Winning Space Race with Data Science

Phan Yen Thy Bui
April 22, 2024

Outline

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

Executive Summary

- This report follows the following steps to arrive at conclusion:
  - Data Collection (through API and Web Scraping)
  - Data Wrangling
  - Explanatory Data Analysis
  - Interactive Data Visualization
  - Predictive Analysis (ML Modeling)
- Summary of all results
  - As the number of flights increases, the rate of success increases. That being said, most early flights did not bring good results.
  - The launch site KSC LC-39A had the highest number of successful launches, with 41.7% of the total successful launches, and a 76.9% success rate.
  - There were fewer successful launches with Heavier Payload (>4000kg).
  - Decision Tree is the best model to predict the outcome of launches.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars, which is significantly cheaper than other providers' cost of 165 million dollars each. It is claimed that much of the savings is because Space X 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 space X for a rocket launch.

In this report, we will examine the factors affecting the success of a rocket launch from past data, to ultimately build a Machine Learning model to predict if the first stage will land successfully given the data provided.

Section 1

Methodology

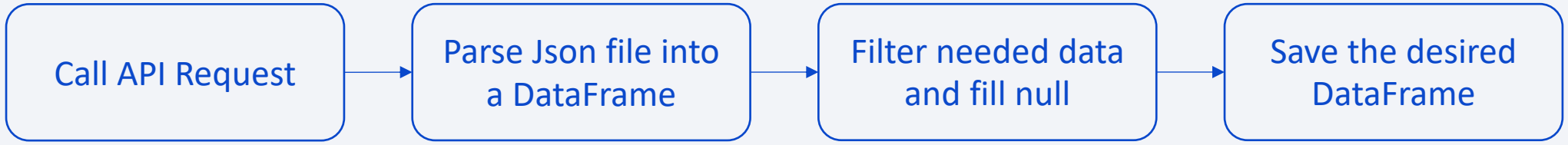
Methodology

Executive Summary

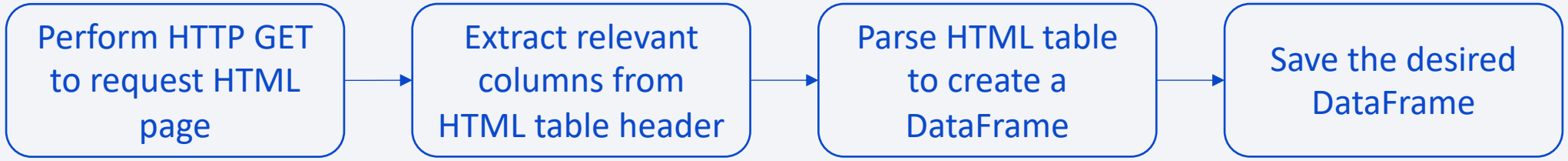
- Data collection methodology:
  - SpaceX Rest API.
  - Web Scaping from a Wikipedia page.
- Perform data wrangling
  - Clean and transform data to determine labels for training Supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Try different hyperparameters for different classification models to choose the best method.

Data Collection

- Datasets were collected from using:
  - SpaceX Rest API (https://api.spacexdata.com/v4/launches/past)

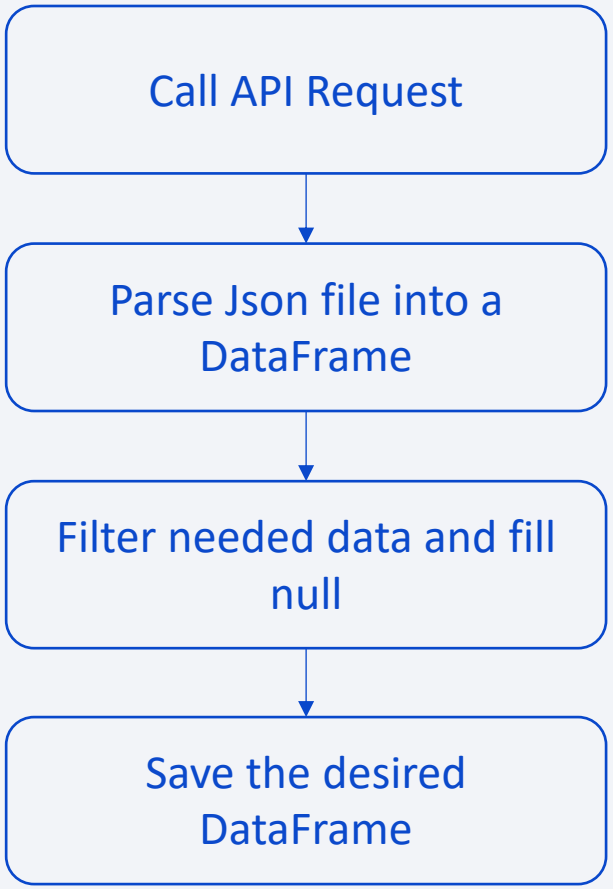


  - Web Scaping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)



Data Collection – SpaceX API

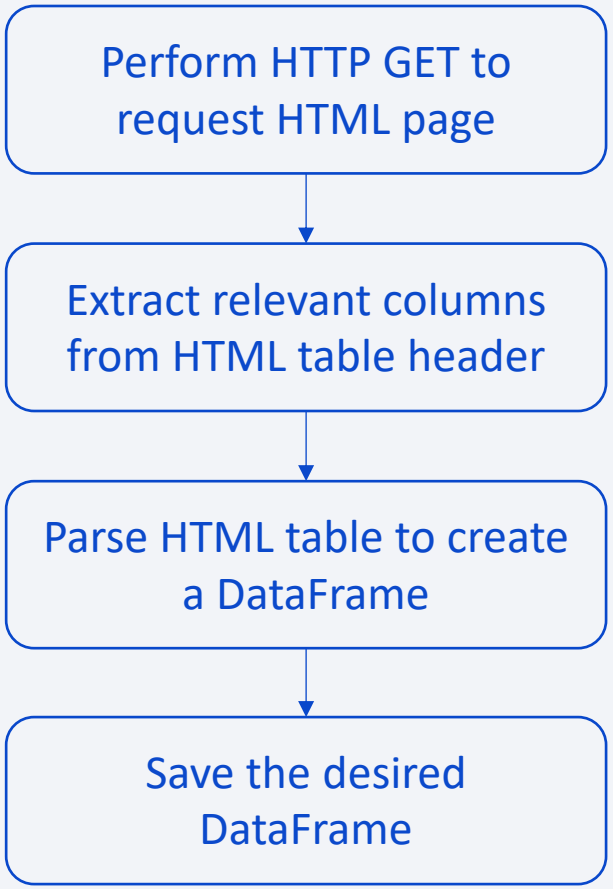
- Retrieve data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- GitHub URL: Data Collection_API



Data Collection - Scraping

- Collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.

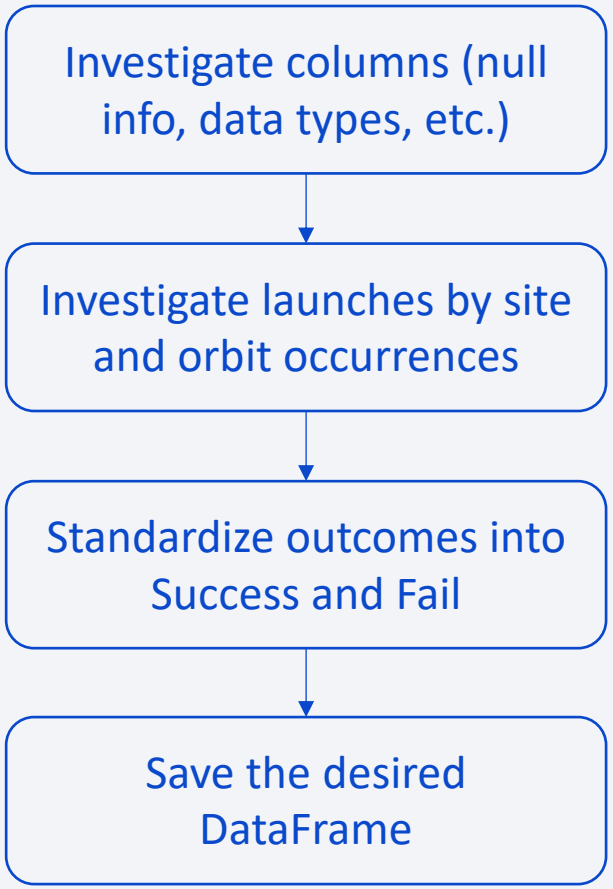
- GitHub URL: Data Collection_Webscraping



Data Wrangling

- The goal is to convert original outcome labels into landing class that represent landing classification which will be our new landing prediction target.
  - 1 for success
  - 0 for failure

- GitHub URL: Data_Wrangling



EDA with Data Visualization

Scatter Plot	Bar Chart	Line Chart
Scatter plots were produced to visualize the relationships between: • Flight Number and Launch Site • Payload and Launch Site • Orbit Type and Flight Number • Payload and Orbit Type	A bar chart was produced to visualize the relationship between: • Success Rate and Orbit Type	Line charts were produced to visualize the relationships between: • Success Rate and Year (i.e. the launch success yearly trend)
		
Scatter charts are useful to observe relationships, or correlations, between two numeric variables.	Bar charts are used to compare a numerical value to a categorical variable.	Line charts contain numerical values on both axes, and are generally used to show the change of a variable over time

- GitHub URL: EDA_Visualization

EDA with SQL

- The following SQL queries were performed:
  1. Display the names of the unique launch sites in the space mission
  2. Display 5 records where launch sites begin with the string 'CCA'
  3. Display the total payload mass carried by boosters launched by NASA (CRS)
  4. Display the average payload mass carried by booster version F9 v1.1
  5. List the date when the first successful landing outcome on a ground pad was achieved
  6. List the names of the boosters which had success on a drone ship and a payload mass between 4000 and 6000 kg
  7. List the total number of successful and failed mission outcomes
  8. List the names of the booster versions which have carried the maximum payload mass
  9. List the failed landing outcomes on drone ships, their booster versions, and launch site names for 2015
  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

- GitHub URL: EDA_SQL

Build an Interactive Map with Folium

- The map was built by doing the following steps:

1. Mark all launch sites on a map

- Initialize the map using a Folium Map object
- Add a folium.Circle and folium.Marker for each launch site on the launch map

2. Mark the success/failed launches for each site on a map

- As many launches have the same coordinates, it makes sense to cluster them together.
- Before clustering them, assign a marker color of successful (class = 1) as green, and failed (class = 0) as red.
- To put the launches into clusters, for each launch, add a folium.Marker to the MarkerCluster() object.
- Create an icon as a text label, assign the icon_color as the marker_color determined previously.

3. Calculate the distances between a launch site to its proximities

- To explore the proximities of launch sites, use Lat and Long values to calculate distances between points.
- After marking a point using the Lat and Long values, create a folium.Marker object to show the distance.
- To display the distance line between two points, draw a folium.PolyLine and add this to the map.

- GitHub URL: Interactive_Folium_Mark

Build a Dashboard with Plotly Dash

- The following charts were used to build the dashboard:
  - Pie chart - px.pie()
    - To show the total successful launches count for all sites, or
    - To show the Success vs. Failed counts if a particular site was selected.
  - Scatter plot - px.scatter()
    - To show the correlation between outcome (success or not) and payload mass (kg)
    - Can be filtered by ranges of payload masses using RangeSlider() object
    - Show points for each Booster Version
- GitHub URL: Interactive_Plotly_Dash

Predictive Analysis (Classification)

Data Preprocessing

- Load Data
- Determine input vs target variables
- Standardize input variables
- Split data into train and test sets

Model Development

- Define models and parameters
- For each model type (Logistic Regression, SVM, Tree Classification, KNN):
  - Create GridSearchCV() object
  - Fit the object using train datasets

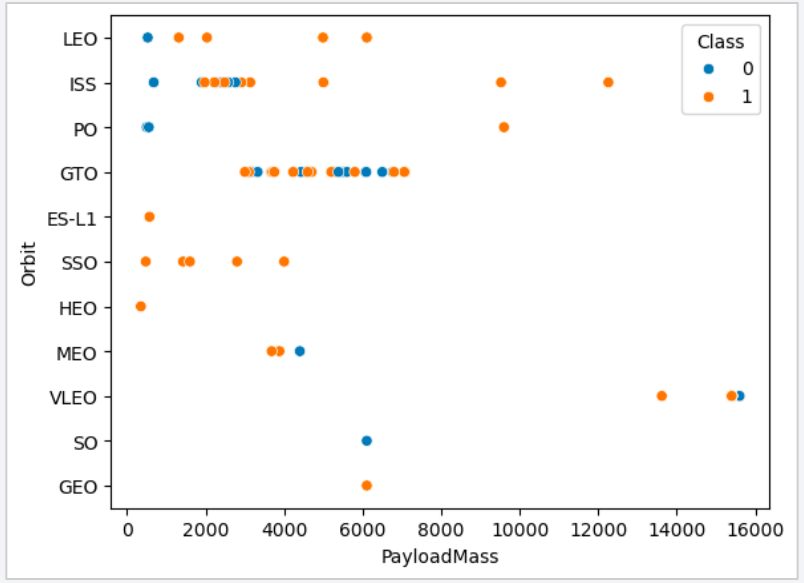
Model Evaluation

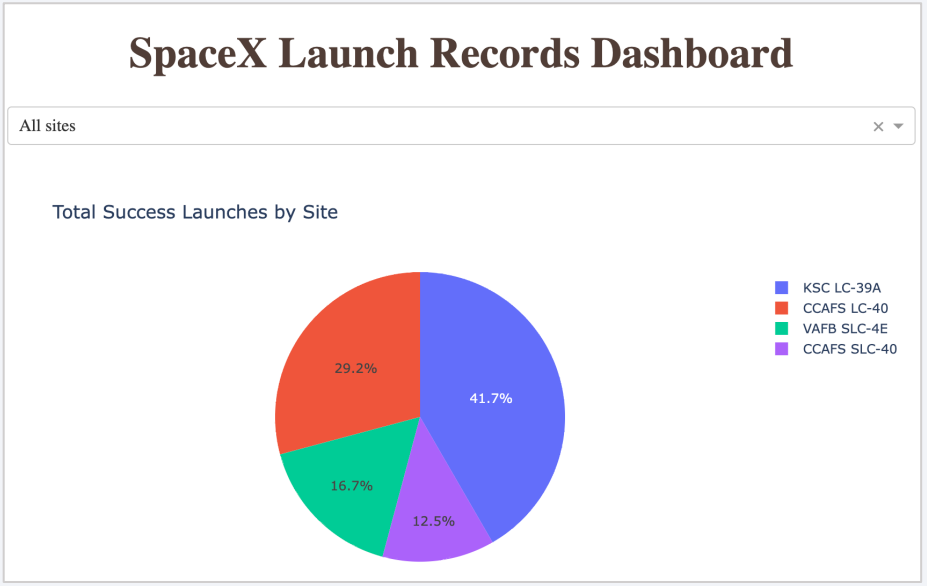
- For each model type:
  - Check the tuned hyperparameters (best_params_)
  - Check the accuracy (score and best_score_)
  - Plot the confusion matrix
- Review all models and determine the best performing one.

- GitHub URL: Machine_Learning_Prediction

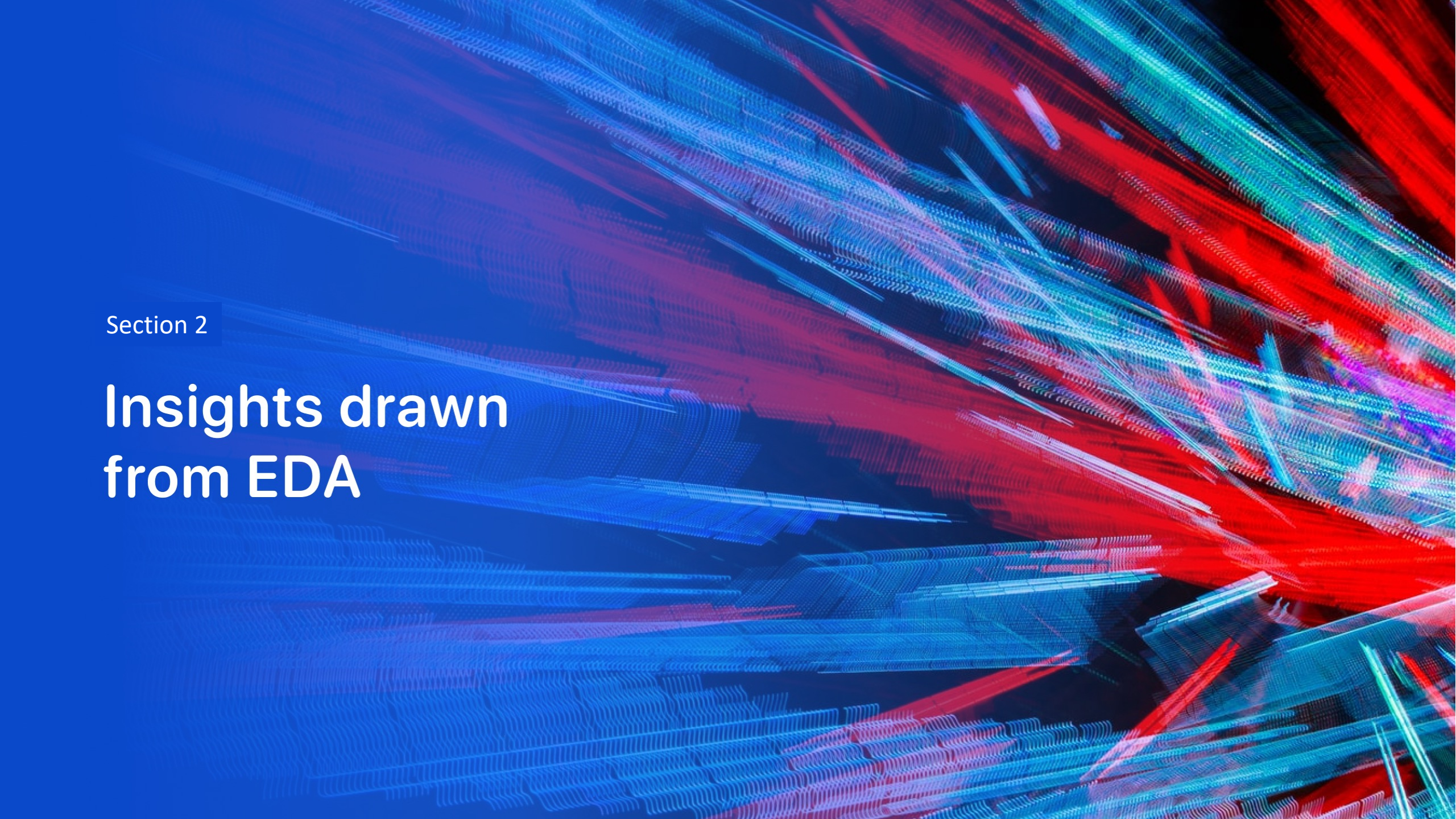
Results

Some snapshots of the results:



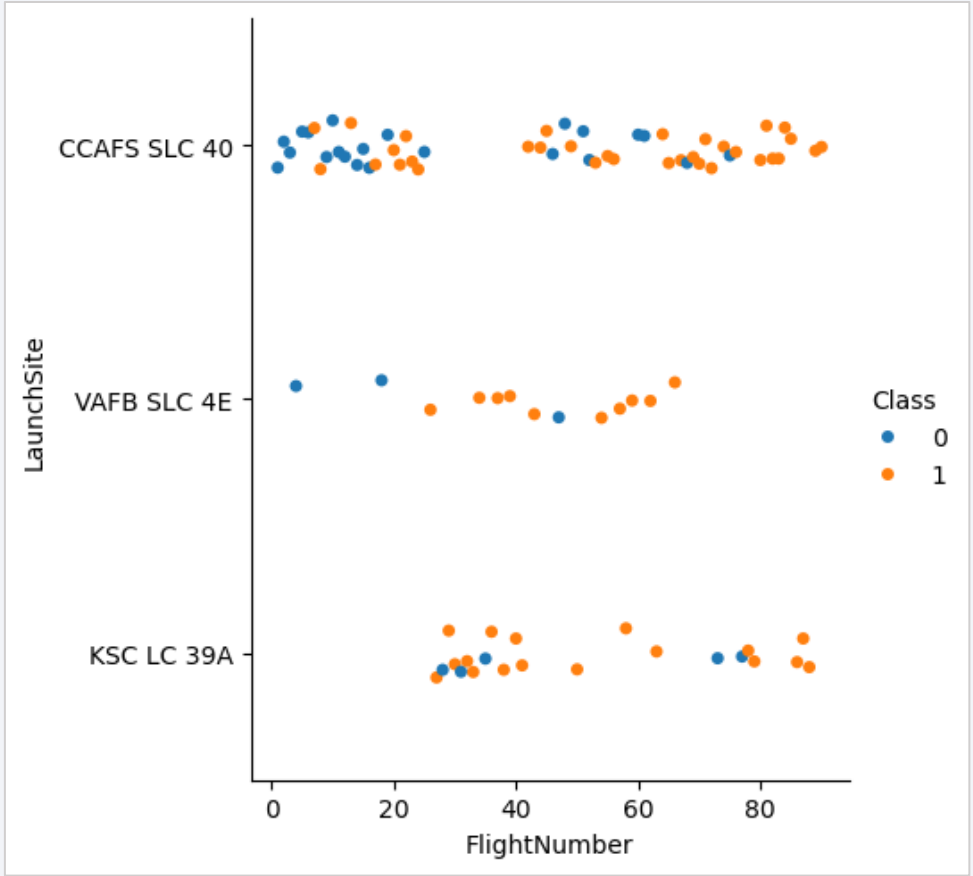


	train accuracy	test accuracy
lr	0.846429	0.833333
svm	0.848214	0.833333
tree	0.900000	0.944444
knn	0.848214	0.833333


Section 2
Insights drawn from EDA

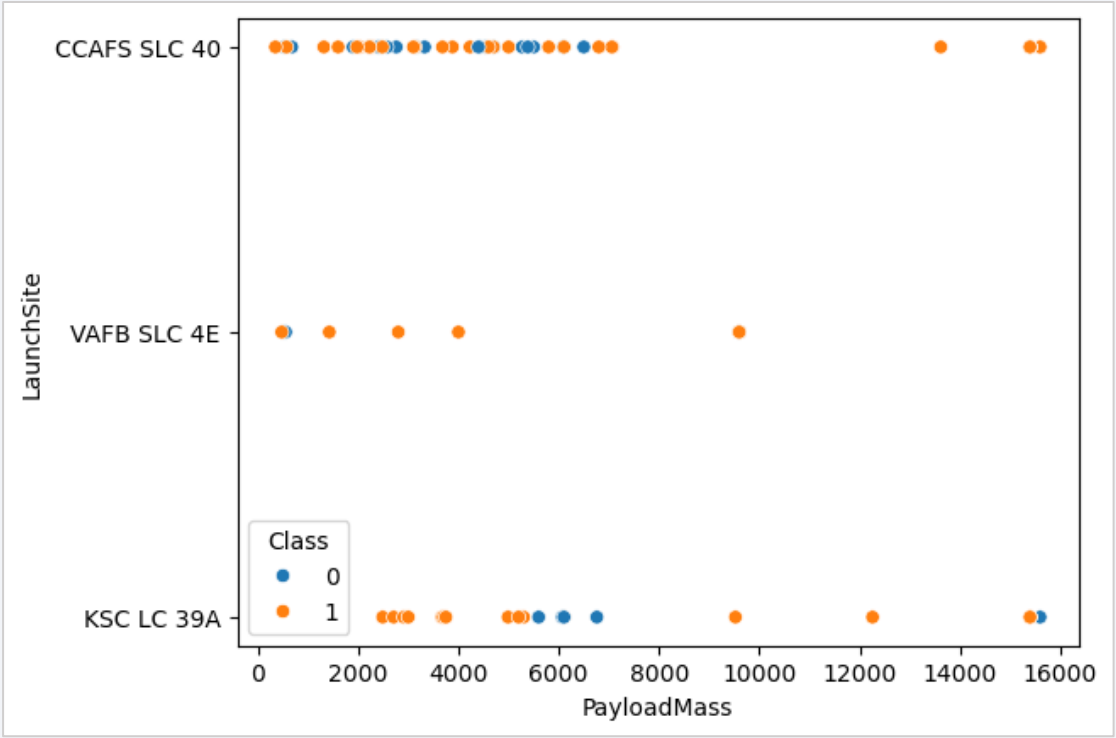
Flight Number vs. Launch Site

- As the number of flights increased, the rate of success at a launch site increased.
- CCAFS SLC 40 observed most of the early flights (flight numbers < 30) and they were mostly unsuccessful.
- There were fewer launches from VAFB SLC 4E and KSC LC 39A, but those that were not early flights had higher success rate.



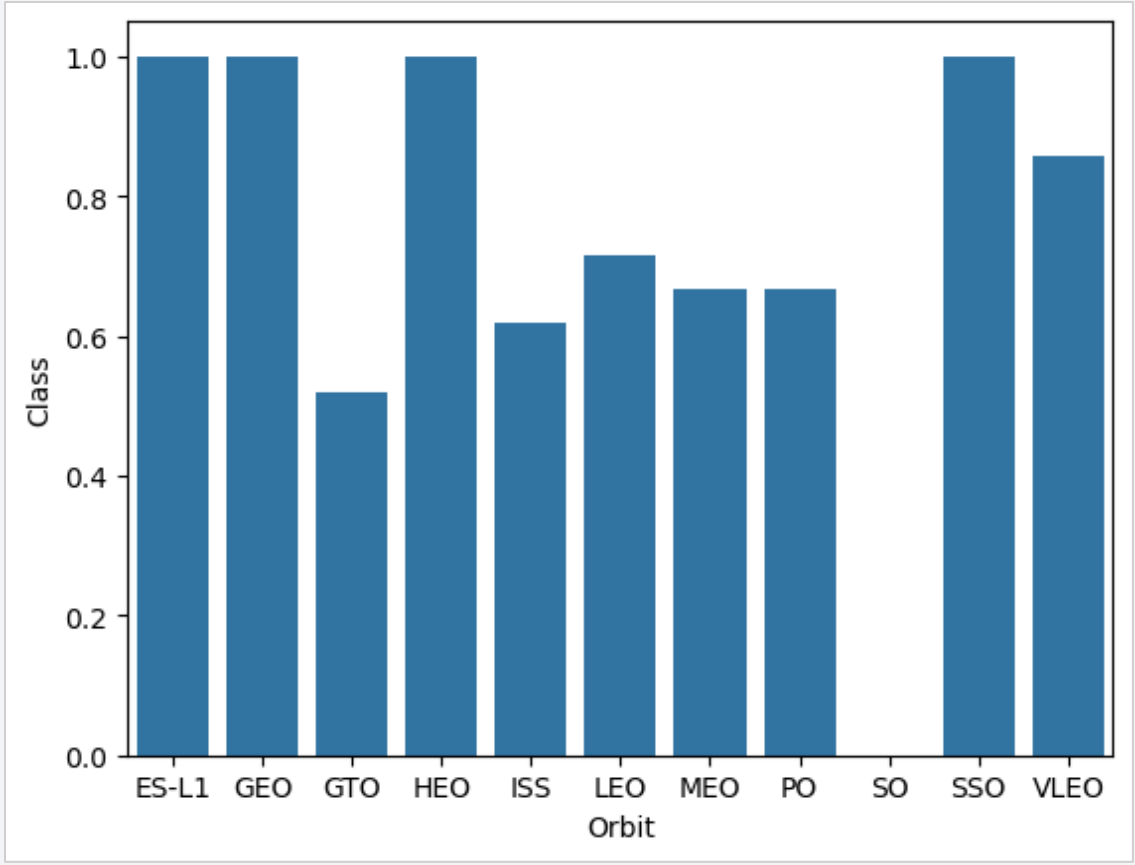
Payload vs. Launch Site

- In general, not many launches happened at a payload mass of around 7000 kg and up even though these launches were mostly successful.
- For the VAFB-SLC 4E launch site there were no rockets launched for heavier payload mass (greater than 10000).
- Most of the launches for CCAFS SLC 40 were with lighter payloads (less than 7500 kg).
- There was no clear correlation between Payload and Launch Site.



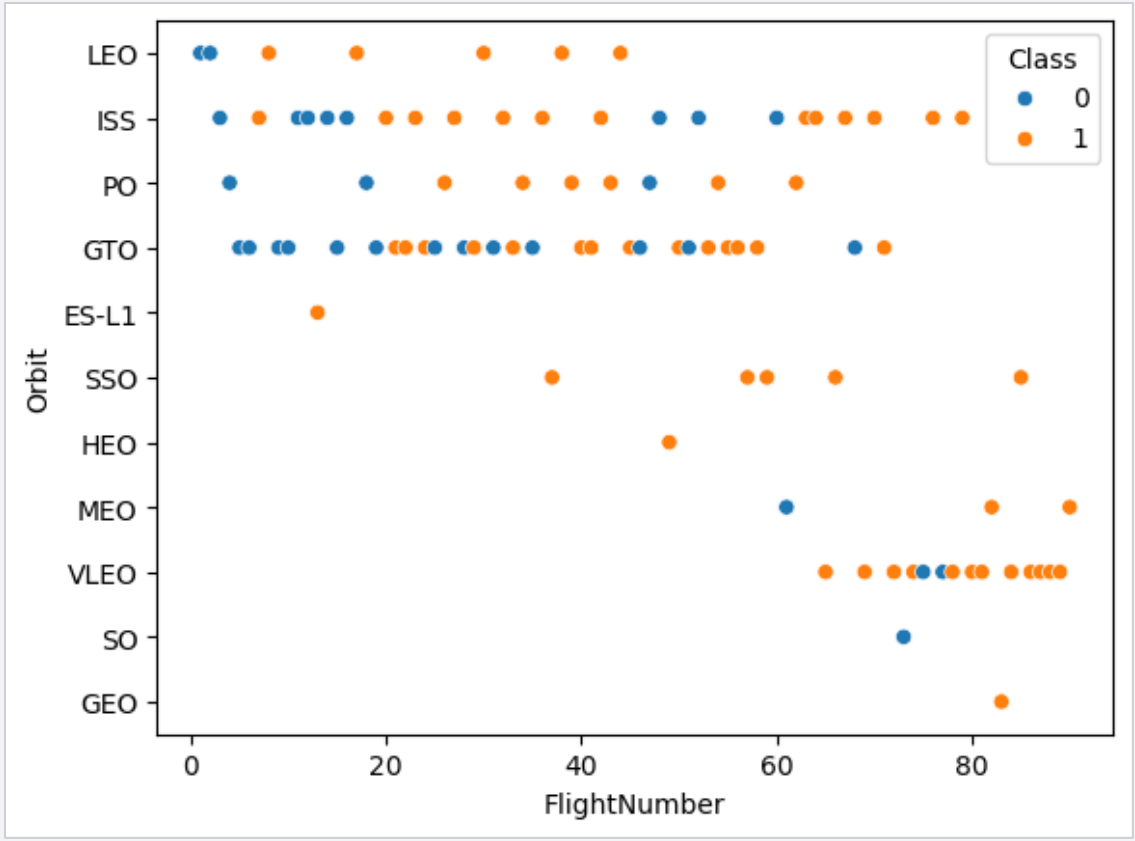
Success Rate vs. Orbit Type

- From the chart, we can see that ES-L1, GEO, HEO, and SSO all had a success rate = 100%.
- SO had a success rate = 0%.
- The other orbit types had a success rate of around 50 to 70%.



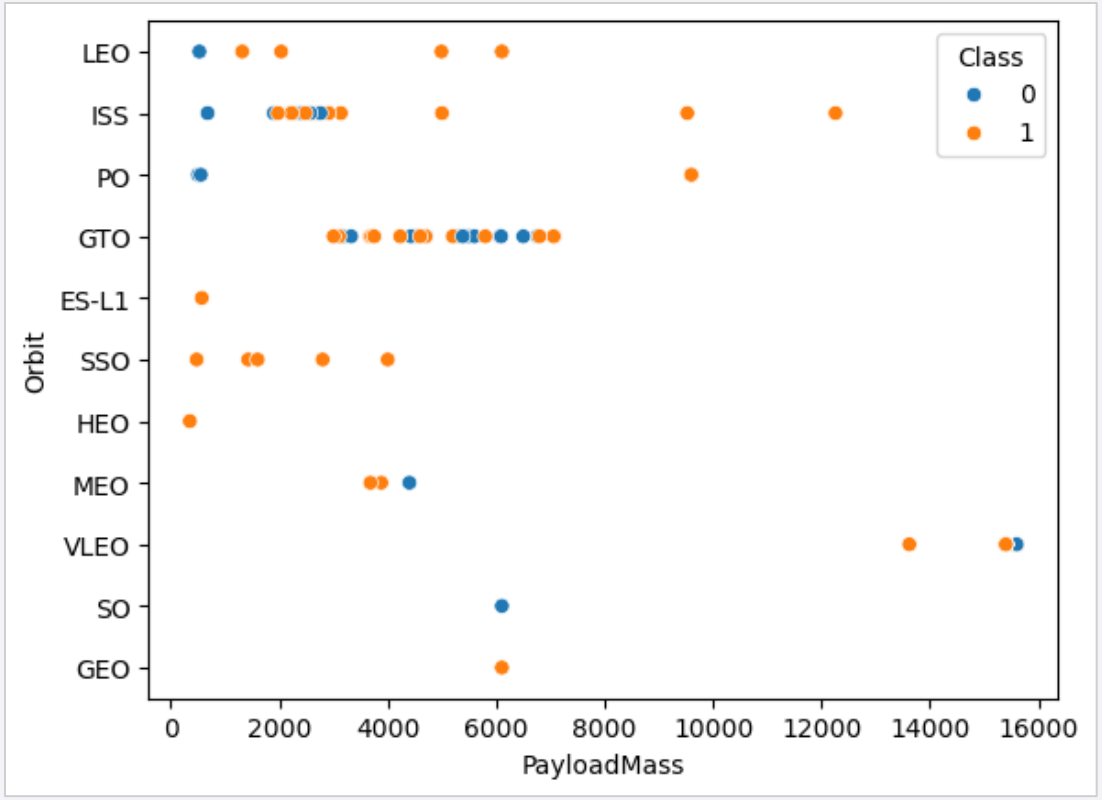
Flight Number vs. Orbit Type

- 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.
- The 100% success rate of GEO, HEO, and ES-L1orbits in the previous plot can be explained by only having 1 flight into the respective orbits, but on the other hand, SSO had 5/5 successful flights.



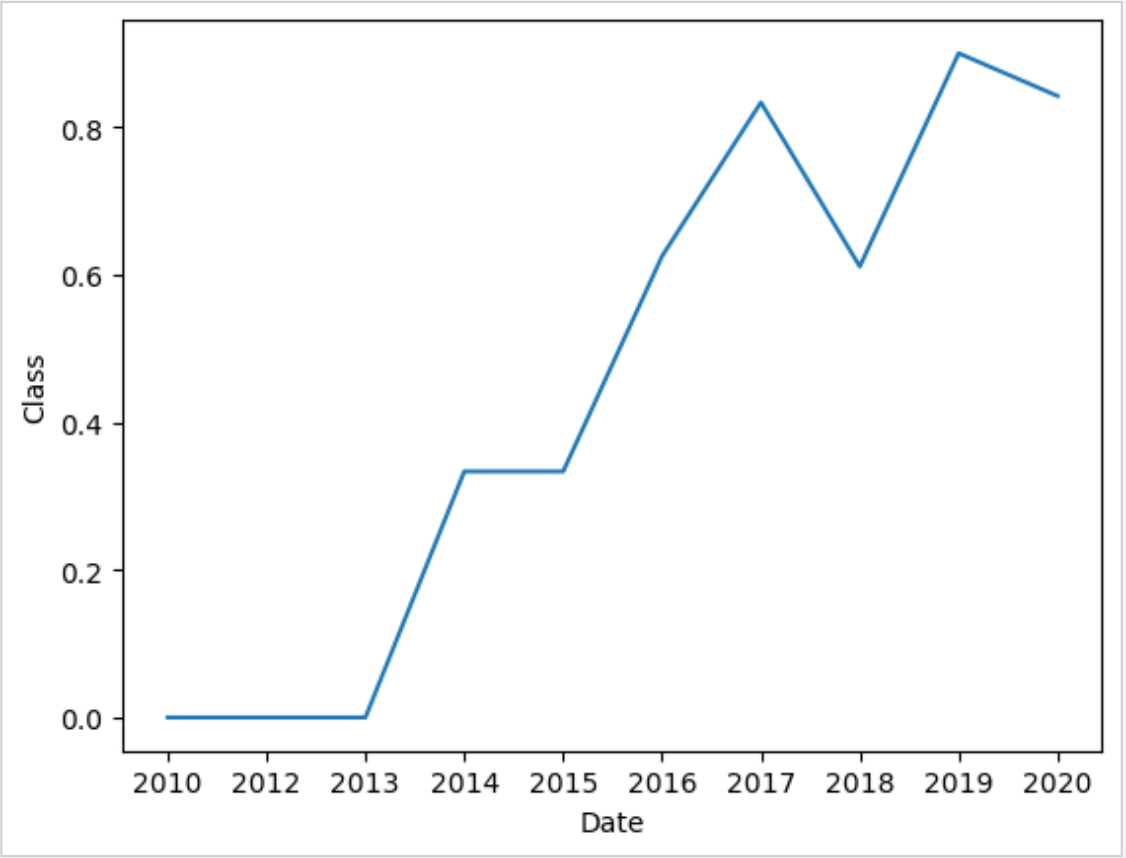
Payload vs. Orbit Type

- With heavy payloads the successful landing rate were more for PO, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing were here and there.
- All SSO launches were with lighter payload (<5000kg), while all VLEO launches were with high payload (>11000kg),



Launch Success Yearly Trend

- All launches from 2010 to 2013 were unsuccessful as the success rate = 0.
- The period of 2013 to 2017 observed continuous increases in success rate (at 80% in 2017).
- In 2018, the success rate decreased to 60%, but then picked up in the next years.



All Launch Site Names

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

```
In [10]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL
```

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

Out[10]:

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

There were four unique launch sites, including CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CAFS SLC-40.

Launch Site Names Begin with 'CCA'

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

In [11]:
```
%sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

Out[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (p
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 (p
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

The first five records that begin with 'CCA' were all from launch sites CCAFS LC-40, happened in LEO Orbit, and 4/5 of them were from Customer NASA.

Total Payload Mass

```
In [12]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
Out[12]:
```

total_payload_mass
45596

The total Payload Mass carried by boosters launched by NASA (CRS) was 45596kg.

Average Payload Mass by F9 v1.1

```
In [14]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1%'
 * sqlite:///my_data1.db
Done.
Out[14]:
```

avg_payload_mass
2534.6666666666665

The average Payload Mass carried by Booster Version F9 v1.1 was 2535kg.

First Successful Ground Landing Date

```
In [16]: %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
 * sqlite:///my_data1.db
Done.
Out[16]: MIN(Date)
         2015-12-22
```

The first successful ground landing date was December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

In [18]:
```
%%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL
    WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

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

Out[18]:

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

There were four unique Booster Versions that landed successfully with Payload between 4000 and 6000kg.

Total Number of Successful and Failure Mission Outcomes

In [23]:
```
%%sql SELECT CASE WHEN Mission_Outcome = 'Success ' THEN 'Success' ELSE Mission_Outcome END as Mission_Outcome,
            COUNT(*) as num_occurrences
    FROM SPACEXTBL GROUP BY 1
```

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

Out[23]:

Mission_Outcome	num_occurrences
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

There were 100 successful missions (among which one was with payload status unclear), and 1 failure.

Boosters Carried Maximum Payload

In [24]:

```
%%sql
WITH max_payload_mass AS (SELECT MAX(PAYLOAD_MASS__KG_) as max_payload FROM SPACEXTBL)

SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT max_payload FROM max_payload_mass)
```

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

Out[24]:

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 were 12 Boosters that carried the maximum payload.

2015 Launch Records

In [39]:
```
%%sql SELECT substr(Date, 6,2) as month, substr(Date,0,5) AS Year_, Booster_Version, Launch_Site
    FROM SPACEXTBL
    WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5) = '2015'
```

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

Out[39]:

month	Year_	Booster_Version	Launch_Site
01	2015	F9 v1.1 B1012	CCAFS LC-40
04	2015	F9 v1.1 B1015	CCAFS LC-40

In 2015, there were two failed drone ship, one in January and the other one in April. They were both launched from site CCAFS LC-40.

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

In [51]:
```
%%sql SELECT Landing_Outcome, COUNT(*) AS num_occurrences
FROM SPACEXTBL
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY 1
ORDER BY 2 DESC
```

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

Out[51]:

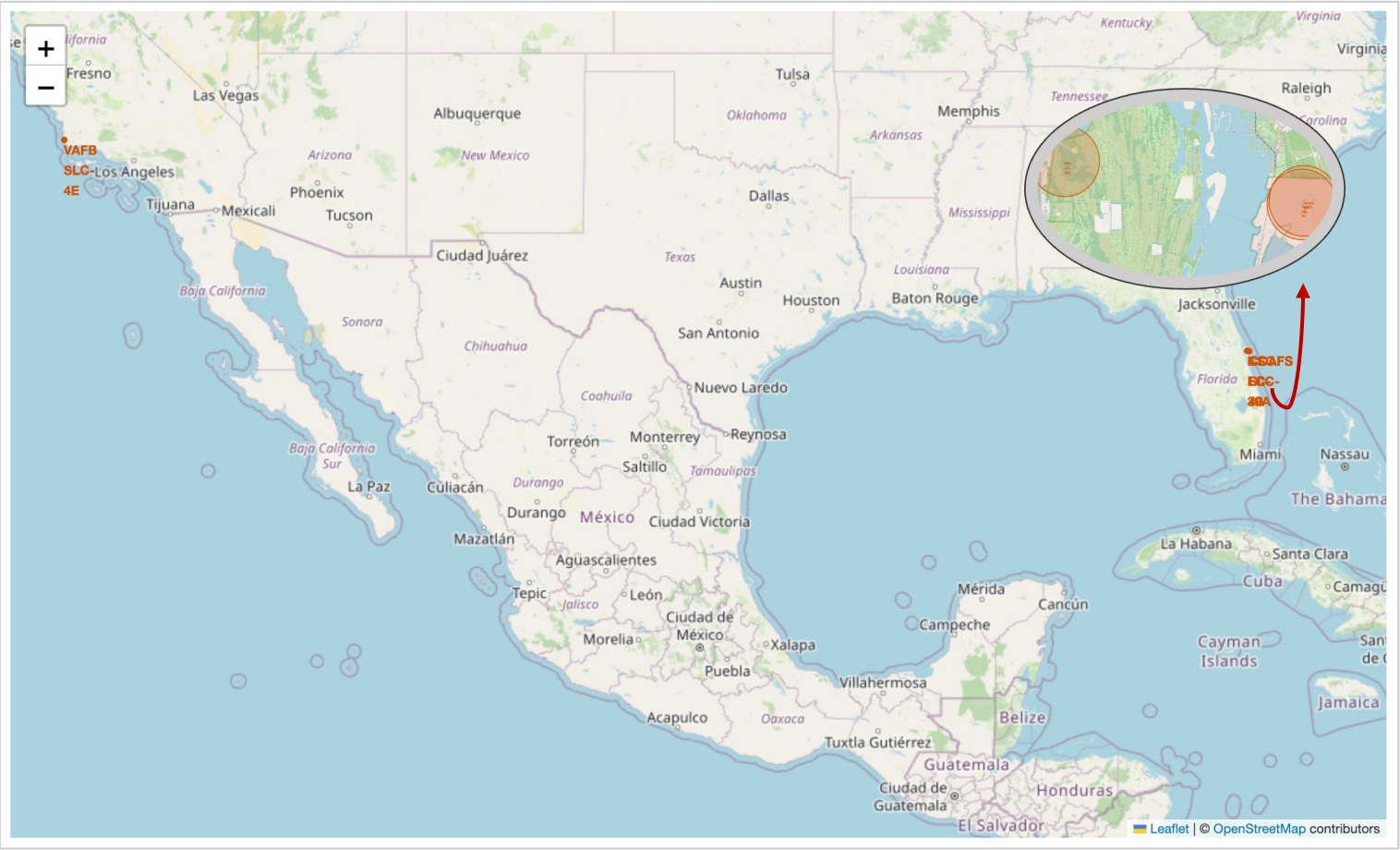
Landing_Outcome	num_occurrences
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

"No attempt" was at the top of the list, followed by "Success (drone ship)" and "Failure (drone ship)" between the period of June 4, 2010 and March 20, 2017.

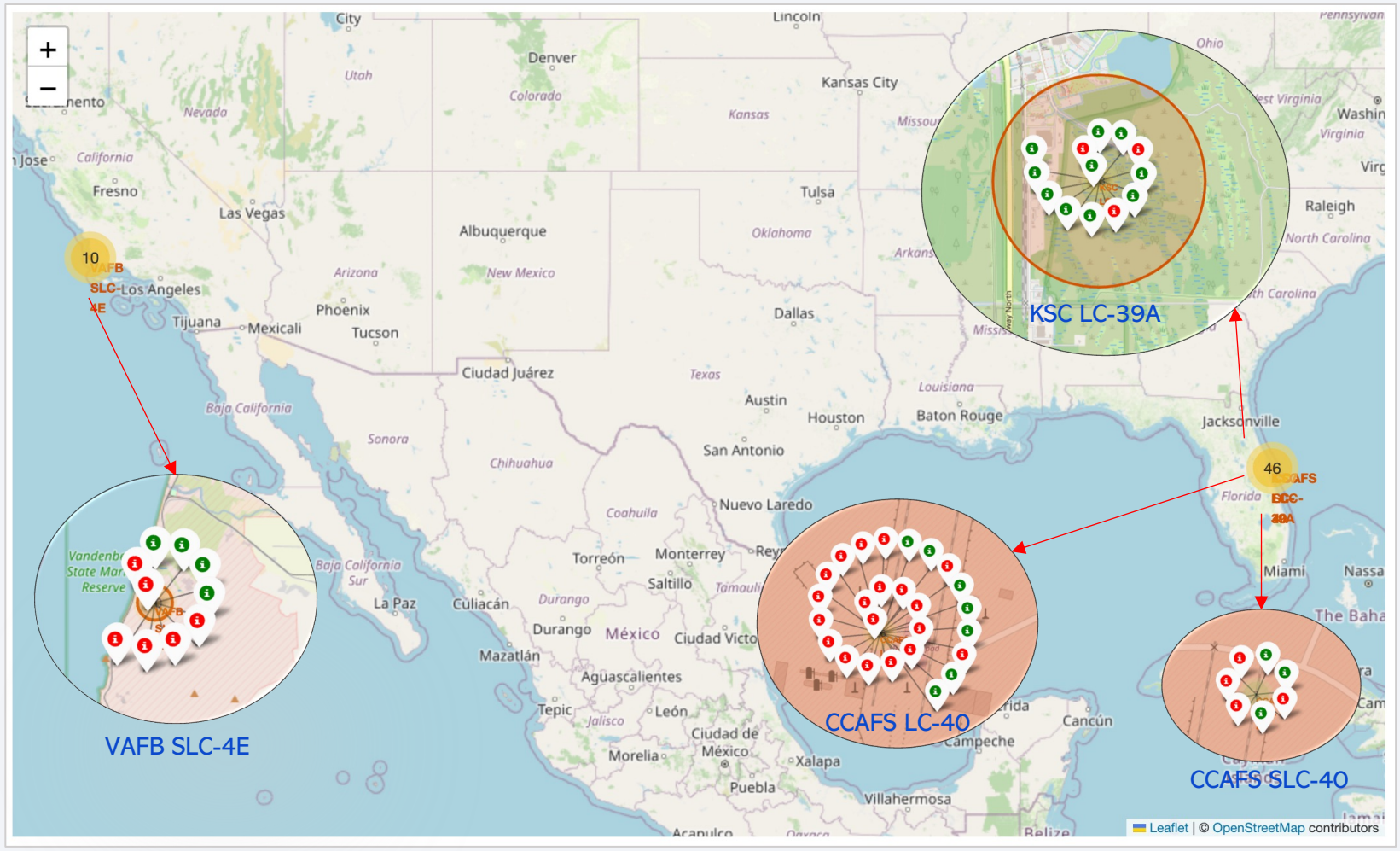

Section 3
Launch Sites
Proximities Analysis

All Launch Sites on a Map

- There were four launch sites. They were all located near the coastline.
- One was on the west coast and three were on the west coast.

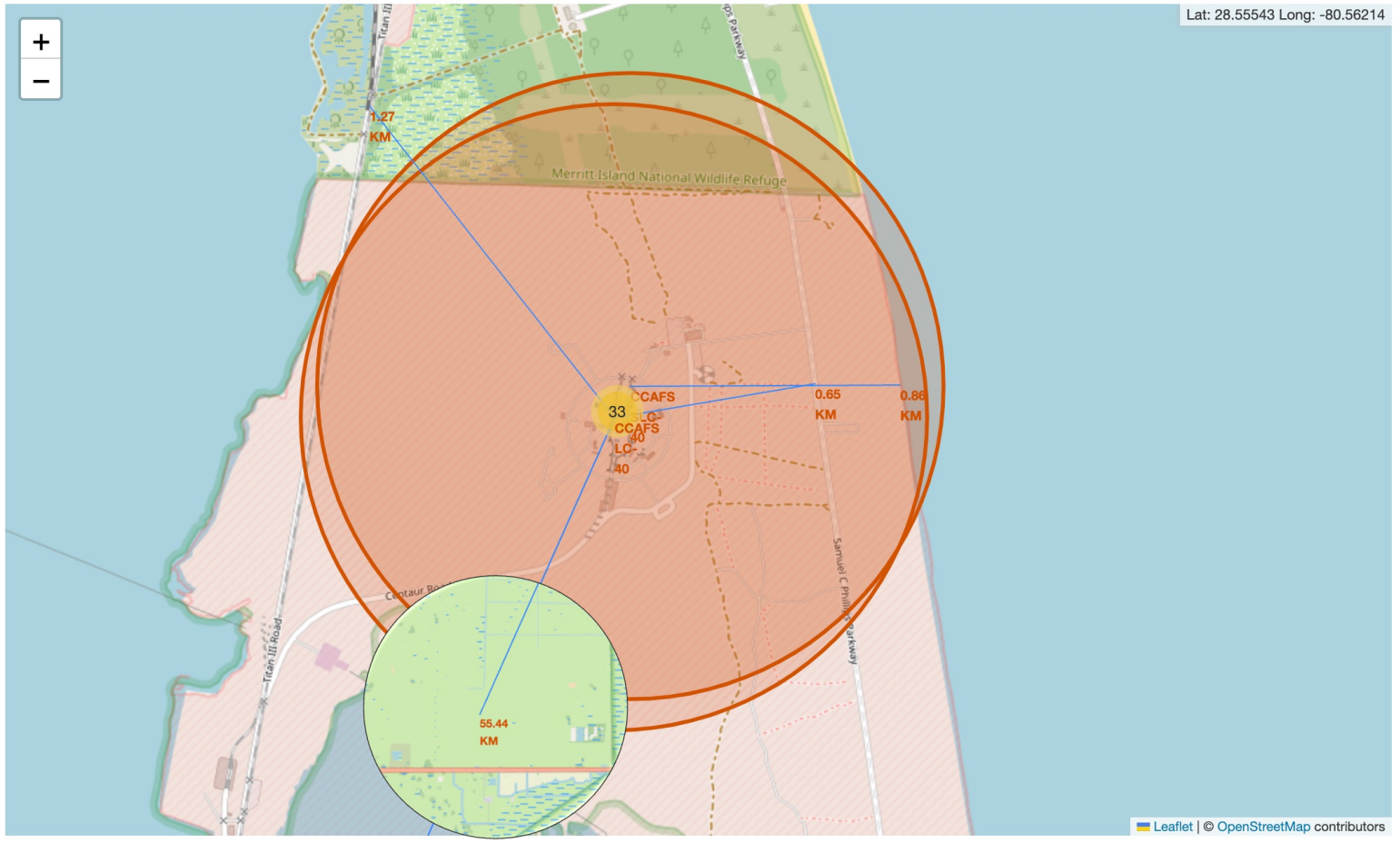


Successful/Failed Launches for Each Site



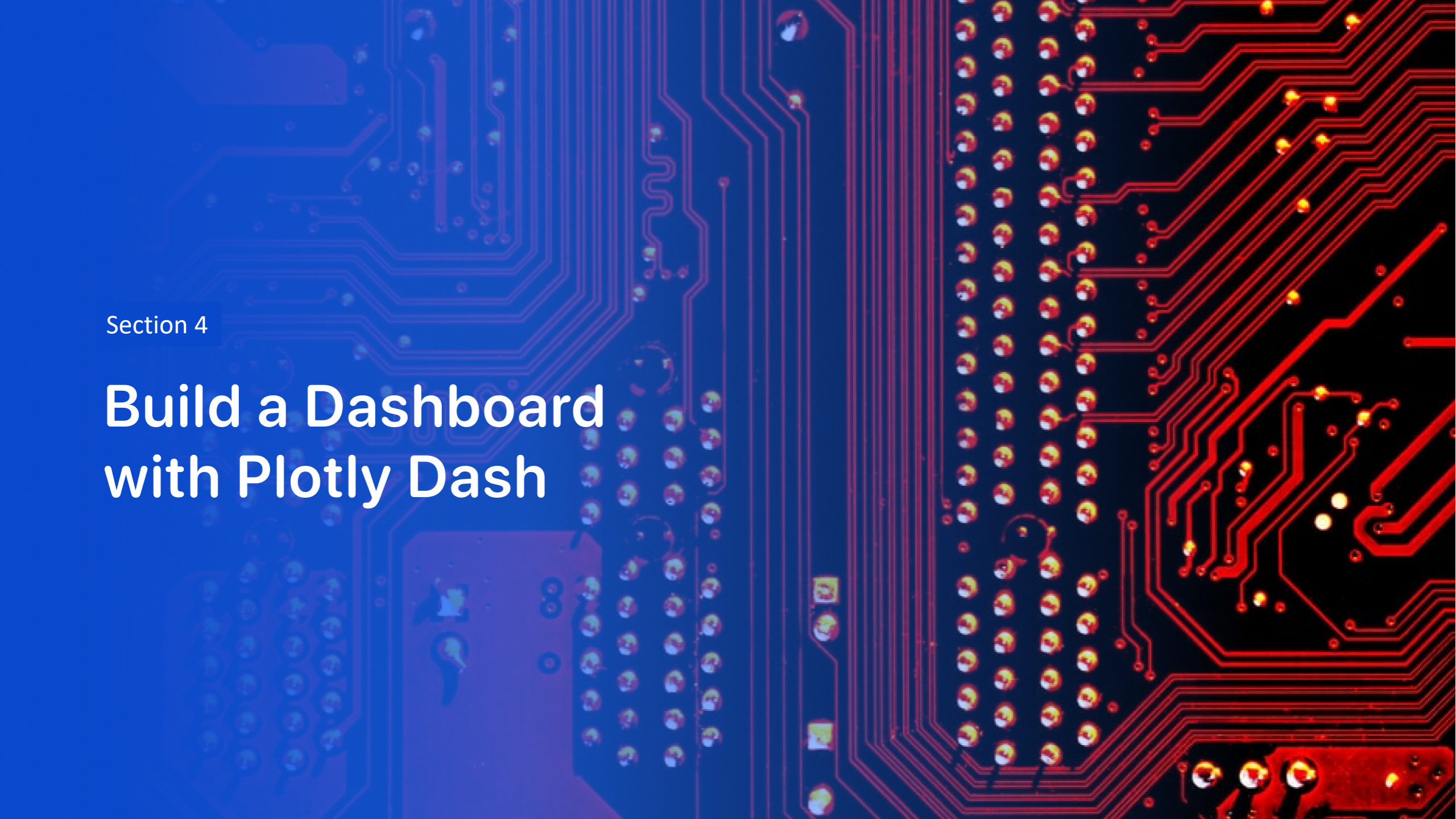
Launches have been grouped into 2 colors: green for success and red for fail. We can see from the map that KSC LC-39A has a higher success rate compared to the remaining three launching sites.

Proximity of Launch Sites to other POIs

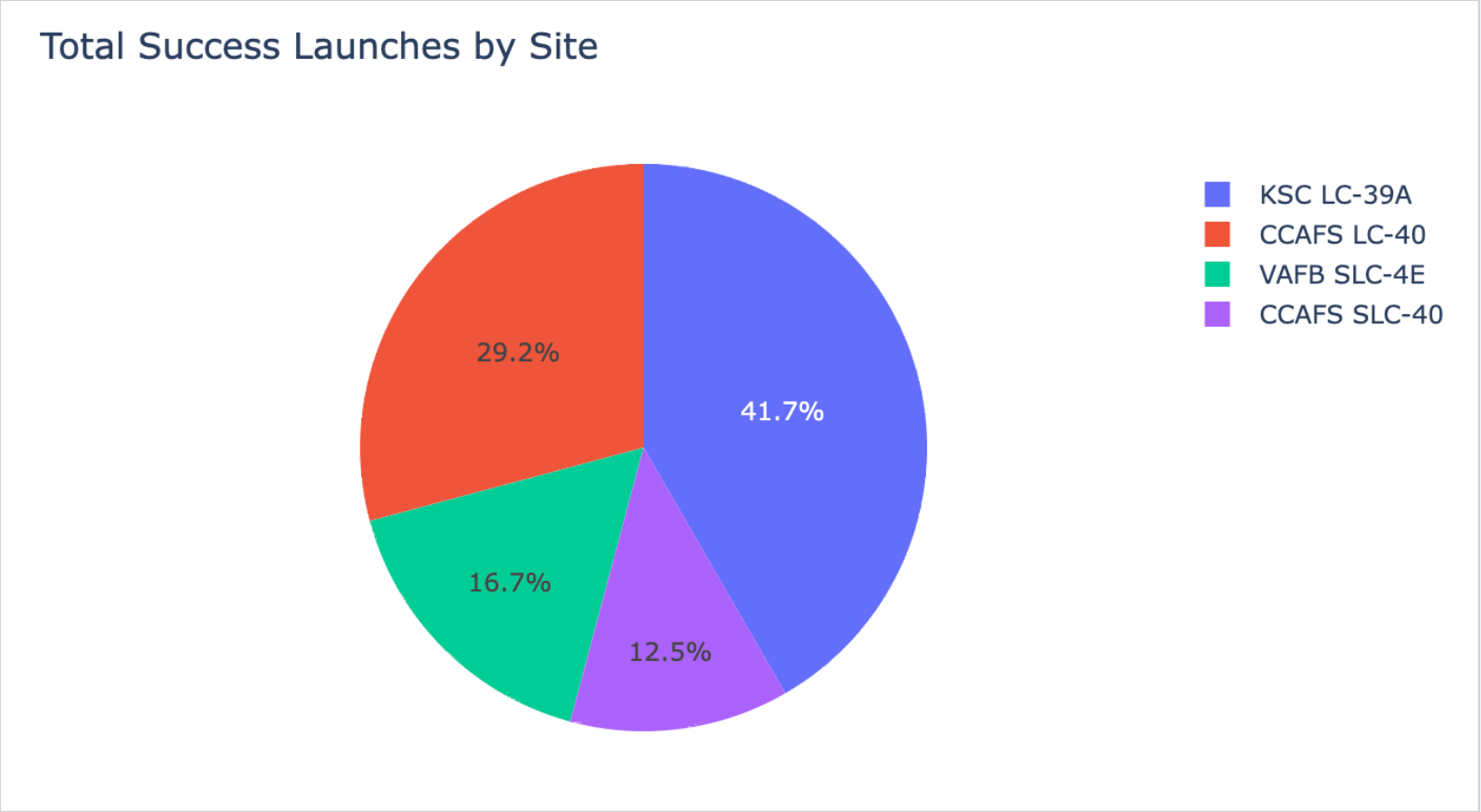


We'll use CCAFS LC-40 as the example site. The distances to other Point of Interest are as follows:

- 650m to Highway
- 1.27km to Railway
- 55.44km to cities

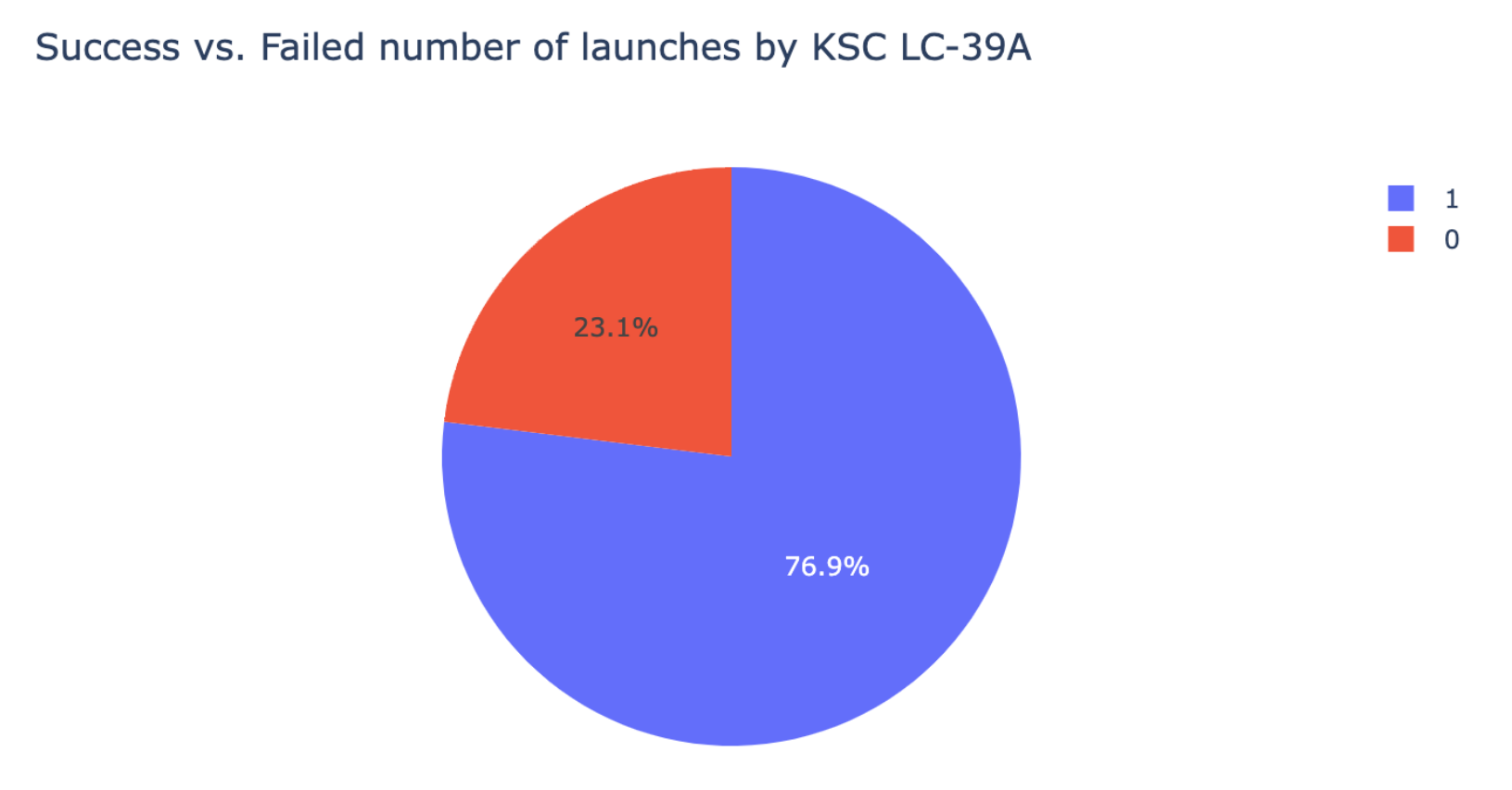

Section 4
Build a Dashboard with Plotly Dash

Launch Success Summary for All Sites



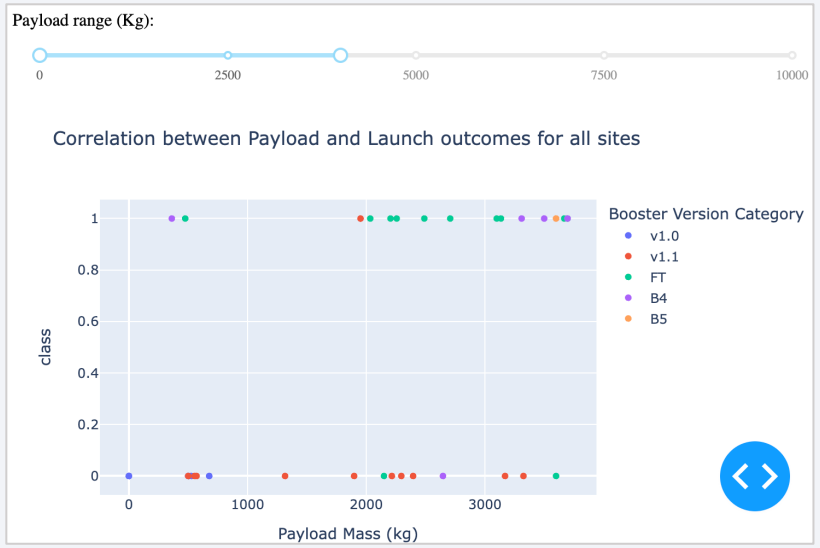
KSC LC-39A had the highest number of successful launches, accounted for 41.7% of total success, while CCAFS SLC-40 has the lowest number of success launches.

Launch Site with Highest Launch Success

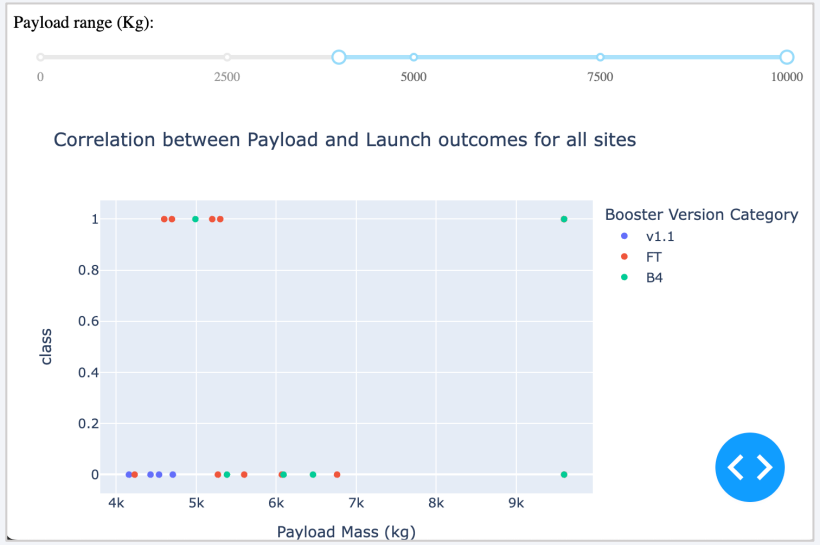


As we look closer to KSC LC-39A's launches, this launch sites has 76.9% of success rate.

The impact of Payload on Launch Outcome



Lighter Payload (<4000kg)



Heavier Payload (>4000kg)

We could divide the range of Payload into Lighter (<4000kg) and Heavier (>4000kg) to investigate the impact of Payload on the Launch Outcome.

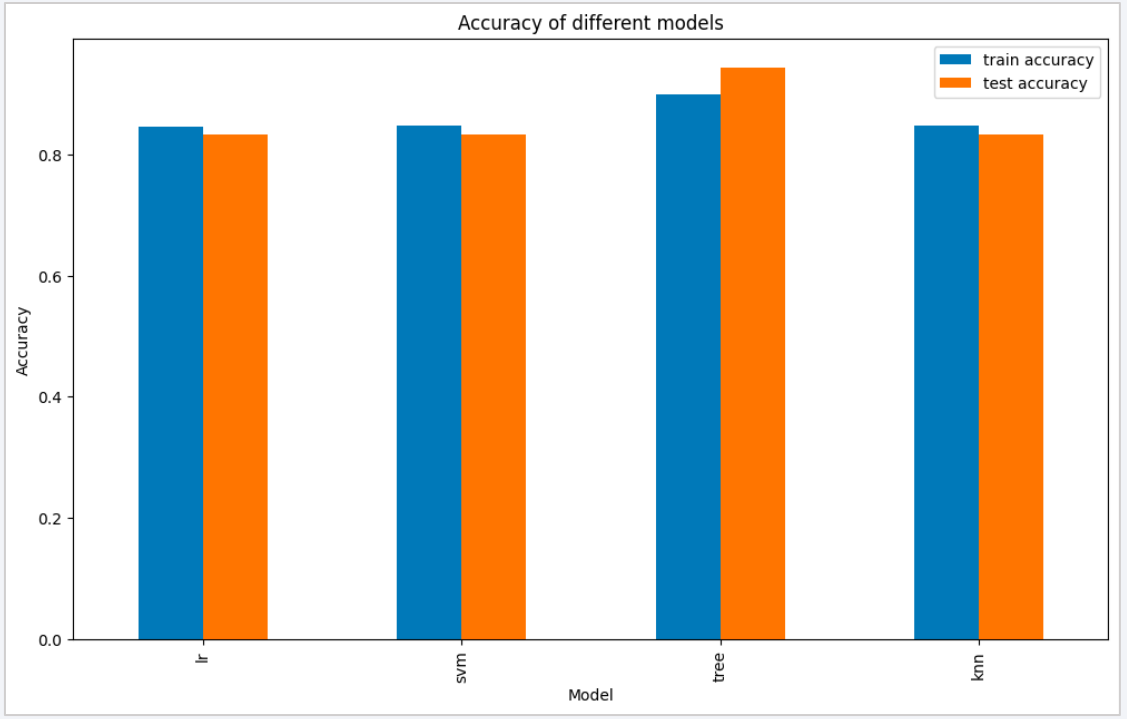
- There were fewer successful launches with Heavier Payload.
- Booster v1.0 and B5 had never been launched with Heavier Payload.
- Booster FT had a higher number of success launches.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

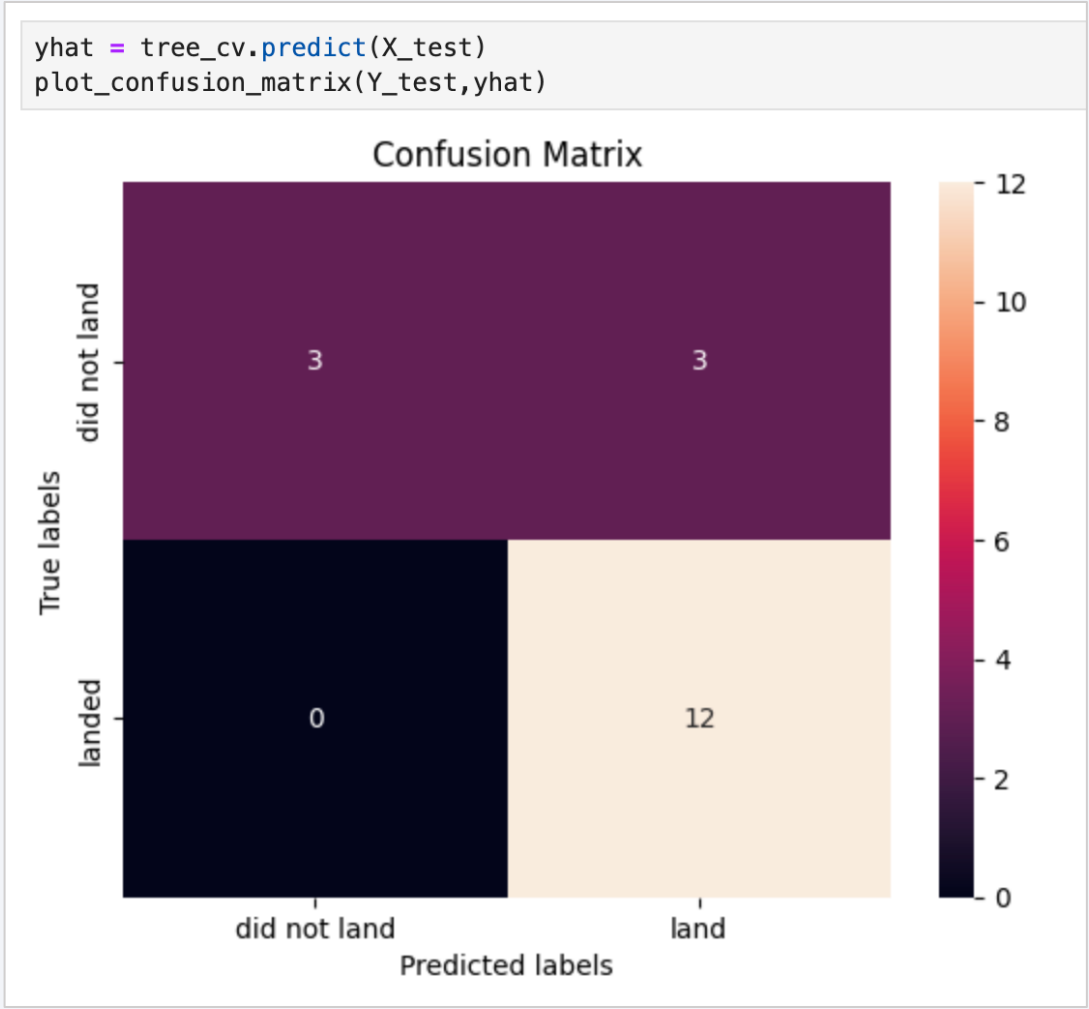
- From the chart, we can see that Tree Classification model had the highest performance among the four models.
  - Training accuracy 0.9
  - Testing accuracy 0.94
- Tuned hyperparameters:
  - {'criterion': 'gini', 'max_depth': 18, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'}



Confusion Matrix

- Confusion Matrix of the highest performing model - Tree Classification.
- As we can see, the model could somehow distinguish different classes pretty well, but it still had some false positives.

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The dataset has 90 samples, with 83 features. With 80/20 split, we have 72 samples for train data and 18 samples for test data. The low number of samples might lead to unstable model performance and possible overfitting.
- Using GridSearchCV, we trained four models on different hyperparameter to find the best combination of parameters for each model.
- Of these models, Tree Classification best predicted landing outcome of rocket.
- However, we still have some problem with false positives which probably would definitely impact our final prediction.

Appendix

- Github Project: See here

Thank you!