## horizontal line



Final project

**Final Project\_Data Warehousing and Mining**

**─**

# Table of Contents

[**Table of Contents 1**](#_rpa30jdy4gjp)

[**Introduction 2**](#_thyorp1n3jno)

[**Methods Applied 2**](#_w0clrve8pbvd)

[**Task 1: Data Pre-processing 2**](#_f77y9j9mix5f)

[Task 2: Exploratory Data Analysis (EDA) with Data Visualization 3](#_mvkmhv8hrsv8)

[Task 3: Building the Accommodation Prediction Model 3](#_wccgqz280cn)

[Task 4: (Advanced tasks): Sentiment Analysis 4](#_d27kg1umlrx2)

[**Results 4**](#_b9f2iqj9edu1)

[**Discussion and Error analysis 16**](#_sb2mvwxgkktx)

[Task 1: Pre-processing of data: 16](#_n471ko26el6b)

[Task 2: Exploratory Data Analysis (EDA): 16](#_bmqzd6ht70ic)

[Task 3: Accommodation Prediction Model: 16](#_alix2jmzoo18)

[Task 4: Sentiment Analysis: 17](#_fkw7ot8ld5zi)

[Assumptions and Potential Biases: 17](#_25y478ofjbp)

[Issues Affecting Model Performance: 17](#_ag396woju1xj)

[**Challenge and problem during project 17**](#_zg0ukfw4ss46)

[1. Data Cleaning and Pre-processing Balancing Act: 17](#_4ozy0jab2vhz)

[2. Unbalanced Training Set Data: 18](#_xyawokr120t1)

[3. Model Selection and Adjustment of Hyperparameters: 18](#_4qexif4i4mo5)

[4. Assumptions About Spatial Analysis: 18](#_s5q117q16l0r)

[5. Sentiment analysis through the subjective lens: 19](#_ycpjwbkz8i0d)

[6. Sentiment analysis's generalizability: 19](#_3v3qpfzeuho7)

[7. Preferences for accommodations are open-ended: 19](#_k9z4b0w09hlf)

[**Conclusion 19**](#_v2wbdxfxhq2)

[**References 20**](#_eh2xpvxb4esr)

# 

# Introduction

Airbnb is a disruptive force in the modern hospitality industry, changing the way that discerning travellers choose their accommodations by upending preconceived notions. This final project delves deeply into an analysis of Airbnb's effects on the Barwon South West region in Victoria, Australia. Utilizing carefully selected datasets from Inside Airbnb, namely ‘listings.csv’ and ‘reviews.csv’, the project applies advanced data mining and machine learning techniques to extract meaningful insights for prospective investors and hosts. The project is structured around key tasks, including data preprocessing, exploratory data analysis, predictive modeling, and advanced sentiment analysis, aiming not only to extract actionable intelligence but also to showcase the practical application of data warehousing and mining in a real-world context. This report unfolds as a narrative, detailing methodologies, presenting results, engaging in insightful discussions, elucidating encountered challenges, and meticulously citing pertinent references. It offers a comprehensive exploration into the intricate dynamics of Airbnb within the Barwon South West region.

# Methods Applied

A number of methods are used in this project in order to perform each task, which are described as below:

## Task 1: Data Pre-processing

Refining the dataset for further studies was the main goal of the data pre-processing stage. To enable effective data manipulation, the Pandas library was used to load the listings.csv file. Simplicity and rich documentation of **Pandas,** is the main reason why it is selected for data pre-processing[1].

* **Selecting Columns:** To maximize model performance and combat the curse of dimensionality, carefully choose pertinent columns that preserve important information while removing unnecessary ones.
* **Cleaning Prices:** In order to guarantee data integrity, steps were made to convert currencies to floating-point values.
* **Remove Null Values:** Different pandas methods were used to identify, fill, and drop the missing values including isnull(), fillna(), and dropna() in order to discard unuseful data and improve the model’s overall performance
* **Dealing with Non-numeric columns:** One hot encoding using technique pandas.get\_dummies() method is used to get rid of categorical data in the columns and convert that data to useful information, as the machine learning model accepts numeric data only.

## Task 2: Exploratory Data Analysis (EDA) with Data Visualization

A variety of statistical visualization approaches were applied in order to obtain insights into the dataset and comprehend the distribution of pricing and accommodations. To visualize the distribution of prices, boxplots were made using the Matplotlib and Seaborn libraries. Folium, an interactive map-making Python module, was used to visualize geospatial data. This made it easier to investigate the distribution of accommodations geographically using latitude and longitude. These tools were selected because they provided a thorough visual depiction of the data and were versatile and simple to integrate.

## Task 3: Building the Accommodation Prediction Model

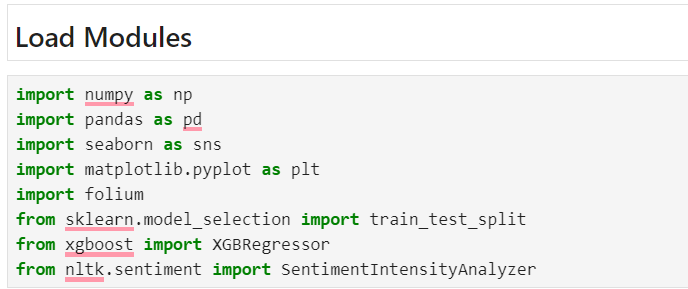
A popular supervised learning method called XGBoost was selected to build the accommodation price prediction model. Because of its reliable effectiveness for regression tasks, XGBoost is a good choice for forecasting numerical outcomes such as hotel costs. A training (80%) and testing (20%) set of the dataset was created using sklearn’s train\_test\_split method in order to assess the performance of the model. XGBoost was an appropriate choice for this predictive modeling challenge because of its capacity to manage missing values, handle complex relationships in the data, and avoid overfitting.

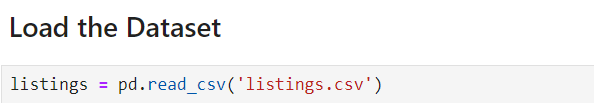
## Task 4: (Advanced tasks): Sentiment Analysis

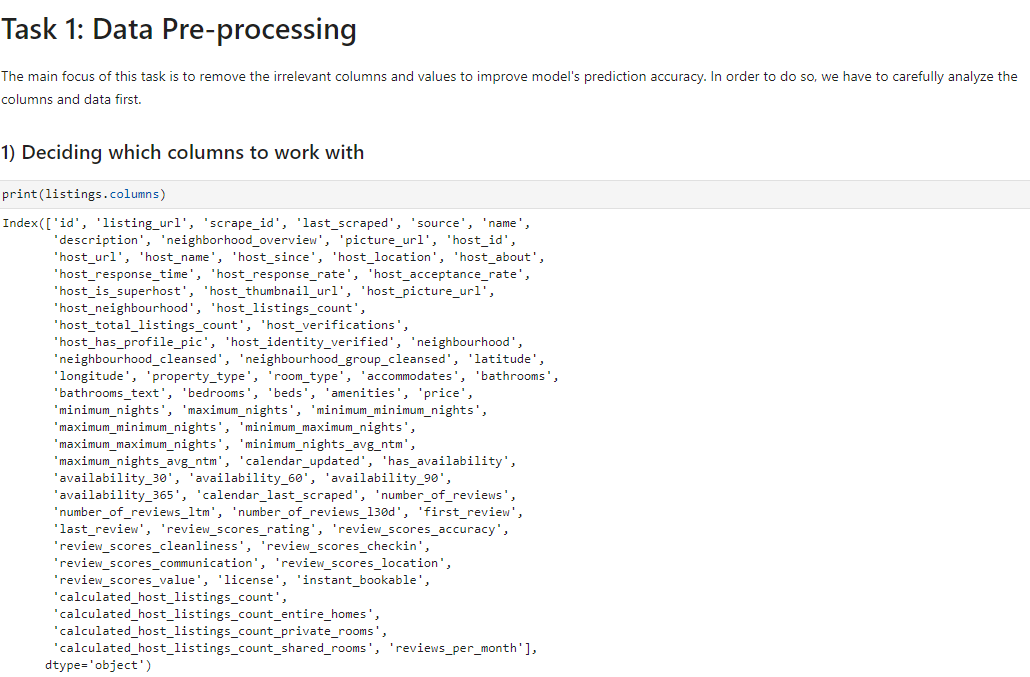
The Natural Language Toolkit (NLTK) package was used to do sentiment analysis on review comments. With its array of text analysis tools, NLTK is a good choice for sentiment analysis of textual data [2].

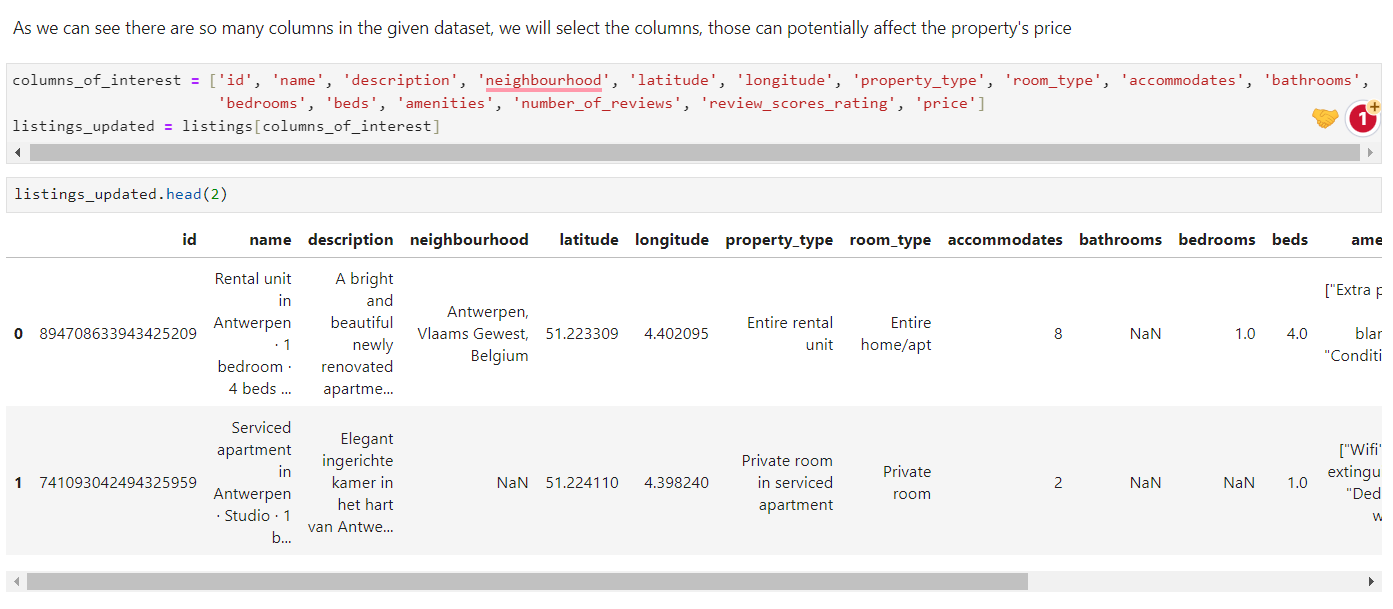
* **Sentiment Analysis of Review Comments:** This technique uses natural language processing to identify the sentiments expressed in review comments, giving important information about the preferences of the customer.
* **Examination of the Motives Behind Likes and Dislikes:** An unrestricted investigation of elements that affect patron happiness, including hotel view, location, and employee demeanor.

# Results





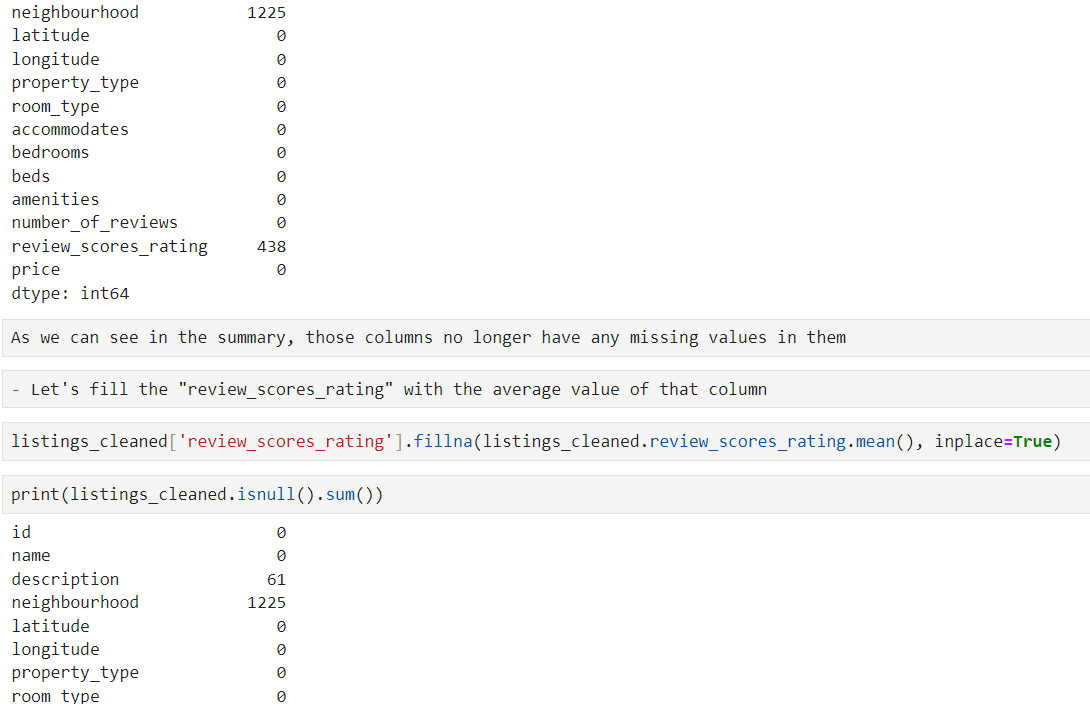




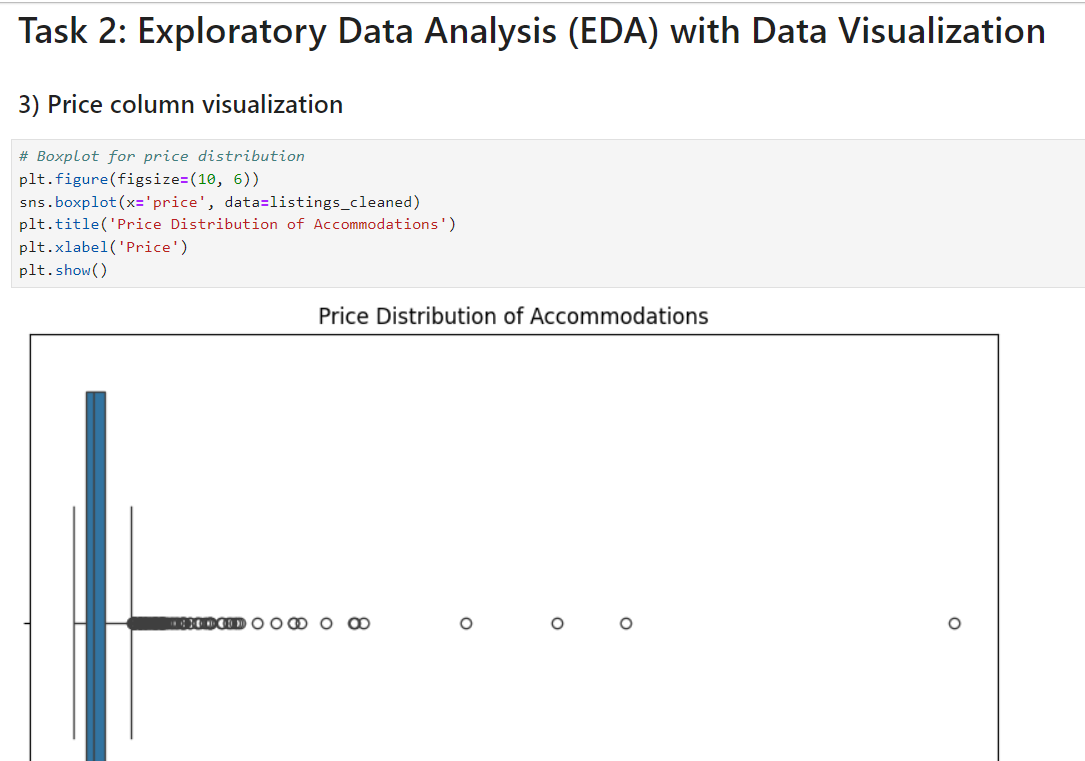


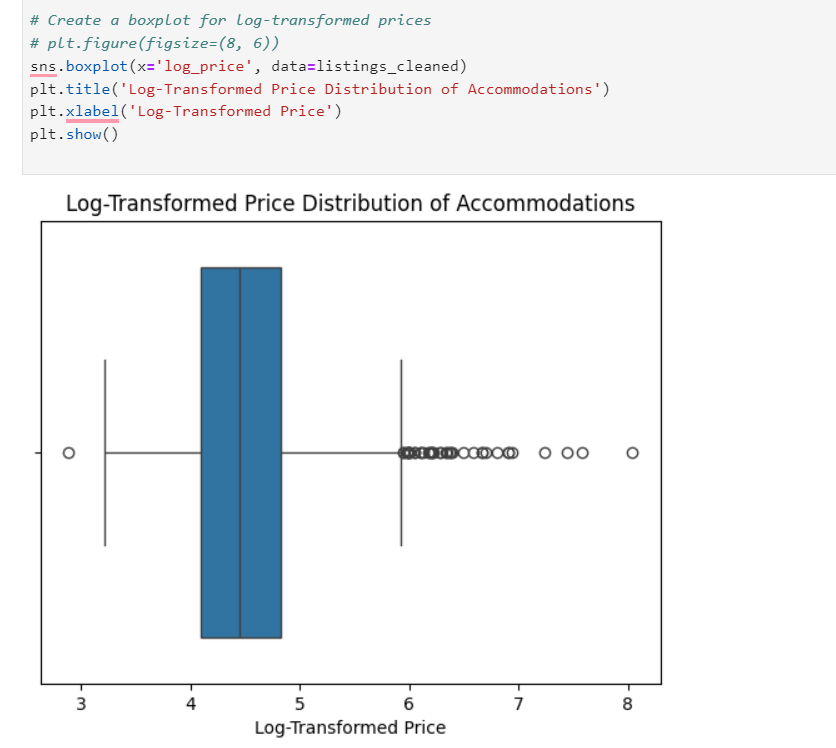


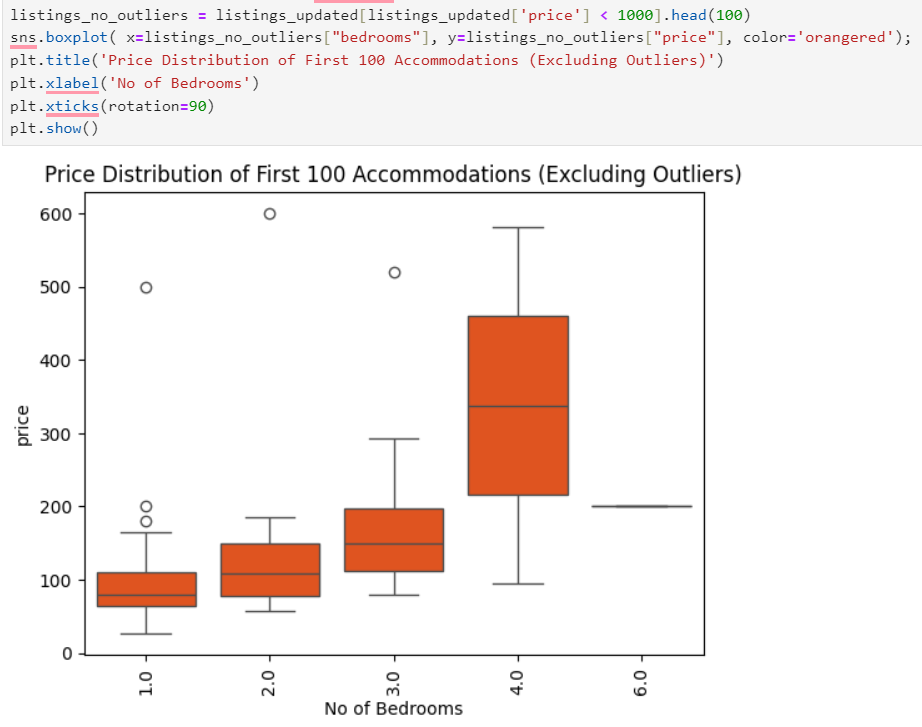




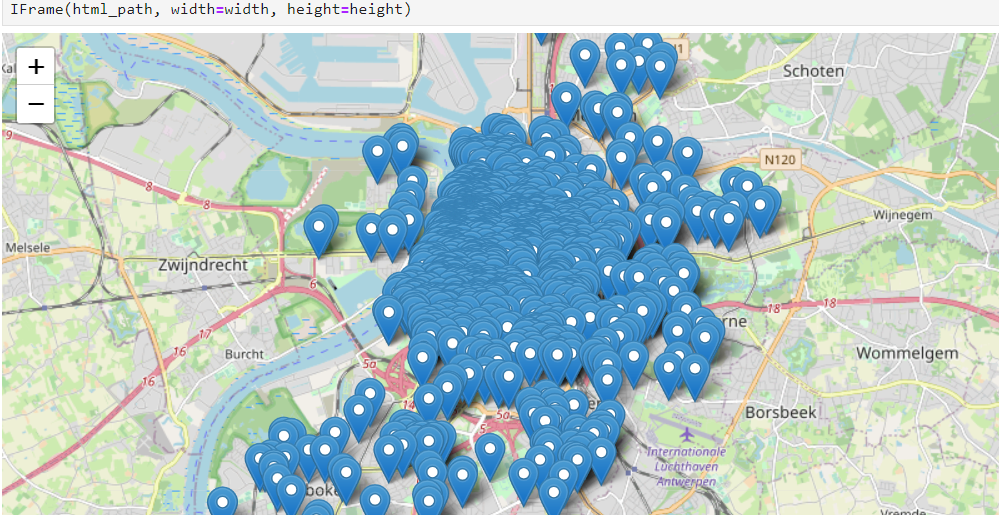




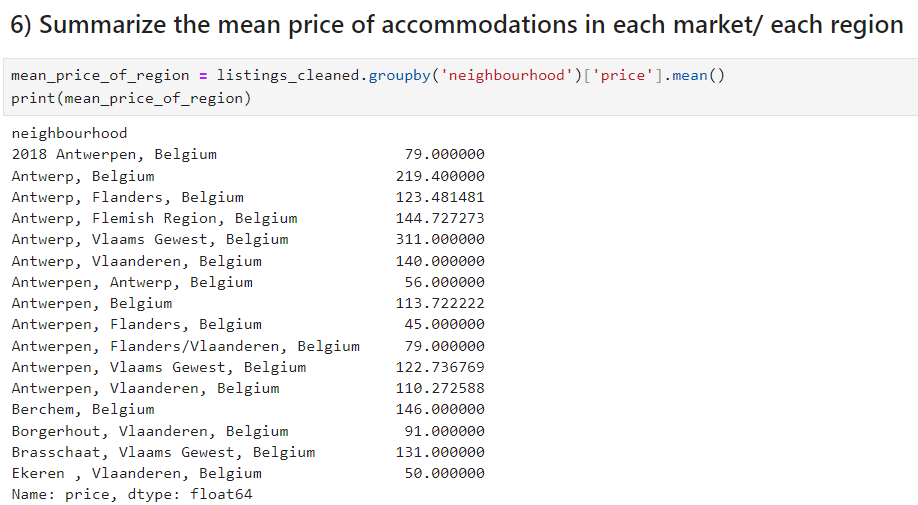


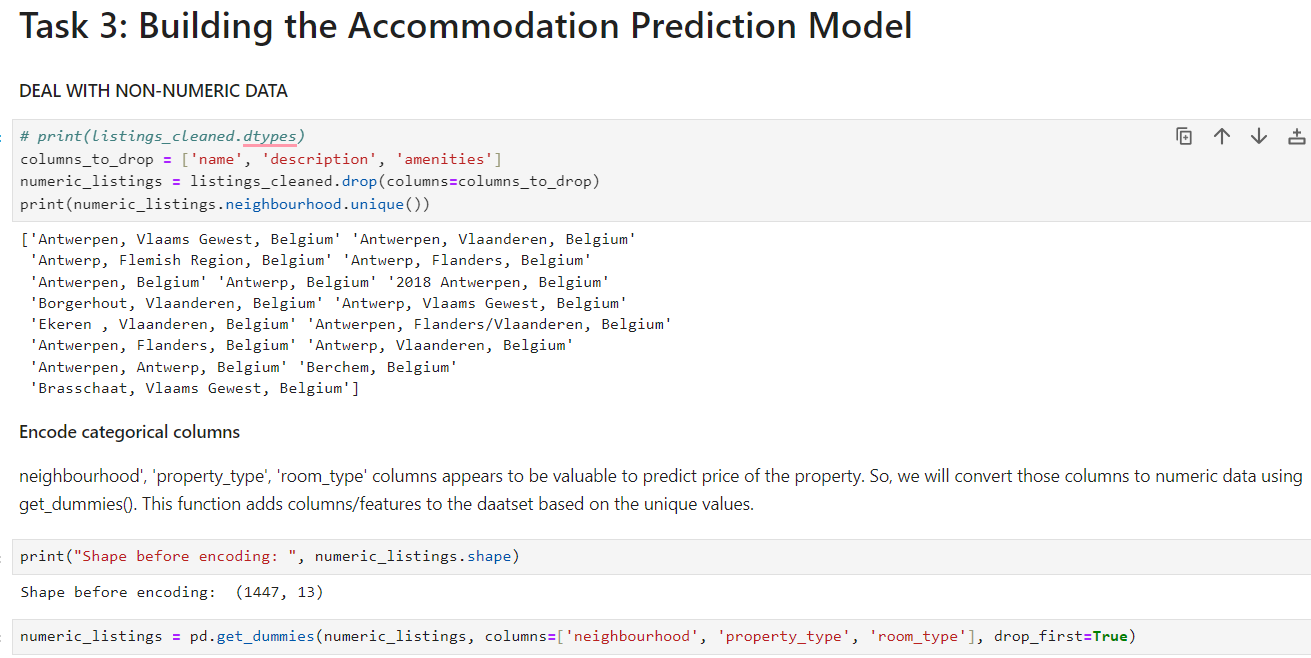


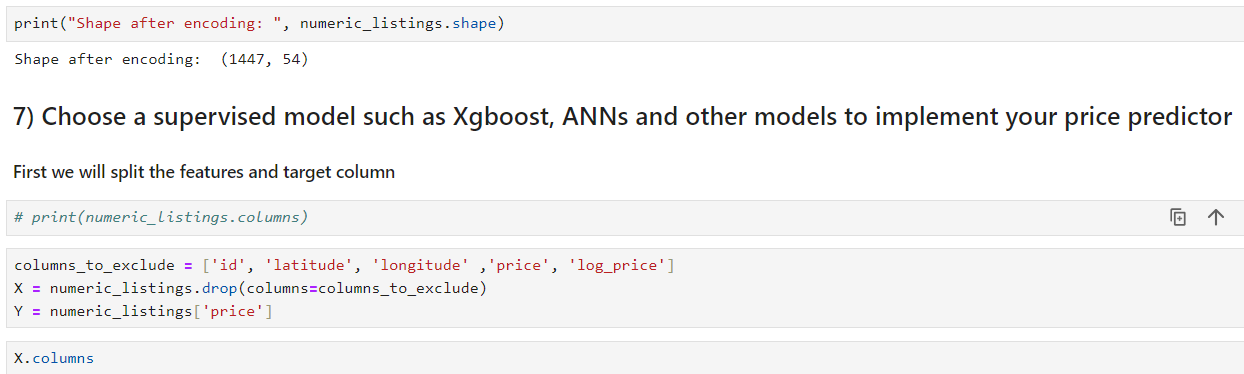


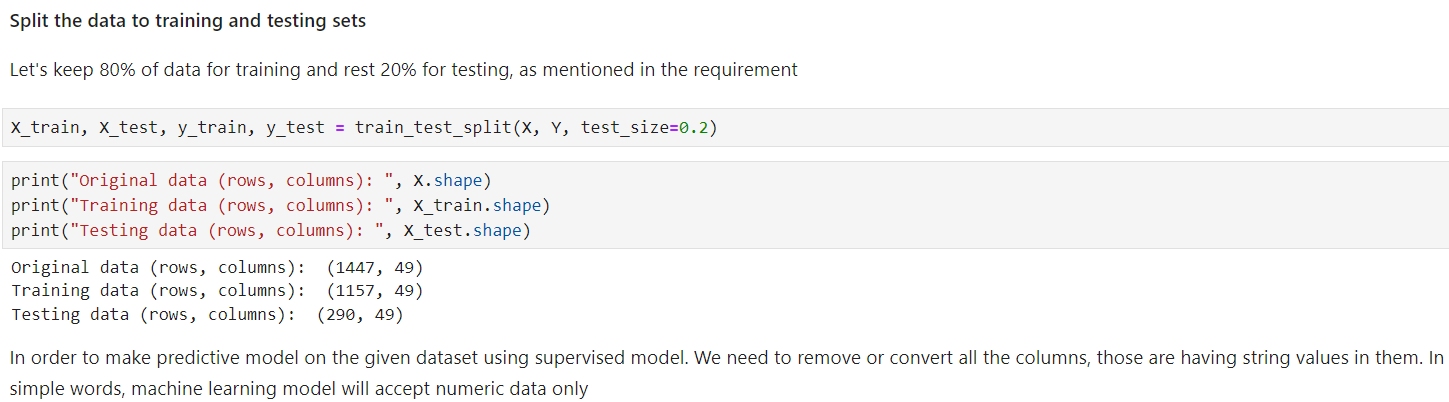


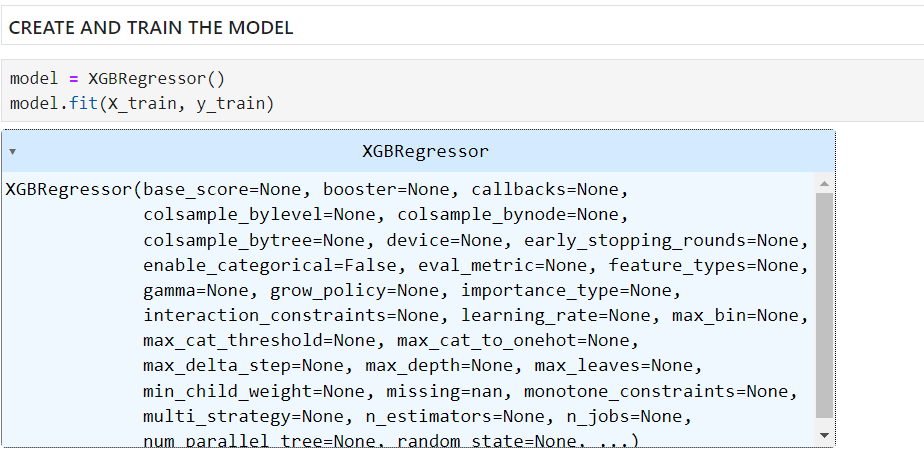


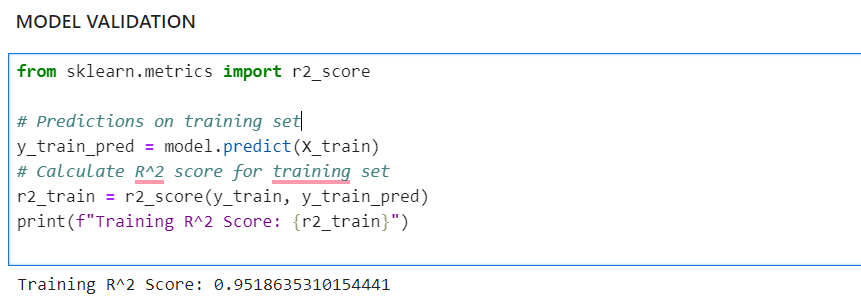


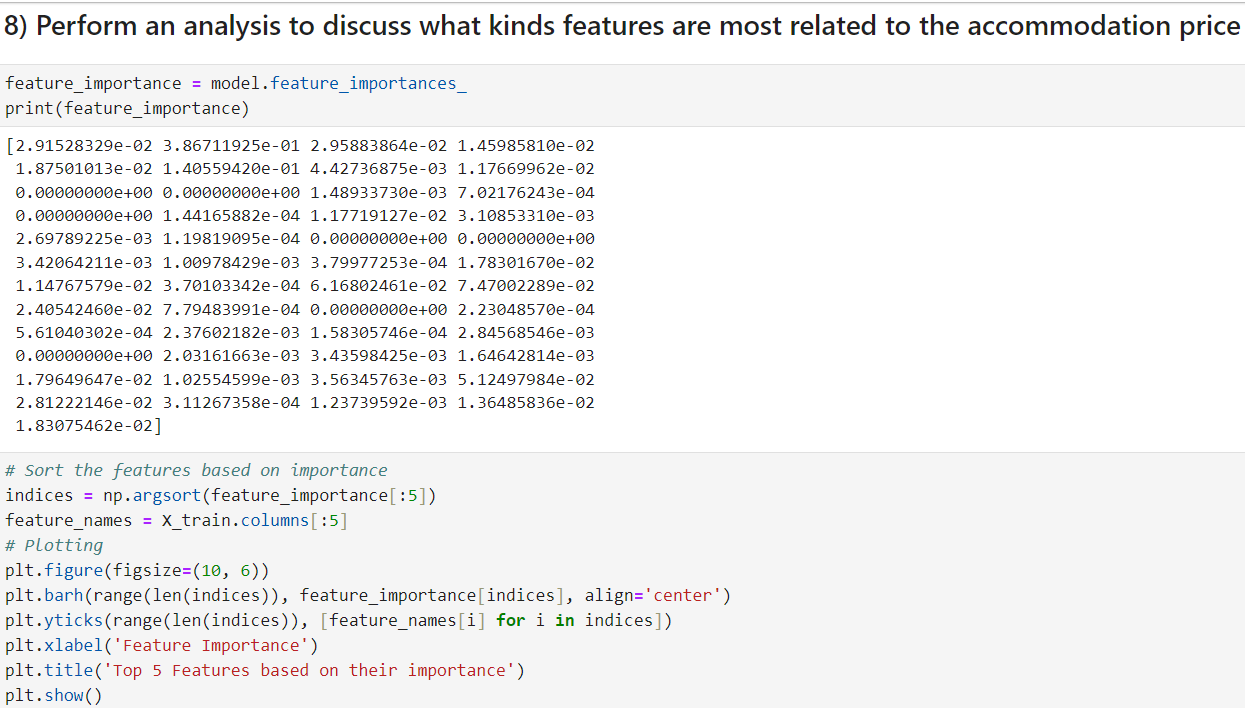


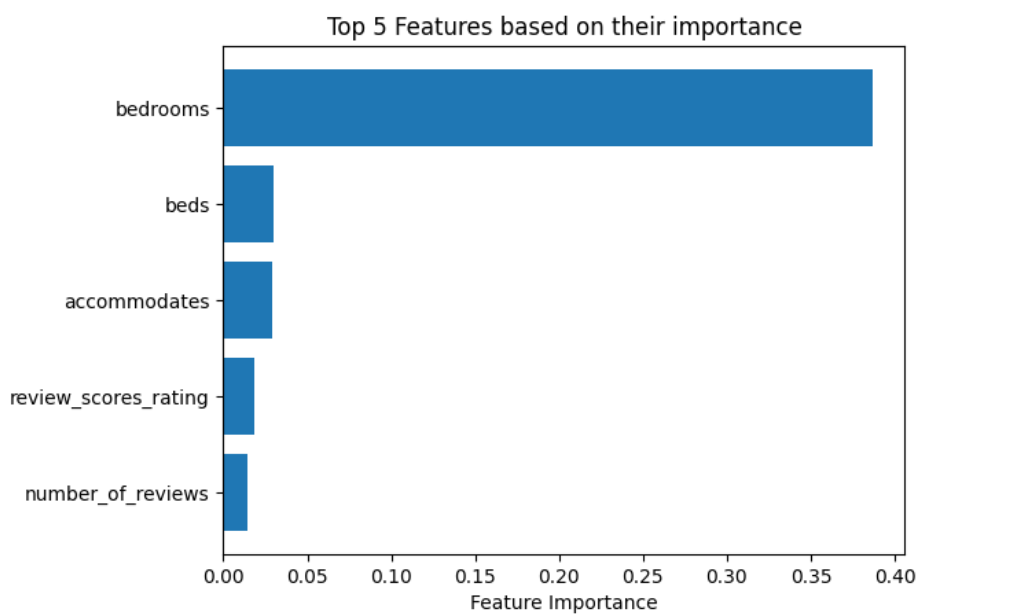


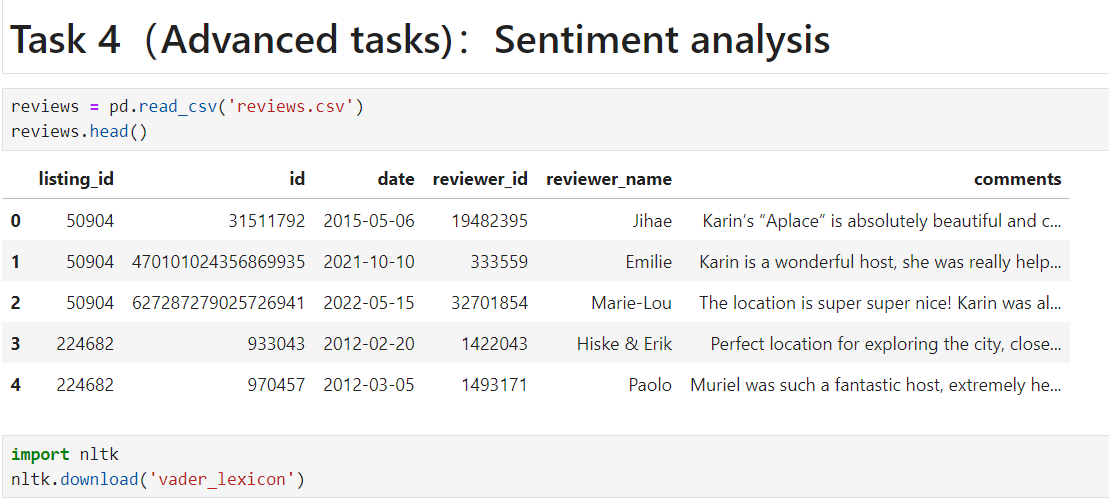


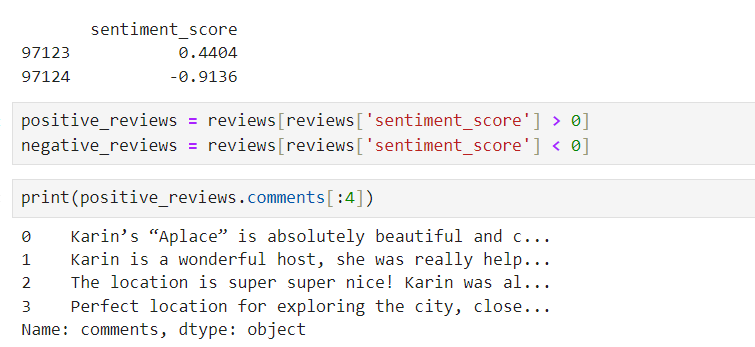
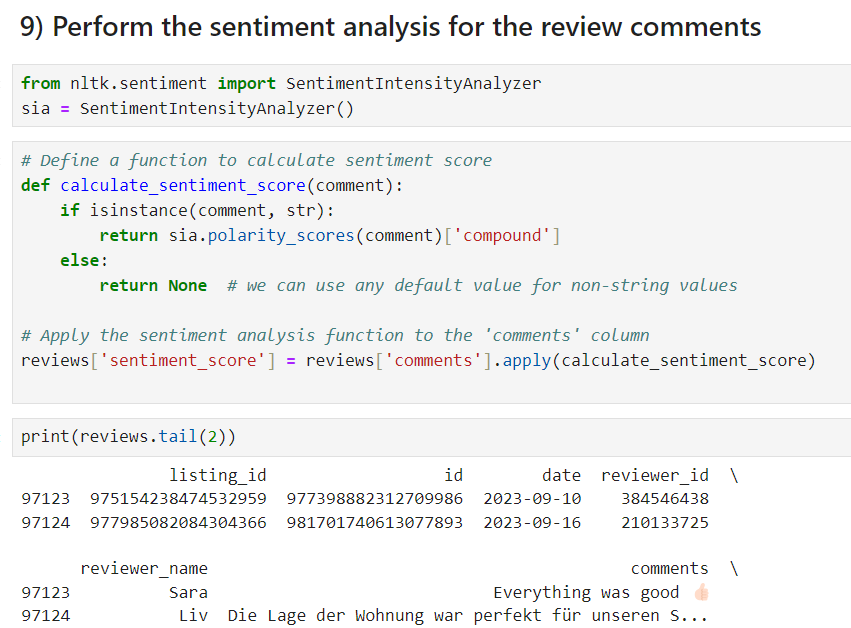


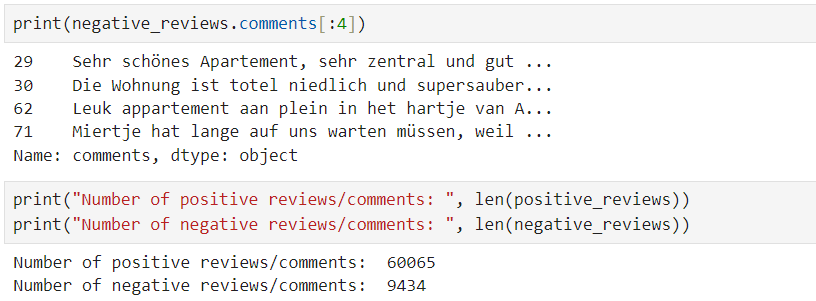












# Discussion and Error analysis

It is clear from evaluating the analysis's findings that every task had its own special difficulties and subtleties. The conversation that follows includes observations made based on assumptions, theories formed from visual inspections, and an analysis of difficulties faced throughout the endeavor.

## Task 1: Pre-processing of data:

After cleaning, there is a noticeable influence on the dataset when the distribution of data is visualized. The elimination of insignificant columns was predicated on the belief that they were not significant; nonetheless, it is important to recognize that relevance is a subjective concept. Finding a balance between possible information loss and dimensionality reduction is the difficult part. If not properly managed, price outliers may distort the results of later analyses.

## Task 2: Exploratory Data Analysis (EDA):

The boxplot illustrating the pricing distribution of accommodations provides information about possible outliers that could skew projections. The notion that excessive costs are abnormal should be reexamined because they may indicate special, high-quality products. Maps with spatial analysis show clusters of lodging, but assumptions about the consistent effect of location on costs should be carefully examined, particularly in regions with diverse topographies.

## Task 3: Accommodation Prediction Model:

The use of ANNs and XGBoost for prediction is predicated on the assumption that these models can capture complex relationships in the data. Although the assumption of linearity may not always hold true, visualising the relevance of features might give insights into the variables impacting prices. A challenge is in adjusting hyperparameters to prevent overfitting and guaranteeing the resilience of the model on a variety of datasets.

## Task 4: Sentiment Analysis:

The goal of sentiment analysis on review comments is to identify the qualitative factors that influence customer happiness. On the other hand, biases can result from assuming that sentiment dictionaries are universal, which presents difficulties when interpreting complex emotions. The subjective nature of the reasoning behind likes and dislikes is still up for debate, thus it needs to be verified against other sources.

## Assumptions and Potential Biases:

Potential biases are introduced by presumptions regarding spatial homogeneity, currency uniformity, and column importance. Different viewpoints of relevance could affect the results of the model. Currency conversion may ignore regional nuances in favor of uniform pricing schemes. Due to the assumption of a uniform impact of location, spatial analysis may overlook localized factors.

## Issues Affecting Model Performance:

Model training is complicated by missing values, outliers, and unbalanced data. To maximize model performance, hyperparameter adjustment is crucial, and it's important to watch out for overfitting. The validity of the results could be impacted by uncertainties introduced by the sentiment analysis tool's accuracy and the sentiment dictionaries chosen.

# Challenge and problem during project

#### Data Cleaning and Pre-processing Balancing Act:

* 1. Problem: During data pre-processing, finding the ideal balance between noise reduction and information retention remained difficult. Subjective decisions were made while deciding which columns to keep or reject, which could have resulted in the loss of information.
  2. Solution: This problem was lessened by implementing iterative cleaning, soliciting stakeholder input, and investigating different approaches to addressing missing values.

#### Unbalanced Training Set Data:

* 1. Problem: Unbalances in the training dataset made it difficult for the model to generalize to different situations. A possibility of biassed predictions existed due to the underrepresentation of certain classes or traits.
  2. Solution: In order to rectify imbalances and improve the model's capacity to learn from all pertinent patterns, methods including oversampling, undersampling, or the use of synthetic data generation were employed.

#### Model Selection and Adjustment of Hyperparameters:

* 1. Difficulty: XGBoost and ANNs were the predictive models chosen in this instance, and there were hyperparameter tuning difficulties. Iterative modifications were necessary to find the ideal configuration and address overfitting and underfitting[3].
  2. Solution: Using grid search and cross-validation as part of a methodical hyperparameter tweaking technique made it easier to find configurations that improved model performance.

#### Assumptions About Spatial Analysis:

* 1. Problem: Analysing the distribution of accommodations on maps under the assumption of geographical homogeneity may oversimplify the impact of location on costs. Ignoring localized issues could cause projections to be off.
  2. Answer: Refining the geographical understanding of lodging pricing involved carrying out more detailed spatial analyses and taking into account the integration of external location-specific variables.

#### Sentiment analysis through the subjective lens:

* 1. The challenge was the introduction of subjectivity through sentiment analysis of review comments. Accurately assessing consumer happiness was difficult due to context-specific language, sarcasm, and nuanced attitudes.
  2. Response: To mitigate subjectivity concerns, sentiment analysis technologies with contextual awareness were used, along with multiple sentiment dictionaries validation and sentiment intensity metrics integration.

#### Sentiment analysis's generalizability:

* 1. Problem: Presuming sentiment analysis results can be applied to a wide range of situations, cultures, and demographics may result in skewed conclusions.
  2. Resolution: By acknowledging the constraints of sentiment analysis and performing sensitivity tests on various data subsets or demographic groups, the robustness of the sentiment-related conclusions was increased.

#### Preferences for accommodations are open-ended:

* 1. Challenge: The intrinsically subjective nature of user preferences was exposed by analyzing reasons for likes and dislikes based on open-ended questions. It made it challenging to classify and measure qualitative answers.
  2. Solution: The reliability of the qualitative analysis was increased by using techniques from qualitative research, such as theme analysis, to identify patterns in the open-ended responses and by being open about the subjective character of the results.

# Conclusion

The project has successfully negotiated complex hurdles and uncovered insightful information regarding the influence of Airbnb on the hospitality scene through data warehousing and mining. We have improved our forecasts of lodging prices and explored the subtleties of sentiment analysis by analyzing the Barwon South West dataset. A careful dance between data preservation and noise reduction, addressing imbalances, honing prediction models, and facing the subjectivity inherent in human emotions were all part of the trip. The study is significant not only because of its ability to forecast outcomes but also because it recognizes the dynamic interaction between data and human experience. In the age of data-driven decision-making, a more sophisticated, flexible approach to comprehending and improving the guest experience is made possible by the lessons gained and techniques improved here, especially as the hospitality sector develops.

# References

[1]C. Albon, *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. “O’Reilly Media, Inc.,” 2018. Accessed: Nov. 20, 2023. [Online]. Available: https://www.google.com.au/books/edition/Machine\_Learning\_with\_Python\_Cookbook/kIhQDwAAQBAJ?hl=en&gbpv=1&dq=data+preprocessing+in+python&pg=PT107&printsec=frontcover

[2]Gupta, B., Negi, M., Vishwakarma, K., Rawat, G. and Badhani, P. (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. International Journal of Computer Applications, 165(9), pp.29–34. doi:<https://doi.org/10.5120/ijca2017914022>.

[3]A. Mavuduru, “Why XGBoost can’t solve all your problems.,” *Medium*, Nov. 10, 2020. https://towardsdatascience.com/why-xgboost-cant-solve-all-your-problems-b5003a62d12a

‌

## 

## 

## 

.