DATA SCIENCE CAPSTONE PROJECT

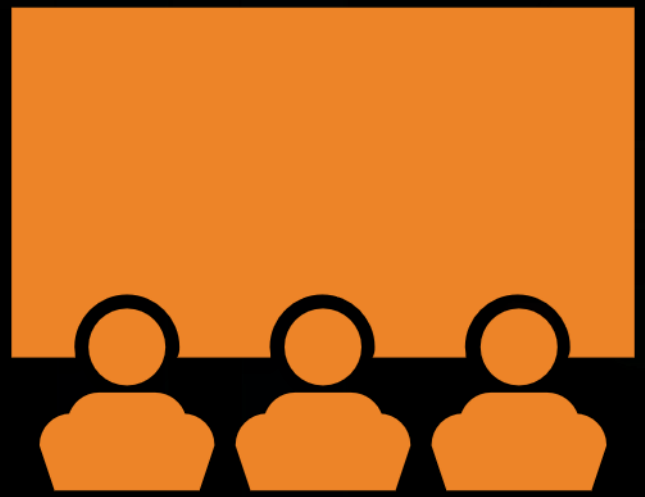
Odiase Eloghosa Paul

June-2024

IBM Developer

SKILLS NETWORK

OUTLINE



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

EXECUTIVE SUMMARY

Summary of Methodologies

- Data collection
- Data wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

INTRODUCTION

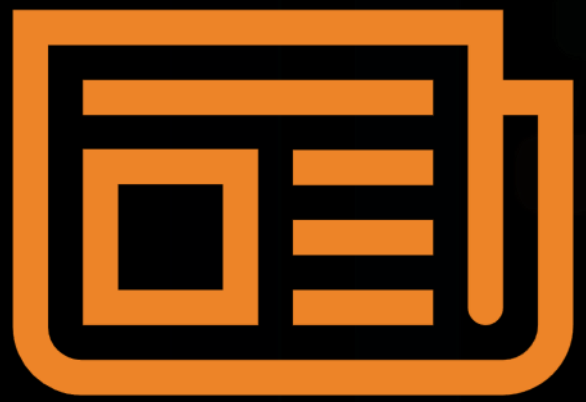
Project background and context:

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Questions to be answered

- - What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.

METHODOLOGY



1. Data collection methodology:
2. Performed data wrangling
3. Performed exploratory data analysis (EDA) using visualization and SQL
4. Performed interactive visual analytics using Folium and Plotly Dash
5. Performed predictive analysis using classification models – Building, tuning and evaluation of classification models to ensure the best results

METHODOLOGY

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

DATA COLLECTION – SPACEX API

- Getting Response from API
- Converting response to a .json file
- Apply custom function to clean data
- Assign list t o a dictionary then a dataframe
- Filter dataframe and export to flat file (.csv)

DATA COLLECTION – WEB SCRAPING

- Getting Response from HTML
- Creating BeautifulSoup Object
- Finding Tables
- Getting Column names
- Creation of dictionary
- Appending data to key
- Converting dictionary to dataframe
- Dataframe to .CSV

DATA WRANGLING

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.

- Perform exploratory data analysis to determine training label
- calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurrence of mission outcome per orbit type
- Creating a landing outcome label from Outcome column

EDA with Data Visualization

Scatter Graphs being drawn:

Flight Number VS. Payload Mass

Flight Number VS. Launch Site

Payload VS. Launch Site

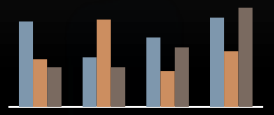
Orbit VS. Flight Number

Payload VS. Orbit Type

Orbit VS. Payload Mass

Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation . Scatter plots usually consist of a large body of data.

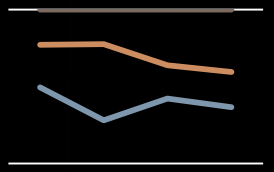
Bar Graph being drawn:



Mean VS. Orbit

A bar diagram makes it easy to compare sets of data between different groups at a glance. The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes. Bar charts can also show big changes in data over time.

Line Graph being drawn:



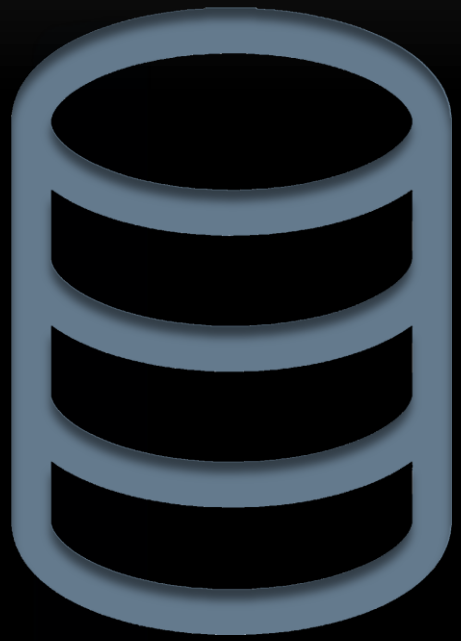
Success Rate VS. Year

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded.

EDA WITH SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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 a



BUILD AN INTERACTIVE MAP WITH FOLIUM

To visualize the Launch Data into an interactive map. We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site.

We assigned the dataframe launch_outcomes(failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()

Using Haversine's formula we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns**. Lines** are drawn on the map to measure distance to landmarks

Example of some trends in which the Launch Site is situated in.

- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

BUILD A DASHBOARD WITH PLOTLY DASH

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

PREDICTIVE ANALYSIS (CLASSIFICATION)

- Creating a NumPy array from the column “Class” in data
- Standardizing the data with StandardScaler, then fitting and transforming it
- Splitting the data into training and testing sets with train_test_split function
- Creating a GridSearchCV object with cv = 10 to find the best parameters
- Finding the method performs best by examining the Jaccard_score and F1_score metrics
- Examining the confusion matrix for all models
- Calculating the accuracy on the test data using the method .score() for all models
- Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

IBM Developer

SKILLS NETWORK

RESULTS

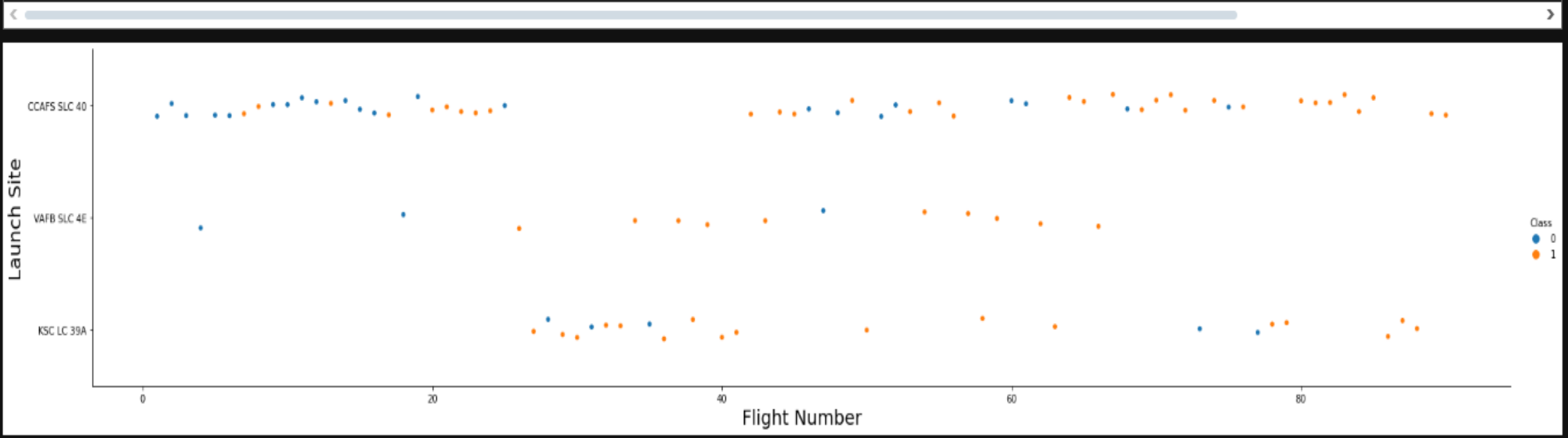
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

EDA WITH VISUALIZATION

IBM Developer

SKILLS NETWORK

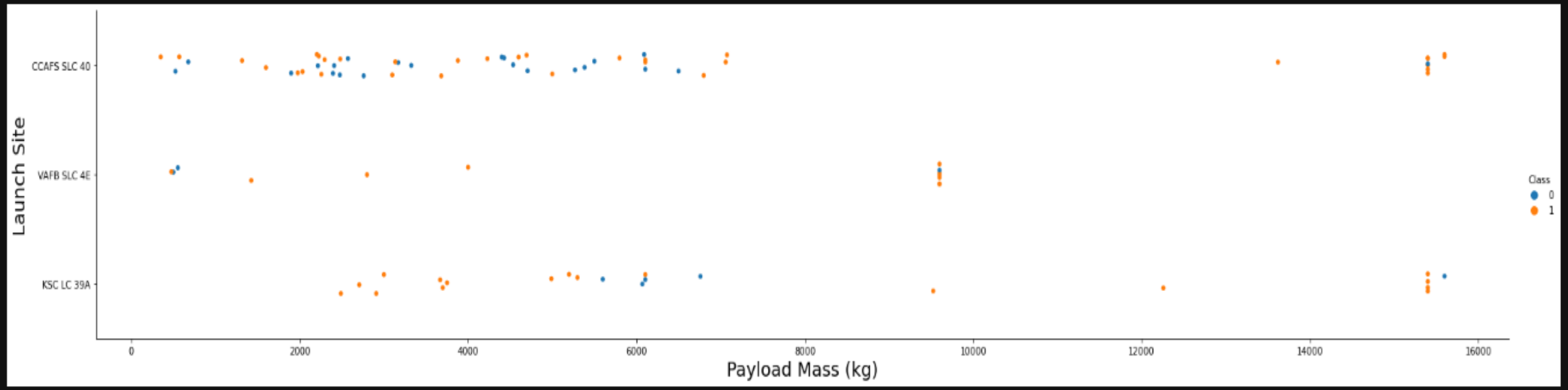
FLIGHT NUMBER VS. LAUNCH SITE



Explanation:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success

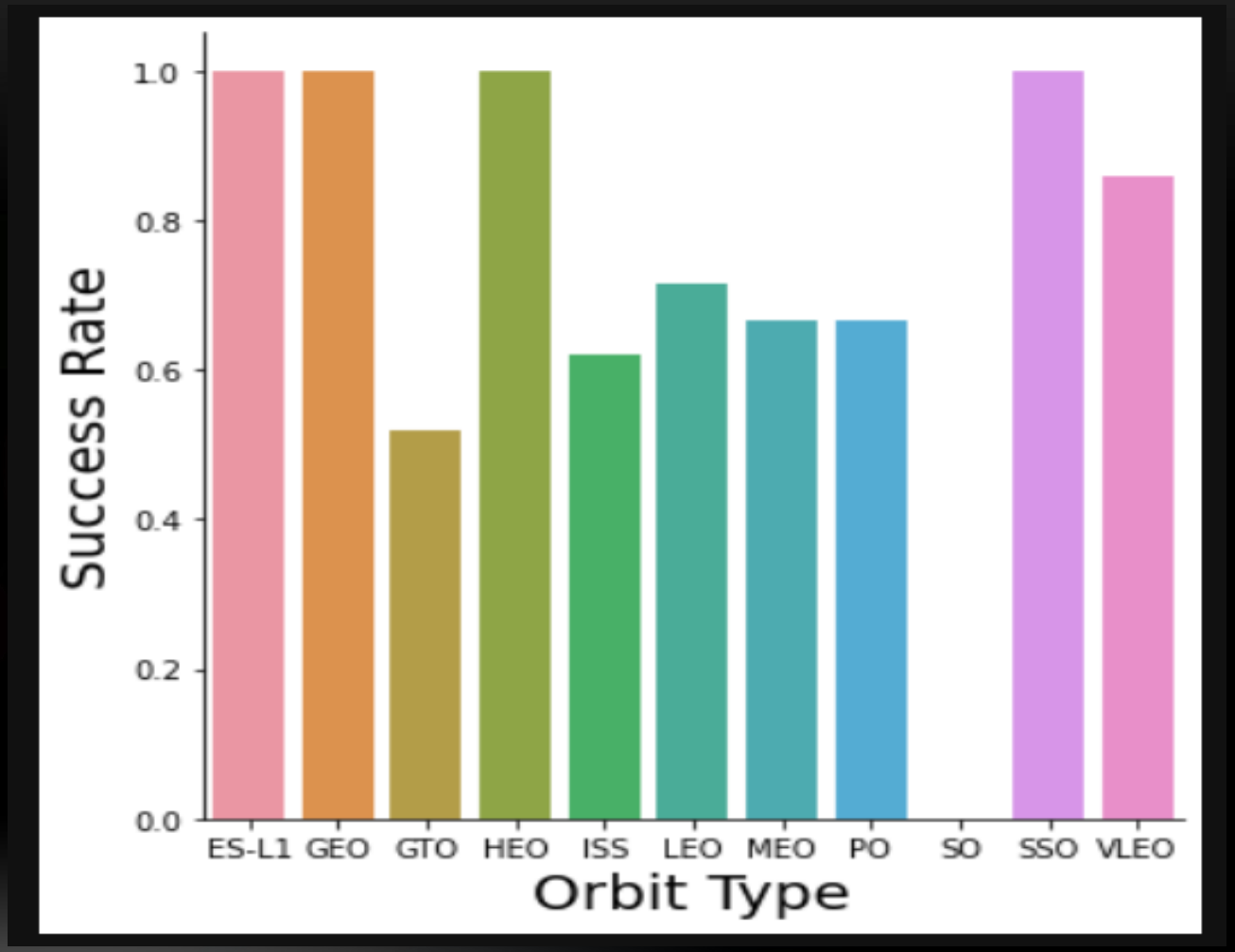
PAYLOAD VS. LAUNCH SITE



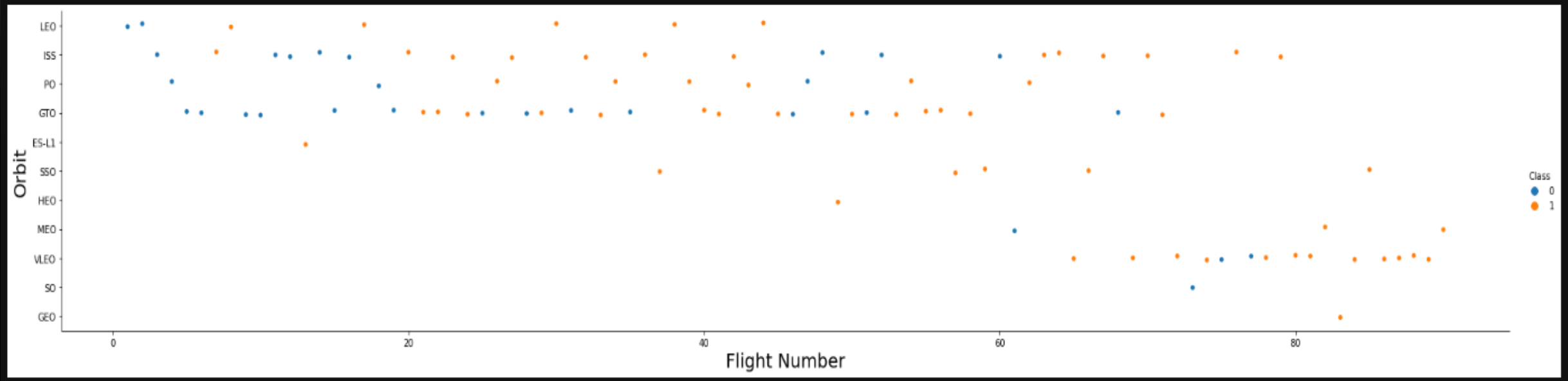
The greater the payload mass for Launch Site CCAFS SLC 40 the higher the success rate for the Rocket. There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

SUCCESS RATE VS. ORBIT TYPE

Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate



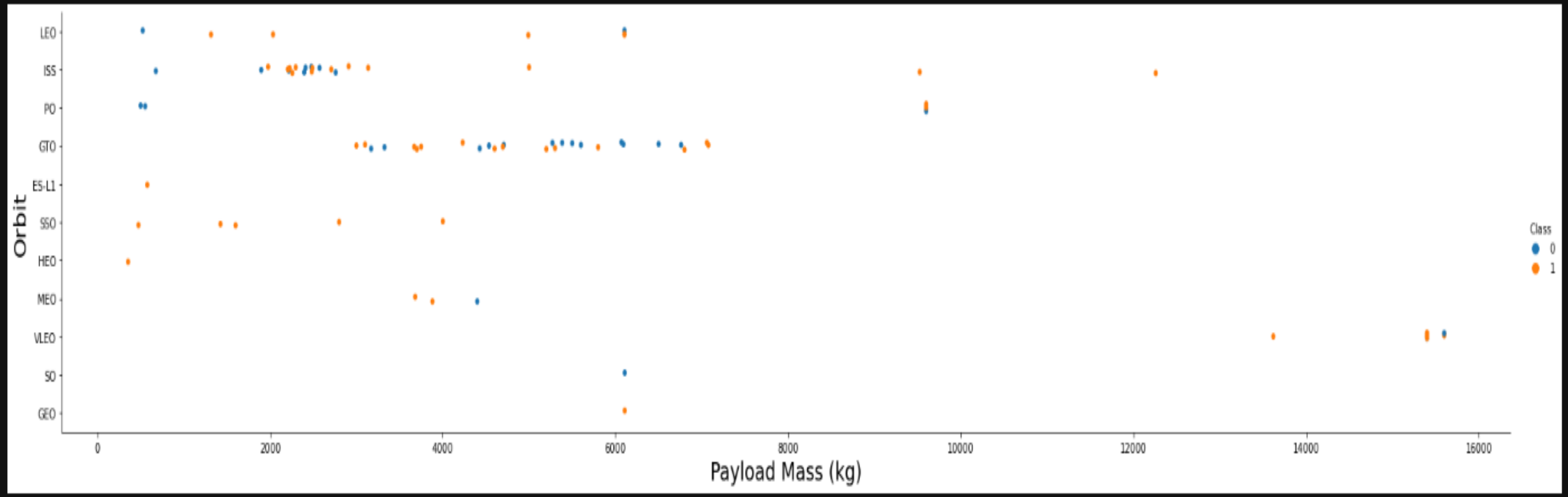
FLIGHT NUMBER VS. ORBIT TYPE



Explanation:

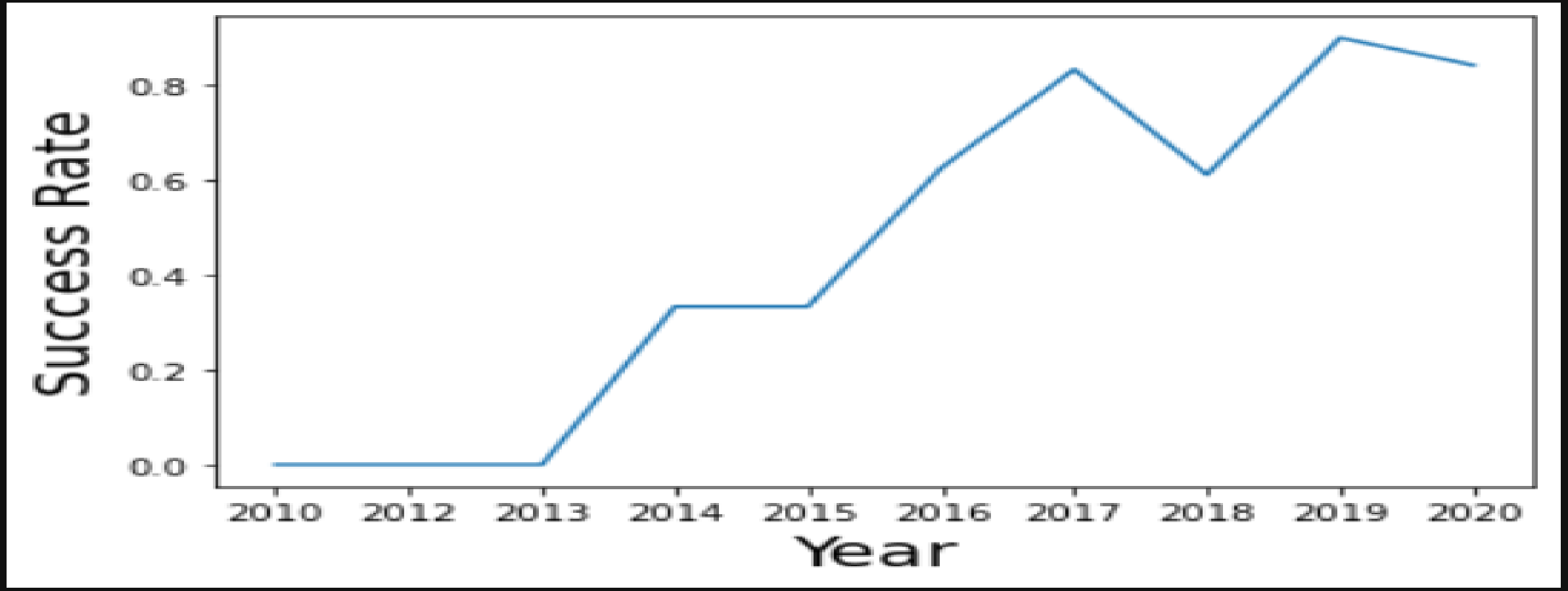
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

PAYLOAD MASS VS. ORBIT TYPE



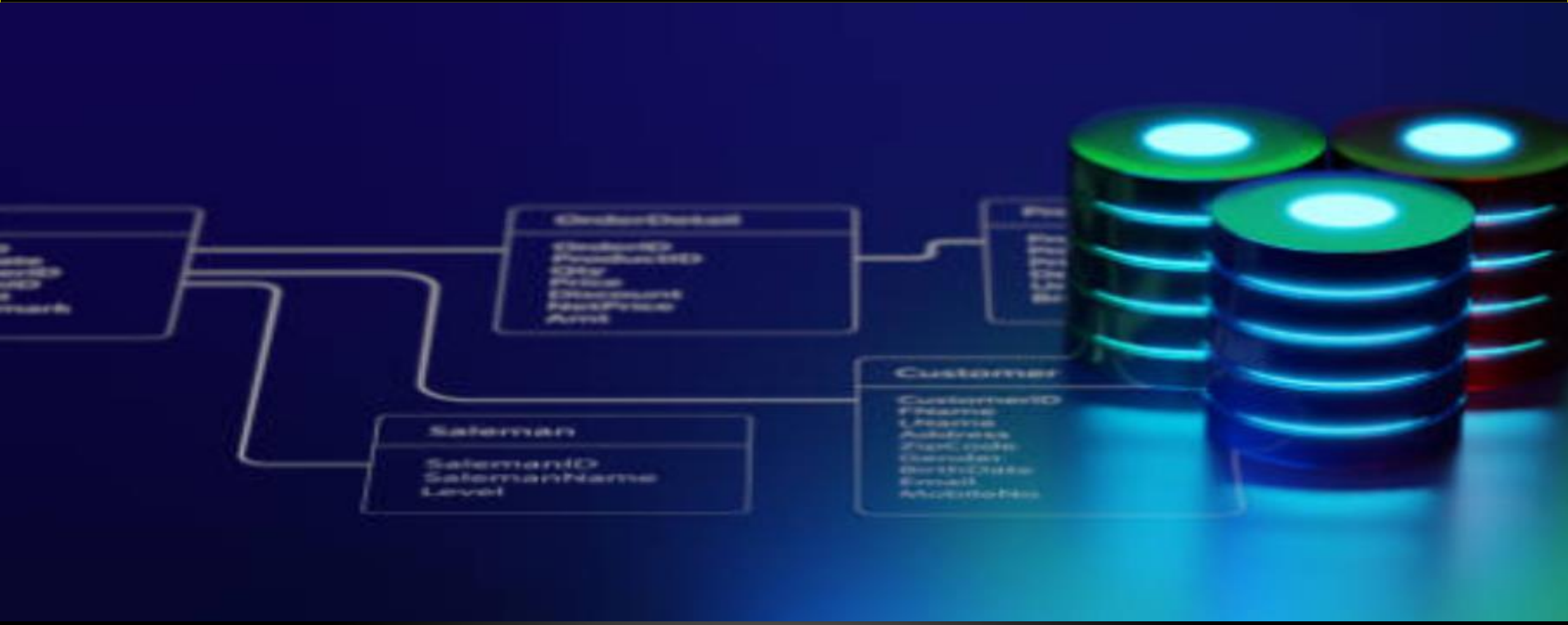
Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

LAUNCH SUCCESS YEARLY TREND



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

EDA WITH SQL



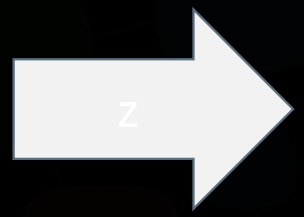
IBM Developer

SKILLS NETWORK

LAUNCH SITE NAMES

SQL QUERY

select DISTINCT Launch_Site from tblSpaceX



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

Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from tblSpaceX

LAUNCH SITE NAMES BEGIN WITH `CCA`

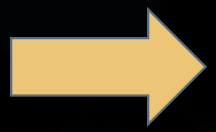
DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Displaying 5 records where launch sites begin with the string 'CCA'.

TOTAL PAYLOAD MASS

SQL QUERY

```
Select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass
from tblSpaceX
where Customer = 'NASA (CRS)'",'TotalPayloadMass
```



total_payload_mass
45596

QUERY EXPLAINATION

Using the function **SUM** summates the total in the column **PAYLOAD_MASS_KG_**

The **WHERE** clause filters the dataset to only perform calculations on Customer **NASA (CRS)**

AVERAGE PAYLOAD MASS BY F9 V1.1

average_payload_mass
2534

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

FIRST SUCCESSFUL GROUND LANDING DATE

first_successful_landing
2015-12-22

- Listing the date when the first successful landing outcome in ground pad was achieved.

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
%sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
%sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

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

- Listing the total number of successful and failure mission outcomes.

BOOSTERS CARRIED MAXIMUM PAYLOAD

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

- Listing the names of the booster versions which have carried the maximum payload mass.

2015 LAUNCH RECORDS

```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

landing__outcome	count_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

- Ranking 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.

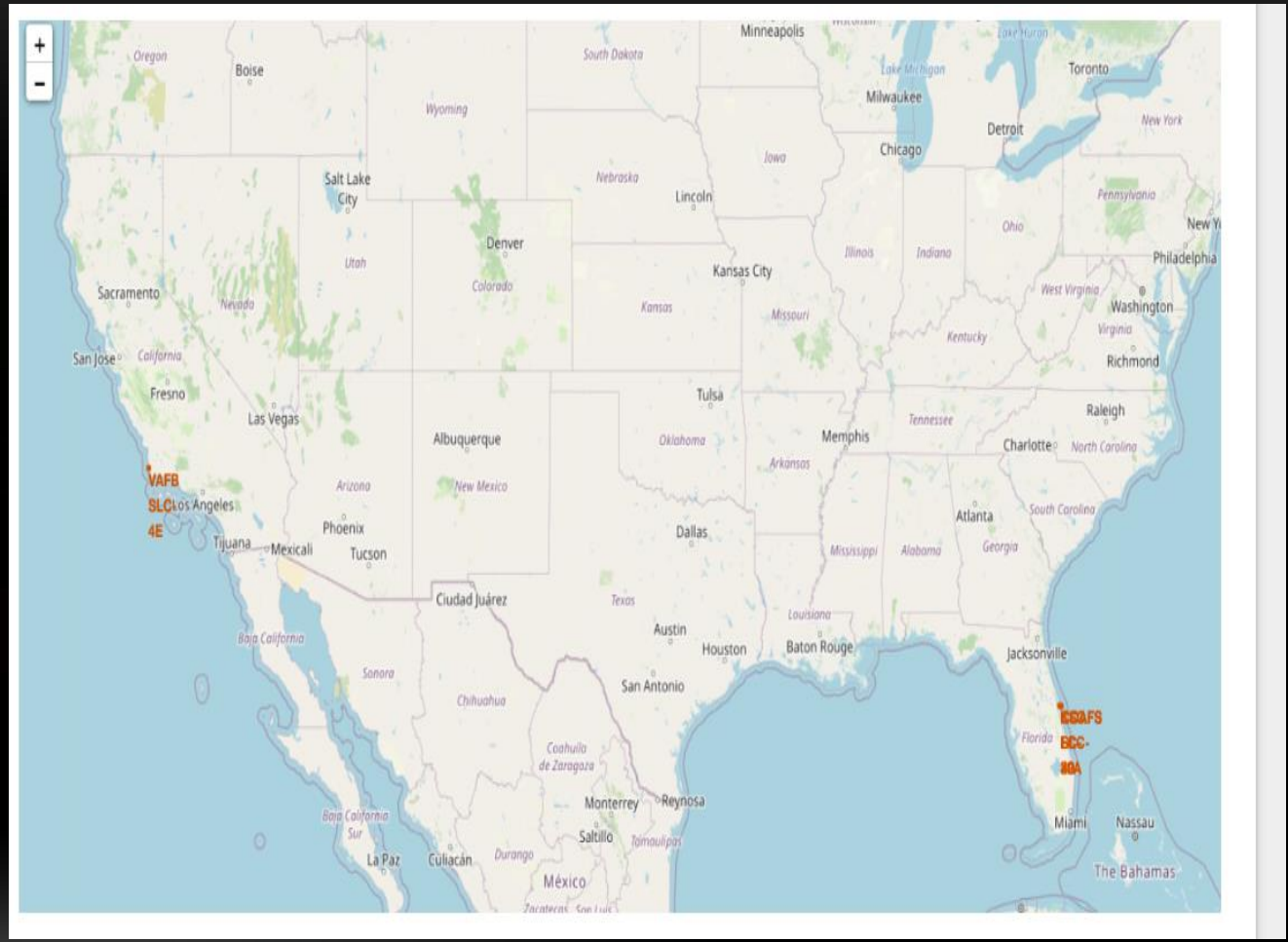
INTERACTIVE MAP WITH FOLIUM

IBM Developer

SKILLS NETWORK

ALL LAUNCH SITES GLOBAL MAP MARKERS

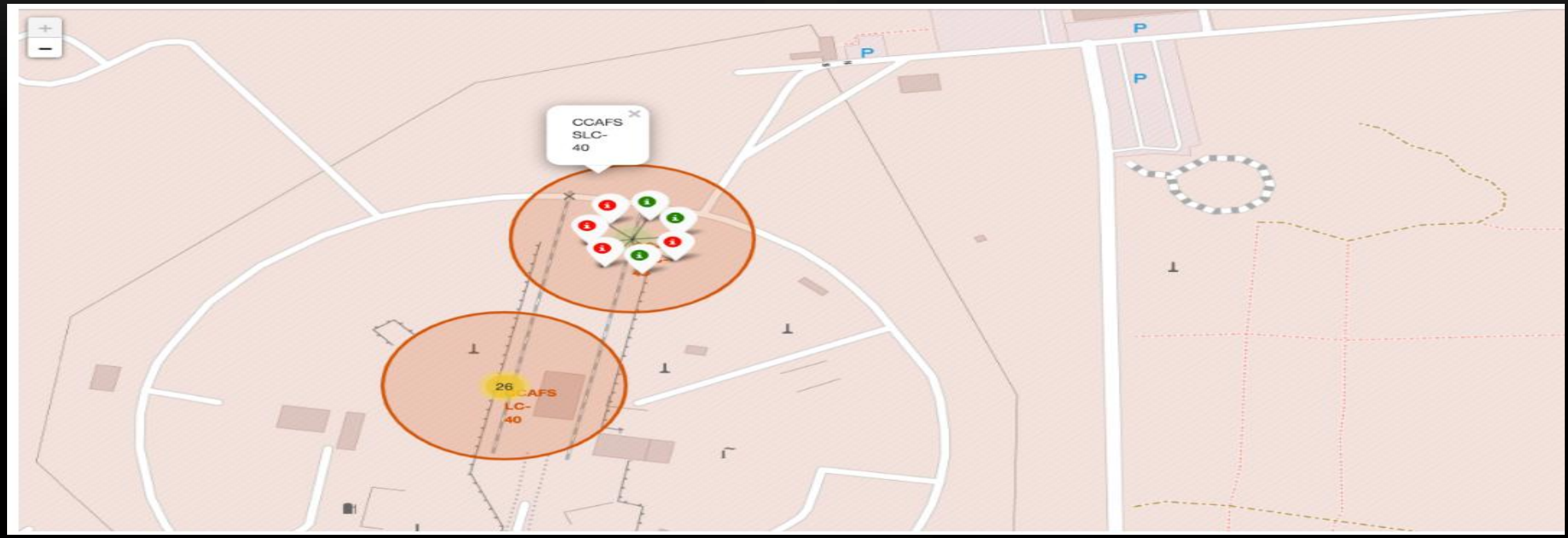
Most of Launch sites are in proximity to the Equator line. The land is moving faster at the equator than any other place on the surface of the Earth. Anything on the surface of the Earth at the equator is already moving at 1670 km/hour. If a ship is launched from the equator it goes up into space, and it is also moving around the Earth at the same speed it was moving before launching. This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit



IBM Developer

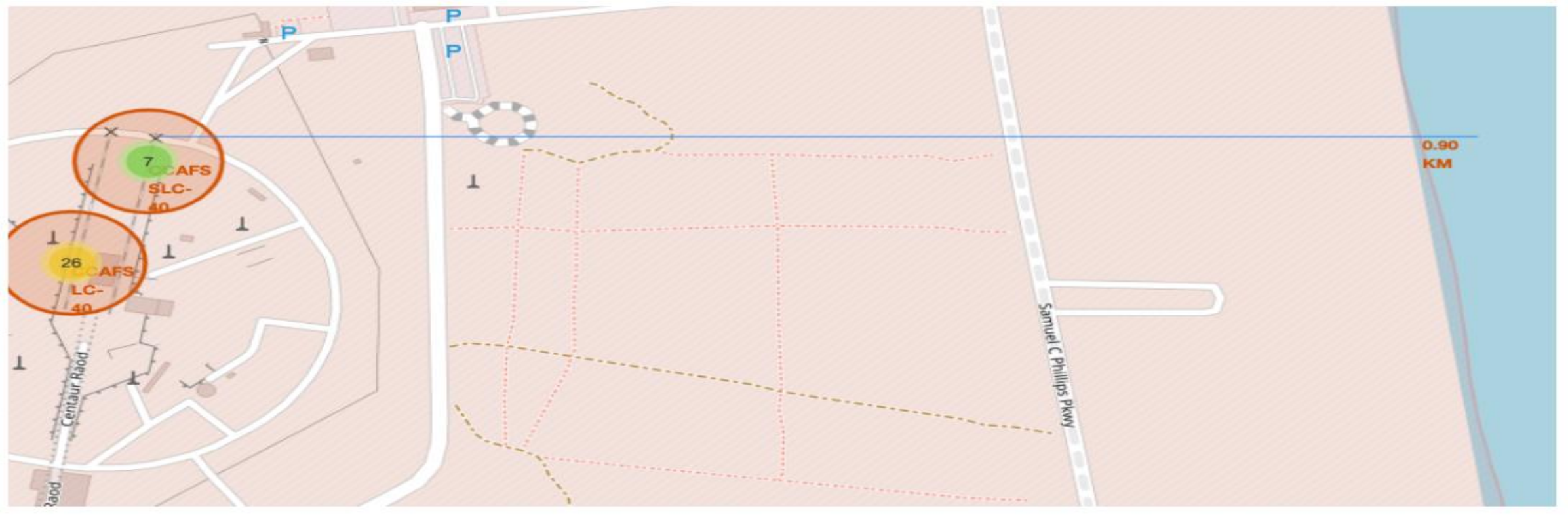
SKILLS NETWORK

COLOR LABELED MARKERS



Green Marker shows successful Launches and Red Marker shows Failures

WORKING OUT LAUNCH SITES DISTANCE TO LANDMARKS TO FIND TRENDS



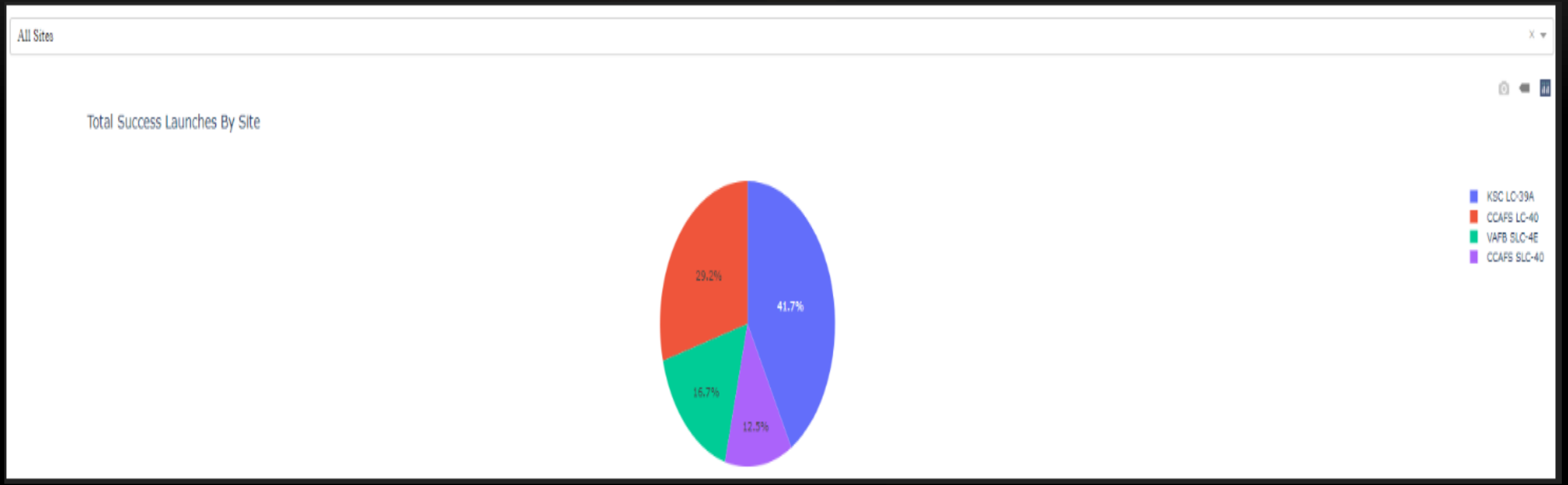
Distance to coast

DASHBOARD WITH PLOTLY DASH

IBM Developer

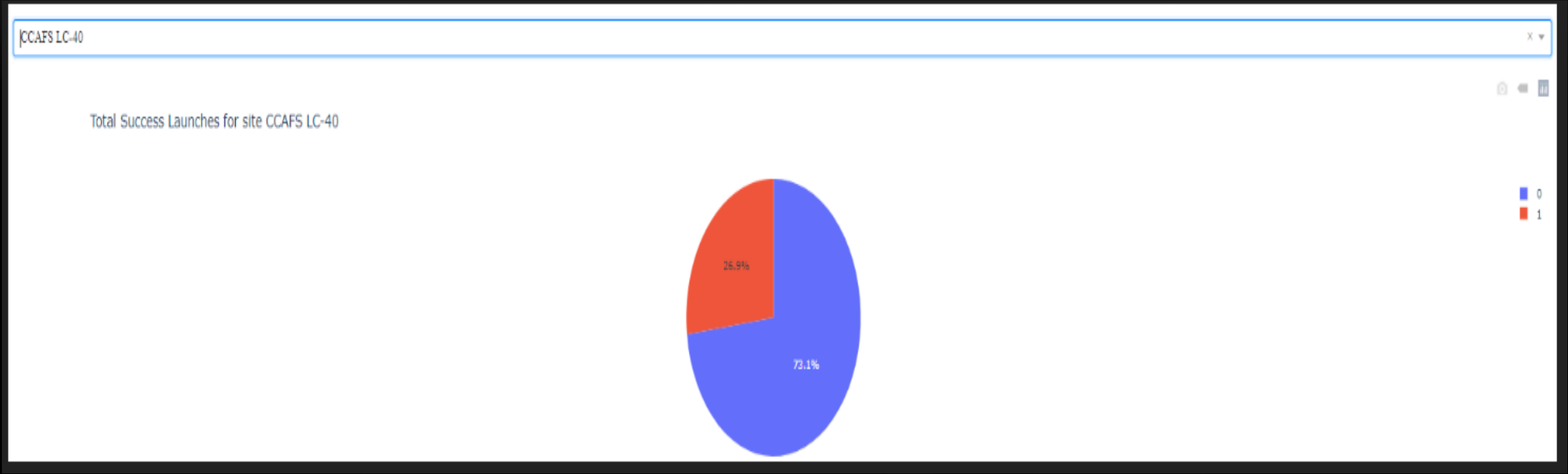
SKILLS NETWORK

LAUNCH SUCCESS COUNT FOR ALL SITES



- We can see that KSC LC-39A had the most successful launches from all the site

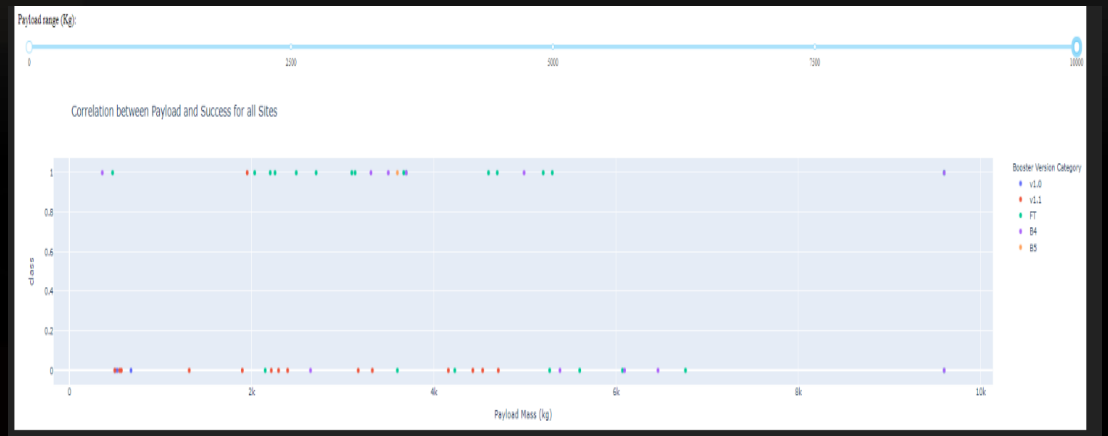
LAUNCH SITE WITH HIGHEST LAUNCH SUCCESS RATIO

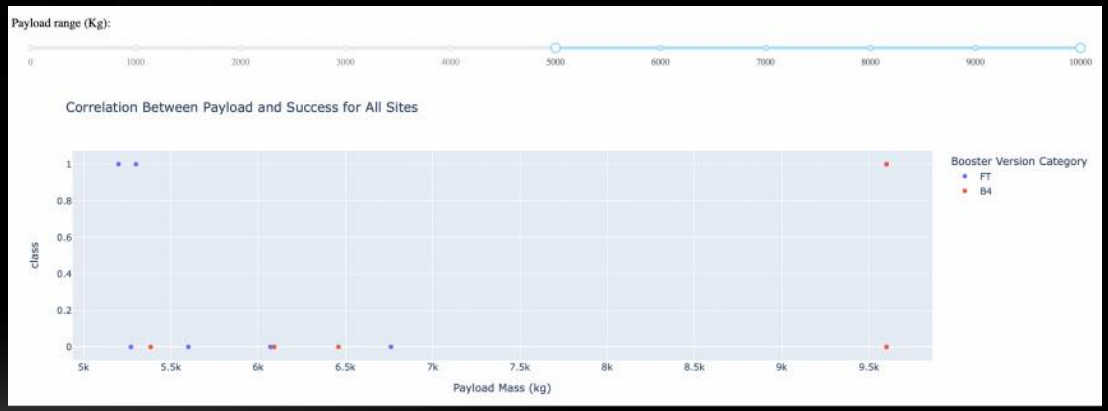


- KSC LC-39A achieved a 73.1% success rate while getting a 26.9% failure rate

PAYLOAD MASS VS. LAUNCH OUTCOME FOR ALL SITES

The charts show that payloads between 2000 and 5500 kg have the highest success rate.





PREDICTIVE ANALYSIS (CLASSIFICATION)

IBM Developer

SKILLS NETWORK

CLASSIFICATION ACCURACY

Scores and Accuracy of the Test Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

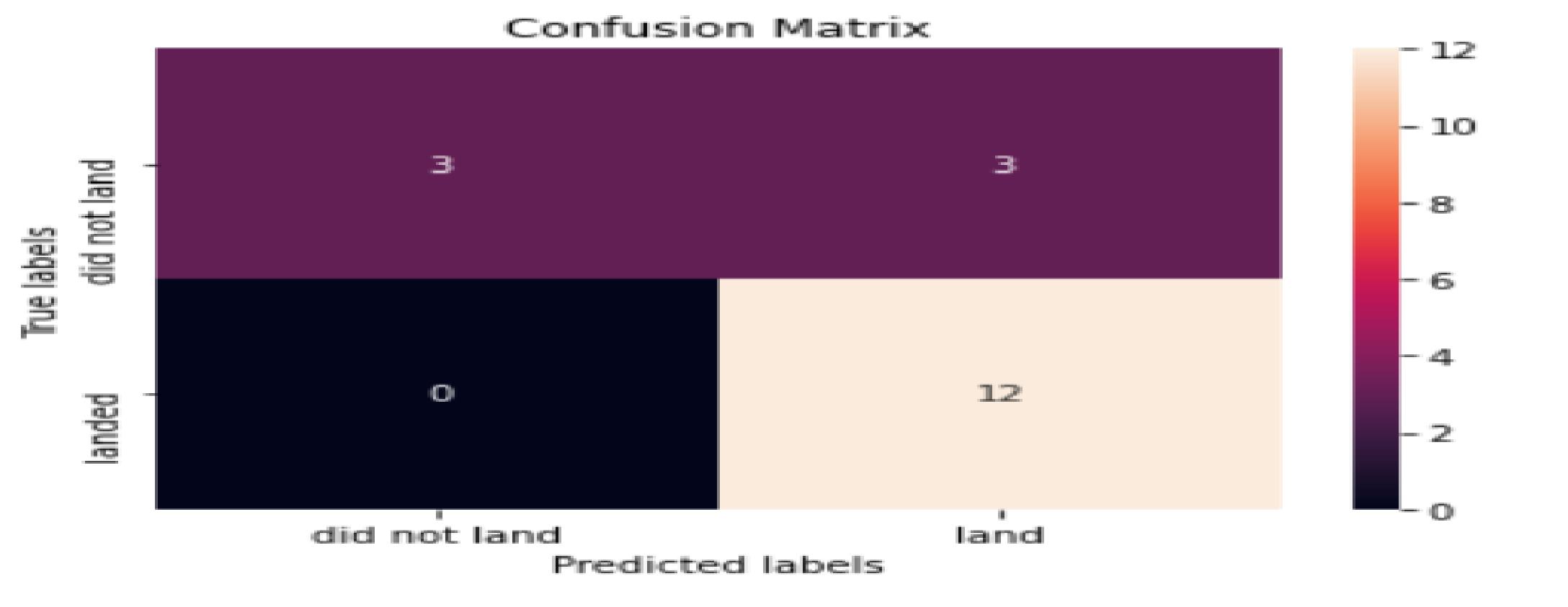
Scores and Accuracy of the Entire Data Set

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Explanation:

- Based on the scores of the Test Set, we can not confirm which method performs best.
- Same Test Set scores may be due to the small test sample size (18 samples). Therefore, we tested all methods based on the whole Dataset.
- The scores of the whole Dataset confirm that the best model is the Decision Tree Model. This model has not only higher scores, but also the highest accuracy.

CONFUSION MATRIX



- Explanation: • Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

CONCLUSION

- Decision Tree Model is the best algorithm for this dataset.
- Launches with a low payload mass show better results than launches with a larger payload mass. •
- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.