A compact study of analysing the trend in World population from 1951 to 2023

By TIASA MALITYA, Id – 24005709
https://github.com/Tiasa24/ASSIGHNMENT5709ADS2

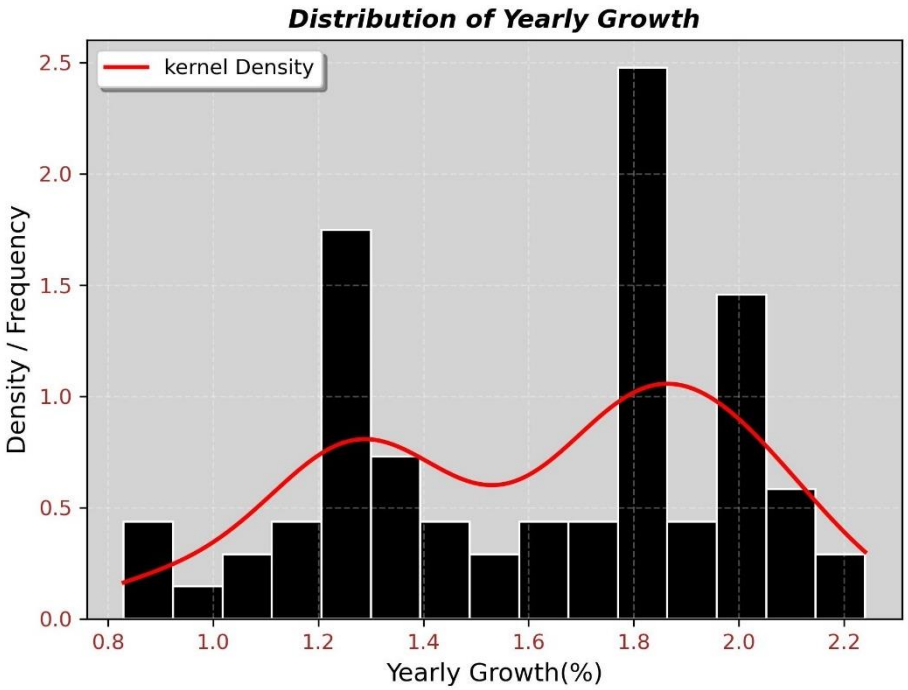
Introduction

The dataset of World Population Growth reflects the four main features which are total population, yearly growth in percentage, yearly increased number and population density measured per square kilometre on worldwide population behaviour over the year 1951 to 2023. This report illustrates the distribution of the attribute, analyses the best line fit to know the trend of population density over the time and explains the use of elbow method to determine the perfect number of clusters to segment the dataset by Kmeans clustering.

The dataset which has been chosen for the study is "*world.csv*".

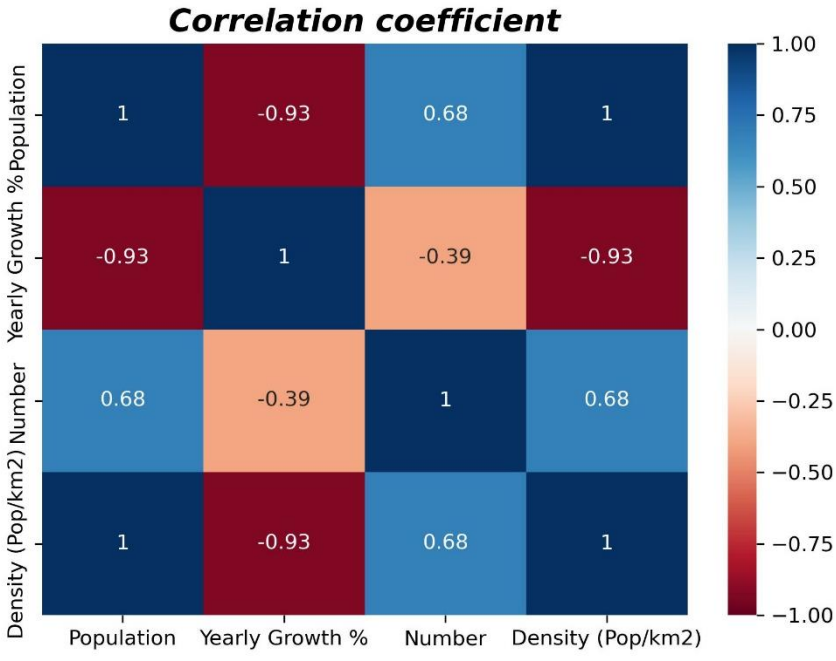
Understanding and Visualisation

Plot 1: Distribution of Yearly Growth % using Kernel Density



First plot explores how yearly growth in percentage is distributed over the years. It is showing a histogram with an overlaid kernel density estimate (KDE) which reveals a slightly left-skewed distribution (**skewness: -0.31**) with a **kurtosis** of **-1.02** The histogram displays a flatter and spread-out distribution compared to a normal curve. This suggests that most years fall within a typical range of growth rates (*1.3, 1.4, 1.8, 2.0*) showing consistent trends over time. However, there are a few years where the growth rate drops significantly lower than usual (*1.0*) which are called outliers. Kernel density of the data shows a density of values around the mean with a long tail going toward smaller growth rates (*1.0, 1.1* etc). The years with notably lower growth rate indicate specific tragedy like war, pandemic and other issues that affects population growth in those years.

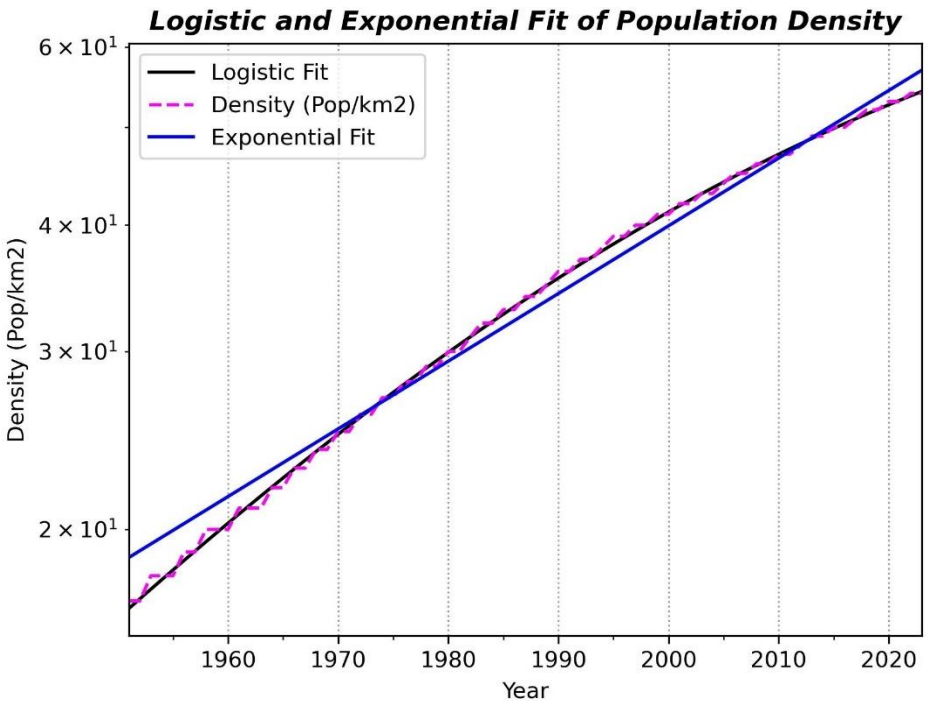
Plot 2: Heatmap showing correlation coefficients between the variables



After building insight about the variable Yearly growth from the Histogram, to know more about the statistical aspect of all four variables, A heatmap is created to visualize the correlations among variables like population, yearly growth, population density, and the number of people added each year. The correlation coefficient of each two key variables which ranges from -1 to 1, gives the idea of linearity and non-linearity among the measurements. Heatmap helps to understand which variables are good for fitting

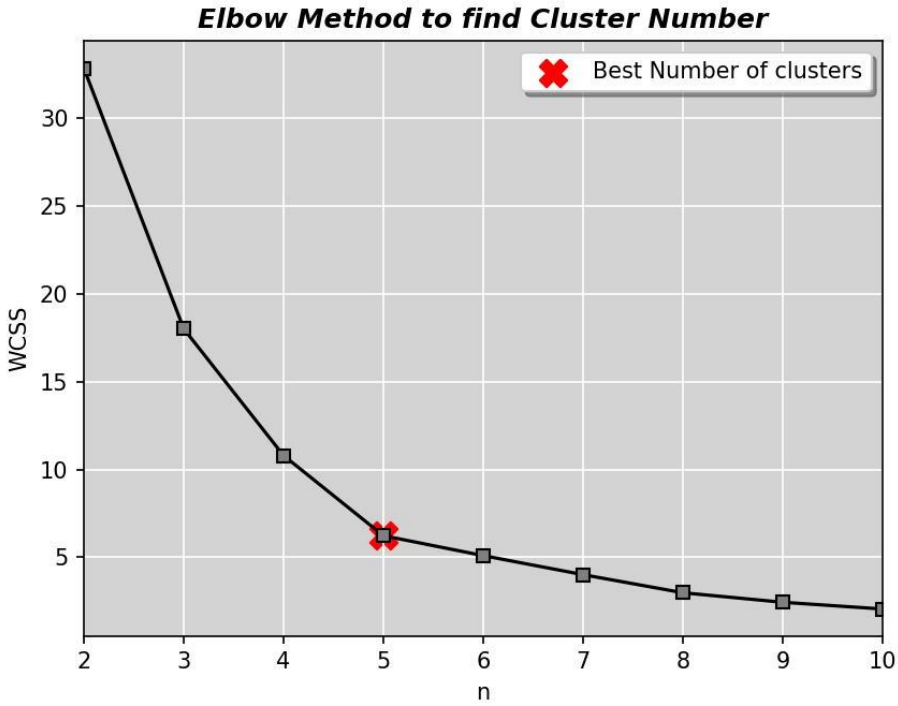
and clustering. From the Heatmap of World Population dataset, it is clear that population size and population density is highly correlated with coefficient value 1 which suggests that areas with higher populations experience more concentrated growth. Whereas Yearly growth rate with population density is negatively correlated (-0.93), Yearly Growth rate and yearly number of people added is nearly uncorrelated as coefficient value close to 0 (-0.39). These relationships were crucial in determining the parameters for clustering and fitting process by highlighting which variables are most and less interconnected.

Plot 3: Line plot to show the best fit of population density



The plot 3 shows the line graph of Population density with respect to years from 1950 to 2023. Previously shown Heatmap gives the idea to perform fitting of Population density. To get the best fit for Population density (nonlinear nature data) two fittings method logistic and exponential are applied. First the exponential fit is developed because population density often increases rapidly in the early stage like exponential growth due to urbanization and development. However, as time progresses, growth rates naturally slow due to different economic, environmental factor. On the other hand, the logistic fit matches with the observed data point of density much better providing more perfect fitting and produces a significantly lower uncertainties/ error compared to the exponential model. Fitting is also useful for finding prediction value.

Plot 4: Elbow plot



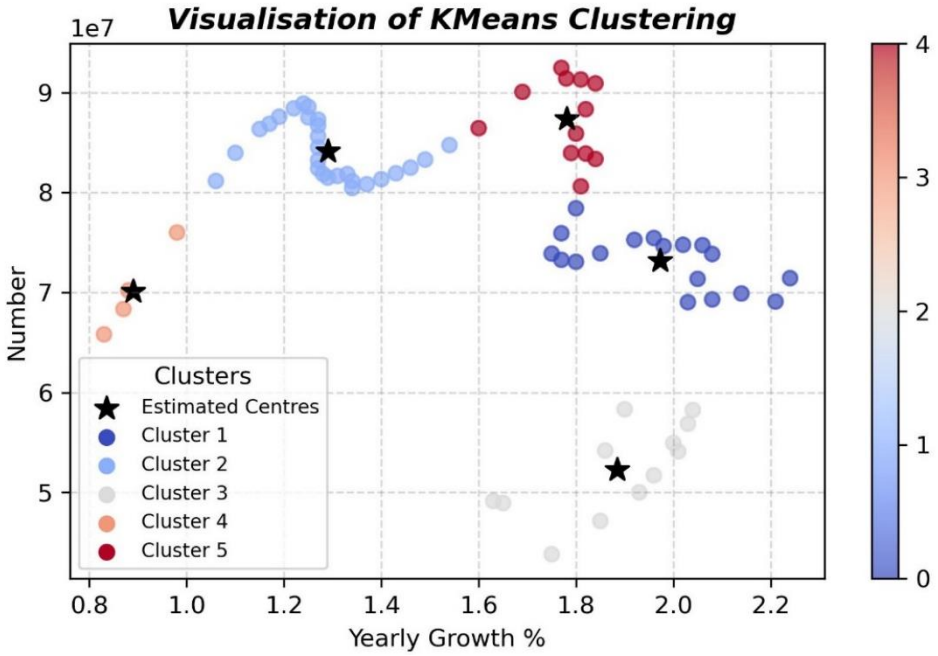
To find out the natural structures within the dataset by grouping data points based on their similarities Clustering method helps to analyse the pattern and relations in data. Creation of Elbow plot visualises the computation of the best number of clusters for K-means clustering. The elbow plot shows how the within-cluster sum of squares (WCSS) changes as the number of clusters increases. The method computes the Sum of the Squared Errors (SSE). This is also referred to as inertia which is the total separation between each sample and its allocated cluster centre. Here

WCSS = $\sum(x_i - a_i)^2$

$x_i : Sample\ points$ & a_i : $Cluster\ Centre$

The point at which WCSS start getting diminishing as more clusters are added creates an *elbow* in the graph. This *elbow* tells us the best number of clusters is 5. In some dataset, cluster number calculated from silhouette score and in elbow plot happens to be different.

Plot 5: K-means Clustering between yearly growth (%) and population increase (number)



After getting the best number of clusters from elbow plot, K-Means clustering is applied to understand the segmentation in the dataset in the basis of yearly growth (%) and population increase (number). Also, in heatmap, it is explained that yearly growth and Number columns are not likely correlated which allows to perform clustering between them. This technique groups years with similar growth and population number added, providing a structured way to identify the similar trends in the dataset. For example, Cluster 3 which groups the data with higher growth rate **(1.6 – 2.0)** and lowest number of populations added period. Cluster 5 represents a steady period of population growth **(1.6-1.8)** close to the average yearly growth rate 1.61 of the dataset and maximum number of populations added. In addition, the estimated centres are shown of each cluster to understand the clusters.

Basic Statistics of all the features of Dataset

The dataset's descriptive statistics reveal a steady global population increase, with the mean population at 5125103969.08, ranging from a minimum of 2543130380.00 to a maximum of 8045311447.00 with Yearly growth averages 1.61% while population density shows moderate variability 11.37. The yearly population number increase with a mean of 75972456.03 maintaining a consistent growth, while deviations in growth rates which is 0.36 shows the changes in the world such as economic, social developments. These statistics provide a solid base for exploring trends, clustering, and fitting techniques.

Conclusion

By using a combination of visualizations and analyses, this report concludes how different techniques can work together to show the basic trend in global population. The histogram with kernel density revealed the distribution of yearly growth rates, while the heatmap highlighted correlations between variables which helps to choose the variable to perform fitting and clustering. The logistic fit for population density demonstrated how growth slows over time, and clustering identified distinct growth patterns among the data.

The elbow method showed that five clusters are the best choice, with one cluster closely matching the average yearly growth rate. These findings provide deep understanding of different useful information for urban planning, resource management. This study highlights how statistical tools and clustering fitting can be applied to understand and address global challenges effectively.