IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Krishna Varma Vetukuri
24th November 2022

Outline of the Presentation

- Executive Summary of Methodologies and Results
- Introduction to project background and problems to find answers
- Methodologies required for analysis and prediction
- Analytical and Explored Results
- Conclusion

Executive Summary of Methodologies and Results

The Summary of methodologies used to analyze data are:

- Data Collection REST API and with Web Scraping
- Data Wrangling using Wikipedia
- EDA (Exploratory Data Analysis) with data visualization and by using SQL
- Building an interactive Visual Analytics to find the launch site location with Folium and Dashboard with Plotly Dash
- Machine Learning Predictive analysis

Analytical and Explored Results Summary:

- EDA (Exploratory Data Analysis) results helped in predicting the successful launches
- Interactive analytics helped in utilizing the data by visualizing
- Predictive analysis helped in predicting future outcomes from the previous outcomes

Introduction to project background and problems to find answers

Project background and context

- The initial cost of the Falcon 9 rocket was 62 million dollars which was advertised by SpaceX. But for other providers it is 165 million dollars, because SpaceX uses reusing of the first stage which saves a lot of money for SpaceX.

Problems you want to find answers

- This Capstone Project is to predict if the Falcon 9 first stage will land successfully, and by what probability.

Section 1

Methodology

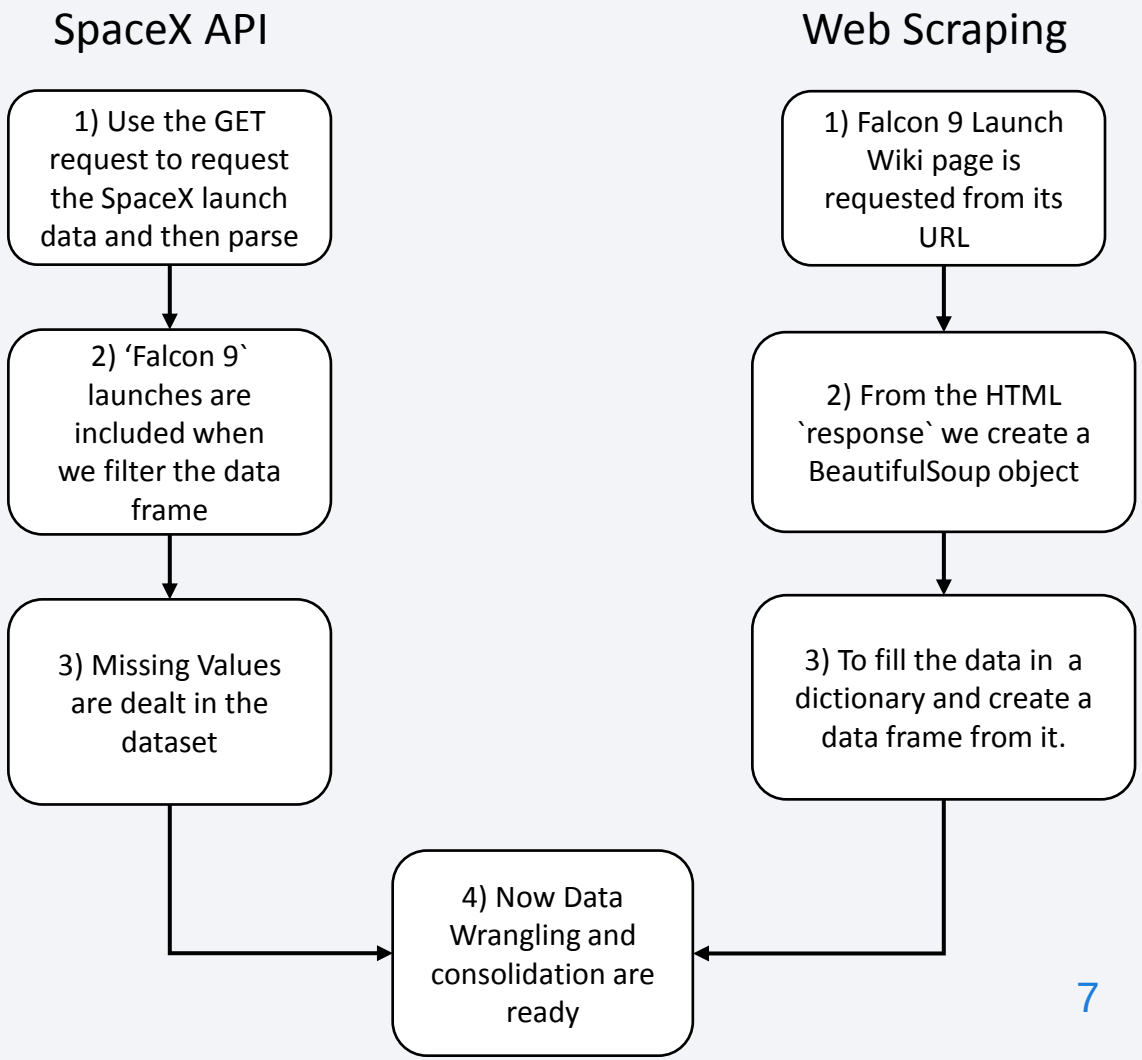
Methodologies required for analysis and prediction

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping form Wikipedia
- Perform data wrangling methodology:
  - Removes errors and combines complex data sets which will make the data easy to analyze
- Using visualization and SQL perform exploratory data analysis (EDA)
- Using Folium and Plotly Dash performing interactive visual analytics to find the launch site location with Folium
- Using classification models performing Machine Learning predictive analysis :
  - For the best classifier the models built and evaluated were KNN, DT, LR, SVM models

Data Collection API and with Web Scraping

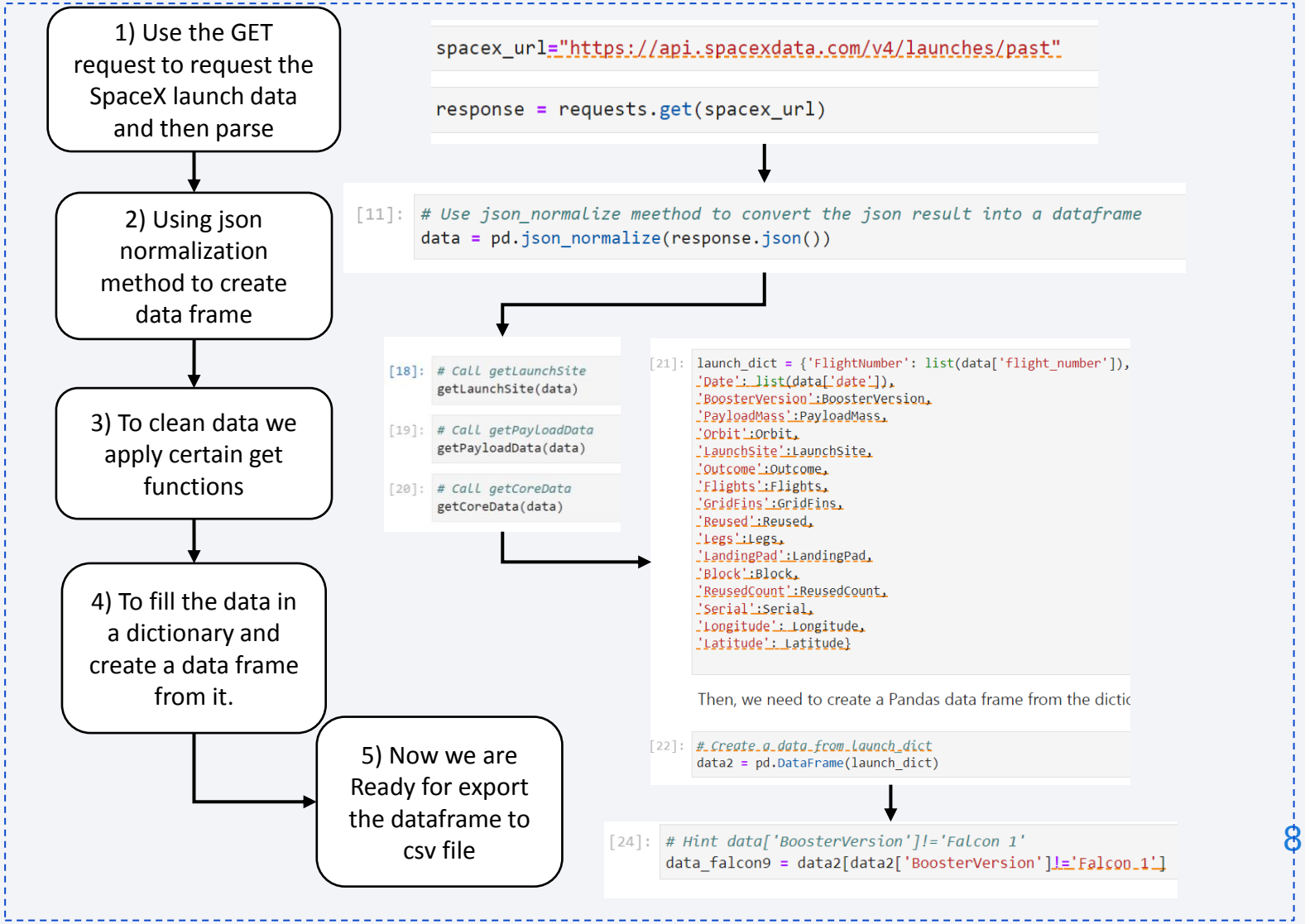
- The data is launched by SpaceX which is gathered by the SpaceX REST API
- This API gives information about launches, which includes information about the number of rockets used, its landing specifications, the payload that has been delivered, outcomes on landing, and specifications on landing.
- Web scraping Wikipedia is a very popular way to obtain Falcon 9 Launch data using BeautifulSoup.



Data Collection with SpaceX REST API

- Data collection using SpaceX REST API
- GitHub URL:

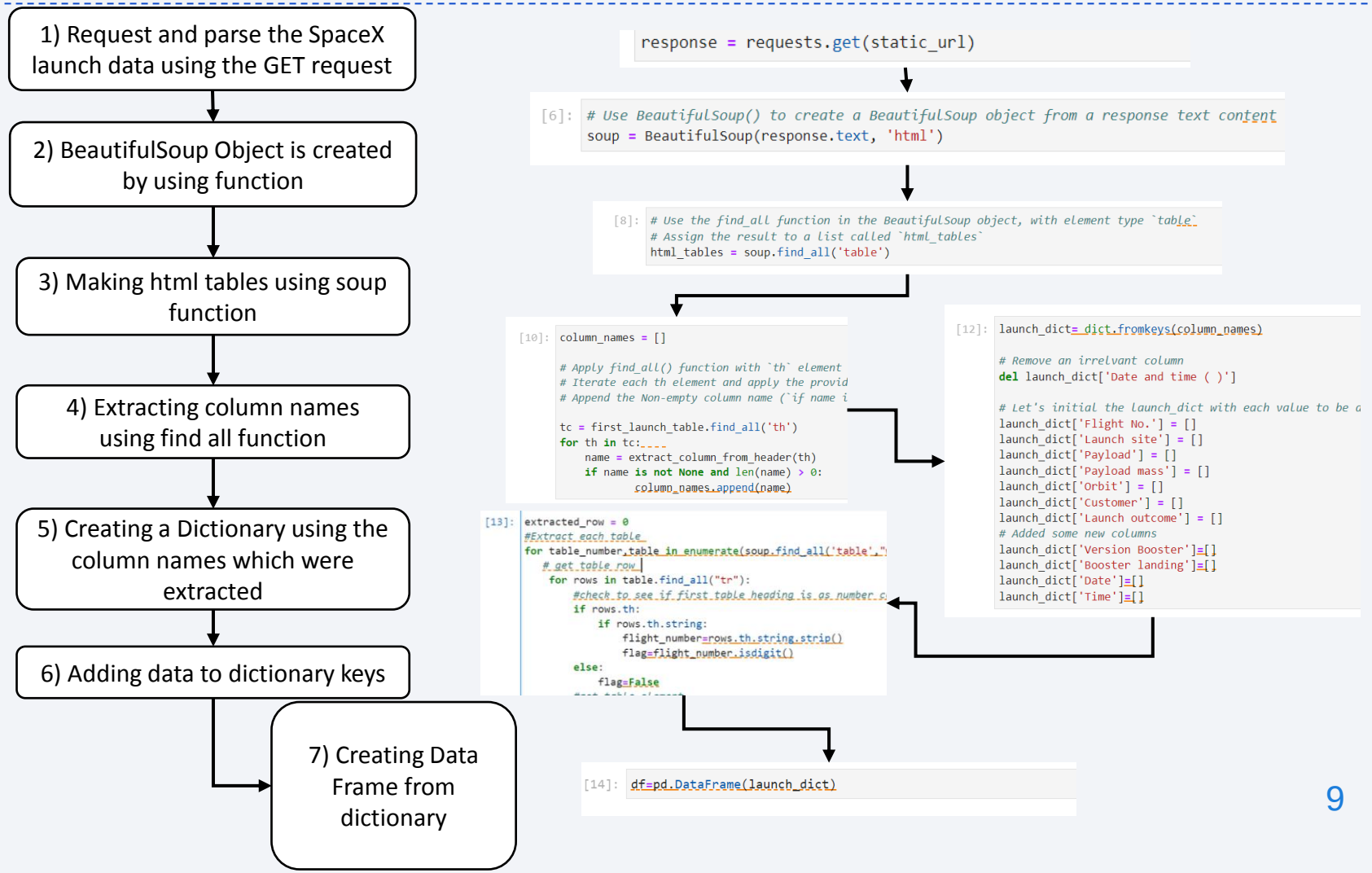
https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/014434476b9361dba5965d106b30a4222c84b5fa/W1_data_collection_api.ipynb



Data Collection Web Scraping from Wikipedia

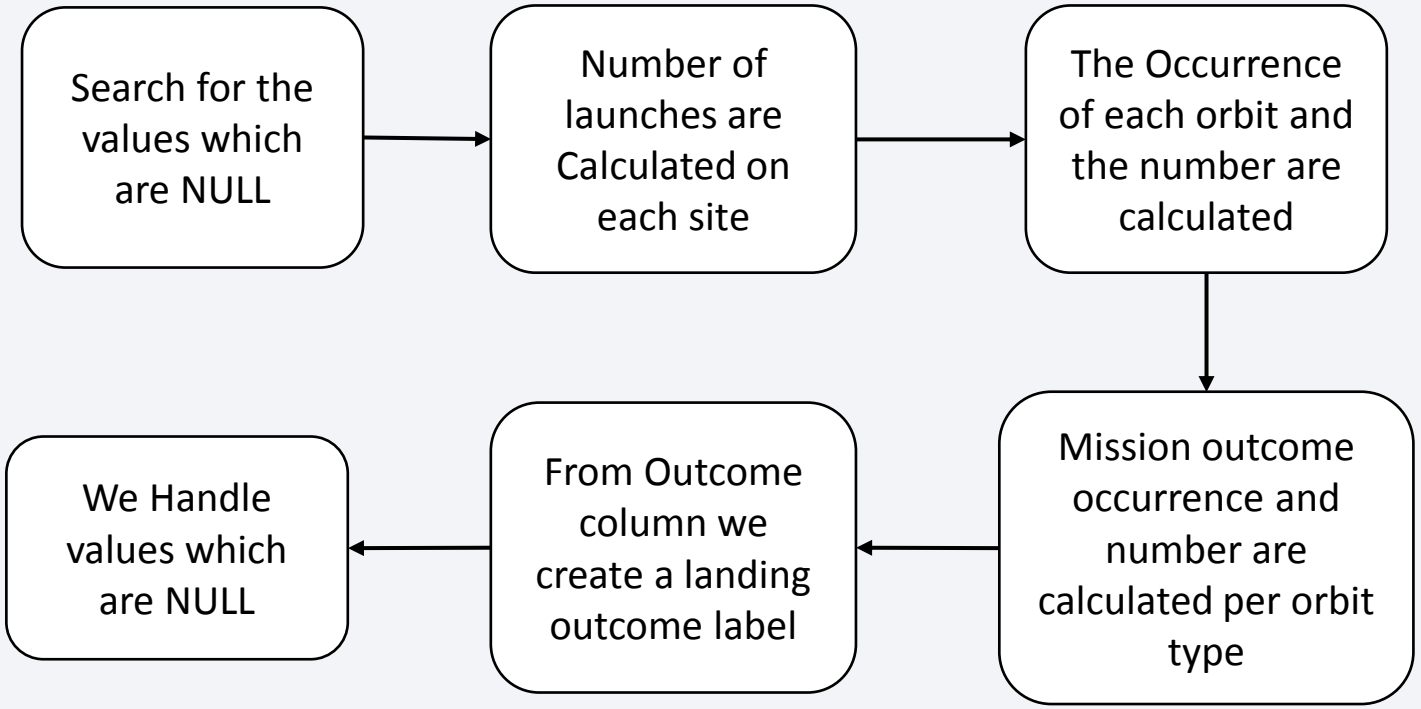
- Data Collection using Web scraping process with the help of Wikipedia
- GitHub URL:

https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/5e335eaf50f8becb27c1f00caf2ed848726371ff/W1_data_webscraping.ipynb



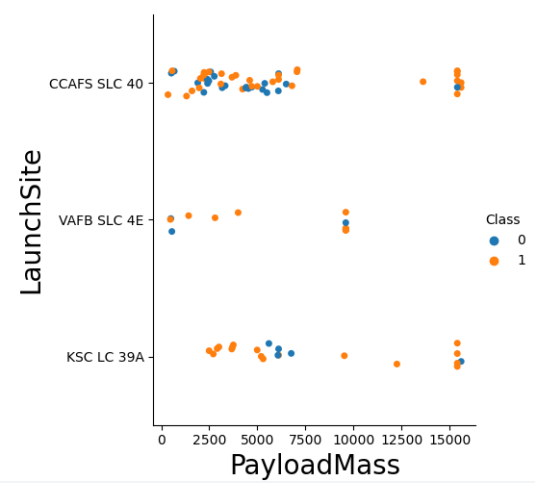
Data Wrangling to check number of occurences

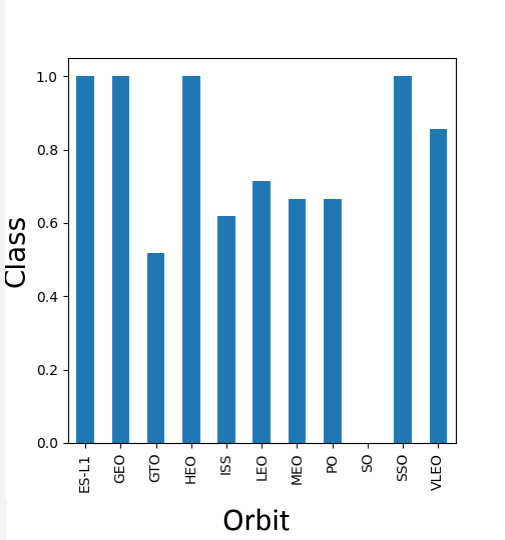
- Data wrangling process is shown in the mind map

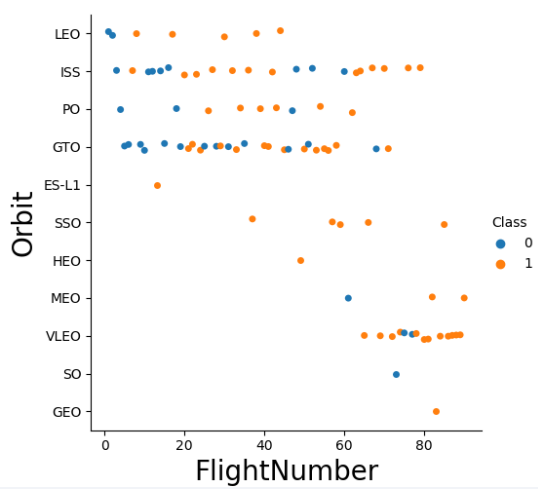


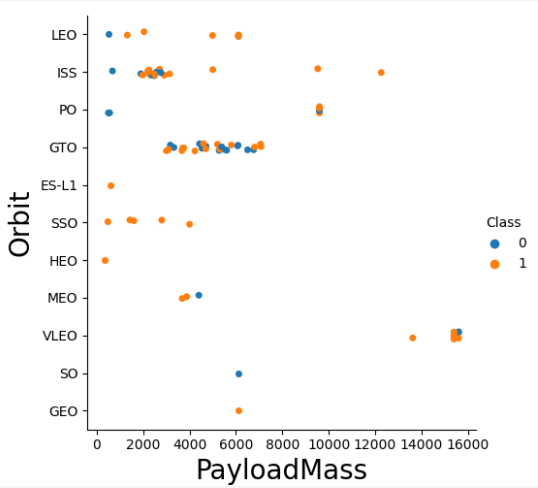
- GitHub URL:
- https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/5e335eaf50f8becb27c1f00caf2ed848726371ff/W1_data_wrangling.ipynb

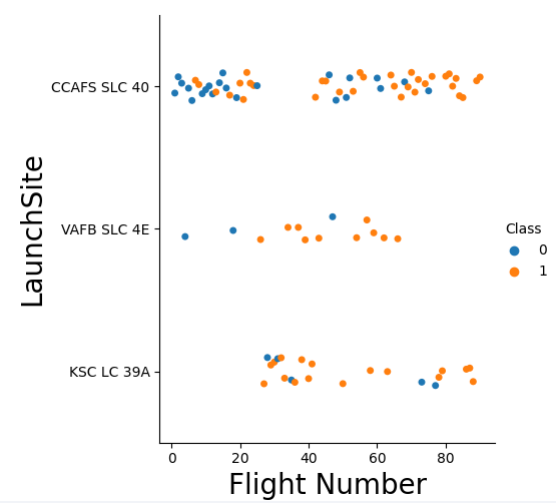
Exploratory Data Analysis with Data Visualization



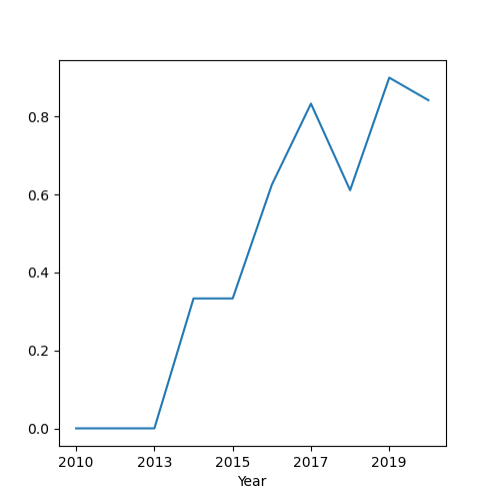








- GitHub URL:
- https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/eb5d587cc9bc19552e90667c65084447407f53fc/W2_eda_data_visualization_lab.ipynb



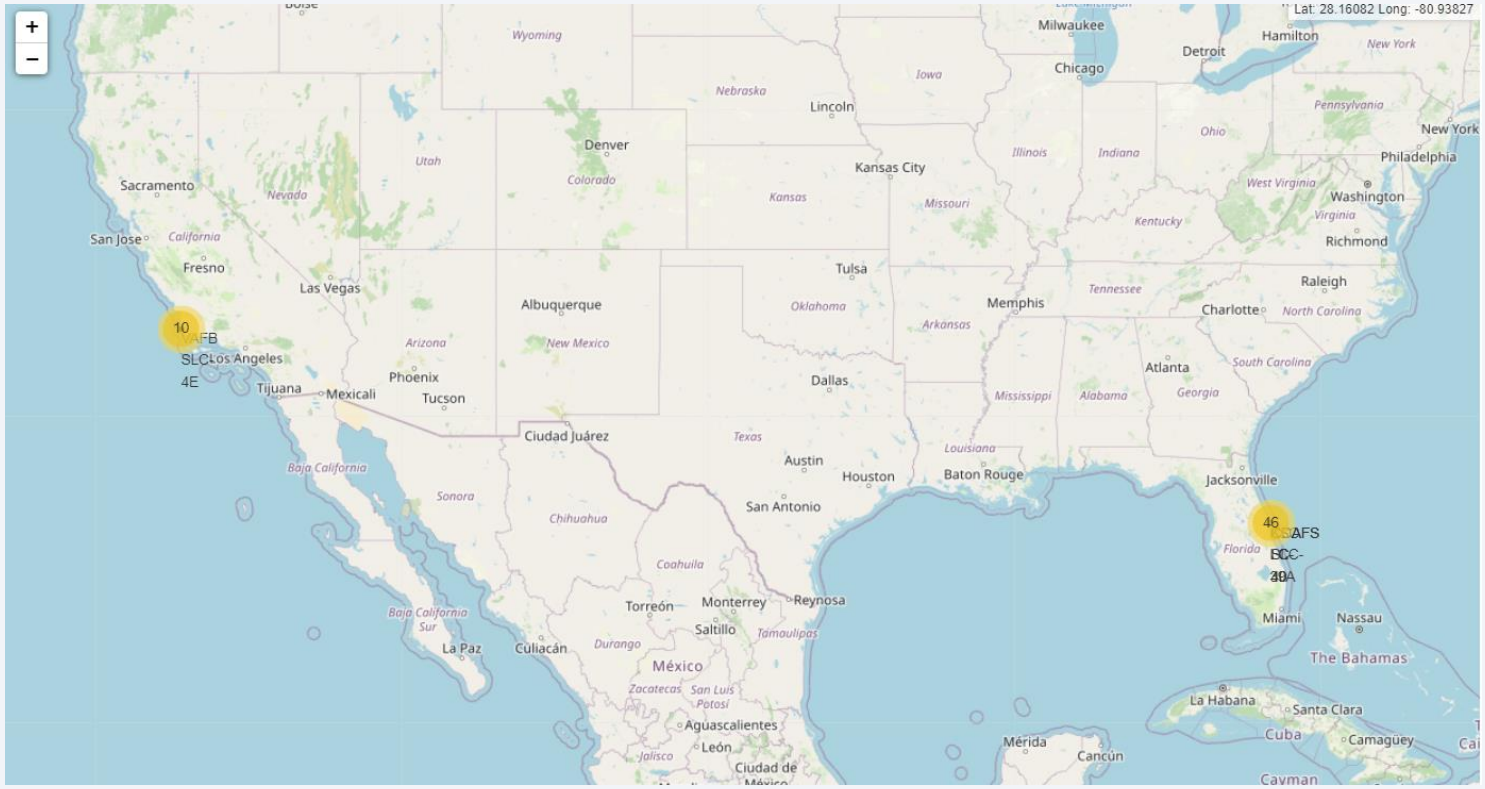
Exploratory Data Analysis with SQL

SQL queries performed are:

- Database connecting and in the space mission, displaying the unique launch sites names.
- Where launch sites begin with the string 'CCA' display the five records.
- Total payload mass launched by NASA (CRS) and which is carried by boosters is displayed
- Average payload mass which is carried by booster version F9 v1.1 is displayed
- The date when the Achievement of the First successful landing outcome in ground pad took place is listed.
- The Boosters names which have payload mass more than 4000 but lower than 6000 and have drone ship success are listed
- The Total number of failure and successful mission outcomes are listed
- Use a subquery to find the list of booster versions names which have carried the highest payload mass .
- For the months in year 2015, display the list that displays month names, failure landing outcomes in drone ship ,booster versions, launch site.
- Landing outcomes counts between the date 2010-06-04 and 2017-03-20 ranking shown in descending order.

GitHub URL:
https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/eb5d587cc9bc19552e90667c65084447407f53fc/W2_eda_sql_sqllite.ipynb

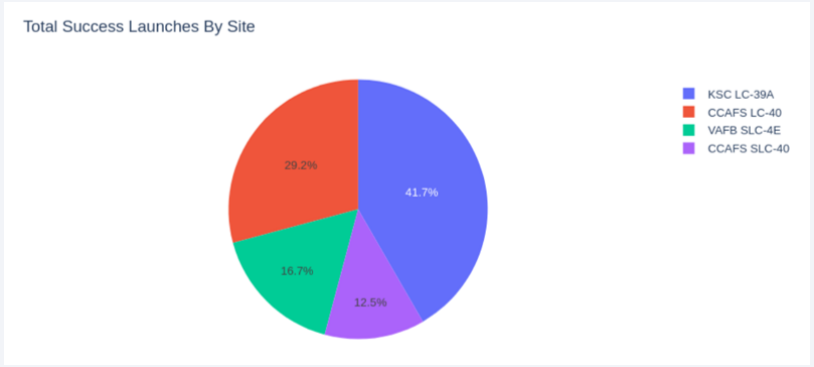
Building an interactive Visual Analytics to find the launch site location with Folium



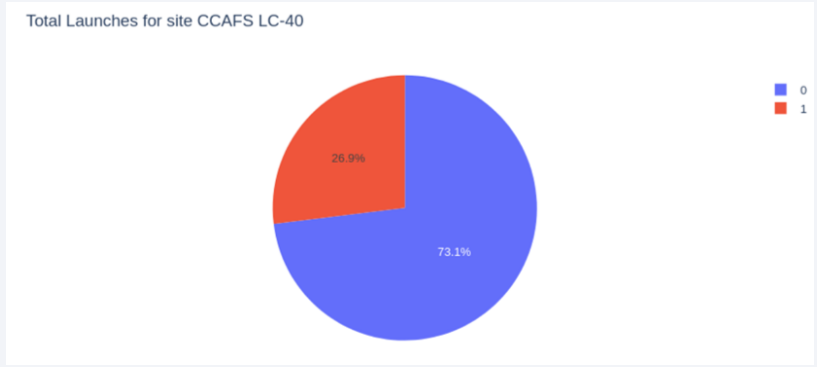
Map marker are used to Find an optimal location for building a launch site

GitHub URL:
https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/eb5d587cc9bc19552e90667c65084447407f53fc/W3_Interactive_Visual_Analytics_with_Folium_Lab_launch_site_location.ipynb

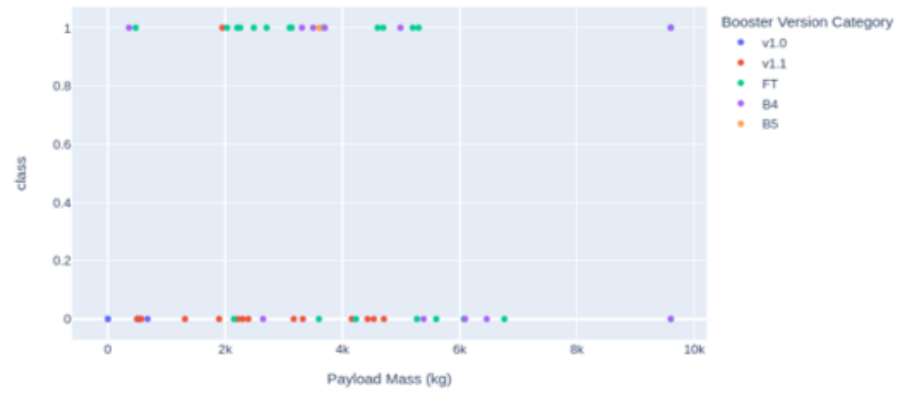
Building an interactive Visual Analytics and Dashboard with Plotly Dash

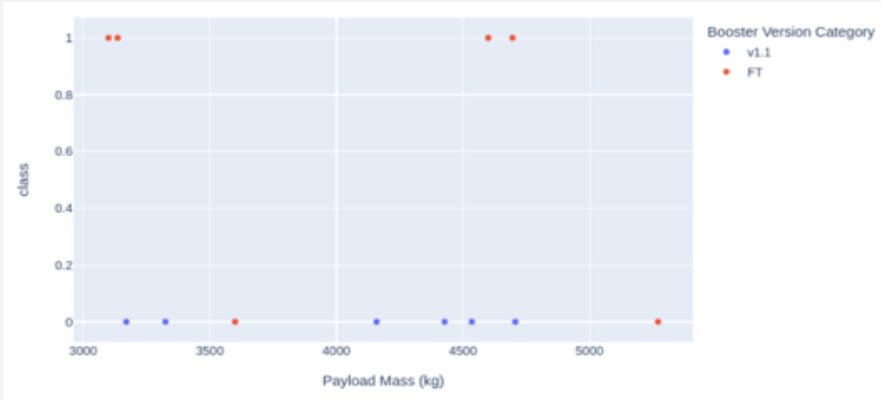


KSC LC-39A has the most number of successful launches from all the sites



Success rate of KSC LC-40 is 73.1% and Failure rate is 26.9% .



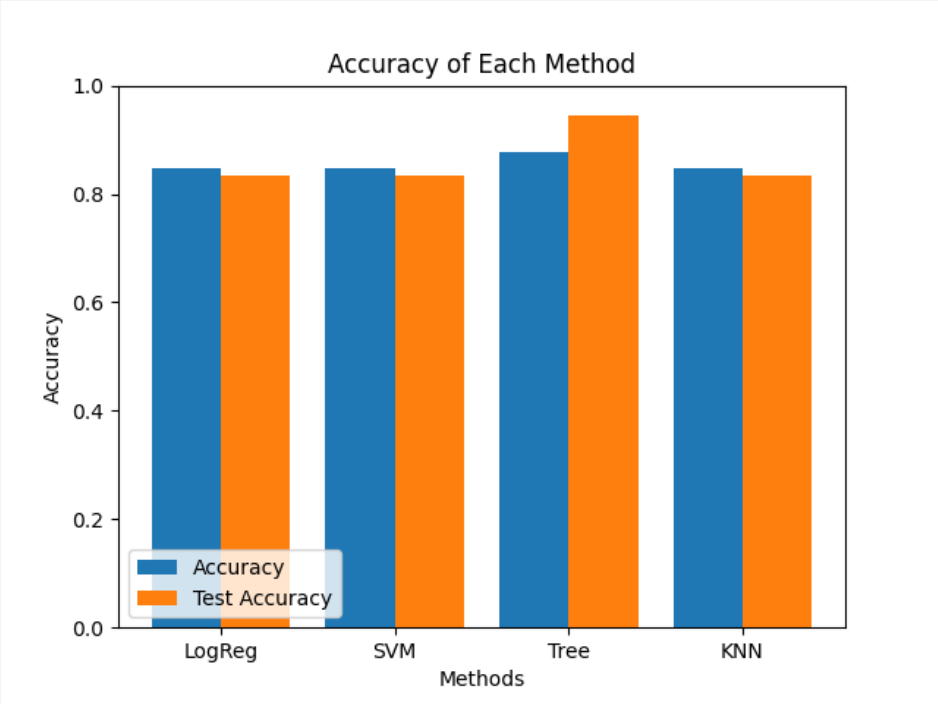


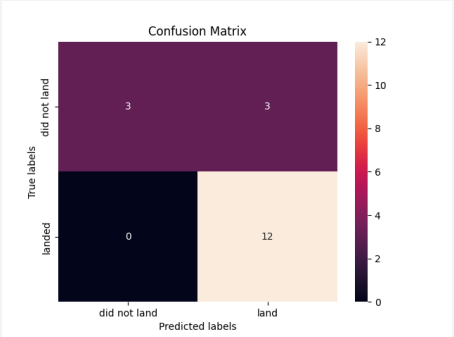
- Performance of Low weighted payloads is better when compared with the heavy weighted payloads.

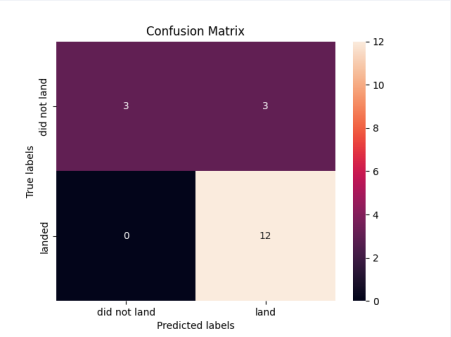
GitHub Url:
https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/eb5d587cc9bc19552e90667c65084447407f53fc/W3_spacex_dashboard_plotlyDash.py

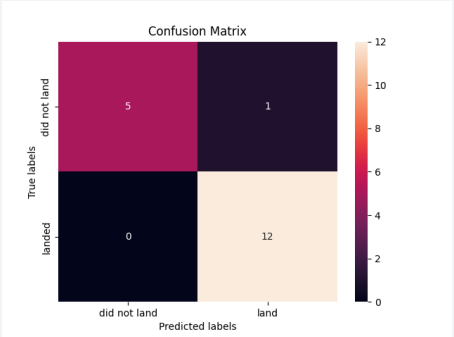
Machine Learning Predictive Analysis

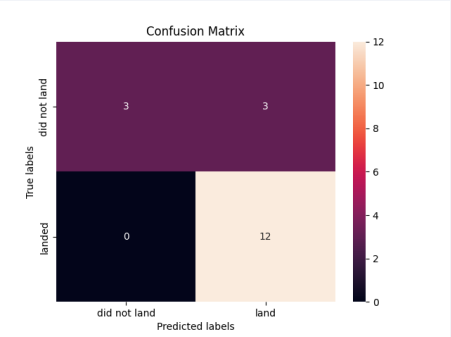
- The SVM, KNN, and Logistic Regression model achieved The highest accuracy is achieved by the Logistic Regression model, KNN, and SVM which is at 83.33%
- The SVM model method is the best in terms of AREA Under the Curve which is nearly 94%











GitHub URL:
https://github.com/krishnavarmavetukuri/Applied_Data_Science_Capstone/blob/ee1e2727b308b3459ca48f4c55205ced8326c624/W4_SpaceX_Machine_Learning_Prediction.ipynb

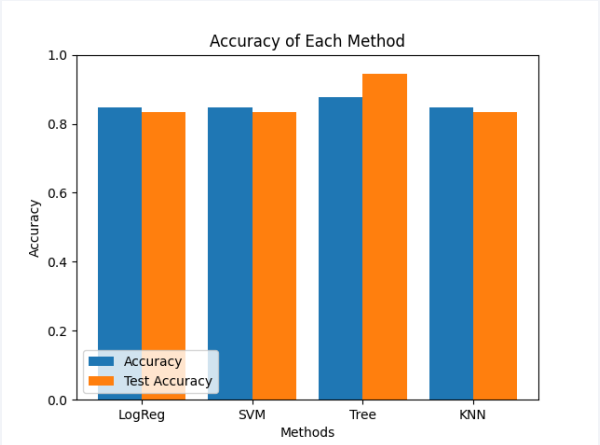
Analytical and Explored Results

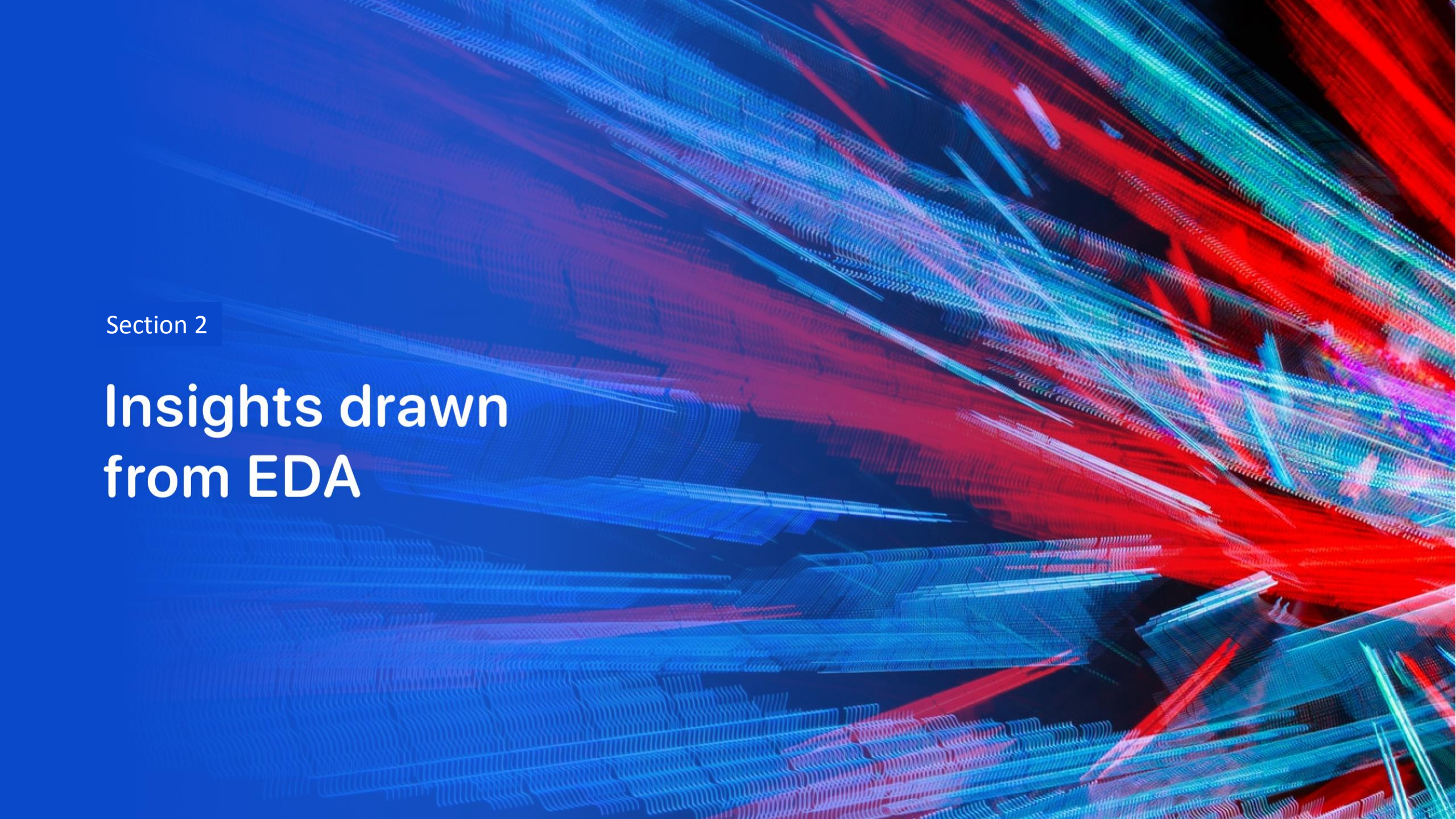
Exploratory data analysis results

- The best in terms of prediction accuracy are the SVM, KNN, and Logistic Regression models for this dataset.
- Performance of Low weighted payloads is better when compared with the heavy weighted payloads.
- The perfection in the launches is due to the direct proportionality of success rates to the time in years for SpaceX launches.
- The most successful launches is by KSC LC 39A when compared from all the sites.
- The best Success Rate is for the Orbits HEO, ES L1, SSO, GEO.
- Tree Classifier is best for prediction of successful profits

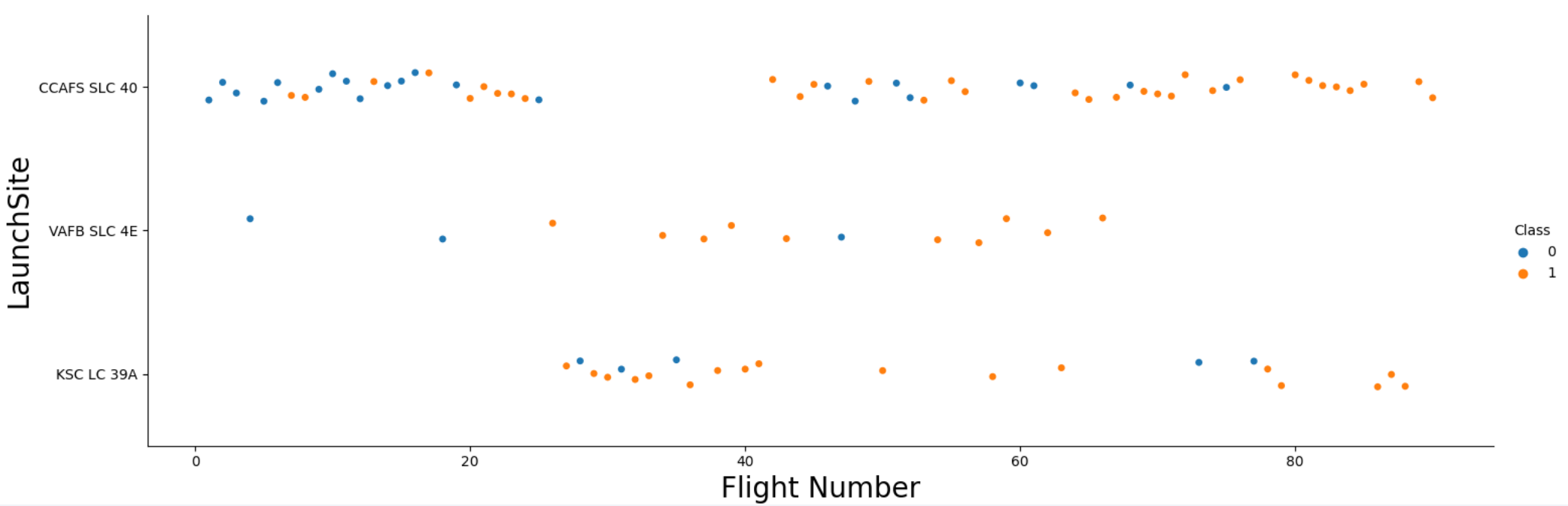
Interactive analytics demo in screenshots

Predictive analysis results



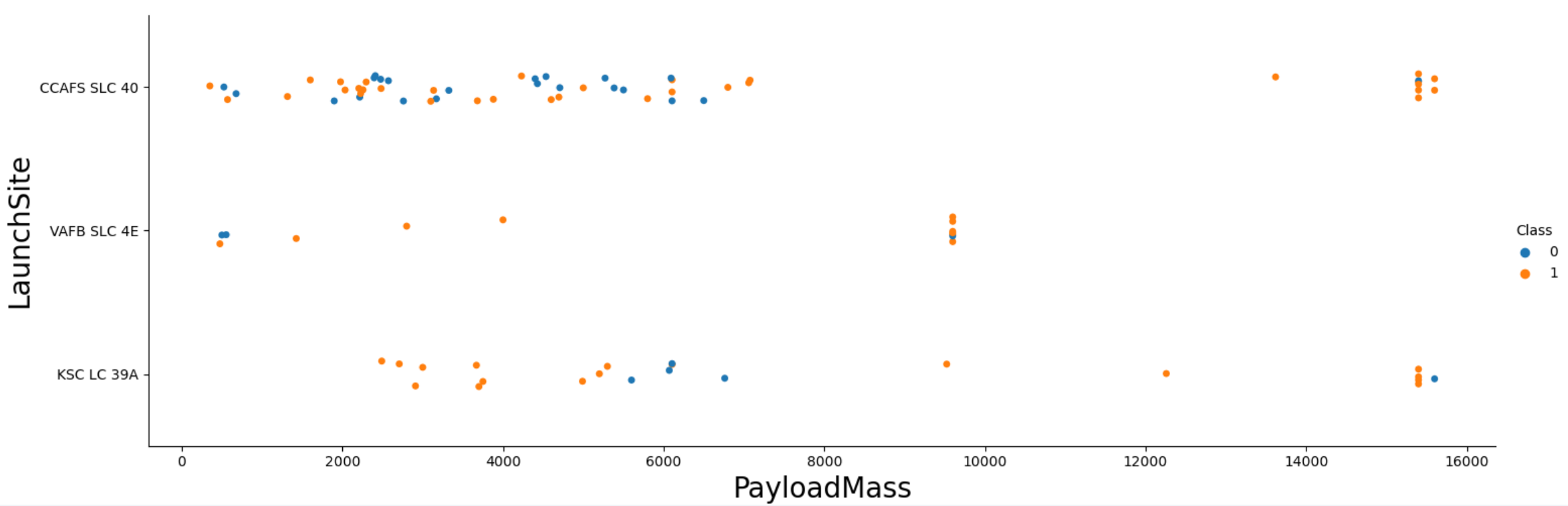

Section 2
Insights drawn from EDA

Launch Site vs. Flight Number of the SpaceX rockets



- CCAFS SLC 40 has high number of launches when compared with the launches from other sites
- The Success rate has been improving and increasing over time.

Launch Site vs. Payload of the SpaceX rockets

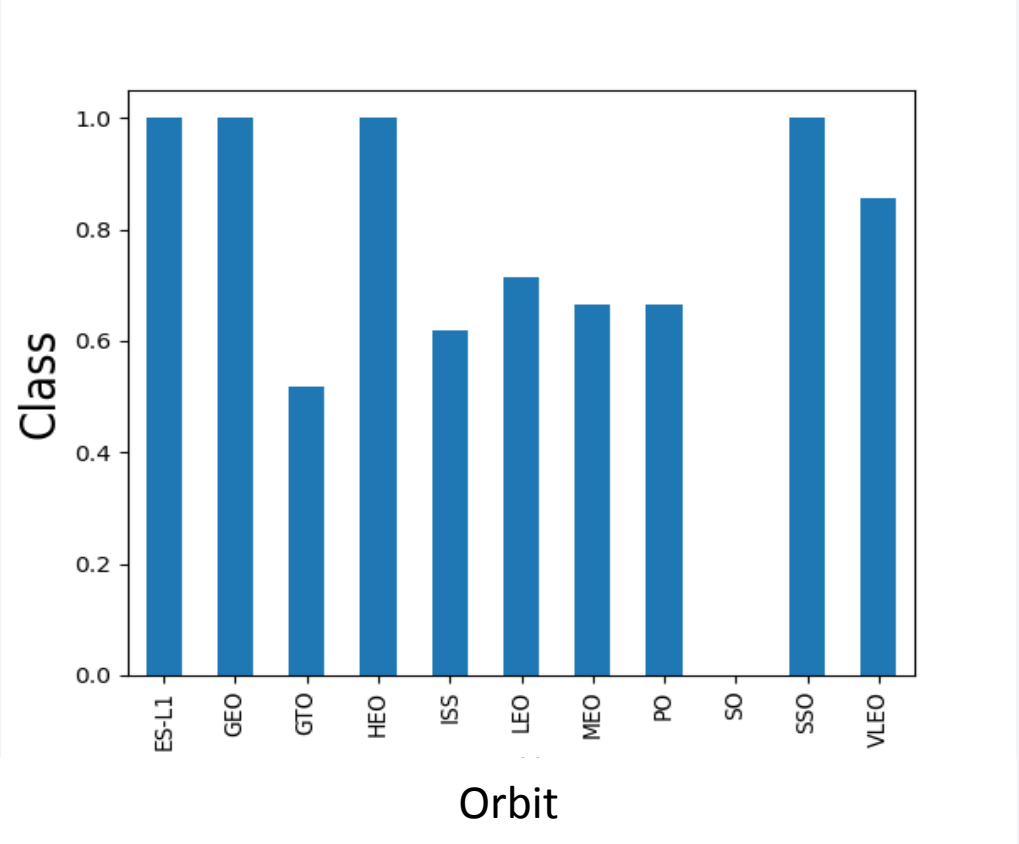


- CCAFS SLC 40 launched the highest number of PayLoads with lower Mass
- VAFB SLC 4E launch sites doesn’t seem to be possible when payloads over 12,000 Kgs.

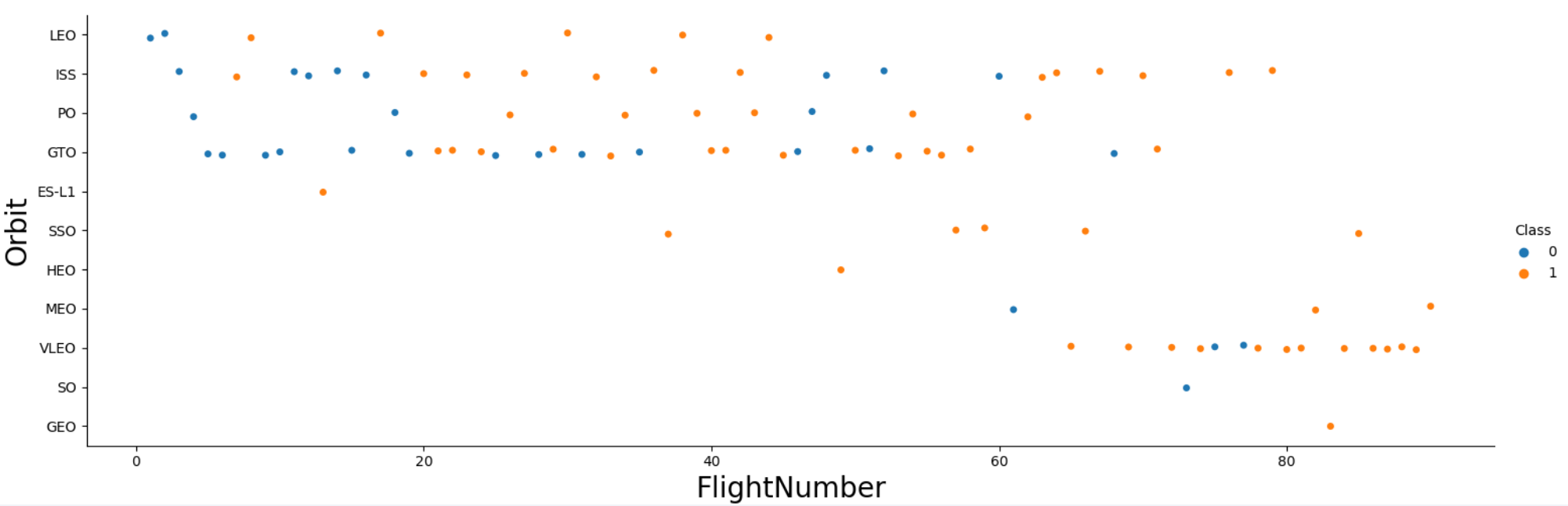
Orbit Type vs. Success Rate of the SpaceX rockets

The highest success rate happens to the orbit types:

- ES-L 1
- GEO
- HEO
- SSO

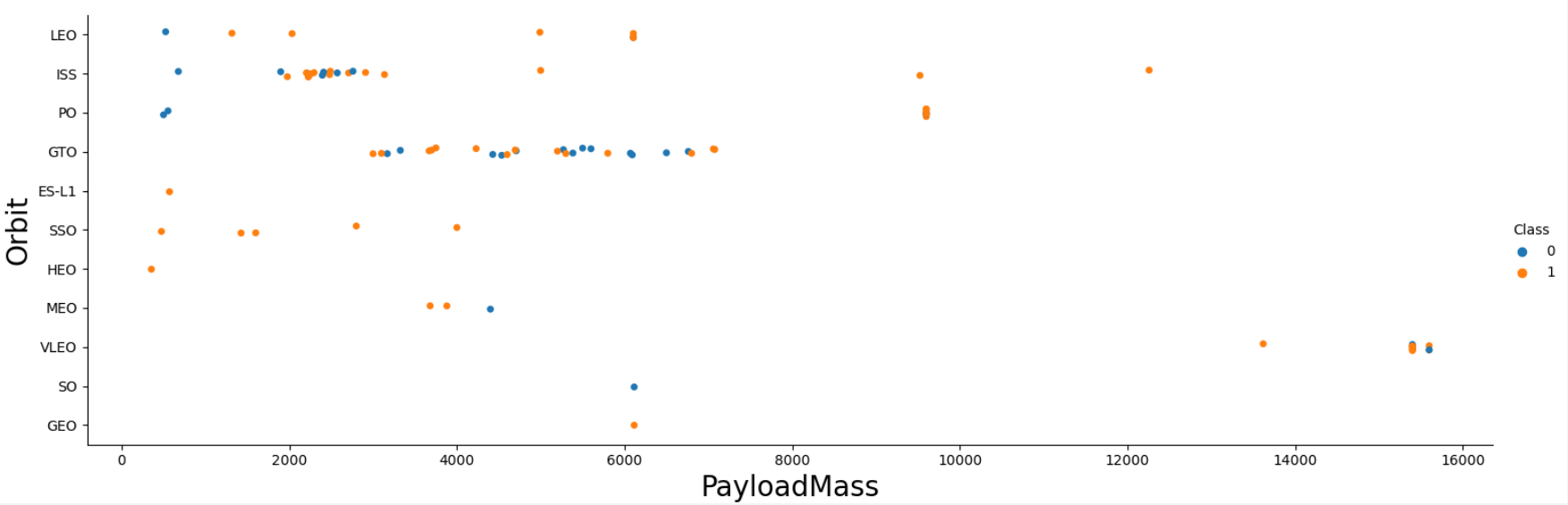


Orbit Type vs. Flight Number of the SpaceX rockets



- Success rate has been improving and increasing over time to all the orbits.
- A trend can be observed of shifting to VLEO launches in recent years

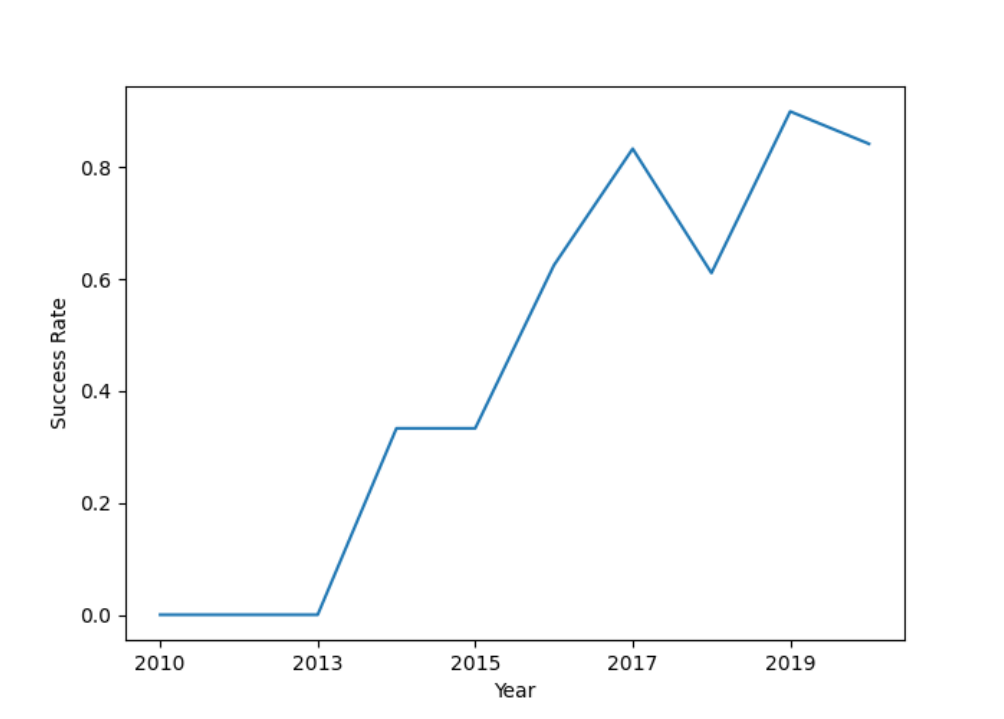
Orbit Type vs. Payload of the SpaceX rockets



- There were many launches for ISS when the Payload mass is around 2000, and the number of launches are high for GTO in the Payload mass is between 4000 and 8000
- There are very few launches for GEO and SO orbits.

Launch Success Yearly Trend of the SpaceX rockets

- After the year 2013 there was high increase in Launch Success rate
- In between the years 2014 and 2015 the success rate was constant but later again boosted
- After the year 2019 there was a decrease in Launch success rate.



All Launch Site Names of the SpaceX rockets

- We can say that there are 4 launch sites in the data and each occurred many times in the data.
  - The Query to get the required output shown below:

sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

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

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

Launch Site Names which Begin with 'CCA'

- Where launch sites begin with the string 'CCA' display the five records
  - The Query to get the required output shown below:

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

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

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 carried by boosters

- Total payload mass launched by NASA (CRS) and which is carried by boosters is displayed

  - The Query to get the required output shown below:

```
sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE
Customer = 'NASA (CRS)';
```

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

TOTAL_PAYLOAD
45596

Average Payload Mass by booster version F9 v1.1

- Average payload mass which is carried by booster version F9 v1.1 is displayed

- The Query to get the required output shown below:

```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM
SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- The date when the Achievement of the First successful landing outcome in ground pad took place is listed.

  - The Query to get the required output shown below:

```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE
LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * sqlite:///my_data1.db
Done.
 FIRST_SUCCESS_GP
------------------
        01-05-2017
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- The Boosters names which have payload mass more than 4000 but lower than 6000 and have drone ship success are listed

  - The Query to get the required output shown below:

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_
BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';
```

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

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

Total Number of Successful and Failure Mission Outcomes

- The Total number of failure and successful mission outcomes are listed

The Query to get the required output shown below:
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

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

]:

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

Boosters Carried Maximum Payload

- Use a subquery to find the list of booster versions names which have carried the highest payload mass .
    - The Query to get the required output shown below:

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records of drone ships

- For the months in year 2015, display the list that displays month names, failure landing outcomes in drone ship ,booster versions, launch site.

The Query to get the required output shown below:
sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND DATE LIKE '%2015'

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- Landing outcomes counts between the date 2010-06-04 and 2017-03-20 ranking shown in descending order.

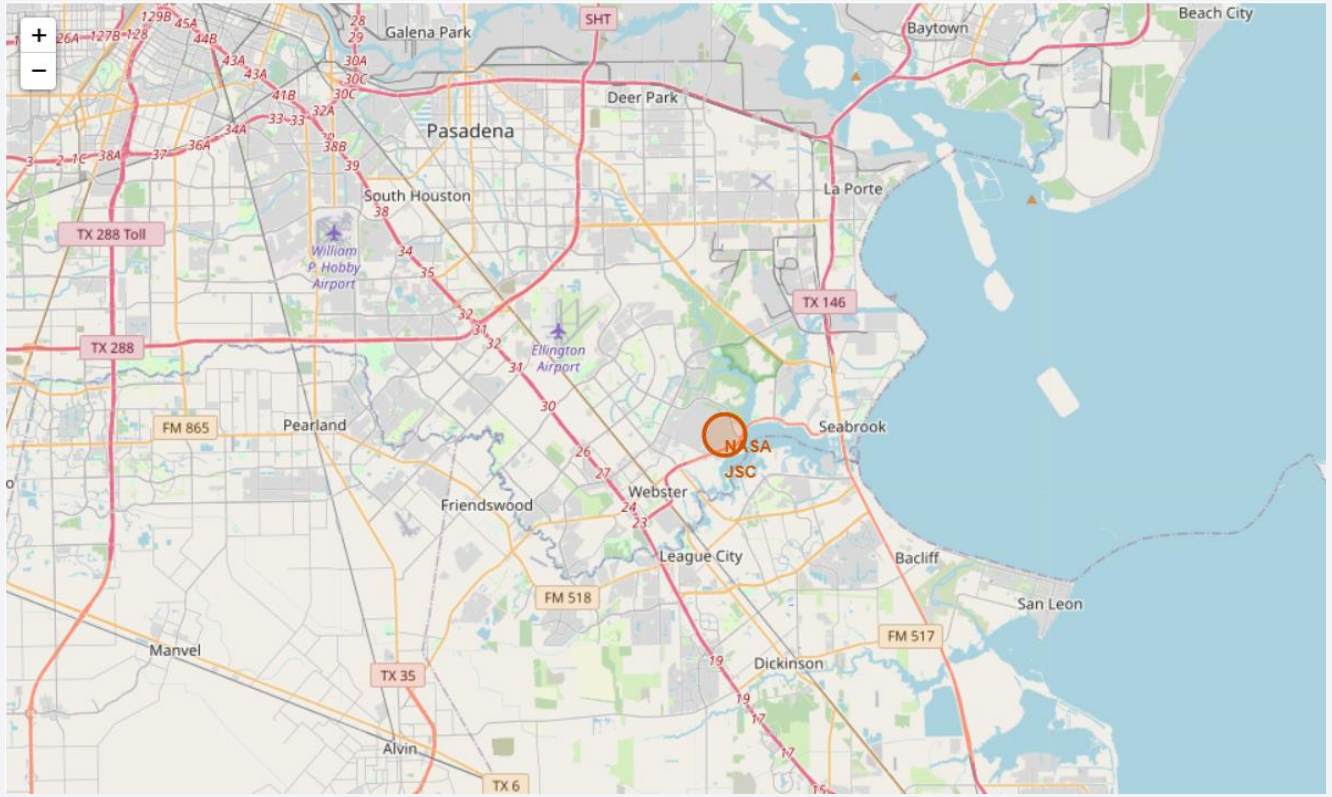
The Query to get the required output shown below:
sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;

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

Landing_Outcome	QTY
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

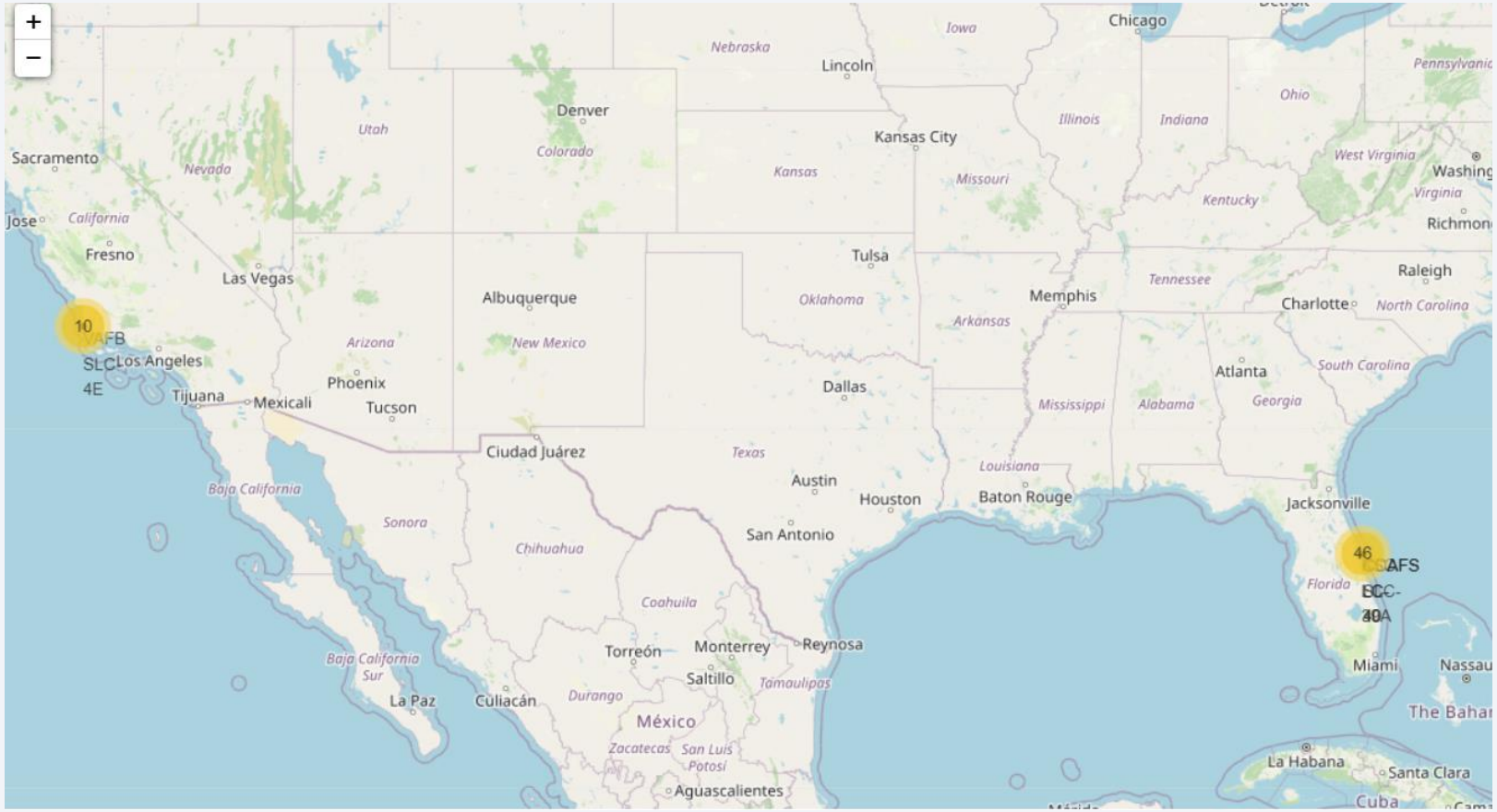

Section 3
Launch Sites
Proximities Analysis

All launch sites marked on a map



- This is the generated folium map with launch sites near the sea.

The success/failed launches is marked for each site on the map



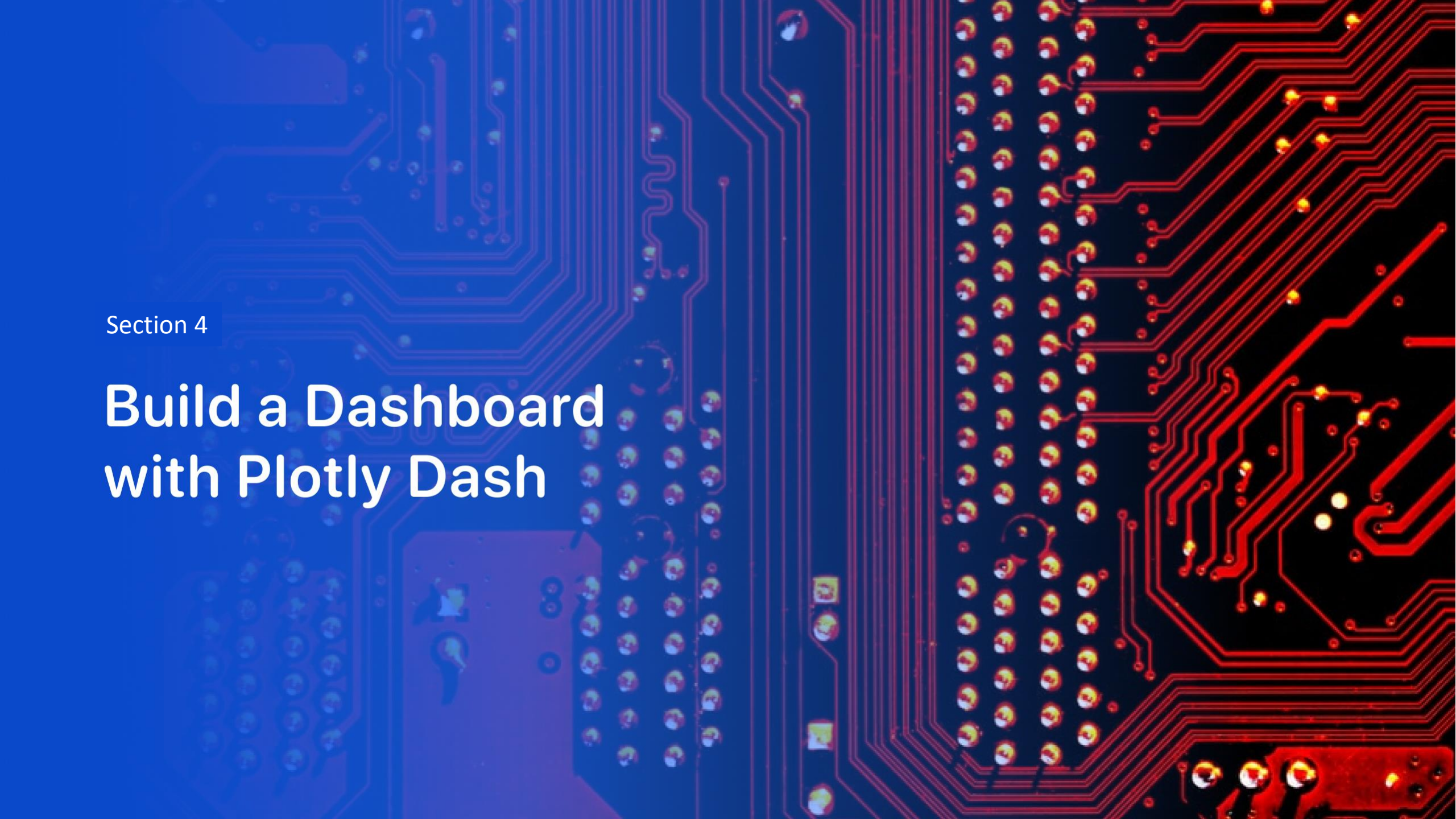
- In this folium map the colour-labeled launch outcomes on the map are shown.

The distances between a launch site to its proximities is calculated

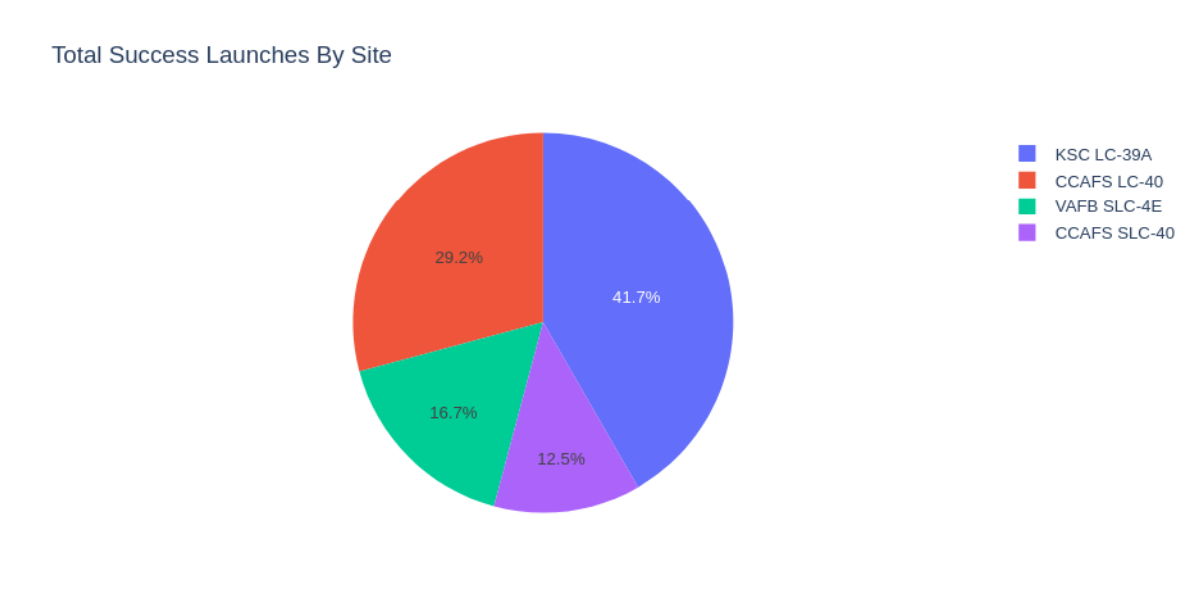
```
[18]: # find coordinate of the closet coastline
      # e.g.,: Lat: 28.56367  Lon: -80.57163
      # distance_coastline = calculate_distance(launch_site_lat, launch_site_lon, coastline_lat, coastline_lon)
      distance = calculate_distance(28.57468,-80.65229,28.573255 ,-80.646895)
      distance

[18]: 0.5503149993453544
```

- Launch site of KSC LC-39A has good logistics aspects though it is near sea and near to railroad and road and also far from inhabited areas

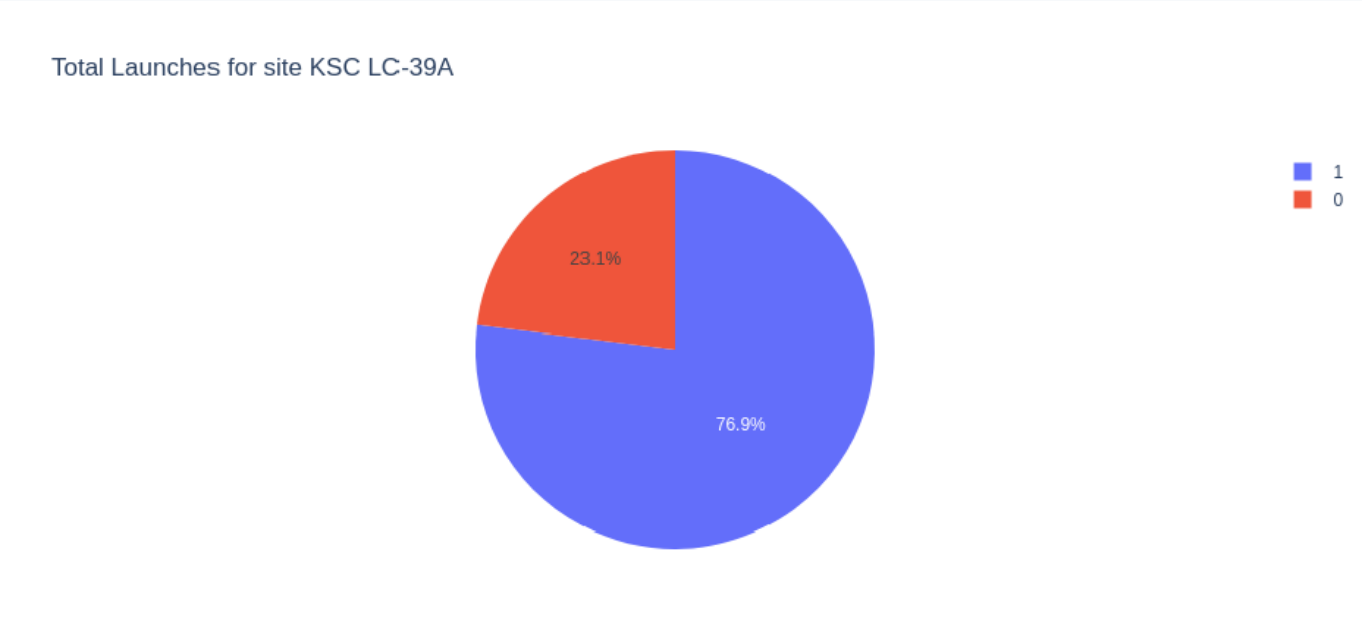

Section 4
Build a Dashboard with Plotly Dash

Four unique sites' total Success launches



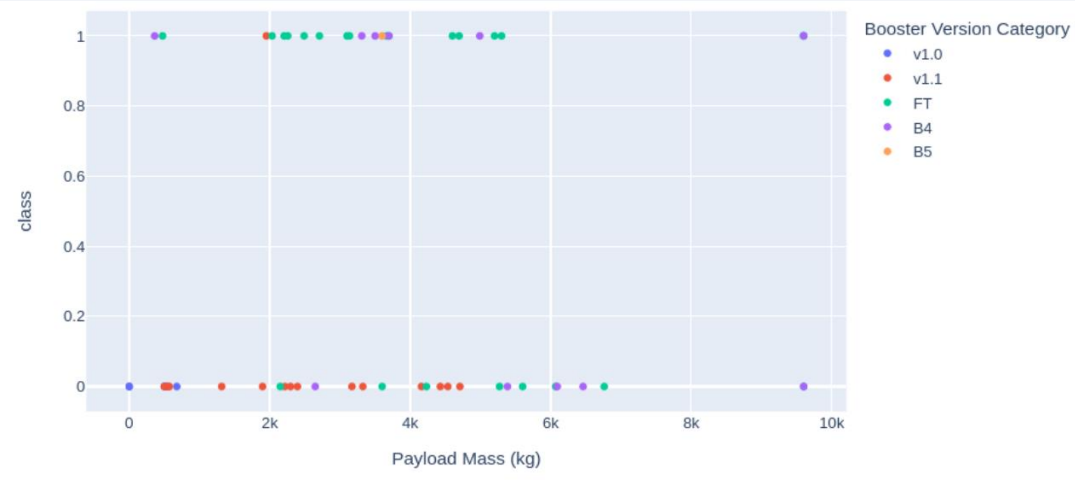
- In this Pie chart we can see that The most successful launches is by KSC LC 39A when compared from all the sites.

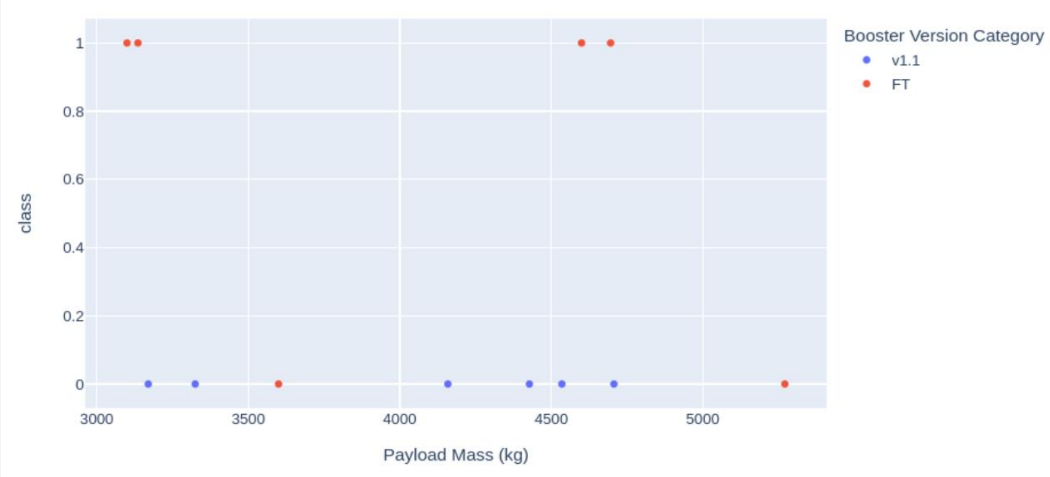
Success rate of the most successful site



- In this Pie chart we can see that success rate of KSC LC-39 is 76.9% and failure rate is 23.1%

Payload vs launch outcome of the SpaceX rockets



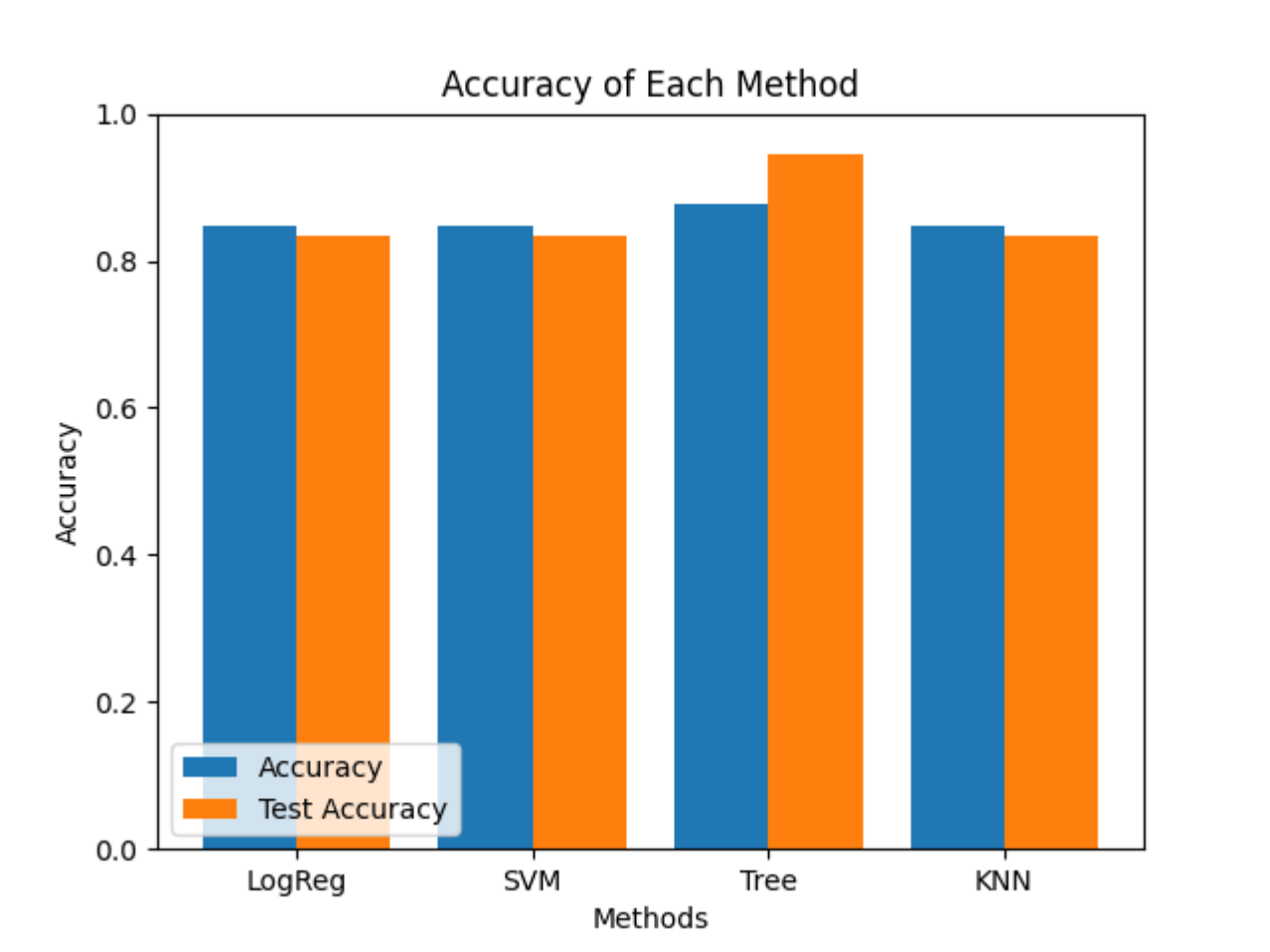


- Performance of Low weighted payloads is better when compared with the heavy weighted payloads.
- Payloads below 7000 kgs are most Successful.
- Payloads over 7000 kgs doesn’t have enough data and is not successful.

Section 5

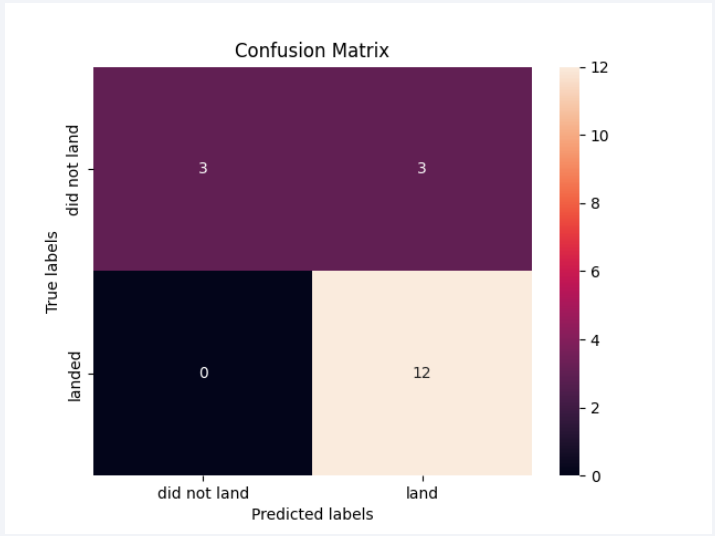
Predictive Analysis (Classification)

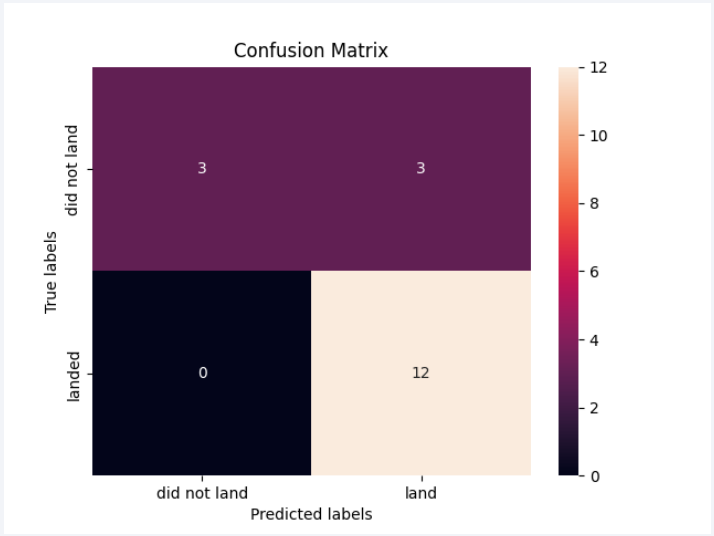
Classification Accuracy of the methods

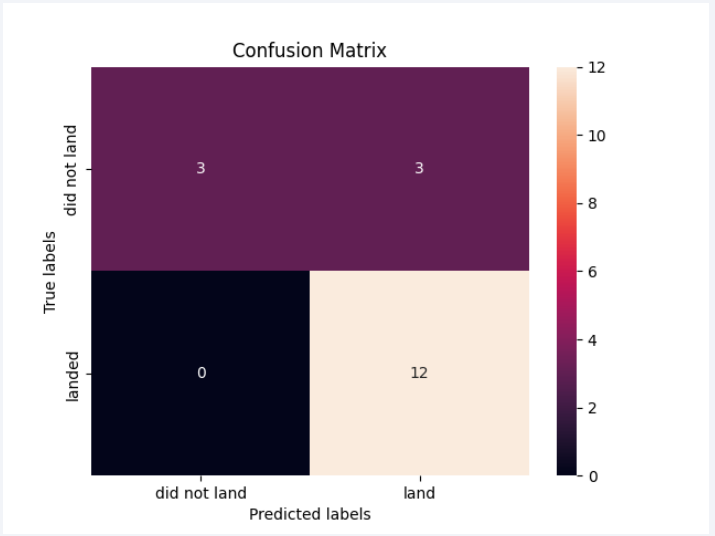


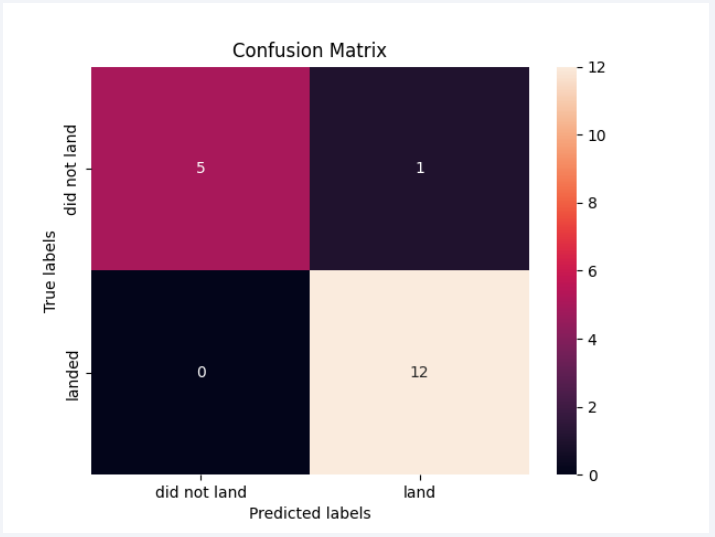
- Four Methods were tested on basis of their accuracies and test accuracies
- Tree Classifier Method has the highest classification accuracy which is 88.9% accurate.

Confusion Matrix of the accuracy methods









- Confusion matrix of the Decision Tree Classifier has the best performing model

Conclusions from the Capstone Project

- The best in terms of prediction accuracy are the SVM, KNN, and Logistic Regression models for this dataset.
- Performance of Low weighted payloads is better when compared with the heavy weighted payloads.
- The perfection in the launches is due to the direct proportionality of success rates to the time in years for SpaceX launches.
- The most successful launches is by KSC LC 39A when compared from all the sites.
- The best Success Rate is for the Orbits HEO, ES L1, SSO, GEO.
- Tree Classifier is best for prediction of successful profits

Thank you!