James Yu

IBM DATA SCIENCE

Space X Falcon 9 Landing Analysis

OUTLINE

EXECUTIVE SUMMARY

Summary of Methodologies

This Project follows these steps:

I. Data Collection
II. Data Wrangling
III. Exploratory Data Analysis
IV. Data Visualization
V. Interactive Visual Analytics
VI. Predictive Analysis (Classification)

Summary of Results

This Project produced the following outputs and visualizations:

I. EDA results
II. Geospatial analysis
III. Interactive dashboard
IV. Predictive analysis of classification models.

INTRODUCTION

- SpaceX launches Falcon 9 rockets at a cost around $62m. It is considerably cheaper than other providers, which costs around $165m. The major reason that lies behind the scene is that SpaceX can land and re-use the first stage of the rocket.

- Thus, if we are able to tell whether a first stage rocket can land, we may help an alternative company to bid SpaceX for a rocket launch.

- This Project will ultimately predict if the SpaceX Falcon9 first stage rocket will land successfully.

METHODOLOGY

METHODOLOGY

1. Data Collection
2. Data Wrangling
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

DATA COLLECTION

- How the data are collected

Through a RESTful API provided by SpaceX, make a get request in order to collect data. First, define several functions that extract information from the launch data, then request data form the SpaceX API url. Decode the response content from SpaceX into JSON and then convert into DataFrame.

The complete notebook for SpaceX API is here.

There is also another data collect method that is using web scraping from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches, those records are store in HTML. Utilizing Beautiful soup and request libraries. I can also extract the table record in HTML and covert them into DataFrame.

The complete notebook for web scraping is here.

DATA WRANGLING

- Data Process and cleaning

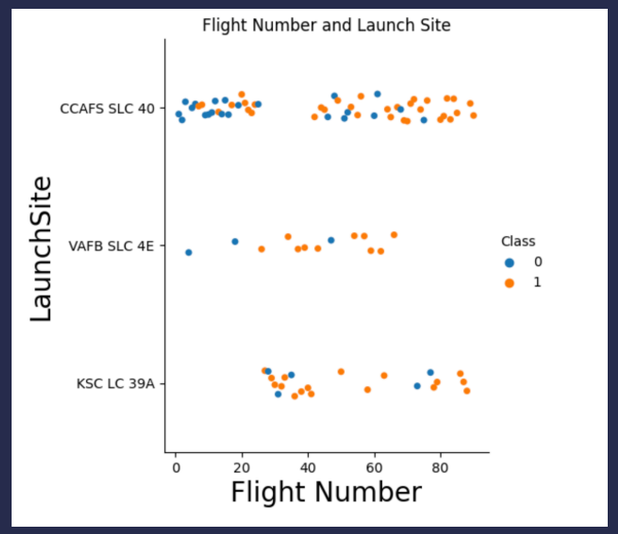
With the DataFrame, we will be able to filtered those data with specific BoosterVersion. Fixing the BoosterVersion on Falcon 9, then deal with the missing values in LandingPad and PayloadMass columns.

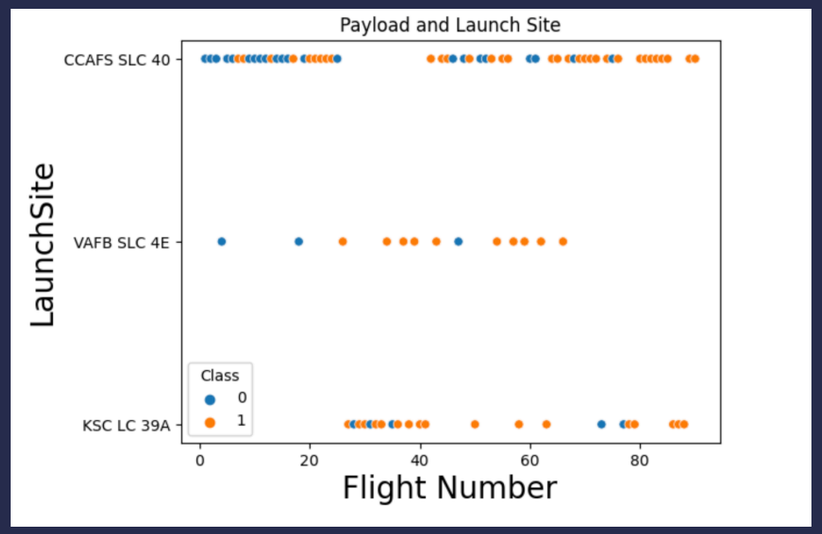
The complete notebook of data wrangling is here.

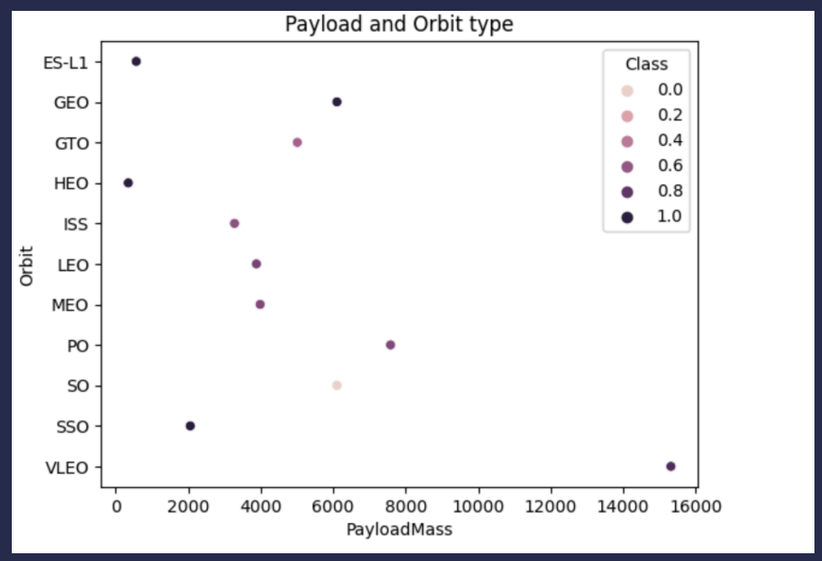
EDA WITH DATA VISUALIZATION

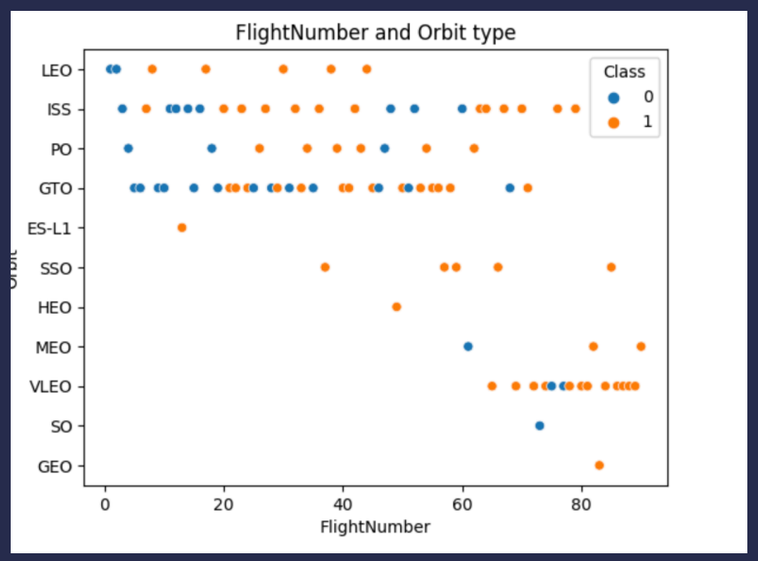
- Perform data Analysis and Feature Engineering using Pandas and Matplotlib

Use Scatter plot to visualize the relationship between Flight Number and Launch Site, Payload and Launch Site, FlightNumber and Orbit type, Payload and Orbit type.







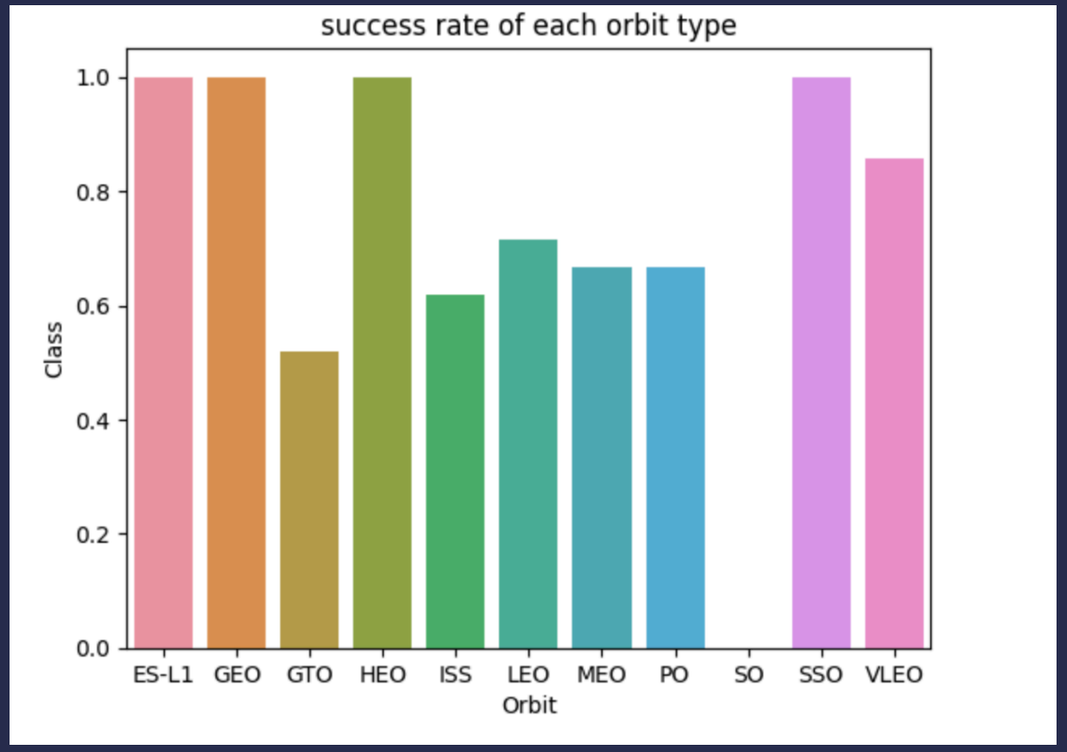


EDA WITH DATA VISUALIZATION

- Perform data Analysis and Feature Engineering using Pandas and Matplotlib

  Use Bar chart to visualize the relationship between success rate of each orbit type.

  Use Line plot to visualize the launch success yearly trend.



EDA WITH SQL

- Perform data Analysis and Feature Engineering using SQL

After attaching the mysql database, we are able to handle these data and use sql query to manage, filter and display those content.

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select Distinct Booster_version, PAYLOAD_MASS__KG_ from "SPACEXTBL" where "Landing_Outcome" = "Success (drone ship)"and
```

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.

```
%sql select COUNT(LANDING_OUTCOME), Landing_OUTCOME from 'SPACEXTBL' where Date between "04/06/2010" and "20/03/2017" group
```

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select Booster_version, PAYLOAD_MASS__KG_ from 'SPACEXTBL' where PAYLOAD_MASS__KG_ = (select Max(PAYLOAD_MASS__KG_) frc
```

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

INTERACTIVE MAP WITH FOLIUM

- Create interactive map with Folium and apply markers, circles and lines to mark the launch sites.
- Create a launch set outcomes
- The complete notebook for Folium map is here.

DASHBOARD WITH PLOTLY DASH

- Build an interactive dashboard application with poorly dash, which should includes dropdown list, callback function, range slider and plot.
- The complete notebook for dashboard is here.

PREDICTIVE ANALYSIS

Build, evaluate, improve and found the best performed model.

I. With Data as Pandas dataframe, we are able to apply data analysis by using Numpy to assign column as variable Y.
II. With the preprocessing.StandardScaler() function from Sklearn, we can standardize the feature dataset(X)
III. Split the data into training and testing sets using function train_test_split from sklearn.model_selection with the test_size parameter set to 0.2 and random_state to 2

PREDICTIVE ANALYSIS

I. Select the best performed model in between different machine learning algorithms.

II. First, create an object for each algorithms then create a GridSearchCV object and assign them a set of parameters.

III. Second, for each model under evaluation, GridsearchCV object was created with cv=10, then fit the training data into the object to find the best hyperparameters.

IV. Output the GridsearchCV to observe the best parameters, accuracy and score through data attribute.

V. Use score to calculate the accuracy on testing data, and build a confusion matrix for each algorithms.

RESULTS


LIFTOFF
STARTUP
MAX-Q
MECO
FAIRING
BOOSTBACK
T+00:00:06
TRANSPORTER-8

RESULT

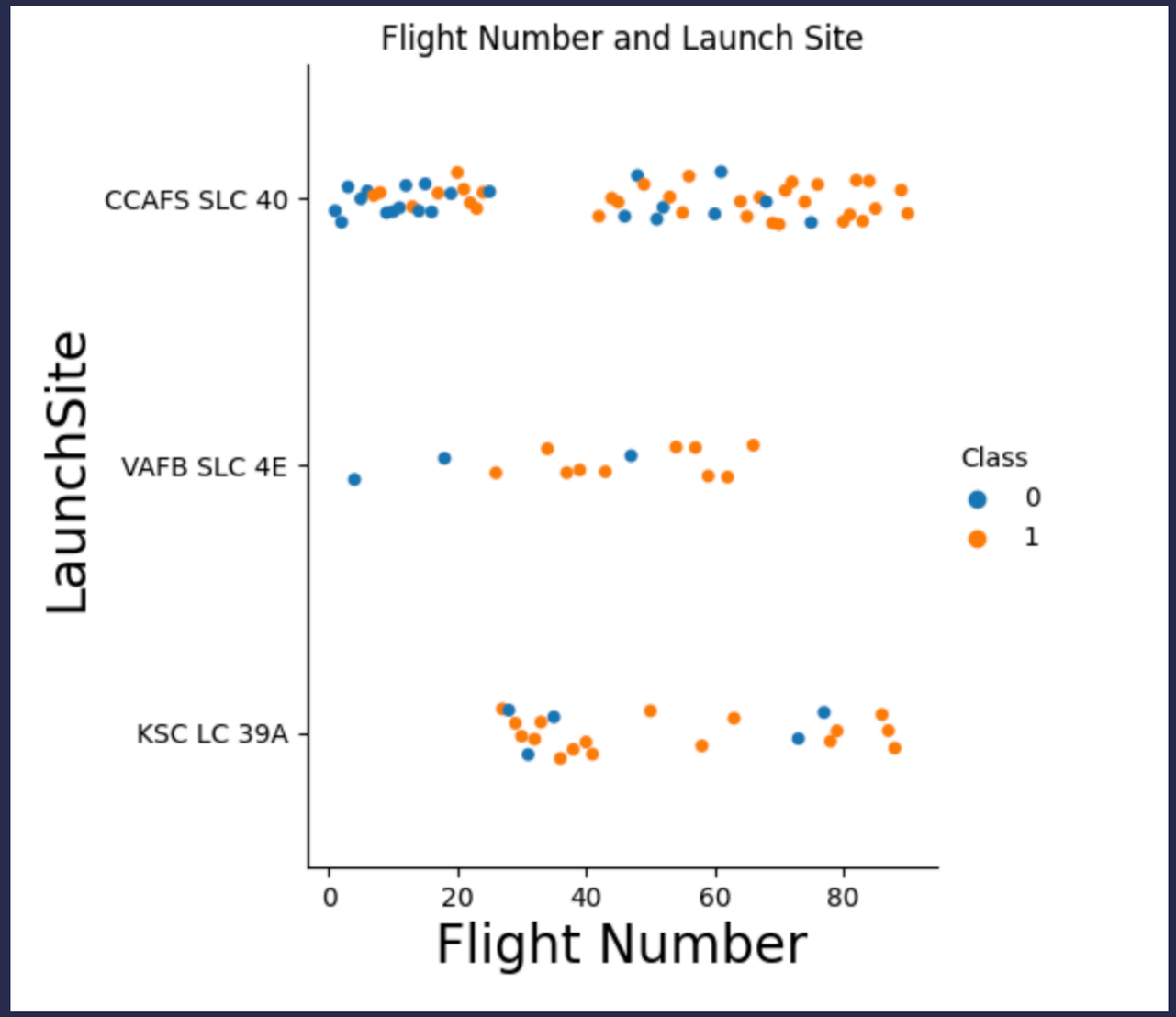
1. Exploratory Data analysis results (EDA)
2. Interactive analysis demo with screenshots
3. Launch Site Proximities Analysis
4. Predictive analysis results

EDA RESULTS WITH PLOTS

- Flight numbers v.s. Launch Site

Use Scatter plot to visualize the relationship between Flight Number and Launch Site.

We can deduce that the success rate of each launch site raises as the number of the flight. For instance, VAFB SLC 4E attend 100% success rate after the flight number reaches 50, and two other sites also reach 100% at the number of 80.

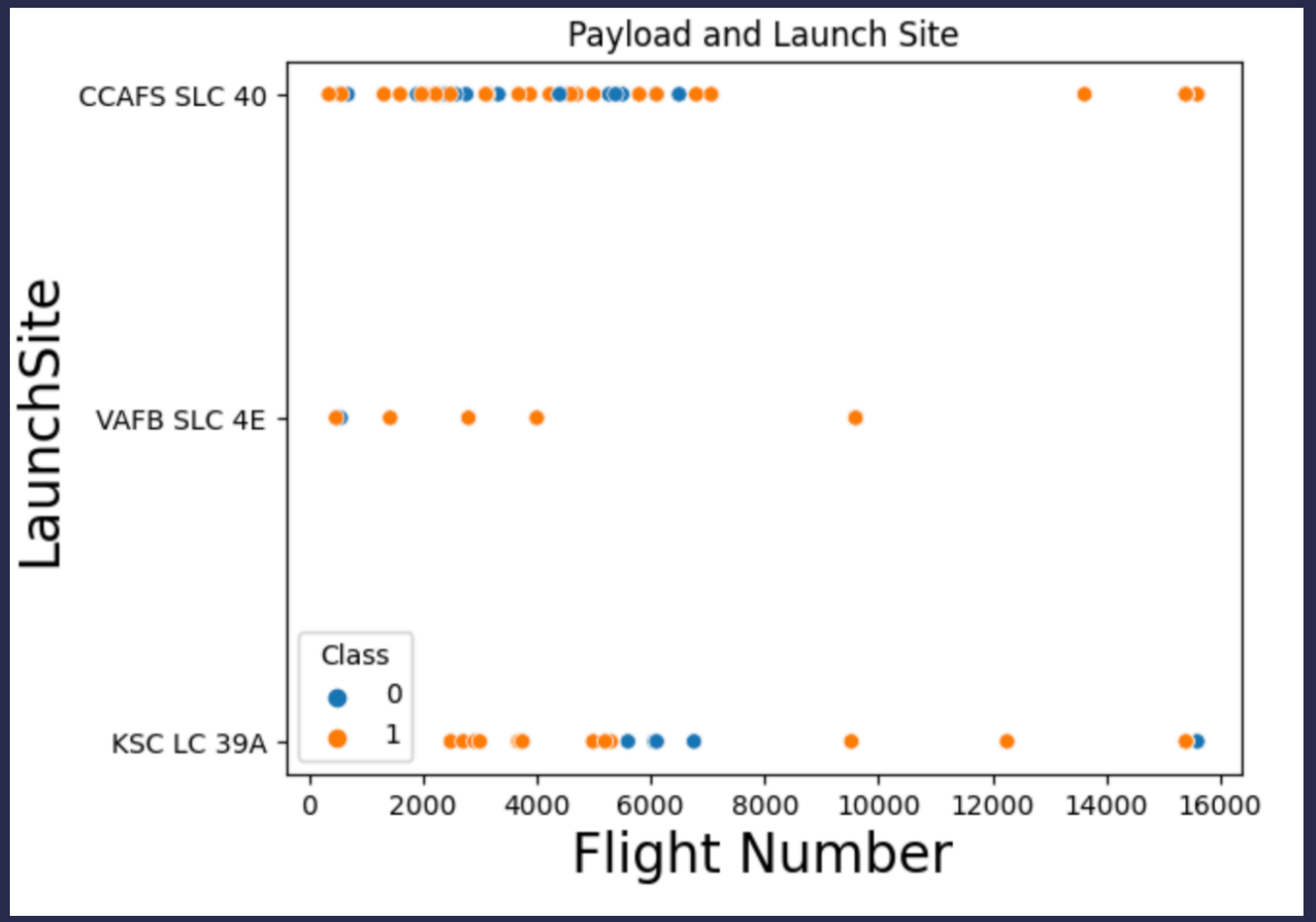


EDA RESULTS WITH PLOTS

- Payload v.s. Launch Site

Use Scatter plot to visualize the relationship between Payload and Launch Site.

We can observe that there are no heavy payload rocket launch in the launch site VAFB SLC 4E



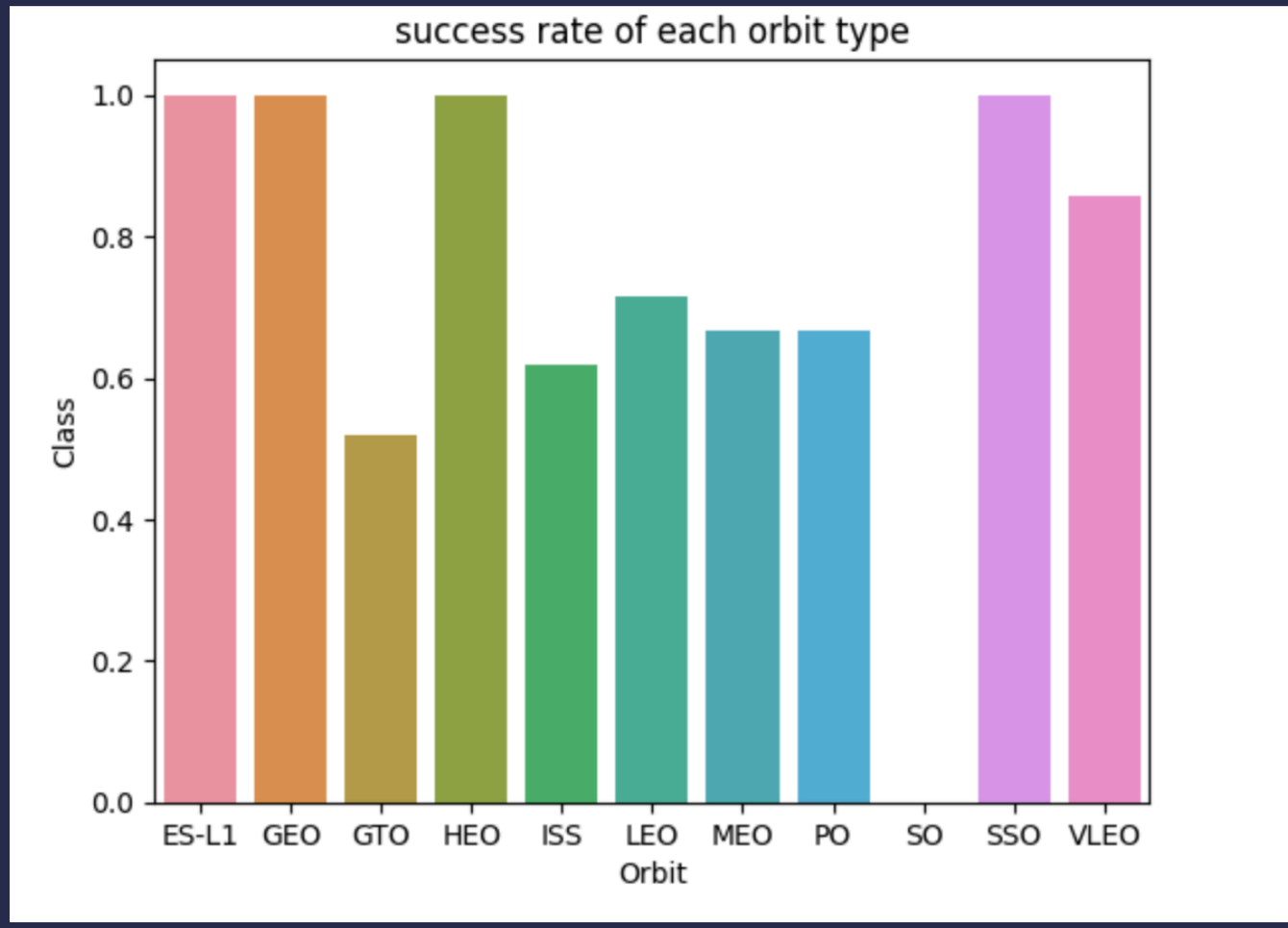
説明

EDA RESULTS WITH PLOTS

- Success Rate v.s. Orbit Type with Explanations

Use bar plot to visualize the relationship and provide explanations

Orbit SSO, HEO, GEO and ES-L1 have 100% success rate while Orbit SO has 0%.

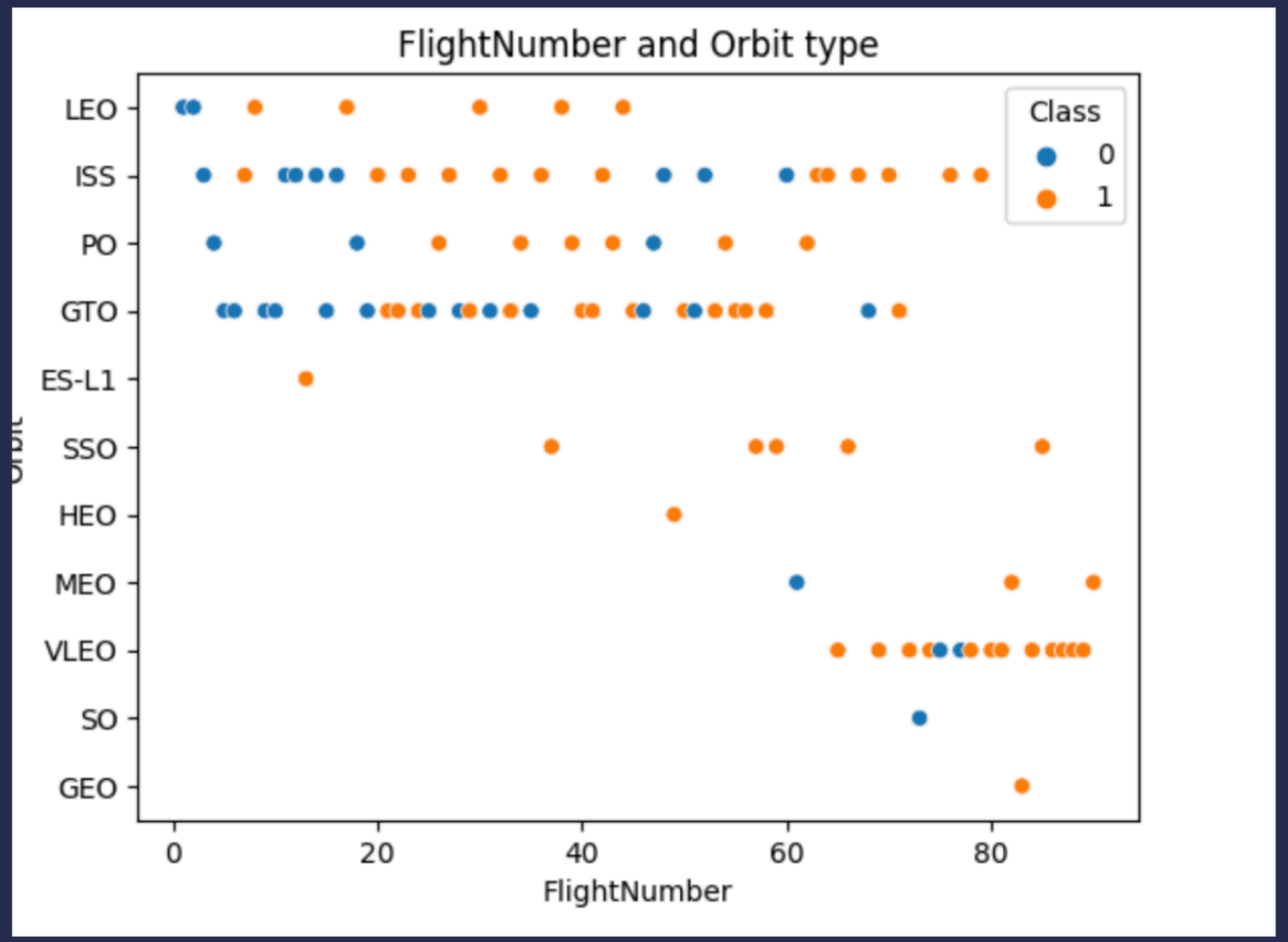


EDA RESULTS WITH PLOTS

- Flight Numbers v.s. Orbit Type with Explanations

Use Scatter plot to visualize the relationship and provide explanations

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



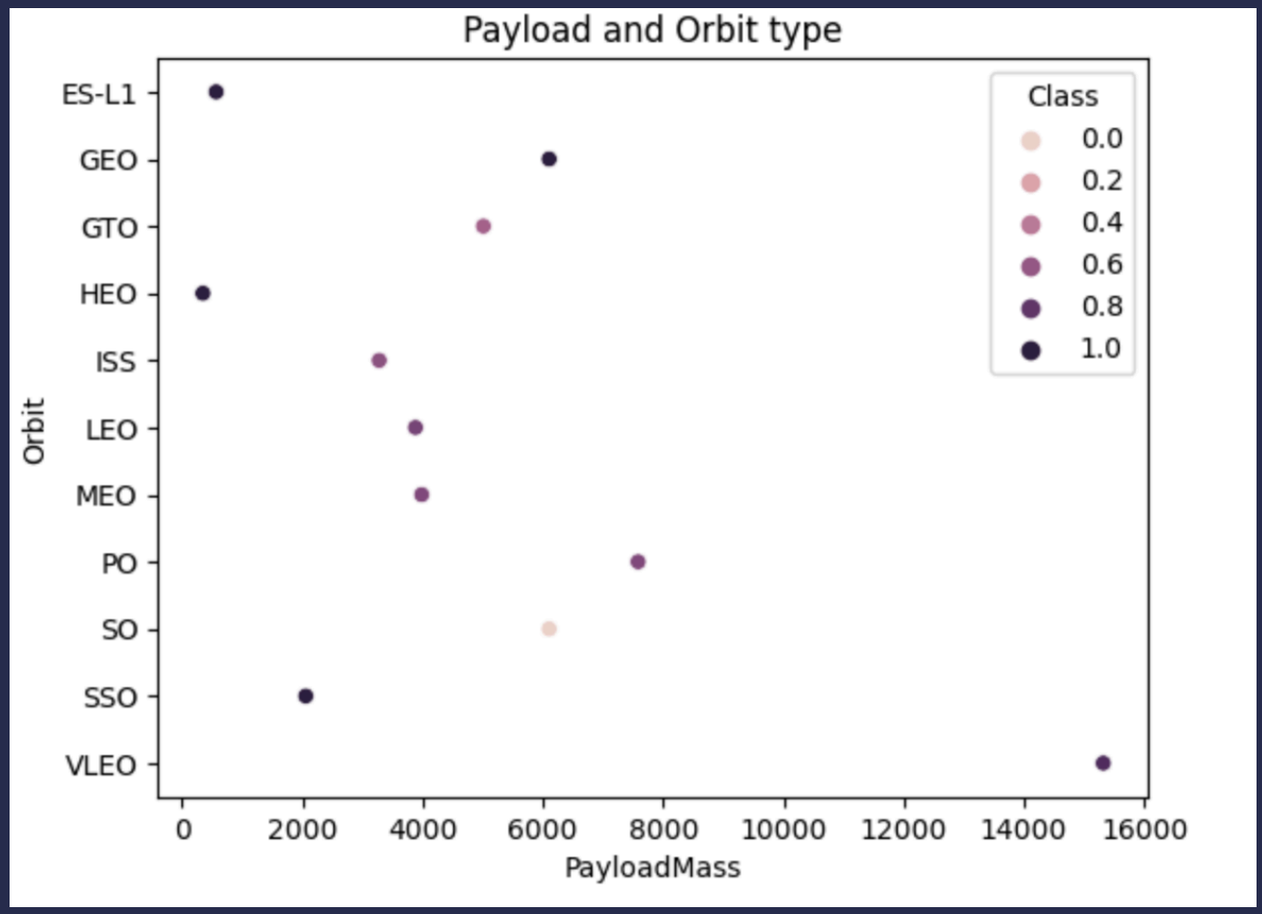
EDA RESULTS WITH PLOTS

- Payload v.s. Orbit Type with Explanations

Use Scatter plot to visualize the relationship and provide explanations

With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

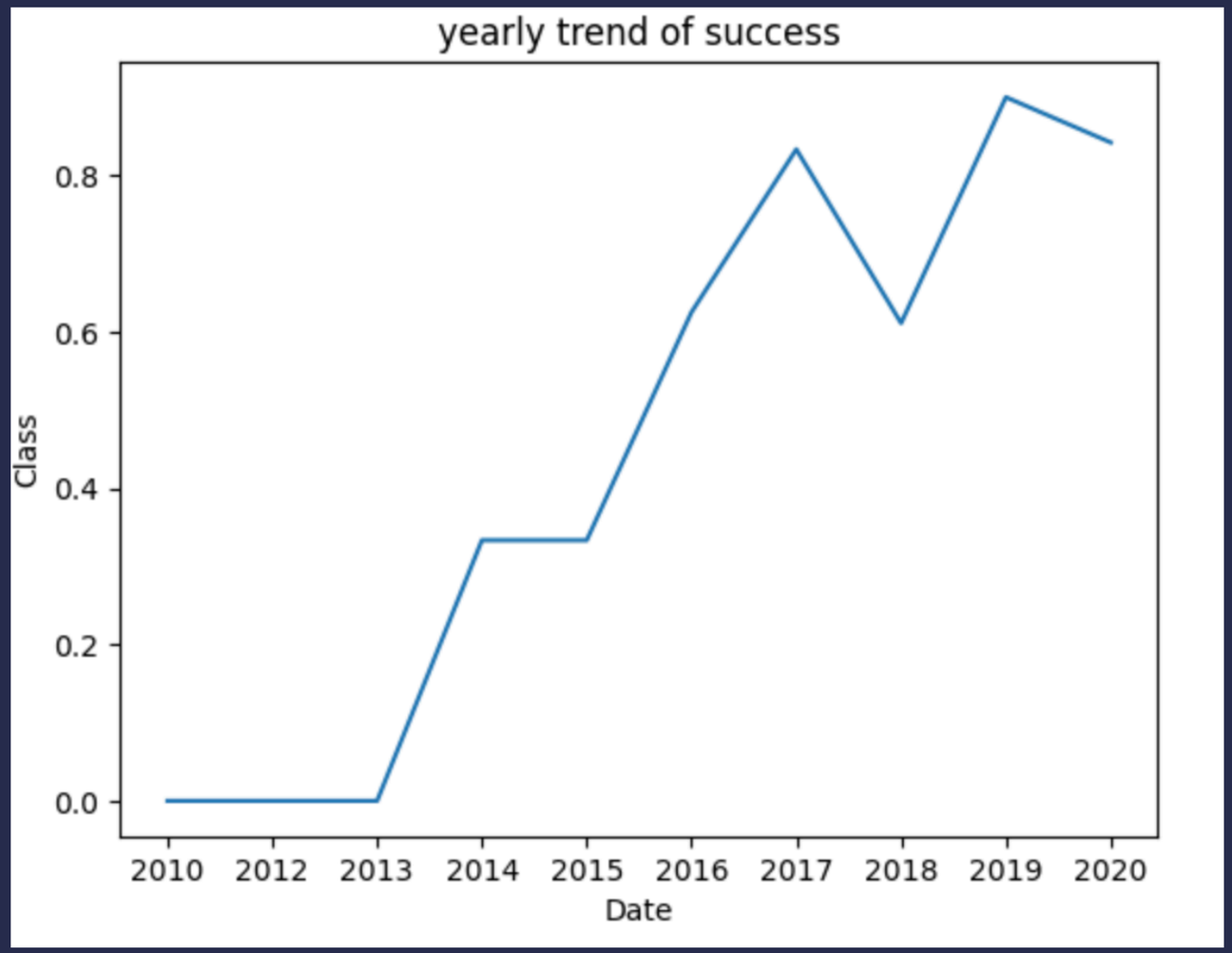
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) both have near equal chances.



EDA RESULTS WITH PLOTS

- Launch Success Yearly Trend

Since 2013, the success rate kept going up till 2020.



INTERACTIVE ANALYSIS WITH SCREENSHOTS

- All Launch Site Names

Display the names of the unique launch sites in the space mission

```
%sql select Distinct(Launch_Site) From SPACEXTABLE
```

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

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

To find all the unique launch site, we can use the sql DISTINCT function.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- 5 records where launch sites begin with `CCA`

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * From SPACEXTABLE Where Launch_site like 'CCA%' Limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Use 'Like' and % operands to help SQL select the data start with specific word.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Calculate and Display the total payload carried by boosters from NASA

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select SUM(Payload_Mass__KG_) as 'Total Payload Mass', Customer From SPACEXTABLE Where Customer = 'NASA (CRS)'
```

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

Total Payload Mass	Customer
45596	NASA (CRS)

Use ‘SUM’ Function to return the total Payload Mass where customer is NASA.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Average Payload Mass by F9 v1.1

```
%sql select AVG(Payload_Mass__KG_), booster_version From SPACEXTABLE Where booster_version ='F9 v1.1'
```

Average Payload Mass by F9 v1.1

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

AVG(Payload_Mass__KG_)	Booster_Version
2928.4	F9 v1.1

Use 'AVG' function to calculate average Payload Mass with specific booster version.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select MIN(Date) From SPACEXTABLE Where Landing_Outcome is 'Success (ground pad)'
```

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

MIN(Date)
2015-12-22

Use 'MIN' function to get the smallest date of successful landing.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Successful Landing Booster with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
select Distinct(Booster_version) From SPACEXTABLE
Where Landing_Outcome is 'Success (drone ship)' and Payload_Mass__KG_ between 4000 and 6000
```

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

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

Use 'Between 'function get the data with payload mass in range 4000 to 6000 kg.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
%sql select Mission_outcome, Count(Mission_outcome) From SPACEXTABLE group by Mission_outcome
```

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

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

Use 'Count ' and 'Group by' function to get the data in Group.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
select Distinct(Booster_version), Payload_Mass__KG_ From SPACEXTABLE
Where Payload_Mass__KG_ = (Select MAX(Payload_Mass__KG_) from SPACEXTABLE)
```

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

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

Use subquery to get maximum payload then get the booster version that carries them.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- 2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

```
%%sql
select  substr(Date, 6, 2) as 'Month',  substr(Date,0,5)as 'Year', Booster_version, Launch_site, Landing_Outcome
From SPACEXTABLE Where Landing_outcome = 'Failure (drone ship)'and substr(Date,0,5) = '2015'
```

[37] Python

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

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

Use ‘substr’ function to extract month and year from the date data.

INTERACTIVE ANALYSIS WITH SCREENSHOTS

- Rank of Landing Outcomes Count in Descending order.

Task 10

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.

```
%%sql
select Count(*) as "COUNT", Landing_outcome From SPACEXTABLE
Where Date Between '2010-06-04' and '2017-03-20' group by Landing_outcome order by Count(*) DESC
```

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

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

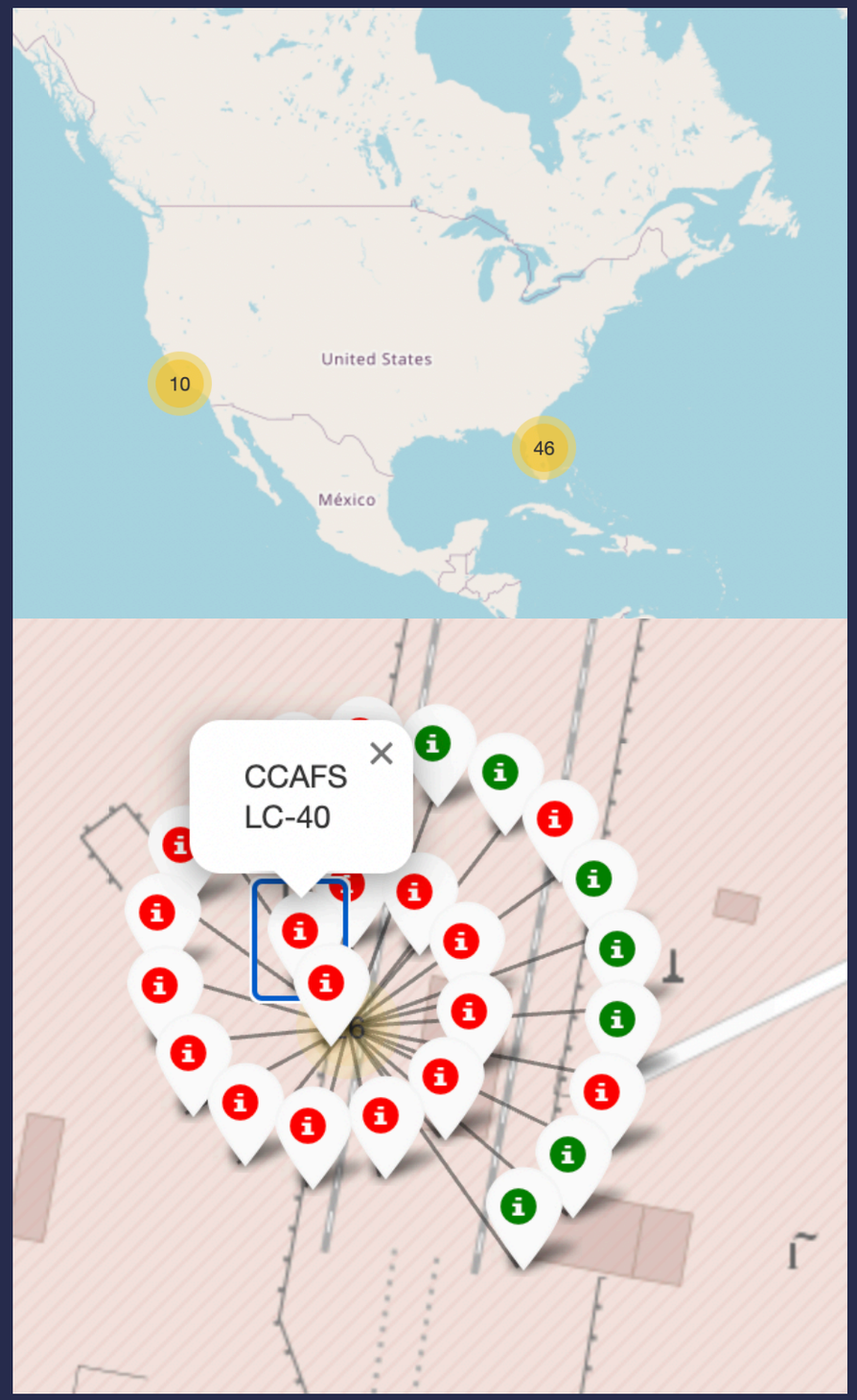
Use Group by, count, and order function to get the requested data.

LAUNCH SITE PROXIMITIES ANALYSIS

First, we mark out all the Launch sites on the map through folium marker.

Afterward, we will be able to classify each launch site by the outcome of each launch.

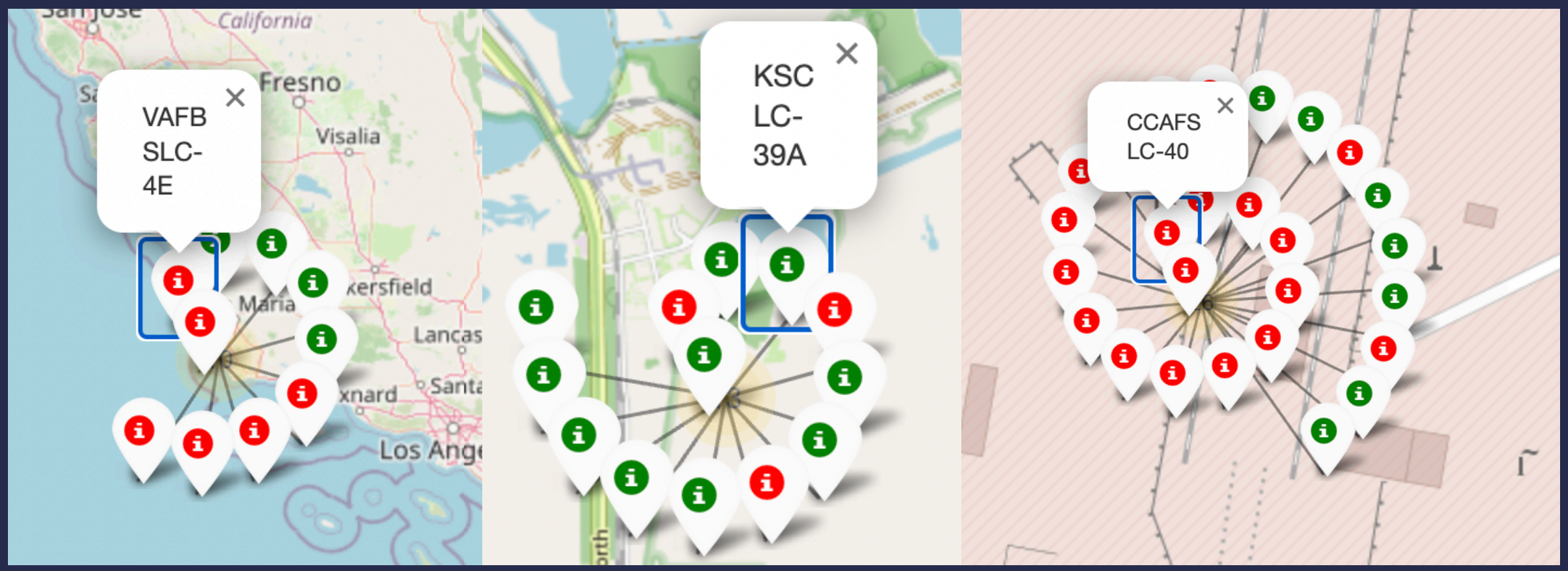
With the red indicates fails and green as success, we can analysis the relation between site and the outcome.



LAUNCH SITE PROXIMITIES ANALYSIS

According to the Picture below, we can deduce that the launch KSC LC-39A has a has relatively high success, while Launch site VAFB SLC-4E has relatively lower success rates.

Which comes to a conclusion that the east coast has higher success rate.

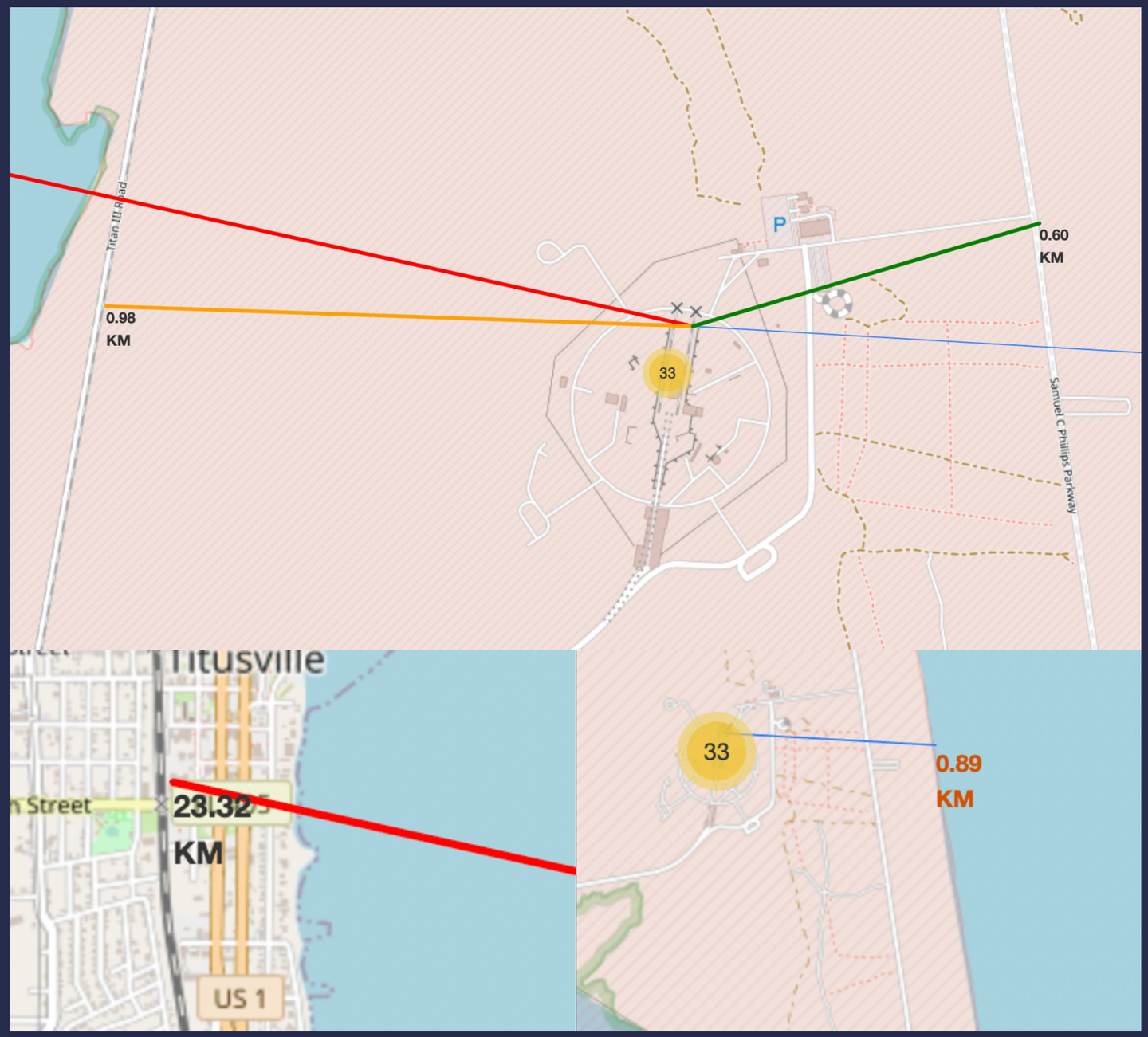


LAUNCH SITE PROXIMITIES ANALYSIS

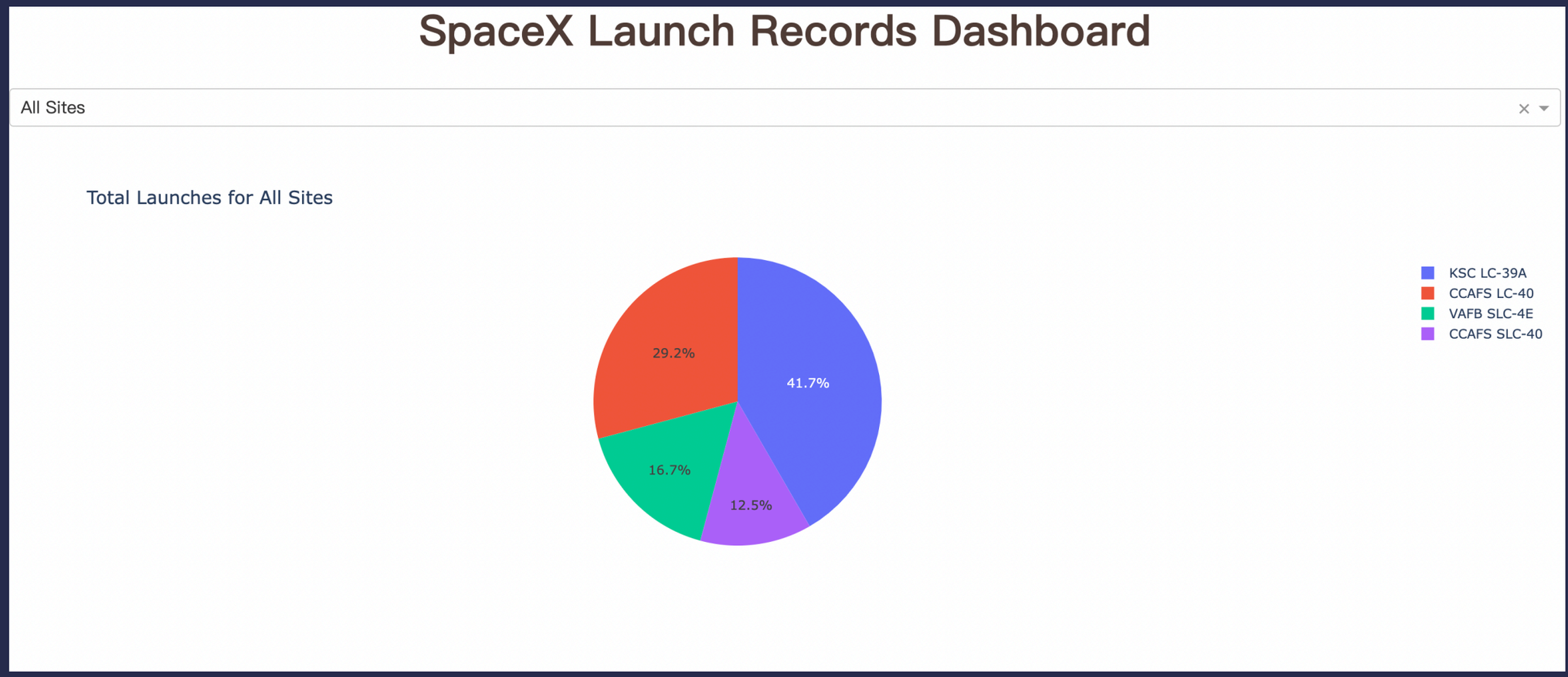
With the map we can calculate the distance between Launch site and different locations.

For instance, the Launch site CCAFS SLC-40 is 0.89 Km away from the coast line and 0.98 Km, 0.80 Km to railway and highway separately.

Titusville is approximately 23.32 km away from the launch site.

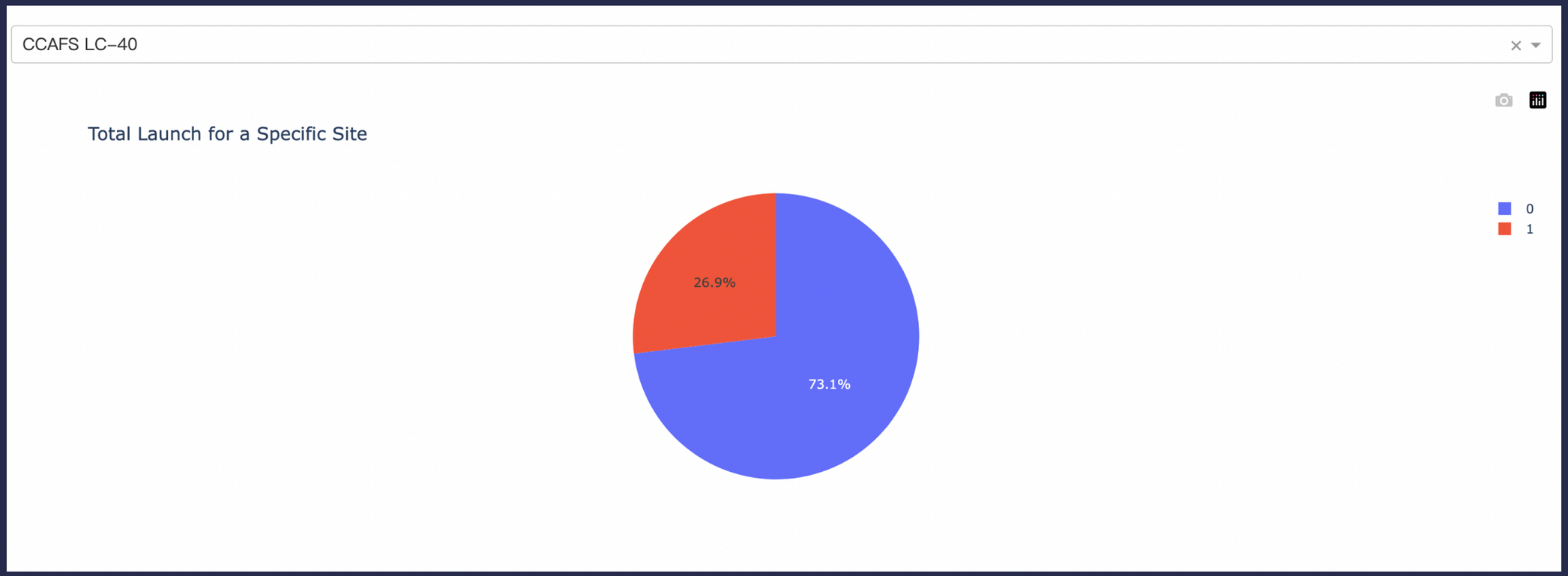


BUILD A DASHBOARD WITH PLOTLY DASH



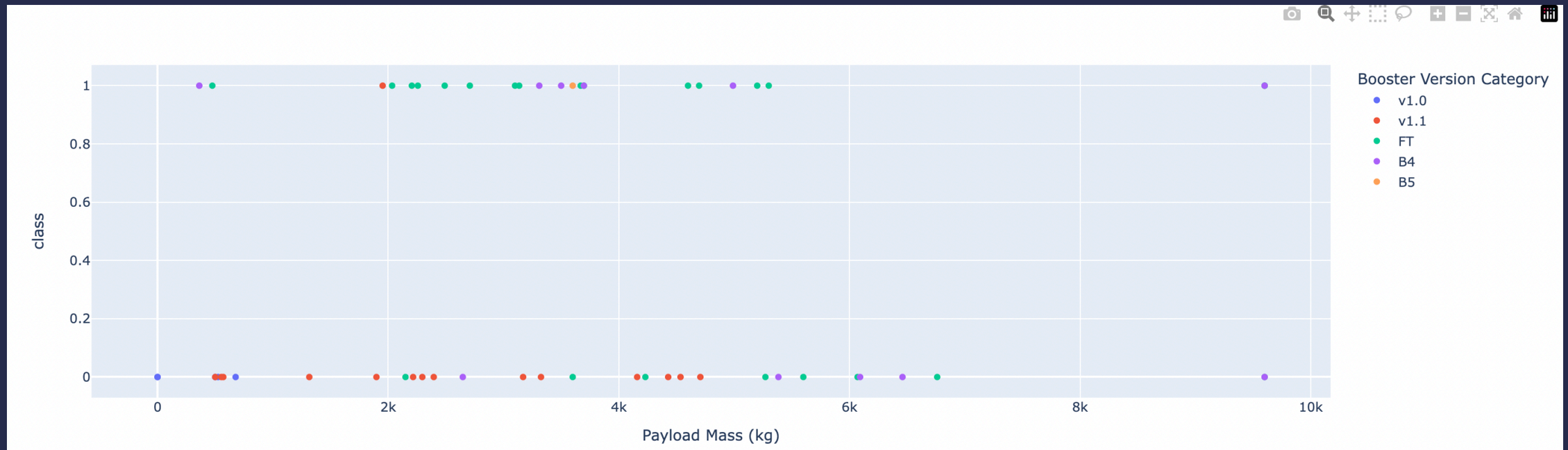
With Plotly Dashboard, we are able to build a interactive panel that can help visualize the project. For example, According to the pie chart, Launch site KSC LC-39A has the highest launch success rate at 42% followed by CCAFS LC-40 at 29%, VAFB SLC-4E at 17% and lastly launch site CCAFS SLC-40 with a success rate of 13%

BUILD A DASHBOARD WITH PLOTLY DASH



By changing the dropdown menu, we can pick the CCAFS LC-40 as an analysis target. Which exposed that Launch site CCAFS LC-40 had the 2nd highest success ratio of 73% success against 27% failed launches.

BUILD A DASHBOARD WITH PLOTLY DASH



With the Payload vs. Launch Outcome scatter plot, we may conclude that for Launch site CCAFS LC-40 the booster version FT has the largest success rate from a payload mass of >2000kg, and after the Payload exceed 6000kg, the success rate drops rapidly.

PREDICTIVE ANALYSIS RESULTS

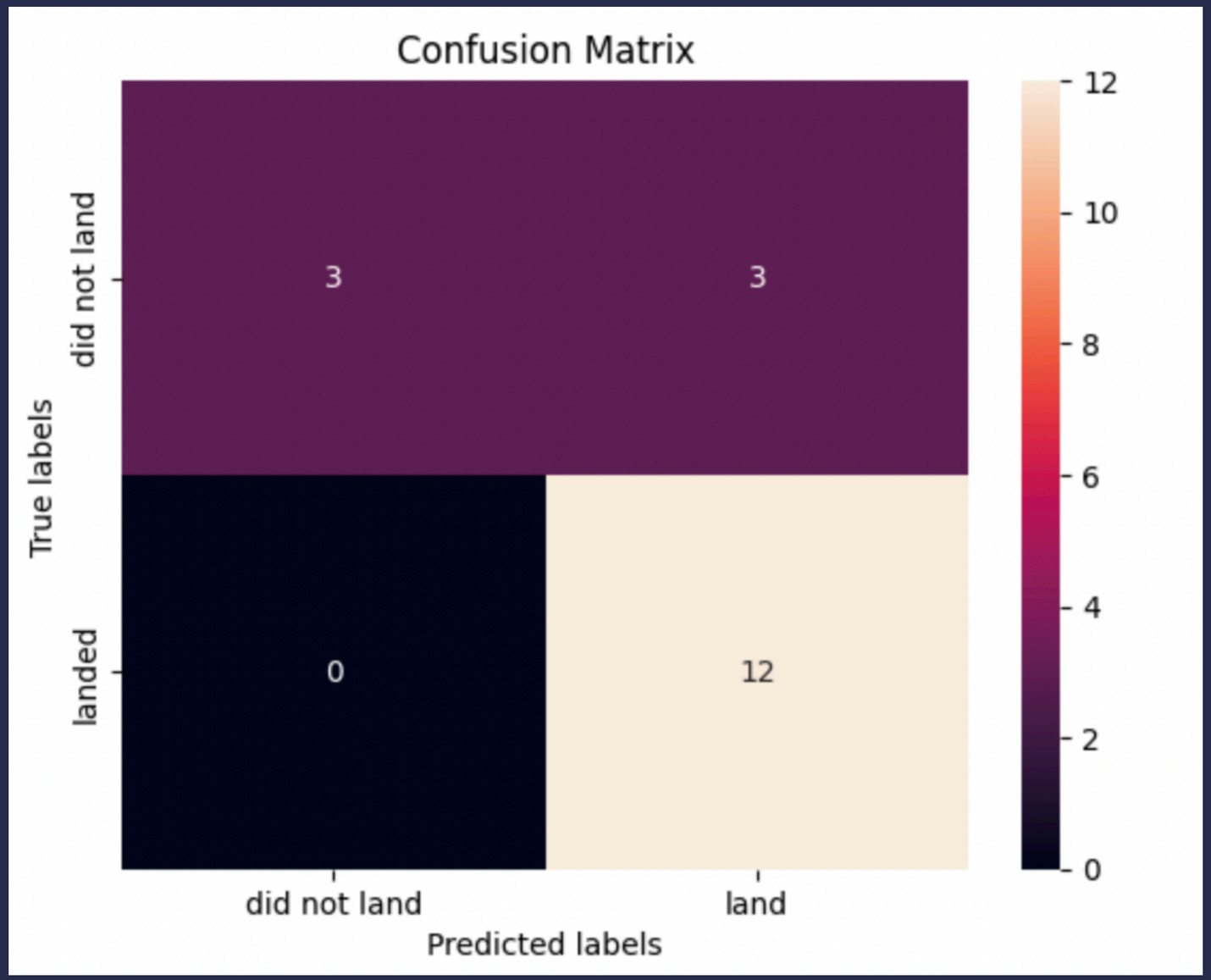
By applying Different classification model, we will be able to compare and pick the one with the best accuracy.

	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

PREDICTIVE ANALYSIS RESULTS

According to the following picture, all the 4 classification model had the same confusion matrixes and were able equally distinguish between the different classes. The major problem is false positives for all the models.



CONCLUSION


LIFTOFF
STARTUP
MAX-Q
MECO
BOOSTBACK
FAIRING
T+00:00:06
TRANSPORTER-8

CONCLUSIONS

- **Model Performance :** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Equator :** Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- **Coast** : All launches are close to coasts.
- **Launch Success**: Increases over time and number of launches.
- **KSC LC-39A**: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Orbits**: ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Payload Mass**: Across all launch sites, the higher the payload mass (kg), the higher the success rate

CONCLUSIONS (THINGS TO CONSIDER)

- **Dataset:** With a larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set.
- **Feature Analysis / PCA**: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- **XGBoost**: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models