IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rehan Zafar
30-DEC-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling / Cleaning
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Building Interactive map with folium
  - Building Interactive Dashboard with Plotly Dash
  - Predictive Analysis
- Summary of all results
  - Results of Exploratory Data Analysis
  - Interactive Analysis Results
  - Predictive Analysis Results

Introduction

- Project background and context
  - SpaceX advertises its 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 could reuse the first stage.
- Problems you want to find answers
  - The project task is predicting if the first stage of SpaceX Falcon 9 rocket will land successfully

Section 1

Methodology

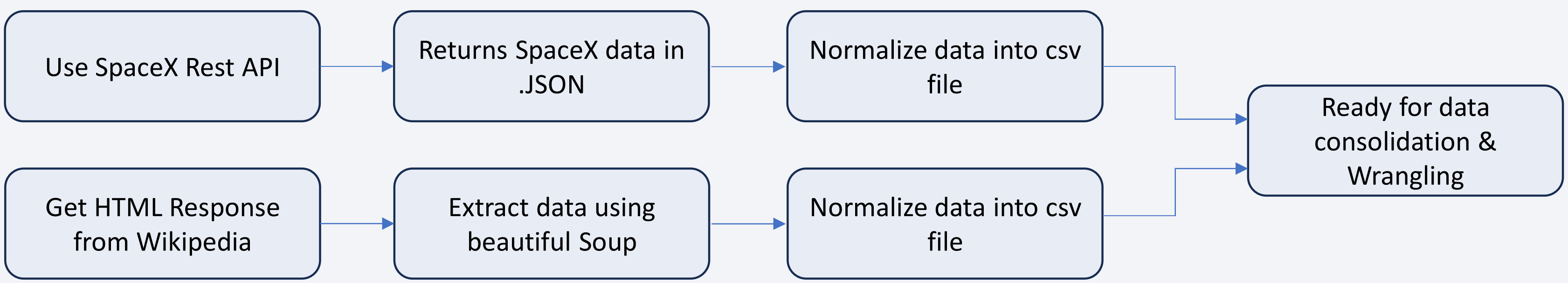
Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - One Hot encoding data fields for Machine Learning and data cleaning of null values and 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
  - Like Logistic Regression, K Nearest Neighbor, SVM and Decision Tree

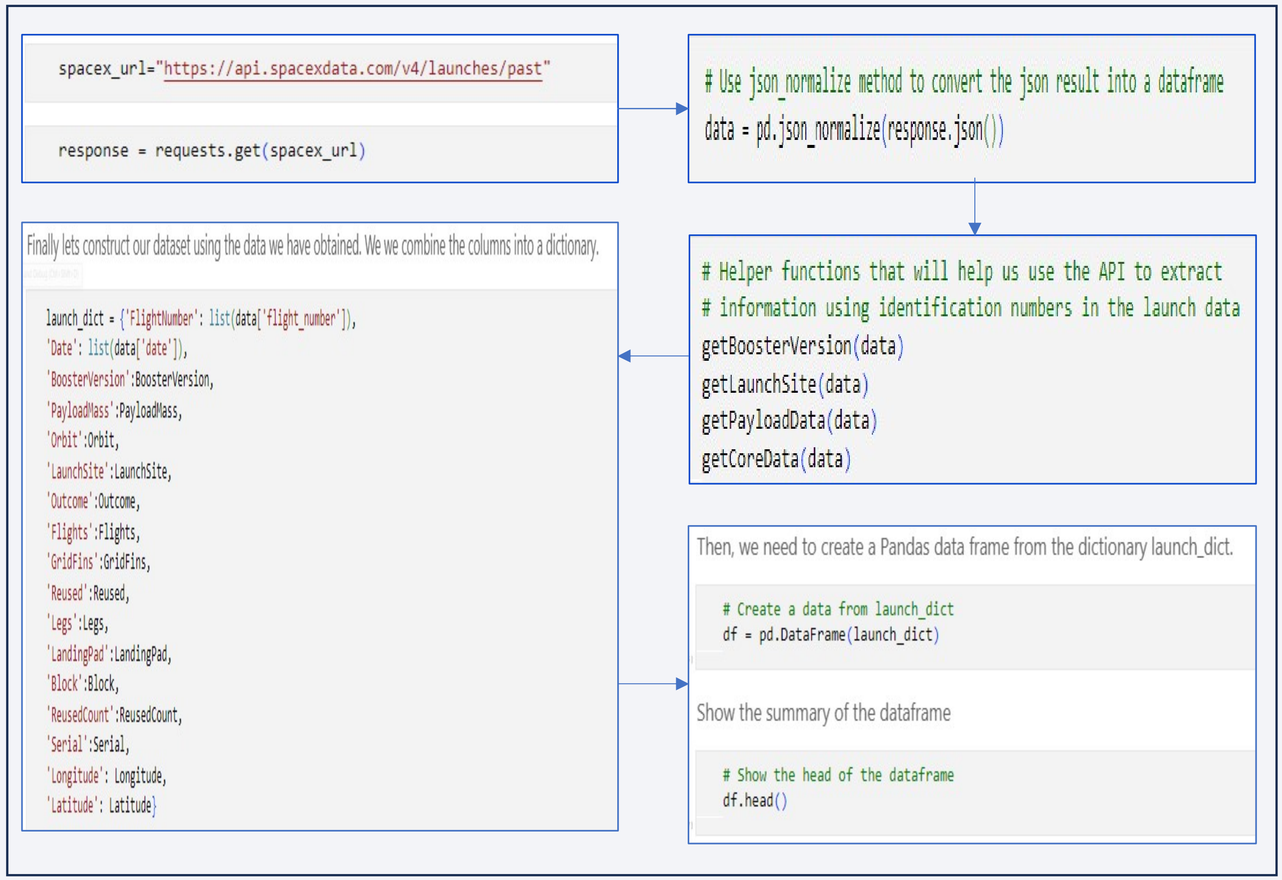
Data Collection

- Data was collected through two approaches:
  - SpaceX launch data was gathered from SpaceX Rest API
  - Also additional Falcon 9 launch data was collected through Web scraping from Wikipedia using Beautiful Soup
- Following is the flow of data collection process:



Data Collection – SpaceX API

- Flow chart of data collection with SpaceX REST API calls using key phrases



- GitHub URL of the completed SpaceX API calls notebook

https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- web scraping from Wikipedia

- https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-webscraping.ipynb

1. Request the Falcon9 Launch Wiki page from its URL

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

2. Create a `BeautifulSoup` object from the HTML `response`

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

3. Extract all column/variable names from the HTML table header

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

4. Create a data frame by parsing the launch HTML tables

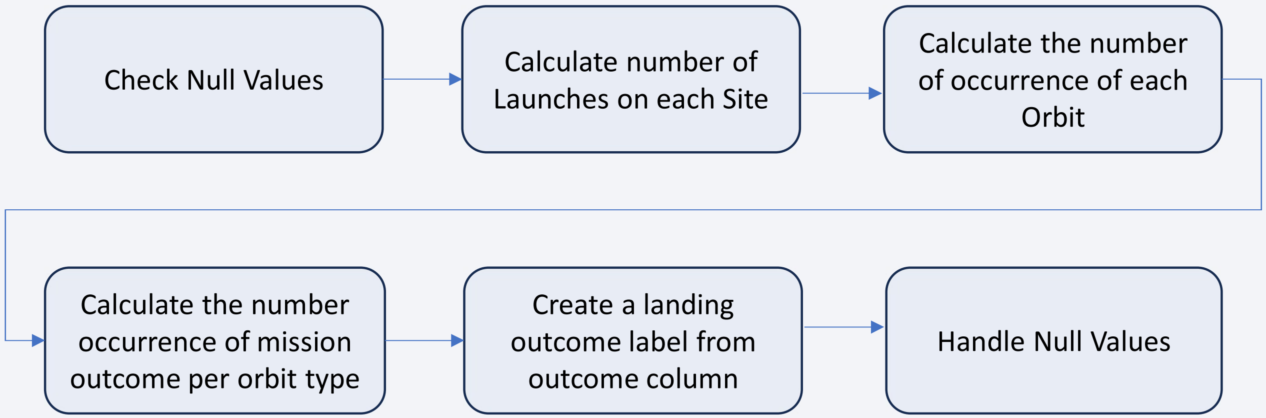
```
launch_dict= dict.fromkeys(column_names)

# 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']=[]
```

Data Wrangling

- EDA Analysis



- https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Task1: Visualize the relationship between Flight Number and Launch Site by plotting scatter point chart
- Task2: Visualize the relationship between Payload and Launch Site to explain the patterns we found in the Flight Number vs. Launch Site scatter point plots
- Task3: Created a bar char to visualize the relationship between success rate of each orbit type to check if there are any relationship between success rate and orbit type
- Task4: Plot a scatter point chart to see if there is any relationship between FlightNumber and Orbit type.
- Task5: Similarly, plotted Payload vs. Orbit scatter point chart to reveal the relationship between Payload and Orbit type
- Task6: Plotted a line chart with x axis to be Year and y axis to be average success rate, to get the average launch success trend.

https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL Queries performed include:

- **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 succesful landing outcome in ground pad was acheived.**
- **List the names of the boosters which have success in drone with 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, launch_site for the months 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.**

https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Created and added folium.Circle and folium.Marker for each launch site on the folium site map
- Created markers for all launch records and adding to Marker Cluster due to many launch coordinates on similar place.
- Added a Mouse Position on the map to get coordinate for a mouse over a point on the map.
- Draw a PolyLine between a launch site to the selected coastline point

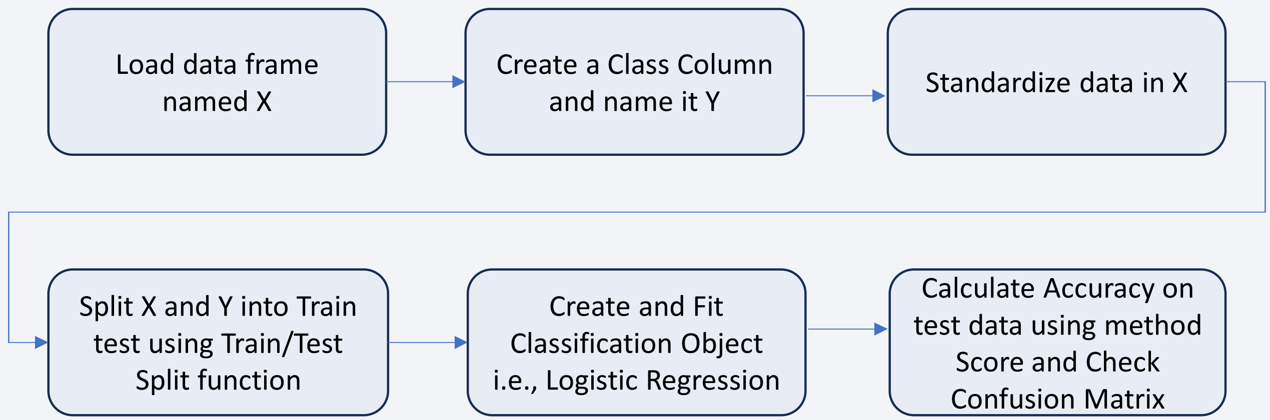
https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/DV0101EN-3-5-1-Generating-Maps-in-Python-py-v2.0.ipynb

Build a Dashboard with Plotly Dash

- TASK 1: Add a dropdown list to enable Launch Site selection in order to show success/failure status of launch for specific sute
- TASK 2: Add a pie chart to show the total successful launches count for all sites. If a specific launch site was selected, show the Success vs. Failed counts for the site
- TASK 3: Add a slider to select payload range to show success/failure scatter plot for specific payload range
- TASK 4: Add a scatter chart to show the correlation between payload and launch success
- https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

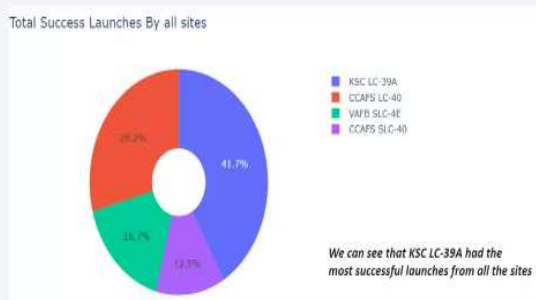
- We created a machine learning pipeline to predict if the first stage will land given the data from the preceding labs.
- Performed exploratory Data Analysis and determine Training Labels: 1) creating a column for the class 2) Standardize the data 3) Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data

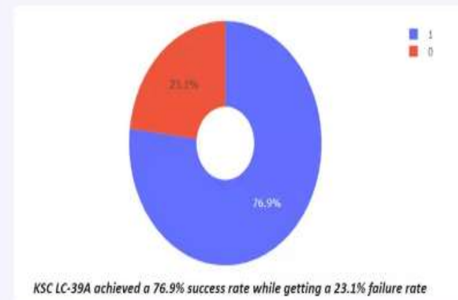


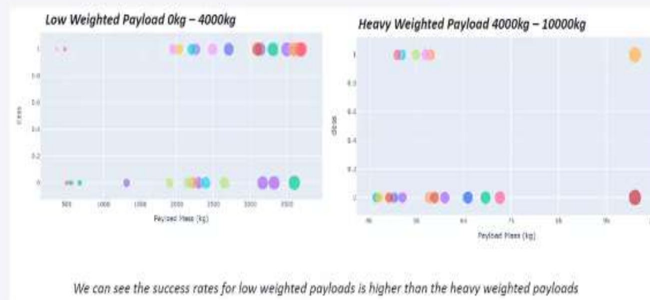
- https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results







- Predictive analysis results

Find the method performs best:

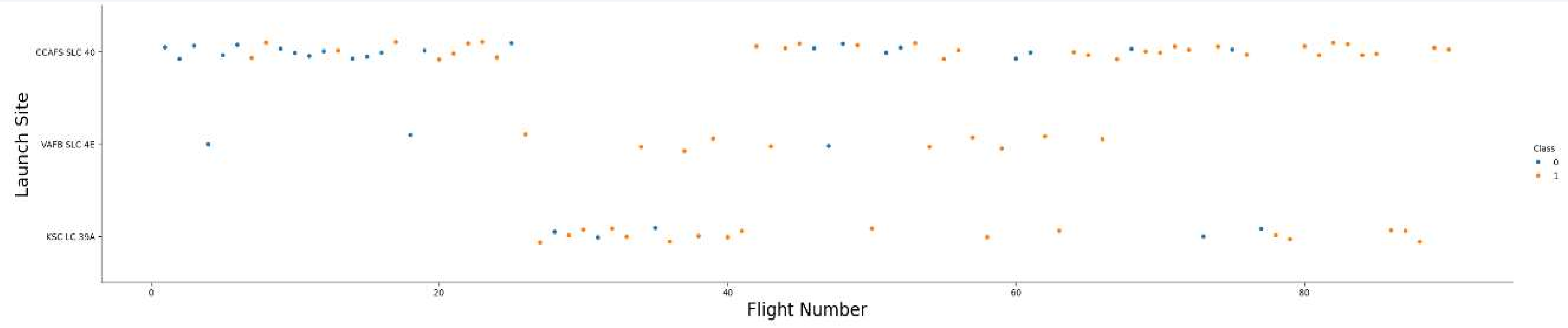
```
[46]: print(logreg_cv.score(X_test, Y_test))
      print(svm_cv.score(X_test, Y_test))
      print(tree_cv.score(X_test, Y_test))
      print(knn_cv.score(X_test, Y_test))

0.8333333333333334
0.8333333333333334
0.8333333333333334
0.8333333333333334
```

Section 2

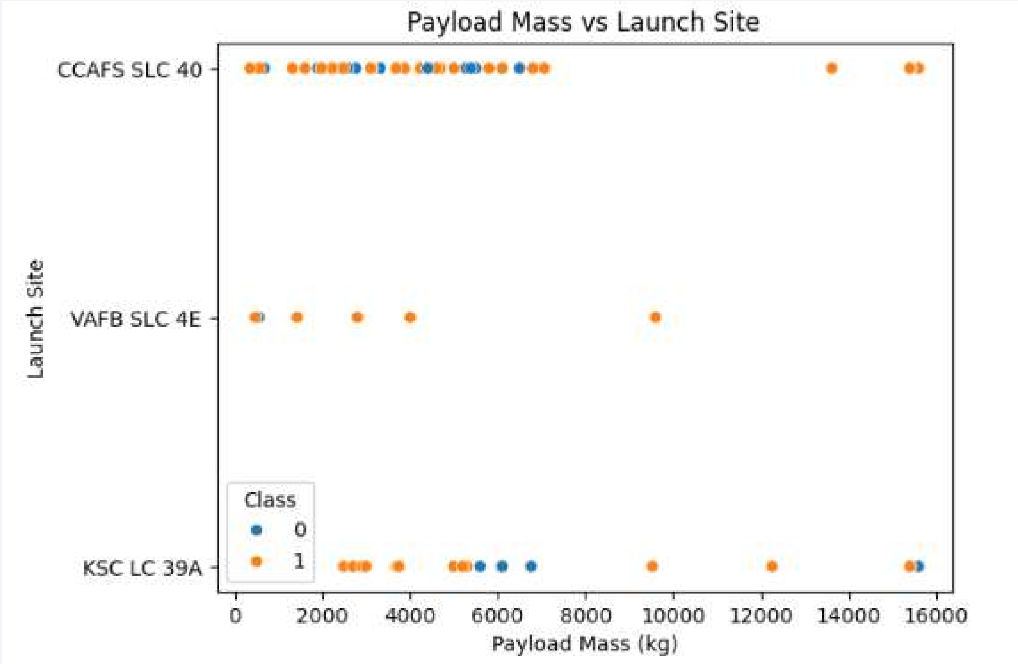
Insights drawn from EDA

Flight Number vs. Launch Site



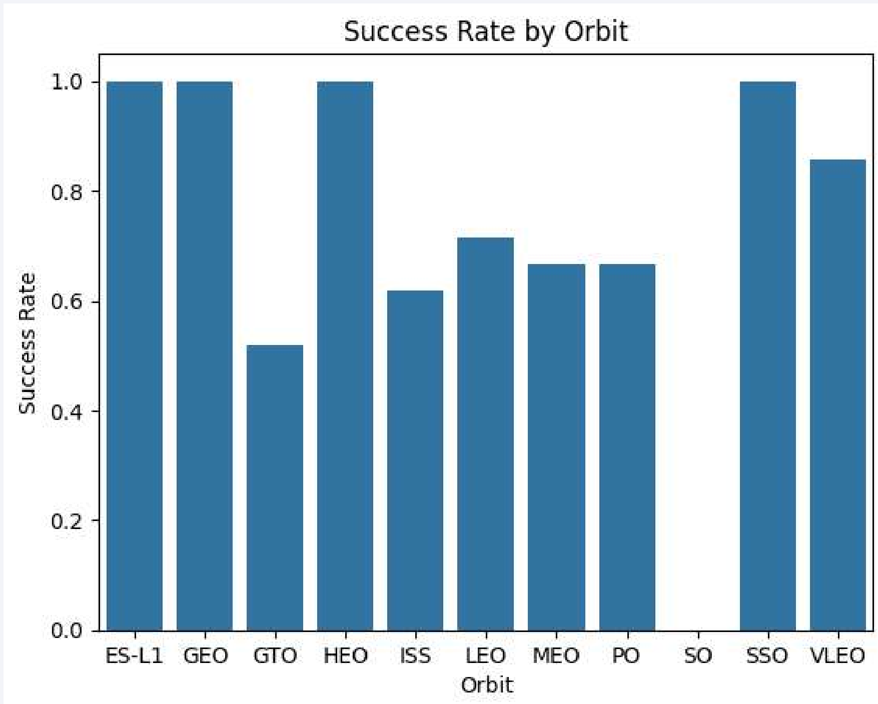
The above screen shot is a chart to visualize the relationship between Flight Number and Launch Site to check if a particular launch has specific success/failure impact

Payload vs. Launch Site



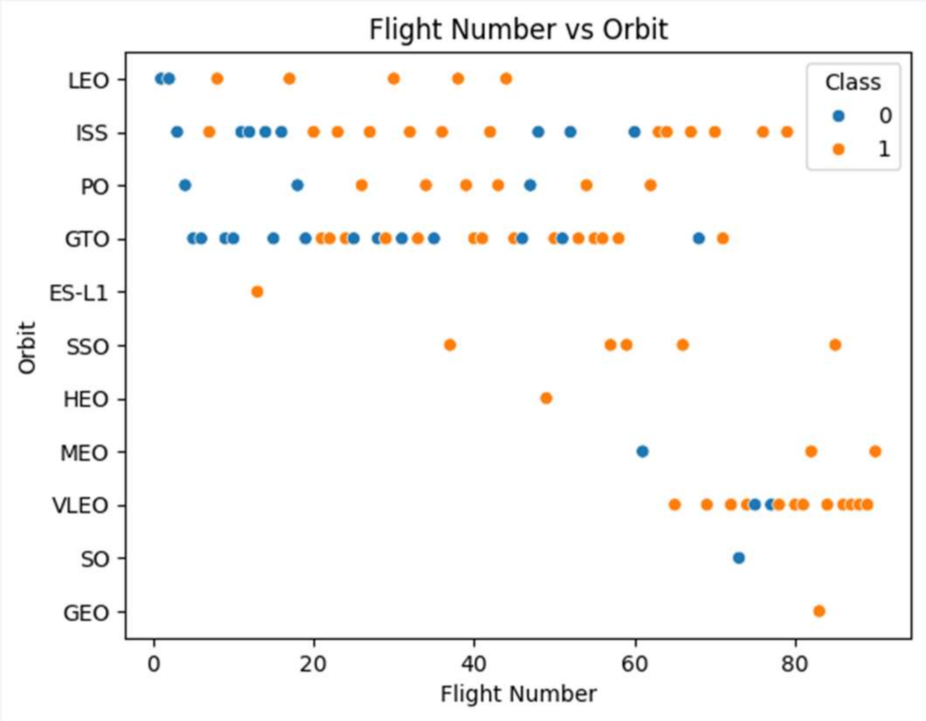
- The above chart is to observe if there is any relationship between launch sites and their payload mass.
- We can observe that for the VAFB-SLC launchsite there are no rockets launched for heavy pay load mass(greater than 10000).

Success Rate vs. Orbit Type



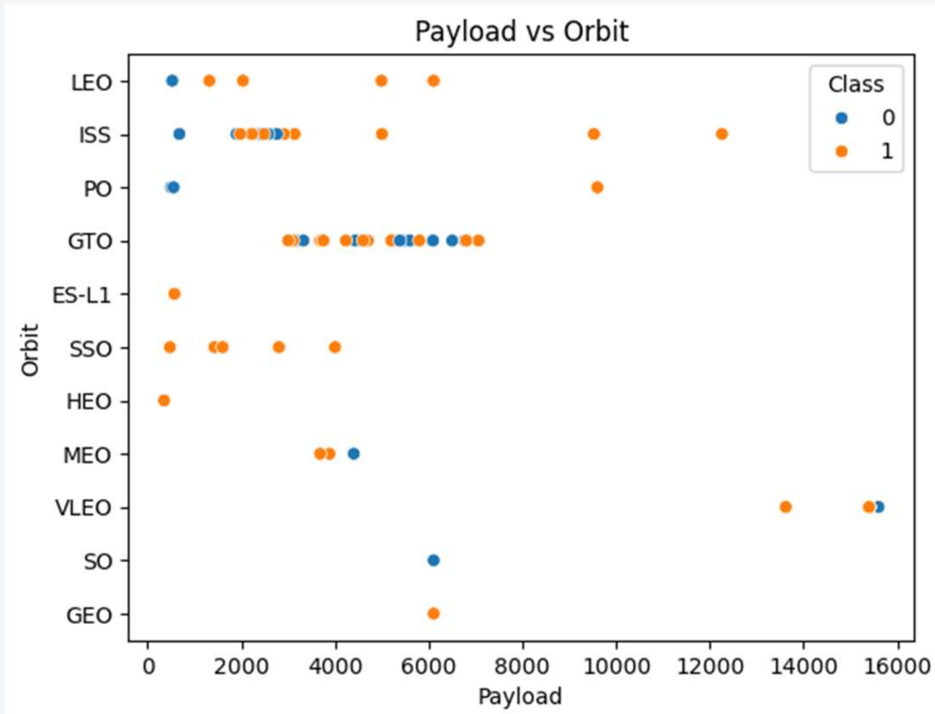
- The above bar chart is to visually check if there is any relationship between success rate and orbit type. As we can see ES-L1, GEO, HEO and SSO orbits have high success rates.

Flight Number vs. Orbit Type



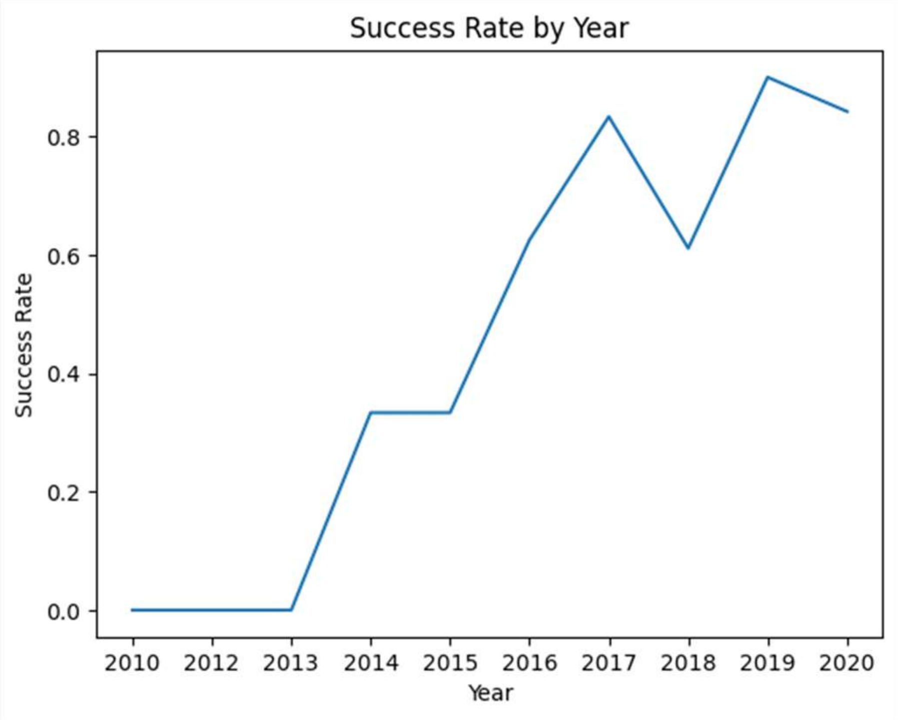
- You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

Launch Success Yearly Trend



- We can observe that the success rate since 2013 kept increasing till 2020

All Launch Site Names

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

In [11]:
```
%sql SELECT DISTINCT Launch_Site from SPACEXTABLE
```

```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

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

Launch Site Names Begin with 'CCA'

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

```
In [10]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like 'CCA%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

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

```
In [13]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer  = 'NASA (CRS)'

 * sqlite:///my_data1.db
Done.

Out[13]:
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

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

```
In [15]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.0%'
 * sqlite:///my_data1.db
Done.
```

Out[15]:

avg(PAYLOAD_MASS__KG_)
340.4

First Successful Ground Landing Date

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

Hint:Use min function

In [24]:
```
%sql select MIN(Date) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[24]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [26]:
```
%sql select distinct Booster_Version from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

```
 * sqlite:///my_data1.db
Done.
```

Out[26]:

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

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
In [27]: %sql select Mission_Outcome,count(*) from SPACEXTABLE group by Mission_Outcome;

 * sqlite:///my_data1.db
Done.
```

Out[27]:

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

Boosters Carried Maximum Payload

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

```
%sql select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_
                                        = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

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

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015. ¶

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE
       where Landing_Outcome = 'Failure (drone ship)' and substr(Date,0,5)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

Month	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

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.

```
%sql select Landing_Outcome,count(*) as CNT from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20'
        GROUP BY Landing_Outcome ORDER BY CNT DESC
```

```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	CNT
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

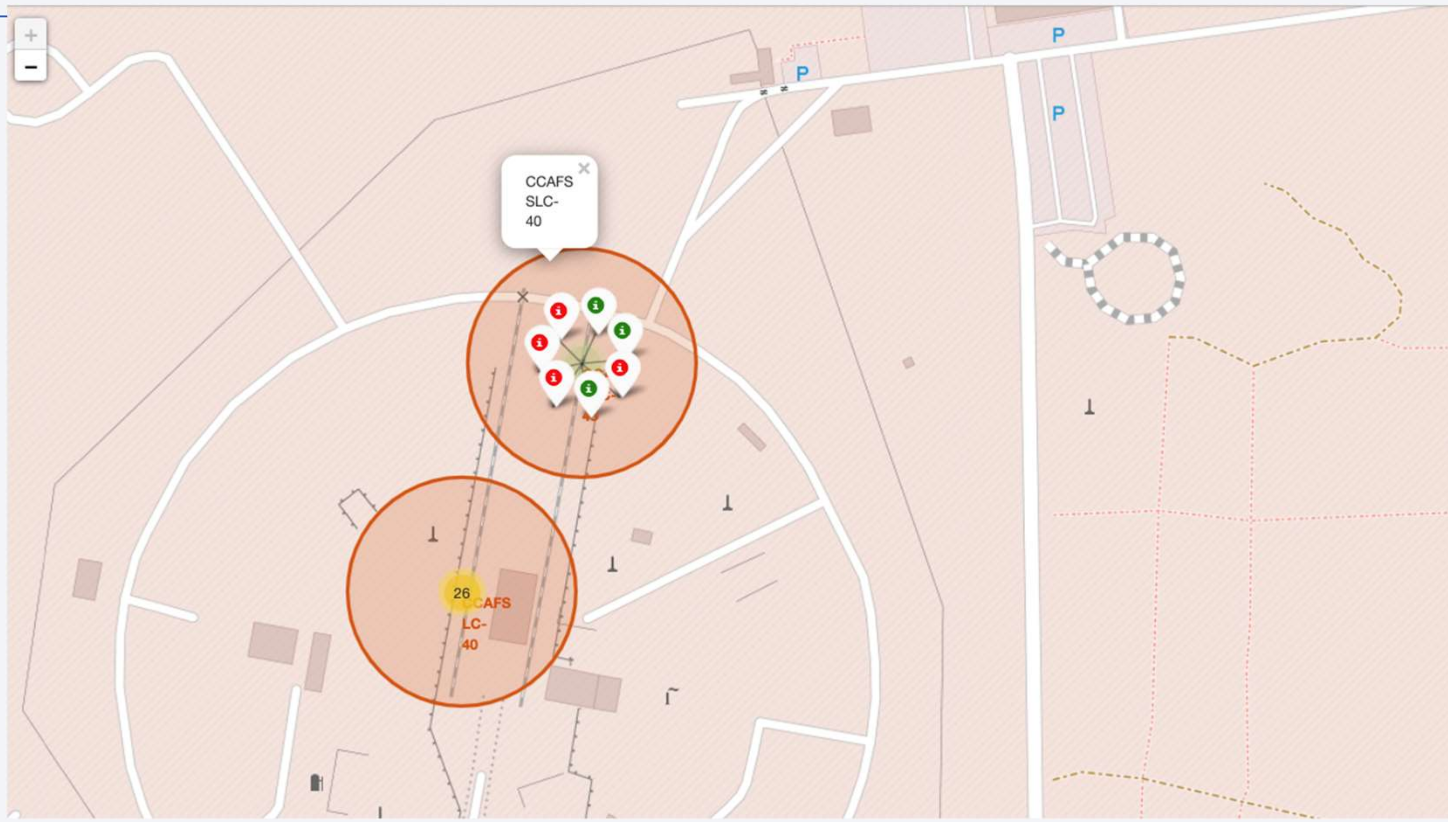
Section 3

Launch Sites Proximities Analysis

All Launch Sites marked on a Map

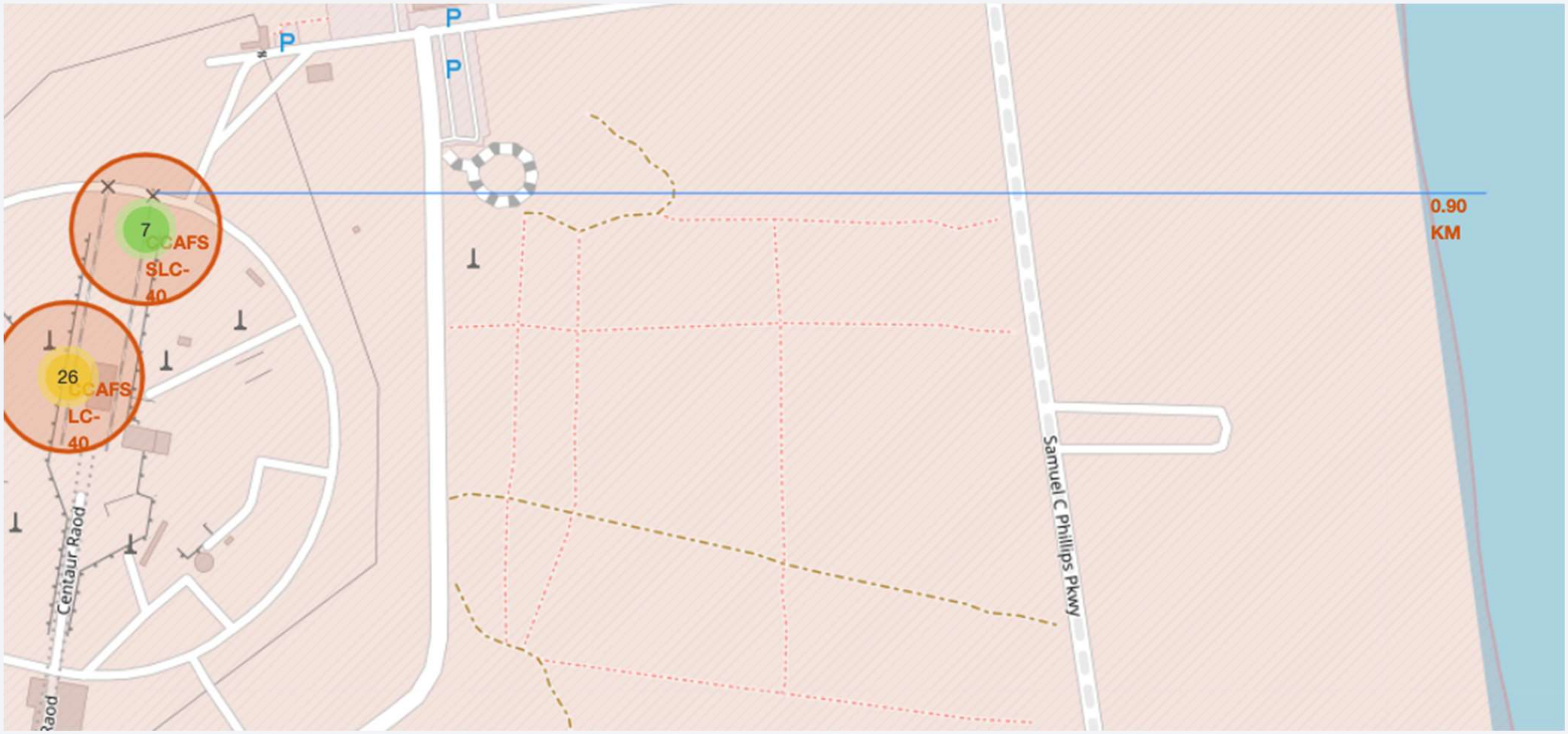
- There are launch sites on East and West coast of US.

Success / Failed Launched marked on the map



- Success Outcome shown in Green while failure launch are marked as Red

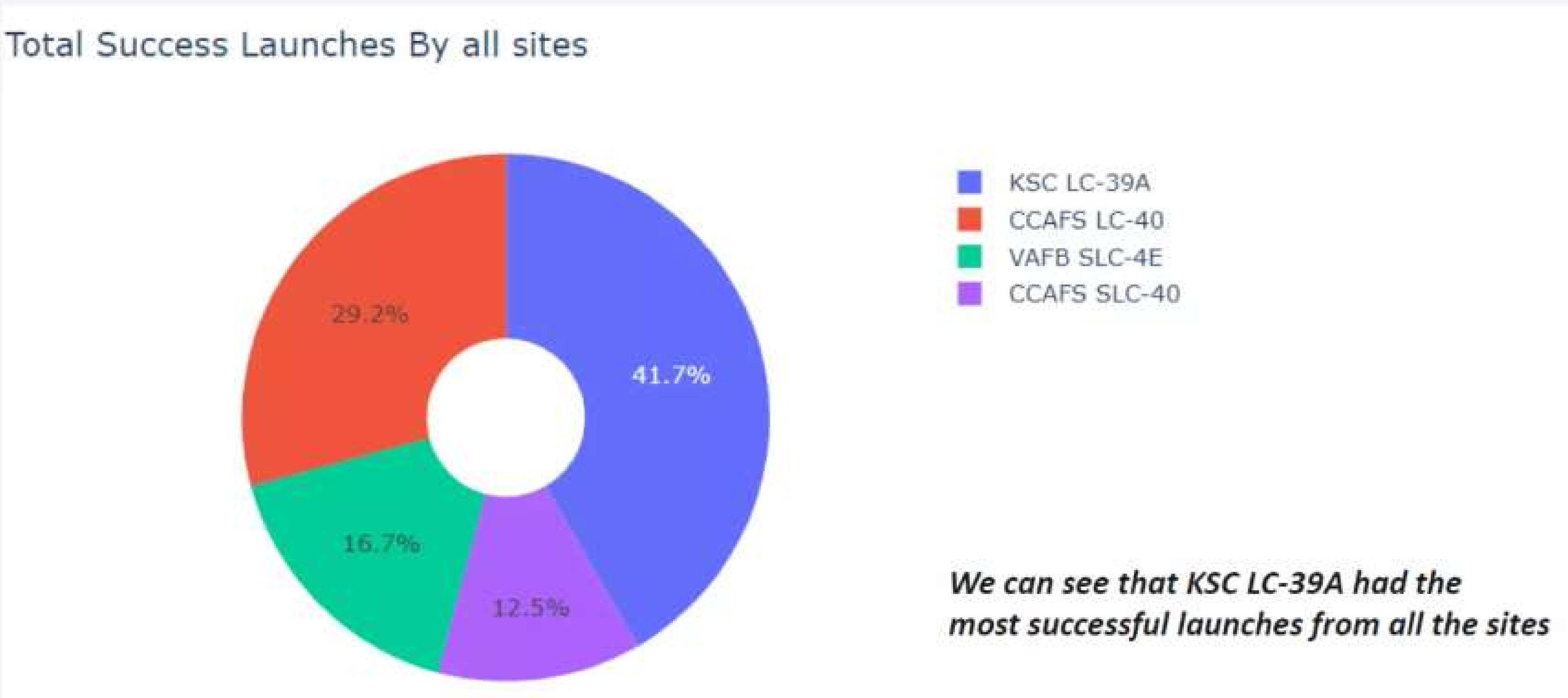
Launch Site with proximities



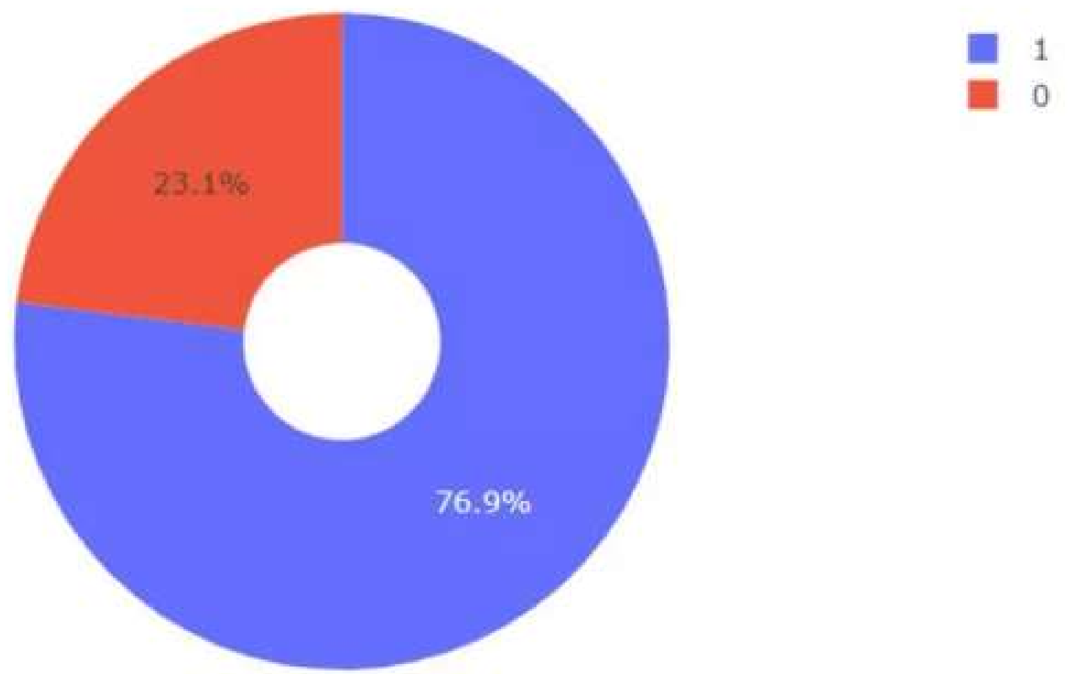
Section 4

Build a Dashboard with Plotly Dash

Total Success Launch by All Sites

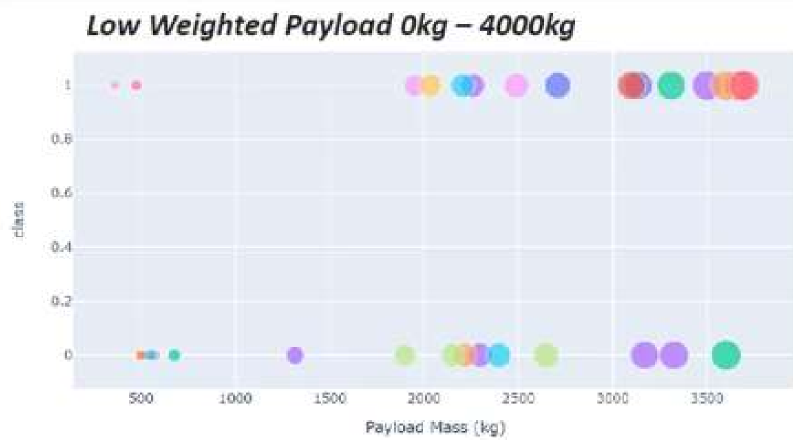


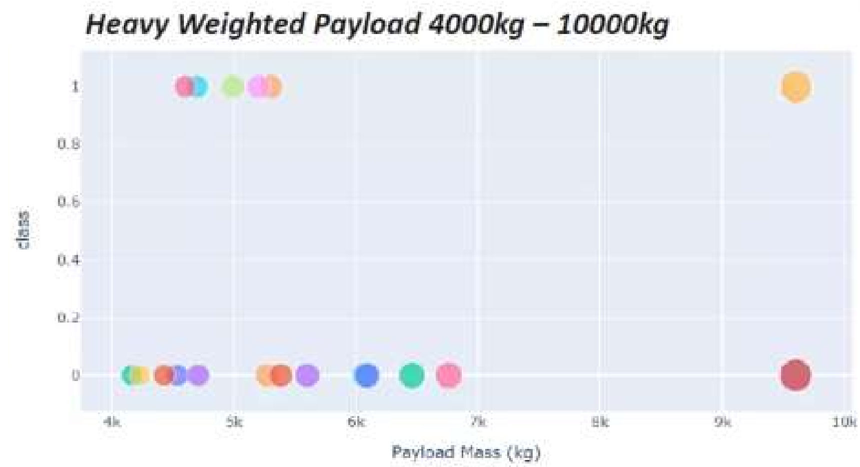
Success Rate by Site



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

Payload vs Launch Outcome





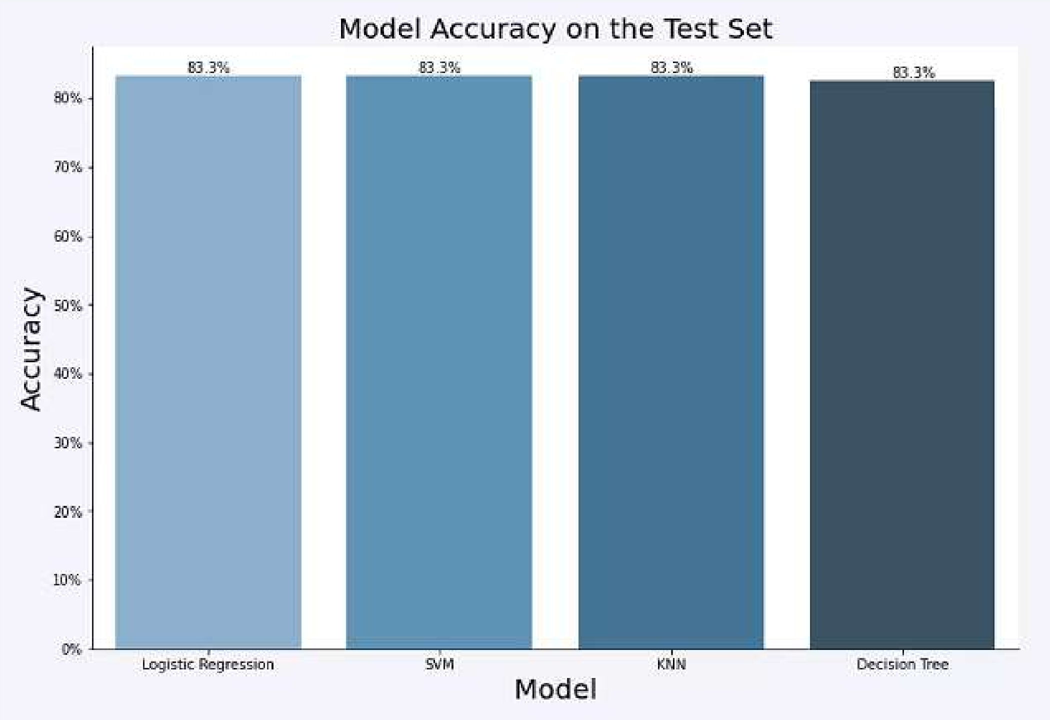
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

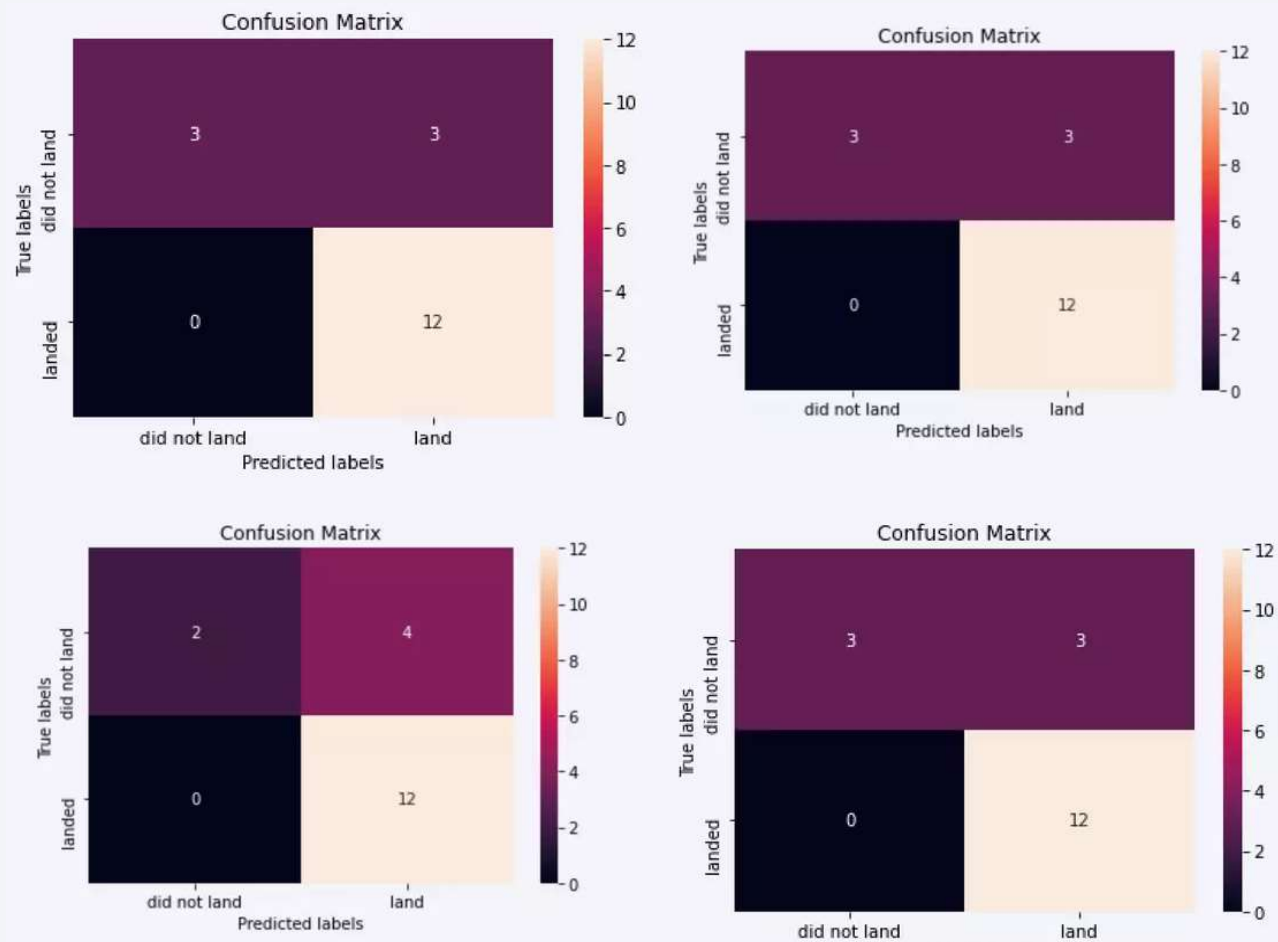
Classification Accuracy

- All models whether Logistic Regression, Support Vector, Decision Tree or K-Nearest Neighbor performed equally well.



Confusion Matrix

- All models confusion matrix yield the same result whether LR, KNN, SVM or DC.



Conclusions

- All the classification models SVM, KNN, LR and Decision Tree provided equally good results.
- Low weight payloads perform better than heavier payloads.
- The success rates for SpaceX launches is directly proportional to time in years its takes to perfect the launches.
- Orbits ES-L1, GEO, HEO and SSO has the highest success rate

Appendix

All supporting notebooks or python snippets have been uploaded into following GitHub Link:

https://github.com/mrehanzafar/IBM-Data-Science-Professional-Certificate/tree/main


Thank you!