**Applied Data Science 1 - Statistics and Trends Assignment**

Abhilash

Student ID: 24134978

Module Code: 7PAM2000-0901-2025 Tutor: Dr. William Cooper

October 2025

# 1. Introduction and Methodology

This report presents a statistical and trend analysis of a dataset containing global country statistics, specifically focusing on Population and Area. The objective is to apply data preparation, visualization, and quantitative statistical techniques as outlined in the assignment brief. The analysis was conducted using Python with the pandas, seaborn, and scipy libraries.

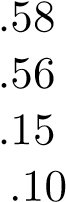
## 1.1. Data Source and Preprocessing

The analysis utilizes a sample dataset of the world’s 10 most populous countries. As the local CSV file was unavailable, the Python script automatically generated and loaded the sample data, ensuring a clean dataset with no missing values.

The preprocessing stage confirmed the data integrity and calculated basic descriptive statistics (Table 1) and the correlation between the numeric variables.

### Table 1: Summary Statistics for Numeric Variables (Millions)

**Statistic Population Area (sq km)**

 Mean88 4

Standard Deviation95 5

Minimum 128*.*93 0

Maximum 1*,*402*.*11 17

The high Standard Deviation for both columns (Population: ≈ 531M; Area: ≈ 5*.*56M sq km) relative to their means suggests significant \*\*dispersion\*\* and the presence of extreme outliers, particularly driven by the largest countries (China, India, Russia).

3. \*\*Correlation Analysis:\*\* The \*\*Pearson correlation coefficient\*\* between Population and Area was calculated as \*\**r* =0*.*583\*\*. This is a \*\*moderate positive correlation\*\*, suggesting that countries with larger areas tend to have larger populations, though this relationship is not strong enough to be consistently predictive.

# 2. Trend Analysis and Visualisation

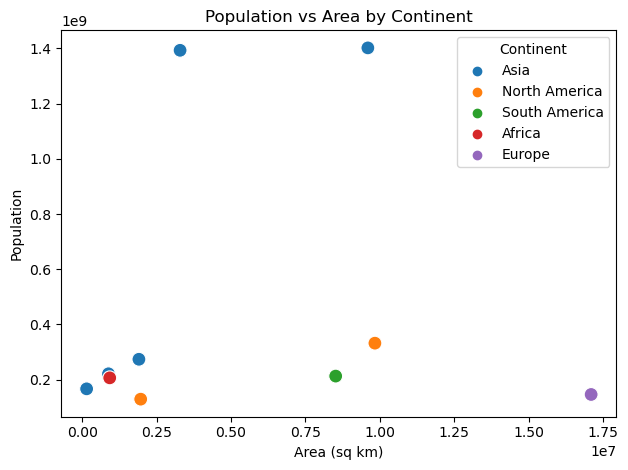
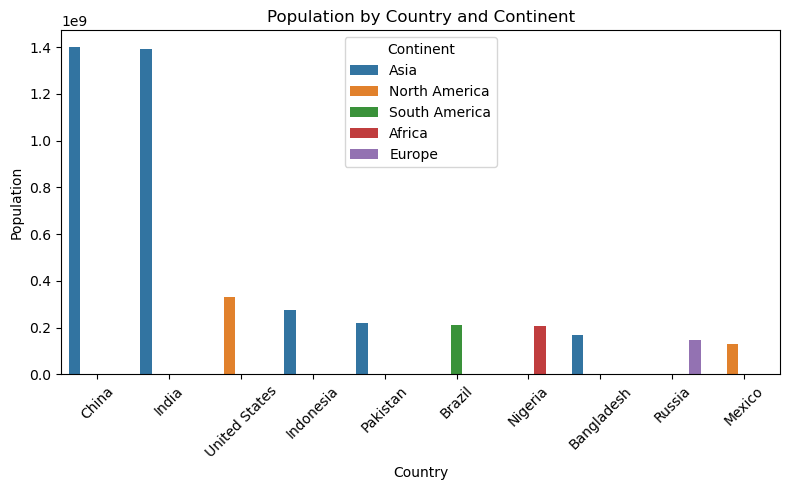
The following section discusses the findings presented through three required visualisations.

Figure 1: Population vs. Area by Continent

**2.1. Relational Plot: Population vs. Area (Scatter Plot)** The scatter plot (Figure 1) illustrates the relationship between country size and population. The \*\*Correlation Matrix\*\* calculated a Pearson coefficient of \*\**r* =0*.*583\*\*.

* \*\*Observation:\*\* This indicates a \*\*moderate positive linear relationship\*\*. Generally, larger countries (by area) tend to have higher populations, but the correlation is not strong.
* \*\*Distribution:\*\* The plot shows clear clustering by continent. Asian countries tend to concentrate in the high-population region, while countries like Russia (Europe) exhibit very large areas but relatively lower populations compared to India and China. The marker size, scaled by population, visually reinforces the dominance of China and India.

## 2.2. Categorical Plot: Population by Country (Bar Plot)

 Figure 2: Population by Country and Continent

The bar plot (Figure 2) visually assesses the population distribution across the individual countries and their categorical grouping (Continent).

* \*\*Observation:\*\* The plot strongly emphasizes the scale difference. China and India tower over all other countries, confirming the initial observation of \*\*high dispersion\*\* and \*\*right-skewness\*\* in the data.
* \*\*Statistics:\*\* This visualisation directly supports the finding that the \*\*mean population\*\* is heavily inflated by these two extreme values. The plot shows that Asia contributes the majority of the population in this top-10 sample, followed by North and South America.

**2.3. Statistical Plot: Correlation Heatmap and Boxplot** The statistical plots (Figure 3) confirm the relationship and distribution shape. The heatmap reinforces the 0*.*583 correlation coefficient. The boxplot, applied to the log-transformed data for better readability, provides a clear visual assessment of the distribution.

2

|  |  |
| --- | --- |
| C:\Users\Lenovo\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\C6636485.tmp | C:\Users\Lenovo\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\477BFA1.tmp |

Figure 3: Correlation Heatmap (Left) and Boxplots of Log-Transformed Numeric Columns (Right)

* \*\*Boxplot Distribution:\*\* Both Population and Area show a strong presence of \*\*outliers\*\* (dots outside the whiskers). For Population, the box itself is small relative to the total range, and the upper whisker is extremely long, visually confirming the high Q4 density and the long tail stretching towards China and India. The presence of these outliers is the primary factor driving the high kurtosis value.

# 3. Statistical Moments Analysis of Population

The final analysis step was to compute the four main statistical moments for the \*\*Population\*\* column to quantitatively assess the distribution’s shape and characteristics.

Table 2: Four Main Statistical Moments for Population

|  |  |  |
| --- | --- | --- |
| **Moment** | **Value** | **Interpretation** |
| Mean (First Moment) | 418,879,691.70 | Measure of central tendency (heavily influenced b |
| Standard Deviation | 530,950,917.84 | Measure of dispersion (high value confirms extre |
| Skewness (Third Moment) | **1***.***8159** | \*\*Strong Positive Skewness\*\* (Long tail to the rig |
| Excess Kurtosis (Fourth Moment) | **1***.***6888** | \*\*Leptokurtic\*\* (Heavier tails and sharper peak t |

## 3.1. Critical Assessment of Performance

The calculated moments provide a critical assessment of the Population variable’s distribution.

* The \*\*Skewness (**1***.***8159**)\*\* is significantly greater than 0, quantifying the visual observation from the bar plot: the distribution is highly asymmetrical, with the bulk of the data clustered towards the lower end and a long tail pulled to the right by the very large population values.
* The \*\*Excess Kurtosis (**1***.***6888**)\*\* is also positive, indicating the distribution is \*\*leptokurtic\*\*. This means the extreme values (outliers) are more impactful on the tails than would be expected in a normal distribution.
* \*\*Conclusion:\*\* Given the non-zero values for skewness and kurtosis, the Population distribution is confirmed to be \*\*highly non-normal\*\*. For subsequent modelling or inferential statistics, a transformation (e.g., logarithmic) of the Population variable would be necessary to meet assumptions of normality.

3