IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Chandan Singh

18th Oct 2023

# OUTLINE

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

# EXECUTIVE SUMMARY

- **Summary of methodologies-**

  In this project, I started by collecting data using APIs and web scraping. After that, I cleaned up the data to prepare it for use in machine learning. Then, I dug into the data using SQL to find interesting patterns. I made the findings easy to understand by creating visual representations of the data. I also made interactive maps to showcase the information. Finally, I used machine learning to make predictions based on the data.

- **Summary of all results-**

  The project extracted insights through Exploratory Data Analysis, presented information via screenshots, and generated predictions using predictive analytics.

# INTRODUCTION

- **Project background and context-**

  SpaceX promotes Falcon 9 rocket launches on its website, priced at 62 million dollars, while other providers charge over 165 million dollars. The cost difference is mainly because SpaceX can reuse the first stage of the rocket. If we can predict whether the first stage will land properly, we can estimate the launch cost. This prediction is valuable for other companies bidding against SpaceX for a rocket launch. The project's aim is to build a machine learning process that forecasts the success of the first stage landing.

- **Problems you want to find answers-**

  Which elements influence a successful rocket landing?

  How do different features interact to affect the landing success rate?

  What operational conditions are necessary to guarantee a successful landing program?

Section 1

# Methodology

# METHODOLOGY

## Executive Summary

- Data collection methodology:
  - Data was collected through SpaceX data collection API and by webscrapping of Wikipedia.
- Performed data wrangling
  - Checked for null values and data types that can hinder ML process. Scaled the data for smooth gradient descent. Did one-hot encoding of categorical data.
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Data was splitted into train and test data. Train data was used for training of model and test data was used for the testing the accuracy of model on unseen data

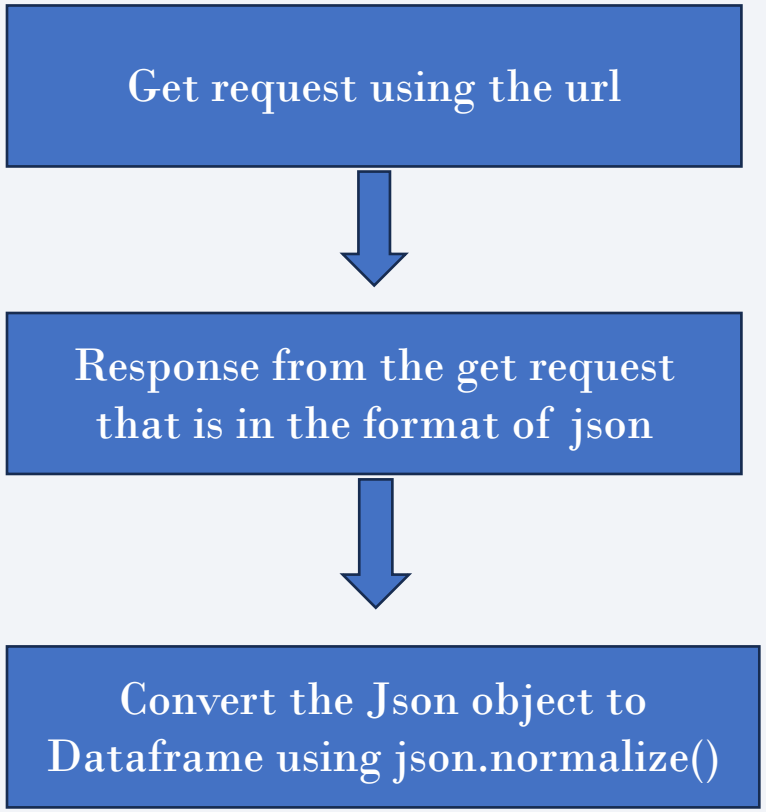
# DATA COLLECTION

**The process is explained step by step  as follows:**

- Data collection: Utilized a GET request to SpaceX API.
- Data transformation: Decoded API response using .json() and transformed it into a pandas dataframe using .json_normalize().
- Web scraping: Extracted Falcon 9 launch records from Wikipedia using BeautifulSoup.
- Web data transformation: Parsed the HTML table and converted it into a pandas dataframe for analysis.

# DATA COLLECTION – SPACEX API

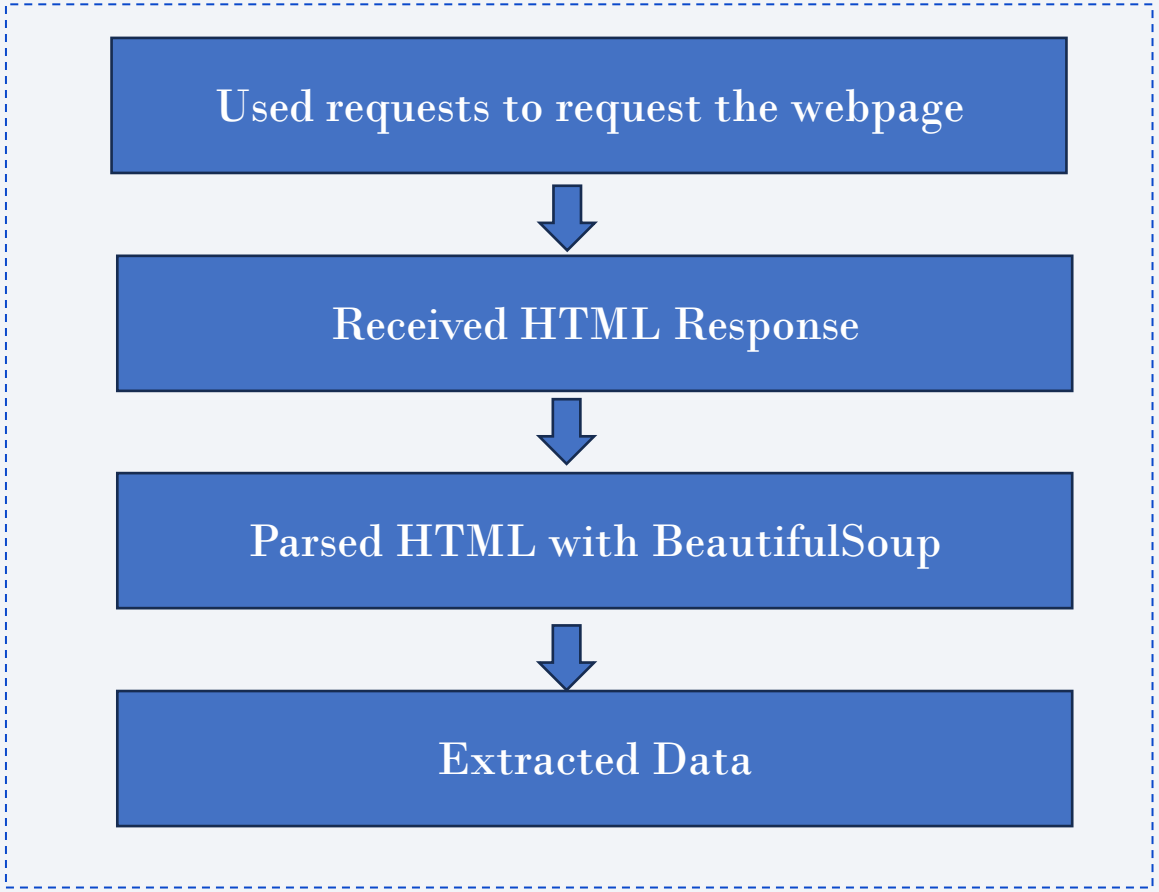
- We employed a GET request for the rocket launch API, which provided a JSON-formatted response. Subsequently, we utilized the json_normalize method to convert the response into a dataframe.
- The link to my notebook is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/spacex-data-collection-api%20.ipynb*



# DATA COLLECTION - SCRAPING

- Present your web scraping process using key phrases and flowcharts

- The link to my notebook is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/Webscraping.ipynb*



# Data Wrangling

- Computed launch counts for each launch site.
- Analyzed the frequency of each orbit type.
- Examined the occurrence of mission outcomes based on orbit types.
- Generated a landing outcome label using the "Outcome" column.
- My notebook link is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/spacex-data_wrangling_jupyterlite.jupyterlite.ipynb*

# EDA with Data Visualization

- We examined the data through visualizations, investigating correlations between flight number and launch site, payload and launch site, success rates across orbit types, flight number and orbit type, as well as the annual trend in launch success.
- We used catplots, scatter plots, bar charts, and line charts to visualize the data for their specific advantages in showcasing categorical relationships, variable correlations, category comparisons, and trends over time, respectively.
- The link to my notebook is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/eda-dataviz.ipynb.jupyterlite.ipynb*

# EDA WITH SQL

Employed SQL for exploratory data analysis (EDA), where we utilized queries to uncover insights, including:

- Identifying unique launch site names in space missions.
- Calculating the total payload mass of NASA's (CRS) launched boosters.
- Determining the average payload mass for booster version F9 v1.1.
- Counting successful and failed mission outcomes.
- Identifying failed landing outcomes on drone ships, along with associated booster versions and launch site names.
- My notebook link is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/eda-sql-coursera_sqllite.ipynb*

# BUILD AN INTERACTIVE MAP WITH FOLIUM

- Marked all launch sites and incorporated map elements like markers, circles, and lines to represent launch outcomes (success or failure) for each site on the folium map.
- Categorized launch outcomes as class 0 for failure and class 1 for success.
- Utilized color-coded marker clusters to identify launch sites with notable success rates.
- Computed distances between launch sites and nearby features, addressing inquiries such as proximity to railways, highways, coastlines, and distance from urban areas.
- The link to my notebook is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/eda-sql-coursera_sqllite.ipynb*

# BUILD A DASHBOARD WITH PLOTLY DASH

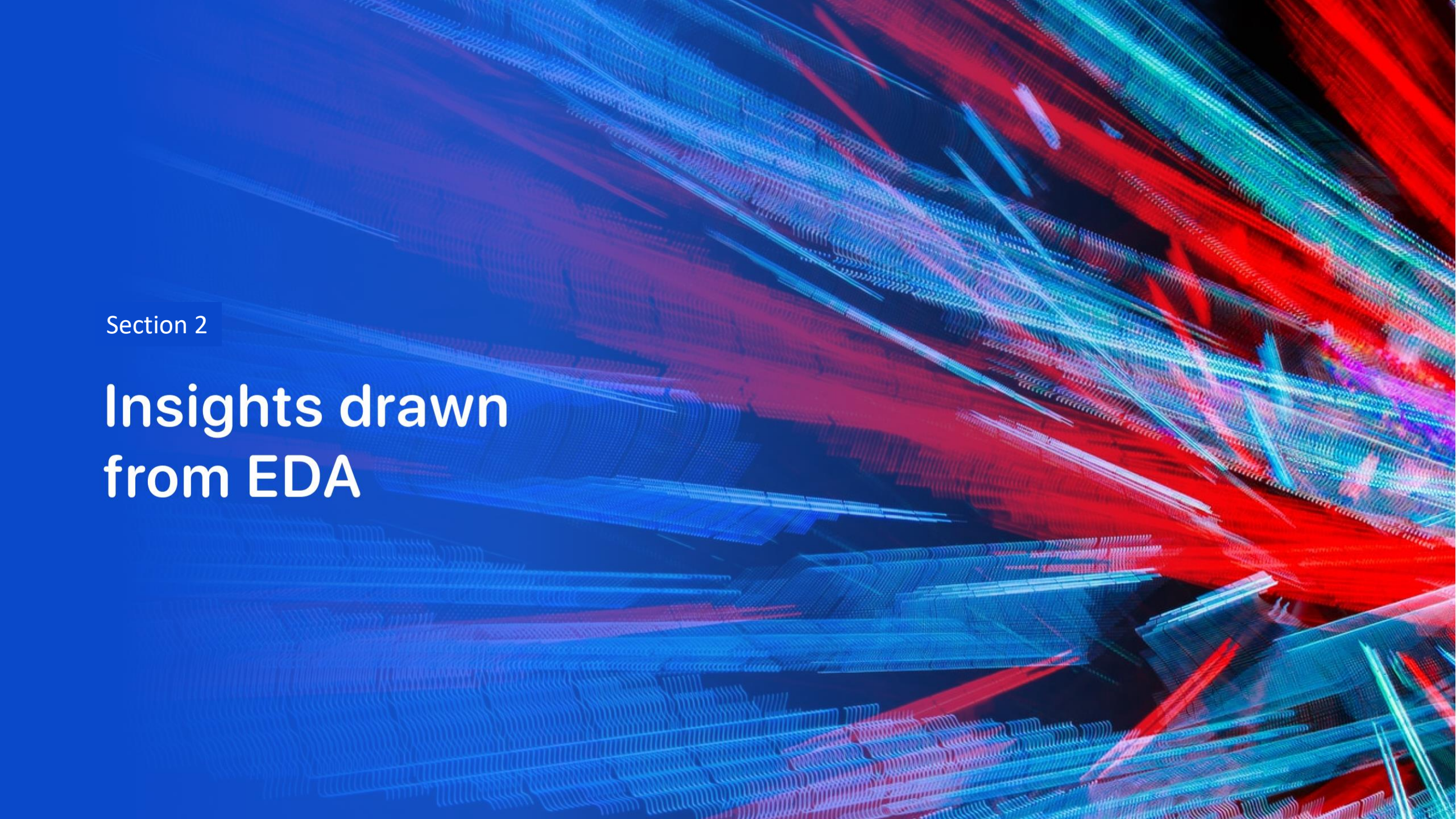
- Constructed an interactive dashboard using Plotly Dash.
- Generated pie charts illustrating the overall launches categorized by specific sites.
- Created scatter plots to display the correlation between Outcome and Payload Mass (Kg) across various booster versions.

# PREDICTIVE ANALYSIS (CLASSIFICATION)

- Utilized numpy and pandas for data loading and transformation.
- Conducted data splitting into training and testing subsets.
- Constructed varied machine learning models, tuning hyperparameters via GridSearchCV.
- Employed accuracy as the primary metric, enhancing the model through feature engineering and algorithm fine-tuning.
- Determined the top-performing classification model, identifying Decision tree as the most accurate.
- The link to my notebook is: *https://github.com/chandan160/IBM-Data-Science-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb*

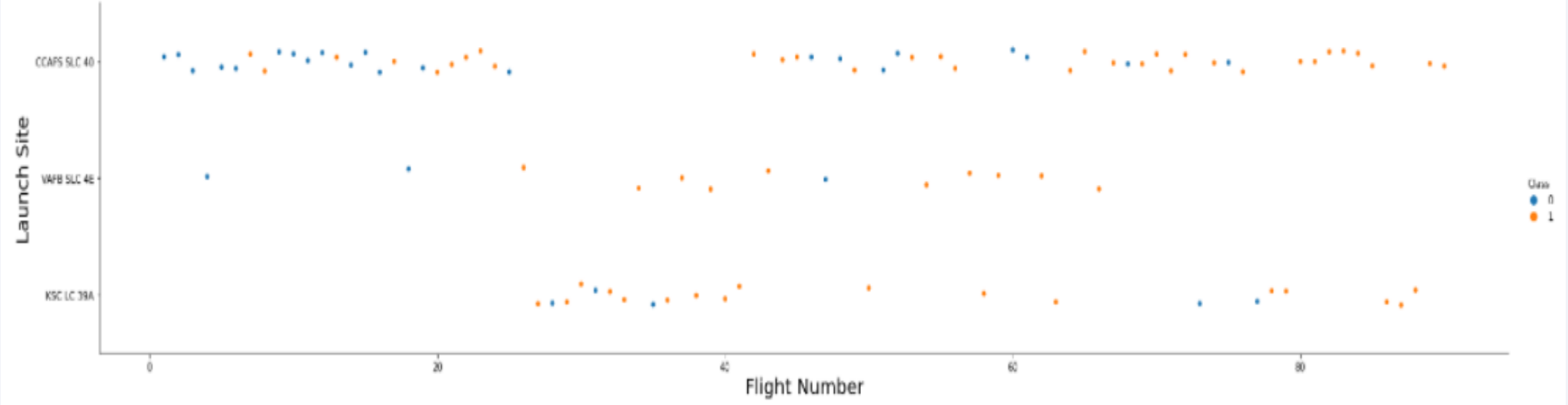
# RESULTS

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


Section 2
Insights drawn from EDA

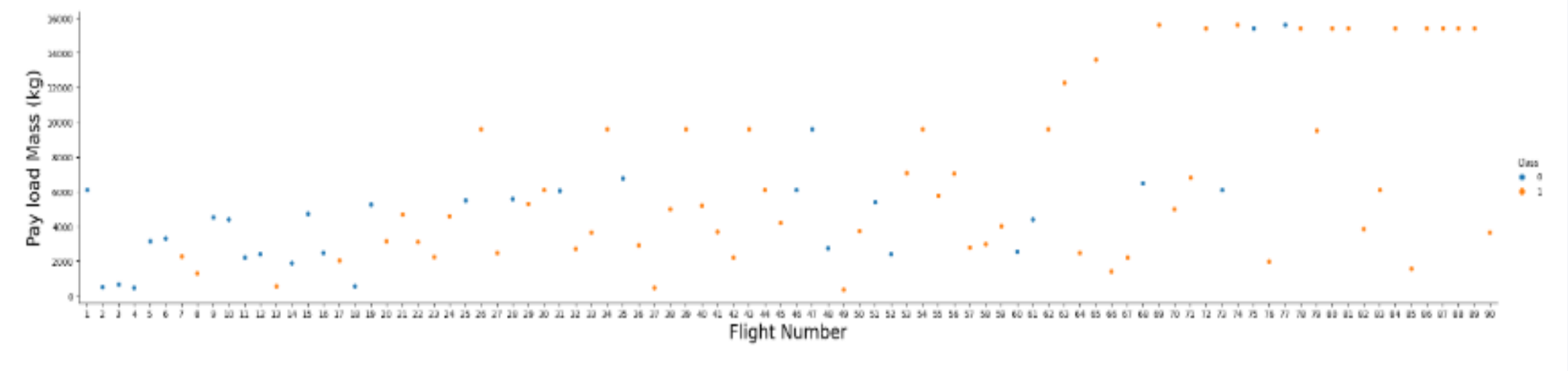
# FLIGHT NUMBER VS. LAUNCH SITE

- An observed trend indicates that success rates tend to increase with higher flight numbers at a given launch site. This suggests that larger flight numbers correspond to newer rockets equipped with enhanced and more reliable technologies.



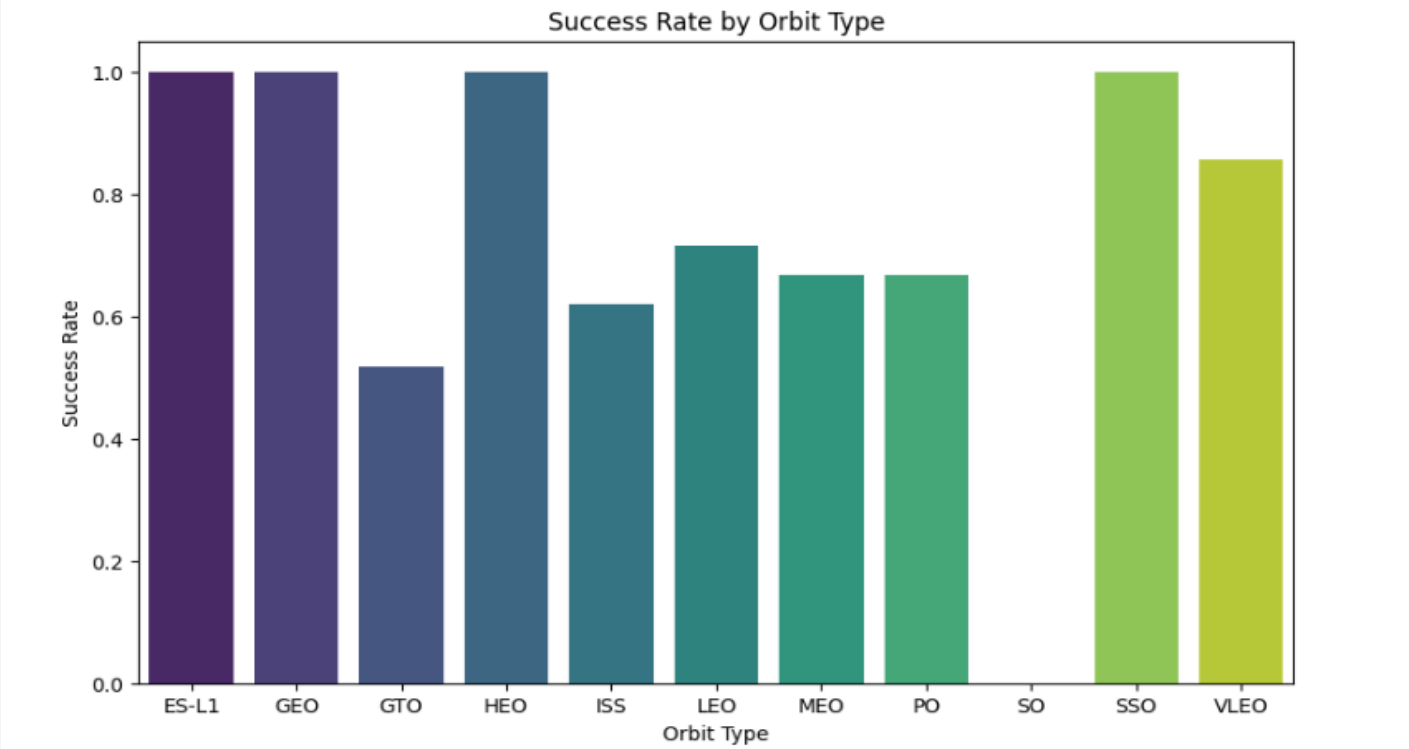
# PAYLOAD VS. LAUNCH SITE

- It was found that success rate is proportional to the Payload mass. higher payload mass necessitates advancements in design, technology, propulsion, and structural strength—all of which collectively contribute to a higher likelihood of success in reusing a SpaceX Falcon 9 rocket.



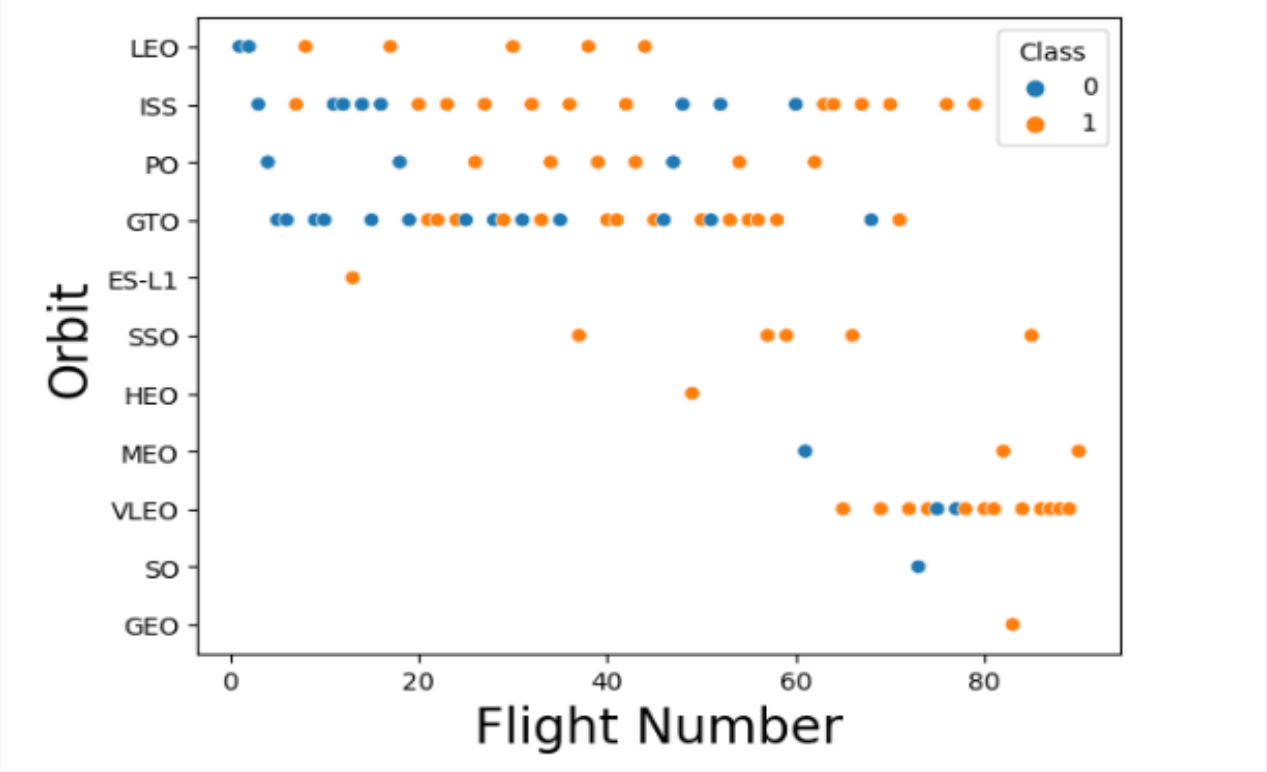
# SUCCESS RATE VS. ORBIT TYPE

- From the Bar graph it is clear that orbits ES-L1, GEO, HEO and SSO has the highest success rate.



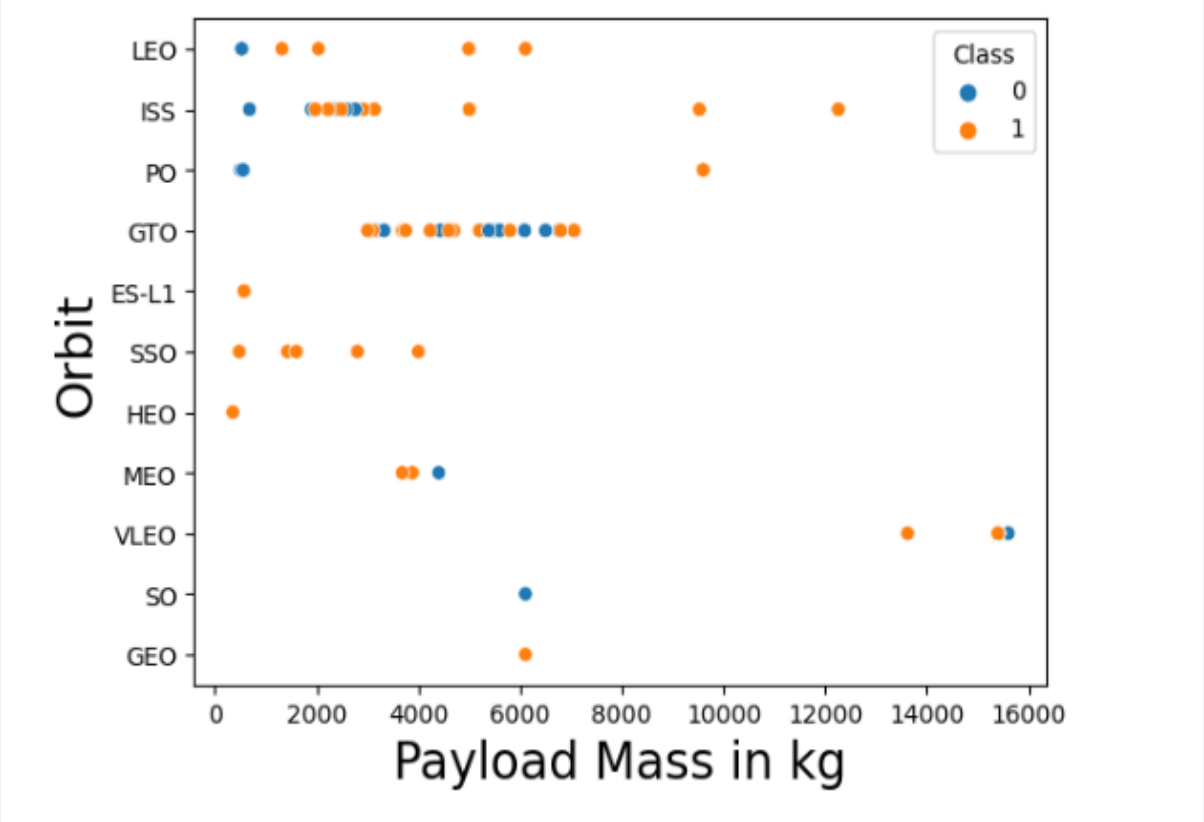
# FLIGHT NUMBER VS. ORBIT TYPE

- The graph displayed depicts Flight Number against Orbit type. It's evident that success in the LEO orbit is linked to flight frequency, whereas such a correlation is absent in the GTO orbit.



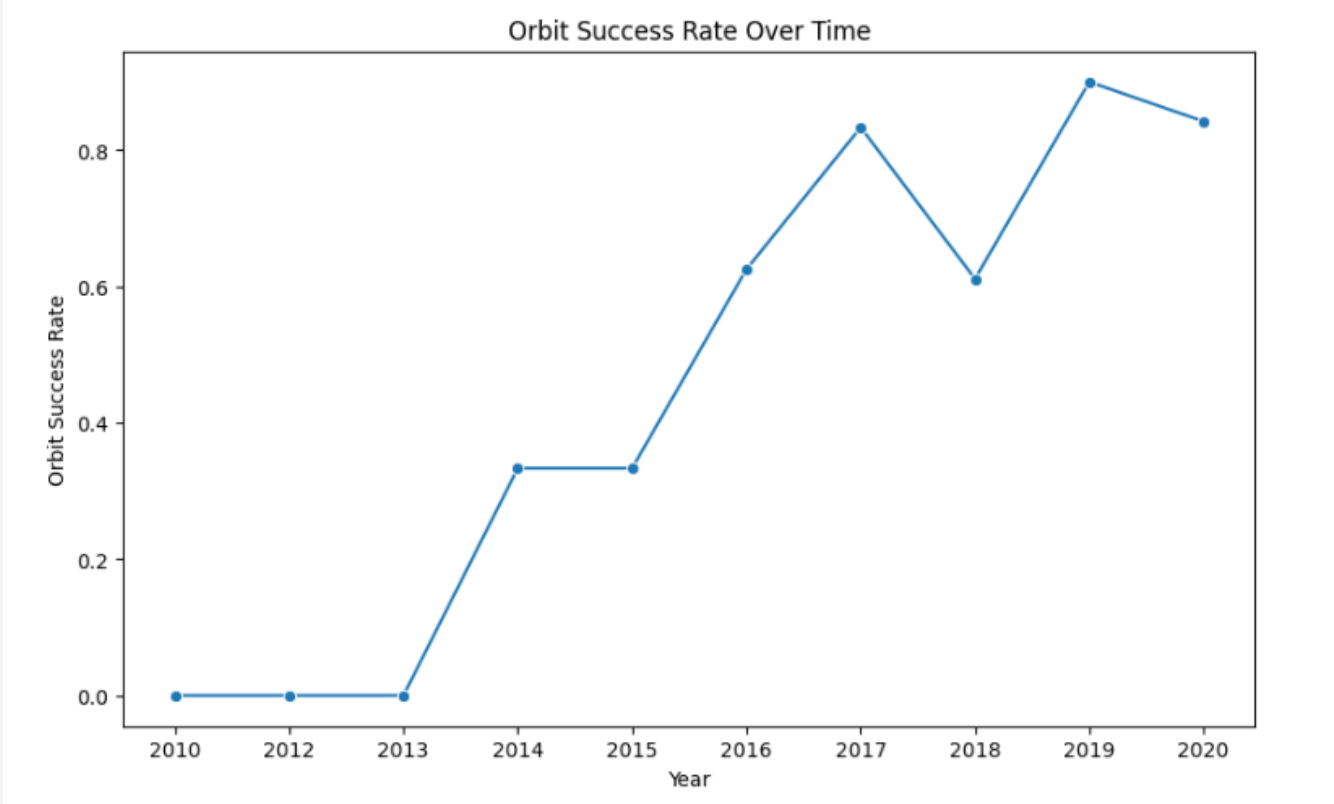
# Payload Vs. Orbit Type

- It is evident from the plot that the success rate for launches for LEO, ISS and PO orbits are proportional to payload mass.



# LAUNCH SUCCESS YEARLY TREND

- It can be observed that the success rate has increased from 2013 till 2020.



# ALL LAUNCH SITE NAMES

- We used the key word DISTINCT to show only unique launch sites from the SpaceX data. The unique launch sites are as follows-

| Launch_Site | count |
|---|---|
| CCAFS LC-40 | 26 |
| CCAFS SLC-40 | 34 |
| KSC LC-39A | 25 |
| VAFB SLC-4E | 16 |

# LAUNCH SITE NAMES BEGIN WITH 'CCA'

- We used the LIKE operator and WHERE clause to find the Launch site that begin with CCA.

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

[10]:
```
%%sql
SELECT * FROM SPACEXTABLE
WHERE "Launch_Site" LIKE 'CCA%'
LIMIT 5;
```

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

[10]:

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

# TOTAL PAYLOAD MASS

- The total payload carried by boosters from NASA is 45596 which was calculated with the following query:

```
%%sql
SELECT SUM("PAYLOAD_MASS__KG_") as total_payload_mass
FROM SPACEXTABLE
WHERE "Customer" = 'NASA (CRS)';
```

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

| total_payload_mass |
|---|
| 45596 |

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4. which was calculated using the following query:

```
%%sql
SELECT AVG("PAYLOAD_MASS__KG_") as average_payload_mass
FROM SPACEXTABLE
WHERE "Booster_Version" = 'F9 v1.1';
```

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

| average_payload_mass |
| --- |
| 2928.4 |

# FIRST SUCCESSFUL GROUND LANDING DATE

- The date of the first successful landing outcome on ground pad was 12 December, 2015.

```
%%sql
SELECT MIN("Date") as first_successful_landing_date
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (ground pad)';
```

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

| first_successful_landing_date |
| --- |
| 2015-12-22 |

# 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. Used the comparison operator to compare the payload mass.

```
%%sql
SELECT "Booster_Version"
FROM SPACEXTABLE
WHERE "Landing_Outcome" = 'Success (drone ship)'
  AND "PAYLOAD_MASS__KG_" > 4000
  AND "PAYLOAD_MASS__KG_" < 6000;
```

```
 * 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

- It was found that successful mission outcome is 100 while the number of failure mission outcome is 1.

List the total number of successful and failure mission outcomes

In [16]:
```
task_7a = '''
        SELECT COUNT(MissionOutcome) AS SuccessOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Success%'
        '''

task_7b = '''
        SELECT COUNT(MissionOutcome) AS FailureOutcome
        FROM SpaceX
        WHERE MissionOutcome LIKE 'Failure%'
        '''
print('The total number of successful mission outcome is:')
display(create_pandas_df(task_7a, database=conn))
print()
print('The total number of failed mission outcome is:')
create_pandas_df(task_7b, database=conn)
```

Out[16]:
```
The total number of successful mission outcome is:
```

| | successoutcome |
|---|---|
| 0 | 100 |

```
The total number of failed mission outcome is:
```

| | failureoutcome |
|---|---|
| 0 | 1 |

# BOOSTERS CARRIED MAXIMUM PAYLOAD

- The list the names of the booster which have carried the maximum payload mass

| 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 |
| F9 B5 B1049.7 |

# 2015 LAUNCH RECORDS

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015. Both the failed ones are from the launch site CCAFS LC-40.

```
%%sql
SELECT Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure (drone ship)%'
  AND Date BETWEEN '2015-01-01' AND '2015-12-31';
```

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

| Booster_Version | Launch_Site | Landing_Outcome |
|---|---|---|
| F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| F9 v1.1 B1015 | CCAFS LC-40 | Failure (drone ship) |

# 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 is depicted in the image.
- We utilized the GROUP BY clause to categorize landing outcomes and employed the ORDER BY clause to arrange them in descending order.

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome)
FROM SPACEXTABLE
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC;
```

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

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

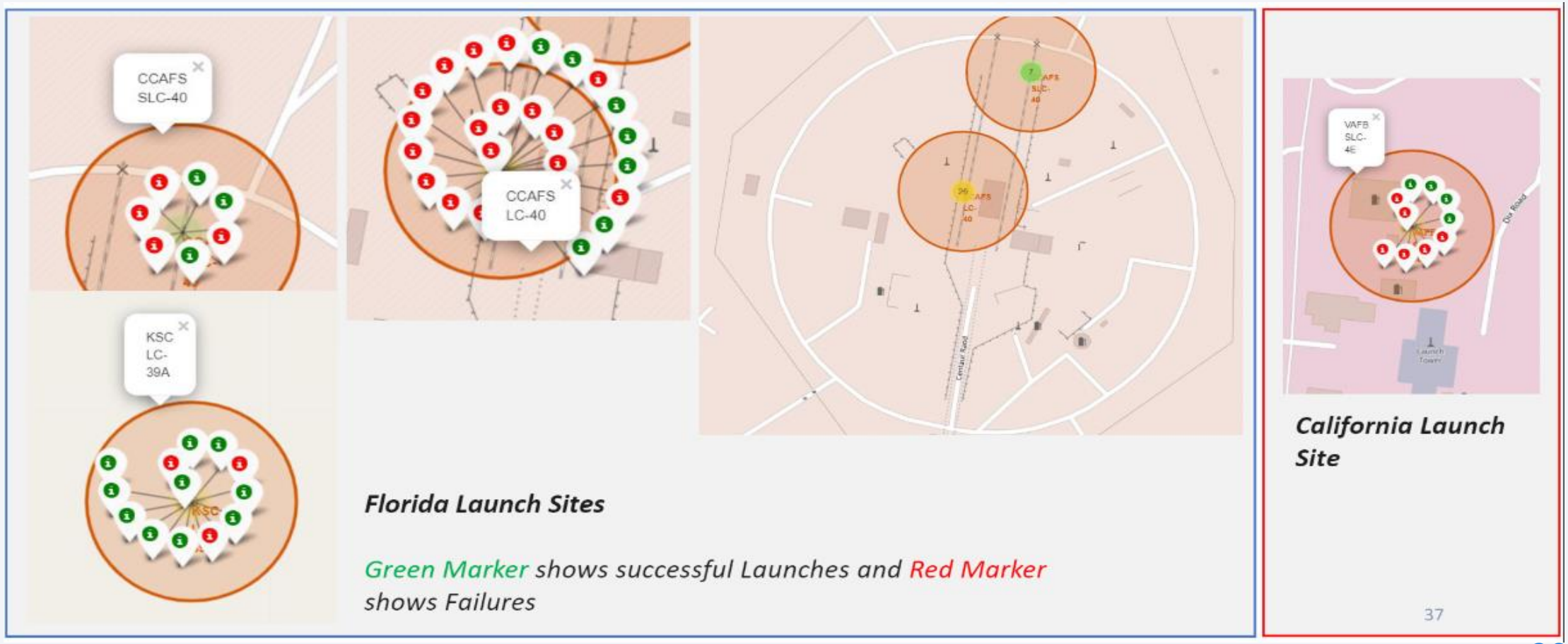
Section 3

# Launch Sites Proximities Analysis

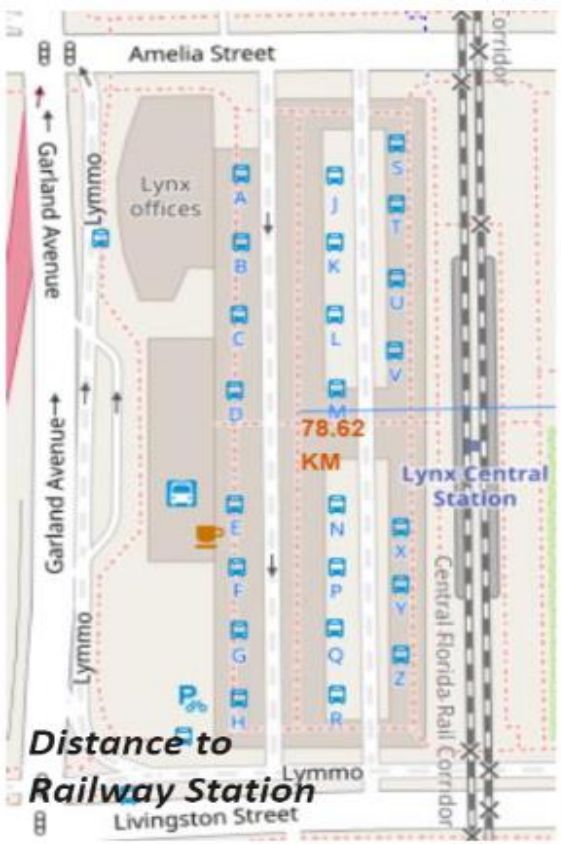
# SPACEX LAUNCH SITES


We can see that SpaceX launch sites are at the coasts at Florida and California.

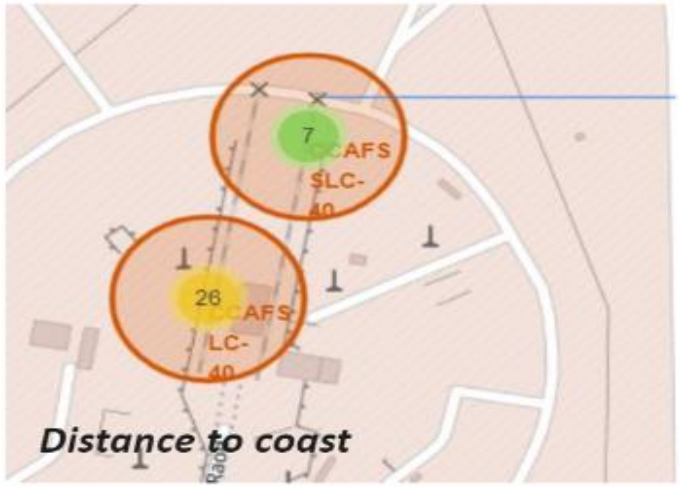
# MARKERS SHOWING LAUNCH SITES WITH COLOR LABELS

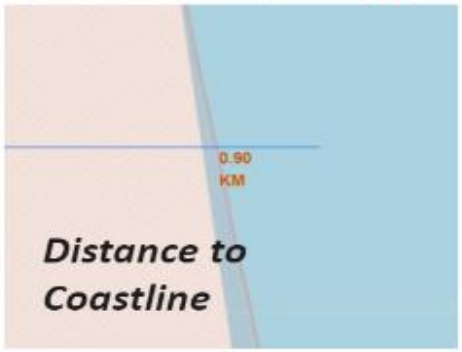


# LAUNCH SITE DISTANCE TO LANDMARKS



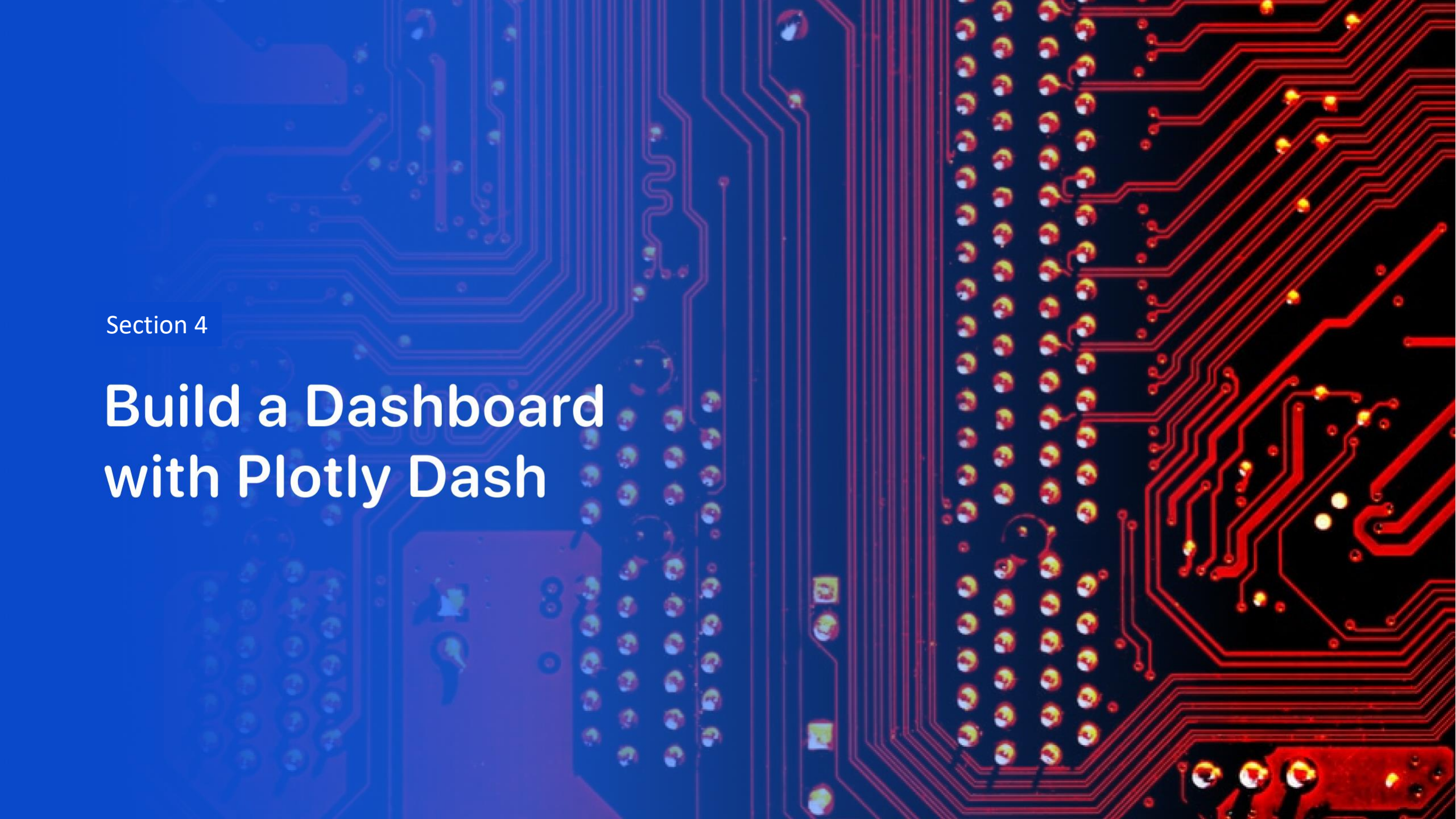




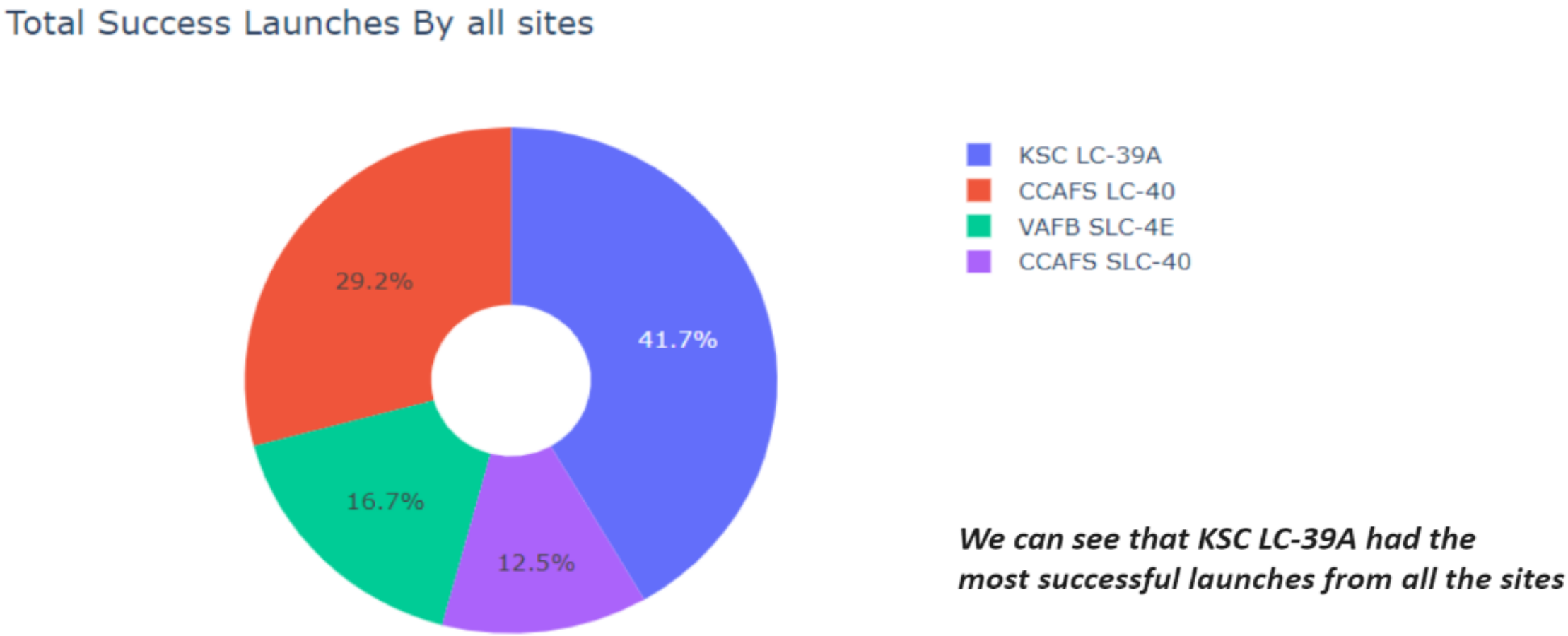





- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes

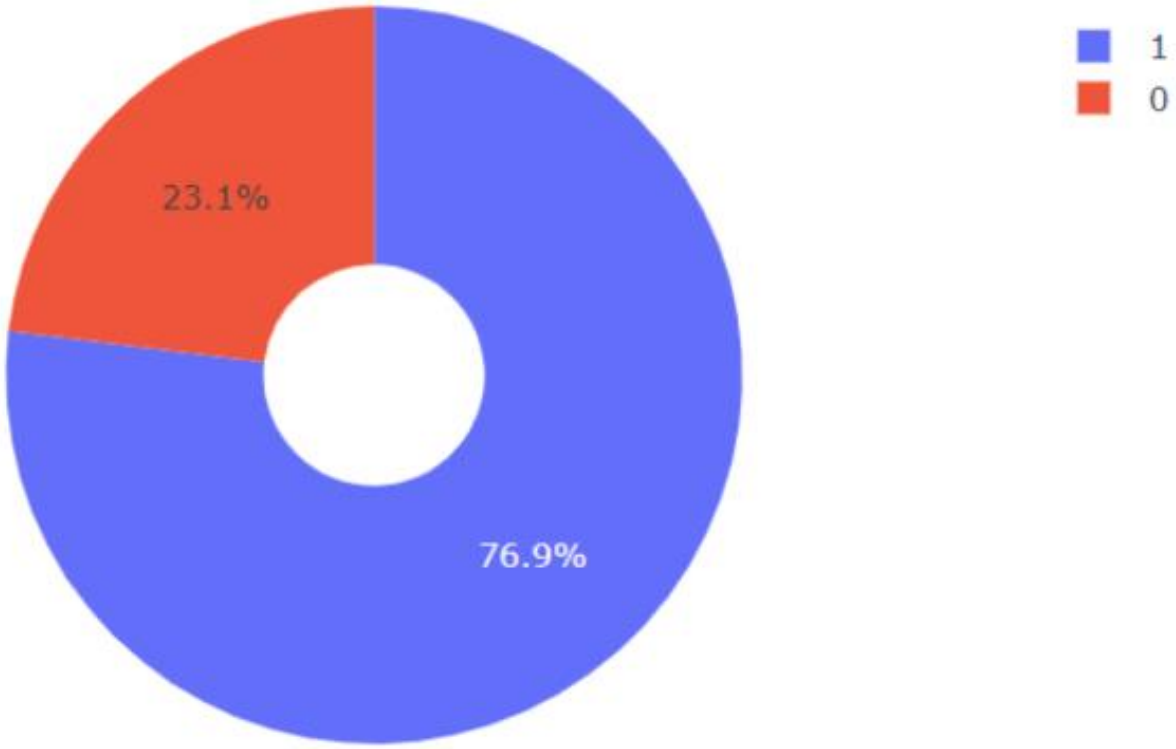

Section 4
Build a Dashboard with Plotly Dash

# PIE CHART SHOWING THE SUCCESS PERCENTAGE ACHIEVED BY EACH LAUNCH SITE



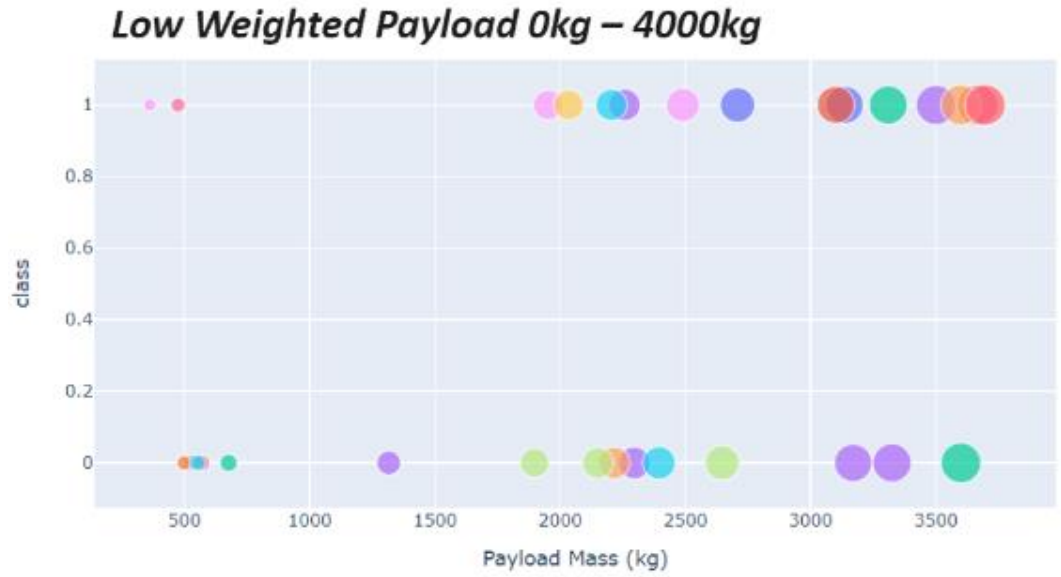

***We can see that KSC LC-39A had the most successful launches from all the sites***

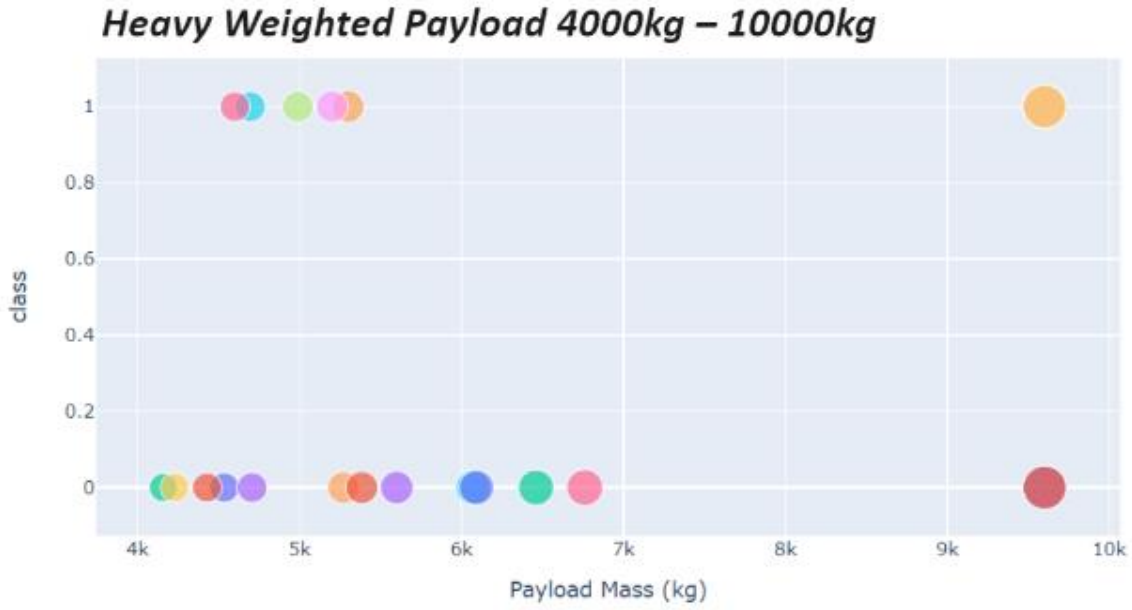
# PIE CHART SHOWING THE LAUNCH SITE WITH THE HIGHEST LAUNCH SUCCESS RATIO




*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



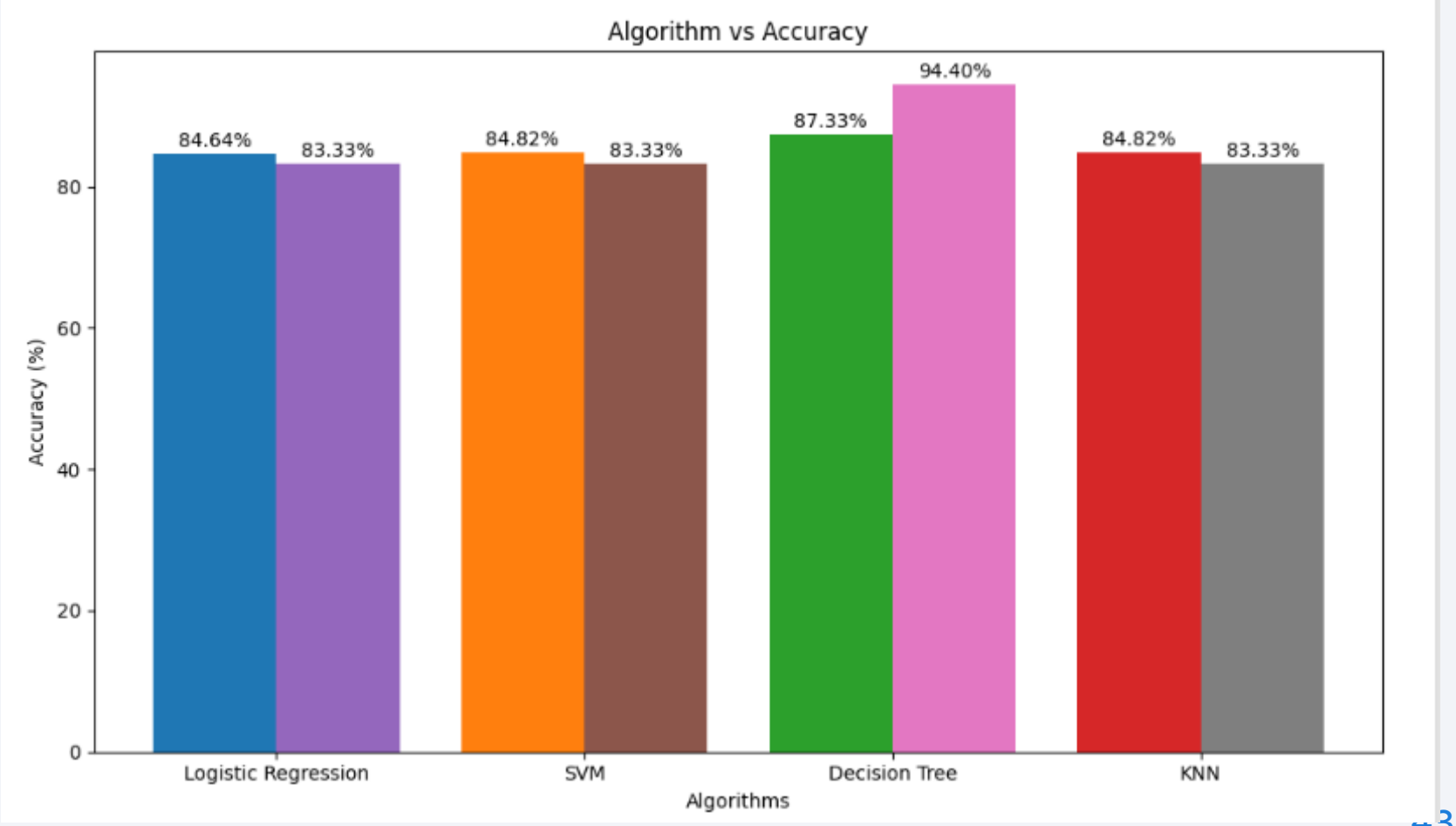

*We can see the success rates for low weighted payloads is higher than the heavy weighted payloads*

Section 5

# Predictive Analysis (Classification)

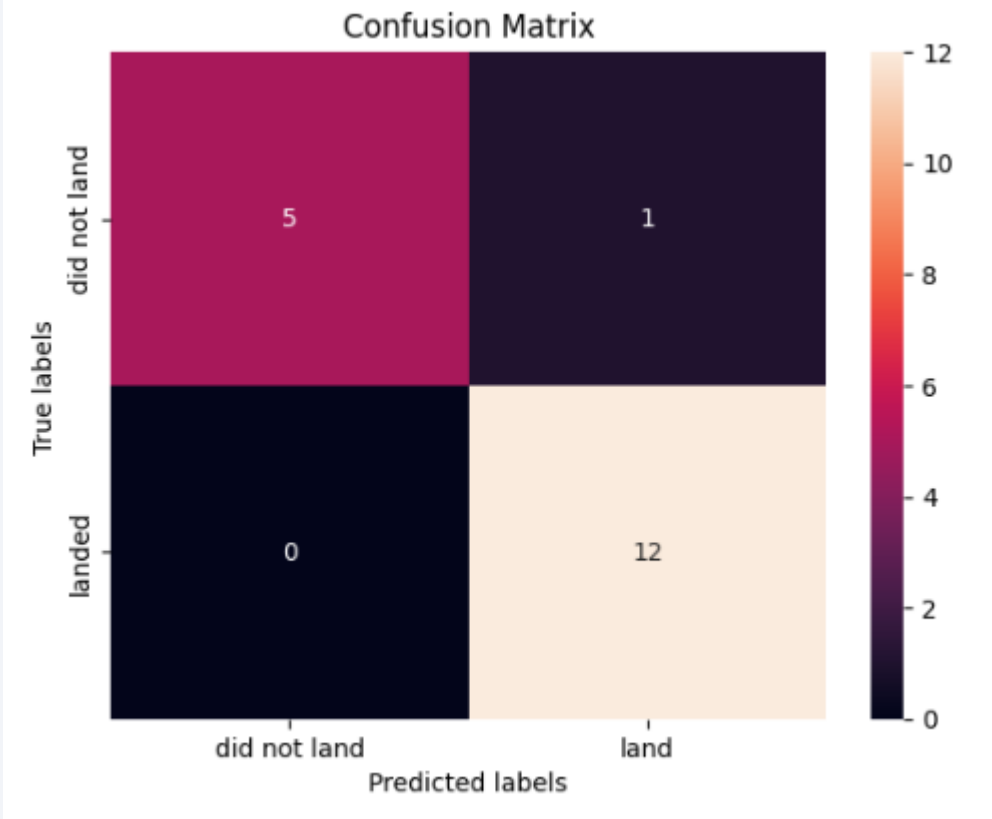
# CLASSIFICATION ACCURACY

- We can observe from the graph that Descision tree has the highest Training and test accuracy.



# CONFUSION MATRIX

- The Decision tree demonstrates the most accurate classification. Examining the confusion matrix reveals nearly perfect predictions, with only one false negative. This indicates just a single inaccurate prediction, suggesting a successful rocket landing. This is preferable to a false positive, which could incur significant losses by predicting a successful landing when it may not happen, impacting the company adversely.



# CONCLUSIONS

- It is evident that a higher number of flights at a launch site corresponds to an increased success rate at that site.
- The payload mass is directly proportional to the success rate of the rocket.
- 2. The success rate of launches showed a consistent upward trend from 2013 to 2020.
- 3. Orbits ES-L1, GEO, HEO, SSO, and VLEO demonstrated notably high success rates.
- 4. Among all launch sites, KSC LC-39A boasted the highest number of successful launches.
- 5. The Decision tree classifier emerged as the most effective machine learning algorithm for this specific task.

# Thank you!