IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shiv Prakash
15/01/2022

Outline

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

Executive Summary

- The goal of this project is to estimate if falcon 9 stage 1 would land successfully or not, as it plays a major role in predicting the price of its relaunch. The data for this project was sourced from space x REST API, and Wikipedia. After performing some data wrangling In order to determine the best predictors for our outcome, EDA and feature scaling were done with the help of visualization using scatter and line plots. Later some ML models were created to predict future outcomes.
- The results showed that the outcome was dependent on the orbit, mass of payload, launch site, and various other technical factors such as gridfins, cores etc. The results show that there is huge progress in space x regarding this space race.

Introduction

- The evolution of technologies has changed the lives of people a lot, and with the current technologies, we are on the verge of building commercial space flights. Which can make humans multi-planetary species. There are major companies in this space race, namely blue origin, virgin galactic, and space x. The current leader in this race seems to be space x, and the reason behind that is the reusability of their stage 1. Which reduces the cost of launch from a minimum of 165 million to around 62 million per launch.
- The problem that we are trying to answer is that can the stage 1 launch of falcon 9 of spacex will successful or not , so that we can use this data for companies that want to compete with space x. Predicting that whether stage 1 will land successfully or not, plays a crucial role in predicting the launch price. As that stage can be reused again with different payloads, thus reducing the cost by more than half of original.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected from space x using REST API and from Wikipedia, using web scrapping frameworks such as BeautifulSoup.
- Performed data wrangling
  - The null values were handled at the time of performing web scrapping, one hot encoding was done on categorical variables such as orbit, launch site, landing pad and serial.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Various models like SVM, logistic regression, tree classifier and k nearest neighbours were used.

Data Collection

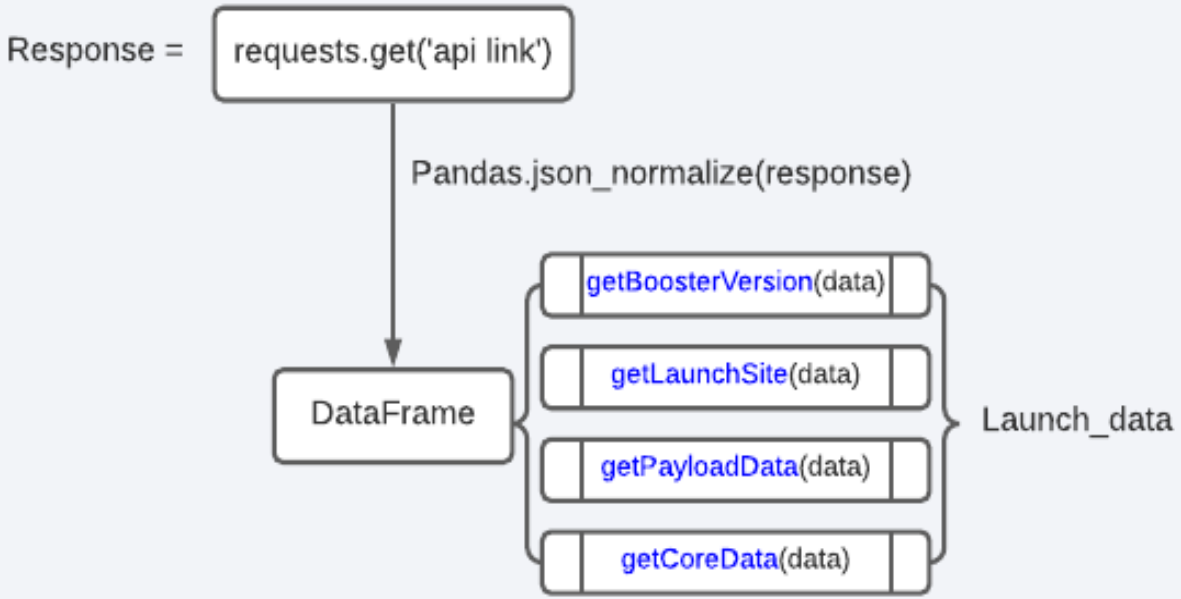
The data was first collected from the space x rest API with endpoint: Click me most of the data of certain attributes was encoded like its name so it was decoded with the help of rest api connection to the specific attributes detail's endpoint like for example to get the rocket's name a connection to endpoint click me was used. The data was also sourced from the following webpage using web scraping, with the help of BeautifulSoup framework.

Wikipedia: Falcon 9 and falcon heavy list

Data Collection – SpaceX API

- A request object was created using the space x api's end point The response object was converted to a dataframe using pandas.json_normalize(response.json())

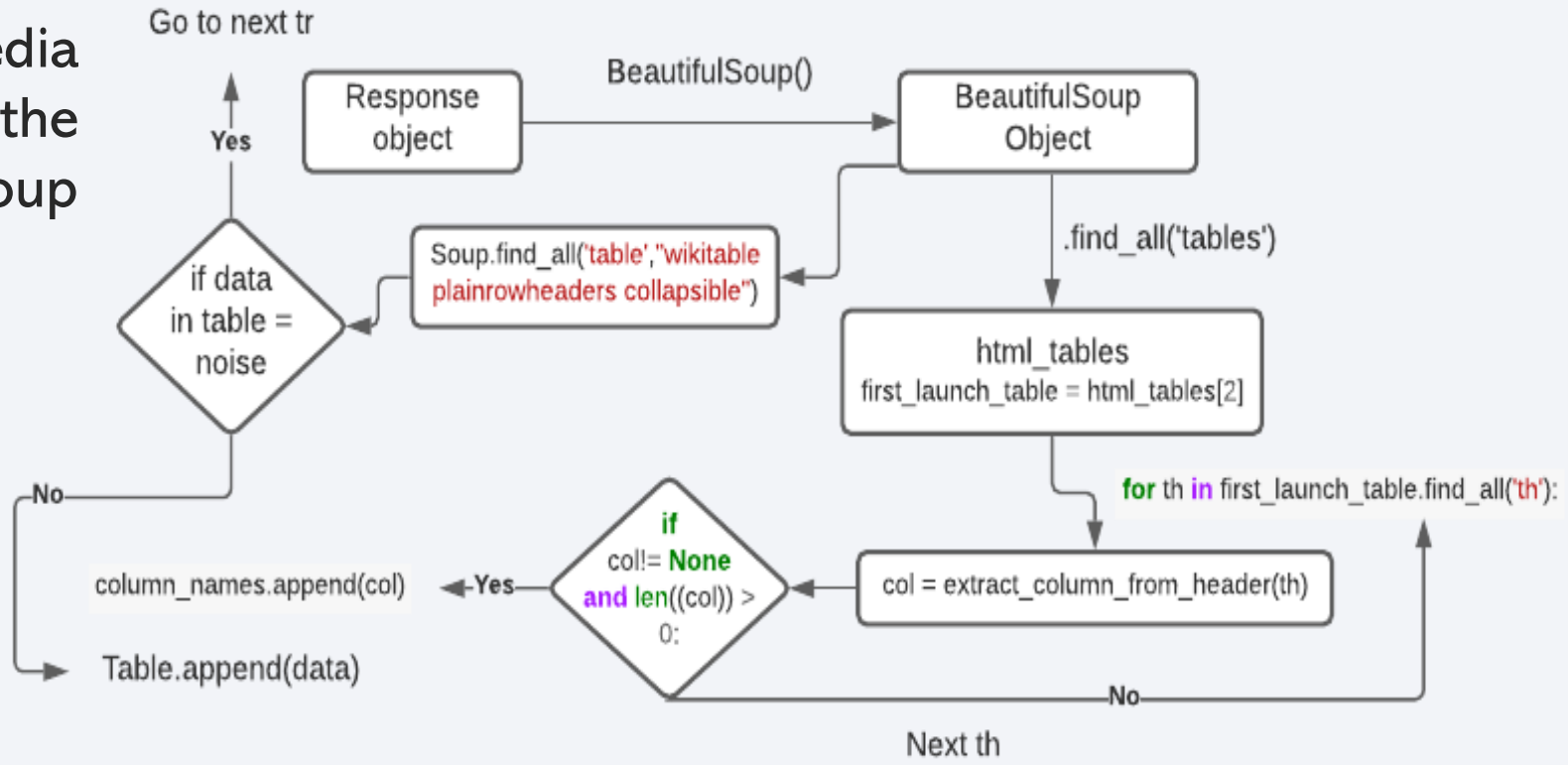
- Github url: Click me



Data Collection - Scraping

- Some of the essential data was collected from Wikipedia using web scrapping with the help of beautiful soup framework.

- Github link: [Click me]



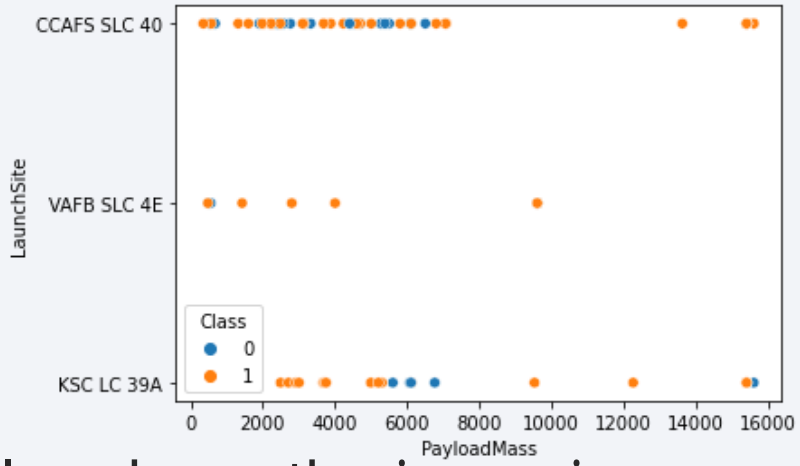
Data Wrangling

- Initially, the data were filtered to contain only the records of falcon 9, later it was found that there were some null values in the data as shown in the figure: Landing pad would remain null as It represents that the landing pad wasn't used. Null values in the column PayloadMass will be replaced by the mean value.
- One hot encoding was done for some categorical varibles in order to feed them to the ML algorithm after performring feature scaling.
- The class variable had the values {'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'} These values were converted to 0 and {'True ASDS', 'True Ocean', 'True RTLS'} to 1 representing successful landing of stage 1. Additionally, success rate was found out to be: 66%
- Guthub link: Click me

```
FlightNumber      0
Date              0
BoosterVersion    0
PayloadMass       5
Orbit             0
LaunchSite        0
Outcome           0
Flights           0
GridFins          0
Reused            0
Legs              0
LandingPad       26
Block             0
ReusedCount       0
Serial            0
Longitude         0
Latitude          0
dtype: int64
```

EDA with Data Visualization

- Bar graphs and Scatter plots for launch sites w.r.t number of launches, payload and number of launches w.r.t payload mass was created with hue set to class variables(it represents colors according to the values of class which in our case was 0 for blue and 1 for yellow) as example:



- The visualisation also shows the increasing progress of space x with each year, Github link: Click me

EDA with SQL

The following SQL quries were executed.

- %sql select DISTINCT Launch_site from SPACEXTBL
- %sql select * from SPACEXTBL where Launch_site LIKE 'CCA%'
- %sql select SUM(PAYLOAD_MASS__KG_) as TOTAL_MASS from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
- %sql select AVG(payload_mass__kg_) as Average_payload from SPACEXTBL where booster_version = 'F9 v1.1'
- %sql select min(DATE) as first_succesful_landing from SPACEXTBL where mission_outcome = 'Success'
- %sql select booster_version from SPACEXTBL where Landing__outcome = 'Success (ground pad)' AND payload_mass__kg_ >4000 AND payload_mass__kg_ <6000
- %sql select booster_version, Launch_site, Landing__Outcome from SPACEXTBL where Landing__outcome ='Failure (drone ship)' AND Year(DATE)=2015

Github Link: click me

Interactive Map with Folium

- An interactive map was created for visualizing the various factors, markers, and highlighted circles with popups were added for different launch sites to easily spot them on the map. Cluster object of markers was created to show lunch outcomes, with red being 0 I.e unsuccessful, and 1 being green I.e., successful landing. At the end a polyline object was added to the map to show the distance of the launch site from its proximites such as nearest railway station.
- GitHub URL: Click me

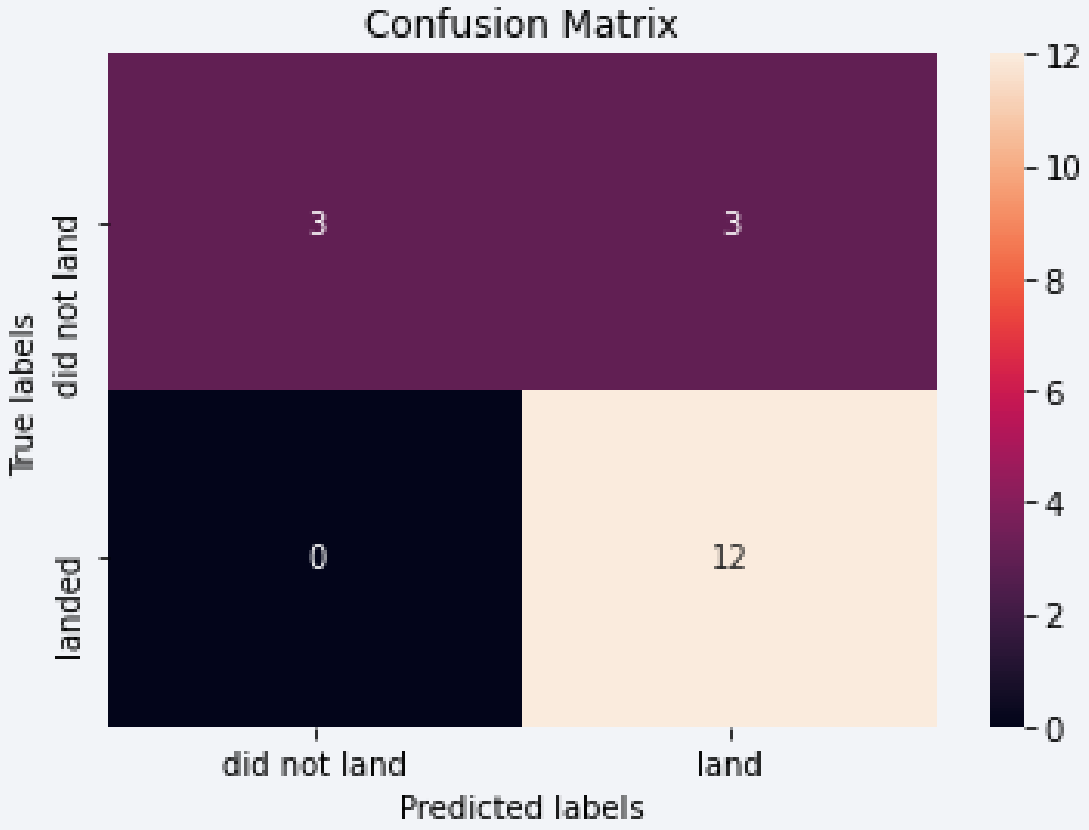
Build a Dashboard with Plotly Dash

- An interactive web application was created using dash, with a drop-down menu to select the launch site, and range-slider to choose the range of payload. Interactive pie chart showing success rate of all launch sites by default, and a scatter plot showing launch outcomes of all sites according to their payloads in the default range(0-10000) were added.
- Dropdown menu would allow the user to choose the launch site that would alter the figure of pie chart and scatter plot to show outcomes of that launch site, and through the range-slider user can select the range of payload on the x-axis of scatter plot. These interactions would allowthe user to visualise the data more in depth ccording to his needs.
- GitHub URL: Click me

Predictive Analysis (Classification)

- Many various types of ML model was used with in addition with GridSearchSV. These ML algorithms were logistic regression, SVM, decision tree and KNN. The model with best out of sample accuracy were KNN, logistic regression and SVM. The R2 score was around .83
- GitHub URL: Click me

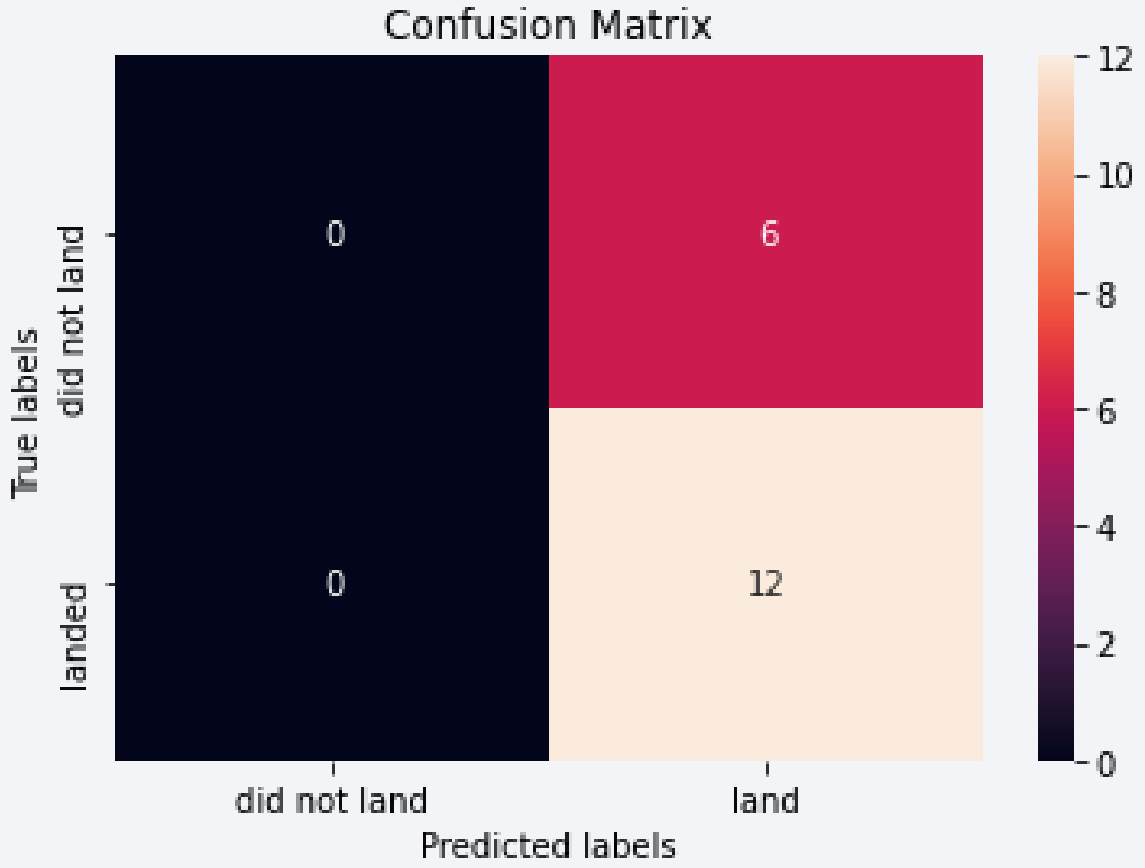
Results



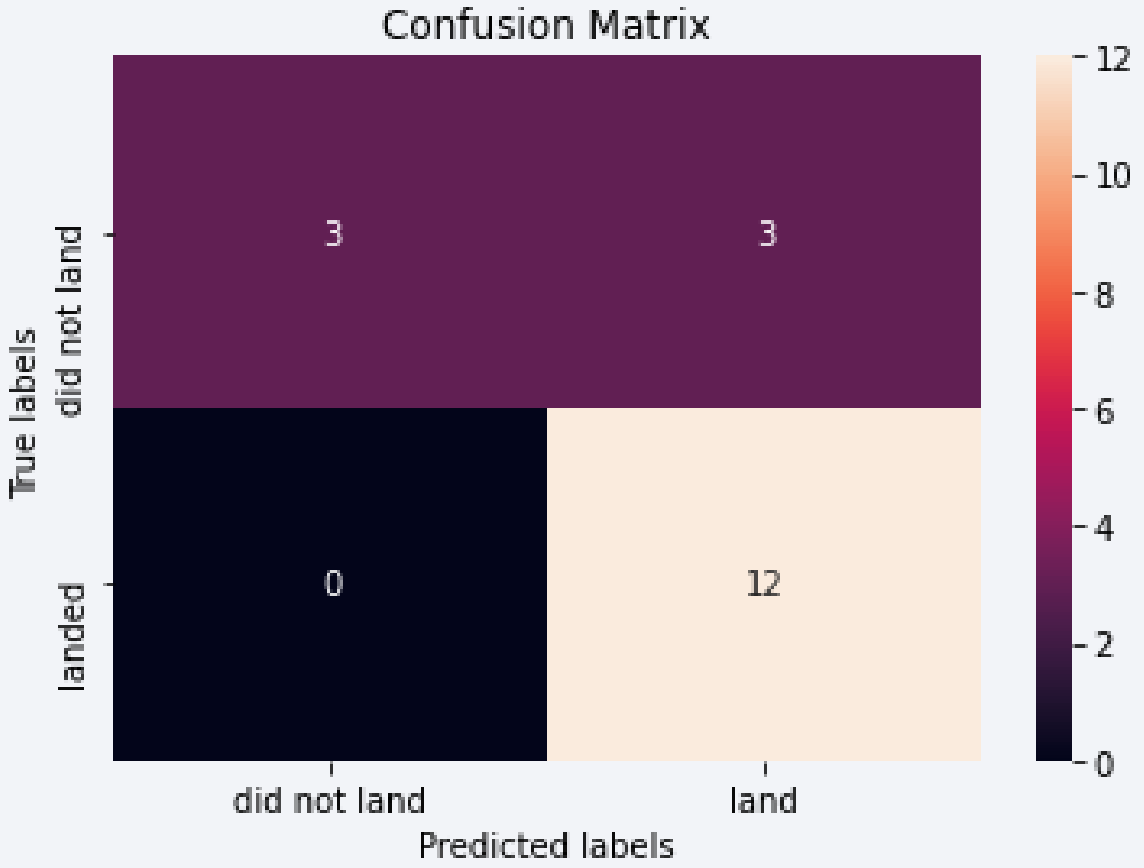
Logistic regression

SVM

Results



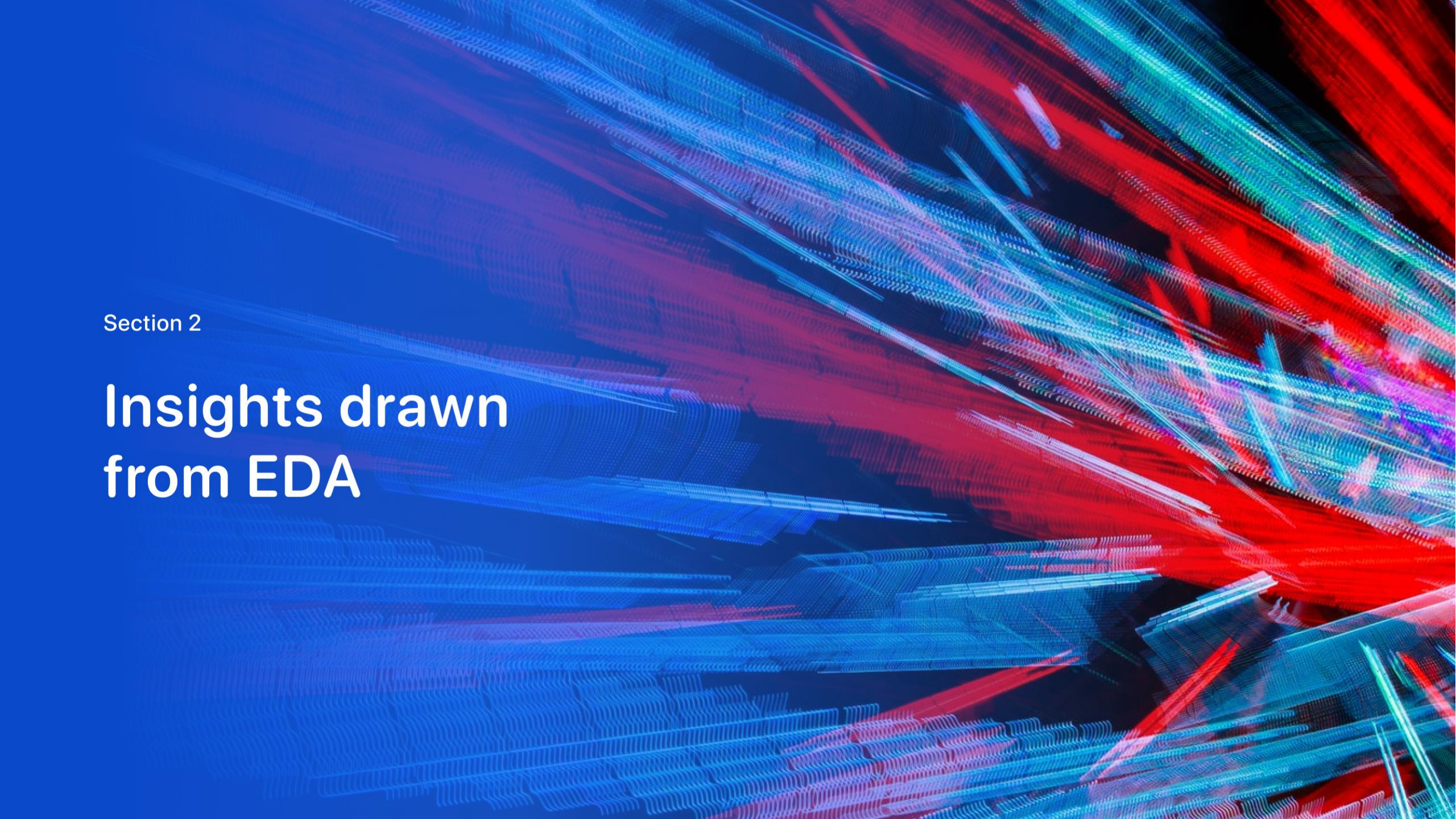
Decision Tree



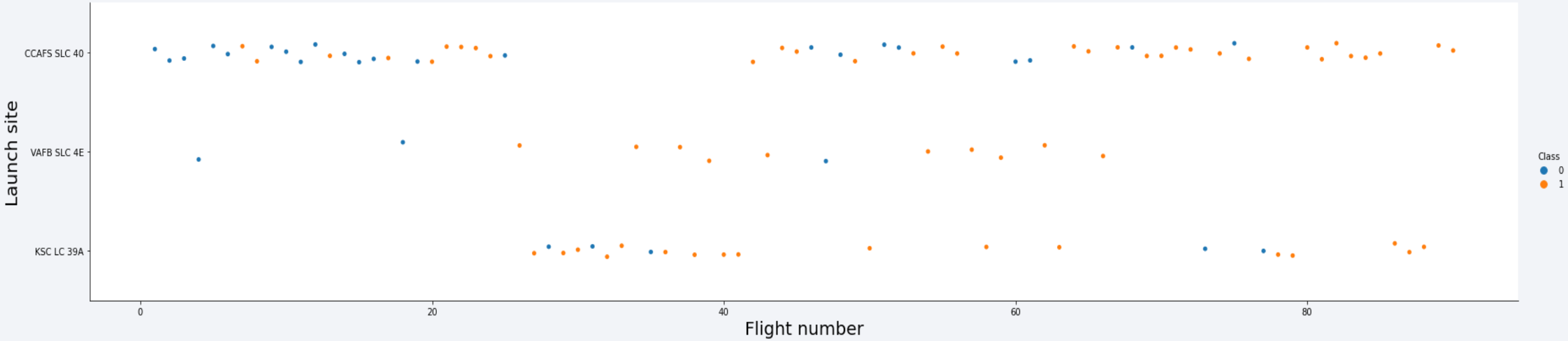
K Nearest Neighbors

Section 2

Insights drawn from EDA

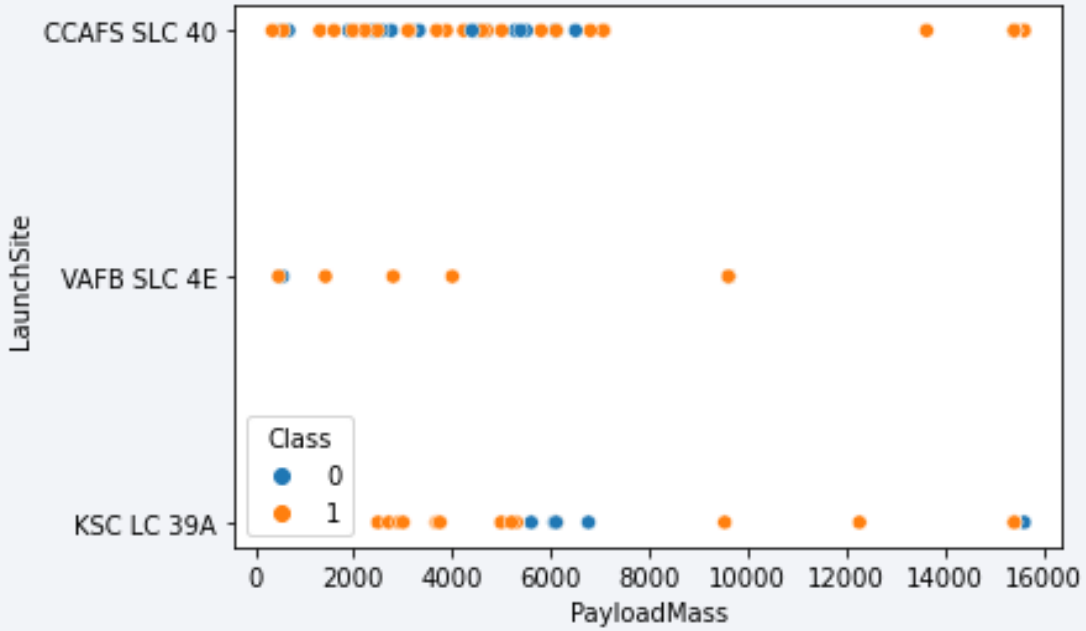


Flight Number vs. Launch Site



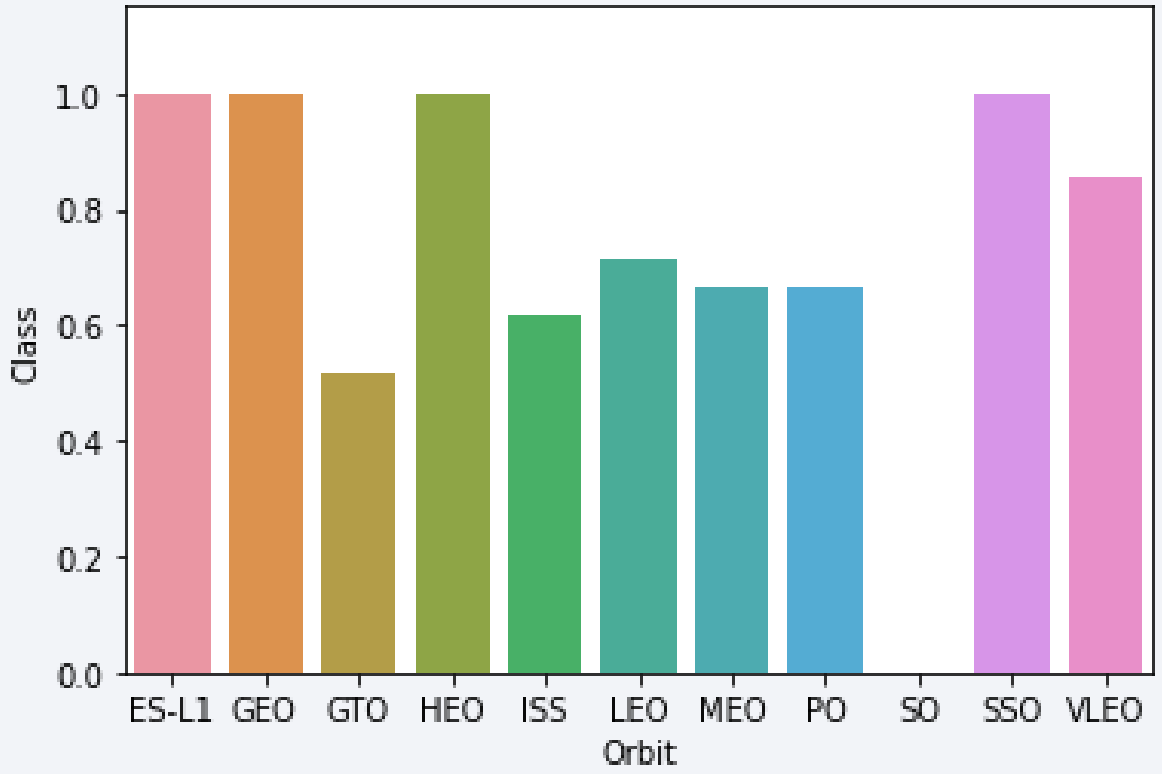
We can see above that CCAFS SLC 40 have a higher chance of success with increasing number of launches. Same goes for the KSC LC 39A.

Payload vs. Launch Site



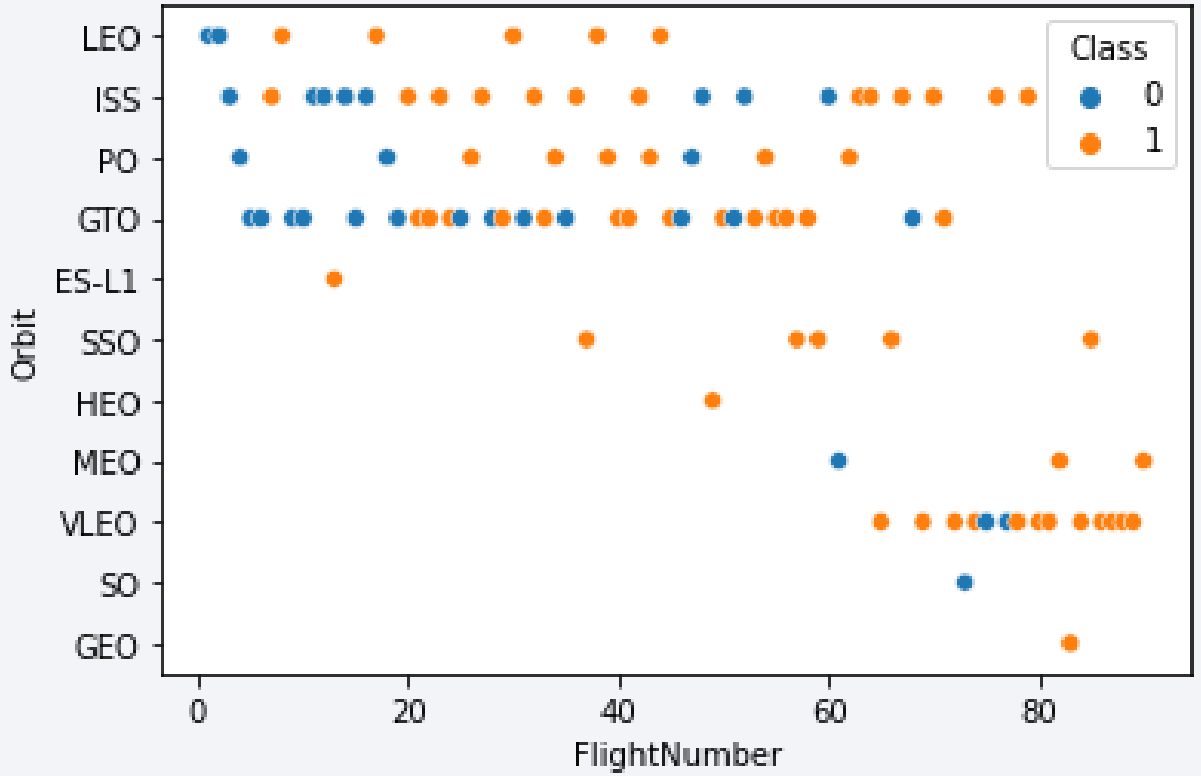
The launch site CCAFS SLC 40 had a high chances of success, when the payload mass was lower. We can also see that VAFB SLC 4E was not used much in comparison to other launch sites.

Success Rate vs. Orbit Type



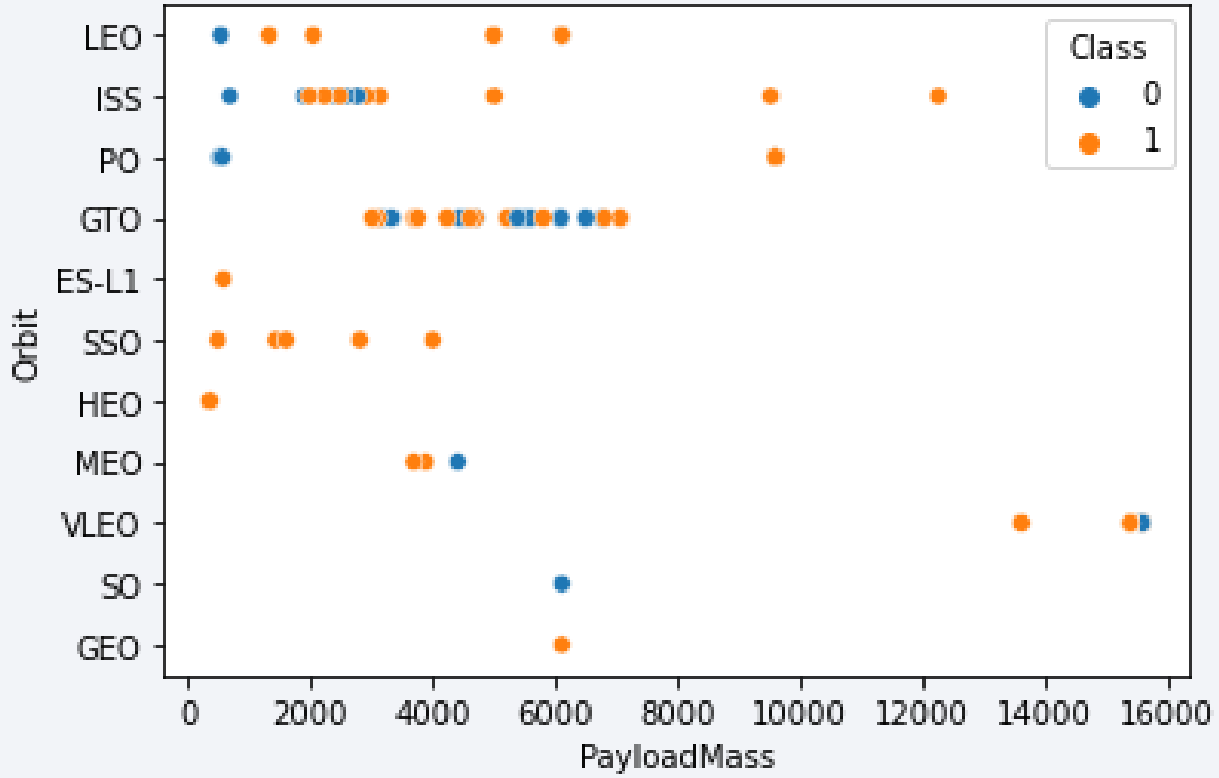
Here we can notice that launch w.r.t SSO, HEO, ES-L1 and GEO orbits have a higher chances of success

Flight Number vs. Orbit Type



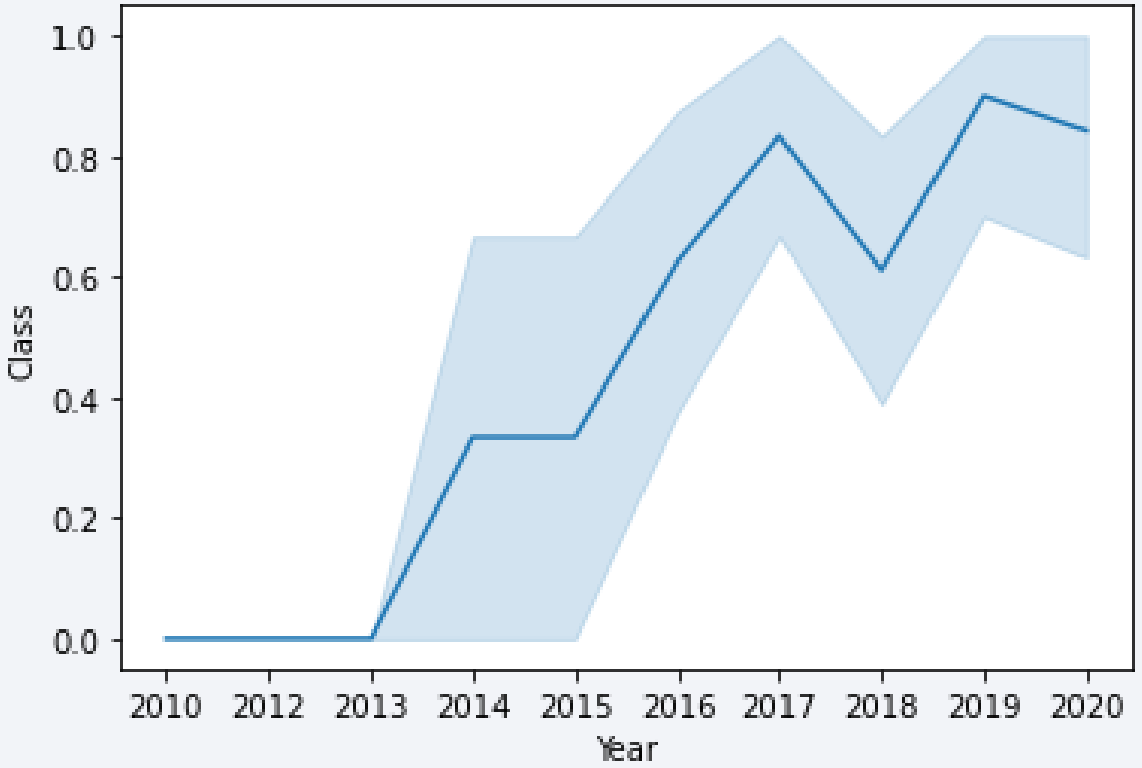
in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

Payload vs. Orbit Type



We can observe that Heavy payloads have a negative influence on GTO orbits and positive on ISS and LEO orbits.

Launch Success Yearly Trend



We can observe that the sucess rate since 2013 kept increasing till 2020. It shows how much progress had space x done in recent times

All Launch Site Names

- The names of the all launch sites are:

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

- We used pandas.groupby() funtion passing 'Launch Sites' as the parameter, and then we printed the first element in those group using .first() function on our groupby object.

```
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
```

Launch Site Names Begin with 'CCA'

- In order to find the records that begin with 'CCA' first we need to create a mapping list, which can be passed to dataframe spacex_df to get the records having launch site that begin with 'CCA' we can do that using:

```
flag = spacex_df['Launch Site'].str.contains('CCA')
```

- After that we can pass it to the dataframe spacex to get all the records who's launch site attribute starts with 'CCA' and we use .head(5) to show only the first 5 records:

```
spacex_df[flag].head(5)
```

	Unnamed: 0	Flight Number	Launch Site	Mission Outcome	class	Payload Mass (kg)	Booster Version	Booster Version Category
0	0	1	CCAFS LC-40	Success	0	0.0	F9 v1.0 B0003	v1.0
1	1	2	CCAFS LC-40	Success	0	0.0	F9 v1.0 B0004	v1.0
2	2	3	CCAFS LC-40	Success	0	525.0	F9 v1.0 B0005	v1.0
3	3	4	CCAFS LC-40	Success	0	500.0	F9 v1.0 B0006	v1.0
4	4	5	CCAFS LC-40	Success	0	677.0	F9 v1.0 B0007	v1.0

Total Payload Mass

The total payload carried by boosters from NASA, can be calculated using groupby() function on the dataframe with passing parameter as 'Launch Site' and then performing aggregate function sum on the feature payload mass. After that we convert the result to datframe using .to_frame() pandas function. Later after performing reset_index() on the obtained dataframe we can search for the row that contains the NASA Launch site namely KSC LC-39A.

```
x = spacex_df.groupby('Launch Site')['Payload Mass (kg)'].sum().to_frame()
```

```
x.reset_index(inplace = True)
x
```

	index	Launch Site	Payload Mass (kg)
0	0	CCAFS LC-40	67363.00
1	1	CCAFS SLC-40	24616.65
2	2	KSC LC-39A	56894.65
3	3	VAFB SLC-4E	58138.00

```
x[x['Launch Site'].str.contains('KSC')]
```

	index	Launch Site	Payload Mass (kg)
2	2	KSC LC-39A	56894.65

Average Payload Mass by F9 v1.1

We can calculate the average payload mass where booster version is f9 v1.1 by using groupby function on 'BoosterVersion' key and the aggregate the value of payload using mean()

```
Boosters = spacex_df.groupby('Booster Version')['Payload Mass (kg)'].mean().to_frame()
Boosters.reset_index(inplace = True)
Boosters[Boosters['Booster Version'].str.contains('F9 v1.1')]
```

	Booster Version	Payload Mass (kg)
41	F9 v1.1	2928.4
42	F9 v1.1 B1003	500.0
43	F9 v1.1 B1010	2216.0
44	F9 v1.1 B1011	4428.0
45	F9 v1.1 B1012	2395.0
46	F9 v1.1 B1013	570.0
47	F9 v1.1 B1014	4159.0
48	F9 v1.1 B1015	1898.0
49	F9 v1.1 B1016	4707.0
50	F9 v1.1 B1017	553.0
51	F9 v1.1 B1018	1952.0

Total Number of Successful and Failure Mission Outcomes

- We can use the value_counts() function on the column spacex_df['class'] to get the total number of successful/unsuccessful outcomes represented as 0 and 1 as shown below:

```
spacex_df['class'].value_counts()
```

```
0    32
1    24
Name: class, dtype: int64
```

- The above result shows that there were 32 outcomes as failure and 24 falcon 9 stage 1, were successful at landing.

Boosters Carried Maximum Payload

- To get the names of the booster which have carried the maximum payload mass. We can compare each value in payload mass column to maximum mass and then pass the result to spacex_df data frame to get the respective records and then we can print the booster version from them.

```
max_payload = spacex_df['Payload Mass (kg)'].max()
spacex_df[spacex_df['Payload Mass (kg)'] == max_payload]['Booster Version']
```

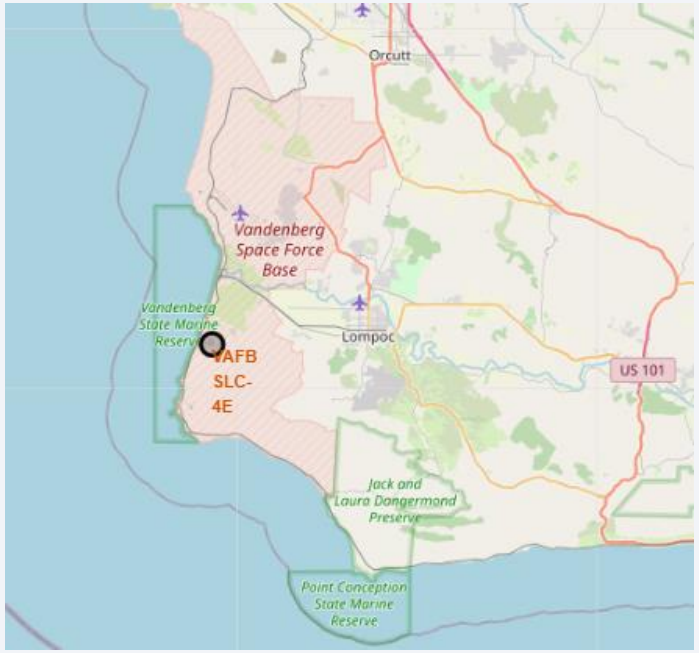
```
28     F9 FT B1029.1
29     F9 FT B1036.1
31     F9 B4 B1041.1
32    F9 FT  B1036.2
34    F9 B4  B1041.2
Name: Booster Version, dtype: object
```

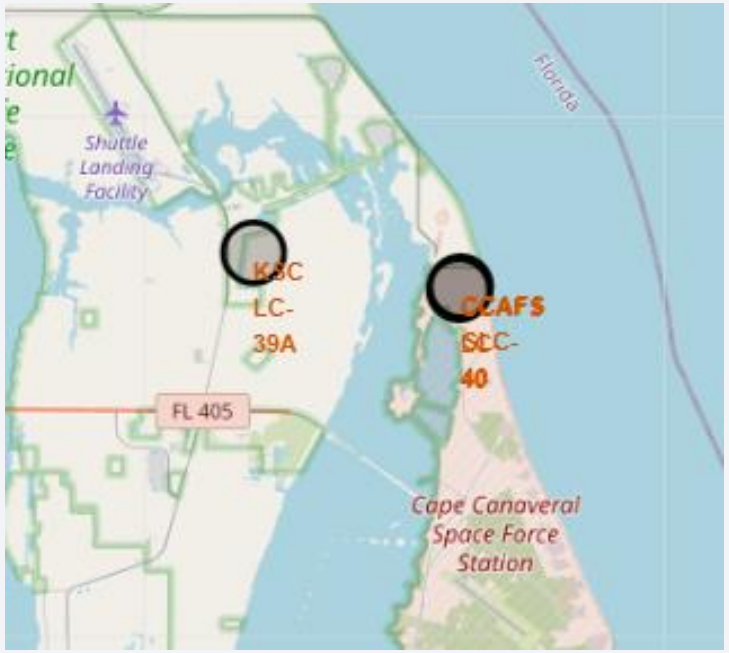
Section 4

Launch Sites Proximities Analysis

Plotting launch sites on map

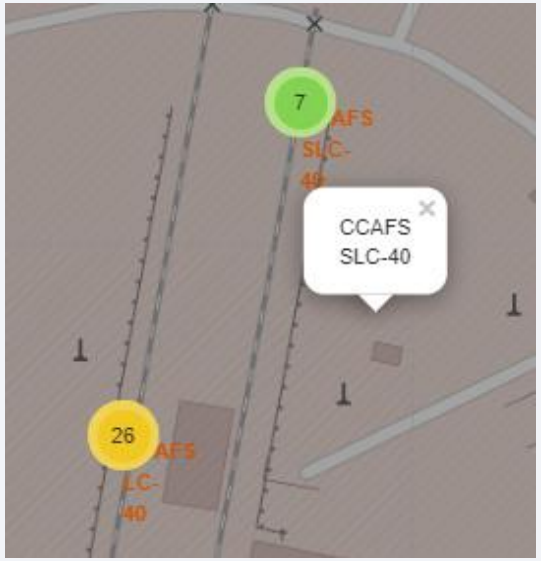
- We'll be plotting all the launch sites on the map by using the marker and circler objects of folium map, and then label them too.

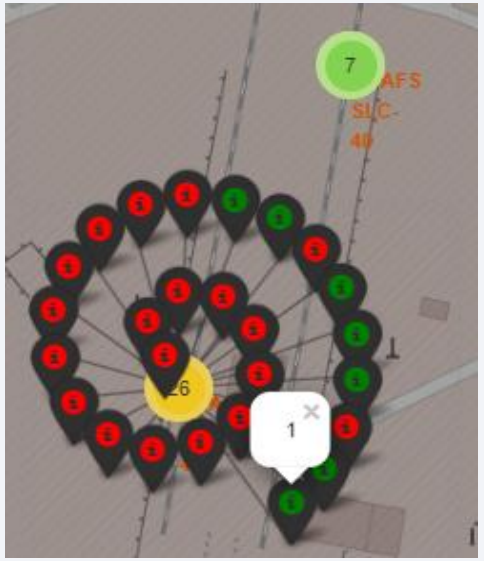




Marking outcomes

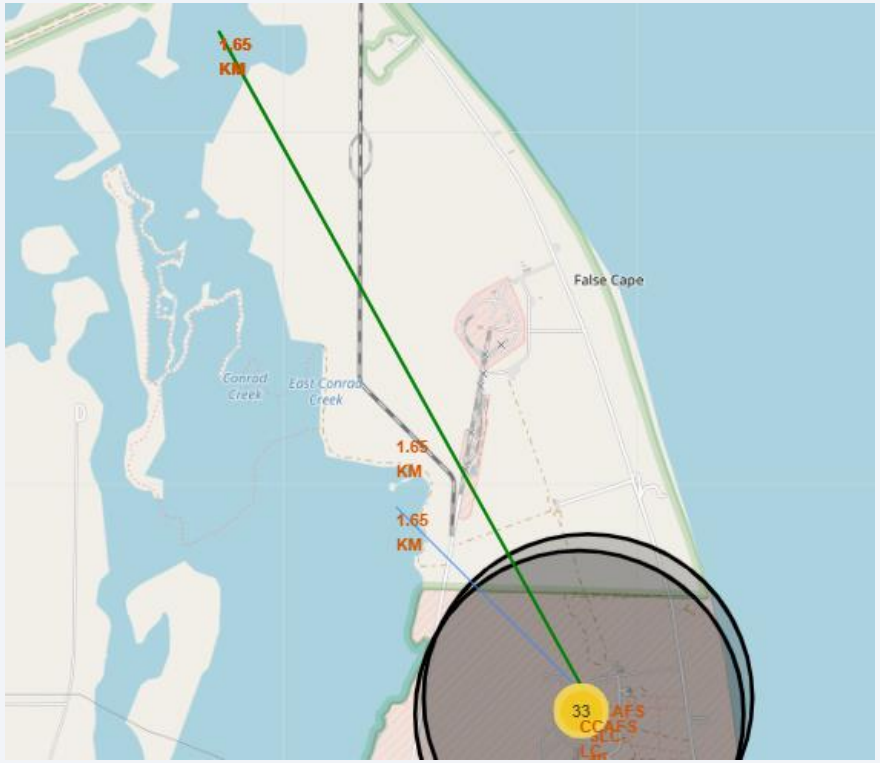
- Next, we will be adding a marker cluster representing total number of launches and their outcome with red representing failure and green success.





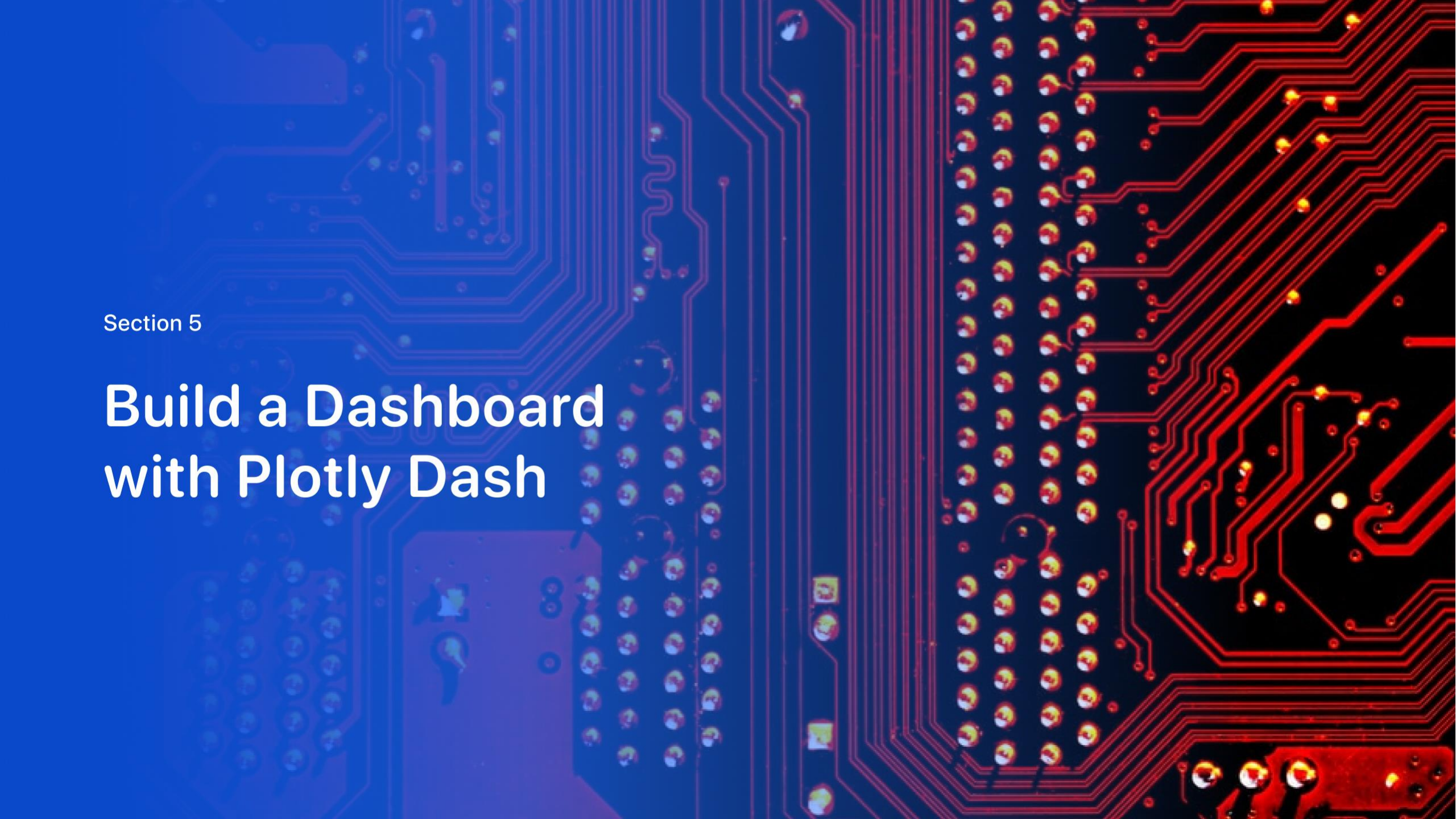
Adding distance to proximites

- In the end we'll be adding a polyline to the folium map representing the distance of launch sites from its nearby locations such as ocean and railway station etc.



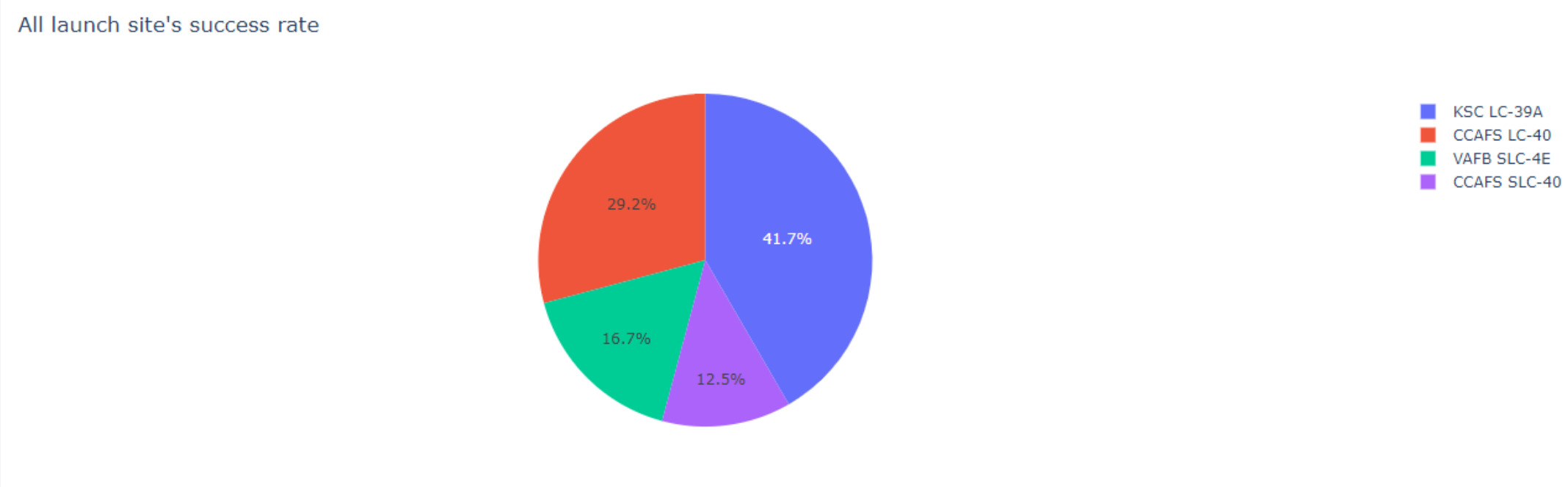
Section 5

Build a Dashboard with Plotly Dash



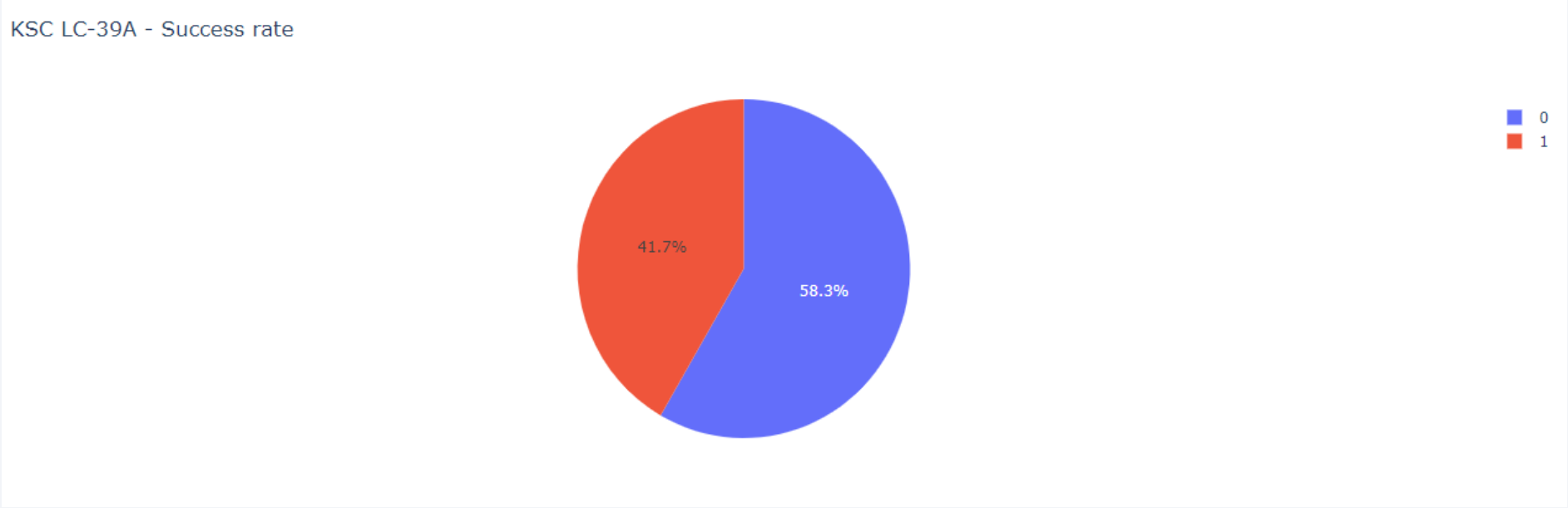
Launch success rate

- The below pie chart shows the success rate of alll the launch sites, we used plotly along with dash to create a web application to show these interactive visual graphs.



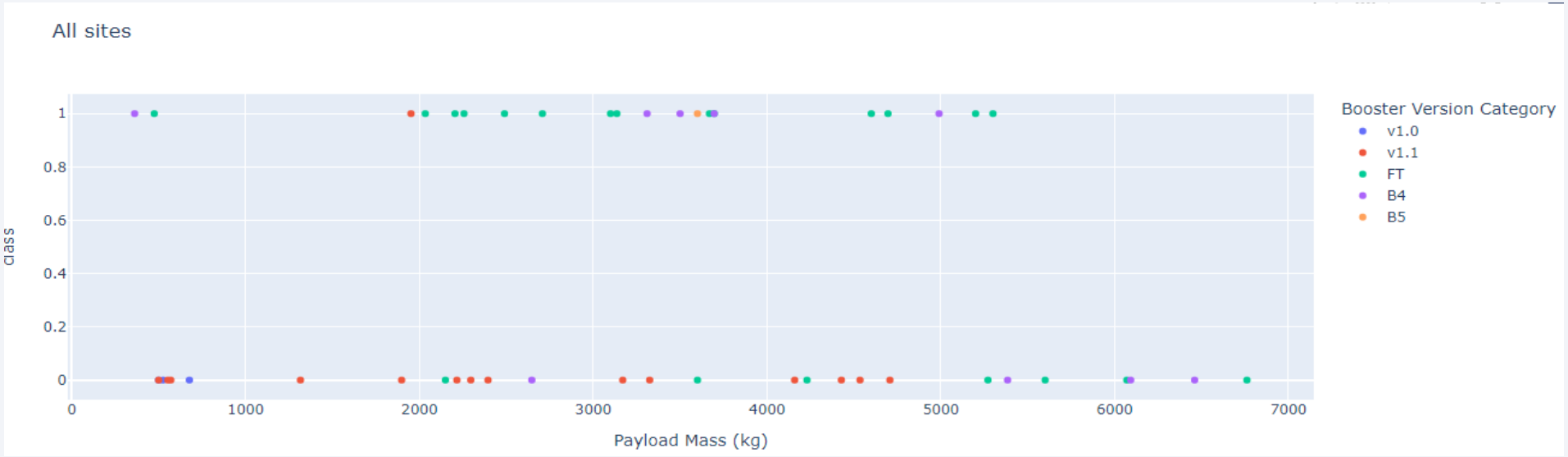
Launch site with highest success

The below pie chart shows the success rate of the launch site KSC LC-39A which had the highest success rate of around 41.7% in comparison to all the other launch sites.



Payload vs Launch outcome

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



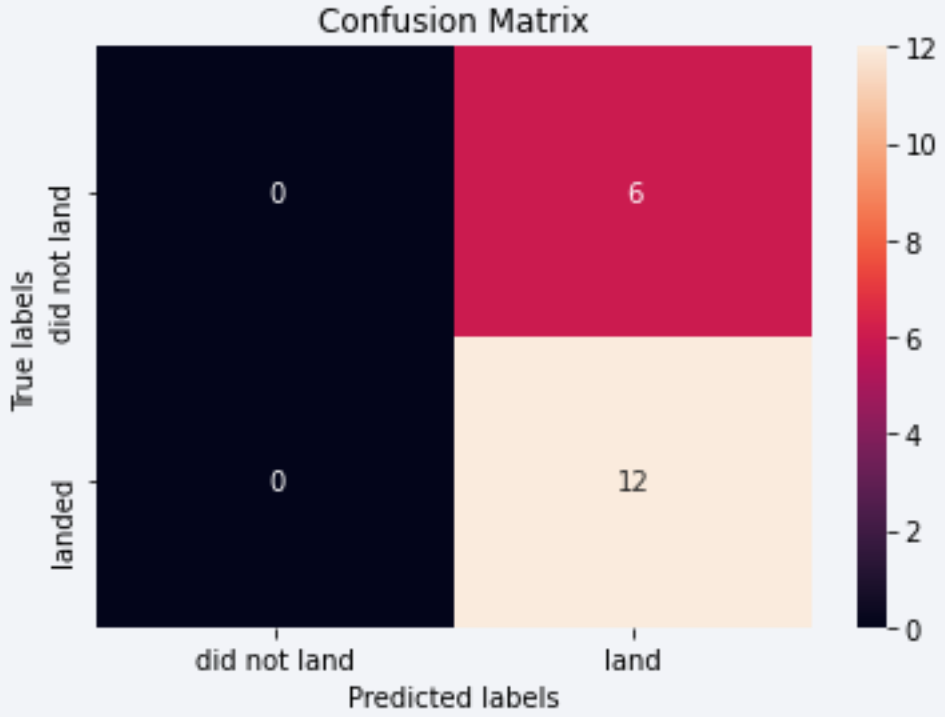
- From the above plot we can see that booster verison FT had the highest success rate in comparision to other booster versions.

Section 6

Predictive Analysis (Classification)

Confusion Matrix

- The best performing model is Decision Tree. For the outcome prediction Decision Tree would be a good model as it's a decent model for data with high dimensions, like in our case. The model achieved a accuracy of 87.5% on test data.



Conclusions

- There is a good amount of progress in launch outcomes w.r.t increasing number of launches each year.
- Some sittes such as KSC LC-39A had the highest success rate in comparision to other launch sites.
- We can clearly see that space x is leading the space race, and there are some major improvments in the rate of success of landing of falcon 9 stage one.
- The success rate was also dependent on the orbit and payload mass, we saw that ISS and VLEO orbits had a good success rate.
- Decision Tree is a suitable model to predict if the stage one would land or not, it had an accuracy of 87.5%

Thank you!