Principal Component, Factor Analysis and Principal Component Regression
Adinda Herawati - 425148

A. Sourced Data

The data for this research are sourced from Kaggle (https://www.kaggle.com/augustus0498/life-expectancy-who). There are 22 columns and 2938 rows. Here is the description of the data

No	Variables	Description	No	Variables	Description
1	Country	List of copuntries in the world	12	Under Five Deaths	Number of under-five deaths per 1000 population
2	Year	Year	13	Polio	immunization coverage among 1-year-olds (%)
3	Status	Status of the countries (Developing or Developed)	14	Total Expenditure	General government expenditure on health as a percentage of total government expenditure (%)
4	Life Expectacy	In Age	15	Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
5	Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	16	HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
6	Infant Deaths	Number of Infant Deaths per 1000 population	17	GDP	Gross Domestic Product per capita (in USD)
7	Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	18	Population	Population of the country
8	Percentage Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)	19	Thinness 1-19 Years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
9	Hepatitis B	Hepatitis BHepatitis B (HepB) immunization coverage among 1-year-olds (%)	20	Thinness 5-9 Years	Prevalence of thinness among children for Age 5 to 9(%)
10	Measles	number of reported cases per 1000 population	21	Income Composition of Resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

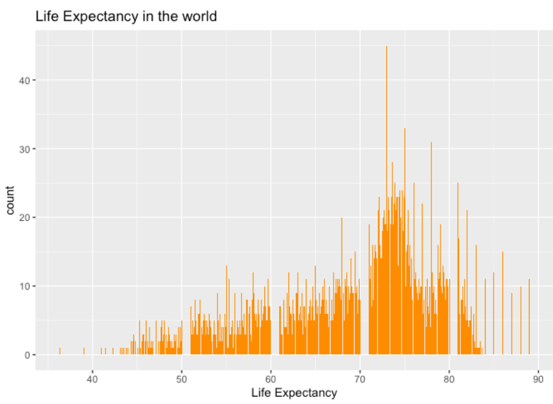
11	BMI	Average Body Mass Index of entire population	22	Schooling	Number of years of Schooling(years)

B. Analysing Data

in this first analysing data, I did some pre-processing data such as descriptive statistics but focused on missing values because if we involved the missing values in our data it will affects of the results.

No	Variables	Total of Missing Values	No	Variables	Total of Missing Values
1	Life Expectacy	10	11	Total Expenditure	226
2	Adult Mortality	10	12	Diphtheria	19
3	Infant Deaths	-	13	HIV/AIDS	-
4	Alcohol	194	14	GDP	448
5	Percentage Expenditure	-	15	Population	652
6	Hepatitis B	553	16	Thinness 1-19 Years	34
7	Measles	-	17	Thinness 5-9 Years	34
8	BMI	34	18	Income Composition of Resources	163
9	Under Five Deaths	-	19	Schooling	167
10	Polio	19			

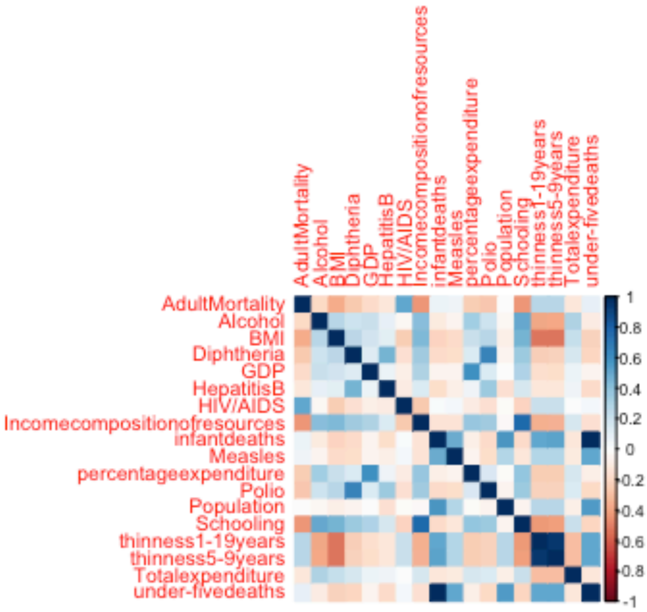
Based on the Table 2, as we can see there are some missing values in most of variables. Hence, I used imputation data using median to fill missing values in variables which have missing values. Therefore, my data are more clean now without missing values. Following that, I tried to use descriptive statistics in some variables.



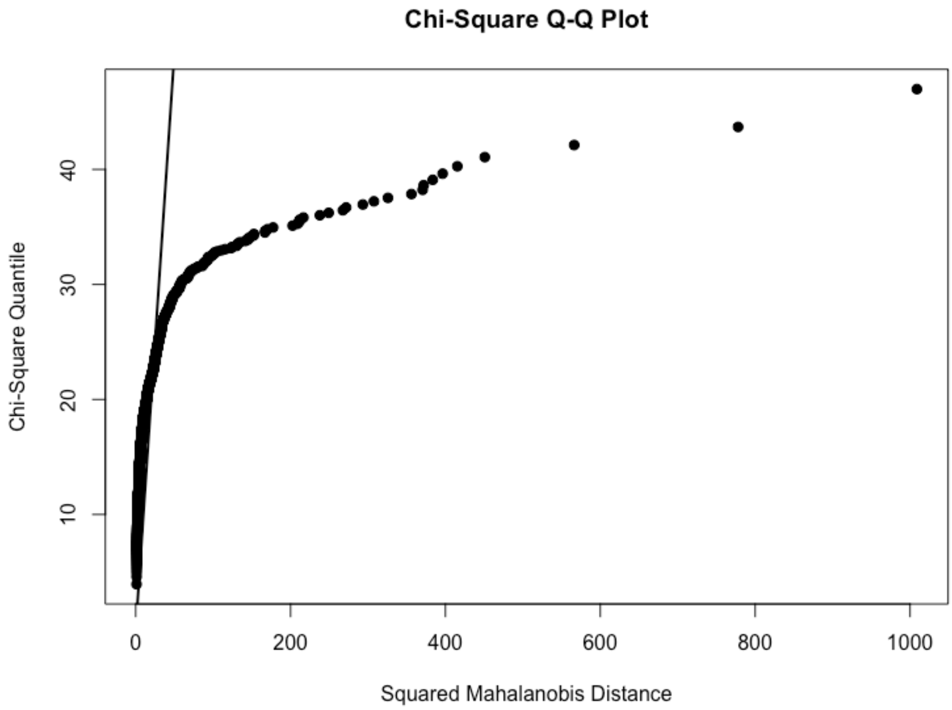
Graph Life Expectancy In The World

From Graph 1 as we can see that people have expectation that they will live in the world in range 36-89 years old. However from the graph also we can say that mostly life expectancy for people around 70-80 years old.

After doing some visualisation in data, next step is check the assumption. First I would like to check whether the data has multicollinearity or not, then I will check for multivariate normality assumption



Based on Graph 1 mostly in some variables, they have high collinearity with other variables even though some of the variables have low correlation also in some variables. Then we can try to use PCA for this data although this data is required data should has high multicollinearity. Second, I would like to check the multivariate normality in this data using mvn packages.



Based on Graph above, it shows that there are so many outliers in that data and those of data are stay away from normality line. Using Mardia Test for Multivariate Normality Assumptions, it clearly shows that Life Expectancy Data doesn't have multivariate normality assumption because their p-value is less than 5%.

		Test Statistic	P-Value	Result
1	Mardia Skewness	533357.799259453	0	NO
2	Mardia Kurtosis	1628.41697420486	0	NO
3	MVN	<NA>	<NA>	NO

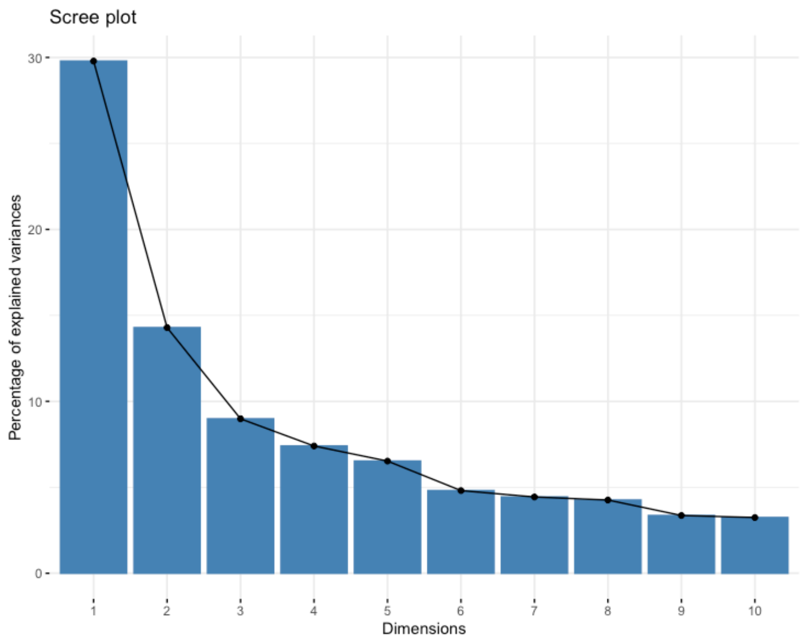
1. Results for Variables

After all those assumption test then we continue to calculations of PCA. Using Eigen values, we can say that we got 5 new factors from Life Expectancy Data because their Eigen values are more than 1. Based on their cumulative variance percent, it clearly shows that using 5 factors, they can explain about 66.9% of total variance of data.

	eigenvalue	variance.percent	Cumulative Variance Percent
Dim.1	5.362254256	29.79030142	29.79030
Dim.2	2.571961857	14.28867698	44.07898
Dim.3	1.617178434	8.98432463	53.06330
Dim.4	1.332474045	7.40263358	60.46594
Dim.5	1.174768212	6.52649006	66.99243
Dim.6	0.865713581	4.80951989	71.80195
Dim.7	0.798827140	4.43792856	76.23988
Dim.8	0.767130424	4.26183569	80.50171
Dim.9	0.605115042	3.36175023	83.86346
Dim.10	0.583895149	3.24386194	87.10732
Dim.11	0.506343867	2.81302148	89.92034
Dim.12	0.447317242	2.48509579	92.40544
Dim.13	0.407413030	2.26340572	94.66885
Dim.14	0.378240900	2.10133833	96.77018
Dim.15	0.323240205	1.79577892	98.56596
Dim.16	0.209702559	1.16501422	99.73098
Dim.17	0.046043942	0.25579968	99.98678

Dim.18	0.002380116	0.01322287	100.00000

Graph below explain about the scree plot which based on Eigen values score. To interpret the scree plot, we can say that it suits with the results before that in this paper I got 5 new factors because after 5 factors, the other factors become more constant values.



Besides explanation based on cumulative proportion, we can check from their standard deviation for 5 first factors. Standard deviation would explain the dispersion of the data, In PC1 the dispersion of the data would around 2.3157, in PC2 they have dispersion of the data about 1.603, in PC3 their standard deviation is about 1,27168, in PC4 their standard deviation is about 1.15433 and last for PC5 is about 1.08387.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Standard Deviation	2.3157	1.6037	1.27168	1.15433	1.08387
Proportion of Variance	0.2979	0.1429	0.08984	0.07403	0.06526
Cumulative Proportion	0.2979	0.4408	0.53063	0.60466	0.66992

In all tables below show the correlation between variables to their factors. Based on lectures, we can say that the more important the variable in a given factor when they have min 0.3-0.5. In general, almost in all factors have difference variables which have high correlation.

	PC1	PC2	PC3	PC4	PC5
AdultMortality	0.2168619	-0.232313066	0.058091710	0.409100338	-0.20241890
infantdeaths	0.2627607	0.439918900	0.006911180	-0.038901066	-0.14546115
Alcohol	-0.2209314	0.176668402	0.198285926	0.224729089	-0.29640154
percentageexpenditure	-0.1933403	0.226326064	0.236366558	0.357143123	0.36322573
HepatitisB	-0.1414904	0.001035664	-0.492189813	0.171060232	-0.01896068
Measles	0.1792380	0.280556482	0.070678744	-0.028071765	-0.13916995
BMI	-0.2761531	0.093519201	0.093383507	-0.213157450	-0.15734680
under-fivedeaths	0.2670372	0.433236042	0.018492307	-0.032636789	-0.14986553
Polio	-0.2247388	0.110724056	-0.468139017	0.115918443	-0.08307804
Totalexpenditure	-0.1579560	0.028220097	0.132832003	0.183412153	-0.47367080
Diphtheria	-0.2266890	0.108542234	-0.512465325	0.130516729	-0.10551198
HIV/AIDS	0.1396912	-0.160542614	0.087029396	0.577704142	-0.26770667
GDP	-0.1551970	0.213483531	0.182376751	0.361408239	0.47851081
Population	0.1476645	0.376415239	0.009524495	-0.067327437	-0.16369938
thinness1-19years	0.3383551	0.118587437	-0.227589528	0.144920193	0.19805815
thinness5-9years	0.3382405	0.122505483	-0.225342462	0.145940463	0.19473260
Incomecompositionofresources	-0.2867264	0.261915044	0.021116612	-0.005040098	0.05881552
Schooling	-0.3076825	0.243657991	0.041412059	0.034493239	-0.02631689

Based on Eigen values we got 5 factors only than in table above, we would like to know the contribution of each variables in 5 factors respectively. Let's say in PC 1, there are three variables, thinness1-19years, thinness5-9years and schooling, which have positive correlation and contribution to create new factor in PC1. Following that, we would like to name it as **Education**.

However in PC2, only one variable ,Population, has positive correlation and contribution to create new factor, PC2. Therefore, the new variable would name it as **Population**.

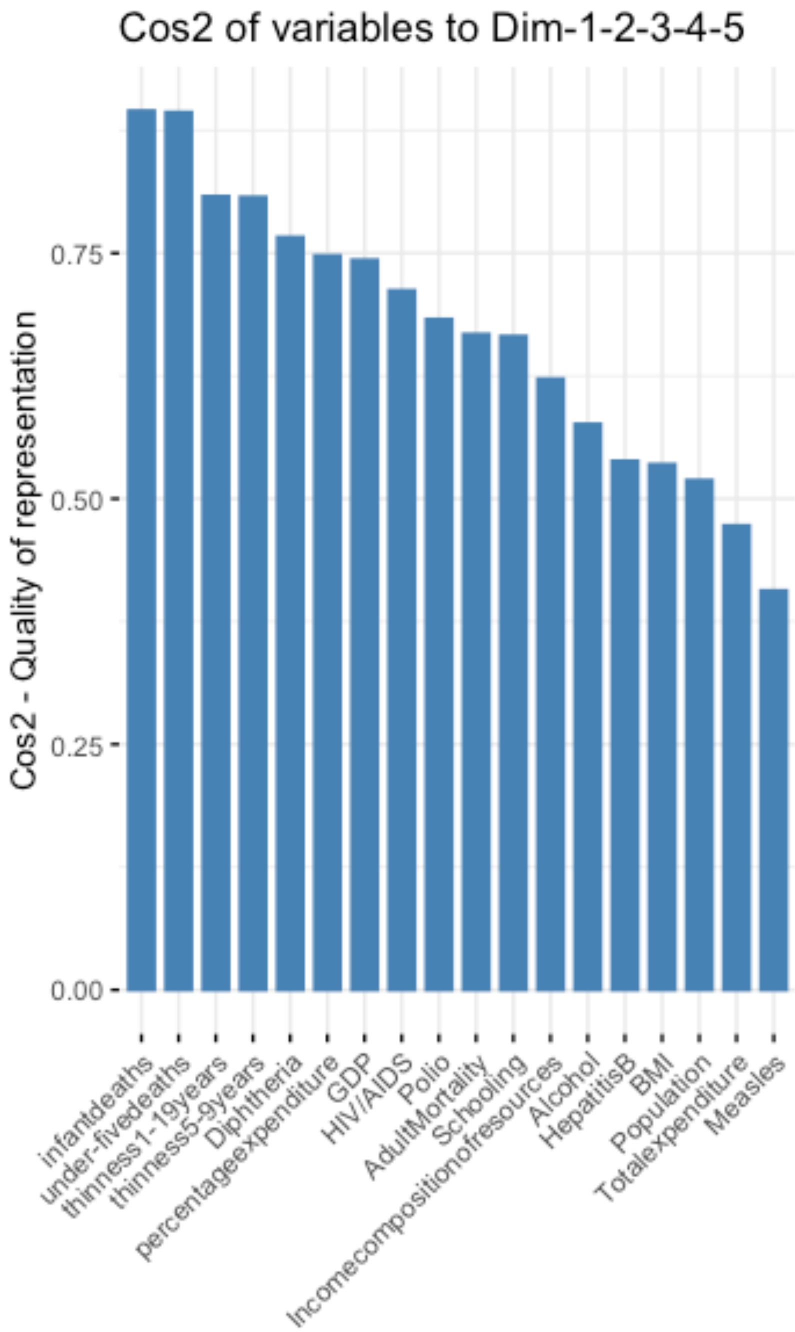
Variables Hepatitis B, Polio and Diphtheria have negative correlation and contribute to create new factor, PC3 then we name the new variables as **Diseases**.

Variables Adult Mortality, Percentage Expenditure, HIV/AIDS and GDP have positive correlation and contribute to create new factor in PC4. As a result we give name to new factor as **Economic Effects**.

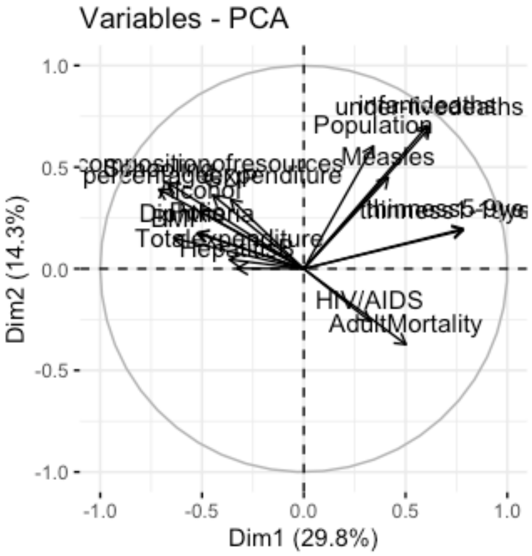
Percentage Expenditure and GDP have positive correlation to PC5, nevertheless Total Expenditure has negative correlation to PC5. However, we will give the name for PC5 as **Economic Reasons**.

Table below shows the summary of the new factors's name and what kind of variables have influenced to create 5 new factors.

Principal Component	Variable's Name	Loading Factors	Variance
Education	thinness1-19years	0.3383551	29.79030142
	thinness5-9years	0.3382405	
	Schooling	-0.3076825	
Population	Population	0.376415239	14.28867698
Diseases	Hepatitis B	-0.492189813	8.98432463
	Polio	-0.468139017	
	Diphteria	-0.512465325	
Economic Effects	Adult Mortality	0.409100338	7.40263358
	Percentage Expenditure	0.357143123	
	HIV/AIDS	0.577704142	
	GDP	0.361408239	
Economic Reasons	Percentage Expenditure	0.36322573	6.52649006
	GDP	-0.47367080	
	Total Expenditure	0.47851081	



Graph Cos2 about shows Infant Deaths, Under Five Deaths, Thinnes 1-19 years, Thinnes 5-9 years, Diphteria, Percentage Expenditure and so on have good representation of the variables to principal components. The highest of the cos2 means that will have good represent the variables to principal components.



Based on that correlation plot, as we can see that variables Income Composition of resources, schooling, percentage expenditure, alcohol, BMI, Diphtheria, Polio, Total Expenditure and Hepatitis B become one group and they have positive relationship. However the next group, there are population, measles, infant deaths, under five deaths, thinness 1-19 years and thinness 5-9 years also have positive relationship. Last group, there are HIV/AIDS and adults mortality have positive relationship.

These tables below show contribution of each observation to their new factors. Sign minus (-) in front of the values, it means the negative correlation otherwise plus (+) it means that they have positive correlation.

Observations	PC1	PC2	PC3	PC4	PC5
1	-0.9914374	0.2733730	-0.5061362	-0.9749248	-0.20493593
2	-0.4055571	-0.1045145	0.2732151	-0.6887726	-0.31826694
3	0.6312606	-0.7171200	3.3689505	-1.5345755	0.12924916
4	-0.4503224	-0.3904740	0.3495748	-0.5810087	-0.22621336
5	0.6273970	-0.7297799	-1.0349147	0.2579379	0.05102953
6	0.5324157	-0.5286082	-1.0172413	-0.6079566	0.54814256

In table above shows contribution for each observations to new factors. This table above only explain about observation from number 1 until 6. Take an example in PC1, from 6 observations, we can say that observation number 3, 5 and 6 have positive correlation and contribution to PC1 about 0.6312606, 0.6273970 and 0.5324157 respectively. Otherwise, observation number 1,2, and 4 have negative correlation and contribution to PC1 about -0.9914374. The interpretation would be similar for other PCs.

2. Results for Individual

In this individual Results mean that we would know the contribution for each observation to their dimensions.

Observations	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	0.007800373	0.0012364569	0.006740762	0.030353980	1.521305e-03
2	0.001305235	0.0001807262	0.001964191	0.015150436	3.669129e-03
3	0.003162293	0.0085084652	0.298650647	0.075205553	6.051115e-04
4	0.001609281	0.0025226201	0.003215544	0.010780495	1.853602e-03
5	0.003123702	0.0088115309	0.028182744	0.002124731	9.432418e-05
6	0.002249502	0.0046231226	0.027228403	0.011803709	1.088346e-02

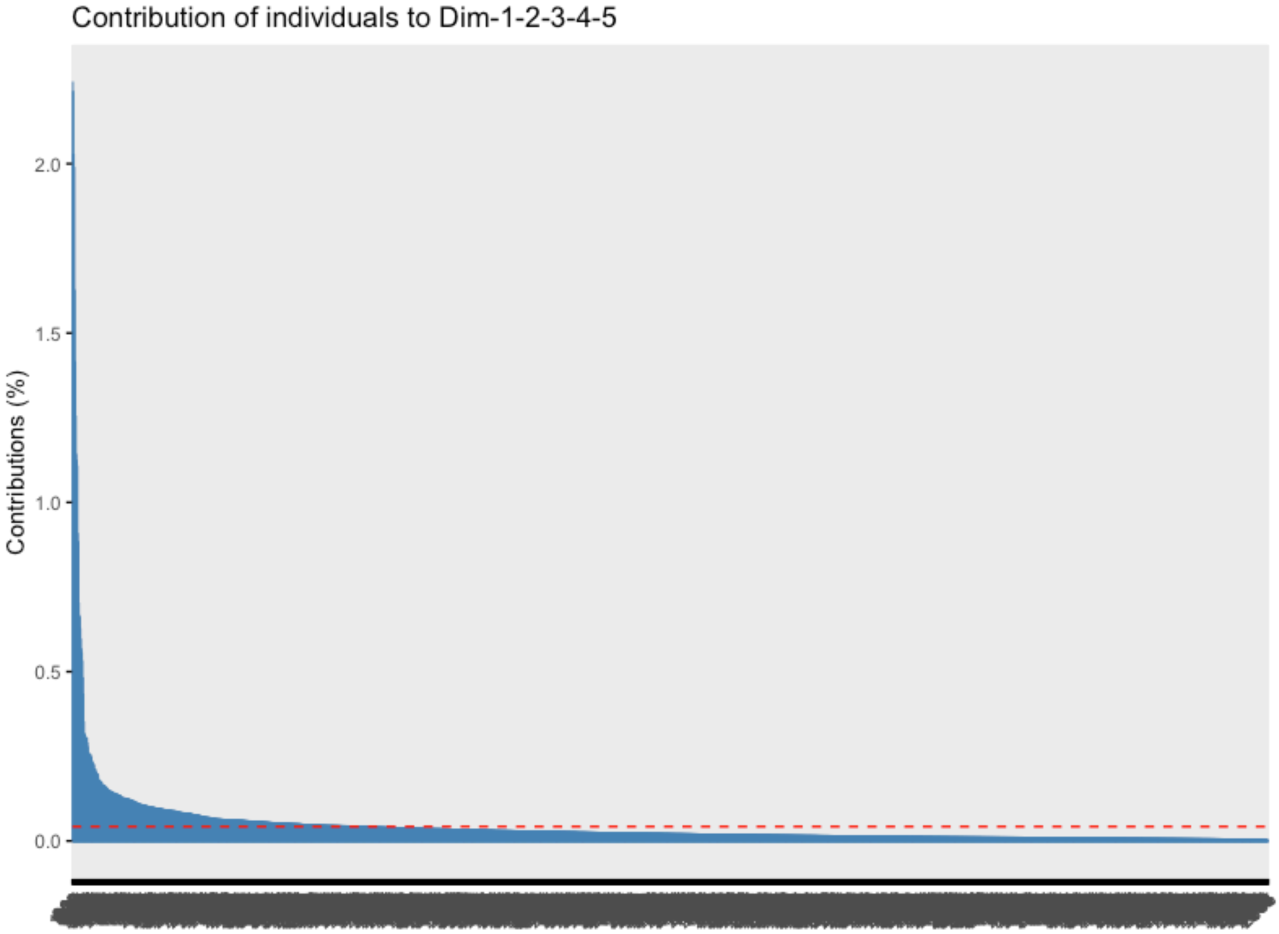
In table above shows contribution for each observations to new dimensions. This table above only explain about observation from number 1 until 6 in percentages. Take an example in Dim 1. the contribution for observation 1,2,3,4,5 and 6 are about 0.7%, 0.1%, 0.3%, 0.3% and 0.2%. For the other dimensions, they have similar interpretation based on the values.

Observations	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	-0.9914374	0.2733730	-0.5061362	-0.9749248	-0.20493593
2	-0.4055571	-0.1045145	0.2732151	-0.6887726	-0.31826694
3	0.6312606	-0.7171200	3.3689505	-1.5345755	0.12924916
4	-0.4503224	-0.3904740	0.3495748	-0.5810087	-0.22621336
5	0.6273970	-0.7297799	-1.0349147	0.2579379	0.05102953
6	0.5324157	-0.5286082	-1.0172413	-0.6079566	0.54814256

Table above shows coordinates from individual results for creating scatterplot

Observations	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	0.28826634	0.021916676	0.07512735	0.27874403	0.0123168466
2	0.11452308	0.007605756	0.05197546	0.33032441	0.0705296883
3	0.01825462	0.023558038	0.51992986	0.10787779	0.0007652634
4	0.18209998	0.136913718	0.10973443	0.30312940	0.0459514519
5	0.15040997	0.203505188	0.40926143	0.02542273	0.0009950269
6	0.05947096	0.058623401	0.21709568	0.07754402	0.0630362329

Table above shows quality of representations for variable on the factor map (cos2), all those values come from square coordinate. Based on the contribution values above, it can get plot like this

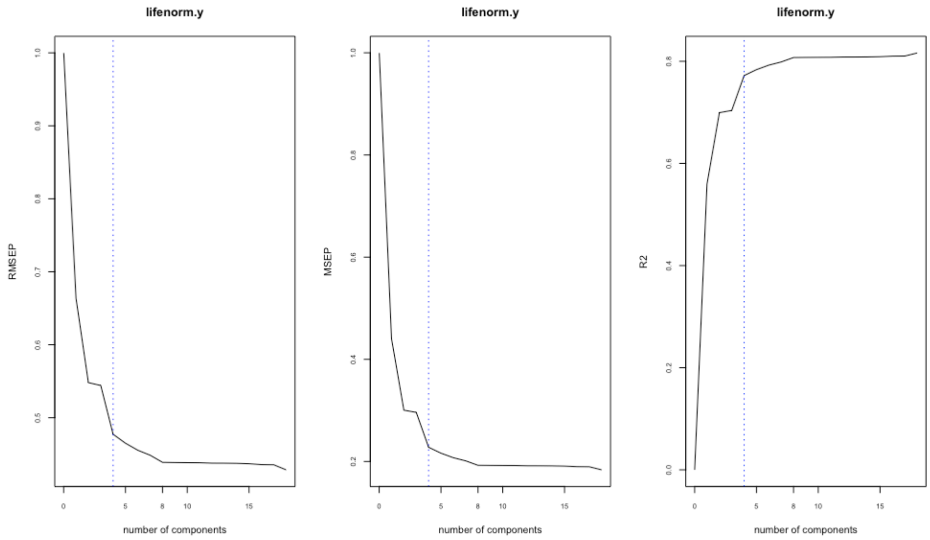


C. Principal Component Regressions

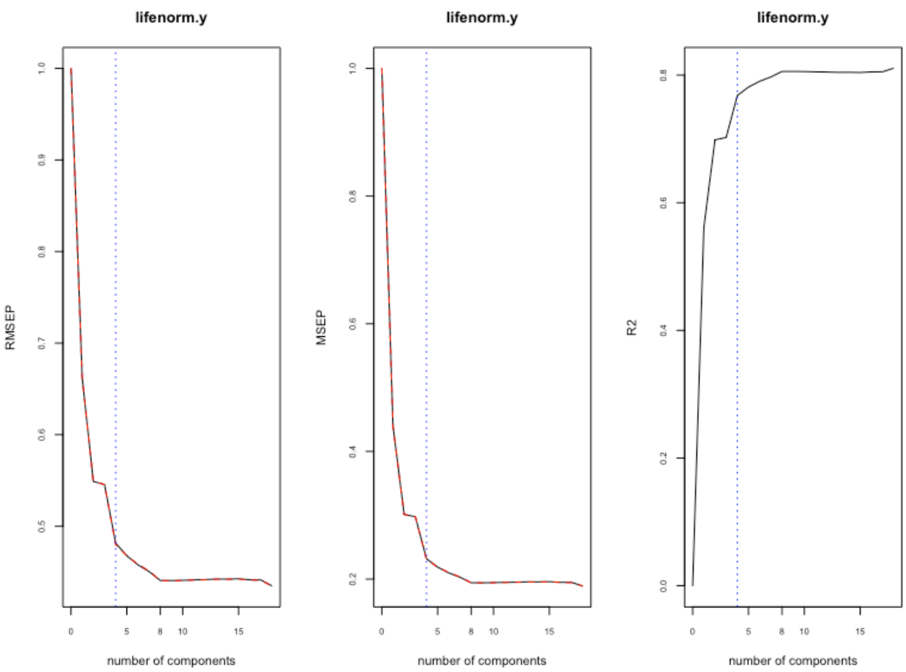
In this paper also tried to analysis data using Principal Component Regressions, based on this results shows that almost all factors have significant influenced to variable dependent/ life expectancy.

	Estimate	Standar Error	T-Value	P-Value
(Intercept)	4E-13	9E+00	0	1.000000
PC1	-3E+02	4E+00	-84.258	< 2e-16
PC2	2E+02	6E+00	42.114	< 2e-16
PC3	-5E+01	7E+00	-7.411	1.75e-13
PC4	-2E+02	8E+00	-29.433	< 2e-16
PC5	1E+02	8E+00	12.107	< 2e-16
PC6	-1E+02	1E+01	-10.608	< 2e-16
PC7	-9E+01	1E+01	-8.866	< 2e-16
PC8	1E+02	1E+01	10.487	< 2e-16
PC9	2E+01	1E+01	1.663	0.096366
PC10	-2E+01	1E+01	-1.354	0.175894
PC11	-1E+01	1E+01	-1.189	0.234672
PC12	-3E+01	1E+01	-2.508	0.012199
PC13	1E+01	1E+01	718	0.472844
PC14	2E+01	1E+01	1.430	0.152880
PC15	-4E+01	2E+01	-2.432	0.015093
PC16	7E+01	2E+01	3.672	0.000247
PC17	-7E+01	4E+01	-1.613	0.106871
PC18	2E+03	2E+02	8.723	< 2e-16

to choose appropriate factors in PCR, I analysed with R2, RMSEP, and MSEP plots.



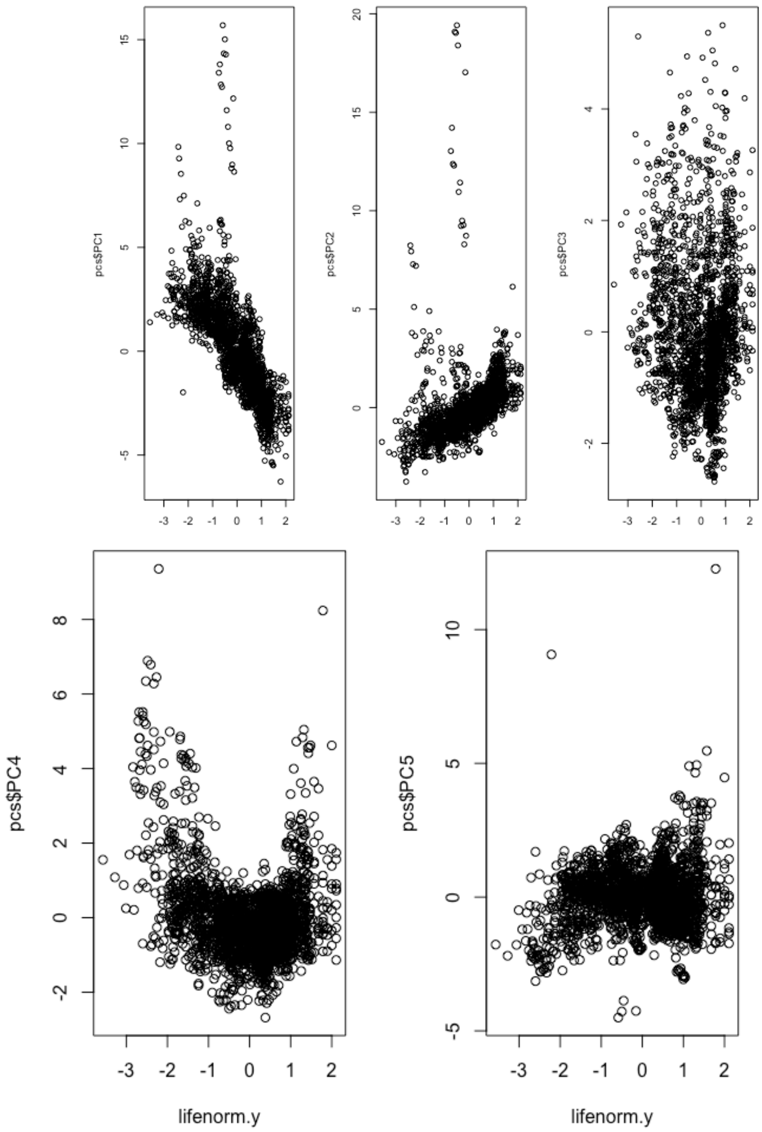
Based on those three plots, we can say that those of the plots give similar results that all the points change dramatically in 4 factors. It means that using PCR, we reduce our variables become 4 factors only. Then I tried again to get the conclusions using cross validation and the results still the same that using when we use Cross Validation to determine number of new factors.



3. Comparison between Linear model without PCA, PCR and after PCA

	Before PCA	PCR	After PCA
Multiple R-squared	0.964	0.8163	0.7832
Adjusted R-squared	0.9613	0.8148	0.7836
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16

based on R-squared, using original data, I got R-square about 96.4%, higher than the model after reducing variables using PCA. Moreover, the PCA R-squared explains that the model can explain 96.4% variation of the data that affect to dependent variables which is life expectancy variable. I also tried to calculate R-square in PCR methods and the result is 81.63% which means that model can explain 81.63% variation of the data.



5 scatterplots above shows correlation between 5 new factors to life expectancy as variable dependent. In the first plot, it shows low negative correlation between PC1 and variable dependent (life expectancy). Following that, second scatterplot it shows low

positive correlation between PC2 with variable dependent (life expectancy). however in other scatterplots, such as scatterplots between PC3 with life expectancy, PC4 with life expectancy and PC5 with life expectancy, I assume that they have no correlation because their plots spreads and they don't have patter in those three plots.

D. Conclusion

After all the analysis above, we can conclude that:

1. From 18 variables, I reduce the variables using PCA until I got 5 factors only. Those 5 factors are created from variables which have high correlation. The new variables name are Education, population, Diseases, Economic Effects, Economic Reasons. When I tried to compare R-squared between regression linear model using original data, after reducing the variables using PCA and PCR, I got the results that regression linear before PCA treatments has high R-squared than after PCA and PCR.

2. However, I also tried to use PCR for reducing the variables and I got 4 factors from this method based on Cross Validation and fit plots.

3. Both PCA and PCR did not give satisfied results, these results might happen because in these analysis the data did not meet the assumptions such as High Multicollinearity in all variables and Multivariate Normal Distribution. However, before to start Principal Component Analysis, usually researcher besides check the multicollinearity and multivariate normal distribution, they should check whether our data enough for PCA analysis using Kaiser Meyer Oikin (KMO) method.