IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Gunawan
28 October 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Methodologies to analyze data

- Data Collection using web scraping and SpaceX API; data wrangling
- Exploratory Data Analysis (EDA) for data visualization and interactive dashboard
- Machine Learning Prediction.

Summary of all results

- EDA is a useful method to identify which features best predict launch success;
- Machine Learning Prediction shows the best model to predict which characteristics are important to predict the launch success.

Introduction

Project background and context

- Space X 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 Space X 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 space X for a rocket launch. The project is aimed to predict if the Falcon 9 first stage will land successfully.

Problems:

1. What factors determine if the rocket will land successfully?
2. The interaction amongst various features that determine the success rate of a successful landing.
3. What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - summarizing and analyzing features
  - creating a landing outcome label based on outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- The predictive analysis was performed using classification models, which are decision tree, K-Neighbors, Logistic regression and support vector machine.

Data Collection

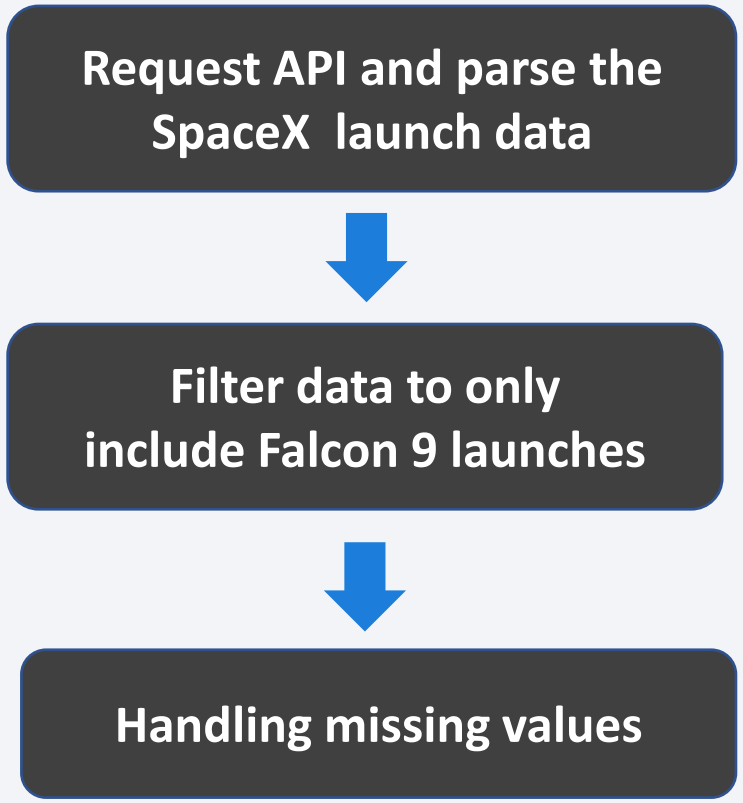
Data sets were collected from

- Space X API (https://api.spacexdata.com/v4/rockets/)
- Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection – SpaceX API

- The process of collecting data using the API is presented in a flowchart.
- The SpaceX API notebook is stored on the GitHub with this link:

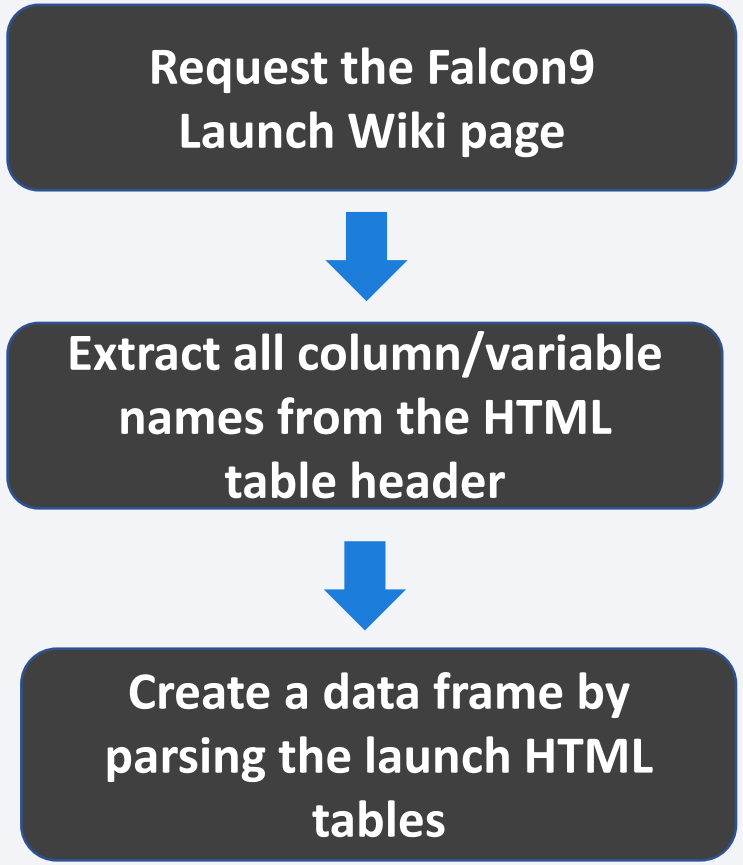
https://github.com/unyil01/edx-IBM-capstone/blob/main/01%20Data%20Collection%20API.ipynb



Data Collection - Scraping

- Data are scrapped from Wikipedia followed the process as outlined in the flowchart
- The notebook is stored on the GitHub with this link:

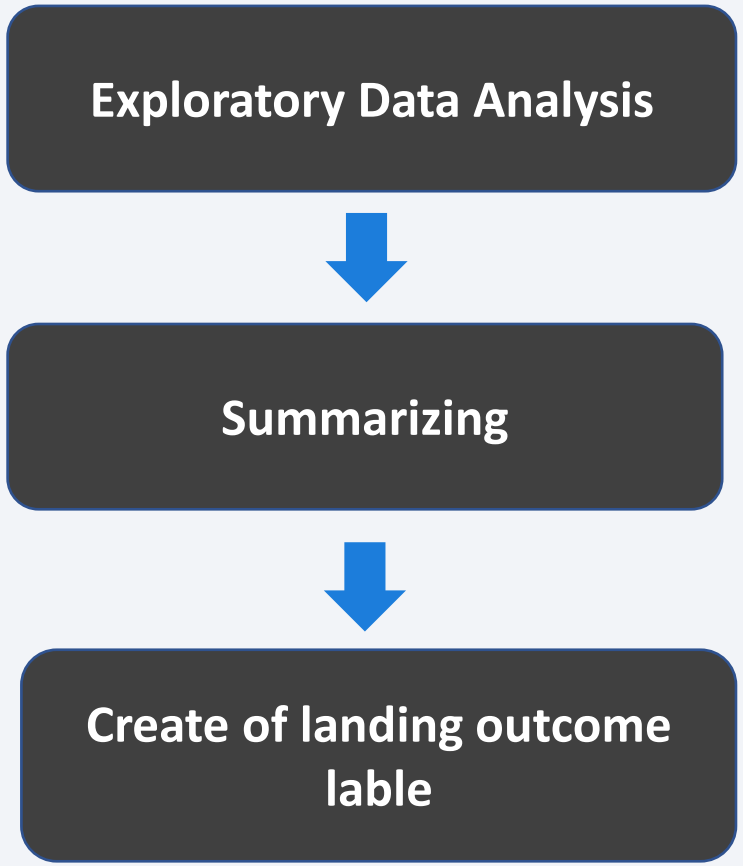
https://github.com/unyil01/edx-IBM-capstone/blob/main/02%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

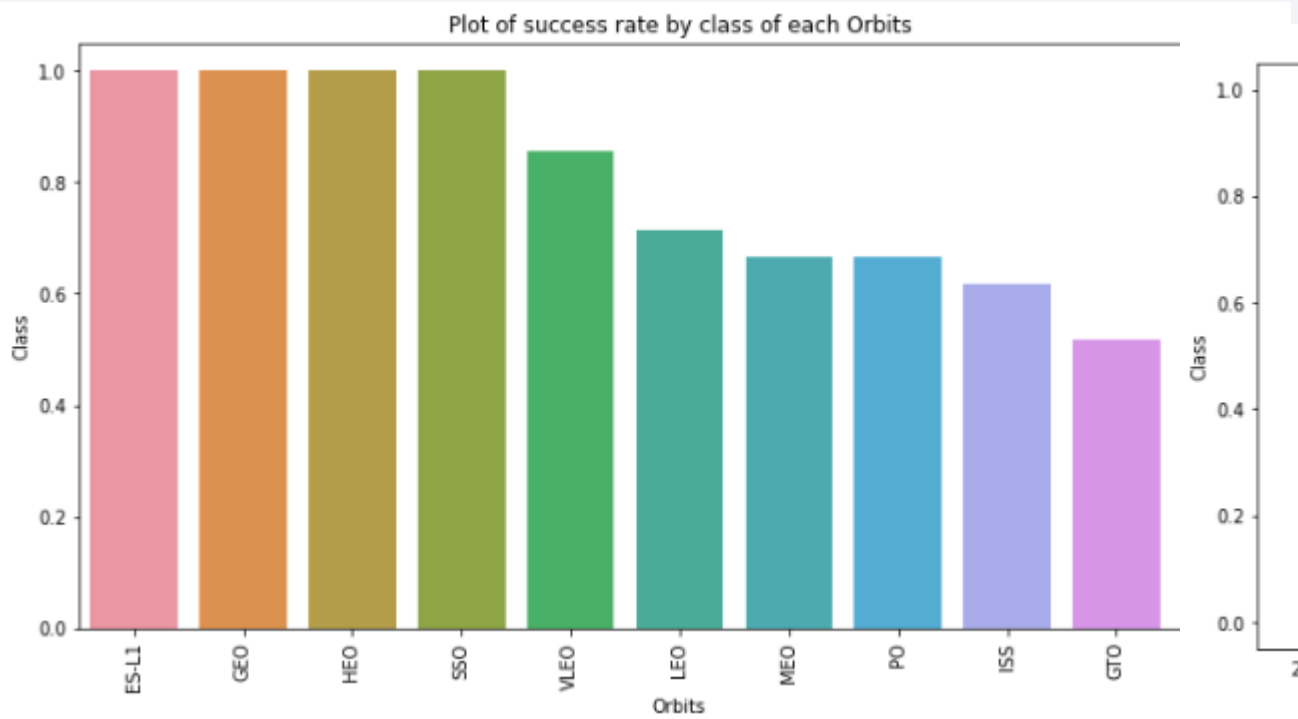
- EDA was performed and training labels were determined.
- The number of launches at each site, and the number and occurrence of each orbits were calculated
- Landing outcome label from outcome column was made. The result was exported to csv.
- The notebook is stored on the GitHub with this link:

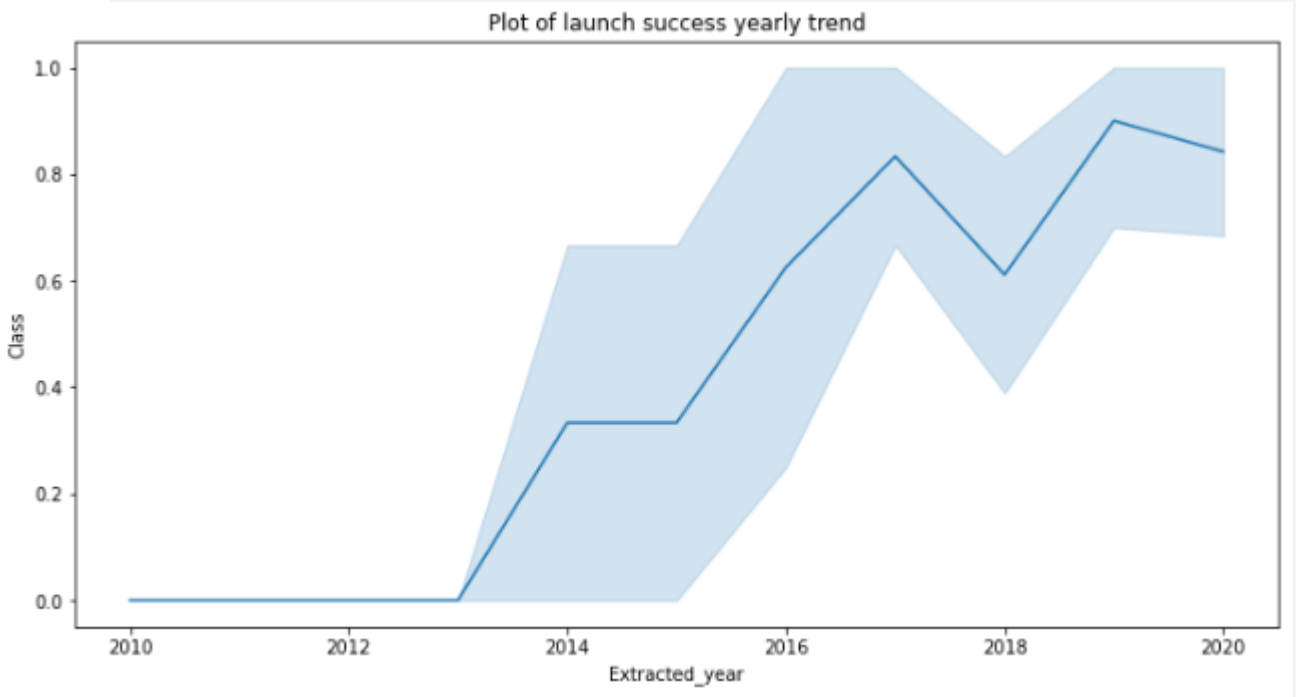
https://github.com/unyil01/edx-IBM-capstone/blob/main/03%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Data visualization was performed 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.
- Here is the notebook link on GitHub: https://github.com/unyil01/edx-IBM-capstone/blob/main/05%20EDA%20with%20Data%20Visualization.ipynb





EDA with SQL

- The SpaceX dataset was exported into a PostgreSQL database.
- EDA with SQL was implemented to get insight from the data. The queries were constructed:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- Here is the link of the notebook https://github.com/unyil01/edx-IBM-capstone/blob/main/04%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- 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.
- The feature launch outcomes (failure or success) were transformed into class 0 and 1.i.e., 0 for failure, and 1 for success.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- The distances between a launch site to its proximities were computed for answering such questions:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.
- The notebook link is https://github.com/unyil01/edx-IBM-capstone/blob/main/06%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Creating an interactive dashboard with Plotly dash
- Plotting pie charts presenting the total launches by a certain sites
- Plotting scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The link to the notebook is
- https://github.com/unyil01/edx-IBM-capstone/blob/main/07%20spacex_dash_app.py

Predictive Analysis (Classification)

- Loading the data using numpy and pandas, transformed the data, split our data into training and testing.
- Building different machine learning models and tune different hyperparameters using GridSearchCV.
- Using accuracy as the metric for our model, improving the model using feature engineering and algorithm tuning.
- Selecting the best performing classification model.
- Here is the link to the notebook
    - https://github.com/unyil01/edx-IBM-capstone/blob/main/08%20Machine%20Learning%20Prediction.ipynb

Results

Exploratory data analysis results:

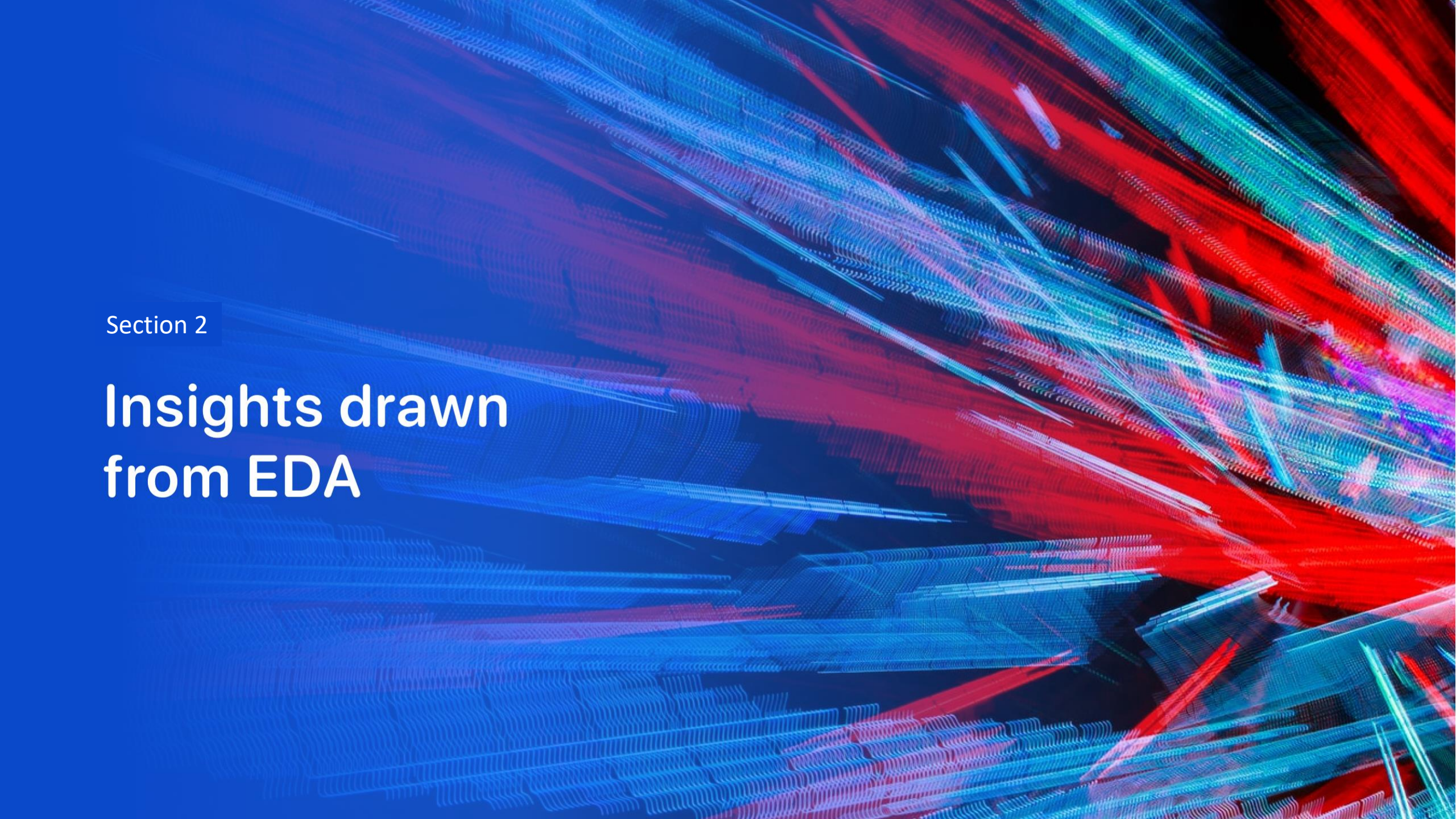
- Space X uses 4 different launch sites;
- The first launches were done to Space Xitself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

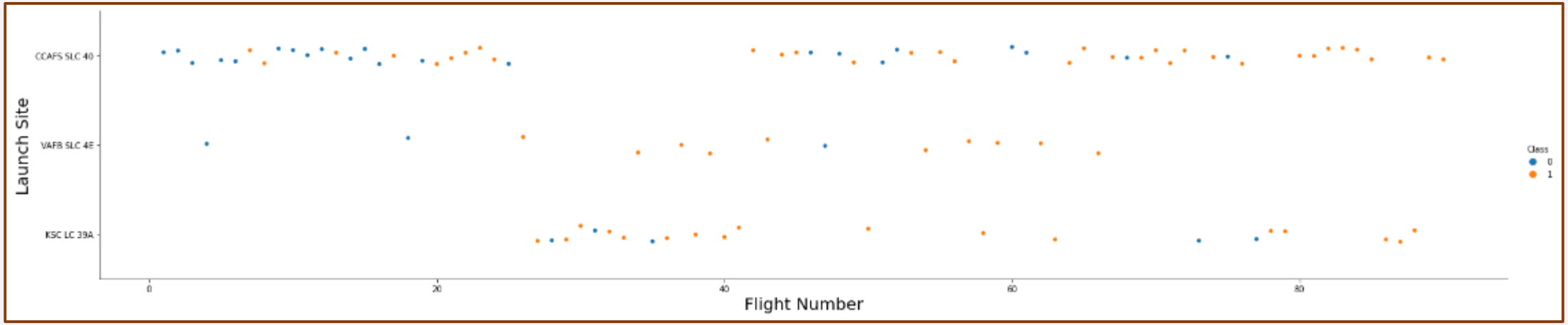
- Interactive analytics demo in screenshots
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.

Results

Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 94%.

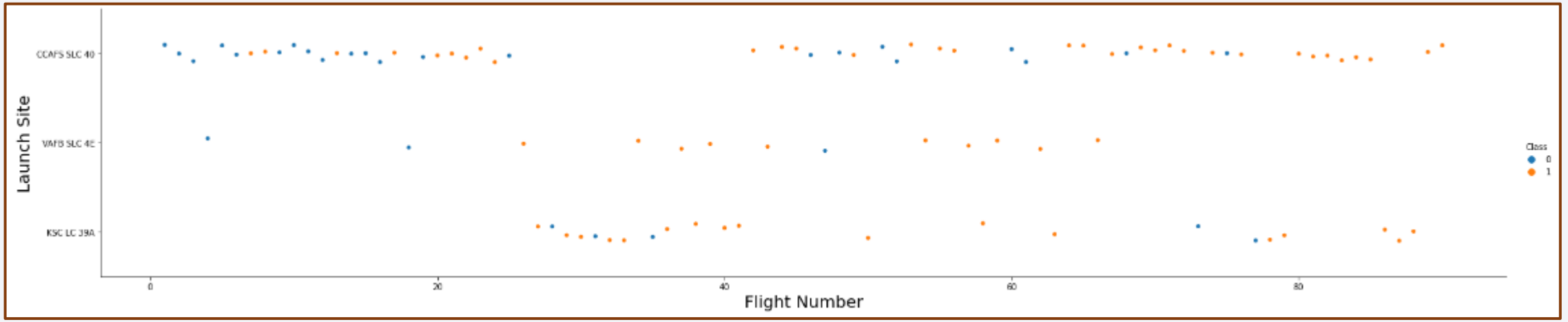

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



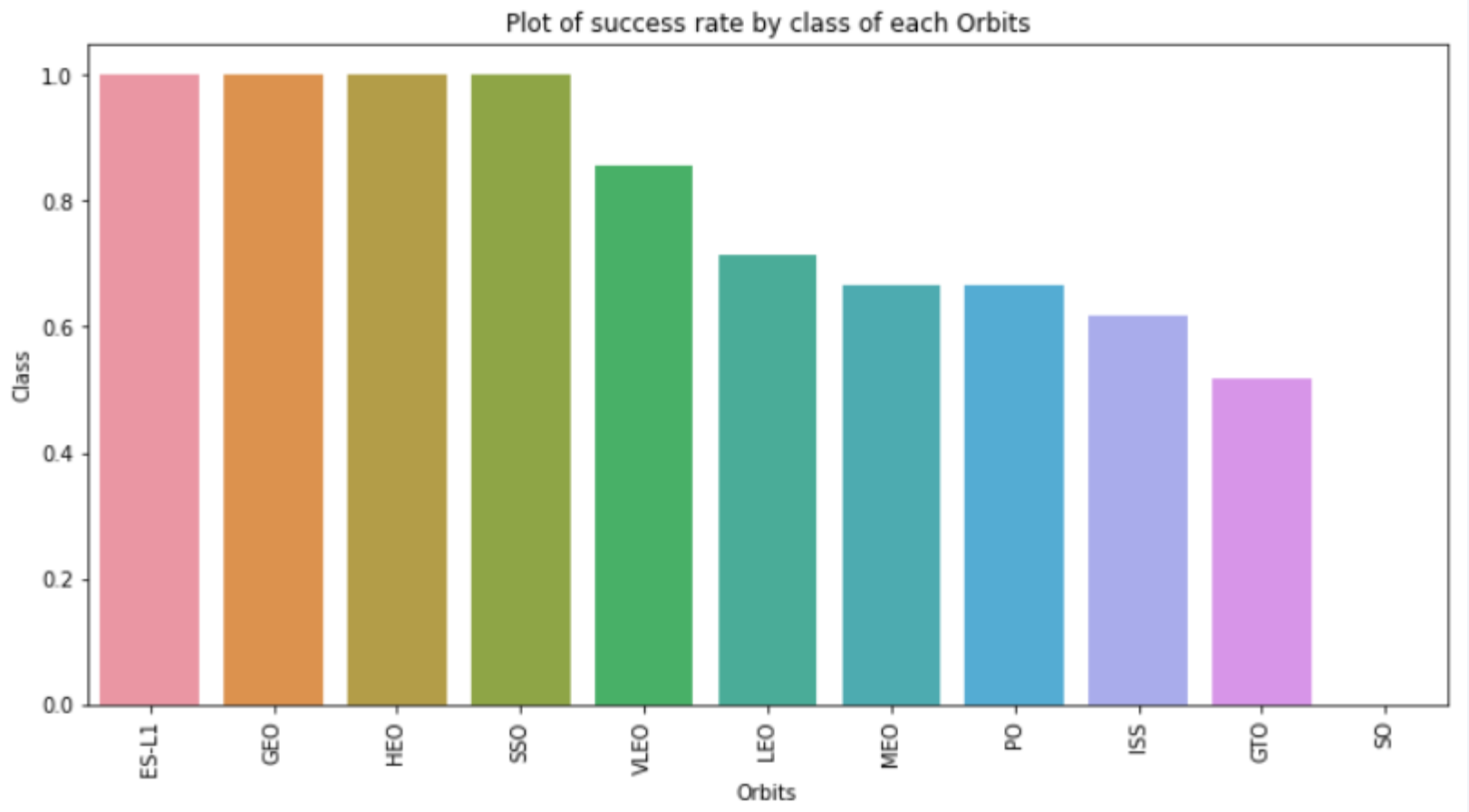
- The plot indicates that the larger the flight amount at a launch site, the greater the success rate at a launch site.
- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful;
- In second place VAFB SLC 4E and third place KSC LC 39A;
- It's also possible to see that the general success rate improved over time.

Payload vs. Launch Site



- Payloads over 9,000kg have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.

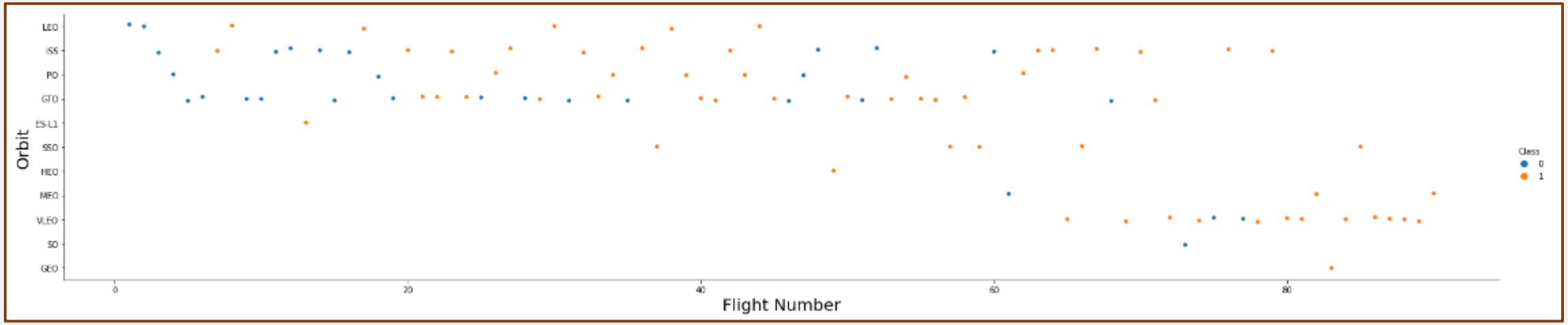
Success Rate vs. Orbit Type



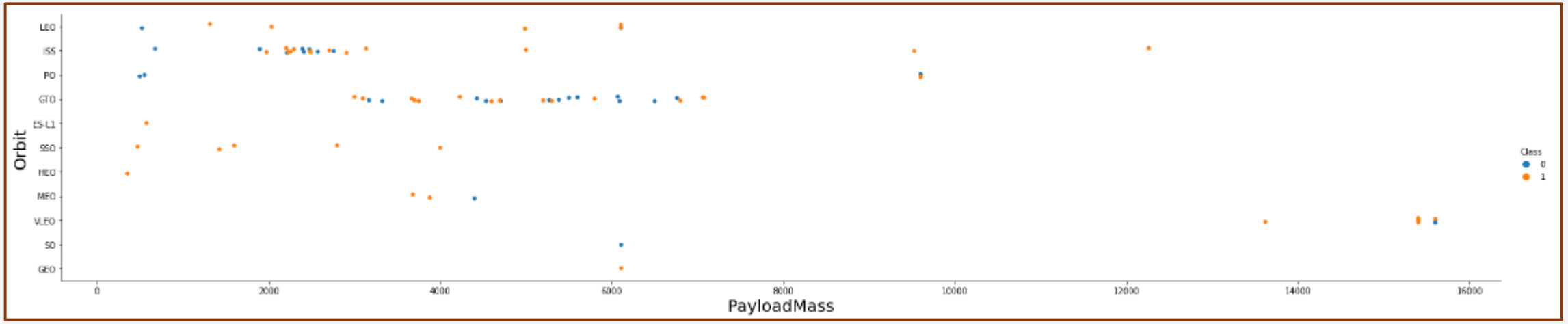
- The plot indicates that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.

Flight Number vs. Orbit Type

The plot below shows the Flight Number vs. Orbit type. It appears that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.

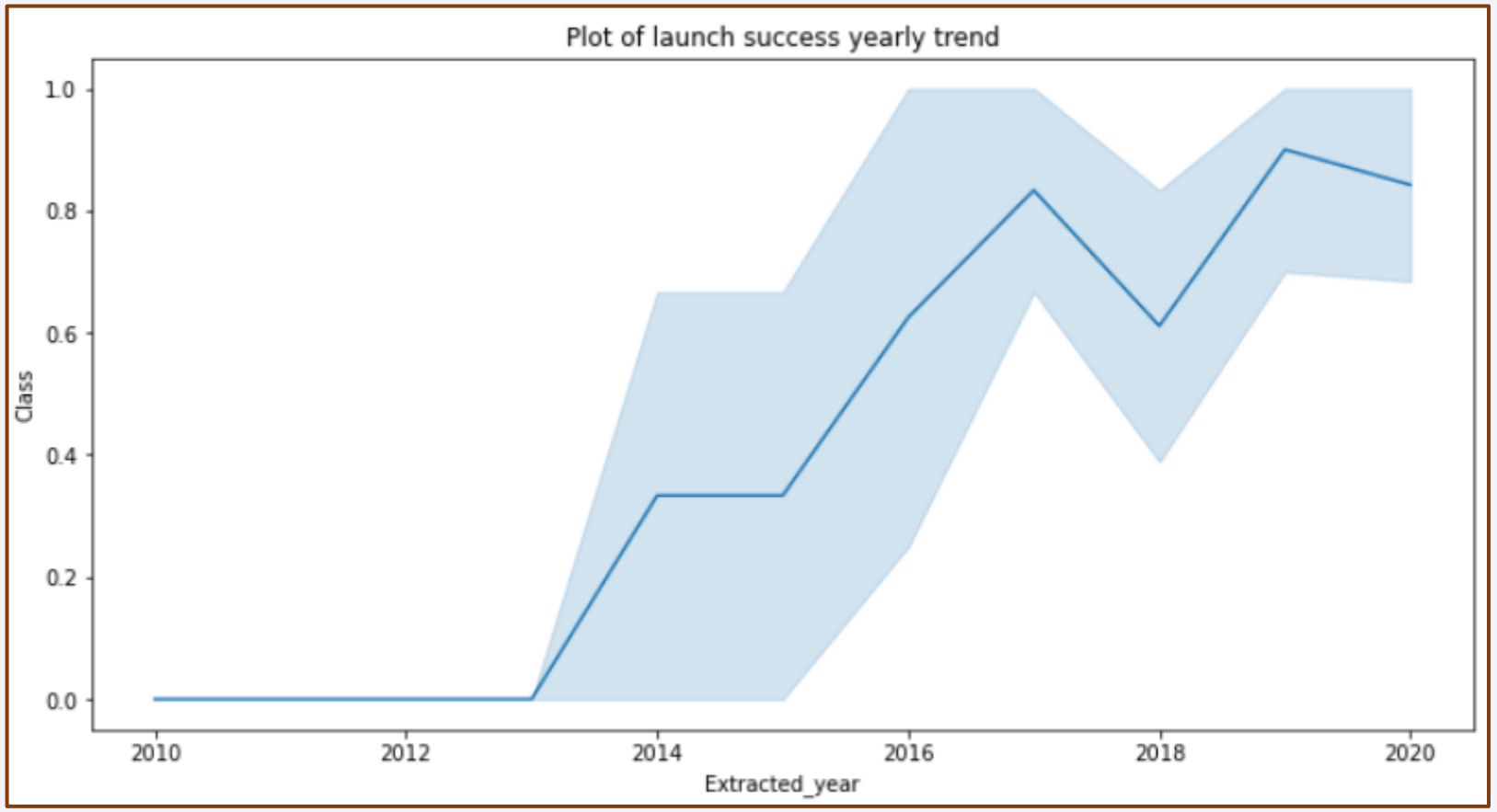


Payload vs. Orbit Type



- The relationship between payload and success rate to orbit GTO was not apparent;
- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.

Launch Success Yearly Trend



- Success rate started increasing in 2013 and kept until 2020;
- The first three years were a period of adjusts and improvement of technology.

All Launch Site Names

- We used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.

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

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

	launchsite
0	KSC LC-39A
1	CCAFS LC-40
2	CCAFS SLC-40
3	VAFB SLC-4E

Launch Site Names Begin with 'KSC'

The query was to present 5 records where launch sites begin with `CCA`

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

In [11]:
```
task_2 = '''
        SELECT *
        FROM SpaceX
        WHERE LaunchSite LIKE 'CCA%'
        LIMIT 5
        '''
create_pandas_df(task_2, database=conn)
```

Out[11]:

	date	time	boosterversion	launchsite	payload	payloadmasskg	orbit	customer	missionoutcome	landingoutcome
0	2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

The total payload carried by boosters from NASA as 45596 was computed using the query below.

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

In [12]:
```
task_3 = '''
        SELECT SUM(PayloadMassKG) AS Total_PayloadMass
        FROM SpaceX
        WHERE Customer LIKE 'NASA (CRS)'
        '''
create_pandas_df(task_3, database=conn)
```

Out[12]:

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 as 2928.4 was calculated

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

In [13]:
```
task_4 = '''
        SELECT AVG(PayloadMassKG) AS Avg_PayloadMass
        FROM SpaceX
        WHERE BoosterVersion = 'F9 v1.1'
        '''
create_pandas_df(task_4, database=conn)
```

Out[13]:

	avg_payloadmass
0	2928.4

First Successful Ground Landing Date

The dates of the first successful landing outcome on ground pad was 22nd December 2015 were identified.

In [14]:
```
task_5 = '''
        SELECT MIN(Date) AS FirstSuccessfull_landing_date
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Success (ground pad)'
        '''
create_pandas_df(task_5, database=conn)
```

Out[14]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

In [15]:

```
task_6 = '''
        SELECT BoosterVersion
        FROM SpaceX
        WHERE LandingOutcome = 'Success (drone ship)'
            AND PayloadMassKG > 4000
            AND PayloadMassKG < 6000
        '''
create_pandas_df(task_6, database=conn)
```

Out[15]:

	boosterversion
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Wildcard like '%' was used to filter for **WHERE** MissionOutcome was a success or a failure.

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:

The total number of successful mission outcome is:

	successoutcome
0	100

The total number of failed mission outcome is:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- Determining the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.
- List the names of the booster which have carried the maximum payload mass

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

In [17]:

```
task_8 = '''
        SELECT BoosterVersion, PayloadMassKG
        FROM SpaceX
        WHERE PayloadMassKG = (
                                SELECT MAX(PayloadMassKG)
                                FROM SpaceX
                                )
        ORDER BY BoosterVersion
        '''
create_pandas_df(task_8, database=conn)
```

Out[17]:

	boosterversion	payloadmasskg
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
5	F9 B5 B1051.3	15600
6	F9 B5 B1051.4	15600
7	F9 B5 B1051.6	15600
8	F9 B5 B1056.4	15600
9	F9 B5 B1058.3	15600
10	F9 B5 B1060.2	15600
11	F9 B5 B1060.3	15600

2015 Launch Records

- We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

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

In [18]:
```
task_9 = '''
        SELECT BoosterVersion, LaunchSite, LandingOutcome
        FROM SpaceX
        WHERE LandingOutcome LIKE 'Failure (drone ship)'
            AND Date BETWEEN '2015-01-01' AND '2015-12-31'
        '''
create_pandas_df(task_9, database=conn)
```

Out[18]:

	boosterversion	launchsite	landingoutcome
0	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Selecting Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- Using the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad))

In [19]:
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

Out[19]:

	landingoutcome	count
0	No attempt	10
1	Success (drone ship)	6
2	Failure (drone ship)	5
3	Success (ground pad)	5
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Precluded (drone ship)	1
7	Failure (parachute)	1

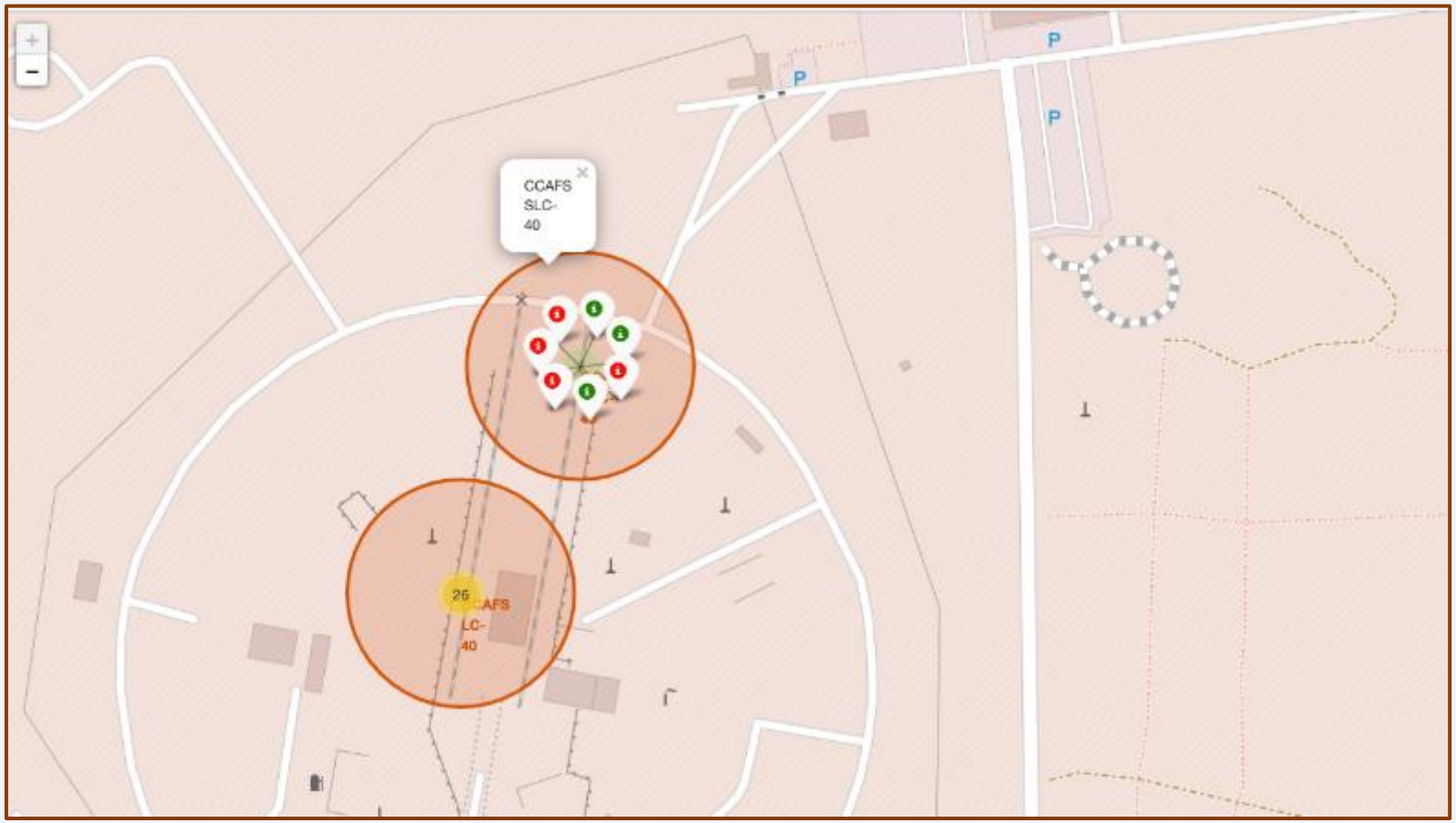
Section 3

Launch Sites Proximities Analysis

All launch sites

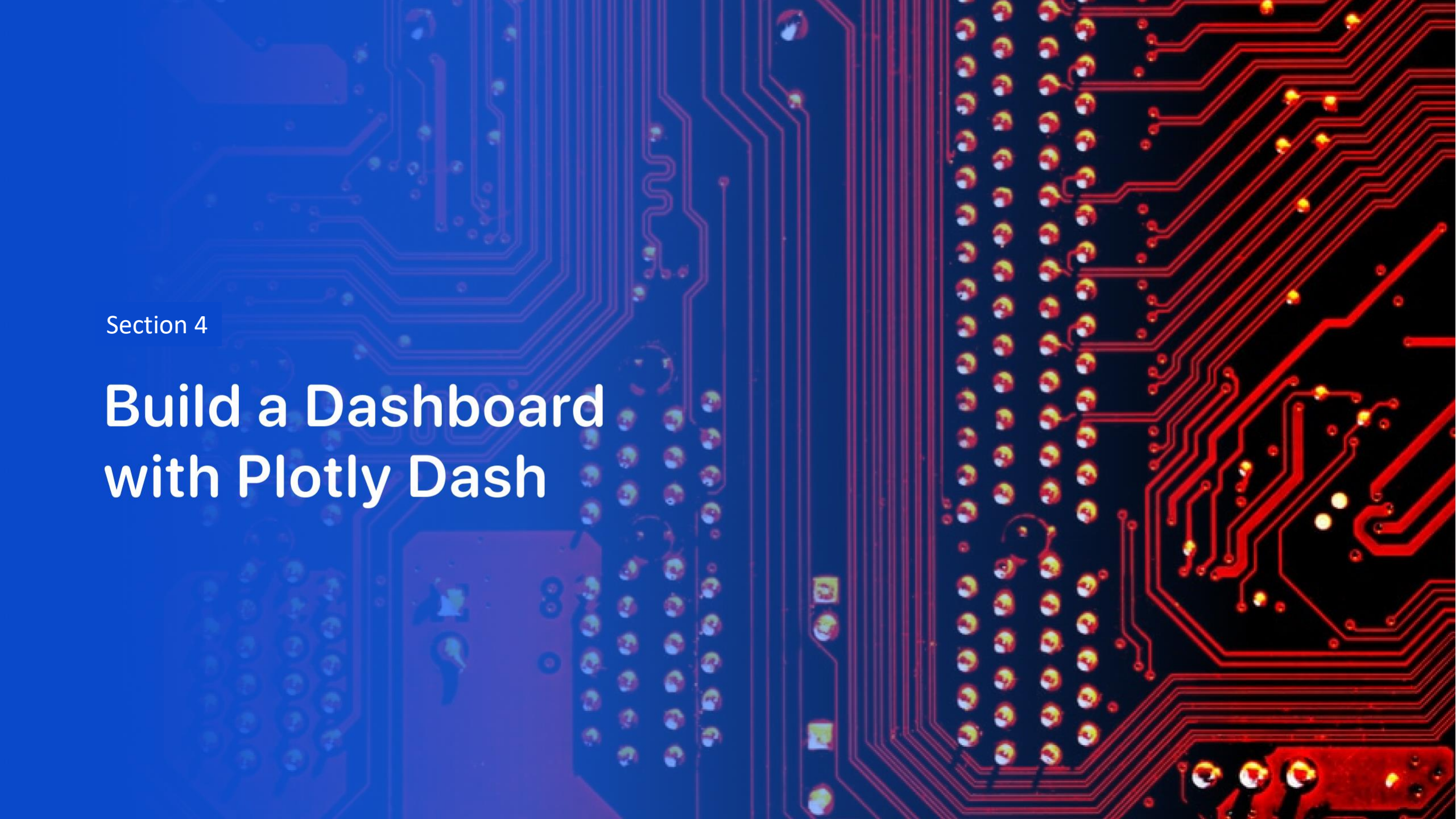
- Launch sites are near sea, probably by safety, but not too far from roads and railroads.

Markers showing successful and failure launches

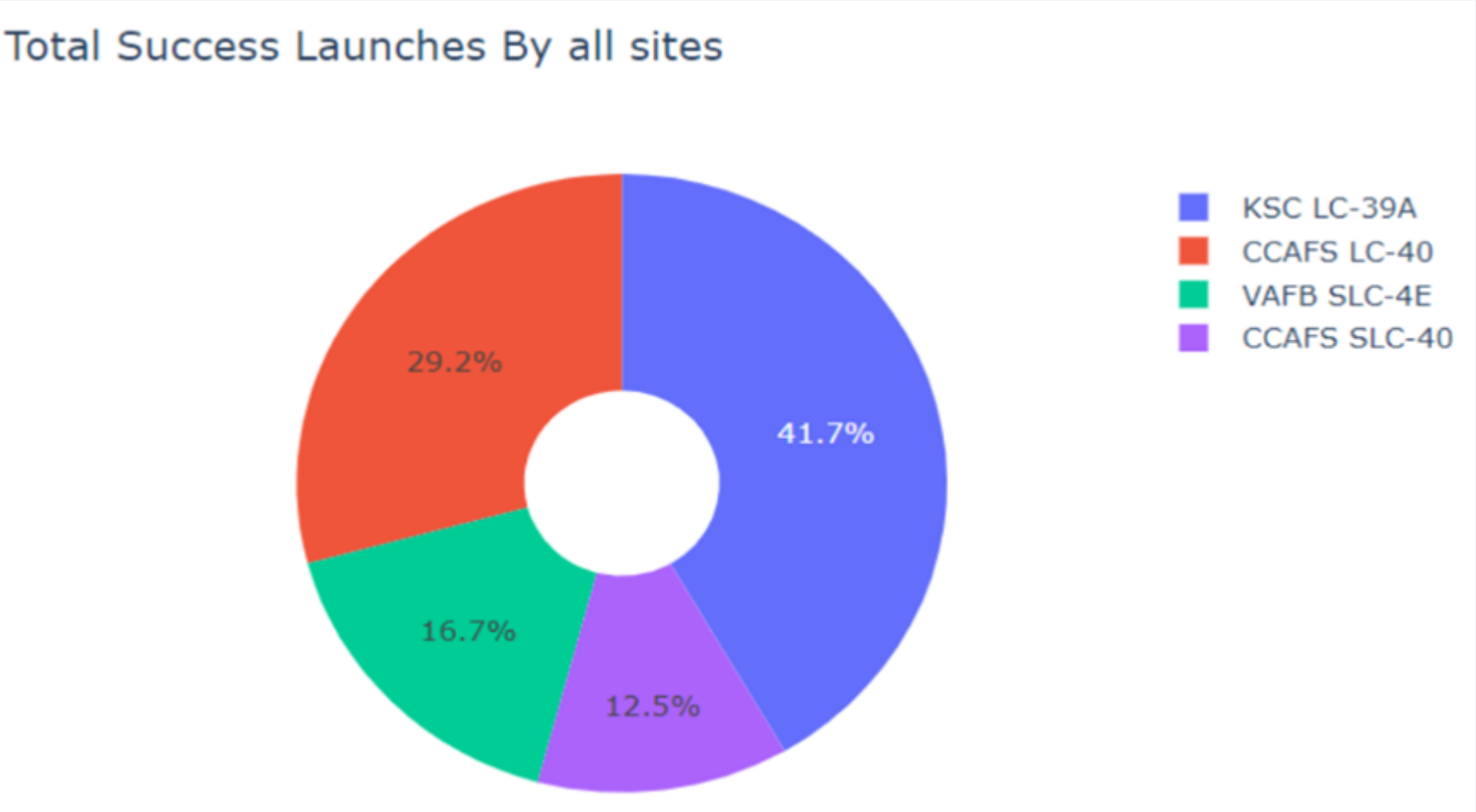


Green markers indicate successful Launces and Red markers failure ones.

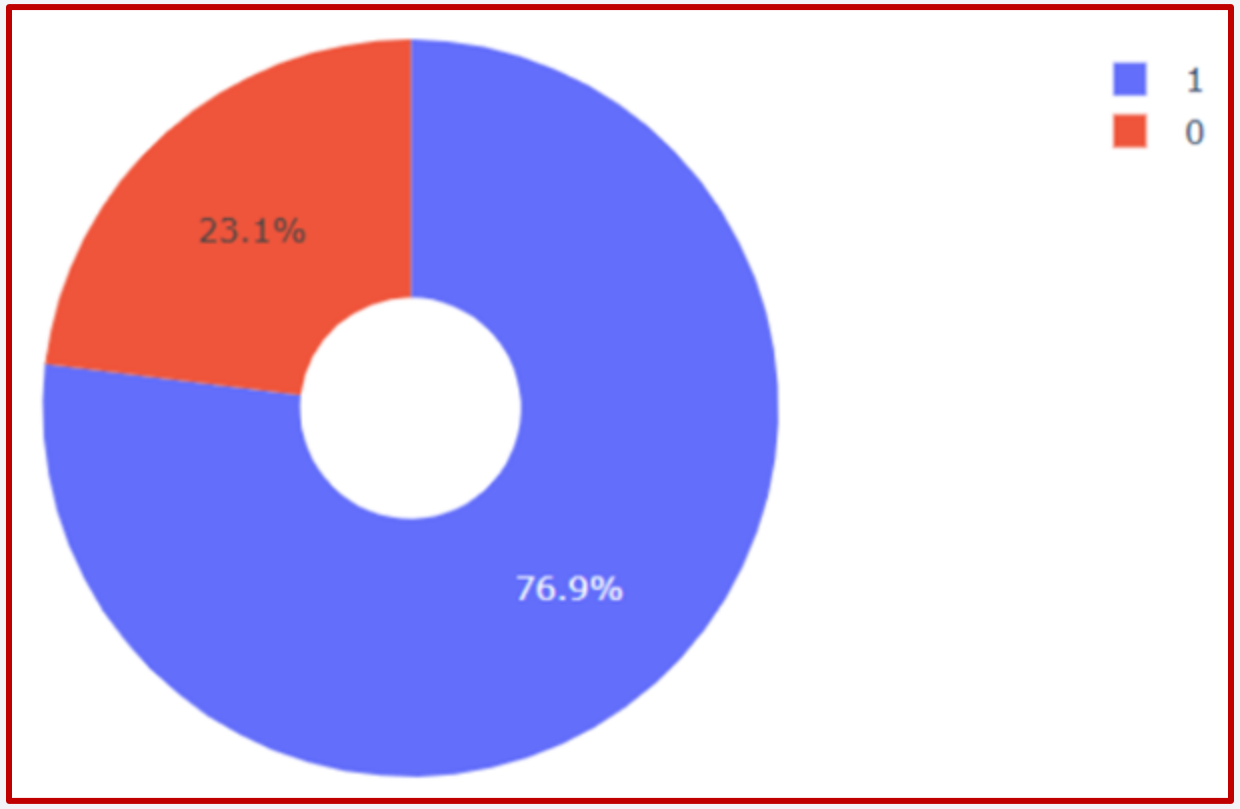
Launch Site distance to landmarks


Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

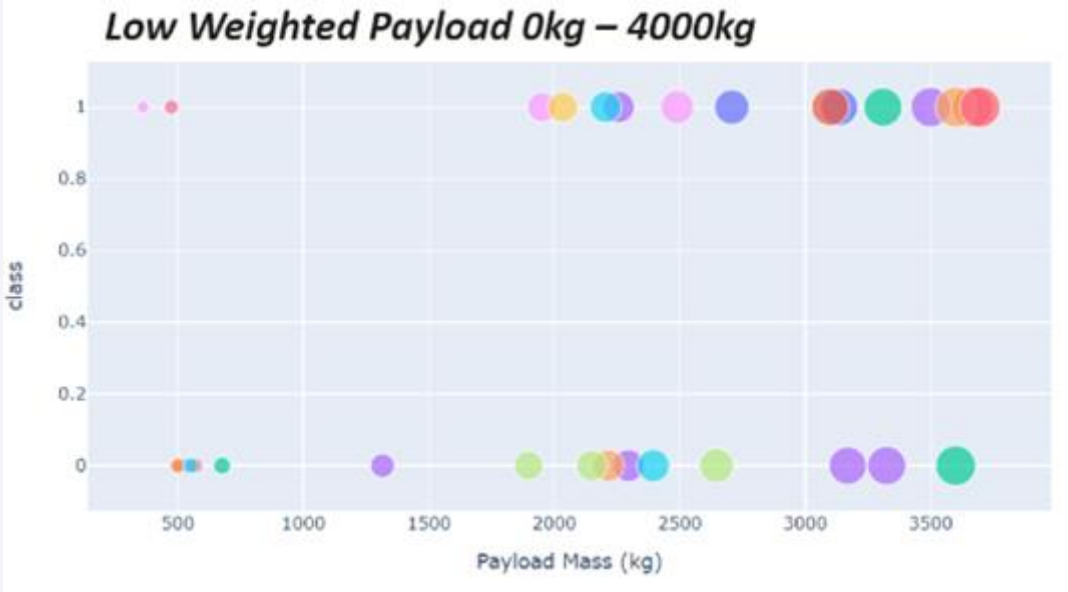


Pie chart showing the Launch site with the highest launch success ratio

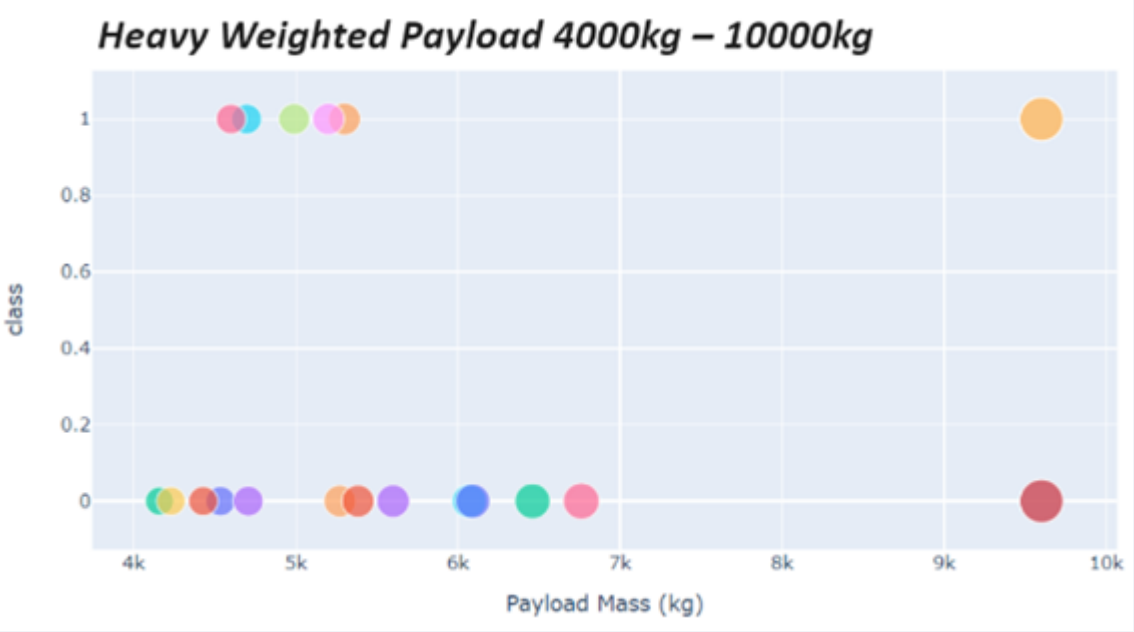


This pie chart shows KSC LC-39A with 76.9% success rate (blue) and 23.1% failure (red)

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



Success rate for low weighted payload is higher than heavy ones

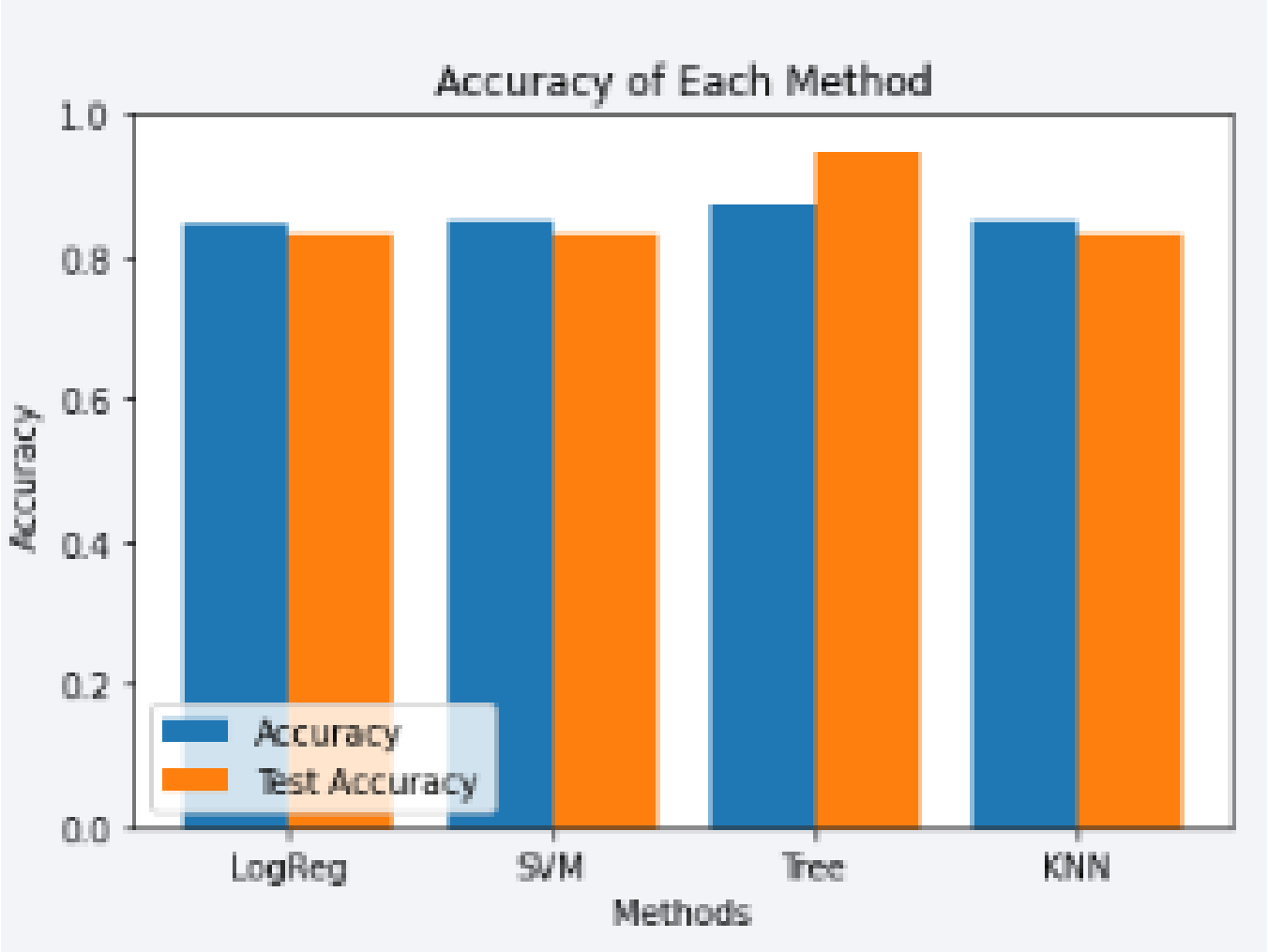


Section 5

Predictive Analysis (Classification)

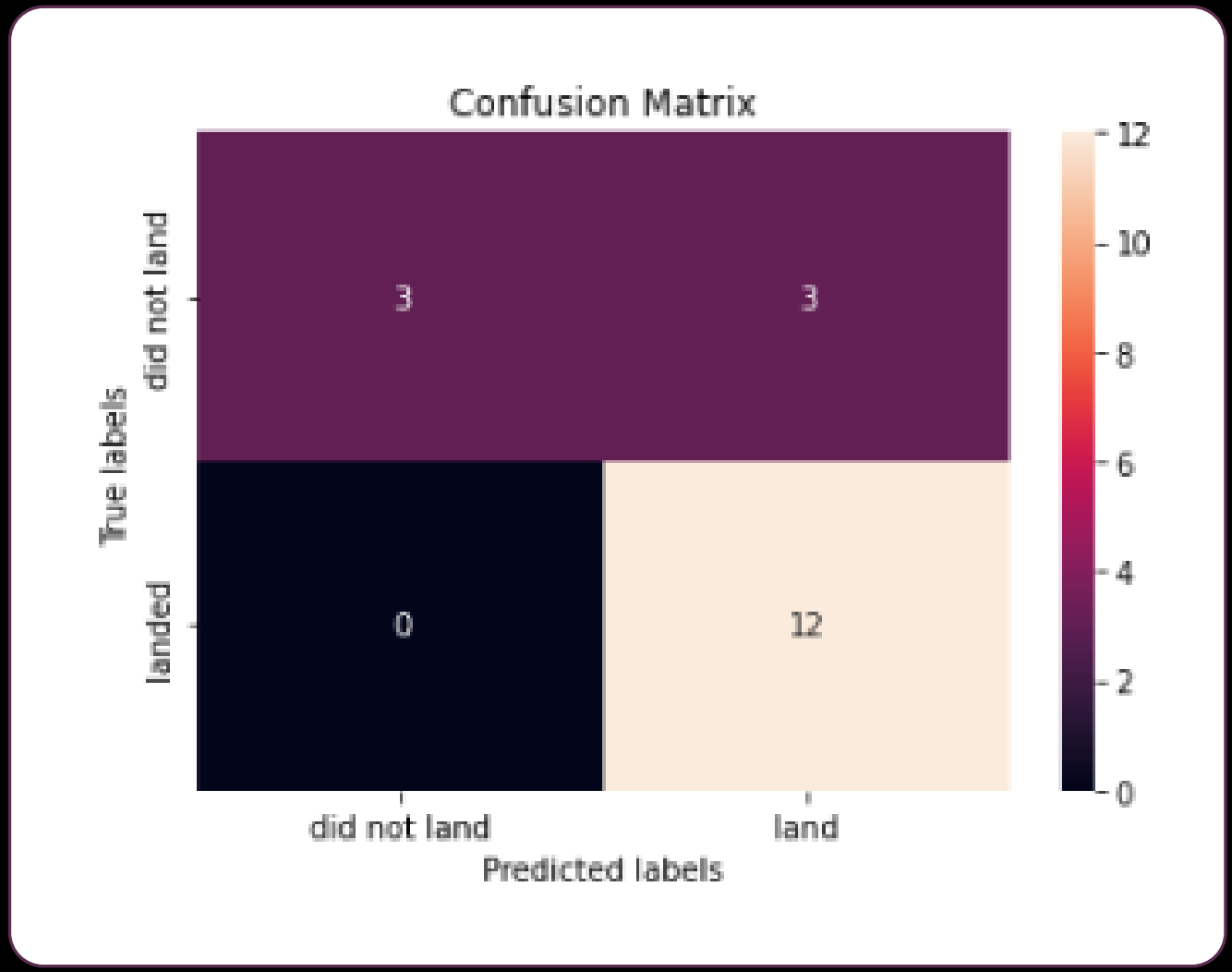
Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside;
- The model with the highest classification Accuracy, over than 87%, is Decision Tree Classifier



Confusion Matrix

- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- Successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- The Decision tree classifier is the best machine learning algorithm for this task.

Appendix

- All links of notebook presented in this report are provided in this repository
- https://github.com/unyil01/edx-IBM-capstone

Thank you!