IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Allegra Weston

13/03/2025

Outline

Executive Summary

This project analyzes SpaceX launch data to identify trends, predict mission outcomes, and assess payload impact.

Key Findings

High Success Rate: SpaceX consistently demonstrates reliable launch performance.

Launch Site Efficiency: Some sites achieve higher success rates, optimizing mission planning.

Payload Influence: Heavier payloads correlate with lower success rates.

Orbit Type Trends: Certain orbits have higher success probabilities.

Booster Reusability: High success in drone ship landings enhances cost-effectiveness.

Predictive Analysis

A classification model predicts launch outcomes with strong accuracy, showcasing the potential for data-driven decision-making.

Conclusion

This study provides actionable insights into SpaceX's operations, leveraging data analysis, interactive tools, and machine learning to enhance future mission planning.

Introduction

SpaceX Launch Data Analysis

This project aims to analyze SpaceX launch data to gain insights into various aspects of their launch operations, including launch trends, mission outcomes, and payload characteristics.

Project Goals

- **Data Exploration:** Explore SpaceX launch data to identify patterns and trends related to launch success, payload mass, orbit types, and other relevant factors.
- **Interactive Visualizations:** Create interactive visualizations using Folium and Plotly Dash to provide intuitive interfaces for exploring the data.
- **Predictive Modeling:** Develop a classification model to predict launch outcomes based on relevant features.
- **Performance Evaluation:** Evaluate the performance of the classification model and identify areas for improvement.
- **Insight Generation:** Generate actionable insights that can be used to improve future launch planning and operations.

Data Sources

The data for this project was obtained from two primary sources:

- **SpaceX API:** The SpaceX API provides access to real-time and historical data on SpaceX launches, rockets, and payloads.
- **Web Scraping:** Web scraping was used to extract additional data from websites (e.g., Wikipedia) that are not available through the SpaceX API.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology

We collected SpaceX launch data using both the SpaceX API and web scraping. The API provided structured data on launches, rockets, and payloads, while web scraping from sources like Wikipedia supplemented missing details. The data was cleaned, transformed, and enriched by handling missing values, converting data types, and creating new features.

Exploratory Data Analysis (EDA) leveraged visualizations (scatter plots, bar charts, line charts) and SQL queries to uncover trends and patterns. Classification models were then developed to predict launch outcomes, evaluated using accuracy, precision, recall, and F1-score. The process involved selecting models, training on the dataset, tuning hyperparameters, and assessing performance to enhance predictive accuracy.

The project follows a structured methodology that includes:

- **Data Collection:** Gathering data from the SpaceX API and web scraping.
- **Data Wrangling:** Cleaning, transforming, and preparing the data for analysis.
- **Exploratory Data Analysis (EDA):** Exploring the data using visualizations and SQL queries.
- **Interactive Visual Analytics:** Creating interactive visualizations using Folium and Plotly Dash.
- **Predictive Analysis:** Building and evaluating a classification model.

Data Collection – SpaceX API

The main objective was to query the SpaceX Api. The endpoints were as follows:

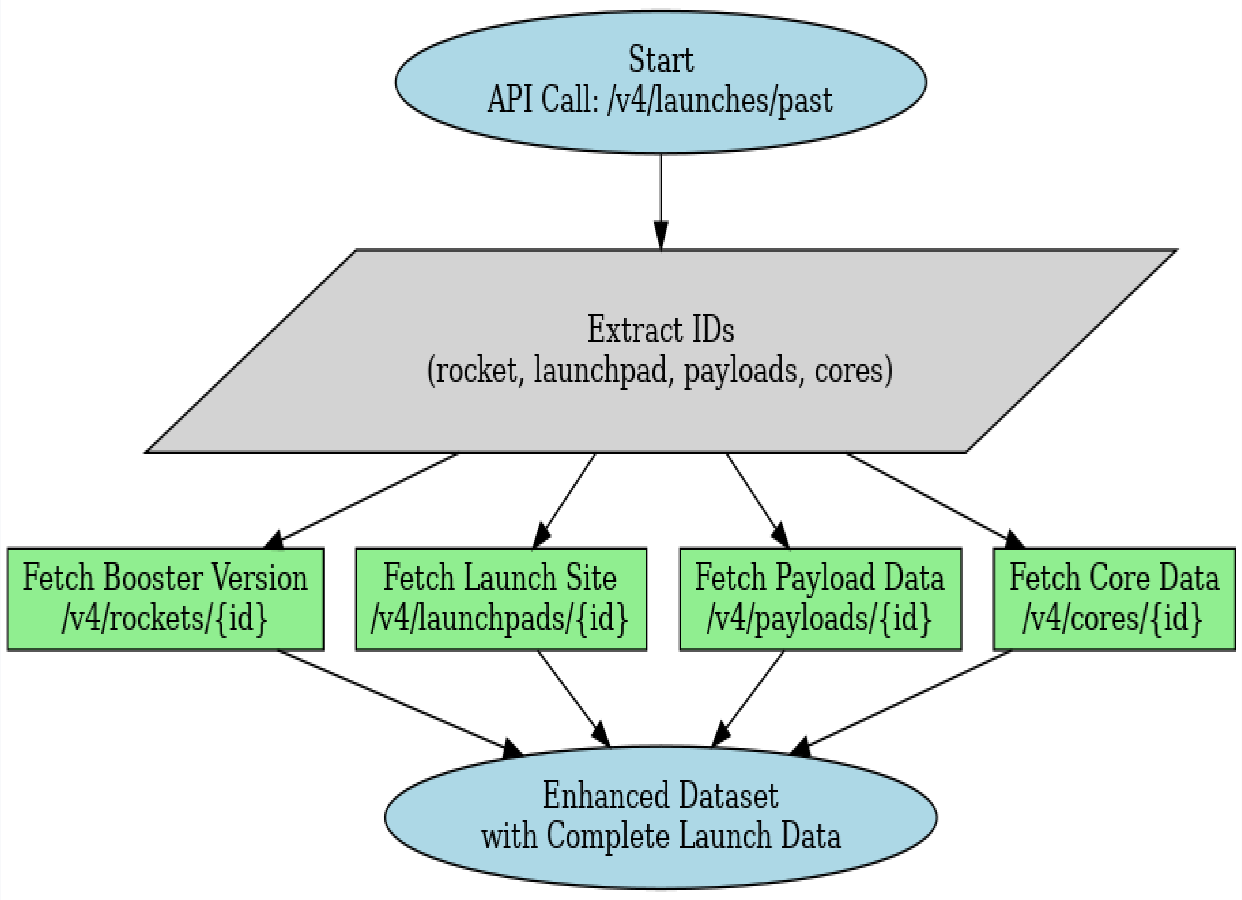
Booster Version: Extracted from /v4/rockets/

Launch Site: Coordinates and name from /v4/launchpads/

Payload Data: Mass and orbit from /v4/payloads/

Core Data: Block, reuse count, and landing success from /v4/cores

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

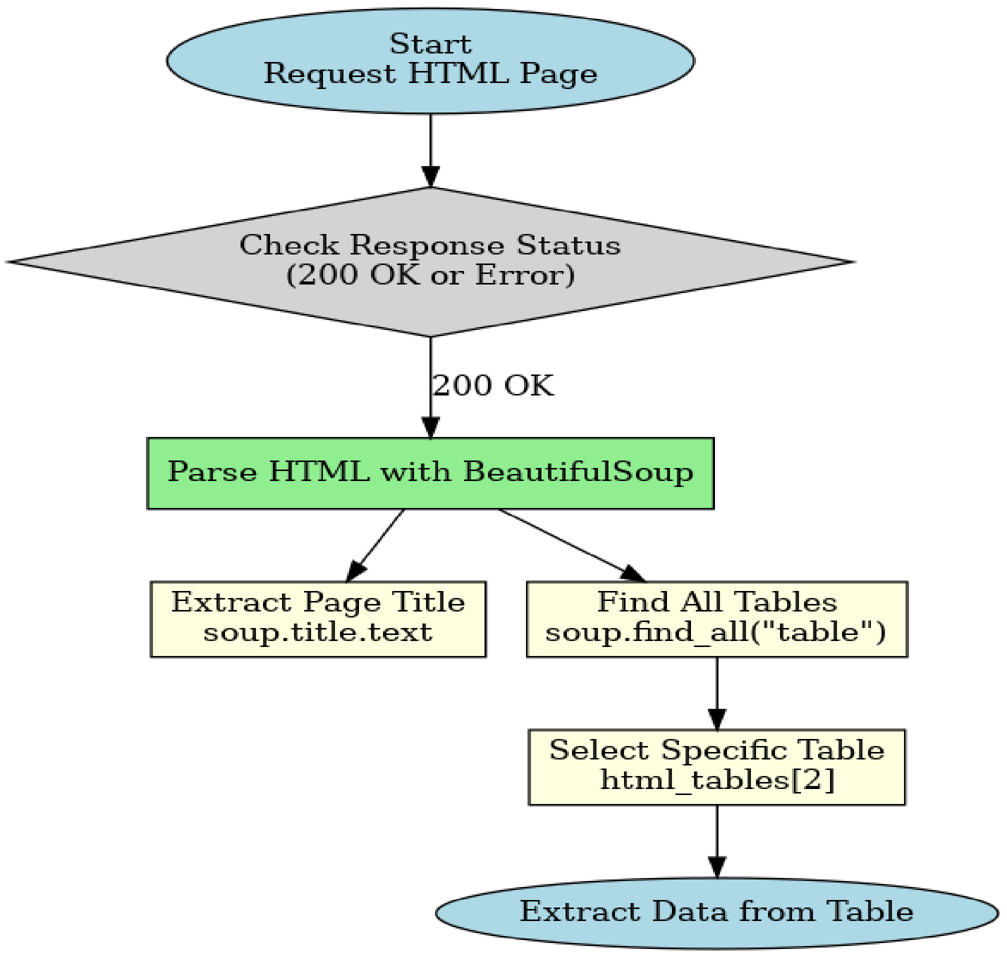
To extract SpaceX launch data web scraping techniques were used on Wikipedia.

First the webpage was retrieved using requests.get(), then the response was parsed using BeautifulSoup.

The page title was extracted, and I identified all tables within the page

From these I was able to select the relevant launch data table for further analysis.

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

The dataset includes information on SpaceX launches, including launch sites, orbits, payload mass, and mission outcomes. The data was processed to extract insights and prepare it for predictive modeling.

Key steps included: Data Extraction and importation of dataset containing SpaceX launch records.

Identification of key columns: LaunchSite, Orbit, Outcome, PayloadMass, etc.

Data Cleaning & Transformation which included removed missing values and standardized column formats.

Data analysis was then conducted using value_counts() to analyze launch sites and orbit types. Defining mission outcome categories (e.g., True ASDS, False Ocean). Created landing_class to classify successful vs. unsuccessful landings. Mapped mission outcomes into binary labels (1 = Success, 0 = Failure) and finally calculated success rate using df['Class'].mean().

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Data visualizations were created using the following key graphs:

Scatter Plot: Showing payload mass vs. flight number. This plot helps visualize how payload mass varies across different flight numbers and how it correlates with successful or failed launches. This identifies trends in payload capacity and success rates over multiple flights.

Scatter Plot: Flight Number vs. Launch Site To help observe the distribution of launches across different launch sites. This graph highlights the frequency and success rate of launches at specific locations.

Scatter Plot: Payload mass vs orbit type To depict the success/ failures of different masses and whether these affected the outcome.

Line Chart: Average Launch Successes This line chart gives a representation of how successful launches were.

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/edadataviz.ipynb

EDA with SQL

Summary of SQL Queries Performed:

•Created a new table SPACEXTABLE from SPACEXTBL while ensuring non-null dates

•Retrieved distinct launch sites from SPACEXTBL.

•

•Filtered launch records for specific sites (e.g., CCA%) with a limit

•Computed average payload mass

•Identified the first successful landing date on a ground pad.

•Calculated total payload mass for NASA CRS missions.

•Analyzed booster versions and launch site distribution

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

The following map objects were used to help in analysing the geographical distribution of launch sites, their accessibility, and potential risks associated with launches.

- **Markers:** Placed at SpaceX launch sites to visualize their exact locations.
- **Circles:** Added around launch sites to indicate impact zones or proximity analysis.
- **Lines:** Drawn to represent distances between launch sites and other points of interest (e.g., nearby cities or water bodies).
- **Mouse Position Plugin:** Enabled to display real-time latitude and longitude coordinates for interactive exploration.
- **Marker Cluster Plugin:** Used to group nearby markers for better visualization.

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

A dashboard was built with the following components:

- **Pie Chart:** Displays the success rate of launches, either for all SpaceX launch sites or for a selected launch site. This helps in understanding which sites have higher or lower success rates.
- **Scatter Plot:** Shows the relationship between payload mass and launch success, categorized by launch site. This visualization helps analyze how payload weight affects launch outcomes.
- **Dropdown Menu:** Allows users to select a specific launch site to filter data accordingly, providing insights into site-specific performance.
- **Payload Slider:** Enables users to adjust the payload range to examine how different payload sizes impact launch success.
- These visualizations and interactions enhance data exploration, making it easier to identify trends and factors affecting SpaceX launch success.

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

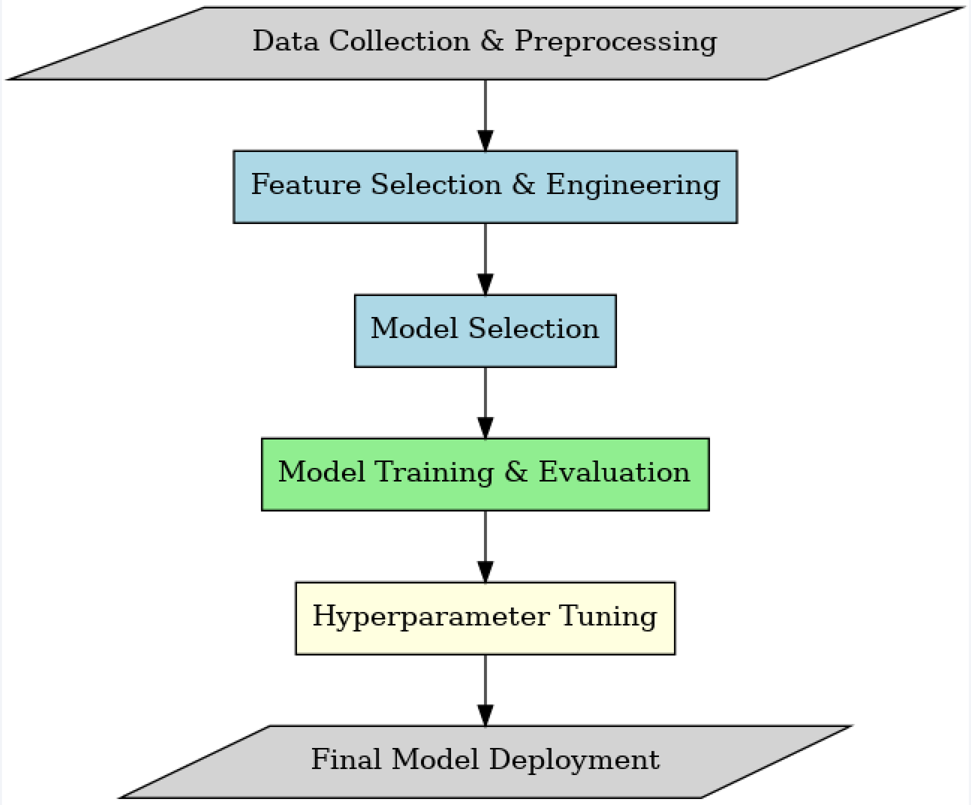
The classification model was developed using SpaceX launch data.

The process began with data preprocessing, where missing values were handled, categorical features were encoded, and numerical features were standardized.

Next, feature selection identified the most impactful variables affecting launch success. Multiple classification algorithms, including Logistic Regression, Support Vector Machine (SVM), Decision Trees, and Random Forest, were trained and evaluated using accuracy, precision, recall, and F1-score.

Hyperparameter tuning was conducted using GridSearchCV to optimize model performance. Finally, the best-performing model was selected based on evaluation metrics and cross-validation results.

https://github.com/AlleWeston-hash/DataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

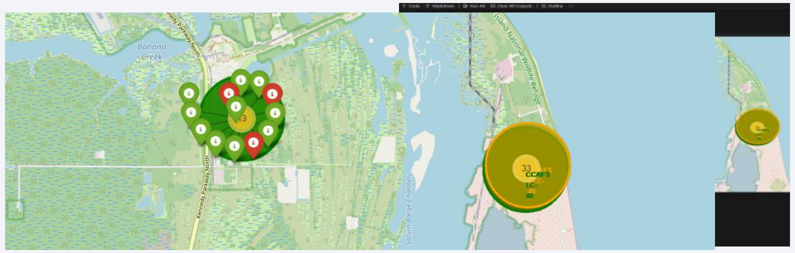


Results

Exploratory Data Analysis Results:

- Identified key trends and patterns in SpaceX launch data.
- Analyzed launch success rates by site, payload mass, and booster version.
- Visualized correlations between payload and launch success.

Interactive Analytics Demo in Screenshots:



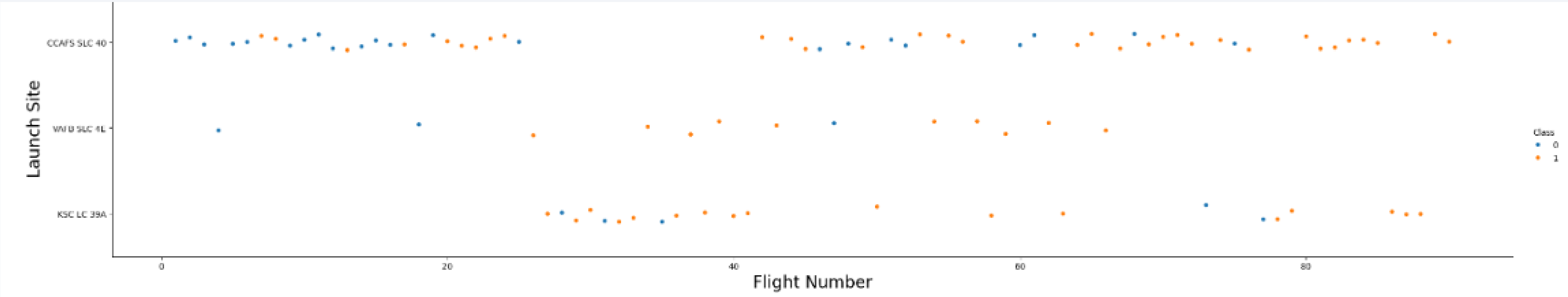
Predictive Analysis Results:

- Evaluated different classification models to predict launch success.
- Optimized model parameters to improve accuracy.
- Identified the best-performing model for predicting launch outcomes.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

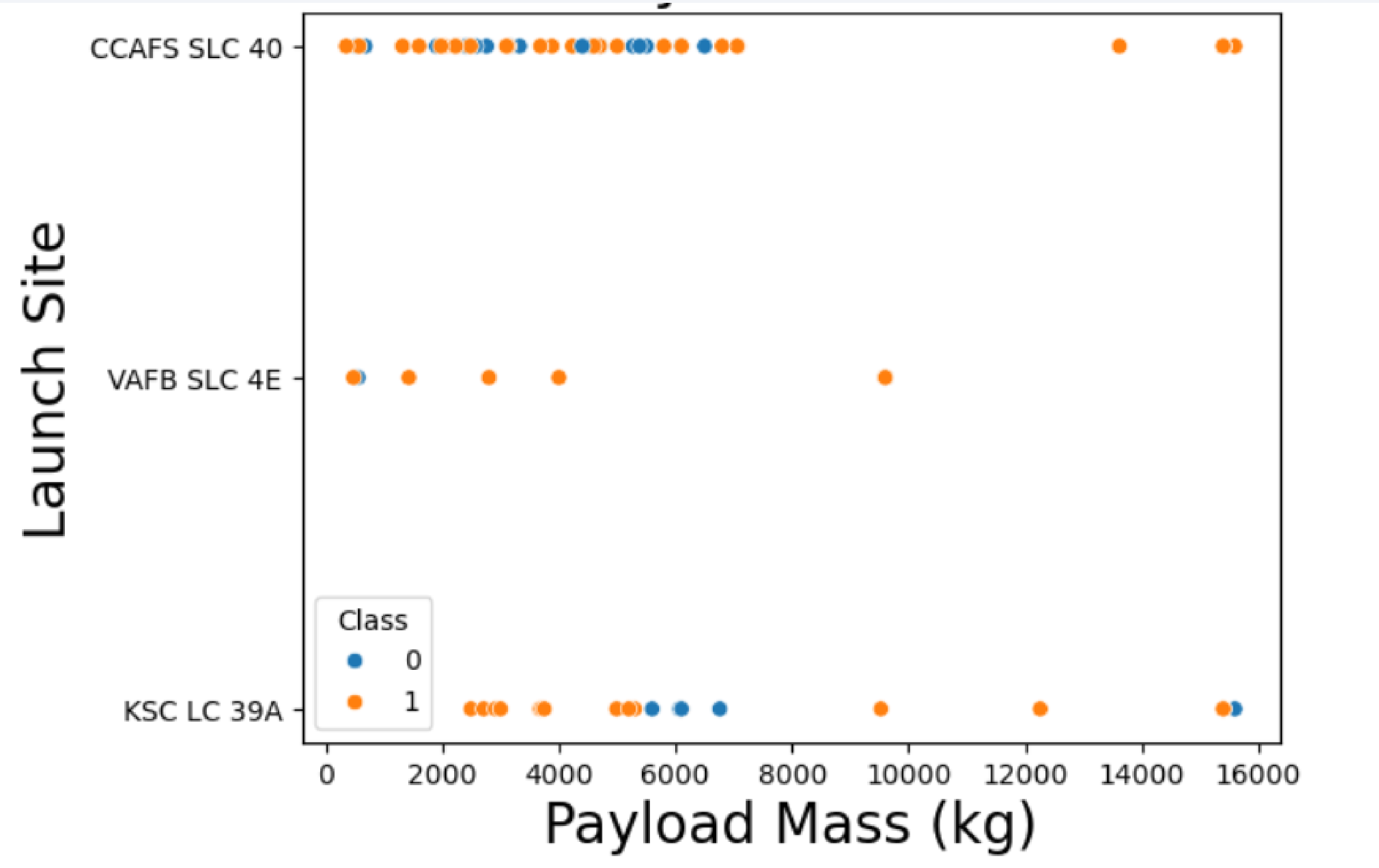


•**XAxis:** Flight number (chronological order of launches).
•**YAxis:** Launch site (categorical).
•**Data Points:** Each dot represents a launch, with color indicating success (1) or failure (0).
Insights:
•Success rates improve over time.
•Some sites show higher reliability.
•Later flights tend to have more successes, reflecting experience gains.

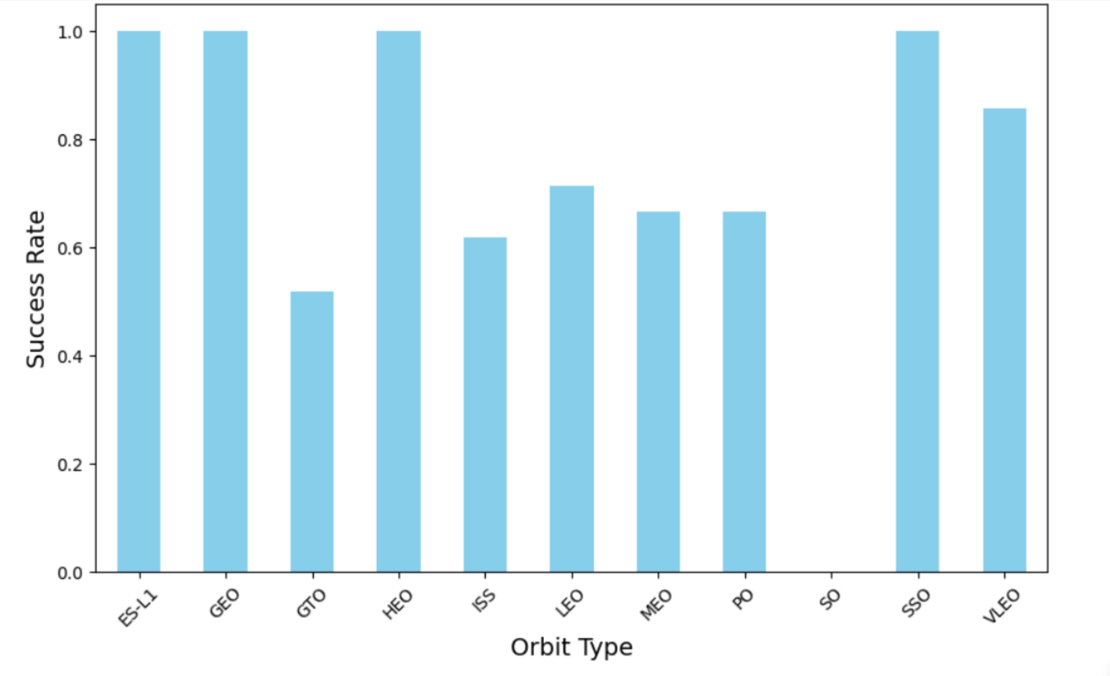
Payload vs. Launch Site

XAxis: Payload mass (kg).
YAxis: Launch site (categorical).
Data Points: Each dot represents a launch, with color indicating success (1) or failure (0).

Insights:
- Higher payloads have varying success rates across sites.
- Some launch sites handle heavier payloads more frequently.
- Success rates may decline slightly at extreme payload values.



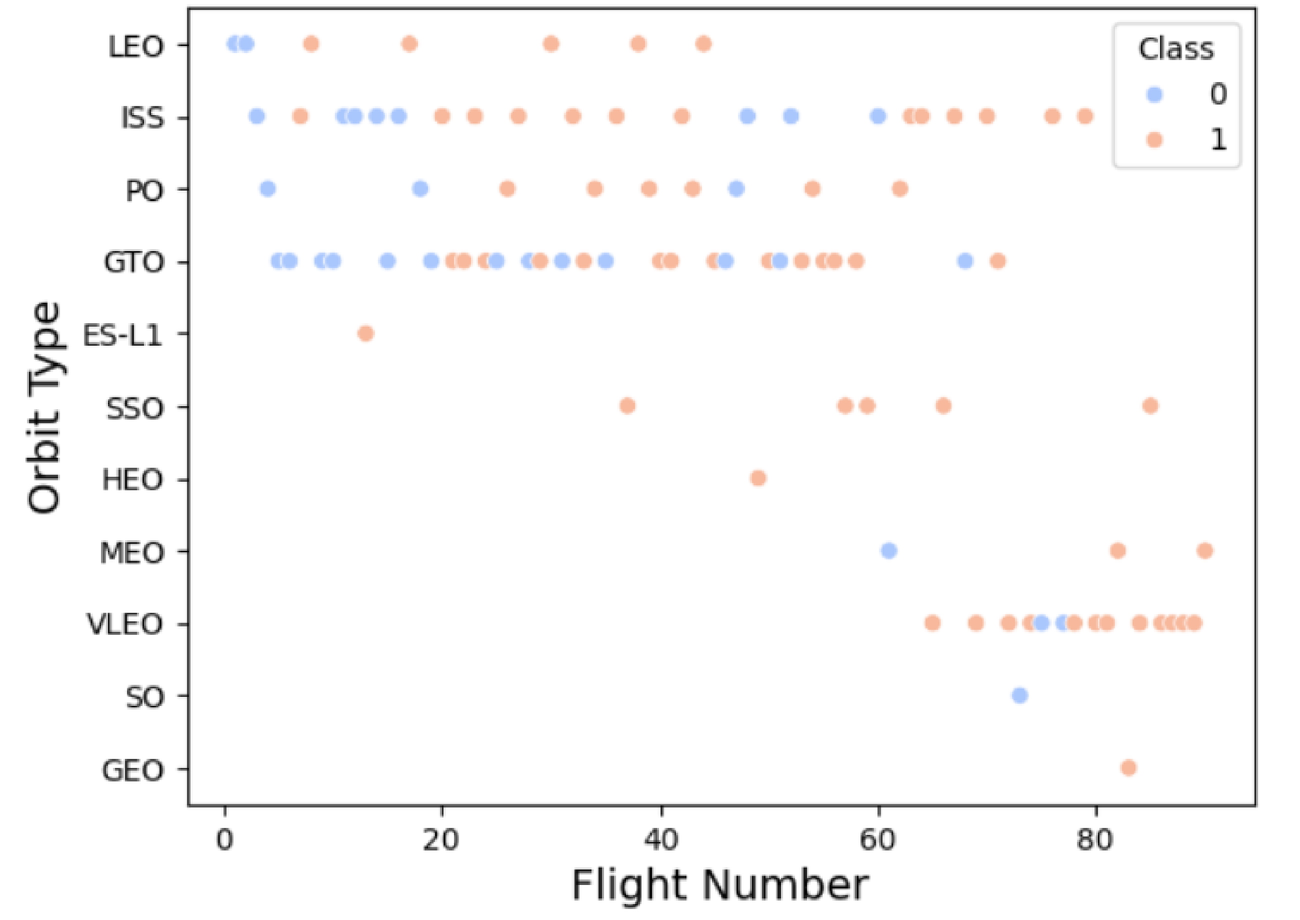
Success Rate vs. Orbit Type



The Success Rate vs. Orbit Type graph visualizes how launch success varies across different orbit destinations. Each bar (or point) represents the proportion of successful launches for a given orbit type. Orbits with frequent launches, such as LEO (Low Earth Orbit), may show higher success rates due to experience and optimization, while more challenging orbits, like GTO (Geostationary Transfer Orbit), could have lower success rates due to complex mission requirements. This analysis helps identify trends in mission reliability across different orbital destinations.

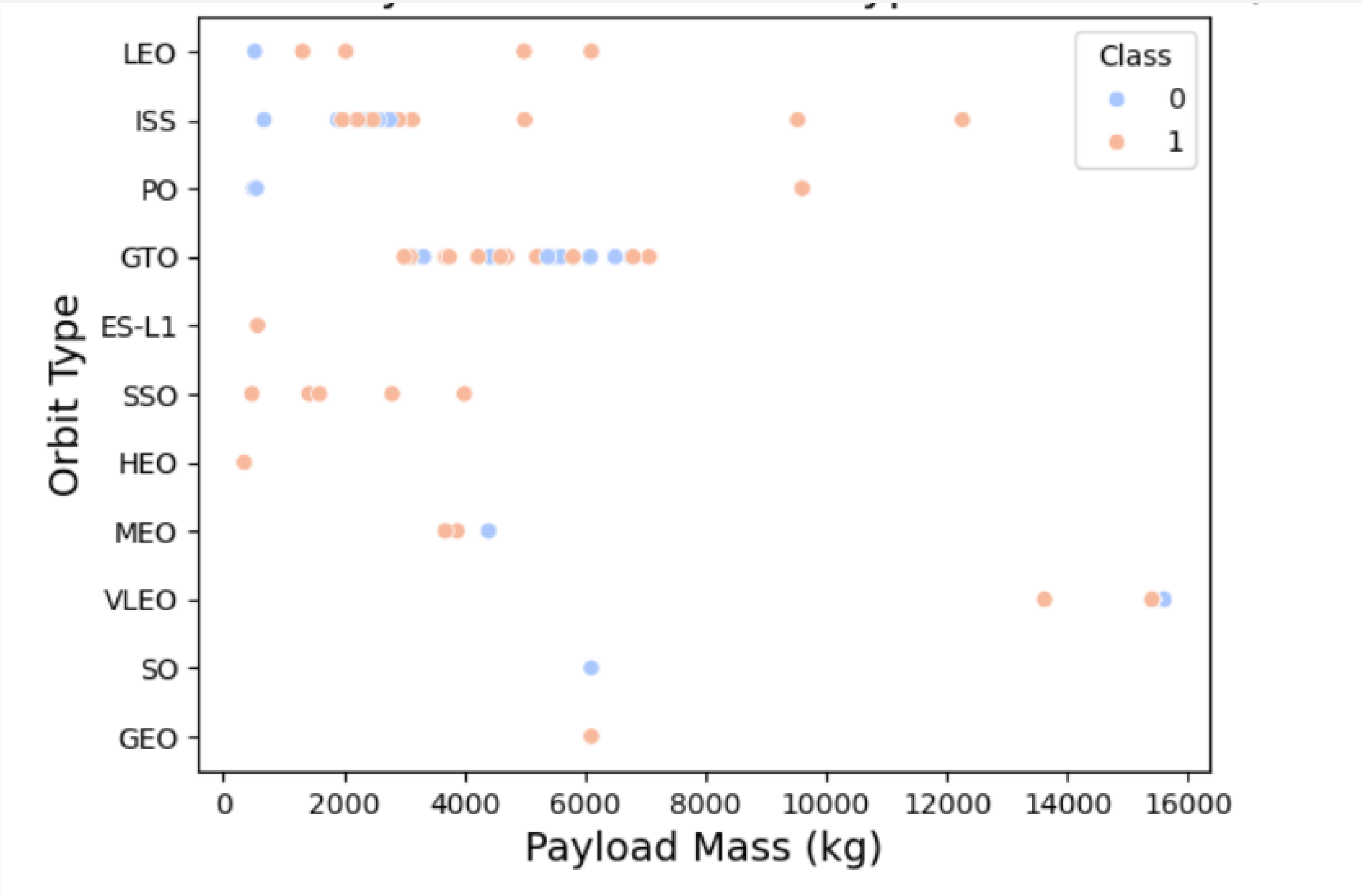
Flight Number vs. Orbit Type

The Flight Number vs. Orbit Type scatter plot shows how different SpaceX missions are distributed across various orbit types over time. Each point represents a launch, with flight numbers on the x-axis and orbit types on the y-axis. This visualization helps track how SpaceX has expanded its mission portfolio, showing trends such as increasing launches to specific orbits as experience grows. It may also highlight whether certain orbits were targeted more frequently in earlier or later missions.



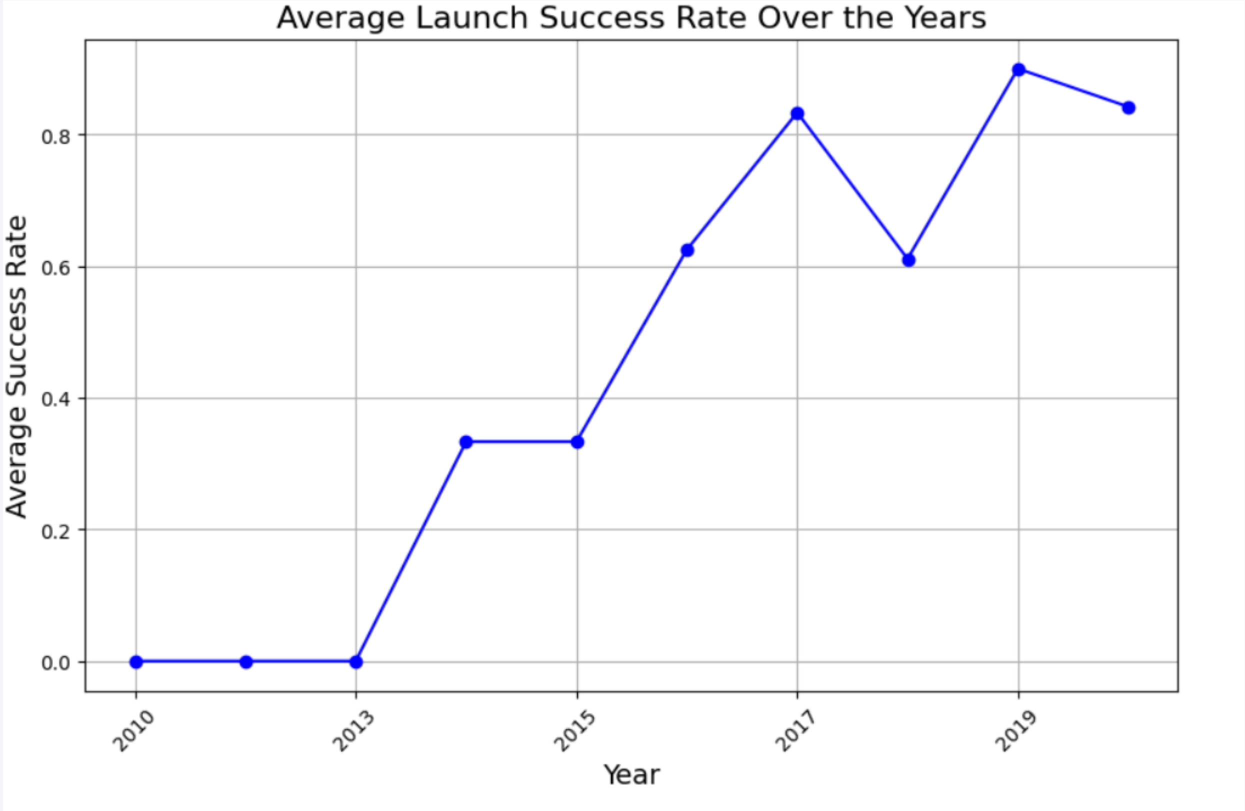
Payload vs. Orbit Type

The Payload vs. Orbit Type scatter plot illustrates the relationship between payload mass and the types of orbits SpaceX missions targeted. Each point represents a launch, with payload mass on the x-axis and orbit type on the y-axis. This visualization helps identify trends, such as which orbits accommodate heavier payloads and whether certain orbit types have stricter payload limitations. It can also highlight patterns in SpaceX's launch strategy for different payload capacities.



Launch Success Yearly Trend

The Launch Success Yearly Trend graph showcases how SpaceX's launch success rate has evolved over the years. By plotting the number of successful launches per year, this visualization highlights improvements in reliability and advancements in technology. An upward trend suggests growing expertise, better engineering, and enhanced operational efficiency, while any dips might indicate technical setbacks or testing phases. This trend is crucial for understanding SpaceX's progress in achieving consistent launch success.



All Launch Site Names

This command retrieves all distinct launch site names from the dataset. Identifying unique launch sites helps analyze SpaceX's operational range and the frequency of launches per location.

This insight is useful for understanding site preferences, logistical advantages, and historical performance trends.

These were:

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

```
df.columns
query = "select distinct Launch_Site from SPACEXTBL"
launch_sites_df = pd.read_sql(query, con)
print(launch_sites_df)
```

```
    Launch_Site
0   CCAFS LC-40
1   VAFB SLC-4E
2    KSC LC-39A
3  CCAFS SLC-40
```

Launch Site Names Begin with 'CCA'

This query filters the dataset for launch sites beginning with "CCA" (likely referring to Cape Canaveral sites) and returns the first five matching records. Analyzing these records helps understand launch frequency from this site, its success rates, and historical trends.

The screenshot to the right shows the rows of the first 5 launch sites whose names begin with CCA.

```python
query = "select * from SPACEXTBL where Launch_Site like 'CCA%' limit 5"

launch_sites_df = pd.read_sql(query, con)
print(launch_sites_df)
```

```
         Date Time (UTC) Booster_Version  Launch_Site  \
0  2010-06-04   18:45:00  F9 v1.0  B0003  CCAFS LC-40
1  2010-12-08   15:43:00  F9 v1.0  B0004  CCAFS LC-40
2  2012-05-22    7:44:00  F9 v1.0  B0005  CCAFS LC-40
3  2012-10-08    0:35:00  F9 v1.0  B0006  CCAFS LC-40
4  2013-03-01   15:10:00  F9 v1.0  B0007  CCAFS LC-40

                                              Payload  PAYLOAD_MASS__KG_  \
0                Dragon Spacecraft Qualification Unit                  0
1  Dragon demo flight C1, two CubeSats, barrel of...                  0
2                               Dragon demo flight C2                525
3                                        SpaceX CRS-1                500
4                                        SpaceX CRS-2                677

       Orbit         Customer Mission_Outcome      Landing_Outcome
0        LEO           SpaceX         Success  Failure (parachute)
1  LEO (ISS)  NASA (COTS) NRO         Success  Failure (parachute)
2  LEO (ISS)      NASA (COTS)         Success           No attempt
3  LEO (ISS)       NASA (CRS)         Success           No attempt
4  LEO (ISS)       NASA (CRS)         Success           No attempt
```

Total Payload Mass

This query filters the dataset to include only launches where NASA was the customer and sums up the payload mass for those launches. This helps assess NASA's total payload contribution and its significance in SpaceX's overall launch history.

The total mass was calculated 45596

```
query = """
select sum(payload_mass__kg_) as total_payload_mass
from spacextbl
where customer = 'NASA (CRS)'
"""

total_payload_mass_df = pd.read_sql(query, con)

print(total_payload_mass_df)
```

```
   total_payload_mass
0               45596
```

Average Payload Mass by F9 v1.1

- This query filters the dataset to include only launches where the booster version is F9 v1.1 and then calculates the mean payload mass. This helps evaluate the performance and capacity of this specific booster version compared to others.

The average was calculated to be 2928.4

```
query = """
select avg(payload_mass__kg_) as average_payload_mass
from spacextbl
where booster_version = 'F9 v1.1'
"""
average_payload_mass_df = pd.read_sql(query, con)
print(average_payload_mass_df)
```

```
   average_payload_mass
0                2928.4
```

First Successful Ground Landing Date

This query retrieves the earliest date when a successful landing occurred on a ground pad. Ordering by date in ascending order ensures that we get the first recorded success. This information helps track SpaceX's progress in achieving ground landings.

The date was 2015-12-22

```
query = """
select min(date) as first_successful_landing_date
from spacextbl
where landing_outcome = 'Success (ground pad)'
"""
first_successful_landing_date = pd.read_sql(query, con)
print(first_successful_landing_date)
```

```
first_successful_landing_date
                   2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

This query filters the dataset to include only boosters that:

- Had a successful landing on a drone ship
- Carried a payload between 4000 and 6000 kg

This helps analyze the boosters' performance under specific payload conditions and track their reusability.

The result are in the screenshot to the right

```
query = """
select booster_version
from spacextbl
where landing_outcome = 'Success (drone ship)'
and payload_mass__kg_ > 4000
and payload_mass__kg_ < 6000
"""
boosters = pd.read_sql(query, con)
print(boosters)
```

```
  Booster_Version
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

This query groups all launch records by mission outcome (e.g., "Success", "Failure")

It counts the number of occurrences for each outcome

This provides insight into SpaceX's overall mission success rate, helping assess launch reliability over time.

The total number of successful mission outcomes were 100 with 1 failure

```
query = """
select mission_outcome, count(*) as total_count
from spacextbl
group by mission_outcome
"""
mission_outcomes = pd.read_sql(query, con)
print(mission_outcomes)
```

```
                Mission_Outcome  total_count
0              Failure (in flight)            1
1                        Success           98
2                       Success             1
3  Success (payload status unclear)            1
```

Boosters Carried Maximum Payload

This query retrieves the booster version(s) associated with the highest recorded payload mass

The subquery finds the maximum payload mass in the dataset

The outer query filters for the booster(s) that carried this maximum payload

This helps identify the most capable boosters in terms of payload capacity.

The result are in the screenshot to the right

```
query = """
select booster_version
from spacextbl
where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
"""
booster_versions = pd.read_sql(query, con)
print(booster_versions)
```

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

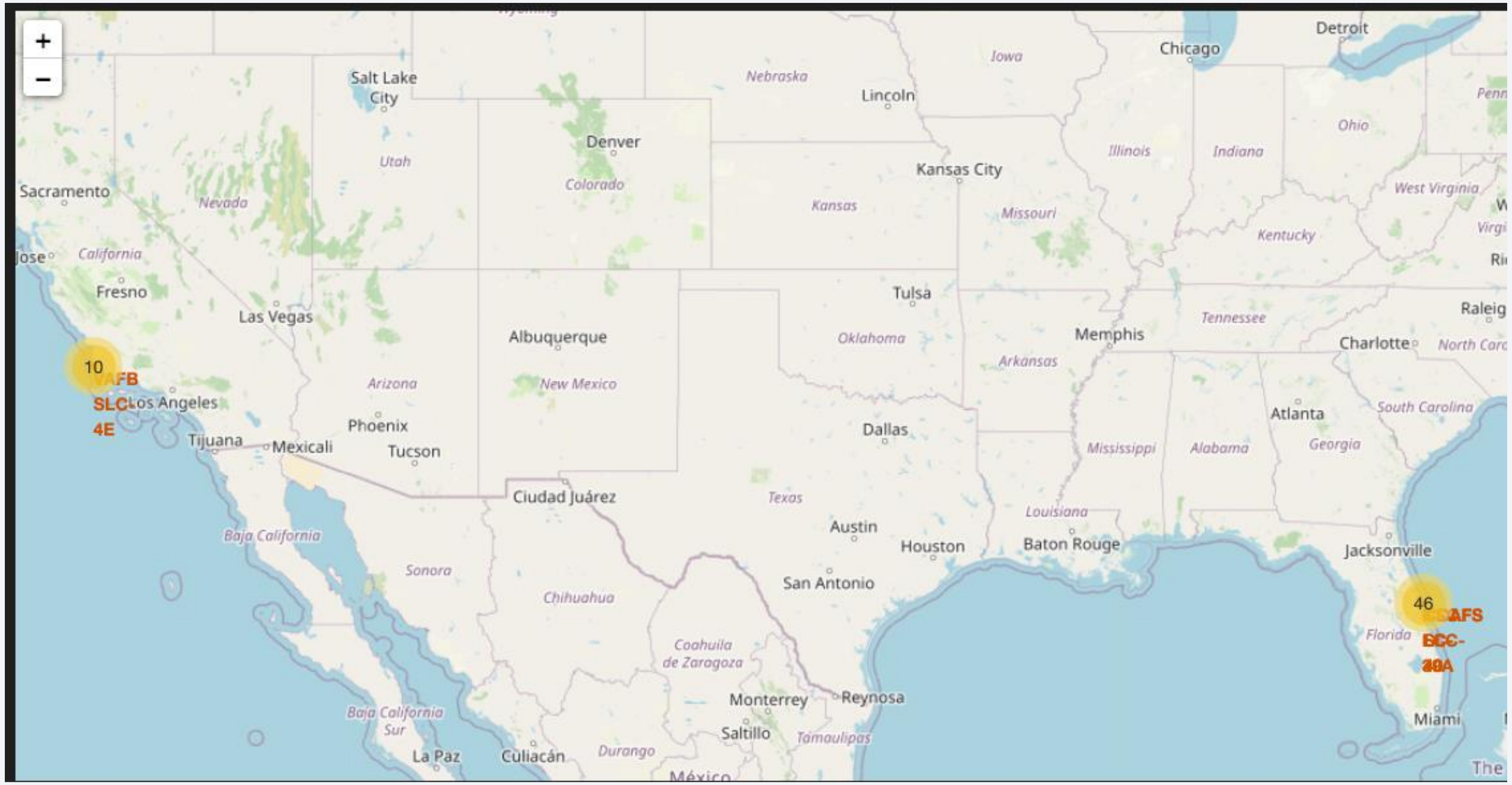
Section 3

Launch Sites Proximities Analysis

Mission Launch Sites

Most SpaceX launch sites are concentrated in the United States, particularly in Florida and California.

This geospatial visualization helps analyze how launch locations impact mission success and trajectory planning.



Launch Site Success Failures Map

- Each site is marked on the map, showing SpaceX's key launch locations.
- Color Labels for Outcomes:
- Green markers indicate successful launches.
- Red markers indicate failed launches.
- Orange markers may represent partial success or uncertain outcomes.

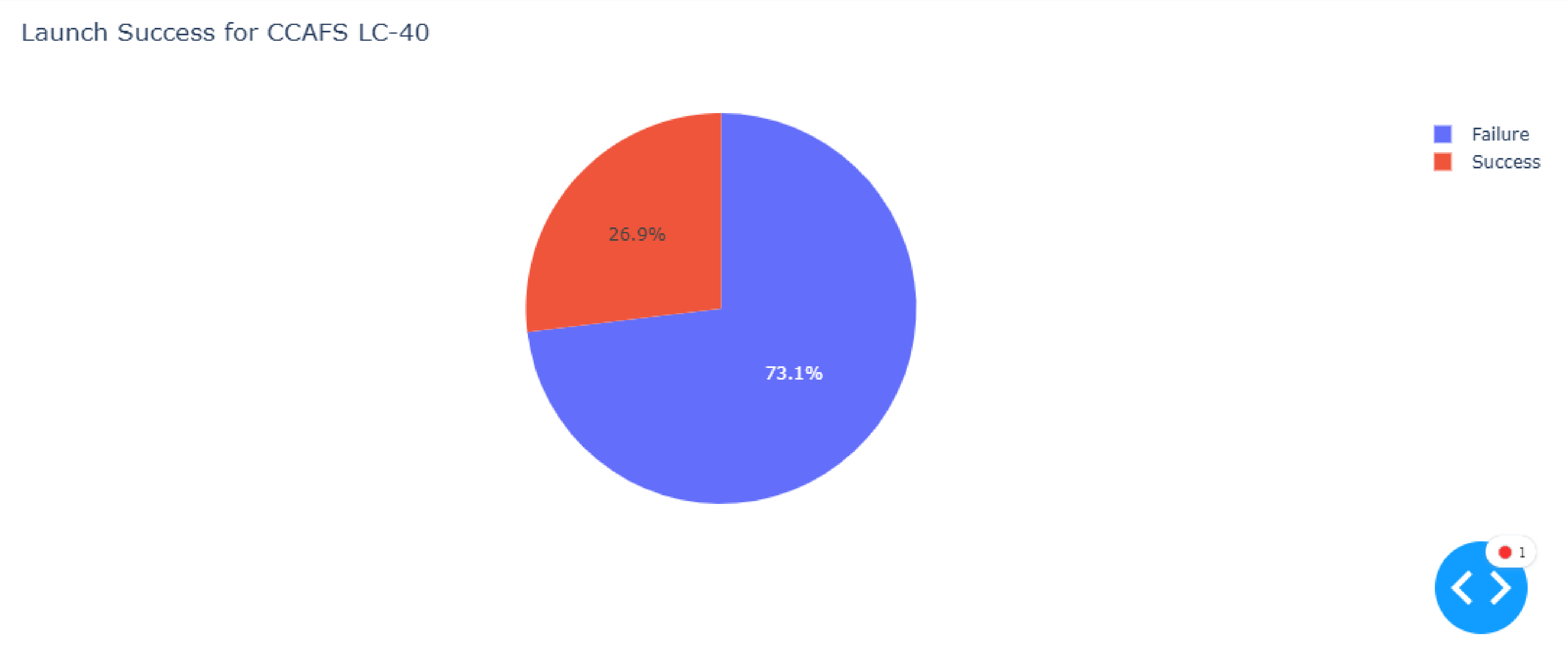
Obstacle proximity map

The map also shows obstacles in close proximity such as airports, or coastlines. By calculating these distances, we gain insights into site efficiency, safety, and logistics advantages.

Section 4

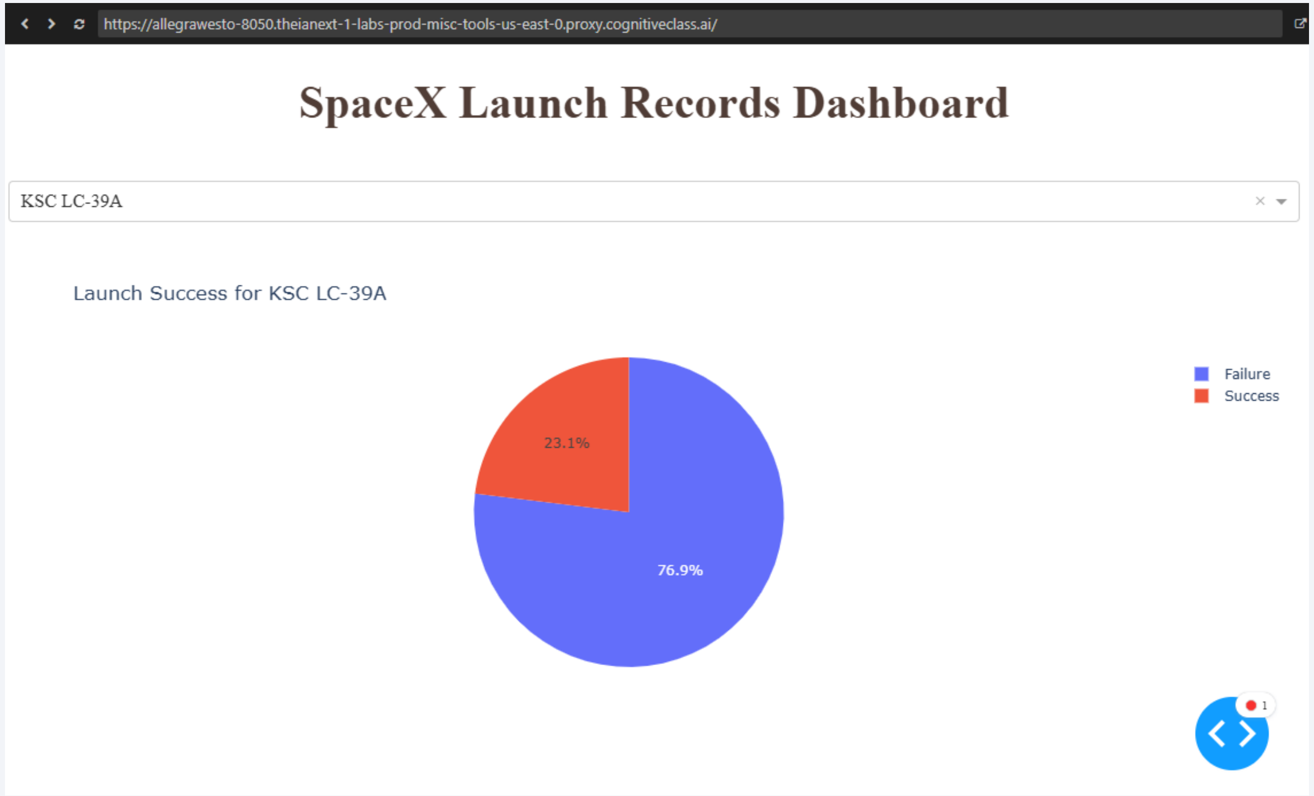
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



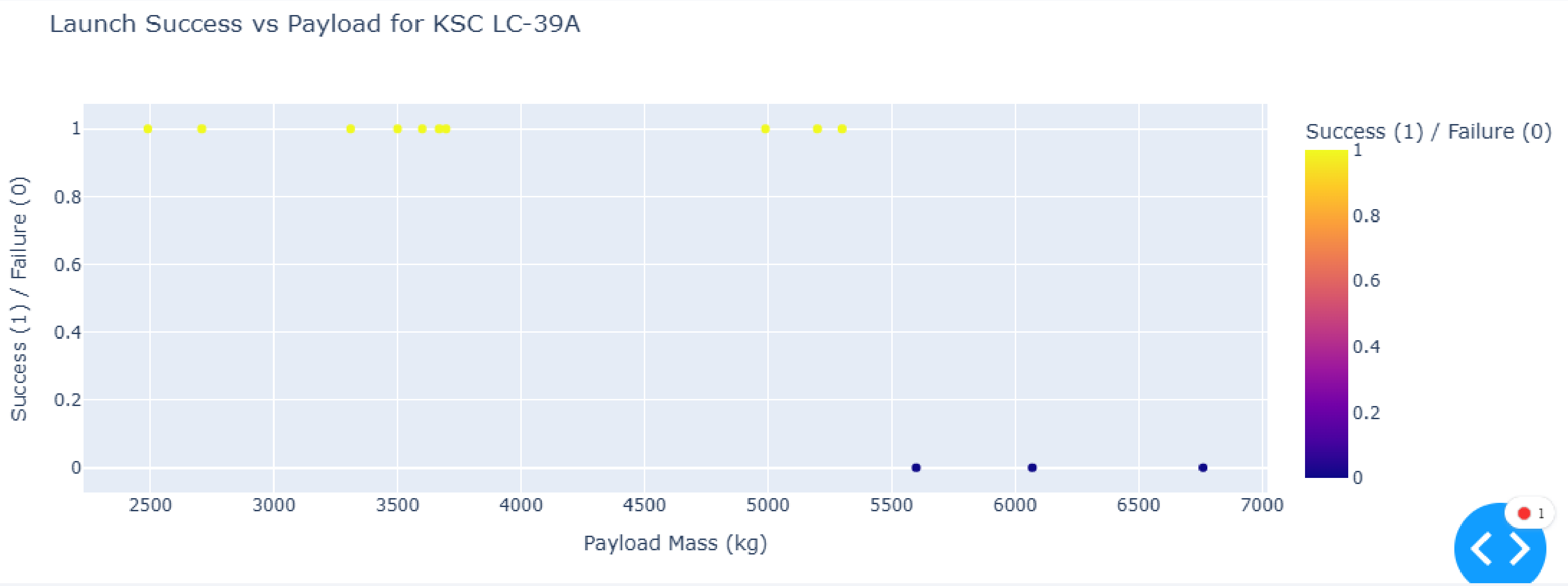
This section presents a pie chart illustrating the number of successful launches at each launch site.

Highest Success LaunchSite



The pie chart visualizes the success rate of launches for the SpaceX launch site with the highest launch success ratio.

<Dashboard Screenshot 3>



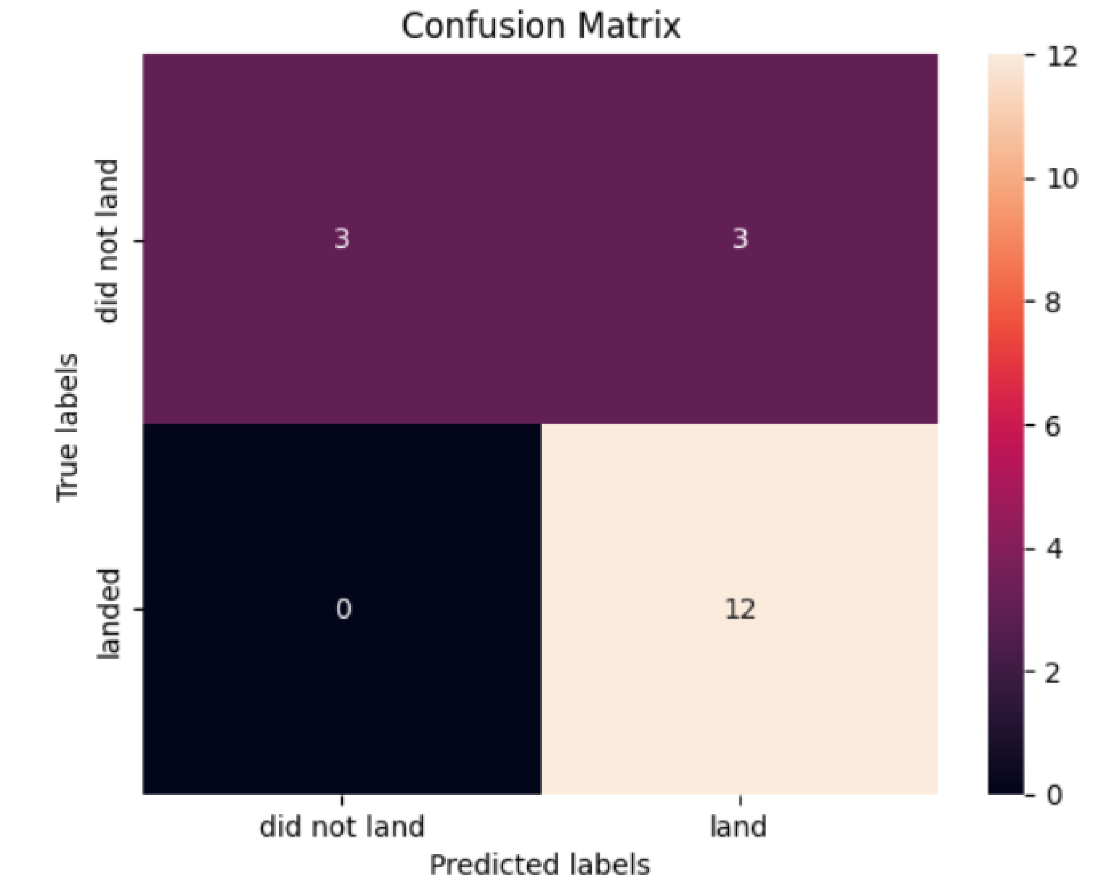
This section presents the Payload vs. Launch Outcome scatter plot for all SpaceX launch sites, using different payload ranges selected in the range slider.

Section 5

Predictive Analysis (Classification)

Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's classification performance by comparing predicted vs. actual outcomes.



Conclusions

The following Conclusions can be made from the findings in this report:

- SpaceX launch success rates vary across sites, with some locations performing better than others.
- Payload mass plays a role in determining launch outcomes, as seen in the scatter plot analysis.
- The interactive dashboard enables users to filter data and explore site-specific trends effectively.
- Understanding these patterns can aid in optimizing future launch strategies for higher success rates.
- Further analysis could include additional factors like weather conditions or rocket type to refine predictions.

Thank you!