IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Yufan Li
July 28, 2023

# Outline

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

# Executive Summary

- Summary of methodologies
  - Data collection
  - Data wrangling
  - Exploratory data analysis (EDA)
  - Interactive visual analytics
  - Predictive analysis
- Summary of all results
  - Exploratory data analysis results
  - Interactive analytics demo in screenshots
  - Predictive analysis results

# Introduction

- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

- Problems we want to find answers
  - Which factors impacts the success rate?
  - How these factors impacts the success rate?
  - What are the characteristics of the location of the launch sites?
  - What model performances best when predicting the success rate?

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology:
  - Using SpaceX API and web scraping
- Perform data wrangling
  - Using Pandas and auxiliary functions
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using scikit-learn, GridSearchCV and confusion matrix

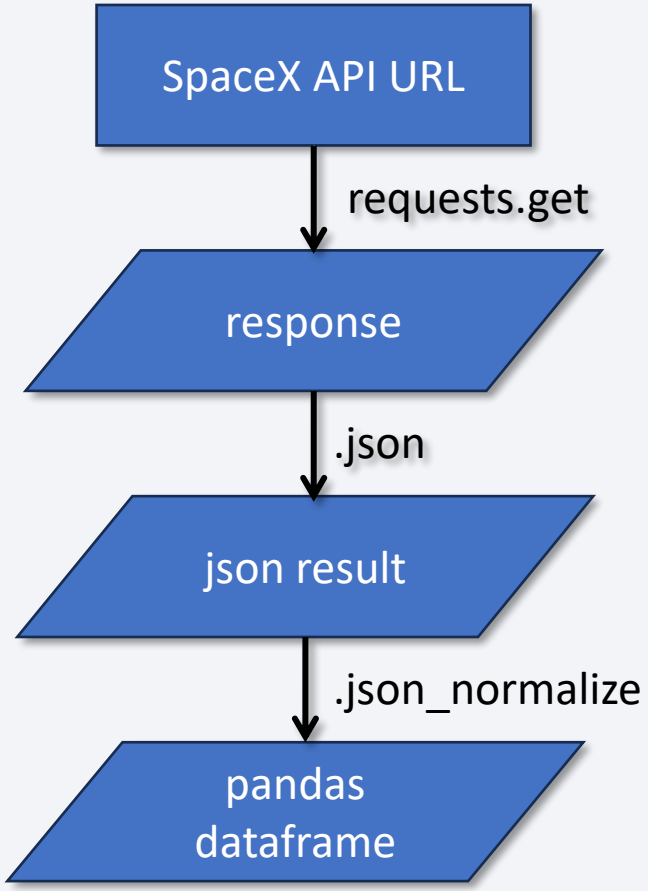
# Data Collection

- Request data from SpaceX REST API
- Web Scraping from List of Falcon 9 and Falcon Heavy launches Wikipedia page

# Data Collection – SpaceX API

- Get JSON data from SpaceX API using requests
- Load data into a Pandas DataFrame
- Manipulate data using Pandas and defined auxiliary functions

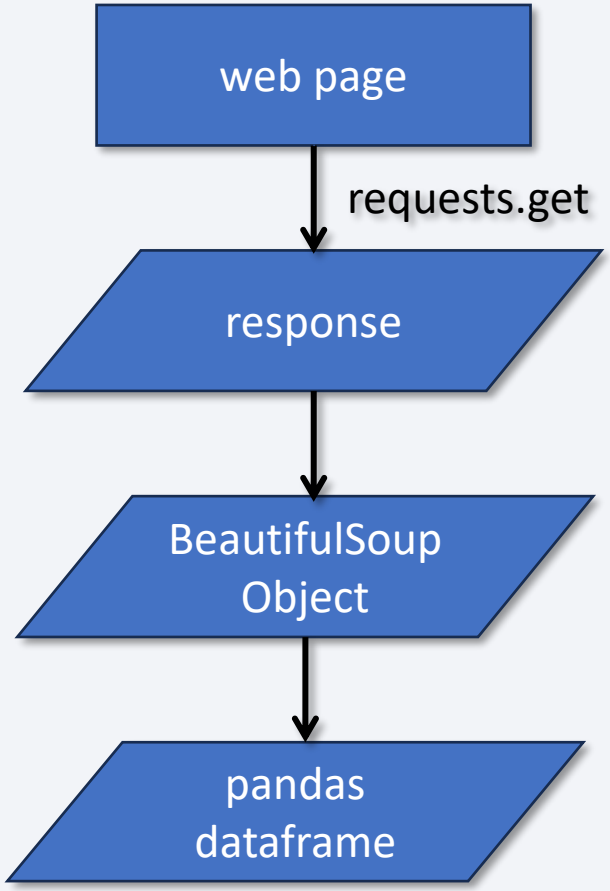
https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week1/Collecting_the_Data.ipynb



# Data Collection - Scraping

- Get HTML page from Wikipedia using requests
- Create a BeautifulSoup Object from the response
- Extract tables from the page
- Load data into a Pandas DataFrame

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week1/Data_Collection_with_Web_Scraping.ipynb



# Data Wrangling

- Load SpaceX dataset
- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week1/Data_Wrangling.ipynb

# EDA with Data Visualization

- Scatter plot is suitable for visualizing the relationship between an independent variable and an independent variable. For example, I used it to visualize the relationship between Flight Number and Launch Site.
- Bar chart can represent numerical and categorical variables. I used it to visualize the relationship between success rate of each orbit type.
- Line plot is best suited for trend-based visualizations of data over a period of time. I used it to visualize the launch success yearly trend.

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week2/Exploratory_Analysis_Using_Pandas_and_Matplotlib.ipynb

# EDA with SQL

- Explore what launch sites there are
- Explore the payload mass with different conditions
- Explore the landing outcomes with different conditions

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week2/Exploratory_Analysis_Using_SQL.ipynb

# Build an Interactive Map with Folium

- Create Circle and Marker objects to mark all launch sites on map.
- Create Cluster objects to add all launch outcomes for each site, Cluster can simplify map since many launch records will have the exact same coordinate.
- Create MousePosition object just so we can easily find the coordinates of any points of interests.
- Create PolyLine objects so we can clearly see the distance between a location to the launch sites.

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week3/Launch_Sites_Locations_Analysis_with_Folium.ipynb

# Build a Dashboard with Plotly Dash

- Add a Launch Site Drop-down Input Component so we can specify whether a pie chart of all sites or a specific site to display.
- Add a pie chart where its values based on the selected site, so we can see the proportions of success or failure.
- Add a Range Slider to select Payload withing specific range.
- Add a scatter plot shows the relationship between payload mass and outcome.

https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week3/spacex_dash_app.py

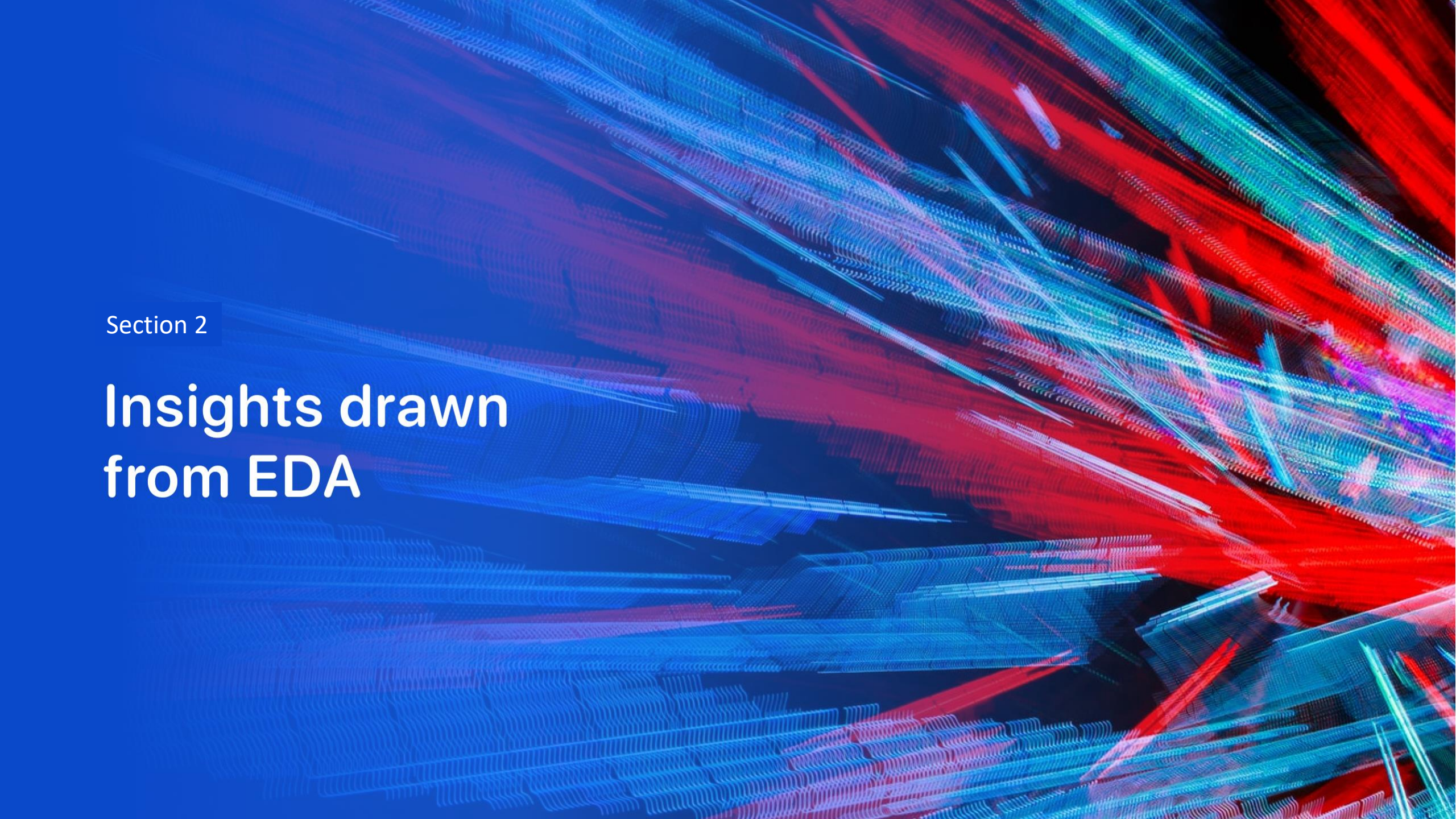
# Predictive Analysis (Classification)

- Load the DataFrame and define an auxiliary function to plot the confusion matrix.
- Standardize the independent variables.
- Split the data into training and testing data.
- Create models such as SVM, Classification Trees and Logistic Regression.
- Fit the object, find the best parameters.
- Calculate the accuracy and look at the confusion matrix.

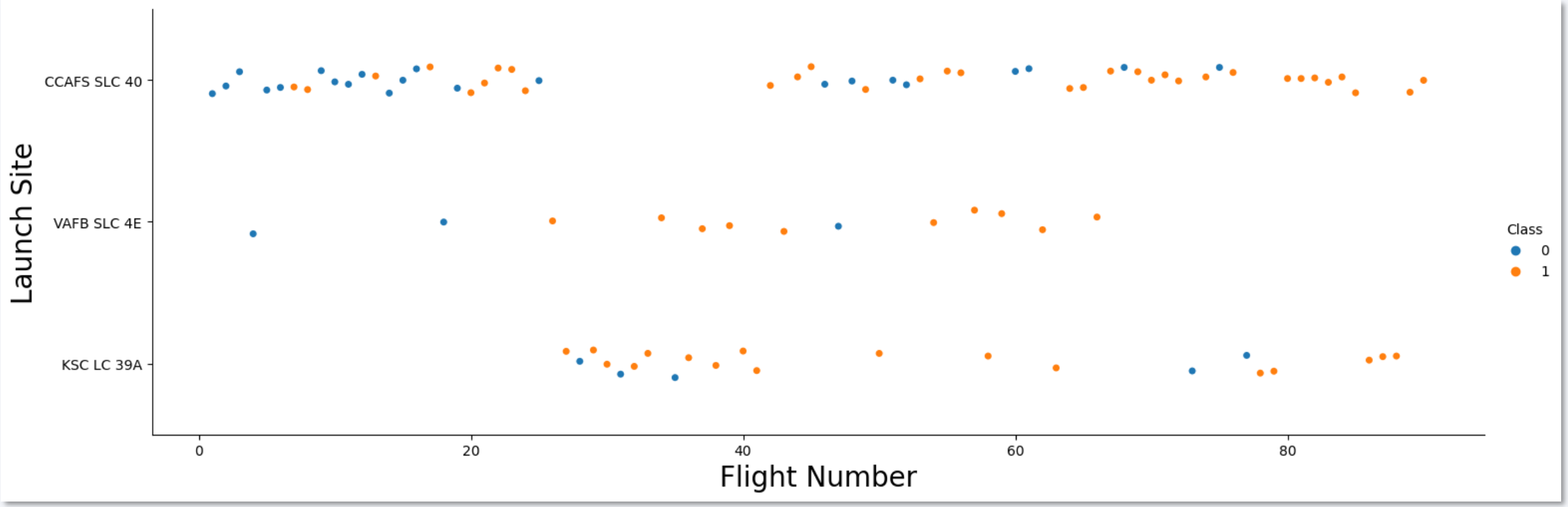
https://github.com/imyufanli/ibm-data-science-assignments/blob/main/10.Applied_Data_Science_Capstone/week4/SpaceX_Machine_Learning_Prediction.ipynb

# Results

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

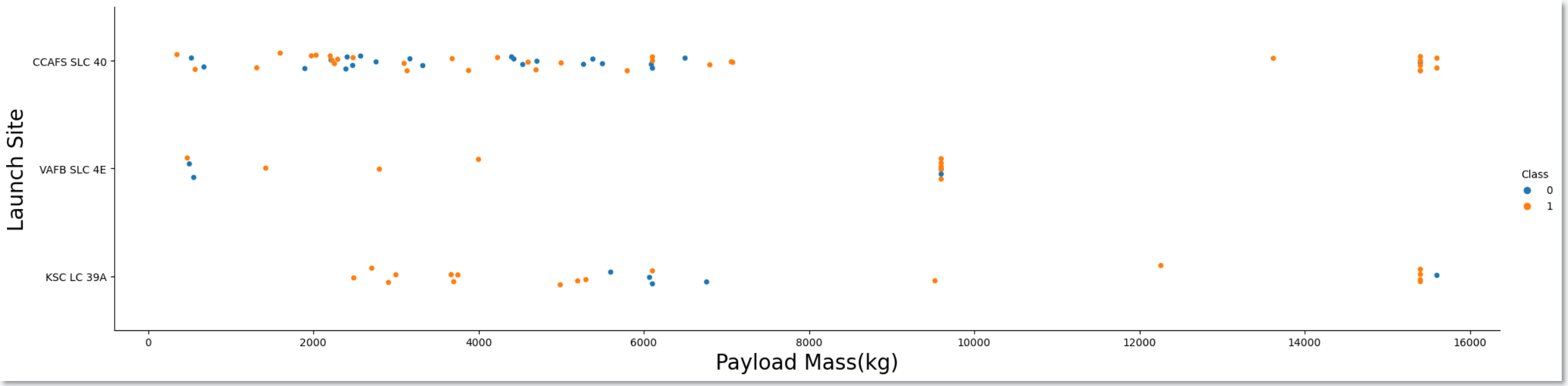

Section 2
Insights drawn from EDA

# Flight Number vs. Launch Site




We can see it appears that as the number of launch attempts increases, the success rates also increase.

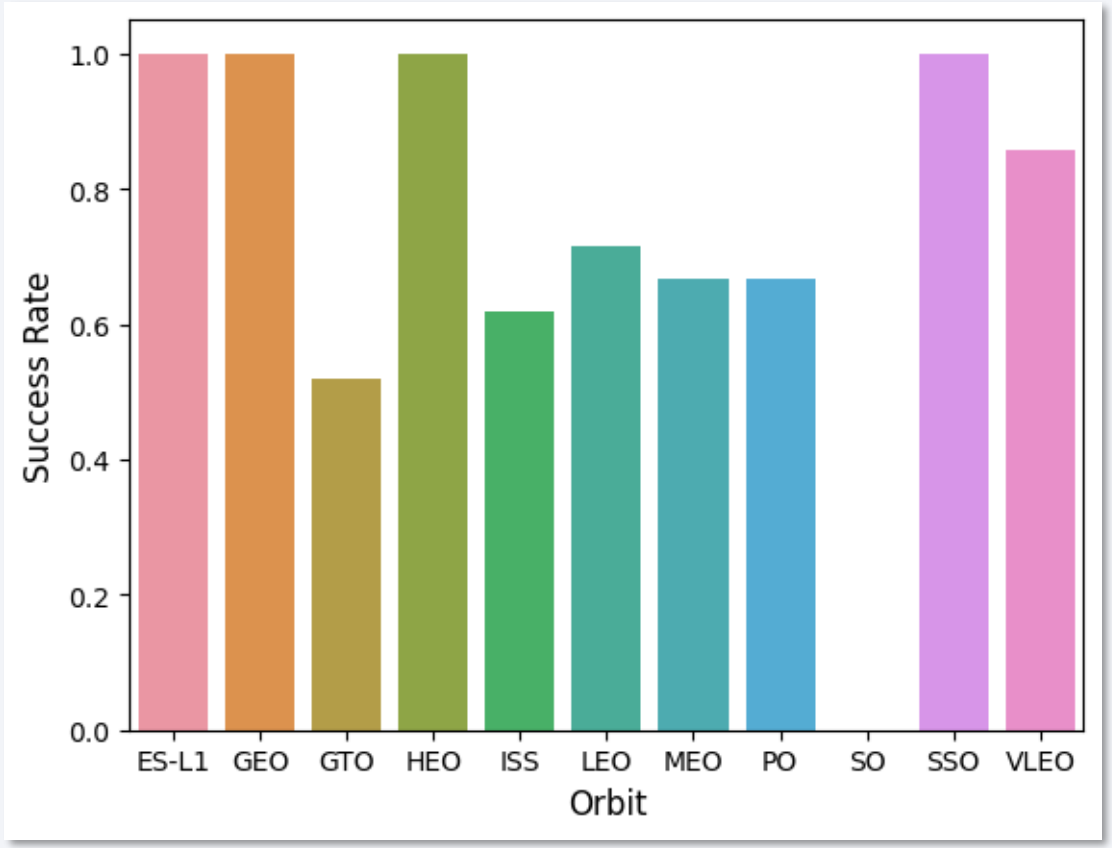
# Payload vs. Launch Site




We can see that most rocket launches load below 8000kg.

Also, for CCAFS SLC 40 rocket launches with payload mass below 8000kg, the success rate is 50-50.

# Success Rate vs. Orbit Type

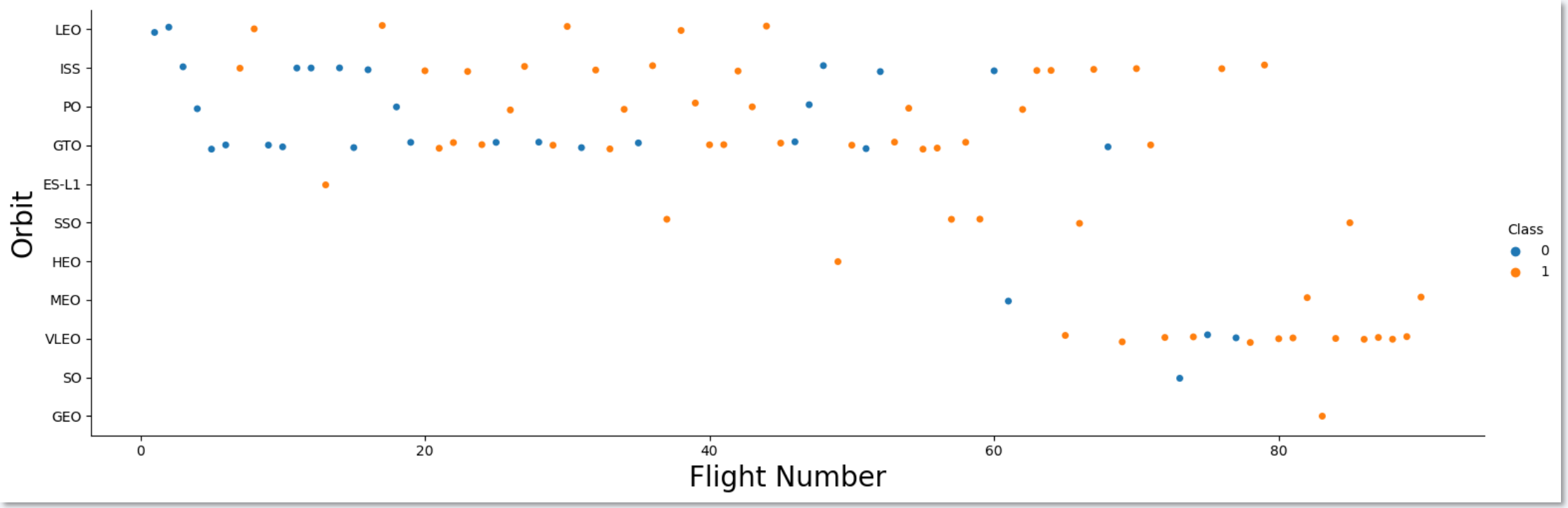



ES-L1, GEO, HEO, and SSO seem to have a success rate of about 100%.

GTO, ISS, LEO, MEO, PO, and VLEO has a success rate of around 50% to 90%.

SO has the lowest success rate, which is 0%.

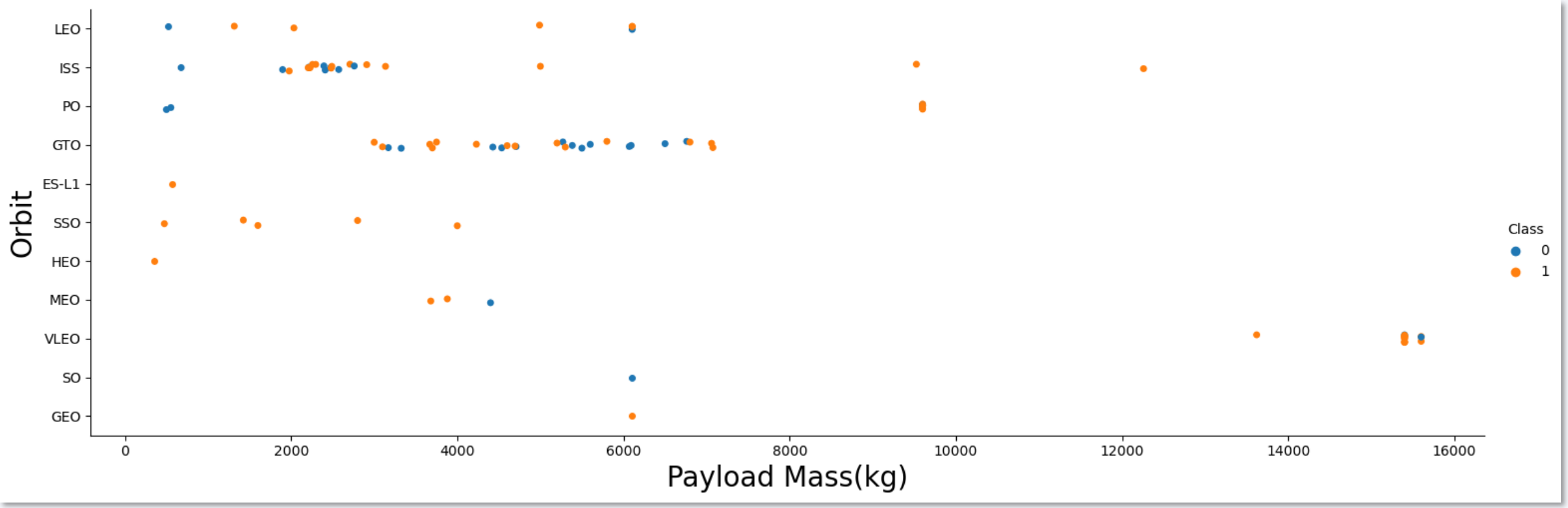
# Flight Number vs. Orbit Type




SpaceX start to launch into LEO, ISS, PO, GTO, and ES-L1 at the very beginning.

But they start to launch into SSO and HEO after the 35$^{th}$. Start to launch into MEO, VLEO, SO, and GEO very late, till the 60$^{th}$ attempt.

# Payload vs. Orbit Type



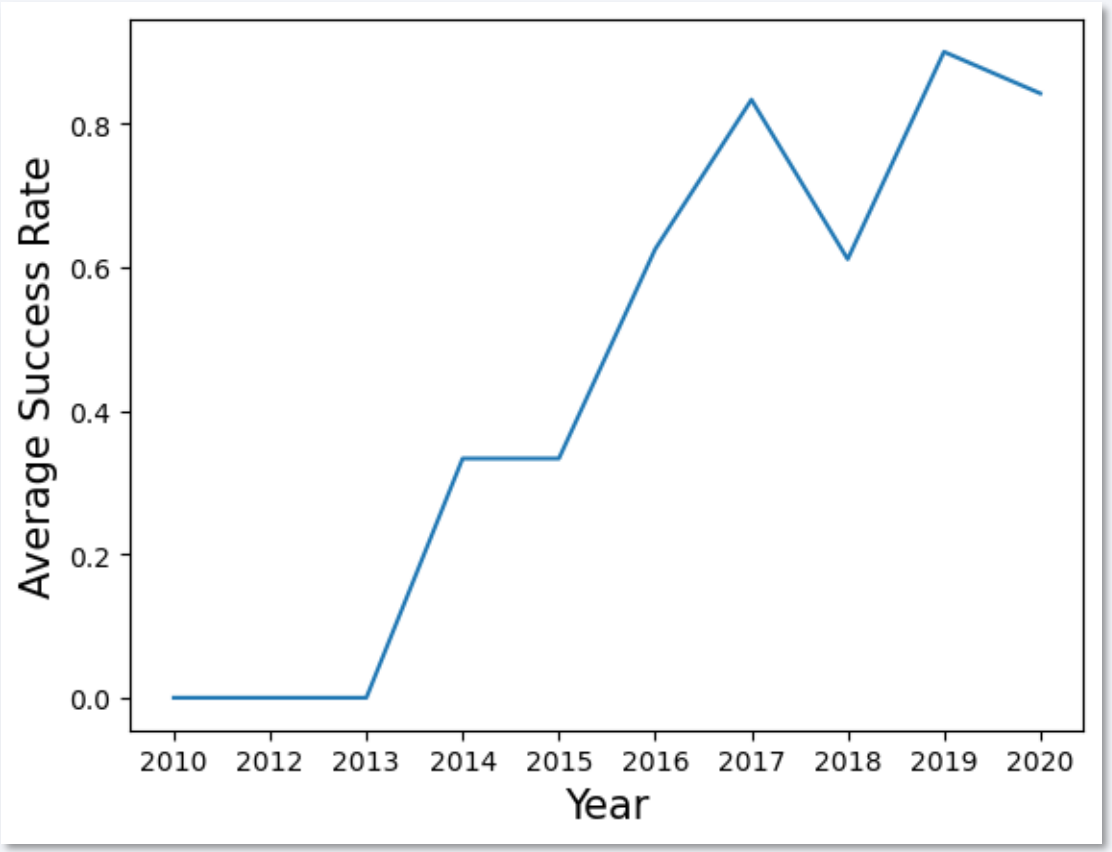

Most of the launch loads are concentrated below 8000kg.

Especially for ISS, loads of about 1500kg to 3700kg, and for GTO loads of about 2000kg to 8000kg.

# Launch Success Yearly Trend

We can clearly see that the average success rate trend increases over time.

Although in the years 2018 and 2020, there was a minor decrease.



# All Launch Site Names

There are 4 launch sites in total.

| Launch_Site |
| --- |
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |
| None |

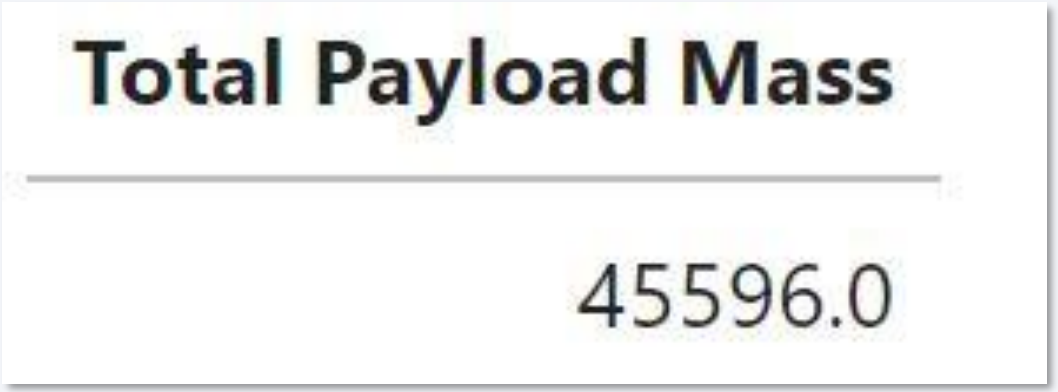
# Launch Site Names Begin with 'CCA'

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD_MASS__KG_ | Orbit | Customer | Mission_Outcome | Landing_Outc |
|---|---|---|---|---|---|---|---|---|---|
| 06/04/2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0.0 | LEO | SpaceX | Success | Failure (paracl |
| 12/08/2010 | 15:43:00 | F9 v1.0 B0004 | CCAFS LC-40 | Dragon demo flight C1, two CubeSats, barrel of Brouere cheese | 0.0 | LEO (ISS) | NASA (COTS) NRO | Success | Failure (paracl |
| 22/05/2012 | 7:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525.0 | LEO (ISS) | NASA (COTS) | Success | No atte |
| 10/08/2012 | 0:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500.0 | LEO (ISS) | NASA (CRS) | Success | No atte |
| 03/01/2013 | 15:10:00 | F9 v1.0 B0007 | CCAFS LC-40 | SpaceX CRS-2 | 677.0 | LEO (ISS) | NASA (CRS) | Success | No atte |

5 Launch records where launch sites begin with the 'CCA'.

# Total Payload Mass

SpaceX launched rockets for NASA with a total payload mass of 45596kg.



# Average Payload Mass by F9 v1.1

Booster version F9 v1.1 has an average payload mass of 2928.4kg.

| Average Payload Mass |
| --- |
| 2928.4 |

# First Successful Ground Landing Date

| Date |
| --- |
| 22/12/2015 |

The date of the first successful landing outcome on ground pad was Dec 22, 2015.

# Successful Drone Ship Landing with Payload between 4000 and 6000

| Booster_Version |
| --- |
| F9 FT B1022 |
| F9 FT B1026 |
| F9 FT B1021.2 |
| F9 FT B1031.2 |

There are 4 boosters that have successfully landed on drone ships and had payload mass greater than 4000 but less than 6000.

# Total Number of Successful and Failure Mission Outcomes

Almost all missions have succeeded, only one was a failure in flight.

| Mission_Outcome | count |
|---|---|
| Failure (in flight) | 1 |
| Success | 100 |

# Boosters Carried Maximum Payload

Boosters that have carried the maximum payload mass are shown on the right.

| 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

| Month | Landing_Outcome | Booster_Version | Launch_Site |
|---|---|---|---|
| 10 | Failure (drone ship) | F9 v1.1 B1012 | CCAFS LC-40 |
| 04 | Failure (drone ship) | F9 v1.1 B1015 | CCAFS LC-40 |

There were 2 failure landing outcomes in drone ships in 2015, one in April, and another in October, both launched from CCAFS LC-40.

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

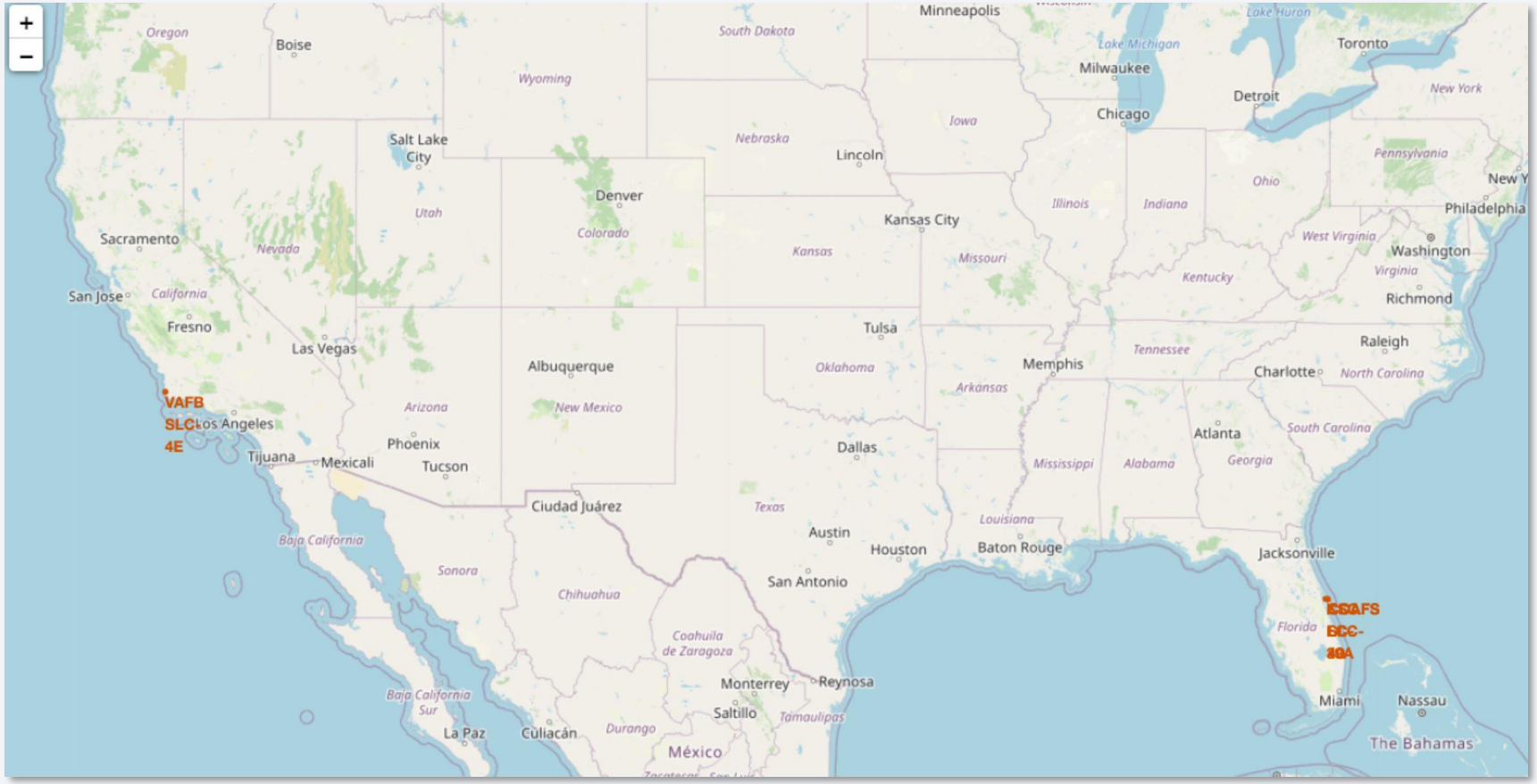
Here is the rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.

| Date | Landing_Outcome | total |
|---|---|---|
| 08/07/2018 | Success | 20 |
| 10/08/2012 | No attempt | 9 |
| 04/08/2016 | Success (drone ship) | 8 |
| 18/07/2016 | Success (ground pad) | 7 |
| 14/04/2015 | Failure (drone ship) | 3 |
| 12/05/2018 | Failure | 3 |
| 06/04/2010 | Failure (parachute) | 2 |
| 18/04/2014 | Controlled (ocean) | 2 |
| 08/06/2019 | No attempt | 1 |

Section 3

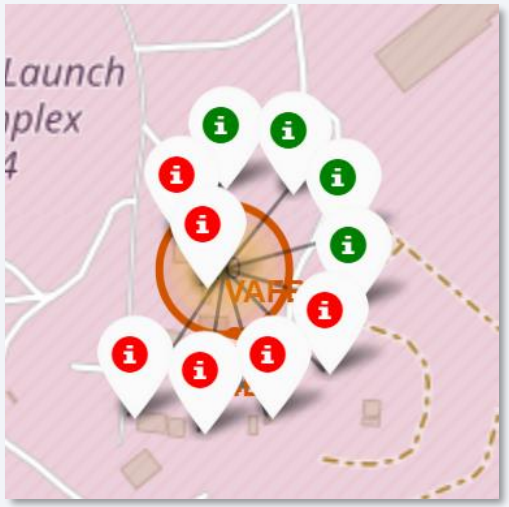
# Launch Sites Proximities Analysis

# All Launch Sites on the Map

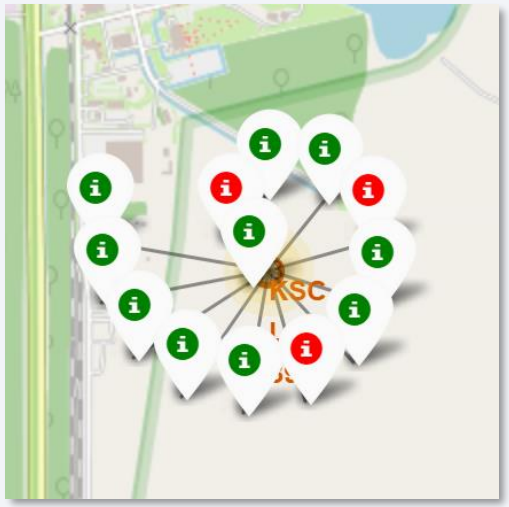



All launch sites are based near coastlines, one located in California, along the Pacific Coast, and three located in Florida, along the Atlantic Coast.

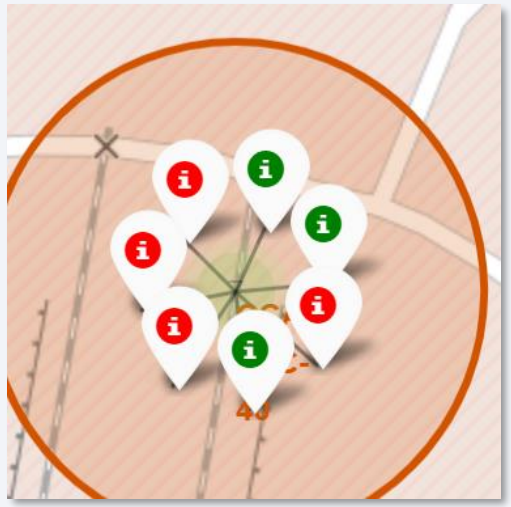
# Success/Failed Launches on the Map



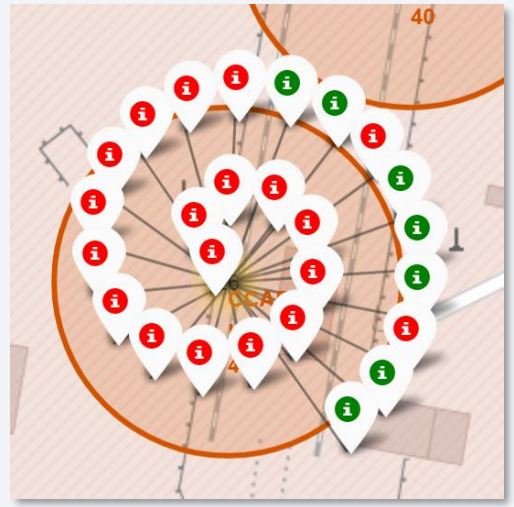

VAFB SLC-4E




KSC LC-39A



CCAFS SLC-40



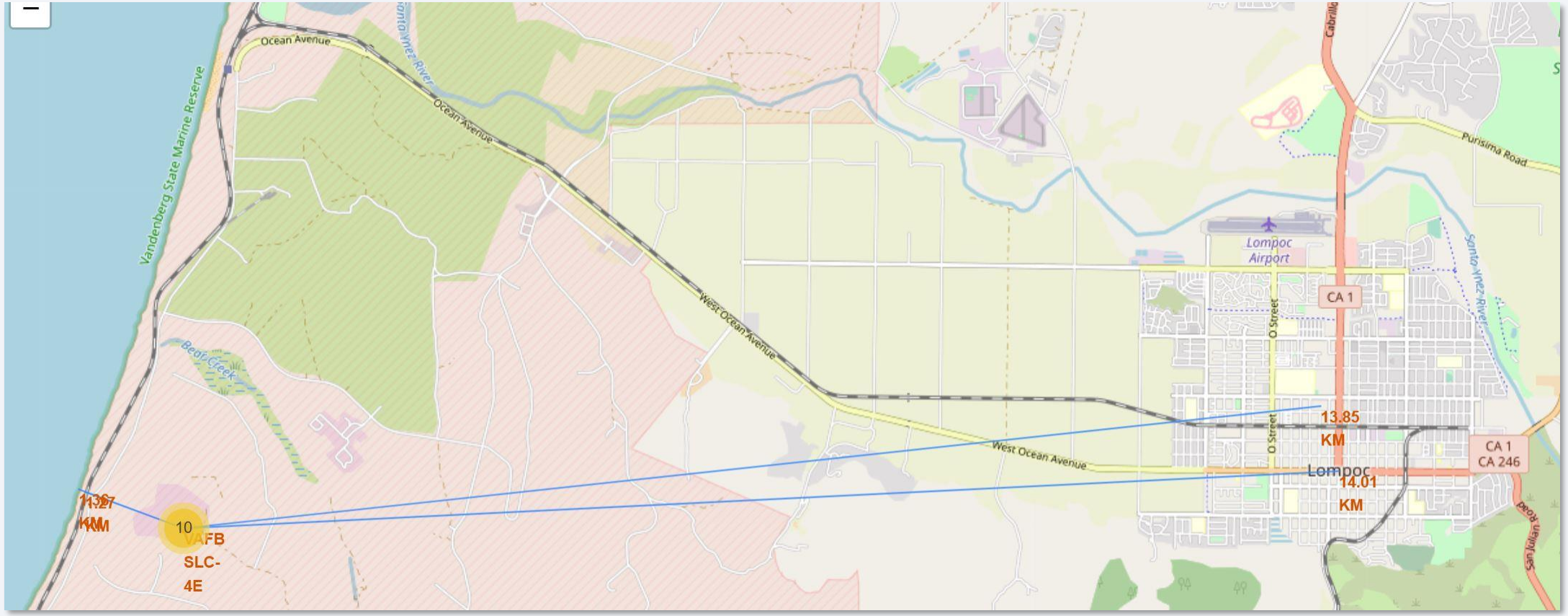

CCAFS LC-40

Obviously, KSC LC-39A has the highest launch success rate.

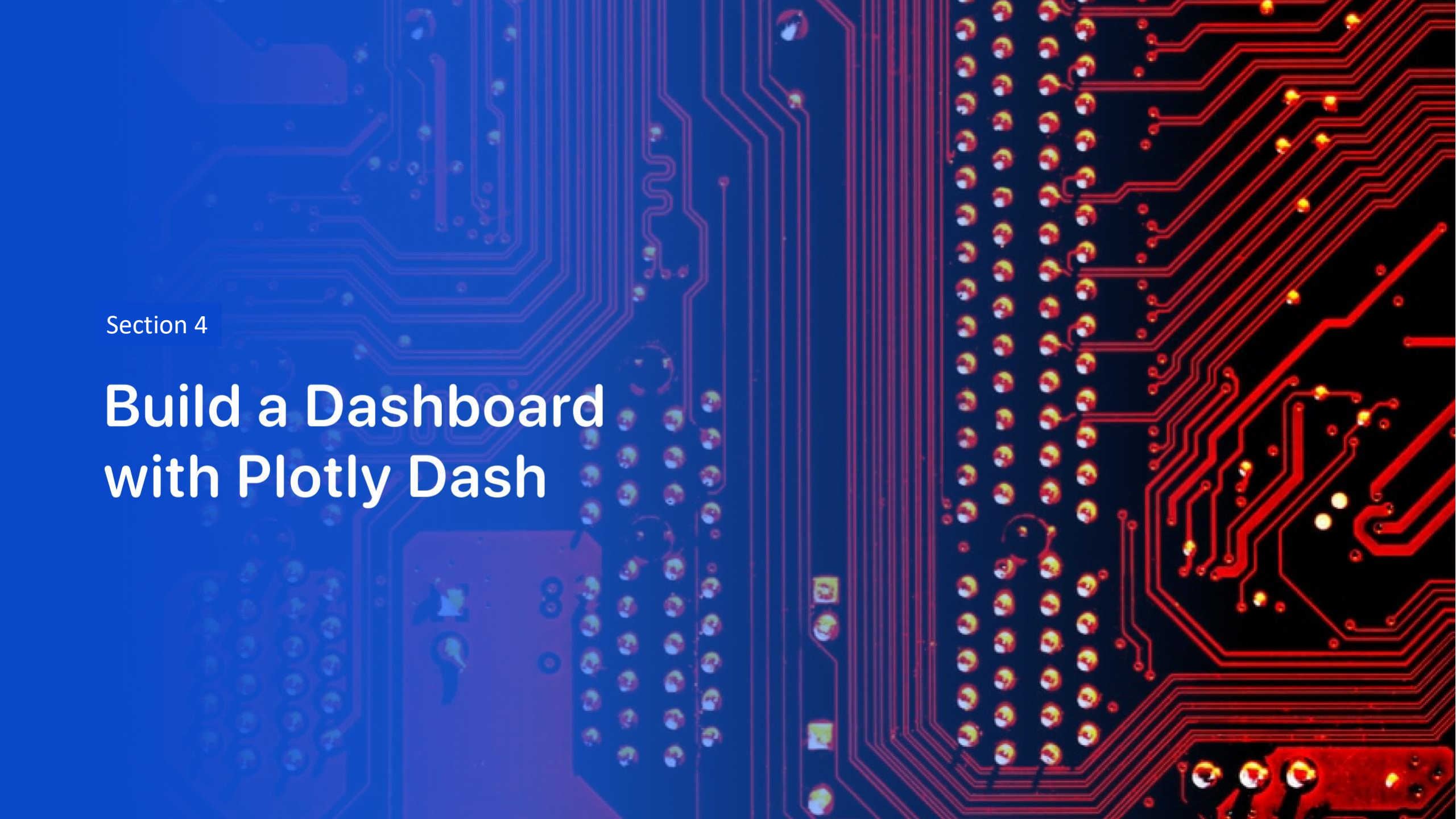
CCAFS SLC-40 has the highest launches in number, success and failed launches are near equal.

CCAFS LC-40 has the lowest launches in number, with a bad success rate.

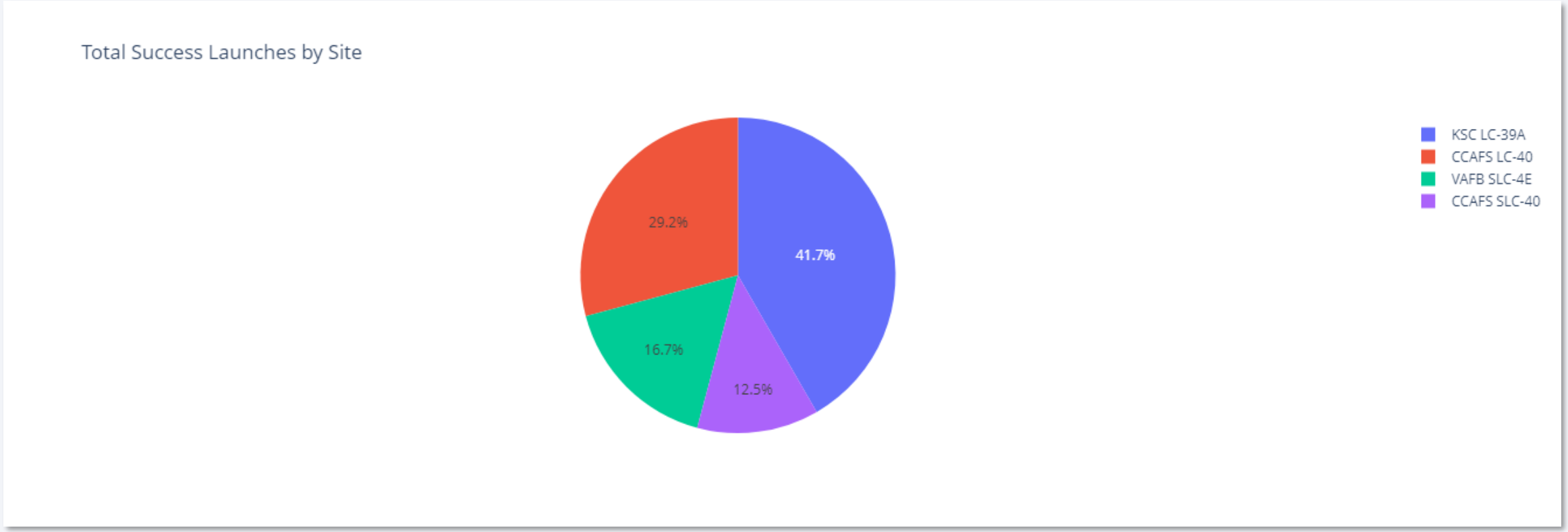
# Distances to its Proximities




All launch sites have convenient transportation, with railways and roads nearby, and keep a certain distance from residential areas.

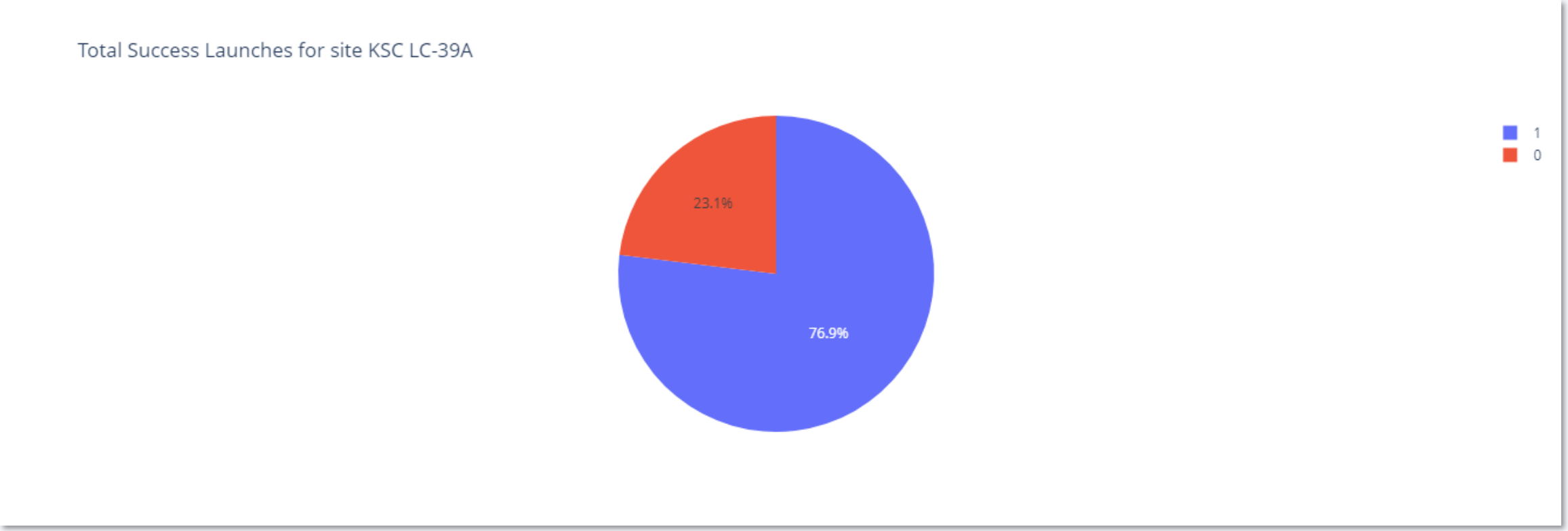

Section 4
Build a Dashboard with Plotly Dash

# Launch Success Count of All Sites



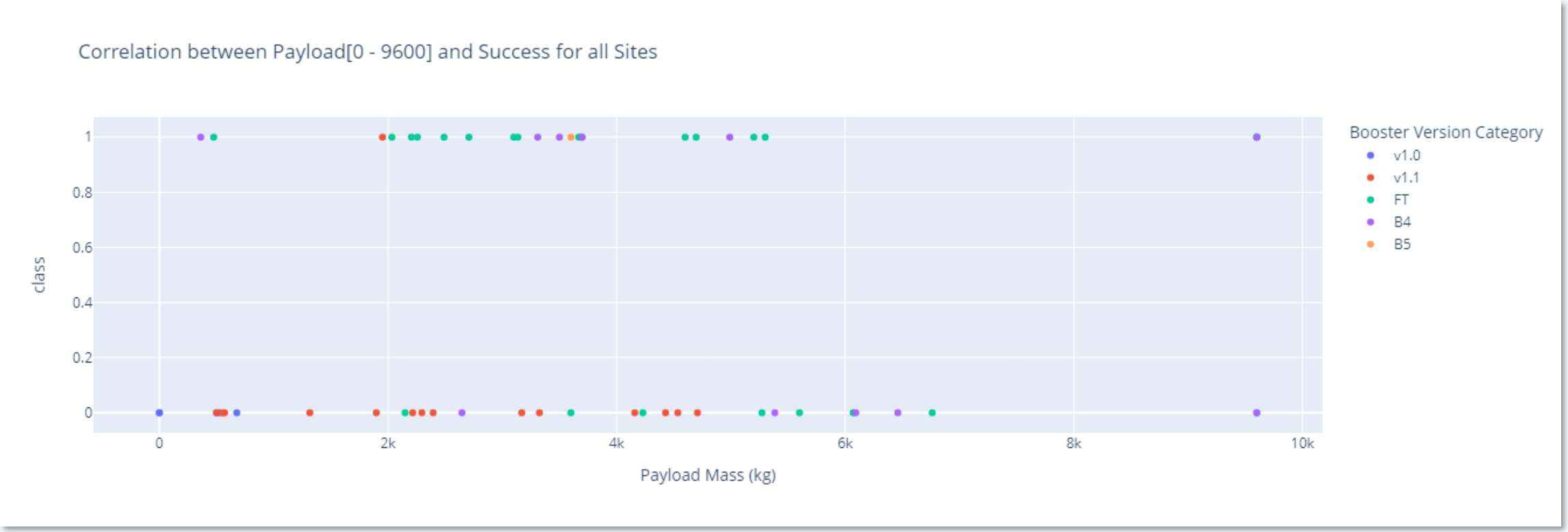

KSC LC-39A has the highest number of successful launches, accounting for 41.7% of the total, followed by CCAFS LC-40 with 29.2%. VAFB SLC-4E and CCAFS SLC-40 accounted for 16% and 12% respectively.

# Launch Site with Highest Launch Success Ratio




KSC LC-39A has the highest launch success ratio, with 76.9% successful.

# Payload vs. Launch Outcome for All Sites

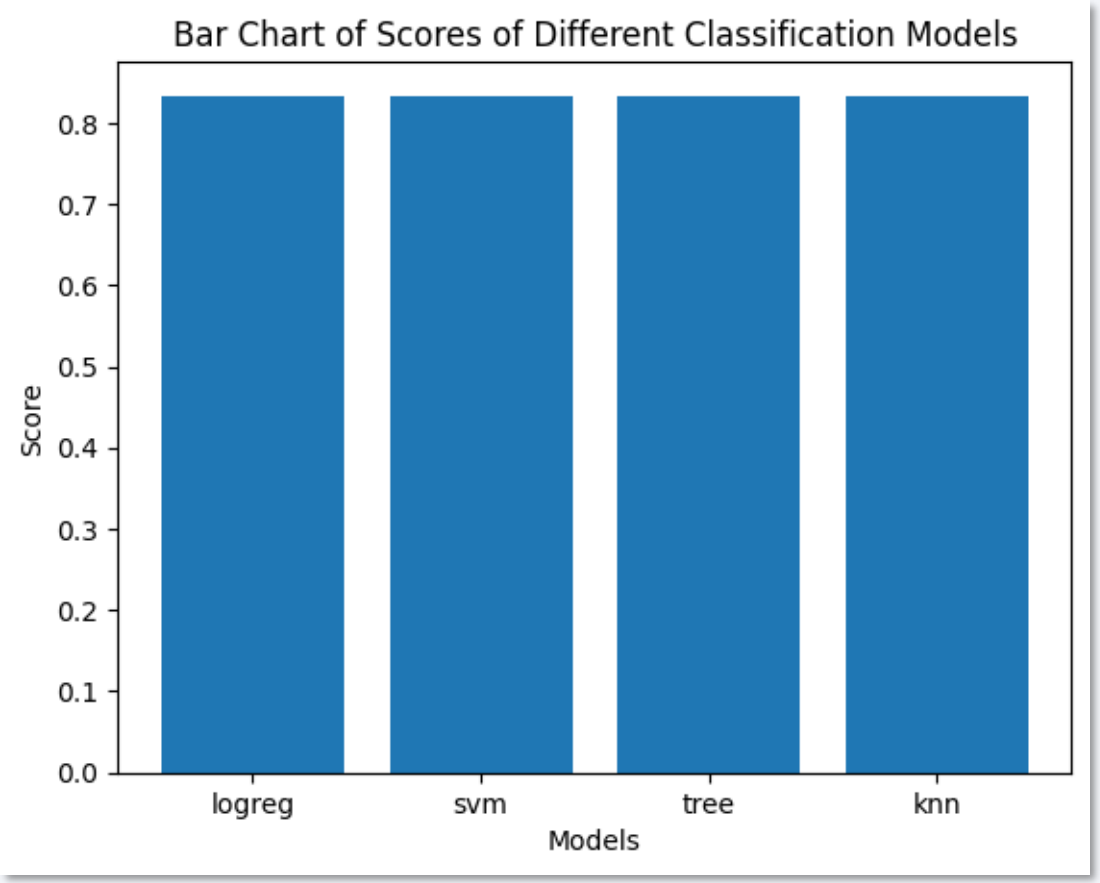



Launches with a payload mass range between 0 to 6000kg has the highest success rate.

Section 5

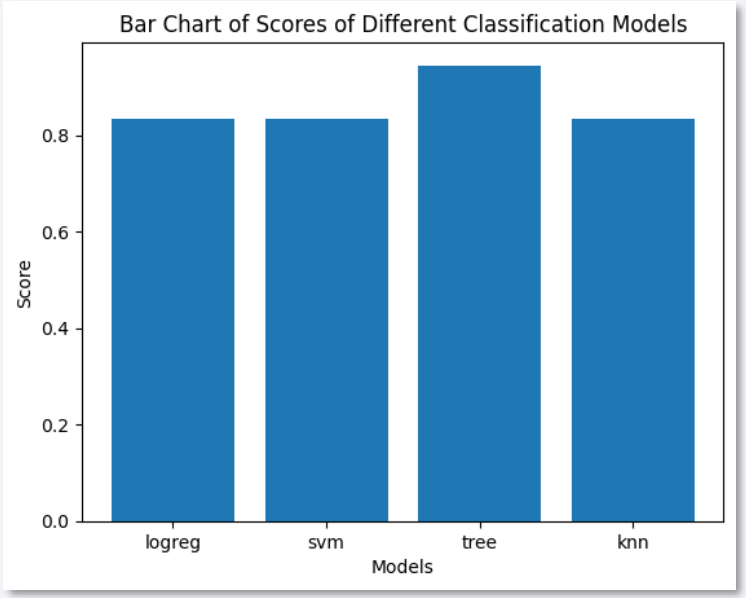
# Predictive Analysis (Classification)

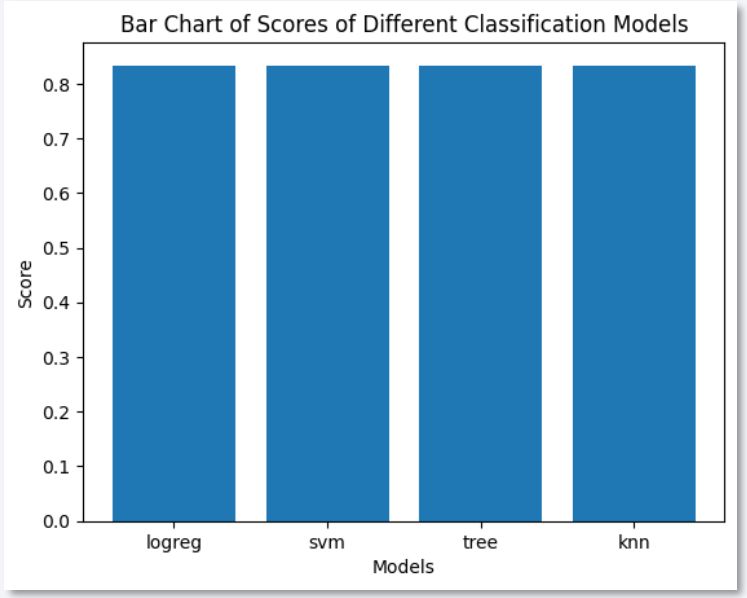
# Classification Accuracy

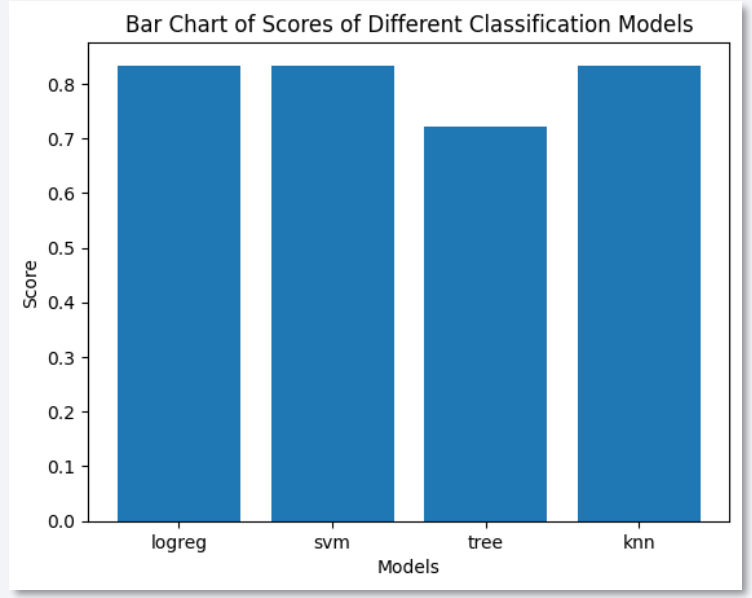



Logistic Regression, SVM, and K-Nearest Neighbors have equal accuracy of around 83%.

# Classification Accuracy



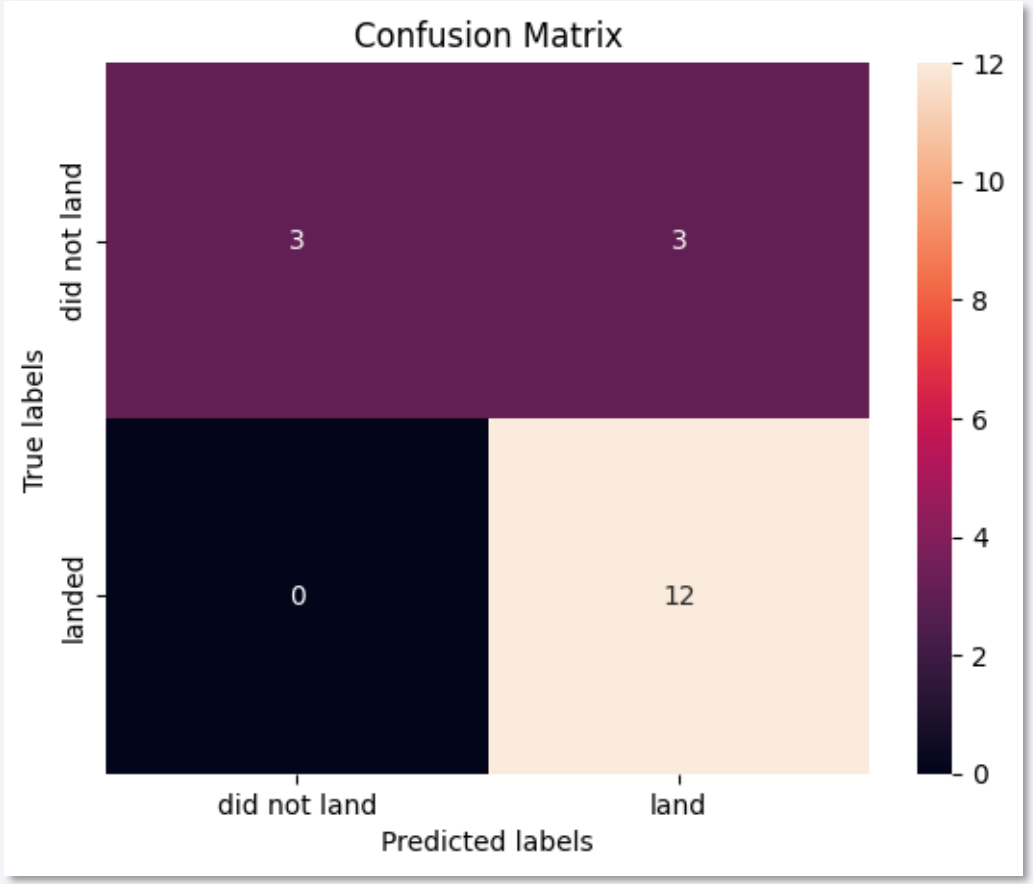




But the Decision Tree results in different accuracy almost every time, which could be 72% to 94%.

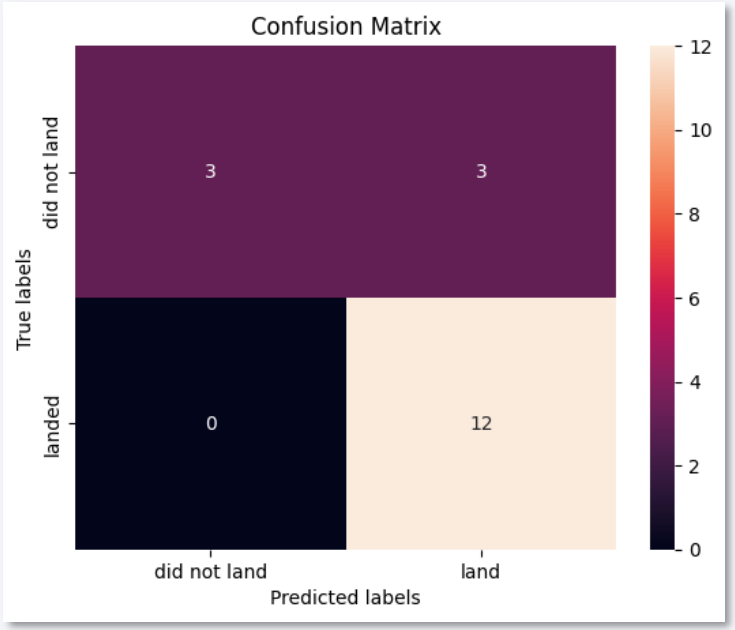
# Confusion Matrix

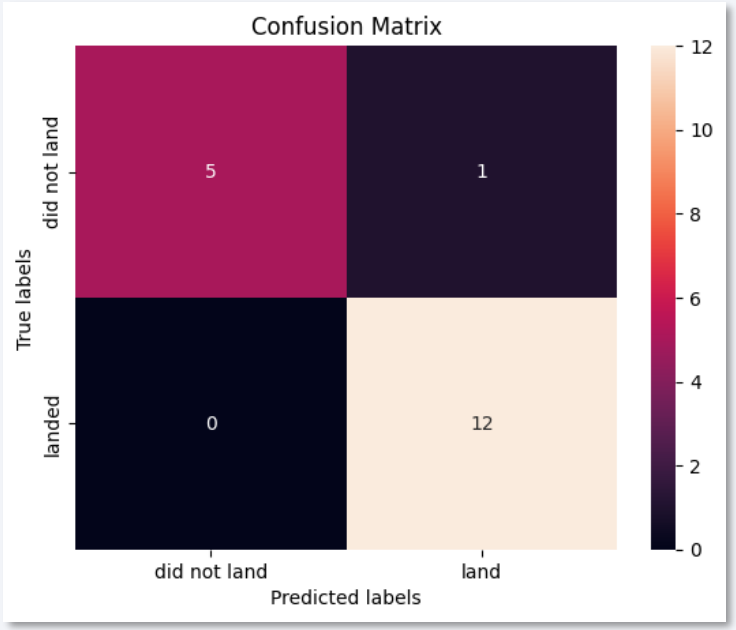
Logistic Regression, SVM, and K-Nearest Neighbors result in the same confusion matrix.

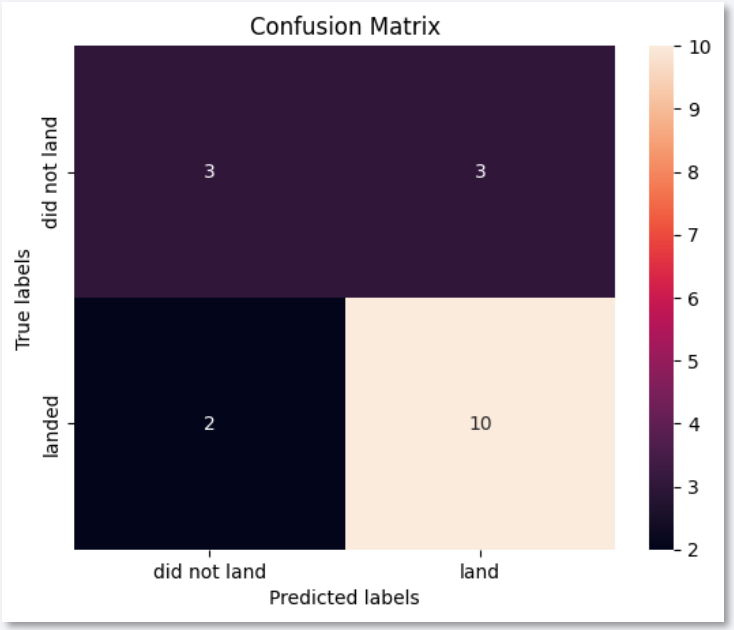
These 3 models did a great job in predicting landed launches but were not so good at not landing launches prediction.



# Confusion Matrix




Accordingly, the Decision Tree results in a different confusion matrix every time, we can tell it performs quite unstable.

# Conclusions

- The launch outcomes could depend on launch sites, time, payload mass, and orbit type.
- Booster version F9 v1.1 has an average payload mass of 2928.4kg.
- SpaceX launched rockets for NASA with a total payload mass of 45596kg.
- SpaceX did a great job at landing outcomes control.
- All launch sites are based near coastlines, have convenient transportation, with railways and roads nearby, and keep a certain distance from residential areas.
- KSC LC-39A has the highest number of successful launches and the highest launch success ratio.
- Logistic Regression, SVM, and K-Nearest Neighbors have equal accuracy around 83%, on the other hand, the Decision Tree performs quite unstable.

# Appendix

References:

SpaceX – Falcon 9

SpaceX API Docs

List of Falcon 9 and Falcon Heavy launches – Wikipedia

Vandenberg Space Force Base – Wikipedia

Kennedy Space Center – Wikipedia

Cape Canaveral Space Force Station - Wikipedia

# Thank you!