M.SC. DATA ANALYTICS & TECHNOLOGIES

ASSESSMENT 1- PORTFOLIO 1

DAT7301: DATA ANALYSIS AND VISUALISATION

REPORT 1: STATISTICAL STUDY DESIGN

BY: OKIKE U. J.

INSTRUCTOR: DR EUGEN HARINDA

DATE. 09-02-2025

ABSTRACT

Internally Generated Revenue (IGR) is a critical component of financial sustainability for sub-national governments in Nigeria. This study presents a comprehensive statistical analysis plan to analyse IGR trends in Nigeria for a period of four years, from 2019 to 2023 using data-driven methods to identify key patterns and determinants of revenue generation. Using a longitudinal research design, the study uses descriptive statistics and inferential analysis and methods to assess revenue volatility, tax efficiency and economic impact on IGR performance. The study further explores potential biases in income reporting and suggests strategies to minimise errors while ensuring data integrity and reliability. Ethical considerations, including data transparency and confidentiality, are also addressed to maintain research credibility. The outcomes provide valuable insights for policymakers, financial analysts and stakeholders to enhance tax policies, optimise revenue mobilisation and improve financial autonomy at the state level. By leveraging statistical modelling and data visualisation techniques, this study contributes to evidence-based policy formulation aimed at strengthening Nigeria's financial sustainability.

Keywords: Statistical Analysis; Machine Learning; Data Visualisation; Taxation Efficiency; Financial Data Analytics; Data Management; Data Security; Data Analytics; Ethical considerations; Time-Series Analysis; Public Finance.

TABLE OF CONTENTS

ABSTRACT.......II

1. INTRODUCTION IV

1.1 Overview of Data Analysis and Visualisation iv

1.2 Importance of Data-Driven Problem Solving.......iv

2. RESEARCH QUESTIONS.......VI

2.1 Developed Research Questions And Their Implications To The Study vi

2.2 Research Question and Objectives vi

2.3 Justification of the Report vii

3. STUDY DESIGN PLAN VIII

3.1 Type of Study and Justification.......viii

3.2 Target Population and Sampling Methods.......viii

4. ERRORS, BIASES AND MITIGATION STRATEGIES X

4.1 Identifying Potential Sources of Error.......x

4.2 Common Biases and Their Impact.......x

4.3 Strategies to Minimise Errors and Biases.......xi

5. STATISTICAL ANALYSIS PLAN.......XII

5.1 Overview of Statistical Approach xii

5.2 Descriptive Statistics.......xii

6. ETHICAL CONSIDERATIONS XV

6.1 Data Integrity and Transparency.......xv

6.2 Privacy, Confidentiality and Bias Prevention xv

6.3 Compliance with Ethical Research Standards.......xv

7. CONCLUSION XVII

REFERENCES XVIII

1. INTRODUCTION

1.1 Overview of Data Analysis and Visualisation

In today's digital age, data analysis and visualisation have become central and indispensable tools for making informed decisions in a variety of fields including business, healthcare, defence, space exploration, government and academia. The ever-increasing volume of data requires robust methodologies to extract meaningful insights, enabling organisations and researchers to make data-driven decisions.

Data analysis involves the systematic application of statistical and computational methods to identify trends, correlations and patterns present in data sets (Provost & Fawcett, 2013). Using statistical methods, researchers and analysts can summarise information, test hypotheses and develop predictive models that aid in decision making. At the same time, data visualisation plays an important role in presenting complex data in an easy-to-understand format using graphical elements such as charts, graphs and heat maps to enhance understanding and communication (Few, 2017).

Employing data analytics and visualisation to transform raw data into actionable insights, enables more effective problem solving and strategic planning. As organisations continue to leverage big data, mastering these analytical techniques becomes essential to generating reliable data and driving innovation.

1.2 Importance of Data-Driven Problem Solving

Data-driven methods have transformed problem solving by replacing instinctive solutions with empirical evidence (Field, 2018). By using structured data sets and statistical tools, researchers can reduce bias, test hypotheses and increase the reliability of results.

Organisations use data analytics to track trends, allocate resources and predict future events. In business, it helps improve customer understanding and operational efficiency; in management, it informs economic policies and strategies.

Compliance with ethical standards is important to ensure fair and transparent interpretation of data. Observing frameworks such as the FAIR principles (findability, accessibility, interoperability and reusability) maintains data integrity and builds public trust (Wilkinson et al., 2016). Understanding these methodologies helps make responsible and effective data-driven decisions.

2. RESEARCH QUESTIONS

2.1 Developed Research Questions And Their Implications To The Study

Revenue generation is an essential aspect of economic sustainability for sub-national governments. In Nigeria, Internally Generated Revenue plays a fundamental role in funding public services and reducing reliance on federal allocations. However, there are significant differences in IGR contributions across states, raising concerns about economic productivity, tax efficiency and governance structures (Adeniran & Osakede, 2021). A country's ability to mobilise internal revenue depends on many factors including industrialisation, tax discipline and financial policy, making it necessary to analyse these changes using data-driven methods. This report attempts to examine the patterns, determinants and trends of IGR in Nigeria from 2019 to 2023. The research problem focuses on understanding the economic and structural causes of revenue disparities across states and assessing the sustainability of current revenue generation strategies.

2.2 Research Question and Objectives

The following research question is proposed to guide the report: What are the key trends and determinants influencing changes in Internally Generated Revenue (IGR) in Nigerian states from 2019 to 2023?

Based on this, the study aims to achieve the following objectives:

i. Analyse IGR trends over a four-year period to identify revenue growth and volatility patterns.

ii. Examine major revenue components such as PAYE tax, direct assessments and other taxes.

iii. Evaluate the economic and political factors influencing revenue generation in states.

iv. Identify potential inefficiencies and biases in data collection and reporting that may impact revenue estimates.

2.3 Justification of the Report

Understanding the differences in IGR is important for policymakers, economists and financial planners. A well-structured analysis provides insights into effective revenue mobilisation strategies, helping states optimise tax policies and improve financial autonomy. Furthermore, identifying differences in income reporting ensures better accountability and financial management (Eze & Ogundipe, 2020). With increasing dependence on internal sources of funding for development projects, this report contributes to evidence-based policy making to enhance financial sustainability at the state level.

3. STUDY DESIGN PLAN

3.1 Type of Study and Justification

This study adopts a quantitative research approach by analysing secondary data to examine trends in Internally Generated Revenue (IGR) in Nigerian states from 2019 to 2023. A longitudinal design is used to observe changes in revenue over multiple years, allowing for trend analysis and identification of patterns (Creswell & Creswell, 2018). Using secondary data obtained from government publications and official revenue reports ensures reliability and comparability across states. This approach minimises data collection costs while maximising analytical depth.

3.2 Target Population and Sampling Methods

The study encompasses all 36 states of Nigeria and the Federal Capital Territory (FCT), making it more of a census study than a sample analysis. This comprehensive approach eliminates sampling bias and provides a comprehensive picture of IGR dynamics across the country. The dataset includes variables such as total IGR, revenue components (PAYE, direct assessments, road tax, etc.) and state-level economic indicators.

3.3 Determination of Sample Size and Power

Since this study covers the entire population of Nigerian states, there is no need to calculate sample size. The power of the analysis is enhanced by the large-scale nature of the dataset, which provides high statistical power for trend observations and inter-state comparisons (Babbie, 2020).

3.4 Identification of Variables and Measurement Methods

The table below outlines the variables and their details, being considered for usage in the study plan, it includes the variable names, types and measurement methods.

S/N	Variable Type	Variable Name	Measurement Method
1	**Dependent Variable**	Total IGR (2019-2023)	Naira (₦), annual state-level data
2	**Independent Variables**	PAYE, Direct Assessment, Road Taxes, Other Taxes	Naira (₦), disaggregated revenue components
3	**Control Variables**	Population, Economic Activity, Policy Interventions	Percentage growth, policy classifications

Table 1. Variables and their measurement methods (Source: Author)

3.5 Data Collection Approach and Justification

The study relies on secondary data sources, mainly official revenue reports from the National Bureau of Statistics (NBS), verified at the Joint Tax Board (JTB) and the National Revenue Authority. These government-verified sources will enhance the reliability and credibility of the data (Kumar, 2019).

Since the data were collected pre-collected, this study acknowledges potential limitations such as inconsistencies in reporting and updating revenue figures over time. However, cross-referencing across multiple government publications can mitigate these concerns, ensuring the integrity of the data in a reliable manner.

4. ERRORS, BIASES AND MITIGATION STRATEGIES

4.1 Identifying Potential Sources of Error

Data collection and reporting errors can significantly impact the accuracy of results, in this case, internally generated revenue (IGR) analysis. The following are the main potential sources of error in this study:

- Data reporting errors: Inconsistencies in income reporting across states can lead to differences in the overall IGR calculation. These errors can be due to administrative errors, misclassification of income sources, or incorrect calculations (Kumar, 2019).
- Data entry and processing errors: Human errors in recording and processing tax data can skew results. Differences in accounting practices at the state level can lead to inconsistencies (Biemer & Lyberg, 2003).
- Impact of time lags: Delays in compiling, collating and releasing official financial data can result in the use of outdated data, potentially leading to misinterpretations of current trends.
- Missing data: Some states may not provide complete records for all years, resulting in gaps that can impact longitudinal trend analysis. Missing values may require imputation methods to make additional assumptions (Little & Rubin, 2019).

4.2 Common Biases and Their Impact

Several biases can influence the interpretation of IGR data, potentially skewing the results. These biases include, but are not limited to:

- Selection bias: States with more structured financial systems and better tax administration may report revenues more accurately, while others with weaker institutions may underreport, leading to overestimation of differences in IGR.

- Measurement bias: Variations in tax classification and collection methods across states may lead to differences in reported revenues that do not reflect actual differences in tax efficiency.
- Political and administrative bias: Government agencies may over or under report revenue figures for political reasons, affecting the legitimacy of the reported data (Bartlett & Burton, 2016).
- Economic bias: Economic blows such as recessions, inflation and even the economic downturn caused by the COVID-19 pandemic, can impact tax collections differently across states, leading to misleading interpretations if the economic context is not considered.

4.3 Strategies to Minimise Errors and Biases

To ensure the consistency of the results, the following strategies can be implemented:

- Data cross-checking: Income reports from multiple sources (e.g. JTB, NBS, state tax authorities) can be compared to identify any discrepancies.
- Data cleaning methods: outlier detection, missing data imputation and normalisation methods can be applied to improve data quality (Gelman & Hill, 2007).
- Sensitivity analysis: Alternative statistical models can be used to test the versatility of the results to various assumptions and potential biases.
- Contextual adjustment: Economic and policy-related factors can be included to contextualise income trends and account for macro-economic impacts.

5. STATISTICAL ANALYSIS PLAN

5.1 Overview of Statistical Approach

To analyse the trend of Internally Generated Revenue (IGR) in Nigerian states over a period of four years (2019 - 2023), this study uses a blend of descriptive statistics, inferential analysis and advanced statistical modelling using R programming language. These methods facilitate trend identification, correlation analysis and predictive modelling of revenue generation patterns (Field, 2018).

5.2 Descriptive Statistics

Descriptive statistics summarise key revenue metrics, providing basic information about the distribution and trends of IGR:

- Measures of central tendency, which include mean, median and mode of IGR by state and year.
- Measures of dispersion, which involve the standard deviation and interquartile range to assess revenue variability.
- Data visualisation, which implies time series charts, box plots and histograms to illustrate changes in revenue and emissions (Few, 2017).

5.3 Inferential Statistics

Inferential methods can be used to determine the relationship and statistical significance in the income trend:

- Correlation analysis: Pearson or Spearman correlation to assess the relationship between IGR and independent variables such as economic activity and tax policy.
- Regression analysis: Multiple linear regression to model the impact of tax components (PAYE, direct assessment, etc.) on overall IGR.

- Hypothesis testing: ANOVA or t-test to compare income differences between states with different economic profiles (Gelman & Hill, 2007).

5.4 Advanced Statistical Techniques

To provide deeper insights, the study includes advanced analytics:

- Time series analysis: ARIMA model to forecast future IGR trends based on historical data.
- Cluster analysis: K-means clustering to group states based on income measures and economic similarities.
- Machine learning methods (if applicable): decision trees or random forests for income forecasting and pattern recognition (James et al., 2021).

5.5 Justification for Statistical Methods

The choice of these statistical techniques is consistent with the research objectives, ensuring robust and data-driven conclusions:

- Analysis describes key revenue trends and variables across states.
- Correlation and regression analysis explores the relationship between IGR and economic factors.
- Time series models project future model revenues to support key goals in the financial planning set.
- Analysis techniques categorise states based on IGR performance, allowing appropriate policy recommendations to be made.

5.6 Software and Tools

All statistical analysis will be performed using R programming language, in RStudio incorporating some R packages and libraries designed to facilitate analytical computations. The following table contains the statistical analysis and the accompanying libraries:

S/N	**Methodology**	**R Package**
1	Descriptive Statistics	dplyr, ggplot2
2	Regression Analysis	lm, car
3	Time Series Analysis	forecast, tseries
4	Machine Learning	randomForest, caret

Table 2. Statistical Analysis Approaches and R packages (Source: Author).

6. ETHICAL CONSIDERATIONS

6.1 Data Integrity and Transparency

Ensuring data integrity is fundamental to obtaining reliable and valid research results. As this study relied on secondary data sources, accuracy in data collection, cleaning and reporting was critical. Any changes to the dataset, such as handling missing values or adjusting for outliers, will be transparently documented to maintain reproducibility (Babbie, 2020). Additionally, multiple data sources (e.g. General Tax Council and Office for National Statistics reports) will be cross-referenced to minimise potential reporting bias (Zook et al., 2017).

6.2 Privacy, Confidentiality and Bias Prevention

While the dataset used in this study consists of publicly available government records, ethical concerns still exist. No personally identifiable information (PII) is involved; however, precautions will be taken to avoid bias or misinterpretation of state-level income data. Bias in the interpretation of data can lead to misleading conclusions that may influence policy recommendations. This study will use statistical methods to minimise subjective interpretations and ensure neutrality of results (Kitchin, 2014).

6.3 Compliance with Ethical Research Standards

This study adheres to established ethical principles for data analysis, including:

- FAIR (Findability, Accessibility, Interoperability and Reusability) principles: ensuring that data and analytical methods are transparent and accessible for review (Wilkinson et al., 2016).
- Academic integrity: ensuring the proper citation of all sources accurately, avoiding plagiarism and adhering to university research ethics guidelines (Resnick, 2020).

By adhering to ethical principles, this study maintains the highest standards of integrity, ensuring that research findings are reliable, reproducible and useful to policymakers and stakeholders.

7. CONCLUSION

This report provides a structured research design to analyse Internally Generated Revenue (IGR) trends in Nigerian states using data sets collected by the National Bureau of Statistics for a period of four years (2019 - 2023). The study aims to identify key factors influencing income formation and inequality across states by integrating statistical methods with a clearly defined research design.

A systematic study of data analysis, research design, error correction and statistical modelling comprehensively assesses revenue trends. The clearly defined research problem focuses on identifying key revenue patterns and economic and political determinants of state-level financial performance. The research design established a rigorous methodological framework that incorporated quantitative methods to generate meaningful insights while accounting for errors and biases inherent in secondary data sources.

In addition, a statistical analysis plan built around descriptive and inferential methods ensures that the conclusions drawn from the data are objective and reproducible. Ethical considerations, including data integrity, confidentiality and research transparency, enhance trust in the research while adhering to best practices for conscientious data analysis.

Ultimately, this study provides valuable insights that can be used to reform policies and improve revenue generation strategies at the state level. Using evidence-based insights, stakeholders, including policymakers and financial analysts, can develop targeted approaches to optimise tax efficiency, improve financial planning and ensure sustainable economic development across all states in Nigeria.

REFERENCES

1. Adeniran, A., & Osakede, K. (2021). State financial autonomy and revenue generation in Nigeria. Journal of Public Economics and Policy, 45(2), 112-135.
2. Babbie, E. (2020). The Practice of Social Research (15th ed.). Cengage Learning.
3. Bartlett, D., & Burton, D. (2016). Introduction to Data Analysis for Public Policy. Routledge.
4. Biemer, P. P., & Lyberg, L. E. (2003). Introduction to Survey Quality. Wiley.
5. Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (5th ed.). SAGE Publications.
6. Eze, C., & Ogundipe, A. (2020). Revenue performance and tax administration efficiency in Nigeria: An empirical review. African Economic Review, 37(1), 56-78.Few, S. (2017). Data Visualisation: Best Practices for Communicating Information. Analytics Press.
7. Few, S. (2017). Data Visualisation: Best Practices for Communicating Information. Analytics Press.
8. Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. SAGE Publications.
9. Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
10. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer.
11. Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. SAGE Publications.
12. Kumar, R. (2019). Research Methodology: A Step-by-Step Guide for Beginners (5th ed.). SAGE Publications.
13. Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data. Wiley.

14. Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.

15. Resnik, D. B. (2020). The Ethics of Research with Human Subjects. Springer.

16. Wickham, H., & Grolemund, G. (2017). R for Data Science. O'Reilly Media.

17. Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018.

18. Zook, M., et al. (2017). Ten simple rules for responsible big data research. PLoS Computational Biology, 13(3), e1005399.