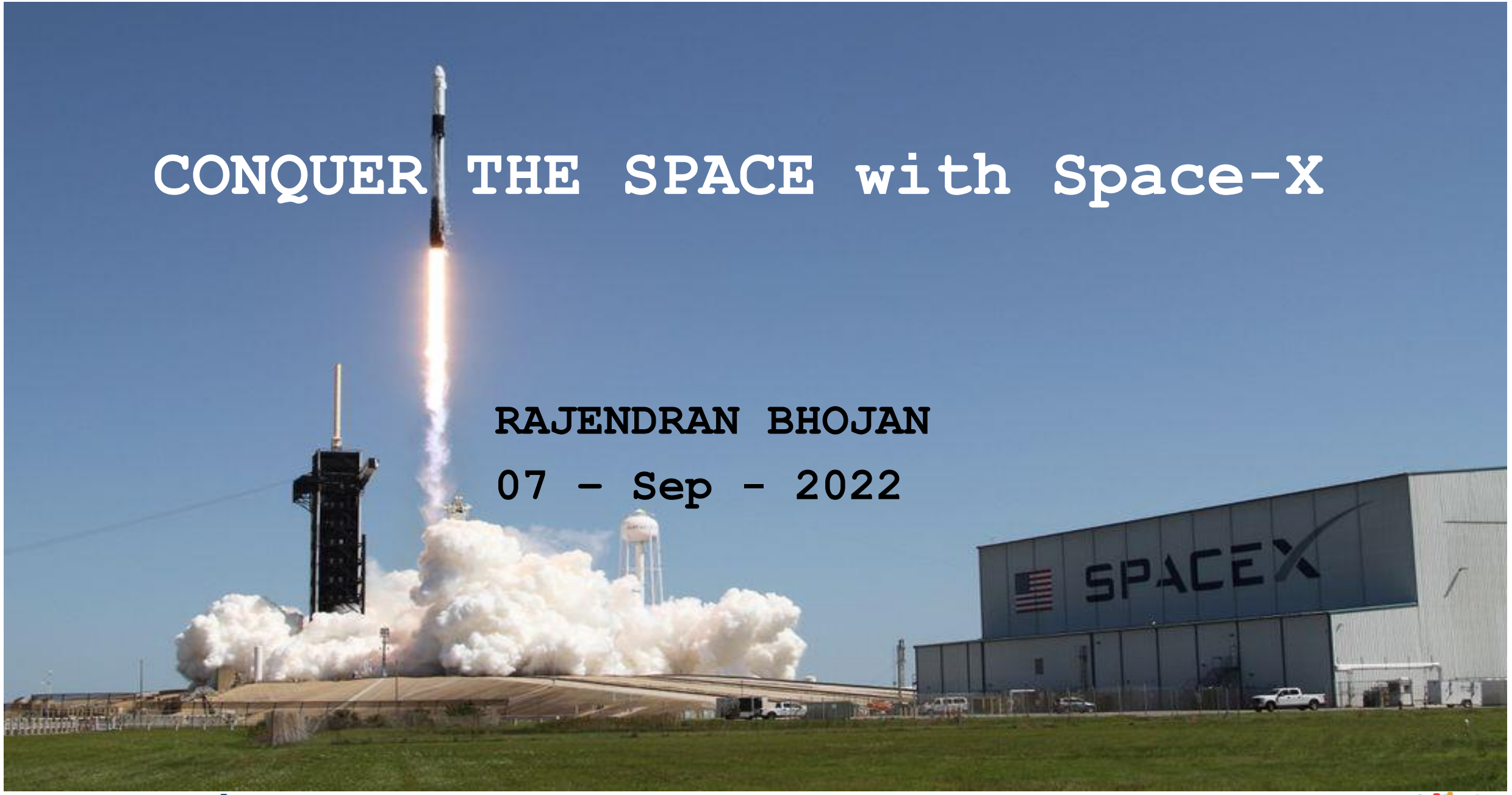
CONQUER THE SPACE with Space-X

RAJENDRAN BHOJAN

07 - Sep - 2022



IBM Developer

SKILLS NETWORK

OUTLINE

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

EXECUTIVE SUMMARY

- Methodologies Summary
  - Collecting Data through API
  - Collecting Data through Web Scrapping
  - Data Wrangling
  - Using SQL for Exploratory Data Analysis
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Results Summary
  - Results of Exploratory Data Analysis
  - Screenshots of Interactive Analytics
  - Results of Predictive Analytics from Machine Learning Lab

INTRODUCTION

Rocket revolution company, Space-X having disrupted the space industry by making available in launching rockets specifically Falcon 9 as minimum as $62 million even as other companies provide $165 million each. Space-X startling ideas to reuse the first stage launching by re-landing the rocket to be used for next mission, has done with most savings.

Launch mission prices go down further by repeating this process. My objective of this project is to create the machine learning pipeline to analyse and predict future outcomes of first stage landing as a data scientist of a rival start-up. This is an important project to identify the price right for bidding against Space-X for launching rockets.

The issues are:

- To identify all factors that influence landing outcome.
- The links between every variables and how those affect the outcomes.
- The best probability needed to increase the accuracy of a successful landing


Section 1
Methodology

METHODOLOGY

Executive Summary

- Data collection
  - Data collection using SpaceX REST API and web scrapping
- Data wrangling
  - Data processing done by one-hot encoding for categorical features
- To perform exploratory data analysis (EDA) using visualization and SQL
- To Perform interactive visual analytics using Folium and Plotly Dash
- To Perform predictive analysis using classification models
  - To build, tune, evaluate classification model

DATA COLLECTION

The process of data collection is gathering data and measuring information on intended variables through the system established, which then facilitates one to answer important questions and evaluate outcomes. Dataset was collected by REST API and Web Scrapping from Wikipedia

For performing REST API, done using get request. Next, decoded the response content as Json and converted to pandas dataframe using json_normalize(). Next, cleaned data, checked for missing values and filled with whatever required.

For web scrapping, used BeautifulSoup to select launch records as HTML table, parsing the table and converting into pandas dataframe for further analysis.

Data Collection – SpaceX API

Get request for rocket launch data using API

Use json_normalize method to convert json result to dataframe

Perform data cleaning and filling for the missing values

Link:

https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final_Assignment_C10_RajendranB.ipynb

Data Collection – Web Scraping

Get by request the Falcon9 Launch Wiki page from given url

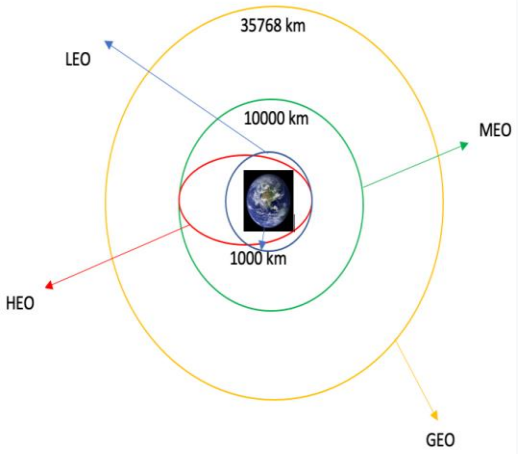
Create a BeautifulSoup from HTML response

Extract variable names of all columns from the HTML header

Link:

https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final_Assignment_Lab02_DataCollection_WebScrapping.ipynb

Data Wrangling



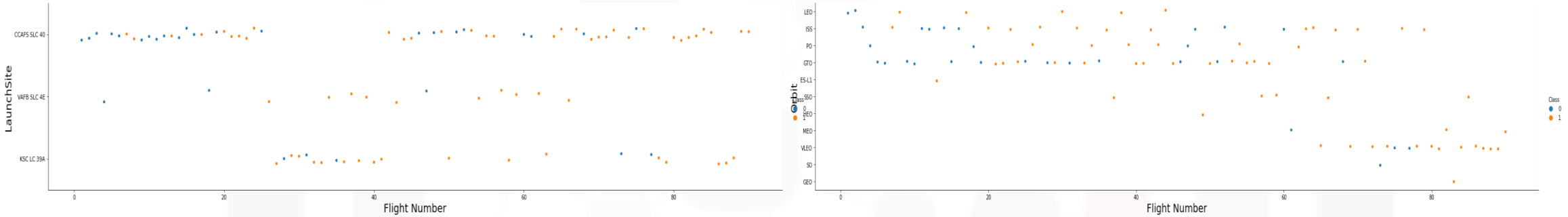
Data Wrangling is the process of cleaning and unifying messy and complex data sets for easy access and Exploratory Data Analysis (EDA).

We will first calculate the number of launches on each site, then calculate the number and occurrence of mission outcome per orbit type.

We then create a landing outcome label from the outcome column. This will make it easier for further analysis, visualization, and ML. Lastly, we will export the result to a CSV

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20Lab03%20Data%20Wrangling%20RB.ipynbb

EDA with Data Visualization

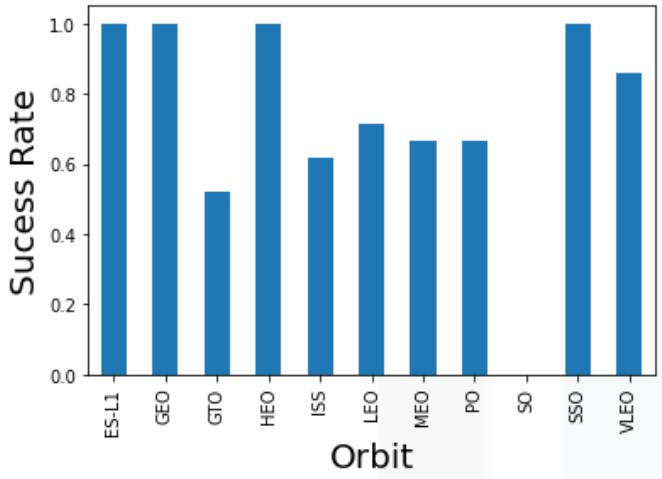


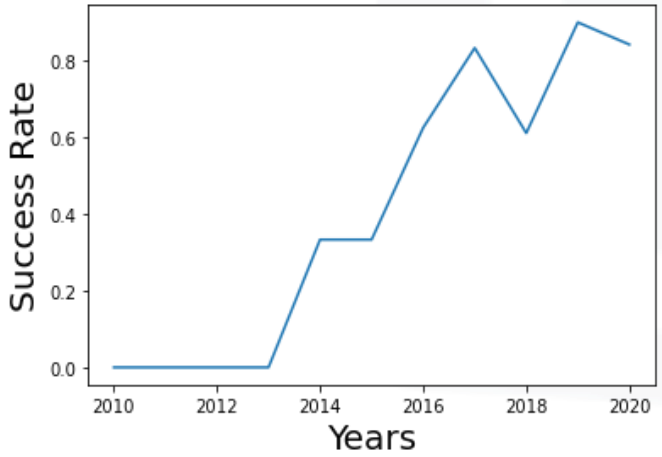
- **We first begin by using scatter graph to find the relationship between the attributes:**
- **Payload and Flight Number.**
- **Launch Site vs. Flight Number.**
- **Payload and Launch Site.**
- **Flight Number vs. Orbit Type.**
- **Payload and Orbit Type.**

Scatter plots show dependency of attributes on each other. Once a pattern is determined from the graphs. It's very easy to see which factors affecting the most to the success of the landing outcomes

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20W02%20Lab02%20EDA%20with%20Visuali%20RB.ipynb

EDA with Data Visualization





While we get few insights of the relationships using scatter plots, We will then use visualization tools such as bar graph and line plots for further analysis. Bar graphs is one of the easiest way to portray the relationships between the attributes. In this case, we will use the bar graph to determine which orbits have the highest probability of success.

We then use the line graph to show a trends or pattern of the attribute over time which in this case, is used for see the launch success yearly trend. We then use Feature Engineering to be used in success prediction in the future module by created the dummy variables to categorical columns.

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20W02%20Lab02%20EDA%20with%20Visuali%20RB.ipynb

EDA using SQL

Using SQL, we had performed many queries to incur good understanding of dataset as:

- Displaying the names of the launch sites.
- Displaying 5 records where launch sites begin with the string 'CCA'.
- Displaying the total payload mass carried by booster launched by NASA (CRS).
- Displaying the 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 sites names for in year 2015.
- Rank the count of landing outcomes or success between the date 2010-06-04 and 2017-03-20, in descending order.

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20W02%20Lab01%20EDA%20with%20SQL%20RB.ipynb

Building Interactive Map using Folium

In order to visualize the launch data into an interactive map, we take the latitude and longitude coordinates at each launch site and add a circle marker around each launch site with a label of the name of the launch site.

We assign the dataframe launch outcomes(failure,success) to classes 0 and 1 with Red and Green markers on the map in MarkerCluster().

We then use the Haversine's formula to calculate the distance of the launch sites of various landmarks to find results to the questions:

- How close the launch sites with railways, highways and coastlines?
- How close the launch sites with nearby cities?

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20W03%20Lab01%20LaunchSite%20LocAF%20RB.ipynb

Building Dashboard using Plotly

- We built an interactive dashboard with Plotly dash which allowing the user to play around with the data as they need.
- We plotted pie charts showing the total launches by a certain sites.
- We then plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/DashBoard_SpaceX_RajendranBhojan.py

Predictive Analysis (Classification)

Model Build

- Load the dataset into NumPy and Pandas
- Transform the data and then split into training and test datasets
- Decide which type of ML to use
- set the parameters and algorithms to GridSearchCV and fit it to dataset.

Model Evaluation

- Check the accuracy for each model
- Get tuned hyper parameters for each type of algorithms.
- plot the confusion matrix.

Model Improvement

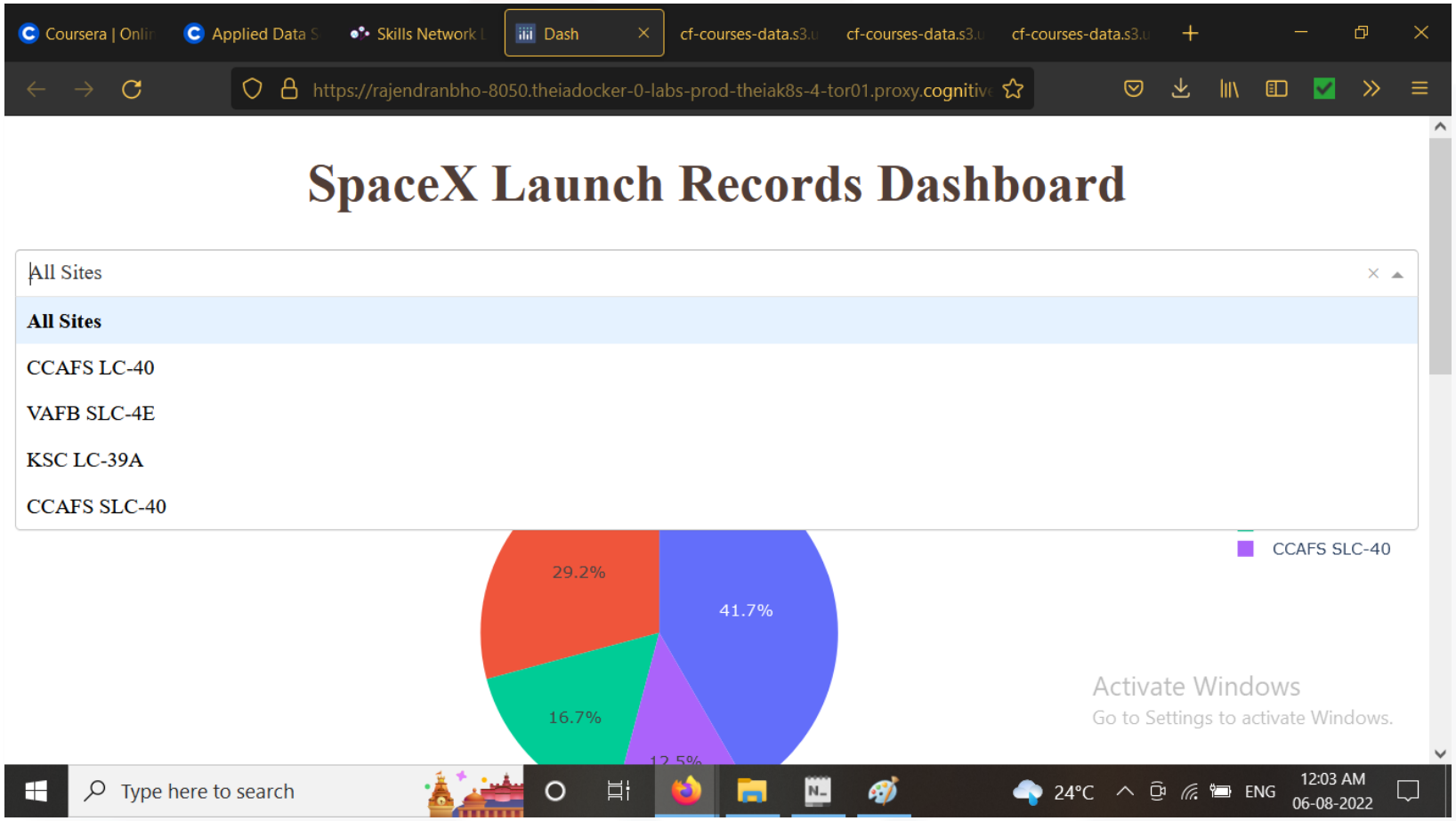
- Use Feature Engineering and Algorithm Tuning

Find Best Model

- The model with the best accuracy score will be the best performing model.

Link: https://github.com/RajendranBhojan/Test-Repo-Course-10/blob/master/Final%20Assignment%20C10%20W04%20Lab01%20MachineLearningPred.ipynb

Space X Launch Records Dashboard



RESULTS

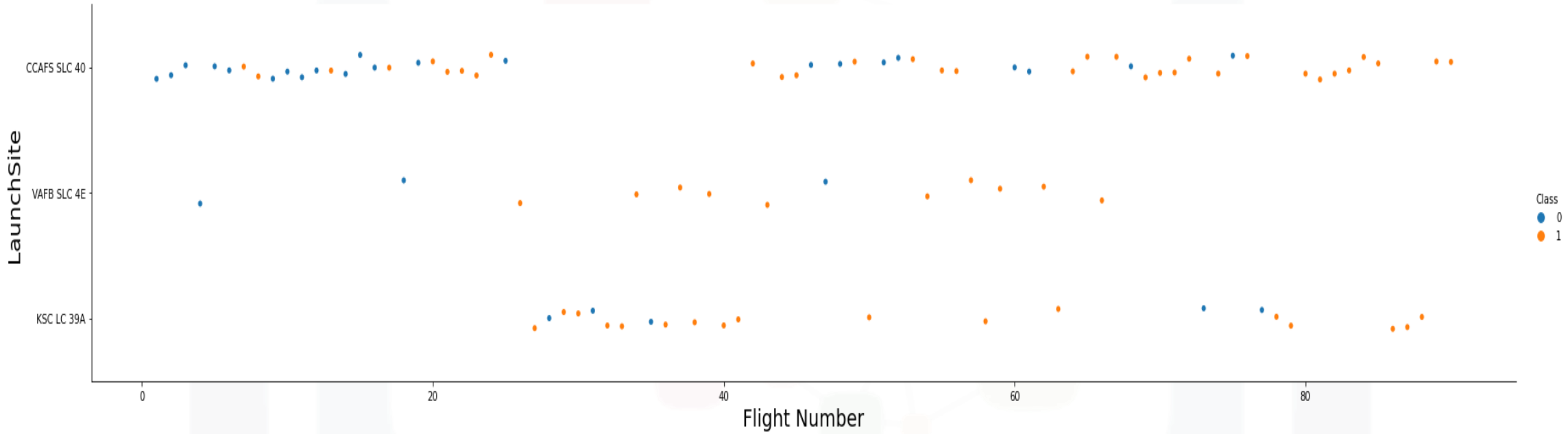
The results will be categorized to 3 main results which is:

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

Section 2

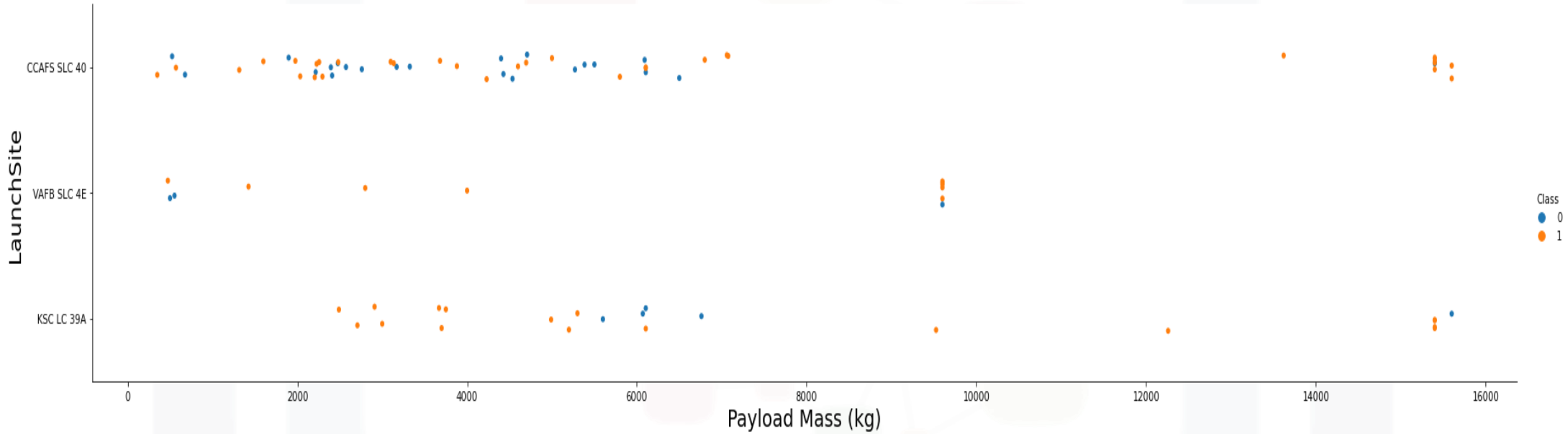
Insights drawn from EDA

Flight Number VS Launch Site



- Above scatter plot shows that many the flights of the launch site, the greater the success rate.
- However, site CCAFS SLC40 shows the least pattern of this.

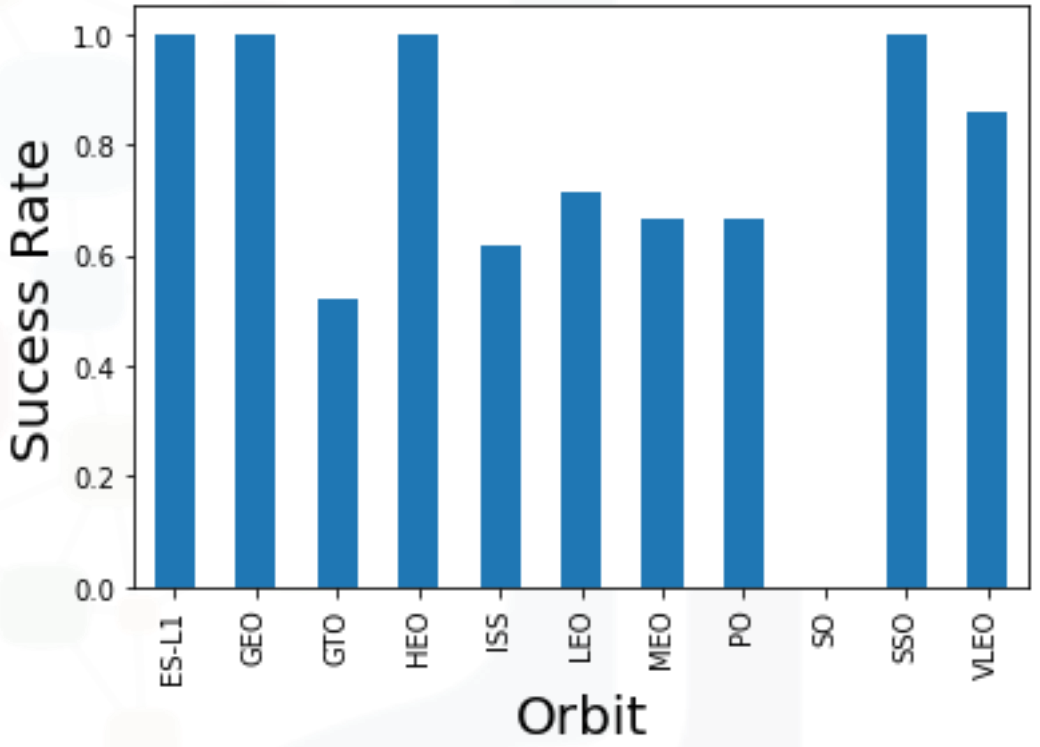
Payload VS Launch Site



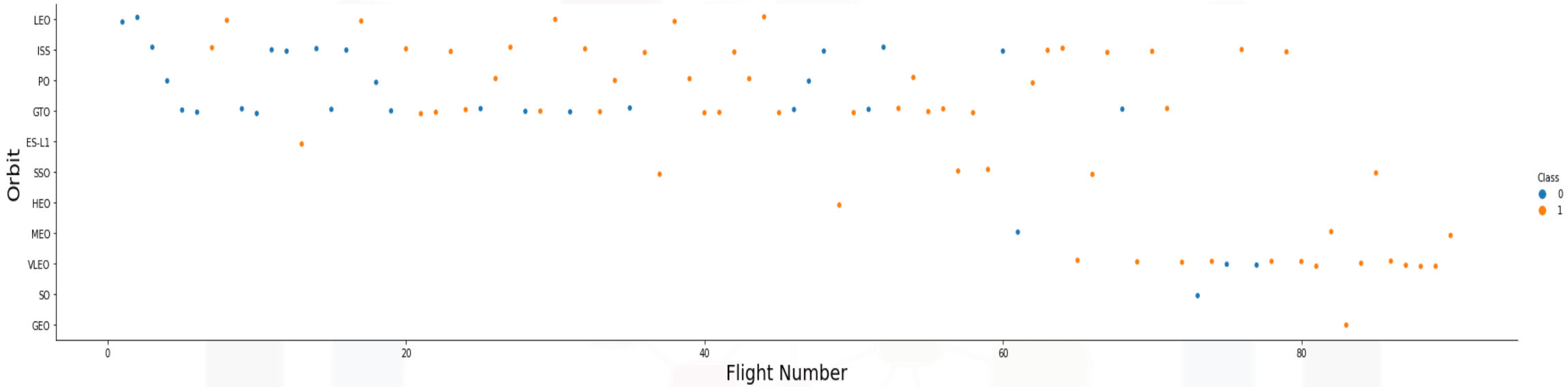
- Above scatter plot shows when the pay load mass is greater than 7000kg, the probability of the success rate can be increased highly.
- However, there is no clear pattern to conclude the launch site is dependent on the pay load mass for the success rate.

Success rate VS Orbit Type

- This figure shows the possibility of the orbits to influence landing outcomes as some orbits have 100% success rate such as ES-L1, GEO, HEO, and SSO whereas SO orbit produced 0% success rate.
- However, deeper analysis show that some of this orbits have only 1 occurrence such as GEO, SO, HEO and ES-L1 which mean this data need more dataset to see pattern or trend before we draw any conclusion.

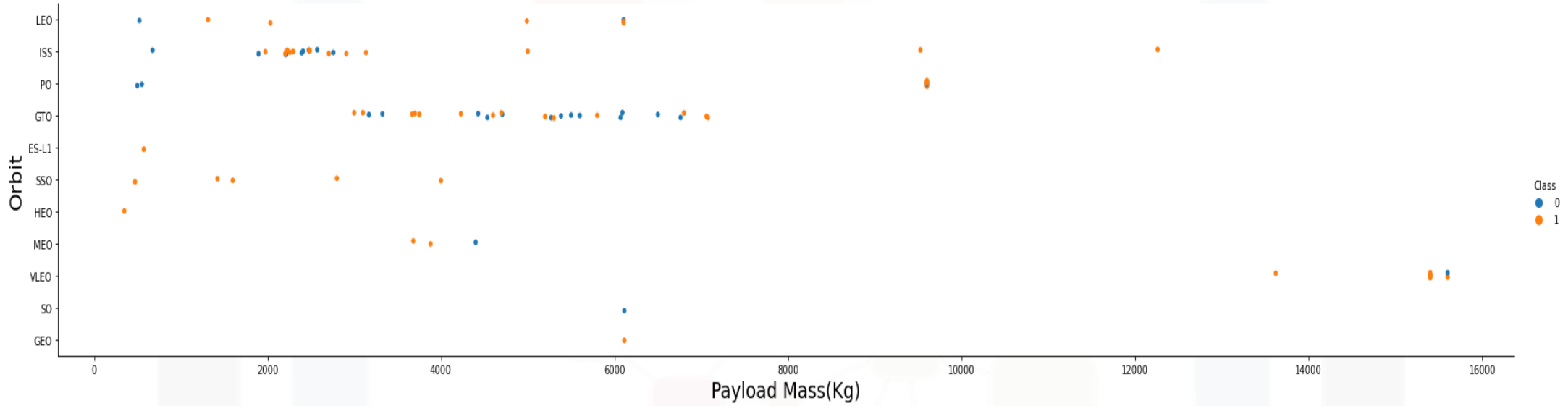


Flight No VS Orbit Type



- Above scatter plot shows that in general, the more the flight numbers on each orbits, the greater the success rate (especially LEO orbit) except for GTO orbit which depicts no relationship between both attributes.
- Orbit that only has 1 occurrence should also be excluded from above statement as it's needed more dataset.

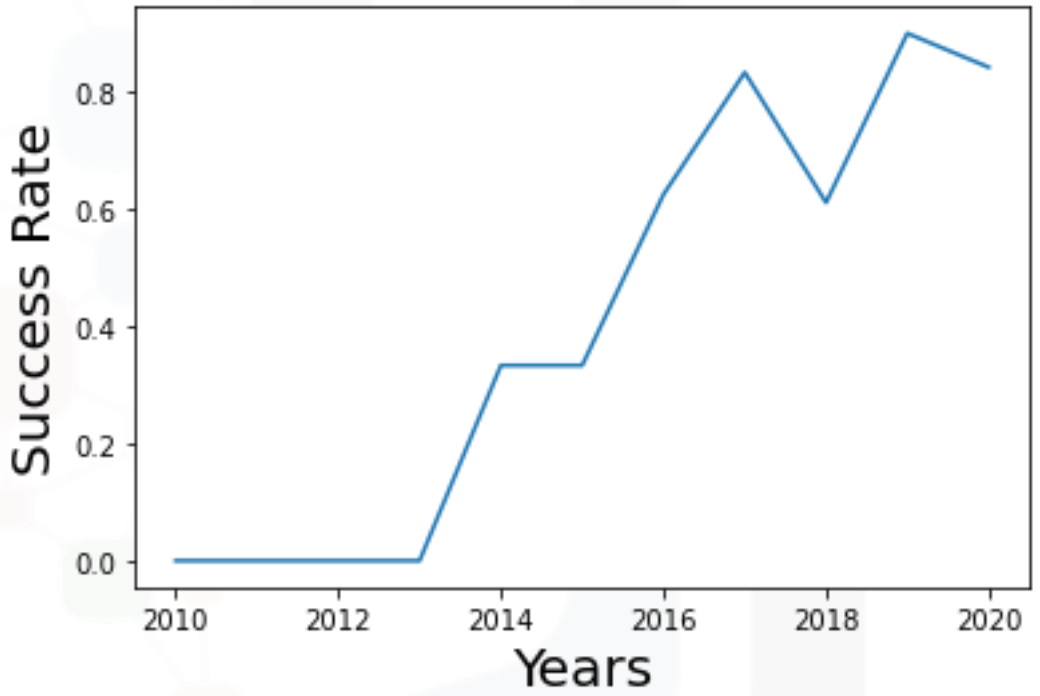
Payload VS Orbit Type



- Heavier payload has positive impact on LEO, ISS and P0 orbit. However, it has negative impact on MEO and VLEO orbit.
- GTO orbit seem to depict no relation between the attributes.
- Meanwhile, again, SO, GEO and HEO orbit need more dataset to see any pattern or trend.

Launch Success Yearly Trend

- The graph clearly shows the increasing trend from year 2013 until 2020.
- If this trend continues subsequently, the success rate will steadily increase until reaching success rate of 1/100%



Names of Launch Sites

Task 1

Display the names of the unique launch sites in the space mission

In [6]:
```
%sql select DISTINCT(LAUNCH_SITE) as "Launch_Sites" from SPACEXTBL;
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[6]:

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

Used the key word DISTINCT to show only unique launch sites from the SpaceX data.

Launch Site names beginning with CCA

Task 2

Display 5 records where launch sites begin with the string 'CCA'

In [7]:
```
%sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[7]:

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

Used the query above to display 5 records where launch sites begin with 'CCA'

Total Payload Mass

Task 3

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

In [8]:
```
%sql select sum(PAYLOAD_MASS__KG_) as Totalpayloadmass from SPACEXTBL;
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[8]:

totalpayloadmass
619967

Total payload was calculated, carried by boosters from NASA using the query above

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [9]:
```
%sql select avg(PAYLOAD_MASS__KG_) as Avgpayloadmass from SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/blud
b
Done.
```

Out[9]:

avgpayloadmass
2928

The average payload mass carried by booster version F9 v1.1 as calculated by query above.

First Successful Ground Landing Date

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

In [10]:
```
%sql select varchar_format((min(Date)), 'DD-MM-YYYY') as Sdate from SPACEXTBL where LANDING_OUTCOME = 'Success (ground pac
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/blud
b
Done.
```

Out[10]:

sdate
22-12-2015

Using min() function to find the result it is observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

In [11]:
```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[11]:

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

Using WHERE clause to filter for boosters which have successfully landed on drone ship and applied AND condition to determine successful landing with payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [12]:
```
%sql select count(MISSION_OUTCOME) as SuccessfulMission from SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/blud
b
Done.
```

Out[12]:

successfulmission
100

In [13]:
```
%sql select count(MISSION_OUTCOME) as FailureMission from SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Failure%';
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/blud
b
Done.
```

Out[13]:

failuremission
1

We used to filter through GROUP BY Mission Outcome to get a success (100) & a failure(1).

Boosters carrying Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [14]:
```
%sql select DISTINCT BOOSTER_VERSION as Boosterversions_Carried_MaxPayloadMass from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[14]:

boosterversions_carried_maxpayloadmass
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

Using sub-query in WHERE clause with the MAX() function, boosters carrying maximum payload have been found.

Launch Records of 2015

Task 9

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

In [16]:
```
%sql SELECT BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND LANDING_OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[16]:

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

Using combinations of WHERE, LIKE, AND clauses and BETWEEN conditions to filter the failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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

Task 10

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

In [17]:
```
%sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS "Total_Count" FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY LANDING_OUTCOME ORDER BY COUNT(LANDING_OUTCOME) DESC;
```

```
 * ibm_db_sa://htz48873:***@8e359033-a1c9-4643-82ef-8ac06f5107eb.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30120/bludb
Done.
```

Out[17]:

landing_outcome	Total_Count
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

Selecting Landing outcomes, using COUNT of landing outcomes from data and using the WHERE clause to filter the landing outcomes BETWEEN 2010-06-04 to 2010-03-20.

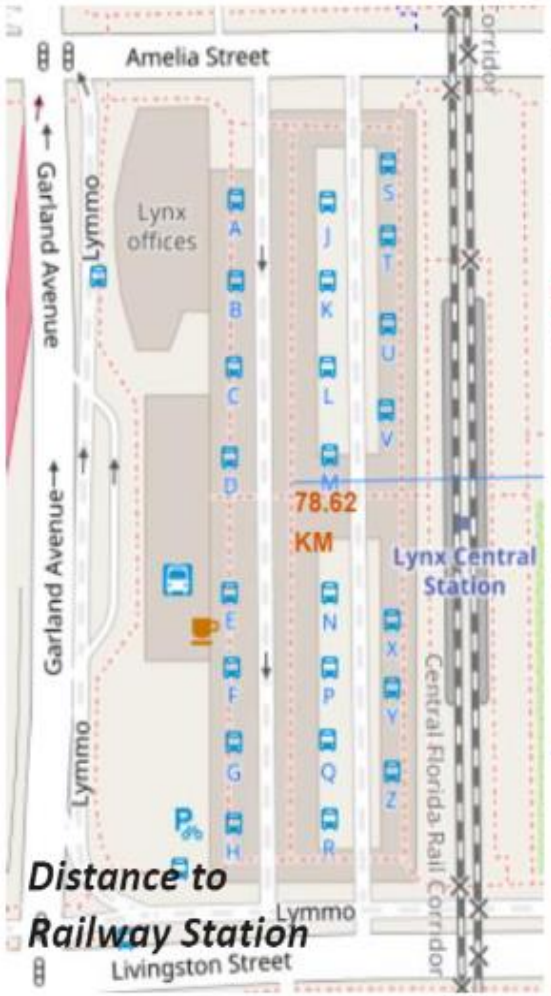
Section 3

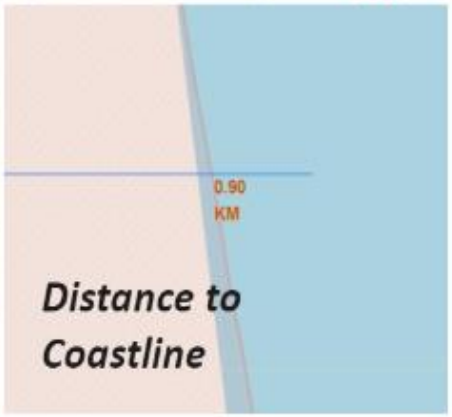
Launch Site Proximities Analysis

Launch Site Locations

Markers showing Launch Sites with color labels

Launch Sites Distance to Landmarks





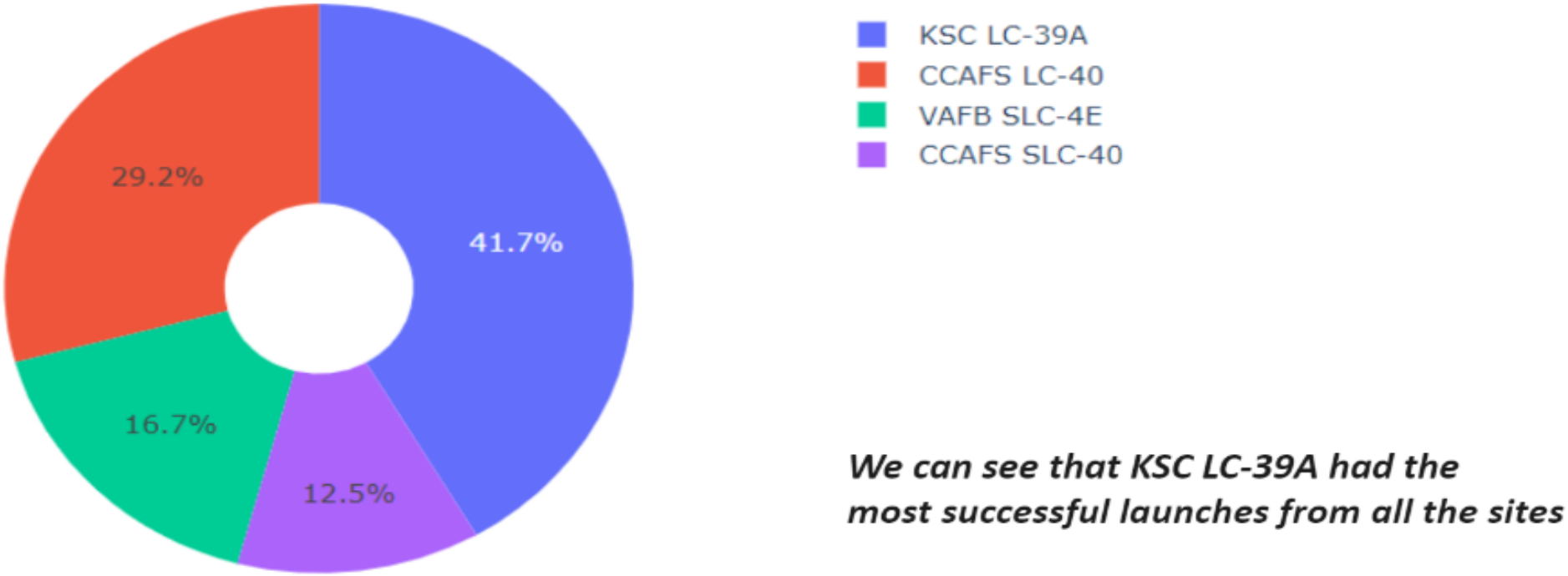
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

Building Dashboard with Plotly Dash

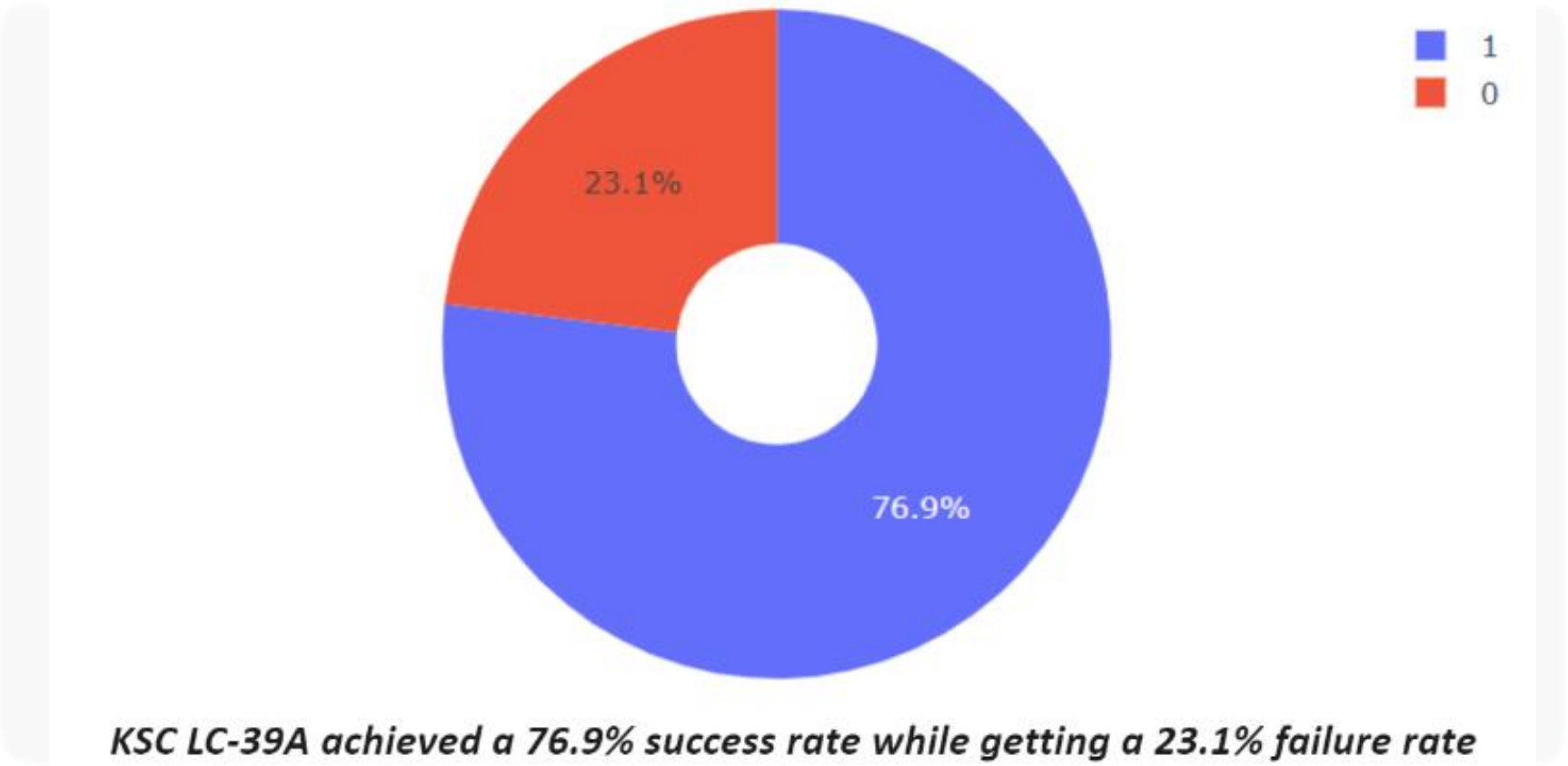
Success Percentage by Each Site

Total Success Launches By all sites



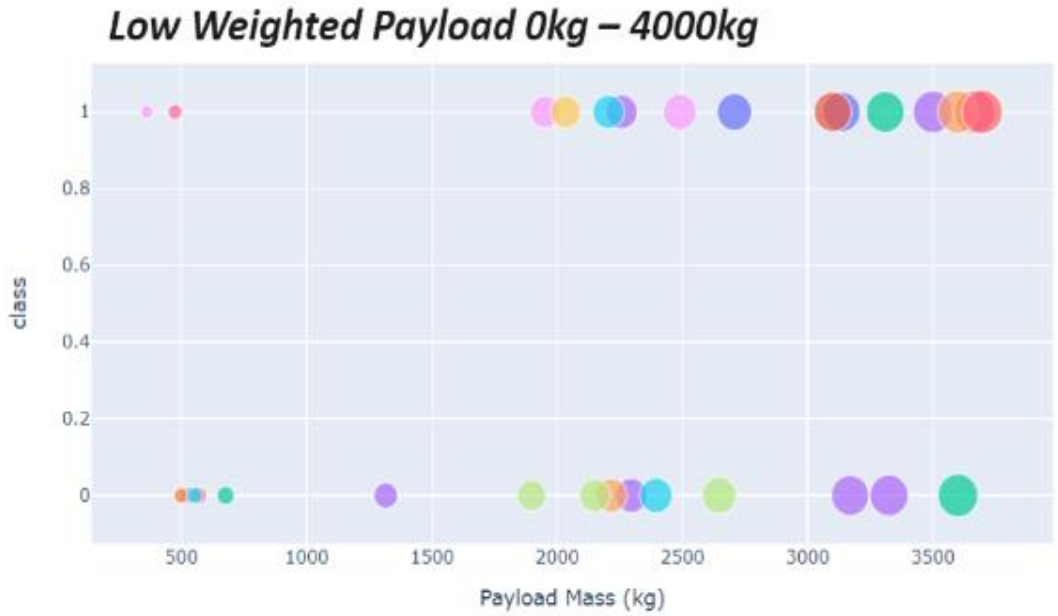
We can see that KSC LC-39A had the most successful launches from all the sites

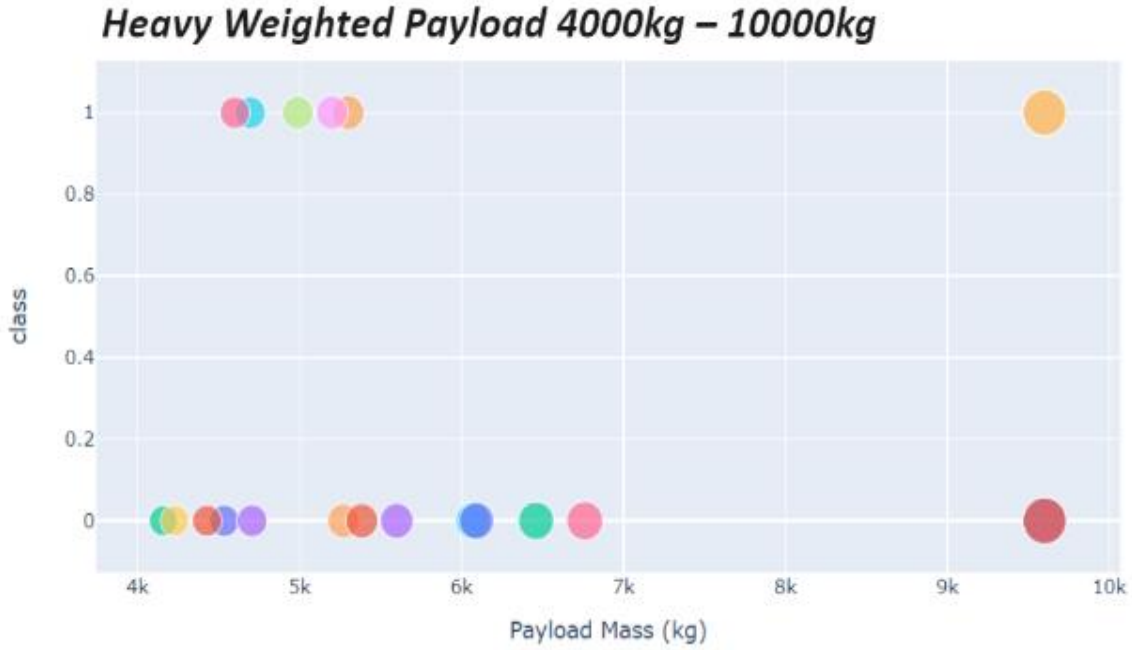
The highest launch-success ratio: KSC LC-39A



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

Payload vs Launch Outcome Scatter Plot





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis

Classification Accuracy

TASK 12

Find the method performs best:

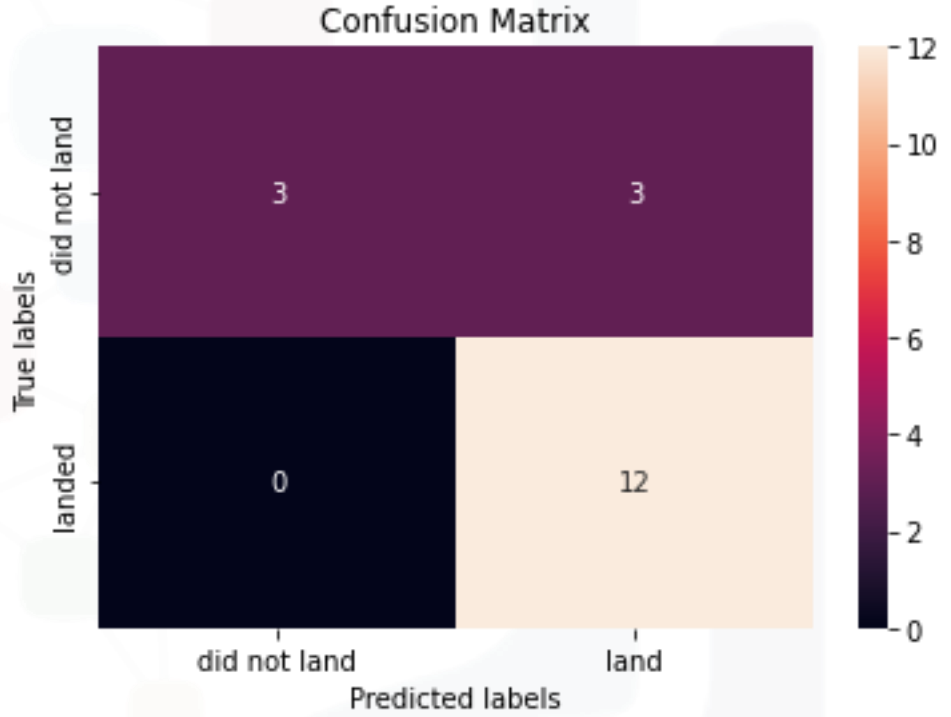
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split':
5, 'splitter': 'random'}
```


- As we can see, by using the code as above: we could identify that the best algorithm as the Decision Tree Classifier Algorithm which has the highest classification accuracy.

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes.
- One major issue is the false positives .i.e., unsuccessful landings are marked as successful landings by the classifier.



CONCLUSION

We can conclude that:

- The more number of flights at a launch site, the greater the success rate at that launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- Of any launch sites KSC LC-39A had the most successful launches.
- The best machine learning algorithm is the Decision tree classifier as observed.

Thank you