IBM DATA SCIENCE CAPSTONE PROJECT – SPACE X

STEFANOVIC – 31/10/21

OUTLINE

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

EXECUTIVE SUMMARY

SUMMARY OF METHODOLOGIES

- Data collection
- Data wrangling
- EDA with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis (Classification)

SUMMARY OF ALL RESULTS

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

INTRODUCTION

PROJECT BACKGROUND

- Space Exploration Technologies Corp. is an American aerospace manufacturer, space transportation services and communications corporation headquartered in Hawthorne, California. SpaceX was founded in 2002 by Elon Musk with the goal of reducing space transportation costs to enable the colonization of Mars.

RESEARCH QUESTIONS

- What influences a successful rocket landing?
- Which factors have an impact on the success rate of the landing?
- What conditions Space X need to accomplish in order to get the best success rate?

METHODOLOGY

1. Data collection by Space X Rest API and Web Scrapping from Wikipedia.
2. Data Wrangling (Transforming the data to be able to run further analysis)
3. Execute exploratory data analysis (EDA) by using several visualization tools and SQL.
4. Performed an interactive visual analytics using Folium and Plotly Dash
5. Performed predictive analytics using classification models

DATA COLLECTION - SPACE X API

1. Getting Response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

2. Converting Reponse to a json file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

3. Apply custom functions to clean data

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
```

4. Assign list to dictionary then data frame

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

5. Filter data frame and export to csv

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

DATA COLLECTION – WEB SCRAPPING

1. Getting Response from HTML

```
page = requests.get(static_url)
```

2. Creating Beautiful Soup Object

```
soup = BeautifulSoup(page.text, 'html.parser')
```

3. Finding tables

```
html_tables = soup.find_all('table')
```

5. Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']


launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Converting to dataframe

```
df = pd.DataFrame.from dict(launch dict)
```

7. Data frame to csv

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

DATA WRANGLING

The data set consists out of several landing data collected for each rocket start.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	Landing
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN

In order to run further analysis it was necessary to proceed first with a data wrangling process

EDA → Data Wrangling → Export dataset as csv

EDA WITH DATA VISUALIZATION

For the EDA we run the following visualizations:

1. Relationship between Flight Number and Launch Site
2. The relationship between Payload and Launch Site
3. Relationship between success rate of each orbit type
4. Relationship between Flight Number and Orbit type
5. Relationship between Payload and Orbit type
6. Launch Success Yearly Trend

EDA WITH SQL

For the EDA with SQL we run the following queries:

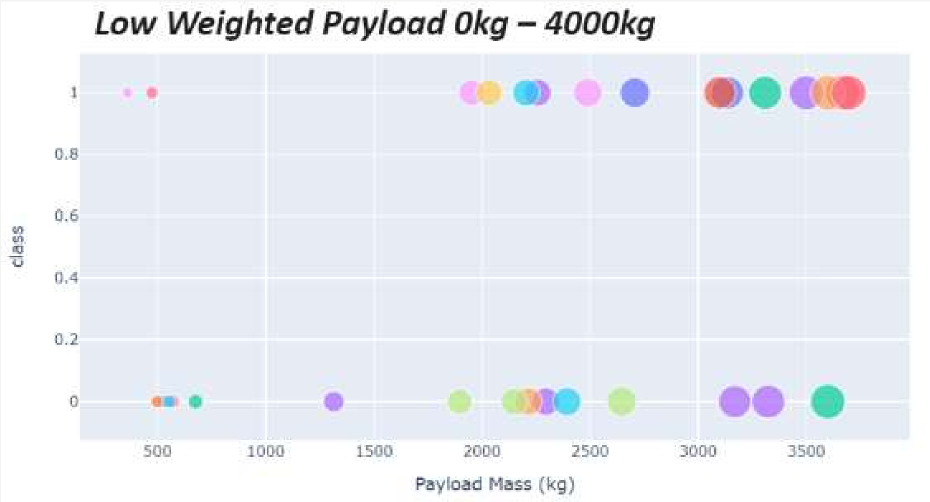
1. Display the total payload mass carried by boosters launched by NASA (CRS)
2. Display average payload mass carried by booster version F9
3. List the date where the successful landing outcome in drone ship was achieved.
4. List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
5. List the total number of successful and failure mission outcomes
6. List the names of the booster versions which have carried the maximum payload mass. Use a subquery
7. List the records which will display the month names, successful landing outcomes in ground pad ,booster versions, launch site for the months in year 2017
8. Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

BUILDING AN INTERACTIVE MAP WITH FOLIUM

- To visualize the launch data I used an interactive map.
- I assigned the launch outcomes to the classes 0 and 1 with green and red markets and display them on the map.
- I also applied Haversine's formula to calculate the distance from the lunch site to landmarks in order to identify trends.

BUILDING AN INTERACTIVE DASHBOARD

- Next I developed an interactive dashboard with Flask and Dash.
- In the dashboard it is possible for the user to adjust the visualization in order to analyze the data.
- Graphs used:
  - Pie Chart
  - Scatter Graph



PREDICTIVE ANALYSIS

1. Building a Model
   - Load dataset with pandas
   - Transform data
   - Split data in training and test sets
   - Decide MLA
   - Set parameters
   - Fit the model

Evaluating the Model
- Check accuracy of the model
- Plot confusion matrix

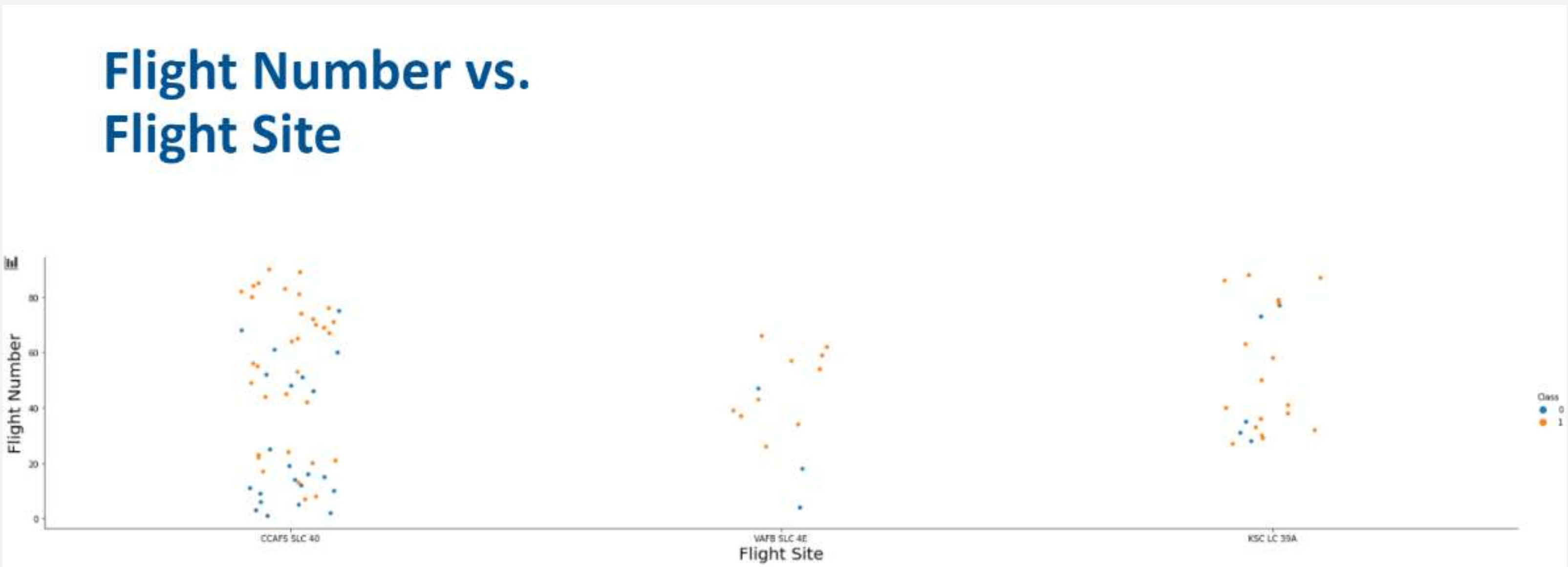
Improving the Model
- Feature Engineering
- Algorithm Tuning

Finding the best performance model
- The model with the best accuracy score wins

RESULTS

- Exploratory data analysis results
- Interactive analytics results
- Predictive analysis results

EDA WITH VISUALIZATIONS



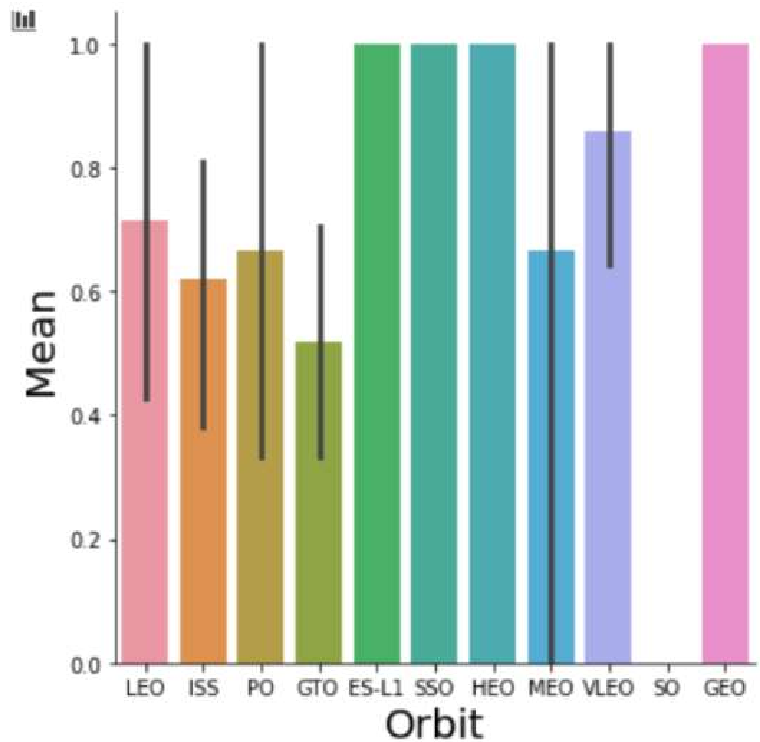
The more amount of flights at launch site the greater the success rate at a lunch site.

EDA WITH VISUALIZATIONS

The greater the payload mass for Lunch Site the higher the success rate for the Rocket.

EDA WITH VISUALIZATIONS

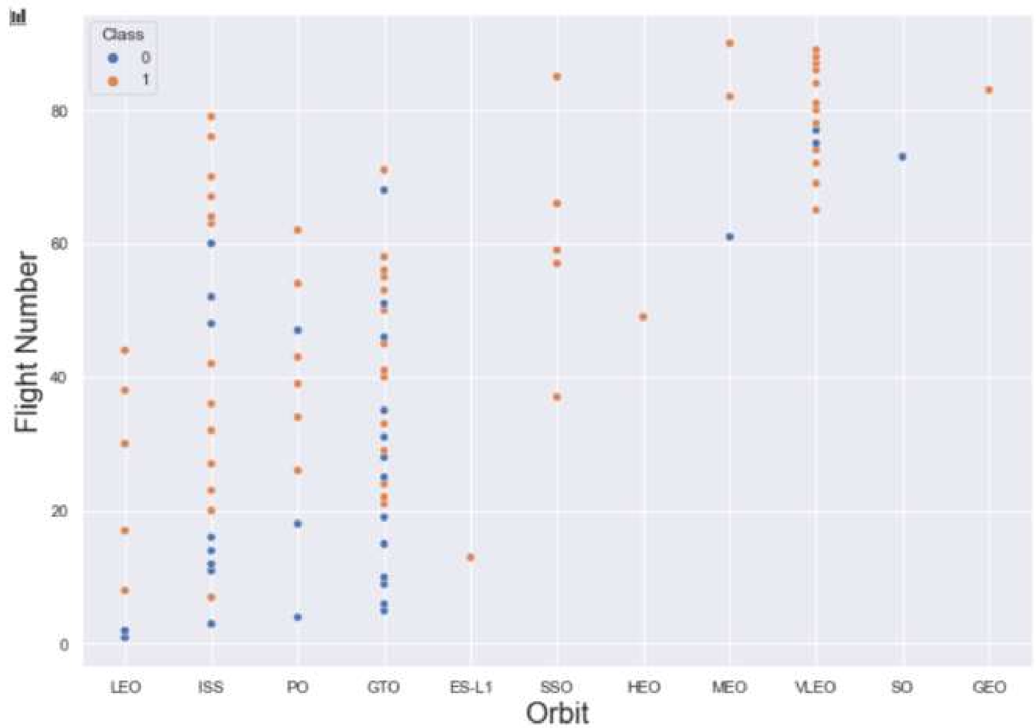
Success rate vs. Orbit type



ORBIT GEO,HEO,SSO,ES-L1 has the best success rate!

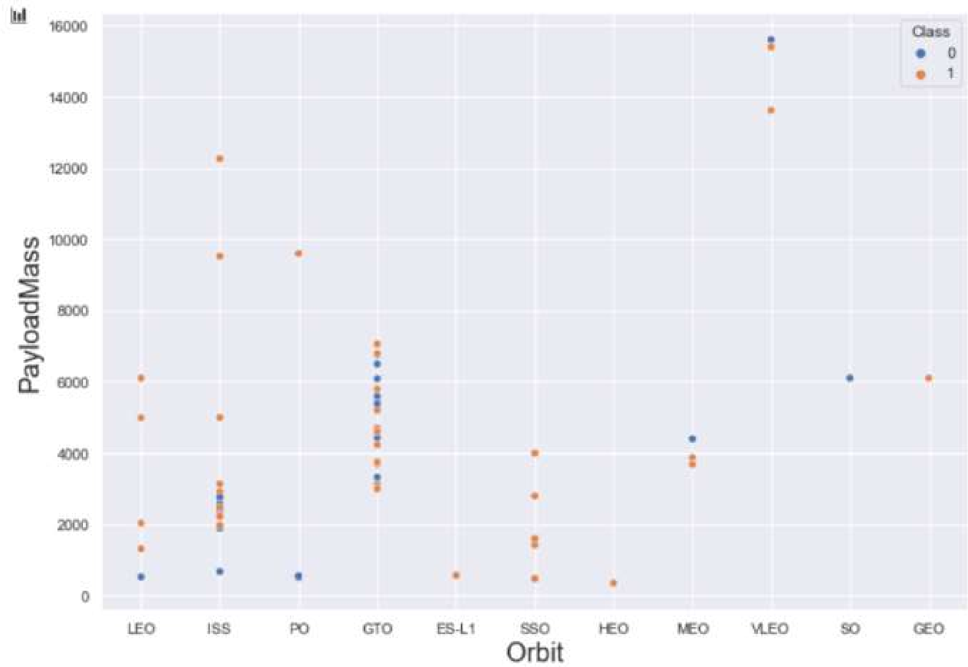
EDA WITH VISUALIZATIONS

Flight Number vs. Orbit type



EDA WITH VISUALIZATIONS

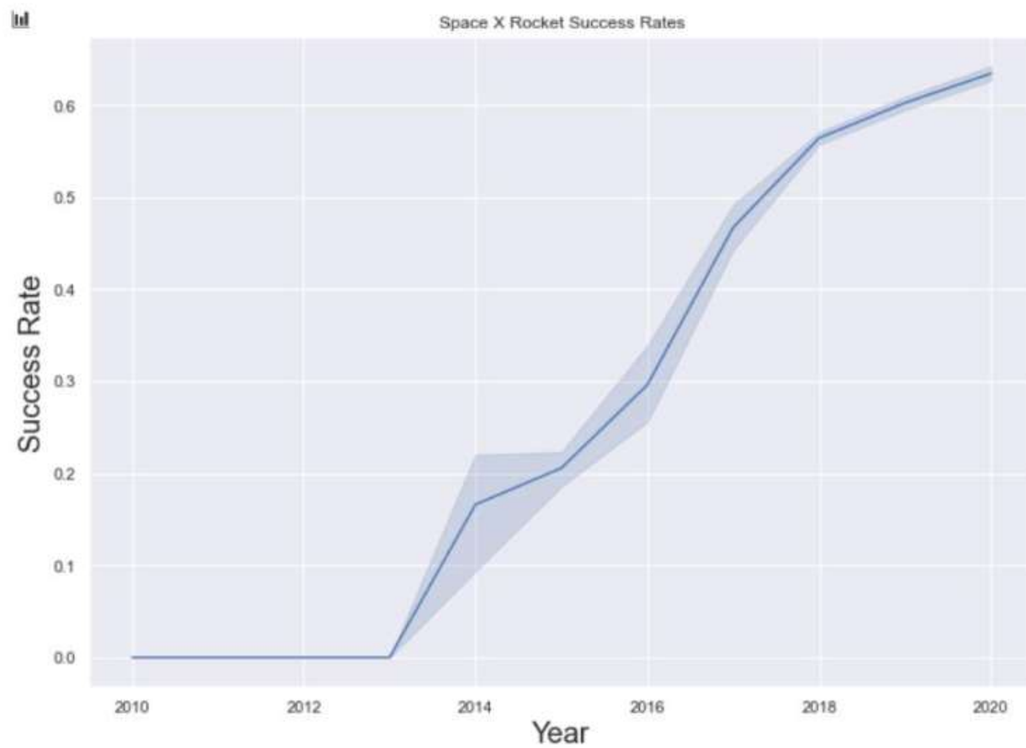
Payload vs. Orbit type



LEO orbit success rate is related to the number of flights.

EDA WITH VISUALIZATIONS

Launch success yearly trend



Lunch success rate increase significant since 2013.

EDA WITH SQL

SQL QUERY

```
select DISTINCT Launch_Site
from tblSpaceX
```

Unique Launch Sites
CCAFS LC-40
CCAFS SLC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

EDA WITH SQL

SQL QUERY

select TOP 5 * from tblSpaceX
WHERE Launch_Site LIKE 'KSC%'

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.0000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.0000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.0000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.0000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.0000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

EDA WITH SQL

SQL QUERY

select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass from tblSpaceX
where Customer = 'NASA (CRS)'",'TotalPayloadMass

	Total Payload Mass
0	45596

EDA WITH SQL

SQL QUERY

select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX
where Booster_Version = 'F9 v1.1'

	Average Payload Mass
0	2928

EDA WITH SQL

SQL QUERY

select MIN(Date) SLO from tblSpaceX where Landing_Outcome = "Success (drone ship)"

	Date which first Successful landing outcome in drone ship was acheived.
0	06-05-2016

EDA WITH SQL

SQL QUERY

select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (ground pad)'
AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

EDA WITH SQL

SQL QUERY

```
SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
LIKE '%Success%') as Successful_Mission_Outcomes,
(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome
LIKE '%Failure%') as Failure_Mission_Coutcomes
```

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

EDA WITH SQL

SQL QUERY

```
SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS
_KG_) AS [Maximum Payload Mass]
FROM tblSpaceX GROUP BY Booster_Version
ORDER BY [Maximum Payload Mass] DESC
```

	Booster_Version	Maximum Payload Mass
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
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows × 2 columns

EDA WITH SQL

SQL QUERY

```
SELECT DATENAME(month, DATEADD(month,
MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month,
Booster_Version, Launch_Site, Landing_Outcome
FROM     tblSpaceX
WHERE  (Landing_Outcome LIKE N'%Success%') AND
(YEAR(CONVERT(date, Date, 105)) = '2017')
```

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

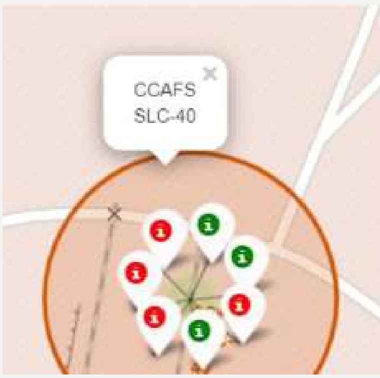
EDA WITH SQL

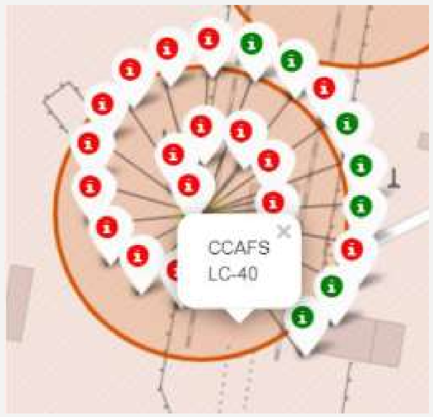
SQL QUERY

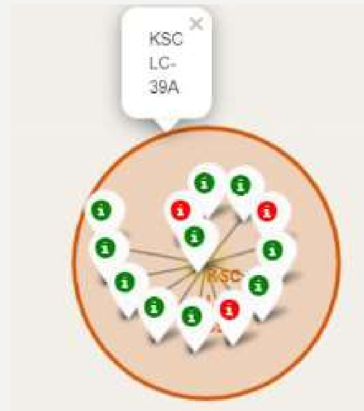
```
SELECT COUNT(Landing_Outcome)
FROM     tblSpaceX
WHERE  (Landing_Outcome LIKE '%Success%')
AND (Date > '04-06-2010')
AND (Date < '20-03-2017')
```

	Successful Landing Outcomes Between 2010-06-04 and 2017-03-20
0	34

INTERACTIVE MAP WITH FOLIUM



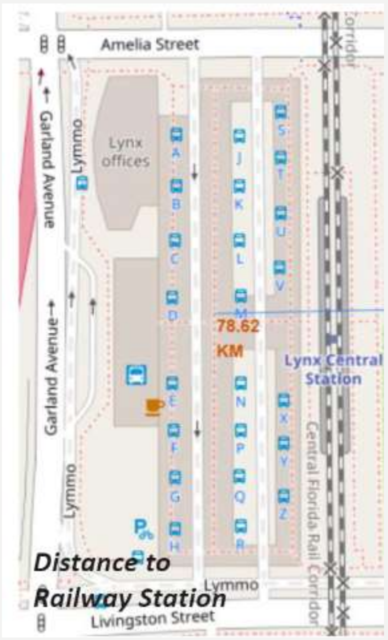




Green Markers: Show the successful Launches

Red Markers: Show the unsuccessful Launches

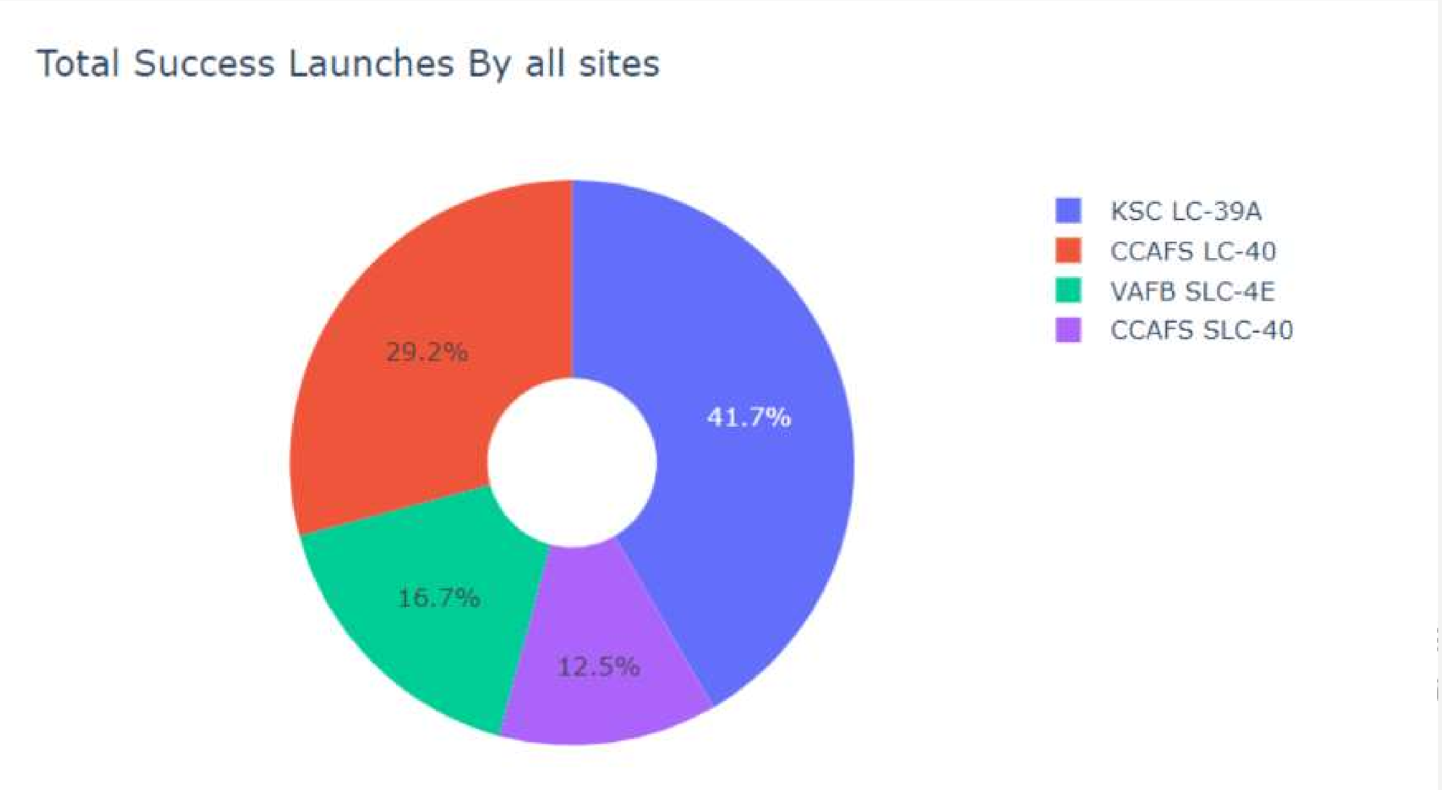
INTERACTIVE MAP WITH FOLIUM



Questions and Answers:

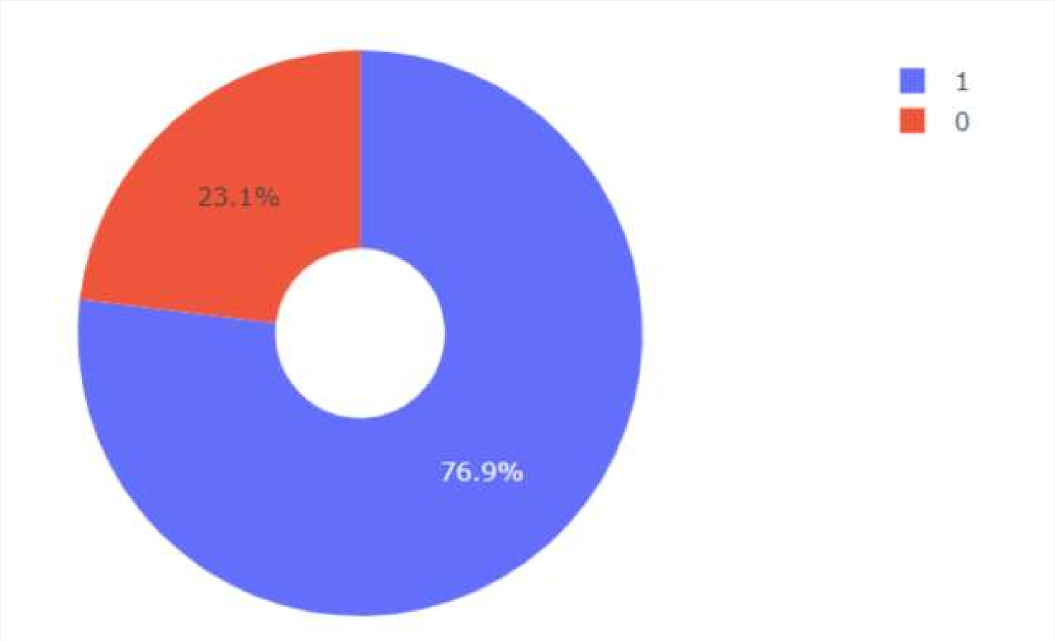
1. Are launch sites in close proximity to railways? No
2. Are launch sites in close proximity to highways? No
3. Are launch sites in close proximity to coastline? Yes
4. Do launch sites keep certain distance away from cities? Yes

DASHBOARD WITH PLOTLY DASH



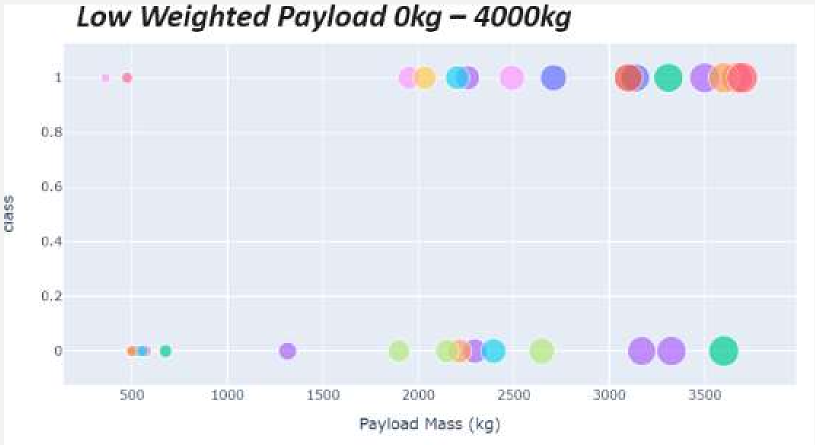
KSC LC-39A is the most successful launch site.

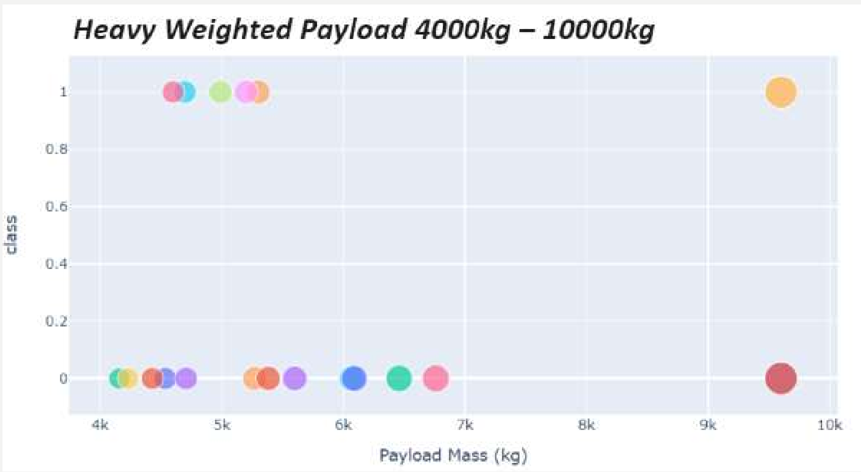
DASHBOARD WITH PLOTLY DASH



KSC LC-39A has a success rate of 76.9%.

DASHBOARD WITH PLOTLY DASH





The success rate for low weighted payloads is higher than for heavy weighted payloads.

PREDICTIVE ANALYSIS (CLASSIFICATION)

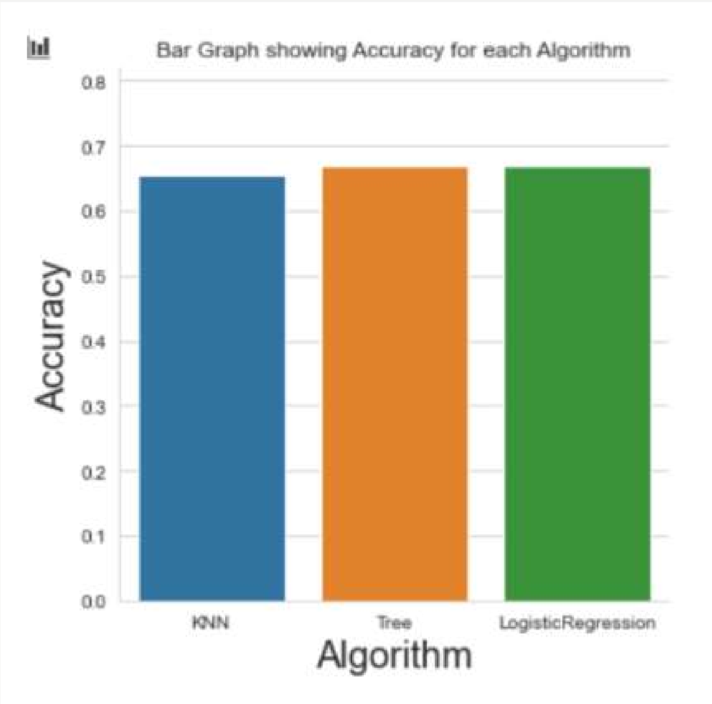
To calculate the best algorithm we used the following function:

```
bestalgorithm = max(algorithms, key=algorithms.get)
```

We received the following result

	Accuracy	Algorithm
0	0.653571	KNN
1	0.667857	Tree
2	0.667857	LogisticRegression

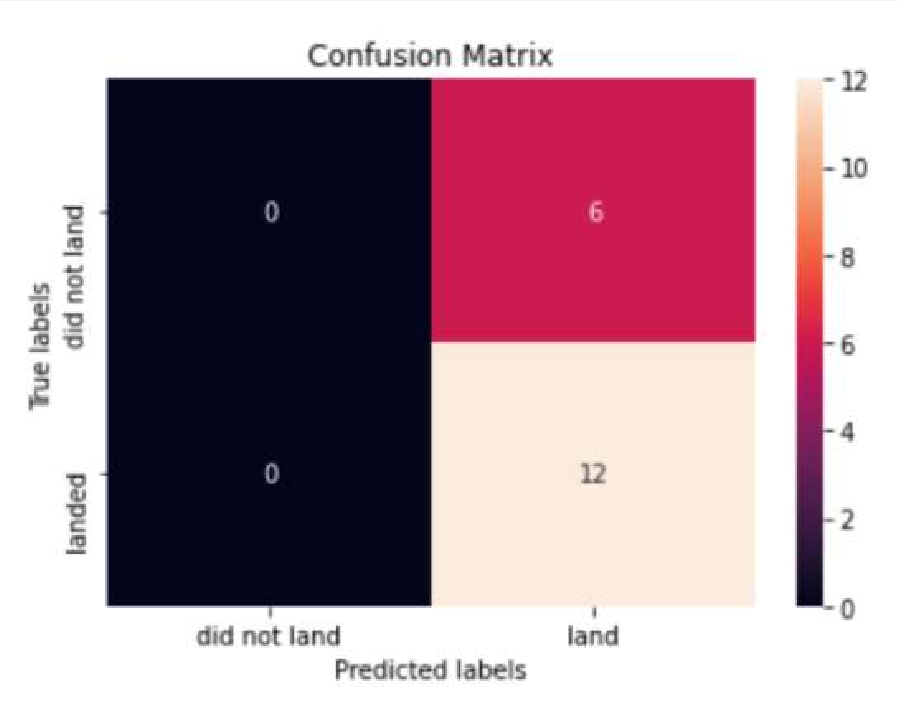
The tree algorithm has the highest accuracy.



PREDICTIVE ANALYSIS (CLASSIFICATION)

The Confusion matrix for the tree algorithm is visualized on the right side.

The major problem is false positives.



CONCLUSION

- The tree classifier algorithm is the best for ML for the data set
- Low weighted payloads perform better than heavy ones.
- KSC LC-39A has the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best success rates
- Based on the data it can be concluded that SpaceX will be able in a few years to significant increase their success rate of rocket launches.

THANK YOU!