Homework_03

Akshaya Mahesh

10/21/2021

Installing packages

```
install.packages(c( "plyr", "dbplyr","usmap"))
```

Adding the necessary libraries

```
library(plyr)
library(tidyverse)
library(ggplot2)
library(readr)
library(dplyr)
library(usmap)
```

Part 1

Importing files and combining them

Dataset after binding rows

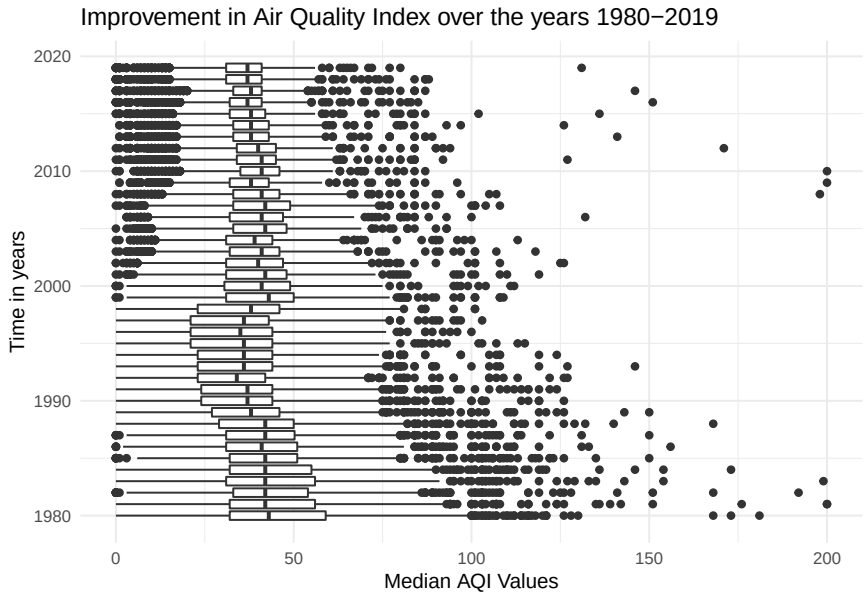
```
files <- list.files(path = "C:\\Users\\maksh\\OneDrive\\Documents\\Datasets\\epa-aqi-data-annual\\epa-aq
                    full.names = T)
data_all <- lapply(files, read.csv, header = TRUE, stringsAsFactors = FALSE)
air_aqi <- bind_rows(data_all)
head(air_aqi)
```

```
##     State     County Year Days.with.AQI Good.Days Moderate.Days
## 1 Alabama    Autauga 1980           179       122            35
## 2 Alabama    Colbert 1980           274       127            45
## 3 Alabama    Jackson 1980           366        85           110
## 4 Alabama  Jefferson 1980           343       171           109
## 5 Alabama Lauderdale 1980           274       120            58
## 6 Alabama    Madison 1980           344       154           125
##   Unhealthy.for.Sensitive.Groups.Days Unhealthy.Days Very.Unhealthy.Days
## 1                                  18              4                   0
## 2                                  63             39                   0
## 3                                  92             79                   0
## 4                                  37             19                   7
```

```
## 5                                  77             19                   0
## 6                                  60              5                   0
##   Hazardous.Days Max.AQI X90th.Percentile.AQI Median.AQI Days.CO Days.NO2
## 1              0     177                  108         40       0        0
## 2              0     200                  165         56       0        0
## 3              0     200                  200         94       0        0
## 4              0     221                  140         51     207        0
## 5              0     200                  139         56       0        0
## 6              0     185                  112         56      28      168
##   Days.Ozone Days.SO2 Days.PM2.5 Days.PM10
## 1        122       57          0         0
## 2          0      274          0         0
## 3          0      366          0         0
## 4        136        0          0         0
## 5          0      274          0         0
## 6        148        0          0         0
```

Including plots to show Median AQI Values Vs Time

```
ggplot(data=air_aqi,mapping = aes(y=Year,x=Median.AQI,group=Year))+geom_boxplot()+
  labs(title="Improvement in Air Quality Index over the years 1980-2019 ",
       x="Median AQI Values",y="Time in years")+theme_minimal()
```



Observation for the above plot:
1.The above plot shows that Air Quality Index in the United States has improved over the 40 year period.
2.The range of spread of Median AQI values has decreased over the years.
3.There is a decrease in the number of outliers having higher Median AQI values and increase in the number of outliers having very lower median AQI values.

Part 2

creating a new variable 'decade'

```
air_aqi$decade <-ifelse(air_aqi$Year>=1980 & air_aqi$Year<1990,
                        "1980-1989",
                        ifelse(air_aqi$Year>=1990 & air_aqi$Year<2000,
                        "1990-1999",
                        ifelse(air_aqi$Year>=2000 & air_aqi$Year<2010,
                        "2000-2009",
                        ifelse(air_aqi$Year>=2010 & air_aqi$Year<2020,"2010-2019","0"))))
head(air_aqi)
```

```
##     State     County Year Days.with.AQI Good.Days Moderate.Days
## 1 Alabama    Autauga 1980           179       122            35
## 2 Alabama    Colbert 1980           274       127            45
## 3 Alabama    Jackson 1980           366        85           110
## 4 Alabama  Jefferson 1980           343       171           109
## 5 Alabama Lauderdale 1980           274       120            58
## 6 Alabama    Madison 1980           344       154           125
##   Unhealthy.for.Sensitive.Groups.Days Unhealthy.Days Very.Unhealthy.Days
## 1                                  18              4                   0
## 2                                  63             39                   0
## 3                                  92             79                   0
## 4                                  37             19                   7
## 5                                  77             19                   0
## 6                                  60              5                   0
##   Hazardous.Days Max.AQI X90th.Percentile.AQI Median.AQI Days.CO Days.NO2
## 1              0     177                  108         40       0        0
## 2              0     200                  165         56       0        0
## 3              0     200                  200         94       0        0
## 4              0     221                  140         51     207        0
## 5              0     200                  139         56       0        0
## 6              0     185                  112         56      28      168
##   Days.Ozone Days.SO2 Days.PM2.5 Days.PM10    decade
## 1        122       57          0         0 1980-1989
## 2          0      274          0         0 1980-1989
## 3          0      366          0         0 1980-1989
## 4        136        0          0         0 1980-1989
## 5          0      274          0         0 1980-1989
## 6        148        0          0         0 1980-1989
```

Selecting the State, Year,Decade and Median AQI columns for observation

```
air_aqi_decade<-air_aqi%>%select(c("State","Year","Median.AQI","decade"))
head(air_aqi_decade)
```

```
##     State Year Median.AQI    decade
## 1 Alabama 1980         40 1980-1989
## 2 Alabama 1980         56 1980-1989
## 3 Alabama 1980         94 1980-1989
## 4 Alabama 1980         51 1980-1989
## 5 Alabama 1980         56 1980-1989
## 6 Alabama 1980         56 1980-1989
```

Including plots for the Average Median AQI values across the states for four decades(1980-2019)

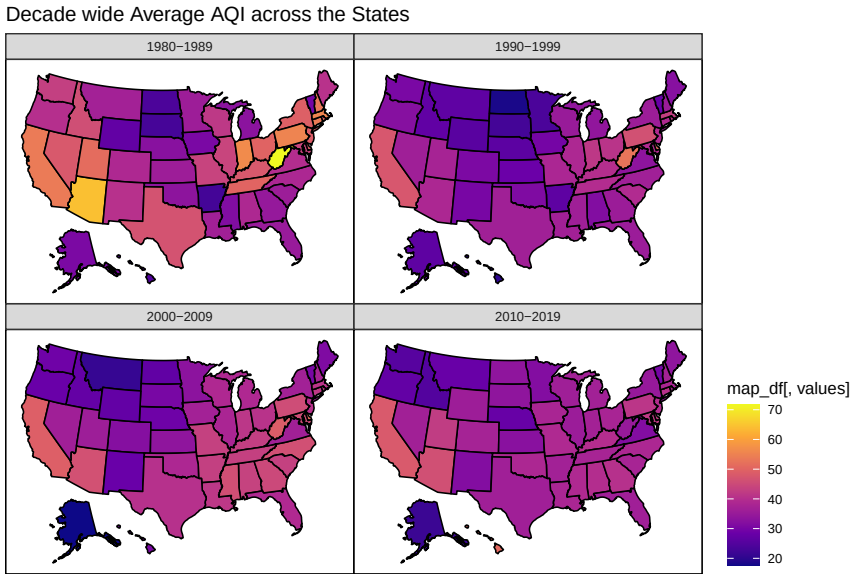
```
my_aqi<-ddply(air_aqi_decade, .(decade,State),summarise,
              Average_MedianAQI=mean(`Median.AQI`,na.rm=TRUE))
my_aqi$region <- tolower(my_aqi$State)
head(my_aqi)
```

```
##      decade      State Average_MedianAQI     region
## 1 1980-1989    Alabama          38.65686    alabama
## 2 1980-1989     Alaska          30.70000     alaska
## 3 1980-1989    Arizona          64.26923    arizona
## 4 1980-1989   Arkansas          22.95833   arkansas
## 5 1980-1989 California          53.69526 california
## 6 1980-1989   Colorado          38.77333   colorado
```

Using Fips to plot the data

```
my_aqi$fips <- fips(my_aqi$region)

plot_usmap(data =my_aqi, values = "Average_MedianAQI", labels=FALSE)+
  scale_fill_viridis_c(option = "plasma")+theme(legend.position = "right")+
  theme(panel.background = element_rect(colour = "black")) +
  labs(title = "Decade wide Average AQI across the States")+
  facet_wrap(~decade)
```



Observation for the above plot:
1. Air Quality Index seems to have improved over the decades across the United States.
2. States such as Arizona, West Virginia and Pennsylvania show great improvement in Air Quality Index over the period.
3. On the contrary, Air Quality index in North Dakota,South Dakota, Arkansas and Hawaii has increased

over the decades.
4. Florida seems to have had a nearly constant Average AQI.
5. In 1980-1989 , the states appear to have a wide range of Air AQI(comprises of both lower and extreme values) whereas in 2010-2019, the states appear to have a uniform AQI in the range of 30-50 (exception being California and Arizona).

Part 3

Importing the country and world_4region file

```
dir1 <- "C:\\Users\\maksh\\OneDrive\\Documents\\Datasets"
dir2 <- "ddf--gapminder--systema_globalis-master\\ddf--gapminder--systema_globalis-master"
path <- file.path(dir1, dir2,"ddf--entities--geo--country.csv" )
countries <- read_csv(path)
head(countries)
```

```
## # A tibble: 6 x 21
##   country   g77_and_oecd_countries income_3groups income_groups    `is--country`
##   <chr>     <chr>                  <chr>          <chr>            <lgl>
## 1 abkh      others                 <NA>           <NA>             TRUE
## 2 abw       others                 high_income    high_income      TRUE
## 3 afg       g77                    low_income     low_income       TRUE
## 4 ago       g77                    middle_income  lower_middle_in~ TRUE
## 5 aia       others                 <NA>           <NA>             TRUE
## 6 akr_a_dhe others                 <NA>           <NA>             TRUE
## # ... with 16 more variables: iso3166_1_alpha2 <chr>, iso3166_1_alpha3 <chr>,
## #   iso3166_1_numeric <dbl>, iso3166_2 <chr>, landlocked <chr>, latitude <dbl>,
## #   longitude <dbl>, main_religion_2008 <chr>, name <chr>, un_sdg_ldc <chr>,
## #   un_sdg_region <chr>, un_state <lgl>, unicef_region <chr>,
## #   unicode_region_subtag <chr>, world_4region <chr>, world_6region <chr>
```

```
path<-file.path(dir1, dir2, "ddf--entities--geo--world_4region.csv")
world_region <- read_csv(path)

head(world_region)
```

```
## # A tibble: 4 x 11
##   world_4region color   description    `is--world_4reg~ latitude longitude name
##   <chr>         <chr>   <chr>          <lgl>               <dbl>     <dbl> <chr>
## 1 africa        #00d5e9 The entire Af~ TRUE              -14.3       28.5 Afri~
## 2 americas      #7feb00 North, South ~ TRUE                8.99     -79.5 The ~
## 3 asia          #ff5872 Asia as defin~ TRUE               16.2      108.  Asia
## 4 europe        #ffe700 West & East E~ TRUE               50.8        4.5 Euro~
## # ... with 4 more variables: name_long <chr>, name_short <chr>, rank <dbl>,
## #   shape_lores_svg <chr>
```

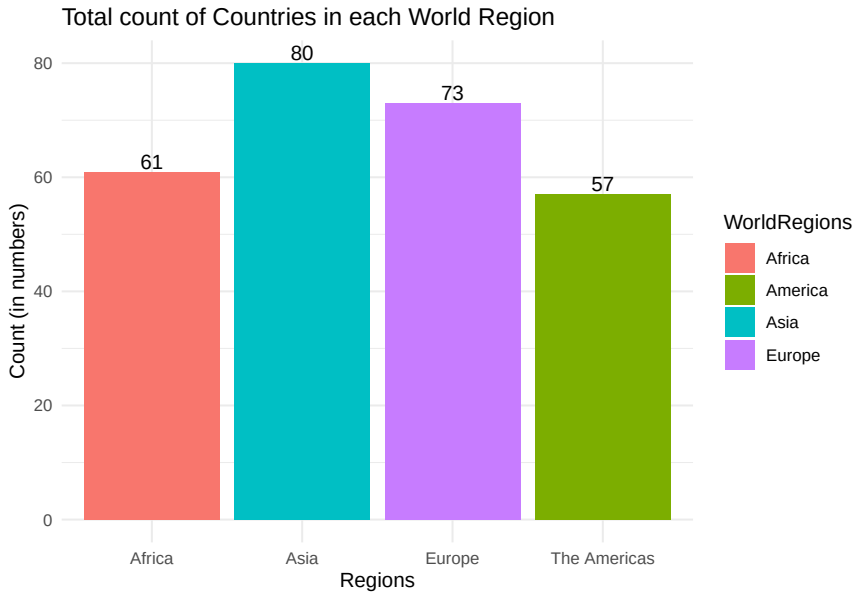
Using INNER JOIN to join the tables Countries and World_Region to get countries that have been assigned a world region

```
country_count<- inner_join(countries,world_region,by="world_4region")
head(country_count)
```

```
## # A tibble: 6 x 31
##   country   g77_and_oecd_countries income_3groups income_groups    `is--country`
##   <chr>     <chr>                  <chr>          <chr>            <lgl>
## 1 abkh      others                 <NA>           <NA>             TRUE
## 2 abw       others                 high_income    high_income      TRUE
## 3 afg       g77                    low_income     low_income       TRUE
## 4 ago       g77                    middle_income  lower_middle_in~ TRUE
## 5 aia       others                 <NA>           <NA>             TRUE
## 6 akr_a_dhe others                 <NA>           <NA>             TRUE
## # ... with 26 more variables: iso3166_1_alpha2 <chr>, iso3166_1_alpha3 <chr>,
## #   iso3166_1_numeric <dbl>, iso3166_2 <chr>, landlocked <chr>,
## #   latitude.x <dbl>, longitude.x <dbl>, main_religion_2008 <chr>,
## #   name.x <chr>, un_sdg_ldc <chr>, un_sdg_region <chr>, un_state <lgl>,
## #   unicef_region <chr>, unicode_region_subtag <chr>, world_4region <chr>,
## #   world_6region <chr>, color <chr>, description <chr>,
## #   is--world_4region <lgl>, latitude.y <dbl>, longitude.y <dbl>, ...
```

Including Plot to visualize the number of countries across regions

```
ggplot(data=country_count,mapping=aes(x=name.y,fill=name_short))+geom_bar()+
 labs(title="Total count of Countries in each World Region",x="Regions",
      y="Count (in numbers)",fill="WorldRegions")+
  geom_text(stat='count',aes(label=..count..),vjust=-0.2)+theme_minimal()
```



Observation of above plot:
Asia seems to have the highest number of countries and The Americas the least.

Part 4

Importing the infant mortality rate datapoints table

```
dir3 <- "ddf--gapminder--systema_globalis-master\\ddf--gapminder--systema_globalis-master\\countries-et
path<-file.path(dir1, dir3,"ddf--datapoints--infant_mortality_rate_per_1000_births--by--geo--time.csv")
inf_mortality_rate <- read_csv(path,guess_max=1000)
colnames(inf_mortality_rate)[colnames(inf_mortality_rate)=='geo']<-'country'
head(inf_mortality_rate)
```

```
## # A tibble: 6 x 3
##   country  time infant_mortality_rate_per_1000_births
##   <chr>   <dbl>                                 <dbl>
## 1 afg      1960                                  245
## 2 afg      1961                                  240.
## 3 afg      1962                                  236.
## 4 afg      1963                                  232.
## 5 afg      1964                                  228.
## 6 afg      1965                                  225.
```

Using INNER JOIN to join the country_count and inf_mortality_rate tables to get region wise infant mortality data

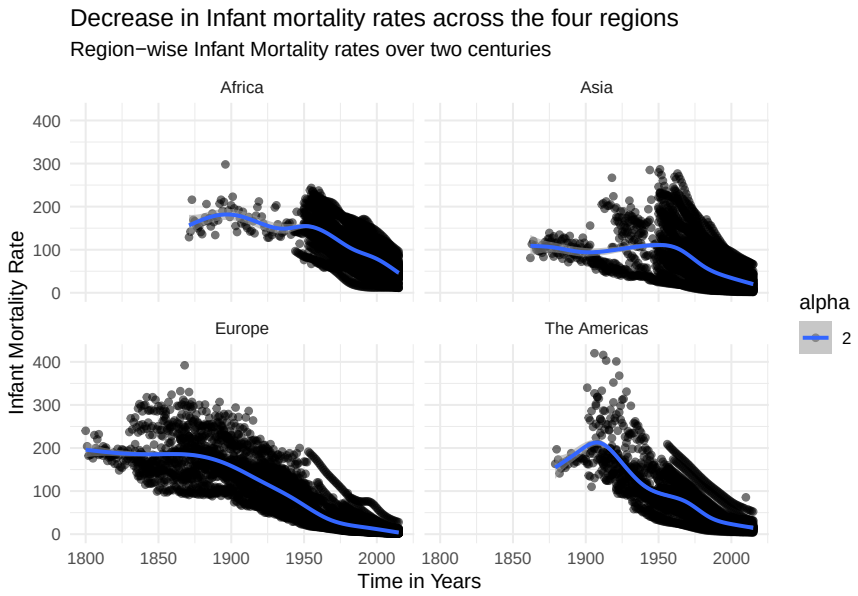
```
inf_region<-inner_join(inf_mortality_rate,country_count,by="country")
head(inf_region)
```

```
## # A tibble: 6 x 33
##   country  time infant_mortality_~ g77_and_oecd_co~ income_3groups income_groups
##   <chr>   <dbl>              <dbl> <chr>            <chr>          <chr>
## 1 afg      1960               245  g77              low_income     low_income
## 2 afg      1961               240. g77              low_income     low_income
## 3 afg      1962               236. g77              low_income     low_income
## 4 afg      1963               232. g77              low_income     low_income
## 5 afg      1964               228. g77              low_income     low_income
## 6 afg      1965               225. g77              low_income     low_income
## # ... with 27 more variables: is--country <lgl>, iso3166_1_alpha2 <chr>,
## #   iso3166_1_alpha3 <chr>, iso3166_1_numeric <dbl>, iso3166_2 <chr>,
## #   landlocked <chr>, latitude.x <dbl>, longitude.x <dbl>,
## #   main_religion_2008 <chr>, name.x <chr>, un_sdg_ldc <chr>,
## #   un_sdg_region <chr>, un_state <lgl>, unicef_region <chr>,
## #   unicode_region_subtag <chr>, world_4region <chr>, world_6region <chr>,
## #   color <chr>, description <chr>, is--world_4region <lgl>, ...
```

```
colnames(inf_region)[which(names(inf_region)=="name.y")]<-"Region"
```

Including plots for region-wise infant mortality rates

```
ggplot(data=inf_region,mapping=aes(x=time, y=infant_mortality_rate_per_1000_births,alpha=2)) +
    geom_point()+geom_smooth()+
  theme_minimal()+
labs(title = "Decrease in Infant mortality rates across the four regions",
    x="Time in Years",y="Infant Mortality Rate",
    subtitle="Region-wise Infant Mortality rates over two centuries")+
  facet_wrap(~Region)
```



Observation of the above plot:
Overall infant mortality rate seems to have reduced over the given period across all the four regions.

Africa: This region has had Infant mortality rate of 200-250 and has now come down to 5-100(approx).This range seems to be quite high when compared to other regions.

Asia: This region seems to have had a nearly constant mortality rate for the period of 1850-1950 and then a slight increase during 1950s and then has started to reduce to 0-50(approx) at year 2000.

Europe :Data for infant mortality for the 1800-1850 is available only for Europe out of the four regions. This region has had a nearly constant infant mortality rate of 200(approx.) for the years(1800-1900)and then starts to reduce constantly and now it seems to be 0-20(approx).

The Americas: This region shows an increase in the rates for the period of 1850-1900 and the starts to decrease over the following years and at 2000, it seems to have an infant mortality rate of 0-50(approx.)

Part 5

Importing the life expectancy datapoints table

```
path<-file.path(dir1,dir3,"ddf--datapoints--life_expectancy_years--by--geo--time.csv")
life_exp<-read_csv(path)
colnames(life_exp)[colnames(life_exp)=='geo']<-'country'
head(life_exp)
```

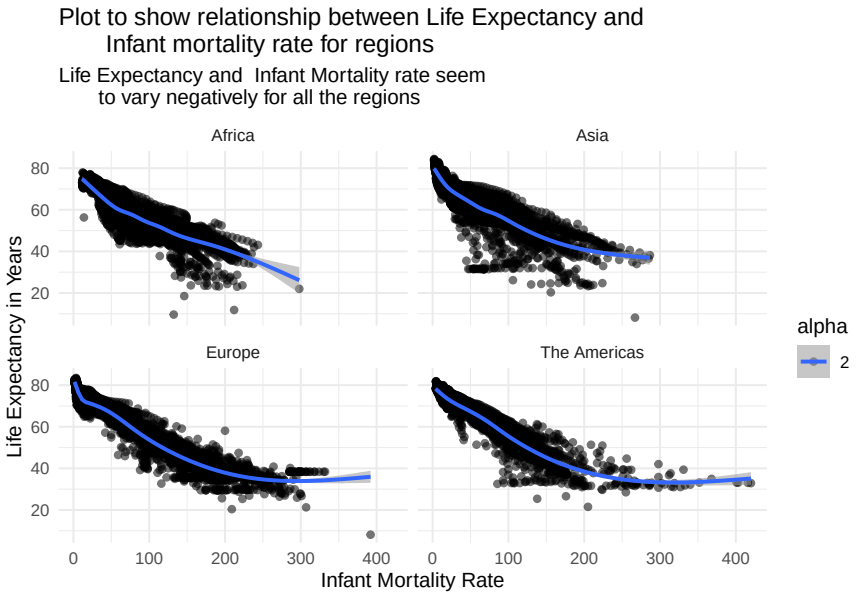
```
## # A tibble: 6 x 3
##   country  time life_expectancy_years
##   <chr>   <dbl>                 <dbl>
## 1 afg      1800                  28.2
## 2 afg      1801                  28.2
## 3 afg      1802                  28.2
## 4 afg      1803                  28.2
## 5 afg      1804                  28.2
## 6 afg      1805                  28.2
```

Using INNER JOIN to combine the life_exp and inf_region table to help visualize Life Expectancy Vs Infant Mortality Rate

```
life_exp_Inf_rate<-inner_join(life_exp,inf_region,by=c("country","time"))
head(life_exp_Inf_rate)
```

```
## # A tibble: 6 x 34
##   country  time life_expectancy_years infant_mortality_rate_~ g77_and_oecd_coun~
##   <chr>   <dbl>                 <dbl>                   <dbl> <chr>
## 1 afg      1960                  39.3                    245  g77
## 2 afg      1961                  40.0                    240. g77
## 3 afg      1962                  40.8                    236. g77
## 4 afg      1963                  41.5                    232. g77
## 5 afg      1964                  42.2                    228. g77
## 6 afg      1965                  43.0                    225. g77
## # ... with 29 more variables: income_3groups <chr>, income_groups <chr>,
## #   is--country <lgl>, iso3166_1_alpha2 <chr>, iso3166_1_alpha3 <chr>,
## #   iso3166_1_numeric <dbl>, iso3166_2 <chr>, landlocked <chr>,
## #   latitude.x <dbl>, longitude.x <dbl>, main_religion_2008 <chr>,
## #   name.x <chr>, un_sdg_ldc <chr>, un_sdg_region <chr>, un_state <lgl>,
## #   unicef_region <chr>, unicode_region_subtag <chr>, world_4region <chr>,
## #   world_6region <chr>, color <chr>, description <chr>, ...
```

```
ggplot(data=life_exp_Inf_rate,mapping=aes(x=infant_mortality_rate_per_1000_births
                                          ,y=life_expectancy_years,alpha=2))+
  geom_point()+geom_smooth()+facet_wrap(~life_exp_Inf_rate$Region)+
  labs(title = "Plot to show relationship between Life Expectancy and
       Infant mortality rate for regions",x="Infant Mortality Rate"
       ,y="Life Expectancy in Years",
       subtitle="Life Expectancy and  Infant Mortality rate seem
       to vary negatively for all the regions")+
  theme_minimal()+facet_wrap(~Region)
```



Observation for the above plot:
1.Life Expectancy and Infant Mortality rate seems to show negative correlation.
2.For a period where Infant mortality rate is between 0-200, the data points are widely distributed for Africa and Asia whereas, Europe and The Americas seem to have a much lesser range.
3.All the regions seem to have a life expectancy range of (30-80).