
# Winning Space Race with Data Science

Adnan Amro
15-05-2022

# Outline

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

# Executive Summary

- Methodologies used in this analysis were as follows:
  - Web scraping and SpaceX API for data collection
  - Data wrangling, data visualization and interactive maps
  - Machine Learning
- Summary of all results
  - All 3 methodologies resulted in positive outcome, data were collected then wrangled and visualized and at the end a machine learning model was created.

# Introduction

- The aim of this analysis is to train a machine learning model using public data of SpaceX to predict the best launch locations and successful landings of first stage of rockets.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology
  - Data were collected using two methods: SpaceX API and Web Scraping
- Perform data wrangling
  - Loaded Space X dataset, performed Exploratory Data Analysis then created a landing outcome label from Outcome column
- 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

- Two methods were used to collect data:

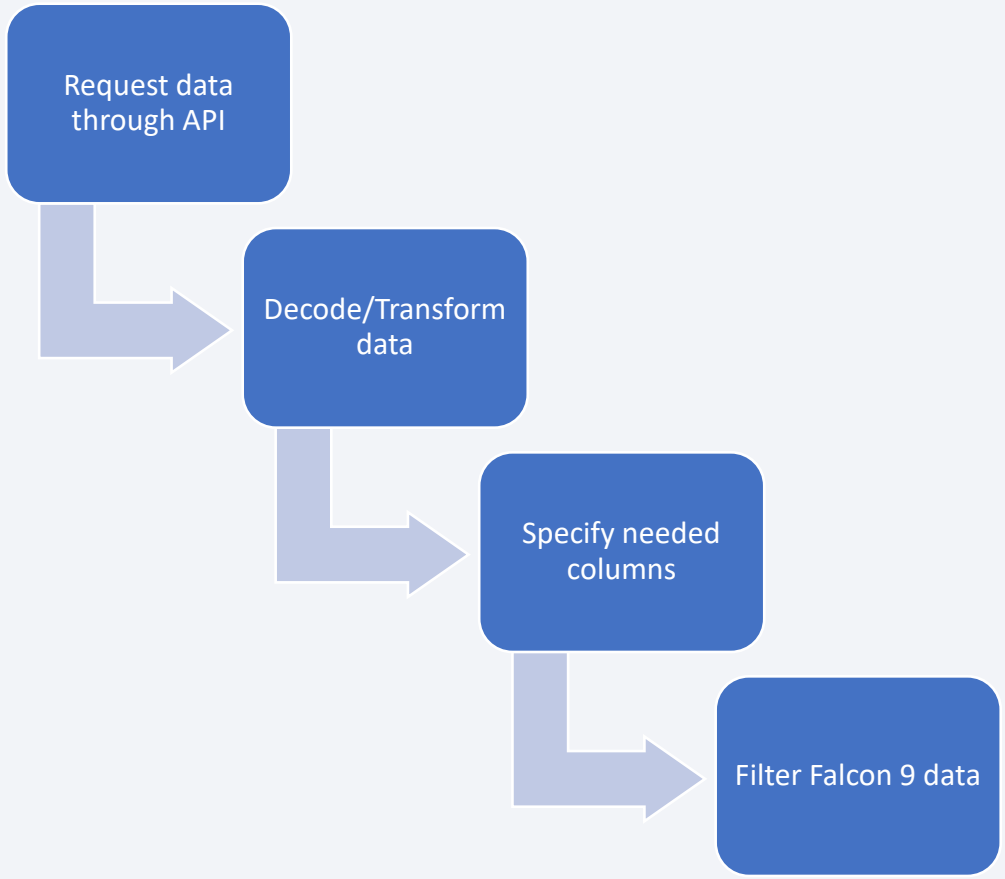
1- API:

“https://api.spacexdata.com/v4/launches/past"

2- Web Scraping:
https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

# Data Collection – SpaceX API

- Requested rocket launch data from SpaceX API using URL: "https://api.spacexdata.com/v4/launches/past"
- Decoded the response content as a Json using .json() and turned it into a Pandas DF using .json_normalize()
- Specified needed columns (rocket, payloads, launchpad, cores)
- Removed Falcon 1 launches keeping only the Falcon 9 launches




Notebook link:

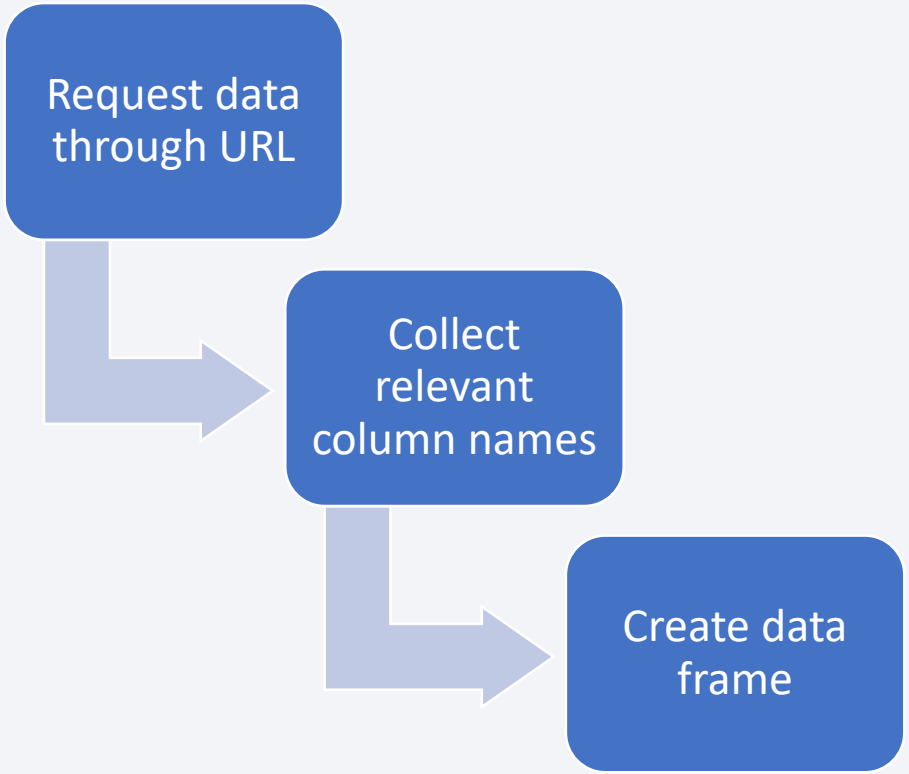
https://github.com/adnanamro/Capstone/blob/master/01.%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb

# Data Collection - Scraping

- Requested data from Falcon9 Launch Wiki page through URL:

"https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

- Extracted all column/variable names from the HTML table header
- Created a data frame by parsing the launch HTML tables




Notebook link:

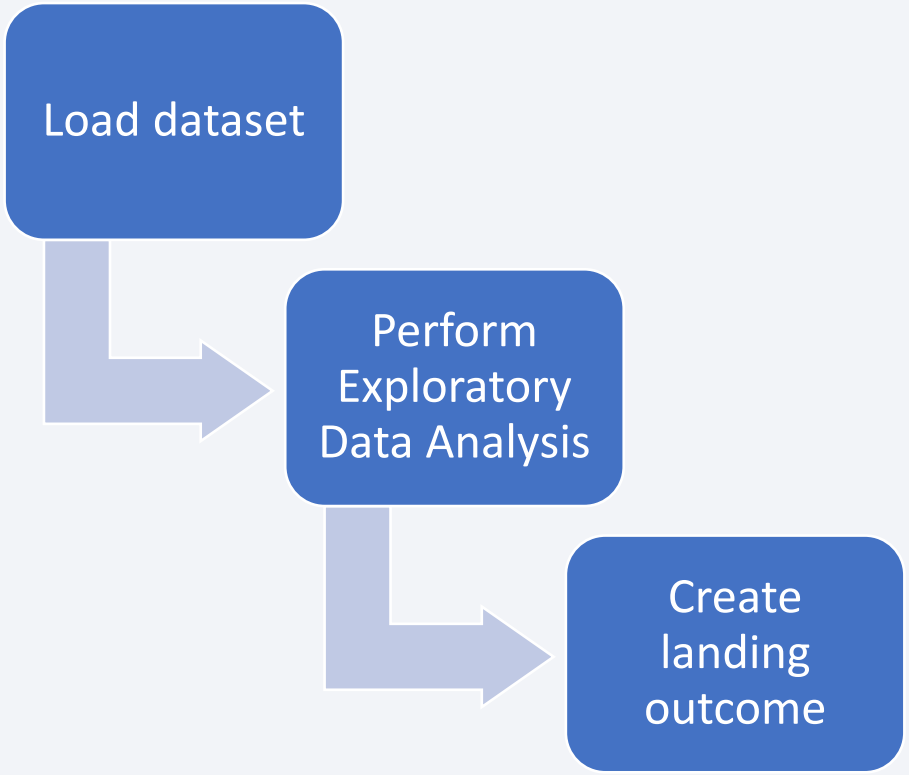
https://github.com/adnanamro/Capstone/blob/master/02.%20Data%20Collection%20with%20Web%20Scraping.ipynb

# Data Wrangling

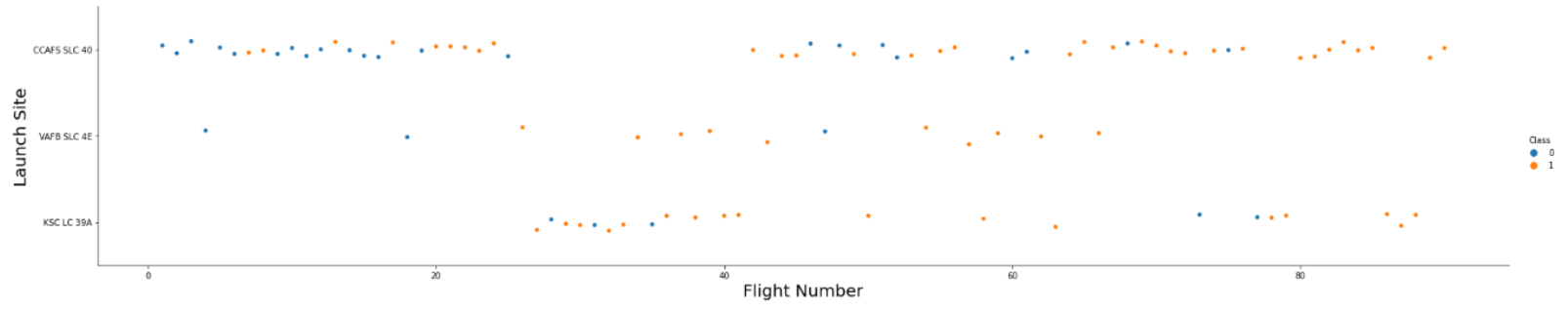
- Loaded Space X dataset
- Performed Exploratory Data Analysis:
  - Calculated the number of launches on each site
  - Calculated the number and occurrence of each orbit
  - Calculated the number and occurrence of mission outcome per orbit type
- Created a landing outcome label from Outcome column

Notebook link:

https://github.com/adnanamro/Capstone/blob/master/03.%20Data%20Wrangling.ipynb



# EDA with Data Visualization

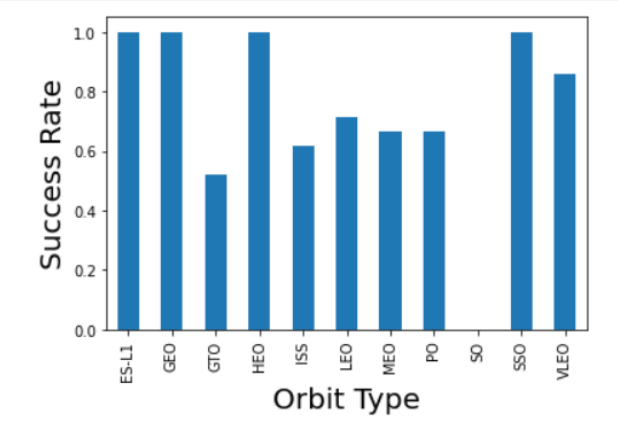



Scatter Plot was used several times to describe relations between the following pairs:

FlightNumber vs. PayloadMass

FlightNumber vs LaunchSite

LaunchSite vs PayloadMass




Bar chart was used to describe relation between:

Success Rate vs Orbit Type

## Notebook link:

https://github.com/adnanamro/Capstone/blob/master/04.%20Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

# EDA with SQL

SQL Queries Performed:

- (SELECT DISTINCT … FROM …) Display the names of the unique launch sites in the space mission
- (SELECT … FROM … WHERE … LIKE … LIMIT …) Display 5 records where launch sites begin with the string 'CCA'
- (SELECT SUM … FROM … WHERE …) Display the total payload mass carried by boosters launched by NASA (CRS)
- (SELECT AVG … FROM … WHERE … LIKE …) Display average payload mass carried by booster version F9 v1.1
- (SELECT MIN … FROM … WHERE …) List the date when the first successful landing outcome in ground pad was achieved
- (SELECT … FROM … WHERE … AND …) List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- (SELECT … COUNT … AS … FROM … GROUP BY …) List the total number of successful and failure mission outcomes
- (SELECT DISTINCT … FROM … WHERE … SELECT MAX … FROM …) List the names of the booster versions which have carried the maximum payload mass
- (SELECT … FROM … WHERE … AND … YEAR) List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- (SELECT … FROM … WHERE … BETWEEN … AND … GROUP BY … ORDER BY … DESC) 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.

Notebook link:

https://github.com/adnanamro/Capstone/blob/master/05.%20Complete%20the%20EDA%20with%20SQL.ipynb

# Build an Interactive Map with Folium

- Marked all launch sites on a map
  - Added each site's location on a map using site's latitude and longitude coordinates
  - Created a folium Map object
  - Added a highlighted circle area with a text label on a specific coordinate
- Mark the success/failed launches for each site on the map
  - Adding the launch outcomes for each site
  - Created markers for all launch records
  - Created a MarkerCluster object
  - Calculated the distances between a launch site to its proximities
  - Added a MousePosition on the map
  - Calculated the distance between two points on the map

Notebook link:

https://github.com/adnanamro/Capstone/blob/master/06.%20Complete%20the%20Data%20Visualization%20with%20Folium.ipynb

# Build a Dashboard with Plotly Dash

The following Plots/Graphs were added to the dashboard:

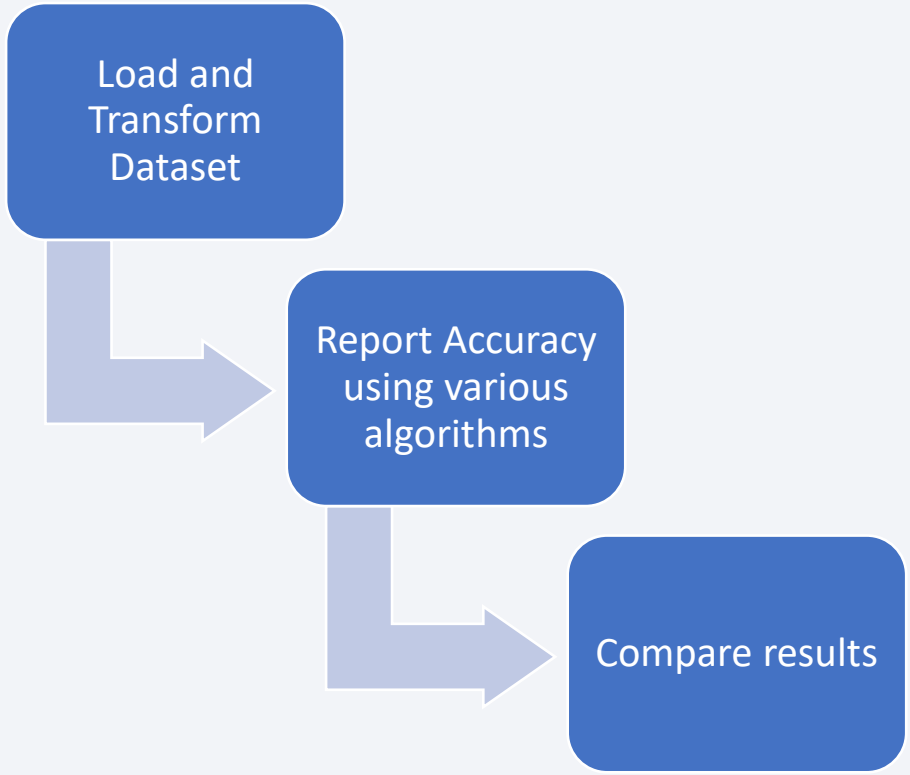
- Drop-down Input Component: Launch Site
- Pie Chart: Total Success Launches by Site
- Range Slider: Select Payload
- Scatter Plot: Payload Mass (kg) vs Class

Notebook link:

https://github.com/adnanamro/Capstone/blob/master/07.%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.py

# Predictive Analysis (Classification)

- Prepared Dataset
- Used the test set to report the accuracy of the model using the following algorithm:
  - K Nearest Neighbor (KNN)
  - Decision Tree
  - Support Vector Machine
  - Logistic Regression
- Compared results in a tabular format




Notebook Link:

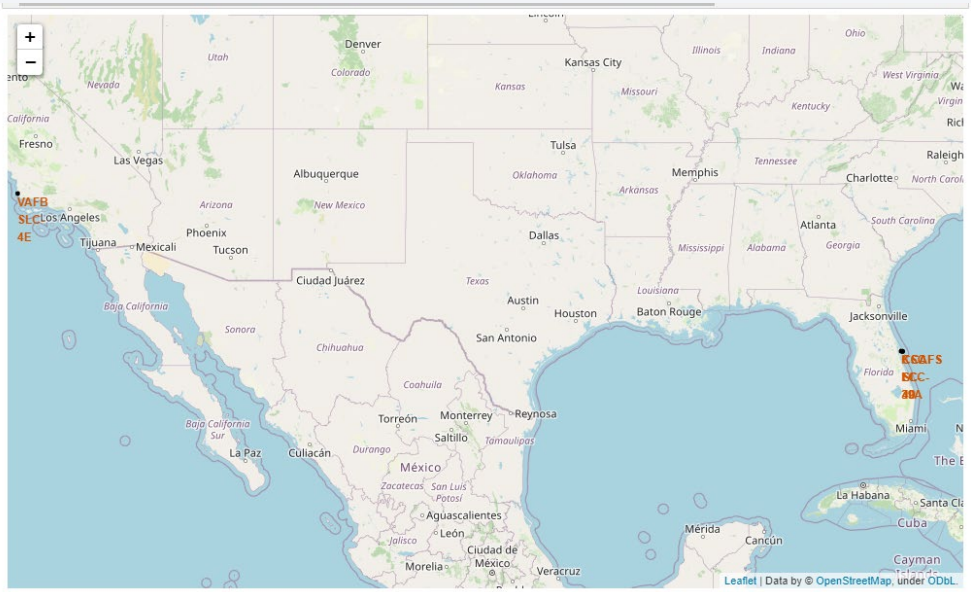
https://github.com/adnanamro/Capstone/blob/master/08.%20Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

# Results

Launch locations seems to be in both ends of South of USA near the coasts, most probably for safety issues and better transportation.

Decision Tree Classifier should be used to predict successful landings.

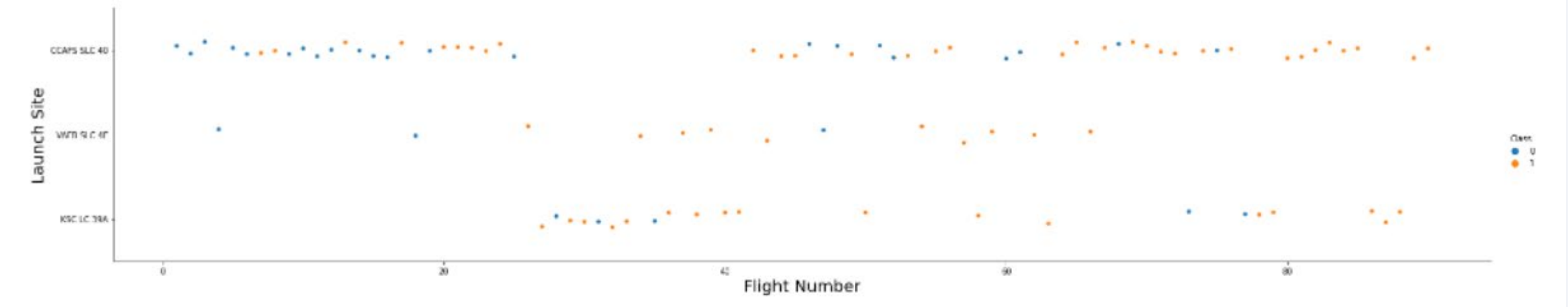
Out[8]:



Section 2

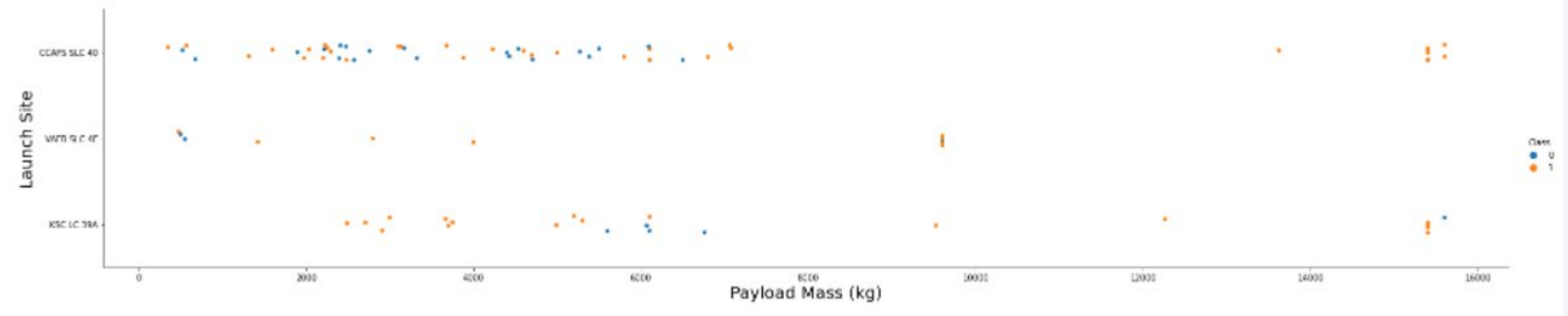
# Insights drawn from EDA

# Flight Number vs. Launch Site



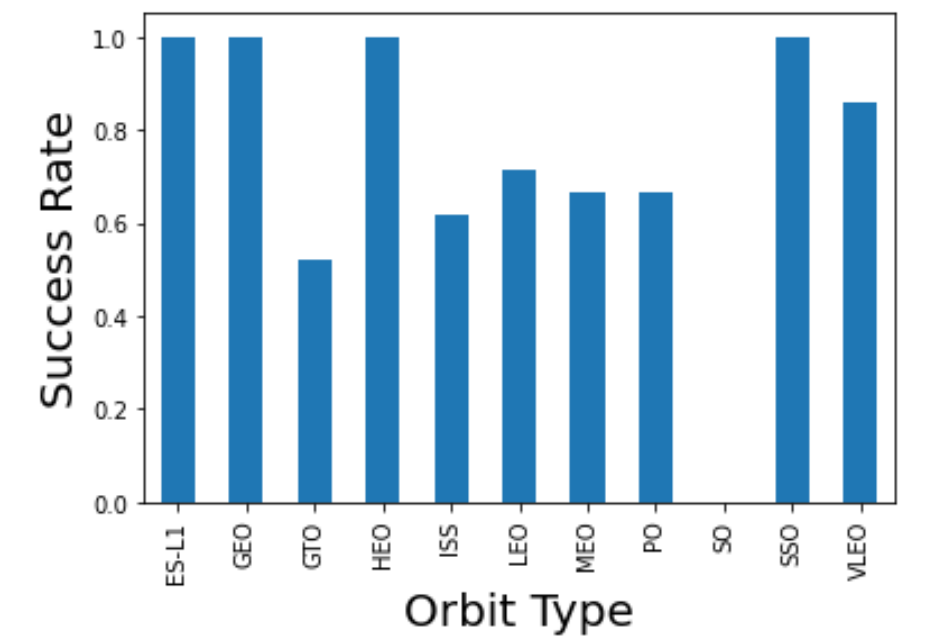

- Most successful Launch Site is CCAFS SLC 40
- As more attempts were made, more successful launches were achieved

# Payload vs. Launch Site



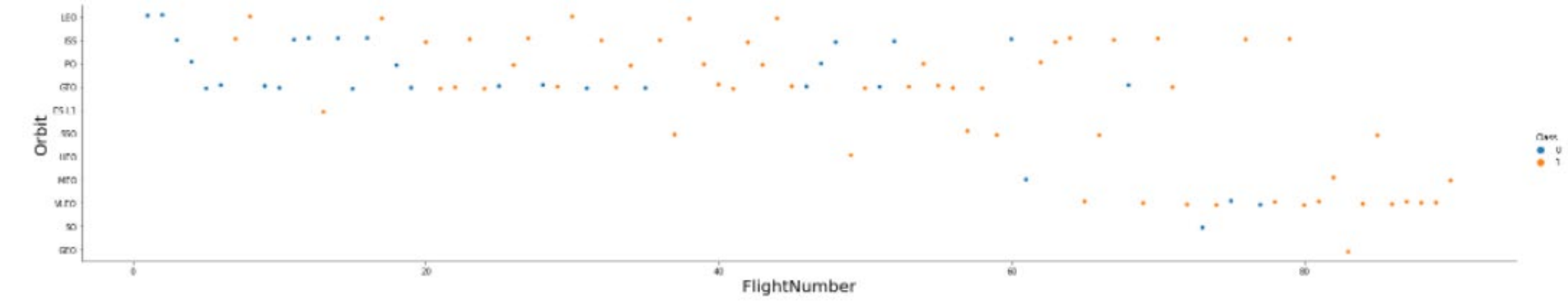

- Launches at sites CCAFS SLC 40, KSC LC 39A had a high success rate with Payloads above 10,000 kg.
- Launches at site VAFB SLC 4E had a high success rate with Payloads below 10,000 kg.

# Success Rate vs. Orbit Type



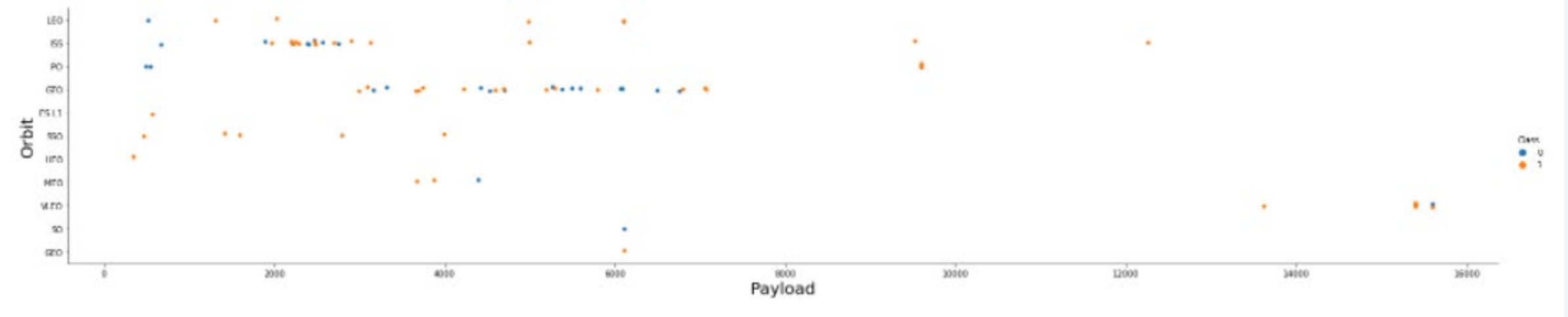

- Orbits with highest success rates are ES-L1, GEO, HEO, SSO.

# Flight Number vs. Orbit Type




- This explains the high success rate of Launch Site GEO, and low success rate of SO since only 1 flight was done on each.

# Payload vs. Orbit Type




- Less launches were done on SO and GEO orbits, regardless of payload

# All Launch Site Names

## Query

```
In [23]:
%%sql
SELECT DISTINCT LAUNCH_SITE
FROM SPACEXTBL;
```

## Output

Out[23]:

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

- There are 4 unique launch sites

# Launch Site Names Begin with 'CCA'

## Query

```
In [24]:
%%sql
SELECT LAUNCH_SITE
FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

## Output

Out[24]:

| launch_site |
|---|
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |
| CCAFS LC-40 |

- The same launch site is repeated 5 times

# Total Payload Mass

## Query

```
In [25]:
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer = 'NASA (CRS)';
```

## Output

```
Out[25]:  1
          45596
```

- Circa 46 tons were carried by boosters launched by NASA (CRS)

# Average Payload Mass by F9 v1.1

## Query

```
In [26]:
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

## Output

```
Out[26]:  1
          340
```

- Average of 340 kg payload mass was carried by booster version F9 v1.1

# First Successful Ground Landing Date

## Query

```
In [28]:
%%sql
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship)'
    AND 4000 < PAYLOAD_MASS__KG_ < 6000;
```

## Output

```
Out[27]:   1
           2015-12-22
```

- First successful landing outcome in ground pad was achieved in 2015-12-22

# Successful Drone Ship Landing with Payload between 4000 and 6000

## Query

In [28]:

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

## Output

Out[28]:

| booster_version |
|---|
| F9 FT B1021.1 |
| F9 FT B1023.1 |
| F9 FT B1029.2 |
| F9 FT B1038.1 |
| F9 B4 B1042.1 |
| F9 B4 B1045.1 |
| F9 B5 B1046.1 |

- The 7 boosters with success in drone ship landing are listed as a result

# Total Number of Successful and Failure Mission Outcomes

## Query

```
In [29]:
%%sql
SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
GROUP BY MISSION_OUTCOME;
```

## Output

Out[29]:

| mission_outcome | total_number |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- A categorized table is listed as a result

# Boosters Carried Maximum Payload

## Query

```
In [30]:
%%sql
SELECT DISTINCT BOOSTER_VERSION
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTBL);
```

## Output

Out[30]:

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

- A table of boosters is listed as a result

# 2015 Launch Records

## Query

```
In [31]:
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

## Output

Out[31]:

| landing__outcome | booster_version | launch_site |
|---|---|---|
| Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

- The query resulted in two failures with different booster version at the same launch site

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

## Query

```
In [32]:
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

## Output

Out[32]:

| landing__outcome | total_number |
|---|---|
| No attempt | 10 |
| Failure (drone ship) | 5 |
| Success (drone ship) | 5 |
| Controlled (ocean) | 3 |
| Success (ground pad) | 3 |
| Failure (parachute) | 2 |
| Uncontrolled (ocean) | 2 |
| Precluded (drone ship) | 1 |

- The query resulted in a table with categorized list of landing outcomes and related number


Section 3
Launch Sites
Proximities Analysis

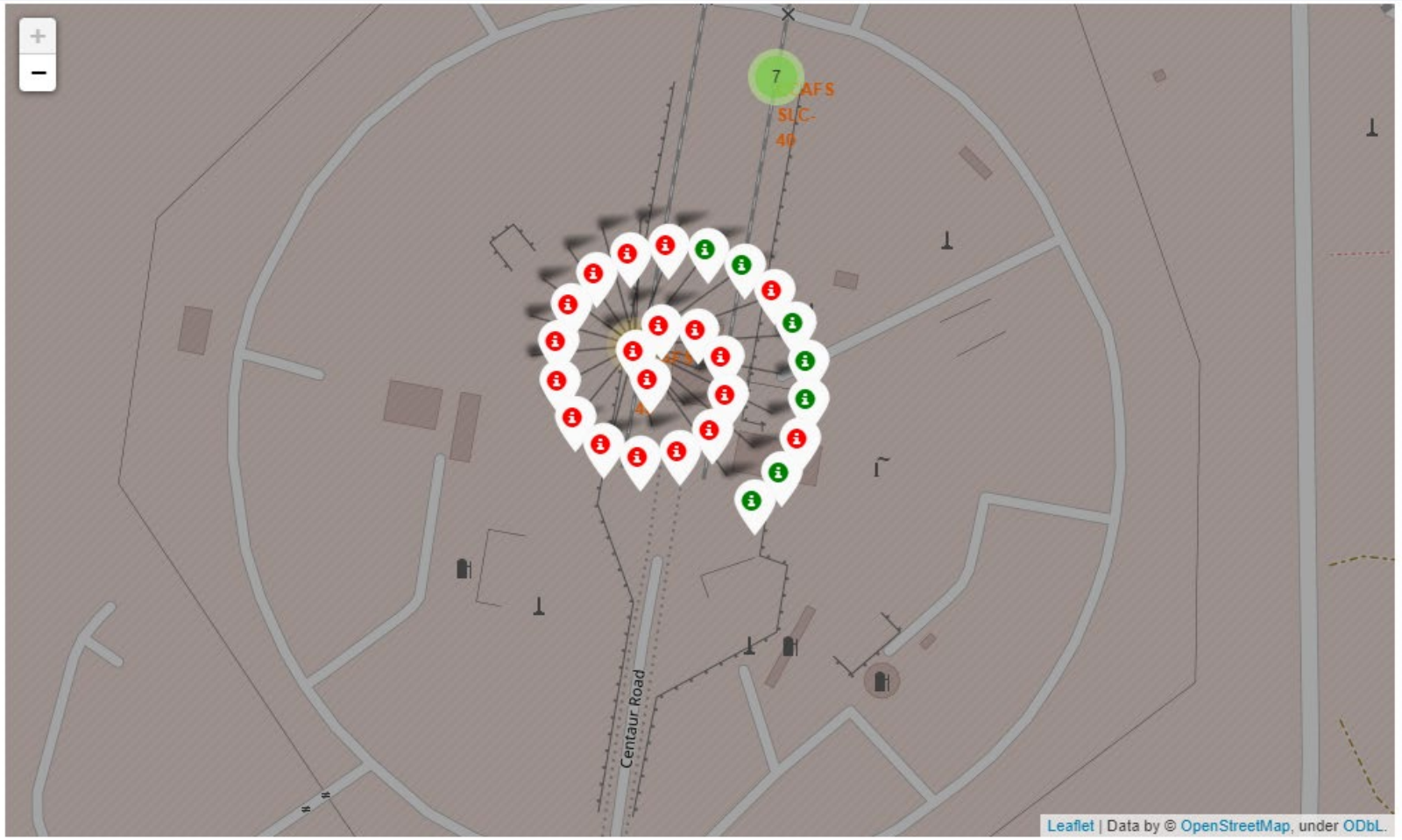
# Launch Sites Location

Out[8]:


- Launch locations seems to be in both ends of South of USA

# Color Labeled Launch Outcomes

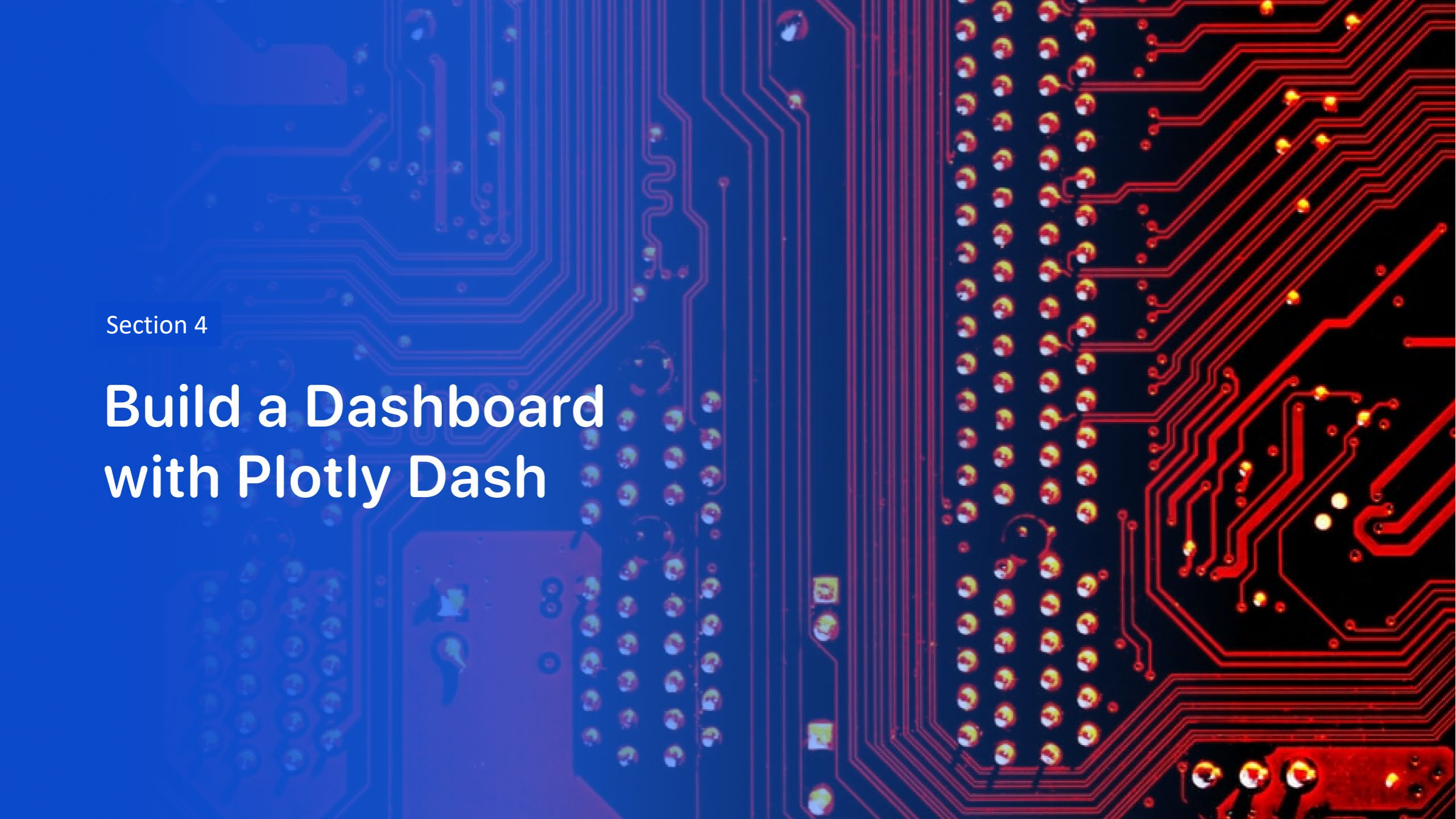
Out[18]:



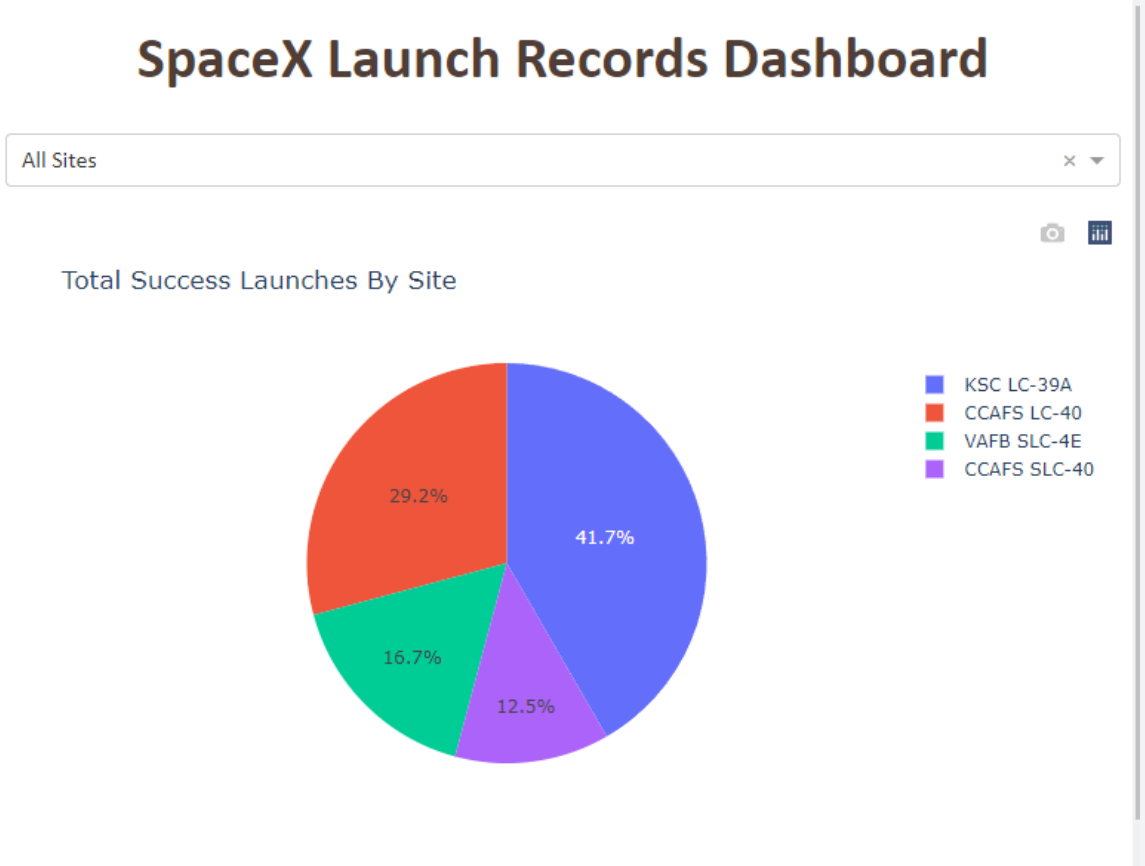

- Launch outcomes are color coded:

Success: Green

Fail: Red

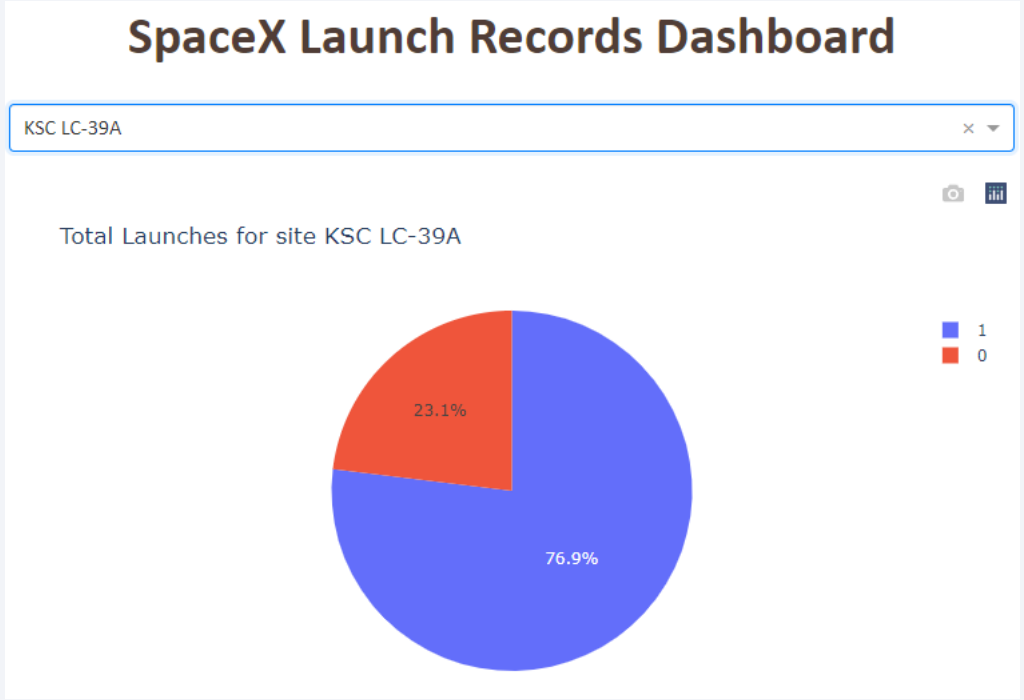

Section 4
Build a Dashboard with Plotly Dash

# Total Successful Launches by Site




- The graph shows the percentage of successful launches by site

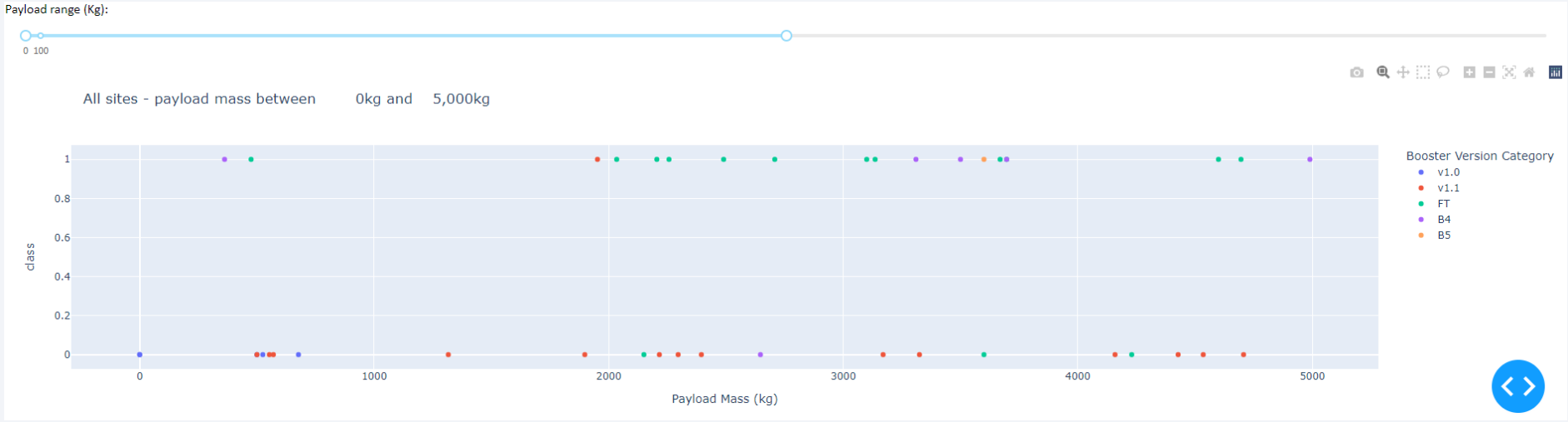
# KSC LC-39A Launches Status



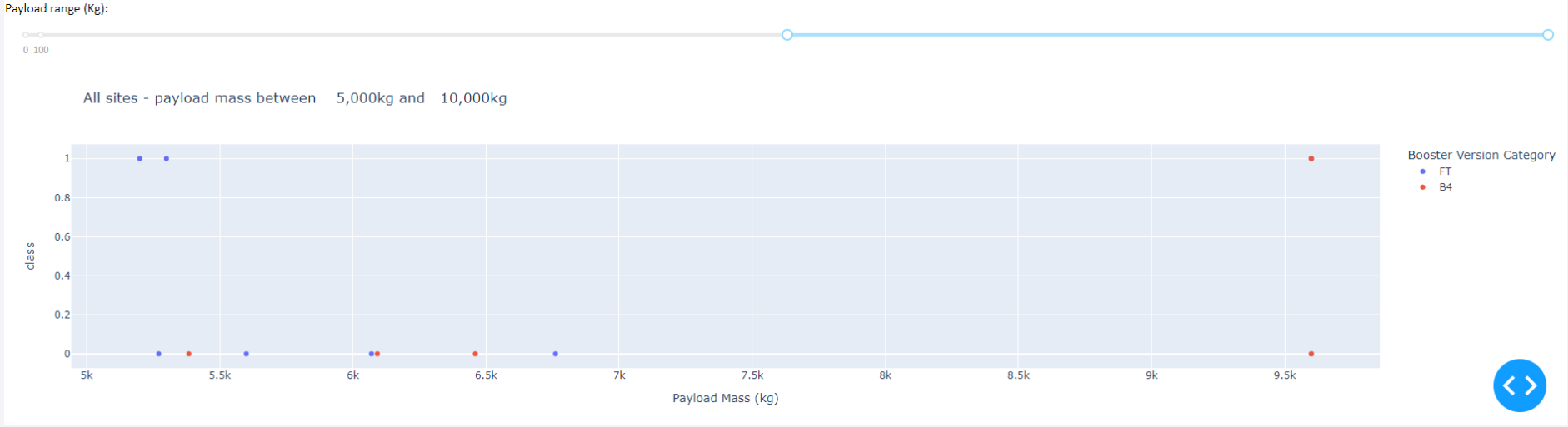

- The graph shows the percentage of successful launches against unsuccessful ones in filtered site

# Launch Outcomes by Payload

0-5,000 kg



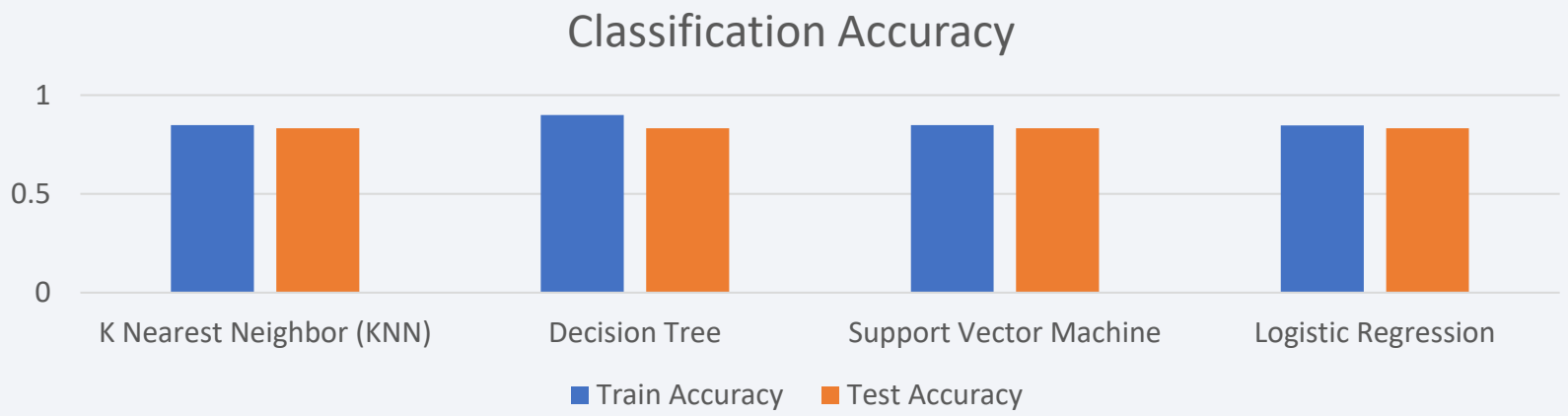

5,000-10,000 kg



Section 5

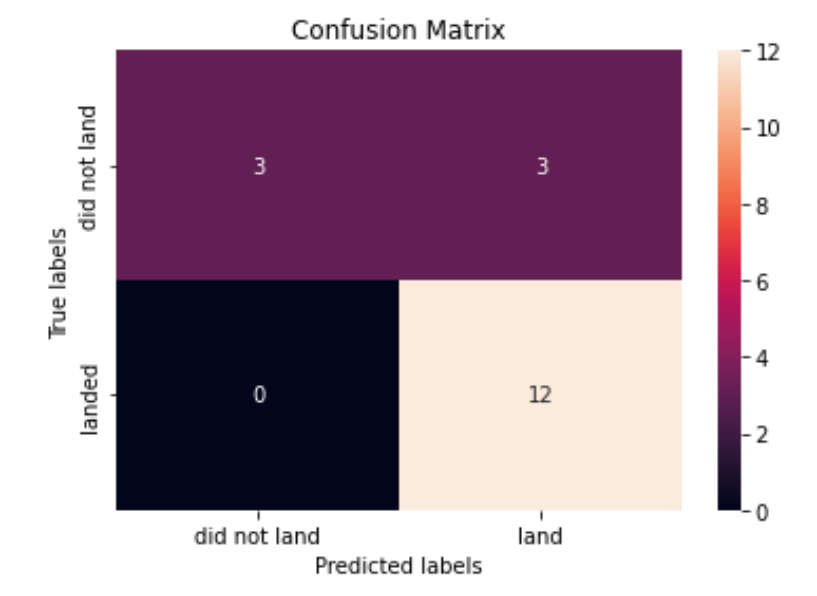
# Predictive Analysis (Classification)

# Classification Accuracy




- Decision Tree has the highest classification accuracy with 90%

# Confusion Matrix




- Decision Tree is the best performing model, with highest accuracy.

# Conclusions

- The best launch site is KSC LC-39A
- The higher the payload the more success rate is achieved
- Decision Tree Classifier should be used to predict successful landings

# Appendix

## Notebooks Links

- https://github.com/adnanamro/Capstone/blob/master/01.%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb
- https://github.com/adnanamro/Capstone/blob/master/02.%20Data%20Collection%20with%20Web%20Scraping.ipynb
- https://github.com/adnanamro/Capstone/blob/master/03.%20Data%20Wrangling.ipynb
- https://github.com/adnanamro/Capstone/blob/master/04.%20Complete%20the%20EDA%20with%20Visualization%20lab.ipynb
- https://github.com/adnanamro/Capstone/blob/master/05.%20Complete%20the%20EDA%20with%20SQL.ipynb
- https://github.com/adnanamro/Capstone/blob/master/06.%20Complete%20the%20Data%20Visualization%20with%20Folium.ipynb
- https://github.com/adnanamro/Capstone/blob/master/07.%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.py
- https://github.com/adnanamro/Capstone/blob/master/08.%20Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

# Thank you!