IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Marco Cutraro
09/03/2024

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

Executive Summary

Summary of methodologies

- Data Collection through API
- Data Collection with Web Scraping
- Data Wrangling
- Exploratory Data Analysis with SQL
- Exploratory Data Analysis with Data Visualization
- Interactive Visual Analytics with Folium
- Machine Learning Prediction

Summary of all results

- Exploratory Data Analysis result
- Interactive analytics in screenshots
- Predictive Analytics result

Introduction

Project background and context

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because 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. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

Problems you want to find answers

- What factors determine if the rocket will land successfully?
- The interaction amongst various features that determine the success rate of a successful landing.
- What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology – Executive summary

- Data collection
- Web scraping
- Data wrangling
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

Data collection was done using get request to the SpaceX API.

Decode the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

Clean the data, checked for missing values and fill in missing values where necessary.

Web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup.

Extract the launch records as HTML table, parse the table and convert it to a pandas dataframe.

Data Collection – SpaceX API

- Request to the SpaceX API to collect data, clean the requested data and did some data wrangling and formatting.
- Github: https://github.com/marcocutraro/capstone_project_space/blob/main/2_data_collection.ipynb

Data Collection - Scraping

- Web scraping Falcon 9 launch records with BeautifulSoup, parse the table and converte it into a pandas dataframe.
- Github: https://github.com/marcocutraro/capstone_project_space/blob/main/3_web_scraping.ipynb

Data Wrangling

- Exploratory data analysis
- We calculated the number of launches at each site, and the number and occurrence of each orbits, we also created landing outcome label from outcome column and exported the results to csv.
- Github: https://github.com/marcocutraro/capstone_project_space/blob/main/4_data_wrangling.ipynb

EDA with Data Visualization

- We explored the data by visualizing the relationship between:
  - flight number and launch site, payload
  - launch site, success rate of each orbit type, flight number and orbit type
  - the launch success yearly trend.

Github: https://github.com/marcocutraro/capstone_project_space/blob/main/6_EDA_dataviz.ipynb

EDA with SQL

- EDA with SQL to get insight from the data:
  - The names of unique launch sites in the space mission.
  - The total payload mass carried by boosters launched by NASA (CRS)
  - The average payload mass carried by booster version F9 v1.1
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.
- Github: https://github.com/marcocutraro/capstone_project_space/blob/main/5_EDA_sql.ipynb

Build an Interactive Map with Folium

- **TASK 1:** Mark all launch sites on a map
- **TASK 2:** Mark the success/failed launches for each site on the map
- **TASK 3:** Calculate the distances between a launch site to its proximities

to find some geographical patterns about launch sites.

After you plot distance lines to the proximities, you can answer the following questions easily:

- Are launch sites in close proximity to railways?
- Are launch sites in close proximity to highways?
- Are launch sites in close proximity to coastline?
- Do launch sites keep certain distance away from cities?

Build a Dashboard with Plotly Dash

- We built an interactive dashboard with Plotly dash
- We plotted pie charts showing the total launches by a certain sites
- We plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
  - Github: https://github.com/marcocutraro/capstone_project_space/blob/main/8_Visual_Analytics_Plotly_Dash.ipynb

Predictive Analysis (Classification)

- Perform exploratory Data Analysis and determine Training Labels
- create a column for the class
- Standardize the data
- Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
- Find the method performs best using test data
  - Github: https://github.com/marcocutraro/capstone_project_space/blob/main/9_Machine_Learning_Pipeline.ipynb

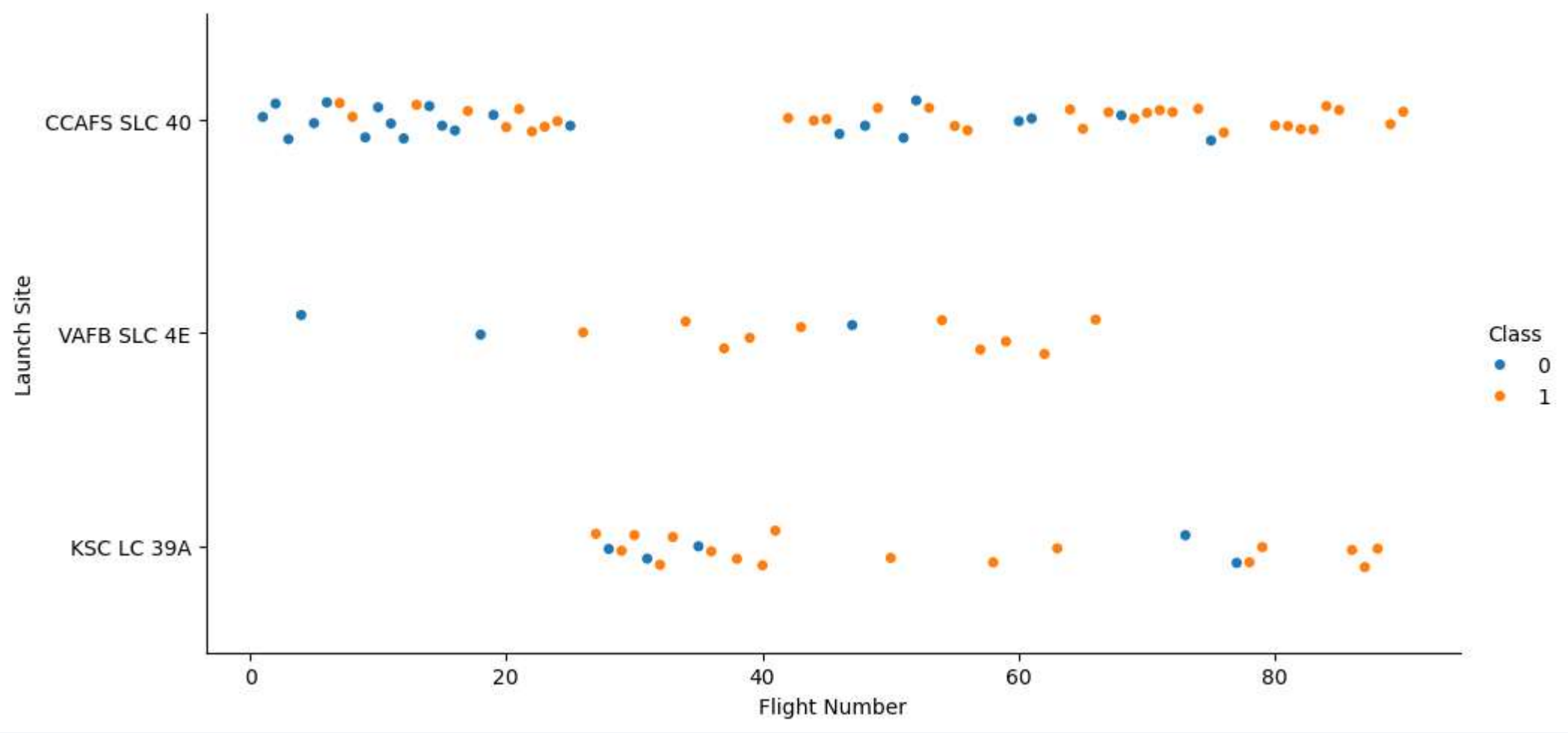
Results

Best model is DecisionTree with a score of 0.8732142857142857

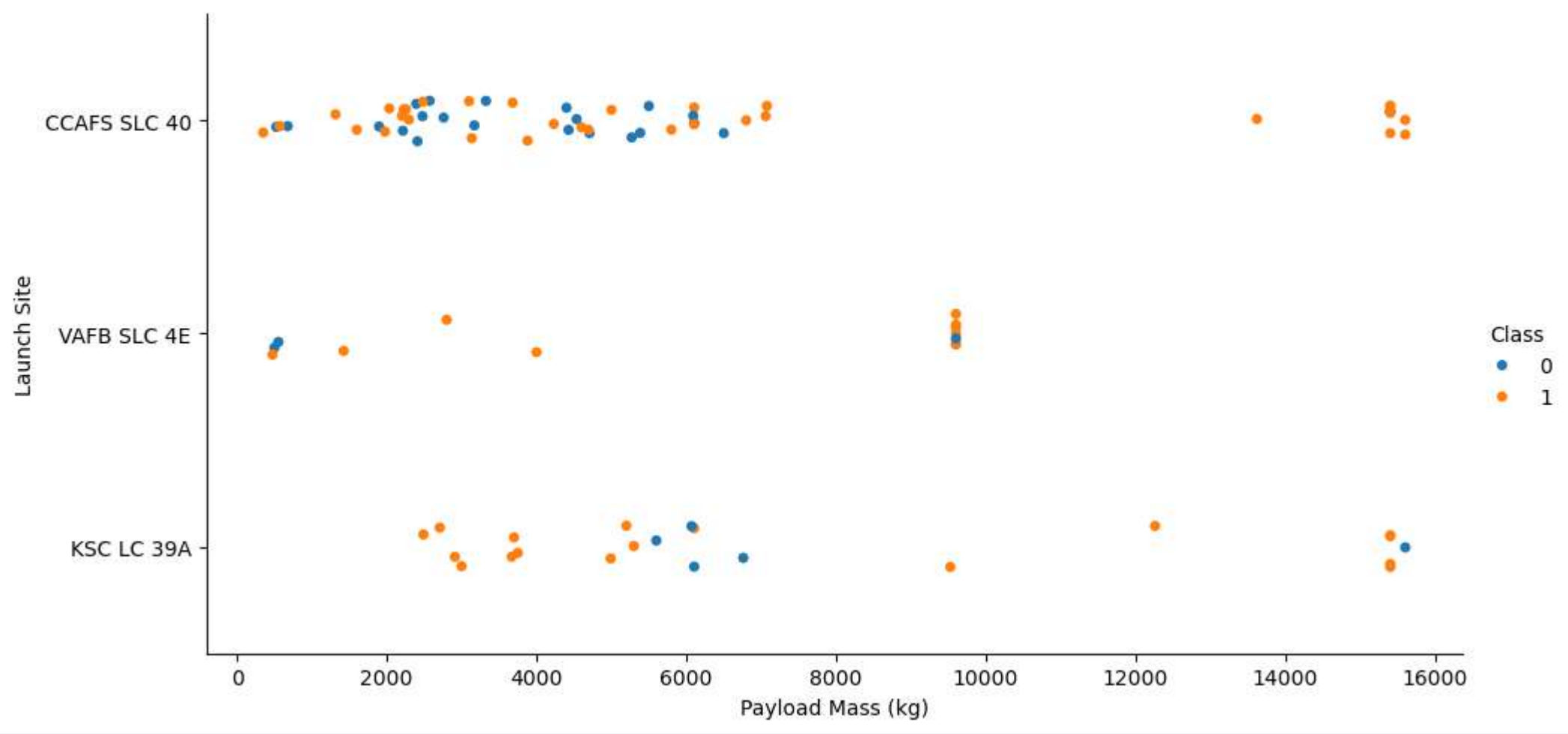
Section 2

Insights drawn from EDA

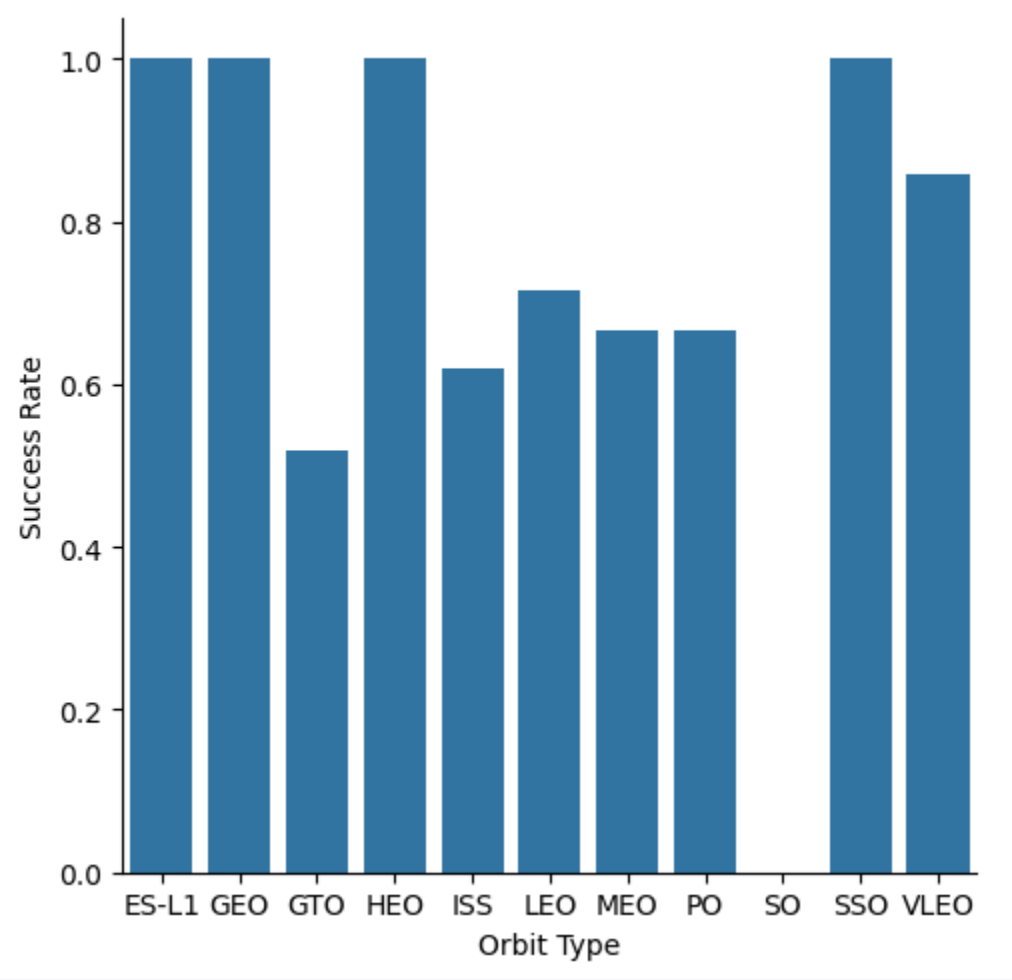
Flight Number vs. Launch Site



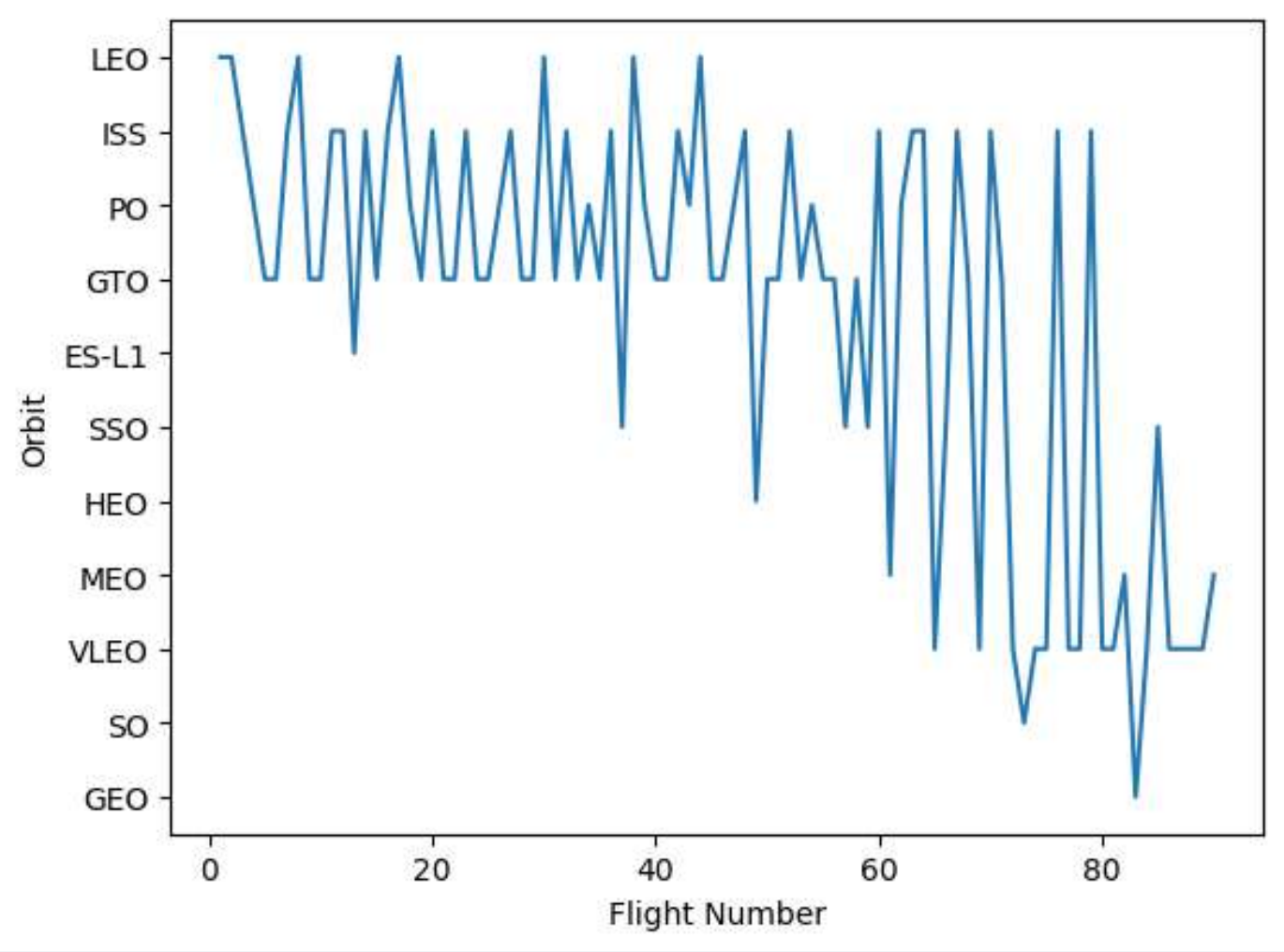
Payload vs. Launch Site



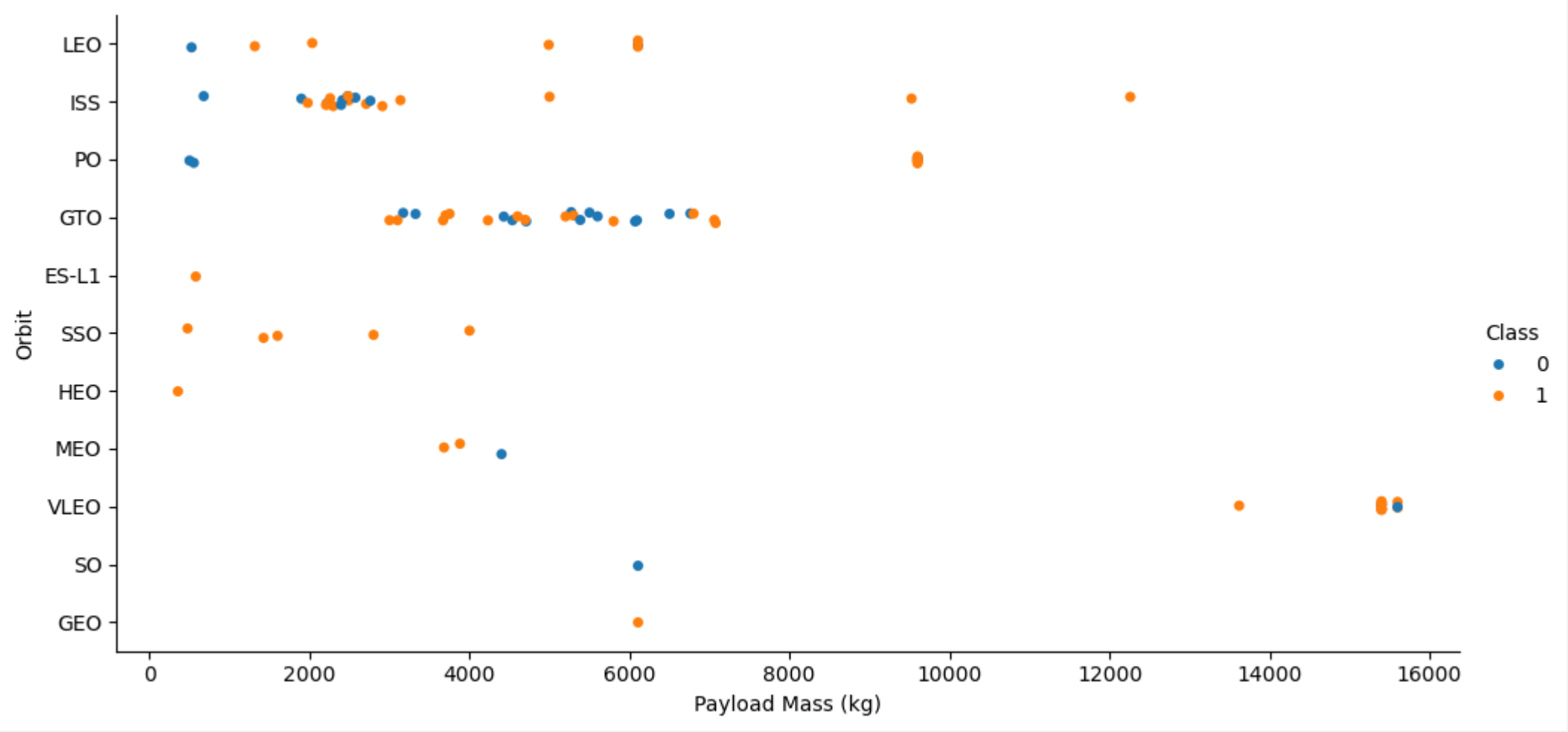
Success Rate vs. Orbit Type



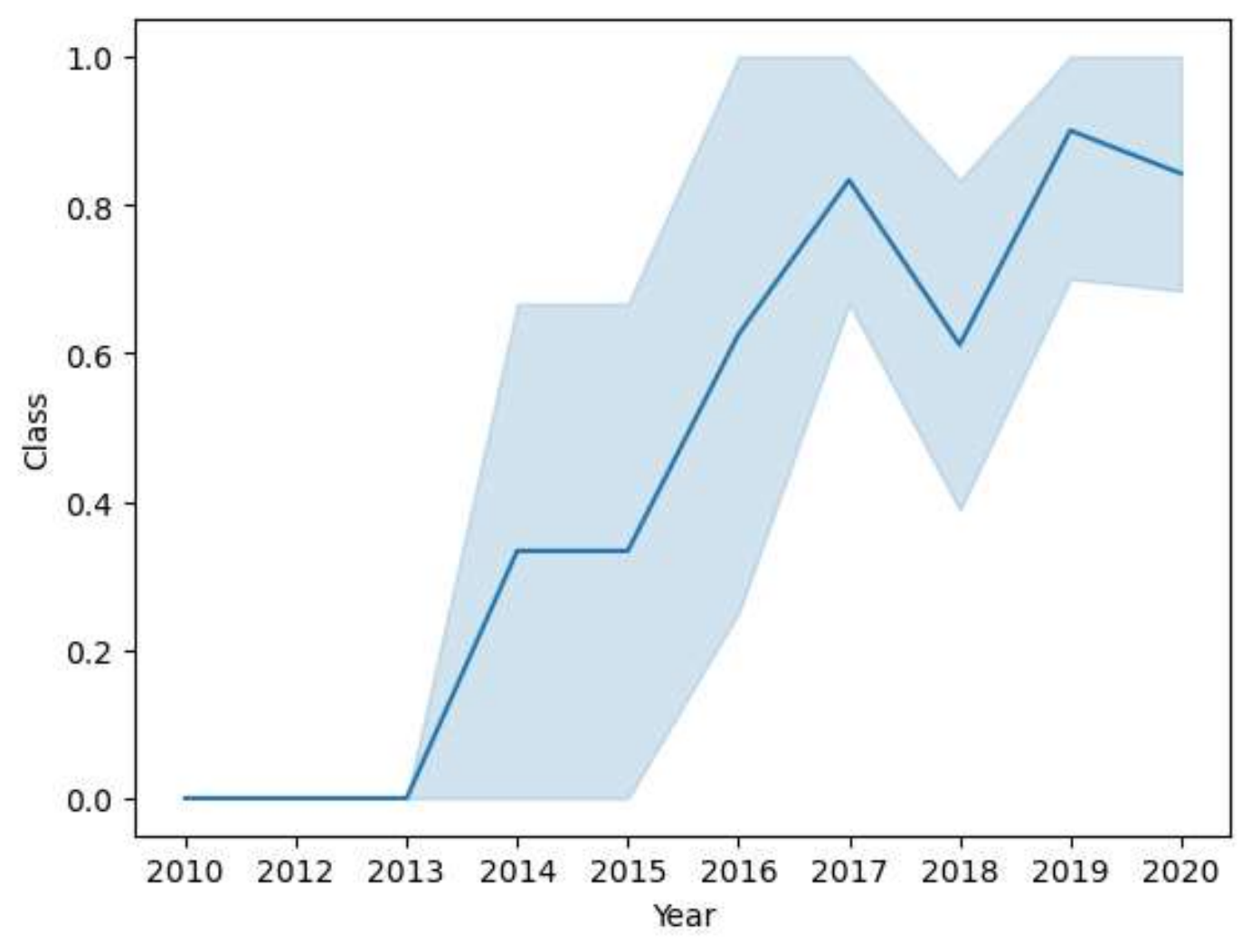
Flight Number vs. Orbit Type



Payload vs. Orbit Type



Launch Success Yearly Trend



All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

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

```
%sql
SELECT Unique(LAUNCH_SITE)
FROM SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

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

```
%sql
SELECT * \
FROM SPACEXTBL \
WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;
```

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

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

```
%sql
SELECT SUM(PAYLOAD_MASS__KG_) \
FROM SPACEXTBL \
WHERE CUSTOMER = 'NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

Display average payload mass carried by booster version F9 v1.1

```
%sql
SELECT AVG(PAYLOAD_MASS__KG_) \
FROM SPACEXTBL \
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

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

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

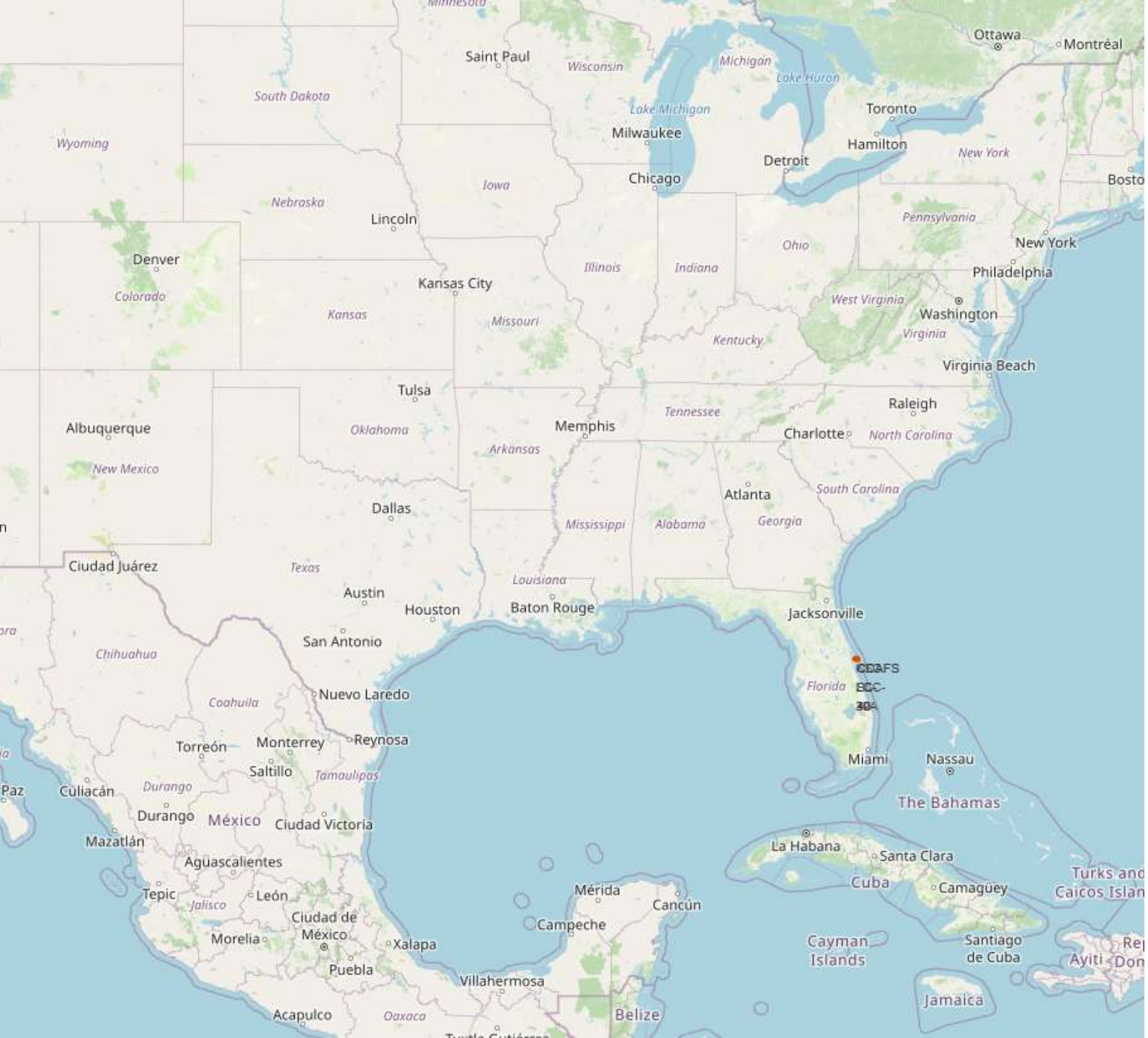
- Present your query result with a short explanation here

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

```
%sql
SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

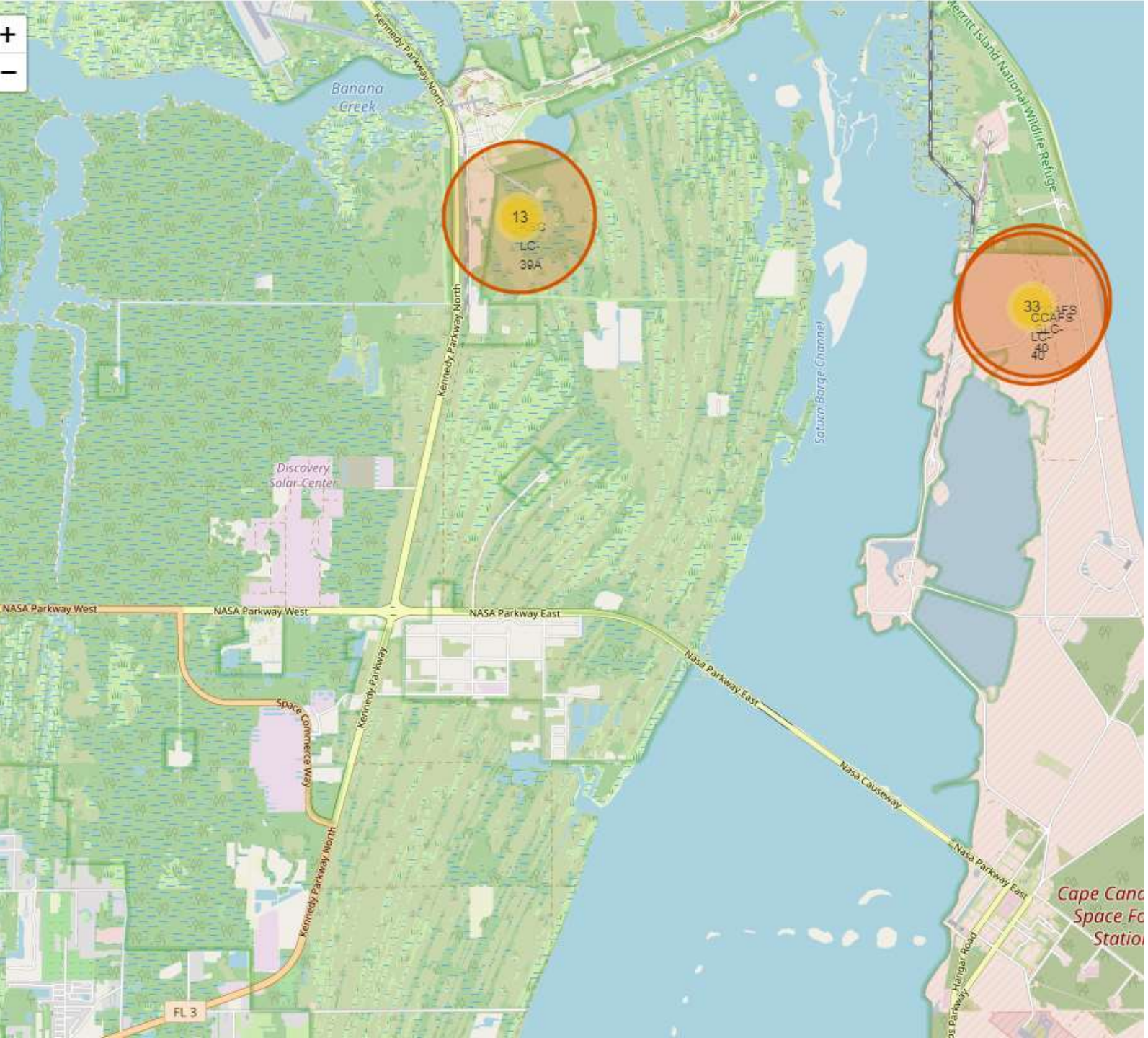
Section 3

Launch Sites Proximities Analysis



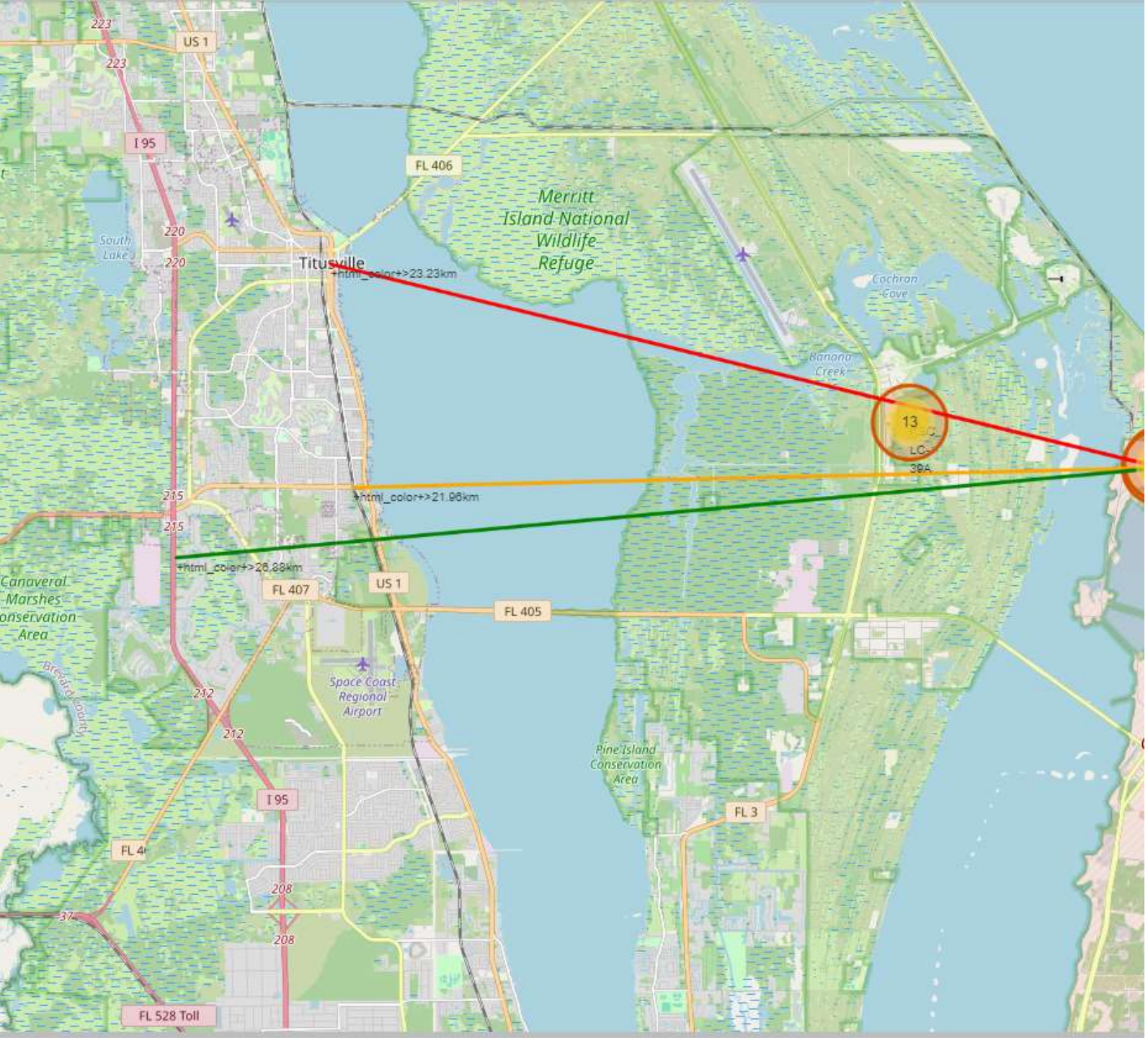
Launch sites' location markers

- Now, you can explore the map by zoom-in/out the marked areas , and try to answer the following questions:
- Are all launch sites in proximity to the Equator line?
- Are all launch sites in very close proximity to the coast?



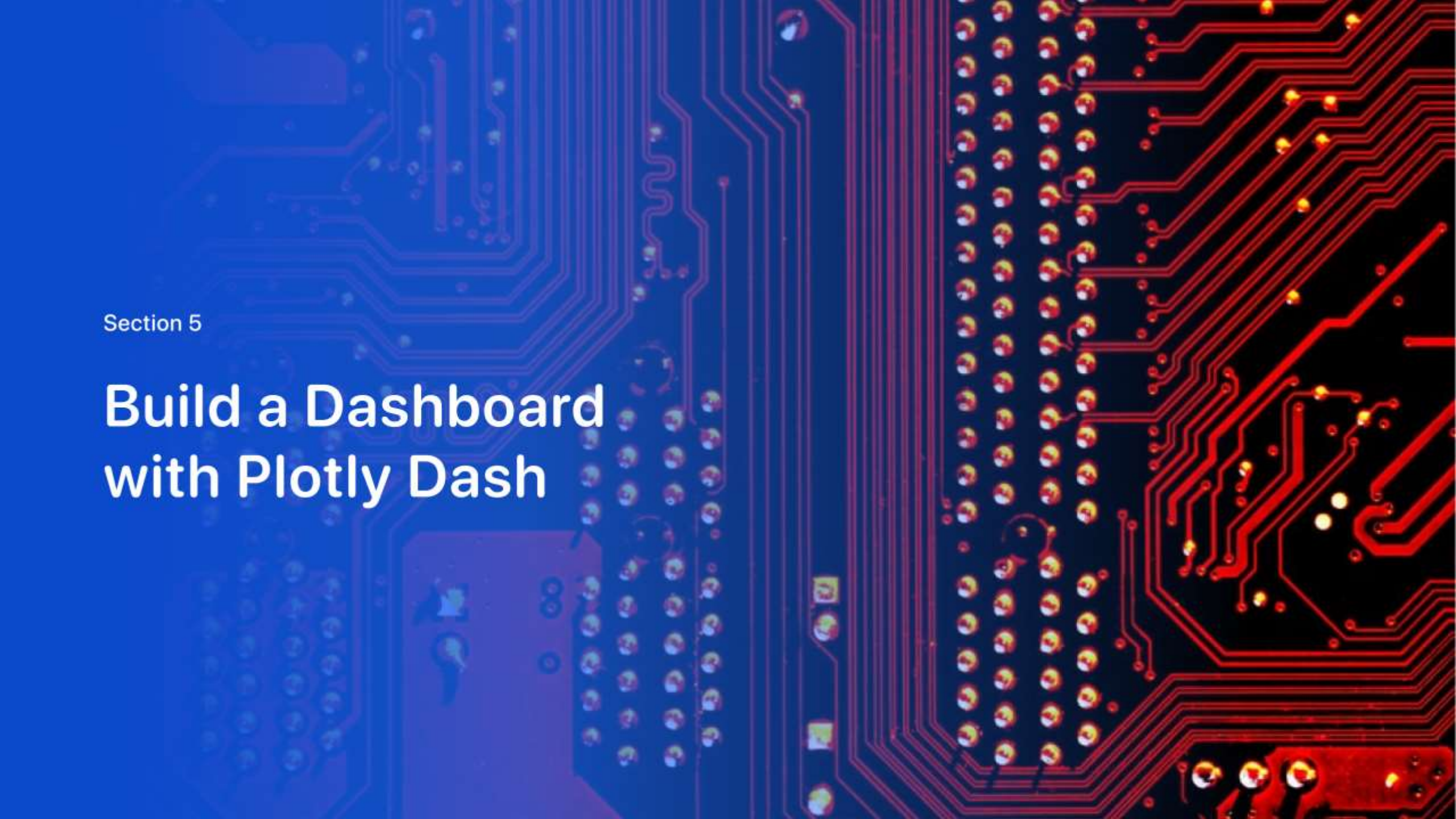
Launch sites' color-labeled markers

- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates

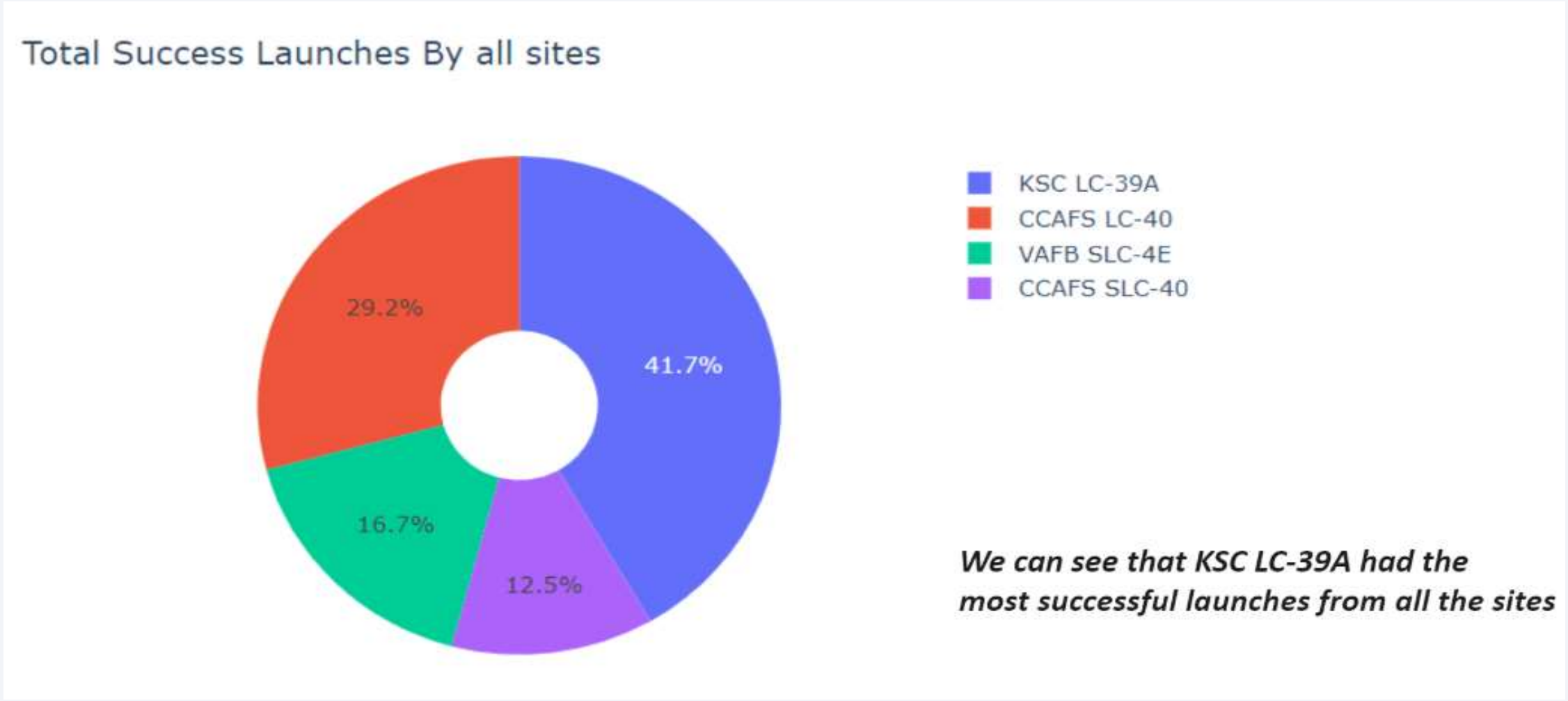


Launch sites' to its proximities

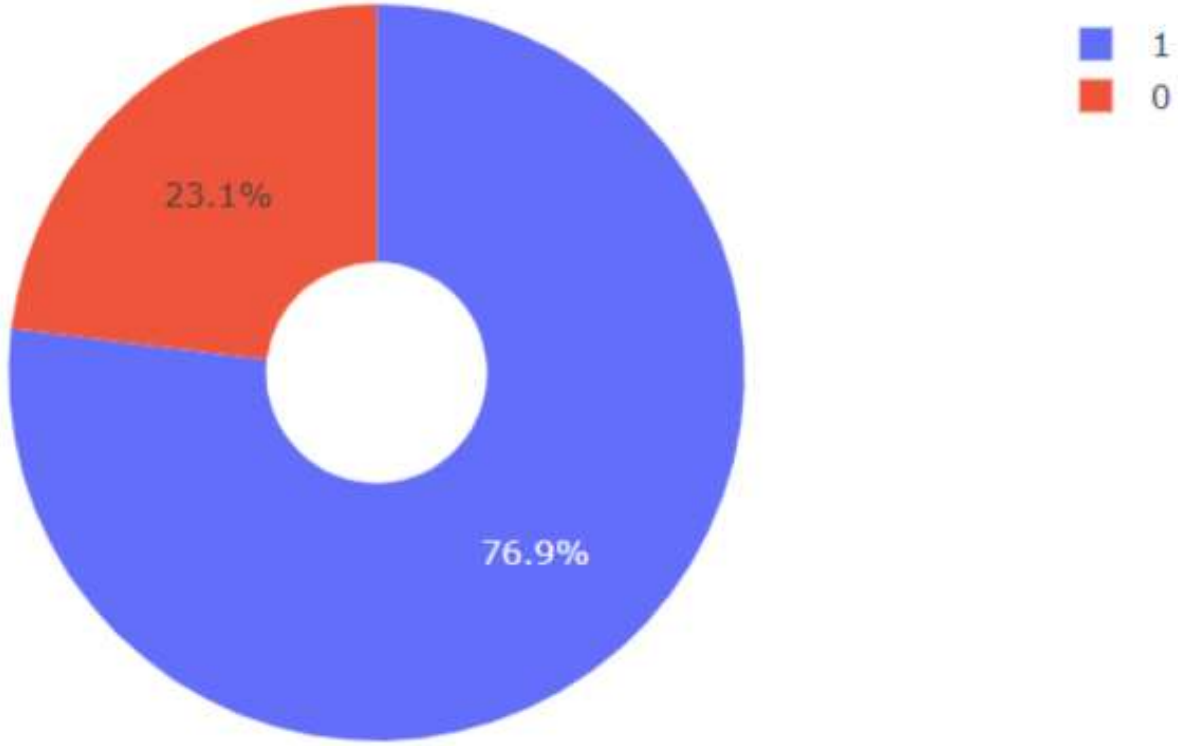
- Now zoom in to a launch site and explore its proximity to see if you can easily find any railway, highway, coastline, etc. Move your mouse to these points and mark down their coordinates (shown on the top-left) in order to the distance to the launch site.


Section 5
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

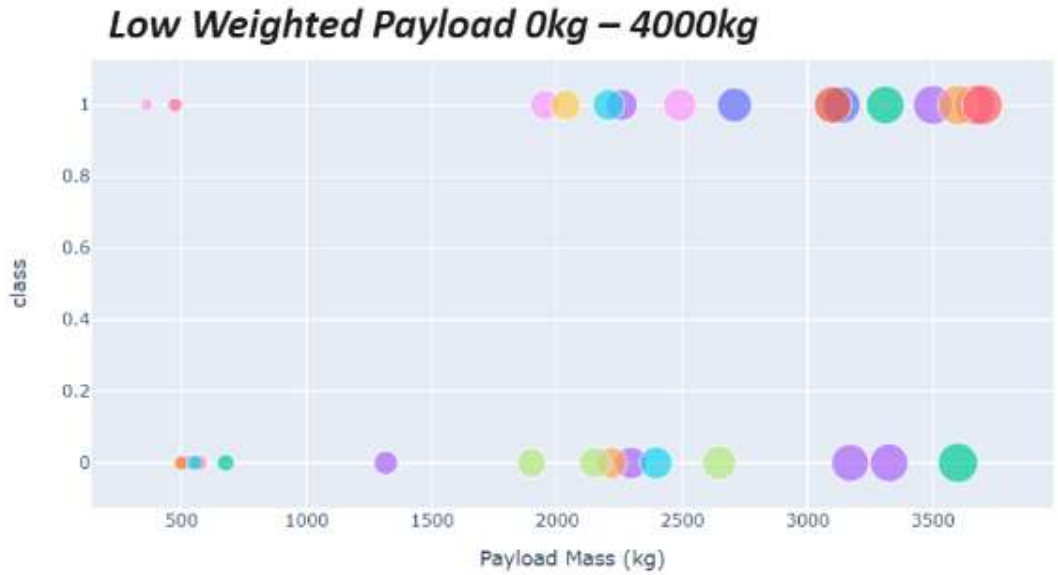


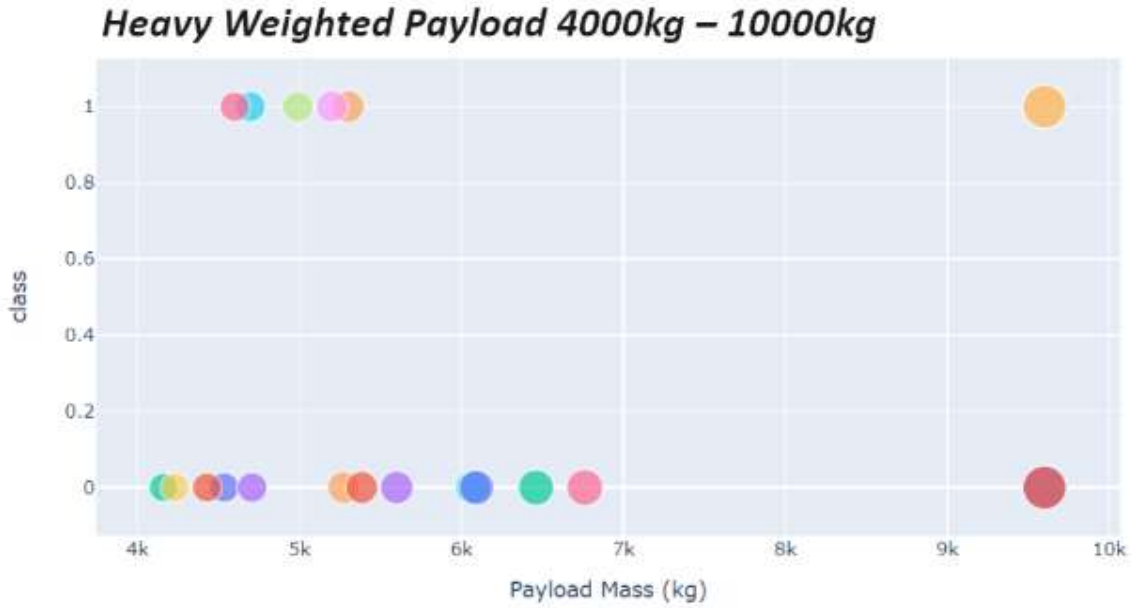
Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

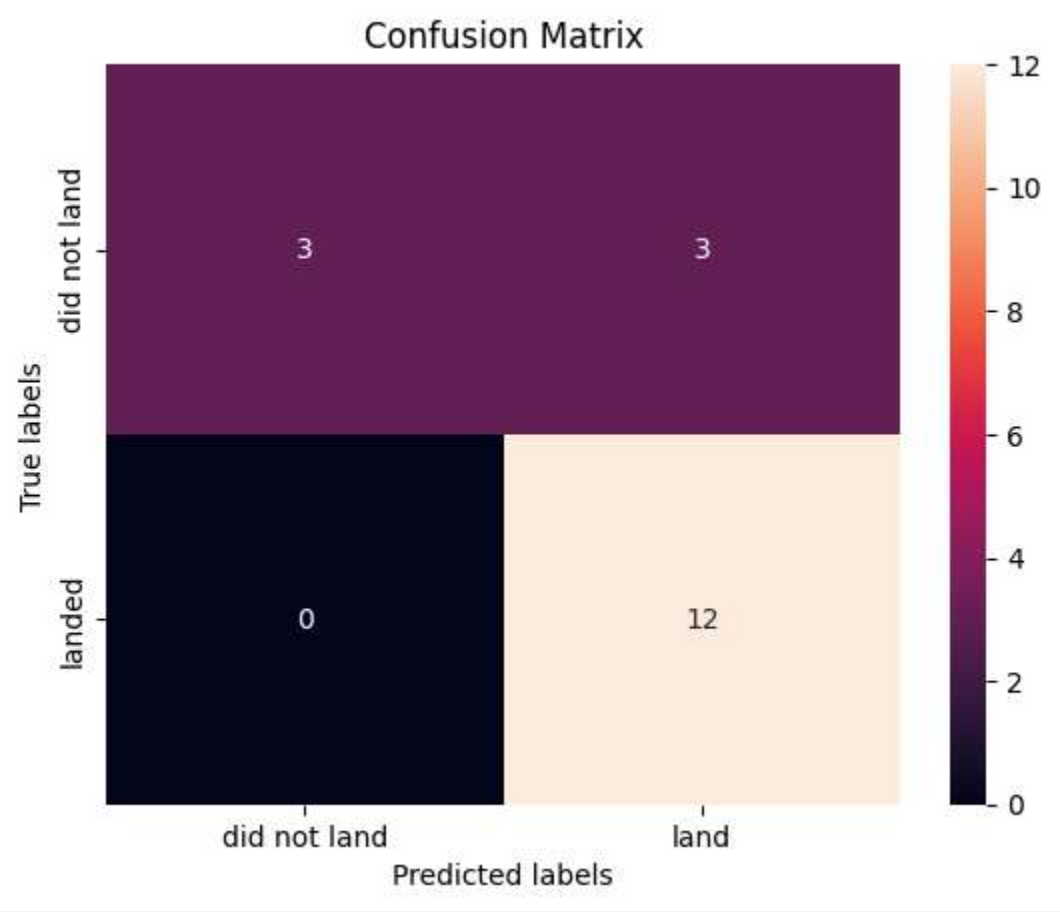
Classification Accuracy

```
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[bestalg
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 0.8732142857142857
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_feature
s': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'splitte
r': 'random'}
```

Confusion Matrix



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!