HM3

Mu Cheng

3/1/2021

Problem 1

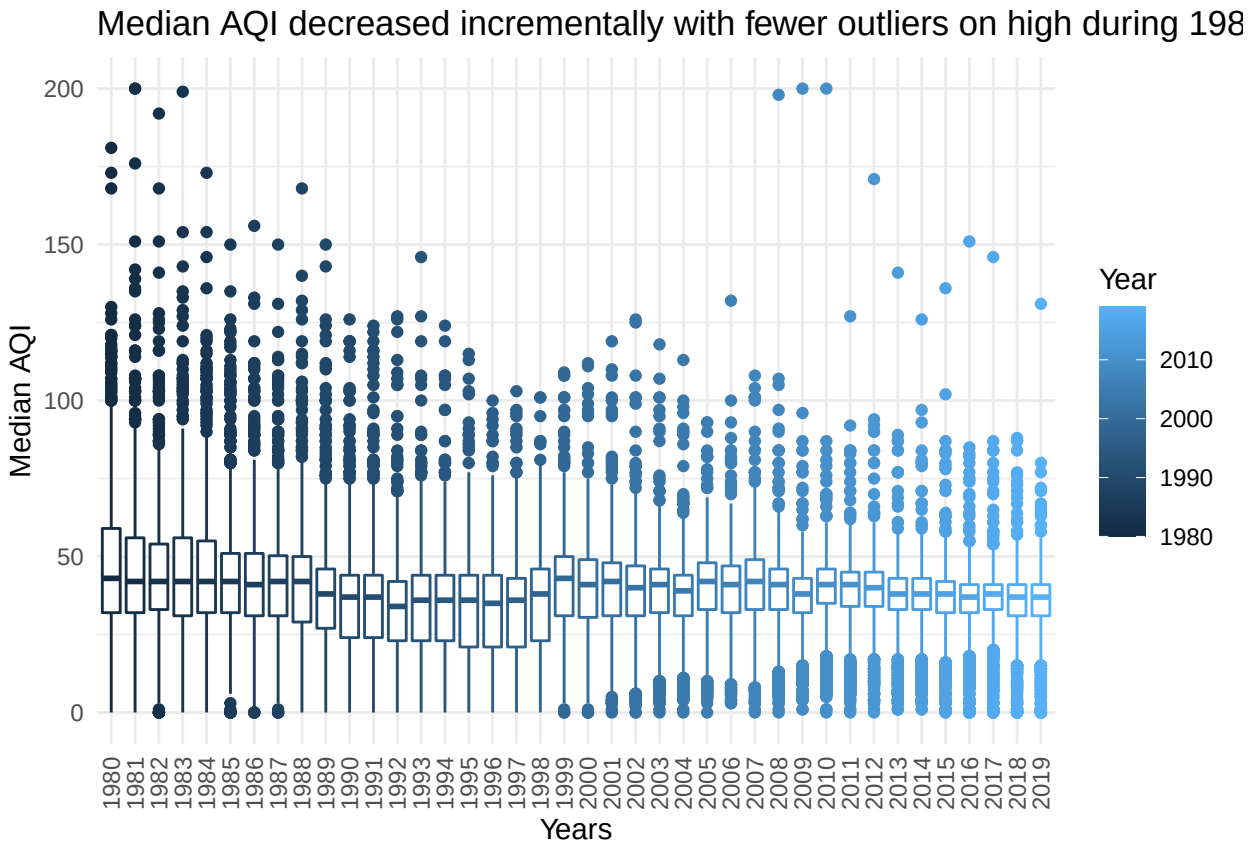
```
getwd()
```

```
## [1] "/Users/mucheng/Desktop/5110/HM3"
```

```
setwd("~/Desktop/5110/HM3")
library("readr")
library(plyr)
library(dplyr)

mydir = "epa-aqi-data-annual"
myfiles <- list.files(path=mydir, pattern="*.csv", full.names=TRUE) %>%
  lapply(read_csv) %>%
  bind_rows()
```

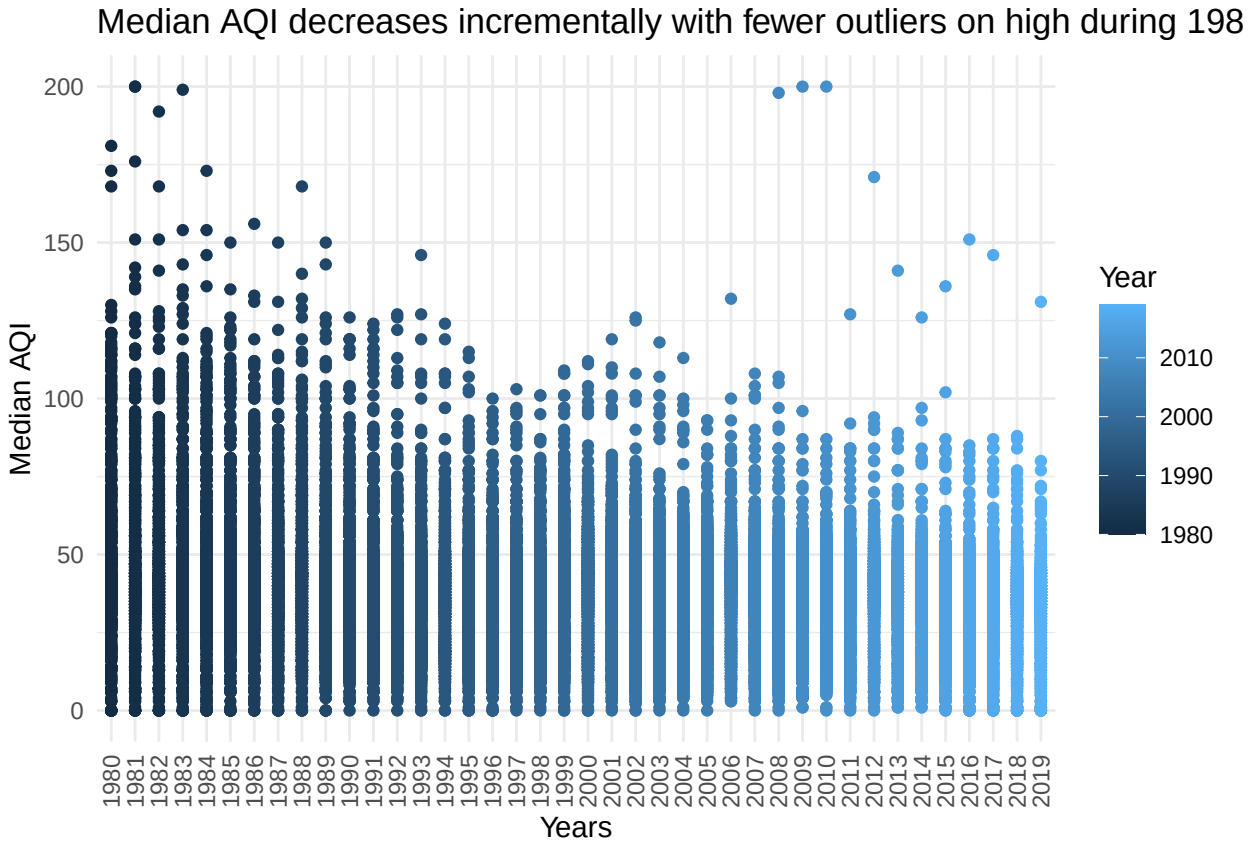
1st visulization with boxplot

```
library(ggplot2)
ggplot(myfiles, mapping=aes(x=as.factor(Year), y=`Median AQI`, color=`Year`)) +
  geom_boxplot() + labs(x="Years",
                        y="Median AQI",
                        title=
"Median AQI decreased incrementally with fewer outliers on high during 1980-2019 ") + theme_minimal() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5,  hjust=1))
```



2nd Visulization with Scatter-plot. The second graph is made to assure the conclusion addressed on outliers that are getting fewer than before is true in the above analysis.

```
library(ggplot2)
ggplot(myfiles, mapping=aes(x=as.factor(Year), y=`Median AQI`, color=`Year`)) +
  geom_point() + labs(x="Years",
                      y="Median AQI",
                      title=
"Median AQI decreases incrementally with fewer outliers on high during 1980-2019") + theme_minimal() +
  theme(axis.text.x = element_text(angle=90, vjust=0.5,  hjust=1))
```



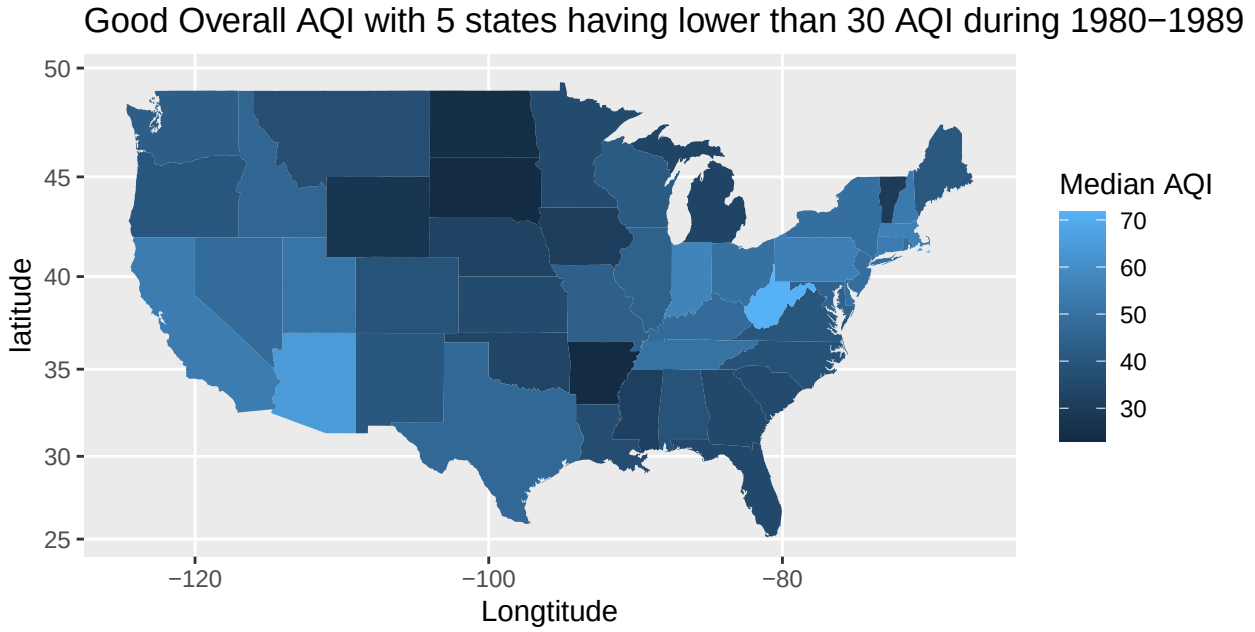
Problem 2

```
lowerLetters <- function(s) {
    paste(tolower(substring(s, 1, 20)))
}

myfiles$State <- lowerLetters(myfiles$State)


decade1 <- filter(myfiles, `Year` >= 1980 & `Year` <= 1989)

decade2 <- filter(myfiles, `Year` >= 1990 & `Year` <= 1999)

decade3 <- filter(myfiles, `Year` >= 2000 & `Year` <= 2009)

decade4 <- filter(myfiles, `Year` >= 2010 & `Year` <= 2019)
```

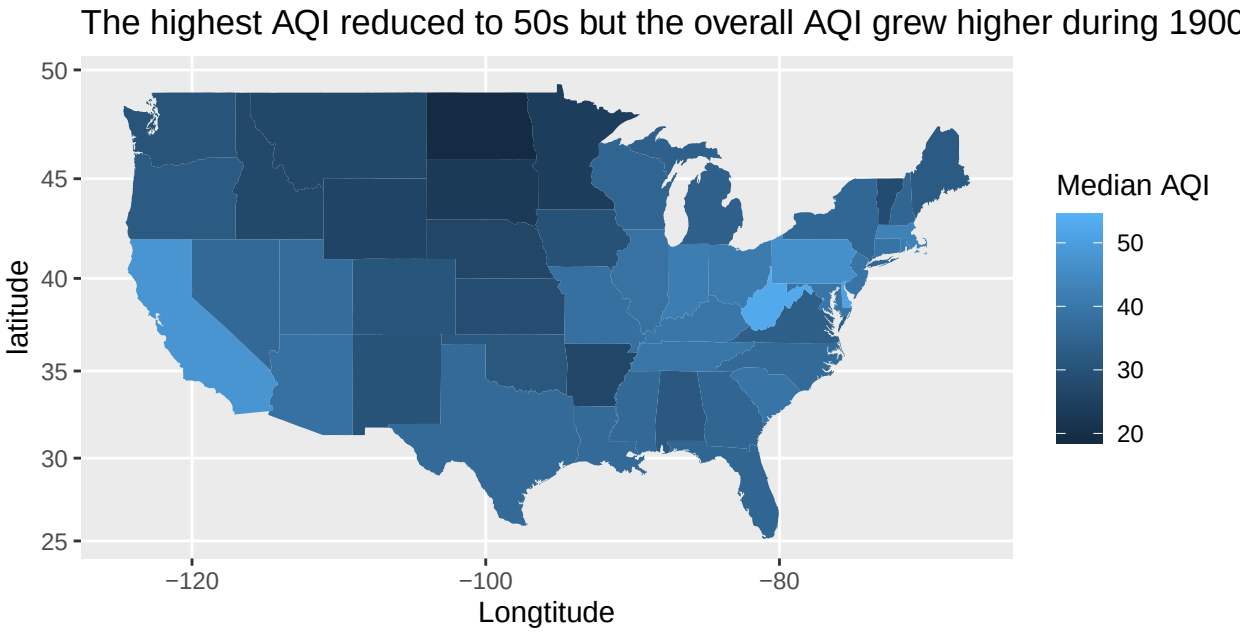
```
decade1_state <-select(decade1, `State`, `Median AQI`) %>%
  group_by(`State`) %>%
  summarize_each(funs(mean(`Median AQI`, na.rm=TRUE)))

decade2_state <-select(decade2, `State`, `Median AQI`) %>%
group_by(`State`) %>%
  summarize_each(funs(mean(`Median AQI`, na.rm=TRUE)))

decade3_state <-select(decade3, `State`, `Median AQI`) %>%
  group_by(`State`) %>%
  summarize_each(funs(mean(`Median AQI`, na.rm=TRUE)))
```

```
decade4_state <-select(decade4, `State`, `Median AQI`) %>% group_by(`State`) %>%
  summarize_each(funs(mean(`Median AQI`, na.rm=TRUE)))
```

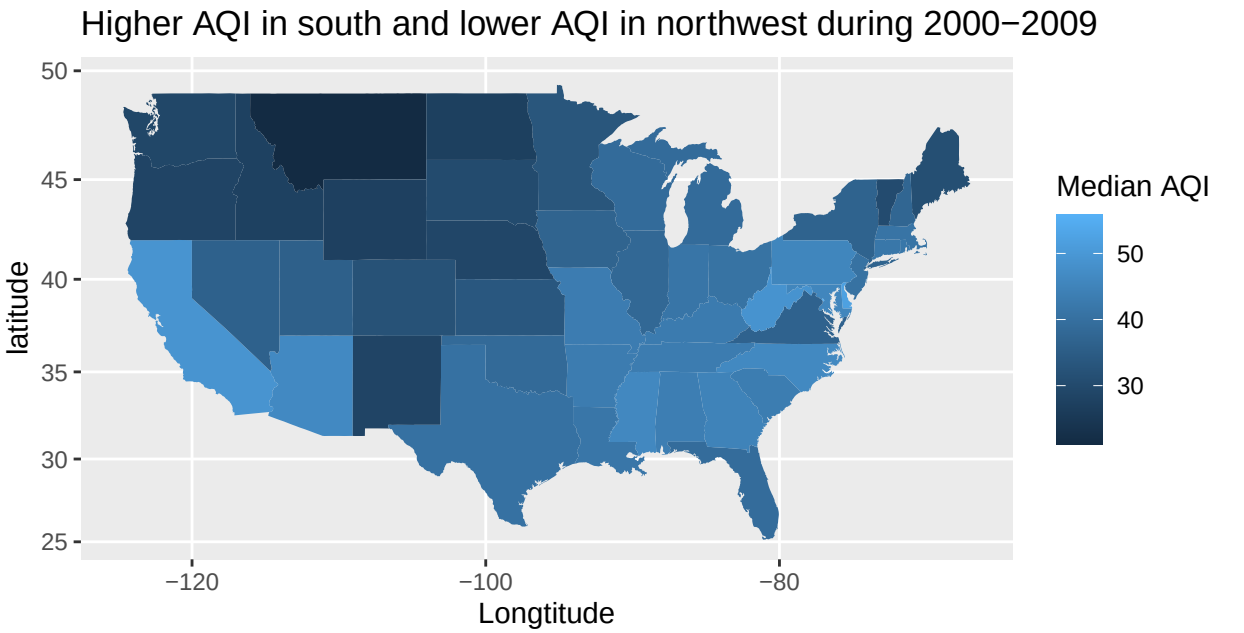
```
library(maps)

us_map <- map_data("state")

us_map1 <-inner_join(us_map, decade1_state, by=c("region"="State"))

ggplot() +
  geom_polygon(data = us_map1, aes(x = long,
                                   y = lat, group = group, fill=`Median AQI`)) +
  coord_map() + labs(x="Longtitude",
                     y="latitude",
                     title=
"Good Overall AQI with 5 states having lower than 30 AQI during 1980-1989")
```




```
us_map2 <-inner_join(us_map, decade2_state, by=c("region"="State"))

ggplot() +
  geom_polygon(data = us_map2, aes(x = long,
                                   y = lat, group = group, fill=`Median AQI`)) +
  coord_map() + labs(x="Longtitude",
                     y="latitude", title=
"The highest AQI reduced to 50s but the overall AQI grew higher during 1900-1999")
```

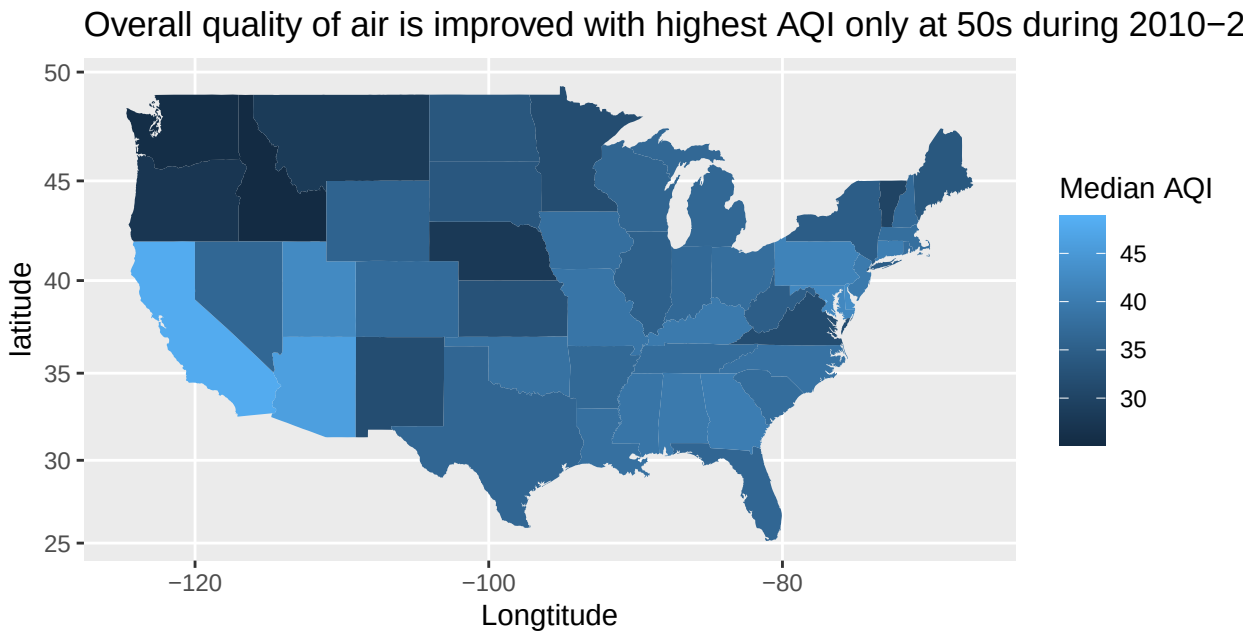



```
us_map3 <-inner_join(us_map, decade3_state, by=c("region"="State"))

ggplot() +
  geom_polygon(data = us_map3, aes(x = long,
                                   y = lat, group = group, fill=`Median AQI`)) +
  coord_map() + labs(x="Longtitude",
                     y="latitude",
                     title=
"Higher AQI in south and lower AQI in northwest during 2000-2009")
```




```
us_map4 <-inner_join(us_map, decade4_state, by=c("region"="State"))

ggplot() +
  geom_polygon(data = us_map4, aes(x = long,
```

```
                                    y = lat, group = group, fill=`Median AQI`)) +
  coord_map() + labs(x="Longtitude",
                     y="latitude", title=
"Overall quality of air is improved with highest AQI only at 50s during 2010-2019")
```

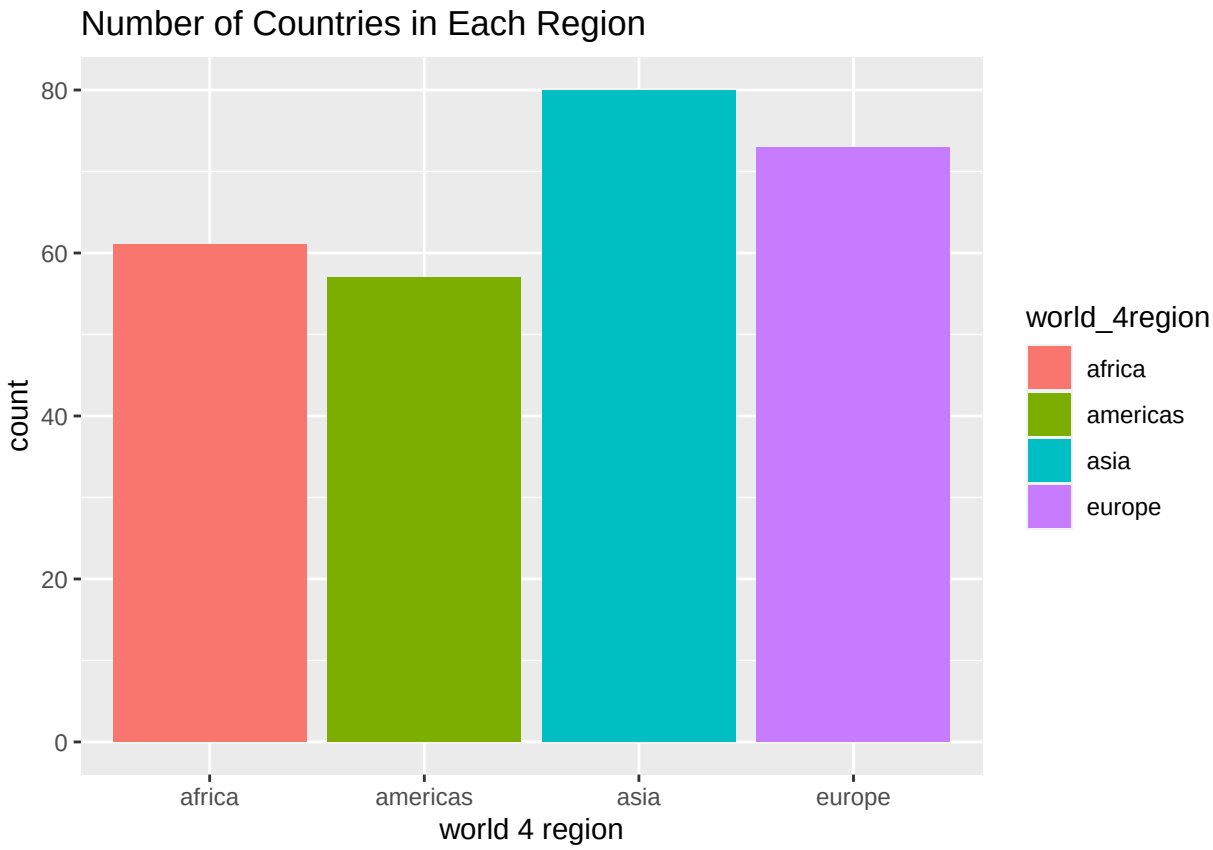


Problem 3

```
getwd()
```

```
## [1] "/Users/mucheng/Desktop/5110/HM3"
```

```
setwd("~/Desktop/5110/HM3/ddf--gapminder--systema_globalis-master")
library("readr")
library(ggplot2)

country_entities <- read_csv("ddf--entities--geo--country.csv")

country_region_match <- select (country_entities, `name`, `world_4region`)

a <- na.omit(country_region_match)

country_num_each_region <- ggplot(data=a, mapping=aes(x=`world_4region`, fill=`world_4region`)) +
geom_bar() + labs(x="world 4 region",
                  y="count",
                  title="Number of Countries in Each Region")
country_num_each_region
```



Problem 4

In Africa:The overall infant mortality rate has decreased during 1950-2015. Twice had the rate happened to increase, they were at 1875-1900 and late 1930s-1960s.

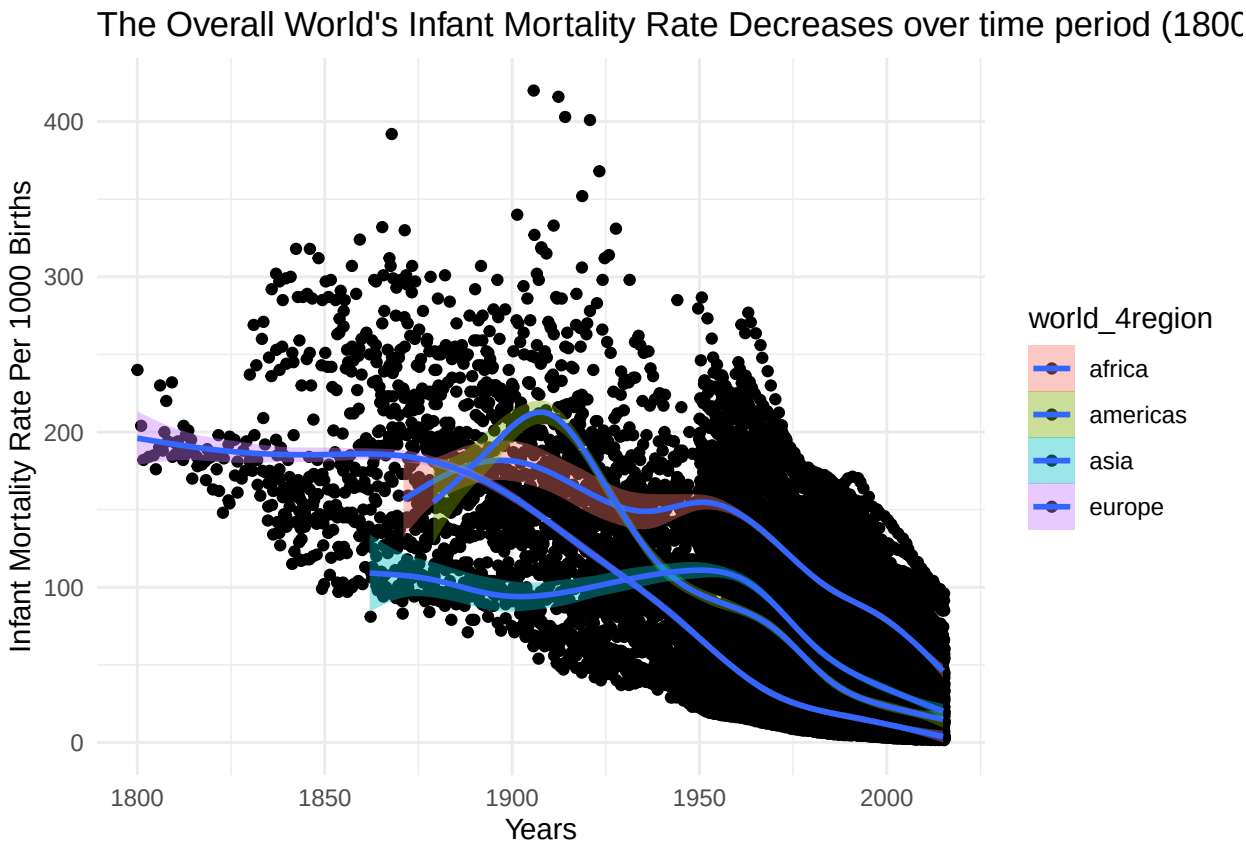
In America:The overall infant mortality rate has decreased during 1950-2015. One major increase happened at 1875-1912 time-period.

In Asia:The overall infant mortality rate has decreased during 1950-2015. One major increase happened from 1900 to early 1960s time-period.

In Europe: The only region world-wide that has stored data all the way from 1800s. The infant mortality rate is quite stable during 1800-1875. Slightly decreased and then slightly increased. Yet, drastically decreased during 1875-1960s. Then slowly decreased after 1960s.

```
library(RSQLite)
library(dbplyr)
library(DBI)
library(ggplot2)

setwd("~/Desktop/5110/HM3/ddf--gapminder--systema_globalis-master")
library("readr")
country_entities <- read_csv("ddf--entities--geo--country.csv")


setwd("~/Desktop/5110/HM3/ddf--gapminder--systema_globalis-master/countries-etc-datapoints")
infant_mortality_rate <- read_csv(
  "ddf--datapoints--infant_mortality_rate_per_1000_births--by--geo--time.csv")
```

```
con <- dbConnect(SQLite(), ":memory:")
dbWriteTable(con, "infant_mortality_rate", infant_mortality_rate)
dbWriteTable(con, "country_entities", country_entities)

joined_infant_mortality_dt <-dbGetQuery(con, "SELECT *
                                        FROM infant_mortality_rate
                                        JOIN country_entities
                                        ON geo=country
                                        ORDER BY time DESC")

dbWriteTable(con, "joined_infant_mortality_dt", joined_infant_mortality_dt)

ggplot(data=joined_infant_mortality_dt,
       mapping=aes(x=time,
                   y=infant_mortality_rate_per_1000_births,
                   fill=world_4region)) +
  geom_point(position="jitter") +
  geom_smooth(method = 'gam') +
  labs(x="Years",
       y="Infant Mortality Rate Per 1000 Births", title=
"The Overall World's Infant Mortality Rate Decreases over time period (1800-2015)") + theme_minimal()
```



Problem 5

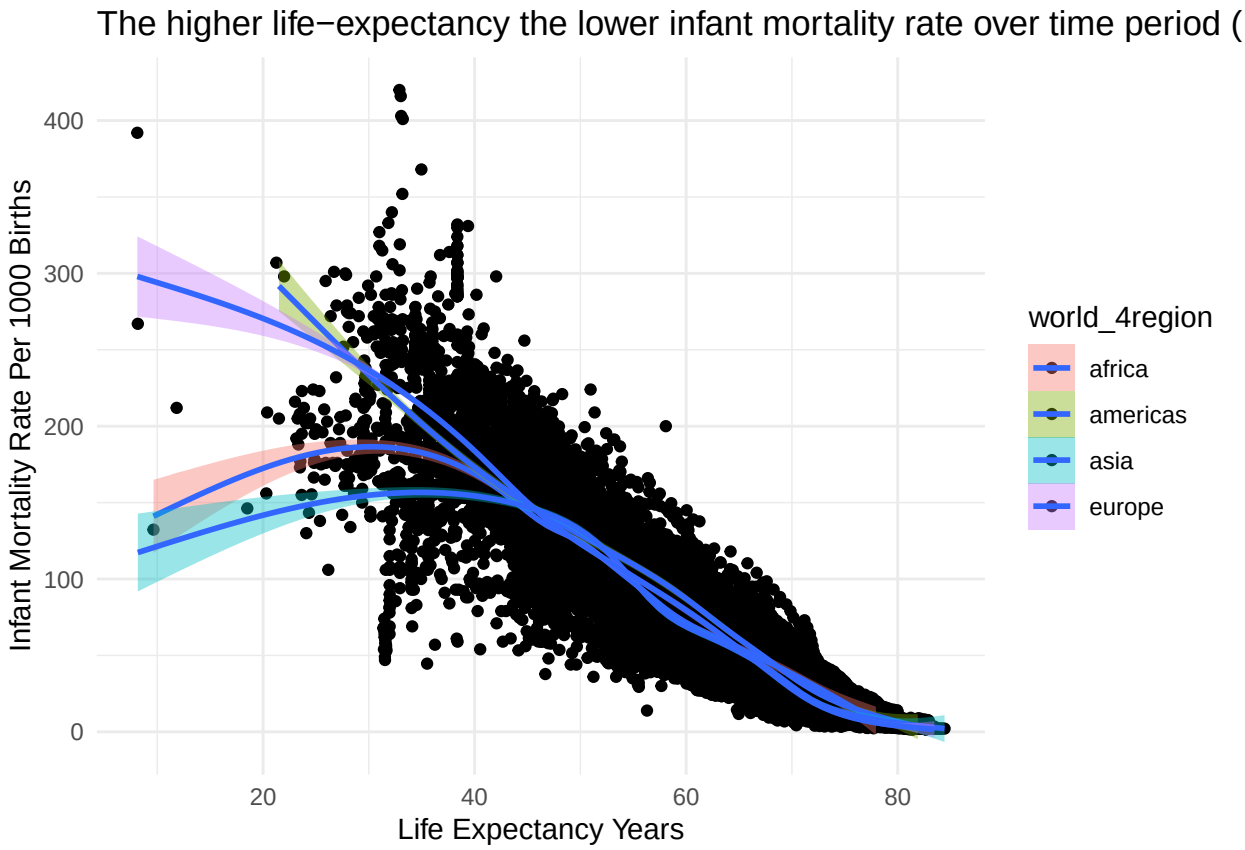
In Europe: The graph shows consistancy of the trend that when life expectancy increases, infant mortality rate decreases over time.

In America: Same as the trend in Europe.

In Asia: Both infant mortality rate(IMR) and life expectancy years(LEY) increase when life-expectancy is around 0-30. Then IMR decreased while LEY increased.

In Africa:Both infant mortality rate(IMR) and life expectancy years(LEY) increase when life-expectancy is around 0-35. Then IMR decreased while LEY increased.

```
setwd("~/Desktop/5110/HM3/ddf--gapminder--systema_globalis-master/countries-etc-datapoints")
life_expectancy_years <- read_csv(
  "ddf--datapoints--life_expectancy_years--by--geo--time.csv")
dbWriteTable(con, "life_expectancy_years", life_expectancy_years)

library(dplyr)

lifeE_infantM <- dbGetQuery(con, "SELECT DISTINCT
joined_infant_mortality_dt.geo,
joined_infant_mortality_dt.time,
joined_infant_mortality_dt.world_4region,
joined_infant_mortality_dt.infant_mortality_rate_per_1000_births,
life_expectancy_years.life_expectancy_years
FROM joined_infant_mortality_dt
INNER JOIN life_expectancy_years
WHERE life_expectancy_years.geo=joined_infant_mortality_dt.geo
AND life_expectancy_years.time=joined_infant_mortality_dt.time
")

dbWriteTable(con, "lifeE_infantM", lifeE_infantM)


ggplot(lifeE_infantM, mapping=aes(x=life_expectancy_years,
                                  y=infant_mortality_rate_per_1000_births,
                                  fill=world_4region)) +
  geom_point() + geom_smooth() +
  labs(x="Life Expectancy Years",
       y="Infant Mortality Rate Per 1000 Births",
       title=
"The higher life-expectancy the lower infant mortality rate over time period (1800-2015)") + theme_mini
```


The higher life−expectancy the lower infant mortality rate over time period (
Infant Mortality Rate Per 1000 Births
400
300
200
100
0
20
40
60
80
Life Expectancy Years
world_4region
africa
americas
asia
europe