IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Sahani Guru Prasad
5 January 2024

# Outline

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

# Executive Summary

## Summary of methodologies

- Data Collection Using API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis Using SQL
- Exploratory Data Analysis with Visualization
- Interactive Visual Analytics with Folium
- Interactive Dashboard with Plotly Dash
- Predictive Analysis Using Machine Learning

## Summary of all results

- Exploratory Data Analysis Results
- Interactive Analytics results in Screenshots
- Predictive Analysis Results

# Introduction

## Project background and context

SpaceX is an aerospace company that has changed the dynamics of space industry by introducing a reusable rocket by advertising Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. As a Data Scientist of rivaling company the goal is to create a Machine Learning pipeline which can predict the landing outcome of first stage of falcon 9 by which the company can identify the right price to bid against SpaceX at rocket launch.

## Problems to find answers

- To Find the Factors that influence the Landing Outcome
- To Find the Relationship between Success Rate of each Orbit Type
- To Find the Best Machine Learning Algorithm in Predictive Analysis for Landing Prediction

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Data Collection from the following Sources:
  - ➢ Using SpaceX REST APIs
  - ➢ Web Scraping of SpaceX data from Wikipedia.
- Perform data wrangling
  - Data is processed using different techniques like filtering, dealing with missing data and encoding the data for binary classification where 0 – bad outcomes, 1- rest other outcomes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Exploring and finding insights from data using plots, charts and SQL queries.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning and evaluating the classification models to find the best performing model for prediction.

# Data Collection

In this project the data has been collected using SpaceX REST APIs and Web scraping of Wikipedia website.

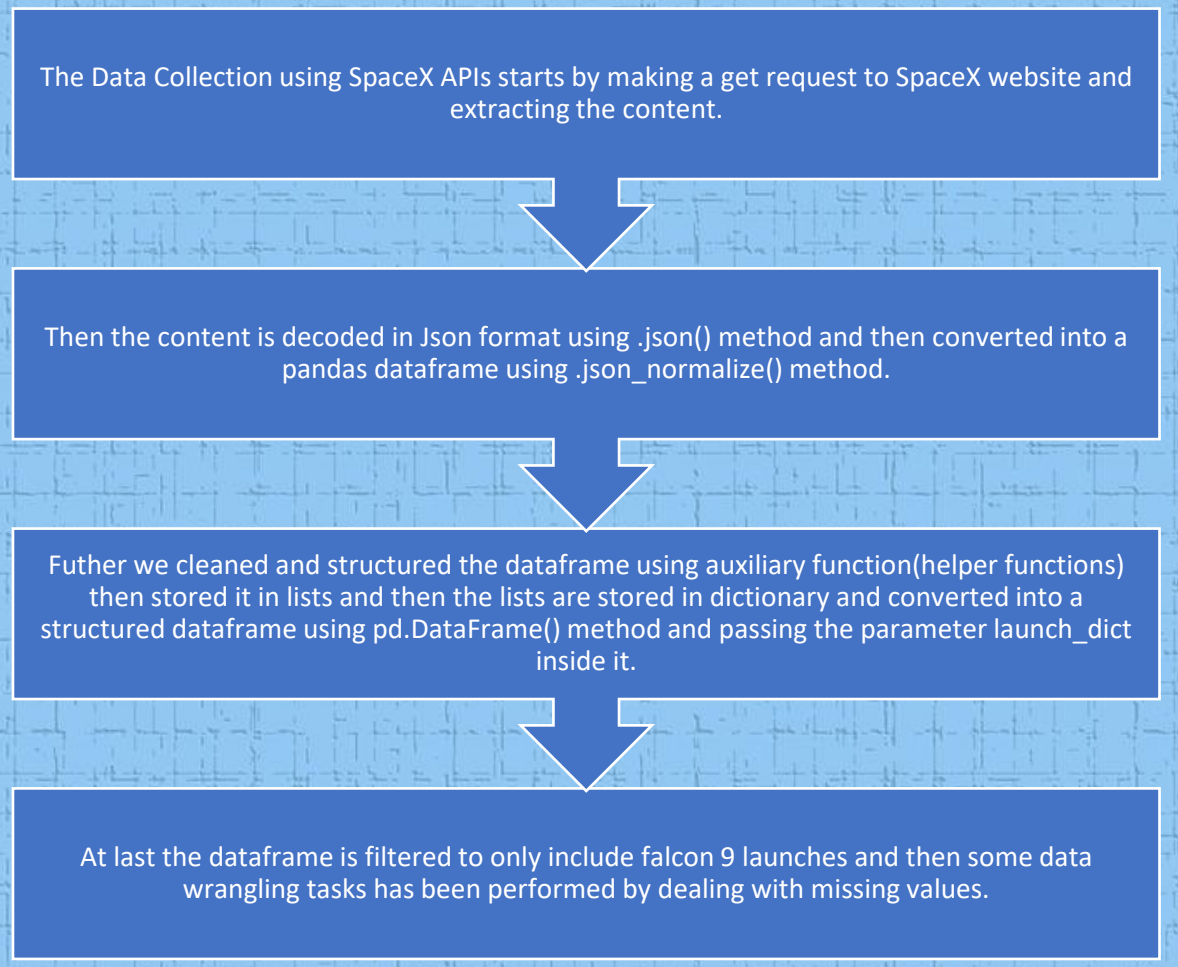
The process get started with SpaceX REST APIs by making a get request and extracting the content. Then the content is decoded in Json format using .json() method and then converted into a pandas dataframe using .json_normalize() method further in this process some data wrangling tasks has been performed.

In web scraping the data has been collected from Wikipedia where the launch records data is extracted using beautifulsoup as HTML tables and then converted into a pandas dataframe.

# Data Collection – SpaceX API

- The Data Collection using SpaceX API is presented in flowchart.

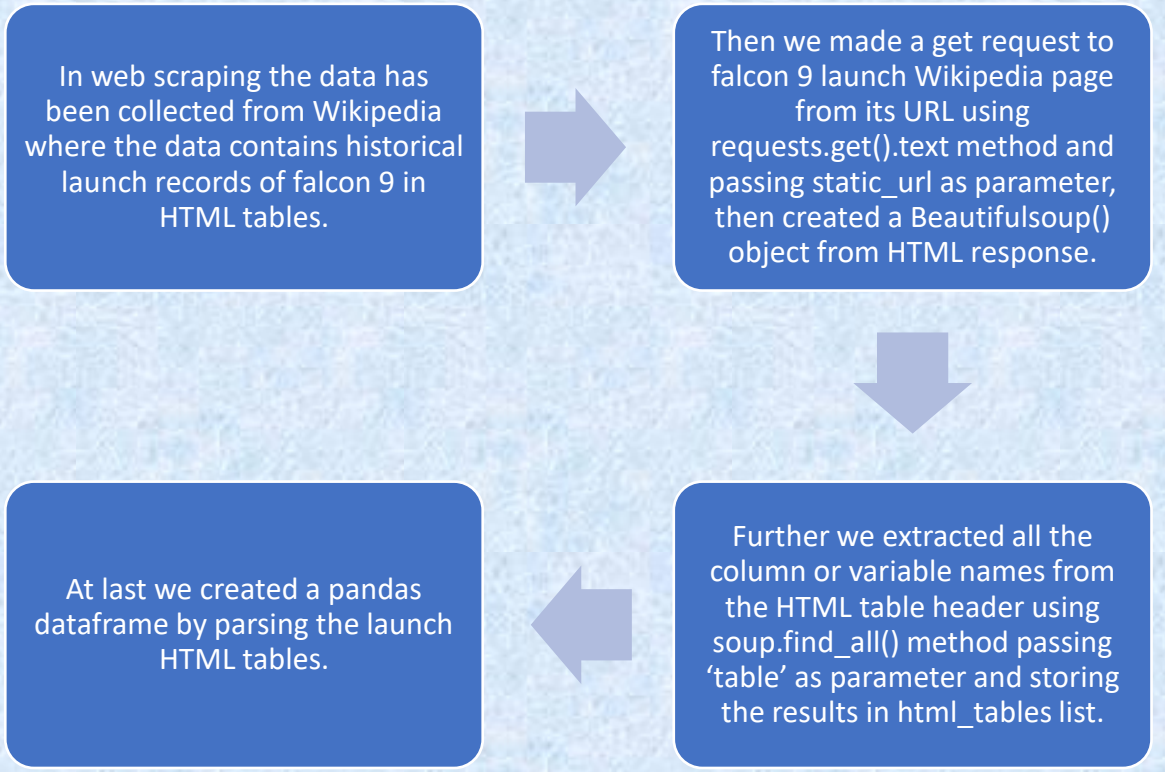
- Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/Data%20Collection%20using%20APIs.ipynb



# Data Collection - Scraping

- Data Collection with Web Scarping is presented using flowchart.

- Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/Data%20Collection%20using%20Web%20Scraping.ipynb



# Data Wrangling

In Data Wrangling, we performed some Exploratory Data Analysis(EDA) to find out some patterns in the data in order to identify and determine the labels suitable for training supervised models.

The process starts with loading and cleaning the dataset by identifying the percentage of missing values in each attributes and verifying the columns with their appropriate datatypes such as numerical or categorical.

Further we calculated the number of launches on each site, occurrence of each orbit and occurrence of mission outcomes of the orbits with the columns ‘Launch site’, ‘Orbit’ and ‘Outcome’ using .value_counts() method.

At last we classify the outcomes using binary classification where bad outcomes are denoted by 0 and rest all outcomes as 1.

Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/Data%20Wrangling.ipynb

# EDA with Data Visualization

In EDA with Data Visualization we used multiple charts and plots such as cat plot, scatterplot, bar chart and line chart using seaborn.

Cat plot shows the relationship between variable Flight Number and Payload Mass.

Scatterplot shows the relationship between variables such as Flight Number and Launch site, Payload Mass and Launch site, Flight Number and Orbit , Payload Mass and Orbit.

Bar chart shows the relationship between success rate of each orbit type.

Line chart shows the launch success yearly trend.

At last we performed feature engineering and one hot encoding on features ‘Orbit’, ‘Launch site’, ‘Landing pad’ and ‘Serial’.

Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb.jupyterlite.ipynb

# EDA with SQL

In EDA with SQL, we performed multiple tasks to explore the data using SQL. The tasks are listed below:

- Displayed the names of the unique launch sites in the space mission.
- Displayed 5 records where launch sites begins with the string 'CCA'.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed the average payload mass carried by booster version F9 v1.1
- Listed the data when the first successful landing outcome in ground pad was achieved.
- Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Listed the total number of successful and failure mission outcomes.
- Listed the names of the booster versions which have carried the maximum payload mass using a sub query.
- Listed the records which displays the month names, failure outcomes in drone ship, booster versions, launch site for the months in year 2015.
- Ranked the count of landing outcomes(such as Failure(drone ship) or success(ground pad)) between the data 2010-06-04 and 2017-03-20, in descending order.
- Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/EDA%20with%20SQL.ipynb

# Build an Interactive Map with Folium

While building an interactive map with folium we use several map objects such as markers, circles, lines, etc. are created and added to a folium map. Below is the description mentioned:

In the first task we mark all the launch sites on the map which allows us to see visually the launch sites on the map:

- By adding a maker with circle, popup label and a text label of NASA Johnson space center using its latitude and longitude coordinates as a start location.
- By adding Markers with circle, popup label and text label of all launch sites using their latitude and longitude coordinates to show their geographical locations and proximity to equator and coasts.

In task 2 we colour the markers of the launch outcomes for each launch sites:

- By adding colour markers of success as Green colour and failed as red colour launches using marker cluster by which we can identify which sites have high success rates.

In task 3 we find out the distance between launch site and its proximities:

- By adding the coloured lines to show the distances between the launch sites and its proximities.

Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.jupyterlite.ipynb

# Build a Dashboard with Plotly Dash

While building a Dashboard with Plotly Dash we used several plots/graphs and interactions that have added to the dashboard. The description is mentioned below:

| | |
|---|---|
| Made a Launch sites Dropdown list: | • Added a dropdown list by which the user can able to select any Launch site option. |
| Made a pie chart showing success launches of all or certain sites: | • Added a pie chart which shows the total successful launches count of all sites and also shows the success and failure counts for the sites if a specific launch site is choosen. |
| Made a slider of payload mass range: | • Added a slider by which the user can select payload range. |
| Made a scatterplot of payload mass and success rate for different Booster versions: | • Added a scatterplot to show the correlation between payload and Launch success. |

Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/spacex_dash_app.py

# Predictive Analysis (Classification)

In predictive analysis we built, evaluated, improved, and found the best performing classification model.The description is mentioned below:

We created a Numpy array from the column class in data using .to_numpy() method.

Then we standardize the data and done fitting and transforming it.

Then performed the model evaluation using Train Test Split.

Trained the model using multiple Machine Learning techniques such as K-Nearest Neighbour(KNN), Decision Trees, Logistic Regression and Support Vector Machine(SVM).

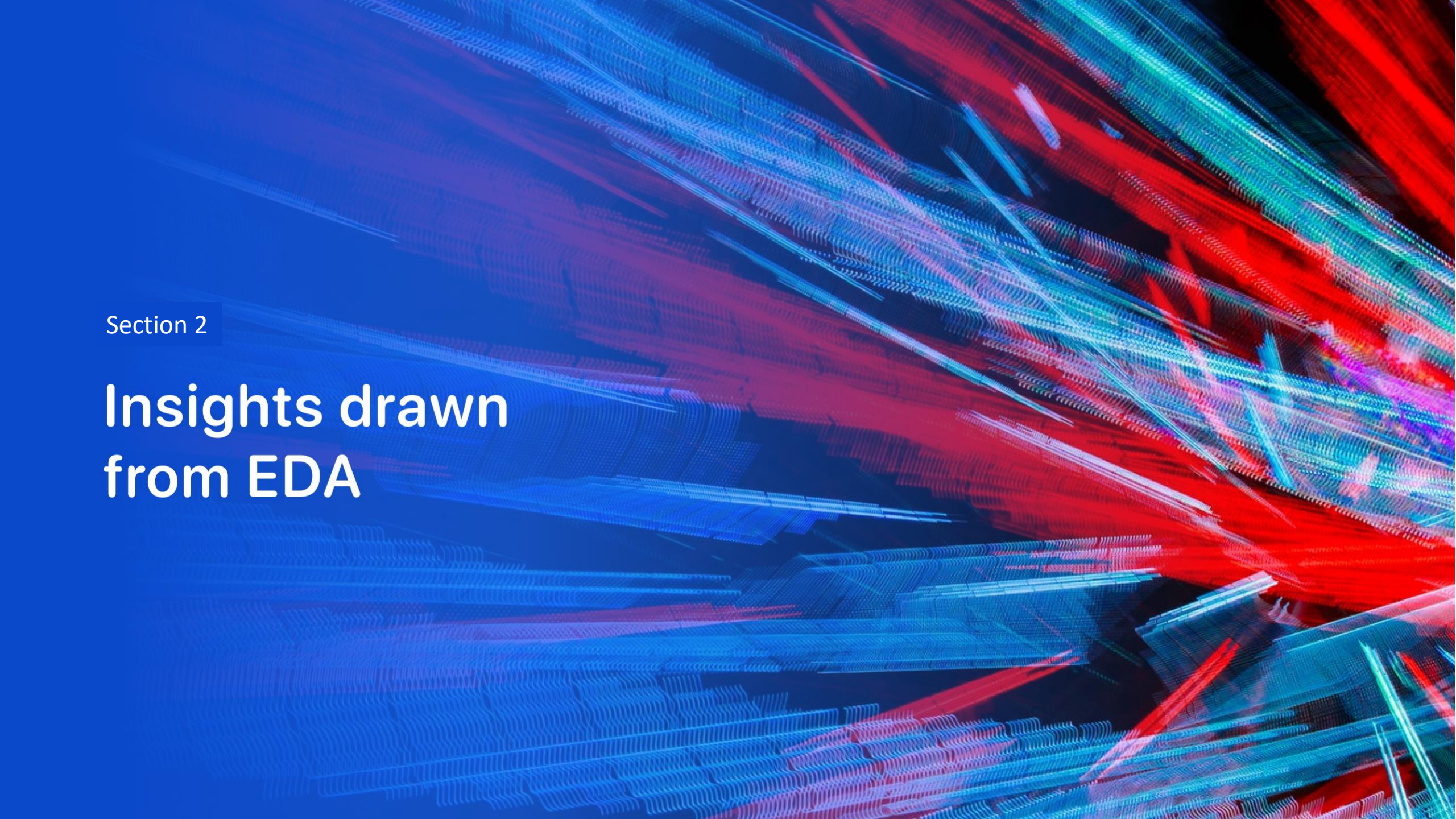
Calculated the accuracy of each technique and measured its predictions using confusion matrix.

Finally, we compared all Techniques to find out the best and optimum algorithm or technique among the multiple techniques for prediction using a bar chart.

Github URL : https://github.com/SahaniGuruPrasad/Applied-Data-Science-capstone/blob/main/Machine%20Learning%20Prediction.jupyterlite.ipynb

# Results

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

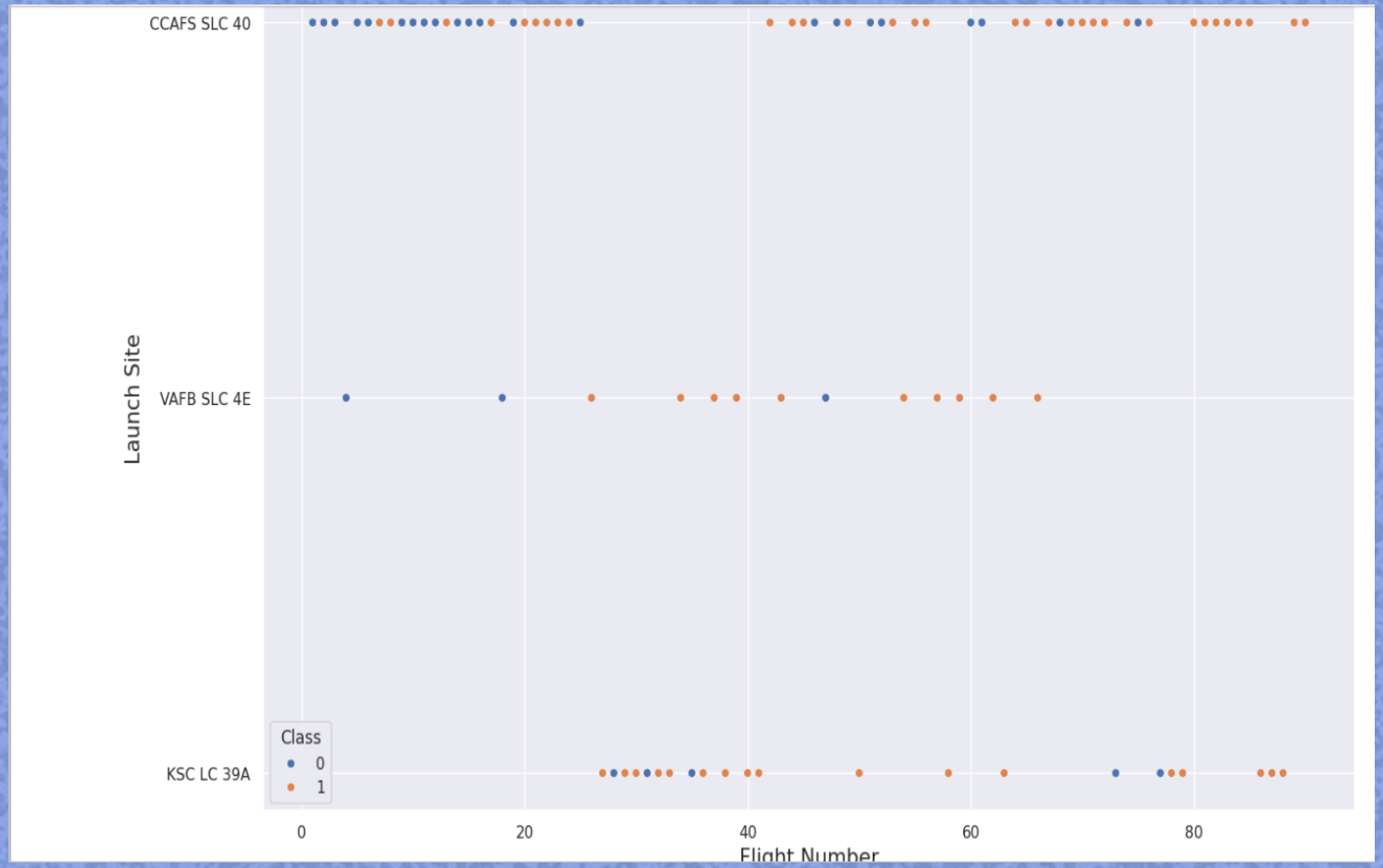

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site

The scatter plot shows the relationship between Flight Number and Launch site.

The Launch site CCAFS SLC 40 has higher number of successful launches as well as higher number of failure launches.

The Launch sites VAFB SLC 4E and KSC LC 39A has more number of success launches than their failure.



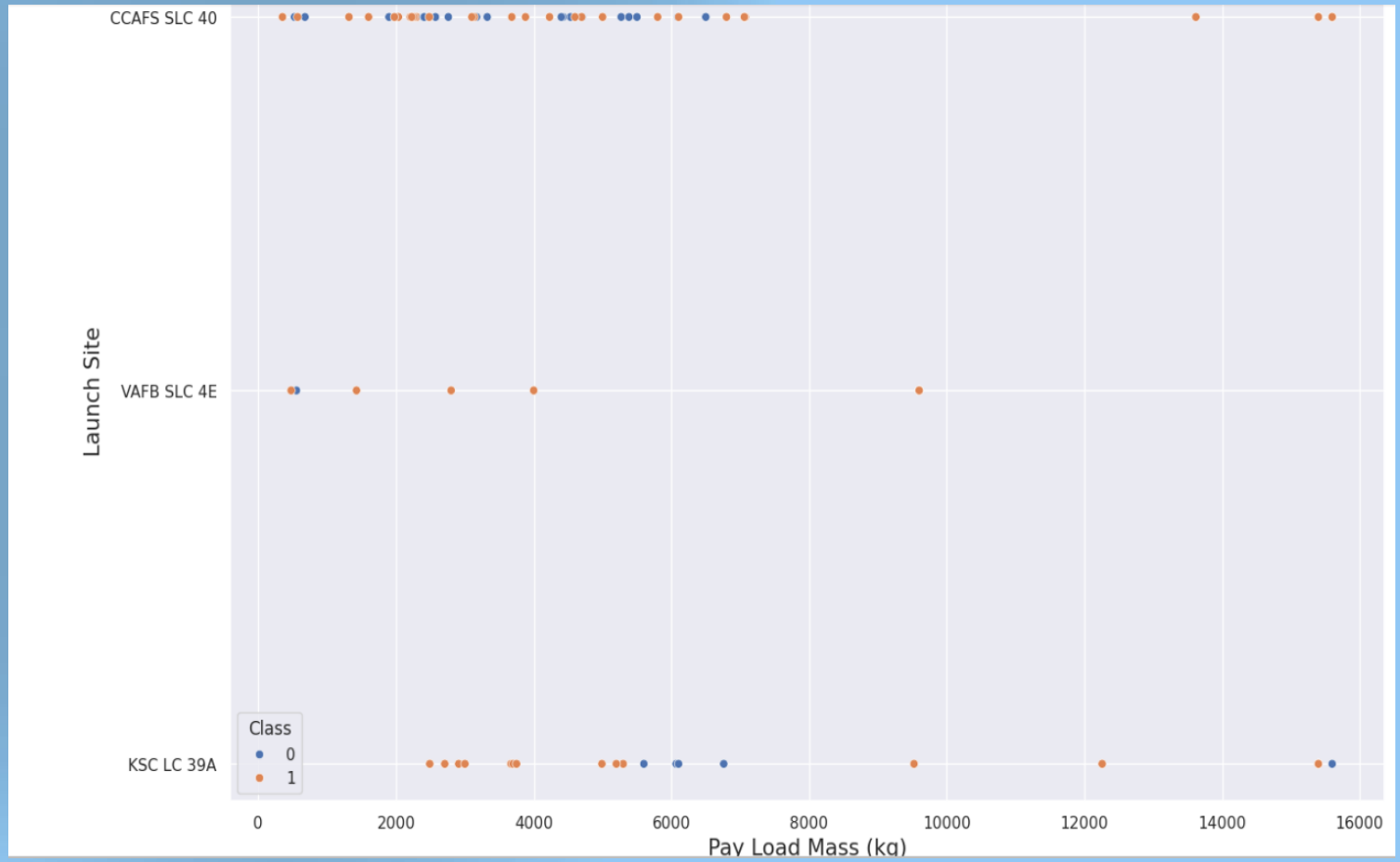
# Payload vs. Launch Site

The scatter plot shows the relationship between the Payload and Launch site.

The CCAFS SLC 40 Launch site has higher number of successful launches as well as higher number of failure launches.

CCAFS SLC 40 and KSC LC 39A Launch sites has more success rate than failure in heavy payload mass in kg.

VAFB SLC 4E has less failures among other Launch sites but has no rocket launches for heavy payload mass.



# Success Rate vs. Orbit Type

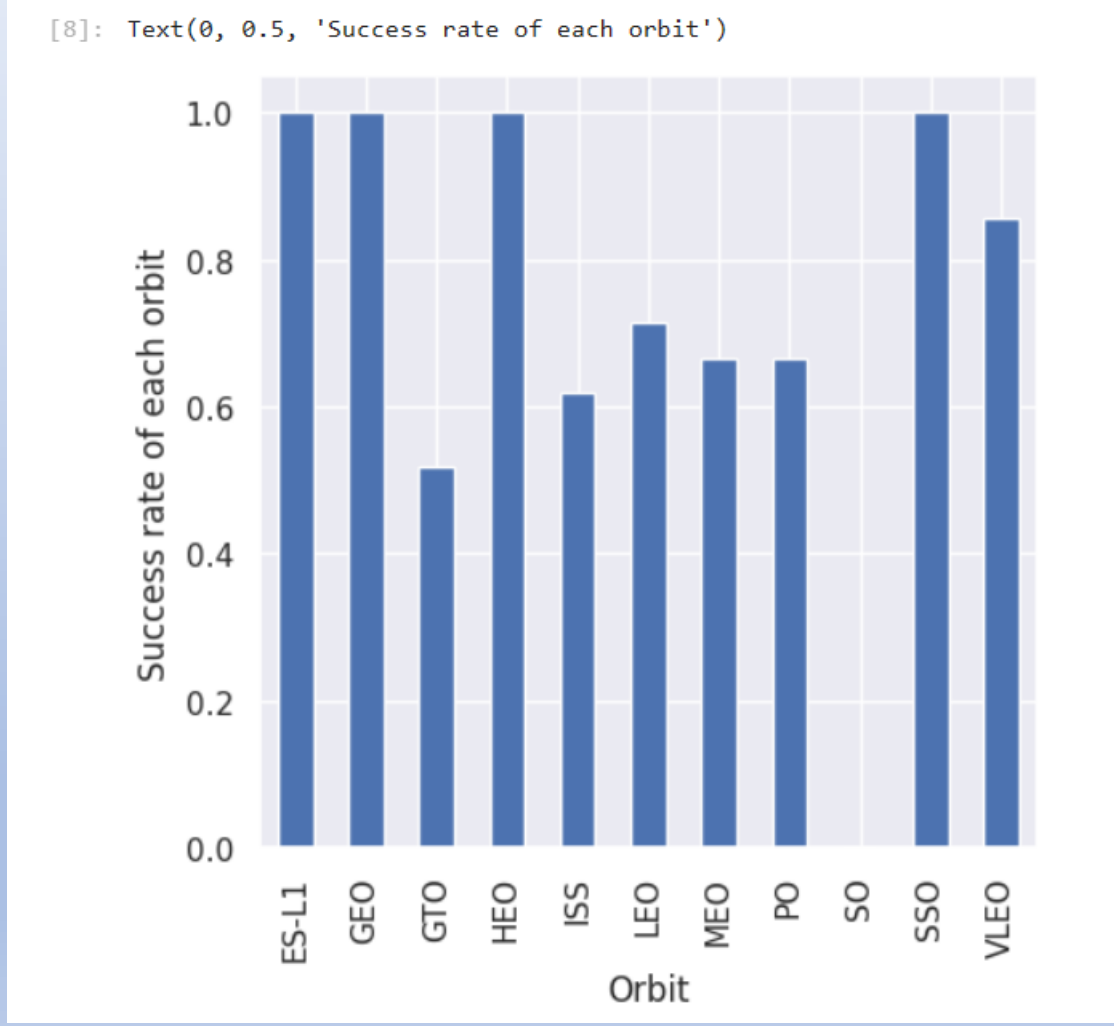
The bar chart shows the relationship between success rate of each orbit.

The orbits ES-L1, GEO, HEO and SSO has highest rate of success as 1.0

The orbits GTO, ISS, LEO, MEO and PO has success rate between 0.4 to 0.8

The orbit VLEO has success rate above 0.8

The orbit SO has lowest success rate among other orbits which has score 0.0

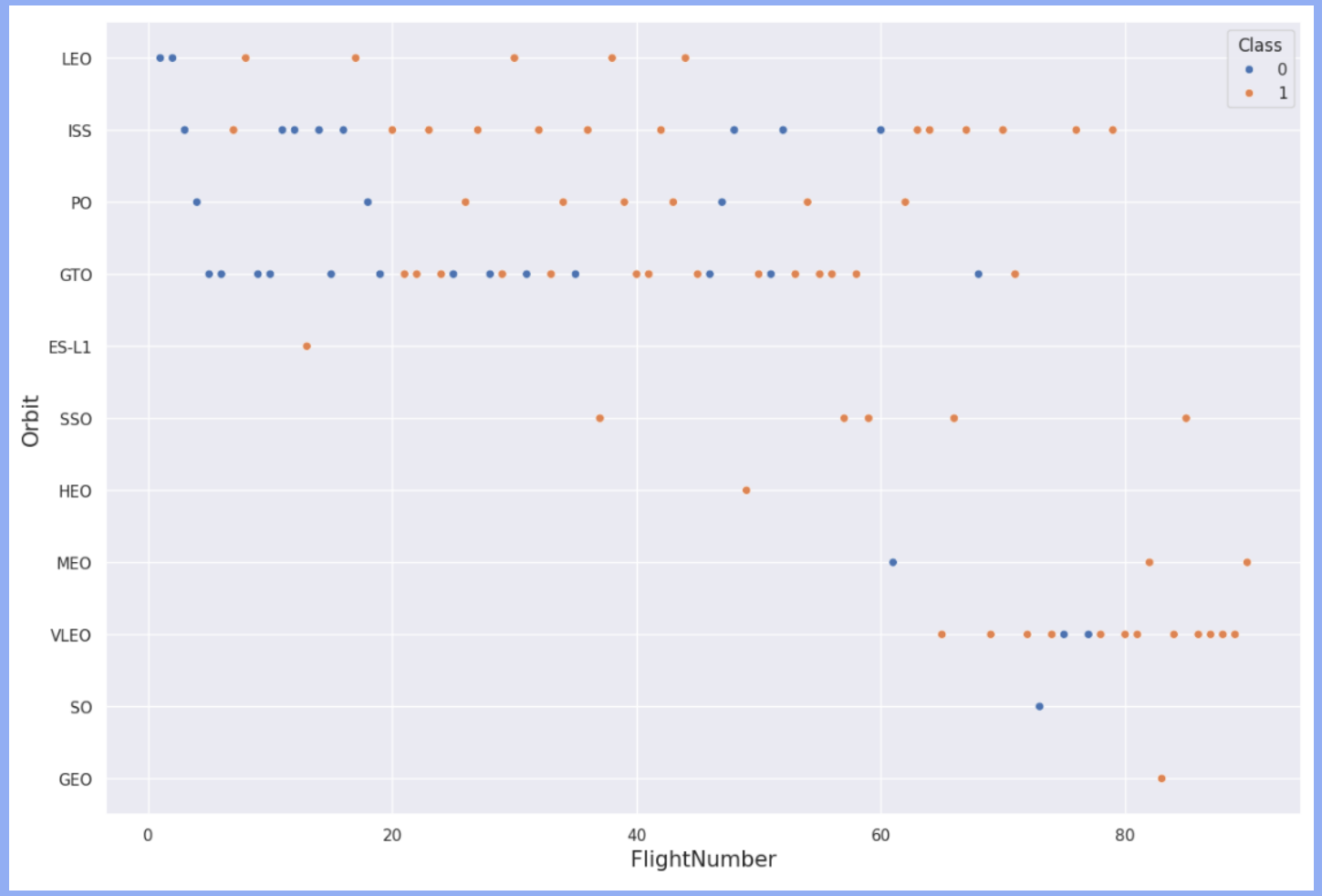


# Flight Number vs. Orbit Type

The scatter plot shows the relationship between Flight Number and Orbit type

The orbits LEO, MEO, VLEO and PO has less failure rate and orbits HEO, SSO and ES-L1 has no failure rates and orbit SO has no success launches.

The LEO orbits success appears related to the number of flights, on the other hand, there seems to be no relationship between fight number when it comes to GTO orbit.



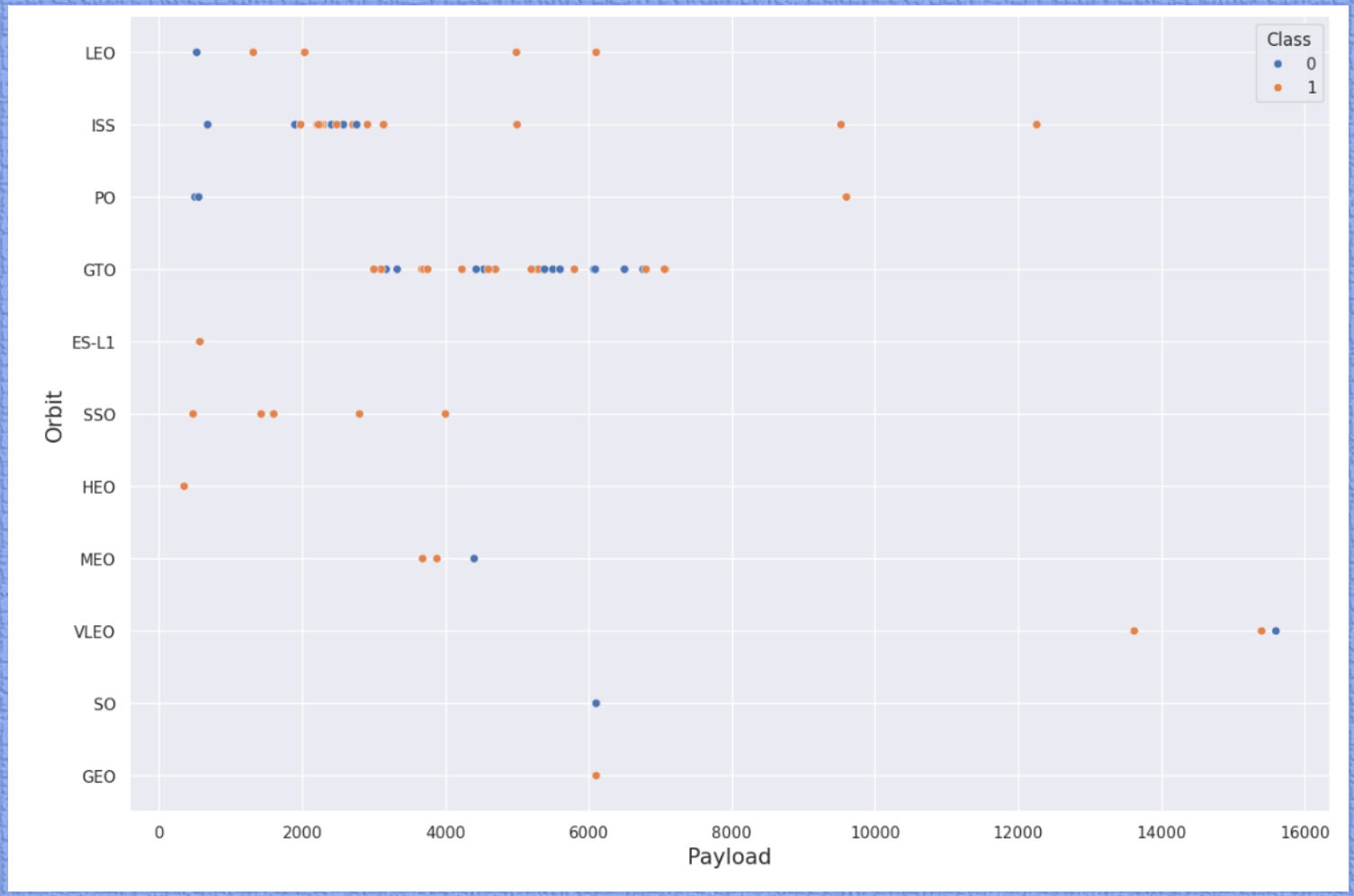
# Payload vs. Orbit Type

The scatter plot shows the relationship between Payload and Orbit type.

The orbits SSO, HEO, ES-L1, GEO has no failure rates.

The orbits VELO, MEO and LEO has less failure rate and SO has no success rate.

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



# Launch Success Yearly Trend


The line chart shows the launch success yearly trend.

The years between 2010 to 2013 their was no trend recorded as rate was 0.0

In years 2013 to 2015 the trend was between 0.0 to 0.4 rate.

In years 2015 to 2020 the trend was between 0.2 to 1.0 rate.

# All Launch Site Names

```
[49]: %sql SELECT DISTINCT LAUNCH_SITE AS "Launch Sites" FROM SPACEXTABLE;
```

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

[49]:

| Launch Sites |
| --- |
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |

The above SQL statement displays the Unique Launch sites records from SPACEXTABLE

The Launch sites were CCAFS LC – 40, VAFB SLC -4E, KSC LC -39A and CCAFS SLC -40

# Launch Site Names Begin with 'CCA'

This Query displays the 5 Launch Sites names which begins with ‘CCA’.

The Launch Site name was CCAFS LC -40.

```
[50]: %sql SELECT LAUNCH_SITE FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE "CCA%" LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

[50]:

| Launch_Site |
|---|
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

# Total Payload Mass

```
[51]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "TOTAL PAYLOAD MASS" FROM SPACEXTABLE WHERE CUSTOMER = "NASA (CRS)";
 * sqlite:///my_data1.db
Done.
[51]:
```

| TOTAL PAYLOAD MASS |
| --- |
| 45596 |

The Query displays the total payload mass from SPACEXTABLE where customer was ‘NASA’.

The total payload mass was 45596.

# Average Payload Mass by F9 v1.1

The Query displays the average payload mass carried by booster version F9 v1.1

The Average Payload Mass was 2928.4

```
[52]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "AVERAGE PAYLOAD MASS" FROM SPACEXTABLE WHERE BOOSTER_VERSION = "F9 v1.1";
 * sqlite:///my_data1.db
Done.
```

[52]:

| AVERAGE PAYLOAD MASS |
| --- |
| 2928.4 |

# First Successful Ground Landing Date

```
[53]: %sql SELECT MIN(DATE) AS "FIRST SUCCESFUL LANDING OUTCOME" FROM SPACEXTABLE WHERE LANDING_OUTCOME = "Success (ground pad)";

 * sqlite:///my_data1.db
Done.

[53]: FIRST SUCCESFUL LANDING OUTCOME
      2015-12-22
```

The Query displays the dates of the first successful landing outcome on ground pad.

The Date of the First Successful Landing Outcome was 2015-12-22.

# Successful Drone Ship Landing with Payload between 4000 and 6000

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

The Booster Version were F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1031.2

[54]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE LANDING_OUTCOME = "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

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

[54]:

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

# Total Number of Successful and Failure Mission Outcomes

```
[61]: %sql SELECT COUNT(MISSION_OUTCOME) AS "TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSIONS" FROM SPACEXTABLE \
      WHERE MISSION_OUTCOME LIKE "Success%" OR MISSION_OUTCOME LIKE "Failure%";
```

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

[61]:

| TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSIONS |
|---|
| 101 |

The Query displays the total number of successful and failure mission outcomes

The Total Number of Successful and Failure Missions are 101 where number success missions were 100 and failure missions were 1.

# Boosters Carried Maximum Payload

This Query displays the List of the names of the booster which have carried the maximum payload mass

The Booster Versions were F9 B5 B1048.4, F9 B5 B1049.4 till F9 B5 B1049.7

```
[62]: %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTABLE \
      WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

[62]:

| 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

```
[63]: %sql SELECT substr(Date,6,2) AS MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE \
      WHERE substr(Date,0,5)='2015' AND LANDING_OUTCOME = "Failure (drone ship)";
 * sqlite:///my_data1.db
Done.
```

[63]:

| 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 |

This Query lists out the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

The Month number was 01 and 04, Landing Outcome was Failure(drone ship), Booster Version was F9 v1.1 B1012 and F9 v1.1 B1015 and Launch Site was CCAFS LC -40

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

```
[64]: %sql SELECT LANDING_OUTCOME, COUNT(LANDING_OUTCOME) AS "TOTAL COUNT OF LANDING OUTCOMES" FROM SPACEXTABLE \
      WHERE DATE BETWEEN "2010-06-04" AND "2017-03-20" \
      GROUP BY LANDING_OUTCOME \
      ORDER BY COUNT(LANDING_OUTCOME) DESC;
```

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

[64]:

| Landing_Outcome | TOTAL COUNT OF LANDING OUTCOMES |
|---|---|
| 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 |

This Query displays the Rank 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

Section 3

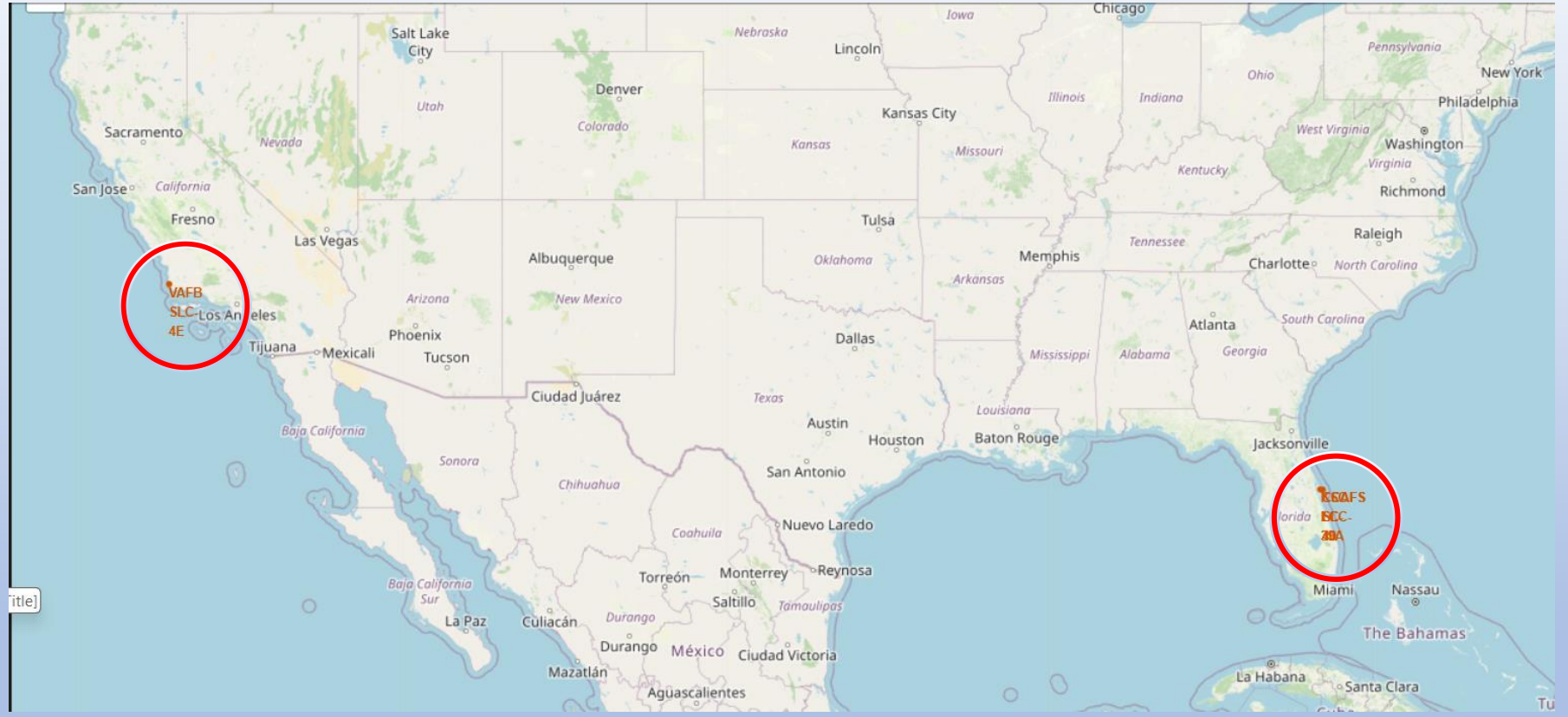
# Launch Sites Proximities Analysis

# Location of All Launch Sites on Map

The Global Map shows all launch sites location of SpaceX falcon 9 rocket launches.

Upon exploring the map we found that VAFB SLC-4E launch site is located near the western coastline while other launch sites like KSC LC-39A, CCAFS LC-40 and CCAFS SLC-40 are located near eastern coastline.

Upon zooming the map near launch sites at eastern coastline we found that launch sites CCAFS LC-40 and CCAFS SLC-40 are located very close to each other.



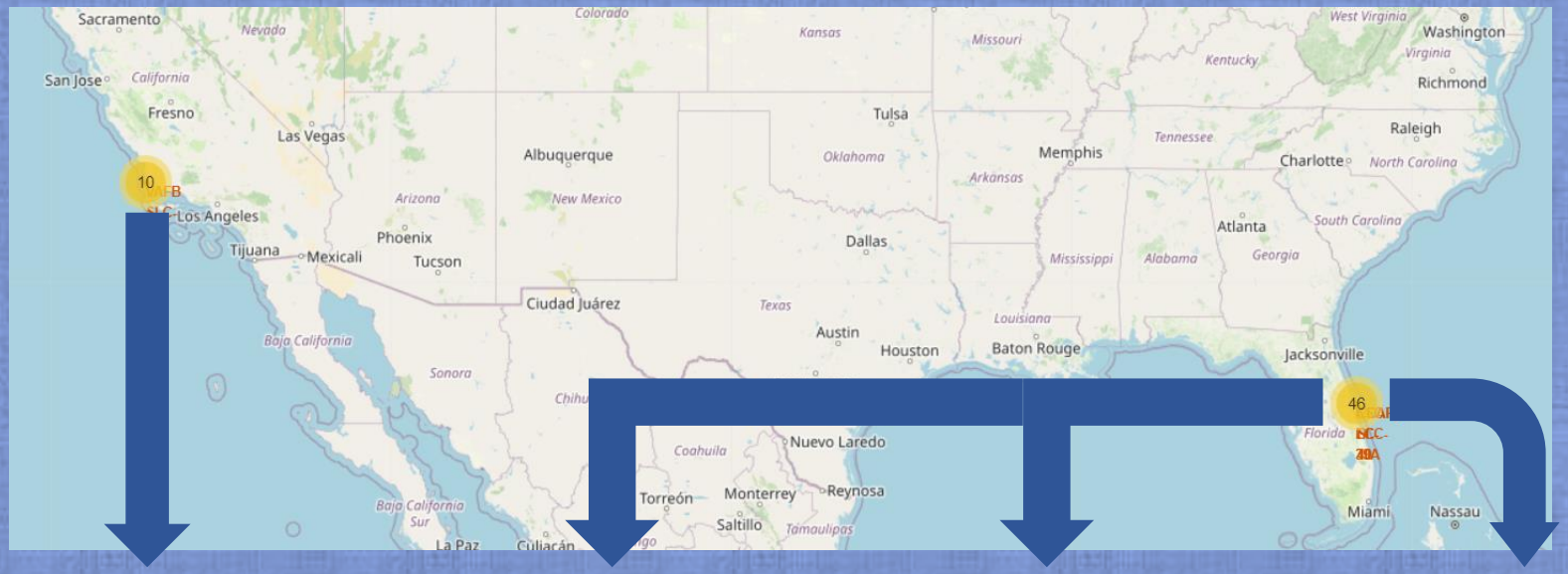
# Colour-Labeled Launch Outcomes of Launch Sites

The Global Map shows the colour-labeled launch outcomes on the map.

In the map the colour-labeled markers represents the launch outcomes

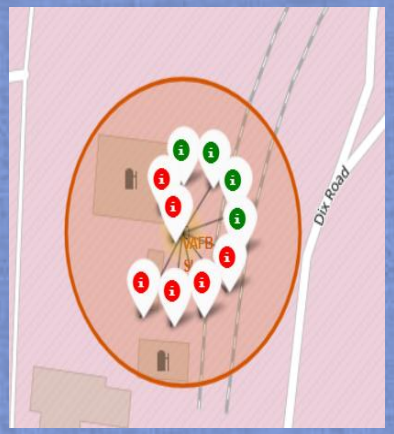
Red Marker shows Failed Lauches and Green Marker shows Success Launches.

The Launch site KSC LC-39A has highest launch success rate while CCAFS LC-40 has highest launch failure rate.

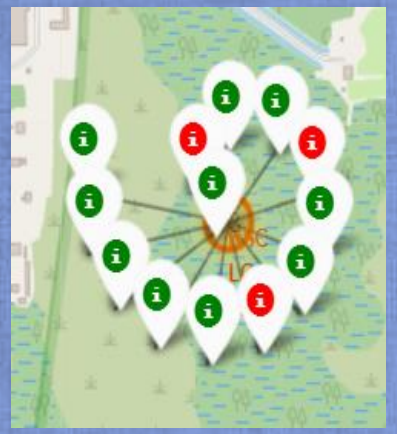



Failed Launch Outcomes

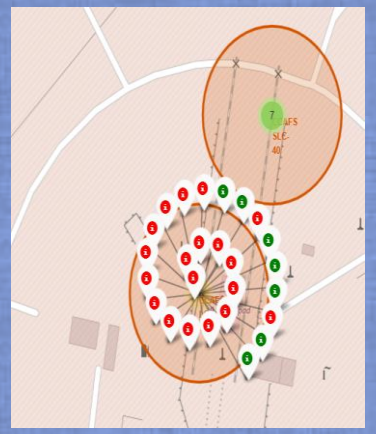
Success Launch Outcomes



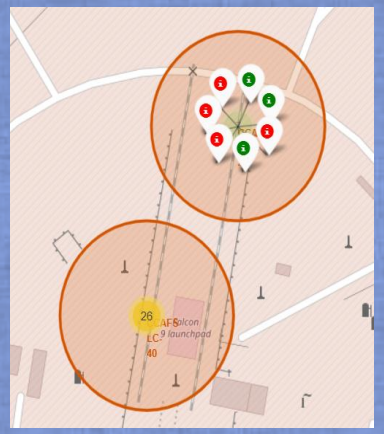

VAFB SLC-4E



KSC LC-39A

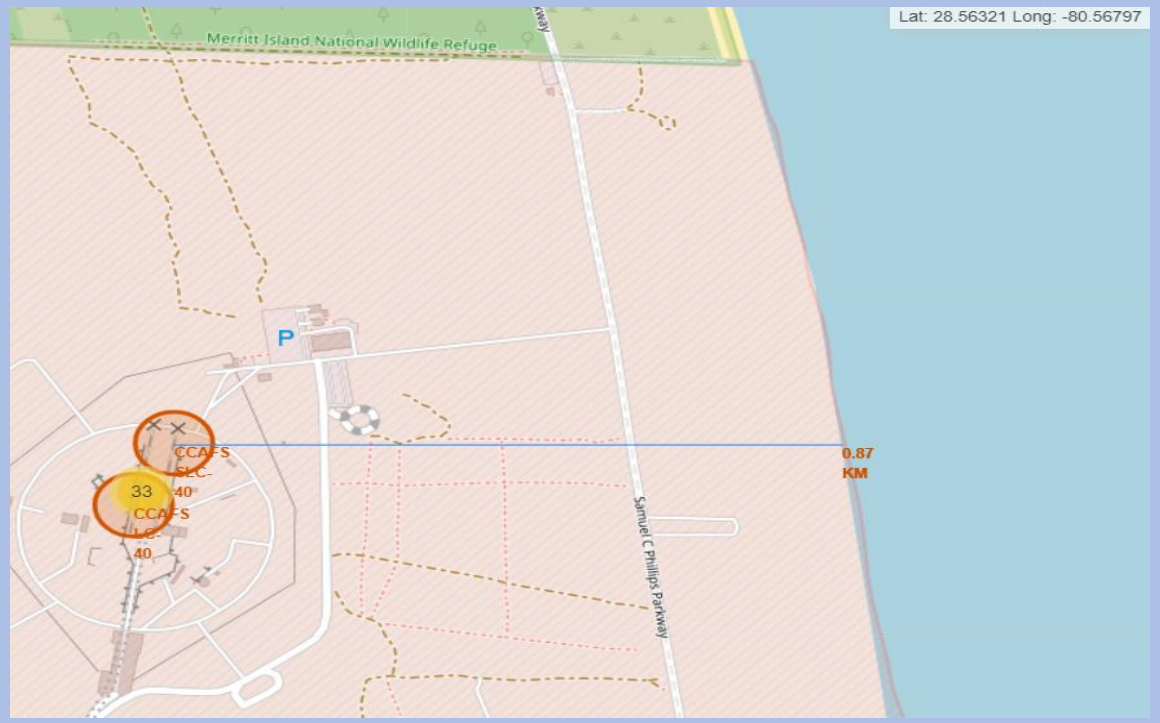


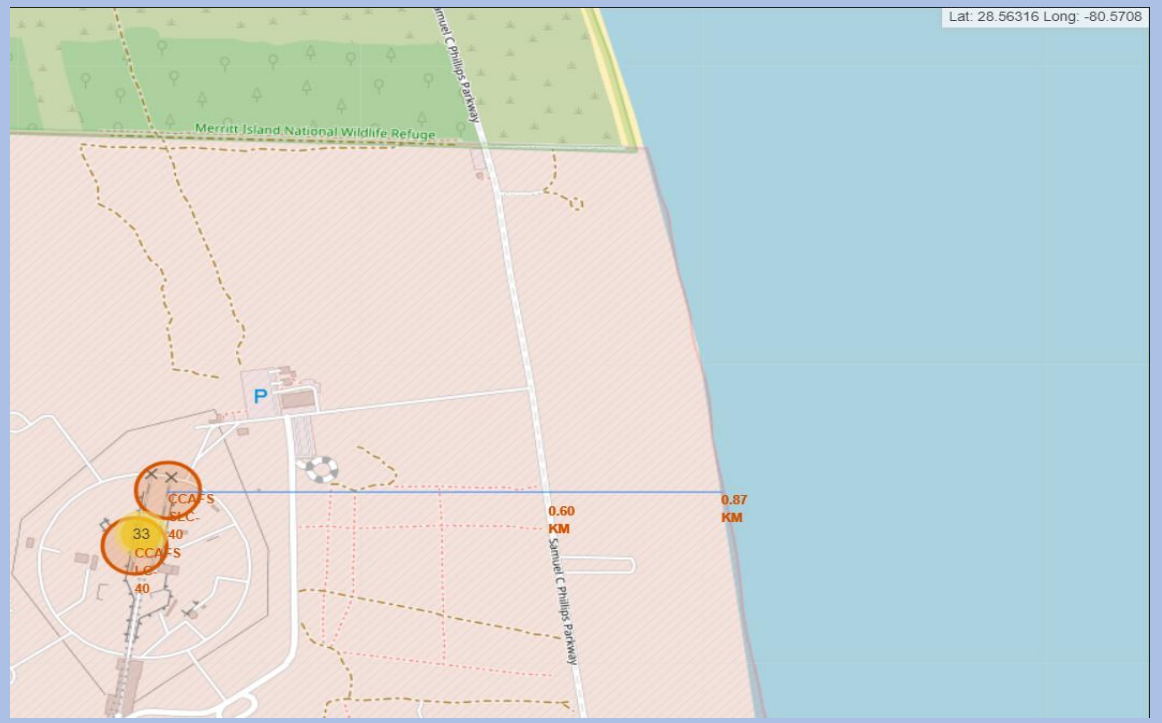
CCAFS LC-40



CCAFS SLC-40

# Selected Launch Site to its Proximities

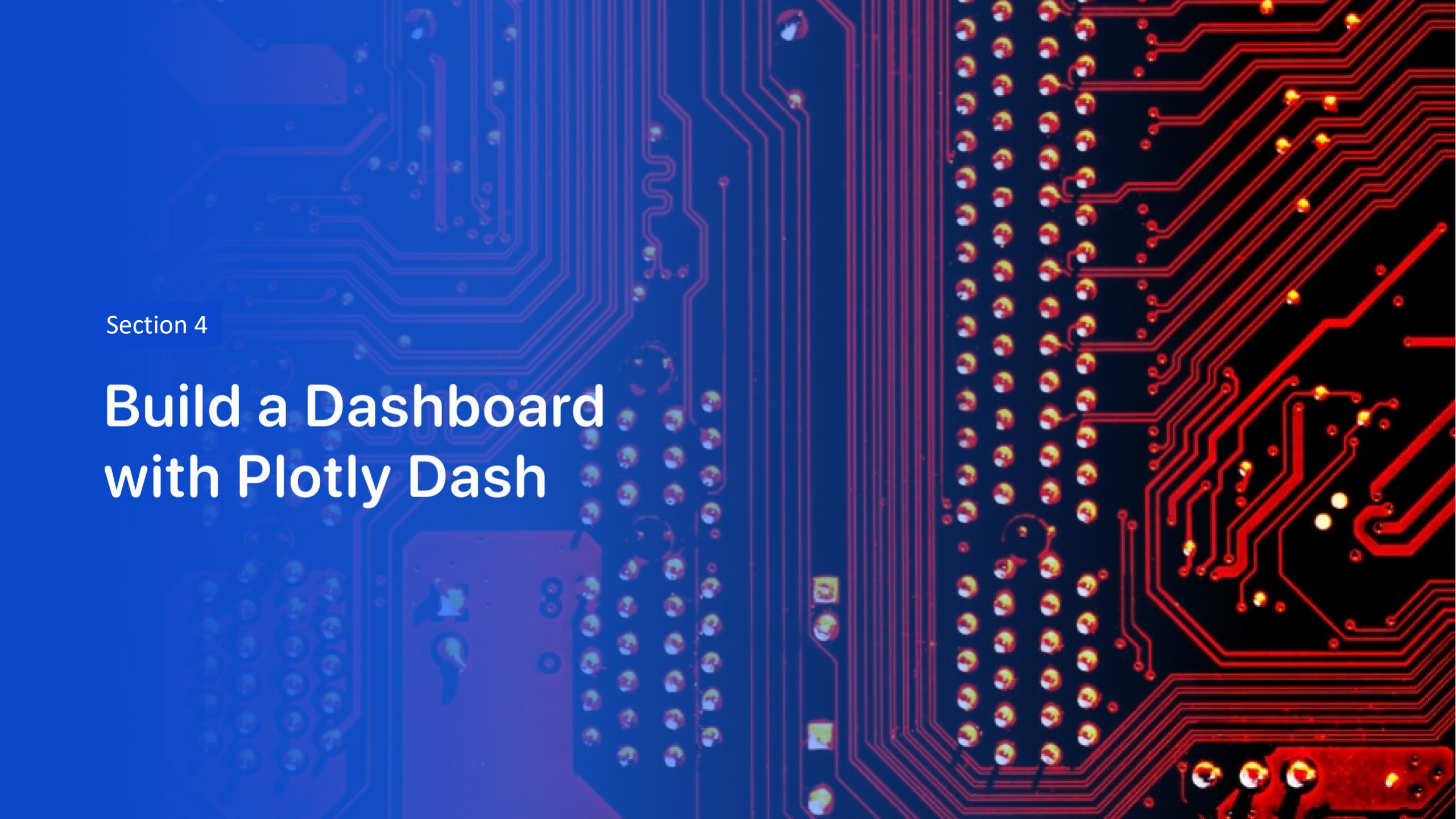


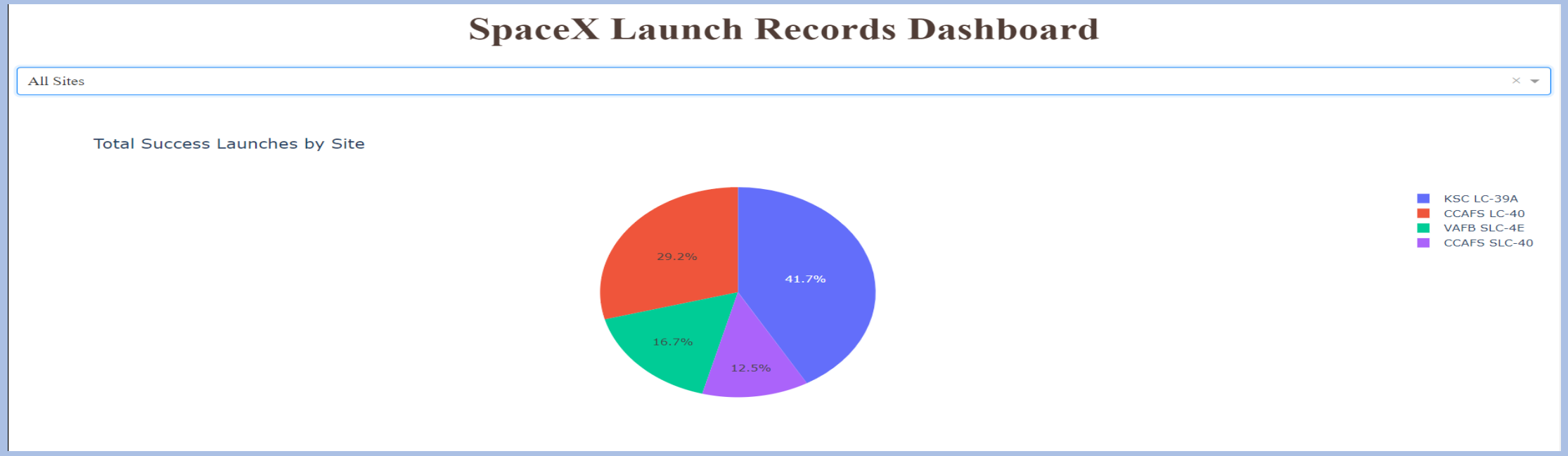

The above two pictures of map shows the selected launch site to its proximities such as railway, highway or coastline with distance calculated.

The selected launch site CCAFS SLC-40 is 0.87 km away from the coastline of eastern coastline.

The Latitude of coastline is 28.56321 and Longitude of coastline is -80.56797.


Section 4
Build a Dashboard with Plotly Dash

# Launch Success count for all sites




The Pie chart shows the Launch Success count for all sites of SpaceX Launch records in the Dashboard.

We can clearly see that KSC LC-39A Launch Site has most successful launches among all other sites where the Launch count percentage is 41.7%.

While other Launch sites like CCAFS LC-40 having 29.2% (Second highest) , VAFB SLC-4E (Third highest) having 16.7% and CCAFS SLC-40 having 12.5% which is lowest among all other launch sites.

# Launch Site with Highest Launch Success Ratio


The Pie chart shows the Launch site with highest launch success ratio among all the launch sites in the dashboard.

Among all the launch sites, KSC LC-39A launch site has highest launch success rate when compare to other launch sites which is 76.9%

# Payload vs Launch Outcome for all Launch sites


The Scatterplot shows the correlation between Payload Mass and launch outcome for all launch sites.

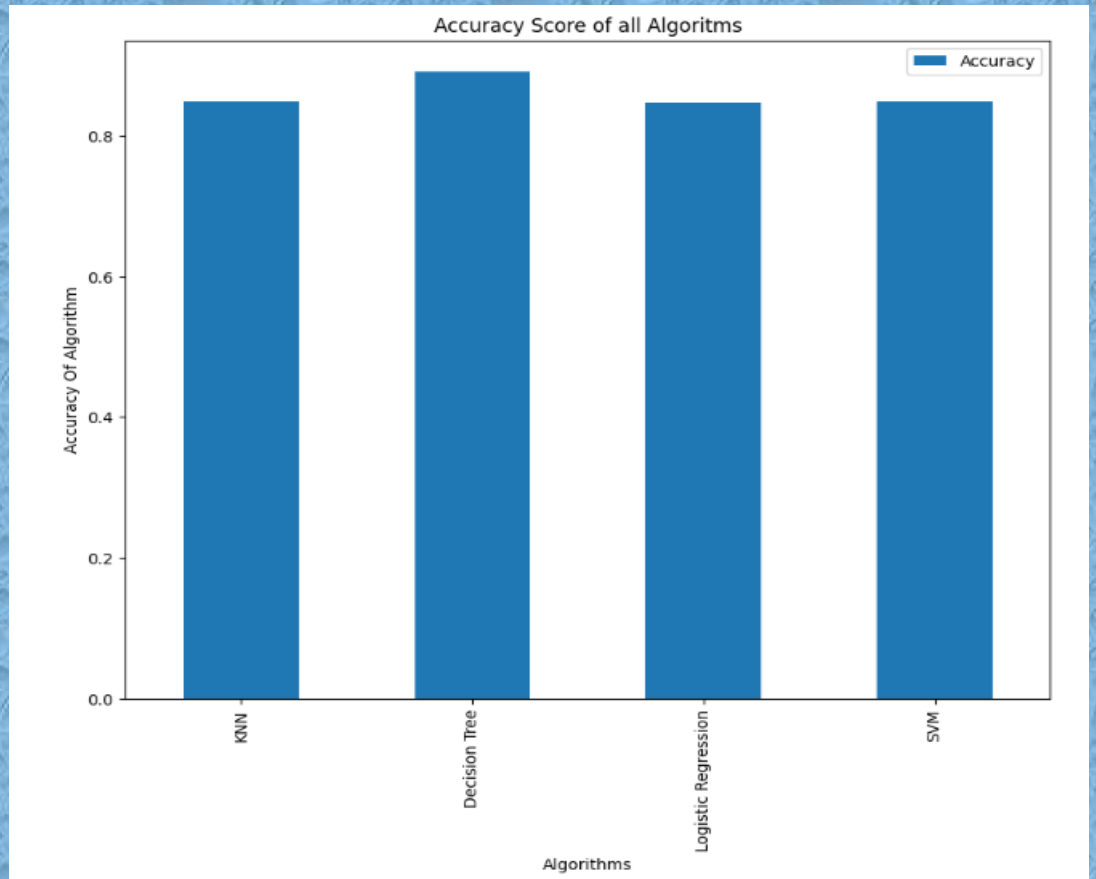
In the scatterplot, we see that the Payload Mass between 2000 kg to 5500 kg has highest launch success rate among all other payload masses.

Section 5

# Predictive Analysis (Classification)

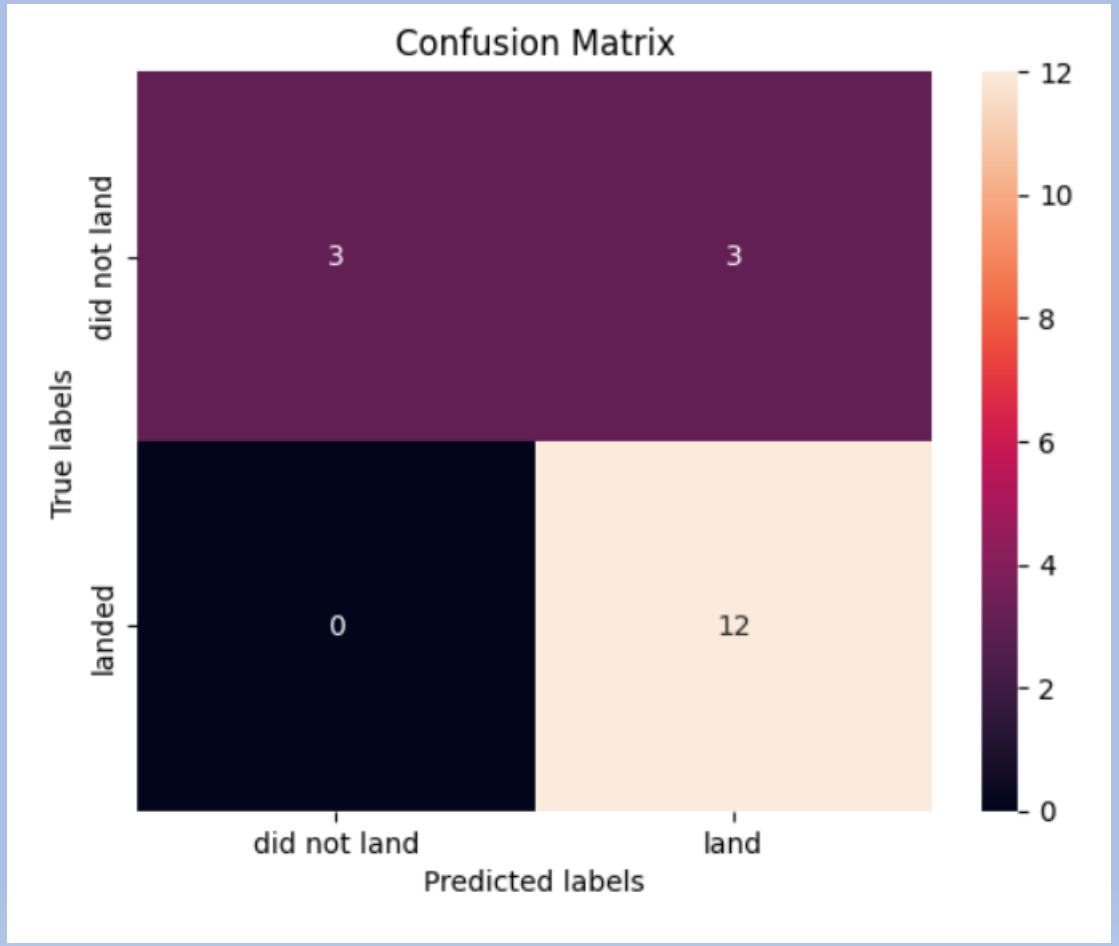
# Classification Accuracy

- The bar chart shows the Accuracy results of all model.
- Among all the classification models the Decision Tree model performs better than other models.
- The Accuracy Score of Decision tree is 0.8910714285714286 which is near to 89%.



# Confusion Matrix

- This Confusion Matrix shows the predicted labels performed by the best classification model which is Decision tree in this project.
- The model correctly predicted the results of True Positives, True Negatives and False Negatives.
- The model incorrectly predicated the results of False Positives by misclassifying 3 labels.



# Conclusions

Here are some Conclusions drawn from this project.

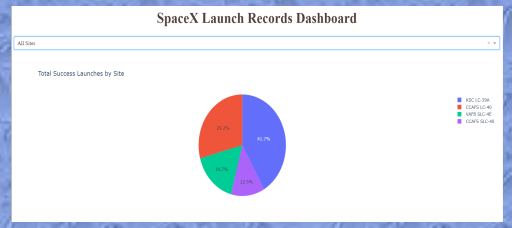
- The orbits ES-L1, GEO, HEO and SSO has highest rate of success as 1.0
- With heavy payloads the successful landings or positive landings rate are more for Polar, LEO and ISS.
- The launch success trend increases over every year.
- The launch site KSC LC-39A has highest launch success rate while CCAFS LC-40 has highest launch failure rate.
- Most of the launch sites are situated near the coastlines.
- The payload mass which are low in weight has higher success rate than heavier payload mass.
- The Decision tree classification model performs better for prediction compare to other classification models.

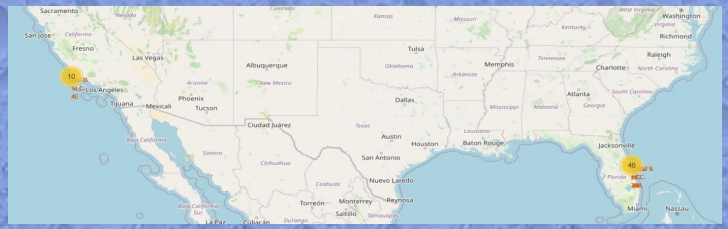
# Appendix

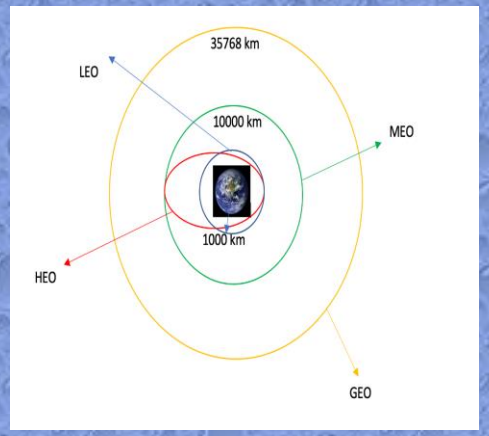
- In this project we find out several insights from the data by performing multiple data analysis methodologies like EDA with data visualization and SQL queries, dashboards, interactive maps, predictive analysis etc. Few of the code snipets and results are mentioned below.

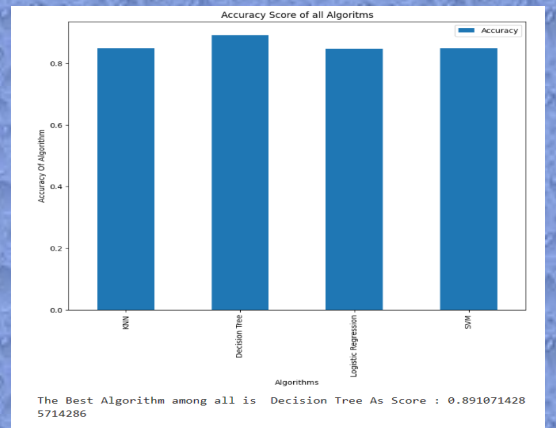
```
[61]: %sql SELECT COUNT(MISSION_OUTCOME) AS "TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSIONS" FROM SPACEXTABLE \
      WHERE MISSION_OUTCOME LIKE "Success%" OR MISSION_OUTCOME LIKE "Failure%";

       * sqlite:///my_data1.db
      Done.
[61]:
```

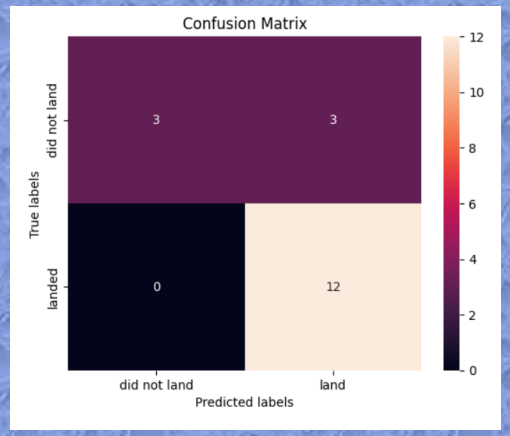
| TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSIONS |
|---|
| 101 |






# Thank you!