# SpaceX Falcon 9 First Stage Landing Prediction

A Comprehensive Data Science Analysis

**Author:**

Gabriel Santiago Murillo Barragán

**Affiliation:**

Biomedical Engineering, Universidad de los Andes

Medical Student, Universidad Nacional de Colombia

**Email:**

gsmurilloba@unal.edu.co

July 17, 2025

# Contents

# Executive Summary

This report presents a thorough analysis of SpaceX Falcon 9 launch data with the objective of predicting whether the first stage will land successfully. By collecting and cleaning data from multiple sources (including the official SpaceX API and Wikipedia), performing exploratory data analysis (EDA), and applying various predictive models, we identify the critical factors that affect landing success. The findings highlight:

- **The critical role of launch site characteristics**, with certain sites like KSC LC-39A demonstrating significantly higher success rates.
- **The impact of payload mass and orbit type**, revealing complex relationships where higher payloads do not strictly correlate with failure, but certain orbits present greater challenges.
- **The iterative improvements of the booster versions**, which show a clear positive trend in landing success over time, underscoring the value of operational learning.

The predictive analysis culminated in a Decision Tree model as the most effective classifier after hyperparameter tuning, providing a robust framework for forecasting landing outcomes. This study provides actionable, data-driven insights that can inform operational strategies and enhance risk management in future space launch planning.

# Chapter 1

# Introduction

## 1.1 Background

The advent of reusable rocket technology has marked a paradigm shift in the aerospace industry, significantly reducing launch costs and improving the sustainability of space exploration. Historically, space launch systems were expendable, meaning each mission required a newly manufactured rocket, which incurred enormous costs and resource consumption [1].

The introduction of SpaceX's Falcon 9 revolutionized the industry by demonstrating that a reusable first stage could land successfully and be relaunched multiple times. This breakthrough has fundamentally altered the economics of spaceflight, reducing the cost per kilogram of payload to orbit and enabling more frequent launches [2]. Reusability has now become a core design feature for modern launch vehicles, with companies like Blue Origin, Rocket Lab, and Relativity Space investing in similar technologies [3, 4].

However, first-stage landing success is influenced by multiple variables, including fuel efficiency, atmospheric conditions, thrust vector control, and landing pad precision [5]. The development of machine learning models to predict landing success has gained traction, as these algorithms can analyze vast datasets to identify patterns and optimize landing parameters [6]. Advances in deep learning and reinforcement learning have improved trajectory optimization and landing stability, making automated rocket landings more reliable than ever before [7].

The objective of this project is to analyze key historical launch data to identify determinants of Falcon 9 first-stage landing success using data science and machine learning techniques. By leveraging exploratory data analysis (EDA) and predictive modeling, we aim to extract insights that could contribute to future advancements in reusable launch systems.

## 1.2 Objectives

### 1.2.1 General Objective

To identify and characterize the key factors influencing the successful landing of the Falcon 9 first stage through data-driven analysis, enabling the development of predictive models that enhance landing reliability and contribute to cost-effective space operations.

### 1.2.2 Specific Objectives

1. **To determine the influence of technical, environmental, and operational variables on the success rate of Falcon 9 first-stage landings.** By analyzing historical launch data, this objective aims to establish statistical relationships between independent factors (e.g., launch site, payload mass, orbit type) and landing outcomes.

2. **To assess the predictive power of machine learning models in forecasting landing success.** This includes evaluating the accuracy, interpretability, and applicability of different algorithms in predicting first-stage recovery, contributing to improved decision-making in launch operations.

## 1.3 Report Structure

This document is organized as follows:

- **Chapter 2: Methodology** - Describes the research approach, including data sources, preprocessing steps, feature selection, and the machine learning models employed.
- **Chapter 3: Results** - Presents the outcomes of exploratory data analysis (EDA), statistical insights, and the predictive model performance.
- **Chapter 4: Discussion** - Interprets the results in the context of aerospace operations, discusses the practical implications, and identifies limitations.
- **Chapter 5: Conclusion** - Summarizes the major contributions of the study and proposes directions for future work.

# Chapter 2

# Methodology

## 2.1 Data Collection

To compile a comprehensive dataset of Falcon 9 launches, multiple data sources were leveraged to ensure data completeness and accuracy:

- An open-source SpaceX API providing launch data in JSON format.
- HTML tables from a publicly available Wikipedia page.
- Supplementary datasets from the IBM Data Scientist Coursera Capstone in CSV format.

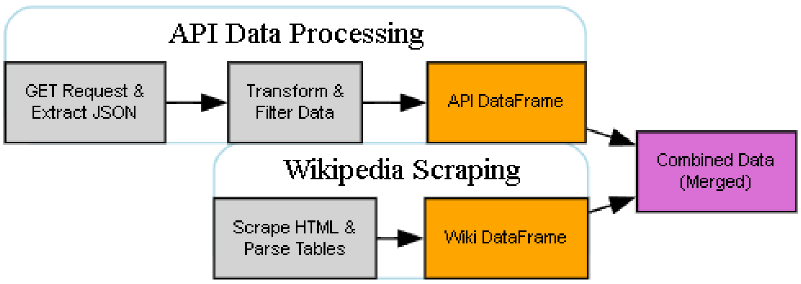
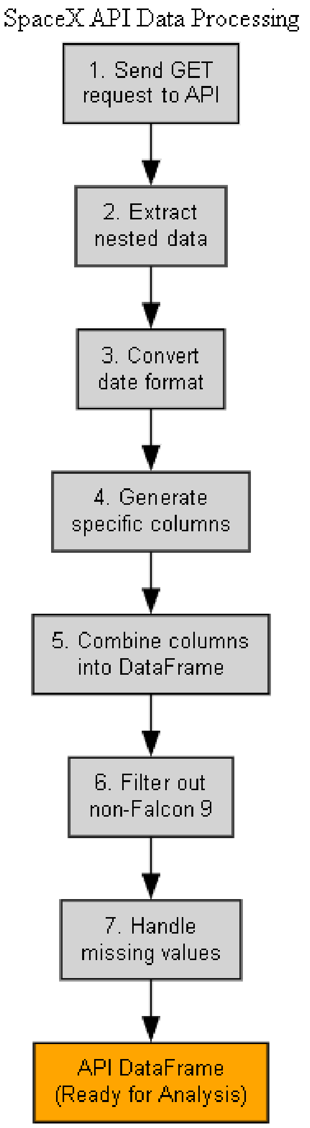


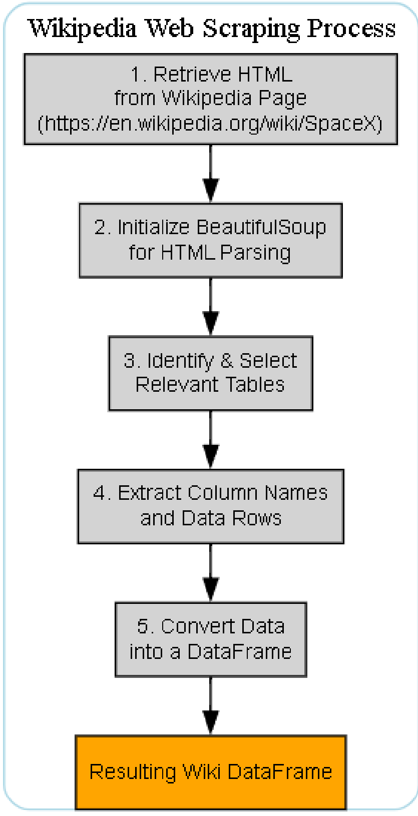

Figure 2.1: Data collection and processing workflow.

### 2.1.1 Data Acquisition via SpaceX API and Web Scraping

Data from the SpaceX API (https://api.spacexdata.com/) was extracted to obtain structured information on rocket configurations and mission outcomes. Complementary historical launch data was obtained by scraping HTML tables from Wikipedia using `BeautifulSoup`. This dual approach enriched the dataset by including additional mission details and contextual information.




(a) API Data Processing




(b) Web Scraping Workflow

Figure 2.2: Comparative workflows for data acquisition.

## 2.2 Data Wrangling

A dedicated data wrangling process was conducted to produce a consolidated and analytically robust dataset. Key steps included:

- **Handling Missing and Inconsistent Records:** Imputing or removing records with incomplete information to maintain data reliability.
- **Data Type Alignment:** Casting attributes like `launch_date_utc` into `datetime` objects for chronological analyses.
- **Categorical Encoding:** Mapping mission outcomes into binary indicators (1

for success, `0` for failure) and encoding other categorical variables using one-hot schemes.

- **Data Integrity Checks:** Identifying and removing duplicate records and performing outlier analysis on key metrics.

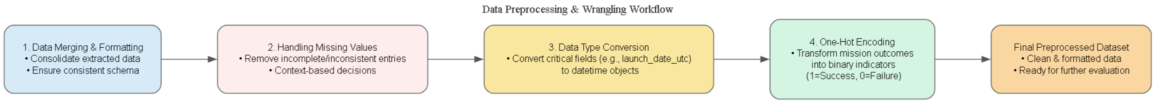



Figure 2.3: Data Wrangling workflow.

## 2.3 Exploratory Data Analysis (EDA)

This study employed a two-tiered exploratory approach: (1) descriptive statistics to summarize overall data characteristics and (2) visual analytics to uncover underlying relationships and patterns. This included creating bar charts, scatter plots, and time series plots to analyze trends related to launch sites, orbit types, and payload mass. SQL queries were also used to extract targeted insights and validate data integrity.

## 2.4 Predictive Modeling

### 2.4.1 Model Selection

To predict the success of the first-stage landing, several supervised learning algorithms were evaluated: Logistic Regression, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (kNN).

### 2.4.2 Data Splitting and Hyperparameter Optimization

The dataset was partitioned into training and testing sets using an 80/20 split. Hyperparameter tuning was performed using `GridSearchCV` to identify the optimal parameter combinations for each algorithm, significantly enhancing model performance.

### 2.4.3 Evaluation Metrics

The performance of the predictive models was assessed using standard classification metrics: Accuracy, Precision, Recall, and F1-Score. A confusion matrix was also generated to visualize the performance of the best model.

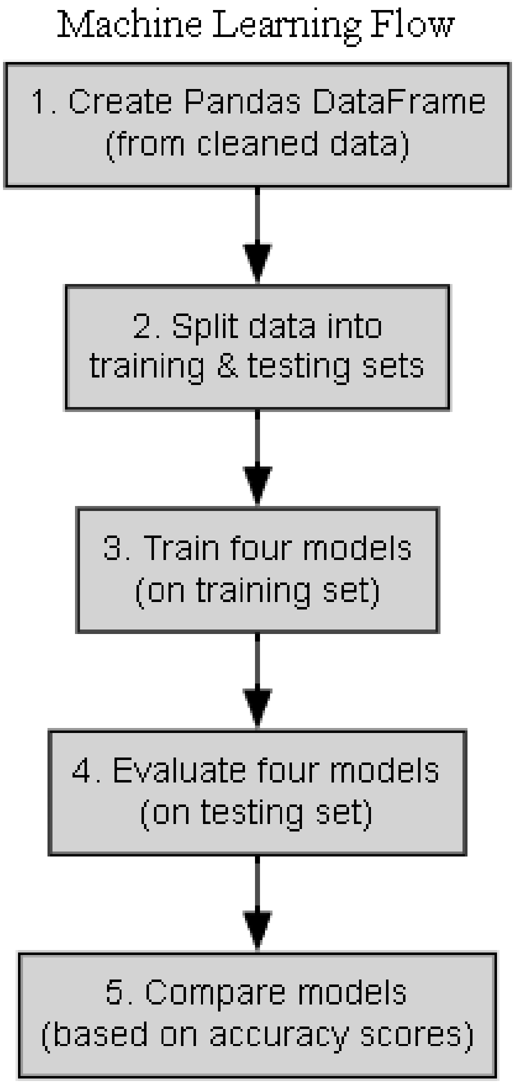



Figure 2.4: Machine learning workflow.

# Chapter 3

# Results

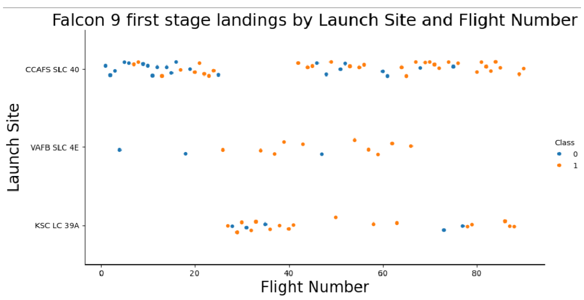
## 3.1 Exploratory Data Analysis Findings

The EDA revealed several significant trends influencing the success of Falcon 9 first-stage landings. Visual analysis highlighted strong correlations between landing outcomes and operational variables such as launch site, payload mass, and orbit type.

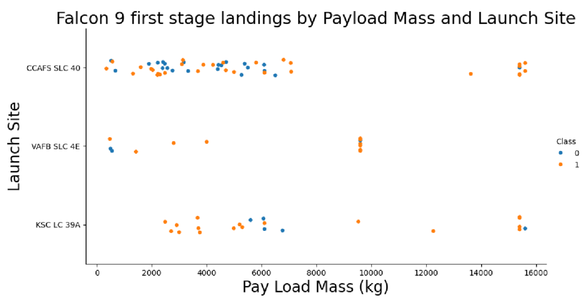
### 3.1.1 Launch Site and Payload Analysis

As shown in Figure 3.1a, a clear learning curve is evident, with a higher proportion of successful landings (Class 1) at higher flight numbers across all sites. The CCAFS SLC-40 site has the highest number of launches, while KSC LC-39A shows a high density of successful landings in its operational history.

Figure 3.1b illustrates the relationship between payload mass and landing success. Notably, there is no simple linear relationship suggesting heavier payloads lead to failure. Instead, successful landings are distributed across a wide range of payload masses, although certain sites, like CCAFS SLC-40, handle a broader spectrum of payloads.




(a) Landing outcomes by Flight Number and Launch Site.



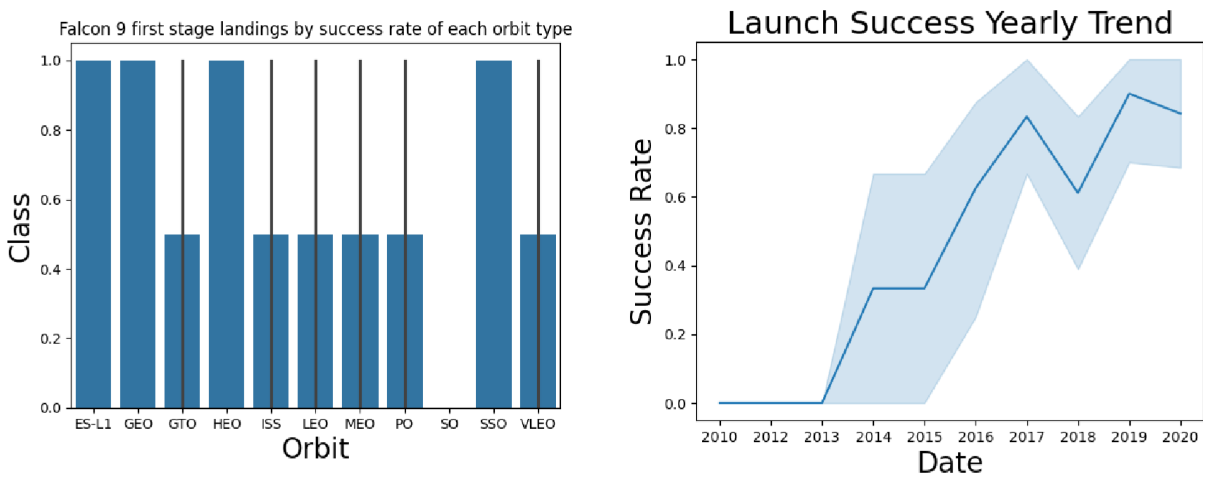

(b) Landing outcomes by Payload Mass and Launch Site.

Figure 3.1: Analysis of landing success based on launch site.

### 3.1.2 Orbit and Temporal Trend Analysis

The success rate varies significantly by orbit type (Figure 3.2a). Orbits such as ES-L1, GEO, HEO, and SSO demonstrate a 100

The yearly trend, depicted in Figure 3.2b, confirms a steady improvement in landing success over time. The success rate climbed from zero in the initial years to consistently above 80




(a) Success rate for each orbit type.

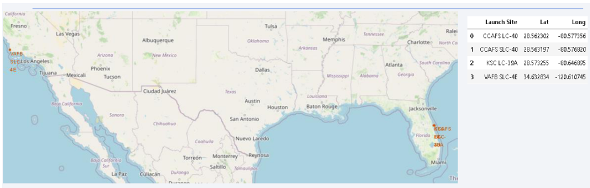
(b) Yearly trend of launch success rate.

Figure 3.2: Analysis of landing success based on orbit and time.

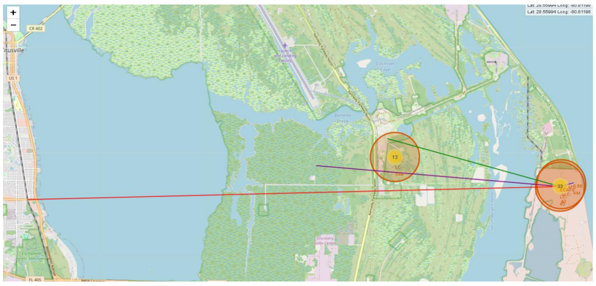
### 3.1.3 Geospatial Analysis with Folium

To better understand the operational context of the launch sites, an interactive geospatial analysis was performed using the Folium library. This allowed for the visualization of launch pads in relation to key infrastructure and geographical features, as shown in Figure 3.3.

The analysis confirmed that launch sites are strategically located with close access to essential transport infrastructure while maintaining a safe buffer from major urban centers. For instance, the CCAFS LC-40 site is approximately 1.33 km from the nearest rail line and only 0.19 km from a perimeter road, ensuring logistical efficiency. Its proximity to the coastline (0.92 km) is advantageous for over-water launch trajectories, minimizing risks to populated areas.




(a) Geographic location of Falcon 9 launch sites.



(b) Proximity analysis from CCAFS LC-40 to nearby infrastructure.

Figure 3.3: Geospatial visualizations of launch sites and infrastructure.

## 3.2 Predictive Model Performance

After training and hyperparameter tuning, the four selected classification models were evaluated on the test set. While all models performed well, the Decision Tree classifier was identified as the optimal model after hyperparameter tuning.

Table 3.1 summarizes the performance metrics for each model before tuning. All four models achieved an identical accuracy of 83.33

Table 3.1: Comparison of Classification Model Performance.

| Metric | LogReg | SVM | Decision Tree | kNN |
|---|---|---|---|---|
| Accuracy | 0.8333 | 0.8333 | 0.8333 | 0.8333 |
| Jaccard Score | 0.8000 | 0.8000 | 0.7500 | 0.8000 |
| F1-Score | 0.8889 | 0.8889 | 0.8571 | 0.8889 |

The confusion matrix for the best-performing model (Decision Tree after tuning) is shown in Figure 3.4. The model correctly identified 12 out of 12 successful landings (perfect recall for the positive class) but misclassified 3 failed landings as successful (false positives). This indicates the model is highly effective at identifying successful landings but has a tendency to be optimistic in its predictions.

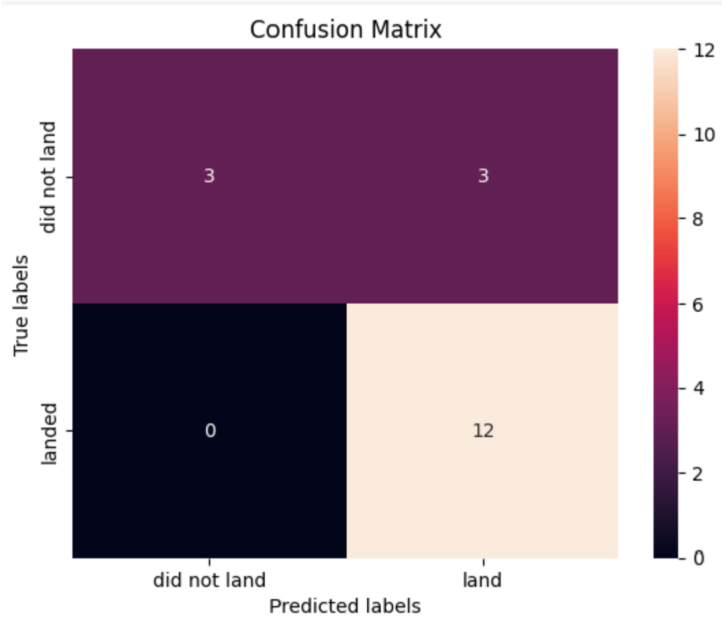



Figure 3.4: Confusion matrix for the best performing model (Decision Tree).

# Chapter 4

# Discussion

## 4.1 Interpretation of Findings

The results of this study provide a data-driven confirmation of several widely understood principles in aerospace operations while also revealing nuanced insights. The strong correlation between increased flight numbers and higher success rates (Figure 3.1a) is a clear quantitative representation of the "learning by doing" philosophy that underpins SpaceX's iterative design and operational strategy. Each launch provides valuable data that informs subsequent engineering and procedural refinements.

The variability in success rates across different launch sites suggests that factors beyond the rocket itself, such as ground support infrastructure, local atmospheric conditions, and range safety protocols, play a significant role. The high performance of KSC LC-39A, a site with a long history of high-profile launches, may reflect a more mature operational environment compared to other sites.

Perhaps one of the most interesting findings is the complex relationship between payload mass, orbit, and success. The absence of a simple negative correlation between payload mass and landing success indicates that SpaceX has successfully engineered the Falcon 9 to be robust across a wide operational envelope. However, the lower success rates in certain orbits like GTO highlight that the mission profile—which dictates the velocity, altitude, and fuel margins required for the landing burn—is a more critical determinant of success than payload mass alone.

The high accuracy (83.33

## 4.2 Limitations of the Study

Despite the comprehensive approach, this study is subject to several limitations that should be considered when interpreting the results:

- **Data Granularity:** The dataset, while comprehensive, lacks highly granular telemetry data from the launch vehicles, such as real-time engine performance, fuel levels, or atmospheric sensor readings during descent. The inclusion of such data would likely improve model accuracy significantly.

- **Exclusion of External Variables:** Key external factors, most notably weather conditions (e.g., wind shear, precipitation, sea state for drone ship landings), were not included in the dataset. These variables are known to have a substantial impact on landing operations.
- **Static Dataset:** The analysis is based on a historical snapshot of launches. As SpaceX continues to launch and update its hardware and software, the statistical patterns identified in this report may evolve.

## 4.3 Future Work

Building on the findings and limitations of this study, several avenues for future research are proposed:

- **Integration of Weather Data:** Future work should prioritize enriching the dataset with historical weather data corresponding to each launch, including upper-level winds, sea states, and visibility.
- **Advanced Feature Engineering:** Incorporating more complex features, such as the interaction between booster version and orbit type or the time elapsed since a booster's last flight, could uncover more subtle relationships.
- **Exploration of Advanced Models:** While the classic models performed well, exploring more advanced techniques like Gradient Boosting Machines (e.g., XGBoost, LightGBM) or deep learning models (e.g., LSTMs for time-series analysis of launch cadence) could yield higher predictive accuracy.
- **Risk Analysis:** The model's outputs could be used to develop a quantitative risk assessment tool. Instead of a binary prediction, the model could output a probability of success, allowing mission planners to make more informed decisions based on their risk tolerance.

# Chapter 5

# Conclusion

This comprehensive report demonstrates that a multi-faceted data science approach—encompassing multi-source data collection, rigorous data wrangling, detailed exploratory analysis, and predictive modeling—can yield significant insights into the determinants of SpaceX Falcon 9 first-stage landing success. The study successfully identified and characterized key factors such as launch site, payload mass, orbit type, and flight number as strong predictors of landing outcomes.

The exploratory data analysis quantitatively confirmed the positive impact of operational experience over time and highlighted the nuanced interplay between mission parameters. The predictive modeling phase achieved a high accuracy of 83.33

While the study acknowledges limitations, such as the absence of granular telemetry and weather data, it establishes a robust and reproducible framework for future analysis. The findings not only contribute to a deeper understanding of reusable rocket operations but also provide actionable intelligence that can support cost optimization, enhance risk management, and inform strategic decision-making in the increasingly competitive aerospace industry. Future work should focus on enriching the dataset and exploring more complex models to further refine these predictions and build upon the solid foundation established by this research.

# Bibliography

[1] D. Rasky, R. B. Pittman, and M. Newfield, "The reusable launch vehicle challenge," in *Space 2006*, 2006, p. 7208.

[2] E. Musk, "Making humans a multi-planetary species," *New Space*, vol. 5, no. 2, pp. 46–61, 2017.

[3] D. Watson, "Reusable launch vehicles: business & societal interest," 2024.

[4] D. Montes Martínez, "Study of reusable launch vehicles business model: case study of electron from rocket lab," B.S. thesis, Universitat Politècnica de Catalunya, 2024.

[5] L. Blackmore, "Autonomous precision landing of space rockets," *The Bridge*, vol. 46, no. 4, pp. 15–20, 2016.

[6] U. Çelik and M. U. Demirezen, "Optimal reusable rocket landing guidance: A cutting-edge approach integrating scientific machine learning and enhanced neural networks," *IEEE Access*, vol. 12, pp. 16 805–16 829, 2024.

[7] K. Hovell and S. Ulrich, "Deep reinforcement learning for spacecraft proximity operations guidance," *Journal of spacecraft and rockets*, vol. 58, no. 2, pp. 254–264, 2021.

# Acknowledgments

The author wishes to thank IBM and Coursera for providing such a complete coursework that allowed me to truly understand every step of the data management process.

# Appendix A

# Appendix A: SQL Queries

This appendix includes examples of SQL queries used during the exploratory data analysis phase to extract specific information from the dataset.

```
-- Task: Display the names of the unique launch sites
SELECT DISTINCT(LAUNCH_SITE)
FROM SPACEXTBL;

-- Task: Display the total payload mass carried by NASA (CRS)
SELECT SUM(PAYLOAD_MASS_KG_)
FROM SPACEXTBL
WHERE CUSTOMER = 'NASA (CRS)';

-- Task: List the date of the first successful ground landing
SELECT MIN("Date") AS "first_successful_landing"
FROM SPACEXTBL
WHERE "Landing_Outcome" = 'Success (ground pad)';
```