IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Lim Fong

14/2/2024

Table of Content

Executive Summary

Summary of methodologies

1. Data Collection
2. Data Preprocessing
3. Exploratory Data Analysis (EDA) in Python
4. Exploratory Data Analysis (EDA) in SQL
5. Interactive map with Folium
6. Dashboard Development with Plotly Dash
7. Classification Machine Learning

Introduction

SpaceX stands out as the leading company in the era of commercial space exploration, significantly reducing the cost of space travel. The company features Falcon 9 rocket launches on its website, priced at 62 million dollars. In contrast, alternative providers charge over 165 million dollars for each launch. A substantial portion of the cost savings is attributed to SpaceX's innovative practice of reusing the initial rocket stage. Thus, the ability to predict the successful landing of the first stage becomes crucial in estimating the overall launch cost. Leveraging public information and advanced machine learning models, we aim to forecast whether SpaceX will opt to reuse the first stage in upcoming launches.

Section 1

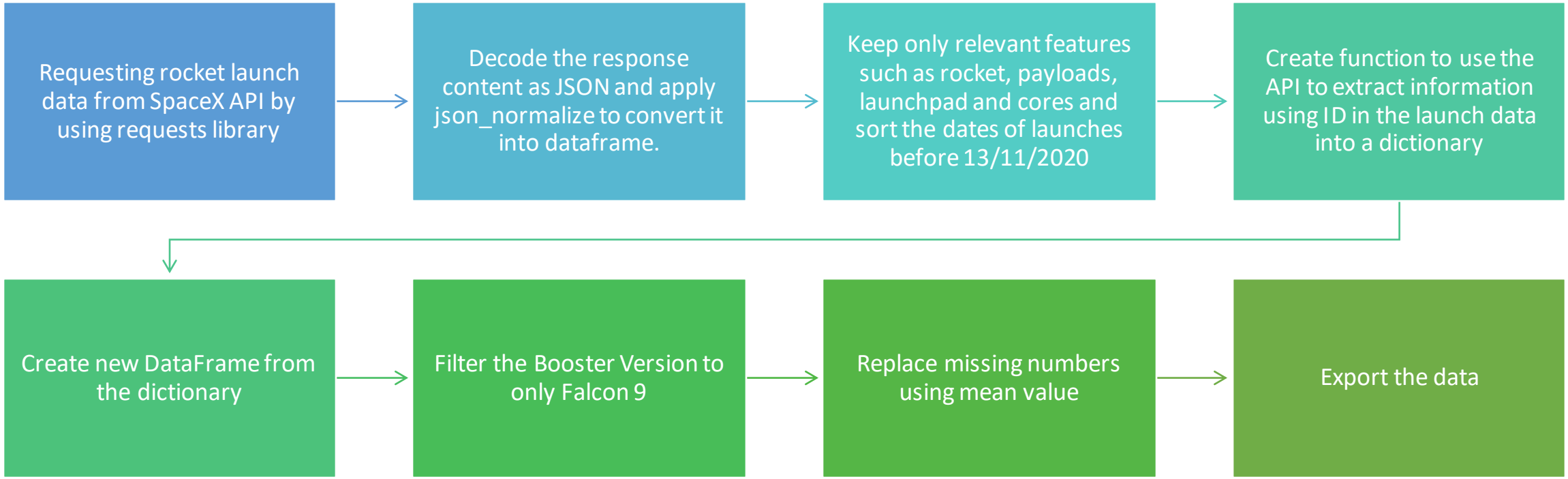
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - By using SpaceX API
  - By web-scraping
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

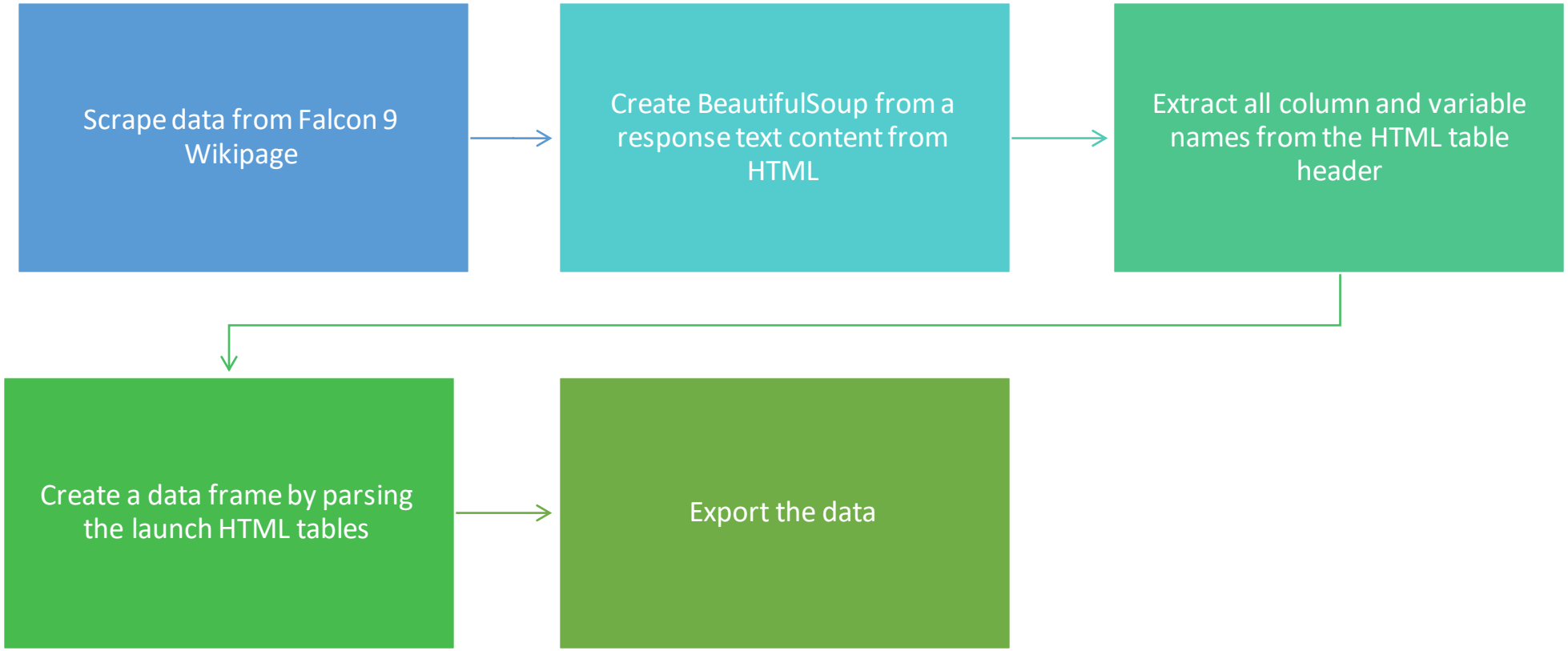
Data Collection – SpaceX API



https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/Data%20Collection%20%E2%80%93%20SpaceX%20API.ipynb

Create a BeautifulSoup object from the HTML response

Data Collection – Scraping



https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/Data%20Collection%20%E2%80%93%20Web%20Scraping.ipynb

Data Wrangling

1) Check for any missing values and the if the data type is correct
2) Identifying the total number of attempt each launch site has launches
3) Identifying the number and occurrence of each orbit
4) Calculate the number of possible outcomes of the orbits
5) Create a variable to group all the variables where the second stage did not land successfully
6) Create a function to convert the variable to 0 if the first stage did not land successfully and variable to 1 if the first stage land successfully under a new column called Class
7) Save and export the dataset

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/Data%20Collection%20%E2%80%93%20Web%20Scraping.ipynb

EDA with Data Visualization

Data Visualization

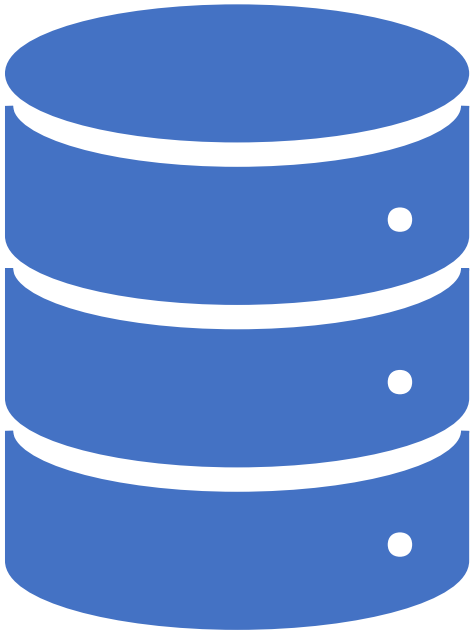
- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit Type vs. Success Rate
- Flight Number vs. Orbit Type
- Payload Mass vs Orbit Type
- Success Rate Yearly Trend

Explanation

- We use scatter plot to show if there is correlation between two variables and the success rate
- Bar plot to show the categories features with the success rate.
- Line chart to analysis if there is any pattern over time.

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/EDA%20with%20Visualization.ipynb

EDA with SQL



- 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. Use a subquery
- 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 landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- We added markers, circle and Pop-up label for each launch site using the Latitude and Longitude on the site map
- To improve the visualisation, we added Marker Cluster to simplify the map that contains multiple markers with similar coordinates
- We added two colours on the Marker's icon to indicate if the launch was successful (green) or failed (red) to determine which launch site has a higher chance of success rate.
- Next, we calculate the distance between the launch site, KSC-LC 39A to the nearest coastline and distance between the launch site, CCAFS to the nearest city, Melbourne.

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Created an interactive dashboard with customizable charts and graphs using dropdown menus and slider bars.
- Added a pie chart depicting the percentage of successful launches versus unsuccessful launches.
- Users can customize the pie chart by selecting different launch sites from the dropdown menu.
- Integrated a scatter plot illustrating the correlation between payload mass and class, with color labels for booster version categories.
-Implemented a slider bar for adjusting the minimal payload mass.
- The scatter plot can be adjusted based on individual launch sites using the dropdown menu.

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/spacex_dash_app.py

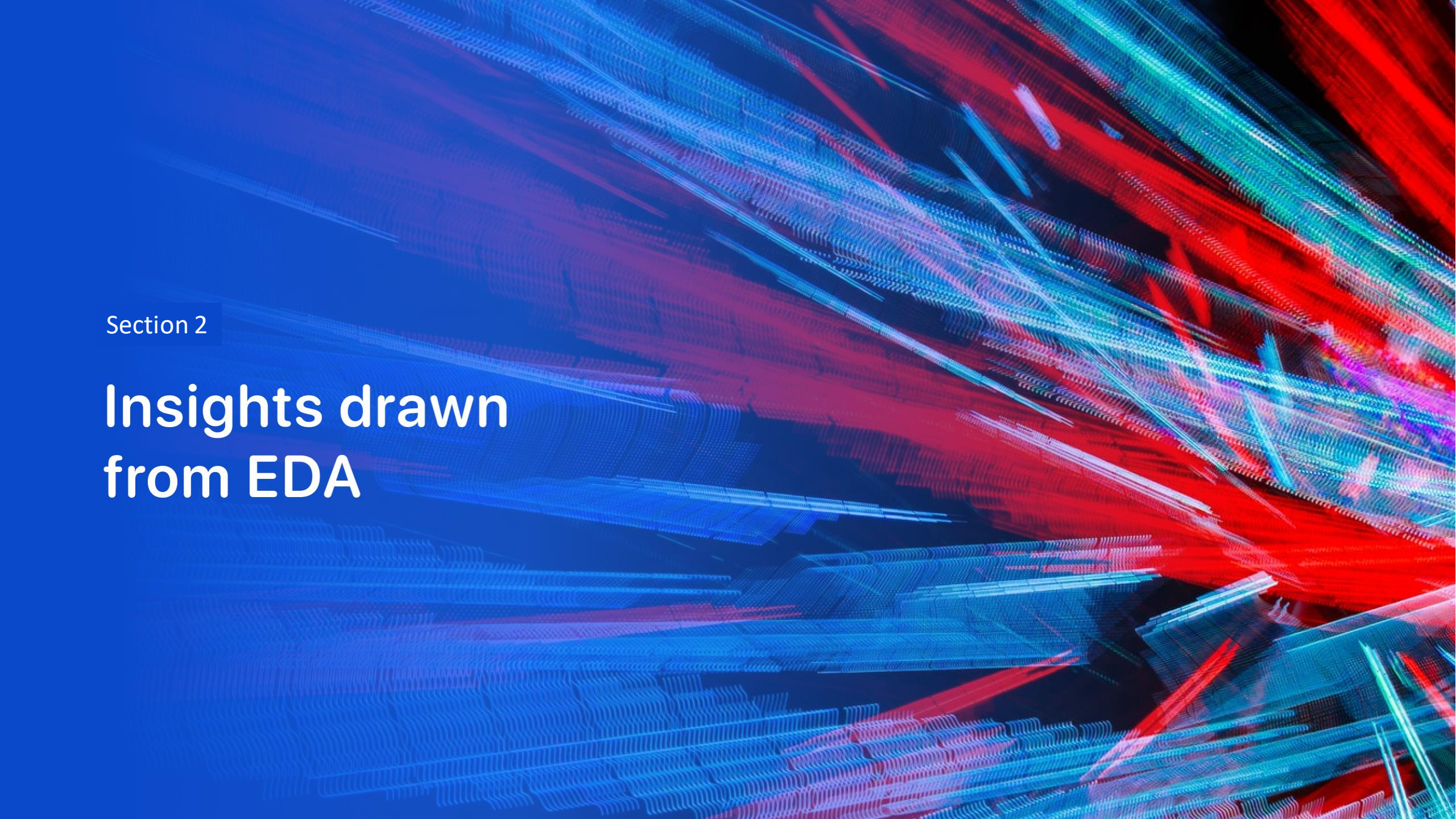
Predictive Analysis (Classification)

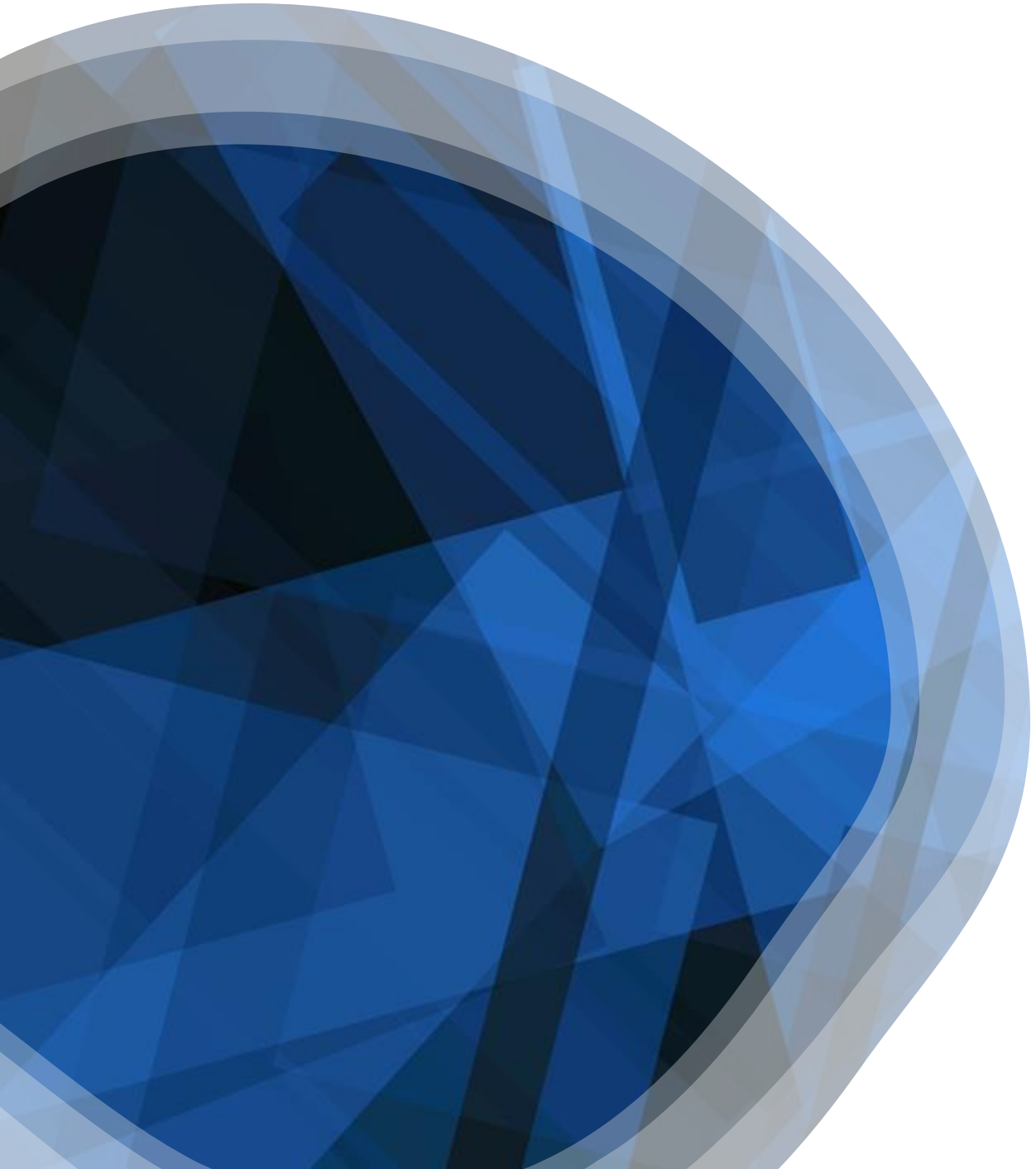
- Convert label to numpy and transform the feature variable using Standard Scaler so that the range is -1 to 1
- Using train_test_split to split the data in 80% training and 20% testing and set random state any number
- Create GridSearchCV and set cross validation to 10 to find the best parameter
- Apply GridSearchCV to all the model (Logistics Regression, SVM, Decision Tree and K-Nearest Neighbour) using the training dataset
- Calculate the accuracy and F1 score and the testing dataset for all the model with a confusion model
- Now, we use all the dataset to train and plot the best model using a bar chart

https://github.com/AsherTeo/IBM_applied_data_science_capstone/blob/main/Machine%20Learning.ipynb

Section 2

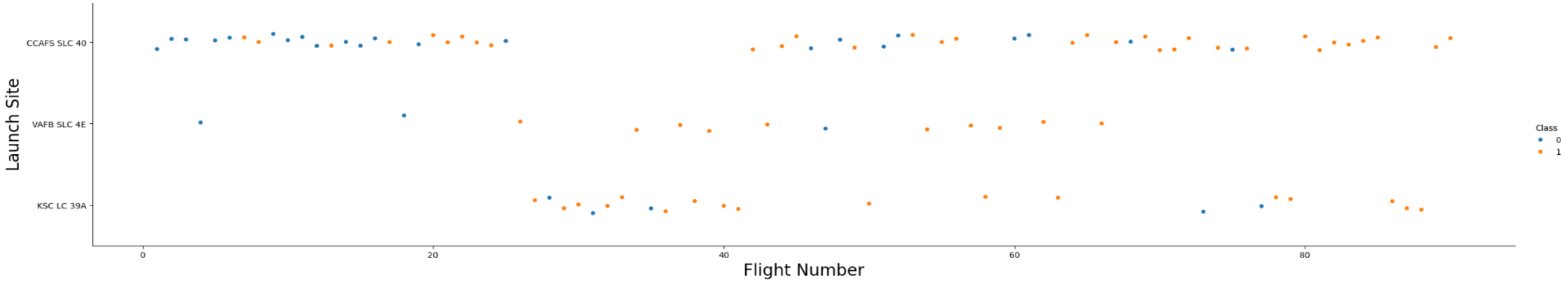
Insights drawn from EDA





Exploratory Data Analysis (EDA) with Python

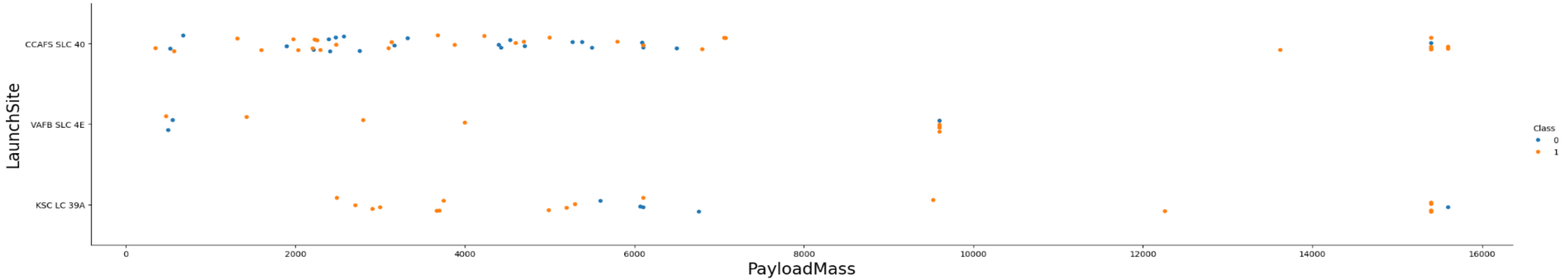
Flight Number vs. Launch Site



Explanation

- Flight Number indicate the continuous attempt.
- There is 100% success rate when the flight number attempt is over 80
- Only 40% success rate when the flight number attempt is below around 25
- VAFB SLC 4E has a higher chance of success rate than the other launch site
- The success rate increase when the flight number attempt is over 50

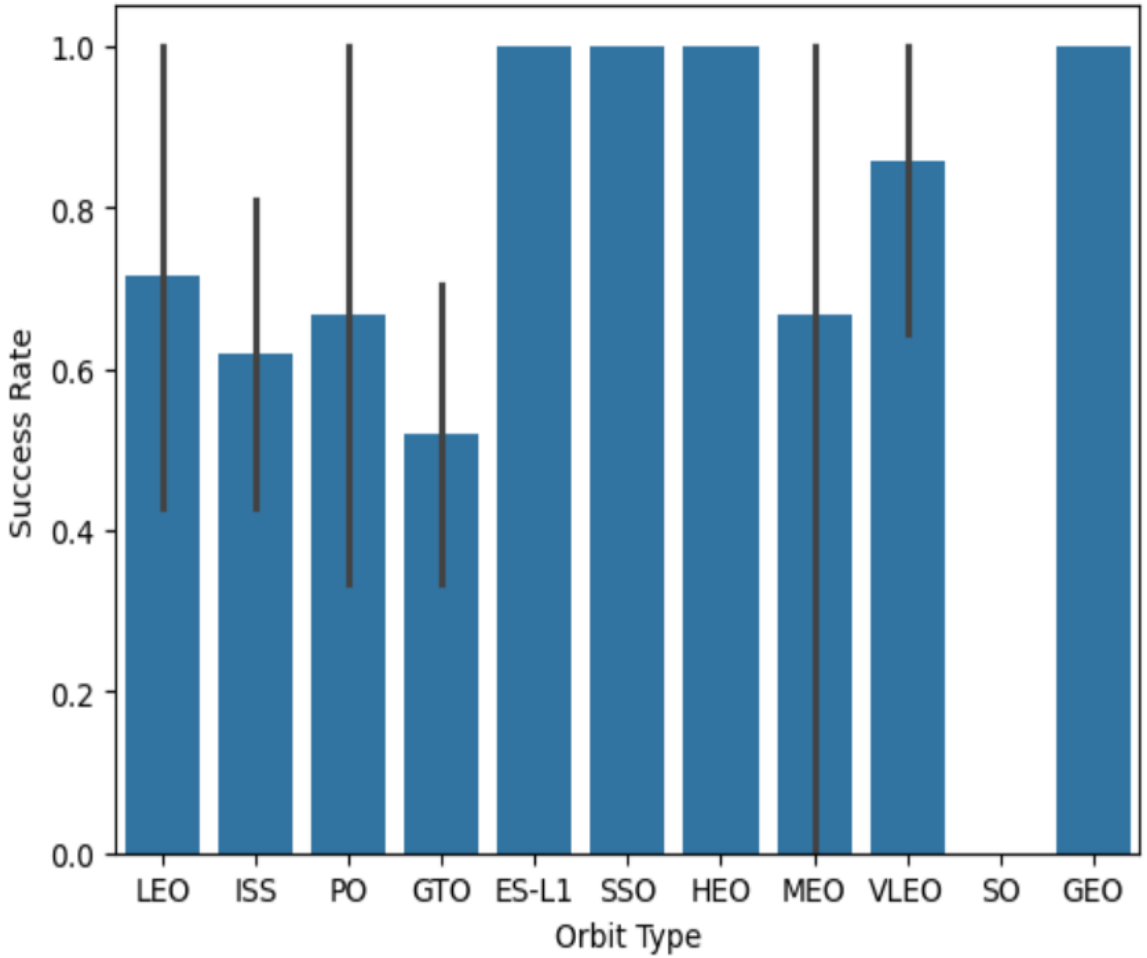
Payload vs. Launch Site



Explanation

- The success rate increase when the Pay load mass is above 8000 Kg
- KSC LC 39A has a 100% success rate from the range of 2000 Kg to around 5500 Kg
- KSC LC 39A has higher chance of failing when the range is between 5800 Kg to 7000 Kg
- We need more data for VAFB SLC 4E in regard of the Payload vs Launch Site
- My hypothesis is that VAFB SLC 4E requires less mass compared to other launch sites, while CCAF SLC 40 requires much more mass to increase the success rate. For KSC LC 39A, we need to have more data in the range of 6000 to 10000 Kg to make any accurate assumption. For instance, the ideal mass is the range of 2000 to 5500 KG for KSC LC 39A launch site

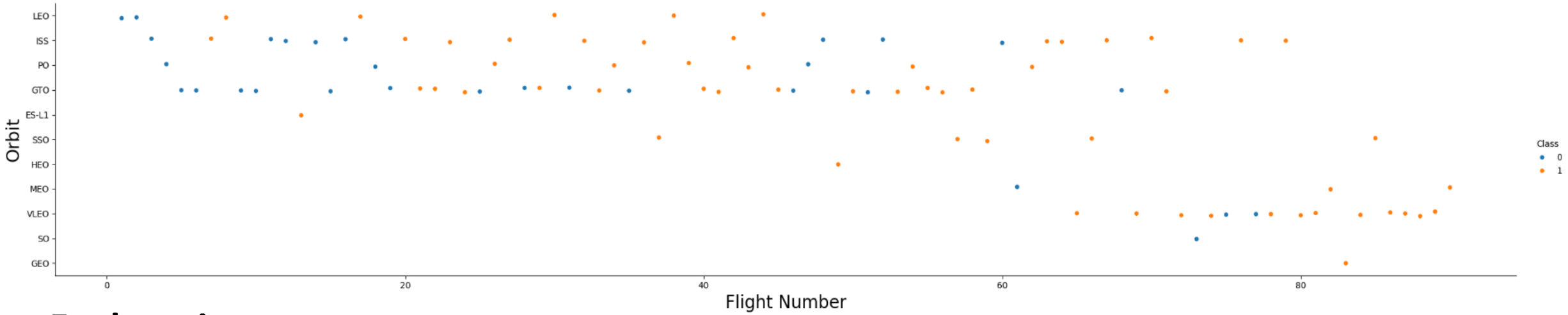
Success Rate vs. Orbit Type



Explanation

- The success rate of ES-L1, SSO, HEO and GEO is 100%
- SO has 0% success rate
- The other Orbit Type has the range of 50 to 85%
- My hypothesis is that all Orbit Types, except for SO, have a success rate exceeding 50%. Therefore, it is advisable to avoid utilizing SO. Furthermore, our next step is to examine whether most of the failures are associated with SO.

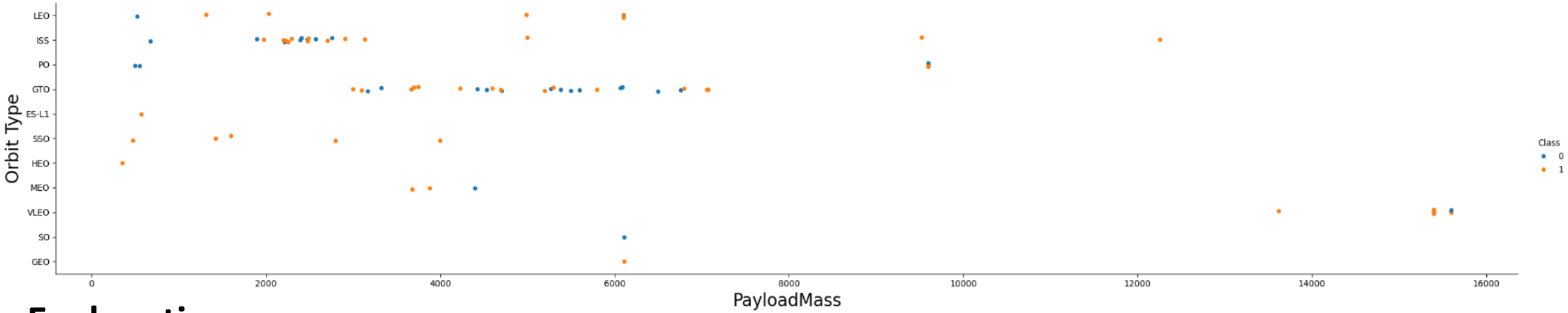
Flight Number vs. Orbit Type



Explanation

- It is not accurate to avoid SO even though has 0% success rate because it only contains one data point
- Like SO, ES-L1 , GEO and HEO also contains one data point which is not reliable to make any assumption
- VLEO has the highest success rate 85 % with 12 success out of 14 attempts which Flight Number attempt more than 65
- There are 100% success rate for LEO, ISS, PO when the Flight Number attempt is between the range of 25 to 45
- Additionally, VELO, MEO, and HEO demonstrate a 100% success rate when the Flight Number attempts surpass 80.
- My hypothesis suggests that there is an increase in success rate with higher flight numbers. However, it's essential to acknowledge that each orbit type has its own ideal flight number attempt

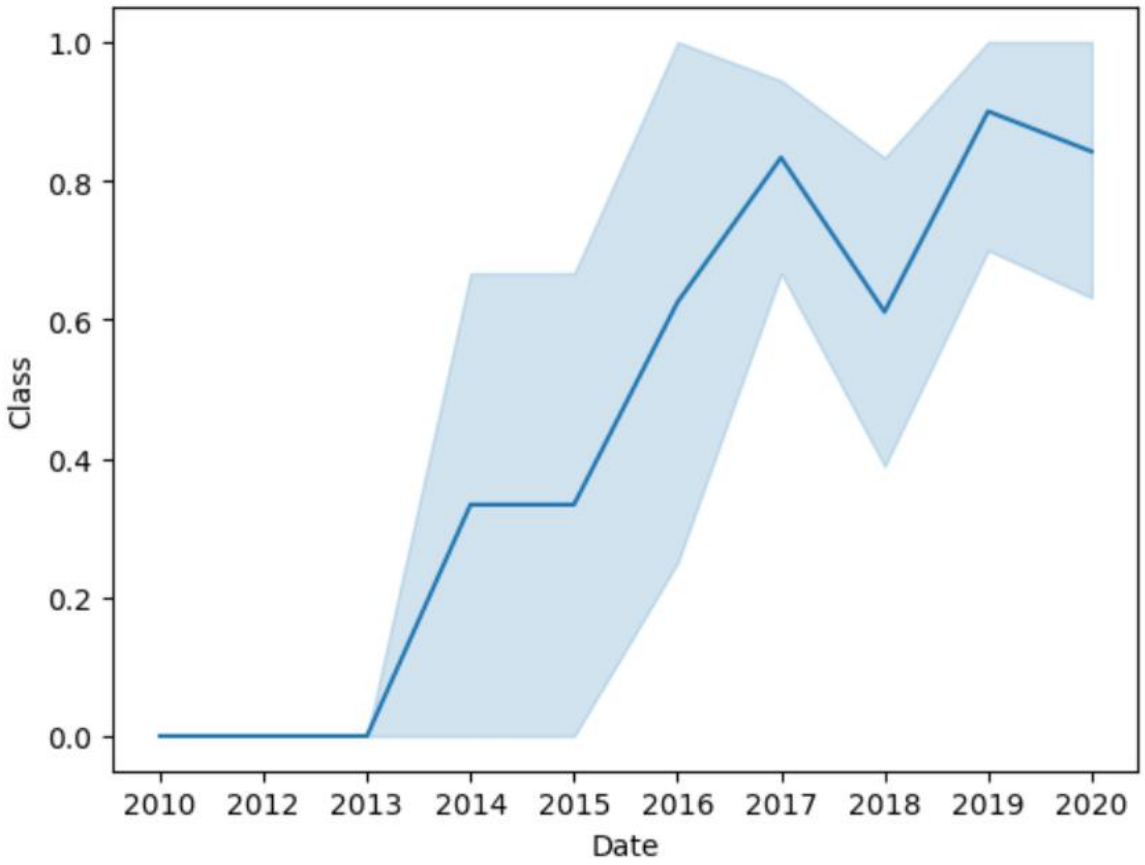
Payload vs. Orbit Type



Explanation

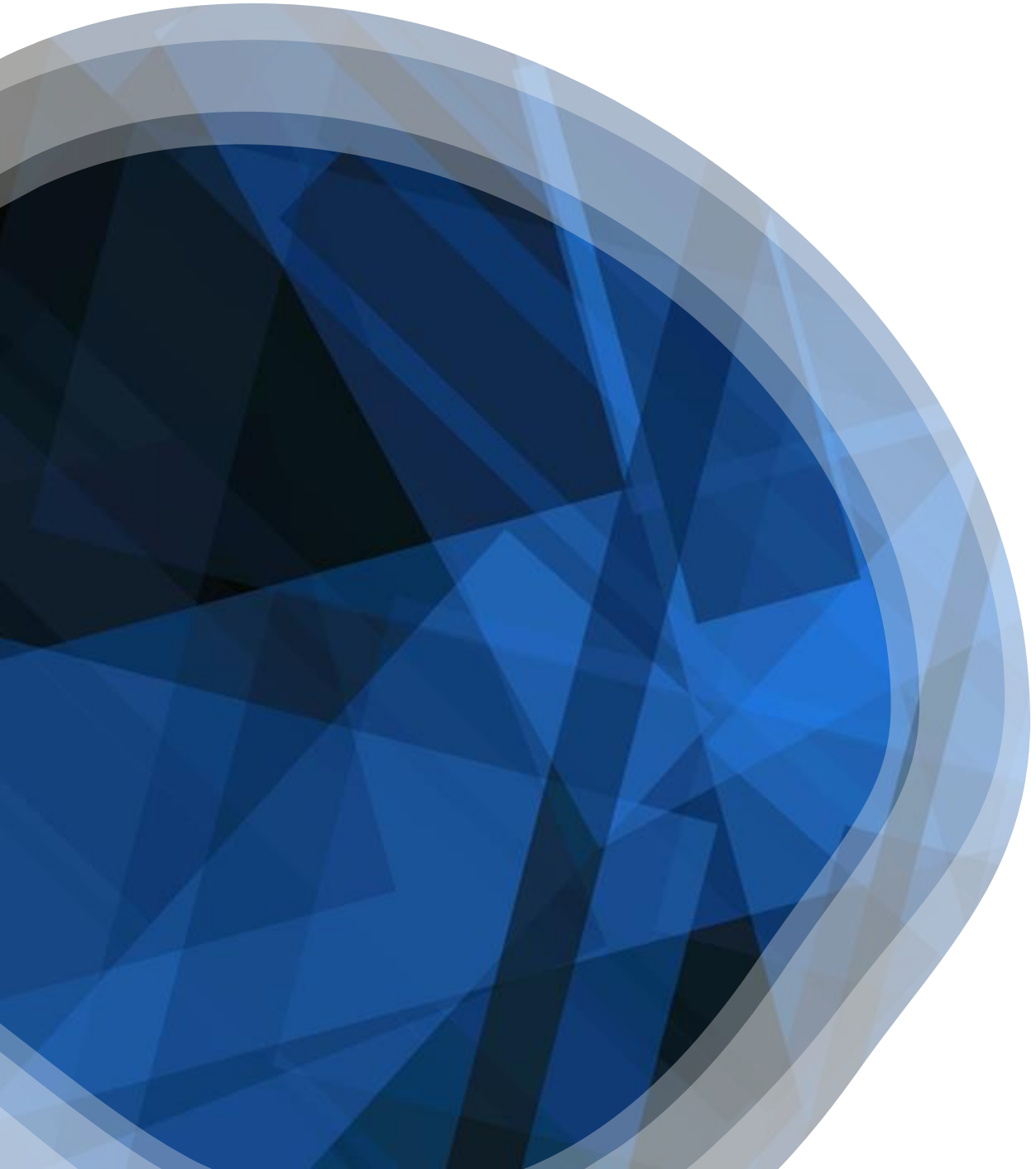
- There is an increase of success rate for ISS when the Pay Load Mass is above 3000 Kg
- For GTO, there are a decrease of success rate when the Pay Load Mass is above 5000 Kg
- 100% success rate for ES-L1, SSO, HEO and MEO when the Pay Load Mass in the range of 500 – 4000 Kg
- Due to insufficient data across the distribution in Pay Load Mass for the type of Orbit, my hypothesis suggests that there is no or close to no correlation between Pay Load mass and Orbit Type

Launch Success Yearly Trend



Explanation

- Overall, there is an increase in success rate since 2013.



Exploratory Data Analysis (EDA) with SQL

All Launch Site Names

Query

```
%sql select distinct Launch_Site from SPACEXTABLE
```

Result

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

Explanation

- We identified different unique launch site in the dataset so that we can narrow it down using map visualization.

- There are 4 launch sites which are CCAFS LC -40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Query

```
%sql select * from SPACEXTABLE where Launch_Site LIKE 'CCA%' limit 5
```

Result

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Explanation

- We extracted the top five launch site beginning with ‘CCA’

Total Payload Mass

Query

```
%sql select SUM(PAYLOAD_MASS__KG_) as Total_Payload_KG from SPACEXTABLE
where Customer = 'NASA (CRS)'
```

Result

Total_Payload_KG
45596

Explanation

- We calculate the total payload carried by boosters NASA which is 45596 Kg

Average Payload Mass by F9 v1.1

Query

```
%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

Result

AVG(PAYLOAD_MASS__KG_)
2928.4

Explanation

- We find the average pay load mass KG for Booster Version is F9 v1.1

First Successful Ground Landing Date

Query

```
%sql select Min(Date) from SPACEXTABLE
where Landing_Outcome LIKE '%Success%'
```

Result

Min(Date)
2015-12-22

Explanation

- The dates of the first successful landing outcome on ground pad is on the 12-22-2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Query

```
%sql select Booster_Version from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
AND (PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000)
```

Result

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

Explanation

- We list the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

Query

```
%sql select Mission_Outcome, count(*) from SPACEXTABLE
group by Mission_Outcome
```

Result

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

Explanation

- We calculate the total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Query

```
%sql select Booster_Version from SPACEXTABLE
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

Result

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

Explanation

- We list the names of the booster which have carried the maximum payload mass

2015 Launch Records

Query

```
%sql select substr(Date,6,2) as Month, Date, Booster_Version, Landing_Outcome, Launch_Site from SPACEXTABLE
where substr(Date,0,5) = '2015' AND Landing_Outcome = 'Failure (drone ship)'
```

Result

Month	Date	Booster_Version	Landing_Outcome	Launch_Site
01	2015-01-10	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	2015-04-14	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

Explanation

- We list the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Apparently, there are two failures. One in the January and one in the April

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

Query

```
%sql select Landing_Outcome, count(*) as Outcome_count from SPACEXTABLE
where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by Outcome_count desc
```

Result

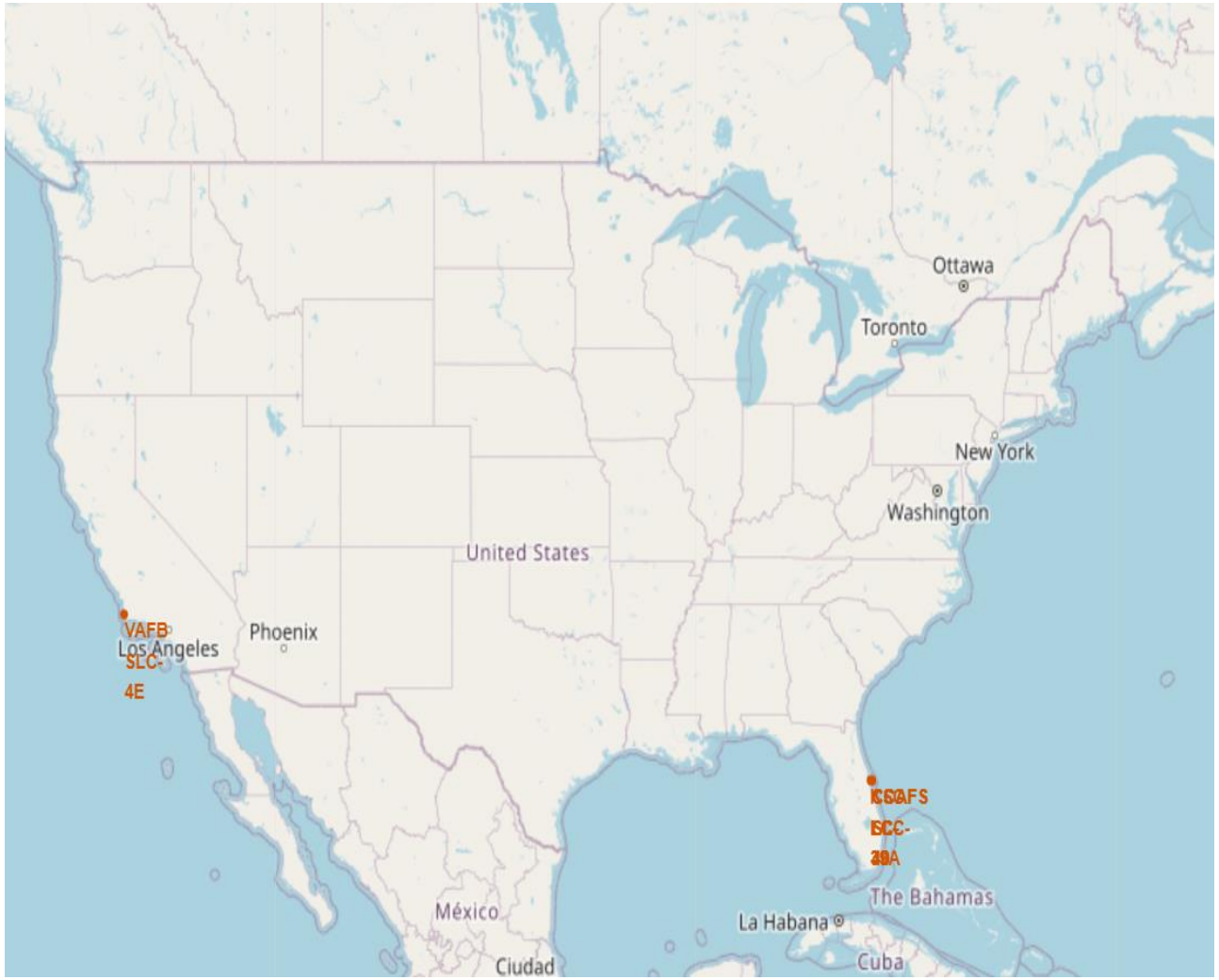
Landing_Outcome	Outcome_count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Explanation

- We rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

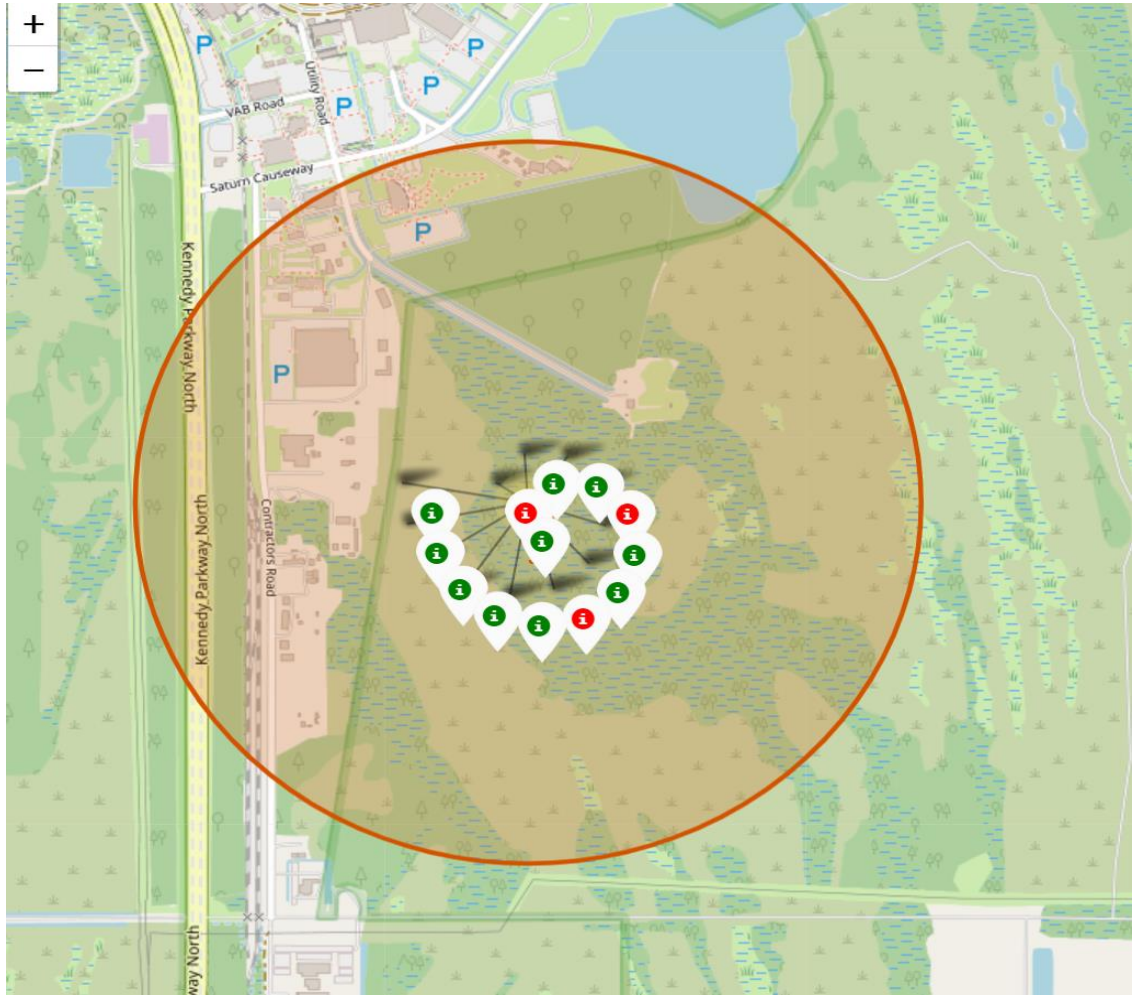

Section 3
Launch Sites
Proximities Analysis

Launch Site using Markers



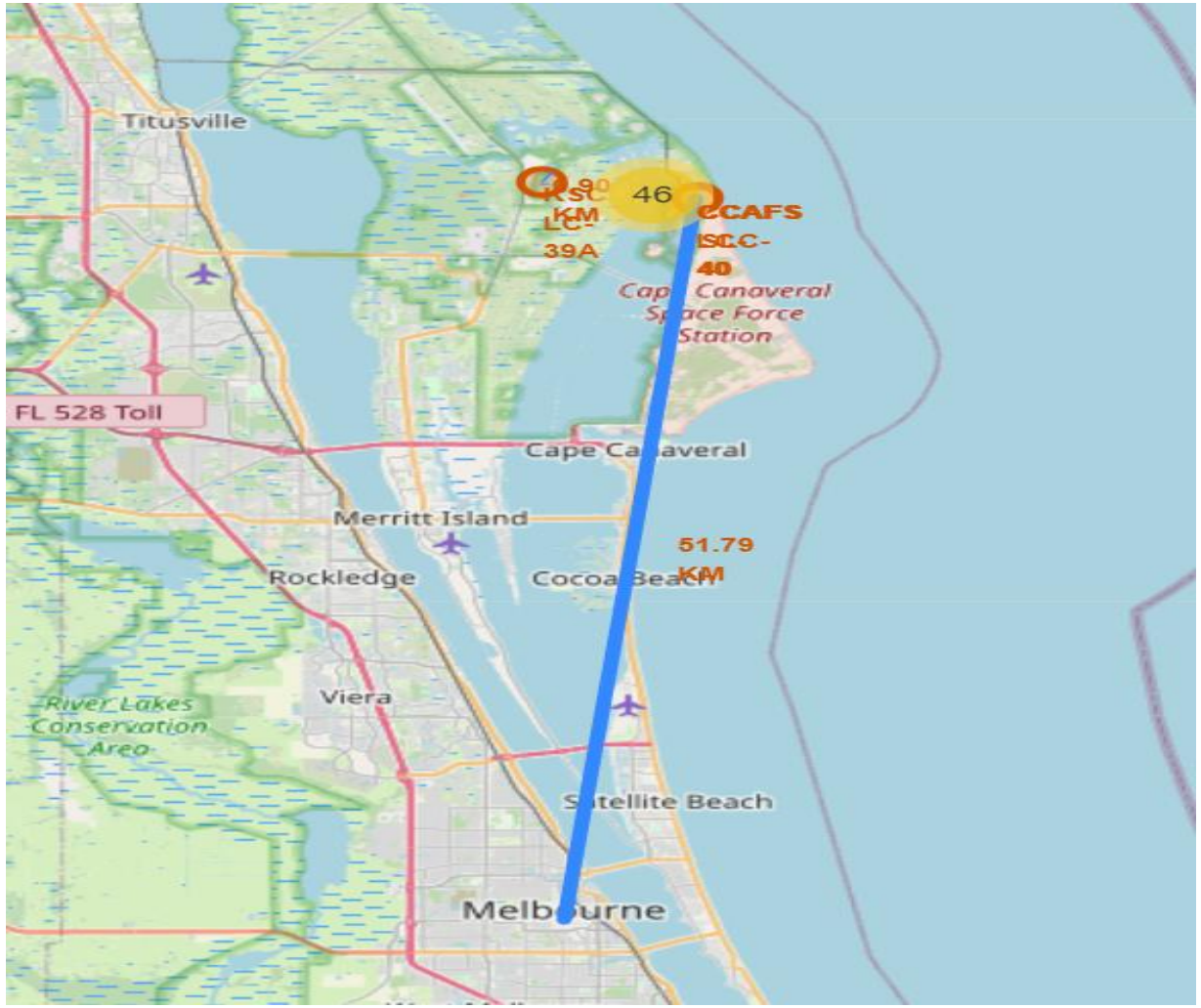
- All launch site are close to the coastline to reduce the risk of explosion or any harm near people
- According the web, it is beneficial for launch sites to be located near the equator - the spin of the Earth can help give an additional push. The surface of the Earth at the equator is moving at 1670 km/hr.
- To determine whether launch site is near the equator, we need to look at the latitude. If the latitude is near 0, it is close to the equator.
- For instance, all the launch site's latitude is close to 0 and less than 45. This mean that all the launch site is close to the equator.

Launch site with label

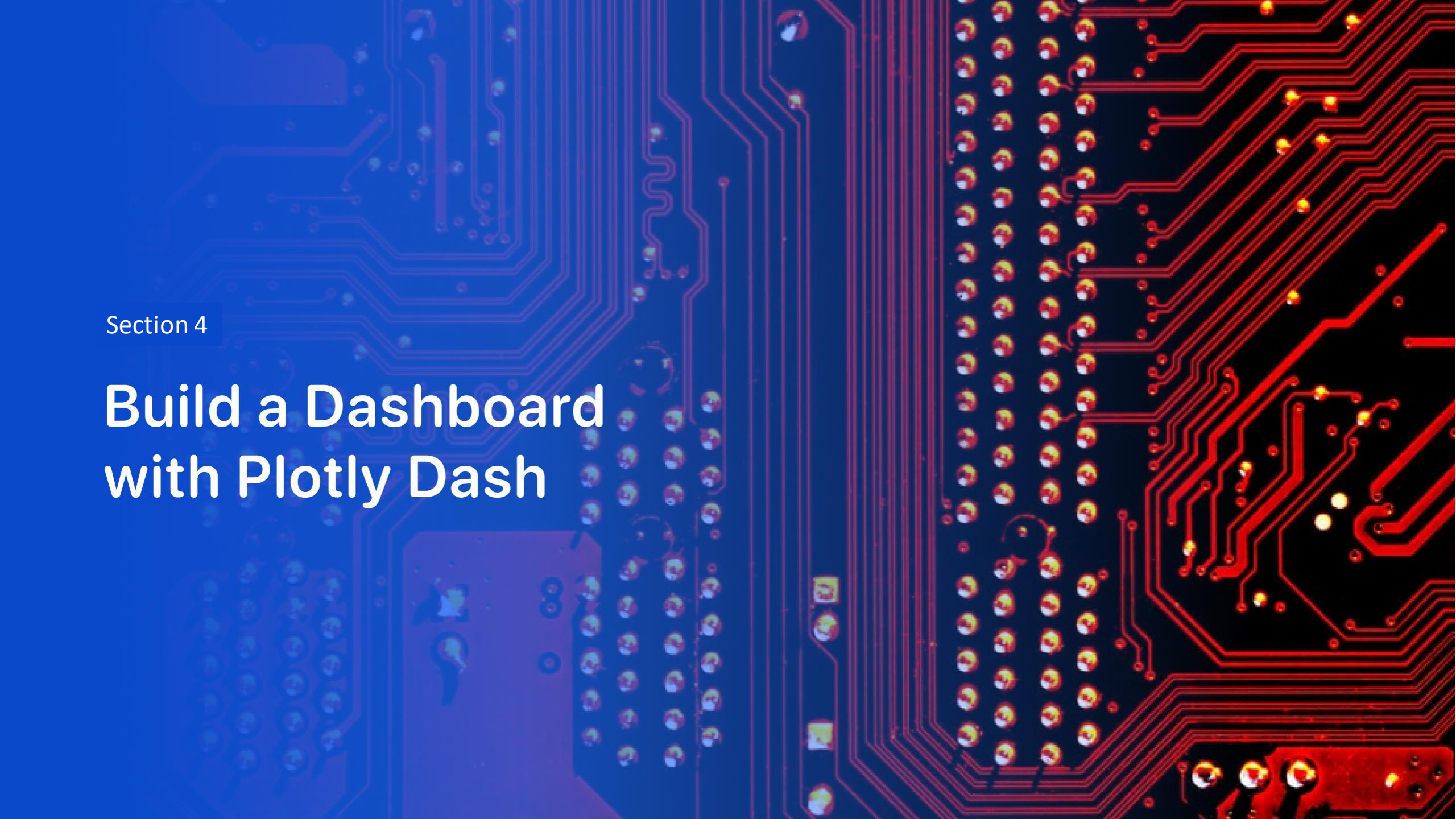


- We use color label to determine if the launch site attempt successful or not successful.
- For instance, green refer to successful and red refer to unsuccessful.
- Among the four launch sites, KSC LC-39A boasts the highest success rate, achieving success in 10 out of 13 attempts.

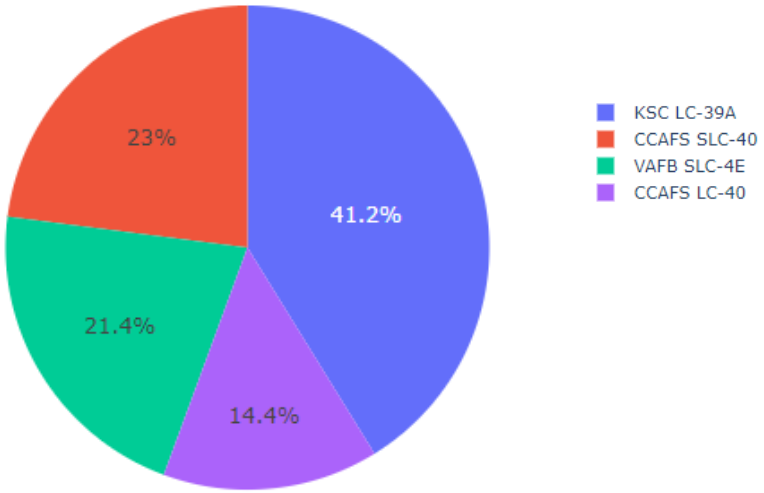
City to CCAFS SLC



- The distance to the nearest city is Melbourne and is around 51.79 Km away.
- It is relatively far from city

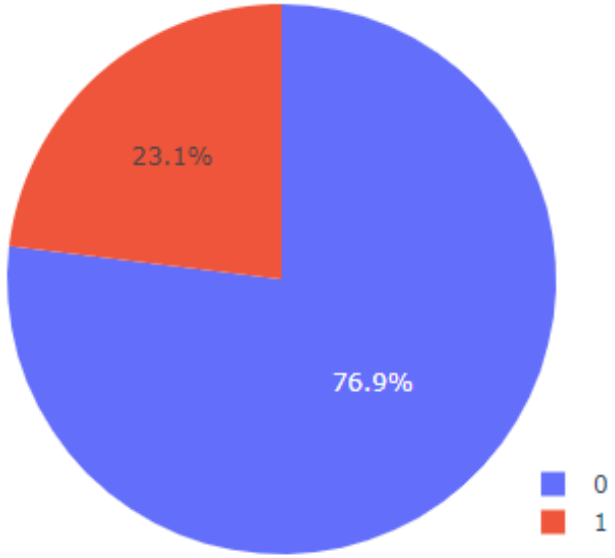

Section 4
Build a Dashboard with Plotly Dash

Successful launch for all launch site



- From the pie chart, we can see that KSC LC-39A has the most successful attempt of 41.2%
- CCAFS LC has the least success

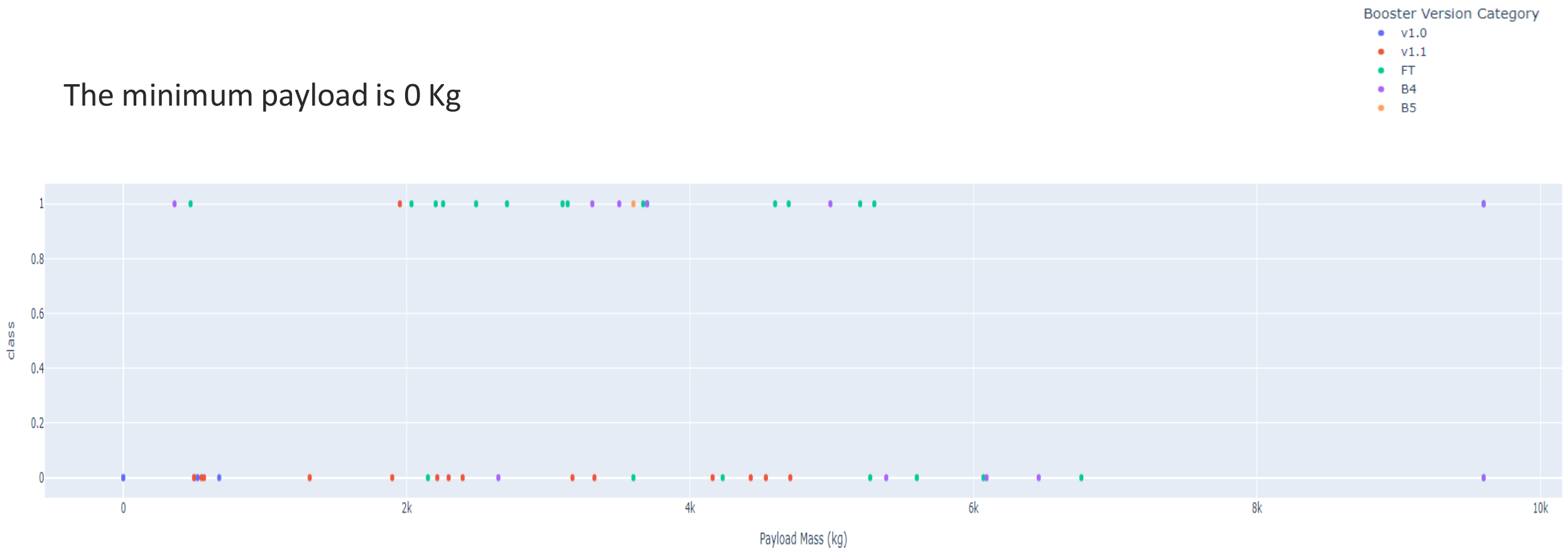
Analyzing the most successful launch site



- KSC LC-39A has 13 attempts
- 10 successful launch and 3 unsuccessful launch

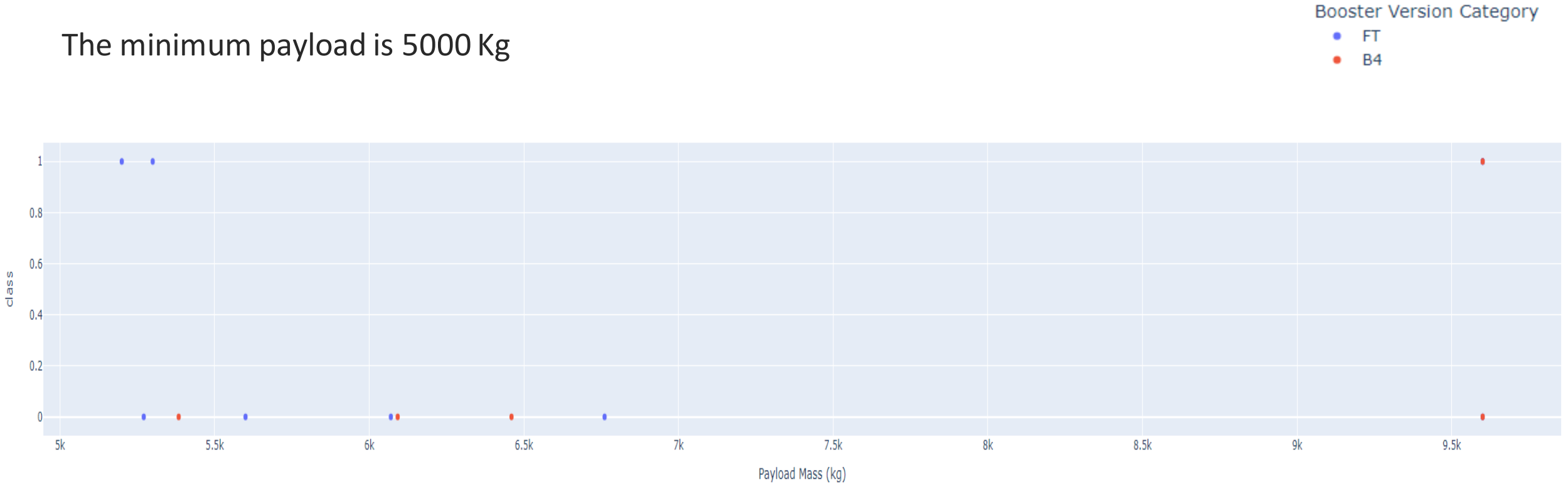
Payload vs Launch Outcome (1)

The minimum payload is 0 Kg



Payload vs Launch Outcome (2)

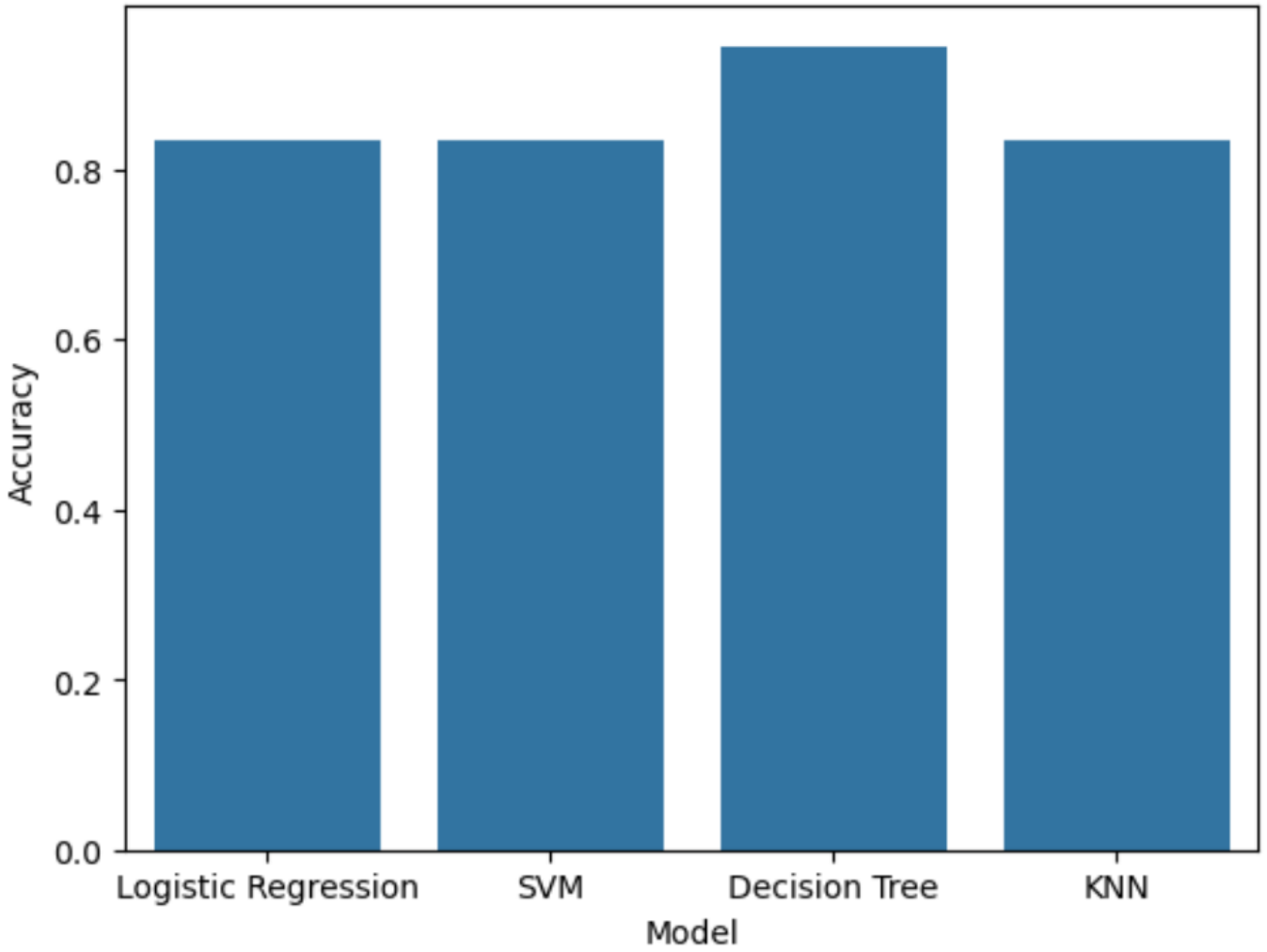
The minimum payload is 5000 Kg



Section 5

Predictive Analysis (Classification)

Classification Accuracy Testing dataset

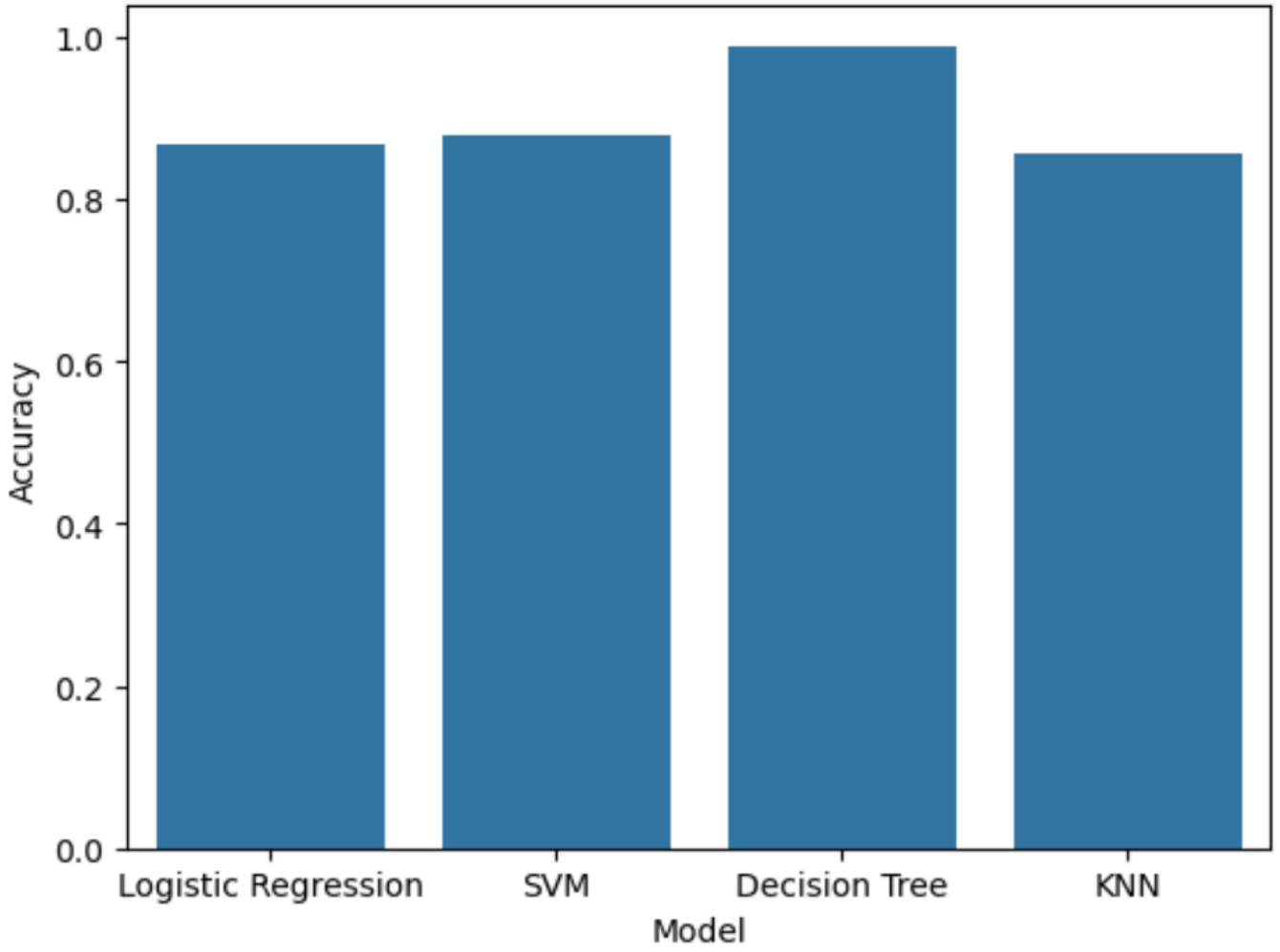


	Model	Accuracy	F1 Score
0	Logistic Regression	0.833333	0.909091
1	SVM	0.833333	0.916031
2	Decision Tree	0.944444	0.991736
3	KNN	0.833333	0.900763

Explanation

- For testing dataset, Decision Tree show a much better performance compared to the other models.
- The reason for the other model to have the same accuracy is due to the small dataset. For instance, it only have 16 training datasets which is insufficient for a typical training task.

Classification Accuracy All dataset



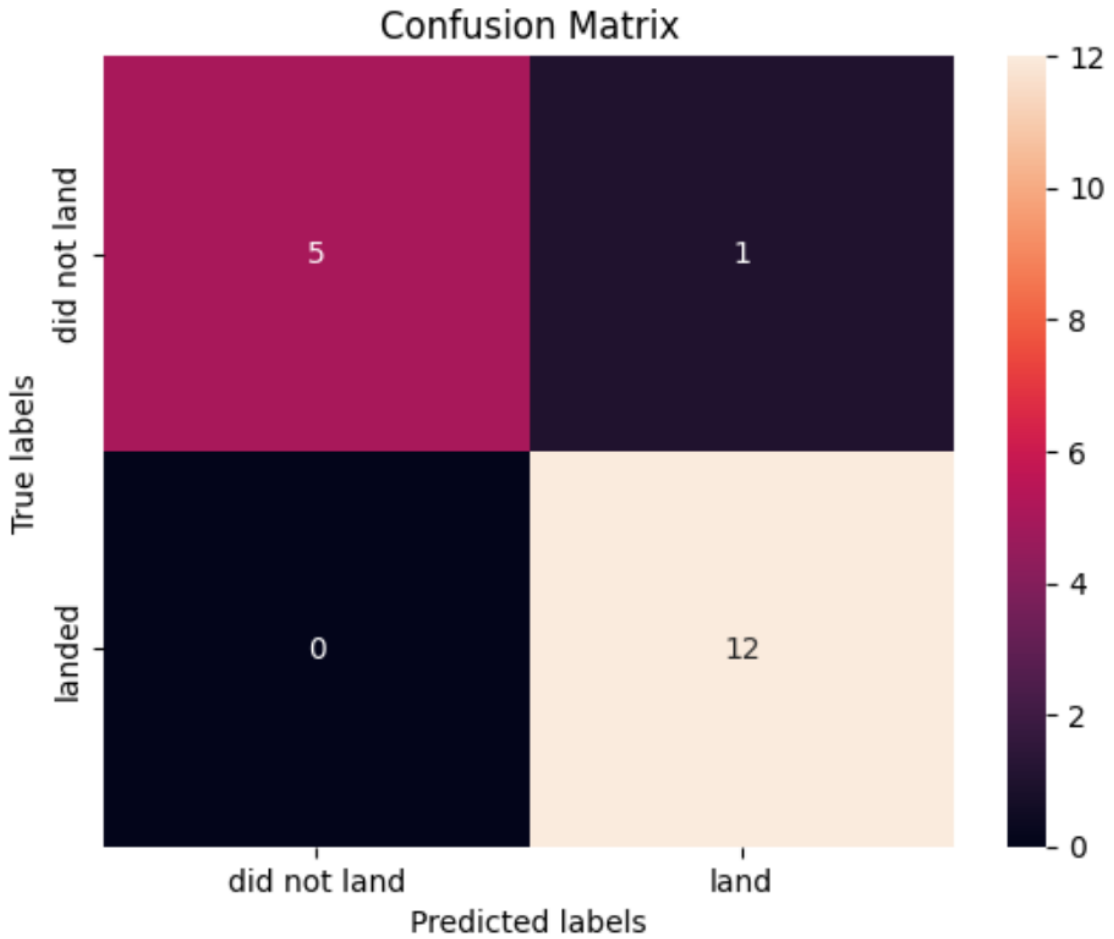
	Model	Accuracy	F1 Score
0	Logistic Regression	0.866667	0.909091
1	SVM	0.877778	0.916031
2	Decision Tree	0.988889	0.991736
3	KNN	0.855556	0.900763

Explanation

- Decision Tree is the best model as it has the accuracy of 98% and 99% for F1 score

- The best parameter for Decision Tree are as follows:
criterion': 'entropy',
'max_depth': 16,
'max_features': 'sqrt',
'min_samples_leaf': 1,
'min_samples_split': 2,
'splitter': 'best'

Confusion Matrix



Explanation

This is the confusion matrix for the decision tree. The optimal confusion matrix would ideally show zero false positives and zero false negatives. In our current matrix, we only have one false positive, which considered a good model.

Conclusions

- Decision Tree is the best model for this dataset of 98% for accuracy score and 99% for F1 score
- KSC LC-39A has the most successful attempt of 41.2%
- The success rate increase when the flight number attempt is over 50
- KSC LC 39A has a 100% success rate from the range of 2000 Kg to around 5500 Kg
- Even though the Orbits of ES-L1, GEO, HEO and SSO have 100% success rate, ES-L1, GEO and HEO only contains one data point which is not reliable to conclude that they are the best orbits.
- 100% success rate for ES-L1, SSO, HEO and MEO when the Pay Load Mass in the range of 500 – 4000 Kg
- VAFB SLC 4E has a higher chance of success rate than the other launch site

Thank you!