Cause of Death data

Data is all about cause of death. There is different type of reason for death. There are 6120 columns & 34 rows in the dataset and mentioned no of death in country in the world.

Observation part:

1. Total no of columns are 6120 & Rows are 34.
2. There are showing how many no. of death is having in which year and for which reason.
3. Country ,code & Year is categorical data and nominal data.
4. There is no missing value is present in the dataset.
5. There is object value is present in two columns Country & Code
6. All column's data is integer value except Country & code columns.
7. Total 30 nos of country are available in the data set and each country has 30 nos unique value.
8. Code column is short form of country's name.

Handle the dataset:

1. We delete two columns one is Country & code due to nominal & categorical date. These are not much important to data analysis & relationship with target.
2. We delete duplicate value if present. There is no duplicate value in the dataset. Because no of rows are heave as earlier rows numbers.

Observation & treat the problem if any:

1. Describe the date to show mean, std & quantile ratio. We can find out the problem from the dataset if present.
2. Year is the categorical & nominal data column. Here data distribution is okay. Mean & std of the data is good.
3. Rest all data column's mean, std & quantile ration is not good. Std is greater than mean of the Data which in not meaningful in every columns.
4. Data is not normal distributed in the dataset in every column.
5. There is have outlier in every column.
6. Skewness is also present in the dataset.

Treat some method with problem:

1. Use Z score method to handle outlier. Z score means only can take data upto 3 std data. 99.24% data can use for create model and rest will delete.
2. It will treat for each and every columns.
3. It will minimize the outlier in the dataset.

Analysis :

1. We are checking relationship each column with other.
2. If correlation score is close to zero means, no good relationship with feature each other. It can find from heatmap.
3. As per map. No good relationship of year with every feature's data.
4. There is skewness, if +->0.5 then there are skewness. In the dataset, all columns values are greater than +->0.5.

5. Use PowerTranformer method to reduce the skewness of dataset. There values are also now greater than +- >0.5. but meanmize the skewness as much as possible and now minimize as earlier as skewness.