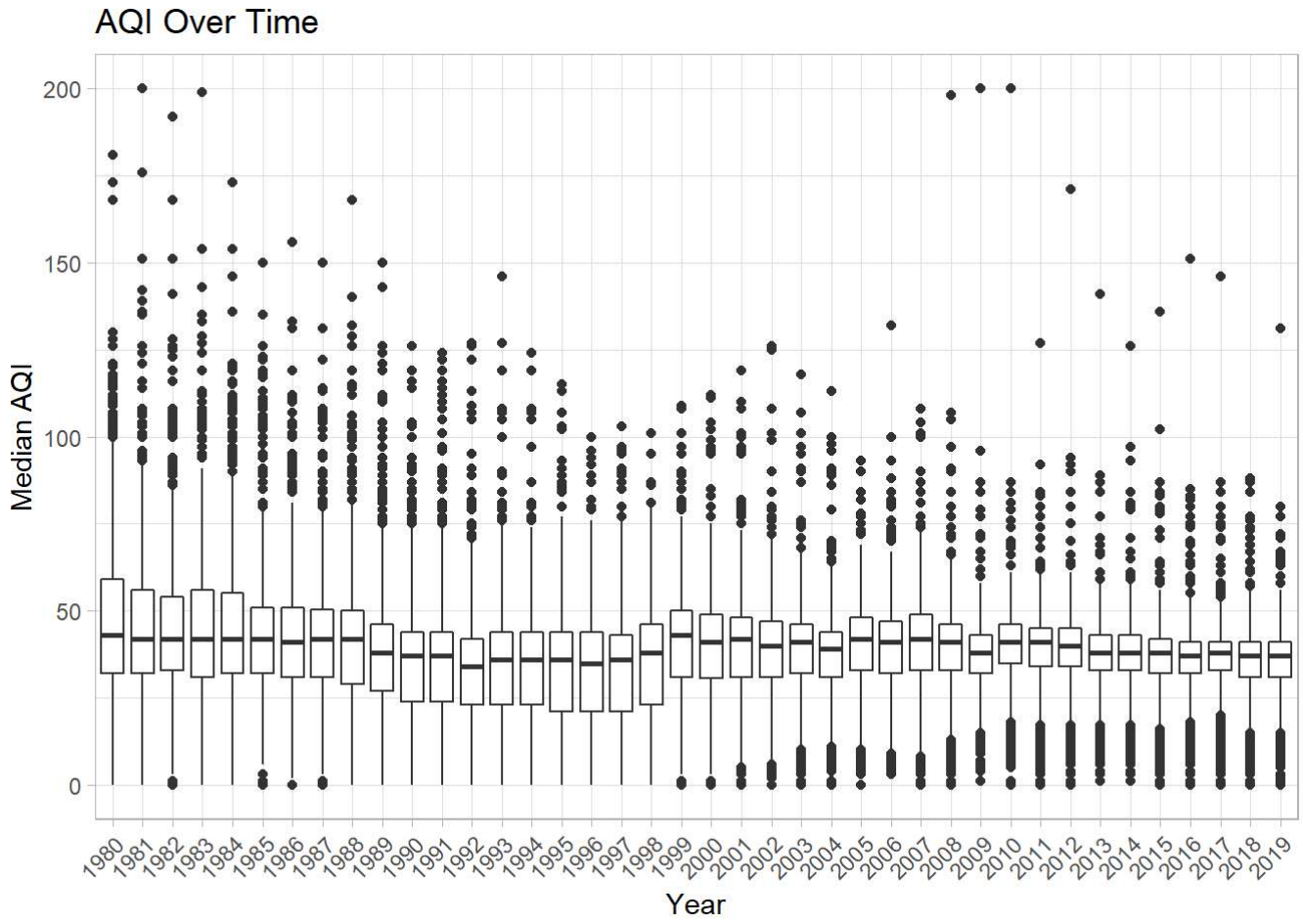
```
library(readr)
library(dplyr)
library(ggplot2)
```

Problem 1

```
file_path <- "D:/Education/MS DS NEU/IDMP/Assignments/HW3/HW3/epa-aqi-data-annual"
data <- list.files(file_path,pattern = "*.csv",full.names = TRUE)
df <- lapply(data, read_csv) %>% bind_rows()
df
```

```
## # A tibble: 38,511 x 19
##    State   County      Year `Days with AQI` `Good Days` `Moderate Days`
##    <chr>   <chr>      <dbl>           <dbl>       <dbl>           <dbl>
##  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
##  7 Alabama Mobile      1980             286         180              62
##  8 Alabama Monroe      1980              90          63              14
##  9 Alabama Morgan      1980             332         207              93
## 10 Alabama Tuscaloosa  1980             132          94              28
## # ... with 38,501 more rows, and 13 more variables:
## #   Unhealthy for Sensitive Groups Days <dbl>, Unhealthy Days <dbl>,
## #   Very Unhealthy Days <dbl>, Hazardous Days <dbl>, Max AQI <dbl>,
## #   90th Percentile AQI <dbl>, Median AQI <dbl>, Days CO <dbl>, Days NO2 <dbl>,
## #   Days Ozone <dbl>, Days SO2 <dbl>, Days PM2.5 <dbl>, Days PM10 <dbl>
```

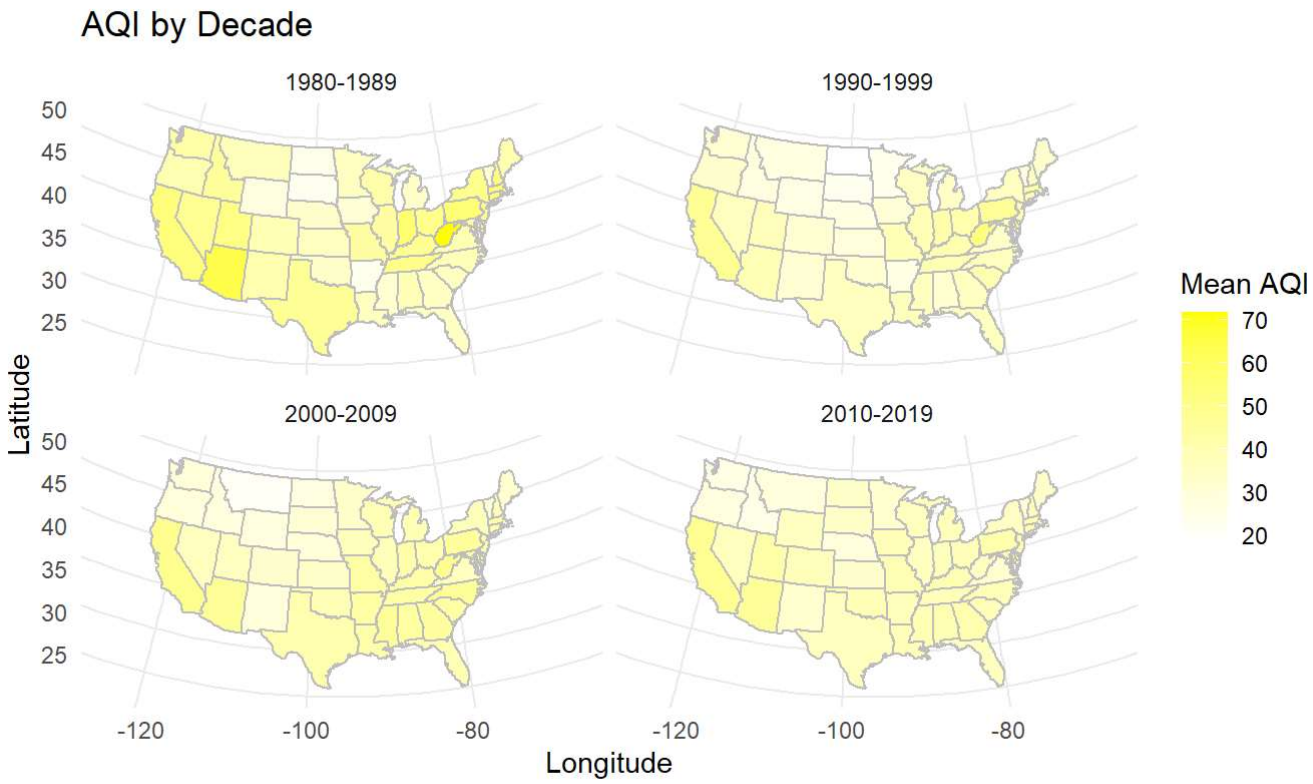
```
ggplot(df,aes(x=factor(Year),y=`Median AQI`,group=Year)) +
geom_boxplot() +
scale_x_discrete(guide = guide_axis(angle = 45)) +
labs(x="Year", y="Median AQI",title = "AQI Over Time") +
theme_light()
```



Observations: Air Quality has gradually increased from 1980 to 2019 with slight ups and downs during the 2000s.

Problem 2

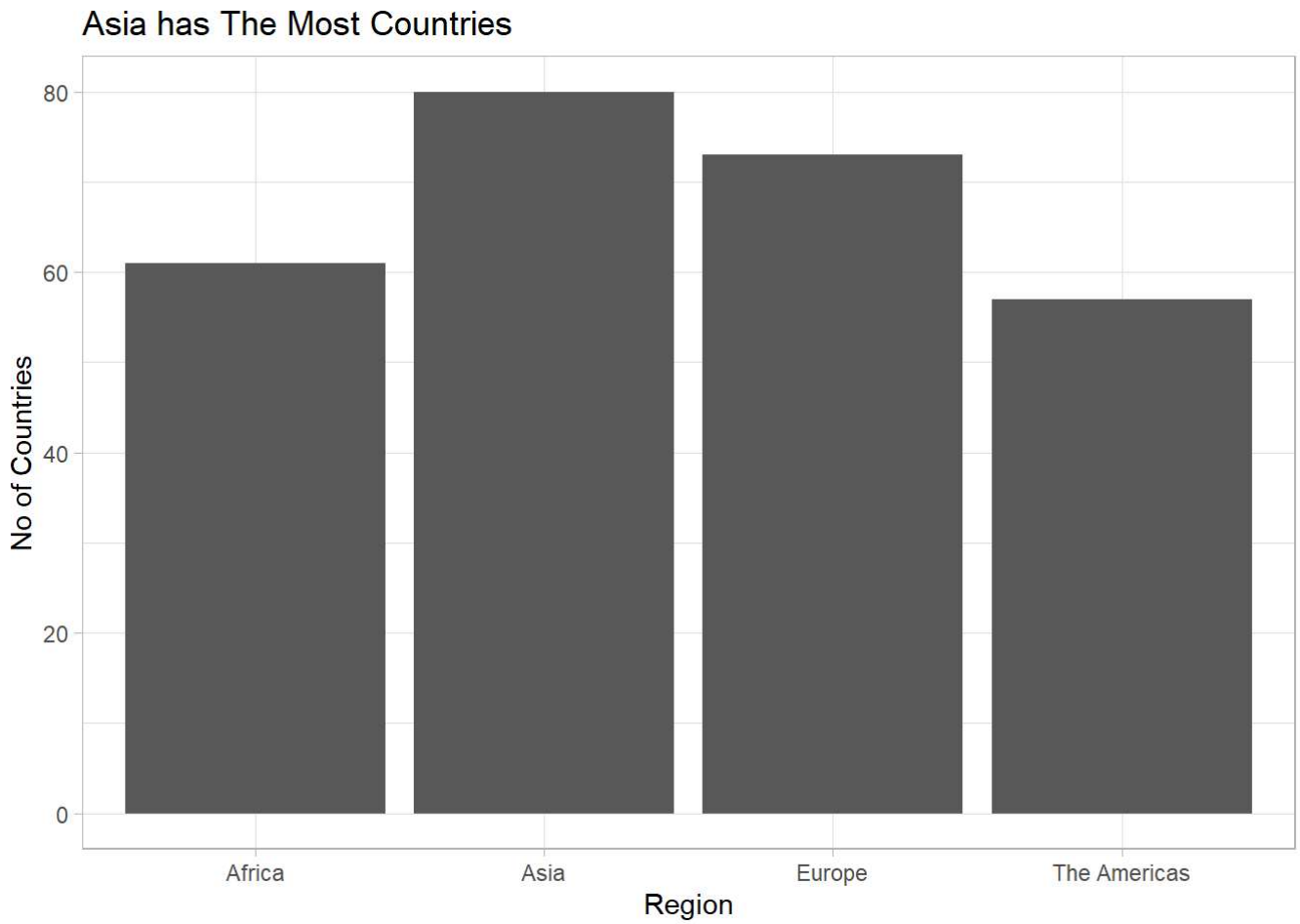
```
aqi_data <- df %>% mutate("Decade"=Year-(Year%%10)) %>%
mutate(Decade = case_when(Decade==1980 ~ "1980-1989",
                          Decade == 1990 ~ "1990-1999",
                          Decade == 2000 ~ "2000-2009",
                          Decade==2010 ~ "2010-2019")) %>%
group_by(Decade,State) %>% summarise("Mean AQI" = mean(`Median AQI`))
aqi_data %>% mutate(State=tolower(State)) %>%
rename(region=State) %>%
inner_join(map_data("state")) %>%
ggplot(aes(long,lat,group=group)) +
geom_polygon(aes(fill=`Mean AQI`),color="grey")  +
facet_wrap(~Decade) +
scale_fill_gradient(high = "yellow",low = "white")+
coord_map("albers",lat0=45.5,lat1=29.5) +
labs(x="Longitude",y="Latitude",title = "AQI by Decade")+
theme_minimal()
```



Observations: Western States of the country demonstrate improved air quality among them. North Eastern states have improved the air quality over the decades Central and Northern states does not have noteworthy increases in AQI.

Problem 3

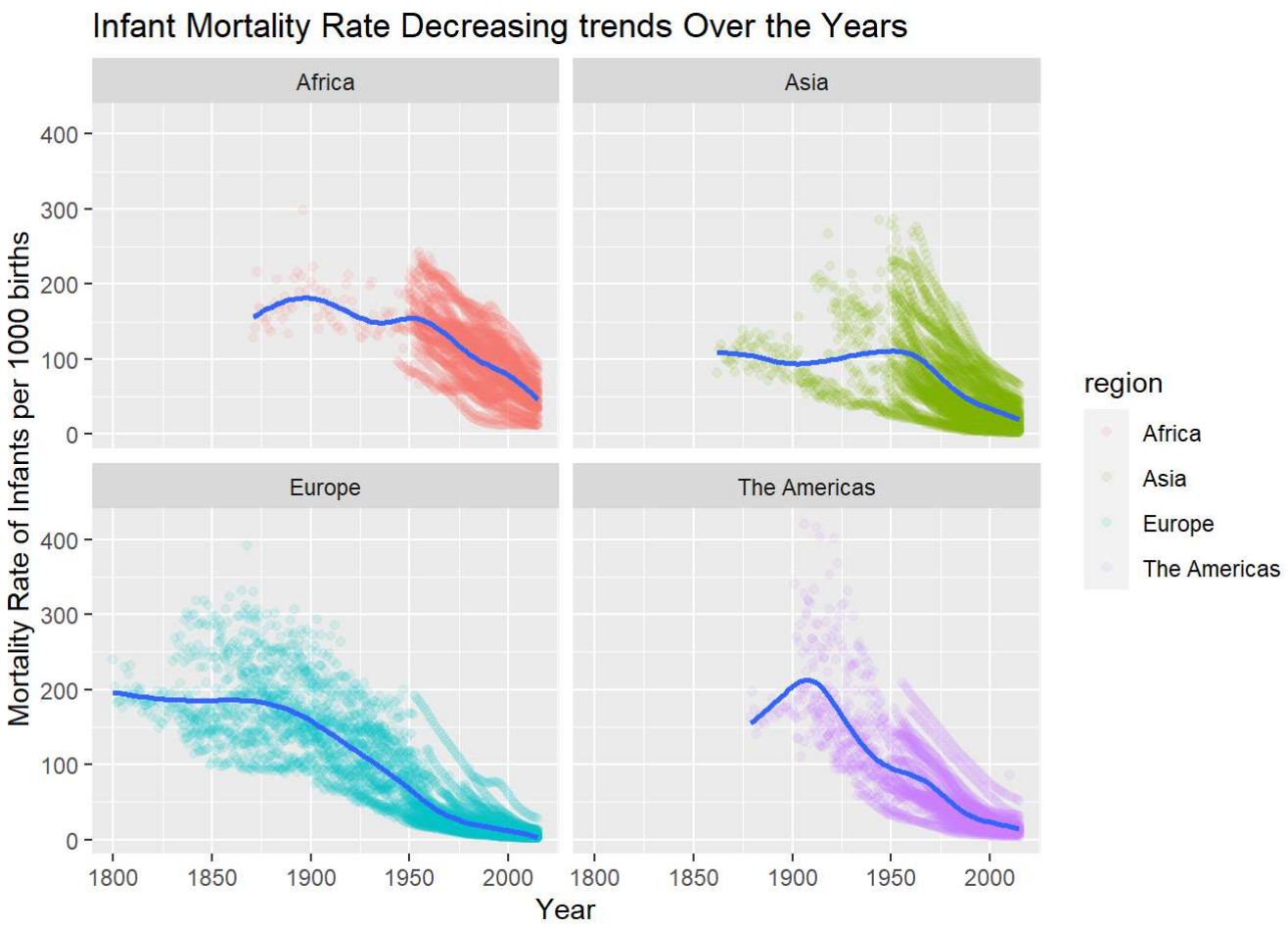
```
file_path <- "D:/Education/MS DS NEU/IDMP/Assignments/HW3/HW3/ddf--gapminder--systema_globalis-m
aster"
country_path <- file.path(file_path,"ddf--entities--geo--country.csv")
world_path <- file.path(file_path,"ddf--entities--geo--world_4region.csv")
country <- read_csv(country_path)
world4Region <- read_csv(world_path)
world4Region<-world4Region %>% select(world_4region,name) %>% rename(region=name)
country_world4region <- inner_join(world4Region,country) %>% select(country,region)
country_world4region %>% ggplot(aes(x=region)) +
geom_bar() +
labs(x="Region",y="No of Countries", title = "Asia has The Most Countries") +
theme_light()
```



Observations: Asia has the most number of countries, Europe second and The Americas is third While Africa has the least number of countries.

Problem 4

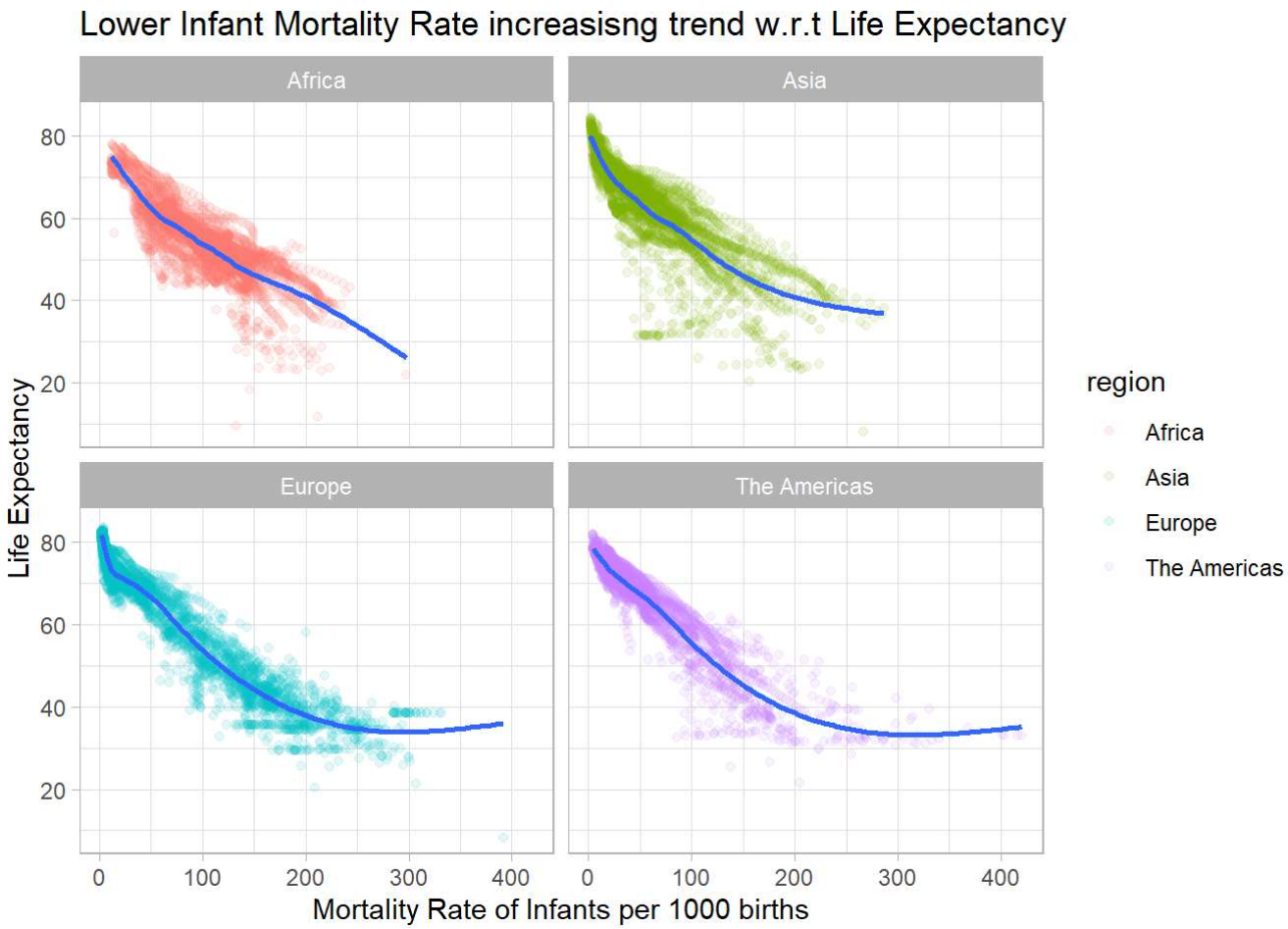
```
path_head="D:/Education/MS DS NEU/IDMP/Assignments/HW3/HW3"
path_tail="/ddf--gapminder--systema_globalis-master/countries-etc-datapoints"
path = file.path(path_head,path_tail,"ddf--datapoints--infant_mortality_rate_per_1000_births--by
--geo--time.csv")
imr=read_csv(path)
left_join(imr,country_world4region,by=c("geo"="country")) %>%
ggplot(aes(x=time,y=infant_mortality_rate_per_1000_births)) +
geom_point(aes(color=region),alpha=0.1) +
geom_smooth(se=FALSE) +
labs(x="Year",y="Mortality Rate of Infants per 1000 births",
     title="Infant Mortality Rate Decreasing trends Over the Years") +
facet_wrap(~region)
```



Observations: Infant mortality has decreased as the time passes. It is same across all regions. American infant mortality rate has short spike between 1850 to 1900 and reduced later in the time. Europe and The Americas saw the deepest dip in the Infant Mortality rate.

Problem 5

```
path = file.path(path_head,path_tail,"ddf--datapoints--life_expectancy_years--by--geo--time.cs
v")
lifeExp = read_csv(path)
lifeExp %>% inner_join(imr) %>%
left_join(country_world4region,by=c("geo"="country")) %>%
ggplot(aes(x=infant_mortality_rate_per_1000_births,y=life_expectancy_years)) +
geom_point(aes(color=region),alpha=0.1) +
geom_smooth(se=FALSE) +
theme_light() +
labs(x="Mortality Rate of Infants per 1000 births",y ="Life Expectancy",
     title = "Lower Infant Mortality Rate increasisng trend w.r.t Life Expectancy") +
facet_wrap(~region)
```



Observations: Increase in Life Expectancy decreases Infant Mortality Rate. It is same across all regions.