IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Here is the methodology we use to get the information we need. First, we scrap the data from API and Web. Then we do data wrangling before applying some EDA using SQL. Moreover, we also do the analysis with visualization. Lastly, we split the data to training and test data to find appropriate hyperparameters for our models (SVM, Classification Trees, and Logistic Regression).
- We achieved 83.33% accuracy for the SVM, logistic regression, KNN, and tree classifier on the test data. From the experiment, we found that the best kernel for SVM is sigmoid, the max depth for tree classifier is 8, the number of neighbors for KNN is 10, and the best regularization strength for logistic regression is 0.01 for our test data.

Introduction

- We would like to collect data from the Falcon 9 then determine if the rocket land successfully. This information will be used for another company which wants to bid against SpaceX for a rocket launch.
- We would like to present our analysis based on exploratory data analysis, visualization analysis, and predictive analysis to determine successful landing for the Falcon 9.

Section 1

Methodology

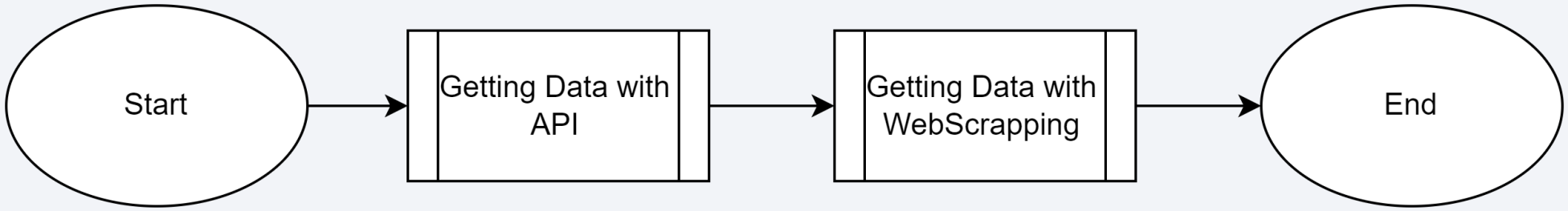
Methodology

Executive Summary

- Data collection methodology:
  - We collected the data from available APIs contain the historical data of Falcon 9 Rocket Launches. Moreover, we also get additional data from scrapping the web.
- Perform data wrangling
  - We removed the Falcon 1 data and only kept the Falcon 9 data. Then we categorize the bad outcome from landing outcomes, we found that there are 5 outcomes which associated with the bad outcome. Otherwise, we assume that it is a successful landing. Moreover, we replaced the null value with averaged value.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - We split the data into two separated datasets, which are training and test dataset. Then we find the best hyperparameter for our models. In this project, we have SVM, classification tree, and logistic regression for the predictive model.

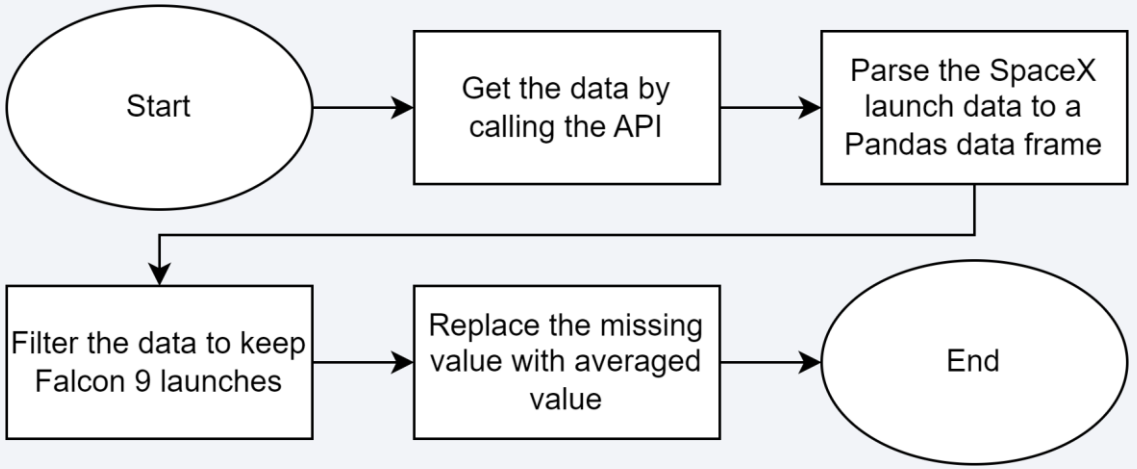
Data Collection

- The data sets are collected by two methods, getting it through API call and Web Scrapping.
- Here is the flowchart to illustrate the process. The detailed process of capturing data from API and web scrapping will be explained in the next process.



Data Collection – SpaceX API

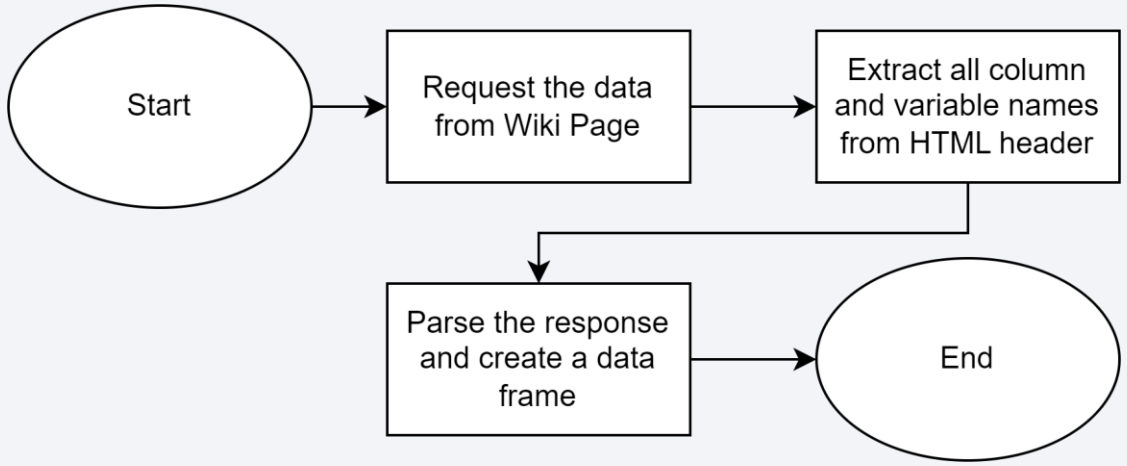
- First, we collect the data by calling API (GET). Then, the JSON format is parsed to pandas' data frame with an existing function. We filter the data so we can keep the Falcon 9 launches data. Moreover, we replace the missing value with averaged value for Payload Mass



Github URL:
https://github.com/Johanesary/TFSTraining_DS/blob/3f7347d7506ec27ea91025af05756626673373 89/Applied%20Data%20Science%20Capstone%20JRY2185/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

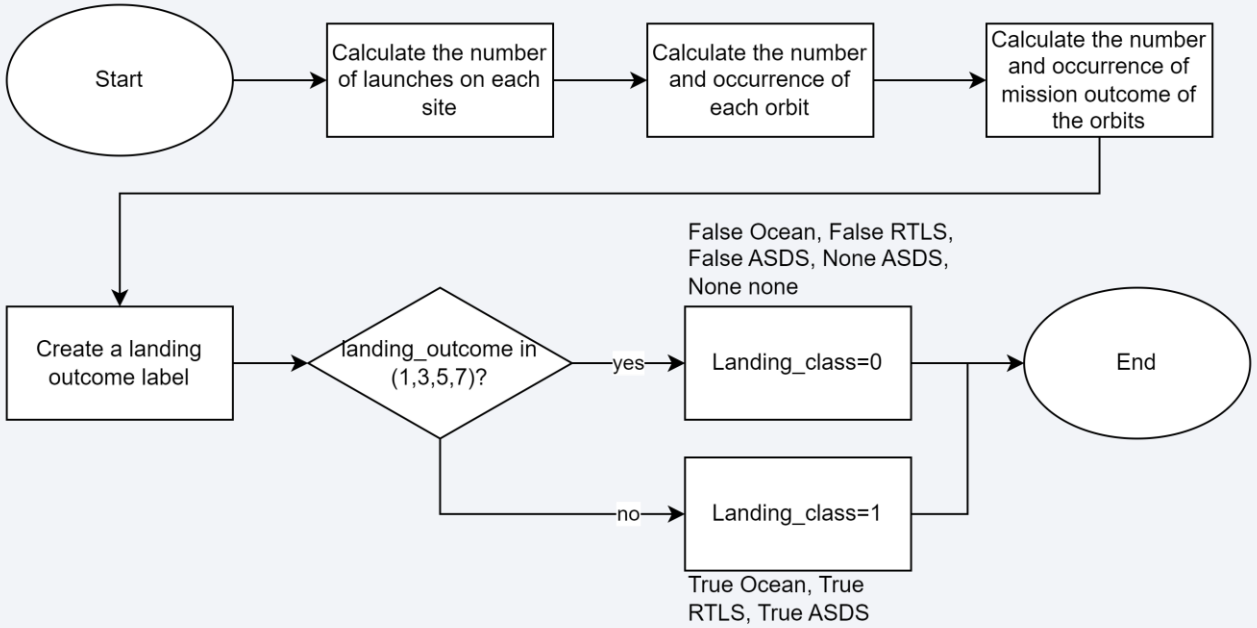
- First, we request the data from URL (in this case, Wikipedia page). Then we extract all column and variables name from HTML header. Finally, the response is parsed so we can convert it to a data frame.



Github URL:
https://github.com/Johanesary/TFSTraining_DS/blob/3f7347d7506ec27ea91025af05756626673373 89/Applied%20Data%20Science%20Capstone%20JRY2185/jupyter-labs-webscraping.ipynb

Data Wrangling

- For data wrangling, we do calculate the following things, number of launches on each site, number and occurrence of each orbit, and the number and occurrence of mission outcome of the orbits. Finally, we create the new landing outcome label to make the data easier to understand. We assume that true Ocean, true RTLS, and true ASDS as successful landing. Otherwise, failure landing.



Github URL:
https://github.com/Johanesary/TFSTraining_DS/blob/3f7347d7506ec27ea91025af05756626673373 89/Applied%20Data%20Science%20Capstone%20JRY2185/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We use multiple types of chart to present our data, scatter plot, bar plot, and line plot.
- Scatter plot is used to see the relationship between two different group such as, flight number vs orbit, payload vs launch site, etc.
- Bar plot is used to compare the value between each group, e.g., comparing the average success rate for each orbit.
- Line plot shows the data changes over time. For example, understanding the success launch rate over some period of time.
- Github URL: https://github.com/Johanesary/TFSTraining_DS/blob/367f97a88457330dae0c15cbc53348783b11b86e/Applied%20Data%20Science%20Capstone%20JRY2185/EDA%20with%20Visualization.ipynb

EDA with SQL

- There are four launch sites, CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40
- The average payload mass for booster version F9 v.1.1 is 2534.667 kg
- Booster version F9 FT B1xxx has successfully landed with payload mass between 4000 kg and 6000 kg
- The mission outcome shows 1 failure (in flight), 99 success, and 1 success with payload status unclear.
- The most landing outcome between 2010, 4th June and 2017, 20th March is No attempt

URL Github:
https://github.com/Johanesary/TFSTraining_DS/blob/367f97a88457330dae0c15cbc53348783b11b86e/Applied%20Data%20Science%20Capstone%20JRY2185/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

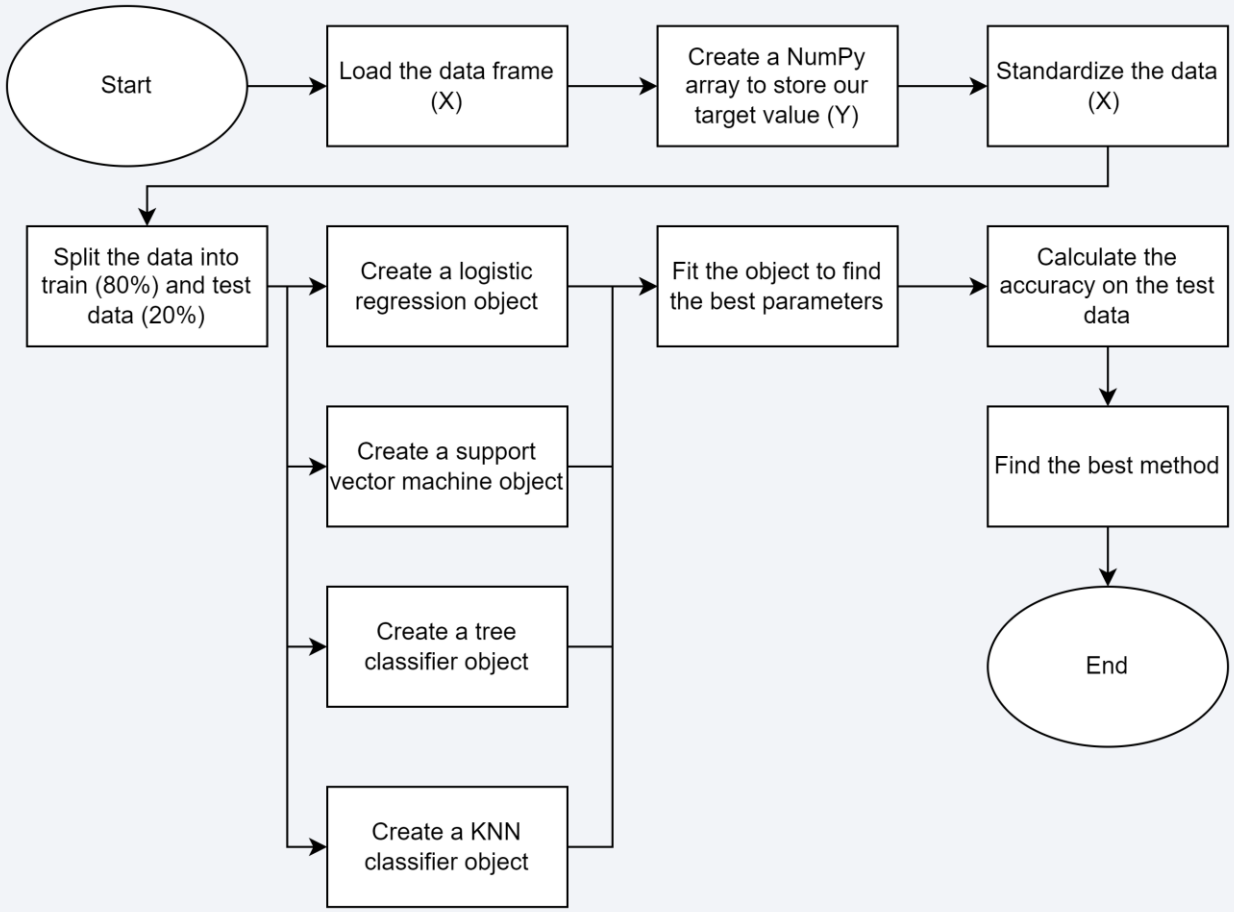
- We use markers and circle to point out our launch site. Moreover, we differentiate the color to show launch outcomes, red for unsuccessful launch (class=0) and green for successful launch (class=1).
- We use line to determine the distance from the launch site to another point. In this case, we find interest on railway, highway, and coastline.
- Github Link: https://github.com/Johanesary/TFSTraining_DS/blob/367f97a88457330dae0c15cbc53348783b11b86e/Applied%20Data%20Science%20Capstone%20JRY2185/Launch%20Site%20Proximities.ipynb

Build a Dashboard with Plotly Dash

- We shows an interactive dashboard in pie chart to understand success launch in each site. When we click to specific site, we can get more understanding about the number of launch and the result in one site.
- Moreover, we have further information to observe correlation between payload and mission outcomes for selected site.

Predictive Analysis (Classification)

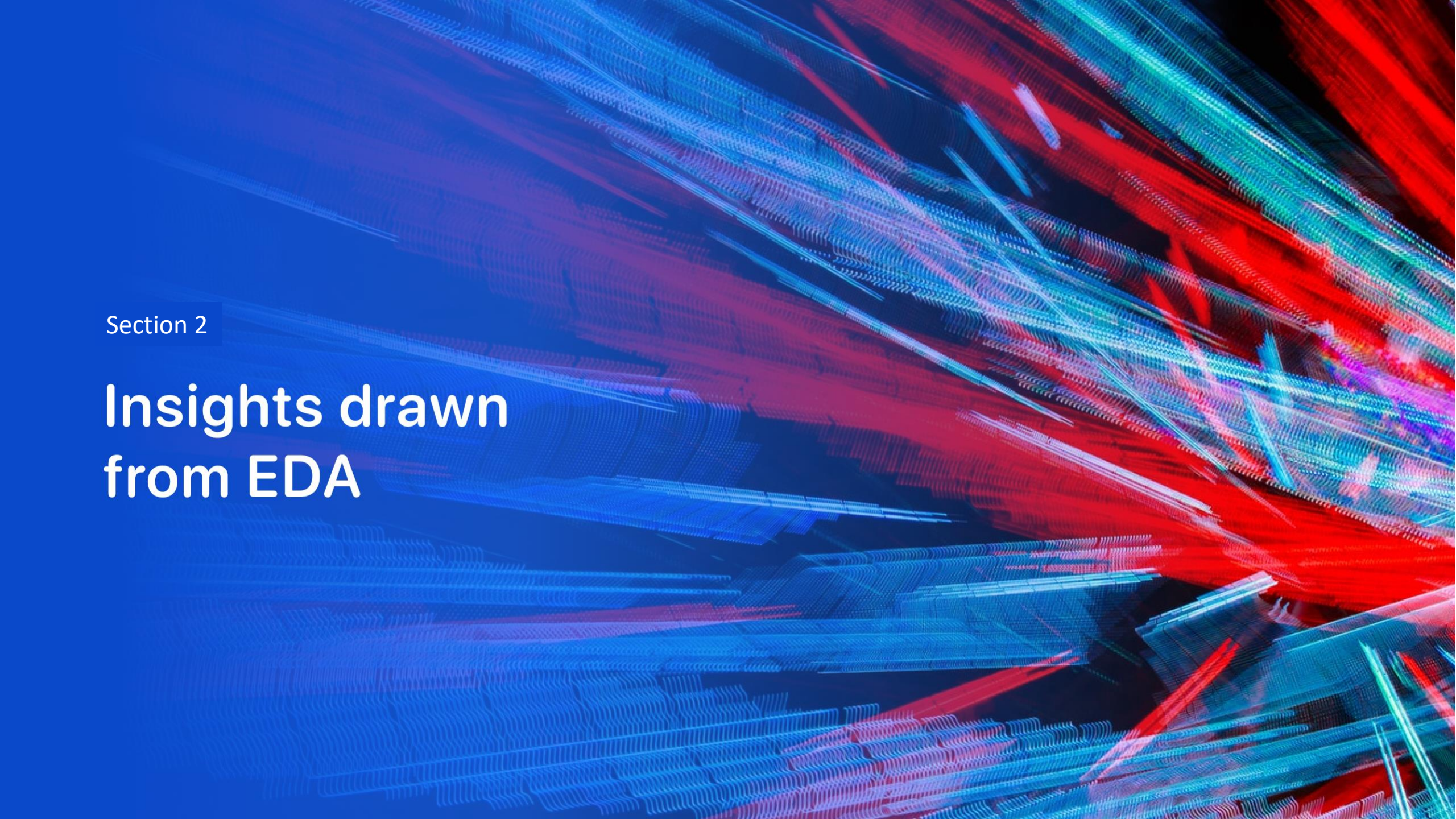
- We load the data and create the NumPy array to store our target value. Furthermore, we standardize the data to make it easier to analyze. Before creating the object model, we split the data into train and test data with proportion of 80% and 20%, respectively.
- Finally, we train our model and find the best parameters for each model. We compare the model with the test data before choosing the best method or model.



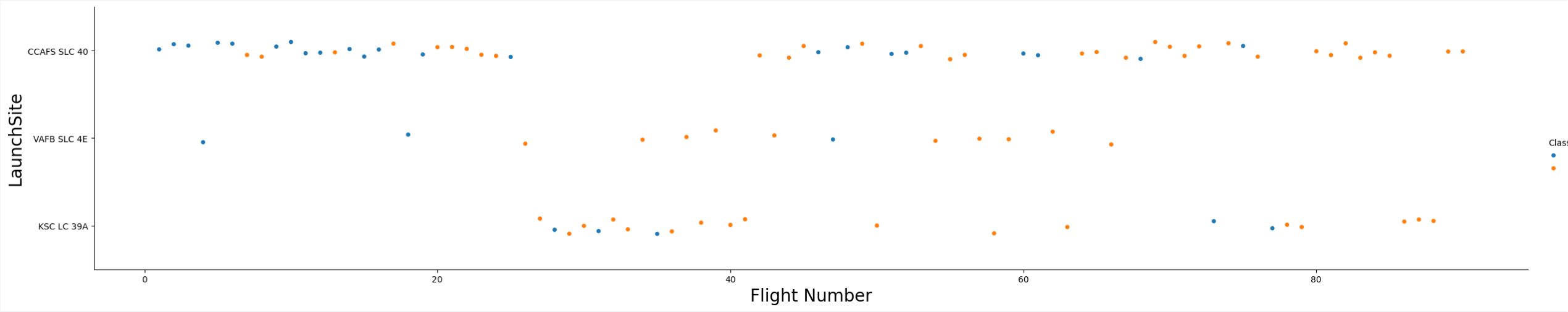
Github URL:
https://github.com/Johanesary/TFSTraining_DS/blob/3f7347d7506ec27ea91025af05756626673373 89/Applied%20Data%20Science%20Capstone%20JRY2185/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

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


Section 2
Insights drawn from EDA

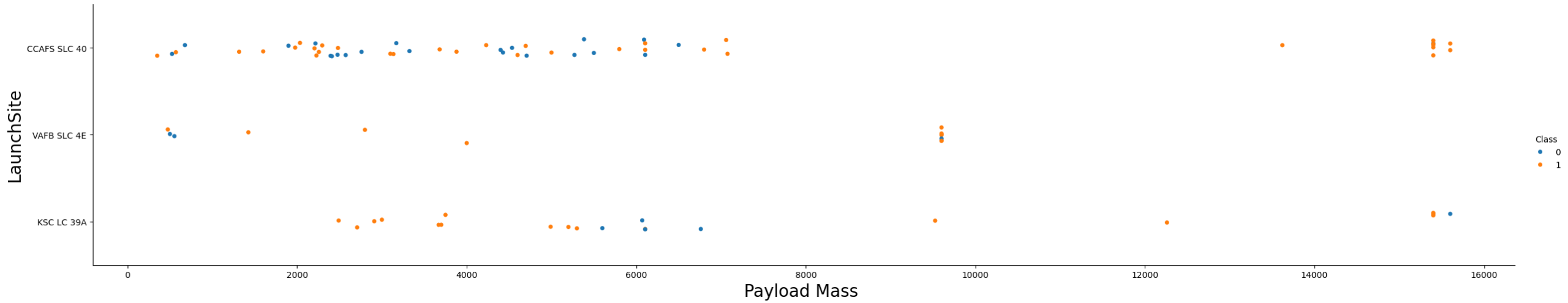
Flight Number vs. Launch Site



- The scatter plot above shows that smaller flight number (1-25) tends to have CCAFS SLC 40 as their launch site.
- Then from flight number 26-41, KSC LC 39A becomes frequent choice for launch site.
- The flight number 42 or higher are scattered along with these three Launch Sites. However, CCAFS SLC 40 is still the most frequent site to launch the Falcon 9.
- The class 1 (yellow) shows successful launch while class 0 (blue) indicates failed launch.

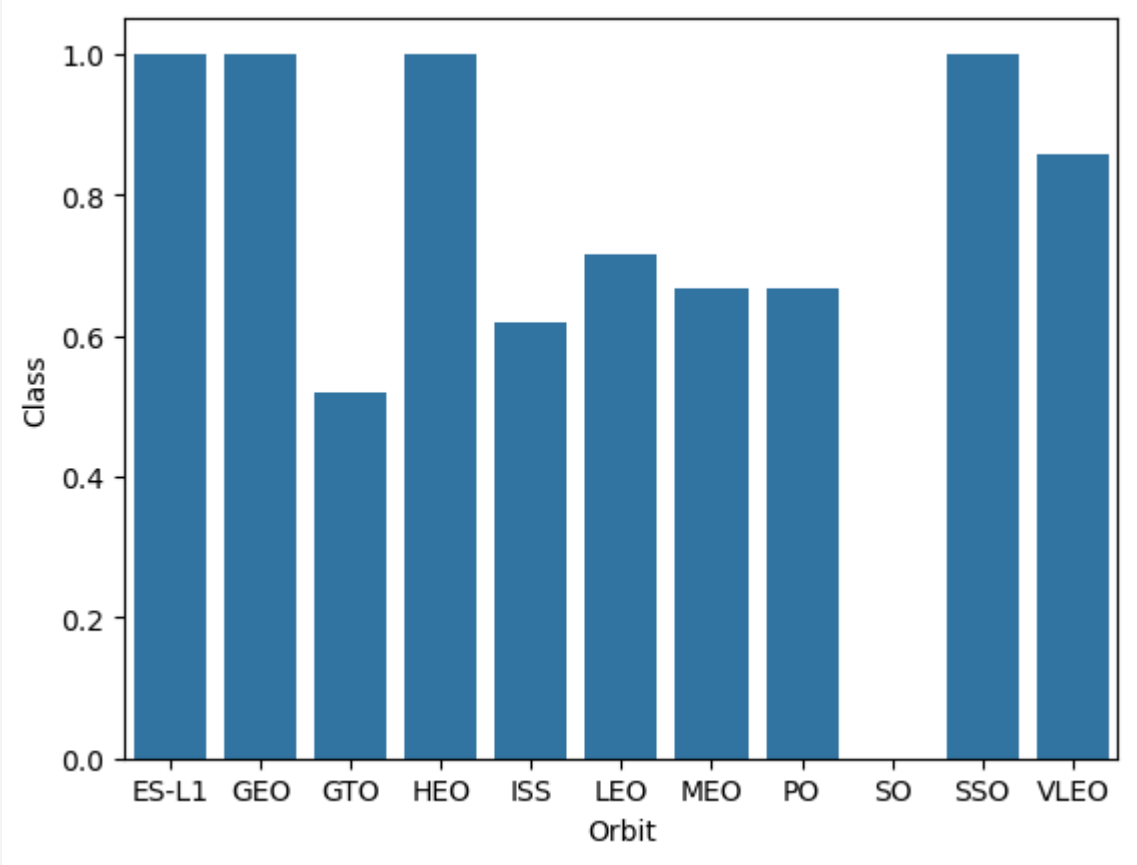
Payload vs. Launch Site

- The picture below shows the correlation between launch site and payload mass. For the payload less than 8000 kg, it is more likely to be launched in CCAFS SLC 40. For payload more than 8000 kg, the launch site is scattered randomly.
- Moreover, the payload more than 8k kg shows likelihood to have desired outcome (class=1)



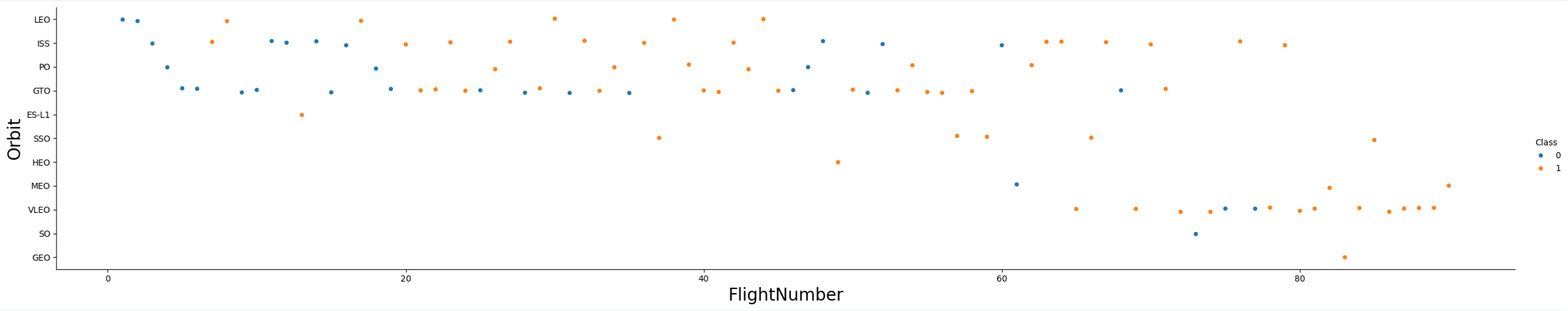
Success Rate vs. Orbit Type

- There several orbits have 100% success rate for launch, such as ES-L1, GEO, HEO, and SSO.
- On the other hand, SO orbit has 0% success rate.
- The rest of orbit have 50% success rate on average except VLEO orbit has more than 80% success rate



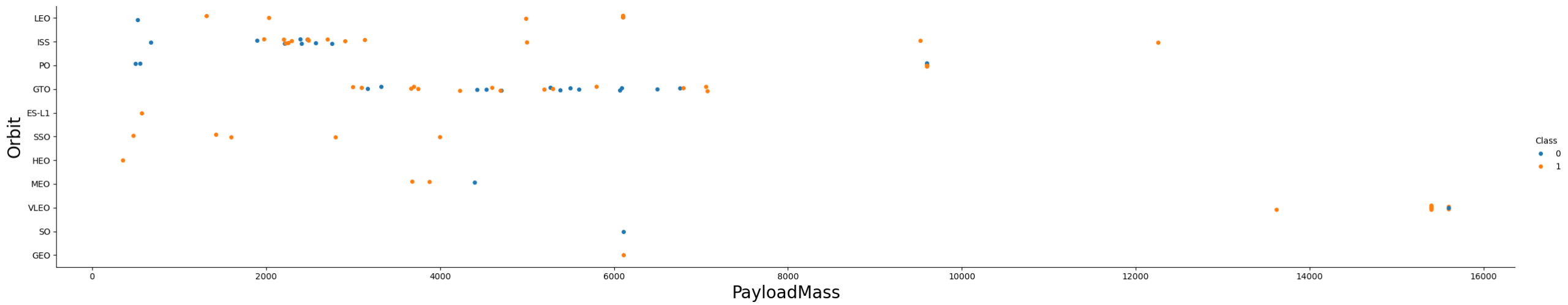
Flight Number vs. Orbit Type

- The initial flight number from 0 to 60 is more focus on several orbits like LEO, ISS, PO, GTO, and ES-L1
- While after flight number 60, the focus changes to other orbits such as SSO, HEO, MEO, VLEO, SO, and GEO



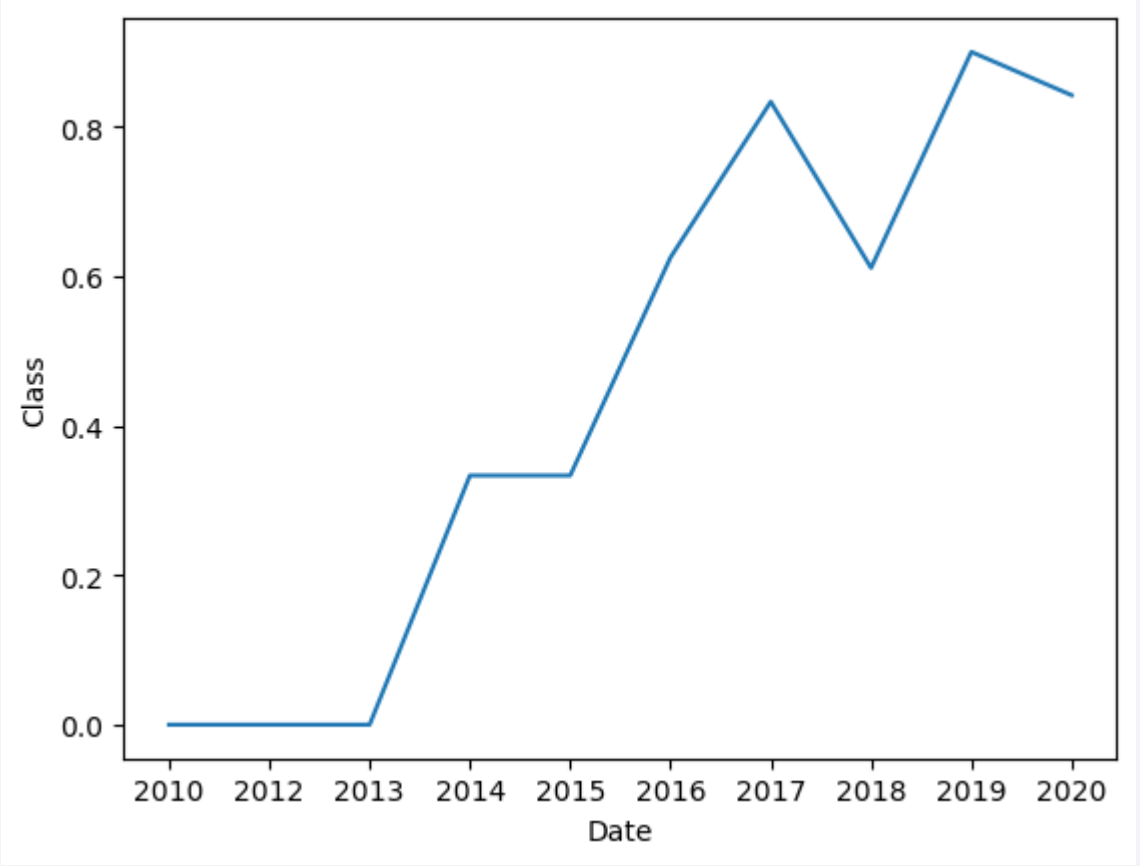
Payload vs. Orbit Type

- Several orbits (LEO, ISS, PO, GTO, ES-L1, SSO) are launched with lighter payload mass (less than 8000 kg)
- While other orbits (HEO, MEO, VLEP, SO, GEO) have heavier payload mass



Launch Success Yearly Trend

- From 2010 to 2013 the launch shows 0% success rate.
- Since 2013, the launch shows positive trend on success rate. It increased over the time and reached the peak in 2019 with almost 100% success rate.



All Launch Site Names

- There four unique launch sites as shown in the image below

```
[14]: %sql select distinct launch_site from spacextable;
 * sqlite:///my_data1.db
Done.
```

[14]:

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

```
[15]: %sql select distinct launch_site from spacextable where launch_site like 'CCA%';
```

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

[15]:

Launch_Site
CCAFS LC-40
CCAFS SLC-40

Total Payload Mass

- The rocket has several booster versions as shown in the tables below.
- From the tables we can see the payload mass for each booster version, for example F9 B4 B1039.2 launch 2647 kg in total

```
[20]: %sql select booster_version, sum(PAYLOAD_MASS__KG_) as 'payload_mas (KG)' from spacextable group by booster_version;
 * sqlite:///my_data1.db
Done.
```

[20]:

Booster_Version	payload_mas (KG)
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.1	3600
F9 B5 B1046.2	5800
F9 B5 B1046.3	4000
F9 B5 B1046.4	12050
F9 B5 B1047.2	5300
F9 B5 B1047.3	6500
F9 B5 B1048.2	3000
F9 B5 B1048.3	4850
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.2	9600
F9 B5 B1049.3	13620
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.6	15440
F9 B5 B1049.7	15600
F9 B5 B1051.2	4200
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.5	14932
F9 B5 B1051.6	15600
F9 B5 B1056.2	2268
F9 B5 B1056.3	6956
F9 B5 B1056.4	15600
F9 B5 B1058.2	5500
F9 B5 B1058.3	15600
F9 B5 B1058.4	2972
F9 B5 B1059.2	1977
F9 B5 B1059.3	15410
F9 B5 B1059.4	3130
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600
F9 B5B1047.1	7075
F9 B5B1048.1	9600
F9 B5B1049.1	7060
F9 B5B1050	2500
F9 B5B1051.1	12055
F9 B5B1054	4400
F9 B5B1056.1	2495
F9 B5B1058.1	12530
F9 B5B1059.1	2617
F9 B5B1060.1	4311
F9 B5B1061.1	12500
F9 B5B1062.1	4311
F9 B5B1063.1	1192
F9 FT B1021.2	5300
F9 FT B1029.2	3669
F9 FT B1031.2	5200
F9 FT B1032.2	4230
F9 FT B1035.2	2205
F9 FT B1036.2	9600
F9 FT B1038.2	2150
F9 FT B1019	2034
F9 FT B1020	5271
F9 FT B1021.1	3136
F9 FT B1022	4696
F9 FT B1023.1	3100
F9 FT B1024	3600
F9 FT B1025.1	2257
F9 FT B1026	4600
F9 FT B1029.1	9600
F9 FT B1030	5600
F9 FT B1031.1	2490
F9 FT B1032.1	5300
F9 FT B1034	6070
F9 FT B1035.1	2708
F9 FT B1036.1	9600
F9 FT B1037	6761
F9 FT B1038.1	475
F9 v1.0 B0003	0
F9 v1.0 B0004	0
F9 v1.0 B0005	525
F9 v1.0 B0006	500
F9 v1.0 B0007	677
F9 v1.1	14642
F9 v1.1 B1003	500
F9 v1.1 B1010	2216
F9 v1.1 B1011	4428
F9 v1.1 B1012	2395
F9 v1.1 B1013	570
F9 v1.1 B1014	4159
F9 v1.1 B1015	1898
F9 v1.1 B1016	4707
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

Average Payload Mass by F9 v1.1

- The average for the booster version F9 v.1.1 is 2534.67 kg.

```
[24]: %sql select avg(PAYLOAD_MASS__KG_) as PAYLOAD_MASS__KG_ from spacextable where Booster_Version like 'F9 v1.1%';
 * sqlite:///my_data1.db
Done.
[24]:
PAYLOAD_MASS__KG_
2534.6666666666665
```

First Successful Ground Landing Date

- The data shows that they successfully did ground landing for the first time on December 22, 2015

```
[29]: %sql select min(date) as first_date from spacextable where Landing_Outcome = 'Success (ground pad)' and mission_outcome = 'Success' ;
 * sqlite:///my_data1.db
Done.
```

[29]:

first_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- There are four booster version, mainly F9 BT, has successfully landed in drone ship with payload between 4000 and 6000

```
[33]: %sql select Booster_Version from spacextable where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS_
 * sqlite:///my_data1.db
Done.
```

[33]:

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

Total Number of Successful and Failure Mission Outcomes

- There are one fail launch, 99 successful launches, and one success with payload status unclear

```
[36]: %sql select mission_outcome, count(1) as total from spacextable group by mission_outcome;
 * sqlite:///my_data1.db
Done.
```

[36]:

Mission_Outcome	total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- F9 B5 B1xxxx is proven to be able carry maximum payload.

```
[39]: %sql select Booster_Version from spacextable where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextable);
 * sqlite:///my_data1.db
Done.
```

[39]:

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

- There are 7 launches in 2015. Failure landing outcome happened in the first month and the fourth month of 2015. They all happened in same launch iste, CCAFS LC-40.

```
[41]: %sql select substr(Date, 6,2) as month, landing_outcome, booster_version, launch_site from spacextable where substr(Date,0,5)='2015';
 * sqlite:///my_data1.db
Done.
```

[41]:

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
02	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
03	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

- No attempt is the most count of landing outcome with 10 times between 4 June 2010 and 20 March 2017. It is followed by success and failure (drone ship) with 5 for both outcomes.

```
[42]: %sql select landing_outcome, count(1) as total from spacextable where date between '2010-06-04' and '2017-03-20' group by landing_outcome
 * sqlite:///my_data1.db
Done.
```

[42]:

Landing_Outcome	total
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

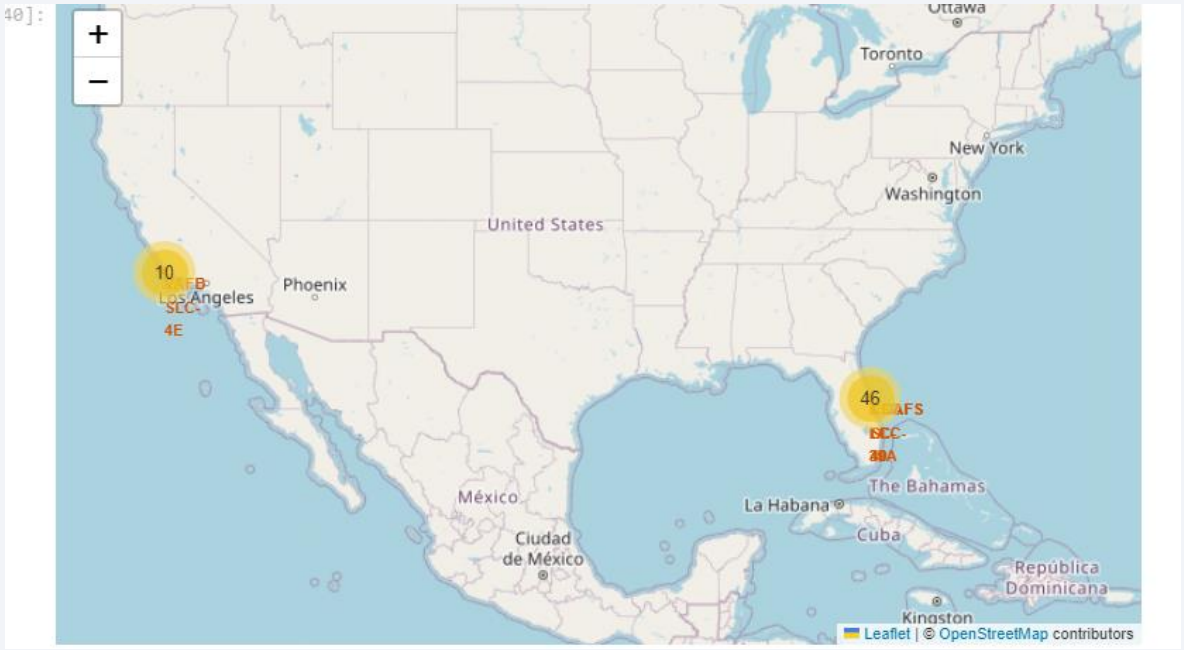
Launch Sites Proximities Analysis

Launch Sites

- The Map shows the location of launch sites. There are two areas, the east and the west of USA. Mark and Circle are shown to indicate the launch site.

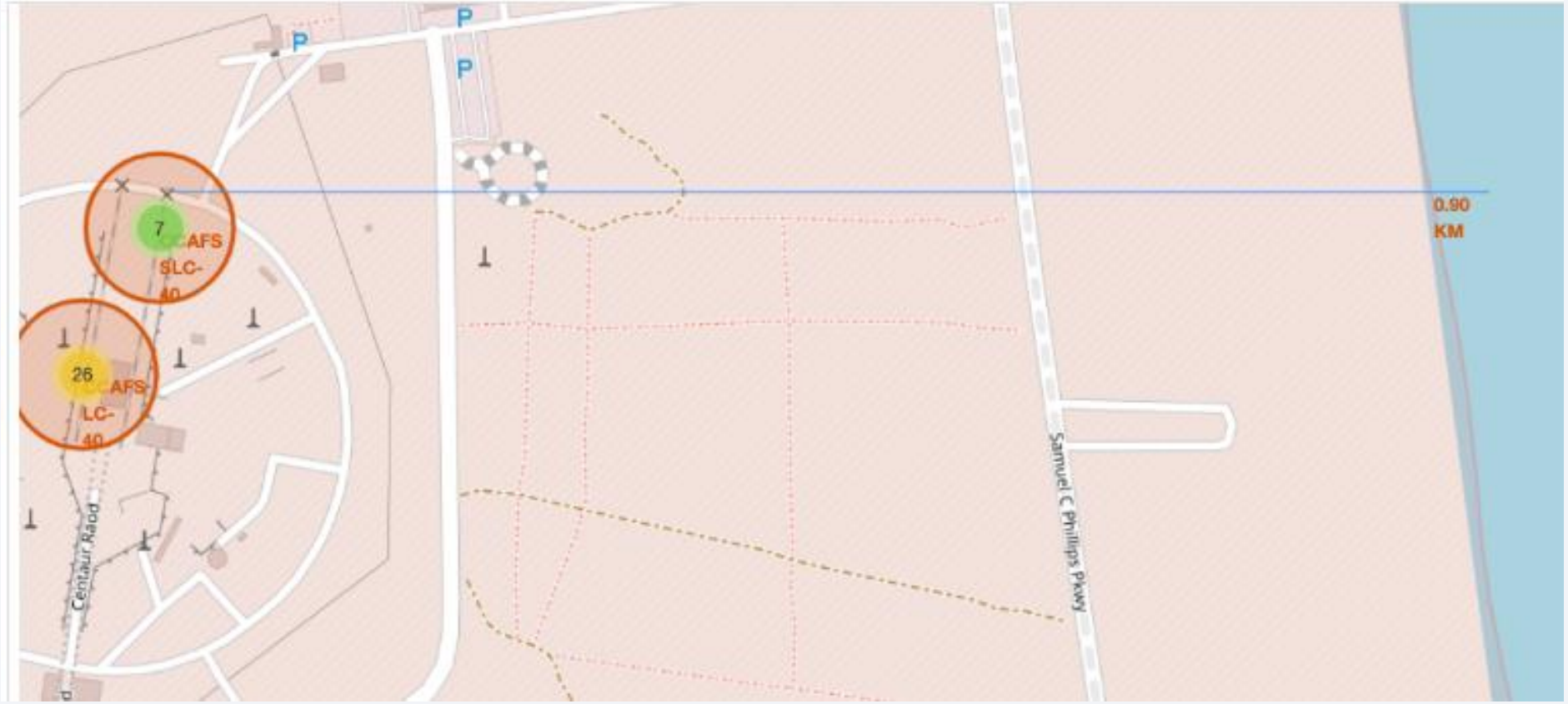
Number of Launch and Launch Outcome

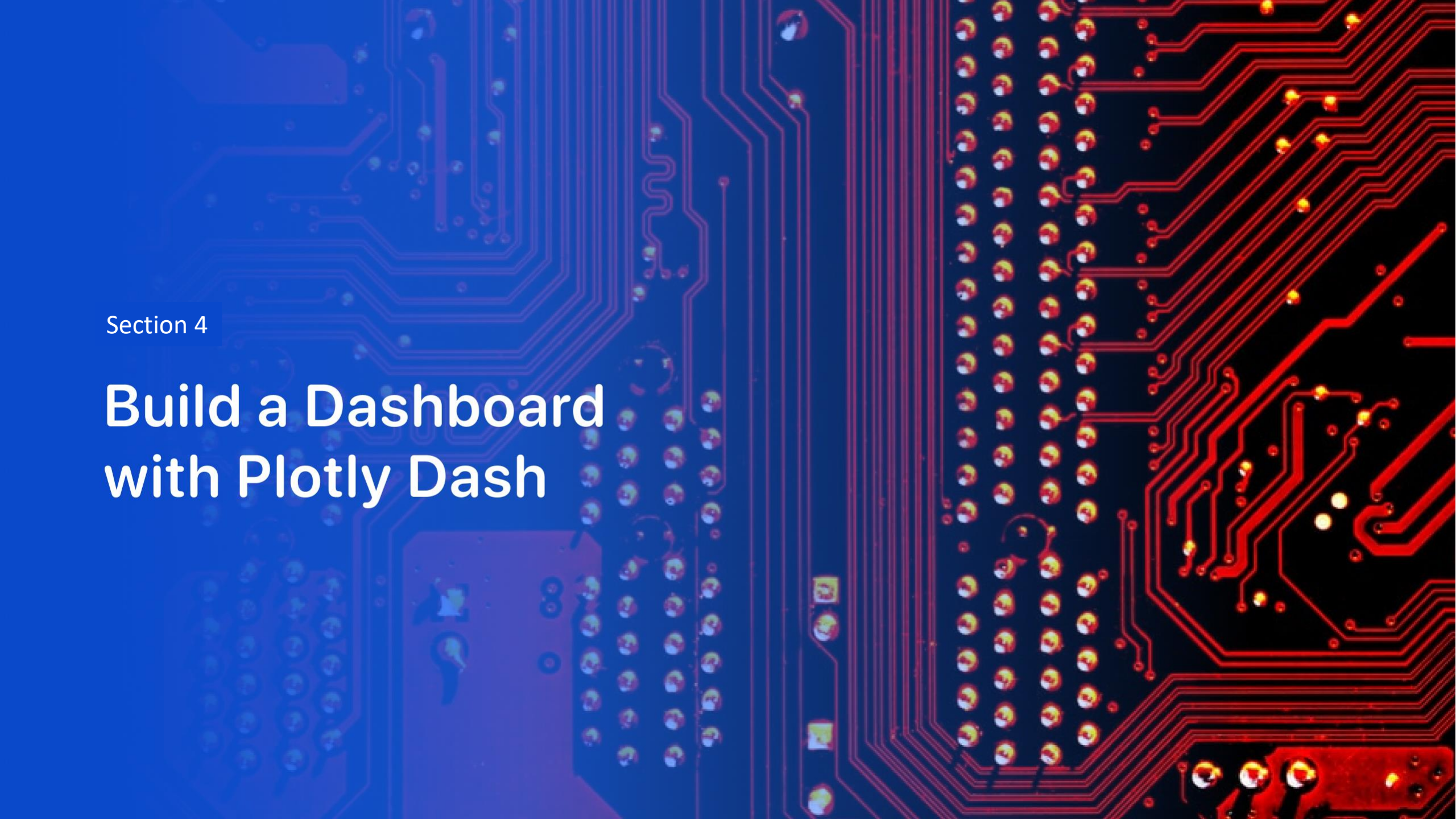
- The West part of USA shows significant number of launch with 46. While the east has 10 launches. The green mark shows successful launch, and the red mark shows unsuccessful launch



Distance between Launch Site with Coastline

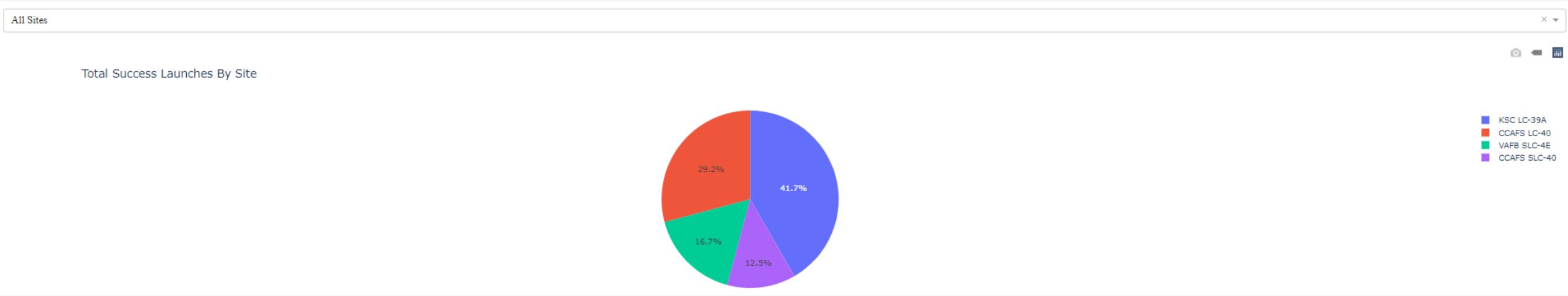
- CCAFS SLC-40 launch site has 0.90 km distance with the nearest coastline. There is a line that shows the distance between two locations




Section 4
Build a Dashboard with Plotly Dash

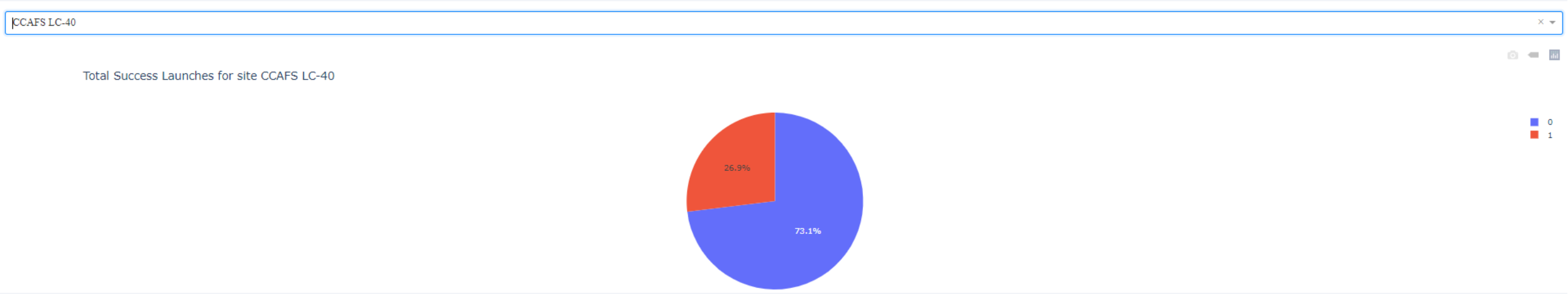
Total Success Launches by Site

- The pie chart show success launches by site. The different site is shown by different colour. On the right side, there is a legend which explain what colour represent the name of site.



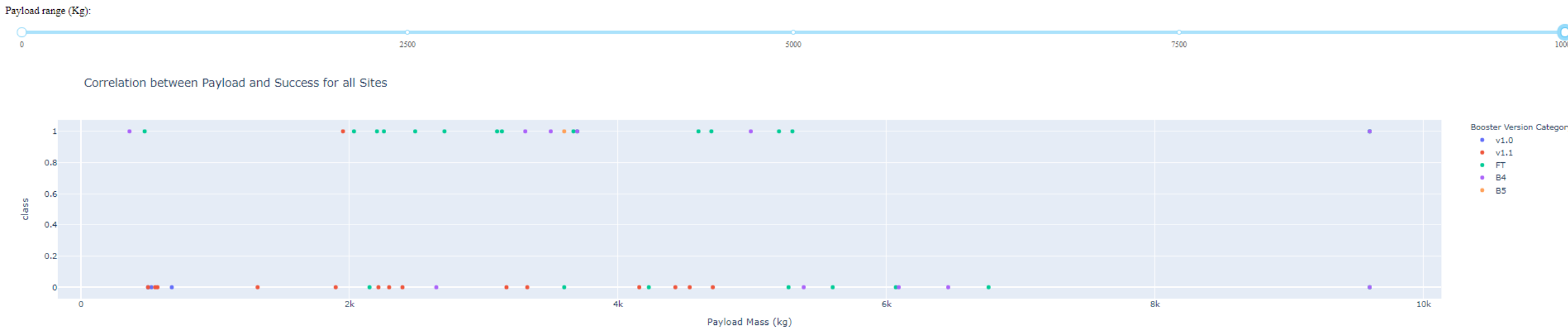
Total Success Launches for selected site

- When we click to one the site, we can see the detail information between the success launch compared to the total launch for that selected site. The legend on the right shows that red colour represents value 1 (successful), and blue colour represents value 0 (unsuccessful) for CCAFS LC-40



Scatter Plot for Payload and Class

- Another interactive dashboard that we can see is correlation between payload and class for sites. It seems there is no real correlation between payload and class. On the right side, there is a legend to show representation of colour to booster version.

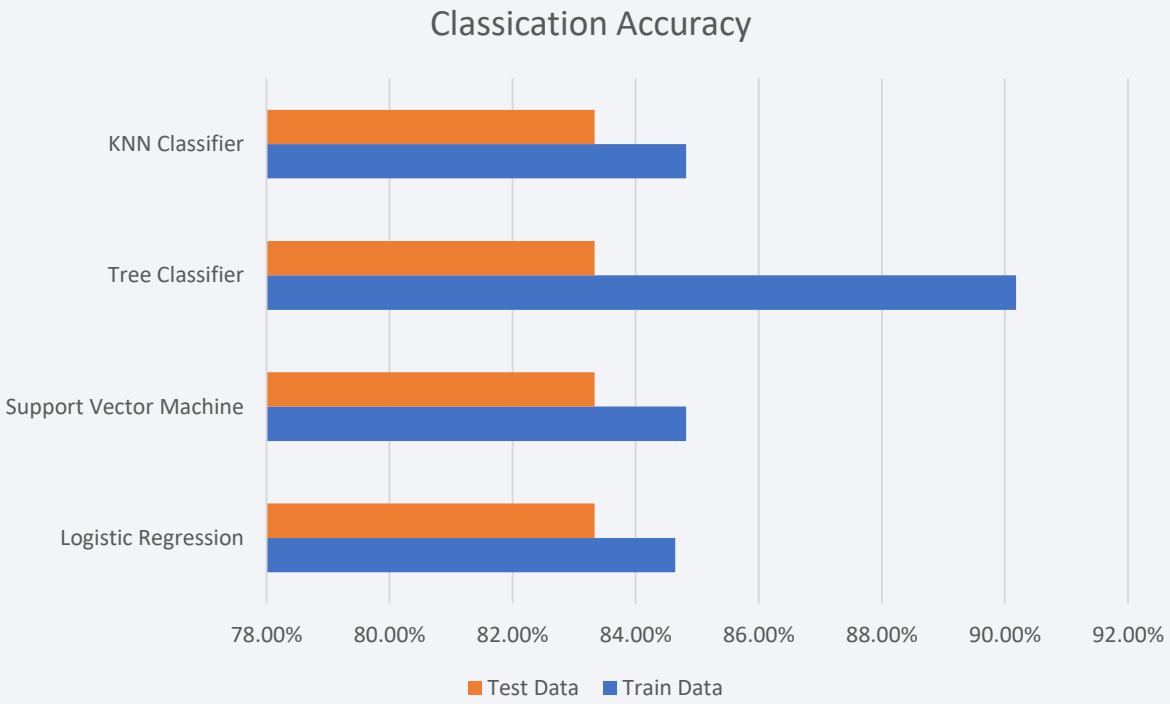


Section 5

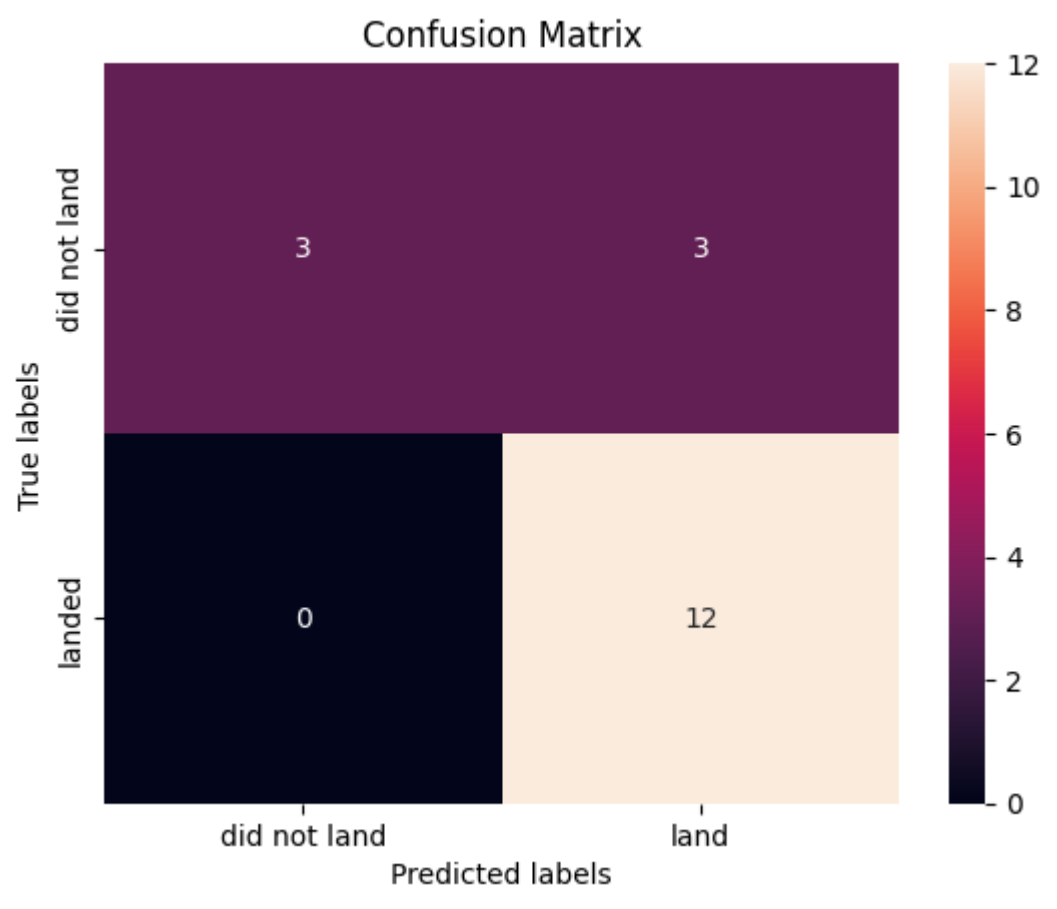
Predictive Analysis (Classification)

Classification Accuracy

- From the graph on the right side, we can conclude that all models have same accuracy for the test data with 83.33%.
- Moreover, the tree classifier shows the best accuracy in training data with more than 90%. It is followed by SVM and KNN classifier with 84.82%. The lowest accuracy is logistic regression with 84.64%.



Confusion Matrix



- For the best performing model, we can see the confirmation matrix on the left side. The picture shows that the model successfully predicts the landed rocket with 80% sensitivity (12/15) and did not land rocket with 100% specificity (3/3). Overall, the model gives 83% accuracy (15/18).

Conclusions

- The number of test data is 18
- The data is split randomly to train and test data with proportion of 80% and 20% of total data.
- Tree classifier shows the most accuracy with more than 90% in train data
- All model shows same accuracy for the test data with 83.33%

Thank you!

Appendix

- Github URL: https://github.com/Johanesary/TFSTraining_DS/tree/367f97a88457330dae0c15cbc53348783b11b86e/Applied%20Data%20Science%20Capstone%20JRY2185