy is not good enough.



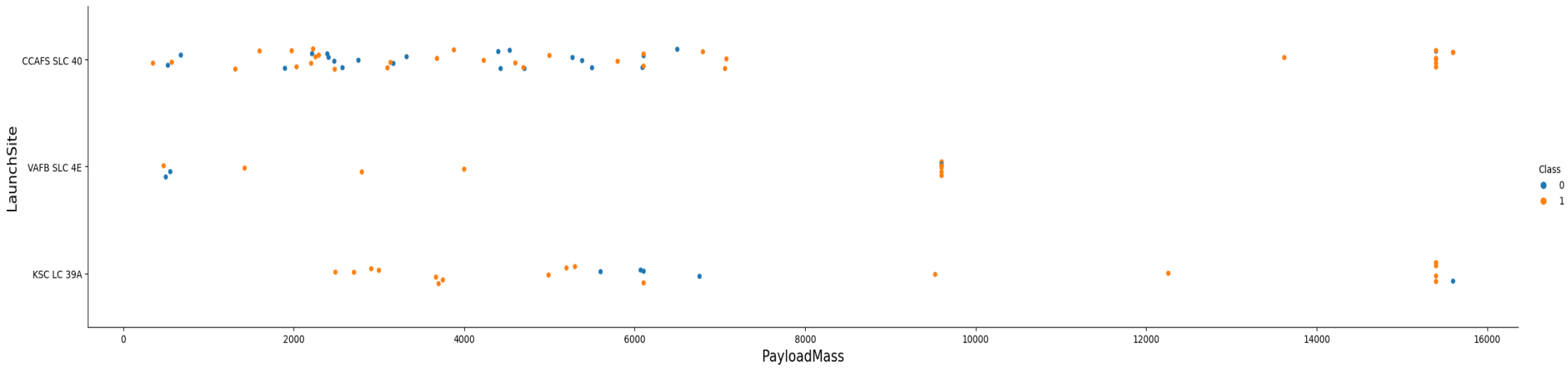

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



CCAF5 SLC 40 was the most successful launch site and shown the improvement over time came along with KSC LC39A and BAFB SLC4E.

Payload vs. Launch Site

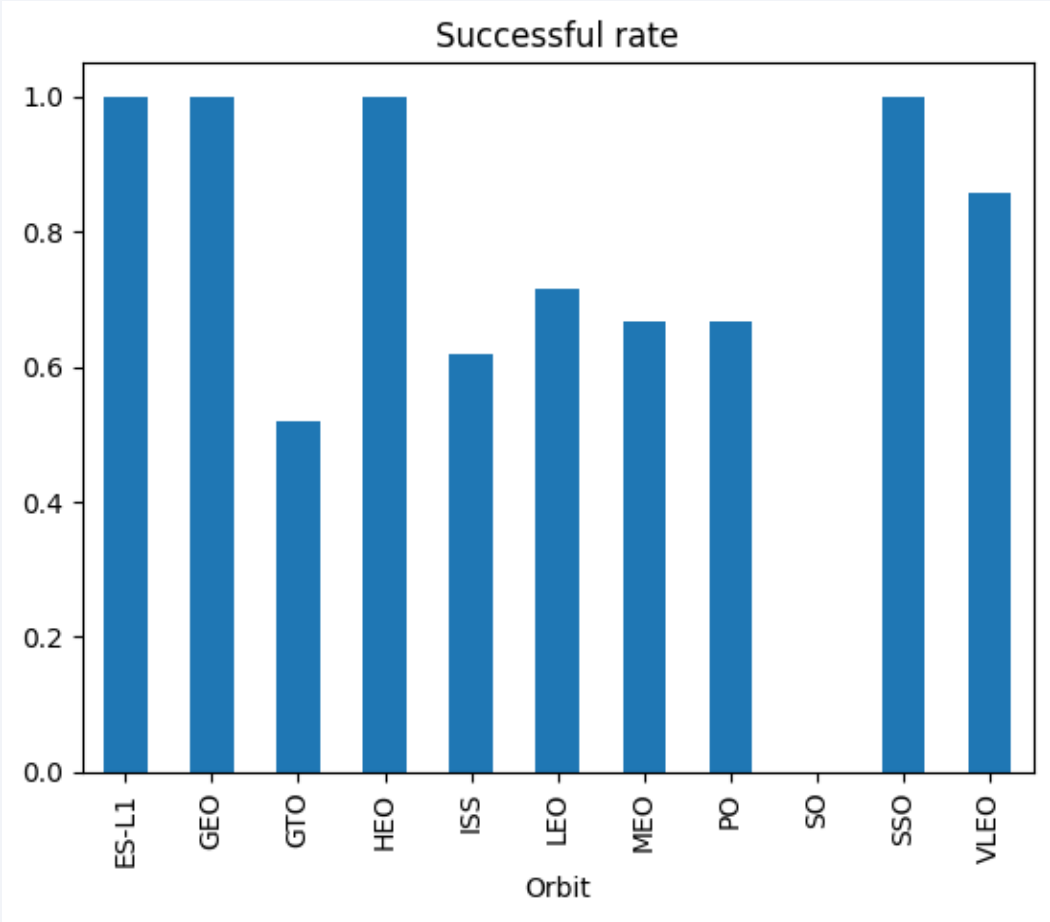


We noticed that the higher PayloadMass, the more chance to be successful launch (over 9,000 kg).

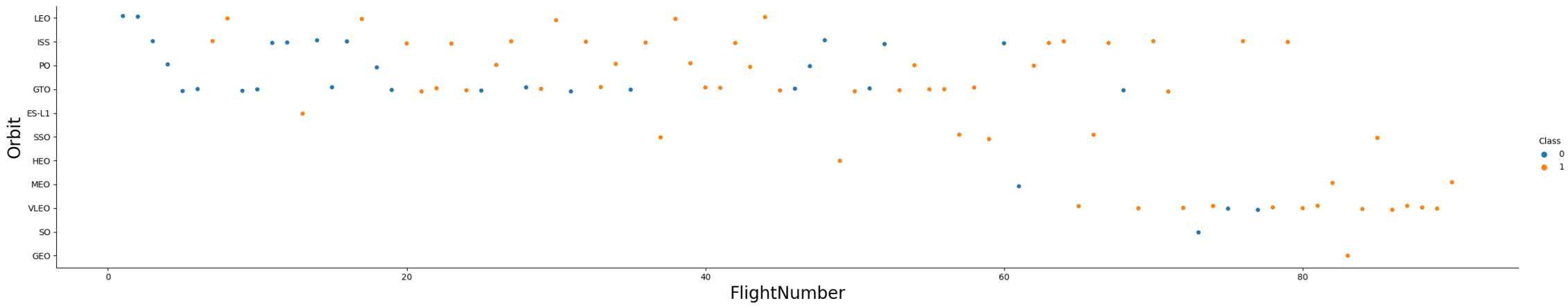
In CCAFS SLC40 with high Payload seemed to have 100% success rate.

Success Rate vs. Orbit Type

- The most successful rate of Orbits were ; ES-L1 GEO, HEO, and SSO.
- The least successful rate of Orbits were GTO.

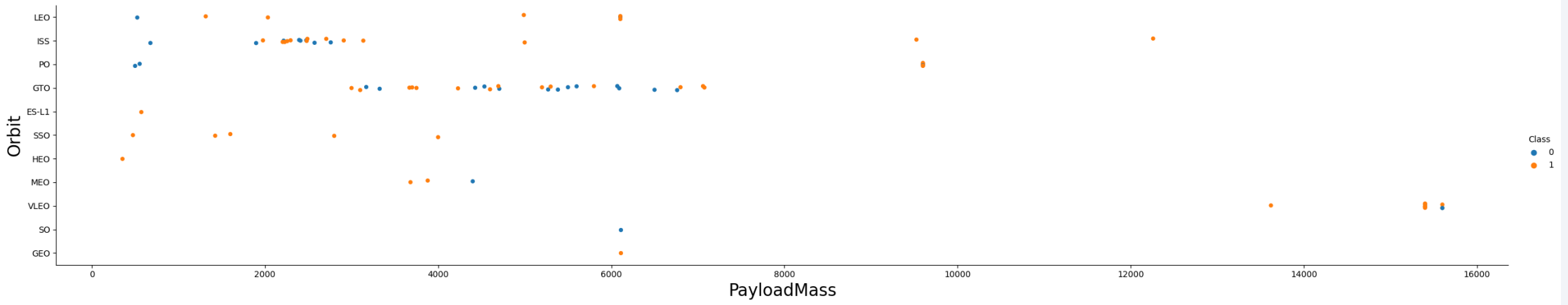


Flight Number vs. Orbit Type



- The increasing of number of flight, the more successful rate we found.
- VELO orbit hit a high successful rate.

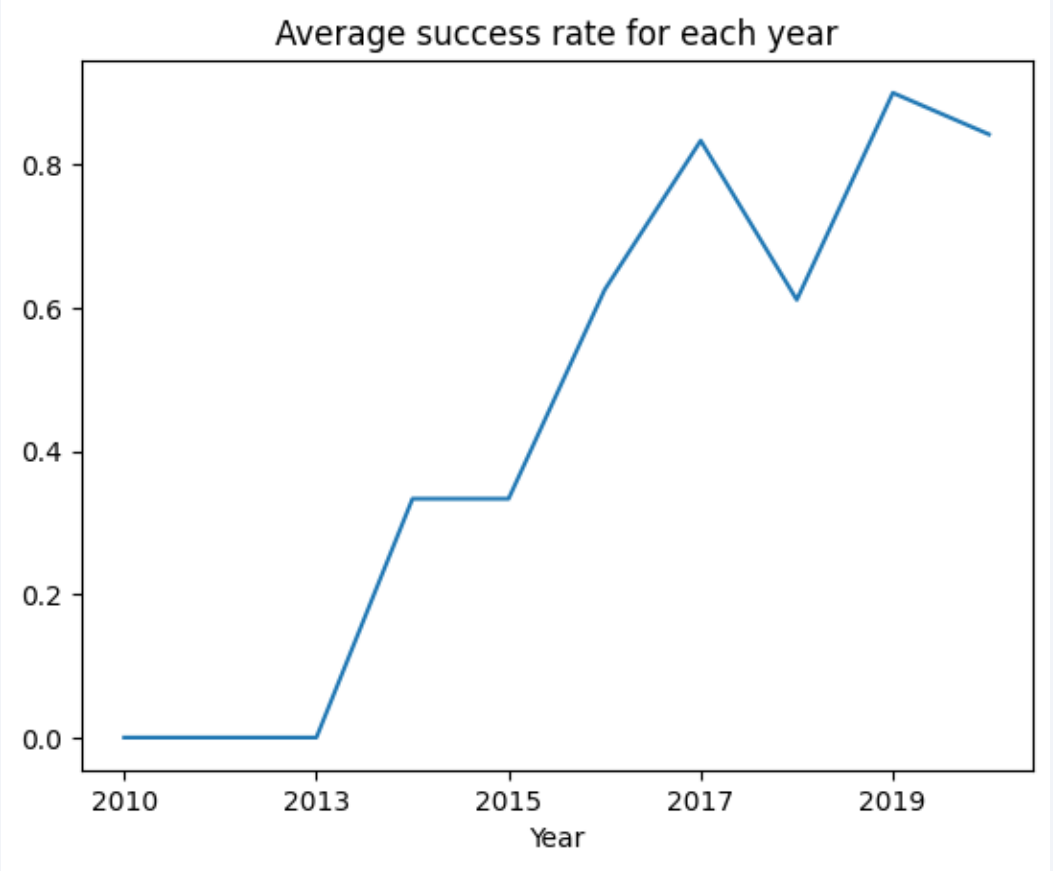
Payload vs. Orbit Type



- SSO showed 100% successful rate over any Payloads.
- The relationship between Payload and Orbit GTO was not show any information.

Launch Success Yearly Trend

- The average success rate for launch show an up forward trend since 2013.



All Launch Site Names

- There are four launch site as the following;

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

- They are obtained by using distinct to the launch_site column.

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_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
2010-12-08	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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

- The like-operation in SQL was used to find the launch sites begin with CCA as 'CCA%'.

Total Payload Mass

- The total payload carried by boosters from NASA (CRS).

TOTAL_PAYLOAD
111268

- The sum function is applied to payload_mass_column and using boosters from NASA(CRS), then we find CRS in payload column using like operation.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1.

AVG_PAYLOAD
2928.4

- The AVG function was used to calculate the mean of payload mass from booster version using where equal to F9 v1.1.

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad.

FIRST_SUCCESS_GP
2015-12-22

- The min function was used to find the first date successful and landing_outcome was set to ‘Success (ground pad)’.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

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

- The distinct function was applied to find the name of boosters that successfully landed on landing_outcome is drone ship and payload mass use between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes.

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

- The count function is used to count the number of successful or failure that group by mission outcomes.

Boosters Carried Maximum Payload

- List the names of the 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
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The subquery technique was used to find the maximum payload then we use distinct of booster_version and set the payload equal to the payload in subquery.

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- We select month by using substr function on Date and set Landing outcome to Failure drone ship and substr date to get year equal to 2015.

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

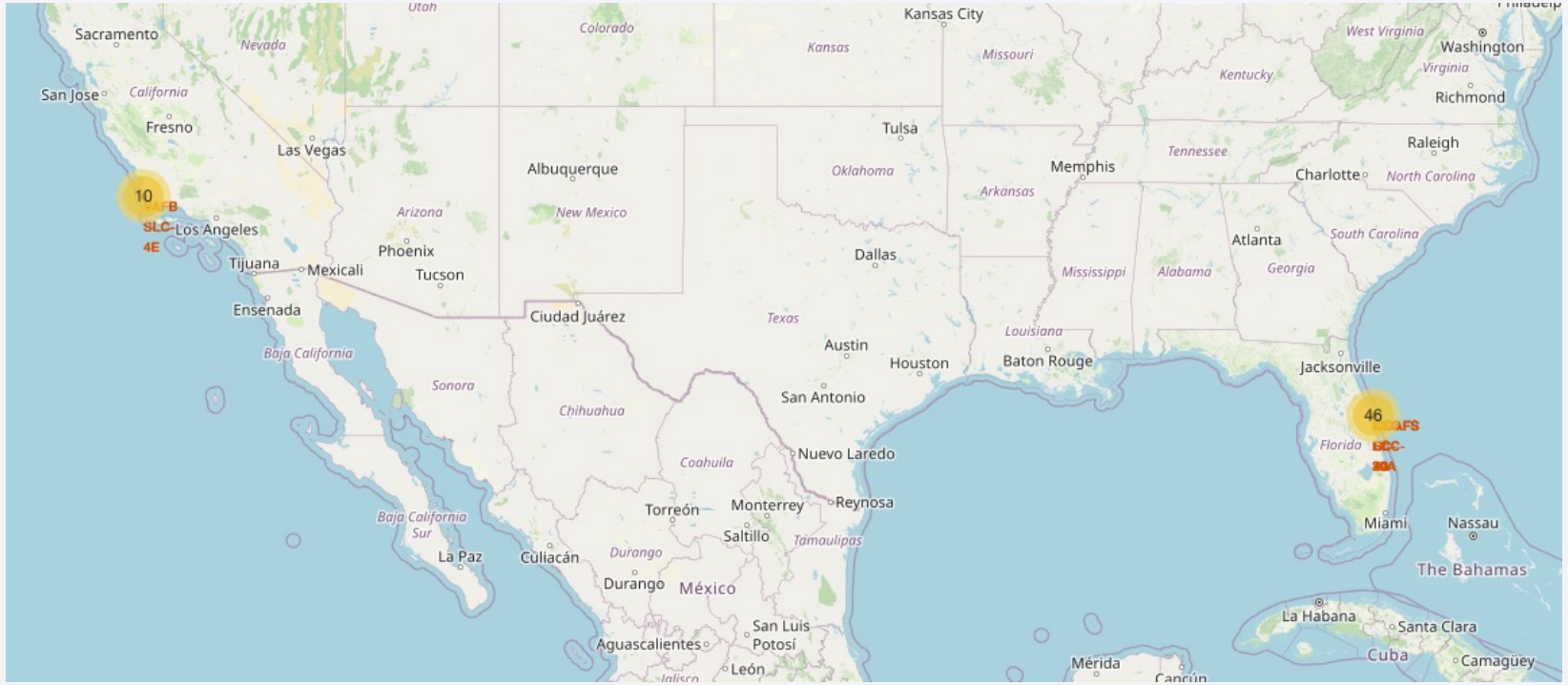
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

- We applied count function to count landing_outcome that happened between date 2010-06-04 and 2017-03-20 and used group by landing_outcome to count and order by to sort the count of landing outcomes.

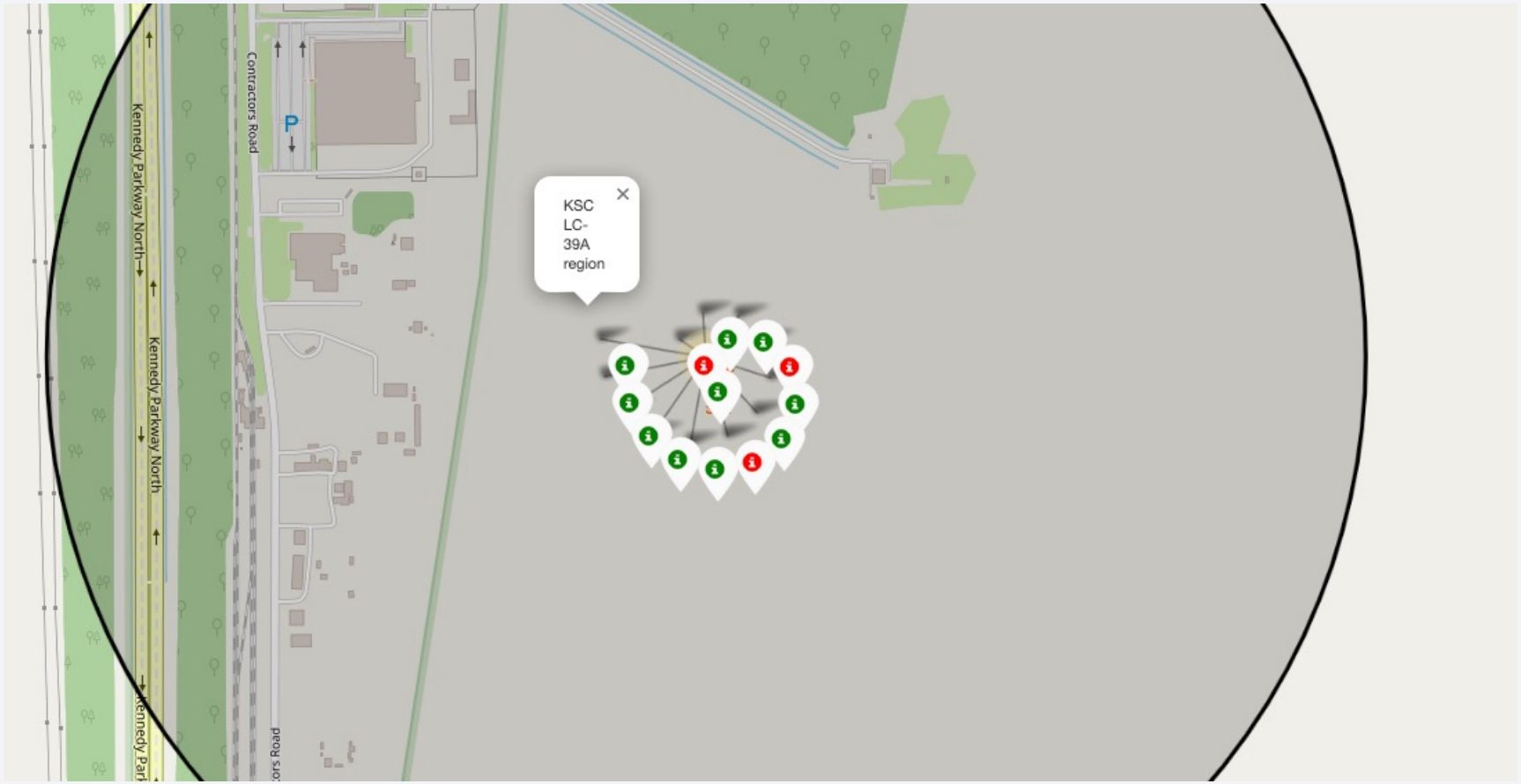

Section 3
Launch Sites
Proximities Analysis

All launch sites with number of launches



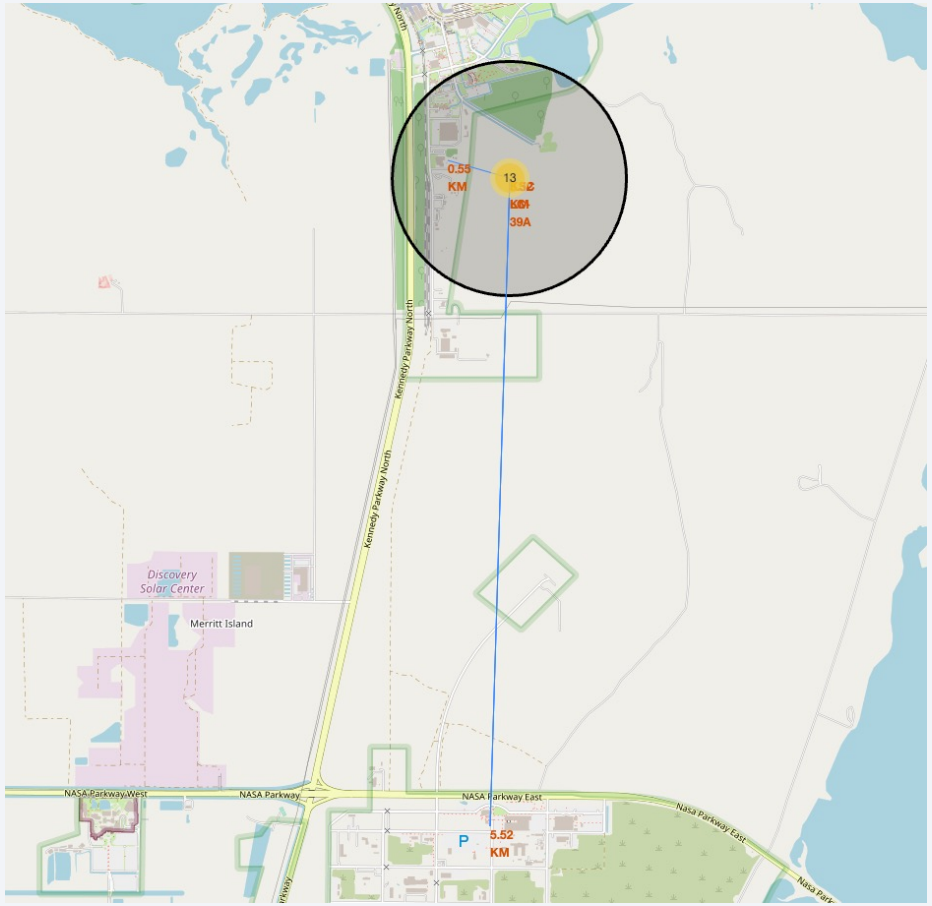
- All launch sites are close to the sea but the most found in east coast.

KSCLC-39A region

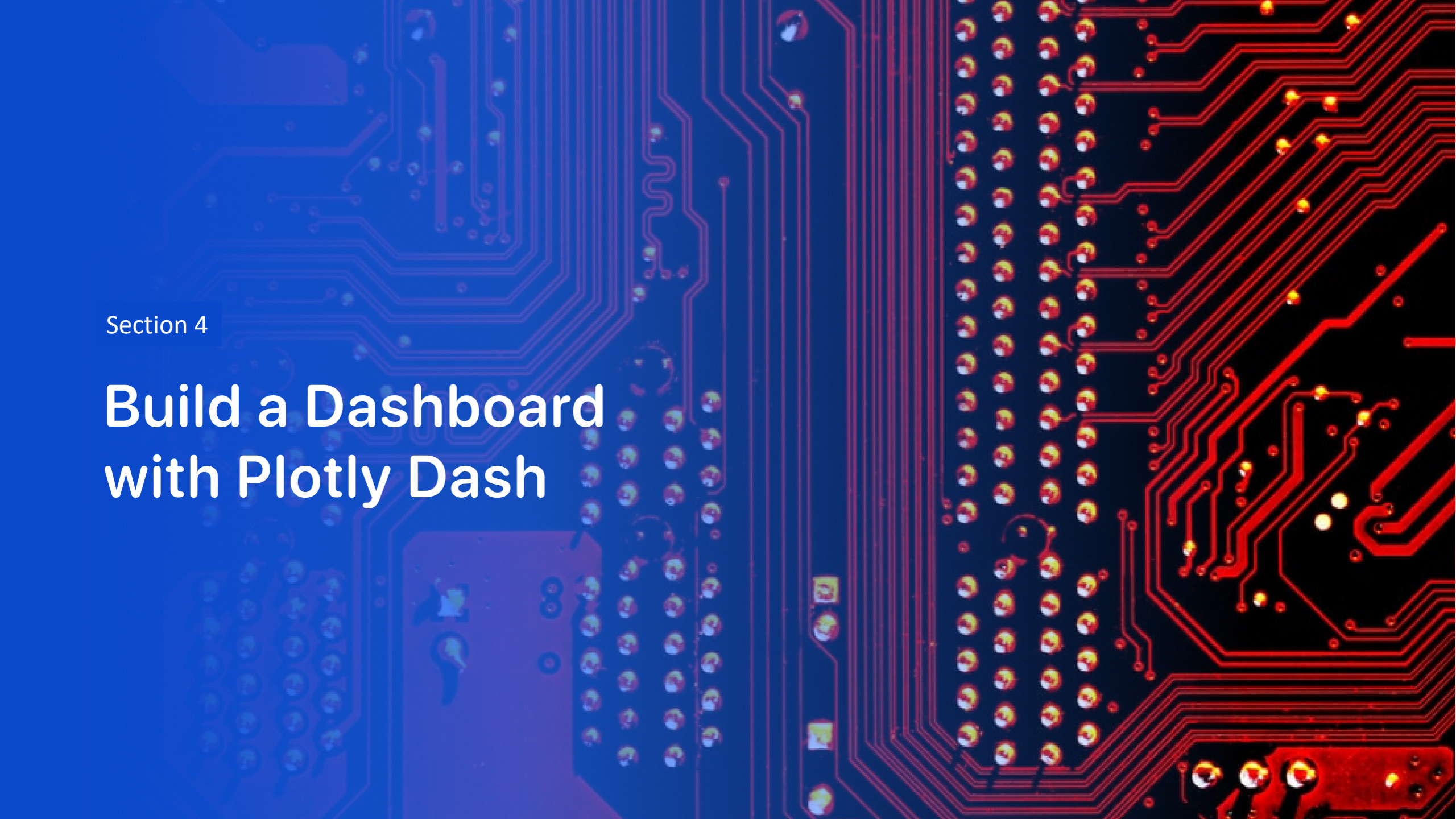


We found that KSCLC-39A region is the most successful launch site with high rate.

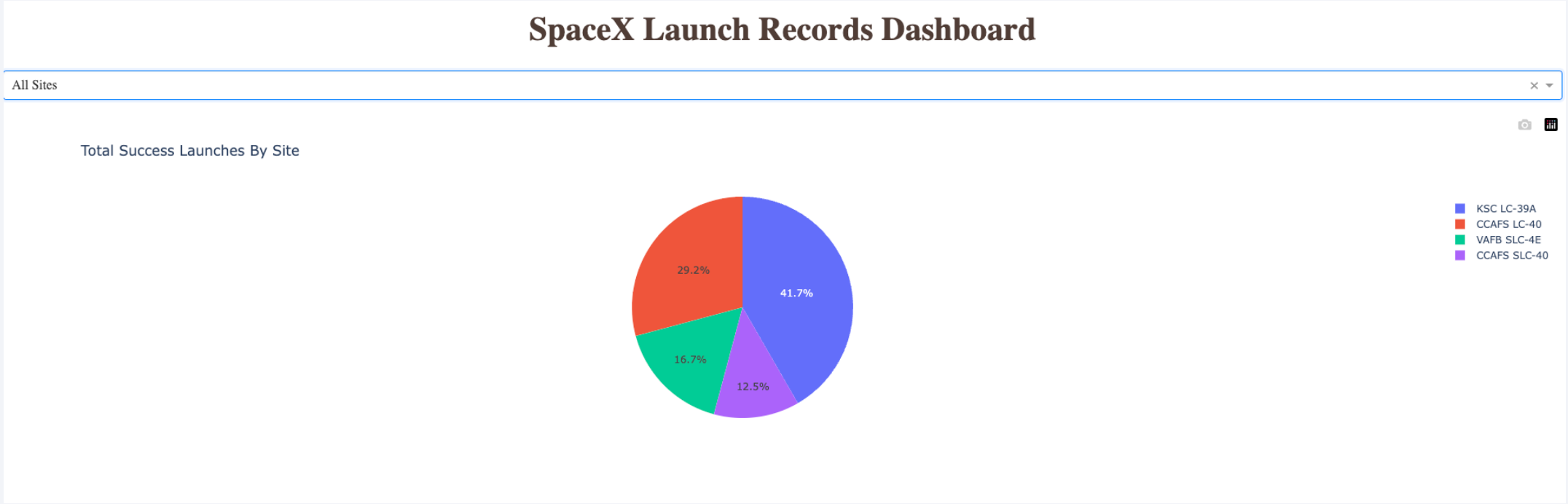
The railways at the best launch site



KSC LC-39A is a good location aspects, that close to the road and railways.

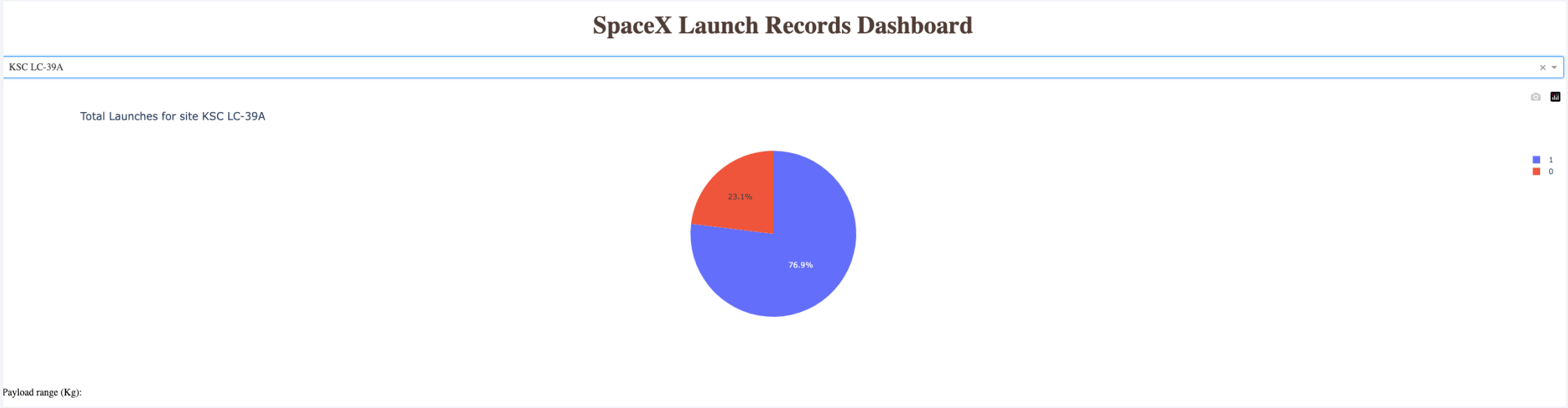

Section 4
Build a Dashboard with Plotly Dash

The proportion of success rate by site



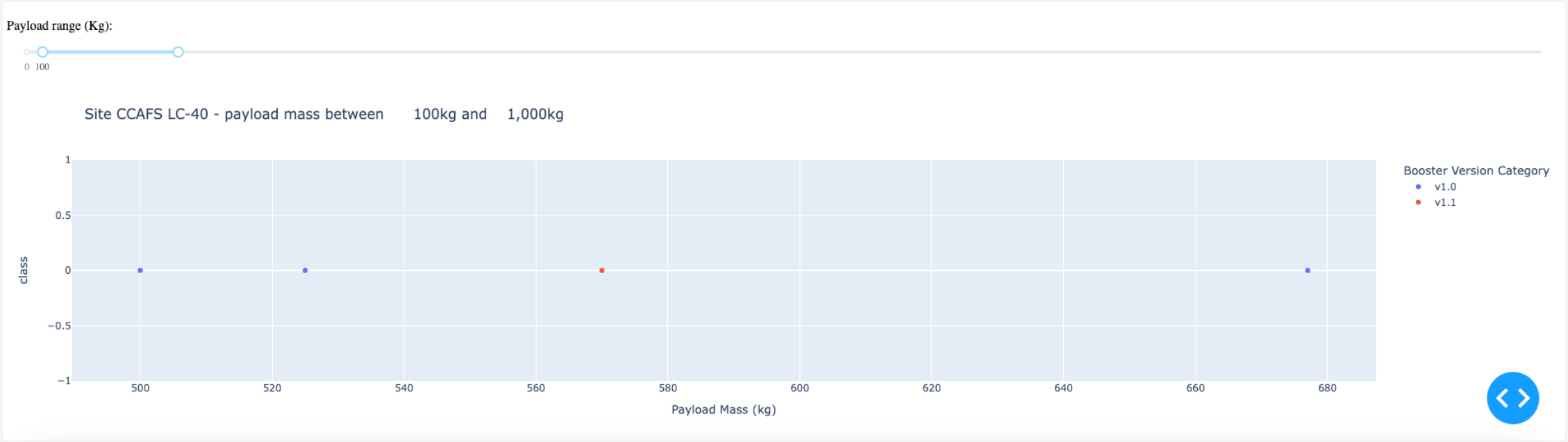
The site play a crucial role in succession of launches. For example, KSC show a high success rate more than CCA 3x.

Success rate for KSC LC-39A



It has success rate of launch is 76.9%

Payload vs Launch Outcome



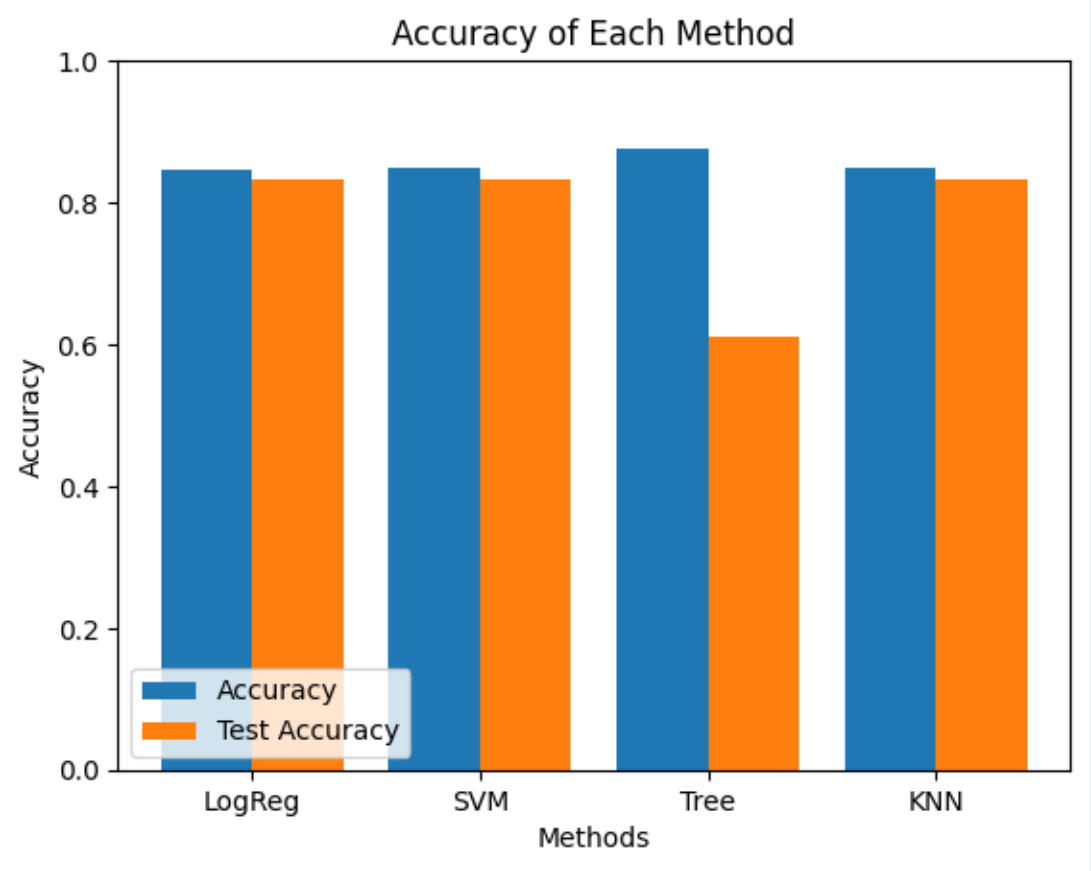
We found that 0% success rate on Payload less than 1,000 kg.

Section 5

Predictive Analysis (Classification)

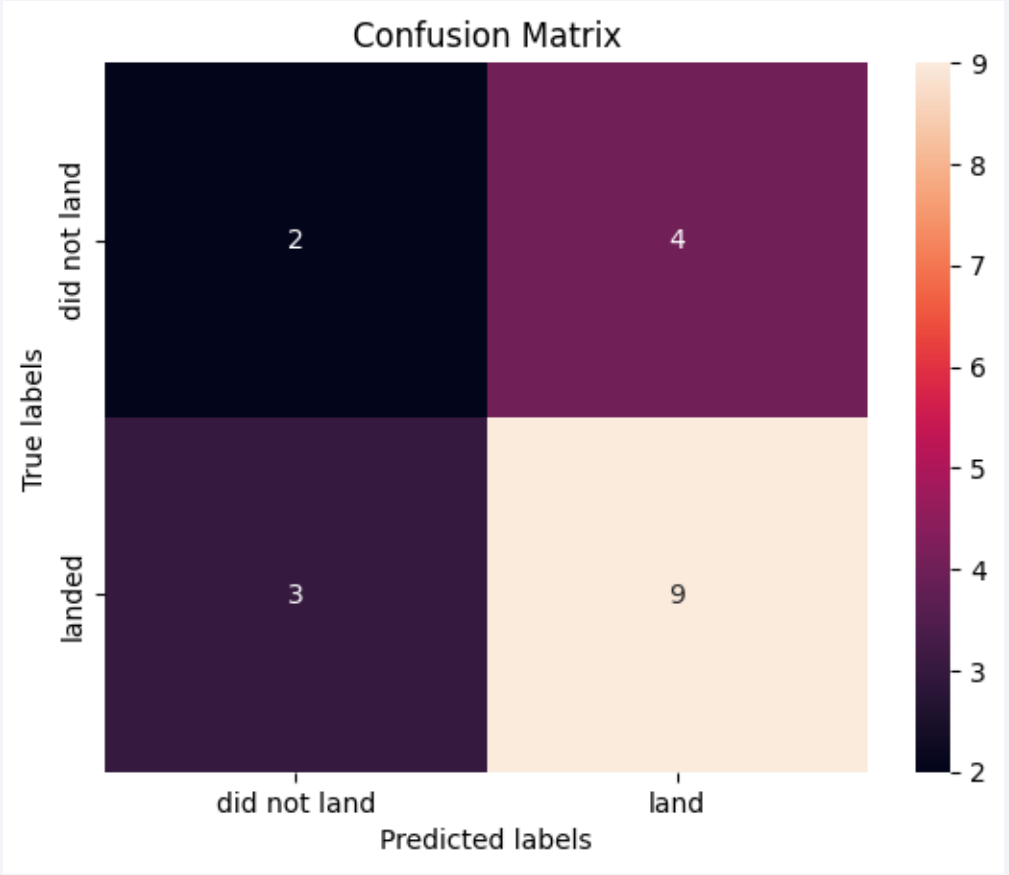
Classification Accuracy

- The 4 machine learning models were evaluated through training and test datasets.
- The model with high accuracy is Decision Tree with over 87%



Confusion Matrix

Confustion matrix of Decision Tree prove its accuracy by showing a high true posive. However, the true negative still small that needs to improve the model further.



Conclusions

- We can access to many sources via API and Web resource to gain more information to perform analysis.
- The best launch site is KSC LC-39A with high success rate.
- The lower Payload show a potential to not success in launch.
- The time also involved regarding to the success rate improves over time.
- The machine learning plays a crucial part to predict the outcome that can reduce cost of loss.

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!