IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Puttaranun Boonchit <25th November 2021>

Outline

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

Executive Summary

The cost of Falcon 9 launches announced by SPACEX is much cheaper than the other providers due to the reuse of the first stage. Therefore, it is possible to predict the cost of the launch from the reusability of the first stage.

In this project, we conducted the whole data science methodologies. Observing various launching features, we could understand the insights of each launching feature, i.e. launching site, payload mass and orbits, affecting the first stage's success rate. For instance, the launch to SSO orbits is the most promising choice. Besides, we investigated the correlation among those features.

Then, the success of the first stage was determined with those features by tuned machine learning methods. By modeling the machine learning methods with various parameters, the result suggests that the Decision tree is overfitting, compared to the others. Therefore, other tuned models are basically efficient in this project scenario.

Introduction

The cost of Falcon 9 launches ..

SPACEX

62

million dollars

< Due to the first stage reuse >

OTHER COMPANIES

165

million dollars

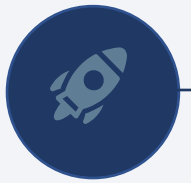
? In order to determine the cost of the launch, this project aims to predict if the first stage will land successfully or not by referring to the SpaceX's information.

Section 1

Methodology

Methodology

1. Data Collection



To explain how data was collected through web-scraping, and REST API.

2. Data Wrangling

To process the collected data for the further use.

3. Exploratory Data Analysis (EDA)

To perform the EDA by using visualization and SQL .

4. Interactive Visual Analytics

To perform the interactive visual analytics by using Folium and Plotly Dash.

5. Predictive Analysis

To evaluate classification models with tuned hyperparameters for predicting the outcome.

Data Collection

The data set was collected by two methods:

REST API

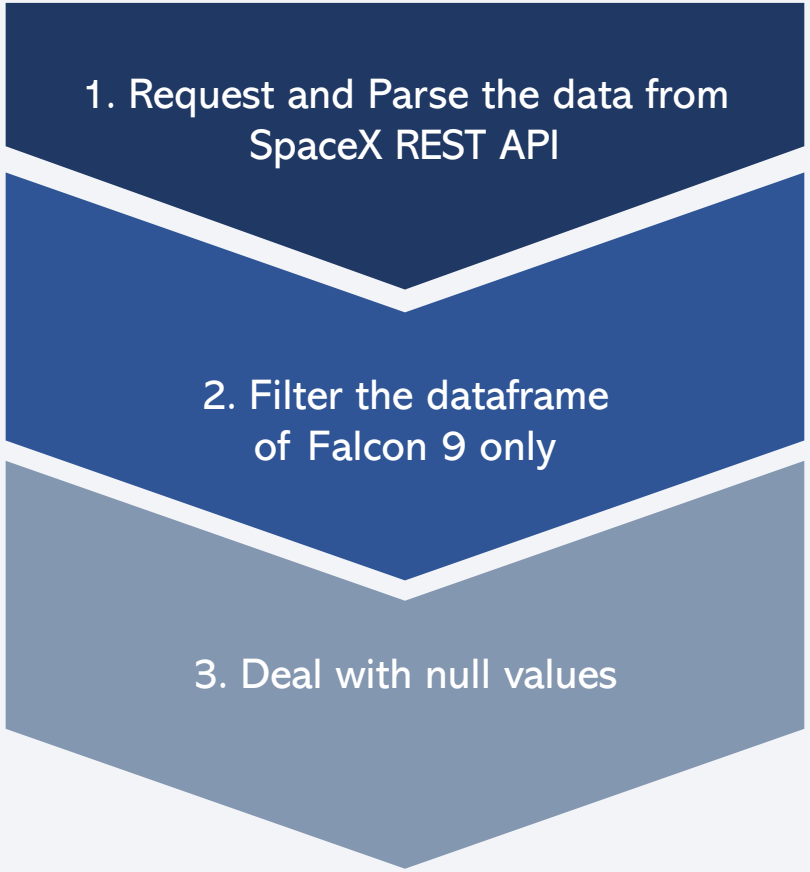
of the SpaceX

WEB SCRAPING

by BeautifulSoup

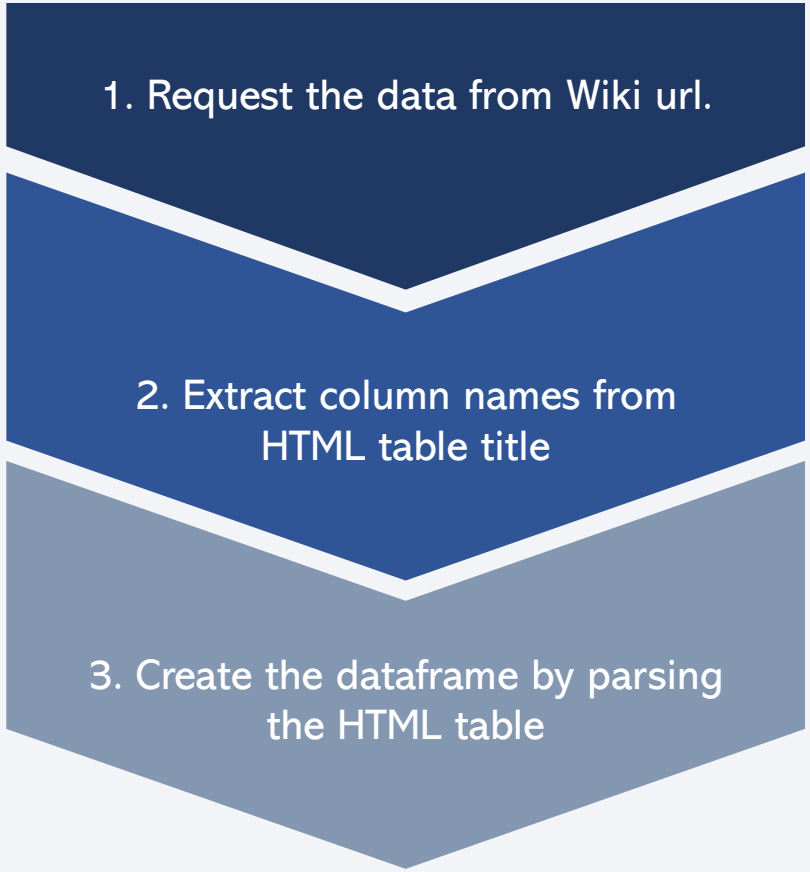
Data Collection – SpaceX API

- Requested by GET request, the data from SpaceX API is obtained in .json, then, it was converted to dataframe. Since the data was mostly in the ID form, we got more information from other APIs.
- The launching data of Falcon 1 and Falcon 9 were in the dataframe, so we filtered to only that of Falcon 9.
- Lastly, we dealt with the null values by using mean values instead.



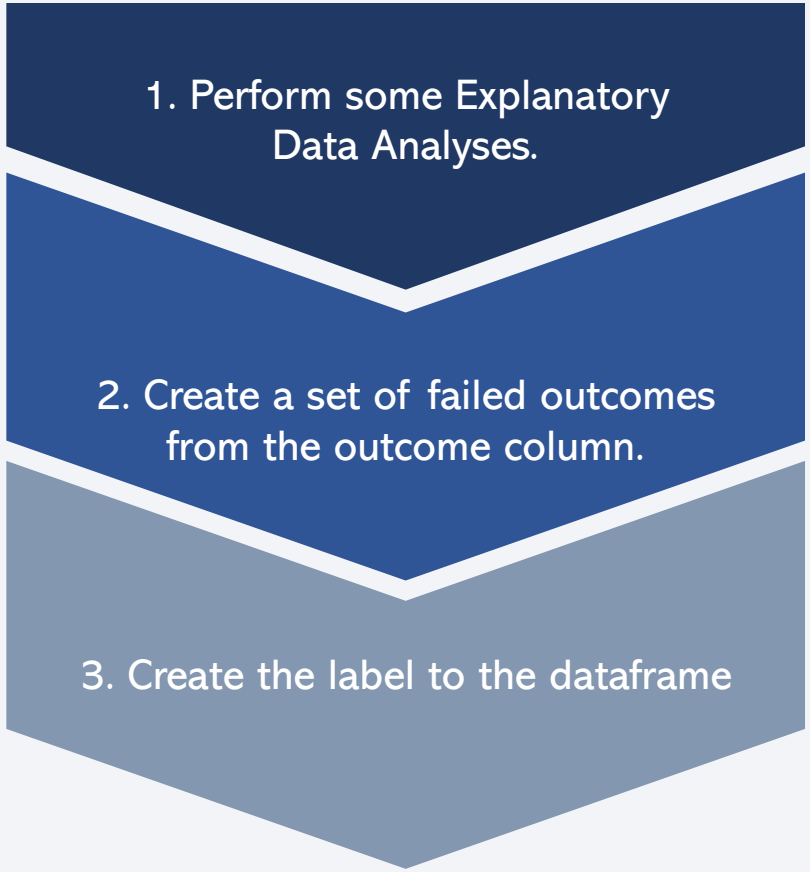
Data Collection – Web Scraping

- We requested the data from Wiki url by GET request.
- Then, we converted the HTML table into the Beautiful Soup object, and extract the table title by using find_all function to be the column names.
- Lastly, we parsed all information in the table to create the dataframe.



Data Wrangling

- Firstly, the fundamental explanatory data analysis was performed to get insights of the data.
- Then, we created a set of failed outcomes from the outcome column.
- Lastly, we created the labels to the dataframe regarding the outcomes (0 = fail, 1 = success).



EDA with Data Visualization

To **understand the relationship of the data features**, we performed the data visualization of

BAR CHART

- Orbit type vs Success rate

SCATTER PLOT

- Flight number vs Launch site
- Payload vs Launch site
- Flight number vs Orbit type
- Payload vs Orbit type

LINE GRAPH

- Year vs Success trend

EDA with SQL

In order to **gather information about the dataset**, we used SQL queries to get answers of these questions:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'KSC’
- 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 where the first succesful landing outcome in drone ship was acheived.
- List the names of the boosters which have success in ground pad 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

Build an Interactive Map with Folium

- Regarding the latitude and longitude, we **marked all launch sites on a map** with the circle area and text label by using *folium.Circle* and *folium.Marker* .

- To **mark the launch outcomes ,** we assigned the marker colors, and then added *folium.Marker* to the *MarkerCluster* object. (1 = success and 0 = failed)

- To calculate the distance on the map, we firstly added a *MousePosition* to obtain coordinate of a mouse, created the function calculating distances on a map, and then marked the distances from a site to its proximities (coastline, railway, highway, city) by using *PolyLine*.

Build a Dashboard with Plotly Dash

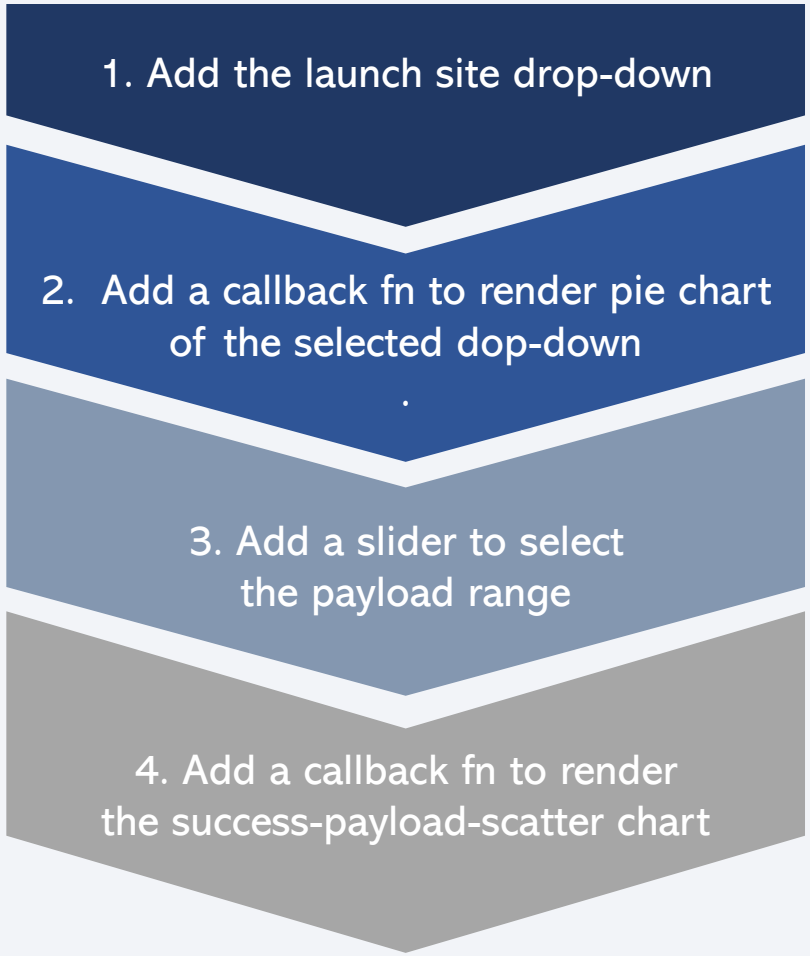
We would like to create a dashboard with components:

1. **Drop-down:** To select the launch site as an input, the *dcc.Dropdown* with the options of *All sites*, and *specific sites* was created.

2. **Success pie chart:** The *callback function* is created *t*o display the success at the site selected from the drop-down, .
   - If *All sites* option is selected, a pie chart with success rate at all sites are shown.
   - If a *specific site* is selected, a pie chart with success and failed outcome are shown.

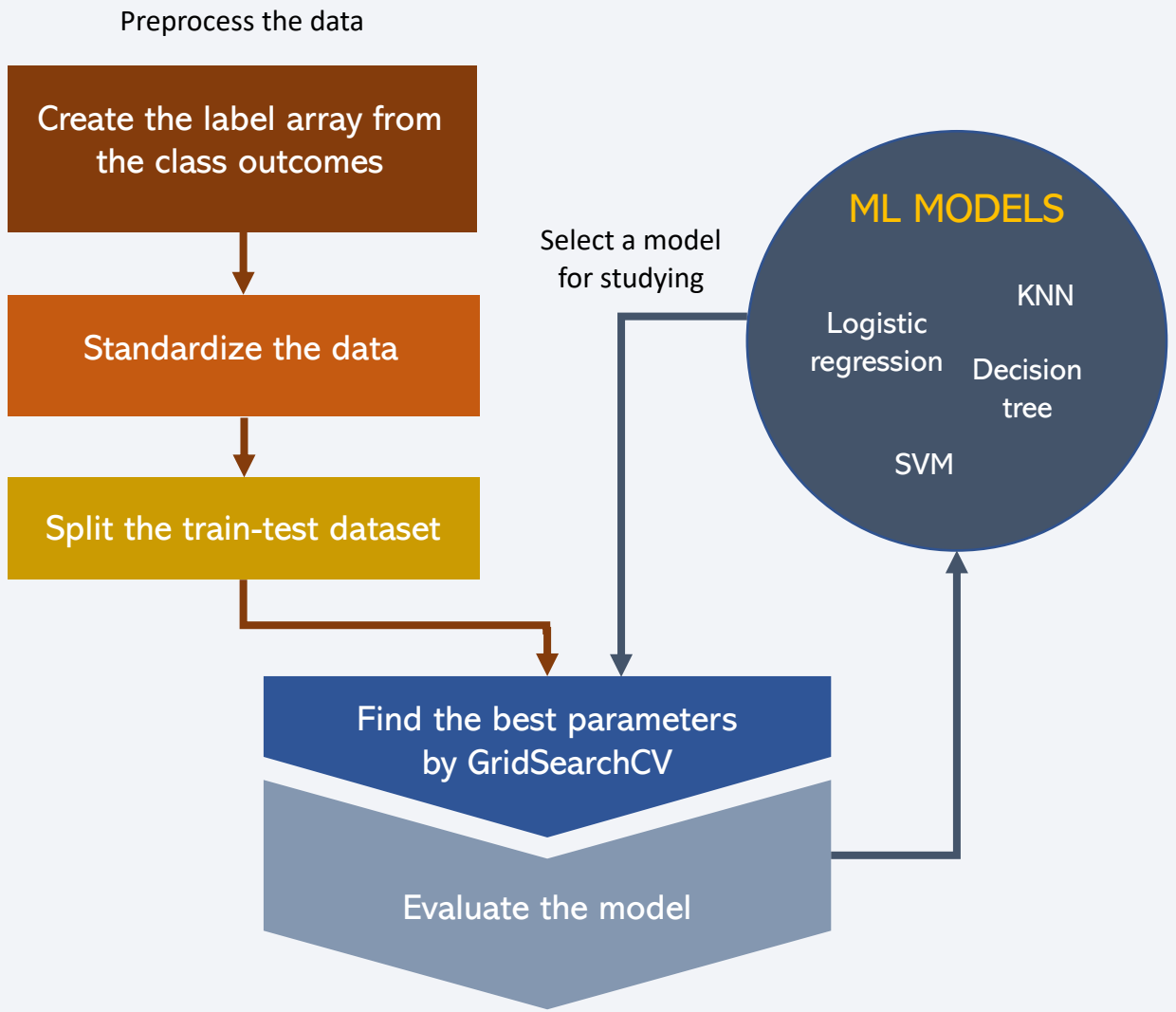
3. **Slider:** To select the payload range as an input (0 – 10,000 with step of 1,000), the *dcc.RangeSlider* was used.

4. **Payload scatter plot:** To display the outcomes at the specific payload range, we created the *callback function.*



Predictive Analysis (Classification)

- We firstly preprocessed the data, and prepare the label array, and train-test dataset.
- For each ML model, we determined the best parameters by GridSearchCV, and then evaluated the model by calculating the accuracy and confusion matrix.
- We compared the performance of 4 models: Logistic regression, KNN, SVM, Decision tress



Results & Discussions

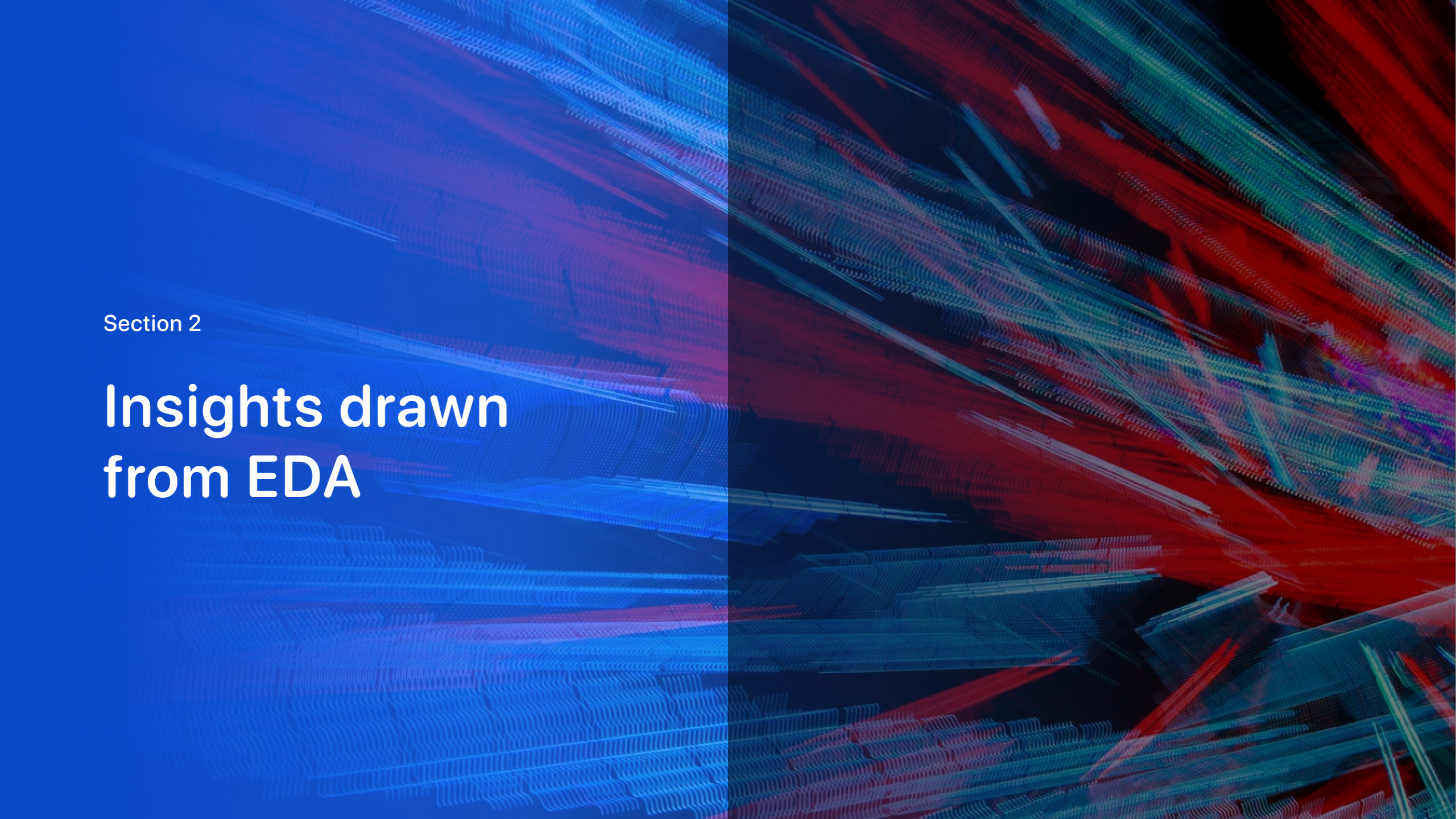
Section
- 2 -
Exploratory
Data Analysis

Section
- 3 -
Interactive
Data Analytics

Section
- 4 -
Predictive
Analysis

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site

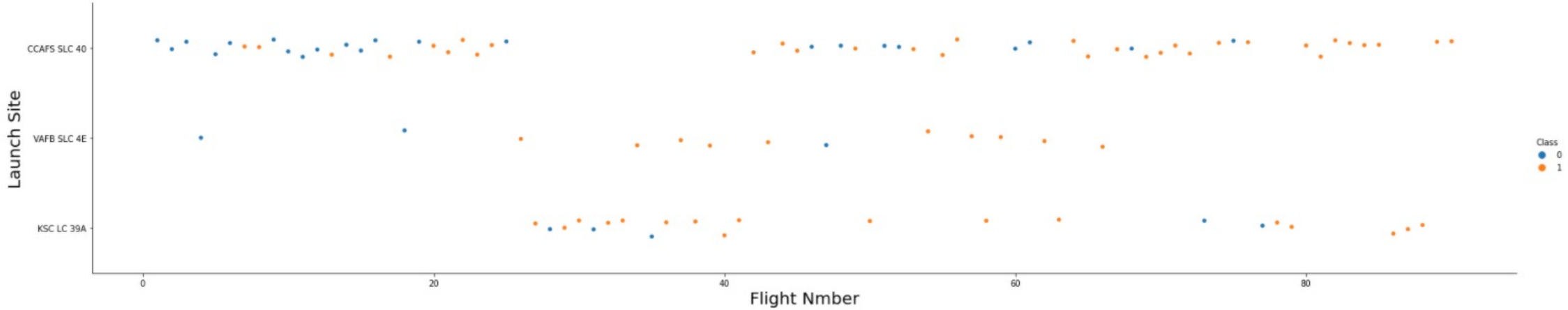


Figure 1: Scatter plot between Flight Number and Launch Site

The number of flights had been launched differently at the different launch site.

- At CCAFS SLC 40: The range of small and large flight numbers.
- At VAFB SLC 4E: The range of medium flight numbers. Launching with the small number all failed.
- At KSC LC 39A: The range of medium to large flight numbers.

At all sites, launching with the large number tends to have higher successful rate.

Payload vs. Launch Site

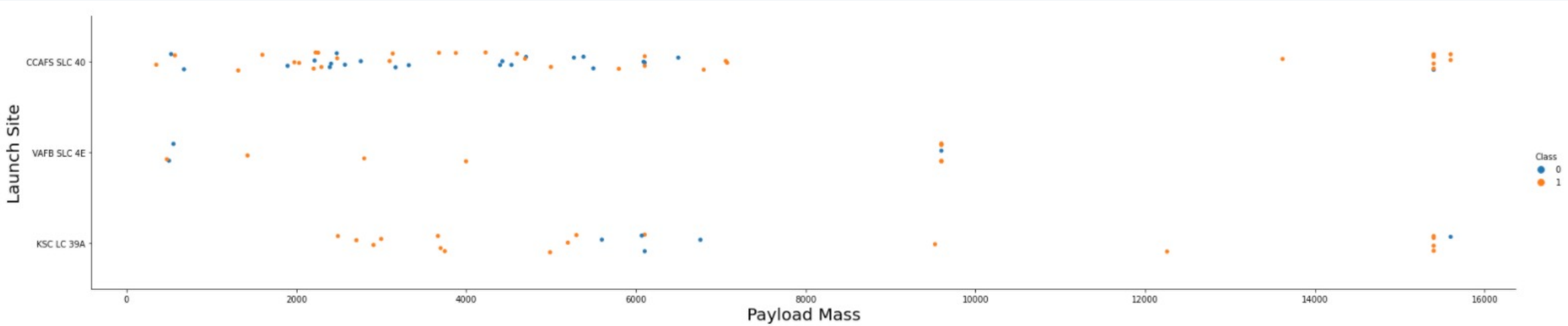


Figure 2: Scatter plot between Payload and Launch Site

- **At all launch sites, the launching had been mostly with the light payload mass (0 – 8,000).**
- **None of the launch at VAFB SLC 4E had heavy payload mass (>10,000).**

Success Rate vs. Orbit Type

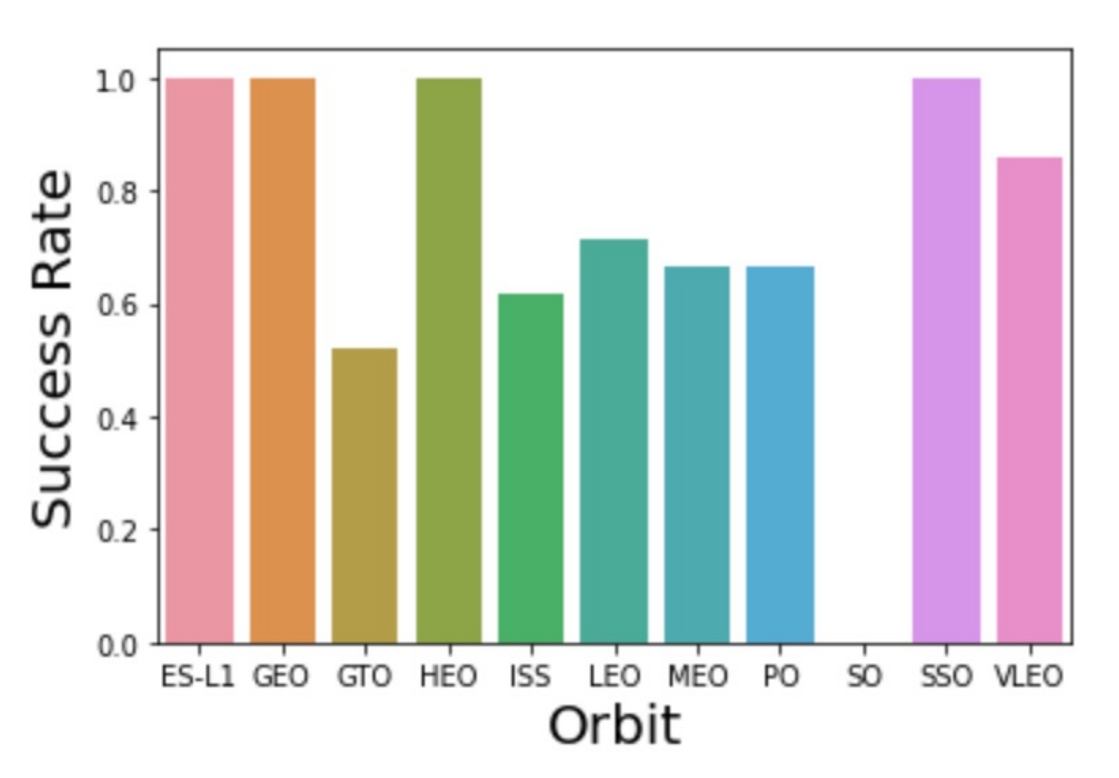


Figure 3: Bar chart representing the success rate of different orbit types

- Launching to the ES-L1, GEO, HEO, SSO orbit types had the highest success rate (at 1.0)
- The success rate was followed by that to the VLEO, and LEO, respectively.
- That of the launches to MEO, PO, ISS, and GTO were in the range of 0.5 – 0.7 .
- The success rate to SO orbit was the lowest (at 0)

Flight Number vs. Orbit Type

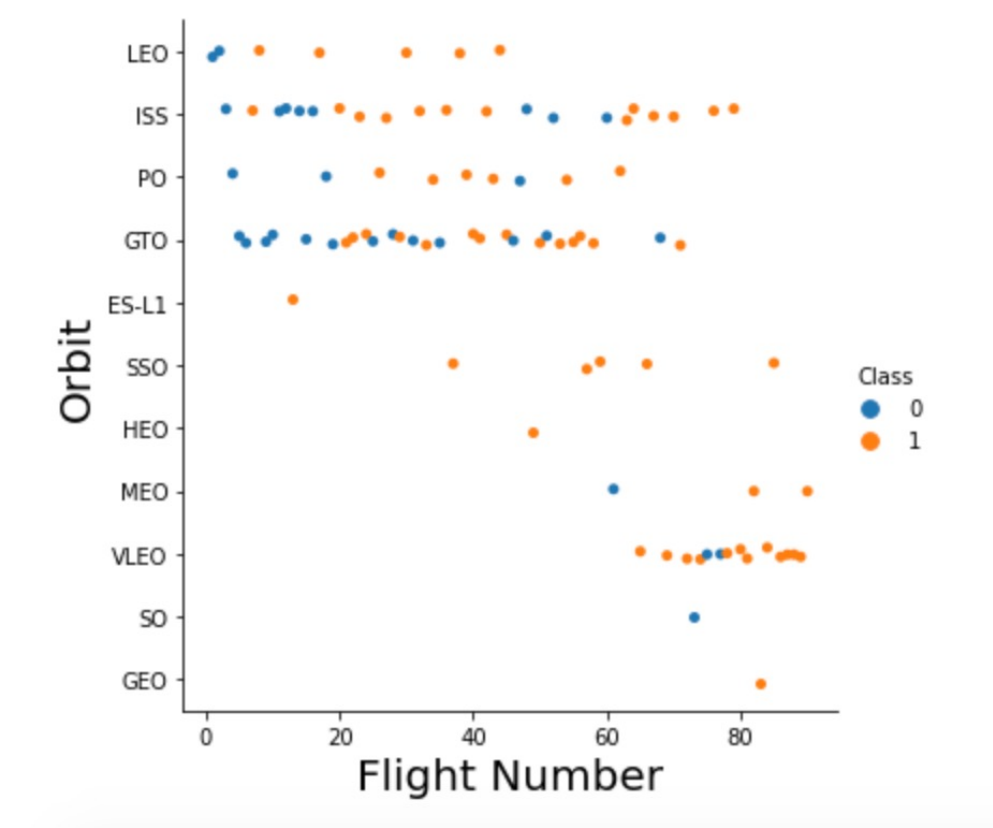


Figure 4: Scatter plot between Flight number and Orbit type

- The flight number launching to ISS has the widest range, followed by that to GTO, and PO, respectively
- The flight to LEO ranged in low to medium. number
- The flight number to SSO, MEO, and VLEO ranged in medium to high, with the narrower ranges respectively.
- There was only one launch to ES-L1, HEO, SO, and GEO; therefore, among those with 1.0 successful rate, the launch to SSO was the most promising.

Payload vs. Orbit Type

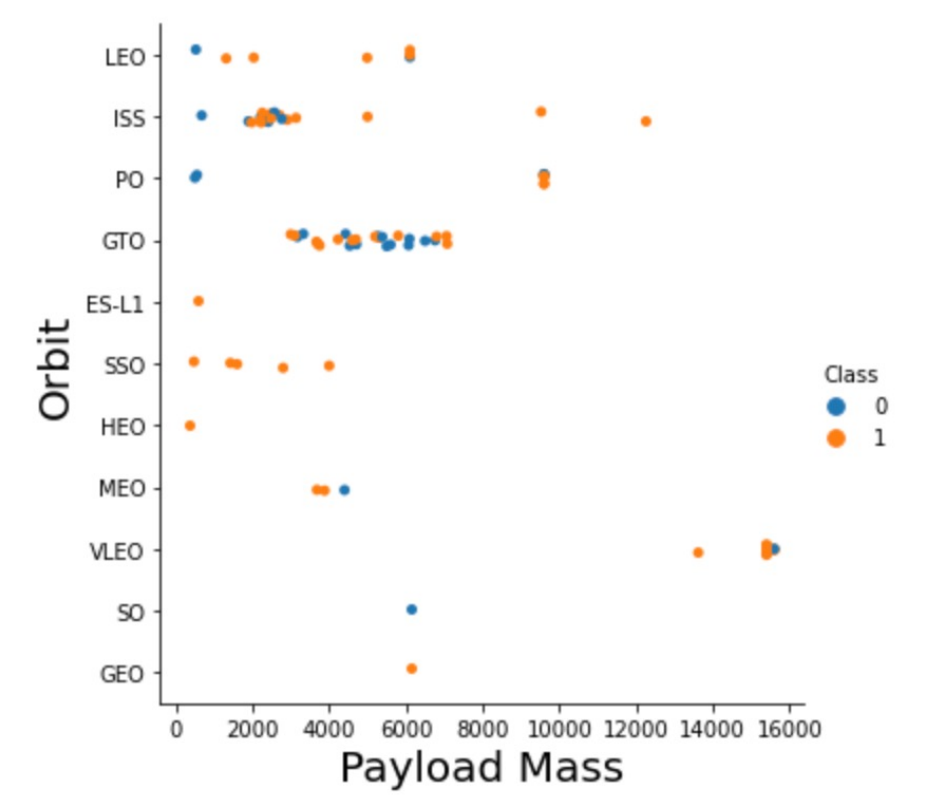


Figure 5: Scatter plot between Payload and Orbit type

- According to Figure 2, most of the launch had been with the light payload mass.
- Figure 5 shows that most of the heavy payload were the launch to VLEO orbit, while some others were that to ISS and PO.
- Launches to every other orbit types had been with light to medium payload mass.

Launch Success Yearly Trend

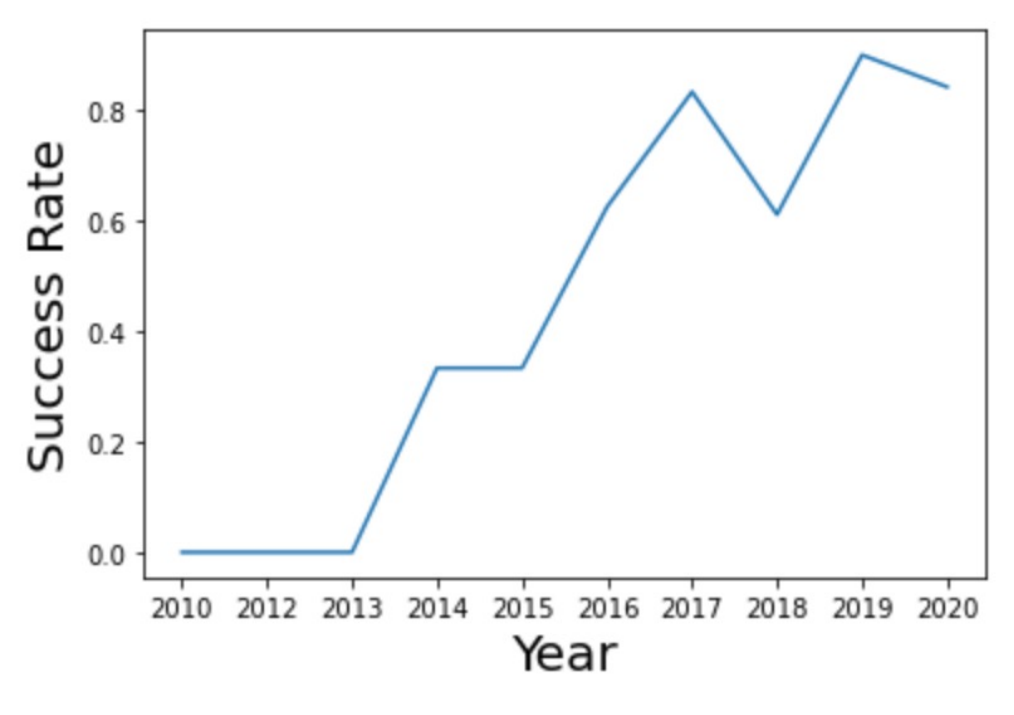


Figure 6: The launch success trend over years

- Overall, the success rate had been increasing from 2013 to 2020.
- It had been increasing to 0.8 in 2017, then it dropped to 0.6 in 2018
- With the increasing trend to the maximum success rate in 2019, it started dropping again.

All Launch Site Names

By performing this SQL query:

```
%sql select unique(Launch_SITE) from SPACEXDATASET;
```

We can know that there are 4 unique launch sites in the space mission:

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

Launch Site Names Begin with 'KSC'

By performing this SQL query:

```
%sql select * from SPACEXDATASET where LAUNCH_SITE LIKE '%KSC%' limit 5;
```

The 5 records which the launch site names begin with ‘KSC’ are displayed:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

By performing this SQL query:

```
%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)';
```

We can know that

45596 is the total payload mass carried by boosters launched by NASA (CRS).

Average Payload Mass by F9 v1.1

By performing this SQL query:

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXDATASET where BOOSTER_VERSION = 'F9 v1.1';
```

We can know that

2928

is the average payload mass carried by booster version F9 v1.1

Successful Drone Ship Landing with Payload between 4000 and 6000

By performing this SQL query:

```
%sql select BOOSTER_VERSION from SPACEXDATASET where LANDING__OUTCOME LIKE '%Success%' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

The booster names of the successful drone ships landing with payload between 4000 – 6000 can be displayed.

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 B4 B1043.1
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

Total Number of Successful and Failure Mission Outcomes

By performing this SQL query:

```
%sql SELECT MISSION_OUTCOME, count(*) from SPACEXDATASET group by MISSION_OUTCOME;
```

We can know that there were 100 successful mission outcome in total (one unclear payload status, and another 1 failure in flight.

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

By performing this SQL query:

```
%sql select BOOSTER_VERSION,PAYLOAD_MASS__KG_ from SPACEXDATASET where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET);
```

We can display a list of booster version which carried the maximum payload at 15600.

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

Section 4

Launch Sites Proximities Analysis

Launch sites marked on a map

Figure 7: The marks of all launch sites on a global map

- All 4 launch sites are located in the nearly equal latitude, around the coastal area (W – E) of the US.
- Three eastern sites are located near one another, which CCAFS LC-40 and CCAFS SLC-40 are almost at the same coordinate.

Outcomes marked at the launch sites

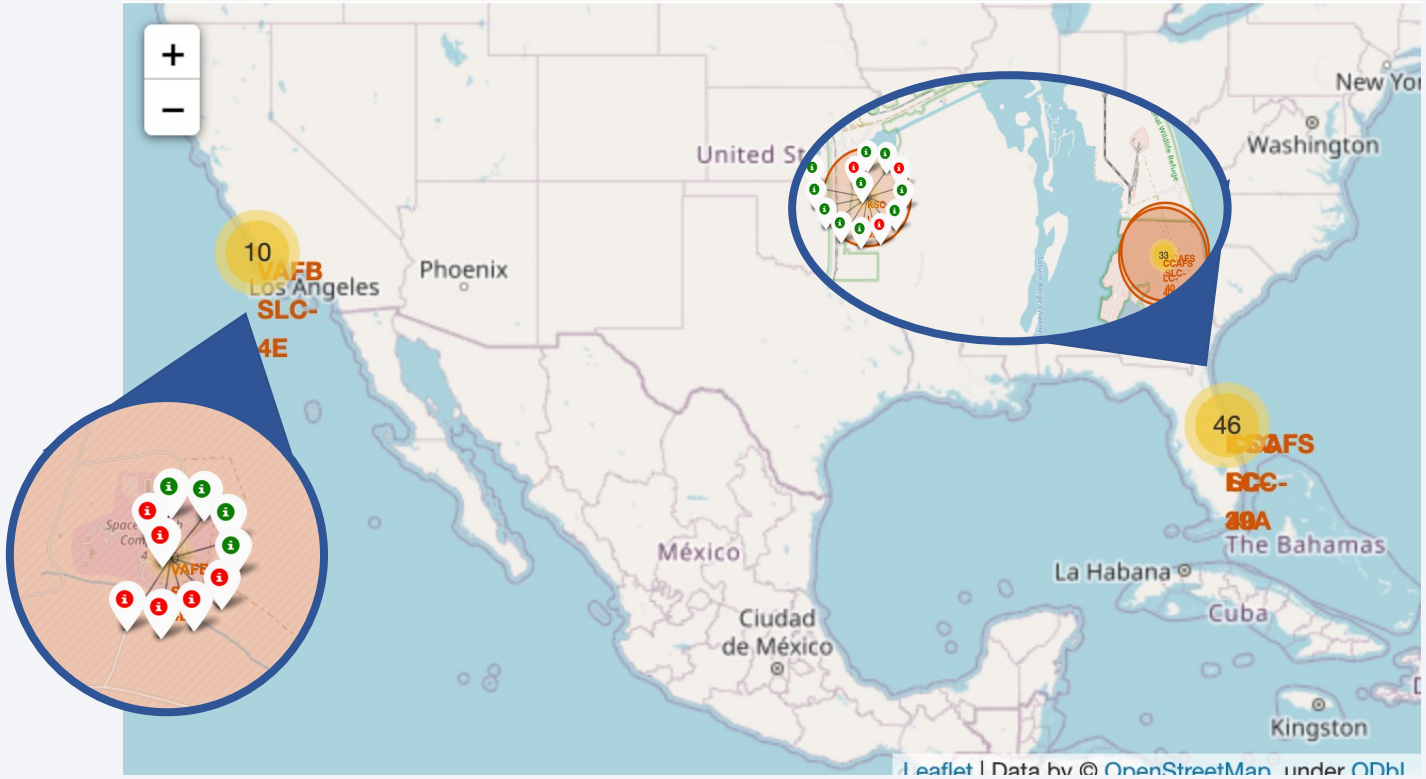


Figure 8: The marks of the outcomes at the launch sites on a global map (Green = Success; Red = Failed)

- **Figure 8** shows the total number of launch at the sites:
  - 10 at the west
  - 46 at the east (for three sites)
- **Table 1** shows the number of success/failed outcomes at each site regarding the counts of the colored markers.

Launch site	Success	Failed
VAFB SLC-4E	4	6
KSC LC-39A	10	3
CCAFS SLC-40	3	4
CCAFS LC-40	7	19

Table 1: Success/Failed outcomes at each site

Distance from a launch site to its proximities

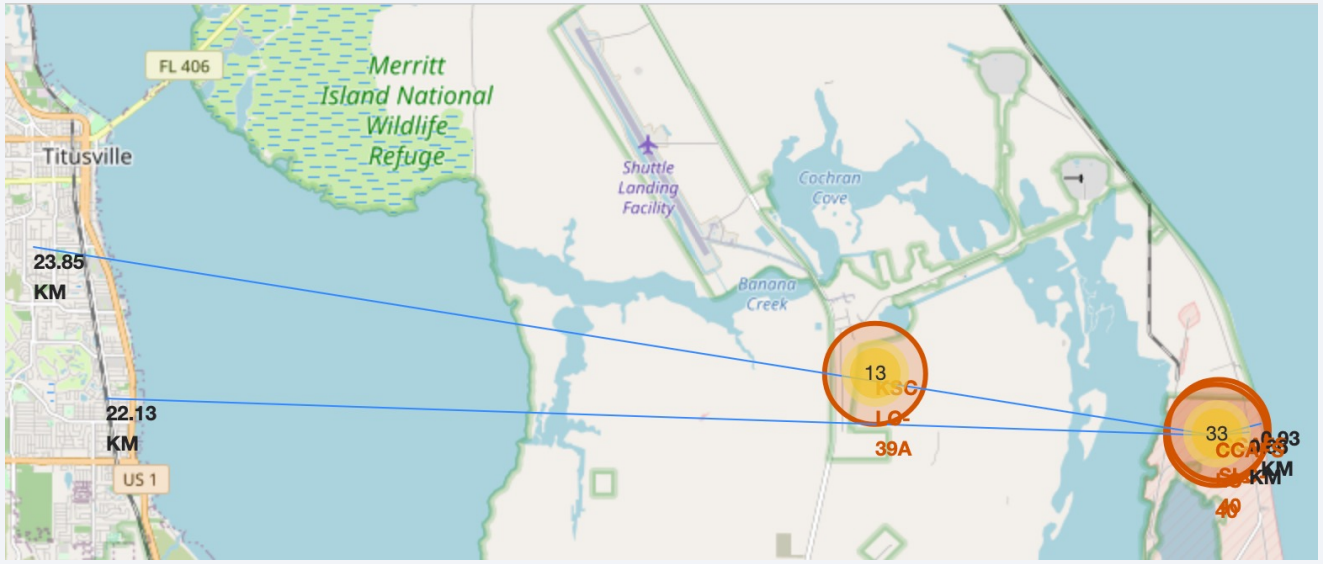
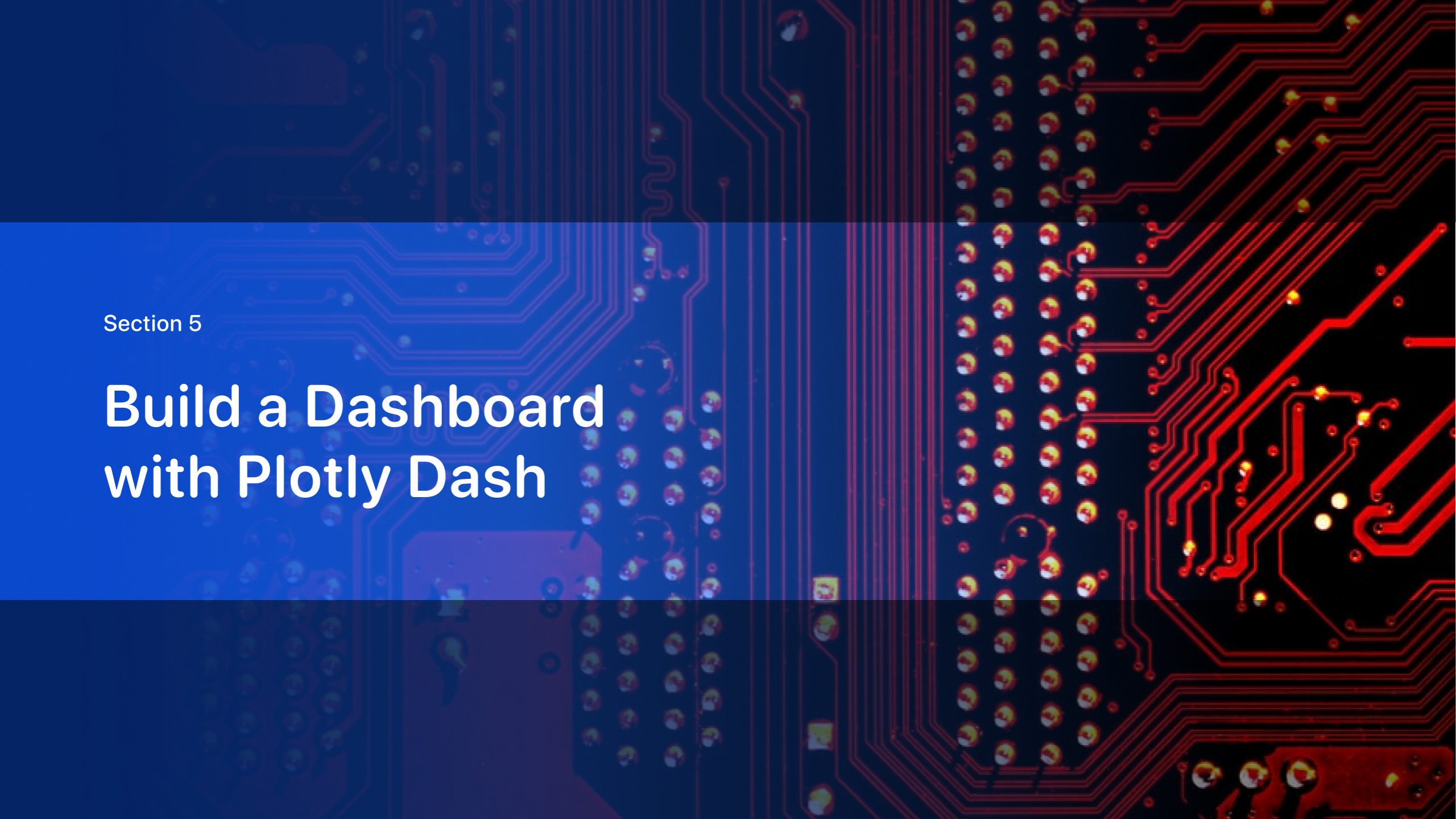


Figure 9: Distance from the CCAFS LC-40 launch site to its proximities

- **Figure 9** shows the marked distance from the CCAFS LC-40 to its nearest
  - Coastal line ~ 0.93 km
  - Railway ~ 22.85 km
  - Highway ~ 0.66 km
  - City ~ 22.13 km
- The results show that the launch site is typically located **near the coastal line.** Besides, it is **far from the city and railway, but not that far from the highway.**



Section 5

Build a Dashboard with Plotly Dash

Dashboard: Pie chart of the success launch for all sites

- Selecting *All sites* from the drop-down, **Figure 10** shows the pie chart of the success launch for all sites.

  - KSC LC-39A : 41.7%
  - CCAFS LC-40 : 29.2%
  - VAFB SLC-4E : 16.7%
  - CCAFS SLC-40 : 12.5%

- The launch at **KSC LC-39K** has the highest success rate, followed by **CCAFS LC-40** and **VAFB SLC-4E,** respectively, The launch at **CCAFS SLC-40** has the lowest success rate.

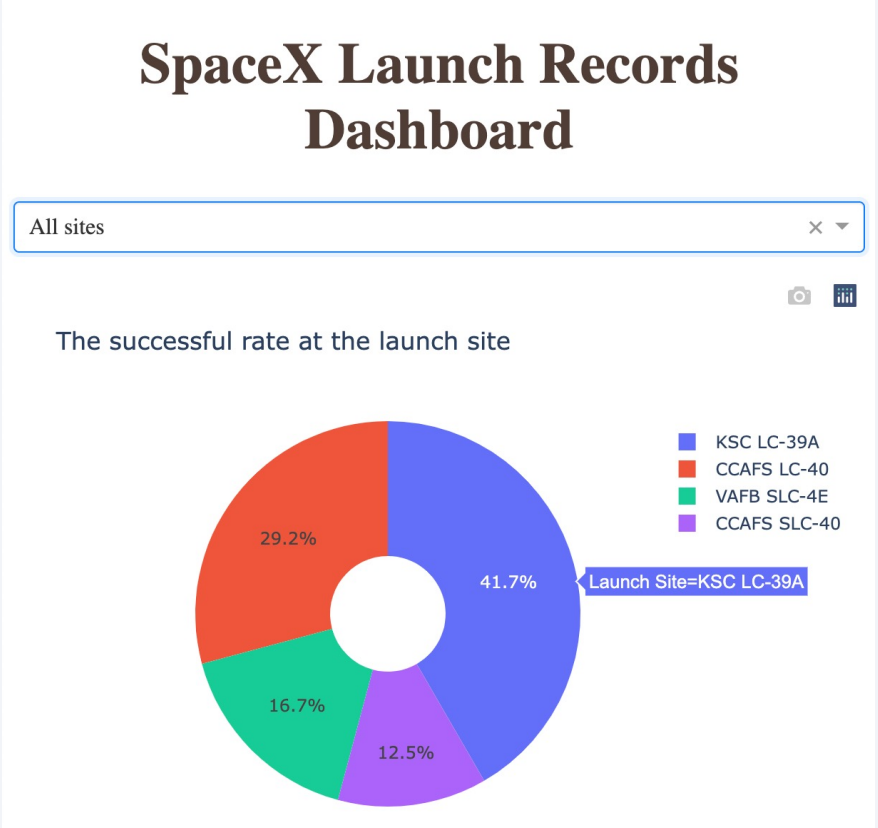


Figure 10: Pie chart of the success launch for all sites

Dashboard: Pie chart for the particular launch site (1)

- Selecting *KSC LC-39A* from the drop-down which has the highest success rate among 4 launch sites, **Figure 11** shows the pie chart of launching outcome at KSC LC-39A.

  - Success (Class = 1) : 76.9 %
  - Failed (Class = 0) : 23.1 %

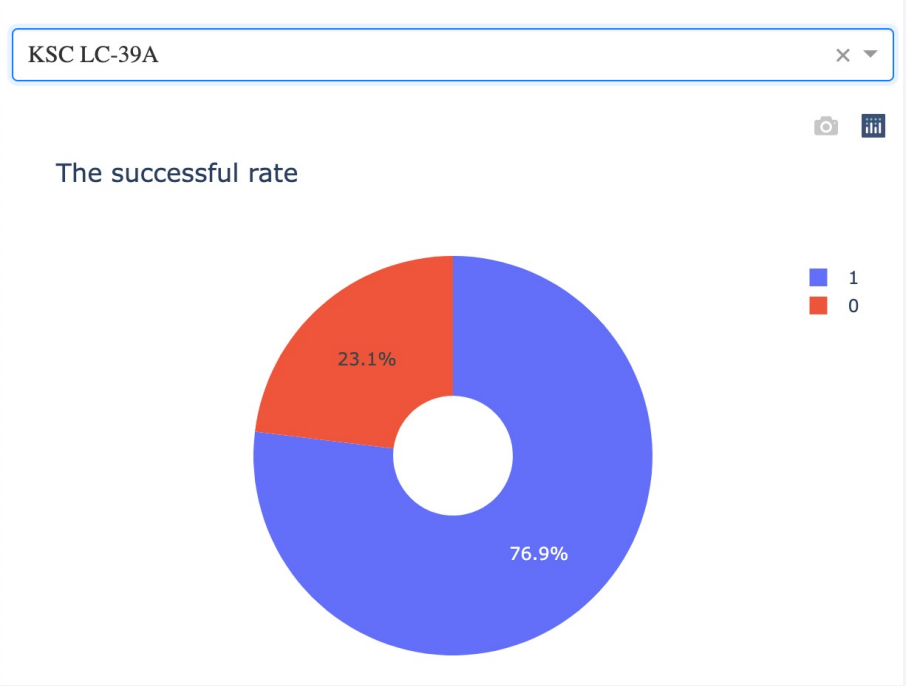


Figure 11: Pie chart of the launch at KSC LC-39A

Dashboard: Scatter plots of Payload vs Outcomes for all sites (1)

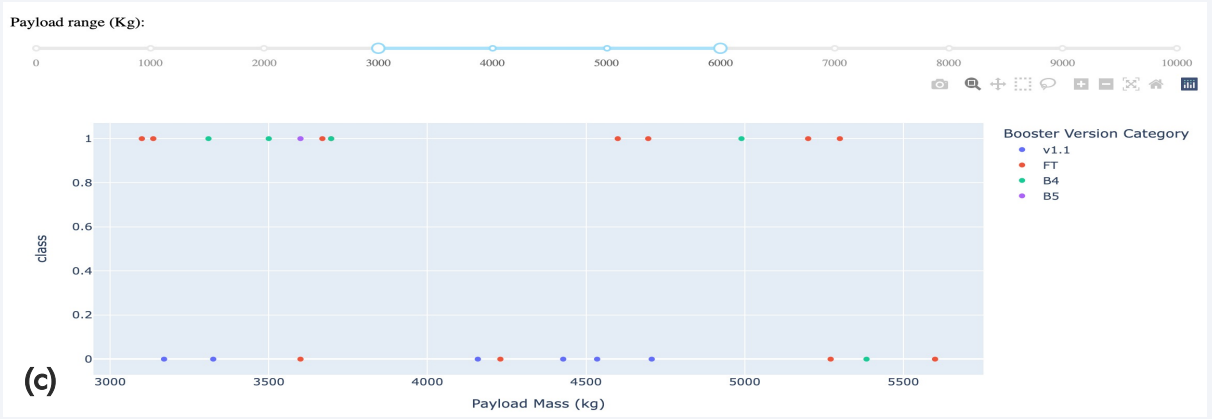


Figure 12: Scatter plots of Payload vs Outcomes for all sites in the different payload ranges
(a) 0 – 10,000 (b) 0 – 3,000 (c) 3,000 – 6,000 (d) 6,000 – 10,000

Dashboard: Scatter plots of Payload vs Outcomes for all sites (2)

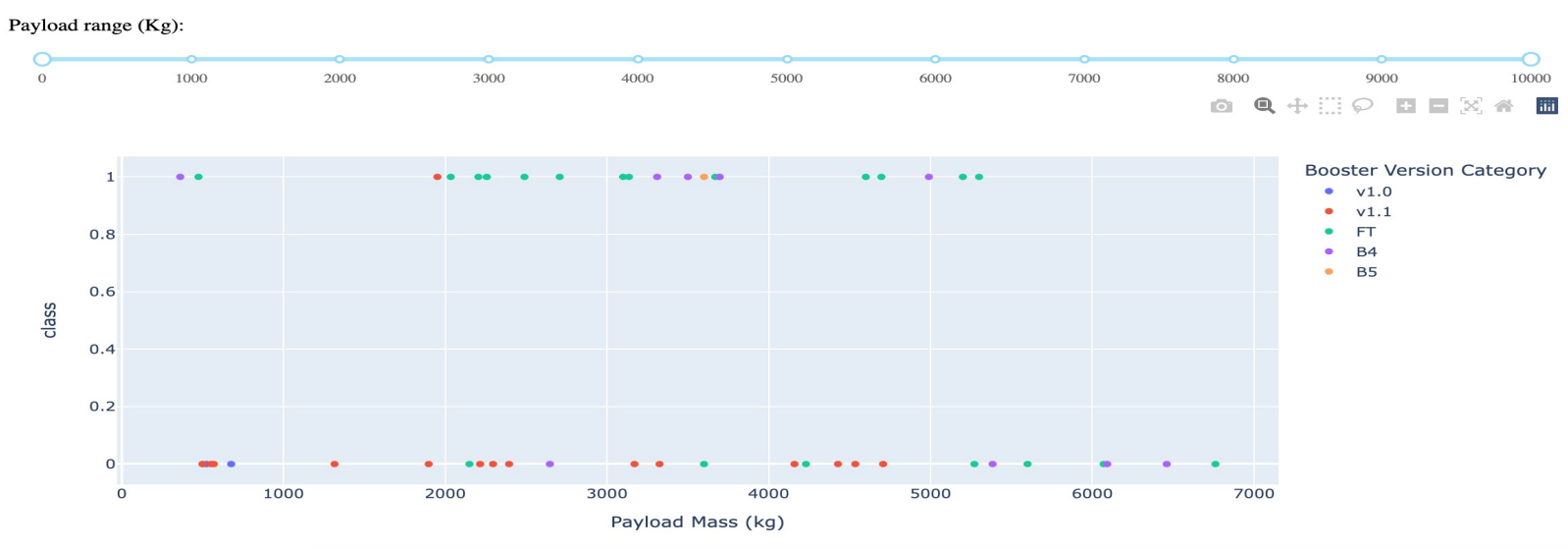


Figure 12(a) : Scatter plots of Payload vs Outcomes for all sites in the payload ranges 0 – 10,000 kg

- Figure 12(a) shows the scatter plot of payload vs outcomes for all sites in the whole payload range (0 – 10,000).
- The launches with **medium payload** (2,000 – 5,500) have the greatest success outcomes. Meanwhile, the launches with **light payload** (< 2,000) are slightly successful, but those with **the heavy payload** (> 5,500) are all failed.
- Most of the success outcomes are the launch with **FT Booster Version Category**, and some of them are with **B4**.

Dashboard: Scatter plots of Payload vs Outcomes for all sites (2)

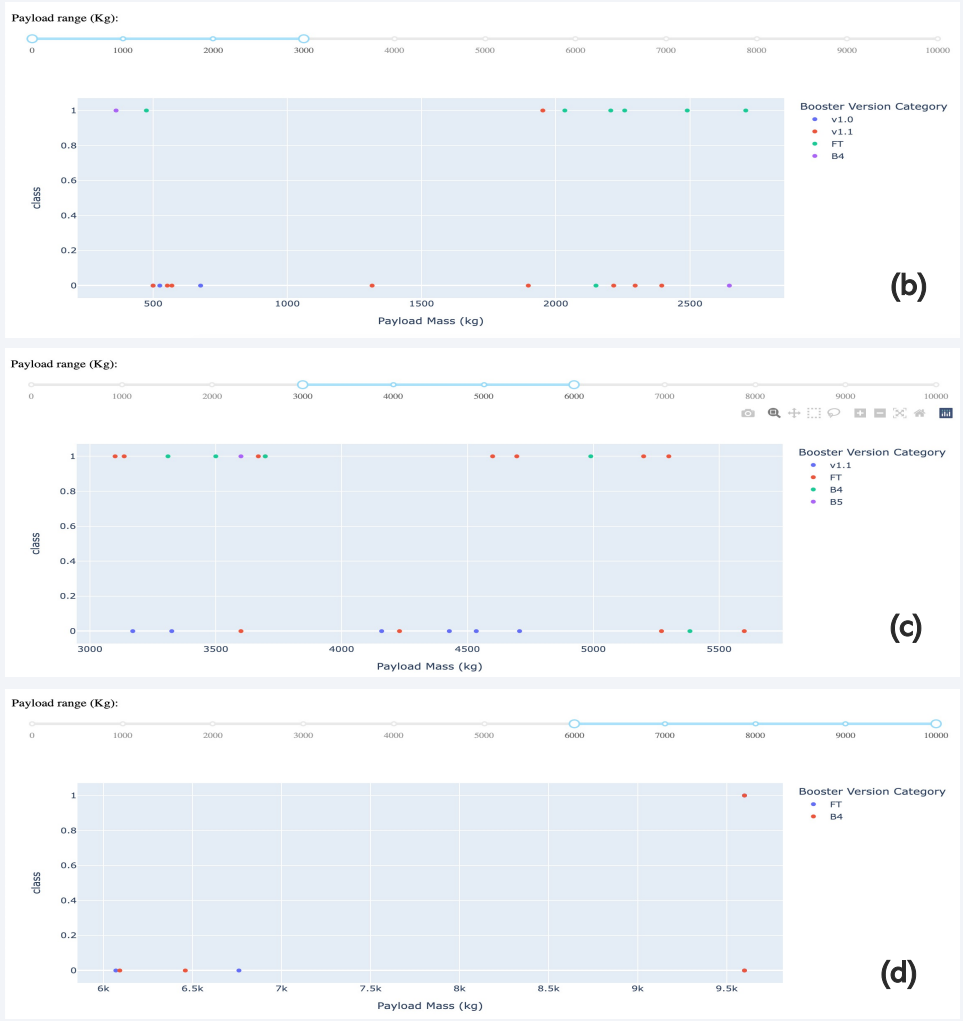


Figure 12: Scatter plots of Payload vs Outcomes for all sites at the particular payload ranges

- We can investigate the plot at a particular payload range by selecting from the slider.
- In the light payload range (0– 3,000 kg), the launches were successful only with the payload mass less than 500, and greater than 1,900. Most of them had FT booster version.
- In the medium payload range (3,000 – 6,000 kg), the launches were all failed in the range 3,800 – 4,500. The success outcomes were with various booster version.
- In the heavy payload range (6,000 – 10,000 kg), the launches were mostly failed. Launching with the same very heavy payload and B4 booster version, there were a success (only a success outcome in the range), and a fail.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

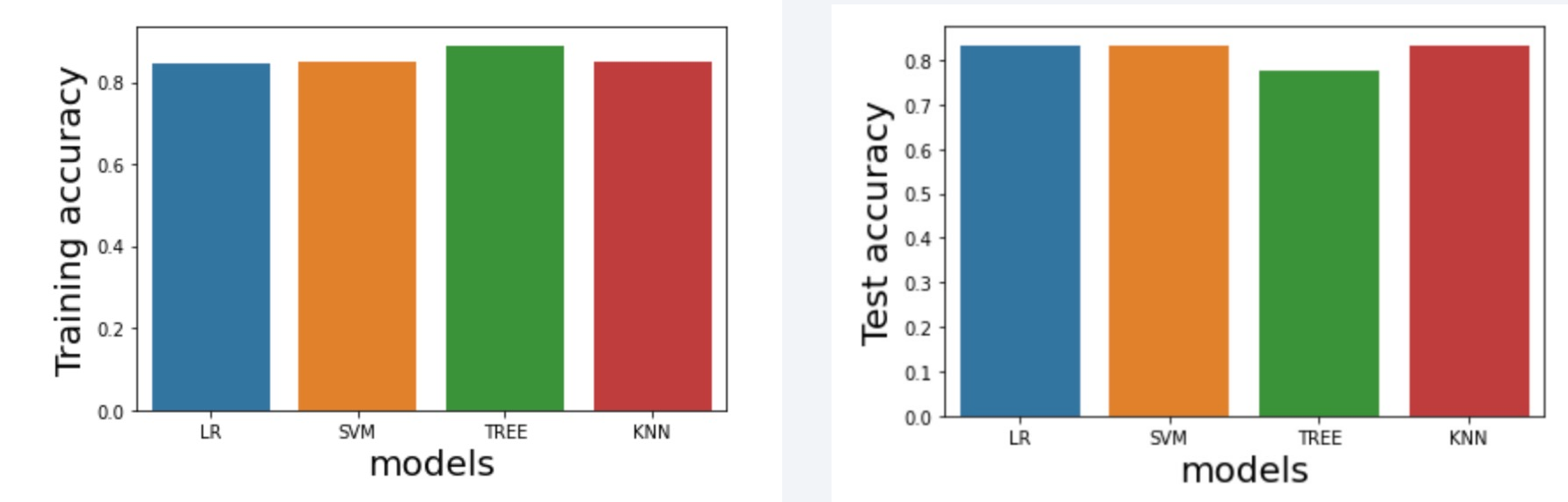


Figure 13: The accuracies of the classification models with the best parameters (left) Training accuracy (right) Test accuracy

- By GridSearchCV, we can find the best parameters among the provided choices of each models.
- Among the model with tuned hyperparameters, the **decision tree performs the best on training dataset.** However, it performs **the least efficiently on test dataset**, which means that decision tree is overfitting.

Confusion Matrix

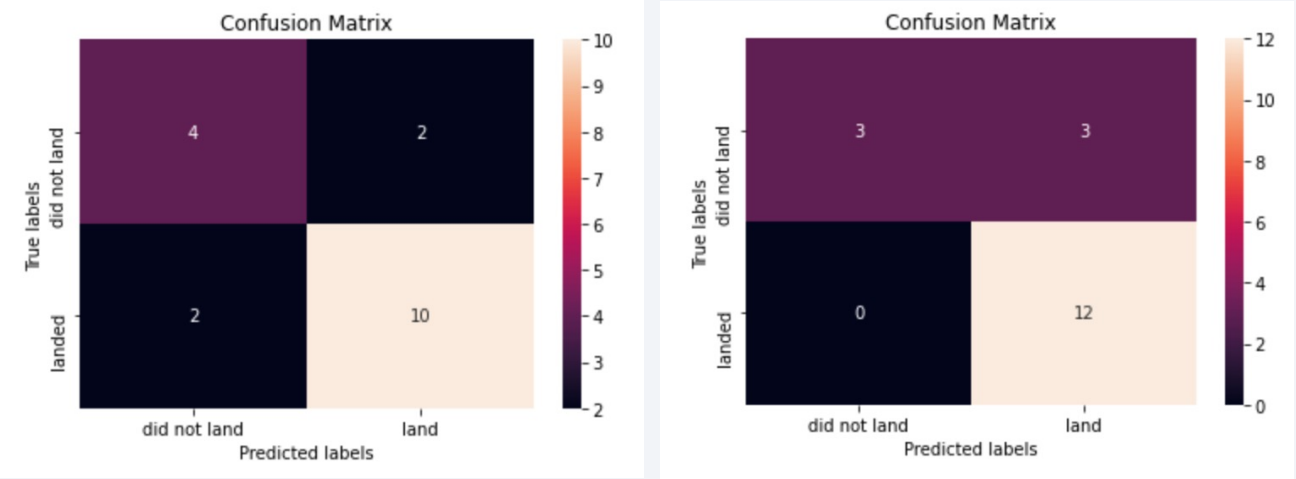


Figure 14: Confusion matrix
(left) Decision Tree (right) Logistic Regression, SVM, KNN

	Decision Tree	All other models
Accuracy	0.833	0.833
F1	0.833	0.889
Precision	0.833	0.8
Recall	0.833	1.0

Table 2: Performance of the models

- By performing the confusion matrix, all other models (Logistic regression, SVM, KNN) have the same performance measures.
- Decision tree has equal accuracy and greater precision to other models; however, it has lower f1-score and recall.
- Overall, all other models with tuned parameters are comparably efficient. They are more suitable for prediction than the decision tree.

Conclusions

- As for the 4 launch sites, all of them are located at the nearly equal latitude, which is near to the coastal line and highway, but far from the city and railway. Each one typically has different ranges of flight number and payload masses. One of the launch sites, named VAFB SLC 4E, has no heavy payload mass.
- The greater flight number, the higher success rate.
- The launch mostly has the light payload mass, which was found that the one with heavy payload mass was all failed. Meanwhile, the one with medium payload mass had the highest success rate.
- By exploring the success rate of the launch to each orbit, the result suggested that SSO is the most promising launching target.
- It was found that the success trend of the launch had been increasing from 2010 to 2020.
- From the interactive dashboard, we could investigate the launch at each particular site, payload range, and booster version category. It was found that most of the successful launch was with FT Booster version category.
- By modeling and tuning, it was found that all other machine learning models, but the decision tree, were comparably efficient in this project scenario. The tuned decision tree was overfitting.

Appendix

The Python codes, Notebook outputs, and dataset can be found in this Github url:

https://github.com/Puttaranun/IBM-DataScience-Capstone.git

Thank you!