IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nirali Hemant Dave
14/10/2022

Outline

Executive Summary

- Summary of methodologies

  -*Data Collection through API*
  - *Data Collection with Web Scraping*
  - *Data Wrangling*
  - *Exploratory Data Analysis with SQL*
  - *Exploratory Data Analysis with Data Visualization*
  - *Interactive Visual Analytics with Folium*
  - *Machine Learning Prediction*

- Summary of all results
  - *Exploratory Data Analysis result*
  - *Interactive analytics in screenshots*
  - *Predictive Analytics result*

Introduction

- **Project background and context**

SpaceX is the most successful for making space travel affordable. Its accomplishment include:

-Sending spacecraft to the International Space Station.

-Sending manned missions to Space.

SpaceX get can do because of inexpensive rocket launch. In website SpaceX mention that it cost around 62 million dollars for Falcon 9 rocket launch which is comparatively very less as compare to other company as other provide cost upward of 165 million dollars each.

SpaceX do much saving by reusing 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 companywants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- **Problems you want to find answers**

1. For successful landing which factor to be consider.

2. Rate of successful Landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We gathered data from SpaceX REST API and web scraping from wikipedia
- Perform data wrangling
  - One hot encoding was applied to categorical features
- 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

Data Collection from SpaceX REST API

1. Data is collected by requesting using request library from SPACEX API
2. After receiving response in form of a JSON the data is been converted in to dataframe using json_normalization function
3. After Normalization data will be in table format and have to deal with missing values and clean data where it is necessary

Below data is being used

Flight Number, Date, Booster Version, Payload Mass, Orbit, Launch Site, Outcome, Flights, Grid Fins, Reused, Legs, Landing Pad, Block, Reused Count, Serial, Longitude, Latitude

Data Collection from web scraping related wikipedia

1. BeautifulSoup package is used to web scrape HTML table to retrieve Falcon 9 launch data, further data is parse the data from table and convert into Pandas data frames for future consideration to analysis.

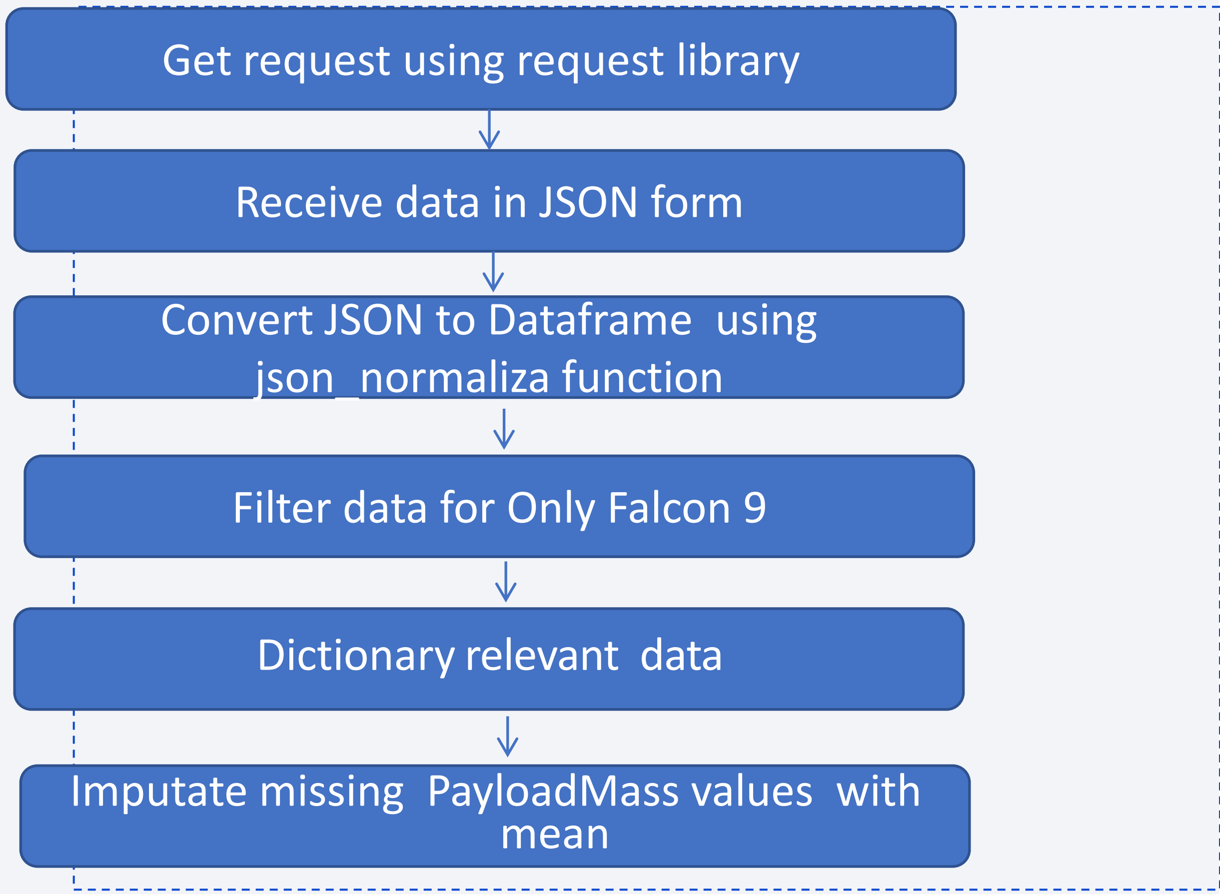
Below data is being used

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

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

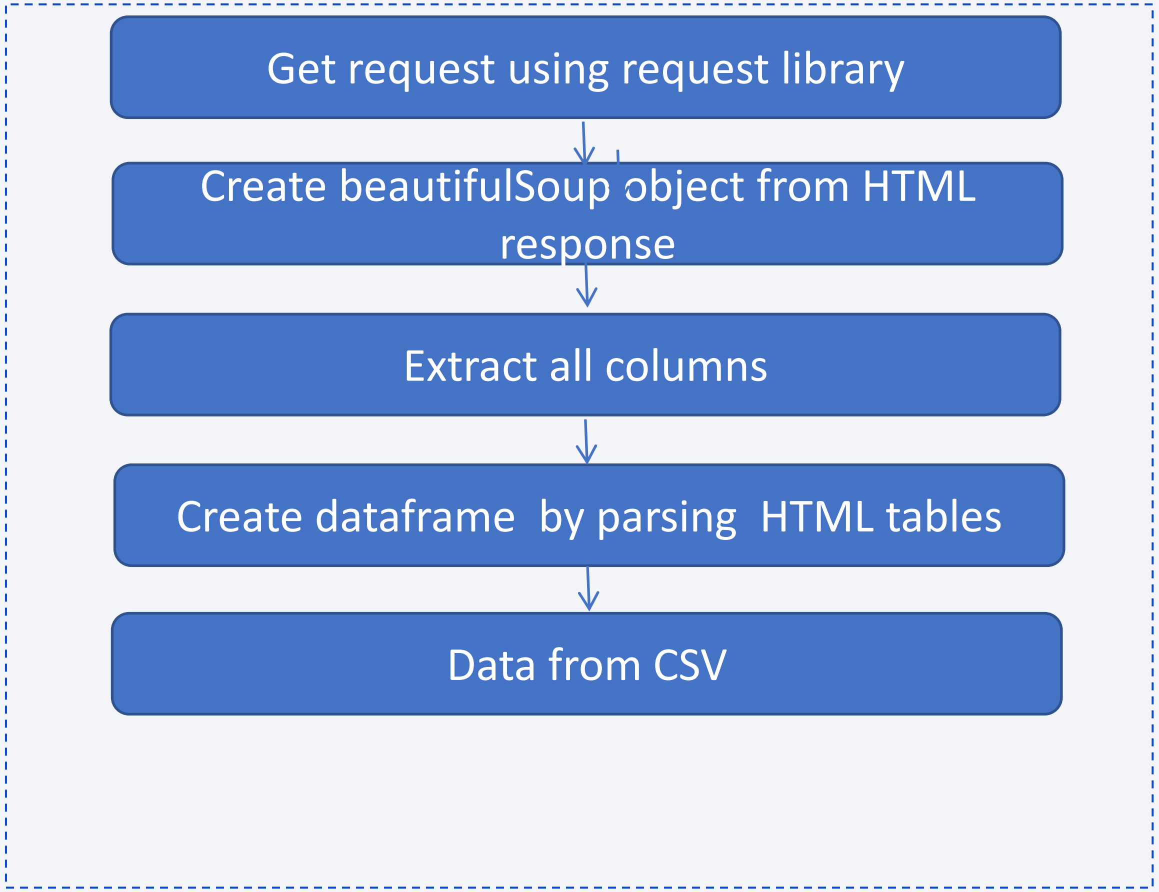
- https://github.com/Nirali0/My-submission/blob/main/jupyter-labs-spacex-data-collection-api(2).ipynb



Data Collection - Scraping

- We applied web scrapping to webscrap Falcon 9 launch records with BeautifulSoup
  • We parsed the table and converted it into a pandas dataframe.

- https://github.com/Nirali0/My-submission/blob/main/10.%20Captone%20Week1Data%20Collection%20with%20Web%20scrapering.ipynb



Data Wrangling

- We will perform Exploratory Data Analysis (EDA) patterns in the data.
- Calculate no of launches at each site, number and occurrence of each orbits.
- We covert landing outcome

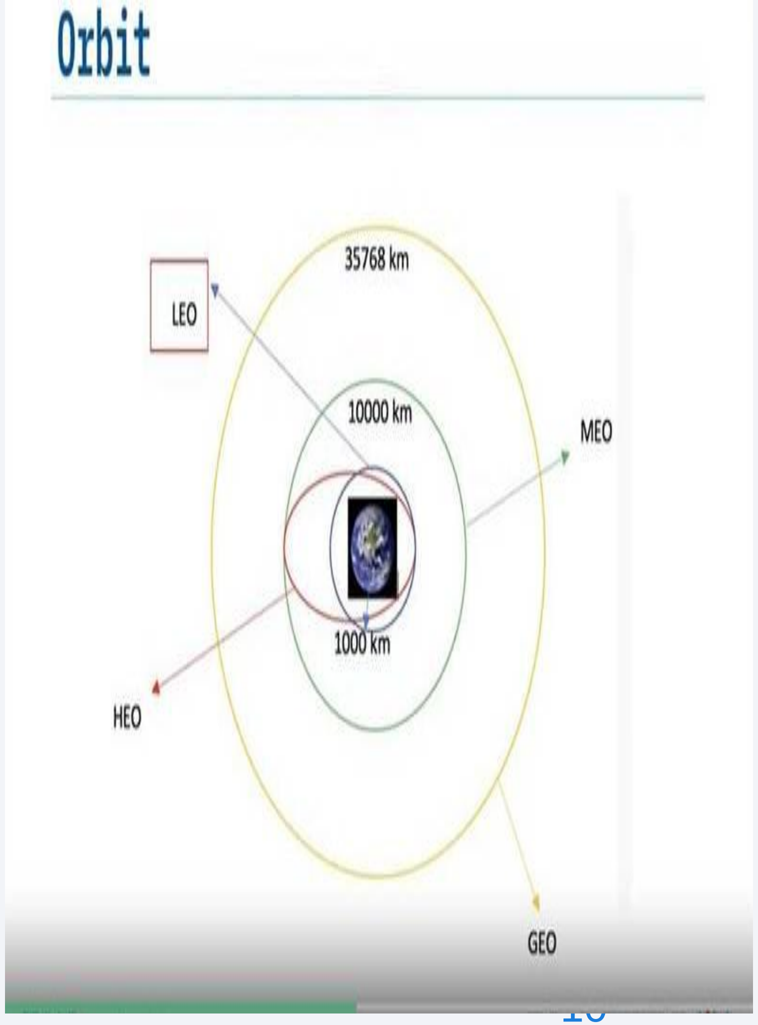
  1 Means the booster successfully landed

  0 Means it was unsuccessful

  Result will be in CSV form

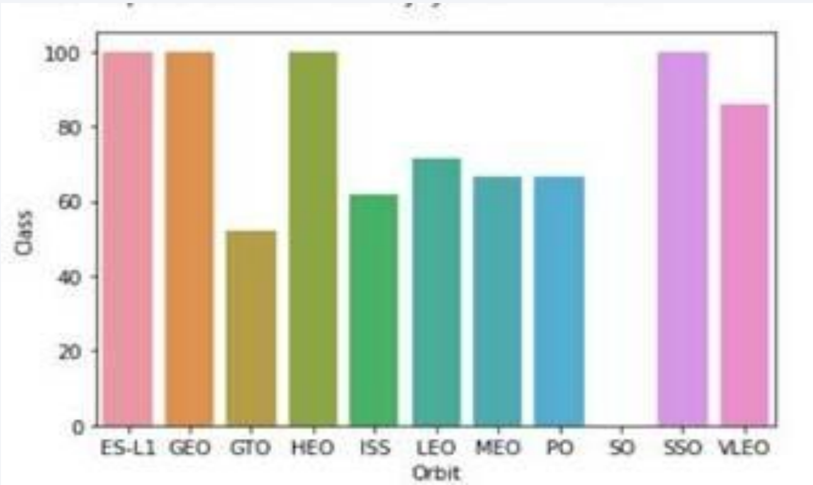
Below is notebook link:

https://github.com/Nirali0/Mysubmission/blob/main/10.%20Captone%20Week1Data%20Collection%20with%20Web%20scrapering.ipynb



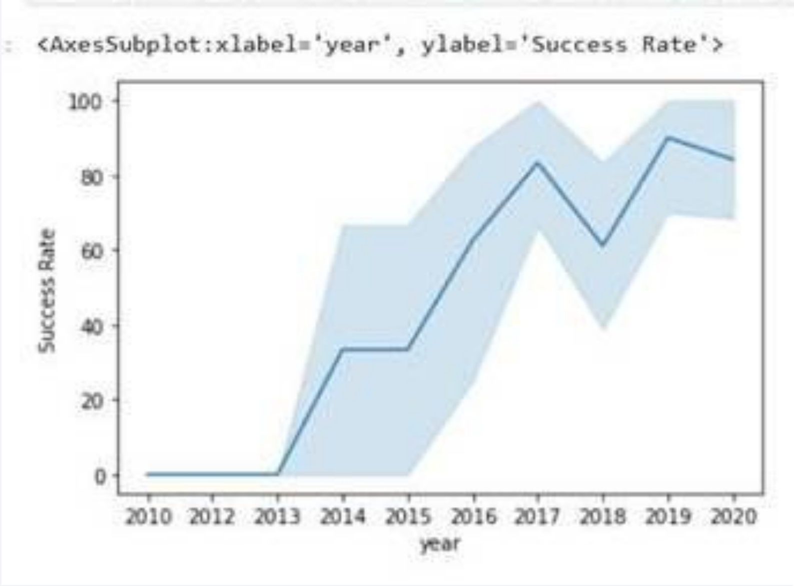
EDA with Data Visualization

- We explored the data by visualizing the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend



- the average launch success yearly trend

Notebook link: https://github.com/Nirali0/My-submission/blob/main/10.Captone%20Proejct%20Week2%20EDA%20with%20Visualization.ipynb



EDA with SQL

- We loaded the SpaceX dataset into a PostgreSQL database without leaving the jupyter notebook.

- We applied EDA with SQL to get insight from the data. Following are the queries for executing
  Display unique Launch site
- Retrieve total and average payload mass by boosters
- Retrieve total number of successful and failure mission outcome.
- Retrieve month names, failure landing_outcomes in drone ship, booster versions, lauch_site

Notebook link:

https://github.com/Nirali0/My-submission/blob/main/10.%20Capston%20project_week2_eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- We create mark on each launch site using folium map and folium circle to highlight circle area.
- Create folium.marker as well for each launch site.
- Explore as which launch site has success rate of launching by using class column.

  If a launch was successful - (Class=1)

  If a launch was failed – (Class=0)

- We calculate the distance between launch site to its proximity. We get to know whether the site is near to railways, coastline or highway. Folium marker is used to show distance.

Notebook link:

https://github.com/Nirali0/Mysubmission/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

Notebook link:

- https://github.com/Nirali0/Mysubmission/blob/main/10.%20capstone%20week3%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash

Predictive Analysis (Classification)

- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.

Notebook link:

- https://github.com/Nirali0/Mysubmission/blob/main/10.%20Capstone%20Week%204%20Space%20X%20Falcon%209%20First%20Stage%20Landing%20Prediction.ipynb

Results

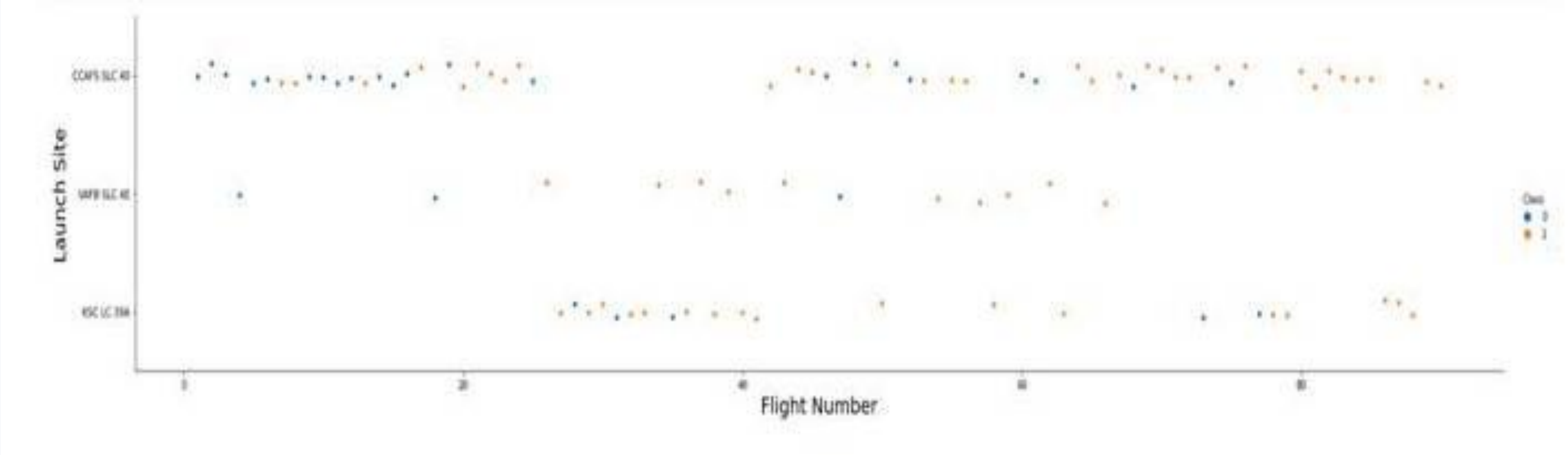
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Show a scatter plot of Flight Number vs. Launch Site



- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.

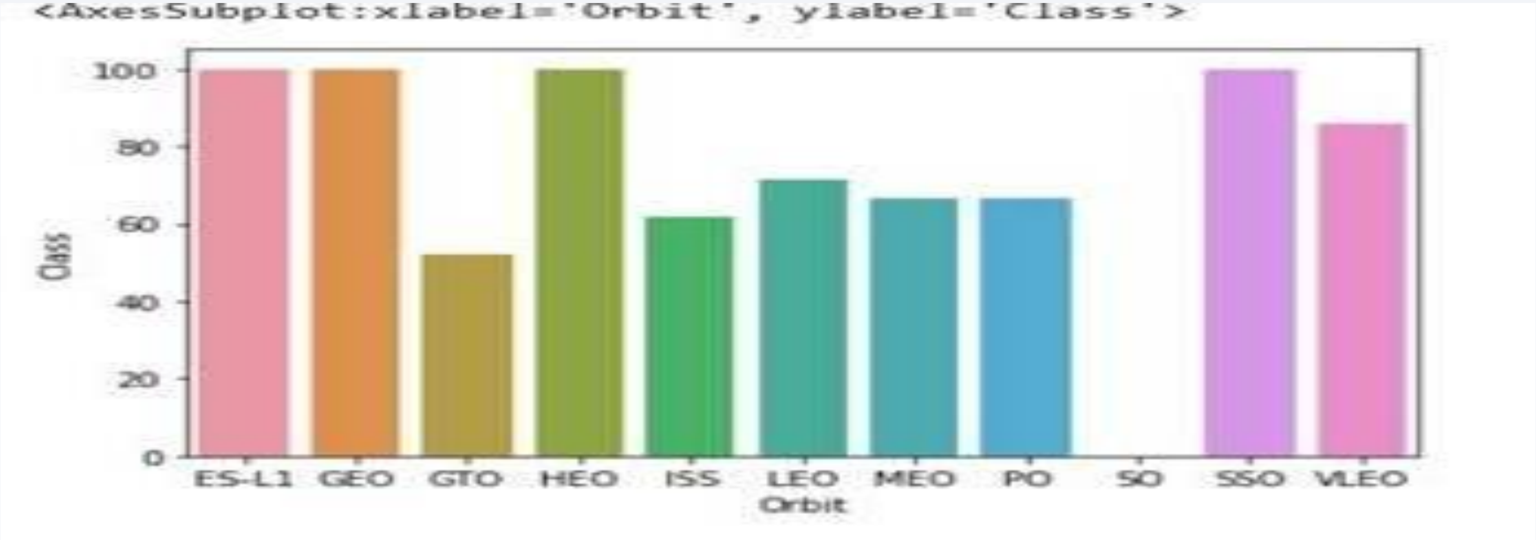
Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

- CCAFS SLC 40 higher success rate as the payload mass is greater
- If payload is greater the success rate would be higher.

Success Rate vs. Orbit Type

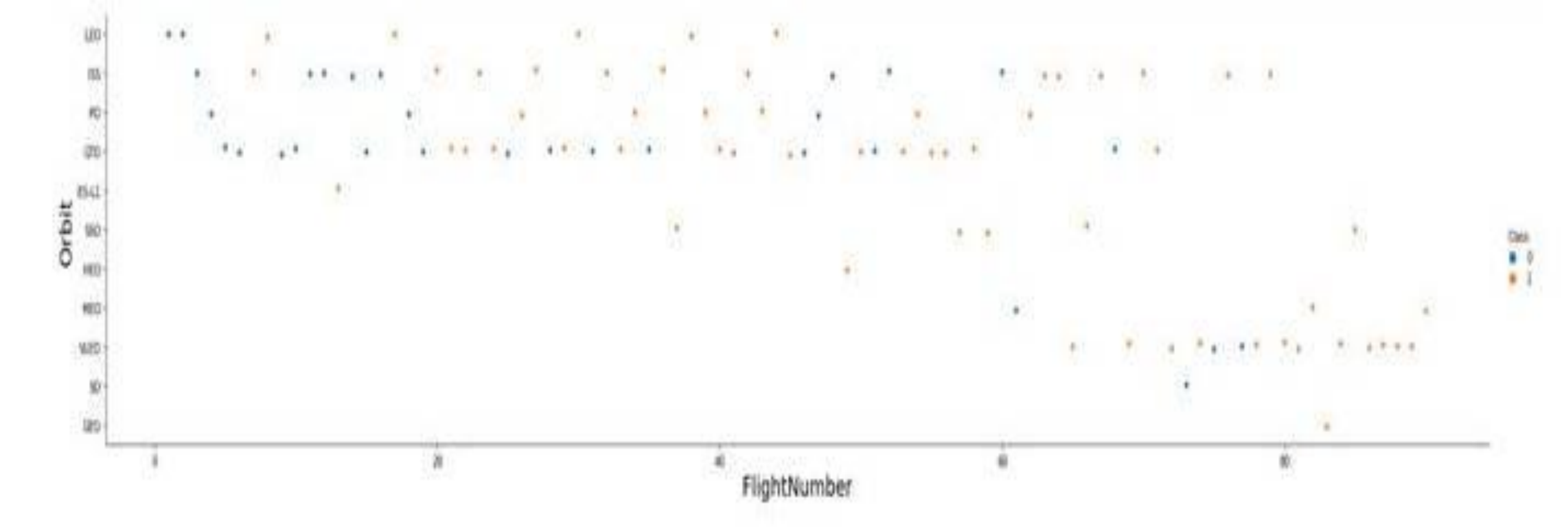
- Show a bar chart for the success rate of each orbit type



- Orbit type like ES-L1,GEO, HEO, SSO has highest success rate.

Flight Number vs. Orbit Type

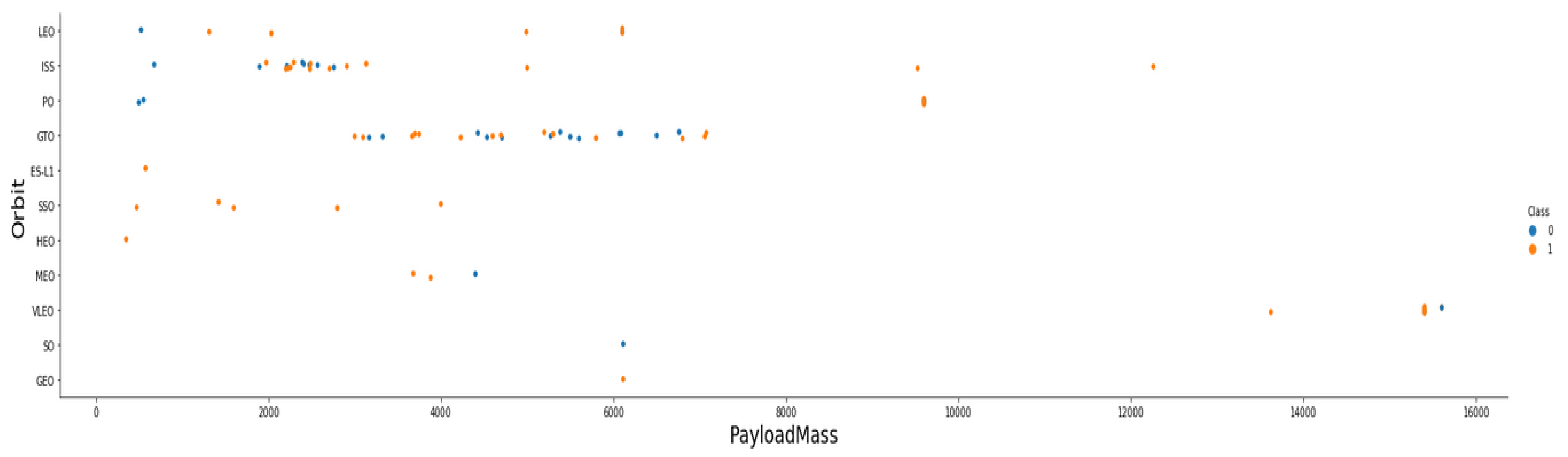
- Show a scatter point of Flight number vs. Orbit type



- 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 vs. Orbit Type

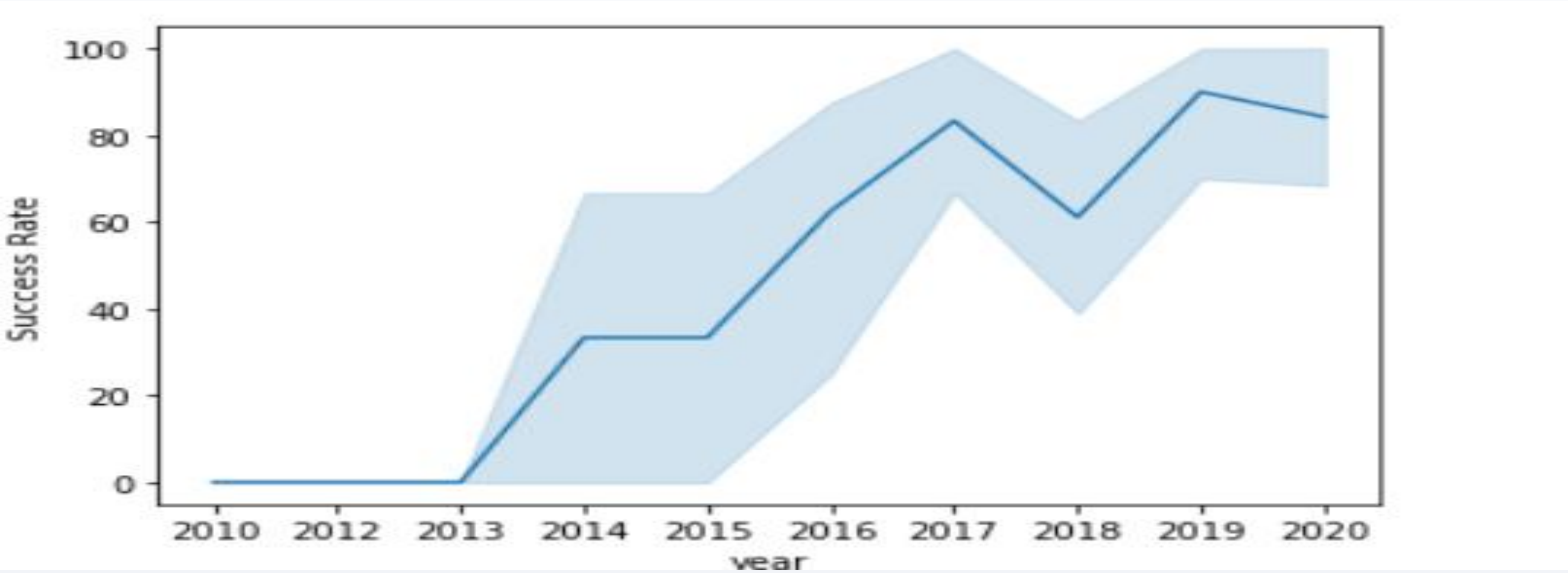
- Show a scatter point of payload vs. orbit type



- LEO, ISS and PO has successful landing rate where as GTO is not clear because positive and negative landing is there.

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- There is continuous rise in success rate from 2013.

All Launch Site Names

Task 1

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

[6]:
```
%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET
```

Total five Launch site

Launch Site Names Begin with 'CCA'

Task 2

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

```
In [9]: %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

Out[9]:

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

- We execute query to display Launch site which begin with ‘CCA’

Total Payload Mass

Task 3

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

[10]:
```
%sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

[10]:

payloadmass
619967

- Calculate the total payload carried by boosters from NASA

Average Payload Mass by F9 v1.1

Task 4

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

[11]:

```
%sql select avg(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL;
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

t[11]:

payloadmass
6138.287128712871

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

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
[12]: %sql select min(DATE) from SPACEXTBL;
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

t[12]:

min(DATE)
01-03-2013

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

[17]:

```
%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
```

```
 * ibm_db_sa://nxs27972:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/BLUDB
Done.
```

[17]:

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

[27]:
```
%sql select count(MISSION_OUTCOME) as missionoutcomes from SPACEXTBL GROUP BY MISSION_OUTCOME;
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

[27]:

missionoutcomes
1
98
1
1

Details of Mission Outcomes

Boosters Carried Maximum Payload

Task 8

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

In [28]:
```
%sql select BOOSTER_VERSION as boosterversion from SPACEXTBL where PAYLOAD_MASS__KG_=(select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
   ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
 * sqlite:///my_data1.db
Done.
```

Out[28]:

boosterversion
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

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for yea

```
%sql SELECT MONTH(DATE),MISSION_OUTCOME,BOOSTER_VERSION,LAUNCH_SITE FROM SPACEXTBL where EXTRACT(YEAR FROM DATE)='2015';
```

```
  ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb
* sqlite:///my_data1.db
```

Out[49]:

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- 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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

[33]:
```
%sql SELECT LANDING__OUTCOME FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY DATE DESC;
```

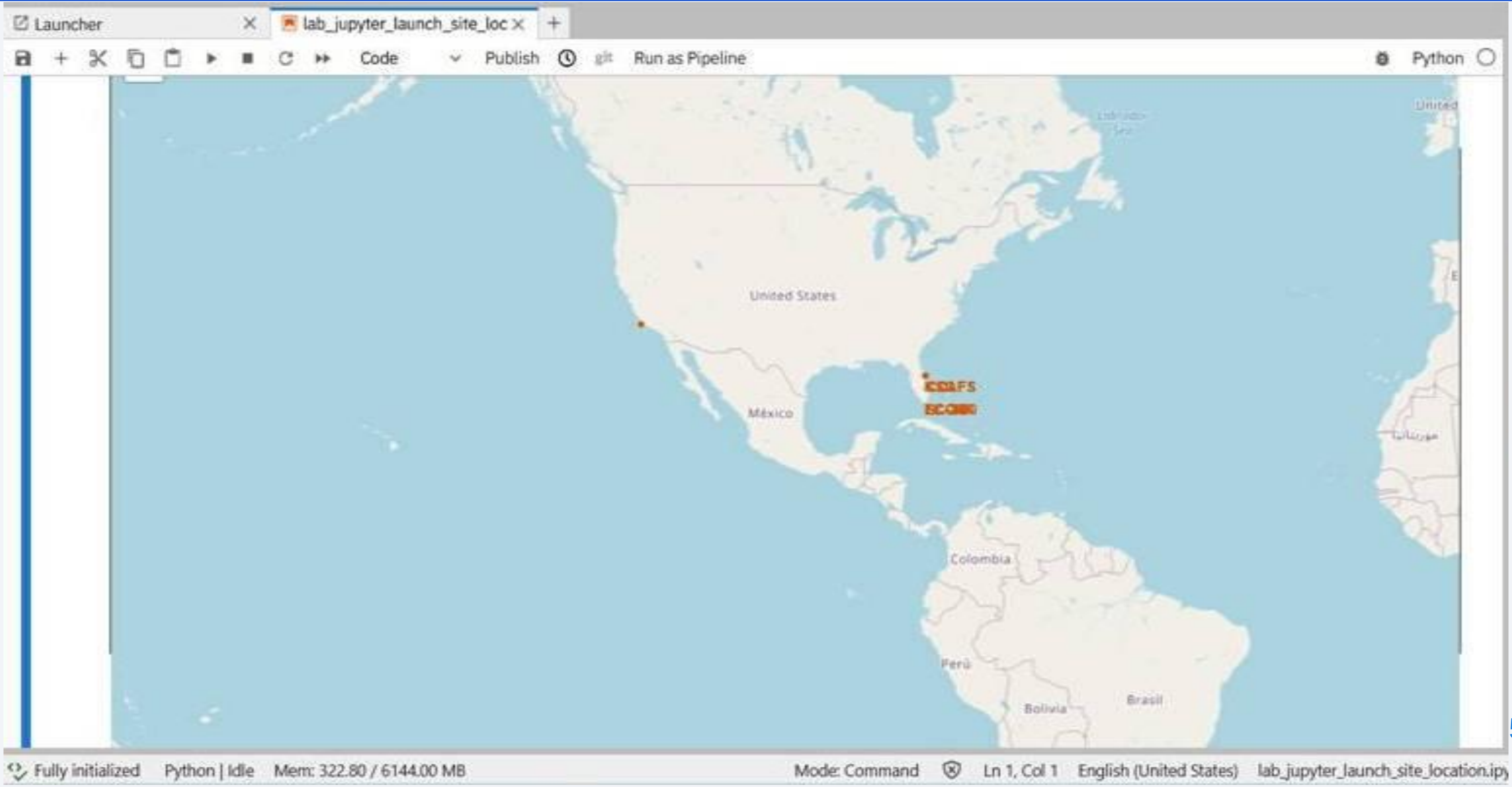
ibm_db_sa://tkk07684:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/bludb

landing__outcome	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

Section 3

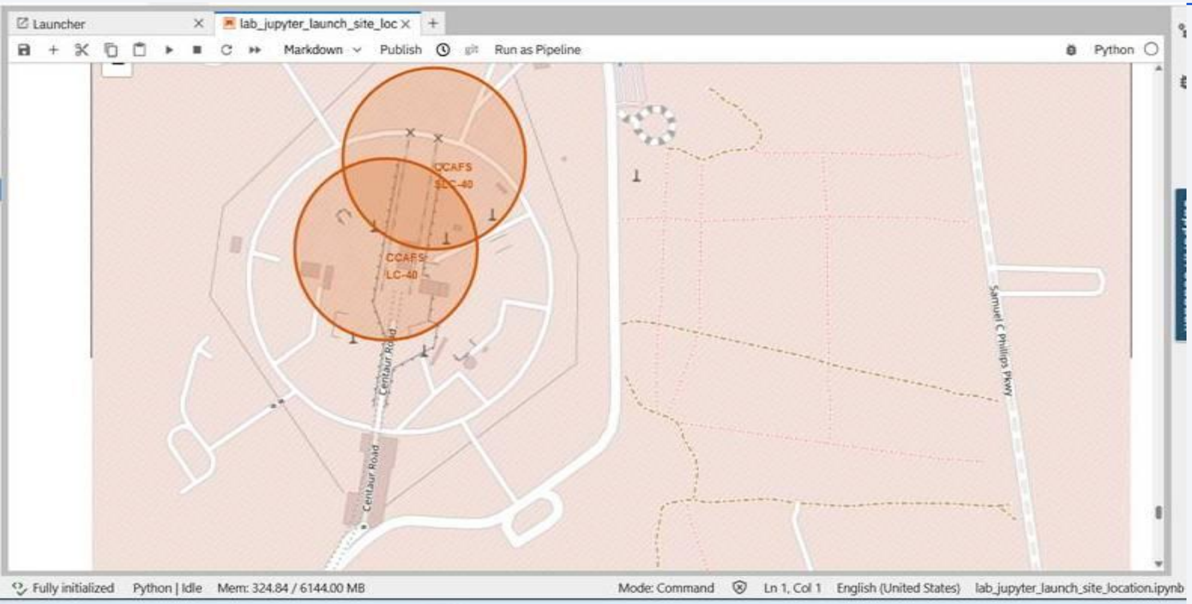
Launch Sites Proximities Analysis

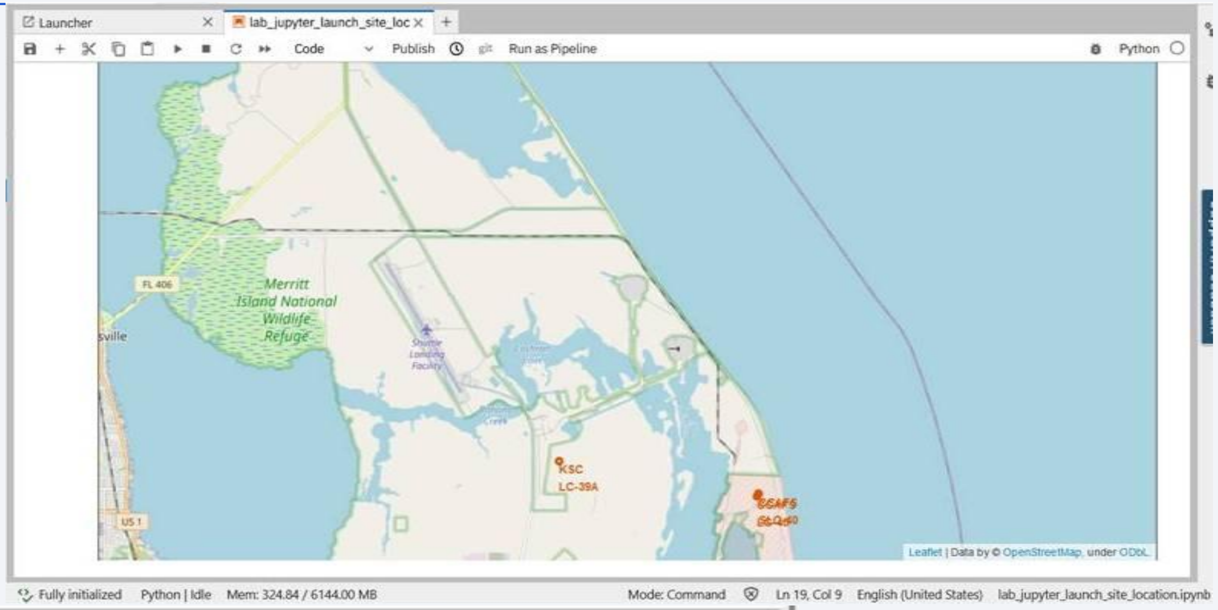
Two site are display in the region of American coaster Califonia and Florida

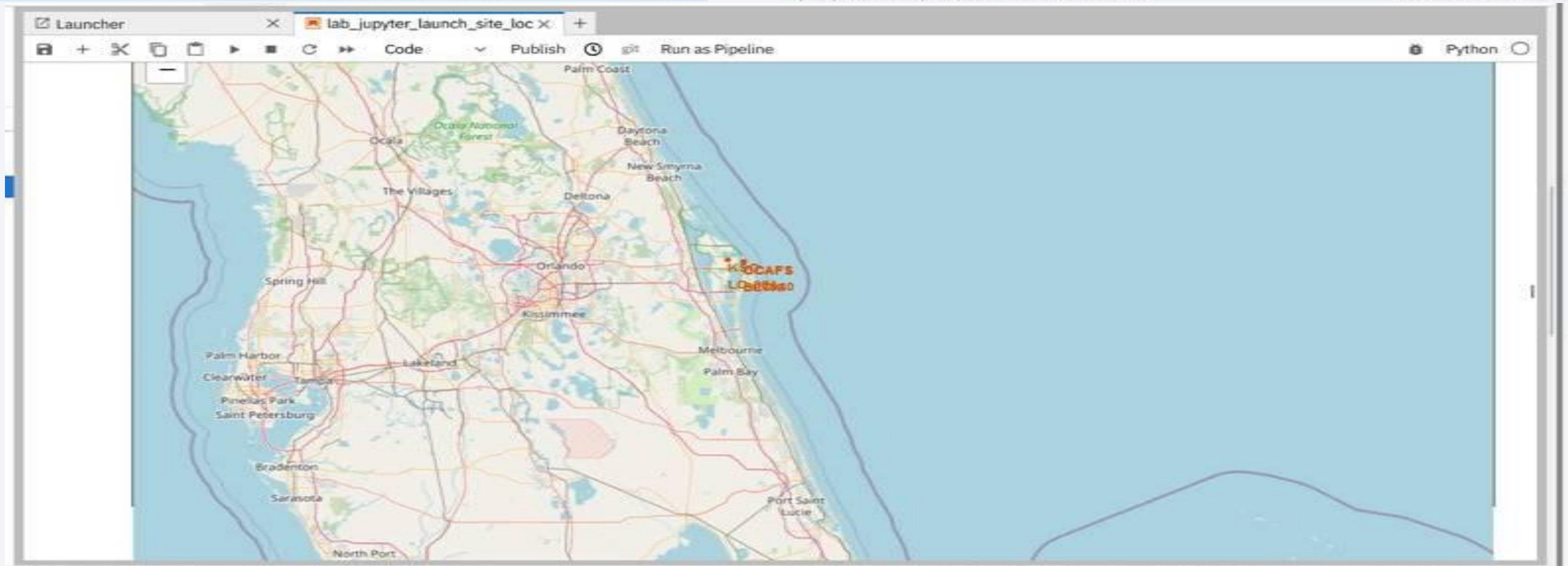


From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates. which launch sites have relatively high success rates.

Launch site distance from landmark



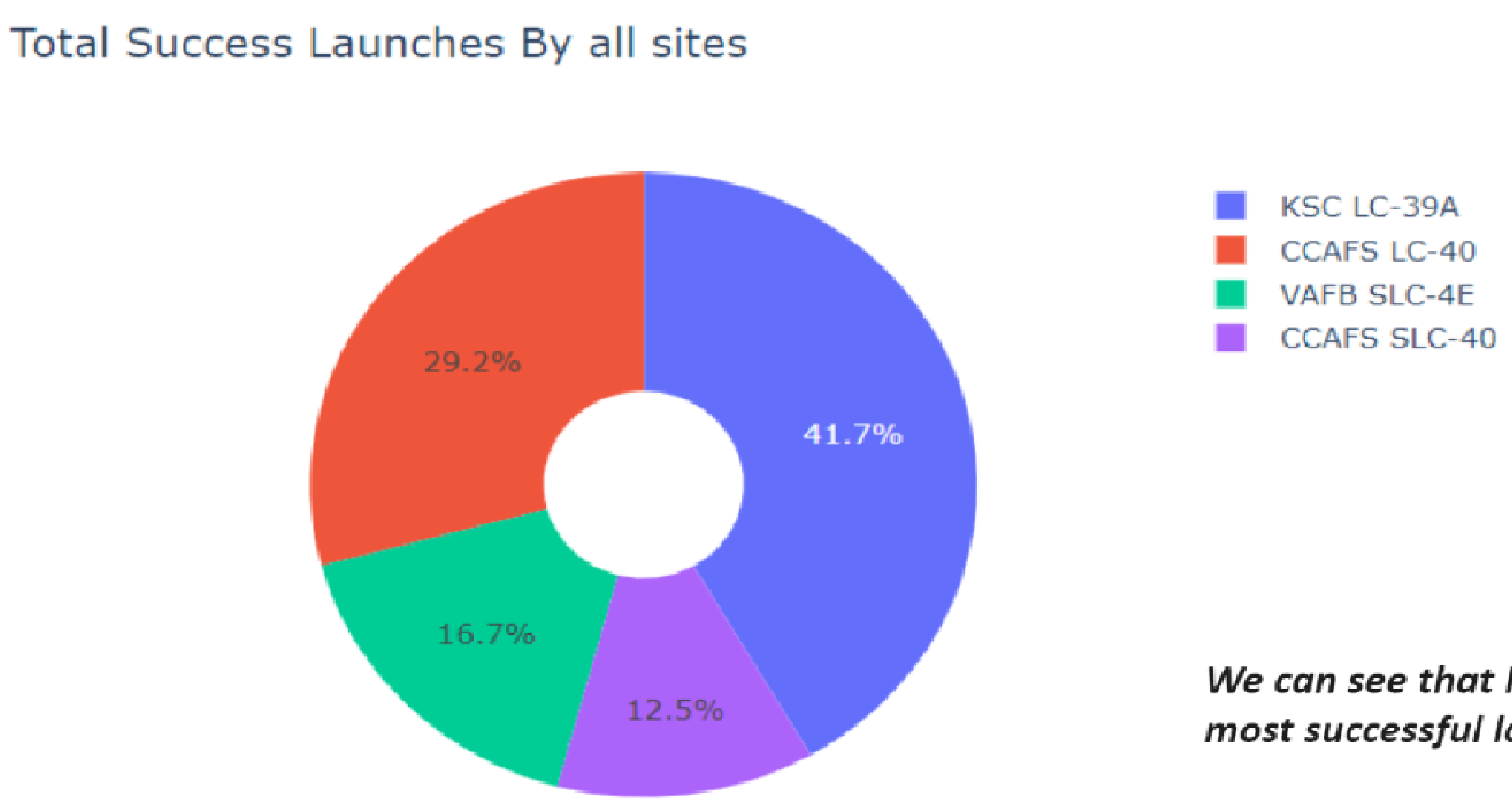




Section 4

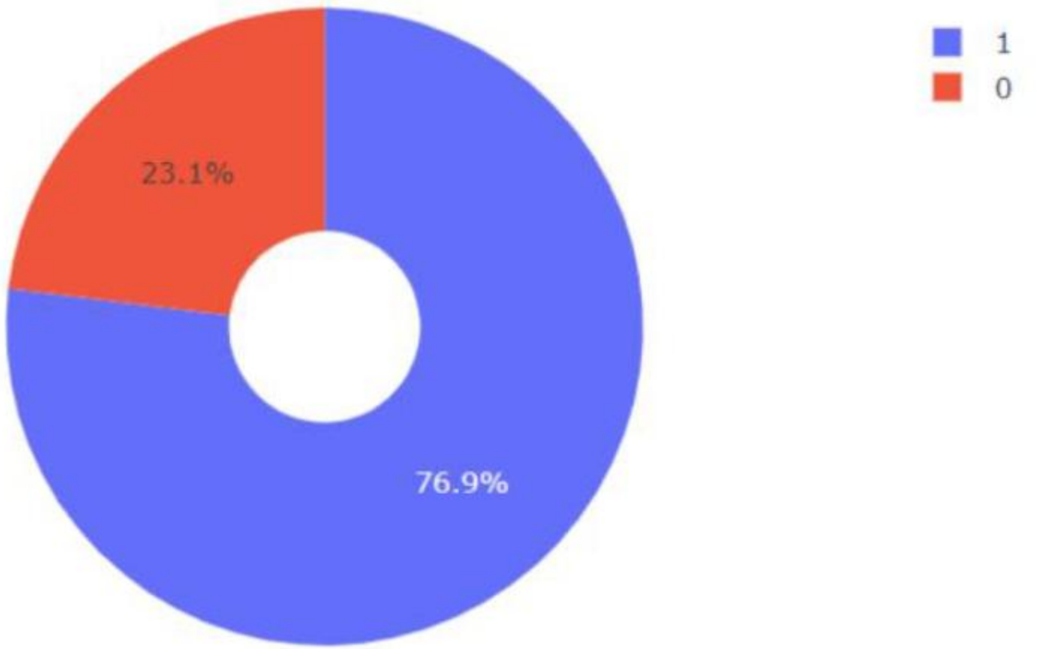
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



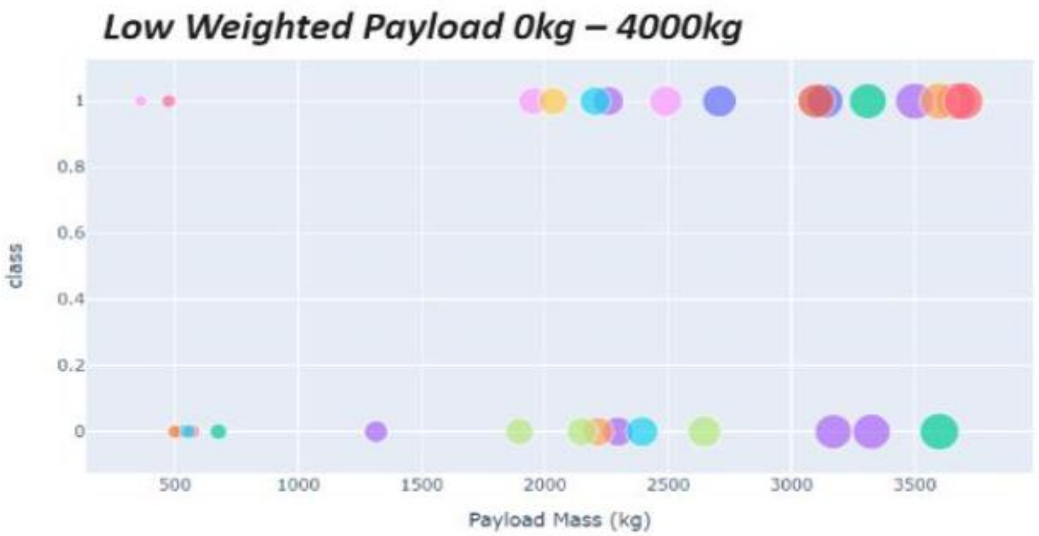
We can see that KSC LC-39A had the most successful launches from all the sites

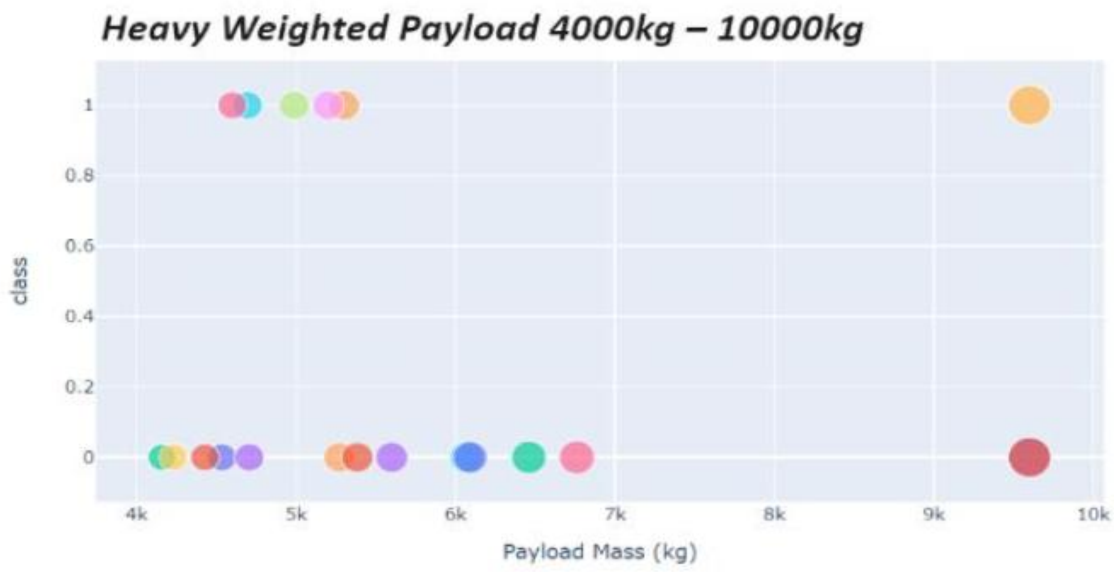
Pie chart showing the Launch site with the highest launch success ratio



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

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





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

Section 5

Predictive Analysis (Classification)

Classification Accuracy

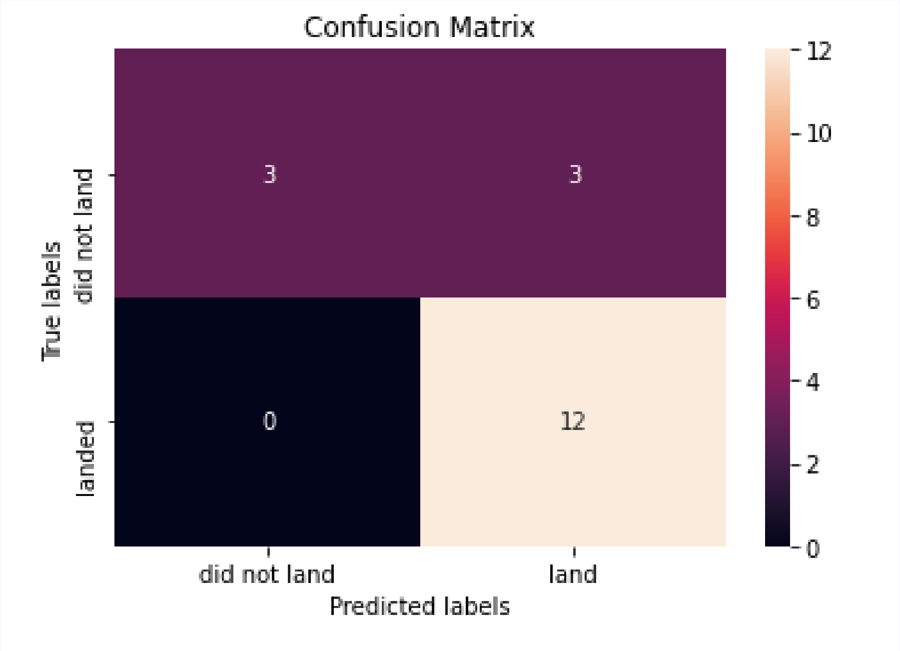
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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'}
```

Confusion Matrix

- The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier



Conclusions

- The is rise in launch success rate from 2013
- Orbit type like ES-L1,GEO, HEO, SSO has highest success rate.
- CCAFS SLC 40 higher success rate as the payload mass is greater.
- Higher amount of flight have a greater success rate.

Thank you!