Winning Space Race with Data Science

Richard Lung Wicaksono
October 4, 2023

Outline

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

Executive Summary

The data collection methodology for this project involves retrieving and preparing data from two sources: the SpaceX API and web scraping Falcon 9 launch data from Wiki pages. The API data is retrieved and converted into a structured JSON format, which is then transformed into a tabular format using *json_normalize*. Null values in the PayloadMass column are handled by replacing them with the mean. Falcon 1 launches are filtered out if present. Additional Falcon 9 launch data is scraped from relevant Wiki pages and converted into a Pandas DataFrame. The dataset is prepared for analysis, leaving *NULL* values in the LandingPad column and performing one-hot encoding for LandingPad values.

Following dataset preparation, data wrangling is performed, including calculating the number of launches on each site, the number and occurrence of each orbit, and the number and occurrence of mission outcomes per orbit type. A landing outcome label is created from the Outcome column. Exploratory data analysis (EDA) is conducted using visualization and SQL, and interactive visual analytics are performed using Folium and Plotly Dash. Predictive analysis is carried out using classification models.

The data preprocessing stage involves standardizing the data for consistency and splitting it into training and testing sets. Model training and hyperparameter tuning are conducted with various algorithms, and Grid Search is used to optimize hyperparameters for each algorithm. Model evaluation is performed, assessing Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors, and outputting a confusion matrix for model evaluation.

Introduction

- In this capstone project, our goal is to forecast the successful landing of the Falcon 9 first stage. SpaceX promotes Falcon 9 rocket launches on its website at a price of $62 million, a significantly lower cost compared to other providers, who charge upwards of $165 million for each launch. This cost advantage primarily stems from SpaceX's ability to recycle the initial stage of the rocket. Consequently, by ascertaining the likelihood of a successful first-stage landing, we can effectively estimate the overall launch cost. This data becomes invaluable for potential competitors looking to bid against SpaceX for rocket launch contracts.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    1. Data Retrieval and Preparation:
        - Access SpaceX launch data from the API (api.spacexdata.com/v4/launches/past).
        - Retrieve and convert the data into a structured format (JSON).
    2. Data Transformation and Cleaning:
        - Normalize the JSON data into a tabular format using json_normalize.
        - Handle NULL values, focusing on the PayloadMass column by replacing them with the mean.
        - Filter out Falcon 1 launches if present.
    3. Alternative Data Source:
        - Scrape additional Falcon 9 launch data from relevant Wiki pages.
        - Convert the scraped data into a Pandas DataFrame for analysis.
    4. Dataset Preparation and Finalization:
        - Ensure the dataset is clean and meaningful for analysis.
        - Leave the LandingPad column with NULL values (representing no landing pad usage).
        - Prepare the data for further analysis, including one-hot encoding for LandingPad values.

Methodology (continued)

Executive Summary

- Perform data wrangling
  1. Calculate the number of launches on each site
  2. Calculate the number and occurrence of each orbit
  3. Calculate the number and occurrence of mission outcome per orbit type
  4. Create a landing outcome label from Outcome column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology (continued)

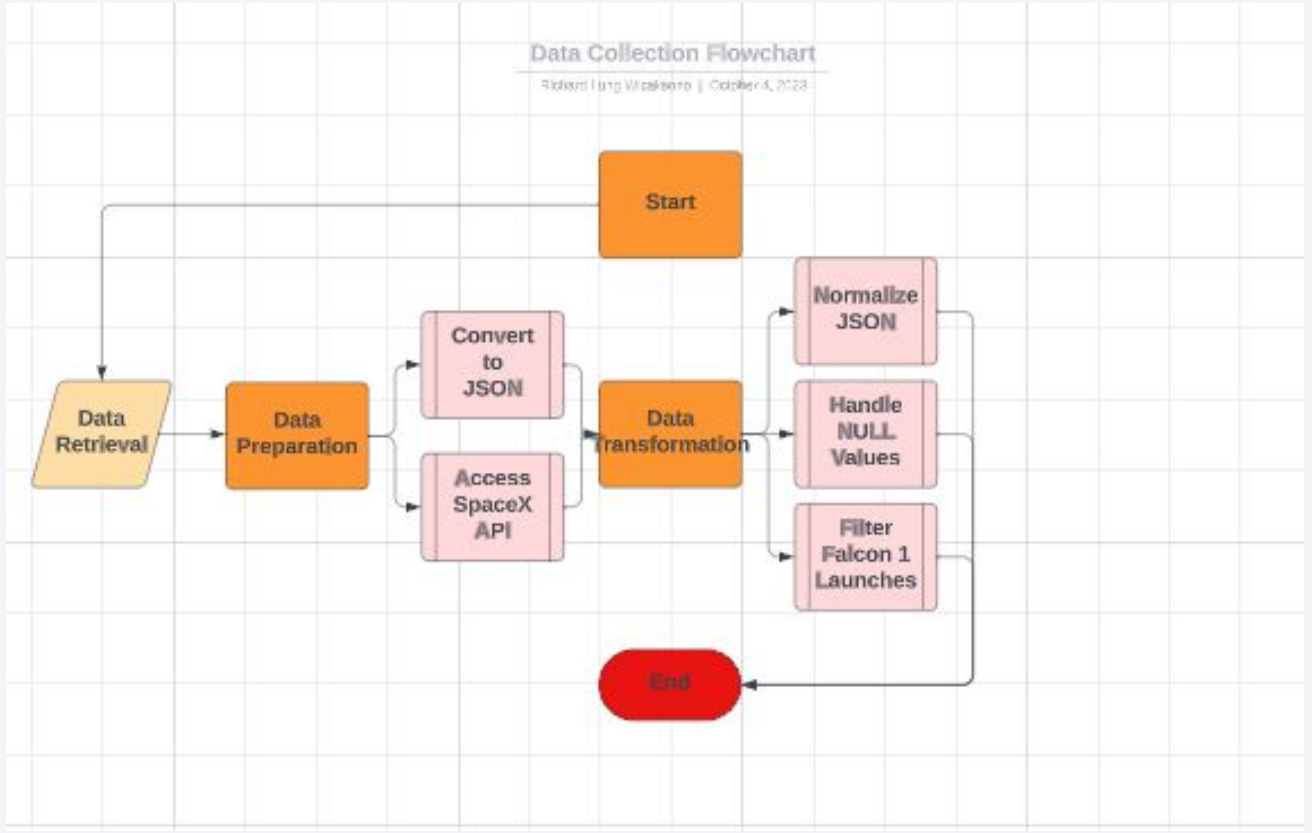
Executive Summary

- Perform predictive analysis using classification models
  1. Data Preprocessing and Splitting:
     - Standardize the data for consistency.
     - Split the data into training and testing sets.
  2. Model Training and Hyperparameter Tuning:
     - Train the model using various algorithms.
     - Conduct Grid Search to optimize hyperparameters for each algorithm.
  3. Model Evaluation and Reporting:
     - Evaluate the models using the best hyperparameter values.
     - Assess Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors.
     - Output the confusion matrix for model evaluation.

Data Collection

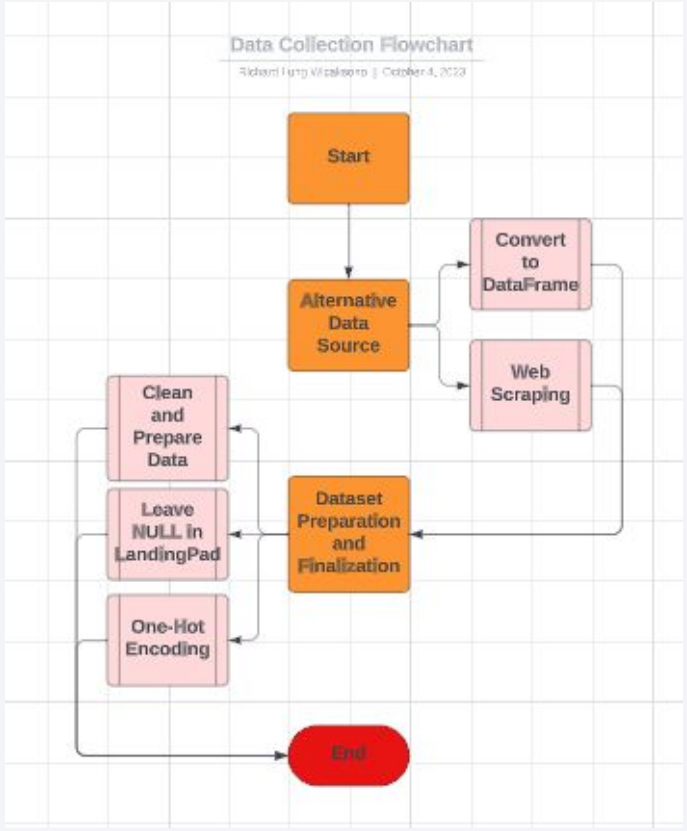
1. **Data Retrieval and Preparation:**
    a. Access SpaceX launch data from the API (api.spacexdata.com/v4/launches/past).
    b. Retrieve and convert the data into a structured format (JSON).
2. **Data Transformation and Cleaning:**
    a. Normalize the JSON data into a tabular format using json_normalize.
    b. Handle NULL values, focusing on the PayloadMass column by replacing them with the mean.
    c. Filter out Falcon 1 launches if present.
3. **Alternative Data Source:**
    a. Scrape additional Falcon 9 launch data from relevant Wiki pages.
    b. Convert the scraped data into a Pandas DataFrame for analysis.
4. **Dataset Preparation and Finalization:**
    a. Ensure the dataset is clean and meaningful for analysis.
    b. Leave the LandingPad column with NULL values (representing no landing pad usage).
    c. Prepare the data for further analysis, including one-hot encoding for LandingPad values.

Data Collection – SpaceX API



https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/Data%20Collection%20API.ipynb

Data Collection - Scraping



https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column

https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/Data%20Wrangling.ipynb

EDA with Data Visualization

- Charts that were used:
  - Scatter plot
    - To find the relationship between various independent and dependent variables
  - Bar plot
    - To find the relationship between the success rate and the orbit type
  - Line plot
    - To find the trend of the success rate throughout the years

https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/EDA%20with%20Visualization%20Lab.ipynb

EDA with SQL

- SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
- SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
- SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"
- SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)"
- SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (drone ship)" AND ("PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
- SELECT "Mission_Outcome", COUNT("Mission_Outcome") FROM SPACEXTABLE GROUP BY "Mission_Outcome"
- SELECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
- SELECT SUBSTR("Date", 6, 2) AS "Month_Names", "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = "Failure (drone ship)" AND SUBSTR("Date", 1, 4) = '2015'
- SELECT "Landing_Outcome", COUNT("Landing_Outcome") FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing_Outcome" ORDER BY COUNT("Landing_Outcome") DESC

https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Circles:
  - To mark the launch sites on the map
- Markers:
  - To mark the launch site names and distances
- Lines:
  - To mark the distances between the launch sites and various geographical sites

https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/e3dd5b10f10ac9eae6a18b8041fbdb7cd3d29476/Applied%20Data%20Science%20Capstone/Interactive%20Viusal%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie chart:
  - To show the successful launch counts for all sites
- Scatter plot:
  - To show the correlation between payload and launch success

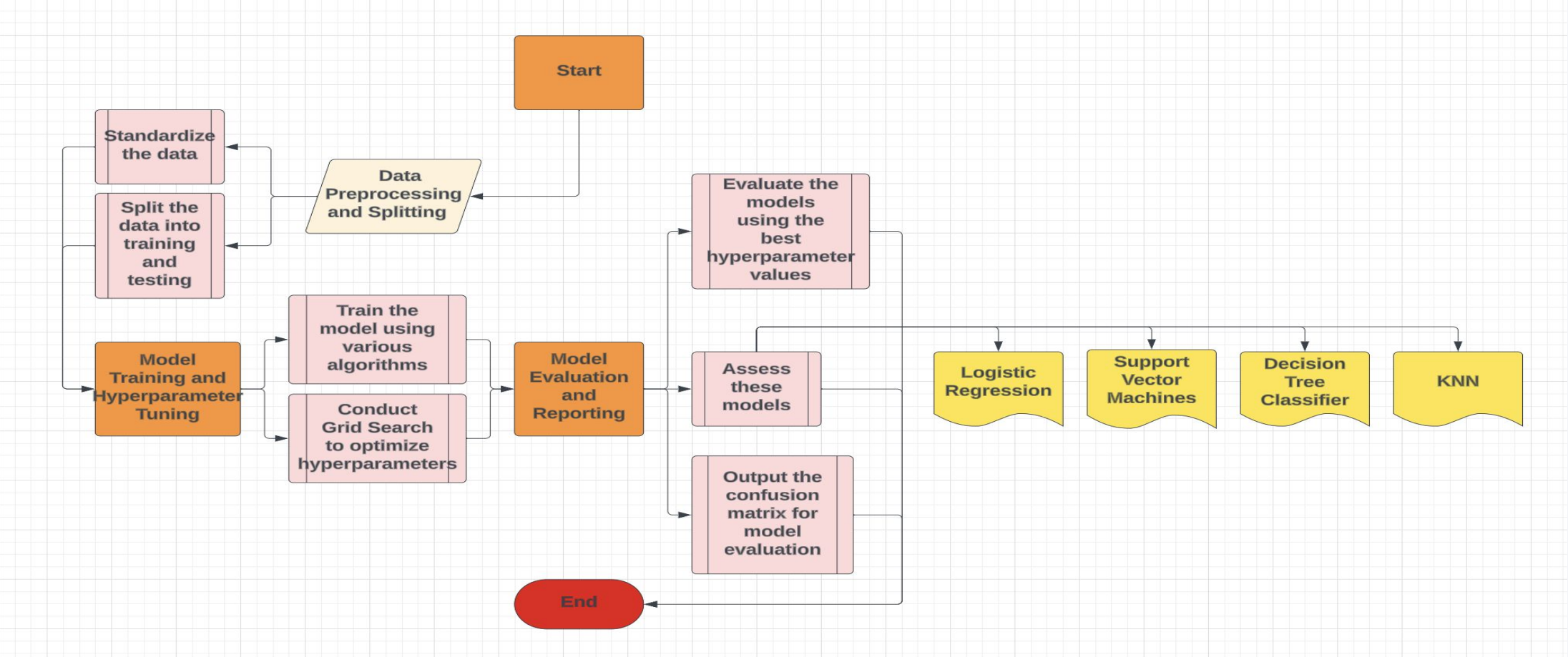
https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/70df6abfaba647d9e690b60a720c65608b790fec/Applied%20Data%20Science%20Capstone/Interactive%20Dashboard%20with%20Plotly%20Dash.py

Predictive Analysis (Classification)

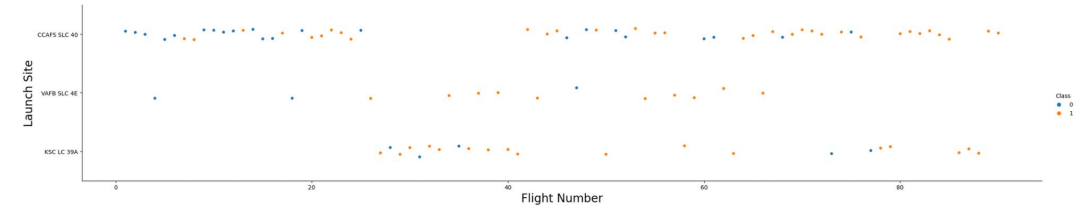
1. Data Preprocessing and Splitting:
    - Standardize the data for consistency.
    - Split the data into training and testing sets.
2. Model Training and Hyperparameter Tuning:
    - Train the model using various algorithms.
    - Conduct Grid Search to optimize hyperparameters for each algorithm.
3. Model Evaluation and Reporting:
    - Evaluate the models using the best hyperparameter values.
    - Assess Logistic Regression, Support Vector Machines, Decision Tree Classifier, and K-nearest neighbors.
    - Output the confusion matrix for model evaluation.

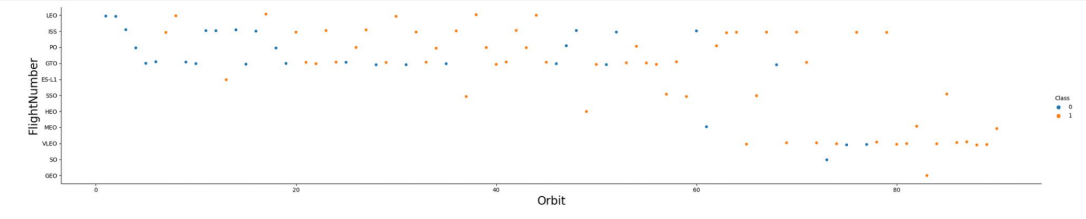
https://github.com/richardlw14/coursera-ibm-data-science-professional-certificate/blob/70df6abfaba647d9e690b60a720c65608b790fec/Applied%20Data%20Science%20Capstone/Machine%20Learning%20Prediction.ipynb

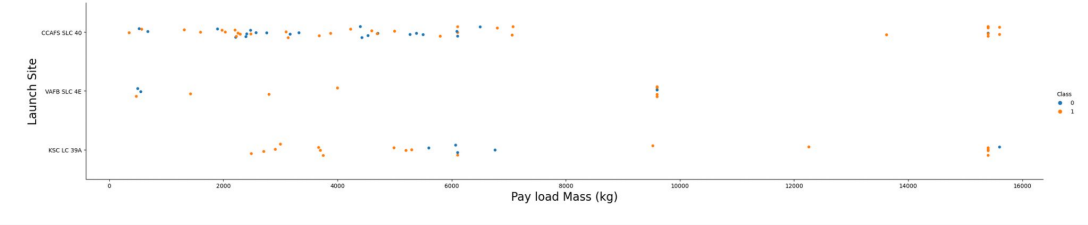
Predictive Analysis (Classification) (continued)

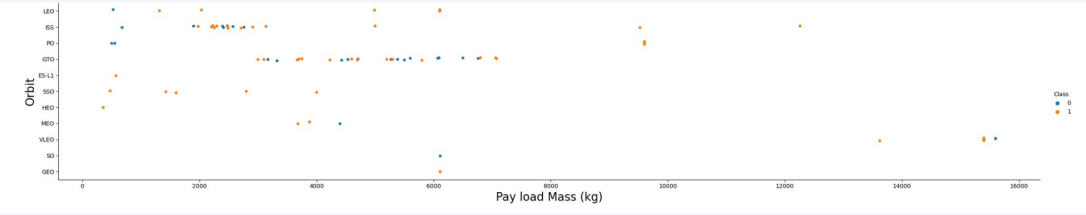


Results - EDA Scatter Plots

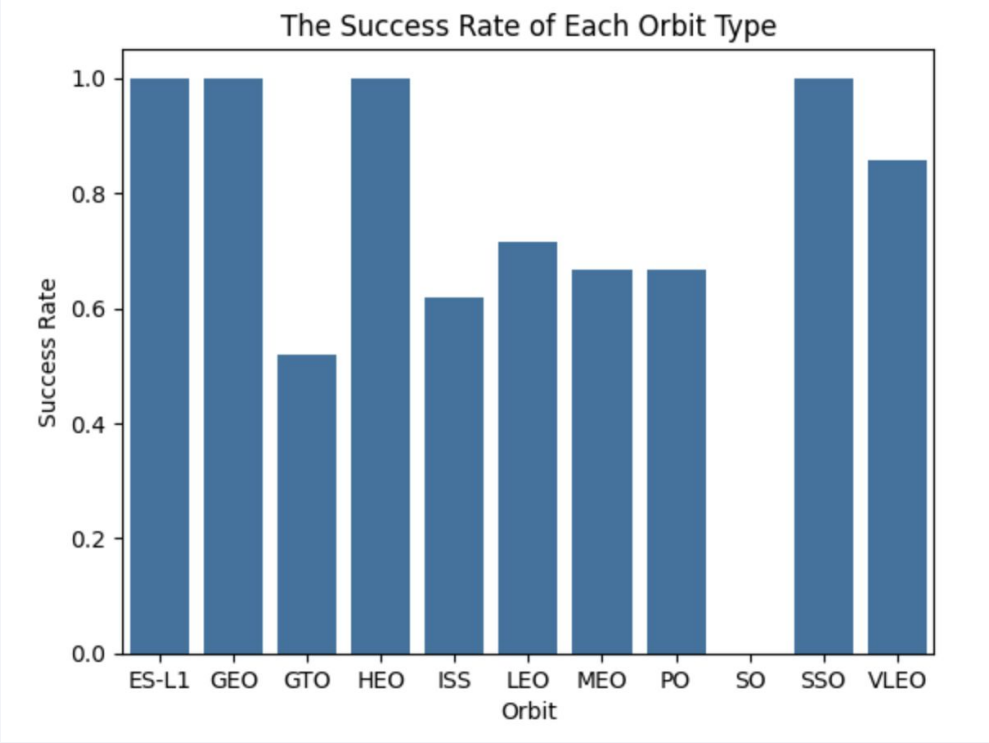


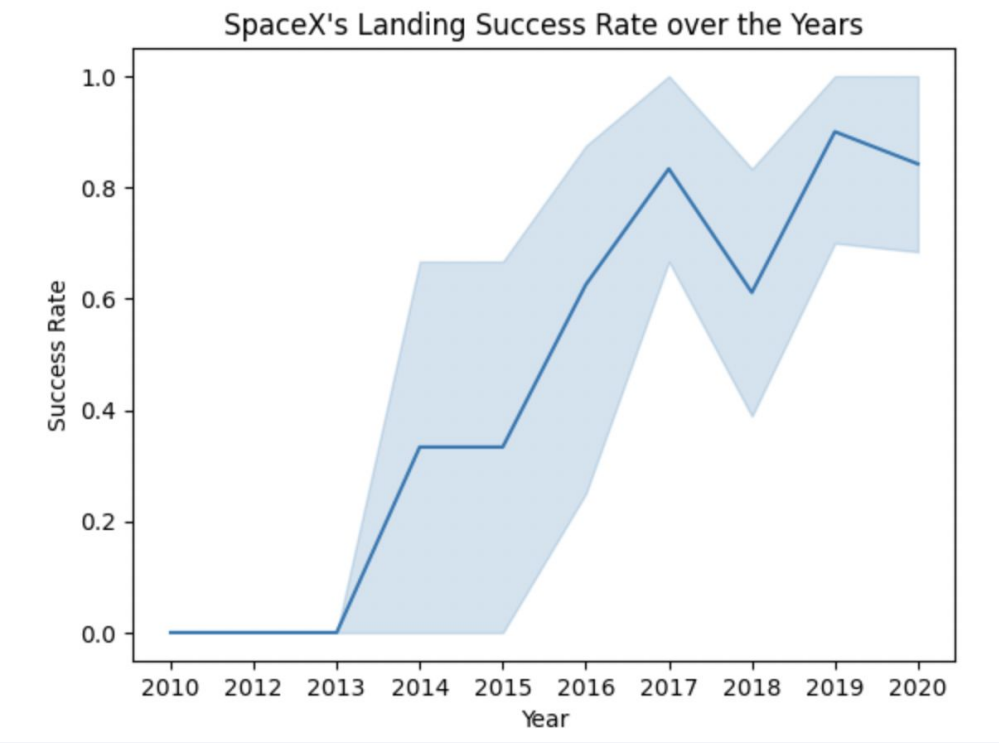




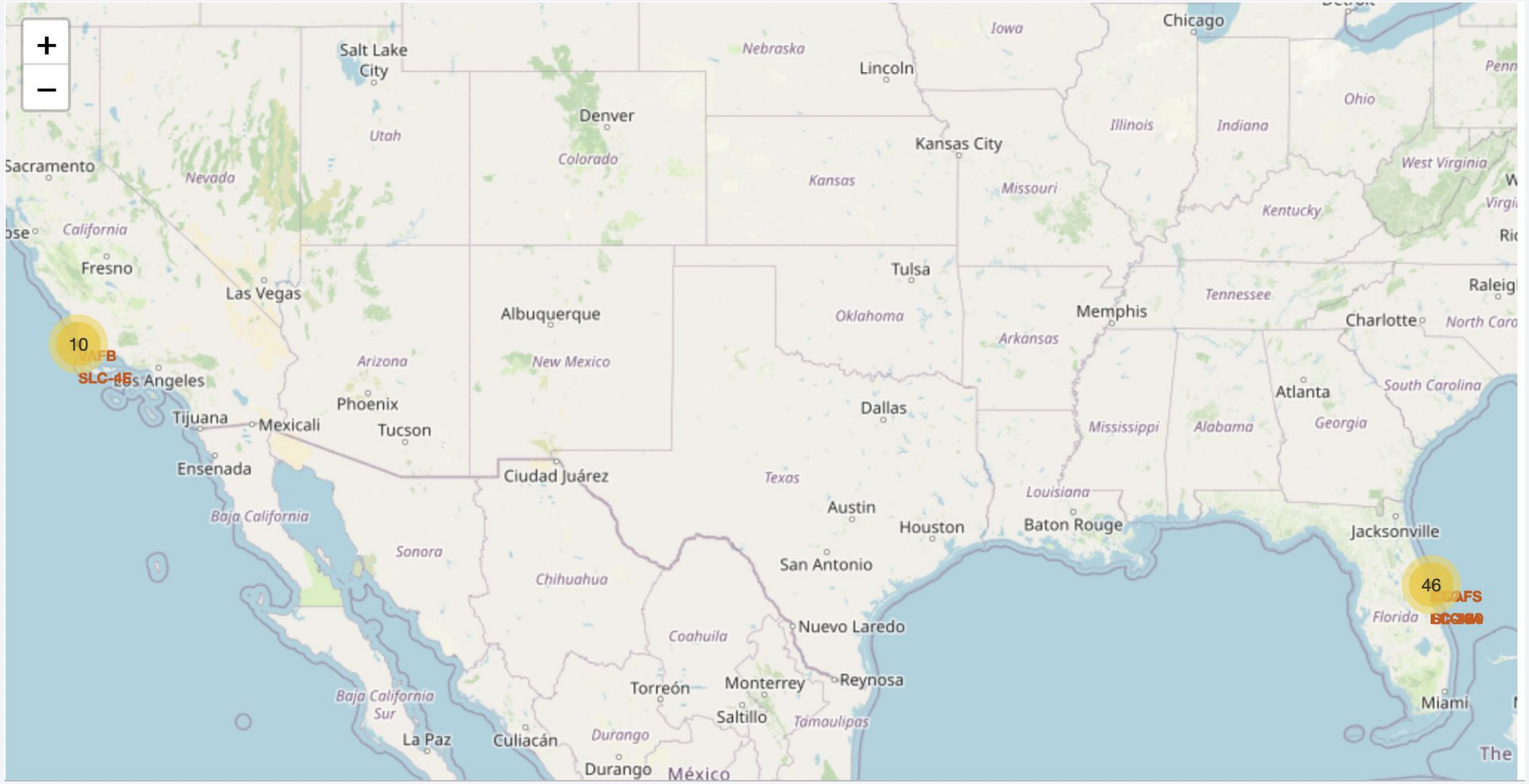


Results - EDA Charts

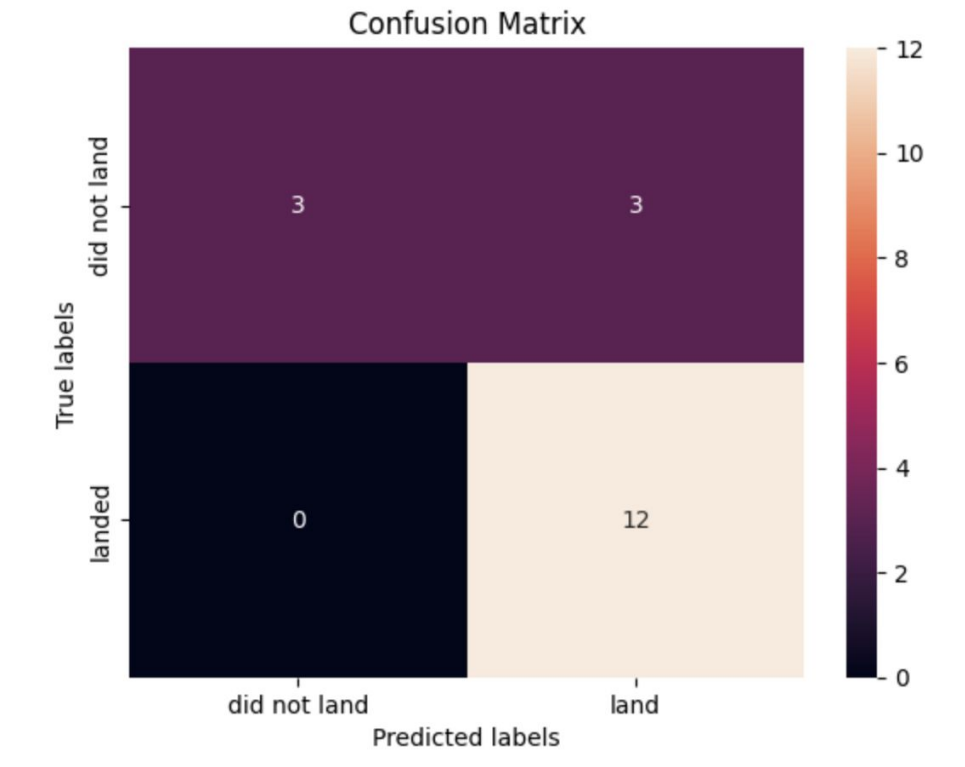


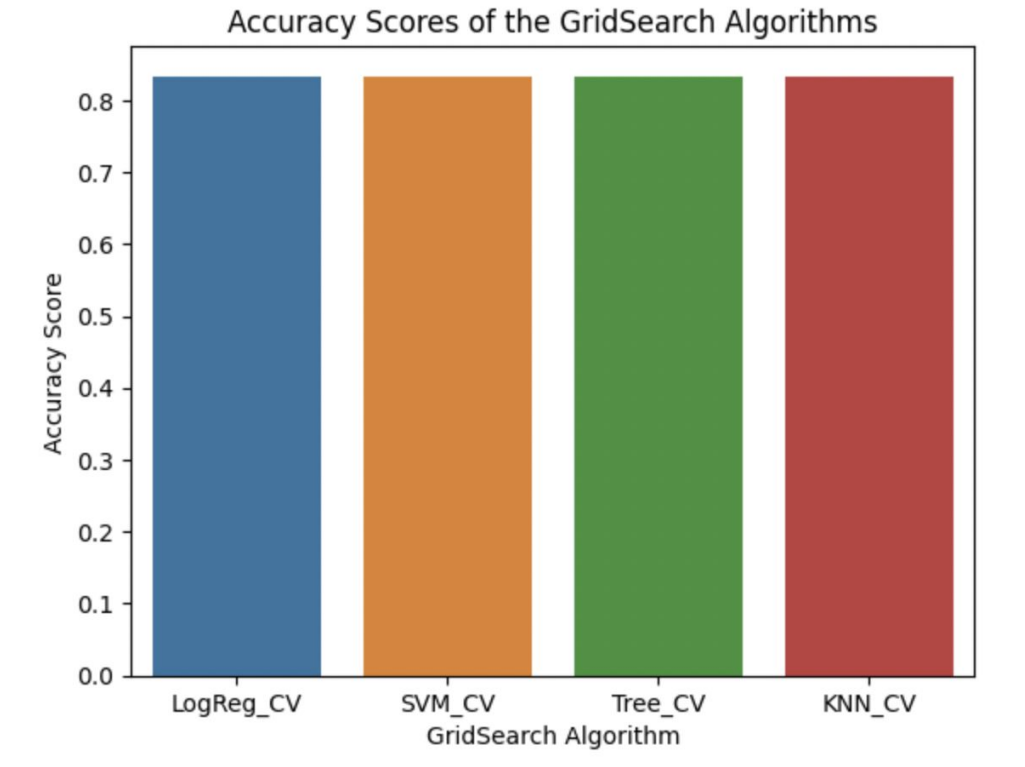


Results - Interactive Analytics (Folium)



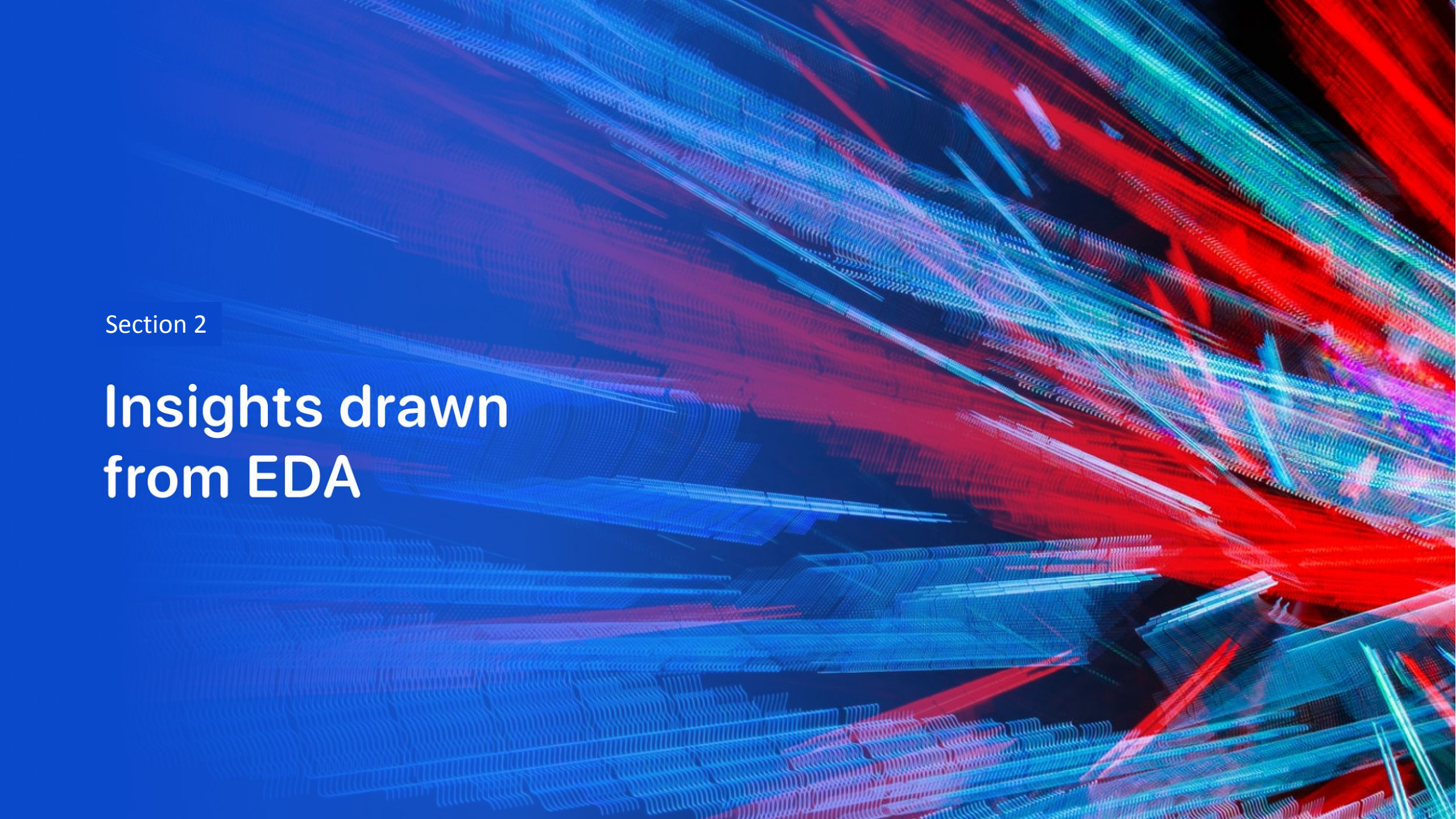
Results - Predictive Analysis (Classification)



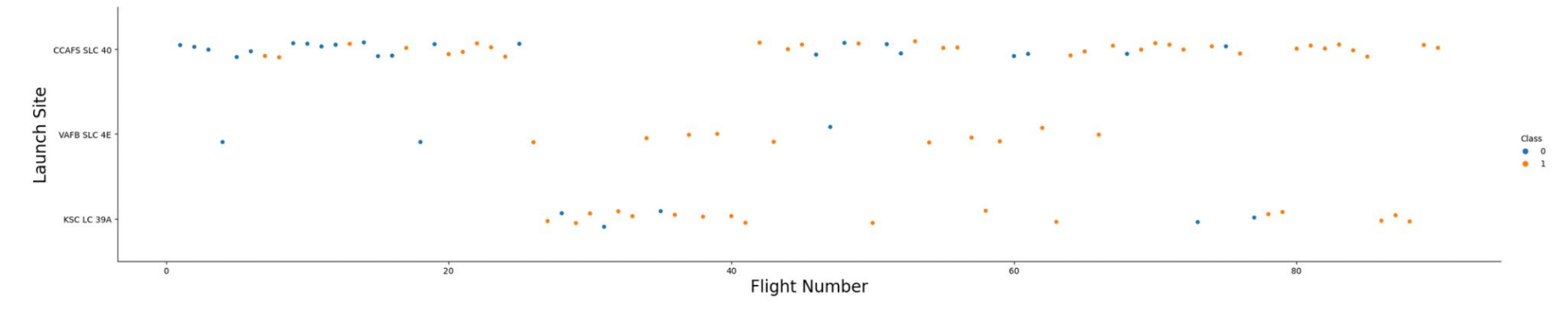


Section 2

Insights drawn from EDA

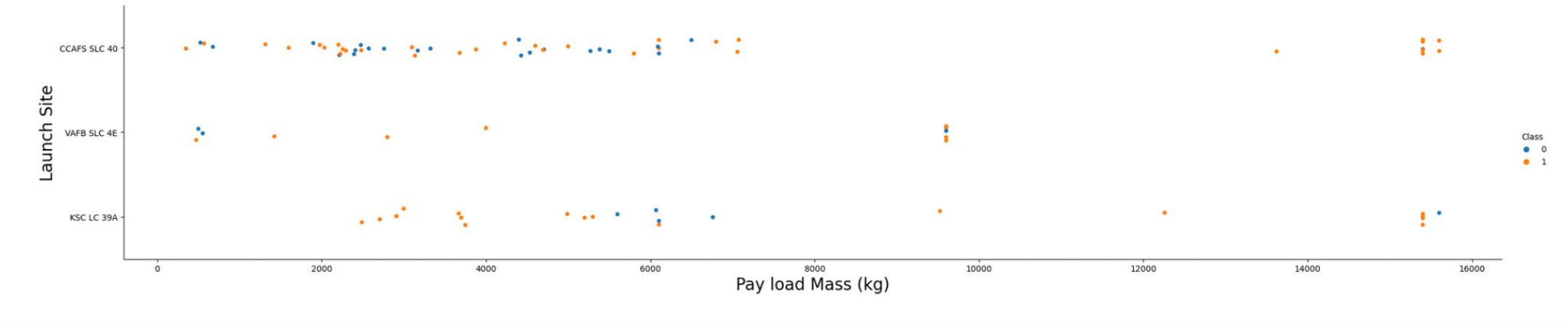


Flight Number vs. Launch Site



There is no correlation between Flight Number and Launch Site, but we can see that Launch Site CCAFS SLC 40 launched more rockets compared to the other two sites.

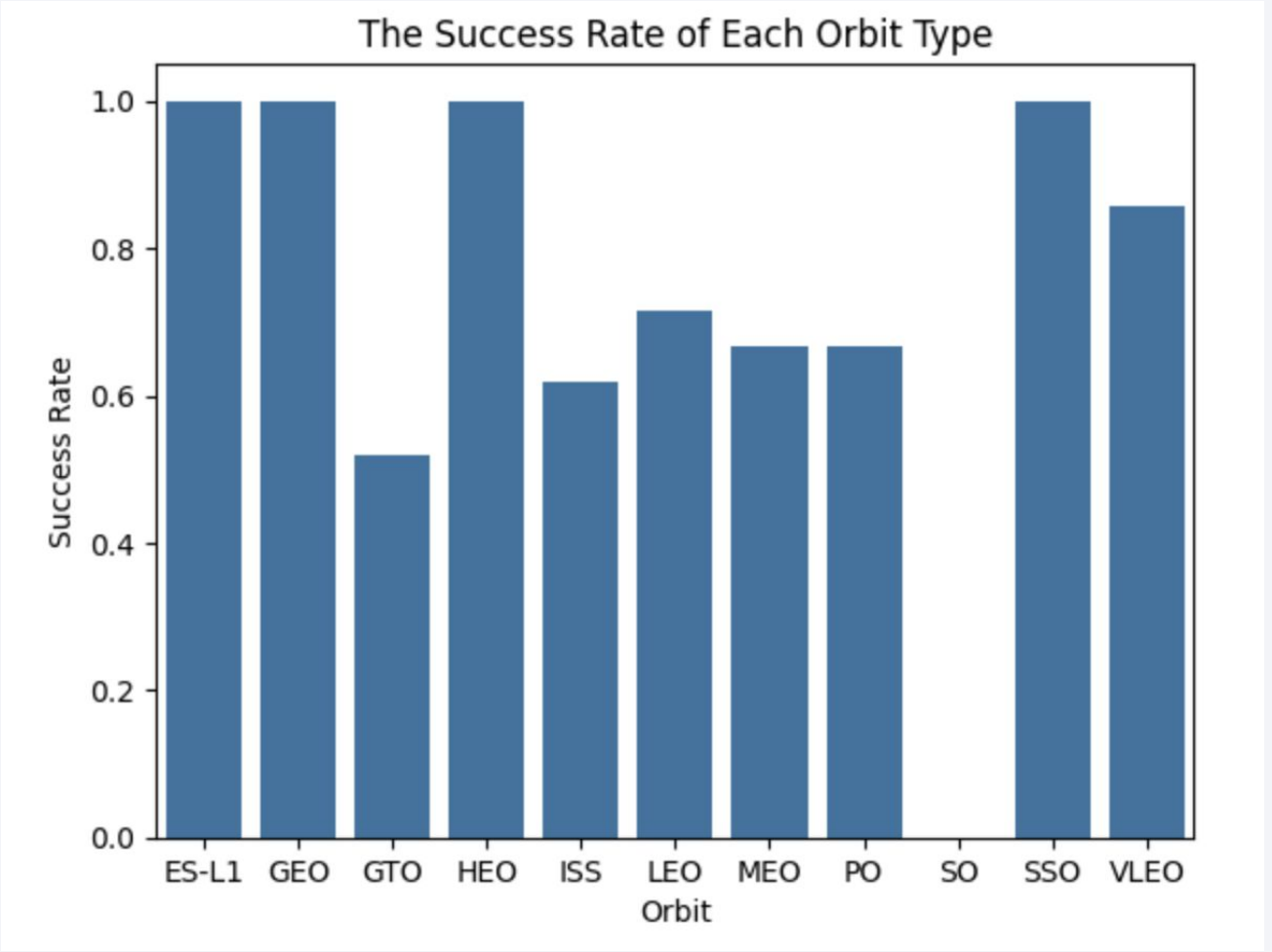
Payload vs. Launch Site



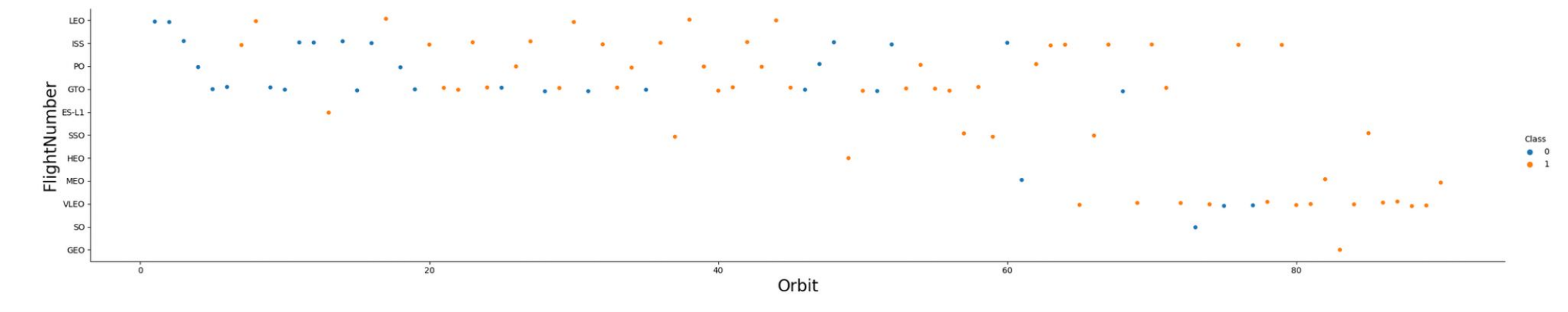
For the VAFB-SLC launchsite, there are no rockets launched for heavy payload mass (> 10,000kg).

Success Rate vs. Orbit Type

- Out of 11 Orbit types, only 4 has a 100% success rate of landing:
  - ES-L1
  - GEO
  - HEO
  - SSO
- Meanwhile, Orbit type SO has 0% success rate of landing a rocket.

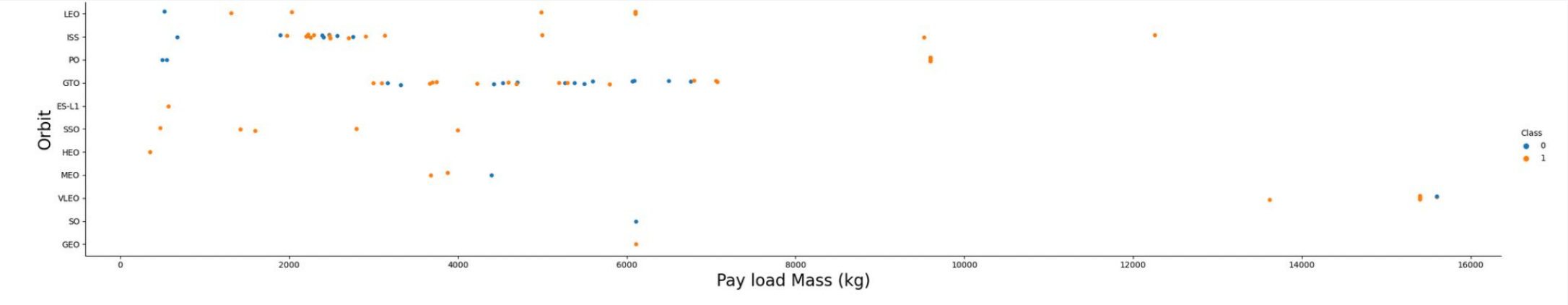


Flight Number vs. Orbit Type



No relationship found in this scatter plot. We can see that ISS and GTO have the most number of rocket launches.

Payload vs. Orbit Type

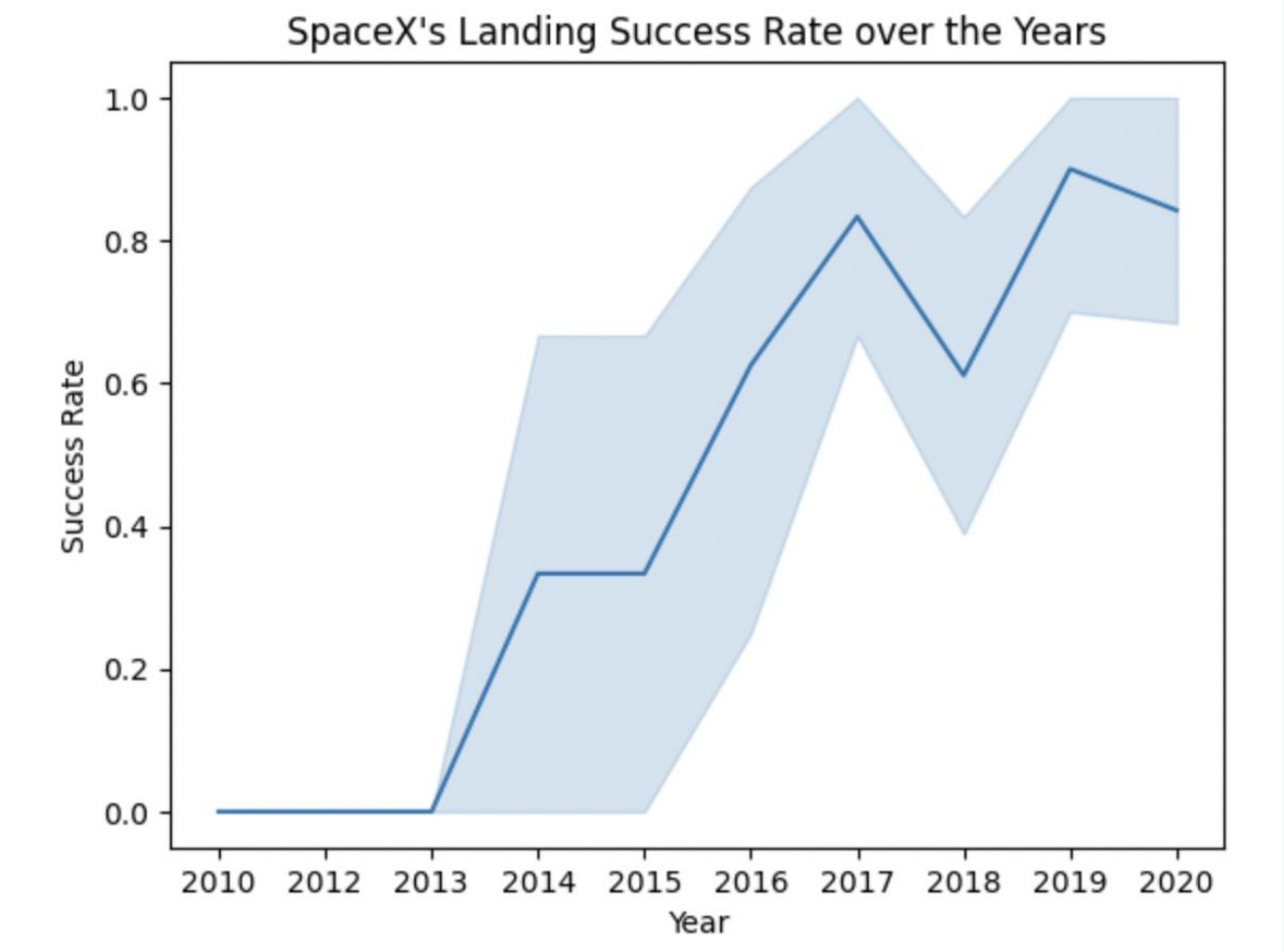


With heavy payloads, PO, LEO, and ISS have a more successful or positive landing rate.

However, for GTO, we cannot quite determine if it has a more successful or unsuccessful landing rate, because both classes are clustered together in the scatter plot.

Launch Success Yearly Trend

- Since 2013, the success rate gradually increased for 7 years.



All Launch Site Names

```
[9]: %sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[9]:

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

SpaceX has 4 different launch sites: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40.

Launch Site Names Begin with '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

From the 5 records shown above, there are no successful landings at the Launch Site CCAFS LC-40.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

For the customer NASA (CRS), SpaceX had launched a total of 45,596 kg worth of payload mass.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE WHERE "Booster_Version" = "F9 v1.1"
```

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

AVG(PAYLOAD_MASS__KG_)
2928.4

For booster version F9 v1.1, the average payload mass that was carried by SpaceX's rockets was 2,928.4 kg.

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = "Success (ground pad)"
```

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

MIN("Date")
2015-12-22

The first successful ground landing date for SpaceX occurred on December 22nd, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = "Success (drone ship)"
AND ("PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000)
```

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

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

The 4 booster versions that had a successful drone ship landing with payload between 4,000 and 6,000 kg are: F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2.

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

Out of 101 SpaceX's missions, it has successfully completed 99 missions with no issues at all, successfully completed 1 mission with unclear payload status, and failed 1 mission in flight.

NOTE: Not sure why it shows 2 different "Success" mission outcomes here. I have double checked the data with DISTINCT "Mission_Outcome" and the query still returns 4 different outcomes.

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

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

There 12 boosters which carried the maximum amount of payload. They are shown in the results above.

2015 Launch Records

```
%%sql
SELECT SUBSTR("Date", 6, 2) AS "Month_Names", "Landing_Outcome", "Booster_Version", "Launch_Site"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = "Failure (drone ship)" AND SUBSTR("Date", 1, 4) = '2015'
```

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

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

In 2015, there were 2 unsuccessful landing outcomes that happened in April and October. Both of these outcomes occurred at landing site CCAFS LC-40.

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

Between June 4th, 2010 and March 20th, 2017, no attempt has been made more than each distinct landing outcome. However, there was also an equal amount of successful landing outcomes between those dates, which were accomplished on a ground pad and a drone ship.

```
%%sql
SELECT "Landing_Outcome", COUNT("Landing_Outcome")
FROM SPACEXTABLE
WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY "Landing_Outcome"
ORDER BY COUNT("Landing_Outcome") DESC
```

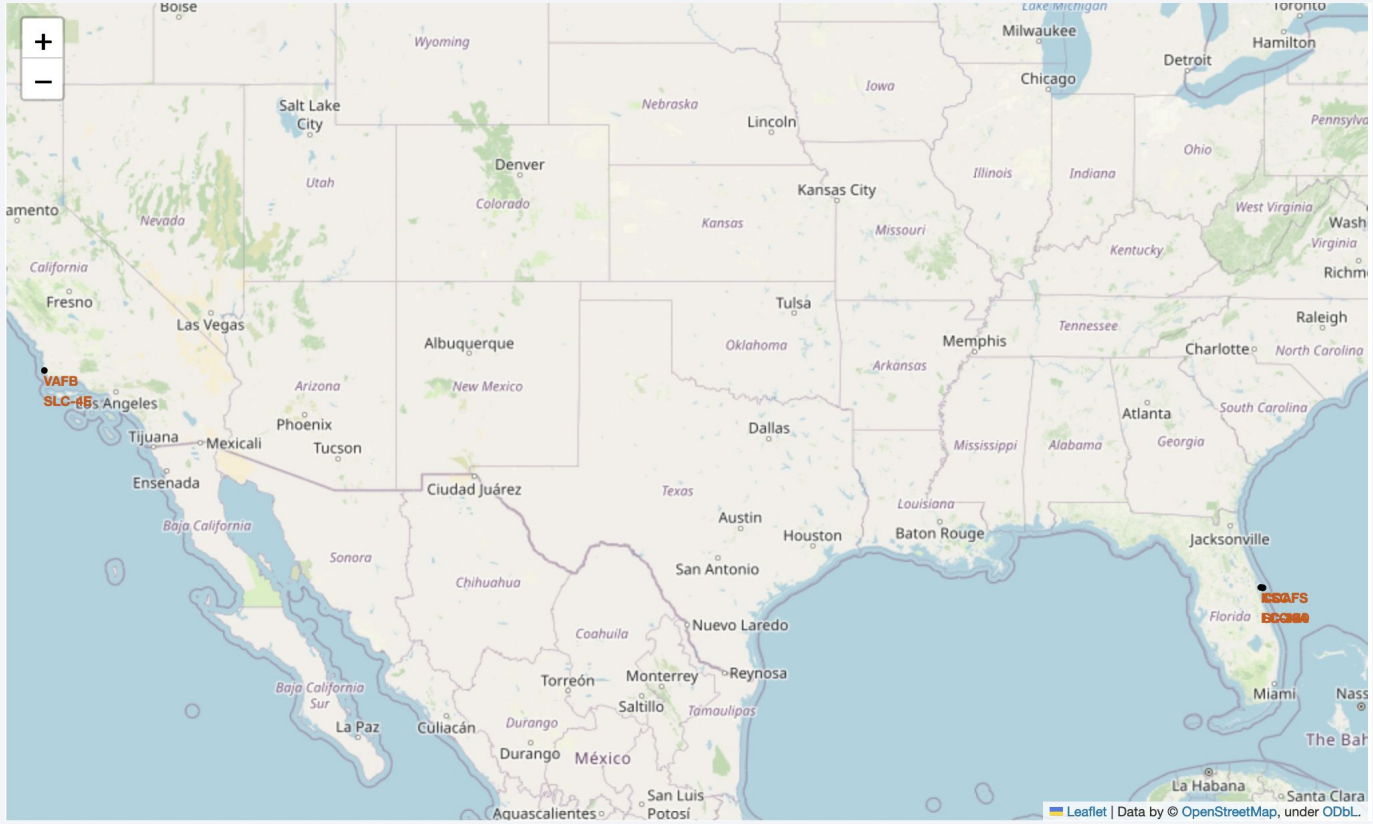
```
 * sqlite:///my_data1.db
Done.
```

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


Section 3
Launch Sites
Proximities Analysis

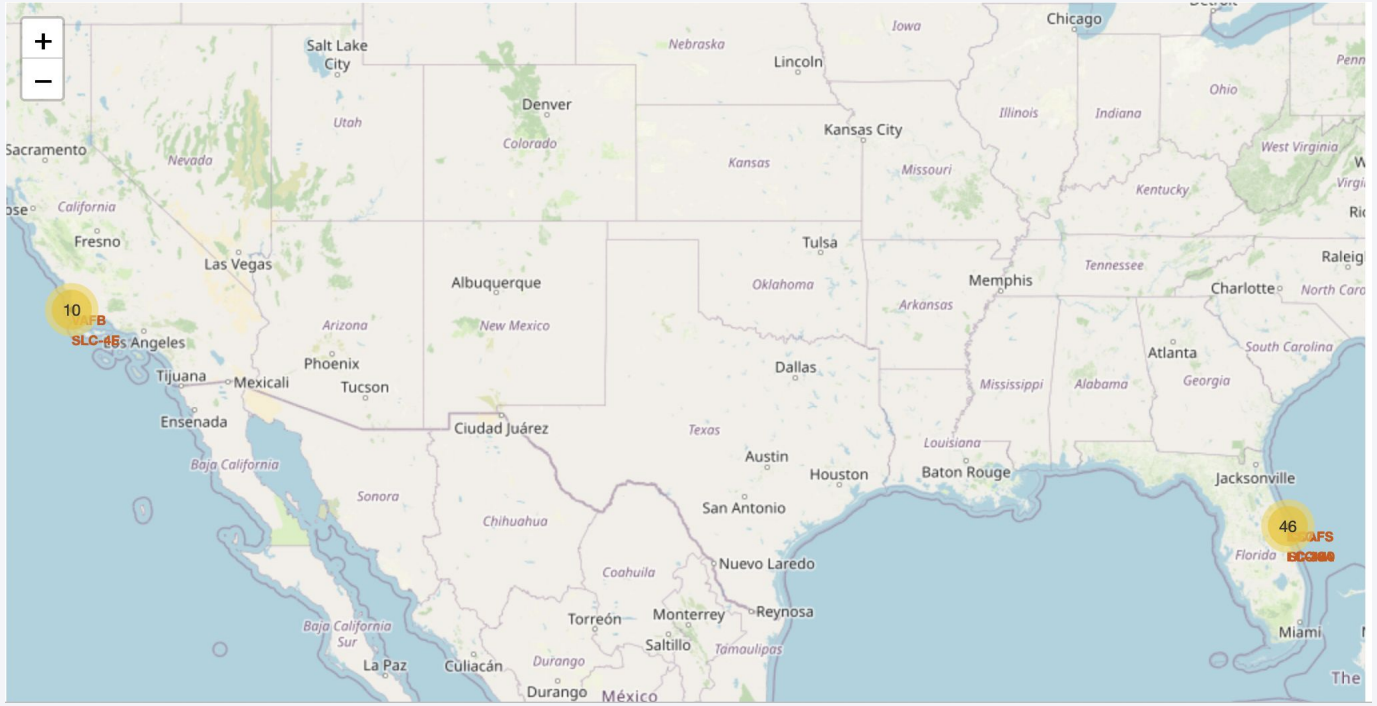
SpaceX’s Falcon 9 Launch Sites on Folium

- SpaceX’s launch sites are located in 2 states:
  - California
  - Florida
- All sites are located close to the coastlines.
- All sites are located in the Northern Hemisphere.



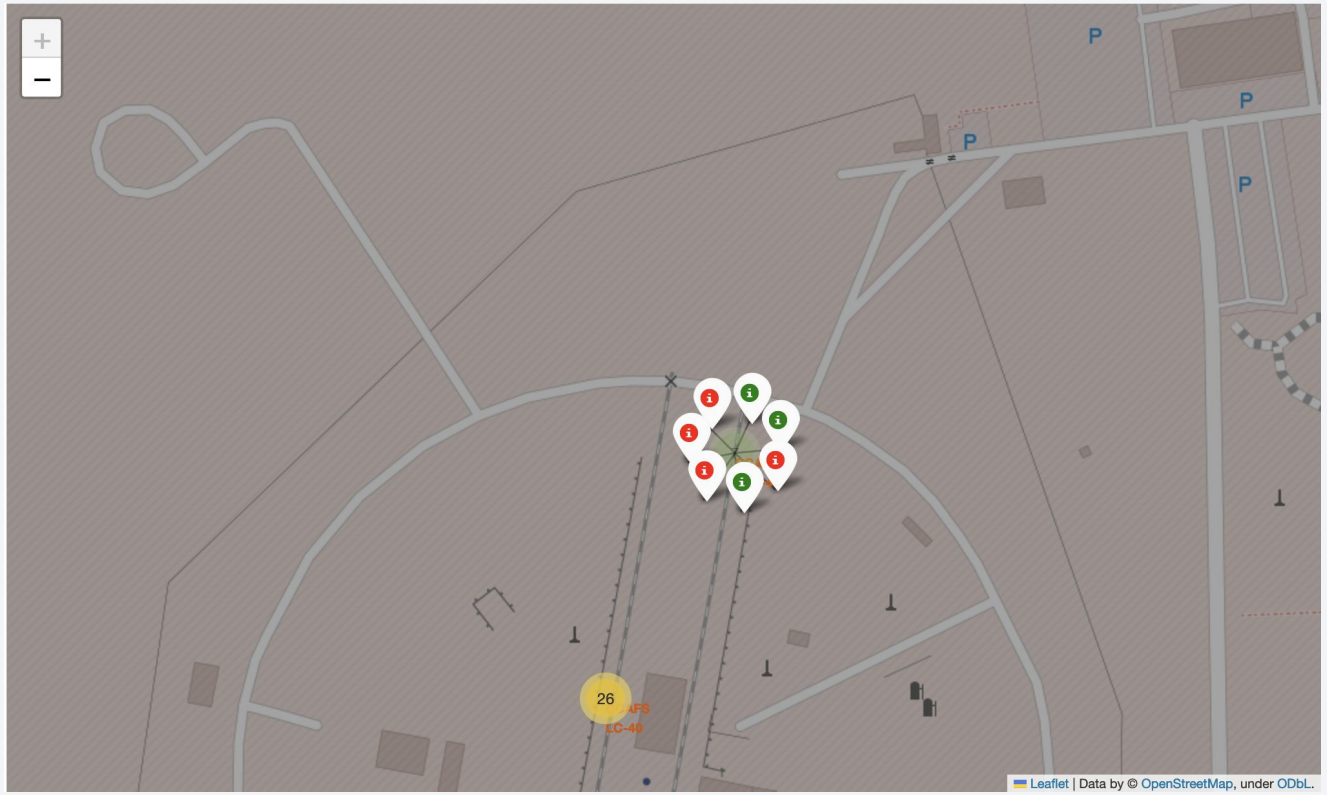
Successful/Failed Launches for each Site on Folium

- A total of 56 launches:
  - 10 in the West Coast
  - 46 in the East Coast
- All 10 launches in the West Coast are launched from the same site



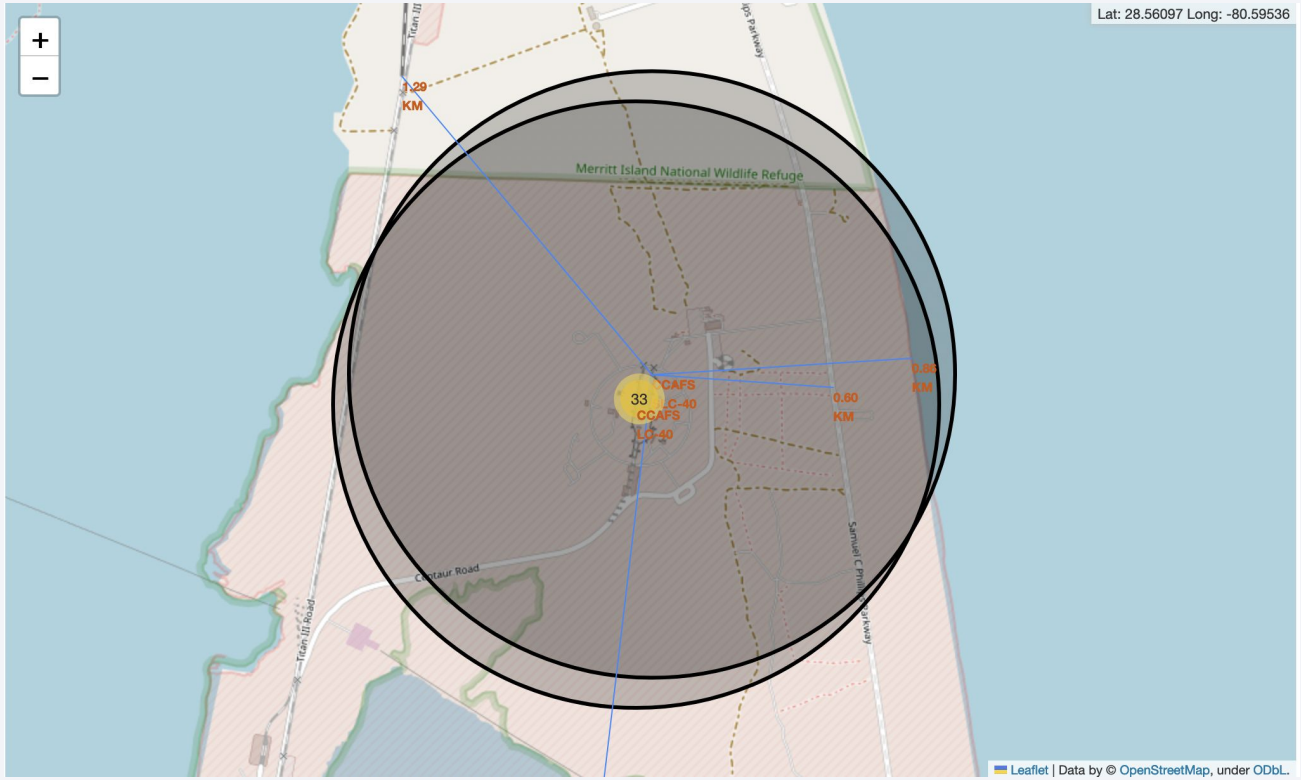
Successful/Failed Launches for each Site on Folium (continued)

- An example of a landing site: CCAFS SLC 40
- CCAFS SLC 40:
  - Located in Florida
  - ~42% successful launches
- Successful outcomes marked in green, failed outcomes marked in red

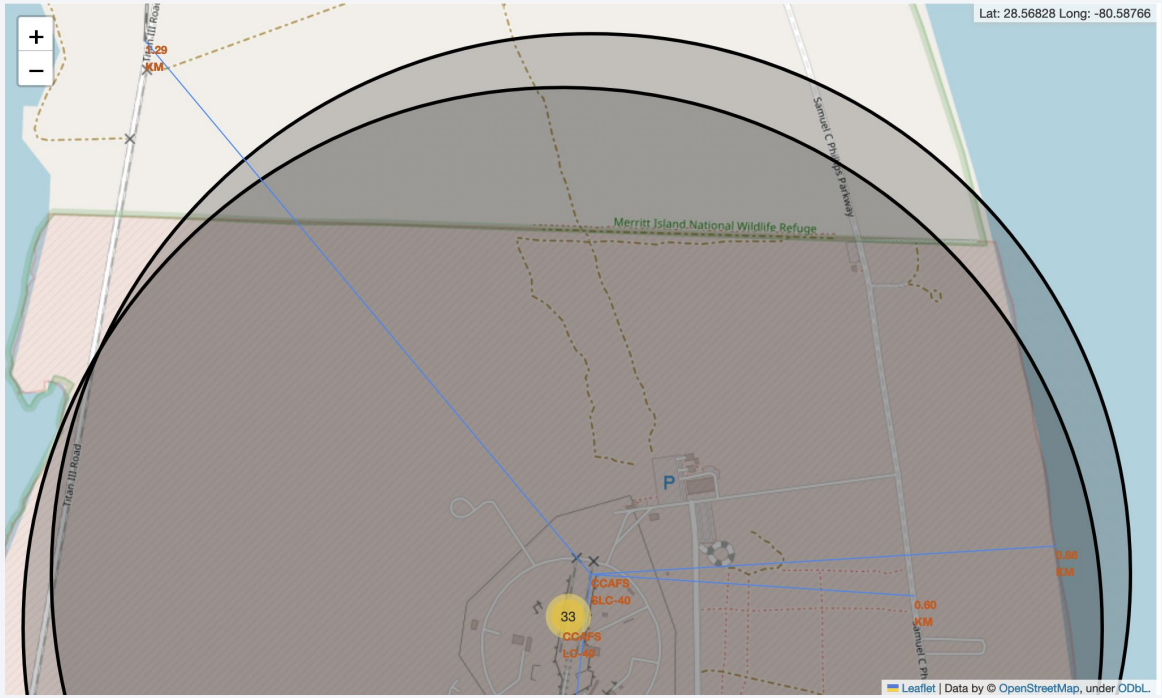


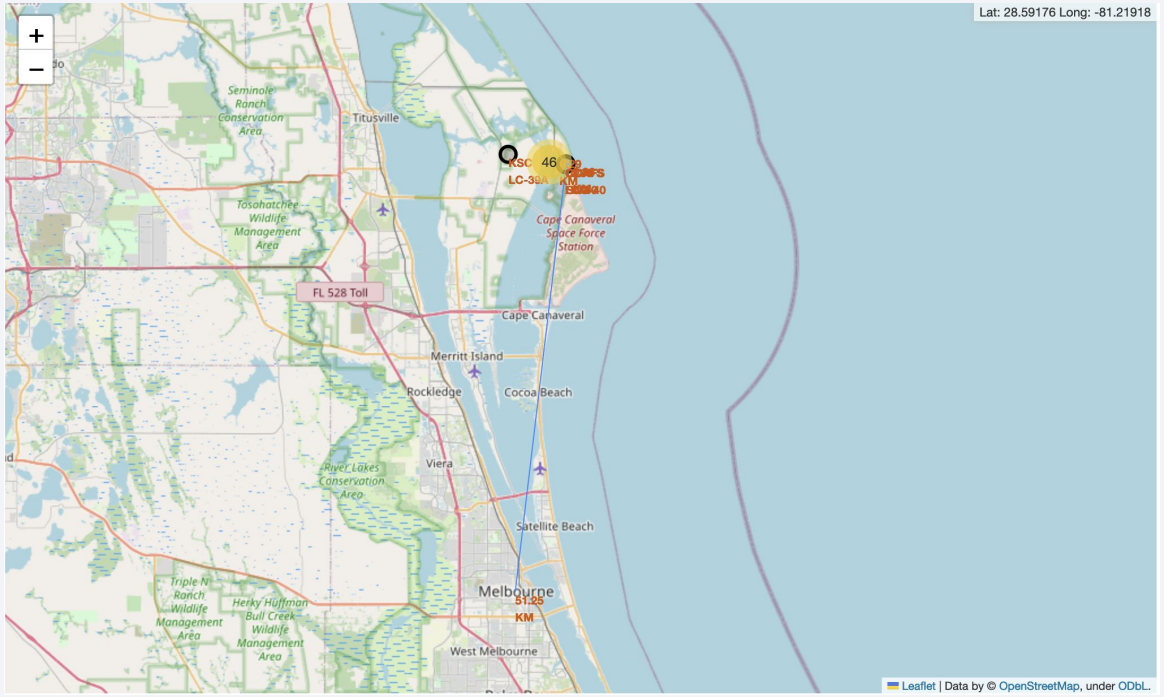
Distances between a Launch Site to its Proximities

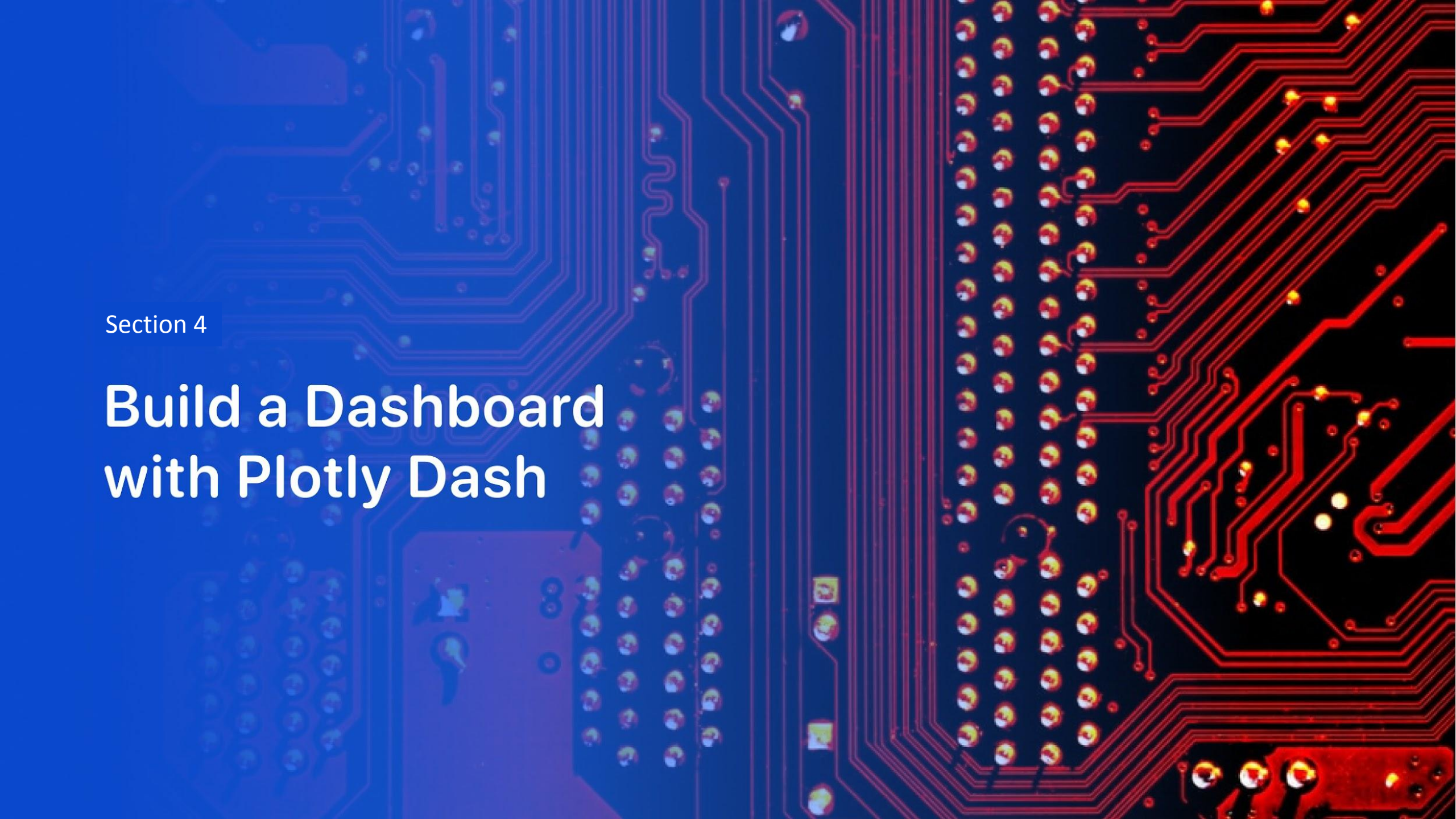
- CCAFS SLC 40
- Coastline:
  - 0.86 km
- Highway:
  - Samuel C Phillips Parkway
  - 0.60 km
- Railway:
  - NASA Railroad
  - 1.29 km
- City:
  - Melbourne
  - 51.25 km
- Close to railway, highway, and coastline, but furthest from the nearest city.



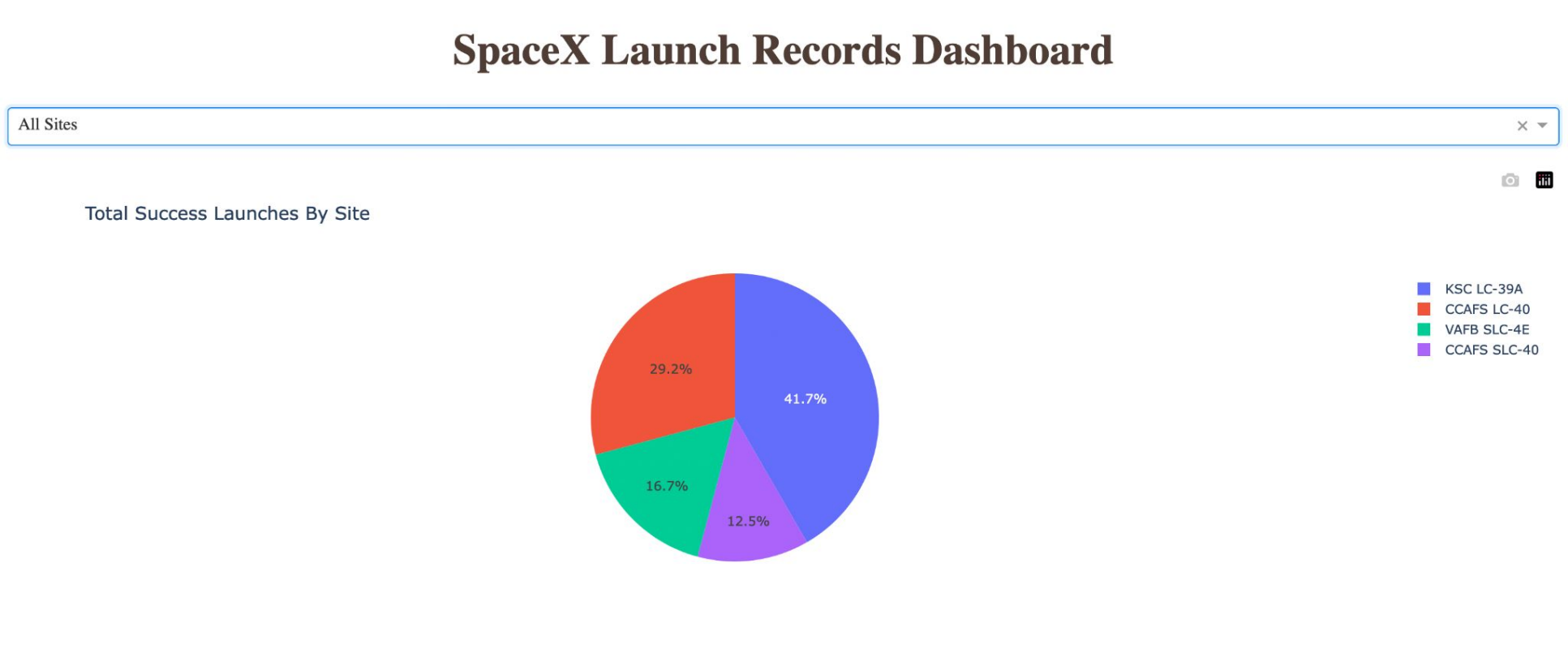
Distances between a Launch Site to its Proximities





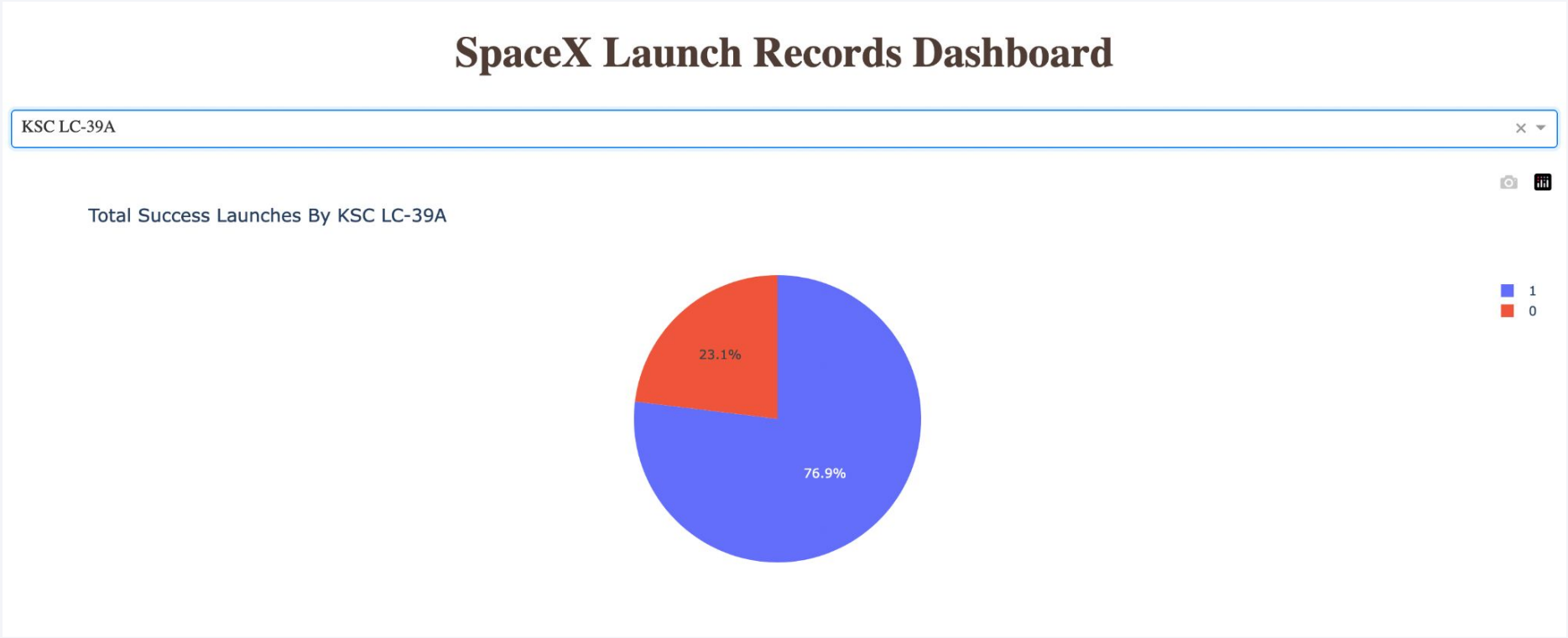

Section 4
Build a Dashboard with Plotly Dash

Total Success Launches for All Sites



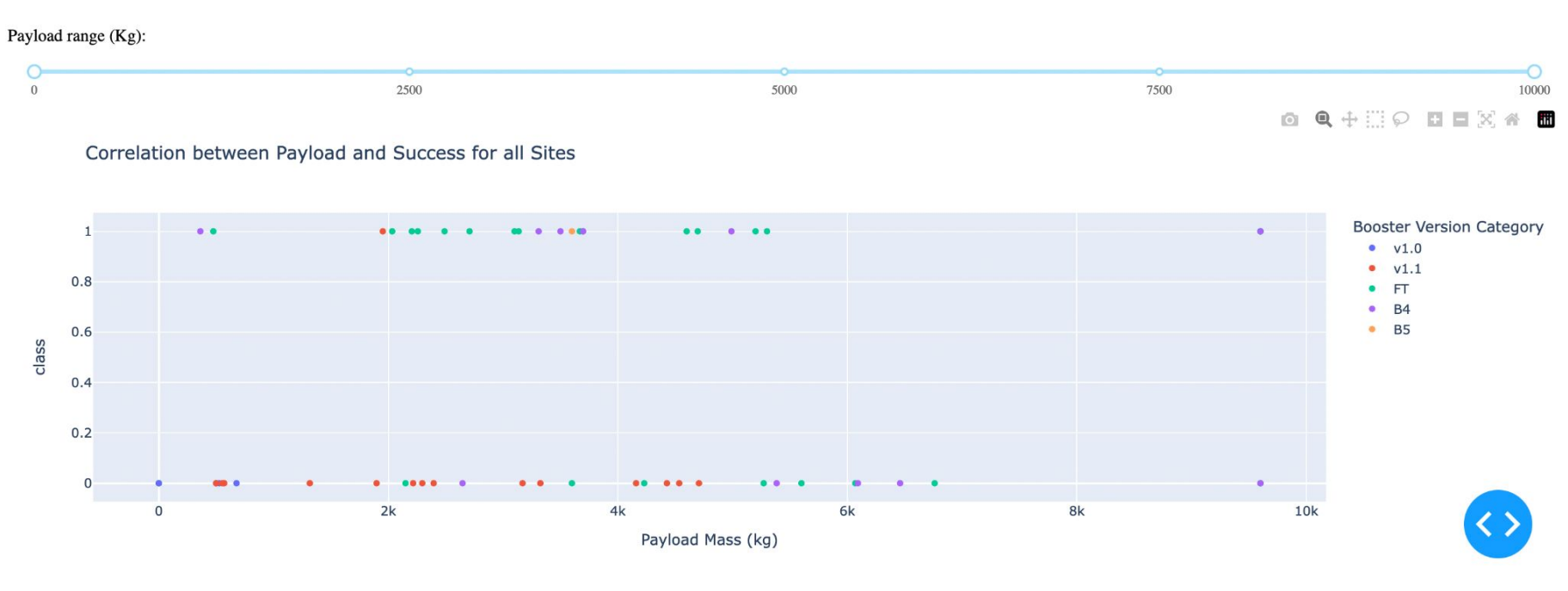
Out of the 4 launch sites, KSC LC-39A has the most successful launches, dominating 41.7% of all successful launches made.
CCAFS SLC-40 has the least successful launches, measuring only 12.5% of all successful launches made.

Launch Site with the Highest Launch Success Ratio



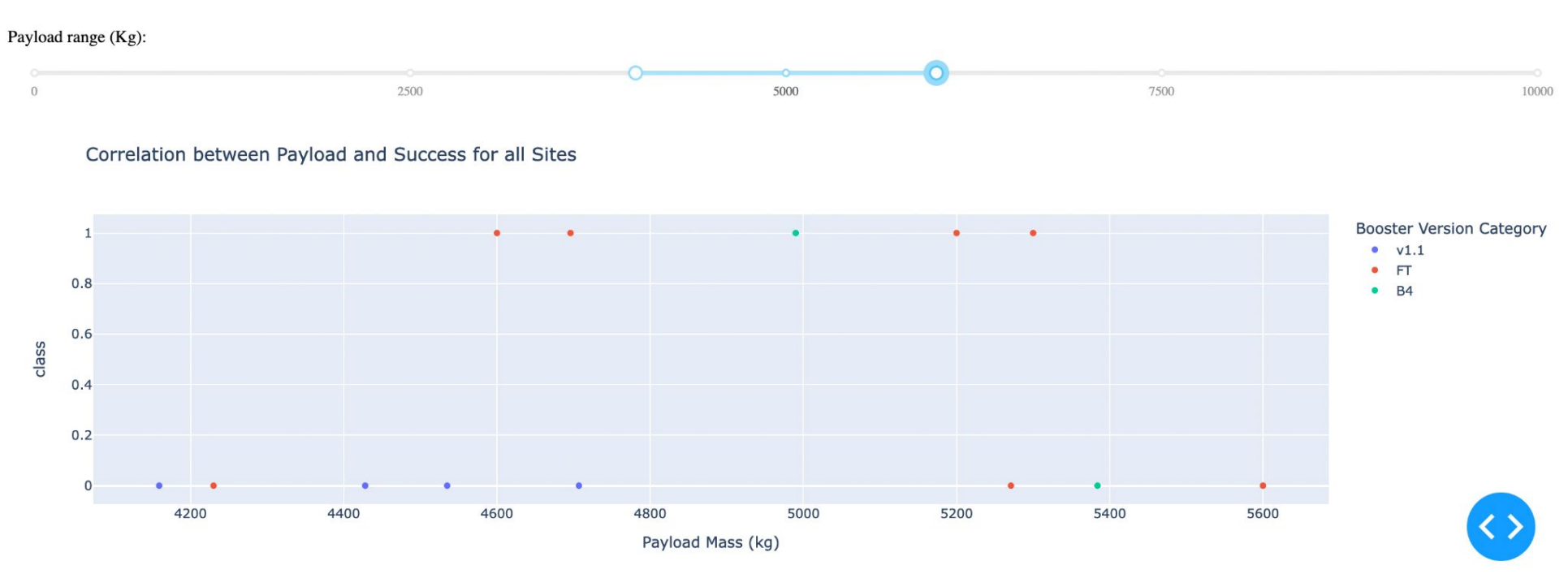
KSC LC-39A has a 76.9% success rate of launching SpaceX's rockets.

Payload vs. Launch Outcome for All Sites



Judging from the scatter plot above, the booster version FT had the most successful launches for all sites, with a payload mass ranging from 0-10000 kg. Meanwhile the booster version v1.1 had the least successful launches for all sites.

Payload vs. Launch Outcome for All Sites (Continued)



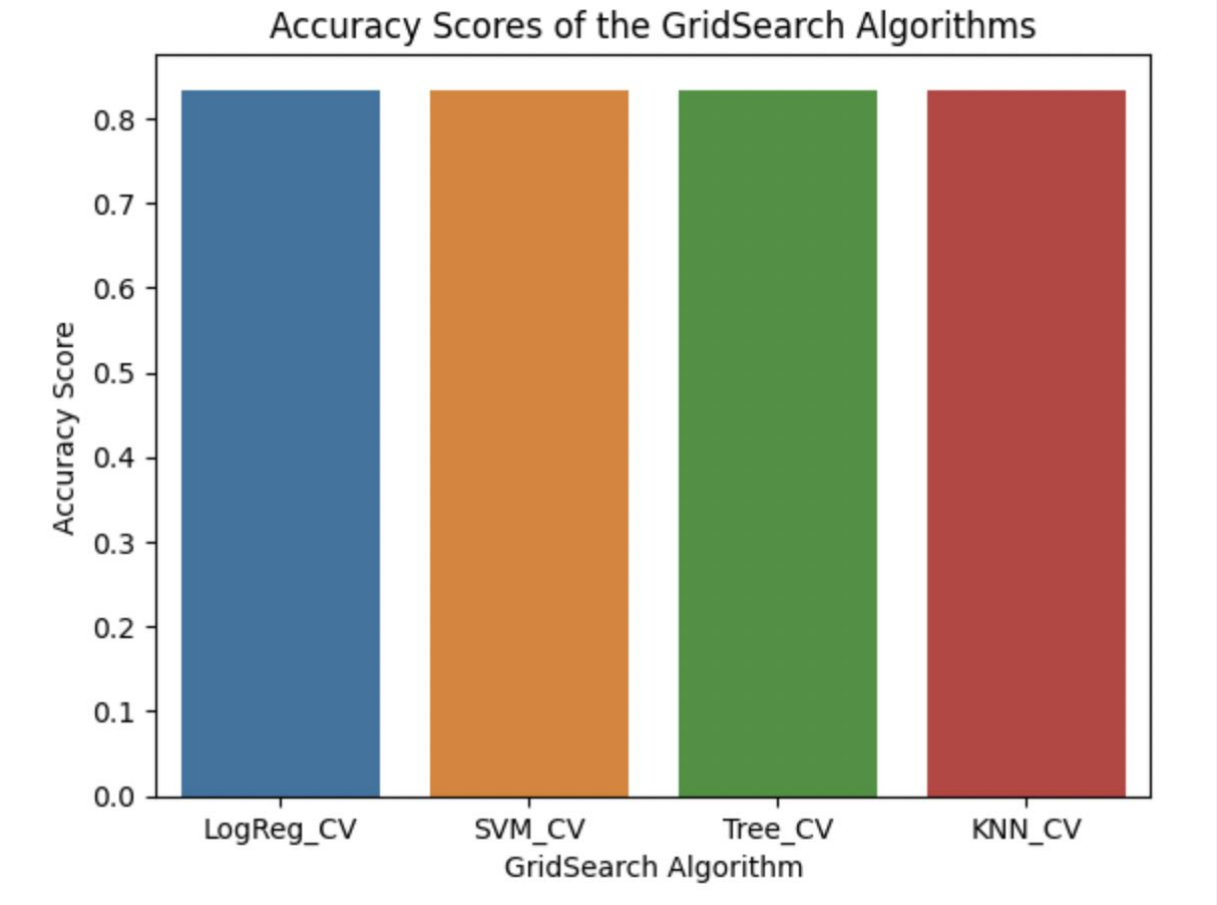
There were only 3 booster versions for the payload mass between 4000 - 6000 kg. The booster version FT still had the most number of successful launches for all sites, while booster version v1.1 still had the most unsuccessful launches for all sites.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

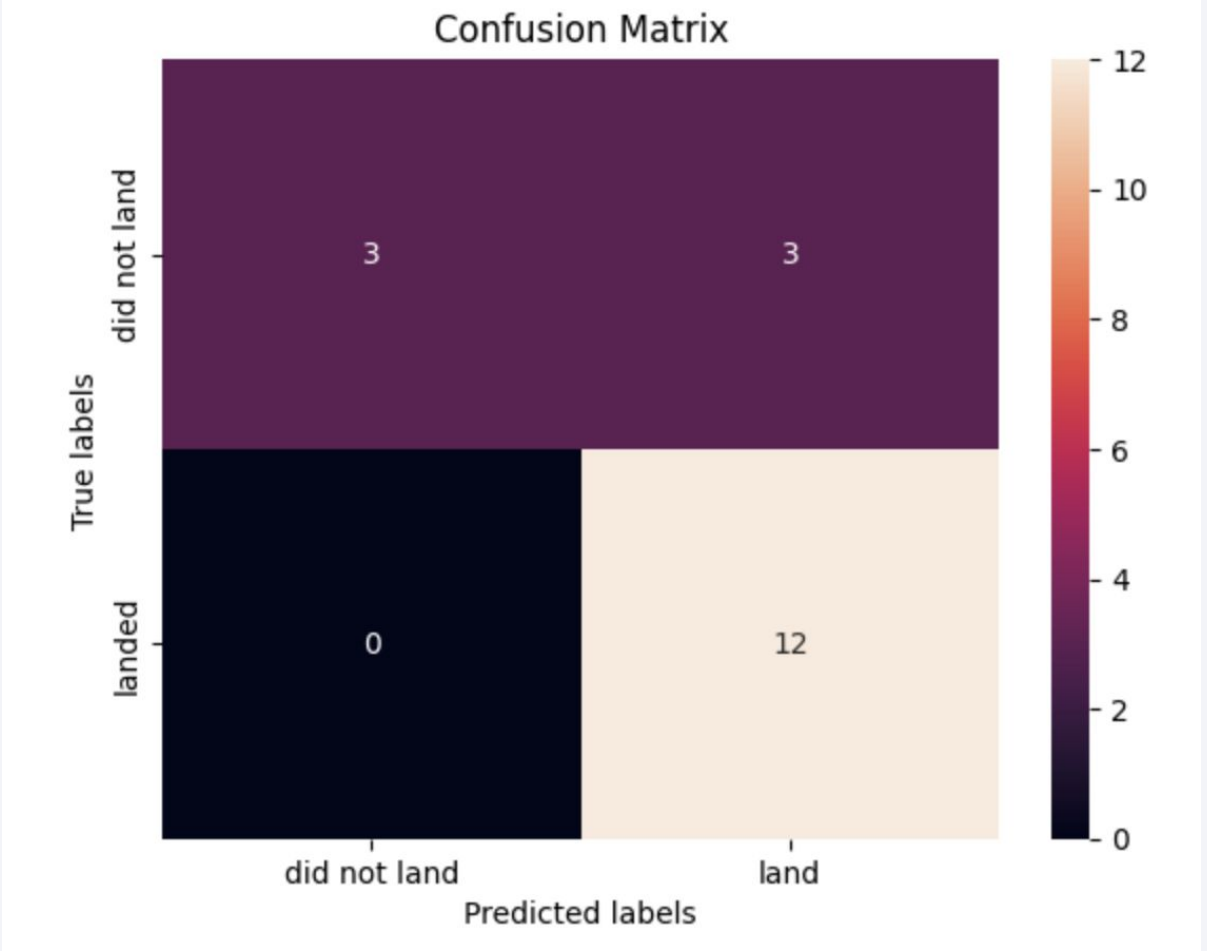
They all have the same accuracy score: 83.3%



Confusion Matrix

	Measurements	Values
0	Recall	1.000000
1	Precision	0.800000
2	Accuracy	0.833333
3	F-1 Score	0.888889

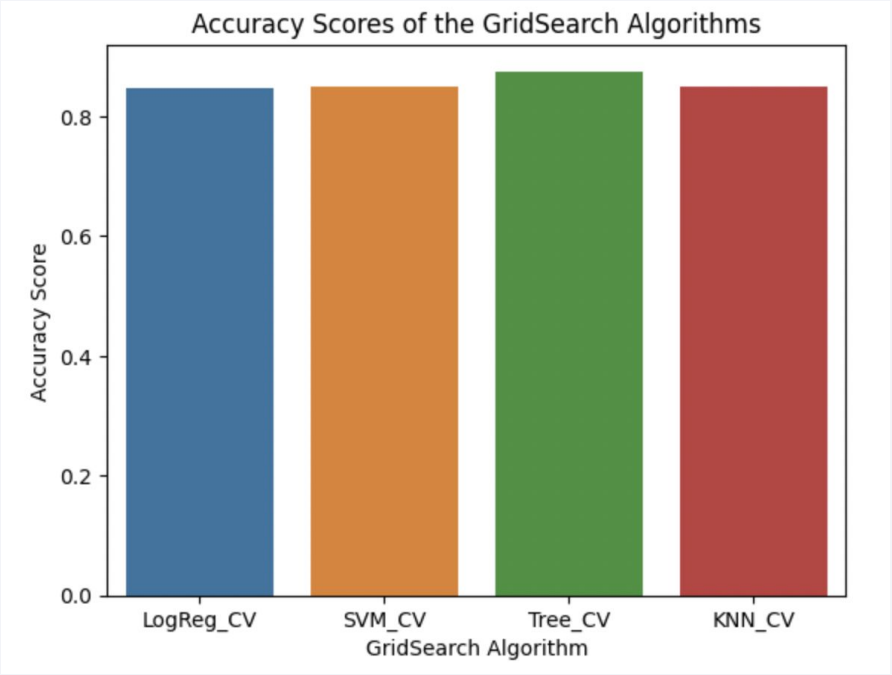
- **True Positive**: 12
- **True Negative**: 3
- **False Positive**: 0 *(Type 1 Error)*
- **False Negative**: 3 *(Type 2 Error)*



Conclusions

- Despite some failures, SpaceX's success rate of launching their rockets gradually increased from 2013 - 2020.
- The most successful launch site for SpaceX's Falcon 9 was the KSC LC-39A, dominating 41.7% of all the successful launches made from this project's dataset.
- According to the GridSearch algorithms and their confusion matrices, they all have a 83.3% accuracy in predicting the successful landing outcomes of all sites.
- A customer would want to look into SpaceX's launches at KSC LC-39A, given the data analysis and results performed in this capstone.

Appendix



The accuracy scores (best_score_ function) of the GridSearch algorithms

```
#Evaluation Metrics
Recall = 12/(12+0)
Precision = 12/(12+3)
Accuracy = (12+3)/18
F_1 = (2 * Recall * Precision) / (Recall + Precision)

eval_dict = {'Measurements': ['Recall', 'Precision', 'Accuracy', 'F-1 Score'], 'Values': [Recall, Precision, Accuracy, F_1]}
eval_df = pd.DataFrame(data = eval_dict)
eval_df
```

	Measurements	Values
0	Recall	1.000000
1	Precision	0.800000
2	Accuracy	0.833333
3	F-1 Score	0.888889

Python functions/methods used to produce the evaluation metrics.

Thank you!