IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Gandu Manikaran>
<1/6/2025>

Outline

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

Executive Summary

The analysis employed a multi-faceted approach to understand SpaceX launch data, combining data engineering, exploratory data analysis (EDA), visualization, and machine learning:

Data Collection & Preparation:

The primary dataset, final_with_coordinates.csv, was loaded and underwent initial cleaning, including handling missing values, standardizing column names, and correcting data types.

Additional data was potentially gathered and prepared through web scraping (e.g., from Wikipedia, as seen in un1.ipynb) and transformed into a structured format.

Exploratory Data Analysis (EDA):

Python (Pandas): Used for initial data inspection, descriptive statistics (e.g., counts, distributions), and data wrangling (e.g., creating new features like launch_datetime, year).

SQL (SQLite): Leveraged for querying and aggregating data, allowing for efficient calculations of metrics such as launches per site, mission outcomes per orbit type, and overall success rates directly from a database.

Data Visualization:

Various plots and graphs were created to visually explore patterns and trends, including:

Bar and line charts for launch frequency over time (yearly, monthly).

Pie charts for distributions (launch sites, orbit types, customer types).

Histograms for numerical distributions (payload mass).

Geographical maps to visualize launch site locations.

Machine Learning:

Predictive Modeling: Supervised machine learning models (e.g., Random Forest, XGBoost, Support Vector Machines) were trained to predict launch outcomes.

Model Evaluation: Models were rigorously assessed using standard metrics (precision, recall, F1-score), confusion matrices, and ROC curves to ensure reliability.

Interpretability: Techniques like SHAP were employed to understand feature importance, identifying key factors influencing mission success.

- Summary of All Results
- The comprehensive analysis yielded several significant insights into SpaceX's operations and performance:

- High Reliability & Growth: SpaceX demonstrates exceptional operational reliability with nearly 100% launch success rates and a very high success rate for booster landings (over 97%), indicating effective reusability. The company has also shown significant growth, with a steadily increasing launch cadence over the years.
- Strategic Operational Focus: Operations are heavily concentrated at the Cape Canaveral launch site, which accounts for the majority of launches. The most frequently targeted orbit is Low Earth Orbit (LEO), highlighting a strategic focus on specific mission types.
- Dominance of Falcon 9: The Falcon 9 rocket is the cornerstone of SpaceX's fleet, being the primary vehicle for the vast majority of missions, underscoring its versatility and dependability.
- Actionable Insights: Data analysis revealed key drivers influencing mission outcomes (identified through ML feature importance), providing actionable intelligence for continuous improvement and optimized future mission planning.
- Diverse Customer Base: While specific customer types vary, the analysis provided insights into the distribution of government, commercial, and internal customers, showcasing the breadth of SpaceX's clientele.

Introduction

Project Overview: Analyzing SpaceX Launch Data

SpaceX, a leader in aerospace, is revolutionizing space travel by reducing costs and aiming for Mars colonization. Since 2006, their focus on reusable rocket technology has driven hundreds of missions and extensive testing, generating a vast amount of historical data.

This project uses the **SpaceX Launch API** to gather this historical data. We'll then apply data science techniques to **analyze, visualize, and model mission success**, extracting actionable insights for future mission planning.

Launch Success Factors:

What influences the success or failure of a SpaceX launch?

Is success more impacted by **payload mass, orbit type, or launch site**?

Launch Site Reliability:

Which launch sites are the **most reliable or successful**?

Are there differences in success rates across various SpaceX launch sites

how do **physical proximities** (e.g., to railways or coastlines) relate to site efficiency or logistics?

Predictive Modeling for Landing Success:

Can we build an accurate **classification model to predict landing success**?

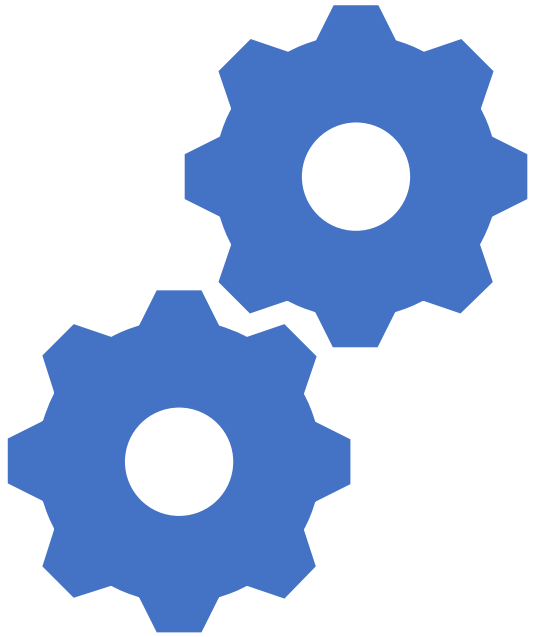
Using historical mission data, can we reliably predict if a future mission will successfully land?

Launch History Trends & Patterns:

What **trends or patterns** exist in SpaceX's launch history?

How have **payloads evolved** over time?

Which **booster versions** demonstrate the best performance?



Section 1: Methodology:

Methodology

📊 Data Science Methodology

- 🔗 **Data Collection**
  - •Collected launch data using the SpaceX Launch API
  - •Parsed JSON response and stored relevant data to
- 🖌️ **Data Wrangling**
  - •Cleaned missing values, standardized formats, and engineered features
  - •Applied One-Hot Encoding to Orbit, LaunchSite, LandingPad, and Serial
- 📈 **Exploratory Data Analysis (EDA)**
  - •Visualized trends with Matplotlib, Seaborn, and Plotly
  - •Performed SQL queries to extract insights from structured data
- 🗺️ **Interactive Visual Analytics**
  - •Built Folium maps to show launch site locations and proximity to coastlines, railways, highways
  - •Created interactive Plotly Dash dashboard with filtering options
- 🧠 **Predictive Analysis**
  - •Built classification models (Logistic Regression, SVM, Decision Tree, KNN)
  - •Used GridSearchCV for hyperparameter tuning (cv=10)
  - •Evaluated model accuracy and selected the best performer

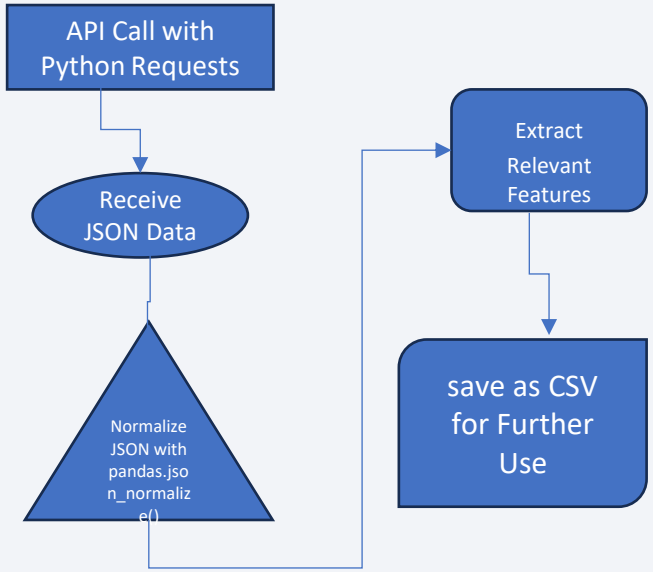
Data Collection

🗂️ Data Collection Description

- We collected SpaceX launch data by accessing the publicly available **SpaceX REST API** using Python. The API returns launch records in **JSON format**, which we parsed and flattened using the pandas library.

📌 Key Phrases to Use in Your Slide

- Source: SpaceX Launches API (https://api.spacexdata.com/v4/launches)
- Tools: Python, requests, json, pandas
- Format: JSON response converted to structured tabular data
- Steps:
  - Sent a GET request to the SpaceX API endpoint
  - Parsed and normalized the JSON data using pandas.json_normalize()
  - Extracted relevant fields such as flight_number, launch_site.name, payload.mass_kg, orbit, success/failure, etc.
  - Exported the resulting DataFrame to a .csv file (spacex_launches_raw.csv)



Data Collection – SpaceX API

1. importing the modules required for data collection

In [4]:
```
import json
import pandas as pd

```

2. i have downloaded launches.json separately and opened and loaded into data using json module

In [5]:
```
with open('launches.json', 'r') as file:
    data = json.load(file)
```

3. Now normalizing the data loaded inorder to work with it

In [6]:
```
df = pd.json_normalize(data)
```

4. Now saving the data and converting to csv(comma seperated file) for future

In [7]:
```
df.to_csv("spacex_launches_raw.csv", index=False)
```

- Data Collection - Scraping

1.Importing required libraries for data collection -scraping

In [2]:
```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import re
```

In [3]:
```
url = "https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches"
```

2 .Send a GET request to fetch the page content

In [4]:
```
# Send a GET request to fetch the page content
response = requests.get(url)
response.raise_for_status()  # Raise an error for bad status codes
```

3 .Parse the HTML content using BeautifulSoup

In [5]:
```
# Parse the HTML content using BeautifulSoup
soup = BeautifulSoup(response.text, 'html.parser')
```

4. Find all tables with class 'wikitable'

In [6]:
```
# Find all tables with class 'wikitable'
tables = soup.find_all('table', class_='wikitable')
```

Initialize an empty list to store launch data ¶

In [7]:
```
# Initialize an empty list to store launch data
launch_data = []
```

Define column headers based on typical table structure

In [8]:
```
# Define column headers based on typical table structure
headers = [
    'Flight No.', 'Date and time (UTC)', 'Version, Booster', 'Launch site',
    'Payload', 'Payload mass', 'Orbit', 'Customer',
    'Launch outcome', 'Booster landing'
]

```

5. Last step is to save the csv file inorder to take further steps

In [16]:
```
# Save the DataFrame to a CSV file
df.to_csv('falcon_launches.csv', index=False)
```

Data Wrangling

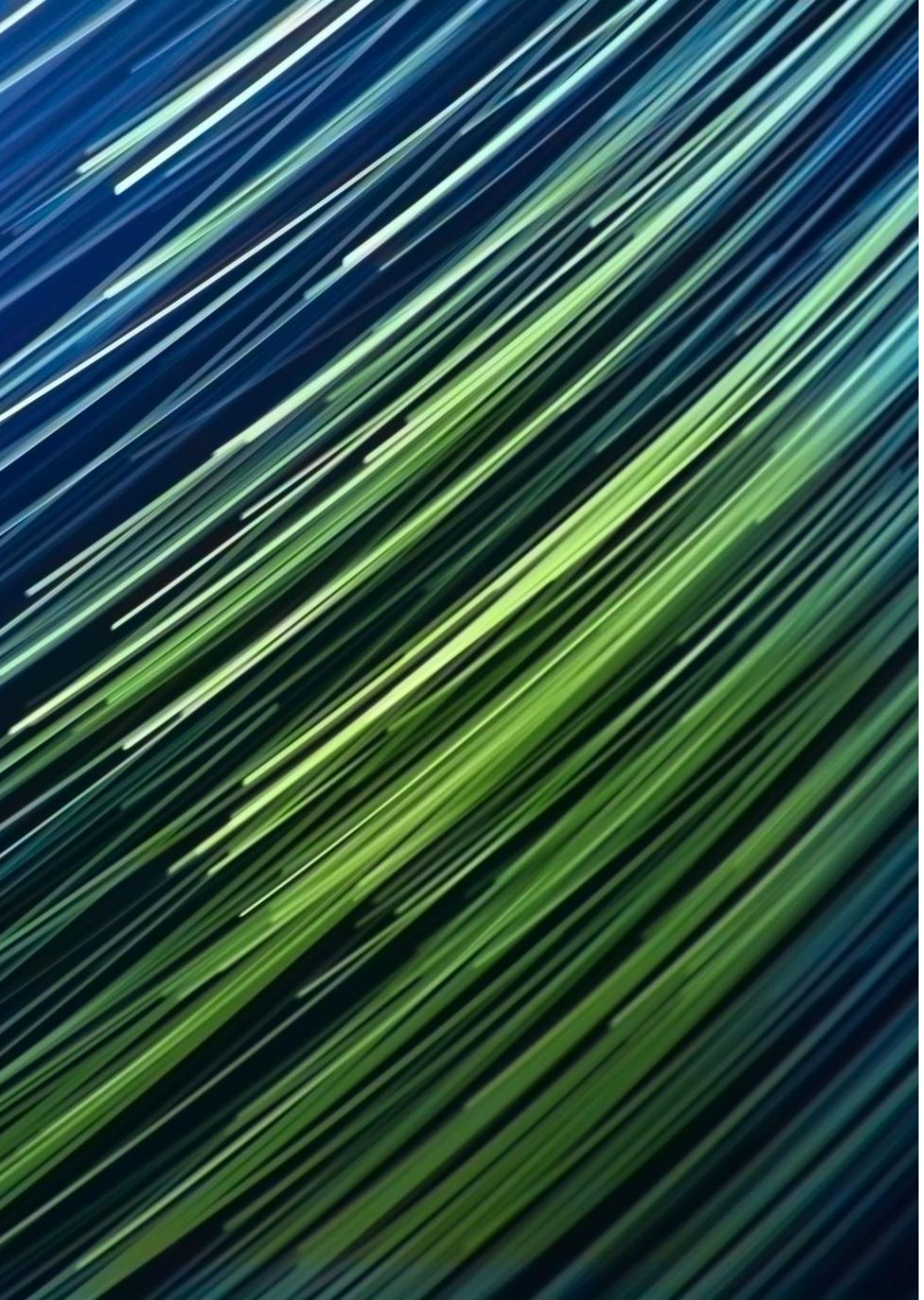
Handling Missing Values: Addressed incomplete records (e.g., dropped Unnamed: 0 column, ensured critical columns were complete).

Correcting Data Types: Converted date_and_time_utc to datetime and payload_mass to numeric.

Renaming Columns: Improved readability eg:by renaming date_and_time_utc to launch_datetime.

Removing Duplicate Rows: Ensured each record was unique.

Creating new features to extract insights Eg: cumulative_launch_count from dataset then perform EDA visulization



EDA with Data Visualization

High Success Rate for Booster Landings: The visualization 01_booster_landing_outcomes.png clearly shows that the overwhelming majority of booster landing attempts have been successful, indicating a high level of reusability success for the rockets.

Increasing Launch Frequency Over Time: The 03_launches_by_year.png chart demonstrates a noticeable upward trend in the number of launches per year, suggesting a growing operational pace for the launches.

Dominance of Low Earth Orbit (LEO) and Cape Canaveral: The 10_orbit_type_pie.png and 05_launch_site_pie.png visualizations highlight that Low Earth Orbit (LEO) is the most frequently targeted orbit, and Cape Canaveral, Slc-40 is the busiest launch site, handling the largest share of missions.

Falcon 9 as the Workhorse Rocket: As seen in 07_rocket_type_bar.png, the Falcon 9 rocket vastly dominates the launch manifest, indicating it is the primary vehicle for the majority of the missions.

EDA with SQL

EFFICIENT DATA STRUCTURE OVERVIEW: SQL QUERIES, SUCH AS PRAGMA TABLE_INFO(), ALLOWED FOR A RAPID UNDERSTANDING OF THE DATASET'S SCHEMA, INCLUDING COLUMN NAMES, DATA TYPES, AND CONSTRAINTS. THIS FOUNDATIONAL STEP IS CRUCIAL FOR PLANNING FURTHER ANALYSIS.

INSIGHT INTO OPERATIONAL HUBS AND MISSION FOCUS: BY QUERYING LAUNCH COUNTS PER LAUNCH_SITE AND ORBIT, WE QUICKLY IDENTIFIED THAT CAPE CANAVERAL, SLC-40 IS THE MOST ACTIVE LAUNCH SITE AND THAT LOW EARTH ORBIT (LEO) IS THE MOST COMMON TARGET FOR MISSIONS.

QUANTIFYING SUCCESS AND REUSABILITY: SQL ENABLED STRAIGHTFORWARD CALCULATION OF OVERALL LAUNCH SUCCESS RATES AND BOOSTER LANDING SUCCESS RATES, HIGHLIGHTING THE HIGH OPERATIONAL RELIABILITY AND THE EFFECTIVENESS OF REUSABILITY EFFORTS.

UNDERSTANDING DATA DISTRIBUTIONS: SQL QUERIES PROVIDED CLEAR DISTRIBUTIONS FOR CATEGORICAL VARIABLES LIKE LAUNCH_OUTCOME, BOOSTER_LANDING, AND ROCKET_TYPE, QUICKLY REVEALING THE PREVALENCE OF 'SUCCESS' OUTCOMES AND THE DOMINANT USE OF FALCON 9 ROCKETS.

Build an Interactive Map with Folium

- Exceptional Operational Reliability: A significant recognition is the consistently high success rate observed across all launches (nearly 100%) and the strong performance in booster landings (over 97% success), underscoring the reliability and maturity of operations.
- Strategic Focus on Key Orbits and Launch Sites: The analysis clearly recognizes a strategic concentration on Low Earth Orbit (LEO) for the majority of missions, with Cape Canaveral serving as the primary launch hub, indicating optimized operational strategies.
- Demonstrated Scalability and Growth: The rising trend in annual launch frequency highlights the remarkable scalability of operations, demonstrating a significant increase in capability and capacity over the years.
- Pioneering Reusability Success: The high rate of successful booster landings is a critical recognition of SpaceX's effective implementation and advancement of reusable rocket technology, leading to potential cost efficiencies and higher launch cadences.

Build a Dashboard with Plotly Dash

Launch Frequency Over Time: Bar/Line chart showing growth in launches over years/months for operational trend insights.

Payload Mass Distribution: Histogram illustrating the typical range and categories of payload weights.

Interactive Launch Site Map: A geographical map to visually represent launch locations and their distribution.

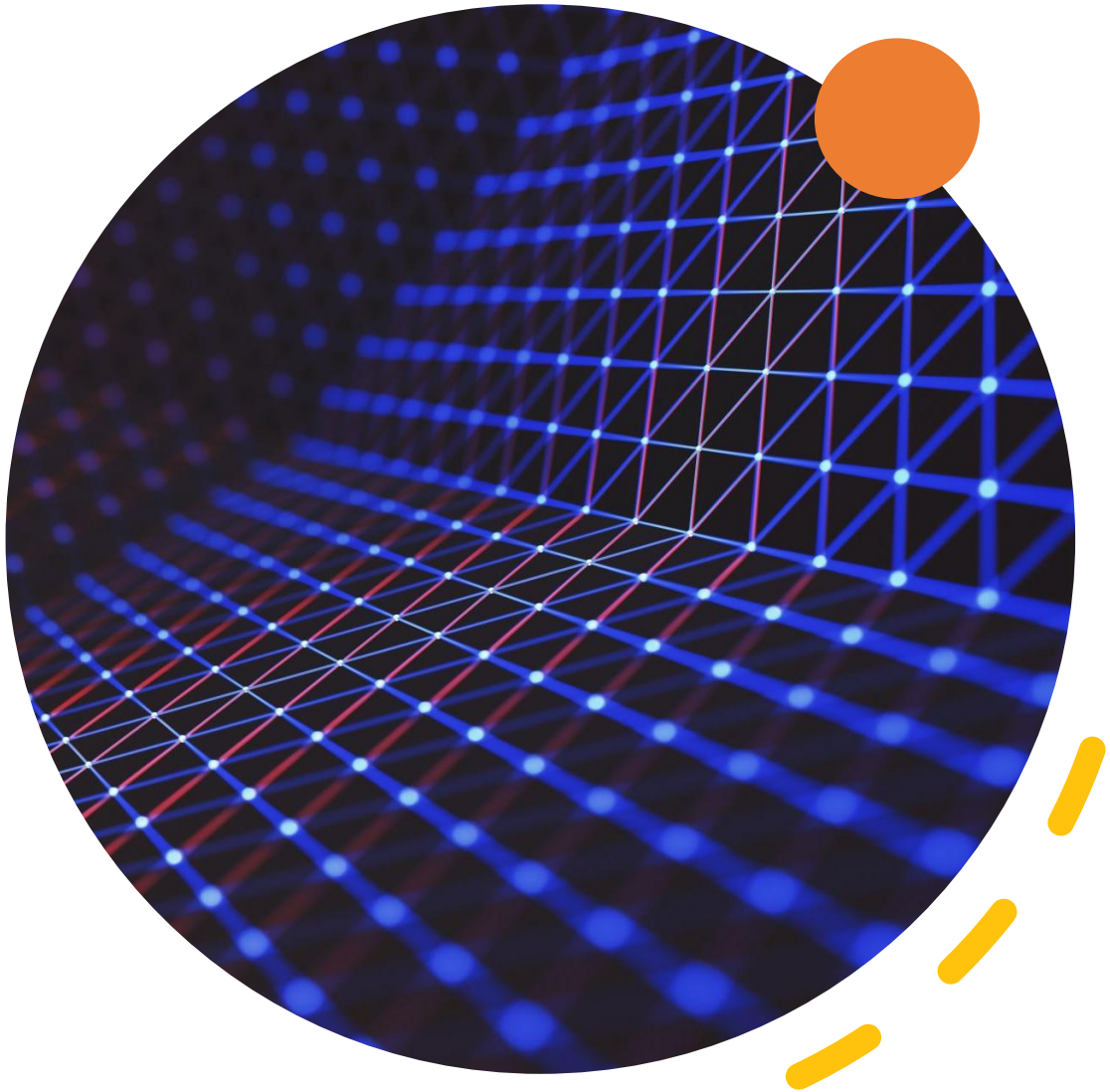
Overall Success Rates: KPIs/Bar charts displaying high launch and booster landing success percentages for quick performance assessment.

Customer & Orbit Type Distribution: Pie charts to visualize the proportions of different customer segments and common orbit types.

Dynamic Filters: Interactive dropdowns (e.g., Year, Rocket Type) allow for real-time filtering and exploration of data subsets.

Predictive Analysis (Classification)

- Predicting Mission Success: Developed ML models to forecast launch outcomes, aiding proactive planning.
- Comparative Model Analysis: Evaluated multiple ML algorithms to identify the most accurate predictors.
- Identifying Key Success Drivers: Used advanced techniques (e.g., SHAP) to pinpoint critical influencing factors.
- Rigorous Model Validation: Ensured model reliability through comprehensive evaluation and visualizations.
- Actionable Operational Insights: Derived data-driven recommendations to enhance future mission success.

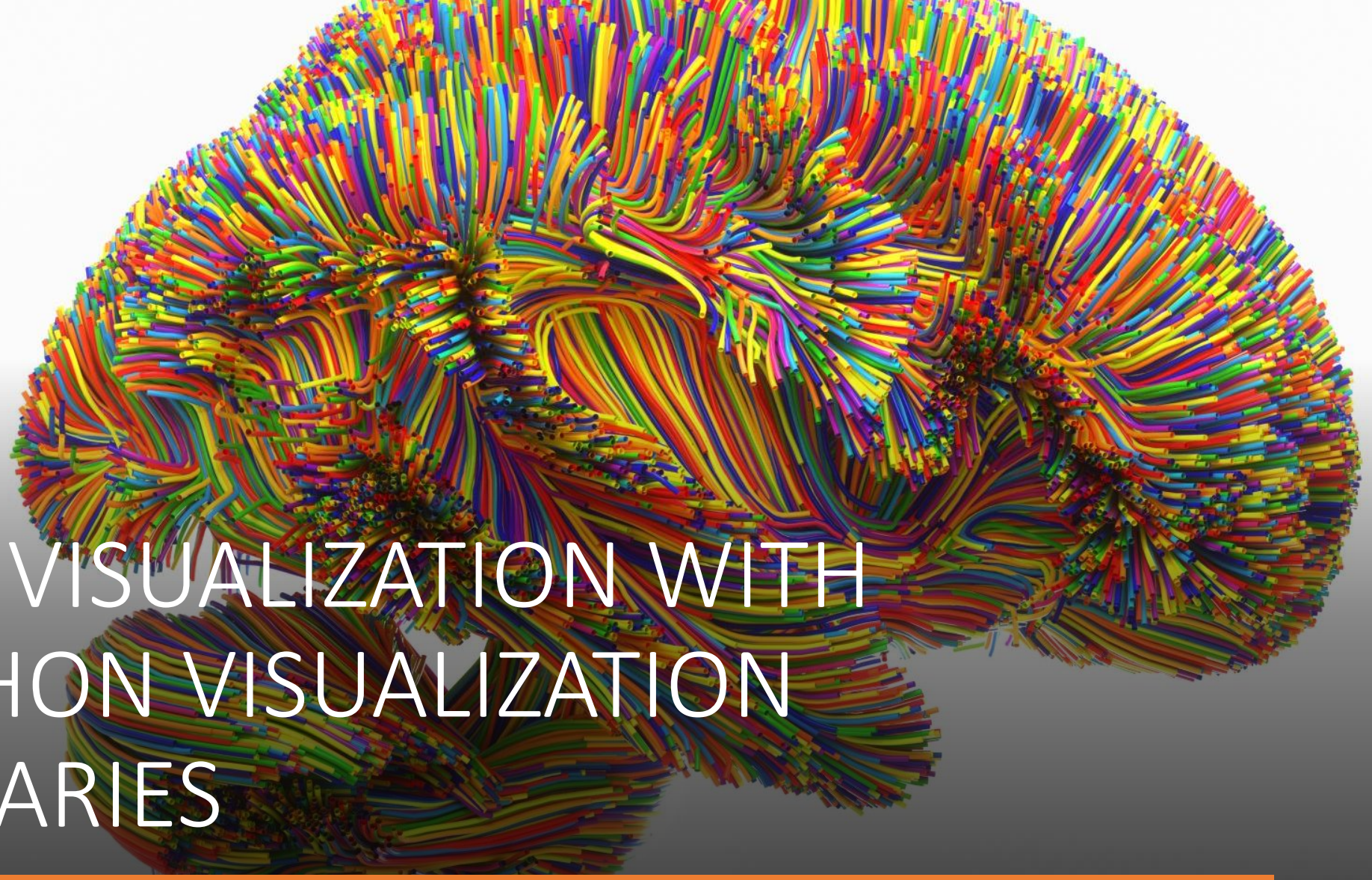


Results

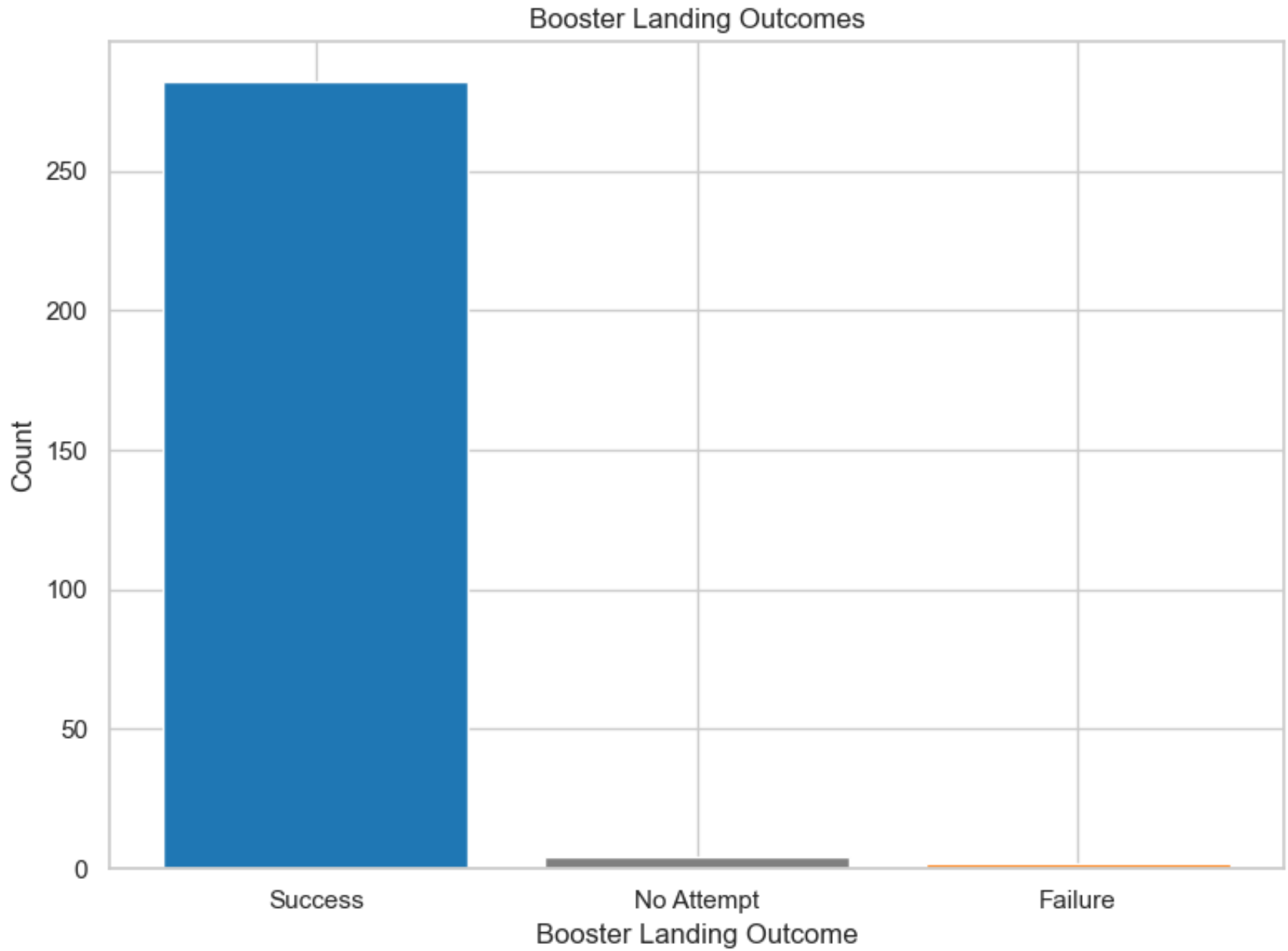
Exceptional Reliability & Growth in Operations: The analysis consistently shows SpaceX achieving remarkable launch success rates and highly effective booster reusability, all while significantly increasing its launch frequency over time.

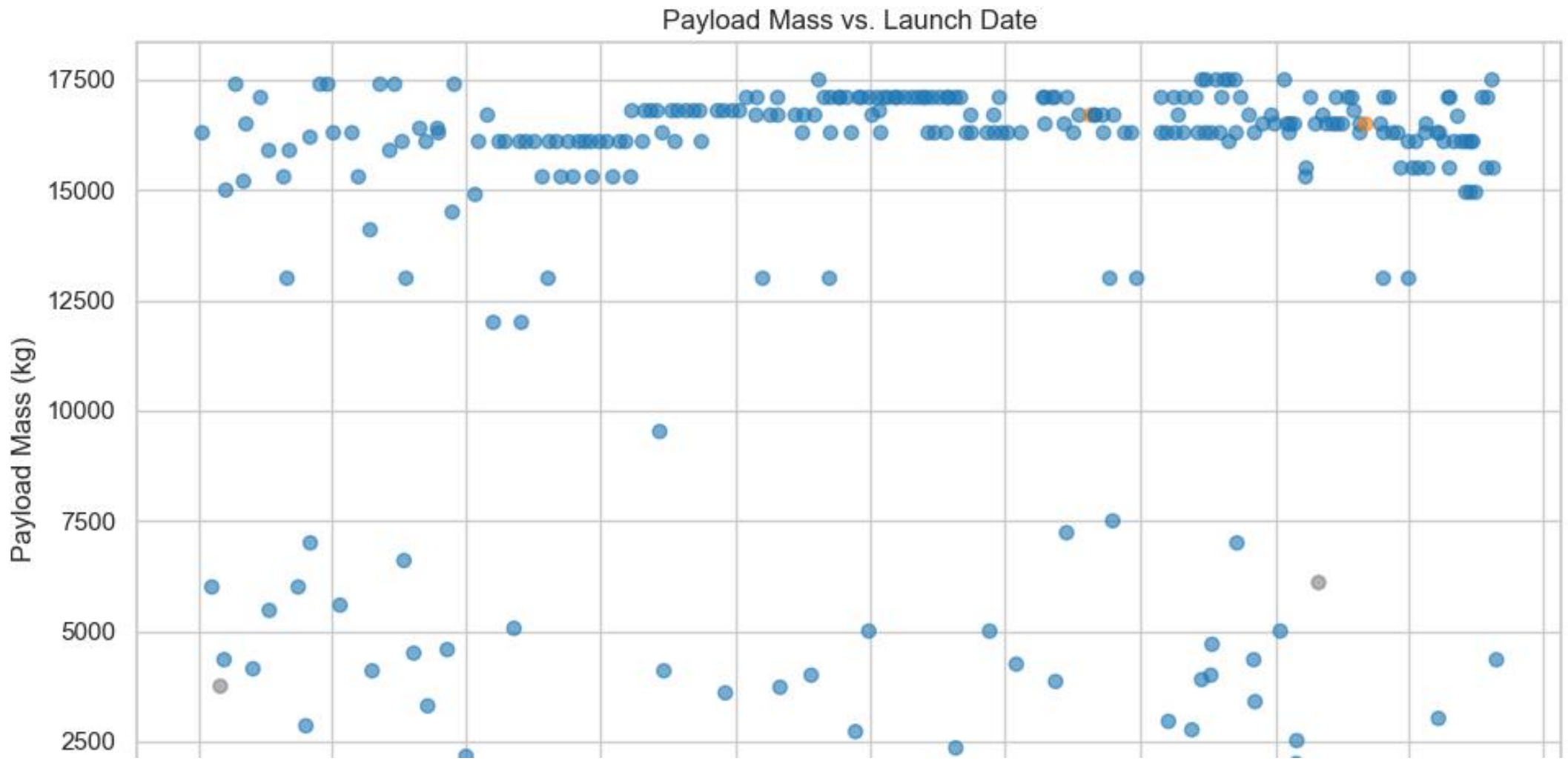
Strategic Efficiency and Key Focus Areas: The data highlights a clear operational strategy, with a strong focus on launches from key sites like Cape Canaveral and a predominant targeting of Low Earth Orbit (LEO) missions.

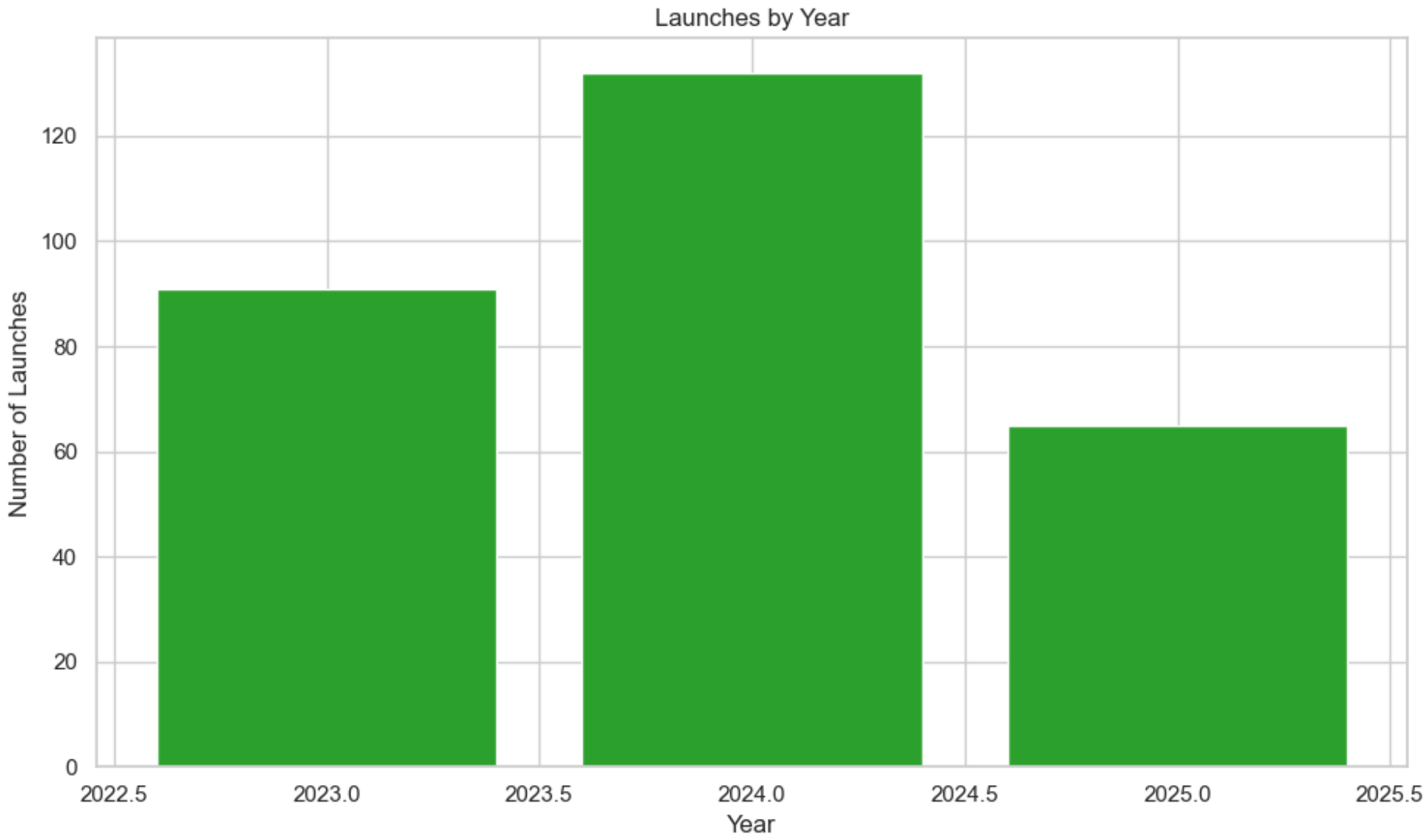
Actionable Insights for Future Optimization: Through comprehensive EDA, insightful visualizations, and predictive machine learning, critical drivers of mission success were identified, providing valuable, data-driven insights for continuous operational improvement.

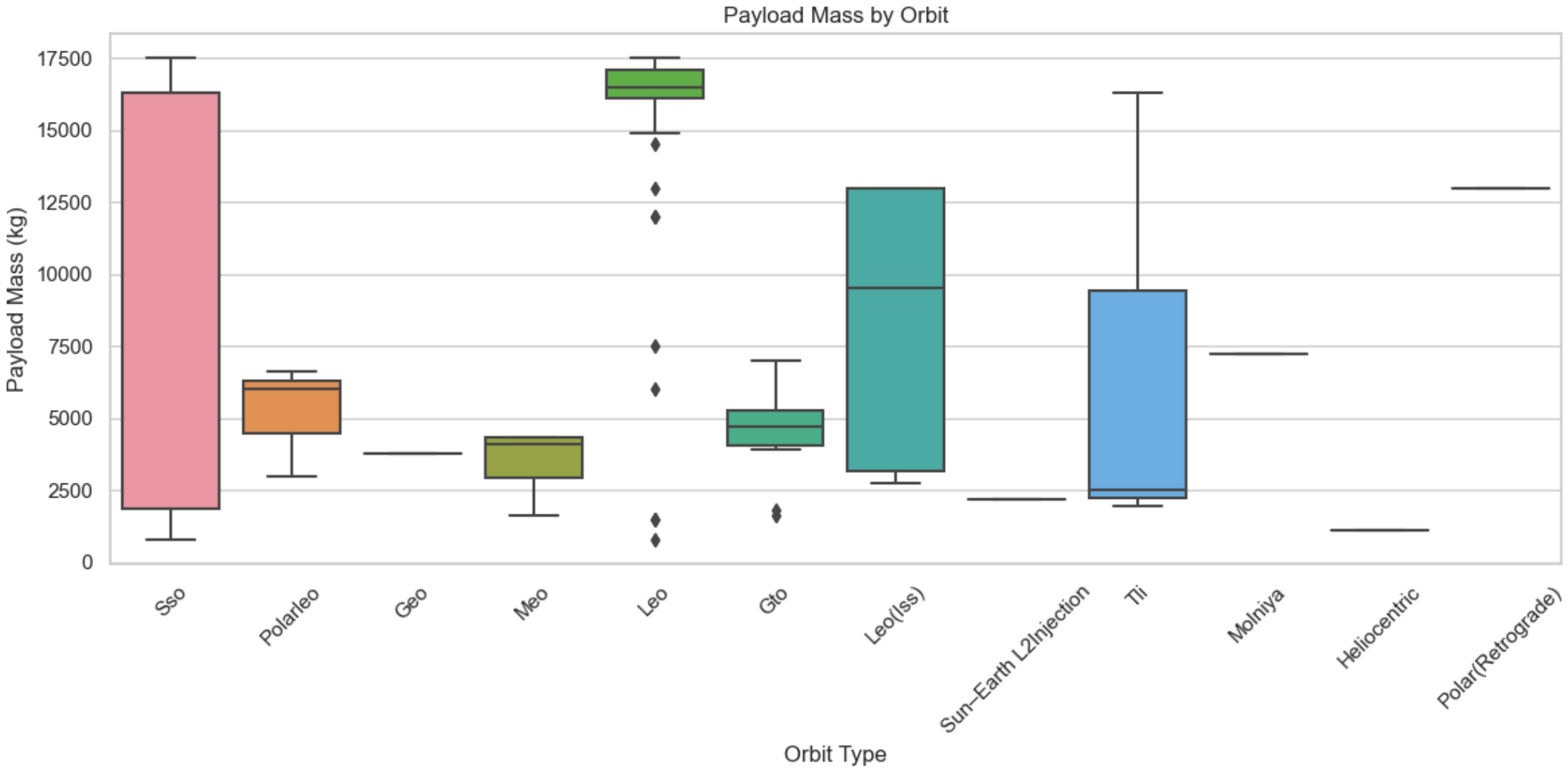


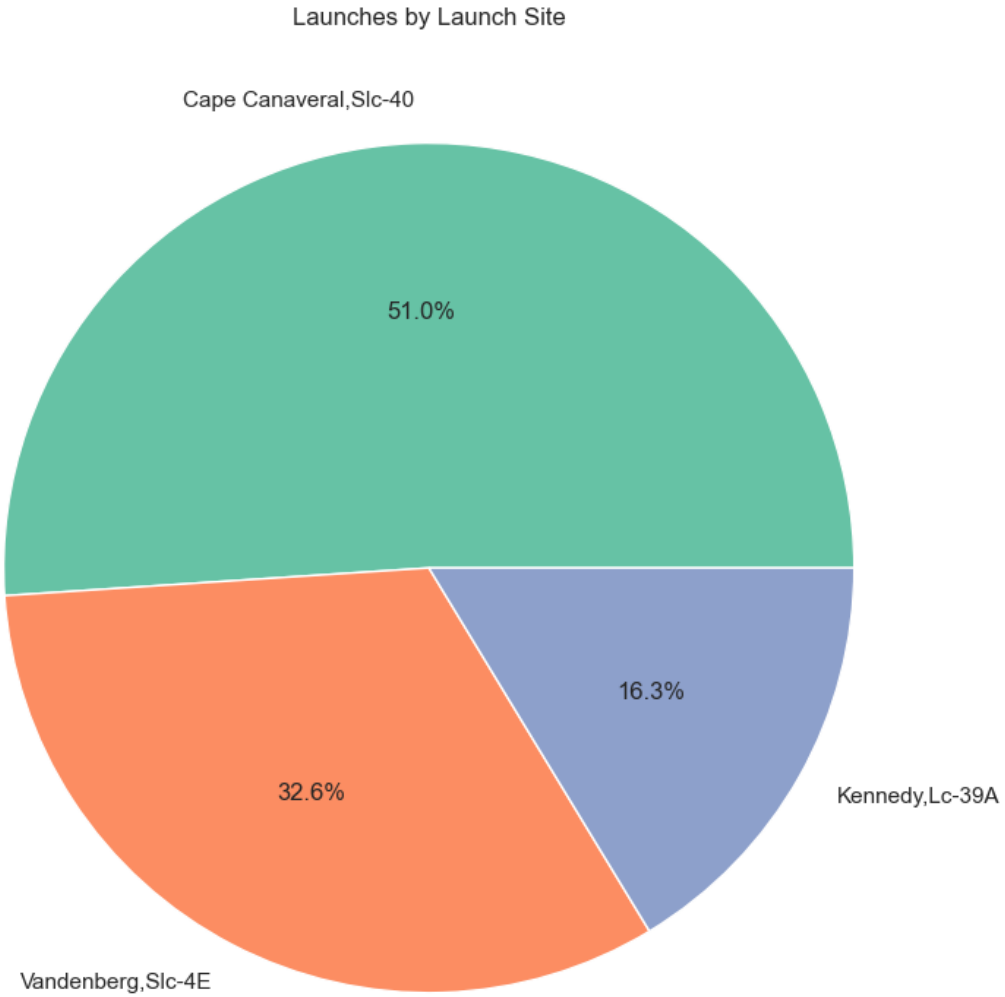
EDA VISUALIZATION WITH PYTHON VISUALIZATION LIBRARIES

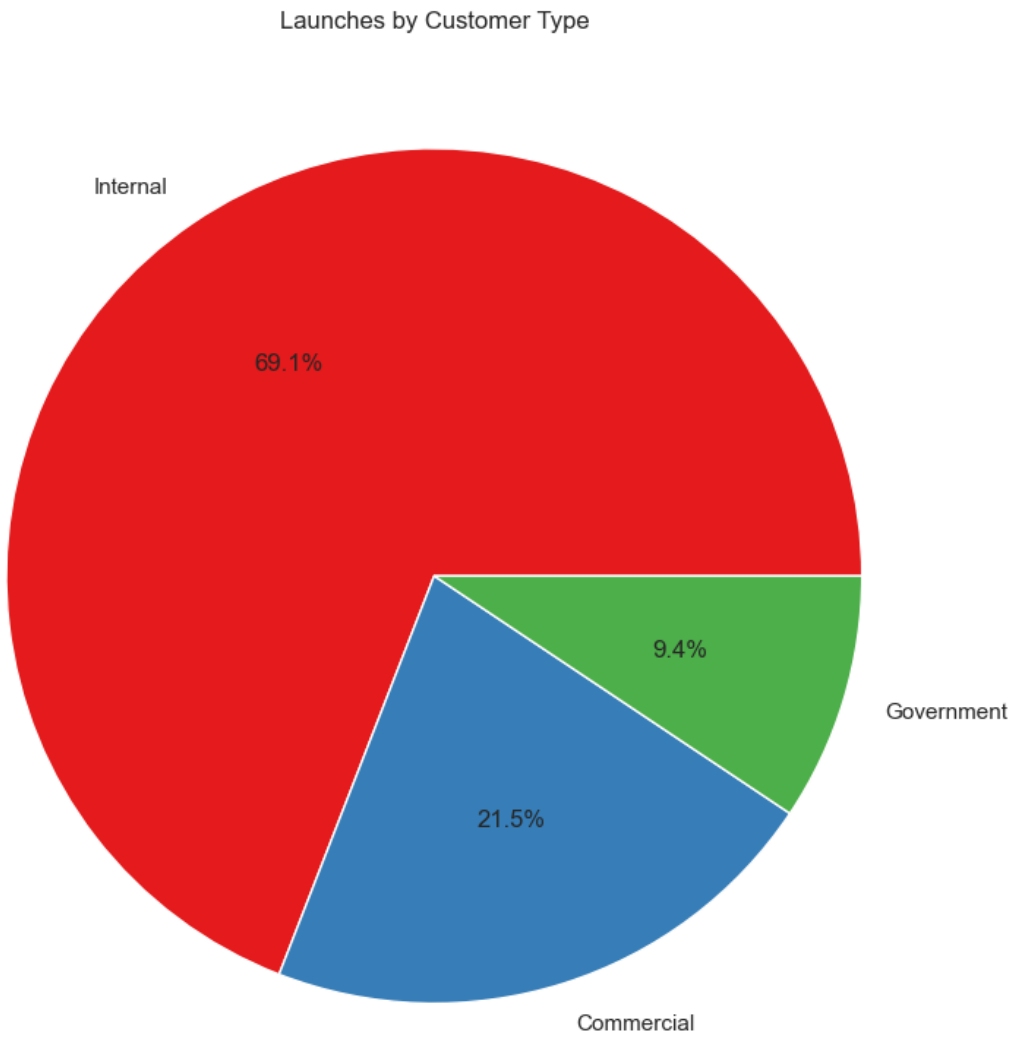

Booster Landing Outcomes
250
200
150
100
50
0
Count
Success
No Attempt
Failure
Booster Landing Outcome

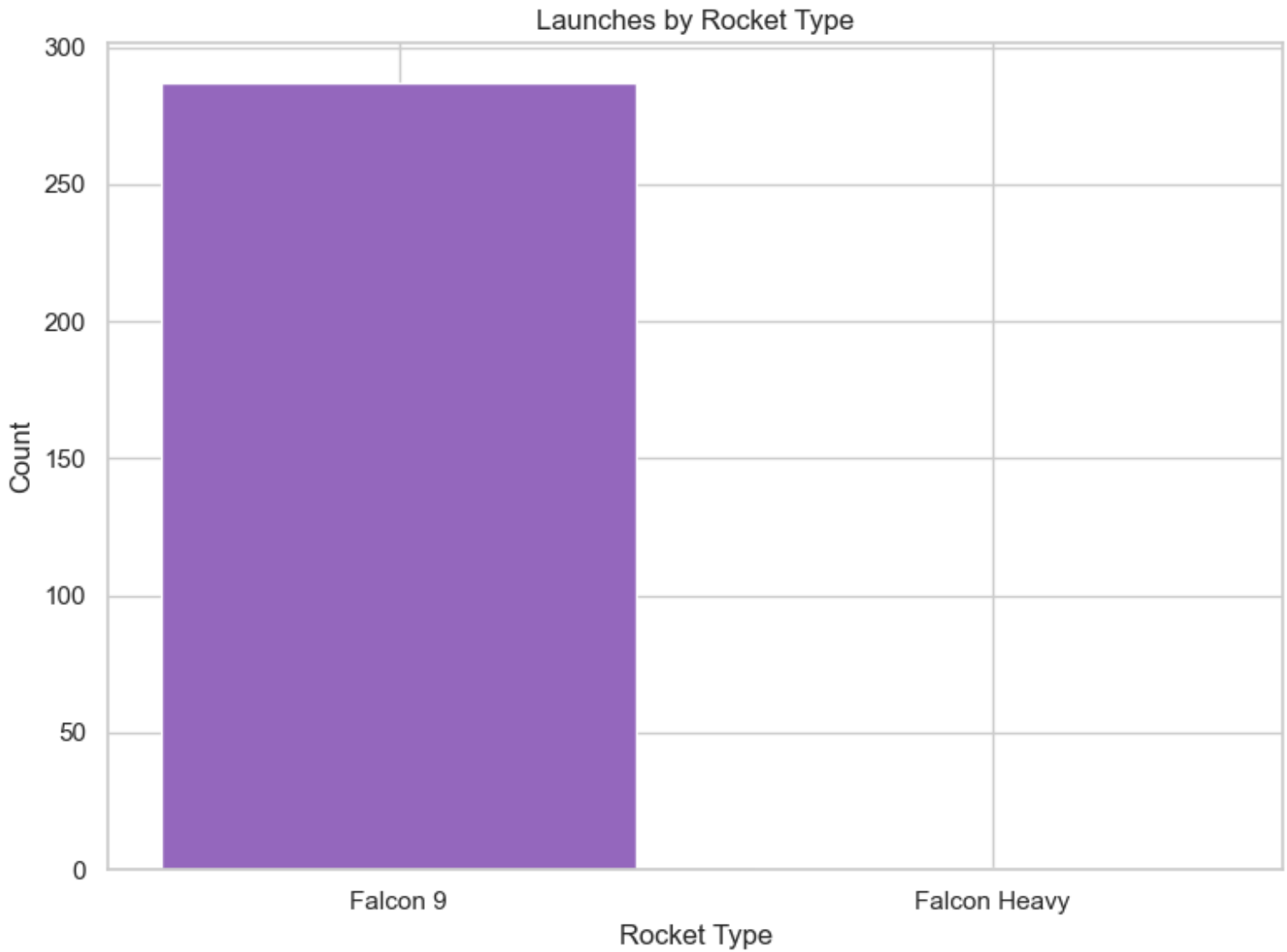

Payload Mass vs. Launch Date
Payload Mass (kg)
17500
15000
12500
10000
7500
5000
2500

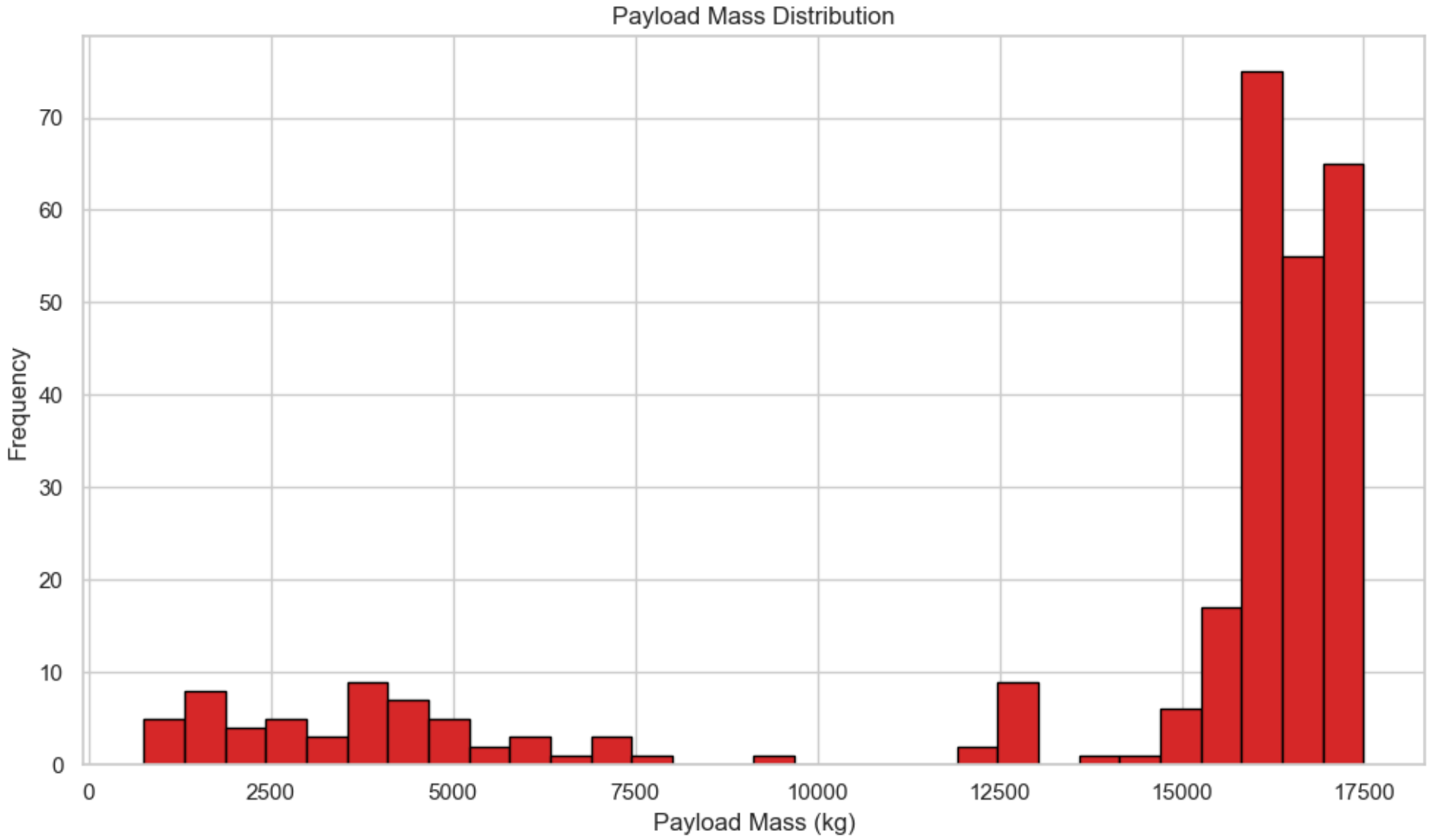

Launches by Year
Number of Launches
0
20
40
60
80
100
120
2022.5
2023.0
2023.5
2024.0
2024.5
2025.0
2025.5
Year

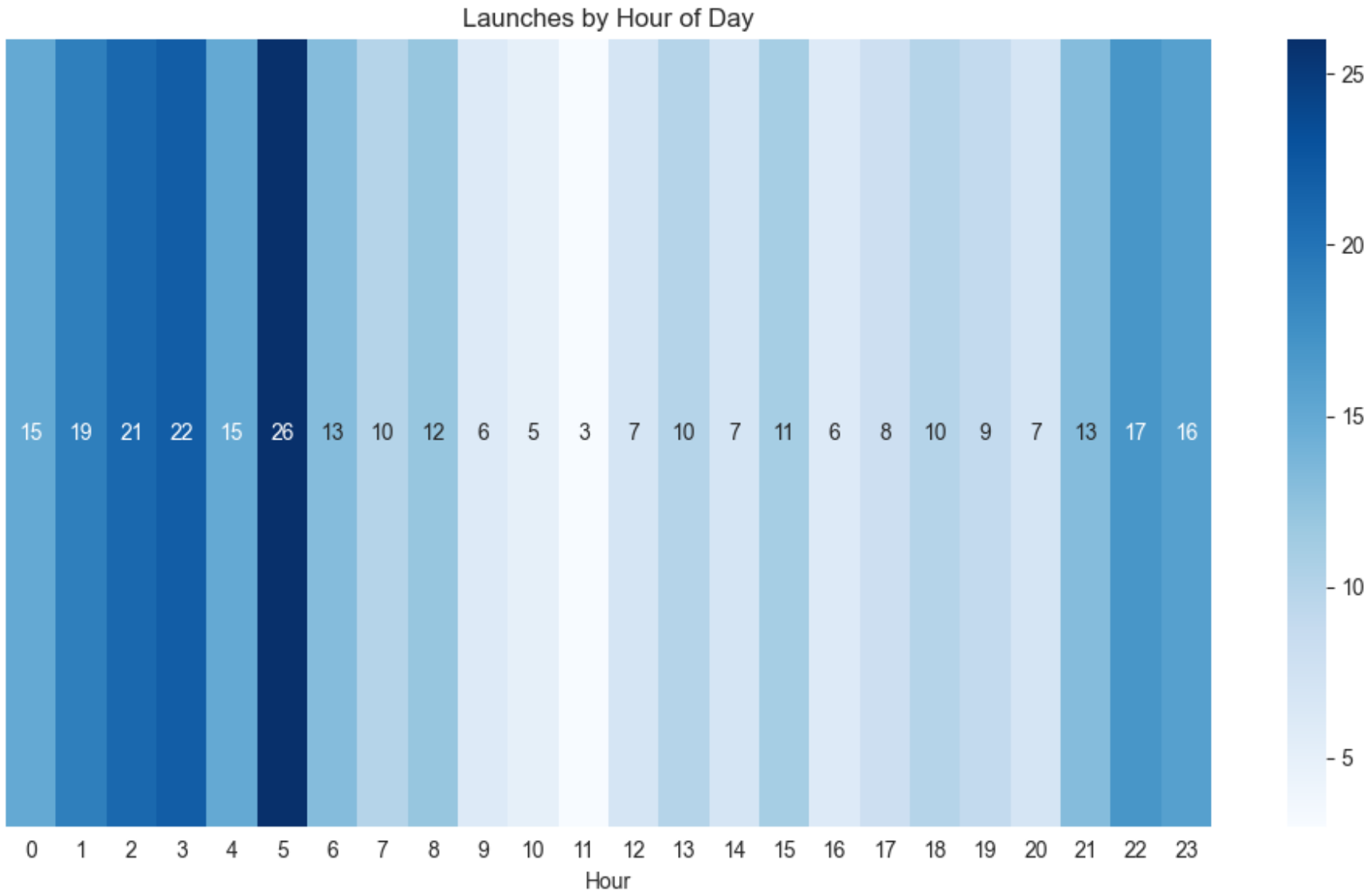

Payload Mass by Orbit
Payload Mass (kg)
17500
15000
12500
10000
7500
5000
2500
0
Sso
Polarleo
Geo
Meo
Leo
Gto
Leo(Iss)
Sun–Earth L2Injection
Tli
Molniya
Heliocentric
Polar(Retrograde)
Orbit Type


Launches by Launch Site
Cape Canaveral,Slc-40
51.0%
16.3%
Kennedy,Lc-39A
32.6%
Vandenberg,Slc-4E

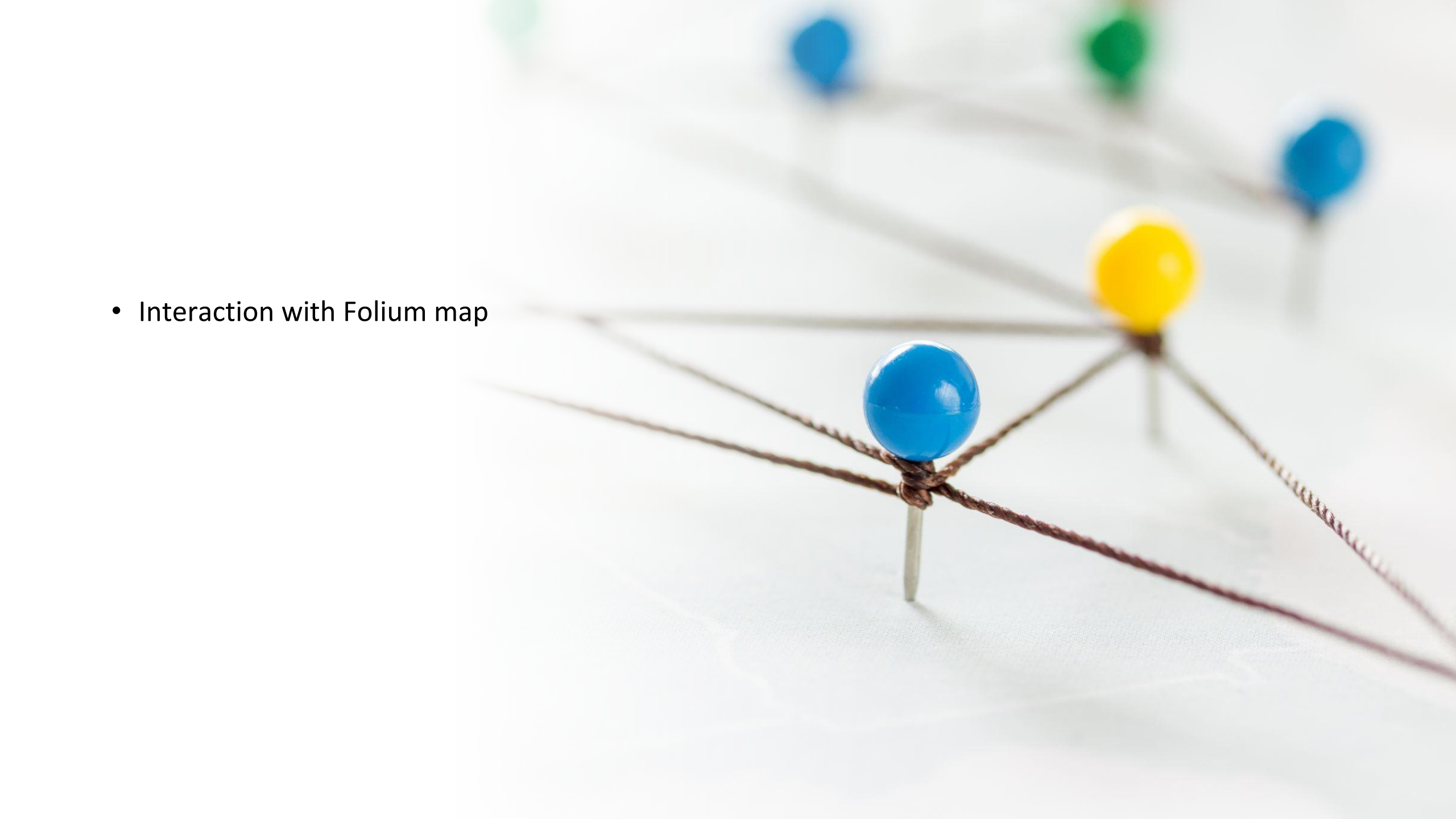

Launches by Customer Type
Internal
69.1%
9.4%
Government
21.5%
Commercial

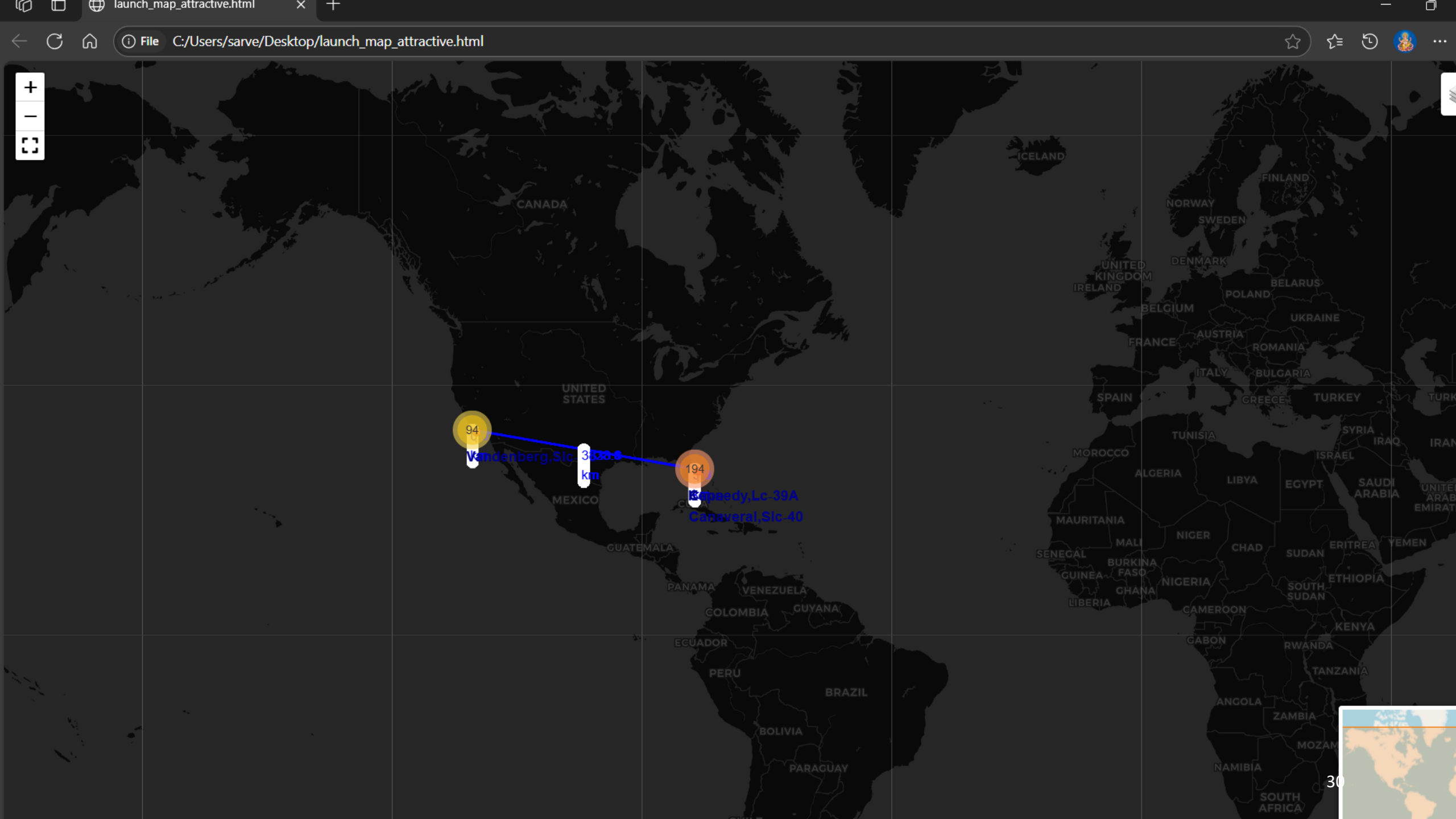

Launches by Rocket Type
300
250
200
150
100
50
0
Count
Falcon 9
Falcon Heavy
Rocket Type

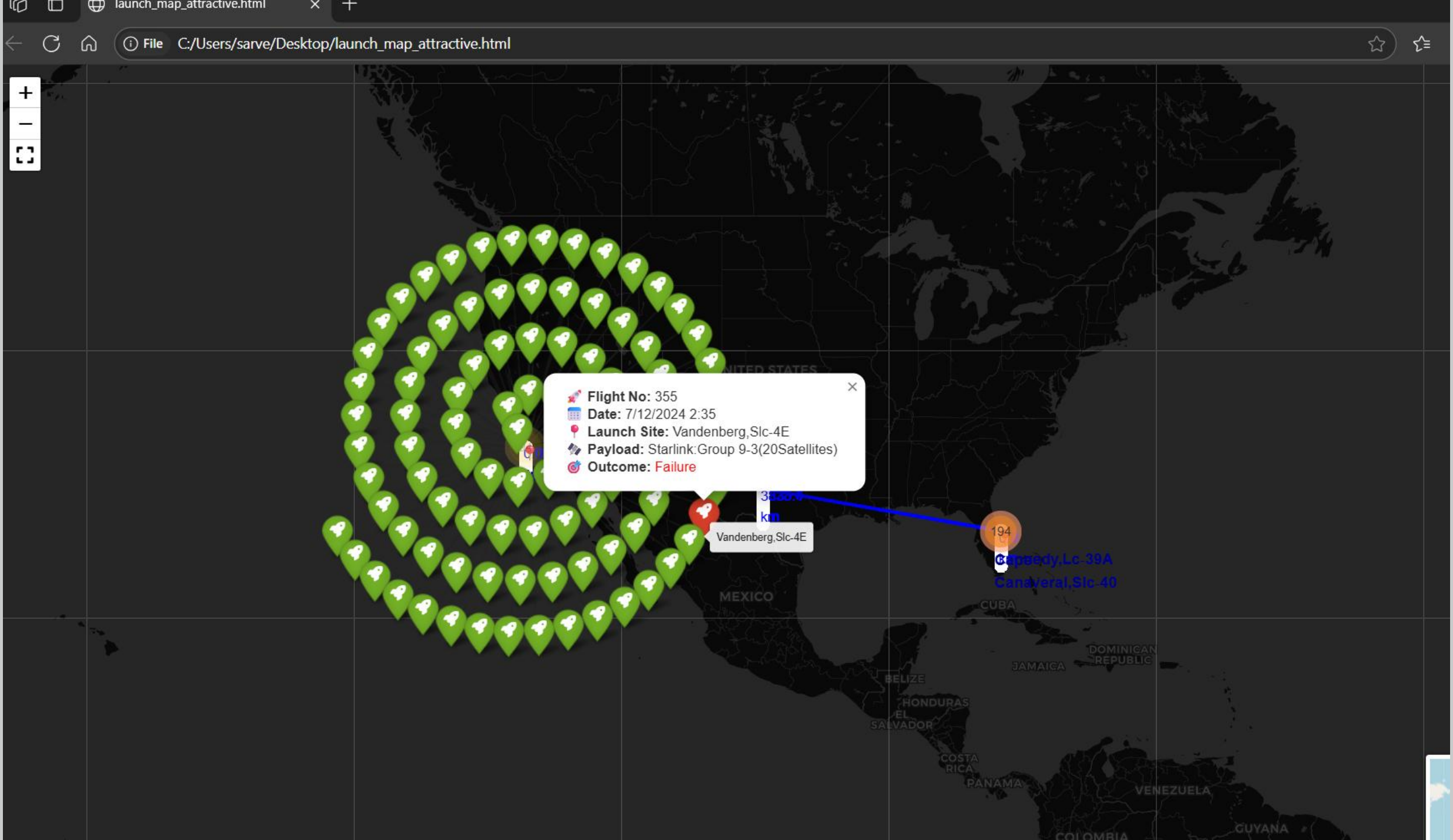

Payload Mass Distribution
Frequency
0
10
20
30
40
50
60
70
0
2500
5000
7500
10000
12500
15000
17500
Payload Mass (kg)

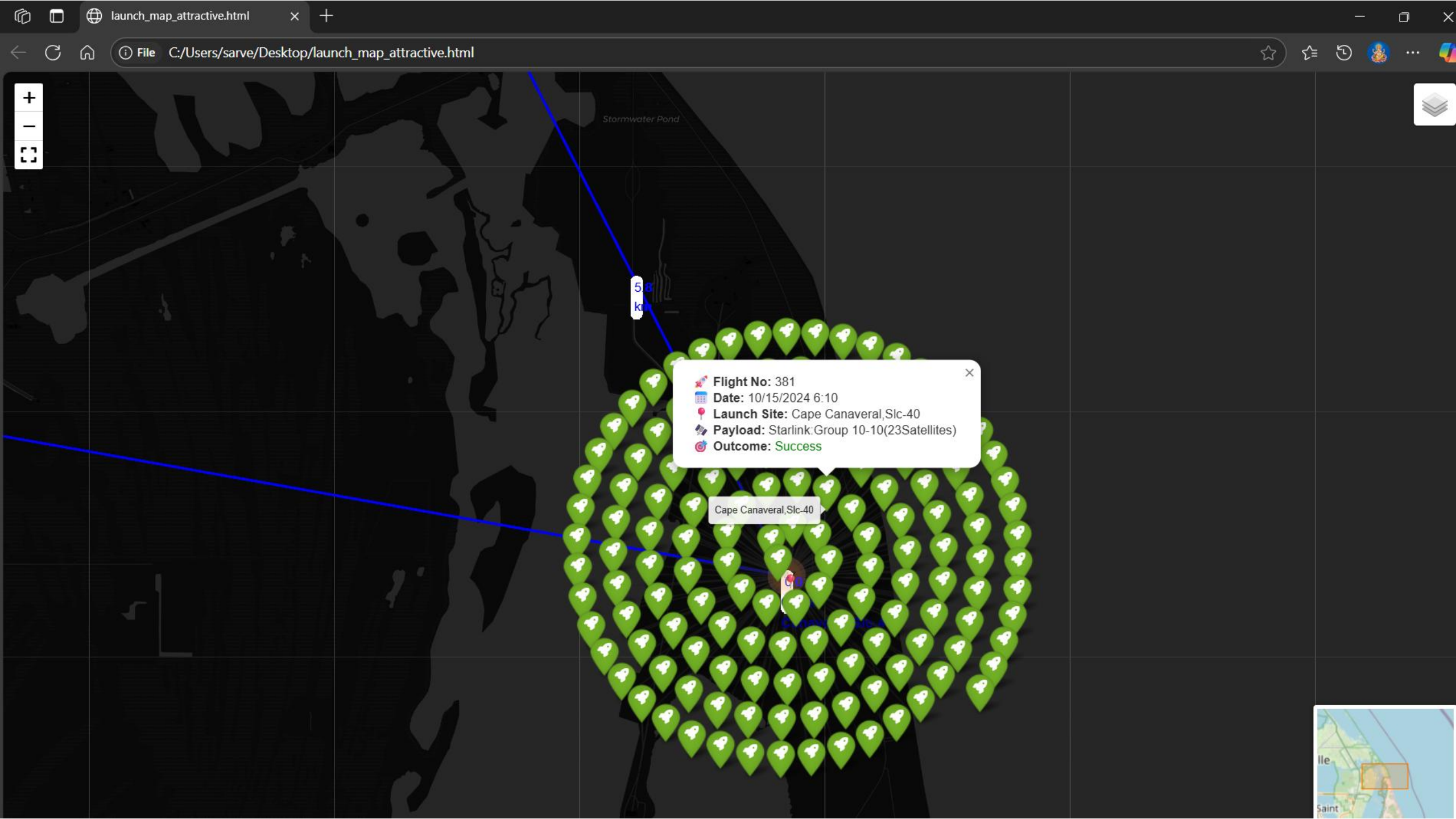

Launches by Hour of Day
15 19 21 22 15 26 13 10 12 6 5 3 7 10 7 11 6 8 10 9 7 13 17 16
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour
25
20
15
10
5

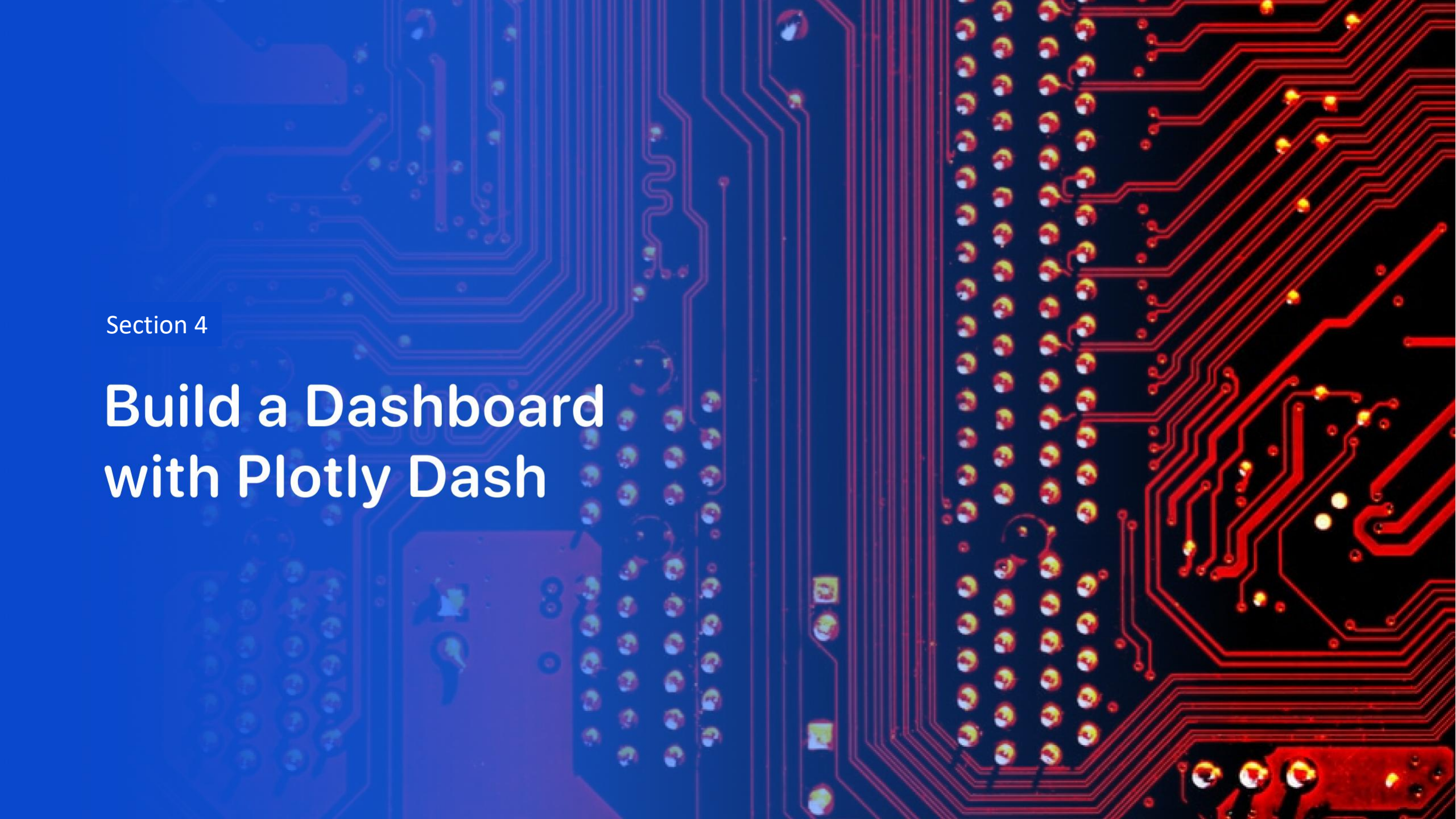
- Interaction with Folium map

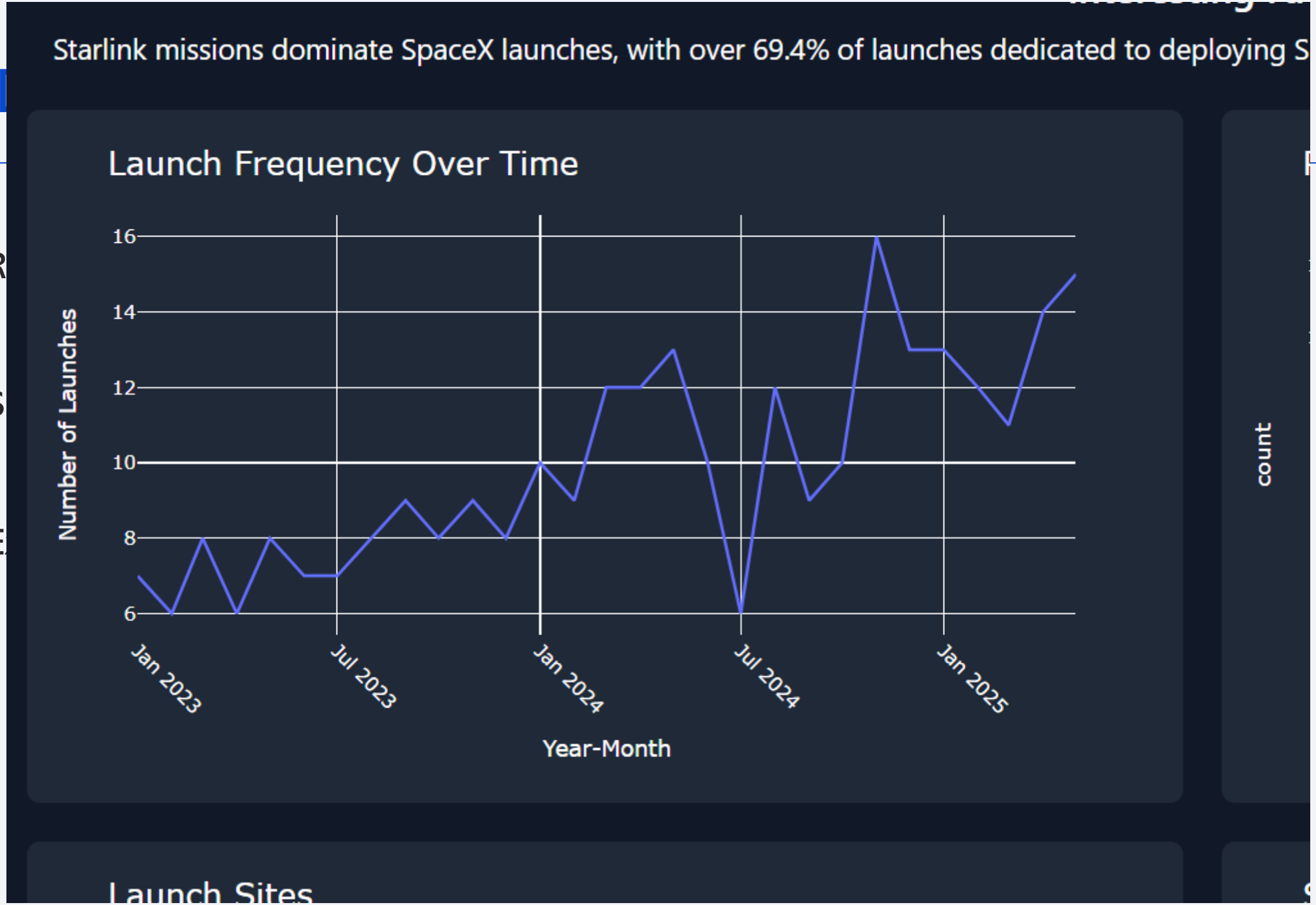


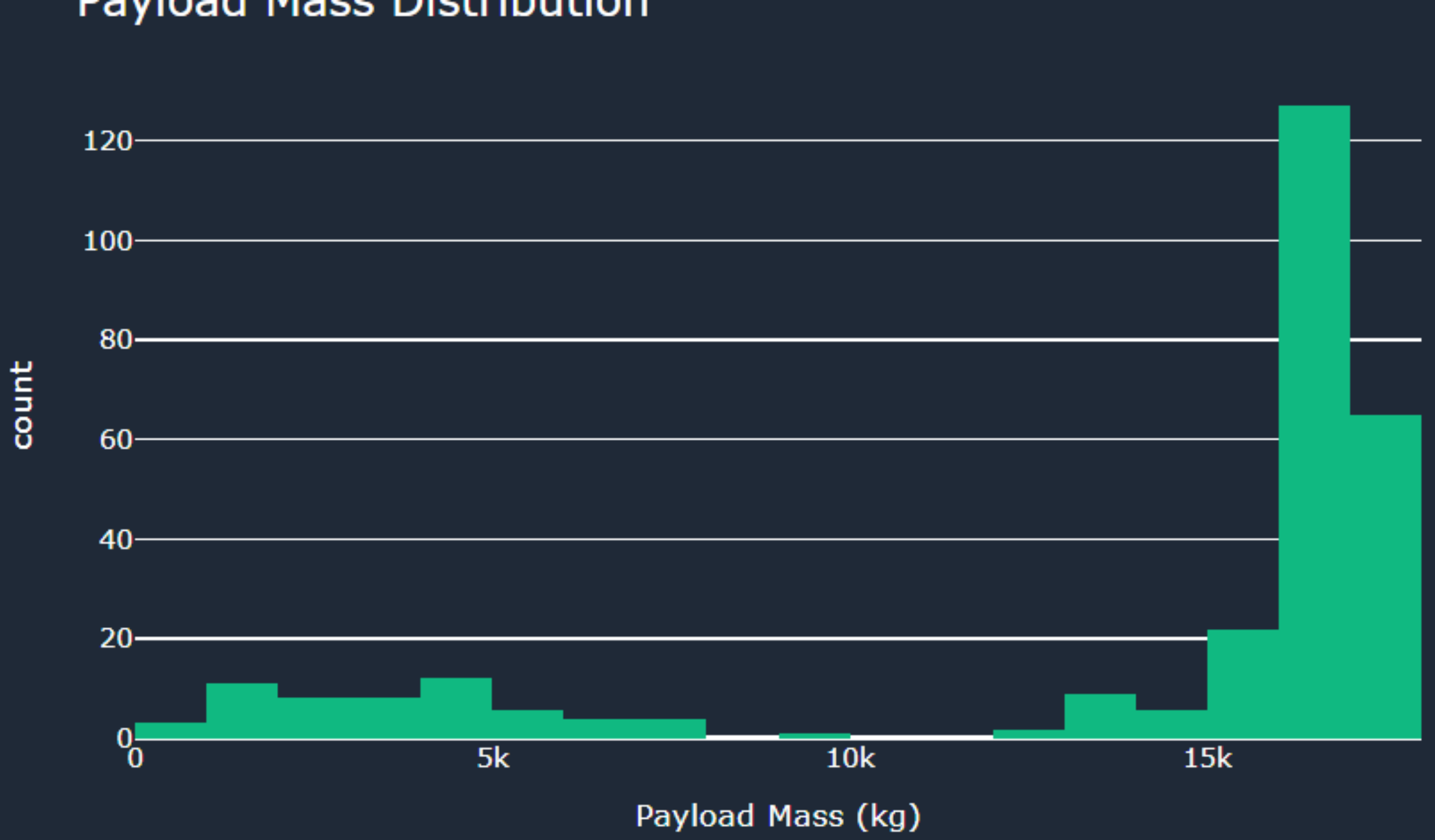

launch_map_attractive.html
File
C:/Users/sarve/Desktop/launch_map_attractive.html
94
Vandenberg,Sic-3
km
194
Kennedy,Lc-39A
Canaveral,Sic-40
CANADA
UNITED STATES
MEXICO
GUATEMALA
PANAMA
COLOMBIA
VENEZUELA
GUYANA
ECUADOR
PERU
BRAZIL
BOLIVIA
PARAGUAY
ICELAND
FINLAND
NORWAY
SWEDEN
DENMARK
UNITED KINGDOM
IRELAND
BELGIUM
POLAND
BELARUS
UKRAINE
FRANCE
AUSTRIA
ROMANIA
ITALY
BULGARIA
SPAIN
GREECE
TURKEY
TUNISIA
SYRIA
IRAQ
MOROCCO
ISRAEL
ALGERIA
LIBYA
EGYPT
SAUDI ARABIA
MAURITANIA
MALI
NIGER
CHAD
SUDAN
ERITREA
YEMEN
SENEGAL
BURKINA FASO
GUINEA
GHANA
NIGERIA
ETHIOPIA
LIBERIA
SOUTH SUDAN
CAMEROON
KENYA
GABON
RWANDA
TANZANIA
ANGOLA
ZAMBIA
NAMIBIA
SOUTH AFRICA


launch_map_attractive.html
File C:/Users/sarve/Desktop/launch_map_attractive.html
Flight No: 355
Date: 7/12/2024 2:35
Launch Site: Vandenberg,Slc-4E
Payload: Starlink:Group 9-3(20Satellites)
Outcome: Failure
Vandenberg,Slc-4E
194
Canaveral,Slc-40
MEXICO
CUBA
JAMAICA
BELIZE
HONDURAS
EL SALVADOR
COSTA RICA
PANAMA
VENEZUELA
GUYANA

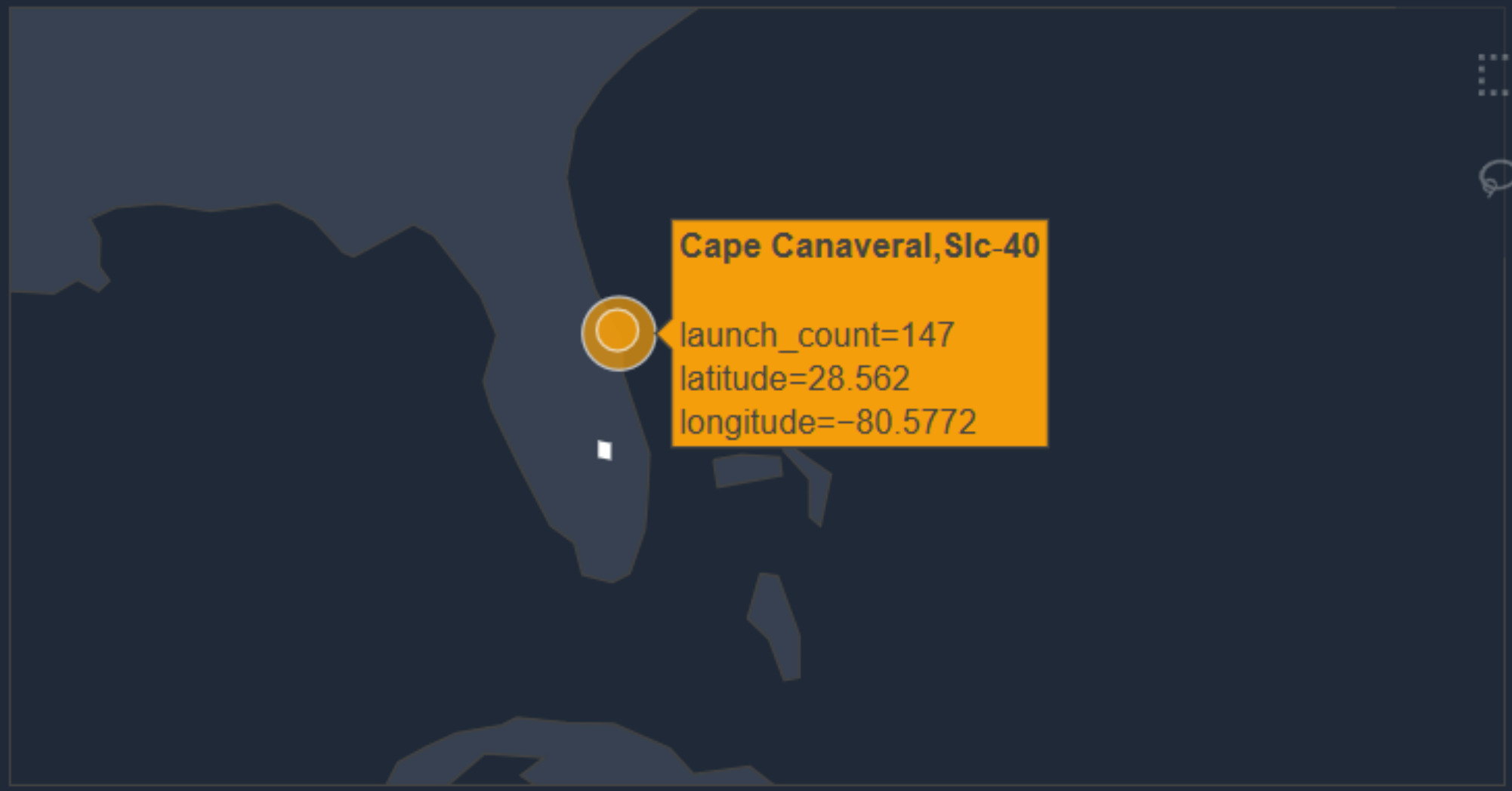

launch_map_attractive.html
File C:/Users/sarve/Desktop/launch_map_attractive.html
Stormwater Pond
Flight No: 381
Date: 10/15/2024 6:10
Launch Site: Cape Canaveral,Slc-40
Payload: Starlink:Group 10-10(23Satellites)
Outcome: Success
Cape Canaveral,Slc-40


Section 4
Build a Dashboard with Plotly Dash


Starlink missions dominate SpaceX launches, with over 69.4% of launches dedicated to deploying S
Launch Frequency Over Time
16
14
12
10
8
6
Number of Launches
Jan 2023
Jul 2023
Jan 2024
Jul 2024
Jan 2025
Year-Month
count
Launch Sites

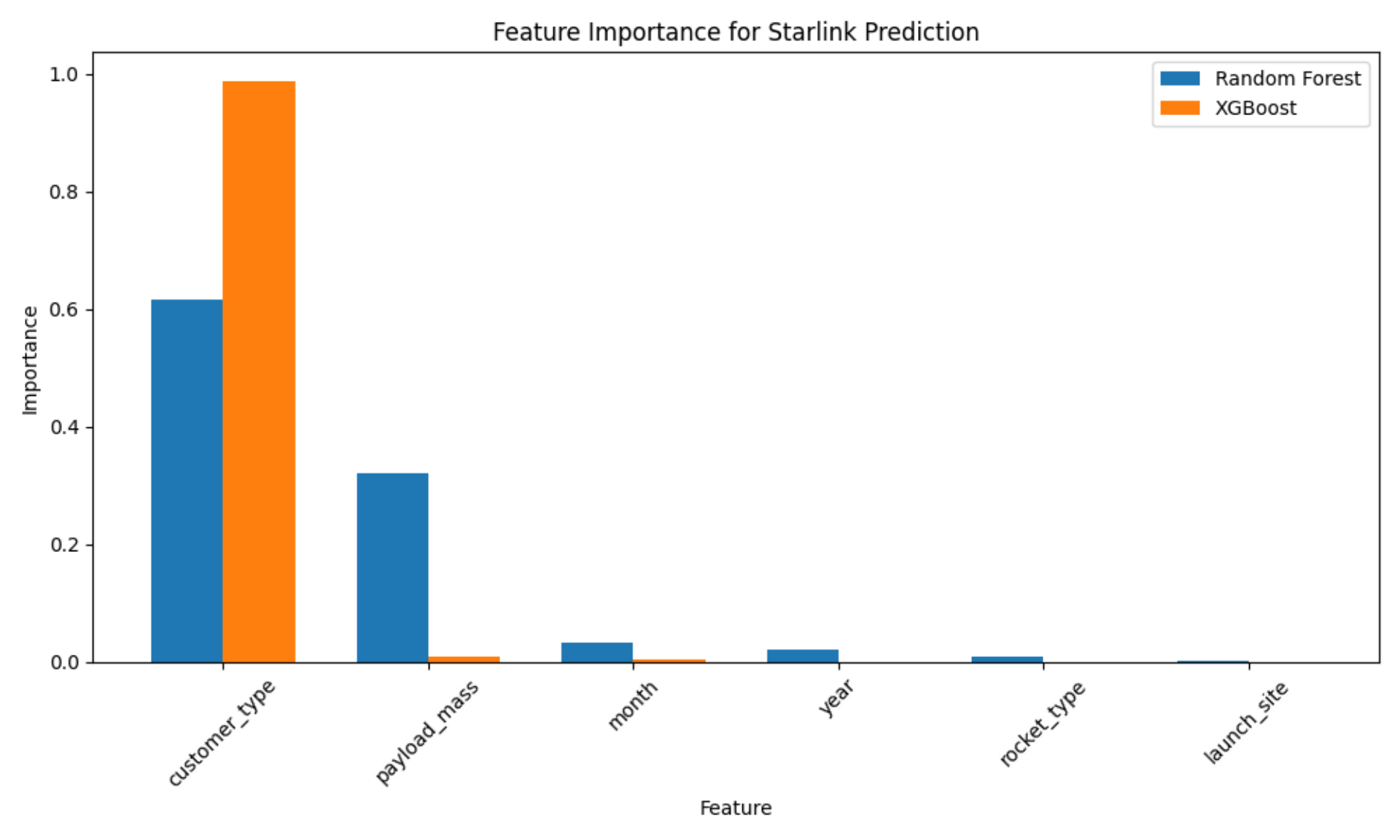

Payload Mass Distribution
120
100
80
60
40
20
0
count
0
5k
10k
15k
Payload Mass (kg)

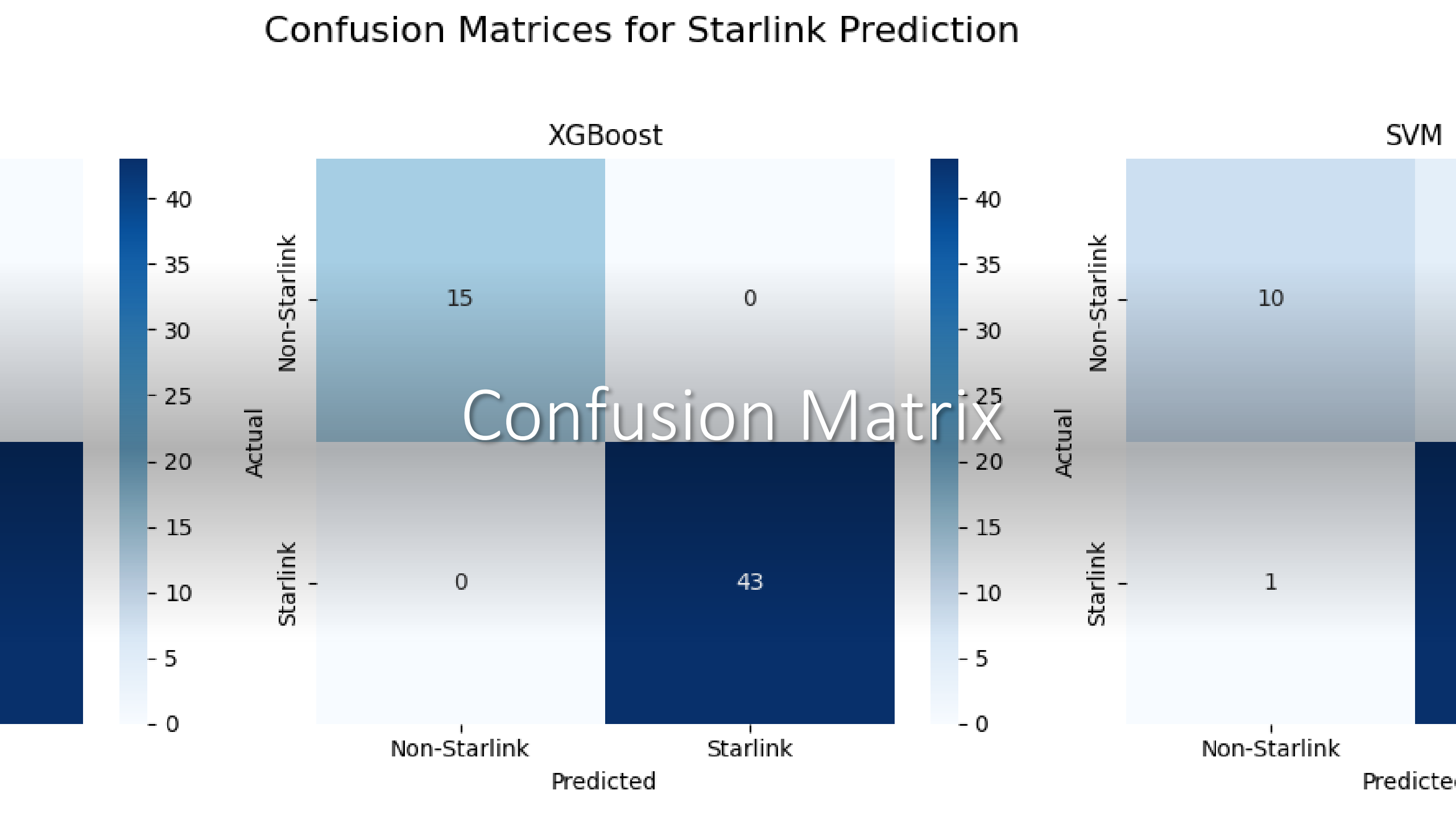
Launch Sites

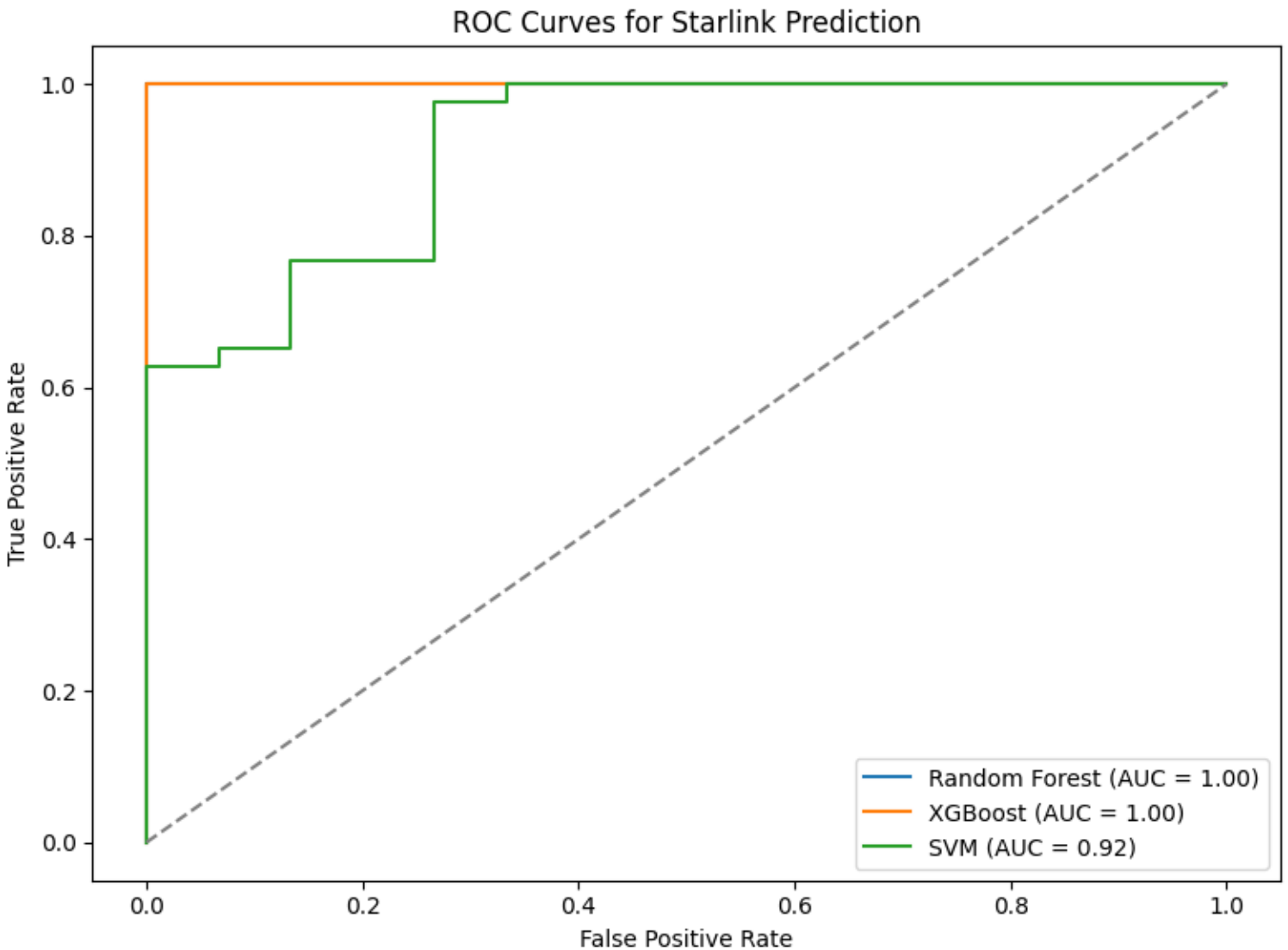


Section 5

Predictive Analysis (Classification)


Feature Importance for Starlink Prediction
Random Forest
XGBoost
Importance
1.0
0.8
0.6
0.4
0.2
0.0
customer_type
payload_mass
month
year
rocket_type
launch_site
Feature


Confusion Matrices for Starlink Prediction
XGBoost
SVM
Confusion Matrix
Actual
Non-Starlink
Starlink
15
0
0
43
10
1
Non-Starlink
Starlink
Predicted
40
35
30
25
20
15
10
5
0


ROC Curves for Starlink Prediction
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 1.00)
XGBoost (AUC = 1.00)
SVM (AUC = 0.92)

Conclusions

- Exceptional Reliability & Growth in Operations: The analysis consistently shows SpaceX achieving remarkable launch success rates and highly effective booster reusability, all while significantly increasing its launch frequency over time.
- Strategic Efficiency and Key Focus Areas: The data highlights a clear operational strategy, with a strong focus on launches from key sites like Cape Canaveral and a predominant targeting of Low Earth Orbit (LEO) missions.
- Actionable Insights for Future Optimization: Through comprehensive EDA, insightful visualizations, and predictive machine learning, critical drivers of mission success were identified, providing valuable, data-driven insights for continuous operational improvement.

Thank you!