**Assignments: ALY6080 90325 Integrated Experiential Learn SEC 03 Summer 2023 CPS [BOS-1-HY]**

**Module 2 Assignment — XN Project: Pros & Cons**

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Zihan Ma

ma.zihan1@northeastern.edu

**Instructor**

 Prof. Valerie A. Atherley

**Date**

07/18/2023

**Locally Inspired** **XN Project: Pros & Cons**

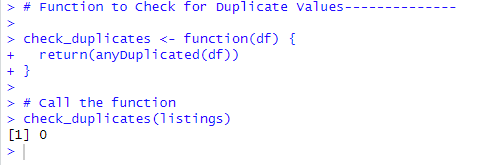
**Was there missing data? Duplications? How clean was the data?**

Yes, there was missing data in the dataset, specifically in the name, last\_review, and reviews\_per\_month columns. The missing values in last\_review and reviews\_per\_month could be attributed to some listings not having any reviews.

A computer screen shot of a missing values

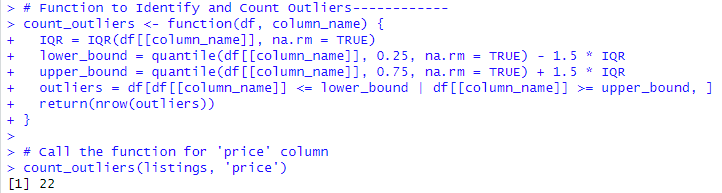
Description automatically generated

There were no duplications in the dataset, as verified by the function check\_duplicates. In general, the dataset can be considered relatively clean, apart from the missing values. The missing value in the name column might need some further investigation, as this is unusual.



**Were there outliers or suspicious data?**

Yes, there were outliers in the dataset, specifically in the 'price' column. These outliers were identified and removed using the Interquartile Range (IQR) method. After the removal of these outliers, there might be some changes in the dataset's distribution and characteristics, which could influence the subsequent analysis and predictive modeling results. So we will perform modeling twice on both datasets to see if it improves the model significantly.



A computer code with blue text

Description automatically generated

**What did you find? What intrigued you about the data? Why does that matter?**

In terms of intriguing findings, the substantial number of missing values in the last\_review and reviews\_per\_month columns suggests that many listings have not yet been reviewed. This could potentially impact the ability to accurately predict price since customer reviews and the frequency of reviews can be indicative of a listing's popularity or quality, and thereby its price. The encoding of the room\_type feature into a binary format and normalization of latitude and longitude will provide a simplified and more uniform dataset for modeling.

**What would your proposed next steps be? How do you plan to approach the cleansing of the data?**

The next steps in this process could involve handling the missing values, possibly by imputing them or perhaps excluding those rows or columns, depending on the specific use case and the acceptable level of data loss. For the name column, as it's generally used for identification rather than prediction, we could potentially ignore the missing value. For the last\_review and reviews\_per\_month columns, depending on the objective of the analysis, imputation methods or exclusion could be considered. It's also crucial to consider other aspects of data cleansing and transformation, such as ensuring data types are appropriate for each column and additional feature engineering to improve the model's accuracy.

**What feature engineering would you perform?**

More detailed feature engineering would depend on the specific analysis or prediction task, but some potential steps could include creating new features from existing ones (like extracting specific date components from a date column, if available), creating interaction features (multiplying or dividing two columns), or even binning numerical columns into categorical ones. For example, the host\_id might be used to derive a feature representing hosts with multiple listings. Additionally, for the columns with missing values, depending on the handling method, new features indicating "was reviewed" or "not reviewed" could be created. More advanced techniques, like clustering latitude and longitude to identify specific neighborhoods or regions, could also be beneficial, depending on the scope of the project. The specific transformations and feature engineering tasks we need to perform may depend on the model we plan to use. We will keep all features for now until we get started with the modeling.

