Winning Space Race with Data Science

IBM Data Science Capstone Project – Space X

Jay Soni

Outline

Executive Summary

- First, collect data from official website then preprocess the data after that we visualize the data using matplot, plotly and folium to get better insights of data pattern after that apply machine learning algorithms like logistic regression , support vector machine , decision tree and knn after we select best model according to accuracy.
- For visualization , we use matplot-lib to knows the patterns then use plotly to get information about data after that we use folium to create dashboard.

Introduction

- Project background and context:
- - Project is based on finding the best circumstances to land a rocket successfully so that we can reuse parts which will cut down price of rocket launch compare to other companies.

- Problems you want to find answers:
- - Success ratio of rocket landing.
- - Best place for rocket launch.

SPACEX

Space Exploration Technologies

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Data is collected from Past launches of SpaceX data using SpaceX rest API.
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection – SpaceX API

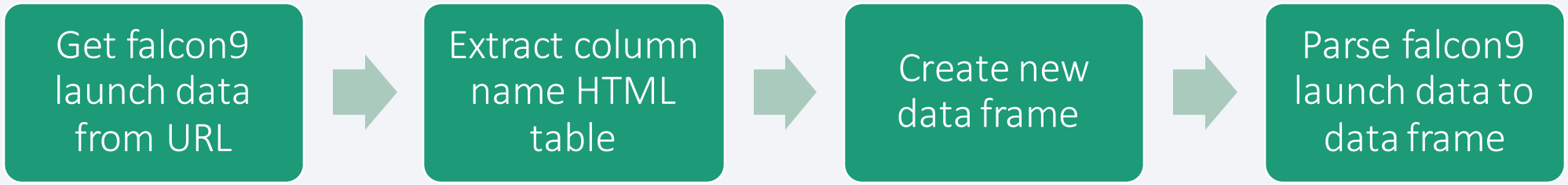
- Github URL to Notebook

Request	Request SpaceX API to gather data
Clean	Clean the collected data
Filter	Filter the only Falcon 9 data

Data Collection - Scraping

Github URL to Notebook

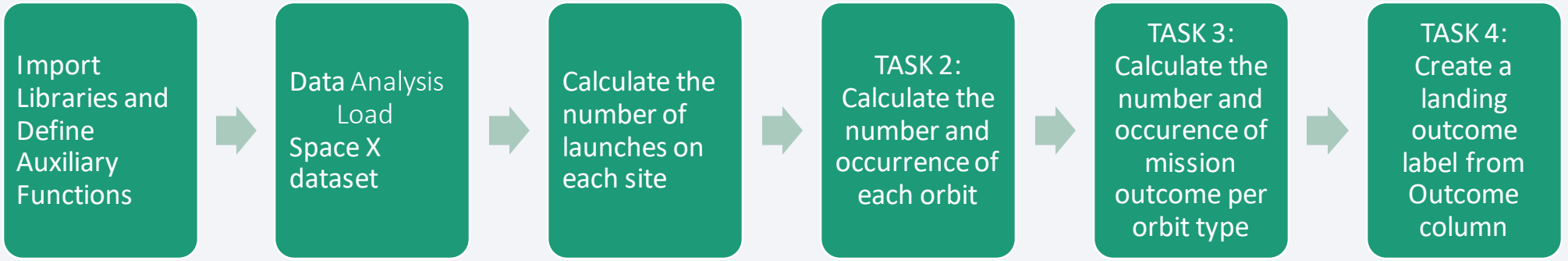
Flow-chart of data scraping:



Data Wrangling

Github URL to notebook

Flow-cahrt of Data Wrangling:



EDA with Data Visualization

Github URL to notebook

- Use catplot to plot Flight Number vs Launch Site to Visualize the relationship between Flight Number and Launch Site.
- To Visualize the relationship between Payload and Launch Site we plot a scatter plot.
- Plot bar chart to Visualize the relationship between success rate of each orbit type.
- Plot another Scatter plot to Visualize the relationship between FlightNumber and Orbit type.
- Visualize the relationship between Payload and Orbit type we plot scatter plot.
- Visualize the launch success yearly trend, plot line chart.

EDA with SQL

Github URL to link

- For example of some questions we were asked about the data we needed information about. Which we are using SQL queries to get the answers i n the dataset :

- 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 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.

Build an Interactive Map with Folium

- Create circle of different colours to identify different space center
- Use marker for different to represent launch with color, for successful launch use green color and for un-successful launch use red color.
- Use mousepoint feature to get cordinates of proximites and then use line object to calculate distance
- We also use line to vizualize dsitance between launch site to airpot

- Github URL to Notebook

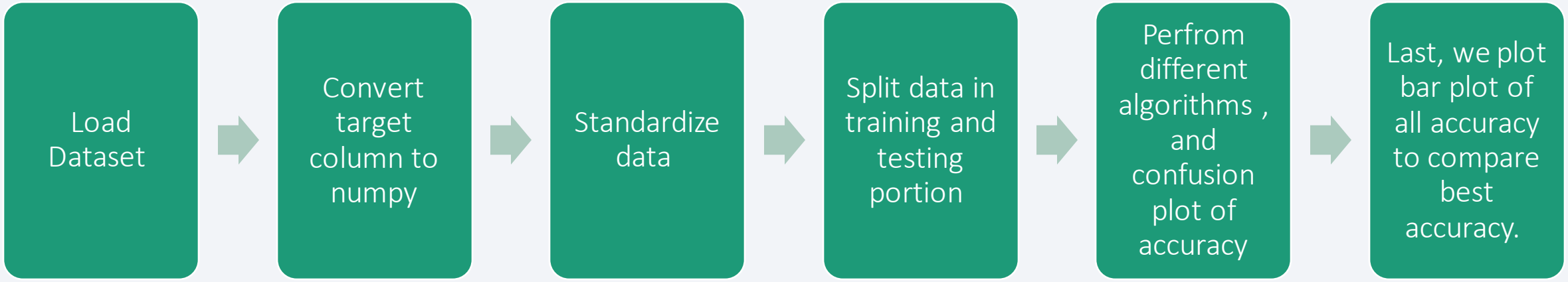
Build a Dashboard with Plotly Dash

- Here, We plot a pie chart of all launch site and their success ratio using plotly.
- We also add scatter plot to dashboard to check the different payload mass range compared to payload orbit with marker which is dynamic so that it became easy to get insights from scatter plot.

- Github URL to notebook

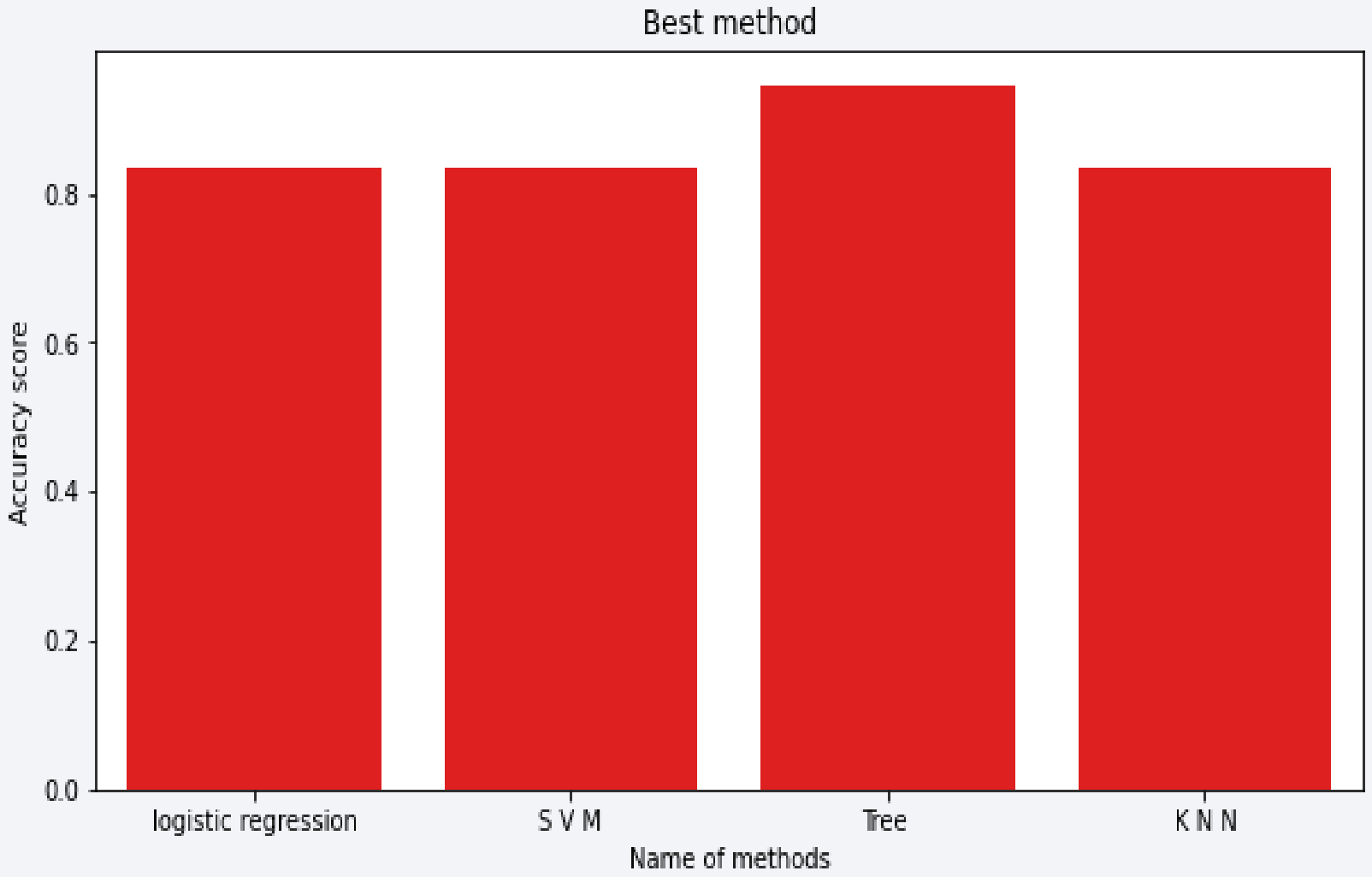
Predictive Analysis (Classification)

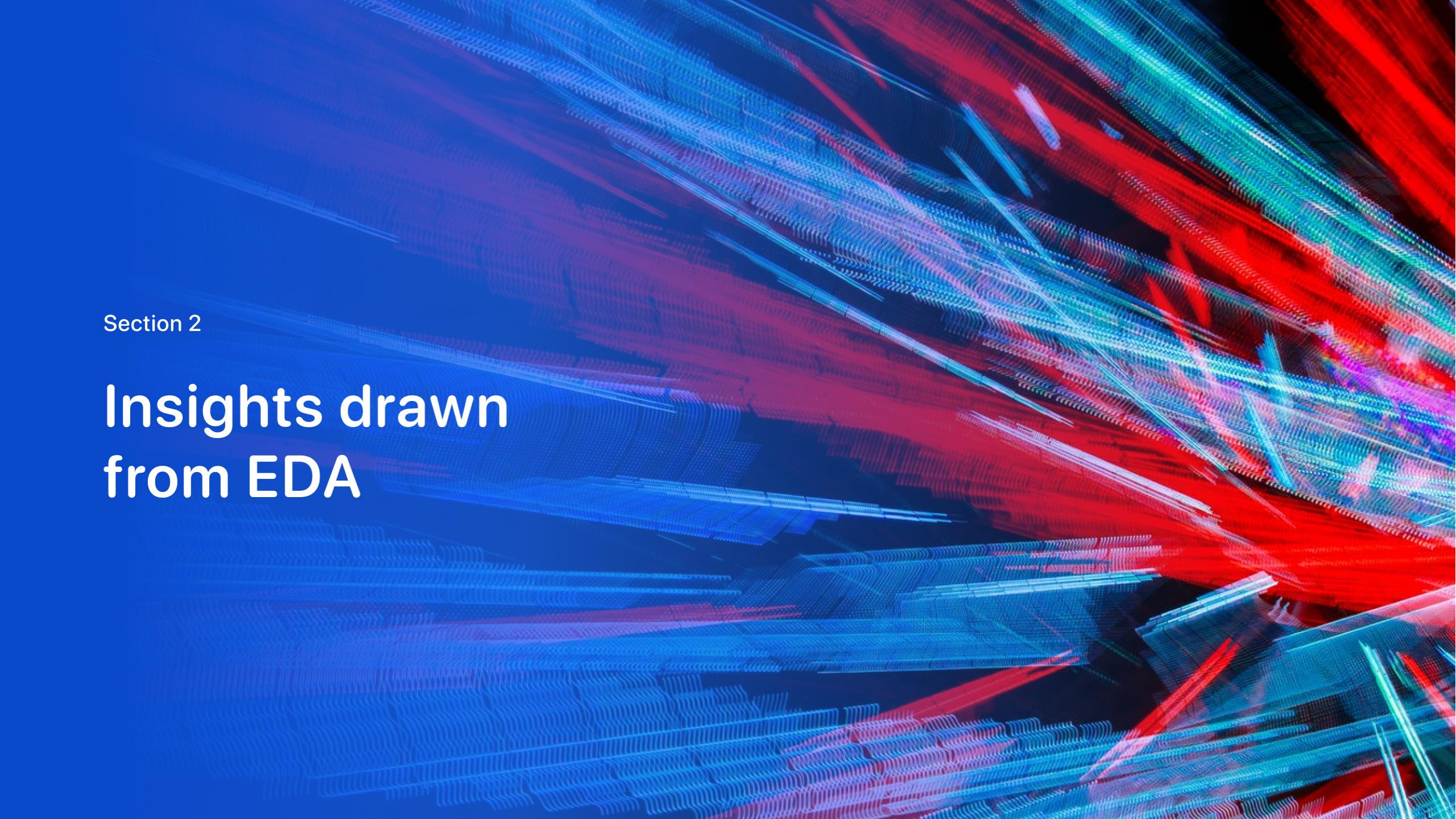
Github URL to notebook



Results

- Decision Tree has the highest accuracy among all the algorithms. (95%)



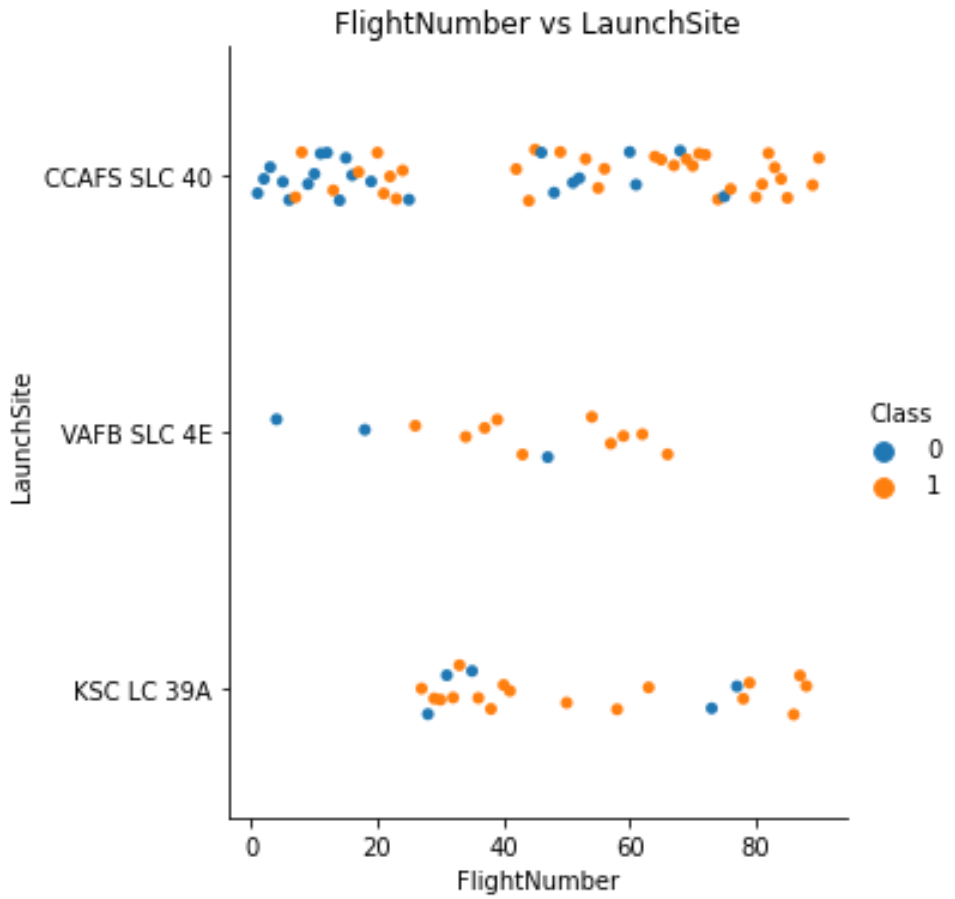


Section 2

Insights drawn from EDA

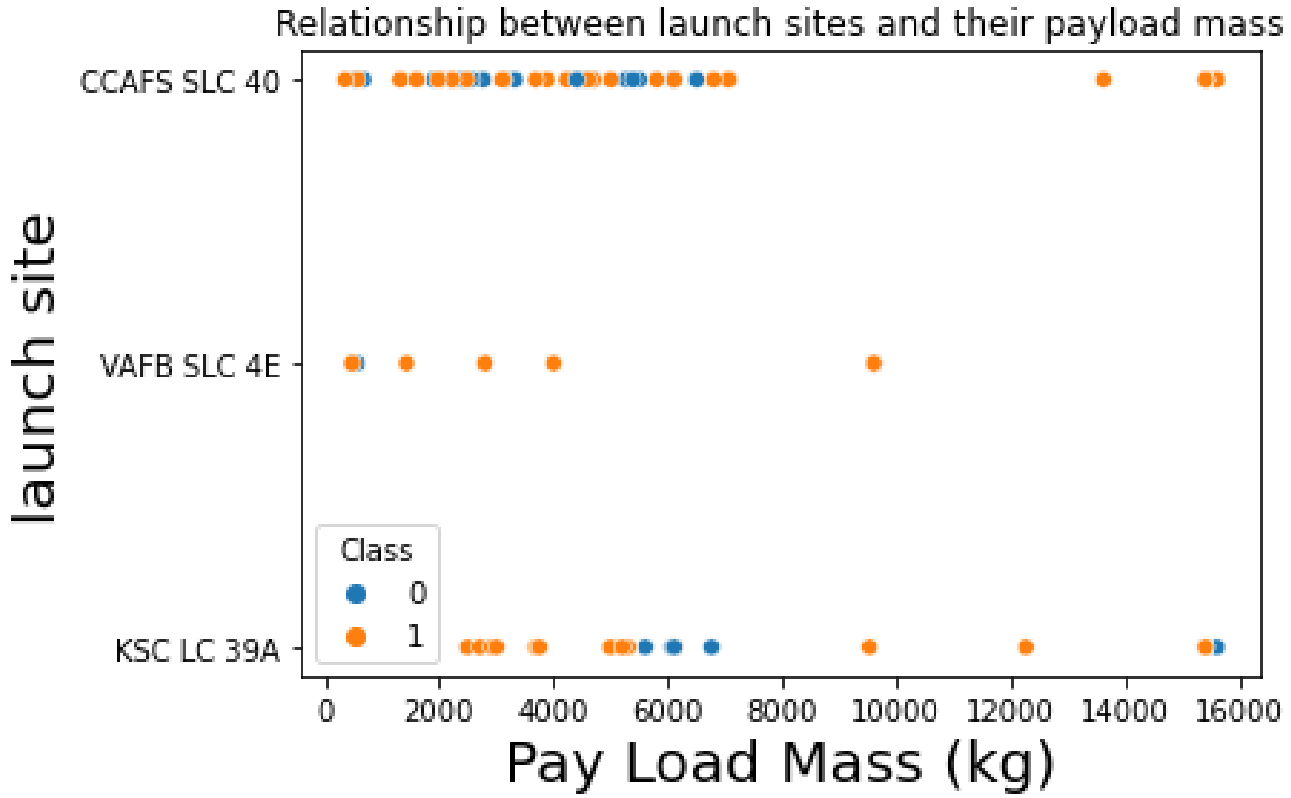
Flight Number vs. Launch Site

- From the scatter plot we can say that most of the rockets are launched from launch site "CCAFS SLC 40" and it has mix ratio of



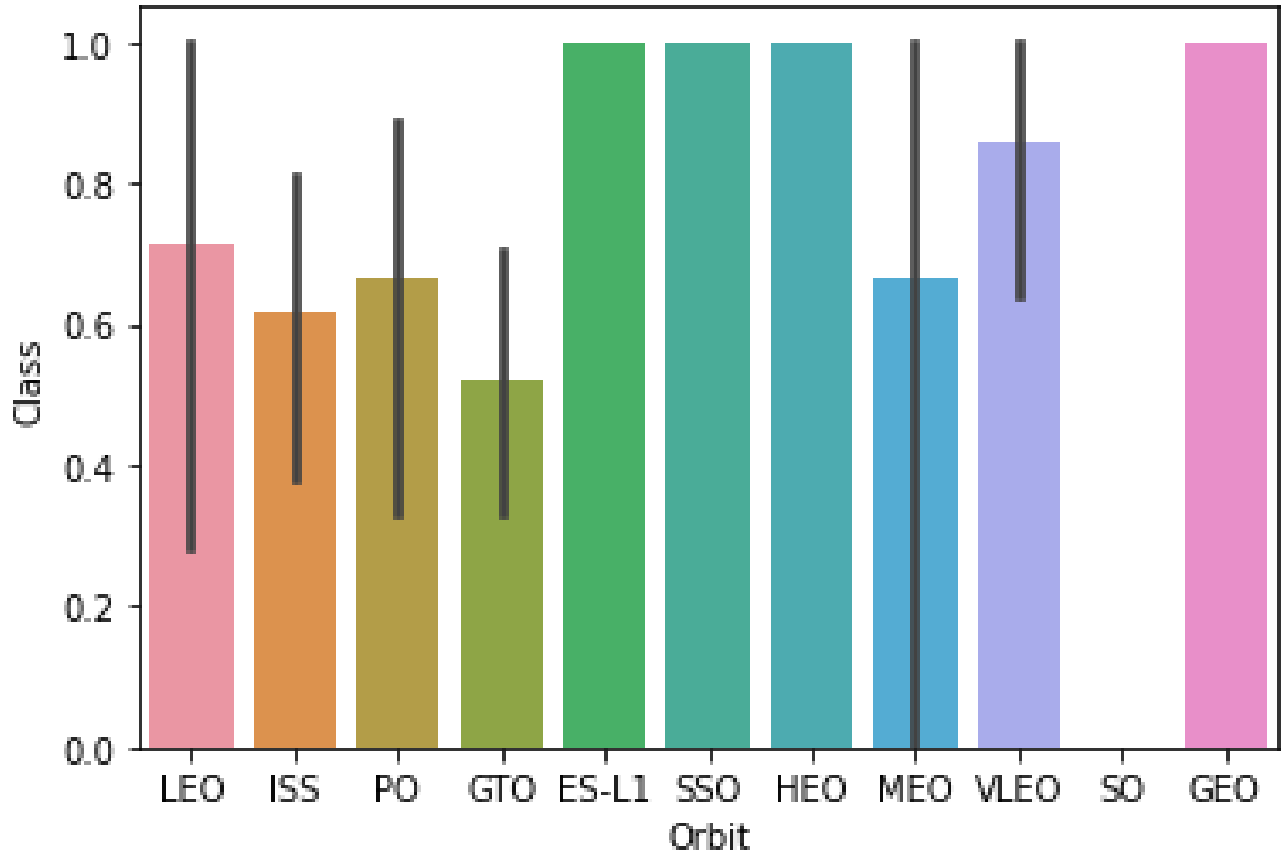
Payload vs. Launch Site

- From this scatter plot we observe that most of rockets are launch within 10,000 kg of mass
- If you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launch-site there are no rockets launched for heavy payload mass(greater than 10000).



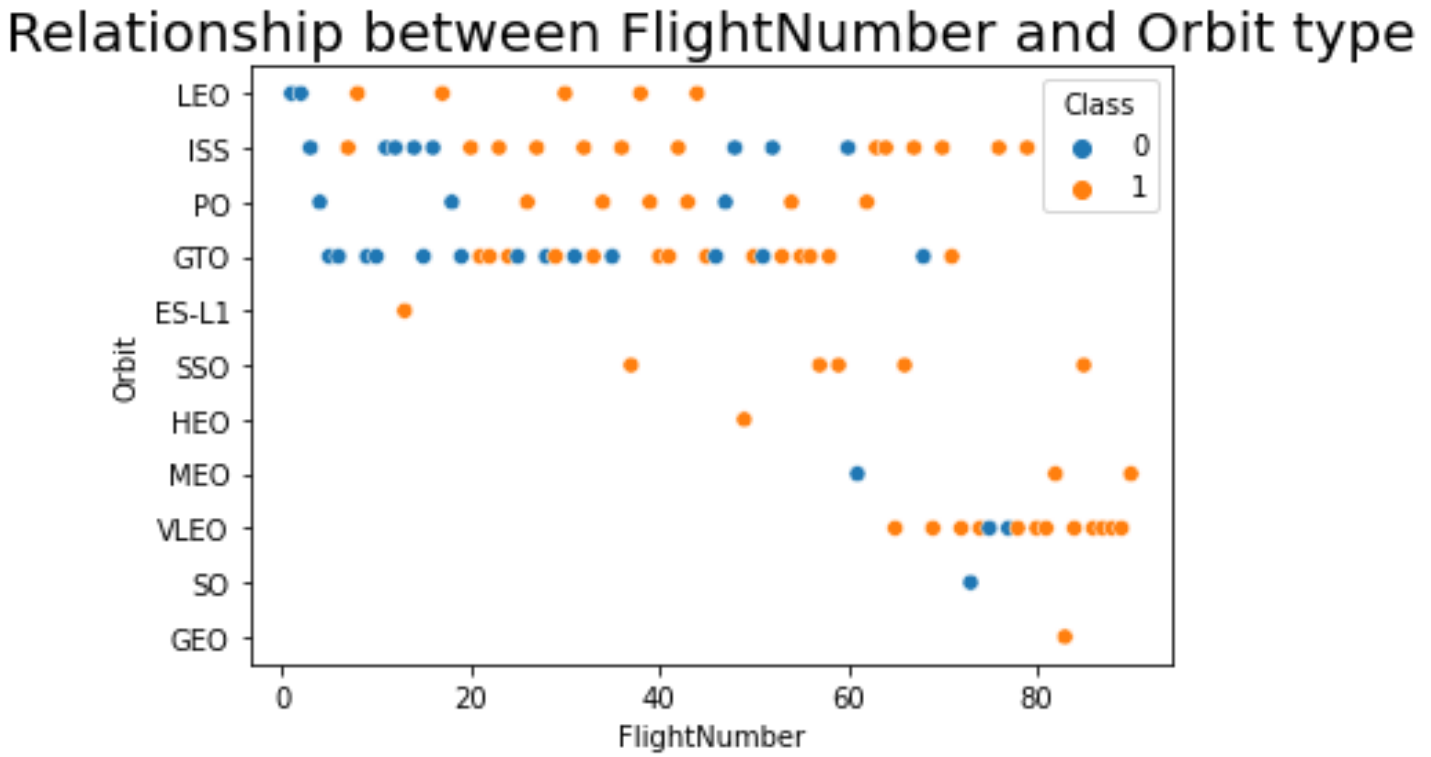
Success Rate vs. Orbit Type

- After analyzing the bar plot we can see that orbit "SO" has the zero success rate.



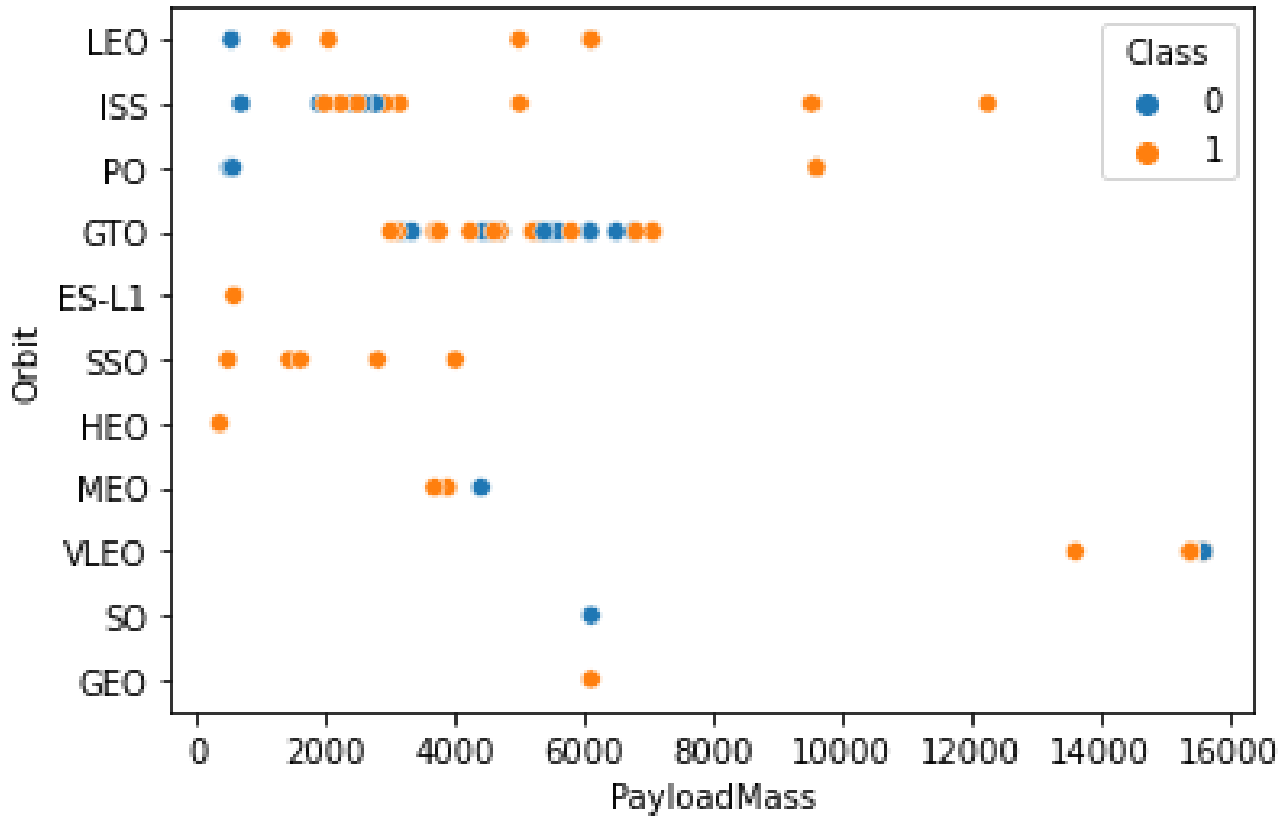
Flight Number vs. Orbit Type

- As we can see that in 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.



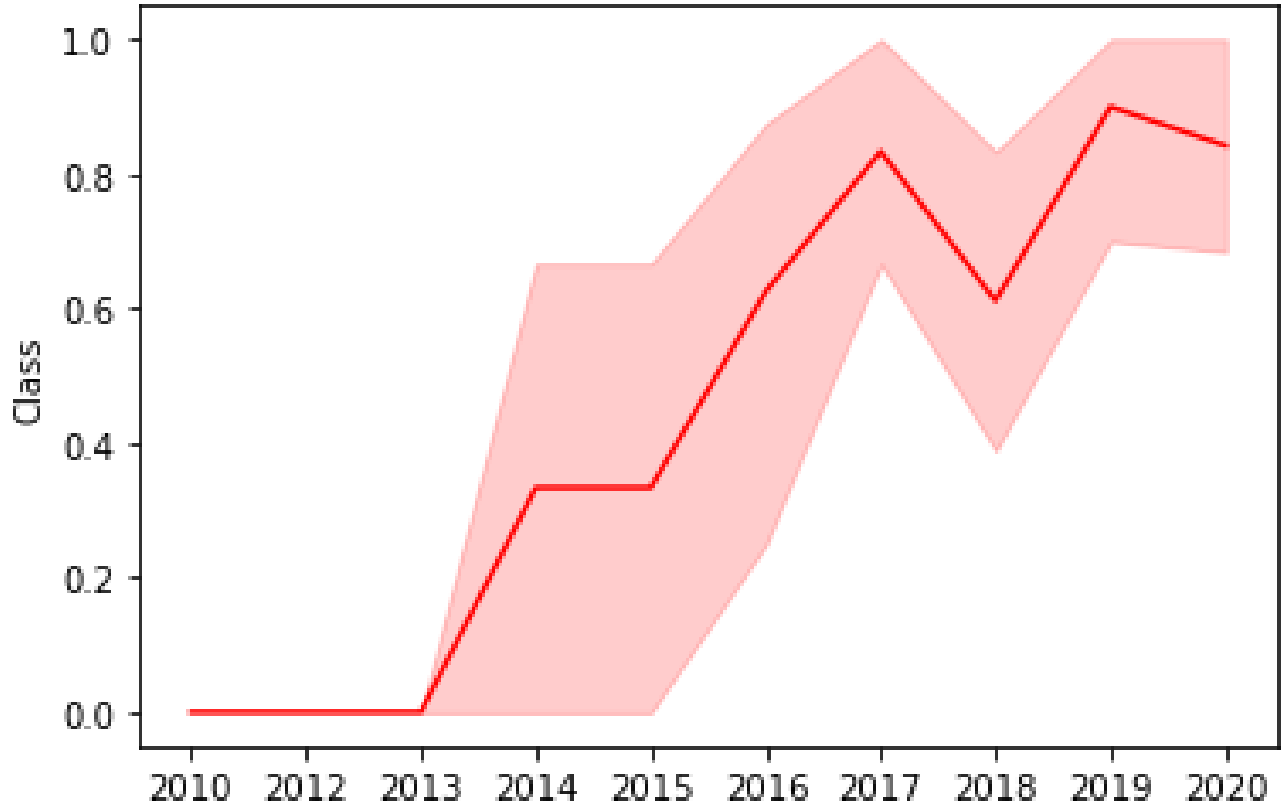
Payload vs. Orbit Type

- After analyzing this scatter-plot we can say that with heavy payloads the successful landing or positive landing rate are more for Polar , LEO and ISS.
- However, 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

- We can say from this trend chart that after 2013 the trend of success ratio is keep increasing.



All Launch Site Names

- Query:
  - **%sql** select DISTINCT LAUNCH_SITE from SPACEXDATASET

- Output :

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

Launch Site Names Begin with 'CCA'

- Query :
  - **%sql** select * from SPACEXDATASET where launch_site like 'CCA%' limit 5

This output table shows the only launch site which are only begin with 'CCA'.

- Output

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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 :
  - **%sql** select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
  -
  - Output

SUM
45596

- The amount 45596 is the number which shows the total payload mass.

Average Payload Mass by F9 v1.1

- Query:
  - **%sql** select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'
- Output:

Average
2534

- Here this query shows 2534 is the average payload mass by booster version F9 v1.1.

First Successful Ground Landing Date

- Query :
  - **%sql** select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'
- Output :

DATE
2010-06-04

- On 4-6-10 first space craft was landed success fully.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- Query :
  - **%sql** select booster_version from SPACEXDATASET where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')
- Here on the table the names of the booster version which has payload mass bettwen 4000-6000 and landed successfully

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 :
  - **%sql** SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome
- Output :

mission_outcome	***COUNT***
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

  - Here Total 101 count and of 101 99 were success and 1 was Failure and another one was succeed but its payload status is unclear.

Boosters Carried Maximum Payload

- Query :
  - maxm = **%sql** select max(payload_mass__kg_) from SPACEXDATASET
  - maxv = maxm[0][0]
  - **%sql** select booster_version from SPACEXDATASET where
  - payload_mass__kg_=(select max(payload_mass__kg_) **from** SPACEXDATASET)

Here , we use subquery to gather name of booster who have carried out Maximum payload.

Outcome :

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

- Query :
- **%sql** select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
- **from** SPACEXDATASET where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
- Output :

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Here , above table shows the data of failed launch in year 2015. This table shows month of launch, landing outcome , booster version and launch site.

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

- Query :
- **%sql** select landing__outcome, count(*) as count from SPACEXDATASET
- where Date **>=** '2010-06-04' AND Date **<=** '2017-03-20'
- GROUP by landing__outcome ORDER BY count Desc
- After running above query we get output table which has two column landing outcome and count which is group by landing outcome.

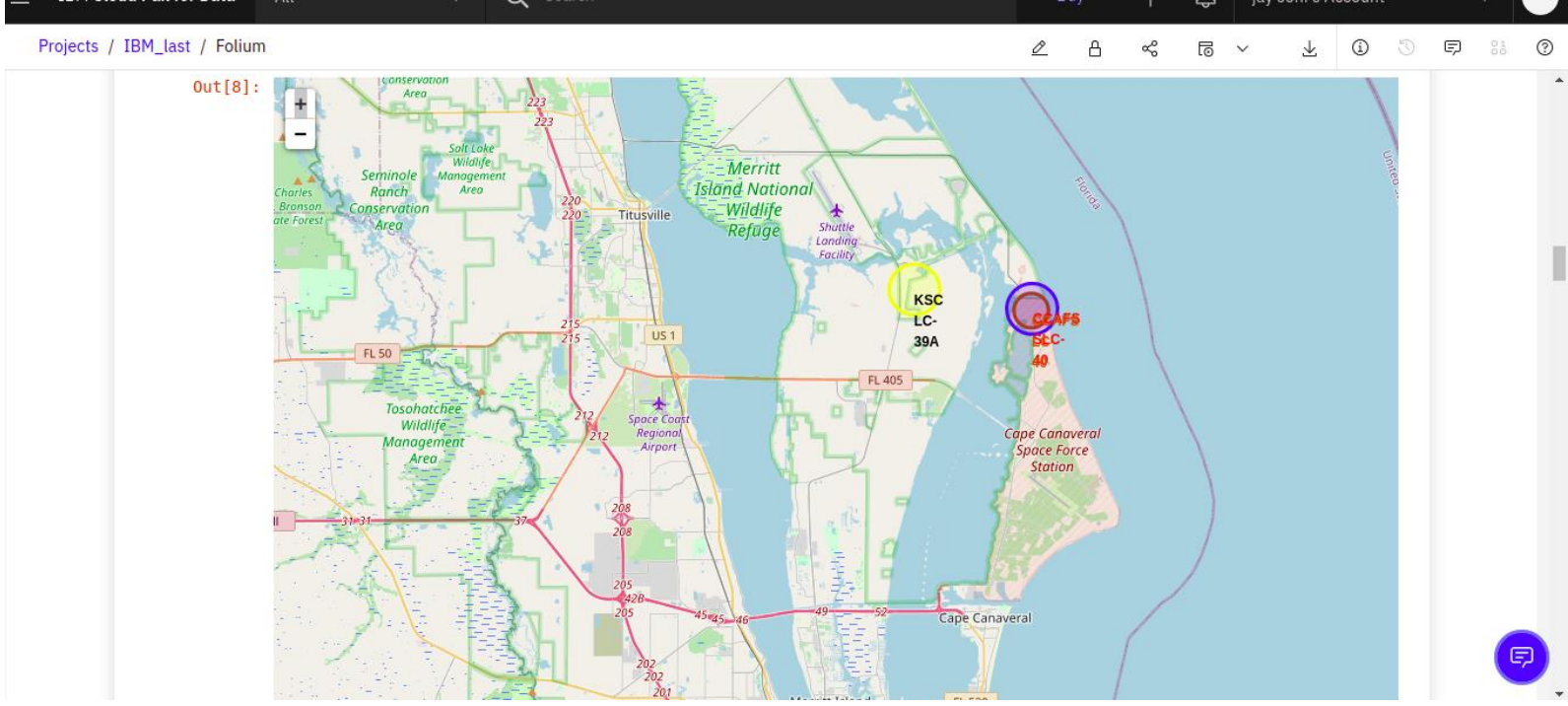
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

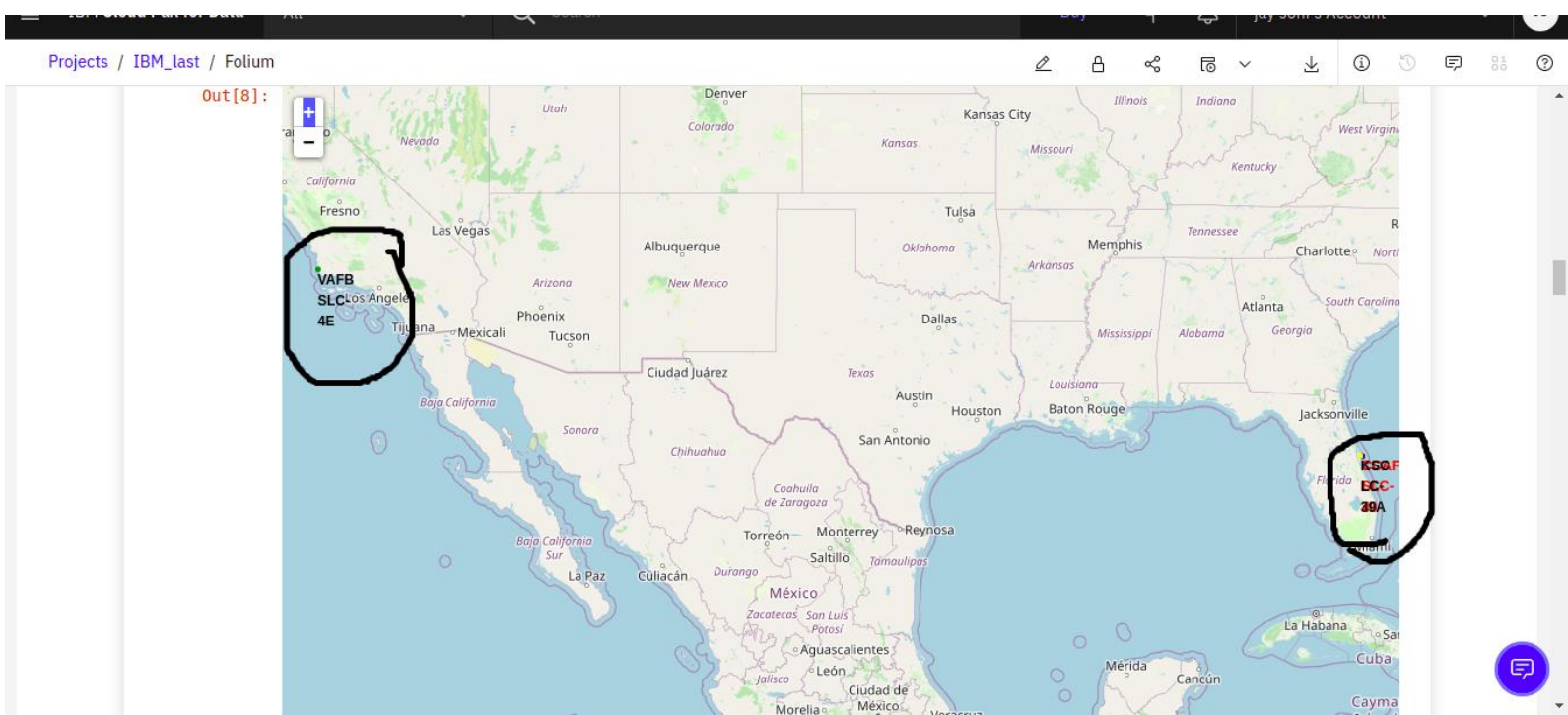
Section 4

Launch Sites Proximities Analysis

Locations of Space Launch site

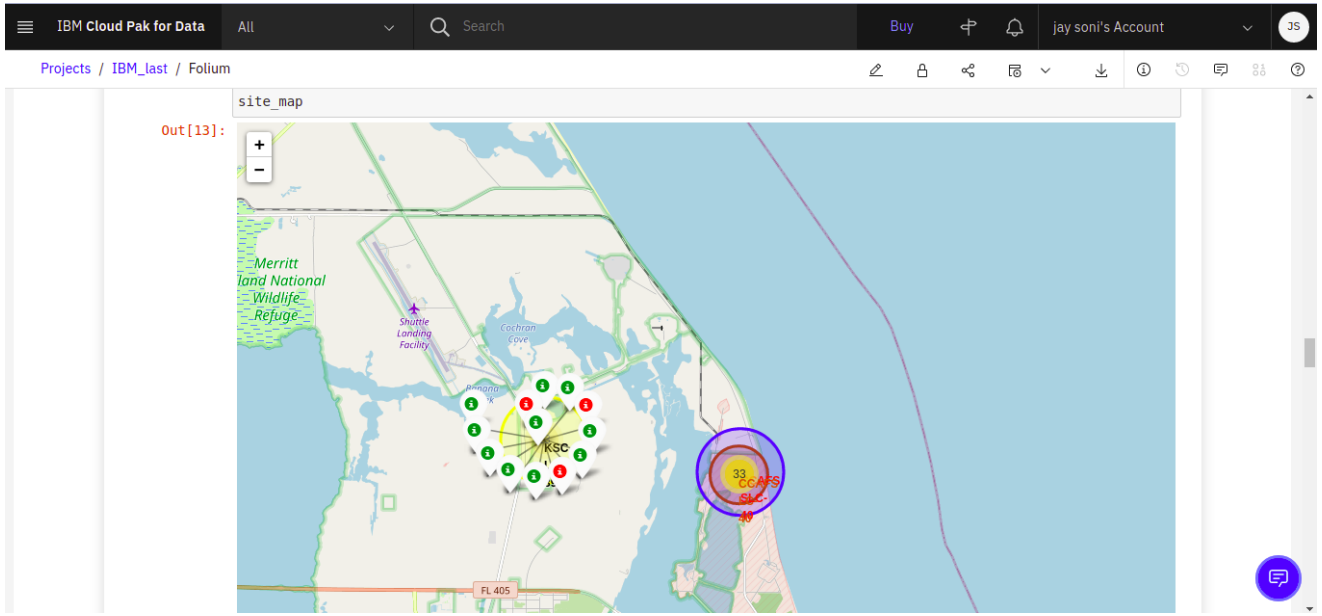
- Here in the first screenshot we can see that there are three launch site one in yellow , another in red and blue.
- In the second screenshot we can see all the launch site which are in black circle.





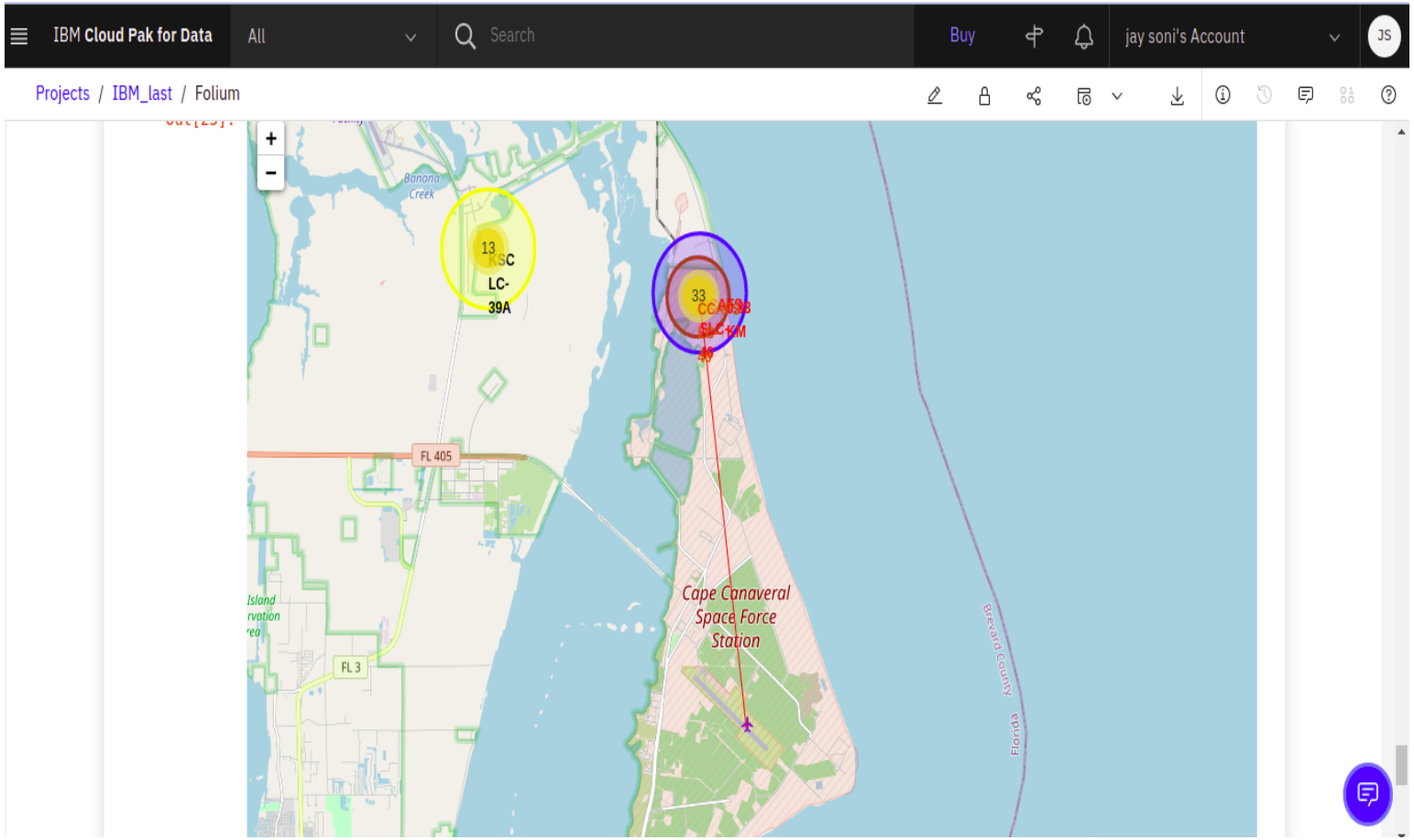
Ratio of result of rocket launch on each site

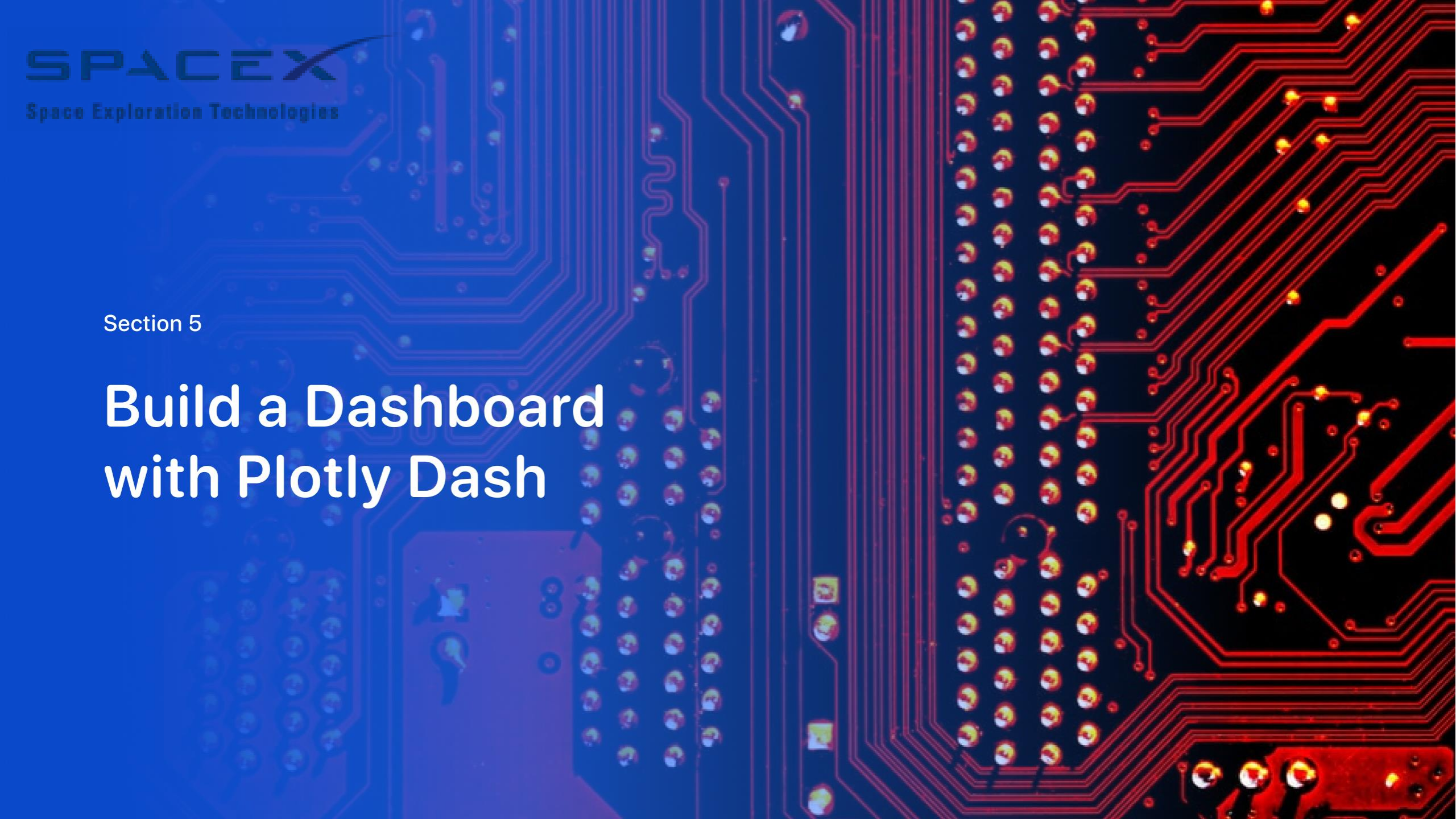
- Here in this map we can see green and red marker when we zoom in on that location which represents the success and failure.
- We can also see the total launch on that location.



Distance between Airport and launch site

- Here red line on map shows the distance between the launch site and airport



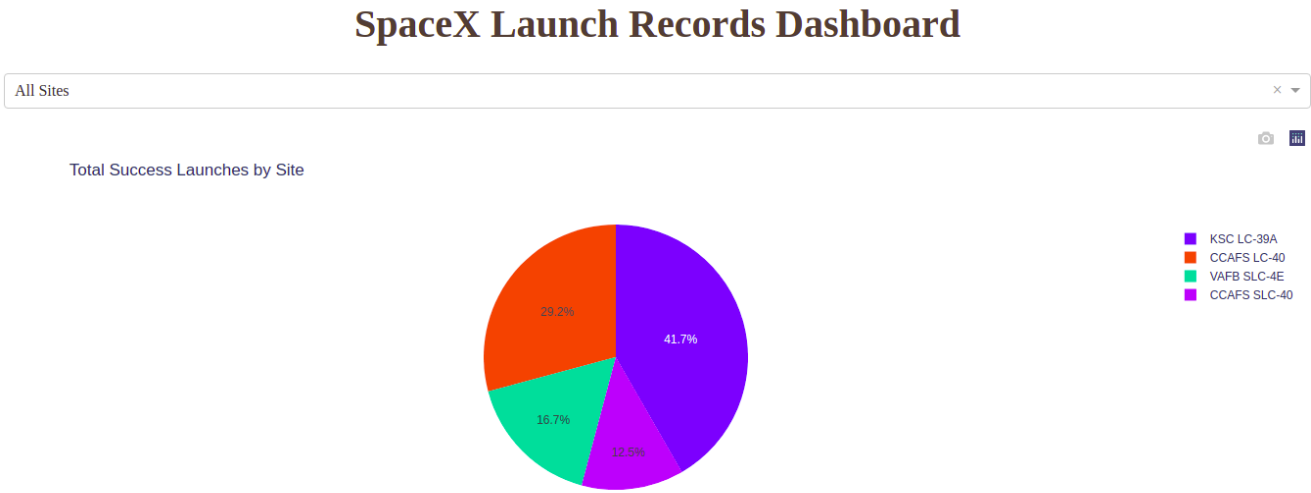


Section 5

Build a Dashboard with Plotly Dash

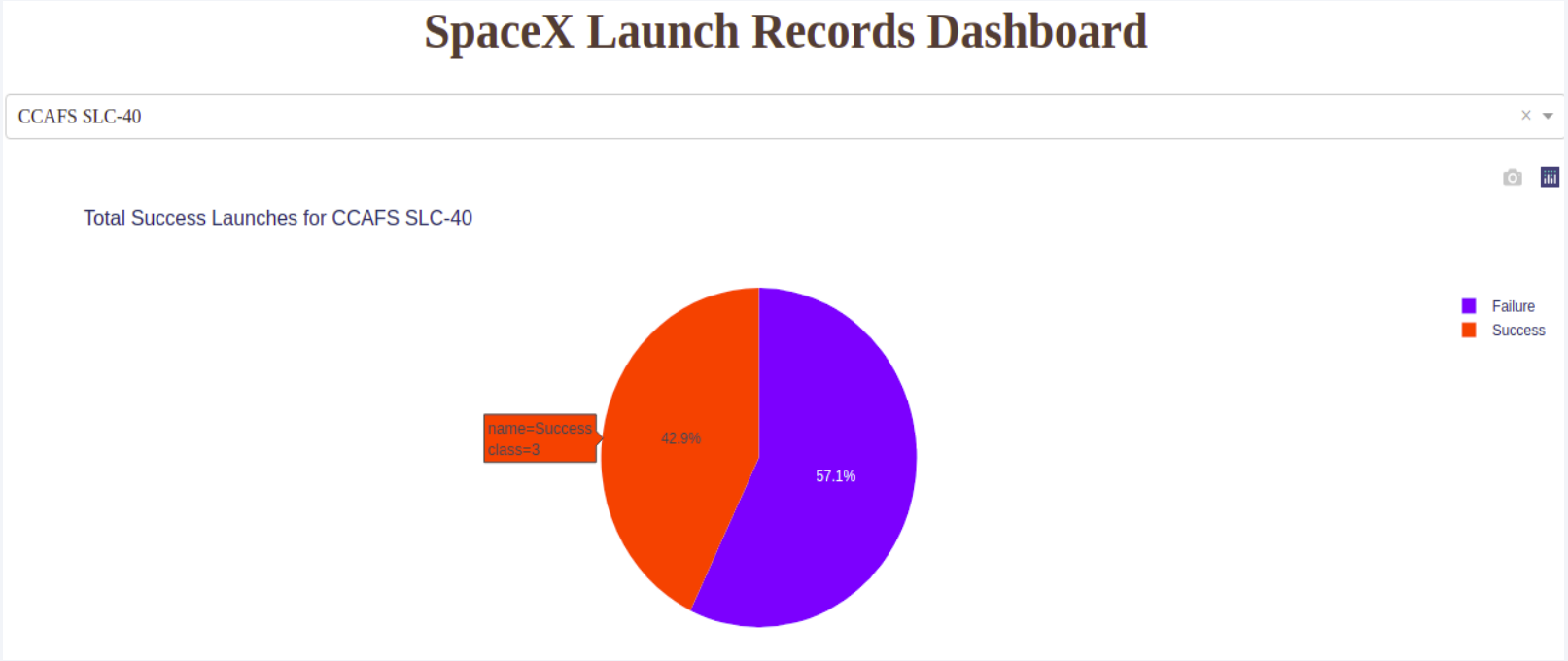
DASHBOARD -
Pie chart showing the success percentage achieved by each launch site

- In this pie chart shows the ratios of all site we can also change All site to only one site using dropdown menu.



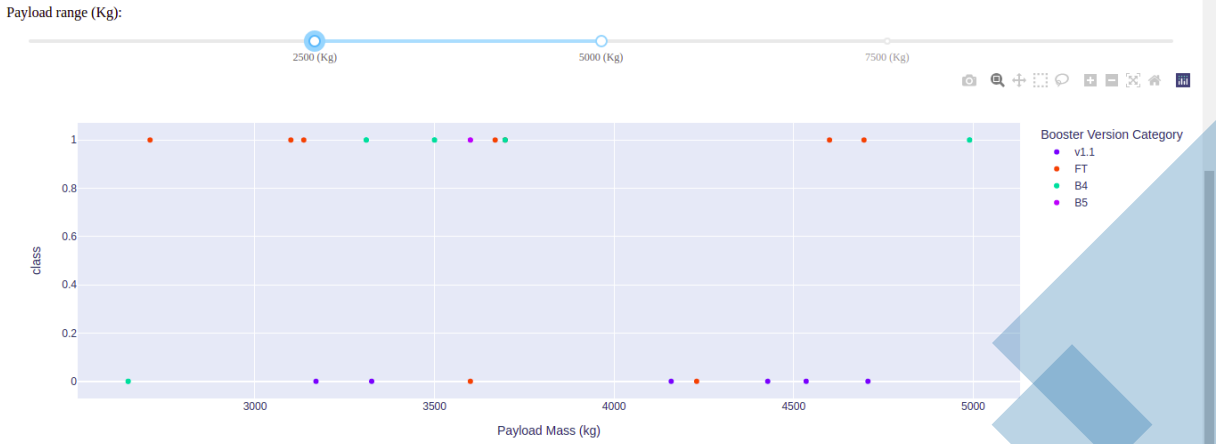
DASHBOARD – Pie chart for the launch site with highest launch success ratio

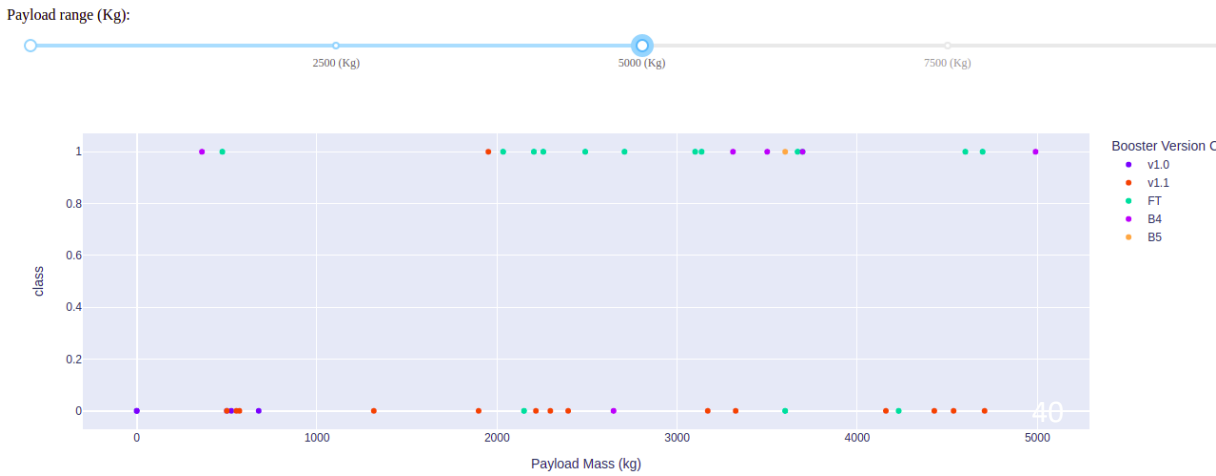
- This is the pie chart of launch site with most success rate, it has almost 43% of success rate.



DASHBOARD –
Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- The first image is of payload range between 2500 to 5000 kg and more points are on the class one which shows success ratio.
- Next image is of from 0 - 8000 kg range payload
- And in the last pic of payload range 5000 kg.
- So, from that we can say that success ratio is higher for low-weigh than heavy-weight payload.





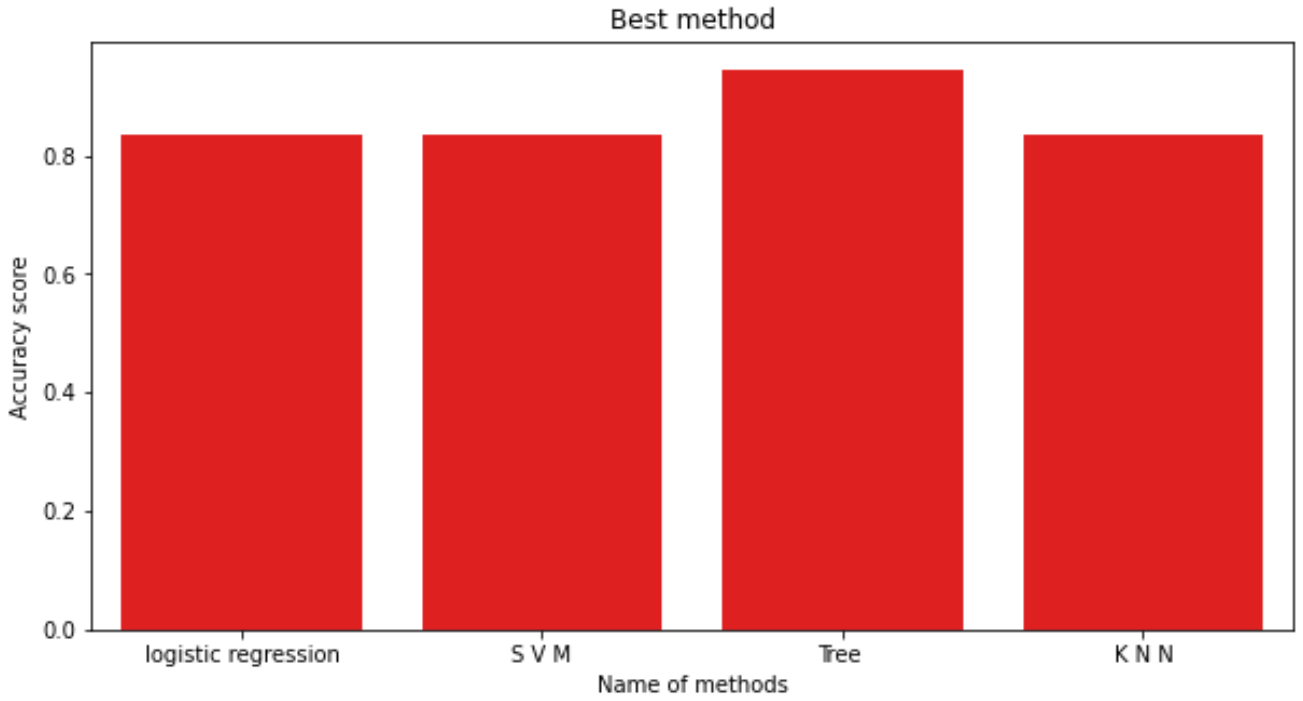
Section 6

Predictive Analysis (Classification)

Classification Accuracy

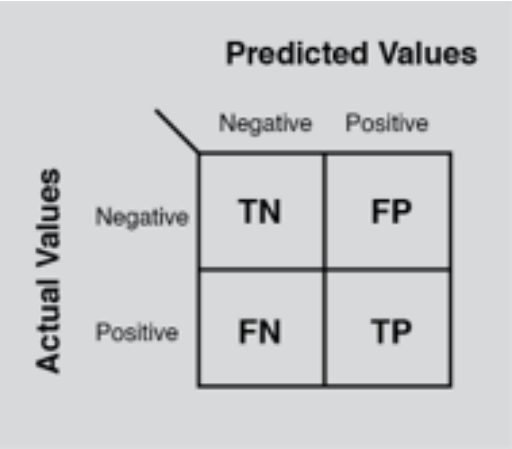
- Here bar chart compare all the accuracy of different methods.

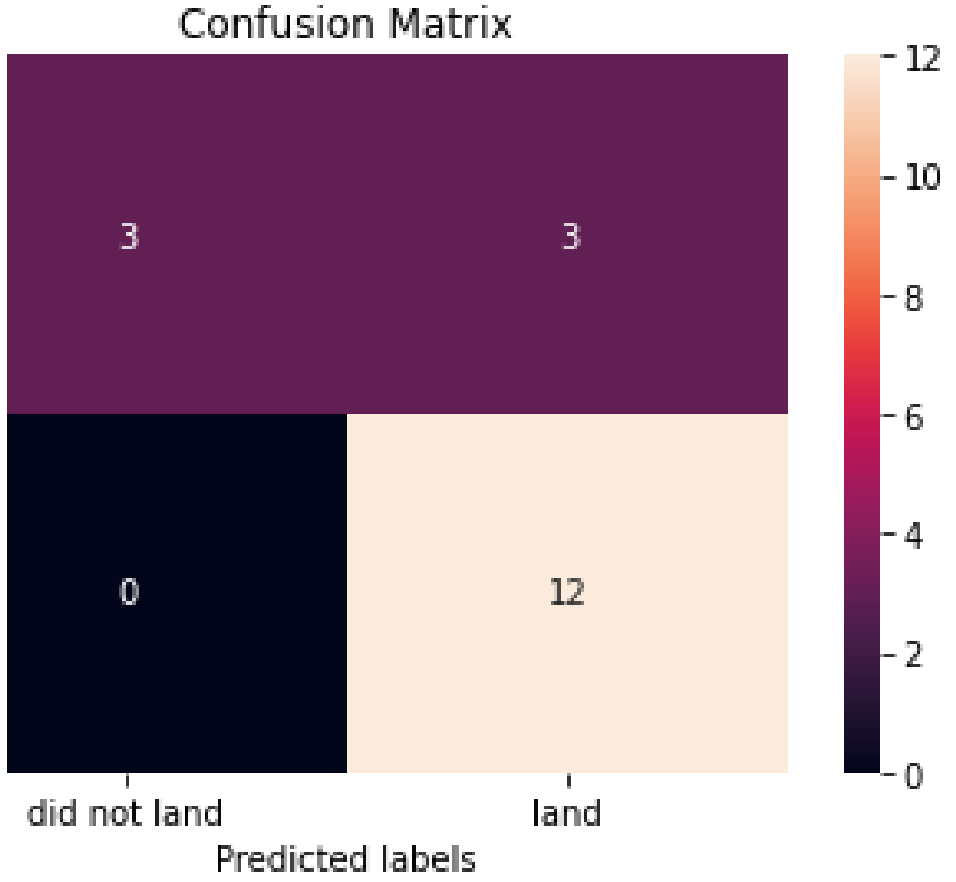
- From bar chart Decision tree classifier has the highest accuracy compared to other classifier.



Confusion Matrix

- Here in this confusion matrix represents the success ratio of Decision tree algorithm where out of 18 predictions 15 were predict truly and the other 3 were not accurate.
- Bellow in the given pic which gives idea about confusion matrix.





Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Thank you!