IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Javeria Majeed
28th Oct 2023

Outline

Executive Summary

- Summary of methodologies

Data Collection

- Data collection using API
- Data collection using web scrapping

Data Wrangling

Exploratory data analysis

- EDA with visualization
- EDA with SQL

Interactive visual analysis with folium

Machine Learning Prediction

- Summary of all results

Collected from multiple resources

Using EDA we were able to tell which features are important for prediction of successful launching

Predictive analytic result

Introduction

- Project background and context

Space X decided to launch falcon 9 at the cost of 65 million dollars, other providers cost up to 165 million dollars. Space X is reusing is first stage after successful landing, that's why it is able to launch falcon 9 at a low cost compare to others.

So if we determine if the first stage will land successfully or not, we will be able to find 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 goal of this project is to create a machine learning pipeline to predict if the first stage will land successfully or not.

- Problems you want to find answers

What condition favors successful launching

What is the best place launching

What operating conditions needs to be fulfilled in order to have a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using spaceX API
- Perform data wrangling
  - Data was analyzed by creating a landing outcome variable based on outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data was standardized first, then we divided it into training and testing set. After that we apply machine learning algorithm on training datasets. Lastly we test those model on our testing datasets and find the best model

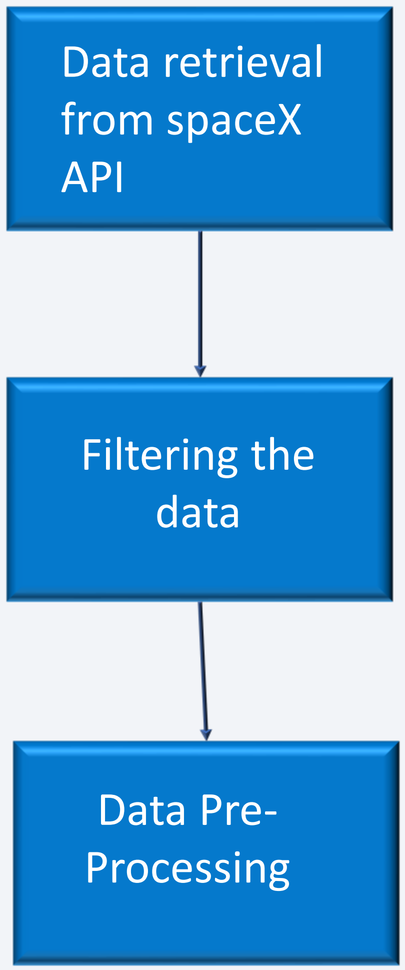
Data Collection

- Dataset was collected using spaceX API.
- Utilize the '.json_normalize' command, to convert the JSON data into a structured data frame
- Then we deal with the missing values in the data
- We also perfromed web scrapping fro falcon9 records with BeautifulSoup

Data Collection – SpaceX API

- Access the public API provided by SpaceX to retrieve data in .json format
- Utilize the '.json_normalize' command, to convert the JSON data into a structured data frame for further analysis.
- Filter the data frame to only include Falcon9 launches.
- Apply data preprocessing techniques (i.e. dealing with missing values)
- Link to the notebook

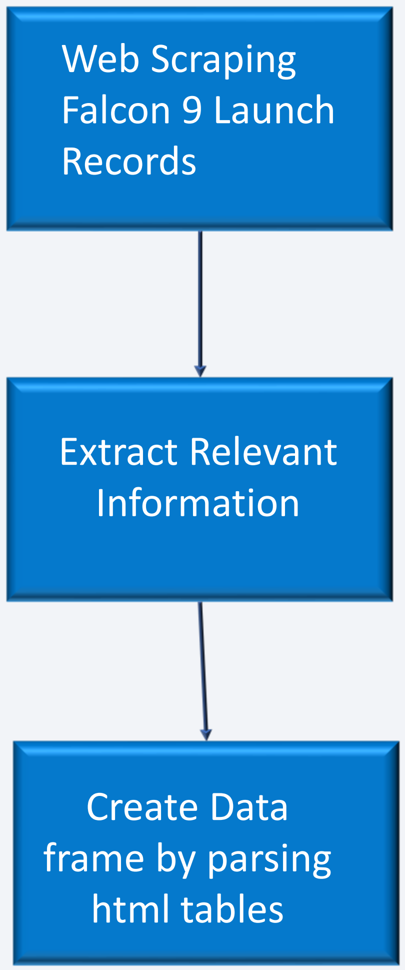
https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/Data%20Collection%20API.ipynb



Data Collection - Scraping

- Implement web scraping using Beautiful Soup to extract Falcon 9 launch records from the web.
- Extract data from the HTML table headers, including launch date, mission name, launch site, and other pertinent details.
- Utilize data parsing results to create a data frame that holds the Falcon 9 launch records.
- Link to notebook

https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

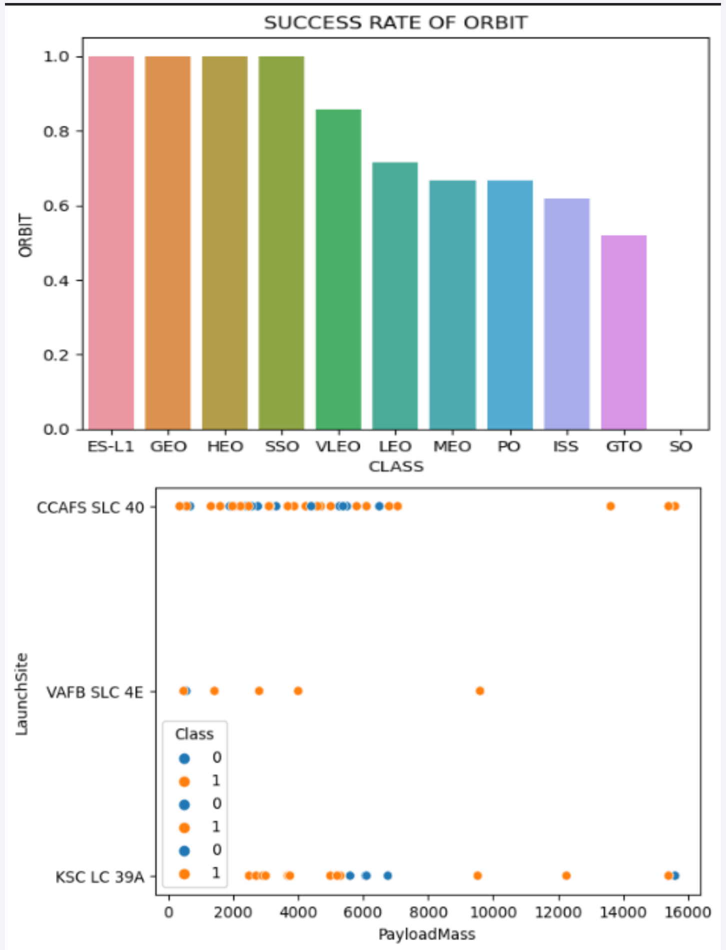
- We performed exploratory data analysis and determined the training labels.
- Calculate the number of launches that took place at each launch site
- Investigate the dataset to identify the different orbits used and the frequency of each orbit in the SpaceX missions.
- Convert the outcomes into a training label where '1' represents a successfully landed booster and '0' represents an unsuccessful landing.
- The link to the notebook:

https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Utilized various data visualization techniques such as scatter plots, bar plots, and other relevant visualization tools to explore and understand the dataset.
- Explored the data by visualizing the relationship between flight no and launch sites, payload and launch sites, success rate of each orbit, flight no and orbit type, yearly trend of launch site
- The link to the notebook is

https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/EDA%20with%20Visualization.ipynb



EDA with SQL

We performed EDA using SQL. The following queries were performed

- names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- total payload mass carried by boosters launched by NASA (CRS)
- average payload mass carried by booster version F9 v1.1
- date when the first successful landing outcome in ground pad was acheived.
- 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
- names of the booster versions which have carried the maximum payload mass. Use a subquery
- failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

The link of notebook is

https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

- Markers, circles, Marker clusters and lines were used in building interactive map with folium
    - Markers were used to indicate points like launch sites
    - We calculated the distances between a launch site to its proximities, indicated by lines
    - Markers cluster used to indicate event in each coordinate, like launch in launch site
- Code: https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/Interative%20visusl%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Built an interactive dashboard using plotly
- Built a pie chart to show the successful launches by each sites
- Plotted a scatter chart which shows relation between outcome and payload mass for different booster version.
- Code: https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/app.py

Predictive Analysis (Classification)

- First we standardized the data
- Then split the data into test and train datasets using scikit learn
- Then we apply logistic regression, support vector machine(SVM) and decision trees.
- Plot confusion matrix to see which model gives the best accuracy
- Also find tuned hyper parameters for each model using GridSearchCV function
- Code: https://github.com/JaveriaMajeed/IBM-Data-Science-Capstone-SpaceY/blob/master/Machine%20Learning%20Prediction%20Lab.ipynb

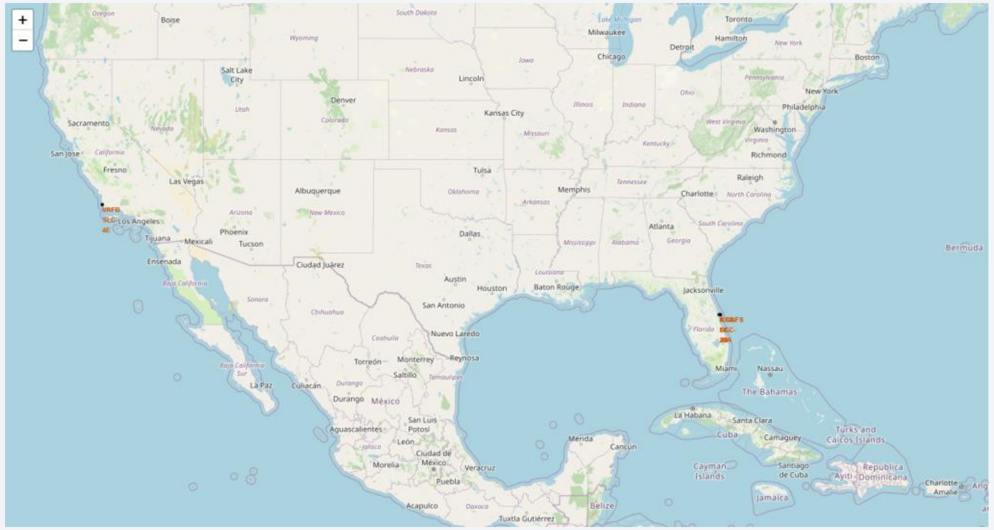
Results

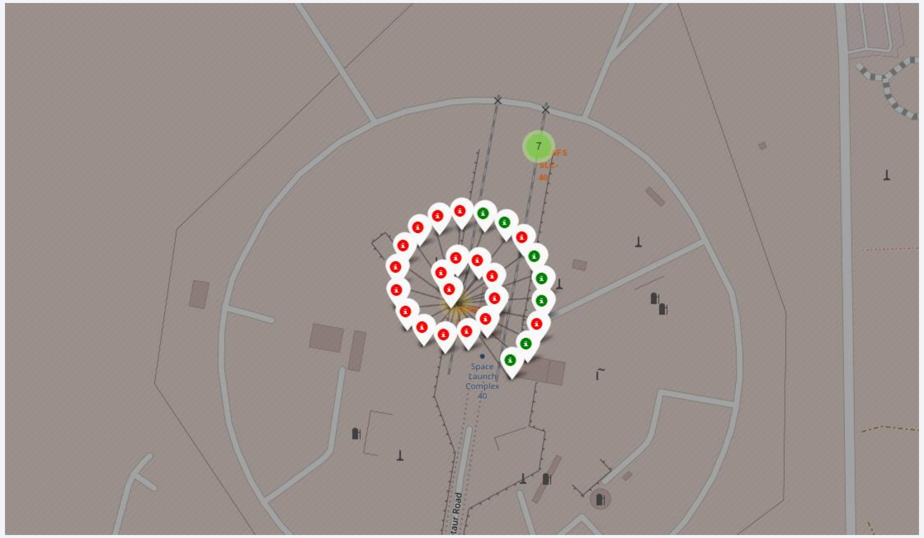
- Exploratory data analysis results

    - SpaceX uses 4 different Launch sites
    - The success rate of landing is almost 100%
    - The maximum payload mass was carried by F9B5
    - The first successful landing in ground pad was achieved in 2015
    - CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
    - The orbit ES-L1, GEO, HEO SSO has high rate of successful landing.
    - The no of successful landing kept increasing since 2013 until 2020.

Results

- Interactive analytics demo in screenshots
- Most of the launch sites are built near the sea.





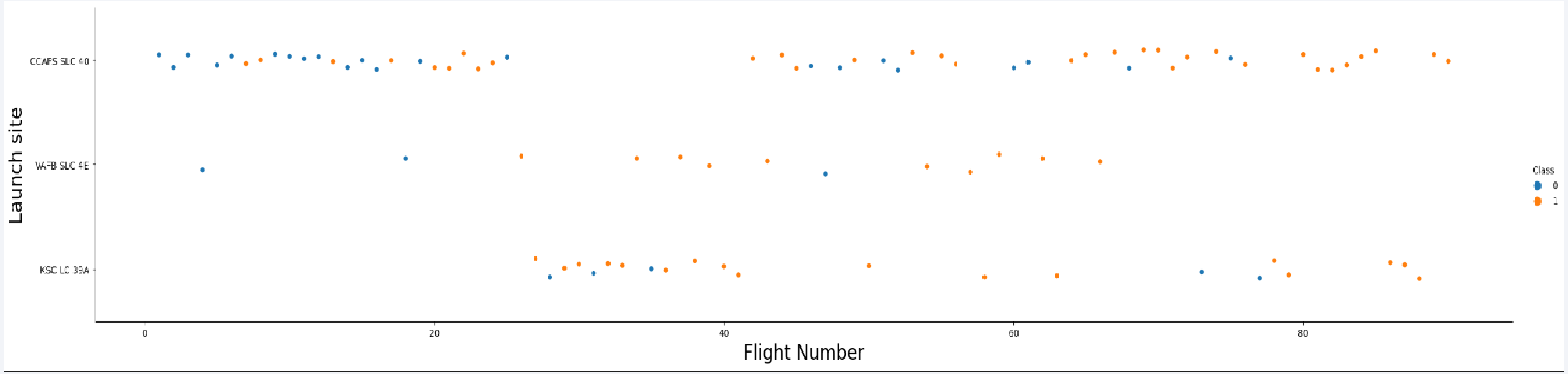
Results

- Predictive analysis results
  - We trained our data using three classifiers Logistic regression, support vector machine and decision trees.
  - Predictive analysis shows that decision tree has a high accuracy in predicting the successful launches

Section 2

Insights drawn from EDA

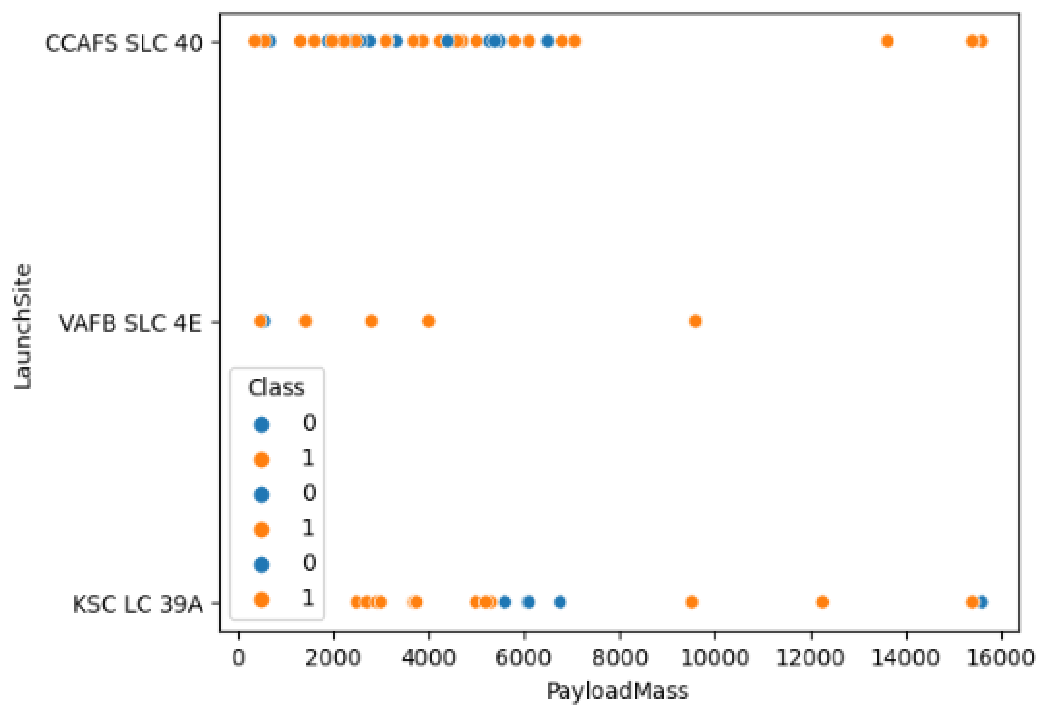
Flight Number vs. Launch Site



- Most of the rocket launches at CCAFS launch site, it has high success rate in recent year.
- Few rockets were launched at VAFB, but it also have good success rate

Payload vs. Launch Site

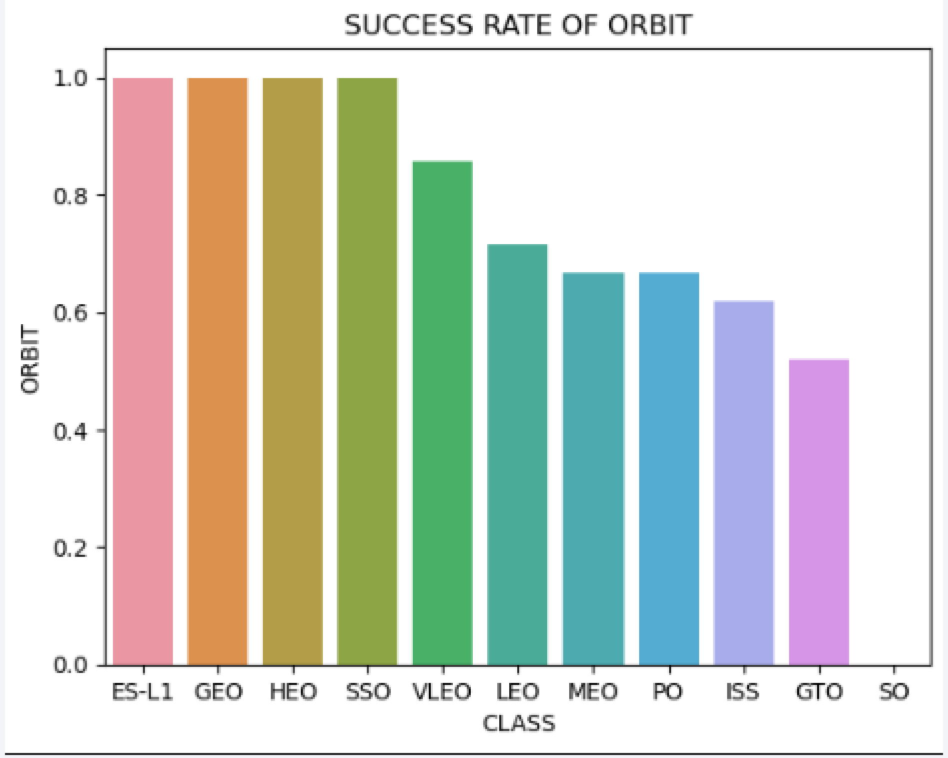
- For the VAFB-SLC launch site there are no rockets launched for heavypayload mass(greater than 10000).



Success Rate vs. Orbit Type

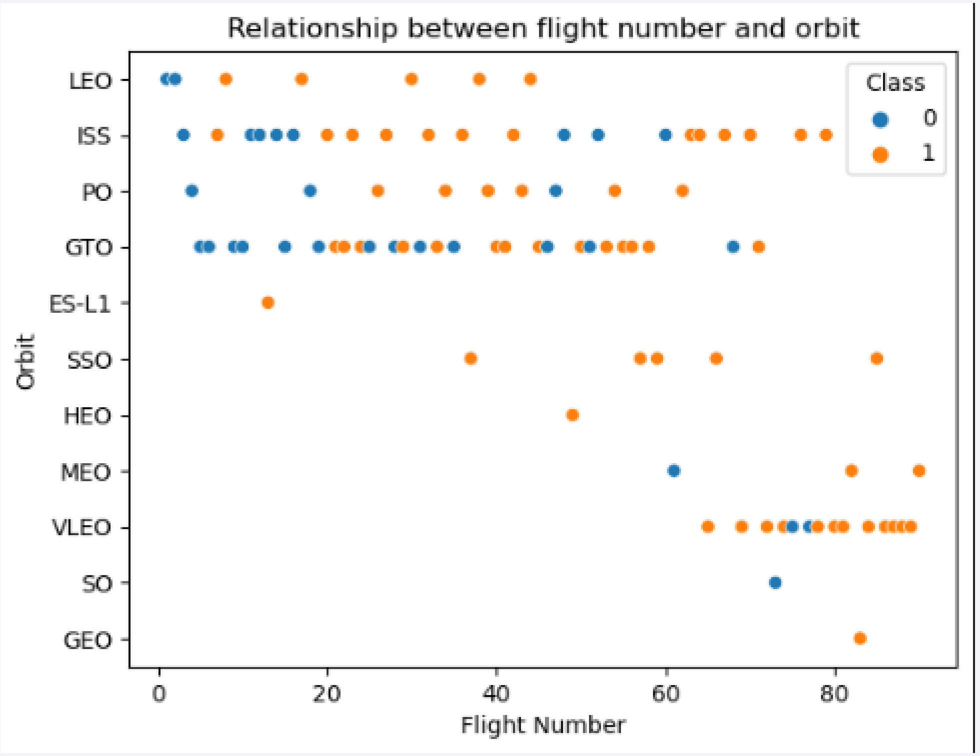
Following orbits have high success rate

- ES-L1
- GEO
- HEO
- SSO

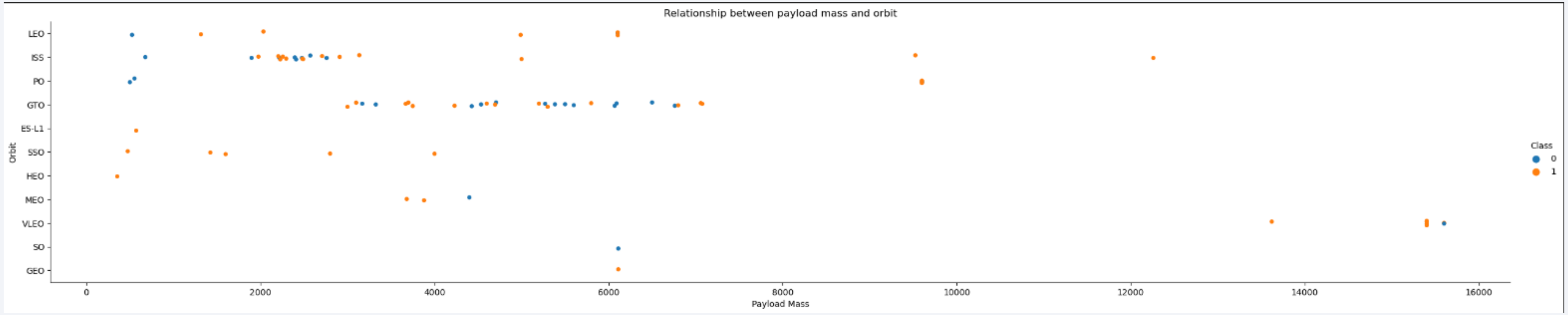


Flight Number vs. Orbit Type

- VLEO has a high success rate in recent launches
- 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.



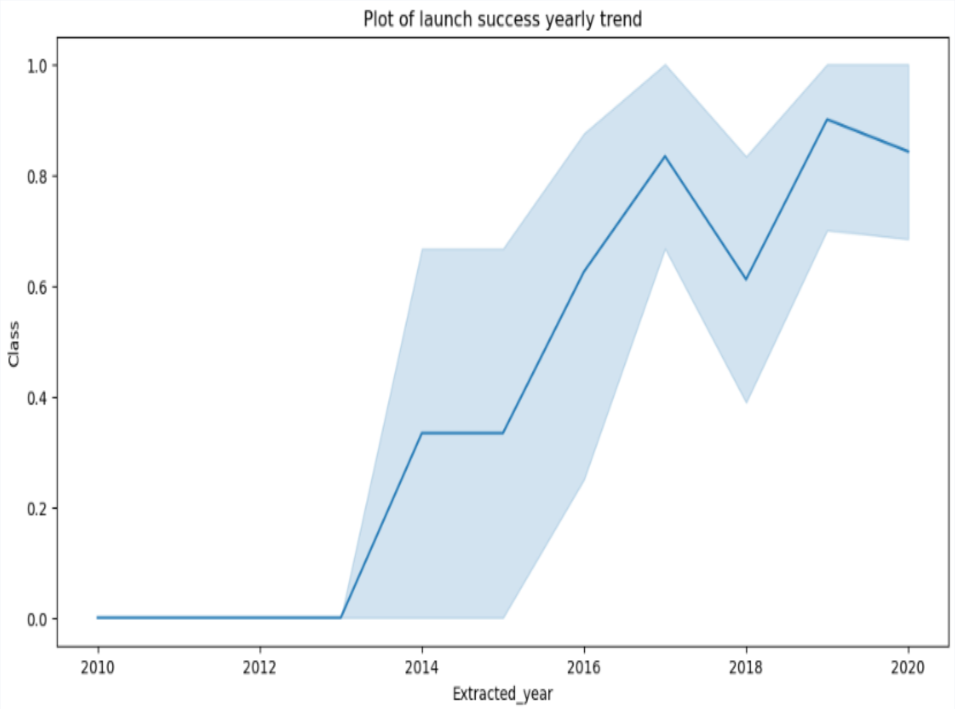
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,VLEO and ISS.
- There are few launches to the orbit SO and GEO

Launch Success Yearly Trend

- From the plot we can see that the no of successful landing kept increasing since 2013 until 2020.



All Launch Site Names

- We used DISTINCT function to select the unique launch sites from the data

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

```
pd.read_sql_query("SELECT DISTINCT Launch_Site FROM SPACEX",cnn)
```

[8]

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

Launch Site Names Begin with 'CCA'

```
pd.read_sql_query("SELECT * FROM SPACEX WHERE Launch_Site LIKE 'CCA%' LIMIT 5;",cnn)
```

	index	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	0	4/6/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	1	8/12/2010	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	2	5/22/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	3	8/10/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	4	1/3/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- In this query we used WHERE clause to specify the condition, in order to only show those rows where launch site starts from CCA
- We also used LIMIT function to display only 5 records

Total Payload Mass

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

```
pd.read_sql("""SELECT SUM(Payload_Mass__KG_) as totalpayloadmass
from spacex WHERE Customer = 'NASA (CRS)' """,cnn)
```

	totalpayloadmass
0	45596

- We used above queru to display the total payload mass carried by nasa
- In WHERE clause we specify the condition to only sum those payload mass where customer is NASA(CRS)

Average Payload Mass by F9 v1.1

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

```
[11]
pd.read_sql("""SELECT AVG(Payload_Mass__KG_) as avgpayloadmass
from spacex WHERE Booster_Version = 'F9 v1.1'""",cnn)
```

	avgpayloadmass
0	2928.4

- We calculated the average payload mass by using AVG function

First Successful Ground Landing Date

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

Hint:Use min function

```
]: pd.read_sql("""SELECT MIN(Date) AS FIRSTSUCCESSFULL_LANDING
FROM SPACEX WHERE Landing_Outcome = 'Success (ground pad)';""", cnn)
```

FIRSTSUCCESSFULL_LANDING
2015-12-22

- It shows that first successful landing happened in 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
pd.read_sql("""SELECT Booster_Version FROM SPACEX
WHERE LANDING_OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000""", cnn)
```

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

- We used BETWEEN function to define the range for payload mass

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

List the total number of successful and failure mission outcomes

```
success = pd.read_sql("""SELECT COUNT(Mission_Outcome) AS SuccessCount
FROM SPACEX WHERE Mission_Outcome like "%Success%"
;""", cnn)
failure = pd.read_sql("""SELECT COUNT(Mission_Outcome) AS FailureCount
FROM SPACEX WHERE Mission_Outcome like "%failure%"
;""", cnn)

print(success)
print(failure)
```

```
   SuccessCount
0           100
   FailureCount
0             1
```

Boosters Carried Maximum Payload

- We used subquery to determine the determine the booster version which carry the max payload mass

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

```
pd.read_sql("""SELECT Booster_Version, Payload_Mass__kg_ from SPACEX
WHERE Payload_Mass__kg_ = (SELECT MAX(Payload_Mass__kg_) FROM SPACEX)""", cnn)
```

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

2015 Launch Records

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

```
pd.read_sql("""SELECT Booster_Version, Launch_Site, Landing_Outcome, Date
            FROM SPACEX WHERE substr(DATE,1,4) = '2015'
            AND substr(DATE,6,2)
            AND Landing_Outcome = 'Failure (drone ship)';""", cnn)
```

Booster_Version	Launch_Site	Landing_Outcome	Date

BOOSTER_VERSION	LAUNCH_SITE	LANDING_OUTCOME	DATE
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	2015-10-01
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	2015-04-14

- We used functions like WHERE, AND and SUBSTR to filter for the condition given

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

```
pd.read_sql('''SELECT Landing_Outcome, COUNT(Landing_Outcome) AS TOTALCOUNT FROM SPACEX
 WHERE DATE BETWEEN '2010-06-04'AND '2017-03-20'
 GROUP BY Landing_Outcome
ORDER BY TOTALCOUNT DESC;''', cnn)
```

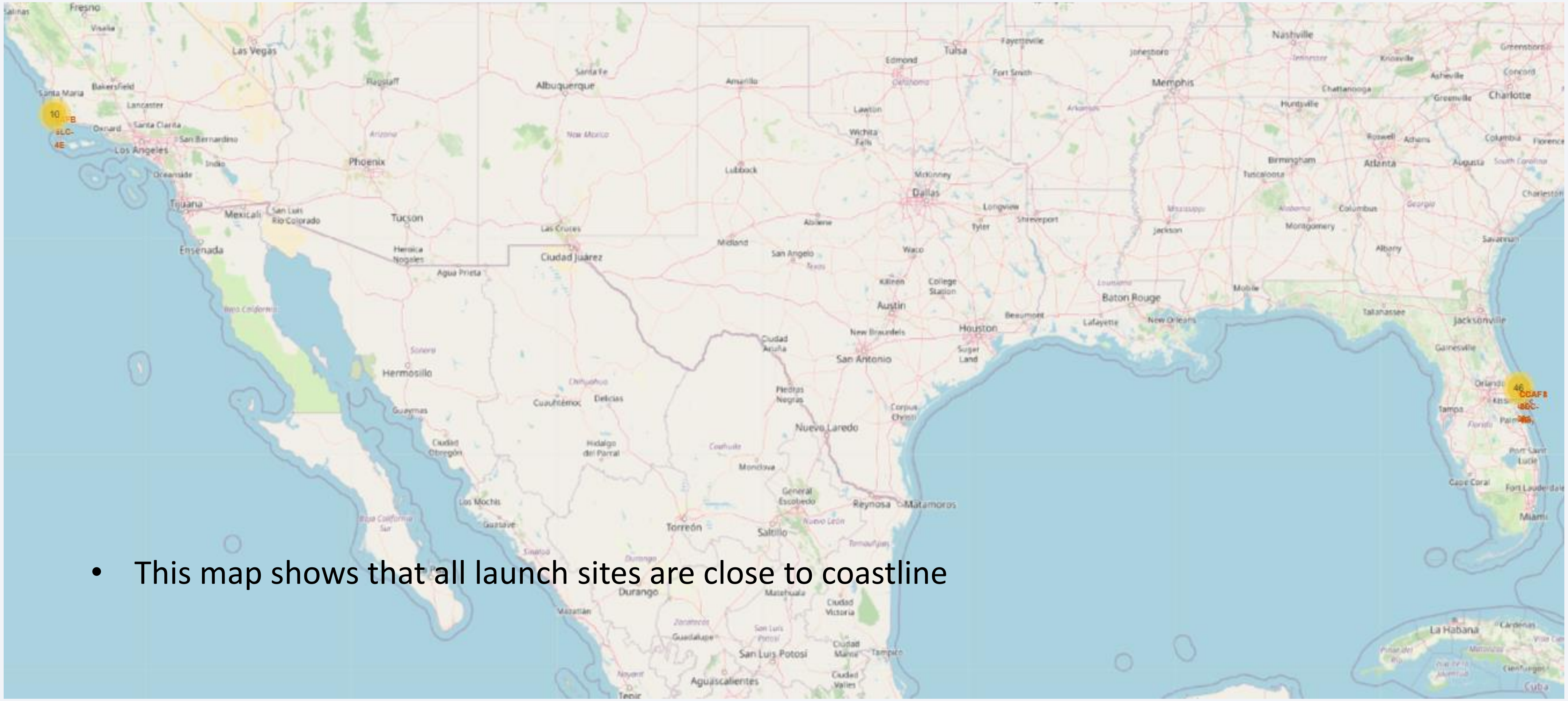
Landing_Outcome	TOTALCOUNT

LANDING_OUTCOME	TOTALCOUNT
No attempt	20
Failure (drone ship)	10
Success (drone ship)	10
Success (ground pad)	10
Controlled (ocean)	6
Uncontrolled (ocean)	4
Failure (parachute)	2
Precluded (drone ship)	2

Section 3

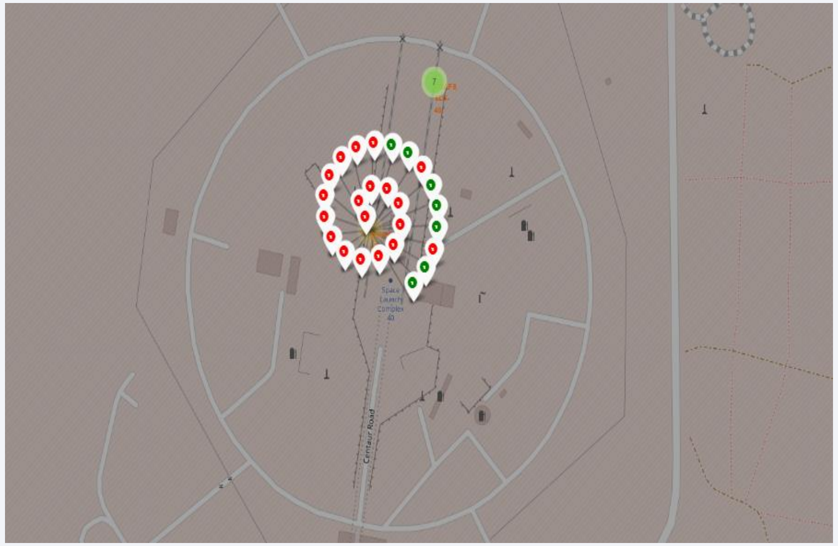
Launch Sites Proximities Analysis

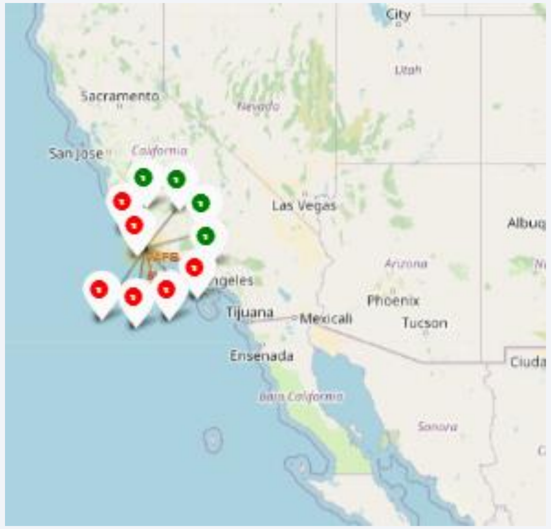
ALL LAUNCH SITES

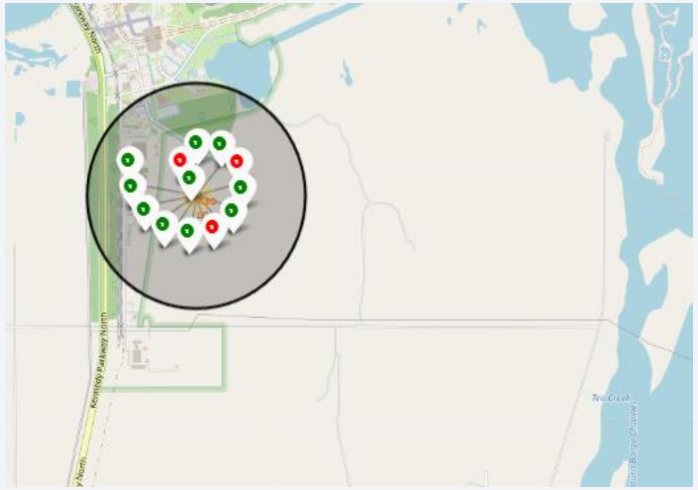


- This map shows that all launch sites are close to coastline

LAUNCH SITES WITH COLOR LABELS

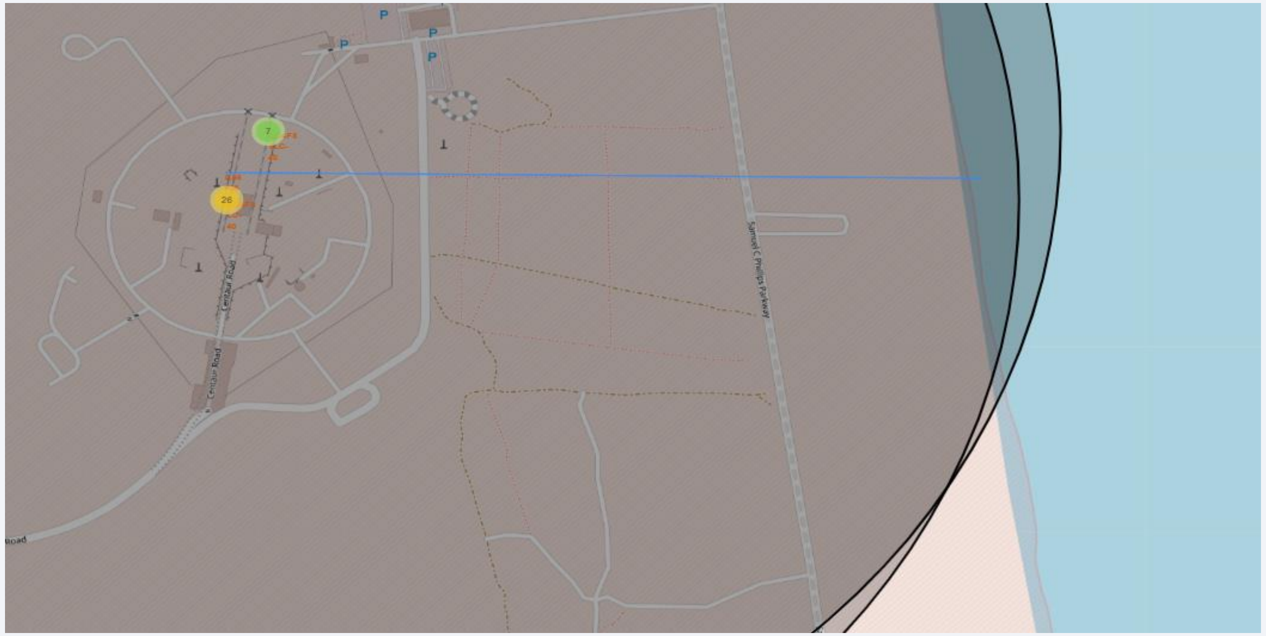






Green markers shows successful landing and red marker shows failure

LAUNCH SITE DISTANCE TO COASTLINE

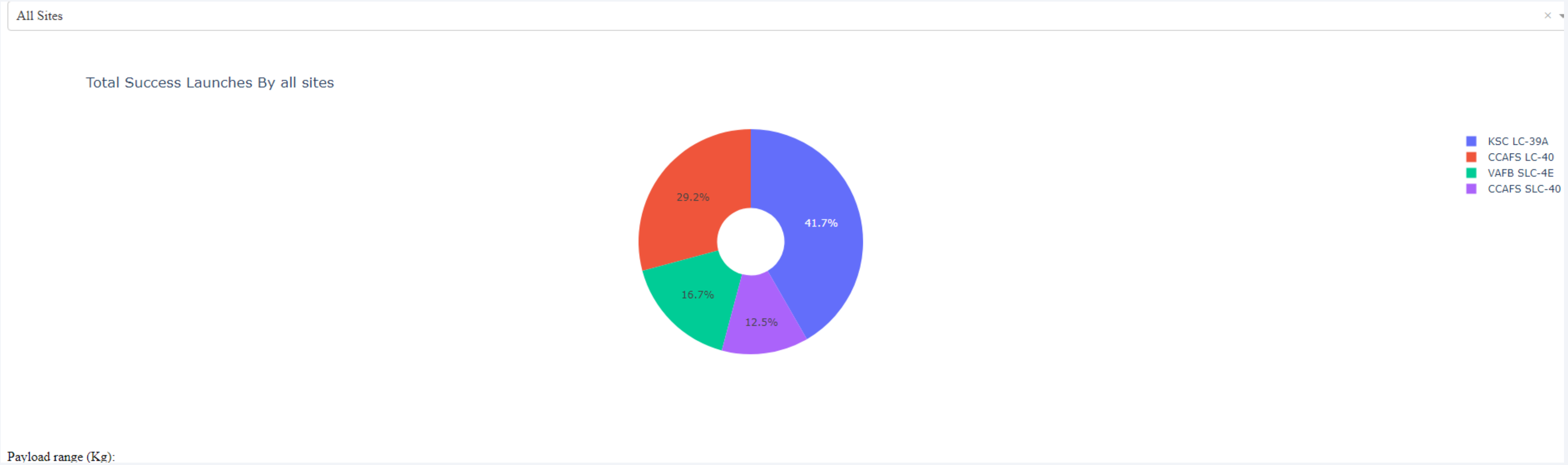


- The launch site distance from coastline is 0.953km.

Section 4

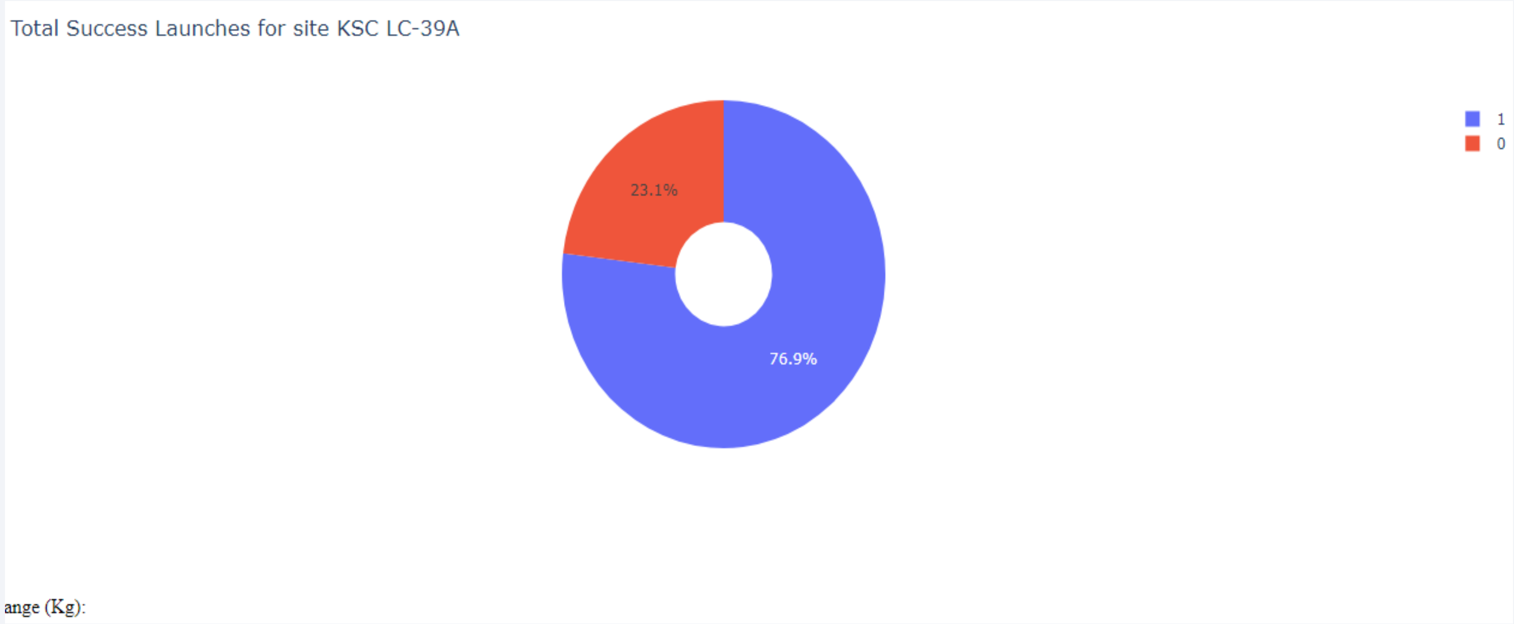
Build a Dashboard with Plotly Dash

SUCCESSFUL LAUNCHES BY ALL SITES



KSC LC-39A has most successful launches

HIGHEST SUCCESS RATE OF A LAUNCH SITE

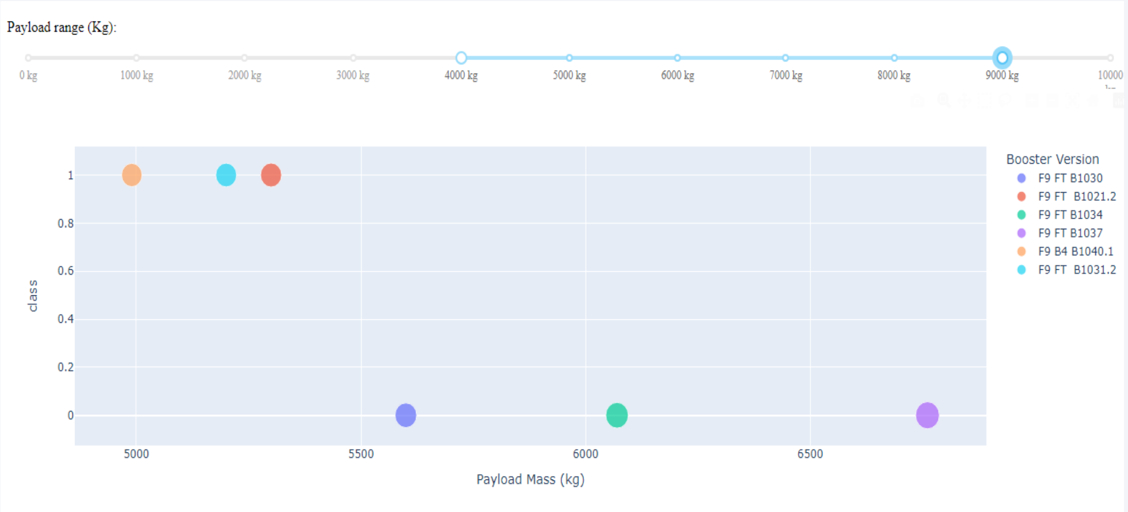


- **The success rate of KSC LC-37A IS 76.3%**

<Dashboard Screenshot 3>

Payload mass 0 – 400kg

Payload mass 400 – 900kg

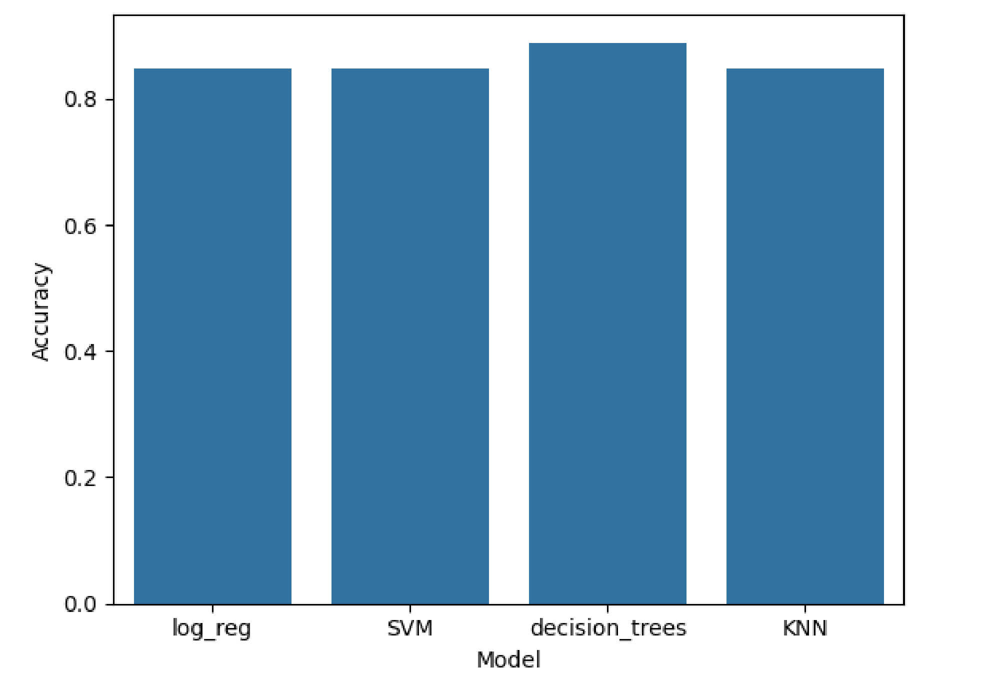


Section 5

Predictive Analysis (Classification)

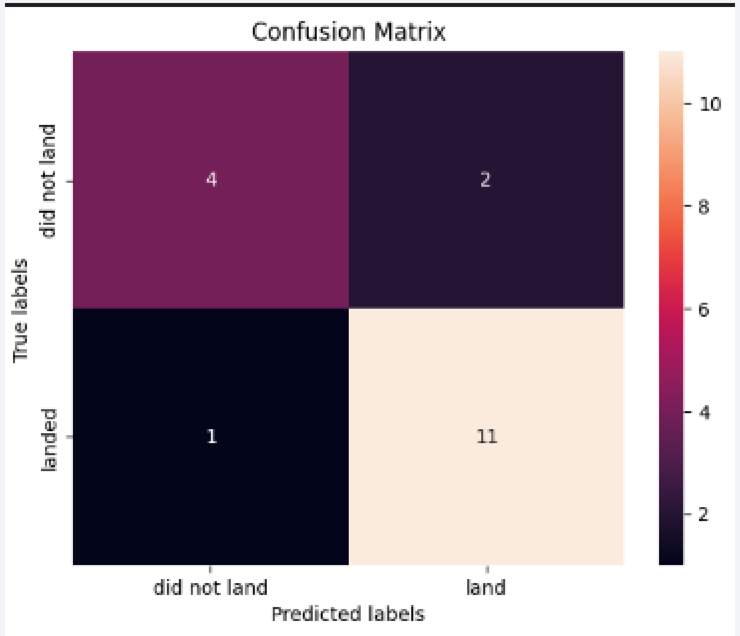
Classification Accuracy

- According to the bar plot, decision has the best accuracy



Confusion Matrix

- The confusion matrix shows that it successfully distinguish between the different classes. Out of 6 did not land it guess 4 correct and out of 12 land it guess 11 correct



Conclusions

- The first successful landing happened in 2015.
- Success rate has increased since 2013 till 2020.
- Launch site KSC-LC 39A has the most successful launches.
- Decision tree classifier has the best accuracy for this task.
- The orbit ES-L1, GEO, HEO SSO has high rate of successful landing.

Thank you!