IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Mira Ehab
25-9-2022

# Outline

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

# 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**

Worked on a new rocket company named "SpaceY" that would like to compete with SpaceX founded by Billionaire industrialist Allon Musk. 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.

- **Problems you want to find answers**
- - Find the features that affect the first stage landing success
- - predict if the first stage will land successfully ?

Section 1

# Methodology

# Methodology

## Executive Summary

- **Data collection methodology:**
  - Collected the rocket launch data from the SpaceX API
  - Collect Falcon 9 historical launch records using Web scraping
- **Perform data wrangling:**
  - Used one hot encoding to deal with categorical data
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**

# Data Collection

- Collected the rocket launch data from the SpaceX API
- Web-scraped the SpaceX table in the Wikipedia page to collect Falcon 9 historical launch records

# Data Collection – SpaceX API

- the GitHub URL https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/collectingTheData.ipynb

# Data Collection - Scraping

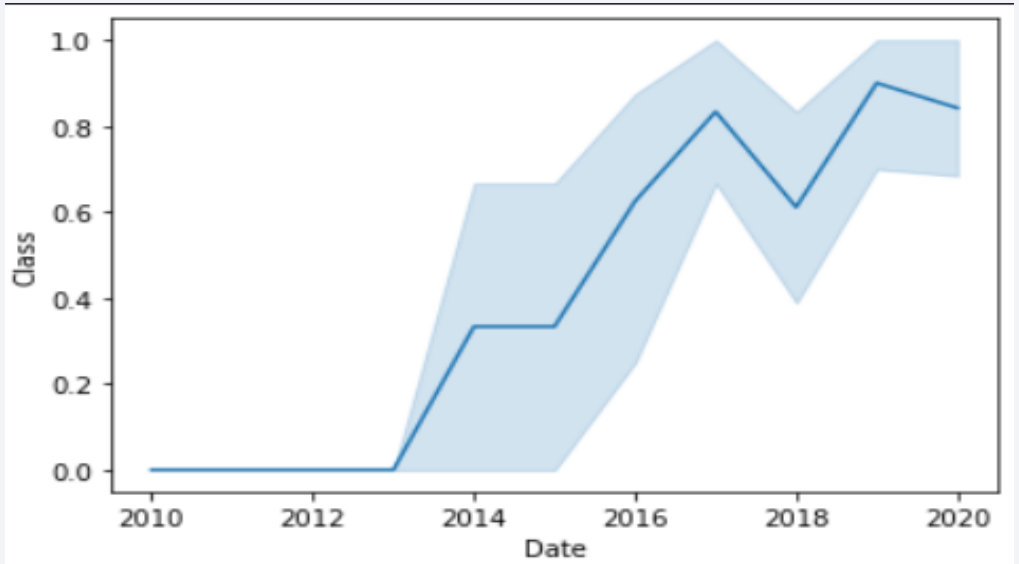
- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/collectingTheData-webscraping.ipynb

# Data Wrangling

- Calculated the number of launches on each site
- Calculated the number and occurrence of each orbit
- Calculated the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/Data%20wrangling.ipynb

# EDA with Data Visualization



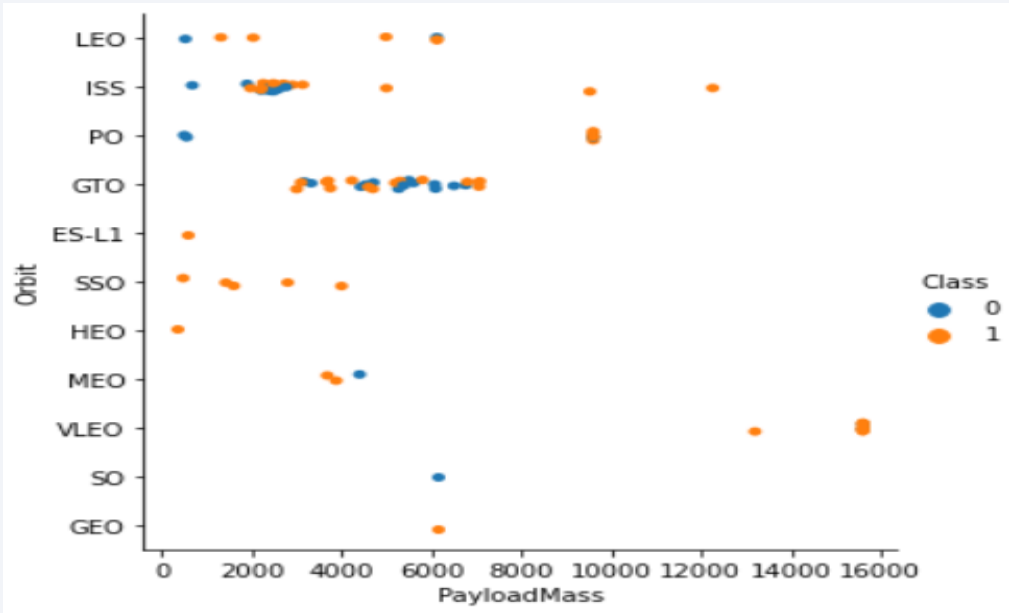

Visualized the relation between the date and the average success rate across the years.

**Conclusion:** the success rate since 2013 kept increasing till 2020

- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/eda-dataviz.ipynb

Visualized the relation between the Orbit and the payload mass.

**Conclusion:** With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.



# EDA with SQL

- Loaded the SpaceX dataset into a sqllite database.
- We applied EDA with SQL to get insight from the data. We wrote queries to find out for instance:
  - 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.
- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/eda-sql.ipynb

# Build an Interactive Map with Folium

- Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.
- Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- Using the color-labeled marker clusters, we identified which launch sites have relatively high success rate.
- We calculated the distances between a launch site to its proximities. We answered some question for instance:
  - Are launch sites near railways, highways and coastlines.
  - Do launch sites keep certain distance away from cities.

- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/site_location.ipynb

# Build a Dashboard with Plotly Dash

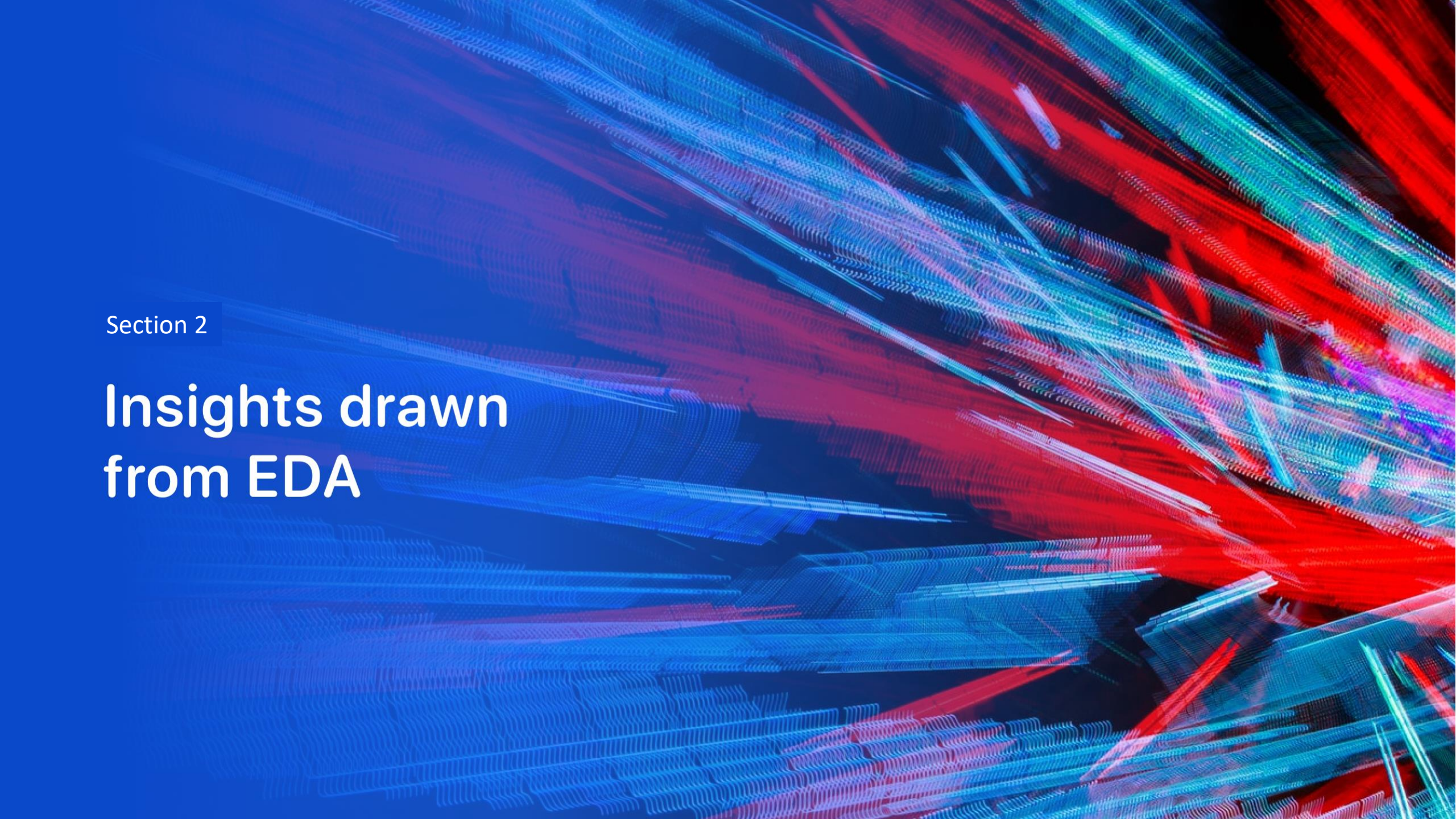
- Built an interactive dashboard with Plotly dash
- Plotted pie charts showing the total launches by a certain sites
- Plotted scatter graph showing the relationship with Outcome and Payload Mass (Kg) for the different booster version.
- The GitHub URL:

# Predictive Analysis (Classification)

- Loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- Built different machine learning models and tune different hyperparameters using GridSearchCV.
- Used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- Found the best performing classification model.
- The GitHub URL: https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

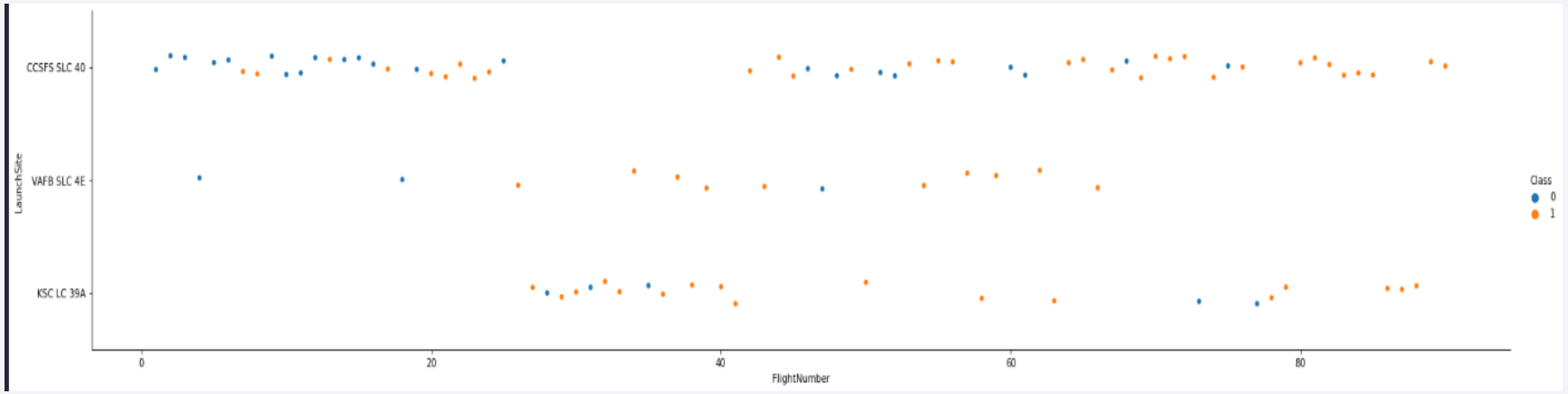
# Results

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


Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site

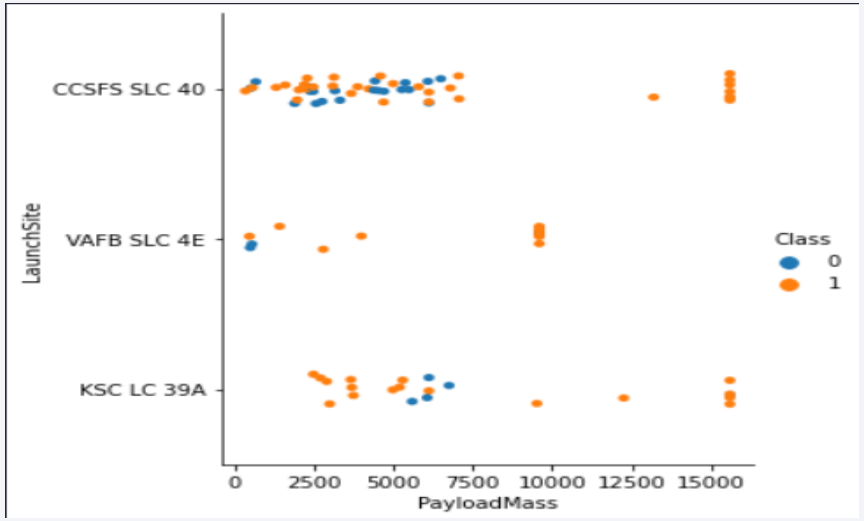
- Show a scatter plot of Flight Number vs. Launch Site




- **Explanation:** when the flight number is > 80 the success rate is 100%

# Payload vs. Launch Site

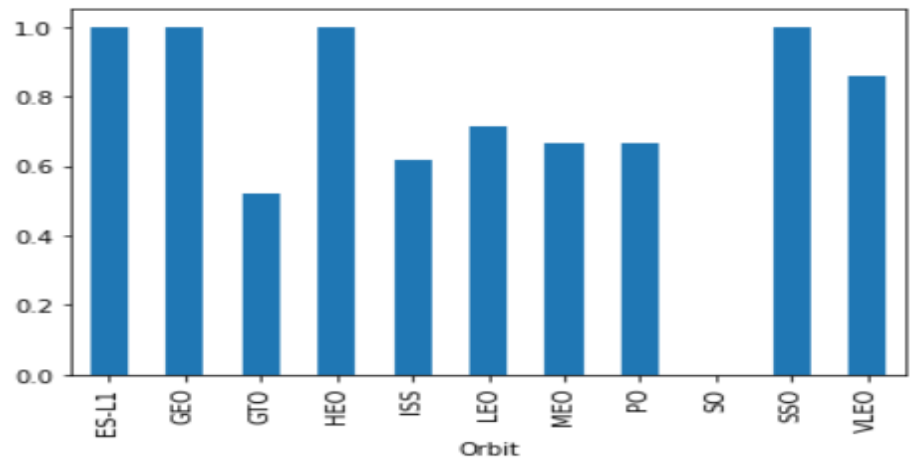
- Show a scatter plot of Payload vs. Launch Site




- **Explanation:** the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

# Success Rate vs. Orbit Type

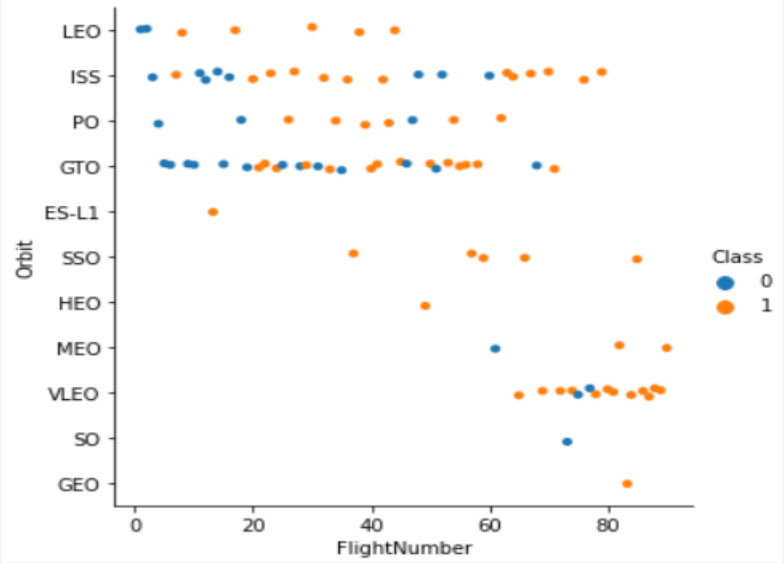
- Show a bar chart for the success rate of each orbit type




- **Explanation**: The ES-L1, GEO, HEO  and SSO orbit have the maximum success rate.

# Flight Number vs. Orbit Type

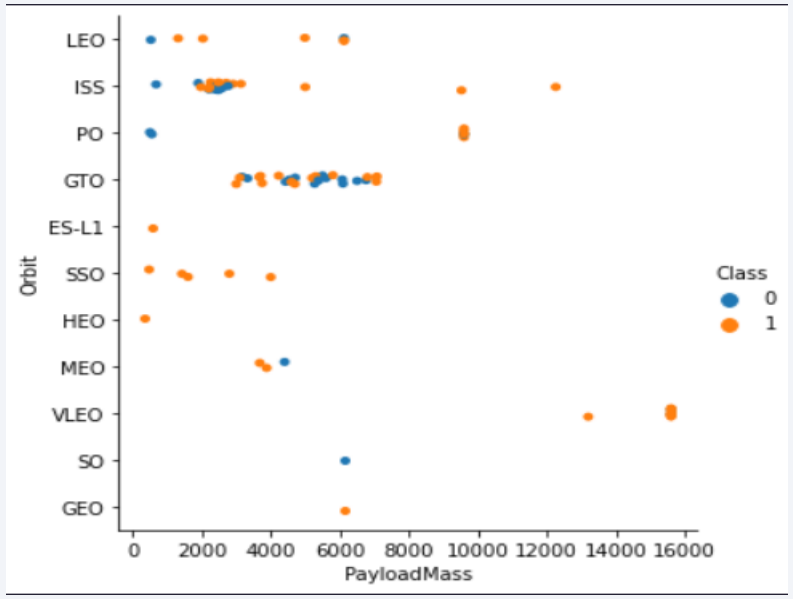
- Show a scatter point of Flight number vs. Orbit type




- **Explanation:** 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

- Show a scatter point of payload vs. orbit type

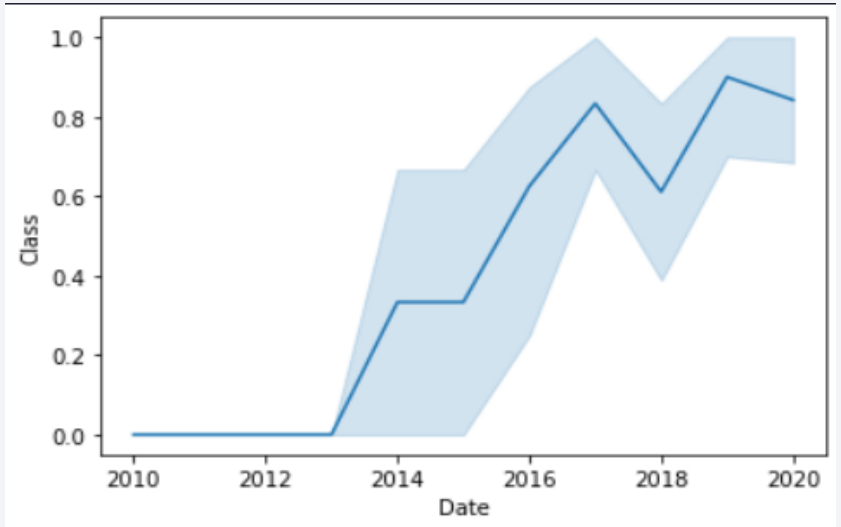



- **Explanation:** With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

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

- Show a line chart of yearly average success rate




- **Explanation:** the sucess rate since 2013 kept increasing till 2020

# All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEX_DATA
```

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

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

# Launch Site Names Begin with 'CCA'

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

```
%sql SELECT * FROM SPACEX_DATA WHERE Launch_Site LIKE 'CCA%'LIMIT(5)
```

[12] Python

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

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 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) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass FROM SPACEX_DATA WHERE Customer == 'NASA (CRS)'
```

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

| Total_Payload_Mass |
| --- |
| 45596 |

# Average Payload Mass by F9 v1.1

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

```
[21] %sql SELECT AVG(PAYLOAD_MASS__KG_) as Avg_Payload_Mass FROM SPACEX_DATA WHERE Booster_Version ='F9 v1.1'
```

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

| Avg_Payload_Mass |
|---|
| 2928.4 |

# First Successful Ground Landing Date

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

```
%sql SELECT MIN(Date) AS FirstSuccessfull_landing_date FROM SPACEX_DATA WHERE [Landing _Outcome] == 'Success (ground pad)'
```

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

| FirstSuccessfull_landing_date |
| --- |
| 01-05-2017 |

# Successful Drone Ship Landing with Payload between 4000 and 6000

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 Booster_Version FROM SPACEX_DATA WHERE [Landing _Outcome] = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000  AND PAYLOA
```

Python

```
* sqlite:///SpaceX.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

List the total number of successful and failure mission outcomes

```
%sql SELECT (SELECT COUNT(*) FROM SPACEX_DATA WHERE [Landing _Outcome] LIKE 'Success%') as Success_outcomes, (SELECT COUNT(*) FROM SPA(
```

Python

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

| Success_outcomes | Failure_outcomes |
|---|---|
| 61 | 10 |

# Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEX_DATA WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX_D
```

Python

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

| Booster_Version | PAYLOAD_MASS__KG_ |
|---|---|
| F9 B5 B1048.4 | 15600 |
| F9 B5 B1049.4 | 15600 |
| F9 B5 B1051.3 | 15600 |
| F9 B5 B1056.4 | 15600 |
| F9 B5 B1048.5 | 15600 |
| F9 B5 B1051.4 | 15600 |
| F9 B5 B1049.5 | 15600 |
| F9 B5 B1060.2 | 15600 |
| F9 B5 B1058.3 | 15600 |
| F9 B5 B1051.6 | 15600 |
| F9 B5 B1060.3 | 15600 |
| F9 B5 B1049.7 | 15600 |

# 2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

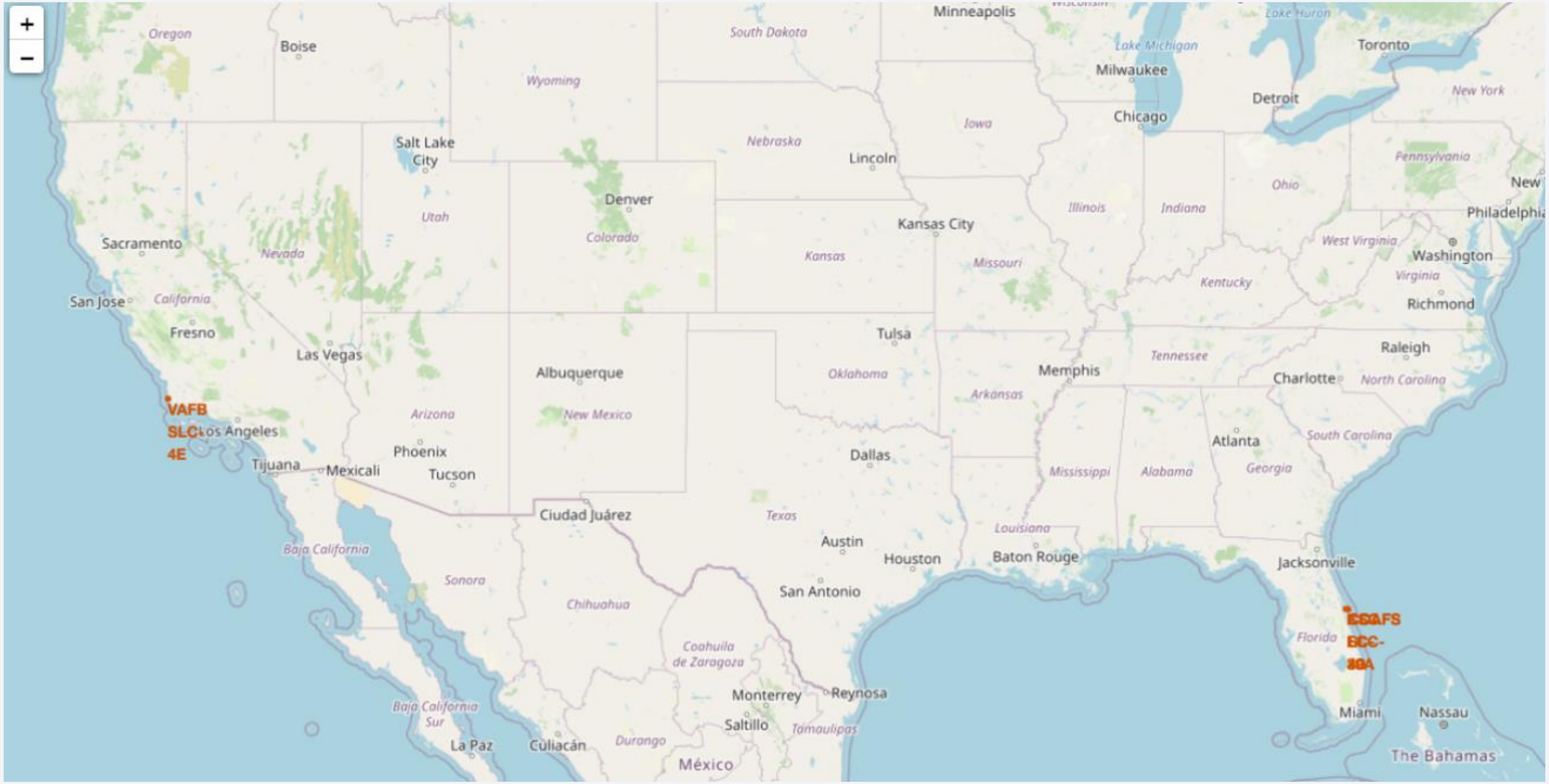
# Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- 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

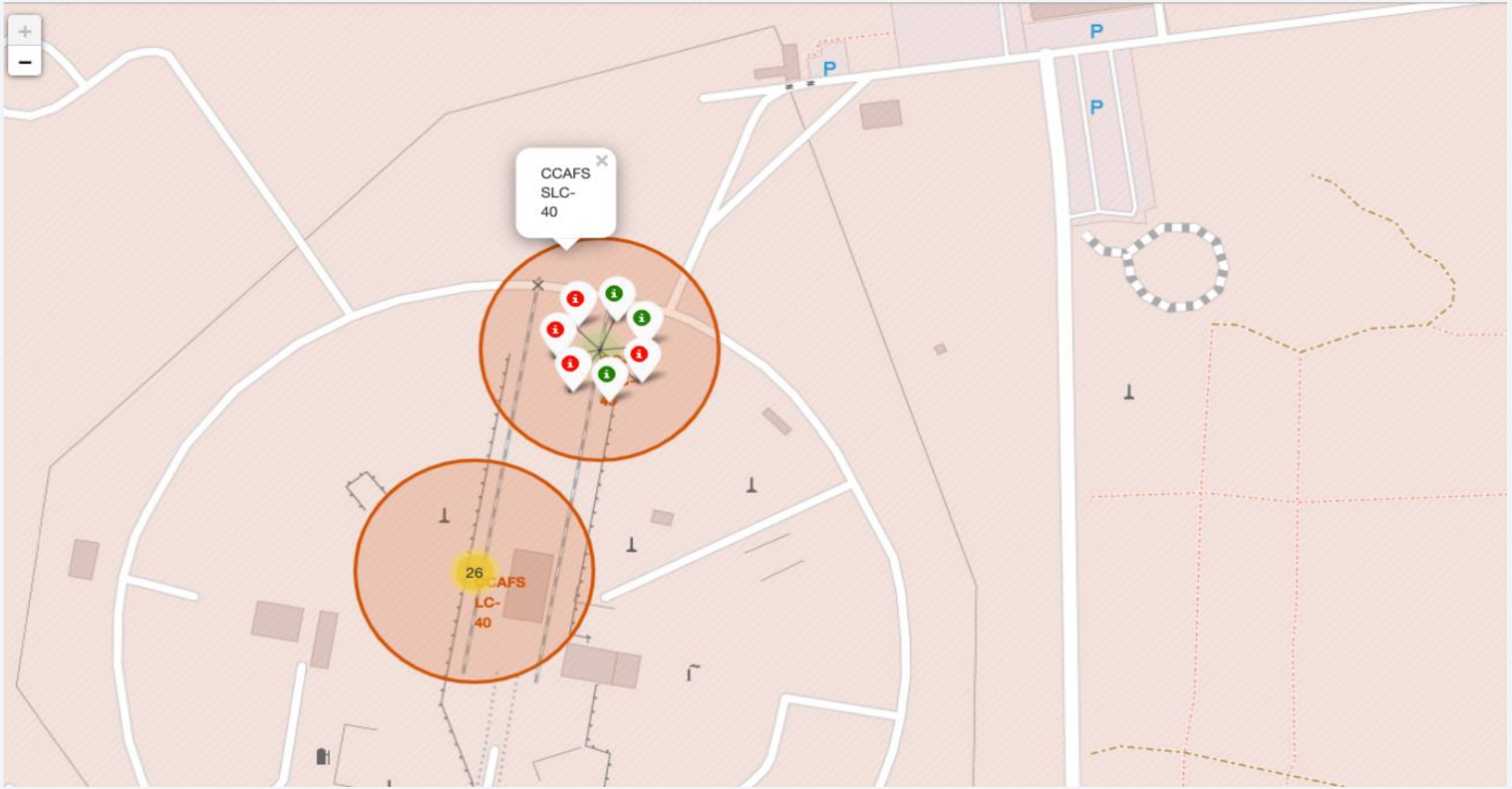
Section 3

# Launch Sites Proximities Analysis

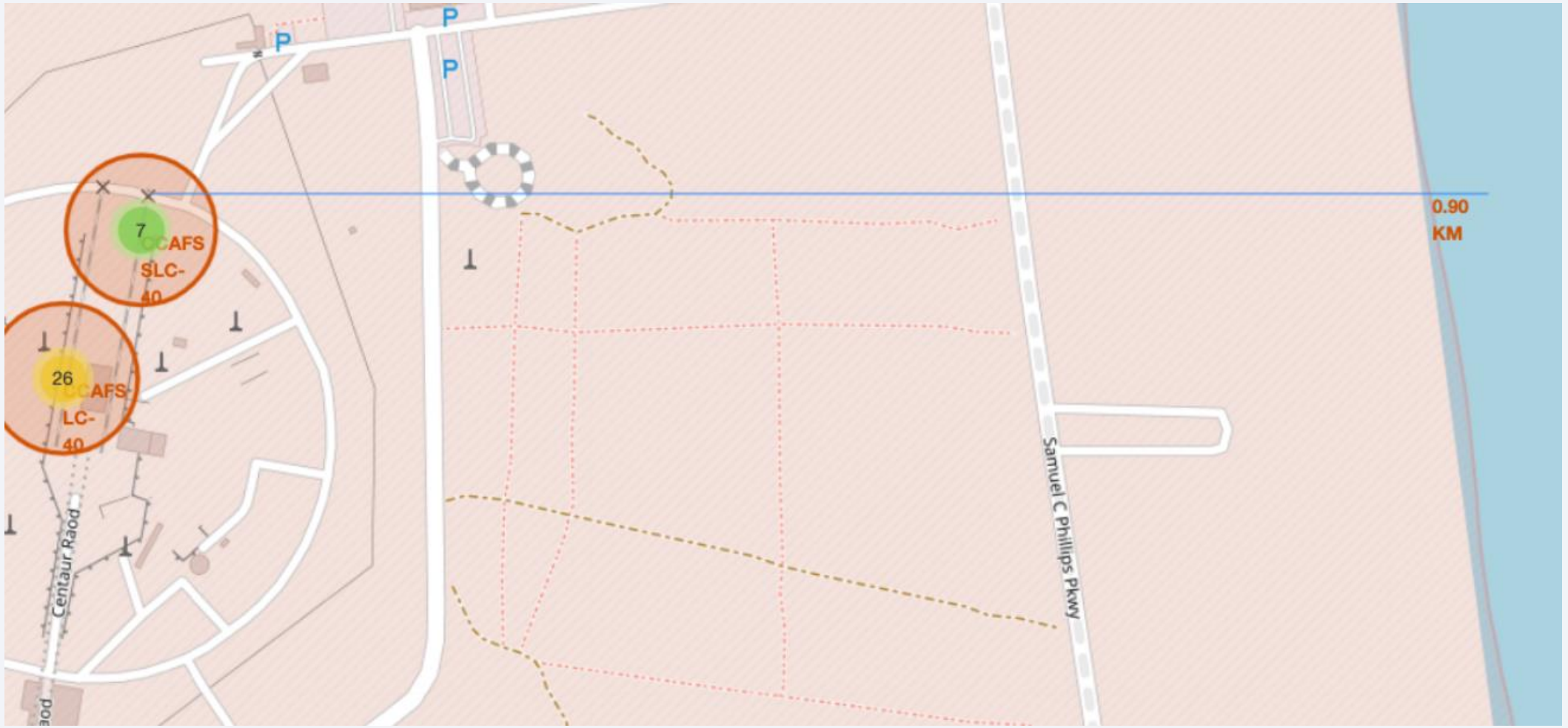
# All launch sites global map markers

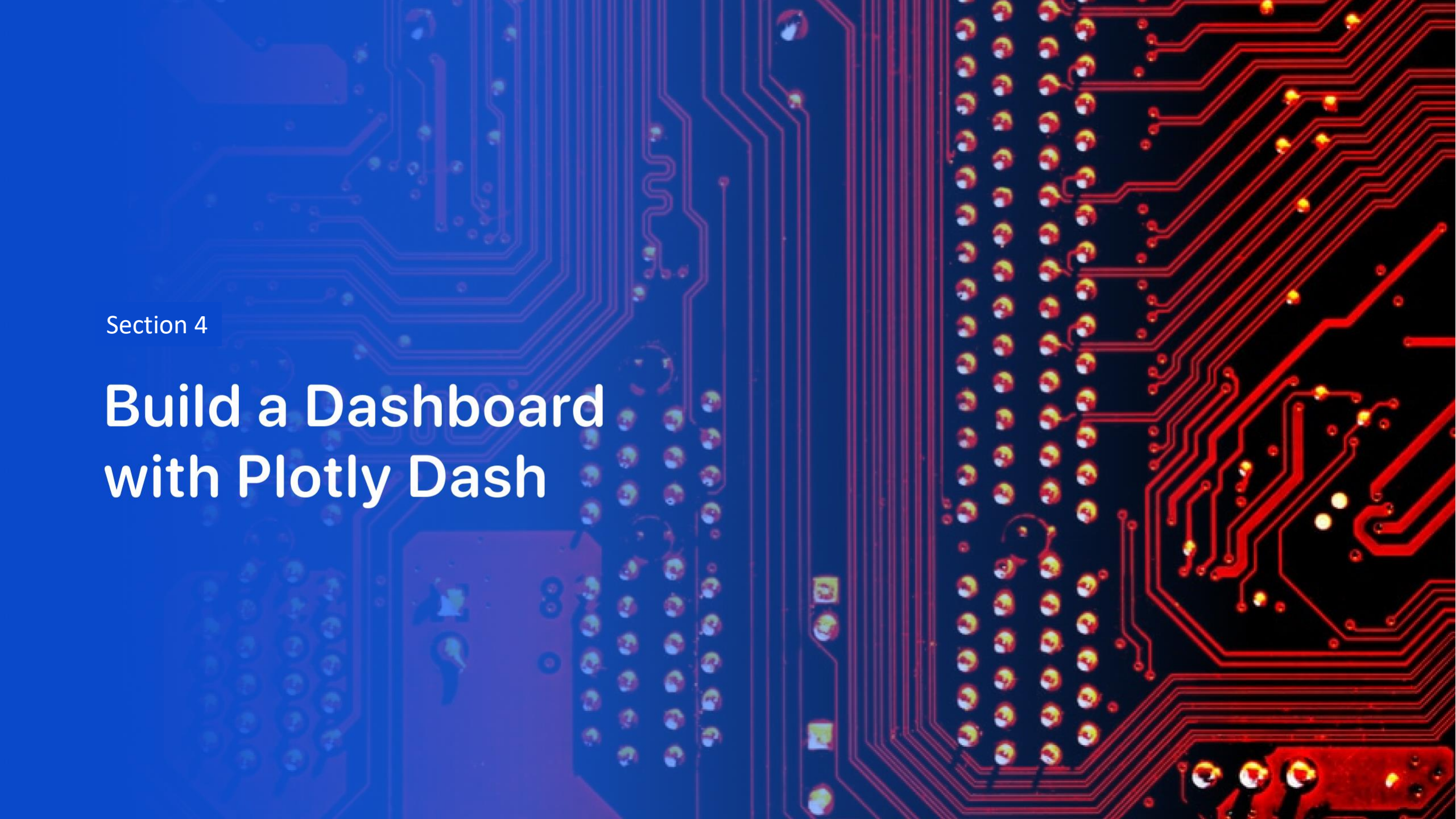


# Markers showing launch sites with color labels

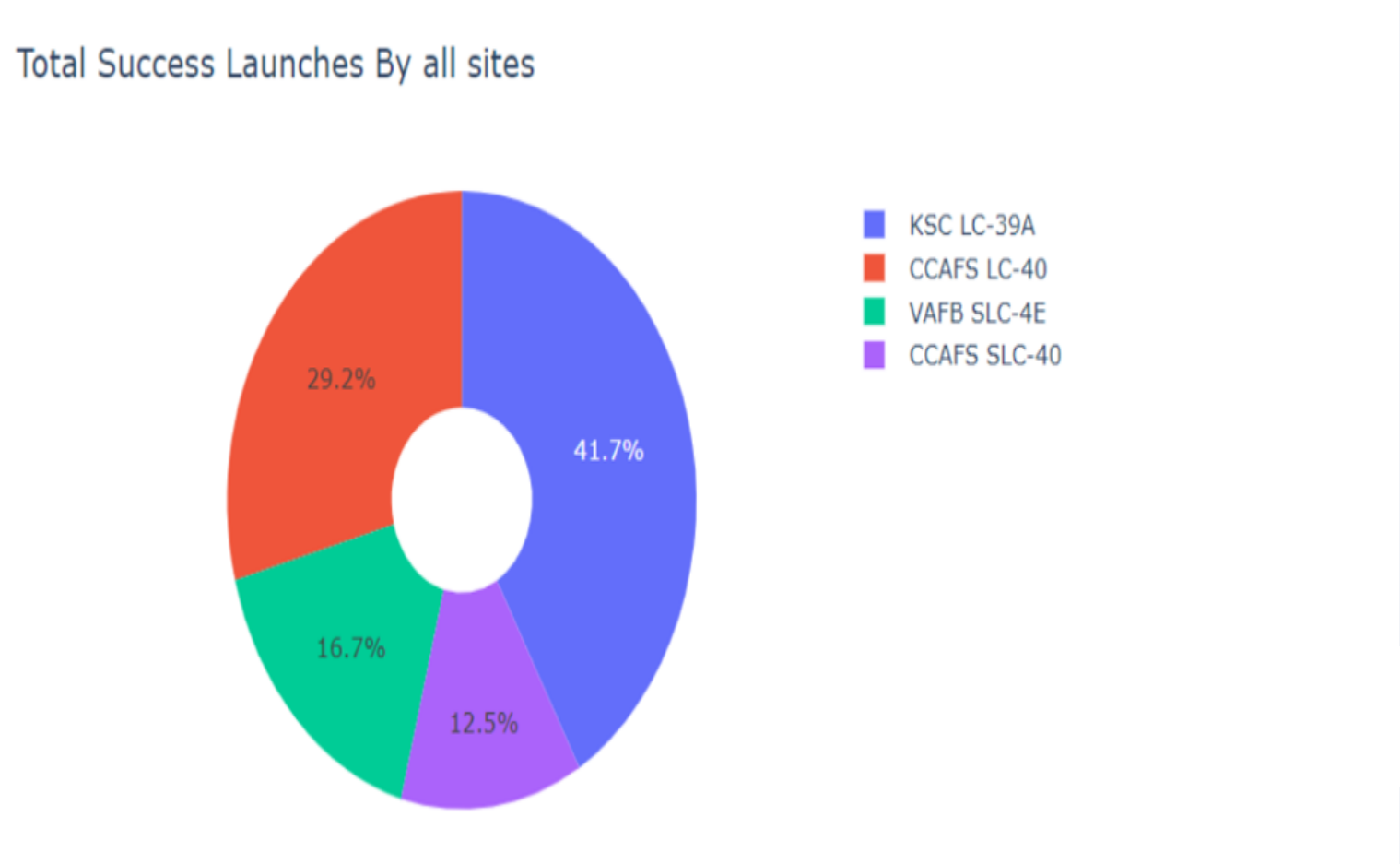


# Launch Site distance to landmarks




Section 4
Build a Dashboard with Plotly Dash

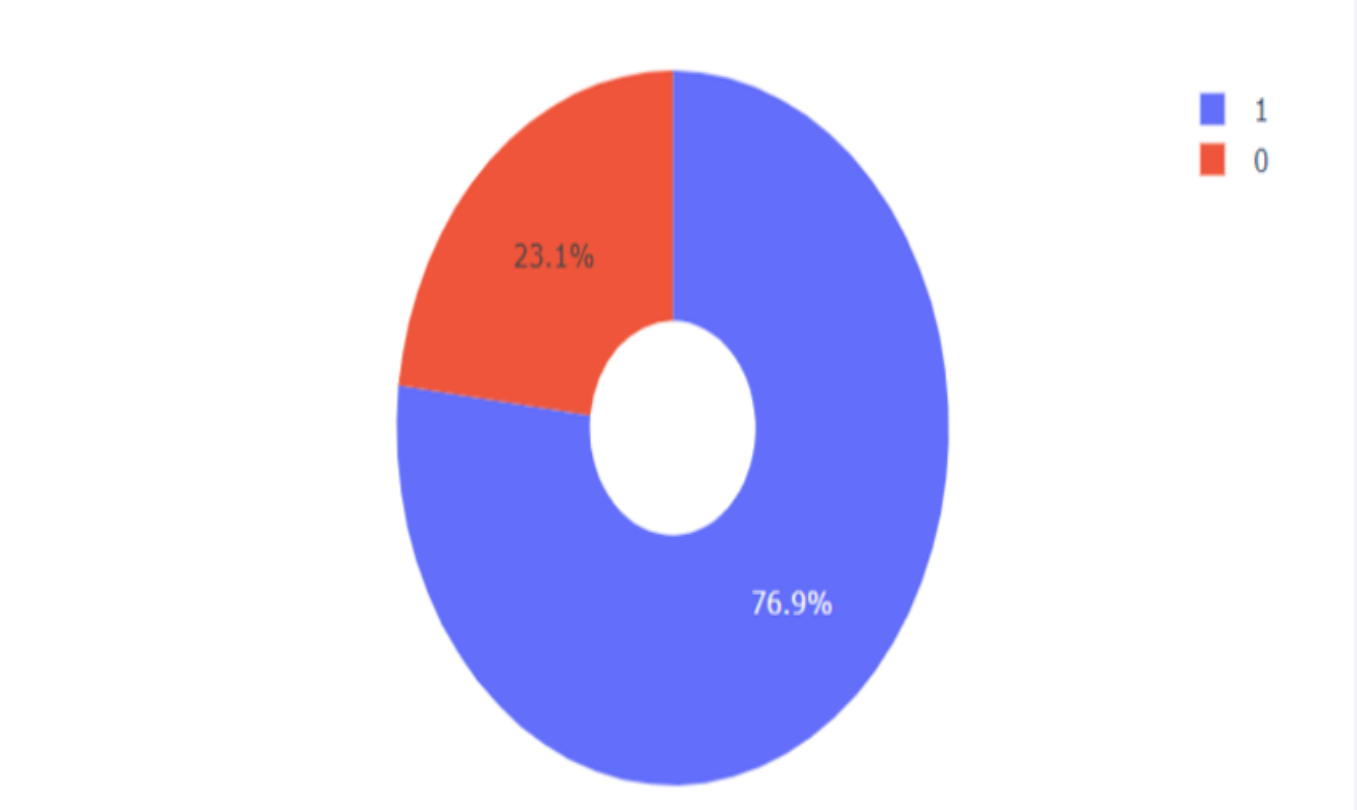
# Pie chart showing the success percentage achieved by each launch site



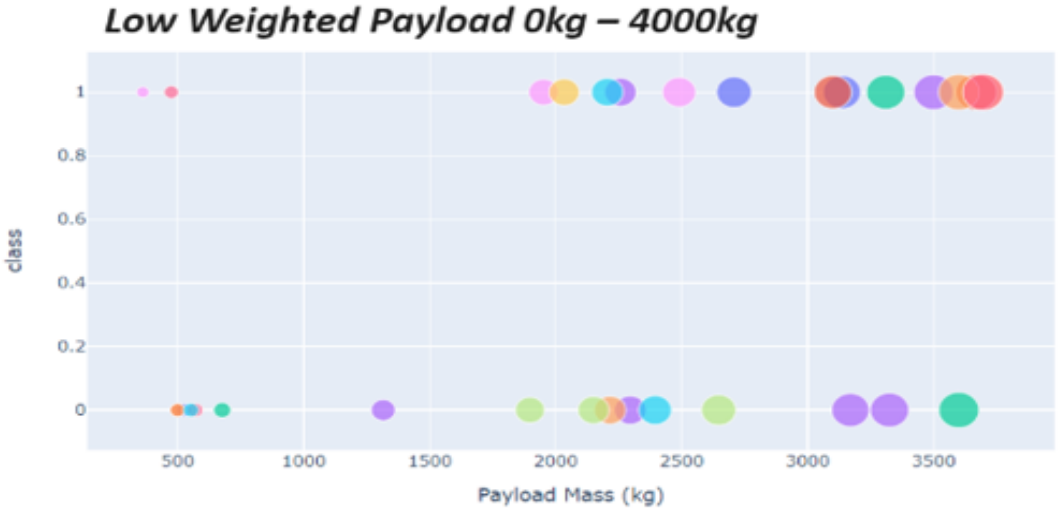

We conclude that the KSC LC-39A have the most successful launches

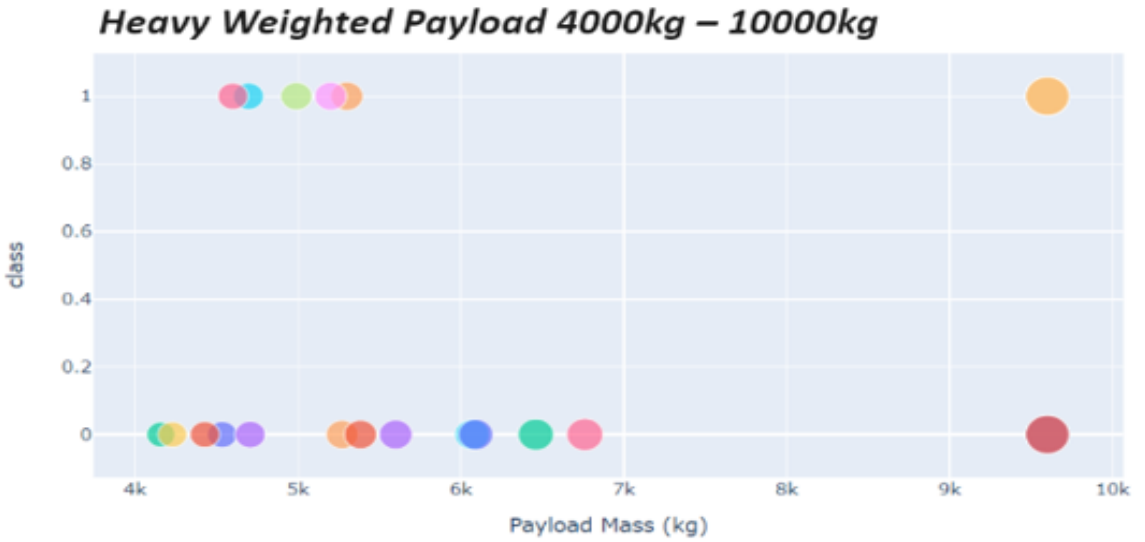
# Pie chart showing the Launch site with the highest launch success ratio

the KSC LC-39A success is 76.9%



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



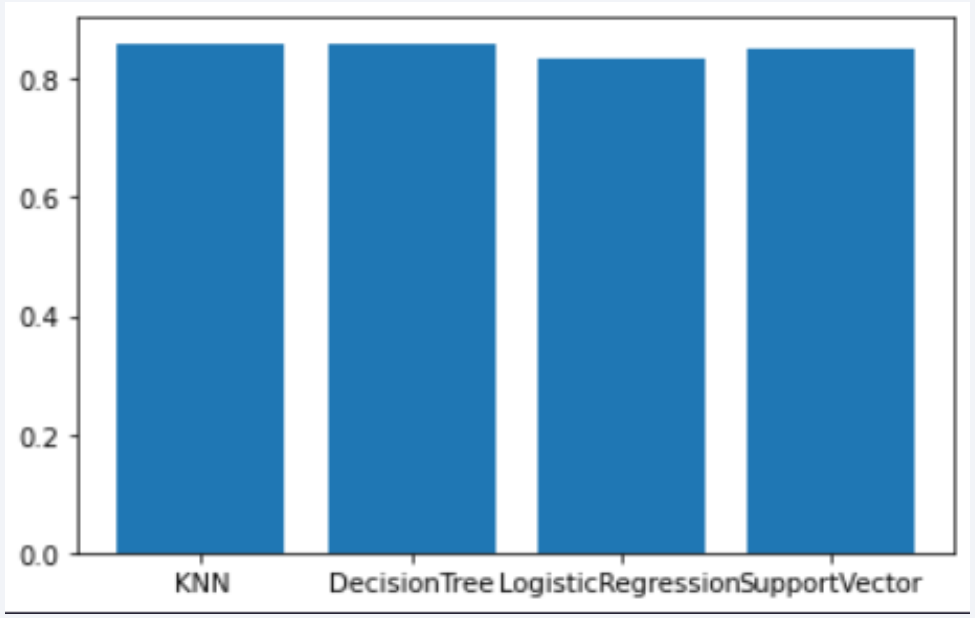

The success rate of low weighted payload is heigher than high weighted payload

Section 5

# Predictive Analysis (Classification)

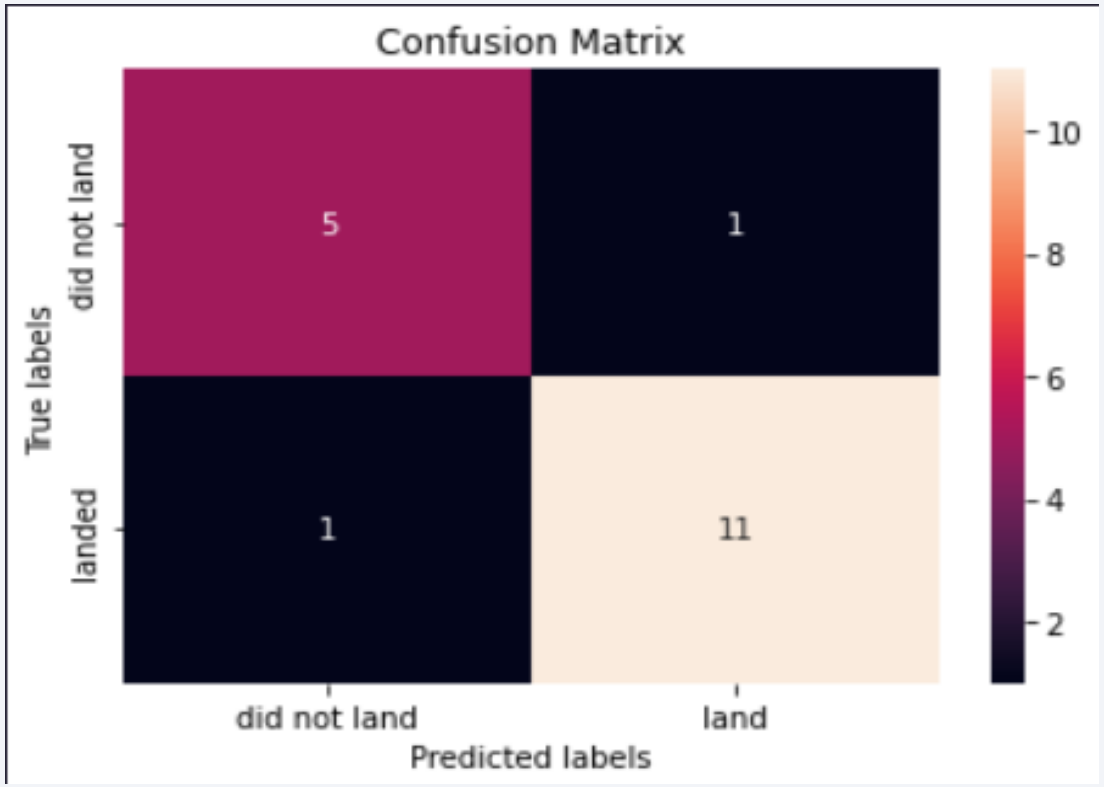
# Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart




- Find which model has the highest classification accuracy: KNN

# Confusion Matrix




The knn model classify well the True positive and True negative and the percentage of False positive and false negative is too small.

# Conclusions

- 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 KNNclassifier is the best machine learning algorithm for this task.

# Appendix

Data Collected Links:

- https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/dataset_part_1.csv
- https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/dataset_part_2.csv
- https://github.com/miraehab/IBM-Applied-Data-Science-Capstone/blob/main/dataset_part_3.csv

# Thank you!