A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Lorrany Inacio Cardoso dos Santos |
| *Student Number* | 2024062 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA2 Capstone Report |
| *Assessment Due Date* | 19th May 2024 |
| *Date of Submission* | 19th May 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Summary

[1. Introduction 1](#_Toc167002702)

[1.1. Objective 2](#_Toc167002703)

[1.2. Problem Definition 2](#_Toc167002704)

[2. Development 2](#_Toc167002705)

[2.1. Data Understanding and Data Preparation 2](#_Toc167002706)

[2.2. Modeling 4](#_Toc167002707)

[2.3. Data Sources and Ethical Considerations 6](#_Toc167002708)

[Conclusion 7](#_Toc167002709)

[References 9](#_Toc167002710)

# Introduction

Digital technologies have revolutionized organizational operations, prompting the need for innovative solutions and diverse applications. Despite the broad scope of technology, this research project focuses specifically on Data Analysis as the proposed solution to the identified challenge. According to the UCD website, Data Analytics involves analyzing raw data to draw conclusions about business decisions. Data Analytics is essential for predicting trends, understanding customer behavior, and optimizing budget allocation.

Karen and Henry explain that "data-informed" and "data-driven" are often used to discuss how data analytics helps organizations make decisions. Despite that, these two terms mean different things. Then, it is important to understand their differences and similarities (WEBBER & ZHENG,2020). It is about using decision algorithms, heuristics, and rules to make decisions. They also explain that this approach focuses on letting data guide decisions, minimizing the influence of human factors. "Data-Informed Decision Making" (DIDM) is a newer concept. It is about using data to understand situations better and provide evidence for decisions. Using data to figure out what it tells us, helps us make better decisions (WEBBER & ZHENG,2020).

Harold suggests that effective project management commences with delineating the project itself. A project, in essence, embodies a sequence of tasks and activities with a distinct goal to be achieved within predefined parameters, such as specified objectives, clear start and end dates, and, if pertinent, financial constraints. The essence of project management lies in optimizing resource utilization by fostering both horizontal and vertical flow of work within the organization. This methodology doesn't seek to dismantle the hierarchical structure but rather encourages interdepartmental collaboration to ensure seamless progress across the organization (KERZNER,2017).

According to Nick in the realm of data science, project manager holds the responsibility of overseeing the successful execution of advanced analytics and AI/Machine Learning projects. While the essence of their role aligns with traditional IT project management, it is distinguished by a focus on the intricacies of data science applications. Key tasks encompass: communicating comprehensive project roadmaps, coordinating and supervising the day-to-day activities and workflows of project teams, addressing stakeholders, defining project tasks in line with the project's vision, managing and documenting scope, and identifying and collating essential datasets required for project implementation. Additionally, some roles may necessitate technical proficiency, where the data science project manager is expected to be adept in Python, and visualization tools (HOTZ,2024).

Nick says that clients anticipate that project management will guarantee the precise fulfillment of their needs regarding the delivered data analytic solutions. This entails solutions that proficiently tackle the designated business issues or opportunities. Data analytics initiatives must harmonize with the overarching business aims and strategic objectives of the company. Clients expect project management to certify that projects are synchronized with these objectives and play a role in their attainment (HOTZ,2024).

# Objective

The main objective of this project is to develop and implement a plan aimed at using data tools to enhance decision-making processes. This includes improving decision-makers' abilities to make informed choices and achieve better overall outcomes. The plan encompasses various aspects of data analytics, such as understanding data, utilizing appropriate tools, and selecting optimal machine learning models. Additionally, a key secondary objective involves conducting thorough data analysis to identify patterns and trends among competitors, thereby refining the project's scope and providing a more accurate and objective outcome.

# Problem Definition

Therefore, with the increasing demand for utilizing data and taking advantage of its opportunities, organizations are seeking clear and simple solutions and guidelines for data management. Accordingly, the project problem of this paper is " The multifaceted influence of various attributes such as location, property type, and amenities, on the dynamic pricing trends within the Airbnb platform ".

## Development

## Data Understanding and Data Preparation

Before diving into characterization, it is essential to understand the nature and context of the data. To understand the dataset we use some codes that make it possible to analyze each piece of information in depth. In the figure below used .head() that shows each column and what they represent.

Understanding the business context helps ensure that the data analytics project is aligned with the organization's overarching objectives. This alignment ensures that the insights derived from the analysis are relevant and actionable for the business. This phase is understanding the goals and constraints and defining the common problems the AI model needs to solve.

Rose emphasizes the importance of understanding data before diving into analysis. Doug Rose stresses the significance of data understanding as a foundational step in the data science process. He emphasizes that before diving into analysis or modeling, it is crucial to thoroughly understand the data being used. Rose discusses various aspects of data understanding that will be talked about in this project.

Using data analysis project management methodologies, it is possible to achieve good development throughout a data lifecycle. Initially, it is necessary to select a dataset to serve as the basis for the entire project. To understand the dataset, you need some codes to deeply understand each data. The most important ones will only be discussed below:

• First code: “.head()”, this function shows the first n rows and all the columns. It is useful for quickly checking of data in it, as shown in table 1 below;

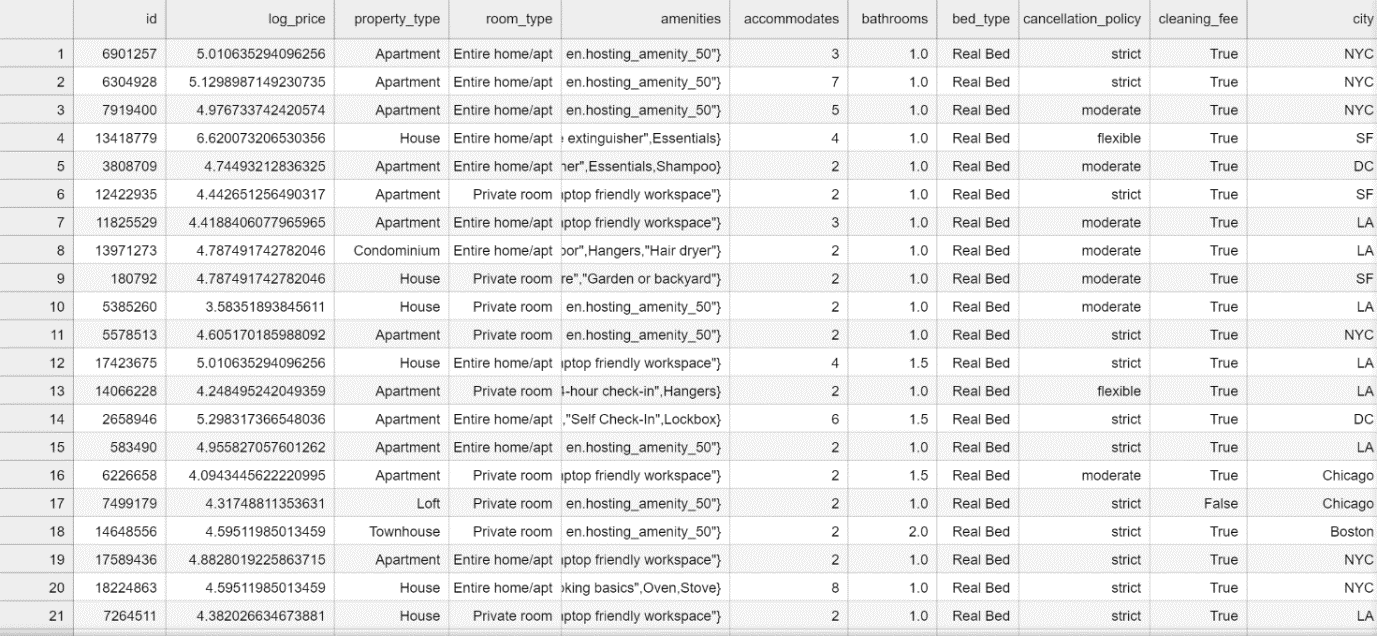


Table 1 (Jupyter Notebook)

• Second code: “.shape”, this function shows the total of rows and columns;

• Third: “.dtypes”, with this code you can identify if columns are numerical or categorical.

Deepening the analysis and making understanding more visual, there are some functions that can be used to facilitate this action. Conducting EDA to gain insights into the structure, distribution, and relationships within the data.

Using the graphical library matplotlib, it is possible to visualize the data relating to the price column through a histogram in Figure 1.

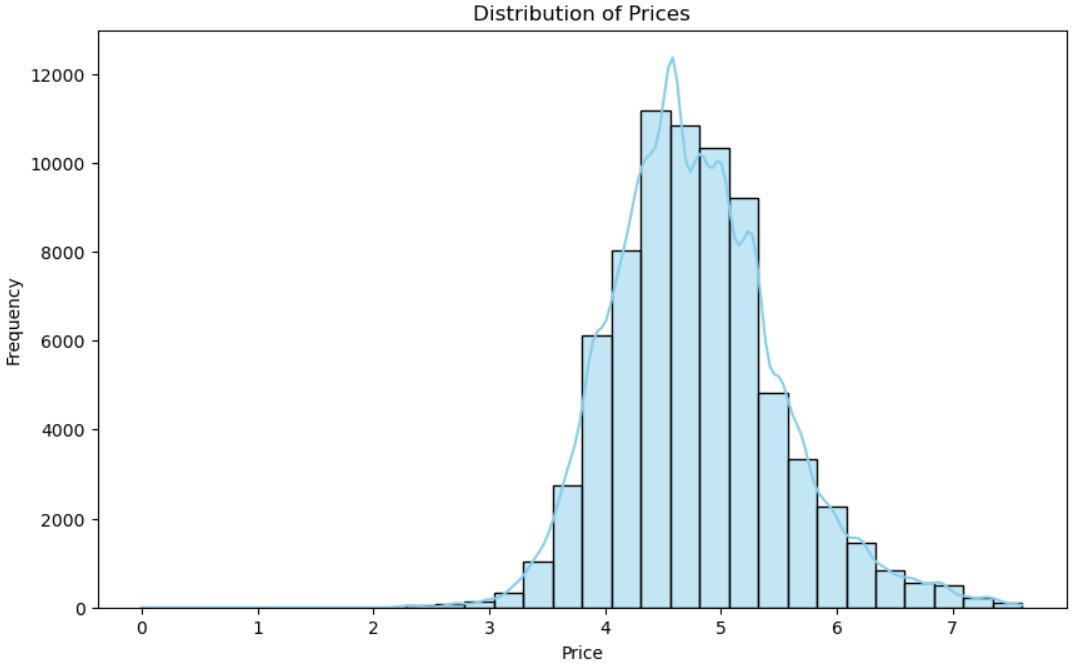


Figure 1 (Jupyter Notebook, 2024)

Through box plot shows the prices by room type, followed by the figure 2.

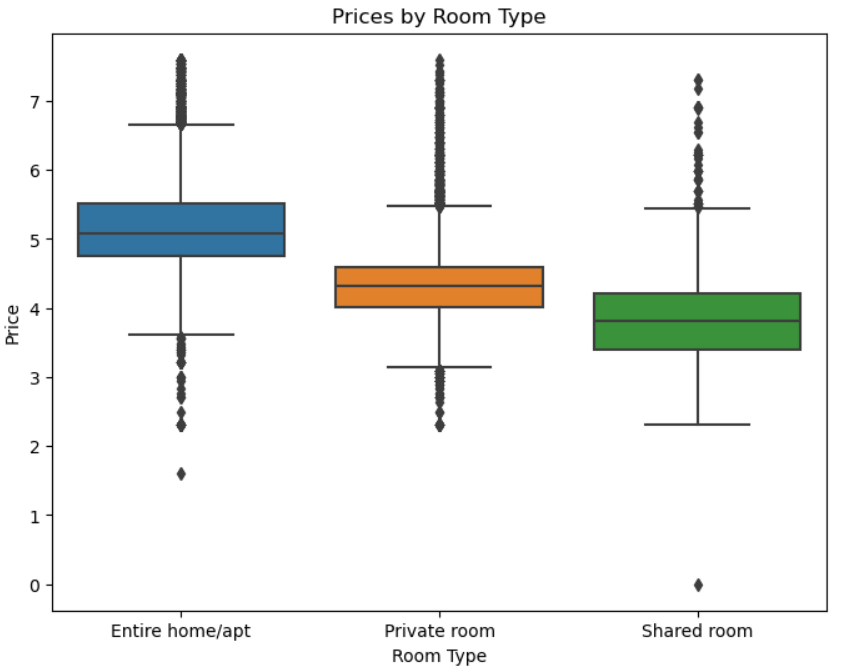


Figure 2 (Jupyter Notebook, 2024)

Another fundamental point in all data analysis is Data Preparation. Moment to which most of the project is dedicated. Knowing this, a good data project manager needs to use means that shorten this data cleaning time and this can be done through using precise codes. Table 2 below obtained using the same code (.head) is the result of a complete data cleaning, where columns were removed, null values ​​were removed, words were exchanged for numbers and several other necessary actions, which only during the entire process will you identify, and adapting to the project.

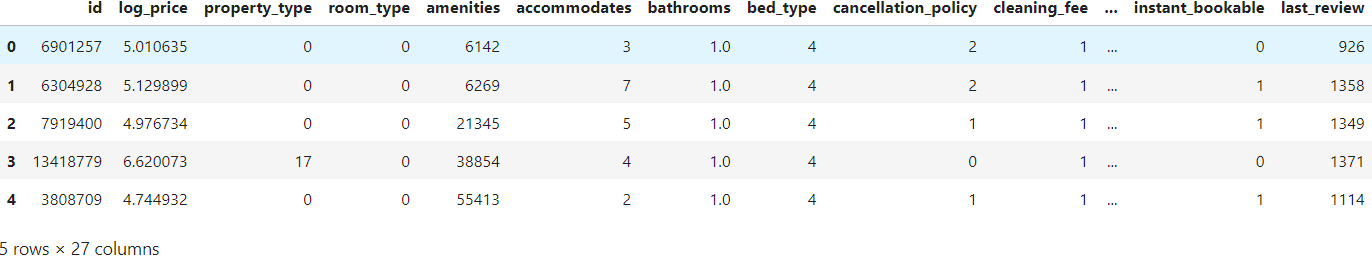


Table 2 (Jupyter Notebook, 2024)

## Modeling

First of all, it is necessary to introduce what Machine Learning is. According to Zhi-Hua, Machine learning is a method that enhances how systems work by learning from past experiences through computer-based techniques. In computer systems, these past experiences are in the form of data. The primary goal of machine learning is to create algorithms that learn from this data to build models. By giving the learning algorithm access to this data, we create a model capable of making predictions based on new observations (ZHOU, 2016).

Which algorithms to try?

Considering the objective of the project, a Linear Regression was chosen for modeling. According to the Scikit-Learn library Linear Regression constructs a linear model represented by coefficients w = (w1, …, wp) to minimize the sum of squared differences between the actual target values in the dataset and the target values predicted by the linear model.

The following Figure 3 shows the relationship between two important variables that make us understand the dataset even more.

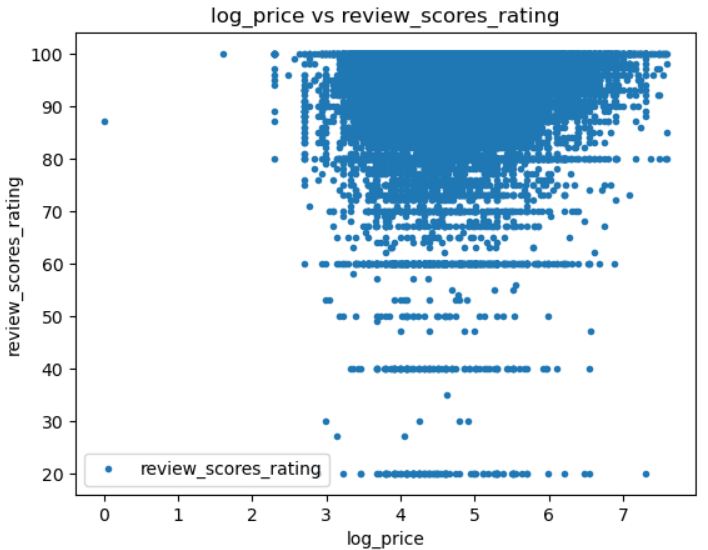


Figure 3 (Jupyter Notebook, 2024)

## Data Sources and Ethical Considerations

In conducting this capstone project, General Data Protection Regulation (GDPR) will address several key ethical considerations. Firstly, data privacy measures must be implemented to safeguard sensitive information, including obtaining appropriate permissions for data access, anonymizing personal data, and securely storing data to prevent unauthorized access.

The dataset in the study will be collected for free on the Kaggle website. Kaggle is a place where people who work with data and machine learning can compete and connect online. On Kaggle, you can find data, create and test models, and collaborate with others to solve problems (MUO. 2023-04-17).

Through the Kaggle website, an excellent platform for data research, from which the dataset under study was taken. Paramvir the owner provides a comprehensive view of a wide range of Airbnb accommodations across the globe, spanning from lively urban apartments to peaceful rural getaways. Equipped with detailed data on property attributes, pricing, guest reviews, and host profiles, researchers and enthusiasts can delve into trends and preferences within shared accommodations, enhancing their grasp of contemporary hospitality. Whether examining market shifts, evaluating the economic ramifications of tourism, or investigating worldwide travel tendencies, this dataset is an invaluable tool for exploration and analysis (PARAMVIR,2024).

Firstly, Arihant says that data privacy measures must be implemented to safeguard sensitive information, including obtaining appropriate permissions for data access, anonymizing personal data, and securely storing data to prevent unauthorized access. Additionally, it is crucial to reflect on the potential societal impacts of the project, including addressing biases in data collection or analysis and mitigating any adverse consequences, such as discrimination or unfair treatment, because if there is no regulation of those important aspects, people can have very bad consequences with online entertainment. Following that Arihant proposes some keys to how to do it:

* Embracing an ethical framework – Establishing a robust ethical framework guiding the entire data science process is imperative. This framework should encompass clearly defined principles, values, and best practices emphasizing fairness, transparency, and accountability across the data lifecycle.
* Detecting and mitigating bias – Proactively identifying and mitigating biases within algorithms is essential. Techniques like fairness-aware machine learning and routine model audits to ensure equitable outcomes among diverse demographic groups can effectively handle challenges associated with algorithmic biases.
* Implementing ethical training initiatives – Equipping team members with a thorough understanding of the established ethical framework cultivates a culture of ethical decision-making throughout the data science process.
* Regularly updating ethical guidelines – Given the evolving nature of technology, it is crucial to periodically update ethical guidelines;
* Advocating for external audits – External audits, verifying adherence to ethical guidelines and fostering transparency in data science endeavors (PATNI, 2023).

## Conclusion

The report provides a summary of how the project came about, highlighting important goals that were achieved during the modeling stage. There are some difficulties, especially during the data cleaning phase, where most of the time is spent on project management. The modeling results will be interpreted, considering the important things learned by analyzing the data and showing any major trends or patterns found.

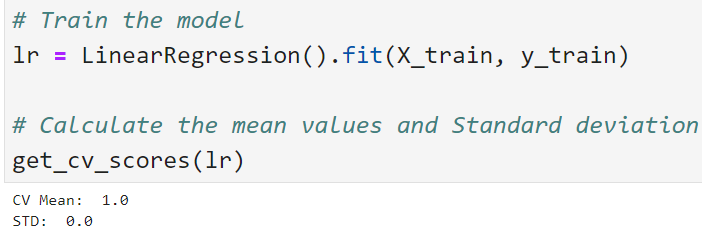


Figure 4 (Jupyter Notebook, 2024)

A result of "CV Mean: 1.0" and "STD: 0.0" in a linear regression model indicates that the model is providing perfect accuracy on the cross-validation data. Here's what each part means:

The chosen machine learning model, Linear Regression, proves how it can provide results with good accuracy. Because "CV Mean: 1.0" means that the average of performance metrics (like R², for example) calculated across multiple cross-validation folds is 1.0. A score of 1.0 indicates that the model can predict the target values ​​in the validation data perfectly. Zero standard deviation (STD) suggests that there is no variation in the performance score between the different cross-validation folds. This implies that the model is consistently achieving the maximum score across all cross-validation folds.

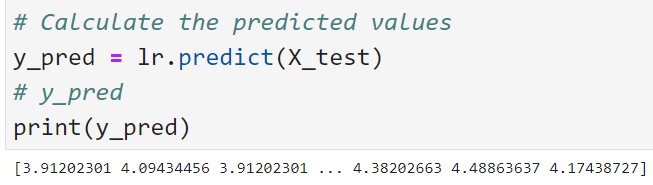


Figure 5 (Jupyter Notebook, 2024)

**Evaluation**

To understand the results obtained, you need to understand the following terms:

The R² metric, also known as the coefficient of determination assesses how well the predictions align with the actual values, indicating the degree of "goodness of fit." It ranges from zero to one, where zero signifies no fit and one denotes a perfect fit.

The Mean Squared Error quantifies the average squared difference between predicted and observed values (residuals). Similar to mean absolute error (MAE), it offers an overall estimation of the error's magnitude. Taking the square root of MSE restores the units to the original scale of the output variable, enhancing interpretability.

Mean Absolute Error computes the average of the absolute differences between predictions and actual values. It's a linear score, treating all differences equally in the average calculation. MAE provides insight into the extent of prediction errors without considering their direction (e.g., over- or underpredictions).

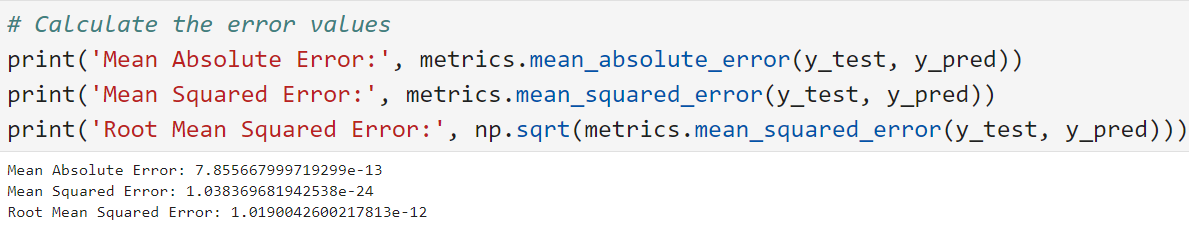


Figure 6 (Jupyter Notebook, 2024)

# References

Arihant Patni, “Ethical Considerations And Challenges In Data Science”. Linkedin. <https://www.linkedin.com/pulse/ethical-considerations-challenges-data-science-arihant-patni-sbnlc/>

[*"A Beginner's Guide to Kaggle for Data Science"*](https://www.makeuseof.com/beginners-guide-to-kaggle/). MUO. 2023-04-17*. Retrieved 2023-06-10*.

Kerzner, H. (2017). PROJECT MANAGEMENT A Systems Approach to Planning, Scheduling, and Controlling. New Jersey, Tenth Edition, page 29.

Karen L. Webber, Henry Y. Zheng (2020). Big Data on Campus: Data Analytics and Decision Making in Higher Education, pages 6-7.

Nick Hotz. “The Project Manager in Data Science”. Data Science Process Alliance. <https://www.datascience-pm.com/data-science-project-manager/>

Provost, F. (2013) and Fawcett, T.(2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. United States of America, O’Reilly Media, page 41.

PARAMVIR, “AirBNB\_Data”. Kaggle Website. <https://www.kaggle.com/datasets/paramvir705/airbnb-data>

Unidentified Author. “What is Data Analytics?”. UCD Professional Academy. <https://www.ucd.ie/professionalacademy/resources/what-is-data-analytics/>

Unidentified Author, “Model Evaluation”. Scikit Learn. <https://scikit-learn.org/stable/modules/model_evaluation.html#mean-squared-error>

Zhou, Zhi-Hua. (2016). Machine Learning, pages 2-3.