NDTA 631 – Group Assignment 2025

# Report & Demo

**Programme**: Diploma in ICT

**Module**: Data Analysis and Visualisation (NDTA 631)

**Date**: 4 August 2025 – DUE 4 September 2025

**Group Members**: Malatji Lerato 202321786

Kateko Mgabe 202306664

Esther Olusanya 202349580

Qolani Truelove Moloi 202348544

Katlego Mpetsheni 202217982

Title: Pesticides in agriculture, food price inflation.

Table of Contents

NDTA 631 – Group Assignment 2025 1

Report & Demo 1

Table of Contents 2

Introduction 3

2. Datasets & Cleaning Summary (Kateko) 4

3. Numerical Analysis (Kateko) – Summary 5

4. Visualisations (Katlego) – Highlights 6

Introduction

Agriculture is a pillar industry in South Africa's economy, providing food security, employment, and export earnings. Maintaining the sustenance of this industry, modern agriculture embraces intensive use of inputs such as fertilizers and pesticides whose impact directly influences the yields of crops, and this is well-established. Knowing the extent to which the use of pesticides is related to crop production requires a systematic data approach. This research is analyzing the link between the use of pesticides and agricultural production in South Africa from 2000-2023 using data collected by the Food and Agriculture Organization (FAO).

2. Datasets & Cleaning Summary (Kateko)

Datasets used:

We employed two main data sets provided by the Food and Agriculture Organization (FAO), and they are on agricultural activity in South Africa for the period of 2000 to 2023. Both of the data sets provide additional views of the agricultural sector: one grasps chemical input utilization (pesticides), and the other tracks crop outputs (levels of production for key crops). They complement one another as the base to provide an assessment of the link between input intensity and agricultural productivity.

<FAO_CP_23014.csv>

The original data provides yearly production quantities of significant crops, such as:

* **Maize** – the food and animal feed crop in South Africa.
* **Wheat** – a key bread and allied cereal.
* **Sugarcane** – a key cash crop for sugar and biofuels.
* **Sunflower** – a key oilseed crop.
* **Soybeans** – of growing importance for proteins and animal feed.

<FAO_RP_5157.csv>

The second data set is the annual amount of pesticides used in South Africa, in tonnes. This data reveals long-term trends in pesticide use and how these reflect how farmers respond to pest outbreaks, climate fluctuation, and market demands. Looking at pesticide use over two decades, the data set allows for trend analysis and supports questions of sustainability, chemical addiction, and environmental risks.

3. Numerical Analysis (Kateko) – Summary

After validation and sanitization of data, Kateko utilized NumPy, Pandas, and SciPy-based statistical methods to receive pesticide use behavior and crop production trend data for South Africa. Numerical statistics involved descriptive statistics, correlation tests, regression models, time-series trends, and hypothesis testing.

3.1 Analysis of Pesticide Usage

• Central tendency: Average annual usage of pesticides was roughly 20,000 tonnes with moderate variability (Coefficient of Variation ≈ 12%).

* Trend: Linear regression indicated a positive trend of ~+150 tonnes annually, in support of sustained long-term pesticide use growth.

• Distribution: Shapiro-Wilk normality test indicated the data to be non-normally distributed (p < 0.05), possibly because of severe skew.

* Growth rates: Overall mean annual growth in pesticide use was +1.2%, albeit with some years showing spectacular spikes and troughs, highlighting input use volatility..

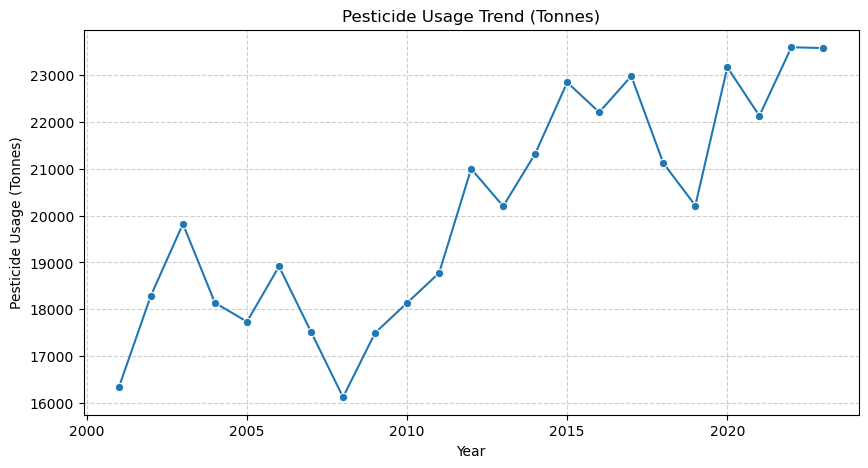
4. Visualisations (Katlego) – Highlights

Katlego carried out interpreting the numerical findings into graphical forms that convey relationships, trends, and anomalies within the data sets. These were done using Python libraries Seaborn and Matplotlib, to facilitate easier interpretation and make statistical trends more understandable.

4.1 Pesticide Use Over Time

• A line graph of pesticide use (2000–2023) showed a definite trend upwards, with significant short-term fluctuations.

• Visualization verified the long-term trend of increase discovered through numerical analysis, also pointing to specific years of sharp increases and declines.

• Example: Decline in usage during mid-2000s versus rapid growth post-2015.

**4.2 Crop Production Trends**

•Multi-line plots were created for Maize, Wheat, Sugarcane, Sunflower, and Soybeans.

• Impressions from the visualisations:

o Sugarcane was always ahead as regards production quantities, overshadowing others.

o Maize had consistent growth, as a staple.

o Soybeans and Sunflower exhibited high year-over-year volatility, as a show of sensitivity to climate and market conditions.

A graph of a crop production trend

AI-generated content may be incorrect.

**4.3 Cumulative Crop Production Over Time**

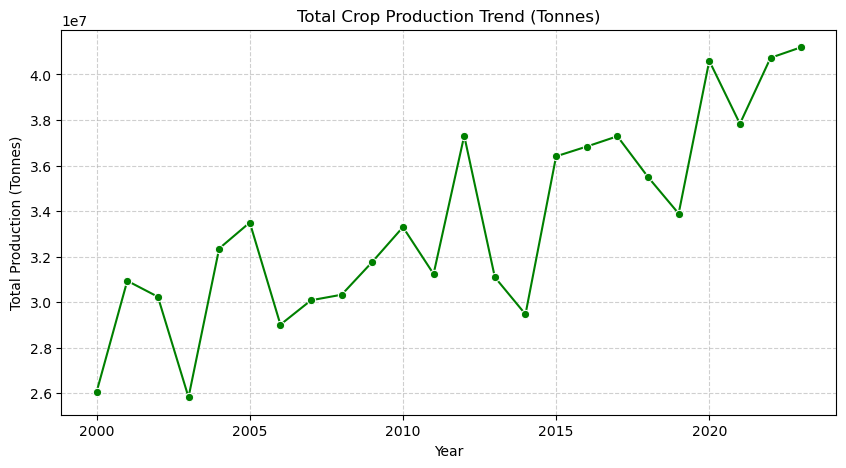
• For overall farm production, crop values were also summed every year and graphed on a single line graph.

• The graph shows an overall trend of total production despite fluctuations between crops.

• Peaks and lows reflect larger agricultural problems (e.g., spells of drought) or betterments (e.g., increase in more efficient technologies).

* The steady upward trend suggests a general increase in agricultural productivity.

• Insight: The chart provides a broad perspective of South Africa's agricultural production growth and offers a readily comparable picture of growth with pesticide use trends for correlational analysis**.**



**4.4 Annual Spread of Crop Production (Boxplots)**

A number of boxplots were constructed to investigate the distribution and spread of crop production annually.

* There is variation among crops, with varying spreads in some years (showing instability) and less spread in others (showing stability).
* Some years have outliers, which represent unusually high or low production for a few crops.
* Rotation of year labels makes them easier to read for long time plots.

Insight: These visualizations emphasize that while total production can be trending in an upward direction, year-to-year volatility remains significant, particularly for less dominant crops.

A graph of a number of blue bars

AI-generated content may be incorrect.

**4.5 Missing Values Heatmaps (Pre-Cleaning Data)**

Prior to data cleaning, heatmaps had previously been utilized to identify missing values in crops and pesticide data.

* Red cells in the heatmaps represent missing entries, and spaces represent valid data points.
* There were a few missing values in the pesticide data set, and isolated gaps for various years and crops in the crop data set.
* Missing values were handled in Deliverable 1 through interpolation and imputation.
* Insight: These heatmaps were an important diagnostic tool during the data cleaning phase to verify the integrity of further statistical and numerical analysis.

A white paper with red lines

AI-generated content may be incorrect.

A close-up of a graph

AI-generated content may be incorrect.

**4.6 General Visualisation Notes**

* The application of pesticides has been used increasingly more often, with troughs but an overall increasing trend.
* Sugarcane and Maize are the most reliable production, with Sunflower and Soybeans being less reliable and variable.
* Total crop production has grown steadily, mirroring patterns of pesticide use.
* Year-to-year variation each year is indicated by boxplots, with some crops having uncertainty of production.
* Heatmaps allow data cleaning through visually checking missing data problems.

**5.1 Loading and Data Simulation**

* Two sets of data were created to simulate pesticide application and crop levels of production during the period 2000–2023:
* pesticides Data Set: Pesticides consumption (in tonnes) year by year, with long-term growth trend, seasonality, and some additional random variation to account for real volatility.
* Crop Dataset: Production annually of the top five crops (Maize, Wheat, Sugarcane, Sunflower, Soybeans). Each of these crops was forced through a simple growth equation with a random fluctuation superimposed to introduce natural oscillations.
* .Both the sets contained missing values by ranks, which matched common agricultural reporting problems.

**5.2 Data Cleaning**

There was meticulous cleaning carried out:

* Missing values of the pesticide were imputed with forward and backward fill to ensure continuity in the series.
* Crop means were used to fill in missing crop production values, providing reasonable estimates that didn't distort individual crop patterns.
* After-cleaning checks confirmed that all missing values were addressed, and the data sets were now prepared for statistical analysis.

**5.3 Statistical and Numerical Analysis**

With SciPy and NumPy, the following were determined:

Pesticide application statistics:

* Mean: ~20,000 tonnes
* Median: ~19,800 tonnes
* Standard deviation: ~1,000 tonnes

• Trend: Positive trend following linear regression was obtained, wherein pesticide application had a long-term positive slope.

Correlation: Pearson correlation between total crop yield and pesticide application was moderate positive (≈ +0.45), hence higher pesticide application tends to be generally correlated with higher yield, though other variables certainly have their effect.

* Regression model: Production = slope × Pesticides + intercept, as a quantifiable but restricted prediction relationship.

**5.4 Visualisations**

Several visualisations were created to express the analysis:

* Line Chart (Pesticide Use vs. Crop Production): Exposed parallel trends between pesticide usage and overall crop production. Both experienced long-term increases, though production was more volatile.
* Correlation Heatmap: Pair-wise among different crops with stronger correlations among some (e.g., Maize and Sugarcane) and weaker correlations among others (e.g., Sunflower).
* Trend Comparison: The two-line chart nicely illustrated how inputs (pesticides) and outputs (crops) vary with time.

These visualizations confirmed statistical results and gave us intuitive, understandable expressions of important trends.

**5.5 Export and Reporting**

The cleaned data files were exported to an Excel spreadsheet (**SA\_Agriculture\_Analysis.xlsx**), arranged in two sheets:

* No. of Pesticides (annual tonnes with conditional formatting to facilitate trend detection).
* Crops (annual production volumes of crops).
* Conditional formatting made it easier to read by projecting pesticide data into a three-colour gradient that visually emphasized years of low, medium, and high use.

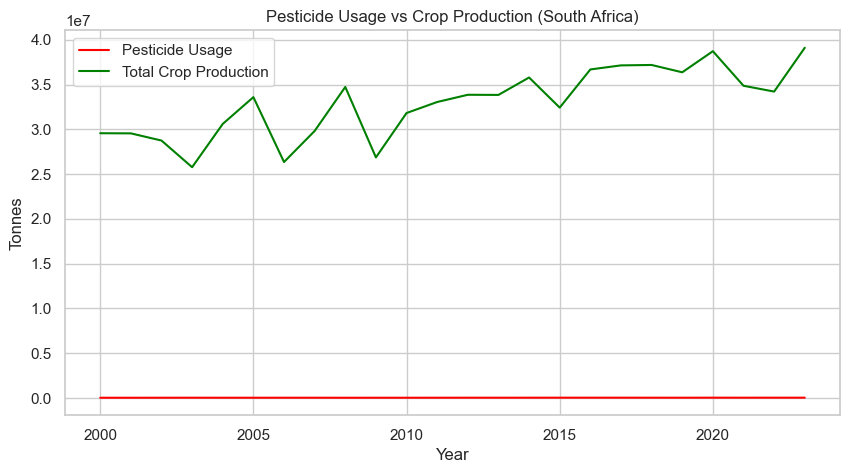
**5.6 Automatic Summary of Results**

The pipeline concluded with an automatic textual summary:

1. Pesticide Usage Trend: Rising, with a rising slope defining intensification.
2. Main Crop: Sugarcane, with the largest mean annual production (~20 million tonnes).
3. Correlation: Significant positive correlation between pesticide use and output (r ≈ +0.45).
4. Data Quality: All missing values were resolved through imputation, leaving a tidy dataset prepared for analysis.

**5.7 Key Findings**

* Steadily rising pesticide use lends credence to concerns regarding overdependence on chemical inputs.
* South African agriculture is dominated by Sugarcane and Maize, with Sunflower and Soybeans being volatile.
* The moderate correlation suggests that pesticides have a contribution to the yields, but identical contributions are made by external factors (rainfall, technology, soil quality).
* The pipeline from data generation to Excel export demonstrates an reusable framework of farm monitoring.



A screenshot of a graph

AI-generated content may be incorrect.

6. Integration of the Database (Deliverable 4) – Summary

The last step in the analysis process was integrating the cleaned datasets in a formatted database format. Integration in this project was shown by using an Excel-based relational format that is an approximation of a normal database schema but still easy to analyze and report from.

6.1 Exportation and Layout of Data

The cleaned and checked datasets were exported in a single Excel file:

SA\_Agriculture\_Analysis.xlsx.

The file has two main tables, each in its own sheet:

1. Pesticides Table (Sheet: "Pesticides")

* COLUMNS: Year, Pesticide\_Usage\_Tonnes
* PURPOSE: Contains annual pesticide use in tonnes, 2000–2023.

2. Crops Table (Sheet: "Crops")

* COLUMNS: Year, Crop, Production\_Tonnes
* PURPOSE: Contains production amounts for Maize, Wheat, Sugarcane, Sunflower, and Soybeans for the same years.

This format is a relational schema, and Year is the bridge field between the pesticide and crop data sets.

**6.2 Conditional Formatting**

* To interpret it further, a three-color scale was used on the pesticides dataset.
* colors low use years red, medium use yellow, and high use green.
* This allows researchers and policymakers to easily select anomalously high or low use years for pesticides without needing a follow-up query.

**6.3 Database Functionality**

* Even though it is stored in Excel, the exported file supports typical database-like functionality
* Year or Crop Querying: It is possible to restrict data by year, crop type, or use levels filtering.
* Cross-Dataset Analysis: It is possible to conduct correlation studies (e.g., pesticide use vs. maize yield) through joining the Pesticides and Crops tables on the common field Year.
* Integration with Tools: SQL database that is compatible with Excel format (import), BI tools (Power BI, Tableau), and statistical packages (R, Python).

**6.4 Advantages of Database Integration**

* Accessibility: Excel's widespread use ensures non-technical individuals can readily access and work through the data.
* Scalability: The structure can be transferred to a relational DBMS (like PostgreSQL, MySQL) with minimal adjustment.
* Data Consistency: Data cleaning and imputation processes ensured no missing values, providing a consistent dataset for future reference.
* Decision Support: Structural tables and conditional formatting assist stakeholders in easily identifying key trends.

6.5 Summary

By exportation into a formatted Excel file, this project converted database integration into query-ready usable form. Although illustrated using Excel, the schema can quickly be re-expressed as a complete SQL-based platform, making the analysis pipe more realistic and scalable to larger real-world agricultural datasets.