Visualisation in R Part 3

Lab 8

In this lab you will use ggplot and sqldf to analyse the results of the Winter Olympics over multiple years.

1. Plotting maps and geographic data in ggplot

In order to plot geographical data using R, we will use the package ‘*maps*’ and ‘*ggplot*’. Other packages that provide similar and extended mapping functionality are: ggmap, sp, rgdal, rgeos, RgoogleMaps, maptools, Rjsonio and OpenStreetMap amongst others.

The package ‘*maps*’ contains basic geographical information included in databases for the world (countries and cities), USA (states, counties and cities), Canada, France, Italy, New Zeland and lakes.

```
library(reshape2)

library(maps)

crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests) crimesm <- melt(crimes,
id = 1)
```

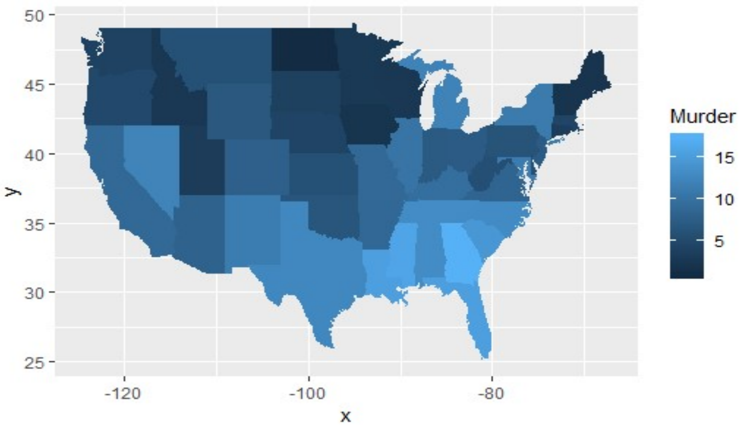
Original Table

state	Murder	Assault	UrbanPop	Rape
alabama	13.2	236	58	21.2
alaska	10.0	263	48	44.5
arizona	8.1	294	80	31.0
arkansas	8.8	190	50	19.5

state	variable	value
alabama	Murder	13.2
alaska	Murder	10.0
arizona	Murder	8.1
arkansas	Murder	8.8

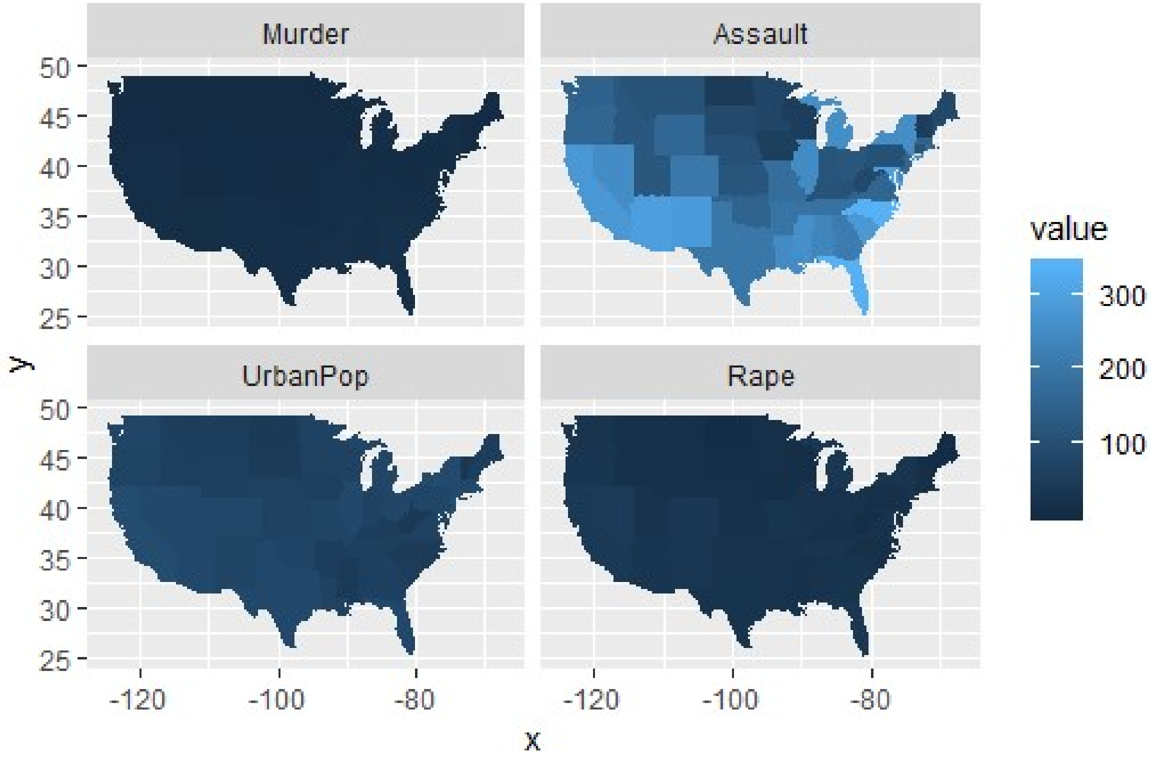
```
states_map <- map_data("state")

ggplot(crimes, aes(map_id = state)) +  geom_map(aes(fill = Murder), map = states_map) +
expand_limits(x = states_map$long, y = states_map$lat)
```



Creating small multiples with maps

```
ggplot(crimesm, aes(map_id = state)) +    geom_map(aes(fill = value), map = states_map)
+ expand_limits(x = states_map$long, y = states_map$lat) +    facet_wrap( ~ variable)
```



2. The Assignment Dataset

The data (OlympicGames.csv) is available from Brightspace.

3. Import the data

```
olympicdata<-  read.csv( 'C:\\.your location..\\OlympicGames.csv'  ,  sep= ','  , header=T  )
```

4. Data manipulation

4.1 Calculating the number of gold, bronze and silver medals per country

We can run sql queries using sqldf to find the number of gold medals per country :

```
resultsmedalsgold<-sqldf("select country, count(Medal) as gold from olympicdata where
Medal=='gold' group by country")
```

4.2 Repeating the process for silver and bronze:

```
resultsmedalssilver<-sqldf("select country, count(Medal) as silver from olympicdata where
Medal=='silver' group by country") resultsmedalsbronze<-sqldf("select country, count(Medal) as
bronze from olympicdata where Medal=='bronze' group by country")
```

Now we have three dataframes with information about the medals won by the different countries. If we want to combine all the information about medals we can create a new data frame merging the three dataframes:

```
resultsmedals<-merge(resultsmedalsgold,resultsmedalssilver,by="Country",all=TRUE)
resultsmedals2<-merge(resultsmedals,resultsmedalsbronze,by="Country",all=TRUE)
```

As a result of the merge, some cells will contain NA values for those countries without any medal of that particular type (all=TRUE performs the full outer join keeping all rows from both data frames). We want to replace those values with 0s.

#Option 1

```
resultsmedals2[is.na(resultsmedals2)] <- 0
```

#Option 2

```
resultsmedals2$gold <- ifelse(is.na(resultsmedals2$gold), 0, resultsmedals2$gold)
resultsmedals2$silver <- ifelse(is.na(resultsmedals2$silver), 0, resultsmedals2$silver)
resultsmedals2$bronze <- ifelse(is.na(resultsmedals2$bronze), 0, resultsmedals2$bronze)
```

5. Visualisation Exercises

Please answer the following questions. Use a visualisation to support the answer for each question. Experiment with the different types of charts and options ggplot offers. Please place the visualisations underneath each question.

1) What country has won most Silver medals since 2000? (2 marks)
2) How is the gender balance amongst the United States gold medalist? (2 marks)
3) What are the best sports for Sweden, USA, Austria and Switzerland? (2 marks)
4) Who has the most total medals? (2 marks)
5) What is the variation and spread of ages amongst gold and silver medalists? (2 marks)