Project 1

Aadil Noon

3/8/2020

```
#Data sets
library(ggplot2)
library(readxl)
library(dplyr)
```

```
##
## 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(cluster)
library(graphics)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x plyr::arrange()   masks dplyr::arrange()
## x purrr::compact()  masks plyr::compact()
## x plyr::count()     masks dplyr::count()
## x plyr::failwith()  masks dplyr::failwith()
## x dplyr::filter()   masks stats::filter()
## x plyr::id()        masks dplyr::id()
## x dplyr::lag()      masks stats::lag()
## x plyr::mutate()    masks dplyr::mutate()
## x plyr::rename()    masks dplyr::rename()
## x plyr::summarise() masks dplyr::summarise()
## x plyr::summarize() masks dplyr::summarize()
```

```
lifeexpec <- read_xlsx('Life expectancy (3).xlsx')
airpol <- read_xlsx('AirPollution (2).xlsx')
```

Introduction: The two datasets that I decided to take a look at are both from the world health association. One of them looks into the different levels of air pollution in different nations by looking at the concentration of fine particulate matter in the air. The second dataset looks at life expectancies of different countries. I chose these two becuase I am very interested in medicine and the intersections of politics, geography, and health outcomes. Although there are a host of variables which will influnce life expectancies I anticipate that the abundance of pollution will decrease life expectancies due to various health concerns, including respiratory disease.

```
#in order to make it easier to organize, I decided to remove the ranges from the data
airpol<-airpol%>%separate(`Total Concentrations of fine particulate matter (PM2.5)`,into=c("A","B"))
airpol<-airpol%>%unite(A, B, col="Total Concentrations of fine particulate matter (PM2.5)",sep=".")
airpol<-airpol%>%separate(`Urban Concentrations of fine particulate matter (PM2.5)`,into=c("A","B"))
airpol<-airpol%>%unite(A, B, col="Urban Concentrations of fine particulate matter (PM2.5)",sep=".")
airpol<-airpol%>%separate(`Rural Concentrations of fine particulate matter (PM2.5)`,into=c("A","B"))
airpol<-airpol%>%unite(A, B, col="Rural Concentrations of fine particulate matter (PM2.5)",sep=".")
#This got rid of the ranges but also split up the data by whole numbers and decimals, so I united them
airpol$`Total Concentrations of fine particulate matter (PM2.5)`<-as.numeric(airpol$`Total Concentratio
airpol$`Urban Concentrations of fine particulate matter (PM2.5)`<-as.numeric(airpol$`Urban Concentratio
airpol$`Rural Concentrations of fine particulate matter (PM2.5)`<-as.numeric(airpol$`Rural Concentratio
#I got rid of all the rows that were years other than 2016 so that I could compare the two datasets
lifeexpec<-lifeexpec%>%pivot_wider(names_from = 'Year',values_from = 'Country')
lifeexpec<-lifeexpec%>%select(-c('2000':'2015'))
lifeexpec<-lifeexpec%>%na.omit()
lifeexpec<-lifeexpec%>%rename(c('2016'='Country'))
```

```
#Joining
#Because I only wanted data for countries in both sets, I used inner_join
airpollifeexpec<-inner_join(lifeexpec,airpol,by="Country")
```

```
#Wrangling
#Decided to take a look at the relationship between pollution and life expectancy
airpollifeexpec%>%select(`Both Sexes' Life expectancy at birth in years`,`Total Concentrations of fine p
```

```
## # A tibble: 183 x 3
## # Groups:   Continent [6]
##    `Both Sexes' Life expectancy a~ `Total Concentrations of fine part~ Continent
##                              <dbl>                               <dbl> <chr>
##  1                            52.9                                27.8 Africa
##  2                            53                                  49.5 Africa
##  3                            53.1                                20.6 Africa
##  4                            54.3                                53   Africa
##  5                            54.6                                23.7 Africa
##  6                            55.2                                48.7 Africa
##  7                            55.4                                29.5 Africa
##  8                            57.7                                16.3 Africa
##  9                            58                                  31.2 Africa
## 10                            58.1                                65.3 Africa
## # ... with 173 more rows
```

```
#Made another collum to see which sex lives longer in which country
airpollifeexpec<-airpollifeexpec%>%mutate(`Sex with higher life expectancy at birth` =case_when(`Female
airpollifeexpec%>%summarize_all(n_distinct)
```

```
## # A tibble: 1 x 12
##   `Both Sexes' Li~ `Male Life expe~ `Female Life ex~ `Both Sexes' Li~
##              <int>            <int>            <int>            <int>
## 1              128              134              140               93
## # ... with 8 more variables: `Male Life expectancy at age 60 in years` <int>,
## #   `Female Life expectancy at age 60 in years` <int>, Country <int>, `Total
## #   Concentrations of fine particulate matter (PM2.5)` <int>, `Urban
## #   Concentrations of fine particulate matter (PM2.5)` <int>, `Rural
## #   Concentrations of fine particulate matter (PM2.5)` <int>, Continent <int>,
## #   `Sex with higher life expectancy at birth` <int>
```

```
#Looked at Asian Countries
airpollifeexpec %>% filter( Continent == "Asia")
```

```
## # A tibble: 48 x 12
##    `Both Sexes' Li~ `Male Life expe~ `Female Life ex~ `Both Sexes' Li~
##               <dbl>            <dbl>            <dbl>            <dbl>
##  1             62.7             61               64.5             16.3
##  2             74.8             71.2             78.1             19.6
##  3             73.1             70.3             75.7             18.9
##  4             79.1             78.6             79.6             21.7
##  5             72.7             71.1             74.4             19.6
##  6             70.6             70.4             70.8             20.7
##  7             76.4             75.3             77.6             20.4
##  8             69.4             67.3             71.2             17.4
##  9             76.4             75               77.9             19.9
## 10             80.7             78.4             83.1             22.8
## # ... with 38 more rows, and 8 more variables: `Male Life expectancy at age 60
```

```
## #   in years` <dbl>, `Female Life expectancy at age 60 in years` <dbl>,
## #   Country <chr>, `Total Concentrations of fine particulate matter
## #   (PM2.5)` <dbl>, `Urban Concentrations of fine particulate matter
## #   (PM2.5)` <dbl>, `Rural Concentrations of fine particulate matter
## #   (PM2.5)` <dbl>, Continent <chr>, `Sex with higher life expectancy at
## #   birth` <chr>
```

```
#summary of variables
summary(airpollifeexpec$`Both Sexes' Life expectancy at birth in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52.90   66.10   73.40   71.79   76.85   84.20
```

```
summary(airpollifeexpec$`Both Sexes' Life expectancy at age 60 in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.30   17.35   19.60   19.76   21.95   26.40
```

```
summary(airpollifeexpec$`Male Life expectancy at birth in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   63.80   70.30   69.44   74.70   81.20
```

```
summary(airpollifeexpec$`Female Life expectancy at birth in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.80   68.25   76.40   74.17   79.60   87.10
```

```
summary(airpollifeexpec$`Male Life expectancy at age 60 in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.30   15.95   17.60   18.33   20.40   24.30
```

```
summary(airpollifeexpec$`Female Life expectancy at age 60 in years`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.40   18.20   21.00   21.05   23.65   28.90
```

```
summary(airpollifeexpec$`Total Concentrations of fine particulate matter (PM2.5)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.70   14.45   21.00   25.60   32.90   94.30
```

```
summary(airpollifeexpec$`Urban Concentrations of fine particulate matter (PM2.5)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.80   14.90   21.50   26.16   32.95   99.50
```

```r
summary(airpollifeexpec$`Rural Concentrations of fine particulate matter (PM2.5)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.20   12.55   19.50   24.44   35.00   81.30
```

```r
#summary by grouping by Continent
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Both Sexes' Life expectancy at birth in y
```

```
##       mean       sd
## 1 71.78798 7.633225
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Both Sexes' Life expectancy at age 60 in
```

```
##       mean       sd
## 1 19.75519 2.994145
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Male Life expectancy at birth in years`,r
```

```
##      mean       sd
## 1 69.4377 7.361501
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Female Life expectancy at birth in years
```

```
##       mean       sd
## 1 74.16503 8.024932
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Male Life expectancy at age 60 in years`
```

```
##       mean       sd
## 1 18.33224 2.743624
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Female Life expectancy at age 60 in years
```

```
##       mean       sd
## 1 21.04699 3.347868
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Total Concentrations of fine particulate
```

```
##       mean       sd
## 1 25.59617 16.53754
```

```r
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Urban Concentrations of fine particulate
```

```
##       mean      sd
## 1 26.16393 16.8956
```

```
airpollifeexpec%>% group_by(Continent) %>%summarize(mean=mean(`Rural Concentrations of fine particulate
```

```
##      mean       sd
## 1 24.4388 15.94046
```

```
#Summary Grouping by Country
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Both Sexes' Life expectancy at birth in yea
```

```
##       mean       sd
## 1 71.78798 7.633225
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Both Sexes' Life expectancy at age 60 in y
```

```
##       mean       sd
## 1 19.75519 2.994145
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Male Life expectancy at birth in years`,na
```

```
##      mean       sd
## 1 69.4377 7.361501
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Female Life expectancy at birth in years`,
```

```
##       mean       sd
## 1 74.16503 8.024932
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Male Life expectancy at age 60 in years`,na
```

```
##       mean       sd
## 1 18.33224 2.743624
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Female Life expectancy at age 60 in years`
```

```
##       mean       sd
## 1 21.04699 3.347868
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Total Concentrations of fine particulate ma
```

```
##       mean       sd
## 1 25.59617 16.53754
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Urban Concentrations of fine particulate ma
```

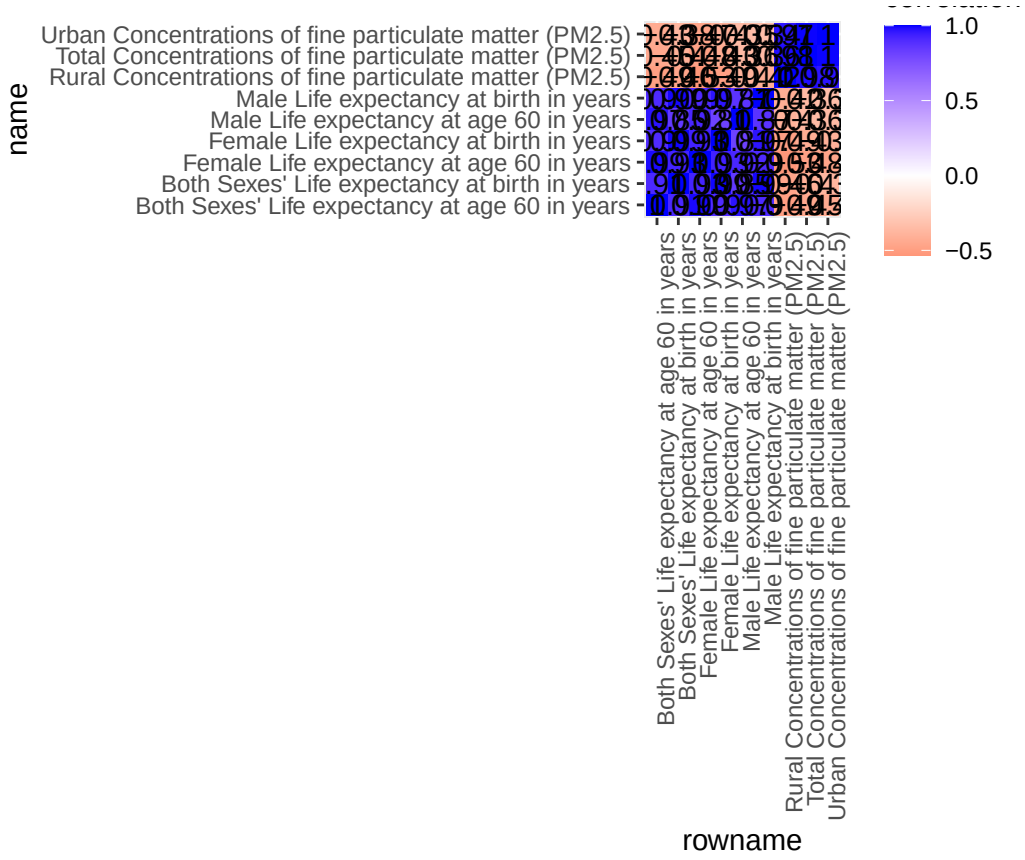
```
##       mean      sd
## 1 26.16393 16.8956
```

```
airpollifeexpec%>% group_by(Country) %>%summarize(mean=mean(`Rural Concentrations of fine particulate ma
```

```
##      mean       sd
## 1 24.4388 15.94046
```

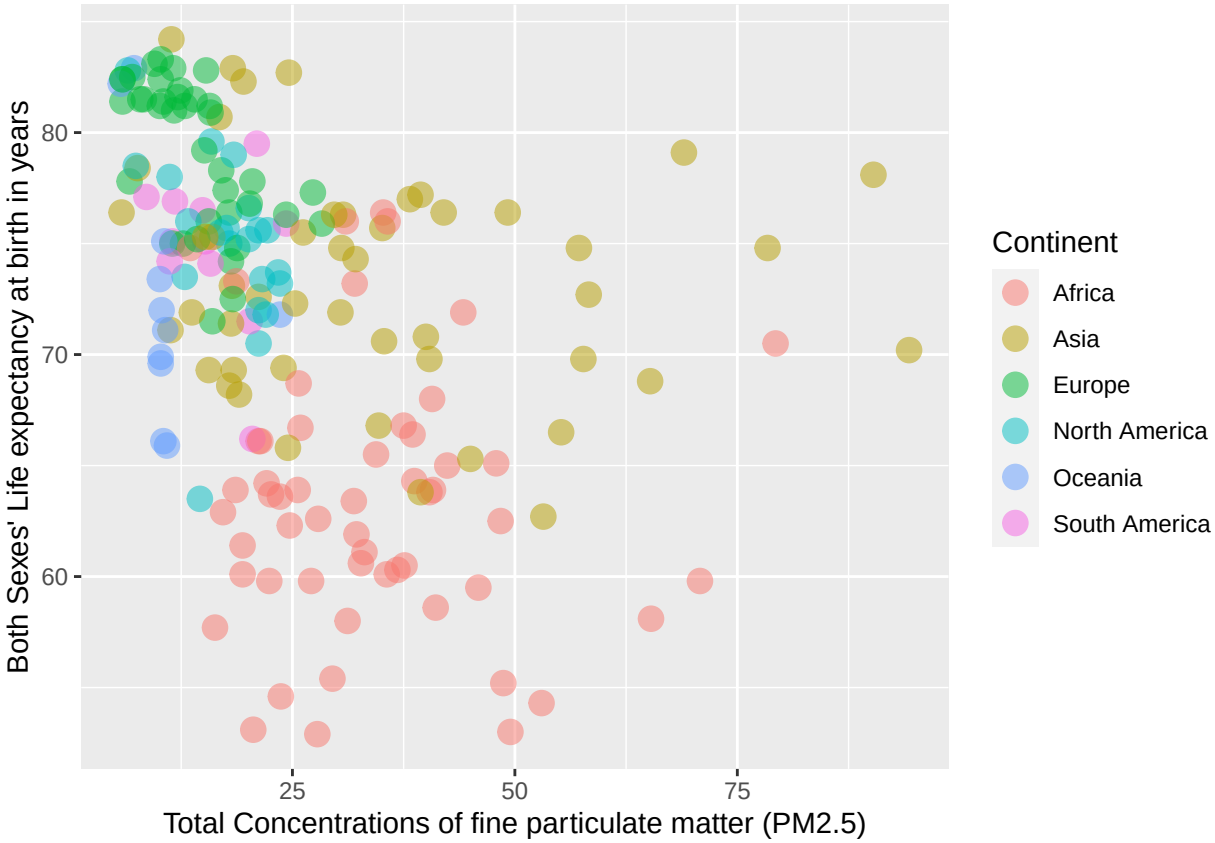
Some of the statistics I found include that the global average life expectancy at birth is 71.79 years. Furthermore, females tend to live longer than males, with expectancies of 74.14 and 69.44 years respectivaley.In addition, the global average concentration of fine particulate matter is 25.60. Urban areas have a higher concentration on average than rural, with 26.16 and 24.44 repesctivaley.

```
#Visualizing
df<-airpollifeexpec%>%na.omit%>%select_if(is.numeric)
cor<-cor(df)%>%as.data.frame%>%rownames_to_column%>%pivot_longer(-1,names_to="name",values_to="correlati
cor%>%ggplot(aes(rowname,name,fill=correlation))+geom_tile()+scale_fill_gradient2(low="red",mid="white"
```



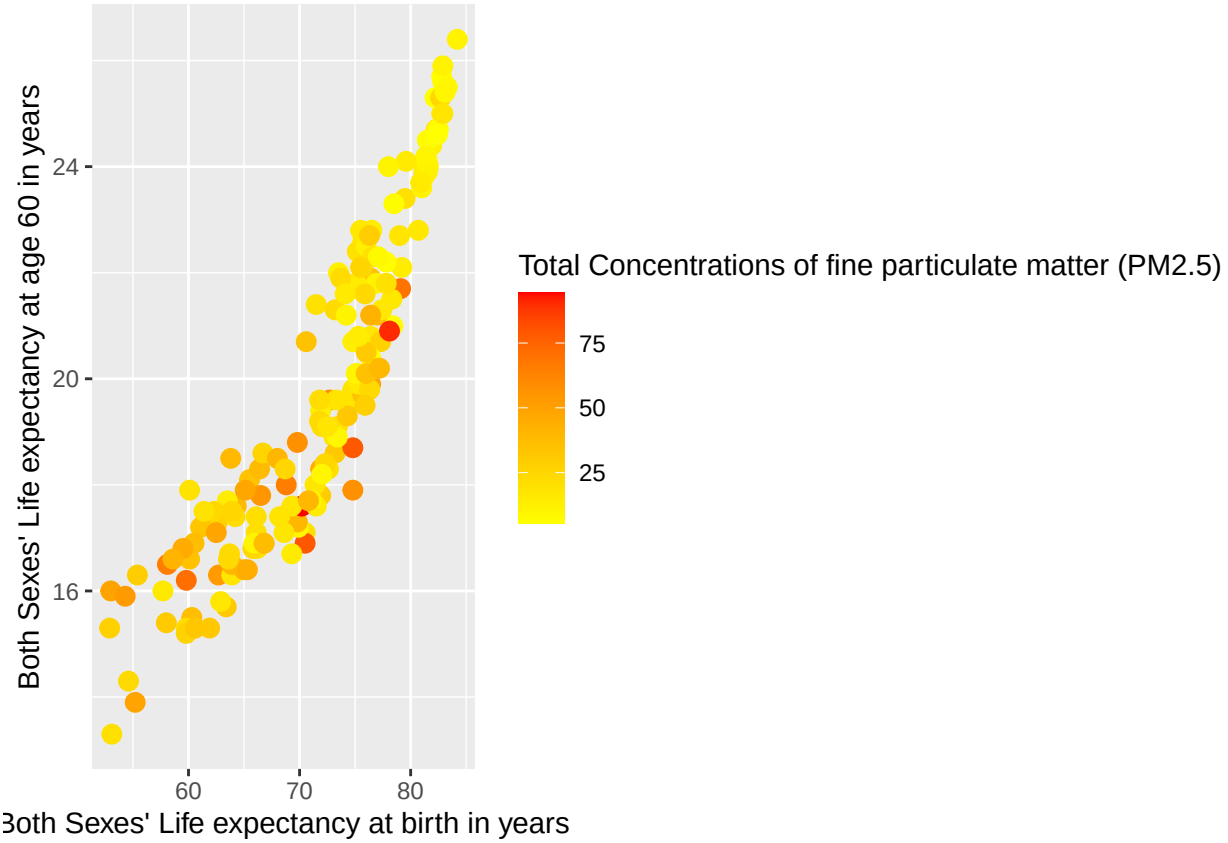
The correlation heatmap above indicates that there is a negative correlation between the concentration of fine particulate matter and life expectancy. There is around a -0.5 correlation between the variables looking at pollution levels and those looking at life expectancies. In addition, all 6 forms of life expectancy are very correlated with one another. This makes logical sense and indicates that countries which have a higher life expectancy at birth have a higher one at age 60 and countries tend to have similar patterns in regard to sex. The variables which measure life expectancy are very correlated, as are those that measure pollution and those two groups of variables have a negative correlation.

```
#Visualizing
ggplot(airpollifeexpec, aes(`Total Concentrations of fine particulate matter (PM2.5)`, `Both Sexes' Life
```



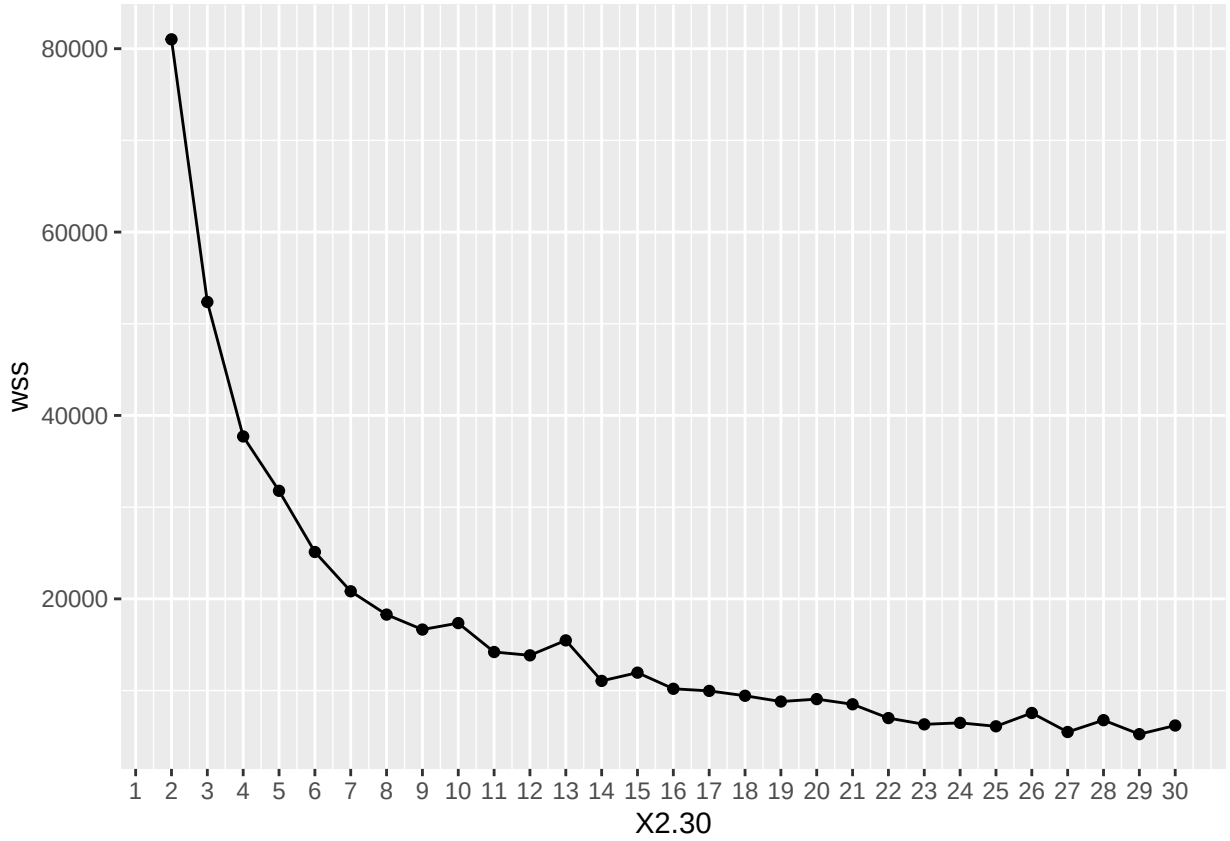
Firstly, the graph visually presents how lower concentrations of particulate matter are generally associated with longer life spans. It is also interesting how the majority of the countries are clustered with high life expectancies and lower pollution, which could indicate a possible bias in the dataset, because these kind of countries tend to be more industrially deveolped and are therfore more avaliable for research. Further, it also seems as though a lot of African countries tend to have lower life expectancies despite having similar levels of pollution. This speaks to the plethora of factors which impact how long one lives. Finally it appears as though Asian countries tend to be the most polluted.

```
#Visualizing
ggplot(airpollifeexpec, aes(`Both Sexes' Life expectancy at birth in years`,`Both Sexes' Life expectancy
```



Firstly, this graphs speaks to the strong correlation between life expectancy at birth and that at 60. Secondly, the graph further shows how the level of fine particulate matter is negatively correlated with life expectancy. This adds to the data above and further provides validity to the argument that the level of pollution has a directly negative effect on how long people in a given country live for.

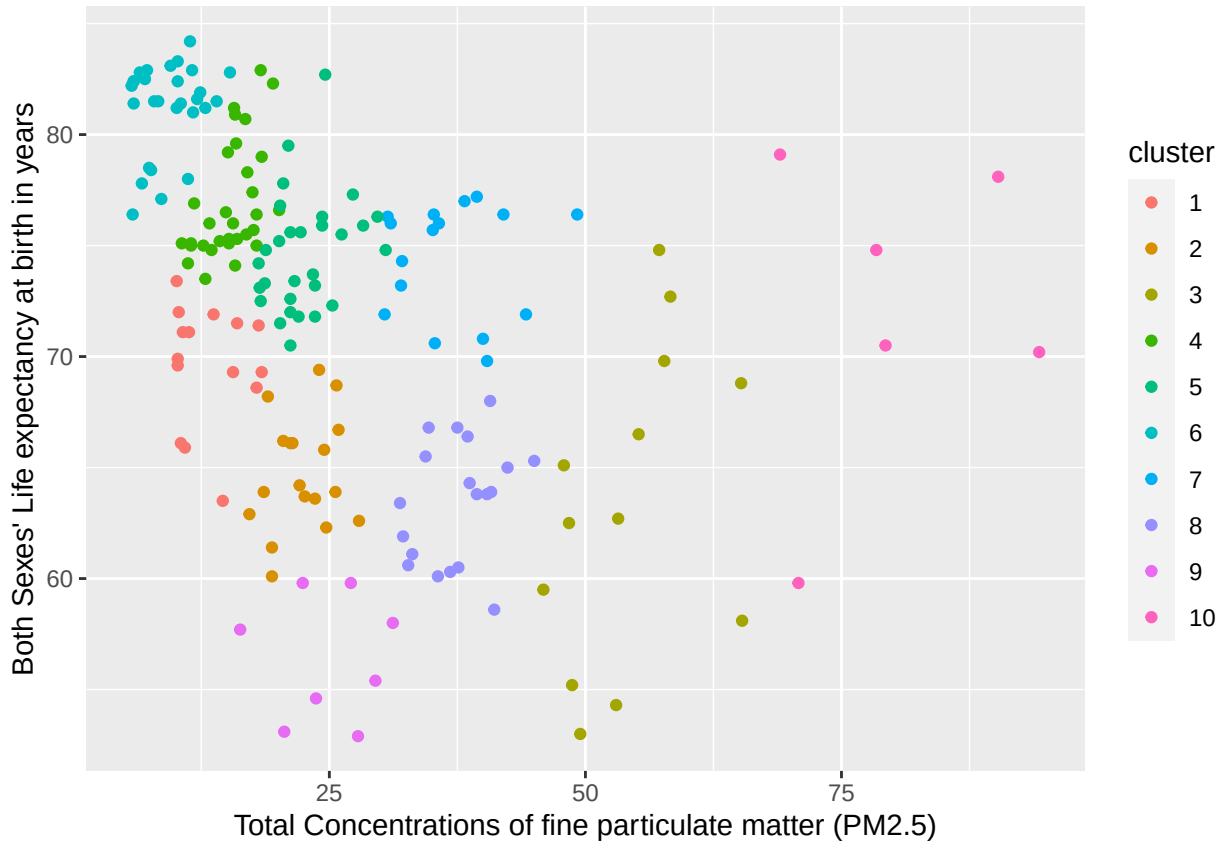
```
#Dimentionality Reduction
pc_cluster <-kmeans(df, 5)
kmean_withinss <- function(k) {
cluster <- kmeans(df, k)
return (cluster$tot.withinss) }
wss <- sapply(2:30, kmean_withinss)
elbow <-data.frame(2:30, wss)
ggplot(elbow, aes(x = X2.30, y = wss)) +geom_point() +geom_line() +scale_x_continuous(breaks = seq(1, 3
```



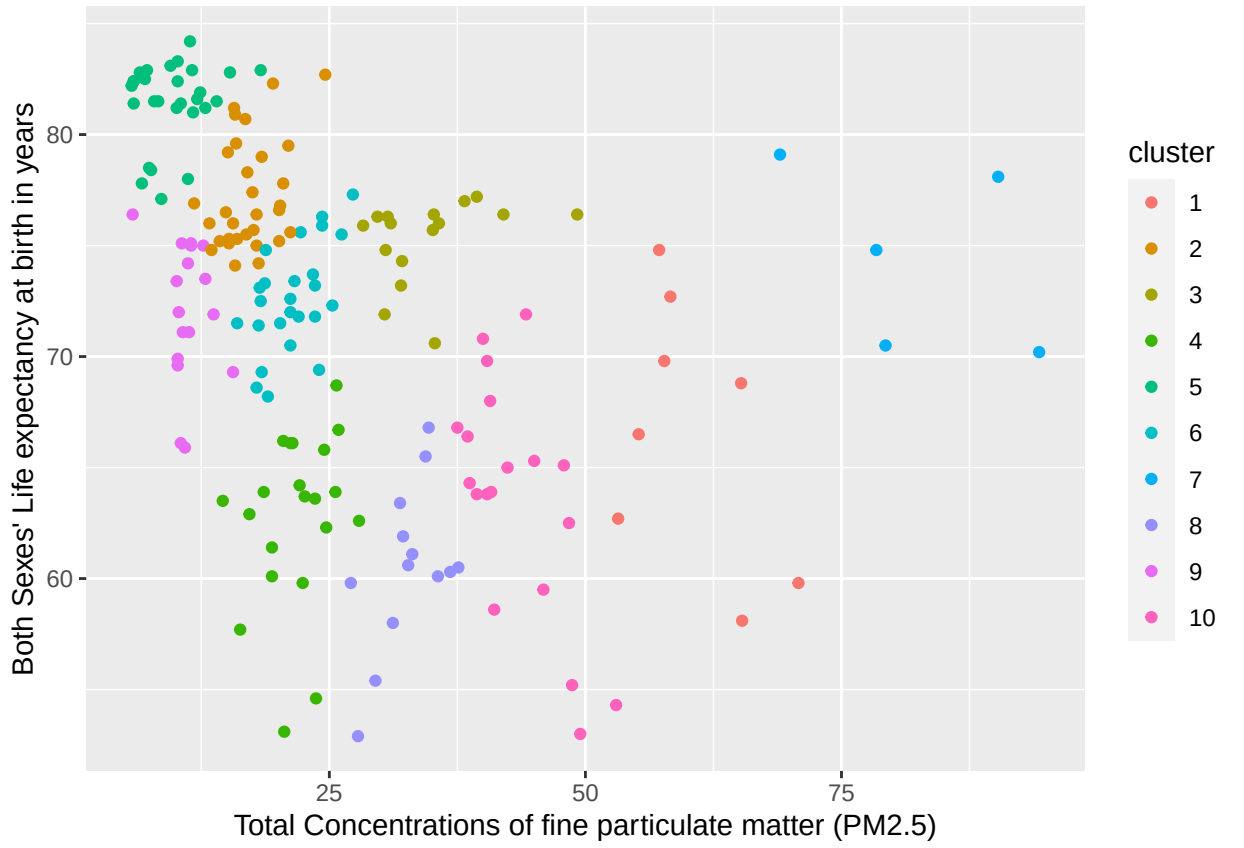

```
#according to elbow, curve has most dimishing return at 10
kmeans1 <-kmeans(df, 10)
kmeans1$size
```

```
##  [1] 15 18 13 31 29 28 16 19  8  6
```

```
#pretty good homegoneity between clusters
kmeansclust <- df %>% mutate(cluster=as.factor(kmeans1$cluster))
kmeansclust %>% ggplot(aes(`Total Concentrations of fine particulate matter (PM2.5)`,`Both Sexes' Life
```




```
pam1 <- df %>% pam(k=10)
pamclust<-df %>% mutate(cluster=as.factor(pam1$clustering))
pamclust %>% ggplot(aes(`Total Concentrations of fine particulate matter (PM2.5)`,`Both Sexes' Life exp
```




```
pamclust %>% group_by(cluster) %>% summarize_if(is.numeric,mean,na.rm=T)
```

```
## # A tibble: 10 x 10
##    cluster `Both Sexes' Li~ `Male Life expe~ `Female Life ex~ `Both Sexes' Li~
##    <fct>              <dbl>            <dbl>            <dbl>            <dbl>
##  1 1                   66.6             65.3             68.1             17.6
##  2 2                   77.3             74.3             80.2             22.0
##  3 3                   75.3             73.3             77.3             20.2
##  4 4                   62.7             60.8             64.6             16.8
##  5 5                   81.5             79.1             83.8             24.4
##  6 6                   72.6             69.5             75.8             19.3
##  7 7                   74.5             73.3             76.1             19.2
##  8 8                   60.5             59.1             61.8             16.1
##  9 9                   72.0             69.1             75.0             19.0
## 10 10                  63.6             61.7             65.6             17.1
## # ... with 5 more variables: `Male Life expectancy at age 60 in years` <dbl>,
## #   `Female Life expectancy at age 60 in years` <dbl>, `Total Concentrations of
## #   fine particulate matter (PM2.5)` <dbl>, `Urban Concentrations of fine
## #   particulate matter (PM2.5)` <dbl>, `Rural Concentrations of fine
## #   particulate matter (PM2.5)` <dbl>
```

The data did a good job of clustering into a group of 10. This could indicate that the previous way of labeling the countries by continent is limiting and is not a very accurate way of assesing a country's level of pollution or life expectancy. This speaks to the reductive nature of taking a diverse group of countries and putting them into a geographically assigned boundary which does not accurately reflect the political and cultural reality of the

country. I believe this is one of the reasons that the world health organiztion and various other organiztions have begin to look at countries through the more specific lens of regions as opposed to looking at them through the reductive framework of continents.