IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexander Rafael

29th March 2025

Outline

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

Executive Summary

This project aims to predict whether a rocket launching is success or failed to help the business understanding and requirements of SpaceX Coorporation based on the historical records of SpaceX Rocket Launches. In order to predict it, there are several details that needs attention.

Methodology

The historical data will be collected by using SpaceX API and scrap the records from the Wikipedia. To ensure its quality, the data will being wrangled with some manipulation and exploration. There are several visualization that ease to understand the pattern of data's distribution. Finally, there are 4 classification models : logistic regression, K nearest neighbours, support vector machine and decision tree that will be fitted and evaluated to get the best accuracy score.

Results

Based on the fitted model, there are 3 models that has the same accuracy's score : Logistic Regression, Support Vector Machine and Decision Tree with 100% Accuracy. Because the small size of dataset, it may lead to a limitation on this project which doesn't rule out the possibility that the models is not generalized since less data that can be used for training the model.

Introduction

Background

- SpaceX has the Falcon 9 Rocket Launches that can reuse its first stage for another launching if the launch was successful.
- Therefore, SpaceX may reduce the cost of a launch to 62 Million Dollars, rather than the other providers upward of 165 million dollars

Problems

- To have an effective advertisement of the cheaper launch cost, SpaceX needs to predict the launches outcome.
- The prediction will be used when another alternate company wants to bid agains SpaceX to a rocket launch.

So, to predict the outcome, we will use the **Machine Learning** approach using the historical launches data.

Section 1

Methodology

Methodology

Executive Summary

1. Data collection methodology

    - The data was collected by extracting the SpaceX API and scrapping Falcon 9 and Falcon Heavy Launches on Wikipedia Website

2. Perform data wrangling

    - The data was processed by exploring the null values, the cardinaility of each category, and develop a new numerical target label

3. Perform exploratory data analysis (EDA) using visualization and SQL

4. Perform interactive visual analytics using Folium and Plotly Dash

5. Perform predictive analysis using classification models

    - Build several models, logistic regression, Support Vector Machine, Decision Tree, and K Nearest Neighbours, doing the hyperparameter tuning, fitting the data, and evaluate the results

Data Collection

The dataset collected by 2 methods from 2 different sources, which are:

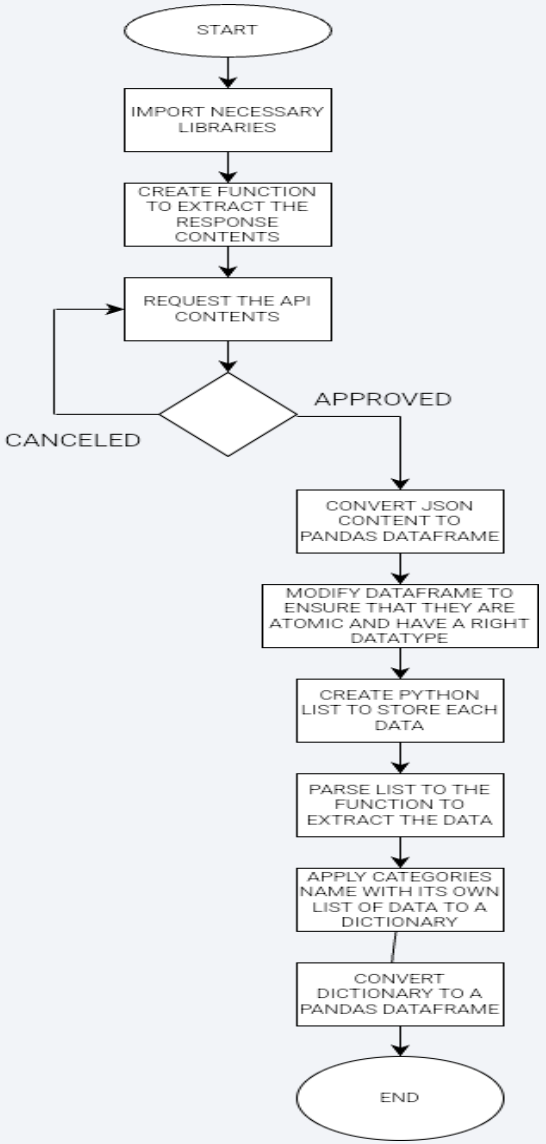
Application program Interface (API)

A SpaceX API that connect the program with the data of SpaceX Rocket Launching Records and Histories

Web based

A wikipedia site of Falcon 9 and Falcon Heavy Launches records that includes many informations and categories about the launching records. The data will be scrapped from the webpage using a beautiful soup object

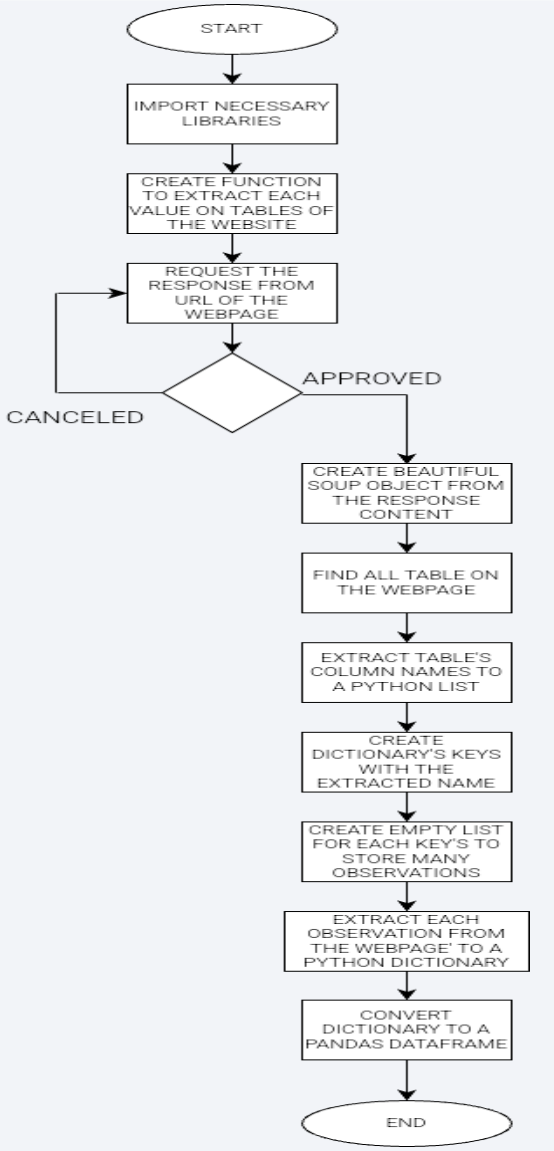
Data Collection – SpaceX API



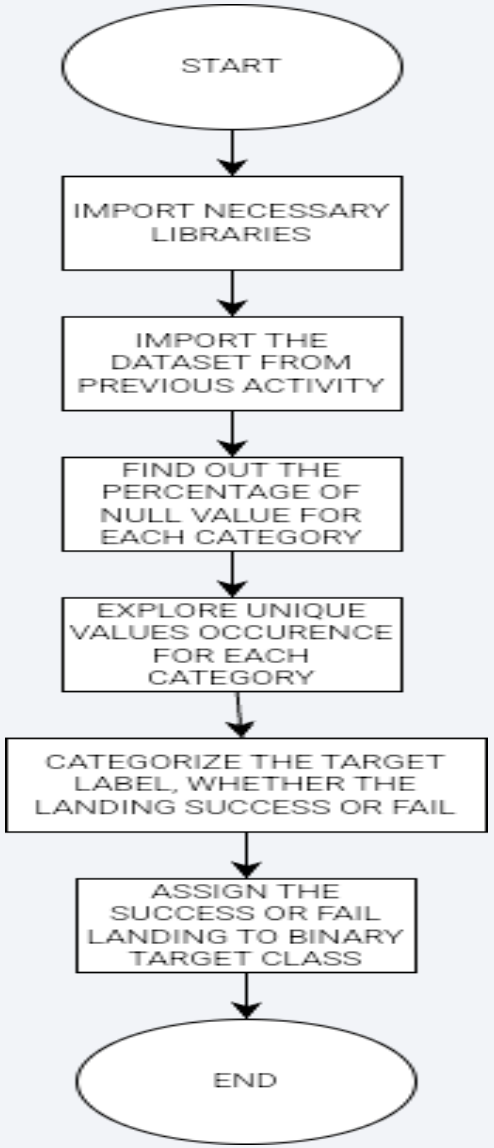
The historical data was collected by extracting the records from the SpaceX's Application Programming Interface that connected with the python programming languages

Data Collection - Scraping

The historical data was collected by scrapping the wikipedia's site tables that relates to the Falcon 9 and Heavy Falcon Launches Records



Data Wrangling



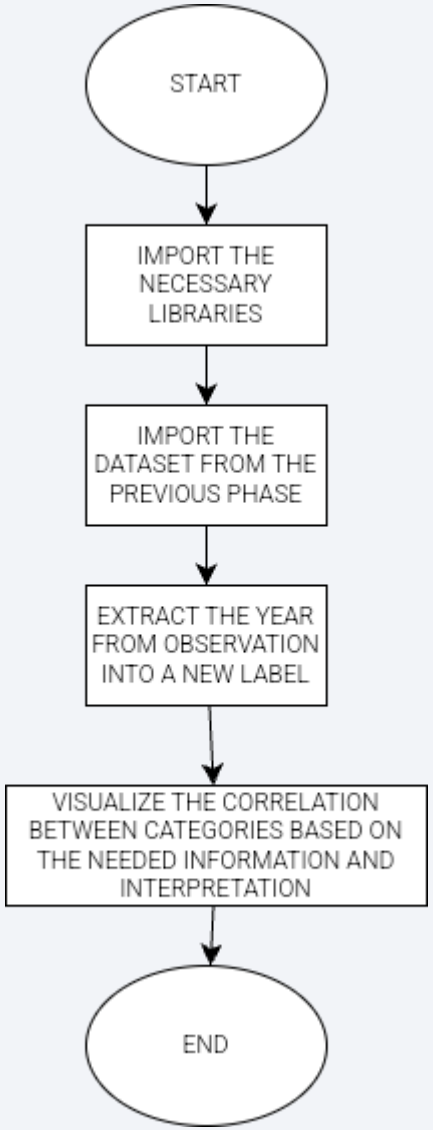
There are several actions to wrangling the data :

1. **NULL Values**, the empty record on the observation. Knowing the number of NULL Values can ease the interpretation
2. **Cardinality**, which is the number of unique values in a category, to help find inconsistent data
3. **Labelling,** which is aggregating many label into a binary target label to ease the classification from small data

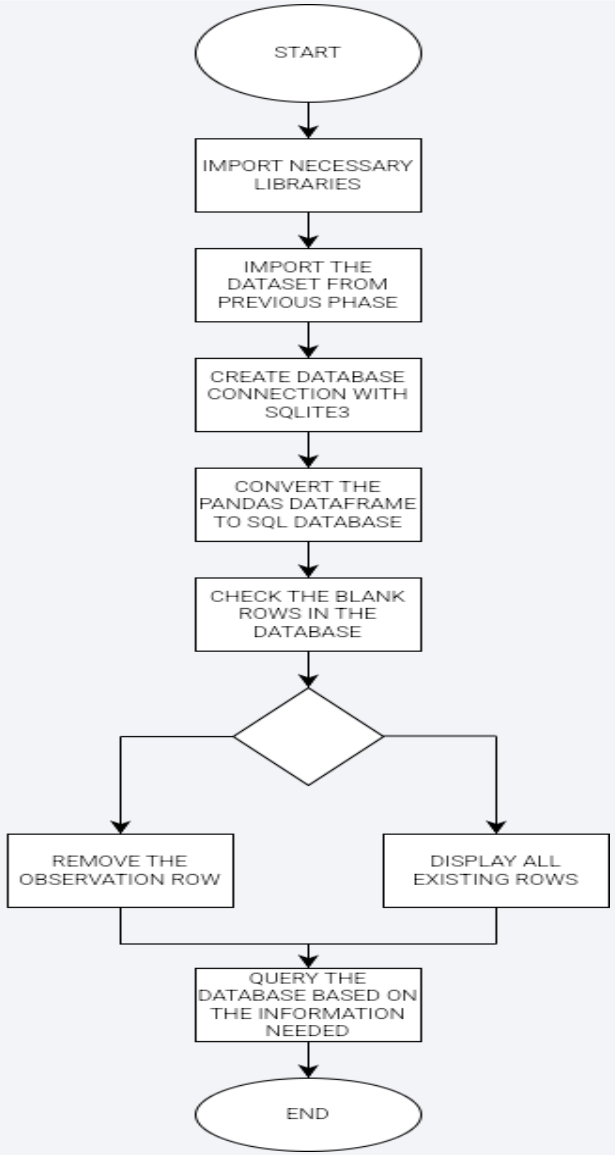
EDA with Data Visualization

There are several kinds of graphs that used in this activity, such as :

1. **Scatter plot,** aims to visualize the correlation/pattern between 2 numerical variables
2. **Bar plot**, aims to visualize the number on each categorical values, a numerical and categorical's data distribution
3. **Line chart**, same as bar plot, to visualize a numerical and categorical, but this is to more highlight about a progress in a certain period



EDA with SQL



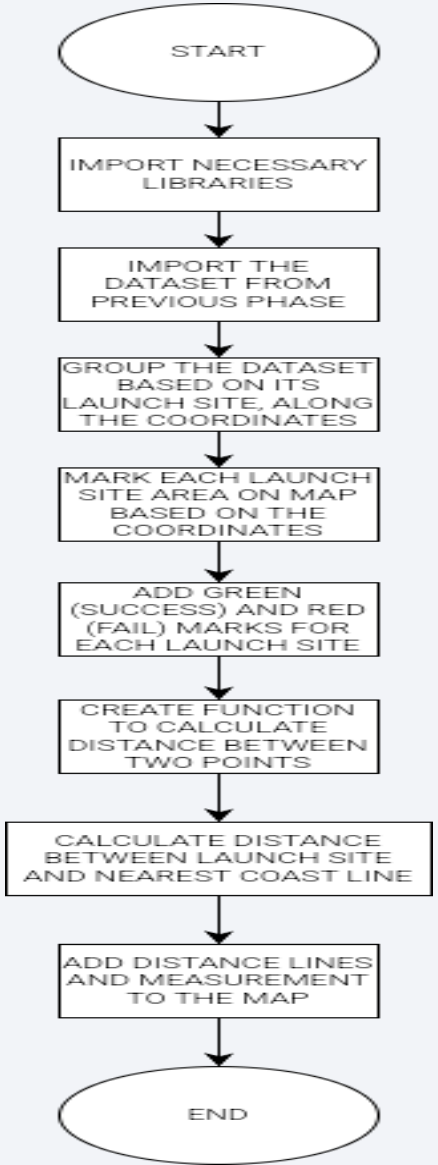
The Exploratory Data Analysis with SQL means to query data with some condition, aims to :

1. Knows deeper about the distribution of data in certain situation
2. Find out the statistical calculation and information on some categories
3. Perform a Data Manipulation Language with the SQL Statement

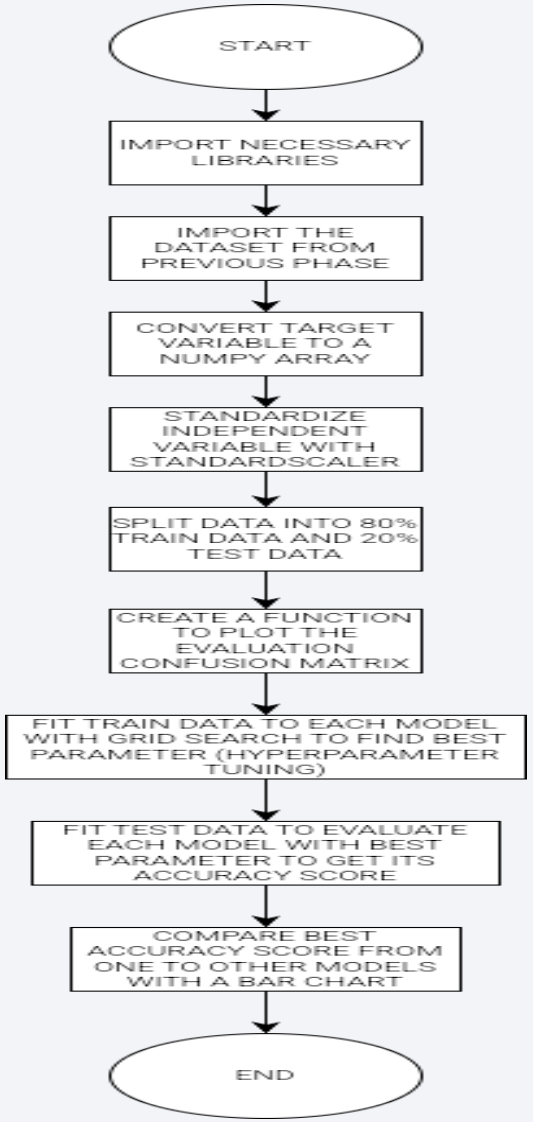
Build an Interactive Map with Folium

There will be several object that occurs in the map, such as :

1. **Circle**, indicates an area of the launch site of the launching rockets
2. **Green mark,** indicates the successful landing of the rocket
3. **Red mark,** indicates the failed landing of the rocket
4. **Line**, indicates the distance between a launch site area to another object



Predictive Analysis (Classification)



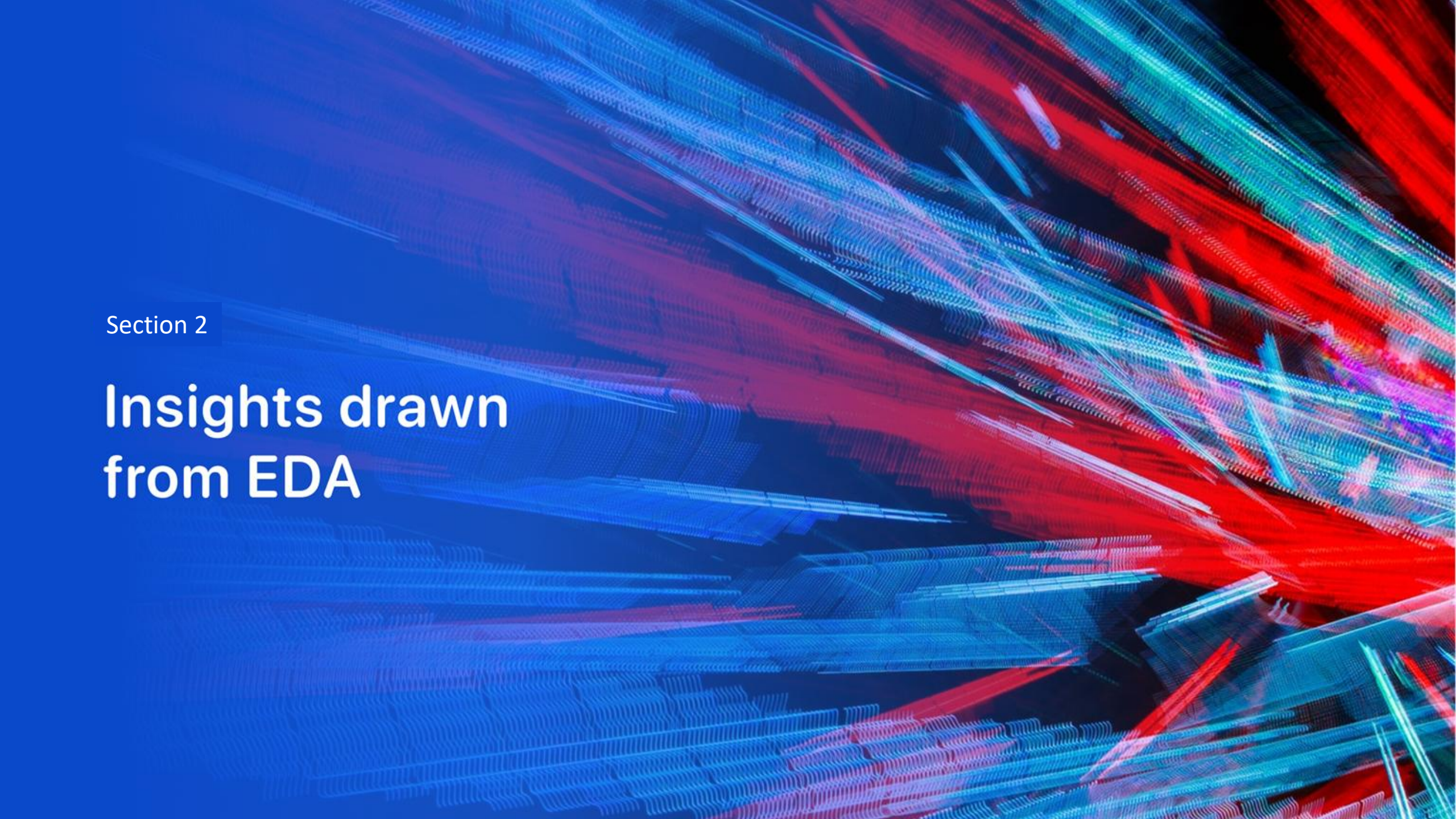
Because the target label either success or fail (binary), then it is categorized as classification. There are 4 common models that used for it, such as :

- Logistic Regression
- Support Vector Machine
- K Nearest Neighbours
- Decision Tree

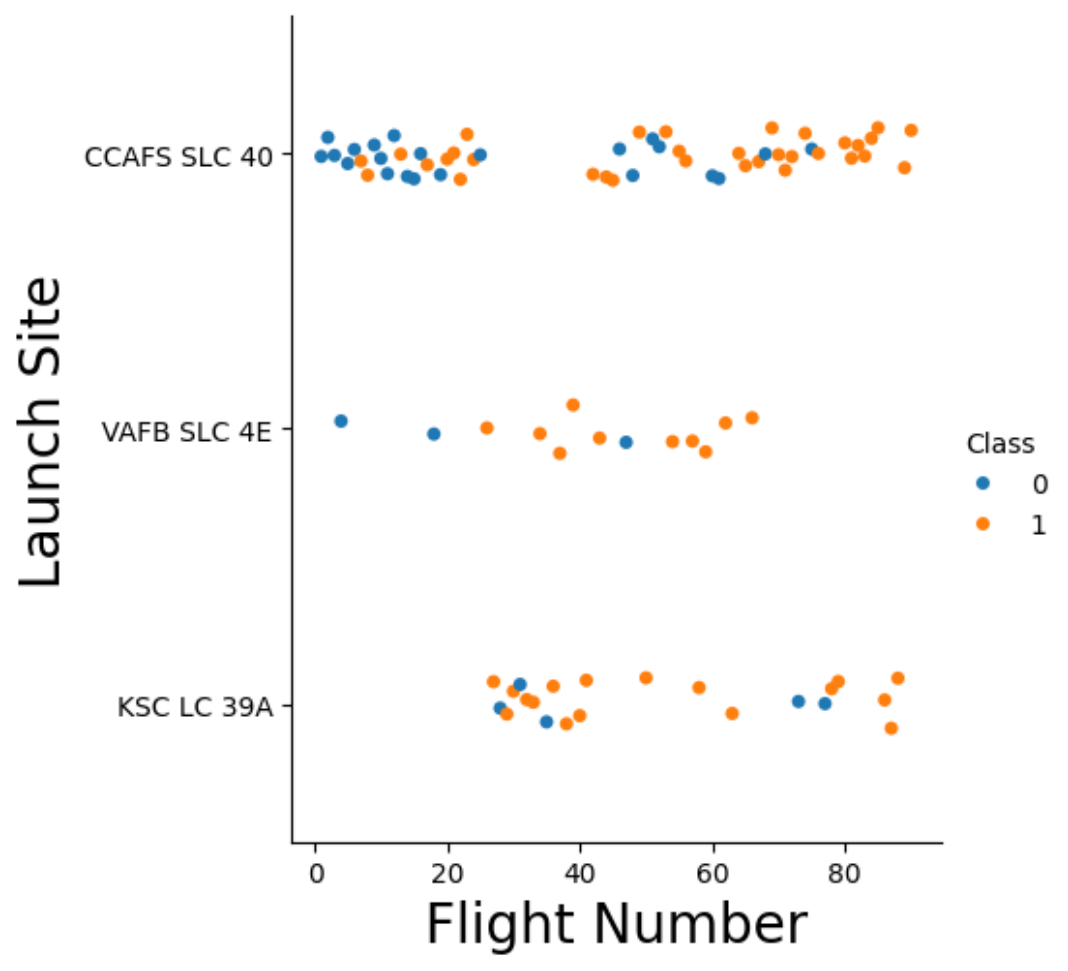
This analysis aims to determine which model give it highest accuration to the validation data

Section 2

Insights drawn from EDA

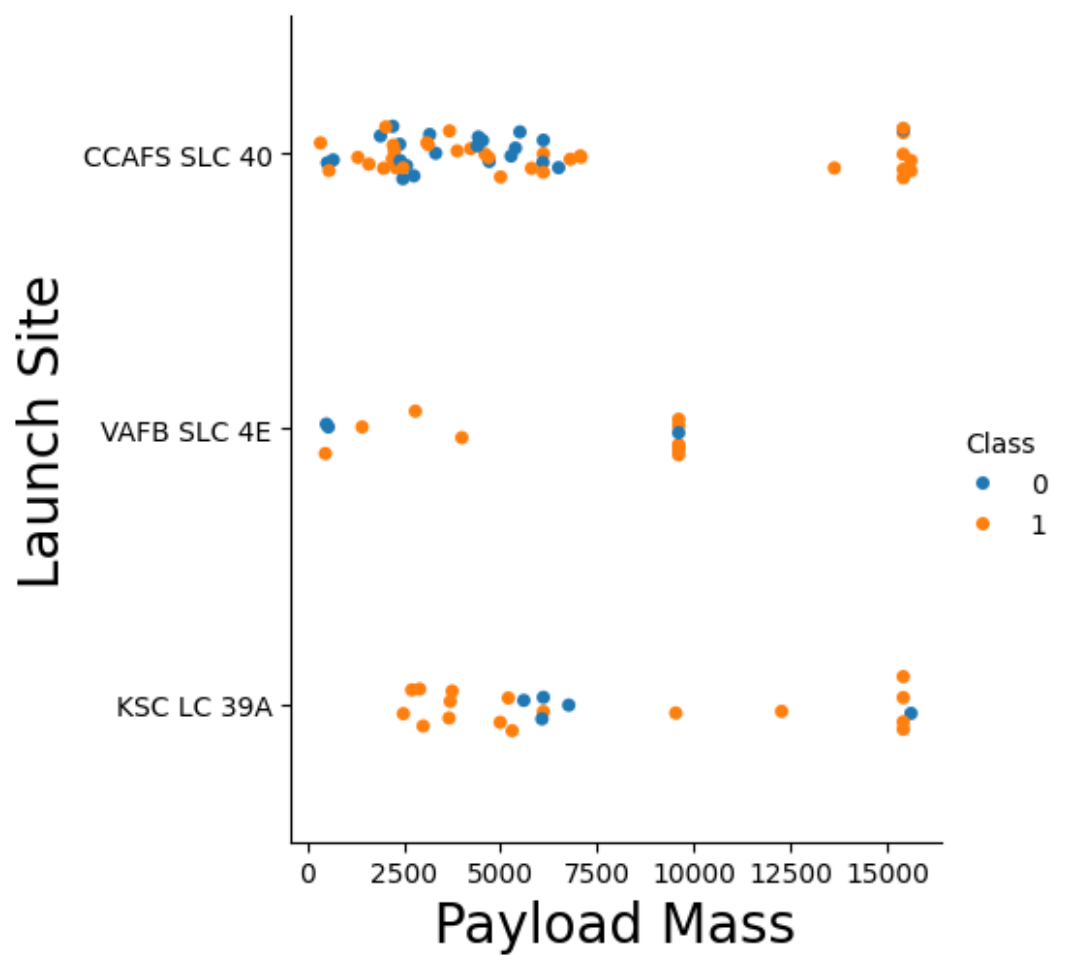


Flight Number vs. Launch Site



On CCAFS SLC 40 site has more numbers of flight, where the rest are lesser than it. Overall, there are more numbers of successful flight than the failed one.

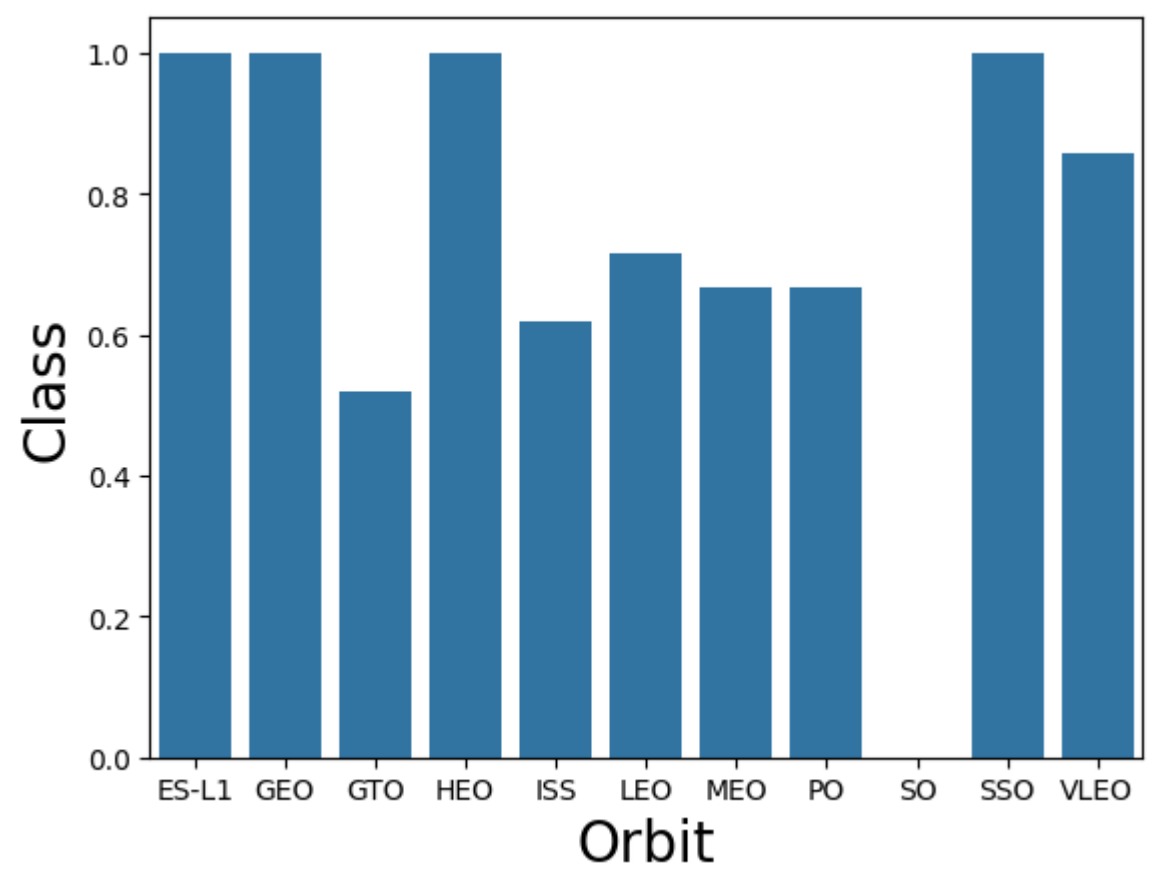
Payload vs. Launch Site



On CCAFS SLC 40 site, the launching rockets are most likely come from the light to medium mass, while for the others, there are most likely have same distribution from the level of payload mass.

The number of failed launch mostly comes from the low payload mass.

Success Rate vs. Orbit Type

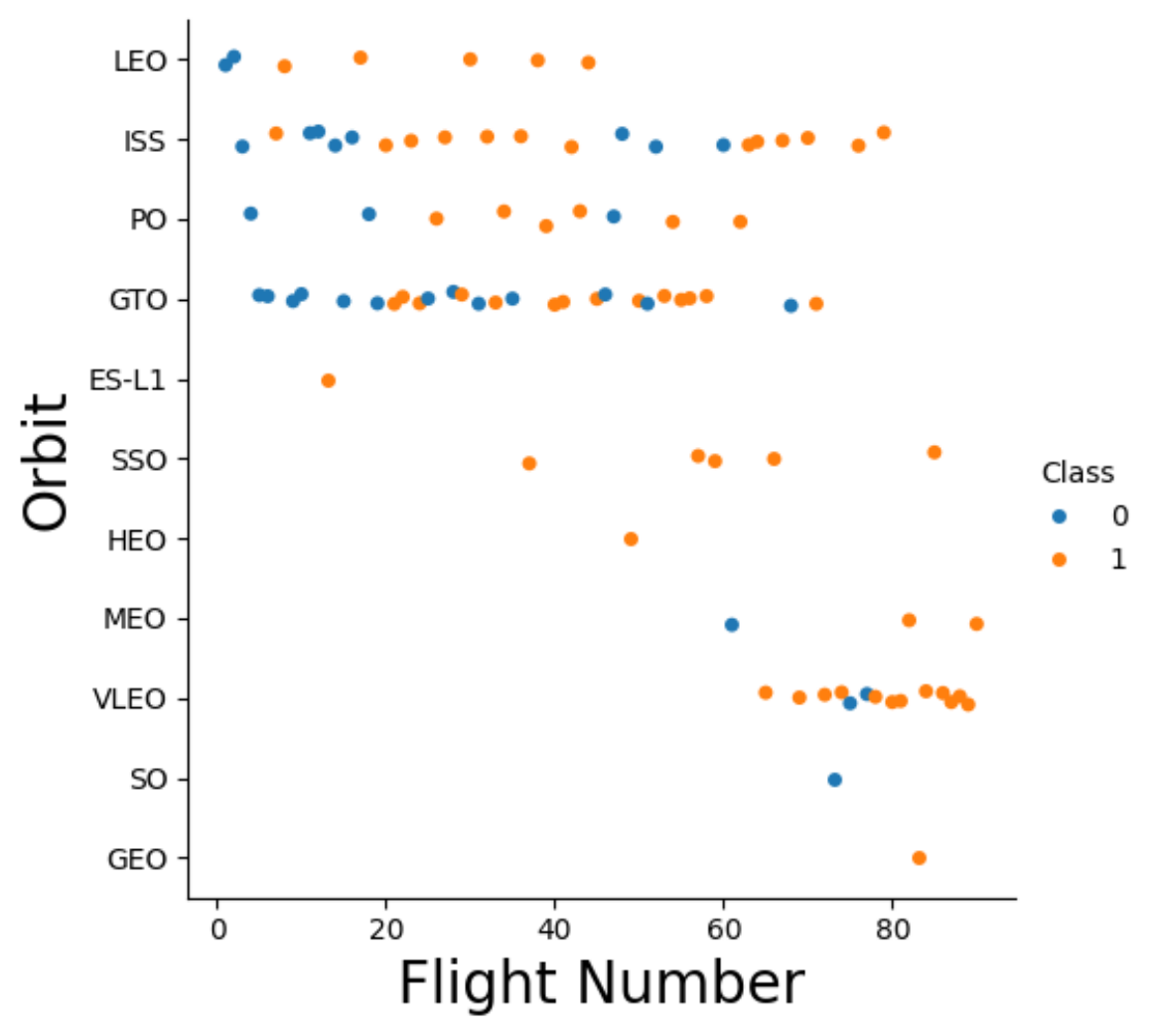


The ES-L1, GEO, HEO and SSO orbits has the most success rate rocket launch, which is 1, or 100% success.

The SO orbit has the lowest, which there are no success launch at all.

The rest are diverse at the range 1 to 0 success rate.

Flight Number vs. Orbit Type

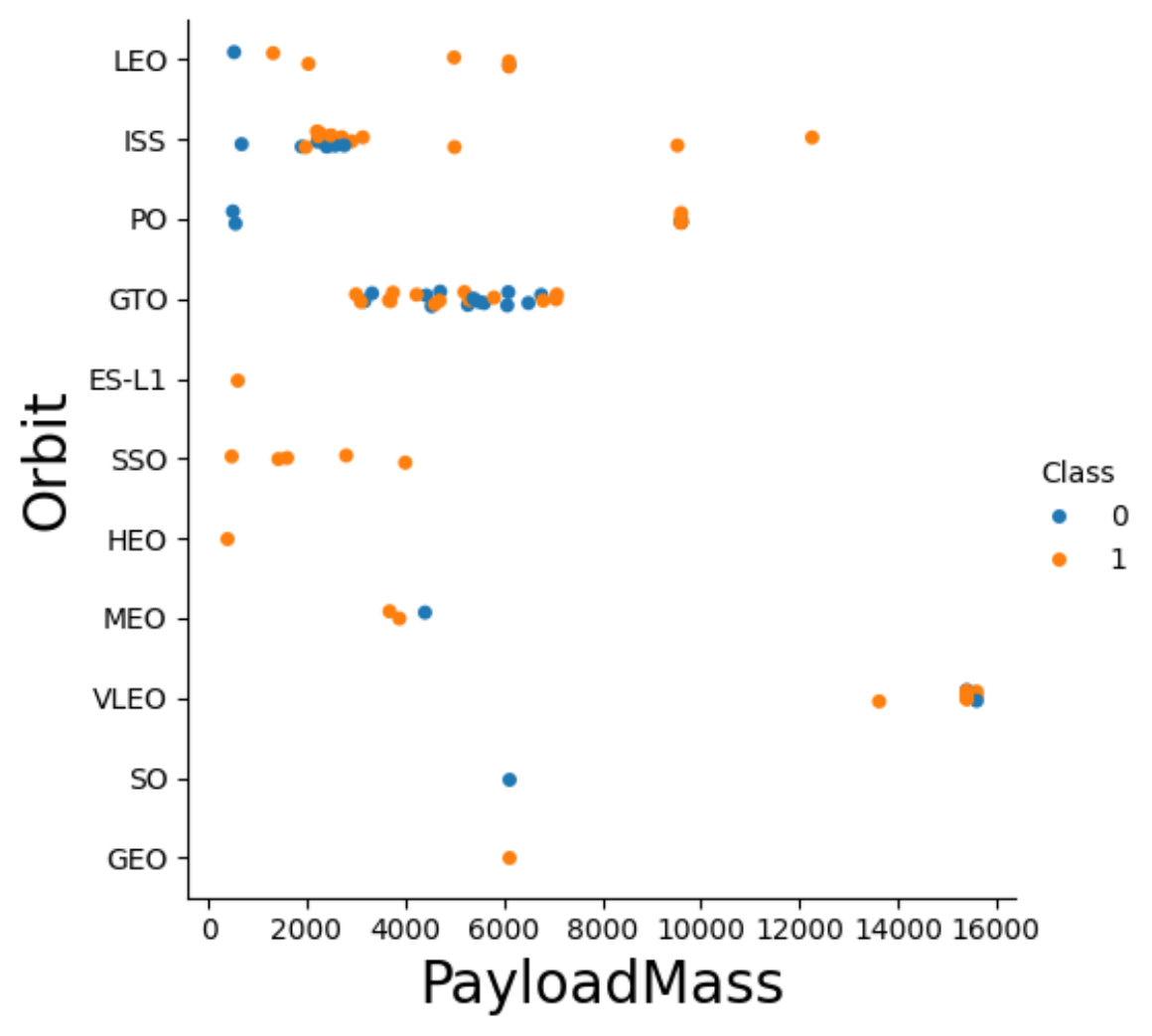


The LEO, ISS, PO and GTO orbits has lots of flight number of the rocket, with its each success and fail numbers.

While the rest, only has a few of flight number.

This condition bring the analysis of data may not be relevant and accurate, because of the difference distribution of the data

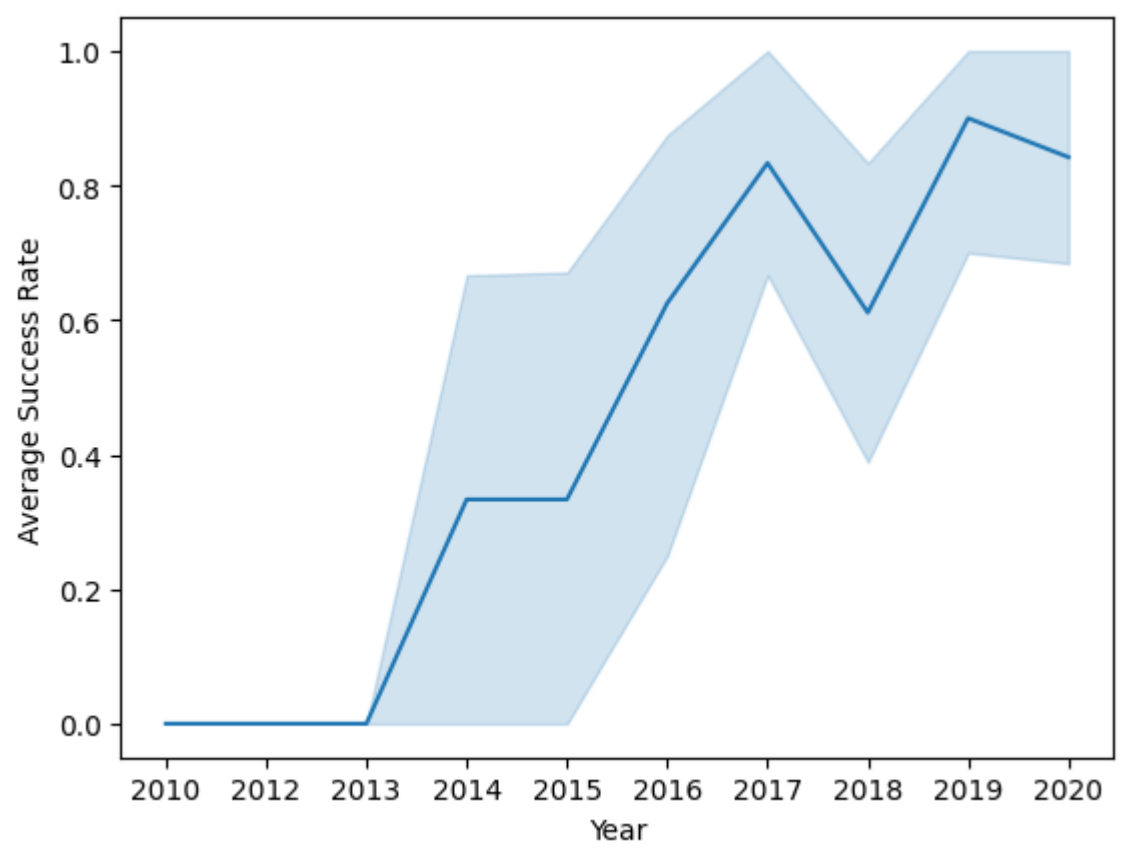
Payload vs. Orbit Type



There are very many diverse distributions between the payload mass on each orbit, indicating that each rocket is unique each other.

In GTO orbit, the distributed rockets are most likely comes from 2000 to 8000 payload mass indicates the unique patterns and consistency there.

Launch Success Yearly Trend



From year to year,

The overall success rate of rocket launched are increases in the average, this is a good milestones that indicating the better security and performance on the systems that they are using.

All Launch Site Names

```
[42] %sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTABLE
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

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

Based on the snippet code, There are four (4) unique launch site names on the dataset as mentioned above.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

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

Based on the snippet code, There are several flight that the launch site names begin with the “CCA” characters. In this case, only the top 5 observations displayed as the limitation on the questions

Total Payload Mass

```
%sql SELECT BOOSTER_VERSION, SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE CUSTOMER LIKE 'NASA%' GROUP BY BOOSTER_VERSION
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

Booster_Version	SUM(PAYLOAD_MASS__KG_)
F9 B4 B1039.2	2647
F9 B4 B1039.1	3310
F9 B4 B1045.1	362
F9 B4 B1045.2	2697
F9 B5 B1046.4	12050
F9 B5 B1056.2	2268
F9 B5 B1058.4	2972
F9 B5 B1059.2	1977
F9 B5B1050	2500
F9 B5B1051.1	12055
F9 B5B1056.1	2495
F9 B5B1058.1	12530
F9 B5B1059.1	2617
F9 B5B1061.1	12500
F9 B5B1063.1	1192
F9 FT B1035.2	2205
F9 FT B1021.1	3136
F9 FT B1025.1	2257
F9 FT B1031.1	2490
F9 FT B1035.1	2708
F9 v1.0 B0004	0
F9 v1.0 B0005	525
F9 v1.0 B0006	500
F9 v1.0 B0007	677
F9 v1.1	2296
F9 v1.1 B1010	2216
F9 v1.1 B1012	2395
F9 v1.1 B1015	1898
F9 v1.1 B1017	553
F9 v1.1 B1018	1952

Based on the snippet code, There are the group of the total payload mass, for each booster version type that carried by boosters from NASA.

Average Payload Mass by F9 v1.1

```
%sql SELECT BOOSTER_VERSION, AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE 'F9 V1.1%' GROUP BY BOOSTER_VERSION
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

Booster_Version	AVG(PAYLOAD_MASS__KG_)
F9 v1.1	2928.4
F9 v1.1 B1003	500.0
F9 v1.1 B1010	2216.0
F9 v1.1 B1011	4428.0
F9 v1.1 B1012	2395.0
F9 v1.1 B1013	570.0
F9 v1.1 B1014	4159.0
F9 v1.1 B1015	1898.0
F9 v1.1 B1016	4707.0
F9 v1.1 B1017	553.0
F9 v1.1 B1018	1952.0

Based on the snippet code, There are the group of payload mass average for each of the F9 V1.1 booster version type.

First Successful Ground Landing Date

```
%sql SELECT * FROM SPACEXTABLE WHERE LANDING_OUTCOME like "Success (ground pad)" order by date asc limit 1
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites 2034		LEO	Orbcomm	Success	Success (ground pad)

Based on the snippet code, There is the first observation of the successful rocket launch on the ground landing.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE "Success (drone ship)" AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

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

Based on the snippet code, there are 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

TASK 7 : List the total number of successful and failure mission outcomes

```
[48] %sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) FROM SPACEXTABLE GROUP BY MISSION_OUTCOME
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

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

Based on the snippet code, There are the total of calculations number of successful and failure rocket launching mission outcomes

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE) GROUP BY BOOSTER_VERSION
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

Based on the snippet code, There are the group of booster's name with its each total of payload mass for each names.

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) as MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTABLE WHERE LANDING_OUTCOME LIKE "Failure (drone ship)" AND substr(DATE
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

MONTH	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Based on the snippet code, There are the list of observations which has failed the landing outcomes during the 2015 period.

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

```
%sql SELECT LANDING_OUTCOME, COUNT(*) AS count, RANK() OVER (ORDER BY COUNT(*) DESC) AS Ranking FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP
```

```
 * sqlite:///ExerciseDB1.db
Done.
```

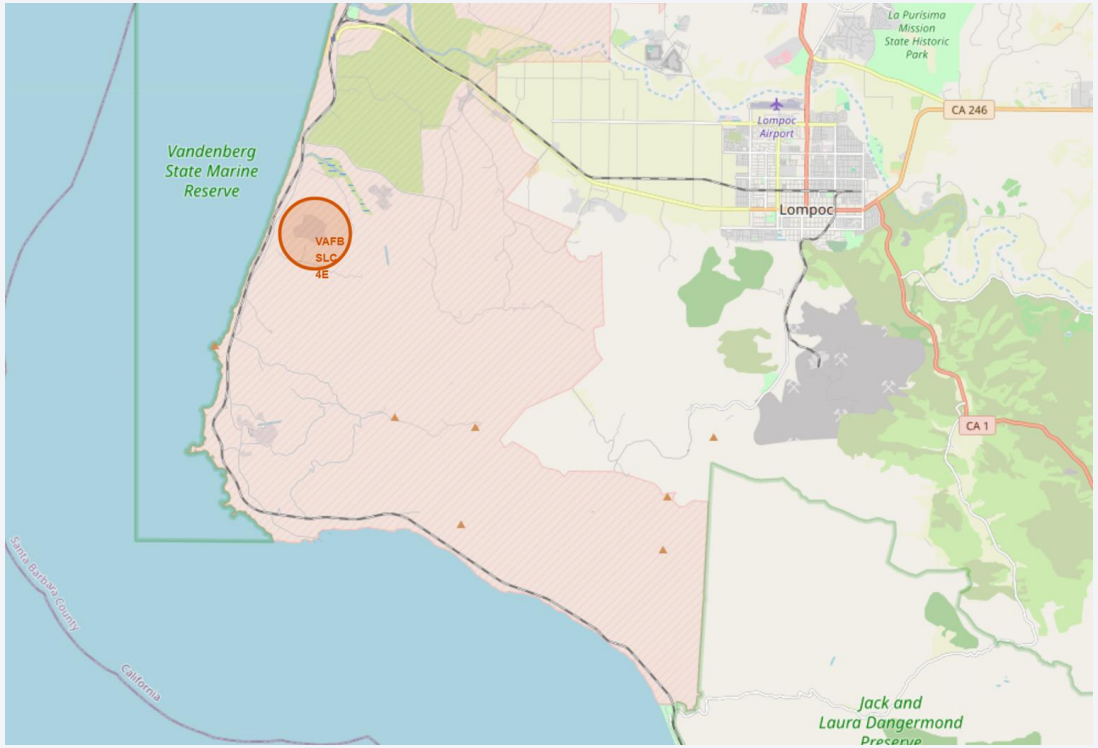
Landing_Outcome	count	Ranking
No attempt	10	1
Success (drone ship)	5	2
Failure (drone ship)	5	2
Success (ground pad)	3	4
Controlled (ocean)	3	4
Uncontrolled (ocean)	2	6
Failure (parachute)	2	6
Precluded (drone ship)	1	8

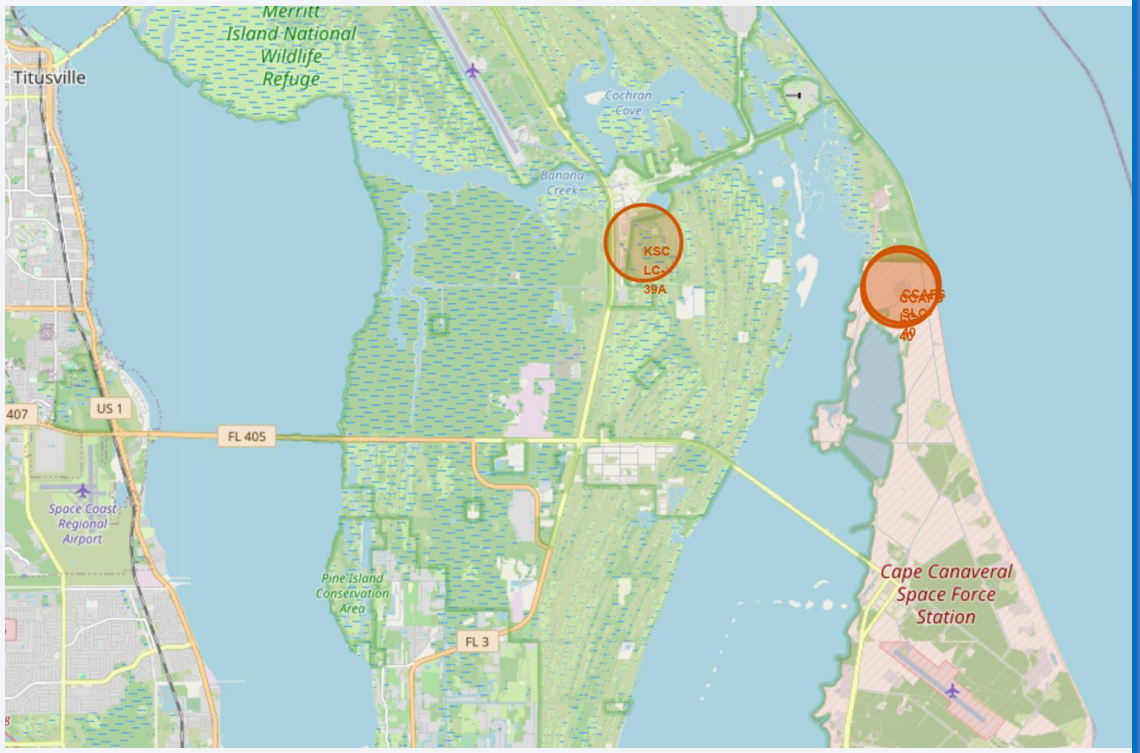
Based on the snippet code, there are the rank's records of landing outcomes between the date 2010-06-04 and 2017-03-20 in a descending order

Section 3

Launch Sites Proximities Analysis

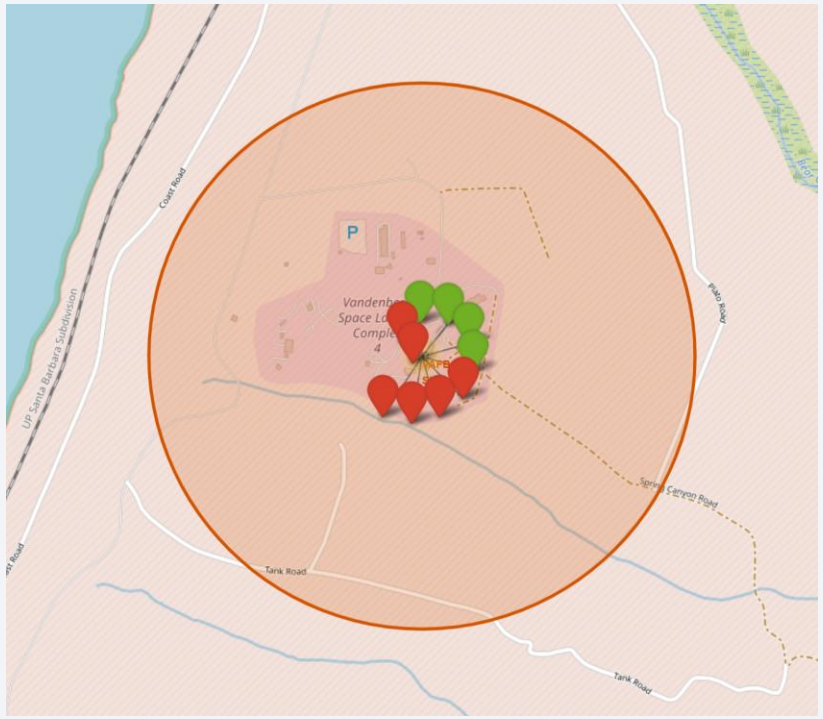
Launch Sites' Area Distribution

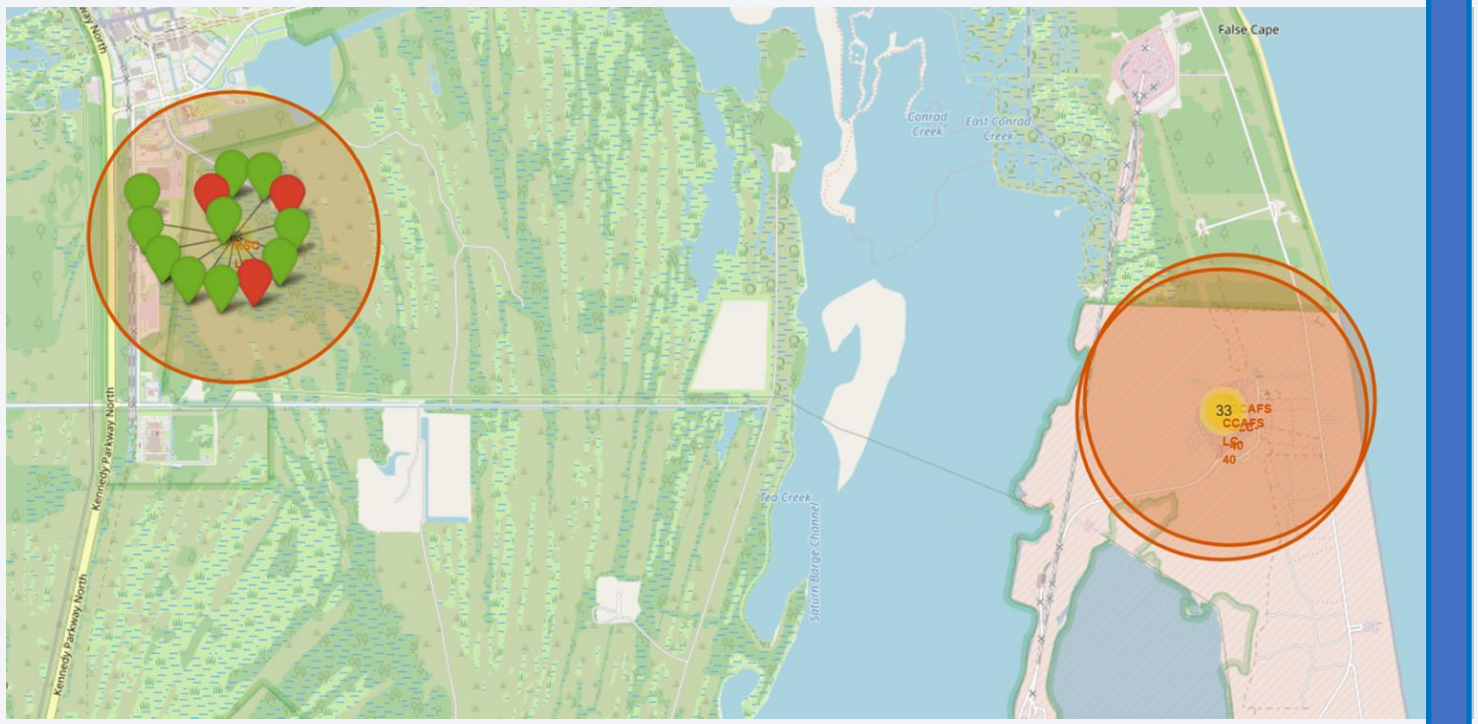




Based on the map, there are the geographical places from each launch sites based on the longitude and latitude coordinates.

Flight’s Distribution on Each Launch Site





Based on the map, there are the detailed geographical place from each record of rocket launch on each launch site area, where the green mark indicates a successful launch, otherwise the red indicated the failed.

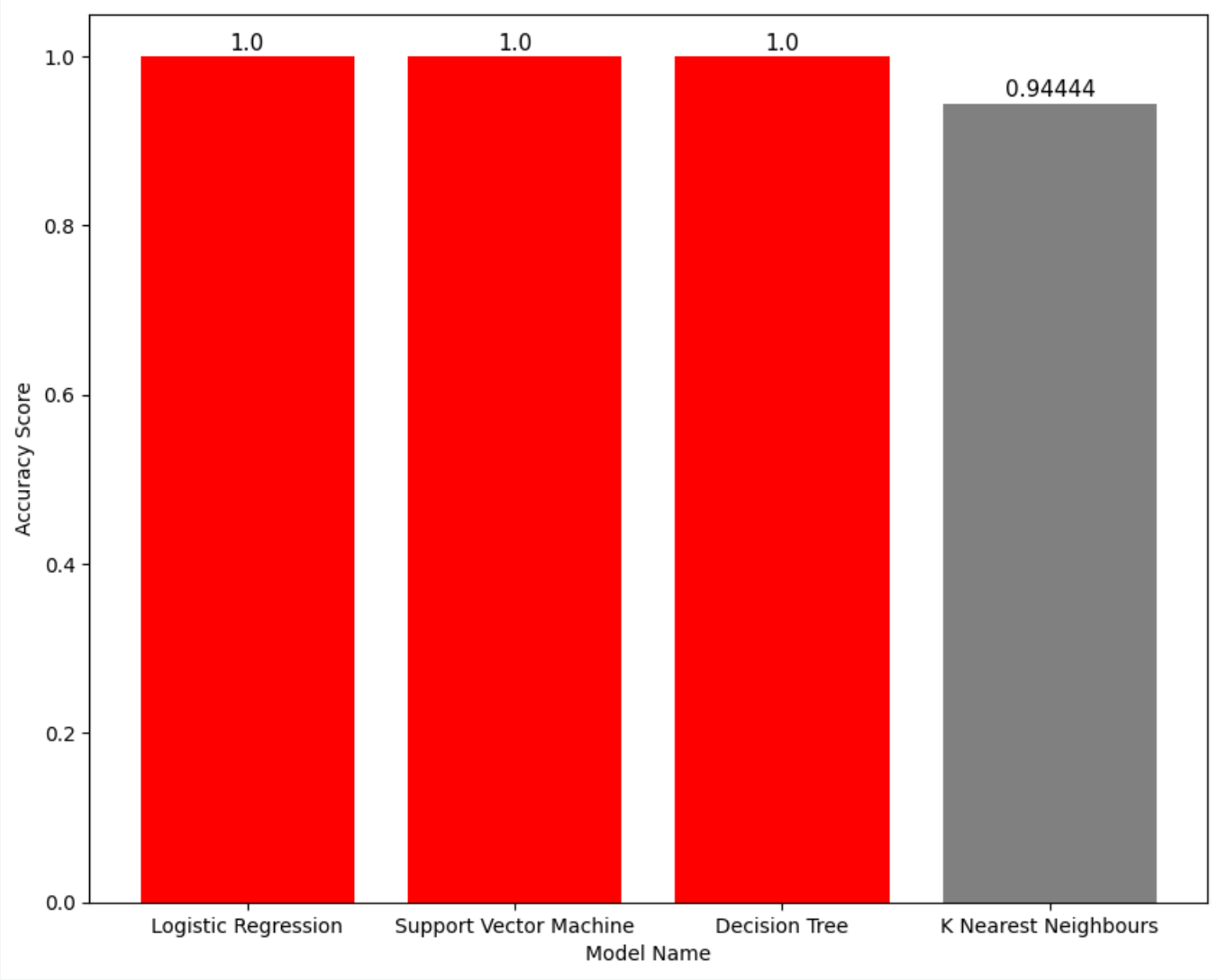
Launch Site's Distance with Nearest CoastLine

Based on the map, there is a path of the nearest **coastline** from the CCAFS SLC-40 Launch Sites Area, which is 0.86 KM

Section 5

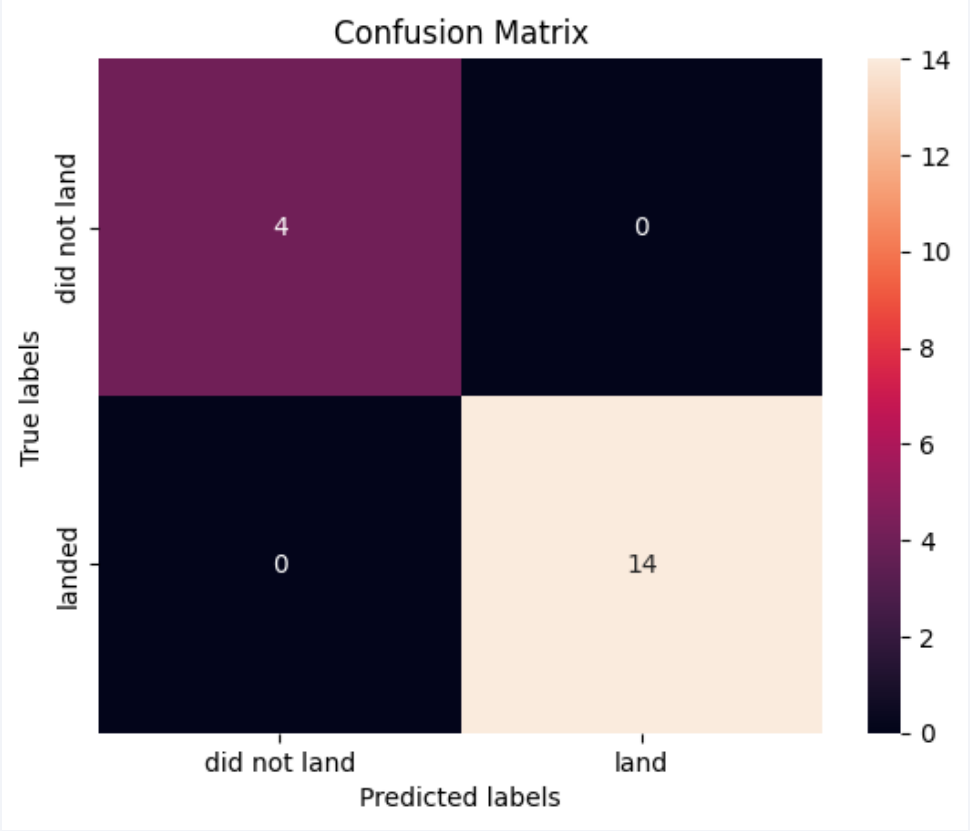
Predictive Analysis (Classification)

Classification Accuracy



There are 3 models at total that has the highest score of accuracy on the validation data, which consists of **Logistic Regression** and **Support Vector Machine**

Confusion Matrix



Because the model has 100% Accuracy on testing data, its mean that all predictions all truly predicted. In matrix, we can see that there are 14 values true positive, and 4 values false negative.

Conclusions

- The dataset has been cleaned enough that bring simplicity to the exploration, but the small size of dataset can be an obstacle for the model to learning its pattern and generalizing it.
- The distribution of the data is not normal-distributed enough, because there is not equivalent count of each value on a category that may bring to less accurate results because the model doesn’t have the same number of data from each class as the learning media
- Eventhough the evaluation results show a very high accuracy score, which is 100% on the 3 different models of classification, but it doesn’t rule out the possibility if there will be a miss-prediction later on, because of the limitation of provided dataset and records.

Appendix

Used Wikipedia's Dataset :
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Full Version of Python Source Code :
https://colab.research.google.com/drive/1t6AAdDuHLeohyWQdqF4tSNu5zVCdO0Fm?usp=sharing

Thank you!