SOUTH AFRICA: ANALYZING THE RELATIONSHIP BETWEEN CARBON EMISSIONS AND FOOD AFFORDABILITY (2017-2023)

NDTA631 - Data Analysis and Visualization

Group Assignment Report

Programme: Diploma in ICT

Module Code: NDTA631

Group Members:

Kamogelo Bantsheng - 202311455

Lebajoa Ramakatane - 202219950

Moepeng Bokang Khakhau - 202324860

Luyanda Nhlapho - 202213238

Ofentse Batlang – 202300977

Submission Date: 4 September 2025

EXECUTIVE SUMMARY

Our group undertook this comprehensive research project to examine the complex relationship between environmental impact and social welfare in South Africa. We analysed the correlation between carbon dioxide emissions and food affordability from 2017 to 2023 using robust data analysis methodologies and advanced visualization techniques with World Bank Open Data.

Our Methodological Approach: As a team, we developed a systematic pipeline including data acquisition, rigorous cleaning processes, statistical analysis using Python's scientific computing stack, SQLite database integration, and comprehensive visualizations. We emphasized reproducibility, scalability, and professional coding standards.

Key Findings: Our research revealed a weak negative correlation ($r = -0.28$) between CO_2 emissions and diet unaffordability, suggesting periods of higher economic activity showed slight association with improved food affordability. However, this relationship was not statistically significant ($p = 0.54$). During our study period, CO_2 emissions decreased by 8.7% while diet unaffordability increased by 1.5%, highlighting divergent trends.

Technical Achievements: Our group successfully implemented a complete data analysis pipeline including automated processing, statistical modelling, database management, advanced visualizations, and professional documentation with industry-standard practices.

Implications: From our discussions, we believe this research provides valuable insights for policymakers. Our findings suggest that economic growth alone may not directly translate to improved food security, emphasizing the need for targeted social interventions alongside economic development strategies.

1. INTRODUCTION AND RESEARCH CONTEXT

1.1 Background and Motivation

As a group, we recognized that South Africa represents a compelling case study for examining the relationship between economic development and social welfare. Through our initial discussions, we identified this as both an academic exercise and a real-world problem with significant policy implications.

1.2 Research Questions

Our team collaboratively developed three primary research questions:

1. What is the statistical relationship between CO2 emissions and food affordability in South Africa?
2. How have these variables evolved over the 2017-2023 period?
3. What methodological approaches are most effective for analysing socio-environmental data relationships?

1.3 Dataset Selection Rationale

CO2 Emissions Data: We chose total CO2 emissions including land-use change (OWID_CB_CO2) as a proxy for economic and industrial activity to capture overall economic activity.

Diet Affordability Data: We selected the percentage of population unable to afford a healthy diet (FAO_CAHD_7005) because it directly measures food security at the household level, which we felt was more meaningful than production metrics.

Timeframe Selection: Our choice of 2017-2023 represents the most recent complete data and covers significant economic and social changes in South Africa, including COVID-19 impacts and recovery.

1.4 Ethical Considerations

Throughout our project, we maintained high ethical standards. All data was sourced from publicly available, authoritative sources (World Bank) with appropriate citations. We maintained data integrity and worked exclusively with aggregated national-level data.

2. METHODOLOGY AND TECHNICAL IMPLEMENTATION

2.1 Technical Stack Selection

Our group carefully selected our technology stack to meet academic and industry standards:

- **Python 3.x:** For extensive data science ecosystem and reproducibility
- **Pandas:** For efficient data manipulation and cleaning
- **NumPy:** For numerical computations and array operations
- **Matplotlib/Seaborn:** For publication-quality visualizations
- **SQLite:** For lightweight, file-based database management
- **SciPy:** For advanced statistical testing and analysis

2.2 Data Acquisition and Validation Process

Our data acquisition involved multiple validation steps: direct download from World Bank, hash verification for integrity, cross-referencing with alternative sources, and manual metadata inspection.

2.3 Data Cleaning Decisions and Rationale

Handling Missing Values: We used inner join merge strategy focusing on overlapping period (2017-2023), ensuring data consistency, and avoiding bias from imputation. This created a clean dataset with seven complete annual records.

Data Transformations: We implemented Z-score normalization for outlier detection, 3-year moving averages for trend identification, percentage changes for year-over-year analysis, and categorical binning for comparative analysis.

2.4 Database Design and Statistical Methodology

Our database schema included primary keys, timestamp fields, and appropriate data types following professional practices. Our analytical approach was multi-faceted: descriptive statistics, Pearson correlation, linear regression, rolling statistics, and significance testing with emphasis on transparency and reproducibility.

3. DATA PREPARATION AND CLEANING PROCESS

3.1 Initial Data Assessment

When our group examined the raw datasets, we found:

- **CO2 Data:** 174 records spanning 1850-2023, with missing values in early years.
- **Diet Data:** eight records from 2017-2024 with complete data.
- **Structural Issues:** Both datasets in long format with extensive metadata columns

3.2 Filtering and Transformation Strategy

We implemented multi-stage filtering: country isolation (South Africa using REF_AREA = 'ZAF'), indicator selection (OWID_CB_CO2), and temporal alignment (2017-2023). Our transformation pipeline included structural changes from long to wide format, mathematical calculations for derived metrics, statistical processing, and categorical creation.

3.3 Validation Procedures

We implemented comprehensive validation: cross-field checks, summary statistics comparison, visual distribution inspection, correlation consistency checks, and unit verification. This resulted in a clean, analysis-ready dataset with complete transformation documentation.

4. ANALYTICAL APPROACH AND RESULTS

4.1 Statistical Analysis Framework

Our group employed a hierarchical approach: descriptive statistics for understanding, correlation analysis for relationships, trend analysis for temporal patterns, predictive modelling for insights, and validation through significance testing.

4.2 Key Findings

Descriptive Statistics Analysis:

- **CO2 Emissions (2017-2023):** Mean 429.41 million tonnes (±21.27 std dev), overall decrease of 8.7%, high volatility with significant COVID-19 impacts.
- **Diet Affordability:** Mean 60.97% unable to afford healthy diet (±0.64 std dev), overall increase of 1.5%, remarkably stable with minimal fluctuation.

Correlation Analysis Results:

- **Pearson Correlation Coefficient:** -0.28 (weak negative relationship)
- **P-value:** 0.54 (not statistically significant at α=0.05)
- **R-squared:** 0.08 (only 8% of variance explained)
- **Confidence Interval:** [-0.85, +0.49] at 95% confidence level

Our Interpretation: The weak negative correlation suggests higher CO2 emissions slightly associate with better diet affordability, but the relationship is not statistically significant and explains very little variation.

4.3 Time Series and Category Analysis

Our temporal analysis showed divergent trends: CO2 emissions showed decreasing trend with high volatility, while diet affordability showed increasing trend with low volatility. Category analysis by emission levels revealed minimal variation (61.2%, 60.8%, 60.9% diet unaffordability for low, medium, high emission years respectively), further supporting weak relationships.

5. VISUALIZATION STRATEGY AND TECHNICAL CHALLENGES

5.1 Visualization Design Philosophy

Our approach emphasized clarity for immediate comprehension, accuracy without distortion, comparability across variables and time, and professional aesthetics suitable for academic publication.

5.2 Key Visualizations

- **Dual-Axis Time Series:** Showed opposing trends between decreasing emissions and increasing diet issues.
- **Scatter Plot with Regression:** Visual representation of weak negative correlation with confidence intervals.
- **Year-over-Year Change Comparison:** Highlighted volatility differences between variables
- **Distribution Box Plots:** Showed different variability patterns and central tendencies.

5.3 Technical Challenges and Solutions

Data Limitations: Brief time series limited longitudinal analysis; we compensated with intensive cross-sectional analysis and multiple validation techniques.

Scale Disparities: Vastly different scales required dual axis plotting and derived metrics for comparability.

Database Integration: Ensured seamless Pandas-SQLite integration through comprehensive error handling and parameterized queries.

Statistical Validation: Small sample size addressed through normality testing, non-parametric validation, and bootstrapping for confidence intervals.

6. CONCLUSION AND IMPLICATIONS

6.1 Research Questions Addressed

Our group successfully addressed all research questions:

1. **Relationship:** Weak negative correlation (r = -0.28) but not statistically significant
2. **Trends:** Divergent trends - emissions decreasing, diet affordability worsening
3. **Methodology:** Comprehensive pipeline from data acquisition to insight generation demonstrated

6.2 Key Contributions and Policy Implications

This research contributes methodologically through a complete, reproducible pipeline; technically through tool integration; substantively through insights into South Africa's socio-environmental dynamics; and pedagogically through industry-standard practice demonstration.

Our findings suggest several policy considerations: economic growth alone may not improve food security, targeted social interventions are necessary alongside economic development, environmental and social indicators should be monitored simultaneously, and short-term economic fluctuations do not necessarily impact food affordability.

6.3 Limitations and Future Research

We acknowledge limitations including brief time series, aggregate national data masking regional variations, unaccounted confounding variables, and potential COVID-19 distortions. Future research should include longer time series analysis, regional-level analysis, additional indicators, multivariate modelling, and comparative analysis with other emerging economies.

6.4 Final Reflections

This project demonstrated to our group the power of comprehensive data analysis for understanding complex socio-economic issues. The technical skills we developed are directly transferable to real-world data science challenges. We successfully balanced academic rigor with practical implementation, producing both insightful findings and a robust technical foundation for future research.

7. REFERENCES

World Bank. (2025). World Development Indicators. Washington, DC: World Bank. https://data.worldbank.org/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference.

Virtanen, P., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.