IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marcel Kurniawan
24-5-2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through Rest API and web scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Visual Analytics with Folium
  - Interactive Dashboard with Plotly
  - Predictive Models with Machine Learning
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive visual analytics in screenshots
  - Predictive analysis using classification results

Introduction

Background

SpaceX is a company that aims to make space travel accessible to everyone. It hasachieved many feats, such as sending spacecraft to the space station, launching asatellite network that offers internet service and sending humans to space. SpaceXcan do this because it can reuse the first stage of its Falcon 9 rocket, which makesits rocket launches much cheaper ($62 million per launch) than other companies,which have to discard the first stage after each launch ($165 million or more perlaunch). By predicting if the first stage will land successfully, we can estimate thecost of the launch. We can use machine learning models and public data to do thisprediction for SpaceX and its competitors.

Research Question

- What factors are affect the first-stage landing success
- Does the rate of successful landings increase over the years?
- What classification model that have best performance in this case?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX Rest API
  - Using Web Scrapping from Wikipedia
- Perform data wrangling
  - Filtering data, Handle missing value and apply one-hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune and evaluate the classification models to find best model and parameters

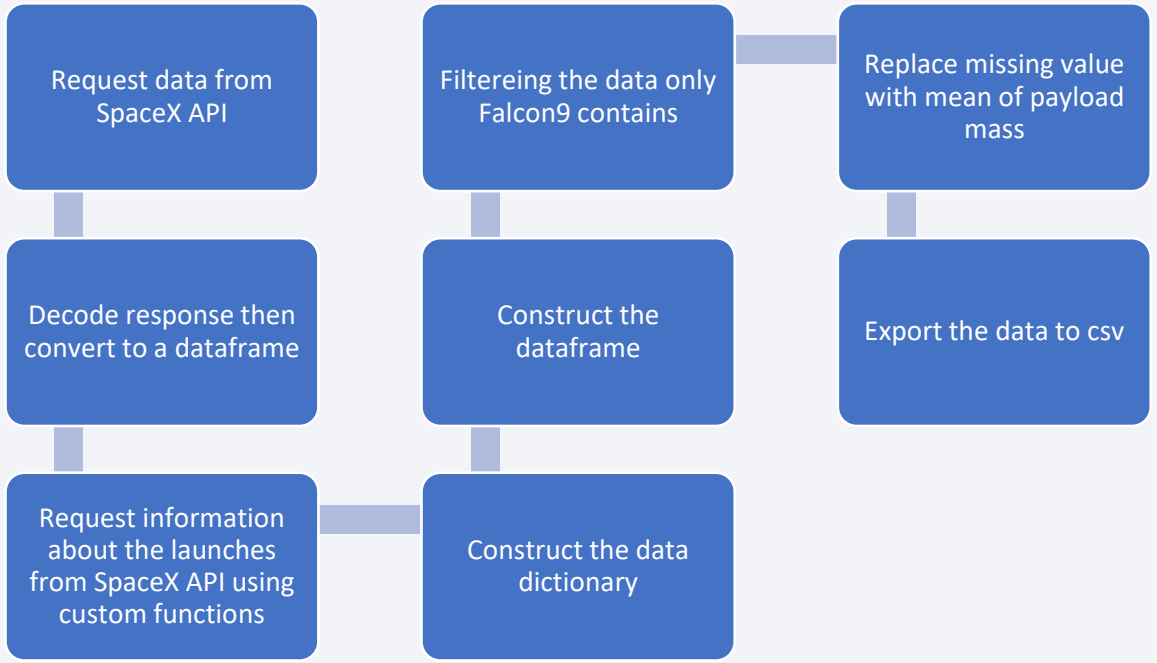
Data Collection

- Data was collected using various techniques
  - Data sets collected by SpaceX API and webscrape from Wikipedia
  - We use beautiful soup to scrape the data from Wikipedia
  - We normalize the json file from SpaceX API

Data Collection – SpaceX API

- Get request from SpaceX API to collect data, normalize, clean, wrangling, filtering and formatting the data.

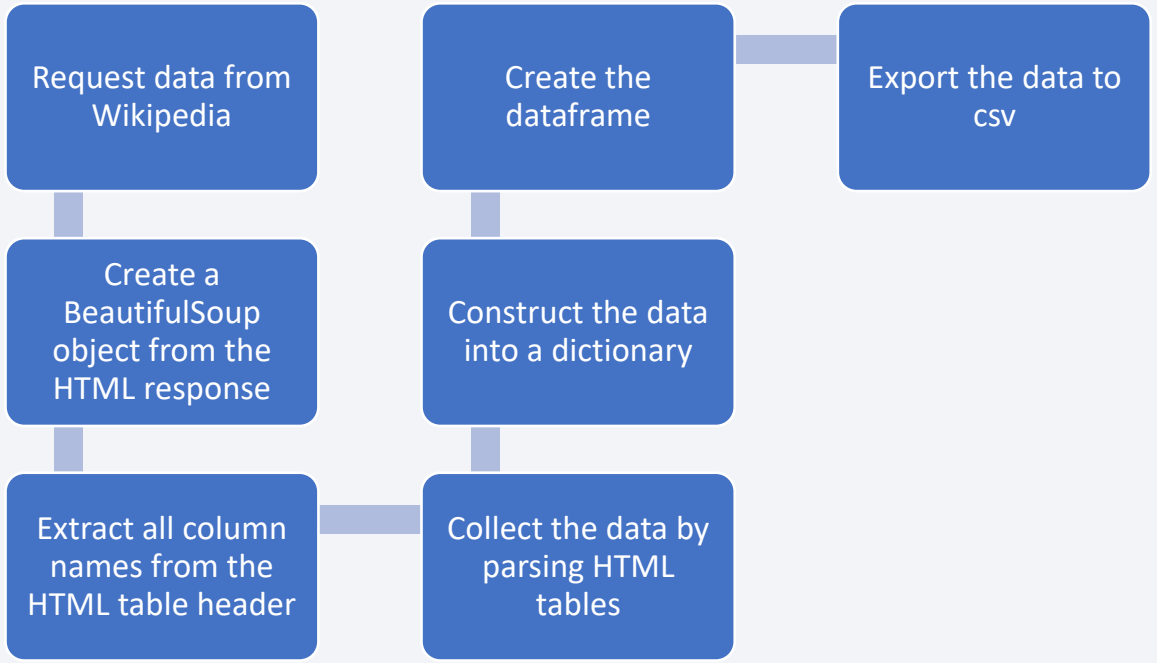
Github link: https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Data_Collection.ipynb



Data Collection - Scraping

- Web scrapping from Wikipedia.com using BeautifulSoup

Github Link:
https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Web_Scraping.ipynb



Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad. True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship. In this case we will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful. GitHub Link: https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

- Plotted charts: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend
- Scatter plots show the relationship between two variables. If they have linear relationship, they could be used in machine learning model.
- Bar charts show comparisons between categories.
- Line charts show trends in data over time (time series).

Github link: https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_EDA_Data_Visualization.ipynb

EDA with SQL

Display:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass
- List the records which display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

GitHub link: https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_EDA_SQL.ipynb

Interactive Map with Folium

- Launch Sites Marked, Distance Marked and Colored Marked on Map with Folium
  - Added a blue circle at the location of NASA Johnson Space Center with its name shown in a popup label using its latitude and longitude coordinates
  - Added red circles at the locations of all launch sites with their names shown in popup labels using their latitude and longitude coordinates Map with Folium Markers of Launch Results
  - Added markers of different colors to indicate successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates Distances from a Launch Site to Nearby Features
  - Added colored lines to show the distance from launch site CCAFS SLC40 to the closest coastline, railway, highway, and city

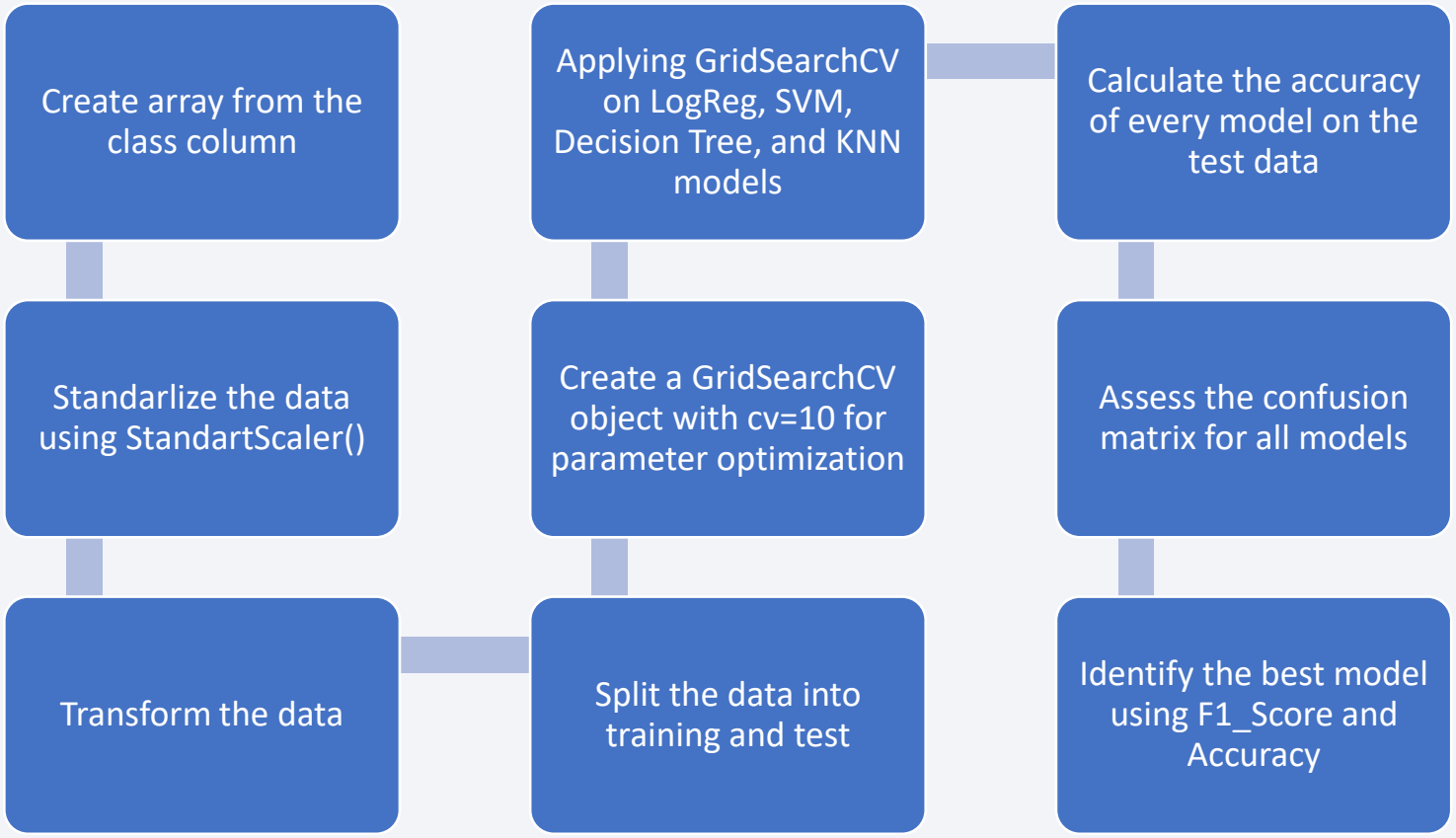
GitHub Link : https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Interative_Visualization.ipynb

Dashboard with Plotly Dash

- Dropdown List to select launch site
- Pie Chart showing Success Launches
- Slider of Payload Mass Range
- Pie Chart Showing Successful Launches
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version to see the correlation and relationship.

GitHub Link : https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Dashboard.py

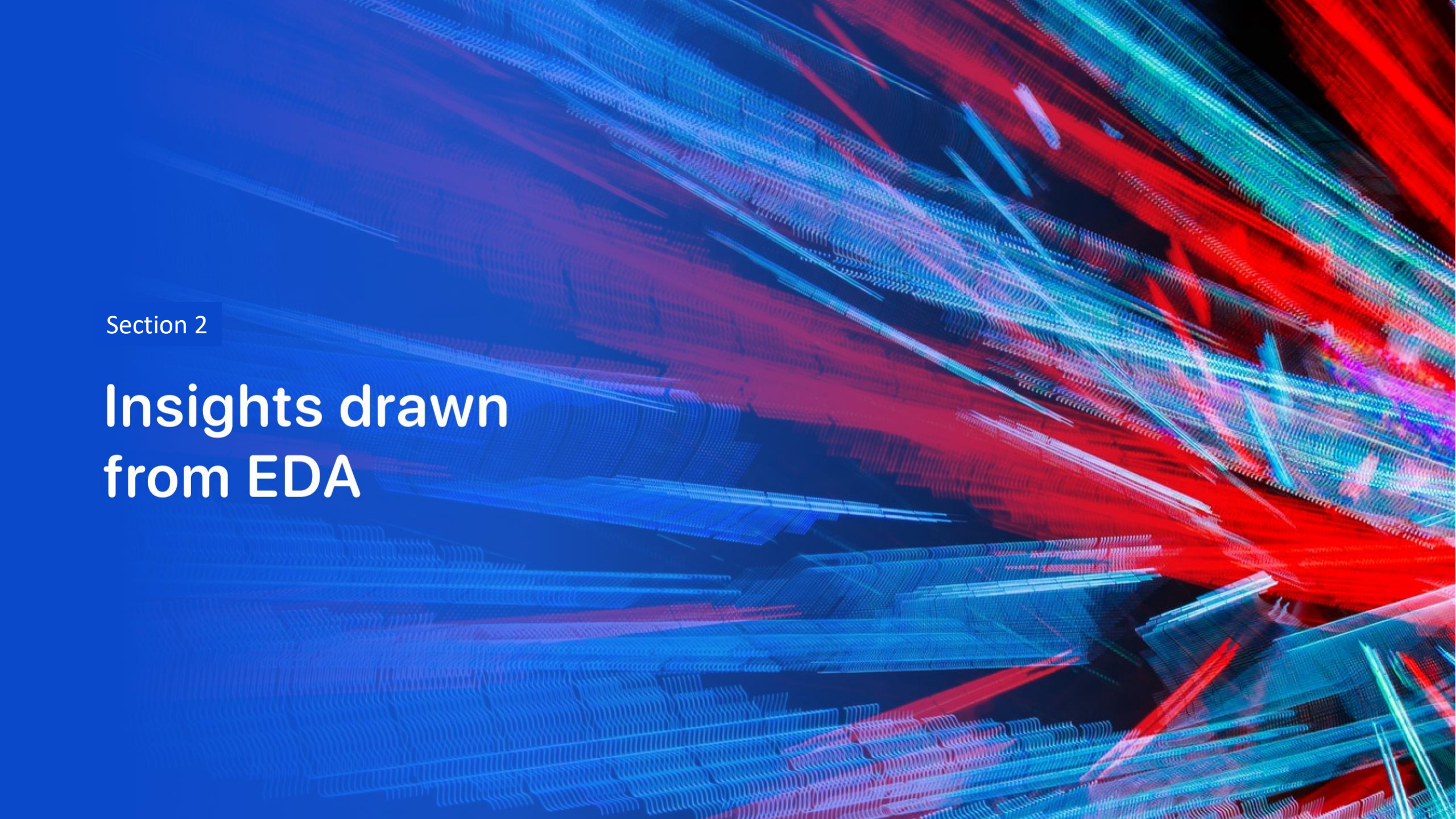
Predictive Analysis (Classification)



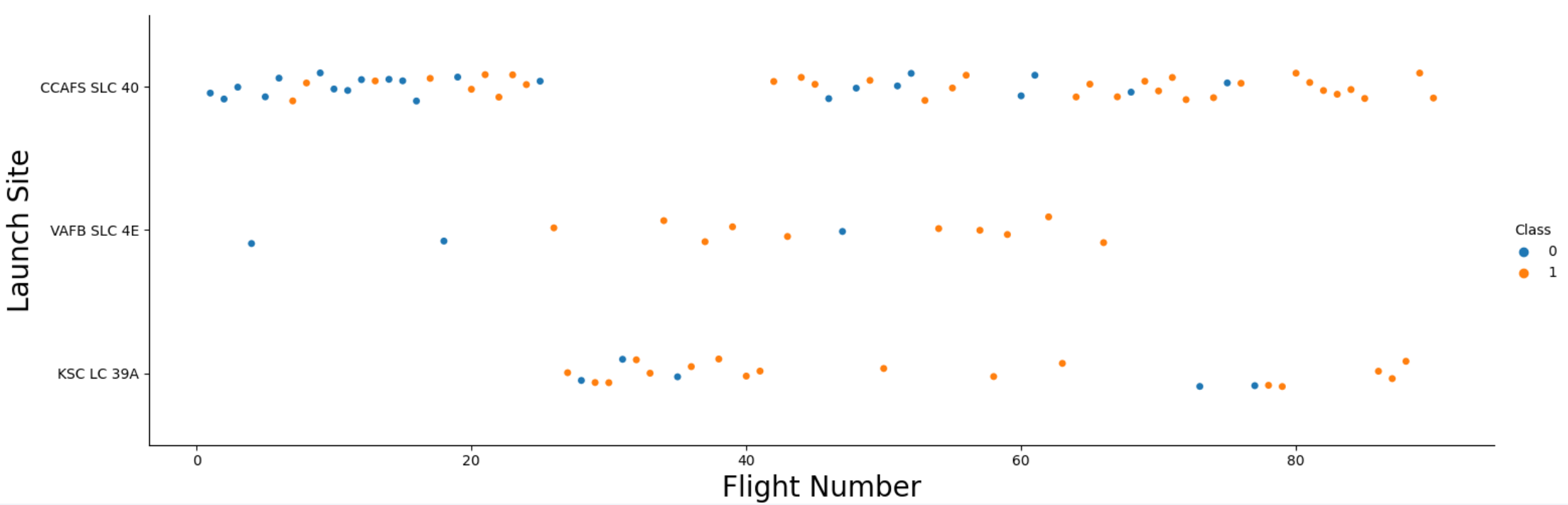
GitHub Link : https://github.com/MarcelKurniawan/IBM-Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

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

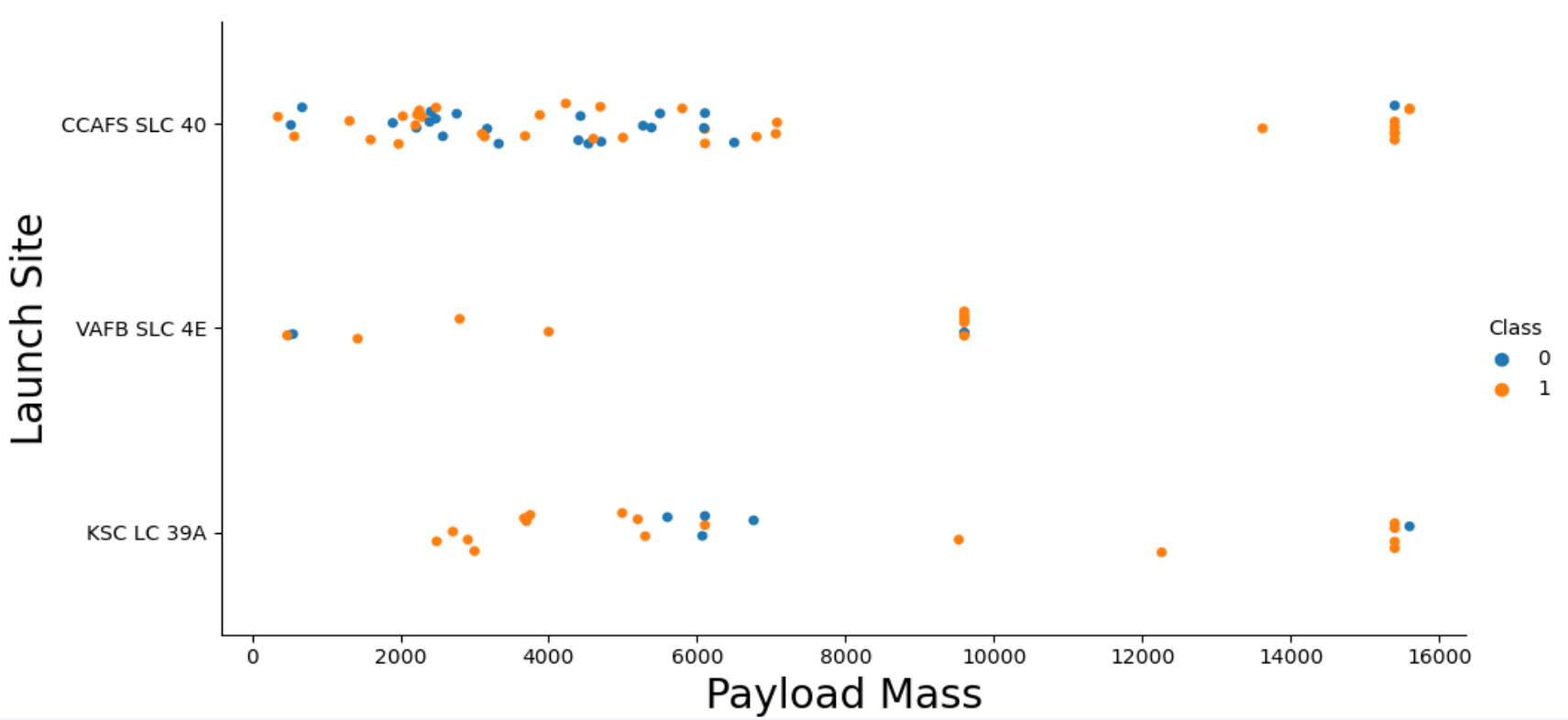

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



- VAFB SLC 4E and KSC LC 39A have higher success rates
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)

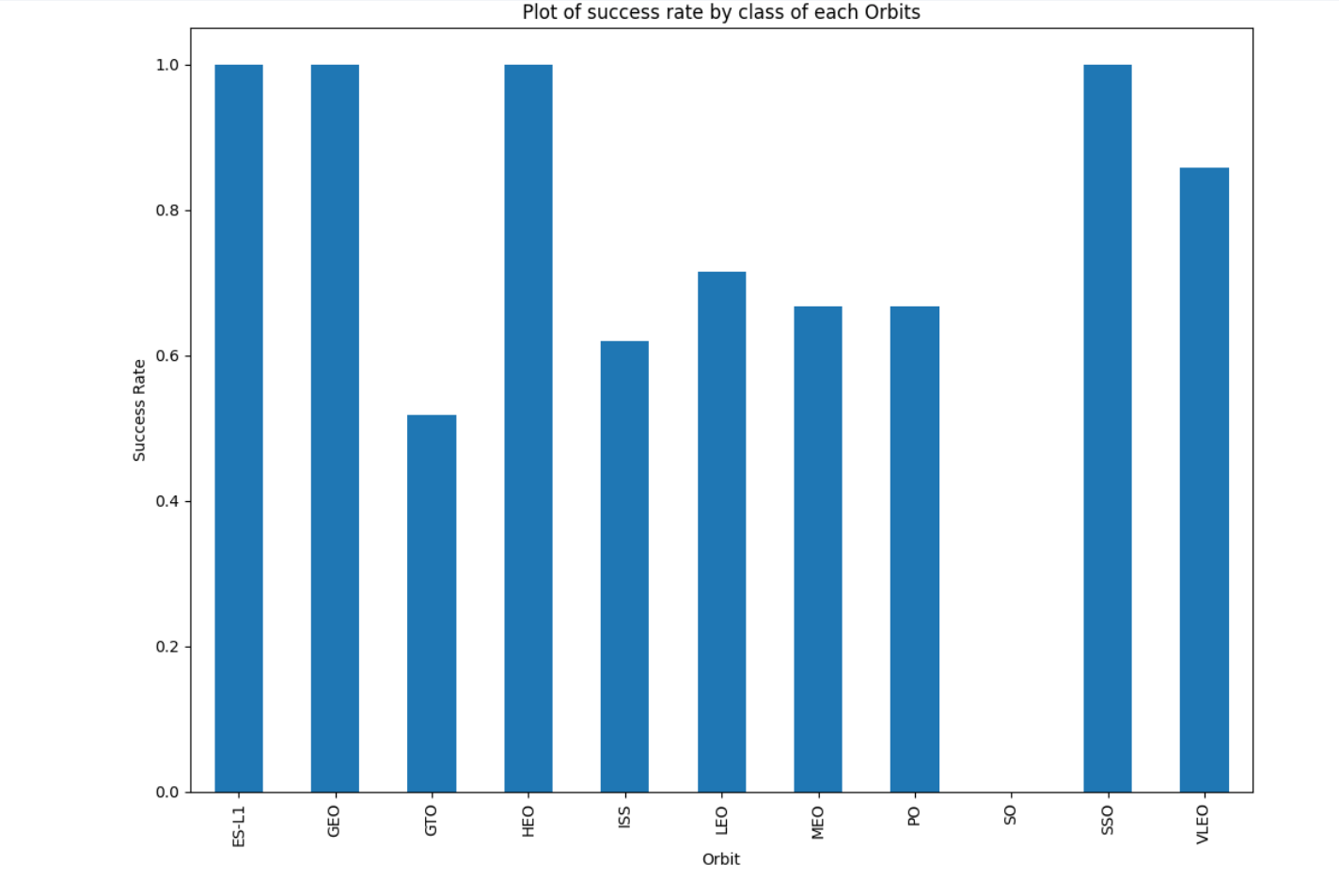
Payload vs. Launch Site



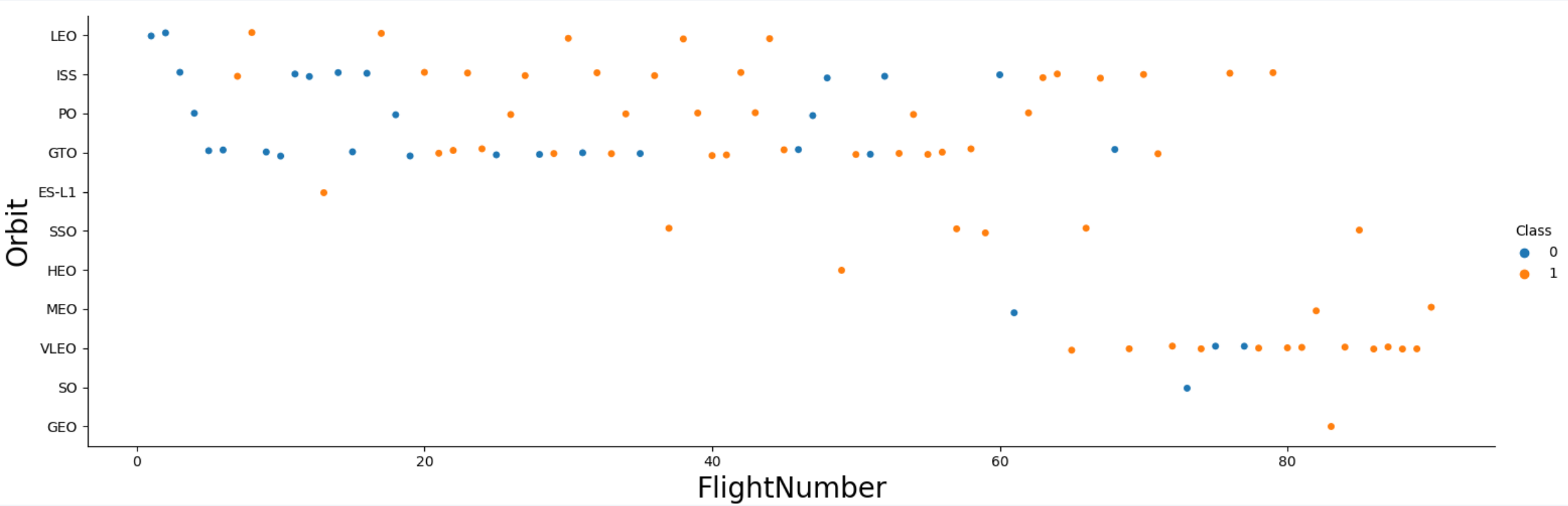
VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type

- The bar chart shows that ES-L1, GEO, HEO, SSO, VLEO had the most success rate

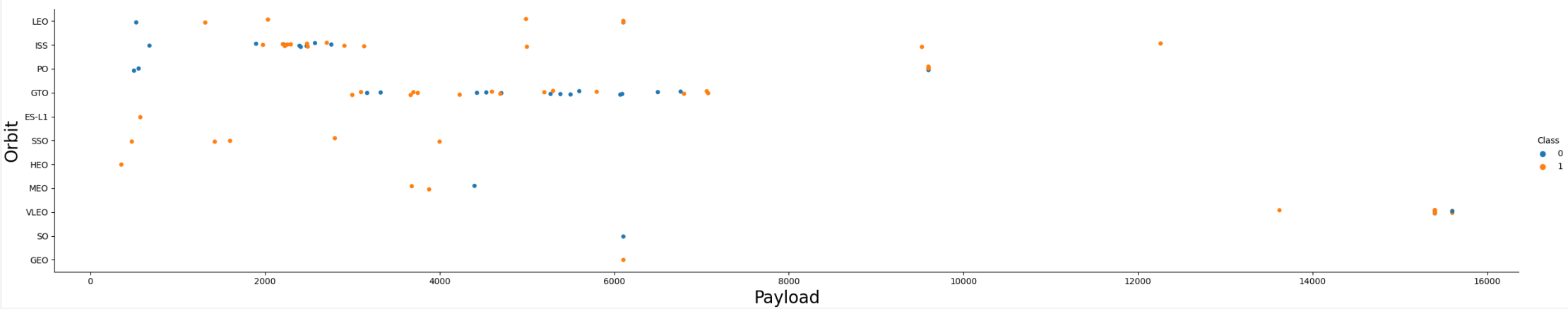


Flight Number vs. Orbit Type



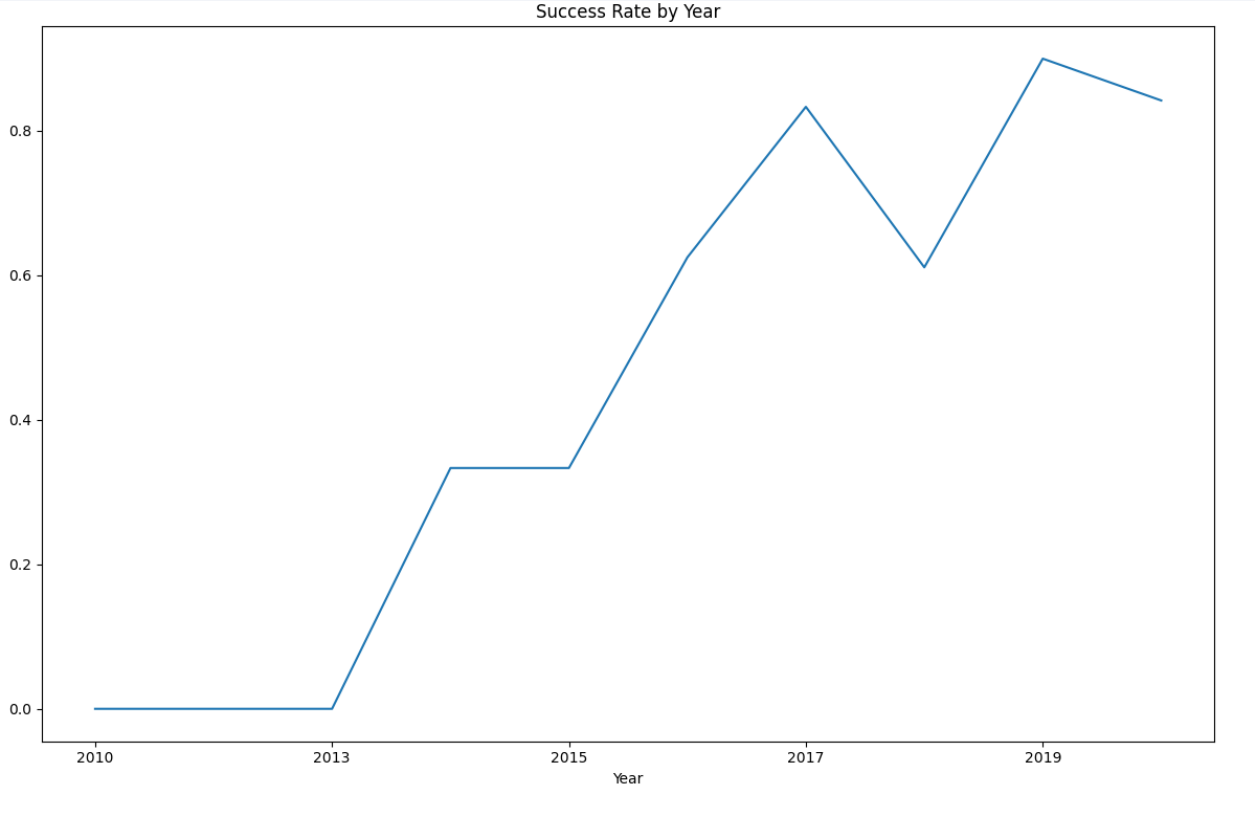
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



The line chats shows that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

```
%%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
None

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer = 'NASA (CRS)'
```

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

SUM(PAYLOAD_MASS__KG_)
45596.0

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

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)'
```

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

MIN(DATE)
01/08/2018

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 4000 AND 6000
```

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

Booster_Version
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(*) AS TOTAL FROM SPACEXTBL GROUP BY Mission_Outcome
```

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

Mission_Outcome	TOTAL
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	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

2015 Launch Records

```
%%sql
SELECT substr(Date,4,2) AS month, DATE, Booster_Version, Launch_Site, Landing_Outcome FROM SPACEXTBL
WHERE Landing_OutcomE = 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

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

month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	01/10/2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14/04/2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
%%sql
SELECT Landing_Outcome, COUNT(*) AS Rank FROM SPACEXTBL
WHERE DATE BETWEEN '04-06-2010' AND '20-03-2017' GROUP BY Landing_Outcome ORDER BY Rank DESC;
```

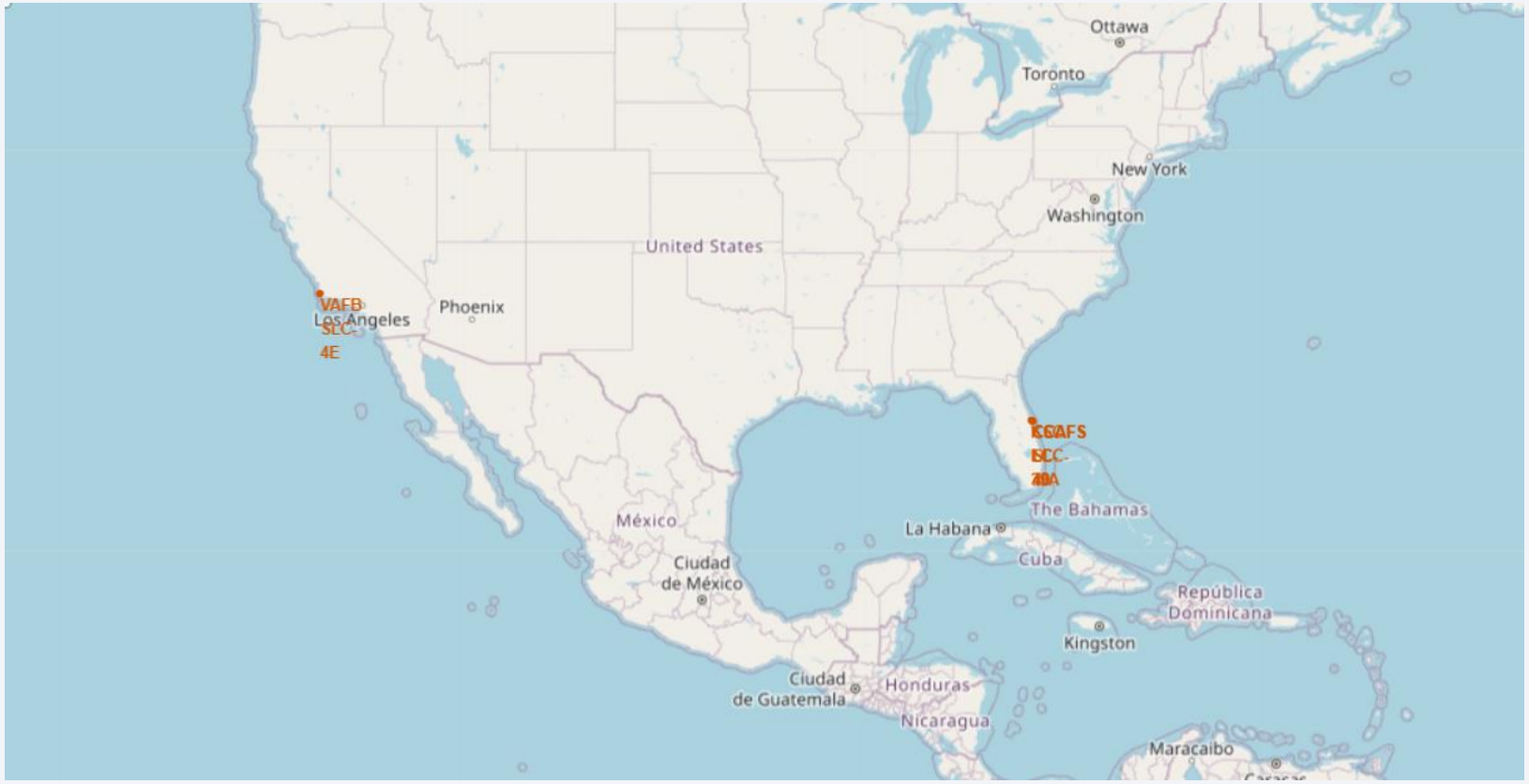
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	Rank
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	7
Failure (drone ship)	3
Failure	3
Failure (parachute)	2
Controlled (ocean)	2
No attempt	1

Section 3

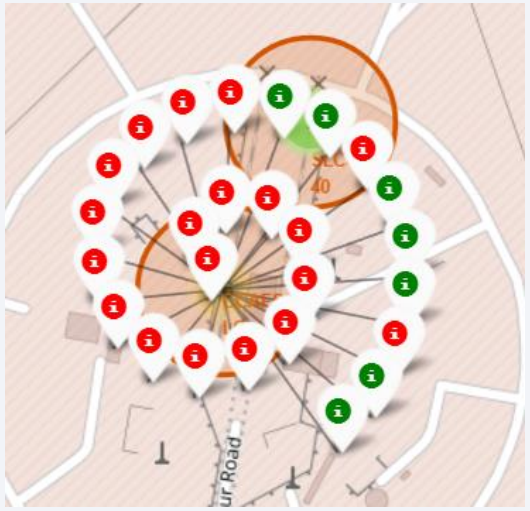
Launch Sites Proximities Analysis

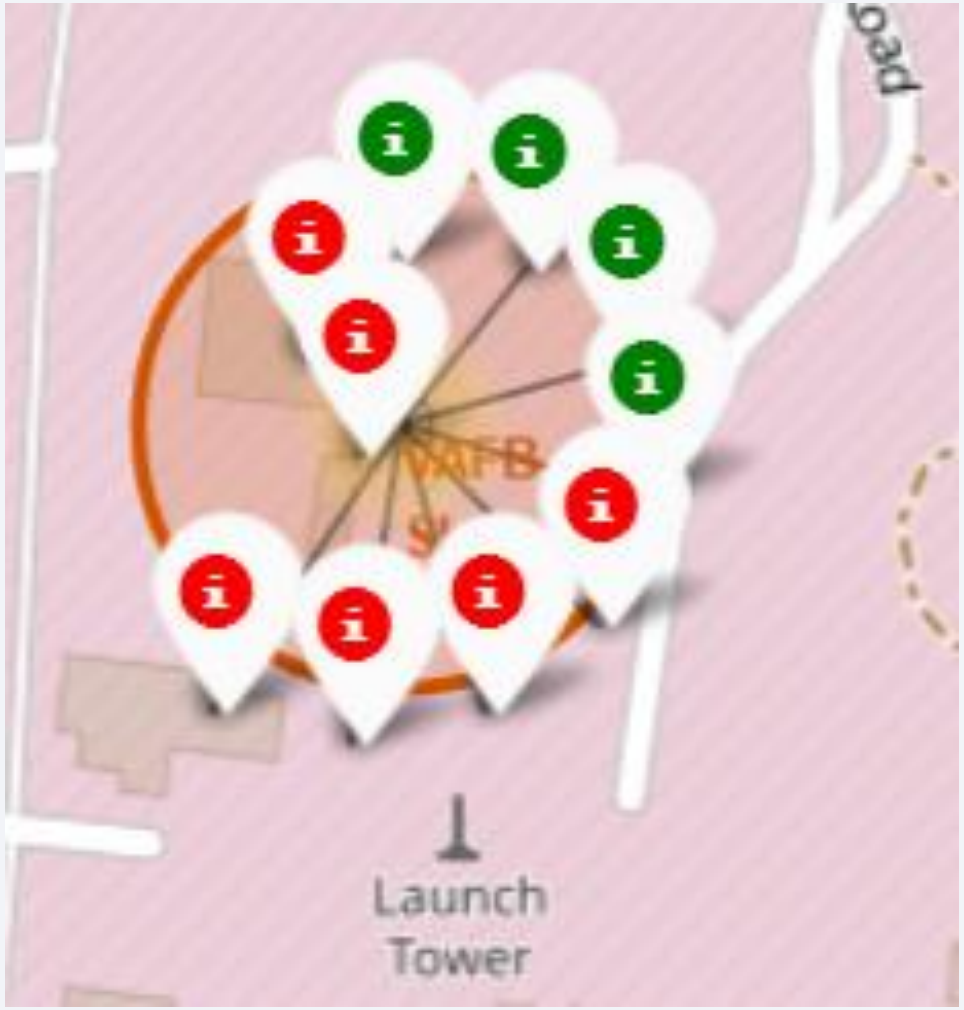
Launch Sites

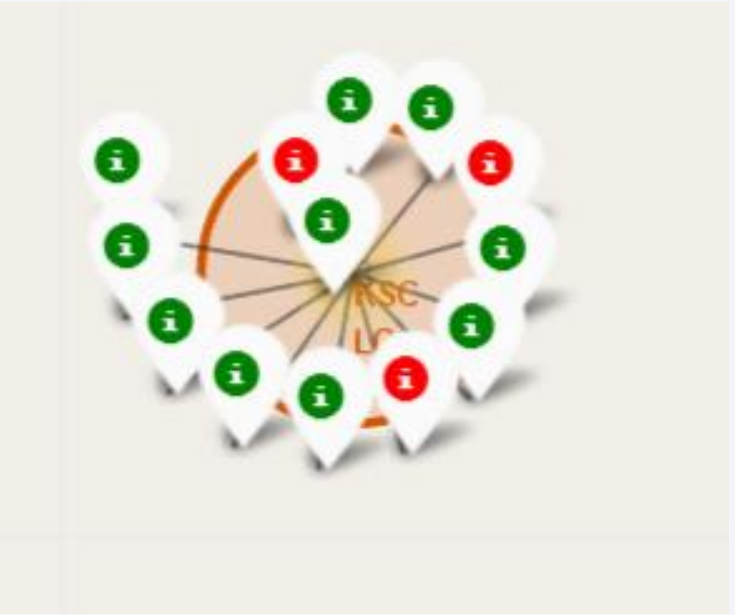


SpaceX launch site located in USA, California and Florida

Launch Outcomes

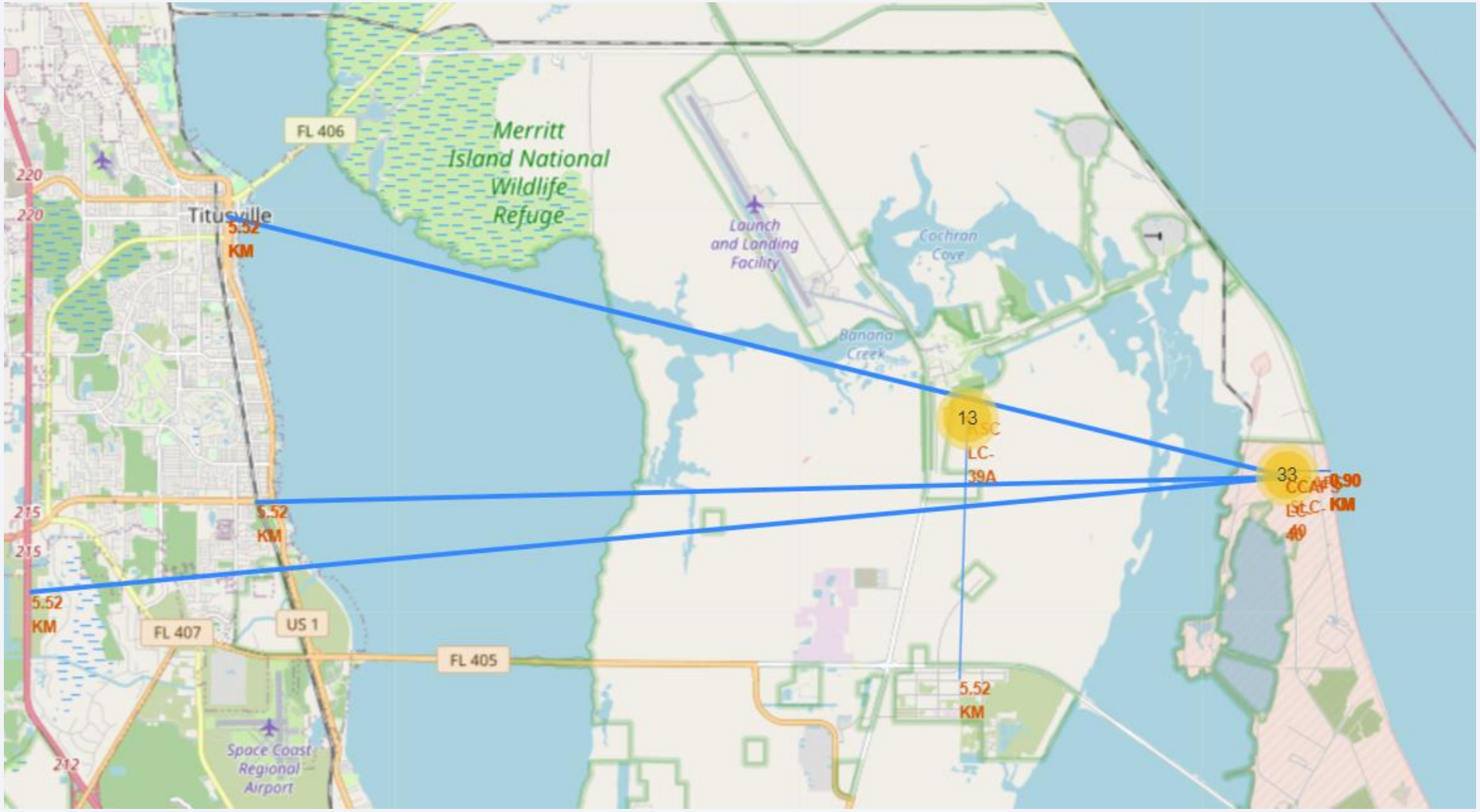


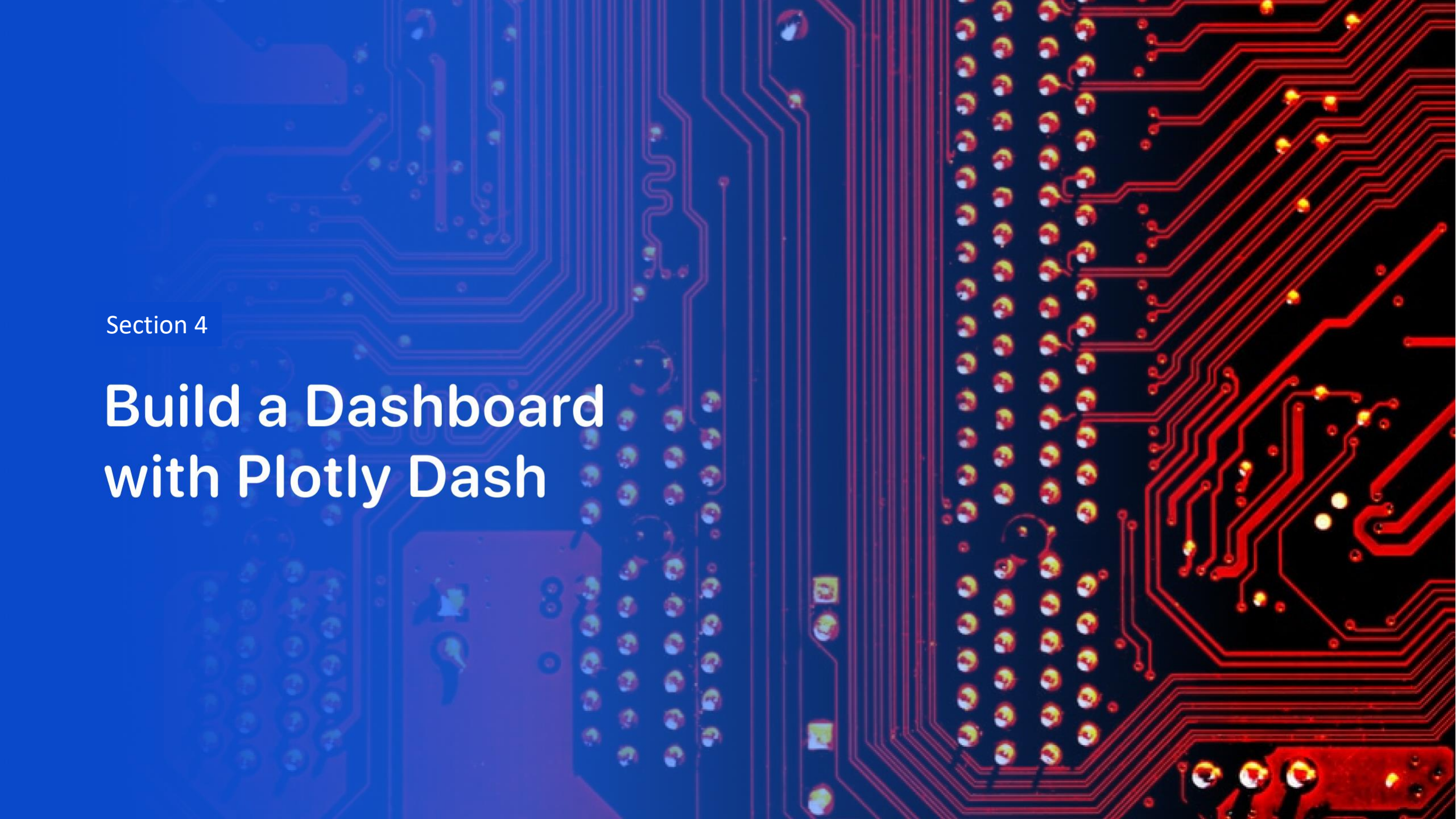




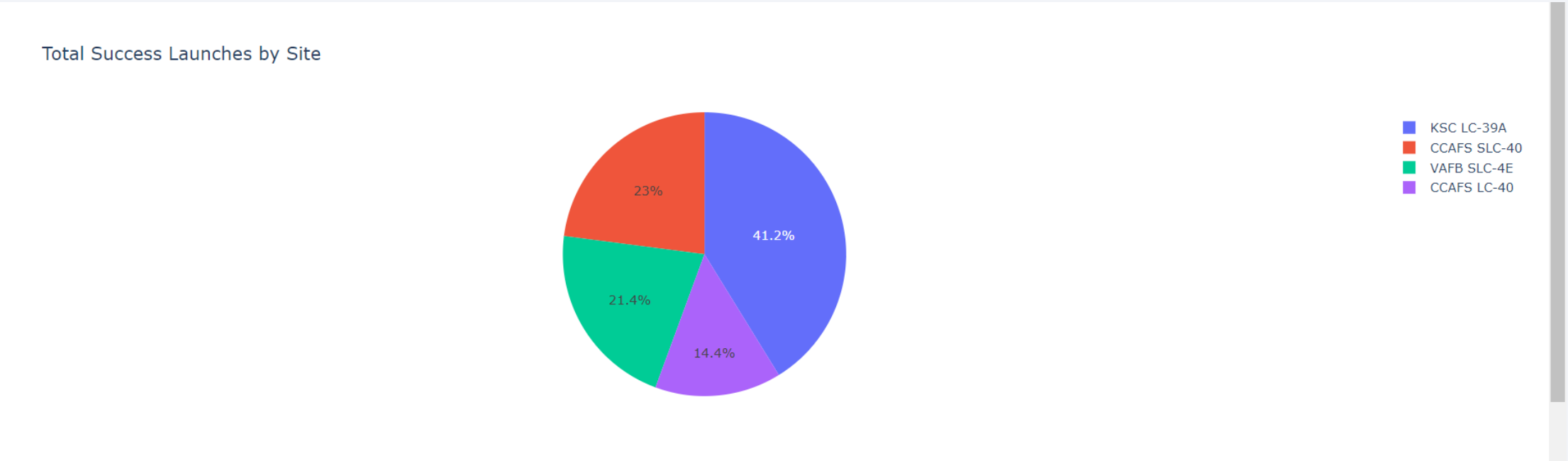
Green markers for successful launches
Red markers for unsuccessful launches

Launch Site distance to landmarks




Section 4
Build a Dashboard with Plotly Dash

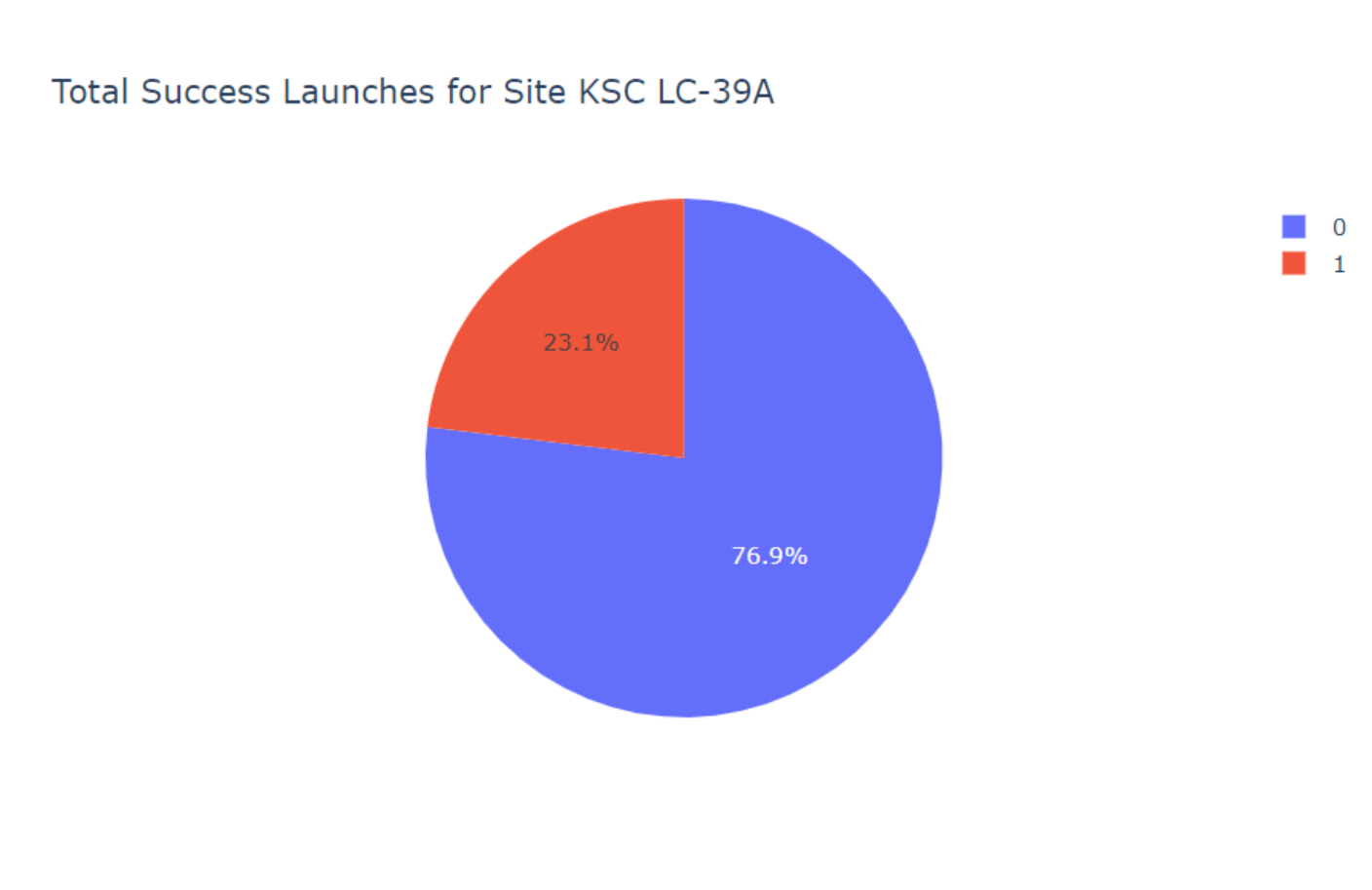
Launch success count for all sites



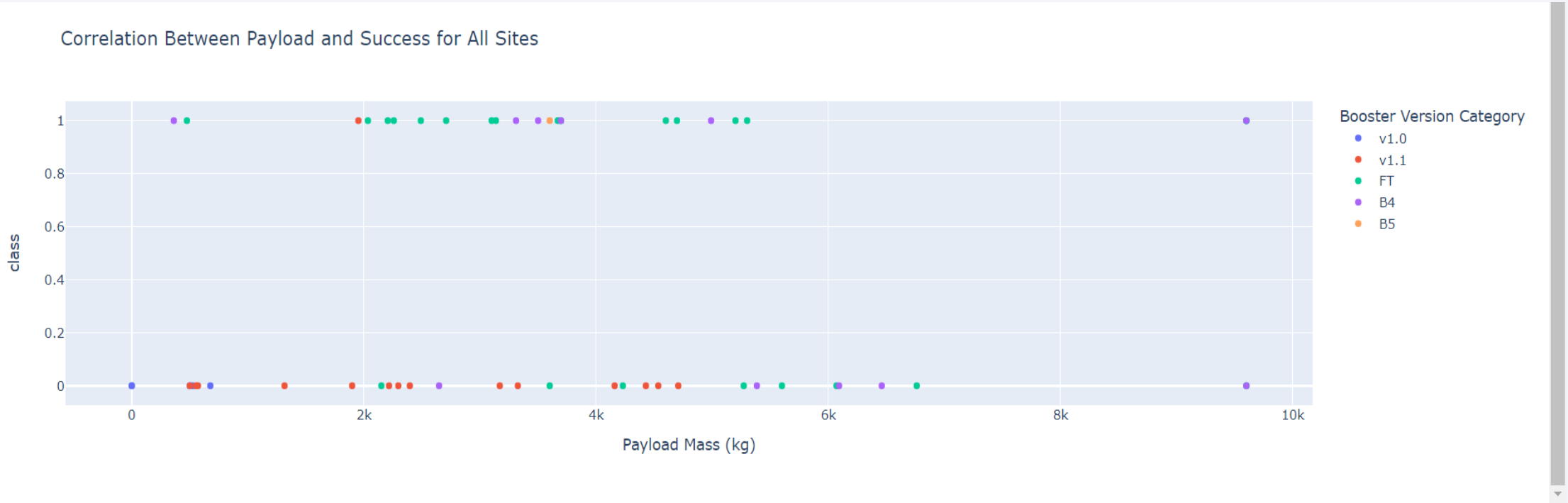
- KSC LC-39A has the higher success launch
- CCAFS LC-40 has the lower success launch

Launch site with highest launch success ratio

- Total KSC LC-39A success rate is 76.9%



Payload Mass vs. Launch Outcome for all sites



- The scatter plot show that payloads between 2000 and 5500 kg have the highest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The model of decision tree classifier is the model with the highest accuracy and F1-Scores

	LogReg	SVM	Tree	KNN
F1_Score	0.888889	0.888889	0.916667	0.888889
Accuracy	0.833333	0.833333	0.888889	0.833333

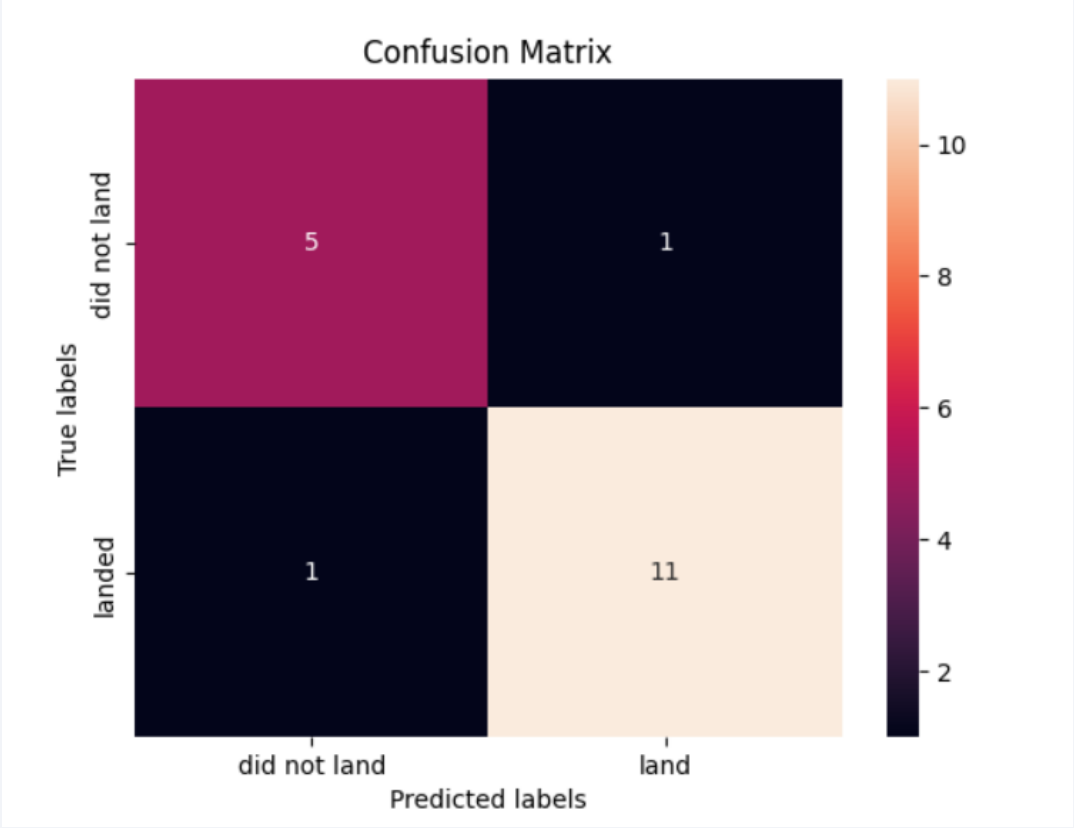
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9017857142857142
Best params is : {'criterion': 'entropy', 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitte
r': 'random'}
```

Confusion Matrix

- Examining the confusion matrix, we see decision tree classifier can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- Decision tree classifier is the best model in this case
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A has the highest successful launches of all sites
- Launch Success Increases over time

Thank you!