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

Abstract

Effective data analysis reflects an improved approach to data collection, cleaning, and screening. However, very few studies reported information on the techniques used in cleaning and screening 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, 367 adapted survey instrument items were administered using Snowballing. Finally, 365 were retrieved from the respondents. Missing data were all replaced using the Series Mean (SMEAN), and outliers were appropriately addressed through z-score values and chi-square criteria. Descriptive statistical measures were used to examine the dataset and presented in several tables, a histogram, a scatterplot and a normal probability plot. The collected, cleaned, and screened data were found to have a normal distribution, helping analyze and understand the parametric data distribution, variation, and normalization. The findings provide valuable information for university students, academics, policymakers, and practitioners to adopt during data collection, cleaning and screening exercises. 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 for better data analysis and 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 looks into the critical structuring, summary and representation of data for better comprehension, explanation and display of information for decision-making. Researchers use it to determine whether the obtained data are reliable with a proposed causal process (Hayes, 2022). Data analysis involves using various methods to examine data and identify characteristics like data type, size, format, patterns and different steps, including finding data issues, doing statistical inspection, 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 usually perform data analysis after the raw data are collected from the field or extracted from a given document. Raw data collected are data that have yet to be used and need

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

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

cleaning analysis, filtering and processing. Proper data qualities are needed before the final data analysis for 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 for better data analysis. Data cleaning is crucial in research analysis as mistakes in the data can make the result application wrong, unfair or even risky (Neutatz et al., 2021). Interestingly, data cleaning fixes issues with data, like missing, wrong, 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 analyses are closely related. A cleaned dataset improves the data quality and makes the data analysis output more reliable and trustworthy (Alotaibi et al., 2023). Cleaning data before analyzing it is crucial as it involves fixing or removing incorrect, corrupted, wrongly formatted, duplicate or incomplete information in a dataset. This step is necessary for researchers to make better decisions based on the accuracy of the dataset. However, data cleaning aims to prepare data for functional statistical analyses on a given data source. This includes tasks like estimating populations, analyzing specific areas, making composite estimates, and building models like regression (Steorts, 2023).

Data cleaning deals with issues in a given dataset, such as missing values, duplicate data, strange and out-of-range values or unimportant 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 about their scope, frequency, and how they organized and cleaned their data, especially regarding ensuring the integrity of their dataset (Hsieh et al., 2023). Moreover, according to a report by Alotaibi et al. (2023), about 73% of research on fixing data issues only looks at finding outliers, which matters for understanding data, while nearly 13% of recent studies handle multiple issues like missing data, outliers, and duplicates. Alotaibi et al. reported that around 8% of the studies deal with missing values and about 5% concentrate on eliminating the duplicated data. Other issues that cleaning researchers face include low-quality data and challenges in choosing the proper structural parameters for the identification model (Qin et al., 2022). However, proper data cleaning can prevent problems of wrong data analysis, interpretation and conclusion.

Cleaning data can be complex because of the type of data (structured, semi-structured, or unstructured), especially for semi-structured and unstructured data. When cleaning data, problems like missing values, duplicates, outliers, and irrelevant information often appear (Alotaibi et al., 2023). To address data cleaning issues in research, using special tools like artificial intelligence, machine learning, deep learning, statistical methods, and combined and unclassified techniques helps fix abnormal datasets. These techniques make things simpler and more precise, with high accuracy, thereby ensuring the data management system works well within a given dataset (Alotaibi et al., 2023; Lai et al., 2023; Li & Lv, 2023; Rezvani et al., 2022). Even with many methods to clean data, sometimes researchers need to remember to consider critical data-cleaning techniques. This oversight can impact the accuracy and reliability of the final data analysis and result output. As evidence, several studies were recently conducted on data cleaning before data analysis. However, only a few such studies have looked at finding irrelevant data because it is not easy to spot unless, in some cases, the irrelevant 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. It is on this premise that this study sought to examine 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 check for logical sequences and expected ranges and ensures that no critical data is missing (Love et al., 2021). However, researchers must refine their raw data input, as any issues could impact 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 like minimum, maximum, out-of-range values, valid and missing cases, mean, and standard deviation (Pallant, 2020). A thorough review of statistics and graphics data is needed to reduce analysis 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, Black et al., 2019; Tabachnick & Fidell, 2019; Watkins, 2021). Furthermore, before starting data cleaning, researchers must combine their data files, check data quality, and remove or exclude incorrect participants 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 an examination process, exploratory analyses, and an editing process to identify and rectify errors in the dataset (Huang, 2019). Thus, these approaches are highly effective as they provide a brief and 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 fixing mistakes and inconsistencies and dealing with missing information. It begins with examining the univariate outlier at a threshold of ≤ ±3.29 z scores (Tabachnick & Fidell, 2019) and can also be checked at z scores values ≤ ±3.0 (Hair, Black, et al., 2019; Pallant, 2016). The next is examining Mahalanobis Distance (D2) for multivariate outlier detection, which checks for the adequacy of assumptions and preparation for modelling (Pahlevan & Shafir, 2019; Watkins, 2021). Researchers are to examine the value for D2 by comparing the number of independent variables against the degree of freedom (df) within the Chi-square critical values (Pallant, 2016; Pallant, 2020; Watkins, 2021). Besides, outlier cases are usually remedied pre- and post-analysis by removing the observations from the analysis (Hair, Black, et al., 2019). The collected data is subjected to data screening to ensure their normality and appropriateness before final reporting, using the IBM SPSS software package (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 should be normally distributed, sample variances should be similar, suitable variables should be used, samples should be independently selected, and no outliers should be present.

Besides, if the actual values in a dataset deviate from a normal distribution, transforming the data using a natural log scale can render parametric tests appropriate (Alvo & Yu, 2018; Arboretti et al., 2018; Dickhaus, 2018). Equally, the Non-parametric tests are used when parametric tests are not suitable, focusing on ranks and testing for distribution 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 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 if different samples from various populations in a study have the same variance, a concept also known as homoscedasticity. The assumption of homogeneity ensures that all comparison groups (two or more) in a parametric test have equal variance (Pallant, 2020). This assumption is applied in statistical tests that check

for normality, like z-test, t-test, Chi-squared, ANOVA, ANCOVA, and MANOVA (Pallant, 2020; Kumar & Misra, 2020; Tabachnick & Fidell, 2019). However, in t-test analysis, an alternate Levene's test options are provided for selection, based on equally and not equally assumed outputs. The homogeneity test is based on a null hypothesis stating that the samples have equal variances with an alpha value above 0.05. A normal continuous data will accept the null hypothesis with 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 like Hartley's Fmax, Box's, Cochran, Levene, and Bartlett's tests are used to evaluate data homogeneity. If the homogeneity test assumption is violated, the final results of statistical tests become unreliable, necessitating alternative non-parametric 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 depicting cleaned data using scatterplots, boxplots, and distribution tests, and finally, the editing process that involves making necessary adjustments to address and rectify the errors in the dataset. A sample size of 266 undergraduate students was determined using the Anokye (2020) table at a 95% confidence interval and a t-test value of 1.96. It was later increased by 40% (106 more participants), as Salkind (2018) recommended, resulting in a new sample size 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 respondent demographic information items. The questionnaire measurement items were adapted from prior researchers on 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 was first scrutinized for identification of data issues such as inconsistent numbers or missing values that were relatively spotted. Identifying and addressing potential issues in data at this stage is central to practical analysis. Detecting missing or duplicate values is essential, with removal for duplicates and thoughtful filling for missing data. 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 revolves around data visualization using tools like scatterplots, boxplots, and distribution tests, assisting in recognizing patterns within the data and making errors more visibly.

Visualization is a powerful tool for exploring and understanding a given dataset more directly. According to Huang (2019), exploratory analysis methods vary for one-dimensional, two-dimensional, and multi-dimensional data. One-dimensional exploration involves tools like Boxplot and Histogram, which display the distribution of numeric data, as presented in Figure 1.

Scatterplots reveal relationships between two numeric series for two-dimensional analysis, while bar graphs showcase categorical data characteristics as presented 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 start by adjusting data types for consistency, fixing numeric errors, and ensuring compatibility with algorithms. Next, they handle missing or duplicated values based on metadata, like calculating missing pressure values using depth information, as presented in Table 4. Data conversion is employed for unit inconsistency, ensuring datasets with different units are compatible. Additionally, issues may be identified within a dataset for future attention, and replacements can be imputed for missing values based on 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 information on the response rate from the distributed questionnaires. Of the 372 questionnaires distributed, 367 were finally returned, indicating a high response rate of 99%. Among the returned questionnaires, 323 were deemed valid for analysis, representing an 87% validity rate. The obtained 87% valid response rate is above the minimum acceptance rate of 30%, as suggested by Sekaran and Bougie (2016) and Hair, Black, et al. (2019). This means that the higher the response rate in a survey study, the lesser 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 responded to the questionnaires, and the high number of valid responses shows that the students actively participate in e-commerce, providing sufficient valid 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 of the respondents provides a detailed description of the respondents. Table 2 below shows the breakdown of the 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. On the age categorization, 162 respondents,

representing 50%, fall within the age bracket of 15-25 years. The second age category has 133 (41%) respondents, accounting for the age limit of 26-35 years. 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 result clearly shows that the vast majority of the respondents, accounting for 50% of all respondents who are undergraduate students, have an age limit ranging

between 15 and 25 years. They are mainly transiting from adolescence to early adulthood in the universities.

Table 2: Demographic Information of the Respondents

Demographic Variables	Categories	Frequency (N)	Percentage (%)
	Male	185	57
Gender	Female	138	43
Total		**323**	**100**
	15 -25 years	162	50
Age	26 – 35 years	133	41
	36 - 45 years	22	7
	Above 45 years	6	2
Total		**323**	**100**

Missing Data Evaluation

Missing data points are critical issues affecting data analysis. They are identified from empty cells coded as unique values (Hahs-Vaughn & Lomax, 2020). In this study, the missing data cases were 88 out of the total data points of 13,212 data entries, representing 0.67%, as indicated in Table 3. The missing data per cent is far below the standard 10% tolerance from the total data points. Experts assert that missing data cases or observations up to 10% (Hair, Black, et al., 2019) or up to 20% – 30% (Collier, 2020) are generally acceptable. Researchers can handle them using the imputation strategy before the 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 current study, the missing data were random, as seen in Table 3 and Table 4. However, missing data are random when they are unrelated to the other variables of the study (Collier, 2020; Hair, Black, et al., 2019; Tabachnick & Fidell, 2019). That is to say, all issues relating to the missing data size and percentage were taken care of based on the minimum and maximum requirements as suggested 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

Items	Valid	Missing Before	Valid	Missing After	Minimum	Maximum
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
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 normal distribution line and are often depicted by dots or asterisks (Hahs-Vaughn & Lomax, 2020). The outliers are in the form of univariate or multivariate, existing in a given data set. Besides, the univariate outliers are more accessible to spot among dichotomous variables. In that regard, 27 z-score values below -3.0 were deleted. Deleting extreme cases at a z-score value below -3.0 helps detect more univariate outliers for deletion than using a z-score value above and 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, Black, et al., 2019; Tabachnick & Fidell, 2019). Mahalanobis D examines the uniqueness of a single observation alongside the differences between the observation's values and the mean values for all other observations across all study independent variables (Hair, Black, et al., 2019; Pallant, 2020). Mahalanobis D was combined with the chi-square table, as depicted in Table 5. The current research study has six variables, with a degree of freedom (df) of 5 at a 0.05 alpha level and was used to arrive at a maximum chi-square value of 11.07.

In comparison, 17 other cases were deleted for multivariate outliers detected above a chi-square value of 11.07, as presented in Table 5. A cumulative total of 44 cases were deleted listwise for having extreme cases of outliers. Additionally, the information further signifies that all extreme cases above and below the recommended outliers' thresholds were eliminated to ensure the normality of the data 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

Univariate Outliers Detected (< - 3.0)		Multivariate Outliers Detected (> 11.07 X^2)	
S/N	**ID**	**ID**	**MAH_1**
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
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, which is normally distributed if it is symmetrical in shape, also known as 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 like the Shapiro-Wilk Test, Kolmogorov-Smirnov Test, and D'Agostino-Pearson Test (Pallant, 2020). If the p-value is less than 0.05, the data is not normally distributed, but if it is above 0.05, it is normally distributed (Hair, Black, et al., 2019; Pallant, 2020; Tabachnick & Fidell, 2019). The information for data normalization of the current study is reported and depicted on the histogram in Figure 1. Histograms typically show the shape of a data distribution, reflecting its nature with skewness and kurtosis values (Hair, Black, et al., 2019). As such, the histogram output further indicates that the screened data has no extreme cases of outliers and, at the same time, is not skewed in either direction. Figure 2 shows that the data is normally distributed as the observed cases were not far from the straight line stretched at an angle of 45o, using the normal probability. The normal probability plot is appropriate if the graphic shows the sample distribution shape to the normal distribution represented by a straight line angled at 45 degrees (Hair, Black et al., 2019). The next is the scatterplot of the residuals. A residuals scatterplot measures the bivariate relationships between multiple continuous variables and is considered normal if the observed cases are trailing off symmetrically from the centre for multivariate

(Tabachnick & Fidell, 2019). The result obtained, as presented in Figure 3, shows that the observed residual values were not sparsely distributed or concentrated in a single place but widely spread in a linear order. That implies that the data used were normally distributed across the scatterplot.

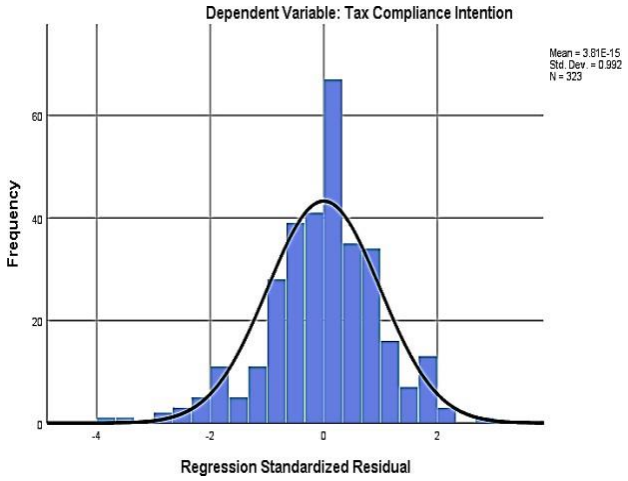

Figure 3: Normal Distribution of Data on Histogram

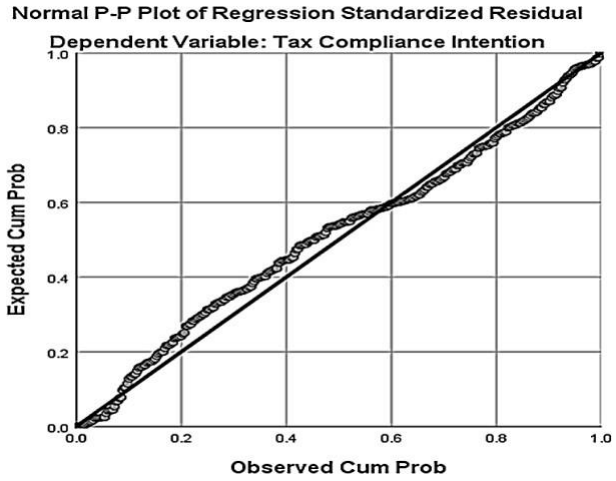

Figure 3: Normal Probability Plot

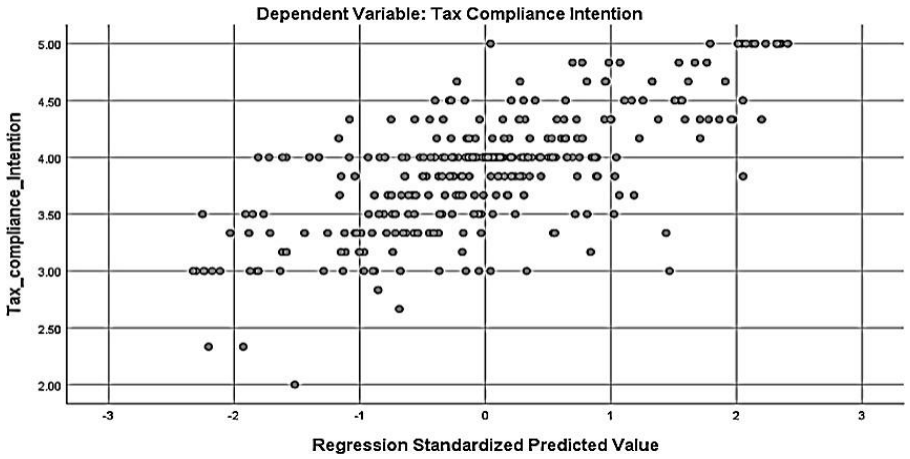

Figure 3: Scatterplot

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 that followed the scientific process of collecting data, screening, cleaning, and finally normalizing the data. Arevalo et al. (2022) also highlighted 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 final results' reliability 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 replaced using SMEAN, and outliers were addressed through z-score values and chi-square criteria. It was inferred that the data was confirmed to be normally distributed graphically through histograms, scatterplots, and normal probability plots, 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, trustful and meaningful research outcomes.

LIMITATION & FURTHER RESEARCH

The limitations of the findings in this study were considered and reported for future adjustments and replication. Despite the efforts to ensure data validity through effective and efficient data cleaning and screening, the study's scope was focused on e-commerce users among Nigerian undergraduate students. Therefore, future researchers can adopt the scientific procedure used in this study and replicate it among undergraduate and postgraduate students in public and private higher institutions alongside 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 contributes valuable information to the targeted context that received less research contributions on data cleaning and reporting.

REFERENCES

Afifi, A., May, S., Donatello, R. A., & Clark, V. A. (2020). *Practical Multivariate Analysis* (Sixth ed.). Taylor & Francis Group, LLC.

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). Trial Design, Measurement, and Analysis of Clinical Investigations. 54-77.e52. 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* 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA),

https://ieeexplore.ieee.org/document/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. 33-50. 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

Hair, J. F. J., Ortinau, D. J., & Harrison, D. E. (2021). *Essentials of Marketing Research* (M. Fossel, L. H. Spell, N. Young, J. McAtee, & E. Windelborn, Eds. Fifth ed.). McGraw-Hill Education.

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. 1-4. 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. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 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., & M Abadilla, M. E. (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's Third-Party Mobile and Online Payment Platform Based on Alipay. *International Journal of Informatics and Computation: Nanjing University of Information Science and Technology*, *1*(1), 1-11.

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

Ringle, C. M., Sarstedt, M., Sinkovics, N., & Sinkovics, R. R. (2023). A perspective on using partial least squares structural equation modelling in data articles. *Data Brief*, *48*, 109074. https://doi.org/10.1016/j.dib.2023.109074

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. www.cbn.gov.ng

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. 505-512.

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 PressTaylor &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