SpaceX Falcon 9 first stage Landing Prediction

Shubham Oulkar

2022-08-20

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection via API, Web Scraping
- Exploratory Data Analysis (EDA) with Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboards with Plotly Dash
- Predictive Analysis
- Summary of all results
- Exploratory Data Analysis results
- Interactive maps and dashboard
- Predictive results

INTRODUCTION

Project background and context

The aim of this project is to predict if the Falcon 9 first stage will successfully land. SpaceX says on its website that the Falcon 9 rocket launch cost 62 million dollars. Other providers cost upward of 165 million dollars each. The price difference is explained by the fact that SpaceX can reuse the first stage. By determining if the stage will land, we can determine the cost of a launch. This information is interesting for another company if it wants to compete with SpaceX for a rocket launch.

Problems you want to find answers

- What are the main characteristics of a successful or failed landing ?
- What are the effects of each relationship of the rocket variables on the success or failure of a landing ?
- What are the conditions which will allow SpaceX to achieve the best landing success rate ?

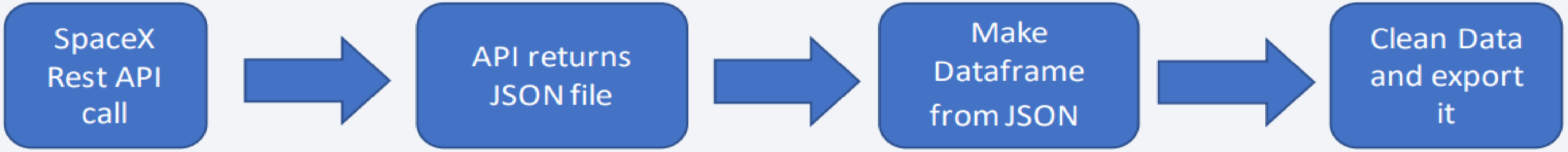
METHODOLOGY

Executive Summary

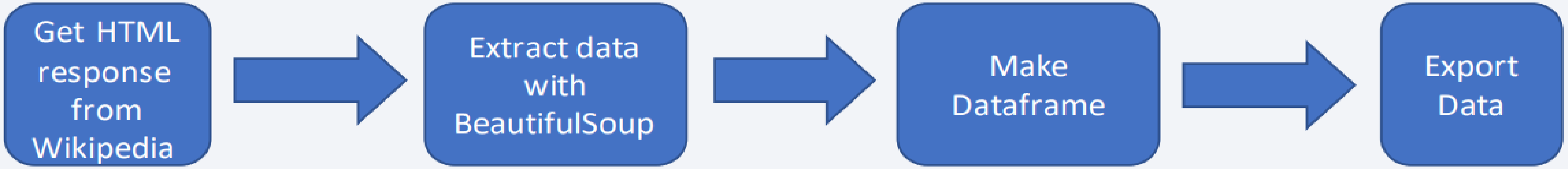
- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Datasets are collected from Rest SpaceX API and webscrapping Wikipedia
  - The information obtained by the API are rocket, launches, payload information.
    - The Space X REST API URL is api.spacexdata.com/v4/



  - The information obtained by the webscrapping of Wikipedia are launches, landing, payload information.
    - URL is https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922



Data Wrangling

- In the dataset, there are several cases where the booster did not land successully.
  - True Ocean, True RTLS, True ASDS means the mission has been successful.
  - False Ocean, False RTLS, False ASDS means the mission was a failure.
- We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

1. Calculate launches number for each site

```
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

2. Calculate the number and occurence of each orbit

```
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
SO        1
```

3. Calculate number and occurrence of mission outcome per orbit type

```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
None ASDS       2
False Ocean     2
False RTLS      1
Name: Outcome, dtype: int64
```

4. Create landing outcome label from Outcome column

```
landing_class = []
for key,value in df["Outcome"].items():
    if value in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
df['Class']=landing_class
```

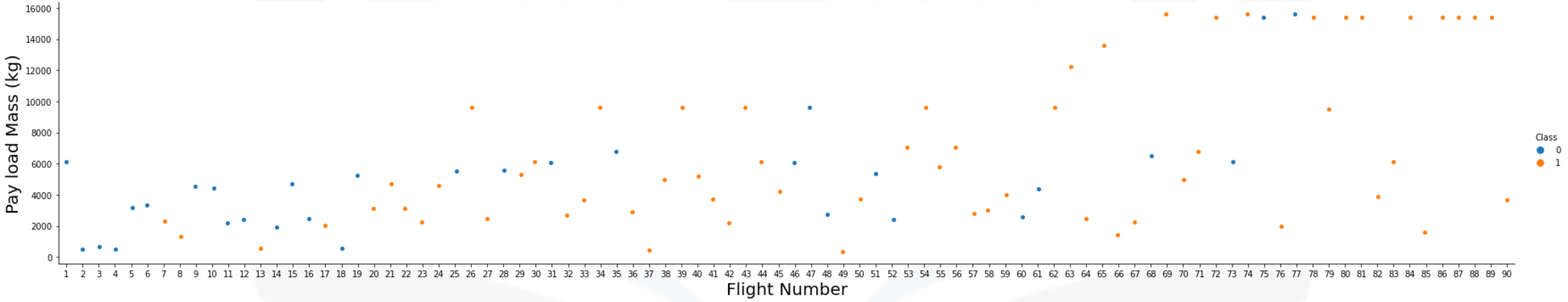
5. Export to file

```
df.to_csv("dataset_part_2.csv", index=False)
```

EDA with Data Visualization

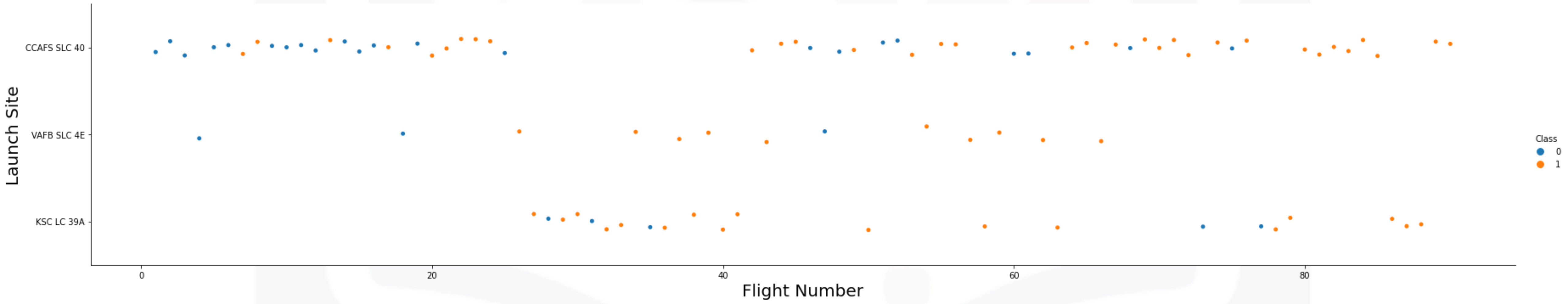
Flight Number vs Pay load Mass (kg)

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.



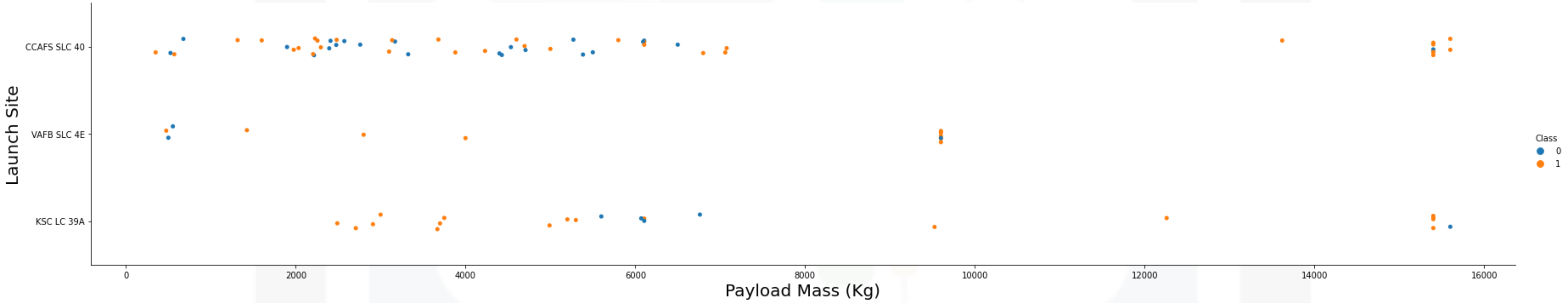
Flight Number and Launch Site

Launch Site VAFB SLC 4E have 13 launches, 3 were unsuccessful landings. KSC LC 39A had 5 unsuccessful landings. CCAFA SLC 40 had most of the launches and most of the unsuccessful landings.



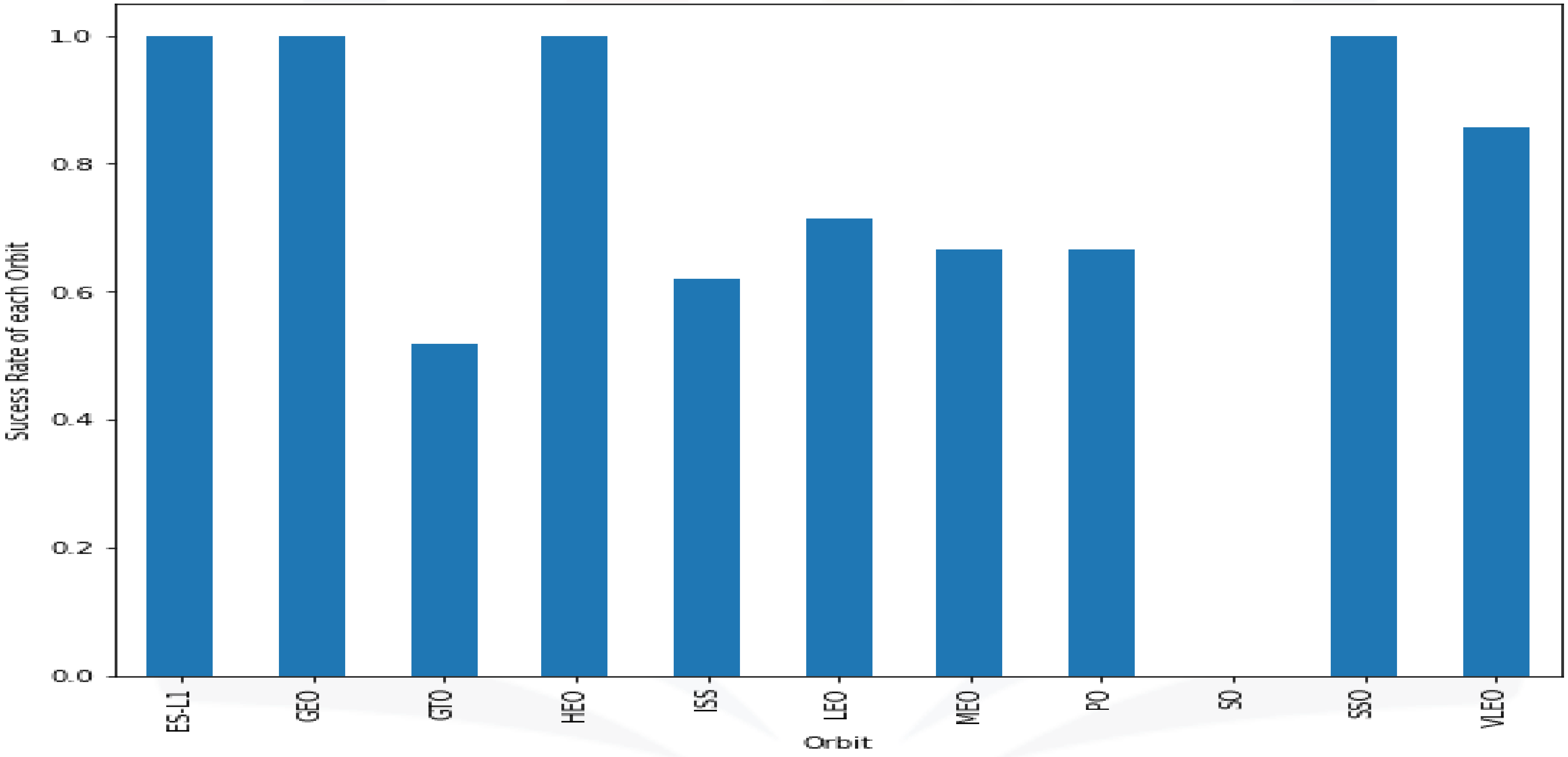
Payload and Launch Site

Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).



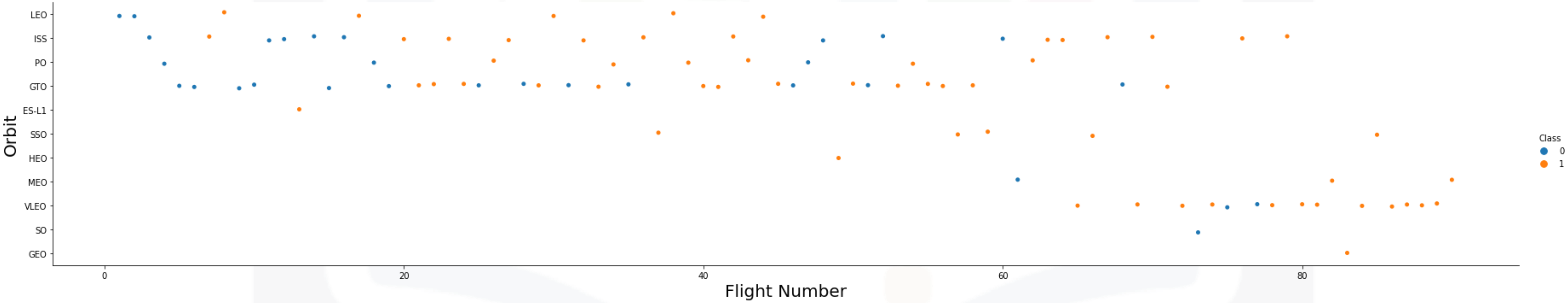
success rate of each orbit type

ES-L1, GEO, HEO, SSO Orbits with higher success rate of landing.



Flight Number and Orbit type

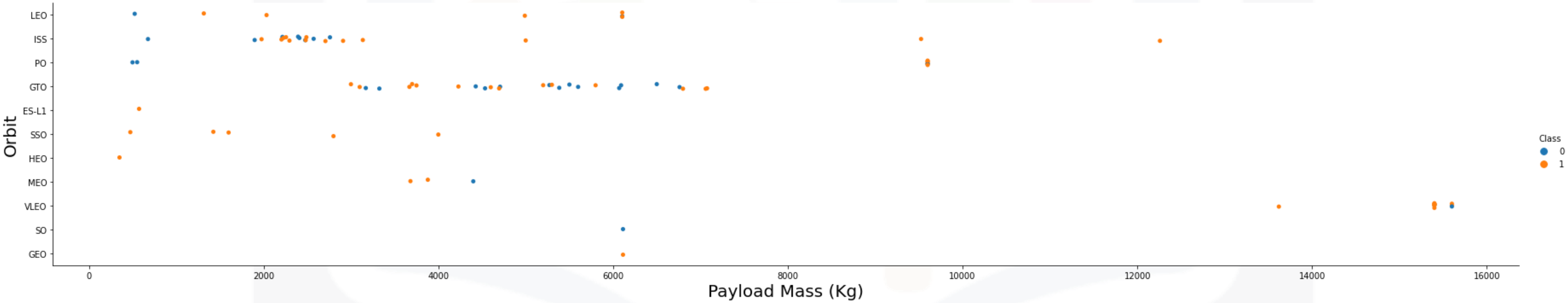
You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



Payload and Orbit type

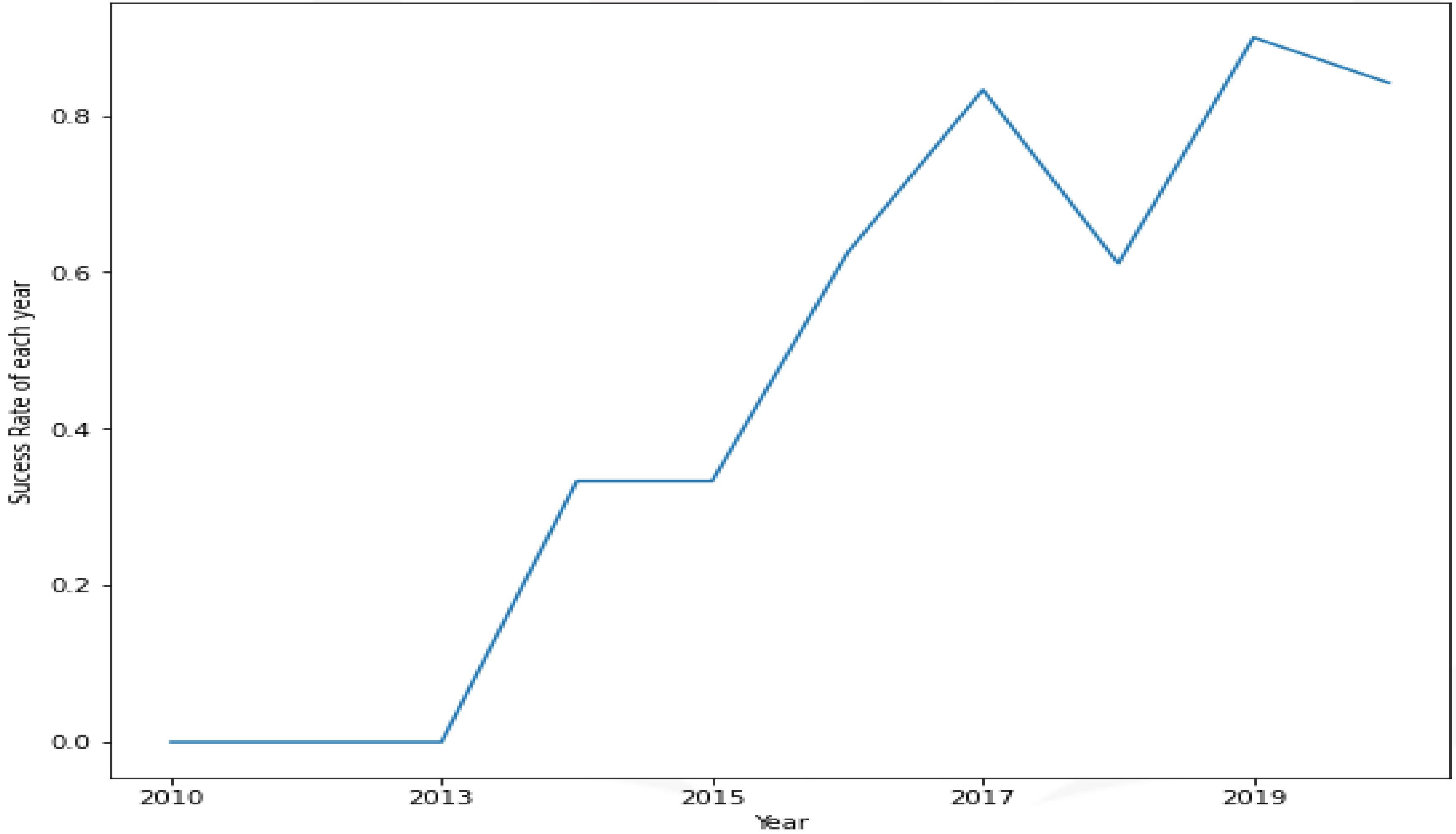
With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch success yearly trend

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



Build an Interactive Map with Folium

• Folium map object is a map centered on NASA Johnson Space Center at Houson, Texas

• Red circle at NASA Johnson Space Center's coordinate with label showing its name (folium.Circle, folium.map.Marker).

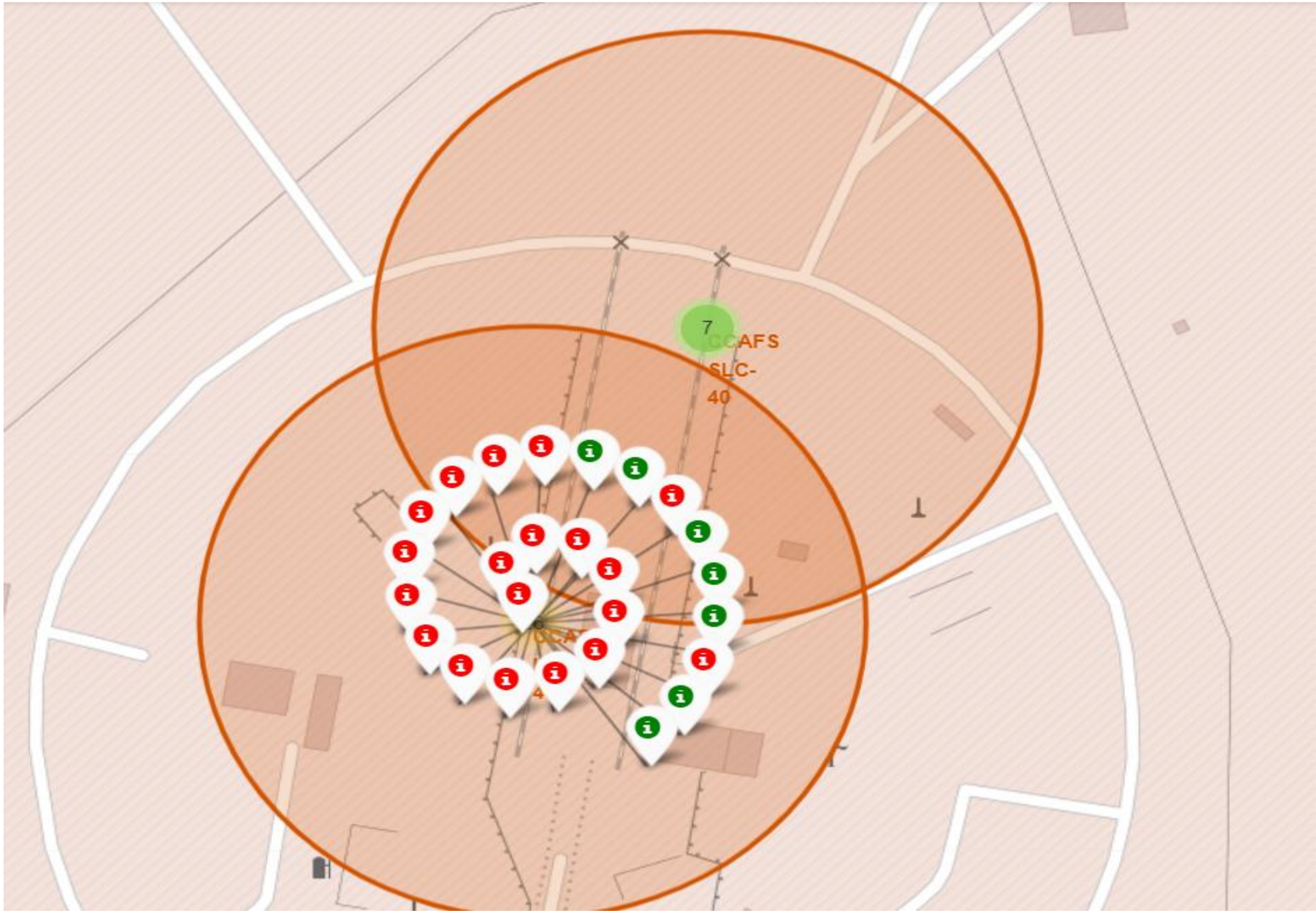
• Red circles at each launch site coordinates with label showing launch site name (folium.Circle, folium.map.Marker,
folium.features.DivIcon).

• The grouping of points in a cluster to display multiple and different information for the same coordinates (folium.plugins.MarkerCluster).

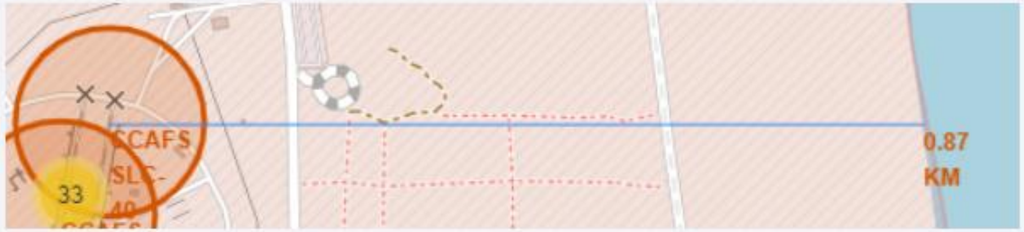
• Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing.
(folium.map.Marker, folium.Icon).

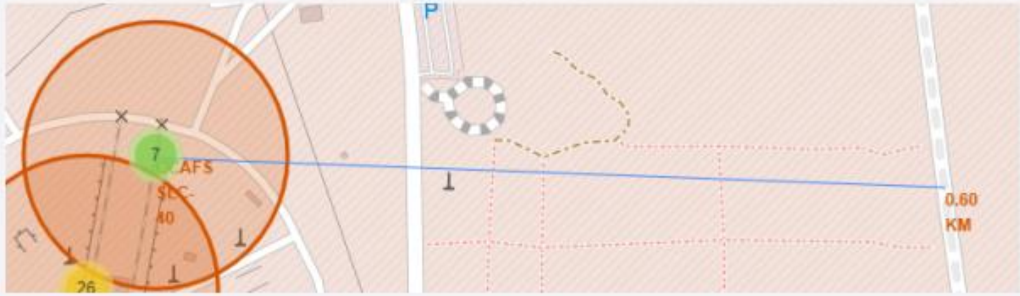
• Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them.
(folium.map.Marker, folium.PolyLine, folium.features.DivIcon)

• These objects are created in order to understand better the problem and the data. We can show easily all launch sites, their surroundings and the number of successful and unsuccessful landings.


7
CCAFS
SLC-
40

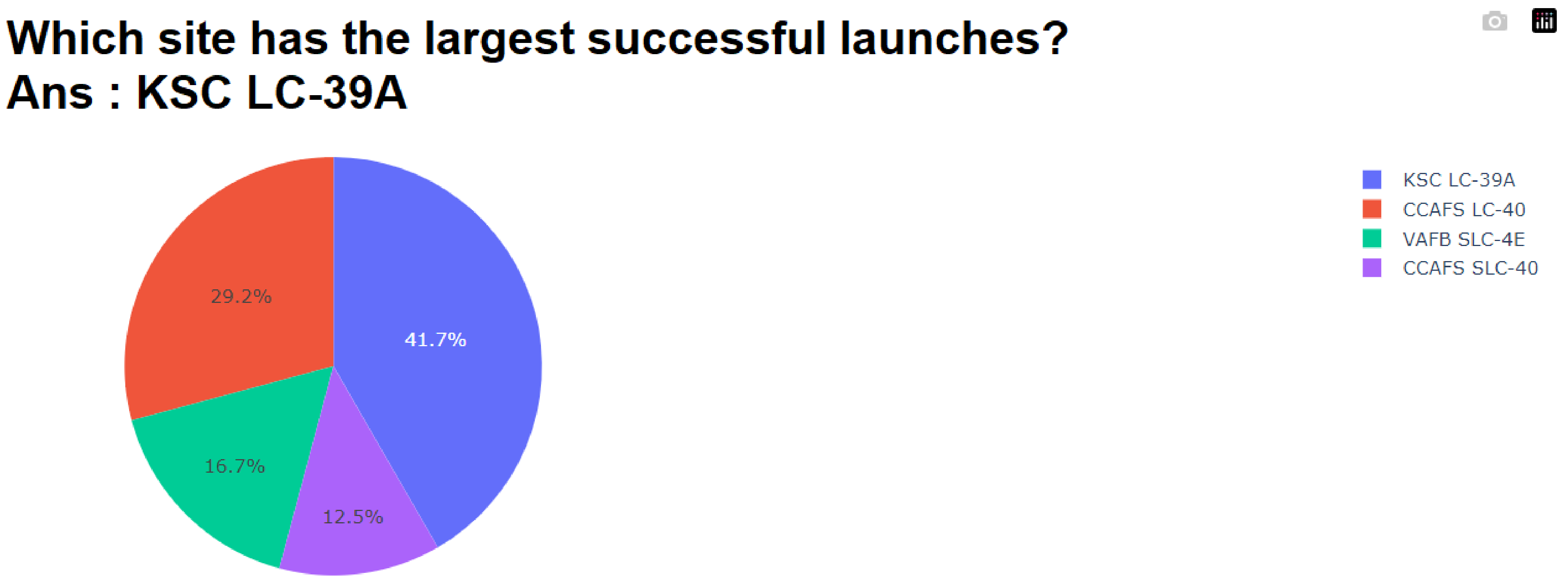
Folium Map – Distances between CCAFS SLC-40 and its proximities





Is CCAFS SLC-40 in close proximity to railways ? Yes
Is CCAFS SLC-40 in close proximity to highways ? Yes
Is CCAFS SLC-40 in close proximity to coastline ? Yes
Do CCAFS SLC-40 keeps certain distance away from cities ? No

Dashboard – Total success by Site



Which site has the highest launch success rate?

Total Success Launches for Site KSC LC-39A

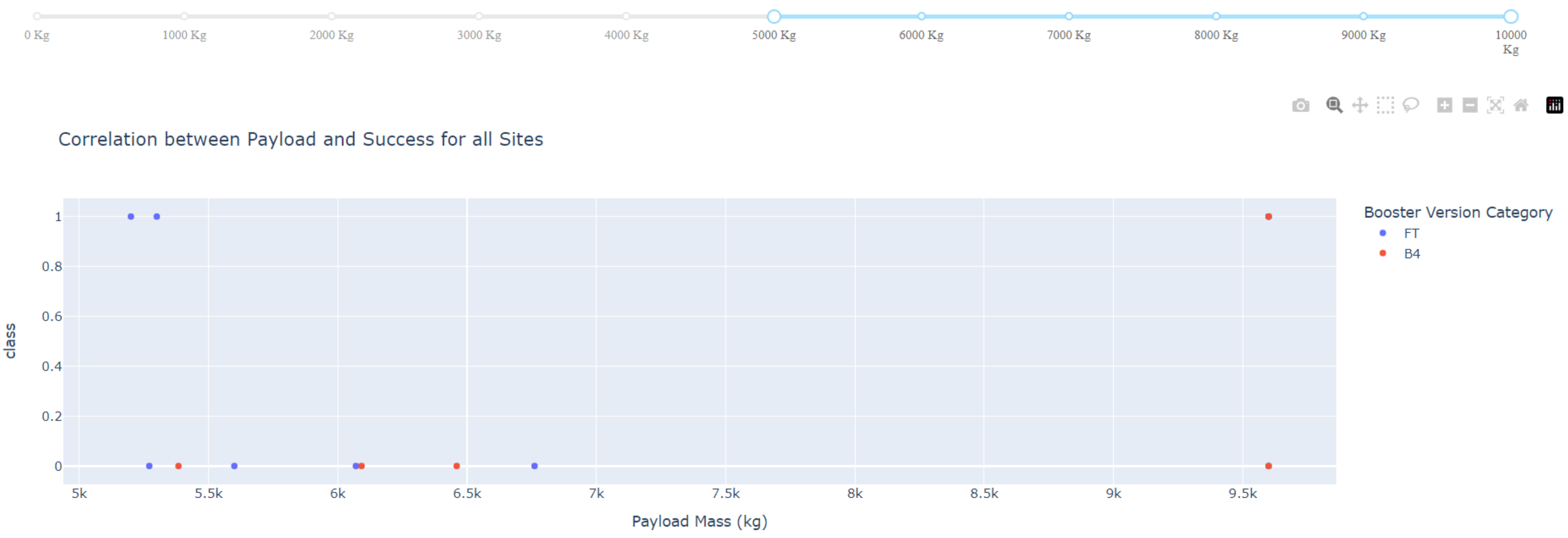
KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.

Low weighted payload (0 – 5000 kg)

Payload range (Kg):

Heavy weighted payload (5000 – 10000 kg)

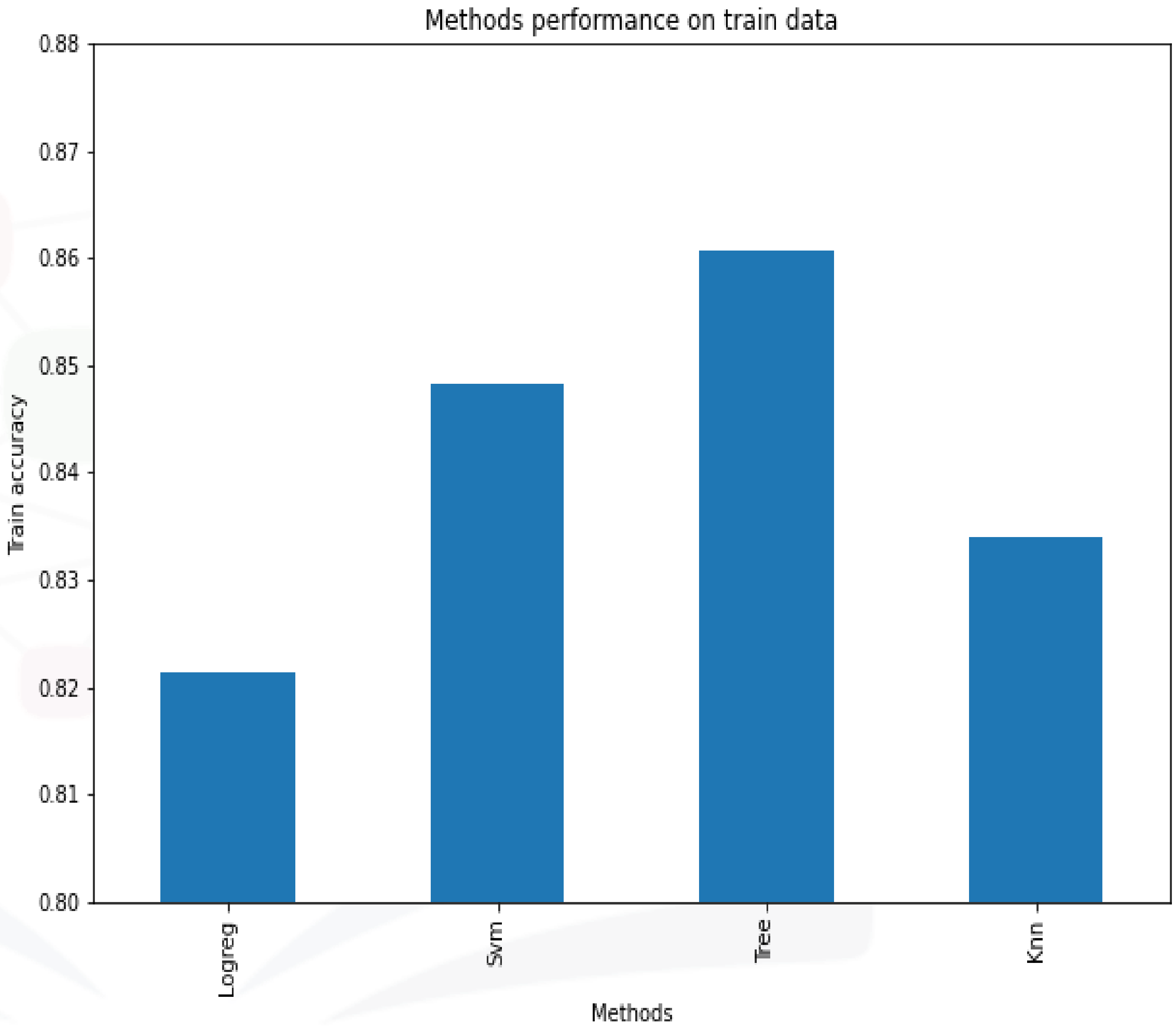
yload range (Kg):

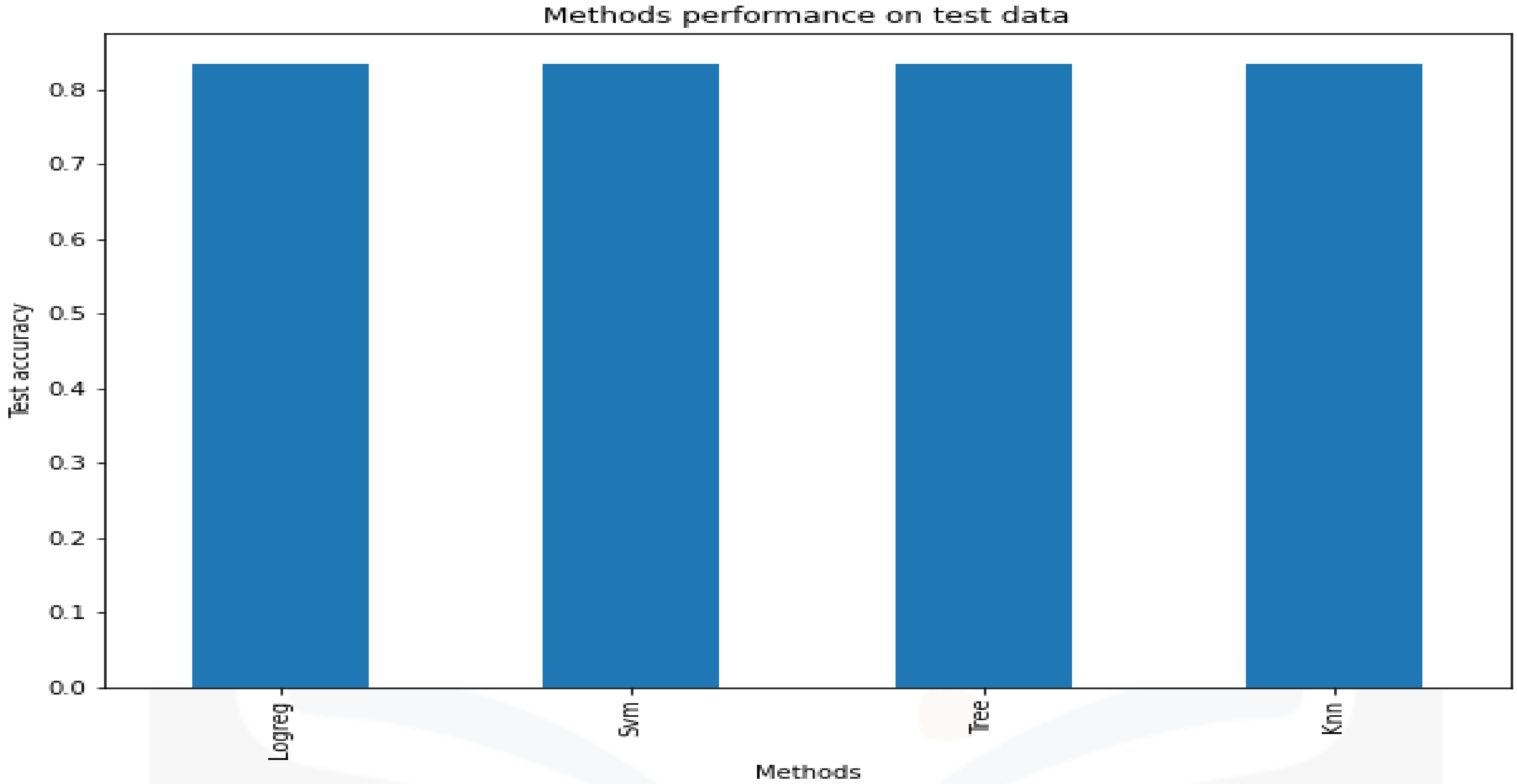


Predictive Analysis (Classification)

- Data preparation
  - Load dataset
  - Normalize data
  - Split data into training and test sets.
- Model preparation
  - Selection of machine learning algorithms
  - Set parameters for each algorithm to GridSearchCV
  - Training GridSearchModel models with training dataset
- Model evaluation
  - Get best hyperparameters for each type of model
  - Compute accuracy for each model with test dataset
  - Plot Confusion Matrix
- Model comparison
  - Comparison of models according to their accuracy
  - The model with the best accuracy will be chosen (see Notebook for result)

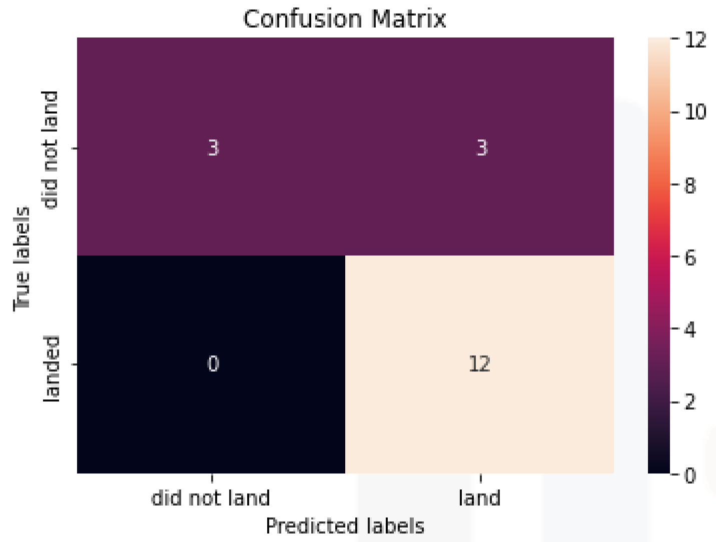
	Accuracy Train	Accuracy Test
Logreg	0.821429	0.833333
Svm	0.848214	0.833333
Tree	0.860714	0.833333
Knn	0.833929	0.833333

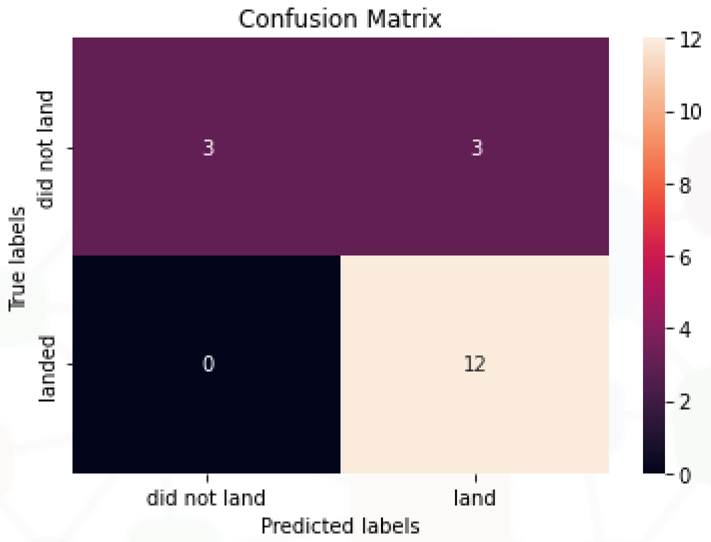


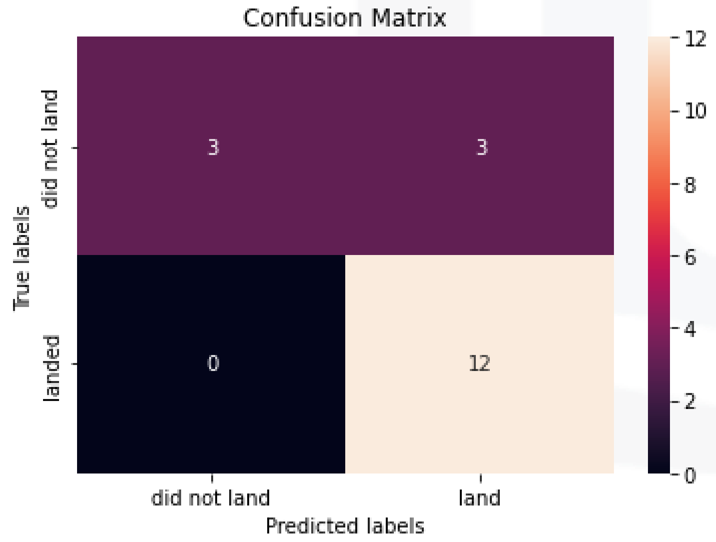


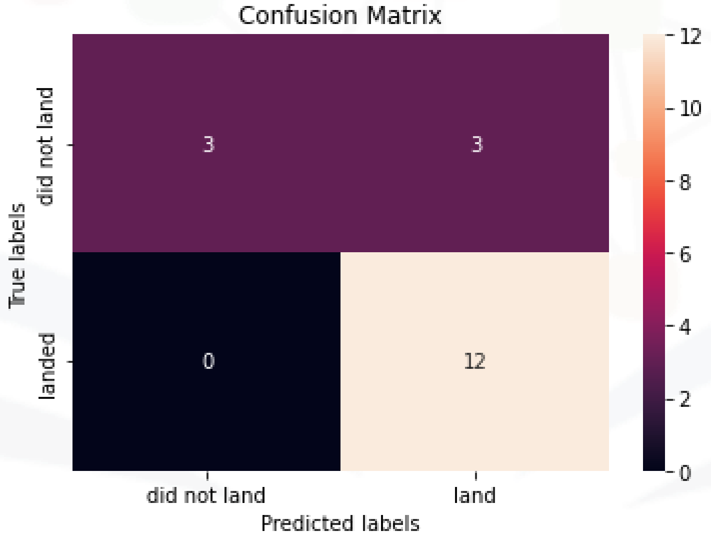
Methods performance on test data
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Test accuracy
Logreg
Svm
Tree
Knn
Methods

Confusion Matrix









As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models are false positives.

Conclusions

- The success of a mission can be explained by several factors such as the launch site, the orbit and especially the number of previous launches. Indeed, we can assume that there has been a gain in knowledge between launches that allowed to go from a launch failure to a success.
- The orbits with the best success rates are GEO, HEO, SSO, ES-L1.
- Depending on the orbits, the payload mass can be a criterion to take into account for the success of a mission. Some orbits require a light or heavy payload mass. But generally low weighted payloads perform better than the heavy weighted payloads.
- With the current data, we cannot explain why some launch sites are better than others (KSC LC-39A is the best launch site). To get an answer to this problem, we could obtain atmospheric or other relevant data.
- For this dataset, we choose the Decision Tree Algorithm as the best model even if the test accuracy between all the models used is identical. We choose Decision Tree Algorithm because it has a better train accuracy.