

Data mining project report

WORLD LIFE EXPECTANCY STATISTICAL ANALYSIS

Roshan

Contents

[Project Overview 1](#_Toc445370626)

[Data Summary 2](#_Toc445370627)

[Analysis Summary 2](#_Toc445370628)

[Money Does NOT Necessarily Buy Your Life 3](#_Toc445370629)

[What effect regions have on life expectancy? 5](#_Toc445370630)

[Do you live longer when you reside in a metropolitan city? 7](#_Toc445370631)

[What is the role of annual population growth rate on life expectancy? 8](#_Toc445370632)

[Why is the life expectancy of people in countries who spend more on health, lesser than that of people in countries who spend less on health? 9](#_Toc445370633)

[Conclusion 11](#_Toc445370634)

[References 12](#_Toc445370635)

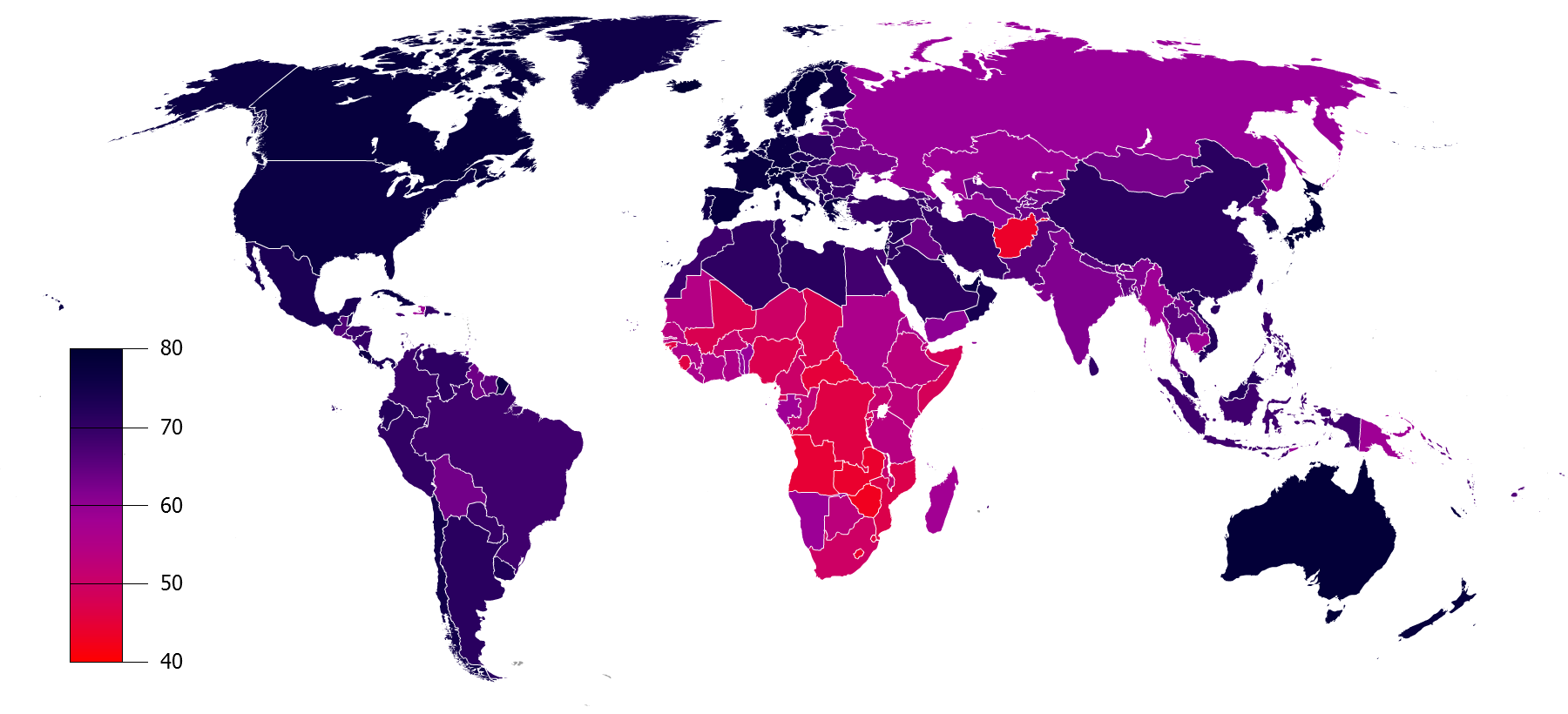
[Appendix 13](#_Toc445370636)

# Project Overview

Life expectancy is a very complicated topic. It relies on a lot of factors, which may interact with each and change over time or upon certain conditions either internal or external. Life expectancy of individuals is almost unpredictable because every human being is unique and has more or less free personal choices during his life. However, life expectancy worldwide may appear certain patterns and features from a large amount of data. Statistics is a right tool to mine such data and find the contributing factors. Furthermore, we can investigate the reason behind the significant factors and thus improve the life expectancy worldwide/region/country. Maybe from here, we can find a way to improve the life expectancy of each individual.

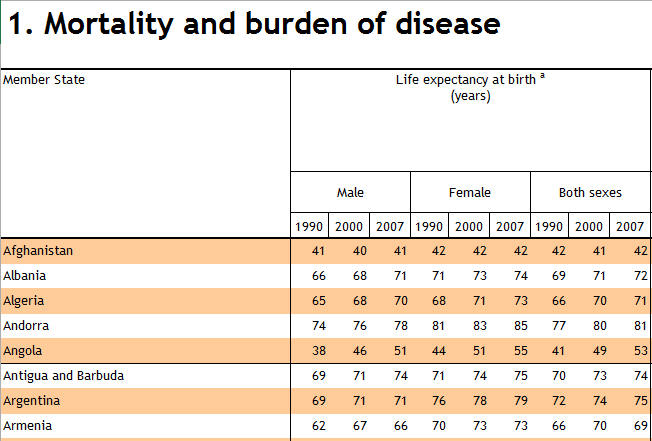
In this project, I utilize the World Health Statistics 2009 from WHO (World Health Organization), which includes the life expectancy at birth and various factors from 193 countries. I also use the global location of each county from statvision.com, trying to understand if location is significant.

The graph of the life expectancy of each country displayed below shows very contrasting patterns, for example, sub-Saharan countries usually have a smaller life expectancy while European countries have the opposite. Afghanistan is an exception comparing to its neighboring countries due to known reasons. We can exclude Afghanistan from the analysis to improve the regression but I keep it for the completeness of the data.



# Data Summary

World Health Statistics 2009 from WHO covers 193 countries and 9 sections. In each section, data are presented in multiple years, which I choose the latest one for our analysis. Our dependent variable, the life expectancy at birth of each country is found in section 1 Mortality and burden of disease (shown below) and categorized in male, female and both sexes. I choose the data of both sexes since it appears that female averagely lives longer than male by a constant number of years globally. Hence the investigation the data of both sex will cover male and female with constant offset.

In section 1. Mortality and burden, I choose the following factors as independent variables:

* Infant mortality rate (probability of dying between birth and age 1 per 1000 live births)
* Under-5 mortality rate (probability of dying by age 5 per 1000 live births)
* Adult mortality rate probability of dying between 15 to 60 years per 1000 population)

In section 5. Risk factors, I choose the following factors as independent variables:

* Access to improved drinking-water source (%)
* Access to improved sanitation (%)
* Alcohol consumption among adults aged ≥ 15 years (litres per person per year)
* Prevalence of current tobacco use (%) Adults (≥ 15 years)
* Prevalence of current tobacco use (%) Adolescents (13-15 years)

In section 6. Health workforce, infrastructure, essential medicines, I choose the following factors as independent variables:

* Physicians Density (per 10 000 population)
* Nursing and midwifery personnel Density (per 10 000 population)
* Hospital beds (per 10 000 population)

In section 7. Health expenditure, I choose the following factors as independent variables:

* Total expenditure on health as % of gross domestic product
* Per capita total expenditure on health at average exchange rate (US$)

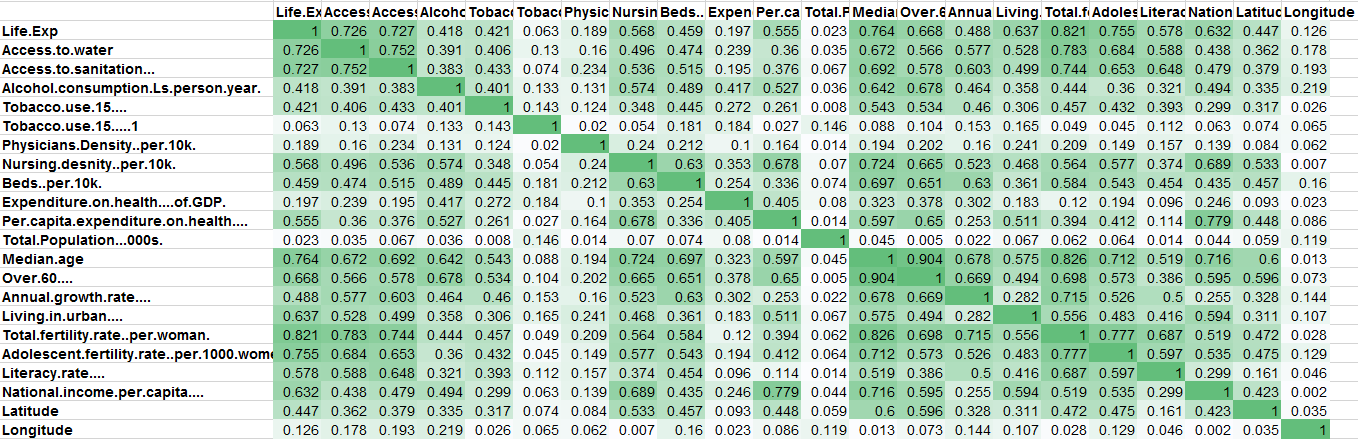
In section 9. Demographic and socioeconomic statistics, I choose the following factors as independent variables:

* Total Population ('000s)
* Median age
* Under 15 (%)
* Over 60 (%)
* Annual growth rate (%)
* Living in urban areas (%)
* Total fertility rate (per woman)
* MDG 5 Adolescent fertility rate (per 1000 women)
* Adult literacy rate (%)
* Gross national income per capita (PPP int. $)

Section 2. Cause-specific mortality and morbidity, 3. Selected infectious diseases, 4. Health service coverage and 8. Health inequities are ignored due to too many categorical data or too many missing data, which either may complicate or invalidate our analysis. We can include them in the further analysis beyond this project. In the selected data set, there are still some data not available, which the average data of all other countries are substituted. I also regroup all countries in the following regions: Africa, Asia&Oceanania, Europe, LatinAmerica, MiddleEast, NorthAmerica for categorical (dummy) analysis.

# Analysis Summary

The selected data set from World Health Statistics 2009 from WHO is further simplified by removing some apparently highly correlated or duplicated columns and keeping only one. The global locations, Latitude and Longitude, are appended in the data set. The table below shows the absolute values of correlation matrix of remaining factors with darker green highlights denotes higher correlation (closer to 1).



It is interesting to find out that the total population has a very weak correlation (close to 0) with all other variables including the life expectancy. It is among the first few insignificant variables being removed in the following multiple regression analysis.

After removing 16 factors one step a time, the final regression analysis results are shown below:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 65.0231009 2.2658864 28.697 < 2e-16 \*\*\*

RegionAsia&Oceanania 6.2792720 1.1423372 5.497 1.31e-07 \*\*\*

RegionEurope 2.6811576 1.6528272 1.622 0.106518

RegionLatinAmerica 6.8123695 1.2718562 5.356 2.57e-07 \*\*\*

RegionMiddleEast 6.7284726 1.4277681 4.713 4.88e-06 \*\*\*

RegionNorthAmerica -0.0683534 3.4001096 -0.020 0.983983

Per.capita.expenditure.on.health.... 0.0010954 0.0003266 3.354 0.000971 \*\*\*

Over.60.... 0.4019978 0.1136510 3.537 0.000515 \*\*\*

Annual.growth.rate.... 2.0118984 0.4376528 4.597 8.05e-06 \*\*\*

Living.in.urban.... 0.0431094 0.0177368 2.431 0.016059 \*

Total.fertility.rate..per.woman. -2.5939495 0.4526208 -5.731 4.13e-08 \*\*\*

Adolescent.fertility.rate..per.1000.women. -0.0544877 0.0113491 -4.801 3.31e-06 \*\*\*

Longitude -0.0225955 0.0070288 -3.215 0.001548 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.978 on 180 degrees of freedom

Multiple R-squared: 0.8581, Adjusted R-squared: 0.8486

F-statistic: 90.68 on 12 and 180 DF, p-value: < 2.2e-16

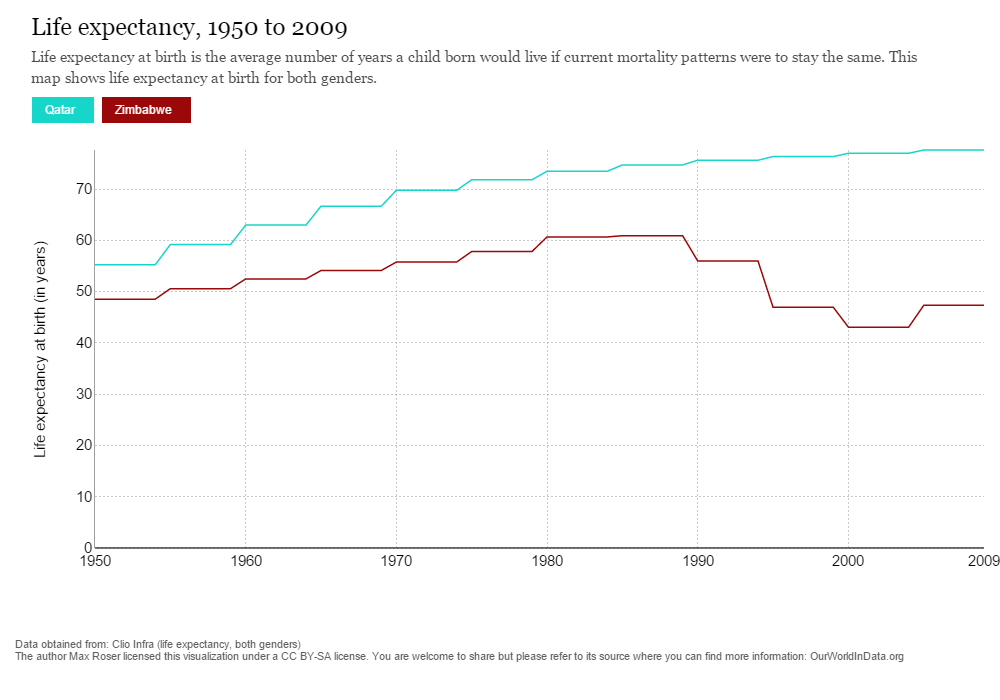
Out of 26 originally select variables, only 7 plus region dummy variables remain significant to the dependent variable, life expectancy at birth. I will explore the possible reasons behind these variable that they are significantly contribute to the life expectancy worldwide.

# What is the role of annual population growth rate on life expectancy?

Annual growth was one of the covariates that that ended up as one of the most significant variables influencing life expectancy during the regression as its p-value was lesser than 0.05. The effect size or the estimate of annual growth rate is 2.012 which is the highest among non-factor variables in the model. This implies us that one unit increase in annual population growth rate of any country would add on an average 2.012 years to the life expectancy of people living in that country.

To get a better insight over this research question, we considered the best and the worst growth rate countries in our data set.

|  |  |  |
| --- | --- | --- |
| Country | Region | Annual growth rate (%) |
| Qatar | Middle East | 4.2 |
| Zimbabwe | Africa | 0.4 |



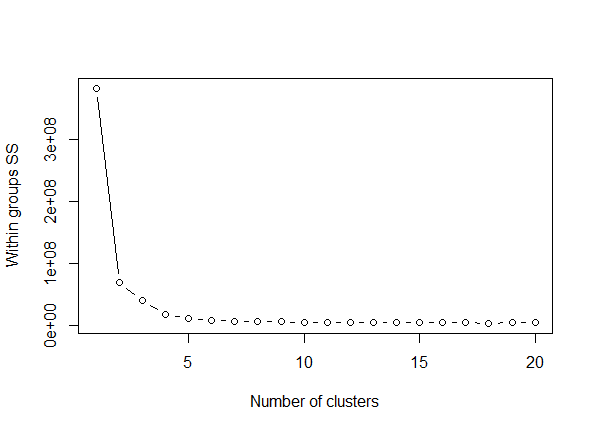
In the above chart, the life expectancy of Qatar, which is the best annual growth rate country in our data-set is clearly above 75 years while the life expectancy of Zimbabwe, which is the worst growth rate country is approximately 45 years. The above chart data supports out claim that annual growth rate influences life expectancy.

Statistically speaking, the p-value of the interaction term between Middle East region and annual growth rate was 0.014 (lesser than 0.05) which shows us how significant annual growth rate was in this model. Also anova test resulted in a p-value lesser than 0.05 when two models, one with annual growth rate and one without annual growth rate were compared.

As a proof from the literature supporting this claim, the book ‘A Continental Overview of Environmental Issues’ written by Kevin Hillstrom and Laurie Collier Hillstrom discusses on how life expectancy saw an increase in Middle East during the early 2000s. According to the book, Middle East was among the fastest growing regions in the World due to the presence of World’s largest oil reserves. As a result of this, there was notable increase in the standard of living, infrastructure, health and education.

# Why is the life expectancy of people in countries who spend more on health, lesser than that of people in countries who spend less on health?

To answer this research question, I ran a cluster analysis to group countries that spend more and less on their health. The below graph helped us to determine the optimal k value which was 4.



The mean values of significant variables in the under each of the clusters are shown below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Life.Exp | Per.capita.expenditure.on.health | Over.60 | Annual.growth.rate | Living.in.urban | Total.fertility.rate.per.woman. | Adolescent.fertility.rate | Longitude |
| 1 | 74.03 | 916.10 | 13.53 | 0.73 | 68.00 | 1.97 | 29.90 | 16.87 |
| 2 | 64.09 | 146.09 | 7.79 | 1.63 | 47.65 | 3.42 | 70.50 | 21.48 |
| 3 | **80.53** | **3292.79** | 21.21 | 0.86 | 78.95 | 1.65 | **11.68** | 26.90 |
| 4 | **80.29** | **5992.29** | 19.71 | 0.66 | 84.57 | 1.80 | **20.14** | -10.49 |

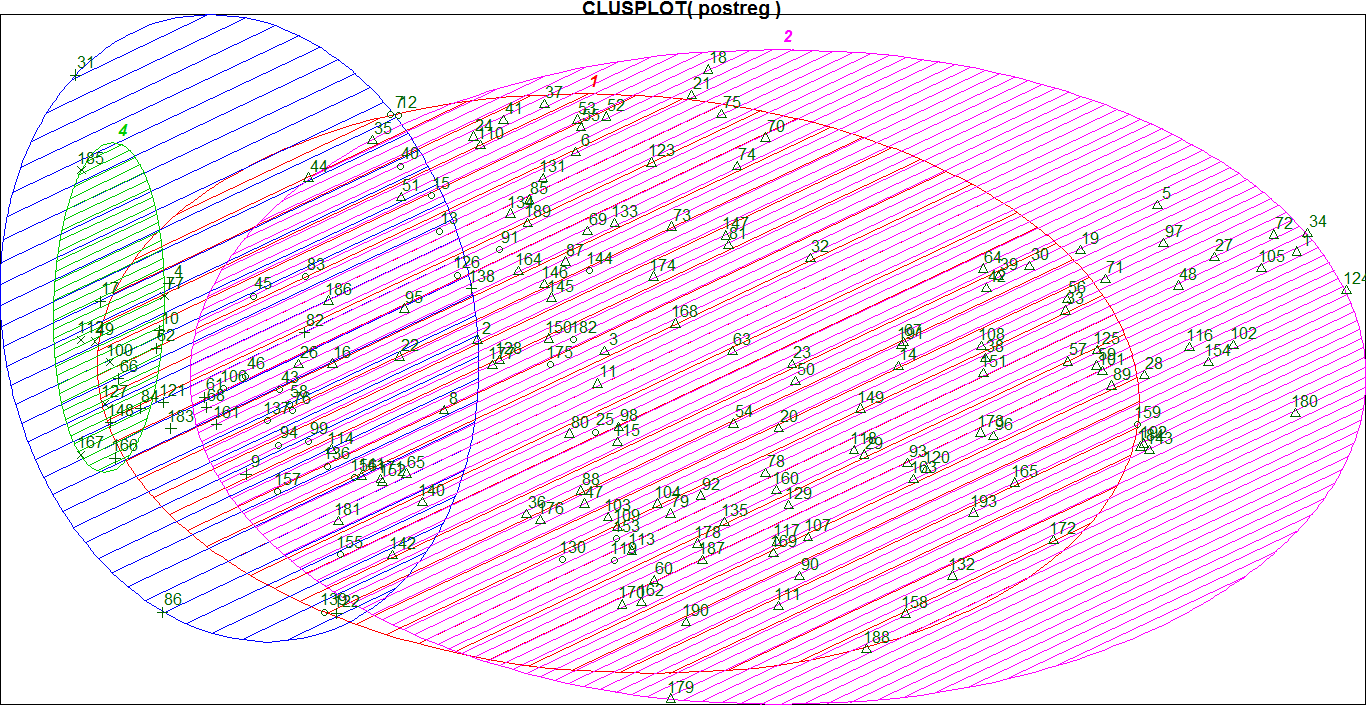
Clusters 1 and 2 are quite dispersed among each other and also with the other clusters.

What interested us was the disparity of life expectancy between cluster 3 and cluster 4. It is evident that there is a very minimum difference among each other in-terms of life expectancy even-though the per capita expenditure on health is very highly dispersed. Researching further to understand why life expectancy is lesser despite high spending on health, I noticed high mean disparity in Adolescent fertility rate between cluster 3 and 4.

High adolescent fertility rate indicates that the number of births among women whose ages are between 15 and 19 are high. Upon trying to understand how this factor influences life expectancy, an article on adolescent fertility stated that, the chances of a teenage mother dying is twice as high as that of a woman who waited until her 20s to begin childbearing.

Also cluster 4 had the US as one of the countries where adolescent fertility rate is high and this helped us to understand the reason behind why life expectancy is less in countries where people spend more on their health. So the implication here is that not only does economic factors influence life expectancy, social factors such as fertility during adolescence also have a huge influence on life expectancy of people.

The below shown cluster plot shows how data points are grouped and we can see that they are overlapping each other which implies us that the data-set considered is not highly dispersed.



# Conclusion

In this project, we started with WHO health data combining with geographic location data from statvision.com to investigate the contributing factors to the life expectancy worldwide by statistics tool. From the analysis results, I further investigated and found out that

* Middle East is among the fastest growing regions in the world and has notable increase in the standard of living, infrastructure, health and education, which is very favourable to immigrants, and boost annual population growth rate. It has positive impact on the life expectancy in the Middle East.
* Cluster analysis shows that the reason behind that two clusters with similar highest life expectancies but quite different amount of health spending may due to factors beyond the economic ones, such as the social factor fertility during adolescence.

# References

* World Health Statistics 2009 - World Health Organization <http://www.who.int/whosis/whostat/EN_WHS09_Tables.xls>
* Countries of The World.xls   
  [www.statvision.com/webinars/countries%20of%20the%20world.xls](http://www.statvision.com/webinars/countries%20of%20the%20world.xls)
* OECD Health Statistics 2014: How does the United States compare? <http://www.oecd.org/unitedstates/Briefing-Note-UNITED-STATES-2014.pdf>
* OECD/WHO (2012), “Life expectancy at birth”, in Health at a Glance: Asia/Pacific 2012, OECD Publishing.
* http://dx.doi.org/10.1787/9789264183902-4-en
* Russian Life Expectancy <https://upload.wikimedia.org/wikipedia/commons/d/d4/Russian_male_and_female_life_expectancy.PNG>
* WORLDHEALTHRANKINGS LIVE LONGER LIVE BETTER   
  <http://www.worldlifeexpectancy.com/cause-of-death/all-cancers/by-country/>
* Rising Health Care Costs   
  <https://www.ahip.org/Issues/Rising-Health-Care-Costs.aspx>
* Why U.S. Health Care Is Obscenely Expensive, In 12 Charts <http://www.huffingtonpost.com/2013/10/03/health-care-costs-_n_3998425.html>
* The Urban-Rural Divide   
  <https://www.mtholyoke.edu/~koyam20m/Urbanruraldivide.html>
* Life Expectancy Gap Between City And Country Folk Widens: How Access To Health Care Plays A Part <http://www.medicaldaily.com/life-expectancy-gap-between-city-and-country-folk-widens-how-access-health-care-plays-part-267929>

# Appendix

R – Code:

file.choose()

DMproj <- read.csv("C:\\Users\\Roshan\\Desktop\\MSM\\Winter '16\\Data Mining\\Project\\my proj.csv", header = TRUE)

str(DMproj)

# To check high correlation

DMproj.cor <- subset(DMproj, select = -c(1, 2))

cor(DMproj.cor)

str(DMproj.cor)

economic.var.names <- c(1:20)

pairs(DMproj.cor[,economic.var.names])

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)

{

usr <- par("usr"); on.exit(par(usr))

par(usr = c(0, 1, 0, 1))

r <- abs(cor(x, y))

txt <- format(c(r, 0.123456789), digits = digits)[1]

txt <- paste0(prefix, txt)

if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)

text(0.5, 0.5, txt, cex = pmax(1, cex.cor \* r))

}

pairs(DMproj.cor[,economic.var.names], lower.panel = panel.cor)

# Regression with all the variables except State

proj.lm1 <- lm(Life.Exp ~ . - State, data = DMproj)

summary(proj.lm1)

# Removing the most insignificant from proj.lm1 - DMproj$Tobacco.use.15.....1

proj.lm2 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1, data = DMproj)

summary(proj.lm2)

# Removing the most insignificant from proj.lm2 - Literacy.rate....

proj.lm3 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate...., data = DMproj)

summary(proj.lm3)

# Removing the most insignificant from proj.lm3 - Total.Population...000s.

proj.lm4 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s., data = DMproj)

summary(proj.lm4)

# Removing the most insignificant from proj.lm3 - Total.Population...000s.

proj.lm4 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s., data = DMproj)

summary(proj.lm4)

# Removing the most insignificant from proj.lm4 - National.income.per.capita....

proj.lm5 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita...., data = DMproj)

summary(proj.lm5)

# Removing the most insignificant from proj.lm5 - Beds..per.10k.

proj.lm6 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - DMproj$Beds..per.10k., data = DMproj)

summary(proj.lm6)

# Removing the most insignificant from proj.lm6 - Beds..per.10k.

proj.lm7 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k., data = DMproj)

summary(proj.lm7)

# Removing the most insignificant from proj.lm7 - Expenditure.on.health....of.GDP.

proj.lm8 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP., data = DMproj)

summary(proj.lm8)

# Removing the most insignificant from proj.lm8 - Nursing.desnity..per.10k.

proj.lm9 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k., data = DMproj)

summary(proj.lm9)

# Removing the most insignificant from proj.lm9 - Access.to.water

proj.lm10 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water, data = DMproj)

summary(proj.lm10)

# Removing the most insignificant from proj.lm10 - Median.age

proj.lm11 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age, data = DMproj)

summary(proj.lm11)

# Removing the most insignificant from proj.lm11 - Tobacco.use.15....

proj.lm12 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age - Tobacco.use.15...., data = DMproj)

summary(proj.lm12)

# Removing the most insignificant from proj.lm12 - Physicians.Density..per.10k.

proj.lm13 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age - Tobacco.use.15.... - Physicians.Density..per.10k., data = DMproj)

summary(proj.lm13)

# Removing the most insignificant from proj.lm13 - Access.to.sanitation...

proj.lm14 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age - Tobacco.use.15.... - Physicians.Density..per.10k. - Access.to.sanitation..., data = DMproj)

summary(proj.lm14)

# Removing the most insignificant from proj.lm14 - Latitude

proj.lm15 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age - Tobacco.use.15.... - Physicians.Density..per.10k. - Access.to.sanitation... - Latitude, data = DMproj)

summary(proj.lm15)

# Removing the most insignificant from proj.lm15 - Alcohol.consumption.Ls.person.year.

proj.lm16 <- lm(Life.Exp ~ . - State - Tobacco.use.15.....1 - Literacy.rate.... - Total.Population...000s. - National.income.per.capita.... - Beds..per.10k. - Expenditure.on.health....of.GDP. - Nursing.desnity..per.10k. - Access.to.water - Median.age - Tobacco.use.15.... - Physicians.Density..per.10k. - Access.to.sanitation... - Latitude - Alcohol.consumption.Ls.person.year., data = DMproj)

summary(proj.lm16)

# After excluding all insignificant variables, reading the new file with just the significant variables.

postreg <- read.csv("C:\\Users\\Roshan\\Desktop\\MSM\\Winter '16\\Data Mining\\Project\\proj\_postreg.csv", header = TRUE)

str(postreg)

# Cluster analysis

postreg$State <- NULL

postreg$Region <- NULL

# Determining optimal number of k

wss <- (nrow(postreg)-1)\*sum(apply(postreg,2,var))

for (i in 2:20) wss[i] <- sum(kmeans(postreg, centers = i)$withinss)

plot(1:20, wss, type = "b", xlab = "Number of clusters", ylab = "Within groups SS")

# kmeans

kmeans.proj <- kmeans(postreg, 4)

kmeans.proj

# Clus-Plot

library(cluster)

clusplot(postreg, kmeans.proj$cluster, color = TRUE, shade = TRUE, labels = 2, lines = 0)

table.with.clusters <- cbind(postreg, kmeans.proj$cluster)

write.csv(table.with.clusters, file = "proj cluster")

# Annual growth rate interaction term

proj.lm17 <- lm(DMproj$Life.Exp ~ DMproj$Region \* DMproj$Annual.growth.rate.... )

summary(proj.lm17)

# ANOVA test

anova(update(proj.lm16, . ~ . - Annual.growth.rate....), update(proj.lm16, . ~ .), test = "Chisq")