Home-Work4 (Balaji Kothandaraman)

Home-Work 4

Section A

Problem 1

Section A uses “Domestic general government health expenditure (GGHE-D) per capita in Purchasing power parity(PPP) int$(Data by country)” and “Adult mortality rate (probability of dying between 15 and 60 years per 1000 population) (Data by country)” from World Health Organization (WHO) Download complete Domestic general government health expenditure(csv table) from: http://apps.who.int/ (http://apps.who.int/) gho/data/view.main.GHEDGGHEDpcPPPSHA2011v To read the data manual: https://www.who.int/data/ (https://www.who.int/data/) gho/indicator-metadata-registry/imr-details/4960 Download complete Adult mortality rate(csv table) from: http://apps.who.int/gho/data/view.main.1360 (http://apps.who.int/gho/data/view.main.1360) To read the data manual: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/64 (https://www.who.int/data/gho/indicator-metadata-registry/imr-details/64) Use read_csv() to import the dataset to R.

Import data to R, and appropriately make these two tables tidy. Then we only want to keep all the information from the table Adult mortality per 1000 population (Data by country)- join these two tables by Country and Year. You will get a new data table. Use head() to present the data.

```
##### Importing libraries

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────────
─── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

```
##### Reading csv
dggh<-read_csv('xmart.csv',skip=1,col_types =cols(Country = col_character(),  `2017` = col_doubl
e(),`2016`=col_double(),`2015` = col_double(),`2014` = col_double(),`2013` = col_double(), `2012
` = col_double(),`2011` = col_double(), `2010` = col_double(), `2009` = col_double(), `2008` = c
ol_double(), `2007` = col_double(), `2006` = col_double(), `2005` = col_double(), `2004` = col_d
ouble(), `2003` = col_double(), `2002` = col_double(),`2001` = col_double(),`2000` = col_double
()))

head(dggh)
```

```
## # A tibble: 6 x 19
##   Country `2017` `2016` `2015` `2014` `2013` `2012` `2011` `2010` `2009`
##   <chr>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Afghan…    NA     8.9    9.5    8.8    8.3    6.3    7.8    7.2    7.8
## 2 Albania   542.  512.   508.   493.   452.   424.   432.   426.   356.
## 3 Algeria   643.  672.   721.   672.   589.   587    483.   448.   465.
## 4 Andorra  2568. 2446   2461.  2372.  2196   2051.  1838   1755   1711.
## 5 Angola     86    81     87.7   95    118.   106.   109.   104.   158.
## 6 Antigu…   504.  484.   559.   598.   590.   547.   518.   556.   508.
## # … with 9 more variables: `2008` <dbl>, `2007` <dbl>, `2006` <dbl>,
## #   `2005` <dbl>, `2004` <dbl>, `2003` <dbl>, `2002` <dbl>, `2001` <dbl>,
## #   `2000` <dbl>
```

```
##### Finding the columns to be pivoted
columns<-colnames(dggh)
columns<-columns[2:19]
```

```
##### converting the columns
dggh_new<-dggh%>%pivot_longer(c(columns),names_to ="Year",values_to = "Expenditure PP")
dggh_new<-dggh_new%>%type_convert(col_types = cols(`Year`=col_double()))
head(dggh_new)
```

```
## # A tibble: 6 x 3
##   Country      Year `Expenditure PP`
##   <chr>       <dbl>            <dbl>
## 1 Afghanistan  2017             NA
## 2 Afghanistan  2016              8.9
## 3 Afghanistan  2015              9.5
## 4 Afghanistan  2014              8.8
## 5 Afghanistan  2013              8.3
## 6 Afghanistan  2012              6.3
```

```
##### Reading csv

AM<-read_csv('WHOSIS_000004.csv',skip=1,col_types = cols(Country = col_character(),
  `Year` = col_double(),
  `Both sexes` = col_double(),
  Male = col_double(),
  Female = col_double()))
```

```
##### tidying the data

AM<-AM%>%pivot_longer(c(`Both sexes`,`Male`,`Female`),names_to='sex', values_to='Adult Mortality
rate')
head(AM)
```

```
## # A tibble: 6 x 4
##   Country      Year sex        `Adult Mortality rate`
##   <chr>       <dbl> <chr>                      <dbl>
## 1 Afghanistan  2016 Both sexes                  245
## 2 Afghanistan  2016 Male                        272
## 3 Afghanistan  2016 Female                      216
## 4 Afghanistan  2015 Both sexes                  233
## 5 Afghanistan  2015 Male                        254
## 6 Afghanistan  2015 Female                      210
```

```
##### joining with Adult Mortality rate

new_df<-AM%>%left_join(dggh_new,by=c('Country','Year'))
head(new_df)
```

```
## # A tibble: 6 x 5
##   Country      Year sex        `Adult Mortality rate` `Expenditure PP`
##   <chr>       <dbl> <chr>                      <dbl>            <dbl>
## 1 Afghanistan  2016 Both sexes                  245              8.9
## 2 Afghanistan  2016 Male                        272              8.9
## 3 Afghanistan  2016 Female                      216              8.9
## 4 Afghanistan  2015 Both sexes                  233              9.5
## 5 Afghanistan  2015 Male                        254              9.5
## 6 Afghanistan  2015 Female                      210              9.5
```

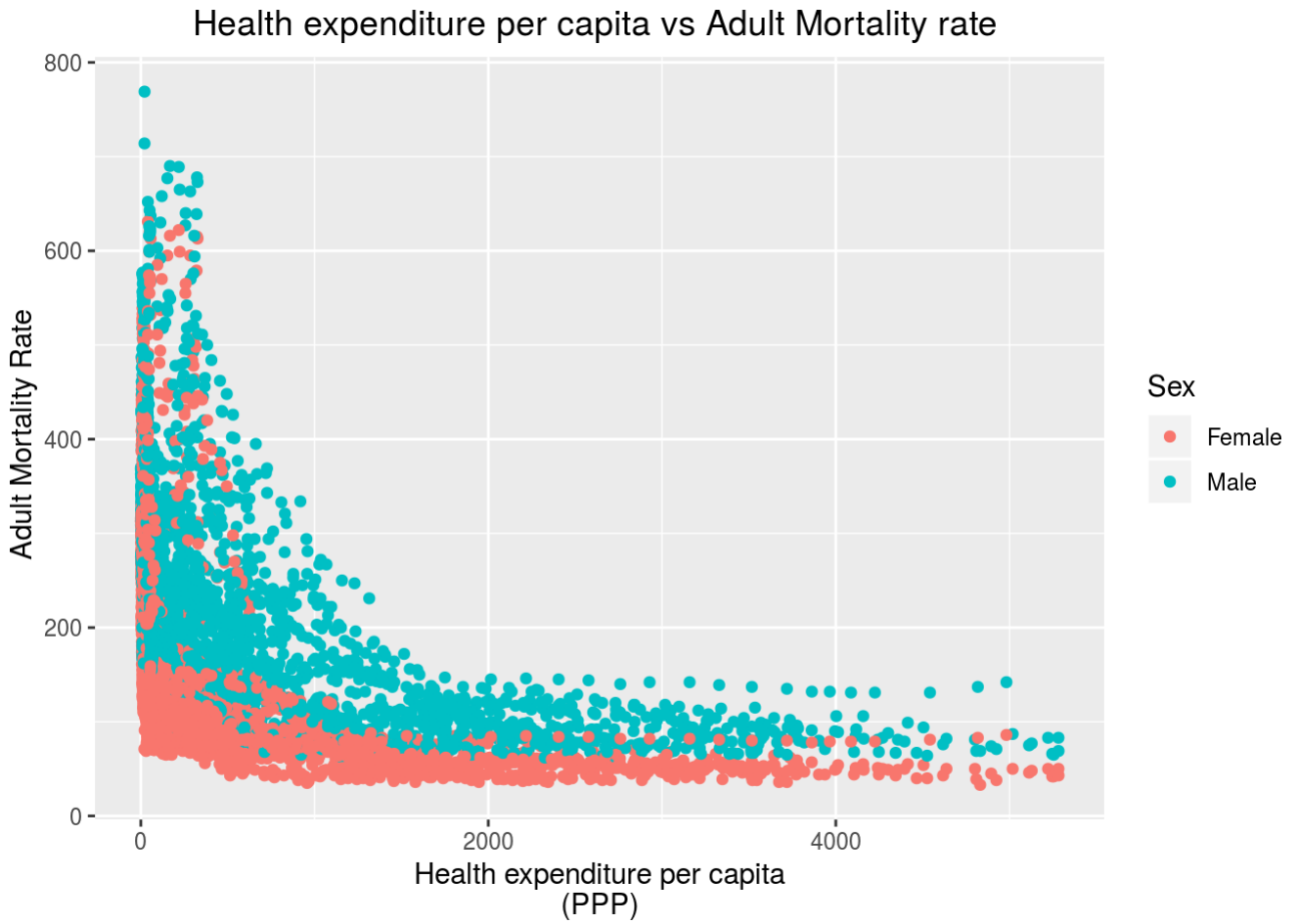
Problem 2

Still working on the tabe created in the problem 1. Create a scatterplot of health expenditure per capita (PPP) (x-axis) versus adult mortality rate(y-axis) over all the countries from 2000 to 2016. In this plot, using color aesthetic to visualize the difference only between “Female” and “Male” (No “Both sexes”). What do you notice about the difference between men and women? Does the gap of life expectancy between women and men (see: https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/ (https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/)) can explain your observation?

hint: Properly deal with the title by using read_csv(skip=?)

```
##### Plotting health expenditure vs adult mortality

new_df%>%filter(`sex`!='Both sexes')%>%ggplot(aes(`Expenditure PP`,`Adult Mortality rate`,color=
`sex`))+geom_point()+xlab('Health expenditure per capita
(PPP)')+ylab(' Adult Mortality Rate')+scale_color_discrete(name='Sex')+ggtitle('Health expenditu
re per capita vs Adult Mortality rate ')+theme(plot.title=element_text(hjust=0.5))
```



Interpretation:

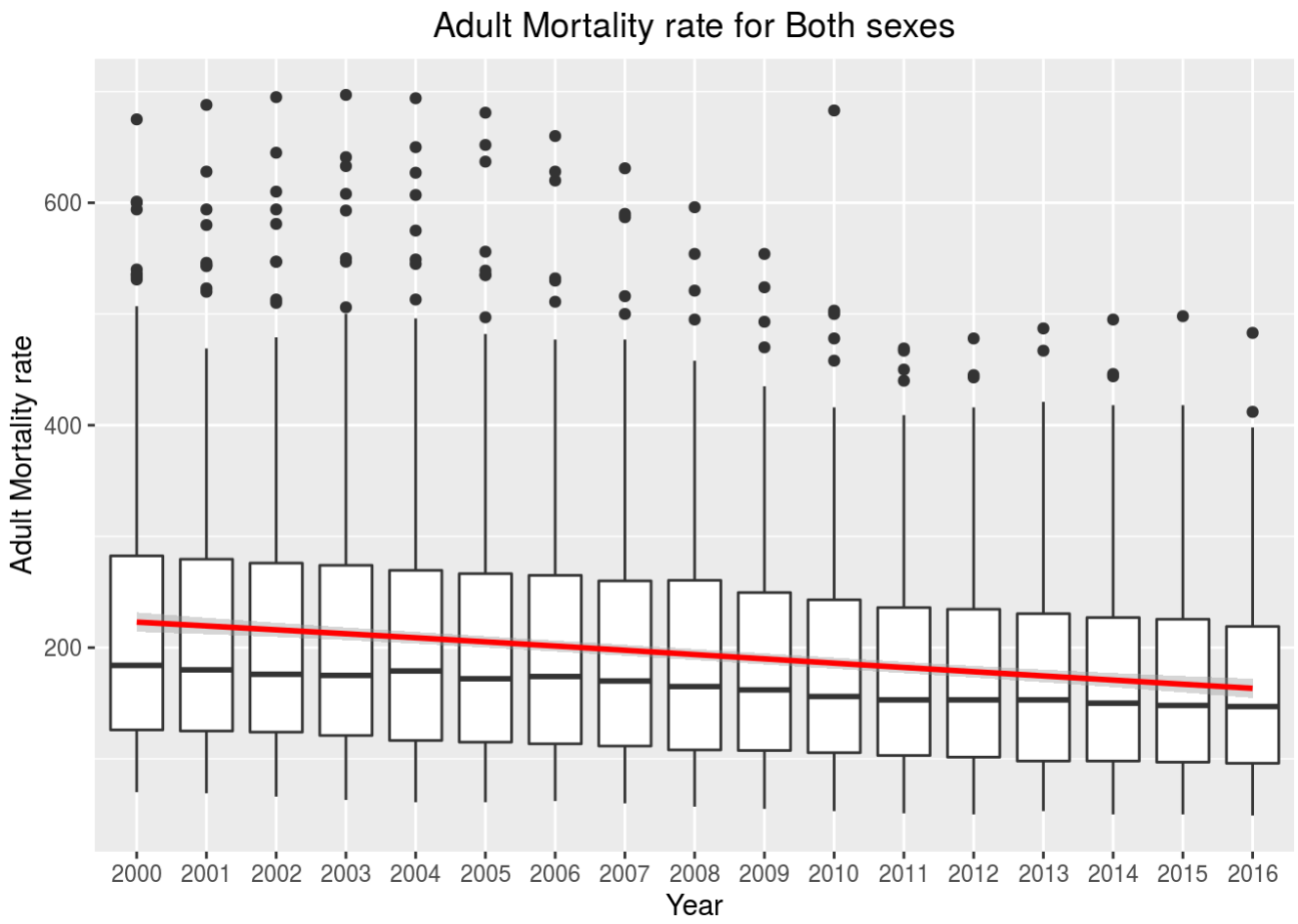
With Amount of Health expenditure spent on both males and females, most of the male have higher probability of death in comparison with female. It is also clear the more money spent on health expenditure, the probability of death is low in comparison with no money spent in health expenditure. The life expectancy of men is lesser in comparison with life expectancy women for most of the cases.

Problem 3

Still working on the above new data table. We would like to see what's the trend of the adult mortality rate of "Both sexes" over all the countries from 2000 to 2016. Visualize year (x-axis) versus Adult Mortality per 1000 population (y-axis) by plotting a boxplot. Include one smooth line (method="auto", aes(group=1)). What do you notice about the chart?

```
##### Plotting box plot
new_df%>%filter(`sex`=='Both sexes')%>%ggplot(aes(factor(`Year`),`Adult Mortality rate`))+geom_b
oxplot()+geom_smooth(method='auto',aes(group=1),color='red')+theme(plot.title=element_text(hjust
=0.5))+ggtitle('Adult Mortality rate for Both sexes')+theme(plot.title=element_text(hjust=0.5))+
xlab('Year')
```

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



Interpretation:

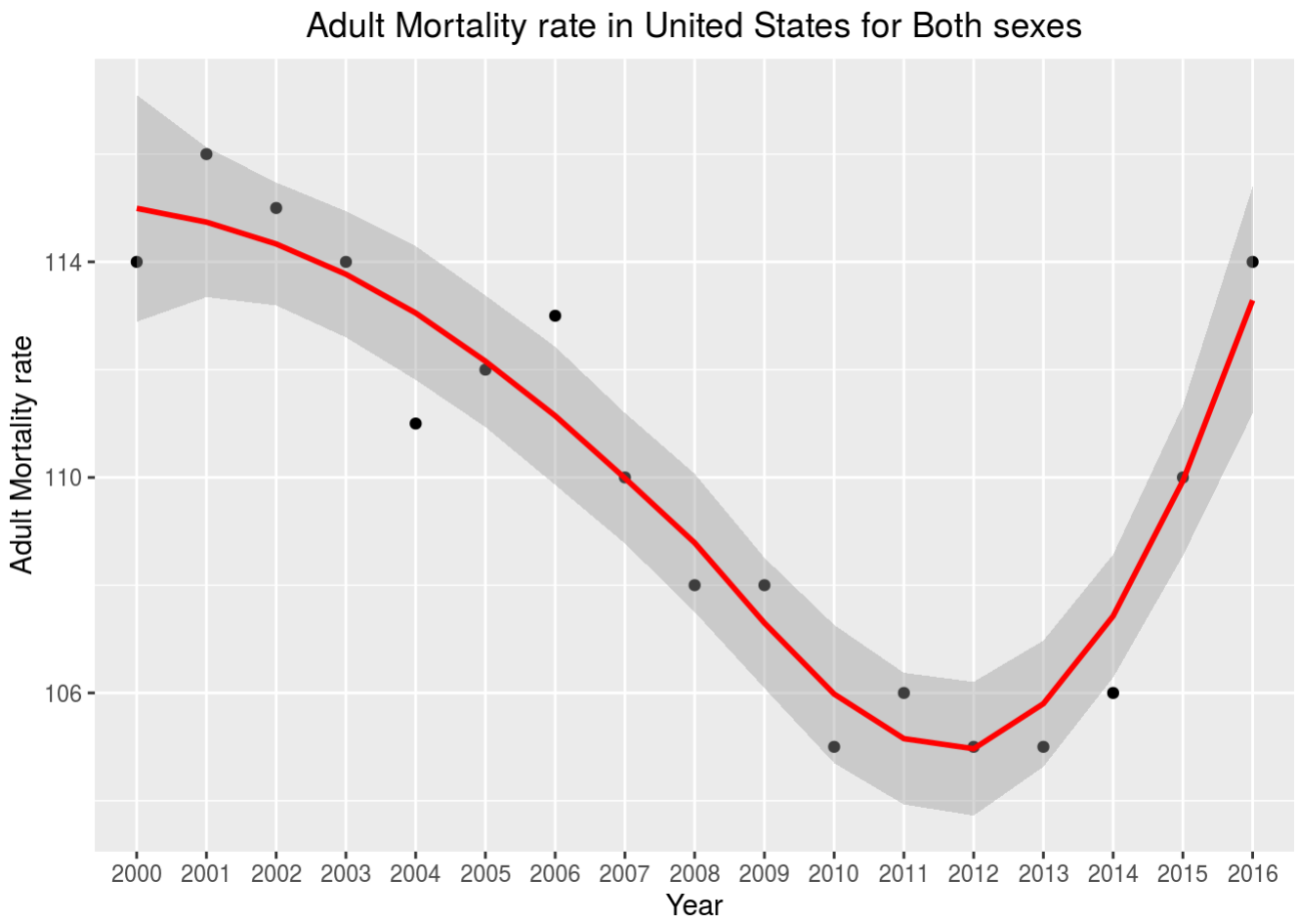
The boxplot shows the Adult mortaliity rate for Both sexes from 2000 to 2016. The 50 % of the people lie in the range of 175 and smooth line indicates how most the data lies and follows a pattern, the box plot also gives the outliers of Adult mortality rate where the most of the points lie above 420. The smooth curve decreases indicating increase in life expectancy as the years changes.

Problem 4

Still working on the data table. We are also interested in the trend of the adult mortality rate of Both sexes in US from 2000 to 2016. Visualize year (x-axis) versus Adult Mortality per 1000 population (y-axis) by plotting a scatterplot. Include one smooth line (method="auto", aes(group=1)). What do you notice about the trend compared to the problem 2? The CDC reported the data of leading causes of death and drug overdose deaths increasing in US (https://www.cdc.gov/nchs/hus/index.htm (https://www.cdc.gov/nchs/hus/index.htm)) and the birth rate decreasing (https://catalog.data.gov/dataset/births-and-general-fertility-rates-united-states-1909-2013 (https://catalog.data.gov/dataset/births-and-general-fertility-rates-united-states-1909-2013)). Intuitively, could this explain the concave-up shape?

```
##### Plotting scatter plot
new_df%>%filter(Country%in%c('United States of America') &`sex`=='Both sexes')%>%ggplot(aes(fact
or(`Year`),`Adult Mortality rate`))+geom_point()+geom_smooth(method='auto',aes(group=1),color='r
ed')+ggtitle('Adult Mortality rate in United States for Both sexes')+theme(plot.title=element_te
xt(hjust=0.5))+xlab('Year')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Interpretation:

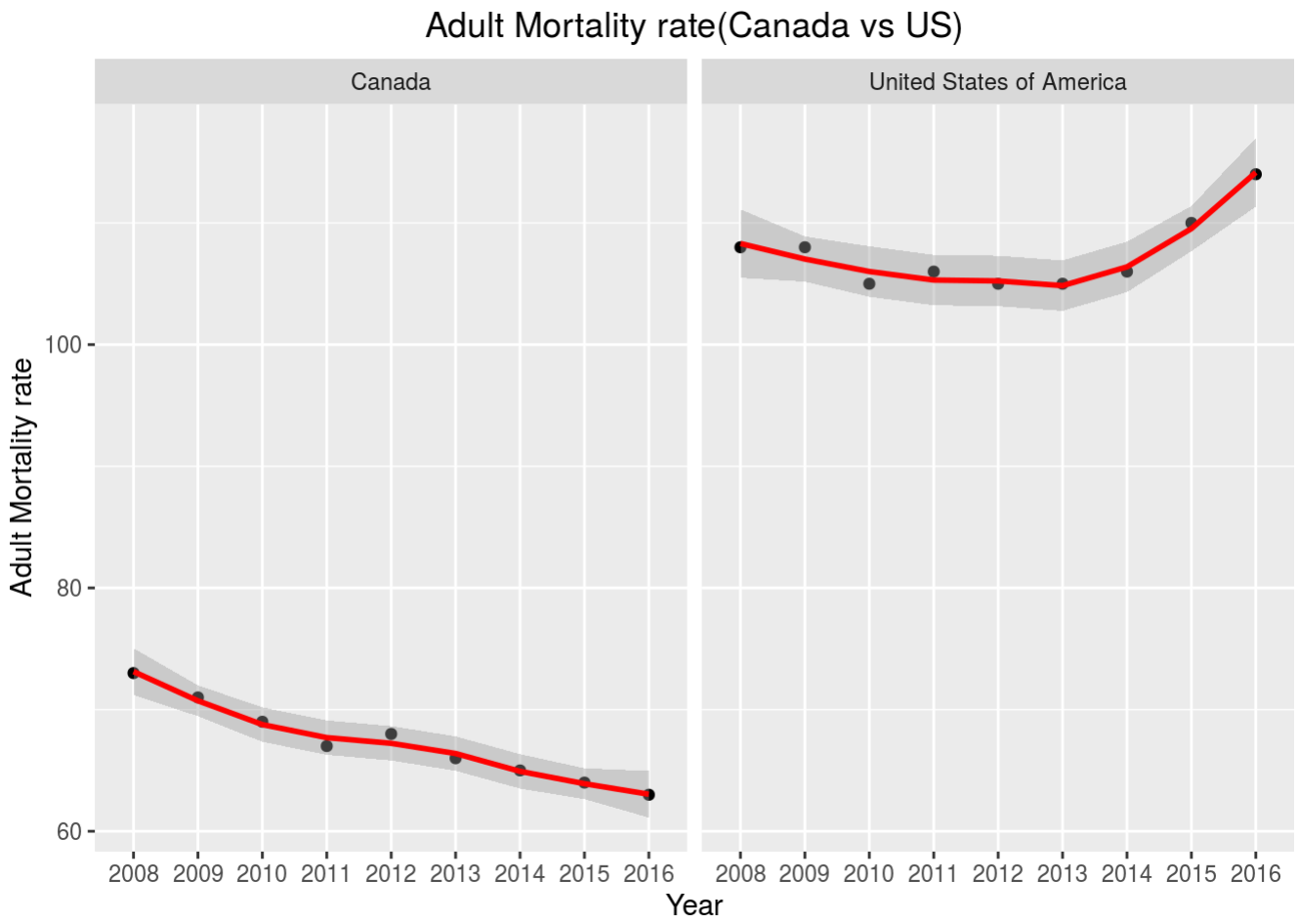
The curve indicates that most of cases in US lies between 100 and 120. The probability of death is decreasing linearly but before peaking at 2011 and then increasing linearly. In comparison with problem 2 where more money spent in health expenditure has resulted higher expectancy but in states the expenditure spent on health is doubled when compared to 2000 and still the number is probability of death is high. This sudden upward movement is due to overdose of drugs and decrease in birth rates.

Problem 5

Still working on the table created in problem 1. Compare the adult mortality rate between US and Canada from 2008 to 2016. Visualize year (x-axis) versus Adult Mortality per 1000 population (y-axis) by plotting a scatterplot. Using faceting to visualize the difference between US and Canada. Include one smooth line (method="auto", aes(group=1)) for each facet. What do you notice about the difference?

```
##### plotting scatter plot
new_df%>%filter(Country%in%c('United States of America','Canada')& Year%in%c(2008:2016)& `sex`==
'Both sexes')%>%ggplot(aes(factor(`Year`),`Adult Mortality rate`))+geom_point()+facet_grid(~Coun
try)+geom_smooth(aes(group=1),method='auto',color='red')+ggtitle('Adult Mortality rate(Canada vs
US)')+theme(plot.title=element_text(hjust=0.5))+xlab('Year')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Interpretation:

The Adult Mortality rate in Canada is comparatively low in comparison with United States of America for the years 2008 to 2016. The probability of rate of death in canada is decreasing reaching the lowest in 2016 almost reaching 60, while the probability of death has increased in United states reaching the maximum in 2016 with number more than 110.

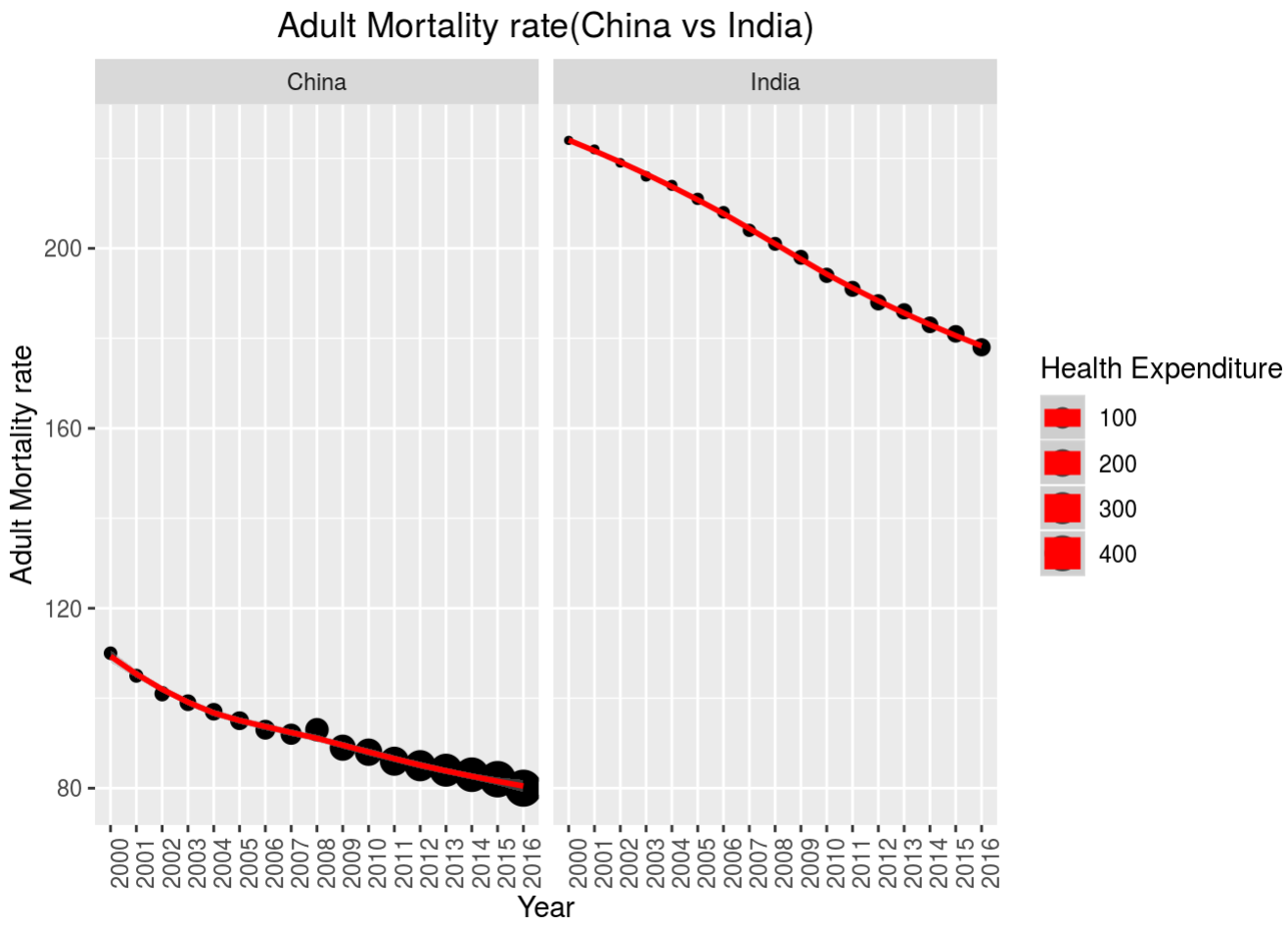
Problem 6

Still working on the table created in problem 1. Select two countries you are interested in, and do the same steps as you did in the problem 5.(Set the period of time from 2000 to 2016. Using size aesthetic to visualize the health expenditure. Properly deal with the label)

```
##### Plotting scatter plot

new_df%>%filter(`Country`%in%c('India','China')&`sex`=='Both sexes')%>%ggplot(aes(factor(`Year
`),`Adult Mortality rate`,size=`Expenditure PP`))+geom_point()+facet_grid(~Country)+geom_smooth
(method='auto',aes(group=1),color='red')+scale_size_continuous(name='Health Expenditure')+ggtitl
e('Adult Mortality rate(China vs India)')+theme(plot.title=element_text(hjust=0.5))+theme(axis.t
ext.x = element_text(angle=90,hjust=1))+xlab('Year')
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


Adult Mortality rate(China vs India)
China
India
Adult Mortality rate
200
160
120
80
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year
Health Expenditure
100
200
300
400