IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammed S AlShaikh
13-Aug-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with SQL
  - EDA with Data Visualization
  - Visualizing Map using Folium
  - Dashboard Building using Plotly Dash
- Summary of all results
  - EDA Result
  - Interactive Dashboard Analytics demo
  - Predictive Analytics Results

Introduction

- Project background and context

- The era of commercial space has arrived, and there are several companies that are making space travel affordable for everyone. SpaceX 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 SpaceX can reuse the first stage.

- Problems you want to find answers
    - Correlation Between Site and Rocket Variables
    - Predict success rate of first stage landing depend on exampled labels available

Section 1

Methodology

Methodology

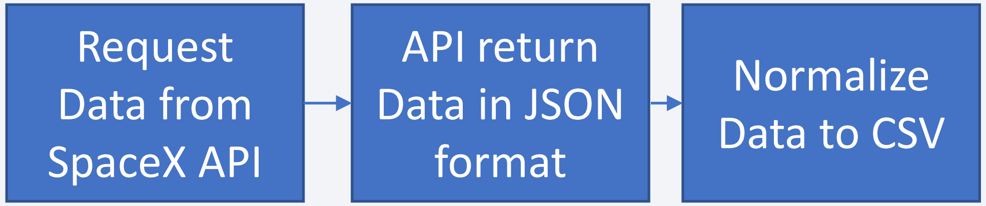
Executive Summary

- Data collection methodology:
    - SpaceX Rest API
    - Web Scrapping Falcon9 from Wikipedia Falcon9
- Perform data wrangling
    - Convert Landing Outcome to Training label ,1 Success Landing , 0 Failed Landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Find best Hyperparameter for SVM, Classification Trees and Logistic Regression Using GridSearchCV

Data Collection

- The Data Collection process include
  - Rest API Calls from SpaceX APIs
  - Web Scrapping Wikipedia Using BeautifullSoup freamwork
- Data collected from Rest API:FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Block, ReusedCount, Serial, Longitude, Latitude
- Data Collected from Web Scrapping from Wikipedia: Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

SpaceX API Flow



Web Scrapping Flow

Data Collection – SpaceX API

- Request Data from Launches API from SpaceX

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

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

- Normalize JSON Response to Pandas Dataframe

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

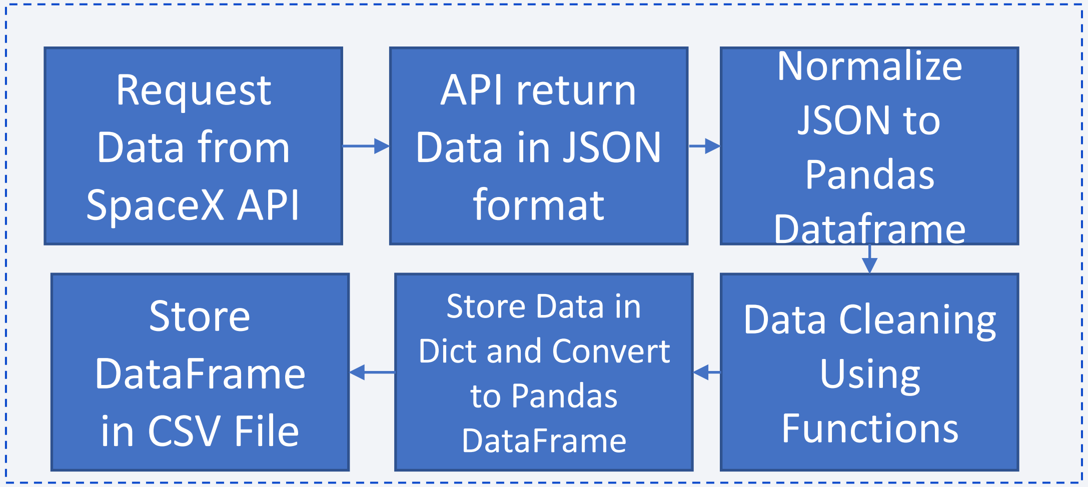
- Data Cleaning

```
# Call getBoosterVersion
getBoosterVersion(data)
```

```
# Call getPayloadData
getPayloadData(data)
```

```
# Call getLaunchSite
getLaunchSite(data)
```

```
# Call getCoreData
getCoreData(data)
```



- Convert Disc to Pandas Dataframe

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

- Store Dataframe in CSV format

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

[Data Collection API GitHub]()

Data Collection - Scraping

- Get HTML Response

```python
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

- Create BeautifulSoup Object

```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.text,'html.parser')
```

- Parse Tables in Response

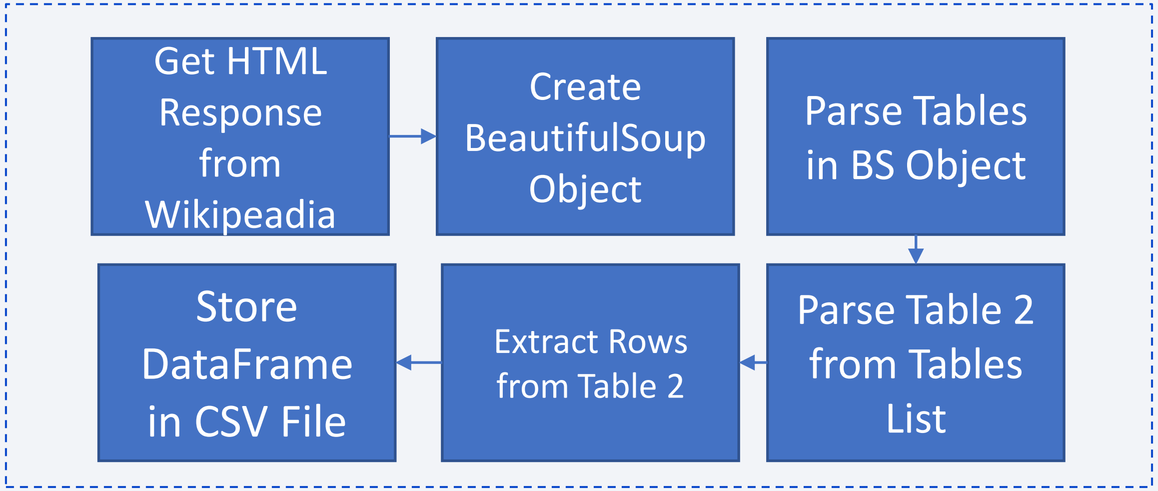
```python
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
type(html_tables)
```

- Parse Table 2

```python
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

- Extract Rows from Table

```python
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()
        else:
            flag=False
```



WebScrapping GitHub

Data Wrangling

- There are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident:
  - True Ocean means the mission outcome was successfully landed to a specific region of the ocean.
  - False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.
  - True RTLS means the mission outcome was successfully landed to a ground pad.
  - False RTLS means the mission outcome was unsuccessfully landed to a ground pad.
  - True ASDS means the mission outcome was successfully landed on a drone ship
  - False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

  - For successful Landing will be Encoded as 1
  - For False Landing will be encoded as 0

Data Wrangling Github

Data Wrangling

- Calculate the number of launches on each site

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

- Calculate the number and occurrence of each orbit

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

- Calculate the number and occurence of mission outcome per orbit type

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

- Create a landing outcome label from Outcome column

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = []
for ele in df['Outcome']:
    if ele in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
landing_class
```

- Calculate Success Rate

```
df["Class"].mean()
```

```
0.6666666666666666
```

EDA with Data Visualization

- CatPlot
  - FlightNumber VS PayloadMass
  - FlightNumber vs LaunchSite
  - Payload vs Launch Site
  - FlightNumber and Orbit type
  - Payload and Orbit type
- Bar Chart
  - success rate of each orbit type
- Line Chart
  - launch success yearly trend

EDA with Data Visualization Github

EDA with SQL

- After loading Data to Database following questions will be Answered
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string 'CCA'**
  - **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - **Display 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 failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015**
  - **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**

[EDA with SQL Github]()

Build an Interactive Map with Folium

- Following Objects added to Map
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities
- Objects added to Map Allow us to answer following Questions
  - Are launch sites in close proximity to railways?Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes

[Interactive Maps with Folium Github]

Build a Dashboard with Plotly Dash

- The dashboard application contains a pie chart and a scatter point chart.
  - Pie chart
    - Shows Total Success Launches for All sites
    - Shows Total Success Launches for Specific Site
  - Scatter Chart
    - Using Slider to Range Payload Mass(kg)
    - Shows Correlation of Success Among Payload Mass(kg),Booster Version for all Sites.
    - Shows Correlation of Success Among Payload Mass(kg),Booster Version for Specific Site.

SpaceX Dash Github

Predictive Analysis (Classification)

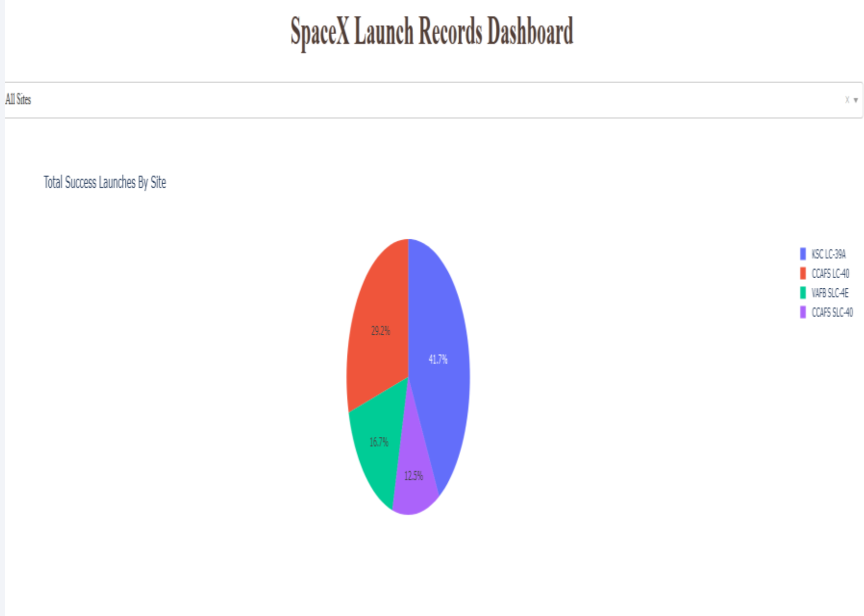
- Perform exploratory Data Analysis and determine Training Labels
  - Create a numpay array using to_numpy method for the class
  - Standardize the data
  - Split into training data and test data

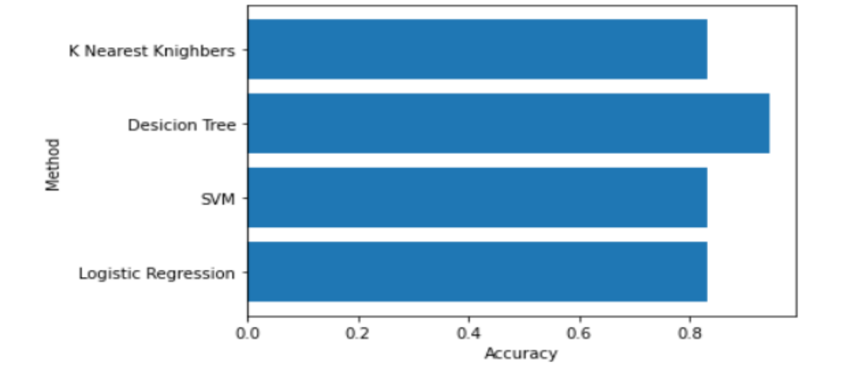
- Find best Hyperparameter for Logistic Regression, SVM, Classification Trees and K Nearest Knighbers
  - Using GridSearchCV
  - Find the method performs best using test data

[Predictive Analysis GitHub]

Results

- Exploratory data analysis results will be show with details later
- Interactive analytics demo in screenshots
- Predictive analysis results show similar results with Decision Tree in lead .





Section 2

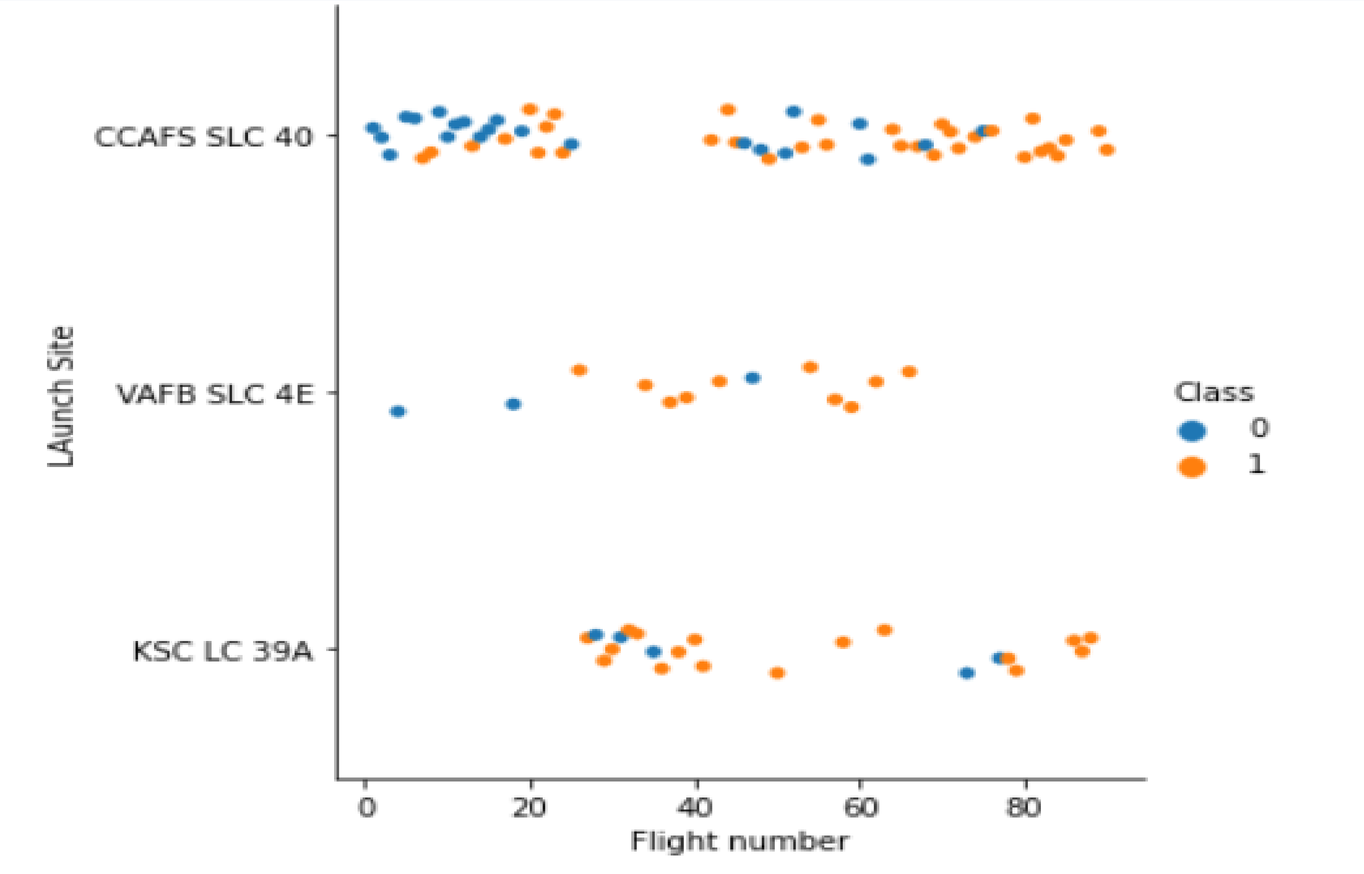
Insights drawn from EDA

Flight Number vs. Launch Site

- `Class 0` (blue) represents unsuccessful launch, and `Class 1` (orange) represents successful launch.

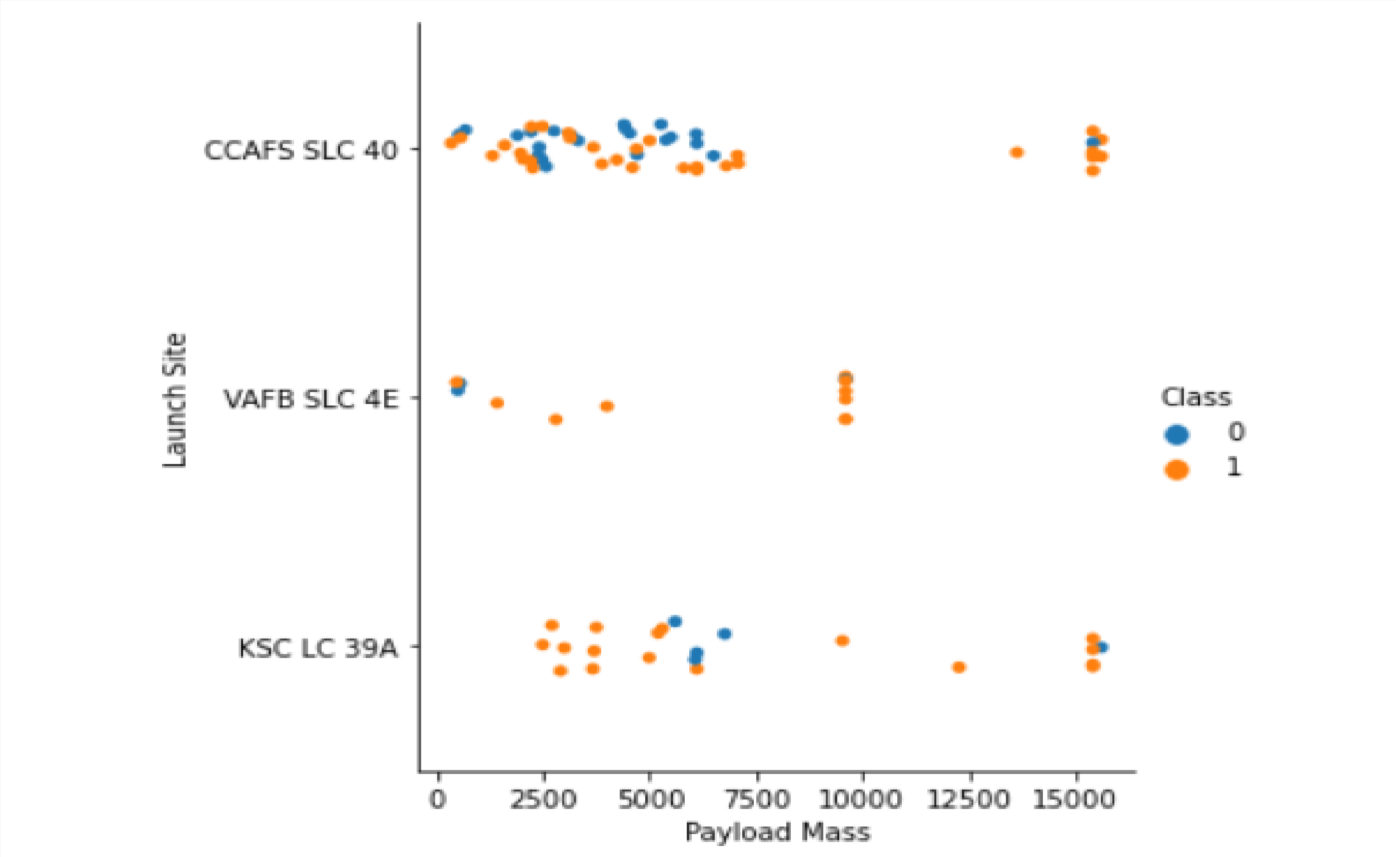
- figure shows that **<u>the success rate increased as the number of flights increased</u>**.

- As the success rate has increased considerably since the *20th* flight, this point seems to be a big breakthrough.



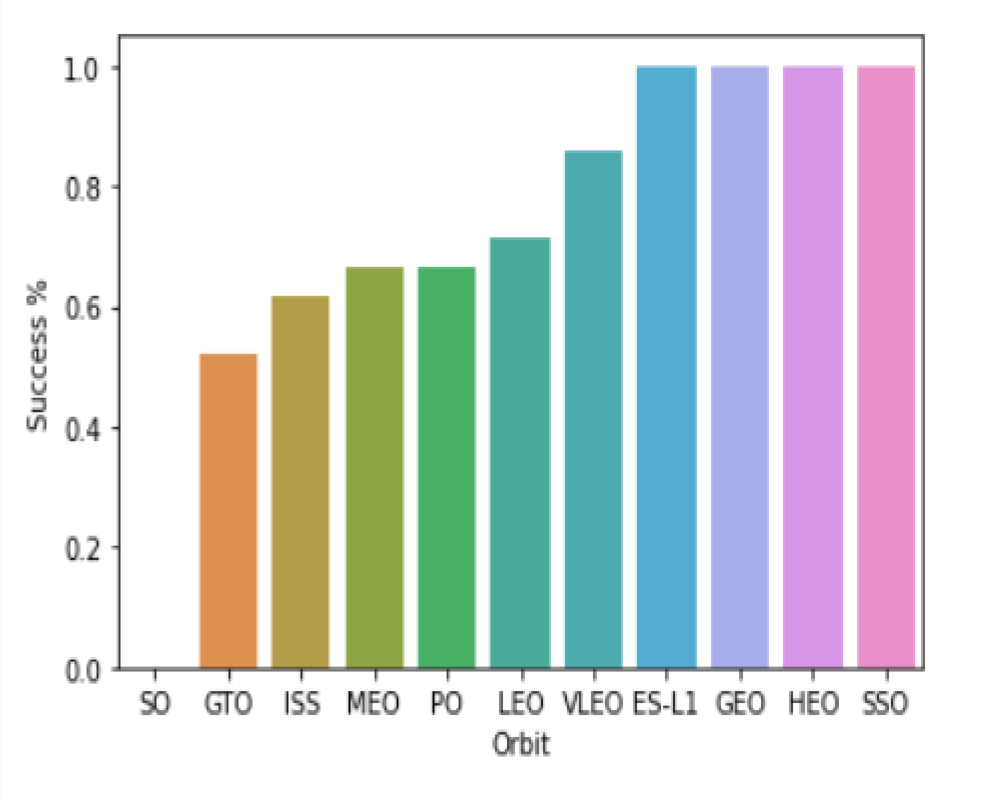
Payload vs. Launch Site

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- CCAFS SLC 40: it shows high failiure rate at LOW AND MID payload and high success rate in HIGH payload
- VAFB SLC 4E: it shows similar success rate at LOW and MID payload, no HIGH payload data available.
- KSC LC 39A: it shows high success rate at LOW payload, high failure rate at MID payload, high success rate at HIGH payload



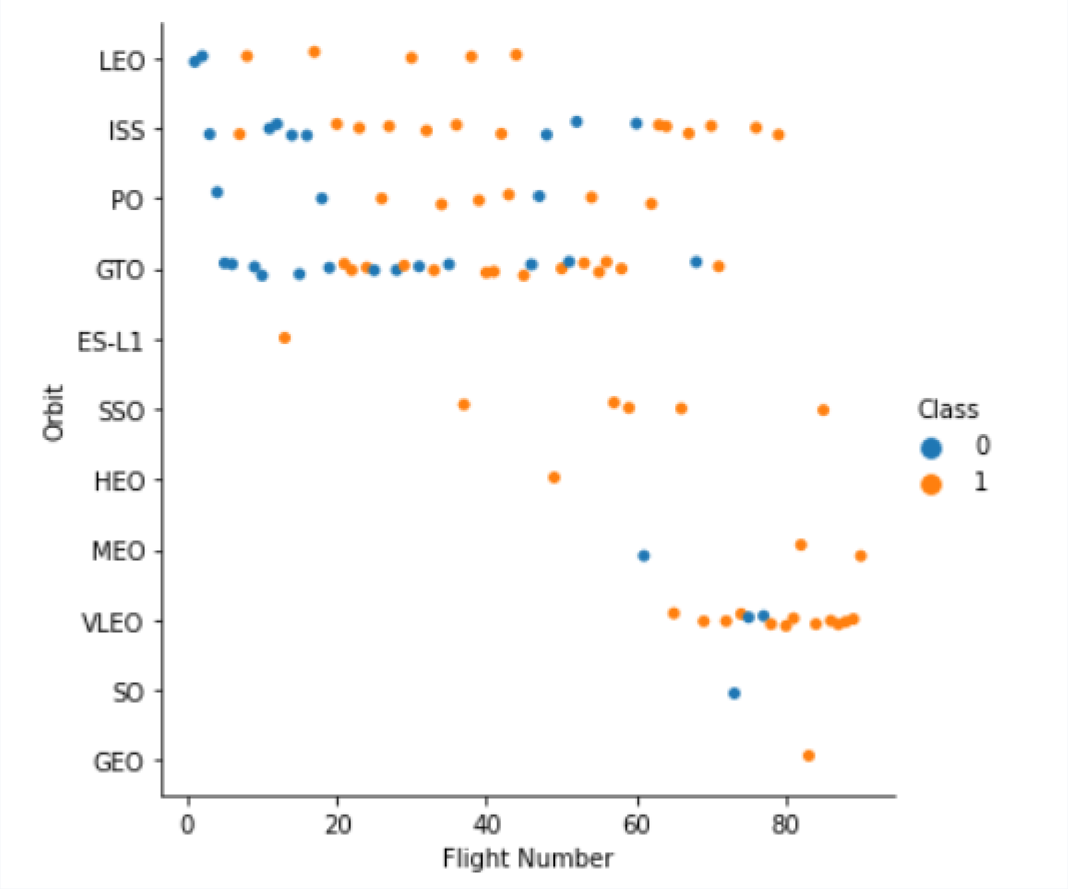
Success Rate vs. Orbit Type

- Bar chart shows 100% success rate in SSO,HEO,GEO and ES-L1 Orbits
- SO orbit 100% failure rate



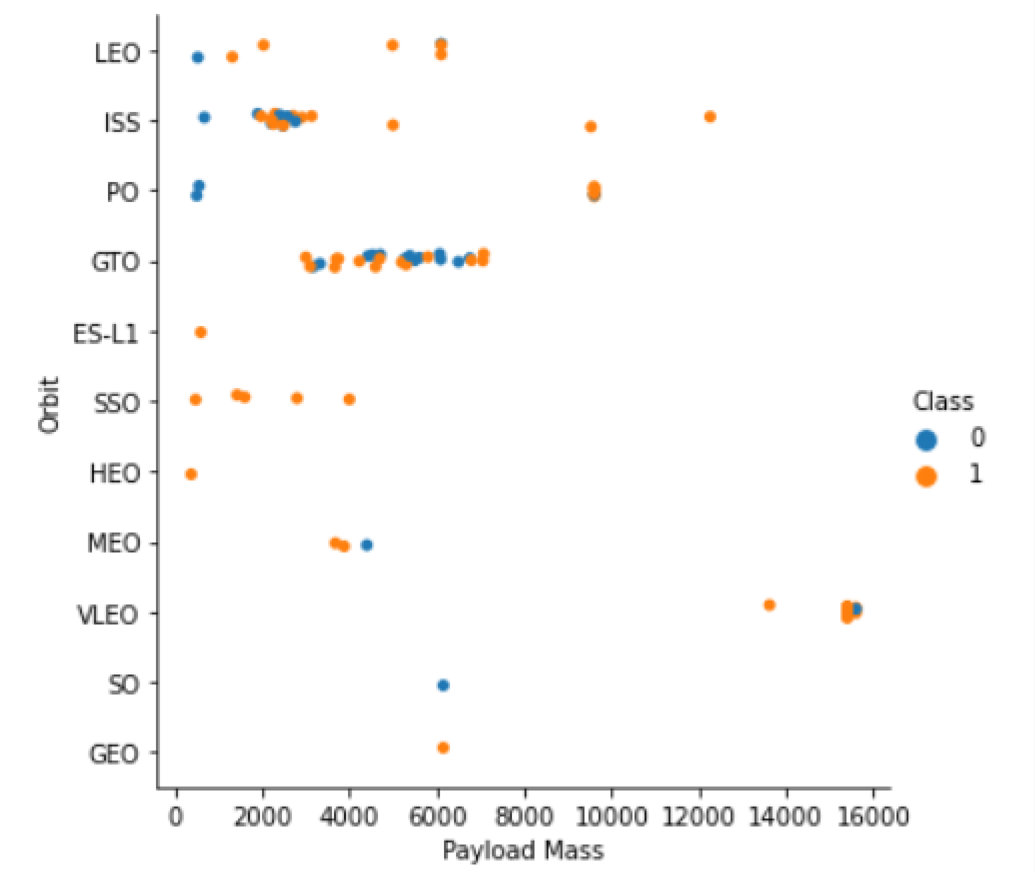
Flight Number vs. Orbit Type

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- LEO,ISS,PO and GTO has must flights with high failure rate at start of each orbit launch ,but rate of success starts to increase
- SSO and VELO consider new orbits with very high success rate
- ES-L1,HEO,SO and GEO consider outliers with single attempt



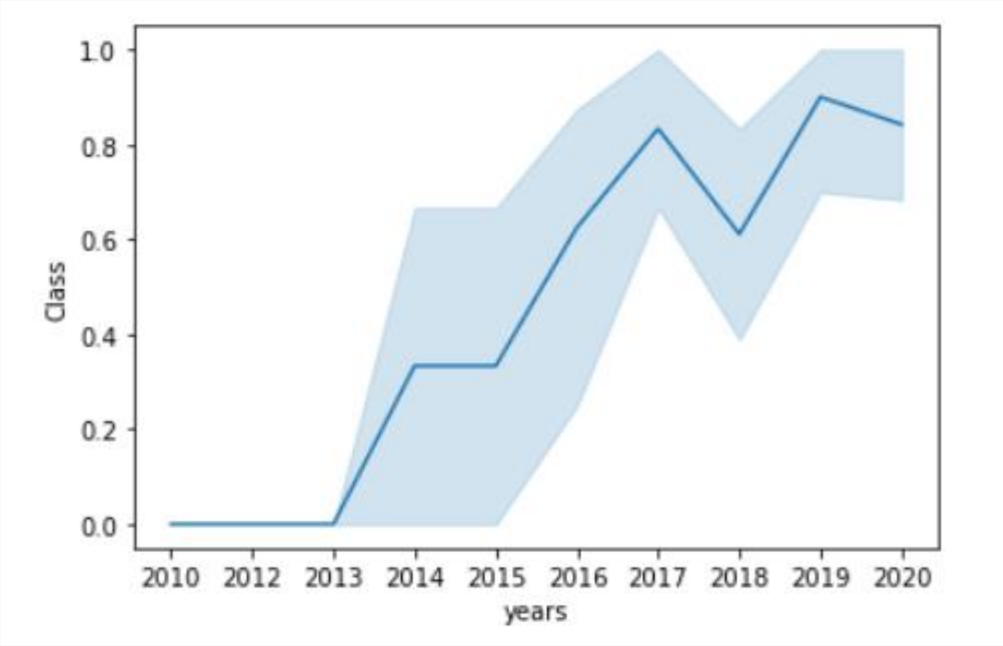
Payload vs. Orbit Type

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- For LOW payload mass ,SSO orbit shows promising performance with 100% success .
- For HIGH payload mass , VLEO orbit shows promising performance with 80% success rate.
- GTO,ISS and LEO orbits show mixed success rate
- PO Orbit shows 100% failure with LOW payload mass and medium success rate in MID payload mass



Launch Success Yearly Trend

- Since 2013, the success rate has continued to increase until 2017.
- The rate decreased slightly in 2018.
- Recently, it has shown a success rate of more than 80%.



All Launch Site Names

- The names of the unique launch sites

```
select distinct launch_site from spacextbl
```

- Result

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

- We have **<u>Four</u>** Unique Launching Sites

Launch Site Names Begin with 'CCA'

- Query for Finding 5 records where launch sites begin with `CCA`

```
select * from spacextbl
where launch_site like 'CCA%'
limit 5
```

- Query Results Using like and wildcard character % to match any string after CCA and Using LIMIT keyword to limit resultset to 5 Records

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	land
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fail
2010-12-08	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	Fail
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Query for Calculating the total payload carried by boosters from NASA

```
select sum(payload_mass__kg_) from spacextbl where customer ='NASA (CRS)'
```

- Result

45596

- By Using Aggregate Function SUM we sum all Payload Mass KG for Customer NASA

Average Payload Mass by F9 v1.1

- Query for Calculating the average payload mass carried by booster version F9 v1.1

```
select avg(payload_mass__kg_) from spacextbl where booster_version like 'F9 v1.1'
```

- Result

```
2928
```

- We use Aggregate function AVG to calculate average payload mass kg for booster version F9 v1.1

First Successful Ground Landing Date

- Query for Finding the dates of the first successful landing outcome on ground pad

```
select * from spacextbl where date = (select min(date) from spacextbl where landing__outcome = 'Success (ground pad)')
```

- Result

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- By using Sub Query and using aggregate function MIN to get minimum date with success ground pad ,we found it was F9 FT B1019 Booster in CCAFS LC-40 launching site in 22-12-2015 (dd-mm-yyyy) to orbit LEO with payload mass of 2034 KG

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
select booster_version from spacextbl
where landing__outcome = 'Success (drone ship)'
and payload_mass__kg_ > 4000
and payload_mass__kg_ < 6000
```

- Results

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

We use Greater than > and Less than < to exclude boundary values 4000 and 6000 , because Between is inclusive

Total Number of Successful and Failure Mission Outcomes

- Query to Calculate the total number of successful and failure mission outcomes

```
select mission_outcome,count(mission_outcome) from spacextbl
group by mission_outcome
```

- Result

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- By using Group by and aggregate function count , we get out of 102 missions , 99 missions with success status , 1 success with unknown payload status , and 1 in Flight Failure

Boosters Carried Maximum Payload

- Query to List the names of the booster which have carried the maximum payload mass

```
select booster_version from spacextbl
where payload_mass__kg_ in(select max(payload_mass__kg_) from spacextbl)
```

- Result

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

Using subquery to get max payload mass kg in spacex table we select booster version

2015 Launch Records

- Query to List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
select  landing__outcome,booster_version,launch_site from spacextbl
where landing__outcome = 'Failure (drone ship)'
and to_char(date,'YYYY') = '2015'
```

- Result

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- We filter landing outcomes to show 'Failure (drone ship)' and limit it to year 2015 by getting year from date using to_char and compare it to '2015'

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

- Query to 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

```
select landing__outcome,count(landing__outcome) as Count,rank() over( order by count(landing__outcome) desc)as Ranking
from spacextbl
where date between to_date('2010-06-04','YYYY-MM-DD') and to_date('2017-03-20','YYYY-MM-DD')
group by landing__outcome
```

- Result

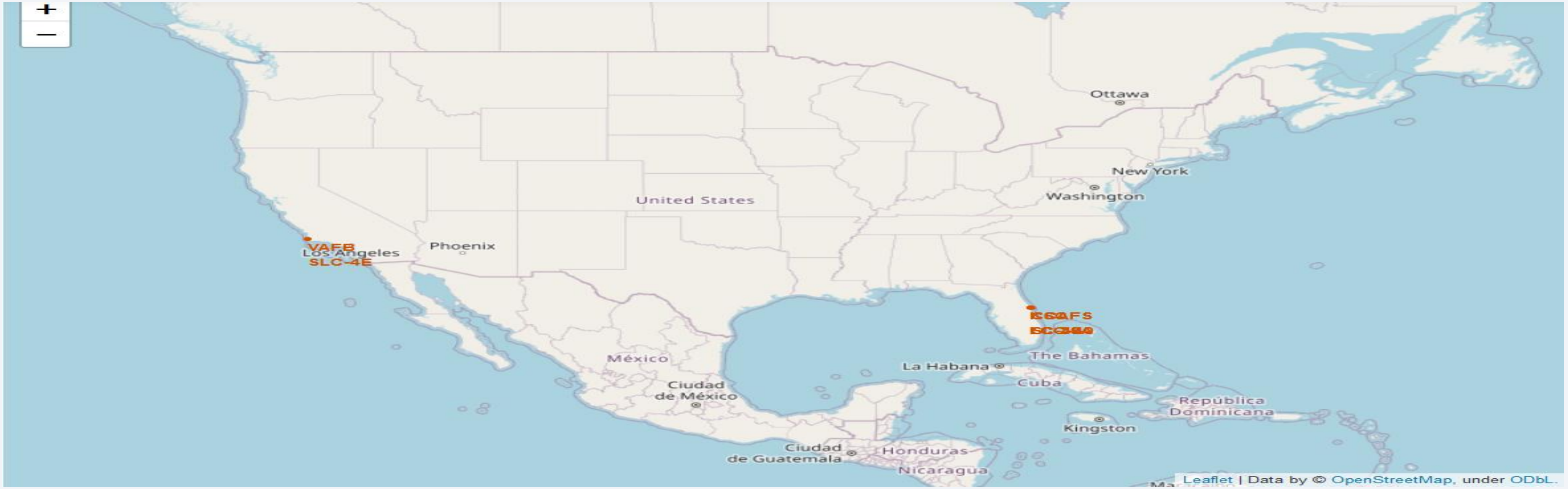
landing__outcome	COUNT	ranking
No attempt	10	1
Failure (drone ship)	5	2
Success (drone ship)	5	2
Controlled (ocean)	3	4
Success (ground pad)	3	4
Failure (parachute)	2	6
Uncontrolled (ocean)	2	6
Precluded (drone ship)	1	8

We use rank to order result in descending order , and we use Between to get result from date to date ,we group landing__outcome to get count of attempts

Section 3

Launch Sites Proximities Analysis

<Launch Site Locations>

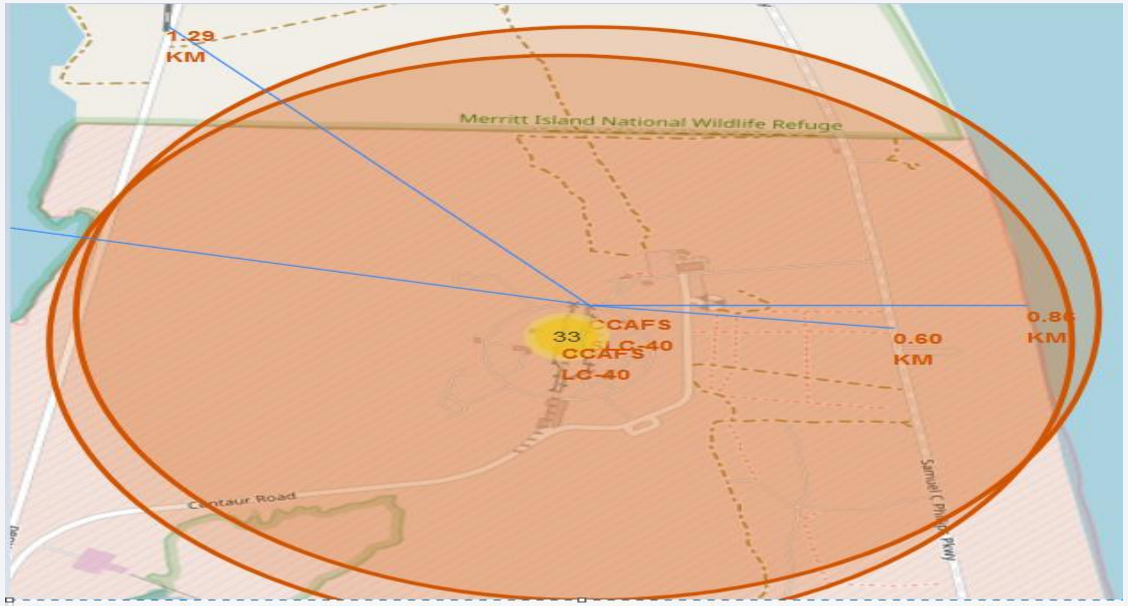


- From map , we could see that all Launch sites are near sea.
- One in California
- Three in Florida

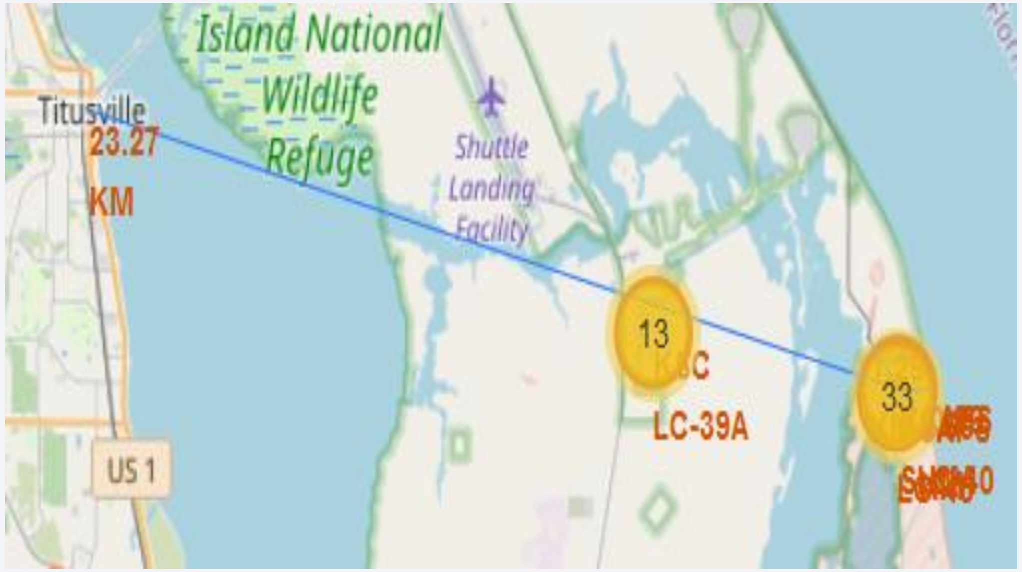
<Launch Sites Outcome>

- Outcomes showed for each launching site
- Green represent Success Landing
- Red represent failure landing

<Launch Sites Proximities>




```
*    Are launch sites in close proximity to railways?YES
*    Are launch sites in close proximity to highways?YES
*    Are launch sites in close proximity to coastline?YES
```



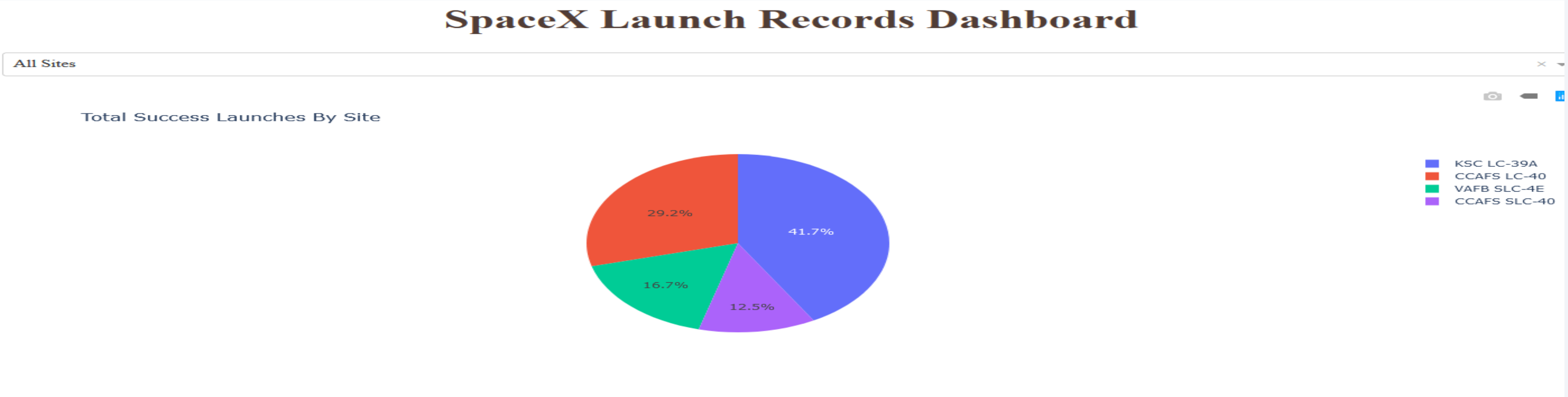

```
*    Do launch sites keep certain distance away from cities?YES
```

- It can be found that the launch site is close to **railways** and **highways** ,and is also close to **coastline** and relatively far from the **cities**

Section 4

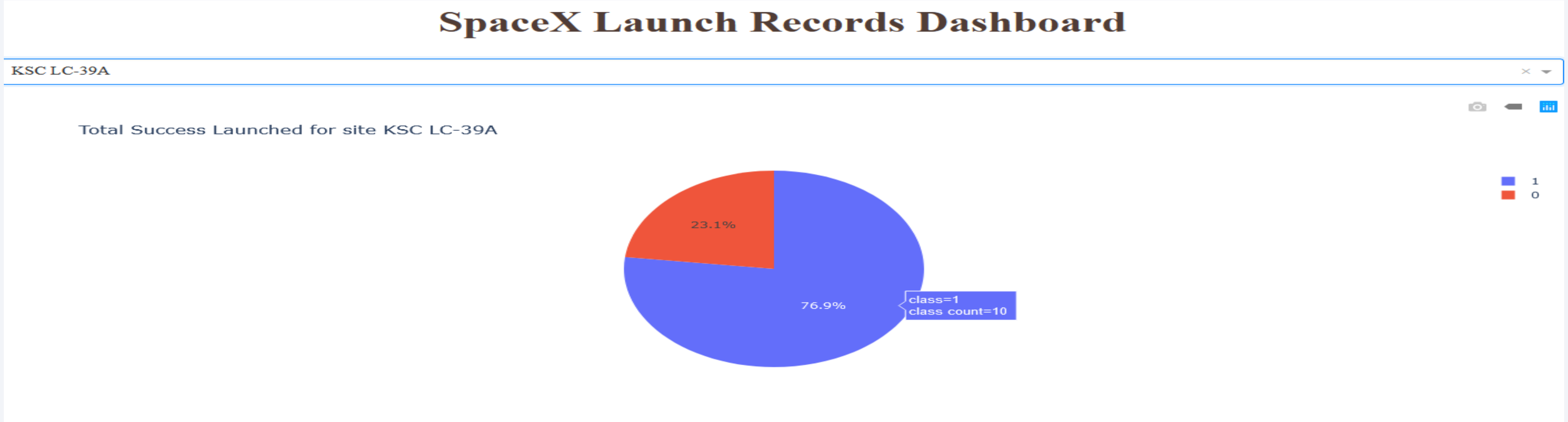
Build a Dashboard with Plotly Dash

<Total Success Launches by all Sites>



- KSC LC-39A has must success cases by 41.7%
- CCAFAS SLC-40 has lowest success rate by 12.5%

<Total Success Launched for Site KSC LC-39A >



- Chart shows 76.9% success with count of 10
- 23.1% failure with count of 3

<Dashboard Screenshot 3>

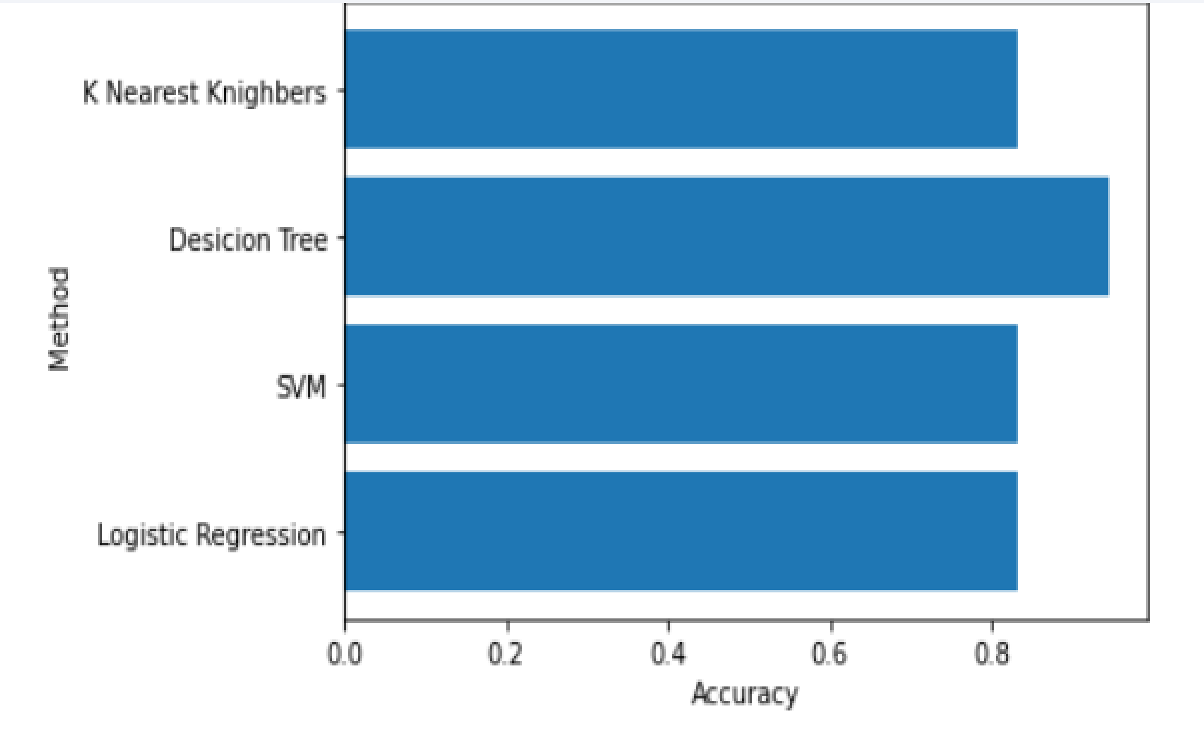
- From chart , we see must success cases has payload mass kg between 2000 and 5500 kg
- There is no success case for payload mass above 5500kg

Section 5

Predictive Analysis (Classification)

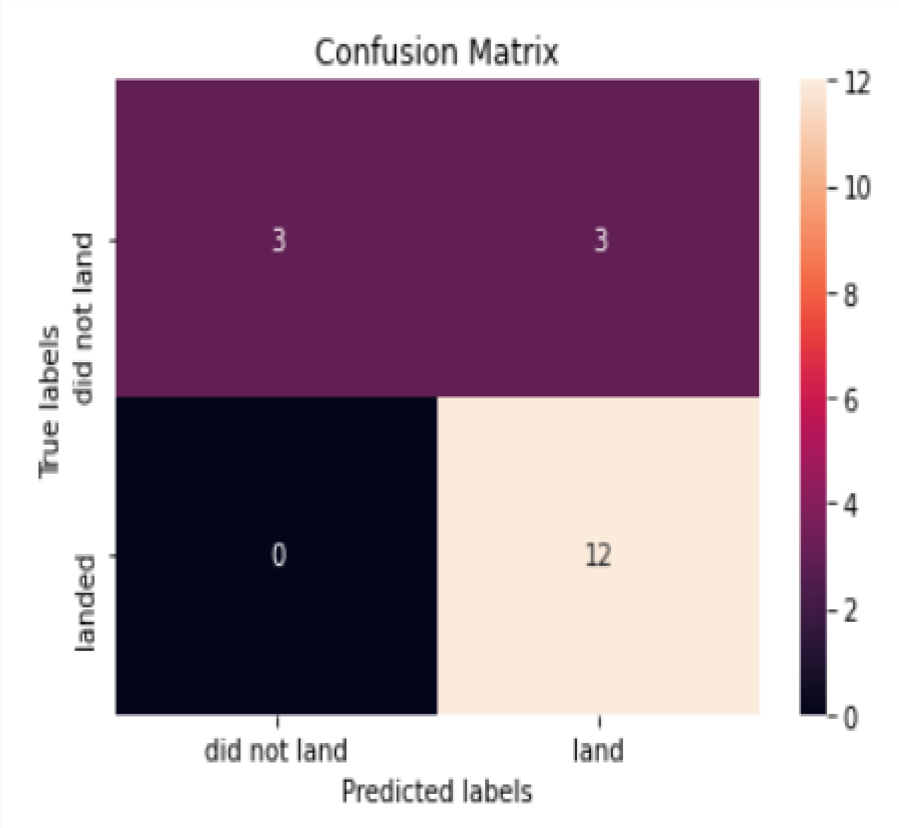
Classification Accuracy

- Decision Tree shows highest score with 94.44%
- Other models show score of 83.33%



Confusion Matrix

- Confusion Matrix shows
  - 12 success landing (True Positive)
  - 3 Failure landing (True Negative)
  - 3 marked as success but they are failure (False Positive)



Conclusions

- Success rate is increasing by year and it reached more than 80% in recent years
- For low payload mass ,SSO orbit has the highest success rate of 100%, for very high payload mass , VLEO orbit has very high success rate.
- Launch sites need to be far from cities , but close to sea and transportation logistics
- Decision tree model show highest score , other models are close behind , test data volume is low at 18 records only.

Thank you!