Final Assigment - Applied Capstone

Joel Caristan Kamlo

Tuesday, May 24th, 2022

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- In this capstone, we will predict if the Falcon 9 first stage will land successfully, by using several machine learning methodologies.
- Main steps on this capstone include :
  - Data collection, wrangling, formatting
  - Exploratory Data Analysis
  - Data Visualization (Matplotlib & Dash)
  - Machine Learning
- The interactive dashboard allows us to know, which site has the largest successful launches
- Graphs obtained, show that some features of Falcon 9 rocket launches, have a correlation with the outcome of launches, ie, success or failure.
- We can conclude that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

INTRODUCTION

- In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.
- Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- In this capstone, we will collect and make sure the data is in the correct format from an API, then explore, analyse and prepare the data to answer the main question.
- The main questions to answer are :
  - For a given set of features of Falcon 9 rocket launches, will the first stage of the rocket land successfully ?
  - Are there a correalation between features of the Falcon 9 rocket and the outcome of launches (success or failure) ?

METHODOLOGY

- Data Collection, Wrangling and Formatting
  - SpaceX API
  - Web Scrapping
- Exploratory Data Analysis (EDA)
  - Pandas & Numpy
  - SQL
- Interactive Data Visualization
  - Matplotlib and Seaborn
  - Folium
  - Plotly Dash
- Machine Learning
  - Logistic Regression (LR)
  - Support Vector Machine (SVM)
  - Decision Tree
  - K-Nearest Neighbors (KNN)

METHODOLOGY

1. Data Collection, Wrangling and Formatting

- Web Scrapping
  - Data are scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - The website contains only data about Falcon 9 launches.
  - After filtering the data, we have 121 rows (instances) and 11

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	[SpaceX]	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	[NASA]	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	[NASA]	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	[NASA]	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[NASA]	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

METHODOLOGY

1. Data Collection, Wrangling and Formatting

- SpaceX API
  - API used is https://api.spacexdata.com/v4/launches/past
  - The API provides data for many types of rocket launches done by SpaceX, then data are filtered to include only Falcon 9's rocket launches.
  - Every missing value is replaced by the mean of the column, where the missing value is belongsto.
  - After filtering the data, we have 90 rows (instances) and 17 columns (features)
- Data is later processed, so that there are no missing entries and categorical features are encoded using one-hot encoding
- An extra column called 'Class' is also added to the dataframe. The column class contains 0 if a given launch is failed and 1 if it's successful
- At the end we have 90 rows and 83 columns

METHODOLOGY

1. Data Collection, Wrangling and Formatting

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	Reus
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
89	86	2020-09-03	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
90	87	2020-10-06	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
91	88	2020-10-18	Falcon 9	15600.0	VLEO	KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	
92	89	2020-10-24	Falcon 9	15600.0	VLEO	CCSFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	
93	90	2020-11-05	Falcon 9	3681.0	MEO	CCSFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	

90 rows × 17 columns

METHODOLOGY

2. Exploratory Data Analysis (EDA)

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
    - The number of launches on each launch site
    - The number of occurrence of each orbit
    - The number and occurrence of each mission outcome
- SQL
  - The data is queried using SQL to answer several questions about the data such as:
    - The names of the unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1

METHODOLOGY

3. Data Visualization

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type
- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
    - Mark all launch sites on a map
    - Mark the succeeded launches and failed launches for each site on the map
    - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

METHODOLOGY

3. Data Visualization

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site

METHODOLOGY

4. Machine Learning

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Find the best combination of hyperparameters for each model
  - Evaluate the models based on their accuracy scores and confusion matrix

RESULTS

- Results are splitted into five(05) sections :
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis

RESULTS

1. SQL (EDA with SQL)

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

[15]:

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

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

[16]:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

RESULTS

1. SQL (EDA with SQL)

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

```
[20]: TOTAL PAYLOAD MASS FOR NASA (CRS)
```

TOTAL PAYLOAD MASS FOR NASA (CRS)
45596

- Display average payload mass carried by booster version F0v1.1

```
[21]: AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9v1.1
```

AVERAGE PAYLOAD MASS CARRIED BY BOOSTER VERSION F9v1.1
2928

- List the date when the first successful landing outcome in ground pad was acheived

```
[53]: DATE OF THE FIRST SUCCESSFUL LANDING OUTCOME IN GROUND PAD
```

DATE OF THE FIRST SUCCESSFUL LANDING OUTCOME IN GROUND PAD
2015-12-22

- List the total number of successful and failure mission outcomes

```
[30]: TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES
```

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES
101

RESULTS

1. SQL (EDA with SQL)

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

[49]:

NAMES OF THE BOOSTERS SUCCESS IN DRONE SHIP
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- 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

[56]:

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

RESULTS

1. SQL (EDA with SQL)

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

BOOSTER VERSION	LAUNCH SITE	LANDING OUTCOME
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- List the names of the booster_versions which have carried the maximum payload mass

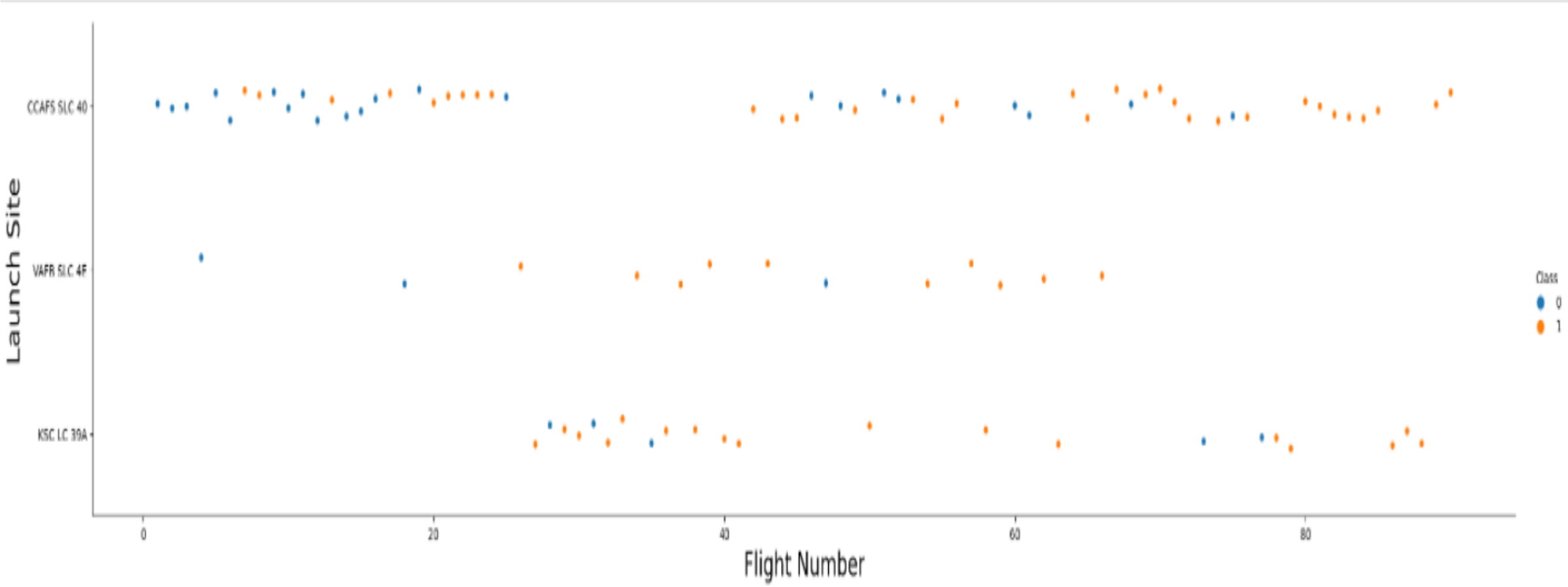
[41]:

NAMES OF THE BOOSTER VERSION WHICH THE MAXIMUM PAYLOAD MASS
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

RESULTS

2. Matplotlib & Seaborn (EDA with Visualization)

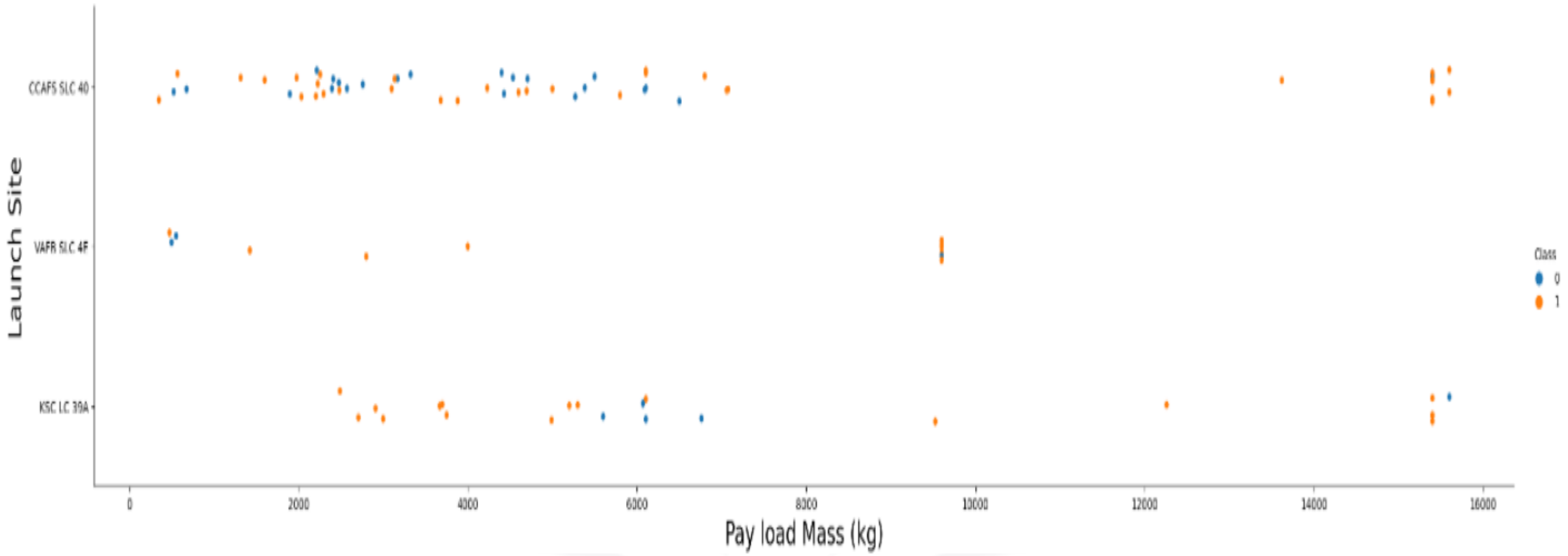
- Visualize the relationship between Flight Number and Launch Site



RESULTS

2. Matplotlib & Seaborn (EDA with Visualization)

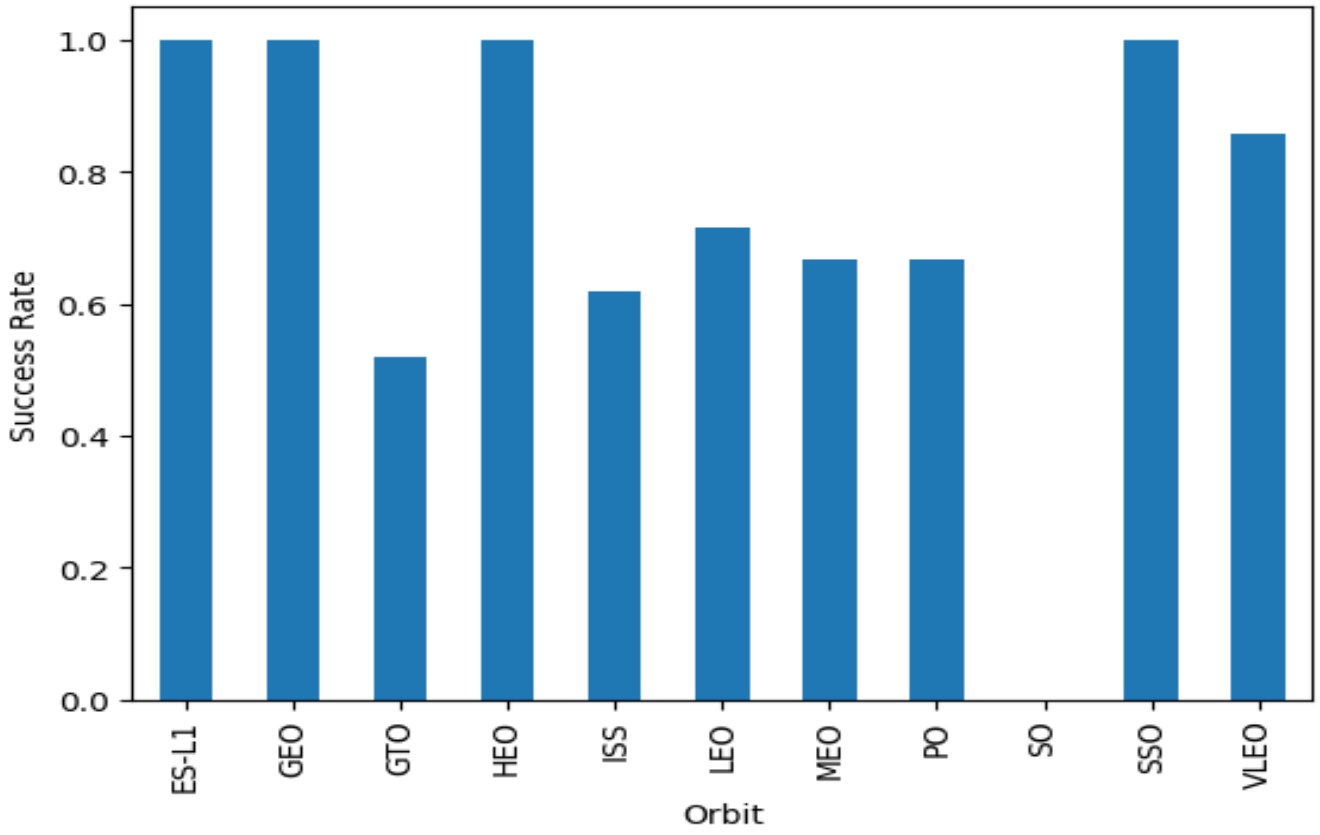
- Visualize the relationship between Payload and Launch Site



RESULTS

2. Matplotlib & Seaborn (EDA with Visualization)

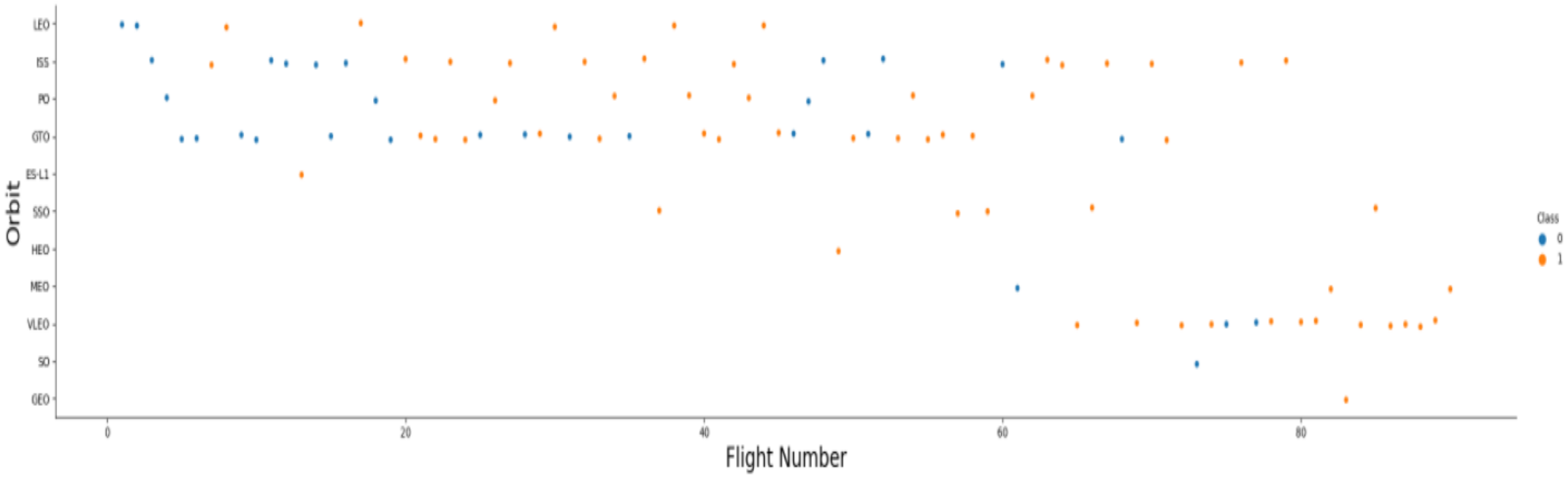
- Visualize the relationship between success rate of each orbit type



RESULTS

2. Matplotlib & Seaborn (EDA with Visualization)

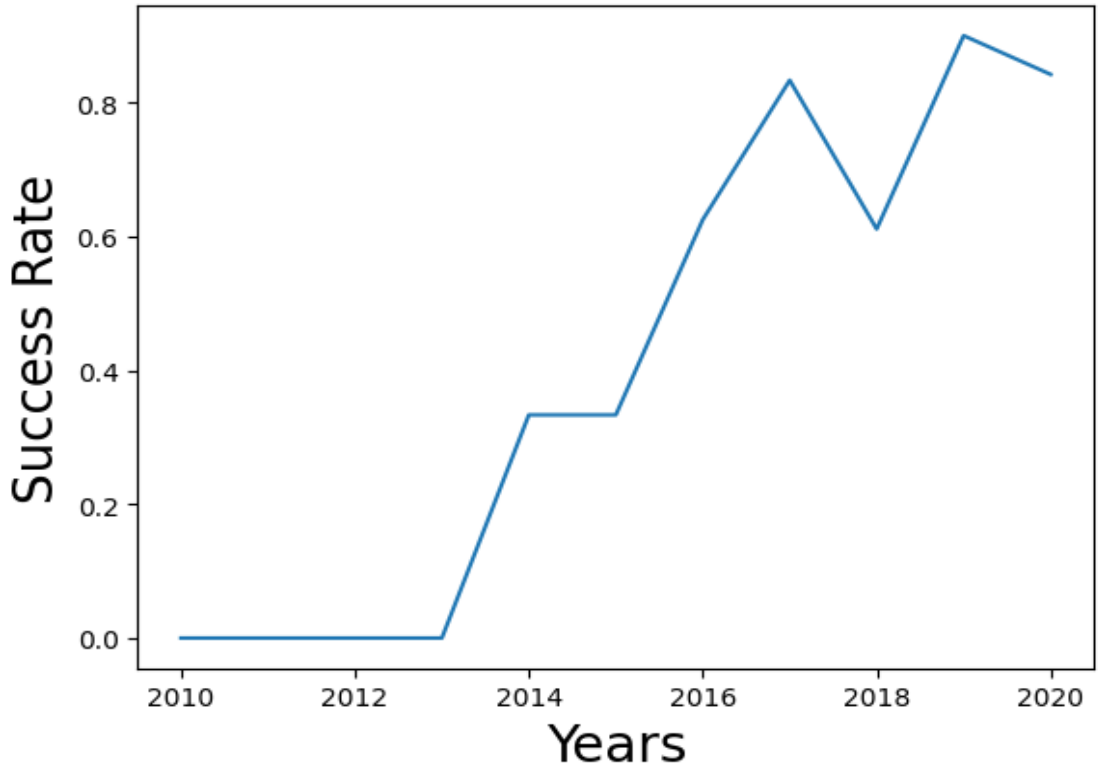
- Visualize the relationship between Flight Number and Orbit type



RESULTS

2. Matplotlib & Seaborn (EDA with Visualization)

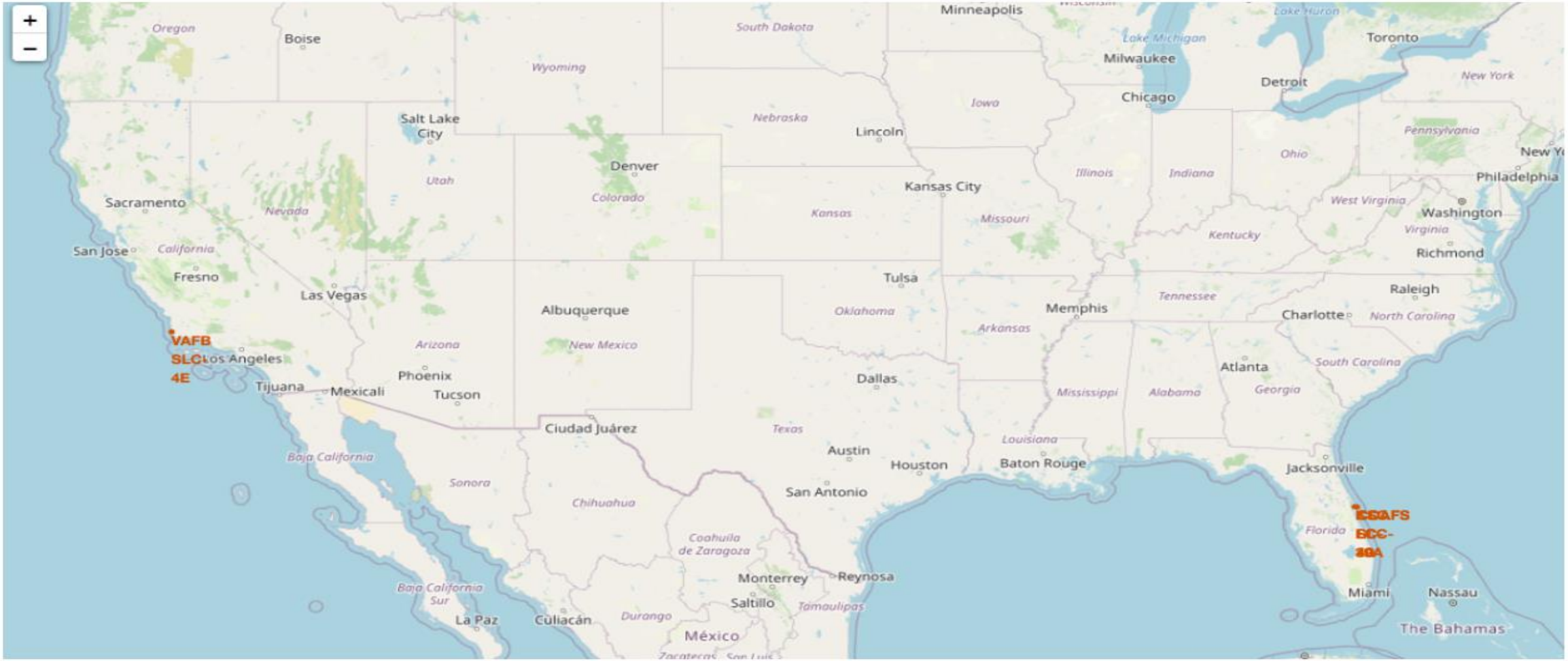
- Visualize the launch success yearly trend



RESULTS

3. Folium

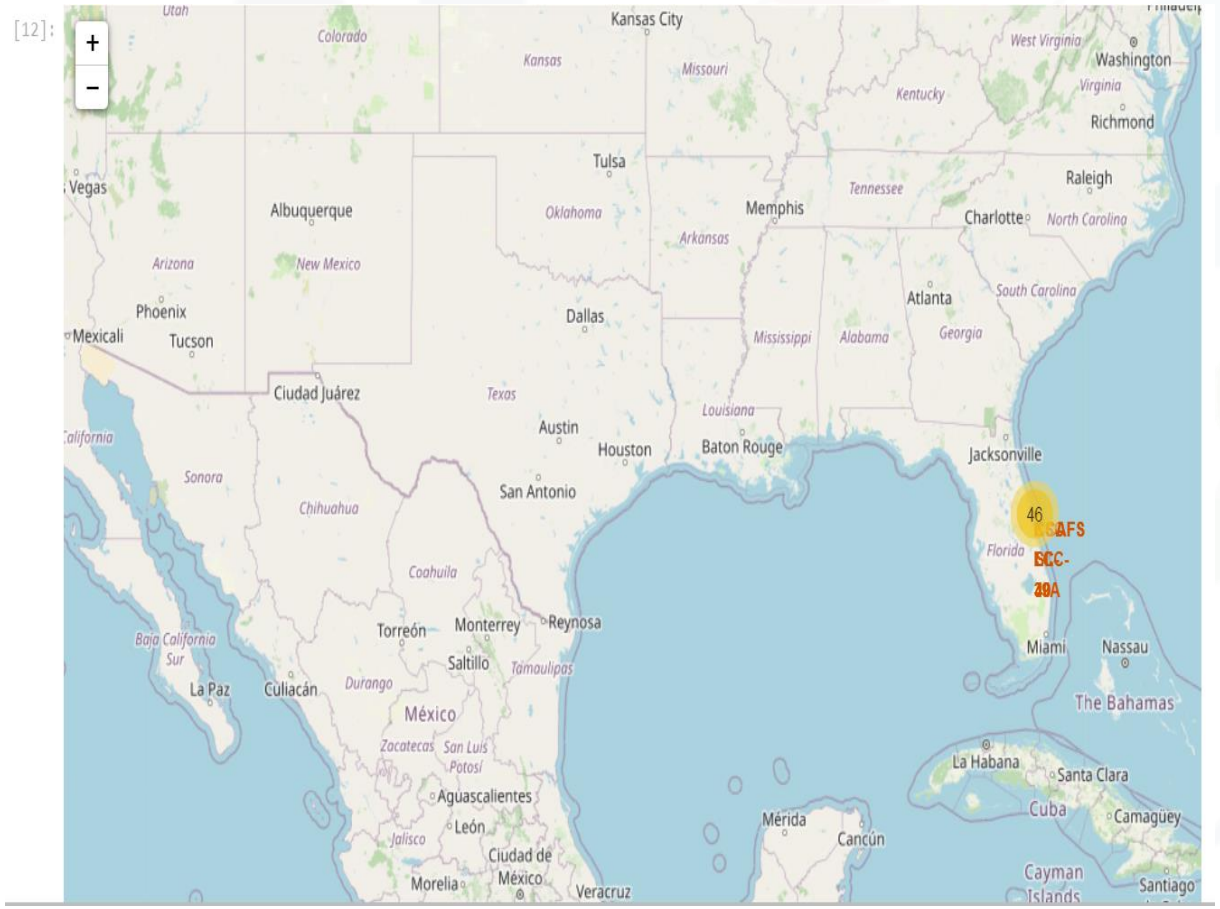
- All sites on the map

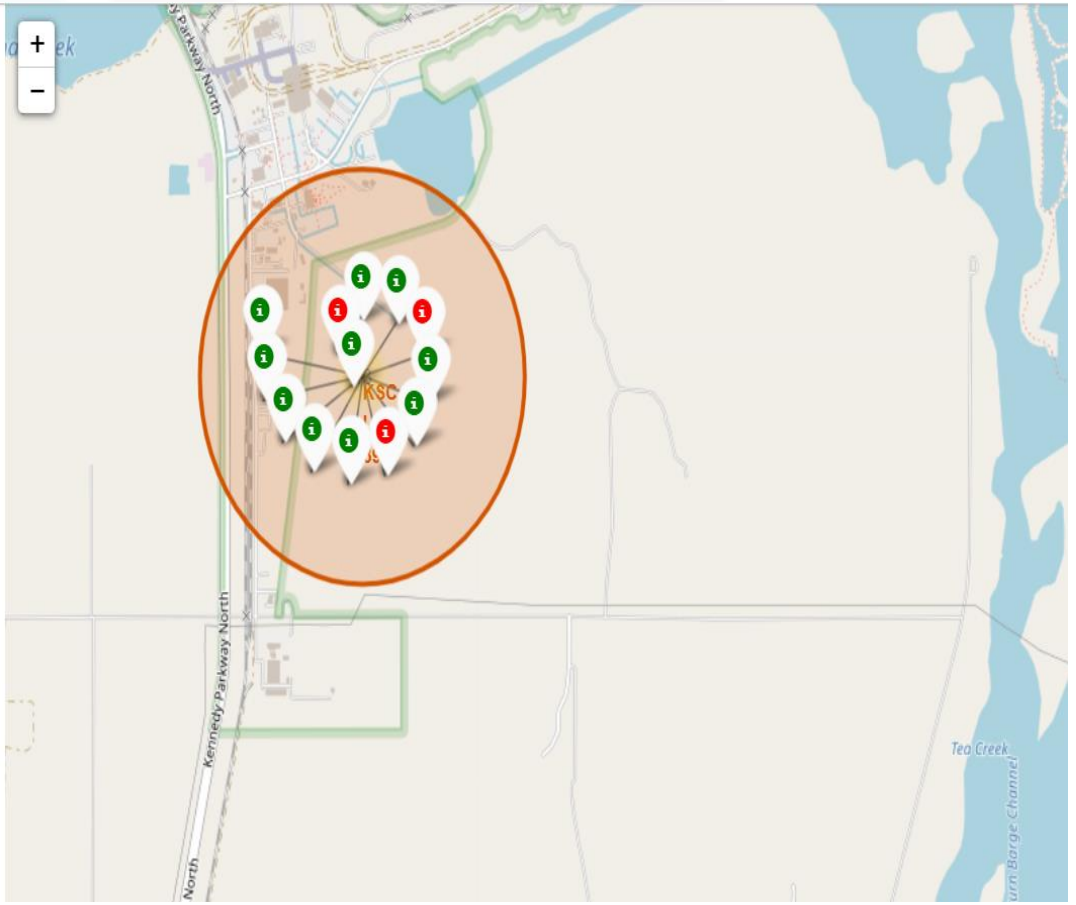


RESULTS

3. Folium

❖ Mark the success/failed launches for each site on the map

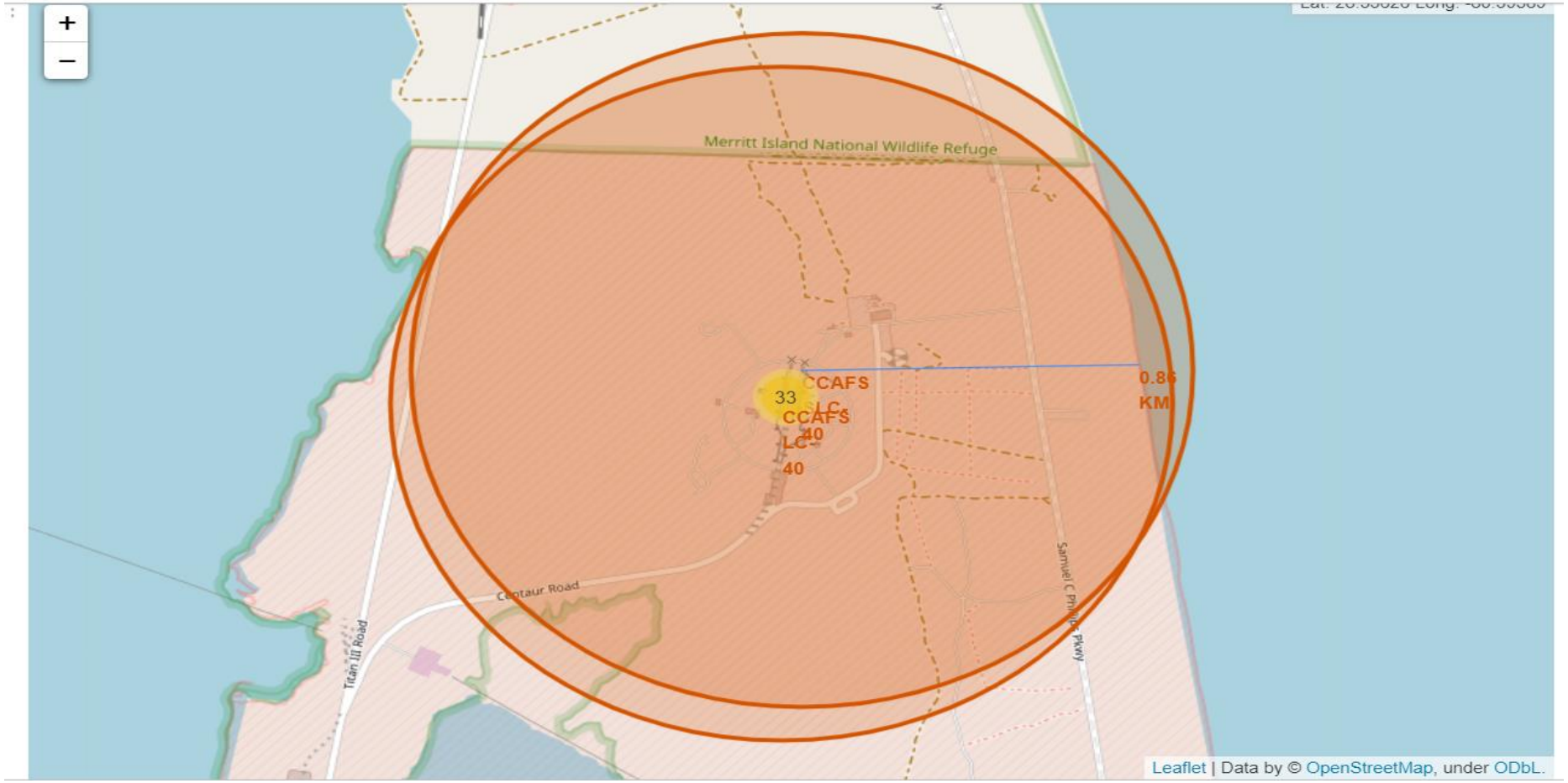




RESULTS

3. Folium

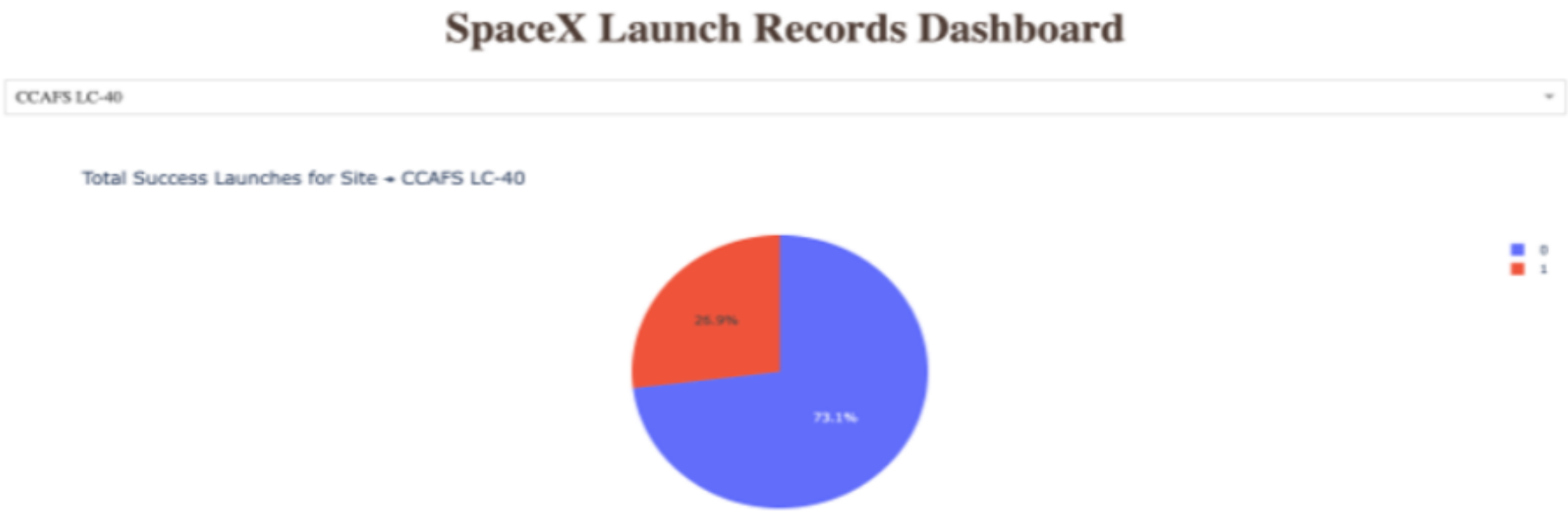
- Calculate the distances between a launch site to its proximities



RESULTS

4. Dash

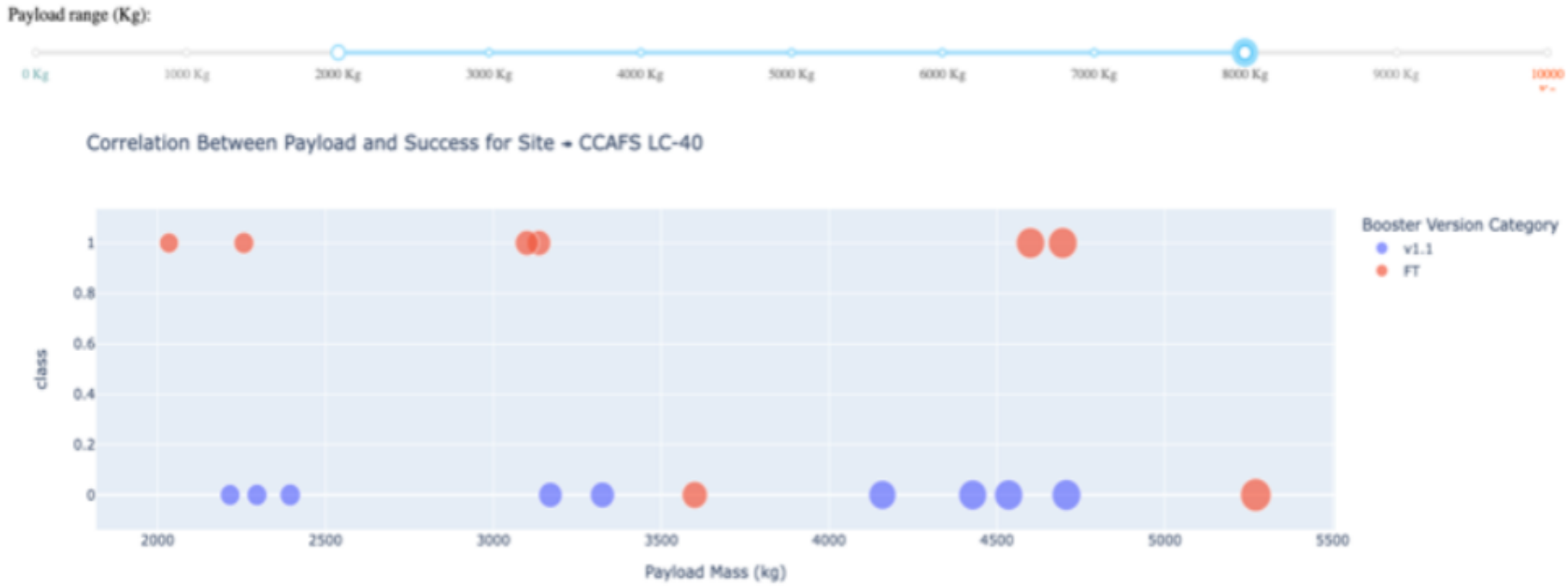
- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



RESULTS

4. Dash

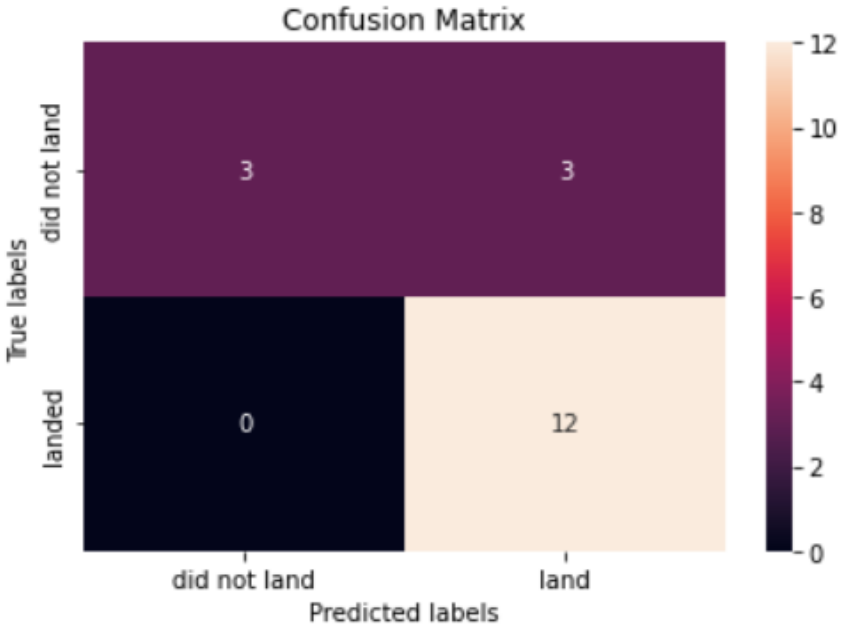
- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.



RESULTS

5. Predictive Analysis

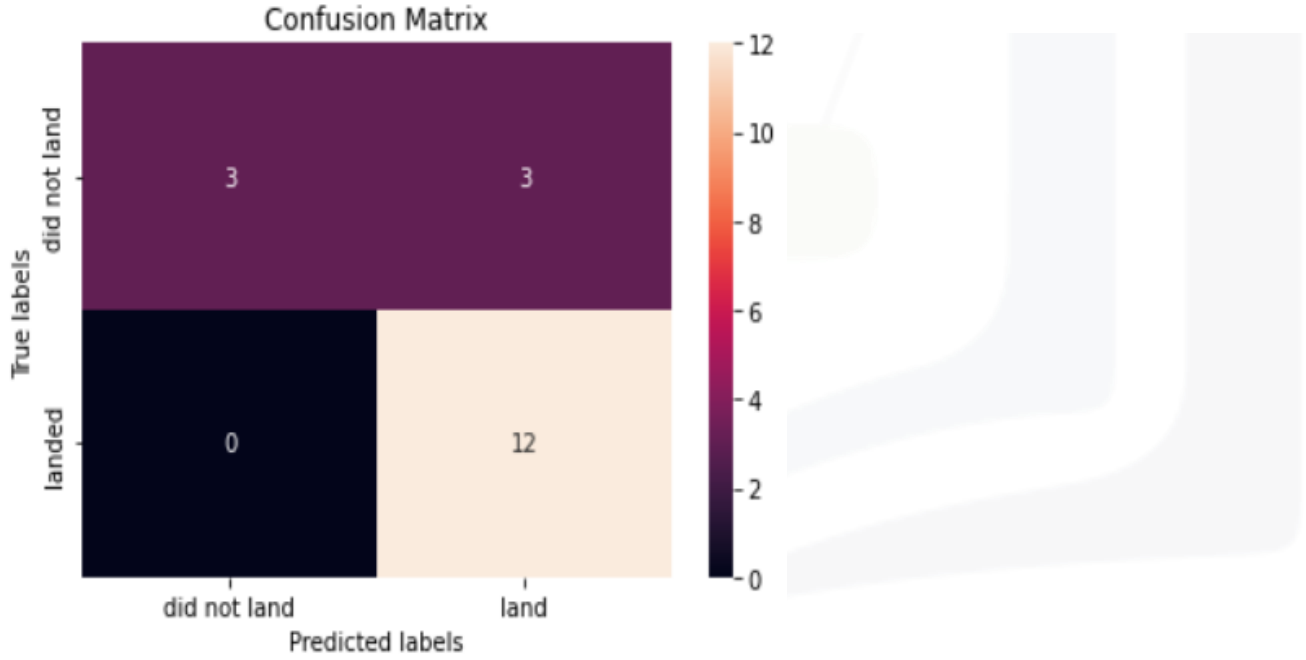
- Logistic Regression
  - GridSearchCV best score on train data : **0.846428571428571 3**
  - Best paramters : **'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'**
  - Accuracy score on test data : **0.8333333333333334**
  - Confusion Matrix :



RESULTS

5. Predictive Analysis

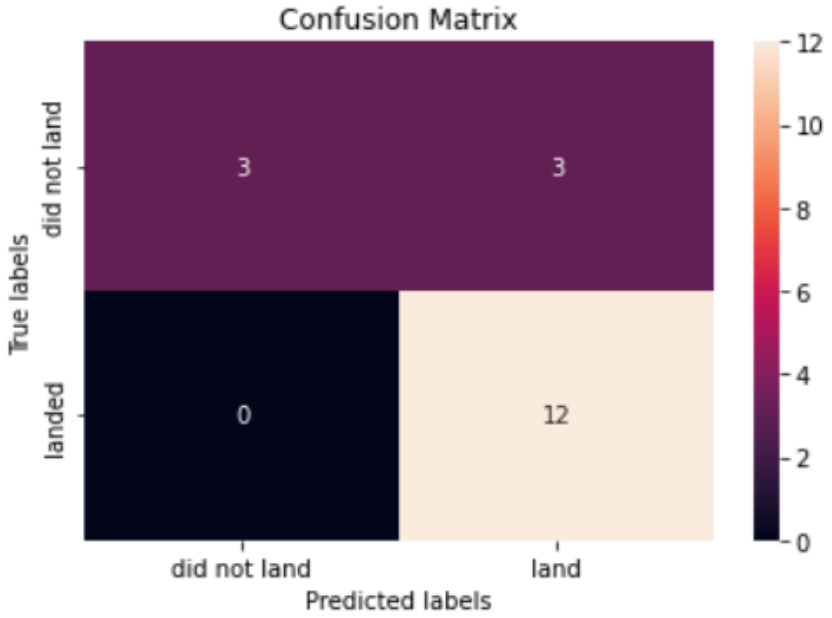
- Support Vector Machine (SVM)
  - GridSearchCV best score on train data : **0.8482142857142856**
  - Best paramters : **'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'**
  - Accuracy score on test data : **0.8333333333333334**
  - Confusion Matrix :



RESULTS

5. Predictive Analysis

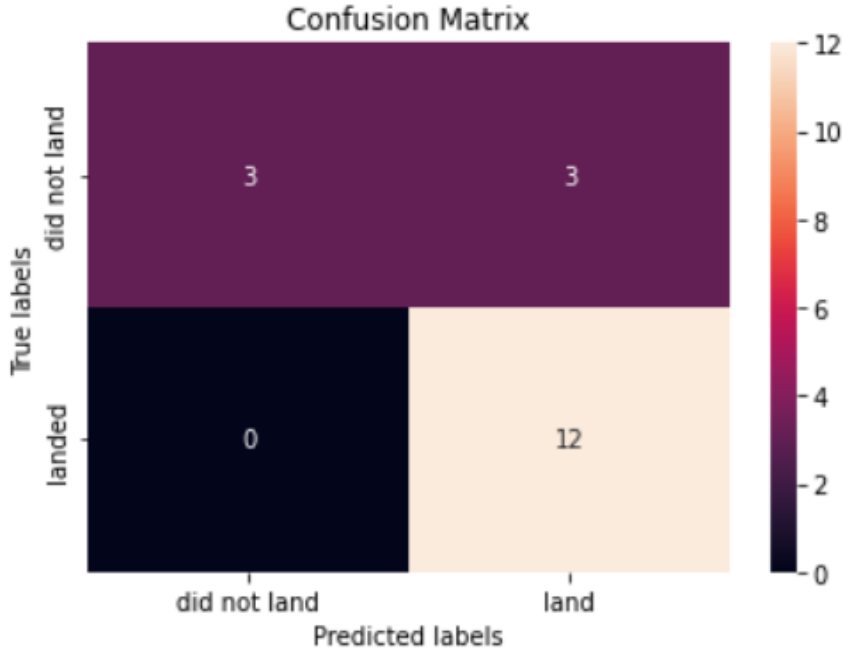
- Decision Tree
  - GridSearchCV best score on train data : **0.875**
  - Best paramters : **'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitter': 'random'**
  - Accuracy score on test data : **0.8333333333333334**
  - Confusion Matrix :



RESULTS

5. Predictive Analysis

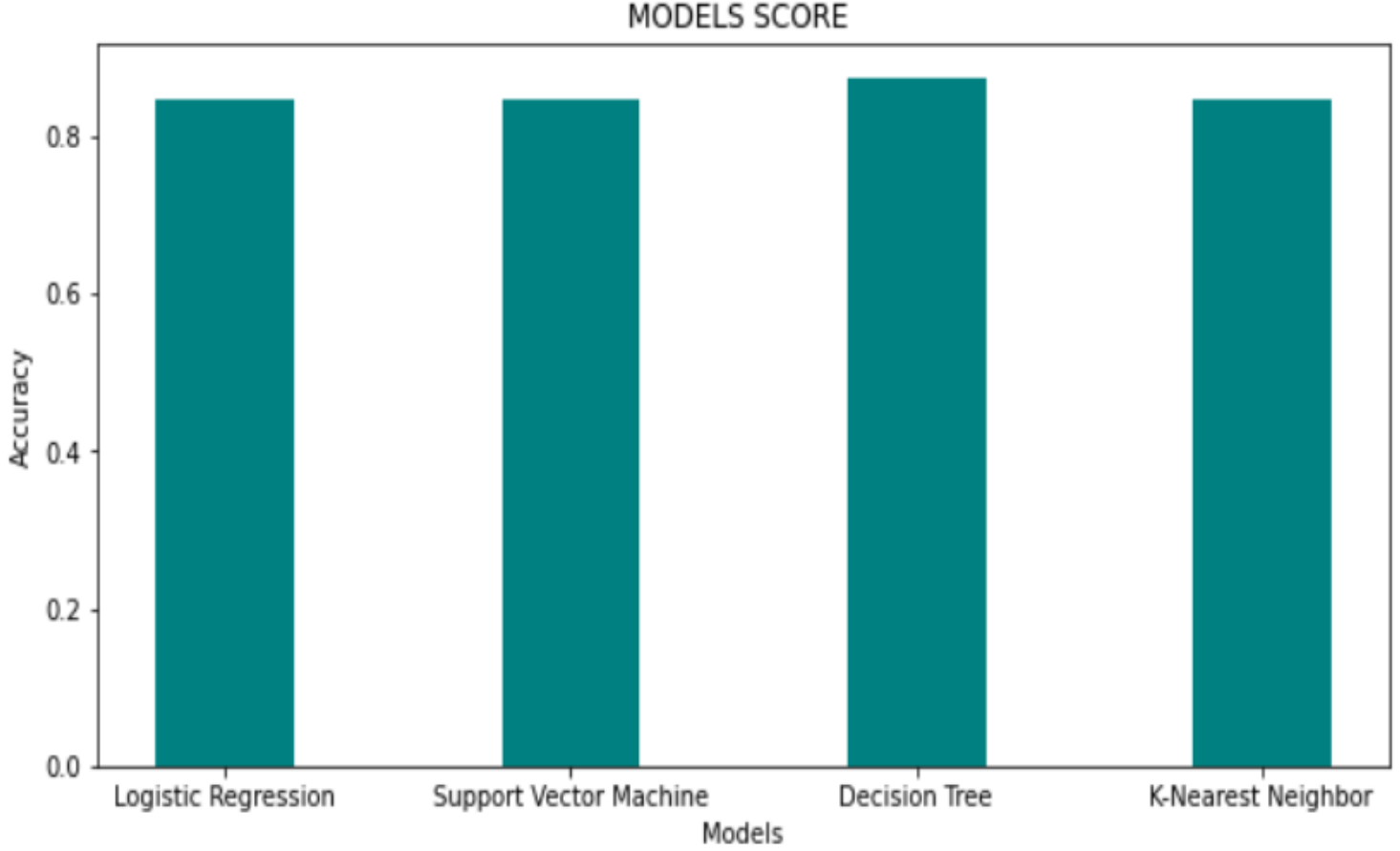
- K-Nearest Neighbors (KNN)
  - GridSearchCV best score on train data : **0.8482142857142858**
  - Best paramters : **'algorithm': 'auto', 'n_neighbors': 10, 'p': 1**
  - Accuracy score on test data : **0.8333333333333334**
  - Confusion Matrix :



RESULTS

5. Predictive Analysis

- Decision Tree is the best classifier to predict if the Falcon 9 rocket , will launch successfully or not



DISCUSSION

- From the data visulaization section, we can see that some features may have a correlation with the mission outcome in several days. For example, with heavy payloads,the successfull landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish well, such as positive and negative landing rate are both presents.

- Therefore, each future may have certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult to satisfy. However, we can use some machien learning algorithms to learn the pattern of the past data and predict wether a mission will be successfull or not, based on the given features.

CONCLUSION

- In this project, we have try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch,
- Each feature of Falcon 9 launch, such as payload mass (kg) or orbit type, may affect the mission outcome in certain way,
- Several machine learning classifier have been used, to learn the patterns of past Falcon 9 launch data, to produce a predictive model.
- By using a new Falcon 9 launch (test data), we outcome with decision tree to be a best classifier to predict if a Falcon 9 rocket launch will land successfully

GITHUB JOB POSTINGS