IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name : Ghadge Saurabh
Date : 06-01-2022

Outline

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

Executive Summary

- Summary of methodologies
  - ➢ Problem Understanding
  - ➢ Collection and Transformation Of Data
  - ➢ Visualization
  - ➢ Predictive Analysis
- Summary of all results
  - ➢ Visualization Results
  - ➢ Predictive Analysis Results

Introduction

- Project background and context

This is IBM Data science Professional certificate last course in which main focus on model building and deploying using concept we learned so Far. In this We predicted 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, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch

- Problems you want to find answers
- The main Focus is on What attributes influenced Launching class of rocket and How?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By space X rest API and from web scrapping through Wikipedia.
- data wrangling
  - Most of it Done using Python but also I used SQL for some extent.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected.

➢ For this Analysis data is collected through space X rest API and from web scrapping of Wikipedia.

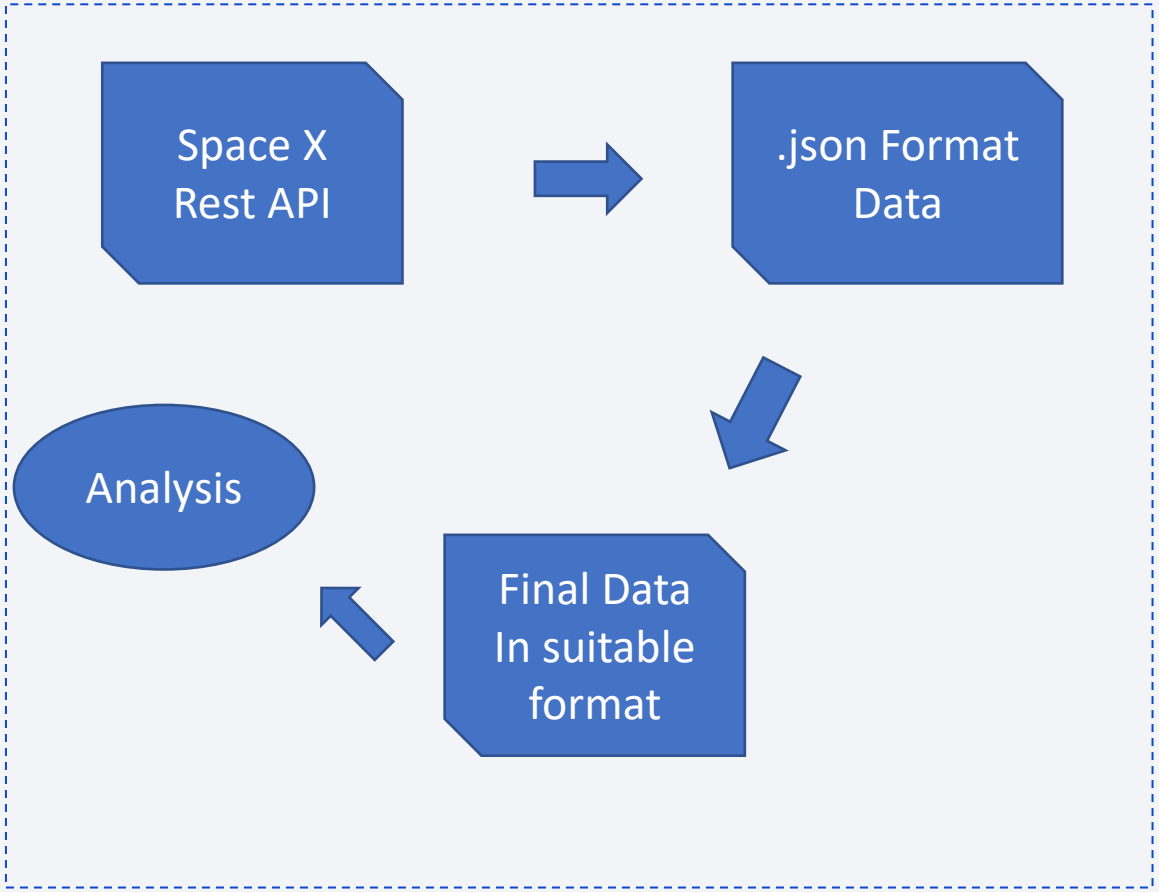
➢ This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

➢ The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

Data Collection – SpaceX API

- Data collection with SpaceX REST calls using key phrases and flowcharts

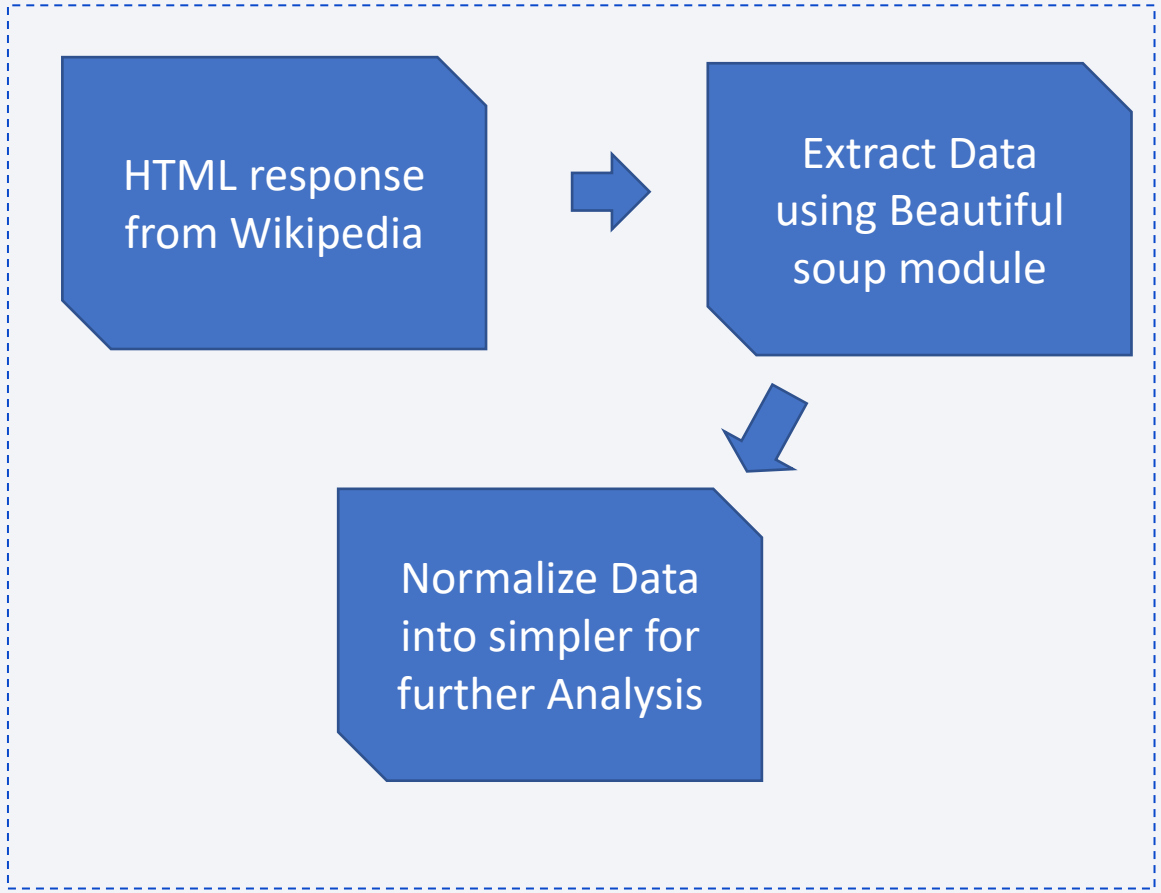
- GitHub URL of the completed SpaceX API calls notebook here



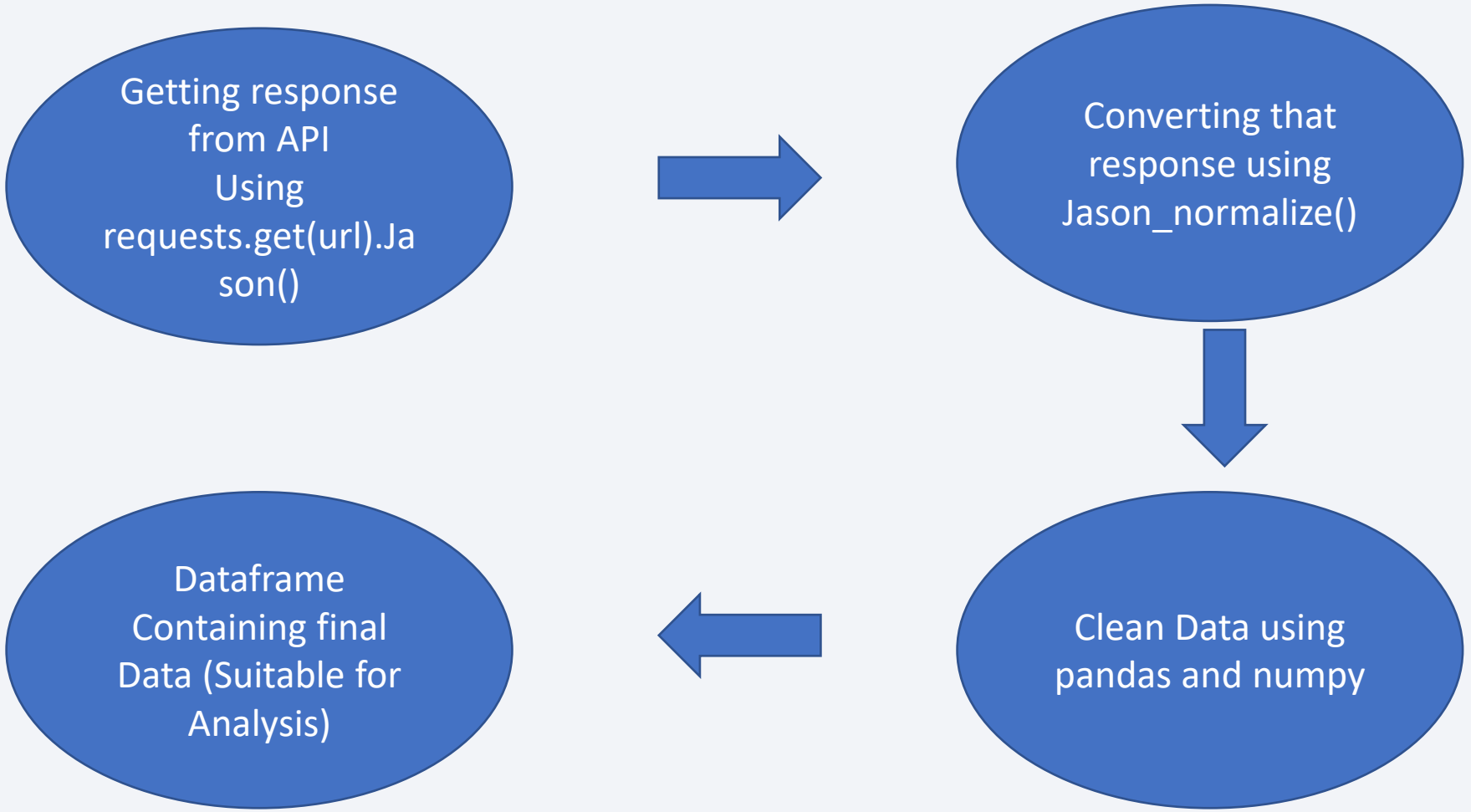
Data Collection - Scraping

- web scraping process using key phrases and flowcharts

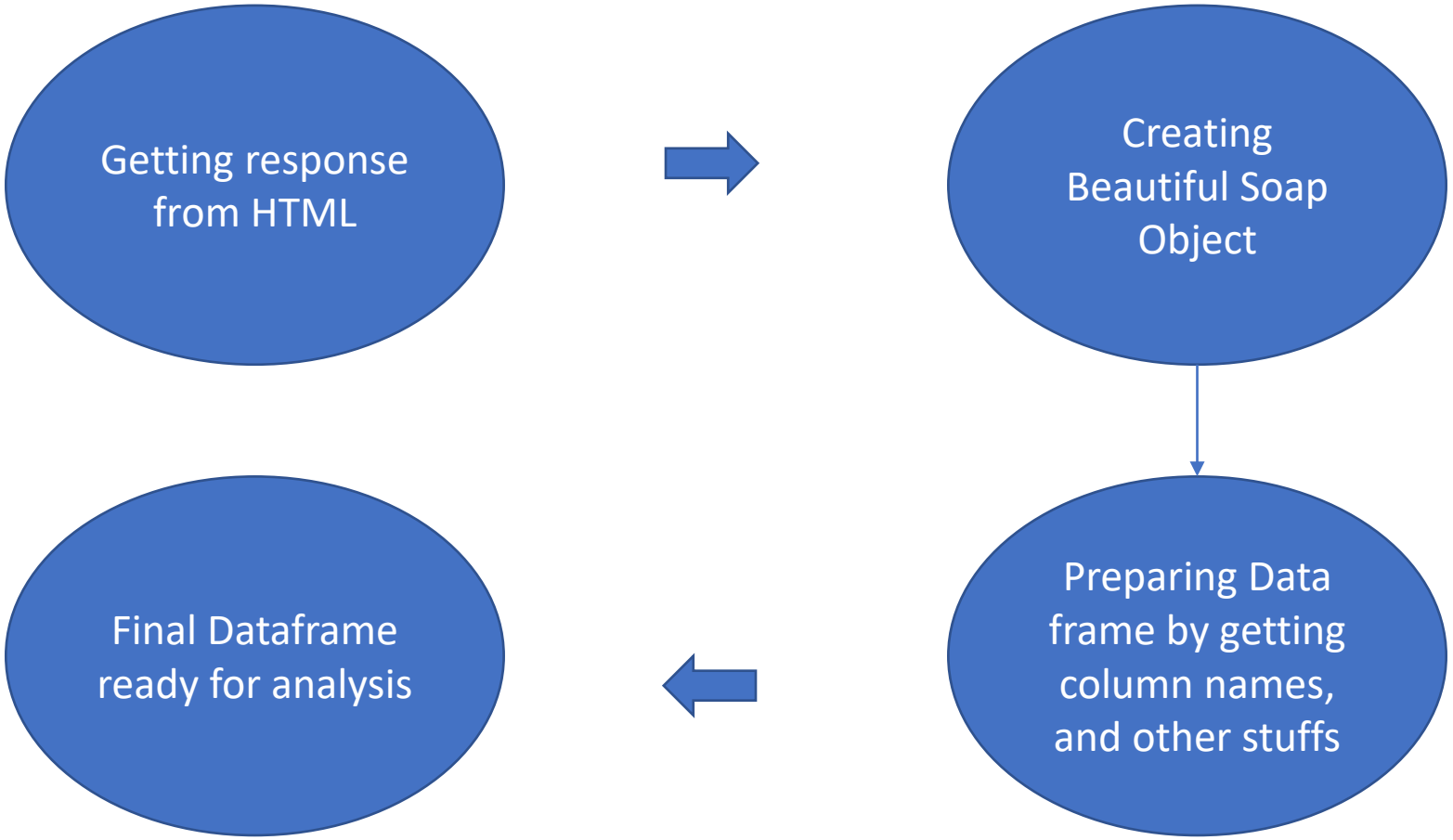
- GitHub URL of the completed web scraping notebook here



Data Wrangling



- Web scrapping



EDA with Data Visualization

- what charts were plotted and why

we use Folium and plotly modules for our interactive visualization about landing pads and distance between them, also we visualize hoe far they are from each other and from other public services like railways and roads.

We try to visualize relation between payloads mass and other attributes and how they are related to launching class of falcon 9 by simple scatter and bar plots.

Also we draw line chart of success rate and year

- GitHub URL of your completed EDA with data visualization notebook

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

  - Displaying the names of the unique launch sites in the space mission • Displaying 5 records where launch sites begin with the string 'KSC’
  - Displaying the total payload mass carried by boosters launched by NASA (CRS) • Displaying average payload mass carried by booster version F9 v1.1
  - Listing the date where the successful landing outcome in drone ship was achieved.
  - Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - Listing the total number of successful and failure mission outcomes
  - Listing the names of the booster_versions which have carried the maximum payload mass.
  - Listing the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
  - Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

- GitHub URL of your completed EDA with SQL notebook here

Build an Interactive Map with Folium

- **To visualize the Launch Data into an interactive map**.
- We took the Latitude and Longitude Coordinates at each launch site and added a Circle Marker around each launch site with a label of the name of the launch site. We assigned the *dataframe launch_outcomes(failures, successes)* to classes 0 and 1 with *Green* and *Red* markers on the map in a *MarkerCluster()*
- Using *Haversine's formula* we calculated the distance from the Launch Site to various landmarks to find various trends about what is around the Launch Site to measure patterns.
- Lines are drawn on the map to measure distance to landmarks Example of some trends in which the Launch Site is situated in.
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Build a Dashboard with Plotly Dash

For building dashboard we use plotly module –

1) *We draw pie chart showing total launches*

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

In a pie chart, the arc length of each slice, is proportional to the quantity it represents.

2) *Scatter Graph showing the relationship with Outcome and Payload Mass (Kg) for the different Booster Versions*

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is **a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data**.

here

Predictive Analysis (Classification)

- **BUILDING MODEL**

  - Load our dataset into NumPy and Pandas
  - Transform Data
  - Split our data into training and test data sets
  - Check how many test samples we have
  - Decide which type of machine learning algorithms we want to use Set our parameters and algorithms to GridSearchCV
  - Fit our datasets into the GridSearchCV objects and train our dataset.

here

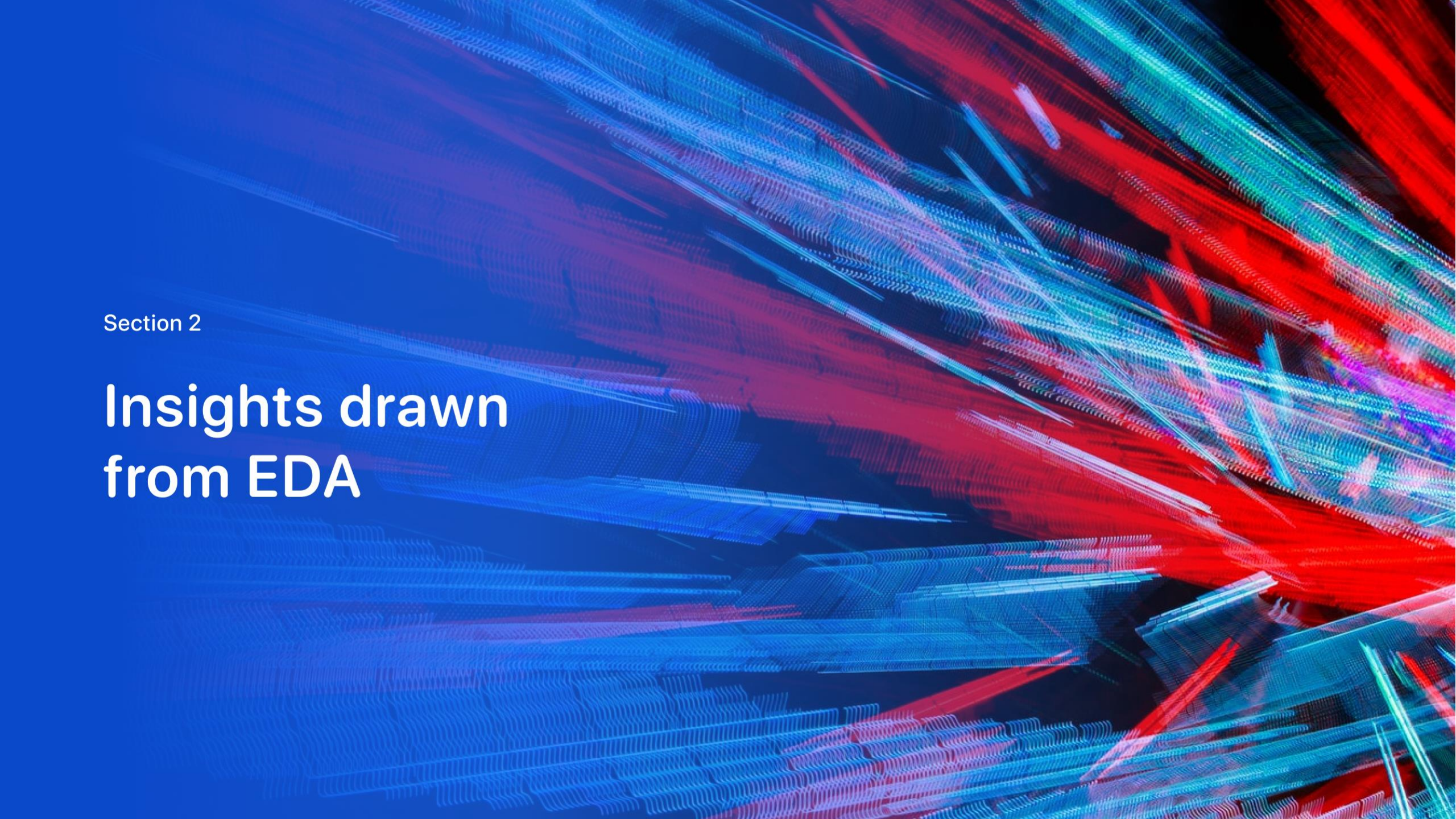
- **EVALUATING MODEL**
  - Check accuracy for each model
  - Get tuned hyperparameters for each type of algorithms
  - Plot Confusion Matrix IMPROVING MODEL
  - Feature Engineering
  - Algorithm Tuning FINDING THE BEST PERFORMING CLASSIFICATION MODEL
  - The model with the best accuracy score wins the best performing model
  - In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook

Results

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

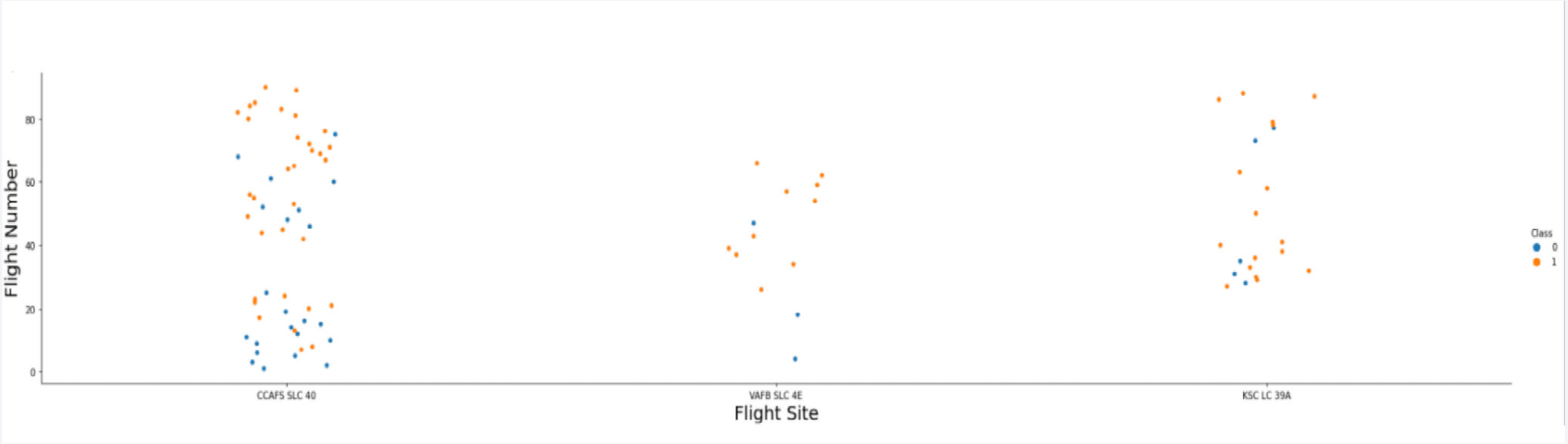
Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

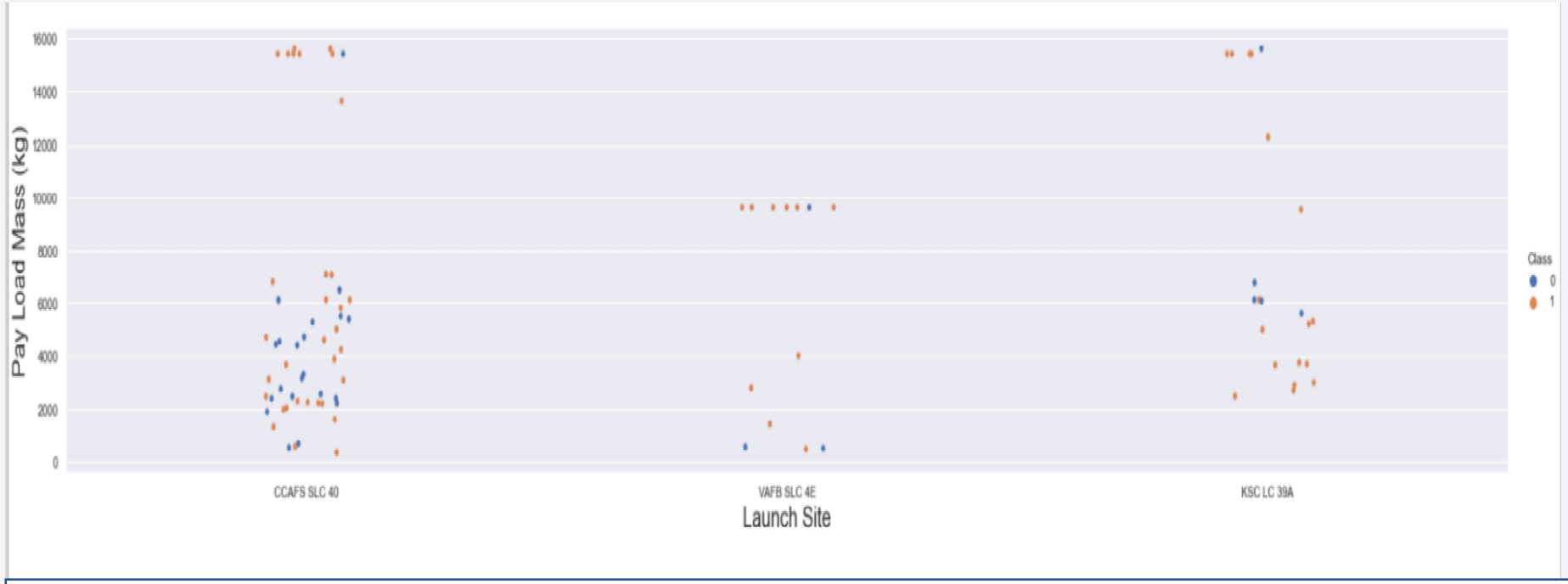
- Scatter plot of Flight Number vs. Launch Site



From above scatter plot we see that CCAFS SLC 40 has maximum number flight launches and it sees that it has maximum successful flight rate among other two launch site.

Payload vs. Launch Site

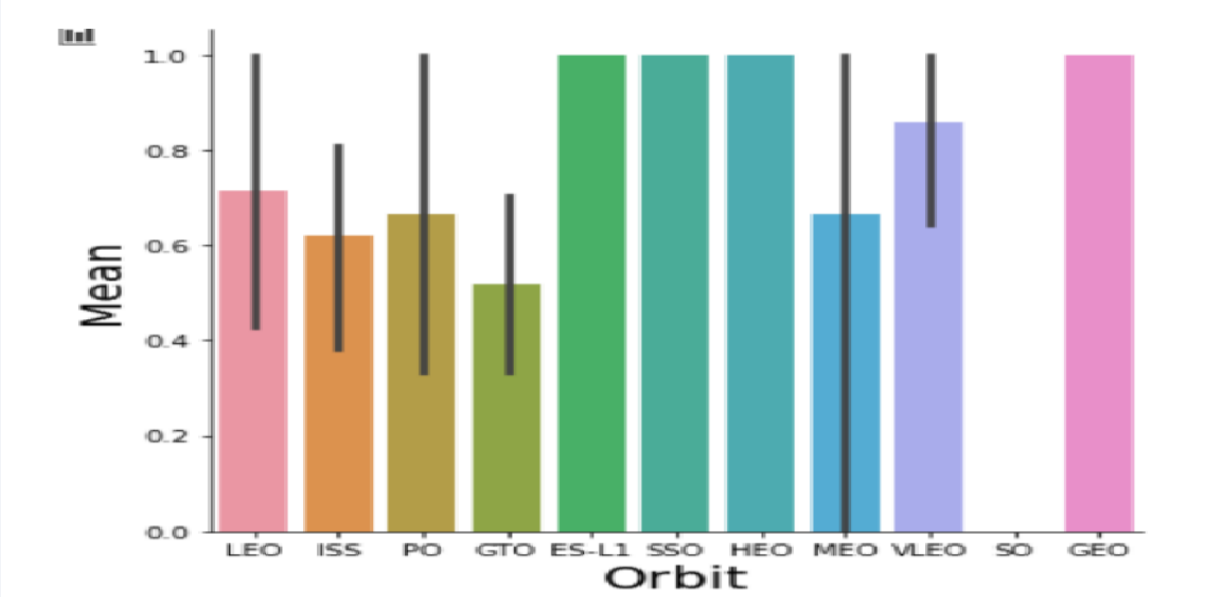
- Scatter plot of Payload vs. Launch Site



From above scatter plot we see that two launch site namely CCAFS SLC 40 and KSC LC 39A carries maximum payload mass during launch.

Success Rate vs. Orbit Type

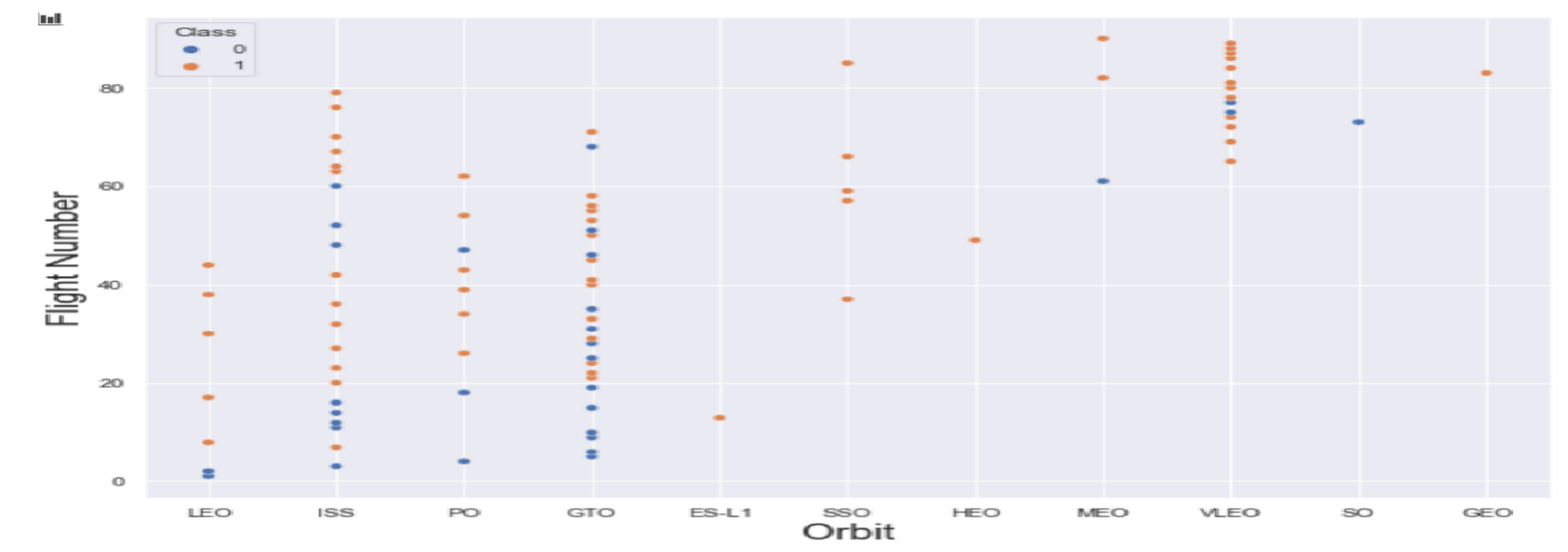
- Bar chart for the success rate of each orbit type



We see that orbit ES-L1,SSO,HEO, and GEO has maximum Success rate compare to other orbit.

Flight Number vs. Orbit Type

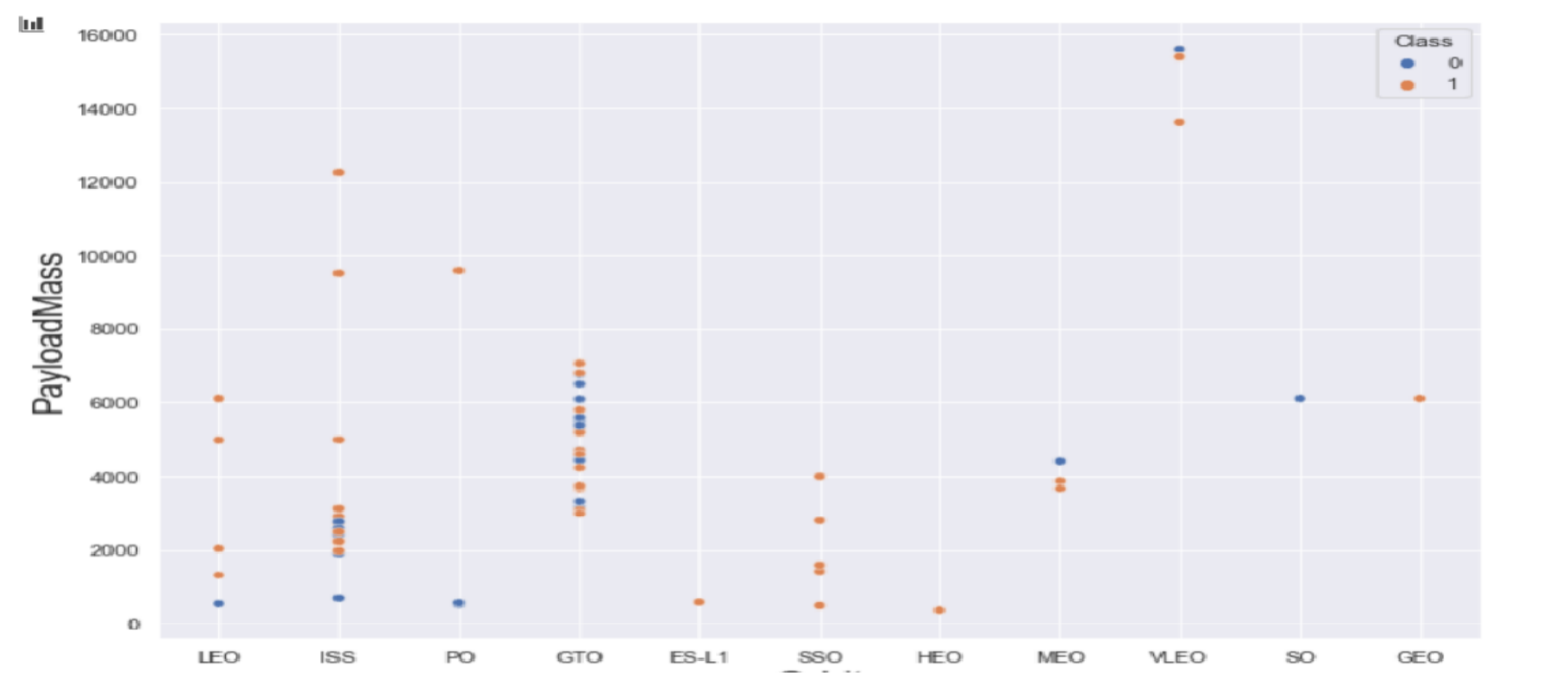
- Scatter plot of Flight number vs. Orbit type



Most of flights are seen in orbit ISS and GTO but it seems that VLEO orbit has most of successful flight launch compare to other orbits even it has low flights with respect to ISS and GTO but its has large success rate than others.
Also we see that there is positive relationship between flights and Orbit Type.

Payload vs. Orbit Type

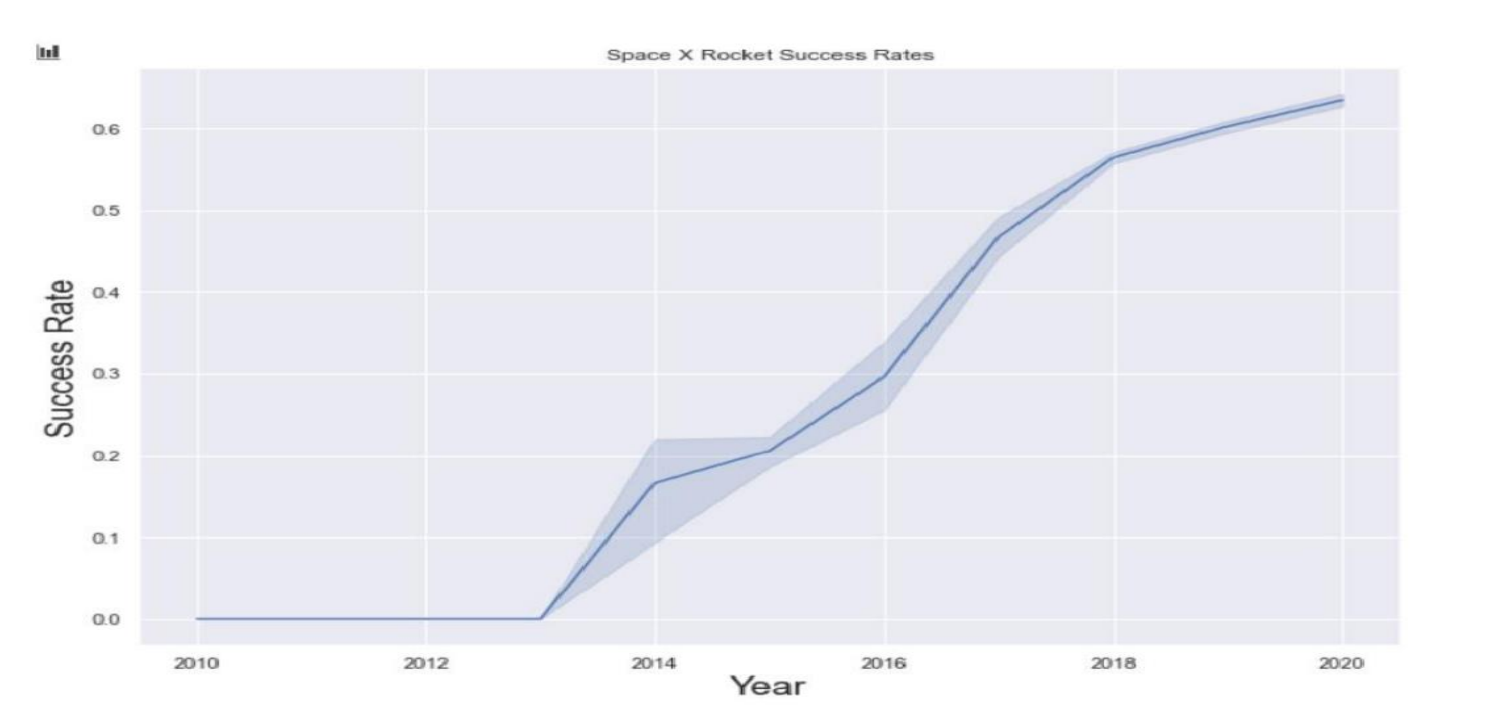
- Scatter point of payload vs. orbit type



Most of flights from orbit ISS and GTO(all most all Orbit) carries specific amount payload mass. Also we see that VLEO orbit Flights maximum Payload mass among all.

Launch Success Yearly Trend

- Trend line chart of yearly average success rate



We see that in year 2010 to 2013 falcon 9 flights success rate is almost zero, But onwards then it follows an increasing trend. As year passes success rate is increases which is indication of successful launches of the Rockets.

Exploratory Data Analysis by SQL

All Launch Site Names

- Names of the unique launch site QUERY:

```
SELECT DISTINCT Launch_site AS UNIQUE_Launch_Sites FROM tblSpaceX;
```

"The `SELECT DISTINCT` statement is used to return only distinct (different) values. Inside a table, a column often contains many duplicate values; and sometimes you only want to list the different (distinct) values."

- Result

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

Launch Site Names Begin with 'KCA'

- QUERY

```
SELECT Launch_site TOP 5 FROM tblSpaceX
    WHERE Launch_site LIKE “KSC%”;
```

The `SELECT TOP` clause is used to specify the number of records to return.
The `SELECT TOP` clause is useful on large tables with thousands of records. Returning a large number of records can impact performance.
The `LIKE` operator is used in a `WHERE` clause to search for a specified pattern in a column.
There are two wildcards often used in conjunction with the `LIKE` operator:

- The percent sign (%) represents zero, one, or multiple characters
- The underscore sign (_) represents one, single character

Total Payload Mass

- Total payload carried by boosters from NASA

```
SELECT SUM(PAYLOAD_MASS_KG_)  AS TotalPayloadMass
        FROM tblSpaceX WHERE Customer = 'NASA (CRS)'
            GROUP BY Customer;
```

The `SUM()` function returns the total sum of a numeric column
The `GROUP BY` statement groups rows that have the same values into summary rows, like "find the number of customers in each country".
The `GROUP BY` statement is often used with aggregate functions (`COUNT()`, `MAX()`, `MIN()`, `SUM()`, `AVG()`) to group the result-set by one or more columns.

- Result

	Total Payload Mass
0	45596

Average Payload Mass by F9 v1.1

QUERY:

```
SELECT AVG(PAYLOAD_MASS_KG_) AS Average_Payload_Mass
    FROM tblSpaceX WHERE Booster_Version = 'F9 v1.1'
        GROUP BY Booster_Version;
```

The `AVG()` function returns the average value of a numeric column.
The `GROUP BY` statement groups rows that have the same values into summary rows, like "find the number of customers in each country".
The `GROUP BY` statement is often used with aggregate functions (`COUNT()`, `MAX()`, `MIN()`, `SUM()`, `AVG()`) to group the result-set by one or more columns.

Result:

	Average Payload Mass
0	2928

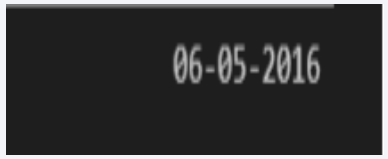
First Successful Ground Landing Date

QUERY:

```
SELECT MIN(Date) SLO FROM tblSpaceX WHERE Landing_Outcome = ”Success (drone ship)” ;
```

The `MIN()` function returns the smallest value of the selected column.
The `MAX()` function returns the largest value of the selected column.

Result:



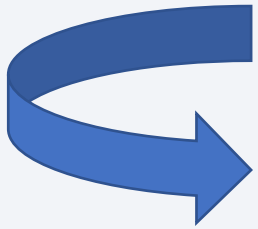
Successful Drone Ship Landing with Payload between 4000 and 6000

QUERY:

```
SELECT Booster_Version FROM tblSpaceX
    WHERE Landing_Outcome = 'Success (ground pad) AND
        Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000
```

WHERE clause select drone with the payload mass between 4000 and 6000 with the help of AND clause in it.

RESULT:



Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

QUERY:

```
SELECT(SELECT Count(Mission_Outcome) FROM tblSpaceX
    WHERE Mission_Outcome LIKE '%Success%') AS Successful_Mission_Outcomes,
        (SELECT  Count(Mission_Outcome) FROM tblSpaceX
            WHERE Mission_Outcome LIKE '%Failure%')
                AS Failure_Mission_Coutcomes
```

Here we used a two subquery for required findings. All other clause have their respective feature as we explained before with the given condition.

RESULT:

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

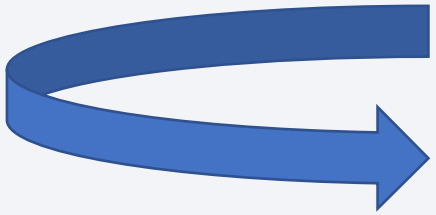
Boosters Carried Maximum Payload

QUERY:

```
SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS _KG_)
    AS [Maximum Payload Mass] FROM tblSpaceX GROUP BY
        Booster_Version ORDER BY [Maximum Payload Mass] DESC
```

Here we select unique booster versions with maximum payload mass and arrange them by payload mass in decreasing order using GROUP BY clause.

RESULT:



	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows × 2 columns

2017 Launch Records

QUERY:

```
SELECT DATENAME(month, DATEADD(month, MONTH(CONVERT(date, Date, 105)), 0) - 1) AS Month,
Booster_Version, Launch_Site, Landing_Outcome FROM tblSpaceX WHERE (Landing_Outcome LIKE
N'%Success%') AND (YEAR(CONVERT(date, Date, 105)) = '2017’)
```

RESULT:

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

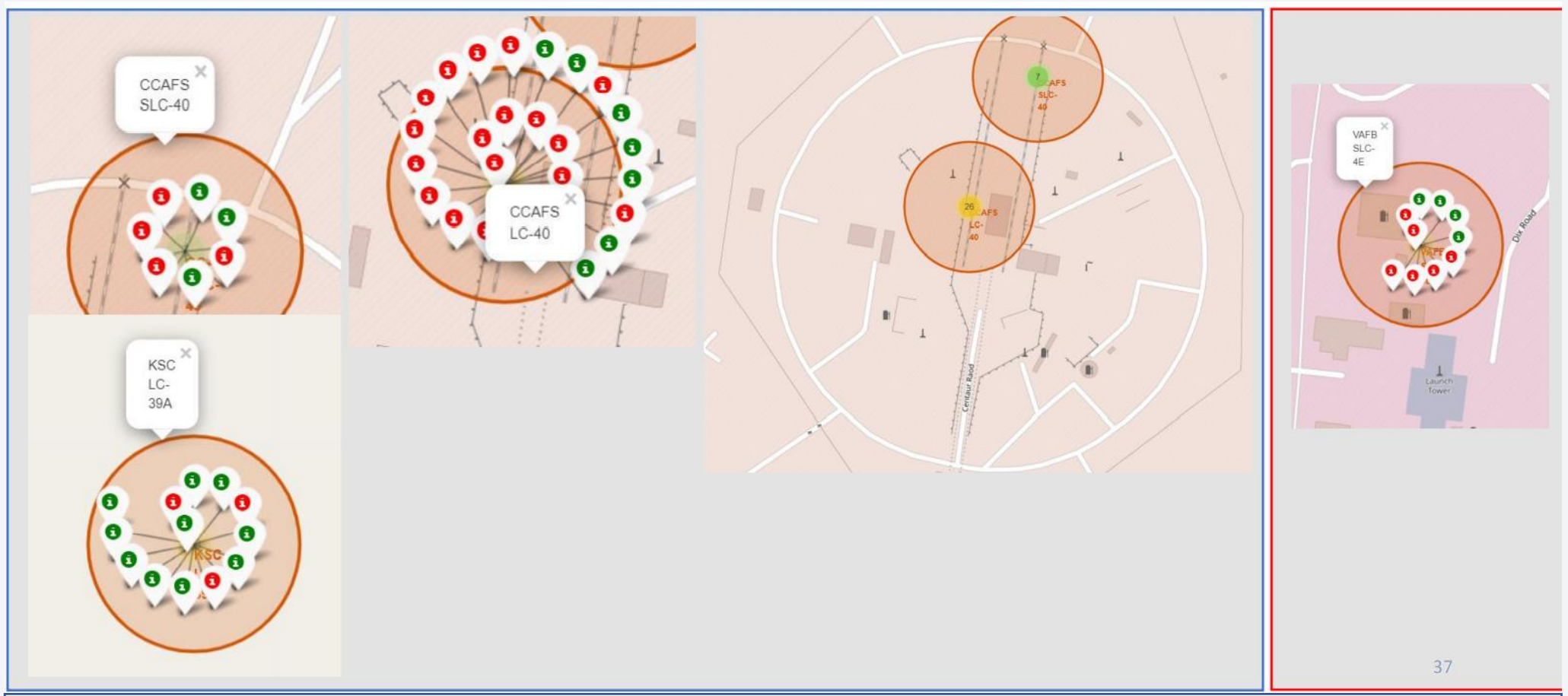
Section 4

Launch Sites Proximities Analysis

Launch Site

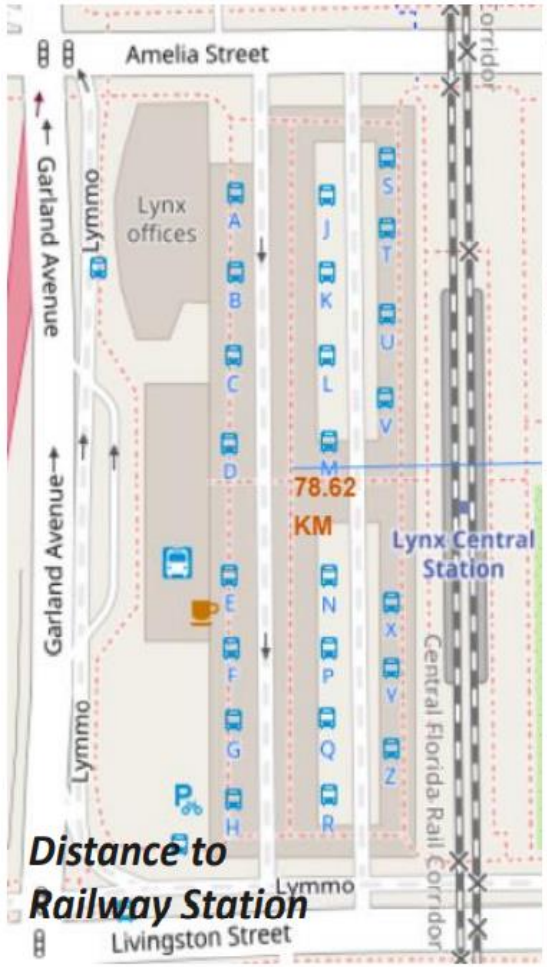
Launch Site of Space X are lies in North America. In California and Florida Region.

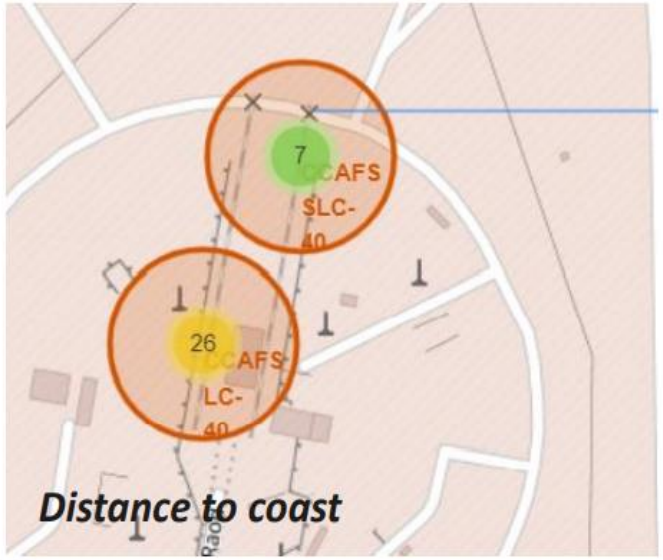
Successful and failed mission on launch site

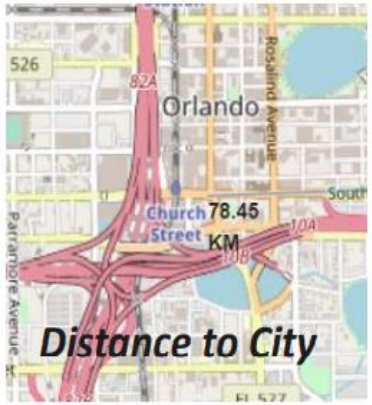


In above Map red markers shows failed mission and green markers shows succeed mission.

Working out Launch Sites distance to landmarks to find trends with Haversine formula using CCAFS-SLC-40 as a reference



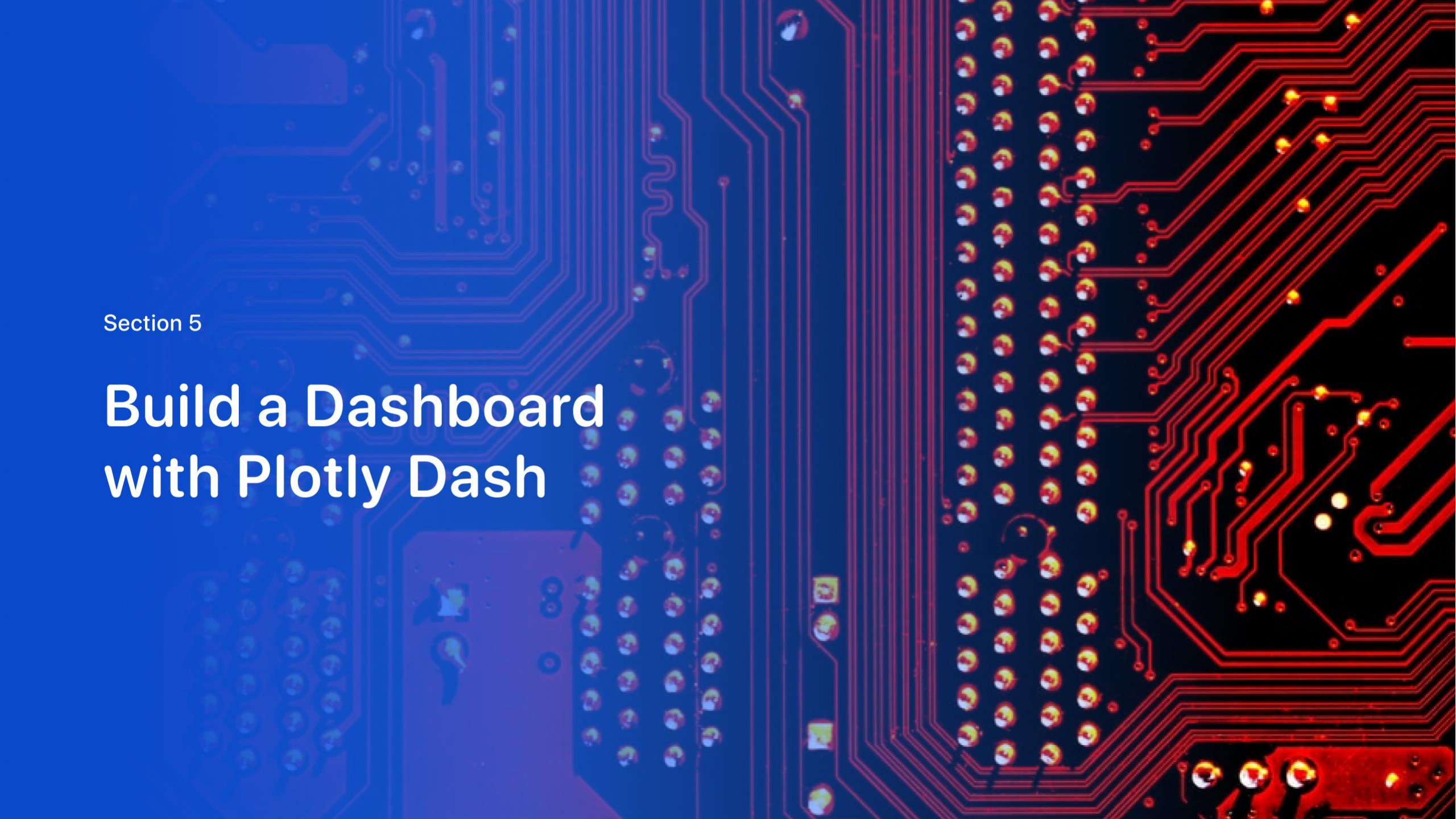




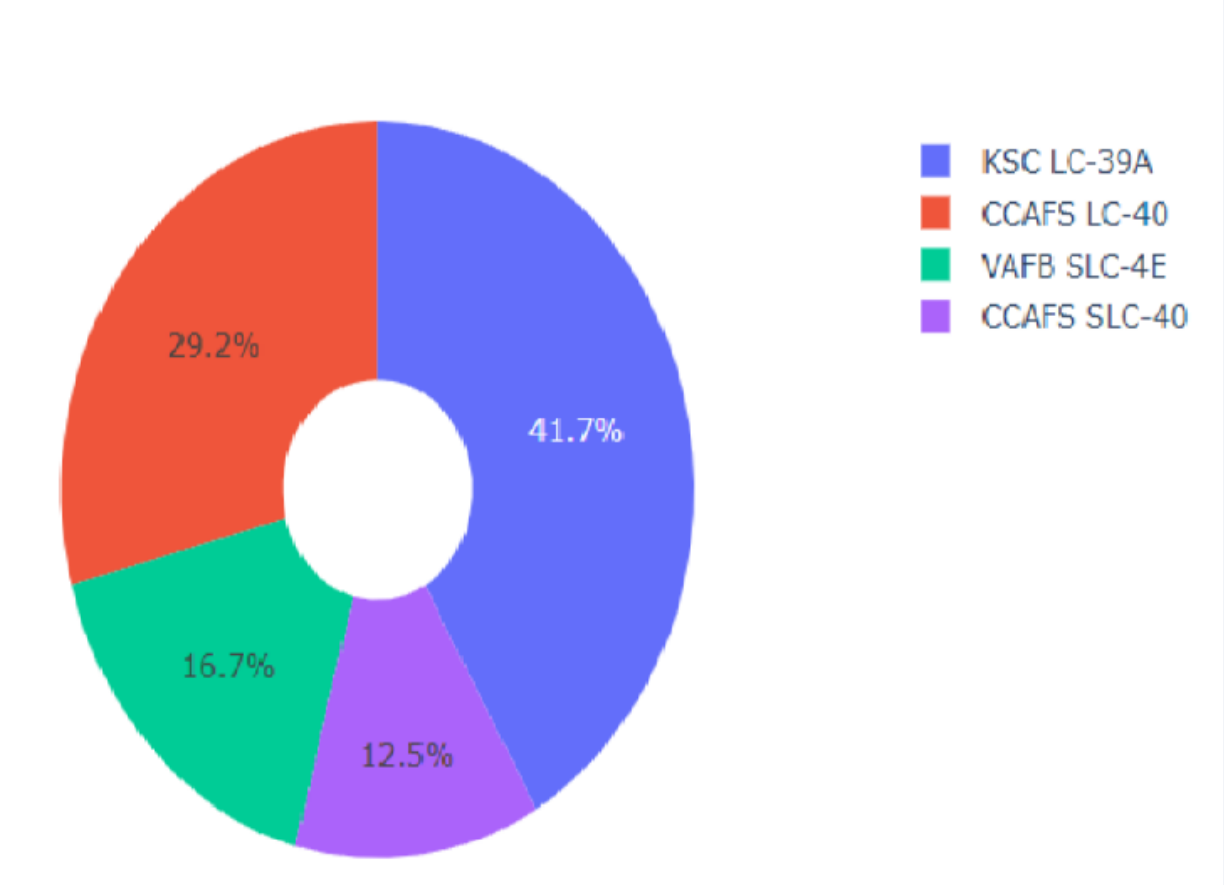
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 5

Build a Dashboard with Plotly Dash

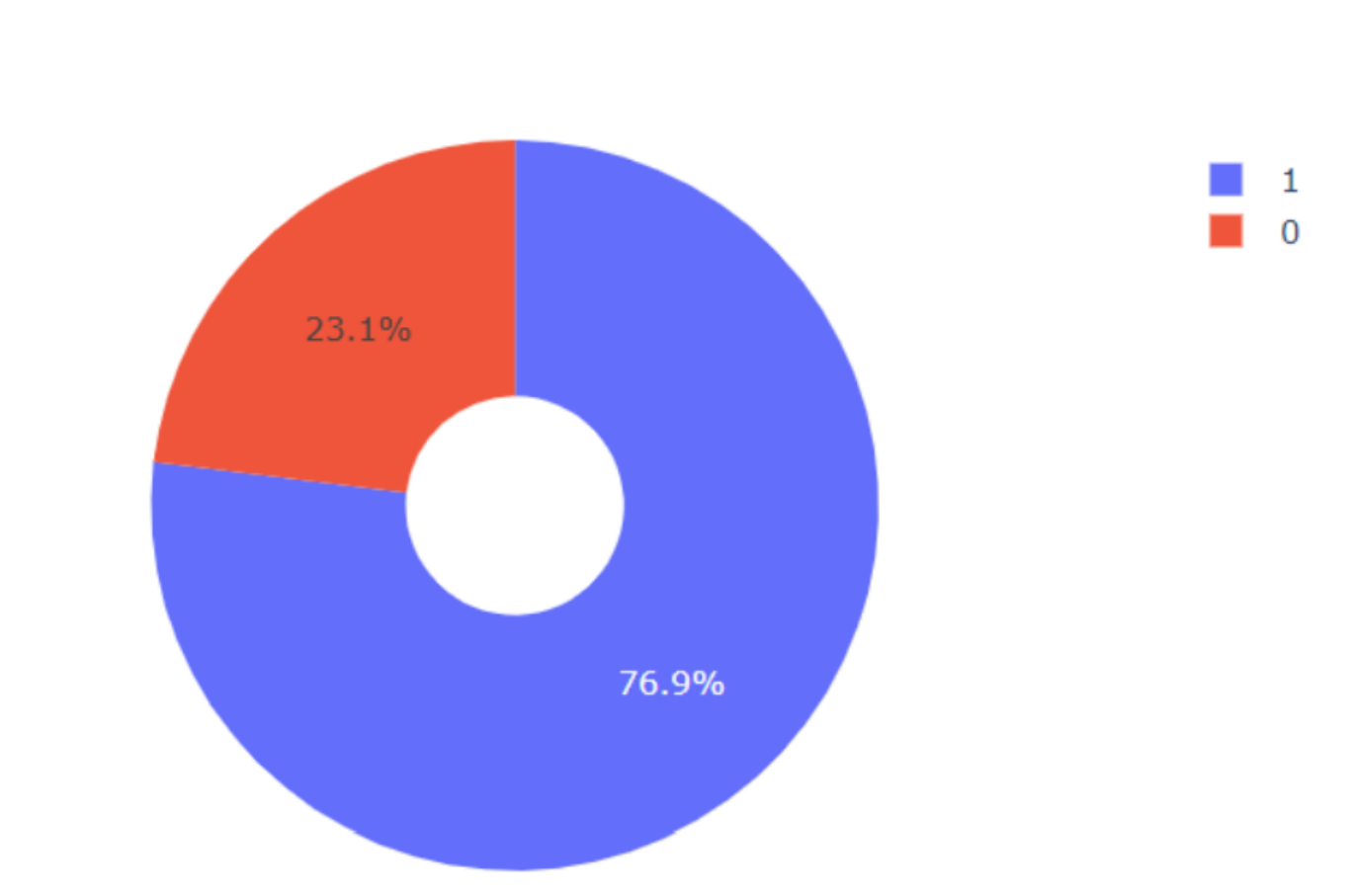


Success rate Pie Chart



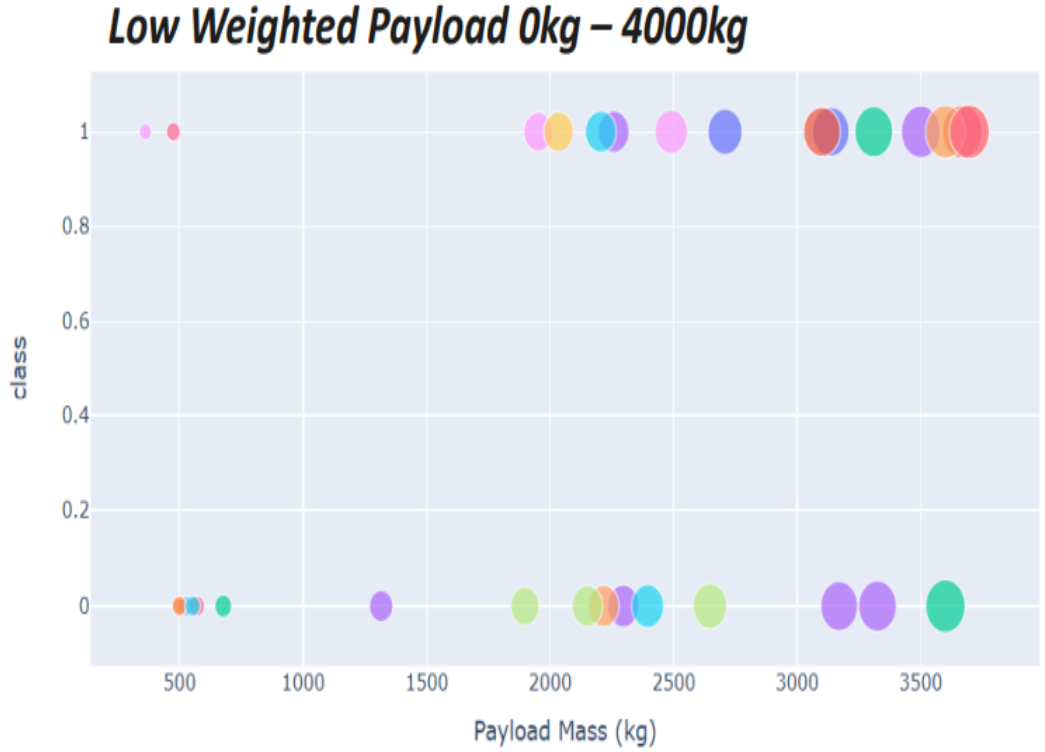
KSC LC-39A has maximum success rate among all launch site.

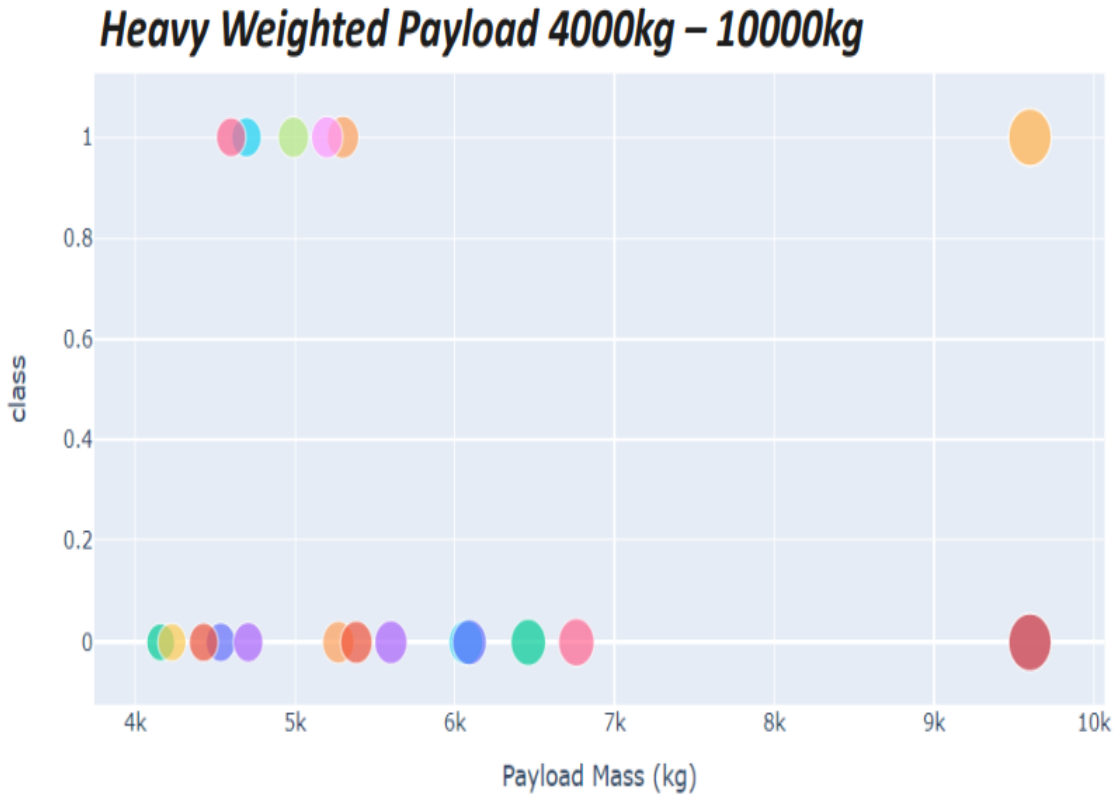
KSC LC-39A success and Failure



About 77 % time launches happened at KSC -39A are succeeded.

Payload vs. Launch Outcome scatter plot





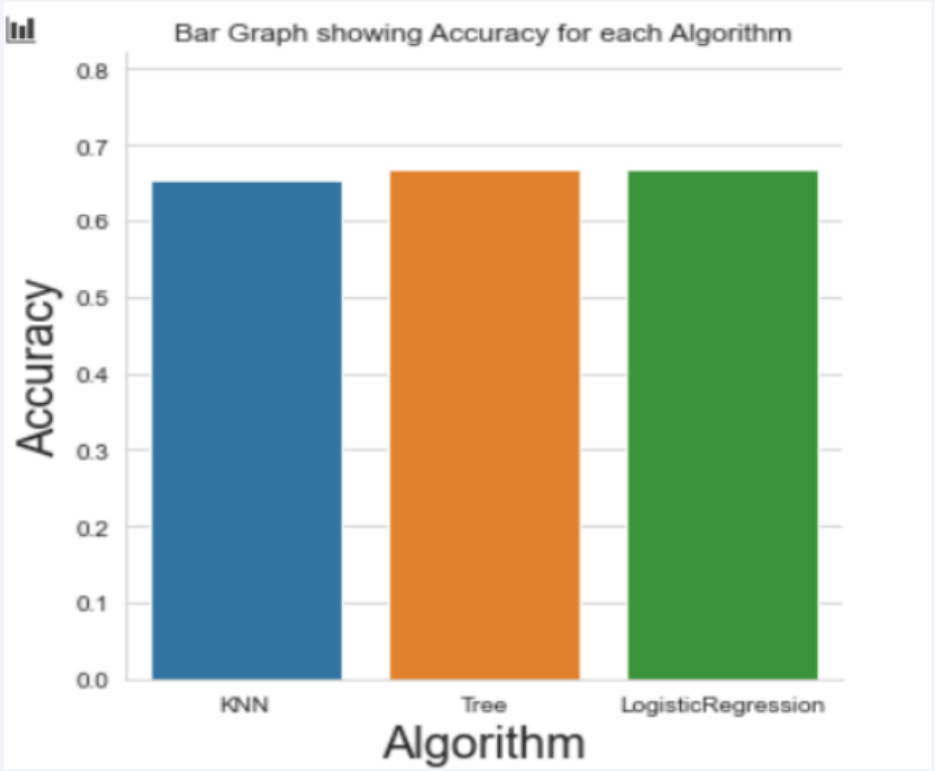
Section 6

Predictive Analysis (Classification)

Classification Accuracy

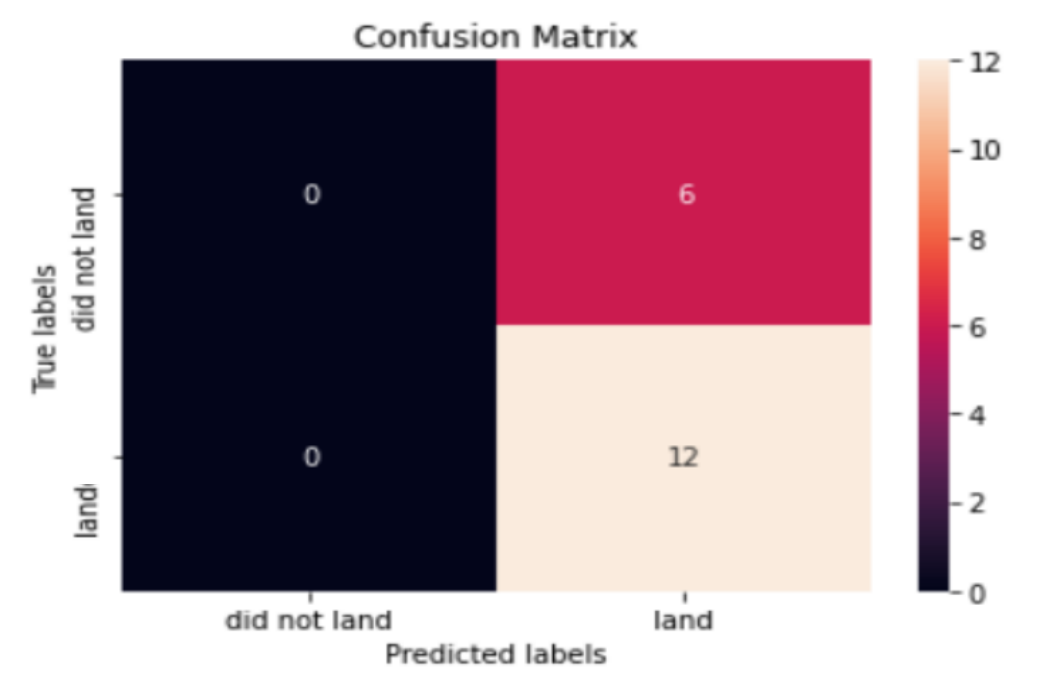
- Visualizing the built model accuracy for all built classification models, in a bar chart

	Accuracy	Algorithm
0	0.653571	KNN
1	0.667857	Tree
2	0.667857	LogisticRegression



- Decision Tree Algorithm is best for given predictive (classification) Analysis.

Confusion Matrix



Our predictive model (decision tree) produces False positive outcomes in case which means that In actual rocket did not land but our model predict that model did land. Which can be Sevier problem later in predicting.

Conclusions

- CCAFS SLC 40 has maximum number flight launches and it sees that it has maximum successful flight rate among other two launch site.
- Two launch site namely CCAFS SLC 40 and KSC LC 39A carries maximum payload mass during launch.
- Year 2010 to 2013 falcon 9 flights success rate is almost zero, But onwards then it follows an increasing trend. As year passes success rate is increases which is indication of successful launches of the Rockets.
- Launch Site of Space X are lies in North America. In California and Florida Region.
- About 77 % time launches happened at KSC -39A are succeeded.
- Decision Tree Algorithm is best for given predictive (classification) Analysis.
- Our predictive model (decision tree) produces False positive outcomes in case which means that In actual rocket did not land but our model predict that model did land. Which can be Sevier problem later in predicting.

Thank you!