Winning Space Race with Data Science

Hari E

11-11-2025

Outline

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

Executive Summary

- The aim of this project is to analyze the SpaceX Falcon9 data, to provide an useful insights and analysis for the company SpaceY for effective bid at other companies for achieving ambitious cost effective space travel program.
- **Summary of methodologies**

  Methodologies include collection of data by web scrapping, and also by an API(Application Programming Interface),then transform the data by Data Wrangling methods. Then will perform Exploratory Data Analysis with SQL and python visualization libraries. Then, using a Plotly dashboard for gaining interactive analysis and also using Folium libraries in Python for interactive visualization. Then, will create a machine learning model for high accuracy prediction of landings of space flights.

- **Summary of all results**

After modelling the suitable machine learning model ,we found that, Decision Tree classification model having an high accuracy score of 90.35%. From the Confusion Matrix on the test data, we observed that there is an accuracy of 83.33%.

Introduction

- **Project background and context**

  SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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. Spaces X's Falcon 9 launch like regular rockets. You will also determine if SpaceX will reuse the first stage. The goal of the project to determine, if the first stage will land successfully by train a machine learning model and use public information to predict if SpaceX will reuse the first stage. So, that SpaceY will have effective competition with SpaceX when competing with rocket launches.

- **Problems you want to find answers**
  - The factors contributing for successful launches ,
  - Reasons behind the failures of some missions,
  - Find the optimal conditions for effective mission launches, and also for reuse capability of first stage of the rockets.

Section 1

Methodology

Methodology

Executive Summary

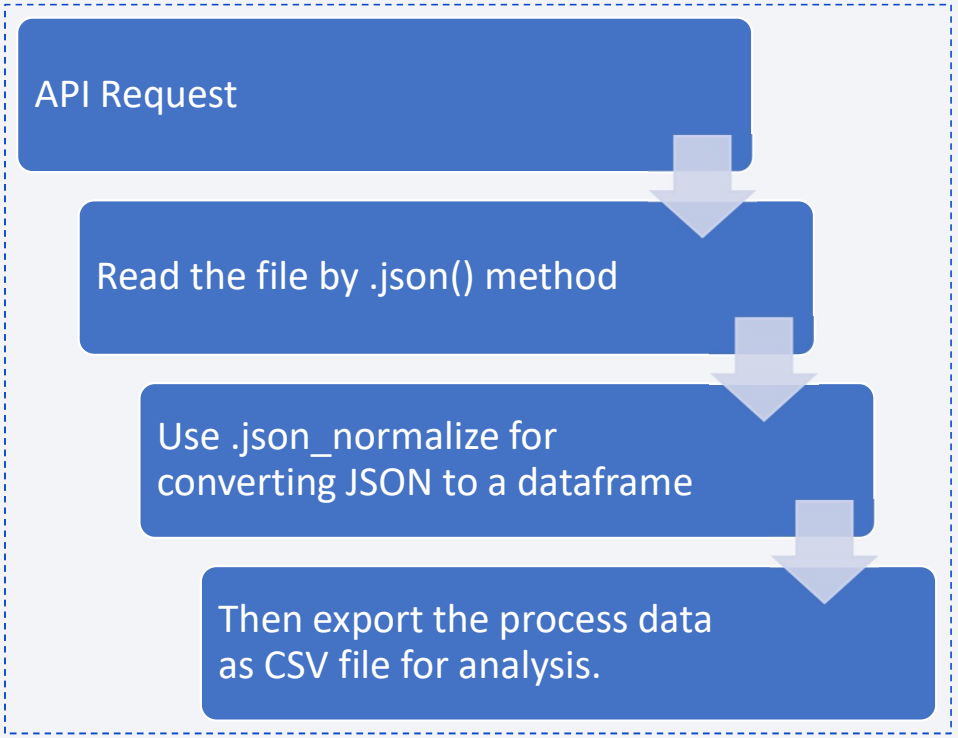
- Data collection methodology:
  - By SpaceX API, and also through , Web Scrapping : Falcon 9 and Falcon heavy launch records from Wikipedia (LINK).
- Perform data wrangling
  - By converting the outcomes into Training Labels with 1 means the booster successfully landed and 0 means it was unsuccessful.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - By create a column for the class ,and standardize the data and then, split into training data and test data. Thereafter, find the method performs best using test data.

Data Collection

- Describe how data sets were collected.
  - SpaceX launch data that is gathered from an API, specifically the SpaceX REST API. This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
  - Also, by web scrapping the Wikipedia pages for relevant SpaceX launches.

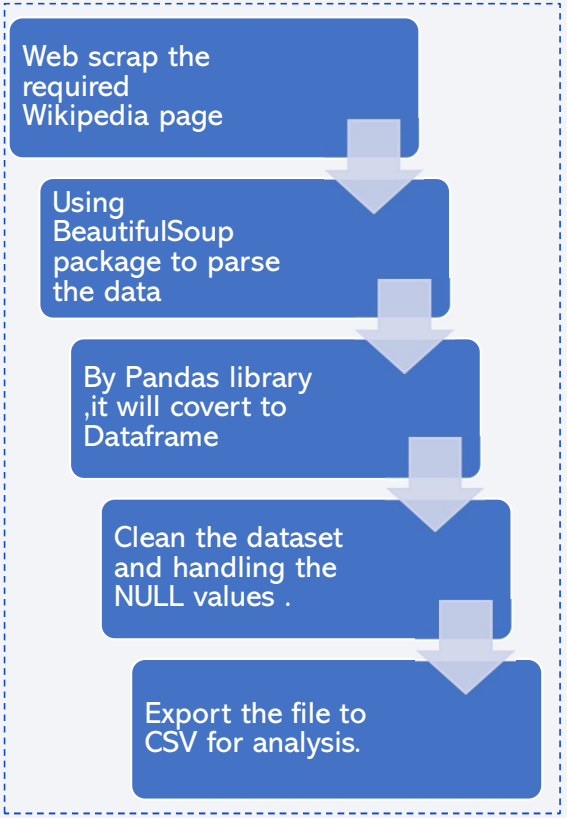
Data Collection – SpaceX API

- We will perform a get request using the requests library to obtain the launch data, which we will use to get the data from the API.
- This result can be viewed by calling the .json() method. Our response will be in the form of a JSON, specifically a list of JSON objects.
- Since we are using an API, you will notice in the lab that when we get a response it is in the form of a JSON.
- Specifically, we have a list of JSON objects which each represent a launch. To convert this JSON to a dataframe, we can use the json_normalize function.
- This function will allow us to "normalize" the structured json data into a flat table. This is what your JSON will look like in a table form.
- GitHub Link



Data Collection – Web Scraping

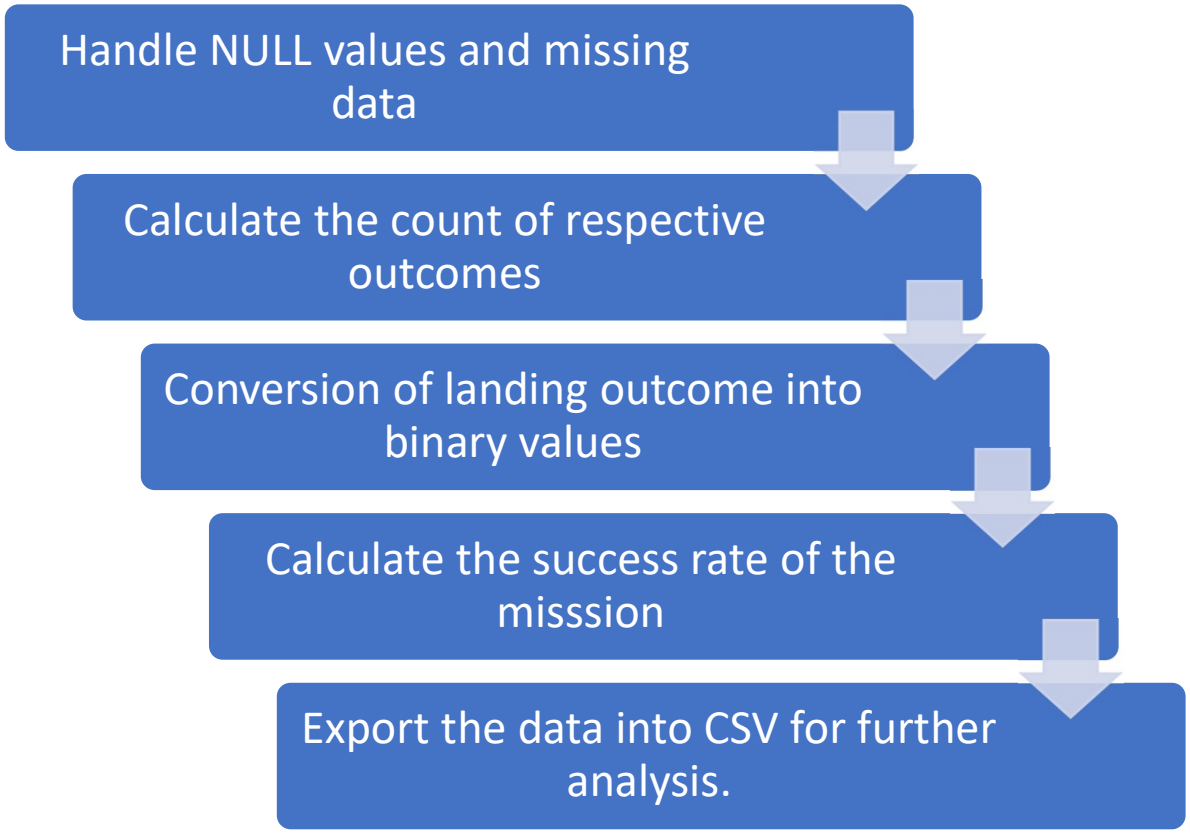
- Now, we are using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.
- Then you need to parse the data from those tables and convert them into a Pandas data frame for further visualization and analysis.
- We want to transform this raw data into a clean dataset which provides meaningful data.
- Finally, not all gathered data is perfect. We may end up with data that contains NULL values. But,we have to deal with these null values in order to make the dataset viable for analysis.
- GitHub Link



Data Wrangling

- Exploratory Data Analysis (EDA) are performed to find some patterns in the data and determine what would be the label for training supervised models.
- In the data set, there are several different cases where the booster did not land successfully.
- True Ocean - means the mission outcome was successfully landed to a specific region of the ocean
- False Ocean - means the mission outcome was unsuccessfully landed to a specific region of the ocean.
- True RTLS - means the mission outcome was successfully landed to a ground pad ,False RTLS means the mission outcome was unsuccessfully landed to a ground pad.
- True ASDS - means the mission outcome was successfully landed on a drone ship ,False ASDS - means the mission outcome was unsuccessfully landed on a drone ship.
- By One hot enconding, the training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.
- GitHub Link

Data Wrangling



EDA with Data Visualization

In the Exploratory Data Analysis(EDA) we will gain insights and further more details from the given dataset by plotting various charts by visualization libraries.

By plotting the Scatter plot for finding the relationship among the following:

- Visualize the relationship between Flight Number and Launch Site
- Visualize the relationship between Payload Mass and Launch Site
- Visualize the relationship between FlightNumber and Orbit type
- Visualize the relationship between Payload Mass and Orbit type

By plotting the Bar chart for finding the relationship among the following:

- Visualize the relationship between success rate of each orbit type

By plotting the Line chart for finding the relationship among the following:

- Visualize the launch success yearly trend
- GitHub Link

EDA with SQL

Now, we are performing the Exploratory Data Analysis(EDA) with SQL . Load the given SpaceX dataset into the corresponding table in a Db2 database, then execute SQL queries for getting the following results.

- 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 successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
- 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.

- GitHub Link

Build an Interactive Map with Folium

Now, we are performing more interactive visual analytics using Folium interactive map and find the following tasks.

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities.

At first we need to create a folium `Map` object, with an initial center location to be NASA Johnson Space Center at Houston, Texas. Then, use `folium.Circle` to add a highlighted circle area with a text label on a specific coordinate. After that, we create a Marker clusters, which can be a good way to simplify a map containing many markers having the same coordinate.

Now, add a MousePosition on the map to get coordinate for a mouse over a point on the map. As such, while you are exploring the map, you can easily find the coordinates of any points of interests.After that, we use folium.PolyLine( ) object to line between the required coordinates. Repeat the streps for all sites such as railways,coastline,highways to get the required results.

- GitHub Link

Build a Dashboard with Plotly Dash

- Now we are building a dashboard application with the Python Plotly Dash package.
- This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.
- After the dashboard is built, you can use it to find more insights from the SpaceX dataset more easily than with static graphs.

GitHub Link For Python code

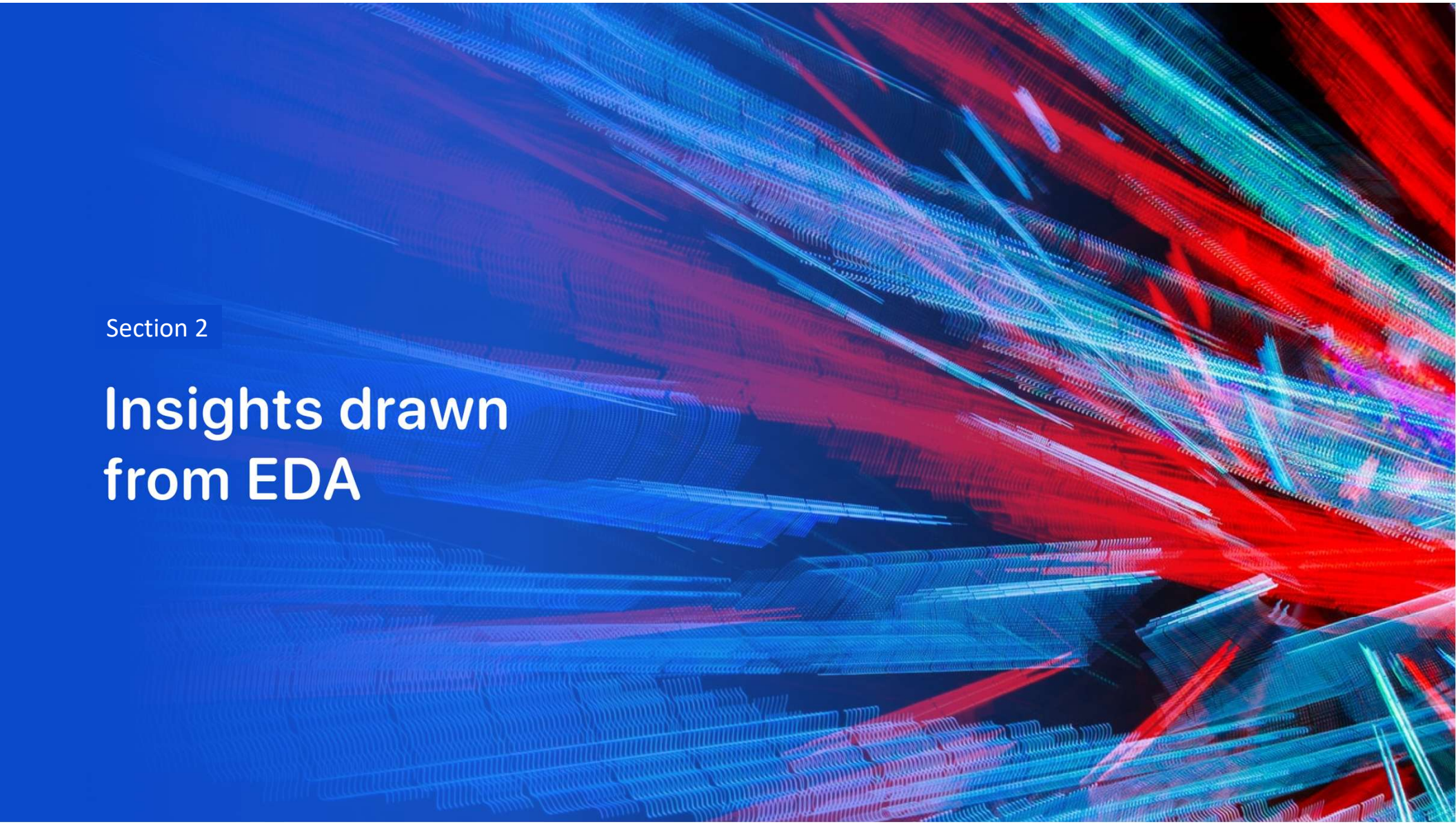
GitHub Link for CSV file

Predictive Analysis (Classification)

- In this stage, we will build a machine learning pipeline to predict if the first stage of the Falcon 9 lands successfully.
- By doing the subsequent process such as Preprocessing, allowing us to standardize our data, and Train_test_split, allowing us to split our data into training and testing data.
- Then, train the model and perform Grid Search, allowing us to find the hyperparameters that allow a given algorithm to perform best.
- Using the best hyperparameter values, we will determine the model with the best accuracy using the training data.
- After that , test the Logistic Regression, using the Support Vector machines, Decision Tree Classifier, and K-nearest neighbors. Finally, we will achieve output with accuracy by the confusion matrix.
- GitHub Link

Results

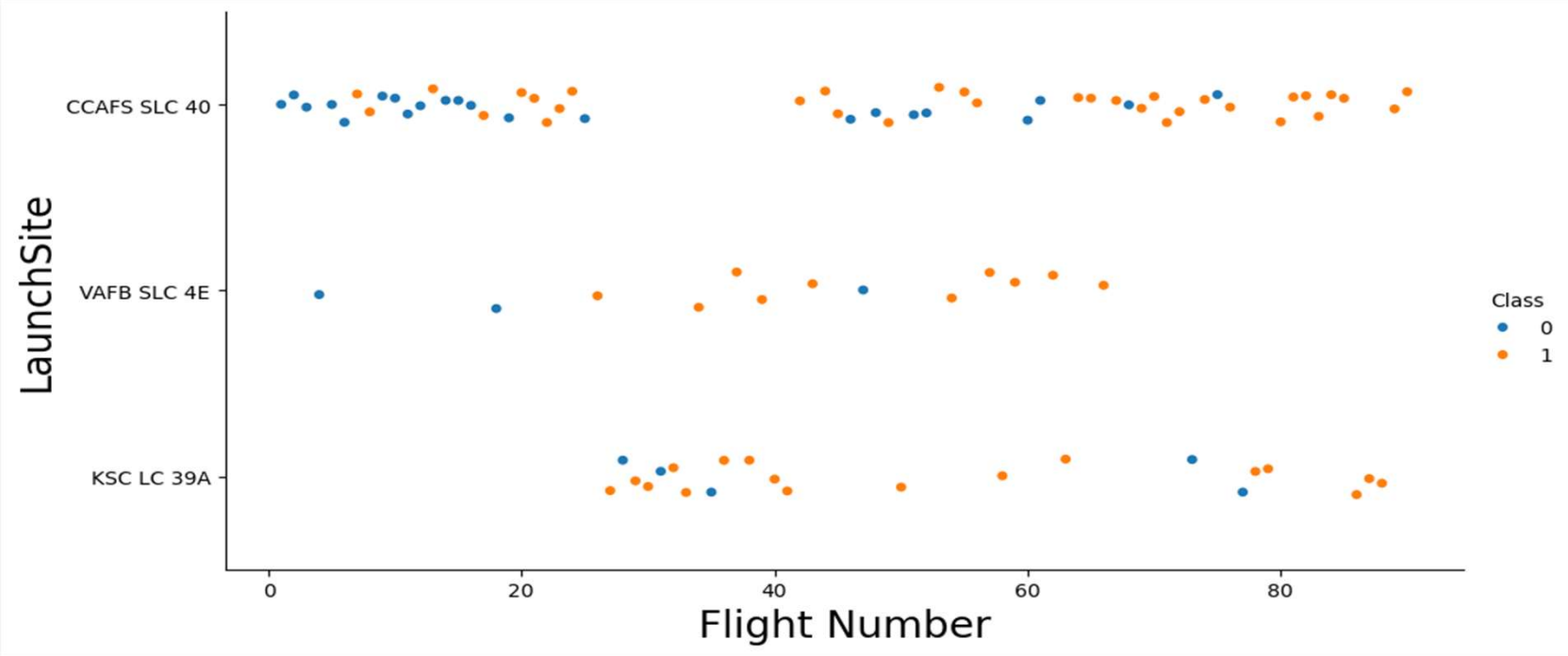
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results
- The Results are analyzed and discussed in the upcoming slides with detailed visualizations.



Section 2

Insights drawn from EDA

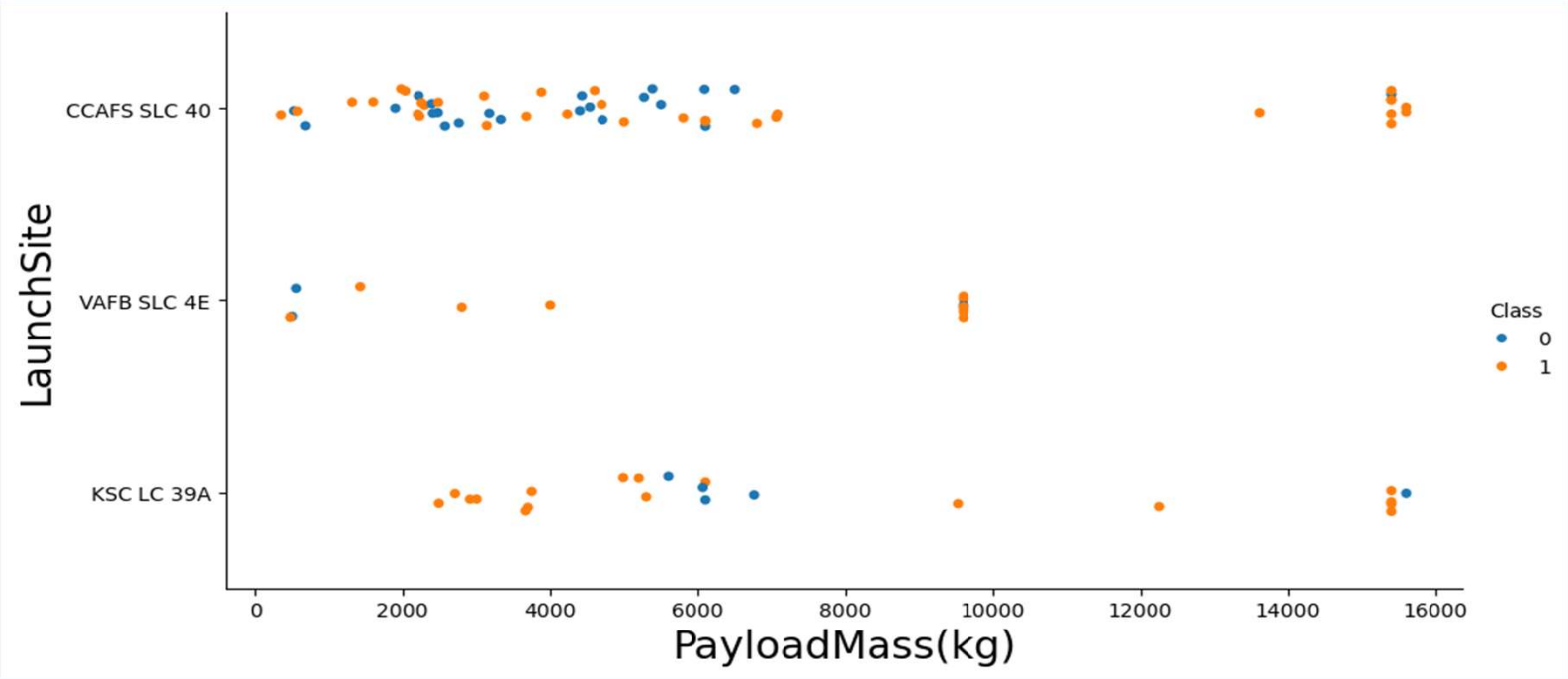
Flight Number vs. Launch Site



Scatter plot of Flight Number vs. Launch Site

- From the above plot, Launch site CCAFS SLC 40 have a significant (class-0) identities that denotes failure launches. On other hand, KSC LC 39A have a high success launches(class-1 identities) among the three launch sites.

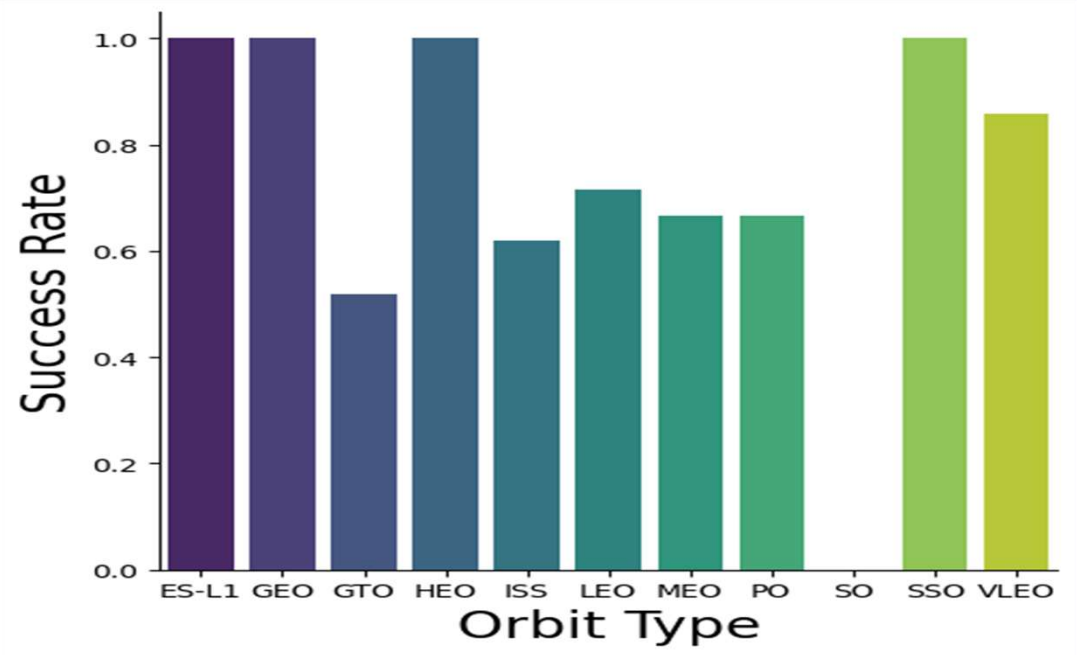
Payload vs. Launch Site



Scatter plot of Payload vs. Launch Site

- From the above plot, we able to understand that, the CCAFS SLC 40 and KSC LC 39A launch sites , have a high successful launches for the Payload Mass greater than 15000 Kg.
- The VAFB SLC 4E launch site have low launches but most are successful , and it cannot have the launches more the Payload Mass of 10000 Kg.

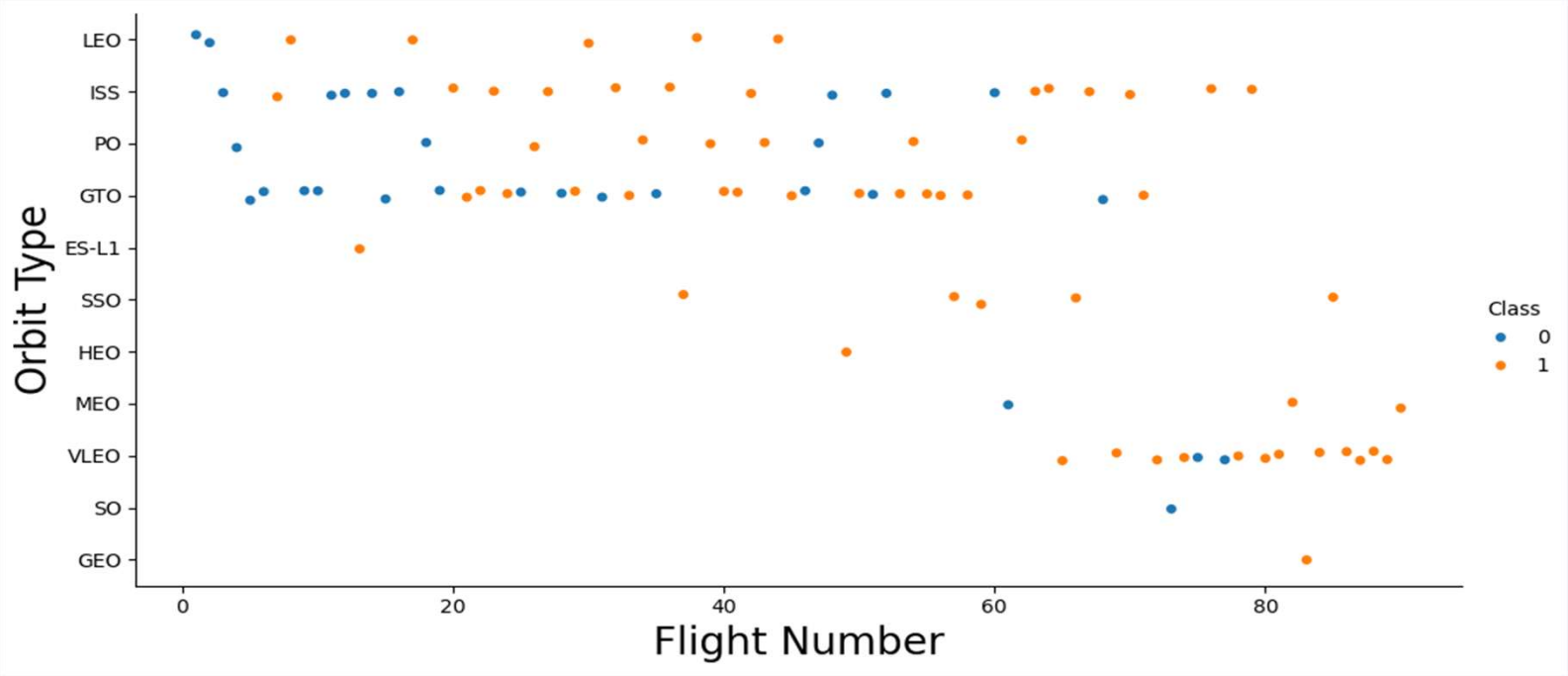
Success Rate vs. Orbit Type



Bar chart for the success rate of each orbit type

- From the bar chart, ES-L1,GEO, HEO, SSO have a 100% success rate. VLEO have a notable success of approx.85% success rate.
- And also, SO have no success rate and GTO have a lowest success rate amongst the other orbit types.

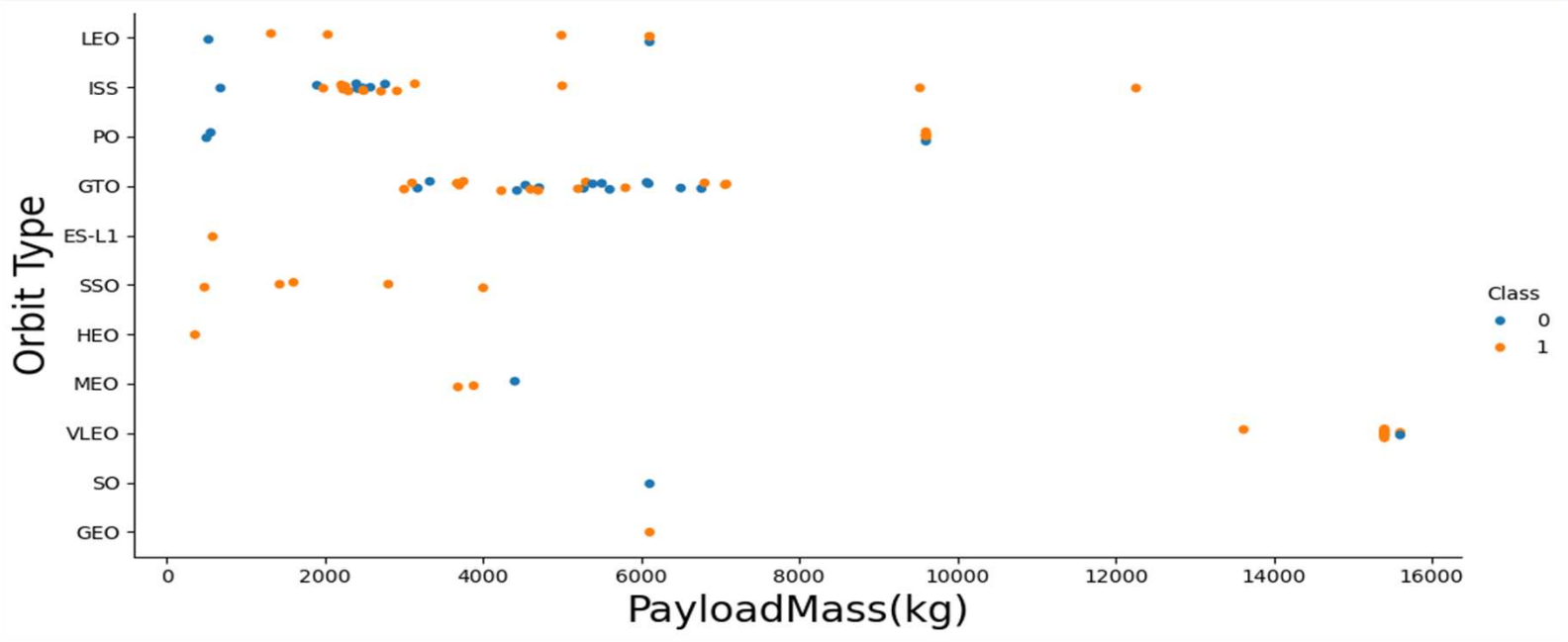
Flight Number vs. Orbit Type



Scatter plot of Flight number vs. Orbit type

- From the above plot, we got an inference that the orbit types ISS and VLEO have a high success with increased flight numbers. The Orbit types LEO and SSO have a lower number of successful flights.

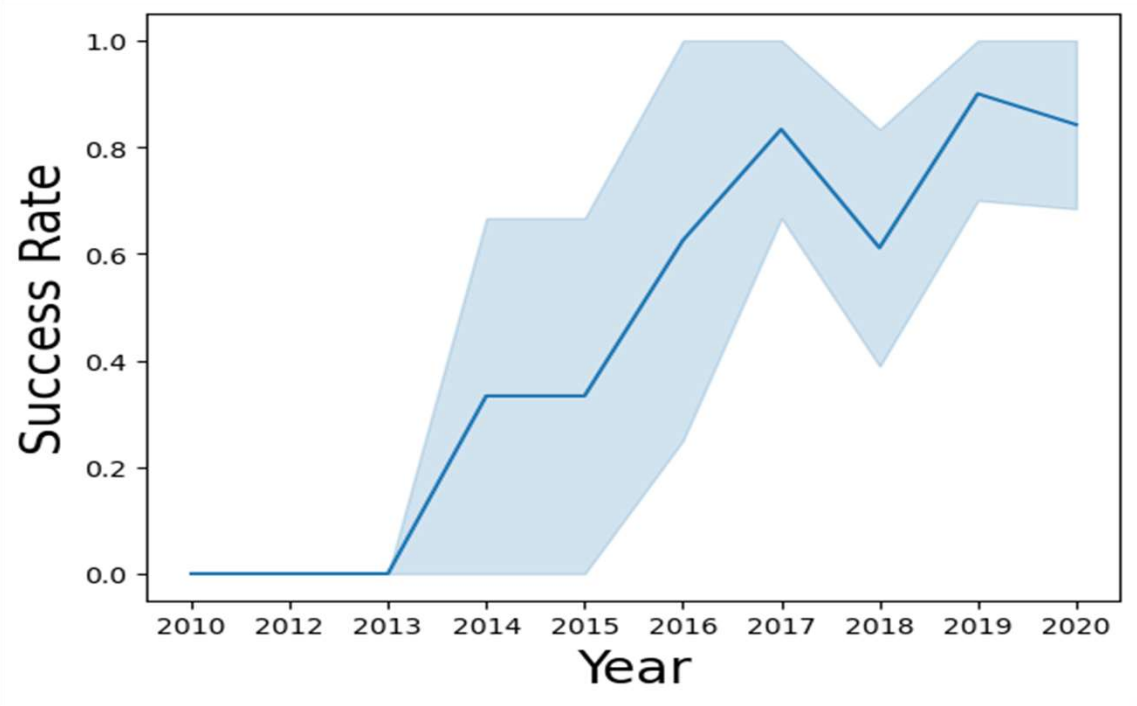
Payload vs. Orbit Type



Scatter plot of payload vs. orbit type

- From the above plot, we understand the orbit type VLEO have a success launch with Payload Mass more than 15000 Kg. The orbits ISS and PO have a successful flight more than Payload mass of 10000 Kg. The GEO orbit have a single and lowest successful flight.

Launch Success Yearly Trend



Line chart of yearly average success rate

- The Line chart illustrates that, the Success Rate increased from 2013 to a highest of 90% in between the years of 2019-20.

All Launch Site Names

- Query used for return unique values from given database column.

```
[10]: %%sql
SELECT DISTINCT "Launch_Site"
    FROM SPACEXTBL;
```

- There are 4 unique values(Launch sites) in column.

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

[10]:

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

Launch Site Names Begin with 'CCA'

- Query used for finding launch sites begin with `CCA` and to limit the returned records to 5.

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

- The result displays 5 records by querying the launch site column.

[12]:

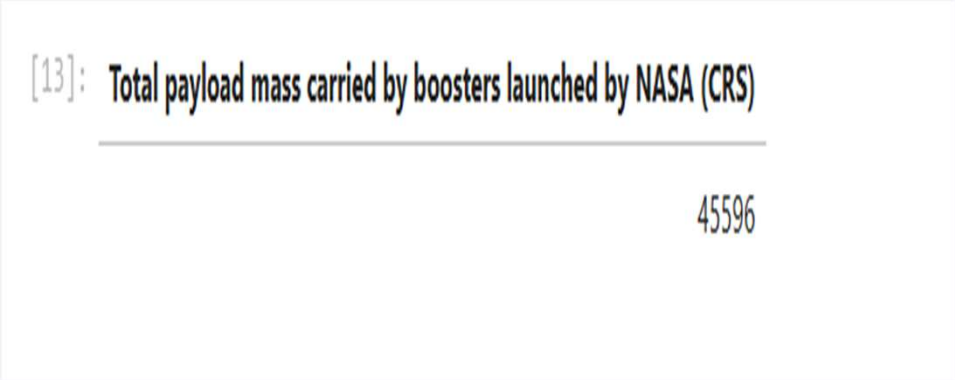
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- Query used for calculate the total payload carried by boosters from NASA is given below:

```
[13]: %%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS " Total payload mass carried by boosters launched by NASA (CRS)" FROM SPACEXTBL WHERE Customer='NASA (CRS)';
 * sqlite:///my_data1.db
Done.
```

- The result obtained by running the query is given below:



Average Payload Mass by F9 v1.1

- Query used for calculate the average payload mass carried by booster version F9 v1.1

```
[14]: %%sql
      select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';

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

- The following result obtained by running the above query:

```
[14]: avg(PAYLOAD_MASS__KG_)
      ----------------------
                      2928.4
```

First Successful Ground Landing Date

- Query used for get the dates of the first successful landing outcome on ground pad.

```
[15]: %sql SELECT MIN(Date) FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'

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

- The result obtained by running the above query:

[15]:

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query used for list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000.

```
[16]: %sql SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
 * sqlite:///my_data1.db
Done.
```

- The following result obtained by running the above query.

[16]:

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

Total Number of Successful and Failure Mission Outcomes

- Query used for calculate the total number of successful and failure mission outcomes.

```
[17]: %sql SELECT Number_of_success_outcomes,Number_of_failure_outcomes FROM (SELECT COUNT(*) AS Number_of_success_outcomes FROM SPACEXTBL WHERE Mission_Outcom

 * sqlite:///my_data1.db
Done.

[17]: ion_Outcome LIKE 'Success%') success_table,(SELECT COUNT(*) Number_of_failure_outcomes FROM SPACEXTBL WHERE Mission_Outcome LIKE 'Failure%') failure_table

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

- The following result obtained by running the above query:

[17]:

Number_of_success_outcomes	Number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- Query used to list the names of the booster which have carried the maximum.

```
[18]: %sql SELECT DISTINCT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
 * sqlite:///my_data1.db
Done.
```

- The following result obtained from running the above query:

[18]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

- Query used to list the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

```
[19]: %sql SELECT substr(Date,6,2)AS Month, Date, Booster_version, Launch_Site,[Landing_Outcome] \
      FROM SPACEXTBL \
      WHERE [Landing_Outcome]='Failure (drone ship)' AND substr(Date,0,5)='2015';
       * sqlite:///my_data1.db
      Done.
```

- The following result obtained by running the above query:

[19]:

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- Query used for 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.

```
[20]: %sql SELECT [Landing_Outcome],count(*) AS count_outcomes FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY [Landing_Outcome] ORDER
 * sqlite:///my_data1.db
Done.
[20]: utcome],count(*) AS count_outcomes FROM SPACEXTBL WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY [Landing_Outcome] ORDER BY count_outcomes DESC
 * sqlite:///my_data1.db
Done.
```

The following result obtained by running the above query:

[20]:

Landing_Outcome	count_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

Section 3

Launch Sites Proximities Analysis

SpaceX – Launch site locations

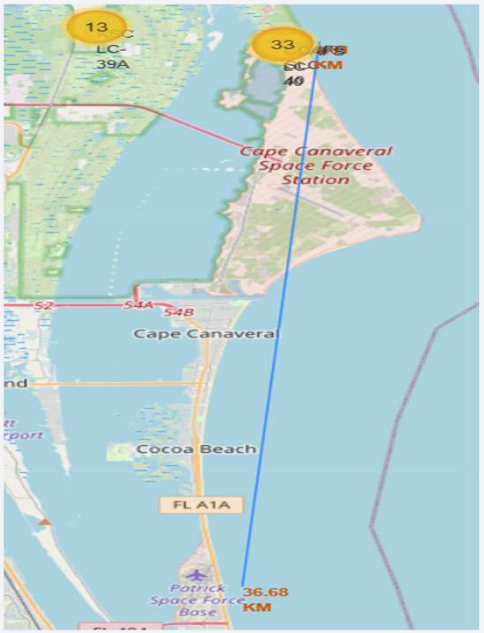
The 4 launch sites of SpaceX Falcon9 are displayed by folium interactive visualization maps. They are:

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

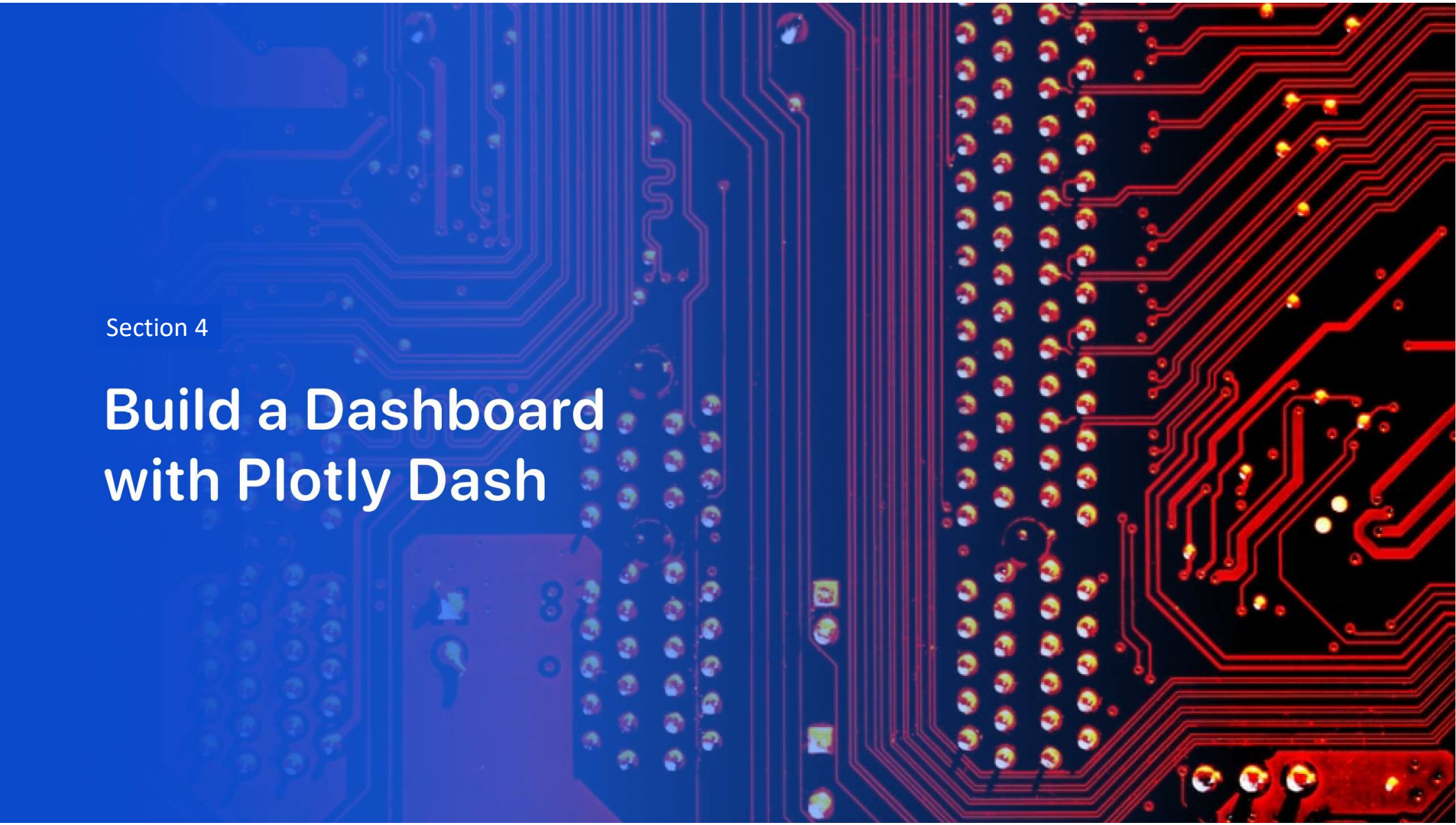
SpaceX Falcon-9 : Success/Failed launches for the launch sites

From the above map , green markers indicate successful launches and red markers indicate failed launches from respective sites.

SpaceX Falcon 9 - Launch site to proximity distance



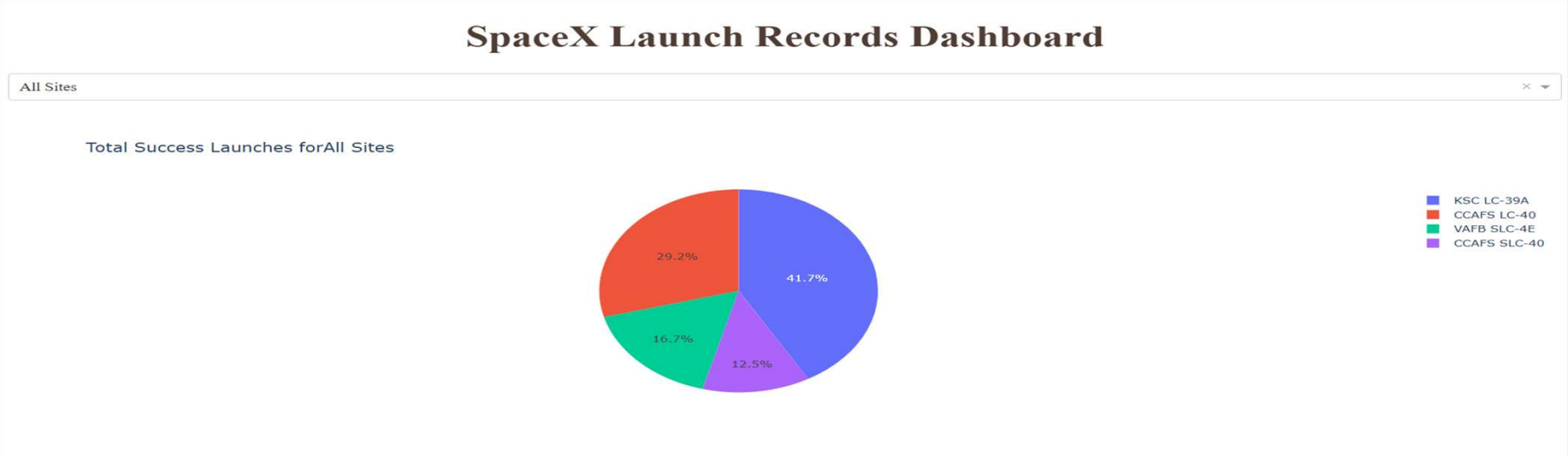
By using Folium.Polyline() object we able extend the line between the two proximities between the launch sites as above.



Section 4

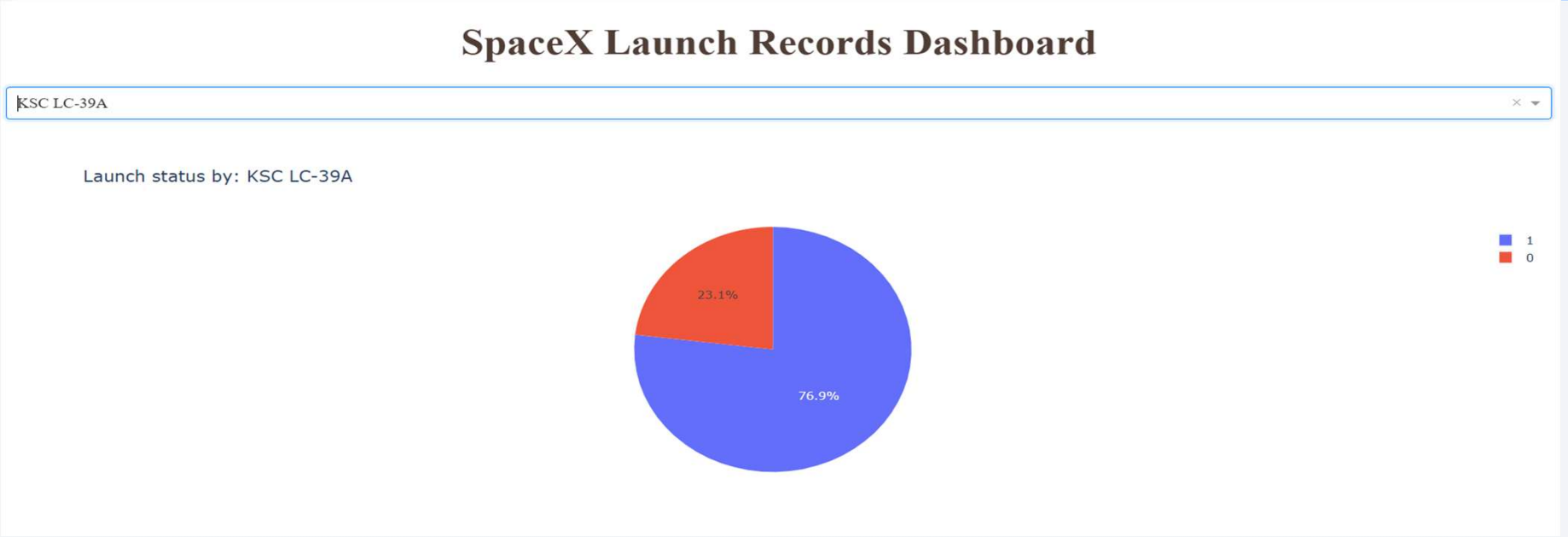
Build a Dashboard with Plotly Dash

Success rates of All Launch sites



- From the generated Plotly dashboard , it is observed that KSC LC -39A launch site had a highest share of successful landings at 41.7% of total among the other launch sites.
- The launch site CCAFS SLC-40 had a lowest success rate of 12.5% of all other launch sites.

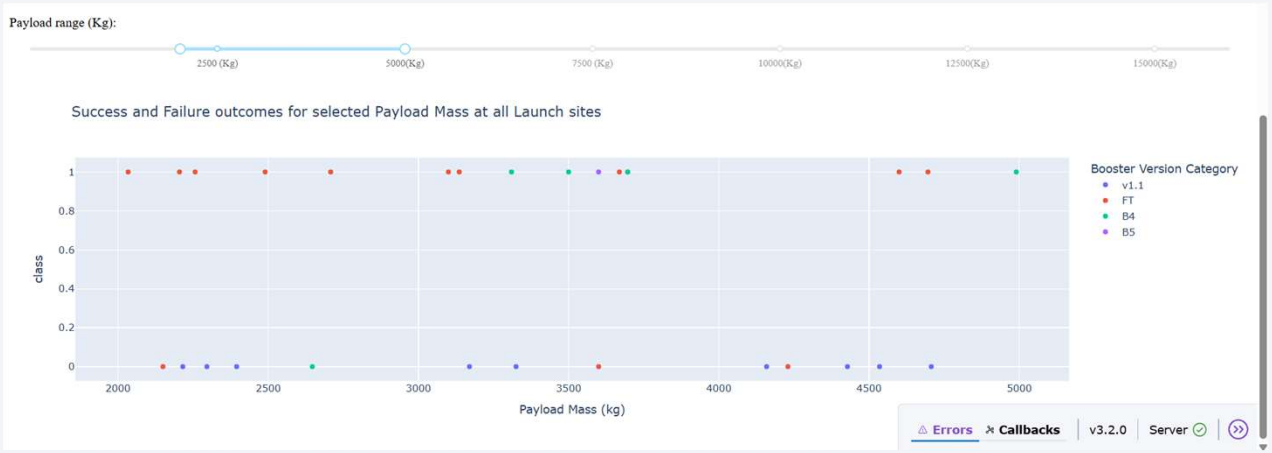
Launch Site with Highest Rate of Successful Landings



- From the previous slide we came know KSC LC-39A has the highest success rate overall among the other sites. But , also the launch site itself has a highest successful landings at 76.9%.

Illustration between Payload Mass and Booster Version Category

From the scatter plot on left at top , we understand that, FT Booster version had the highest rate of success than other booster category

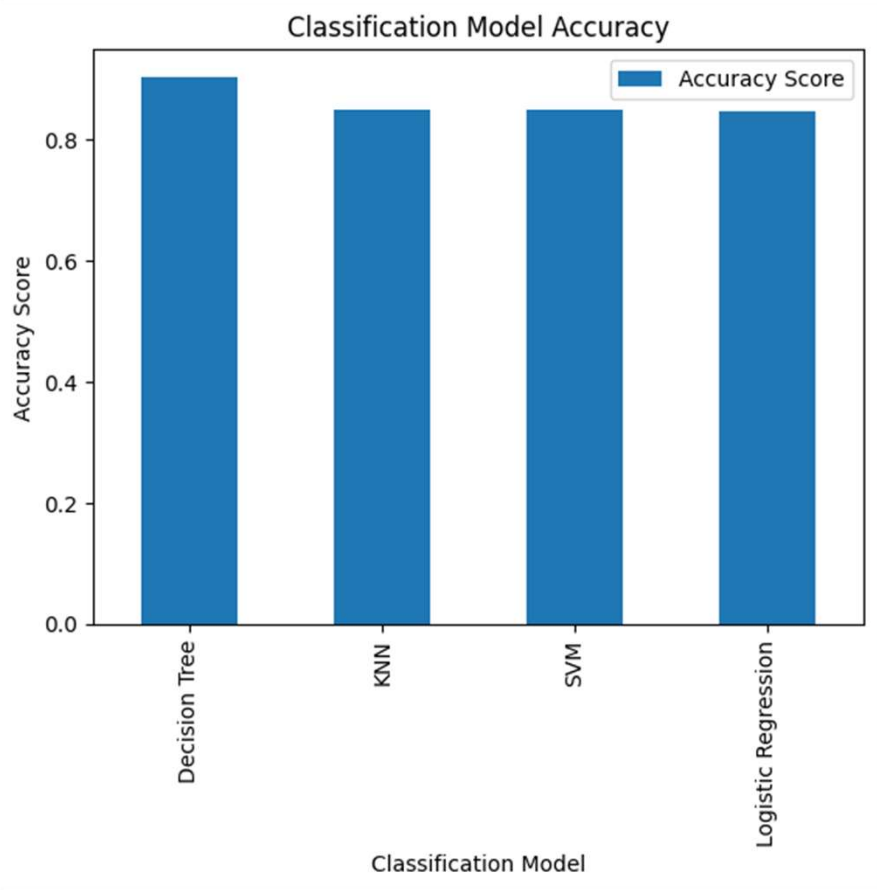


From the scatter plot on left at bottom, we observed that Payload mass raged from 2000 to 5000 Kg have a highest number of successful landings.

Section 5

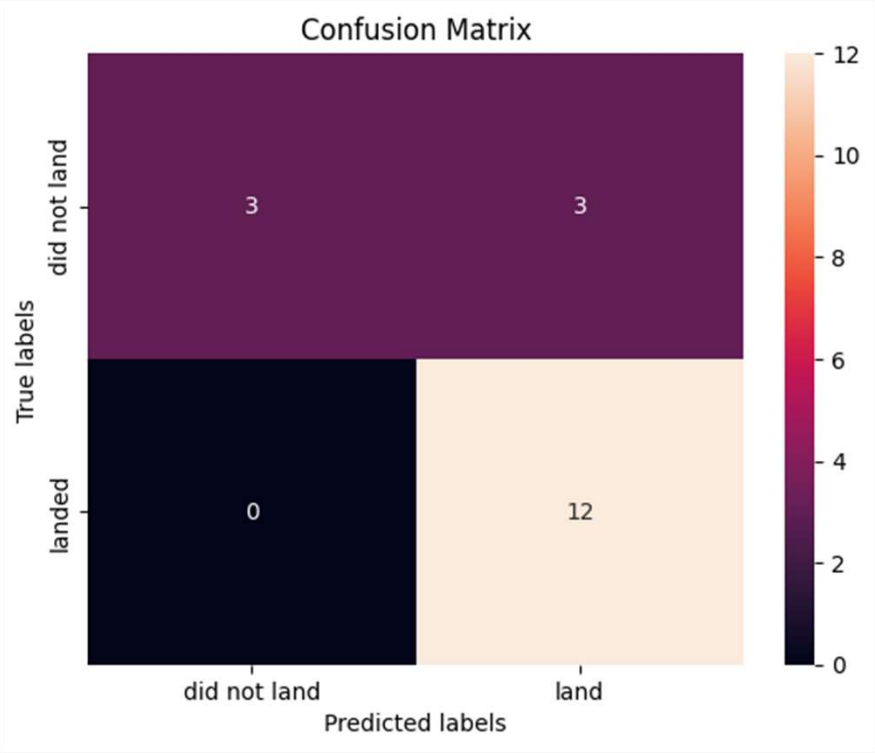
Predictive Analysis (Classification)

Classification Accuracy



- From the Barchart on left, we compared that the machine learning classification model with their respective accuracy score plotted against it.
- So, we inferred that Decision Tree classification algorithm had the highest accuracy score of 90.35%.

Confusion Matrix



- By plotting the confusion matrix between the predicted and true vales, we observed that , it has 3 - False Positives (FP) and 0 –False Negatives (FN). On other side it has 3- True Positive (TP) and 12- True Negatives (TN) .
- The Accuracy of the Confusion Matrix is calculated by the formula given below:
- Accuracy= (TP + TN)/TP+FP+TN+FN
- Therefore, overall the calculated accuracy is 83.33%. This means there is error or misclassification rate of around 16%.

Conclusions

- SpaceX had really achieved a lot in terms of making a cost-effective space travel, with a steadily launch success rate increased from 2013 , and achieved 90% success rate at 2020.
- The launch site ‘KSC LC-39A’ had a highest successful launches , whereas ‘CCAFS SLC-40’ has a lowest successful rate of launches.
- The orbits ES-L1,GEO, HEO, SSO have a 100% success rate. VLEO have a notable success of approx.85% success rate and GTO have a lowest success rate amongst the other orbit types.
- The launch flight with Payload Mass ranged from 2000 to 5000 kg have a highest number of successful launches.
- The best performing Machine Learning classification is Decision Tree classification model with an accuracy score of 90.35%. From the Confusion Matrix on the test data, we observed that there is an accuracy of 83.33%.
- The above findings helpful for put a effective competition by SpaceY for achieving cost-effective , efficient space travel ambition.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project
- The Python, SQL codes, Notebook outputs and other detailed things are used for the project are presented in their respective slides.

Thank you!