IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

UGBOVO YOMA P
20TH JANUARY 2025

Outline

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

Executive Summary

- Summary of Methodologies
- **Data Collection:**
- Utilized the *SpaceX* REST API to fetch launch data.
- Complemented the dataset with web scraping from Wikipedia.
- **Data Wrangling:**
- Filtered out Falcon 1 launches to focus solely on Falcon 9.
- Addressed missing values,.
- **Exploratory Data Analysis (EDA):**
- Conducted visual analysis using charts to uncover trends and relationships.
- Executed SQL queries to derive insights.
- **Interactive Visual Analytics:**
- Created interactive maps using Folium to visualize launch sites and outcomes.
- Built dynamic dashboards with Plotly Dash.
- **Predictive Modeling:**
- Developed classification models (Logistic Regression, Decision Trees, etc.) to predict landing outcomes.

- Summary of Results
- **Insights from Data Analysis:**
- Launch sites such as KSC LC-39A and CCAFS LC-40 emerged as pivotal in *SpaceX's* operations, showing high success rates.
- Payload mass and orbit types demonstrated significant influence on landing outcomes.
- **Interactive Visualizations:**
- Folium maps effectively highlighted geographical patterns in launch site success.
- Plotly Dash dashboards enabled detailed exploration of payload vs.success trends.
- **Predictive Modeling Results:**
- Logistic Regression emerged as the best-performing model, achieving an accuracy of 83.33%.
- Confusion matrix analysis confirmed its reliability in distinguishing successful from failed landings.
- **Actionable Insights:**The findings underscore the importance of payload selection and site optimization in achieving successful launches.

Introduction

- **Project background and context**

SpaceX has been revolutionizing the space industry by reducing the cost of space exploration and improving the success rate of rocket launches. Their reusable rocket technology has set a new benchmark for innovation. This project aims to analyse *SpaceX's* historical launch data to uncover patterns and insights that contribute to their success.

- **Questions to be answered**
- What are the key factors influencing the success or failure of *SpaceX* launches?
- How do payload mass and orbit type impact the likelihood of successful landings?
- Which launch sites contribute the most to *SpaceX's* success?
- Can we predict the success of a launch based on historical data and specific mission parameters?

Section 1

Methodology

Methodology

1. **Data Collection:**

- REST API: Data from *SpaceX* endpoint api.spacexdata.com/v4/launches/past with additional data from /rockets, /payloads, and /launchpads.
- Web Scraping: HTML tables from Falcon 9 Wiki pages cleaned and integrated into pandas dataframes.

2. Data Wrangling:

- Filtered out Falcon 1 launches to focus on Falcon 9.
- Replaced missing PayloadMass values with the mean.

3. Exploratory Data Analysis (EDA):

Visualized data with bar charts, pie charts, scatter plots, and SQL queries for trends and relationships.

4. Interactive Visual Analytics:

Folium: Interactive maps with launch site markers, radii, and trajectories.

- Plotly Dash: Dashboard for filtering by launch sites, payload mass, and visualizing success rates.

5. Predictive Analysis:

Built classification models to predict landing outcomes.

- Tuned models using GridSearchCV and evaluated with accuracy and F1-score.

6. Model Selection:

Compared models via cross-validation.

- Visualized accuracy and selected the best-performing model using confusion matrix analysis.

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX's Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

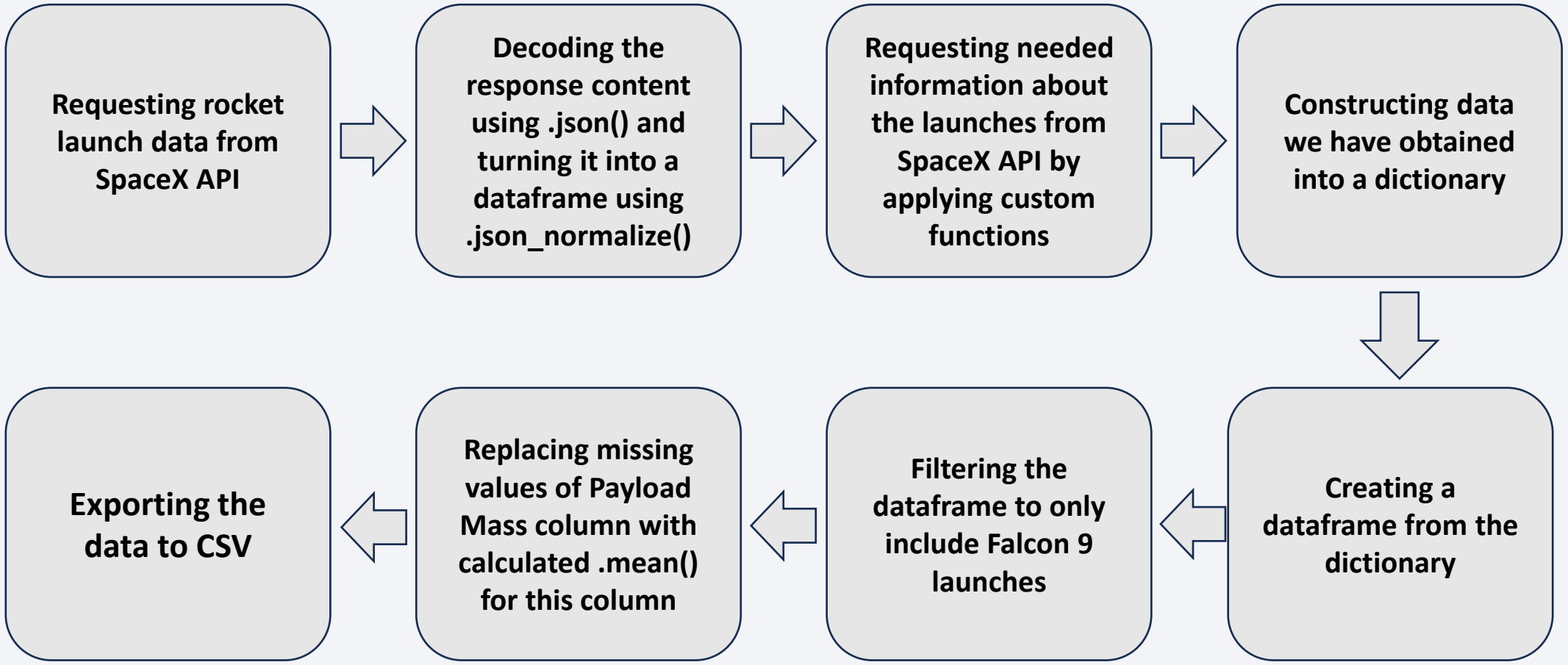
Data Columns are obtained by using SpaceX REST API:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Columns are obtained by using Wikipedia Web Scraping:

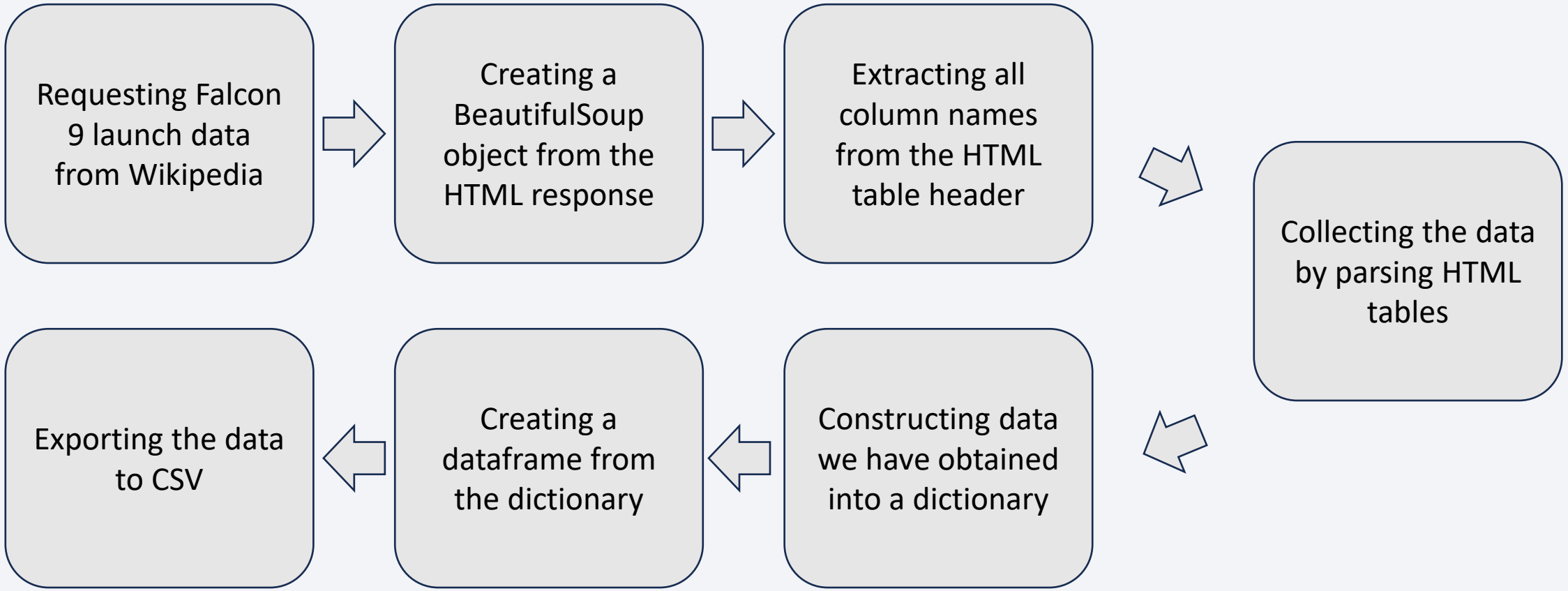
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API



GitHub URL: Data Collection API

Data Collection - Scraping

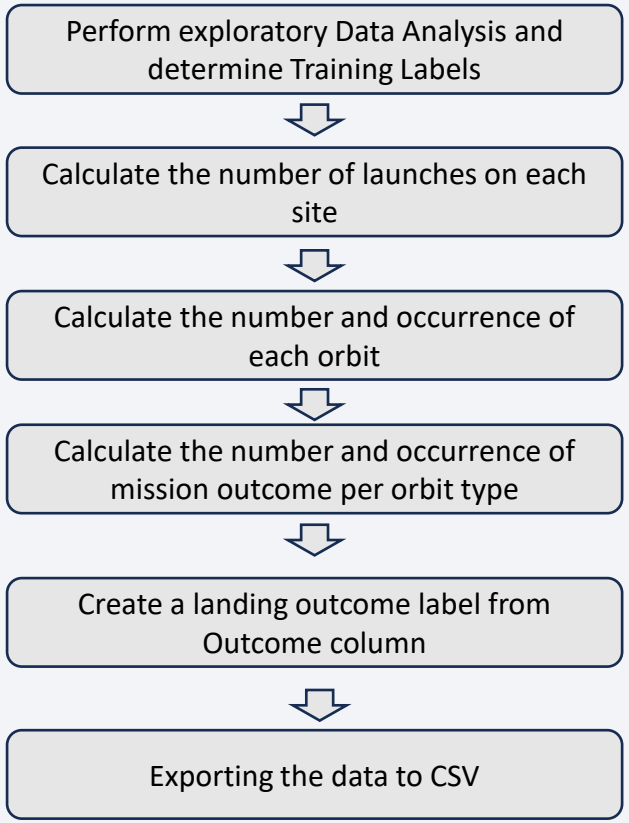


GitHub URL: Data Collection with Web Scraping

Data Wrangling

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

We mainly convert those outcomes into Training Labels with “1” means the booster successfully landed, “0” means it was unsuccessful.



GitHub URL: Data Wrangling

EDA with Data Visualization

- Charts were plotted:

  Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit Type vs. Success Rate, Flight Number vs. Orbit Type, Payload Mass vs Orbit Type and Success Rate Yearly Trend

  Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.

  Bar charts show comparisons among discrete categories. The goal is to show the relationship between the specific categories being compared and a measured value.

  Line charts show trends in data over time (time series).

GitHub URL: EDA with Data Visualization

EDA with SQL

- Performed SQL queries:
    - Displaying the names of the unique launch sites in the space mission
    - Displaying 5 records where launch sites begin with the string 'CCA'
    - Displaying the total payload mass carried by boosters launched by NASA (CRS)
    - Displaying average payload mass carried by booster version F9 v1.1
    - Listing the date when the first successful landing outcome in ground pad was achieved
    - Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - Listing the total number of successful and failure mission outcomes
    - Listing the names of the booster versions which have carried the maximum payload mass
    - Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
    - Ranking 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: EDA with SQL

Build an Interactive Map with Folium

Markers of all Launch Sites:

- Added Marker with Circle, Popup Label and Text Label of NASA Johnson Space Center using its latitude and longitude coordinates as a start location.

- Added Markers with Circle, Popup Label and Text Label of all Launch Sites using their latitude and longitude coordinates to show their geographical locations and proximity to Equator and coasts.

Coloured Markers of the launch outcomes for each Launch Site:

- Added coloured Markers of success (Green) and failed (Red) launches using Marker Cluster to identify which launch sites have relatively high success rates.

Distances between a Launch Site to its proximities:

- Added coloured Lines to show distances between the Launch Site KSC LC-39A (as an example) and its proximities like Railway, Highway, Coastline and Closest City.

Build a Dashboard with Plotly Dash

Launch Sites Dropdown List:

- Added a dropdown list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

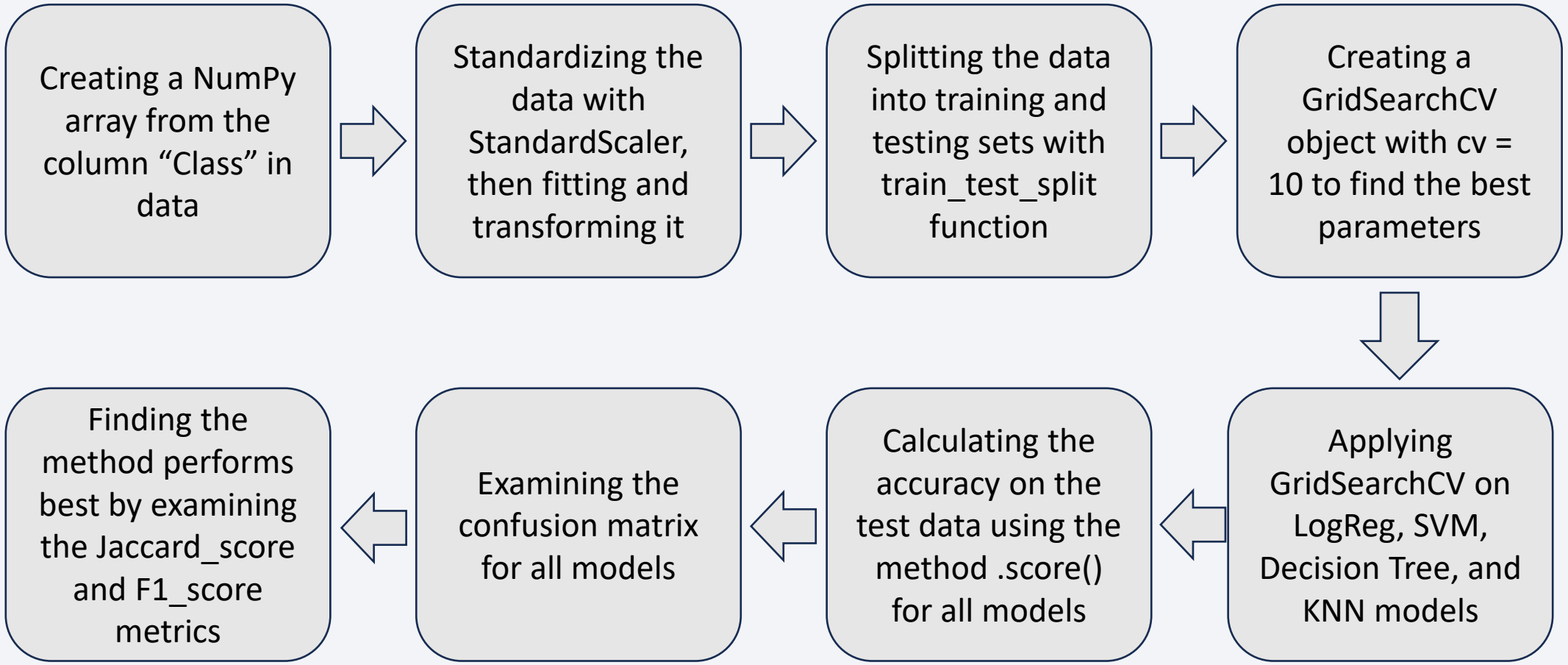
Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

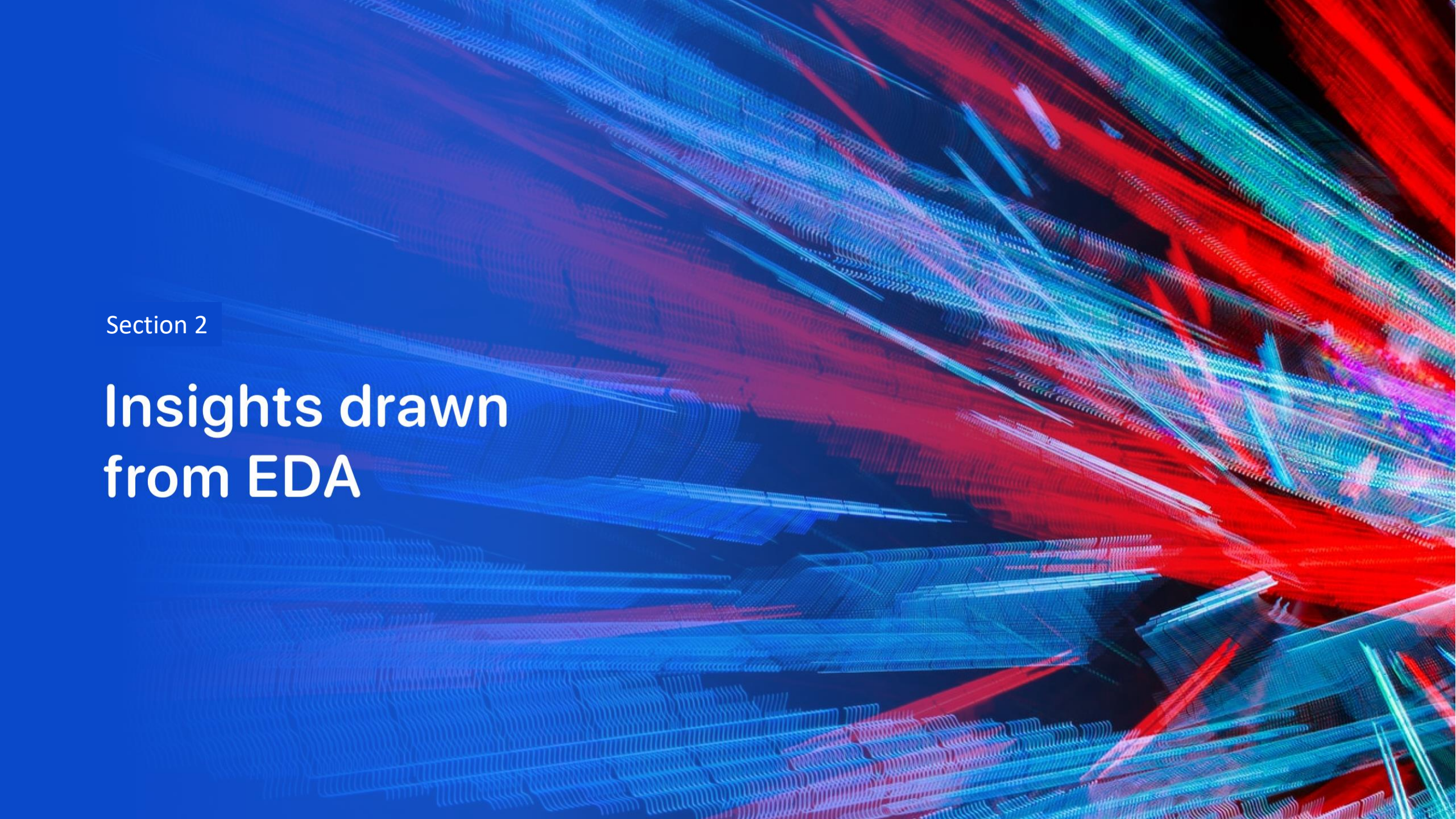
Predictive Analysis (Classification)



GitHub URL: Machine Learning Prediction

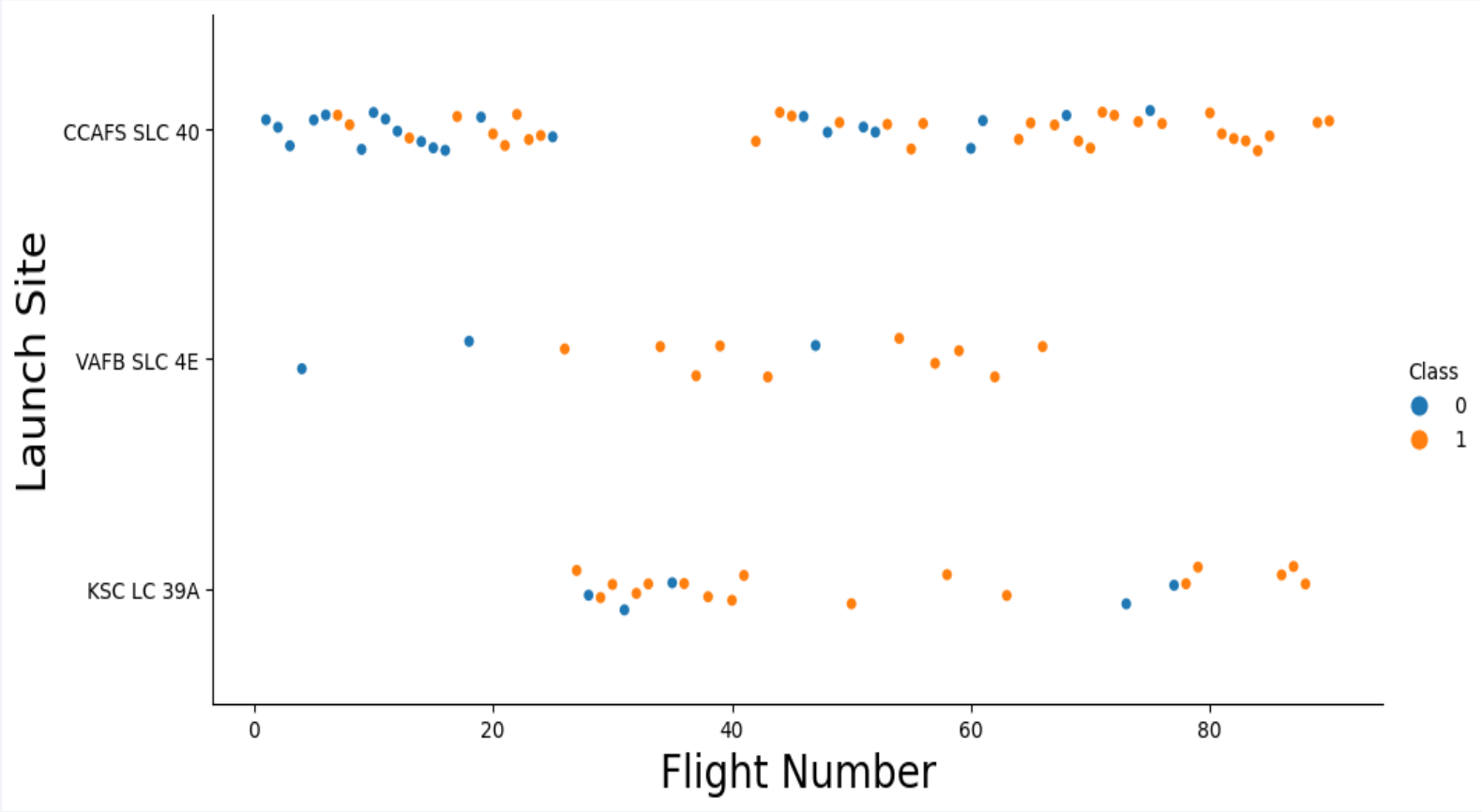
Results

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


Section 2
Insights drawn from EDA

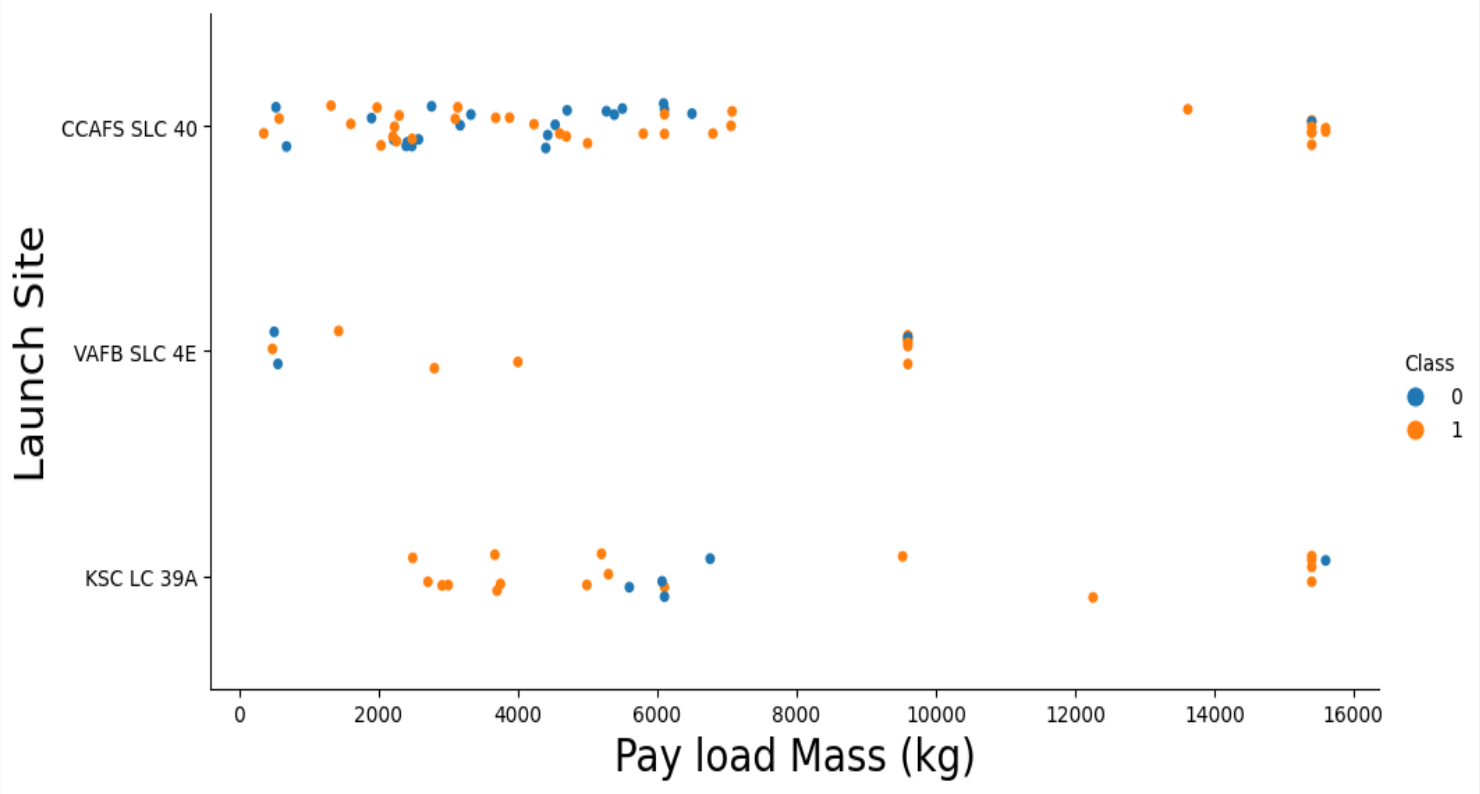
Flight Number vs. Launch Site

This scatter plot demonstrates the relationship between the flight number and the launch site. It highlights the frequency of launches from each site, revealing that CCAFS LC-40 and KSC LC-39A hosted the majority of SpaceX's missions. This insight emphasizes the operational importance of these sites in *SpaceX's* strategy.



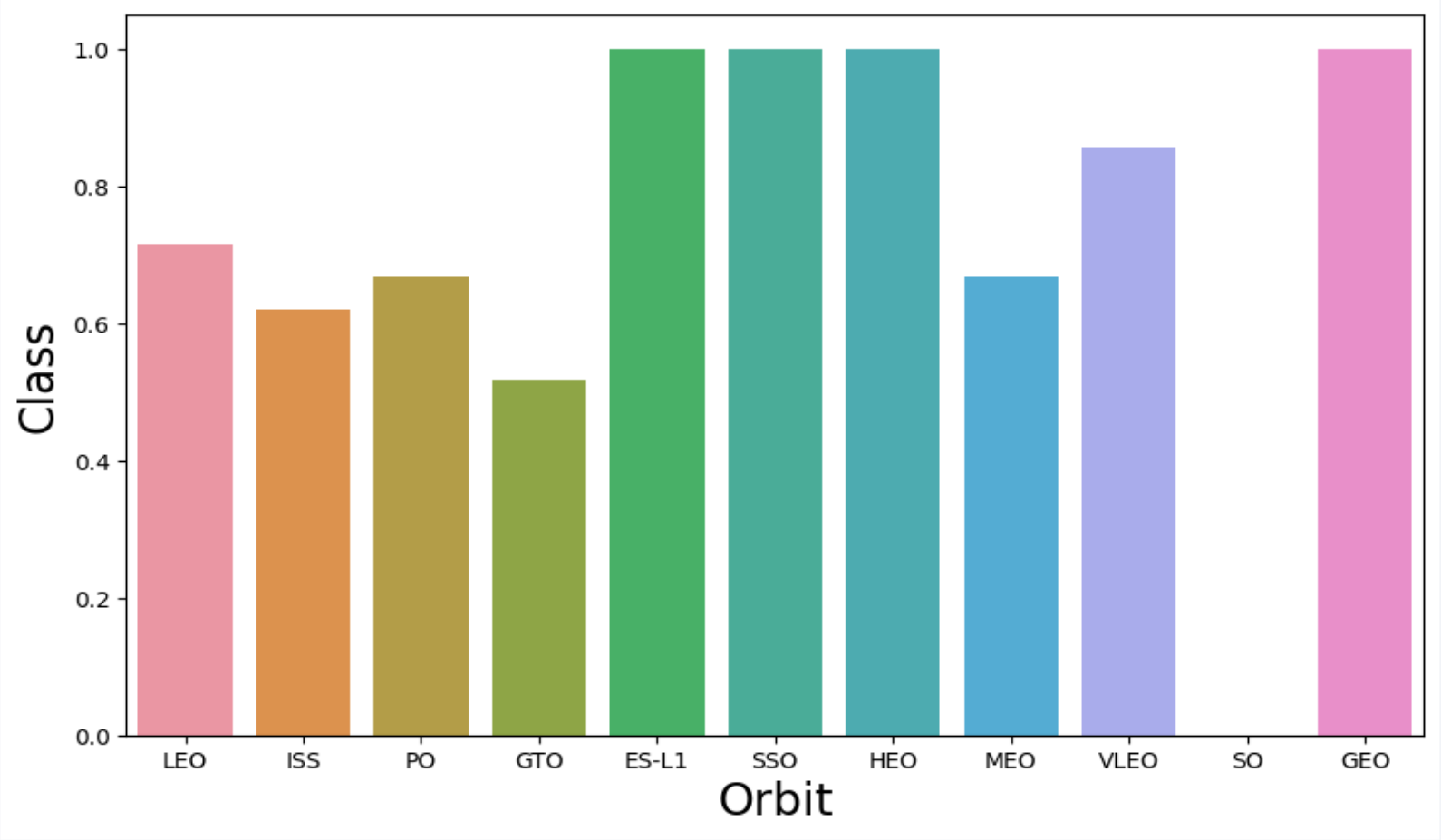
Payload vs. Launch Site

The plot illustrates the payload mass for different launches across *SpaceX's* sites. It shows that heavier payloads were often launched from KSC LC-39A, suggesting that this site is better equipped for missions involving higher payload masses.



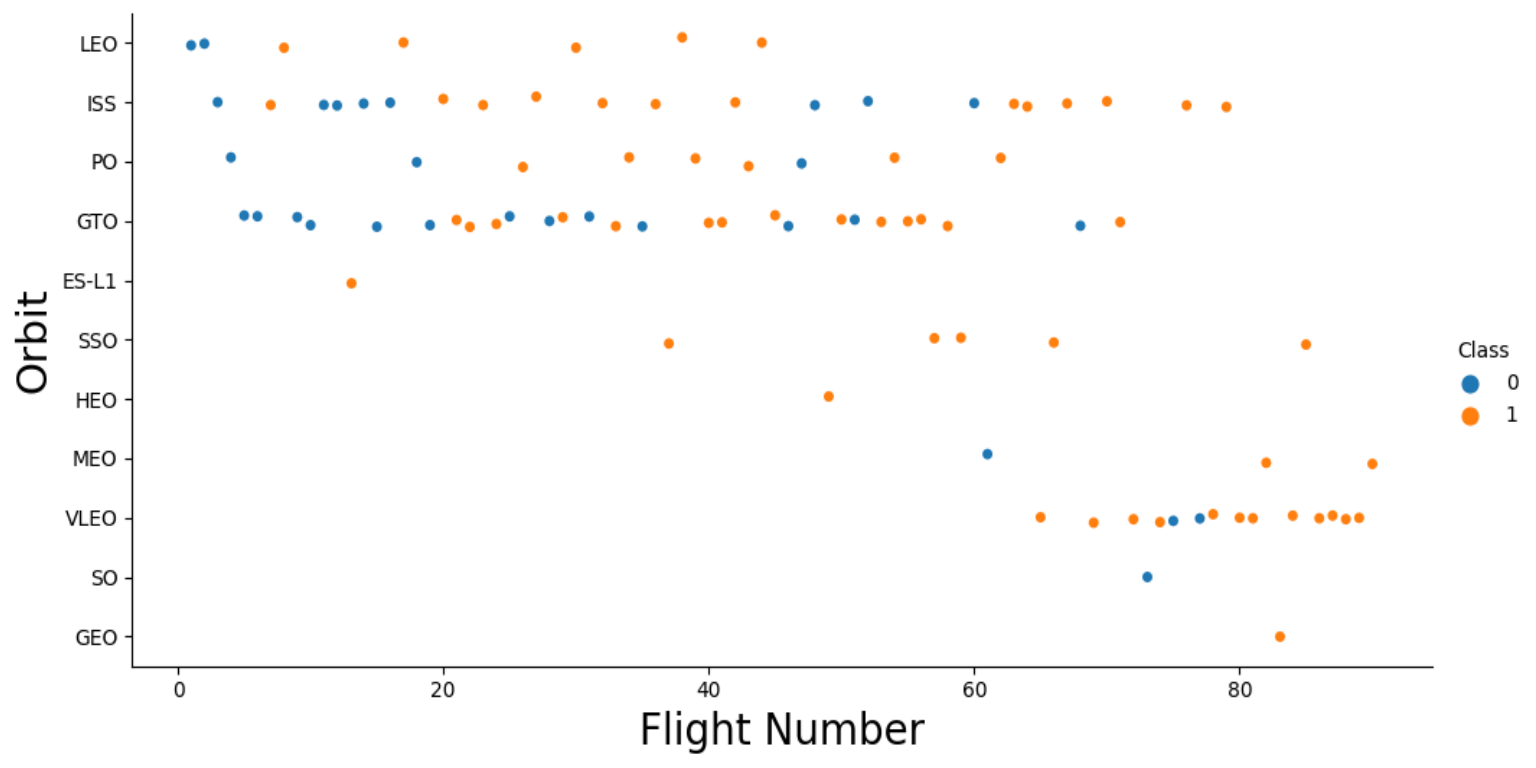
Success Rate vs. Orbit Type

This bar chart showcases the success rate for different orbit types. Geostationary Transfer Orbit (GTO) has a slightly lower success rate compared to Low Earth Orbit (LEO), indicating that GTO missions are more challenging due to higher technical demands.



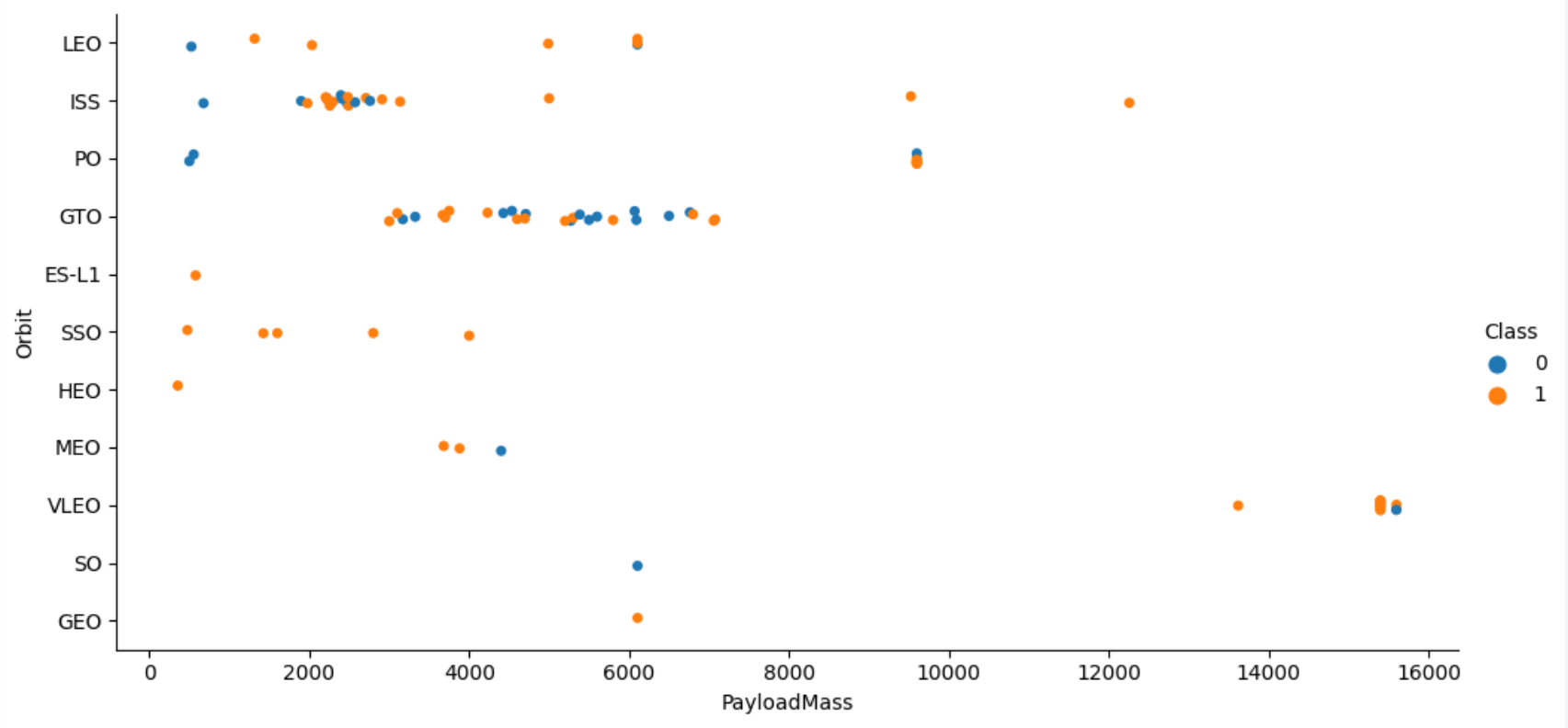
Flight Number vs. Orbit Type

This scatter plot shows how orbit type distribution evolved as *SpaceX* launched more flights. Early missions targeted simpler orbits (LEO), while later flights included more complex missions to GTO and beyond.



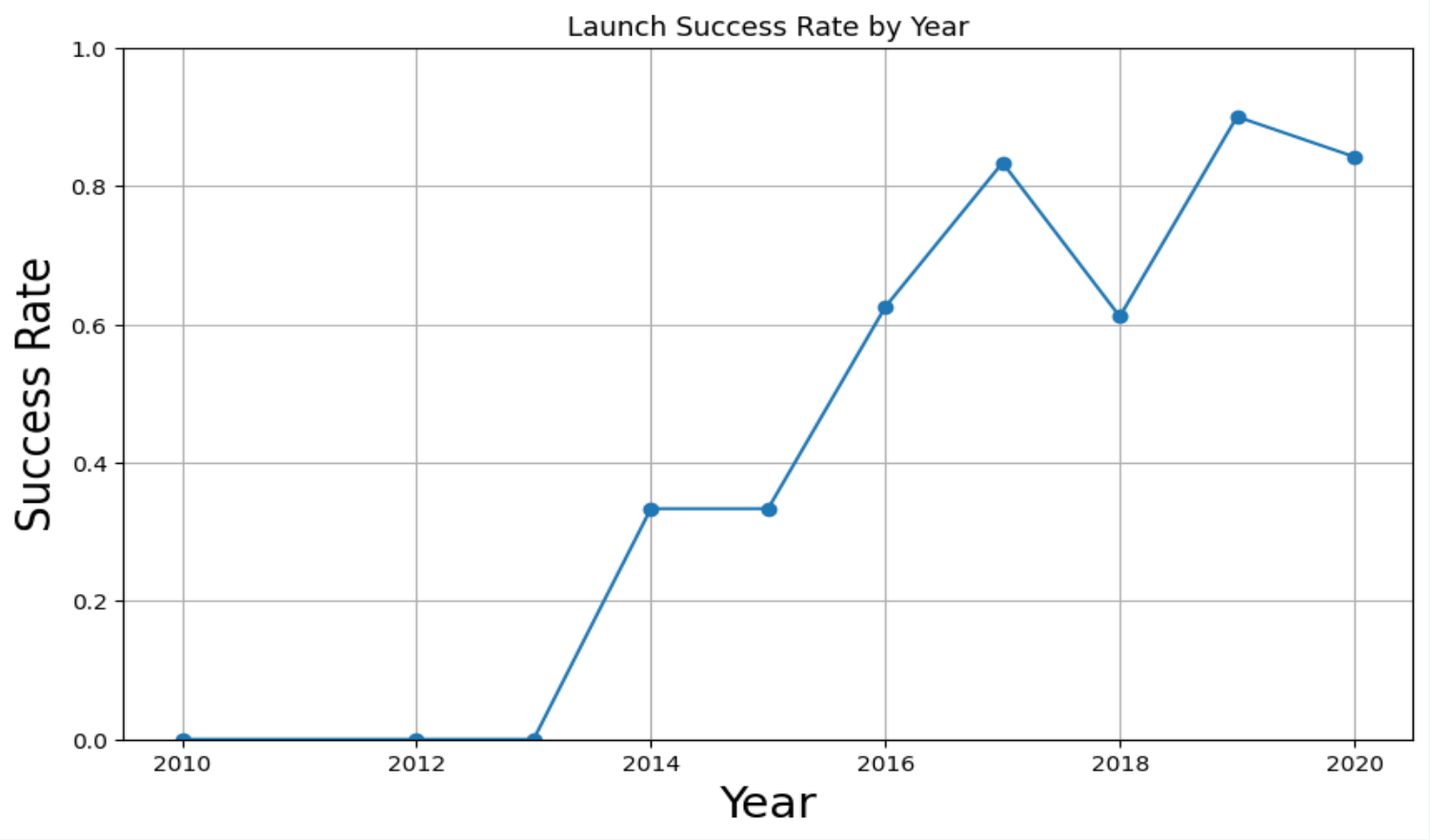
Payload vs. Orbit Type

This scatter plot analyzes how payload mass varies across orbit types. LEO missions can accommodate heavier payloads compared to GTO, reflecting differences in mission requirements and technical constraints.



Launch Success Yearly Trend

The line chart depicts the yearly trend in *SpaceX's* launch success rate. A consistent improvement over time is evident, showcasing the company's learning curve and technological advancements.



All Launch Site Names

```
SELECT
      distinct(Launch_Site)
FROM
      SPACEXTABLE
```

Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

Explanations

This query retrieves the unique names of all launch sites where SpaceX missions have taken place. The result shows sites such as CCAFS LC-40 and KSC LC-39A, highlighting their significance in SpaceX's operations.

Launch Site Names Begin with 'CCA'

```
SELECT
        *
FROM
        SPACEXTABLE

WHERE
        Launch_Site like 'CCA%’  limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Custome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS

Explanations

This query filters launch sites whose names start with ‘CCA’, limiting the result to the first five records. It focuses on sites such as CCAFS LC-40 and CCAFS SLC-40 to analyze their operational patterns.

Total Payload Mass

```
SELECT
        SUM(PAYLOAD_MASS__KG_) as TOTAL_PAYLOAD_MASS_KG_
FROM
        SPACEXTABLE
WHERE
        Customer = 'NASA (CRS)'
```

TOTAL_PAYLOAD_MASS_KG_
45596

Explanations

The query calculates the total payload mass carried by boosters for NASA (CRS) missions. This provides insights into NASA's dependency on SpaceX for delivering heavy payloads.

Average Payload Mass by F9 v1.1

```
SELECT
        AVG(PAYLOAD_MASS__KG_) as AVG_PAYLOAD_MASS_KG_
FROM
        SPACEXTABLE
WHERE
        Booster_Version LIKE 'F9 v1.1%'
```

Date
2015-12-22

Explanations

This query computes the average payload mass for the Falcon 9 v1.1 booster version, offering insights into its operational capacity and performance.

First Successful Ground Landing Date

```
SELECT
        Date
FROM
        SPACEXTABLE
WHERE
        Landing_Outcome LIKE 'Success%' order by Date asc limit 1
```

AVG_PAYLOAD_MASS_KG_
2534.666666666665

Explanations

This query identifies the date of SpaceX's first successful ground landing. It showcases a milestone in SpaceX's reusability goals and technological achievements.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT
        Booster_Version
FROM
        SPACEXTABLE
WHERE
        Landing_Outcome == 'Success (drone ship)’
        AND PAYLOAD_MASS__KG_ > 4000
        AND PAYLOAD_MASS__KG_ < 6000
```

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

Explanations

This query lists the boosters that successfully landed on a drone ship while carrying payloads between 4000 and 6000 kg. It demonstrates SpaceX’s precision in handling medium payloads.

Total Number of Successful and Failure Mission Outcomes

```
SELECT
        Mission_Outcome, COUNT(*) as Count
FROM
        SPACEXTABLE
GROUP BY
        Mission_Outcome
```

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

```
SELECT
        COUNT(CASE WHEN Landing_Outcome LIKE 'Success%' THEN 1 END) AS
        SuccessCount,
        COUNT(CASE WHEN Landing_Outcome LIKE 'Failure%' THEN 1 END) AS FailureCount
FROM
        SPACEXTABLE;
```

Success	Failures
61	10

Boosters Carried Maximum Payload

```
SELECT
        DISTINCT Booster_Version
FROM
        SPACEXTABLE
WHERE
        PAYLOAD_MASS__KG_ = (
SELECT
        MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE)
```

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

Explanations

This query identifies the boosters that carried the heaviest payloads, showcasing their maximum capacity and reliability.

2015 Launch Records

```
SELECT
        substr(Date, 6,2) as month, Landing_Outcome, Booster_Version, Launch_Site
FROM
        SPACEXTABLE
WHERE
        substr(Date,0,5)='2015' AND Landing_Outcome like 'Failure (drone ship)’
ORDER BY
        month;
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Explanations

This query retrieves records of failed drone ship landings in 2015, providing insights into early challenges faced by SpaceX.

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

```
SELECT
        Landing_Outcome, COUNT(Landing_Outcome) as Count
FROM
        SPACEXTABLE
WHERE
        Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY
        Landing_Outcome
ORDER BY
        Count DESC
```

Landing_Outcome	Count
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

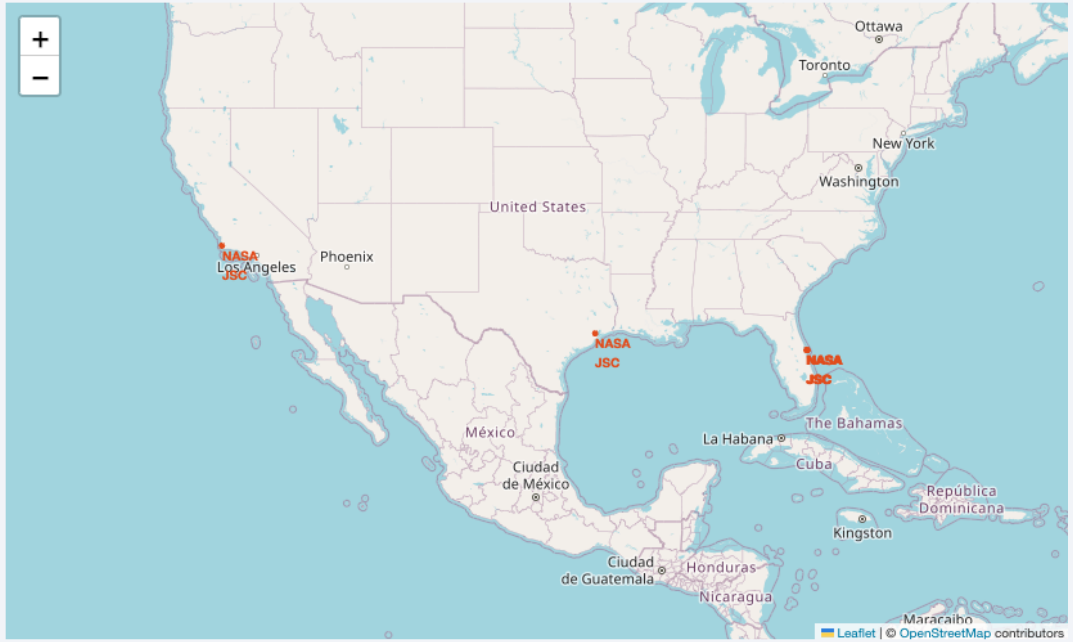
Explanations

This query ranks landing outcomes during a specific timeframe, showing SpaceX's progress and challenges over the years.

Section 3

Launch Sites Proximities Analysis

<Folium Map Screenshot 1>

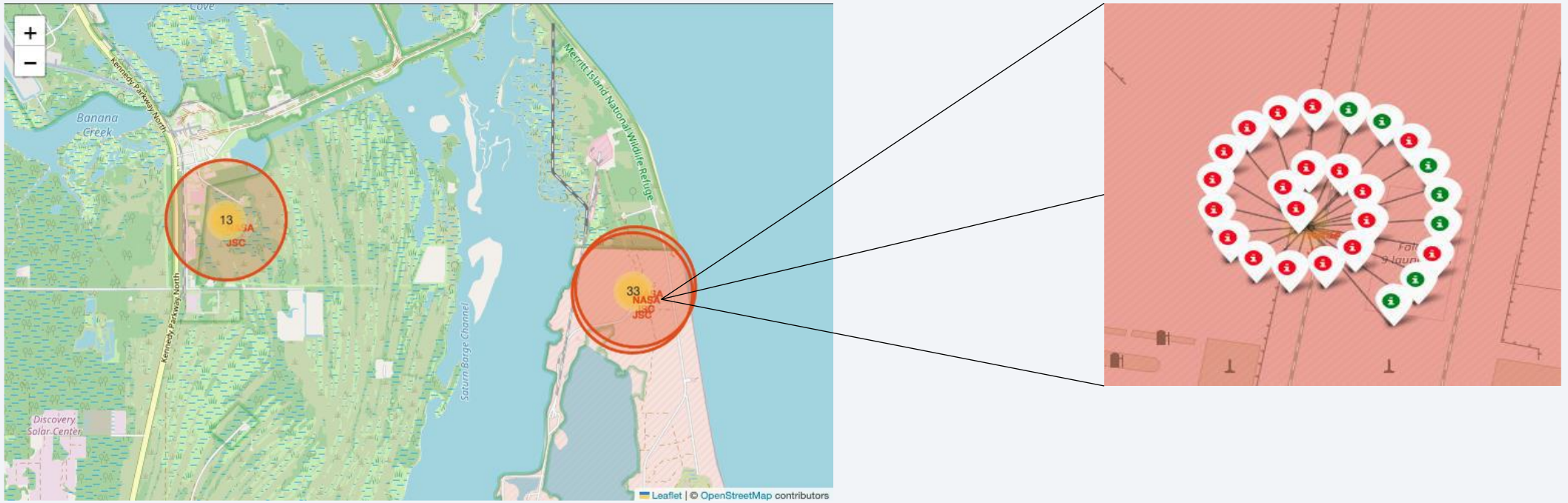


Launch Site	Lat	Long
CCAFS LC-40	28.562302	-80.577356
CCAFS SLC-40	28.563197	-80.576820
KSC LC-39A	28.573255	-80.646895
VAFB SLC-4E	34.632834	-120.610745

Explanations

This interactive map marks all *SpaceX* launch sites. Key locations such as CCAFS LC-40 and KSC LC-39A are highlighted, providing a geographical overview of their strategic positions near the coastline.

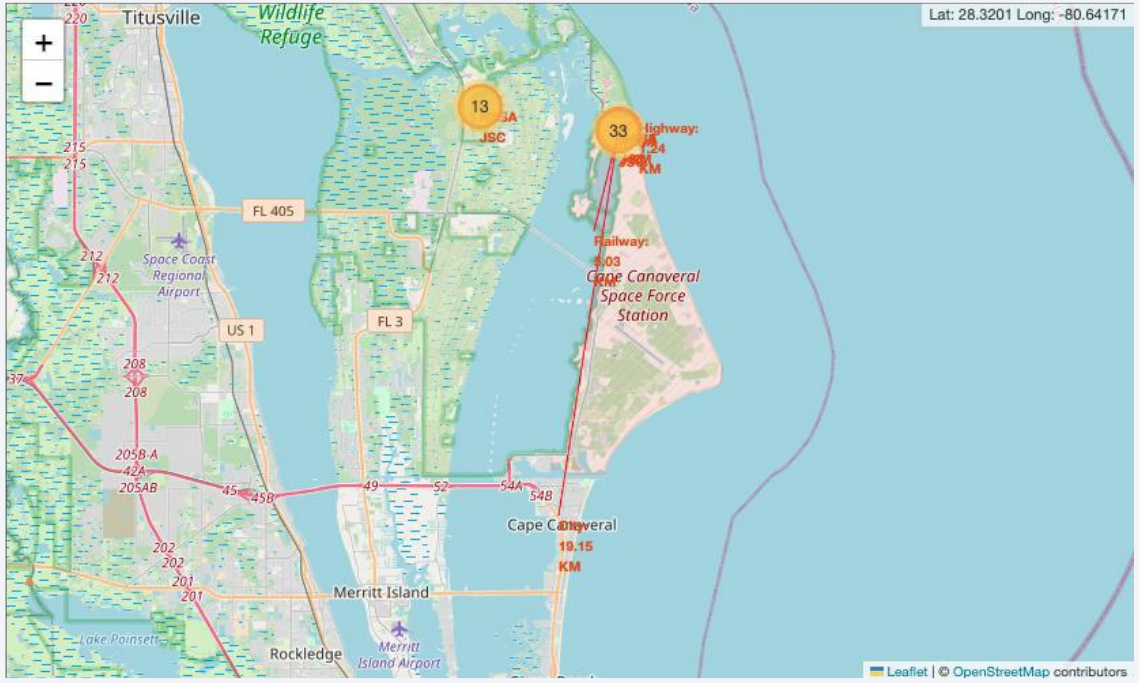
<Folium Map Screenshot 2>



Explanations

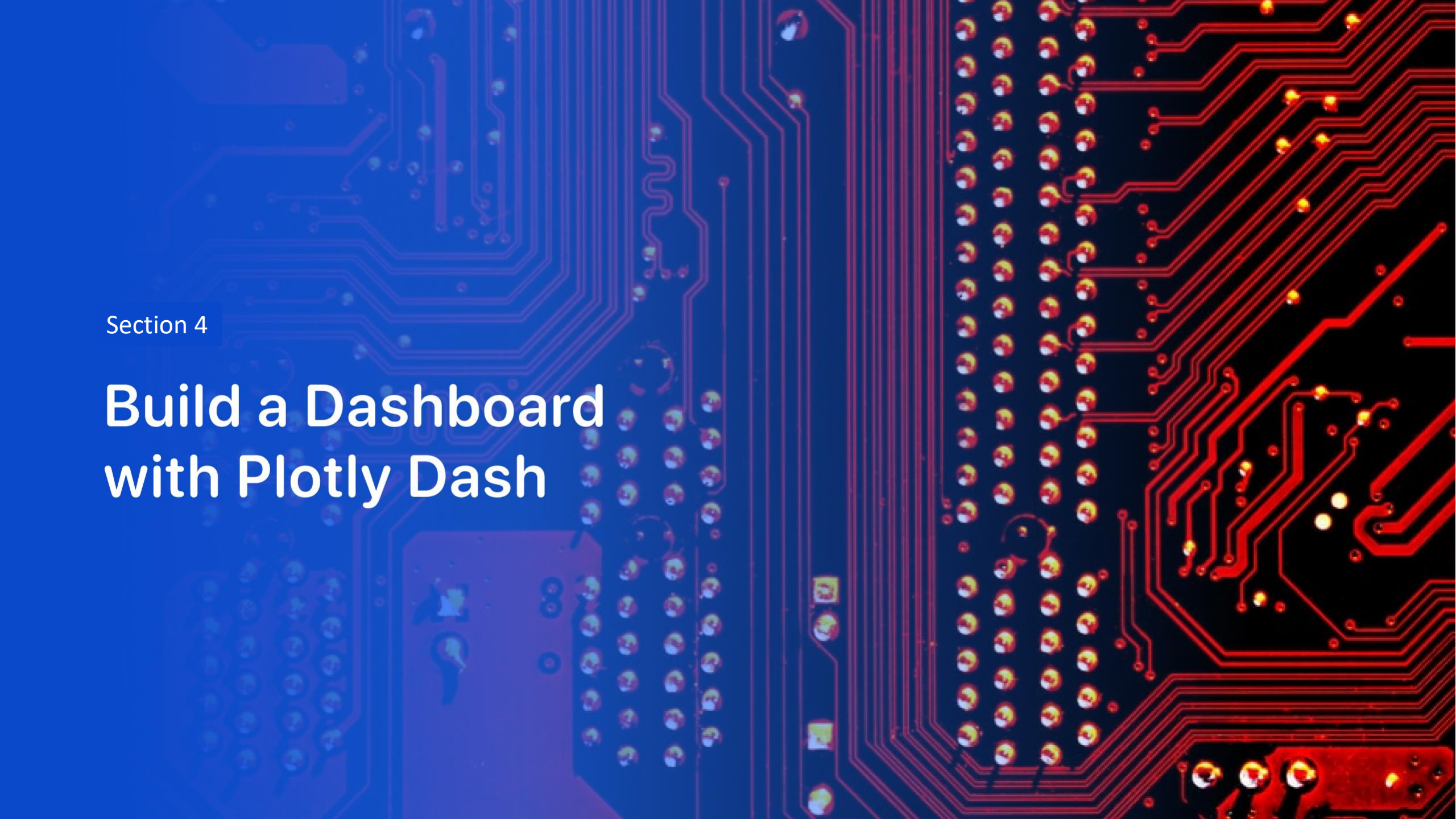
The color-coded markers indicate the success or failure of launches from each site. The clustering of successful launches around KSC LC-39A reflects its reliability and advanced infrastructure.

<Folium Map Screenshot 3>

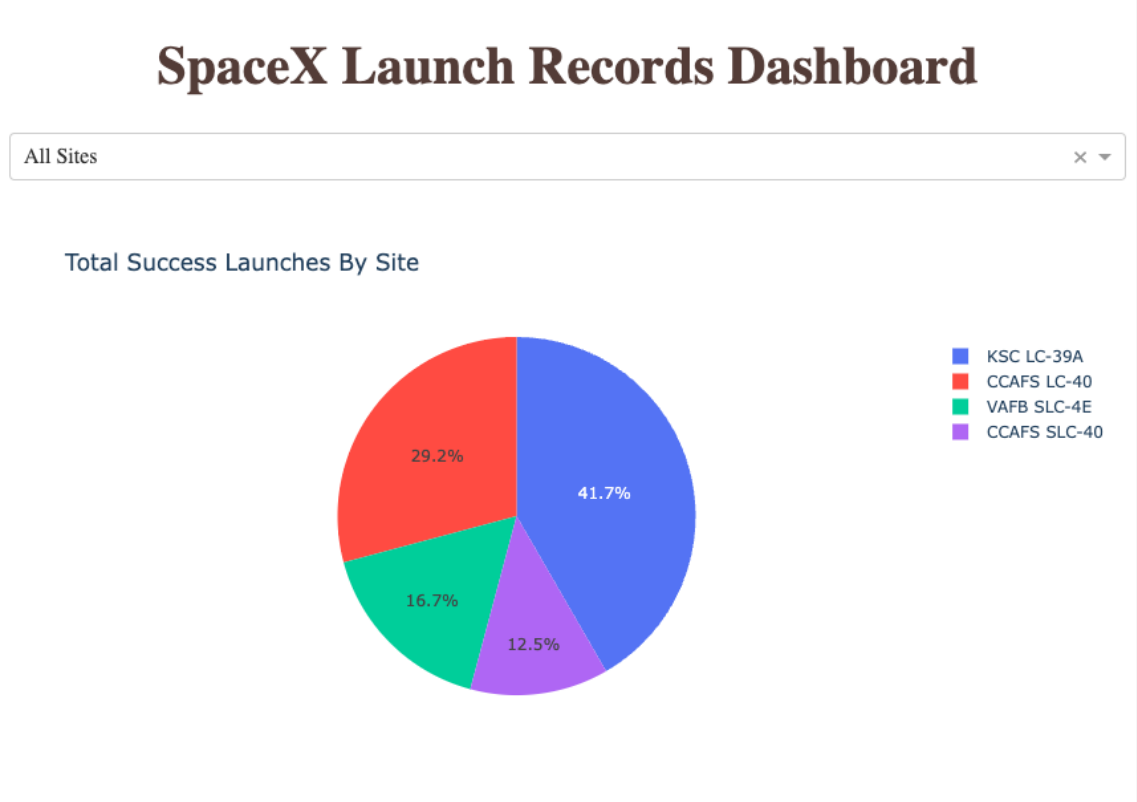


Explanations

This map visualizes the distances between launch sites and nearby infrastructure such as highways, railways, and coastlines. The proximity analysis aids in understanding logistical and safety considerations for site selection.


Section 4
Build a Dashboard with Plotly Dash

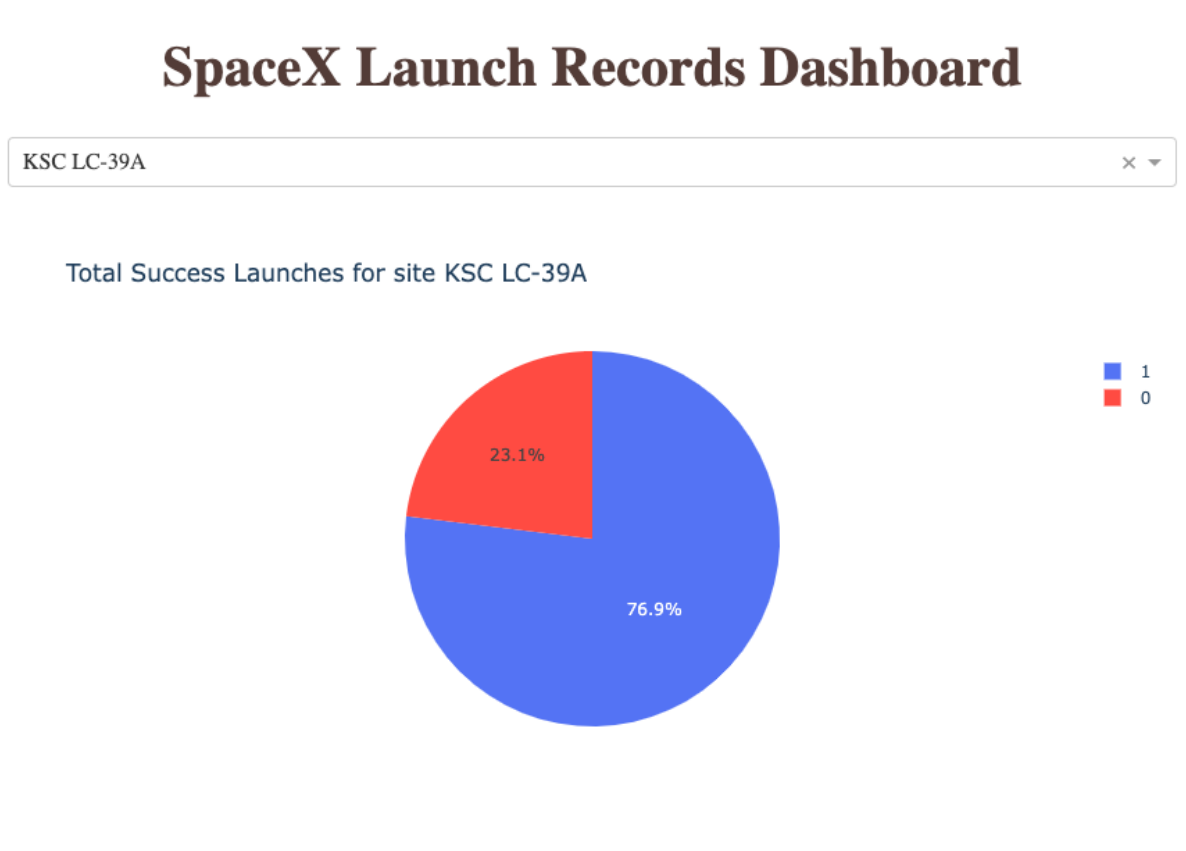
<Dashboard Screenshot 1>



Explanations

The pie chart displays the success rate of launches across all sites. KSC LC-39A leads in successful launches, emphasizing its pivotal role in *SpaceX's* operations.

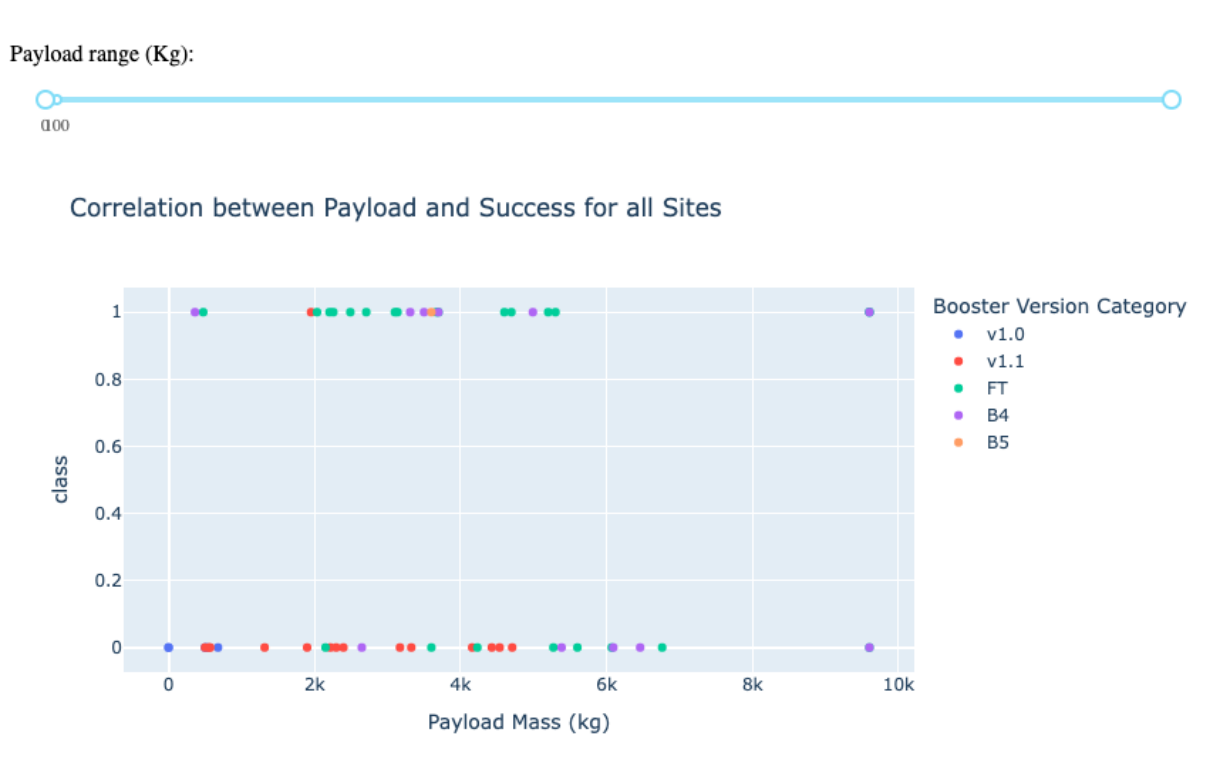
<Dashboard Screenshot 2>



Explanations

This pie chart focuses on the launch site with the highest success ratio. It confirms that KSC LC-39A has the most efficient setup, achieving consistent success over multiple launches.

<Dashboard Screenshot 3>



Explanations

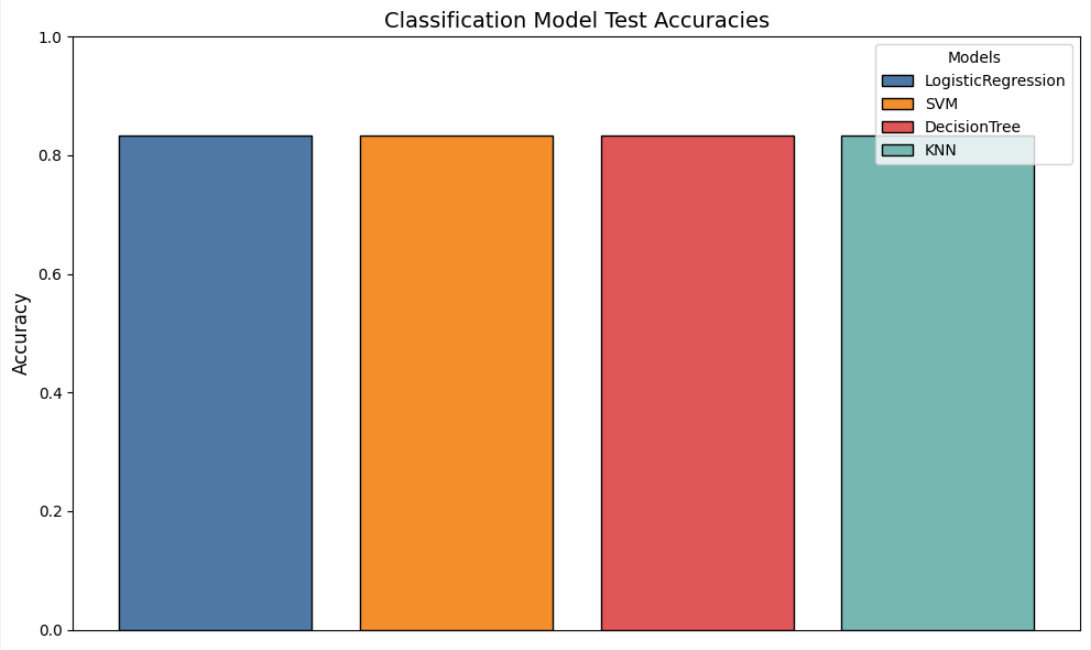
The scatter plot shows the relationship between payload mass and launch outcome for all sites. Higher payloads correlate with a slightly reduced success rate, particularly for complex missions to GTO.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Modelo	Exactitud de Prueba
LogisticRegression	0.8333
SVM	0.8333
DecisionTree	0.8333
KNN	0.8333



Best Model:

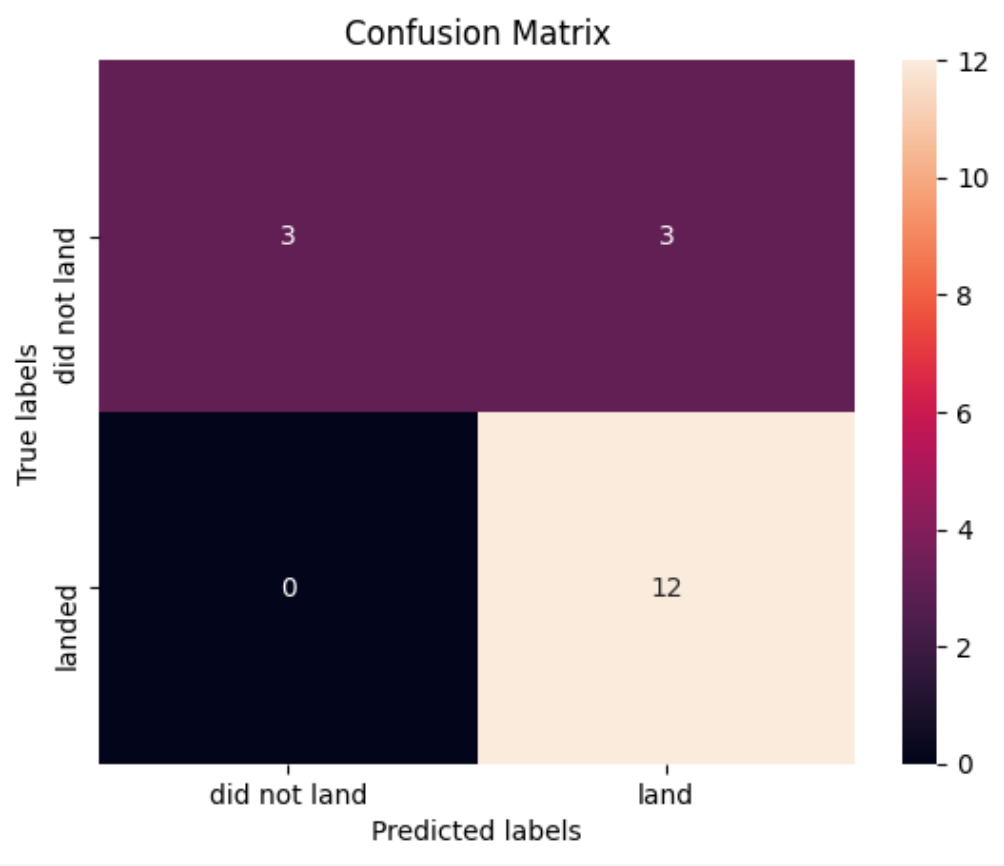
The ***KNN*** model was selected as the best-performing model for the following reasons:

Consistency: It maintained stable performance across all validation datasets.

Simplicity: KNN is computationally less expensive compared to other models like SVM or Random Forests, making it more efficient for real-time predictions.

Interpretability: The coefficients in KNN provide clear insights into feature importance, which aids in understanding the factors influencing landing outcomes.

Confusion Matrix



The confusion matrix above represents the performance of the best-performing classification model, likely Logistic Regression. :

True Positives (Bottom-Right, 12 instances): This high count indicates the model's strength in identifying successful landings.

True Negatives (Top-Left, 3 instances): These results confirm the model's ability to recognize failures.

False Positives (Top-Right, 3 instances): Such misclassifications suggest that the model may overestimate the likelihood of success in borderline cases.

False Negatives (Bottom-Left, 0 instances): Absence of false negatives indicates that the model successfully avoided misclassifying successful landings as failures.

Conclusions

Key Takeaways:

- The integration of API and web-scraped data proved effective in creating a reliable dataset for analysis.
- Interactive visualizations and predictive models provided actionable insights into *SpaceX's* launch strategies.

Recommendations:

- Further exploration of other machine learning models may improve predictive accuracy.
- Enhancing data quality with more granular payload and orbit details could refine analyses.

Future Work:

- Expanding the analysis to include real-time data updates from the *SpaceX* API.
- Leveraging additional visualization techniques to better communicate findings to stakeholders.

Appendix

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

Thank you!