Report of Homework 1

Yin Chenqiao, student number: 1500015533

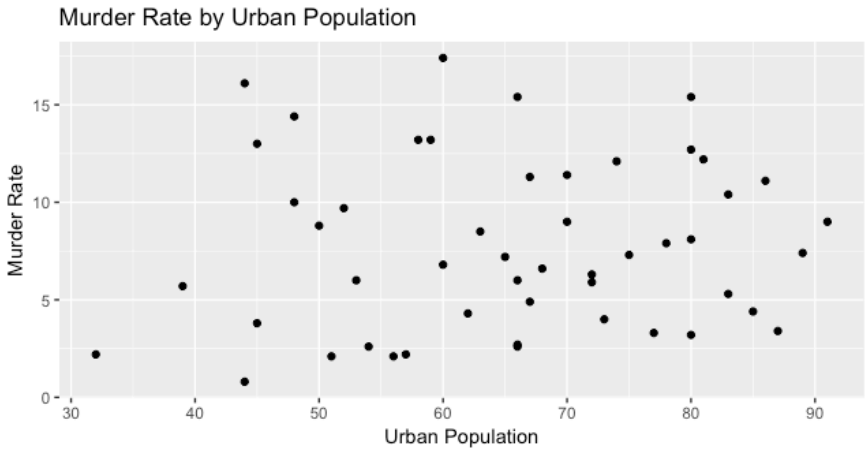
Problem #1

1(a)

First, we should import the data `USArrests` and have a brief view of the data (L10-11). The output is:

```
           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
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

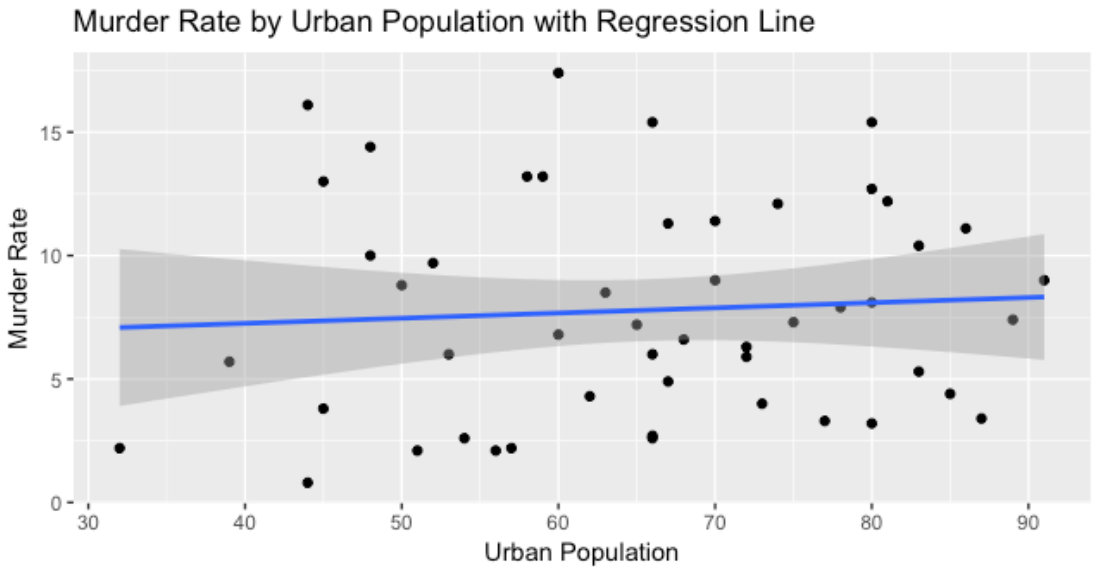
Then use the function `ggplot` to plot the data, and the function `geom_point` to draw the scatterplot (L12-14). And below is the plot:



We have used these two functions many times, so it's easy to finish and I don't need to explain more about the code.

1(b)

We need to add a regression overlay to the plot. It's also easy by using function `geom_smmoth`. As a regression line shows the relationship better, I choose to draw a regression line by using the parameter `method = 'lm'` (L17-19).
Below is the plot:

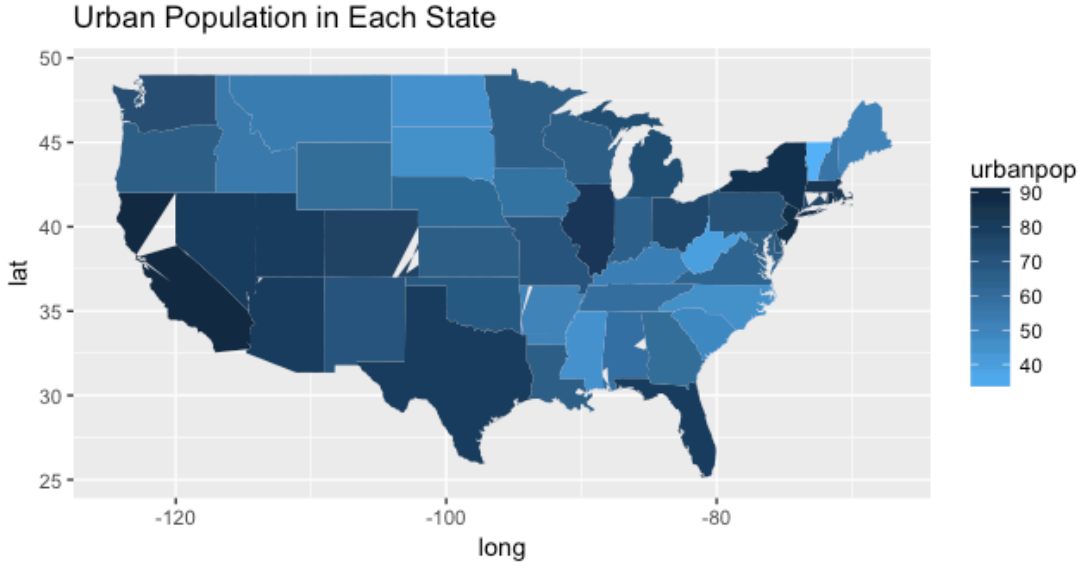


1(c)

From the regression line we can see that there is a slightly positive relationship between urban population and murder rate. That is, a bigger state may have higher murder rate. Maybe a big state has high cost of security, so criminals have lower cost to murder.

1(d)

We have written similar codes on the class. It may take a little time, but it's not hard to finish the code by imitation. Like the final map we drew on the class, I changed the color to make it not strange to read (L23-32). Below is the map:



1(e)

The code here may be hard to read because it has been simplified (L35-40). But first I will show the result of the code. The output is:

```
[1] "california"
```

So California has the largest urban population.

Then I'll try to explain my code. We need to find the state which has the largest population, so first we should find the largest population. We can use the code below:

```
largest.pop = max(choro.1$urbanpop, na.rm = T)
```

Now we have the largest population (actually, the output is 91). Then we should find which state has that population by using the code below:

```
largest.state.id = which(choro.1$urbanpop == largest.pop)
```

Because the return value of `which` function is the row number of the data, now we still don't have the name of the state. But we are close to it. To get the name of the state, we need to use the `region` field in the dataset `choro.1`.

```
largest.state = choro.1$region[largest.state.id]
```

If we run the code above, we can get the output below:

```
  [1] "california" "california" "california" "california"
"california" "california"
  [7] "california" "california" "california" "california"
"california" "california"
```

…… (omitted some output)

```
[505] "california" "california" "california" "california"
"california" "california"
[511] "california" "california" "california" "california"
"california" "california"
```

This is because the dataset has many pieces of data whose value of `field` is `california`. To get only one result, we should use the `unique` function:

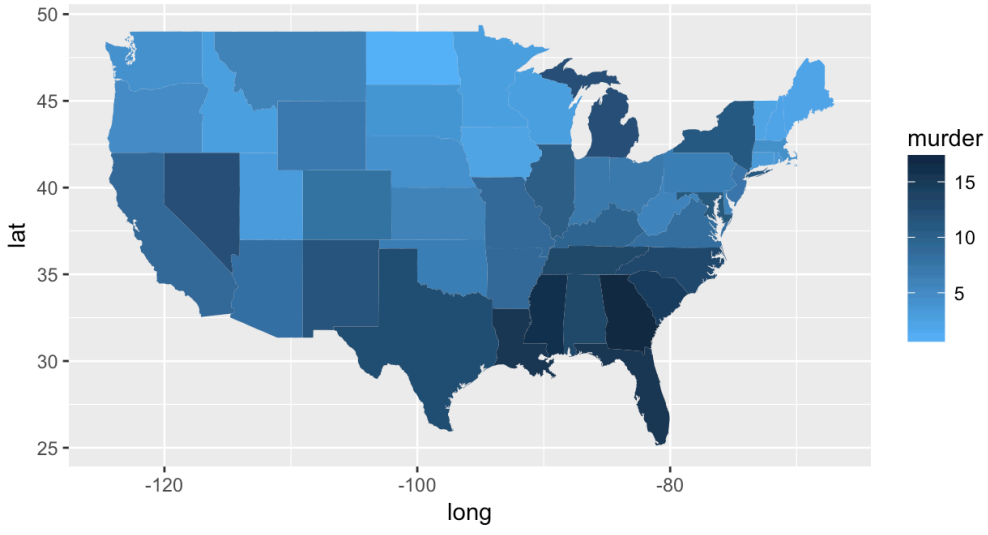
```
largest.state = unique(choro.1$region[largest.state.id])
```

Now the output is:

```
[1] "california"
```

See we get a unique answer. If we simplify the code above, we can get the code in the source file. Actually they are the same.

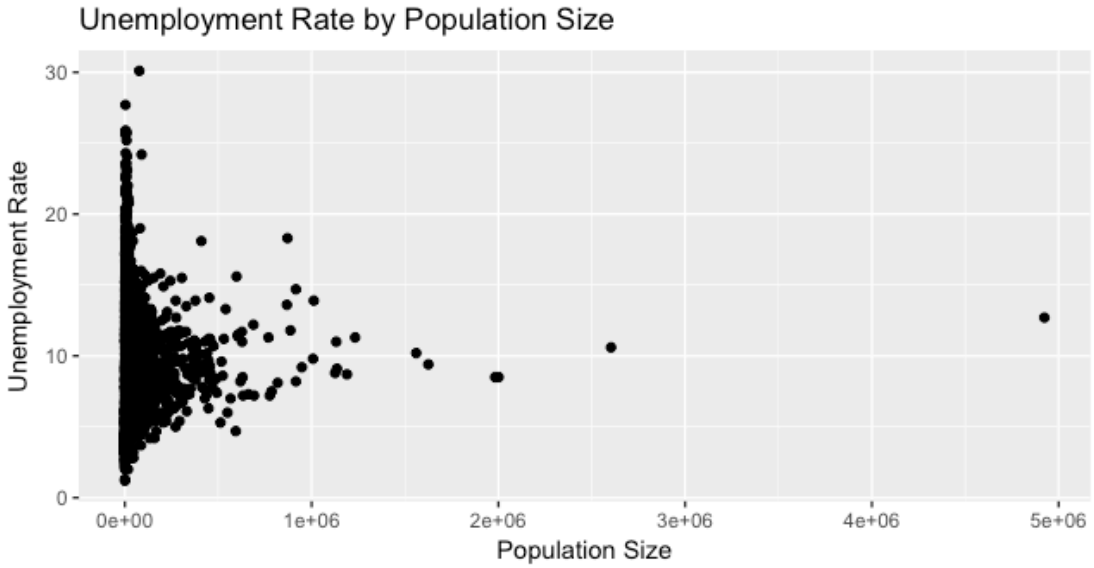
Now take a look at the map on murder rates. We can easily find that the dark blue regions and the light blue regions in the two maps are almost the same. It stands for what we find in the regression line, which is, a bigger state has a higher murder rate.



Problem #2

2(a)

We have already been familiar to the function `ggplot` and `geom_point`, so it's easy to draw the plot below:



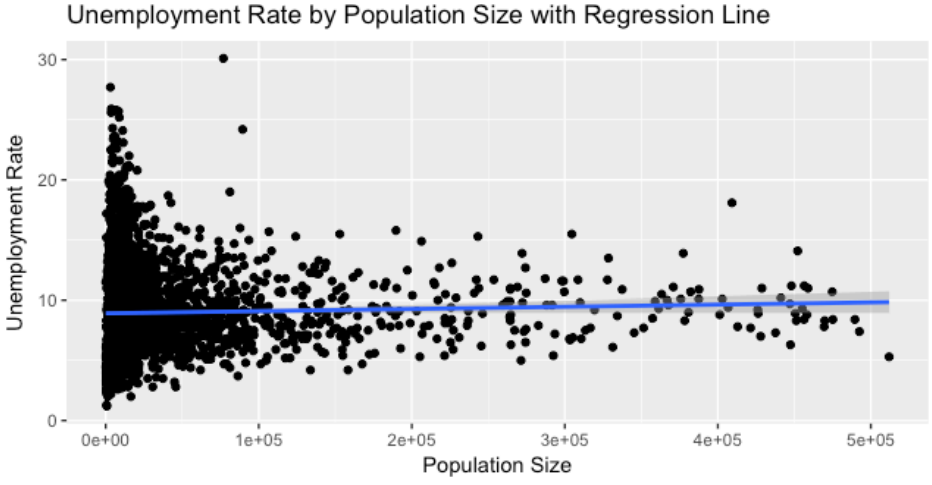
2(b)

Now we need to add a regression line. From the scatter plot we can see there are a few points with extreme values. Before we draw the regression line, we should clear the data first. I choose to delete the data out of 3 times of standard deviation (L58-59). I want to simplify the code, but it's strange that there is an error when I run the code below:

```
unemp.sub = subset(unemp, subset = pop < (mean(unemp$pop)
+ 3*sd(unemp$pop)))
```

Anyway, it's still fine to run the original code.

After clearing the data, we can draw the regression line using the `geom_smooth` function, which we have also been familiar to. Below is the plot:

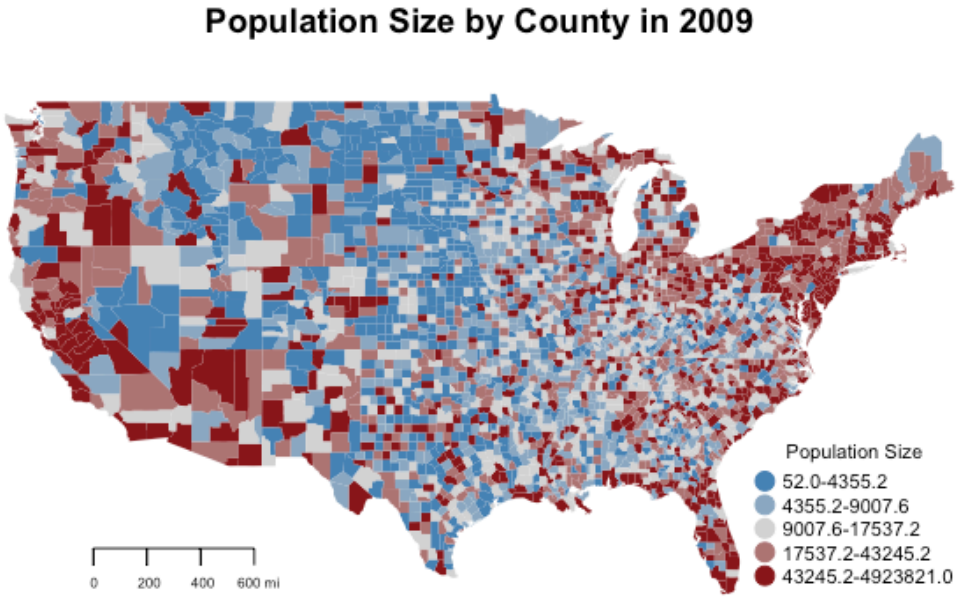


2(c)

There is a so slightly positive relationship that I don't think it's statistically significant. So I think there is no relationship between population size and unemployment rate.

2(d)

We have written similar code on the class, so it isn't hard to write the code. Below is the map.



2(e)

I use similar code as in 1(e). The output is:

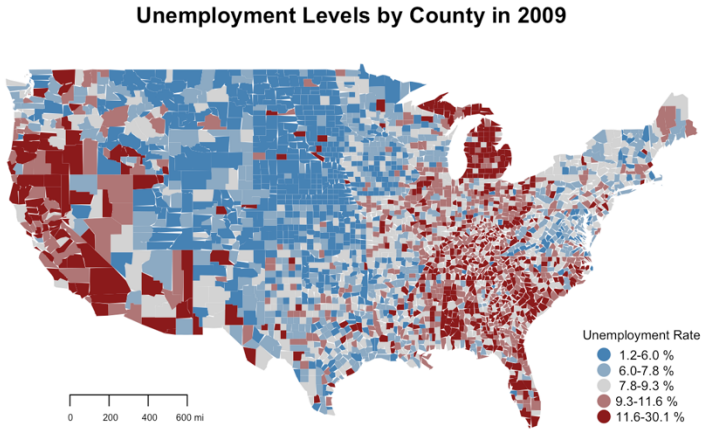
```
[1] "california,los angeles"
```

Because these lines of code are more complex, I'll explain them again. We need to find the county with the largest population. First, we need to find the row number of the county with the largest population by using the code below:

```
largest.county.id = which(unemp$pop == max(unemp$pop,
na.rm = T))
```

If we want to get the name of the county, we should use the field `polyname` in the dataset `county.fips`. To create an relationship between the dataset `unemp` and `county.fips`, note that the two database have a same field `fips`, which maybe a kind of id of the county. To get the fip of the county with the largest population, use the code below:

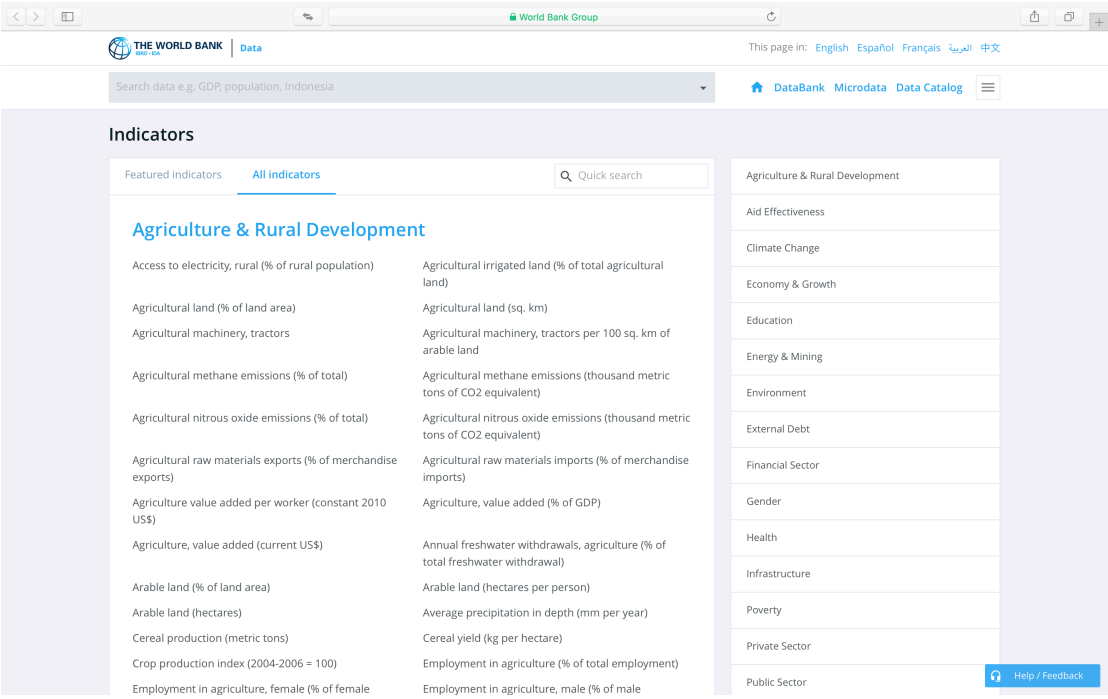
```
largest.county.fip = unemp$fips[largest.county.id]
```
Then we need to find that fip in the dataset `county.fips`. Again we should use `which` function:
```
largest.county.id2 = which(county.fips$fips ==
largest.county.fip)
```
Now we can get the name of that county:
```
largest.county = county.fips$polyname[largest.county.id2]
```
Again, if we simplify these lines of code, we can get the code in the source file.
Now take a look at the map on unemployment rates. We can find that the red regions and the blue regions in the two maps are very similar. So if we just see the map, maybe we can conclude that there is a positive relationship between population size and unemployment level.



Problem #3

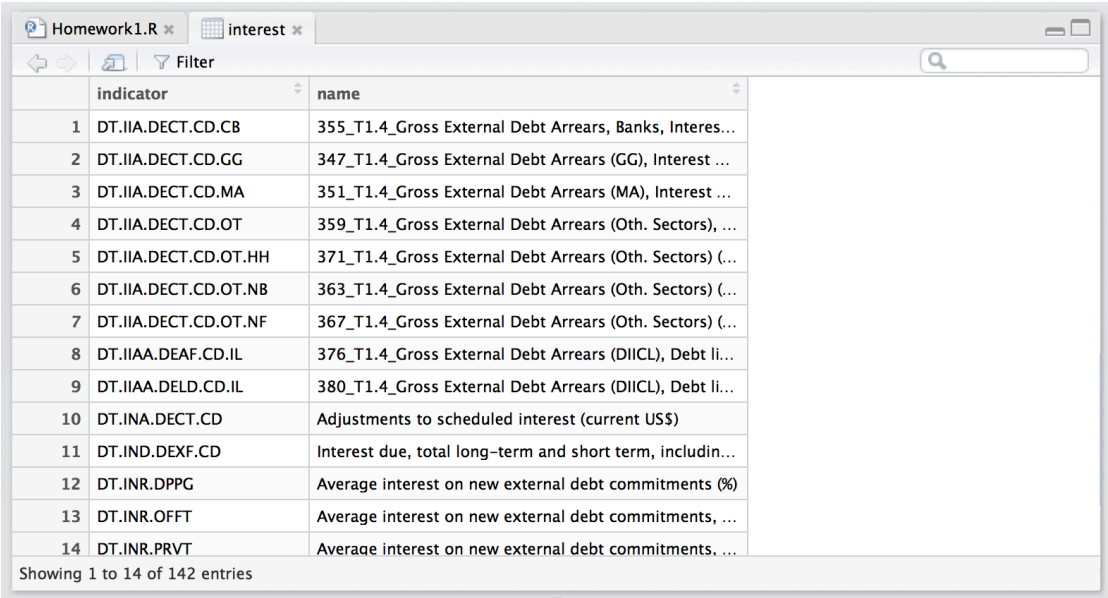
3(a)

Below is World Bank WDI website.



3(b)

I use `WDIsearch` function to search. There are 142 indicators about interest, below are some of them. I choose the indicator `FR.INR.DPST`, which means the deposit interest rate (%).



	indicator	name
1	DT.IIA.DECT.CD.CB	355_T1.4_Gross External Debt Arrears, Banks, Interes...
2	DT.IIA.DECT.CD.GG	347_T1.4_Gross External Debt Arrears (GG), Interest ...
3	DT.IIA.DECT.CD.MA	351_T1.4_Gross External Debt Arrears (MA), Interest ...
4	DT.IIA.DECT.CD.OT	359_T1.4_Gross External Debt Arrears (Oth. Sectors), ...
5	DT.IIA.DECT.CD.OT.HH	371_T1.4_Gross External Debt Arrears (Oth. Sectors) (...
6	DT.IIA.DECT.CD.OT.NB	363_T1.4_Gross External Debt Arrears (Oth. Sectors) (...
7	DT.IIA.DECT.CD.OT.NF	367_T1.4_Gross External Debt Arrears (Oth. Sectors) (...
8	DT.IIAA.DEAF.CD.IL	376_T1.4_Gross External Debt Arrears (DIICL), Debt li...
9	DT.IIAA.DELD.CD.IL	380_T1.4_Gross External Debt Arrears (DIICL), Debt li...
10	DT.INA.DECT.CD	Adjustments to scheduled interest (current US$)
11	DT.IND.DEXF.CD	Interest due, total long-term and short term, includin...
12	DT.INR.DPPG	Average interest on new external debt commitments (%)
13	DT.INR.OFFT	Average interest on new external debt commitments, ...
14	DT.INR.PRVT	Average interest on new external debt commitments, ...

Showing 1 to 14 of 142 entries

3(c)

We need to choose another 1 or 2 countries, so it's necessary to know the short of the countries. First, use the field `country` in the dataset `WDI_data` to browse the country names and the short of country names (L106-107):

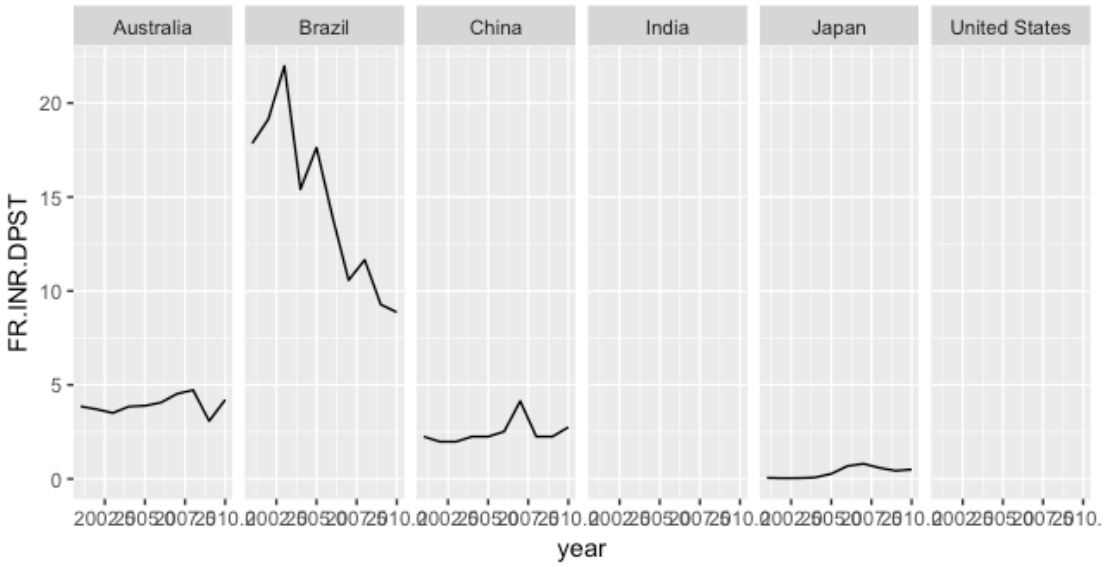
```
country = WDI_data$country  #Browse the country names
View(country)
```

I choose Australia and Brazil to analyze their data. It's easy to download the data by imitating the code on the class.

3(d)

Now we need to plot line graphs for each country. Besides using `ggplot` and `geom_line` function, we also need to use `facet_grid` function. It's strange that the data of India and US is missing, but I have double checked it by viewing the website and find that WDI does not have the data.

From the plot we can see that the deposit interest rate vary among countries. The deposit interest rate in Brazil is dramatically high, and the deposit interest rate in Japan is relatively low (close to 0). The deposit interest rate in Australia and China is stable and appropriate.



3(e)

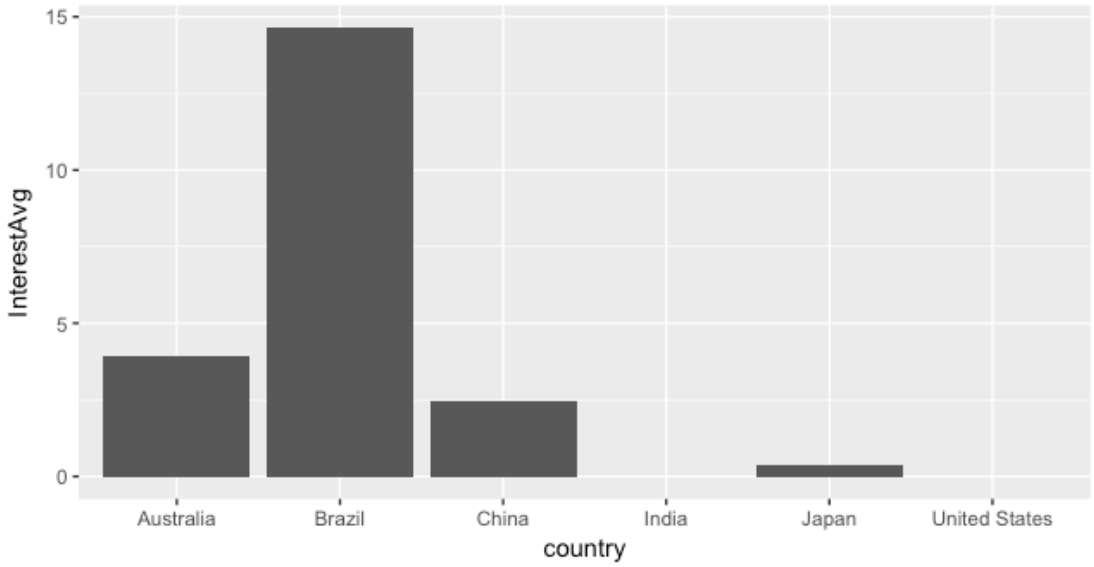
To prevent potential error when using the data with missing values, first I let all the missing values be zero:

```
data.interest$FR.INR.DPST[which(is.na(data.interest$FR.IN
R.DPST) == 1)] = 0
```

Then use `ddply` function to get the average interest of each country:

```
AggrData = ddply(data.interest, .(country), summarize,
                 InterestAvg = mean(FR.INR.DPST, na.rm = T))
```

It's easy to plot by using `ggplot` and `geom_bar` function, below is the plot:



We can see the interest rate in Brazil is very high, and the interest rate in Japan is close to 0. The interest rate in Australia and China falls in the middle.

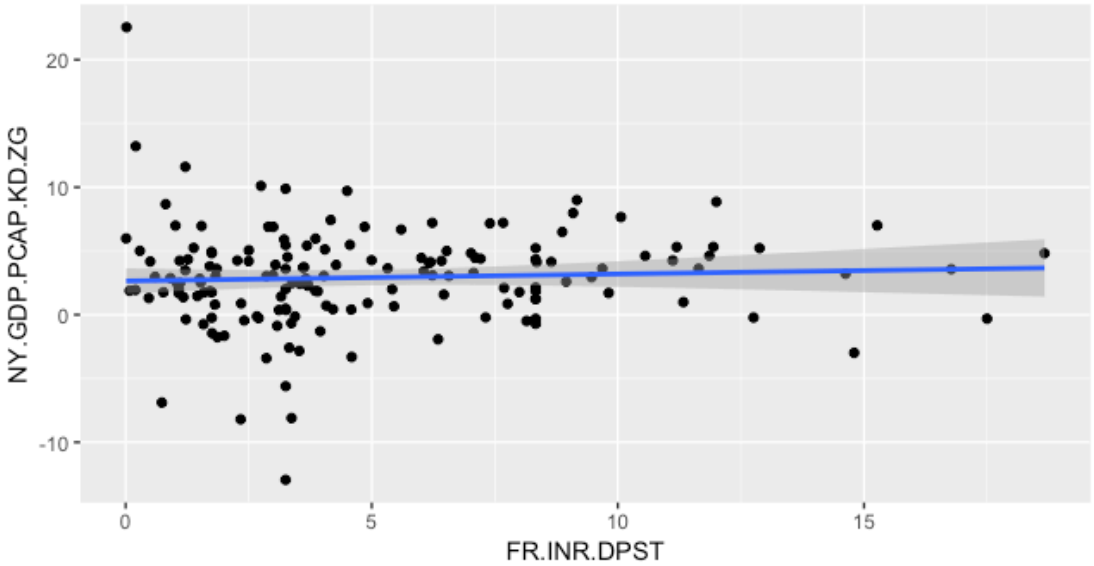
Problem #4

4(a)

We need to download the data `NY.GDP.PCAP.KD.ZG` and `FR.INR.DPST` of all available countries in 2010. It's easy by using `WDI` function and change some of the parameters.

4(b)

Now plot again. I don't think it's necessary to explain the code. Below is the plot:



We can see that there isn't any relationship between interest rate and GDP growth.

4(c)

To finish the problem, we need to imitate the code on the class. We don't know the min and the max of deposit interest rate, so to make each country colored, we need to use `min` and `max` function. Below is the map. Because there are some missing values, some countries remain white.