IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Louis Wellmeyer
Aug 17th, 2025

# Outline

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

# Executive Summary

Summary of methodologies

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- Interactive Visual Analytics
- Predictive Analysis

Summary of all results

- Exploratory Data Analysis (EDA)
- Geospatial analytics
- Interactive dashboard
- Predictive analysis of classification models

# Introduction

- SpaceX lists the cost of Falcon 9 rocket launches on its website at 62 million dollars, while other providers charge over 165 million dollars per launch. Much of this cost difference comes from SpaceX's ability to reuse the first stage. Thus, knowing whether the first stage will land successfully allows us to estimate the launch cost. This insight can be valuable for competitors looking to bid against SpaceX for a rocket launch.
- Most unsuccessful landings are intentional. Occasionally, SpaceX opts for a controlled landing in the ocean.
- The primary question we aim to address is, given various features of a Falcon 9 rocket launch - such as payload mass, orbit type, launch site, and other – can we identify the right circumstances for a successful launch and predict ongoing launch outcomes.

Section 1

# Methodology

# Methodology

## Executive Summary

- Data collection methodology
- Perform 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

# Data Collection

There are various sources for useful data and distinct methods for collection, including but not limited to:

1. APIs
2. WebScraping

# Data Collection – SpaceX API

- The official SpaceX API provides information of past rocket launches including many interesting metadata

- Refer for completed SpaceX API calls notebook https://github.com/Wellbek/IBM-Data-Science/blob/main/10%20Data%20Science%20Captstone/01/jupyter-labs-spacex-data-collection-api.ipynb

# Data Collection - Scraping

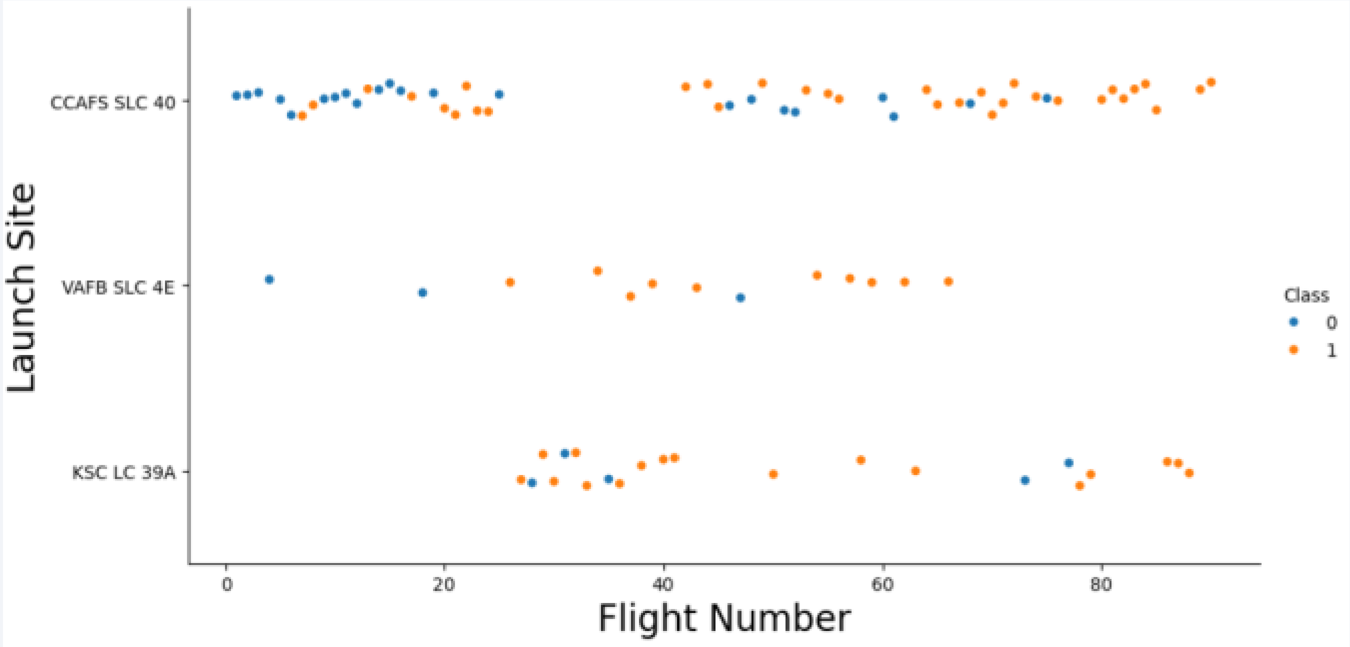
- Web Scraping is the technique of reading DOM-Tree data directly – therefore data can be gathered even when no specific interface is available.
- Refer for completed web scraping notebook https://github.com/Wellbek/IBM-Data-Science/blob/main/10%20Data%20Science%20Captstone/01/jupyter-labs-webscraping.ipynb

# Data Wrangling

- After collection, data is available in **raw format**
- Data wrangling is the process of
    1. Cleaning and transforming data to ensure quality
    2. Combining Data sources – e.g. API + Scraping
    3. Preparing Data for Analysis – e.g. via Pandas dataframes

# EDA with Data Visualization

- To conduct first analysis, scatter plots, bar charts, line charts were used to visibly identify trends, links of variables, and identify outliers.



# EDA with SQL

Following SQL queries were performed to gather information about the transformed rocket dataset:

- Names of unique launch sites
- Total payload mass carried by boosters (CRS)
- Average payload mass by booster version F9 v1.1
- Total number of successes and failures
- Failed landing outcomes in drone ship and linked metadata

# Build an Interactive Map with Folium

- All launch sites were marked, and map elements like markers, circles, and lines were added to the folium map to indicate the success or failure of launches at each location.
- Launch outcomes were categorized into classes, with failures assigned to 0 and successes to 1.
- Launch sites with high success rates were highlighted using color-coded marker clusters.
- Distances from each launch site to nearby points were calculated to determine optimal geopolitical positions.

# Build a Dashboard with Plotly Dash

We developed an interactive dashboard using Plotly and Dash

1. Pie charts displaying the total launches for each site
    - Allows filtering by specific sites
2. Scatter plot illustrating the correlation between launch outcomes and payload mass across various booster versions
    - Includes a rangeslider to filter by payload mass ranges
    - Filterable by booster version

# Predictive Analysis (Classification)

1. The data was loaded with numpy and pandas, then transformed and divided into training and testing sets.
2. Various machine learning models were developed, and hyperparameters were optimized using GridSearchCV.
3. Accuracy was selected as the evaluation metric, and the model was enhanced through feature engineering and algorithm tuning.
4. The top-performing classification model was identified.

# Results

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

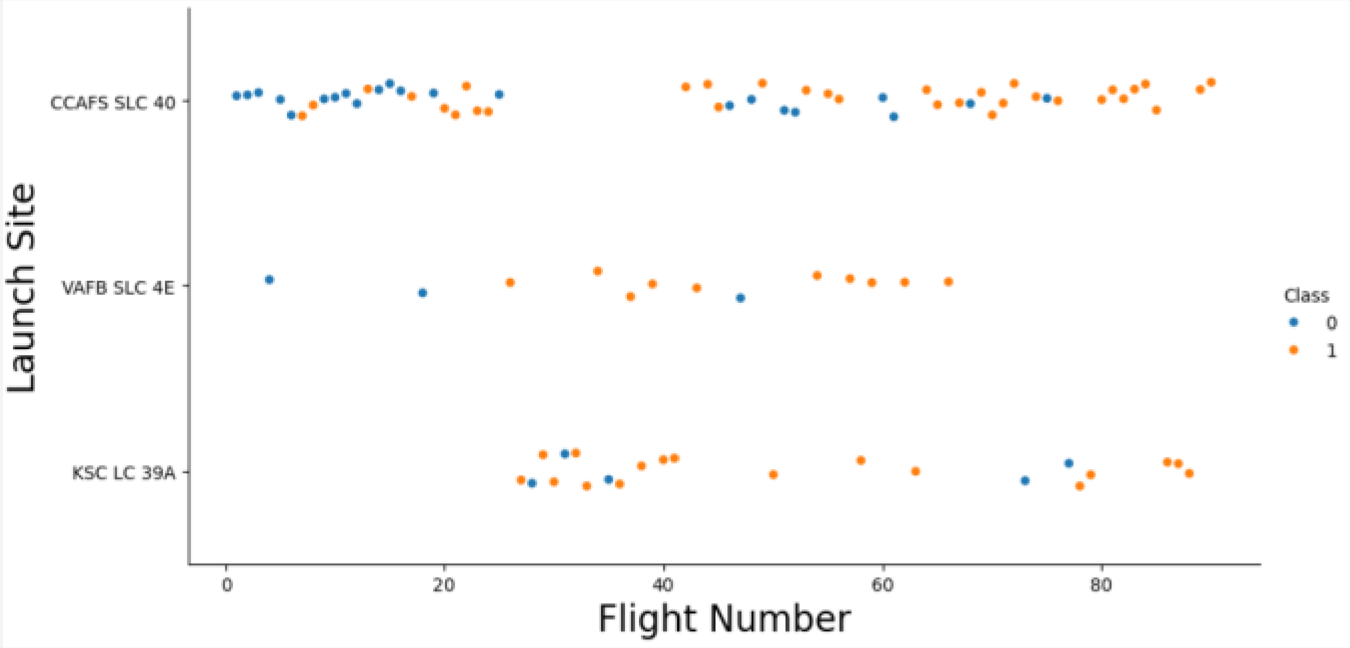
Section 2

# Insights drawn from EDA

# Flight Number vs. Launch Site

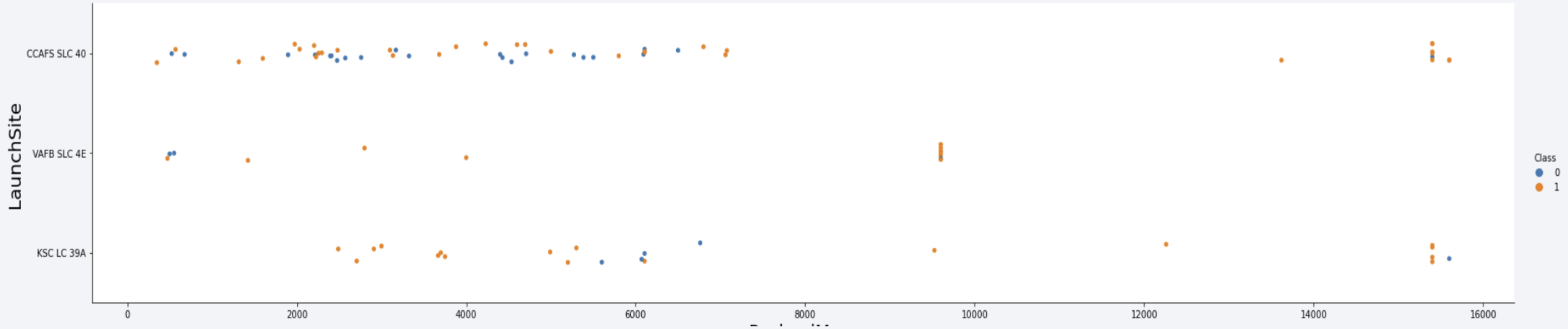
The scatter plot of Launch Site vs. Flight Number indicates that success rates improve with more flights.

- Early flights (under 30) from CCAFS SLC 40 and VAFB SLC 4E were mostly unsuccessful
- KSC LC 39A had no early flights and shows higher success.
- After flight number 30, successful landings increase notably



# Payload vs. Launch Site

- Payloads exceeding 9,000 kg generally achieve a high success rate;
- Payloads above 12,000 kg appear to be feasible only at the CCAFS SLC 40 and KSC LC 39A launch facilities.



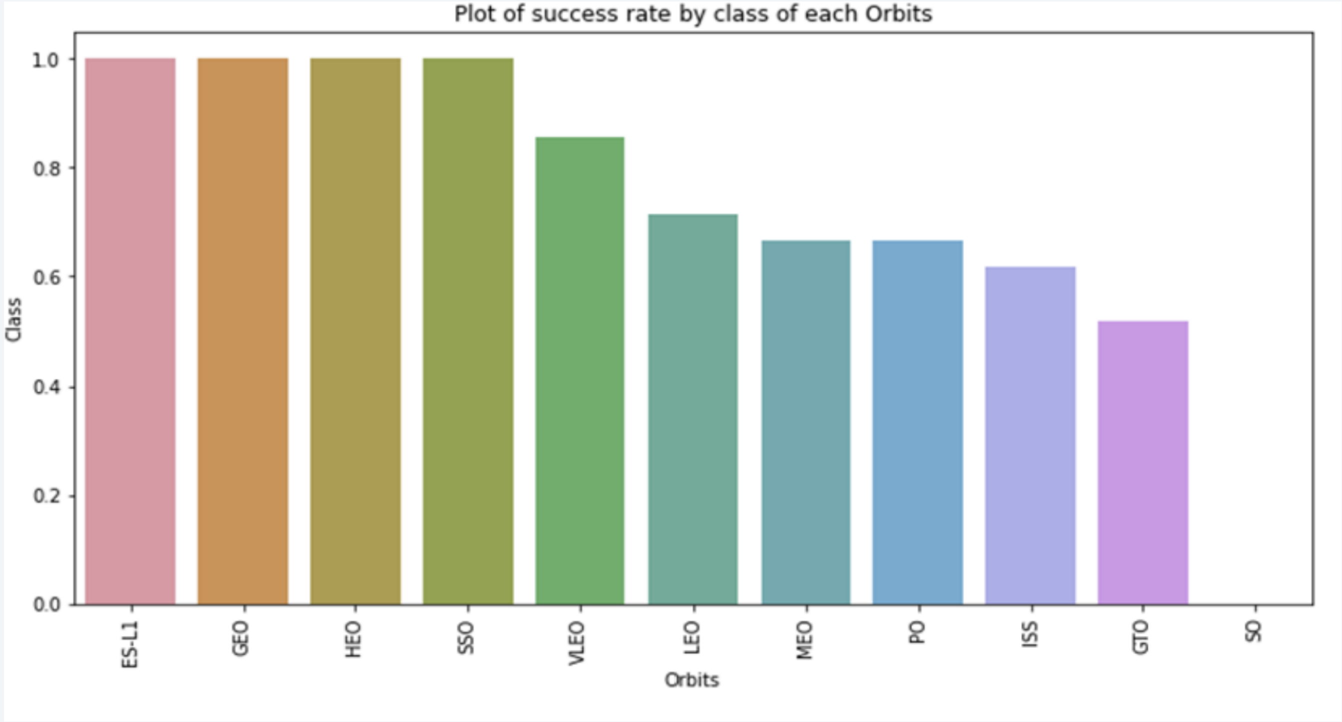
# Success Rate vs. Orbit Type

following orbits have the highest (100%) success rate:

- ES-L1
- GEO
- HEO
- SSO

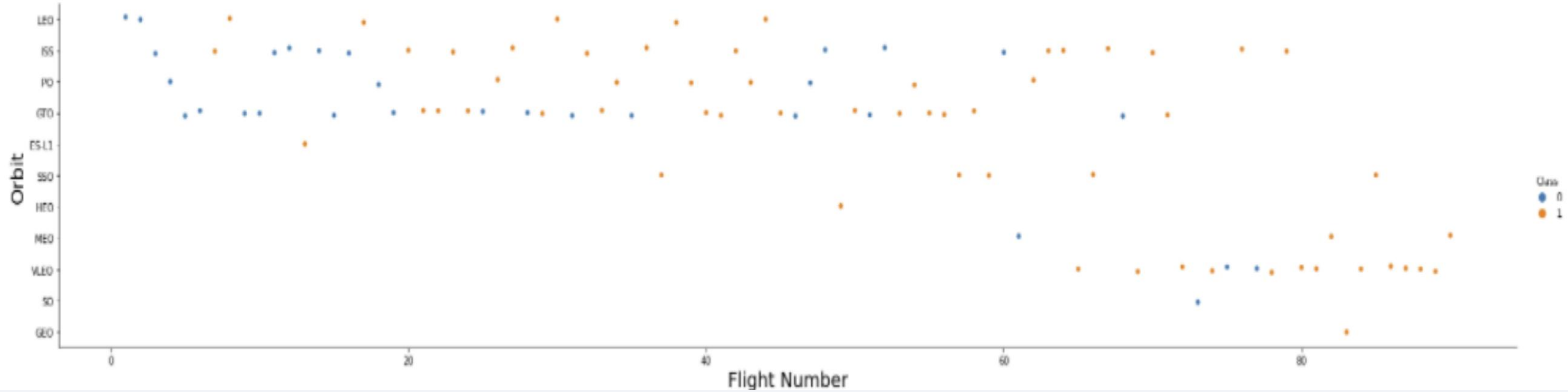
The orbit with the lowest (0%) success rate is:

- SO



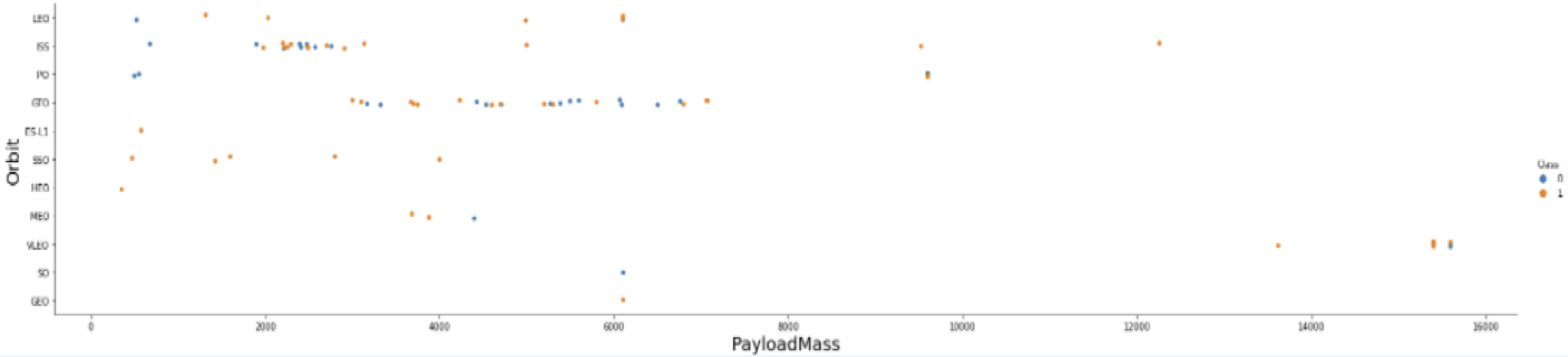
# Flight Number vs. Orbit Type

- Success in LEO correlates with flight number
- in GTO, no such relationship exists.



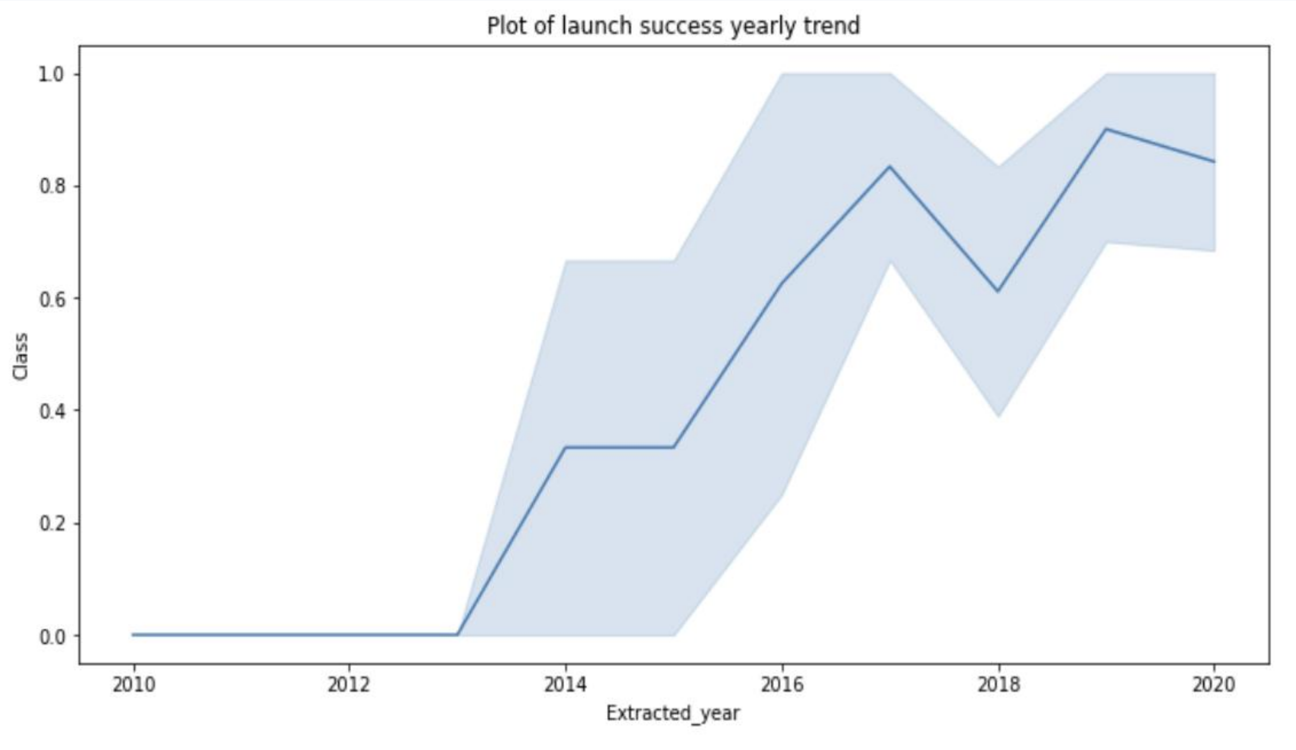
# Payload vs. Orbit Type

- Successful landings with heavy payloads occur more frequently in PO, LEO, and ISS orbits.



# Launch Success Yearly Trend

- Between 2010 and 2013, there were no successful landings, resulting in a 0% success rate.
- After 2013, success rates increased overall, with slight declines observed in 2018 and 2020.
- From 2016 onward, the probability of success has remained above 50%.



# All Launch Site Names

- Find the names of the unique launch sites
- The word DISTINCT returns only unique values from the LaunchSite column of the SPACEX table

Display the names of the unique launch sites in the space mission

In [10]:
```
task_1 = '''
        SELECT DISTINCT LaunchSite
        FROM SpaceX
'''
create_pandas_df(task_1, database=conn)
```

Out[10]:

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

# Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- With LIMIT 5 only 5 records are fetched
- The LIKE keyword was used with the wild card ‘CCA%’ to retrieve string values beginning with ‘CCA’.

| | date | time | boosterversion | launchsite | payload | payloadmasskg | orbit | customer | missionoutcome | landingoutcome |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2010-04-06 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 1 | 2010-08-12 | 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 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 3 | 2012-08-10 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 4 | 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

- Calculate the total payload from NASA boosters
- The total payload from NASA boosters is 45,596
- This was found by summing the LAUNCH column for NASA boosters (CRS) only

# Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass from booster version F9 v1.1 is 2,928.4
- This was found by using the AVG keyword on the PAYLOAD_MASS_KG_ column for F9 v1.1 boosters only

# First Successful Ground Landing Date

- Identify the date of the first successful ground pad landing.
- The initial successful ground pad landing occurred on December 22, 2015.
- The MIN keyword was used on the DATE column, while the WHERE clause filtered for successful ground pad landings only.

## Successful Drone Ship Landing with Payload between 4000 and 6000

- Provide the names of boosters that have successfully landed on a drone ship and carried a payload mass greater than 4000 but less than 6000.
- The WHERE clause filters results to include only those meeting both conditions within the parentheses (with the AND keyword applied). The BETWEEN keyword selects values where 4000 < x < 6000.

| | boosterversion |
|---|---|
| 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

- Determine the overall count of both successful and failed mission outcomes.
- The COUNT keyword was used to tally the total mission outcomes, while a GROUPBY was applied to categorize these results based on the outcome type

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

# Boosters Carried Maximum Payload

- Identify the names of the boosters that have transported the heaviest payload mass.
- A subquery was used to determine the maximum payload, which is then applied in the original WHERE of the master query.
- DISTINCT keyword ensures that only unique booster versions are returned.

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

# 2015 Launch Records

- Identify the failed landing outcomes on drone ships, including their booster versions and launch site names, for the year 2015.

- A combination of WHERE clauses along with LIKE, AND, and BETWEEN conditions was used to filter failed landing outcomes on drone ships, their respective booster versions, and launch site names specifically for 2015.

| | boosterversion | launchsite | landingoutcome |
|---|---|---|---|
| 0 | F9 v1.1 B1012 | CCAFS LC-40 | Failure (drone ship) |
| 1 | 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

- A WHERE clause combined with BETWEEN filters the data to included only dates within the specified range. The results were then grouped using GROUP BY and sorted with ORDER BY, with DESC indicating descending order.

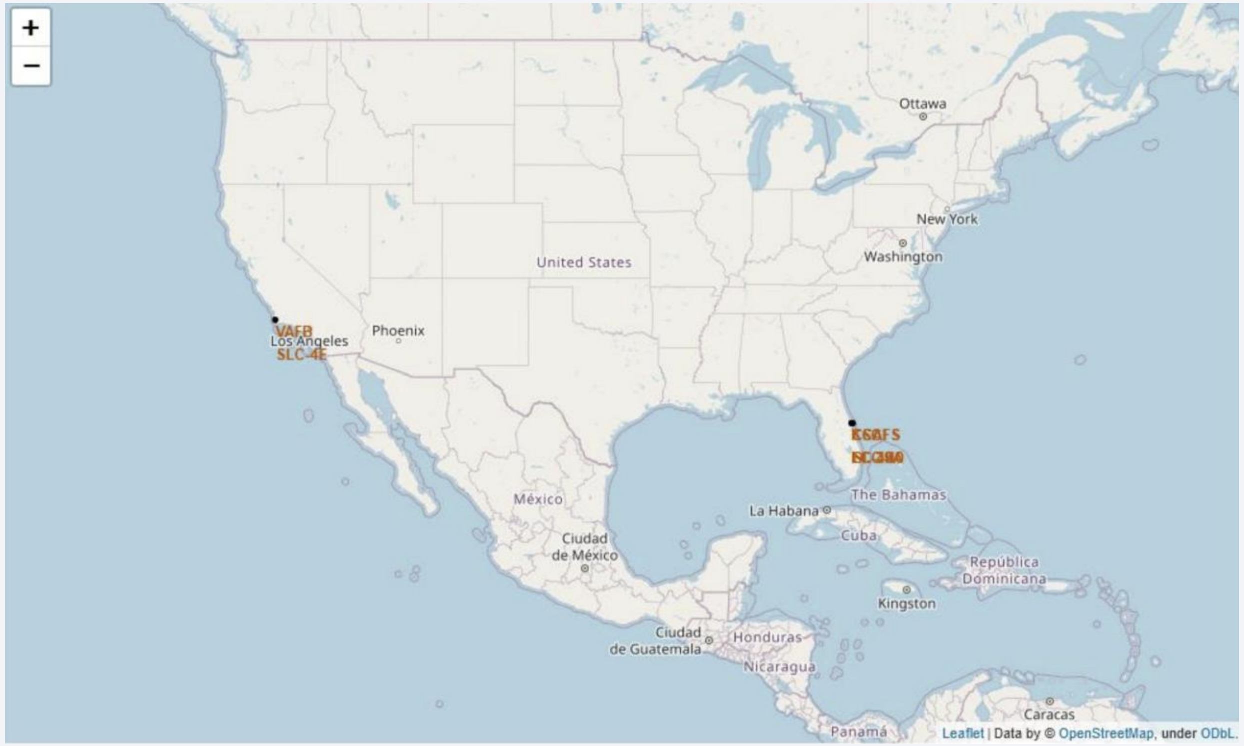
| | landingoutcome | count |
|---|---|---|
| 0 | No attempt | 10 |
| 1 | Success (drone ship) | 6 |
| 2 | Failure (drone ship) | 5 |
| 3 | Success (ground pad) | 5 |
| 4 | Controlled (ocean) | 3 |
| 5 | Uncontrolled (ocean) | 2 |
| 6 | Precluded (drone ship) | 1 |
| 7 | Failure (parachute) | 1 |

Section 3

# Launch Sites Proximities Analysis

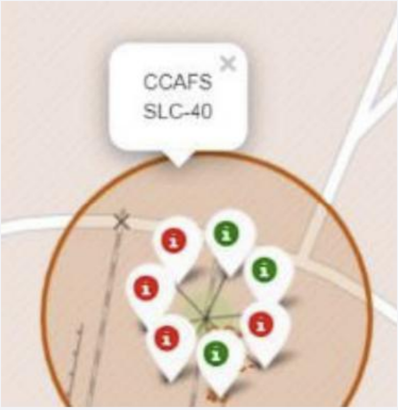
# Launch Site Map

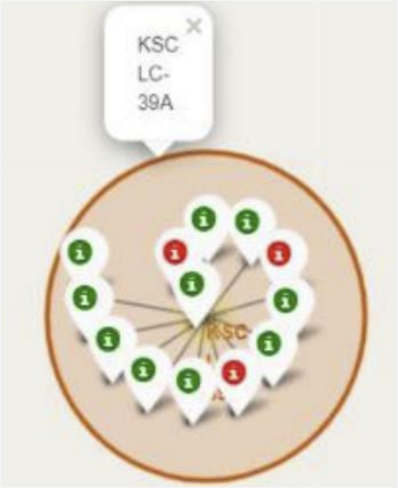
- All SpaceX launch sites are on the coasts of USA
  - Florida
  - California



# Launch Site Successes & Failures

- launches were grouped into clusters to display successful and failed landings at launch sites
- Green representing success
- Red representing failure




# Launch Site Proximities

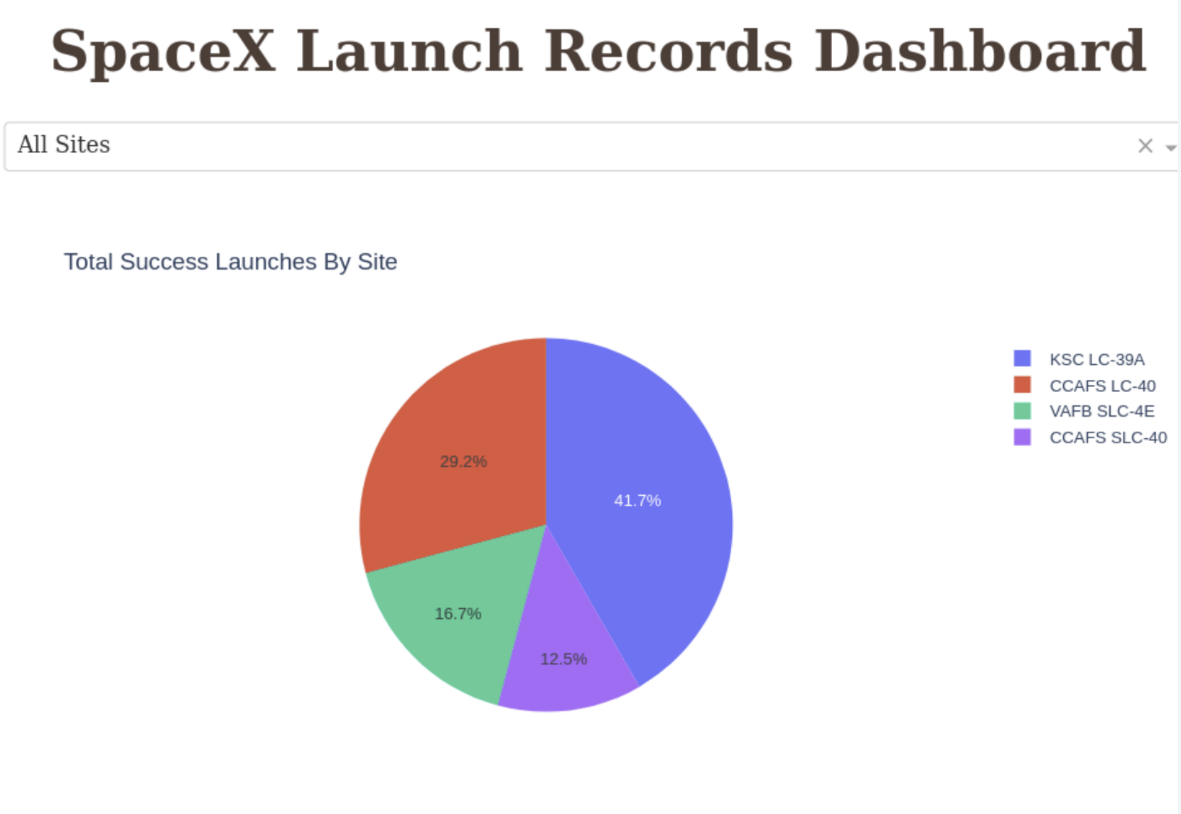
- By examining the CCAFS SLC-40 launch site as a case study, we can gain a better understanding of how launch sites are positioned

- 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 coast lines? Yes
- Do launch sites keep certain distance away from cities? Yes

Section 4

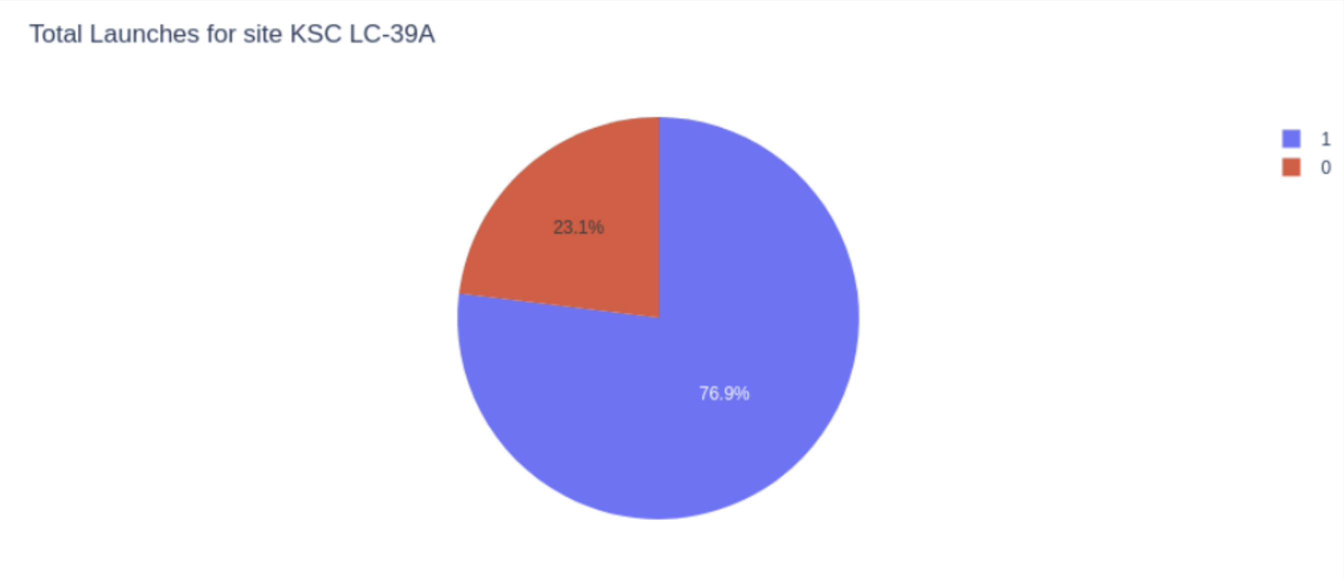
# Build a Dashboard with Plotly Dash

# Successful Launches by Site



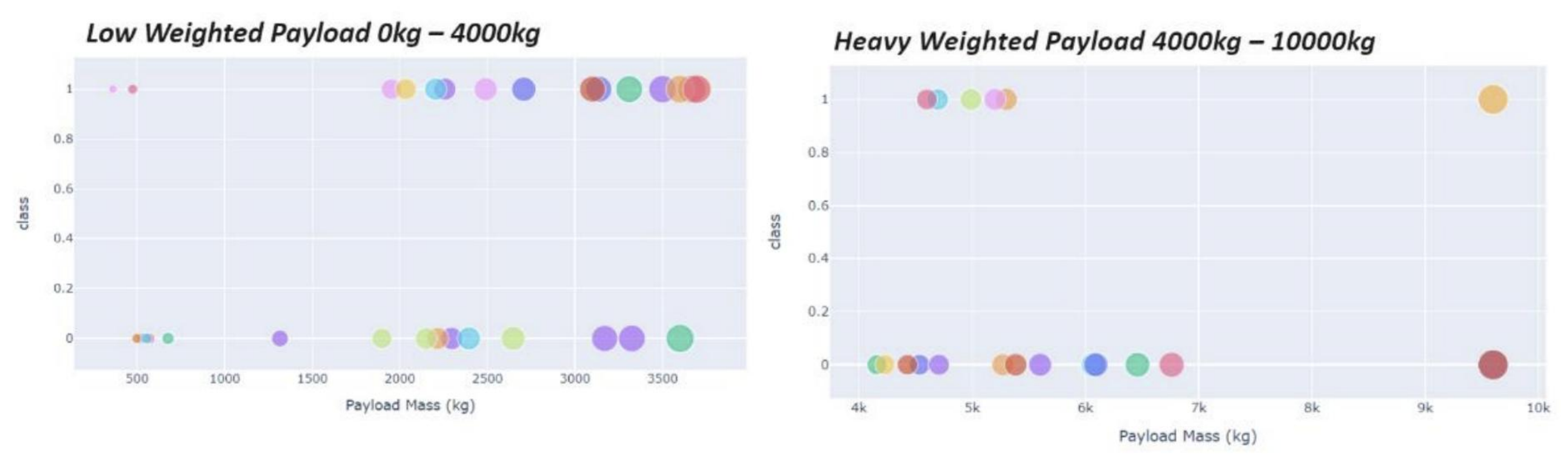

- KSC LC-39A had the most successful launches from all sites

# KSC LC-39A Success Rates




- 76.9% of launches are successful in KSC LC-39A

# Payload vs. Launch Outcome




- Success rates for low weighted payloads are higher than for heavy ones

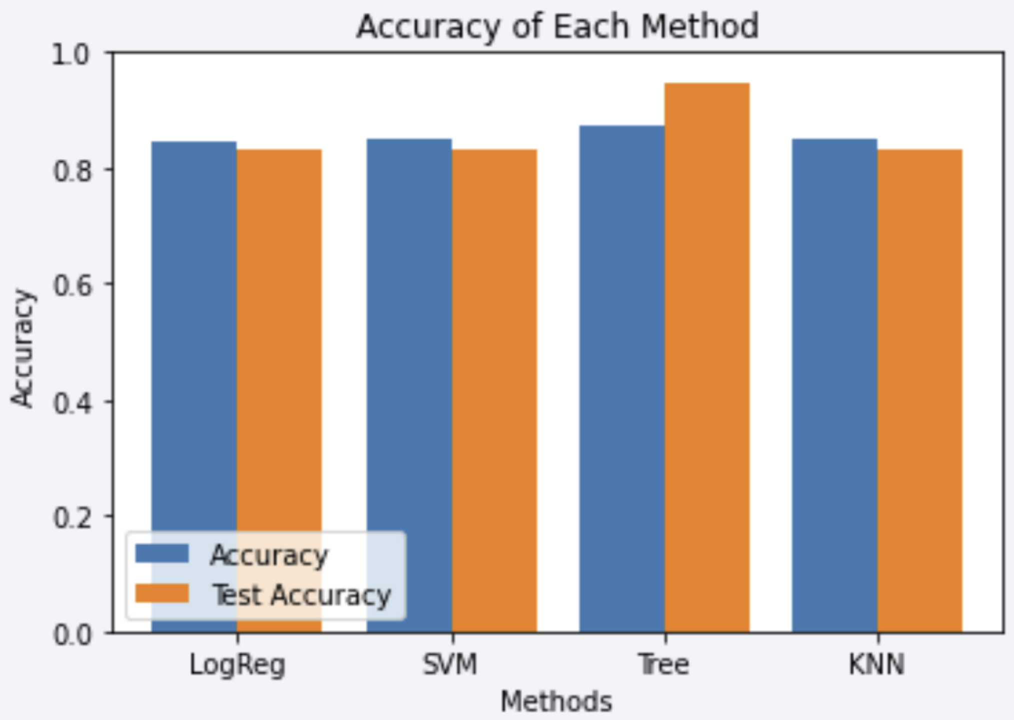
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- Four different classification models were evaluated, and their accuracy results were displayed in a plot

- Among them, the Decision Tree Classifier achieved the highest accuracy, surpassing 87%



# Confusion Matrix

- CM of Decision Tree Classifier demonstrates its accuracy by displaying a significantly higher count of true positives and true negatives compared to false positives and false negatives

# Conclusions

- Higher flight volumes at a launch site are associated with increased success rates.
- Launch success rates showed an upward trend from 2013 to 2020.
- Orbits such as ES-L1, GEO, HEO, SSO, and VLEO demonstrated the highest success rates.
- KSC LC-39A recorded the greatest number of successful launches among all sites.
- The Decision Tree classifier proved to be the most effective machine learning algorithm for this task.

# Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

# Thank you!