IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Abdoul Aziz Gueye
07/12/2024

Outline

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

Executive Summary

- Data Collection: Web scraping and REST API queries.
- Data Wrangling: Launch classification and data standardization.
- Exploratory Data Analysis (EDA): SQL and Python visualizations.
- Interactive Web App: Plotly dashboard for payload and success data.
- Launch Site Exploration: Folium interactive maps.
- Predictive Analysis: Classification models for landing success.
- Results Summary: EDA findings and predictive analysis results.

Introduction

This project focuses on analyzing SpaceX Falcon 9 launch data to gain insights into the factors driving the success of rocket landings, a cornerstone of SpaceX's cost-saving strategy through reusable rockets. By applying advanced data science techniques, we processed raw data into actionable insights, explored trends, and developed predictive models. The results provide a data-driven foundation for improving landing success rates and optimizing launch operations, demonstrating how data science can drive innovation in the aerospace industry.

Section 1

Methodology

Methodology

Data Collection using API and Web Scraping

- Utilized REST API queries to fetch data about SpaceX Falcon 9 launches.
- Extracted details such as launch dates, sites, payloads, and landing outcomes.
- Implemented web scraping techniques to gather supplementary launch data from relevant websites.
- Leveraged libraries like BeautifulSoup and Requests for HTML parsing and data extraction.

Data Wrangling

- Processed raw data to clean inconsistencies and handle missing values.
- Transformed categorical data into numerical format for easier analysis.
- Classified launch data based on success criteria for downstream tasks.

3. Exploratory Data Analysis

- Performed SQL-based exploratory analysis to identify trends in launch success rates.
- Created summary statistics and visualized relationships between payload, site, and outcomse.

Methodology

4. Interactive Dashboard Development

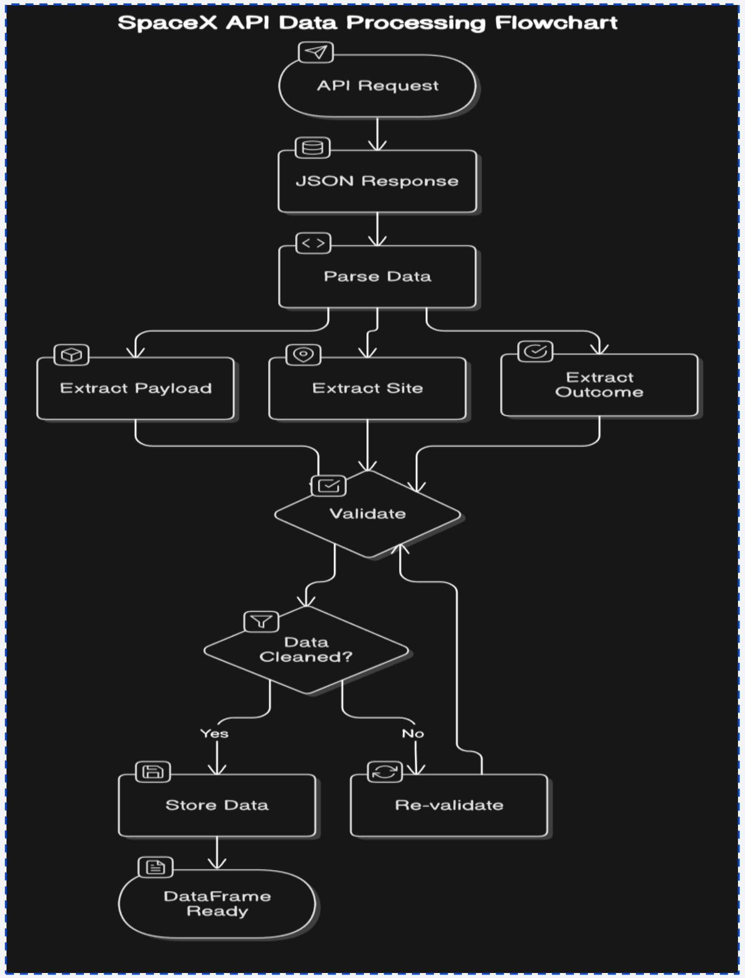
- Built an interactive dashboard using Dash and Plotly for visualizing SpaceX launch data.
- Integrated a dropdown menu for site selection, sliders for payload range, and interactive charts like scatter plots and pie charts.
- Enabled dynamic updates for user-selected filters.

5. Predictive Modeling

- Developed machine learning models to predict rocket landing success.
- Evaluated models based on classification metrics like accuracy, precision, and recall.
- Optimized hyperparameters to enhance model performance.

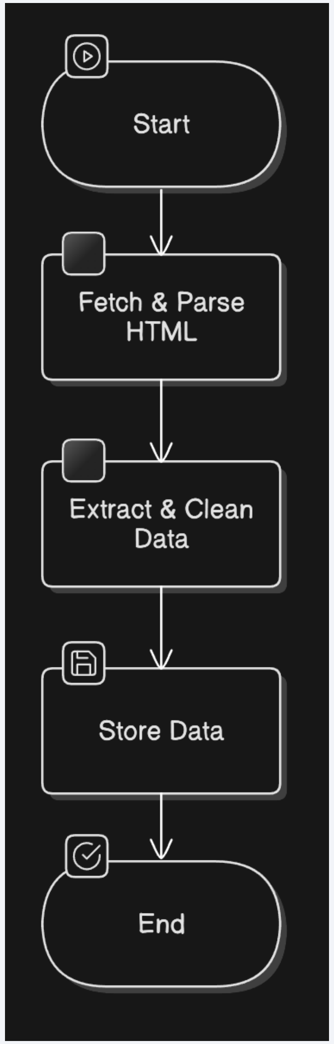
Data Collection – SpaceX API

- API Request: Sent GET request to SpaceX API.
- JSON Response: Received structured JSON data.
- Parse Data: Extracted fields like payload, site, and outcome.
- Validate: Cleaned and checked data accuracy.
- Store Data: Saved in a DataFrame for analysis.
- GitHub URL: [IBM-Capstone-Projects/01_data_collection.ipynb at main · Aziz1234578/IBM-Capstone-Projects]()



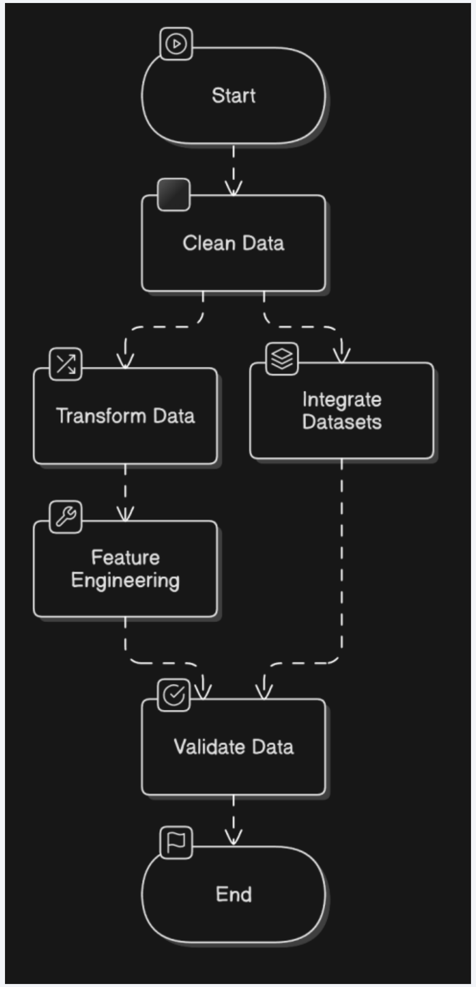
Data Collection - Scraping

- Fetch HTML: Used requests to retrieve webpage content.
- Parse HTML: Utilized BeautifulSoup to parse the HTML structure.
- Extract Data: Extracted required information (e.g., tables, div elements).
- Clean Data: Removed inconsistencies and handled missing values.
- Store Data: Saved clean data in a DataFrame.
- GitHub URL : IBM-Capstone-Projects/01_bis_data_collection_web_scrapping.ipynb at main · Aziz1234578/IBM-Capstone-Projects



Data Wrangling

- **Data Cleaning:** Addressed missing values, duplicates, and inconsistencies in the dataset.
- **Data Transformation:** Standardized categorical fields (e.g., launch outcomes) into numerical values for analysis.
- **Feature Engineering:** Created new features like success rates and payload categories.
- **Integration:** Merged API and web-scraped datasets into a unified structure.
- **Validation:** Ensured data accuracy and consistency across all sources.
- GitHub URL : [IBM-Capstone-Projects/02_data_wrangling.ipynb at main · Aziz1234578/IBM-Capstone-Projects]



EDA with SQL

- **Summary of SQL Queries Performed**
- **Query 1:** Selected launch success counts grouped by site.
- **Query 2:** Retrieved average payload mass for successful launches.
- **Query 3:** Filtered launch data for missions with specific criteria (e.g., orbit type and payload range).
- **Query 4:** Counted total successes and failures for all launches.
- **Query 5:** Calculated success rates for each launch site.
- GitHub URL: IBM-Capstone-Projects/03_data_analysis_eda_sql.ipynb at main · Aziz1234578/IBM-Capstone-Projects

Build an Interactive Map with Folium

Summary of Map Objects Added

- **Markers:** Show SpaceX launch site locations with popups for site details.
- **Circles:** Highlight proximities to infrastructure like highways and coastlines.
- **Polylines:** Connect launch sites to nearby points of interest.
- **Color-Coding:** Differentiate launch outcomes (success/failure).

Purpose

- **Markers:** Display precise locations and launch site details.
- **Circles:** Visualize proximity to critical infrastructure.
- **Polylines:** Show distances between sites and key facilities.
- **Color-Coding:** Easily identify patterns in launch outcomes.

- GitHub URL: IBM-Capstone-Projects/04_interactive_visualization_with_folium.ipynb at main · Aziz1234578/IBM-Capstone-Projects

Build a Dashboard with Plotly Dash

Summary of Dashboard Components

- **Dropdown Menu:** Filter data by launch site.
- **Pie Chart:** Show launch success rates per site.
- **Range Slider:** Adjust payload mass range.
- **Scatter Plot:** Visualize payload vs. success correlation

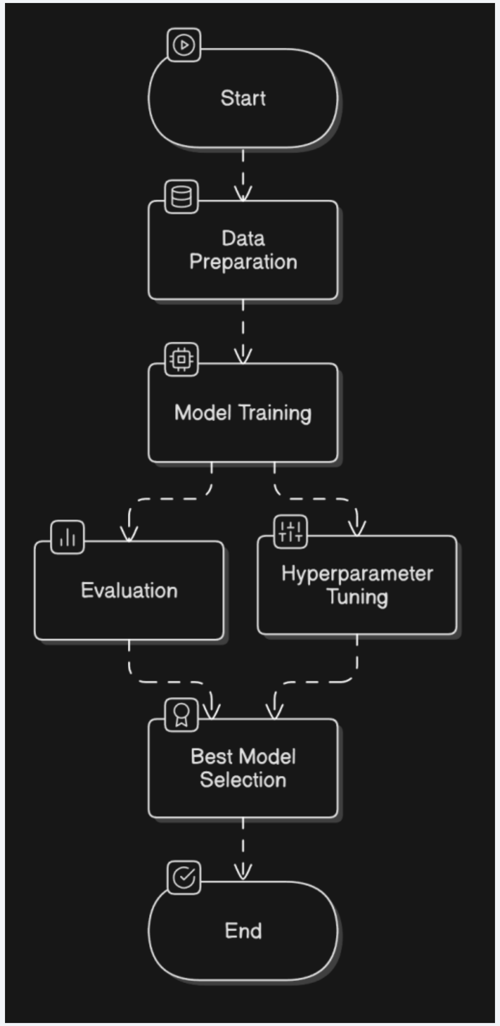
Purpose

- **Dropdown Menu:** Focus on specific sites.
- **Pie Chart:** Compare success rates across sites.
- **Range Slider:** Explore payload impact on success.
- **Scatter Plot:** Analyze payload-success relationship.
- GitHub URL : IBM-Capstone-Projects/Dash_App_Project.py at main · Aziz1234578/IBM-Capstone-Projects

Predictive Analysis (Classification)

Development process:

- Data Splitting: Used train_test_split to divide data into training and testing sets.
- Model Training: Applied classifiers like Logistic Regression and Random Forest.
- Evaluation: Calculated metrics like accuracy, precision, and recall.
- Improvement: Performed hyperparameter tuning (e.g., GridSearchCV) to enhance model performance.
- Best Model Selection: Identified the model with the highest evaluation metrics.
- Github Url : [IBM-Capstone-Projects/05_predictive_modelling.ipynb at main · Aziz1234578/IBM-Capstone-Projects]



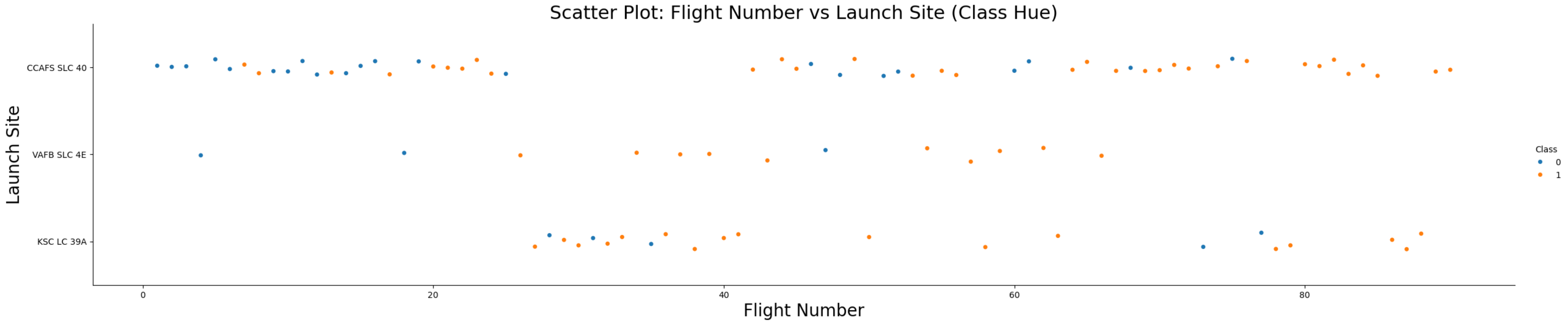
Results

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

Section 2

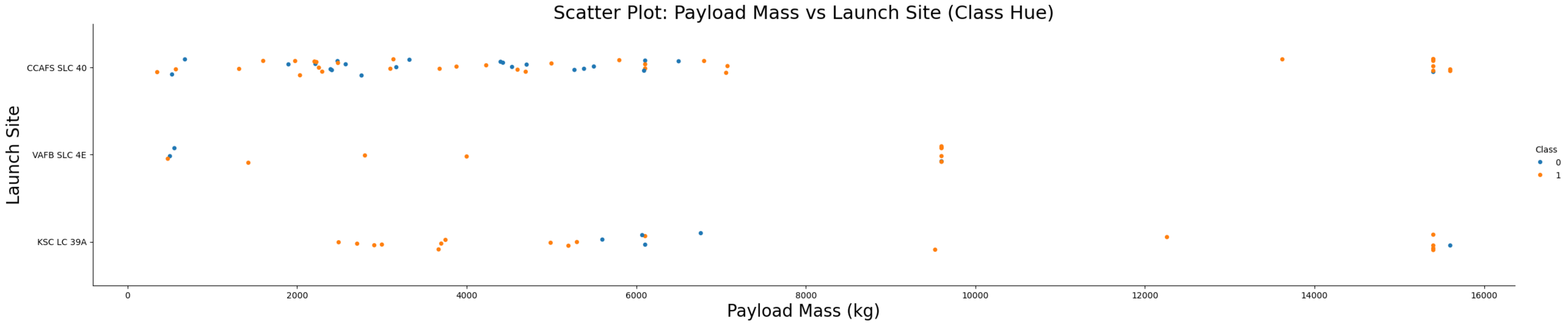
Insights drawn from EDA

Flight Number vs. Launch Site



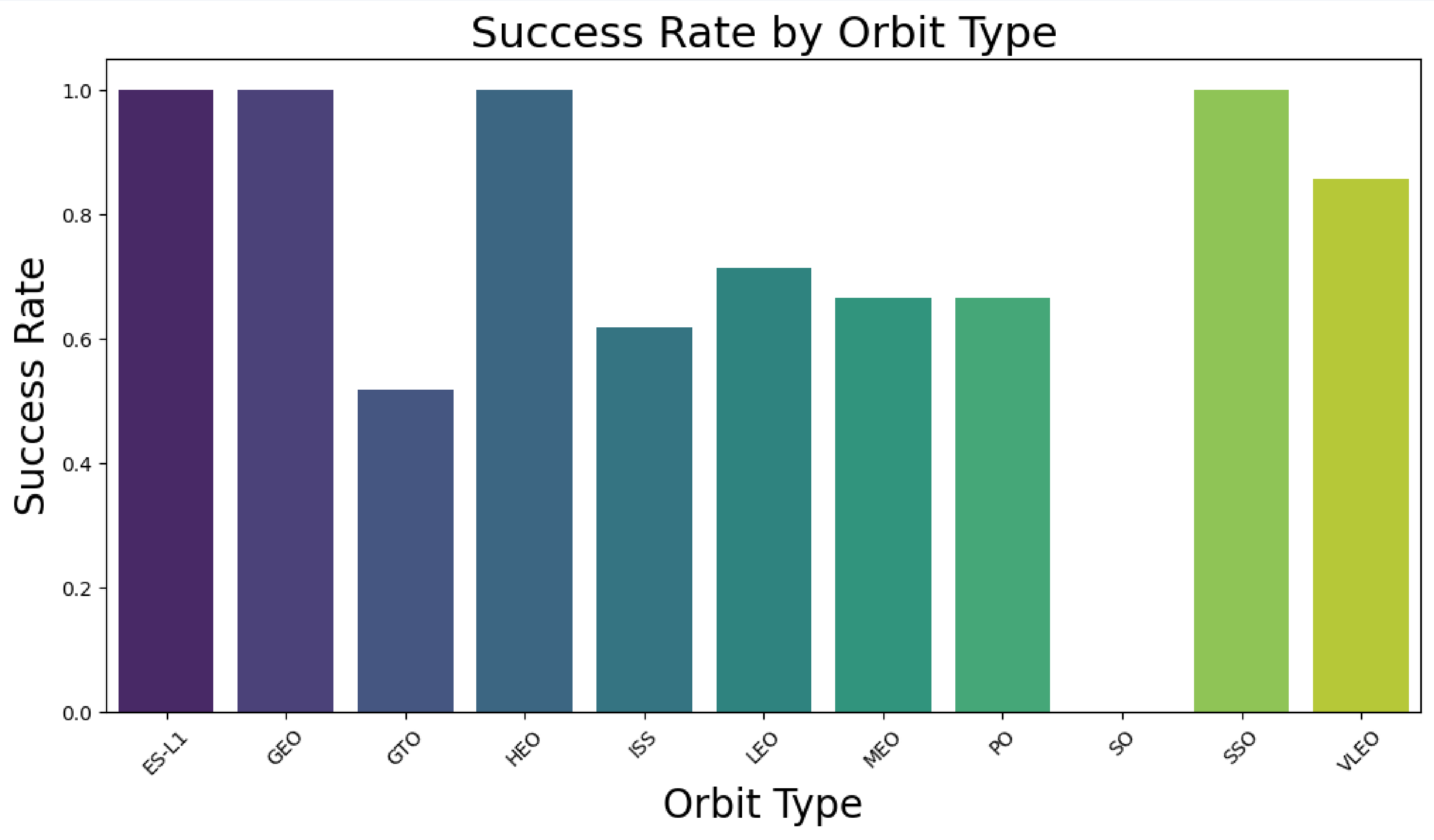
The plot shows the relationship between flight numbers and launch sites, highlighting the success (orange) and failure (blue) of launches. Sites like **KSC LC 39A** and **CCAFS SLC 40** have higher success rates, especially in recent flights, indicating improved reliability over time. In contrast, **VAFB SLC 4E** has fewer launches, making trends less apparent. Overall, the plot demonstrates SpaceX's progress in achieving successful launches.

Payload vs. Launch Site



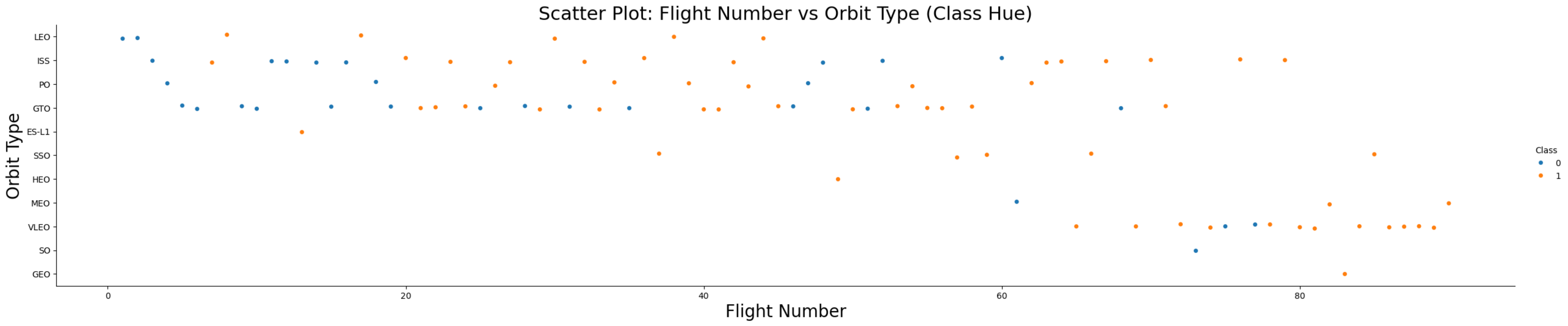
The plot shows the relationship between payload mass (x-axis) and launch sites (y-axis), indicating success (orange) and failure (blue) of launches. Successful launches are more concentrated in moderate payload ranges (2000–6000 kg) across all sites, with **KSC LC 39A** and **CCAFS SLC 40** achieving better success rates even at higher payloads. Overall, this suggests that SpaceX has optimized performance for a wide payload range.

Success Rate vs. Orbit Type



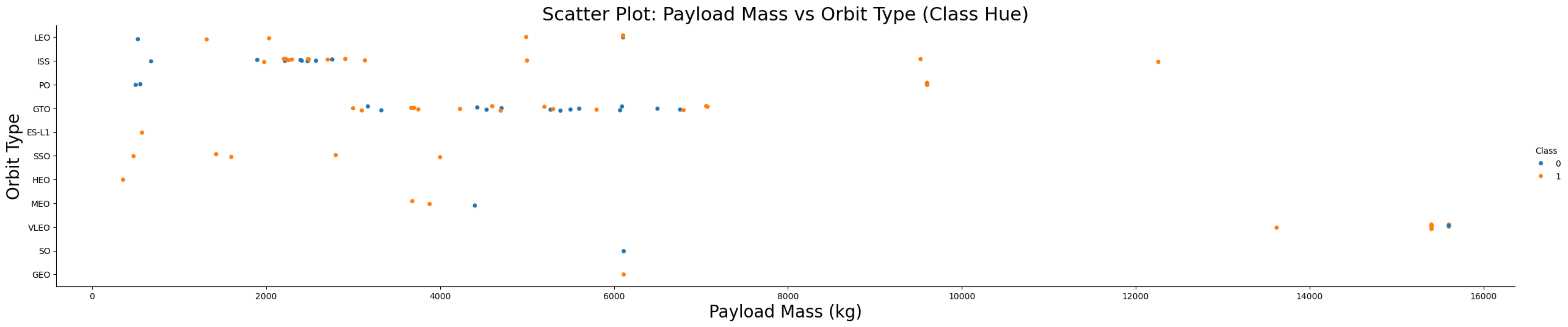
The plot displays the success rate of launches across different orbit types. Orbits such as **ES-L1**, **SSO**, and **HEO** have the highest success rates, nearing 100%, indicating reliable performance. Lower success rates are observed for orbits like **GTO**, suggesting potential challenges in achieving consistent success. This highlights SpaceX's proficiency in specific orbit types while indicating areas for improvement in others.

Flight Number vs. Orbit Type



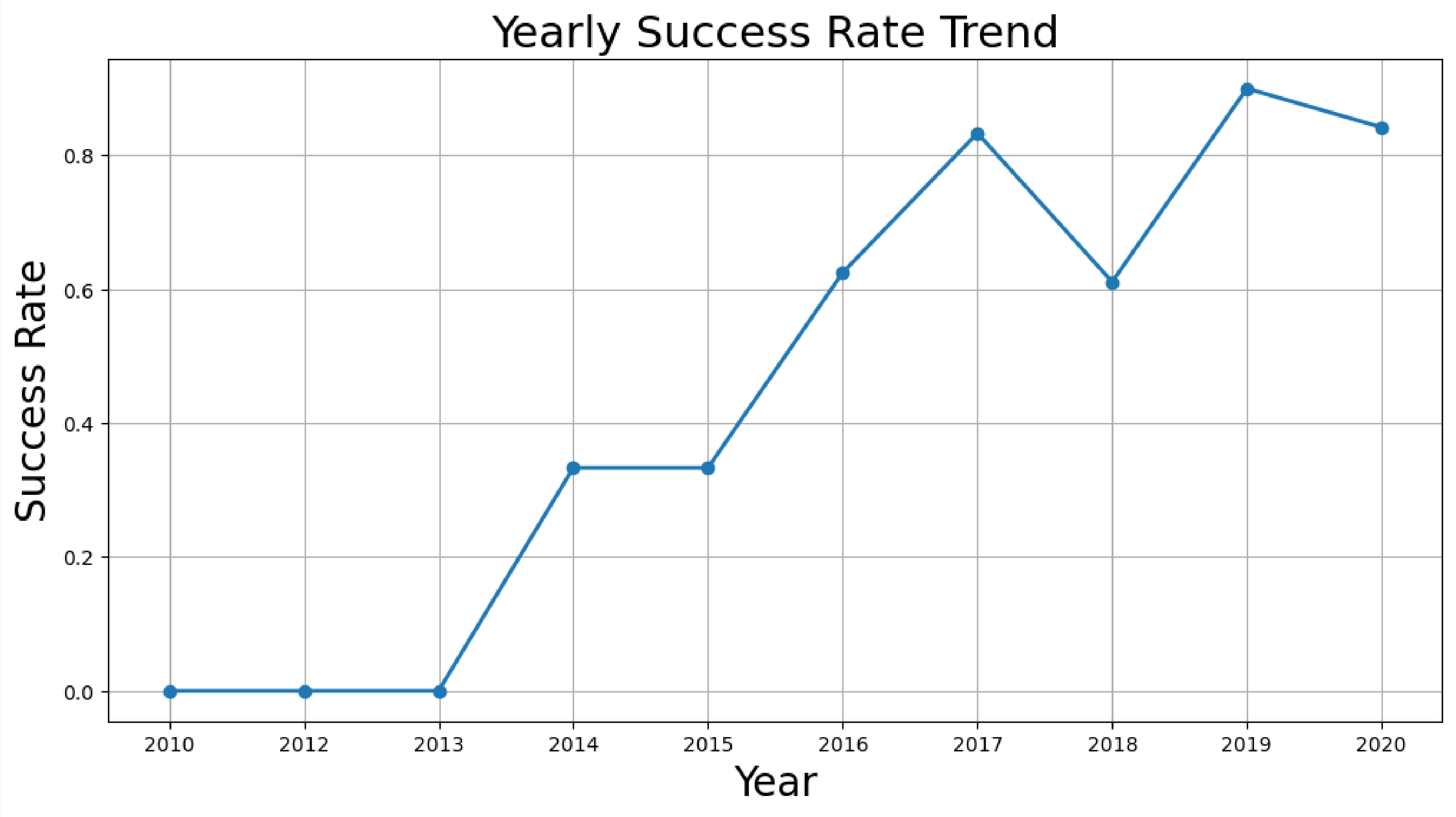
The plot illustrates the relationship between flight numbers and orbit types, showing successful (orange) and failed (blue) launches. Over time, as the flight number increases, there is a noticeable improvement in success rates across most orbit types, particularly in **LEO**, **ISS**, and **SSO**. This indicates SpaceX's growing proficiency in achieving successful launches in diverse orbital categories.

Payload vs. Orbit Type



The plot shows the relationship between payload mass and orbit types, with success (orange) and failure (blue) points. Successful launches are concentrated in payload ranges between 2000–6000 kg for orbits like LEO, ISS, and SSO, highlighting SpaceX's reliability in these categories. Higher payloads also show success but are less frequent, indicating specialized operations for such missions.

Launch Success Yearly Trend



The plot shows the yearly trend in SpaceX's launch success rates. From 2013 onward, there is a consistent improvement, reaching peak success rates around 2018 and 2019. This reflects SpaceX's growing expertise and operational reliability over time, with minor fluctuations indicating areas for refinement.

All Launch Site Names

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Sql Query :

```
SELECT DISTINCT
Launch_Site FROM
SPACEXTABLE;
```

The query retrieves distinct launch site names from the dataset. The results show three unique launch sites where SpaceX conducts its missions: **CCAFS LC-40**, **VAFB SLC-4E**, and **KSC LC-39A**, indicating these are key operational hubs.

Launch Site Names Begin with 'CCA'

```
* sqlite:///my_data1.db
Done.
```

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

SQL QUERY:
SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%’ LIMIT 5;

The query identifies the launch records associated with the **CCAFS LC-40** site. These records are crucial for analyzing mission history and success trends at this specific location. The results show early missions primarily aimed at qualification and testing, with gradual improvements in payload capacity and outcomes.

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE '%NASA (CRS)%';
```

[14]

```
 * sqlite:///my_data1.db
Done.
```

SQL QUERY:
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Mission_Outcome LIKE '%NASA (CRS)%';

The query calculates the sum of payload mass for all launches where the mission outcome is associated with **NASA (CRS)**. This highlights the cumulative payload transported during NASA's Commercial Resupply Services (CRS) missions.

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

[15]

* sqlite:///my_data1.db
Done.

Average_Payload_Mass
2928.4

SQL Query:
SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';

The query calculates the average payload mass carried by the **F9 v1.1** booster version. This indicates the typical payload capacity for this booster during its missions, reflecting its operational efficiency and range.

First Successful Ground Landing Date

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

[16]

```
* sqlite:///my_data1.db
Done.
```

First_Successful_Landing
2015-12-22

The query identifies the earliest date when a SpaceX booster achieved a successful landing on a ground pad. This milestone, achieved on **December 22, 2015**, marks a significant advancement in SpaceX's reusable rocket technology, improving cost efficiency and sustainability in space exploration.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster Versions
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

SQL QUERY :

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

The query retrieves booster versions that successfully landed on a drone ship and carried payloads between **4000 kg and 6000 kg**. These boosters highlight SpaceX's ability to execute precision landings while handling moderate payloads, showcasing technological advancements in reusability and payload delivery.

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT Mission_Outcome, COUNT(*) AS Total_Count
FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

[18]

```
* sqlite:///my_data1.db
Done.
```

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

This query groups mission outcomes and calculates the count for each.
The results show:
•A total of **98 fully successful missions**, demonstrating SpaceX's high success rate.
•A single failure in-flight and one case where the payload's status was unclear, reflecting a minimal number of anomalies in their operations.

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

[19]

```
 * sqlite:///my_data1.db
Done.
```

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

This query identifies boosters that carried the **maximum payload mass**. The result includes multiple boosters, reflecting their advanced payload capacity, which highlights the consistent performance of SpaceX's **Falcon 9 Block 5** series in handling heavy payloads.

2015 Launch Records

```
%%sql
SELECT
    substr(Date, 6, 2) AS Month,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Failure (drone ship)'
  AND substr(Date, 1, 4) = '2015';
```

[20]

```
 * sqlite:///my_data1.db
Done.
```

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 filters for failed landings on drone ships in the year **2015**. The results identify two such incidents:

1. January (Booster Version **F9 v1.1 B1012**)
2. April (Booster Version **F9 v1.1 B1015**)

Both attempts occurred at **CCAFS LC-40**, highlighting challenges faced during early drone ship landing efforts.

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

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

[21]

* sqlite:///my_data1.db
Done.

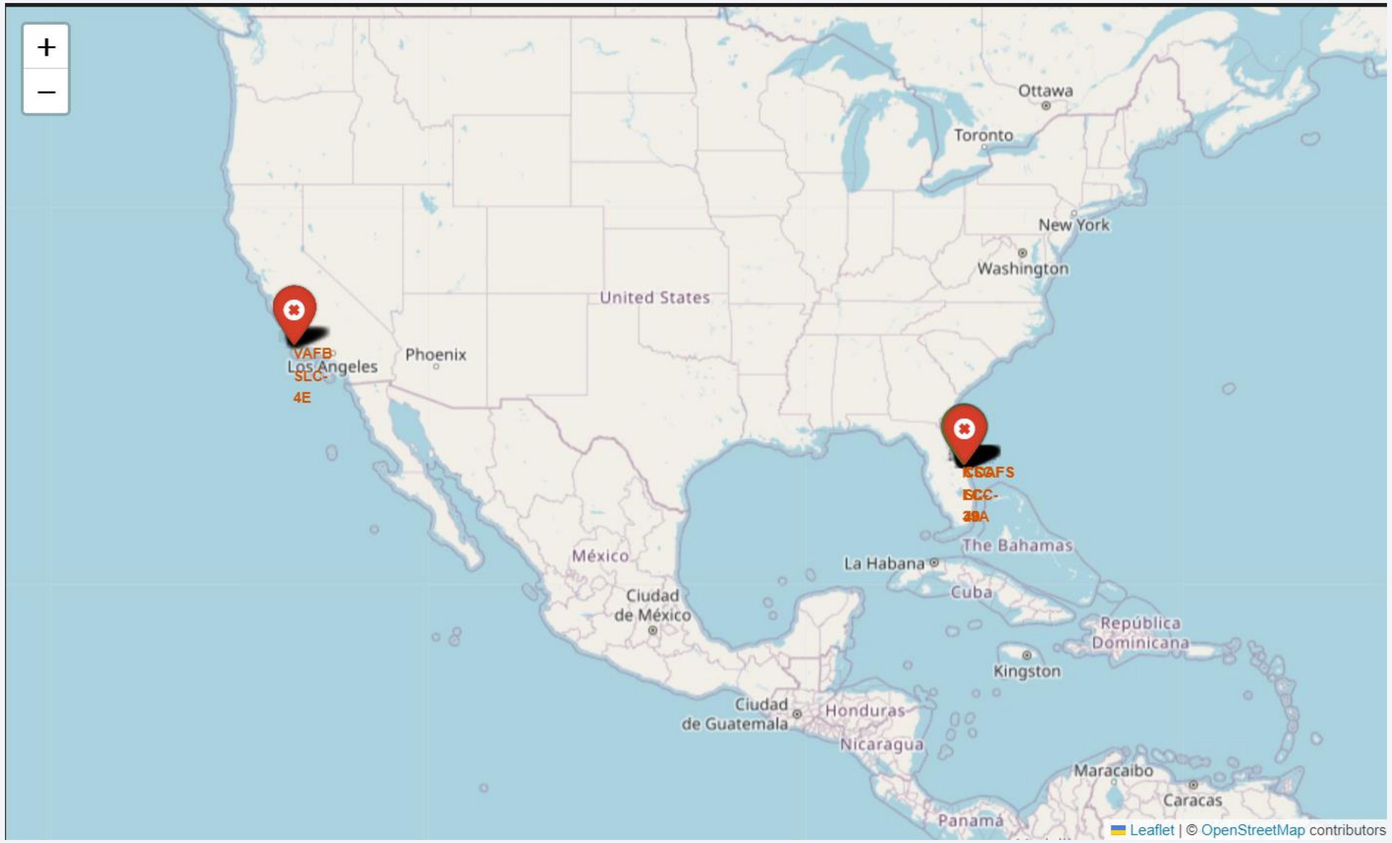
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

This query ranks the count of landing outcomes within the specified time frame (2010-06-04 to 2017-03-20). The highest frequency is **No attempt**, reflecting early launches without landing attempts. Successful landings on drone ships rank second, highlighting significant advancements during this period. Failures and experimental outcomes like "Controlled (ocean)" indicate the iterative process of improving SpaceX's landing technology.

Section 3

Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites



Screenshot Explanation

Markers: Represent SpaceX launch sites:

- **CCAFS LC-40** (Cape Canaveral, Florida)
- **KSC LC-39A** (Kennedy Space Center, Florida)
- **VAFB SLC-4E** (Vandenberg AFB, California)…

Findings

•Launches are U.S.-concentrated, relying on domestic facilities.

•Strategic sites support orbital variety (polar, geostationary) and logistics.

Color-Labelled Launch Outcomes on SpaceX Global Map

Screenshot Explanation

•**Color Labels:**

- **Green:** Successful launches.
- **Red:** Failed launches.
- **Yellow:** Partial/unclear outcomes.

•**Markers:** Represent launch sites with outcomes.

•**Interactive:** Click to view launch details.

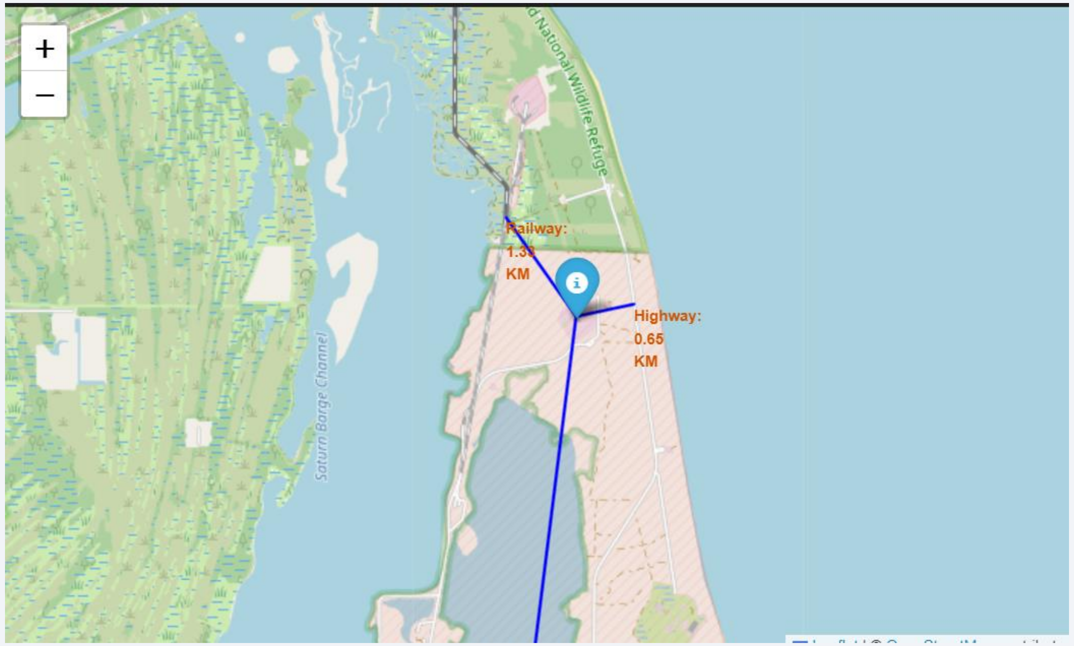
Findings

•Most successes at **KSC LC-39A** and **CCAFS LC-40**.

•Few failures show improved reliability.

•Links geographic locations to outcomes.

Proximity Analysis of Launch Site to Insfracture



Elements Displayed:

- **Railway Distance:** ~1.33 km from the launch site.
- **Highway Distance:** ~0.65 km for site accessibility.
- **City Distance:** ~18.07 km to Cape Canaveral for logistical support.

Lines and Markers:

- Blue lines represent the calculated distances to key infrastructure.
- The map provides an interactive way to view proximity.

Findings:

- Proximity to highways and railways ensures efficient material transport.
- Close coastline proximity optimizes safety and trajectory.
- The nearby city supports operational logistics and staff movement.

Section 4

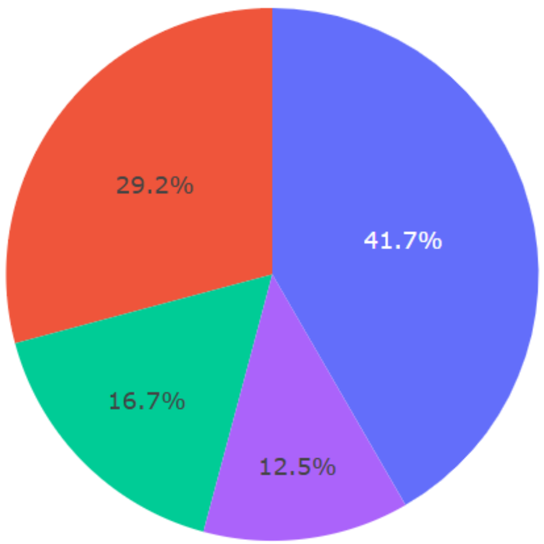
Build a Dashboard with Plotly Dash

Total Success Launches by Site

SpaceX Launch Records Dashboard

All Sites

Total Success Launches by Site



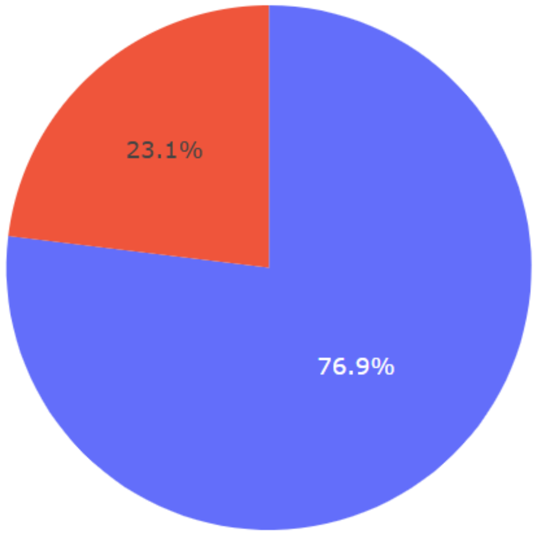
Screenshot Explanation
Pie Chart: Represents the proportion of successful launches at each SpaceX launch site.
KSC LC-39A: 41.7% (highest success rate).
CCAFS LC-40: 29.2%.
VAFB SLC-4E: 16.7%.
CCAFS SLC-40: 12.5%.
Findings
KSC LC-39A dominates in successful launches, indicating its importance in SpaceX operations.
Success is distributed among other sites, showcasing diverse operational capabilities.

Launch Success Ratio for KSC LC-39A

SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A



Screenshot Explanation

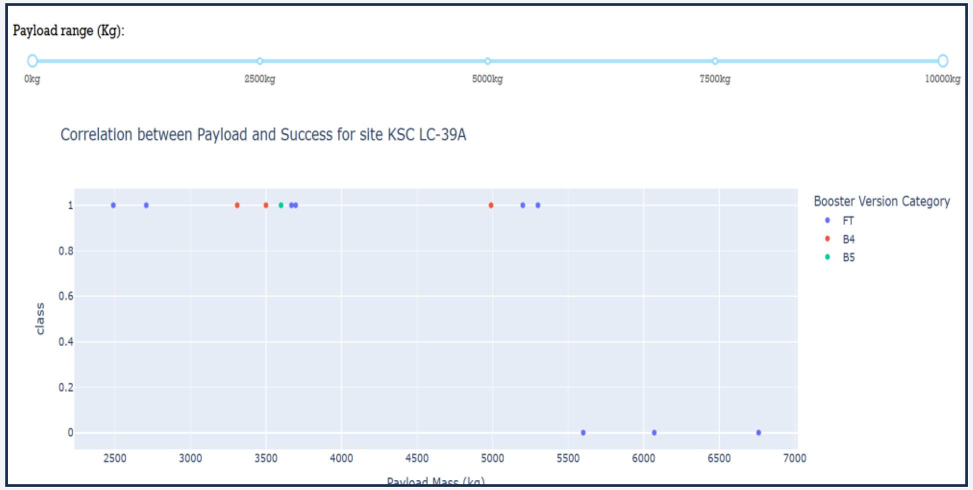
Elements:

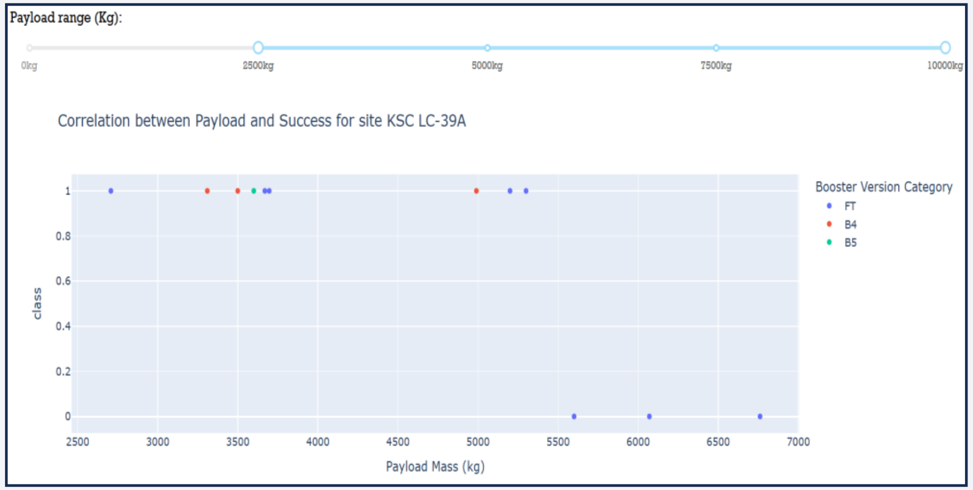
- **Pie Chart:** Visualizes the success ratio of launches from **KSC LC-39A**.
- **Success (Blue):** 76.9% of launches were successful.
- **Failure (Red):** 23.1% of launches failed.

Findings

- **KSC LC-39A** has the highest success ratio among all SpaceX launch sites.
- Highlights the site's reliability and technological maturity in supporting successful missions.

Payload vs Launch Outcome Across Sites with Range Selection





Payload vs Launch Outcome Across Sites with Range Selection

Explanation of Screenshots:

1. **Screenshot 1 (Empty Range):**
   - Payload range is set too narrow (e.g., 0-2500kg), resulting in no visible data points.
   - Highlights the importance of adjusting range for meaningful insights.
2. **Screenshot 2 (Moderate Payload):**
   - Focused on payloads between **5200-6800kg**.
   - Successes dominate, particularly for **FT Booster Versions**.
   - Demonstrates high reliability in handling mid-range payloads.
3. **Screenshot 3 (Expanded Payload):**
   - Payload range widened to **2500-7000kg**.
   - Successful launches observed across multiple booster versions (FT, B4, B5).
   - Some failures noted, particularly at higher payloads.
4. **Screenshot 4 (All Payloads):**
   - Covers the full payload range **0-10,000kg**.
   - Majority of launches with high payloads are successful.
   - Failures align with experimental or new configurations.

Findings:

- **Success Rates:** Best for payloads in the **4000-6000kg** range.
- **Booster Versions: FT and B5** dominate successful launches.
- **Failures:** Increase slightly for extreme payloads (>6000kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

```
Test Accuracies of Models:
Logistic Regression: 0.94
Support Vector Machine: 0.89
Decision Tree: 0.78
k-Nearest Neighbors: 0.94

The best-performing model is: Logistic Regression
Accuracy: 0.94
```

Based on the test accuracies:

•**Logistic Regression** and **k-Nearest Neighbors** both achieved the highest accuracy of **0.94**.

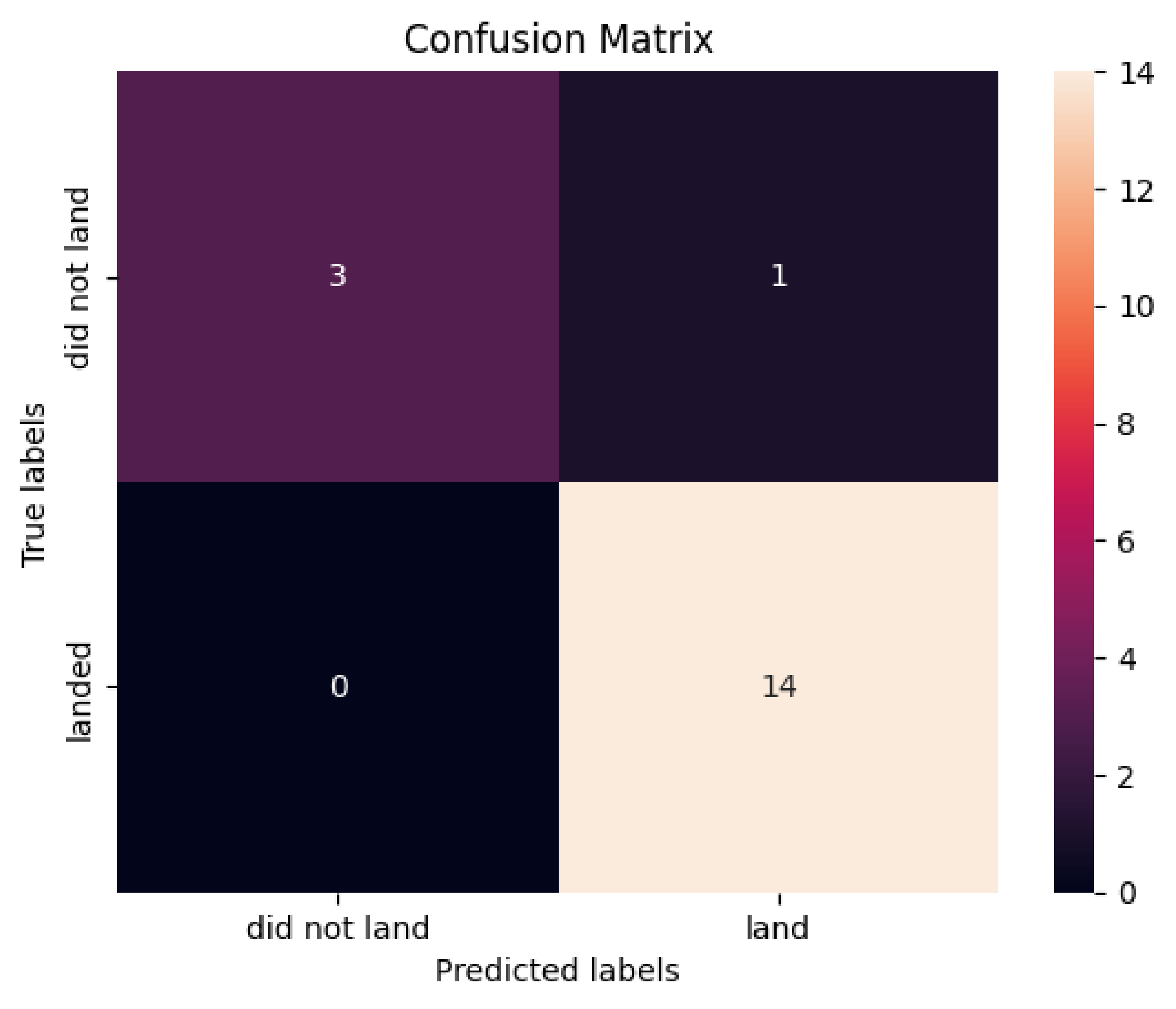
•**Support Vector Machine** followed with an accuracy of **0.89**.

•**Decision Tree** had the lowest accuracy of **0.78**.

Best-Performing Model:

•The **best-performing model** is **Logistic Regression**, given its simpler interpretability and robust performance. However, k-NN could also be considered as it achieved the same accuracy.

Confusion Matrix



This confusion matrix visualizes the performance of the **Logistic Regression** model:

•**True Positives (Bottom Right - 14):** The model correctly predicted 14 cases where the first stage successfully landed.

•**True Negatives (Top Left - 3):** The model correctly predicted 3 cases where the first stage did not land.

•**False Positives (Top Right - 1):** The model incorrectly predicted 1 case as a successful landing, but it did not land.

•**False Negatives (Bottom Left - 0):** The model did not make any errors predicting successful landings as failures.

Key Findings:

•The model shows excellent precision and recall for predicting successful landings.

•Only one false positive indicates the model slightly overestimates landings.

•No false negatives indicate the model reliably identifies successful landings.

Conclusions

SpaceX launch data was collected, cleaned, and analyzed using tools like APIs, SQL, Python, and interactive visualizations. Key trends were identified, such as higher success rates at launch sites like KSC LC-39A and CCAFS LC-40, and the significance of payload mass and orbit types in influencing mission outcomes.
Predictive modeling was conducted using Logistic Regression, SVM, Decision Tree, and k-NN. Logistic Regression and k-NN emerged as the most effective models, achieving an accuracy of **94%**. These models proved valuable in predicting first-stage landings, demonstrating the potential for data-driven optimization in space missions.
SQL-based analysis uncovered that moderate payloads between 2000–6000 kg and launches targeting orbits such as LEO, ISS, and SSO were more likely to succeed. Additionally, insights highlighted the reliability of specific launch sites like CCAFS LC-40 and KSC LC-39A.
Mapping and proximity analyses showcased the strategic placement of launch sites near critical infrastructure such as highways, railways, and coastlines. These placements ensure streamlined logistics, operational efficiency, and cost-effectiveness in launch activities.
Overall, the project successfully integrated data science methods to derive actionable insights from raw data, enhance prediction models, and support operational decision-making. SpaceX's success in reusable rockets reflects the transformative potential of combining technology and data science in the aerospace industry.

Thank you!