APPROACHING SPACEX ROCKET SCIENCE WITH DATA SCIENCE

KARTHIK M K

27/10/2022

IBM Developer

SKILLS NETWORK

OUTLINE

- ❖ Executive Summary
- ❖ Introduction
- ❖ Methodology
- ❖ Results
- ❖ Conclusion

EXECUTIVE SUMMARY

- Summary of methodologies
  - Data collection through web scraping and API
  - Exploratory Data Analysis with SQL and data visualization
  - Maps with folium
  - Dashboards with Plotly Dash
  - Machine Learning Prediction

- Summary of results
  - EDA results
  - Interactive maps and dashboards
  - Prediction results

INTRODUCTION

- Objective:

  The object of this research is to forecast whether the Falcon 9 first stage will successfully land. The launch of the Falcon 9 rocket, according to SpaceX, cost 62 million dollars. The price of a launch can be calculated by assessing if the stage will land. If another business plans to compete with SpaceX for a rocket launch, this information may be useful.

- Desirables
  - What are the features required for predicting successful or failed landings?
  - Which are the best places to achieve successful landing rate?

METHODOLOGY

- Data collection via public sources
  - SpaceX API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Determined training labels
- Perform EDA with SQL and visualization
- Perform visual analytics using folium and Plotly
- Perform predictive analysis using classification models
  - Train test split
  - Model building and finding accuracy of each using different parameters

Data Collection with API

- This API was used according to the flowchart.

Request API and parse the launch data → Filtered data to get Falcon 9 launches → Missing value treatment

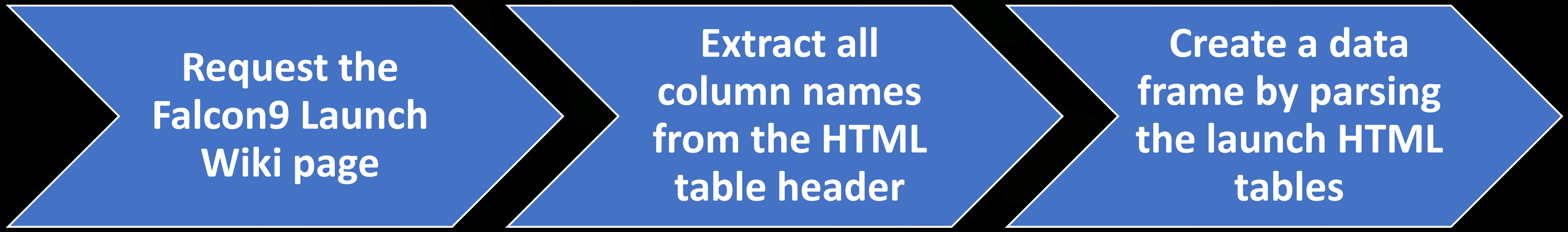
- Source : SpaceY/Data Collection API.ipynb at master · KarthikMK21/SpaceY (github.com)

IBM Developer

SKILLS NETWORK

Data Collection with Web Scraping

- The data can also be collected through web scraping from Wikipedia as shown in the below chart



Source : [SpaceY/Data Collection with Web Scraping.ipynb at master · KarthikMK21/SpaceY (github.com)]

IBM Developer

SKILLS NETWORK

Data Wrangling

- EDA for finding some patterns in the data
- Determining what would be the label for training supervised models was the main objective of this process.
- The summaries 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.

- Source : SpaceY/EDA.ipynb at master · KarthikMK21/SpaceY (github.com)

EDA with SQL

Following SQL queries were performed on the data

- Displaying the names of the unique launch sites in the space mission.
- Display 5 records where launch sites begin with the string 'CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster versions which have carried the maximum payload mass.
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order

Source : SpaceY/EDA with SQL.ipynb at master · KarthikMK21/SpaceY (github.com)

EDA with Visualization

- Scatter plots, Bar Graphs and Line graph were used to perform EDA for following characteristics

- Scatter Plot
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type

- Bar Graph
  - Success Rate vs Orbit Type

- Line Chart
  - Success Rate vs Year

Source : SpaceY/EDA Using Visualization.ipynb at master · KarthikMK21/SpaceY (github.com)

Visualization with Folium

Markers, circles, lines and marker clusters were used with Folium Maps

- Markers indicate points like launch sites
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center
- Marker clusters indicates groups of events in each coordinate, like launches in a launch site
- Lines are used to indicate distances between two coordinates.

Source : Dashoboard_with_folium - Jupyter Notebook

Dashboards using Plotly

- The following graphs and plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This combination allows to quickly analyze the relation between payloads and launch sites, helping to identify which is the best place to launch according to payloads.

Source : SpaceY/spacex_dash_app.py at master · KarthikMK21/SpaceY (github.com)

Predictive Analysis

- Four classification models were built to test the model
  1. Logistic Regression
  2. Support Vector Machine
  3. Decision Tree
  4. K Nearest Neighbors

By processing through data standardization, testing through various hyperparameters and comparing the results

Source: SpaceY/Machine_Learning_Prediction.ipynb at master · KarthikMK21/SpaceY (github.com)

RESULTS

EDA with SQL

- Names of the unique launch sites in the space mission.

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

- Display 5 records where launch sites begin with the string '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	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The total payload mass carried by boosters launched by NASA (CRS) is 111268
- Average payload mass carried by booster version F9 v1.1 is 2928 kg
- The date when the first successful landing outcome in ground pad was achieved on 22-12-2015
- The names of the boosters which have success in drone ship and have 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 names of the booster versions 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 total number of successful and failure mission outcomes

mission_outcome	no_of_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

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

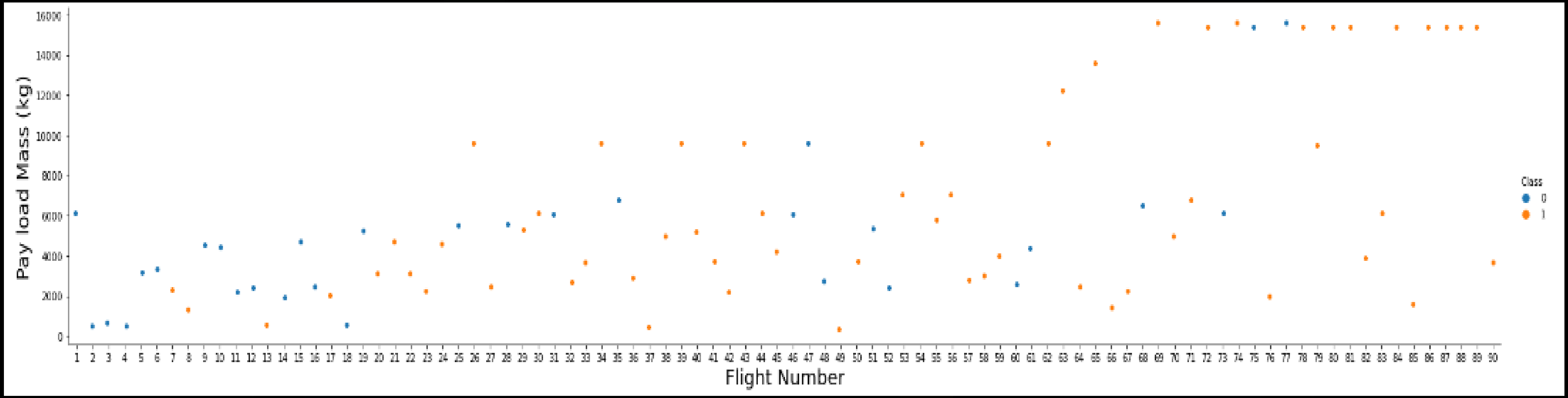
booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- 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	rank_of_outcomes
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

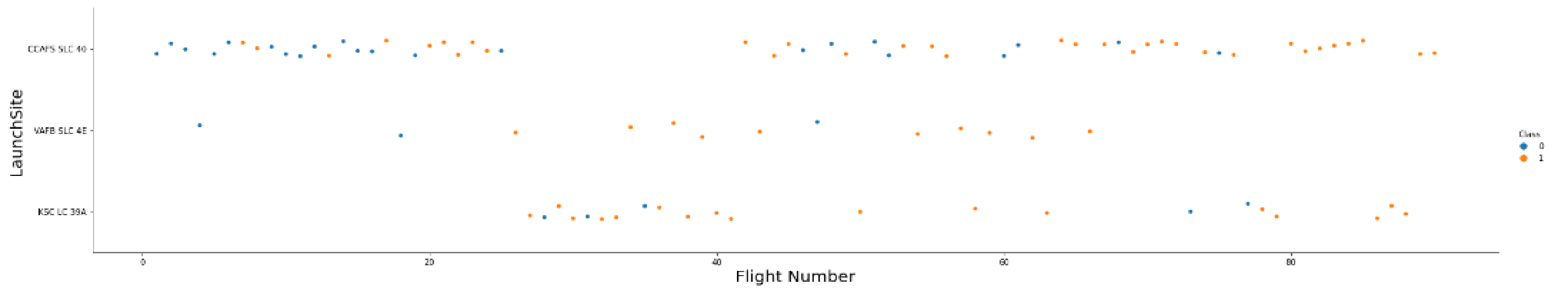
EDA with Visualization

- Flight Number vs PayloadMass



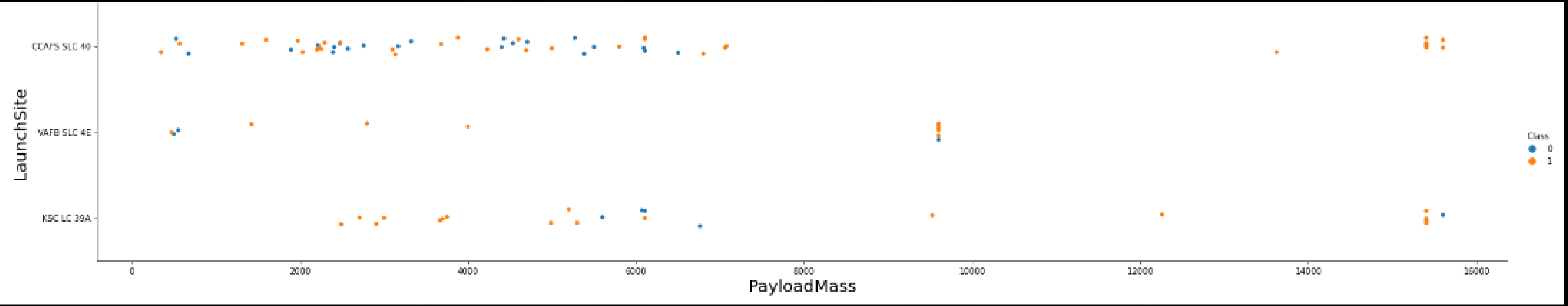
We see that different launch sites have different success rates. CCAFS LC-40 has a success rate of 60% while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

- # Flight Number vs Launch Site



We observe that for each site success rate is increasing

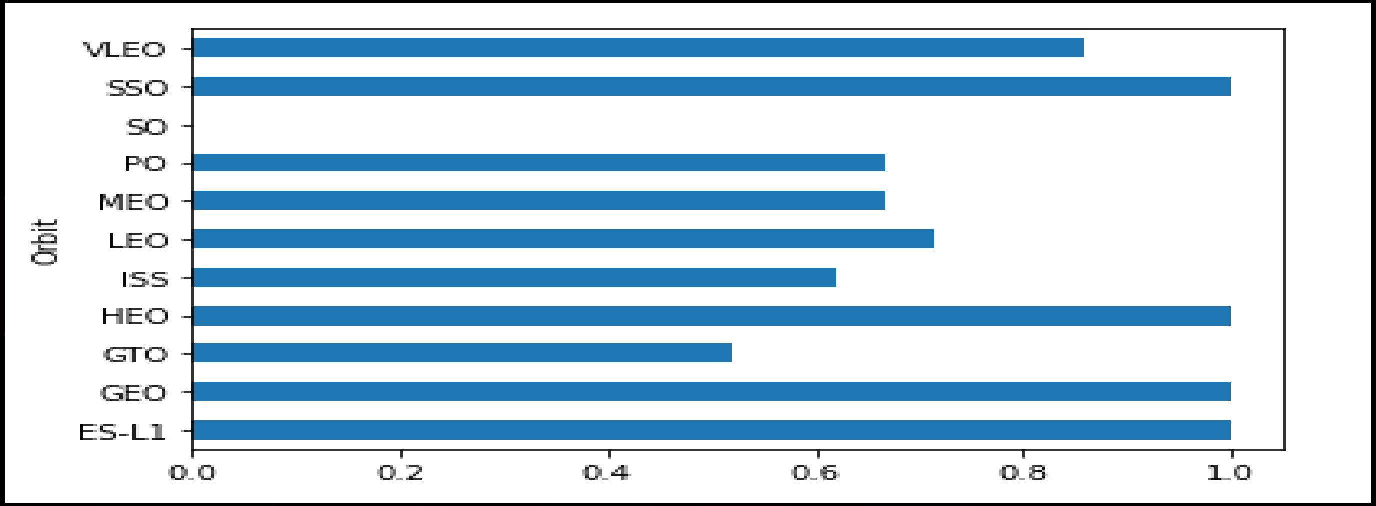
- # Payload vs Launch Site



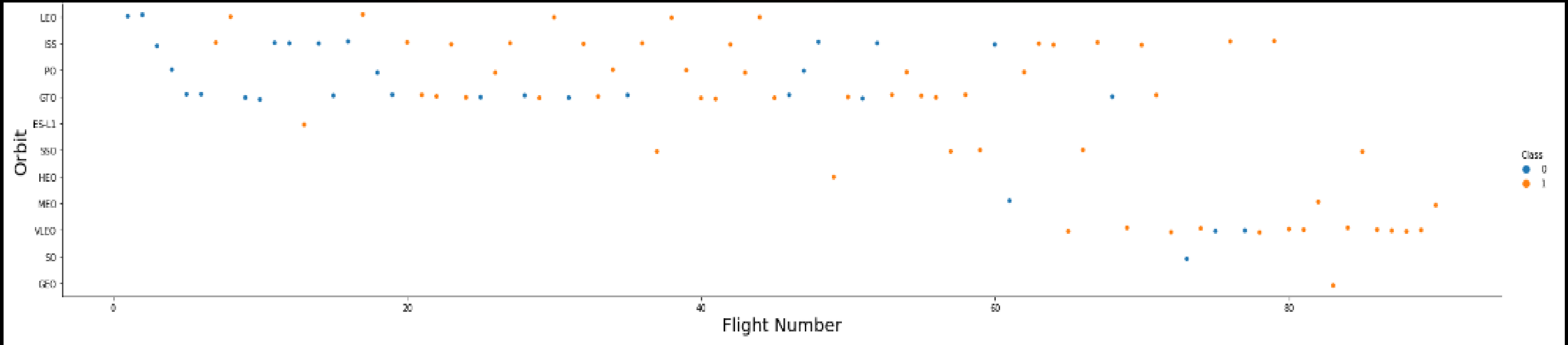
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).

- Success Rate vs Orbit

ES-L1, GEO, HEO, SSO have the best success rate.



- Flight Number vs Orbit Type

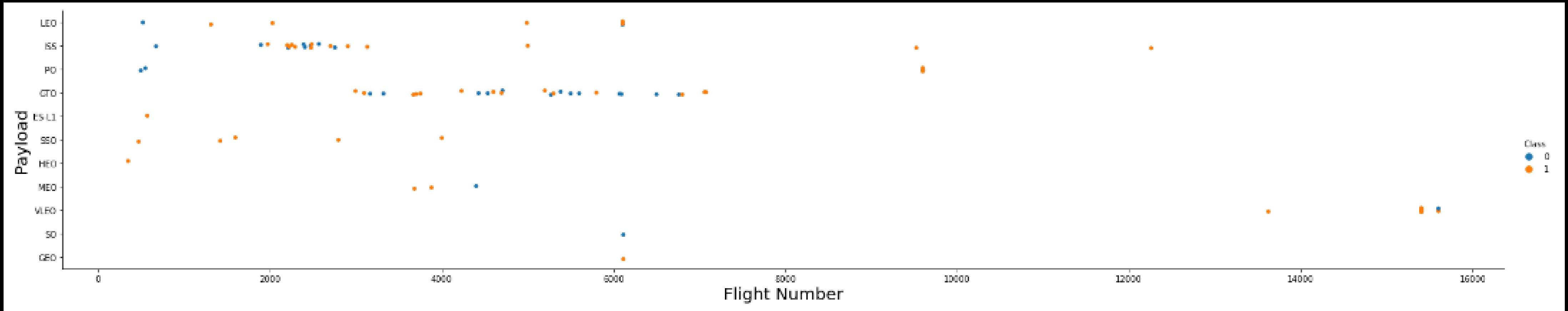


You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

IBM Developer

SKILLS NETWORK

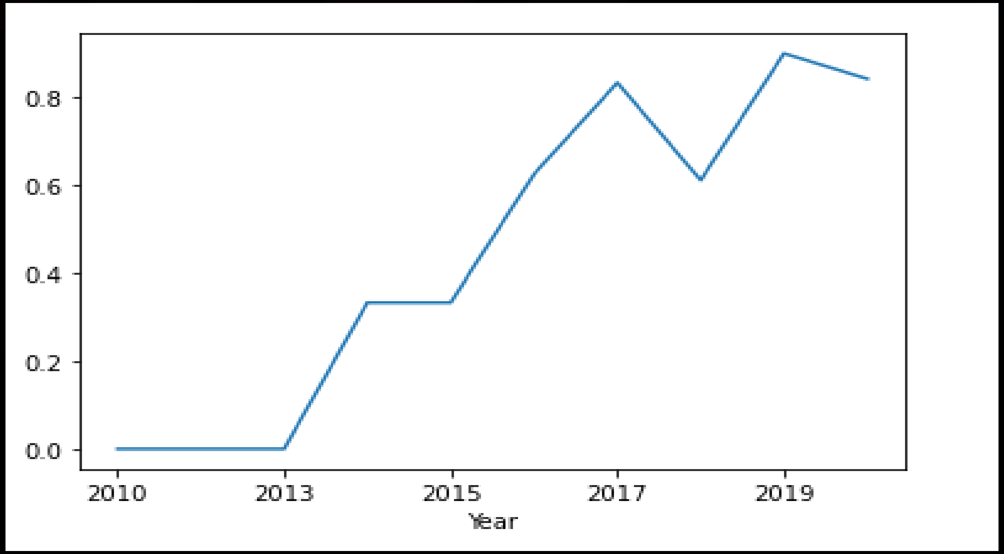
- ## Payload vs Orbit Type



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

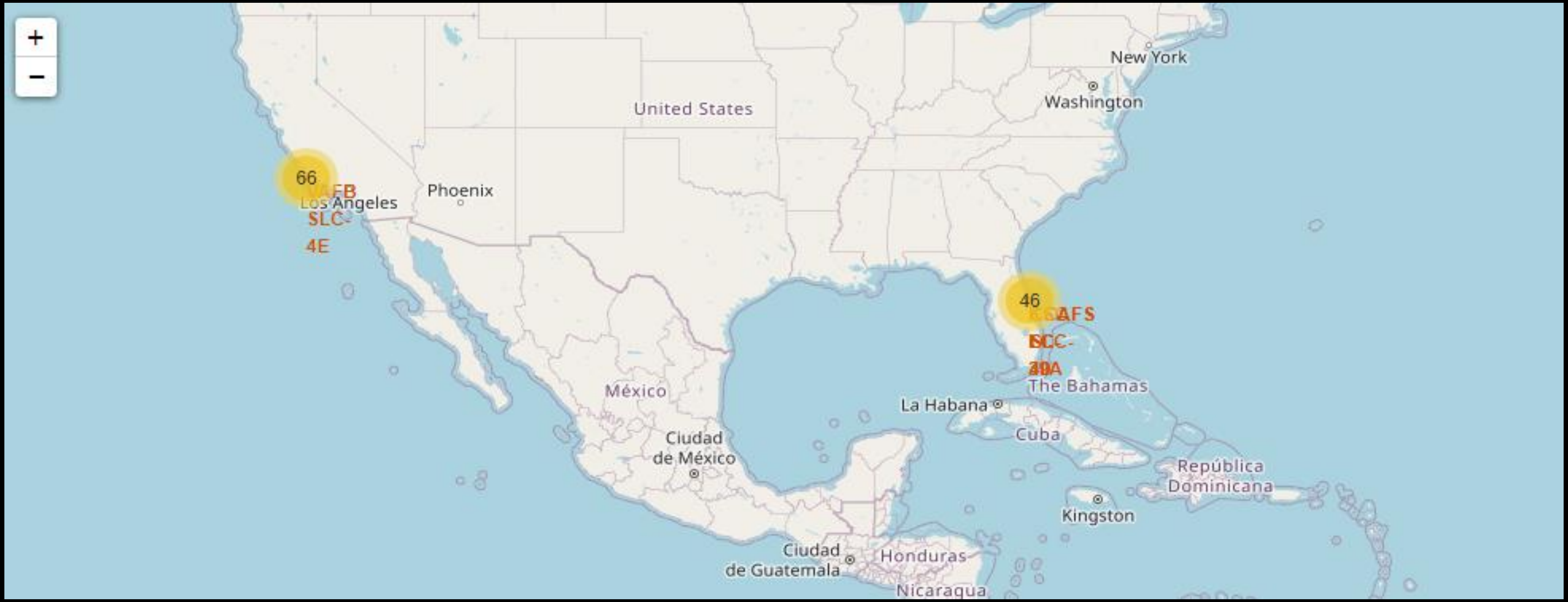
- ## Success Rate vs Year

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



Launch Site Proximity Analysis

- Ground Stations

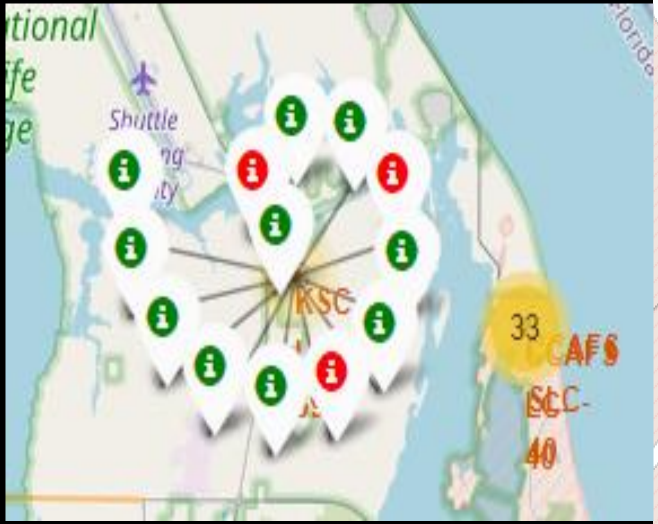


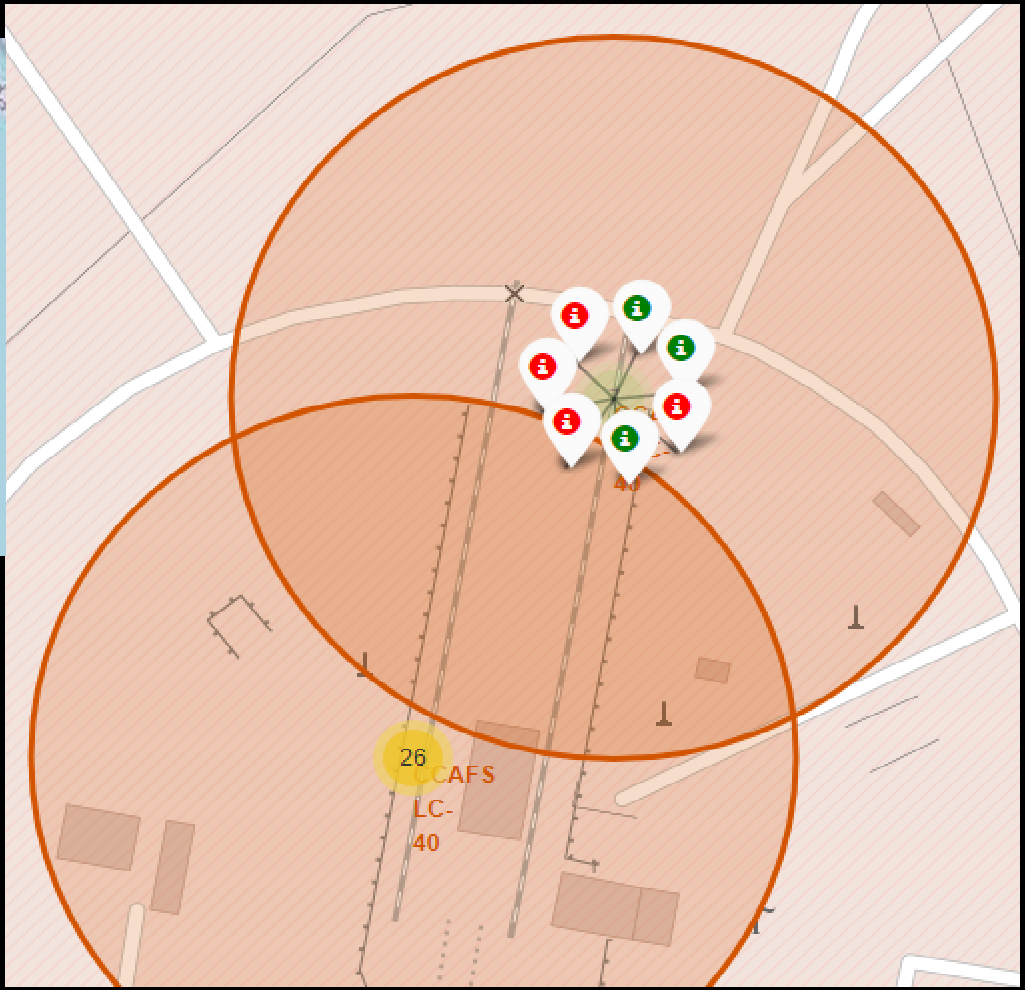
- The Launch Sites are located at coastal regions

IBM Developer

SKILLS NETWORK

Launch Outcomes

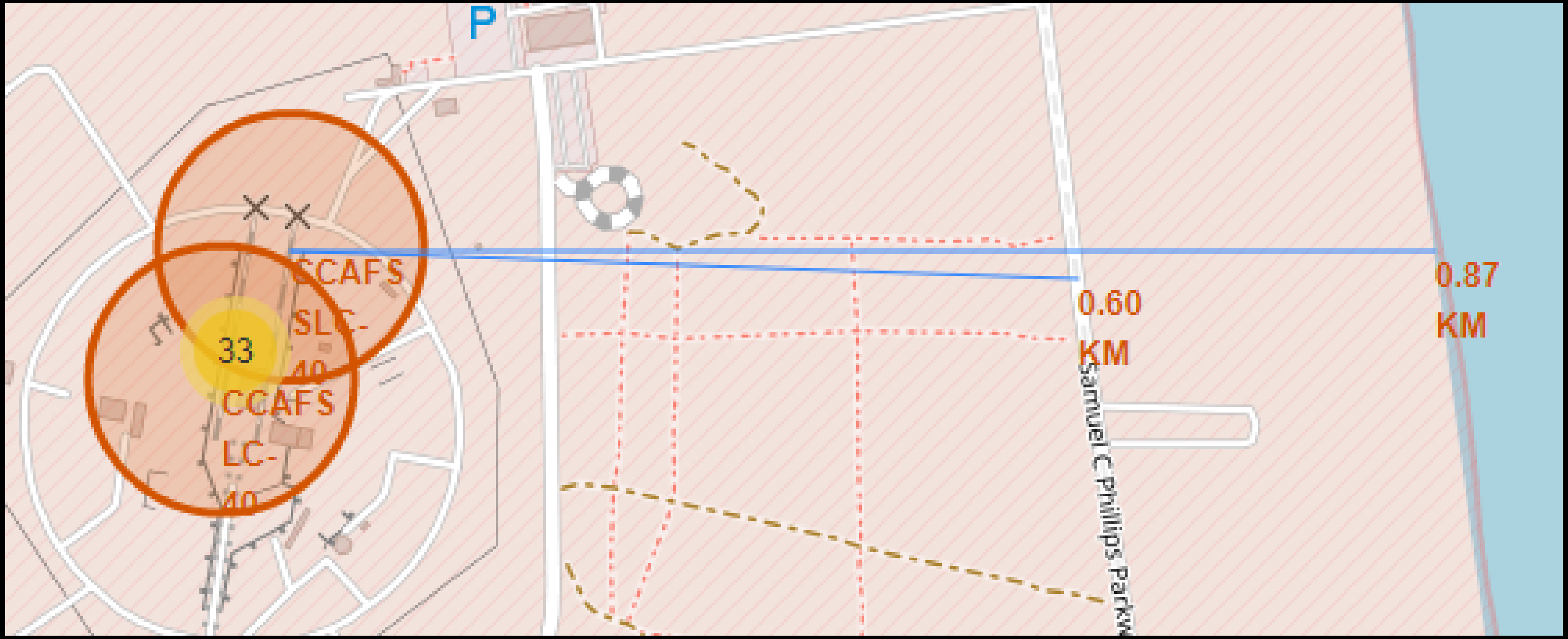




KSC LC-39A launch site outcomes where green markers indicate successful and red markers indicate failed outcomes

Logistics

We can observe that CCAFS SLC-40 launch site has good logistical aspects as it is near to the coastline

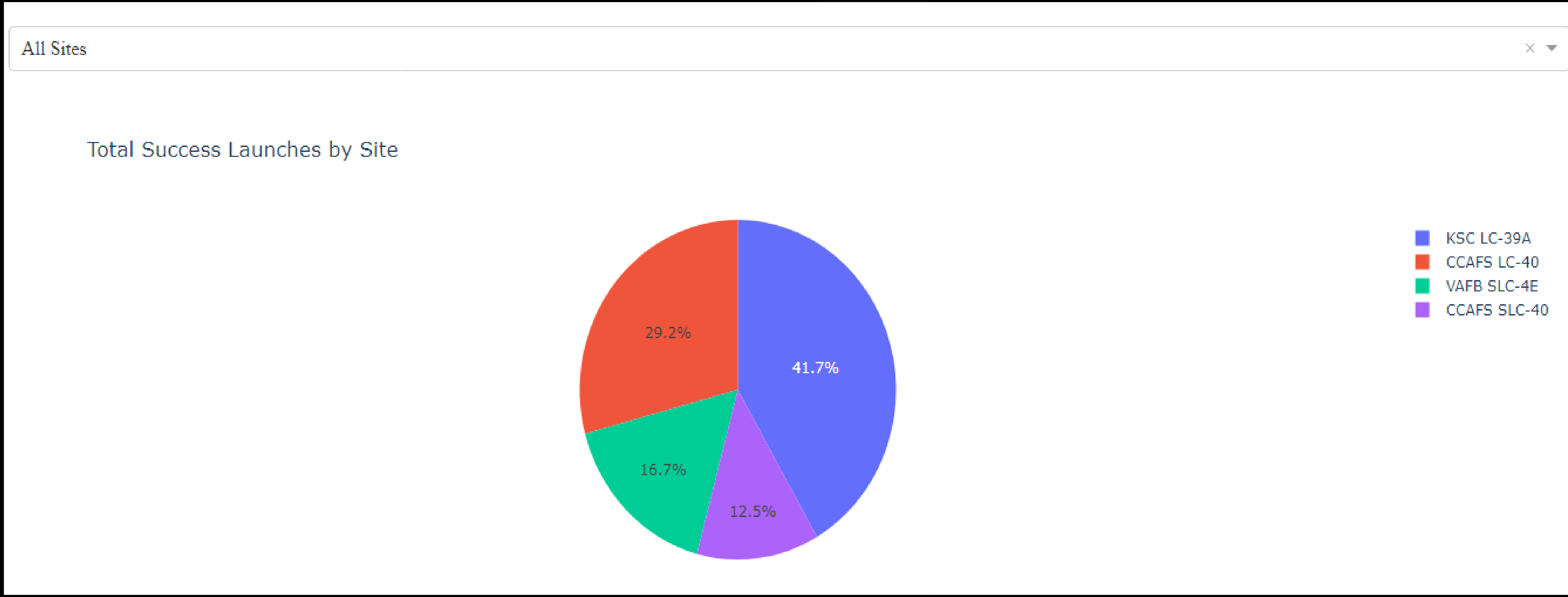


Dashboard

[Dash (cognitiveclass.ai)]

Successful Launches by site

KSC :C-39A has best successful launch rates

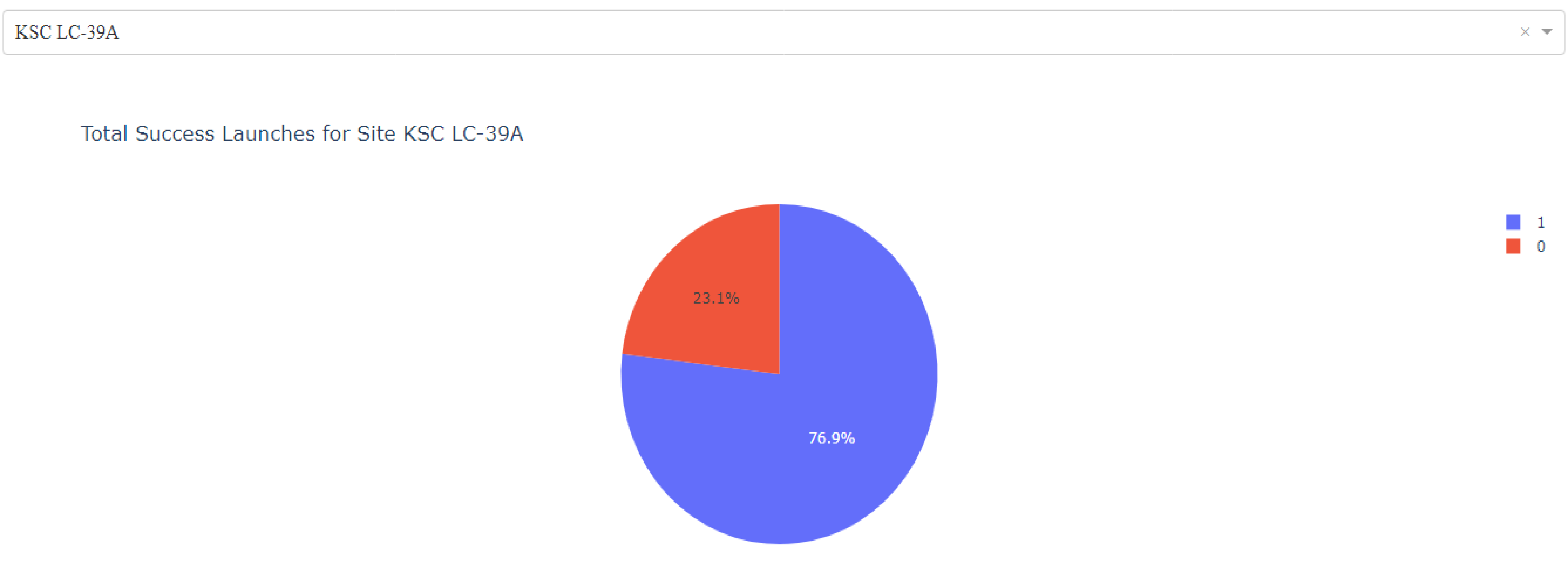


IBM Developer

SKILLS NETWORK

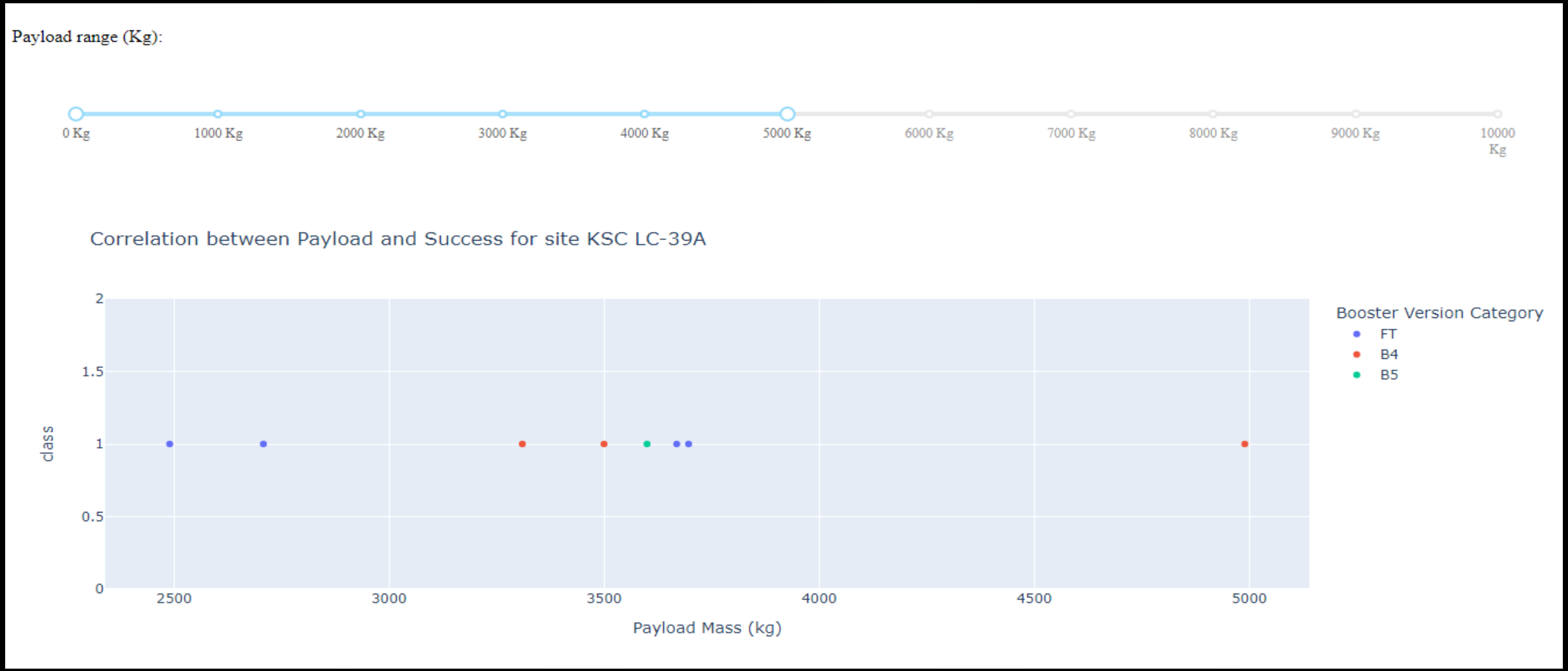
Total successful launches for KSC :C-39A

We can observe that KSC :C-39A has 76.9% successful launch rate



Payload Mass vs Outcome

Low weighted payloads have best success than high weighted payloads



Payload of 0 kg to 5000 kg

IBM Developer

SKILLS NETWORK

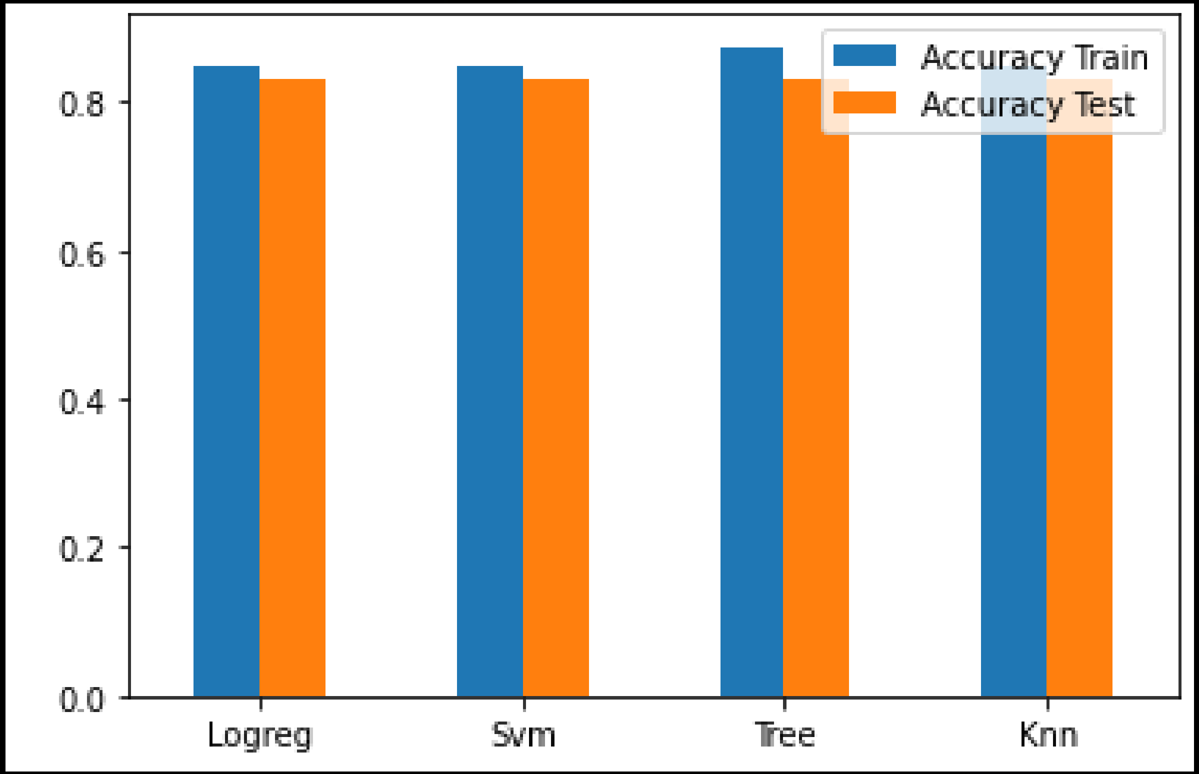
Payload 5000 to 10000 kg

IBM Developer

SKILLS NETWORK

Predictive Analysis

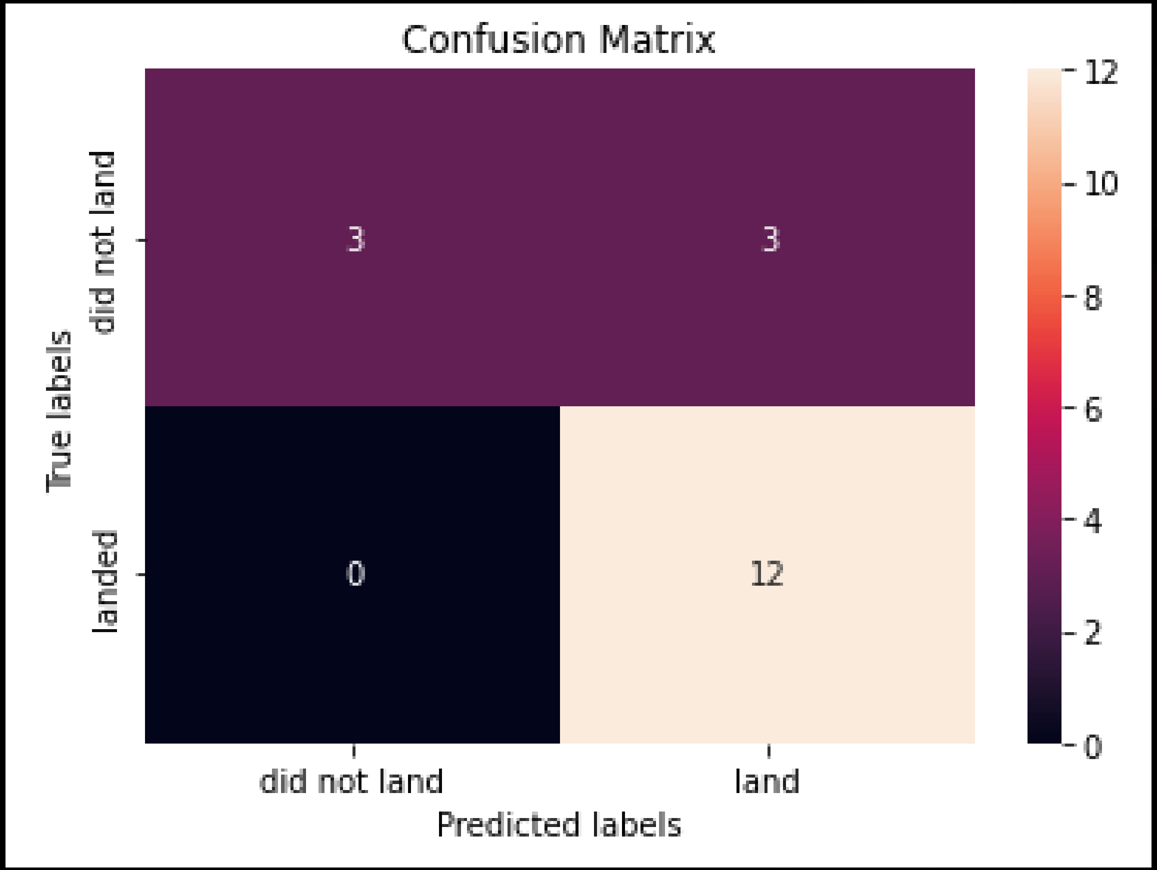
Model Accuracy



For the test data accuracy among all models gave the same results. We can get accurate results using different test size but as per given condition decision tree model has better accuracy compared to rest.

	Models	Accuracy Train	Accuracy Test
0	Logreg	0.846429	0.833333
1	Svm	0.848214	0.833333
2	Tree	0.875000	0.833333
3	Knn	0.848214	0.833333

Confusion Matrix



As the test accuracy of all models are same, the confusion matrix of all 4 models are identical. The model accuracy is less due to the prediction of false positives which is 3 in this case.

CONCLUSION

- The success of this mission can be explained by various factors such as the launch site, the orbit and the number of previous launches.
- GEO, HEO, SSO, ES-L1 are the orbits with best successful launch rates.
- Depending on the orbits, the payload mass can be a criteria for the success of this mission. But the low weighted payloads perform better than the heavy weighted payloads.
- We chose the Decision Tree Algorithm as the best model even if the test accuracy between all the models are identical. We chose Decision Tree Algorithm because it has a better train accuracy.

THANK YOU