IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Name: O Nithin Sharma
Date:12-06-2023

Outline

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

Executive Summary

- Methodologies used in project:
- For Data Collection and Data Set Creation Web Scrapping and Spacex Api have been used.
- Then the collected is wrangled and cleaned using python libraries.
- The modified data is undergone into Exploratory Data Analysis with SQL and Data Visualization Techniques.
- The data is visualized using Folium library and also creation of Dashboard to show changes in the output with respect different inputs.
- Machine Learning Techniques are applied for prediction
- Summary of all results
- Results of Data Analysis
- Some Pictures of Dashboard
- Best Machine Learning for prediction

Introduction

- Project background and context
- SpaceX is most popular and successful company which is an Aircraft Manufacturer, Launcher and Satellite Communication Company.
- Our main object is to find the different aspects to be included in the company SpaceY to be able to compute with SpaceX.
- Problems you want to find answers
- Estimate how different variables such payload mass, launch site, number of flights, orbits for will be by predicting successful landing.
- Which is the best site for launching?
- Best algorithm for classification?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected through two ways

    i) SpaceX API

    ii) Web Scrapping from Wikipedia
- Perform data wrangling
  - Identifying and removing Missing Values
  - Performing One Hot Encoding
  - Filtering the Data
- Perform exploratory data analysis (EDA) using visualization and SQL
- SQL is used to modify the data and python visualization tools are used make analysis

- Perform interactive visual analytics using Folium and Plotly Dash
- Performed visual analysis using Folium to create and justify the observations obtained through data
- Dashboard is created to understand success rate for each site and different Payload Mass Range
- Perform predictive analysis using classification models
- Collected data has been normalized using Python library and various Machine learning techniques for classification are applied for different parameter using Grid Search to obtain the best parameters. And the model is evaluated on the test set obtained by the Train test split library. After total evaluation a particular model is selected and prediction are made for new data.

Data Collection

- The Data is collected through two different techniques using two different sources:

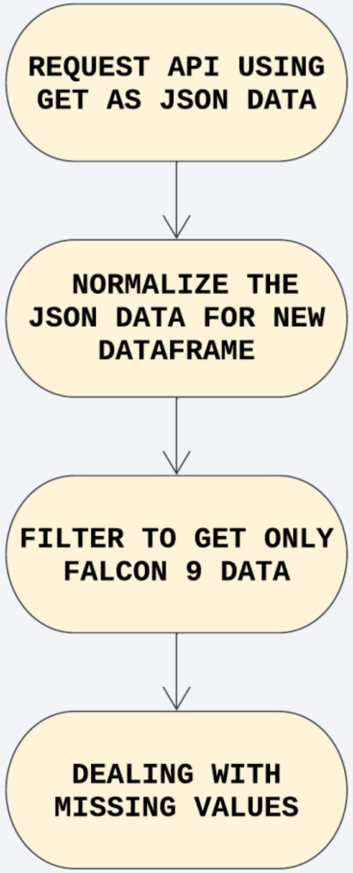
i)Web Scrapping from Wikipedia (Link: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)

ii) SpaceX offers a public API and data is collected through API .

(Link:https://api.spacexdata.com/v4/rockets/)

Data Collection – SpaceX API

- Data collection with SpaceX REST calls is presented in the flow chart showing the important stages of data collection
- GitHub URL:(https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb)



REQUEST API USING GET AS JSON DATA

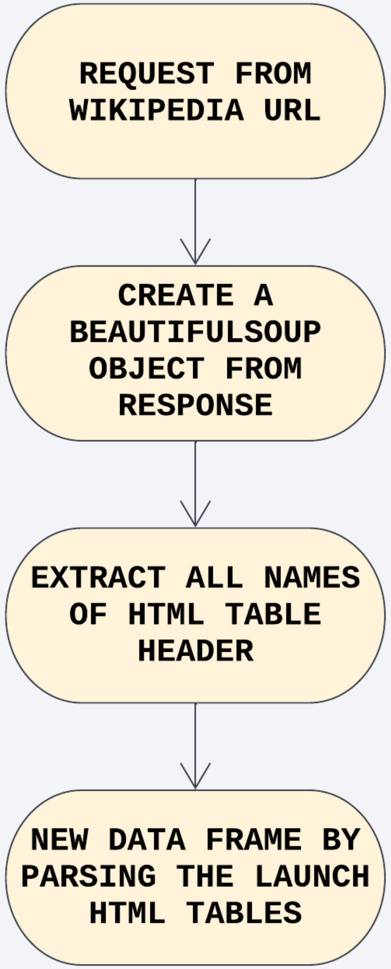
NORMALIZE THE JSON DATA FOR NEW DATAFRAME

FILTER TO GET ONLY FALCON 9 DATA

DEALING WITH MISSING VALUES

Data Collection - Scraping

- Data collection through Scraping has presented in the flow chart with important stages
- GitHub URL : https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



REQUEST FROM WIKIPEDIA URL

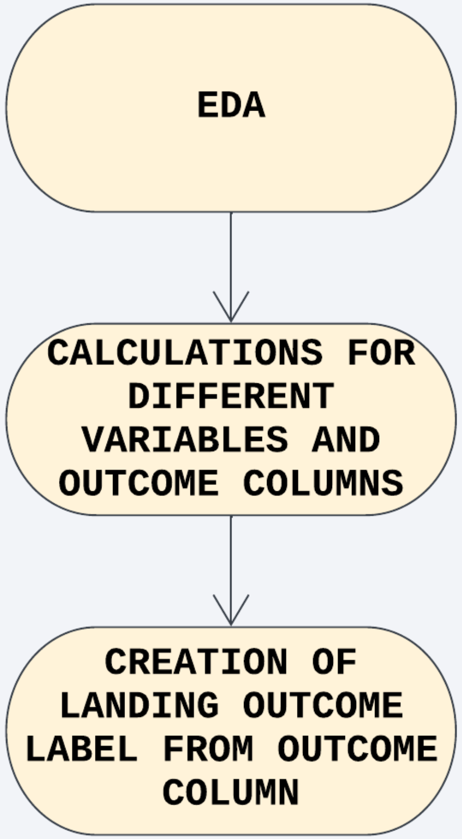
CREATE A BEAUTIFULSOUP OBJECT FROM RESPONSE

EXTRACT ALL NAMES OF HTML TABLE HEADER

NEW DATA FRAME BY PARSING THE LAUNCH HTML TABLES

Data Wrangling

- Some Data Analysis are done on the newly obtained data and some calculation and summaries are made.
- Output landing label is created from Outcome column
- GitHub URL : https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_1_L3_labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb


EDA
CALCULATIONS FOR DIFFERENT VARIABLES AND OUTCOME COLUMNS
CREATION OF LANDING OUTCOME LABEL FROM OUTCOME COLUMN

EDA with Data Visualization

- Relationship between different features like :

1. Flight Number vs Payload Mass
2. Flight Number vs Launch Site
3. Payload Mass vs Launch Site
4. Success rate for different orbit type
5. Flight Number vs Orbit Type
6. Payload Mass vs Orbit Type
7. Launch Success yearly

This graphs are plotted obtained to get more understanding and to do feature engineering for the data.

GitHub URL: https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_2_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- The SQL queries performed are:

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

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

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

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

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

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

➢ List the total number of successful and failure mission outcomes

➢ List the names of the booster_versions which have carried the maximum payload mass

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

➢ Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- GitHub URL : https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Markers, Circles, Make Clusters and lines are used in building and Interactive Map with Folium Library
- Markers indicate different launch sites on the the map with each shown with their respective names.
- Circles and Make clusters are for identifying no.of launches from each site and in specific launch site the no.of successful and unsuccessful launches with different color markers.
- Lines are used to calculate and identify distance between launch sites and nearby land marks.
- GitHub Link: https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A Dashboard is created showing two different plots:

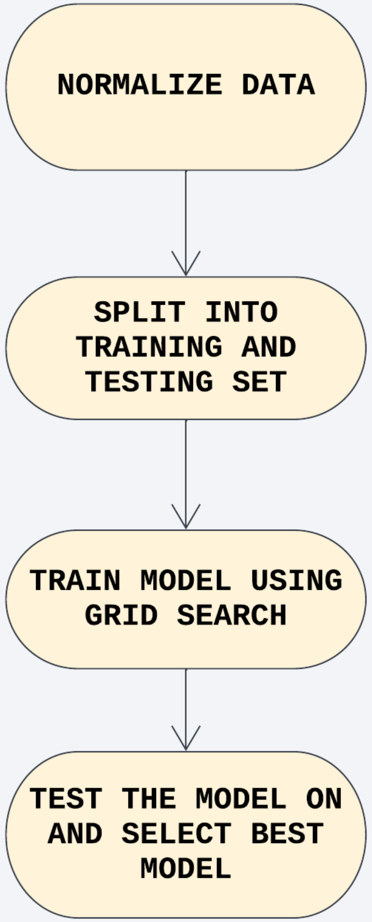
i) Pie Chart : It represents percentage successful launches for all launch sites and also shows percentage of successful and unsuccessful launches for a specific launch sites.

ii) Scatter Plot: It represents scatter plot of successful launches for a range of payload mass and scatter plot of successful and unsuccessful launches for specific launch sites and different range of payload mass.

GitHub URL: https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- The data is normalized and split into training and testing set with 80% used for training and 20% for testing.
- And different models are trained used the training data and performed evaluation using testing set.
- GitHub URL: https://github.com/Nithin029/Coursera-Applied-Data-Science-Capstone/blob/main/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb


NORMALIZE DATA
SPLIT INTO TRAINING AND TESTING SET
TRAIN MODEL USING GRID SEARCH
TEST THE MODEL ON AND SELECT BEST MODEL

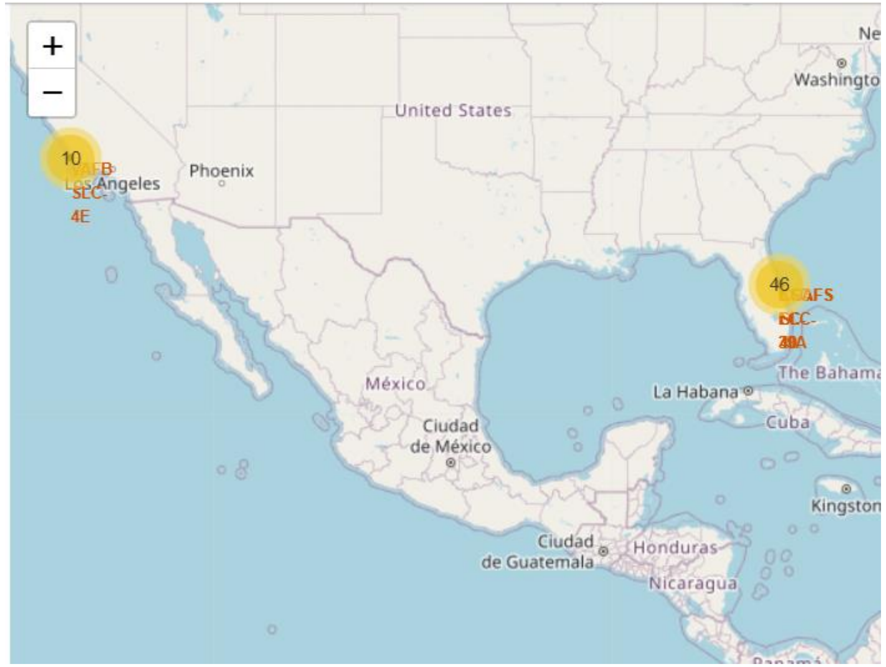
Results

- **Exploratory data analysis results**
- There are 4 different launch sites
- . Average payload mass carried by booster version F9 v1.1 is 2928.5 kg
- The first successful landing outcome in ground pad was achieved in the year 2015
- F9 booster versions have more success rate and the average success rate is increasing year by year.

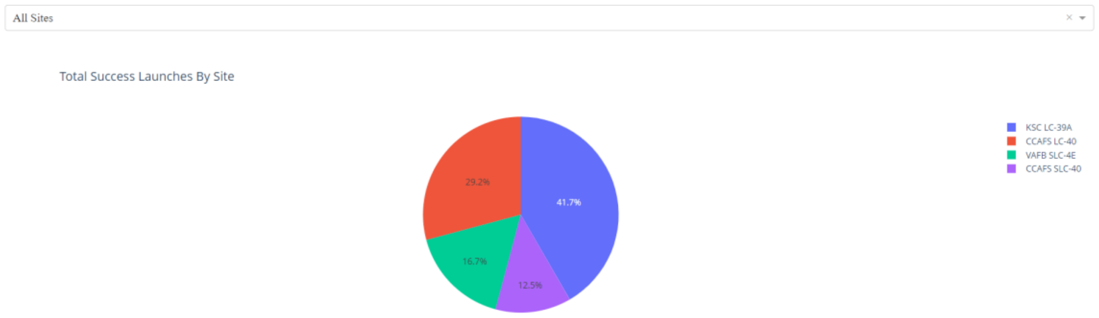
Results

- Interactive analytics demo in screenshots :

Folium Library Screenshot:



Dashboard Screenshot:



Results

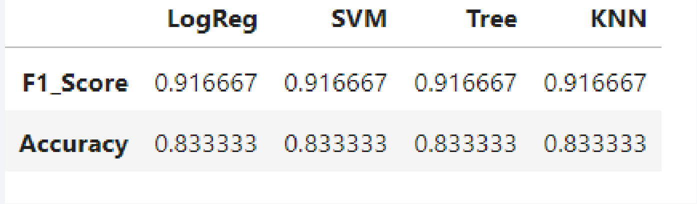
- Predictive analysis results :
- Different classification algorithm are applied like

i) Logistic Regression

ii) Support Vector Machine

iii) K-Nearest Neighbourhood

iv) Decision Trees



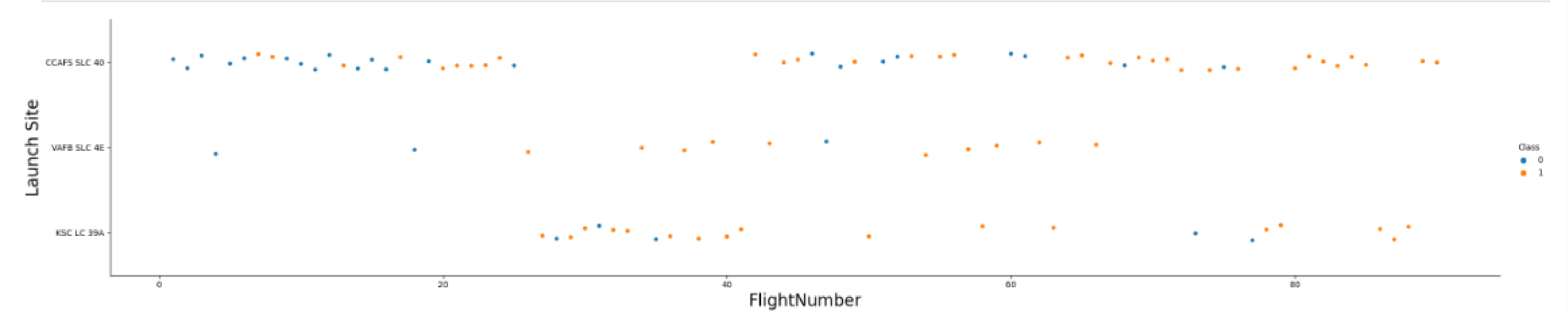
	LogReg	SVM	Tree	KNN
F1_Score	0.916667	0.916667	0.916667	0.916667
Accuracy	0.833333	0.833333	0.833333	0.833333

- The Decision Tree Algorithm is selected has more training accuracy and equal testing accuracy with respect to other algorithms.

Section 2

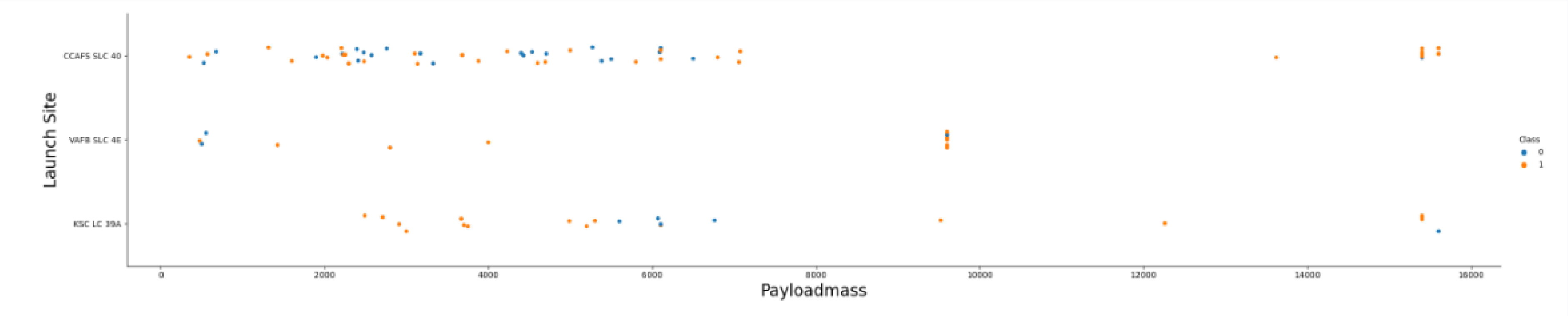
Insights drawn from EDA

Flight Number vs. Launch Site



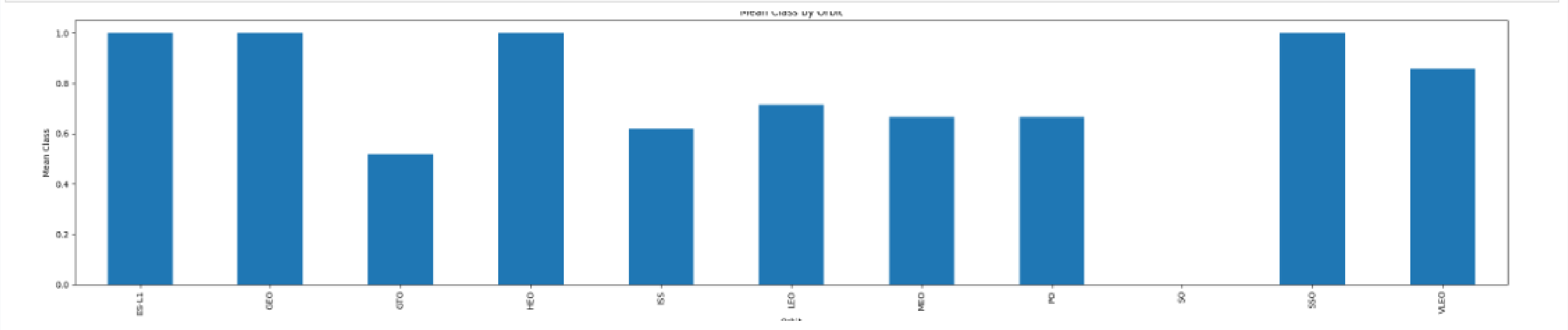
- The Flight Number greater than 80 has 100 success rate while in between 60 to 80 its is nearly 98% and as the Flight number decreases the successrate is decreased and according to the viewpoint of launch site most launch happened in the site CCAFS SLC 40 having approximately equal succes and failure rate compare to other sites where two site where launches are less but have higher success rate

Payload vs. Launch Site



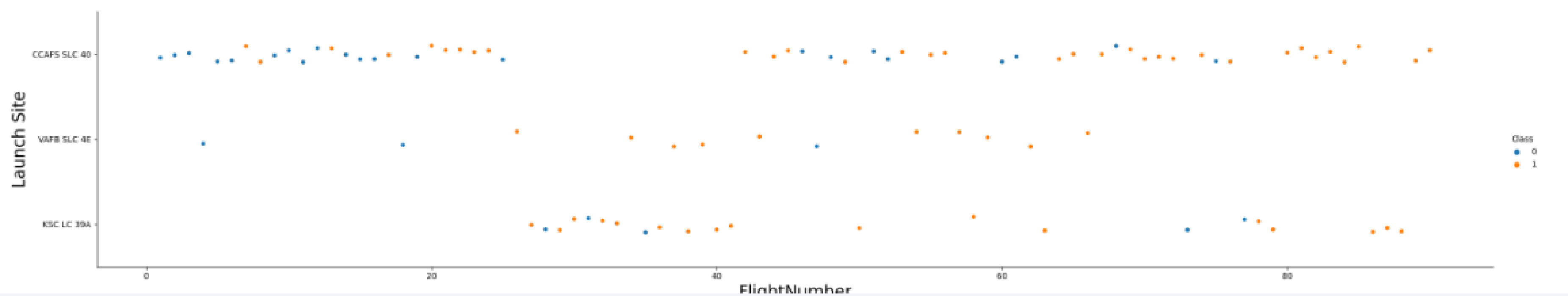
- The VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



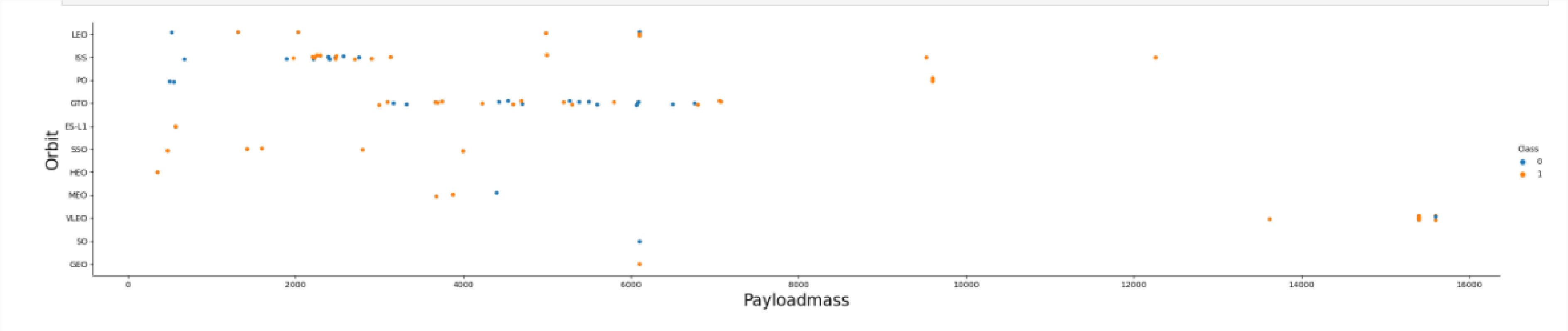
- The biggest success rate are ES-L1,GEO,HEO AND SSO with 100%

Flight Number vs. Orbit Type



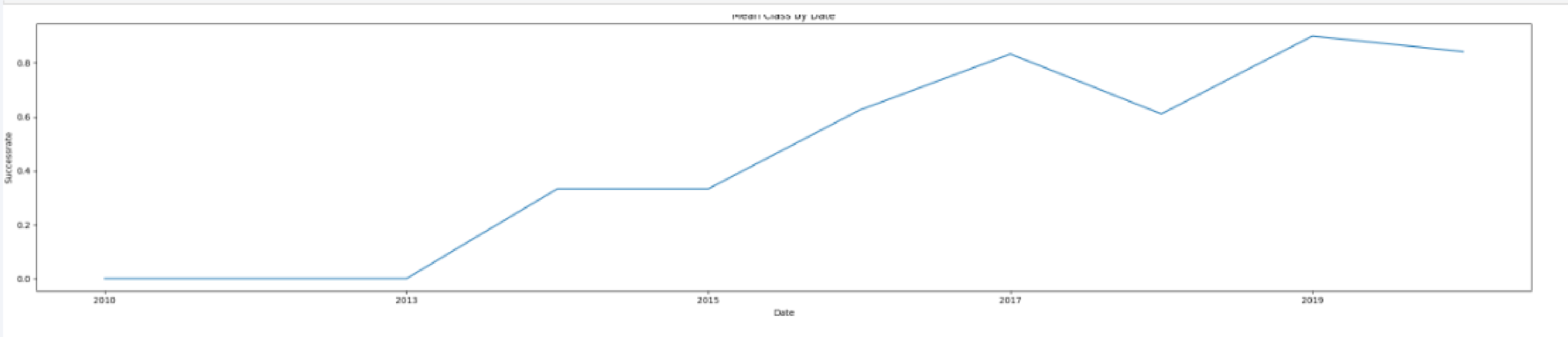
- The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



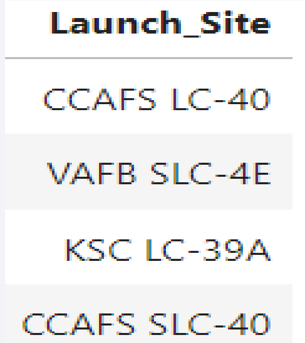
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- For GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- The success rate since 2013 kept increasing till 2020

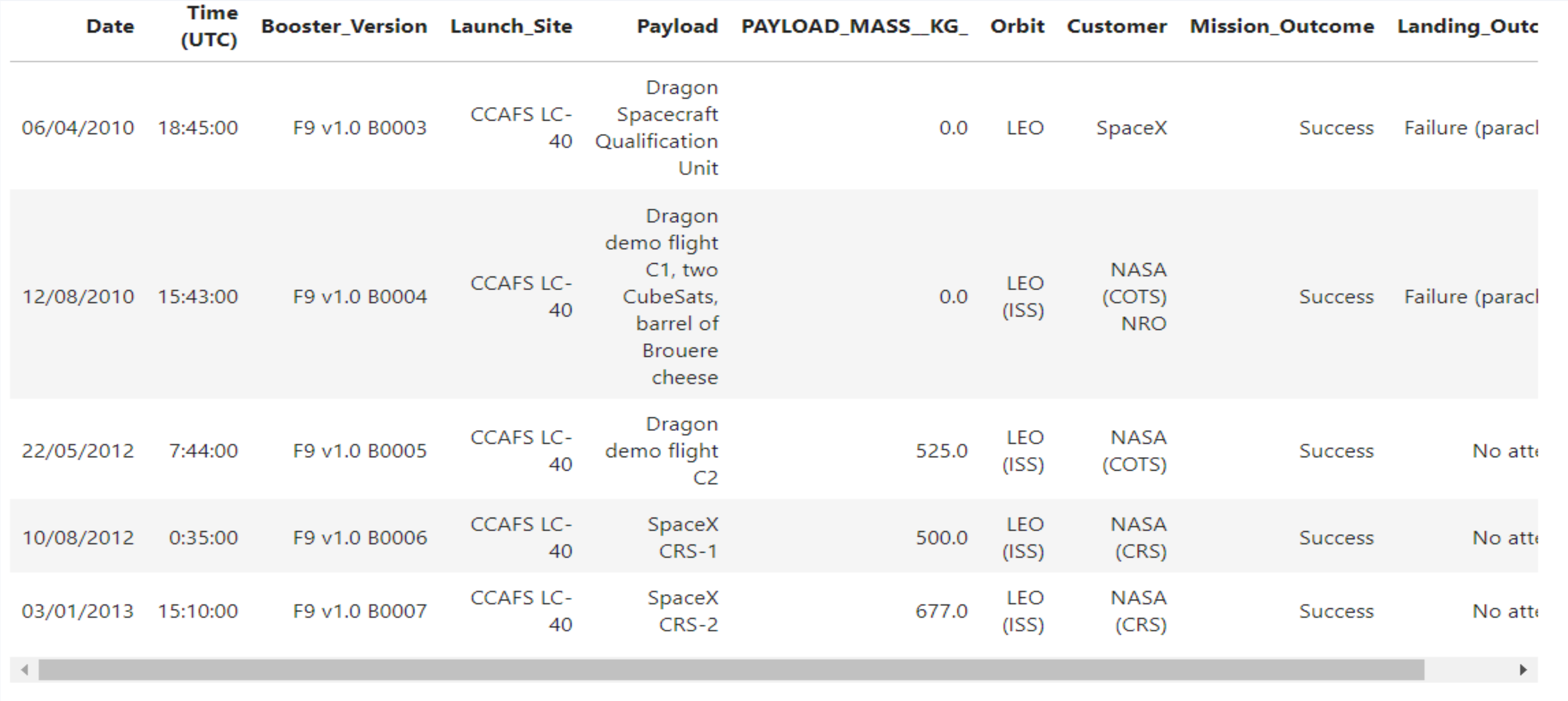
All Launch Site Names



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

There are 4 different launch sites and we can obtain this unique launch site from the column Launch_Site.

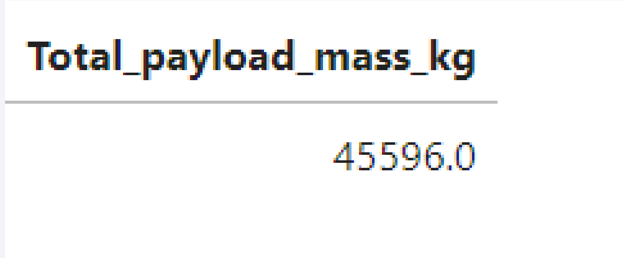
Launch Site Names Begin with 'CCA'



Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (paracl
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (paracl
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No att
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No att
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No att

This the different columns of 5 different rows whose Launch Site Name begins with ‘CCA’

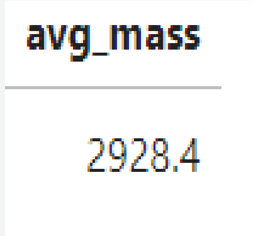
Total Payload Mass



Total_payload_mass_kg
45596.0

- This displays the total payload mass carried by boosters launched by NASA (CRS)

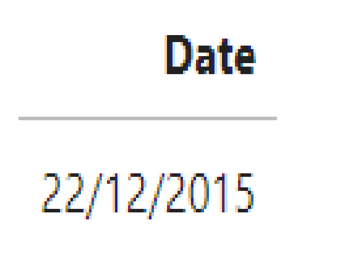
Average Payload Mass by F9 v1.1



avg_mass
2928.4

- The average payload mass carried by booster version F9 v1.1 is displayed above.

First Successful Ground Landing Date

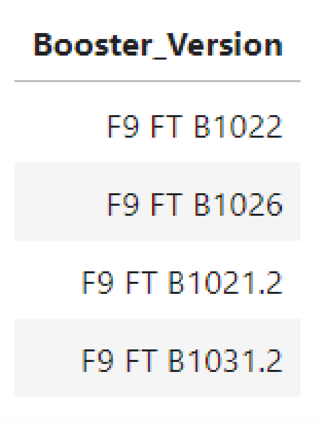


Date
22/12/2015

- The image presents the date when the first succesful landing outcome in ground pad was acheived.

Successful Drone Ship Landing with Payload between 4000 and 6000

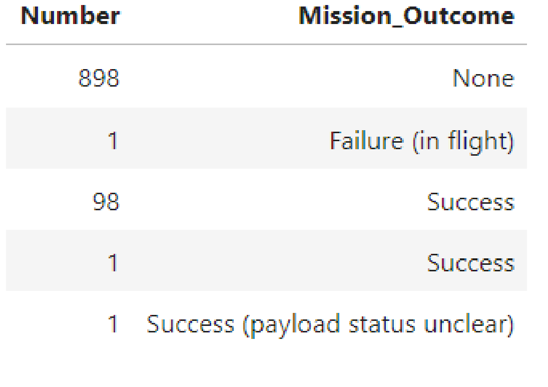
- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are displayed below



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

Total Number of Successful and Failure Mission Outcomes

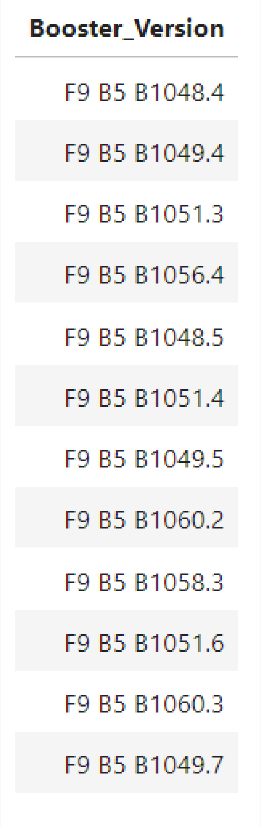
- The total number of successful and failure mission outcomes are presented below



Number	Mission_Outcome
898	None
1	Failure (in flight)
98	Success
1	Success
1	Success (payload status unclear)

Boosters Carried Maximum Payload

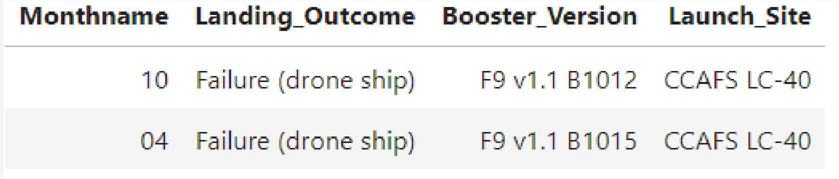
- The names of the booster which have carried the maximum payload mass are listed below



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

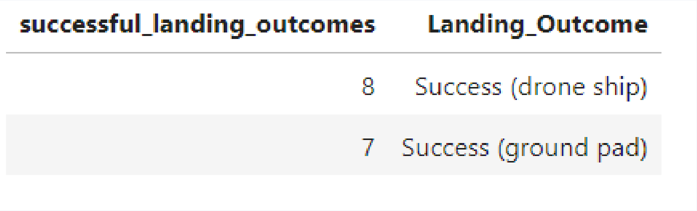
- The failed landing_outcomes in drone ship, their booster versions, and launch site names and month number for in year 2015 are



Monthname	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- The count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order are presented below:

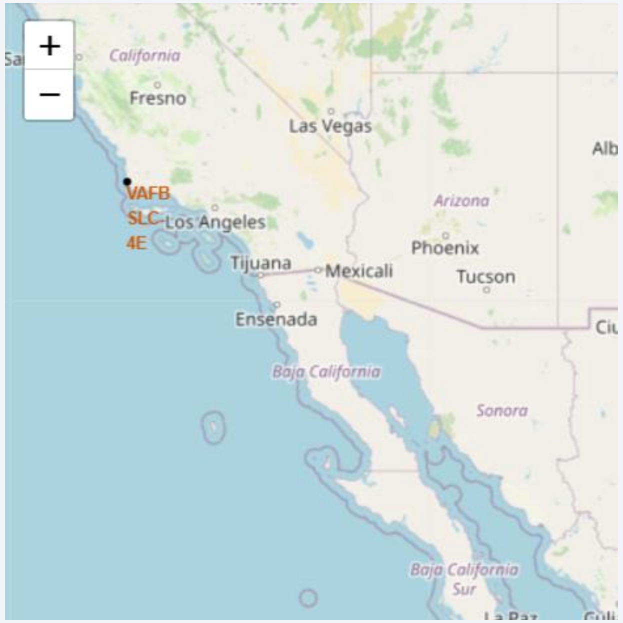


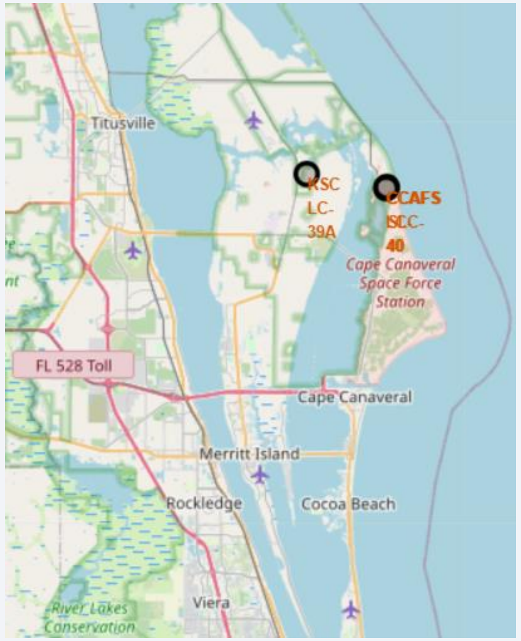
successful_landing_outcomes	Landing_Outcome
8	Success (drone ship)
7	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

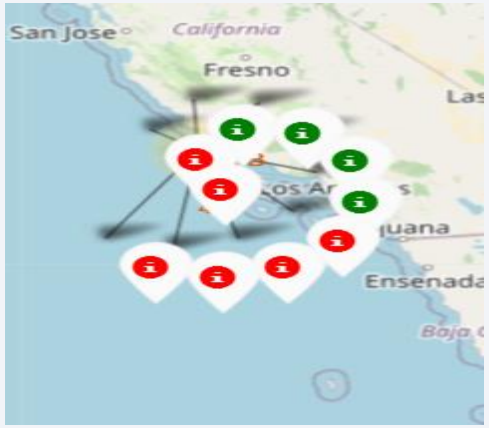
Launch Site Locations

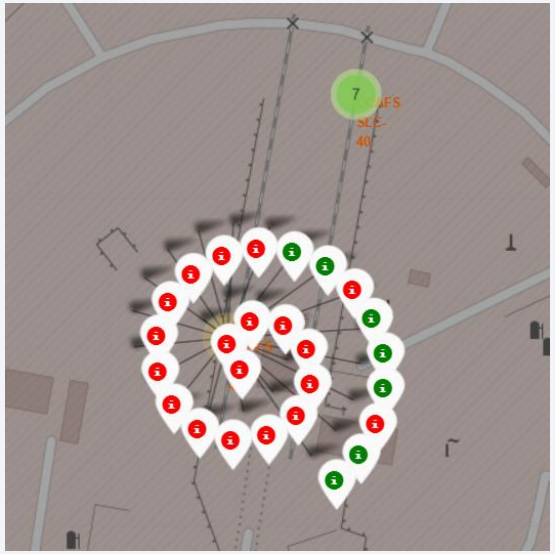


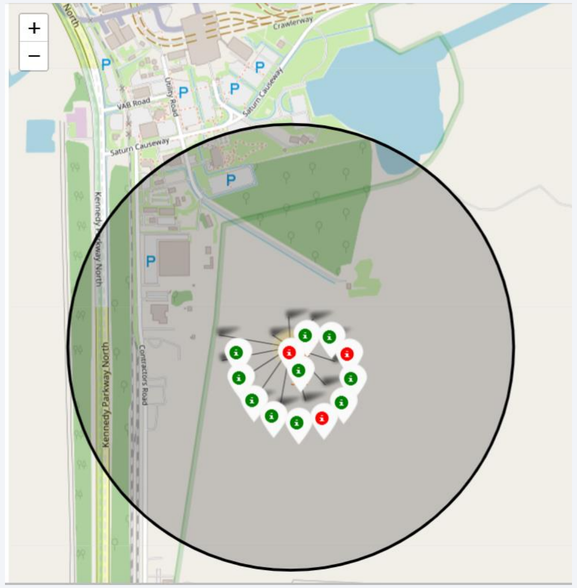


- Left and Right figures different launch sites on the two sides of the America map.

Launch Outcomes

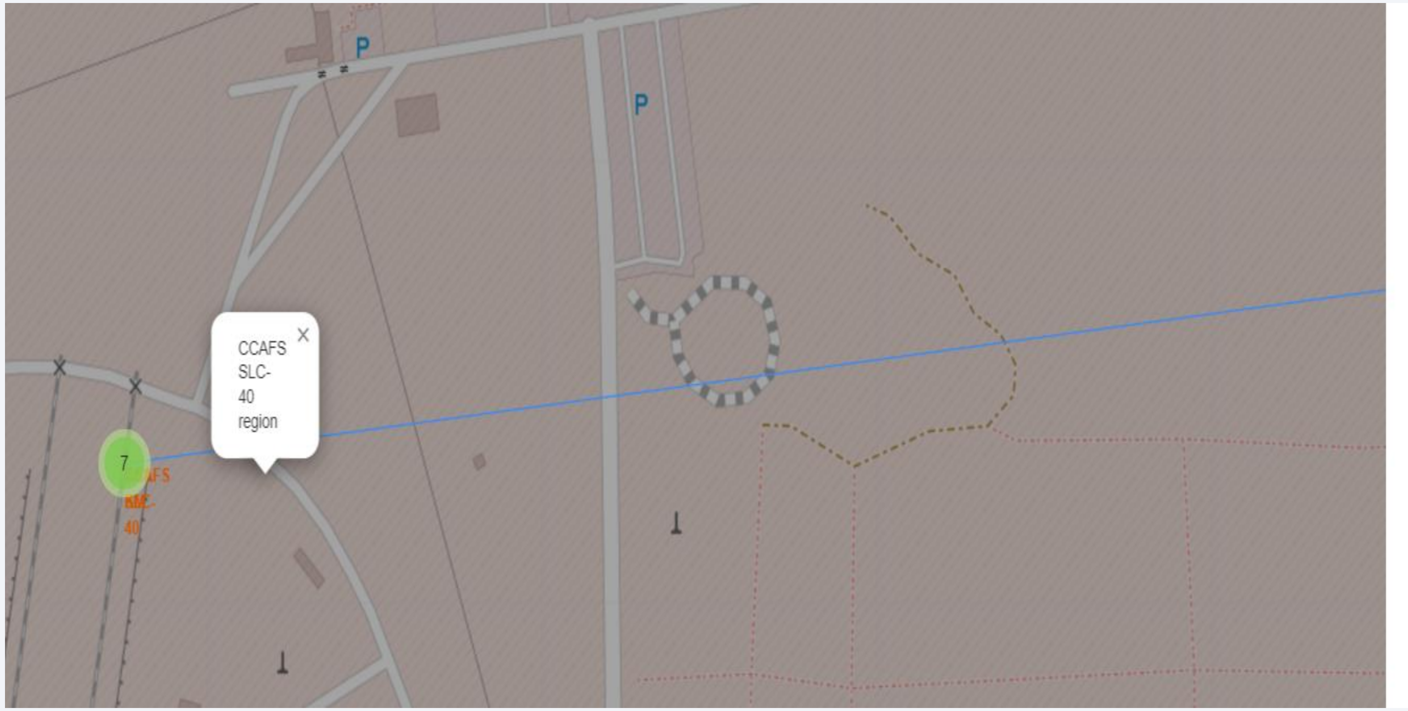






- The green represents successful launch outcome and red represents failure launch outcome
- And we observe KSC-LC 39A more success rate

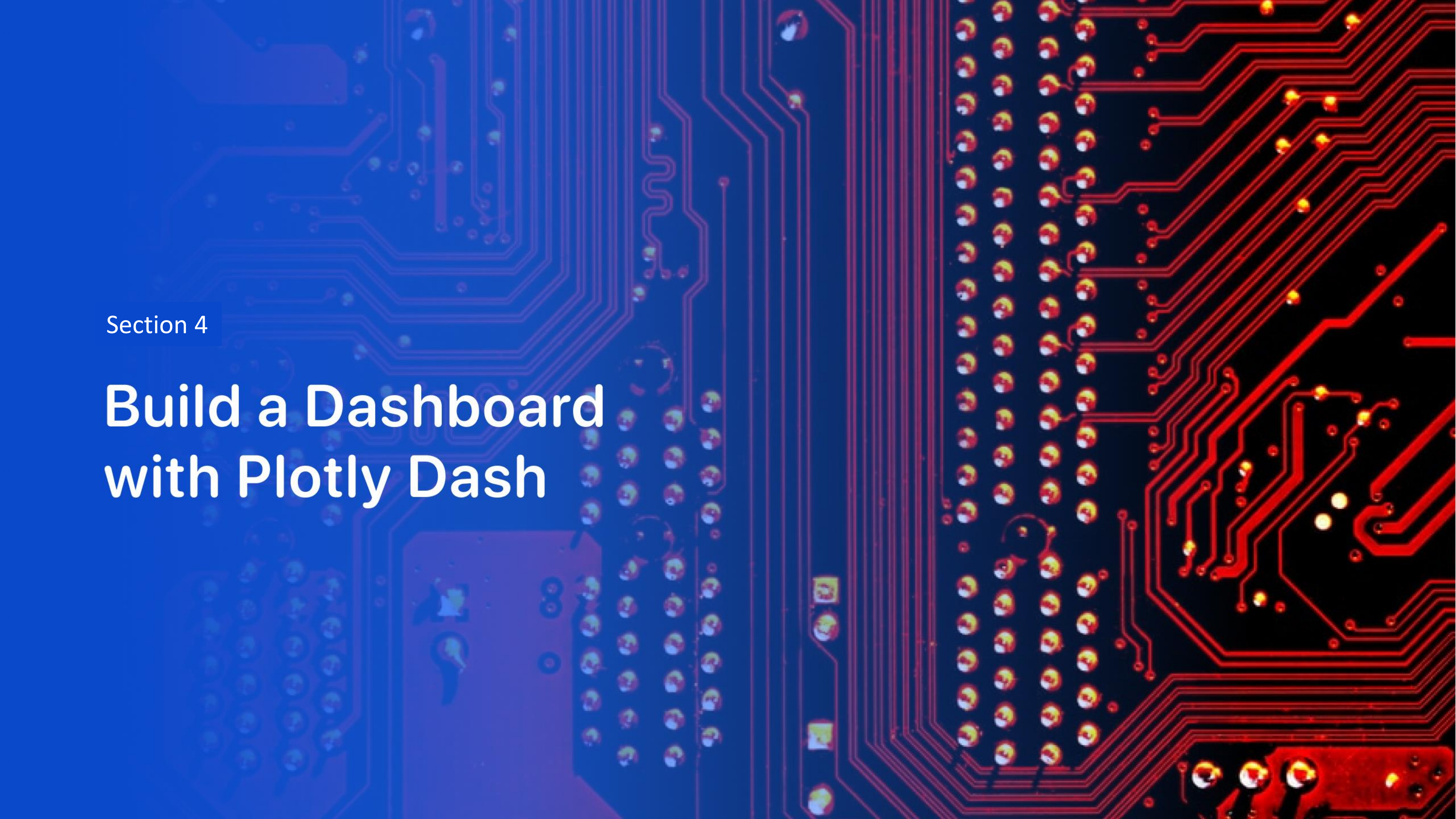
Distance between Launch Site and Coastline



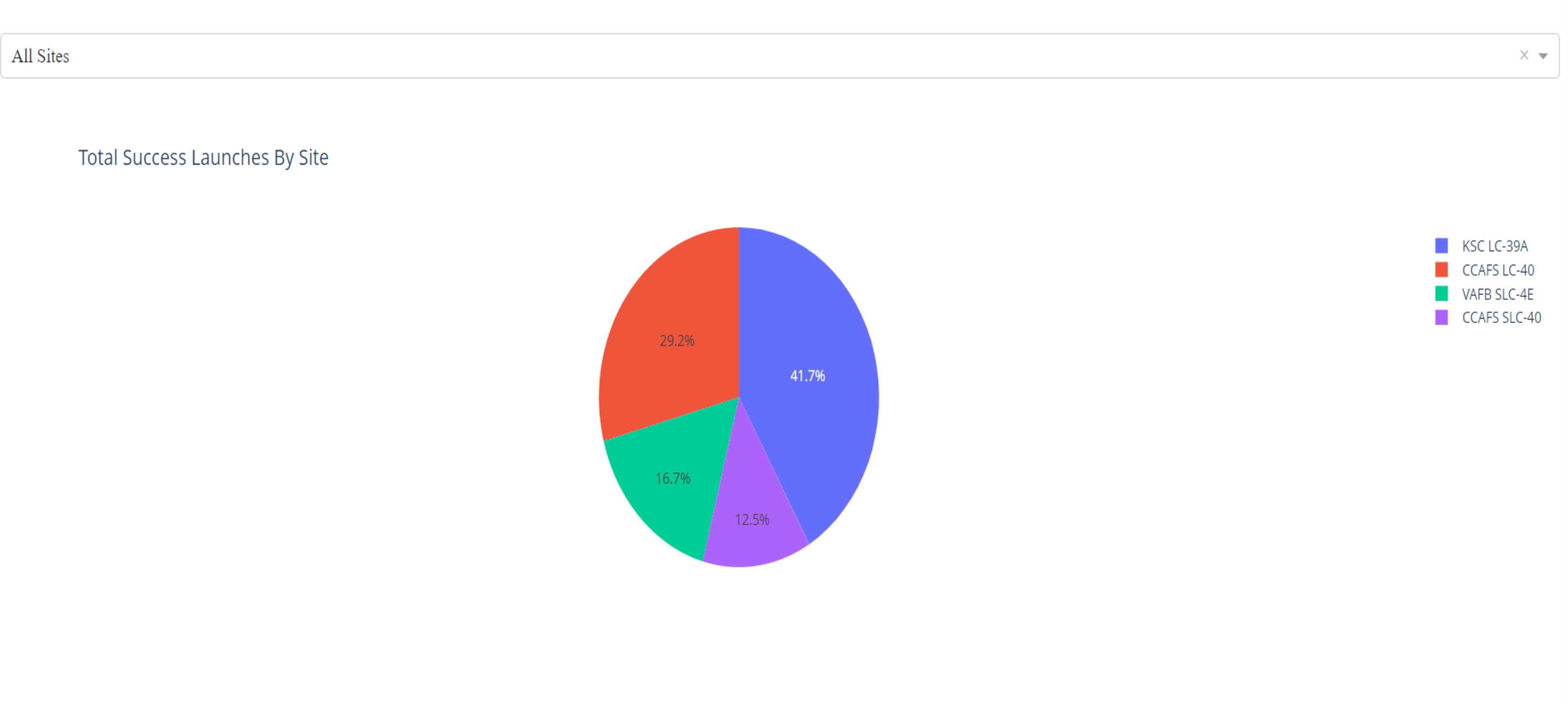
- The distance between Launch Site CCAFS SLC-40 and Nearest Coastline is represented by the blue line and the value is 0.50974

Section 4

Build a Dashboard with Plotly Dash



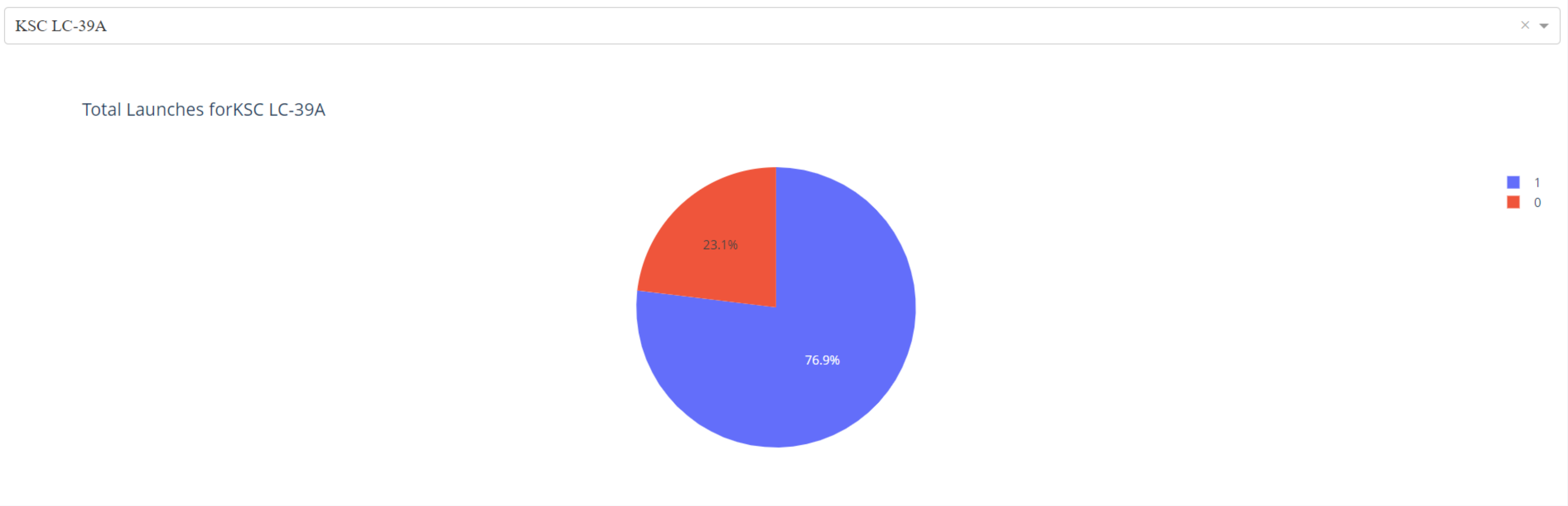
Success Percentage For all Sites


All Sites

- The pie chart shows the percentage of contribution of different launch sites for successful launches.

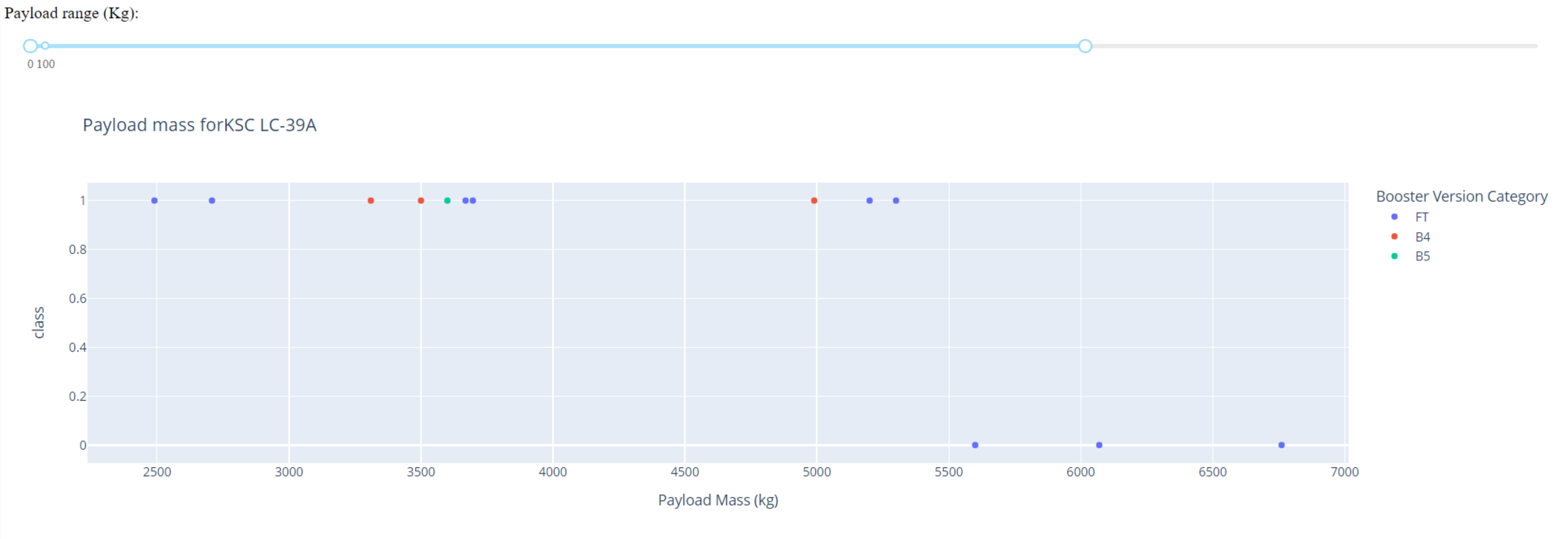
Success Percentage Of Site with highest launch success ratio

- The site with highest launch success ratio is KSC-LC 39A



KSC LC-39A

Payload Range



Payload range (Kg):

0 100

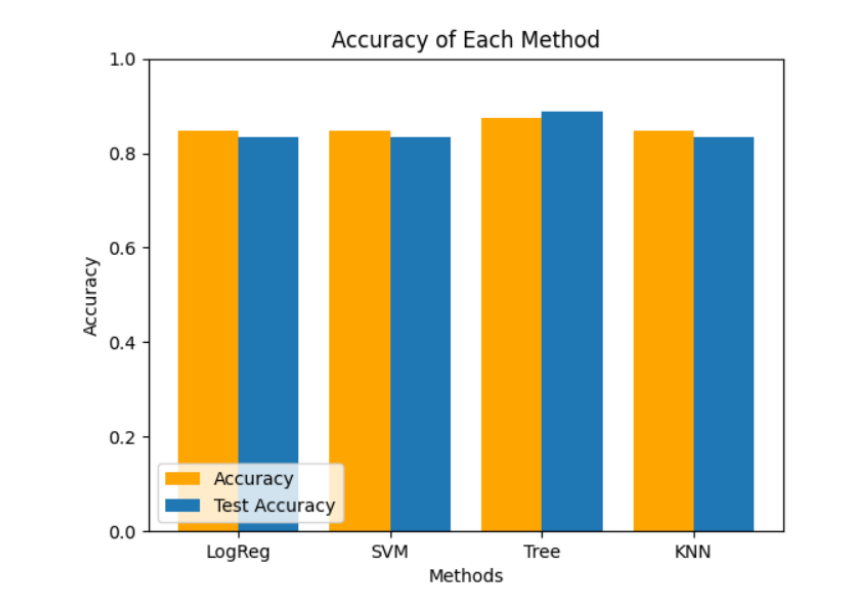
- Payload Mass under 6000 make successful combination for the site KSC LC -39A

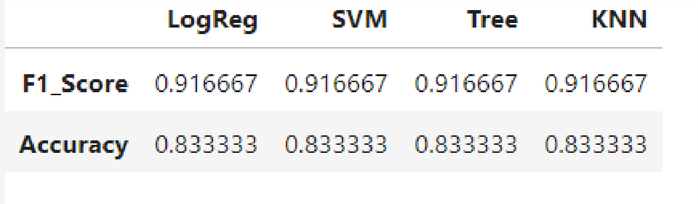
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The images helps to obtain the best model for prediction.
- The Decision Tree is best model because it has more training accuracy and has same f1 score and testing accuracy as for all the methods.

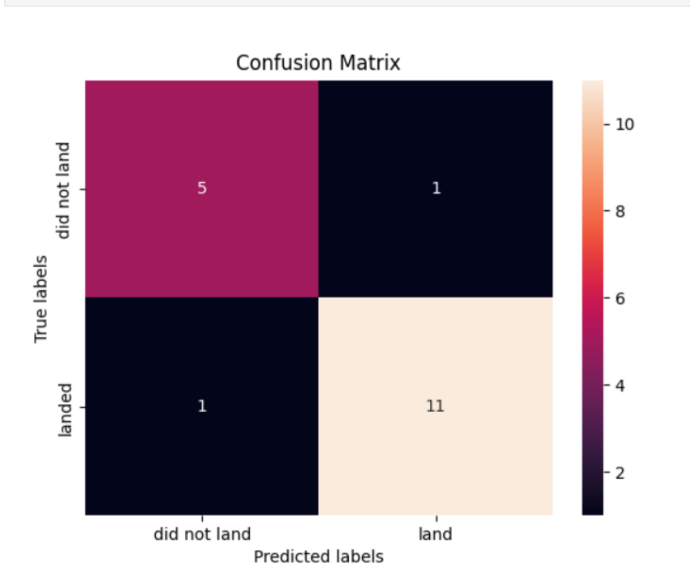




	LogReg	SVM	Tree	KNN
F1_Score	0.916667	0.916667	0.916667	0.916667
Accuracy	0.833333	0.833333	0.833333	0.833333

Confusion Matrix

- The confusion matrix of the best performing model i.e Decision Tree



- The confusion matrix justifies the accuracy calculated by having less false positive and false negative.

Conclusions

- The best launch site is KSC-LC 39A because of its higher success rate
- The best combination of payload mass should be in the range of 2500 and 6000 kg
- The Decision Tree model should able predict pretty well for the unseen data as it provides more accuracy than other models.
- The accuracy can be improved by adding more data and can also establish more logical relationships between different features.

Appendix

- GitHub is unable show the maps so it can be downloaded and verified.
- And I added a code to calculate f1 score to evaluate the model.

Thank you!