IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Anatol Krasowski
26.07.2022

# Outline

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

# Executive Summary

- **Summary of methodologies**
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Visual Analysis with Folium
  - Machine Learning Prediction Analysis
- **Summary of all results**
  - Exploratory Data results
  - Predictive Analytics results

# 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 methodology:
  - SpaceX Rest API
  - Web Crapping (Wikipedia)
- Perform data wrangling
  - 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
  - How to build, tune, evaluate classification models

# Data Collection

- Data was gathered from SpaceX REST API and decoded (json).

- Web scraping from Wikipedia for Falcon 9 launch records with BeautifulSoup (extracted data from HTML-table).

- Checking and cleaning Data with Pandas necessary

# Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.
- The link to the notebook:

https://github.com/AnatoIx/Applied-Data-Science-Capstone-IBM/blob/main/Data%20Collection%20API.ipynb

- Get request

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

- Convert results to dataframe

```
response.json()
data = pd.json_normalize(response.json())
```

- Cleaning data and delete missing values

# Data Collection - Scraping

- Web Scrapping with BeautifulSoup from Wikipedia
- Parsing the html-table and converting it into a pandas dataframe
- The link to the notebook:

https://github.com/AnatoIx/Applied-Data-Science-Capstone-IBM/blob/main/Data%20Collection%20Web%20Scrapping.ipynb

- Get html table as BeautifulSoup object

```
html = requests.get(static_url)
soup = BeautifulSoup(html.content,  'html5lib')
```

- Extract column names and data

```
column_names = []
for row in first_launch_table.find_all('th'):
    col = extract_column_from_header(row)
    print(col)
    if col is not None and len(col)>0:
        column_names.append(extract_column_from_header(row))
```

- Cleaning data and delete missing values

# Data Wrangling

- **Exploratory Data Analysis to find patterns in the data**
- Calculating the number of launches on each site
- Calculating the number and occurrence of each orbit
- Calculating the number and occurence of mission outcome per orbit type
- Creating a landing outcome label from Outcome column

| Orbit | LaunchSite | Outcome | Flights | GridFins | Reused | Legs | LandingPad | Block | ReusedCount | Serial | Longitude | Latitude | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0003 | -80.577366 | 28.561857 | 0 |
| LEO | CCAFS SLC 40 | None None | 1 | False | False | False | NaN | 1.0 | 0 | B0005 | -80.577366 | 28.561857 | 0 |

- Link to Github: https://github.com/Anatolx/Applied-Data-Science-Capstone-IBM/blob/main/Data%20Wrangling.ipynb

# EDA with Data Visualization

- **Visualization to find patterns in the data**
- Flight Number VS. Payload Mass
- Flight Number VS. Launch Site
- Payload VS. Launch Site
- Orbit VS. Flight Number
- Payload VS. Orbit Type
- Orbit VS. Payload Mass
- Mean VS. Orbit
- Success Rate VS. Year

- Link to Github: https://github.com/AnatoIx/Applied-Data-Science-Capstone-IBM/blob/main/EDA%20Visualisation.ipynb

# EDA with SQL

## Using SQL to get important data from dataset:

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'KSC’
- The total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- The date where the successful landing outcome in drone ship was achieved.
- The names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass.
- The records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- The count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Link to Github: https://github.com/Anatolx/Applied-Data-Science-Capstone-IBM/blob/main/EDA%20SQL.ipynb

# Build an Interactive Map with Folium

- All launch sites was added as map objects
- Markers, circles, lines etc. was used 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.
- Calculating of the distances between a launch site to its proximities to ask some questions

Link to Github: https://github.com/AnatolX/Applied-Data-Science-Capstone-IBM/blob/main/Folium_Map.ipynb

# Build a Dashboard with Plotly Dash

- Interactive dashboard with Plotly Dash to visualize the data
- Pie charts showing the total launches by a certain sites
- Scatter graph showing the relationship with Outcome and Payload
- Mass (Kg) for the different booster version.

- Link to Github: https://github.com/Anatolx/Applied-Data-Science-Capstone-IBM/blob/main/Dashboard_Capstone.txt

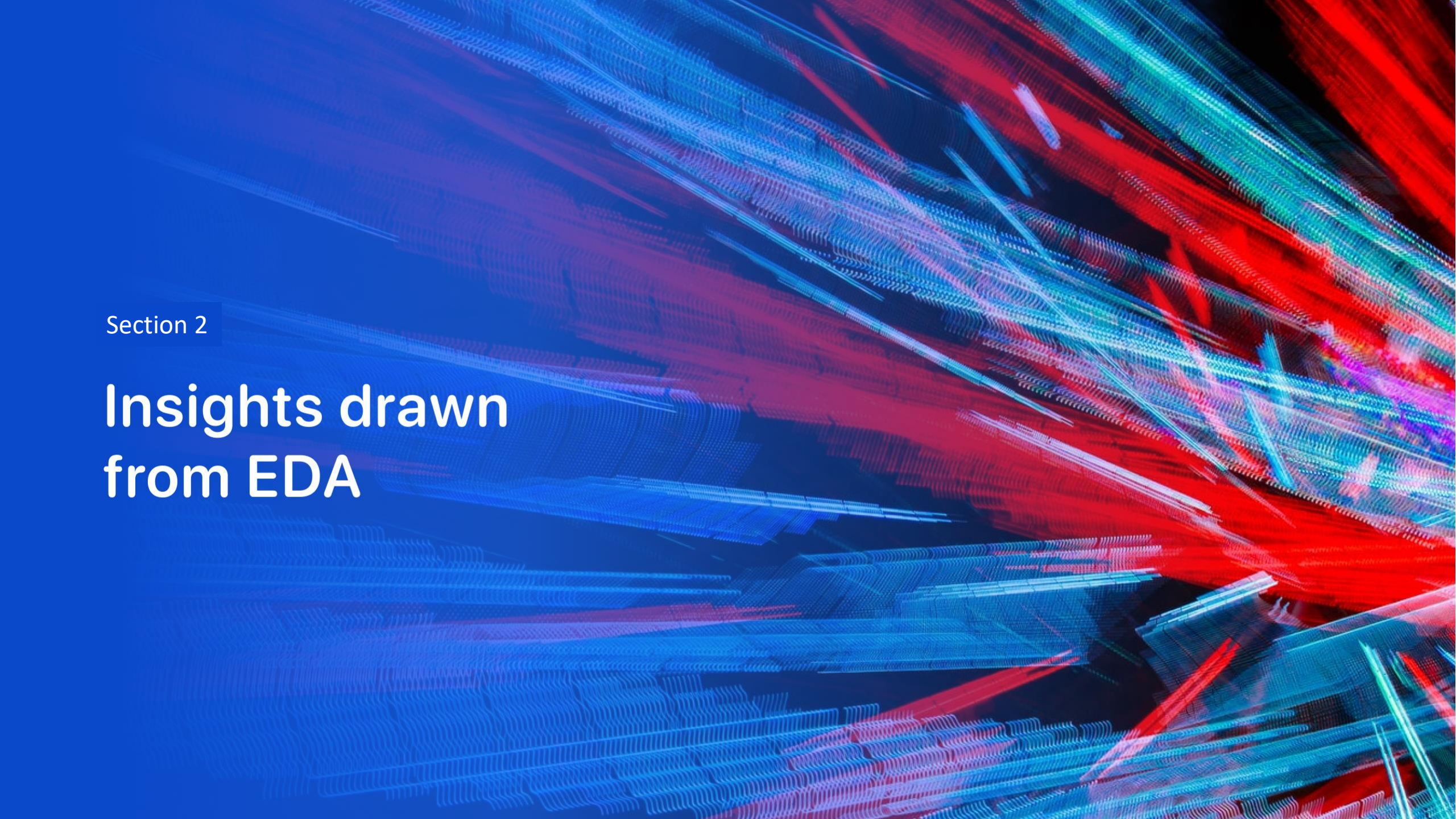
# Predictive Analysis (Classification)

- Transforming data to into training and testing data sets to improve a accurancy.
- Different machine learning models with best hyperparameters (GridSearchCV)
- Accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- The best performing classification model was found.

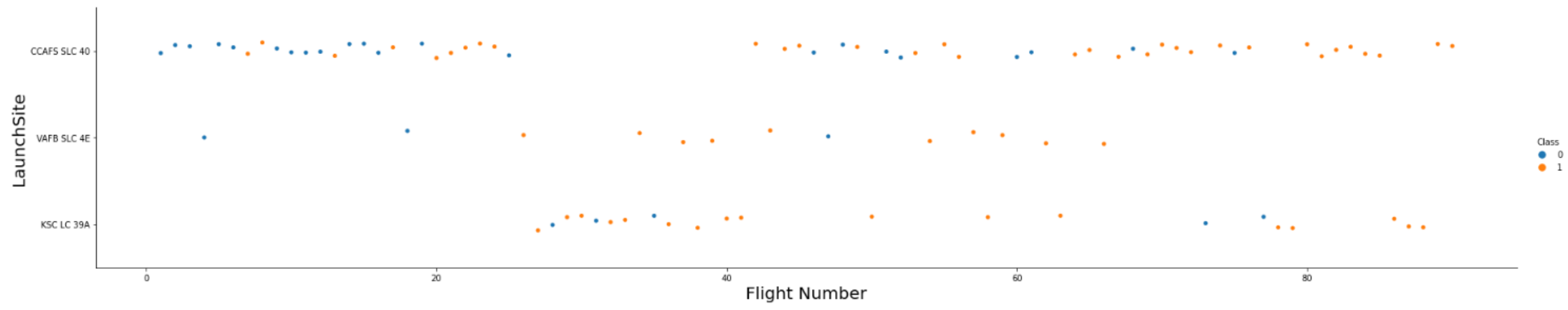
- Link to Github: https://github.com/Anatolx/Applied-Data-Science-Capstone-IBM/blob/main/Predictive_Analyse.ipynb

# Results

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

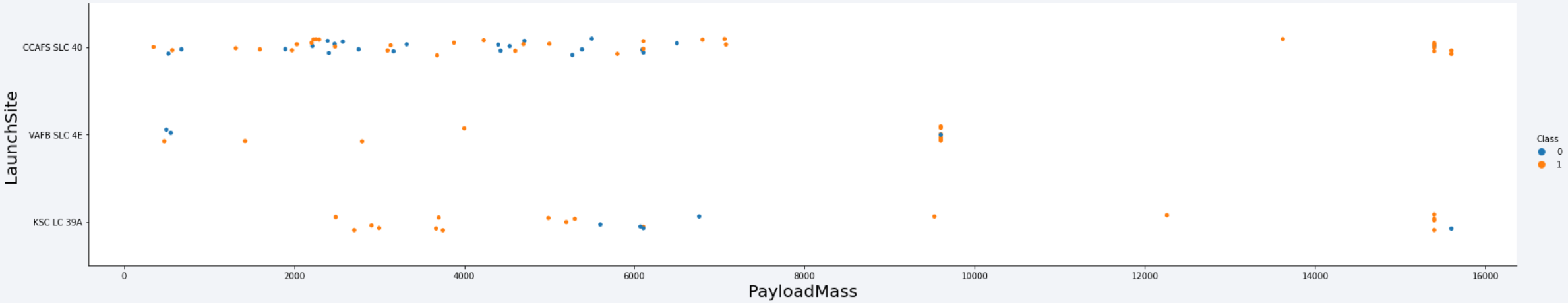

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site




→ The larger the flight amount at a launch site, the greater the success rate at a launch site.

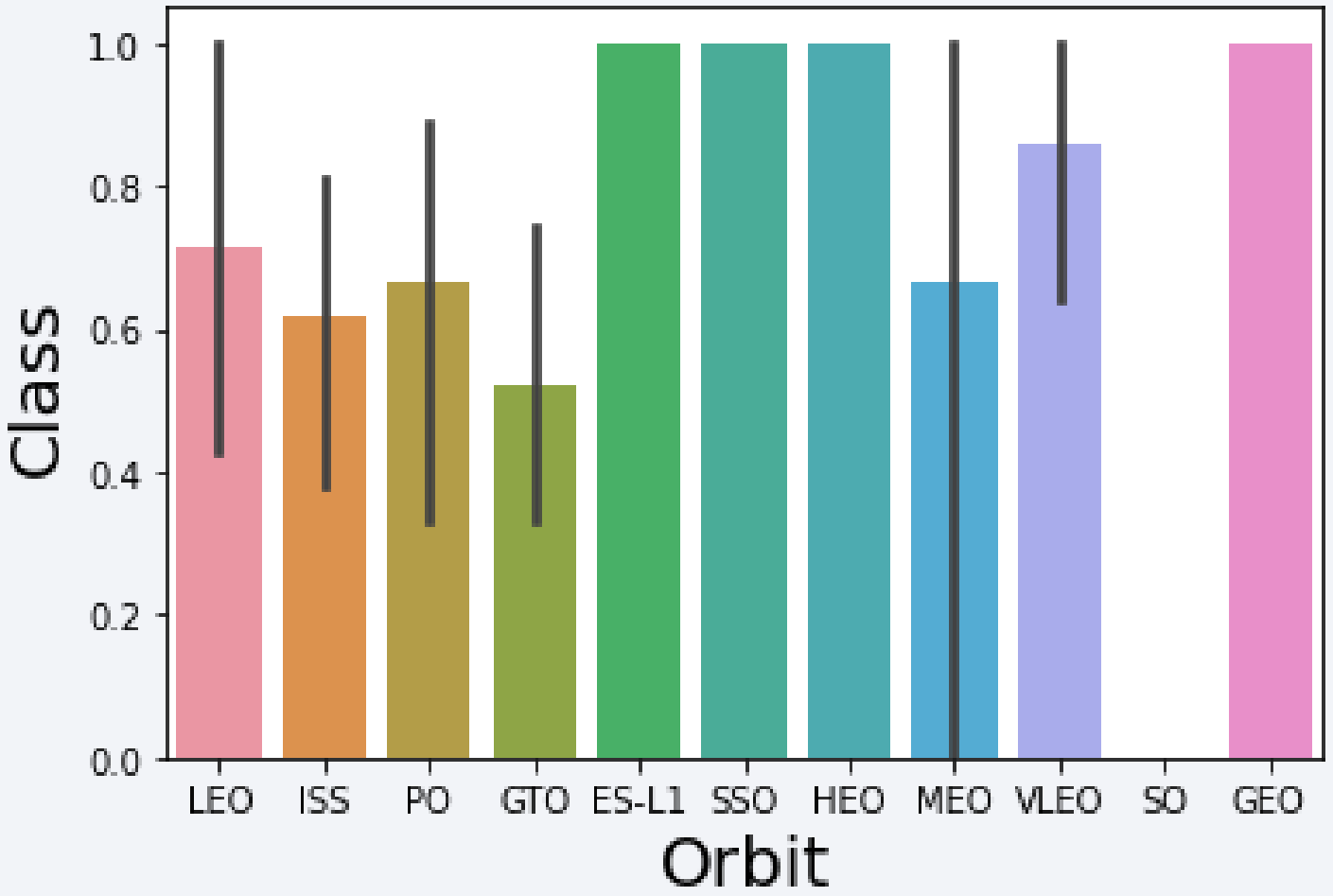
# Payload vs. Launch Site




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

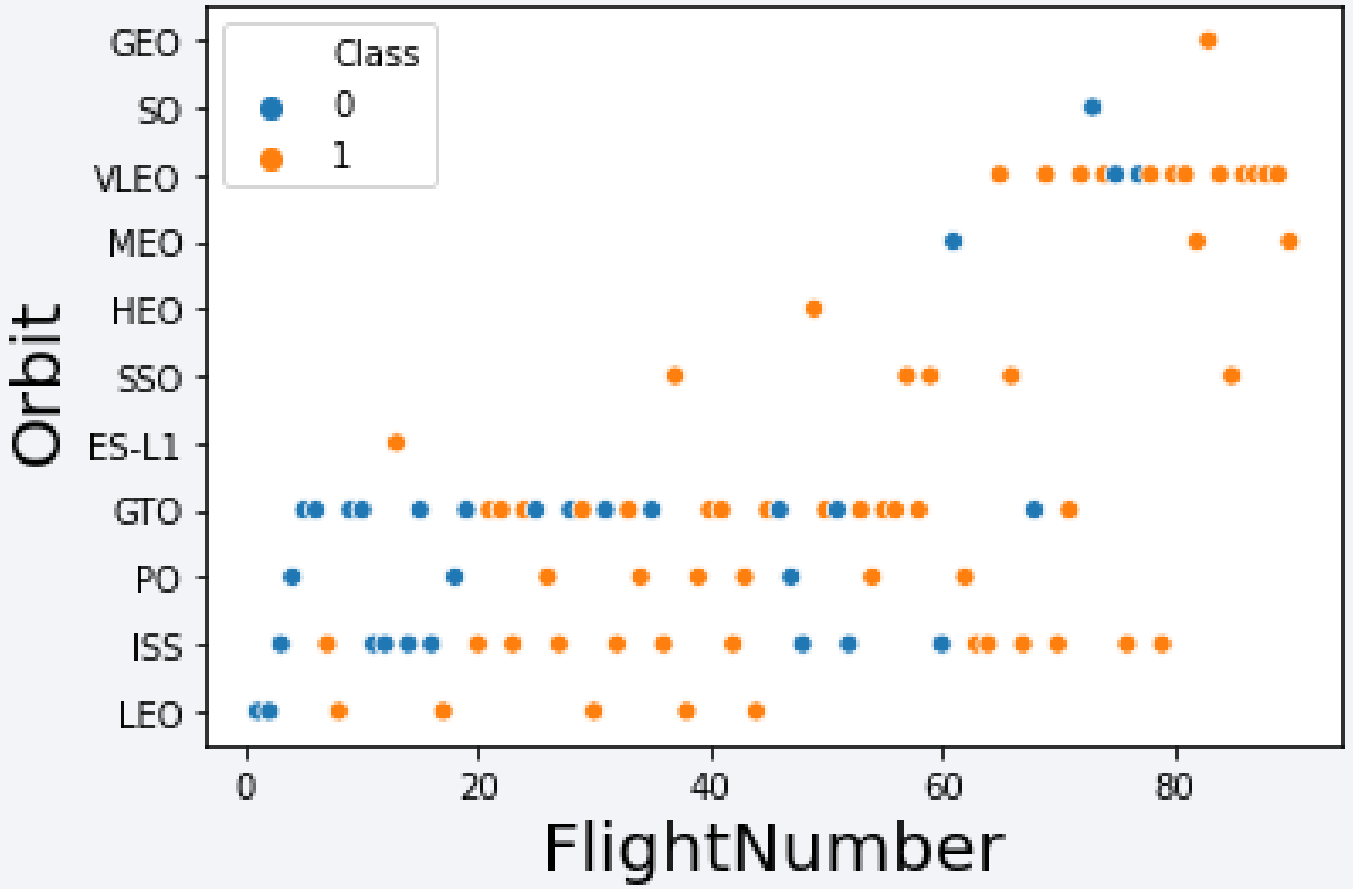
→ launches with big payload mass are more often successful

# Success Rate vs. Orbit Type



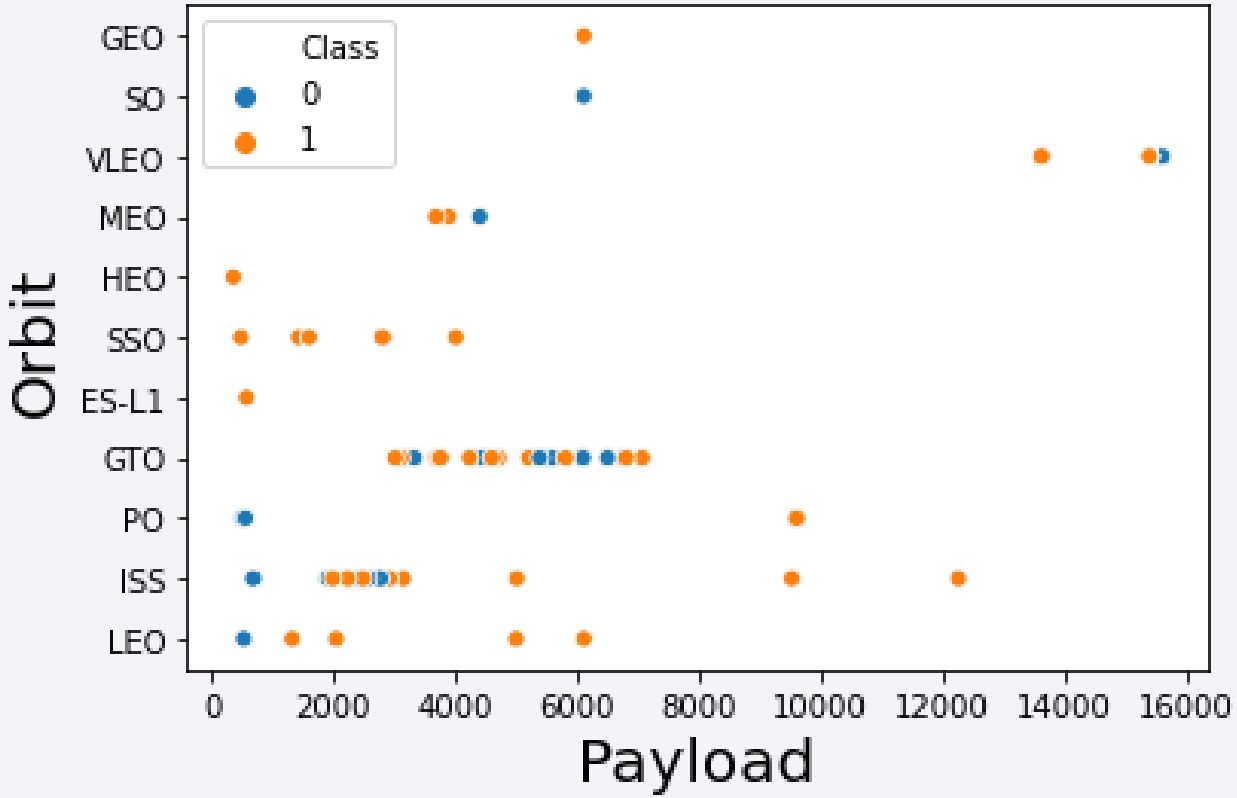

→ Orbits GEO,HEO,SSO,ES-L1 have the best Success Rate

# Flight Number vs. Orbit Type




→ In the LEO orbit the Success appears related to the number of flights
→ there seems to be no relationship between flight number when in GTO orbit.

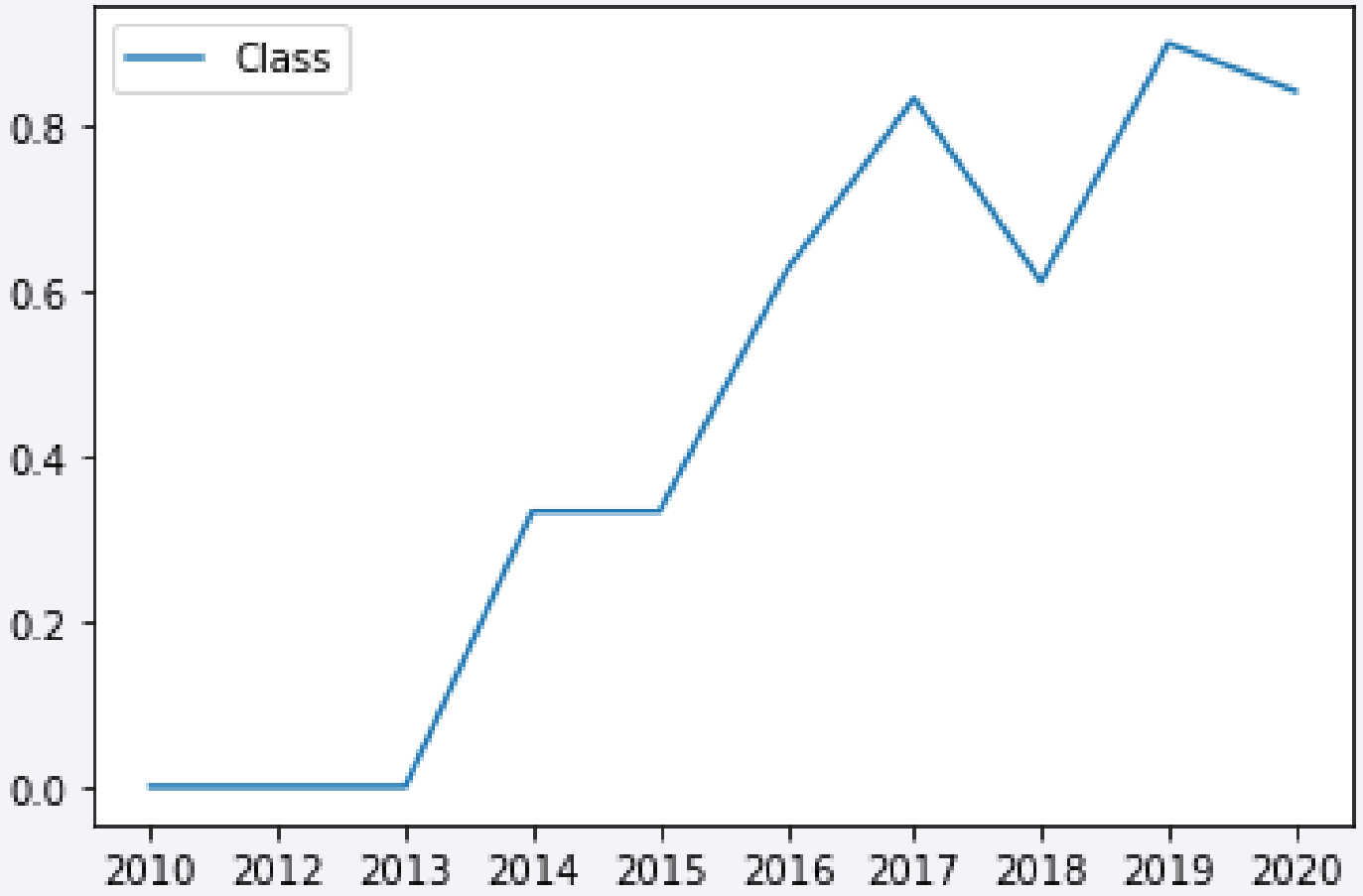
# Payload vs. Orbit Type




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

→ 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 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 |

→ Keyword DISTINCT to show only unique names
→ There are 4 launch sites

# Total Payload Mass

```
%sql select sum(payload_mass__kg_) 'total payload mass' from SPACEXTBL where CUSTOMER='NASA (CRS)';
```

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

| total payload mass |
| --- |
| 45596 |

→ The total payload carried by boosters from NASA is 45596 kg

# Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) 'average payload mass' from SPACEXTBL where booster_version='F9 v1.1';
```

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

| average payload mass |
| --- |
| 2928.4 |

→ The average payload mass carried by boosters with version F9 v1.1 is 2928.4

# First Successful Ground Landing Date

```
%sql select date from SPACEXTBL where date = (select min(date) from SPACEXTBL where [Landing _Outcome]='Success
```

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

| Date |
| --- |
| 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_
```

```
 * 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 [Landing _Outcome], count([Landing _Outcome]) from SPACEXTBL group by [Landing _Outcome] having [Lar
```

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

| Landing _Outcome | count([Landing _Outcome]) |
|---|---|
| Failure | 3 |
| Failure (drone ship) | 5 |
| Failure (parachute) | 2 |
| Success | 38 |
| Success (drone ship) | 14 |
| Success (ground pad) | 9 |

# 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 |

# 2015 Launch Records

```
%sql select substr(Date, 4, 2) month, count([Landing _Outcome]) failure, booster_version, launch_site from SPACEX
```

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

| month | failure | Booster_Version | Launch_Site |
|---|---|---|---|
| 01 | 1 | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | 1 | F9 v1.1 B1015 | CCAFS LC-40 |

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

```
%sql select substr(Date, 4, 2) month, substr(Date,7,4) year, count([Landing _Outcome]) susses from SPACEXTBL wher
```

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

| month | year | susses |
|---|---|---|
| 10 | 2017 | 5 |
| 08 | 2016 | 5 |
| 12 | 2017 | 4 |
| 11 | 2018 | 4 |
| 01 | 2017 | 4 |
| 06 | 2019 | 3 |
| 09 | 2017 | 2 |
| 05 | 2016 | 2 |
| 04 | 2016 | 2 |
| 07 | 2016 | 1 |
| 03 | 2020 | 1 |
| 02 | 2017 | 1 |

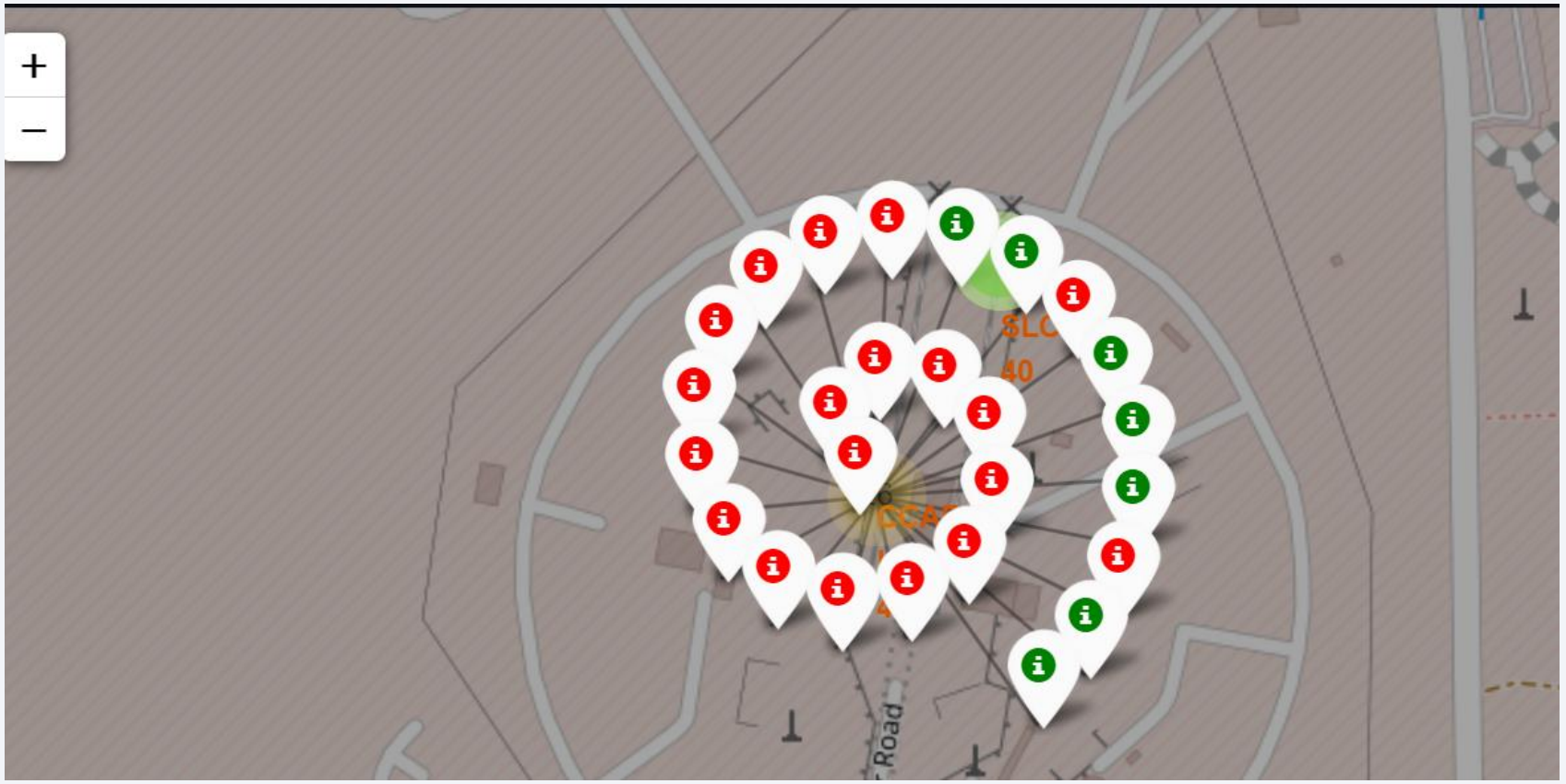
Section 3

# Launch Sites Proximities Analysis

# All launch sites on global map

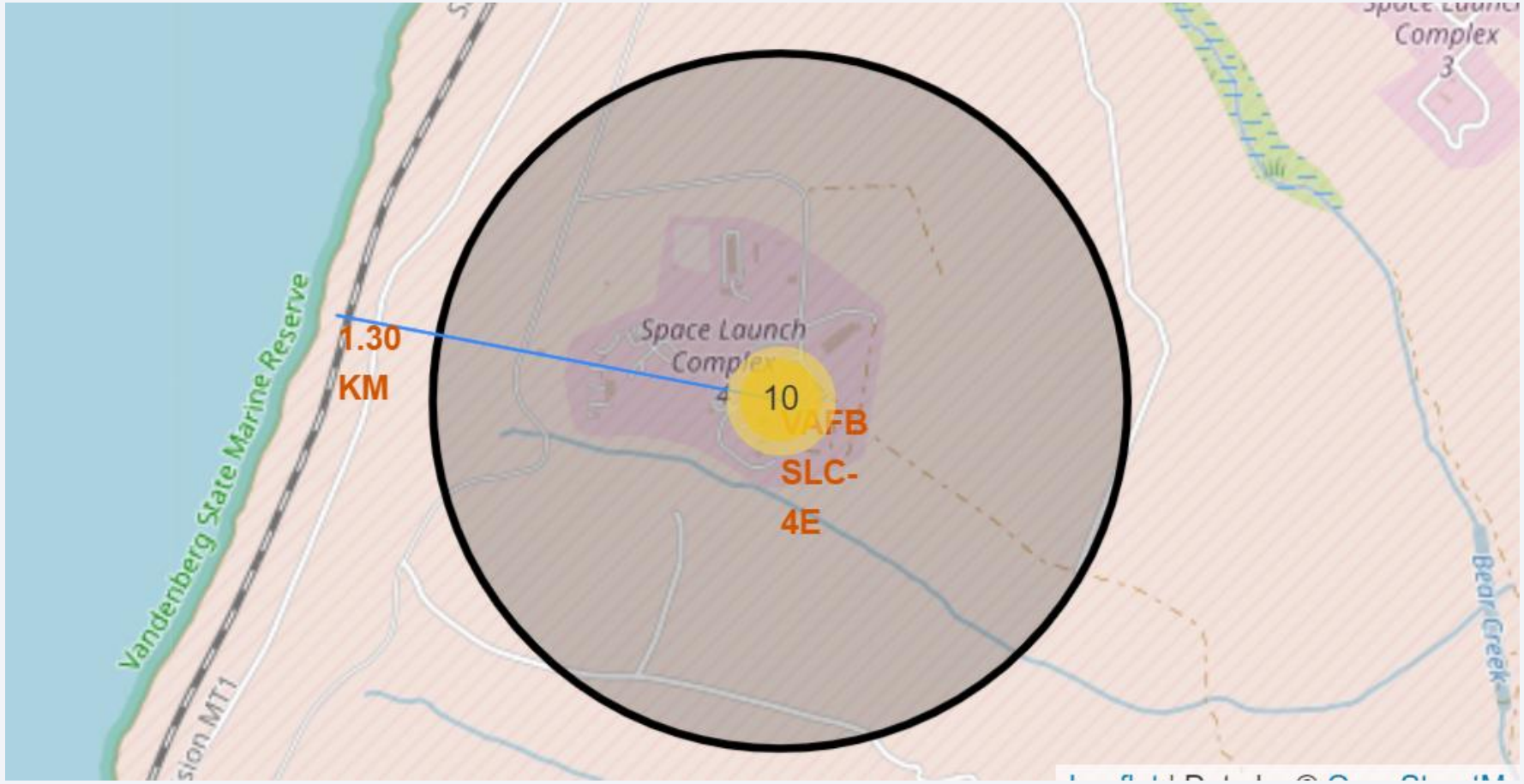

→ All Launch sites are in USA coasts (Florida and California)

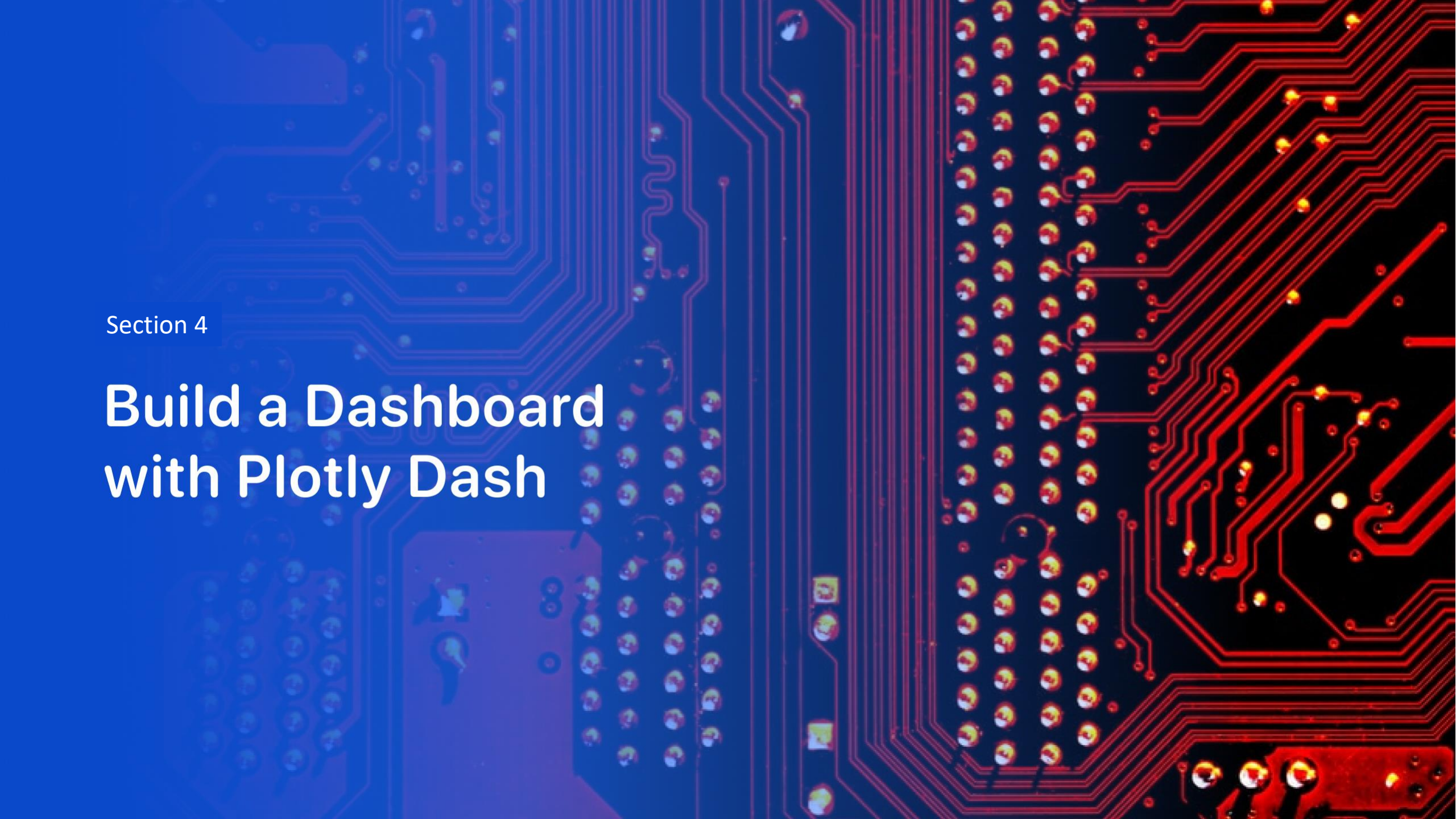
# Colour Labelled Markers for each launch site



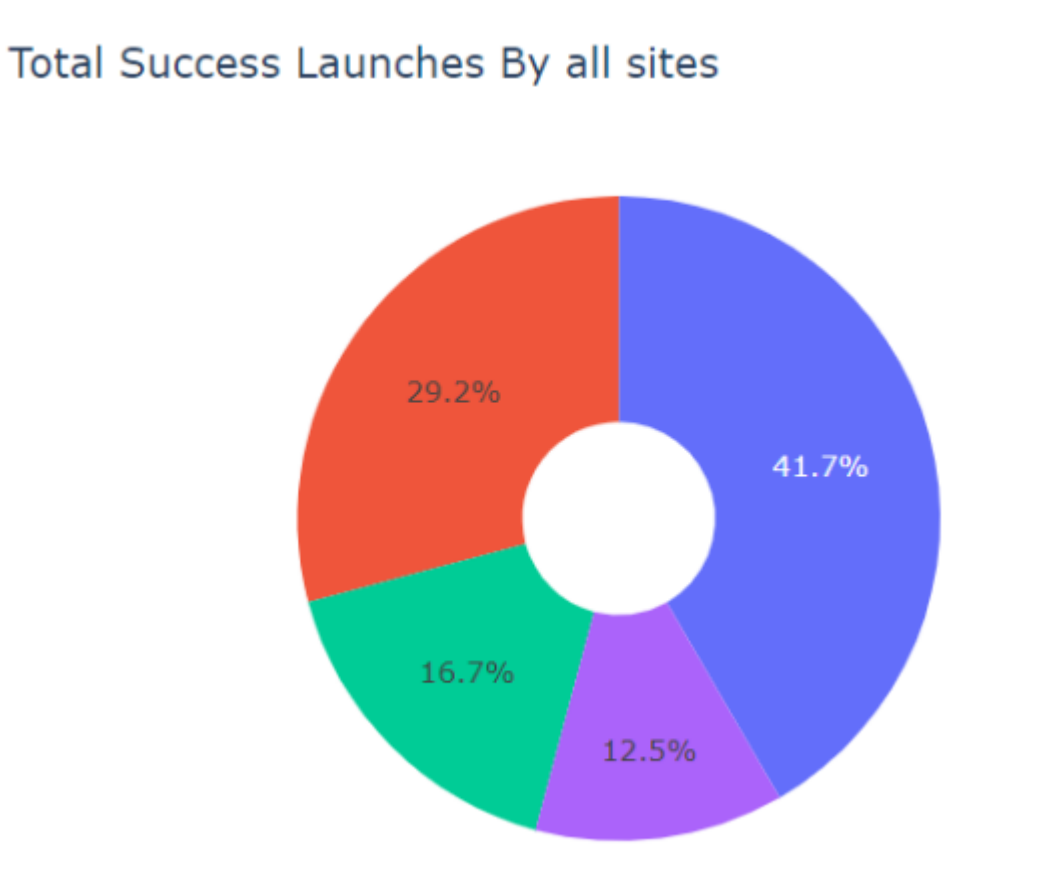

→ success rate is now visible on the map

# Launch Site distance to coast



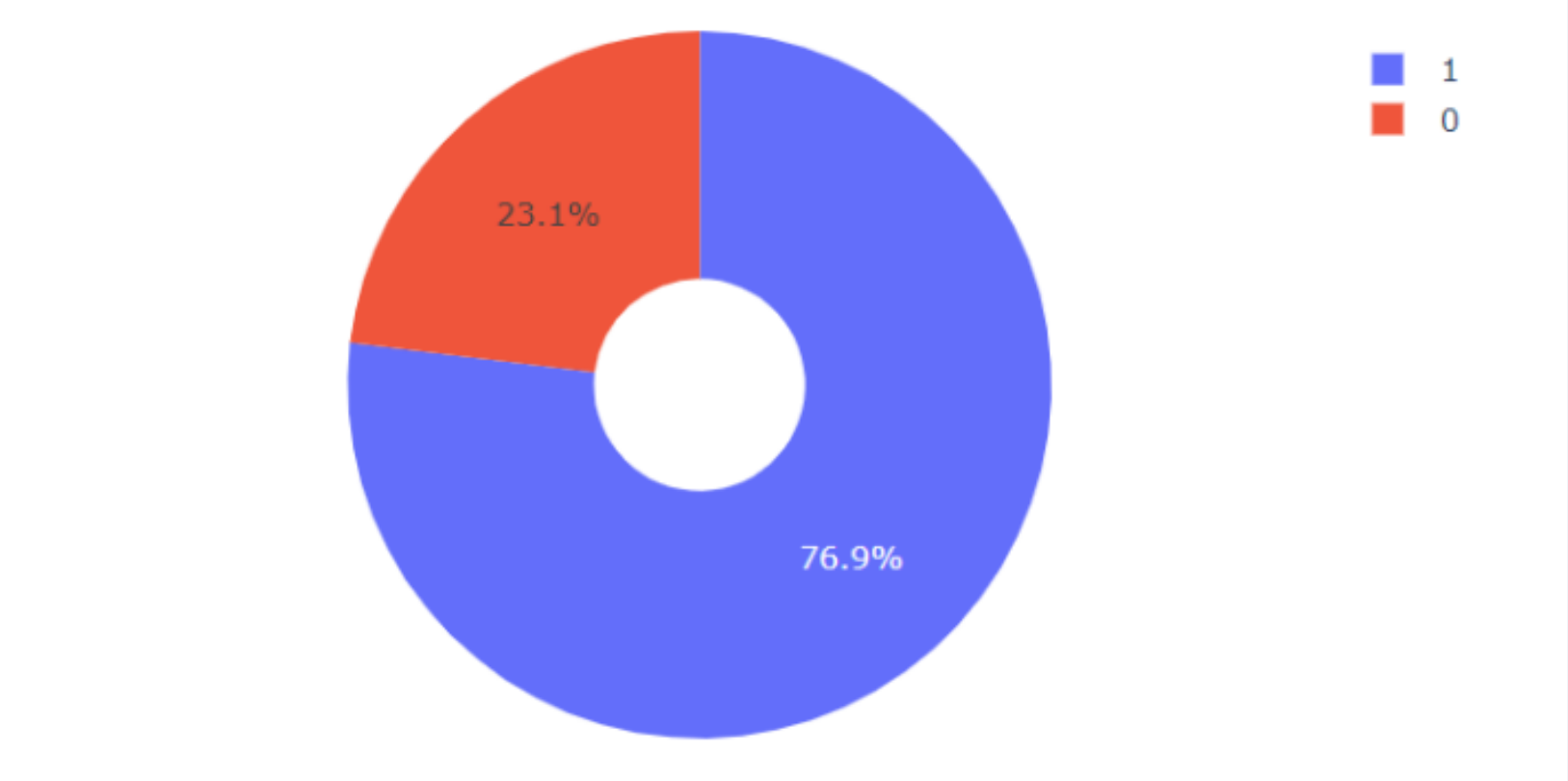

Section 4
Build a Dashboard with Plotly Dash

# Success by all sites



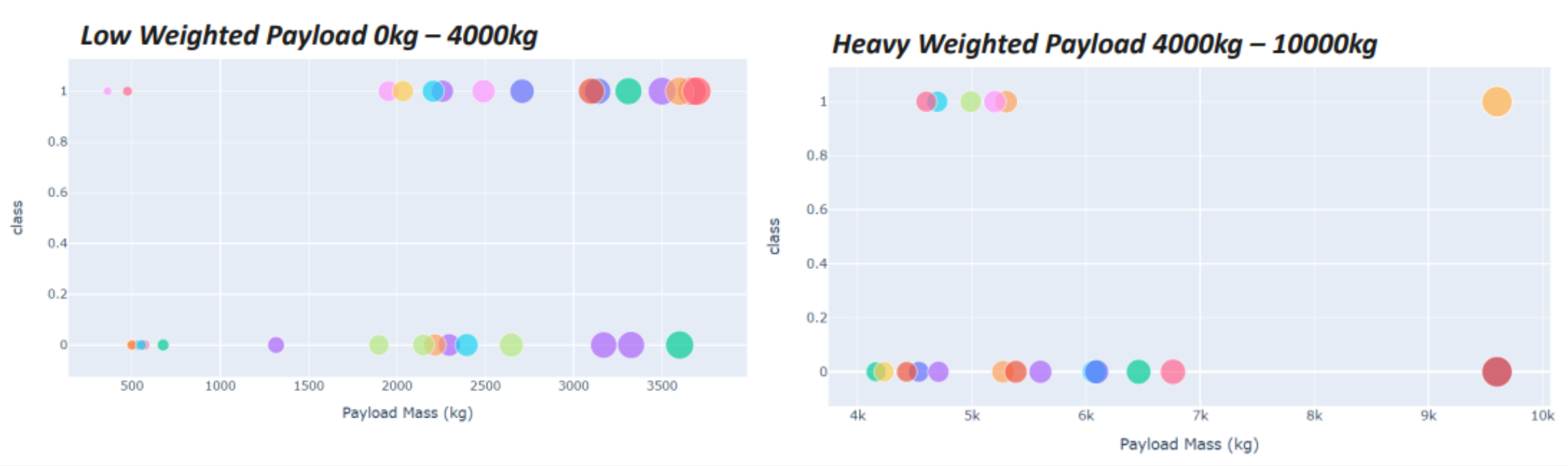

→ Site KSC LC-39A had most successful launches

# Pie Chart for KSC LC 39-A




→ KSC LC-39A had a 76.9% success rate

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




→ The success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

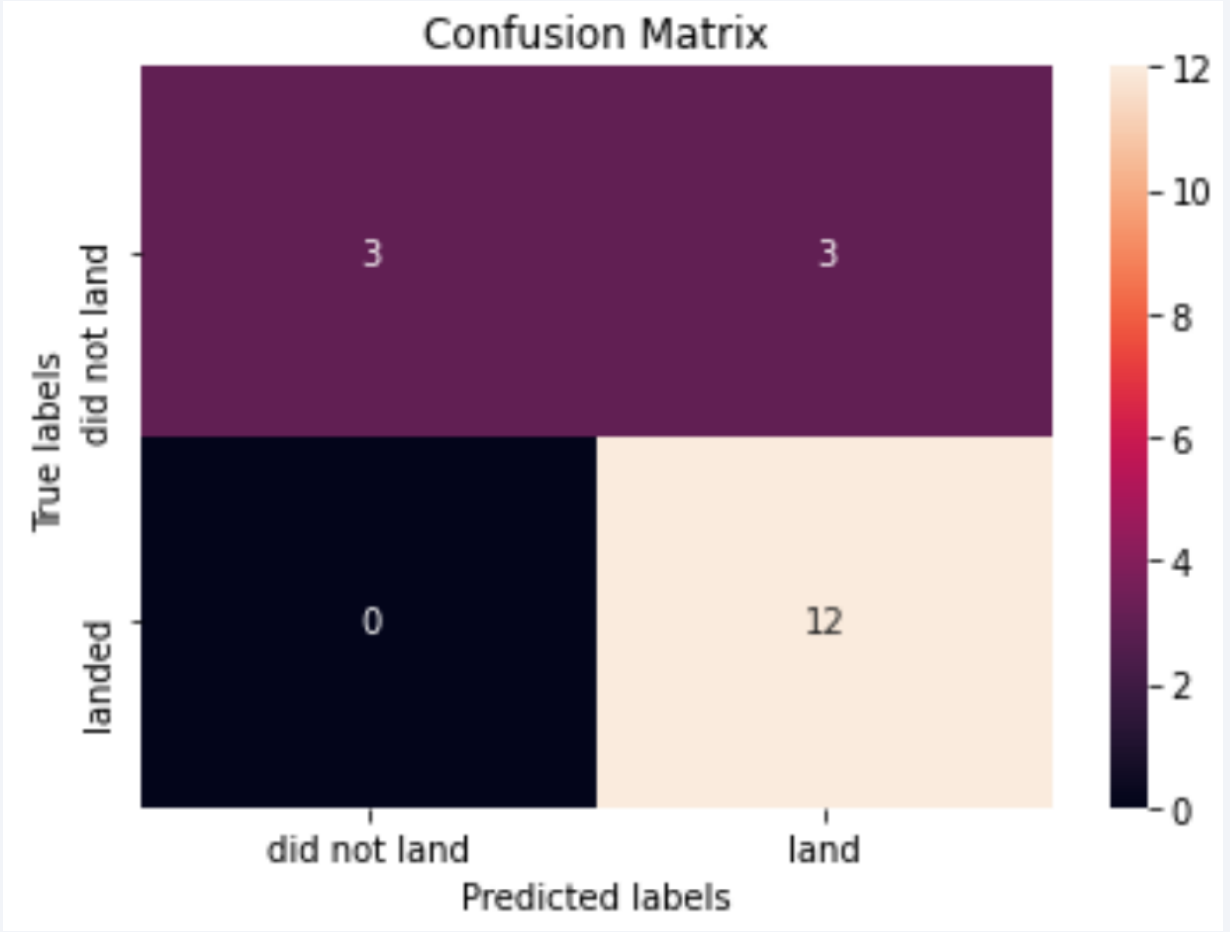
# Predictive Analysis (Classification)

# Classification Accuracy

- The model with best accuracy Is a decision tree model
- 83.33% accuracy on the test data with best hyperparameters

# Confusion Matrix

The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



# 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 Decision tree classifier is the best machine learning algorithm for this task.

# Thank you!