Immigration to Canada

Group C
2022/01/10

The dataset: immigrants count to Canada from 1980 to 2013.
Source: United Nations - Population Division - Department of Economic and Social Affairs.
Consists of: Immigrants record from 150+ countries to Canada between 1980 to 2013.

Import Necessary Libraries

```
library(reshape) #For data manipulation
library(dplyr) #For data manipulation
library(tidyr) #For data manipulation
library(corrplot) #For data plotting
library(ggplot2) #For data plotting
library(ggpubr) #Foe sub plotting
library(waffle) # for plotting waffle chart
library(wordcloud) # for plotting word cloud
```

For simplicity purposes, there are two dataframes after cleaning,
as differnt visualizations require different rows and columns etc.
However, both dataframes are 99% similar with minor differences only.

```
df = read.csv("canadian_immigration_data.csv")
DF = read.csv("UpdatedDF.csv")
```

Renaming Years columns as "charecters" as R does not accept column name to be started with an integer.

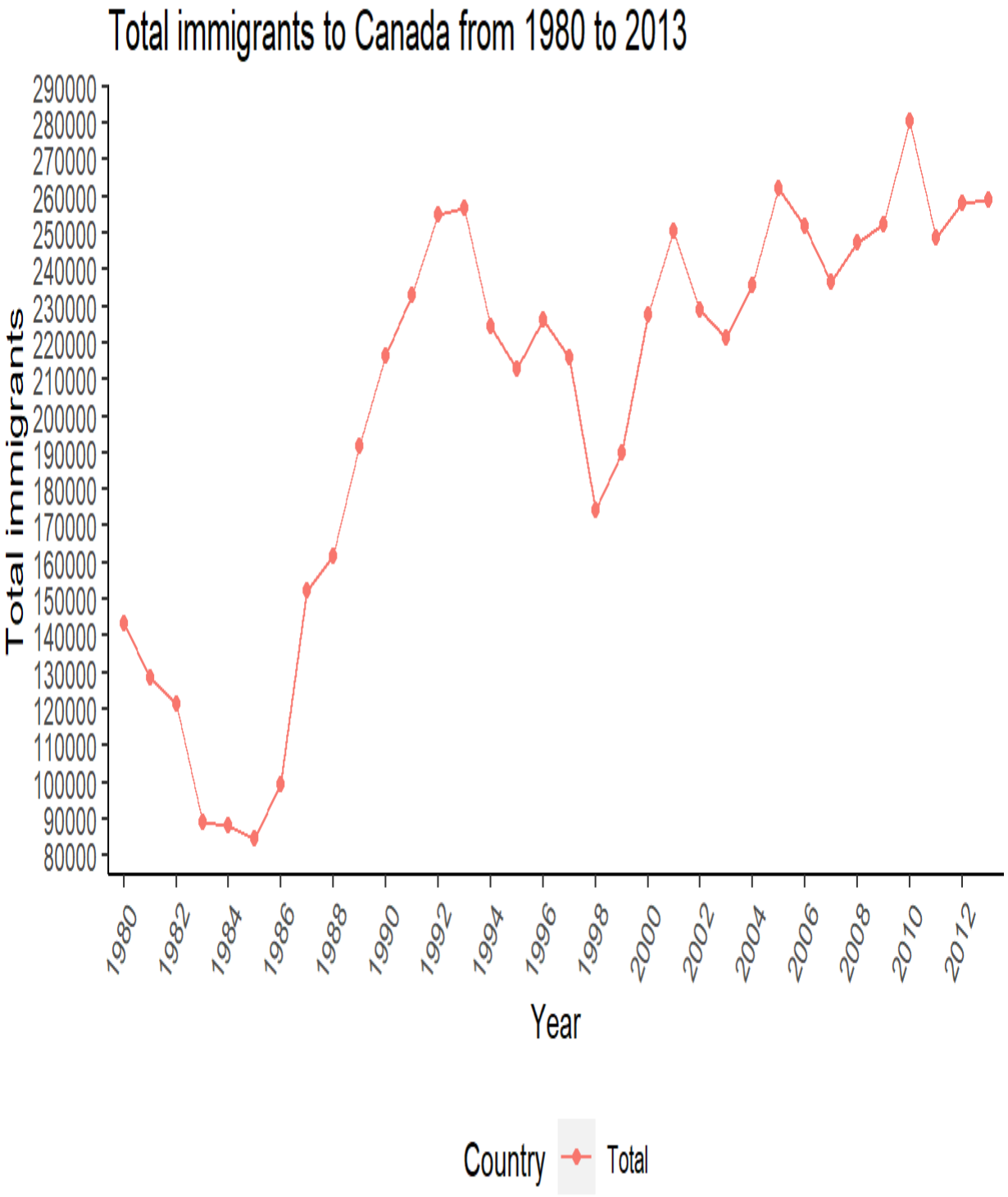
```
names(df)[4:37] <- c("1980", "1981", "1982", "1983", "1984",
                     "1985", "1986", "1987", "1988", "1989",
                     "1990", "1991", "1992", "1993", "1994",
                     "1995", "1996", "1997", "1998", "1999",
                     "2000", "2001", "2002", "2003", "2004",
                     "2005", "2006", "2007", "2008", "2009",
                     "2010", "2011", "2012", "2013")
```

Descriptive Analysis Questions

Q1 - How many total immigrants to Canada from 1980 to 2013.

Visualized by line chart. Displayed.

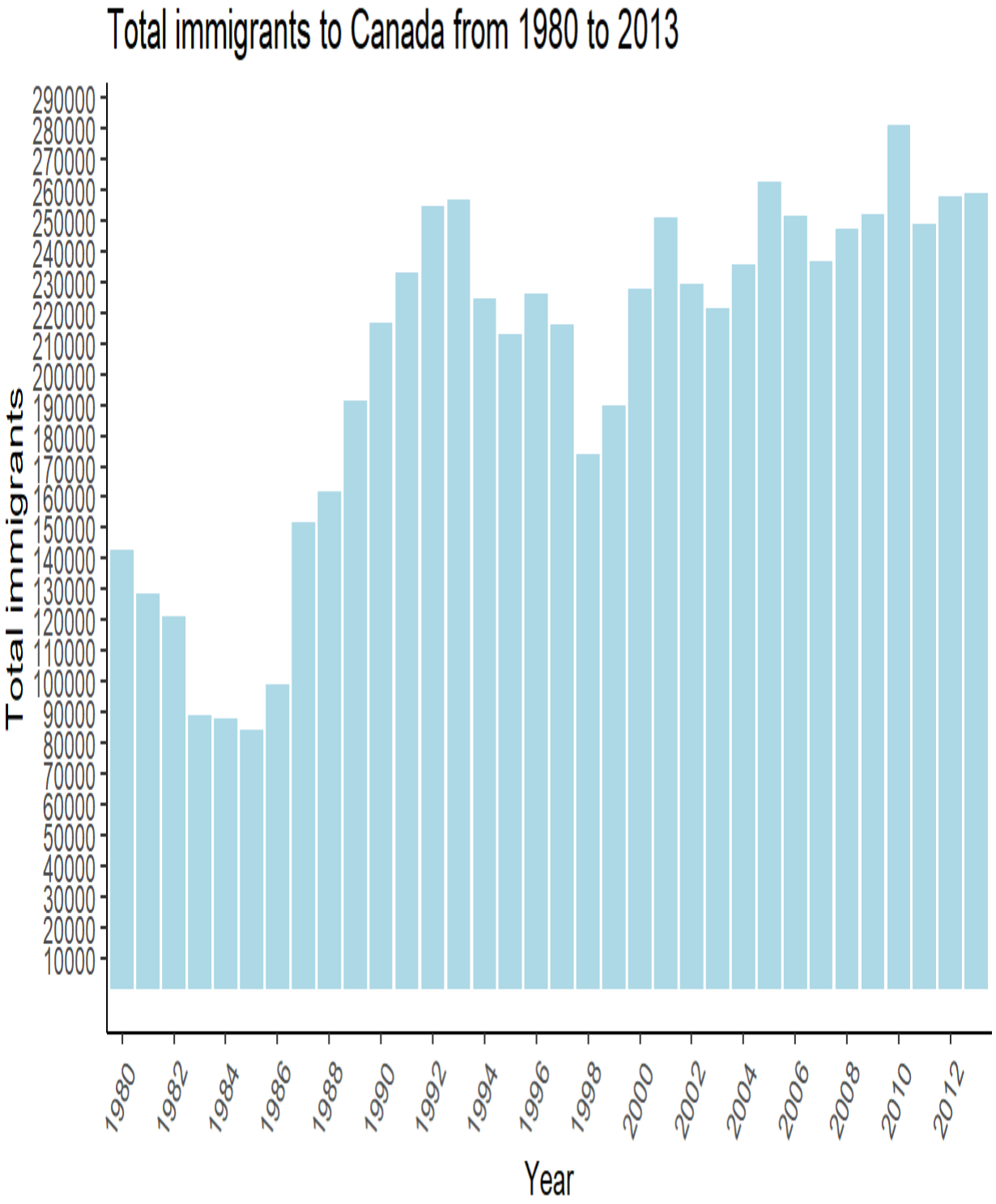
```
Q1Aplot <- df %>%
  filter(Country=="Total") %>%
  select(1,4:37) %>%
  gather(,key = "year", value = "number", 2:35) %>%
  ggplot(aes(x=year, y=number))+
  geom_line(aes(group=Country, color=Country))+
  geom_point(aes(color=Country))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="Total immigrants",title="Total immigrants to Canada from 1980 to 2013")+
  scale_x_discrete(breaks = seq(1980, 2013, by = 2))+ #A jump of 2 years
  scale_y_continuous(breaks = seq(10000, 300000, by = 10000)) #Scale y axis
Q1Aplot
```



Visualized by bar chart. Displayed.

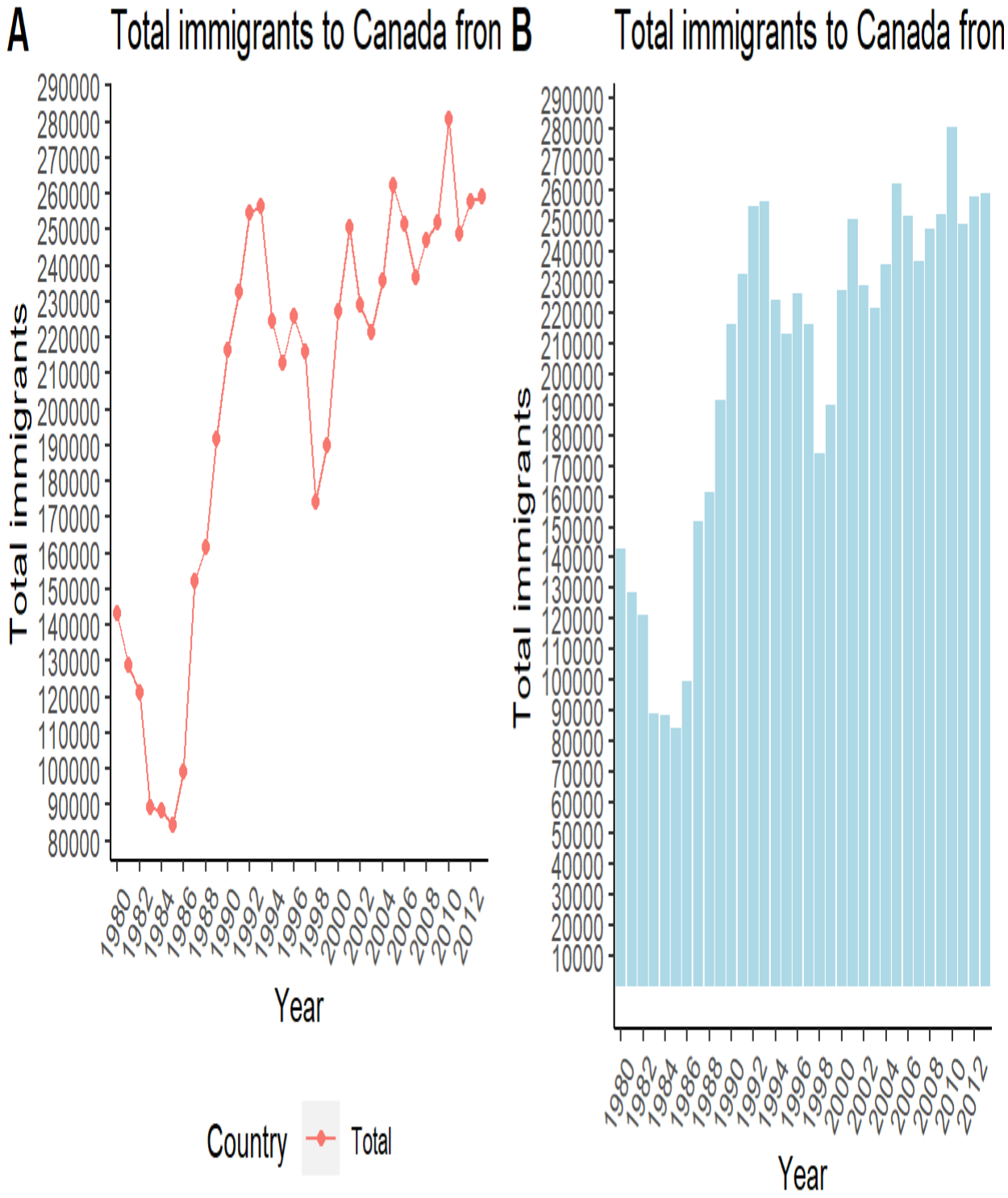
```
Q1Bplot <- df %>%
  filter(Country=="Total") %>%
  select(1,4:37) %>%
  gather(,key = "year",value = "number", 2:35) %>%
  ggplot(aes(x=year,y=number))+
  geom_col(fill="lightblue")+
  # scale_y_continuous(expand = c(0,0))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="Total immigrants",title="Total immigrants to Canada from 1980 to
2013")+
  scale_x_discrete(breaks = seq(1980, 2013, by = 2))+ #A jump of 2 years
```

```
  scale_y_continuous(breaks = seq(10000, 300000, by = 10000)) #Scale y axis
Q1Bplot
```



Combine two plots.

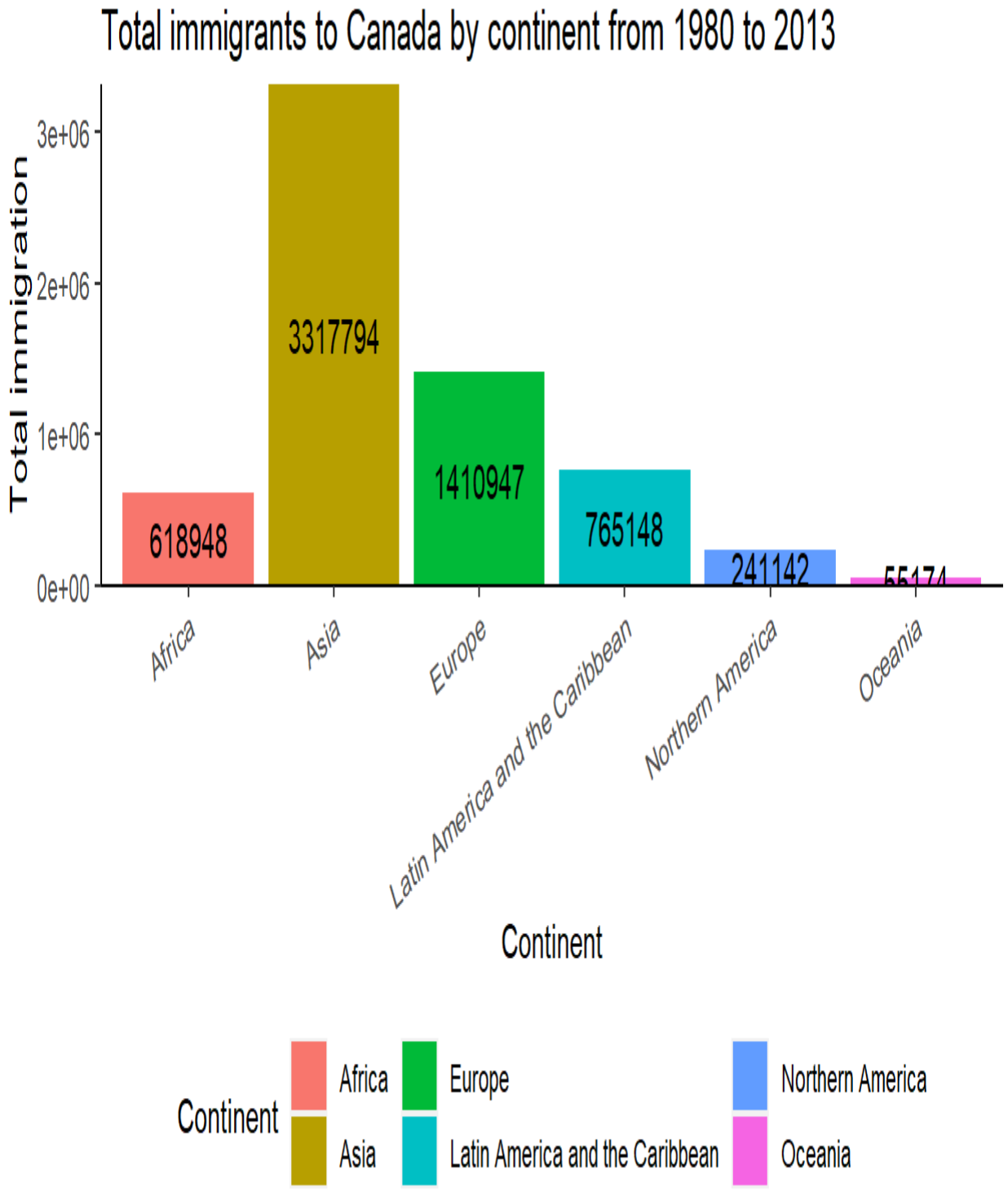
```
 A <- ggarrange(Q1Aplot, Q1Bplot,
           labels = c("A", "B"),
            ncol = 2, nrow = 1)
A
```



Q2 - How many total immigrants to Canada by continent from 1980 to 2013.

Visualized by bar chart. Displayed.

```
Q2Aplot <- ggplot(data = df[196:201,], aes(x=Continent,y=Total, fill=Continent))+
  geom_col()+
  scale_y_continuous(expand = c(0,0))+
  geom_text(aes(label=Total),
            position = position_stack(vjust = 0.5))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        axis.text.x = element_text(angle=30, hjust=1))+
  labs(x="Continent",y="Total immigration",title="Total immigrants to Canada by
continent from 1980 to 2013")
Q2Aplot
```



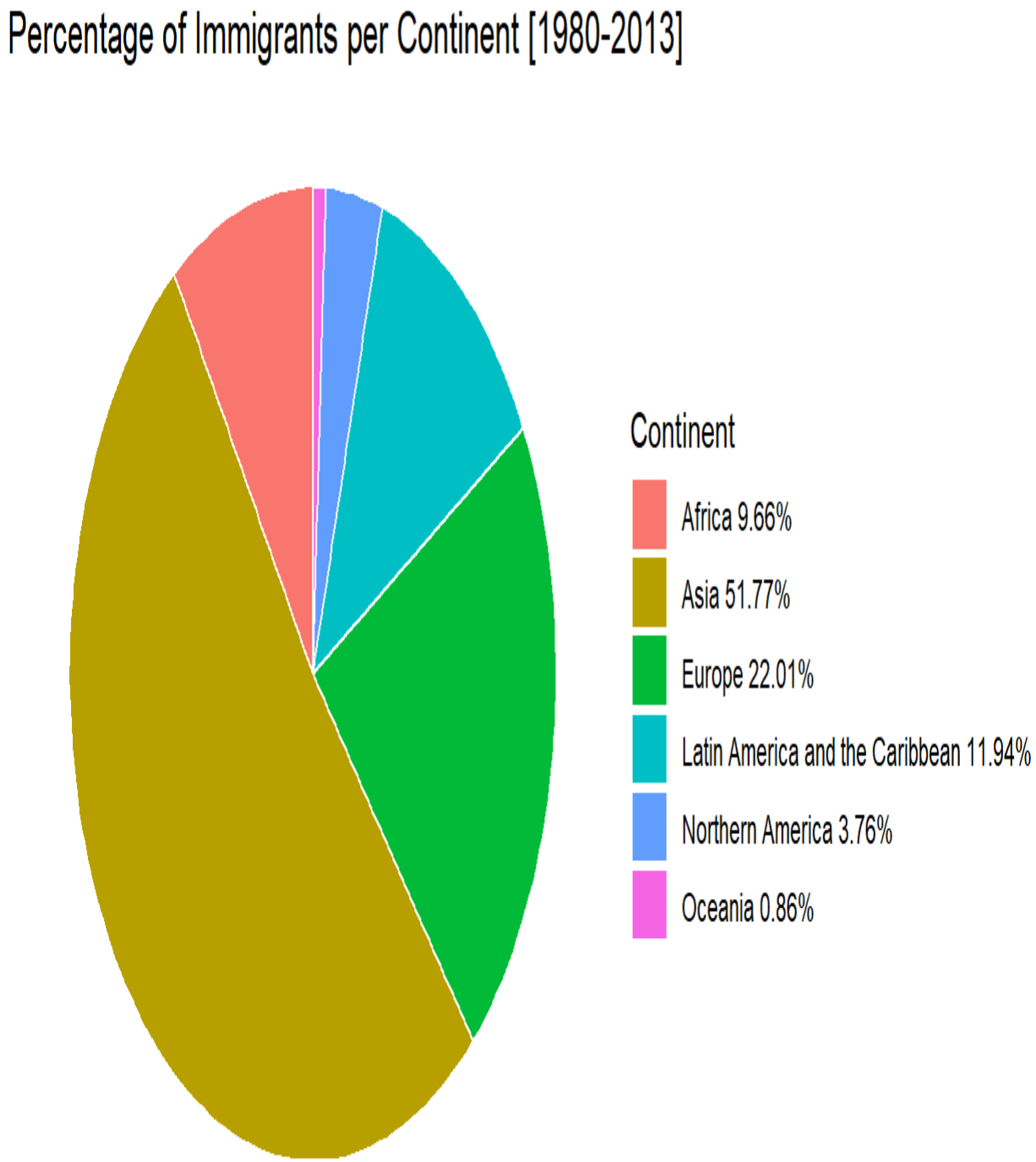
Visualized by pie chart. Displayed.

```
#Creating a sub dataset

Q2DF <- DF[-c(196,197),] #Remove last two rows as they do not have a continent
Q2DF <- tapply(Q2DF$Total,Q2DF$Continent,sum) #Sum total immigrants per continent.
Q2DF <- as.data.frame.table(Q2DF) #Convert it to a data frame
names <- c("Africa", "Asia", "Europe", "Latin America and the Caribbean","Northern
America","Oceania")
percentage <- round(Q2DF$Freq/sum(Q2DF$Freq)*100,2) #Count the percentage of each
Continent
lebals <- paste(names, percentage) # add percents to labels
lebals <- paste(lebals,"%",sep="") # add % to labels
Q2DF$Continent <- lebals #Add the labels as a new column "Continent"
Q2DF<- Q2DF[,-1] #Remove the old column of names
names(Q2DF)[1] <- "Total" #Rename Freq to Total
```

```
Q2DF #Final DF for continent
```

	Total	Continent
1	618948	Africa 9.66%
2	3317794	Asia 51.77%
3	1410947	Europe 22.01%
4	765148	Latin America and the Caribbean 11.94%
5	241142	Northern America 3.76%
6	55174	Oceania 0.86%

```
# Visualization.
Q2Bplot <- ggplot(Q2DF, aes(x="", y=Total, fill=Continent)) +
  geom_bar(stat="identity", width=1, color="White") +
  coord_polar("y", start=0) +
  ggtitle("Percentage of Immigrants per Continent [1980-2013]") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme_void() # remove background, grid, numeric labels
Q2Bplot
```

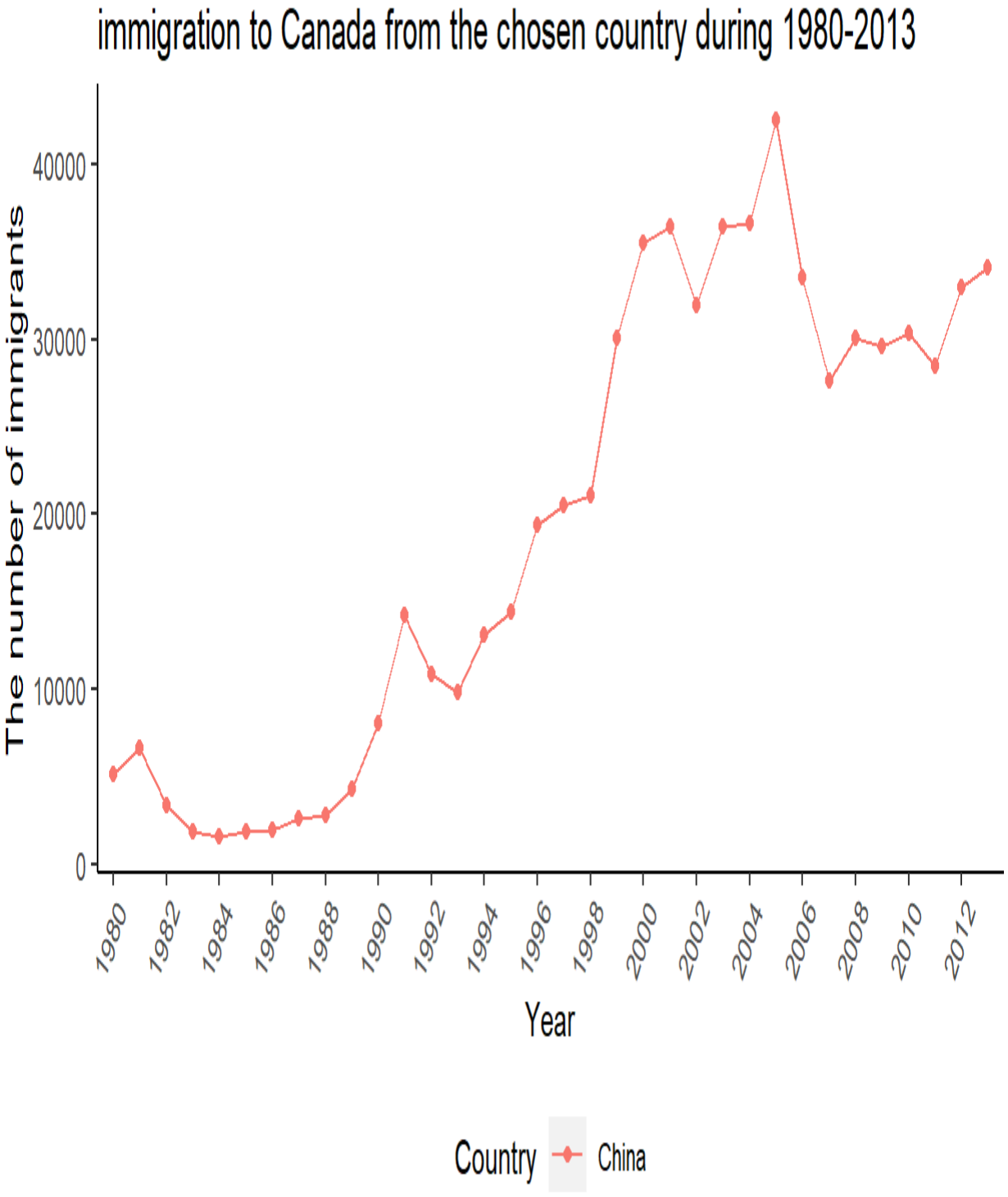


Q3 - How many immigrants to Canada by country from 1980-2013.

Visualized by line chart. AUTOMATED.

```
Q3Aplot <- df %>%
  filter(Country=="China") %>% #Here the country needs to be automated.
  select(1,4:37) %>%
  gather(,key = "year", value = "number", 2:35) %>%
  ggplot(aes(x=year, y=number))+
  geom_line(aes(group=Country, color=Country))+
  geom_point(aes(color=Country))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="The number of immigrants",title="immigration to Canada from the
```

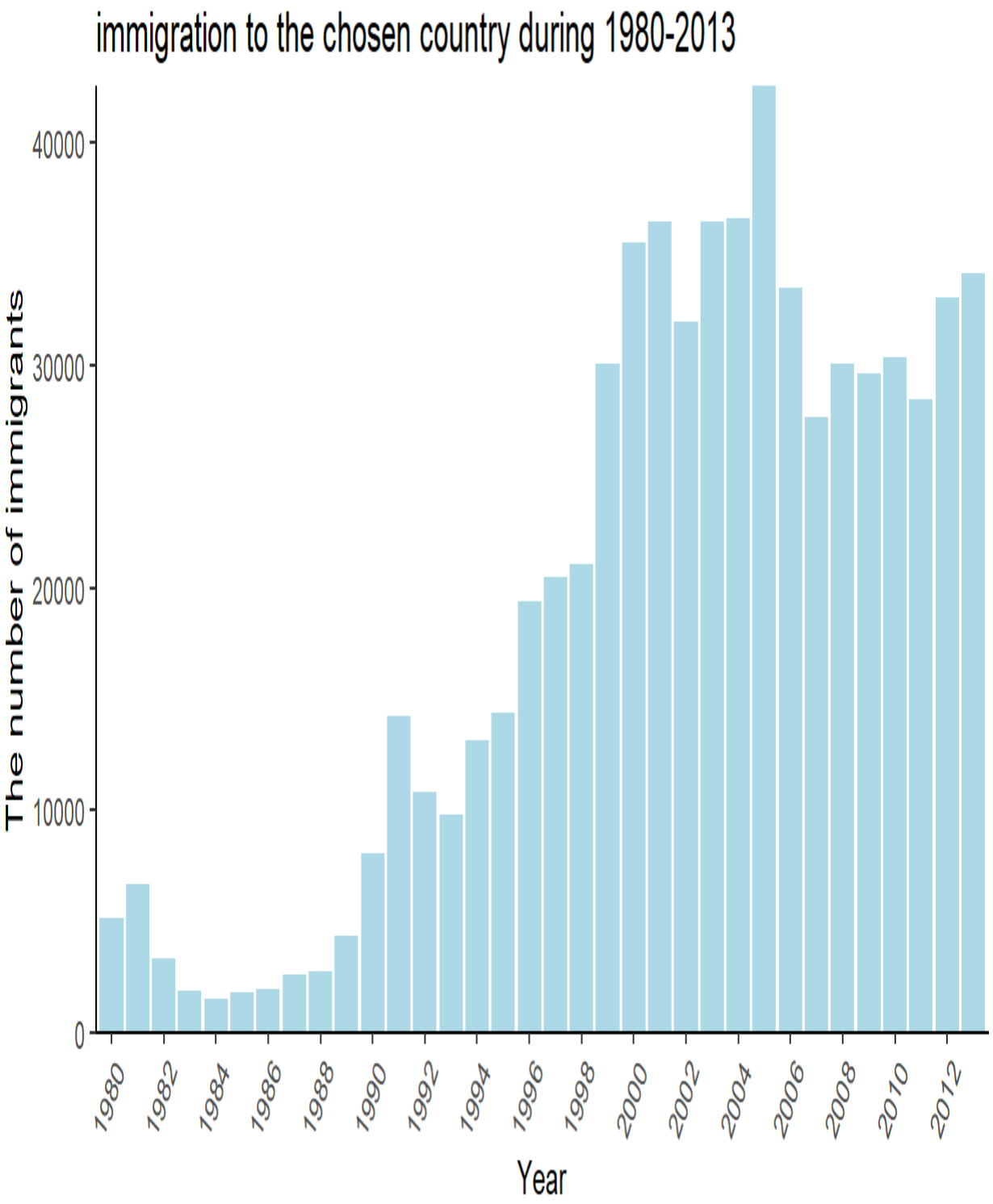
```
chosen country during 1980-2013") +
  scale_x_discrete(breaks = seq(1980, 2013, by = 2)) #A jump of 2 years
Q3Aplot
```



Visualized by bar chart. AUTOMATED.

```
Q3Bplot <- df %>%
  filter(Country=="China") %>%
  select(1,4:37) %>%
  gather(,key = "year",value = "number", 2:35) %>%
  ggplot(aes(x=year,y=number))+
  geom_col(fill="lightblue")+
  scale_y_continuous(expand = c(0,0))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
```

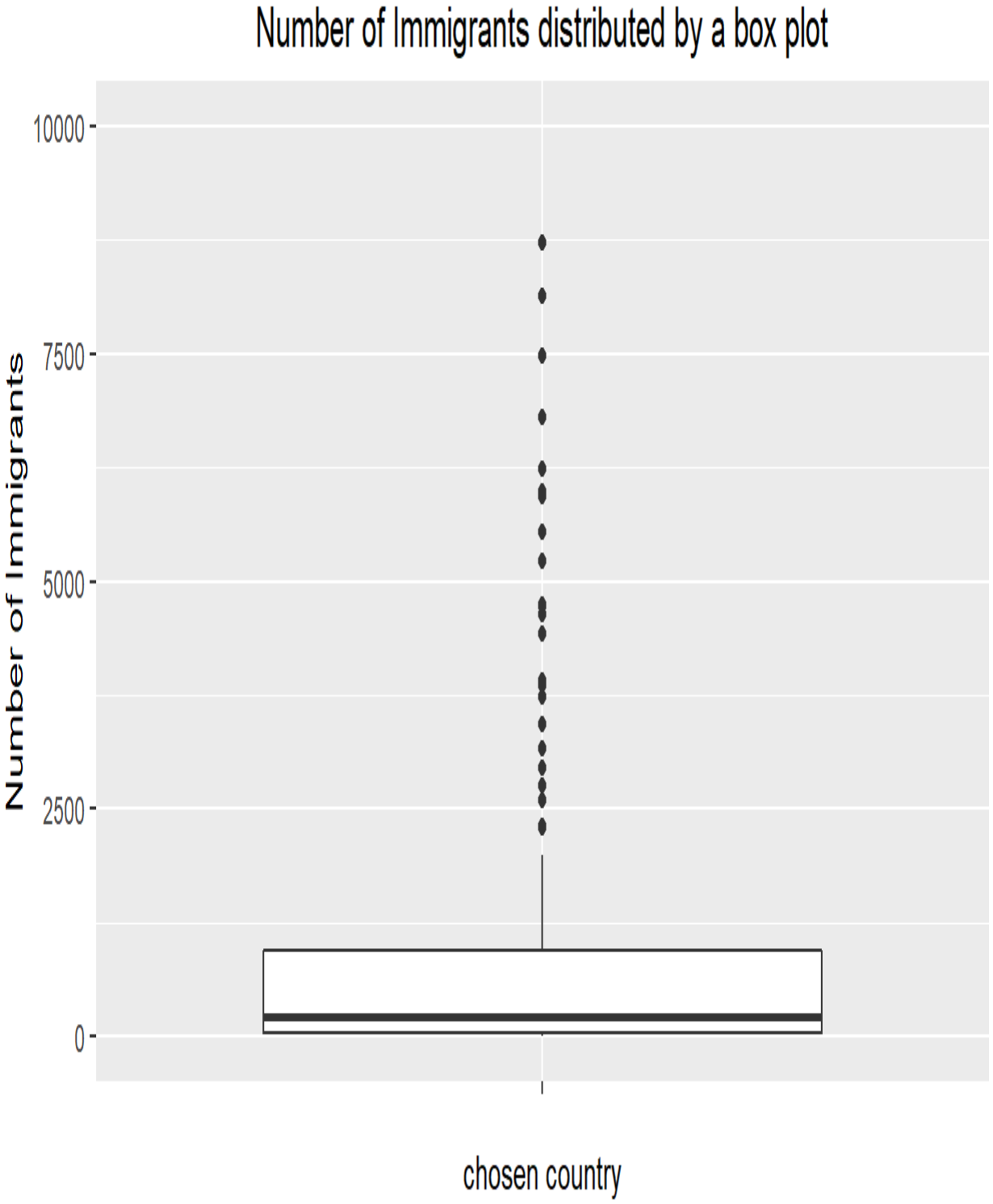
```
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="The number of immigrants",title="immigration to the chosen country
during 1980-2013")+
  scale_x_discrete(breaks = seq(1980, 2013, by = 2)) #A jump of 2 years
Q3Bplot
```



Visualized by box plot. AUTOMATED.

```
# Creating a sub data set
Q3DF <- DF[DF$Country == "Chad"] #This must be automated (User input)
Q3DF <- Q3DF[-nrow(Q3DF),]
Q3DF <- as.data.frame(Q3DF)

# Visualization
Q3Cplot_box <- ggplot(data = Q3DF, aes(x = "", y = Q3DF)) +
  geom_boxplot(fill="white") +
```

```
  coord_cartesian(ylim = c(0,10000)) + # I set the y axis scale so the plot looks
better.
  ggtitle("Number of Immigrants distributed by a box plot") +
    theme(plot.title = element_text(hjust = 0.5))+
  xlab("chosen country") + ylab("Number of Immigrants")
Q3Cplot_box
```

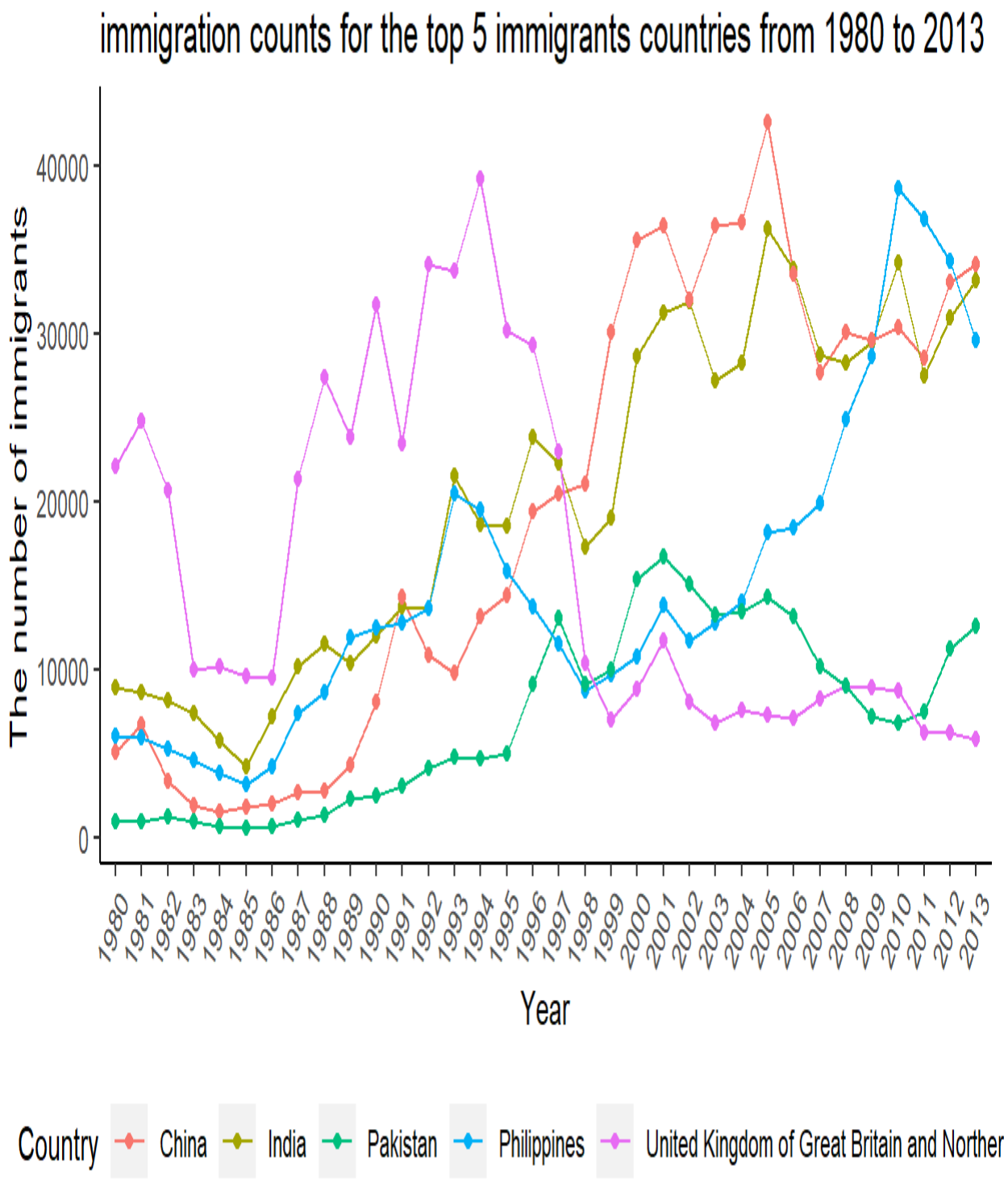


Q4 - What are the immigrants counts for the top 5 immigrants countries from 1980 to 2013.

Visualized by line charts. Displayed.

```
# Visualization
df_top <- df[1:195,] %>%
  arrange(desc(Total)) %>%
  select(1,4:37)
```

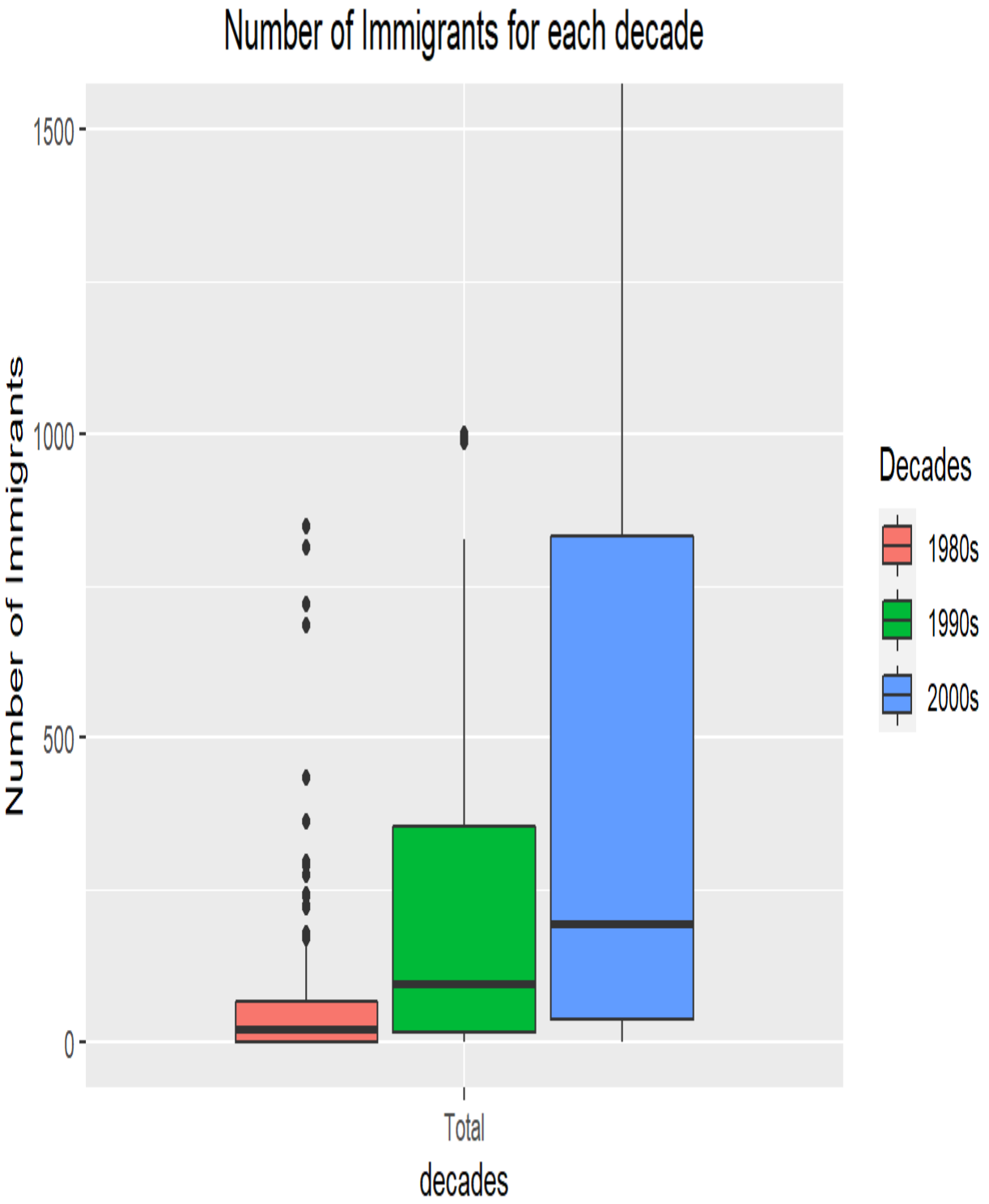
```
Q4plot <- df_top[1:5,] %>%
  gather(,key = "year", value = "number", 2:35) %>%
  ggplot(aes(x=year, y=number))+
  geom_line(aes(group=Country, color=Country))+
  geom_point(aes(color=Country))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="The number of immigrants",title="immigration counts for the top 5
immigrants countries from 1980 to 2013")
Q4plot
```



Q5 - What are the minimum, median, maximum, inter quartile range, and outlier values for immigrants to Canada from 1980 to 2013 per decade.

Visualized by box plot. Displayed.

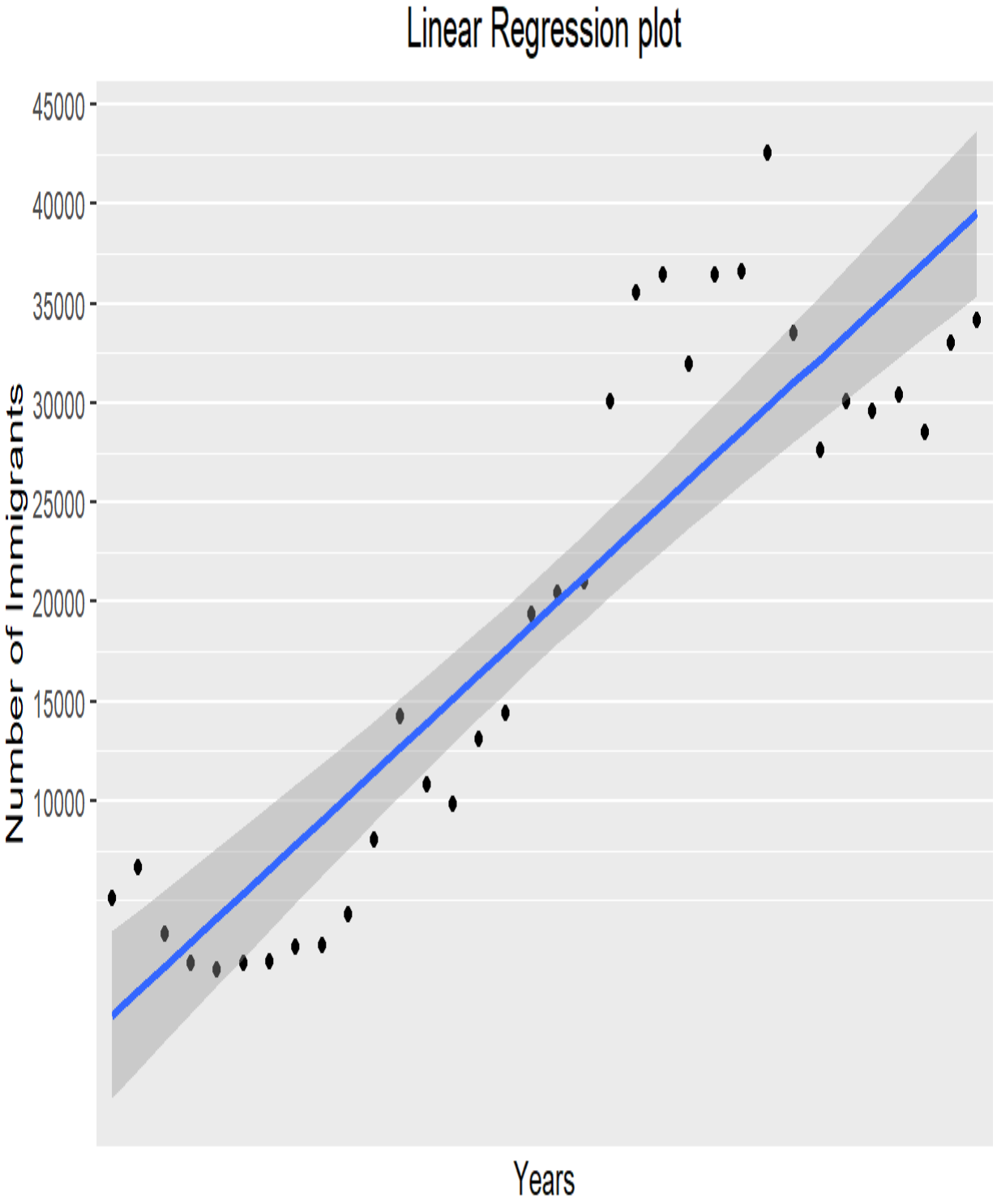
```
# creating empty lists to append the values for each year using for loop
total_decade80s = vector(mode = "list")
total_decade90s = vector(mode = "list")
total_decade2000s = vector(mode = "list")
Years20s <- for (country in DF$Country) {
  a <- select(filter(DF, Country == country), X2000:X2009 )
  total <- apply(a, 1, sum)
  total_decade2000s <- append(total_decade2000s, total)
}

Years90s <- for (country in DF$Country) {
  a <- select(filter(DF, Country == country), X1990:X1999 )
  total <- apply(a, 1, sum)
  total_decade90s <- append(total_decade90s, total)
}

Years80s <- for (country in DF$Country) {
  a <- select(filter(DF, Country == country), X1980:X1989 )
  total <- apply(a, 1, sum)
  total_decade80s <- append(total_decade80s, total)
}

 # making a data frame using the appended lists
total_decade <- as.data.frame(cbind(DF$Country,
total_decade80s,total_decade90s,total_decade2000s))
total_decade <- total_decade[-nrow(total_decade),] #deleting the total row
names(total_decade) <- c("Country", "1980s", "1990s", "2000s")
total_decade$`1980s` <- as.integer(total_decade$`1980s`) #changing the values of each
column to integer
total_decade$`1990s` <- as.integer(total_decade$`1990s`)
total_decade$`2000s` <- as.integer(total_decade$`2000s`)

#filtering the data to avoid outliers
total_decade <- filter(total_decade , `1980s`<1000 & `1990s`<1000 & `2000s`<2500 )
#reshaping the data to a from suitable for box plotting
total_decade <- gather(total_decade, Decades, value, 2:4)

#Visualization.
Q5plot_box <- ggplot(total_decade, aes(x = Decades, y = value, fill = Decades)) +
  geom_boxplot(aes(x=country, y=value)) +
  coord_cartesian(ylim = c(0,1500)) +
  ggtitle("Number of Immigrants for each decade") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("decades") + ylab("Number of Immigrants")
Q5plot_box
```



Exploratory Analysis Questions

```
Q6 - Which countries in the future will have more immigrants,
and which will have less?

Visualized by Scatter & regression plots to observe the immigration patterns.
Automated.

Q6DF <- DF[DF$Country=="China",5:38] #Here "China" is the user input, the purpose is
to automate this option
Q6DF <- gather(Q6DF,Year,Total,1:34) #Transposing the dataset
Q6DF$Year<-gsub("X","",as.character(Q6DF$Year)) #Removing X from years
Q6DF$Year <- as.integer(Q6DF$Year)


Q6Regplot <- ggplot(data = Q6DF, aes(x= Year, y=Total)) +
```

```
geom_point() +
  geom_smooth(method="lm") +
  scale_x_discrete(breaks = seq(1980, 2013, by = 2)) + #A jump of 2 years
  scale_y_continuous(breaks = seq(10000, 50000, by = 5000)) +
  ggtitle("Linear Regression plot") +
  theme(plot.title = element_text(hjust = 0.5))+
  xlab("Years") + ylab("Number of Immigrants")
Q6Regplot
```



Q7 - Which countries would have a dominant immigrants' population?

Visualized by word cloud. Displayed

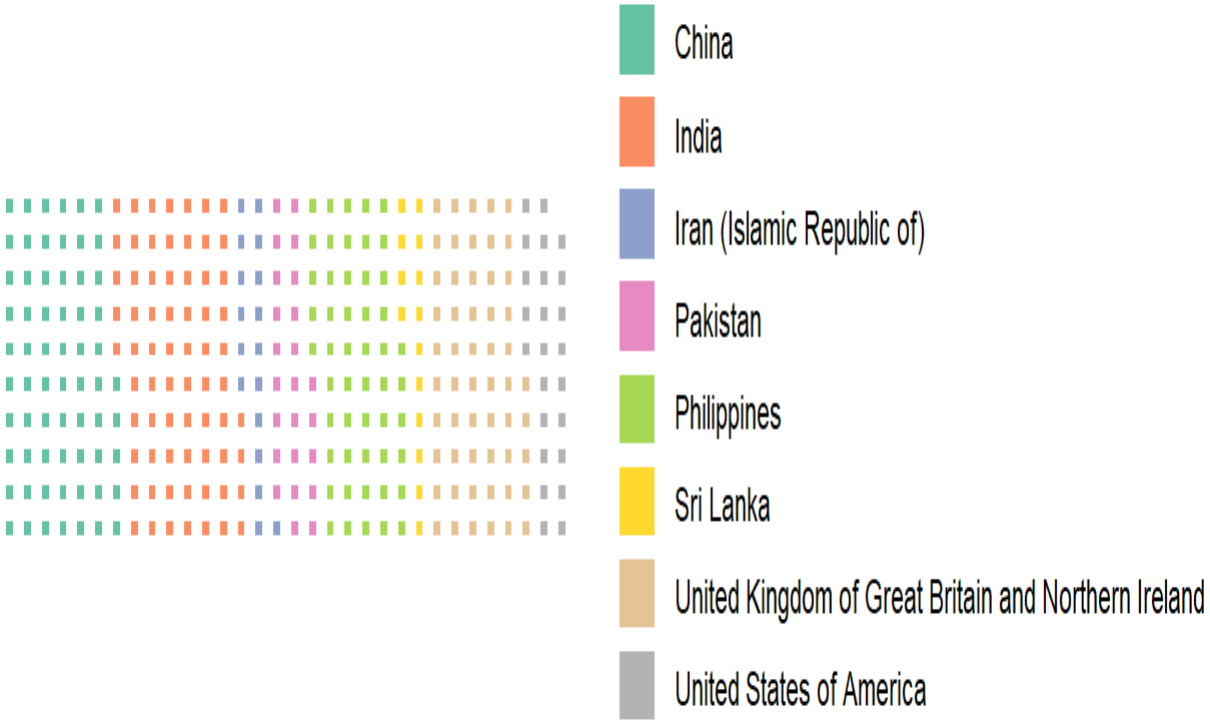
```
Q7DFA <- DF[c("Country", "Total")] #Choosing these two columns only
Q7DFA <- Q7DFA[-c(nrow(Q7DFA), nrow(Q7DFA)-1),] # removing the last two rows
["Unknown" "Total"]
```

```
Q7Word_plot <- wordcloud(words = Q7DFA$Country, freq = Q7DFA$Total, min.freq = 1,
          max.words=2000, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

Visualized by waffle plot. Displayed

```
Q7DFB <-arrange(Q7DFA, desc(Total))
Q7DFB <- Q7DFB[1:8,] #Choosing the first 8 rows >> top 8
Q7DFB$Total <- Q7DFB$Total%/%10000 # scaling down the values by a factor of 10000 to
be plottable
# used split() to make a list that contains countries names and their values which is
the total number of immigrants
waffle_list <- split(Q7DFB$Total, Q7DFB$Country)
waffle_list <- unlist(waffle_list) # unlist the data because waffle() doesn't support
lists
```

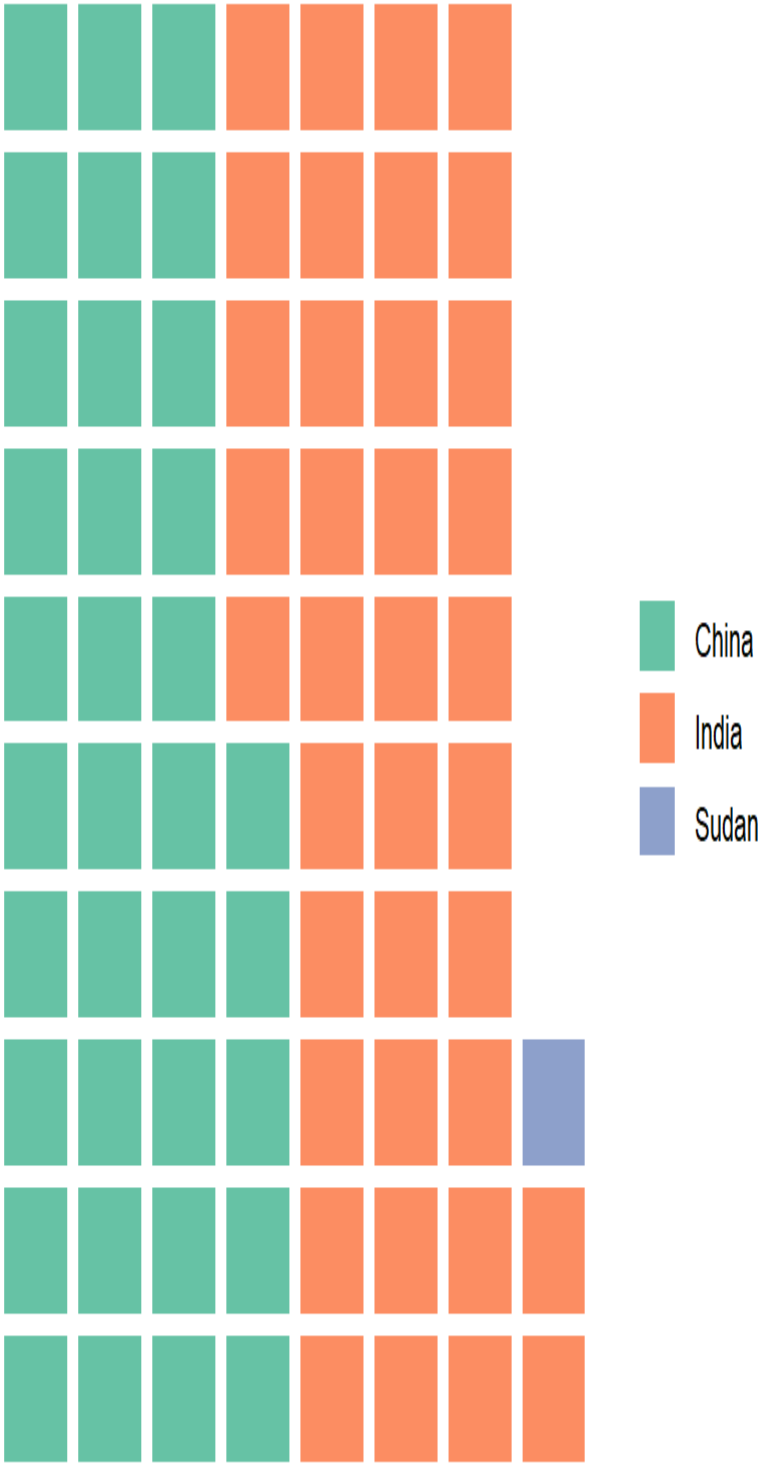
```
waffle(waffle_list)
```



Visualized by waffle plot. Automated

```
#Automated visualization
Q7DFC <- filter(Q7DFA, Total >100)
Q7DFC <- filter(Q7DFC, Country %in% c("India" ,"China", "Sudan")) # the chosen
countries ## To be automated
Q7DFC$Total <- Q7DFC$Total%/%min(Q7DFC$Total)

#used split() to make a lsit that contains countries names and their values which is
the total number of immigrants
waffle_list2 <- split(Q7DFC$Total, Q7DFC$Country )
waffle_list2 <- unlist(waffle_list2) # unlist the data because waffle() doesn't
support lists
Q7waffle <- waffle(waffle_list2)
```

Q7waffle



Q8 - What are the immigration trends of the continents from 1980 to 2013?

Visualized by line chart. Displayed.

```
# Visualized by line chart
df[196:201,] %>%
  select(2,4:37) %>%
  gather(,key = "year", value = "number", 2:35) %>%
  ggplot(aes(x=year, y=number))+
  geom_line(aes(group=Continent, color=Continent))+
  geom_point(aes(color=Continent))+
  theme(panel.background = element_blank(),
        axis.line = element_line(),
        legend.position = "bottom",
```

```
        axis.text.x = element_text(angle=60, hjust=1))+
  labs(x="Year",y="The number of immigrants",title="Immigration trends of the
continents from 1980 to 2013")
```