

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.

TOM	729.89	915.51	185.62	▲ 25.43%
HUM	749.73	924.29	174.56	▲ 23.28%
DMW	833.72	1004.01	170.29	▲ 20.43%
YZJ	903.49	1127.46	223.97	▲ 24.79%
GLY	982.07	1219.39	237.32	▲ 24.17%
VDA	113.74	143.41	29.67	▲ 26.09%
UVV	468.08	535.41	67.33	▲ 14.38%
HJS	545.49	659.05	113.56	▲ 20.82%
EDC	545.49	659.05	113.56	▲ 20.82%

FLR	660.27	745.28	85.01	▲ 12.88%
UVQ	155.59	181.57	25.98	▲ 16.70%
QUV	440.55	540.21	99.66	▲ 22.62%
HZT	285.51	344.98	59.47	▲ 20.83%
PCW	811.44	1029.66	125.73	▲ 23.78%
AIK	361.77	451.39	89.62	▲ 24.77%
ZJJ	858.36	994.57	136.21	▲ 14.11%
RHJ	894.79	1049.68	151.89	▲ 14.71%
VOY	545.49	659.05	113.56	▲ 20.82%

- 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?**

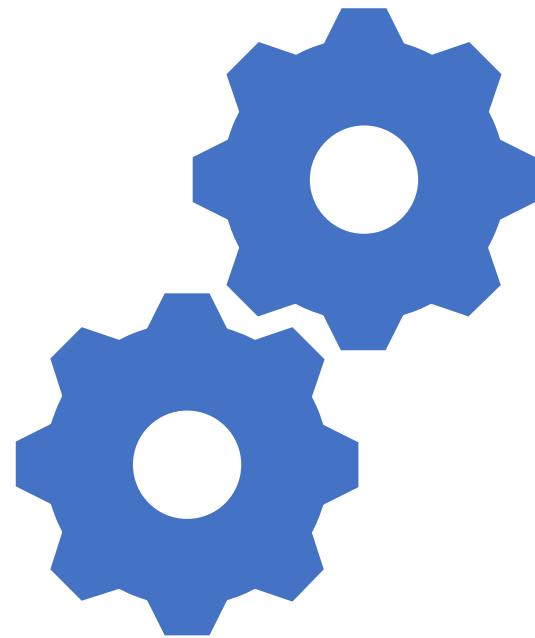
Using historical mission data, can we reliably predict if a future mission will successfully land?

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

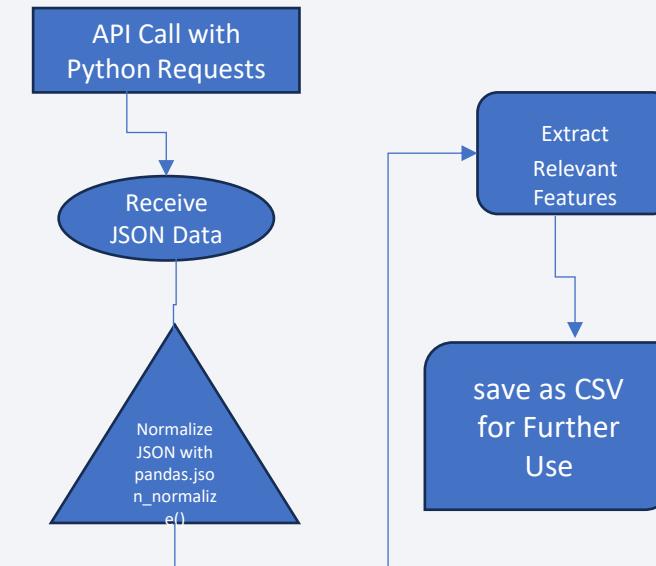
Data Collection

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`)



1. importing the modules required for data collection

In [4]:

```
1 import json  
2 import pandas as pd  
3
```

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

Data Collection – SpaceX API

In [5]:

```
1 with open('launches.json', 'r') as file:  
2     data = json.load(file)
```

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

In [6]:

```
1 df = pd.json_normalize(data)
```

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

In [7]:

```
1 df.to_csv("spacex_launches_raw.csv", index=False)
```

1.Importing required libraries for data collection -scraping

```
In [2]: 1 import requests  
2 from bs4 import BeautifulSoup  
3 import pandas as pd  
4 import re
```

```
In [3]: 1 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]: 1 # Send a GET request to fetch the page content  
2 response = requests.get(url)  
3 response.raise_for_status() # Raise an error for bad status codes
```

3 .Parse the HTML content using BeautifulSoup

```
In [5]: 1 # Parse the HTML content using BeautifulSoup  
2 soup = BeautifulSoup(response.text, 'html.parser')
```

• Data Collection - Scraping

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

```
In [16]: 1 # Save the DataFrame to a CSV file  
2 df.to_csv('falcon_launches.csv', index=False)
```

4. Find all tables with class 'wikitable'

```
In [6]: 1 # Find all tables with class 'wikitable'  
2 tables = soup.find_all('table', class_='wikitable')
```

Initialize an empty list to store launch data

```
In [7]: 1 # Initialize an empty list to store launch data  
2 launch_data = []
```

Define column headers based on typical table structure

```
In [8]: 1 # Define column headers based on typical table structure  
2 headers = [  
3     'Flight No.', 'Date and time (UTC)', 'Version, Booster', 'Launch site',  
4     'Payload', 'Payload mass', 'Orbit', 'Customer',  
5     'Launch outcome', 'Booster landing'  
6 ]  
7
```

Data Wrangling

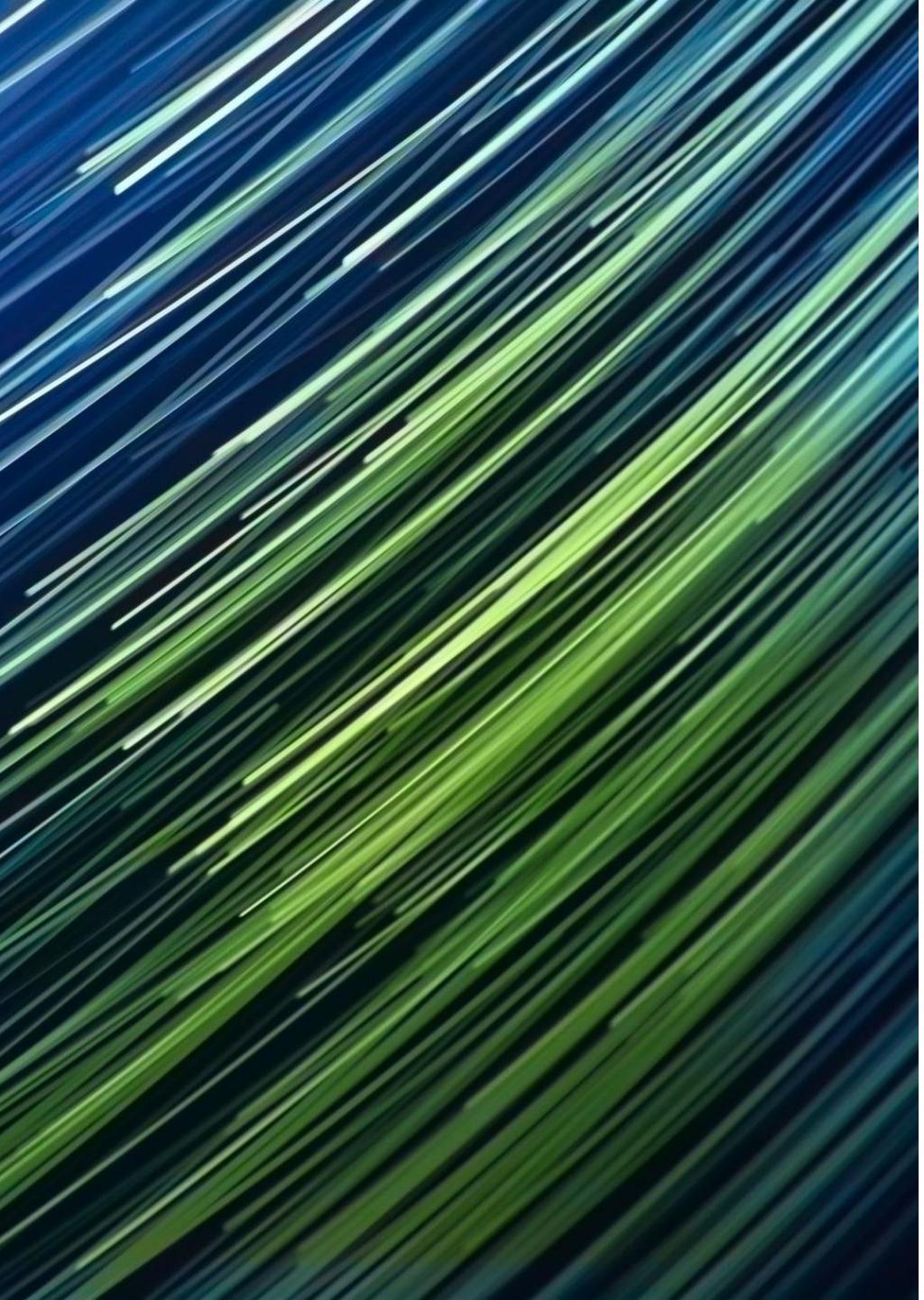
Handling Missing Values:
Addressed incomplete records (e.g., dropped Unnamed: 0 column, ensured critical columns were complete).

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 visualization



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



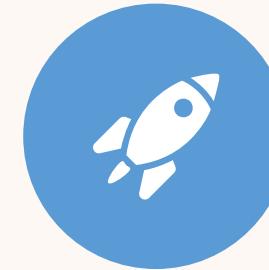
EFFICIENT DATA STRUCTURE OVERVIEW: SQL QUERIES, SUCH AS PRAGMA TABLE_INFO(), ALLOWED FOR A RAPID UNDERSTANDING OF THE DATASET'S SCHEMA, INCLUDING COLUMN NAMES, DATA TYPES, AND CONSTRAINTS. THIS FOUNDATIONAL STEP IS CRUCIAL FOR PLANNING FURTHER ANALYSIS.



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)

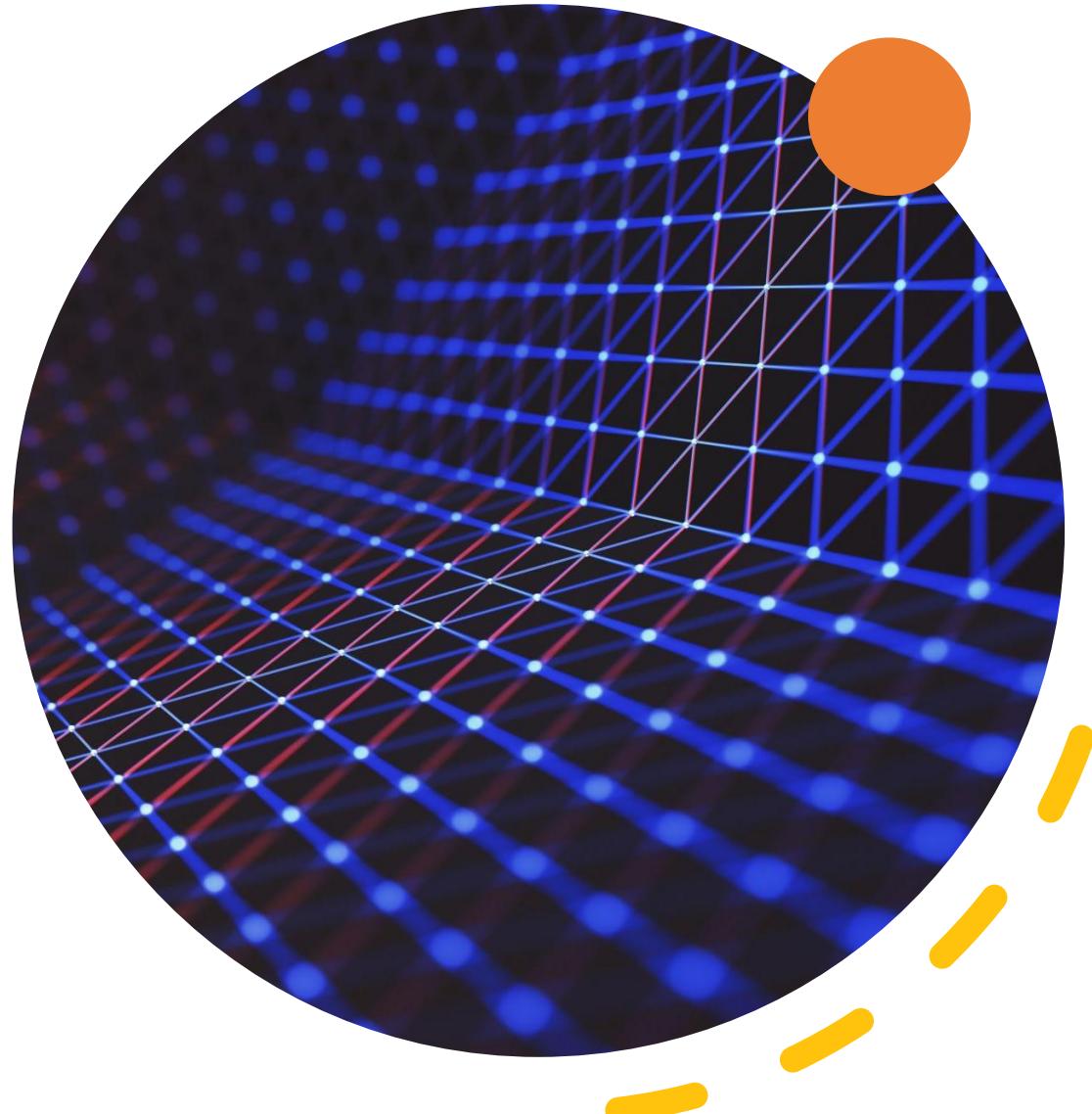
Predicting Mission Success: Developed ML models to forecast launch outcomes, aiding proactive planning.

Comparative Model Analysis: Evaluated multiple ML algorithms to identify the most accurate predictors.

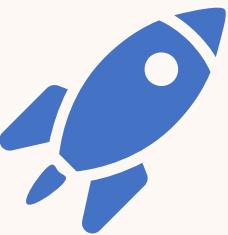
Identifying Key Success Drivers: Used advanced techniques (e.g., SHAP) to pinpoint critical influencing factors.

Rigorous Model Validation: Ensured model reliability through comprehensive evaluation and visualizations.

Actionable Operational Insights: Derived data-driven recommendations to enhance future mission success.



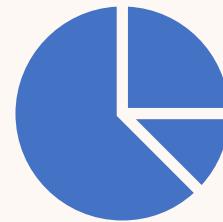
Results



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.

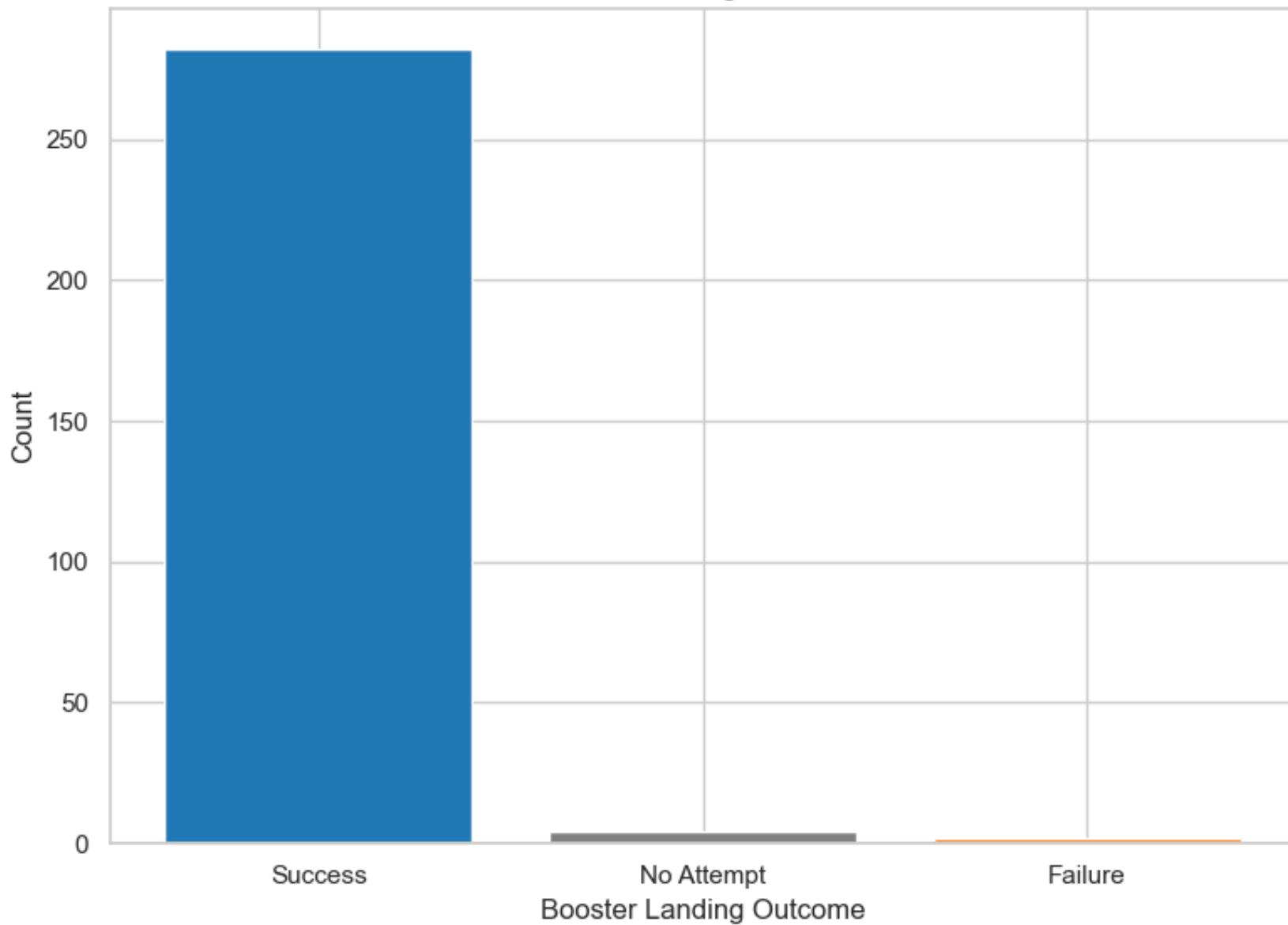


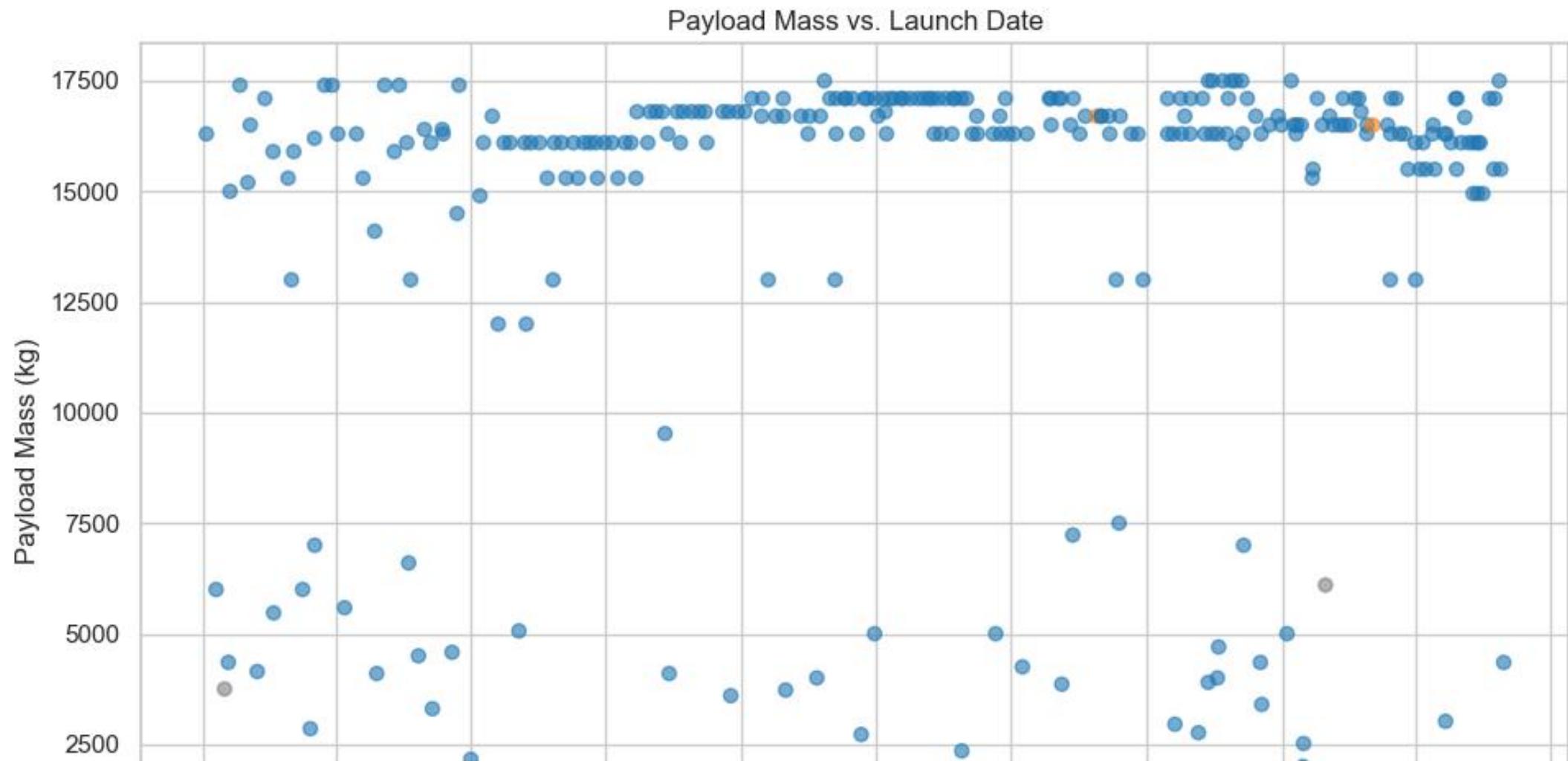
Actionable Insights for Future Optimization:
Through comprehensive EDA, insightful visualizations, and predictive machine learning, critical drivers of mission success were identified, providing valuable, data-driven insights for continuous operational improvement.

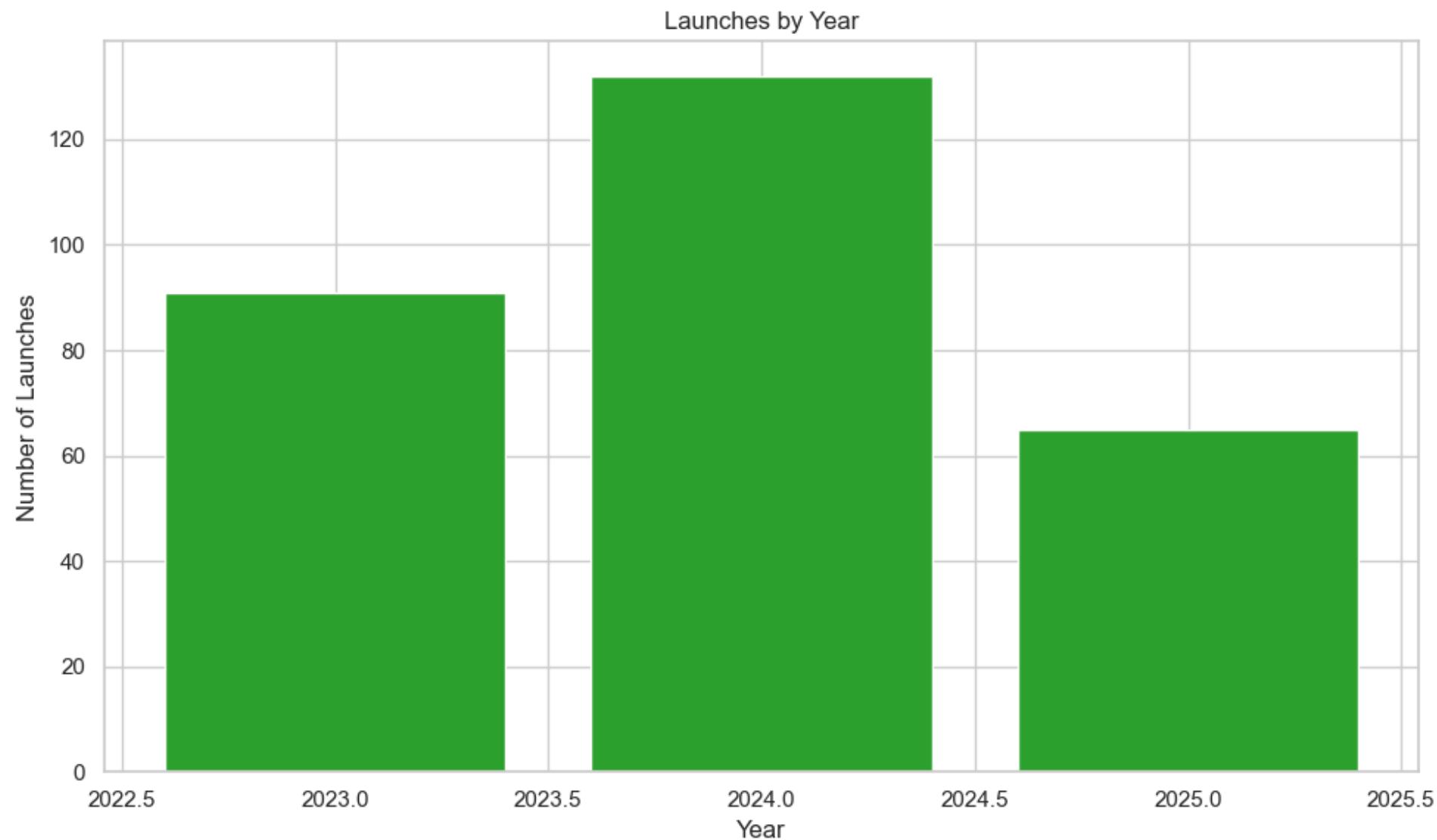


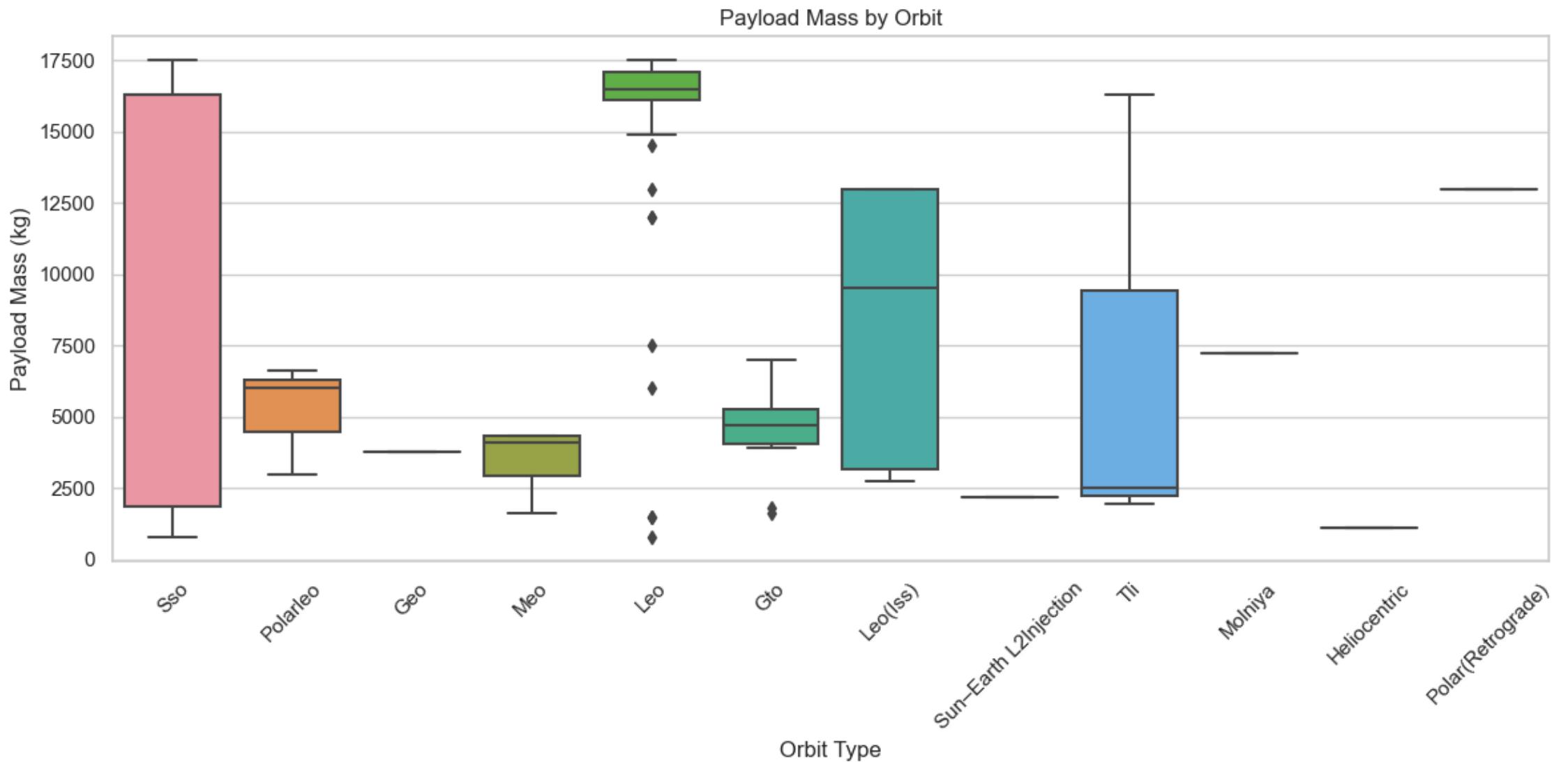
EDA VISUALIZATION WITH PYTHON VISUALIZATION LIBRARIES

Booster Landing Outcomes

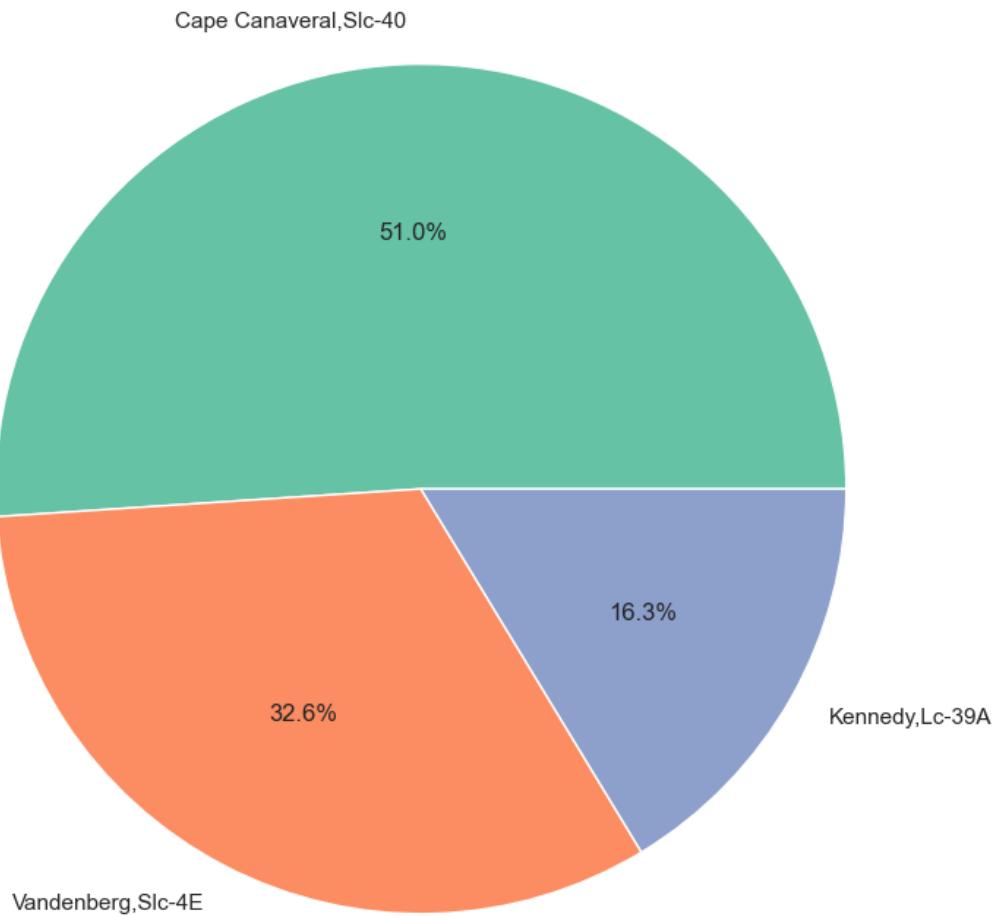




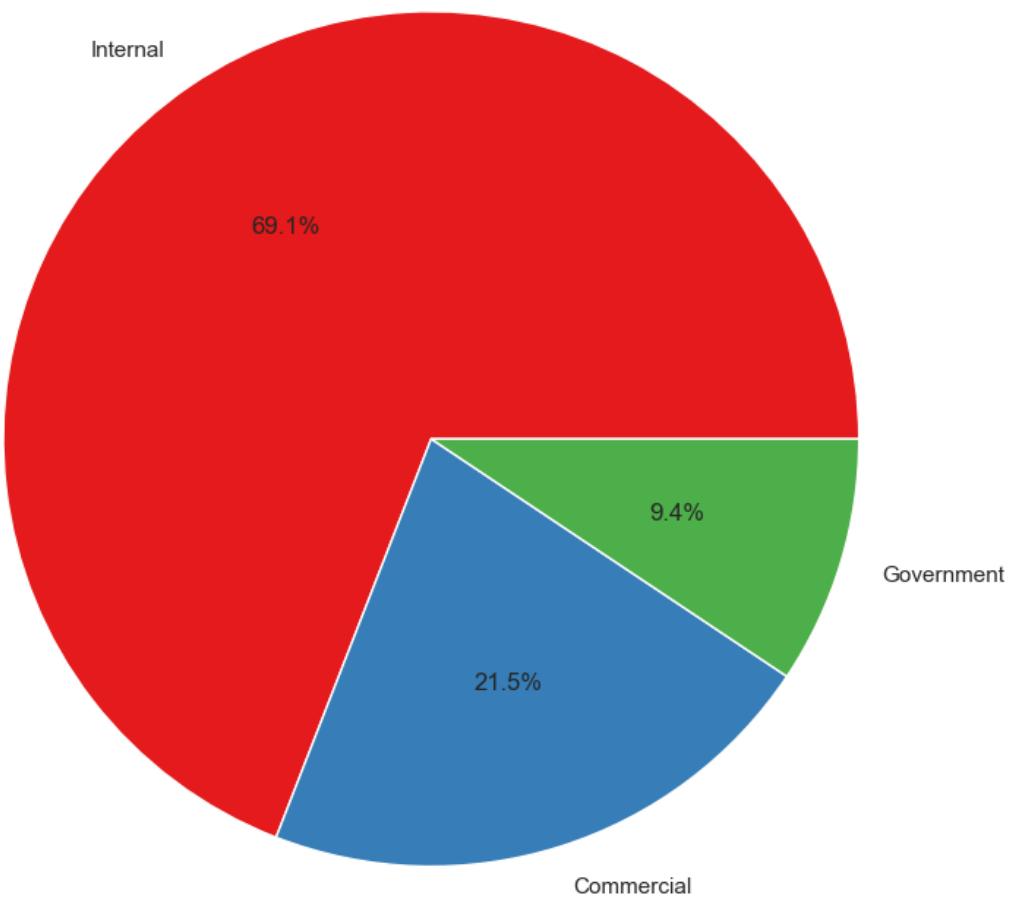




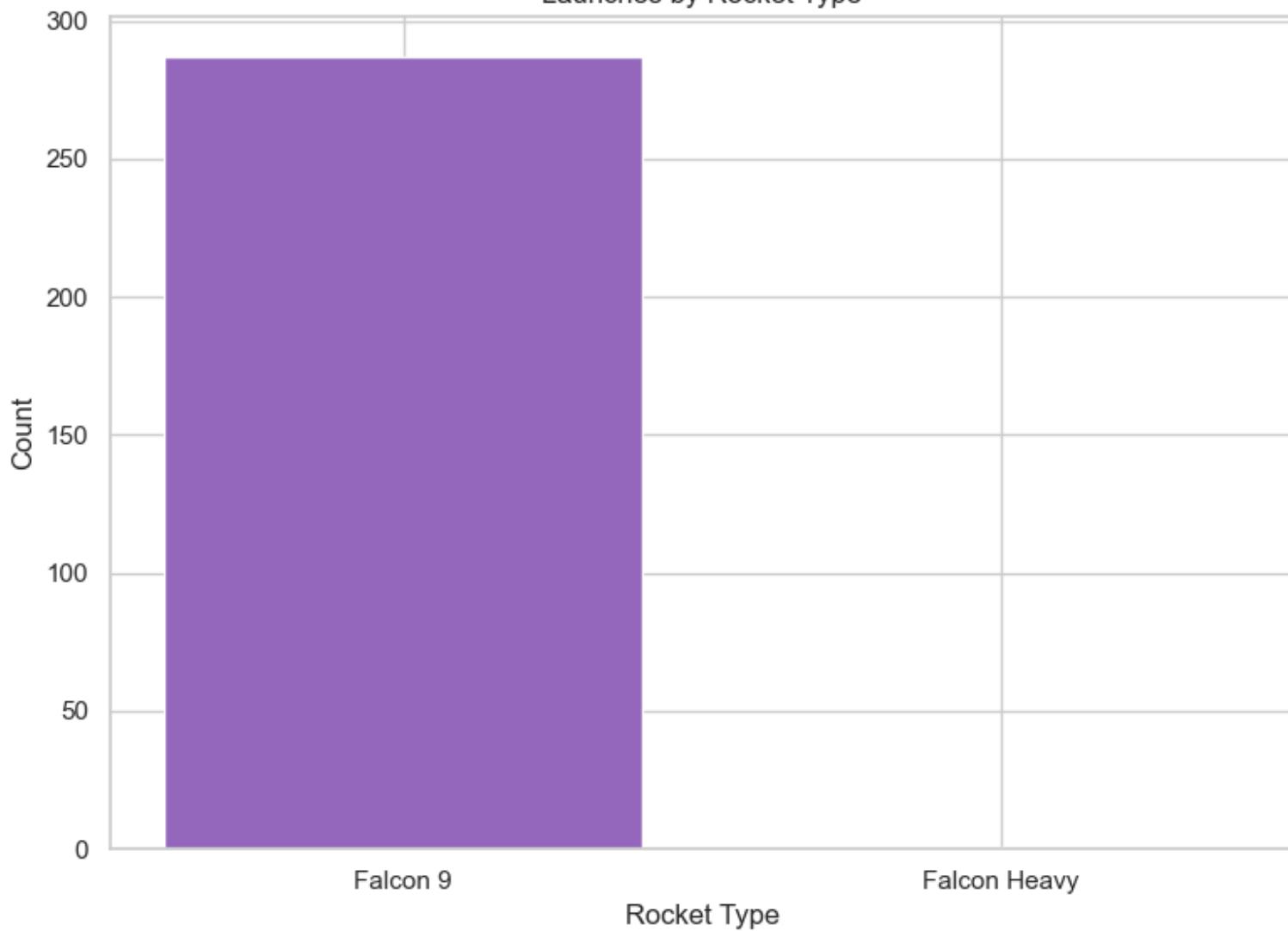
Launches by Launch Site



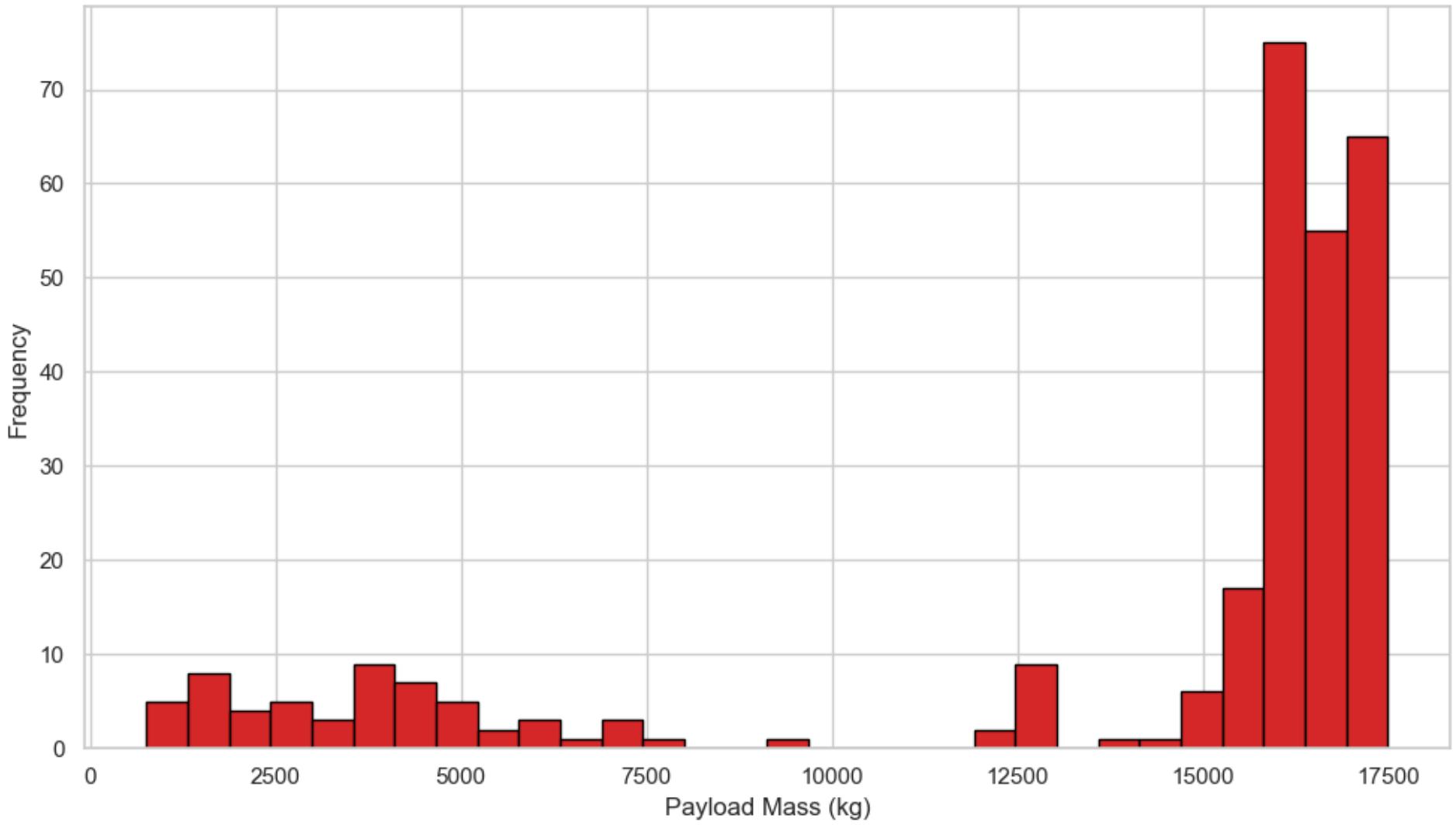
Launches by Customer Type



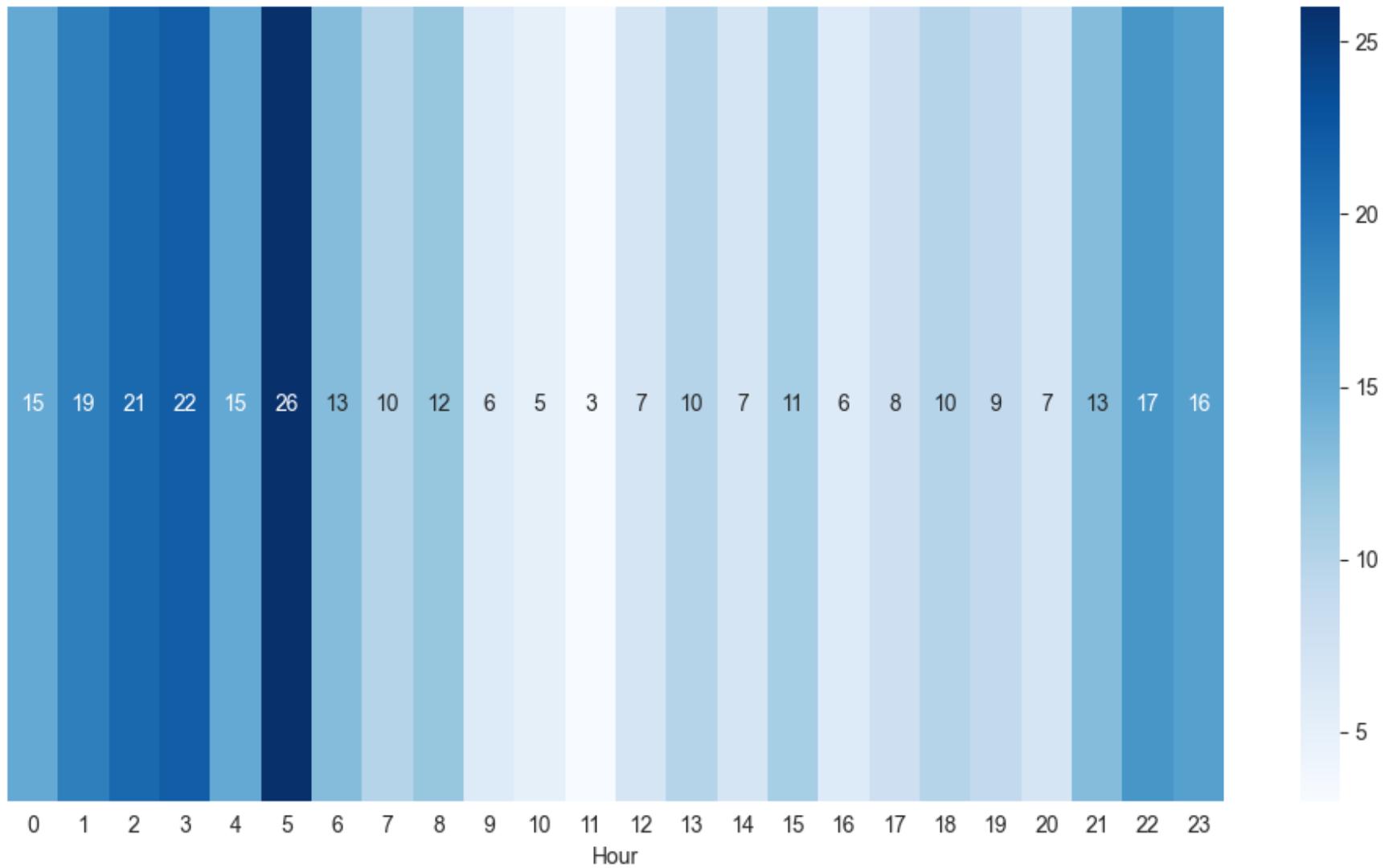
Launches by Rocket Type



Payload Mass Distribution



Launches by Hour of Day





EDA WITH SQL

```
11 | print(yearly_launches.to_string(index=False))
```

Launches per Year:

year	launch_count
2023	91
2024	132
2025	65

Launches by Launch site:

	launch_site	launch_count
Cape Canaveral,Slc-40		147
Vandenberg,Slc-4E		94
Kennedy,Lc-39A		47

Payload Mass Summary:

avg_payload_mass	min_payload_mass	max_payload_mass
13830.19	758.0	17500.0

Launches by Orbit Type:

orbit	launch_count
Leo	216
Gto	19
Sso	19
Leo(Iss)	15
Meo	8
Polarleo	3
Tli	3
Geo	1
Heliocentric	1
Molniya	1
Polar(Retrograde)	1
Sun-Earth L2Injection	1

Launches by Customer Type:

customer_type	launch_count
Internal	199
Commercial	62
Government	27

Success Rates (%):

launch_success_rate	booster_landing_success_rate
99.65	97.92

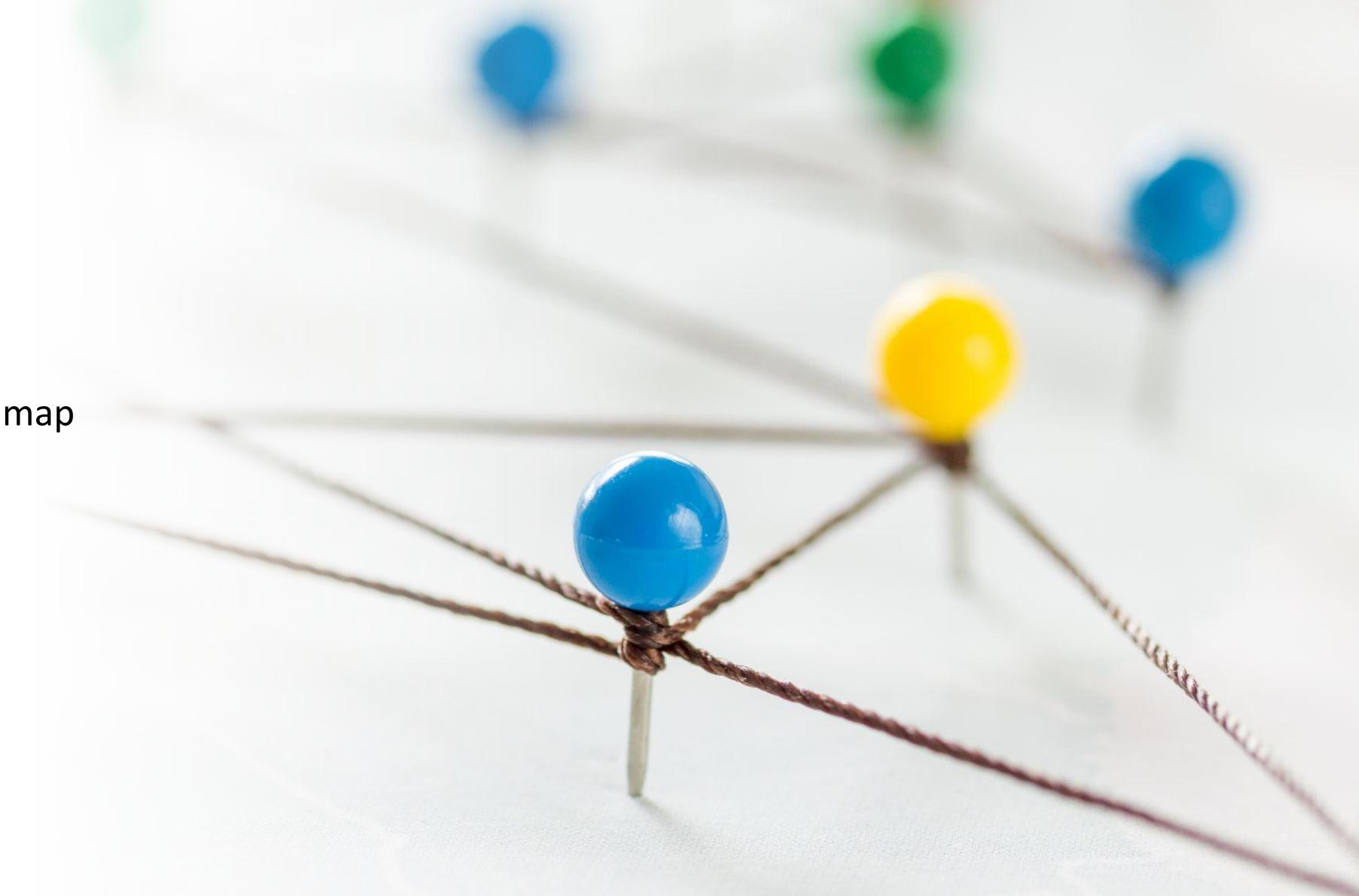
Launches by Launch Site with Coordinates:

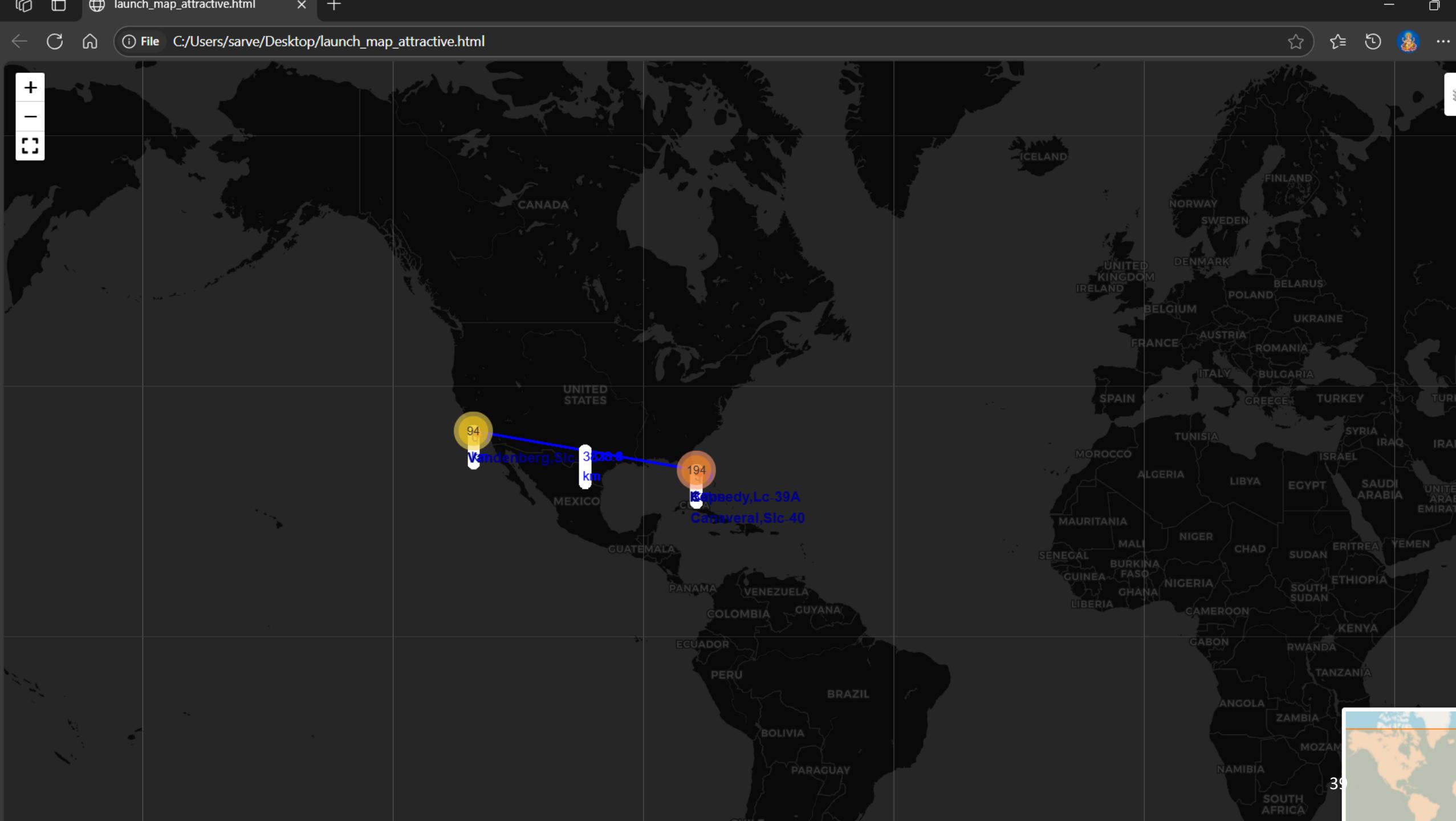
	launch_site	latitude	longitude	launch_count
Cape Canaveral,SLC-40	28.5620	-80.5772		147
Vandenberg,SLC-4E	34.6320	-120.6108		94
Kennedy,LC-39A	28.6082	-80.6041		47

Night vs Day Launches:

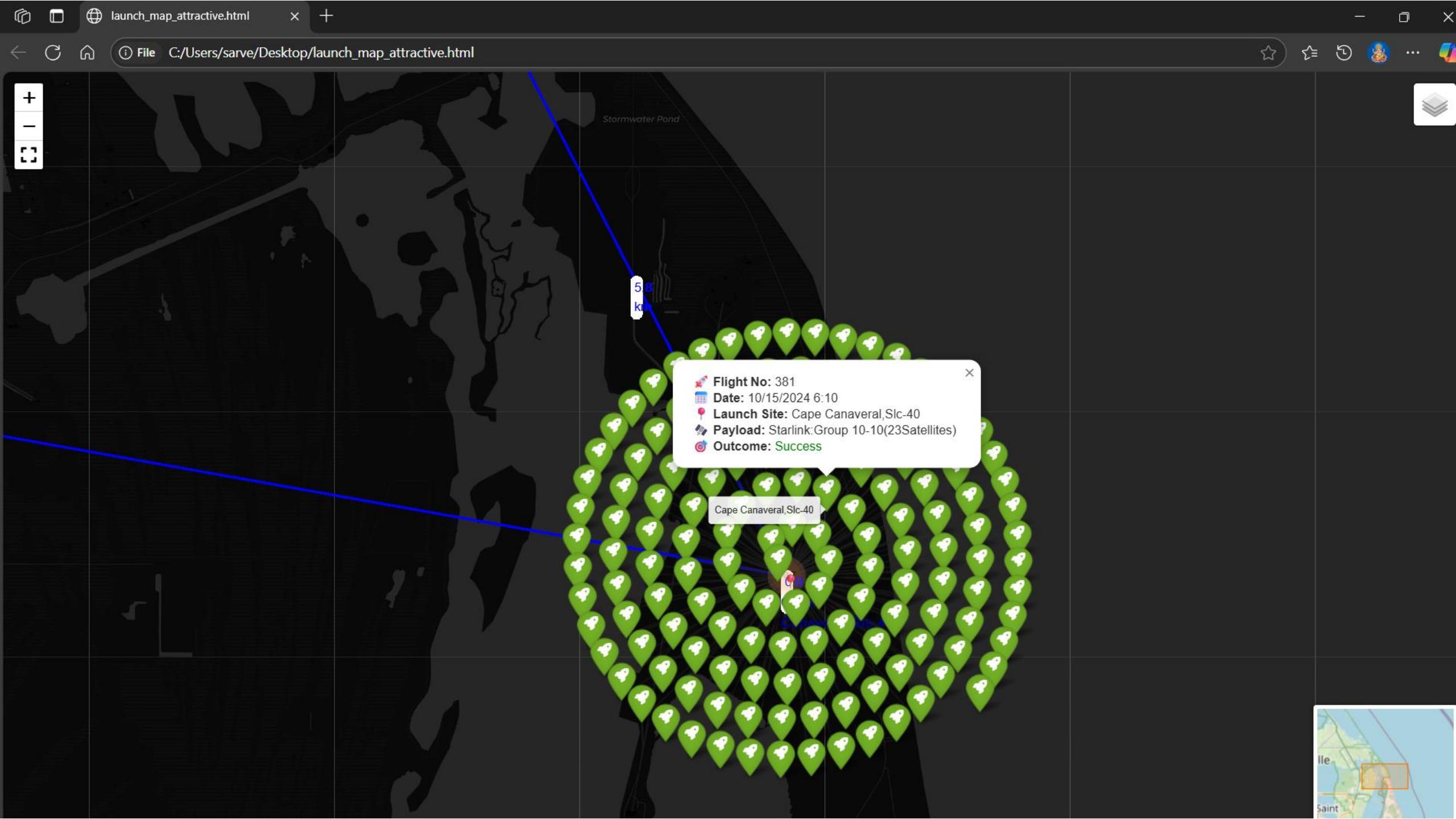
is_night_launch	launch_count
0	103
1	185

- Interaction with Folium map



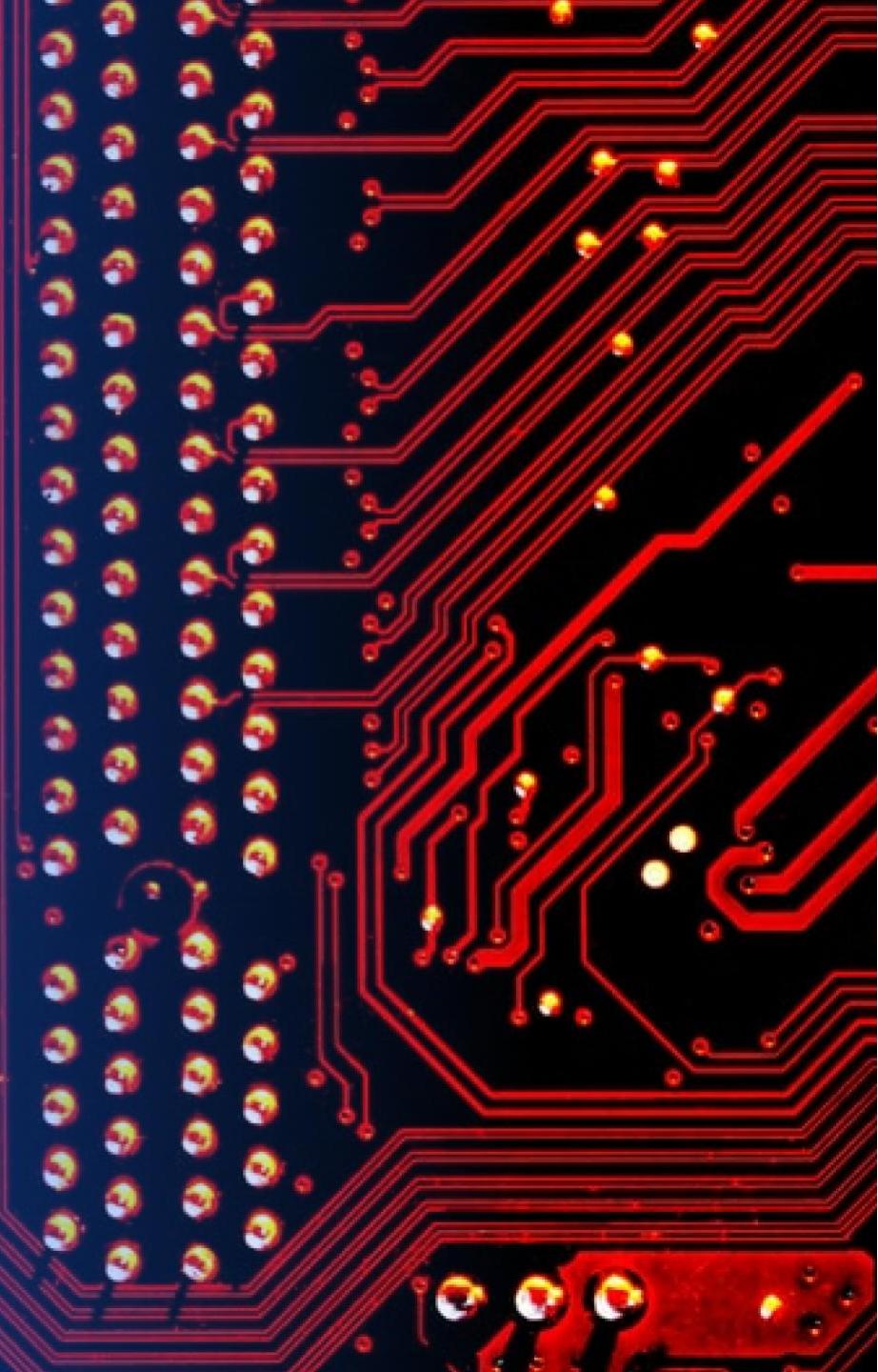






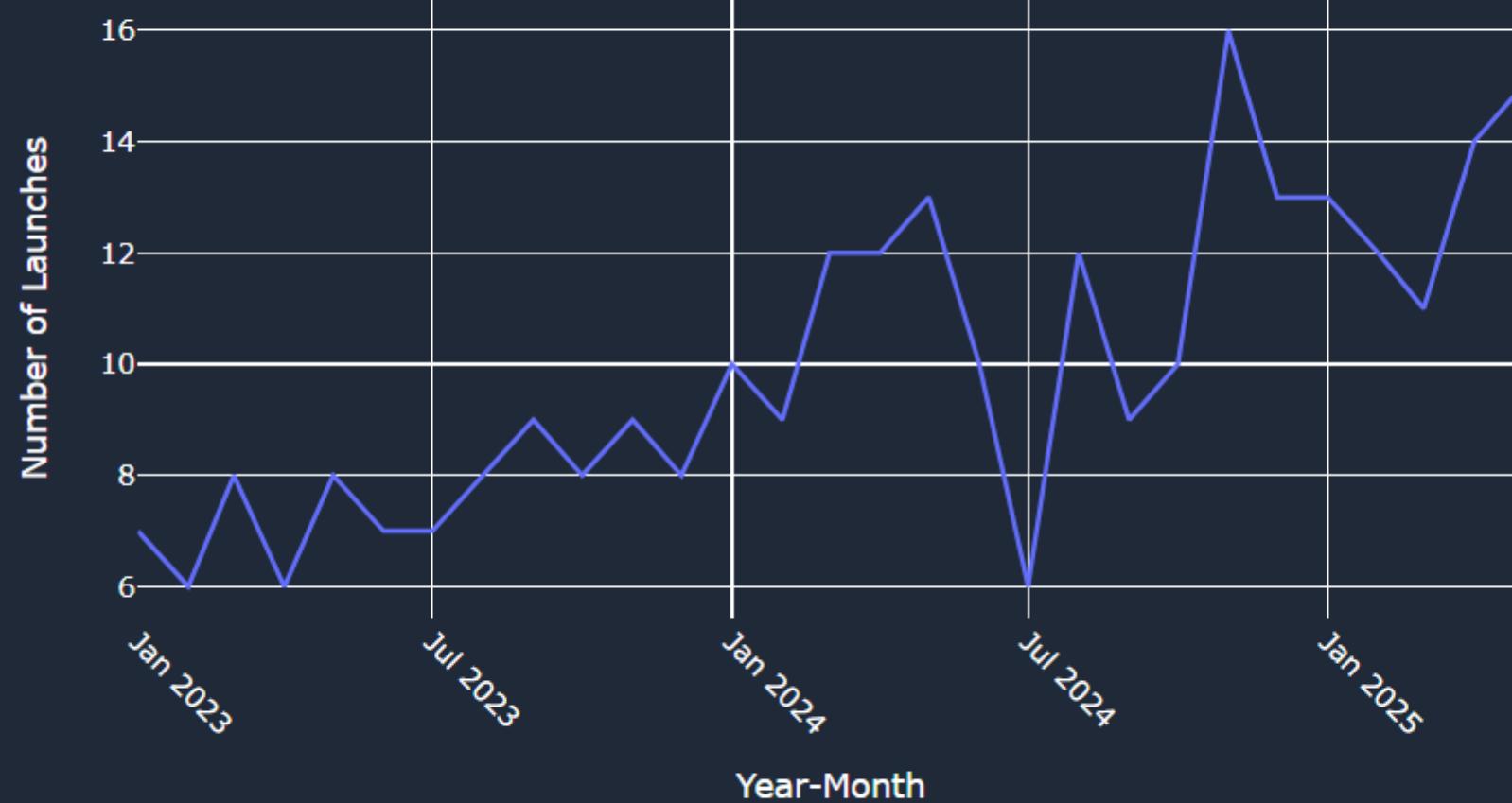
Section 4

Build a Dashboard with Plotly Dash



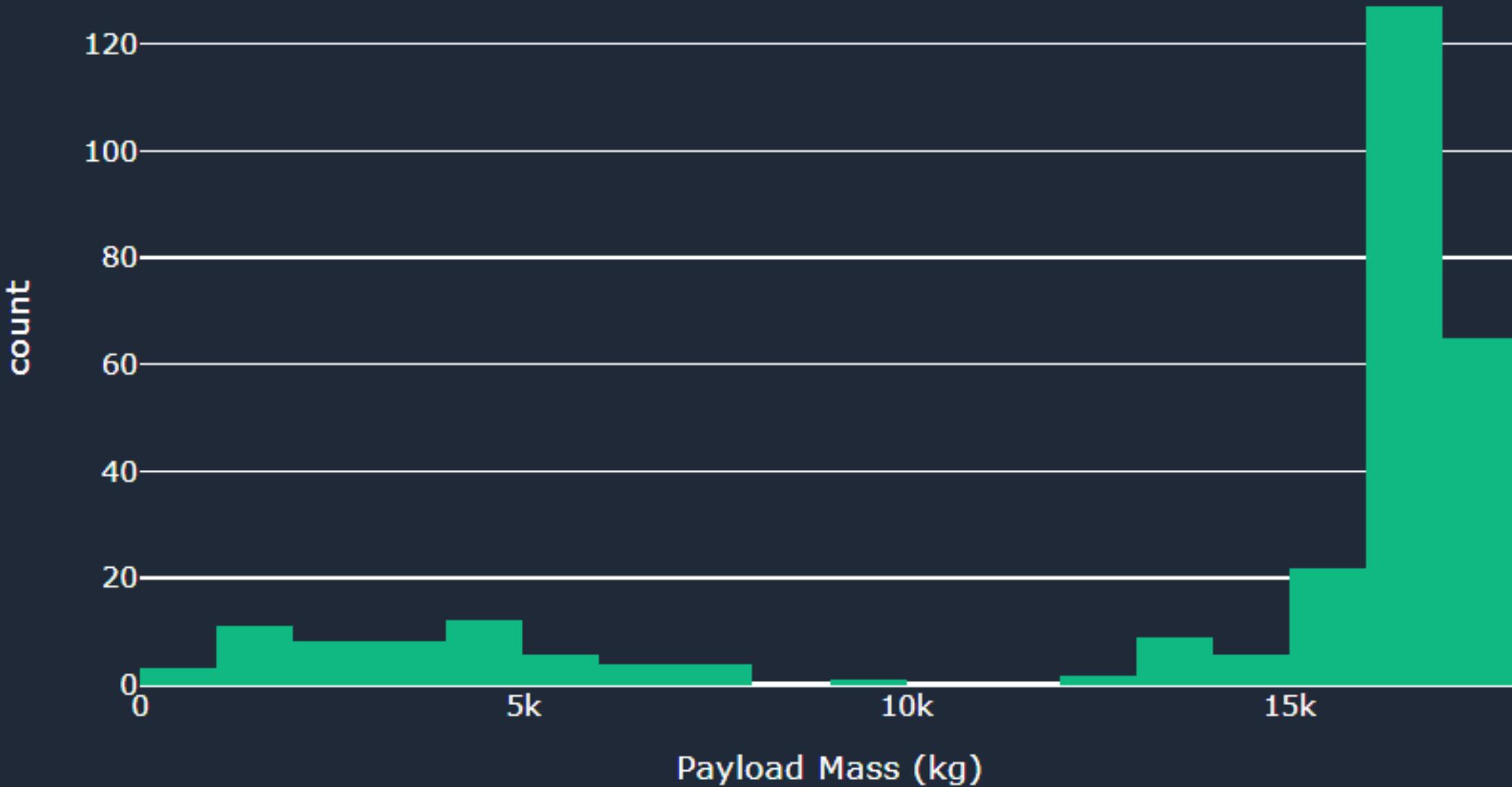
Starlink missions dominate SpaceX launches, with over 69.4% of launches dedicated to deploying S

Launch Frequency Over Time

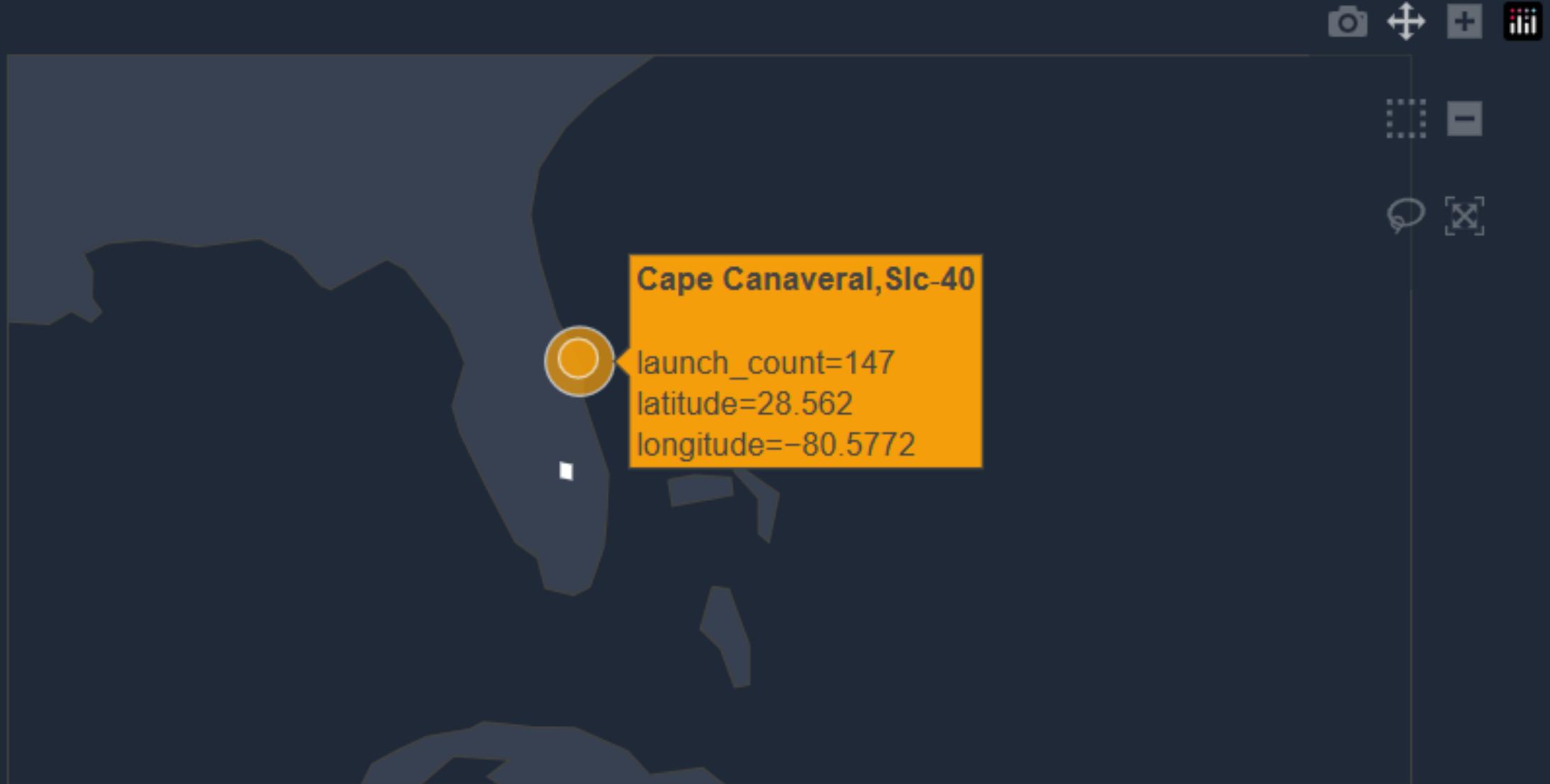


Launch Sites

Payload Mass Distribution



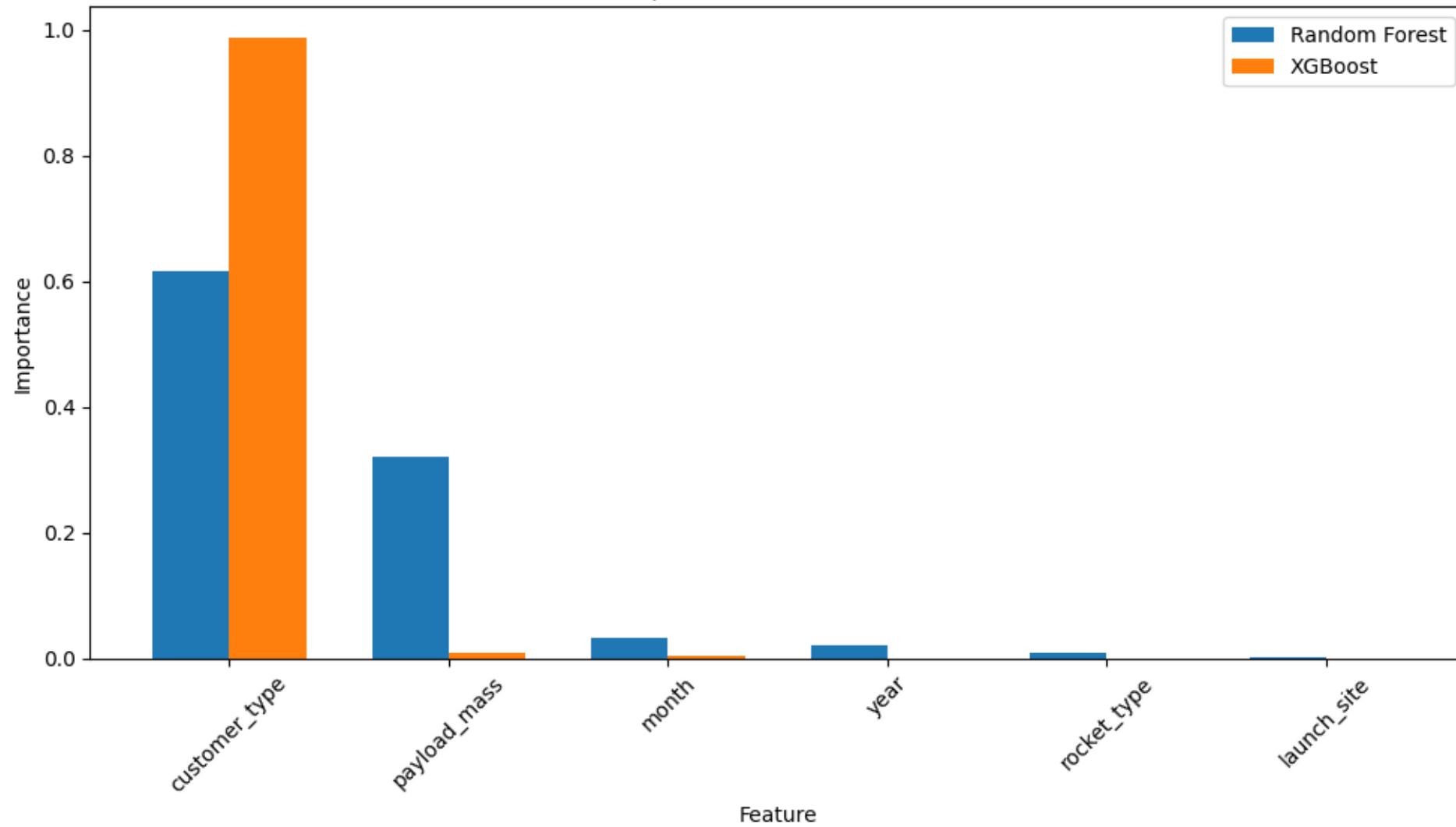
Launch Sites



Section 5

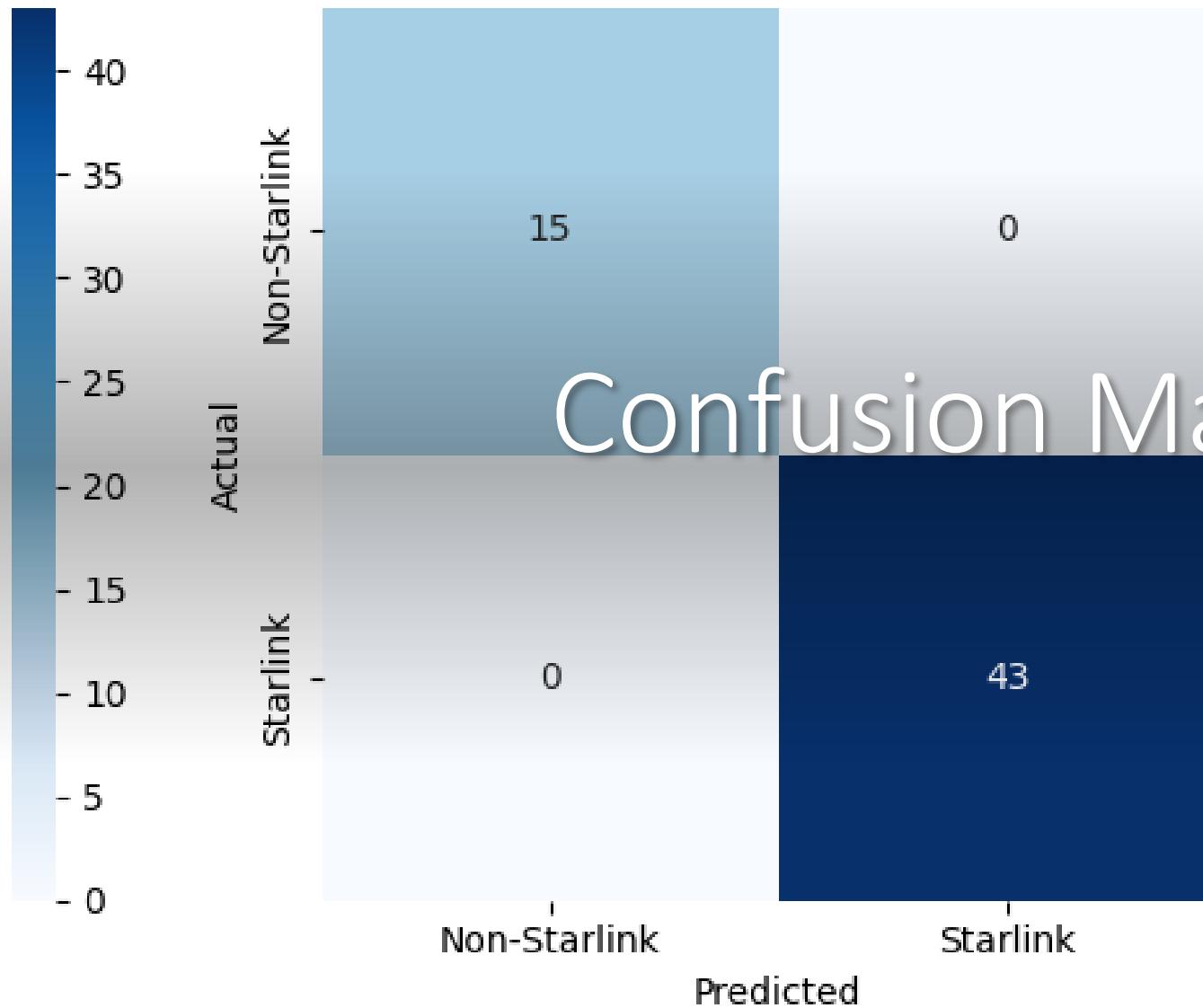
Predictive Analysis (Classification)

Feature Importance for Starlink Prediction

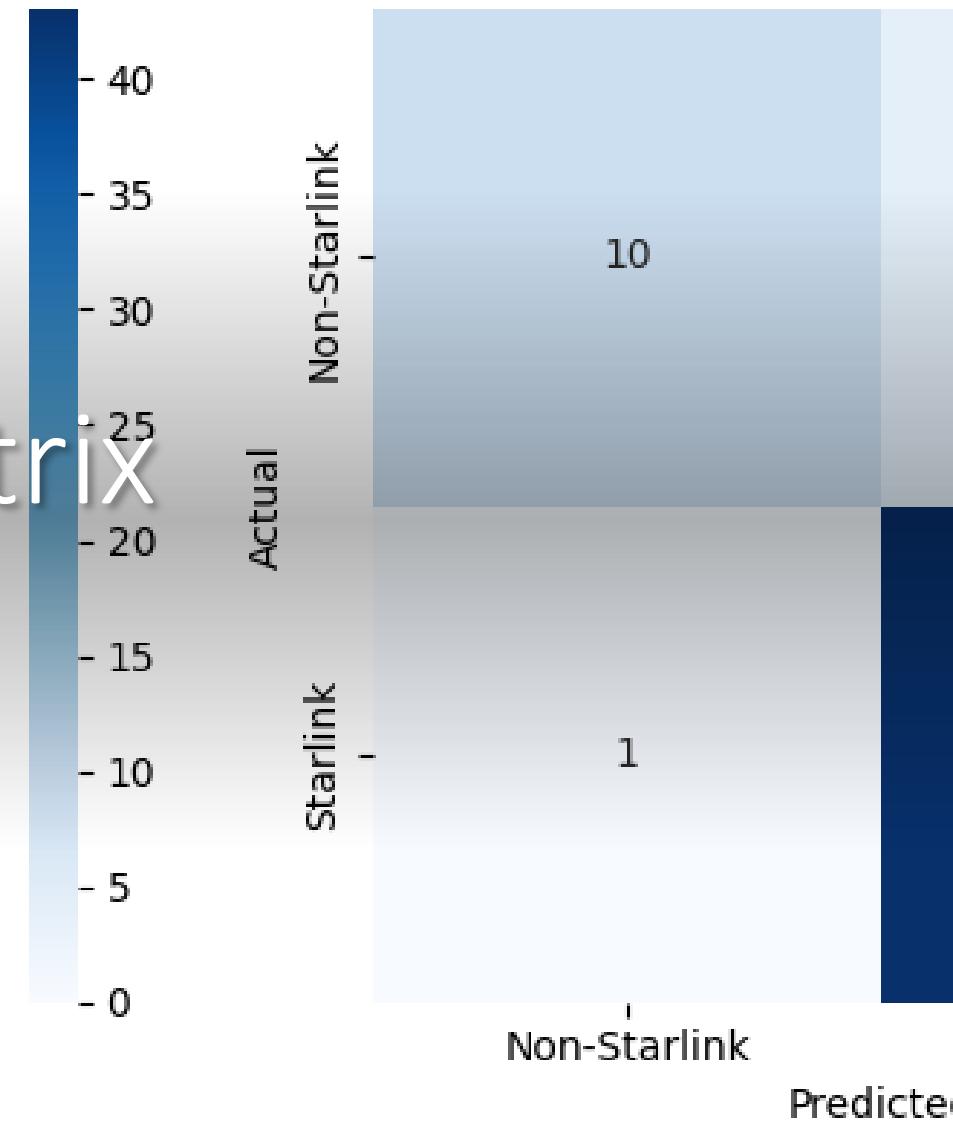


Confusion Matrices for Starlink Prediction

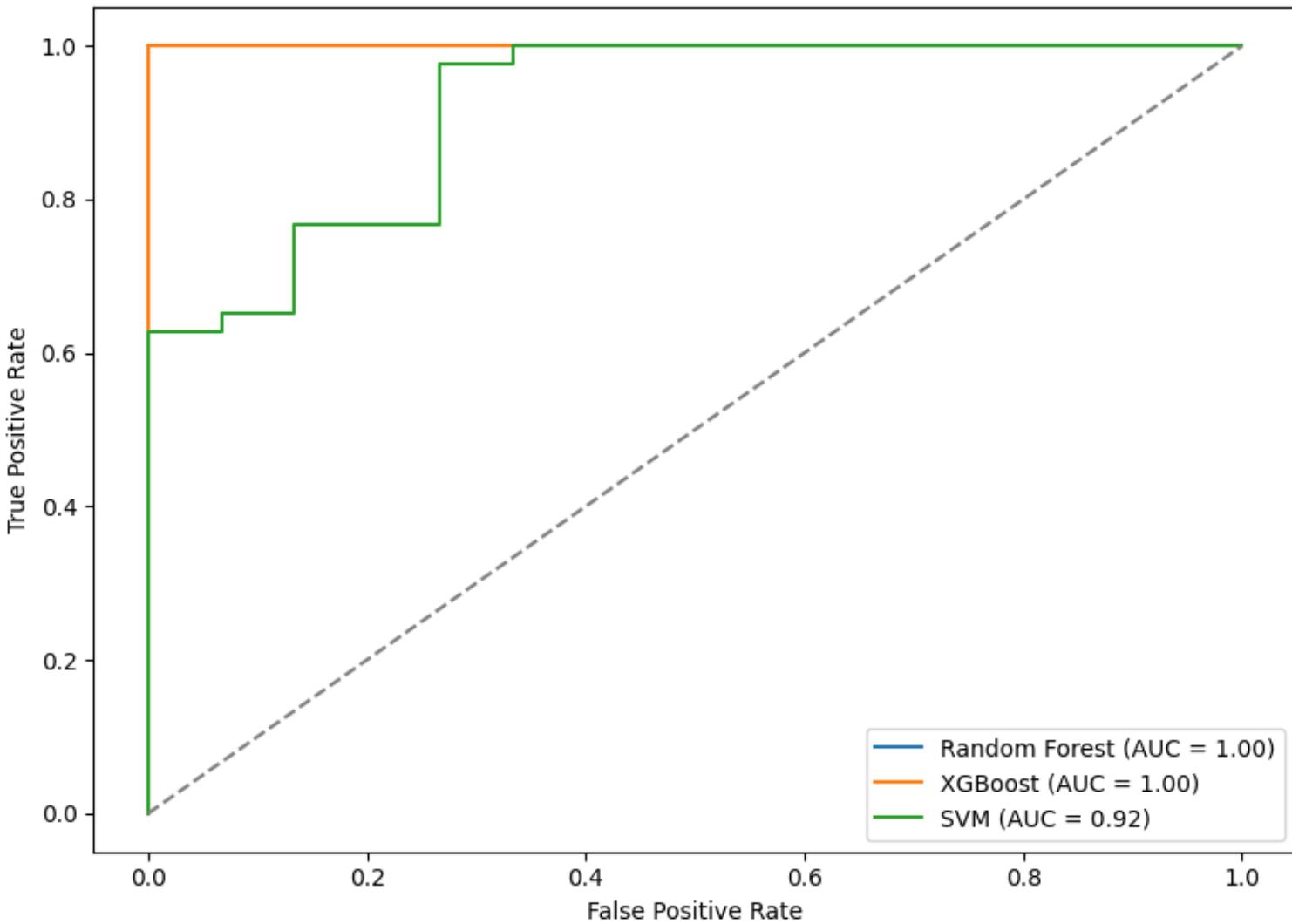
XGBoost



SVM



ROC Curves for Starlink Prediction





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!

