IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Eray Yuztyurk
04.03.2024

# Outline

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

# Executive Summary

## Methodologies

Throughout the project, predictive modeling and data analysis methodologies were used to predict Falcon 9 first stage landings. This involved;

- Acquiring and preprocessing relevant data.
- Performing exploratory data analysis.
- Employing interactive visual analytics tools to enhance understanding and models.
- Applying predictive analysis techniques, such as machine learning algorithms.

## Results

- While KSC LC-39A boasts the highest success rate with 42% among all launch sites (standing at 77% for its own launches), CCAFS LC-40 Site 1 only maintains a success rate of 13% (with a success rate of 27% per its own launches) despite having the highest number of launches at 26, .
- Booster Version 'FT' mostly achieved success while 'v1.1' experienced failures.
- The majority of successful payloads fall within the range of 1800 to 5000.
- The resulting accuracy scores on the test dataset varied across the models, with Decision Trees achieving the highest accuracy of 0.875.

# Introduction

The project aims to predict the successful landing of the Falcon 9 first stage, a critical factor in SpaceX rocket launches. This prediction bears substantial financial implications, given that SpaceX's reusability of the first stage significantly lowers launch costs compared to other providers. Priced at $62 million per launch, Falcon 9 launches stand in stark contrast to competitors, which can cost upwards of $165 million each.

Accurately determining the success of the first stage landing is essential for estimating launch costs, especially when competing companies are vying for rocket launch contracts against SpaceX.

The objective of this project is to analyze data meticulously and predict the outcome of Falcon 9 first stage landings with precision, acknowledging the significant financial ramifications associated with this prediction.

Section 1

# Methodology

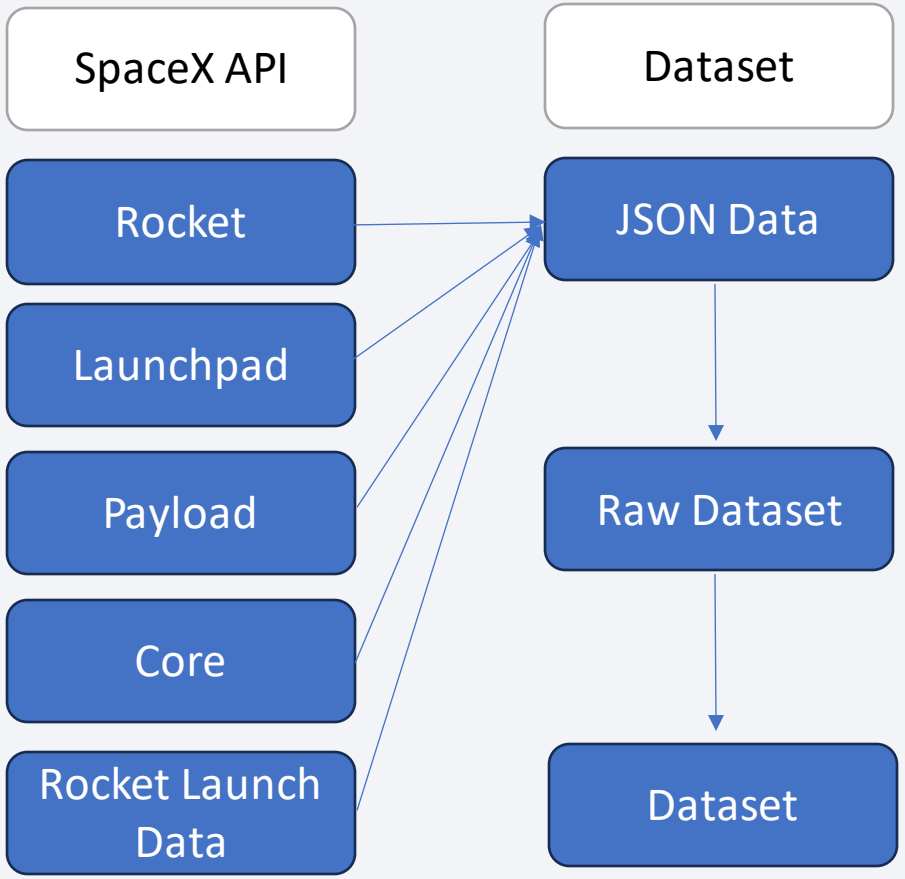
# Methodology

## Executive Summary

- Data collection methodology
  - Requested via SpaceX REST API and Scrapping historical data from Web (Wikipedia)
- Performing data wrangling
  - Cleaning, transforming, and preparing the collected data for analysis, ensuring its quality and suitability for further processing
- Performing exploratory data analysis (EDA) using visualization and SQL
- Performing interactive visual analytics using Folium and Plotly Dash
- Performing predictive analysis using classification models
  - Logistic Regression, Decision Tree, K-Nearest Neighbors and SVM models are built using scikit learn library, searched for best hyperparameters via GridSearchCV with 10-fold, tuned with best hyperparameters found and results are evaluated with accuracy score.

# Data Collection

- Using SpaceX REST API, essential information were retrieved such as rocket, launchpads etc.
- The data being in JSON format, was normalized to convert it into a tabular form.
- From this tabular data, only the necessary information was filtered ('rocket','payloads','launchpad','cores','flight_number','date_utc'), resulting in a dataset ready for analysis.



# Data Collection – SpaceX API

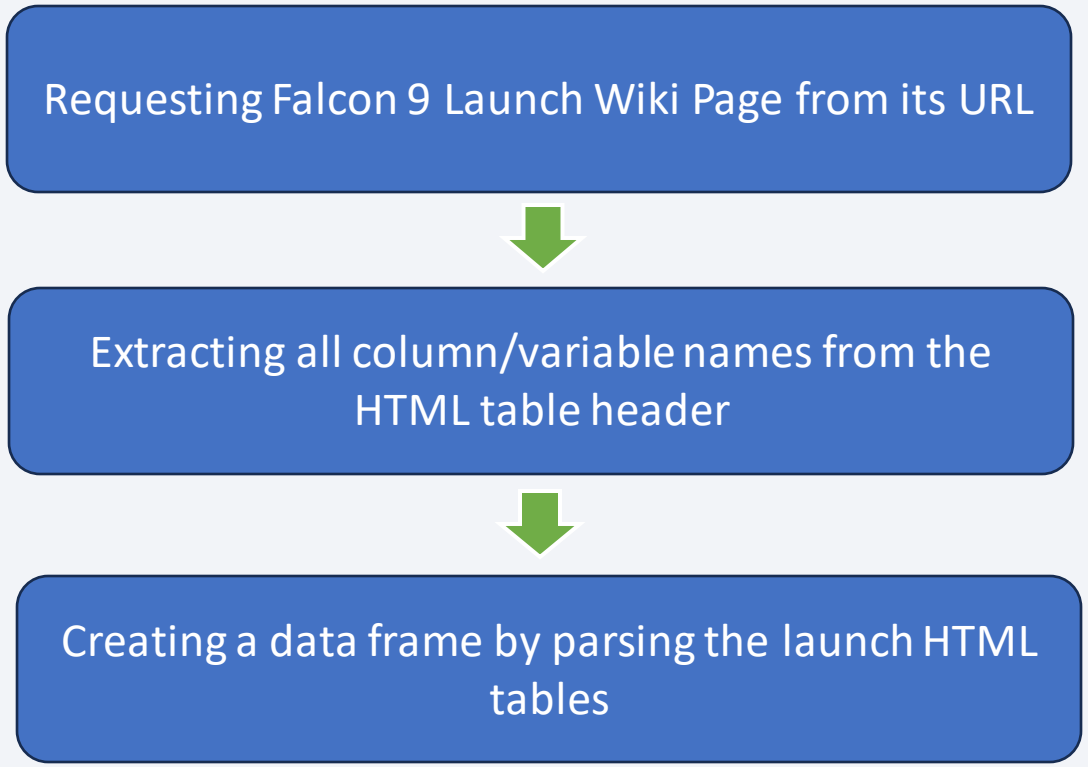
| Source Type | Rockets Info | Launchpads Info | Payloads Info | Cores Info |
|---|---|---|---|---|
| Features | Booster Version | Longitude | Payload Mass | Block |
| | | Latitude | Orbit | Reused Count |
| | | Launch Site | | Serial |
| | | | | Outcome |
| | | | | Flights |
| | | | | GridFins |
| | | | | Reused |
| | | | | Legs |
| | | | | Landing Pad |
| API Keys | "https://api.spacexdata.com/v4/rockets/" | "https://api.spacexdata.com/v4/launchpads/" | "https://api.spacexdata.com/v4/payloads/" | "https://api.spacexdata.com/v4/cores/" |



- *Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# Data Collection - Scrapping

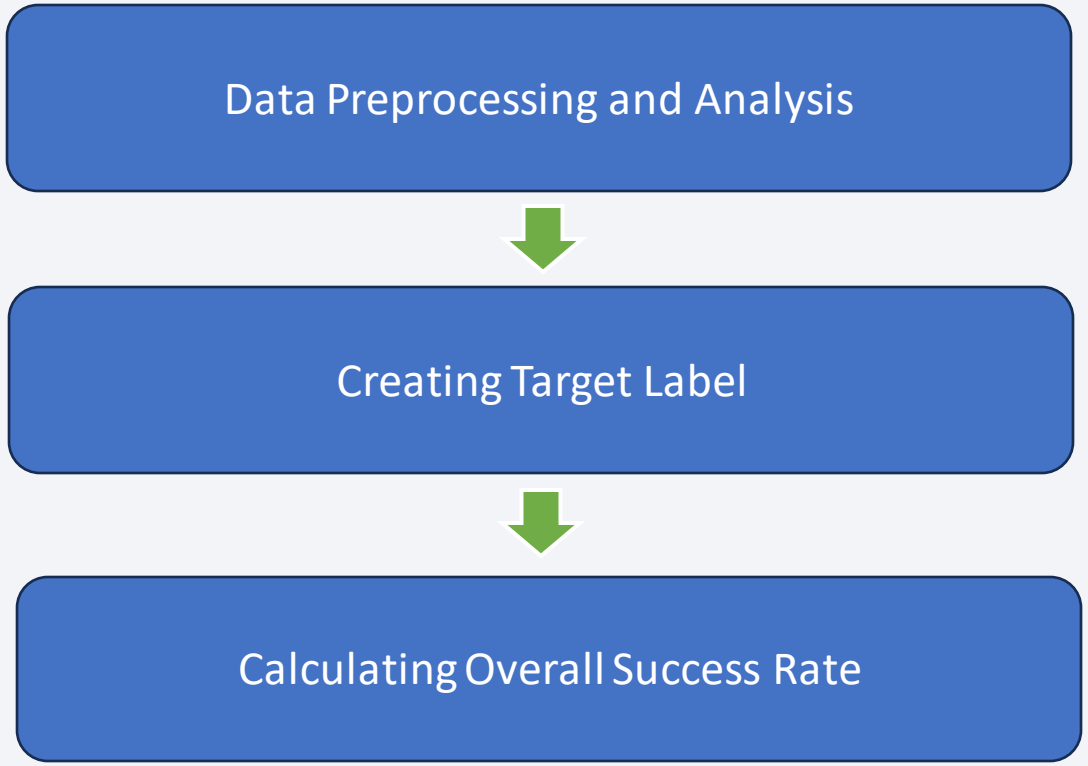
- Falcon 9 launch records were extracted from HTML table of Wikipedia using request and Beautiful Soap libraries and a dataframe was created by parsing extracted data.




*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# Data Wrangling

- Handled missing values within the dataset.
- Created a target label ('class') derived from the Outcome variable, which includes 'True' values.
- Investigated launch sites and orbit types.
- Computed the overall success rate based on the generated target label, resulting in a rate of 66%.




*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# EDA with Data Visualization

- **Flight Number vs. Payload Mass**: This scatter plot was used to observe how the flight number and payload mass affect the success of the launch. It helps visualize trends and patterns in the relationship between these variables and the launch outcome.
- **Payload Mass vs. Launch Site**: This scatter plot was used to examine the relationship between the payload mass and the launch site, considering the launch outcomes. It helps identify any potential correlations or patterns between the payload mass, launch site, and launch success.
- **Success Rates over Orbit Types**: A bar chart was plotted to visualize the success rates for different orbit types. This chart helps identify which orbit types have higher success rates and can provide insights into the factors influencing launch success.
- **Flight Number vs. Orbit Types**: A scatter plot was used to analyze the relationship between the flight number and the orbit type. It helps identify any trends or patterns in the success of launches across different orbit types over time.
- **Payload Mass across Orbit Types**: Another scatter plot was plotted to explore how payload mass varies across different orbit types, considering the launch outcomes. It helps identify any correlations between payload mass, orbit type, and launch success.
- **Success Rate Trend from 2010 to 2020**: A line chart was plotted to visualize the trend in the success rate of launches over the years. It helps identify any overall trends or patterns in the success rates of launches from 2010 to 2020.



*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# EDA with SQL

- SQL queries performed:

- The names of the unique launch sites in the space mission.
- 5 records where launch sites begin with the string 'CCA'.
- Total payload mass carried by boosters launched by NASA (CRS).
- Average payload mass carried by booster version F9 v1.1
- The date when the first succesful landing outcome in ground pad was achieved.
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Total number of successful and failure mission outcomes.
- The names of the booster versions which have carried the maximum payload mass.
- The records which display the months, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015.
- 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.



*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# Build an Interactive Map with Folium

Map objects were added to provide a comprehensive visual representation of launch sites, their success rates, and their proximity to important geographical features. They aid in analyzing factors such as launch site selection, operational logistics, and environmental considerations for space missions.

- **Markers**: Markers were used to represent the launch sites on the map. Each marker corresponds to a specific launch site location, and they are labeled with the name of the launch site. These markers provide a visual reference for the locations of different launch sites.
- **Circles**: Circles were added around the launch sites to highlight their approximate areas of influence or proximity. These circles serve as visual indicators of the potential operational range or coverage of each launch site. They help viewers understand the spatial extent of each site's impact.
- **Marker Clusters**: Marker clusters were utilized to handle multiple markers that share the same coordinates, which is common in launch records where multiple launches occur at the same site. Clustering these markers enhances map readability by preventing overcrowding and providing a clear representation of launch outcomes at each site.
- **PolyLines**: PolyLines were drawn to visually represent distances between launch sites and various points of interest, such as coastline, highway and railway. These lines help viewers understand the spatial relationships between launch sites and their surroundings, facilitating analysis of site accessibility and environmental factors.

*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# Build a Dashboard with Plotly Dash

Plots/graphs and interactions on the dashboard

- **Dropdown List**: Allows selection of launch sites. Enables users to select specific launch sites, allowing for site-specific analysis and comparison. It is in interaction with Pie Chart.
- **Pie Chart**: Displays total successful launches by each site or overall. Offers a clear visualization of the distribution of successful launches across different sites, aiding in understanding the success rates at each site. It is in interaction with Dropdown List.
- **Range Slider**: Selects payload range. Facilitates the selection of a payload range, allowing users to focus their analysis on launches within a specific payload mass range. It is in interaction with Scatter Chart.
- **Scatter Chart**: Shows payload vs. launch success correlation and booster category distribution simultenously. Illustrates the relationship between payload mass and launch success, providing insights into how payload mass impacts launch outcomes. It is interaction with Dropdown List and Range Slider.

*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

# Predictive Analysis (Classification)

**Data Preprocessing:**

- The dataset was loaded from two URLs and preprocessed using techniques such as standardization for the features in X and creating the target variable Y from the "Class" column in the data.

**Model Selection:**

- Four classification algorithms were considered: Logistic Regression, Support Vector Machine (SVM), Decision Trees, and K-Nearest Neighbors (KNN).

**Hyperparameter Tuning:**

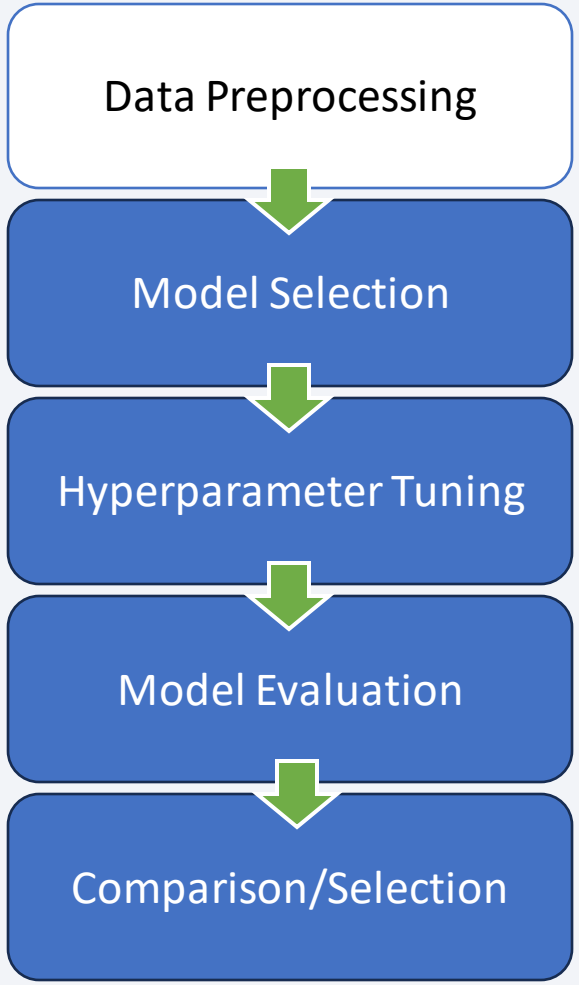
- For each algorithm, a GridSearchCV object was created to tune the hyperparameters using cross-validation (cv=10). The best parameters for each model were determined based on the provided parameter grids.

**Model Evaluation:**

- The accuracy scores of the tuned models were calculated on the test data using the **score** method. Additionally, classification reports were generated to provide a comprehensive evaluation of precision, recall, F1-score, and support for each class.

**Comparison and Selection:**

- Finally, the performance of each model was compared based on their accuracy scores. The model with the highest accuracy was deemed the best performing classification model.




*Please find the related Jupyter notebook for associated operations on my GitHub repository. You can access it by clicking here.*

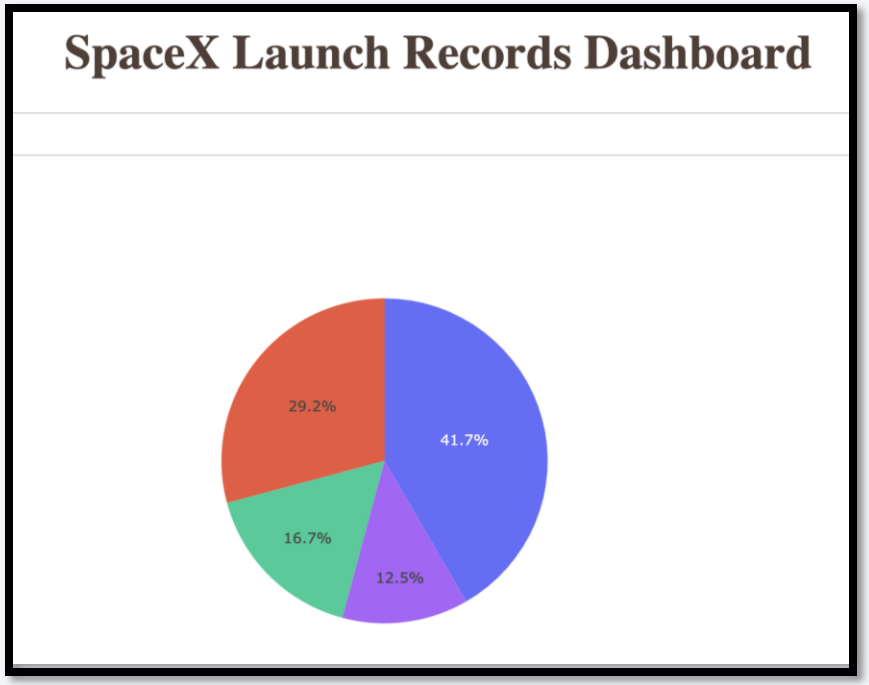
# Results

## Exploratory Data Analysis Results

- The first 20 flights primarily resulted in failures, but success rates increased thereafter (Page 18).
- CCAFS SLC 40 hosts most launches, with payload masses predominantly falling between 0 and 8000 (Page 19).
- Flights targeting ES-L1, SSO, HEO, and GEO orbits show the highest success rates, followed by VLEO (Page 20).
- Success in LEO orbit seems linked to flight number, whereas no such relationship is evident for GTO orbit (Page 21).
- Heavy payloads tend to achieve successful or positive landings in LEO and ISS orbits, while SSO sees success with lighter payloads. However, distinguishing between positive and negative landing rates in GTO orbit is challenging (Page 22).
- A consistent upward trend in launch success is observed from 2010 to 2020 (Page 23).
- KSC LC-39A leads with a 42% success rate overall (77% for its own launches), while CCAFS LC-40 Site 1 has a lower success rate of 13% (27% for its own launches) despite hosting 26 launches. (Page 44/45)

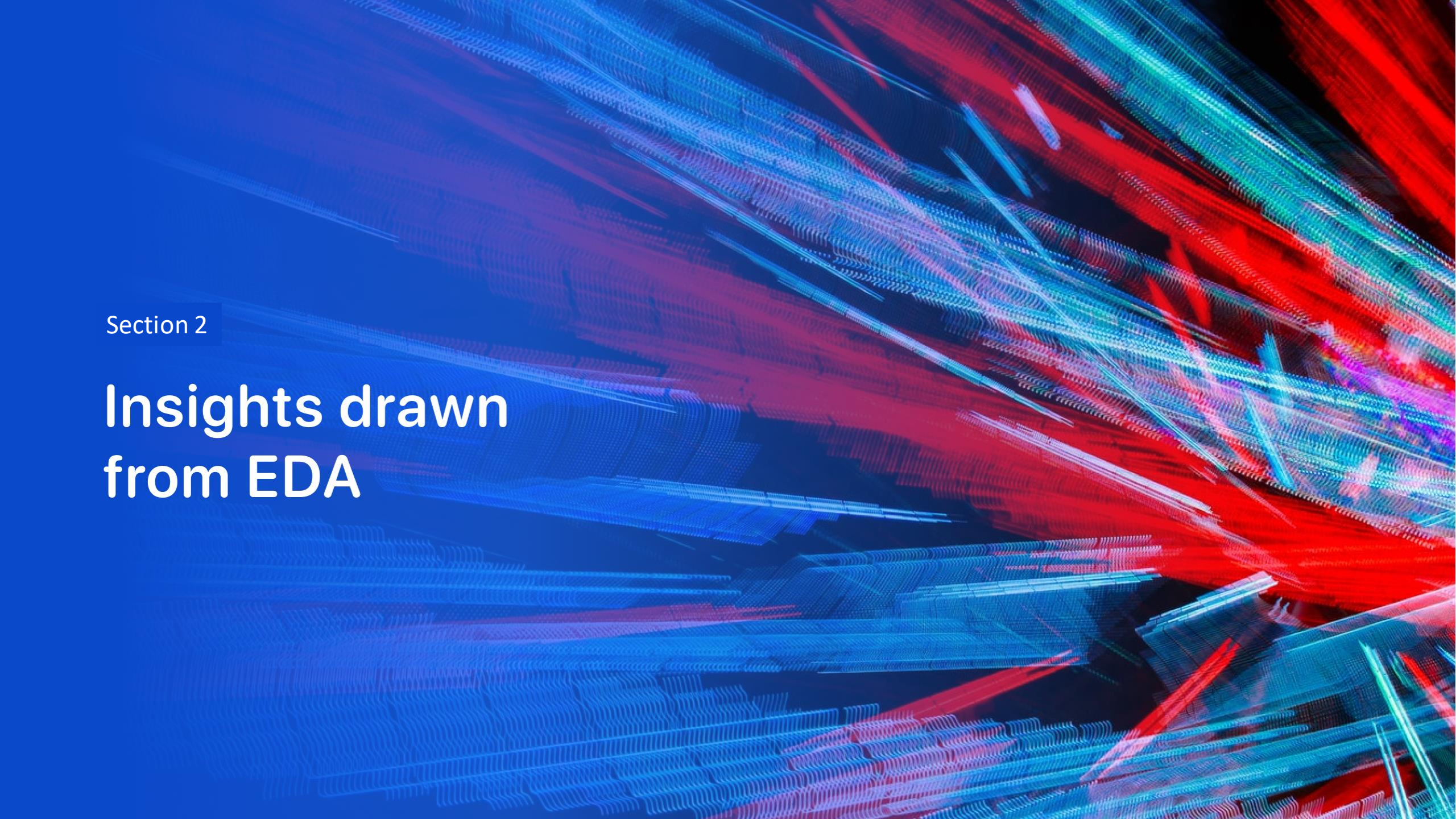
## Predictive Analysis Results

- The test dataset revealed varying accuracy scores among the models, with Decision Trees leading with an accuracy of 0.875, closely trailed by SVM at 0.848. Following suit, Logistic Regression and KNN displayed competitive accuracy scores of 0.846 and 0.848, respectively. (Page 49)



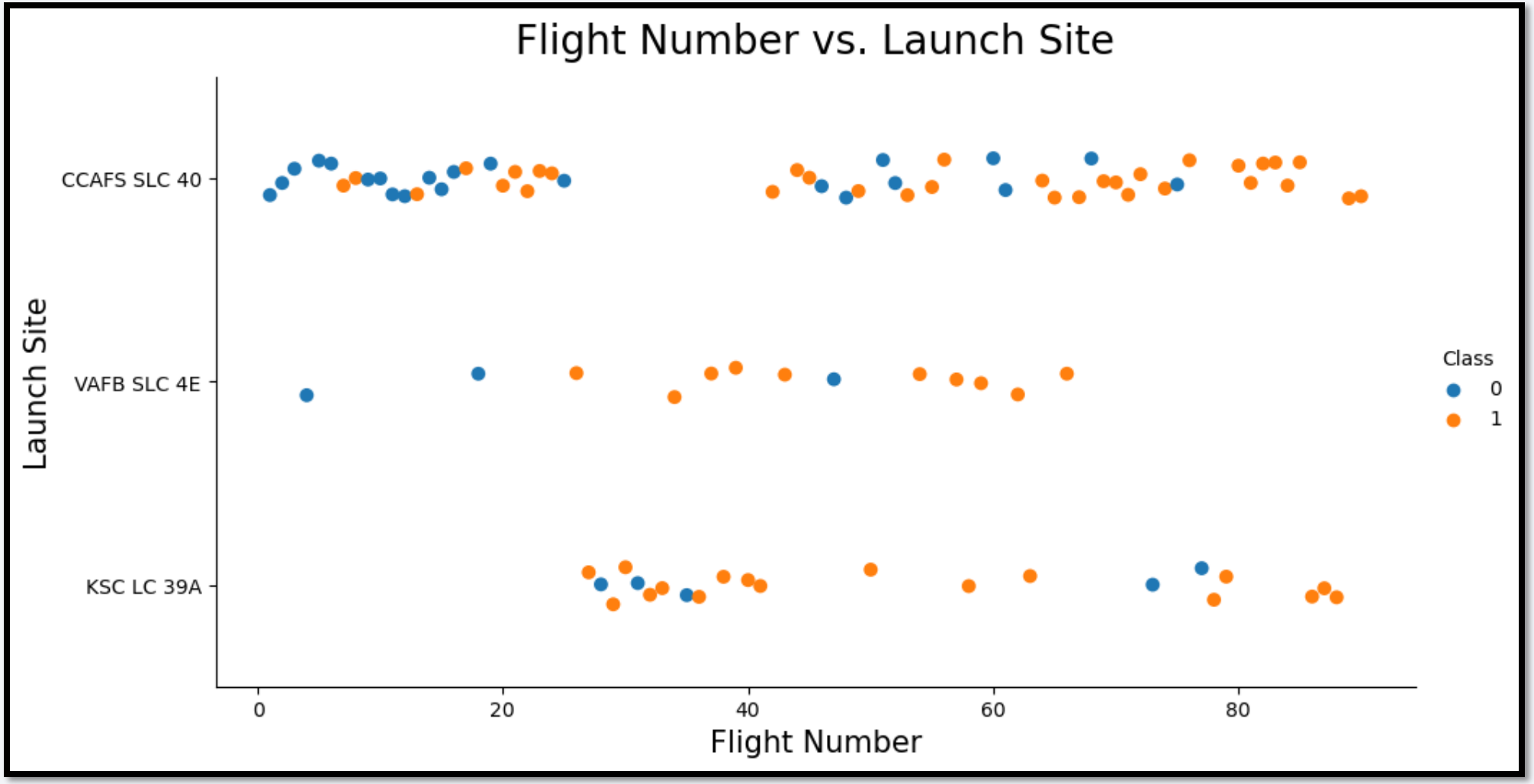
Section 2

# Insights drawn from EDA



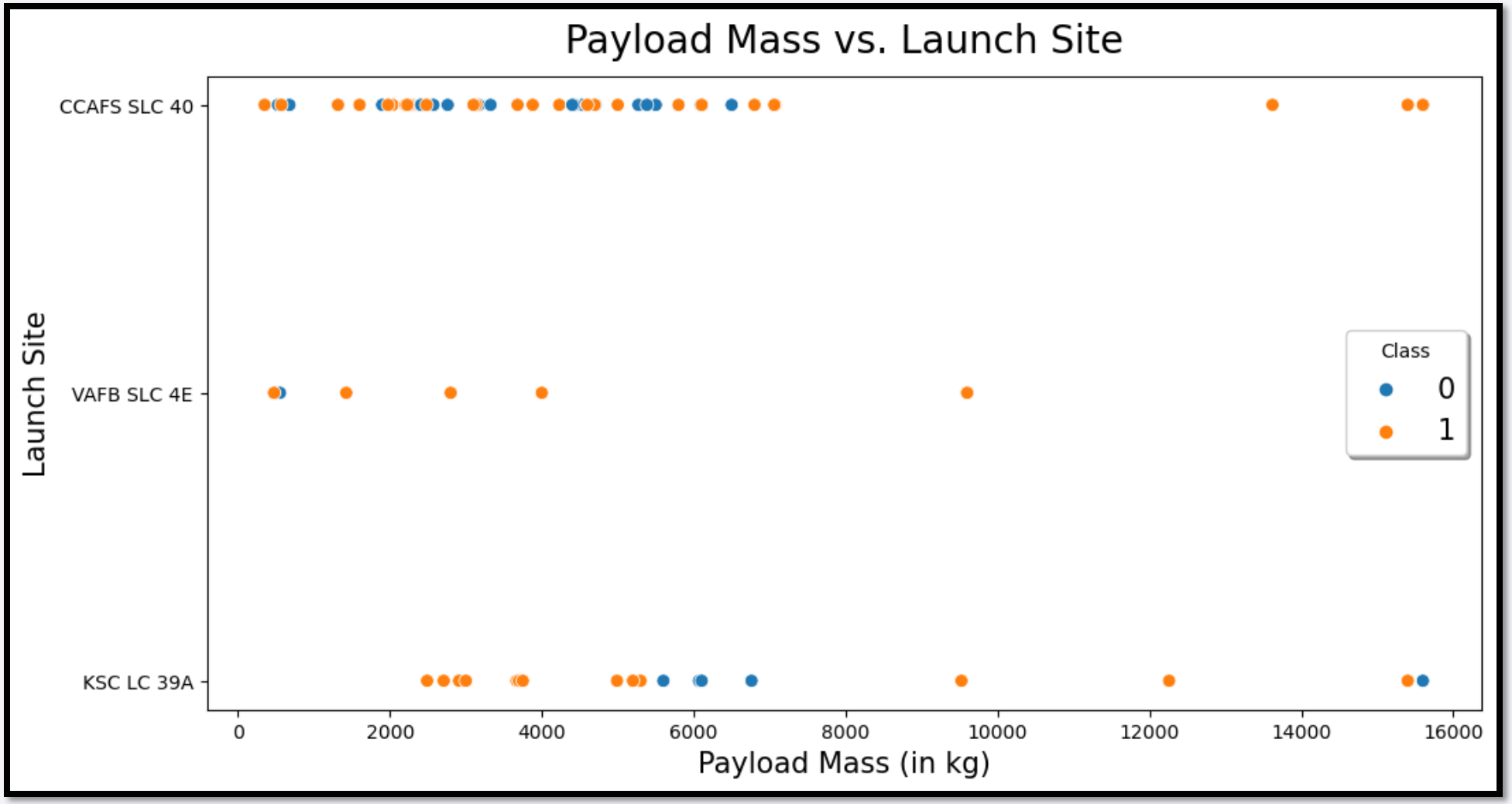
# Flight Number vs. Launch Site

(Success – 1, Fail – 0)

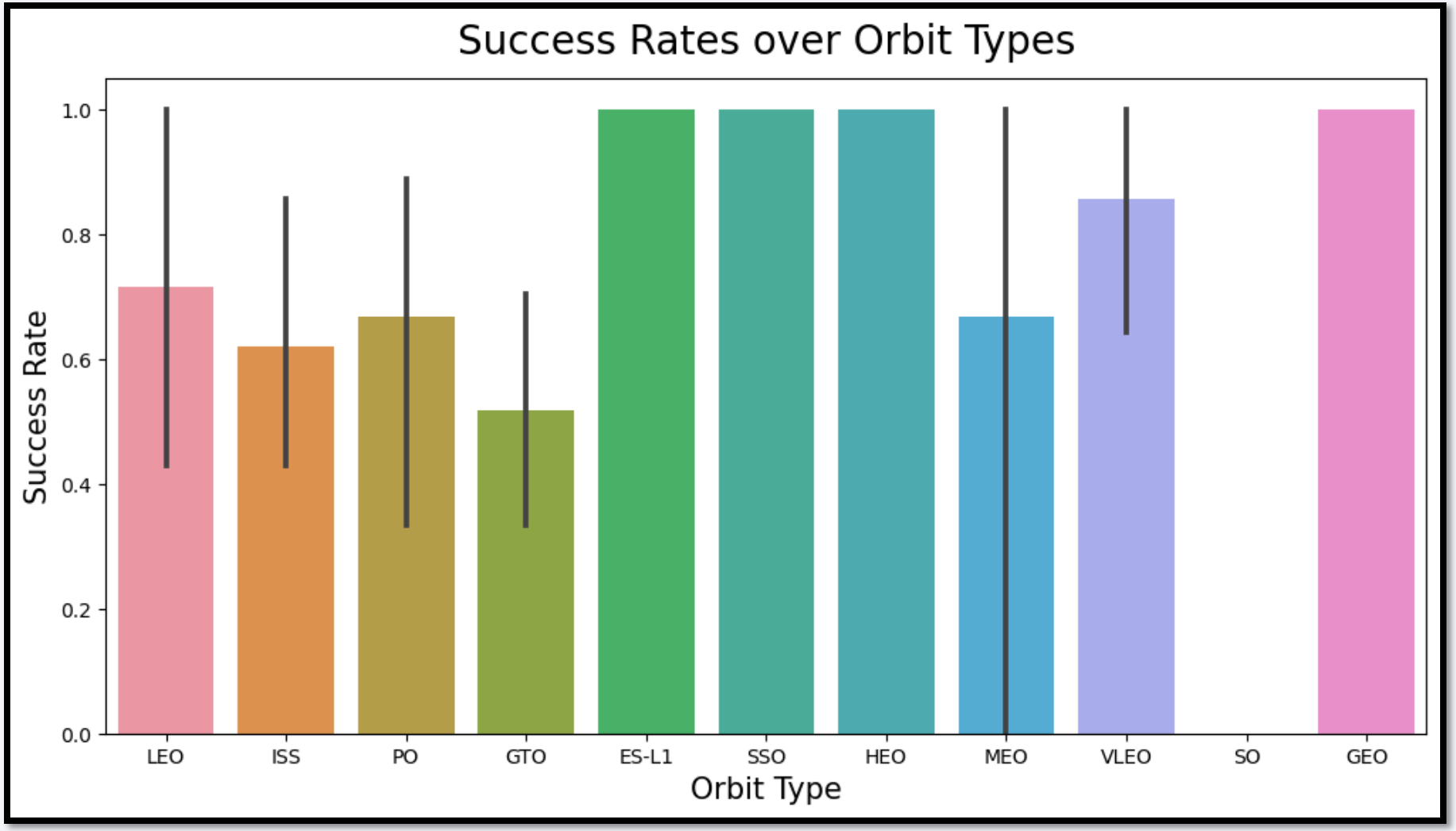


# Payload vs. Launch Site

(Success –1, Fail – 0)

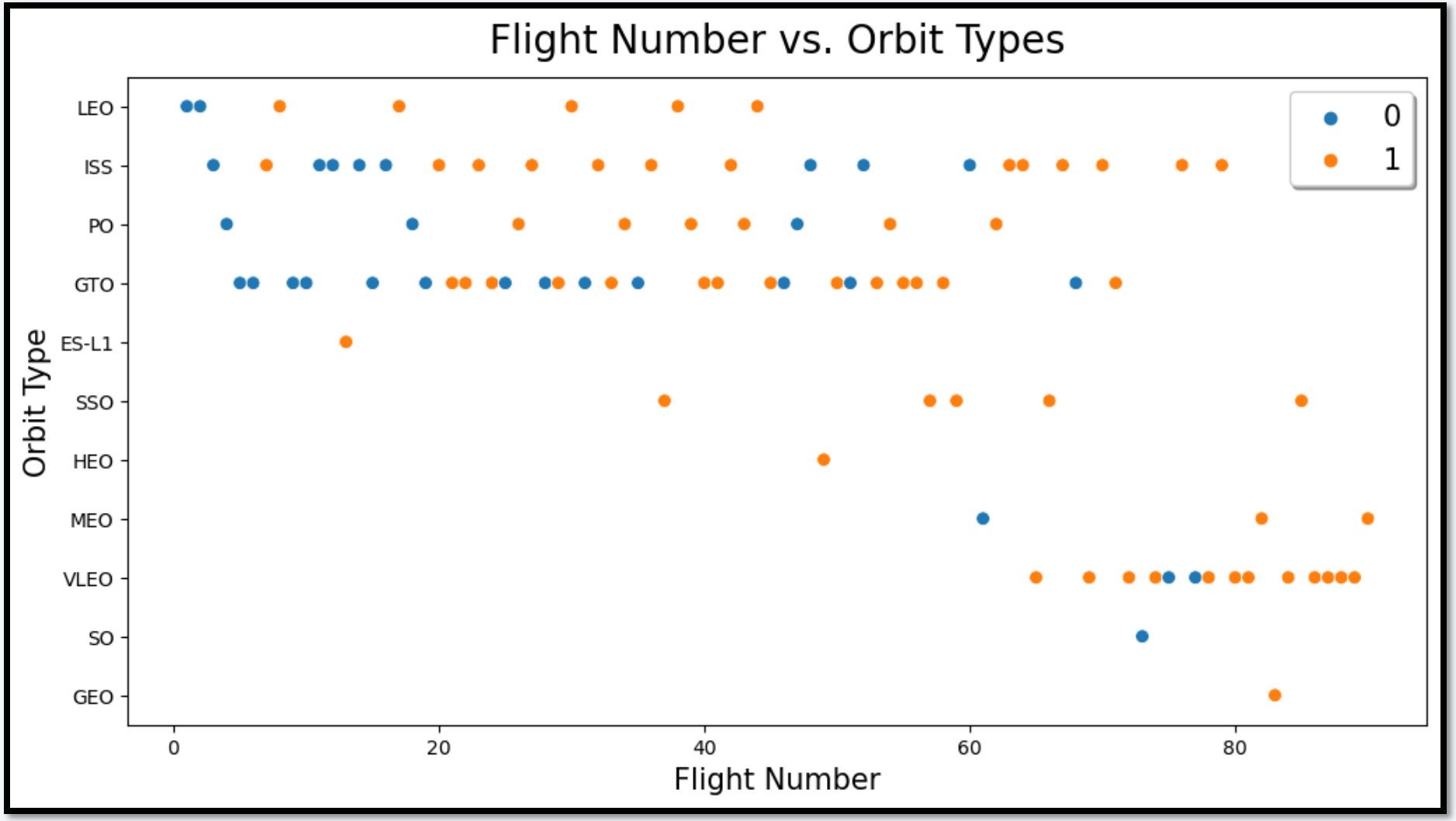


# Success Rate vs. Orbit Type

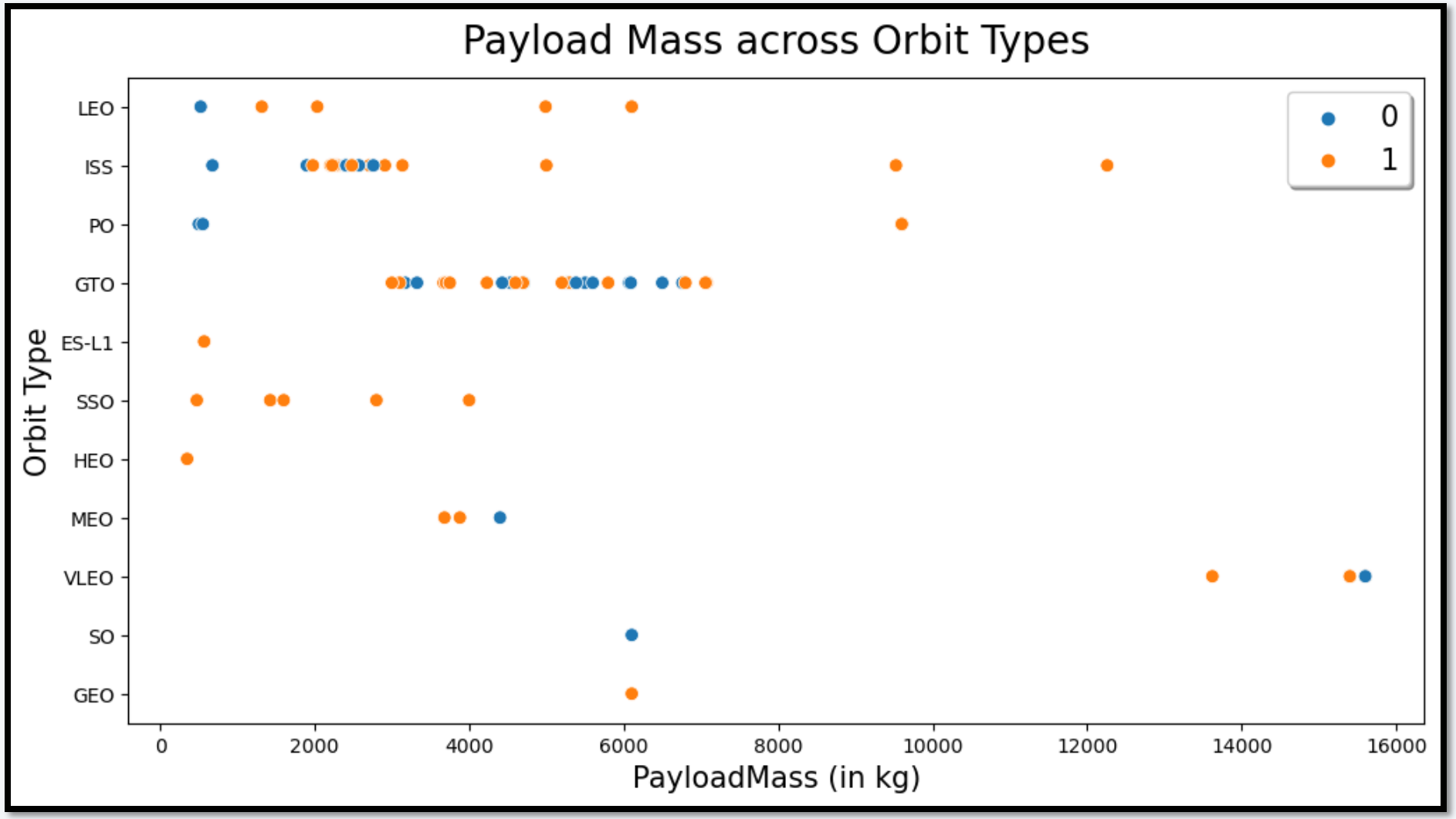


# Flight Number vs. Orbit Type

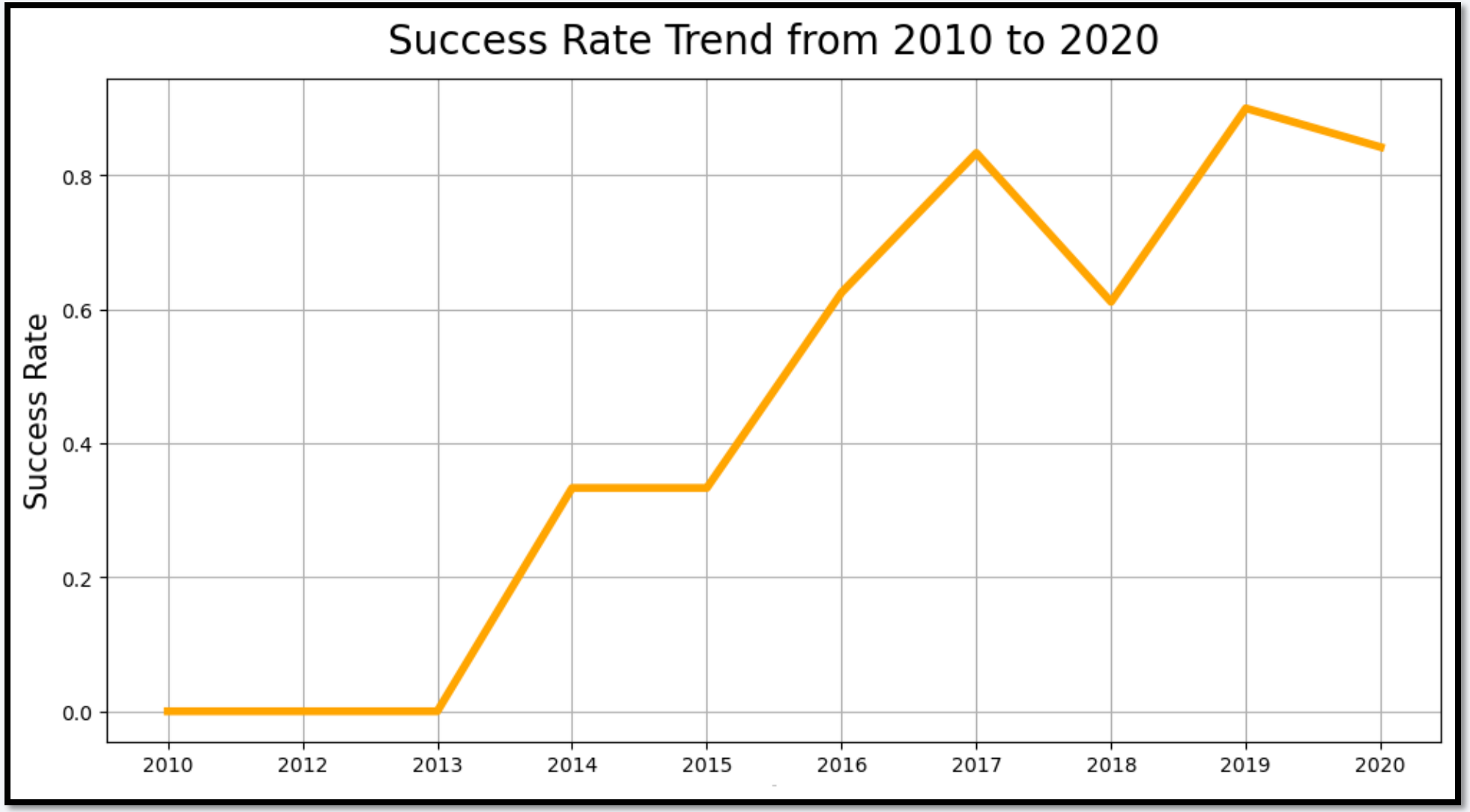
(Success –1, Fail – 0)



# Payload vs. Orbit Type



# Launch Success Yearly Trend



# All Launch Site Names

There are 4 launch sites used by SpaceX,

- CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40 – Site 1)
- VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4E)
- KSC LC-39A (Kennedy Space Center Launch Complex 39A)
- CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40 – Site 2)

**Query and results:**

- [In]: `cur.execute("SELECT DISTINCT(Launch_Site) FROM SPACEXTABLE").fetchall()`
- [Out]: `[('CCAFS LC-40',), ('VAFB SLC-4E',), ('KSC LC-39A',), ('CCAFS SLC-40',)]`

# Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

Query and results:

- [In]:

```
rows = cur.execute("SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5").fetchall()
columns = [desc[0] for desc in cur.description]
pd.DataFrame(rows, columns=columns)
```

- [Out]:

| | Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outcome |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2010-06-04 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 1 | 2010-12-08 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of... | 0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (parachute) |
| 2 | 2012-05-22 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 3 | 2012-10-08 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 4 | 2013-03-01 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677 | LEO (ISS) | NASA (CRS) | Success | No attempt |

# Total Payload Mass

- Total payload carried by boosters from NASA is approx. over **45 tons**.

Query and results:

- [In]: cur.execute("SELECT SUM(PAYLOAD_MASS__KG_), Customer FROM SPACEXTABLE GROUP BY Customer HAVING Customer LIKE 'NASA (CRS)%'").fetchall()
- [Out]: [(45596, 'NASA (CRS)'), (2617, 'NASA (CRS), Kacific 1')]

# Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 is approx. **3 tons**.

Query and results:

- [In]: `cur.execute("SELECT AVG(PAYLOAD_MASS__KG_), Booster_Version FROM SPACEXTABLE GROUP BY Booster_Version HAVING Booster_Version = 'F9 v1.1'").fetchall()`
- [Out]: `[(2928.4, 'F9 v1.1')]`

# First Successful Ground Landing Date

- The date of the first successful landing on the ground pad is December 22, 2015.

Query and results:

- [In]: cur.execute("SELECT MIN(Date), Mission_Outcome, Landing_Outcome FROM SPACEXTABLE WHERE Mission_Outcome LIKE '%Success%' AND Landing_Outcome LIKE '%ground pad%'").fetchall()
- [Out]: [('2015-12-22', 'Success', 'Success (ground pad)')]

# Successful Drone Ship Landing with Payload between 4000 and 6000

- The name of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 is '**F9 v1.1**'.

Query and results:

- [In]: `cur.execute("SELECT Booster_Version, Mission_Outcome, Landing_Outcome FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 GROUP BY Booster_Version AND Mission_Outcome LIKE '%Success%' AND Landing_Outcome LIKE '%drone ship%'").fetchall()`
- [Out]: `[('F9 v1.1', 'Success', 'No attempt')]`

# Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes is **101**.
- **100** of them with success and **1** is with failure.

| Mission_Outcome | |
|---|---|
| Success | 98 |
| Failure (in flight) | 1 |
| Success (payload status unclear) | 1 |
| Success | 1 |

**Query and results:**

- [In]: `cur.execute("SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE '%uccess%' OR Mission_Outcome LIKE '%ailure%'").fetchall()`
- [Out]: `[(101,)]`
- [In]: `cur.execute("SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE '%uccess%'").fetchall()`
- [Out]: `[(100,)]`
- [In]: `cur.execute("SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE '%ailure%'").fetchall()`
- [Out]: `[(1,)]`

# Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass prodominantly belong to F9 B5 B1 series.

Query and results:

- [In]: `cur.execute("SELECT Booster_Version, PAYLOAD_MASS__KG_ FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ IN (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)").fetchall()`
- [Out]:

```
[('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

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 are 'F9 v1.1 B1012' launched by 'CCAFS LC-40' in January and 'F9 v1.1 B1015' launched by 'CCAFS LC-40' in April.

Query and results:

•[In]:
```
cur.execute("SELECT SUBSTR(Date,6,2), Landing_Outcome, Booster_Version, Launch_site \
                    FROM SPACEXTABLE \
                    WHERE Landing_Outcome LIKE '%ailure%' AND Landing_Outcome LIKE '%drone ship%' \
                    AND SUBSTR(Date,0,5) = '2015'").fetchall()
```
•[Out]:

```
[('01', 'Failure (drone ship)', 'F9 v1.1 B1012', 'CCAFS LC-40'),
('04', 'Failure (drone ship)', 'F9 v1.1 B1015', 'CCAFS LC-40')]
```

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

- The ranked 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 are:

**Query and results:**

- [In]:
```
cur.execute("SELECT COUNT(Landing_Outcome), Landing_Outcome FROM SPACEXTABLE GROUP BY Landing_Outcome HAVING Date BETWEEN '2010-06-04' AND '2017-03-20' ORDER BY COUNT(Landing_Outcome) DESC").fetchall()
```
- [Out]:
```
[(21, 'No attempt'),
(14, 'Success (drone ship)'),
(9, 'Success (ground pad)'),
(5, 'Failure (drone ship)'),
(5, 'Controlled (ocean)'),
(2, 'Uncontrolled (ocean)'),
(2, 'Failure (parachute)'),
(1, 'Precluded (drone ship)')]
```

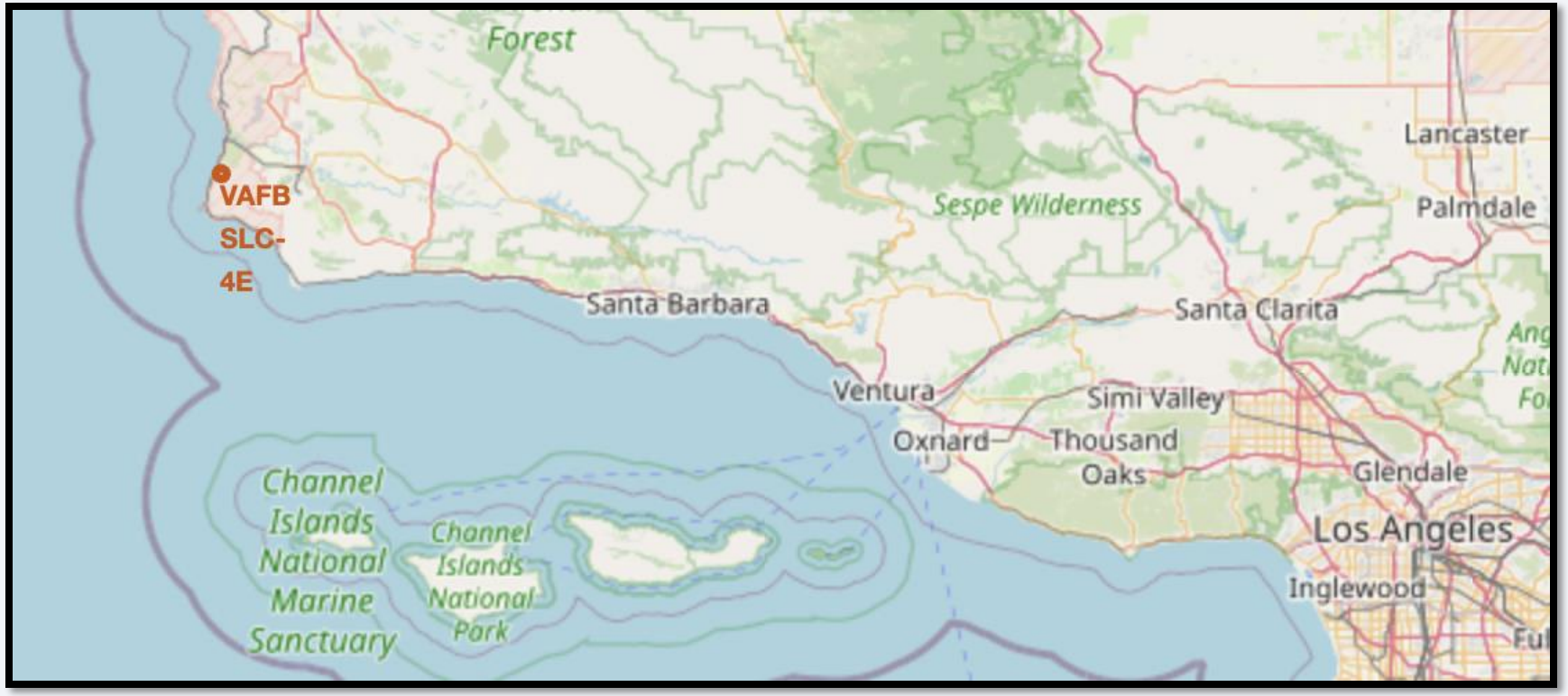

Section 3
Launch Sites
Proximities Analysis

# Launch Sites – General Overview

As we can see below, there are 4 Launch Sites in total, 3 of them on east side and 1 is in west.

# Launch Sites – West

VAFB SLC-4E (Vandenberg Air Force Base Space Launch Complex 4E)

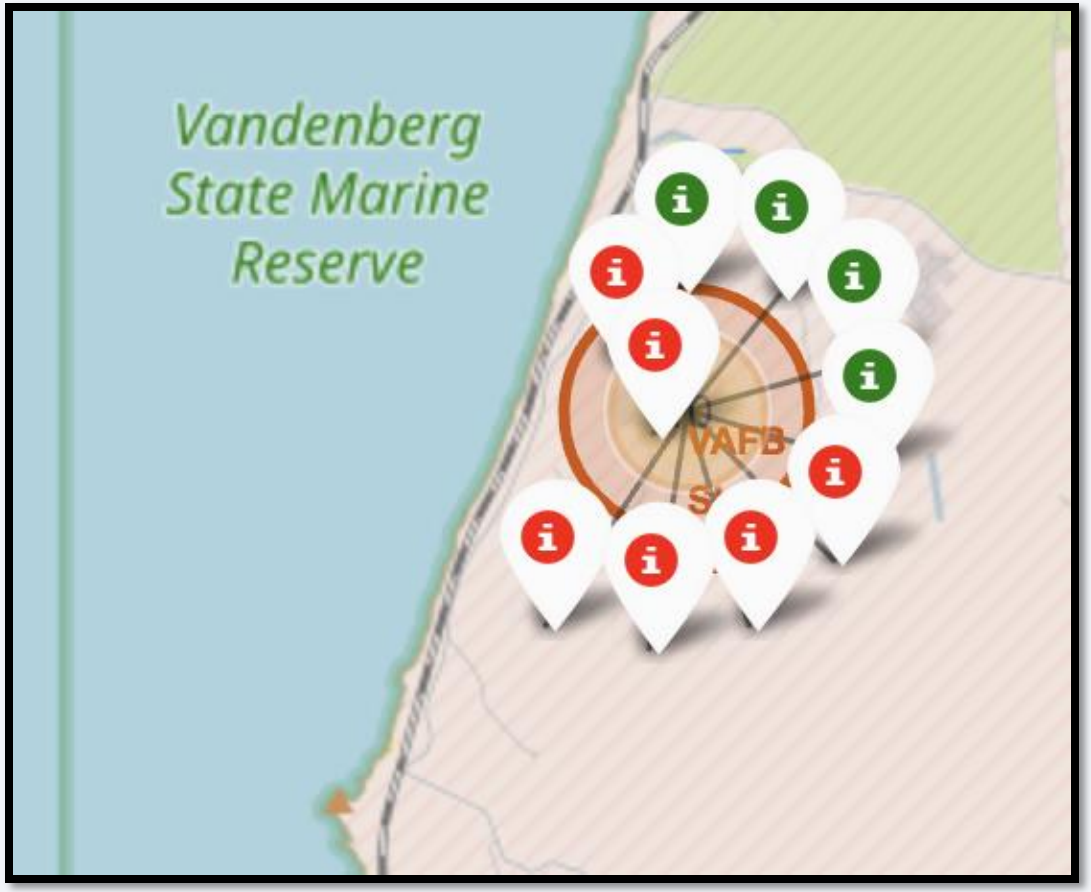


# Launch Sites – East

- CCAFS LC-40 (Cape Canaveral Air Force Station Launch Complex 40 – Site 1)
- KSC LC-39A (Kennedy Space Center Launch Complex 39A)
- CCAFS SLC-40 (Cape Canaveral Air Force Station Space Launch Complex 40 – Site 2)

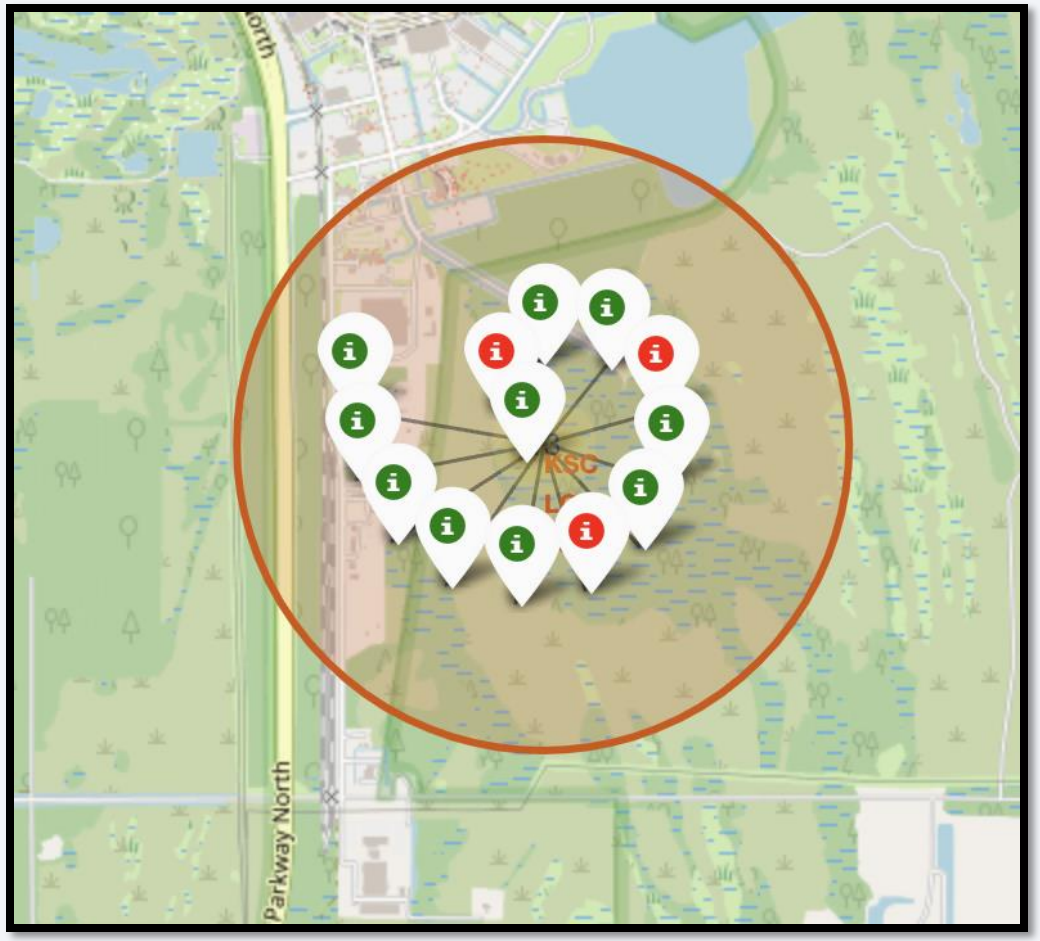
# Launch Outcomes of VAFB SLC-4E

- There has been 10 launches by VAFB SLC-4E.
- 4 of them was successful.
- 6 of them was not successful.
- Success rate is 40 % of the launches by VAFB SLC-4E.



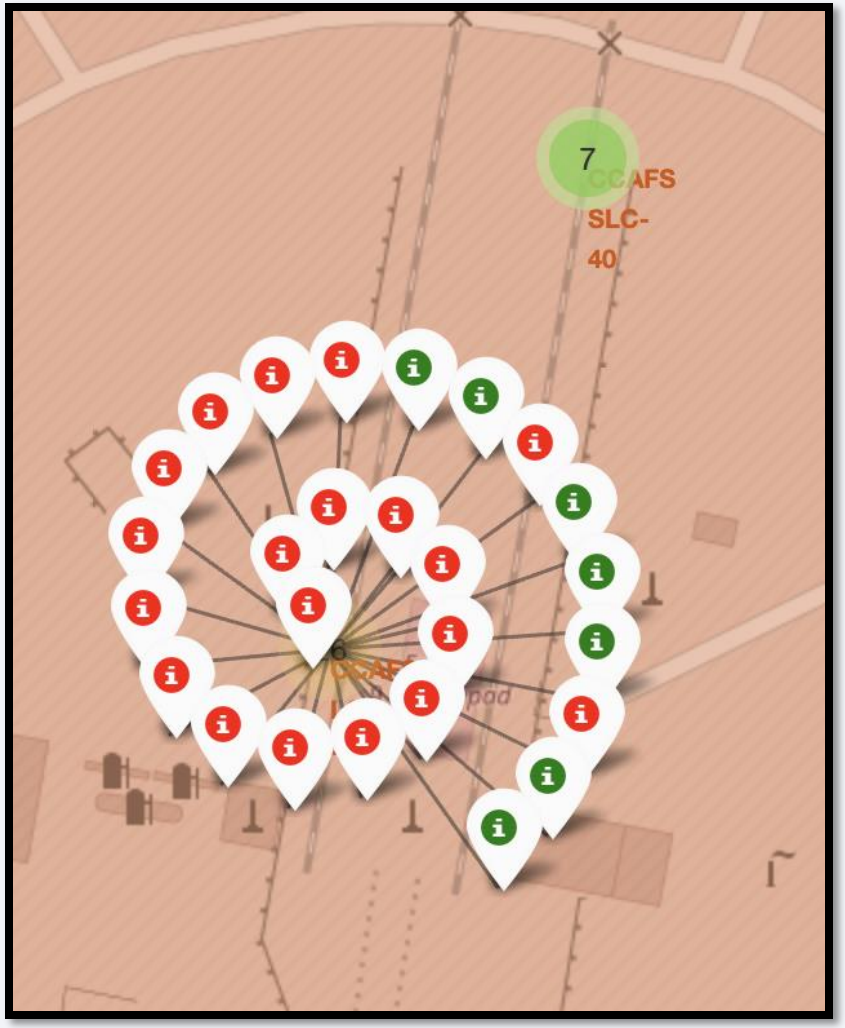
# Launch Outcomes of KSC LC-39A

- There has been 13 launches by KSC LC-39A.
- 10 of them was successful.
- 3 of them was not successful.
- Success rate is 77 % of the launches by KSC LC-39A.



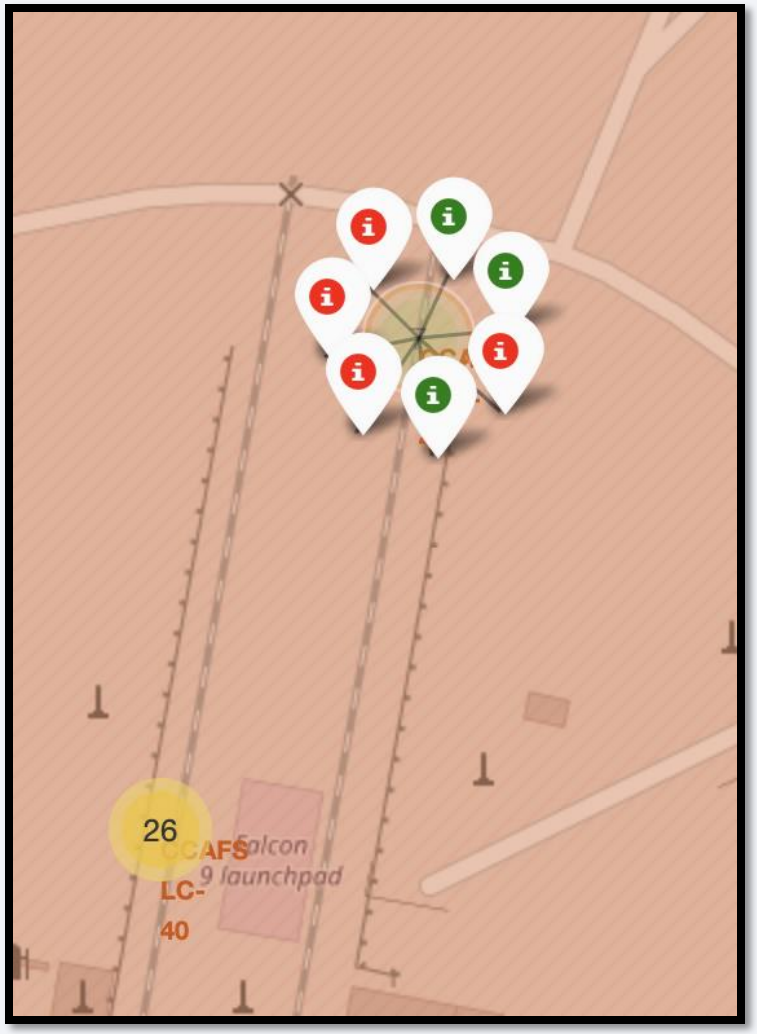
# Launch Outcomes of CCAFS LC-40 Site 1

- There has been 26 launches by CCAFS LC-40 Site 1.
- 7 of them was successful.
- 19 of them was not successful.
- Success rate is 27 % of the launches by CCAFS LC-40 Site 1.



# Launch Outcomes of CCAFS LC-40 Site 2

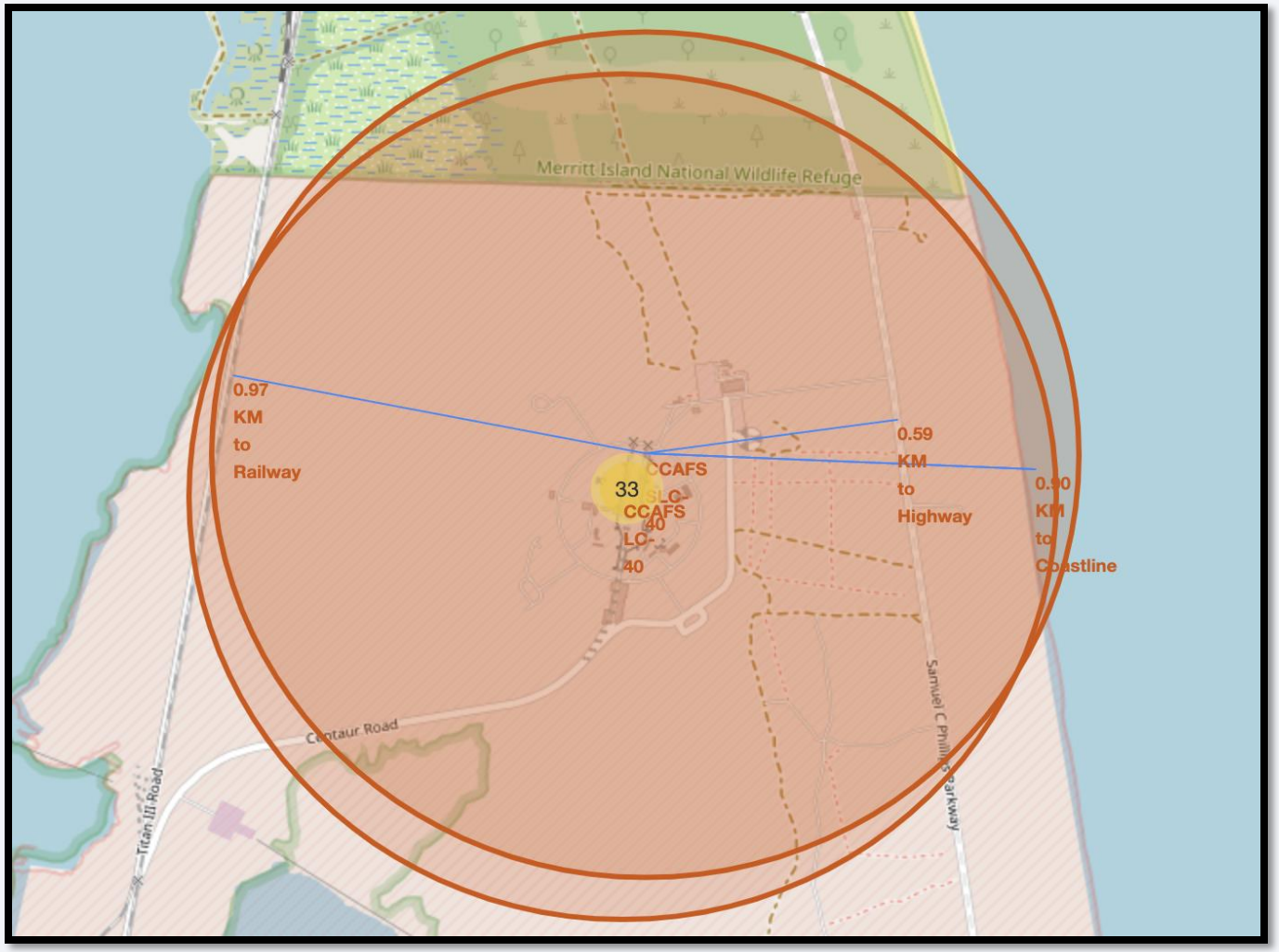
- There has been 7 launches by CCAFS LC-40 Site 2.
- 3 of them was successful.
- 4 of them was not successful.
- Success rate is 43 % of the launches by CCAFS LC-40 Site 2.

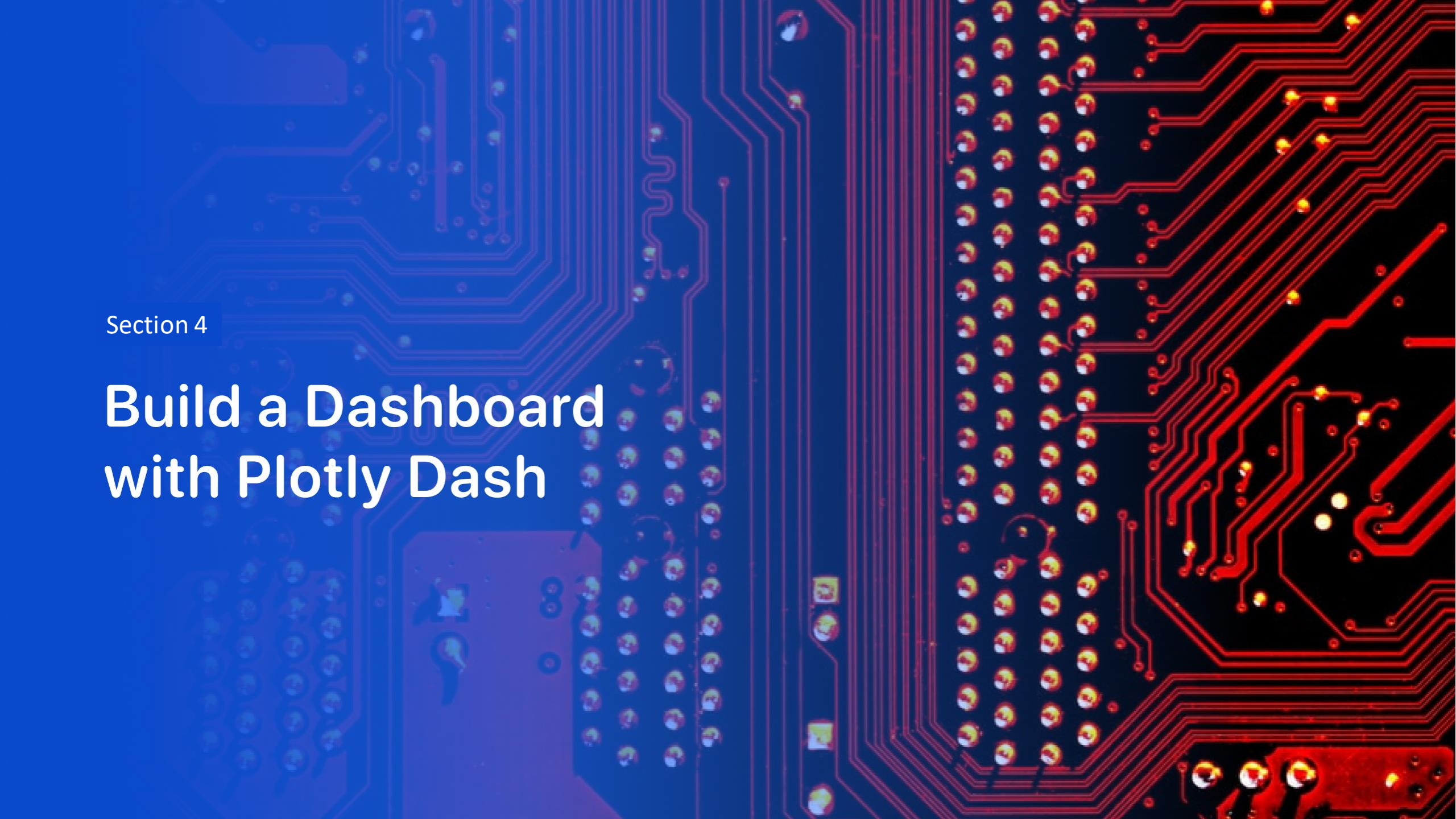


# CCAFS SLC-40 - Proximities

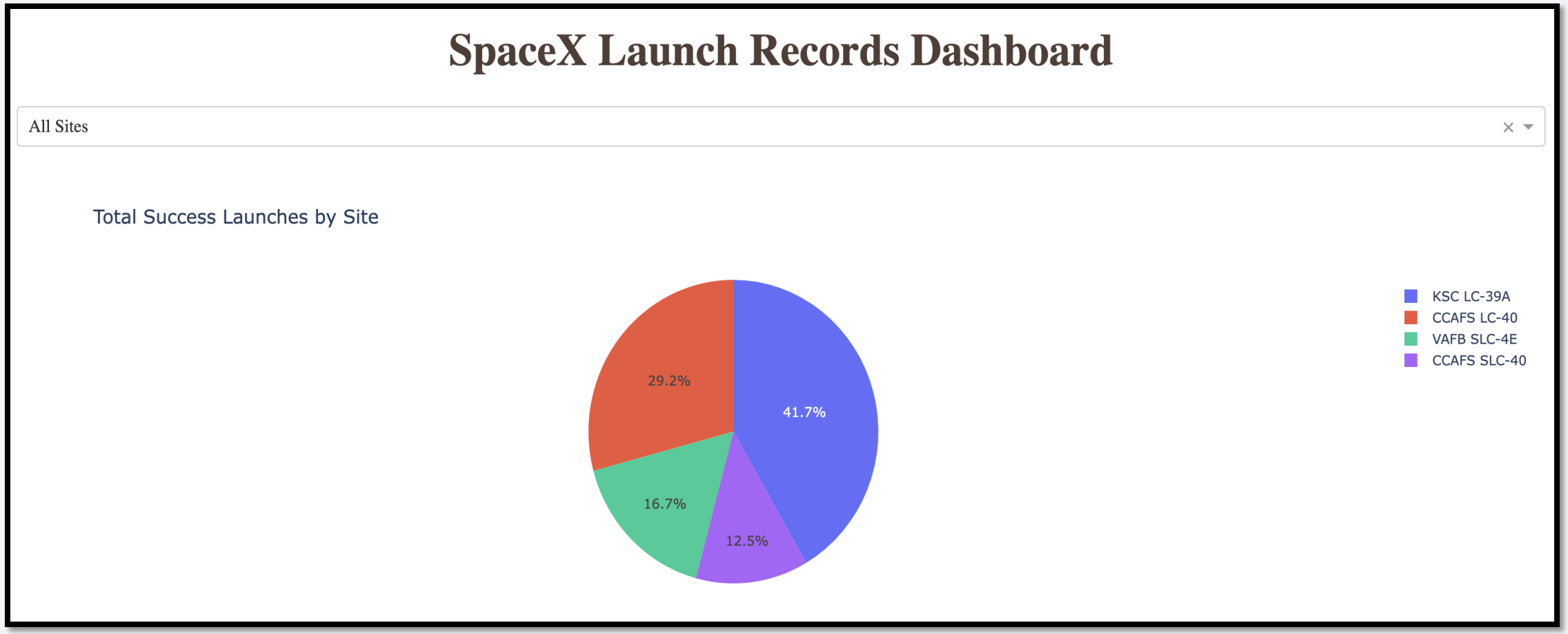
Proximities of CCAFS SLC-40 – Launch 2:

- Railway: 0.97 km
- Highway: 0.59 km
- Coastline: 0.90 km



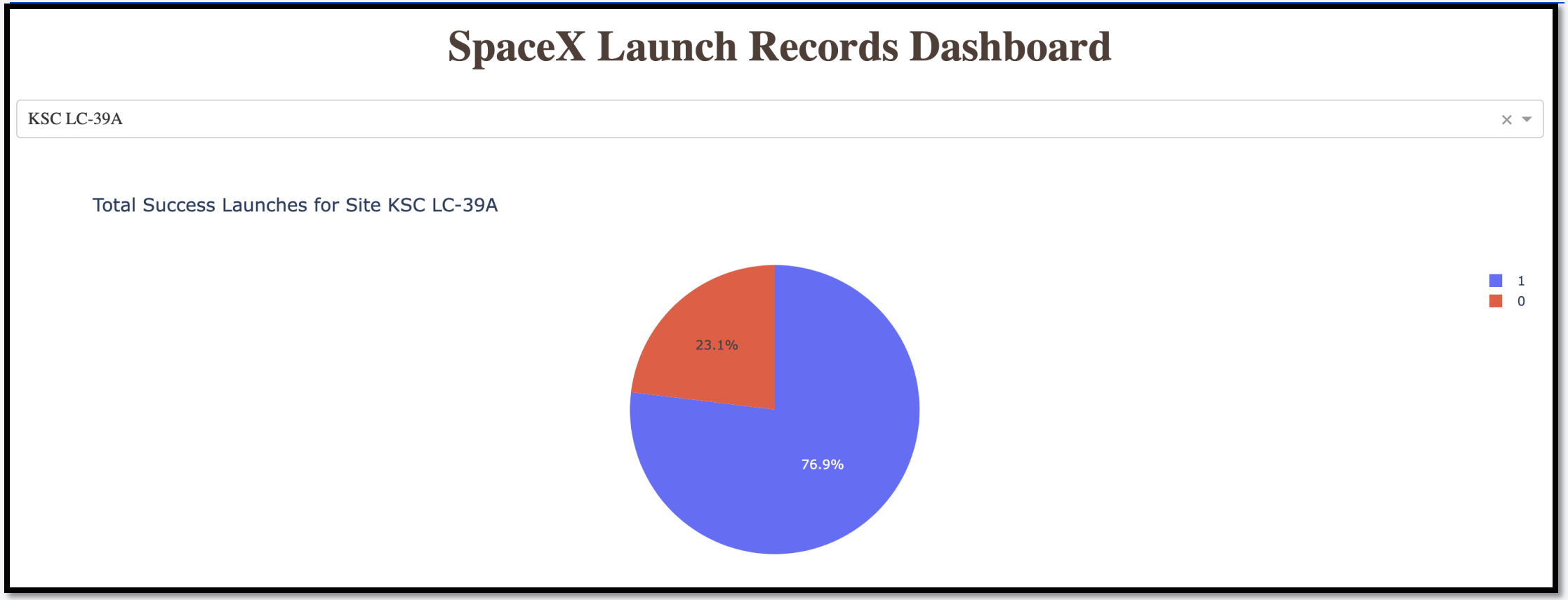

Section 4
Build a Dashboard with Plotly Dash

# Success Rates of Launch Sites



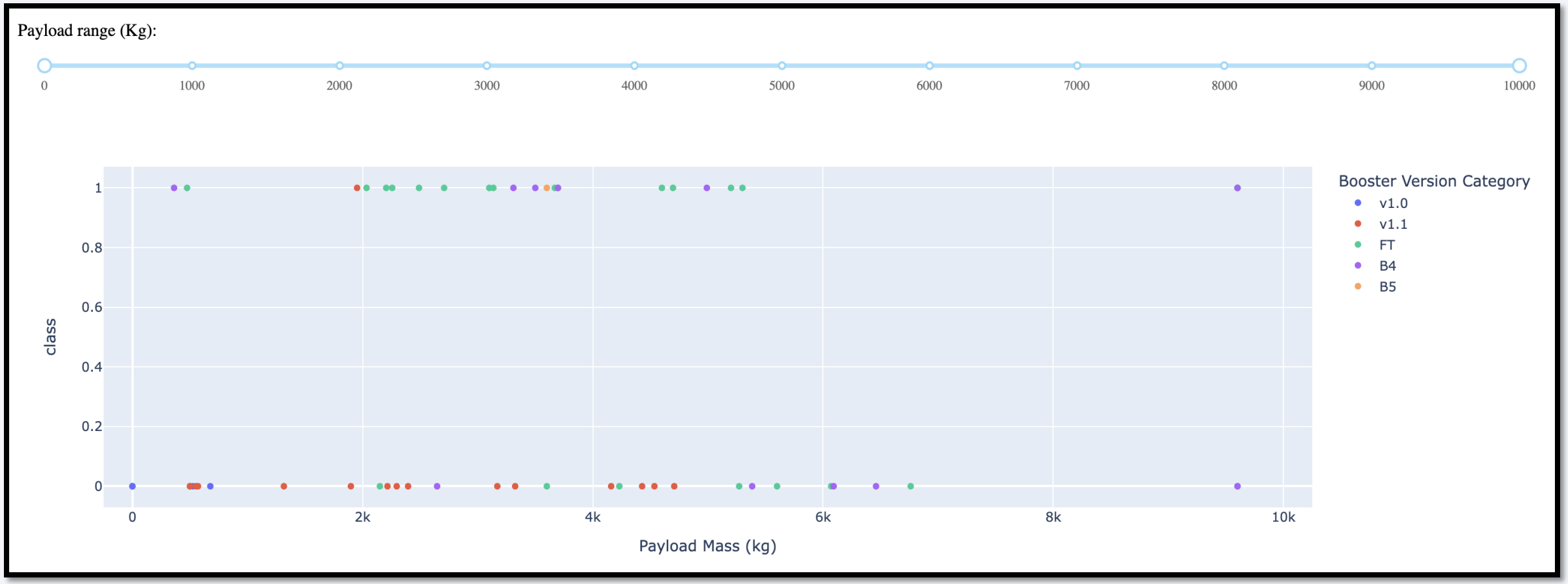

- As we can see above, KSC LC-39A has 41.7% of share for successful launches.

# Success Rates of Launch Sites



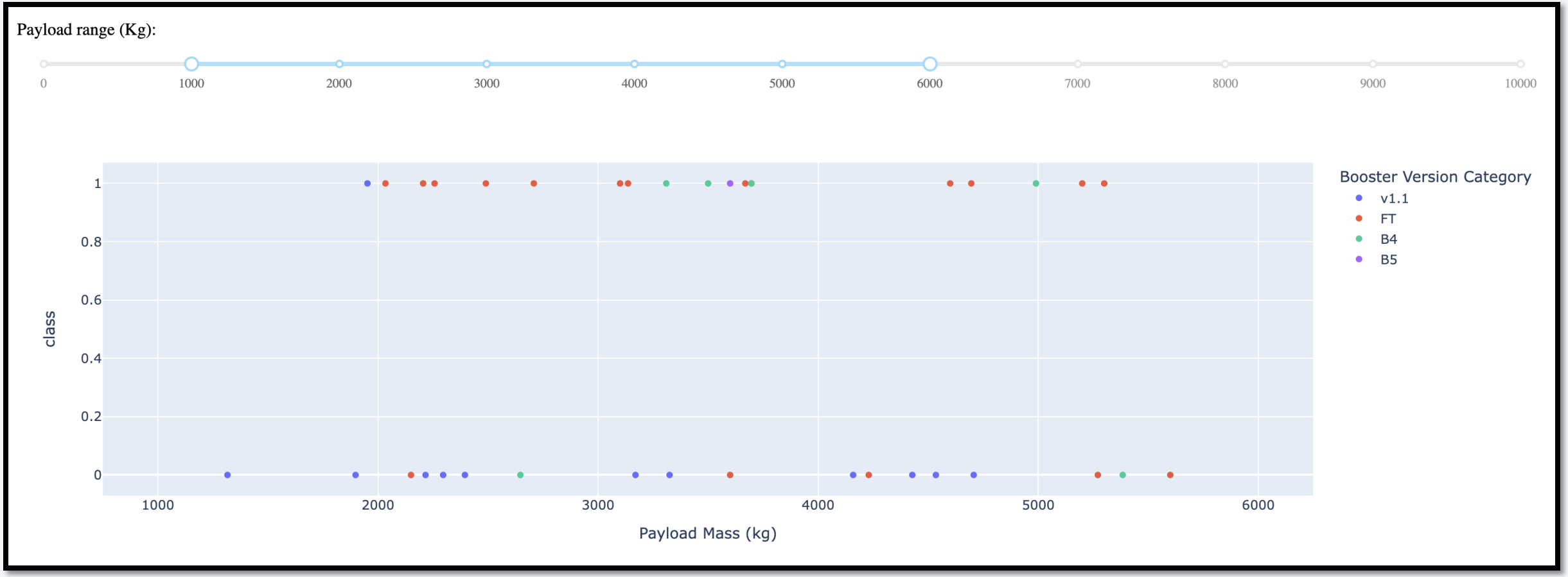

- Launches by KSC LC-39A was successful with 76.9%.

# Success Rates based on Booster Version Category




•Overall, the Booster Version 'FT' mostly achieved success while 'v1.1' experienced failures.

# Success Rates based on Payload Mass



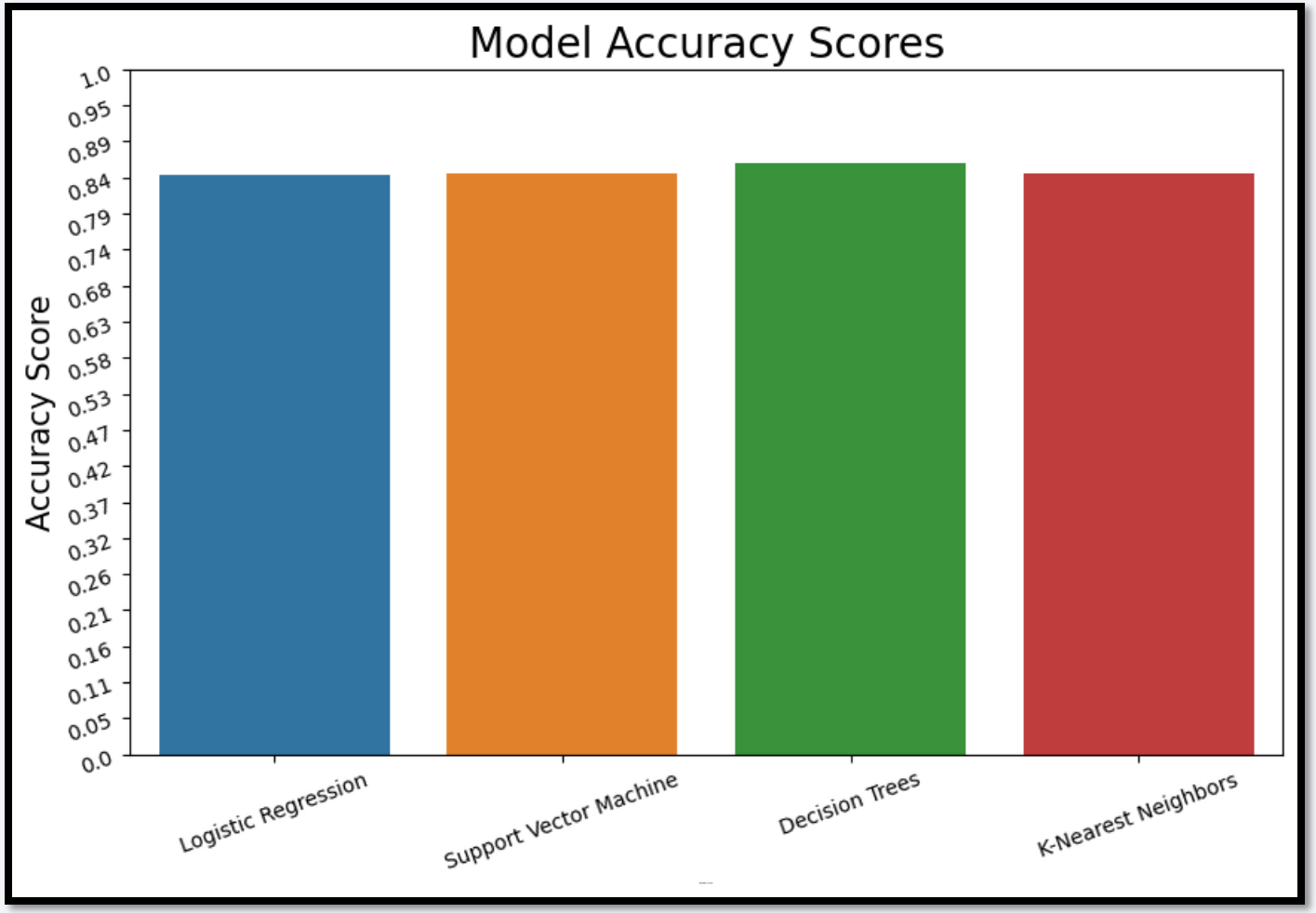

- We observe successful performance in payloads ranging from 1800 to 5000.

Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

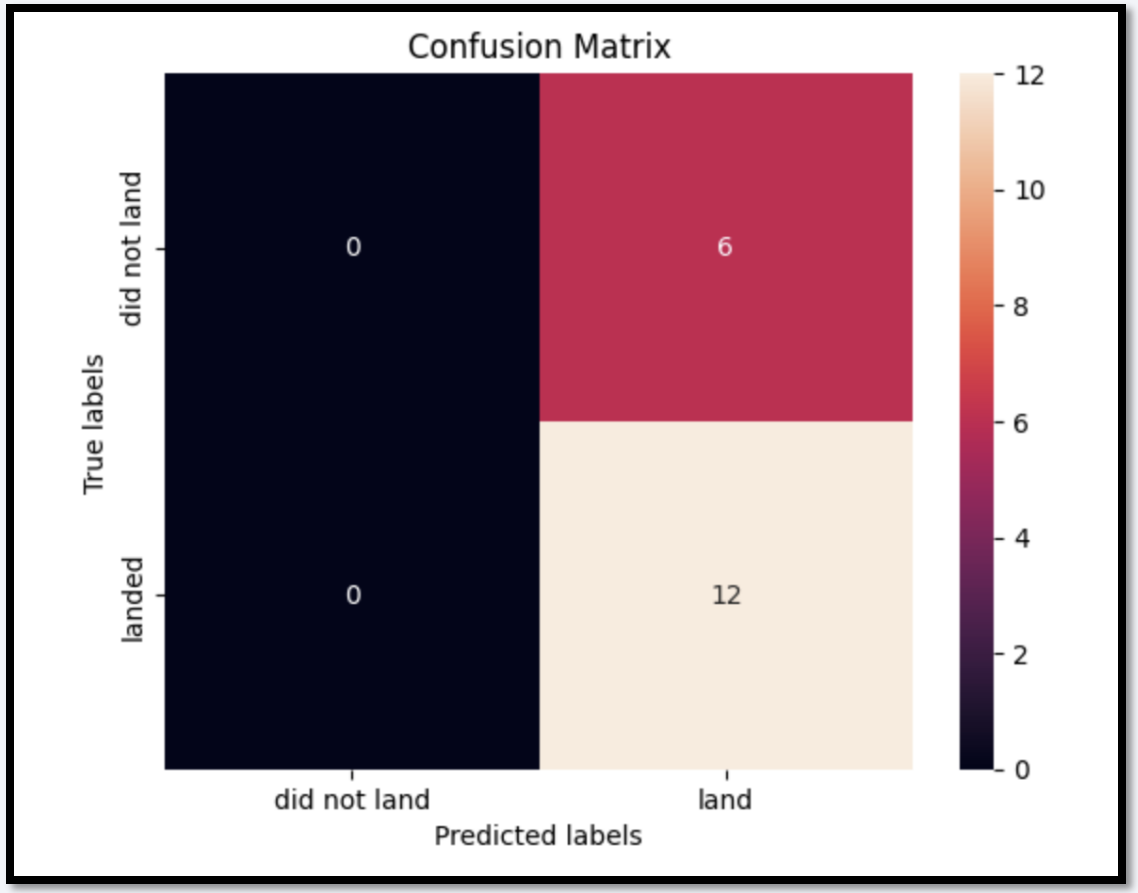
- Logistic Regression, Support Vector Machines, Decision Trees, K-Nearest Neighbors models has been trained.
- Accuracy of Decision Tree has the highest score with 0.86.
- LR Score: 0.85
- SVM Score: 0.85
- KNN Score: 0.85



# Confusion Matrix

Confusion Matrix of the best performing model (Decision Tree).

- There are 12 instances correctly identified as 'landed' (true positives).
- 6 instances are incorrectly identified as positive when they are actually 'did not land' (false positives).
- There are no instances where the model incorrectly identified 'did not landed' as 'landed' (false negatives), and there are no instances where 'did not landed' were correctly identified as 'did not landed' (true negatives).



# Conclusions

- From 2010 to 2020, there has been a consistent upward trend in launch success.
- KSC LC-39A boasts the highest success rate among all launch sites, standing at 42% overall and 77% for its own launches. In contrast, CCAFS LC-40 Site 1 only maintains a success rate of 13%, despite conducting 26 launches.
- The date of the first successful landing on the ground pad is December 22, 2015.
- The 'FT' booster version demonstrated mostly successful outcomes, while 'v1.1' experienced failures.
- Successful payloads predominantly fell within the range of 1800 to 5000.
- The resulting accuracy scores on the test dataset varied across the models, with Decision Trees achieving the highest accuracy of 0.875, followed closely by SVM with an accuracy of 0.848. Logistic Regression and KNN also demonstrated competitive accuracy scores of 0.846 and 0.848, respectively.

# Appendix

*Wikipedia link that Falcon 9 launch records were extracted from HTML table*

- *https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches*

*For more detailed information regarding historical launch data of SpaceX*

- *https://www.spacexstats.xyz/#launchhistory-per-year*

# Thank you!