IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Parveen Kumar Tiwari
22 / 08 /2022

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 by Classification
- Summary of all results
  - EDA results
  - Interactive analytics
  - Predictive analysis

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars, other provides cost upwards of 165 million dollars, much of the savings is because SpaceX can reuse the first stage.

- Problems you want to find answers
  - The project task is to predicting if the first stage of the SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

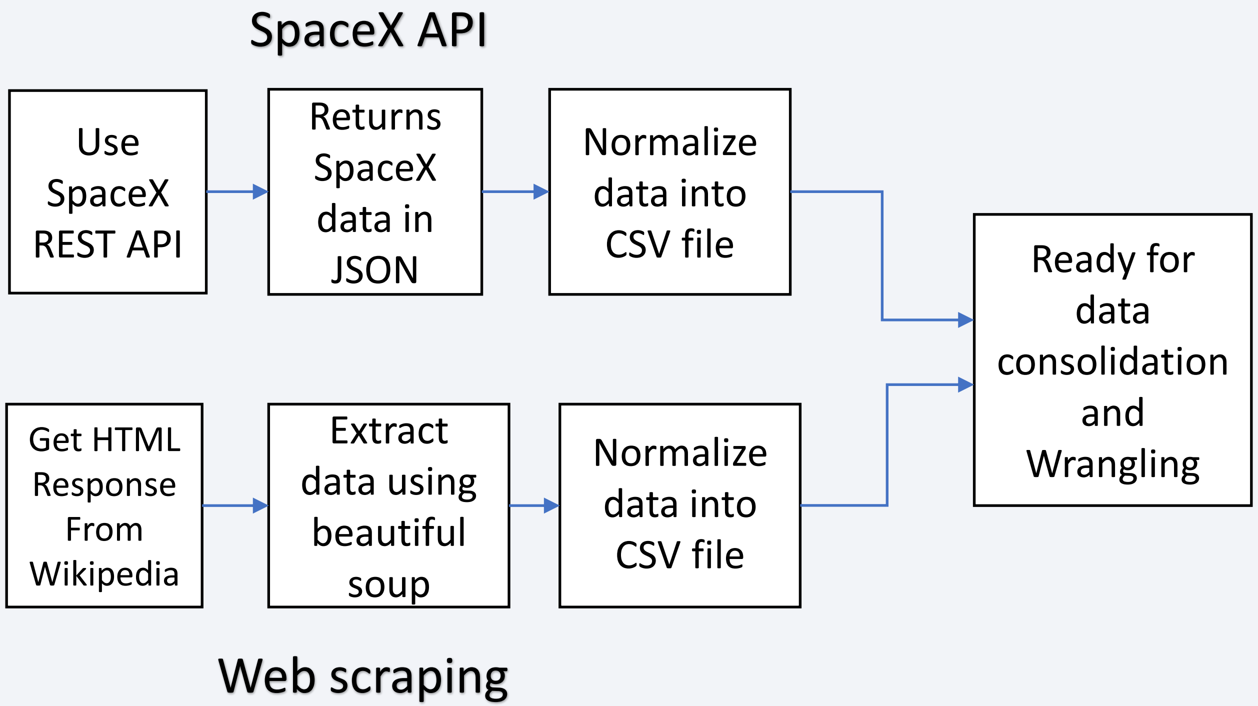
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Null value replacing with mean value, One Hot Encoding data fields for Machine Learning and remove irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - LR, KNN, SVM, DT models have been built and evaluated for best classifier

Data Collection

- The following datasets was collected:
  - SpaceX launch data that is gathered from the SpaceX REST API.
  - This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - The SpaceX REST API endpoints or URL, starts with api.spacexdata.com/v4/.
  - Another popular data source for obtaining Falcon 9 Launch data is web scraping Wikipedia using BeautifulSoup.



Data Collection – SpaceX API

- Data collection with SpaceX REST calls

Github repo : https://github.com/parvntiwari/SpaceX-project-/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

1. Getting response from API

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2. JSON file Normalize into dataframe

```
# Use json_normalize meethod to convert the json result into a dataframe
data=pd.json_normalize(response.json())
```

3.Apply custom function to clean

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

4.Assign list to dictionary then dataframe

```
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 }
```

```
# Create a data from launch_dict
data=pd.DataFrame(launch_dict)
```

5.Filter dataframe and export to flat file(.csv)

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

```
data_falcon9.to_csv('dataset_part\_1.csv', index=False)
```

Data Collection - Scraping

1. Getting response from HTML

```
response= requests.get(static_url).text
```

2. Creating BeautifulSoup Object

```
soup= BeautifulSoup(response, 'html.parser')
```

3. Finding tables

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

4. Gets columns names

```
column_names = []

for row in first_launch_table.find_all('th'):
    name = extract_column_from_header(row)
    if name != None and len(name) > 0:
        column_names.append(name)
```

5. Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

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

# Let's initial the launch_dict with each value to be an empty list
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'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6. Appending data to keys

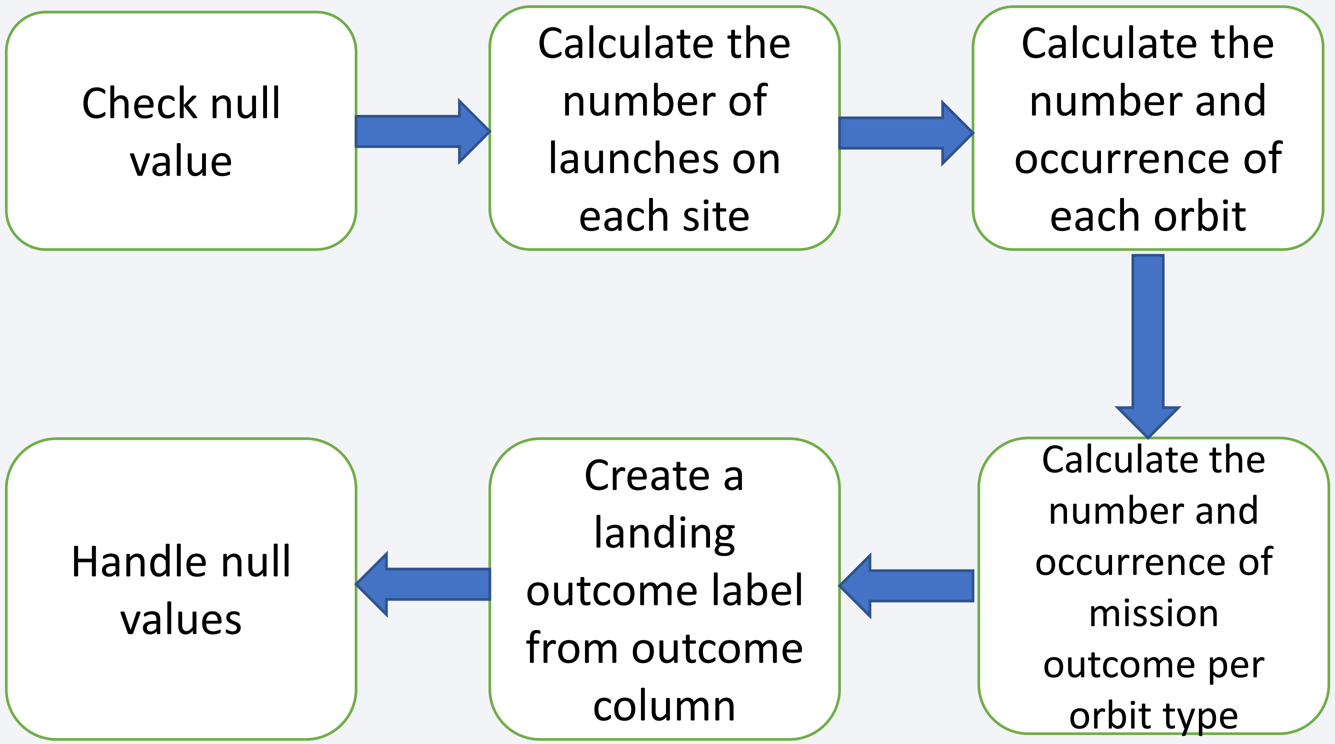
```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is as number corresponding to launch a number
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
```

7. Dataframe to .CSV

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

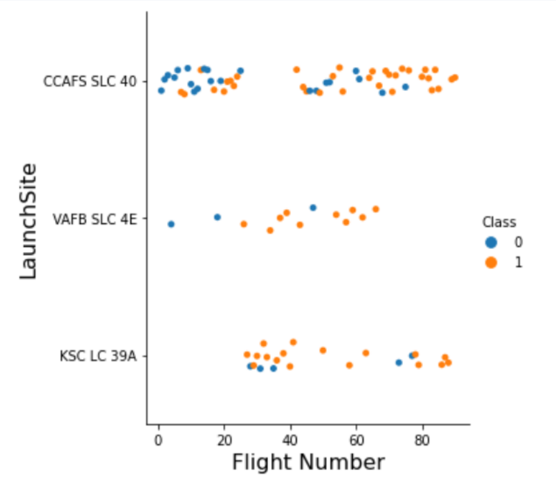
Github repo : https://github.com/parvntiwari/SpaceX-project-/blob/main/jupyter-labs-webscraping%20(1).ipynb

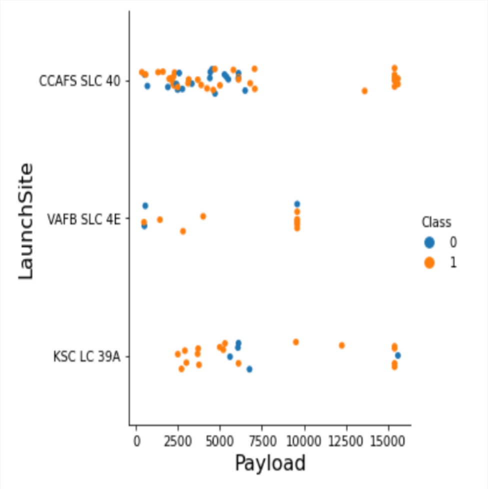
Data Wrangling

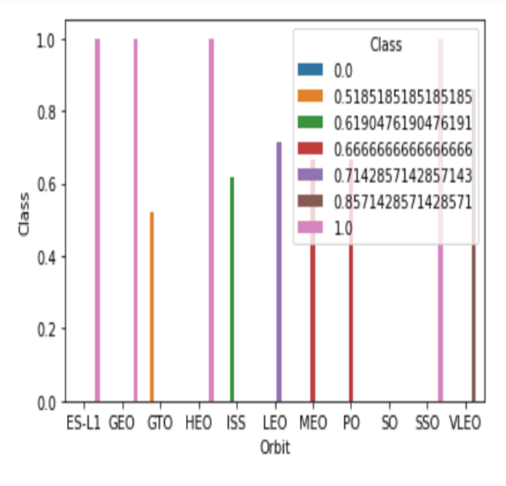


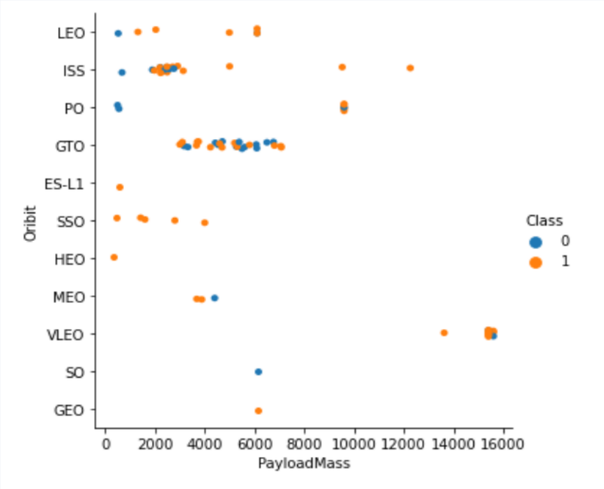
Github repo: https://github.com/parvntiwari/SpaceX-project-/blob/main/labs-jupyter-spacex-Data%20wrangling%20(1).ipynb

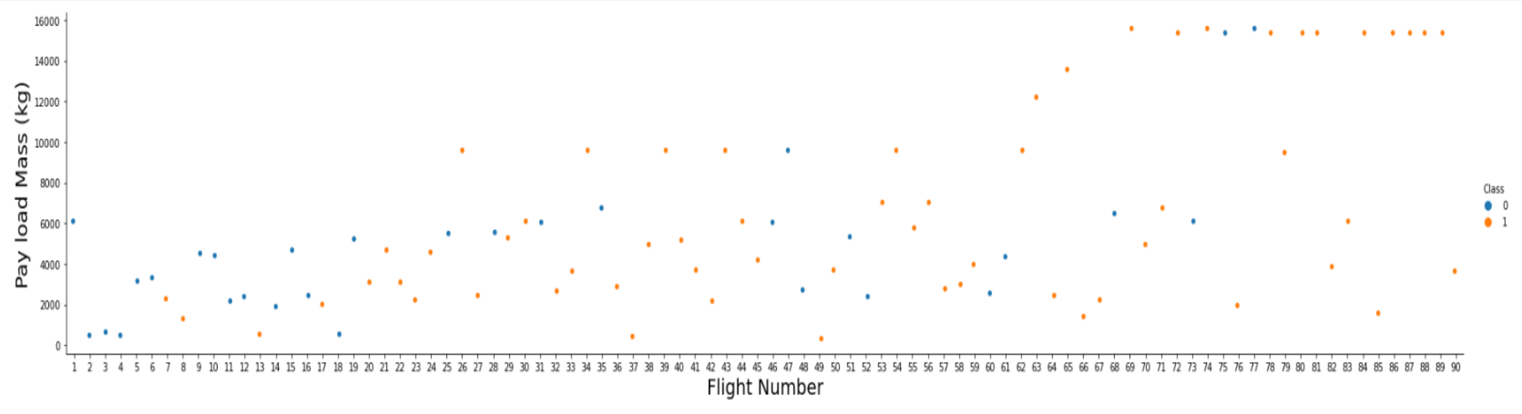
EDA with Data Visualization

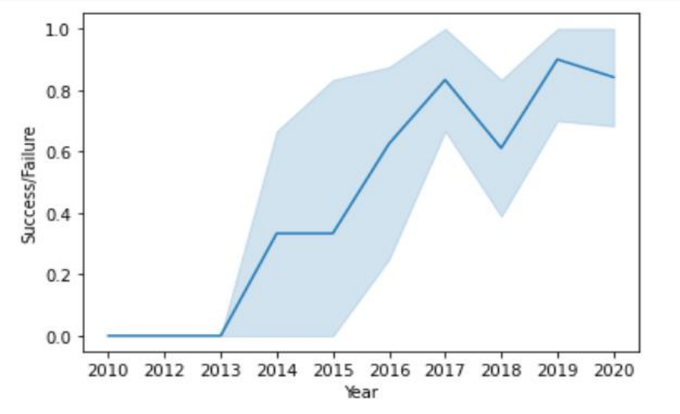












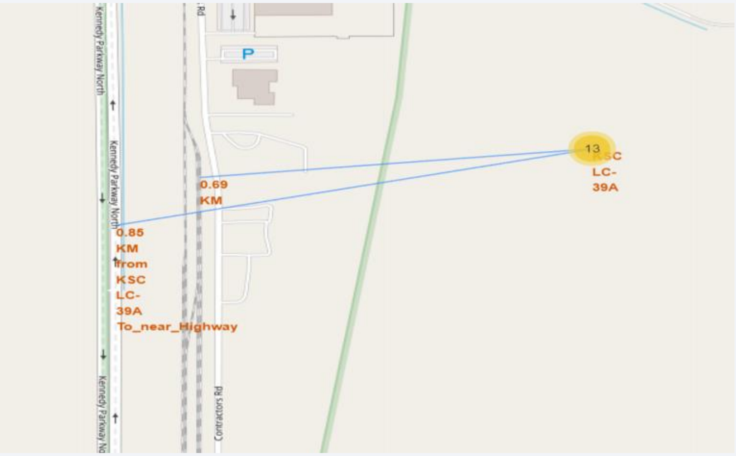
Github repo: https://github.com/parvntiwari/SpaceX-project-/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed include:
  - The names of 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
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.
  - List the records which will display the month names, failure landing_outcomes in drone ship, booster versions, and launch_site for the months in the year 2015.
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Github repo: https://github.com/parvntiwari/SpaceX-project-/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

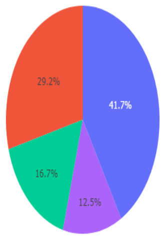


- Map markers have been added to the map with aim to finding an optimal location for building a launch site

Github repo : https://github.com/parvntiwari/SpaceX-project-/blob/main/lab_jupyter_launch_site_location.ipynb

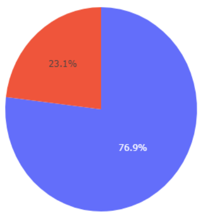
Build a Dashboard with Plotly Dash

Total success by all sites



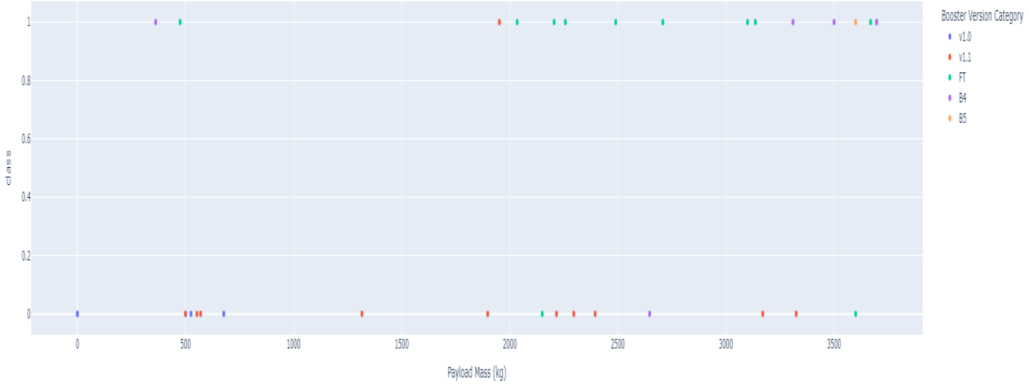
KSC LC-39A had the most successful launch from all the sites

KSC LC-39A Success vs Failure

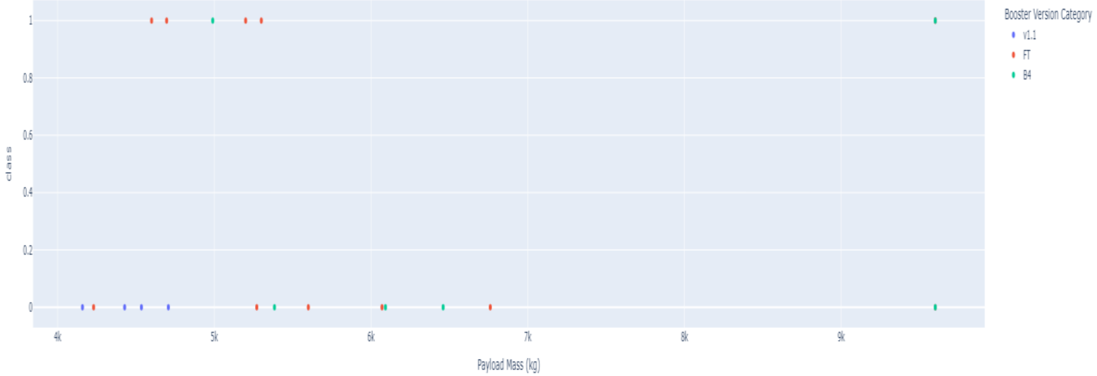


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Low weighted payload 0kg- 4000kg



Heavy weighted payload 4000kg- 10000kg

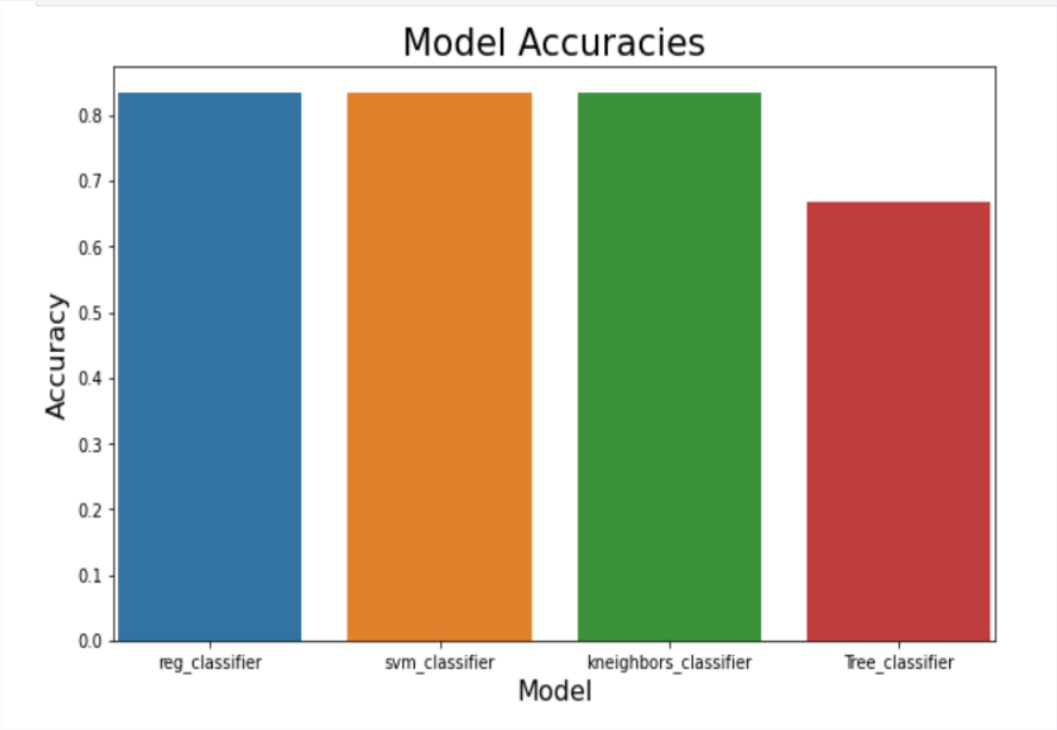


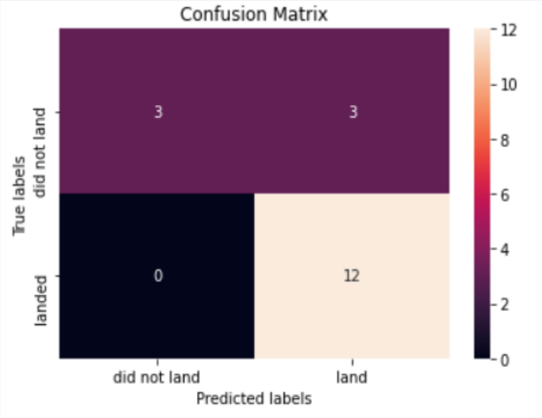
We can see the success rate of the low weighted payloads is higher than the heavy weighted payloads

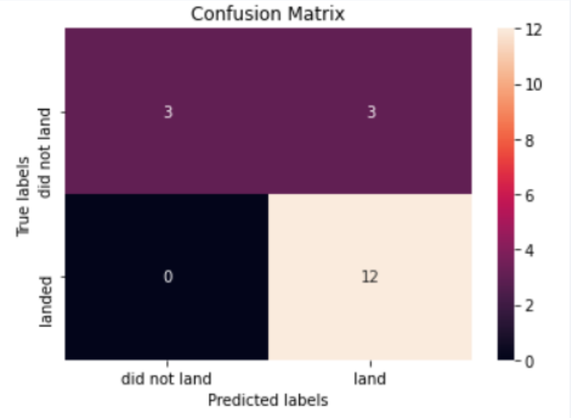
Github repo: https://github.com/parvntiwari/SpaceX-project-/blob/main/spacex_dash_app.py

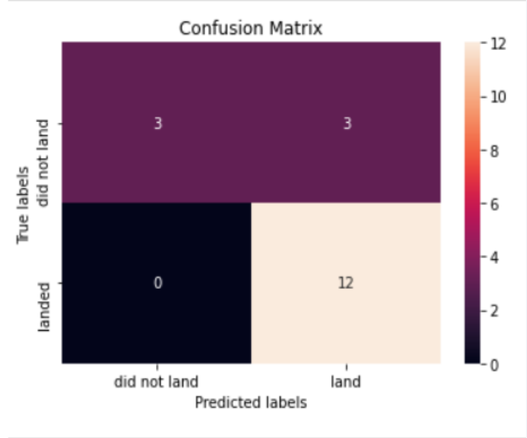
Predictive Analysis (Classification)

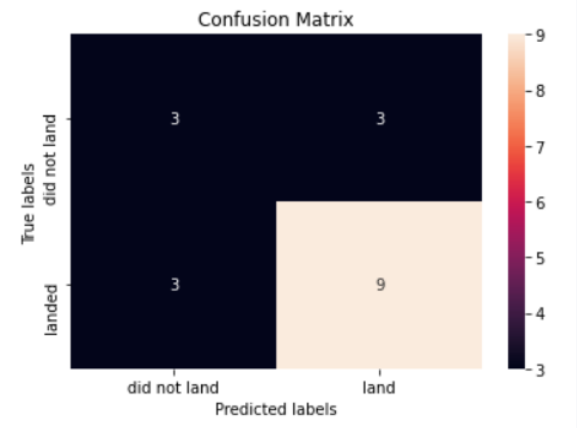
- The SMV, KNN and Logistic Regression models achieved the highest accuracy at 83.3%











Github repo: https://github.com/parvntiwari/SpaceX-project-/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

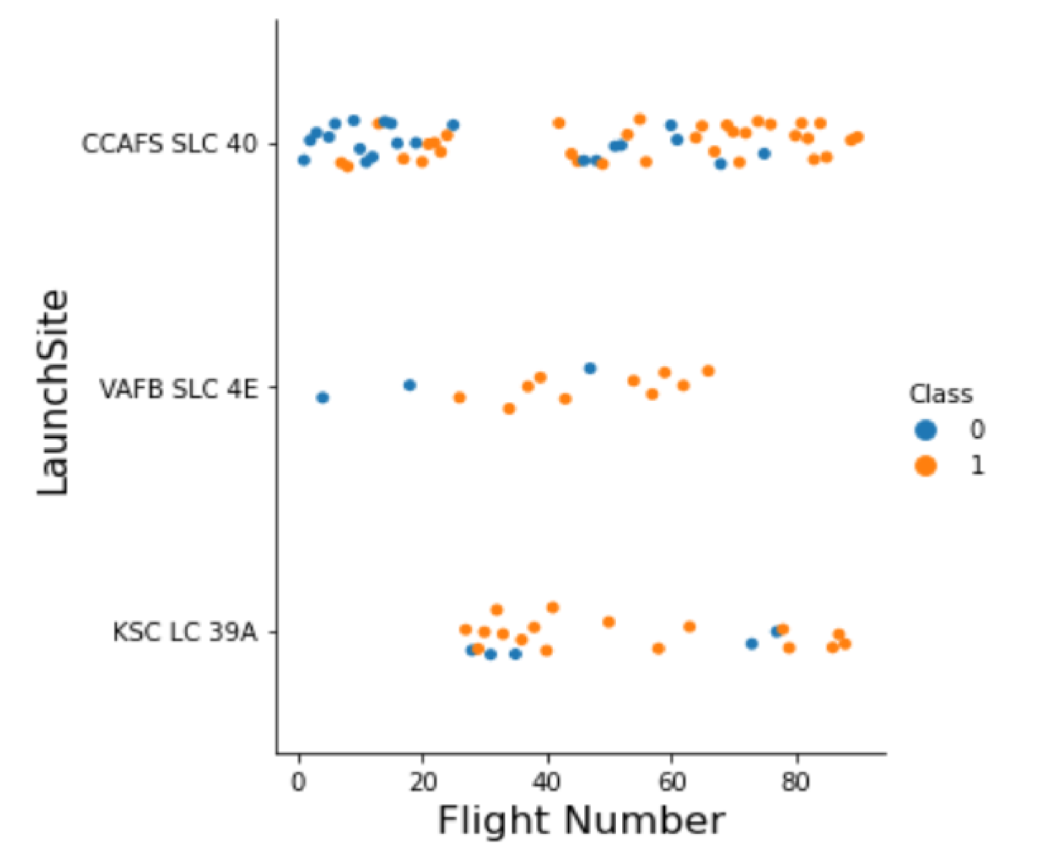
Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low-weight payloads perform better than heavier payloads.
- The success rates for SpaceX launches are directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, and ES-L1 has the best success rates.

Section 2

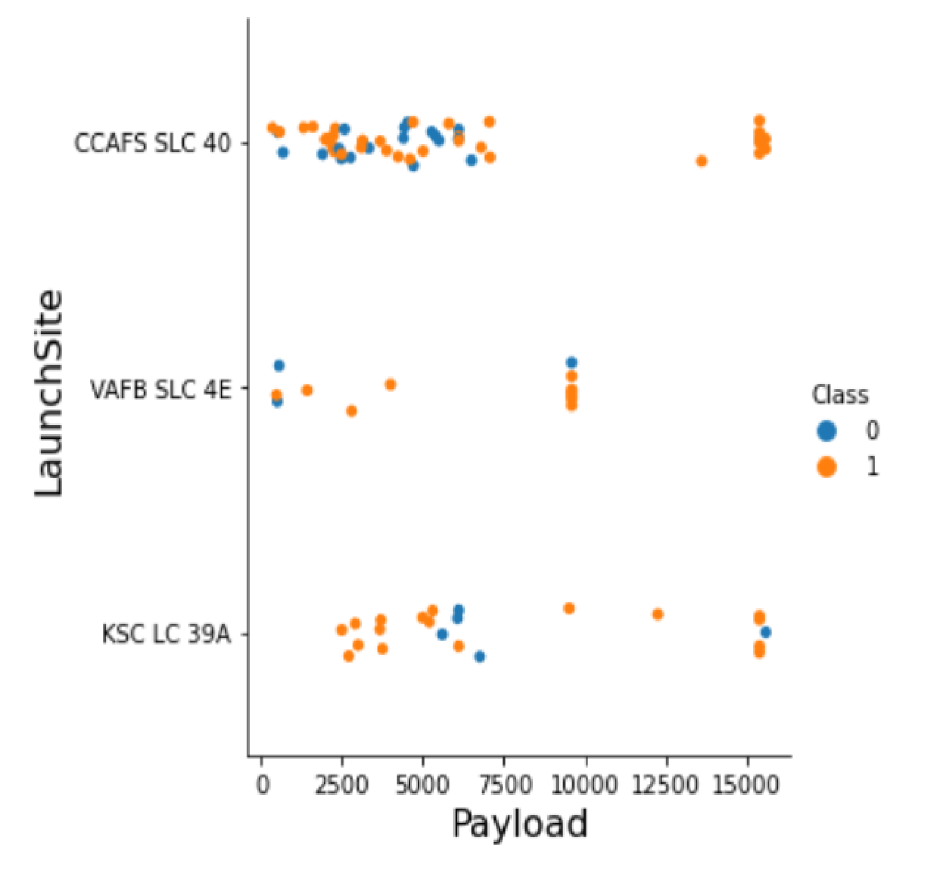
Insights drawn from EDA

Flight Number vs. Launch Site



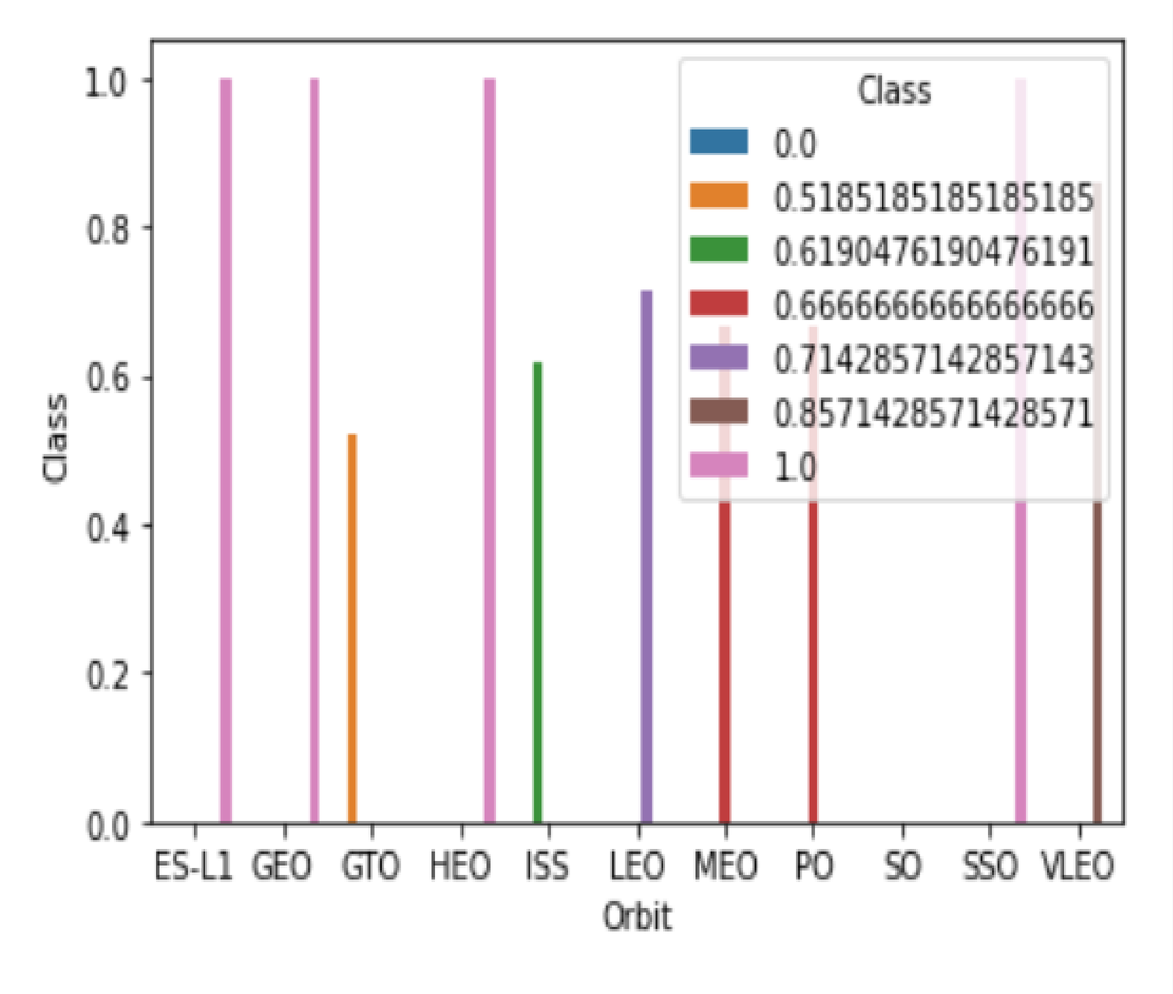
- Launch from the site of CCAFS SLC 40 is Significantly higher than Launch from other sites.
- Launch site KSC LC 39A shows more number of flights in successful Launch.

Payload vs. Launch Site



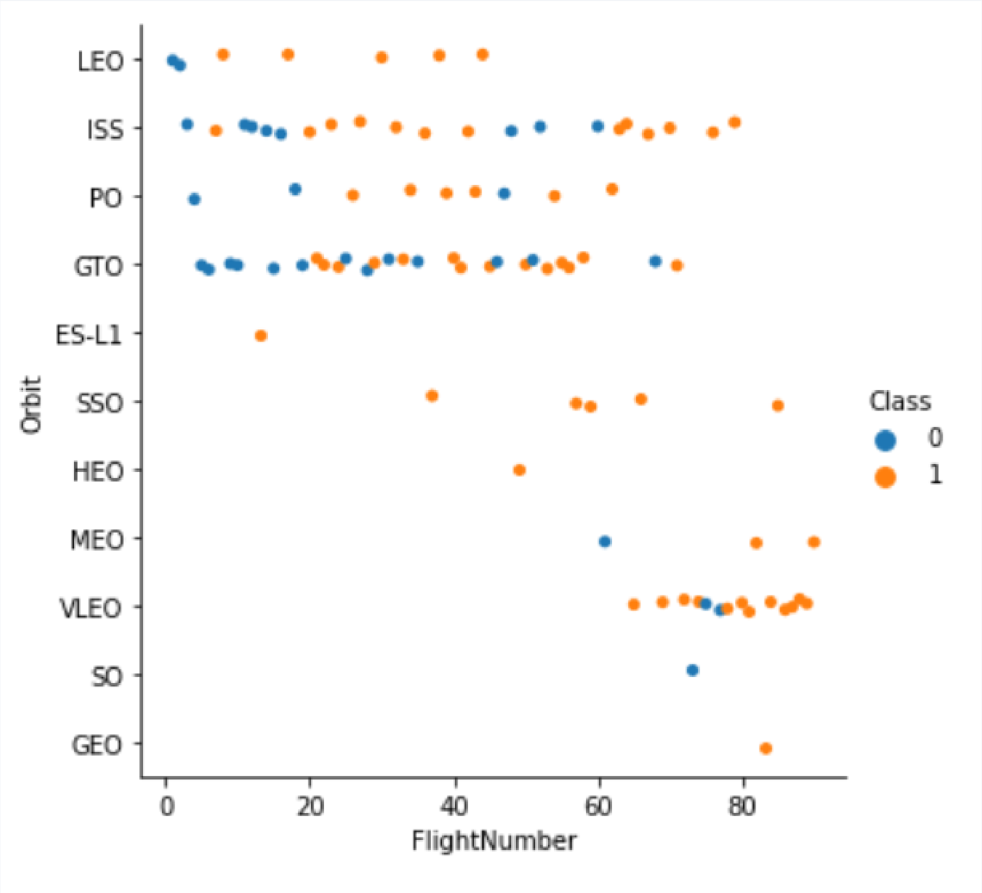
- The majority of payloads with lower mass have been launched from the CCAFS SLC-40 site.

Success Rate vs. Orbit Type



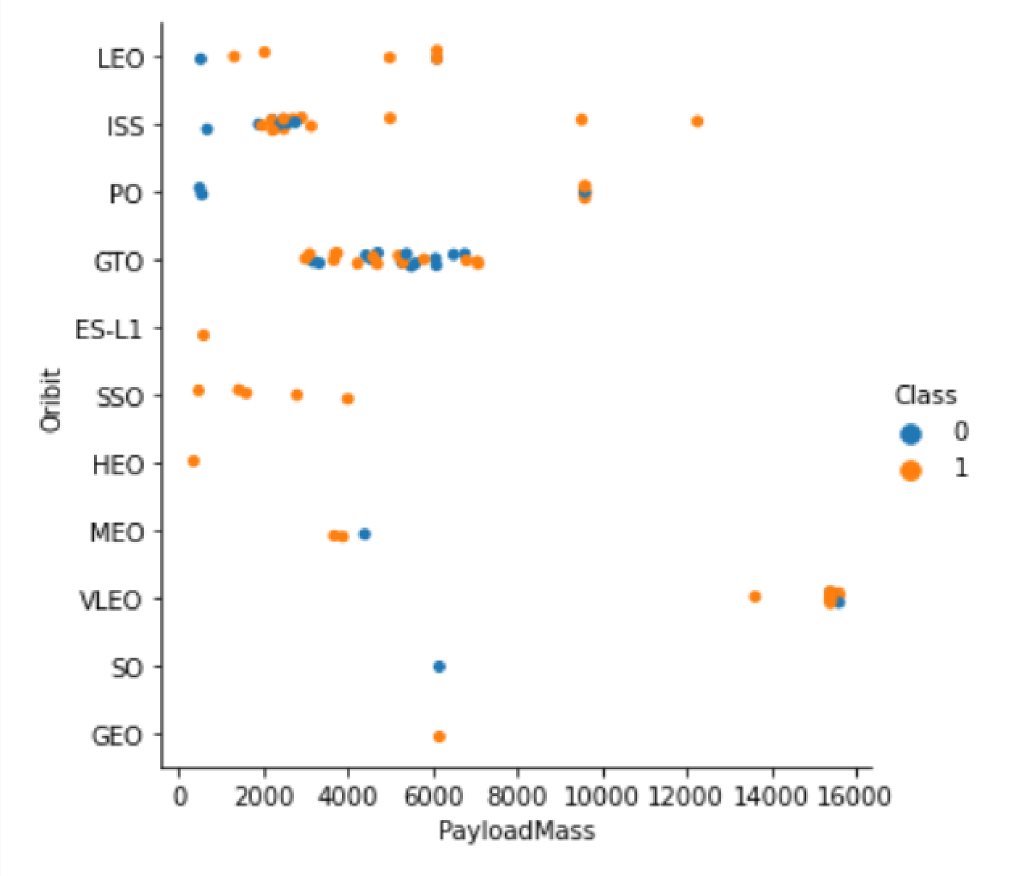
- The Orbit types of ES-L1, GEO, HEO, and SSO are among the highest success rate.

Flight Number vs. Orbit Type



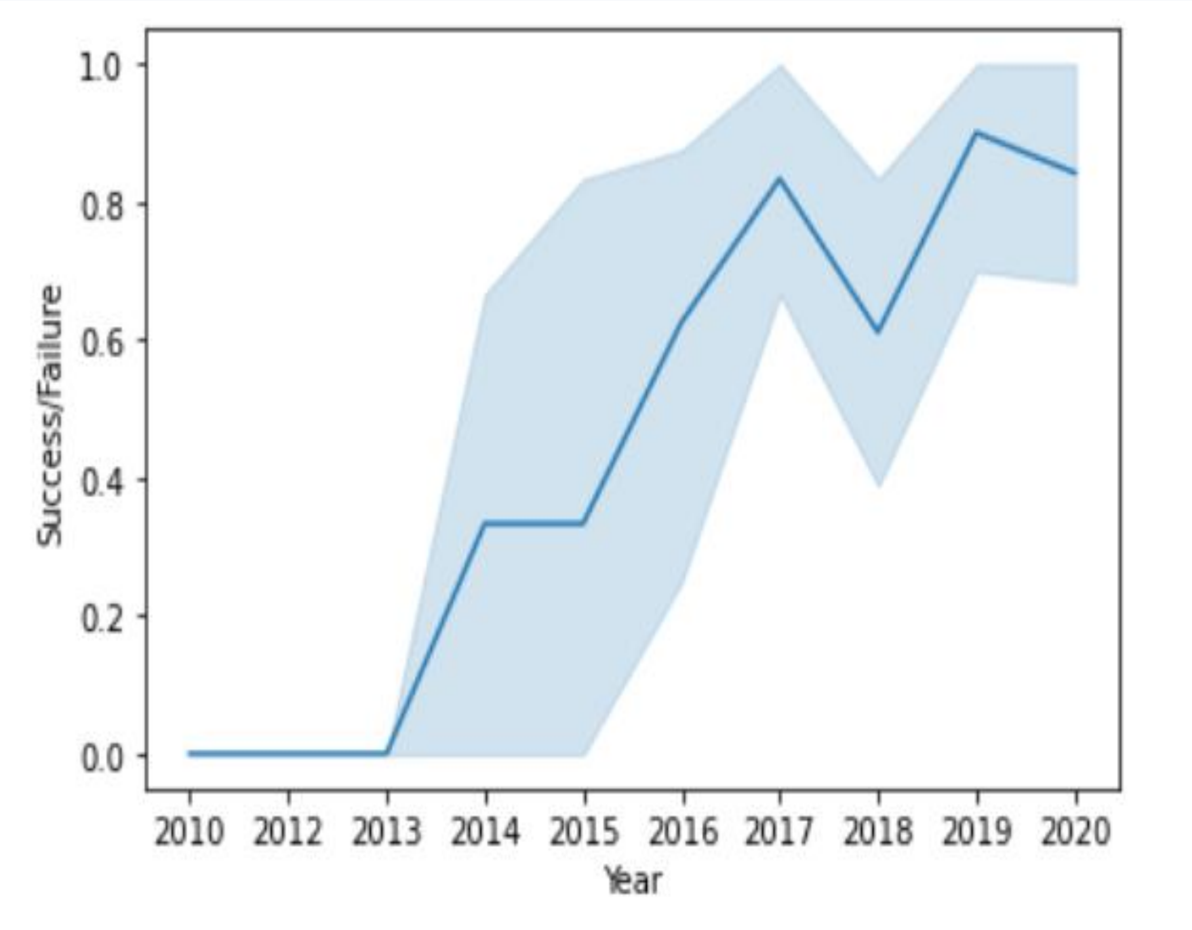
- A trend can be observed of shifting to VLEO launches in recent years with respect to the number of flights.

Payload vs. Orbit Type



- There is a strong correlation between ISS and Payload in the range around 2000, as well as GTO between the range of 4000-8000.

Launch Success Yearly Trend



- Launch success rate has increased significantly since 2013 and has stabilized since 2019, potentially due to advances in technology and lessons learned

All Launch Site Names

- %sql SELECT DISTINCT("Launch_Site") FROM SPACEXTBL;

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS PAYLOAD_MASS_BY_NASA_CRS
FROM SPACEXTBL WHERE CUSTOMER ='NASA (CRS)'
```

PAYLOAD_MASS_BY_NASA_CRS
45596

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE "Booster_Version"='F9 v1.1' ;

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- %sql SELECT min(substr(Date,7,4)||- substr(Date,4,2) ||- substr(Date,1,2)) as First_successful_landing_Date from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)' ;

First_successful_landing_Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE ("Landing _Outcome"= 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" BETWEEN 4000 AND 6000)

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

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT "Mission_Outcome", COUNT("Mission_Outcome") AS OUTCOME FROM SPACEXTBL GROUP BY "Mission_Outcome"

Mission_Outcome	OUTCOME
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" =(SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

Booster_Version
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

- %sql SELECT "Date","Booster_Version" ,"Launch_Site","Landing _Outcome" FROM SPACEXTBL WHERE substr(Date,7,4)='2015' and "Landing _Outcome" ='Failure (drone ship)';

Date	Booster_Version	Launch_Site	Landing _Outcome
10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

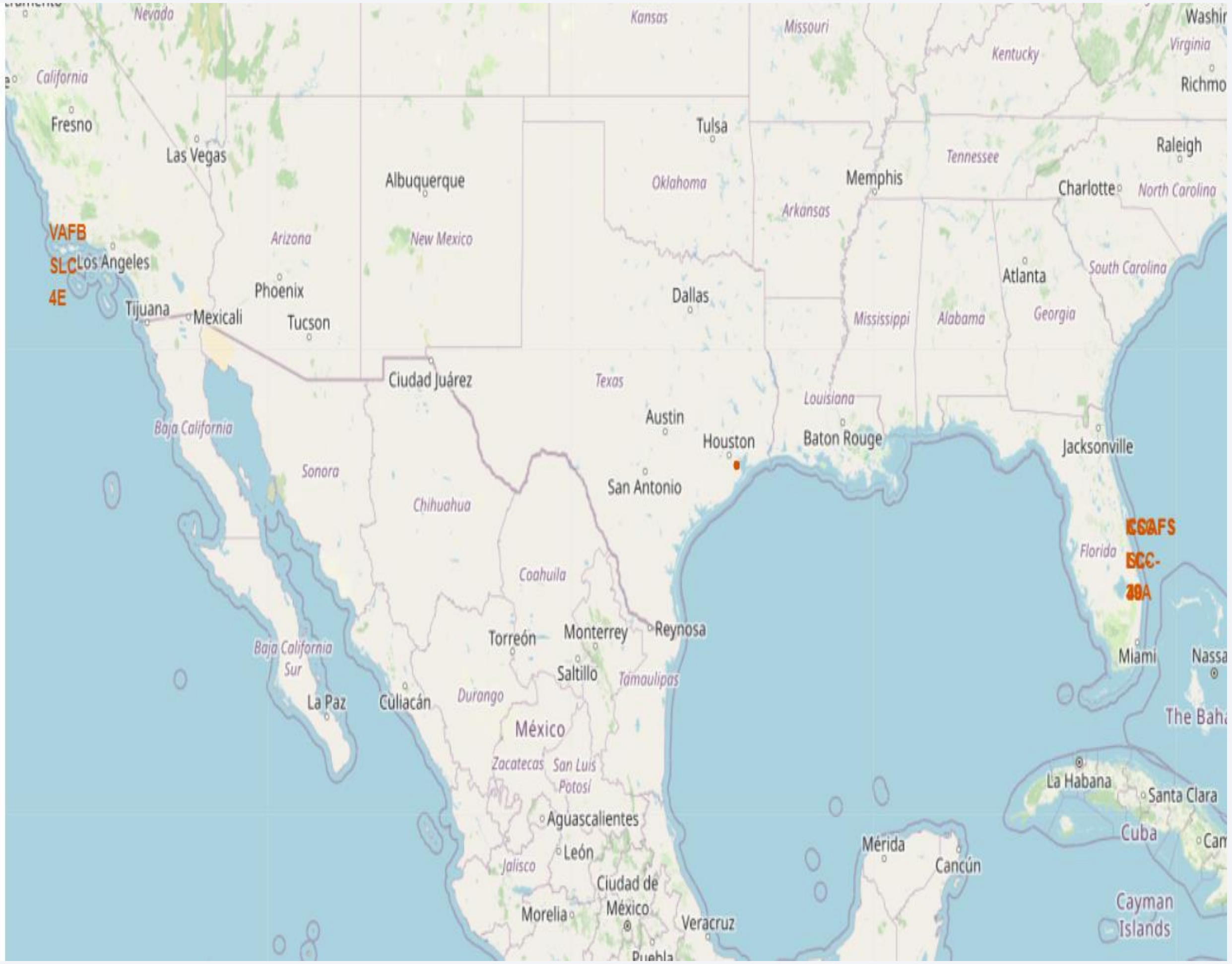
- %%sql SELECT "Landing _Outcome" ,COUNT("Landing _Outcome") AS Counts FROM SPACEXTBL WHERE "Landing _Outcome" LIKE "%Success%" AND "Date" BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY "Landing _Outcome" ORDER BY "Landing _Outcome"DESC;

Landing _Outcome	Counts
Success (ground pad)	6
Success (drone ship)	8
Success	20

Section 3

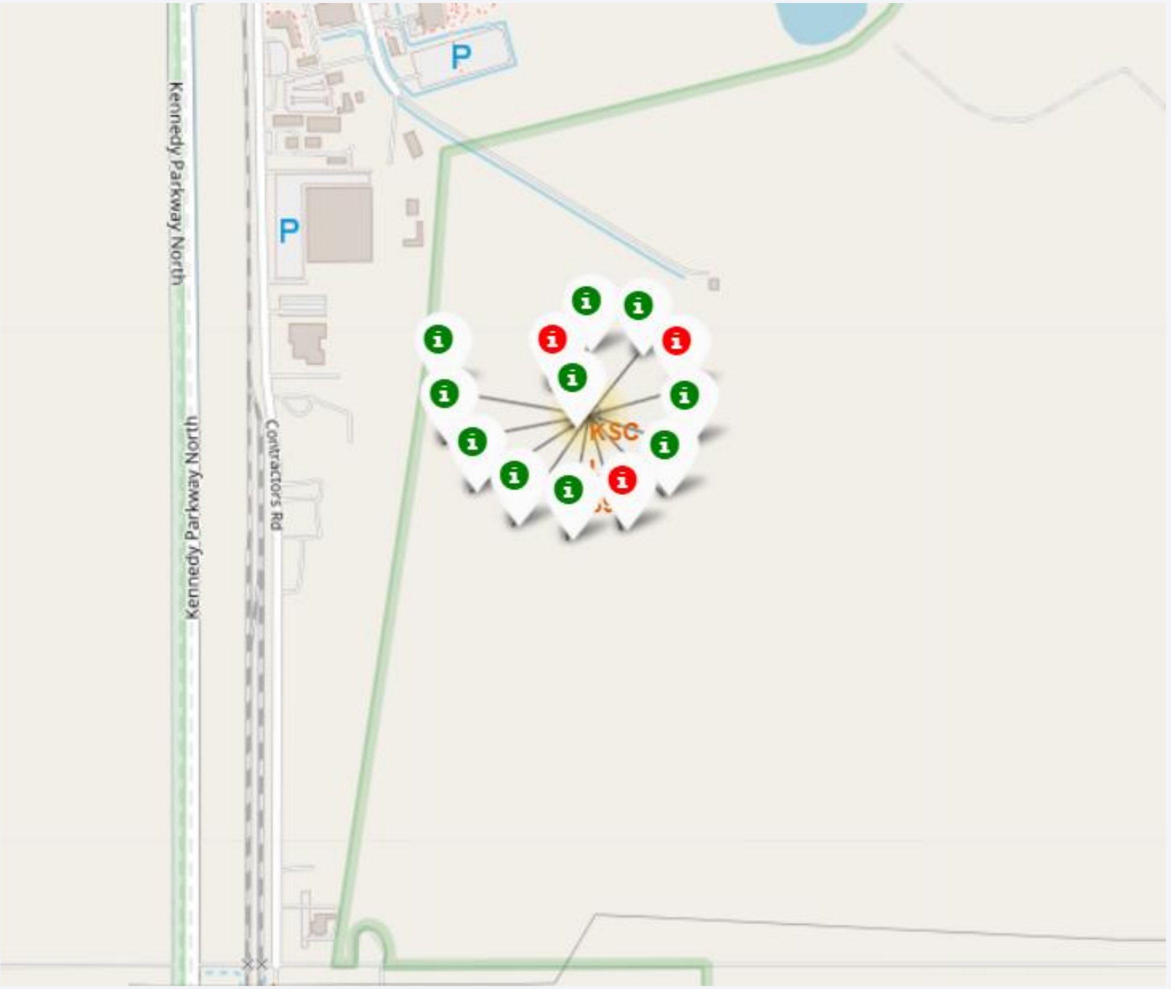
Launch Sites Proximities Analysis

All Launch Sites



- Most of the Launch sites are in proximity to the Equator line.
- When compared to any other place on the Earth's surface the land is moving faster at the equator.
- The speed of anything on the Earth's surface at the Equator is 1670 km/hr.
- Thus if a ship is launched from the equator it goes up into space and it also moves around the Earth at the same speed it was moving before launching.
- This is because of inertia. This speed will help the spacecraft keep up a good enough speed to stay in orbit.
- All launch sites are in very close proximity to the coast while launching rockets towards the ocean minimizes the risk of having any debris dropping or exploding near people

Success/Failed Launches For Each Site



- From the color-labeled markers we should be able to easily identify which launch sites have relatively high success rates.
- Green Marker = Successful launch
- Red marker = Failed launch
- Launch sites KSC LC 39A have a very high success rate.

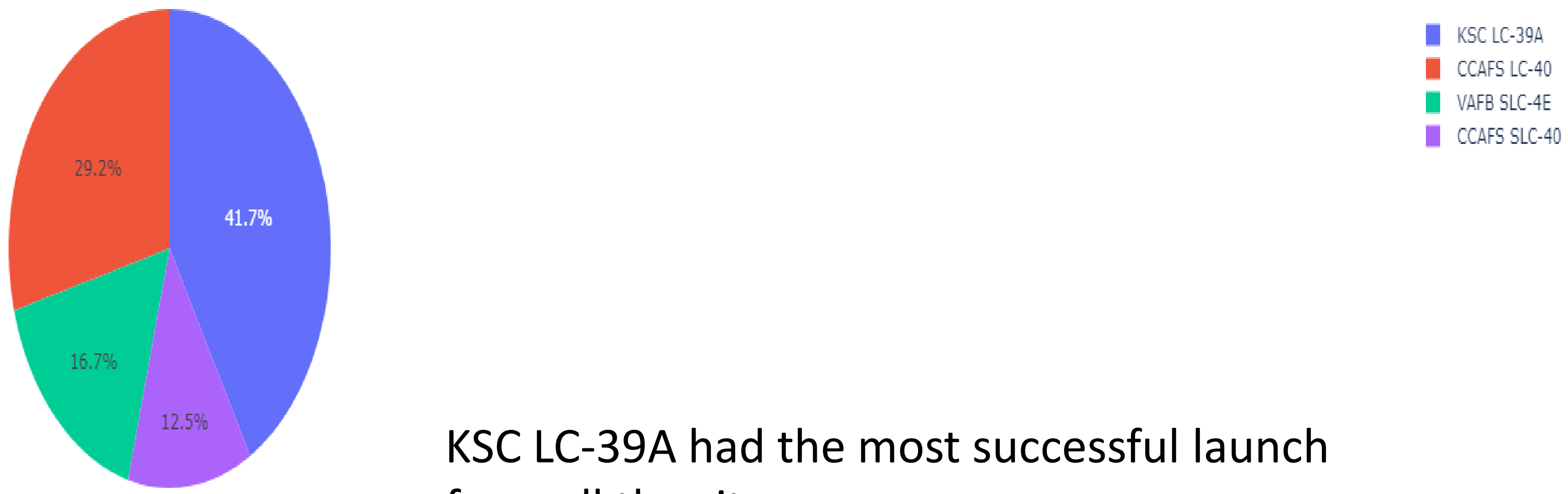
Distance Between A Launch Site To Its Proximities

Section 4

Build a Dashboard with Plotly Dash

Total Success Launches By All Sites

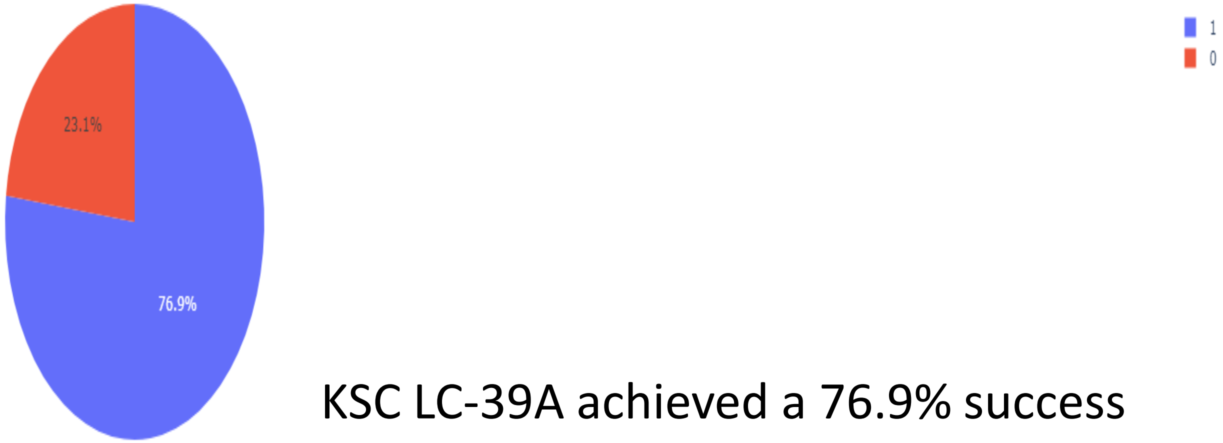
Total success by all sites



KSC LC-39A had the most successful launch from all the sites

Success Rate by Site

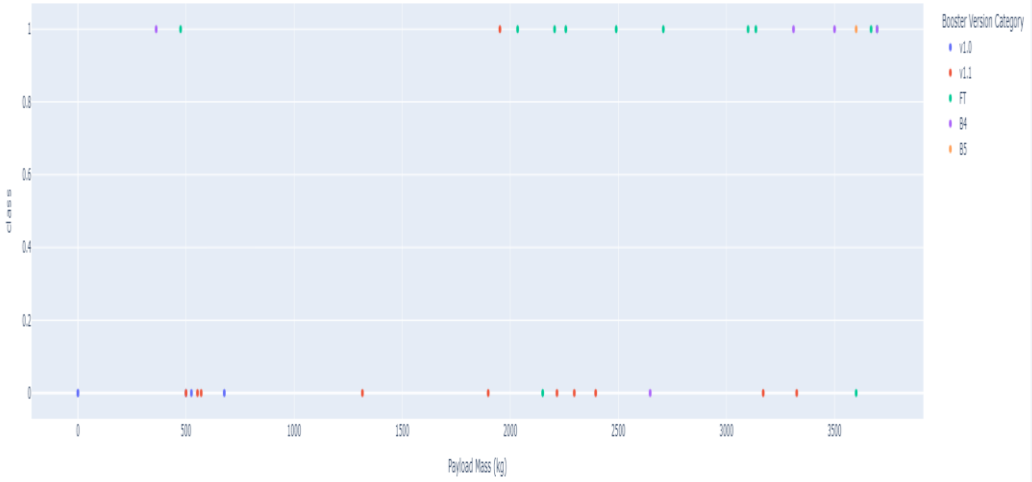
KSC LC-39A Success vs Failure



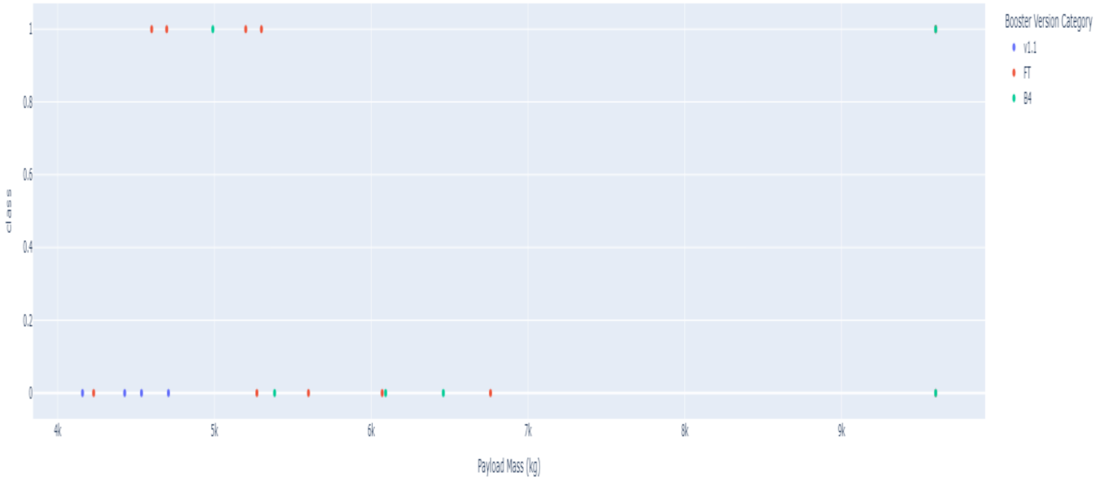
KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload Vs Launch Outcomes

Low weighted payload 0kg- 4000kg



Heavy weighted payload 4000kg- 10000kg



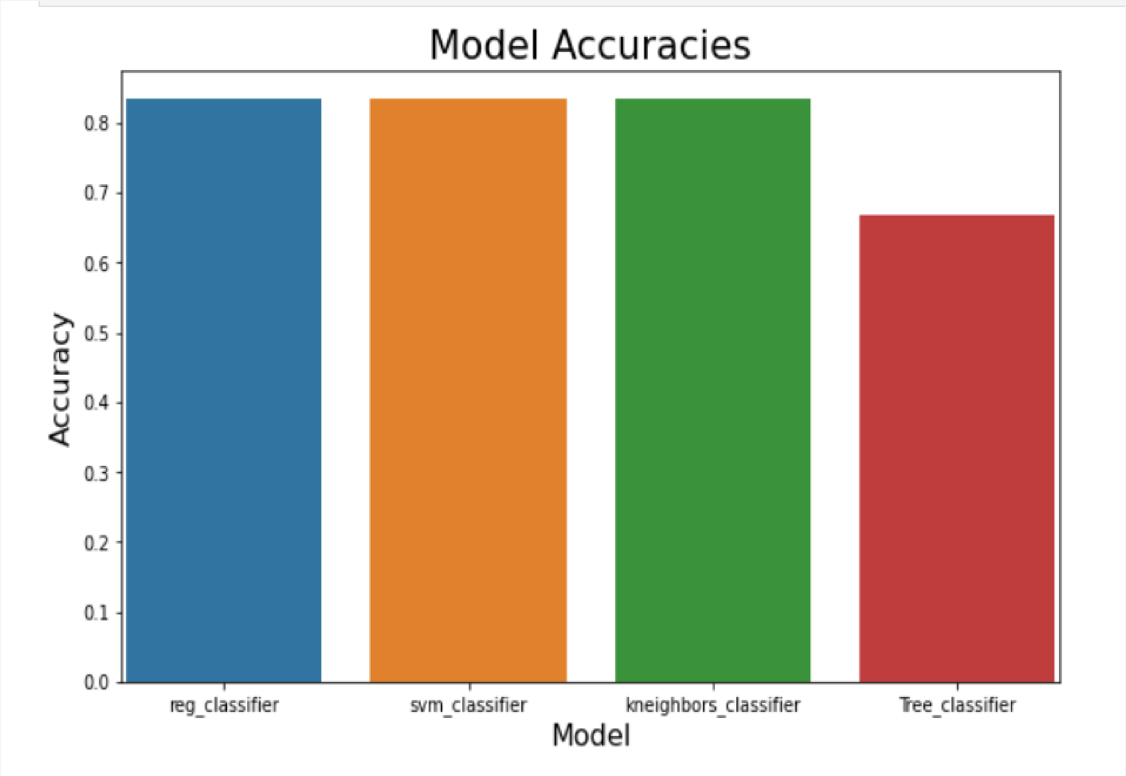
We can see the success rate of the low weighted payloads is higher than the heavy weighted payloads

Section 5

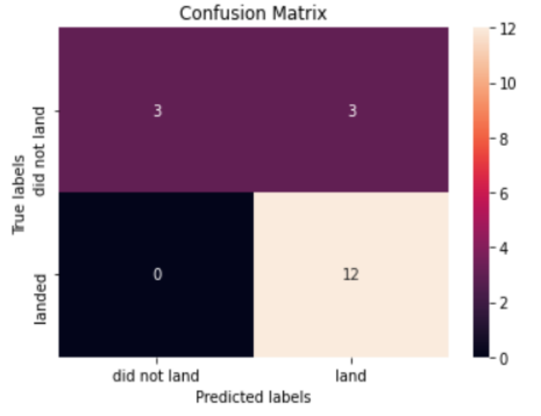
Predictive Analysis (Classification)

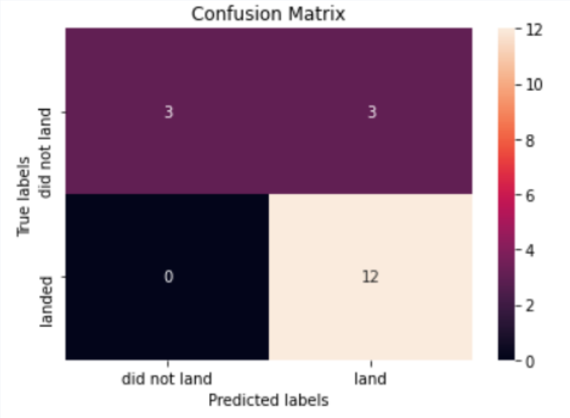
Classification Accuracy

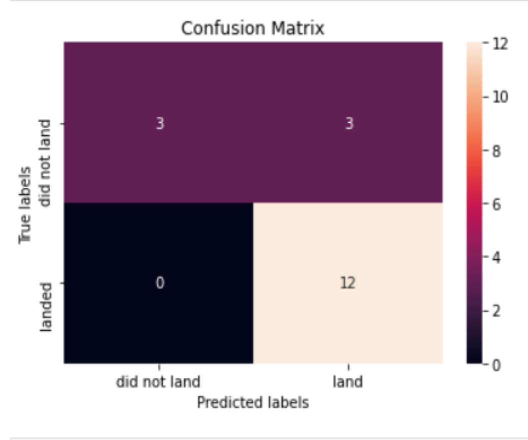
- The SMV, KNN, and Logistic Regression models achieved the highest accuracy at 83.3%

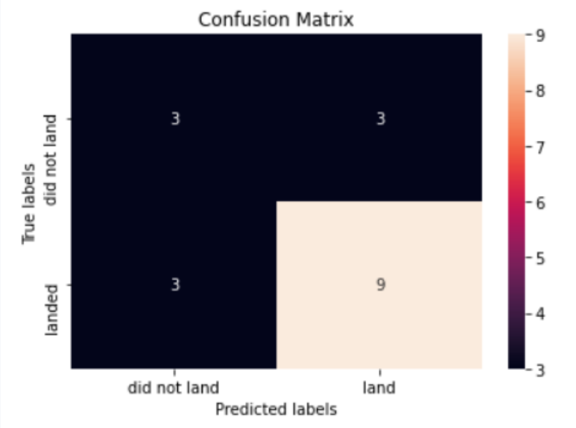


Confusion Matrix









Conclusions

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy for this dataset.
- Low-weight payloads perform better than heavier payloads.
- The success rates for SpaceX launches are directly proportional time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- Orbit GEO, HEO, SSO, and ES-L1 has the best success rates.

Thank you!