IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

John Doe

July 12, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API & Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis using SQL and Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics
  - Predictive Analytics result

Introduction

Falcon 9 Launch Cost:

- SpaceX advertises Falcon 9 rocket launches at $62 million each.
- Competitors charge upwards of $165 million per launch.

Cost Efficiency:

- Significant savings due to SpaceX's ability to reuse the first stage of the rocket.

Project Goal:

- Develop a machine learning pipeline to predict the successful landing of the first stage.
- Utilize predictions to determine launch costs and assist alternate companies in bidding against SpaceX.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - Missing values were resolved, One-hot encoding was applied for categorized 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
  - Using GridSearchCV of scikit-learn to find the best parameters and using metrics like score and confusion metrics to find the best model

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

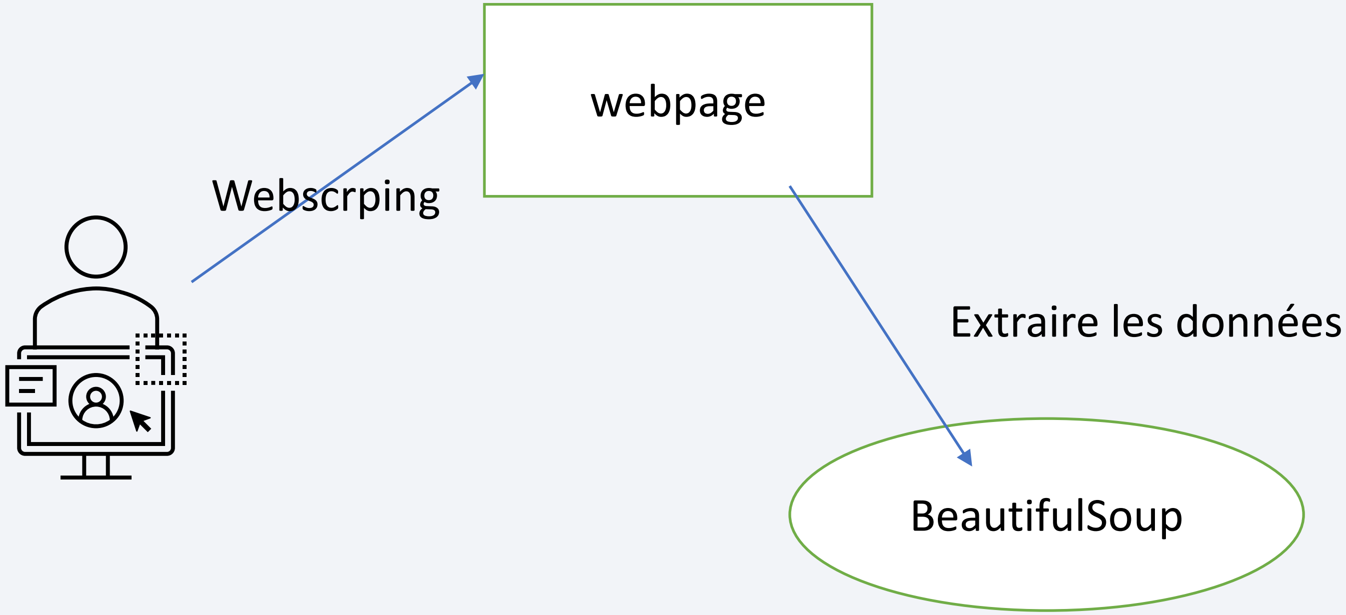
Data Collection – SpaceX API

- To collect and clean data, I used the get request to the SpaceX API to collect data.

Place your flowchart of SpaceX API calls here

Data Collection - Scraping

- We used web scrapping to access to Falcon 9 launch records and we used BeautifulSoup to extract the needed data



Data Wrangling

- Exploratory data analysis to determine the training labels.
- Calculate the number of launches at each site, and the number and occurrence of each orbits
- Used one-shot encoding for categorized data
- Removed inconsistent data
- Create landing outcome label from outcome column and exported findings to csv.

EDA with Data Visualization

- We used histograms and scatterplot to check relationship between features

EDA with SQL

Connected to DB2 using DB2 magic and performed the following SQL queries :

- names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- The failed landing outcomes in drone ship, their booster version and launch site names.
- Rank the count of landing outcomes

Build an Interactive Map with Folium

Launch Sites Marked:
Added map objects such as markers, circles, and lines to indicate launch success or failure.
Launch Outcomes:
Assigned outcomes to classes: 0 for failure and 1 for success.
Color-Labeled Clusters:
Identified launch sites with high success rates using color-coded markers.
Distance Calculations:
Measured distances between launch sites and their proximities.

Build a Dashboard with Plotly Dash

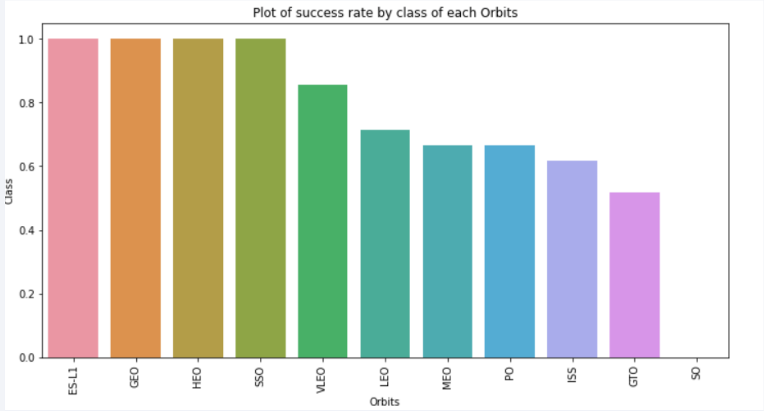
- pie charts showing the total launches by a certain sites
- scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version
- bar chart try to find which orbits have high success rate.

Predictive Analysis (Classification)

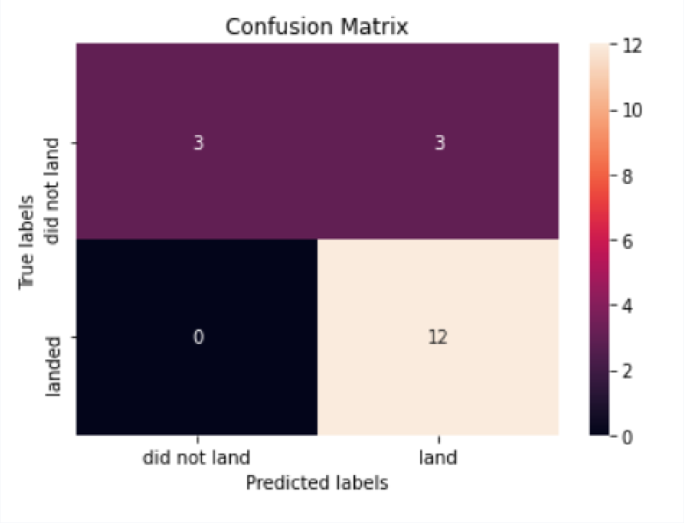
- We used the following process :
  - loaded the data : using numby, panda,s db2 magic and web scraping
  - prepared data : using python and SQL : this is by cleaning and removing inconsistent data
  - Developed our models and evaluated them : using scikitlearn by spliting our data into training and testing. We used prebuilt metrics to evalued our data
  - We tuned different hyperparameters using GridSearchCV.
  - We used accuracy metrics like score, accuracy, confusion matrics and more to evaluate our models
  - Metrics helped us to find the best performing classification model.

Results

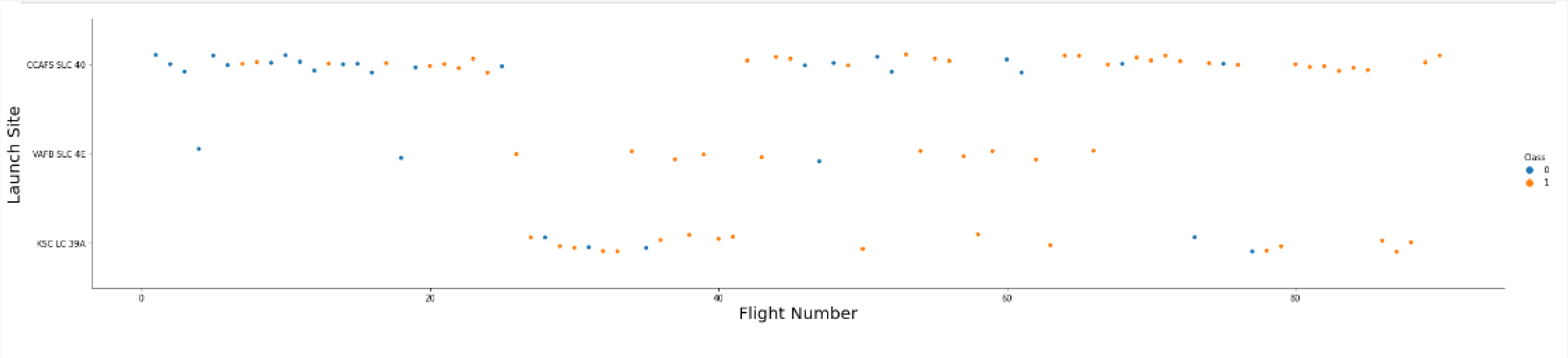
- Visualize the relationship between success rate of each orbit type



- Matrice de confusion



Visualize the relationship between Payload and Launch Site

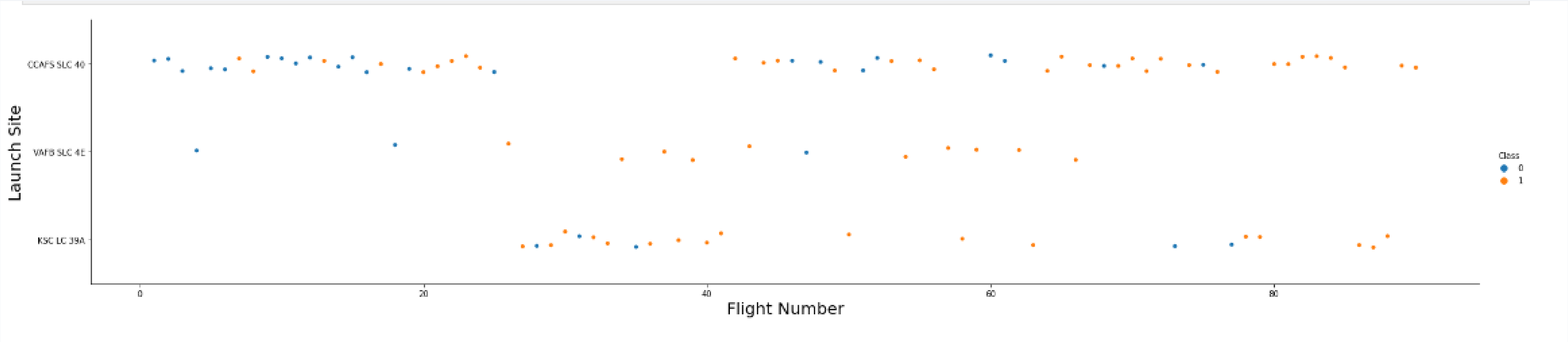


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

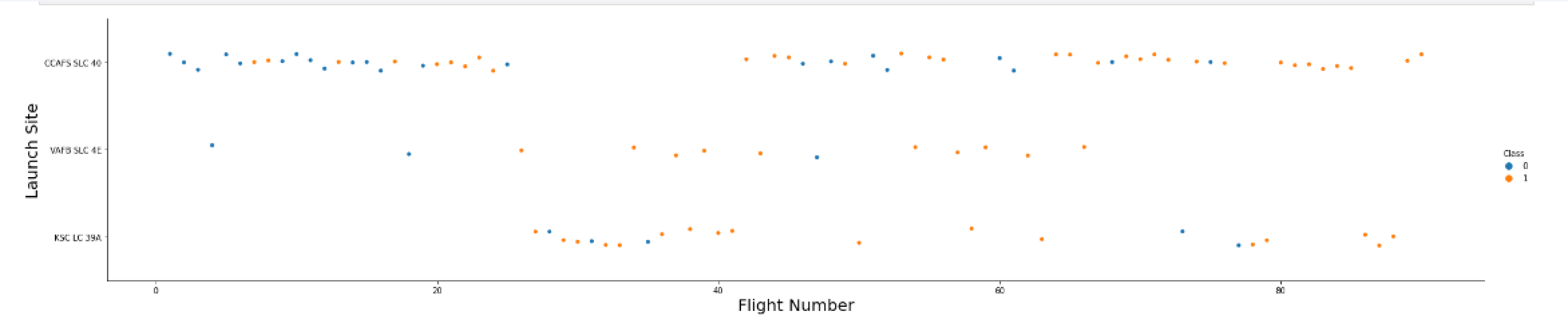
- Scatter plot of Flight Number vs. Launch Site



CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

Payload vs. Launch Site

- Scatter plot of Payload vs. Launch Site

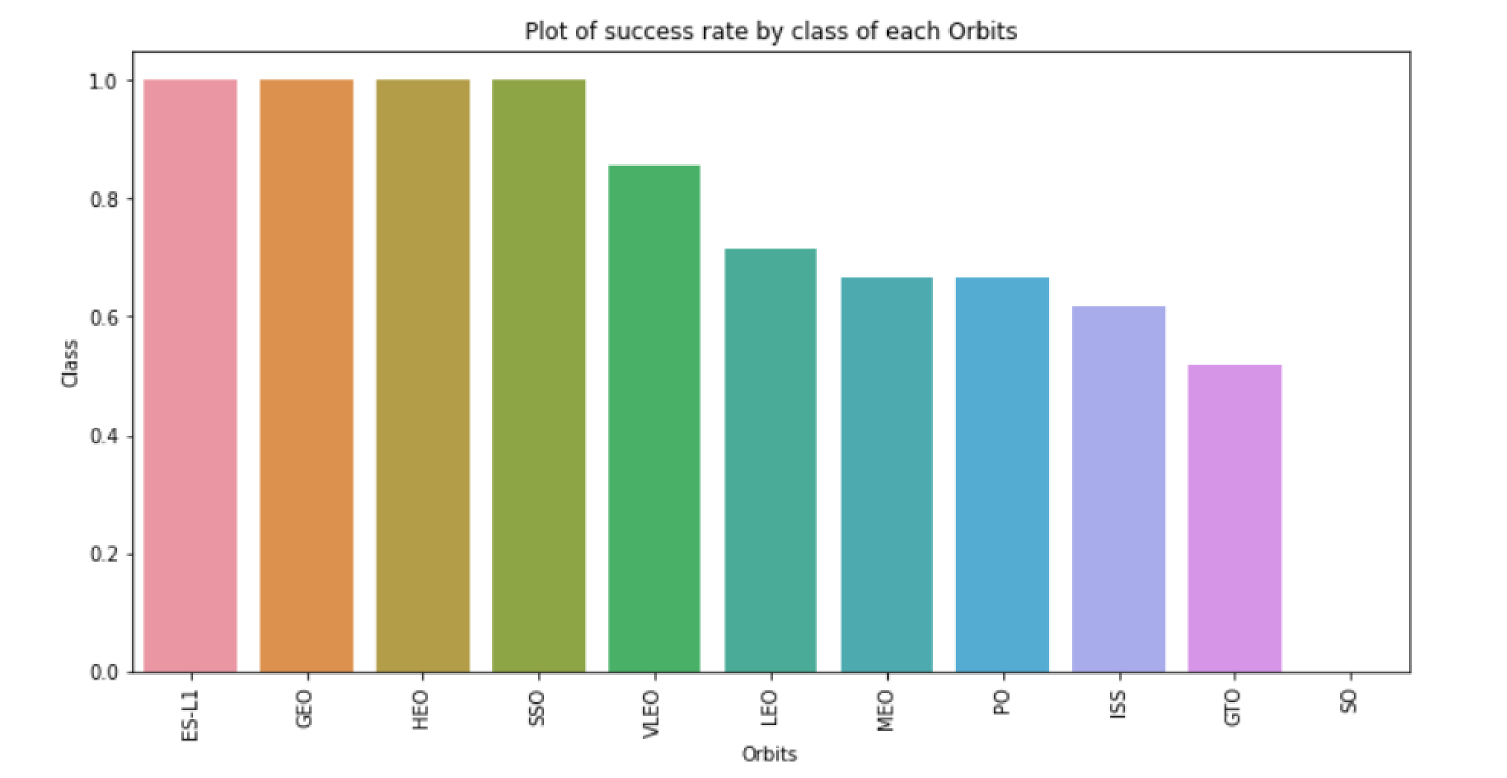


- for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type

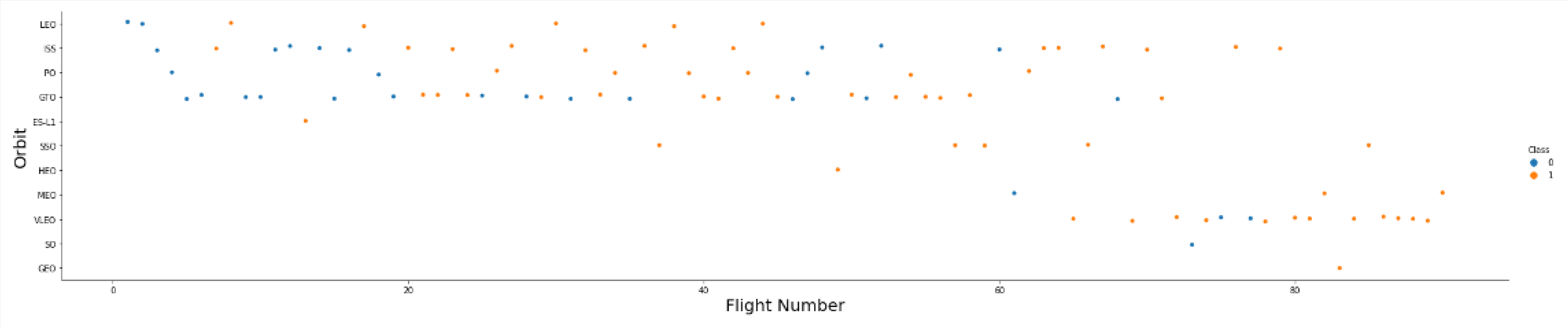
- Bar chart for the success rate of each orbit type

- ES-L1, GEO, HEO, SSO, and VLEO are the Orbits that have high success rate. The SO has the least success rate amongst the orbits.



Flight Number vs. Orbit Type

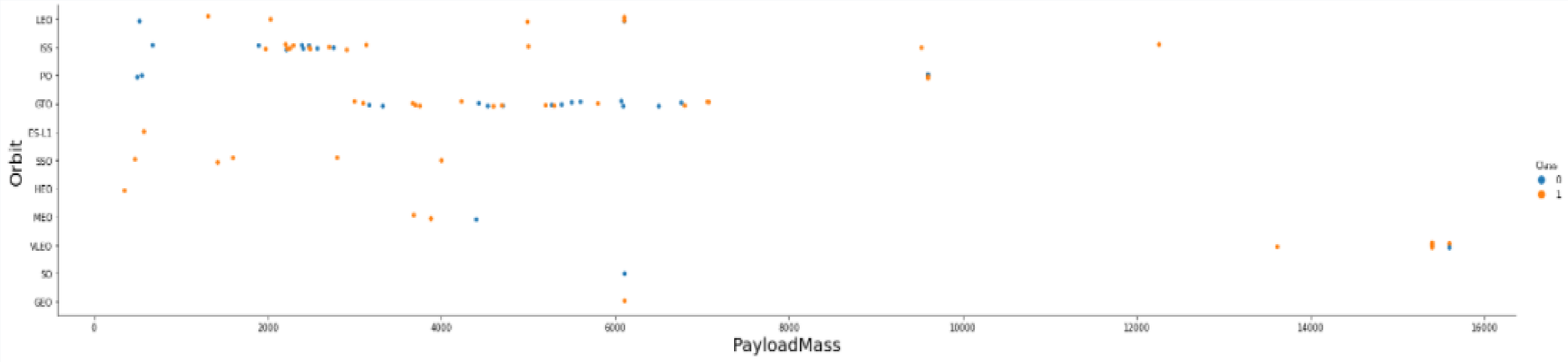
- Scatter point of Flight number vs. Orbit type



- LEO orbit the Success appears related to the number of flights
- no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

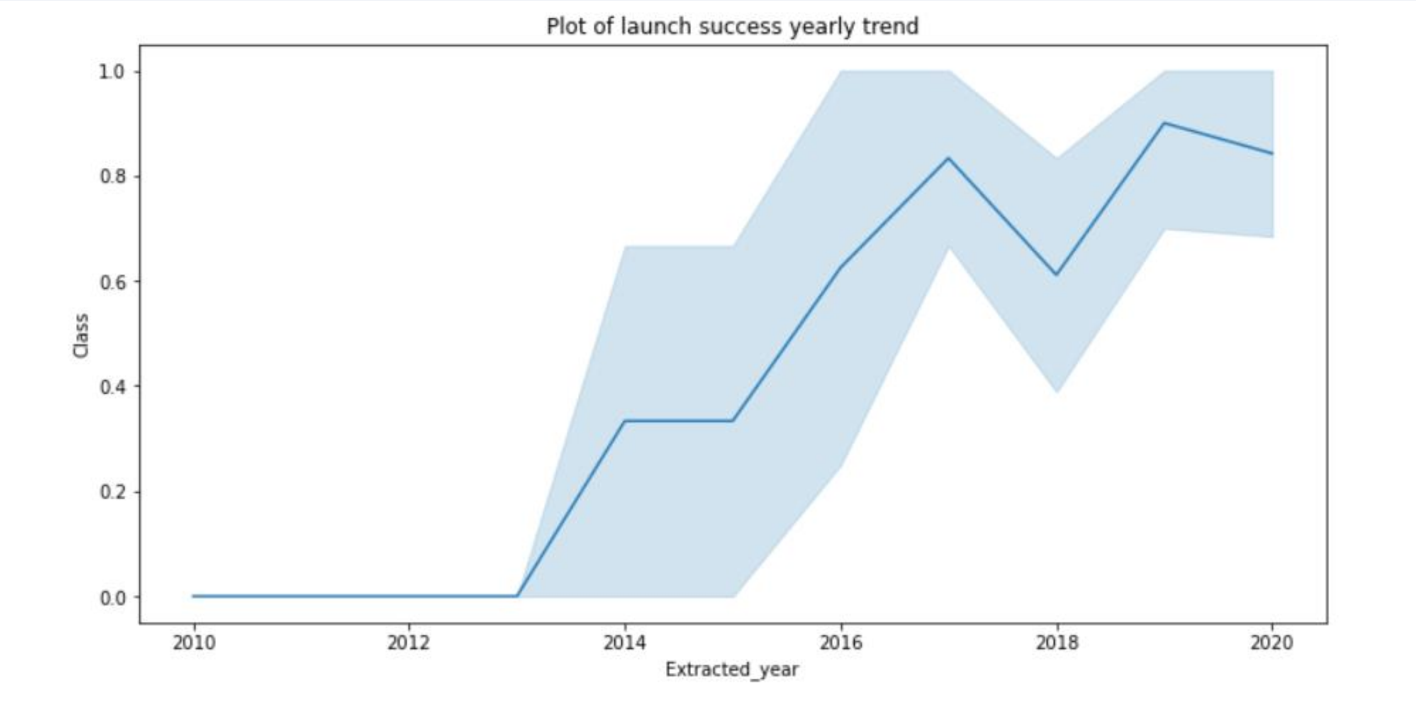
- Scatter point of payload vs. orbit type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

- Line chart of yearly average success rate
- Success rate since 2013 kept increasing till 2020



All Launch Site Names

- We have 4 uniques launches sites in the space mission. This is found by performing distinct in SELECT clause.

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 'CCA'

- Displaying 5 records where launch sites begin with the string 'CCA' by perfoming a filter in WHERE clause

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 is 45 596. This is calculed using the aggregated function SUM.

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

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

	total_payloadmass
0	45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2 928,4

Task 4

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 date of the first successful landing outcome on ground pad is 2015-12-22

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

]:

	firstsuccessfull_landing_date
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are : F9 FT B1022 ; F9 FT B1026 ; F9 FT B1021.2 and F9 FT B1031.2

List the names of the boosters which have success in drone ship and have 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

- The total number of successful is 100 and the total number of failure mission is1

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)
```

```
The total number of successful mission outcome is:
```

	successoutcome
0	100

```
The total number of failed mission outcome is:
```

Out[16]:

	failureoutcome
0	1

Boosters Carried Maximum Payload

- This query gave us the names of 11 boosters 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

- There are 2 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015. Both of these failures are drone ship

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

```
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)
```

	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

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **during the specied period**. Then we grouped the result and sorted it.

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

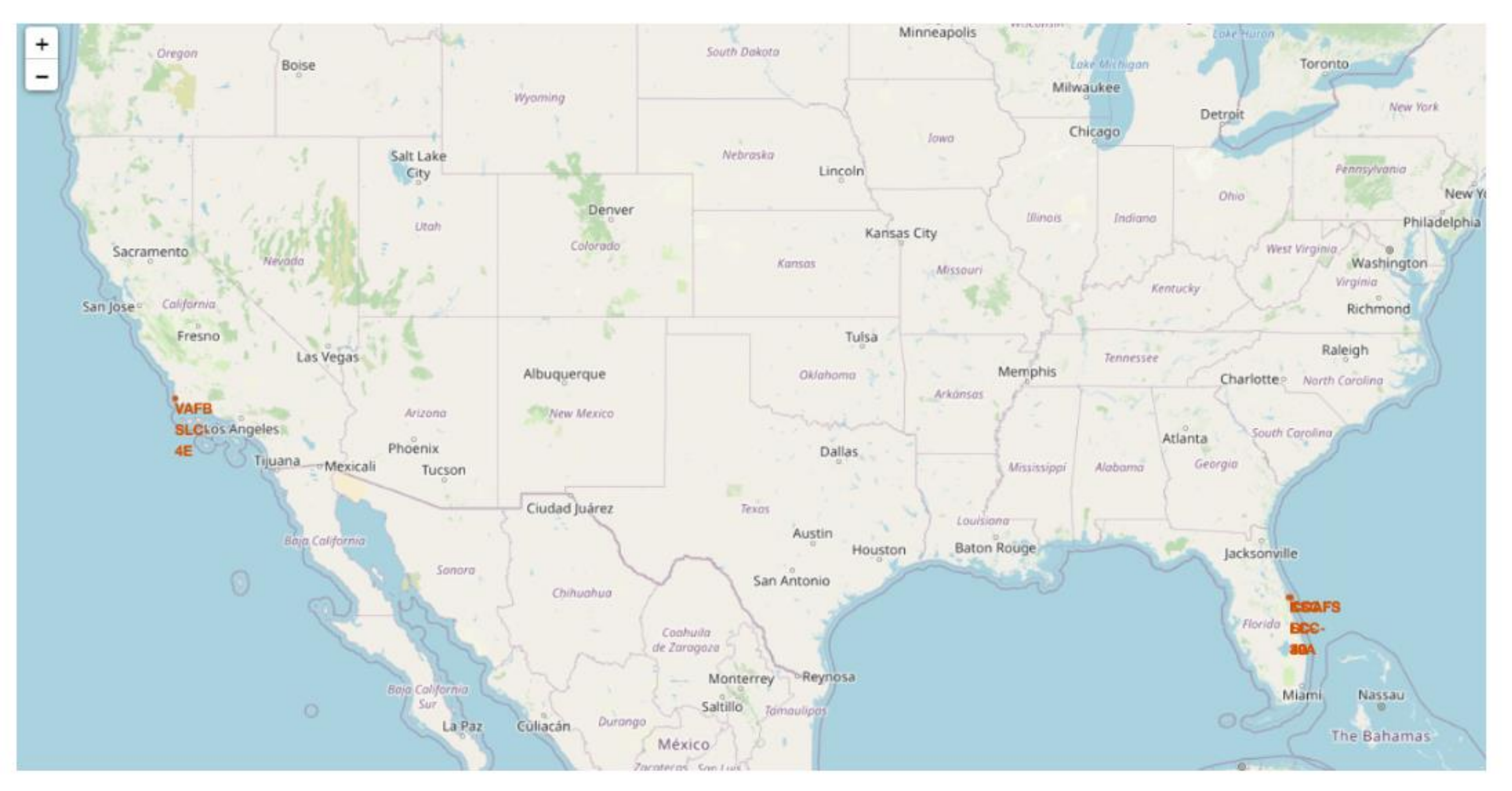
```
task_10 = '''
        SELECT LandingOutcome, COUNT(LandingOutcome)
        FROM SpaceX
        WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
        GROUP BY LandingOutcome
        ORDER BY COUNT(LandingOutcome) DESC
        '''
create_pandas_df(task_10, database=conn)
```

	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’ location markers map



All launch sites are in the USA

Markers launch sites with color labels

- Green for success lunches and red for failures

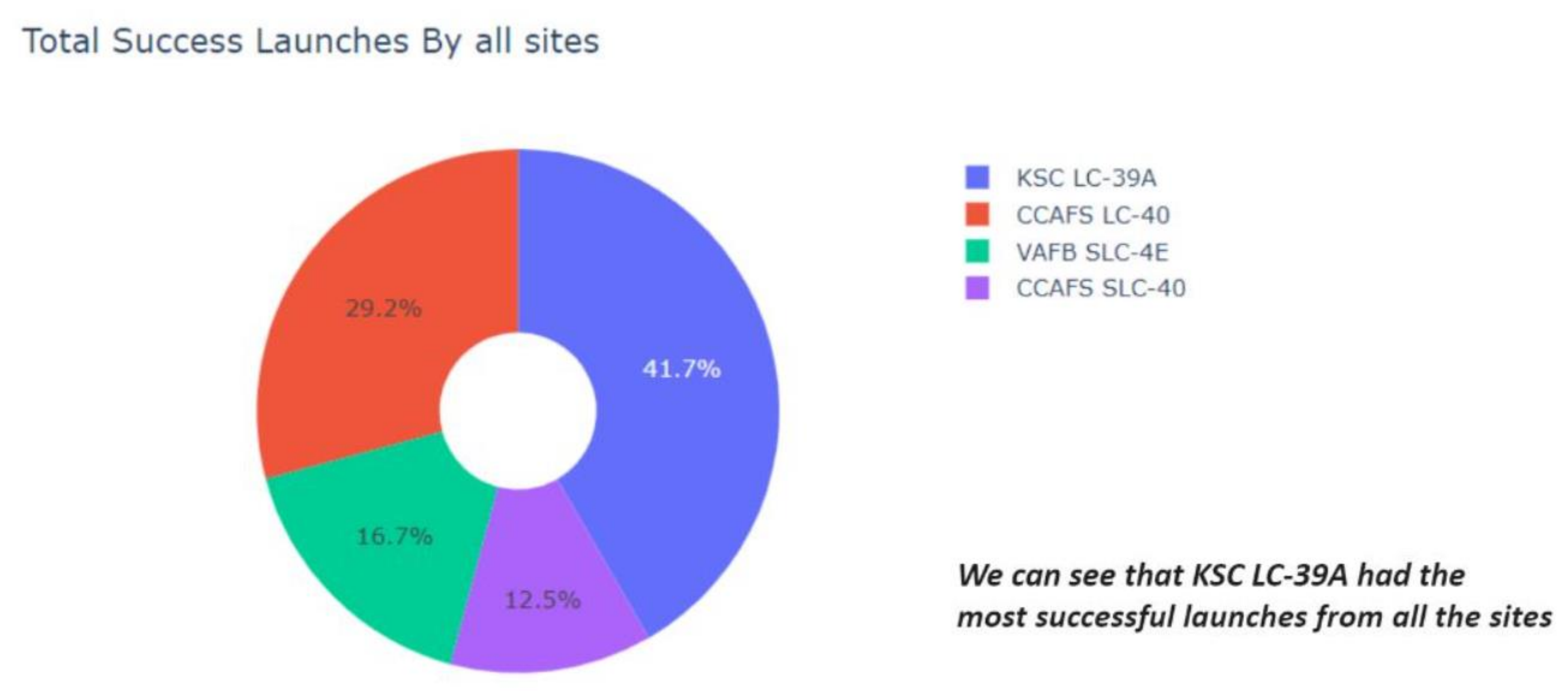
Launch Site distance to landmarks

Section 4

Build a Dashboard with Plotly Dash

Total success lauches by all sites

- KSC LC-39A represent the hight success 41,2% and VAFB SLC-4E represent the low success 16,7%.

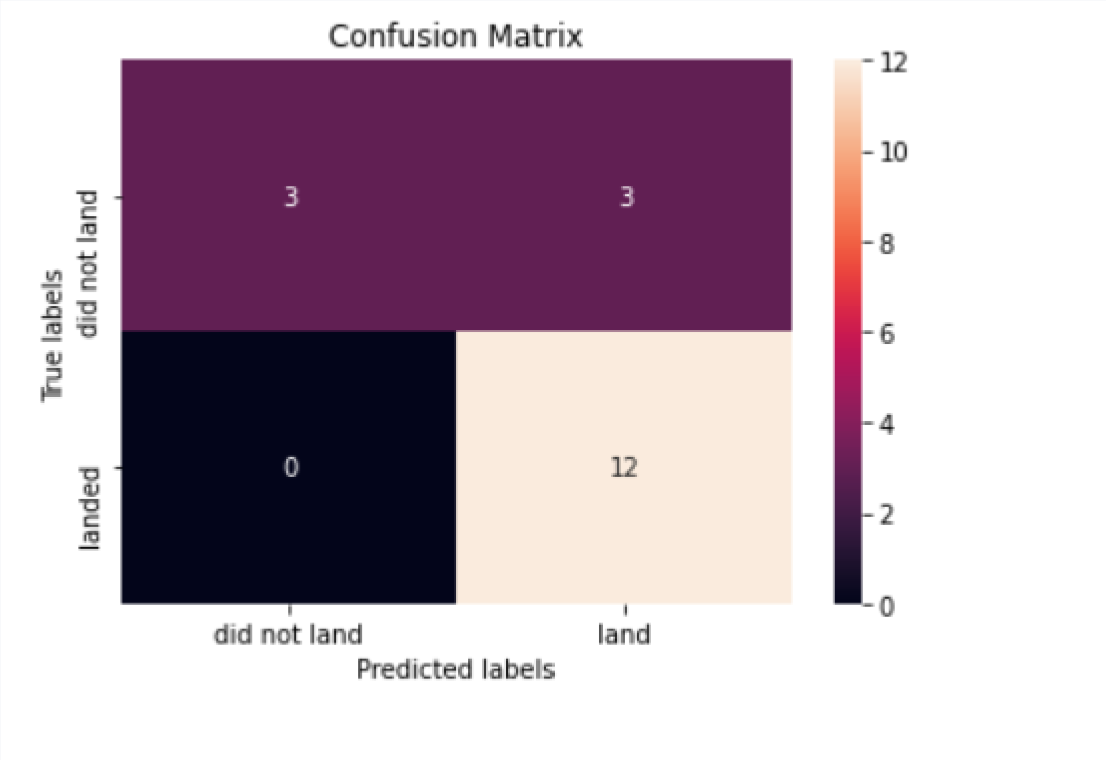


Section 5

Predictive Analysis (Classification)

Confusion Matrix

- Only three are not well classified.



Conclusions

- There are 4 unique sites
- Best algorithm for this problem : the Decision tree classifier with a score of 87%
- Launch success rate started increase from 2013 to 2020.
- Most successful : Orbits ES-L1, GEO, HEO, SSO, VLEO.
- For the 4 sites, the most successful lanches is KSC LC-39A

Appendix

- Here is how to connect to a database

```
# create connection to postgreSQL database
conn = psycopg2.connect(
    host = 'localhost',
    database = 'postgres',
    user = 'postgres',
    password = 'chuksoo',
    port = '5432')
print('Connection to database is successfully')
```

Thank you!