Clustering, Correlation & Data Fitting for world Population rate

By: Hassan Faraz Khan (21089475)

Abstract: In this assessment a story telling through data visualization approach is being used. The dataset selected for this poster is world population dataset from world bank having data rang from 1960 to 2021. the aim of this assessment is to visualize the interesting data patterns in global population. Initially the data patterns were random and irregular, after setting and parameterizing the model the cluster plot get accurate and interesting as shown in the poster.

Introduction: This Datasets of world population rate are collected from the World Bank Site for a number of nations from 1960 to 2021. Population rate dataset contains statistics on each country's overall world population rate as a proportion of its total population. Using this dataset, we plot clusters for some interested data and also plot the correlation matrix and curve fit plot.

Aim - this assessment aims to find out the interesting cluster in the chosen data.

Objectives

- To create the curve fitting model.
- To tell a story using data visualization approach.
- To find out the meaningful clusters in the dataset.
- To correlate the various variables of data.

	1970	1980	1990	2000	2010	2015
count	264.000000	264.000000	264.000000	265.000000	265.000000	265.000000
mean	2.212299	2.044016	1.750571	1.432954	1.379407	1.323819
std	1.709190	1.528791	2.502996	1.296961	1.426037	1.402870
min	-2.562076	-3.924172	-27.722225	-4.075386	-5.343938	-4.415744
25%	1.216862	0.987491	0.781935	0.611851	0.482960	0.500452
50%	2.319867	2.085517	1.924618	1.447088	1.253650	1.140936
75%	2.728104	2.795081	2.734540	2.394192	2.369772	2.200322
max	16.295475	10.996347	10.219241	5.580387	6.611552	9.219918

Methodology: After that using python codes and libraries generate a correlation matrix in figure below that describe the correlation matrix between interested years of world population data. In the next step, the data clusters are found on specific years. In last step the data fitting approach is applied to assess the data on a fitted line.

Summary: In this poster the required goals are achieved . The scatter plots, correlation matrix, cluster plots, and the data is plotted on a fitted line.

Repositories Link:
https://github.com/HassanFrazKhan/ADS-1_Assignment3

Transpose of data

```
[8 rows x 62 columns]
             1970        1980        1990        2000        2010        2015
count  264.000000  264.000000  264.000000  265.000000  265.000000  265.000000
mean     2.212299    2.044016    1.750571    1.432954    1.379407    1.323819
std      1.709190    1.528791    2.502996    1.296961    1.426037    1.402870
min     -2.562076   -3.924172  -27.722225   -4.075386   -5.343938   -4.415744
25%      1.216862    0.987491    0.781935    0.611851    0.482960    0.500452
50%      2.319867    2.085517    1.924618    1.447088    1.253650    1.140936
75%      2.728104    2.795081    2.734540    2.394192    2.369772    2.200322
max     16.295475   10.996347   10.219241    5.580387    6.611552    9.219918
      count      mean       std         min       25%       50%       75%        max
1970  264.0  2.212299  1.709190   -2.562076  1.216862  2.319867  2.728104  16.295475
1980  264.0  2.044016  1.528791   -3.924172  0.987491  2.085517  2.795081  10.996347
1990  264.0  1.750571  2.502996  -27.722225  0.781935  1.924618  2.734540  10.219241
2000  265.0  1.432954  1.296961   -4.075386  0.611851  1.447088  2.394192   5.580387
2010  265.0  1.379407  1.426037   -5.343938  0.482960  1.253650  2.369772   6.611552
2015  265.0  1.323819  1.402870   -4.415744  0.500452  1.140936  2.200322   9.219918
PS D:\University\Assignment\ADS-1_Assignment_3>
```