

Assignment3

Research Question

Twenty years since the end of Apartheid: Did the collapse of Apartheid play a significant role in reducing racial and social inequality in South Africa? Is post-apartheid South Africa better off or worse off than during the apartheid era?

Definition of Racial and social inequality

Before conducting the data analysis to find the answer for our research question, we begin by clarifying the definitions for racial and social inequality. While social inequality broadly refers to the existence of unequal opportunities for different social status/positions within a society, racial inequality can be seen as one of dimensions of social inequality. It thus indicates the discrimination based on race in access to socioeconomic opportunities or services. In our research study, we will specifically look into racial discrimination in terms of employment, education, and income levels. Because these three indicators within the capitalism society can be seen as fundamental yet significant estimators for the quality of human well-being, we decided to include them. In addition, we will try to identify drivers of unequal income distribution by controlling possible factors and variables such as unemployment rate and education level.

Literature review

Before starting investigation, we need to look around previous researches which has been written by various ambitious researchers.

According to Leibbrandt, (see Leibbrandt (n.d.)) 1, Since the fall of Apartheid (1993~2008), overall (include all races) income inequality increased. The same is true among four major racial groups. 2, However, the major driver of inequality increase is intra-African inequality in South-Africa.

The reason why we choose this article as the first reference article for this article is that this is the most cited work in the South-African Inequality Study.

Data Gathering by web-scraping

According to Leibbrandt, inequality has been increased since the fall of Apartheid. We will test this assumption by using other data which is not used in the article.

We found the data of monthly earnings among races and gender. We tried to scraping the data from the website.

```
URL <- 'http://businesstech.co.za/news/wealth/131524/this-is-the-average-salary-in-south-africa-by-race'

RaceEarningsTable <- URL %>% read_html() %>%
  html_nodes('#container > div.content_holder > div.content > div.post_single > div.post_content') %>%
  html_table() %>%
  as.data.frame()

RaceEarningsTable
```

	X1	X2	X3	X4	X5	X6	X7
## 1		Median	Median	Median	Mean	Mean	Mean
## 2	Race	2003	2012	Increase	2003	2012	Increase
## 3	White	14 468	16 581	15%	11 249	11 991	7%

```
## 4 Asian/Indian 7 825 11 701      50% 5 264 8 993      60%
## 5 Coloured 4 241 7 058      66% 2 437 3 897      60%
## 6 Black African 4 059 5 445      34% 2 437 2 998      23%
```

```
URL <- 'http://businesstech.co.za/news/wealth/131524/this-is-the-average-salary-in-south-africa-by-race'
```

```
GenderEarningsTable <- URL %>% read_html() %>%
  html_nodes('#container > div.content_holder > div.content > div.post_single > div.p
  html_table() %>%
  as.data.frame
```

```
GenderEarningsTable
```

```
##      X1      X2      X3      X4      X5      X6      X7
## 1      Median Median      Median Mean Mean      Mean
## 2 Race 2003 2012 Increase 2003 2012 Increase
## 3 Male 5 963 8 299      39% 3 375 4 317      28%
## 4 Female 4 849 6 399      32% 2 435 3 118      28%
```

Data Cleaning and Merging

In this section, we will try to clean the data so that they can be statistical analysed.

Firstly, we use command “summary” to investigate the structure (class of variables, number of vectors) of data frames we got in the previous section.

```
summary(RaceEarningsTable)
```

```
##      X1      X2      X3
## Length:6      Length:6      Length:6
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##      X4      X5      X6
## Length:6      Length:6      Length:6
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##      X7
## Length:6
## Class :character
## Mode :character
```

```
summary(GenderEarningsTable)
```

```
##      X1      X2      X3
## Length:4      Length:4      Length:4
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##      X4      X5      X6
## Length:4      Length:4      Length:4
## Class :character Class :character Class :character
## Mode :character Mode :character Mode :character
##      X7
## Length:4
## Class :character
## Mode :character
```

As shown, every variables has a class of “characters” even though it represents numerical data.

The data we want to have is the mean of earnings among races and gender in 2003, 2012.

Firstly, we make TimeVector and IndividualVector to labeling the data.

```
TimeVector <- c(2003,2012) #numerical vector
IndividualVector <- c("Male","Female","White","Asian/Indian","Coloured","BlackAfrican") #character vector
```

Then, we try to convert character vector to numerical vector.

```
male2003 <- as.numeric(gsub("[0-9]+.*$", "\\1", str_replace_all(GenderEarningsTable$X5[3], fixed(" ")
is.numeric(male2003)
```

```
## [1] TRUE
```

```
male2003
```

```
## [1] 3375
```

As I shown above, the character variable successfully converted to numerical variable. Then, we make function which conduct this sequence.

```
Converter <- function(x){
y <- as.numeric(gsub("[0-9]+.*$", "\\1", str_replace_all(x, fixed(" "), "")))
return(y)
}
test <- Converter(x = GenderEarningsTable$X5[3])
is.numeric(test)
```

```
## [1] TRUE
```

```
test
```

```
## [1] 3375
```

Then, we can apply this function to all data.

```
#definition of vector
Earnings2003 <- c(0,0,0,0,0,0)
Earnings2012 <- c(0,0,0,0,0,0)

#GenderEarnings
for(i in 3:4){
  Earnings2003[i-2] = Converter(x = GenderEarningsTable$X5[i])
  Earnings2012[i-2] = Converter(x = GenderEarningsTable$X6[i])
}
#RaceEarnings
for(i in 3:6){
  Earnings2003[i] = Converter(x = RaceEarningsTable$X5[i])
  Earnings2012[i] = Converter(x = RaceEarningsTable$X6[i])
}
Earnings2003
```

```
## [1] 3375 2435 11249 5264 2437 2437
```

```
Earnings2012
```

```
## [1] 4317 3118 11991 8993 3897 2998
```

```
preEarnings <- data.frame(IndividualVector,Earnings2003, Earnings2012)
preEarnings
```

```
## IndividualVector Earnings2003 Earnings2012
```

```
## 1      Male      3375      4317
## 2      Female    2435      3118
## 3      White    11249     11991
## 4  Asian/Indian    5264      8993
## 5    Coloured    2437      3897
## 6  BlackAfrican    2437      2998
```

The preEarnings is messy data.

So we are going to transform it into tidy data.

```
library(tidyr)
Earnings <- gather(preEarnings, time, mean, Earnings2003:Earnings2012)
Earnings
```

```
##   IndividualVector      time  mean
## 1      Male Earnings2003  3375
## 2      Female Earnings2003  2435
## 3      White Earnings2003 11249
## 4  Asian/Indian Earnings2003  5264
## 5    Coloured Earnings2003  2437
## 6  BlackAfrican Earnings2003  2437
## 7      Male Earnings2012  4317
## 8      Female Earnings2012  3118
## 9      White Earnings2012 11991
## 10  Asian/Indian Earnings2012  8993
## 11    Coloured Earnings2012  3897
## 12  BlackAfrican Earnings2012  2998
```

We succeeded to make the numerical vector showing the earnings among races and genders.

Data Gathering by using Data-API

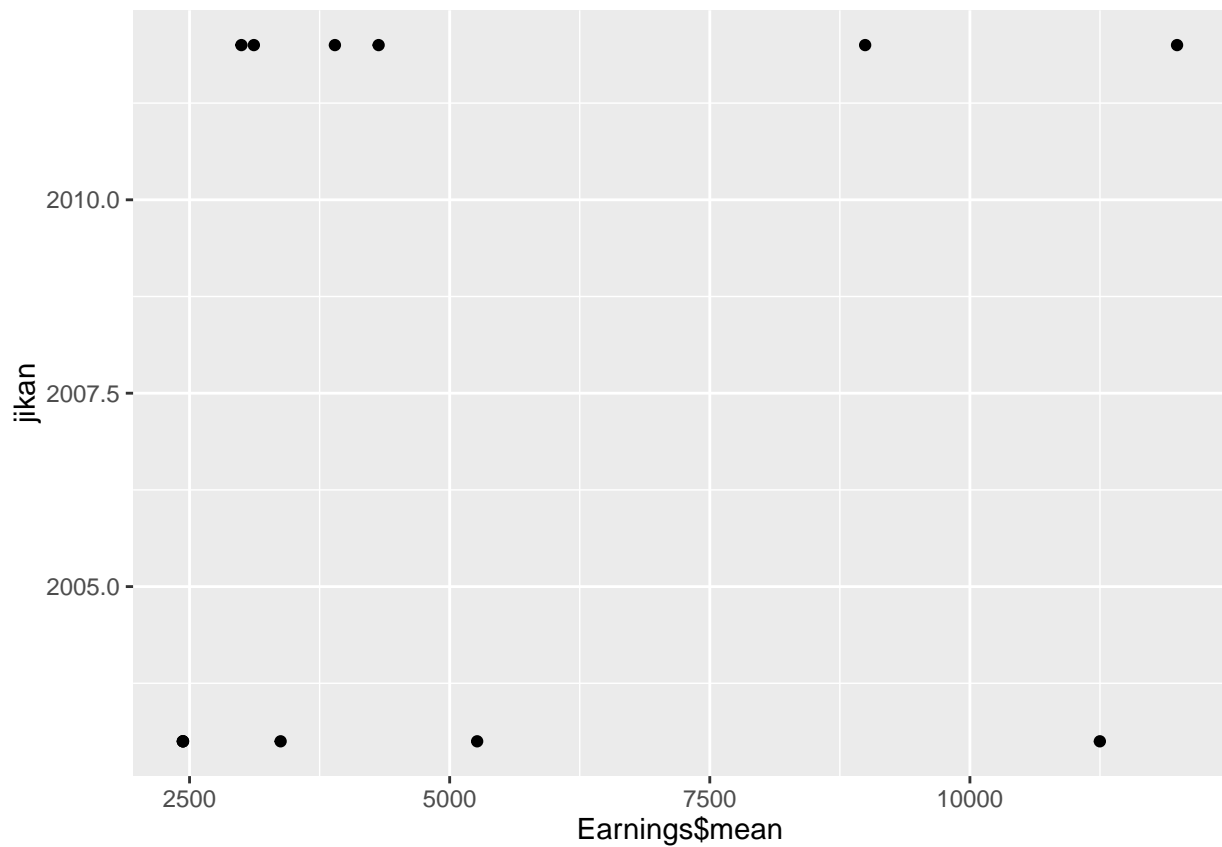
Then, we try to gather data from WorldBank by using Worldbank Data API.

Conduct basic descriptive statistics

The data we gathered in previous section partially statistically analysed (mean and median are already calculated). In this section, we try to figure out the trend of inequality graphically by using descriptive statistics.

We want to plot the data frame in earning mean vs time among each individual.

```
jikan <- c(2003,2003,2003,2003,2003,2003,2012,2012,2012,2012,2012,2012)
qplot(Earnings$mean, jikan)
```



Briefly describing

References

Leibbrandt, et al., <!-- // --> M. n.d. "Trends in South African Income Distribution and Poverty Since the Fall of Apartheid." OECD Publishing. doi:<http://dx.doi.org/10.1787/5kmms0t7p1ms-en>.