IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daria Olszowska
03/18/2023

Outline

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

Executive Summary

- Summary of methodologies
  1. Data collection
  2. Data wrangling
  3. Data visualization
  4. Data prediction
- Summary of all results

Introduction

The aim of this project is to develop a predictive model that can determine the likelihood of a successful landing of the Falcon 9 first stage, which is a crucial component in SpaceX's launch process. The ability to reuse the first stage of the Falcon 9 rocket is a key factor in reducing the cost of launches and enabling SpaceX to offer more competitive pricing than other providers.

We want to identify the critical parameters that determine the successful recovery of Falcon 9 first stage and find relationship between them.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    1. SpaceX API
    2. Webscraping (Wikipedia)
- Perform data wrangling
    1. Extracting Falcon 9 data
    2. Replacing missing values
    3. Translating Failure/Success data into numerical values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    1. Create various classification objects
    2. Testing accuracy of different parameters on validation data
    3. Calculating confusion matrices, calculating accuracy on test data

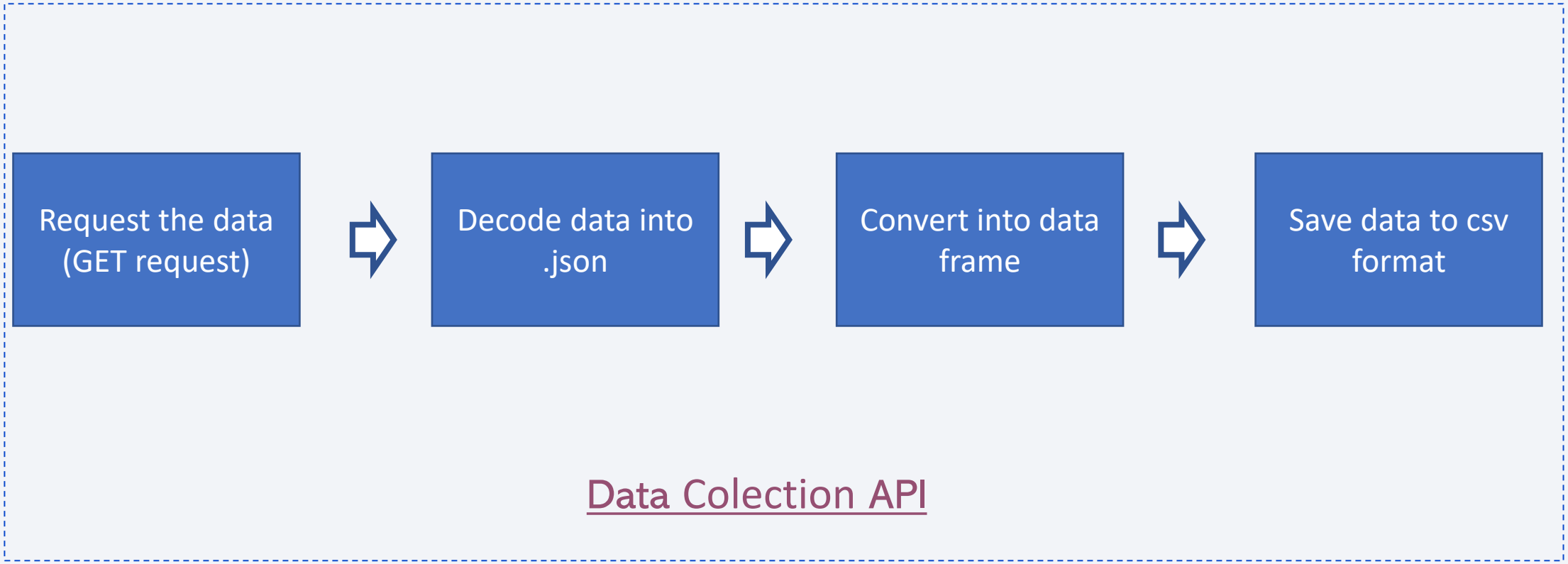
Data Collection

Data used in the following analysis was collected from two sources:

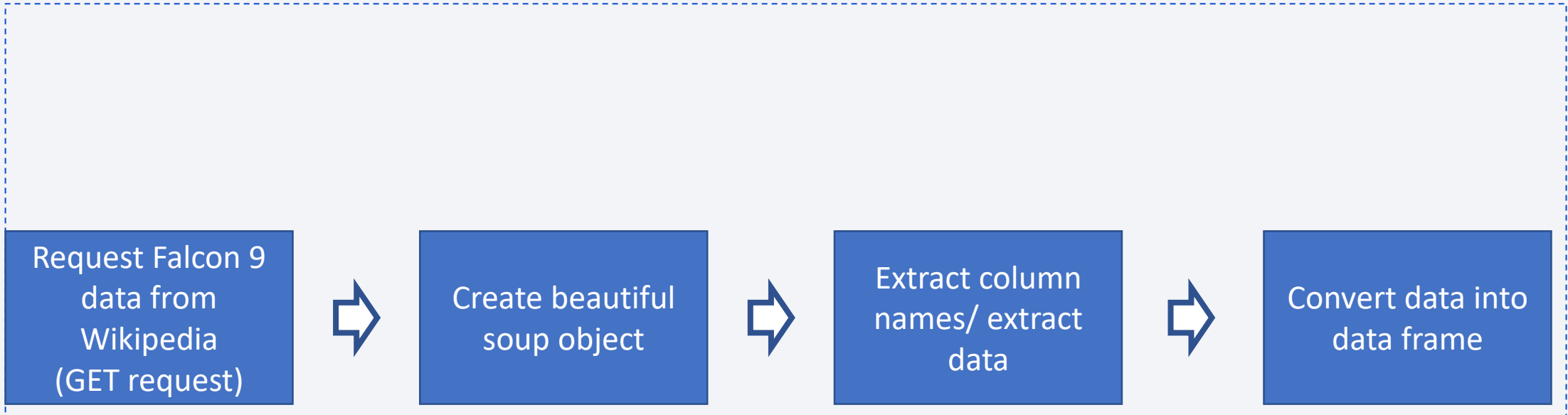
1. SpaceX API
2. Wikipedia

Data is pulled from both sources using GET request, processed, converted into data frames and saved into .csv file

Data Collection – SpaceX API

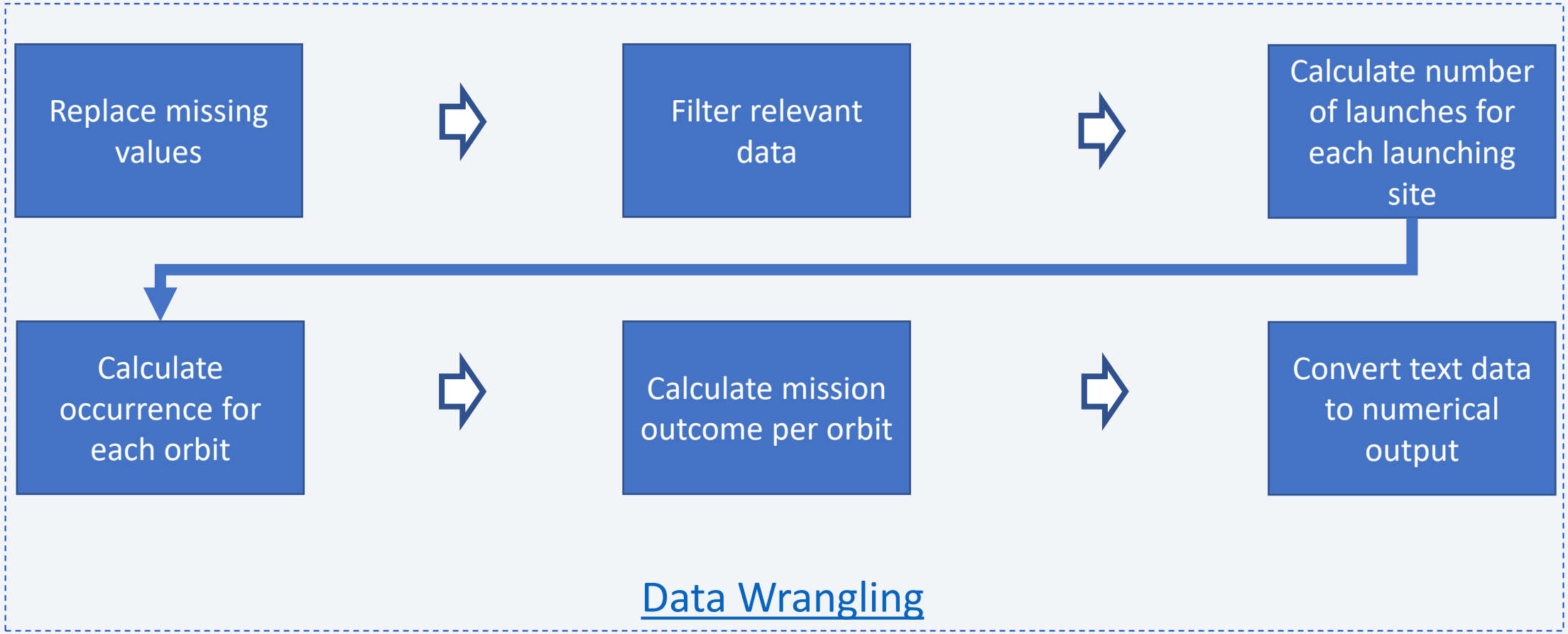


Data Collection - Scraping



Data Collection with Web Scraping

Data Wrangling



EDA with Data Visualization

1. Scatter plots:
   - Flight number vs. payload mass
   - Flight number vs. launch site
   - Payload mass vs. launch site
   - Flight number vs. orbit
   - Payload mass vs. orbit
2. Bar chart:
   - Orbit type vs. success rate
3. Line plot:
   - Date vs. success rate

EDA with Visualization

EDA with SQL

- List of performed SQL queries:

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 average payload mass carried by booster version F9 v1.1,
5. List the date when the first successful landing outcome in ground pad was achieved,
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000,
7. List the total number of successful and failure mission outcomes,
8. List the names of the booster_versions which have carried the maximum payload mass,
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.
10. Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order..

Build an Interactive Map with Folium

- Map objects utilized in this study:
  1. Highlighted circles to visualize launch site location,
  2. Markers to show the relation between successful and failed stage one recovery for each site,
  3. Lines to show the distance to closest coastline, city, highway, railway, etc.,

Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

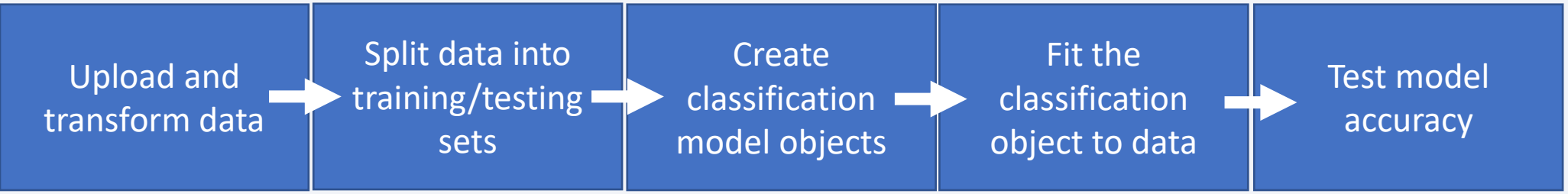
- Using Plotly Dash we created interactive dashboard showing allowing data visualizations for all or specific launching sites using:
  1. Pie chart
     - Success rate per site for all sites
     - Success/failure ratio per specific site
  2. Scatter plot with slider allowing payload values adjustment
     - Payload vs. success rate for all sites
     - Payload vs. success rate for specific site

SpaceX Dash App

Predictive Analysis (Classification)

- To predict the success/failure ratio we used following classification models:
    1. Logistic regression,
    2. Support vector machine,
    3. Decision tree,
    4. K nearest neighbors.

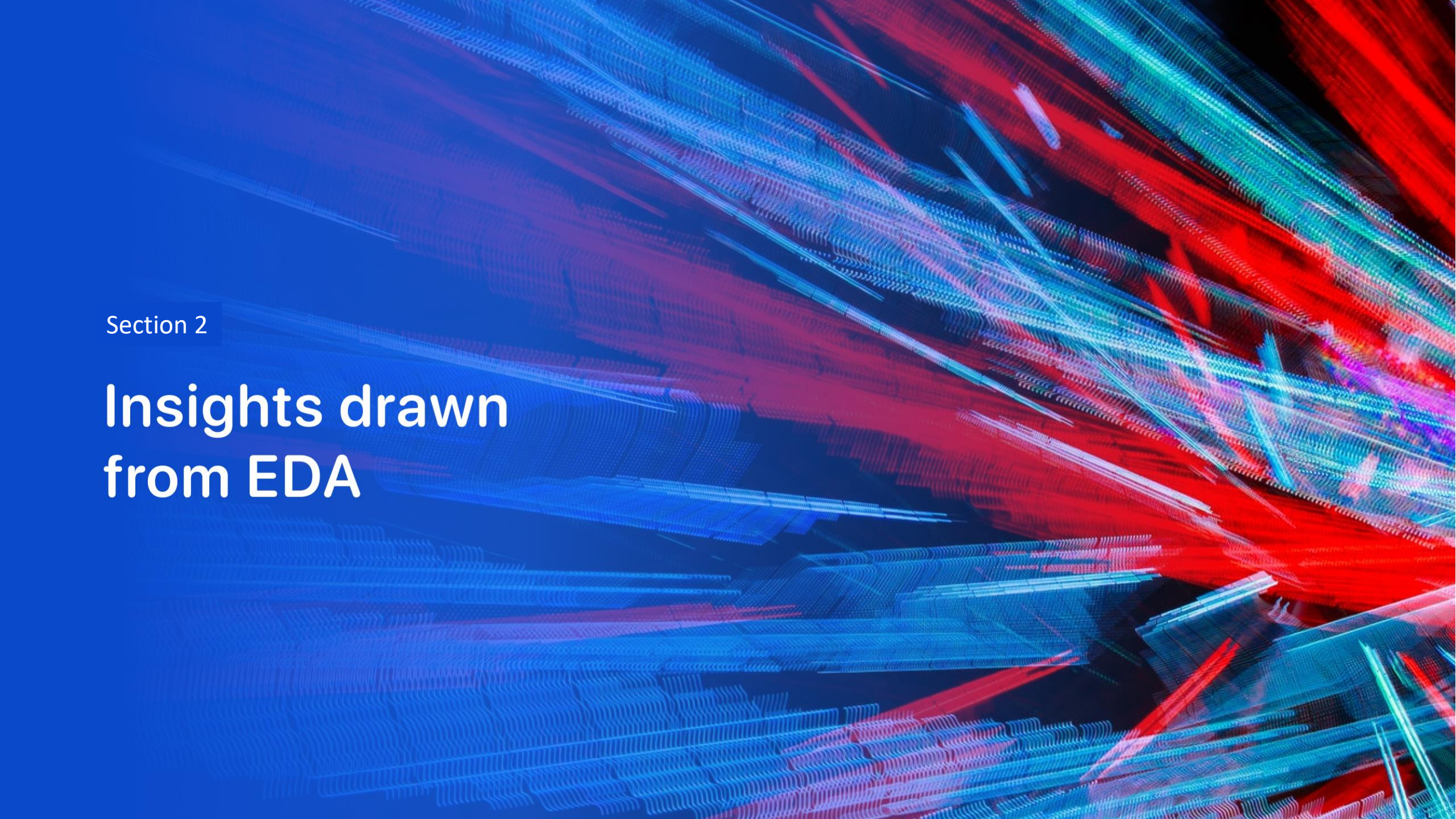
To find the algorithm that performed the best we calculated confusion matrices and accuracy scores for each method used



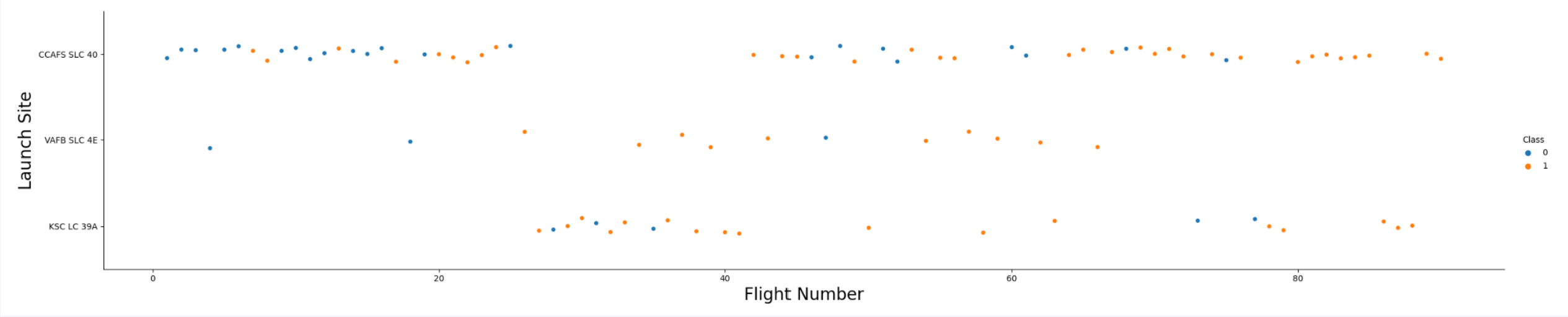
Machine Learning Prediction

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

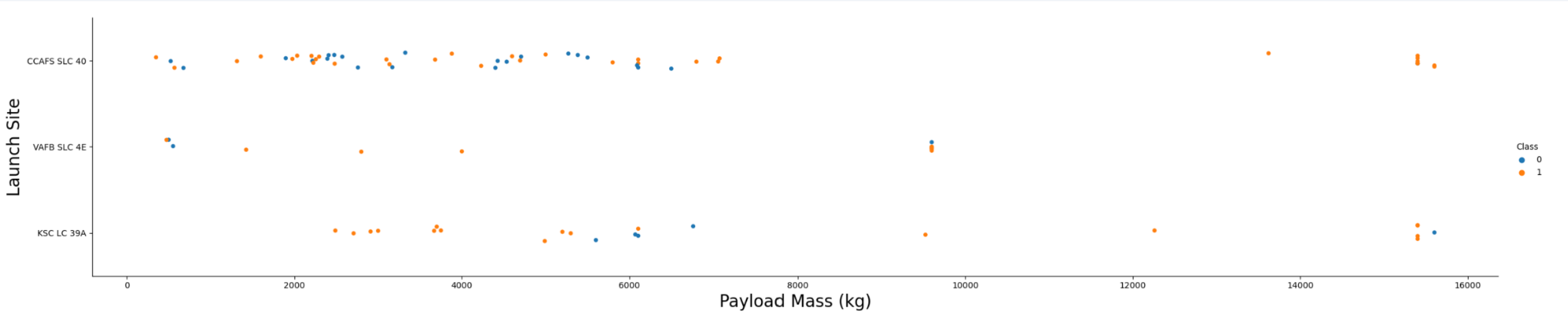

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



The success rate for each launch site increases with a flight number

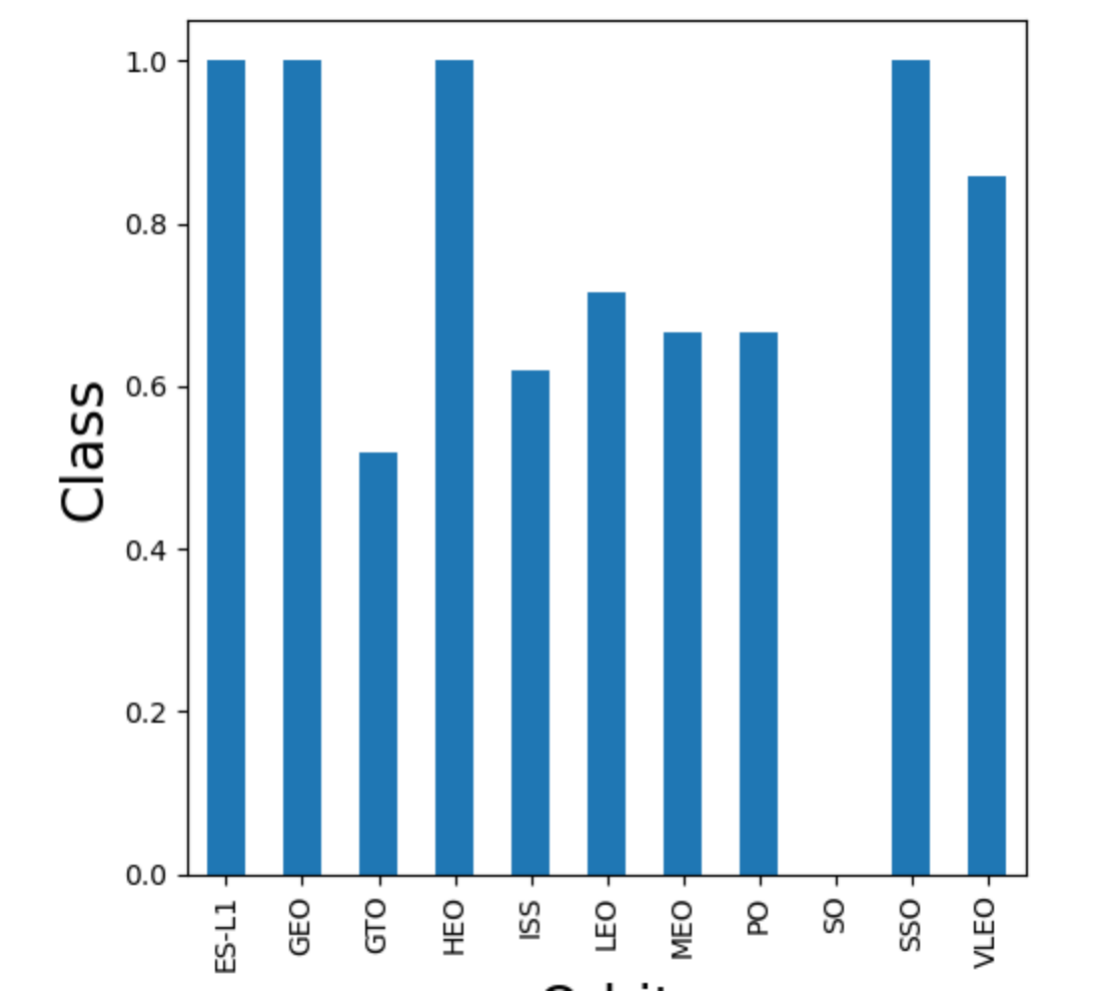
Payload vs. Launch Site



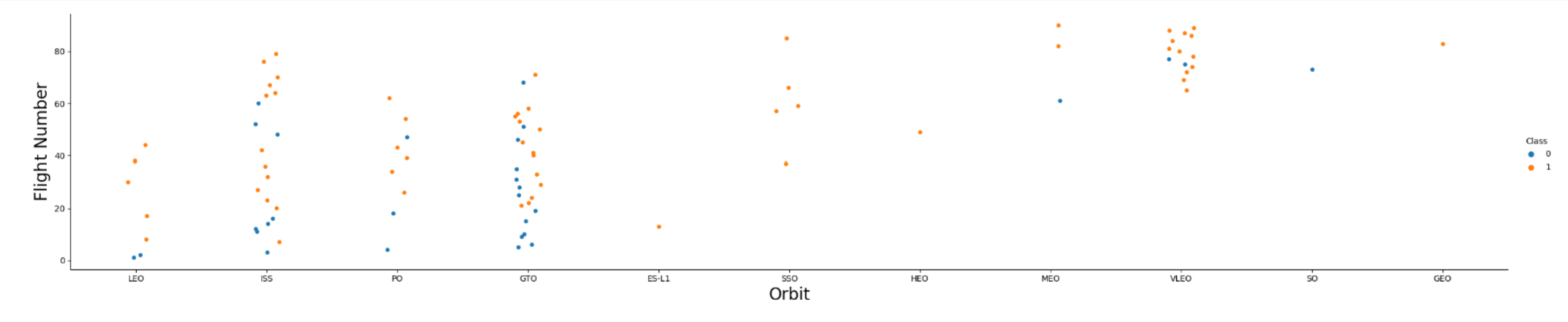
The higher the payload mass the greater the chance for success, especially above 9000 kg

Success Rate vs. Orbit Type

- ES_L1, GEO, HEO and SSO orbits have the highest success rate (1), first stage was recovered for all of the missions
- GTO orbit has the lowest success rate (~0.5), for every second mission first stage was not receovered

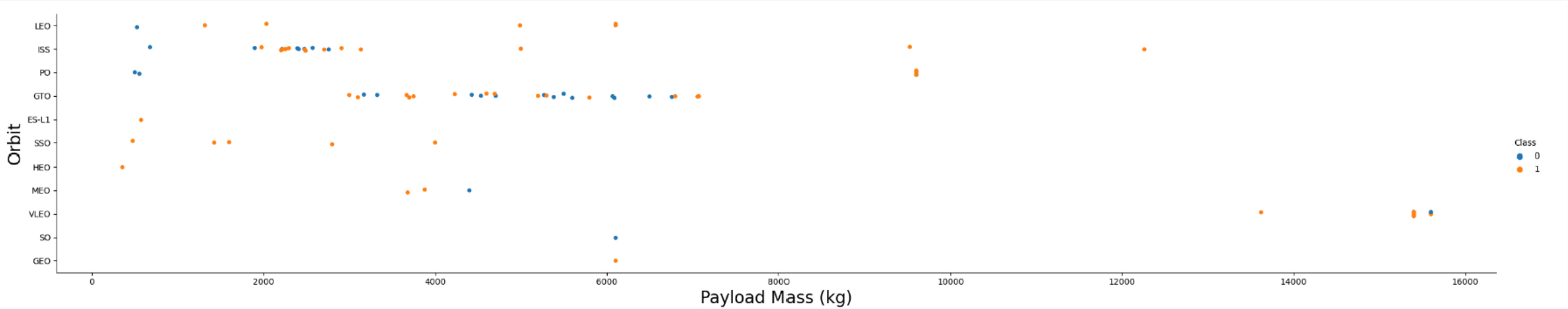


Flight Number vs. Orbit Type



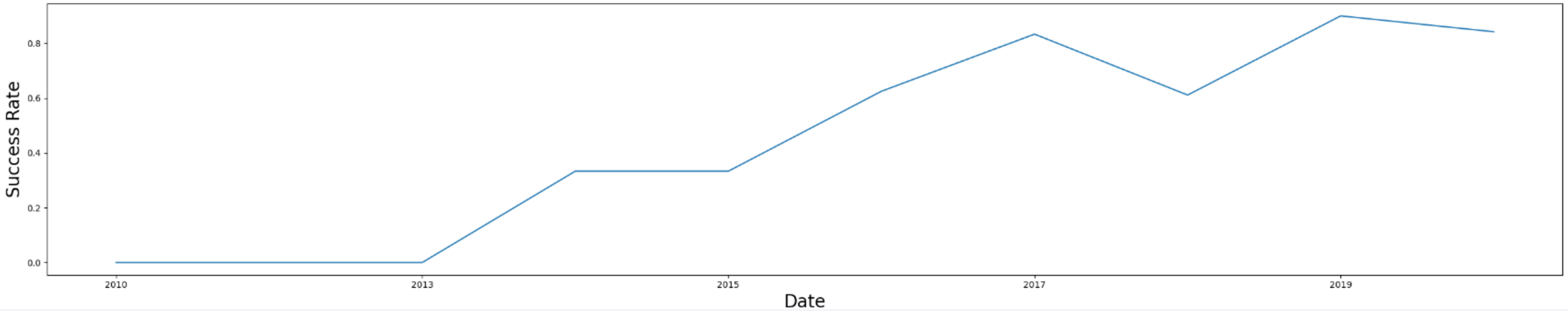
The highest the flight number the greater the chance for successful stage one recovery. Target orbit changed as the Falcon 9 program was developed.

Payload vs. Orbit Type



The greater the payload the greater the chances for success, payload varies for different orbit types.

Launch Success Yearly Trend



The chance for successful recovery increases with time. Between 2010 and 2013 first stage was lost for all of the missions, currently there is 80% chance of success.

All Launch Site Names

- Using the distinct query, we extract only the unique values for the database.
- There are 4 different launch sites

```
%sql SELECT  DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

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

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

Launch Site Names Begin with 'CCA'

- Using command like we only search for values that start with CCA.
- Using limit command, we show only the requested number of results

```
%sql SELECT LAUNCH_SITE FROM SPACEXTBL WHERE LAUNCH_SITE LIKE "CCA%" LIMIT 5
```

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

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXTBL WHERE CUSTOMER = "NASA (CRS)"
```

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

SUM(payload_mass__kg_)
45596

- Sum command calculates the sum of the values in a given column
- Applying where clause we search only for specific type of customer
- The total payload mass for NASA (CRS) customer is **45596 kg**

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXTBL WHERE (booster_version) = 'F9 v1.1'
```

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

AVG(payload_mass__kg_)
2928.4

- AVG command calculates average value for given column
- Where clause select only data fulfilling specific conditions
- Average payload mass for F9 v1.1 booster is **2928.4 kg**

First Successful Ground Landing Date

```
%sql SELECT MIN(Date) from SPACEXTBL where "Landing _Outcome" = 'Success (ground pad)'
```

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

MIN(Date)
01-05-2017

- MIN command returns minimum value for the given column
- Where clause limits the results to those fulfilling specified condition
- The first successful ground pad mission took place on **01/05/2017**

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version from SPACEXTBL where "Landing _Outcome" = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4001 AND 5999
```

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

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

- Using where command we only queried successful drone ship mission
- Between command limited the output to lines where the payload was between specified values
- There were 4 successful landings fulfilling those conditions

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) from SPACEXTBL 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

- COUNT command counts the number of entries
- Using GROUP BY we group the results into different categories
- Number of **successful missions: 100**, **failures: 1**

Boosters Carried Maximum Payload

```
%sql SELECT Booster_Version from SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * 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

- Where clause limits outputs to specified parameters only
- Using subquery, we search for the maximum payload value
- There are **12** booster versions that carried the maximum payload

2015 Launch Records

```
%sql SELECT substr(Date, 4, 2), "Landing _Outcome", Booster_Version, LAUNCH_SITE from SPACEXTBL WHERE substr(Date,7,4)='2015' AND "Landing _Outcome" =
```

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

substr(Date, 4, 2)	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Using SUBSTR command we extract the month/date value from the date
- Using WHERE clause we limit the number of outputs to those fulfilling the criteria
- In 2015 there were 2 drone ship failures, one in **January**, second in **April**

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

```
%sql SELECT "Landing _Outcome", COUNT(*) from SPACEXTBL WHERE "Landing _Outcome" LIKE 'SUCCESS%' \
and Date between '04-06-2010' and '20-03-2017' group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

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

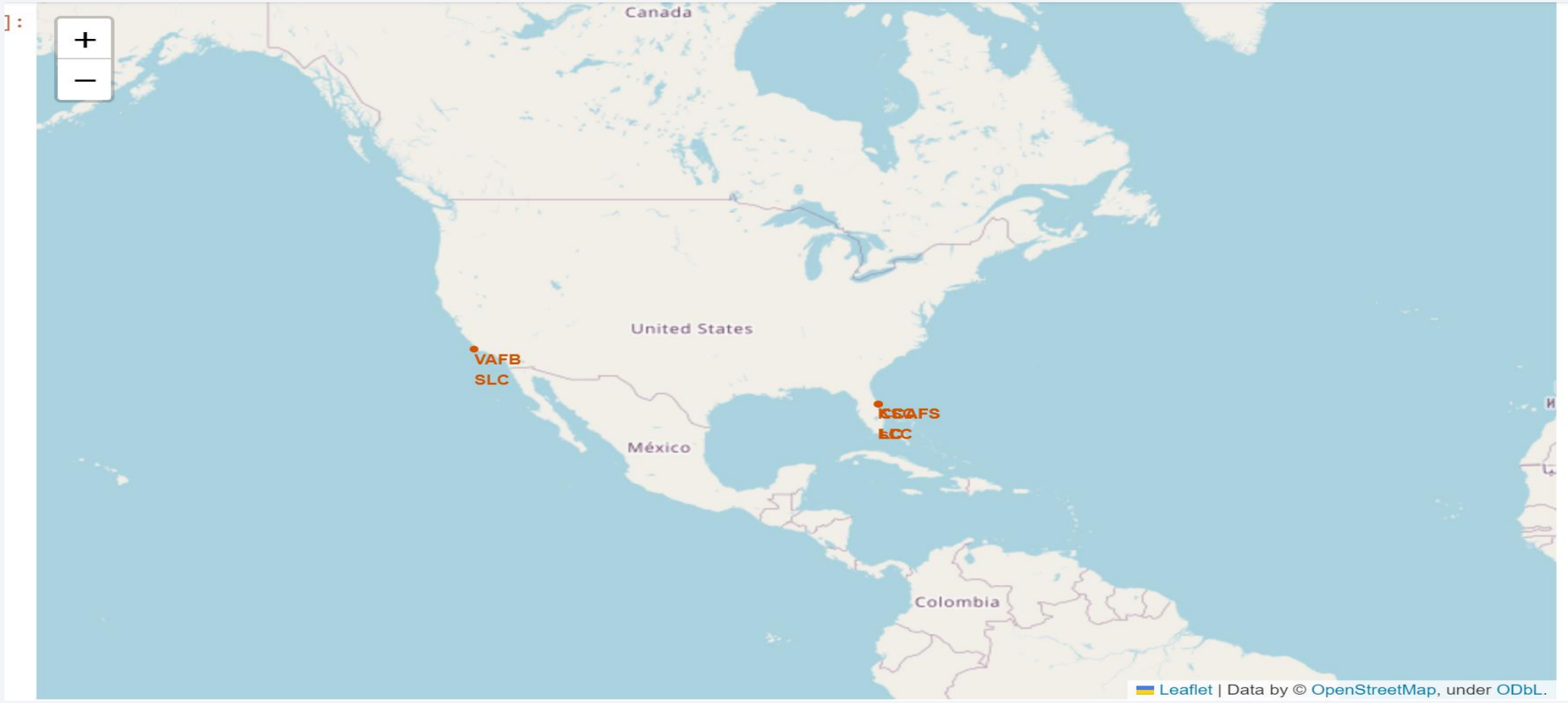
Landing _Outcome	COUNT(*)
Success	20
Success (drone ship)	8
Success (ground pad)	6

- Using WHERE clause we limit data to that fulfilling the criteria
- LIKE command searches for entries starting with Success
- BETWEEN command imposed on Date yields data only for the given time range
- We group the data using the GROUP BY command
- Using ORDED BY we order the data in descending manner

Section 3

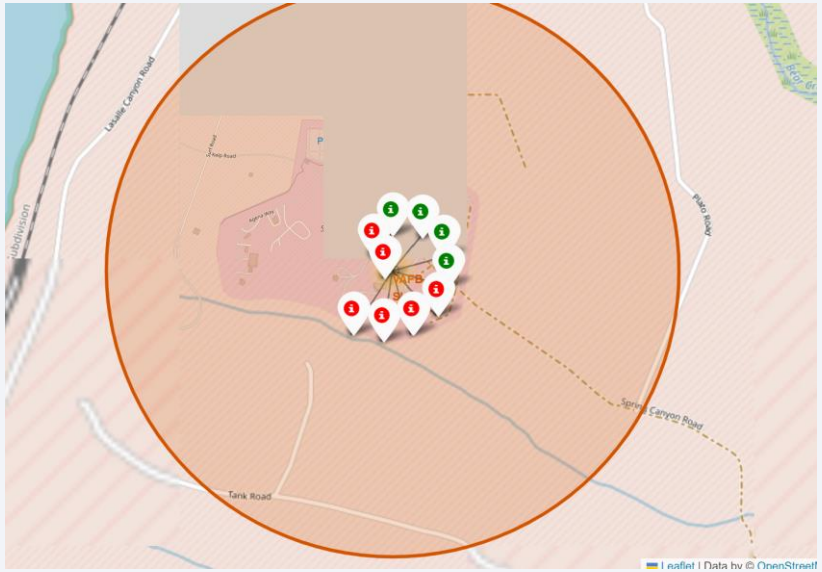
Launch Sites Proximities Analysis

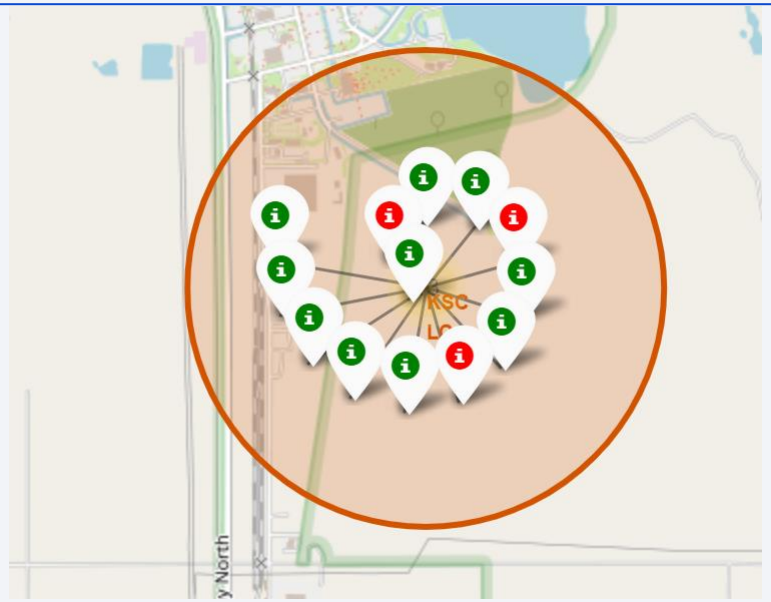
Launch Site Locations

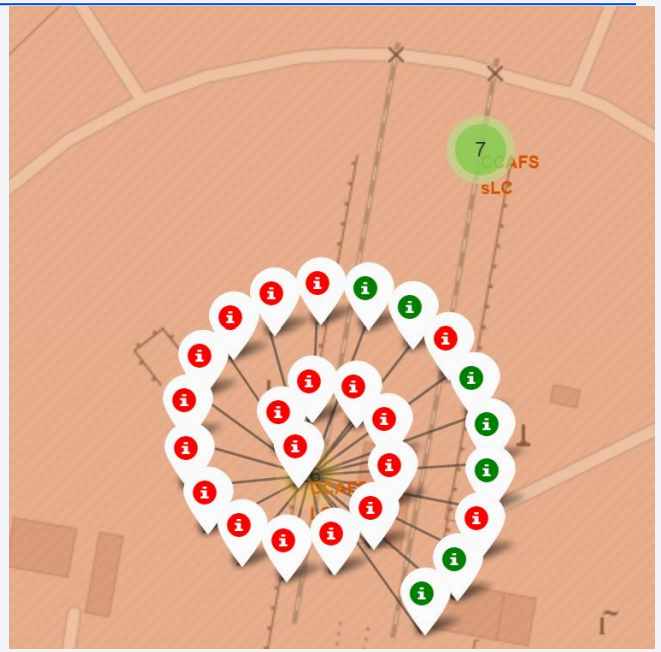


All SpaceX launch Sites are located near the USA coast, in Florida or California

Launch Sites Success/Failure Maps



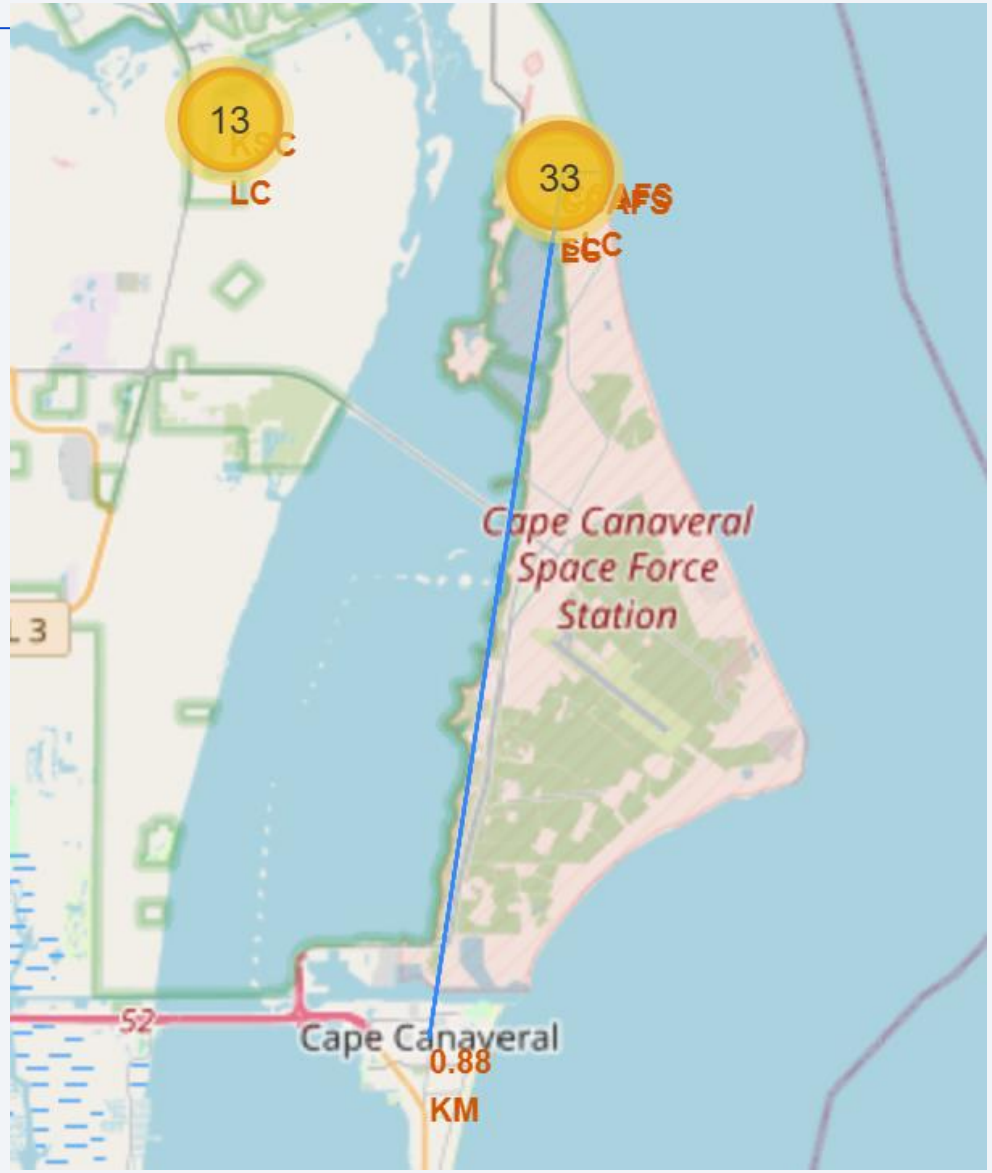


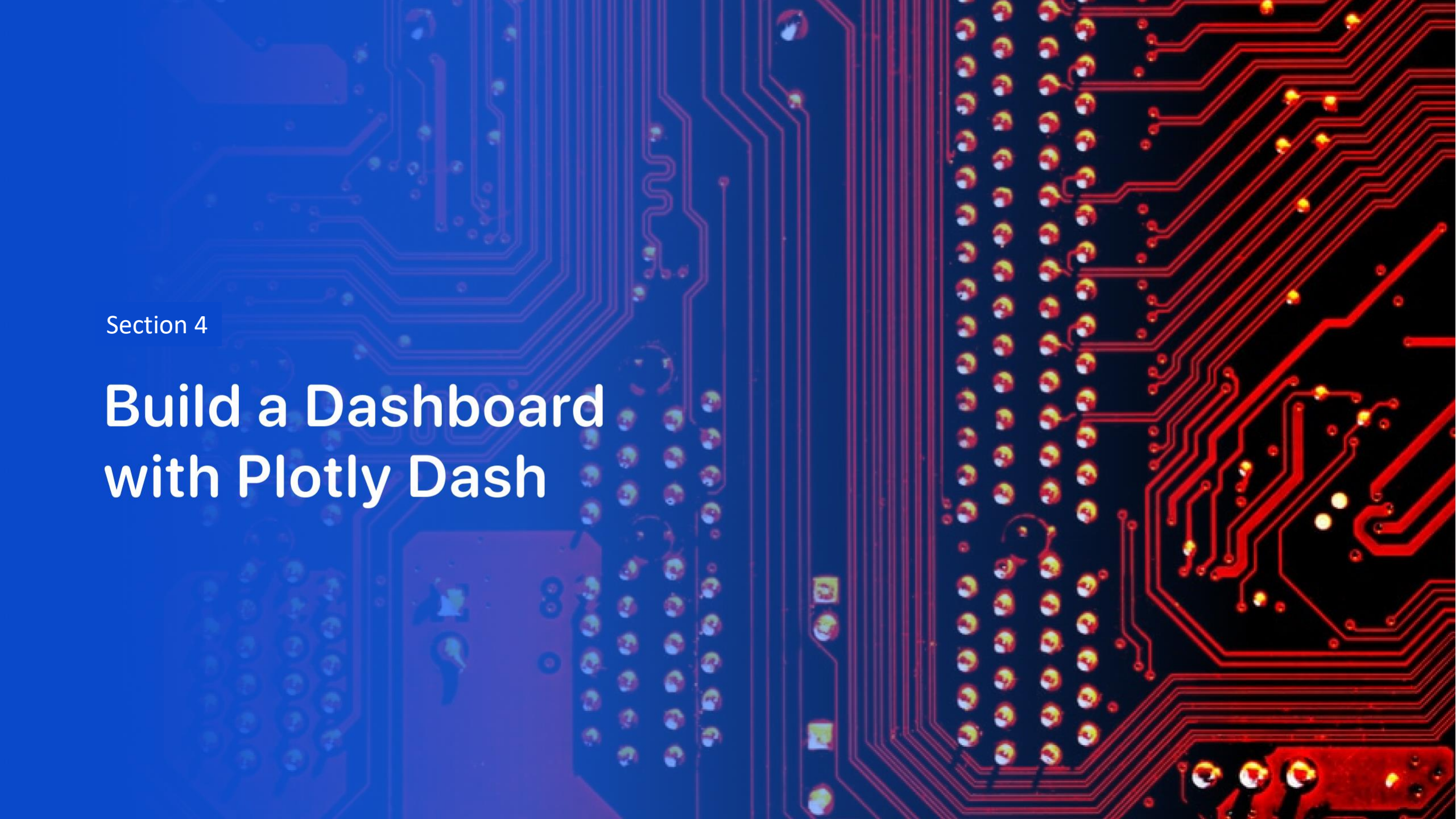


Above maps show Launch Site locations with markers representing success (green) or failure (red) mission outcome

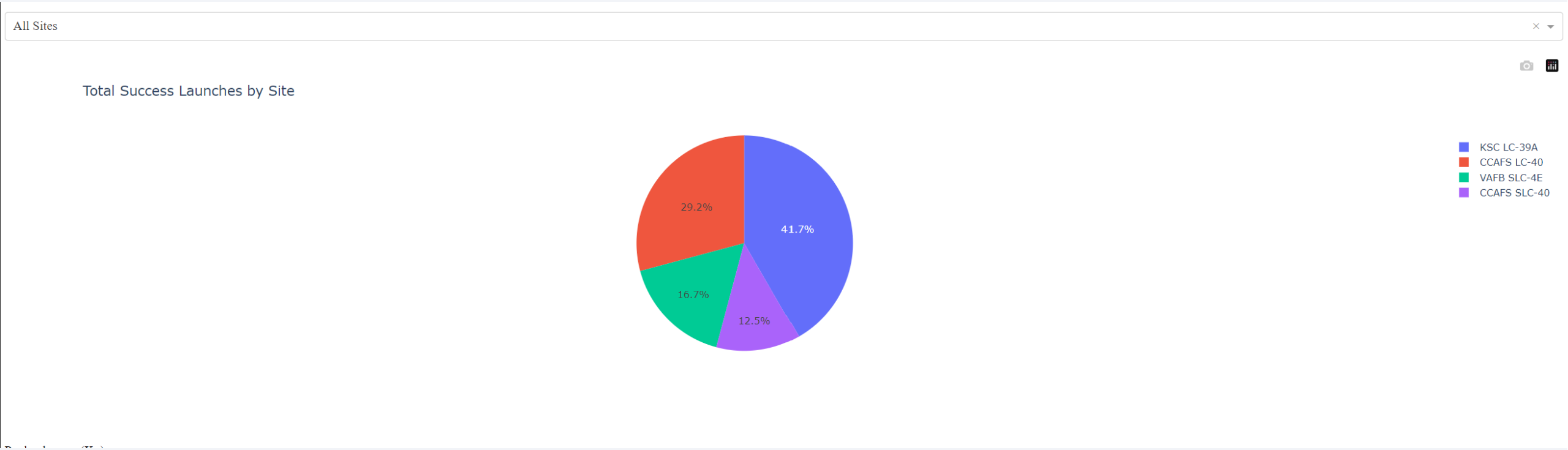
Launch Sites Distance to Nearest Objects

- Using the line objects and markers we find the distance between Launch Site locations and near object.
- Map on the right shows the distance between one of the launch sites and Cape Canaveral, the distance is less than 1 km



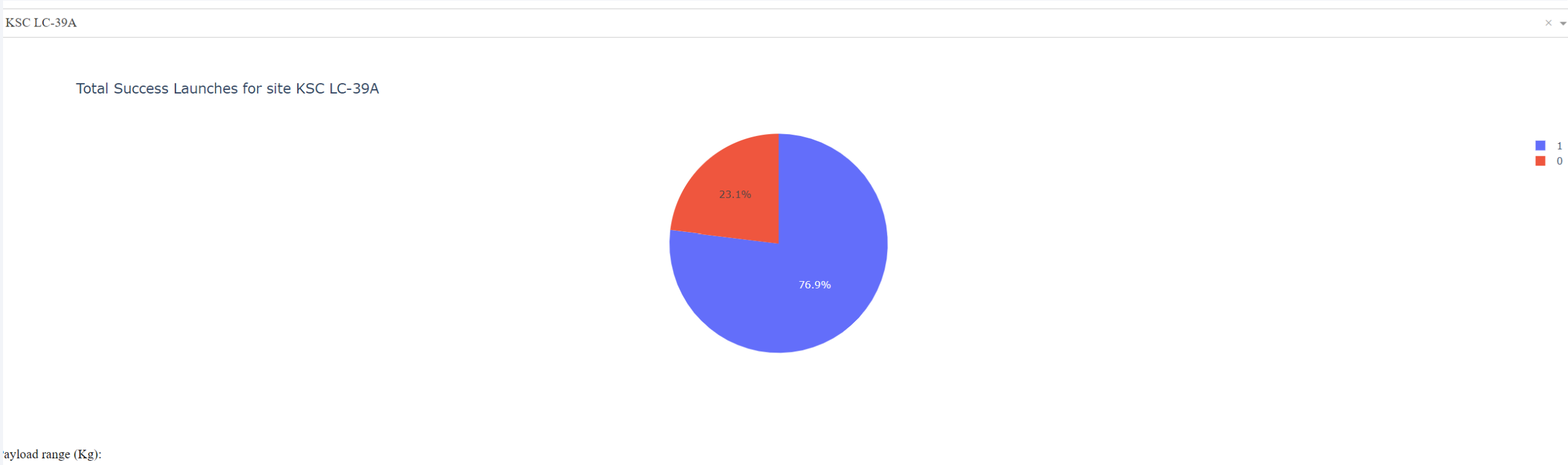

Section 4
Build a Dashboard with Plotly Dash

Success Distribution for each Sites



- KSC LC-39A has the highest number of successful missions
- CCAFS SLC-40 has the lowest numbers of successful recoveries

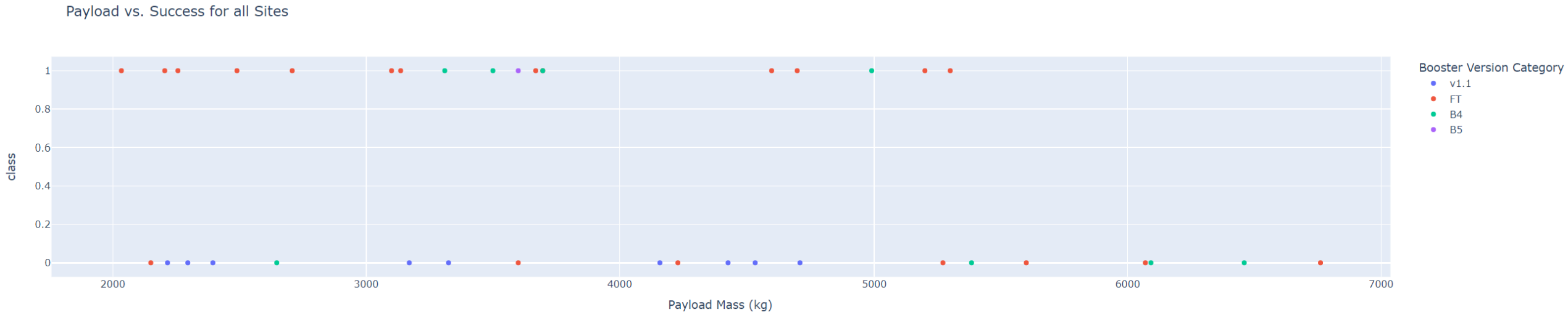
Success/Failure for KSC LC-39A Launch Site

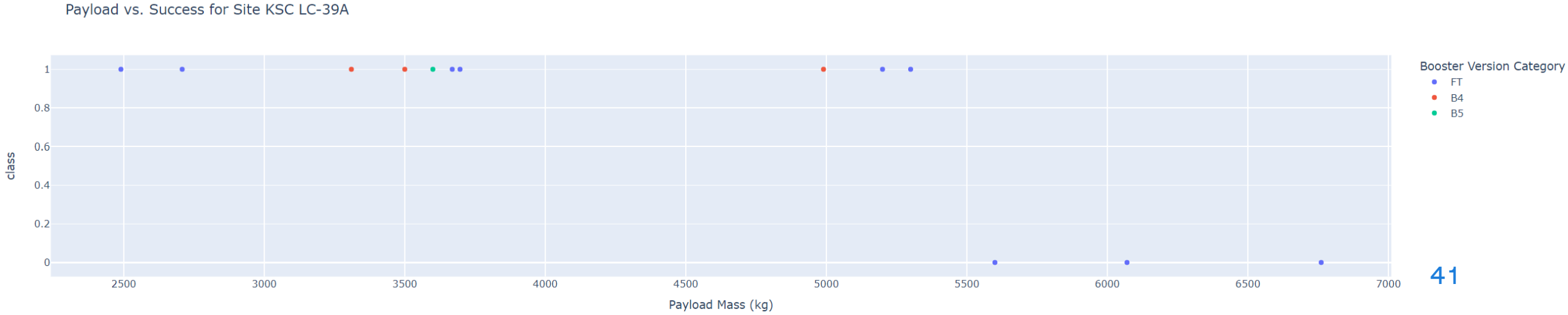


- KSC LC-39A has 76.9% successful recoveries

Payload vs. Succes

- Success probability for lower payload mass is higher



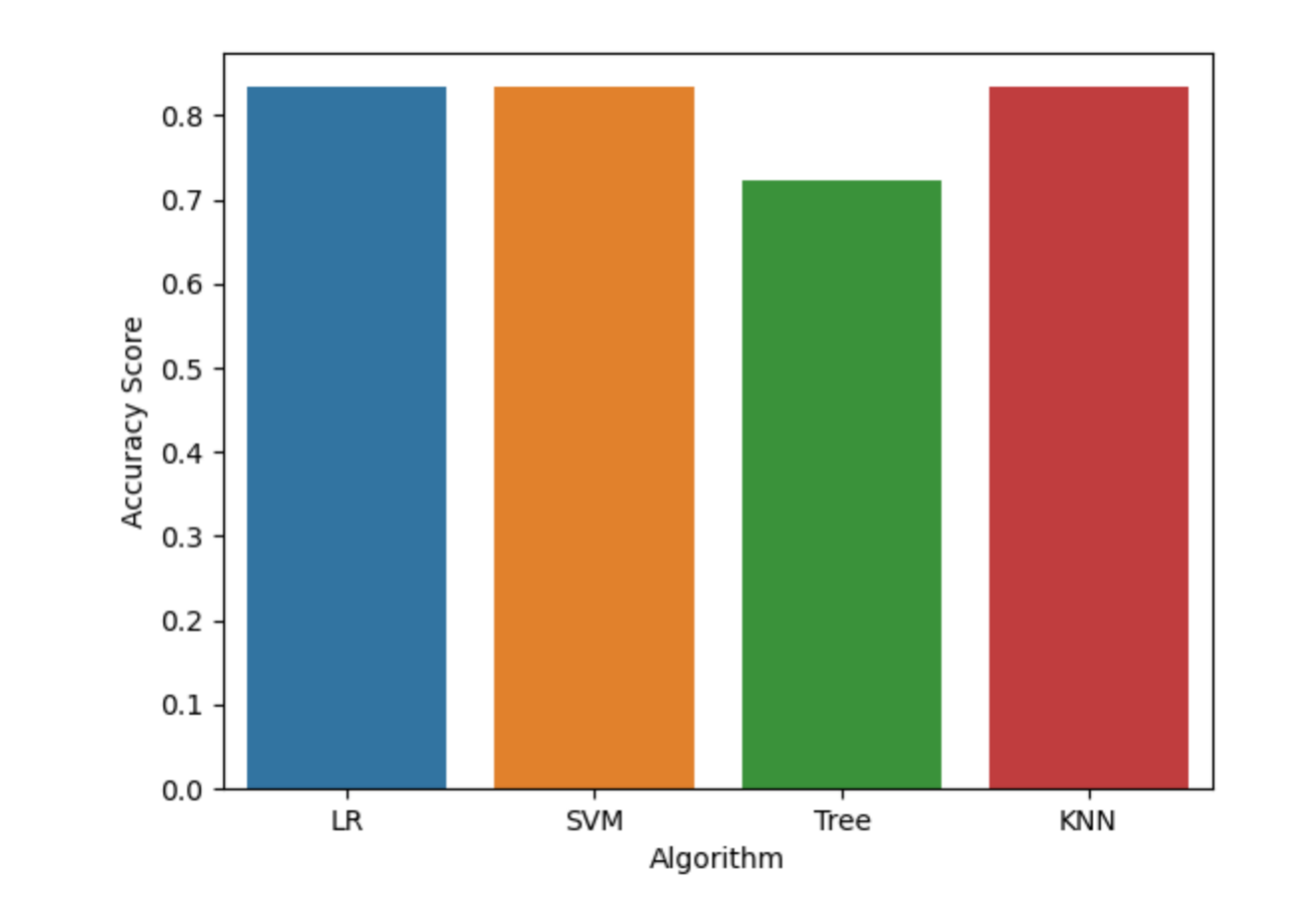


Section 5

Predictive Analysis (Classification)

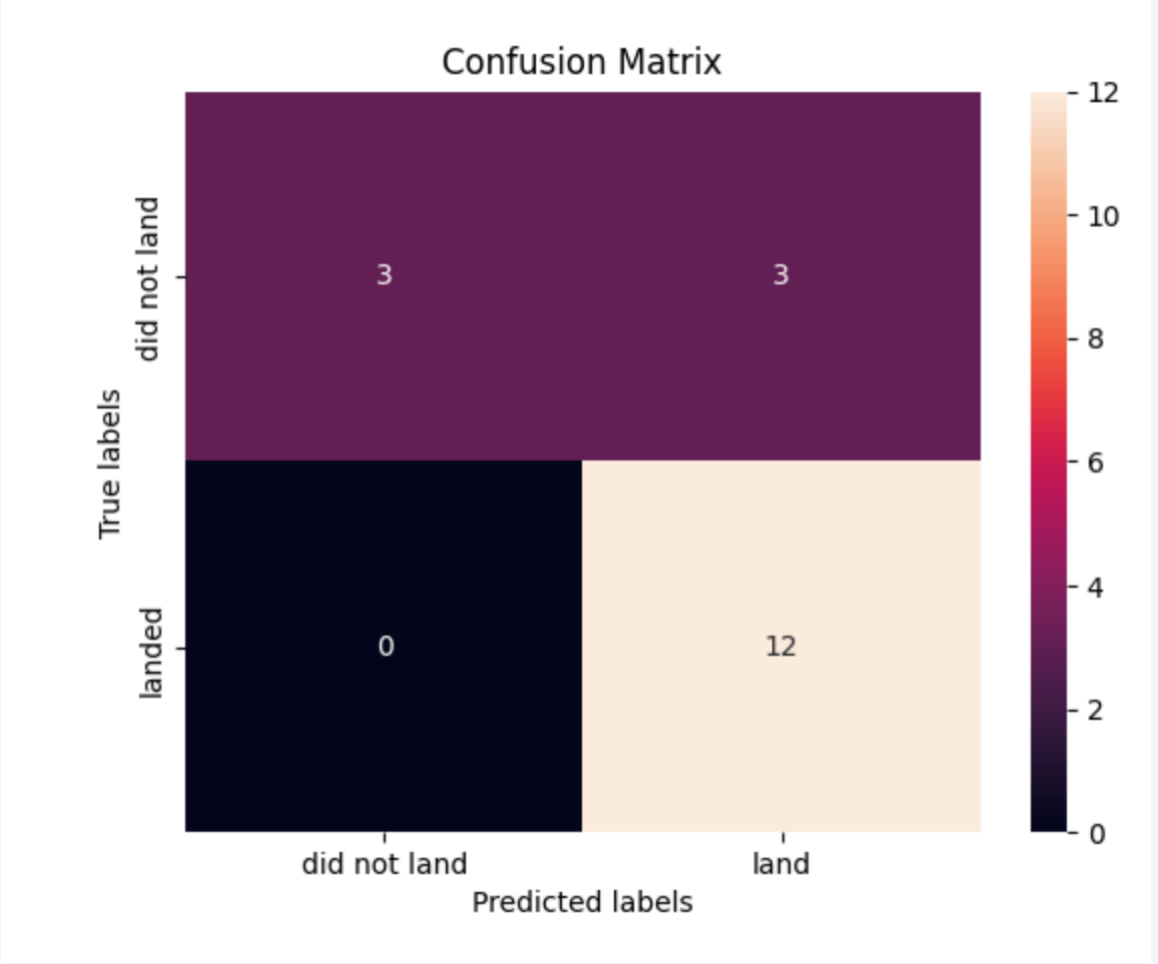
Classification Accuracy

- All models have very similar accuracy.
- LR, SVM and KNN all have accuracy of 0.83.
- Decision Tree accuracy is lower than the rest (0.72)



Confusion Matrix

- Plot shows confusion matrix for KNN.
- The algorithm correctly predicted outcomes for successful missions
- For failed missions the correct prediction ratio is 50%



Conclusions

- Early versions of Falcon 9 had higher failure ratio; with time the system became more reliable
- With program development and increased payload mass Falcon 9 was able to diversify orbit types
- For ES_L1, GEO, HEO and SSO orbits the success rate is equal to 1, stage one was recovered for all of the launches
- All classification models, except decision trees, can predict landing outcome with very high accuracy (83%)

Thank you!