Ass_3

2022-10-17

```
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(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.5
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.3
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(ggplot2)
library(maps)
```

```
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(stringr)
```

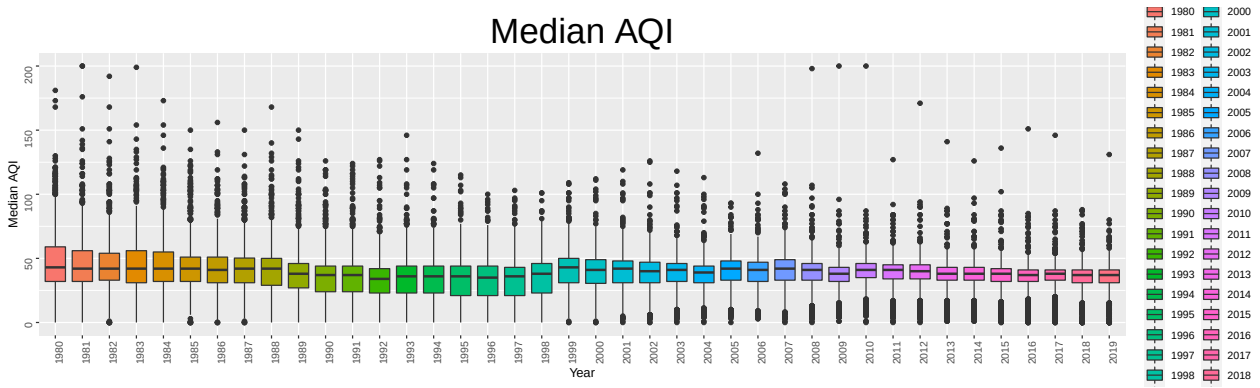
Solution-1

```
temp = list.files(path="home/adityas/Desktop/IDMP/epa-aqi-data-annual", pattern="*.csv", full.names=TRU
df = read_csv(temp[1], show_col_types=FALSE)
for (i in 2:length(temp)){
  df = bind_rows(df, read_csv(temp[i], show_col_types=FALSE))
}
```

```
na.omit(df)
```

```
## # A tibble: 38,511 x 19
##    State   County   Year Days ~1 Good ~2 Moder~3 Unhea~4 Unhea~5 Very ~6 Hazar~7
##    <chr>   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Alabama Autauga  1980     179     122      35      18       4       0       0
##  2 Alabama Colbert  1980     274     127      45      63      39       0       0
##  3 Alabama Jackson  1980     366      85     110      92      79       0       0
##  4 Alabama Jeffer~  1980     343     171     109      37      19       7       0
##  5 Alabama Lauder~  1980     274     120      58      77      19       0       0
##  6 Alabama Madison  1980     344     154     125      60       5       0       0
##  7 Alabama Mobile   1980     286     180      62      35       8       1       0
##  8 Alabama Monroe   1980      90      63      14       7       6       0       0
##  9 Alabama Morgan   1980     332     207      93      32       0       0       0
## 10 Alabama Tuscal~  1980     132      94      28      10       0       0       0
## # ... with 38,501 more rows, 9 more variables: `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>, and abbreviated variable names 1: `Days with AQI`,
## #   2: `Good Days`, 3: `Moderate Days`,
## #   4: `Unhealthy for Sensitive Groups Days`, 5: `Unhealthy Days`,
## #   6: `Very Unhealthy Days`, 7: `Hazardous Days`
```

```
g <- ggplot(df, aes(x=`Median AQI`, fill=as.factor(Year)))

g + geom_boxplot(aes(y=as.factor(Year))) +
theme(axis.text=element_text(angle=90), plot.title = element_text(size=30,hjust=0.5, vjust=0.5)) +
  coord_flip() +
  labs(title="Median AQI",y='Year',x='Median AQI')
```



The Median value stays visually constant till from 1980 to 1988 takes a dip from 1989 to 1992. Then Increases and later stays constant. However, the interquantile range keeps decreasing over the years.

Solution - 2

```
df2 <- df %>% mutate(decade = cut_interval(Year,length = 10, dig.lab=10,right=FALSE))
df2  <- df2  %>% group_by(State,decade) %>% summarize(mean_aqi = mean(`Median AQI`))
```

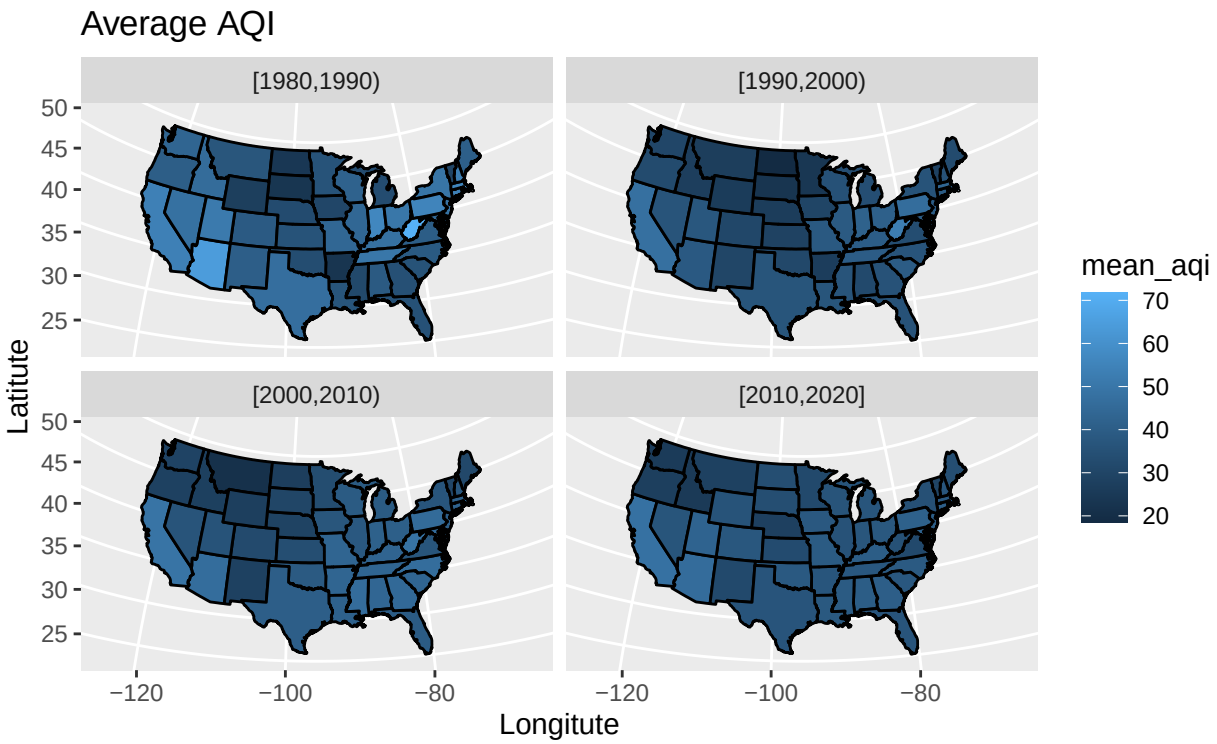
```
## `summarise()` has grouped output by 'State'. You can override using the
## `.groups` argument.
```

```
df2  <- mutate(df2, State = tolower(State))
sp <- map_data('state')
df2 <- rename(df2,'region'='State')
```

```
final_state <- inner_join(df2,sp,on = 'region')
```

```
## Joining, by = "region"
```

```
ggplot(final_state,aes(x=long,y=lat,group = group,fill =mean_aqi))+
geom_polygon(colour='black')+coord_map('polyconic')+facet_wrap(~decade)+
labs(title="Average AQI",y='Latitute',x='Longitute')
```



The average AQI increases in the decade 2000-2010 due to rise in industries but from 2010 to 2020 it again reduces maybe due to rise of polution reducing methods specially in the north region polution has reduced.

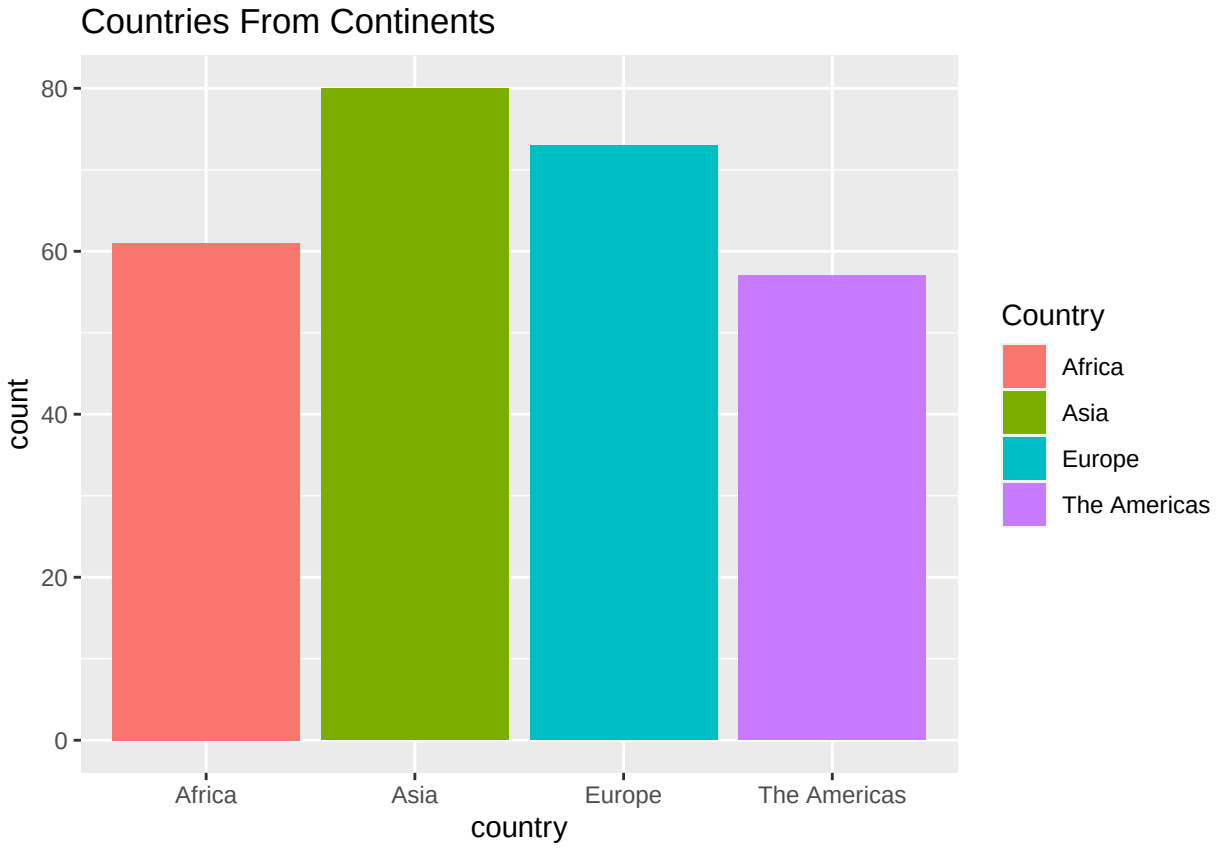
Problem 3

```
country <- read_csv("home/adityas/Desktop/IDMP/ddf--gapminder--systema_globalis-master/ddf--entities--g


na.omit(country) # Removing null values
```

```
## # A tibble: 0 x 21
## # ... with 21 variables: country <chr>, g77_and_oecd_countries <chr>,
## #   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 <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>
```

```
region <- read_csv("home/adityas/Desktop/IDMP/ddf--gapminder--systema_globalis-master/ddf--entities--ge

c_region <- inner_join(country, region, by="world_4region")
```

```
c_region <- c_region %>% select(-latitude.x, -longitude.x)

ggplot(c_region, aes(x=name.y, fill=name.y)) + geom_bar() + scale_fill_discrete("Country") + labs(x="cou
```

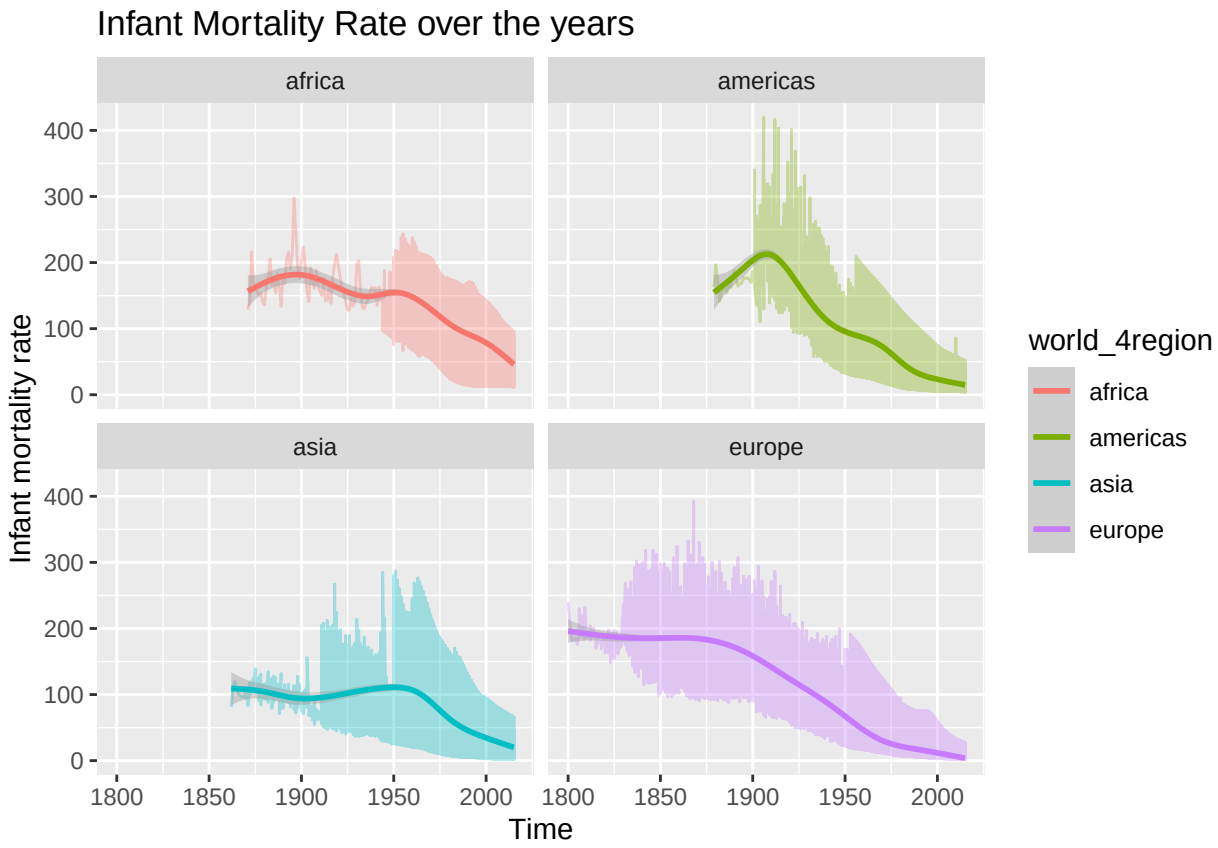


From the plot chart we can infer that most of the countries are from Asia and least from Africa.

Problem 4

```
infant_mortality <- read_csv("home/adityas/Desktop/IDMP/ddf--gapminder--systema_globalis-master/countri

c<-read_csv("home/adityas/Desktop/IDMP/ddf--gapminder--systema_globalis-master/ddf--entities--geo--coun

c <- rename(c, "geo"="country")

df_ <- inner_join(c, infant_mortality, by="geo")


ggplot(df_, aes(x=time, y=infant_mortality_rate_per_1000_births, color=world_4region)) + geom_line(alpha
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Infant mortality has gone down in all the regions with Europe being the least, Americas had a slight elevation in early 1850-1870 however the mortality rate went down fast. Africa still has high mortality rate compared to other regions.

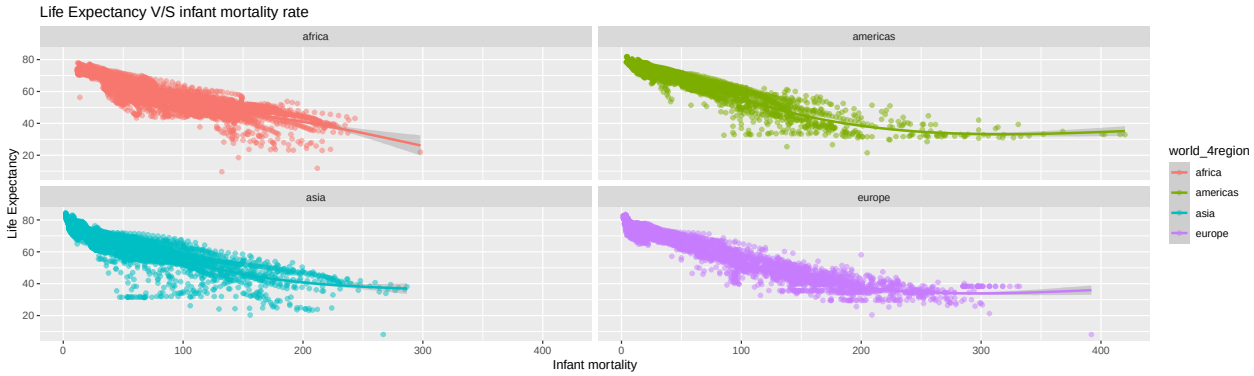
Problem 5

```
life_expentancy <- read_csv("home/adityas/Desktop/IDMP/ddf--gapminder--systema_globalis-master/countries

life_exp_inf_mort <- inner_join(life_expentancy, df_)
```

```
## Joining, by = c("geo", "time")
```

```
ggplot(life_exp_inf_mort, aes(x=infant_mortality_rate_per_1000_births, y=life_expectancy_years, color=w
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



From The above plot we can infer that as the infant mortality rate increases the Life expectancy decreases in ALL the four regions. Therefore we can establish that they have negative correlation.