
# Winning Space Race with Data Science

Dash Abhishek
08-09-21

# OUTLINE

Executive Summary

Introduction

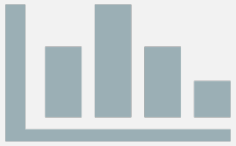
Methodology

Results

Conclusion

Appendix

# EXECUTIVE SUMMARY



## Summary of methodologies

1) Data Collection
2) Data Wrangling
3) EDA With Data Visualization
4) EDA With SQL
5) Interactive Map With Folium
6) Dashboard with Plotly Dash
7) Machine Learning with Classification

## Summary of all results

# INTRODUCTION

## Project background and context

In this project, we will predict whether the first stage a SpaceX designed Falcon 9 rocket will land successfully.  Falcon 9 is competitive in this rocket industry because it can save up to a 100 million dollars because of its ability to reuse the first stage of the rocket. So, in this project we will try to determine some sets of information that could possibly help competitors to determine whether a particular Falcon 9 rocket would succeed at reusing its first stage and thus bid against a SpaceX rocket launch.

## Key problems that need to be solved

- Can we predict if whether the first stage of the Falcon 9 rocket will land or not?
- What variables determine whether the first stage of the rocket will land successfully?
- To what extent does each variable determine the success of the launch?

Section 1

# Methodology

# METHODOLOGY

**Collect data** using SpaceX REST API and Web Scraping from Wikipedia

**Clean Up Data** using data wrangling

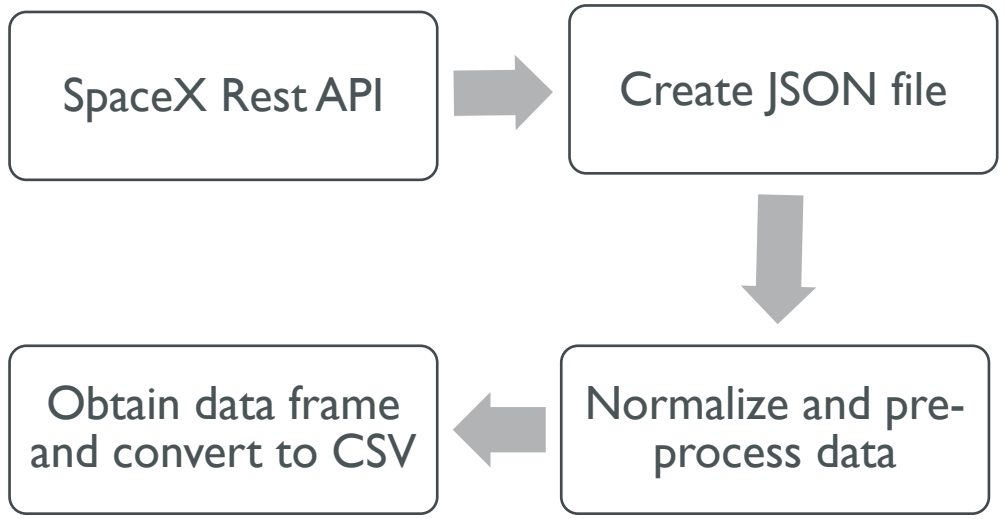
**Perform exploratory data analysis (EDA)** using visualization and SQL

**Perform interactive visual analytics** using Folium and Plotly Dash

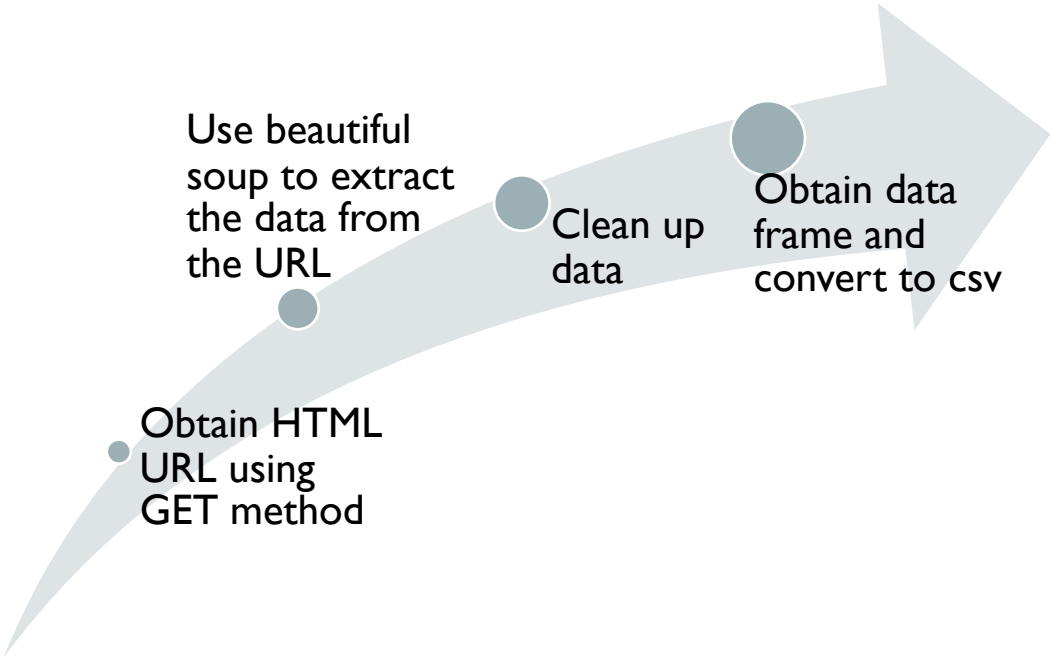
**Perform predictive analysis** using classification models to build, tune, evaluate classification models

# DATA COLLECTION

SpaceX Launch Data was collected using the **SpaceX API** that gave us all relevant data relating to rockets launches that would help determine whether rocket would launch

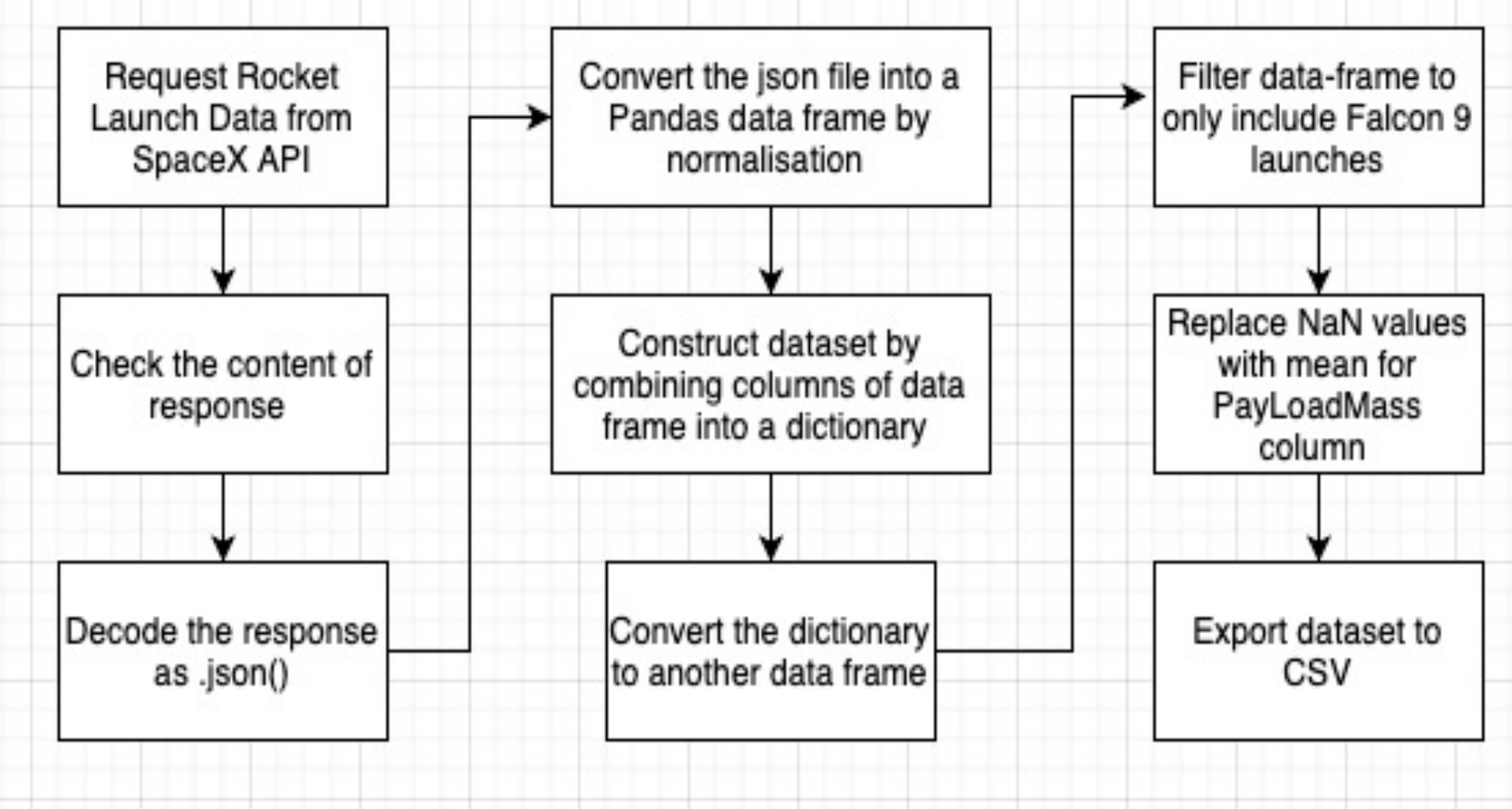



SpaceX launch data could also be collected using **web scraping** using the Beautiful Soup Python library to obtain data from the Wikipedia page of Falcon9



# DATA COLLECTION – SPACEX API

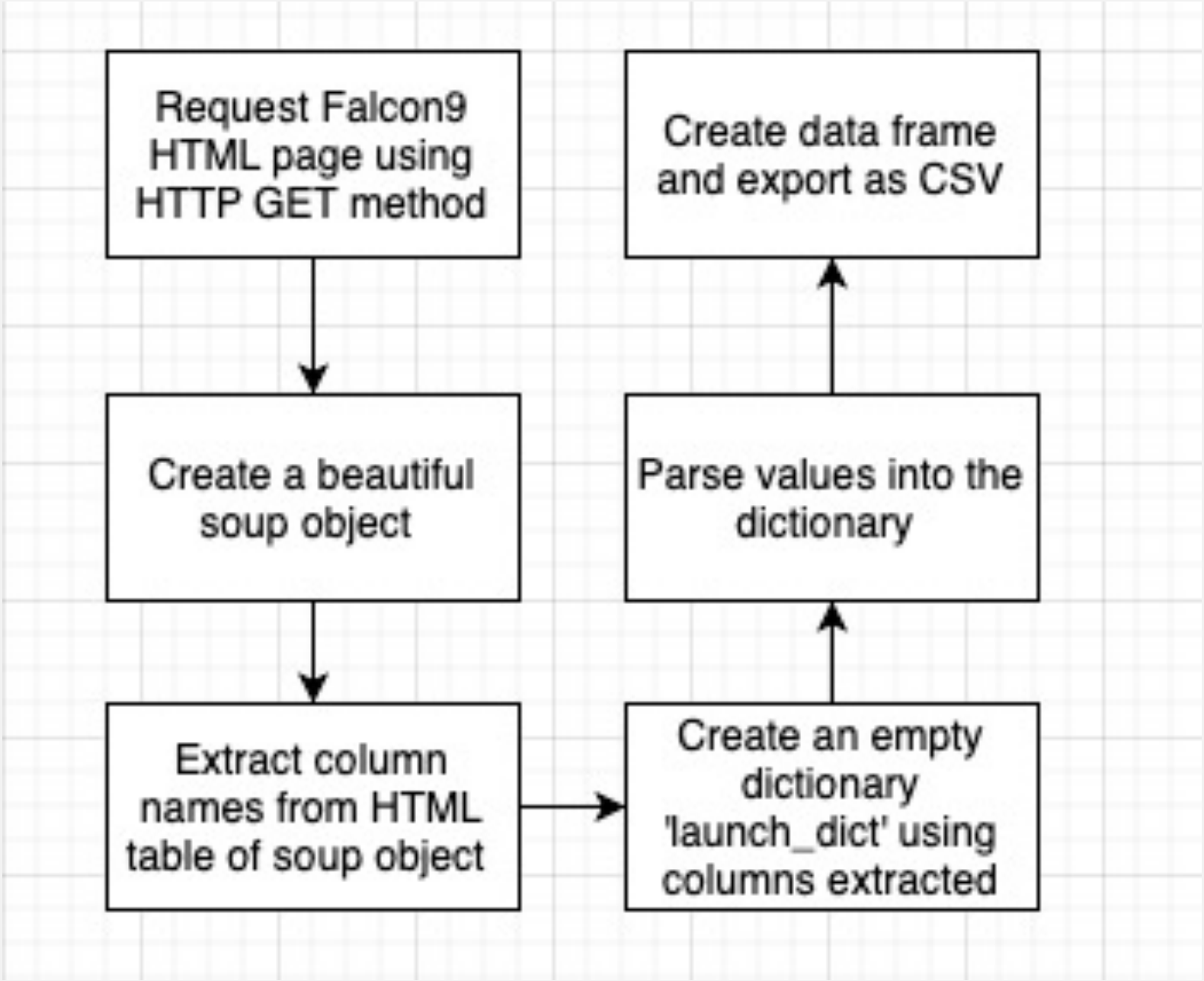
- The detailed algorithm for collecting data from the SpaceX API is as shown




[Link to GitHub Notebook]

# DATA COLLECTION - SCRAPING

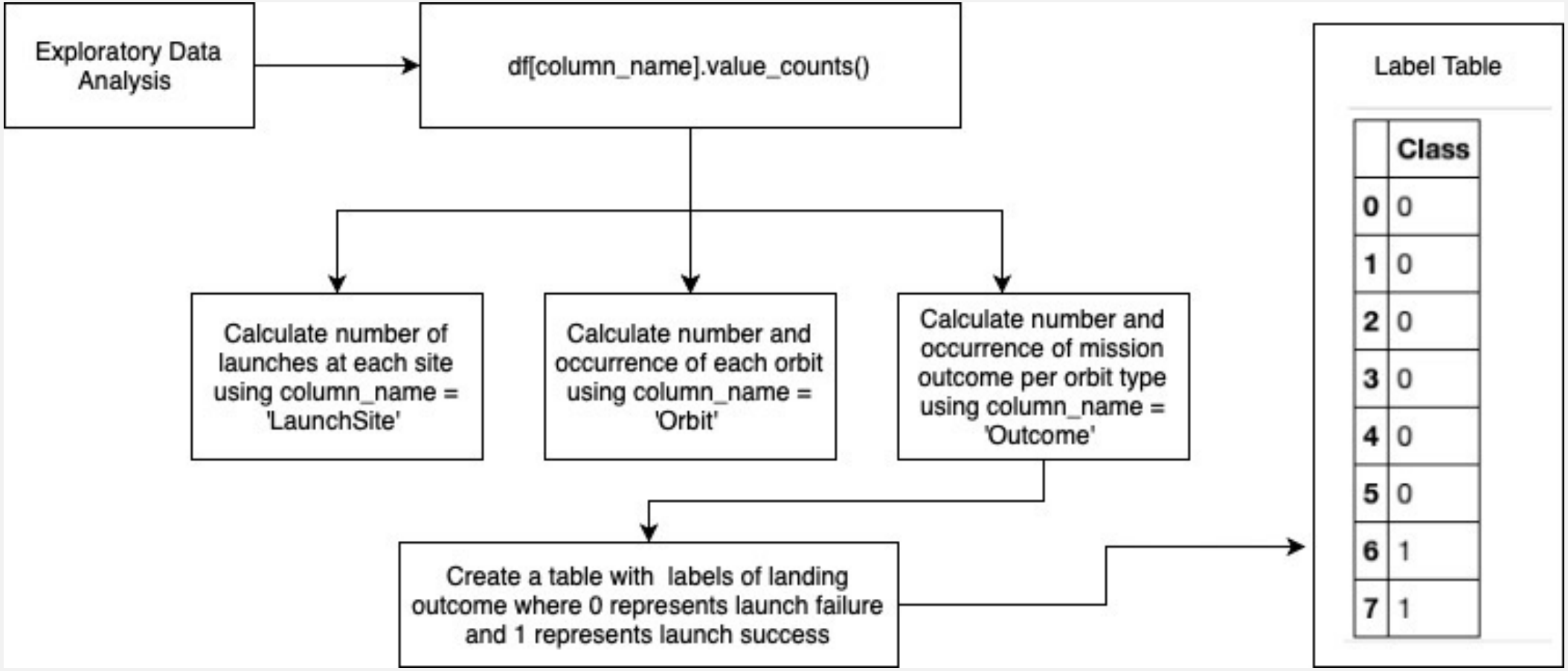
- The detailed algorithm of extracting data using scraping is as shown.




Link to GitHub Notebook

# DATA WRANGLING

- Data wrangling refers to the cleaning up of complex and unorganized data sets for easy access.
- Here, we follow the steps shown in the diagram in order to conduct exploratory data analysis to finally get a table where a successful launch is represented by 1 and an unsuccessful by a 0.

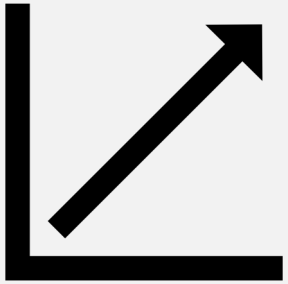



Link to GitHub Notebook

# EDA WITH DATA VISUALIZATION

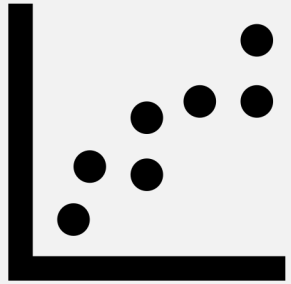
**Bar Graph**: Used to represent categorical data and helps us show the relationship between a categorical value versus a discrete numerical value.

Mean Vs Orbit



**Line Graph**: Used to represent continuous variables and identify trends to help predict one variable from the other. Helps us analyze future trends that we have no data about.

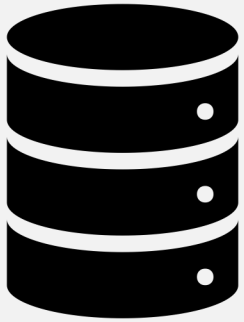
Success Rate Vs Year



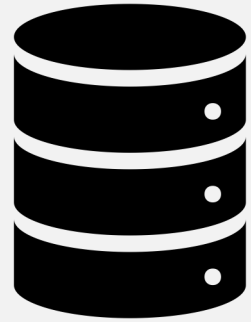
**Scatter Plot**: Used to represent paired numerical data and helps us determine the relationship between two variables to depict correlation.

- Flight Number Vs Launch Site
- Flight Number Vs Payload mass
- Payload Vs Launch Site
- Orbit Vs Flight Number
- Payload Vs Orbit Type
- Orbit Vs Payload Mass

Link to GitHub Notebook



# EDA WITH SQL



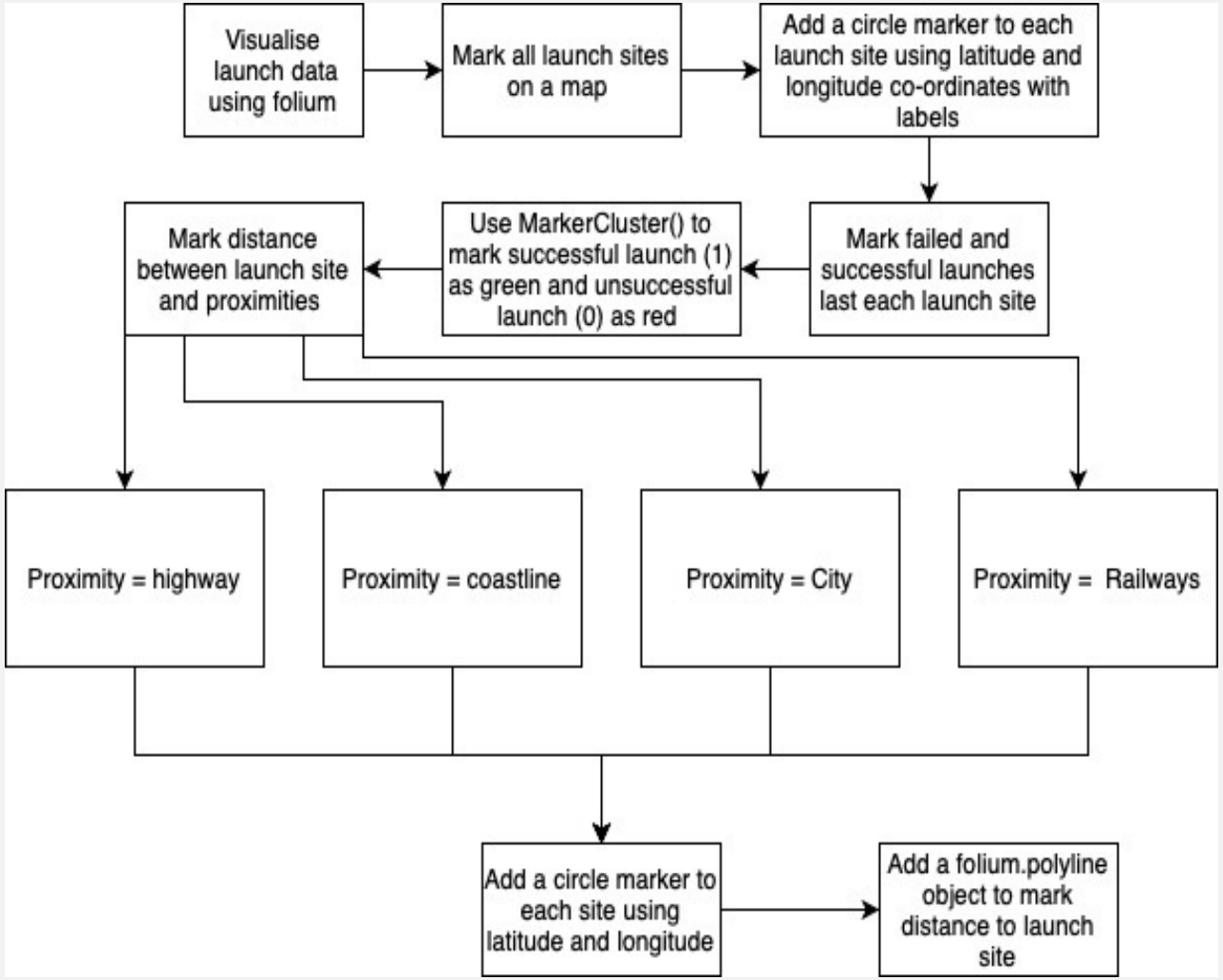
The following queries were performed using SQL to gather useful relationships between variables in the dataset:

- 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)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 6000 kg
- 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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

Link to GitHub Notebook

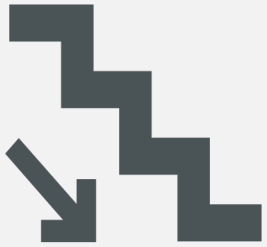
# BUILD AN INTERACTIVE MAP WITH FOLIUM

- The process we follow is as shown in the flowchart to produce three visualizations:
    1. Map of launch sites
    2. Map of successful and failed launches at launch sites
    3. Map of distance between launch site and proximities (4 maps)
- We use this to analyze a few visualization questions:
    1. How many launch sites are there?
    2. Are launch sites near the equator?
    3. Are launch sites near the coastline?
    4. Which are the most successful launch sites?
    5. Are launch sites in close proximity to railways?
    6. Are launch sites in close proximity to highways?
    7. Are launch sites in close proximity to coastline?
    8. Do launch sites keep certain distance away from cities?



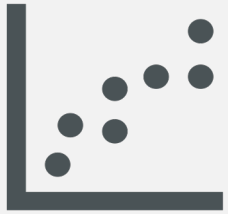

Link to GitHub Notebook

# BUILD A DASHBOARD WITH PLOTLY DASH

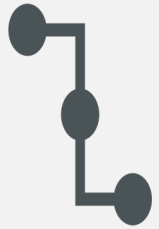


**Drop Down Menu**: Users can select which launch site they want to view or view all launch sites

**Pie Chart:** To represent the most successful launch sites so that the biggest sector is clearly shown



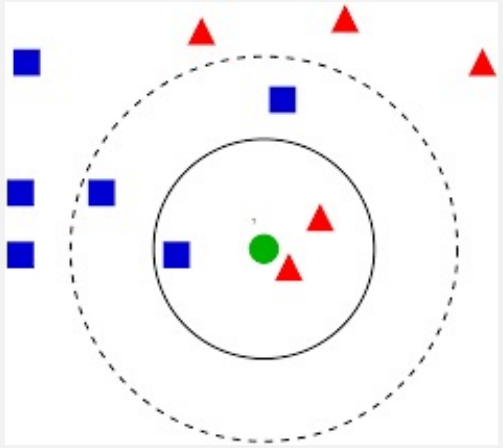
**Scatter Plot:** To represent the most successful payload range for rocket launches



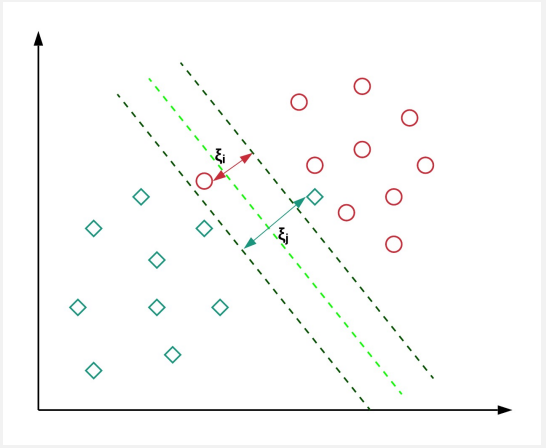
**Slider**: Helps users select the payload range for scatter plot easily

Link to GitHub Notebook

# PREDICTIVE ANALYSIS (CLASSIFICATION)

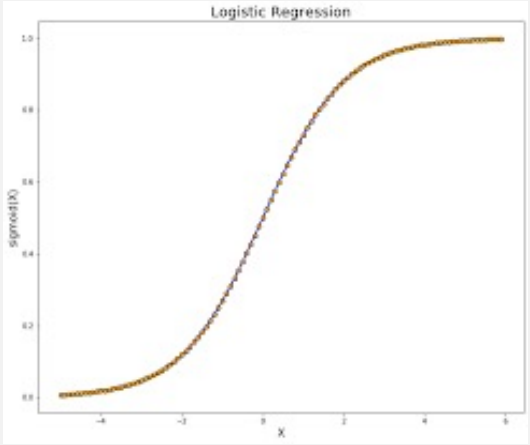


**K Nearest Neighbors**

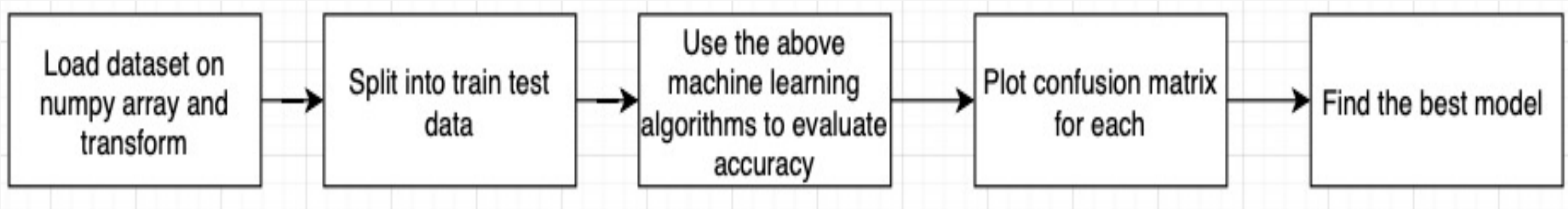


**Support Vector Machine**


**Decision Tree Classifier**




**Logistic Regression**



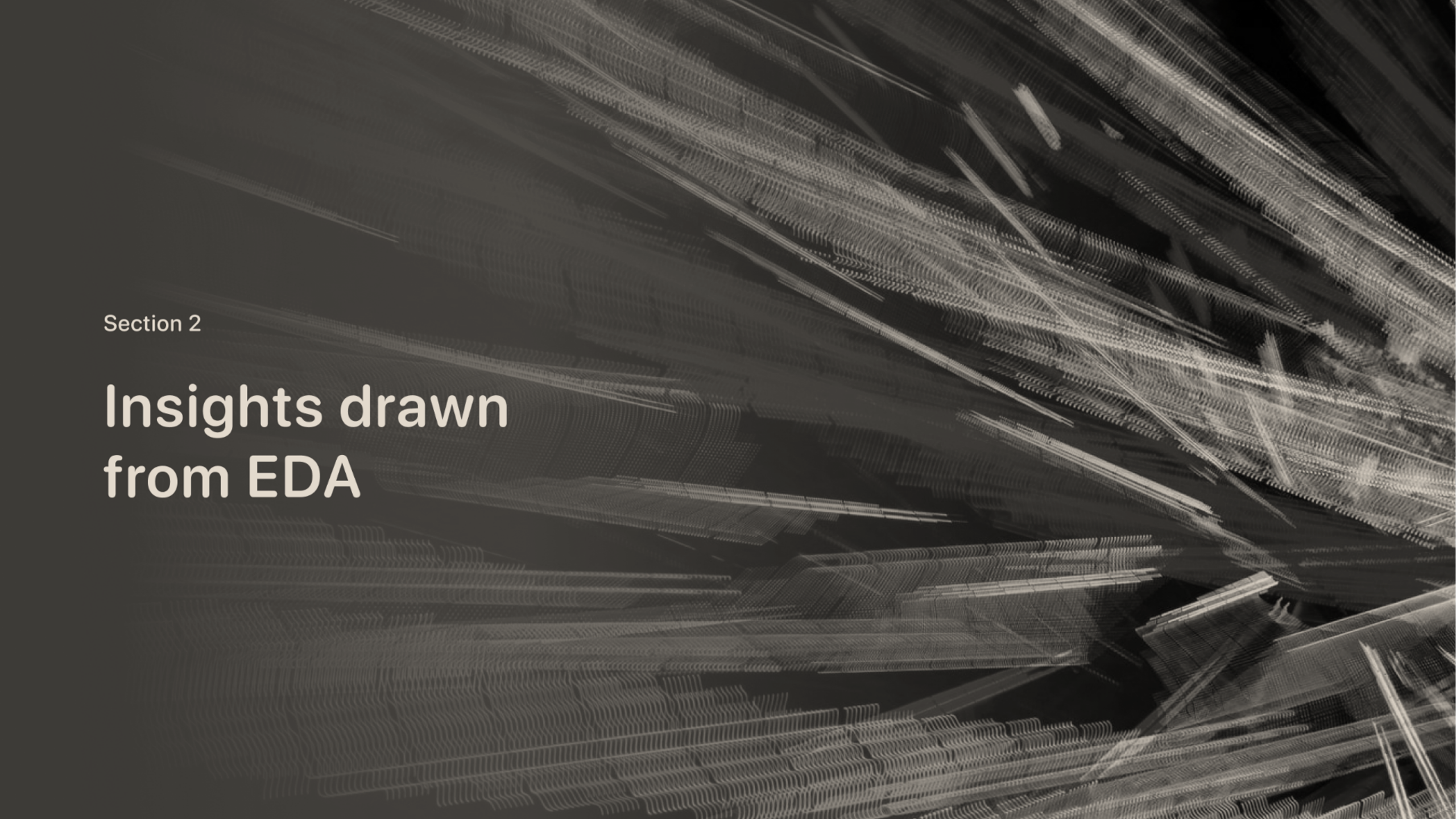

[Link to GitHub Notebook]

# RESULTS

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS



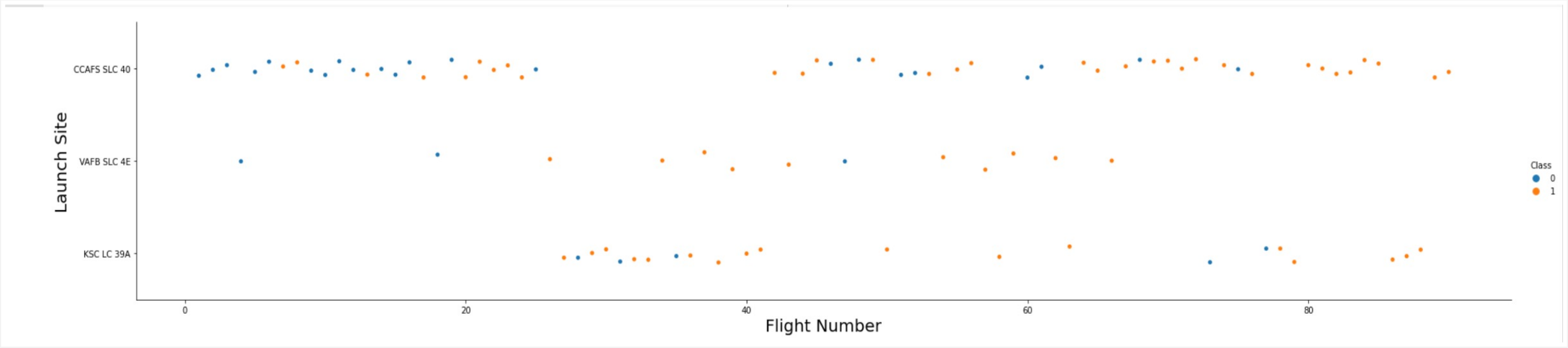
Section 2

# Insights drawn from EDA


EDA WITH DATA VISUALISATION

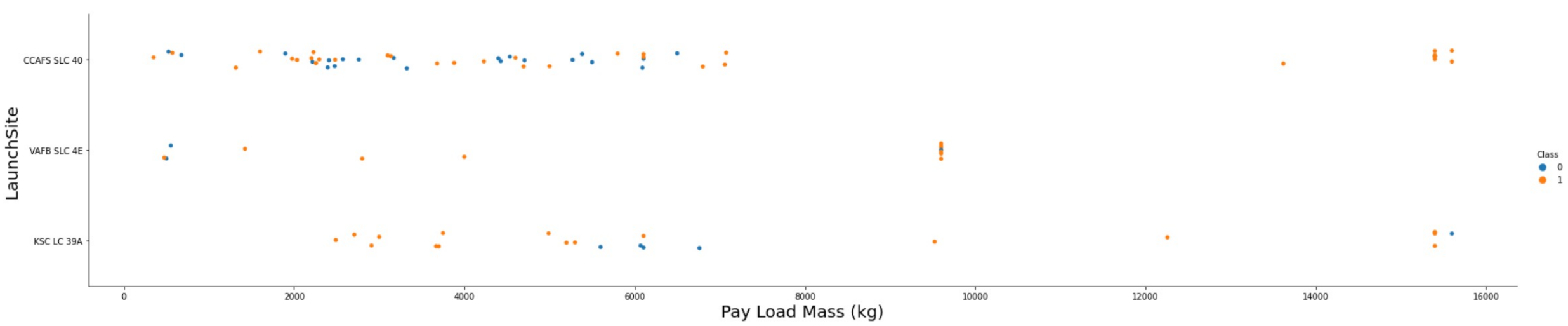
# FLIGHT NUMBER VS. LAUNCH SITE

When more rocket launches take place at a particular site, the number of successful launches also increases. CCAFS SLC 40 has the most launches so it correspondingly also has the most successful launches. Similarly, KSC LC 39A has the second most launches and hence the second most successful launches. Finally, VAFB SLC 4E has the least number of launches hence the least amount of successful launches



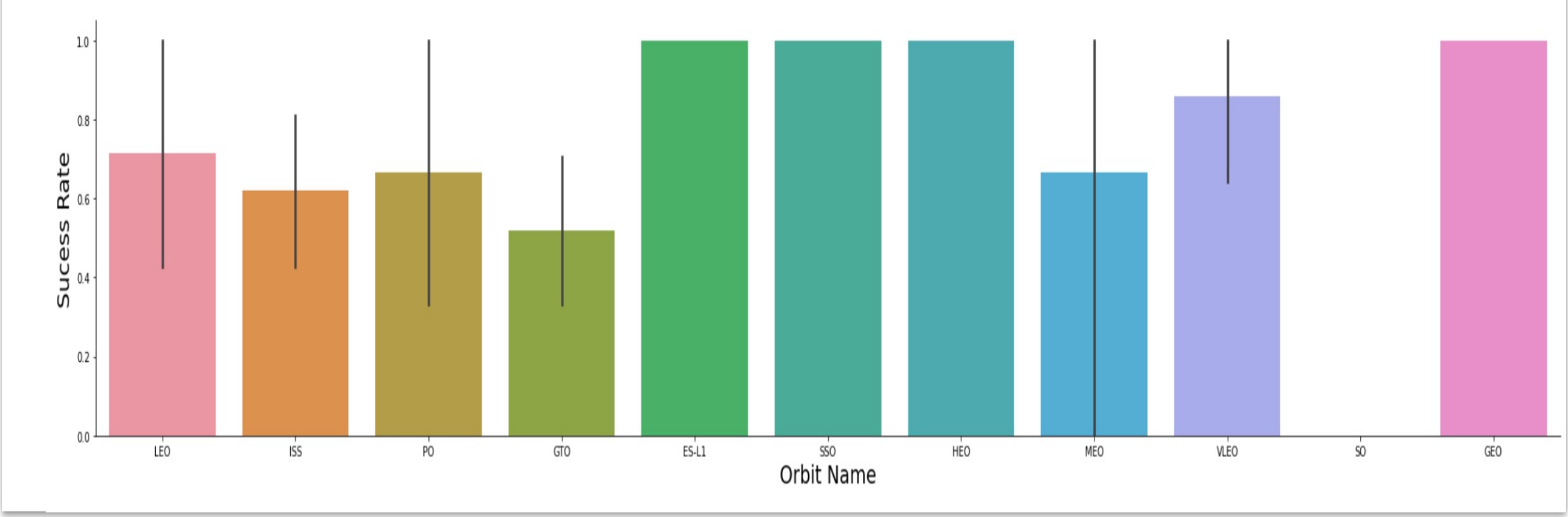
# PAYLOAD VS. LAUNCH SITE

There appears to be no discernible pattern between the launch site and pay load mass.



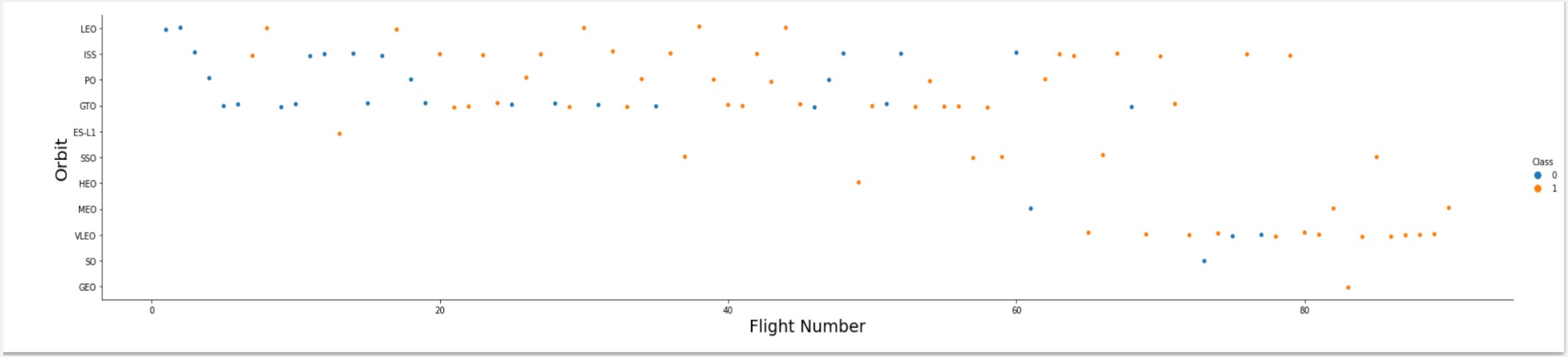
# SUCCESS RATE VS ORBIT TYPE

- The orbits ES-L1, SSO, HEO and GEO have the highest success rate.



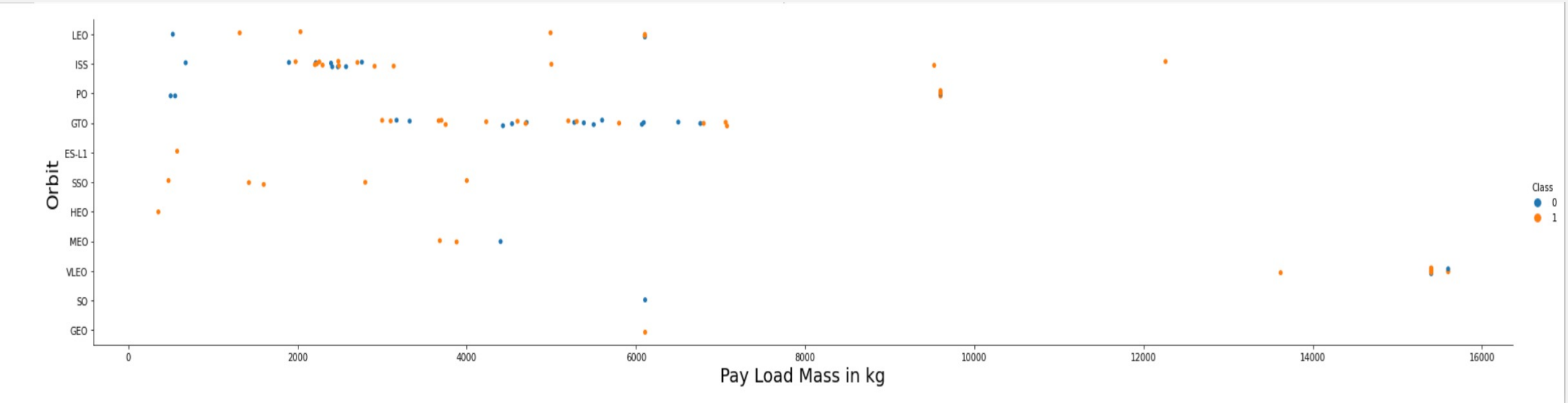
# FLIGHT NUMBER VS ORBIT TYPE

- In the LEO orbit, the success of the launch is related to the number of flights as beyond around 10 flights all the launches are successful. However, it appears there is no relationship between other orbits and the success of launch.



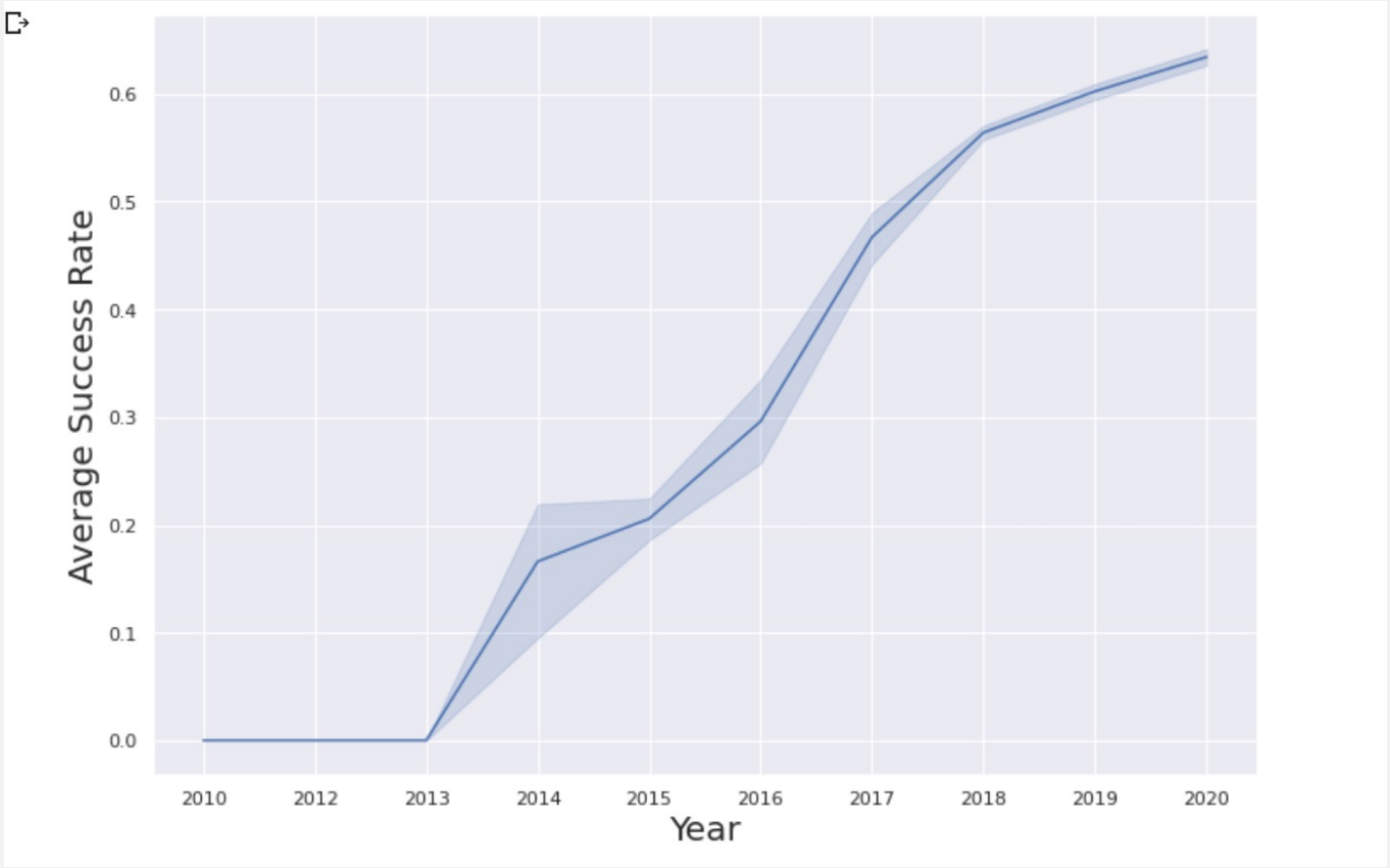
# PAYLOAD VS ORBIT TYPE

- Heavier payloads have a negative influence on GTO Orbits and a positive influence on ISS and GEO orbits.



# YEARLY SUCCESS TREND

- Success rate increases since 2013. Then, the rate of increase of success starts slowing down until the recent year of 2020.




EDA WITH SQL

# ALL LAUNCH SITE NAMES

```
SELECT DISTINCT LAUNCH_SITE FROM
FALCON9
```

Distinct gives us the unique launch sites.

| | LAUNCH_SITE |
|---|---|
| 0 | CCAFS LC-40 |
| 1 | CCAFS SLC-40 |
| 2 | KSC LC-39A |
| 3 | VAFB SLC-4E |

# LAUNCH SITE NAMES BEGIN WITH “CCA”

SELECT * FROM FALCON9 WHERE LAUNCH_SITE LIKE '%CCA%’ LIMIT 5

* Gives us full records, like %CCA% searches for string ‘CCA 'and LIMIT 5 displays 5 records.

| | DATE | TIME__UTC_ | BOOSTER_VERSION | LAUNCH_SITE |
|---|---|---|---|---|
| 0 | 2010-04-06 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 |
| 1 | 2010-08-12 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 |
| 2 | 2012-05-22 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 |
| 3 | 2012-08-10 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 |
| 4 | 2013-01-03 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 |

# TOTAL PAYLOAD MASS

```
SELECT SUM(PAYLOAD_MASS__KG_) as
payload_sum FROM FALCON9 WHERE
CUSTOMER='NASA (CRS)'
```

| | PAYLOAD_SUM |
|---|---|
| 0 | 45596 |

SUM() helps us calculate sum and keyword ‘as’ changes name of result column

# AVERAGE PAYLOAD MASS BY F9 V1.1

```
SELECT AVG(PAYLOAD_MASS__KG_) as payload_avg FROM FALCON9 WHERE BOOSTER_VERSION='F9 v1.1'
```

| | PAYLOAD_AVG |
|---|---|
| 0 | 2928 |

AVG() helps us calculate average and keyword ‘as’ changes name of result column

# FIRST SUCCESSFUL GROUND LANDING DATE

SELECT MIN(DATE) as FIRST_SUCCESS FROM FALCON9 WHERE LANDING__OUTCOME='Success (ground pad)'

| | FIRST_SUCCESS |
|---|---|
| 0 | 2015-12-22 |

# SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

```
SELECT BOOSTER_VERSION FROM
FALCON9 WHERE
LANDING__OUTCOME='Success (drone
ship)' AND PAYLOAD_MASS__KG_ > 4000
AND PAYLOAD_MASS__KG_ < 6000
```

Use AND gate twice for three mutually inclusive conditions.

| | BOOSTER_VERSION |
|---|---|
| 0 | F9 FT B1022 |
| 1 | F9 FT B1026 |
| 2 | F9 FT B1021.2 |
| 3 | F9 FT B1031.2 |

# TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

```
SELECT(SELECT Count(Mission_Outcome) from
FALCON9 where Mission_Outcome LIKE
'%Success%') as
Successful_Mission_Outcomes,(SELECT
Count(Mission_Outcome) from FALCON9 where
Mission_Outcome LIKE '%Failure%') as
Failure_Mission_Outcomes FROM FALCON9 LIMIT 1
```

| | SUCCESSFUL_MISSION_OUTCOMES | FAILURE_MISSION_OUTCOMES |
|---|---|---|
| 0 | 100 | 1 |

We use a nested select statement to combine results from two select queries.

# BOOSTERS CARRIED MAX PAYLOAD

```
SELECT DISTINCT BOOSTER_VERSION,
MAX(PAYLOAD_MASS__KG_) as
Max_Payload_Mass FROM FALCON9
GROUP BY BOOSTER_VERSION ORDER
BY Max_Payload_Mass DESC LIMIT 12
```

| | BOOSTER_VERSION | MAX_PAYLOAD_MASS |
|---|---|---|
| 0 | F9 B5 B1048.4 | 15600 |
| 1 | F9 B5 B1048.5 | 15600 |
| 2 | F9 B5 B1049.4 | 15600 |
| 3 | F9 B5 B1049.5 | 15600 |
| 4 | F9 B5 B1049.7 | 15600 |
| 5 | F9 B5 B1051.3 | 15600 |
| 6 | F9 B5 B1051.4 | 15600 |
| 7 | F9 B5 B1051.6 | 15600 |
| 8 | F9 B5 B1056.4 | 15600 |
| 9 | F9 B5 B1058.3 | 15600 |
| 10 | F9 B5 B1060.2 | 15600 |
| 11 | F9 B5 B1060.3 | 15600 |

ORDER BY DESC helps us get the maximum numbers which are the first 12 entries and that's why we use LIMIT 12

# 2015 LAUNCH OUTCOMES

```
SELECT DATE,
LANDING__OUTCOME,
BOOSTER_VERSION,
LAUNCH_SITE FROM FALCON9
WHERE LANDING__OUTCOME
= 'Failure (drone ship)' AND DATE
LIKE '%2015%'
```

| | DATE | LANDING__OUTCOME | BOOSTER_VERSION | LAUNCH_SITE |
|---|---|---|---|---|
| 0 | 2015-10-01 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 1 | 2015-04-14 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

# RANK LANDING OUTCOMES BETWEEN 2010-06-04 AND 2017-03-20

```
SELECT LANDING__OUTCOME, COUNT(*)
FROM FALCON9 WHERE DATE > '20100604'
AND DATE < '20170320' GROUP BY
LANDING__OUTCOME ORDER BY 2 DESC
```

| | LANDING__OUTCOME | 2 |
|---|---|---|
| 0 | No attempt | 10 |
| 1 | Failure (drone ship) | 5 |
| 2 | Success (drone ship) | 5 |
| 3 | Success (ground pad) | 5 |
| 4 | Controlled (ocean) | 3 |
| 5 | Uncontrolled (ocean) | 2 |
| 6 | Failure (parachute) | 1 |
| 7 | Precluded (drone ship) | 1 |

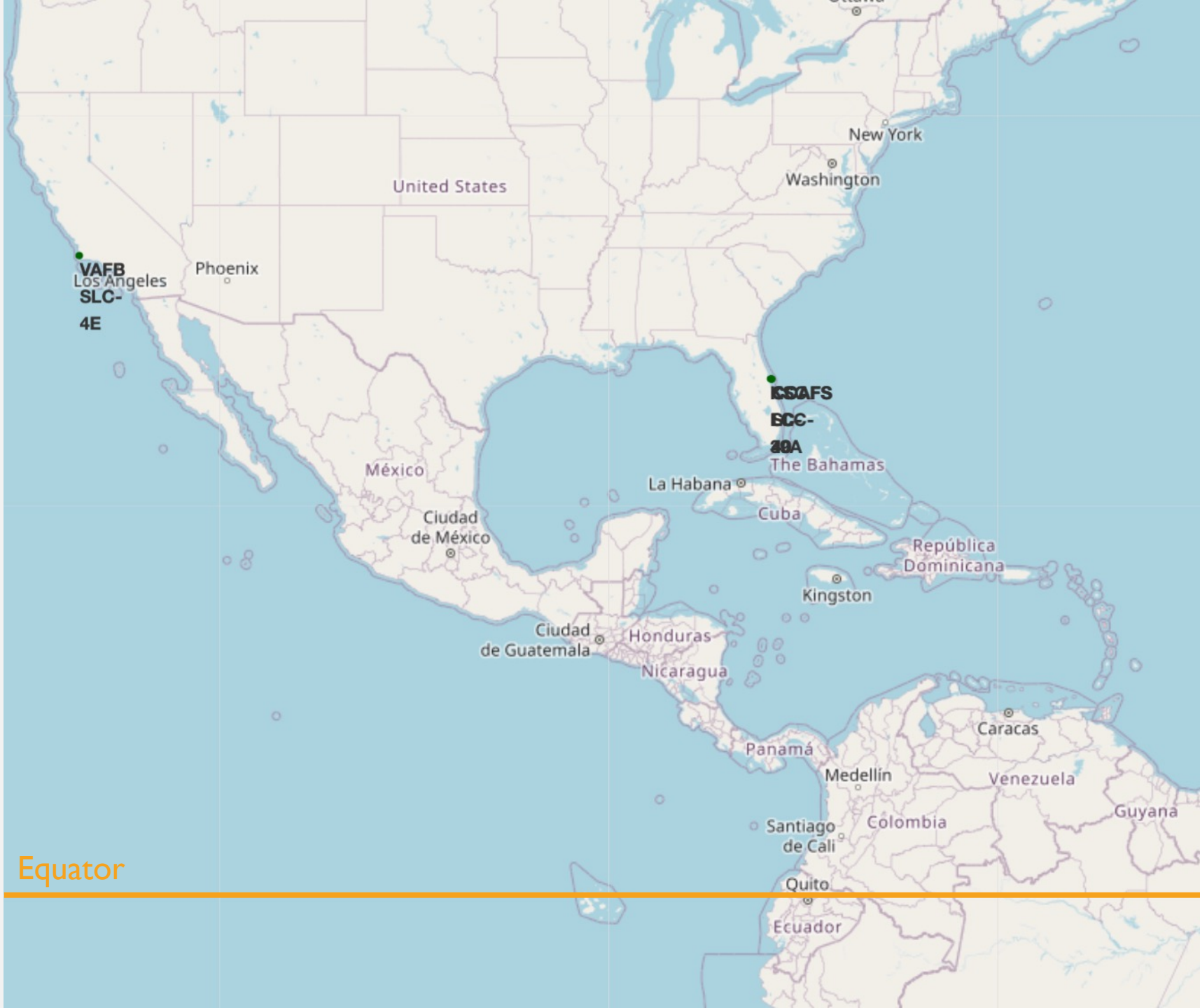
Section 4

# Launch Sites Proximities Analysis

# MAP OF SPACEX FALCON9 LAUNCHSITES

1. The launch sites are not close to the equator which is orange line as shown.

2. The launch sites are near the coasts of America: one site on the West Coast and two sites on the East Coast.

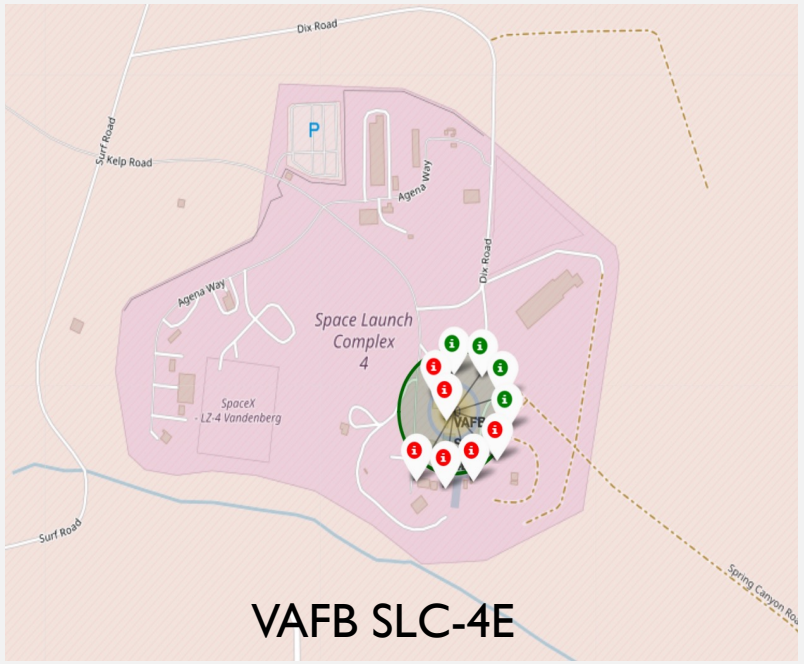


# MAP OF SUCCESSFUL VS UNSUCCESSFUL LAUNCHES AT EACH SITE

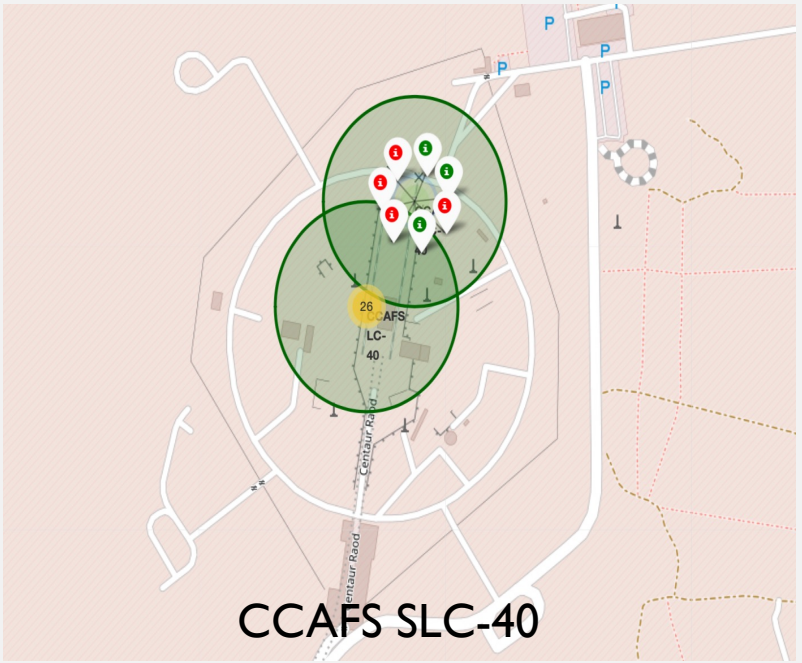
1. The red markers show an unsuccessful launch while the green markers show a successful launch.

2. As we can see from the images, launches from KSC LC 39-A have a greater rate of success than launches at CCAFS LC-40 just by the density of red vs green markers.

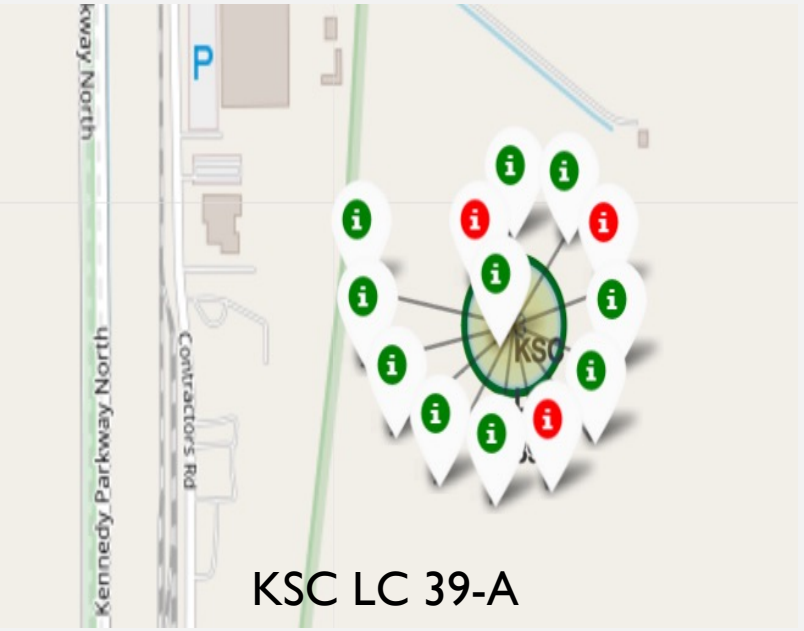
3. CCAFS SLC-40 and VAFB SLC-4E have more unsuccessful launches but by a much lesser degree in comparison to CCAFS LC-40



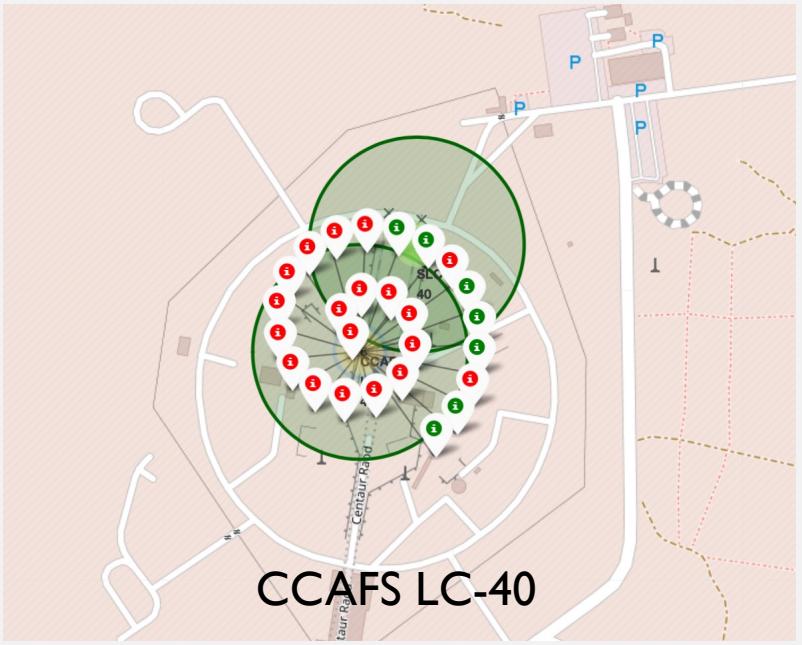
VAFB SLC-4E



CCAFS SLC-40

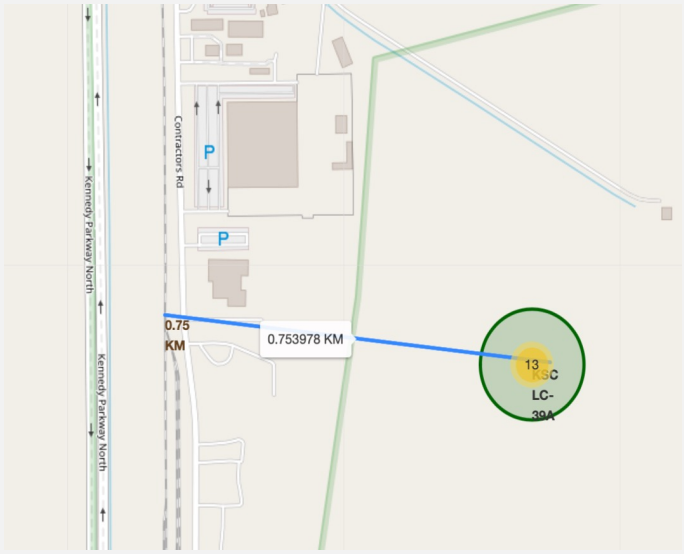


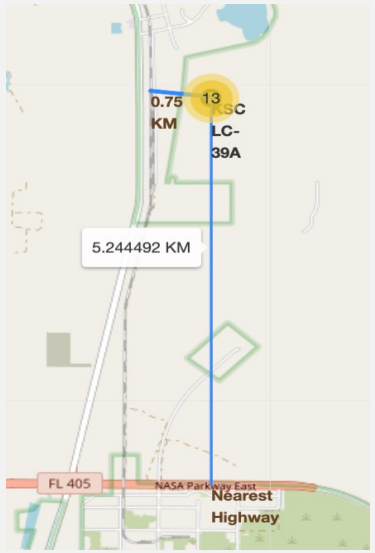
KSC LC 39-A



CCAFS LC-40

# MAPS OF DISTANCE BETWEEN A LAUNCH SITE AND ITS PROXIMITIES



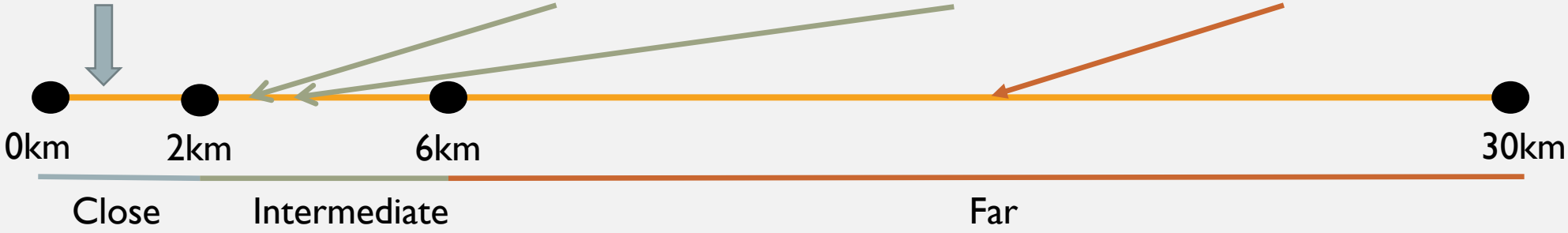


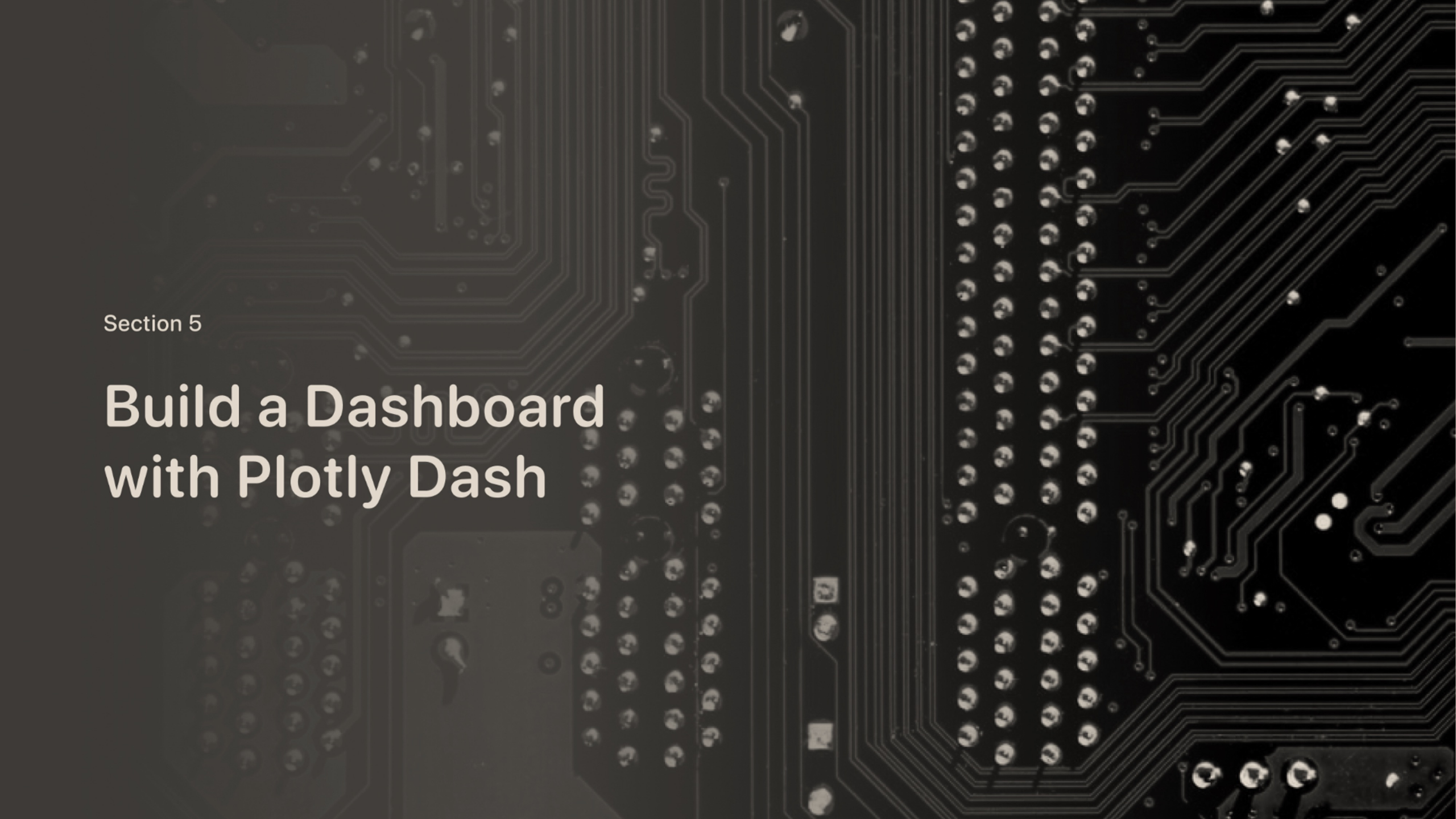

Location: Railway Tracks
Distance: 0.75 km
Proximity: Close

Location: Highway
Distance: 5.24 km
Proximity: Intermediate

Location: Coastline
Distance: 3.91 km
Proximity: Intermediate

Location: Closest City (Cape Canaveral)
Distance: 19.53 km
Proximity: Far



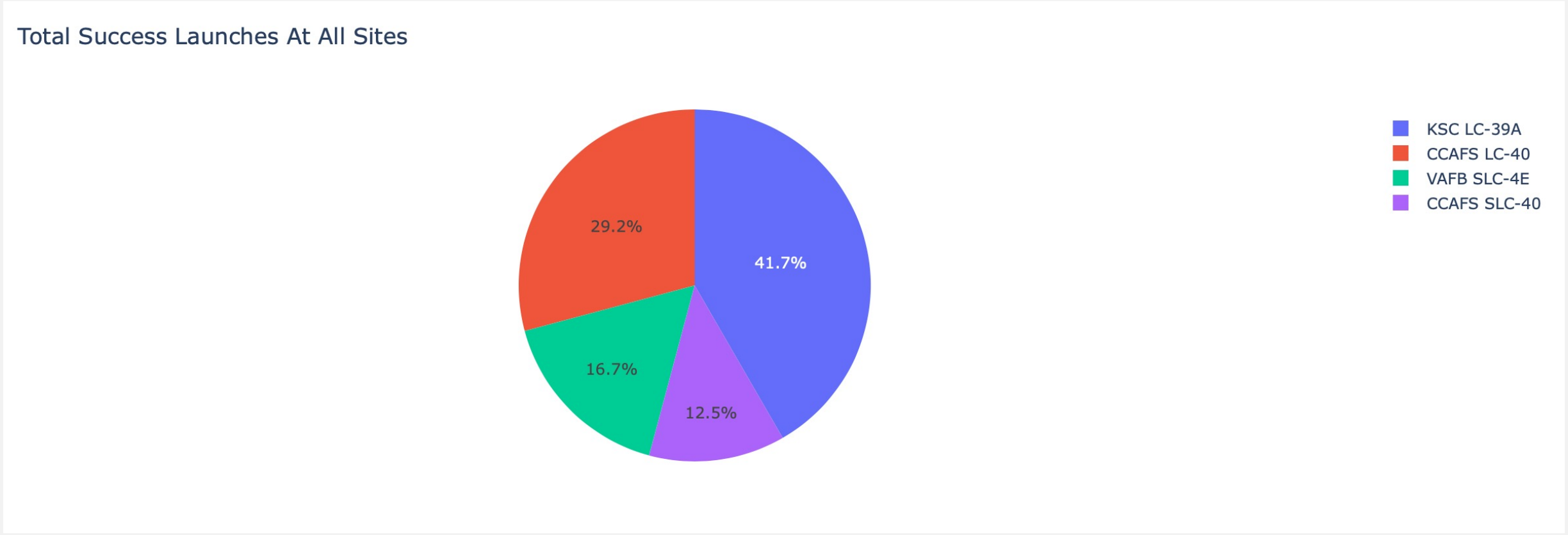


Section 5

# Build a Dashboard with Plotly Dash

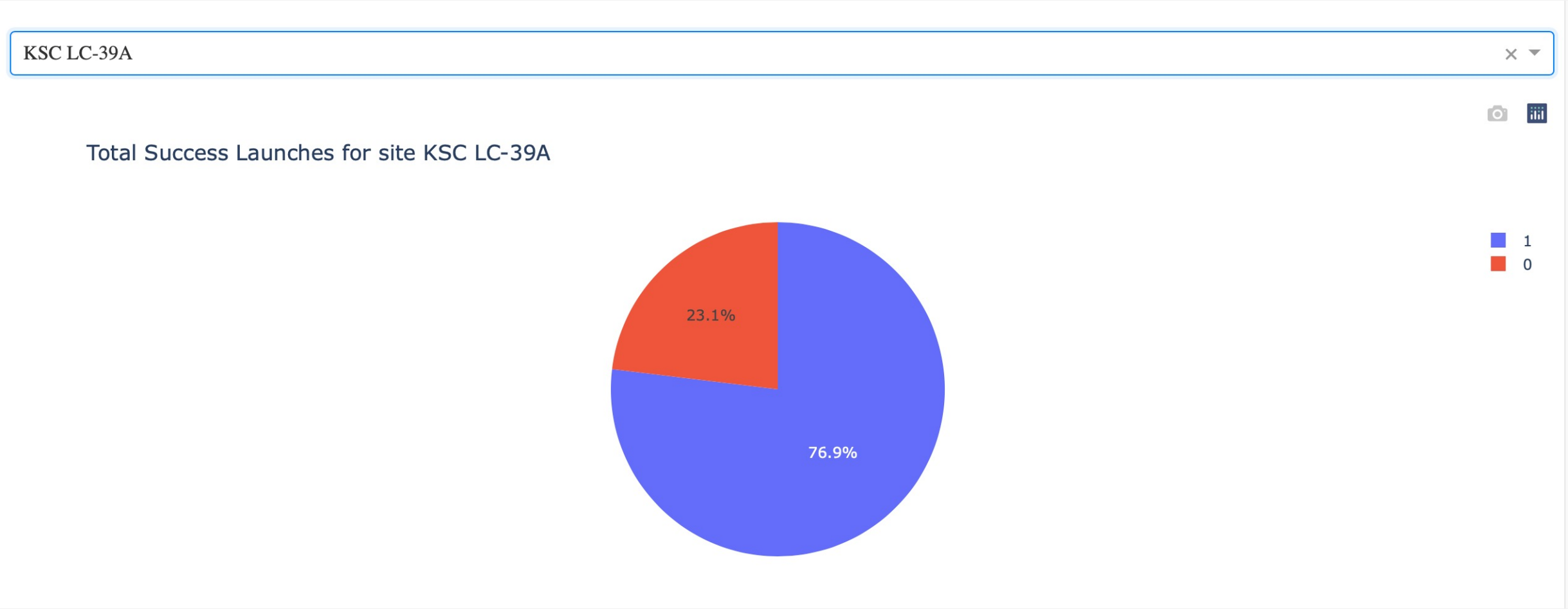
# PIE CHART FOR SUCCESSES BY LAUNCH SITE

KSC LC-39A is the launch site with the most successful launches and accounts for 41.7% of the launches. CCAFS LC-40, VAFB SLC-4E and CCAFS SLC-40 have the second, third most successful and the least successful number of launches, respectively.



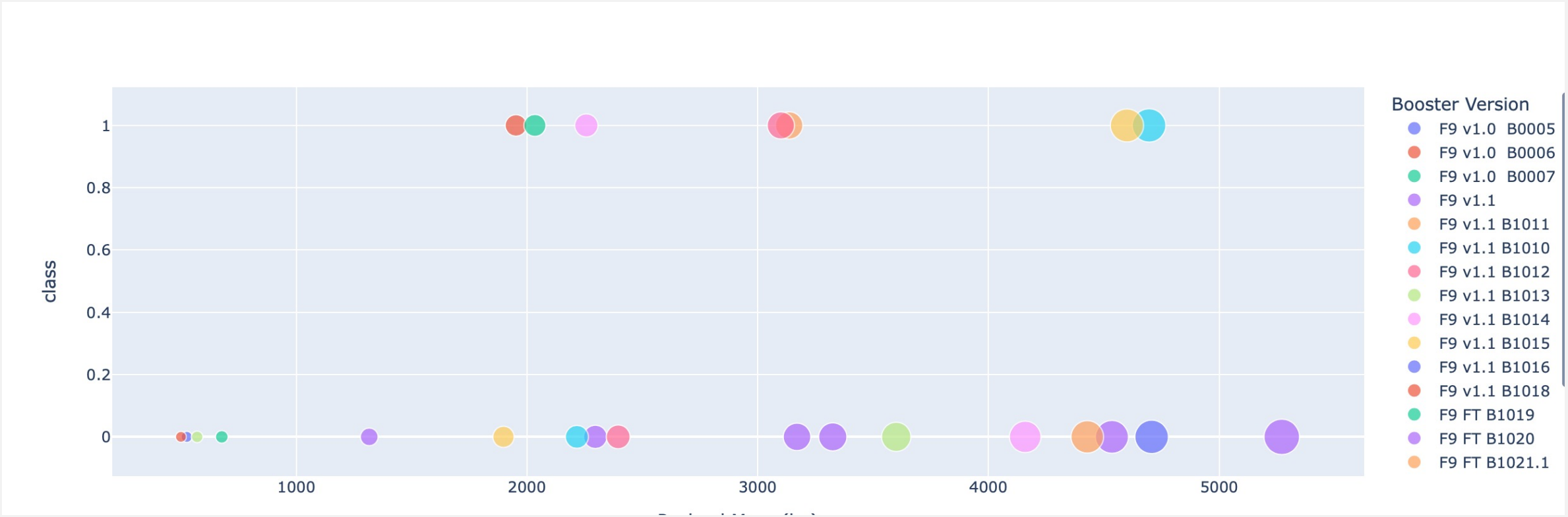
# PIE CHART FOR HIGHEST LAUNCH SUCCESS RATIO

KSC LC-39A is the launch site with the highest success ratio of about 3.33 i.e., success/failure = 76.9/23.1



# PAYLOAD VS LAUNCH OUTCOME SCATTER PLOT

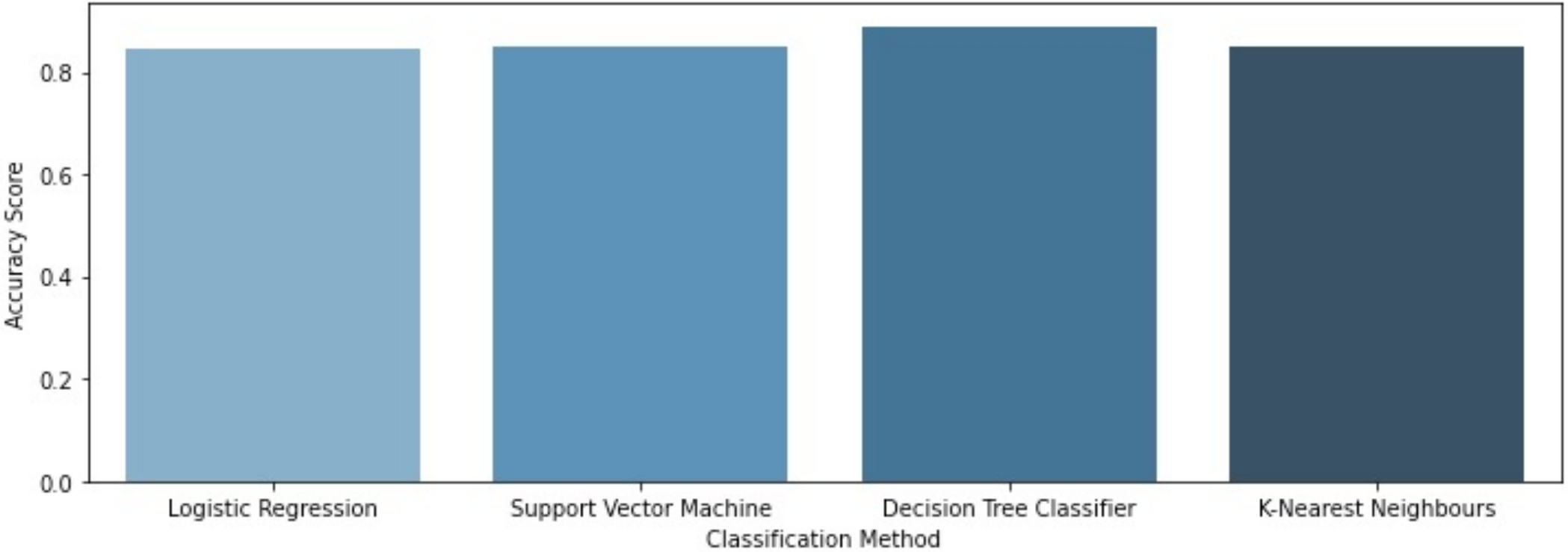
Launches are more successful at lower payloads



Section 6

# Predictive Analysis (Classification)

# CLASSIFICATION ACCURACY



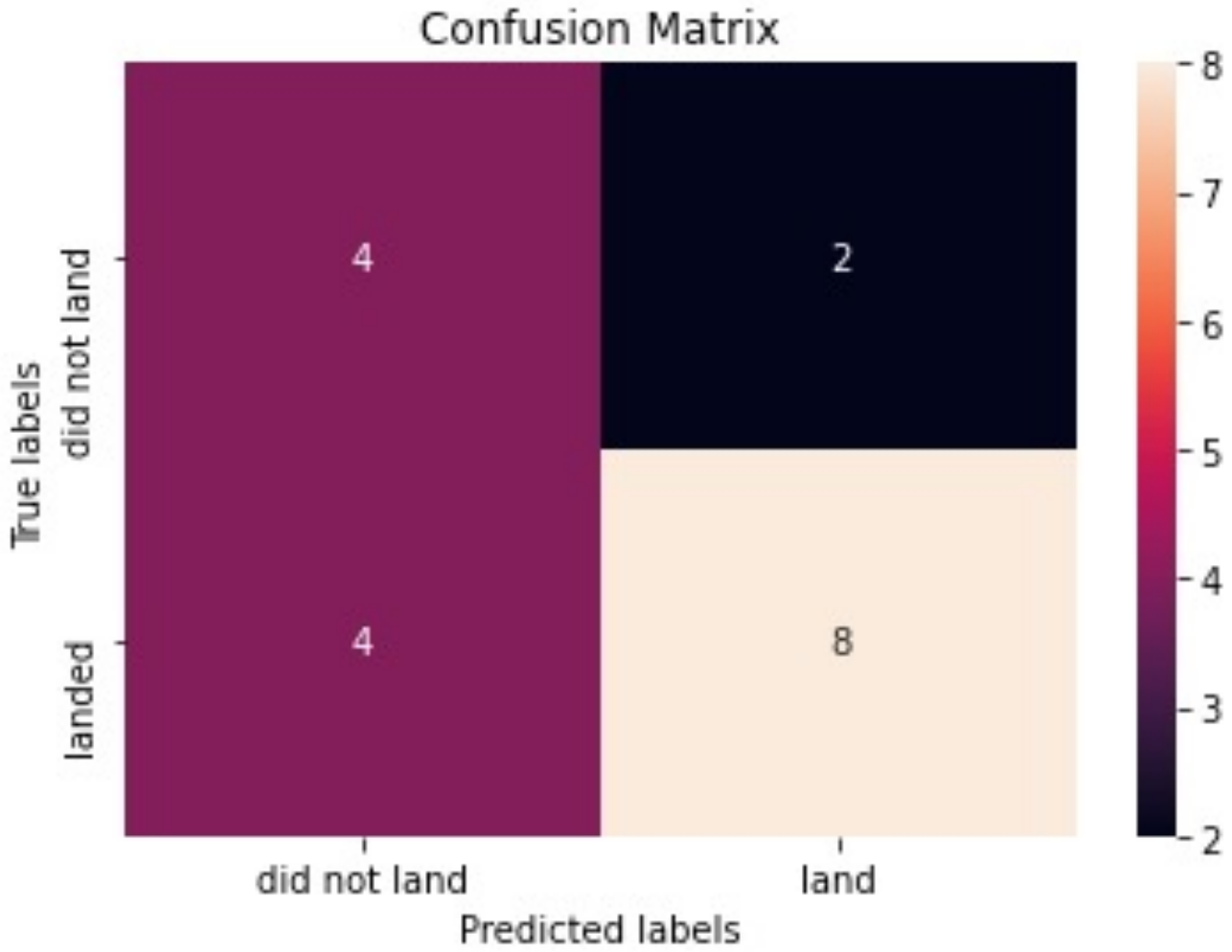

Decision tree classifier is the best classification method with a best score accuracy of about 88.9%.

# CONFUSION MATRIX FOR BEST PERFORMING CLASSIFICATION METHOD (TREE CLASSIFIER)

**Define the following**:

1. True Negatives: 4
2. False Negatives: 4
3. False Positives: 2
4. True Positives: 8

The tree model is generally accurate as there are more correct results (12) than false (6) results,. There are false negatives (4) than there are false positives (2) which is what we want as a false negative i.e., the rocket lands when it was predicted not to land as we do not waste a tremendous amount of money with a rocket. We want as less as false negatives as possible as not landing when predicted to land is a dangerous outcome.




Confusion Matrix

Key:

| | Actually Positive (1) | Actually Negative (0) |
|---|---|---|
| Predicted Positive (1) | True Positives (TPs) | False Positives (FPs) |
| Predicted Negative (0) | False Negatives (FNs) | True Negatives (TNs) |

# Conclusion Part 1

Orbit CCAFS SLC 40 appears to have the highest raw number of successful launches because it has the highest number of launches.

The orbits ES-LI, SSO, HEO and GEO have the highest success rate.

Heavier payloads have a positive influence on ISS and GEO orbits, negative influence on GTO orbits.

KSC LC 39-A is shown to have the absolute highest rate of success from the map.

# Conclusion Part II

Launch sites are close to railway lines so that material can be transported

Launch sites are close to coastlines so that if a crash occurs it can safely crash into the sea

Launch sites are far away from cities to avoid crash risks, high noise pollution and electronic interference from high concentration of networked devices in cities.

KSC LC-39A has the highest success ratio.

# Conclusion Part III

Launches are more successful at lower payloads, about 2000 to 4500 kg range is optimum.

Decision tree classifier is the best classification method for machine learning in this dataset

Decision tree classifier has more true positives than false positives by a ratio of 4:1

# APPENDIX

- GitHub Reference: https://github.com/oasisbeatle/Capstone-Project-IBM

# Thank you!