# DATA SCIENCE ASSIGNMENT 03

## Abstract.

This study looks at the clustering strategy, which is very important in data science when you need to group data in a data set. We used acceptable global urban population data from the World Bank for this project. We only checked the last two years for all countries in the world. This time, we'll show you a few new libraries, including the Python third, which is comparable to SK and is built in for data science-related terminology. Learn about preprocessing for data normalization and Science Pi, where we import the curve fit technique, and cluster, where we use the same library as SK learns k mean cluster. Some of the built-in libraries from previous jobs can also be used. Is it all OK if we go over the findings one by one and show them to you on this poster? We not only got the results, but we also used data visualization to develop a data visualization for more precise learning. We'll now see how many clusters we can create using solely this clustering method from this dataset. On the left side, we explain data in the same way. For the years 2016, 2017, this dataset provides an explanation of the entire dataset.

https://github.com/Ahtshamkarim/ASSIGNMENT3-DATASCIENCE

## Normalization.

The process of turning real-valued numeric properties into a 0 to 1 range is known as normalization. In machine learning, data normalization is used to make model training less sensitive to feature scale. As a result, our model will be able to converge to more accurate weights. Data is normalized using the normalize function from the Python preprocessing package. It converts an array of numbers to a scale of 0 to 1. We divide the result by the range after subtracting the minimal value from each item. The difference between the highest and lowest value is known as range.

## Curve fit method

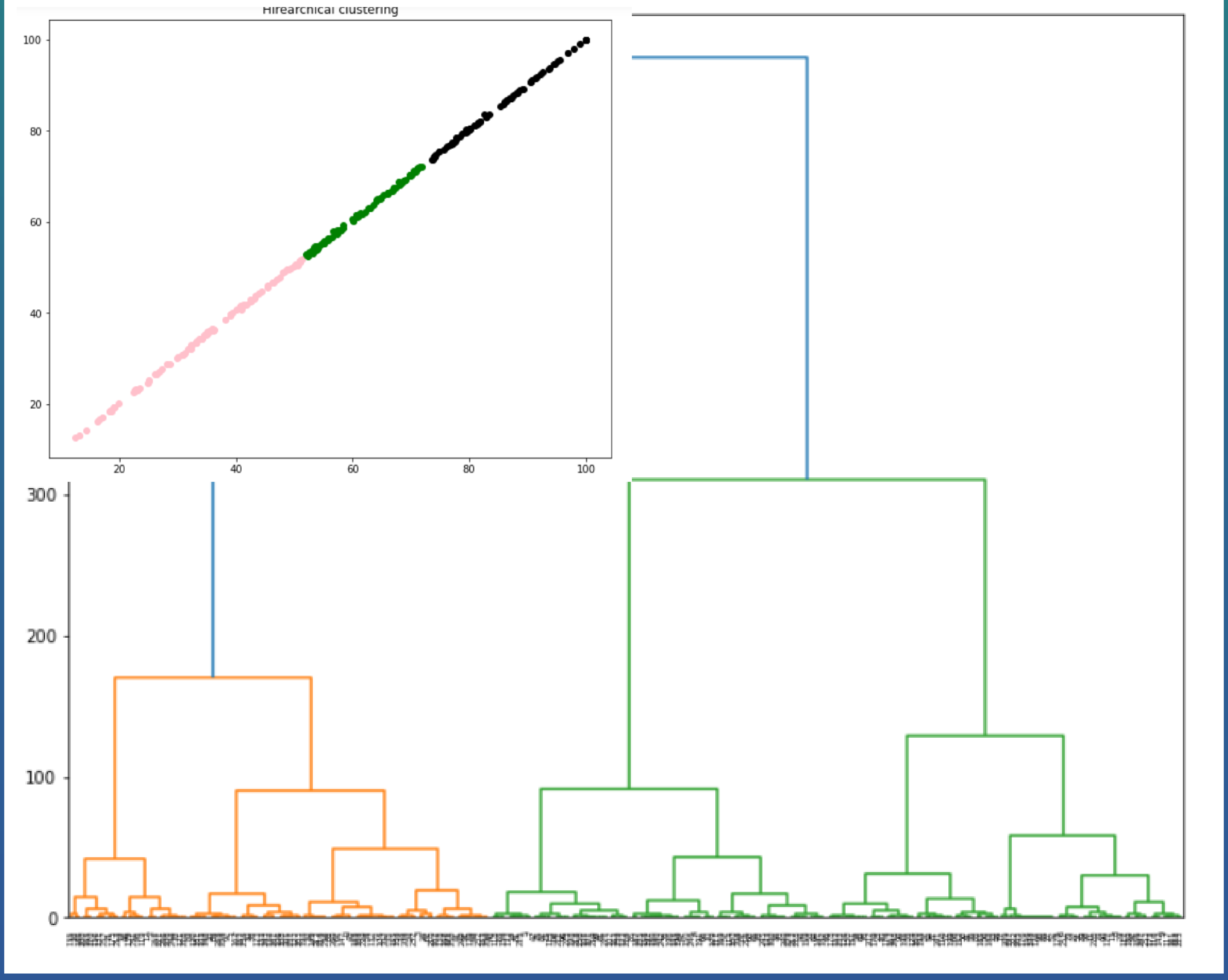
Curve fitting is a type of optimization in which the optimum set of parameters for a given function that best fits a set of data are determined. In contrast to supervised learning, curve fitting requires the declaration of a function to convert input instances to outputs. Built-in libraries are also utilized to fit the dataset using the curve fit method. The mapping function, sometimes called the basis function, can take whatever shape you like, including a straight line (linear regression), a curved line (polynomial regression), and so on. This provides you complete freedom and control over the curve's shape, and an optimization technique is used to get the function's precise ideal parameters.

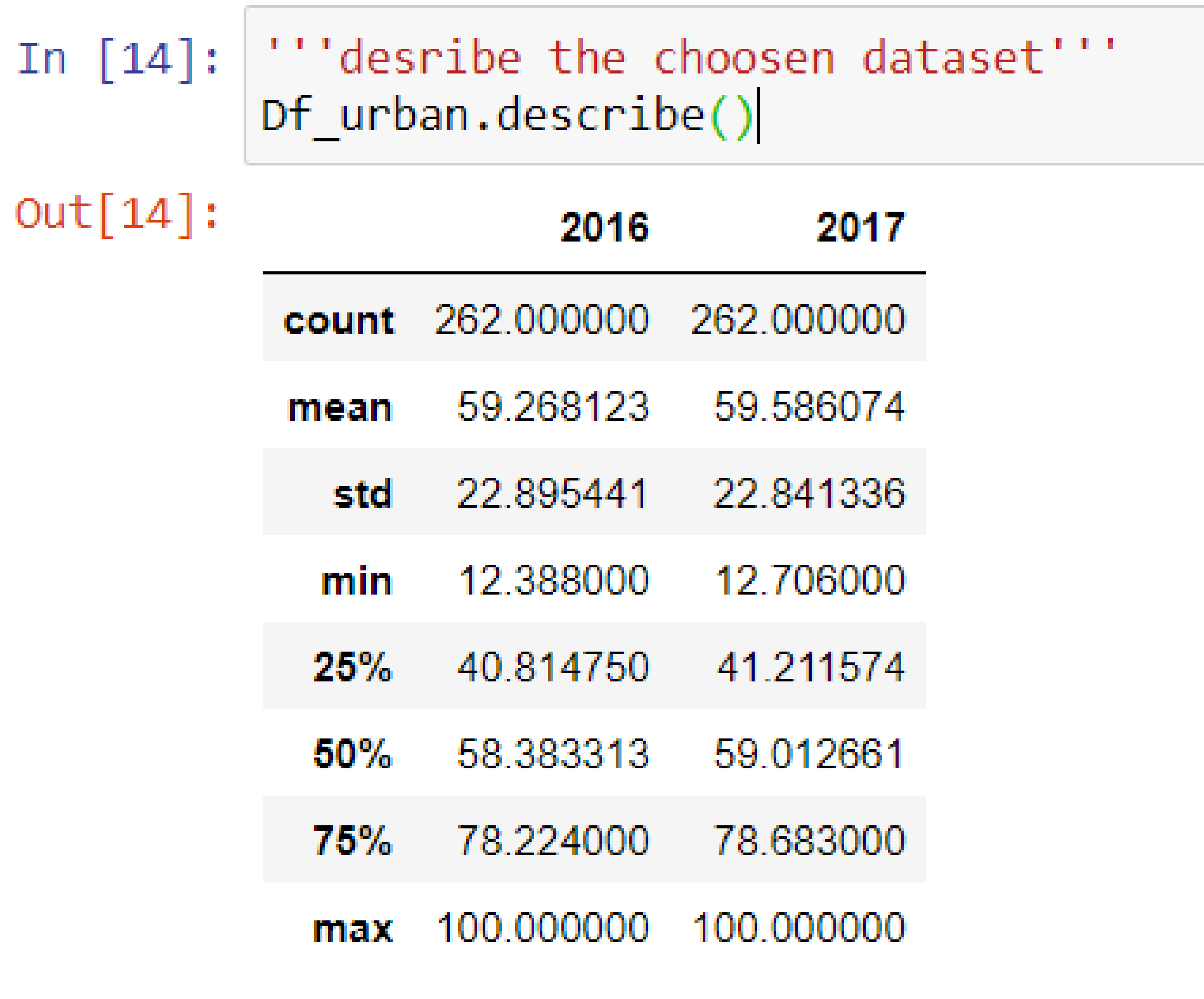
## Clustering.

Clustering is the practice of dividing a population or set of data points into various groups so that data points in the same group are more similar and data points in other groups are more dissimilar. It's just a grouping of objects depending on how similar and distinct they are. It's simply a method of unsupervised learning. Unsupervised learning is a method of obtaining references from datasets with input data but no labelled answers. It's commonly used to find significant structure, explain underlying processes, generate attributes, and group things together.

## Elbow Curve

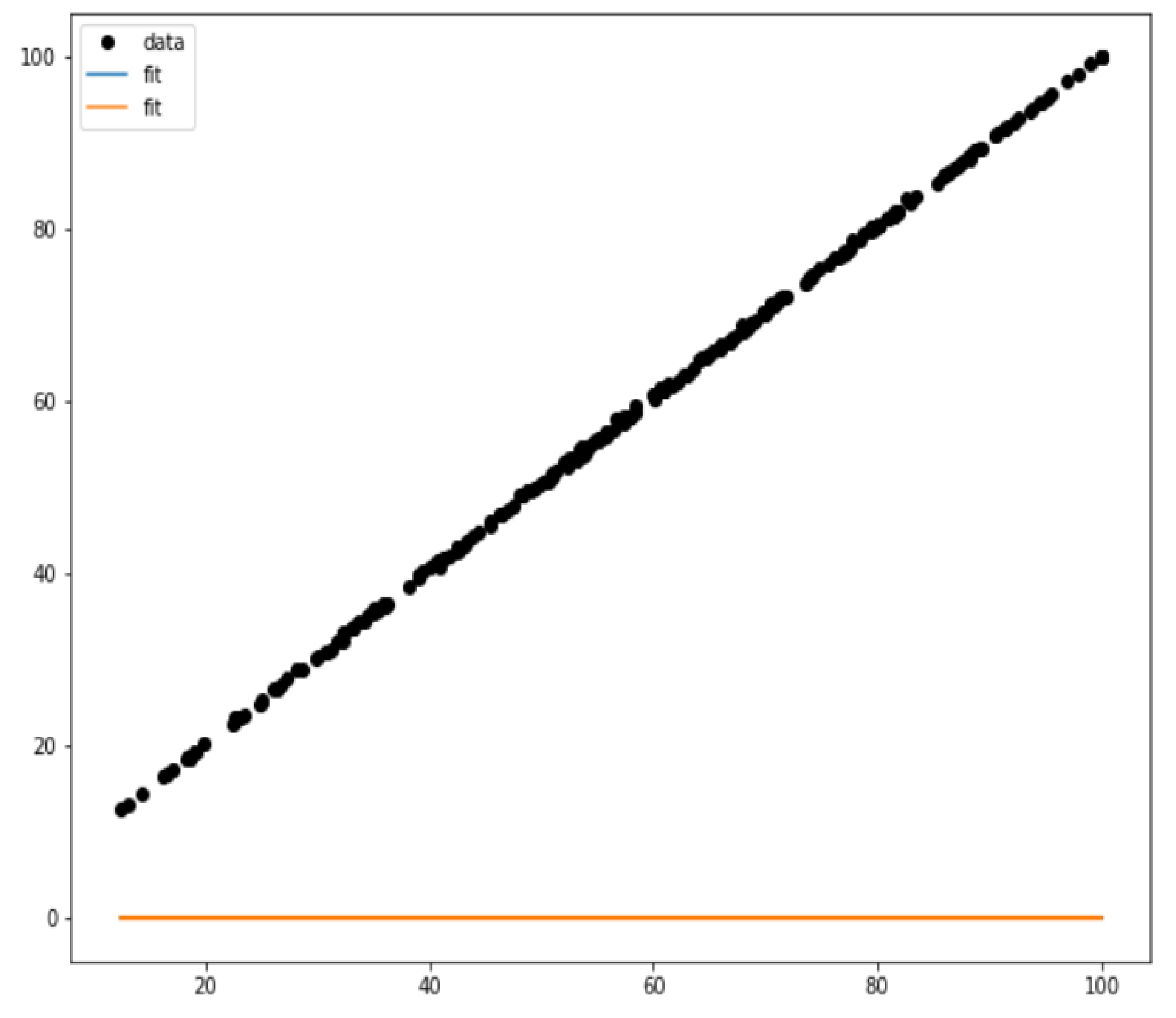
The elbow curve technique to the library SK learn is used before clustering to determine how many clusters are possible in a given data set. Learn that this is a default mode of inertia and add it to the list. Then, using matplotlib, plot the entire list of data and assert that when the line becomes straight till that point, record x and assume that number of clusters.

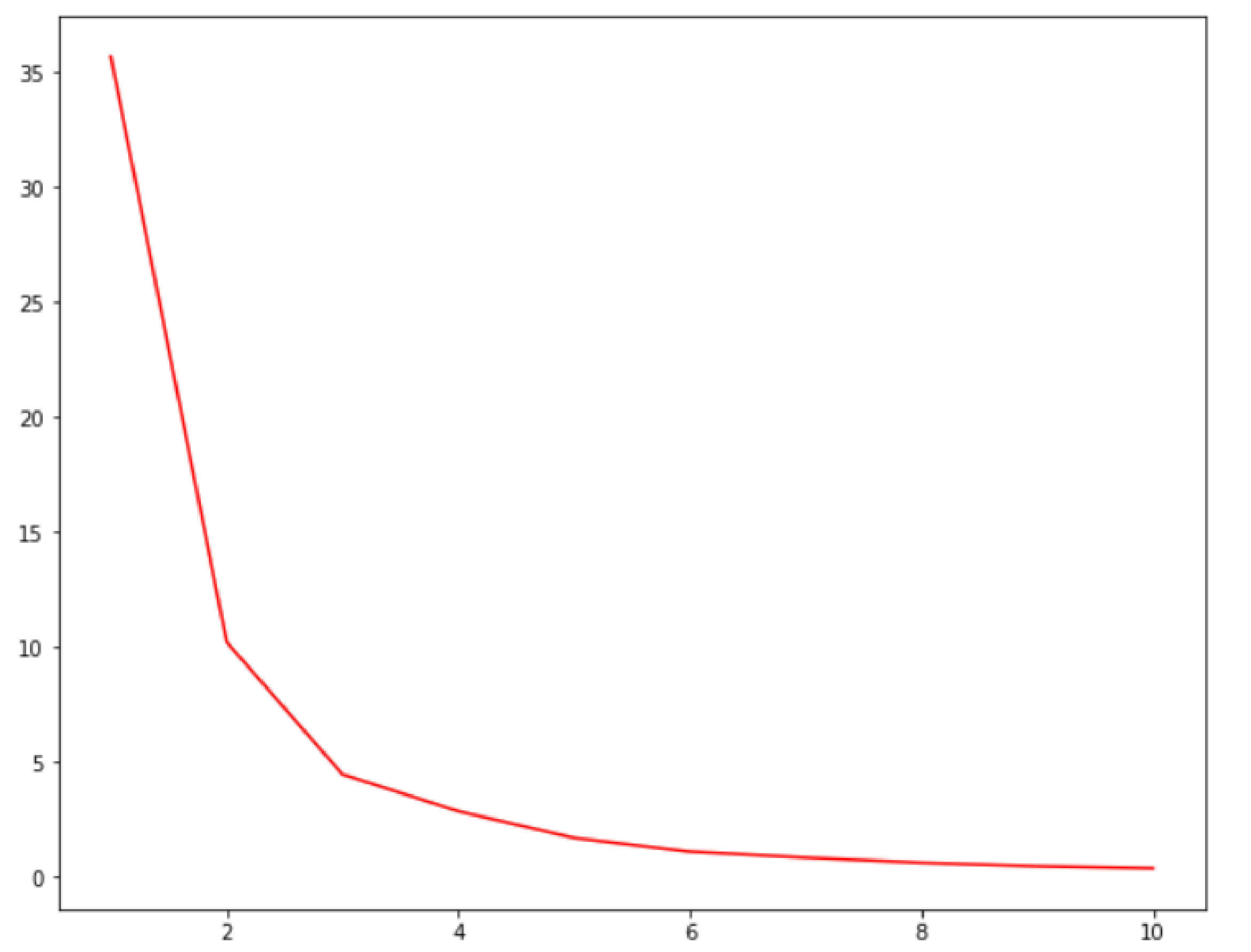

Hirearchical clustering



```
In [14]: '''desribe the choosen dataset'''
         Df_urban.describe()

Out[14]:
                2016         2017
       count  262.000000   262.000000
        mean   59.268123    59.586074
         std   22.895441    22.841336
         min   12.388000    12.706000
         25%   40.814750    41.211574
         50%   58.383313    59.012661
         75%   78.224000    78.683000
         max  100.000000   100.000000
```





## Hierarchical clustering.

Hierarchical clustering, often termed hierarchical cluster analysis, is a method of grouping comparable objects into clusters. The endpoint is a collection of clusters, each of which is distinct from the others yet the items within each cluster are broadly similar. Each observation is treated as a separate cluster in hierarchical clustering. After that, it repeats the next two steps: (1) Find the two clusters that are the closest together, and (2) combine the two clusters that are the most similar. This iterative process is repeated until all of the clusters have been integrated. It is vital to determine where distance is computed after picking a distance measure. the cluster's two least similar bits, or some other requirement

## Interpretation of Results.

Hierarchy of the dataset is defined in dendrogram and we also provide different clusters in scatters from to more precise analysis. In this there is many urban countries and each urban country is value according to years values and hierarchy is also be drawn to place greater urban countries of both years and further little and little.