IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Khaula Khawer
Sep 17, 2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection (both through SpaceX API and Web scraping)
- Data wrangling (Data Cleansing)
- Exploratory data analysis (both through data visualization and SQL)
- Building an interactive map with Folium
- Building Dashboard with Plotly Dash
- Predicting whether first stage landing was successful or not (using various classification methods)

Summary of all results

- Data analysis unveils the patterns and trends of successful landing with respect to launch sites, payload masses and orbit types.
- Maps displays the launch sites location on the globe and their proximities.
- Dashboards provide insights regarding site-wise successful launches and its correlation with payload masses, by offering interactive view.
- Predictive analysis exhibits the effectiveness of the classification models.

Introduction

Project background and context

- This capstone project employs a real-world data with the aim to demonstrate expertise in data science and machine learning techniques and to summarize your findings in a report.
- In this project, we will predict if the Falcon 9 first stage will land successfully. 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. Therefore we can determine the cost of a launch if we can determine whether the first stage landing is successful.
- The report provide details into data collection, data wrangling, exploratory data analysis (EDA), interactive data visualization, ML classification model training and prediction, and model evaluation. Lastly, the accuracy of different ML algorithms are compared for predicting the future successful landing of the Falcon 9 first stage rocket.

Problems you want to find answers

- What are the leading trends of success rate of SpaceX launches and how they have evolved over time?
- How various launch sites and payload distribution contribute to overall success?
- How strategically the launch sites are positioned on the map and how different logistical considerations impact these locations?

Section 1

Methodology

Methodology

Summary

- Data Collection (both through SpaceX API and Web scraping)
- Data wrangling (Data Cleansing)
- Exploratory data analysis (both through data visualization and SQL)
- Building an interactive map with Folium
- Building Dashboard with Plotly Dash
- Predicting whether first stage landing was successful or not (using various classification methods)

Data Collection

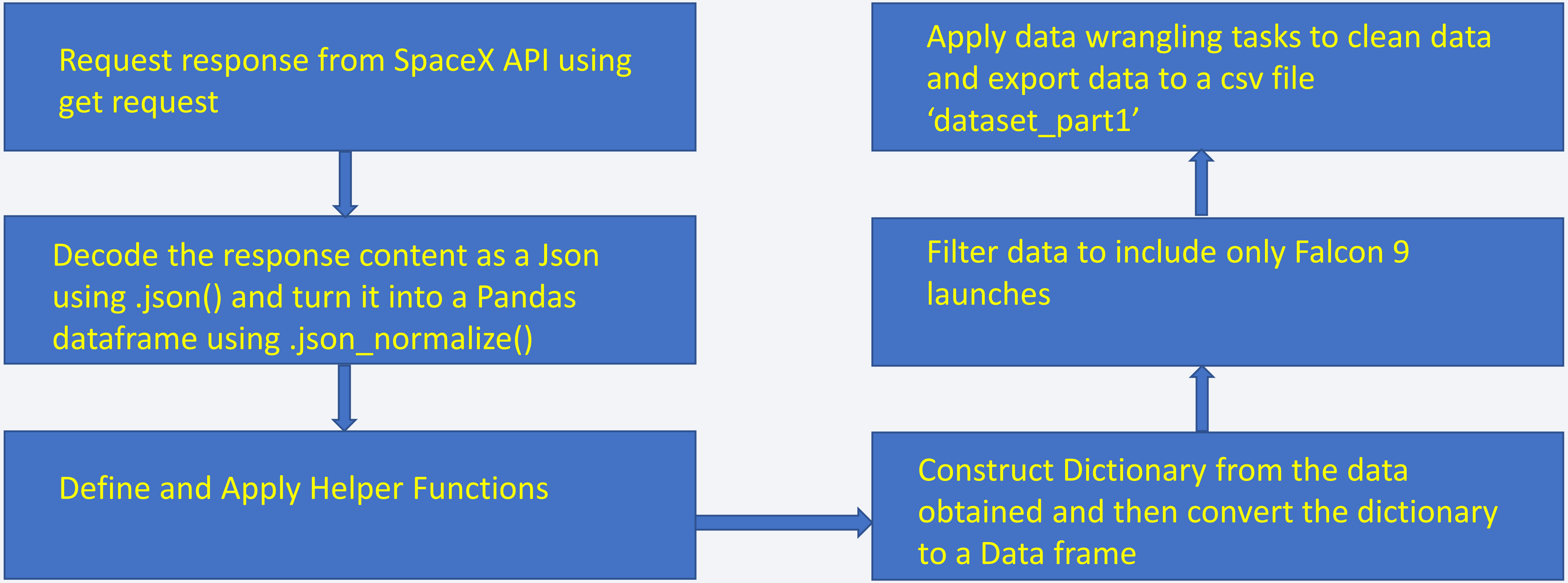
SpaceX API

- Request and get the real-time launch data from SpaceX API
- Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()

Web Scrapping

- Scrap historical data (from Wikipedia) to efficiently extract the structured data from the unstructured HTML text.
- Then filter the structured data to only include Falcon 9 data, assign data to data frame and dictionary, and export data to a csv file

Data Collection – SpaceX API



Data Collection – SpaceX API

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
response = requests.get(static_json_url)
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

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

We will now use the API again to get information about the launches using the IDs given for each launch. Specifically we will be using columns `rocket`, `payloads`, `launchpad`, and `cores`.

[15]:

```
# Lets take a subset of our dataframe keeping only the features we want and the flight number, and date_utc.
data = data[['rocket', 'payloads', 'launchpad', 'cores', 'flight_number', 'date_utc']]

# We will remove rows with multiple cores because those are falcon rockets with 2 extra rocket boosters and rows that have multiple payloads in a single rocket.
data = data[data['cores'].map(len)==1]
data = data[data['payloads'].map(len)==1]

# Since payloads and cores are lists of size 1 we will also extract the single value in the list and replace the feature.
data['cores'] = data['cores'].map(lambda x : x[0])
data['payloads'] = data['payloads'].map(lambda x : x[0])

# We also want to convert the date_utc to a datetime datatype and then extracting the date leaving the time
data['date'] = pd.to_datetime(data['date_utc']).dt.date

# Using the date we will restrict the dates of the launches
data = data[data['date'] <= datetime.date(2020, 11, 13)]
```

Data Collection – SpaceX API

Finally lets construct our dataset using the data we have obtained. We we combine the columns into a dictionary.

```
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}
```

[SpaceX API]

Then, we need to create a Pandas data frame from the dictionary launch_dict.

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

Task 2: Filter the dataframe to only include `Falcon 9` launches

Finally we will remove the Falcon 1 launches keeping only the Falcon 9 launches. Filter the data dataframe using the `BoosterVersion` column to only keep the Falcon 9 launches. Save the filtered data to a new dataframe called `data_falcon9`.

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data[data['BoosterVersion'] == 'Falcon 9']
```

Task 3: Dealing with Missing Values

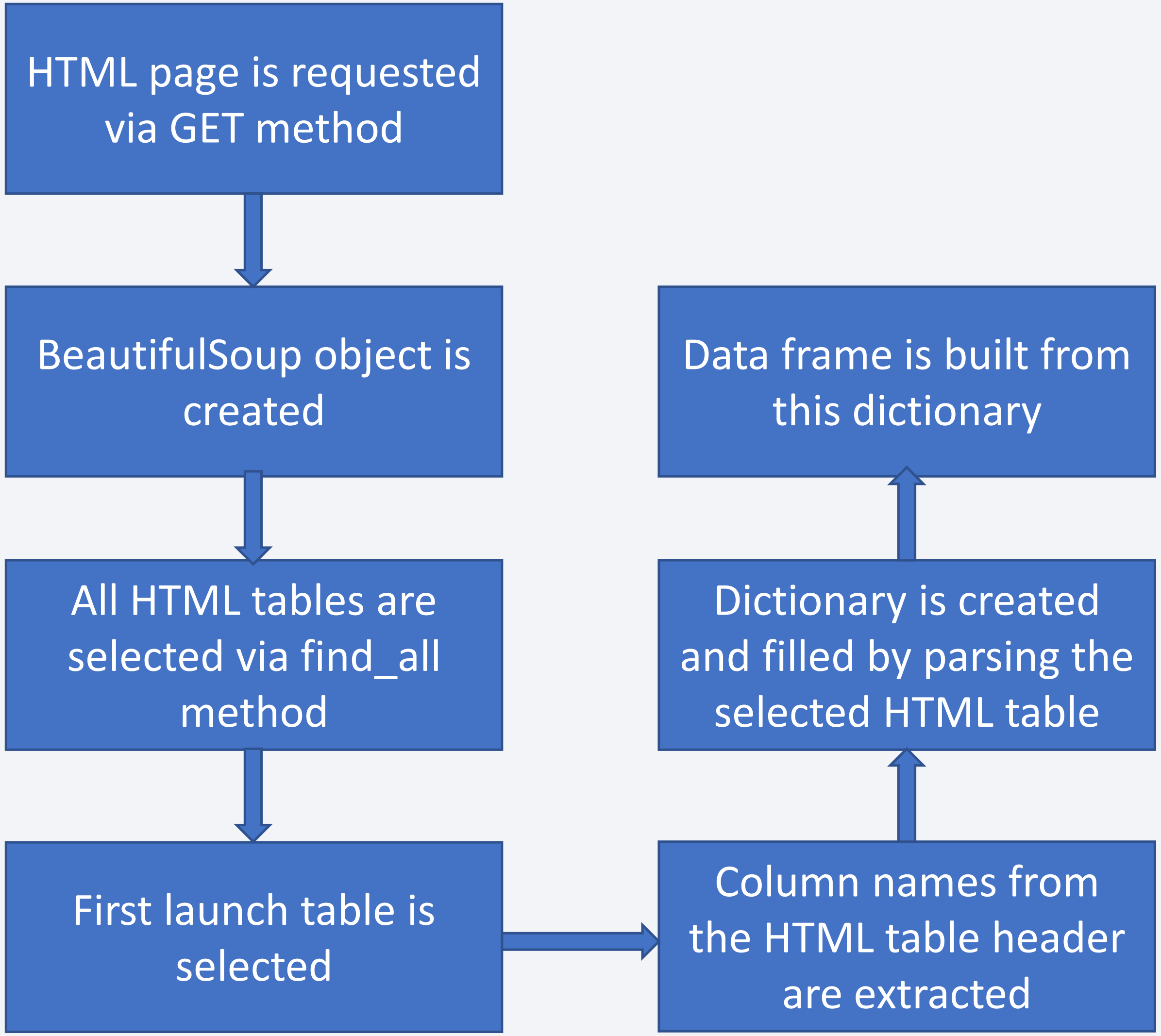
Calculate below the mean for the `PayloadMass` using the `.mean()`. Then use the mean and the `.replace()` function to replace `np.nan` values in the data with the mean you calculated.

```
# Calculate the mean value of PayloadMass column
mean_PayloadMass = data_falcon9['PayloadMass'].astype('float').mean()

# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'].replace(np.nan,mean_PayloadMass, inplace=True)
```

Data Collection - Scraping

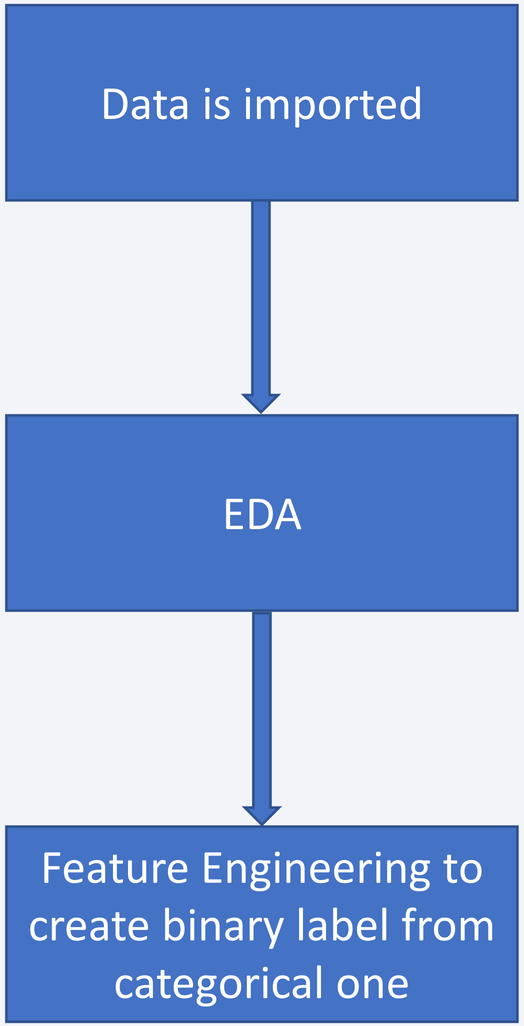
- Firstly, the Falcon9 Launch HTML page is requested via HTTP GET method, assign the HTTP response to a object.
- BeautifulSoup object is created from a response text content
- The required table is selected from HTML (Wiki) page and column names from the HTML table header are extracted
- Dictionary is created and filled by parsing the launch HTML tables and then a data frame is built from this dictionary.
- [Web Scrapping]



Data Wrangling

- Firstly, data is imported, data types of all attributes are identified and missing values for each attribute/column are calculated.
- Then data is analyzed in various aspects. The number of launches per site is calculated, number and occurrence of each orbit are determined, and categorized mission outcomes as successful or unsuccessful.
- Create a landing outcome label (from Outcome column) that represents the outcome of each launch as successful or unsuccessful.

- [Data Wrangling]()



EDA with Data Visualization

Cat Plots

- To check how the Flight Number (indicating the continuous launch attempts.) and Payload variables would affect the launch outcome

Bar Chart

- To visualize the relationship between success rate of each orbit type

Scatter Plot

- To visualize the relationship between Flight Number and Launch Site and between Payload Mass and Launch Site
- To determine whether there is any relationship between Flight Number and Orbit type and also between Payload Mass and Orbit type

Line Chart

- To visualize the launch success yearly trend
- EDA with Data Visualization

EDA with SQL

- The unique launch sites were displayed
- First five records where launch sites begin with the string 'CCA‘ were queried
- Total payload mass carried by boosters launched by NASA (CRS) was calculated
- Average payload mass carried by booster version F9 v1.1 was calculated and displayed
- The date for the first successful ground pad landing is listed
- Boosters with successful drone ship landings, for payload range between 4000 and 6000, are listed
- Successful and failure mission outcomes are counted
- Boosters that carried the maximum payload mass were retrieved
- Failed drone ship landings in 2015 were listed
- Total count of various landing outcomes, between the specified dates, are displayed in descending order

- EDA with SQL

Build an Interactive Map with Folium

- All launch sites on a map are marked
- Circle and Marker on each launch site has been added
- The success/failed launches for each site on the map are marked
- Marker is created at a closest city, railway, highway, coastline from the launch site
- The distances between a launch site to its proximities (highway, railroad, city, coastline) are measured to gain strategic insights
- A line between the marker to the launch site is drawn to demonstrate logistical factors

- Interactive Map

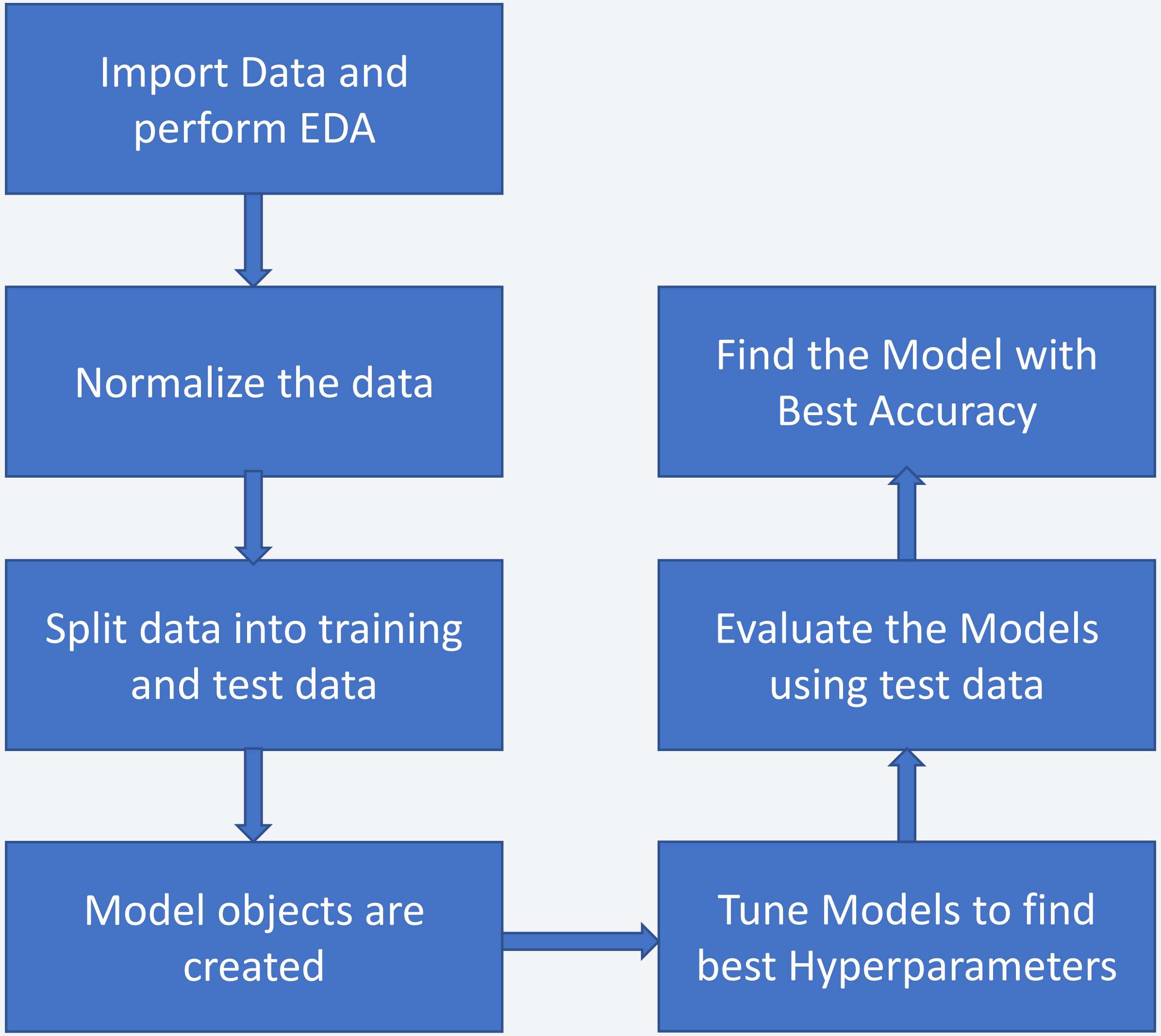
Build a Dashboard with Plotly Dash

- Dropdown menu is added to see which site has the largest success count. Also, to select a specific site for exploring its detailed success rate
- Pie chart is rendered to show the total success launches of all sites or a specific one
- Range Slider is added to select Payload ranges. The purpose is to discover whether the payload is correlated to mission outcome
- Scatter Plot is rendered for the selected payload range and outcomes for all sites or a specific site

- Dashboard

Predictive Analysis (Classification)

- EDA is performed to determine the training labels
- Data is normalized
- Normalized data is then split into training data and test data
- Objects of classification models including Logistic Regression, SVM, Decision Tree, and KNN, are created
- GridSearchCV objects are created and fit (using training data) to find the best Hyperparameters.
- Models are assessed using test data by generating confusion matrix and accuracy scores.
- Decision Tree performed the best, achieving approximately 88.75% accuracy.
- [ML Prediction]



Results

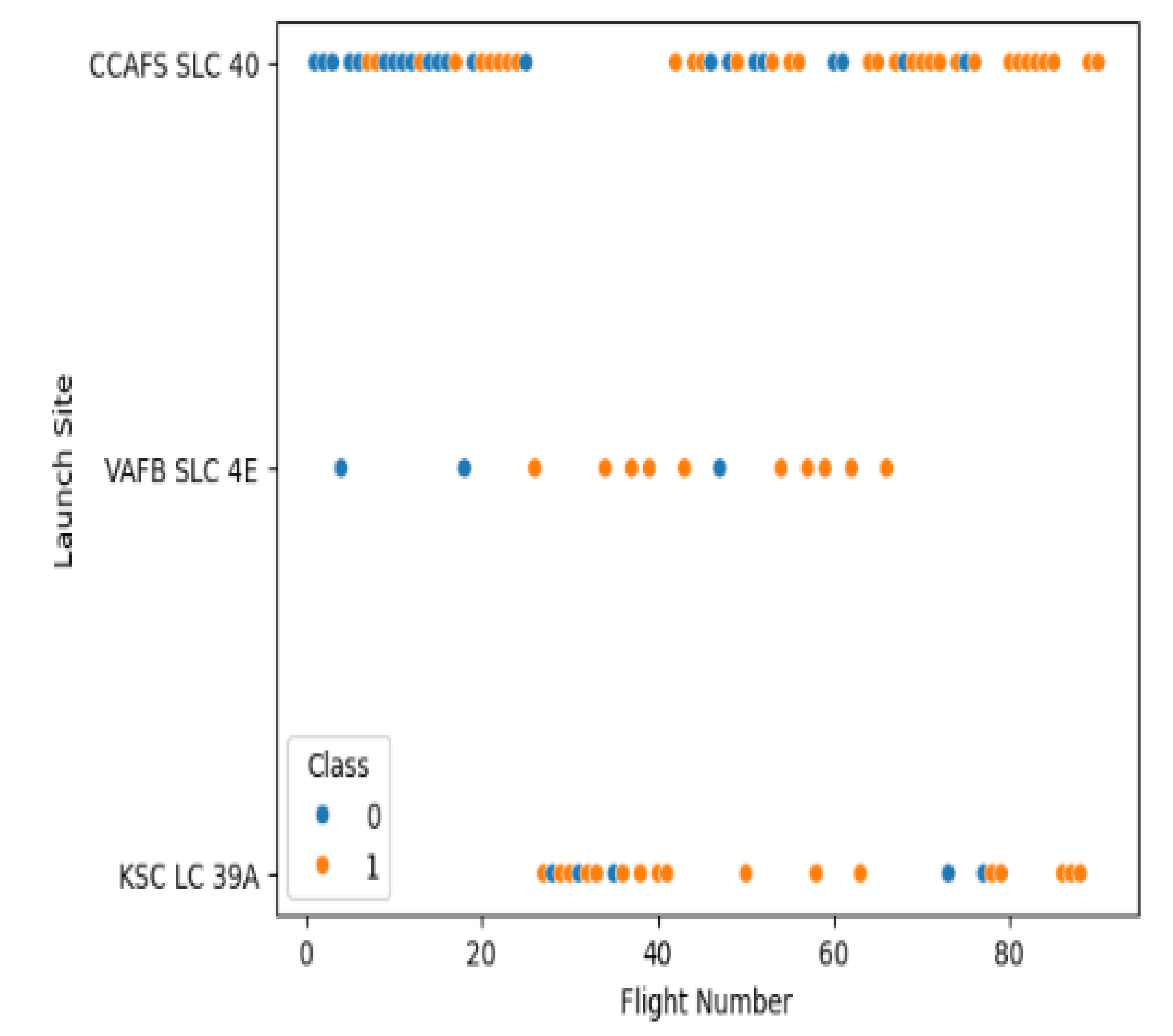
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

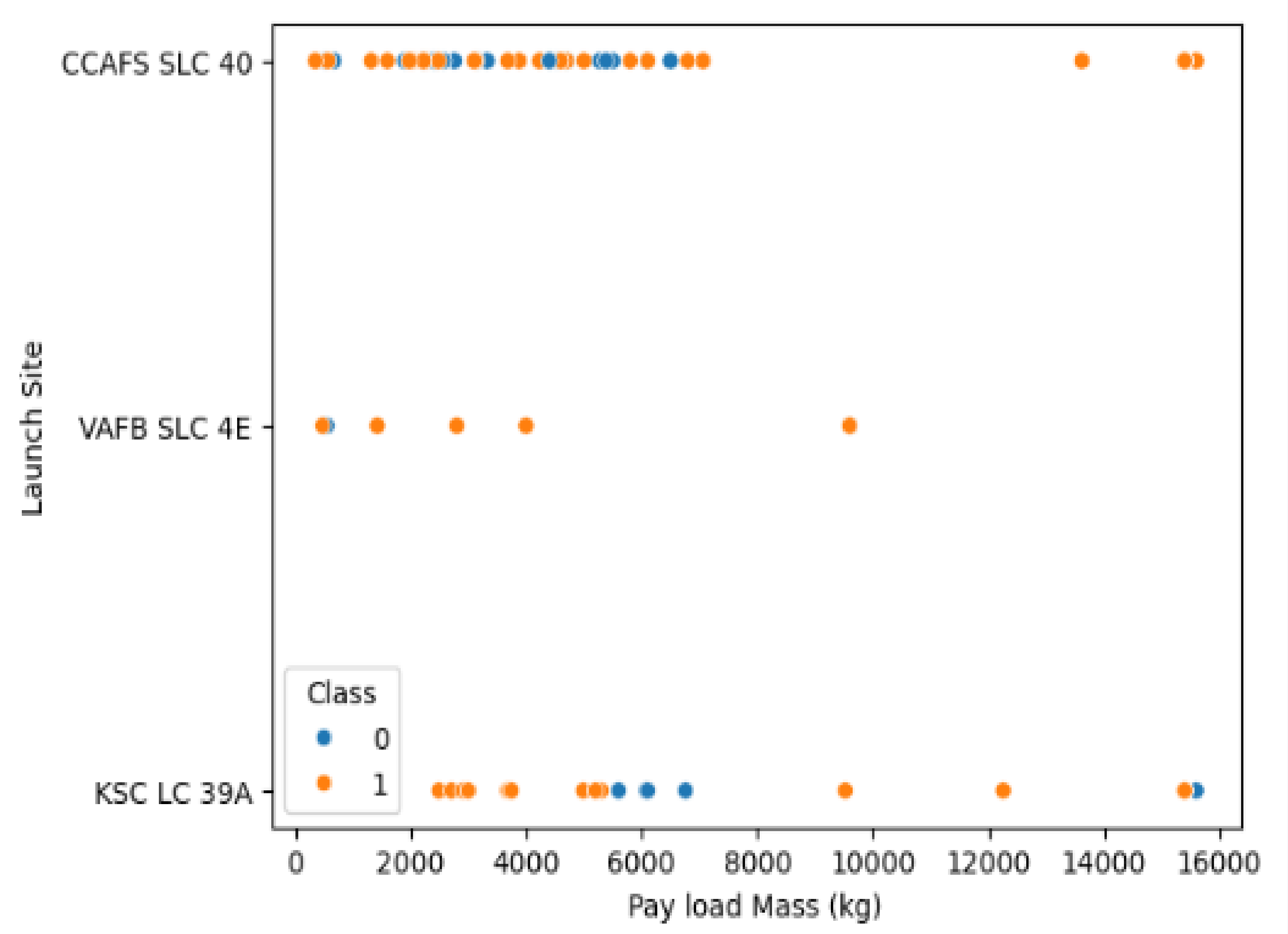
Flight Number vs. Launch Site

- There appears to be more successful landings with the increasing flight numbers, for all sites
- Outcome of launches do not depend upon launch sites as all sites have both outcomes (success and failure)
- VAFB SLC 4E has more Successful Launches in compared to failed ones
- CCAFS SLC 40 has the most number of launches



Payload vs. Launch Site

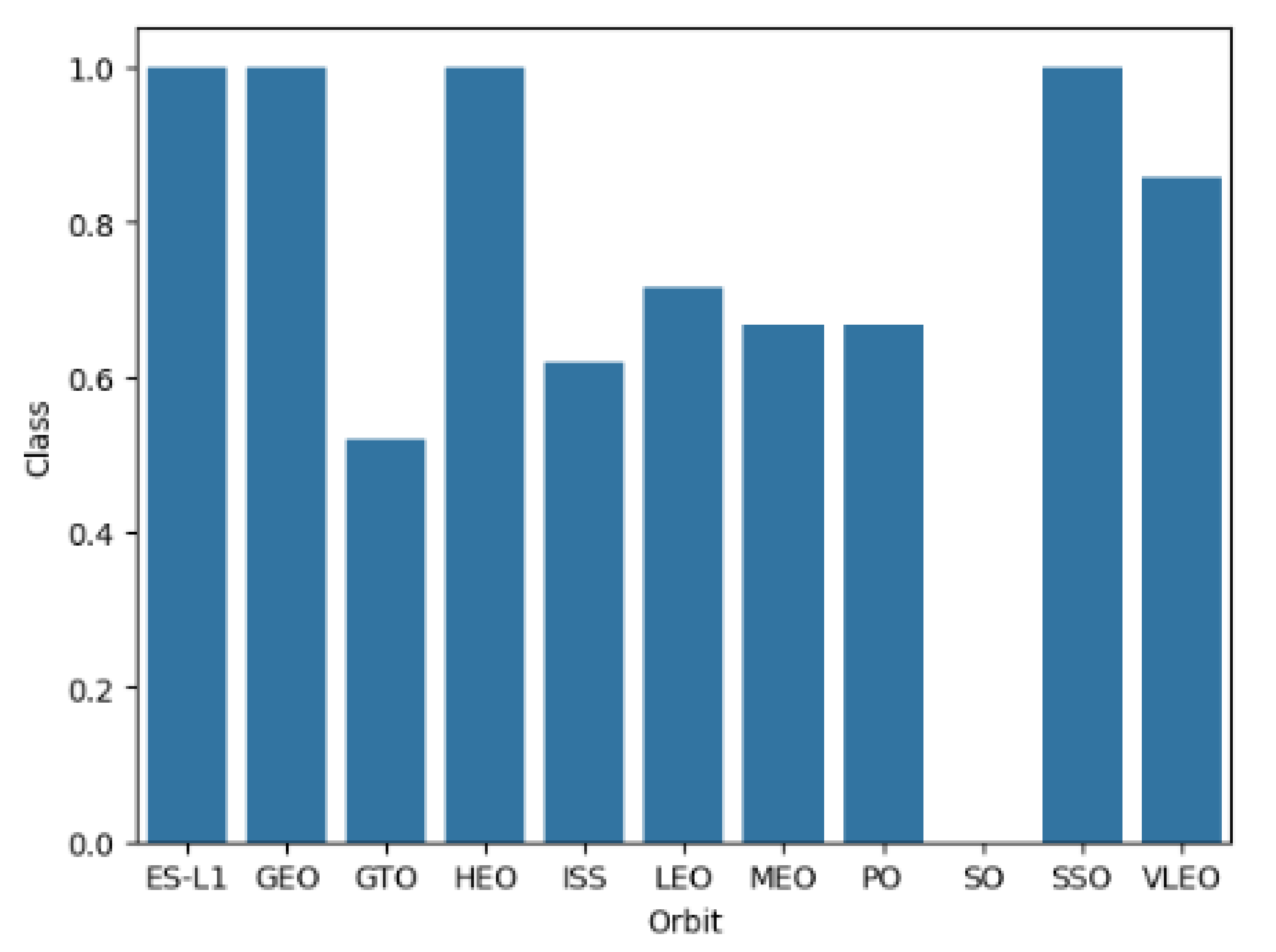
- For site ‘VAFB SLC 4E’, there is no rocket with the payload mass greater than 10,000 kg
- The number of launches decrease as the payload mass of rockets increases
- There seems to be more successful launches with the heavy payload mass



Success Rate vs. Orbit Type

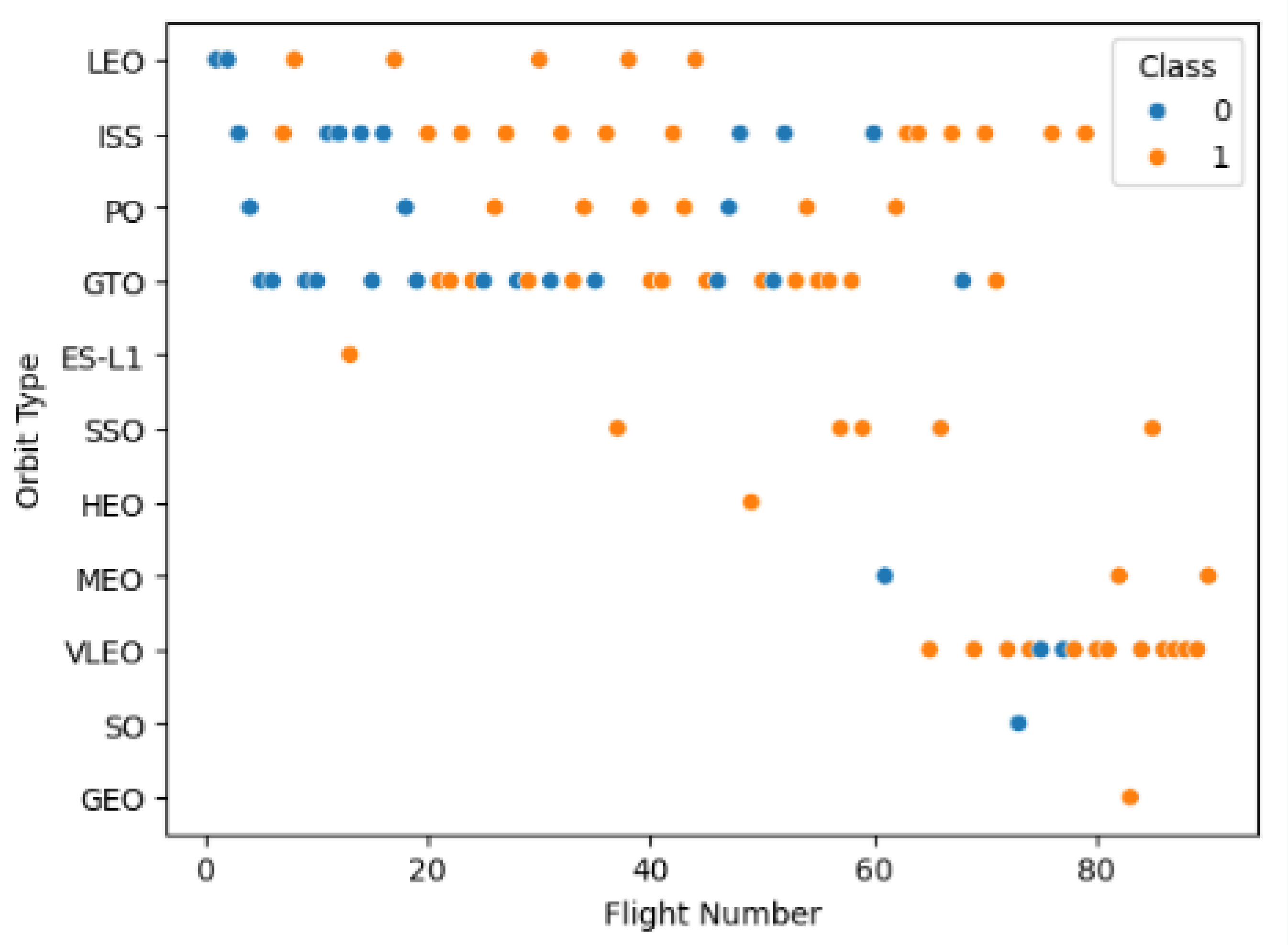
- ES-L1, GEO, HEO and SSO orbits have the highest success rates

- GTO orbit has the lowest success rate



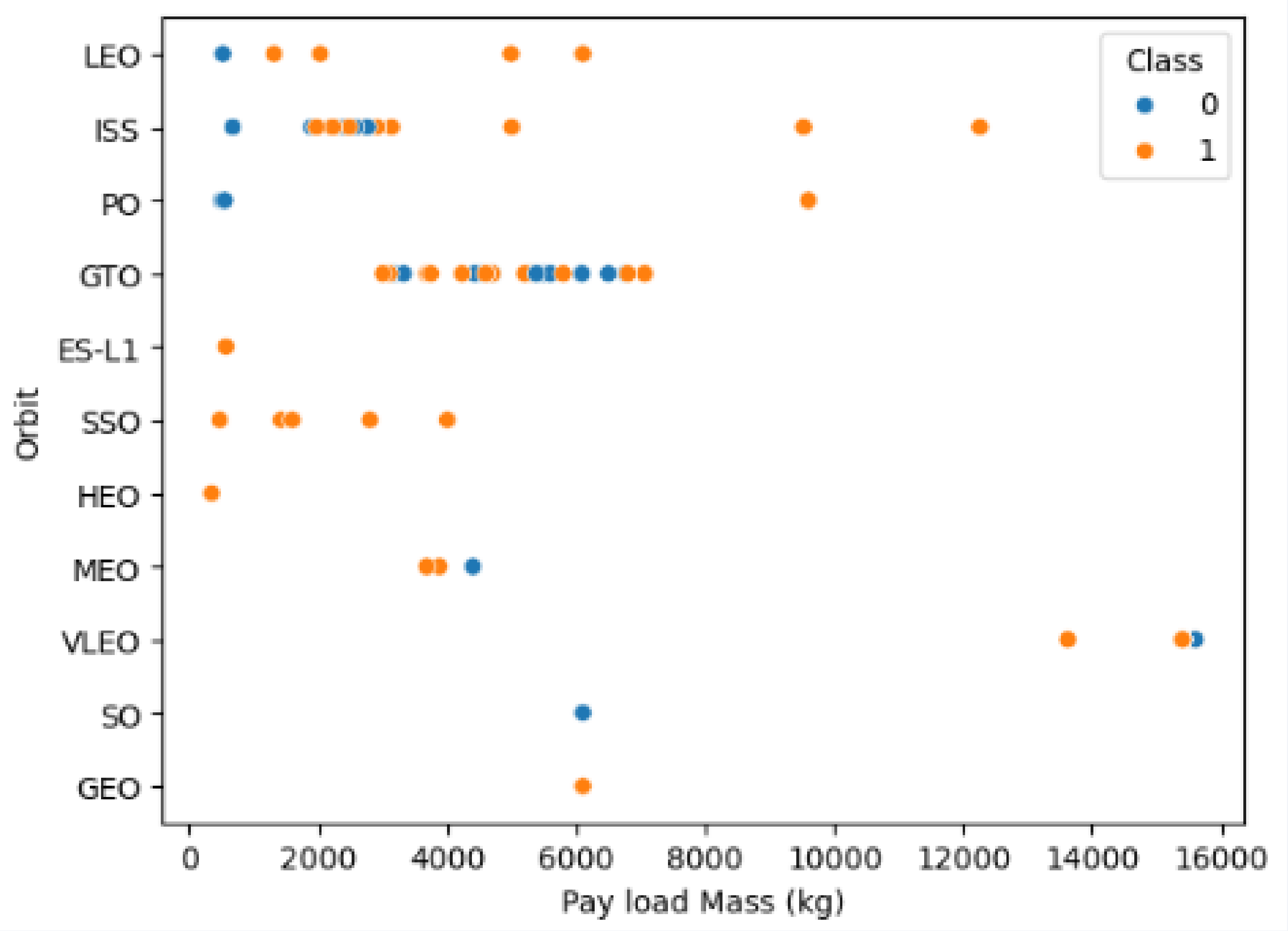
Flight Number vs. Orbit Type

- In LEO orbit, there seems to be more successful launches as the flight number increases
- In most of the orbits, the success rate increases by increasing the number of flights
- Whereas in GTO orbit, the success rate seems to have no relationship with the flight number



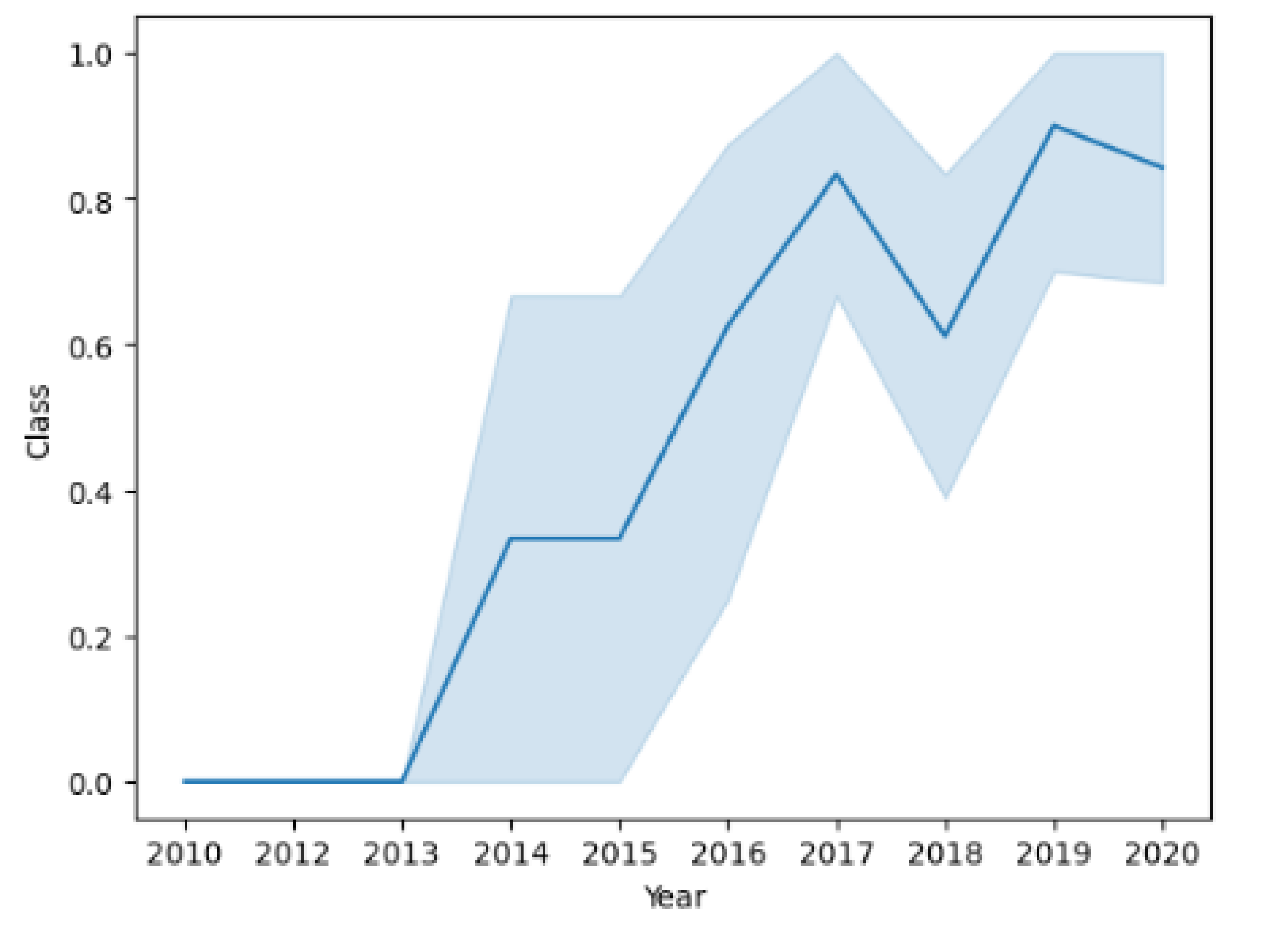
Payload vs. Orbit Type

- For LEO, ISS and PO orbits, success rate seems to be greater as the payload mass becomes heavy
- Whereas the success rate for VLEO orbit seems to have no relationship with the payload mass
- VLEO orbit have the rockets with the heaviest payload mass



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020
- Year 2019 has the highest success rate
- There are no successful launches till year 2013



All Launch Site Names

- Unique launch sites are selected from the column ‘Launch_Site’ by using the ‘**distinct**’ clause

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

```
%sql select distinct s."Launch_Site" from SPACEXTABLE s
```

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

Launch Site Names Begin with 'CCA'

- Substr() function is used on the column ‘Launch_Site’ to retrieve the launch sites whose first three letters are ‘CCA’

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
select * from SPACEXTABLE s
    where substr(s."Launch_Site",1,3) = 'CCA'
limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Sum function is applied on the column ‘PAYLOAD_MASS__KG_’ to get the total payload mass carried by the boosters when the ‘Customer’ is ‘NASA (CRS)’

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

```
%%sql
select round(sum(s."PAYLOAD_MASS__KG_")) Total_Payload_Mass from SPACEXTABLE s
    where s.Customer = 'NASA (CRS)'
```

Total_Payload_Mass
45596.0

Average Payload Mass by F9 v1.1

- Avg function is applied on the column ‘PAYLOAD_MASS__KG_’ to get the mean payload mass carried by the booster version ‘F9 v1.1’

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

```
%%sql
select round(avg(s."PAYLOAD_MASS__KG_")) Avg_Payload_Mass from SPACEXTABLE s
    where s."Booster_Version" = 'F9 v1.1'
```

Avg_Payload_Mass
2928.0

First Successful Ground Landing Date

- Min function is applied on the column ‘Date’ to get the date of first successful landing for the outcome ‘Success (ground pad)’

```
%%sql
select min(s.Date) First_succesful_landing_outcome_groundpad from SPACEXTABLE s
    where s."Landing_Outcome" = 'Success (ground pad)'
```

First_succesful_landing_outcome_groundpad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster versions have been retrieved by putting the conditions on the ‘Landing_Outcome’ to be successfully landed on drone ship and on the ‘PAYLOAD_MASS__KG_’ to be between 4000 and 6000

```
%%sql
select s."Booster_Version" from SPACEXTABLE s
    where s."Landing_Outcome" = 'Success (drone ship)'
    and s."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

- ‘Group by’ clause is used on the column ‘Mission_Outcome’ to get the total count of successful and failed mission outcomes

```
%%sql
select s."Mission_Outcome", count(*) Total_Count
    from SPACEXTABLE s
group by s."Mission_Outcome"
```

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

Boosters Carried Maximum Payload

- Subquery is used on the column 'PAYLOAD_MASS__KG_' to retrieve the boosters with the maximum payload mass

```
%%sql
select s."Booster_Version" from SPACEXTABLE s
    where s."PAYLOAD_MASS__KG_" = (select max(sp."PAYLOAD_MASS__KG_") from SPACEXTABLE sp)
```

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

- Only two launch records have been retrieved by selecting only the year 2015 (by using substr function on 'Date') and the 'Landing_Outcome' to be Failed drone ship landings

```
%%sql
select substr(s.Date, 6,2) Months, s."Landing_Outcome", s."Booster_Version", s."Launch_Site" from SPACEXTABLE s
    where s."Landing_Outcome" = 'Failure (drone ship)'
    and substr(s.Date, 1,4) = '2015'
```

Months	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- ‘Group by’ clause is used on the column ‘Landing_Outcome’ to get the total outcomes grouped by the different Landing Outcomes while selecting the specified data range.
- ‘Order by’ clause is used to rank the Landing Outcomes

```
%%sql
select s."Landing_Outcome", count(*) Total_Count
    from SPACEXTABLE s
    where s.Date between '2010-06-04' and '2017-03-20'
group by s."Landing_Outcome"
order by Total_Count desc
```

Landing_Outcome	Total_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Launch Sites Geographical Positions

- For each launch site, add both Circle and Marker objects based on its coordinates (Lat, Long) values
- The location markers on the global map would illustrate the geographical distribution of these sites, potentially highlighting geopolitical considerations for space launches.
- All sites are located on the coastal lines
- All sites are in proximity to the equator
- VAFB SLC-4E is located on the opposite end of the other two sites

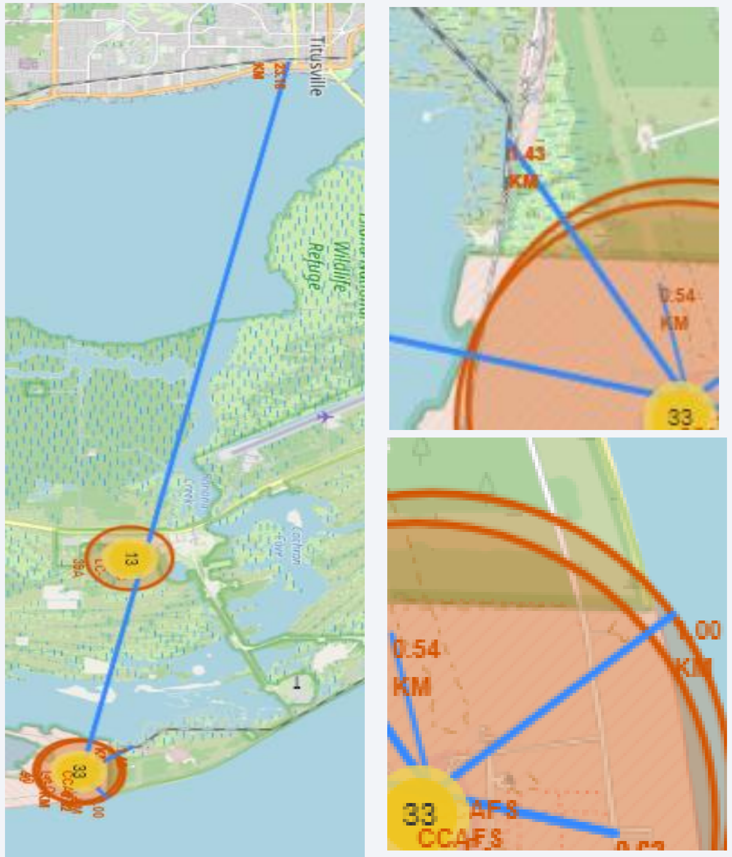
Marking Launch Outcomes for each Site on the Map

- Successful and failed launches of all sites are pinned on the site locations

- KSC LC-39A has the highest success rate

Analyzing the proximities of launch site

- Launch sites are in close proximity to coastline so they can fly over the ocean during launch, for safety reasons like mitigating the potential threat to people and property.
- Launch sites are in close proximity to highways, which allows the rapid transportation of the required people and property.
- Launch sites are in close proximity to railways that aids in easy transportation for heavy cargo.
- **Launch sites are not in close proximity to cities** that minimizes the risk to densely populated areas.

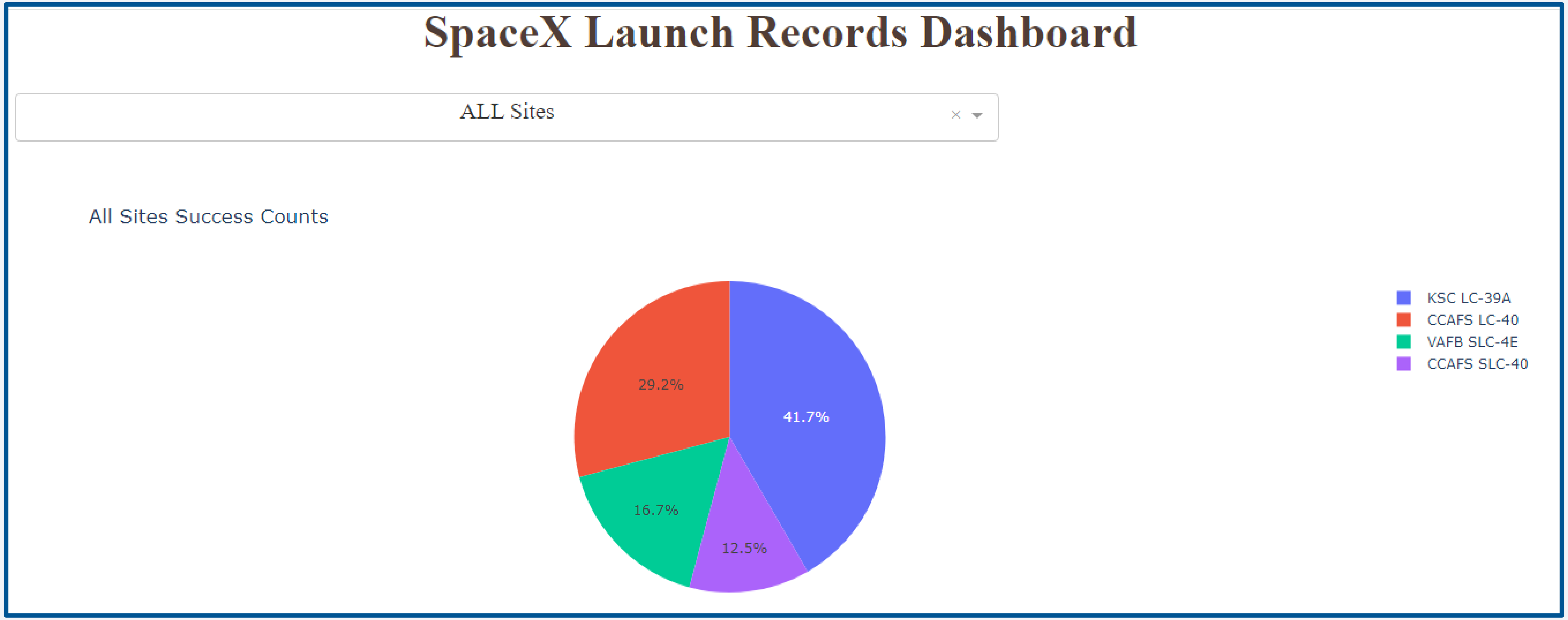


Section 4

Build a Dashboard with Plotly Dash

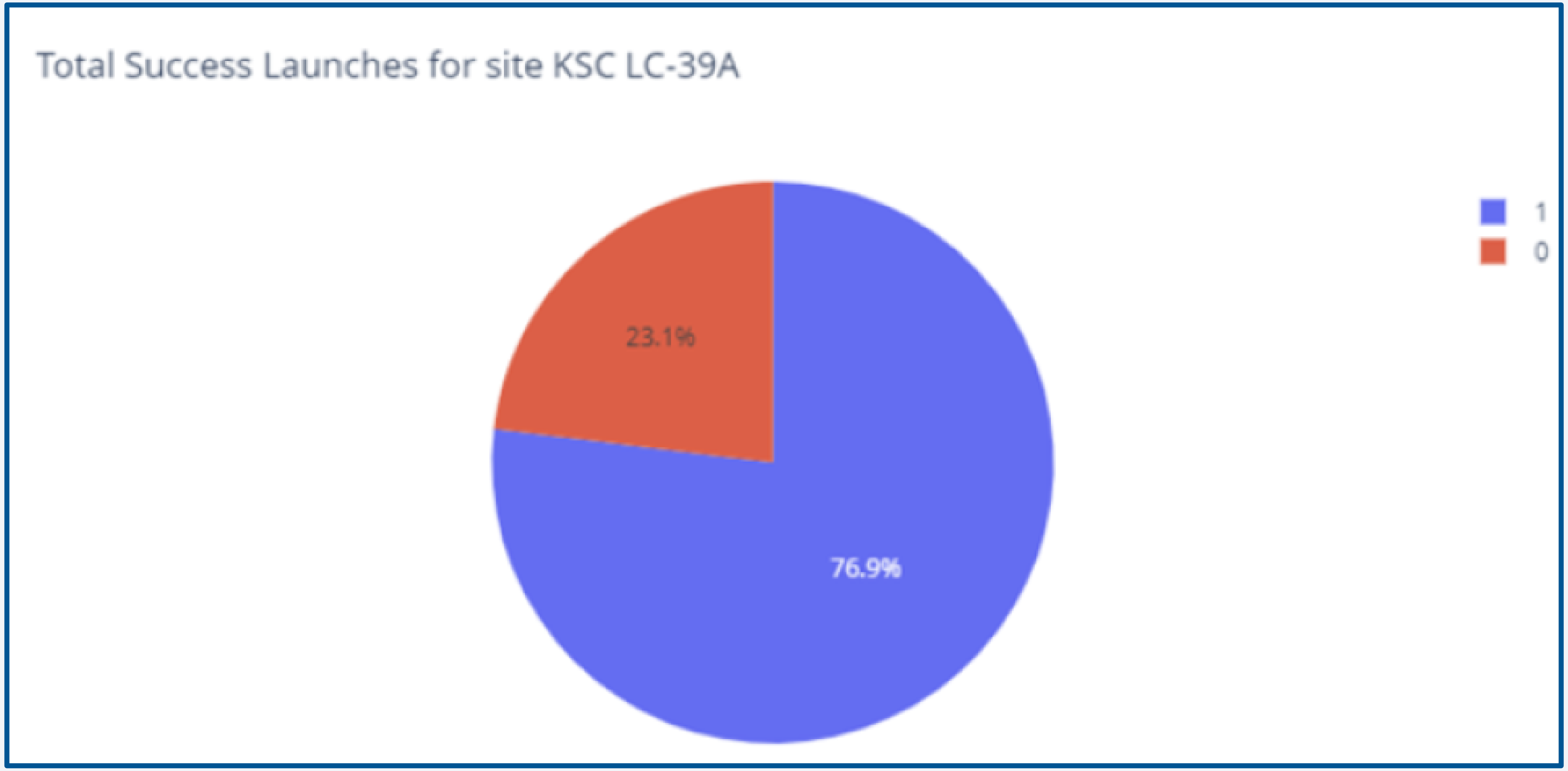
Launch Sites – Success Rate

- KSC LC-39A site has the highest launch success rate
- CCAFS SLC-40 has the lowest successful landings



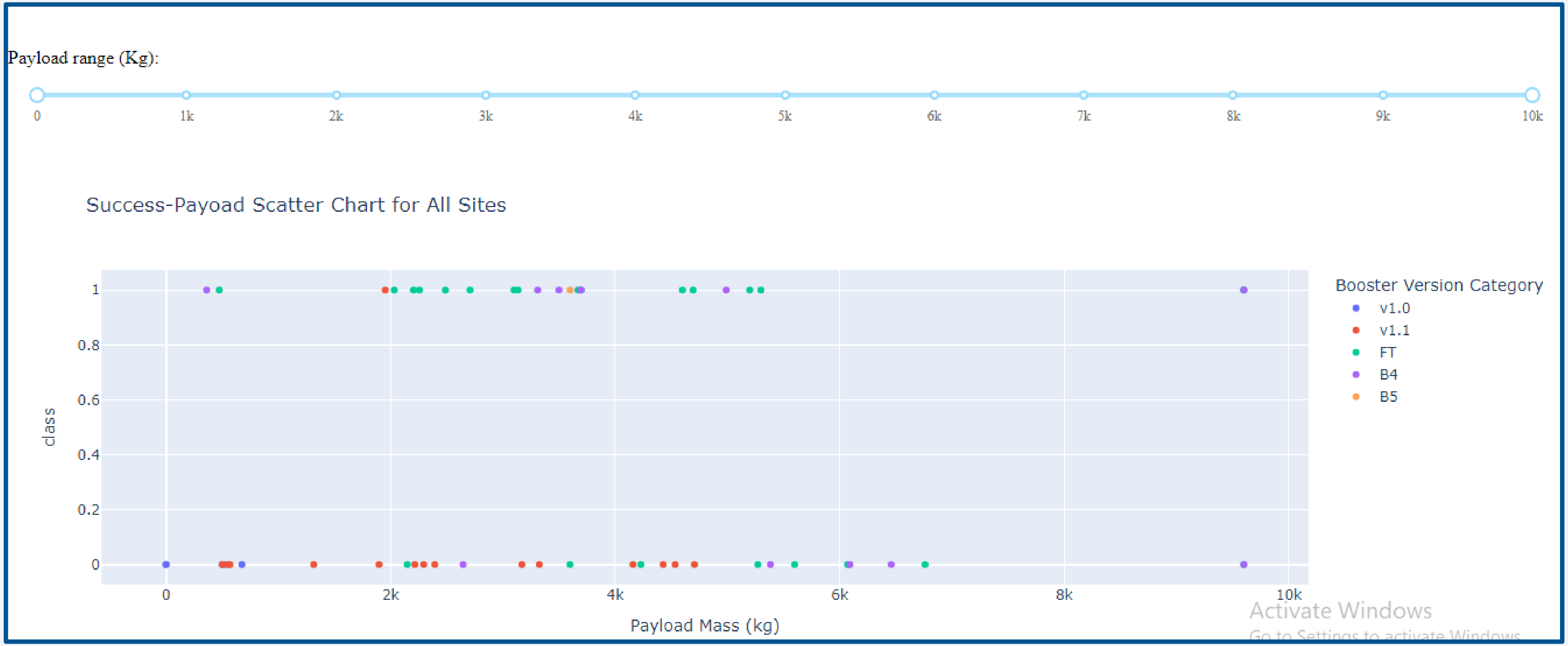
Specific Launch Site – Success Rate

- More than three quarters of landings are successful for the site having highest success rate i.e., ‘KSC LC-39A’



Launch Success Rate against Payload Ranges and Booster Versions

- Payload range of **2K-4K** has the highest launch success rate
- Payload range of **6K-8K** has the lowest launch success rate
- F9 Booster version **'FT'** has the highest launch success rate



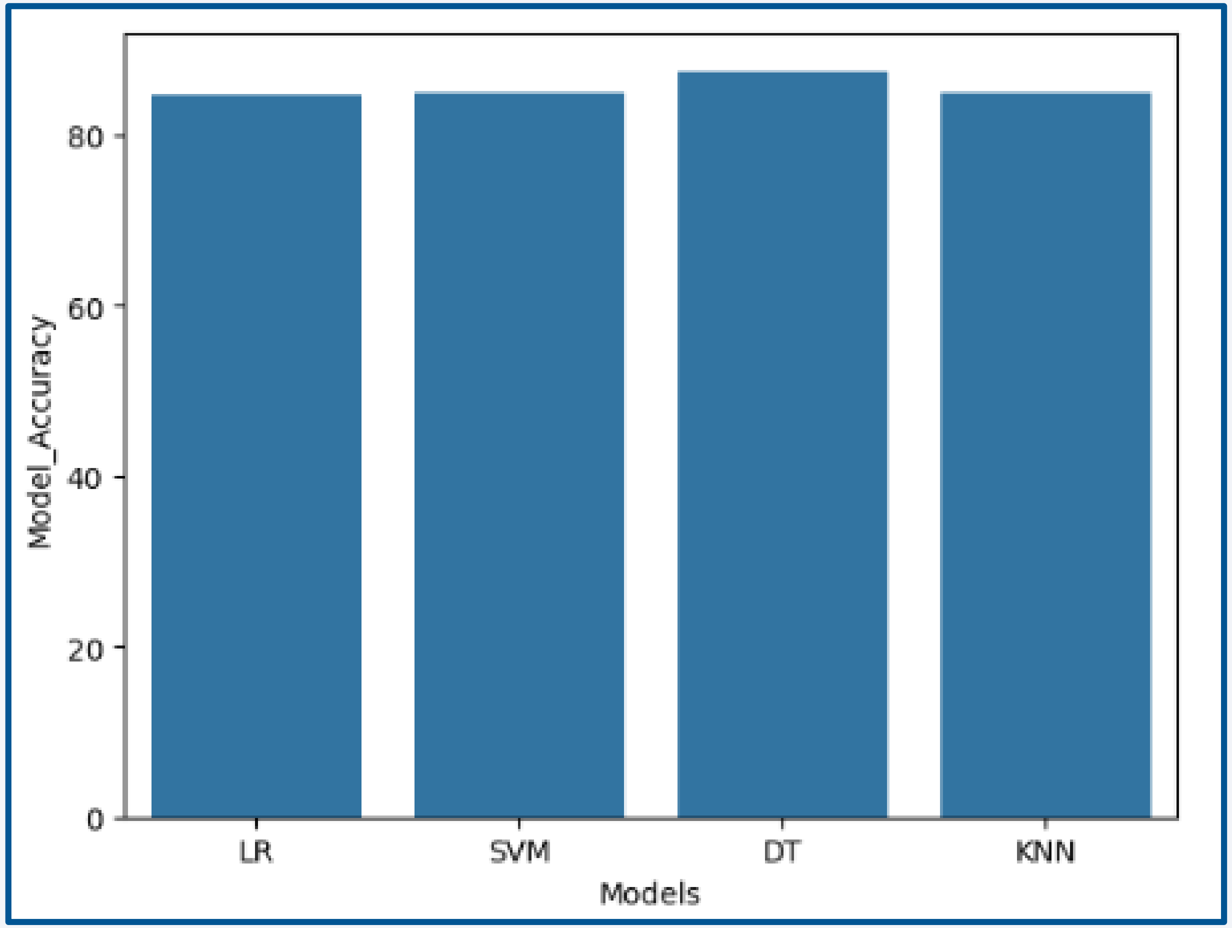
Section 5

Predictive Analysis (Classification)

Classification Accuracy

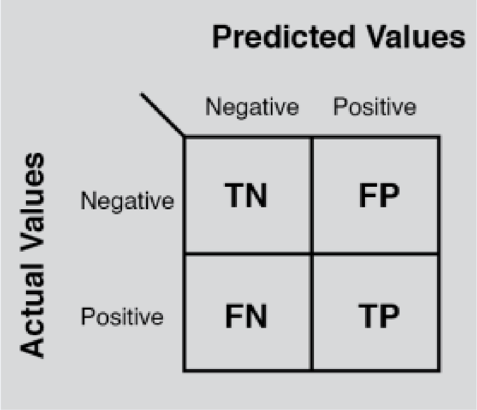
- All the Classifiers (LR, SVM, DT and KNN) displays high accuracy rate greater than 80%, hence ascertains that the models are reliable

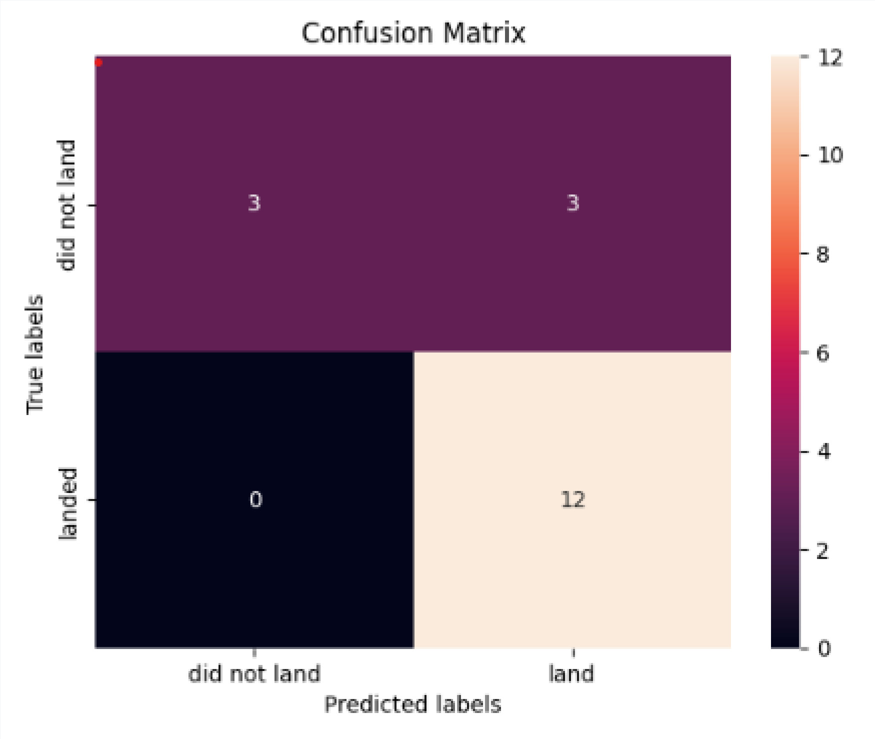
- Decision Tree exhibits the highest accuracy among the classifiers, achieving an accuracy of 87.50%.



Confusion Matrix

- The confusion matrix analysis suggests that all the Classifiers (LR, SVM, DT and KNN) perform the same and demonstrate the good prediction accuracy.
- The confusion matrix predicts 12 true positives, 3 false positives, 3 true negatives, and 0 false negative.





Conclusions

- Both historic and real-time data collection contributed towards better and deep insight into various aspects of the SpaceX launch records
- Interactive maps illustrated strategic positioning of launch sites
- Proximity analyses of launch sites emphasized logistical efficacy and safety.
- Dashboard offered dynamic view of various metrics affecting launch outcomes, hence enhanced user understanding.
- Classifiers (LR, SVM, DT and KNN) exhibited high prediction accuracy (>80%) that made the project weighty for the data science industry
- The valuable insights gained from the project would help for an alternate company to bid against the SpaceX in a more profound manner
- The valued findings would contribute to the space exploration field as well as towards the data science community.

Appendix

- GitHub Repository
- SpaceX API
- Web Scrapping
- Data Wrangling
- EDA – DV
- EDA – SQL
- Site Locations
- ML Prediction

Thank you!