IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Daniel Cristóbal
16/03/2025

Outline

NATIVIDAD 1956

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

Executive Summary

NATIVIDAD 1956

- In this capstone, various methodologies were used to predict whether the first stage of SpaceX's Falcon 9 rocket will land successfully.

- This project demonstrated that it is possible to accurately predict the landing success of the Falcon 9 first stage using machine learning techniques. The results indicate that payload mass, orbit type, and launch site are key factors influencing landing success. Furthermore, the ability to predict landing success allows for estimating launch costs, which is useful for companies looking to compete with SpaceX in the space launch market.

Introduction

The project aims to predict whether the first stage of SpaceX's Falcon 9 rocket will land successfully. This is crucial because first stage reusability is a key factor in reducing the cost of space launches. SpaceX offers launches at a cost of $62 million, while other providers can charge upwards of $165 million per launch. The ability to predict landing success makes it possible to estimate the cost of a launch, which is useful for companies looking to compete with SpaceX in the space launch market.

NATIVIDAD 1956

Problems to be solved:

1. Predicting landing success.
2. Analyzing key factors.
3. Cost estimation.

NATIVIDAD 1956

Section 1

Methodology

Methodology

NATIVIDAD 1956

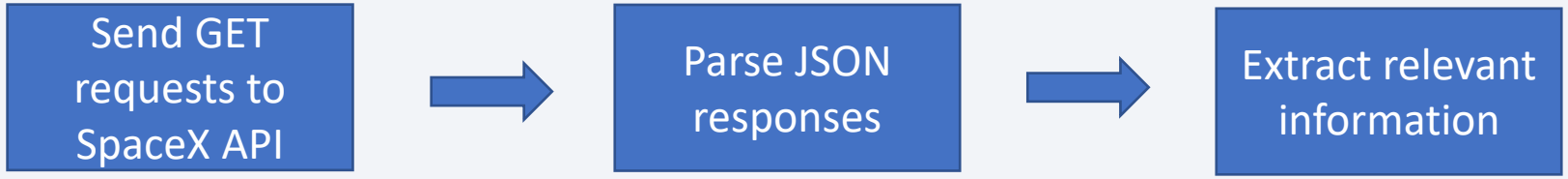
Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping
- Perform data wrangling
  - JSON to dataFrame
  - Data Cleaning
  - Feature Engineering
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data Splitting
  - Model Selection
  - Hyperparameter Tuning
  - Model Evaluation

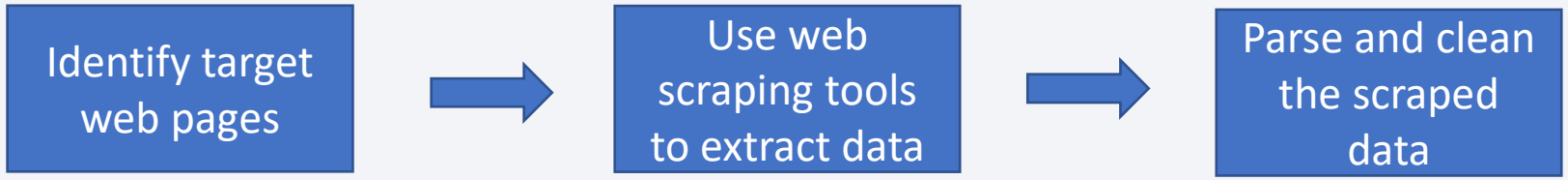
Data Collection

NATIVIDAD 1956

- API Data Retrieval:



- Web Scraping:



Data Collection – SpaceX API

Collection Process:

1. SpaceX API Request

- Base URL: https://api.spacexdata.com/v4/
- Endpoints used: launches/past, rockets, launchpads, payloads, cores

2. Data Processing

- JSON to Pandas DataFrame Conversion
- Falcon 9 Launch Filtering
- Extracting Detailed Information Using Additional Requests

3. Data Cleaning

- Handling Missing Values
- Data Type Conversion

Flowchart

A[Start] --> B[SpaceX API Request]

B --> C[JSON to DataFrame Conversion]

C --> D[Falcon 9 Filtering]

D --> E[Extracting Additional Data]

E --> F[Data Cleaning]

F --> G[Final DataFrame]

G --> H[End]

The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/3.2_Sol_jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Step Description

1.HTTP Request
- •URL: Wikipedia "List of Falcon 9 and Falcon Heavy launches“
- •Using the requests library

2.HTML Parsing
- •Creating a BeautifulSoup object
- •Parsing HTML content

3.Table Localization
- •Identifying the specific launch table

4.Data Extraction
- •Iterating through rows and columns
- •Extracting: date, launch site, payload, etc.

5.Data Cleaning
- •Using helper functions (e.g., date_time, booster_version)
- •Handling inconsistent formats and missing values

6.Creating a DataFrame
- •Organizing data in a dictionary
- •Converting to a Pandas DataFrame

Flowchart

A[Start] --> B[HTTP Request to Wikipedia]

B --> C[HTML Parsing with BeautifulSoup]

C --> D[Locating the Releases Table]

D --> E[Extracting Data by Column]

E --> F[Data Cleaning and Formatting]

F --> G[Creating a Pandas DataFrame]

G --> H[End: Data Ready for Analysis]

The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/3.3_Sol_jupyter-labs-webscraping.ipynb

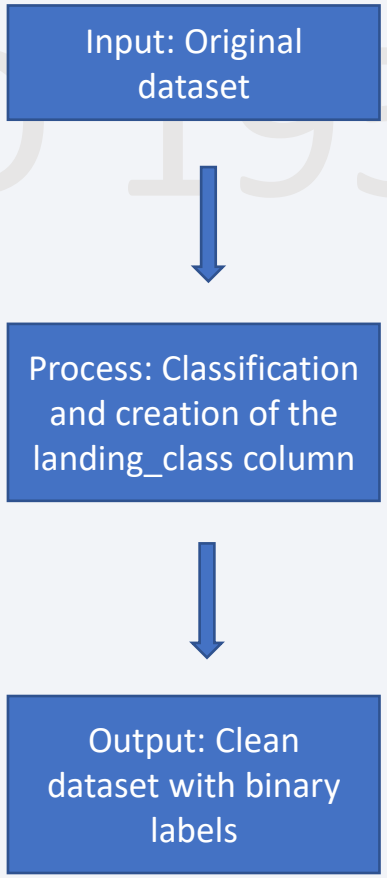
Data Wrangling

Process Description

- Data wrangling is the process of cleaning and transforming data to prepare it for analysis.
- In this case, the data was processed to convert the launch results into binary labels:
  - 1: Successful landing.
  - 0: Failed landing.
- The results were identified and classified as:
  - "True Ocean", "True RTLS", "True ASDS" → Success (1).
  - "False Ocean", "False RTLS", "False ASDS" → Failure (0).

Process Flow

- Data Import: Initial loading of the dataset.
- Result Identification: Classification of successful and failed cases.
- Creation of the landing_class variable
- Transformation Validation



The complete web scraping notebook can be found at the following GitHub URL: https://github.com/seneguehuesca/SpaceY/blob/main/4.2_Sol_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

NATIVIDAD 1956

Summary of the graphs used and their purpose:

- Scatterplots:
  - Analyze the relationship between the number of flights and landing success.
  - Evaluate how payload mass affects success rates in different orbits.
- Bar charts:
  - Compare success rates between different launch sites.
  - Visualize the distribution of successful and unsuccessful missions by orbit.
- Line charts:
  - Show the trend in launch success over time.
- Pie charts:
  - Represent the proportion of successful versus unsuccessful landings on specific platforms.

The types of charts used are appropriate for the questions posed because each provides clarity and specific relevance to data analysis:

- Scatter Plots
  Purpose: Analyze the relationship between two variables, such as the number of flights and landing success, or payload mass and success rates.
- Bar Charts
  Purpose: Compare categories, such as success rates between different launch sites or orbit types.
- Pie Charts
  Purpose: They represent proportions, such as the distribution of successful and unsuccessful landings.
- Line Charts
  Purpose: They analyze temporal trends, such as improvements in success rates over time.

In summary, these charts were selected for their ability to answer specific analysis questions, providing visual clarity, ease of interpretation, and relevance to the variables studied.

The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/2.1_Sol_jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Summary of SQL queries performed:

NATIVIDAD 1956

- Identification of unique launch sites
- Filtering of records by specific launch site
- Calculation of total payload mass for NASA (CRS) missions
- Calculation of average payload mass for F9 rocket version v1.1
- Determination of the first successful landing date on a ground platform
- List of rockets with successful drone landings and specific payload mass range
- Total count of successful and failed missions
- Identification of rocket versions with maximum payload
- Analysis of landing results by month in 2015
- Classification of landing results over a specific period

The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/1.2_Sol_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Objects created and added to the map:
  - Markers for launch sites
  - Circles highlighting launch areas
  - Marker clusters for successes/failures
  - Distance lines to points of interest
- Purpose:
  - Visualize launch site locations
  - Analyze success rates by location
  - Assess proximity to key infrastructure

The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/Sol_mejor_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Charts and interactions added:

- Pie chart to show launch success by site
- Scatter plot to visualize the relationship between payload and launch success
- Dropdown menu to select launch sites
- Slider to filter payload range

Rationale:

- Pie chart allows for a quick comparison of success between sites
- Scatter chart reveals patterns between payload and mission success
- Interactions allow users to explore specific data of interest

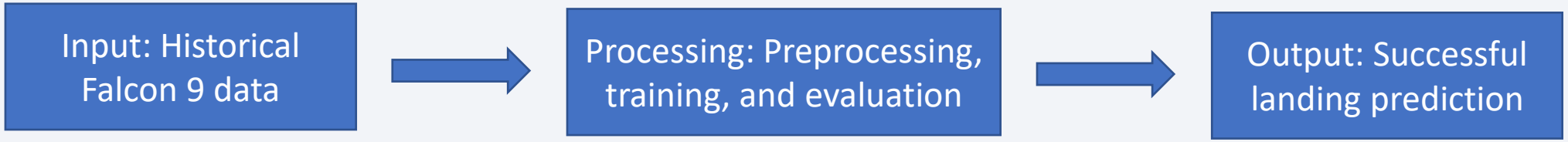
The complete web scraping notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Model development process:

- Data exploration and label creation
- Data standardization
- Splitting into training and test data
- Model evaluation: SVM, Decision Trees, Logistic Regression, and KNN
- Hyperparameter optimization with GridSearchCV
- Selection of the best model based on accuracy

Flowchart:



The complete prediction analysis notebook can be found at the following GitHub URL:
https://github.com/seneguehuesca/SpaceY/blob/main/2_Sol_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

NATIVIDAD 1956

- Exploratory data analysis results
  - Identified patterns in historical Falcon 9 data
  - Created the Class column to classify successful and failed landinds
- Interactive analytics demo in screenshots
  - Visualizations with Seaborn and Matplotlib to understand correlations
- Predictive analysis results
  - Models tested: Logistic regression, SVM, Decision Trees, and KNN
  - Average cross-validation accuracy: ~83%

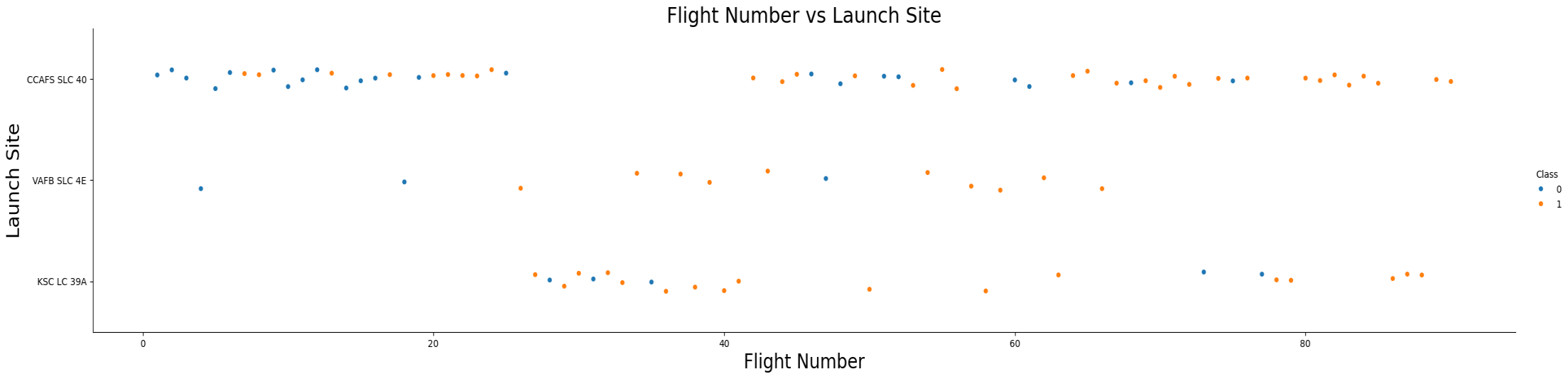
NATIVIDAD 1956

NATIVIDAD 1956

Section 2

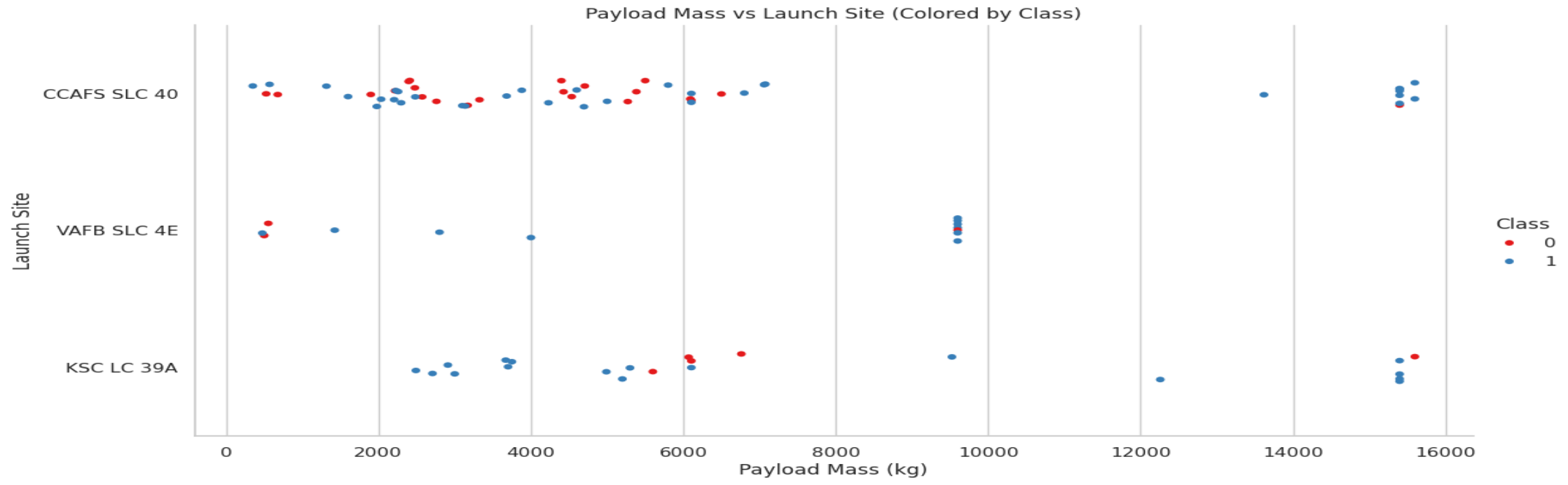
Insights drawn from EDA

Flight Number vs. Launch Site



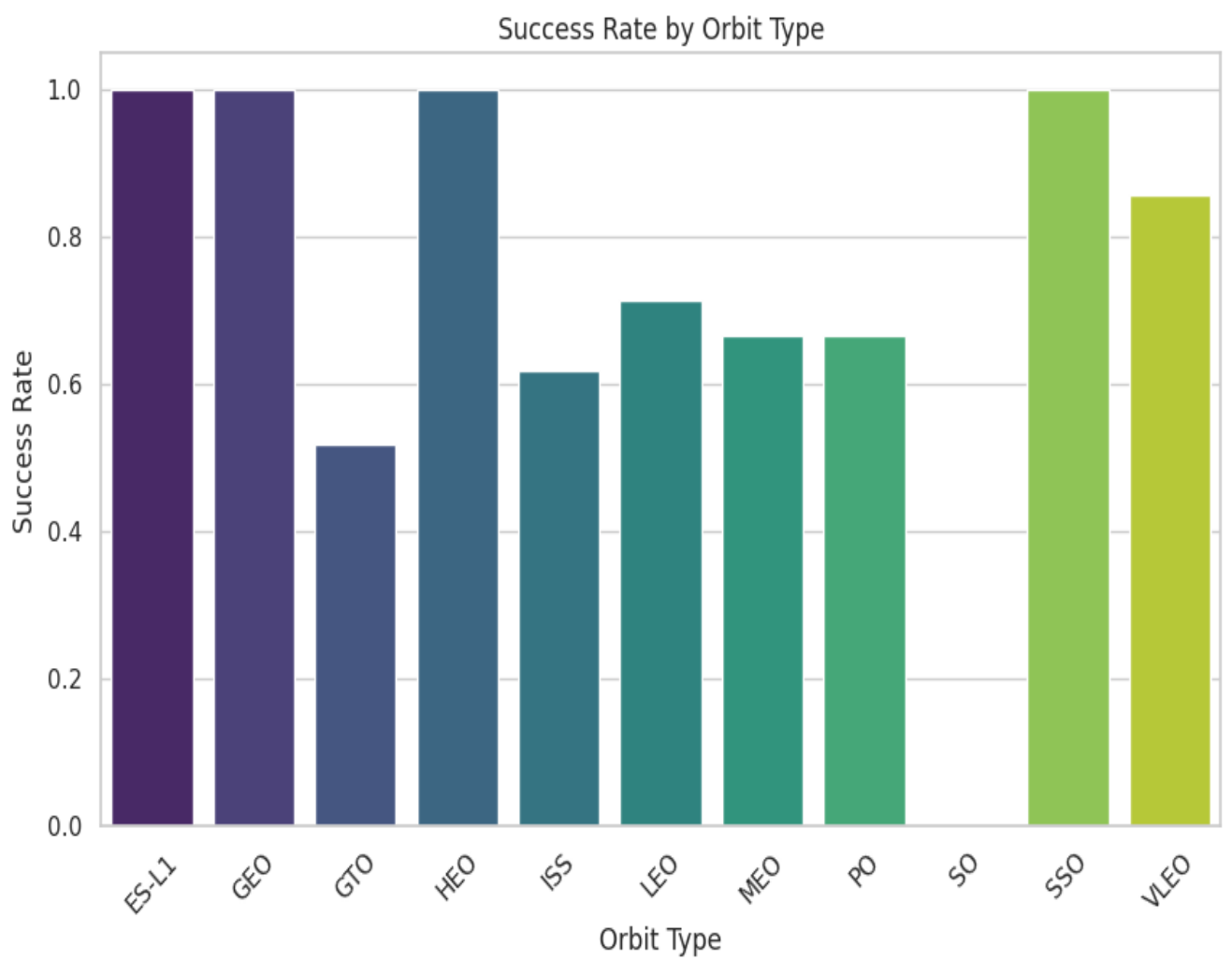
- The graph shows that CCAFS SLC 40 is the most used launch site, with a trend toward greater diversification and flight success as flight numbers increase.

Payload vs. Launch Site



- The graph shows that VAFB SLC 4E is not used for heavy payload launches, suggesting that this site is optimized for specific missions with lighter payloads. On the other hand, CCAFS SLC 40 and KSC LC 39A are the primary sites for heavy payload launches, with a high success rate.

Success Rate vs. Orbit Type

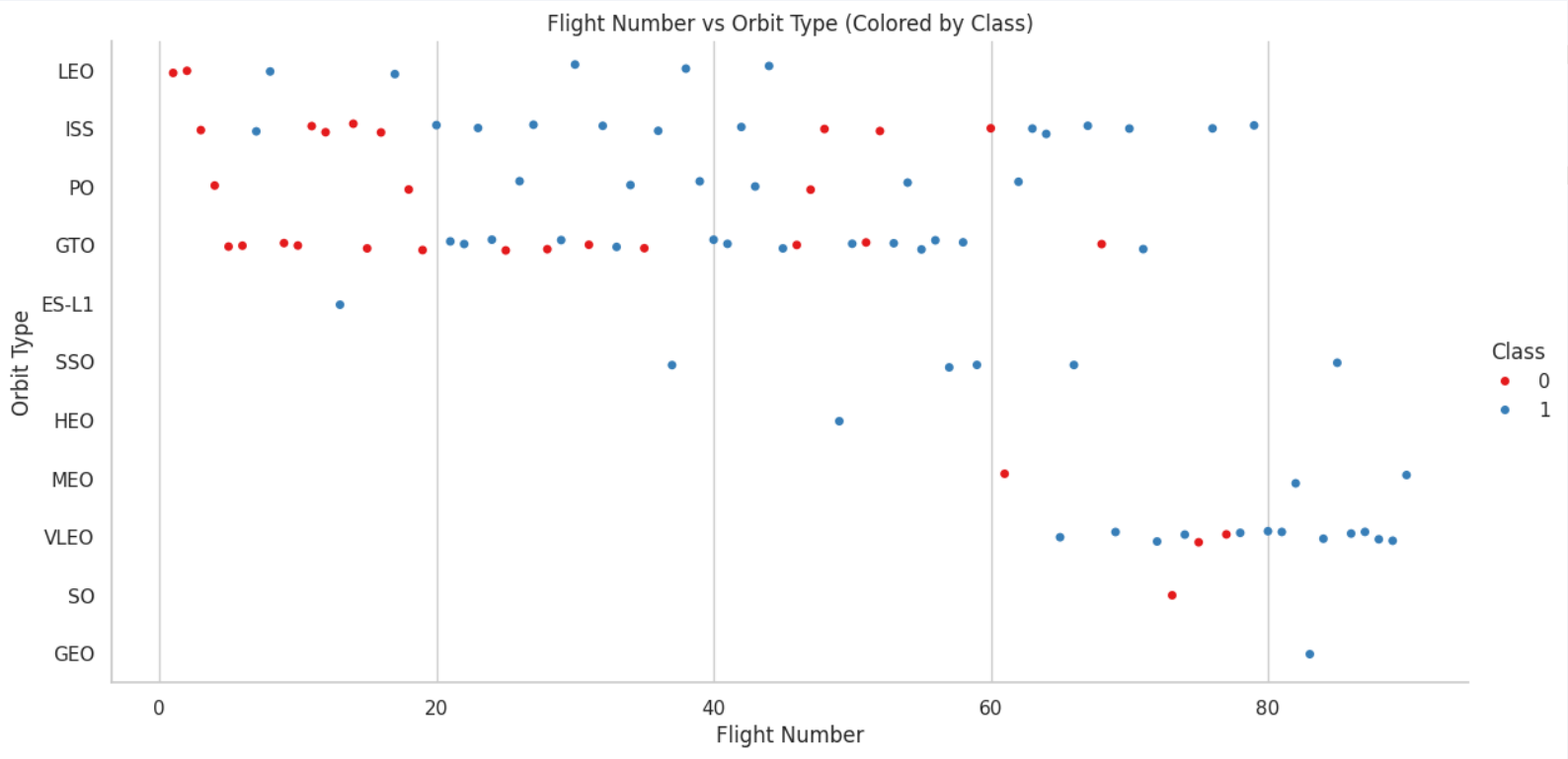


- The orbits with the highest success rates are those with 100% success:

  ES-L1, GEO, HEO, and SSO.

Furthermore, VLEO has a very high success rate (85.7%), making it a very reliable orbit. On the other hand, GTO has a moderate success rate (51.9%), and SO has not had any successful launches in the analyzed data. This suggests that more specialized or less common orbits (such as SO) may be more challenging, while more common or better-optimized orbits (such as LEO and VLEO) have higher success rates.

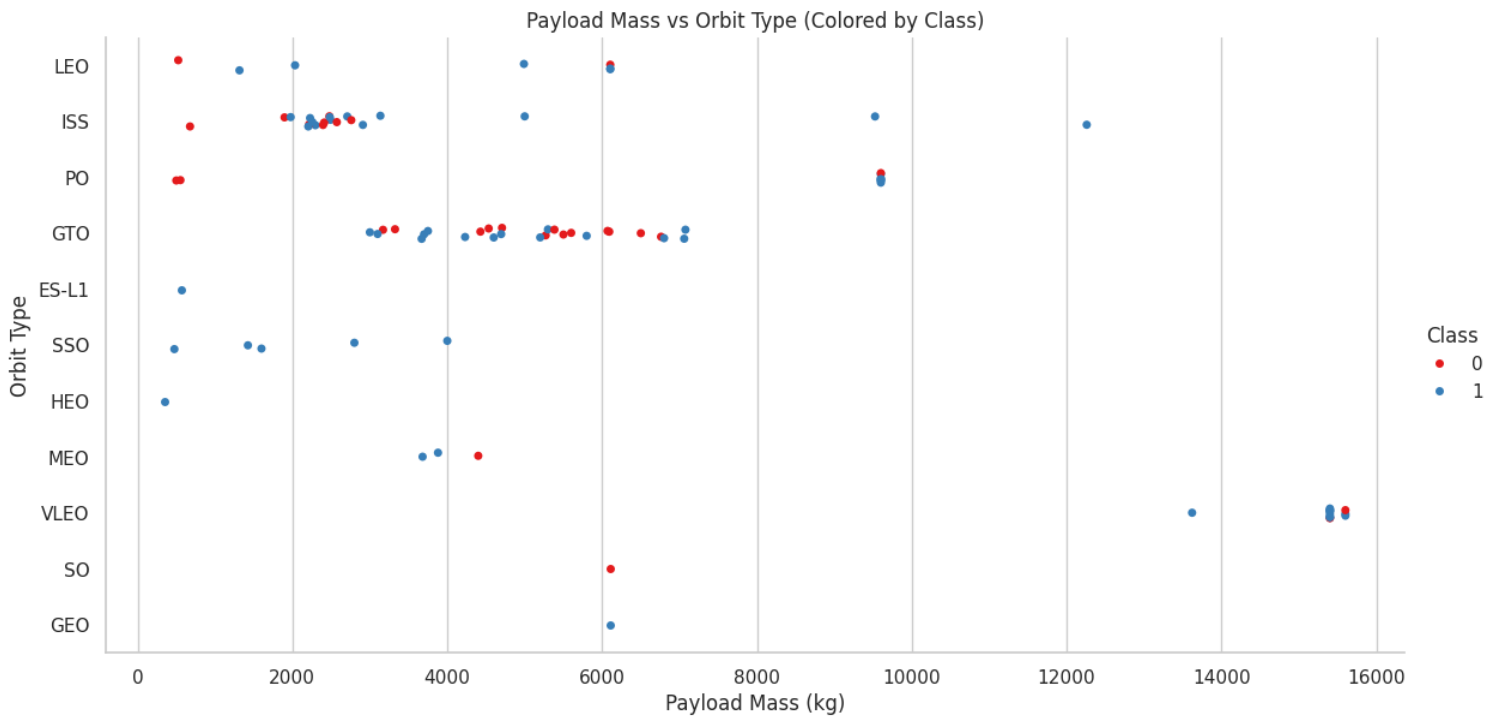
Flight Number vs. Orbit Type



- In LEO orbit, success is correlated with flight number, suggesting that experience and technical improvements have increased the success rate over time.
- In GTO orbit, there is no relationship between flight number and success, indicating that the technical complexity of these launches remains a challenge.
- For other orbits, such as VLEO, a similar trend is observed, while in orbits such as PO and SSO, success does not appear to be related to flight number.

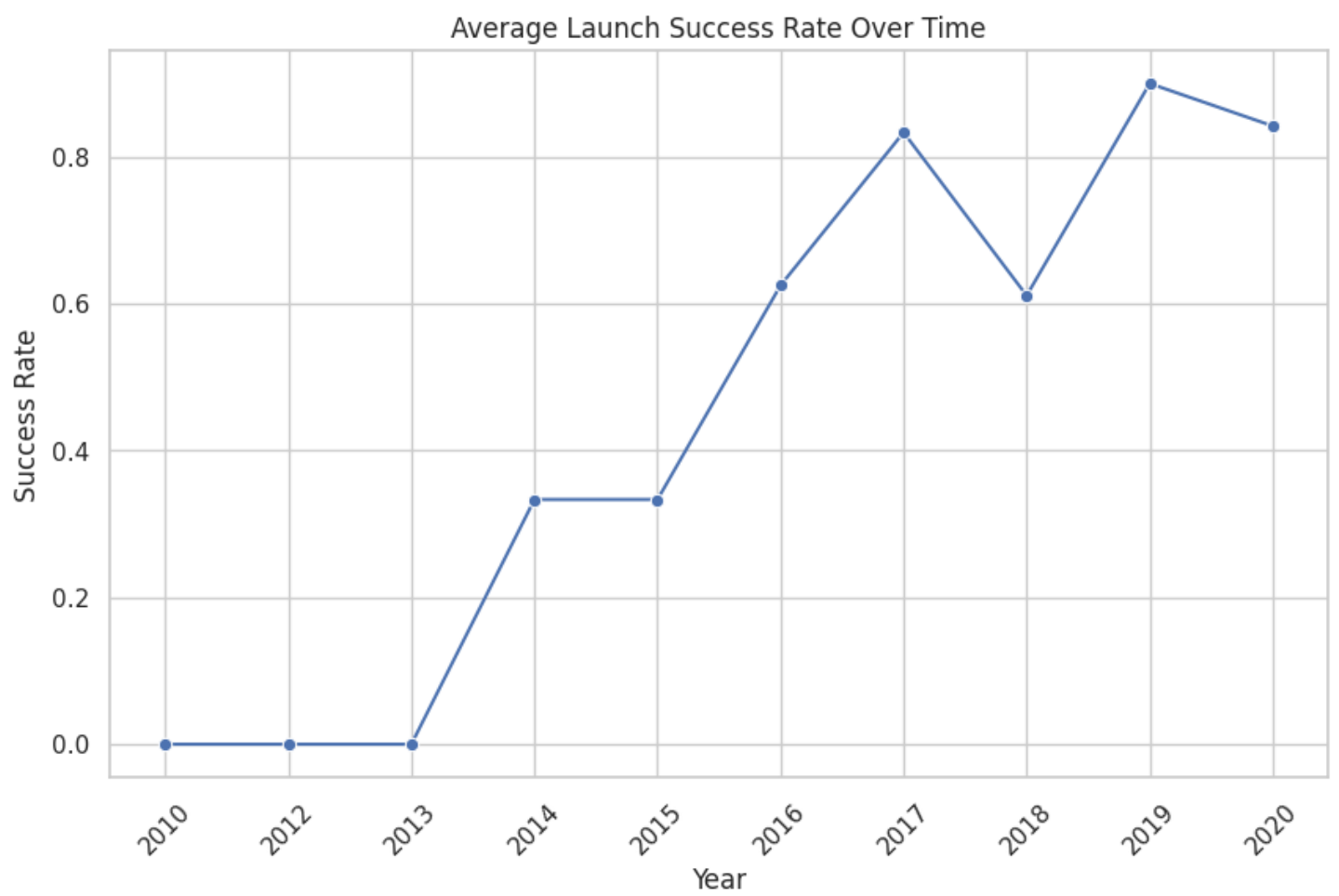
This analysis highlights the importance of considering both accumulated experience and technical complexity when evaluating success rates in different orbits.

Payload vs. Orbit Type



- For orbits such as LEO, ISS, and PO, launches with heavier payloads have higher success rates, suggesting that these orbits are well optimized for handling heavy payloads.
- For the GTO orbit, there is no clear relationship between payload mass and success, indicating that other factors (such as technical complexity) play a more important role in launch outcome.
- In general, the most common and best-optimized orbits tend to have higher success rates with heavy payloads, while more specialized or complex orbits (such as GTO) show a mix of successes and failures regardless of payload mass.

Launch Success Yearly Trend



- The line graph shows that the average launch success rate has increased steadily from 2013 to 2020. This reflects continuous improvements in rocket technology, launch procedures, and accumulated experience. 2020 marks the peak of the success rate, underscoring the maturity and efficiency of launch systems during that period.

All Launch Site Names

SQL QUERY

```
SELECT DISTINCT "LAUNCH_SITE" AS "Launch Sites"
FROM SPACEXTBL;
```

PYTHON - PANDAS

```
df['LaunchSite'].unique()
```

Identified launch sites:

CCAFS SLC 40: Launch site at Kennedy Space Center, Florida.

VAFB SLC 4E: Launch site at Vandenberg Space Force Base, California.

KSC LC 39A: Launch site at Launch Complex 39A at Kennedy Space Center, Florida.

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	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass


```
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTABLE
WHERE CUSTOMER = 'NASA (CRS)';
```

NATIVIDAD 1956

TOTAL_PAYLOAD_MASS
45596

The query will return the total payload mass (in kilograms) carried by boosters launched for NASA (CRS). It means that boosters for NASA (CRS) have carried a cumulative payload mass of 45,596 kg.

Average Payload Mass by F9 v1.1

NATIVIDAD 1956

```
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS
FROM SPACEXTABLE
WHERE BOOSTER_VERSION LIKE '%F9 v1.1%';
```

AVG_PAYLOAD_MASS
2534.6666666666665

The query will return the average payload mass (in kilograms) carried by the F9 v1.1 booster. It means that the F9 v1.1 booster has carried an average payload mass of 2,534.7 kg.

First Successful Ground Landing Date


```
SELECT MIN(Date) AS FIRST_SUCCESSFUL_GROUND_LANDING
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)';
```

FIRST_SUCCESSFUL_GROUND_LANDING
2015-12-22

The query will return the date of the first successful landing on a ground pad. It means that the first successful ground pad landing occurred on December 22, 2015.
This query identifies the historic moment when SpaceX achieved its first successful landing of a Falcon 9 booster on a ground pad. This milestone marked a significant step in SpaceX's efforts to develop reusable rocket technology, reducing the cost of spaceflight and paving the way for more sustainable exploration.

Successful Drone Ship Landing with Payload between 4000 and 6000


```
SELECT DISTINCT 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

The query will return a list of booster versions that meet the specified criteria.

This query identifies reusable boosters that successfully landed on drone ships while carrying payloads in the mid-range mass category (4000–6000 kg).

Total Number of Successful and Failure Mission Outcomes


```
SELECT
   CASE
      WHEN MISSION_OUTCOME LIKE 'Success%' THEN 'Success'
      ELSE MISSION_OUTCOME
   END AS MISSION_OUTCOME,
   COUNT(*) AS TOTAL
FROM SPACEXTABLE
GROUP BY
   CASE
      WHEN MISSION_OUTCOME LIKE 'Success%' THEN 'Success'
      ELSE MISSION_OUTCOME
   END;
```

MISSION_OUTCOME	TOTAL
Failure (in flight)	1
Success	100

This query provides a high-level overview of the mission outcomes in the dataset, helping to understand the reliability and performance of the launch vehicles. The results align with SpaceX's overall success rate, which is consistently high, especially for the Falcon 9.

Boosters Carried Maximum Payload

```
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ =
        (SELECT MAX(PAYLOAD_MASS__KG_)
        FROM SPACEXTABLE);
```

This query is useful for identifying booster versions that have been used on missions with the highest payload mass. This can provide information on the payload capacity of different booster versions and their importance in missions requiring heavy payloads.

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records


```
%sql 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)';
```

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

This query is useful for analyzing failed drone ship landing attempts during 2015. It provides information on the months in which these failures occurred, the booster versions involved, and the launch sites from which the launches were conducted. This can help identify patterns or areas for improvement in SpaceX's recovery operations.

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

```
SELECT Landing_Outcome,
        COUNT(*) AS Landing_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04’
        AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Landing_Count DESC;
```

Landing_Outcome	Landing_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

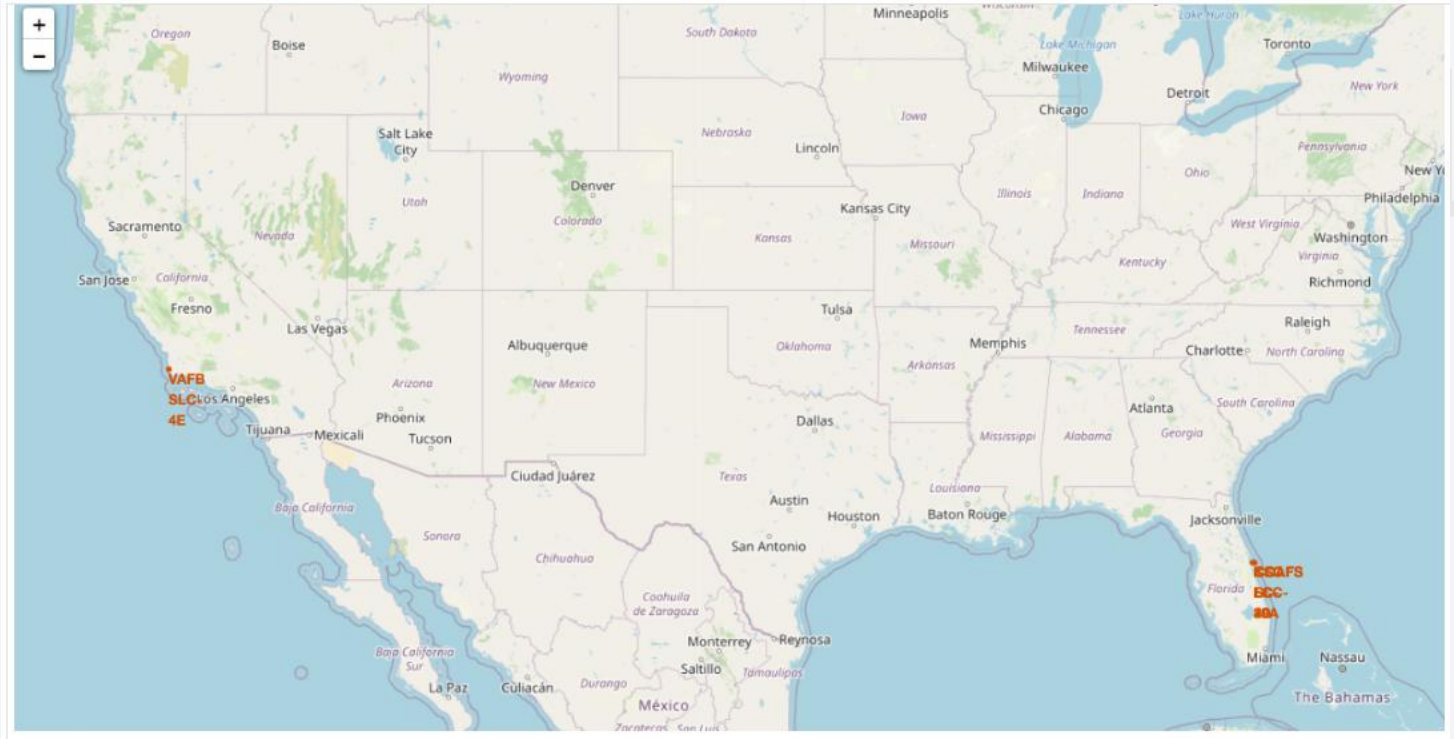
This query is useful for analyzing the distribution of SpaceX landing results within a specific period. The results show how many times each landing type occurred, which can help identify patterns, areas for improvement, and the effectiveness of SpaceX's recovery strategies during that period.

NATIVIDAD 1956
NATIVIDAD 1956

Section 3

Launch Sites Proximities Analysis

Global Launch Site Locations



Key Findings:

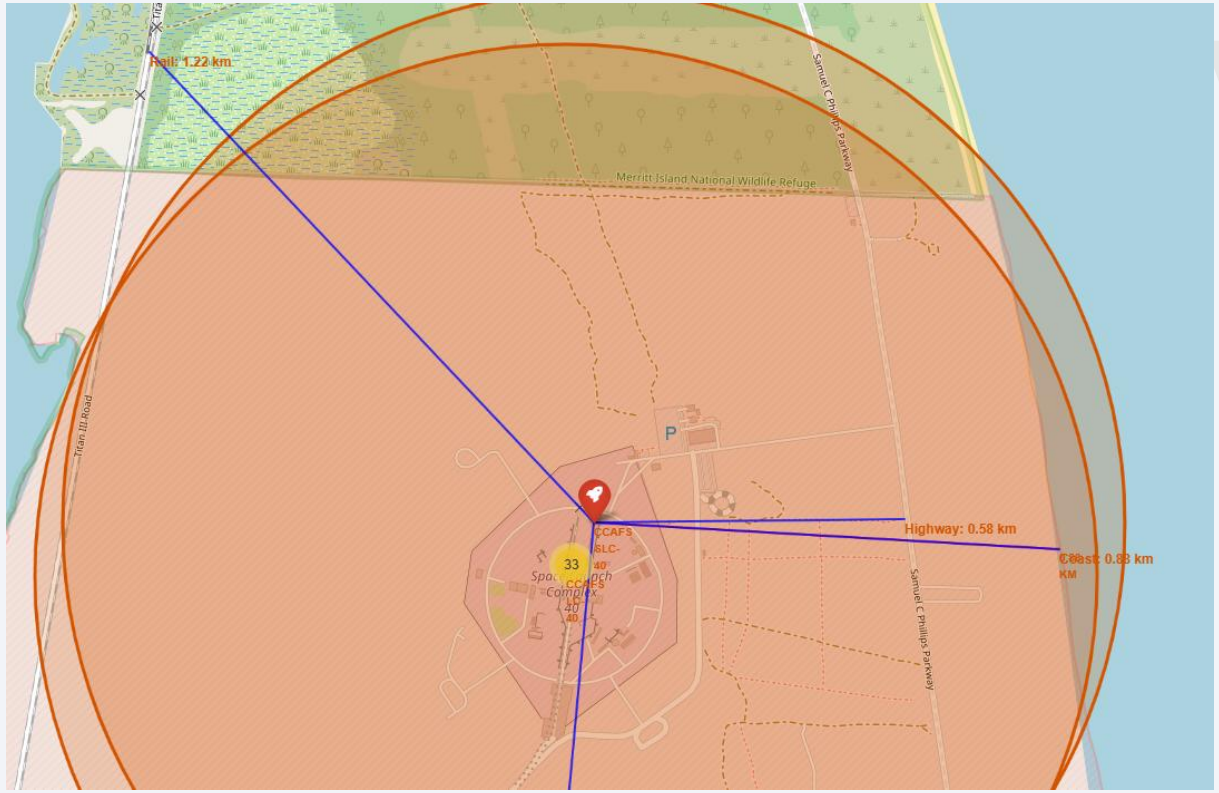
- Sites concentrated in the US: East Coast (Florida) and West Coast (California)
- Strategic locations near coasts for safe trajectories
- Non-equatorial distribution, balancing geographic and logistical advantages

Analysis of Successes and Failures by Site

Important Observations:

- A predominance of green markers indicates a high overall success rate.
- Denser clusters at sites with higher launch frequency.
- Possible correlation between location and success rate.

Proximity to Key Infrastructure



Proximity Analysis for CCAFS SLC-40 :

- Distance to the coast: approximately 0.88 km
- Proximity to rail: approximately 1.22 km
- Proximity to highway: approximately 0.58 km
- Proximity to city: approximately 17.63 km

Conclusions:

- Strategically located sites near coasts for safety
- Good connection to transportation infrastructure
- Prudent distance from urban areas for safety

NATIVIDAD 1956

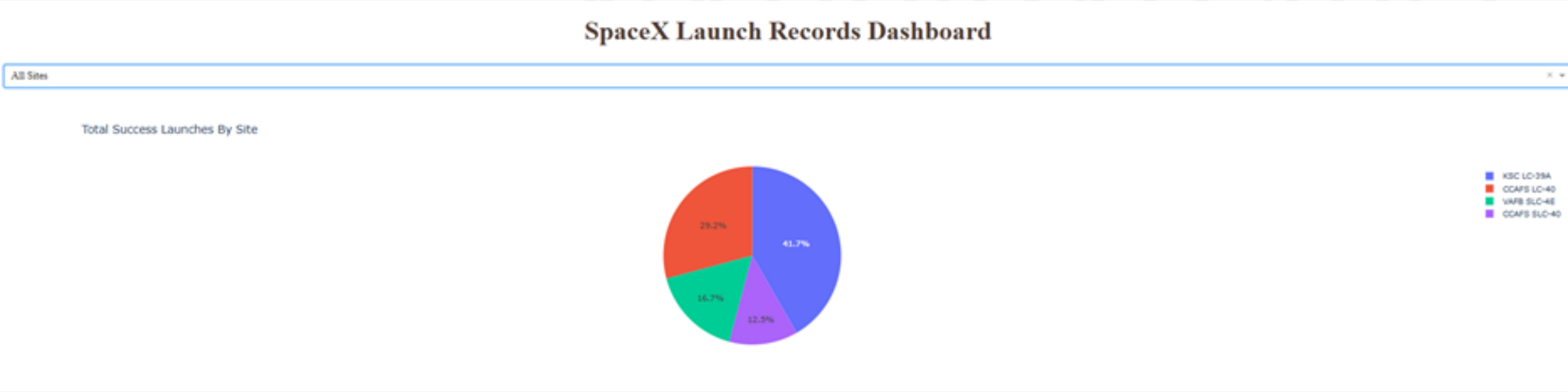
NATIVIDAD 1956

Section 4

Build a Dashboard with Plotly Dash

Launch Success by Site

Distribution of Launch Success by Site

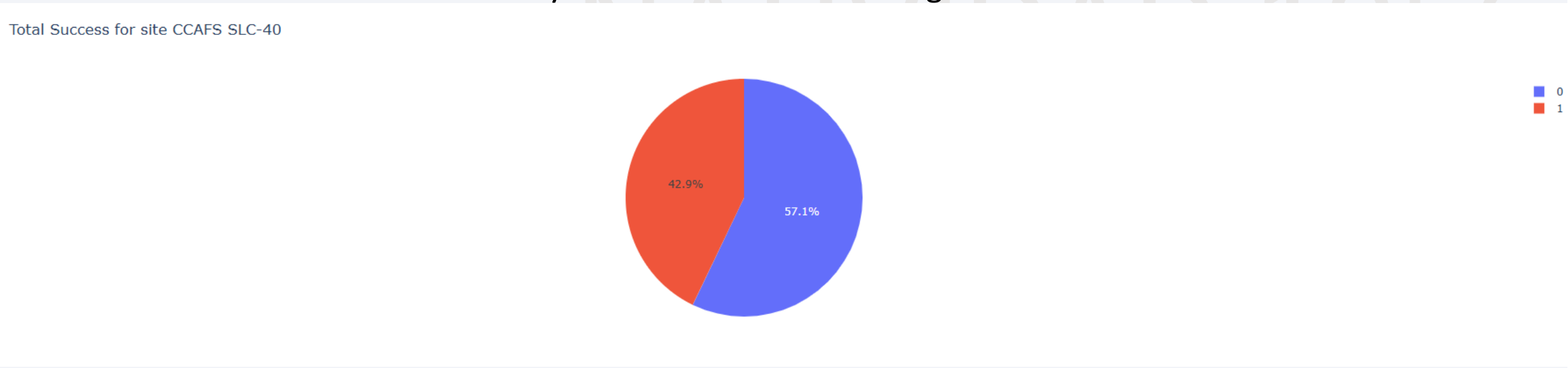


Important elements and findings:

- The chart shows the proportion of successful launches for each site.
- It allows you to quickly identify which site has the highest number of successful launches.
- It facilitates visual comparison between different launch sites.

Site with Highest Success Rate

Detailed Analysis of the Site with the Highest Success Rate

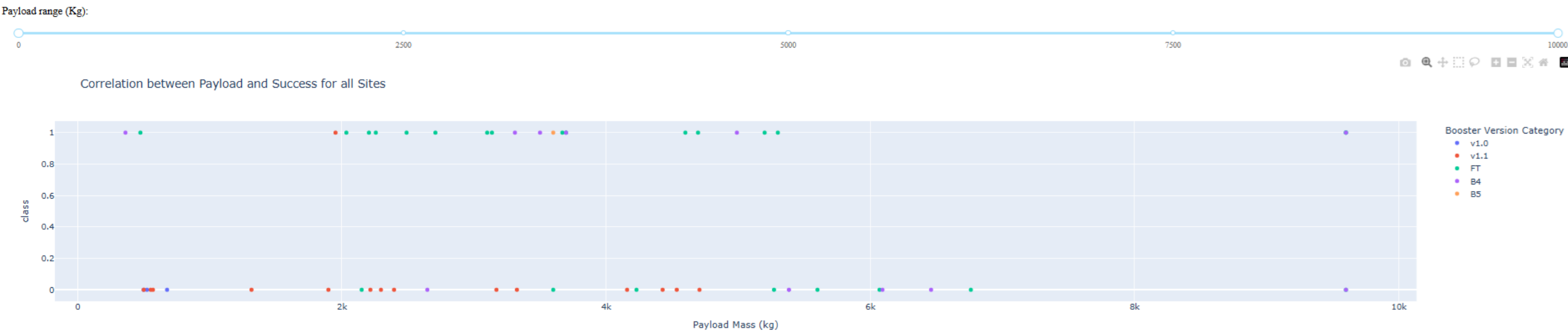


Key Elements and Findings:

- Shows the ratio of successful vs. failed launches for the selected site.
- Allows for a deeper analysis of the success rate of the most effective site.
- Provides context on the reliability of the most successful launch site.

Relationship between payload and launch success

Impact of payload on launch success



Important elements and findings:

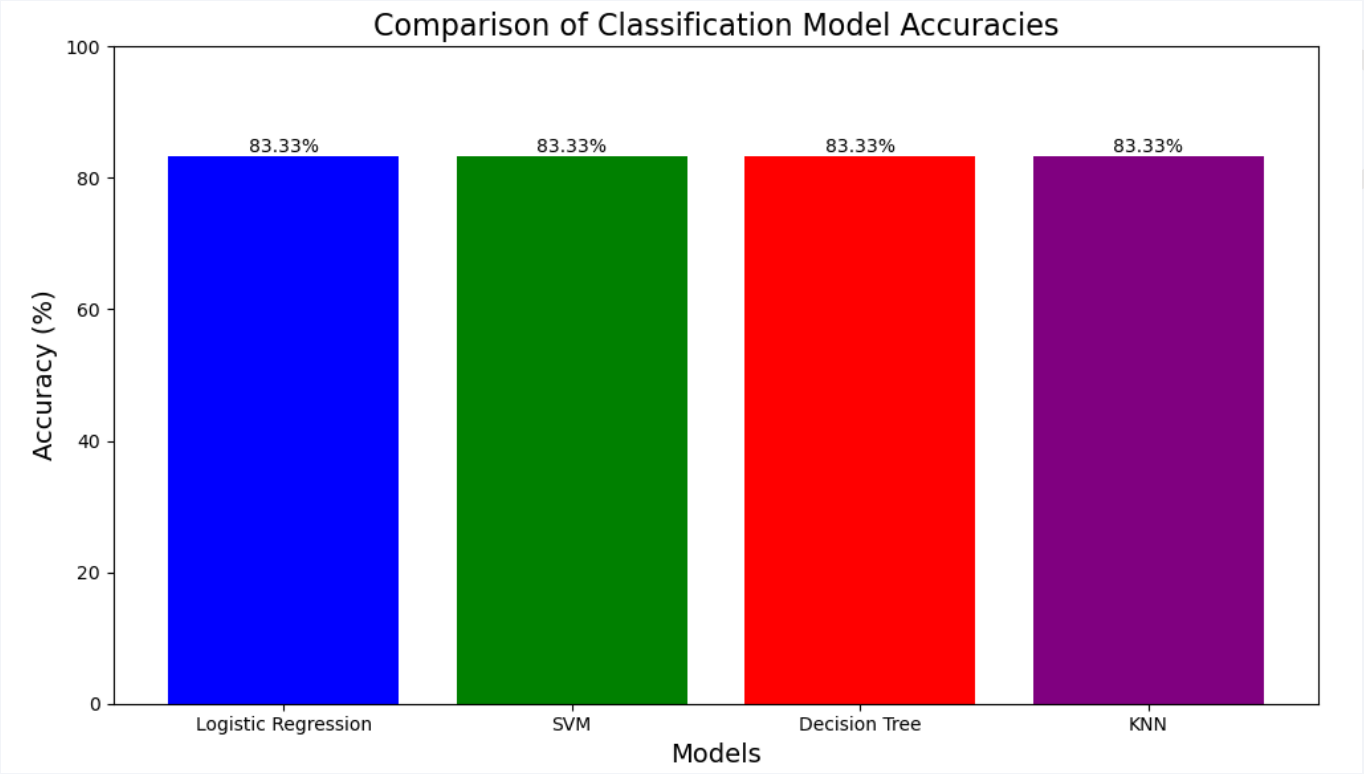
- Visualizes the relationship between payload mass and launch success.
- Dots are colored by booster version, allowing additional patterns to be identified.
- Key observations:
  - Payload ranges with the highest success rate. Between 2000 kg and 4000 kg
  - Booster versions with the best performance. The best is FT
  - Potential correlations between payload mass and mission success.

NATIVIDAD 1956
NATIVIDAD 1956

Section 5

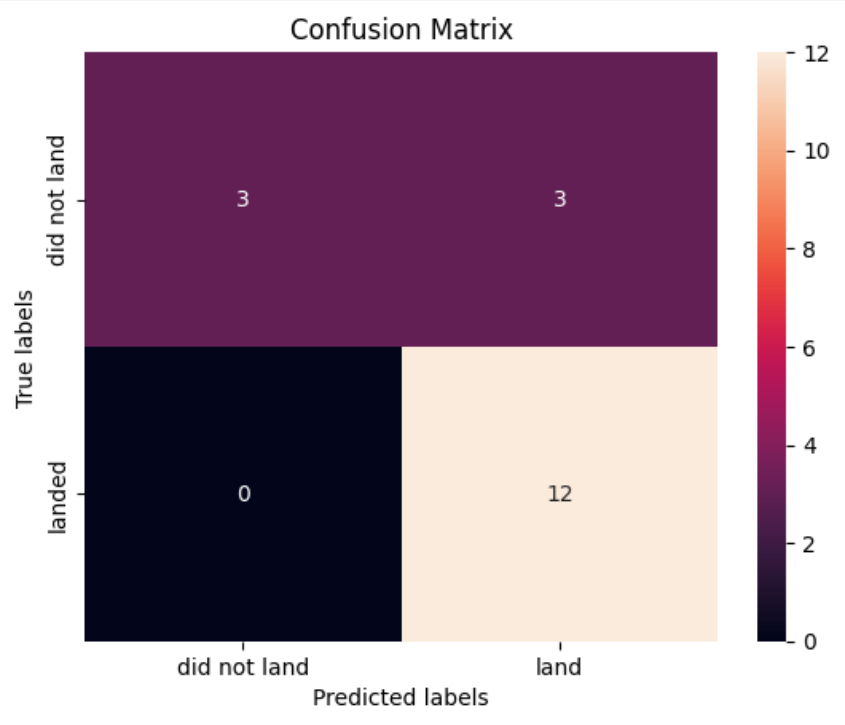
Predictive Analysis (Classification)

Classification Accuracy



All models performed similarly, but logistic regression was selected for simplicity.

Confusion Matrix



- True Positives (12): Successful landings correctly predicted.
- False Positives (3): Incorrect predictions as successful landinds.

Explination:
The model is able to distinguish classes, but has room for improvement in falses positives.

Conclusions

NATIVIDAD 1956

- SpaceX as a leader in innovation: SpaceX has revolutionized the aerospace industry by significantly reducing costs through the reusability of the Falcon 9, setting a new standard for space launches.
- Importance of prediction: The ability to predict the success of a Falcon 9 landing is crucial for assessing the economic viability of launches and for other companies to compete in the market.
- Robust Predictive Models: The evaluated models (Logistic Regression, SVM, Decision Tree, KNN) showed consistent accuracy (~83%), indicating that the data is well-structured and suitable for predictive analysis.
- Logistic Regression as the Preferred Model: Although all models had similar performance, logistic regression was selected for its simplicity and computational efficiency for this dataset.
- Impact on the Space Industry: This predictive approach can be used by competing companies to optimize their strategies and reduce costs, fostering greater competitiveness in the aerospace sector.

Appendix

Code for comparison of Classification Model Accuracies

```
import matplotlib.pyplot as plt

# Datos de precisión
models = ['Logistic Regression', 'SVM', 'Decision Tree', 'KNN']
accuracies = [83.33, 83.33, 83.33, 83.33]

# Crear el gráfico de barras
plt.figure(figsize=(10, 6))
bars = plt.bar(models, accuracies, color=['blue', 'green', 'red', 'purple'])

# Personalizar el gráfico
plt.title('Comparación de Precisión de Modelos de Clasificación', fontsize=16)
plt.xlabel('Modelos', fontsize=14)
plt.ylabel('Precisión (%)', fontsize=14)
plt.ylim(0, 100)  # Establecer el rango del eje y de 0 a 100

# Añadir etiquetas de valor encima de cada barra
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2., height,
             f'{height}%',
             ha='center', va='bottom')

# Ajustar el diseño y mostrar el gráfico
plt.tight_layout()
plt.show()
```

- Limitations and Future Opportunities:
  - Limitations: The false positives identified in the confusion matrix indicate areas where the models can be improved.
  - Opportunities: Incorporating more relevant features on weather conditions, rocket design, and historical data could improve predictions.
- Applications Beyond SpaceX:This analysis can be adapted to other reusable space vehicles, helping to evaluate their performance and economic viability.
- The Value of Machine Learning:This project demonstrates how machine learning algorithms can be applied to real-world problems to make data-driven decisions.

NATIVIDAD 1956

NATIVIDAD 1956

Thank you!