Applied Quantitative Analysis (AQA), Vol. 3 No. 2 (2023) https://doi.org/10.31098/quant.1951

Research Paper

Improving University Students' Data Analysis Outputs through Effective Data Collection, Cleaning, Screening, and Normalisation

Mansur Adam Saidu[1*], Shamsudeen Ladan Shagari[2], Muhammad Auwal Kabir[2], Abdulkadir Abubakar[2]
[1]Abubakar Tafawa Balewa University Bauchi, Nigeria
[2]Bauchi State University, Nigeria

Received: Oct 1, 2023	Revised: Oct 30, 2023	Accepted: Dec 9, 2023	Online: Dec 23, 2023

Abstract

Practical data analysis reflects an improved approach to data collection, cleaning, and screening. However, very few studies reported the techniques used to clean and screen their collected data, leading to questionable final results and interpretations, especially among university students. To address this issue, the current study examines the rigorous data collection, cleaning, and screening processes for data normalization among university students in Nigeria. Using a multi-stage research methodology, 372 adapted survey instrument items were administered via snowball sampling. Finally, 365 were retrieved from the respondents. Missing data were all imputed using the Series Mean (SMEAN), and outliers were appropriately addressed using z-scores and chi-square criteria. Descriptive statistical measures were used to examine the dataset and presented in several tables, a histogram, a scatterplot, and a standard probability plot. The collected, cleaned, and screened data were found to have a normal distribution, facilitating analysis and understanding of the parametric distribution, variation, and normalization. The findings provide valuable guidance for university students, academics, policymakers, and practitioners in data collection, cleaning, and screening. It was recommended that university students, lecturers, researchers, and research institutions prioritize thorough data collection, embrace transparent data cleaning, screening, and reporting practices, and adopt standardized procedures to enhance data accuracy, reliability, and normalization, thereby enabling better data analysis and the interpretation of research findings.

Keywords: *Data Collection, Data Cleaning, Data Screening, Data Normalisation and Data Analysis*

INTRODUCTION

Data analysis is essential in research studies that collect individuals' primary responses. Data analysis involves the critical structuring, summarization, and representation of data to enable better comprehension, explanation, and display of information for decision-making. Researchers use it to determine whether the data obtained are reliable with respect to a proposed causal process (Hayes, 2022). Data analysis involves using various methods to examine data and identify characteristics such as data type, size, format, and patterns, and to perform steps including identifying data issues, conducting statistical inspections, creating models, and testing hypotheses based on the analysis results (Rahul et al., 2020). This helps determine the data quality for use in different settings and applications. Researchers typically conduct data analysis after collecting raw data in the field or extracting it from a document. Raw data are data that have not yet been used and require cleaning, analysis, filtering, and processing. Proper data quality is needed before the final data analysis to enable better data transformation, interpretation, and applications. This can be achieved by appropriately selecting and applying the desirable data cleaning and screening techniques.

Essential data-cleaning methods are the starting point for handling research datasets to improve data analysis. Data cleaning is crucial in research analysis, as mistakes in the data can yield

Corresponding author's email: smadam@atbu.edu.ng

results that are incorrect, unfair, or even risky (Neutatz et al., 2021). Interestingly, data cleaning fixes issues in the data, such as missing, incorrect, or inconsistent research information, before the final analysis and interpretation. Data cleaning before analyzing a survey is vital to ensuring the final data output is reliable. In particular, data cleaning and data analysis are closely related. A cleaned dataset improves data quality and makes analysis outputs more reliable and trustworthy (Alotaibi et al., 2023). Cleaning data before analysis is crucial, as it involves fixing or removing incorrect, corrupted, misformatted, duplicate, or incomplete information in a dataset. This step is necessary for researchers to make better decisions based on the dataset's accuracy. However, data cleaning aims to prepare data for functional statistical analyses on a given data source. This includes tasks such as estimating populations, analyzing specific areas, constructing composite estimates, and developing models, such as regression (Steorts, 2023).

Data cleaning addresses issues in a given dataset, such as missing values, duplicate data, anomalous or out-of-range values, or irrelevant data. However, researchers often must report how they clean and screen their data (Arevalo et al., 2022). For instance, only 8% of recent studies provided complete information on their scope, frequency, and how they organized and cleaned their data, particularly regarding the integrity of their datasets (Hsieh et al., 2023). Moreover, according to Alotaibi et al. (2023), approximately 73% of research on addressing data issues focuses solely on outlier detection, which is important for data understanding, whereas nearly 13% of recent studies address multiple issues, such as missing data, outliers, and duplicates. Alotaibi et al. (2023) reported that around 8% of studies address missing values, and about 5% focus on eliminating duplicate data. Other issues that cleaning researchers face include low-quality data and challenges in selecting appropriate structural parameters for the identification model (Qin et al., 2022). However, proper data cleaning can prevent errors in analysis, interpretation, and conclusions.

Data cleaning can be complex due to data type (structured, semi-structured, or unstructured), particularly for semi-structured and unstructured data. When cleaning data, problems such as missing values, duplicates, outliers, and irrelevant information often arise (Alotaibi et al., 2023). To address data cleaning issues in research, specialized tools such as artificial intelligence, machine learning, deep learning, statistical methods, and combined and unclassified techniques can help handle abnormal datasets. These techniques simplify and improve precision, thereby achieving high accuracy and ensuring that the data management system operates effectively within a given dataset (Alotaibi et al., 2023; Lai et al., 2023; Li & Lv, 2023; Rezvani et al., 2022). Even with many methods for cleaning data, researchers sometimes need to remember to consider critical data-cleaning techniques. This oversight can affect the accuracy and reliability of the final data analysis and the resulting output. As evidence, several recent studies have examined data cleaning prior to data analysis. However, only a few studies have examined the identification of irrelevant data, as it is not easy to detect, except in cases where the data can be detected (Alotaibi et al., 2023; Steorts, 2023).

Furthermore, prior researchers focused more on specific fields of study, leaving many other vital specializations with little or no attention. For instance, most of the previous studies that used and reported data-cleaning information were mainly from Computer Science and Engineering (Lai et al., 2023; Li & Lv, 2023; Liao et al., 2023; Steorts, 2023), Medicine (Arevalo et al., 2022; Hsieh et al., 2023; Love et al., 2021; Rezvani et al., 2022), as well as Artificial Intelligence and Machine learning (Neutatz et al., 2021; Qin et al., 2022). Minimal previous efforts were from business or social sciences (Babagana et al., 2019; Jian Ai et al., 2021; Monsurat et al., 2023), especially the aspect of e-commerce users' behaviour, technology optimism, and tax compliance intention, which contribute to nations' tax revenue generation and economic growth (Saidu & Ladan, 2023; Saidu et al., 2022) as was noted among university students (Chea & Chea, 2022; Parilla & Abadilla, 2021;

Pauch, 2023; Yeo et al., 2019). Therefore, studies on detecting unusual patterns, removing duplicate data, and finding missing values in large and diverse data streams are challenging and still need to be explored in some key fields of study, which need further research, as Alotaibi et al. (2023) suggested. On this premise, this study examined the rigorous data collection, cleaning, and screening processes for data normalization among university students in Nigeria.

LITERATURE REVIEW

The Concept of Data Cleaning

Data cleaning is a crucial step in today's data-driven world. It ensures the reliability of collected data for the final analysis. Data cleaning involves identifying and correcting unusual values, enhancing data features for efficiency and accuracy, and handling missing values and outliers (Liao et al., 2023). The process can be domain-specific, requiring field knowledge, or domain-independent, catering to general database users (Li et al., 2019). Researchers, managers, and statisticians typically perform data cleaning. It helps verify logical sequences and expected ranges and ensures that no critical data is missing (Love et al., 2021). However, researchers must refine their raw data, as any issues could affect their final analysis (Rezvani et al., 2022). Thus, data cleaning prepares raw data for use, organizes it, and assesses its usability, helping researchers identify and eliminate unusual values.

Data Cleaning Process

Data screening and cleaning are critical processes that ensure the validity of the final analysis results in a given research. They involve checking for data entry errors and abnormalities, using descriptive statistics such as minimum and maximum values, out-of-range values, valid and missing cases, mean, and standard deviation (Pallant, 2020). A thorough review of statistical and graphical data is needed to reduce analytical bias and promote data normality (Watkins, 2021). Similarly, researchers are advised to check for missing values, outliers, influential cases, multivariate statistical assumptions for alternative tests, and data normality (Denis, 2020; Hair et al., 2019; Tabachnick & Fidell, 2019; Watkins, 2021). Furthermore, before starting data cleaning, researchers must combine their data files, assess data quality, and remove or exclude participants who are incorrect where necessary (Silvia & Cotter, 2021). The data cleaning process involves pre-processing, anomaly detection and repair, and validation (Li et al., 2019). In research reporting, the data cleaning and screening process involves examination, exploratory analyses, and editing to identify and rectify errors in the dataset (Huang, 2019). Thus, these approaches are highly effective, as they provide a brief, concise procedure and techniques for data cleaning, screening, and visualization.

Measuring Data Screening and Cleaning

Data screening is the first stage of data cleaning. It involves removing irrelevant items and variables that affect the final result. It includes inspecting statistics and graphics data to reduce bias and ensure data normality (Sarstedt & Mooi, 2019; Watkins, 2021). Data cleaning focuses on correcting errors and inconsistencies and handling missing data. It begins by examining univariate outliers at a threshold of ≤ ±3.29 z-scores (Tabachnick & Fidell, 2019) and can also be applied to z-scores ≤ ±3.0 (Hair et al., 2019; Pallant, 2016). The next step is to examine the Mahalanobis Distance (D2) for multivariate outlier detection, which assesses the adequacy of the assumptions and prepares for modelling (Pahlevan & Shafir, 2019; Watkins, 2021). Researchers should examine the value of D2 by comparing the number of independent variables with the degrees of freedom (df) within the Chi-square critical values (Pallant, 2016; Pallant, 2020; Watkins, 2021). Moreover, outlier cases are typically addressed pre- and post-analysis by removing observations from the

analysis (Hair et al., 2019). The collected data are screened for normality and appropriateness prior to final reporting, using IBM SPSS (Arbuckle, 2017; Pahlevan & Shafir, 2019; Salcedo & McCormick, 2020; Watkins, 2021). All items outside the required thresholds are deleted using the listwise procedure (Hair, Marcelo et al., 2019; Pallant, 2020; Tabachnick & Fidell, 2019).

Parametric and Non-Parametric Data Analysis

Parametric and non-parametric tests are essential in research for obtaining valid analysis results. Parametric tests are used when data is assumed to follow a normal distribution. They include the t-test, paired t-test, one-way ANOVA, two-way ANOVA, Analysis of Covariance (ANCOVA), MONOVA, one-sample t-test, two-sample t-test, Pearson's product-moment correlation, and multiple regression (Alvo & Yu, 2018; Arboretti et al., 2018; Bogo et al., 2023; Dickhaus, 2018; Hu et al., 2024; Pallant, 2020; Tabachnick & Fidell, 2019). However, these tests have certain assumptions that must be met to ensure validity (Sarstedt & Mooi, 2019; Van Buren & Herring, 2020). The assumptions include that the population is normally distributed, that sample variances are similar, that suitable variables are used, that samples are independently selected, and that no outliers are present.

Besides, if the actual values in a dataset deviate from normality, transforming the data on a natural log scale can render parametric tests appropriate (Alvo & Yu, 2018; Arboretti et al., 2018; Dickhaus, 2018). Equally, non-parametric tests are used when parametric tests are not suitable, focusing on ranks and testing for distributional anomalies (Alvo & Yu, 2018; Arboretti et al., 2018; Bogo et al., 2023; Dickhaus, 2018; Van Buren & Herring, 2020). They are used when dealing with non-normal or unknown distributions, small sample sizes (<30), and extreme outliers or discrete variables. The standard non-parametric tests include chi-squared, Fisher's exact tests, Wilcoxon's matched pairs, Mann–Whitney U-tests, Kruskal–Wallis tests, and Spearman's rank correlation (Beukelman & Brunner, 2016; Bogo et al., 2023; Dickhaus, 2018; Savani & Barrett, 2009; Sheskin, 2011; Van Buren & Herring, 2020).

Homogeneity Test of Variance

The homogeneity test of variance is a vital part of data analysis, checking whether samples from different populations in a study have the same variance, a property also known as homoscedasticity. The assumption of homogeneity of variance ensures that all comparison groups (two or more) in a parametric test have equal variances (Pallant, 2020). This assumption is used in statistical tests assessing normality, such as the z-test, t-test, Chi-Square test, ANOVA, ANCOVA, and MANOVA (Pallant, 2020; Kumar & Misra, 2020; Tabachnick & Fidell, 2019). However, in the t-test analysis, alternative Levene's test options are provided for selection, depending on whether the outputs are assumed to be equal or unequal. The homogeneity test is based on a null hypothesis that the samples have equal variances, with an alpha value of 0.05 or higher. Normal continuous data will fail to reject the null hypothesis at an alpha value above 0.05. If the alternate hypothesis is significant with an alpha value below 0.05, then it is a sign that there is an absence of homogeneity or normality within the dependent variable (Pallant, 2020; Tabachnick & Fidell, 2019).

Additionally, various tests, such as Hartley's Fmax, Box's, Cochran's, Levene's, and Bartlett's, are used to evaluate data homogeneity. If the homogeneity of variances assumption is violated, the final results of statistical tests become unreliable, necessitating alternative nonparametric measures or data transformations using the natural logarithm (Tabachnick & Fidell, 2019). Thus, the Homogeneity test assumptions must be checked before the final data analysis and interpretation.

METHODOLOGY

A multi-stage, three-step approach to data cleaning was adopted, as provided by Huang (2019). This includes data examination through careful planning and scrutiny, exploratory analyses of cleaned data using scatterplots, boxplots, and distribution tests, and, finally, the editing process that involves making necessary adjustments to address and rectify errors in the dataset. A sample size of 266 undergraduate students was determined using Anokye's (2020) table at a 95% confidence interval and a t-test value of 1.96. It was later increased by 40% (an additional 106 participants) resulting in a new sample of 372 respondents. The respondents were selected using a snowball sampling technique. The snowballing technique was employed in this study to ensure that only university students with e-commerce experience or who had previously made online product purchases from any e-commerce trading platform were included using an adapted questionnaire instrument. The questionnaire includes 23 items and two items on respondent demographics. The questionnaire items were adapted from prior researchers and used a 5-point Likert scale. The attitude and tax awareness items were adapted from Taing and Chang (2020). The behavioural control and subjective norms items were adapted from Pratama and Jin (2019) and Taing and Chang (2020). The tax compliance items were adapted from Pratama and Jin (2019) and Nurlis and Ariani (2020). Finally, the items for technology optimism were adapted from Parasuraman and Colby (2015). The descriptive statistics of the constructs were measured using SPSS version 26.

Step 1: Examining the Data

The collected primary data at this stage were first scrutinized for identification of data issues, such as inconsistent numbers or missing values, which were relatively easy to spot. Identifying and addressing potential issues in data at this stage is central to practical analysis. Detecting missing or duplicate values is essential; duplicates should be removed and missing data filled in thoughtfully. Inconsistencies and conflicts often arise during data merging, requiring careful handling to avoid duplication (Huang, 2019). The relevant information is presented in Table 3.

Step 2: Exploratory analyses of the Data

This stage focuses on data visualization using tools such as scatterplots, boxplots, and distribution tests to identify patterns in the data and make errors more visible. Visualization is a powerful tool for exploring and understanding a given dataset more directly. According to Huang (2019), exploratory analysis methods differ for one-, two-, and multidimensional data. One-dimensional exploration uses tools such as Boxplots and Histograms to display the distribution of numeric data, as shown in Figure 1. Scatterplots reveal relationships between two numeric series in two-dimensional analysis, whereas bar graphs illustrate the characteristics of categorical data, as shown in Figures 2 and 3, respectively. These methods aid in identifying outliers, relationships, patterns, and trends within the data, enhancing its quality for further analysis. Once the risible errors are identified and removed, the data visualization becomes more normally distributed at this level.

Step 3: Editing the Data

The final stage involves making essential adjustments to fix errors in a dataset. At this stage, researchers use various methods to address problems in a dataset. Huang (2019) notes that researchers at this level begin by standardizing data types, correcting numeric errors, and ensuring compatibility with algorithms. Next, they handle missing or duplicated values based on metadata, such as calculating missing pressure values using depth information, as shown in Table 4. Data conversion is used to address unit inconsistencies, ensuring datasets with different units are

compatible. Additionally, issues may be identified within a dataset for future attention, and missing values can be imputed using existing data and relationships in later analyses. The deleted data are also presented in Table 5.

FINDINGS AND DISCUSSION

Findings

Response Rate

Table 1 presents the response rate for the distributed questionnaires. Of the 372 questionnaires distributed, 367 were returned, yielding a 99% response rate. Of the returned questionnaires, 323 were deemed valid for analysis, yielding an 87% validity rate. The obtained 87% valid response rate is above the minimum acceptance rate of 30%, as suggested by Hair et al. (2019). This means that the higher the response rate in a survey study, the lower the risk of non-response bias (Devi et al., 2018). However, 44 questionnaires were excluded, constituting 12% of the returned questionnaires. Five questionnaires were ultimately not returned, comprising a 1% non-response rate. The response rate indicates that many undergraduate students completed the questionnaires, and the high number of valid responses shows that students actively participate in e-commerce, providing sufficient data for further analysis.

Table 1. Summary of the Instrument Response Rate by the University Students

Questionnaires Description	Frequency	Percentages %
Distributed questionnaire	372	100
Returned questionnaires	367	99
Returned and valid questionnaires	323	87
Returned and excluded questionnaires	44	12
Not returned questionnaires	5	1

Demographic Information of the Valid Respondents

Demographic information about the respondents provides a detailed description of them. Table 2 presents a breakdown of respondents' demographic information for the study. The result further indicates that the valid male and female respondents are 185 (57%) and 138 (43%). Male respondents, accounting for 57%, are much higher than female respondents, who represent 43% of the overall valid respondents. Regarding age categorization, 162 respondents (50%) fall within the 15-25 years age bracket. The second age category includes 133 (41%) respondents, within the 26-35 years age range. The third category is 22, within the age limit of 36-45, representing 7%, while the final age categories are 6, representing only 2% with an age limit above 45 years. The results clearly show that the vast majority of respondents, 50% of whom are undergraduate students, are aged 15-25 years. They are mainly transitioning from adolescence to early adulthood at university.

Table 2. Demographic Information of the Respondents

Demographic Variables	Categories	Frequency (N)	Percentage (%)
Gender	Male	185	57
	Female	138	43
Total		**323**	**100**
Age	15 -25 years	162	50
	26 – 35 years	133	41
	36 - 45 years	22	7

Demographic Variables	Categories	Frequency (N)	Percentage (%)
	Above 45 years	6	2
Total		**323**	**100**

Missing Data Evaluation

Missing data points are a critical issue affecting data analysis. They are identified from empty cells coded as unique values (Hahs-Vaughn & Lomax, 2020). In this study, 88 cases (0.67%) were missing out of 13,212 total data points, as indicated in Table 3. The missing data percentage is far below the standard 10% tolerance from the total data points. Experts assert that missing data cases or observations of up to 10% (Hair et al., 2019) or 20%–30% (Collier, 2020) are generally acceptable. Researchers can handle them using an imputation strategy prior to final execution, and they must be represented in the data frame (Hayes, 2022). The 88 missing data frame breakdowns for the current study are presented in Table 4 and replaced using the Series Mean (SMEAN), as Collier (2020) recommended. In this study, the missing data were random, as shown in Tables 3 and 4. However, missing data are random when they are unrelated to the other variables of the study (Collier, 2020; Hair et al., 2019; Tabachnick & Fidell, 2019). That is, all issues related to missing data size and percentage were addressed in accordance with the minimum and maximum thresholds recommended by experts.

Table 3. Missing Data Evaluation Summary

Total Rows	Total Columns	Total Data Points	Total Missing	Per cent of Missing Values (%)
367	36	13,212	88	0.67

Table 4. Missing Data Replacement Frame Before and After Series Mean Replacement

Items	Valid	Missing Before	Valid	Missing After	Minimum	Maximum
ATT3	364	3	367	0	1	5
ATT4	362	5	367	0	1	5
ATT5	364	3	367	0	1	5
SJN2	362	5	367	0	1	5
SJN3	345	22	367	0	1	5
SJN4	366	1	367	0	1	5
SJN5	362	5	367	0	1	5
SJN6	365	2	367	0	1	5
BHC1	365	2	367	0	1	5
BHC2	363	4	367	0	1	5
BHC4	365	2	367	0	1	5
BHC5	365	2	367	0	1	5
TXA2	366	1	367	0	1	5
TXA3	366	1	367	0	1	5
TXA4	352	15	367	0	1	5
TXA5	365	2	367	0	1	5
TCI1	364	3	367	0	1	5
TCI2	366	1	367	0	1	5
TCI3	366	1	367	0	1	5

Items	Valid	Missing Before	Valid	Missing After	Minimum	Maximum
TCI4	366	1	367	0	1	5
TEO1	366	1	367	0	1	5
TEO2	366	1	367	0	1	5
TEO3	362	5	367	0	1	5
Total		**88**		**0**		

Univariate and Multivariate Outliers Analysis

Outliers in a given data set seriously affect data representation and analysis. Outliers are extreme scores that fall beyond the standard distribution line and are often depicted by dots or asterisks (Hahs-Vaughn & Lomax, 2020). Outliers are univariate or multivariate observations in a given data set. Besides, univariate outliers are easier to spot in dichotomous variables. In that regard, 27 z-score values below -3.0 were deleted. Deleting extreme cases with a z-score below -3.0 helps detect more univariate outliers than deleting those with a z-score above or below ±3.26. The list of all the 27 identified and deleted univariate outliers is summarised in Table 5. The second outliers detected and deleted were the multivariate outliers. The multivariate outliers are easily detected using Mahalanobis D (D2) in combination with a chi-square table (Hair et al., 2019; Tabachnick & Fidell, 2019). Mahalanobis D examines the uniqueness of a single observation relative to the differences between its values and the means of all other observations across all study-independent variables (Hair et al., 2019; Pallant, 2020). Mahalanobis D was combined with the chi-square table, as depicted in Table 5. The current research study had six variables, with a degree of freedom (df) of 5 at a 0.05 alpha level, yielding a maximum chi-square value of 11.07.

In comparison, 17 additional cases were excluded due to multivariate outliers with chi-square values exceeding 11.07, as presented in Table 5. A cumulative total of 44 cases were deleted listwise due to extreme outliers. Additionally, the information indicates that all extreme cases above and below the recommended outlier thresholds were removed to ensure the data were normal for further analysis.

Table 5. Deleted Univariate and Multivariate Outliers Detected

Univariate Outliers Detected (<- 3.0)		Multivariate Outliers Detected (> 11.07 X^2)	
S/N	**ID**	**ID**	**MAH_1**
1	12	7	11.24804
2	28	57	12.54536
3	32	59	12.23479
4	44	83	11.6503
5	45	85	12.00504
6	69	130	12.28578
7	73	133	15.8438
8	90	142	14.52531
9	105	204	12.37989
10	119	216	13.97485
11	131	218	11.14934
12	168	241	14.16028
13	175	248	13.22393
14	177	250	22.57726
15	193	304	16.71405

Univariate Outliers Detected (<- 3.0)		Multivariate Outliers Detected (> 11.07 X^2)	
S/N	**ID**	**ID**	**MAH_1**
16	198	311	13.26543
17	225	326	12.48418
18	234		
19	236		
20	297		
21	300		
22	303		
23	351		
24	355		
25	359		
26	361		
27	367		
Individual Total	**27**	**17**	
Overall Total	**44**		

Normal Distribution Evaluation

Eliminating outliers is critical to normalizing a dataset that is normally distributed, which is symmetrical in shape and also known as a Gaussian distribution (Kumar & Misra, 2020). Normality can be tested graphically using a Histogram, Q-Q Plot, Box Plot, and Normal Probability Plot, or analytically using tests such as the Shapiro-Wilk, Kolmogorov-Smirnov, and D'Agostino-Pearson Tests (Pallant, 2020). If the p-value is less than 0.05, the data are not normally distributed; if it is above 0.05, the data are normally distributed (Hair et al., 2019; Pallant, 2020; Tabachnick & Fidell, 2019). The data normalization information for the current study is reported and depicted in the histogram in Figure 1. Histograms typically show the shape of a data distribution, reflecting its skewness and kurtosis (Hair et al., 2019). As such, the histogram output further indicates that the screened data has no extreme outliers and is not skewed in either direction. Figure 2 shows that the data are normally distributed, as the observed cases are not far from a straight line at 45° on the standard probability plot. The normal probability plot is appropriate if the graphic shows the sample distribution matches the normal distribution, as indicated by a straight line at 45 degrees (Hair et al., 2019). The next is the scatterplot of the residuals. A residuals scatterplot measures the bivariate relationships among multiple continuous variables and is considered normal if the observed cases trail off symmetrically from the centre in multivariate space (Tabachnick & Fidell, 2019). The results, as shown in Figure 3 (a-c), indicate that the observed residual values were not sparsely distributed or concentrated in a single location but were widely spread in a linear order. That implies that the data used were normally distributed across the scatterplot.

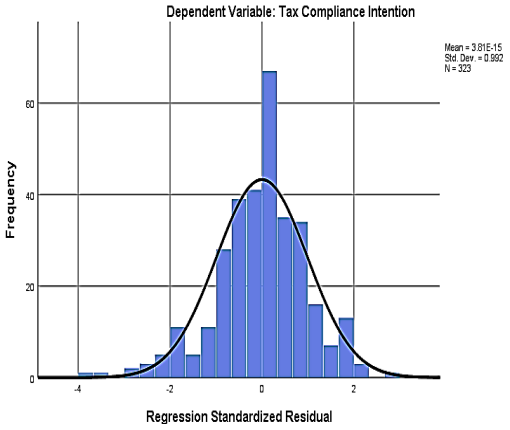

Figure 3a. Normal Distribution of Data on Histogram

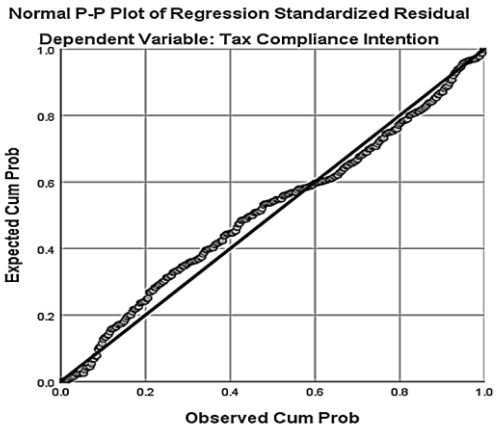

Figure 3b. Normal Probability Plot

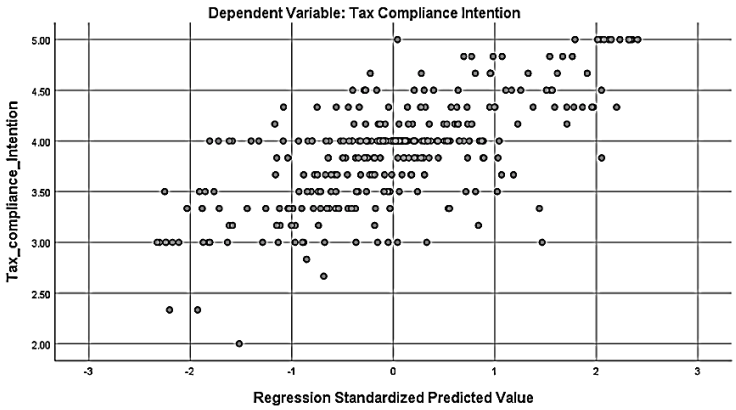

Figure 3c. Scatterplot

Discussion

This study meticulously collected, cleaned, and screened data on e-commerce users' behaviour, technology optimism, and tax compliance intention among Nigerian university students. The process achieved a high valid response rate, efficiently handled missing data and outliers, and confirmed data normalization through various statistical measures. The data were found to be normally distributed, aligning with the findings of Liao et al. (2023) and Li et al. (2019). These findings are consistent with the current study, which followed the scientific process of data collection, screening, cleaning, and normalization. Arevalo et al. (2022) also highlighted the importance of enhancing data quality through systematic data-cleaning procedures in a health survey. Babagana et al. (2019) achieved a highly valid response rate in their study on employee participation and performance appraisal among Nigerian academics. Despite these commonalities, this study focuses on e-commerce users in Nigeria, a demographic context with limited prior data cleaning information. The rigorous handling of missing data and outliers in this study ensures the reliability of the final results and contributes to a broader understanding of the variables under study.

CONCLUSIONS

The study collected, cleaned, and screened data on e-commerce users' behaviour, technology optimism, and tax compliance intention among Nigerian university students. Missing data were imputed using the SMEAN method, and outliers were identified using z-scores and chi-square criteria. The data were confirmed to be normally distributed graphically through

histograms, scatterplots, and standard probability plots, thereby meeting the parametric test assumption. This thorough process resulted in a dataset with high validity and normalization, enhancing the reliability of the findings and contributing to the understanding of e-commerce users in the Nigerian context. It also highlights the importance of effective data collection, cleaning, and screening for accurate, trustworthy, and meaningful research outcomes.

LIMITATION & FURTHER RESEARCH

The limitations of the study's findings were considered and reported to inform future adjustments and replication. Despite efforts to ensure data validity through effective and efficient data cleaning and screening, the study focused on e-commerce users among Nigerian undergraduate students. Therefore, future researchers can adopt the scientific procedure used in this study and replicate it with undergraduate and postgraduate students in public and private higher education institutions, as well as in other related research contexts. Furthermore, the study's cross-sectional nature summarises the participants' behaviours and intentions. However, longitudinal research can be carried out to provide a more comprehensive understanding of these constructs over time. Despite these limitations, the study provides valuable insights into the targeted context, which has received limited research on data cleaning and reporting.

REFERENCES

Alotaibi, O., Pardede, E., & Tomy, S. (2023). Cleaning Big Data Streams: A Systematic Literature Review. *Technologies*, *11*(101), 1-24. https://doi.org/10.3390/technologies11040101

Alvo, M., & Yu, P. L. H. (2018). *A Parametric Approach to Non-parametric Statistics*. Switzerland. https://doi.org/10.1007/978-3-319-94153-0

Anokye, M. A. (2020). Sample Size Determination in Survey Research. *Journal of Scientific Research & Reports*, *26*(5), 90-97. https://doi.org/10.9734/JSRR/2020/v26i530263

Arboretti, R., Bathke, A., Bonnini, S., Bordignon, P., Carrozzo, E., Corain, L., & Salmaso, L. (2018). *Parametric and Non-parametric Statistics for Sample Surveys and Customer Satisfaction Data*. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-91740-5

Arbuckle, J. L. (2017). *IBM® SPSS® Amos™ 25 User's Guide*. Amos Development Corporation.

Arevalo, M., Brownstein, N. C., Whiting, J., Meade, C. D., Gwede, C. K., Vadaparampil, S. T., . . . Christy, S. M. (2022). Strategies and Lessons Learned During Cleaning of Data from Research Panel Participants: Cross-sectional Web-Based Health Behavior Survey Study. *JMIR Form Res*, *6*(6), e35797. https://doi.org/10.2196/35797

Babagana, S. A., Mat, N. B., & Ibrahim, H. B. (2019). Moderating Effect of Employee Participation on Factors that Determine Effective Performance Appraisal (EPA): Data Screening and Preliminary Analysis. *International Journal of Academic Research in Business and Social Sciences*, *9*(4). https://doi.org/10.6007/IJARBSS/v9-i4/5826

Beukelman, T., & Brunner, H. I. (2016). *Chapter 6 - Trial Design, Measurement, and Analysis of Clinical Investigations*, Editor(s): Ross E. Petty, Ronald M. Laxer, Carol B. Lindsley, Lucy R. Wedderburn, Textbook of Pediatric Rheumatology (Seventh Edition), W. B. Saunders. https://doi.org/10.1016/b978-0-323-24145-8.00006-5

Bogo, A. B., Henning, E., & Kalbusch, A. (2023). Statistical Parametric and Non-Parametric Control Charts for Monitoring Residential Water Consumption. *Sci Rep*, *13*(1), 13543. https://doi.org/10.1038/s41598-023-40584-w

Chea, V., & Chea, P. (2022). *Family Background as the Determinant of University Student's Technological Readiness: Evidence from Cambodia*. In 2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) (pp. 322–328). IEEE. https://doi.org/10.1109/SKIMA57145.2022.10029566

Collier, J. E. (2020). *Applied Structural Equation Modeling Using AMOS: Basic to Advanced Techniques*.

Routledge: Taylor & Francis Group.

Denis, D. J. (2020). *Univariate, Bivariate, and Multivariate Statistics Using R Quantitative Tools for Data Analysis and Data Science*. John Wiley & Sons, Inc.

Devi, M., Azfar, M., & Tanwar, N. (2018). *Chapter - 3 Statistical Treatment of Non-Response in Sample Surveys*. Research Trends in Mathematics and Statistics (pp.31-50). AkiNik Publications https://doi.org/10.22271/ed.book14a03

Dickhaus, T. (2018). *Theory of Non-parametric Tests*. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-76315-6

Hahs-Vaughn, D. L., & Lomax, R. G. (2020). *An Introduction to Statistical Concepts* (4th ed.). Taylor & Francis.

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (A. Ainscow, J. Grene, & S. Clarke, Eds. 8th ed.). Cengage Learning EMEA.

Hair, J. F. J., Marcelo, L. D. S. G., da Silva, D., & Braga Junior, S. (2019). Development and Validation of Attitudes Measurement Scales: Fundamental and Practical Aspects. *RAUSP Management Journal*, *54*(4), 490-507. https://doi.org/10.1108/rausp-05-2019-0098

Hayes, A. F. (2022). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (D. A. Kenny & T. D. Little, Eds. 3rd ed.). The Guilford Press.

Hsieh, S. F., Yorke-Edwards, V., Murray, M. L., Diaz-Montana, C., Love, S. B., & Sydes, M. R. (2023). Lack of Transparent Reporting of Trial Monitoring Approaches in Randomized Controlled Trials: A Systematic Review of Contemporary Protocol Papers. *Clin Trials*, *20*(2), 121-132. https://doi.org/10.1177/17407745221143449

Hu, Y., Li, H., & Tan, F. (2024). Testing the Parametric Form of the Conditional Variance in Regressions Based on Distance Covariance. *Computational Statistics & Data Analysis*, *189*, 107851. https://doi.org/10.1016/j.csda.2023.107851

Huang, F. (2019). *Data Cleansing*. In: Schintler, L., McNeely, C. (eds) Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_300-1

Jian Ai, Y., Chew Sze, C., Sook Fern, Y., Hen Toong, T., & Boon Chian, C. (2021). The Use of e-Wallet Among Gen-Y in Malaysia During the Global Pandemic: An Analysis Using PLS-SEM. *Applied Quantitative Analysis, 1*(1), 1-8. https://doi.org/10.31098/quant.597

Kumar, A., & Misra, D. K. (2020). A Review of the Statistical Methods and Implementation of Homogeneity Assessment of Certified Reference Materials in Relation to Uncertainty. *MAPAN*, *35*(3), 457-470. https://doi.org/10.1007/s12647-020-00383-4

Lai, G., Liao, L., Zhang, L., & Li, T. (2023). Wind Speed Power Data Cleaning Method for Wind Turbines Based on Fan Characteristics and Isolated Forests. *Journal of Physics: Conference Series*, *2427*(1), 012001. https://doi.org/10.1088/1742-6596/2427/1/012001

Li, C., Hou, Y., & Yu, Z. (2019). Research on Data Cleaning Technology Based on Instance Level. *Journal of Physics: Conference Series*, *1213*(2), 022021. https://doi.org/10.1088/1742-6596/1213/2/022021

Li, R., & Lv, S. (2023). Research on Data Cleaning Method of Metal Material Corrosion Fatigue Test Data. *Journal of Physics: Conference Series*, *2468*(1), 1-7. https://doi.org/10.1088/1742-6596/2468/1/012097

Liao, L., Liu, X., Wu, Q., Kang, L., & Shang, Y. (2023). Data Cleaning Method of Distributed Photovoltaic Power Generation Based on Clustering Algorithm. *Journal of Physics: Conference Series*, *2474*(1), 012038. https://doi.org/10.1088/1742-6596/2474/1/012038

Love, S. B., Yorke-Edwards, V., Diaz-Montana, C., Murray, M. L., Masters, L., Gabriel, M., . . . Sydes, M. R. (2021). Making a Distinction Between Data Cleaning and Central Monitoring in Clinical Trials. *Clin Trials*, *18*(3), 386-388. https://doi.org/10.1177/1740774520976617

Monsurat, A., Shehu, A., & Usman, N. A. (2023). Relationship Between Consumer Competency, Value, Susceptibility to Control, Communication and Coproduction in MTN in Zaria Local Government

of Kaduna State. *Applied Quantitative Analysis, 2*(2), 14-27. https://doi.org/10.31098/quant.1144

Neutatz, F., Chen, B., Abedjan, Z., & Wu, E. (2021). From Cleaning before ML to Cleaning for ML. *IEEE Data Eng. Bull., 44*(1), 24-41.

Nurlis, N., & Ariani, M. (2020). Tax Awareness Moderates Knowledge and Modernization of Tax Administration on Tax Compliance, Survey on MSME Taxpayers in South Tangerang City, Indonesia. *International Journal of Management Studies and Social Science Research*, *2*(5), 250-259.

Pahlevan, S. S., & Shafir, N. H. (2019). *Exploratory Factor Analysis and Structural Equation Modeling with SPSS and AMOS*. Tehran Artin Teb.

Pallant, J. (2016). *SPSS Survival Manual; A Step by Step Guide to Data Analysis Using IBM SPSS*. Open University Press - McGraw-Hill Education.

Pallant, J. (2020). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS* (7th, Ed.). Open University Press - McGraw-Hill Education.

Parasuraman, A., & Colby, C. L. (2015). An Updated and Streamlined Technology Readiness Index. *Journal of Service Research*, *18*(1), 59-74. https://doi.org/10.1177/1094670514539730

Parilla, E. S., & Abadilla, M. E. M. (2021). Business Students' Assessment of Attitudes and Readiness Towards Online Education. *Applied Quantitative Analysis, 1*(2), 1-17. https://doi.org/10.31098/quant.779

Pauch, D. (2023). Tax Knowledge and Tax Perception by Students at the University of Szczecin. *Zeszyty Teoretyczne Rachunkowości*, *47*(1), 121-133. https://doi.org/10.5604/01.3001.0016.2910

Pratama, A. R. P., & Jin, Z. (2019). Foreign Students Intention towards a China Third Party Mobile and Online Payment Platform Based on Alipay. *International Journal of Informatics and Computation*, *1*(1), 10-20. https://doi.org/10.35842/ijicom.v1i1.8

Qin, B., Luo, Q., Li, Z., Zhang, C., Wang, H., & Liu, W. (2022). Data Screening Based on Correlation Energy Fluctuation Coefficient and Deep Learning for Fault Diagnosis of Rolling Bearings. *Energies*, *15*(2707), 1-21. https://doi.org/10.3390/en15072707

Rahul, K., Banyal, R. K., & Goswami, P. (2020). Analysis and Processing Aspects of Data in Big Data Applications. *Journal of Discrete Mathematical Sciences and Cryptography*, *23*(2), 385-393. https://doi.org/10.1080/09720529.2020.1721869

Rezvani, A., Bigverdi, M., & Rohban, M. H. (2022). Image-Based Cell Profiling Enhancement via Data Cleaning Methods. *PLoS One*, *17*(5), e0267280. https://doi.org/10.1371/journal.pone.0267280

Saidu, M. A., & Ladan, S. S. (2023). A Conceptual Framework on Tax Knowledge and Tax Compliance Intention: The Moderating Effect of Patriotism in Nigeria. *Bullion*, *47*(2), 51-63.

Saidu, M. A., Shagari, S. L., Kabir, M. A., & Abubakar, A. (2022). Perceived effect of e-commerce tax awareness and technology optimism on tax compliance intention. *Journal of Integrated Sciences*, *3*(1), 44-97.

Salcedo, J., & McCormick, K. (2020). *SPSS Statistics for Dummies* (4th ed.). John Wiley & Sons, Inc.

Sarstedt, M., & Mooi, E. (2019). *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics* (3rd ed.). Springer.

Savani, B. N., & Barrett, A. J. (2009). How to Build and Use a Stem Cell Transplant Database. In *Hematopoietic Stem Cell Transplantation in Clinical Practice* (pp. 505-512). Churchill Livingstone. https://doi.org/10.1016/b978-0-443-10147-2.50053-9

Sheskin, D. J. (2011). *Handbook of Parametric and Non-Parametric Statistical Procedures* (5th ed.). CRC Press Taylor & Francis Group.

Silvia, P. J., & Cotter, K. N. (2021). Cleaning and Processing Your Data. In *Researching Daily Life: A Guide to Experience Sampling and Daily Diary Methods* (pp. 93-109). American Psychological Association. https://doi.org/10.1037/0000236-006

Steorts, R. C. (2023). A Primer on the Data Cleaning Pipeline. *Journal of Survey Statistics and Methodology*, *11*(3), 553-568. https://doi.org/10.1093/jssam/smad017

Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics* (Seven, Ed.). Pearson Education, Inc.

Taing, H. B., & Chang, Y. (2020). Determinants of Tax Compliance Intention: Focus on the Theory of Planned Behavior. *International Journal of Public Administration*, *44*(1), 62-73. https://doi.org/10.1080/01900692.2020.1728313

Van Buren, E., & Herring, A. H. (2020). To be Parametric or Non-Parametric, That is the Question: Parametric and Non-Parametric Statistical Tests. *BJOG*, *127*(5), 549-550. https://doi.org/10.1111/1471-0528.15545

Watkins, M. W. (2021). *A Step-by-Step Guide to Exploratory Factor Analysis with SPSS*. Routledge, Taylor & Francis Group.

Yeo, A. A., Lim, T. C., & Azhar, Z. (2019). Exploring Malaysian E-Commerce Taxation: A Qualitative Insight of Online Businesses. *Journal of Contemporary Issues and Thought*, *9*, 75-85. https://doi.org/10.37134/jcit.vol9.8.2019