Project Proposal

Sumukh.Ramagiri And Anirudh.Sunkara

2024-03-01

library(readr)  
Life\_Expectancy <- read\_csv("Life\_Expectancy\_Data.csv")

## Rows: 1649 Columns: 22  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): Country, Status  
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, pe...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#View(Life\_Expectancy)  
str(Life\_Expectancy)

## spc\_tbl\_ [1,649 × 22] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Country : chr [1:1649] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
## $ Year : num [1:1649] 2015 2014 2013 2012 2011 ...  
## $ Status : chr [1:1649] "Developing" "Developing" "Developing" "Developing" ...  
## $ Life expectancy : num [1:1649] 65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...  
## $ Adult Mortality : num [1:1649] 263 271 268 272 275 279 281 287 295 295 ...  
## $ infant deaths : num [1:1649] 62 64 66 69 71 74 77 80 82 84 ...  
## $ Alcohol : num [1:1649] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...  
## $ percentage expenditure : num [1:1649] 71.3 73.5 73.2 78.2 7.1 ...  
## $ Hepatitis B : num [1:1649] 65 62 64 67 68 66 63 64 63 64 ...  
## $ Measles : num [1:1649] 1154 492 430 2787 3013 ...  
## $ BMI : num [1:1649] 19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...  
## $ under-five deaths : num [1:1649] 83 86 89 93 97 102 106 110 113 116 ...  
## $ Polio : num [1:1649] 6 58 62 67 68 66 63 64 63 58 ...  
## $ Total expenditure : num [1:1649] 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...  
## $ Diphtheria : num [1:1649] 65 62 64 67 68 66 63 64 63 58 ...  
## $ HIV/AIDS : num [1:1649] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...  
## $ GDP : num [1:1649] 584.3 612.7 631.7 670 63.5 ...  
## $ Population : num [1:1649] 33736494 327582 31731688 3696958 2978599 ...  
## $ thinness 1-19 years : num [1:1649] 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...  
## $ thinness 5-9 years : num [1:1649] 17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...  
## $ Income composition of resources: num [1:1649] 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...  
## $ Schooling : num [1:1649] 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Country = col\_character(),  
## .. Year = col\_double(),  
## .. Status = col\_character(),  
## .. `Life expectancy` = col\_double(),  
## .. `Adult Mortality` = col\_double(),  
## .. `infant deaths` = col\_double(),  
## .. Alcohol = col\_double(),  
## .. `percentage expenditure` = col\_double(),  
## .. `Hepatitis B` = col\_double(),  
## .. Measles = col\_double(),  
## .. BMI = col\_double(),  
## .. `under-five deaths` = col\_double(),  
## .. Polio = col\_double(),  
## .. `Total expenditure` = col\_double(),  
## .. Diphtheria = col\_double(),  
## .. `HIV/AIDS` = col\_double(),  
## .. GDP = col\_double(),  
## .. Population = col\_double(),  
## .. `thinness 1-19 years` = col\_double(),  
## .. `thinness 5-9 years` = col\_double(),  
## .. `Income composition of resources` = col\_double(),  
## .. Schooling = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## Attaching package: 'dplyr'  
##   
## The following objects are masked from 'package:stats':  
##   
## filter, lag  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(ggplot2)  
  
#data cleaning  
  
Life\_Expectancy <- Life\_Expectancy %>%  
 mutate(across(where(is.numeric), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))   
# Inspecting the dataset  
# by viewing the first few rows of the dataset to get a quick glimpse of the data.  
head(Life\_Expectancy)

## # A tibble: 6 × 22  
## Country Year Status `Life expectancy` `Adult Mortality` `infant deaths`  
## <chr> <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 Afghanistan 2015 Develop… 65 263 62  
## 2 Afghanistan 2014 Develop… 59.9 271 64  
## 3 Afghanistan 2013 Develop… 59.9 268 66  
## 4 Afghanistan 2012 Develop… 59.5 272 69  
## 5 Afghanistan 2011 Develop… 59.2 275 71  
## 6 Afghanistan 2010 Develop… 58.8 279 74  
## # ℹ 16 more variables: Alcohol <dbl>, `percentage expenditure` <dbl>,  
## # `Hepatitis B` <dbl>, Measles <dbl>, BMI <dbl>, `under-five deaths` <dbl>,  
## # Polio <dbl>, `Total expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>,  
## # GDP <dbl>, Population <dbl>, `thinness 1-19 years` <dbl>,  
## # `thinness 5-9 years` <dbl>, `Income composition of resources` <dbl>,  
## # Schooling <dbl>

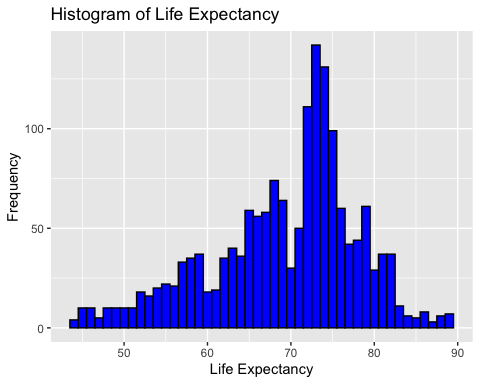
#Using the summary() function to get summary statistics of the dataset.   
#As it is useful for quickly checking the range, median, and presence of missing values in your numeric columns after cleaning.  
summary(Life\_Expectancy)

## Country Year Status Life expectancy  
## Length:1649 Min. :2000 Length:1649 Min. :44.0   
## Class :character 1st Qu.:2005 Class :character 1st Qu.:64.4   
## Mode :character Median :2008 Mode :character Median :71.7   
## Mean :2008 Mean :69.3   
## 3rd Qu.:2011 3rd Qu.:75.0   
## Max. :2015 Max. :89.0   
## Adult Mortality infant deaths Alcohol percentage expenditure  
## Min. : 1.0 Min. : 0.00 Min. : 0.010 Min. : 0.00   
## 1st Qu.: 77.0 1st Qu.: 1.00 1st Qu.: 0.810 1st Qu.: 37.44   
## Median :148.0 Median : 3.00 Median : 3.790 Median : 145.10   
## Mean :168.2 Mean : 32.55 Mean : 4.533 Mean : 698.97   
## 3rd Qu.:227.0 3rd Qu.: 22.00 3rd Qu.: 7.340 3rd Qu.: 509.39   
## Max. :723.0 Max. :1600.00 Max. :17.870 Max. :18961.35   
## Hepatitis B Measles BMI under-five deaths  
## Min. : 2.00 Min. : 0 Min. : 2.00 Min. : 0.00   
## 1st Qu.:74.00 1st Qu.: 0 1st Qu.:19.50 1st Qu.: 1.00   
## Median :89.00 Median : 15 Median :43.70 Median : 4.00   
## Mean :79.22 Mean : 2224 Mean :38.13 Mean : 44.22   
## 3rd Qu.:96.00 3rd Qu.: 373 3rd Qu.:55.80 3rd Qu.: 29.00   
## Max. :99.00 Max. :131441 Max. :77.10 Max. :2100.00   
## Polio Total expenditure Diphtheria HIV/AIDS   
## Min. : 3.00 Min. : 0.740 Min. : 2.00 Min. : 0.100   
## 1st Qu.:81.00 1st Qu.: 4.410 1st Qu.:82.00 1st Qu.: 0.100   
## Median :93.00 Median : 5.840 Median :92.00 Median : 0.100   
## Mean :83.56 Mean : 5.956 Mean :84.16 Mean : 1.984   
## 3rd Qu.:97.00 3rd Qu.: 7.470 3rd Qu.:97.00 3rd Qu.: 0.700   
## Max. :99.00 Max. :14.390 Max. :99.00 Max. :50.600   
## GDP Population thinness 1-19 years  
## Min. : 1.68 Min. :3.400e+01 Min. : 0.100   
## 1st Qu.: 462.15 1st Qu.:1.919e+05 1st Qu.: 1.600   
## Median : 1592.57 Median :1.420e+06 Median : 3.000   
## Mean : 5566.03 Mean :1.465e+07 Mean : 4.851   
## 3rd Qu.: 4718.51 3rd Qu.:7.659e+06 3rd Qu.: 7.100   
## Max. :119172.74 Max. :1.294e+09 Max. :27.200   
## thinness 5-9 years Income composition of resources Schooling   
## Min. : 0.100 Min. :0.0000 Min. : 4.20   
## 1st Qu.: 1.700 1st Qu.:0.5090 1st Qu.:10.30   
## Median : 3.200 Median :0.6730 Median :12.30   
## Mean : 4.908 Mean :0.6316 Mean :12.12   
## 3rd Qu.: 7.100 3rd Qu.:0.7510 3rd Qu.:14.00   
## Max. :28.200 Max. :0.9360 Max. :20.70

#In many cases we observe that there are missing values in a dataset. To check the data if there are any missing values we use the following command.  
  
sum(is.na(Life\_Expectancy))

## [1] 0

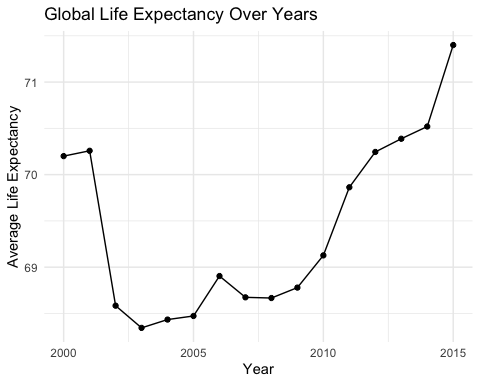
#Visualizing the data can help you spot any anomalies or issues that weren't apparent during the cleaning process.   
#And by plotting histograms of numeric variables can help us understand their distributions and identifying any outliers.  
  
# Histogram of life expectancy  
ggplot(Life\_Expectancy, aes(x = `Life expectancy`)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "Histogram of Life Expectancy", x = "Life Expectancy", y = "Frequency")



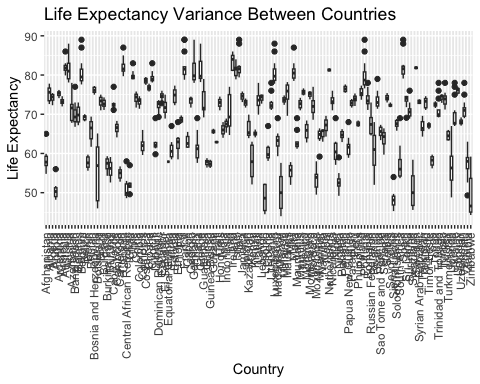
#from the histogram we observed that the data is mostly concentrated from the year 70 to 80.  
#But there is a possibility of outliers in the begining years of the data.  
  
  
#for identifying the outliers let's use the interquartile range (IQR)  
  
# Calculate the IQR for Life expectancy  
Q1 <- quantile(Life\_Expectancy$`Life expectancy`, 0.25, na.rm = TRUE)  
Q3 <- quantile(Life\_Expectancy$`Life expectancy`, 0.75, na.rm = TRUE)  
IQR <- Q3 - Q1  
  
# Define the lower and upper bounds for outliers  
lower\_bound <- Q1 - 1.5 \* IQR  
upper\_bound <- Q3 + 1.5 \* IQR  
  
# Identify outliers  
outliers <- subset(Life\_Expectancy, `Life expectancy` < lower\_bound | `Life expectancy` > upper\_bound)  
  
  
# View outliers  
print(outliers)

## # A tibble: 39 × 22  
## Country Year Status `Life expectancy` `Adult Mortality` `infant deaths`  
## <chr> <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 Angola 2007 Developing 48.2 375 87  
## 2 Botswana 2004 Developing 48.1 652 2  
## 3 Botswana 2003 Developing 46.4 693 2  
## 4 Botswana 2002 Developing 46 699 2  
## 5 Botswana 2001 Developing 46.7 679 2  
## 6 Botswana 2000 Developing 47.8 647 2  
## 7 Lesotho 2008 Developing 47.8 592 5  
## 8 Lesotho 2007 Developing 46.2 633 4  
## 9 Lesotho 2006 Developing 45.3 654 5  
## 10 Lesotho 2005 Developing 44.5 675 5  
## # ℹ 29 more rows  
## # ℹ 16 more variables: Alcohol <dbl>, `percentage expenditure` <dbl>,  
## # `Hepatitis B` <dbl>, Measles <dbl>, BMI <dbl>, `under-five deaths` <dbl>,  
## # Polio <dbl>, `Total expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>,  
## # GDP <dbl>, Population <dbl>, `thinness 1-19 years` <dbl>,  
## # `thinness 5-9 years` <dbl>, `Income composition of resources` <dbl>,  
## # Schooling <dbl>

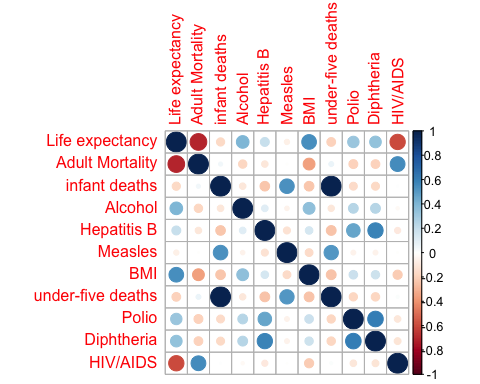
#Based on the values of the outliers, it appears that the outlier life expectancy values are mainly concentrated in the lower range (around 44.5 to 48.2 years), which is significantly below the global average.  
  
#Given the countries and the years mentioned (e.g., Angola, Botswana in the early 2000s), these low life expectancy figures could be genuine due to several factors:  
   
#HIV/AIDS Epidemic: The early 2000s were marked by a severe HIV/AIDS epidemic in Sub-Saharan Africa, which drastically reduced life expectancy in affected countries. The high HIV/AIDS values in the outliers support this.  
#Healthcare System: Developing countries often have healthcare systems that struggle with the burden of infectious diseases, malnutrition, and other health issues, impacting life expectancy.  
  
#As the outliers are genuine we can proceed with the further analysis of the dataset.  
  
#Global Trend in Life Expectancy Over the Years  
global\_trend <- Life\_Expectancy %>%  
 group\_by(Year) %>%  
 summarize(Average\_Life\_Expectancy = mean(`Life expectancy`, na.rm = TRUE))  
  
ggplot(global\_trend, aes(x = Year, y = Average\_Life\_Expectancy)) +  
 geom\_line() +  
 geom\_point() +  
 labs(title = "Global Life Expectancy Over Years", x = "Year", y = "Average Life Expectancy") +  
 theme\_minimal()



#Insight: As the trend shows a global increase in life expectancy, it suggests overall improvements in health outcomes. However, the rate of increase or any plateaus might indicate emerging challenges or diminishing returns on current healthcare strategies.  
#Investment Focus: Countries contributing to a slower rate of increase may need investment in healthcare infrastructure, technologies, and services to address specific health challenges.  
  
#Life Expectancy Variance Between Countries and Regions  
  
ggplot(Life\_Expectancy, aes(x = Country, y = `Life expectancy`)) +  
 geom\_boxplot() +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1)) +  
 labs(title = "Life Expectancy Variance Between Countries", x = "Country", y = "Life Expectancy")

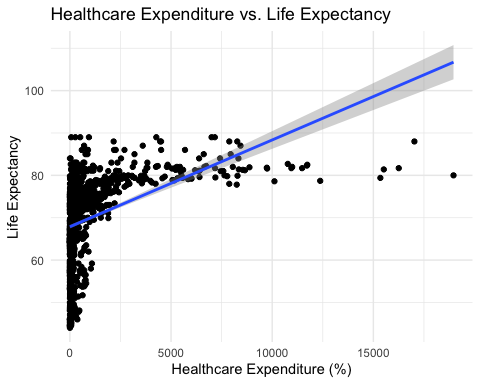


#Insight: Significant variance between countries highlights disparities in health outcomes. Countries with lower life expectancy or high variance within the country suggest inequalities in access to healthcare services, quality of care, or public health challenges.  
#Investment Focus: Countries with lower median life expectancy or wider interquartile ranges present opportunities for impactful investment. These regions may benefit from investments in basic healthcare infrastructure, access to essential medicines, vaccination programs, and public health initiatives.  
  
  
  
#Health-Related Factors Correlation with Life Expectancy  
health\_factors <- Life\_Expectancy %>%  
 select(`Life expectancy`, `Adult Mortality`, `infant deaths`, `Alcohol`, `Hepatitis B`, `Measles`, `BMI`, `under-five deaths`, `Polio`, `Diphtheria`, `HIV/AIDS`)  
  
correlation\_matrix <- cor(health\_factors, use = "complete.obs")  
corrplot::corrplot(correlation\_matrix, method = "circle")



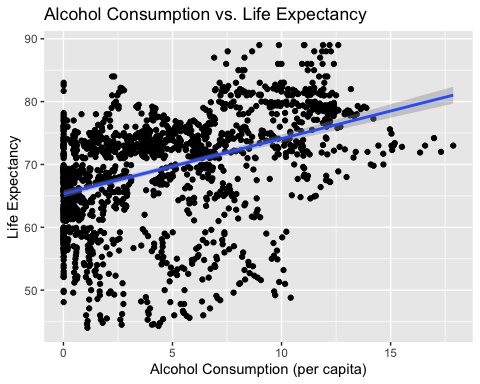
#Insight: Correlations between life expectancy and health-related factors (e.g., adult mortality, HIV/AIDS prevalence) highlight specific health issues that significantly impact longevity.  
#Investment Focus: Investing in countries with strong negative correlations between life expectancy and health challenges (like high adult mortality or high HIV/AIDS rates) can be strategic. Focus on comprehensive healthcare services, disease prevention programs, and education can address these specific challenges.  
  
  
  
  
#Efficiency of Healthcare Expenditure  
  
ggplot(Life\_Expectancy, aes(x = `percentage expenditure`, y = `Life expectancy`)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "Healthcare Expenditure vs. Life Expectancy", x = "Healthcare Expenditure (%)", y = "Life Expectancy") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'



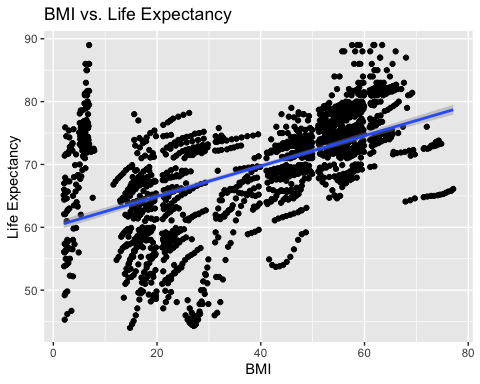
#Insight: The relationship between healthcare expenditure and life expectancy can indicate the efficiency of healthcare spending. A positive correlation suggests that investments in healthcare lead to better health outcomes.  
#Investment Focus: Countries with low life expectancy despite high healthcare expenditure may need investments aimed at increasing the efficiency of healthcare spending, such as through improving healthcare delivery systems, enhancing healthcare worker training, or investing in preventive healthcare.  
  
  
#Impact of Lifestyle Factors  
  
ggplot(Life\_Expectancy, aes(x = Alcohol, y = `Life expectancy`)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "Alcohol Consumption vs. Life Expectancy", x = "Alcohol Consumption (per capita)", y = "Life Expectancy")

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(Life\_Expectancy, aes(x = BMI, y = `Life expectancy`)) +  
 geom\_point() +  
 geom\_smooth(method = "lm") +  
 labs(title = "BMI vs. Life Expectancy", x = "BMI", y = "Life Expectancy")

## `geom\_smooth()` using formula = 'y ~ x'



#Insight: Lifestyle factors like alcohol consumption and BMI affecting life expectancy point to the importance of preventive healthcare and public health initiatives.  
#Investment Focus: Countries where lifestyle factors significantly impact life expectancy could benefit from investments in public health campaigns, lifestyle modification programs, and preventive healthcare services.  
  
  
  
#Predicting Life Expectancy  
model <- lm(`Life expectancy` ~ `Adult Mortality` + BMI + GDP + Schooling + `HIV/AIDS`, data = Life\_Expectancy)  
summary(model)

##   
## Call:  
## lm(formula = `Life expectancy` ~ `Adult Mortality` + BMI + GDP +   
## Schooling + `HIV/AIDS`, data = Life\_Expectancy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.8745 -2.2555 0.1766 2.4917 12.4315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.475e+01 5.661e-01 96.721 < 2e-16 \*\*\*  
## `Adult Mortality` -2.001e-02 1.001e-03 -19.986 < 2e-16 \*\*\*  
## BMI 4.836e-02 5.933e-03 8.151 7.09e-16 \*\*\*  
## GDP 7.934e-05 9.519e-06 8.335 < 2e-16 \*\*\*  
## Schooling 1.364e+00 4.655e-02 29.296 < 2e-16 \*\*\*  
## `HIV/AIDS` -4.511e-01 1.915e-02 -23.564 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.907 on 1643 degrees of freedom  
## Multiple R-squared: 0.8033, Adjusted R-squared: 0.8027   
## F-statistic: 1342 on 5 and 1643 DF, p-value: < 2.2e-16

#Insight:  
#Significant Predictors: The regression model indicates that adult mortality, BMI, GDP, schooling, and HIV/AIDS prevalence are significant predictors of life expectancy. This is evident from their low p-values (p < 0.05) and the confidence intervals not containing zero.  
#Direction of Influence: Adult mortality and HIV/AIDS prevalence have negative coefficients, indicating that higher levels of these factors are associated with lower life expectancy. Conversely, BMI, GDP, and schooling have positive coefficients, suggesting that higher levels of these factors are associated with higher life expectancy.  
#Model Performance: The multiple R-squared value of 0.8033 indicates that approximately 80.33% of the variability in life expectancy can be explained by the combination of these predictors.  
#Investment Focus:  
#Addressing Adult Mortality and HIV/AIDS Prevalence:  
#Insight: Adult mortality and HIV/AIDS have strong negative associations with life expectancy. Investments should focus on strategies to reduce mortality rates and combat the spread of HIV/AIDS.  
#Investment Focus: Healthcare organizations can invest in improving healthcare infrastructure, access to essential medications, and implementing prevention and treatment programs targeted at reducing adult mortality and controlling HIV/AIDS transmission.  
#Improving Socioeconomic Factors:  
#Insight: BMI, GDP, and schooling have positive associations with life expectancy. This suggests that improving socioeconomic factors can positively impact life expectancy.  
#Investment Focus: Investments should be directed towards initiatives aimed at improving education, economic development, and addressing poverty. This could include investing in education systems, vocational training programs, economic empowerment initiatives, and poverty alleviation measures.  
#Promoting Healthy Lifestyles:  
#Insight: BMI, a measure of body mass index, is positively associated with life expectancy. This highlights the importance of promoting healthy lifestyles and reducing obesity rates.  
#Investment Focus: Healthcare organizations can invest in public health campaigns, nutritional education programs, and initiatives aimed at promoting physical activity and healthy eating habits. Investing in facilities and programs that encourage physical activity and access to healthy food options can also be beneficial.  
#Supporting Economic Development:  
#Insight: GDP is positively associated with life expectancy, indicating that economic development plays a crucial role in improving health outcomes.  
#Investment Focus: Investments in economic development projects, infrastructure development, and job creation initiatives can indirectly contribute to improvements in health outcomes by lifting people out of poverty, improving living standards, and increasing access to healthcare services.  
#Enhancing Healthcare Infrastructure and Access:  
#Insight: While not directly included in the model, investments in healthcare infrastructure, access to healthcare services, and quality healthcare delivery are essential for improving overall health outcomes.  
#Investment Focus: Healthcare organizations can invest in building and upgrading healthcare facilities, training healthcare workers, expanding access to essential medicines, and implementing telemedicine and mobile health solutions to reach underserved populations.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.