```
# cleaning the environment
rm(list = ls())

# libraries to be needed
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
```

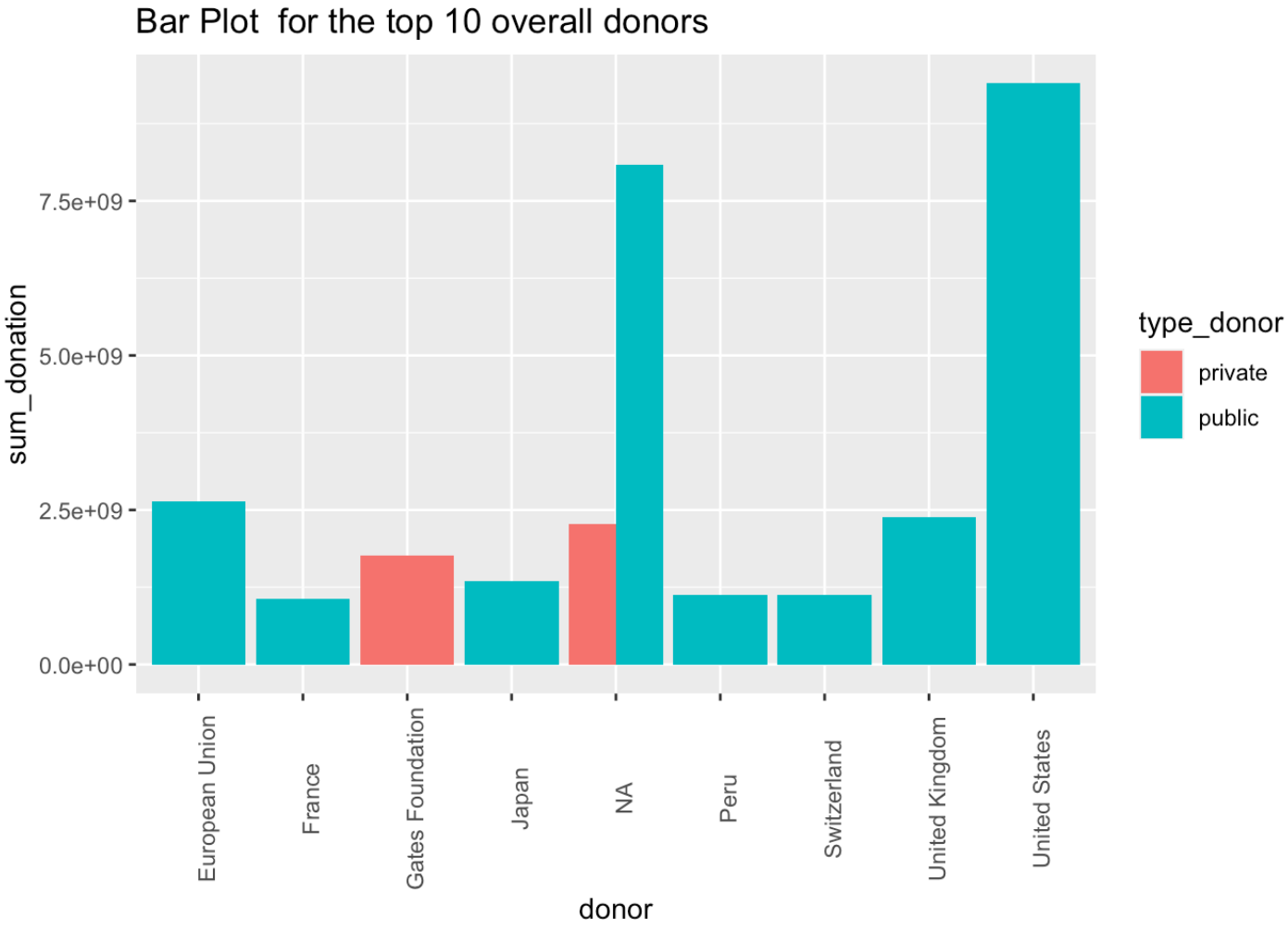
```
library(tidyr)
library(ggplot2)
library(ggpubr)

# importing the data
data=read.csv(file="~/Desktop/R/Material for week 4-20221014/io_income_rs.csv")

# NA has been considered as not applicable by R, changing to character variable
data$donor= ifelse(is.na(data$donor),"NA",data$donor)

# omiting the other NAs
data=na.omit(data)
```

Question 1

```
# grouping by donor and adding the donation amount to get the total, then arrangin
g it into descending order and slicing out the first 10 data
top10= data %>% group_by(donor, type_donor) %>%
  summarise(sum_donation=sum(amount_nominal))
```

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

```
#arranging in descending order
top10=top10%>%
    arrange(desc(sum_donation))
# taking the top 10
top10= top10[1:10,]

# generating the bar plot for top 10 overall donors
ggplot(data=top10, aes(x=donor, y=sum_donation, fill=type_donor),las=2) +
geom_bar(stat="identity", position=position_dodge())+ theme(axis.text.x = element_
text(angle = 90))+ ggtitle("Bar Plot  for the top 10 overall donors")
```

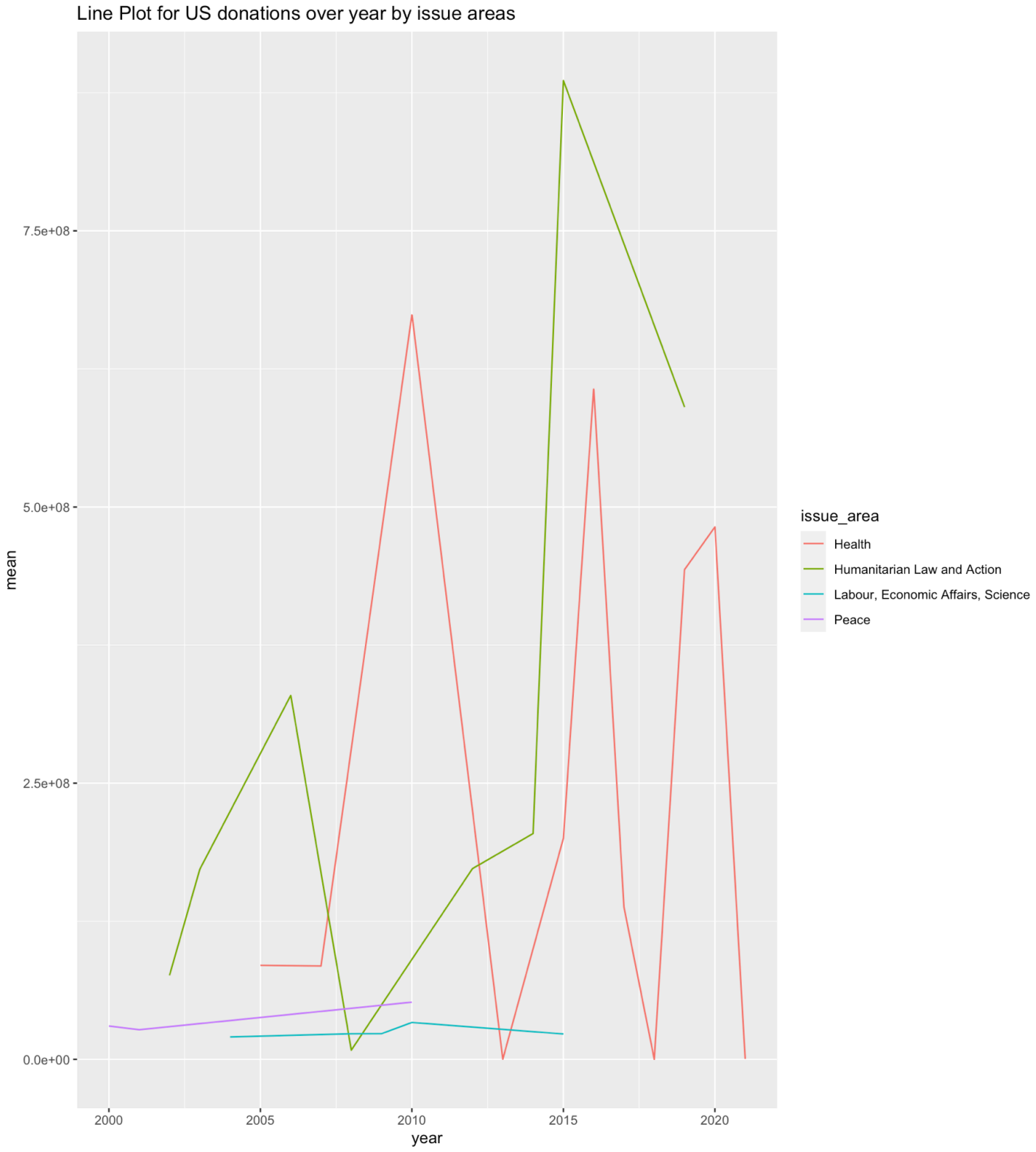


Question 2

```
#filtering the united states from donor
US= data %>% filter(donor=="United States")
# sorting by year
US= US[order(US$year),]
#calculating mean for the repeated measures
US1 = US %>% group_by(year,issue_area) %>% summarise(mean=mean(amount_nominal))
```

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

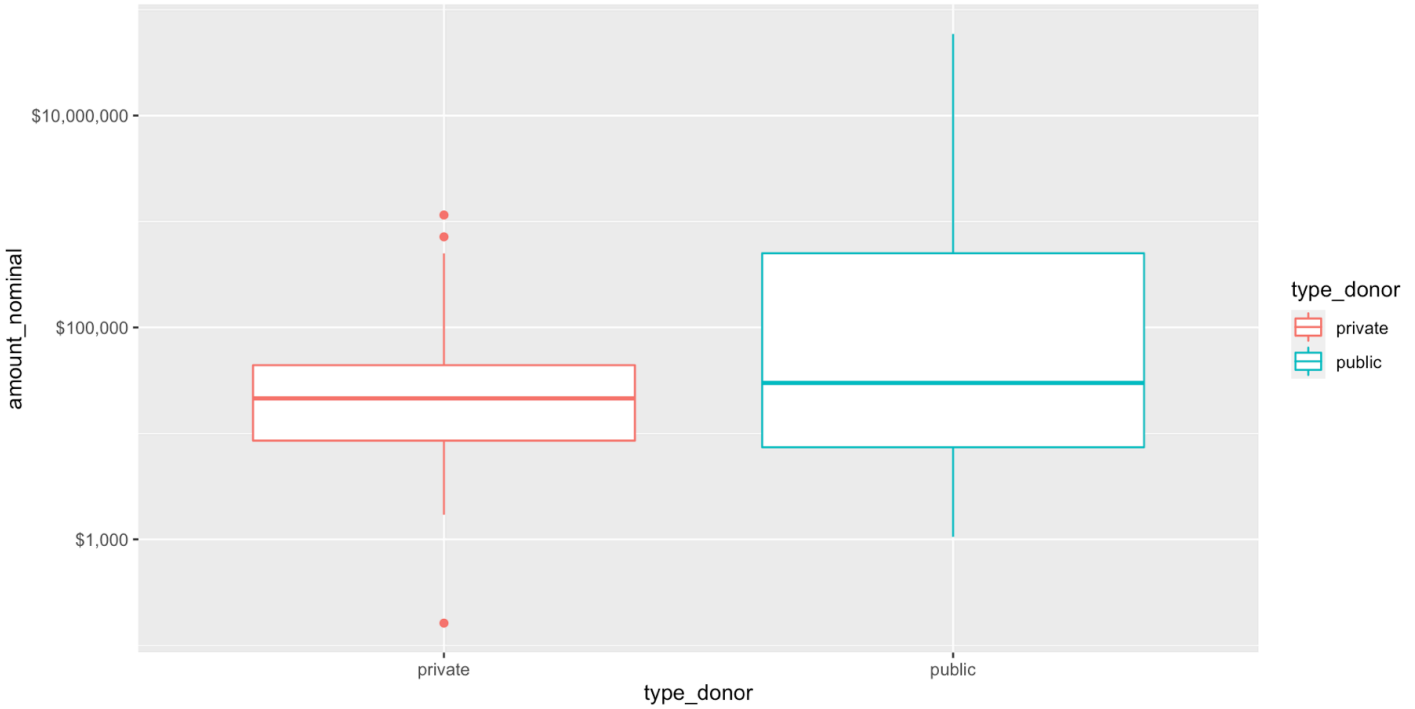
```
# line plot, colored by the respective issue areas
US1 %>%
  ggplot( aes(x=year, y=mean, group=issue_area, color=issue_area)) +
    geom_line()+ ggtitle("Line Plot for US donations over year by issue areas")
```



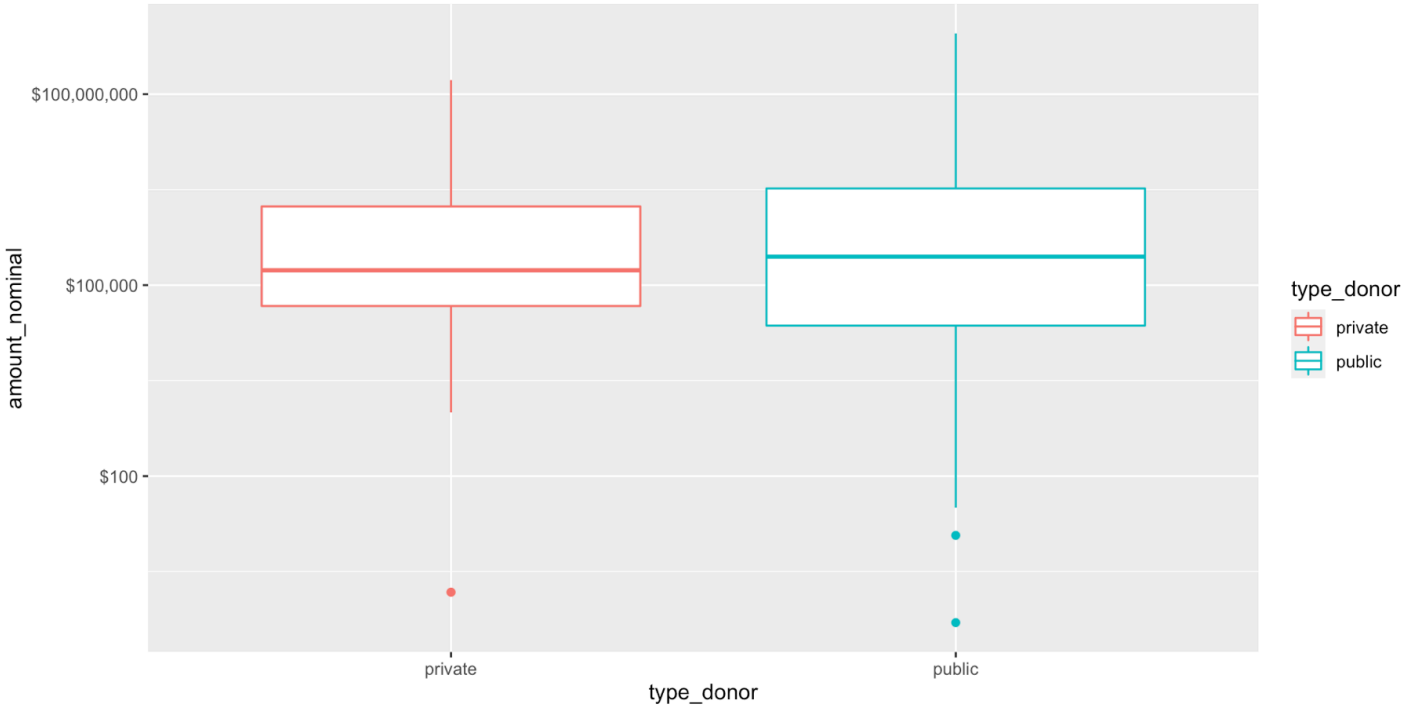
Question 3

```
# filtering public and private donor type from data
box = data %>% filter(type_donor %in% c("public","private"))

# filtering the year 2000
box2000=box %>% filter(year == 2000)

# box plot for the year 2000
box_plot_2000=ggplot(box2000, aes(x = type_donor, y = amount_nominal, color =type_
donor)) +  geom_boxplot()+scale_y_log10(labels = scales::dollar)+ ggtitle("Box Plo
t for distribution of all donations in the year 2000 \n comparing the type of dono
rs")

# filtering the year 2020
box2020=box %>% filter(year == 2020)

# box plot for the year 2020
box_plot_2020=ggplot(box2020, aes(x = type_donor, y = amount_nominal, color =type_
donor)) +  geom_boxplot()+scale_y_log10(labels = scales::dollar)+ ggtitle("Box Plo
t for distribution of all donations in the year 2020 \n comparing the type of dono
rs")

# joining the 2 box plots of year 2000 and 2020 one upon other
ggarrange(box_plot_2000, box_plot_2020,
                    labels = c("2000", "2020"),
                    ncol = 1, nrow = 2)
```

2000 Box Plot for distribution of all donations in the year 2000 comparing the type of donors



2020 Box Plot for distribution of all donations in the year 2020 comparing the type of donors



From the above box plot, we can clearly observe that, there are outliers in the year 2000 for private donor type(dots in red color) and the outliers in the year 2020 for both private(red dots) and public(blue dots) donor type.

Question 4

```
ggplot(data, aes(x = year, y = amount_nominal))+
  geom_point(aes(shape = type_donor)) +
  facet_wrap(~issue_area) +
  geom_smooth(colour="red")+ggtitle("Scatter Plot for all donation by year per don
or type")
```

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