DATA SCIENCE CAPSTONE PROJECT

BY: GORAN LAWATI | 08.05.2024 | https://github.com/Goran17l

OUTLINE

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

EXECUTIVE SUMMARY

Different methods of data collection was used to gather and analyze data from the public SpaceX API and SpaceX page:

- *Data Collection using web scraping*
- *Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics*
- *Machine Learning (Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors)*

Summary:

- *EDA allowed easiest identification of features to help predict the optimal chances of successful launchings*
- *Machine Learning provided the best prediction modelling.*

INTRODUCTION

OBJECTIVES:

- To evaluate the capability of a new Space company (SPACE Y) to compete/ rival with Space X.

- Aiming to find the most pragmatic way to minimise total costs with total budget of $62 million, for launch by using machine learning to predict successful landings during the first stage of rocket launches.

- Finding out what makes the best location to create a launch site.

METHODOLOGY

AN OVERVIEW OF DATA COLLECTION, WRANGLING, VISUALIZATIONS, DASHBOARD, AND MODEL METHODS

METHODOLOGY:

Data collection methodology:

- Data from Space X was acquired from 2 sources:
  - Space X API
  - Web Scraping
- Data wrangling was performed by collecting data and simulating a landing outcome based on the outcome data set, after analysing and summarising the data features.
- Exploratory data analysis (EDA) was conducted using visualizations/SQL.
- Created interactive visual analytics using Folium and Plotly Dash.
- Used classification models to perform predictive analysis:

Normalized data was divided in training/test data sets and assessed into different classification models, with each then being evaluated.

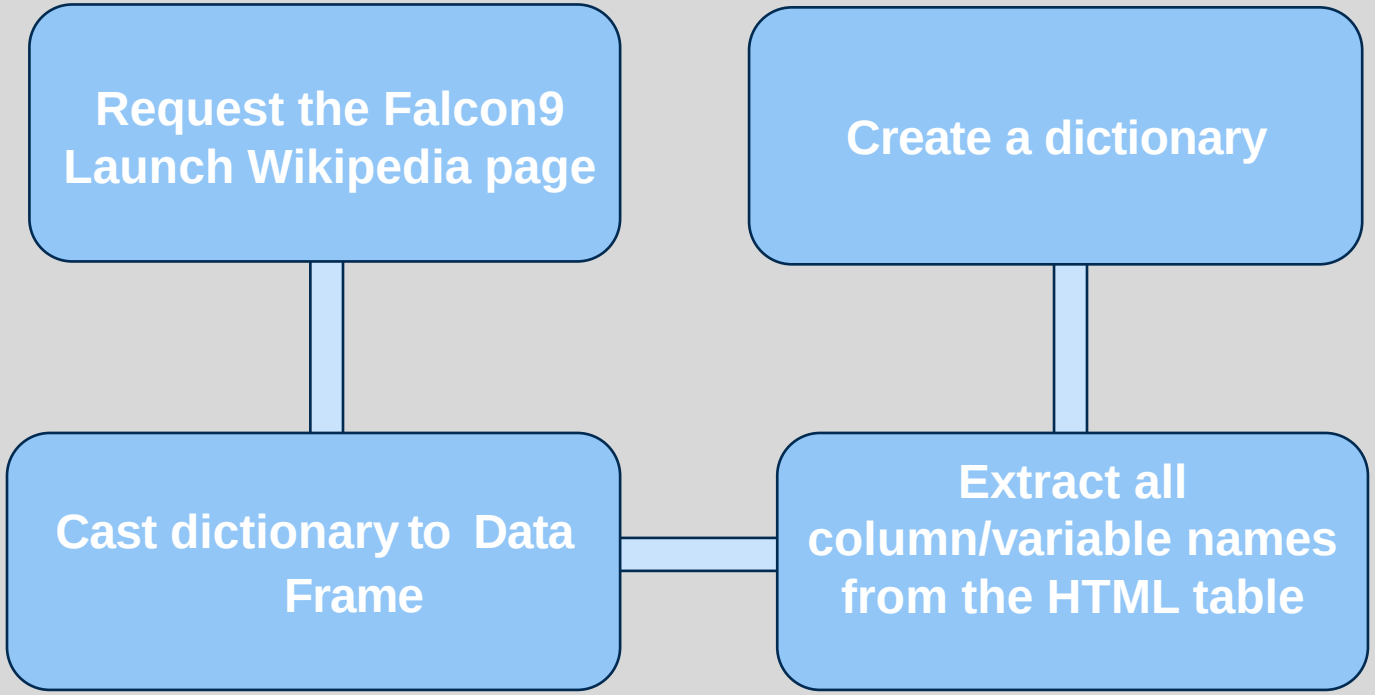
DATA COLLECTION – SpaceX API:

- 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/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/01.%20Data%20Collection/jupyter%20labs%20spacex%20data%20collection%20api.ipynb

DATA COLLECTION – SCRAPING:

- This data about SpaceX was obtained from Wikipedia.
- Data is downloaded, and parsed, like the basic flowchart shows:



Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/01.%20Data%20Collection/jupyter%20labs%20webscraping.ipynb

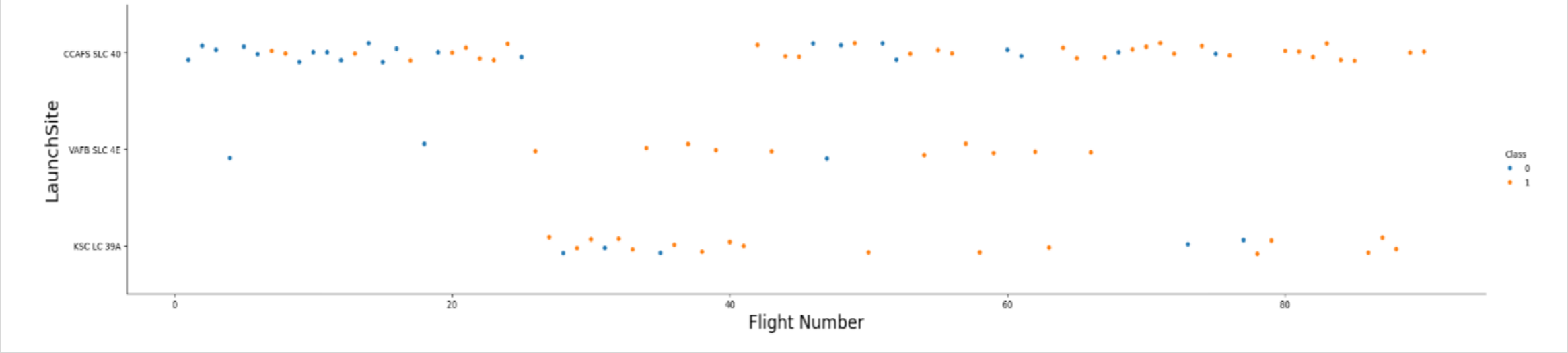
DATA WRANGLING

- Exploratory Data Analysis (EDA) was performed on the dataset: '*Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.*'
- This was used to create summaries and calculations of launches per site, orbits/ orbit type.
- This was then used to create an outcome column.
- Outcome column was used to decide if there was a relationship, and if there was a relationship, it could be used in machine learning.

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/02.%20Data%20Wrangling/EDA%20with%20Visualization%20Lab.ipynb

EDA with Data Visualisation

- Exploration of data with scatterplots and bar plots to visualise the relationship between variables.
- Graph suggests an increase in success rate over time, this is evident with more orange dots which represents Flight Number which were successful.



Blue = Unsuccessful launch | Orange = Successful launch

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/02.%20Data%20Wrangling/EDA%20with%20Visualization%20Lab.ipynb

EDA With SQL

FOLLOWING SQL QUERIES WHICH WERE PERFORMED:

- Displaying names of the unique launch sites in the space mission.
- Top 5 launch sites whose name begin with the string 'CCA'.
- Total payload mass carried by boosters launched by NASA (CRS).
- Finding the average payload mass carried by booster version F9 v1.1.
- Date of the first successful landing outcome in ground pad.
- Names of the boosters which have success in drone ship and have a payload mass between4000 and 6000 kg.
- Total number of successful and failure mission outcomes.
- Names of the booster versions which have carried the maximum payload mass.
- Finding failed landing outcomes, their booster versions, and launch site names in 2015
- The rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad) between 2010-06-04 and 2017-03-2.

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/02.%20Data%20Wrangling/Complete%20the%20EDA%20with%20SQL.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- Markers, circles, lines and marker clusters were used with Folium Maps.
- It was used to show successful and unsuccessful landings:
  - *Markers indicated points like launch sites.*
  - *Circles were used to showcase areas around specific coordinates, i.e. NASA Johnson Space*
  - *Lines were used to show the distances between two coordinates.*
- Through Folium, and the use of markers, circles, and lines we're able to visualizes successful landings relative to location.

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/03.%20Exploratory%20Data%20Analysis/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

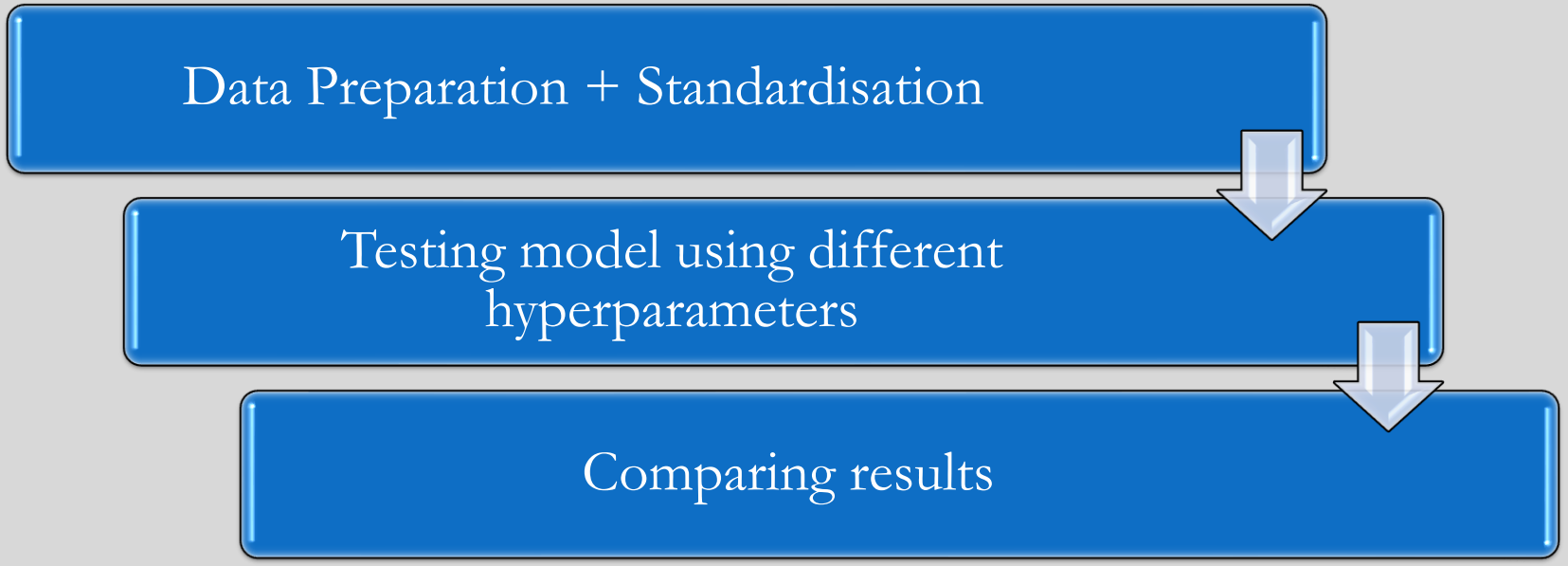
BUILD A DASHBOARD WITH PLOTLY DASH

- Plotly was used to create charts of the SpaceX data.
- A pie chart was used to show the total successful launches count for all sites.
- A scatter chart was also used to show the correlation between payload and launch success.
- This was used to see the relation between payload and launch sites, identifying which was the best location to launch using payloads.

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/03.%20Exploratory%20Data%20Analysis/Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

PREDICTIVE ANALYSIS (CLASSIFICATION)

Classification models used: *Logistic regression, Support vector machine, Decision tree and K nearest neighbours*



Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/04.%20Interactive%20Visual%20Analytics/SpaceX%20Machine%20Learning%20Prediction%20Part%205.ipynb

RESULTS

EXPLORATORY DATA ANALYSIS RESULTS:

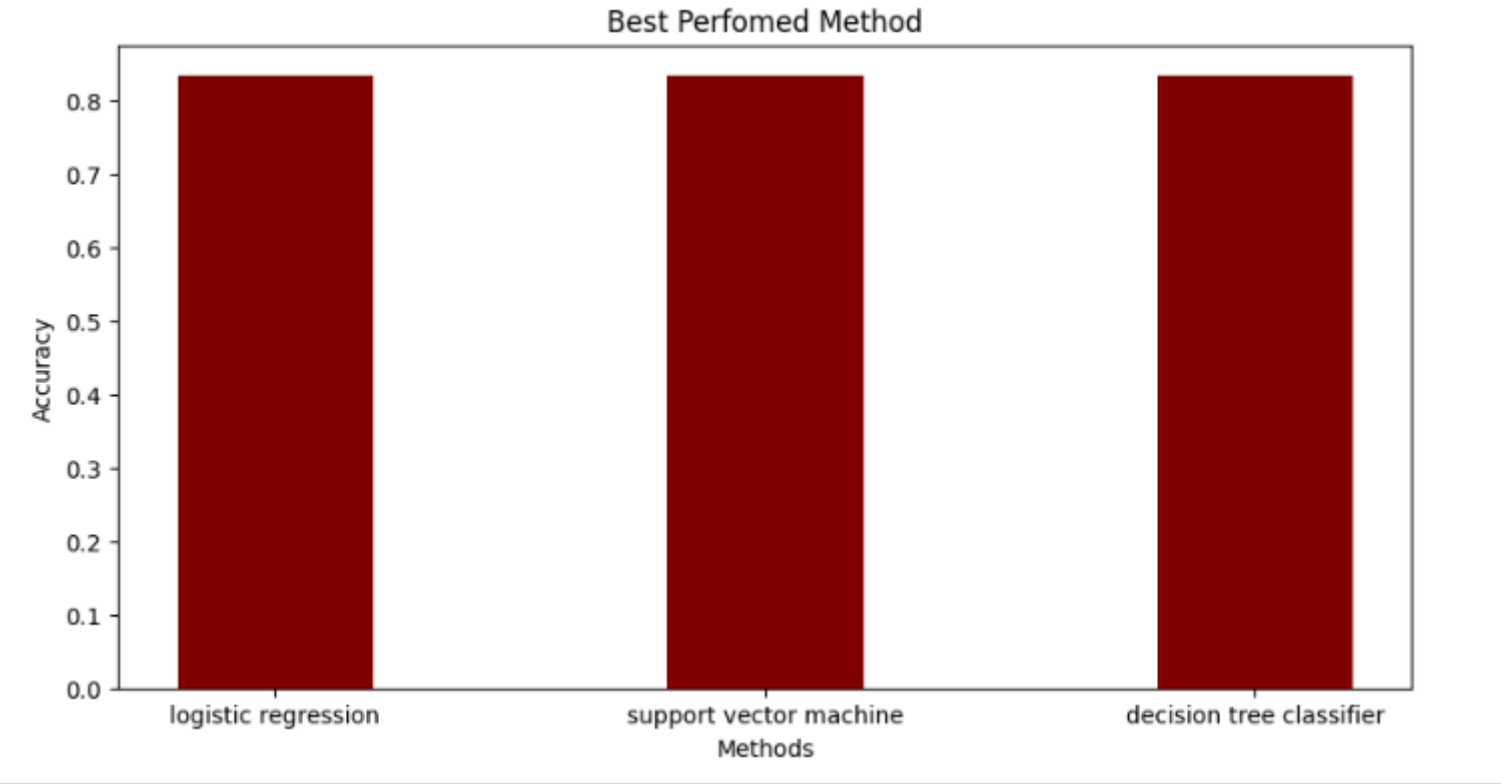
- SpaceX used 4 different launch sites.
- The first launches were done to Space X and NASA.
- AVG payload of F9 v1.1 booster is 2,928 kg.
- First success landing outcome happened in 2015.
- Almost 100% of mission outcomes were successful
- Two booster versions failed at landing in drone ships in 2015.
- The number of landing outcomes became as better as years passed.

RESULTS

- Interactive analytics was used to identify launch sites which where the safest, and closest to the sea.
- Most launches happens near the East Coast, like Miami.

RESULTS

- Predictive Analysis was inaccurate, however attempted.

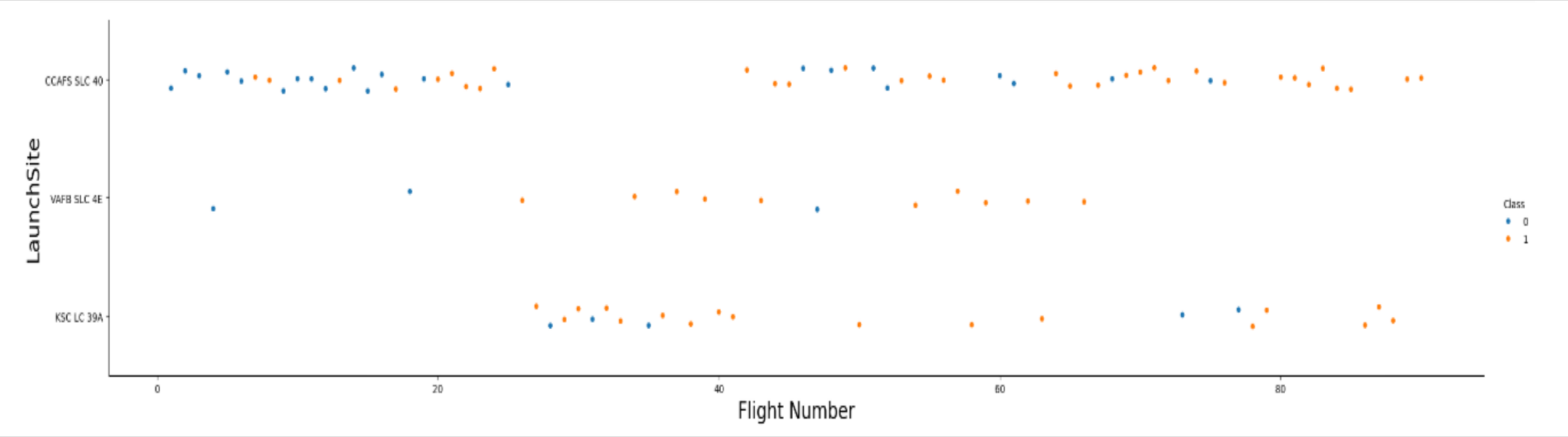


SECTION 2

INSIGHTS DRAWN FROM EDA

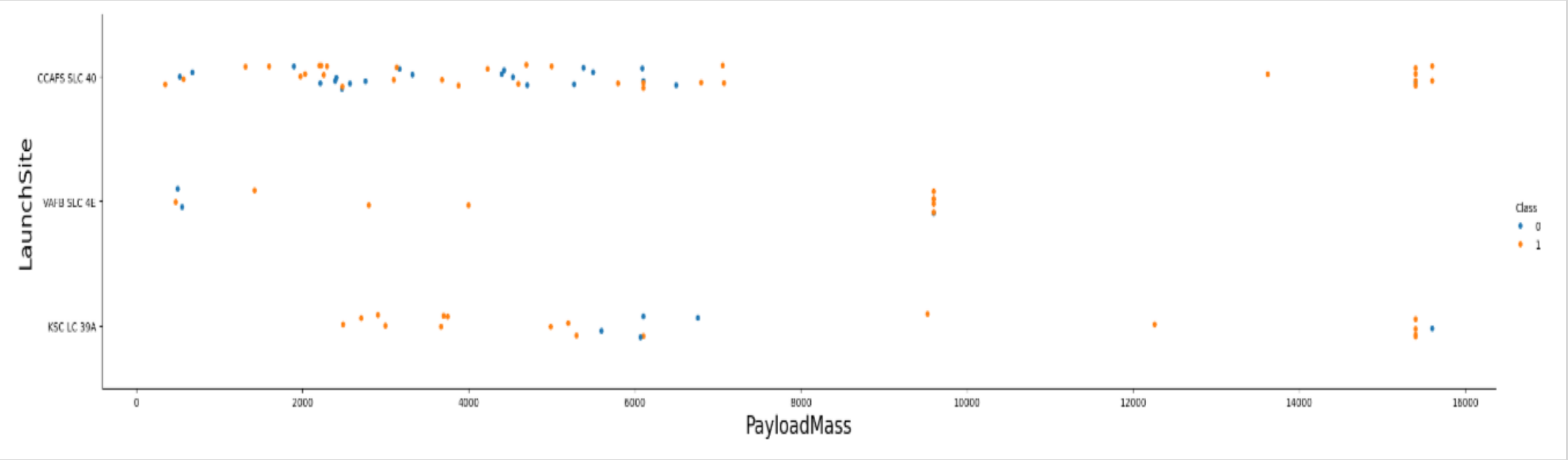
Flight Number vs. Launch Site

- CCAF5 SLC 40 can be considered the best looking at the graph, then VAFB SLC 4E, then at third, KSC LC39A.
- With Orange representing successful launches, you can see the success rate improve with higher frequency of orange dots.



Payload vs. Launch Site

- Orange = Successful launches, with payloads over 9000kg, they have successful launches.
- At 12000 kg, it is still successful but with the most success coming from the CCAFS SLC 40.



Success Rate vs. Orbit Type

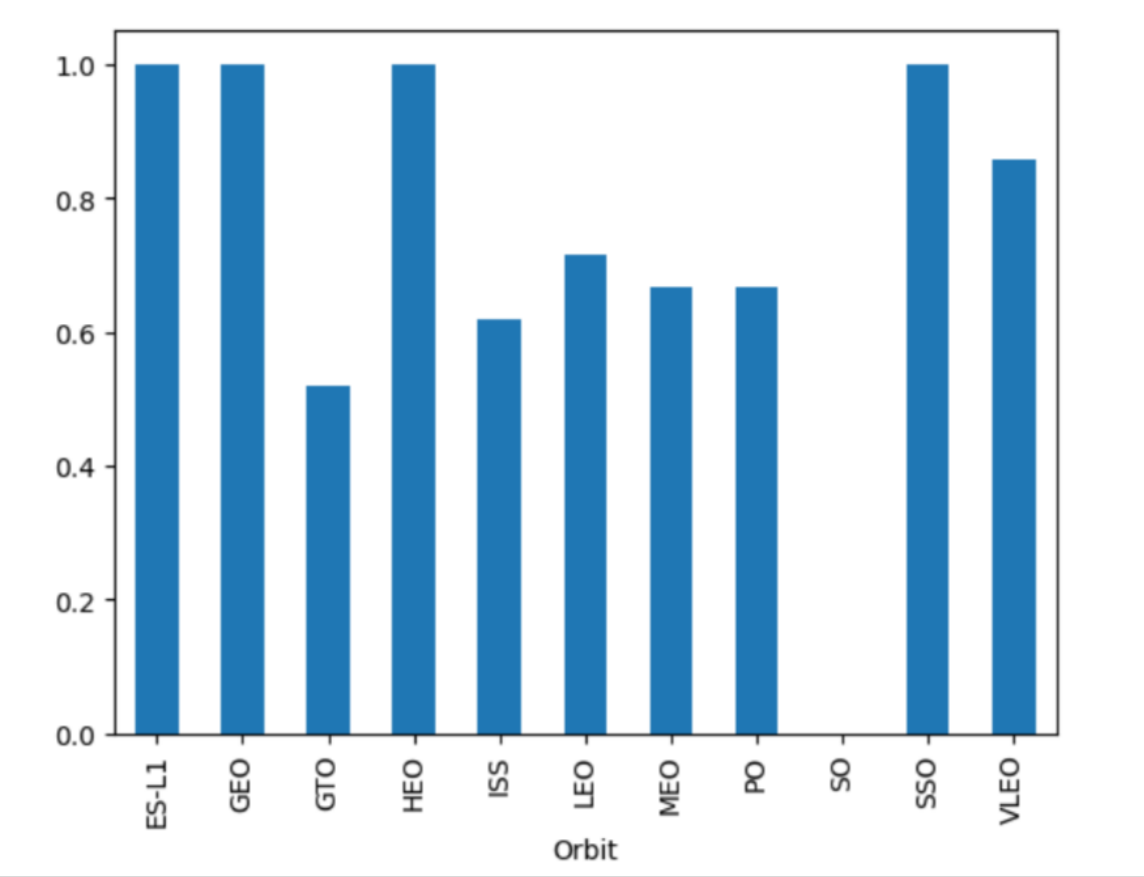
The biggest success rates happens to orbits:

- ES-L1
- GEO
- HEO
- SSO

All with a 1.0

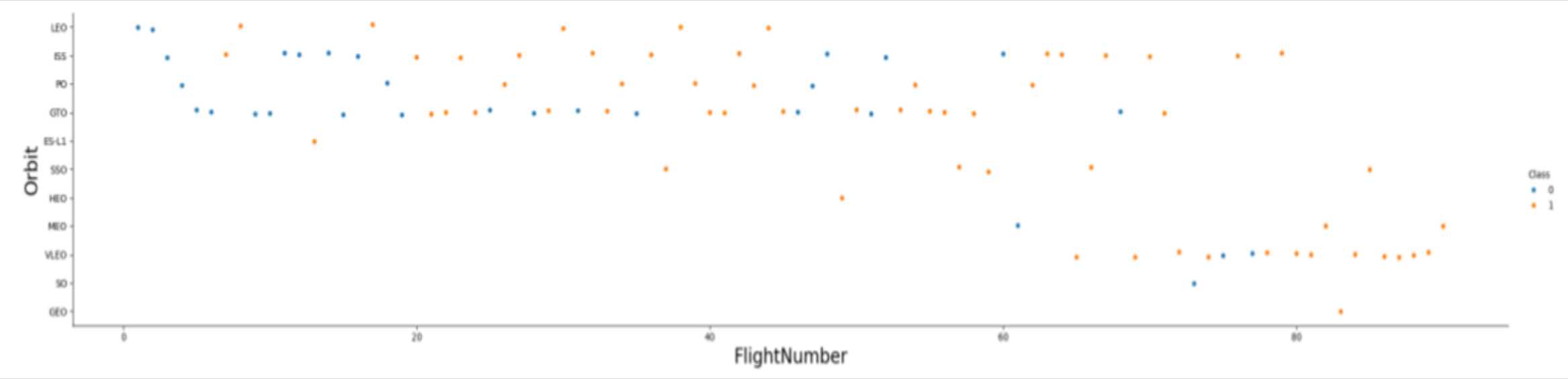
The lowest being SO with 0, then GTO with 0.5, and ISS with 0.6.

The rest fall in the median.



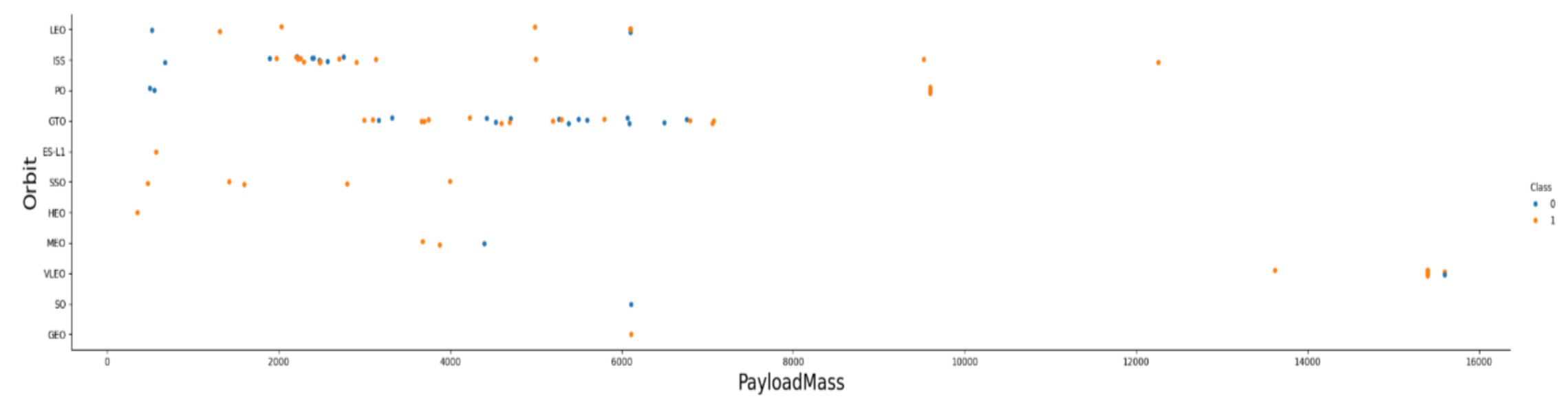
Flight Number vs. Orbit Type

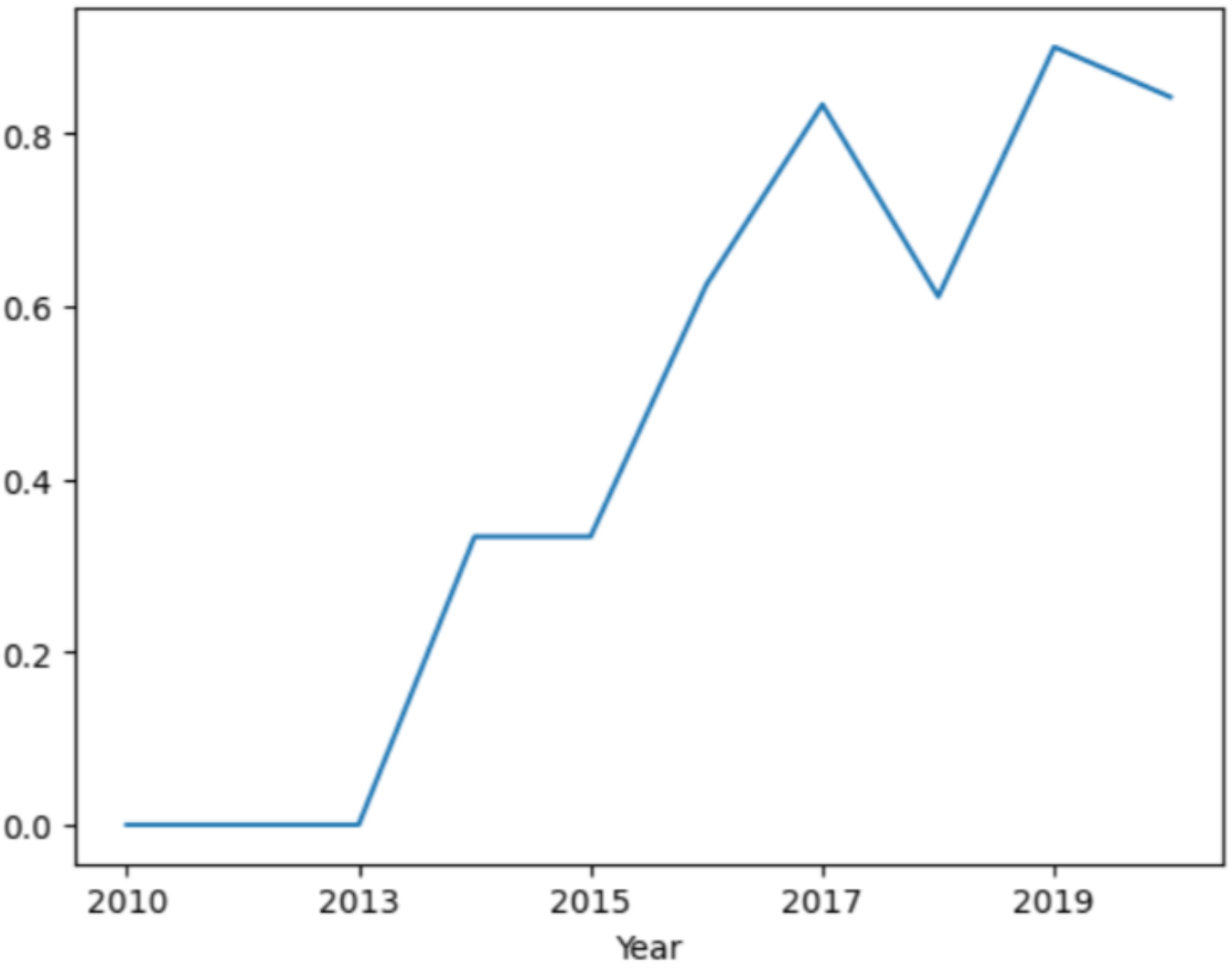
- Success rate is improving, can be justified with orange representing successful orbits.
- VLEO seems to be improving with high increase of successful orbits after 60-100.



Payload Mass vs. Orbit Type

- Little to no correlation between Payload mass and success rate to orbit GTO.
- SO and GEO have the fewest launches, and subsequently launches to the orbit.
- ISS Orbit seems to have the greatest payload in addition to rate of success.





you can observe that the sucess rate since 2013 kept increasing till 2020

Launch Success Yearly Trend

- Success rate starting to increase in 2013 till 2020.
- 2010-2013 can be described as a stagnant period used to starting/ improving technology.

All Launch Site Names

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

- From the gathered data, there seems to be four launch sites:
  - CCAFS LC -40
  - VAFB SLC – 4E
  - KSC LC – 39A
  - CCAFS SLC – 40
- This information was obtained from “launch_site” dataset.

Source Code:
https://github.com/Goran17l/Goran17l/blob/main/10.%20Applied%20Data%20Science%20Capstone/02.%20Data%20Wrangling/Complete%20the%20EDA%20with%20SQL.ipynb

Launch Site Names Begin with 'CCA'

- This shows the 5 records of the launch sites which begins with ‘CCA’
- CCAFS LC -40

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- This query is the total payload mass in Kg carried by NASA (customers).
  - CRS = Commercial Resupply Service

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * sqlite:///my_data1.db
Done.
```

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL \
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

AVERAGE_PAYLOAD_MASS
2928.4

- Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

First successful landing outcome on ground pad:

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL \
    WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
```

FIRST_SUCCESSFUL_GROUND_LANDING
2015-12-22

This query returns the first successful ground pad landing date. The first ground pad landing wasn’t until the end of 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the boosters which have been successful in landing their drone ship while also having a payload mass greater than 4000, but less than 6000.
- Using the filters asked, these are the 4 booter versions.

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

Total Number of Successful and Failure Mission Outcomes

Number of successful and unsuccessful missions:

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
 * sqlite:///my_data1.db
Done.
```

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

Boosters Carried Maximum Payload

These are the boosters which have carried the maximum payload mass:

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- These are the failed landing outcomes:

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

- There are only two occurrences in the data set where this occurs.

Landing_Outcome	TOTAL_NUMBER
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

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

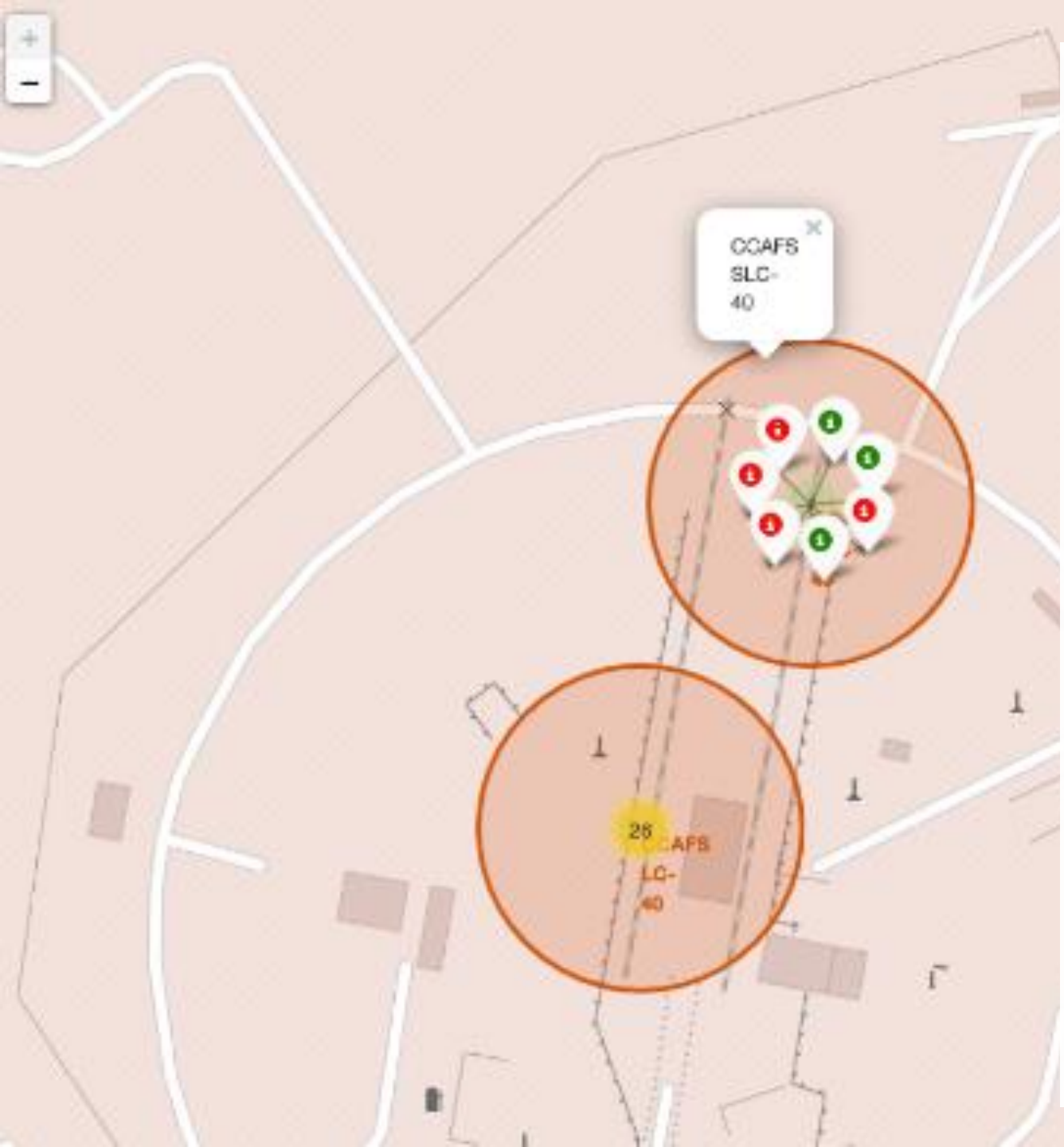
- The ranking of all landing outcomes between the date 2010-06-04 and 2017- 03-20.
- There are 10 occurrences of ‘no attempt’, implying that they are of significance.

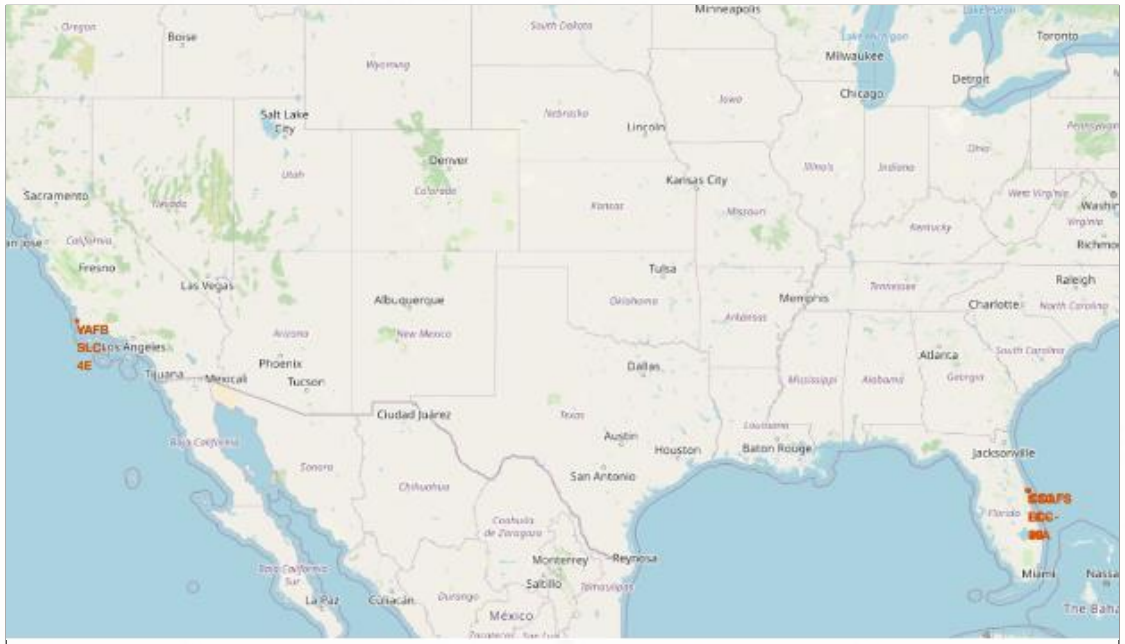
SECTION 4

LAUNCH SITES PROXIMITIES ANALYSIS

ALL LAUCH SITES

Launch sites are located near the coasts.



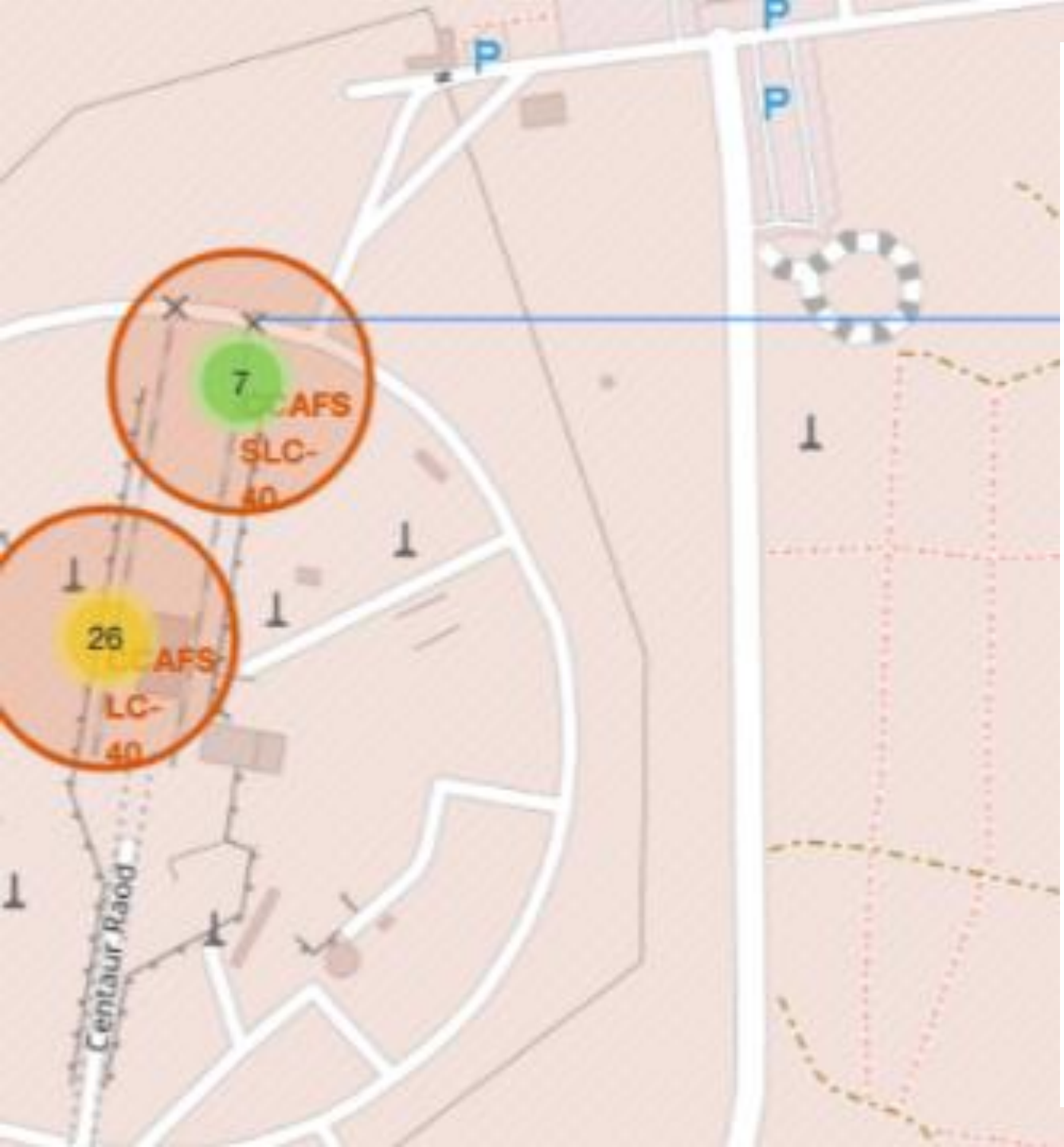


LAUNCH OUTCOMES BY SITE

Green = successful launches

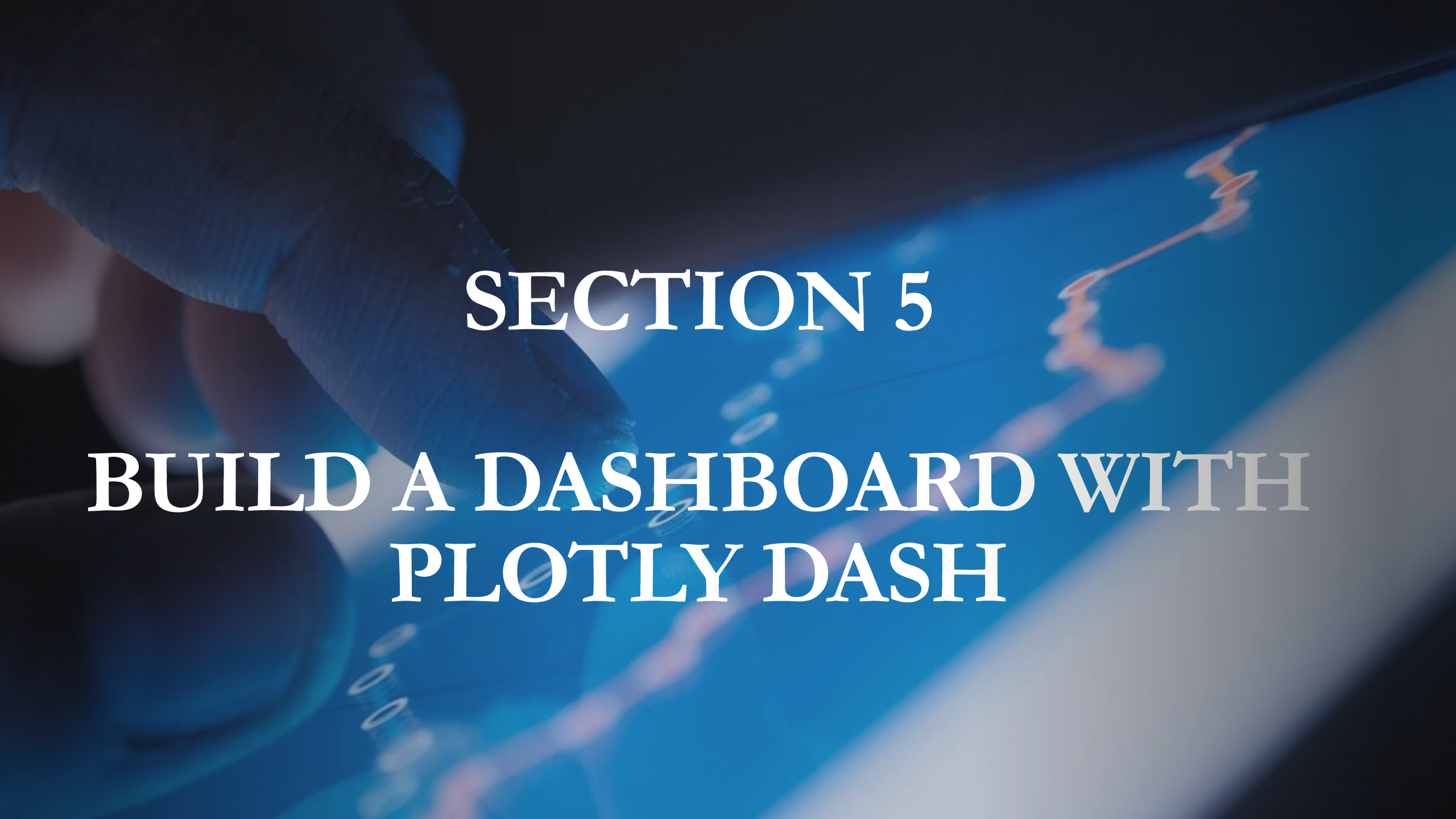
Red = failure

Example of CCAFS SLC - 40



LOGISTICS AND SAFETY

Launch site is located quite close to road, and away from inhabited areas making it a good location.

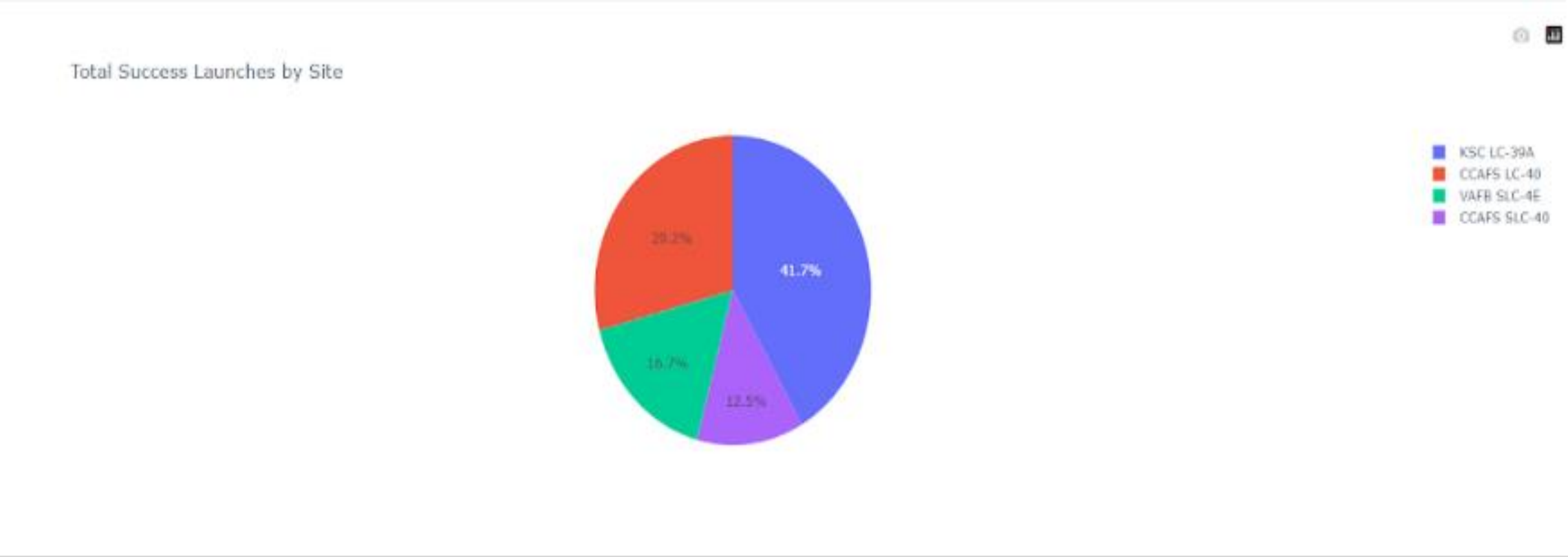


SECTION 5

BUILD A DASHBOARD WITH PLOTLY DASH

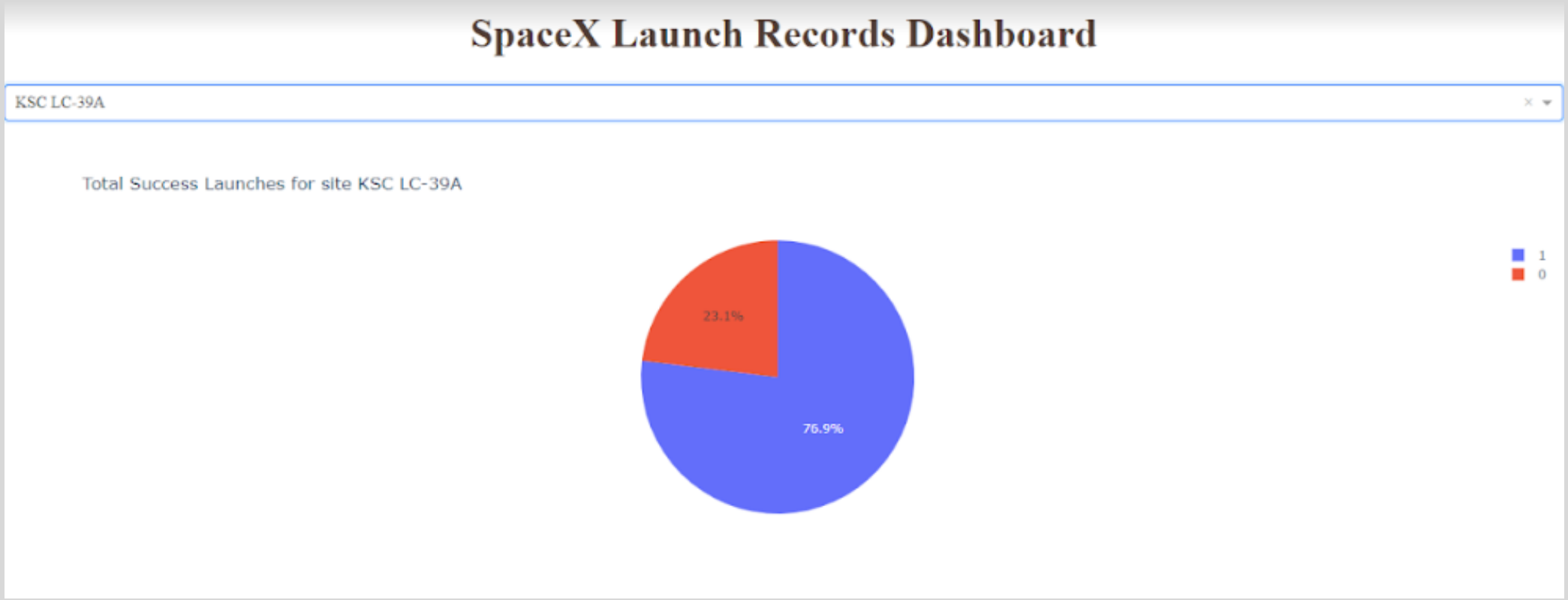
SUCCESSFUL LAUNCHES BY SITE

- Location of launches in relation to successful launches.
- This factor should be included as KSC LC – 39A seems to have the highest success rate.



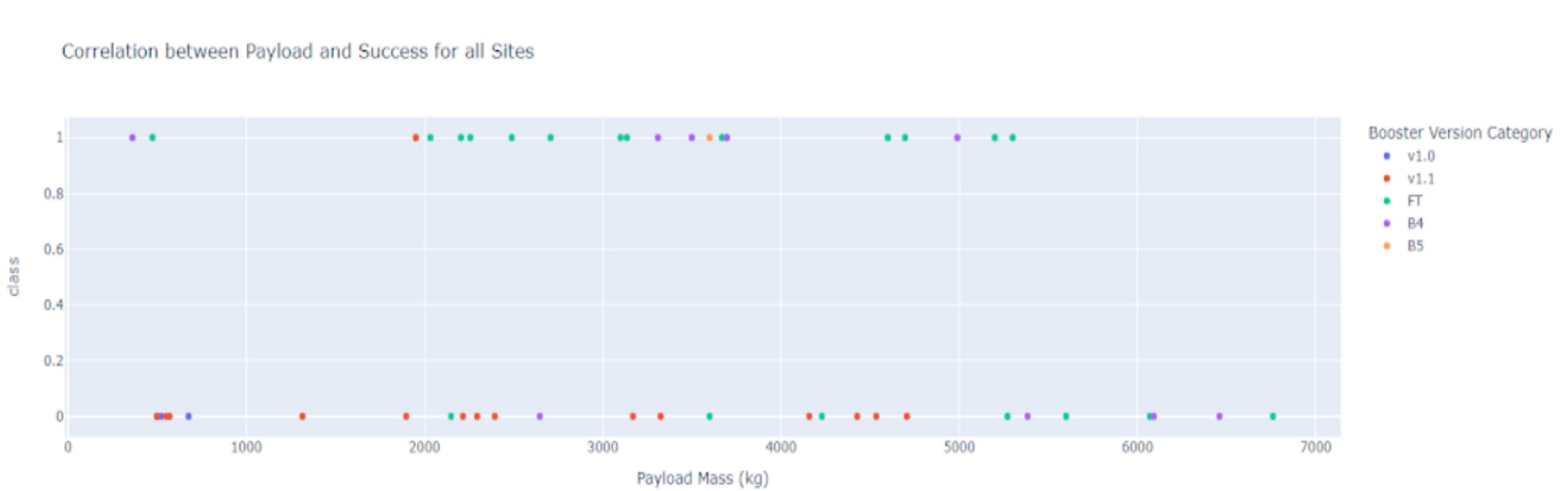
LAUNCH SUCCESS RATIO FOR KSC LC-39A

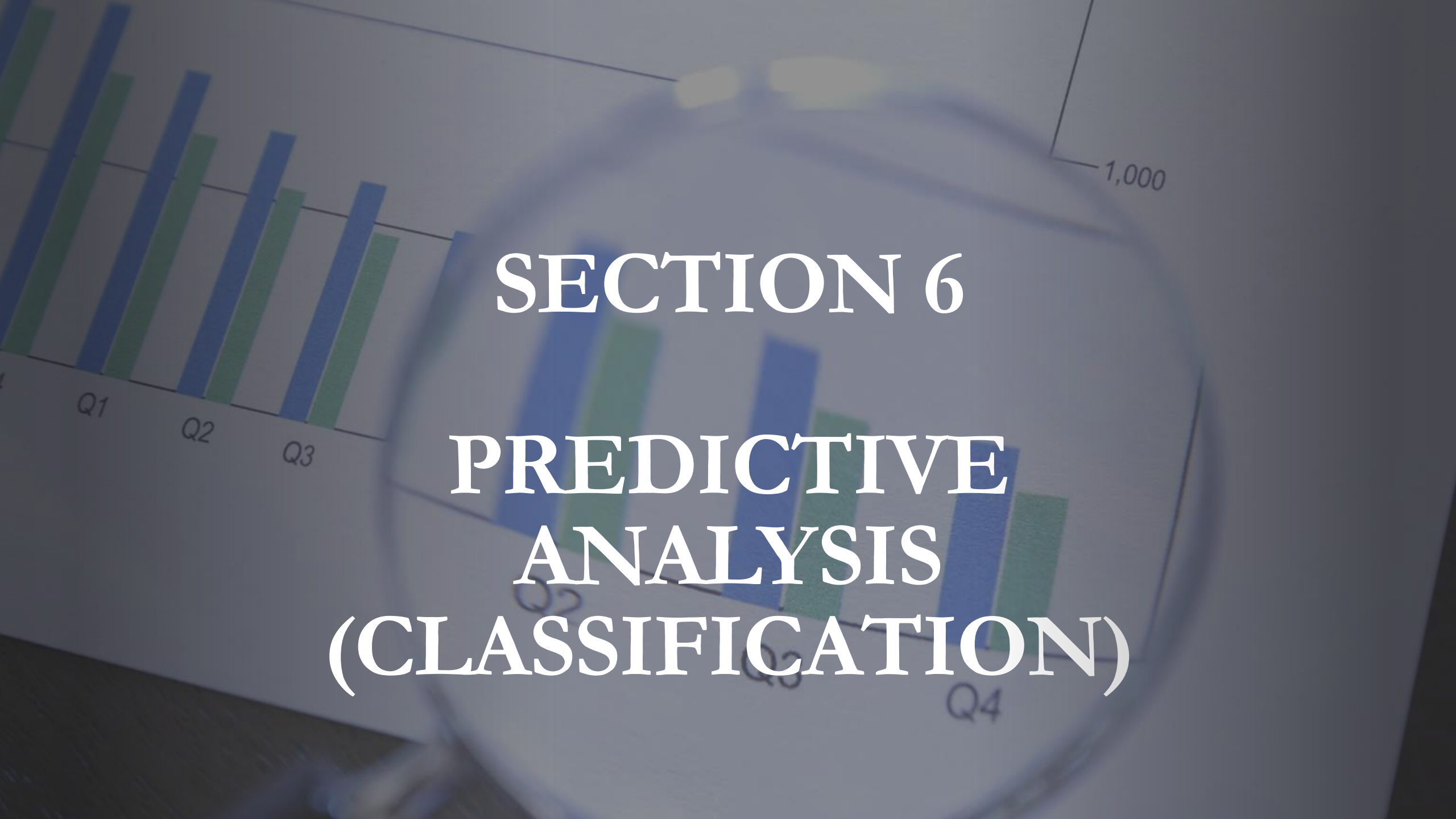
- KSC LC – 39A has a success rate of 76.9% which is represented by blue.



PAYLOAD VS. LAUNCH OUTCOME

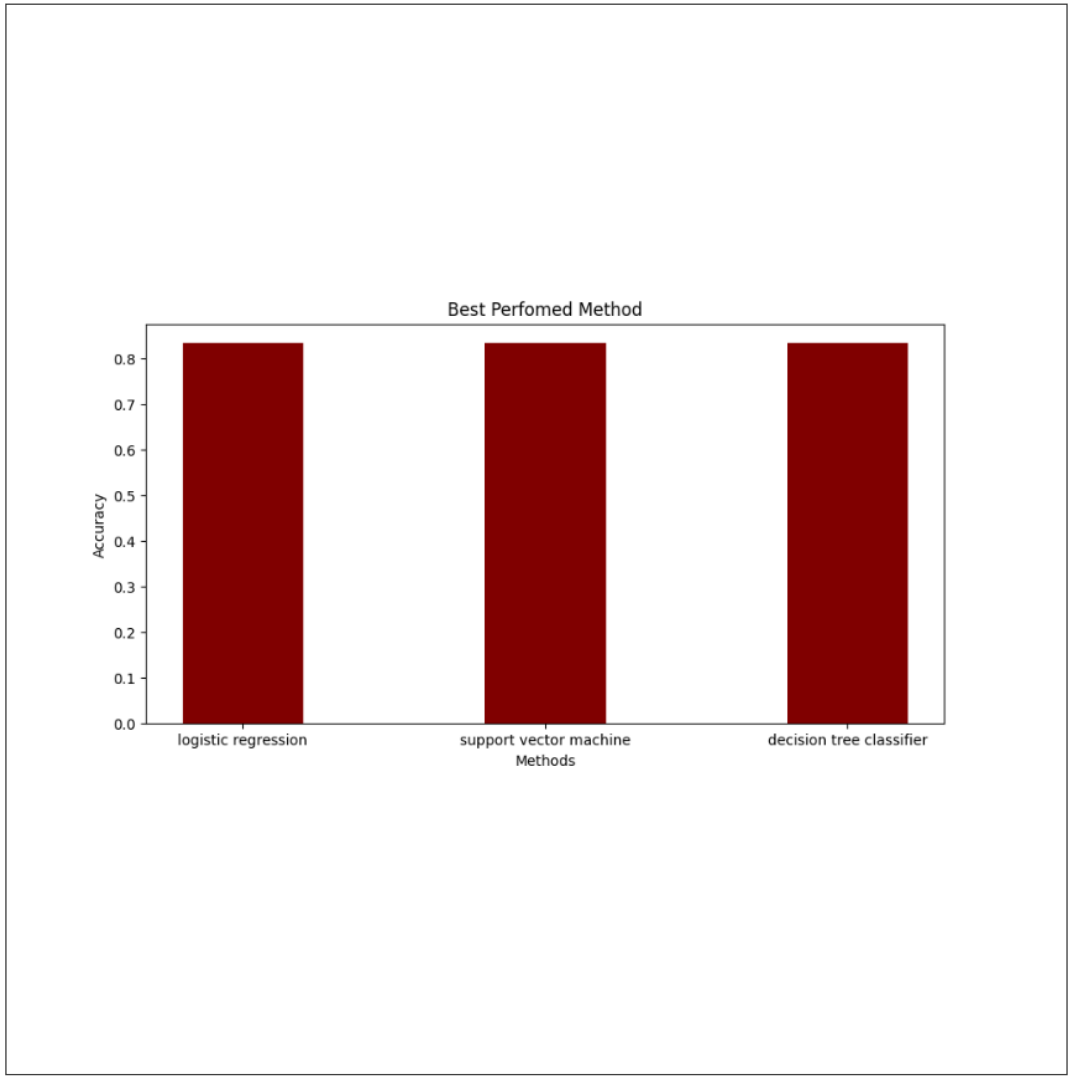
- Plotly dashboard has a Payload range selector.
- Payloads under 6,000kg and FT boosters are the most successful combination.





SECTION 6

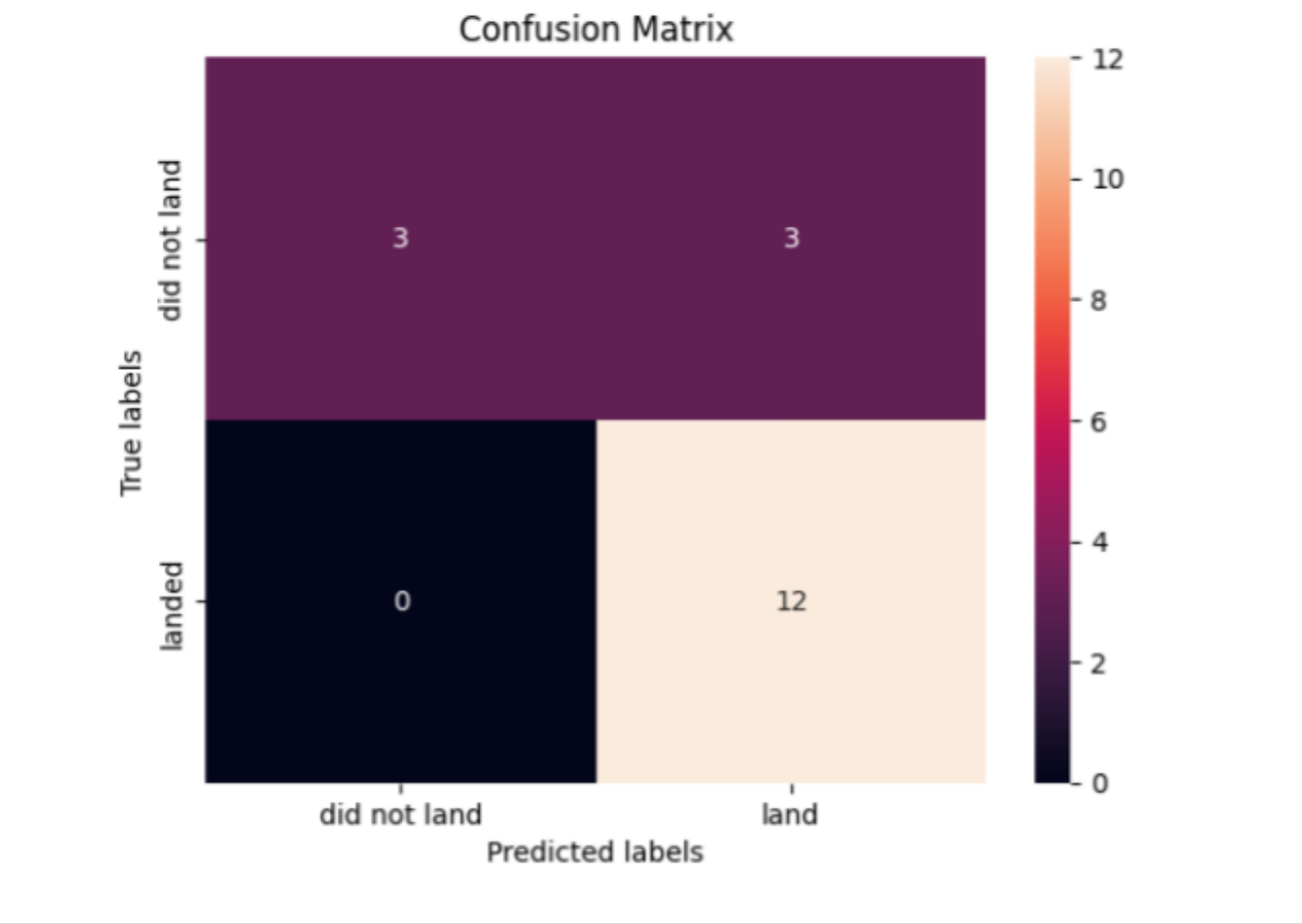
PREDICTIVE ANALYSIS (CLASSIFICATION)



CLASSIFICATION ACCURACY

- Overall, classification was not accurate as graph was unbale to represent which method performed the best.

CONFUSION MATRIX OF DECISION TREE CLASSIFIER



- Matrix shows the quantity of true positive and true negative compared to the incorrect answers.
- The confusion matrix is similar throughout all models.
- There was a prediction of 12 successful landings, which occurred.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.

CONCLUSION

- Aim: Use machine learning for ‘Space Y’ to bid against ‘SpaceX’ and try to save $100 million USD for the company.
- Overall, the best launch site is KSC LC-39A.
- Most of mission outcomes were successful, and the landing outcomes seemed to improve over time, with launches weighing above 7,000kg being the safest and being able to orbit.
- Decision Tree Classifier can be used to predict successful landings and increase profits.
- Next time, more data should be collected to create a better sample to improve overall accuracy of the machine learning model.

APPENDIX

- GitHub repository url:
https://github.com/Goran17l/Goran17l/tree/main/10.%20Applied%20Data%20Science%20Capstone

- Folium maps didn’t appear on GitHub, hence using screenshots.


THANK YOU