IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Shaghayegh Haghbin
21/02/2024

Outline

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

Executive Summary

- Summary of methodologies:
- Week 1: Data Collection API, Data Collection with Web Scraping, Data Wrangling.
- Week 2: Exploratory Data Analysis with SQL, Exploratory Data Analysis with Visualization.
- Week 3: Interactive Visual Analytics with Folium, Building an Interactive Dashboard with Ploty Dash.
- Week 4: Machine Learning Prediction.
- Week 5: final report for peer review

- Summary of all results:
- Obtained critical insights from publicly accessible sources.
- Exploratory Data Analysis enabled the identification of the key factors for predicting launching success.
- Machine Learning Prediction demonstrated an effective model for predicting essential factors, utilizing all gathered data.

Introduction

- Project background and context:
- Utilizing predictive analysis, this project focuses on predicting if the beginning phase of Falcon 9 will have a successful landing or not.
- SpaceX promotes Falcon 9 rocket launch services on its official website, priced at 62 million dollars. In contrast, other providers typically command prices exceeding 165 million dollars per launch. A significant factor contributing to these cost disparities is SpaceX's innovative capability to reuse the first stage of the rocket.
- Problems you want to find answers:
- Factors influencing successful rocket landings.
- Interaction of key features impacting landing success.
- Operating conditions essential for a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The source for SpaceX API is `https://api.spacexdata.com/v4/rockets/`
  - The source for web scraping is https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform data wrangling:
  - Enhanced data collection by developing a landing outcome label through comprehensive analysis and key features summarization.
- Perform exploratory data analysis (EDA) using visualization and SQL.

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash.
- Perform predictive analysis using classification models:
- The gathered data underwent normalization and was partitioned into training and test datasets.
- In the evaluation phase, there was an employment of four distinct classification models.
- Model accuracy was assessed across various parameter combinations.

Data Collection

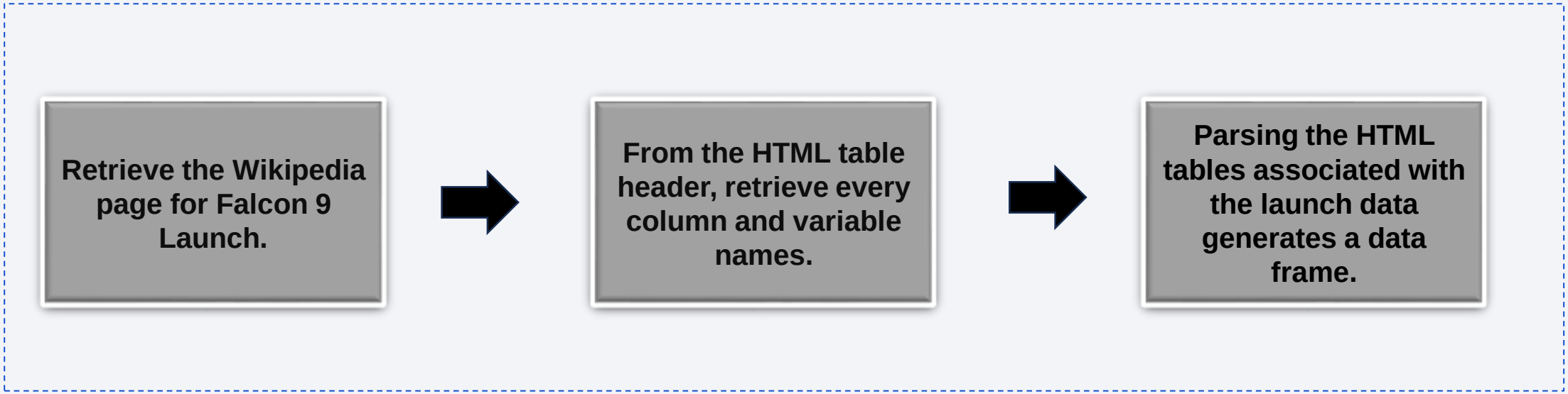
- Describe how data sets were collected.
- The source for SpaceX API dataset collection is: https://api.spacexdata.com/v4/rockets/
- The source for web scraping is: https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Data acquisition was facilitated utilizing SpaceX's public API, following the outlined flowchart, and persisted for further analysis.
- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%201/Data%20Collection%20API%20Lab.ipynb

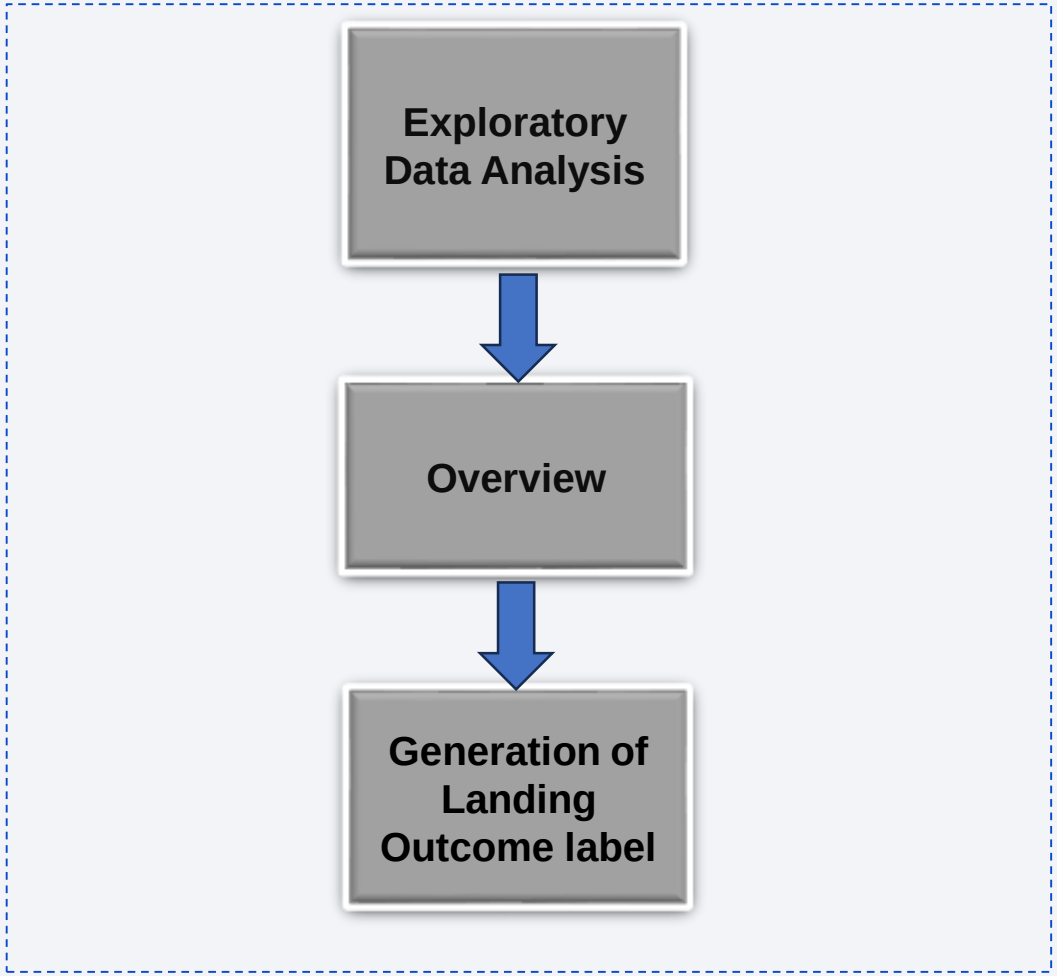
Data Collection - Scraping

- Data sourced from SpaceX launches, including accessible information via Wikipedia are acquired following the outlined flowchart and persisted for analysis.
- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%201/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling

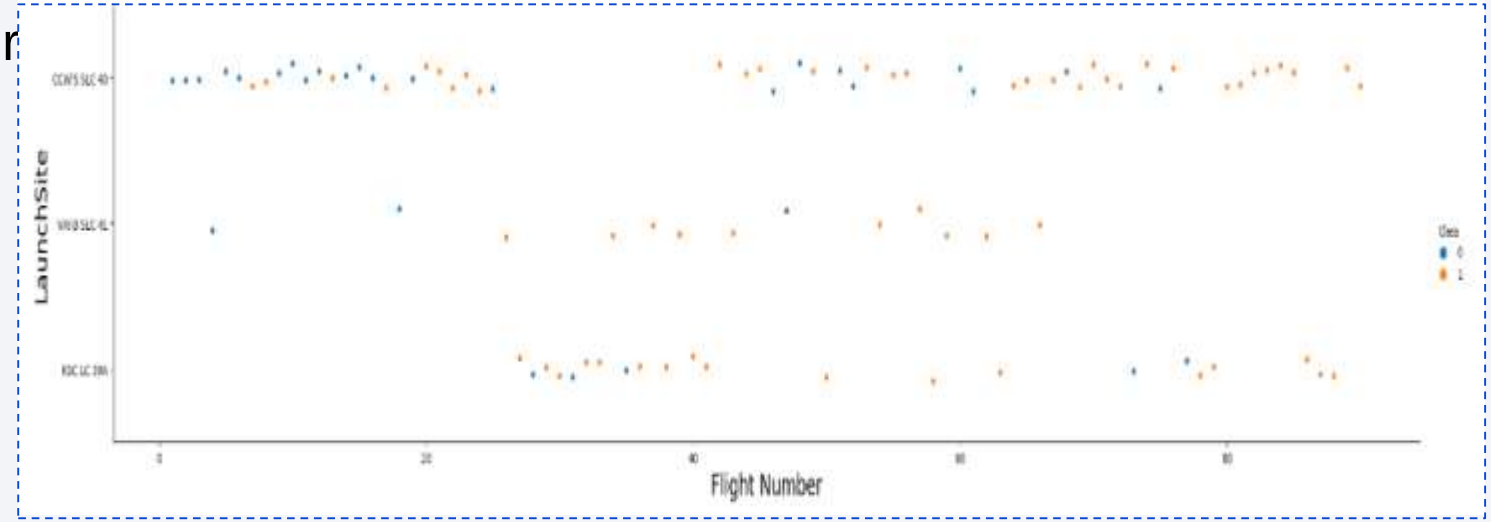
- The dataset underwent initial Exploratory Data Analysis. Following that, analyses were conducted to determine the number of launches per site, instances of each orbit and results categorized by orbit type. Lastly, the landing outcome label was generated based on the data within the Outcome column.
- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%201/Data%20wrangling.ipynb



EDA with Data Visualization

- During the Exploratory Data Analysis with Data Visualization phase, plots like scatterplots and bar plots were employed to investigate data and illustrate the correlation between the sets of variables:

1. Payload Mass X Flight Number
2. Launch Site X Flight Number
3. Launch Site X Payload Mass
4. Orbit and Flight Number
5. Payload and Orbit



- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%202/EDA%20with%20Visualization.ipynb

EDA with SQL

- Task 1: Display the names of the unique launch sites in the space mission
- Task 2: Display 5 records where launch sites begin with the string 'CCA’
- Task 3: Display the total payload mass carried by boosters launched by NASA (CRS)
- Task 4: Display average payload mass carried by booster version F9 v1.1
- Task 5: List the date when the first successful landing outcome in ground pad was achieved.
- Task 6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Task 7: List the total number of successful and failure mission outcomes
- Task 8: List the names of the booster_versions which have carried the maximum payload mass.

EDA with SQL

- Task 9: List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.
- Task 10: 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.

GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%202/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium Maps employed markers, circles, lines, and marker clusters in its design and implementation.
- Markers denote locations such as launch sites.
- Circles designate highlighted zones surrounding coordinates, such as the NASA Johnson Space Center.
- Marker clusters signify collections of occurrences at each coordinate, like launches in a launch site.
- Lines are employed to illustrate the distance between two sets of coordinates.
- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%203/Launch%20Sites%20Locations%20Analysis%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

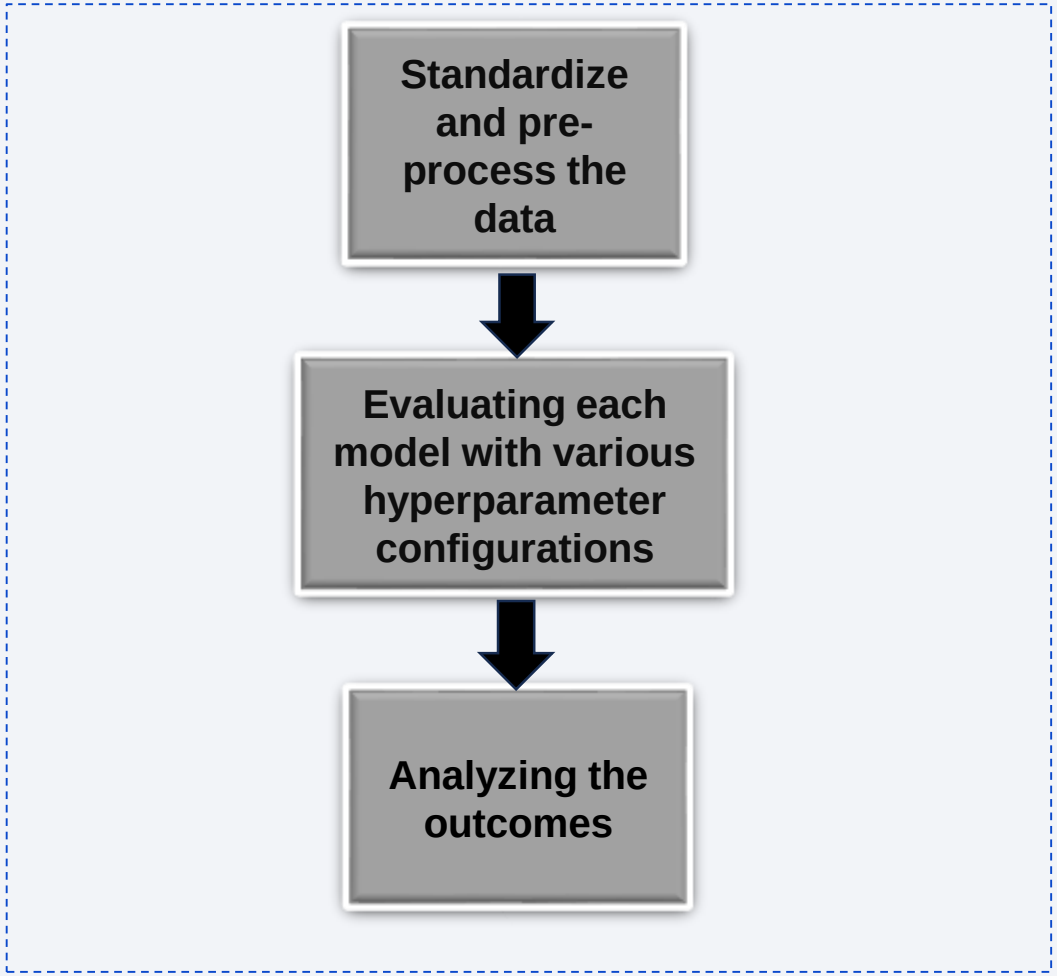
- Graphs and plots were utilized to visually represent data, including the percentage of launches categorized by site and payload range.
- This integration enabled a rapid examination of the correlation between payloads and launch sites, assisting in determining the most suitable launch locations based on payload requirements.
- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%203/spacex_dash_app.py

Predictive Analysis (Classification)

- During the Machine Learning Prediction phase, four types of classification models were evaluated:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. K-Nearest Neighbors

- GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/blob/main/week%204/Machine%20Learning%20Prediction.ipynb

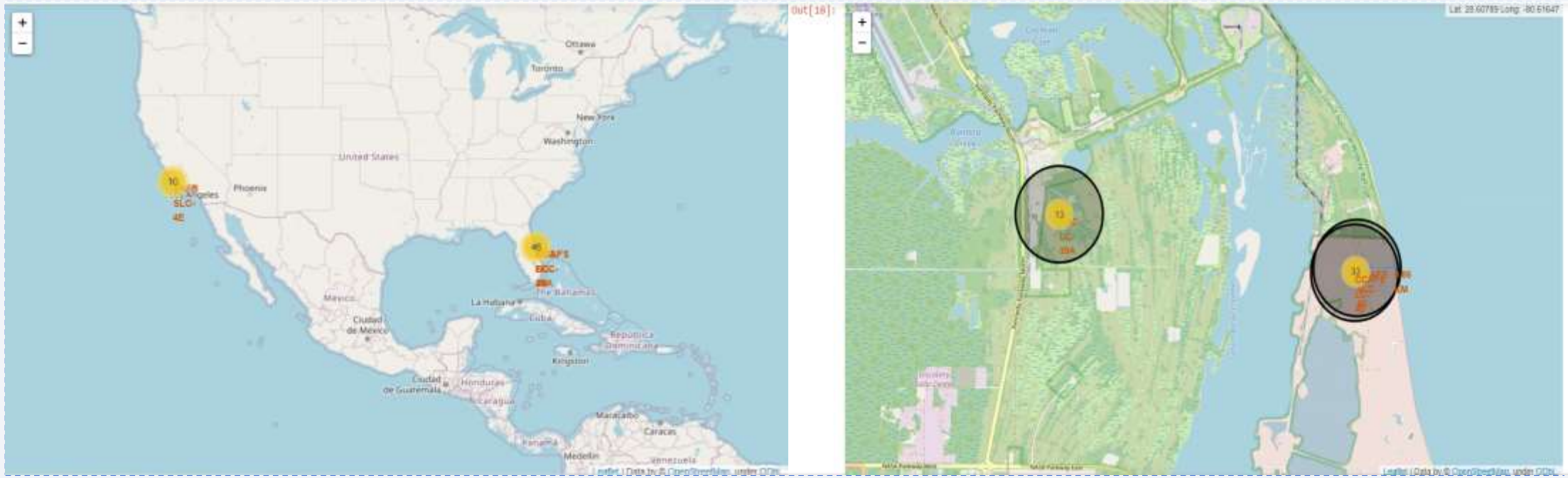


Results

- Exploratory data analysis results:
- SpaceX operates from four distinct launch sites, initially conducting launches both for SpaceX itself and in collaboration with NASA. The average payload for the F9 v1.1 booster stands at 2,928 kg. Notably, the first successful landing occurred in 2015, five years following the inaugural launch. Over time, several Falcon 9 booster iterations achieved successful landings on drone ships, particularly with payloads exceeding the average. Mission success rates approached nearly 100%. However, in 2015, two booster versions, F9 v1.1 B1012 and F9 v1.1 B1015, encountered failed landing attempts on drone ships. Nevertheless, the frequency of successful landings progressively improved with each passing year.

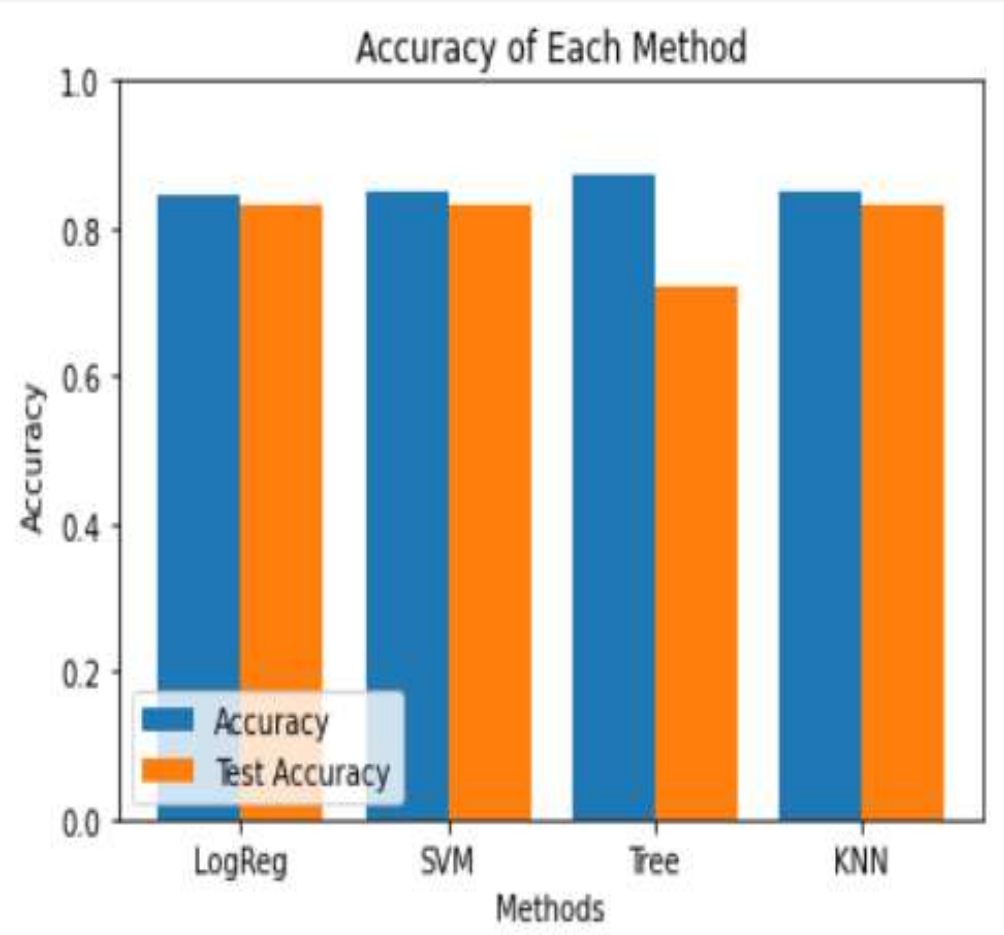
Results

- Leveraging interactive analytics revealed that launch sites tend to be situated in secure locations, often near bodies of water such as the sea, and are surrounded by well-developed logistical infrastructure. Furthermore, maximum launches occur at launch sites situated along the east coast.



Results

- Predictive analysis indicated that the Decision Tree Classifier emerges as the optimal model for forecasting successful landings.
- This model boasts an accuracy exceeding 87% and achieves a test data accuracy surpassing 94%.

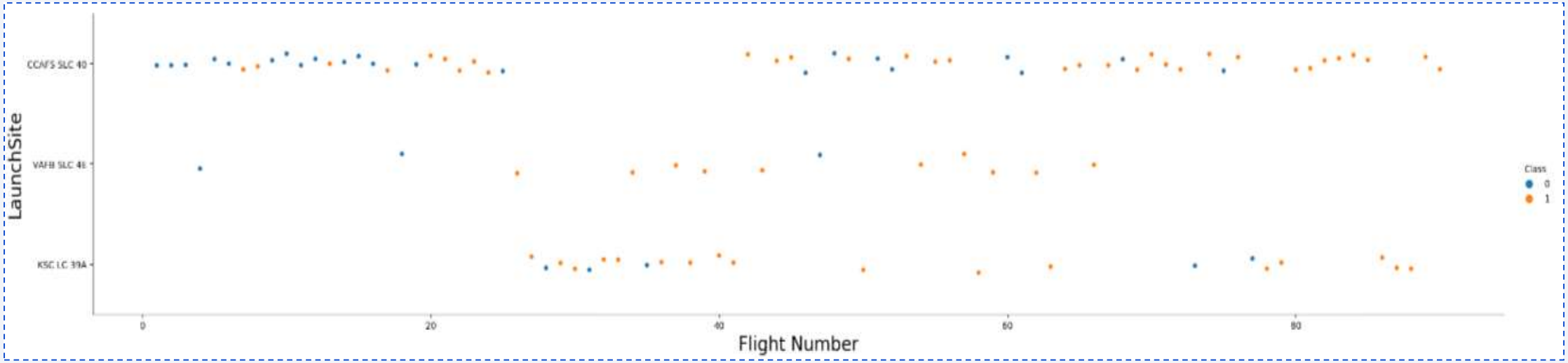


Section 2

Insights drawn from EDA

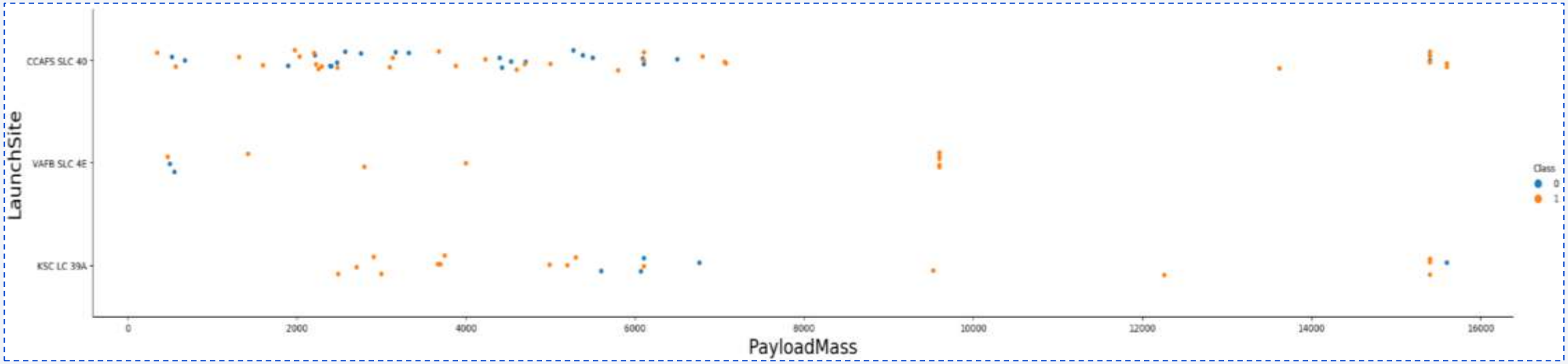
Flight Number vs. Launch Site

- Based on the depicted plot, it is evident that the premier launch site currently is CCAF5 SLC 40, showcasing a predominant success rate in recent launches. Following closely is VAFB SLC 4E, occupying the second position, and KSC LC 39A in third place. Moreover, a notable observation from the plot reveals a progressive enhancement in the overall success rate over time.



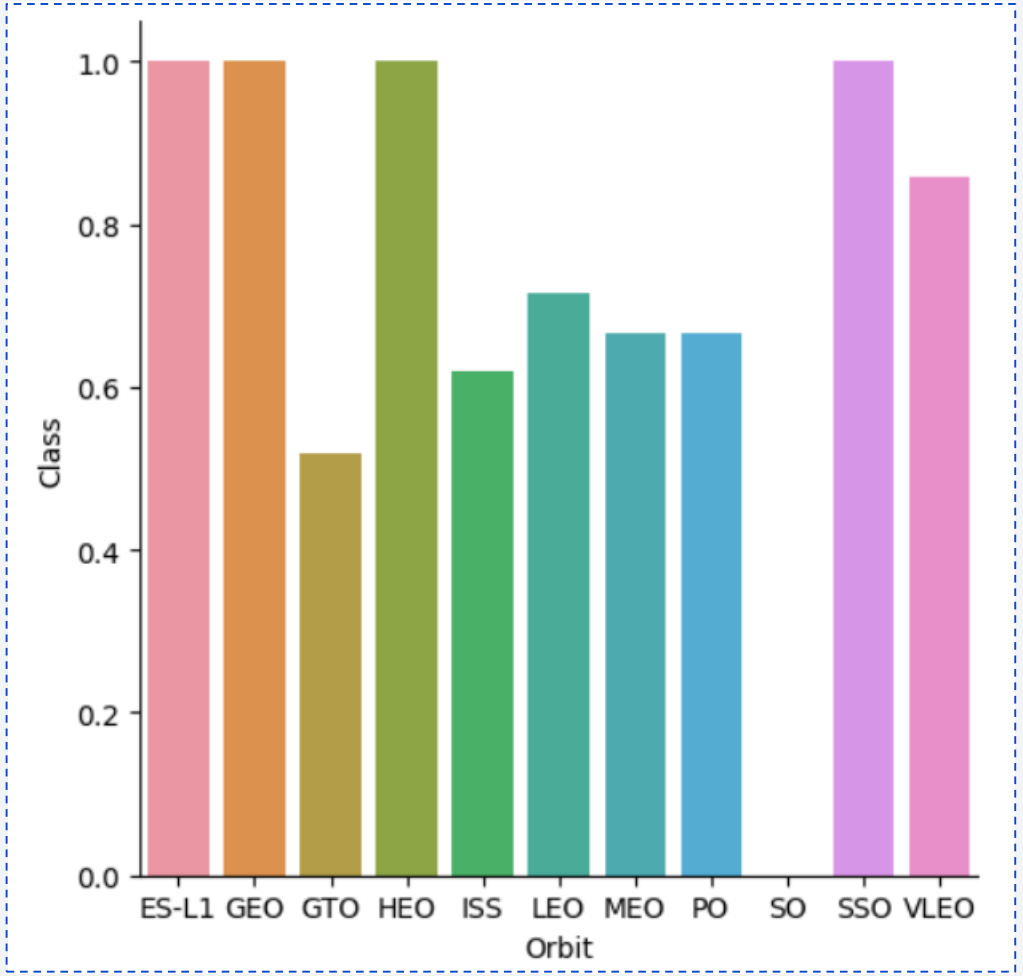
Payload vs. Launch Site

- Payloads exceeding 9,000kg exhibit a notably high success rate, suggesting their robust performance. Furthermore, payloads surpassing 12,000kg appear feasible solely at the CCAFS SLC 40 and KSC LC 39A launch sites, implying their capability to accommodate heavier payloads effectively.



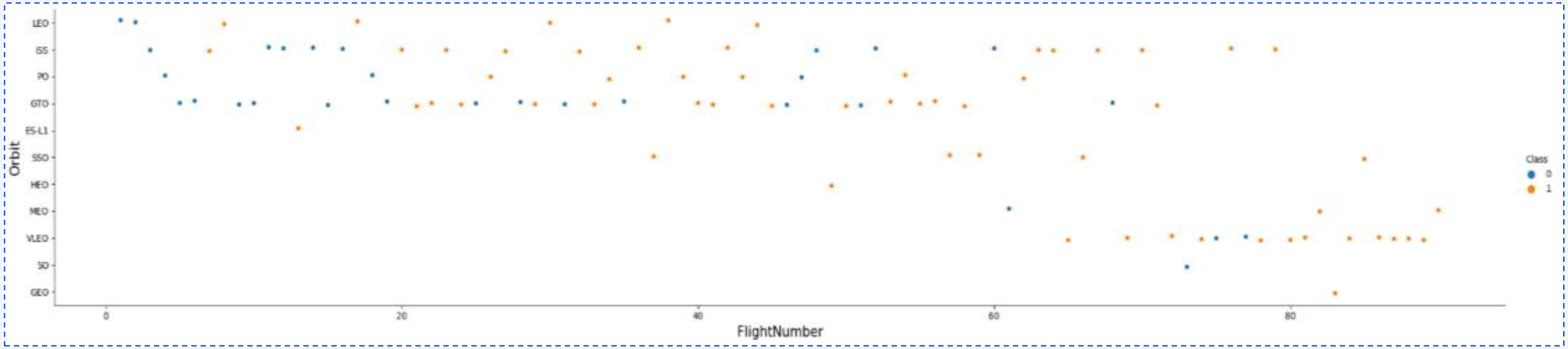
Success Rate vs. Orbit Type

- The highest success rates are observed in specific orbits, including ES-L1, GEO, HEO, and SSO.
- These orbits exhibit the most consistent success rates.
- Following closely are VLEO, with a success rate exceeding 80%, and LFO, boasting a success rate surpassing 70%.



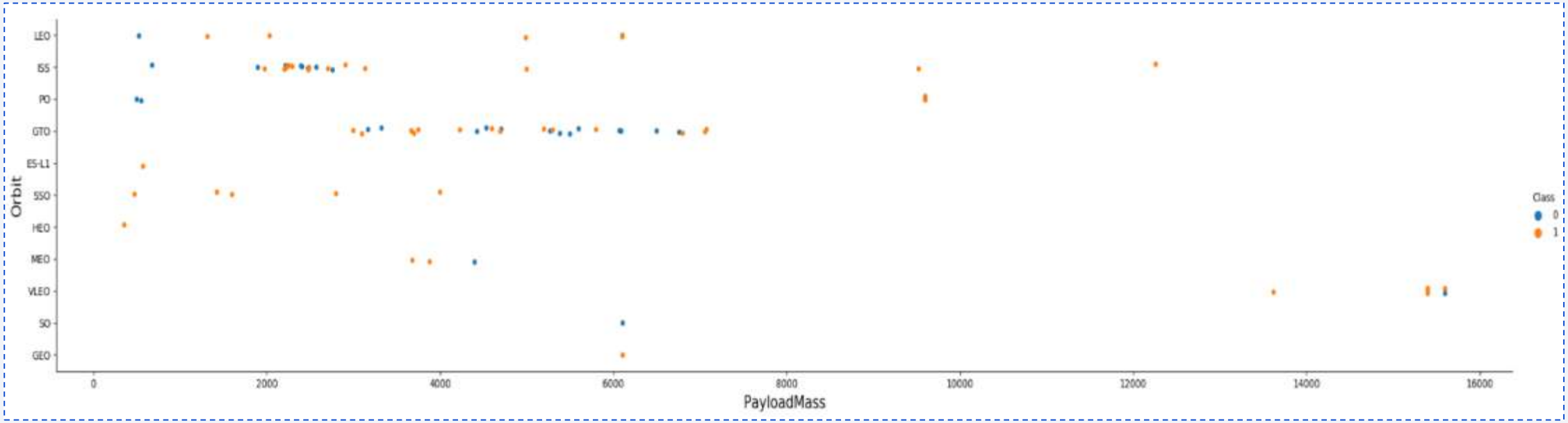
Flight Number vs. Orbit Type

- The success rate has progressively enhanced over time across all orbits, indicating an overall improvement in launch reliability. Notably is the emergence of VLEO orbit as a promising business opportunity, evident from its recent surge in launch frequency. This trend underscores the growing significance and potential profitability associated with VLEO missions in the contemporary space industry landscape.



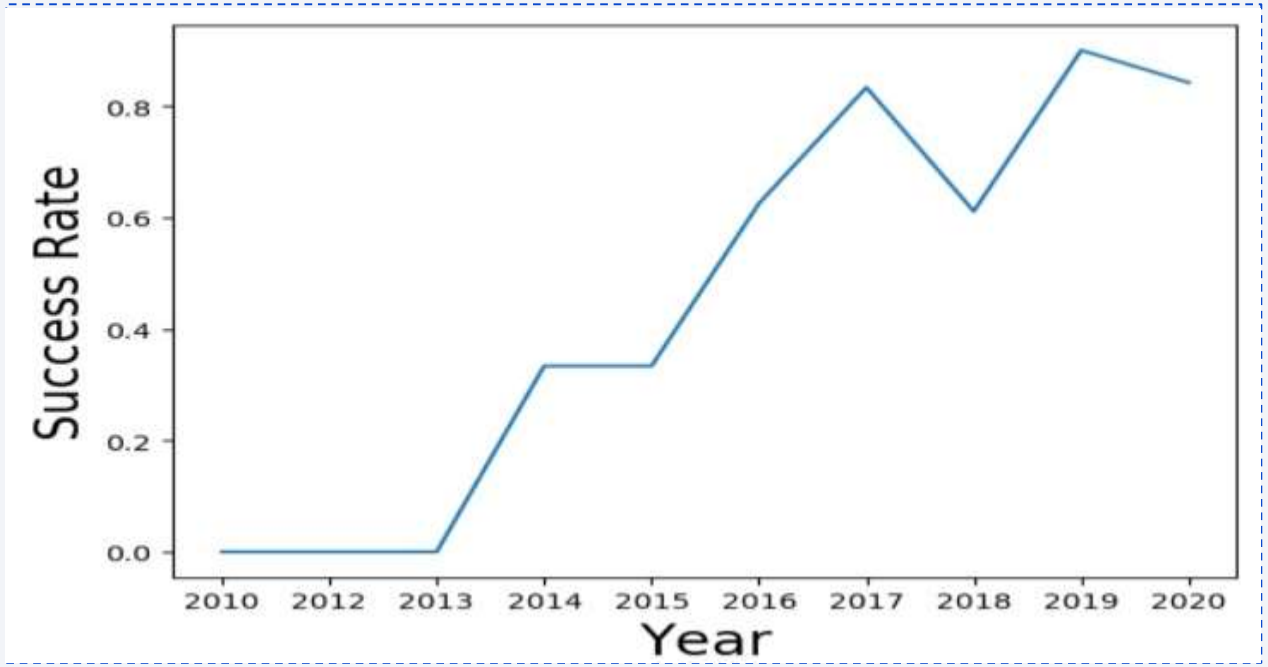
Payload vs. Orbit Type

- There is no discernible correlation between payload size and success rate for launches to the GTO orbit. On the other hand, the ISS orbit exhibits the broadest range of payload capacities while maintaining a commendable success rate. Additionally, there are relatively few launches to the SO and GEO orbits, indicating these orbits are less frequently utilized for space missions.



Launch Success Yearly Trend

The success rate began its upward trajectory in 2013 and continued to rise steadily until 2020, indicating a sustained period of improvement and refinement. The initial three years served as a crucial period for adjustments and technological advancements, laying the groundwork for the subsequent increase in success rates.



All Launch Site Names

The dataset reveals the existence of four distinct launch sites: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, and VAFB SLC-4E. These sites are identified by extracting the unique instances of "launch_site" values within the dataset, indicating the diverse locations from which launches are conducted.

Out[9]:

Launch_Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

Here are 5 records where launch sites begin with 'CCA':

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
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

Out[11]:

TOTAL_PAYLOAD
111268

- Above is the total payload carried by boosters from NASA:
- The total payload calculated previously is derived from the summation of all payloads identified with codes containing 'CRS', indicative of payloads associated with NASA missions.

Average Payload Mass by F9 v1.1

Out[12]:

AVG_PAYLOAD
2928.4

- Above displays Average payload mass carried by booster version F9 v1.1.
- After filtering the data based on the specified booster version and computing the average payload mass, the value amounts to 2928.4 kg

First Successful Ground Landing Date

In [28]:
```
import sqlite3

# Connect to the database
conn = sqlite3.connect('my_data1.db')
cursor = conn.cursor()

# Execute the SQL query
cursor.execute("SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)'")

# Fetch the result
result = cursor.fetchone()
print("First successful landing date on ground pad:", result[0])

# Close the connection
conn.close()
```

```
First successful landing date on ground pad: 2015-12-22
```

First Successful Ground Landing Date

- The first successful ground landing date is 2015-12-22
- The ‘sqlite3’ library connects to an SQLite database named 'my_data1.db’.
- It establishes a connection and creates a cursor object to execute SQL queries.
- The code then executes an SQL query to find the earliest successful landing date on the ground pad from the 'SPACEXTBL' table, filtering by the 'LANDING_OUTCOME' column.
- It fetches the result of the query, which contains the earliest successful landing date, and prints it out.
- Finally, the code closes the connection to the database to release resources and maintain data integrity.

Successful Drone Ship Landing with Payload between 4000 and 6000

- The following four boosters represent the result of selecting distinct booster versions that have successfully landed on a drone ship with a payload mass greater than 4000 but less than 6000.

Out[28]:

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

Total Number of Successful and Failure Mission Outcomes

Out[16]:

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

- Above represents several successful and failure mission outcomes. The summary above is derived from grouping mission outcomes and counting the records for each group.

Boosters Carried Maximum Payload

Out[17]:

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- Above are the boosters representing the maximum payload mass carried, as recorded in the dataset.

2015 Launch Records

Out[31]:

Month	Booster_Version	Launch_Site
01	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

- The provided list contains the sole occurrences of failed landing outcomes on a drone ship, including details such as their booster versions and launch site names, all documented within the year 2015.

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

```
('No attempt', 10)
('Success (drone ship)', 5)
('Failure (drone ship)', 5)
('Success (ground pad)', 3)
('Controlled (ocean)', 3)
('Uncontrolled (ocean)', 2)
('Failure (parachute)', 2)
('Precluded (drone ship)', 1)
```

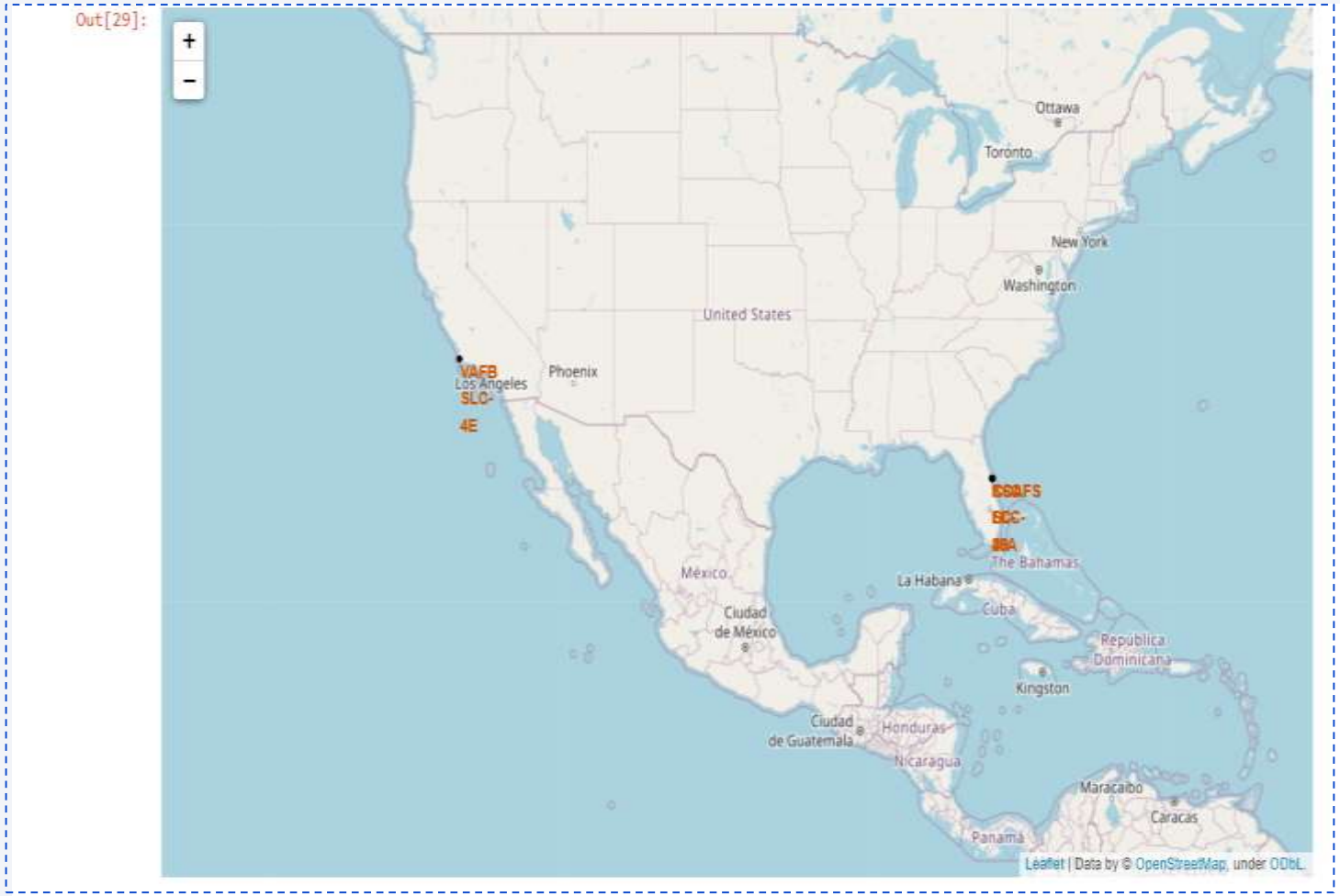
- This code output showcases the ranking of all landing outcomes recorded between June 4th, 2010, and March 20th, 2017. It's crucial to note that the category of "No attempt" should also be considered in this analysis.

Section 3

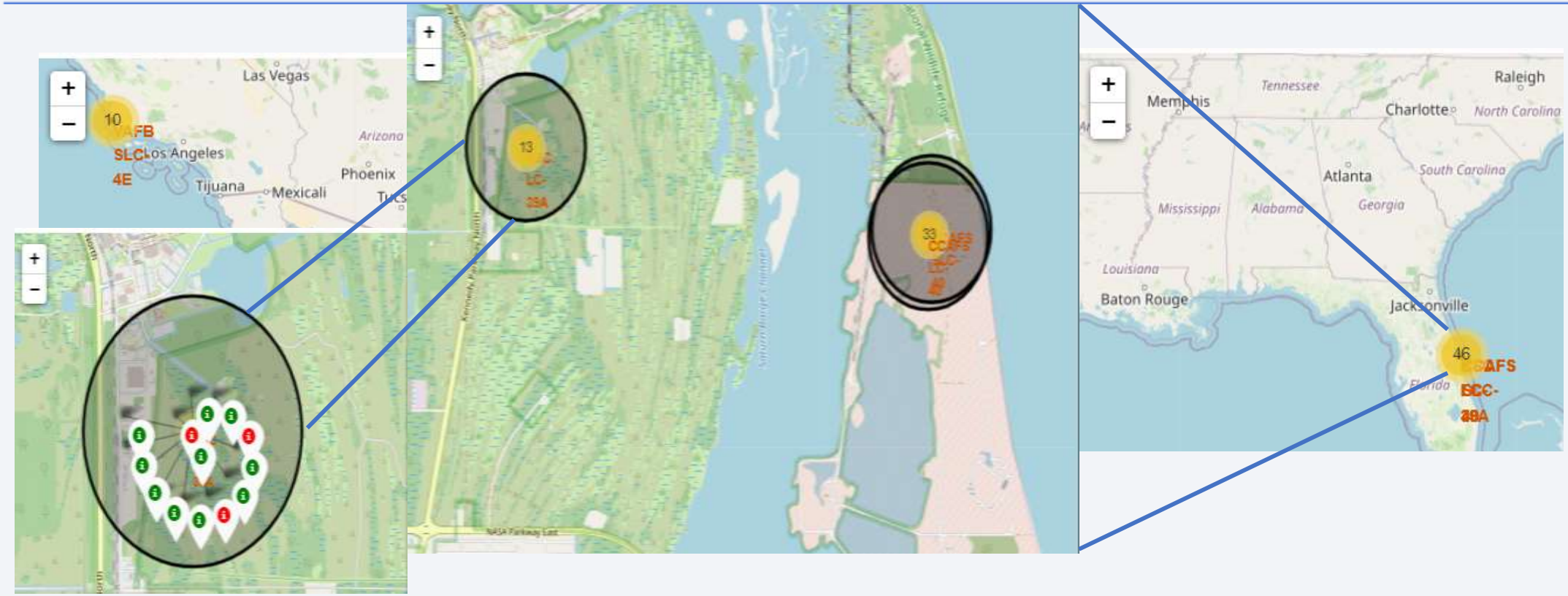
Launch Sites Proximities Analysis

All launch sites

- According to the map, Launch sites are typically situated near the sea, likely for safety considerations, while still being conveniently accessible from nearby roads and rail networks.



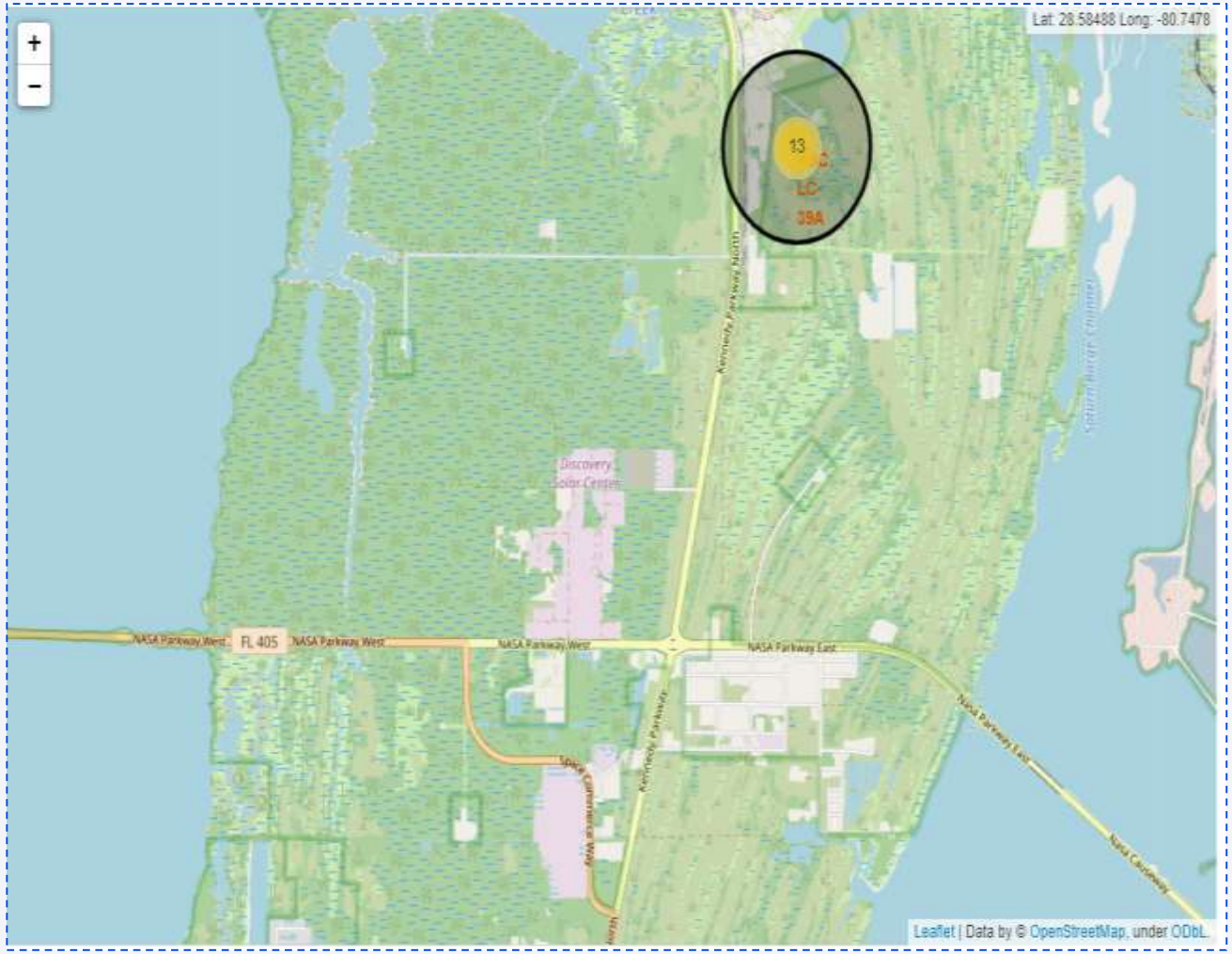
Site Launch Outcomes



- Above is a map illustration showcasing the launch outcomes at KSC LC-39A features green markers denoting successful launches and red markers indicating instances of failure.

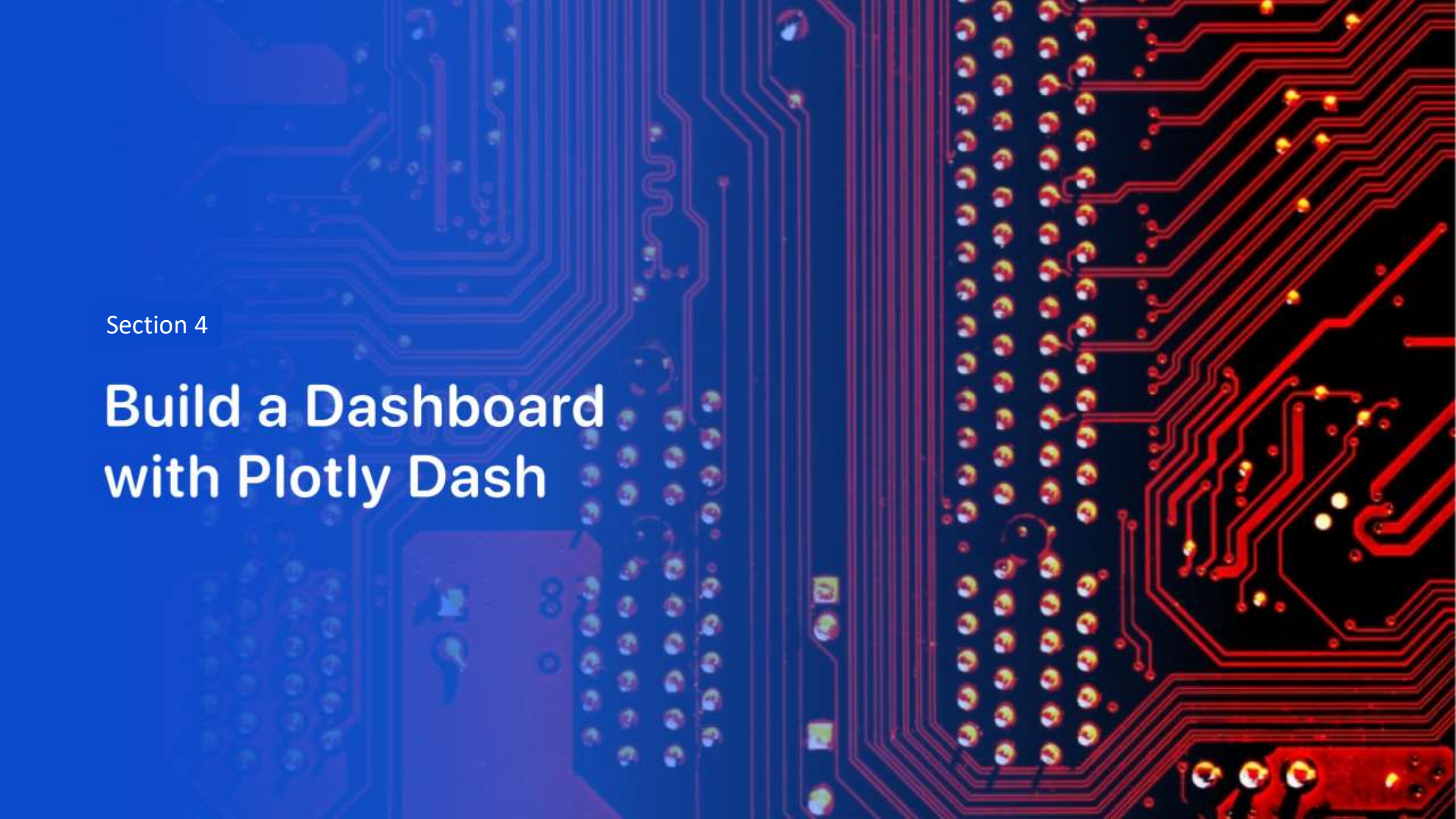
Launch site logistics and safety

- The logistics setup at launch site KSC LC-39A is favorable, benefiting from its proximity to both railroad and road infrastructure, while also maintaining a relatively safe distance from populated areas.

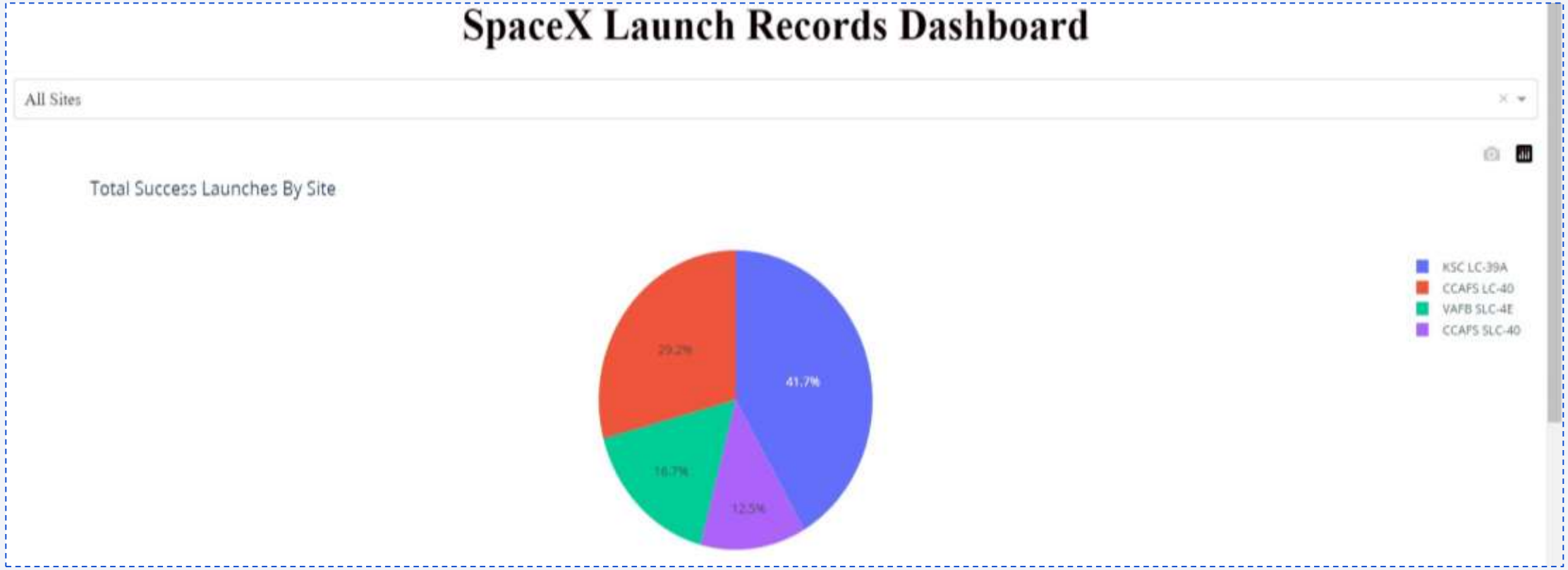


Section 4

Build a Dashboard with Plotly Dash

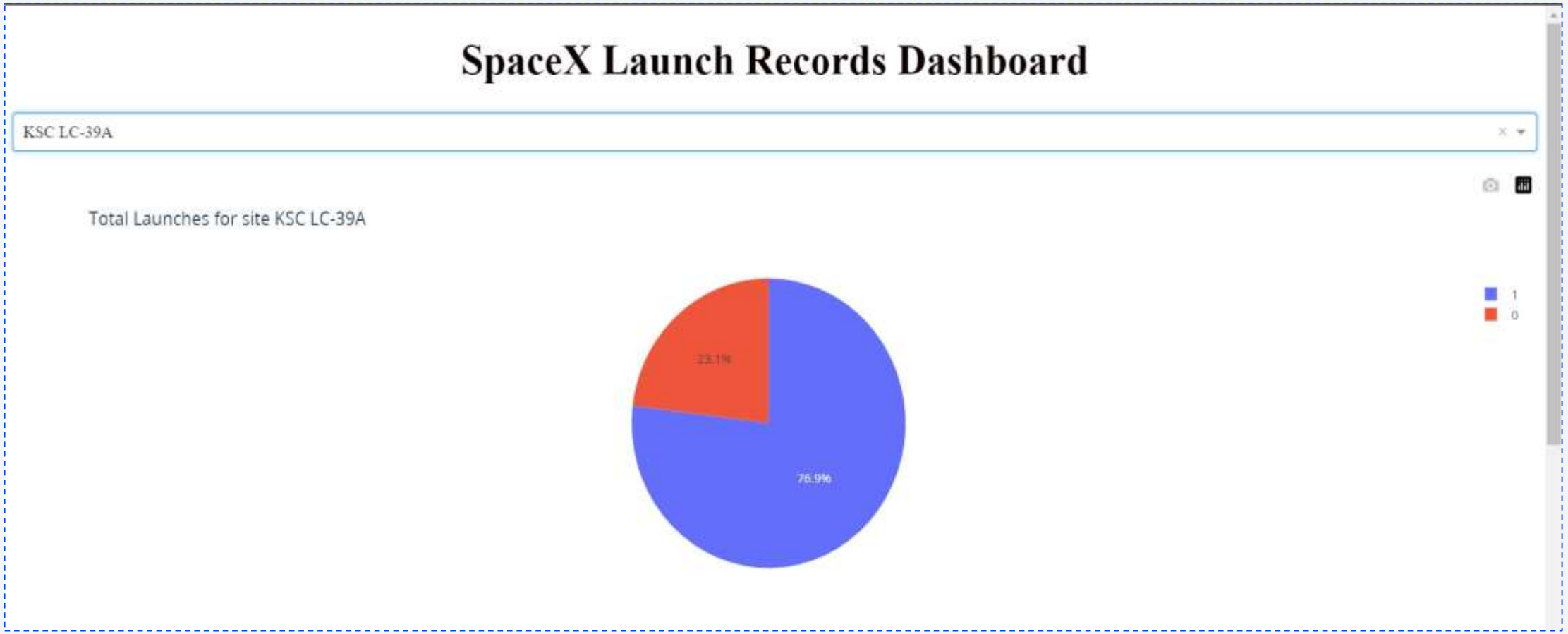


Total successful launches by site



- The screenshot indicates that the launch location plays a crucial role in the success of missions.

KSC LC-39A launch success rate

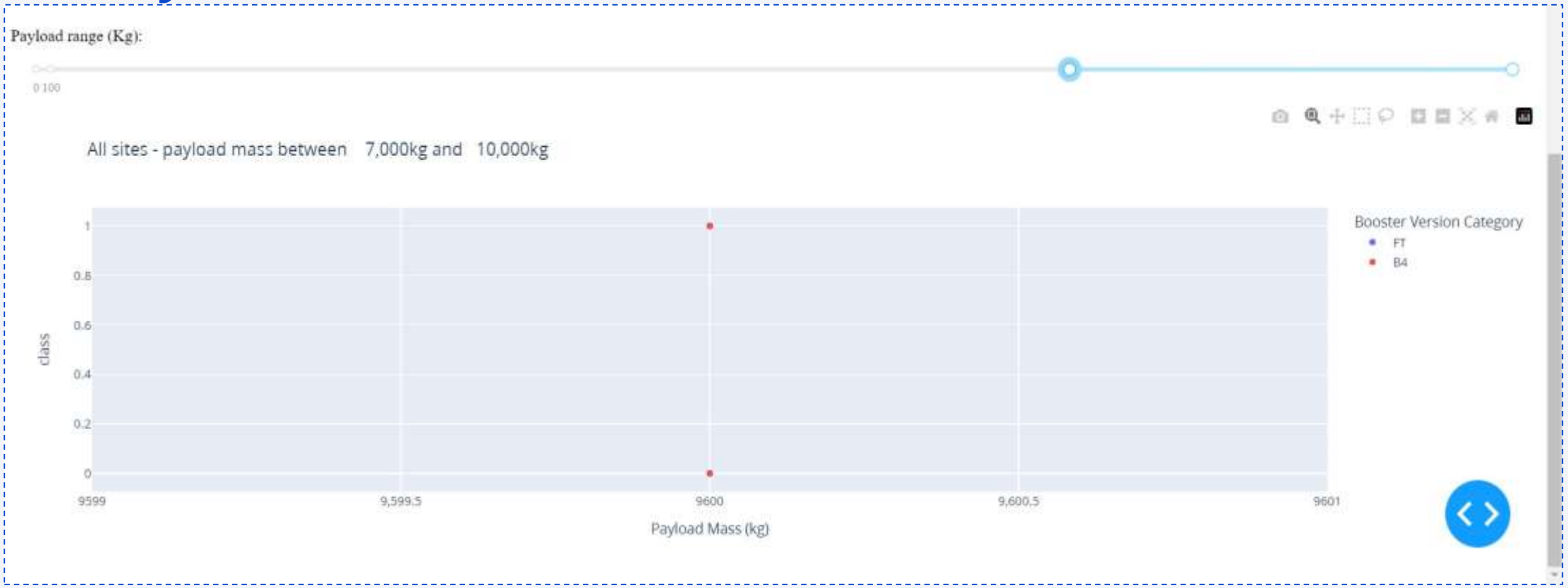


- At this site, the success rate for launches stands at 76.9%.

Payload and Launch Outcome Scatterplot Analysis

- Based on the scatterplot, the combination of payloads weighing less than 6,000kg paired with FT boosters has demonstrated the highest success rate.

Payload and Launch Outcome Scatterplot Analysis

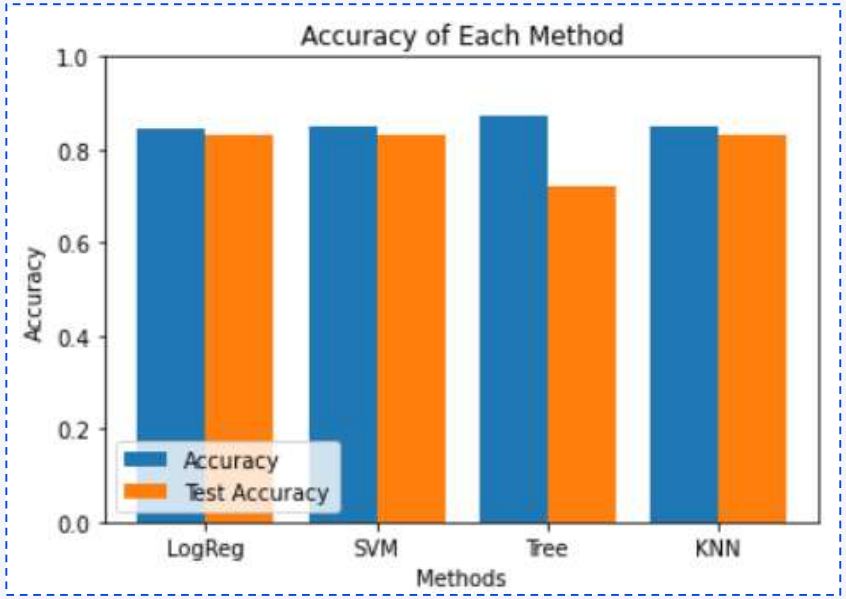


- Based on the scatterplot, Insufficient data is available to accurately assess the risk associated with launches exceeding 7,000kg.

Section 5

Predictive Analysis (Classification)

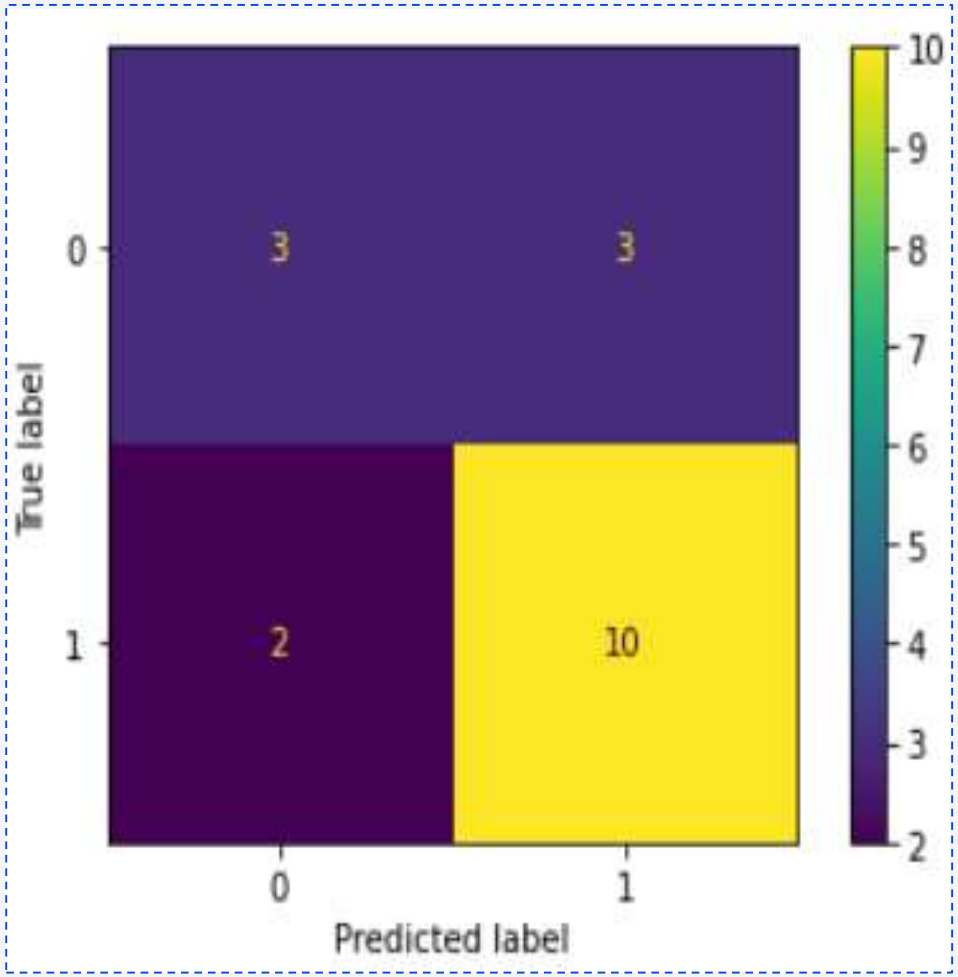
Classification Accuracy



- Four classification models underwent testing, and their respective accuracies are depicted in the accompanying plot. Among these models, the Decision Tree Classifier emerged with the highest classification accuracy, surpassing 87%.

Decision Tree Confusion Matrix

- True Positives for did not land class: 3 instances correctly predicted as did not land.
- False Negatives for landed class: 3 instances incorrectly predicted as did not land when they landed.
- False Positives for did not land class: 2 instances incorrectly predicted as did not land when they landed.
- True Positives for landed class: 12 instances correctly predicted as landed.
- The effectiveness of the Decision Tree Classifier is demonstrated through its confusion matrix, highlighting substantial counts of true positives and true negatives relative to the occurrences of false positives



Conclusions

- The analysis involved the examination of various data sources, leading to refined conclusions.
- After thorough evaluation, KSC LC-39A emerged as the optimal launch site.
- Launches involving payloads exceeding 7,000kg tend to pose lower risks.
- While overall mission success rates are promising, successful landings show improvement over time, reflecting advancements in processes and rocket technology.
- Utilizing the Decision Tree Classifier can effectively forecast successful landings, thereby enhancing profitability.

Appendix

- Full Project GitHub URL: https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone
- Launch Sites Locations Analysis with Folium screenshots (since the images did not load on GitHub): https://github.com/Sherrymdx1377/Applied-Data-Science-Capstone/tree/main/week%203/Interactive%20Visual%20Analytics%20with%20Folium%20lab

Thank you!