

**ICT505 Data Analytics**

Assessment 3-Group Assessment



April 29, 2024

Student NAme

Table of Contents

**1. Report on Data Quality Issues and Data Cleaning Techniques for Airbnb Dataset………….…..2**

**1.1. Introduction……………………………………………………………………………………………………………………….2**

**1.2. Task 1: Identify Common Data Quality Issues……………………………………………...….2  
1.2.1. Missing Values………………………………………………………………………………….…………….2**

**1.2.2. Inconsistent Data Formats…………………………………………………………………………….2**

**1.2.3. Outliers……………………………………………………….………………………………………………….2**

**1.2.4. Duplicate Records…………………………………….…………………………………………………….3**

**1.3. Task 2: Describe Data Cleaning Techniques……………………………………………………………………….3**

**1.3.1. Handling Missing Values……………………………………………………………………………………….3**

**1.3.2. Dealing with Inconsistent Data Formats……………………………………………………………...3**

**1.3.3. Converting Columns to Numeric…………………………………………………………………………3**

**1.4. Task 3: Data Pre-processing…………………………………………………………………………………………….4**

**1.4.1. Selecting Relevant Columns………………………...………………………………………………….4**

**1.4.2. Cleaning the Dataset……………………………………………………………………………………...4**

**1.4.3. Handling Inconsistent Data Formats…………………………………………………………………...4**

**1.4.4. Converting Data Formats…………………………………………………………………...…………….4**

**1.5. Conclusion……………………………………………………………………………………………………………………...5**

**1.6. References……………………………………………………………………………………………………………………...5**

1. **Report on Data Quality Issues and Data Cleaning Techniques for Airbnb Dataset**
   1. **Introduction**

Given the current era of data-centric decision-making, the quality of data utilized significantly influences the accuracy and reliability of the analysis results. Quality aspects identify common inaccuracy sources that occur from various data-related operations. Several sources of data quality challenges include missing data, use of inconsistent data formats, outliers, and duplication of record entries. An effective data cleaning approach ensures that the dataset is accurate, and the analysis insights obtained from it are valid. Therefore, this report presents the data quality issues apparent in the Airbnb dataset and describes the data cleaning and preprocessing methods used to clean the datasets.

Important information pertinent to the analysis of the holiday rental sector is collected by Airbnb and provided in the form of well-structured form data collection. Before beginning any significant study, it is crucial to address frequent limits in data quality and to do data cleaning and pre-preprocessing. The present research provided by the Airbnb needs to be pre-preprocessing and data cleaning methods applied to the Airbnb data set to improve its analytical readiness and integrity.

* 1. **Task 1: Identify Common Data Quality Issues**

The Airbnb data set provides important information relevant to the analysis of the vacation rental market. However, data cleaning and pre-preprocessing is important to first address common data quality limitations and conduct before embarking on any meaningful analysis. To enhance its integrity and analysis readiness, this report provides a description of the data cleaning and pre-preprocessing techniques that were utilized on the Airbnb data set.

* + 1. **Missing Values**

Upon the first look at the dataset, it became evident that there were missing values in several columns. Thus, the ‘last\_review’ column has missing values, which means that some of the listings were not reviewed for a certain period. The ‘license’ column contained the missing value, suggesting that some of the listings are not licensed. Missing values may impact the precision of the analysis due to biased results and a lack of information.

* + 1. **Inconsistent Data Formats**

Another data quality concern identified in the Airbnb dataset is inconsistent data formats. Some examples include cells from the column name containing special characters, ‘★’, punctuation, ‘·’, and more text maintaining the number of bedrooms, beds, and baths available. Uniform data formatting can prevent accurate standardization of data and make it difficult to properly analyze and visualize the data, which is very important in this domain. Consistency of data formats is essential for the data-set to be properly legible and analyzed.

* + 1. **Outliers**

Although this report did not explicitly detect the outliers, which is the major issue that we are facing in the collection of data in now a day, outliers are a major data quality problem that can lead to distorted statistical analyses and machine learning models if left unchecked. For example, outliers are present in the **‘price, minimum\_nights** and **number\_of\_reviews’** of certain columns. Assuring that the dataset is used to produce reliable outcomes is not possible if the outliers are not recognized and dealt with.

There must be the pre-processing is required to make the dataset very efficient, which is be necessary when we are visualizing the data, from this data set, so removal of the outliers is very necessary, or standardize the data from the mentioned dataset.

* + 1. **Duplicate Records**

The final potential problem affecting data quality is duplicate records. Although this report did not find any such records, this would be done when we checking the rank of the matrix, rank provides the duplication of the rows which helps in detecting the records that are most commonly occurring this will lead the dataset to the other dimensions deviate from the actual format, it might be assumed that duplicate **ADEs** are involved in the dataset, so this step should be taken while cleaning of the data from the given dataset. Eliminating duplicates in the data set allows each of the observations to be distinct and thus avoid any biases in the results of the analysis.

* 1. **Task 2: Describe Data Cleaning Techniques**

The following common data cleaning techniques were applied to resolve the data quality issues discovered in the Airbnb dataset, these are the techniques helpful in the detection of the missing values, outliers, inconsistent date formats, and be the no-numeric data from the dataset.

* + 1. **Handling Missing Values**

Firstly, missing values were addressed. Using Dropna function, helps dataset’s rows with any missing value were removed from the dataset. This method guarantees that the dataset used in the analysis is full and contains no missing values that could affect the results and cause bias. However, it is possible to attribute the missing values as well. Depending on the data and analysis’ primary purpose, the mean imputation or predictive modeling can be utilized for this process.

* + 1. **Dealing with Inconsistent Data Formats**

Next, since the data format in the name column was inconsistent, especially the additional characters and the number that denoted the number of **bedrooms, beds, and baths**, the **str.replace()** method was used to remove the inconsistent date format. The data format standardized ensures consistency in the entire dataset and promotes the efficiency of analysis and visualization. Furthermore, it is possible to apply a regular expression to select the targeted part of the text from the column and further enhance the uniformity of the data.

* + 1. **Converting Columns to Numeric**

For the columns that contained numerical data format, this conversion is very necessary, because we have to apply statistical functions, columns provided in our dataset such as **price, minimum\_nights, number\_of\_reviews, reviews\_per\_month, calculated\_host\_listings\_count, availability\_365, and number\_of\_reviews\_ltm**, the resultant numeric data type was utilized. The conversion was achieved by applying the **pd.to\_numeric()** function. The conversion supported numerical analysis and computations as most numerical data can be statistically manipulated for analysis.

* 1. **Task 3: Data Pre-processing**

Cleaning the dataset by dropping irrelevant information, handling missing data, and transforming data formats, as such, the data pre-processing step involved several activities to ready the dataset for analysis. Some of these steps comprised selecting the necessary columns.

* + 1. **Selecting Relevant Columns**

In this instance, columns considered unnecessary such as **‘host\_id’, ‘host\_name’, ‘neighbourhood\_group’, ‘last\_review’,** and **‘license’** were deleted from the dataset. This process reduces the dataset’s dimensionality, hence concentrating the analysis on critical variables.

Hence by choosing the appropriate attributes is very necessary to perform analysis on the data, such as **'host\_id', 'host\_name', 'neighbourhood\_group', 'last\_review',** and **'license'**, were dropped from the dataset for making the dataset be making worth it. Pay more focuses on analysis need most relevant features, by applying this step helps in reducing the dimensionality of the dataset.

* + 1. **Cleaning the Dataset**

After loading the dataset into Google Coolab, the null values were the major issue this is removed by dropping them using the **dropna()** function. Consequently, only the complete cases were utilized in the analysis to prevent biasness and invalid patterns. A variety of techniques, for instance imputation, can be used to handle the null values depending on the type of missing data and the purpose of the analysis.

* + 1. **Handling Inconsistent Data Formats**

This step ensures dependability in the 'name' column across the dataset and improves the interpretability of the data. The inconsistent data formats in the ‘**name’** column were resolved through eliminating the special characters and superfluous text by applying the **str.replace()** method. The method made the data format of the ‘name’ column consistent throughout the dataset and substantially enhanced the data readability.By removing special characters and unnecessary text using the **str.replace()** method, inconsistent data formats in the 'name' column were addressed and removed it through the above mentioned function.

* + 1. **Converting Data Formats**

These are the columns having the numeric data, are **‘price’, ‘minimum\_nights’, ‘number\_of\_reviews’, ‘reviews\_per\_month’, ‘calculated\_host\_listings\_count’, ‘availability\_365’,** and **‘number\_of\_reviews\_ltm’,** **pd.to\_numeric()** helps in order to transformed into the numeric data type by using the built-in function. This conversion streamlines the numerical analysis and modeling operations, allowing to generate more dependable and relevant patterns and insights from the dataset.

* 1. **Conclusion**

To conclude, applying efficient data cleaning techniques is critical to overcoming common data quality deficiencies and achieving the required quality of data for analysis, efficient data cleaning methods are critical to resolving typical problems with data quality and getting the dataset ready for analysis. Through detecting and resolving missing values, data format disparities, and additional data quality deficiencies, we guarantee the dataset is correct, consistent, and suitable for analysis. We guarantee that the dataset is **accurate**, **dependable**, and **appropriate** for analysis by locating and fixing **missing values, inconsistent data formats,** and **other problems** with **data quality.** The methods reviewed in this paper offer a way of structuring and pre-processing numerous kinds of datasets across many domains to blossom adequate data quality outcomes that can lead various meanings among data engineers to distinguish based on high-quality data. In this paper offer a framework for preprocessing and cleaning datasets across a range of **industries**, **empowering** **analysts**, and **data scientists** to extract valuable insights and make defensible judgments using high-quality data.

* 1. **References**
* McKinney, Wes. (2017). "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython." O'Reilly Media, Inc.
* VanderPlas, Jake. (2016). "Python Data Science Handbook: Essential Tools for Working with Data." O'Reilly Media, Inc.
* Wickham, Hadley, Grolemund, Garrett. (2016). "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data." O'Reilly Media, Inc.
* "Pandas Documentation." Accessed from: **https://pandas.pydata.org/docs/**
* "NumPy Documentation." Accessed from: **https://numpy.org/doc/**
* "Scikit-learn Documentation." Accessed from: **https://scikit-learn.org/stable/documentation.html**
* "Regular Expression Operations." Python Documentation. Accessed from: **https://docs.python.org/3/library/re.html**
* Kandel, Sean, et al. (2012). "Wrangler: Interactive Visual Specification of Data Transformation Scripts." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
* Patel, Maulik. (2020). "Data Cleaning with Python: Real-World Examples." Packt Publishing.
* "Airbnb Dataset." Accessed from: **http://data.insideairbnb.com/australia/vic/barwon-south-west-vic/2023-12-26/visualisations/listings.csv**