MAS286 Topic 3: Data Visualisation

Data Origins

After searching on the internet, we finally chose an interesting dataset that includes information on relative levels of income, output, input and productivity for about 183 countries throughout the years (between 1950 and 2019). This data comes from a website called *Penn World Table version 10.0* and is available at https://www.rug.nl/ggdc/productivity/pwt/.

We chose this dataset because it gives enough information that can be helpful for the plots we will make.

The first few rows of the raw data are shown below:

```
## # A tibble: 6 x 52
##   countrycode country currency_unit   year rgdpe rgdpo   pop   emp   avh    hc
##   <chr>       <chr>   <chr>          <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 ABW         Aruba   Aruban Guilder  1950    NA    NA    NA    NA    NA    NA
## 2 ABW         Aruba   Aruban Guilder  1951    NA    NA    NA    NA    NA    NA
## 3 ABW         Aruba   Aruban Guilder  1952    NA    NA    NA    NA    NA    NA
## 4 ABW         Aruba   Aruban Guilder  1953    NA    NA    NA    NA    NA    NA
## 5 ABW         Aruba   Aruban Guilder  1954    NA    NA    NA    NA    NA    NA
## 6 ABW         Aruba   Aruban Guilder  1955    NA    NA    NA    NA    NA    NA
## # ... with 42 more variables: ccon <dbl>, cda <dbl>, cgdpe <dbl>, cgdpo <dbl>,
## #   cn <dbl>, ck <dbl>, ctfp <dbl>, cwtfp <dbl>, rgdpna <dbl>, rconna <dbl>,
## #   rdana <dbl>, rnna <dbl>, rkna <dbl>, rtfpna <dbl>, rwtfpna <dbl>,
## #   labsh <dbl>, irr <dbl>, delta <dbl>, xr <dbl>, pl_con <dbl>, pl_da <dbl>,
## #   pl_gdpo <dbl>, i_cig <chr>, i_xm <chr>, i_xr <chr>, i_outlier <chr>,
## #   i_irr <chr>, cor_exp <dbl>, statcap <dbl>, csh_c <dbl>, csh_i <dbl>,
## #   csh_g <dbl>, csh_x <dbl>, csh_m <dbl>, csh_r <dbl>, pl_c <dbl>, ...
```

Research Questions

Our visualization mainly attempts to address the question "Do workers in richer countries work longer hours?". Hence, we first looked at the relationship between annual working hours and GDP per capita in each country. Simultaneously, through the change in time, it is more intuitive to judge whether there is a certain connection between the GDP growth and average working hours of a country.

A further question, "Do workers in countries with higher labor productivity work longer hours?", may be asked. Therefore, we also dug into the relationship between annual working hours and labor productivity.

Data Preparation

In our raw dataset, the relevant variables are: "real GDP" (*rgdpo*), "average annual hours worked by persons engaged" (*avh*), "population" (*pop*) and the "population engaged" (*emp*). The meanings of these variables are as follows:

- Real GDP (*rgdpo*): real gross domestic product, a measure of a country's economic output.
- Average annual hours worked by persons engaged (*avh*): the amount of real gross domestic product (GDP) produced by an hour of labor, which equals to the total number of hours actually worked per

year divided by the average number of people in employment per year.
- Population (*pop*): the whole population of every country in millions.
- Population engaged (*emp*): the population that is working in the country in millions.

Since the programming was done in two different parts using the same raw data, two different **csv** files are provided in our folder.

In both **csv** files, the raw data has been cleaned by deleting the columns where variables were not needed. GDP per capita was not available in the raw data, therefore, we created an additional column consisting this, dividing the real GDP by the population of the country. Similarly, we added a column consisting "Labor productivity", which was real GDP divided by the multiplication of Annual working hours(per engaged person) and Population engaged.

For the first **csv** file, we deleted all the rows which contained missing values and added a column containing countries' continents.

For the second **csv** file, as we were thinking that we would like to create a choropleth map plot, we had to use the 'maps' package in R, which included a world map, with a data including different county names and their corresponding coordinates. This data and our excel sheet data we had chosen from the website called 'Penn World Table version 10.0' had the names of the countries written differently. Hence, for us to be able to link them together in our R code, we had to link the county names in "maps" package with our excel sheet data in the exact same way, or else the data would not have been shown. We had to manually change some country names in the excel data in order to match the "maps" R package. Moreover, as some data were not present, some counties are still missing.

The first few rows of the processed data:

The First dataset:

```
##   countrycode   country year    rgdpo      pop      emp      avh   rgdpna
## 1         ARG Argentina 1950 50108.76 17.09182 6.608833 2034.000 196688.5
## 2         ARG Argentina 1951 51339.17 17.45758 6.713252 2037.867 208305.2
## 3         ARG Argentina 1952 46855.10 17.81597 6.819321 2041.741 195898.2
## 4         ARG Argentina 1953 49917.15 18.16862 6.927065 2045.622 204164.9
## 5         ARG Argentina 1954 52254.77 18.51717 7.036511 2049.511 210812.6
## 6         ARG Argentina 1955 56072.50 18.86324 7.147688 2053.407 225537.5
##   rgdpna.pop rgdpo.pop rgdpo..avh.emp. rgdpna..avh.emp. continent
## 1   11507.75  2931.739        3.727674         14.63198  Americas
## 2   11932.08  2940.796        3.752668         15.22619  Americas
## 3   10995.66  2629.950        3.365233         14.06983  Americas
## 4   11237.23  2747.438        3.522695         14.40809  Americas
## 5   11384.71  2821.963        3.623416         14.61803  Americas
## 6   11956.45  2972.580        3.820403         15.36661  Americas
```

The second dataset:

```
##   ï..countrycode country  currency_unit year rgdpo pop emp avh rgdpo.pop
## 1            ABW   Aruba Aruban Guilder 1950    NA  NA  NA  NA        NA
## 2            ABW   Aruba Aruban Guilder 1951    NA  NA  NA  NA        NA
## 3            ABW   Aruba Aruban Guilder 1952    NA  NA  NA  NA        NA
## 4            ABW   Aruba Aruban Guilder 1953    NA  NA  NA  NA        NA
## 5            ABW   Aruba Aruban Guilder 1954    NA  NA  NA  NA        NA
## 6            ABW   Aruba Aruban Guilder 1955    NA  NA  NA  NA        NA
##   rgdpo..avh.emp.
## 1              NA
## 2              NA
## 3              NA
## 4              NA
```

```
## 5              NA
## 6              NA
```

Visualisations

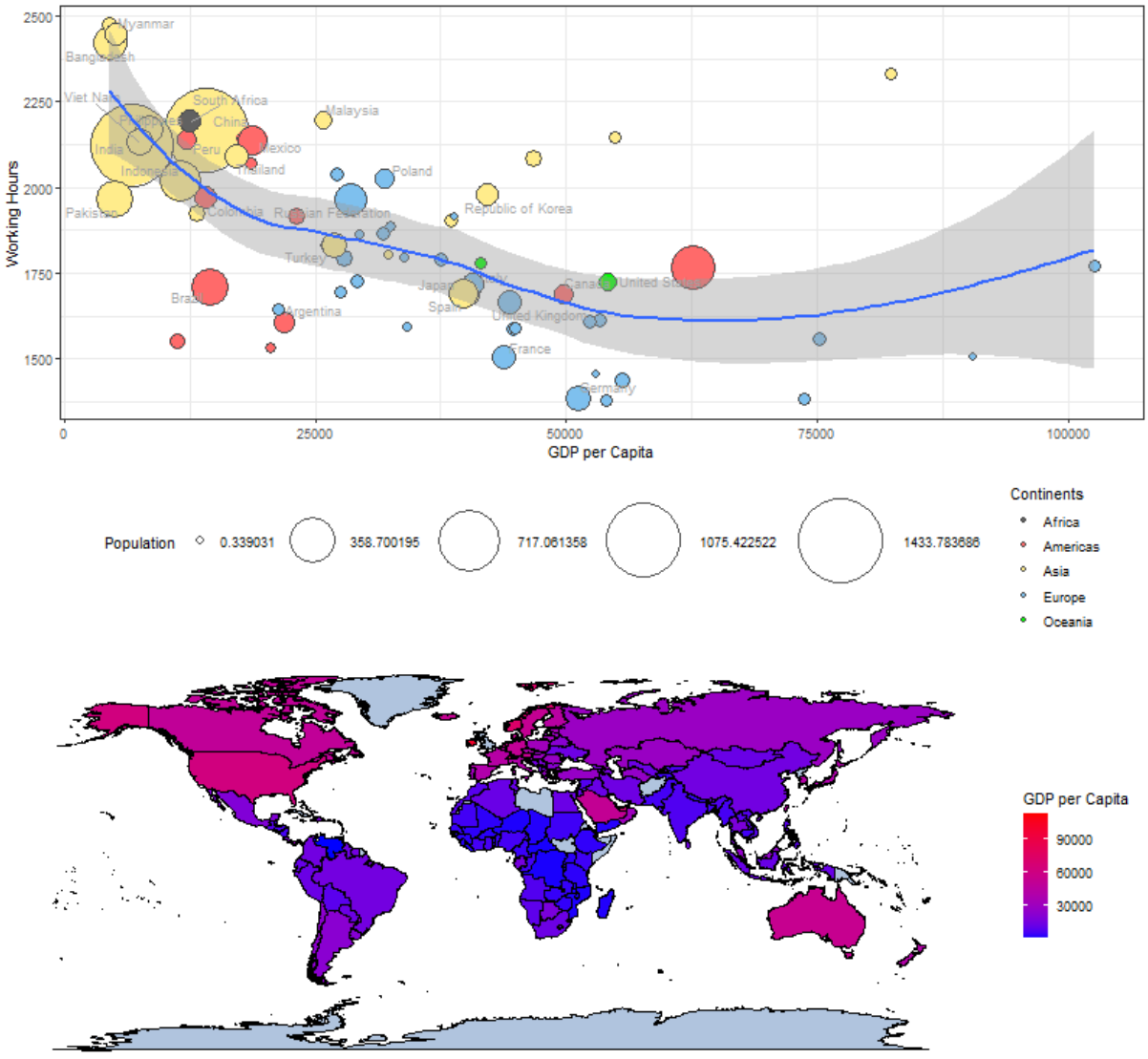

Figure 1: The relationship between annual working hours and GDP per capita, and GDP per capita around the world in 2019.

1st Graph (Scatter plot):

We have used the package shiny to produce an interactive plot application, where the viewer can change and select which variable to look over; the "GDP per capita" and the "Labor productivity". Also, the viewer is invited to look over the change in years. As our plot is interactive, when the viewer can choose between the x-axis being "GDP per capita" or "Labor productivity". In both cases the y-axis stays the same which represents the "Annual working hours (per engaged person). The viewer can determine whether the variable on the x-axis is linear or exponential. Within our first plot, when the"GDP per capita" is chosen, the size of

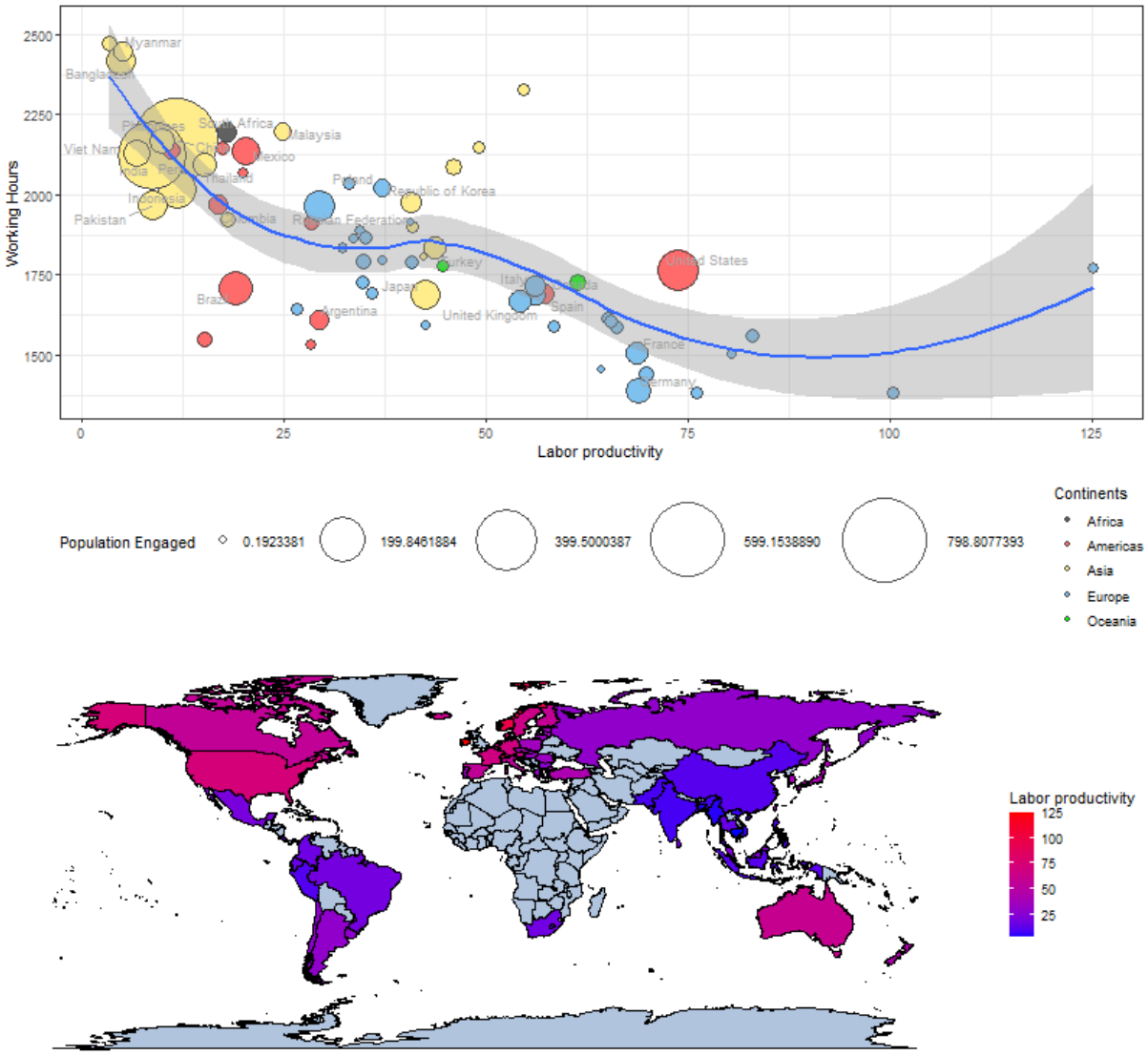

Figure 2: The relationship between annual working hours and labour productivity, and labour productivity around the world in 2019.

the dots stand for the population (in millions). When the "Labor productivity" is chosen, the size of the dots represents the population engaged. Furthermore, the different colors of the circles on the graph, stand for the continent. The blue line on the graph reveals the regression line and the grey section is the confidence interval to 95%.

2nd Graph (Choropleth map) :

This is also an interactive plot, where the viewer can change between "GDP per capita" and "Labor productivity". Any year can be selected by the slider bar. Within this plot, we have shown the GDP per capita or labor productivity using a choropleth map. The reason we have selected to show this information on such plot, is because it visually can look good to the viewer and it makes it easy to understand the bigger picture. Within our choropleth map we can show the data selected as colors. They are shaded using two colors(blue and red), where red represents high numbers and blue represents low numbers (the key explains what the different shades mean). At our attempt to show this in our plot, we figured that some counties had no data, therefore, we illustrated them in a grey color.

The shiny application is attached in the zipped folder.

Summary

In this assignment, we learnt how to make charts using R and how to make visual charts using Shiny applications. We have used ggplot2 to create the chart, which uses buttons to command mobile data from 1950 to 2019. Working on our main question "Do workers in richer countries work longer hours?" the answer we have come out with is that workers in poorer countries actually tend to work more, and sometimes much more.

If we had more time for our project we could have looked into more details and information. Such as the relationship of the GDP per capita and Annual working hours. Also, in regards to the data we have found, we think that if we had complete data, this would have helped us fill in the missing counties.