HW1

Mrinal Soni

10/1/2019

Homework 1

```
# Installing libraries
library(readr)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.1     v stringr 1.3.1
## v ggplot2 3.0.0     v forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.5.2

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
```

Problem 1: Take the Gapminder Test: http://forms.gapminder.org/s3/test-2018. What score did you receive? Did any of the answers surprise you? Choose a question from the test, re-state it, and answer it using visualization and summarization. Provide a figure and any relevant output with your answer.

Interpretation: I got 7 questions right out of 13. A few questions did surprise me, but since I couldnt find the relevant dataset for those questions, I'm answering one of the questions that kind of does make sense: "What does the majority of the world population live in?". This was the one question I was 100% sure of getting right. As expected, majority of the population lives in the middle class (that is made up of lower and upper middle class).

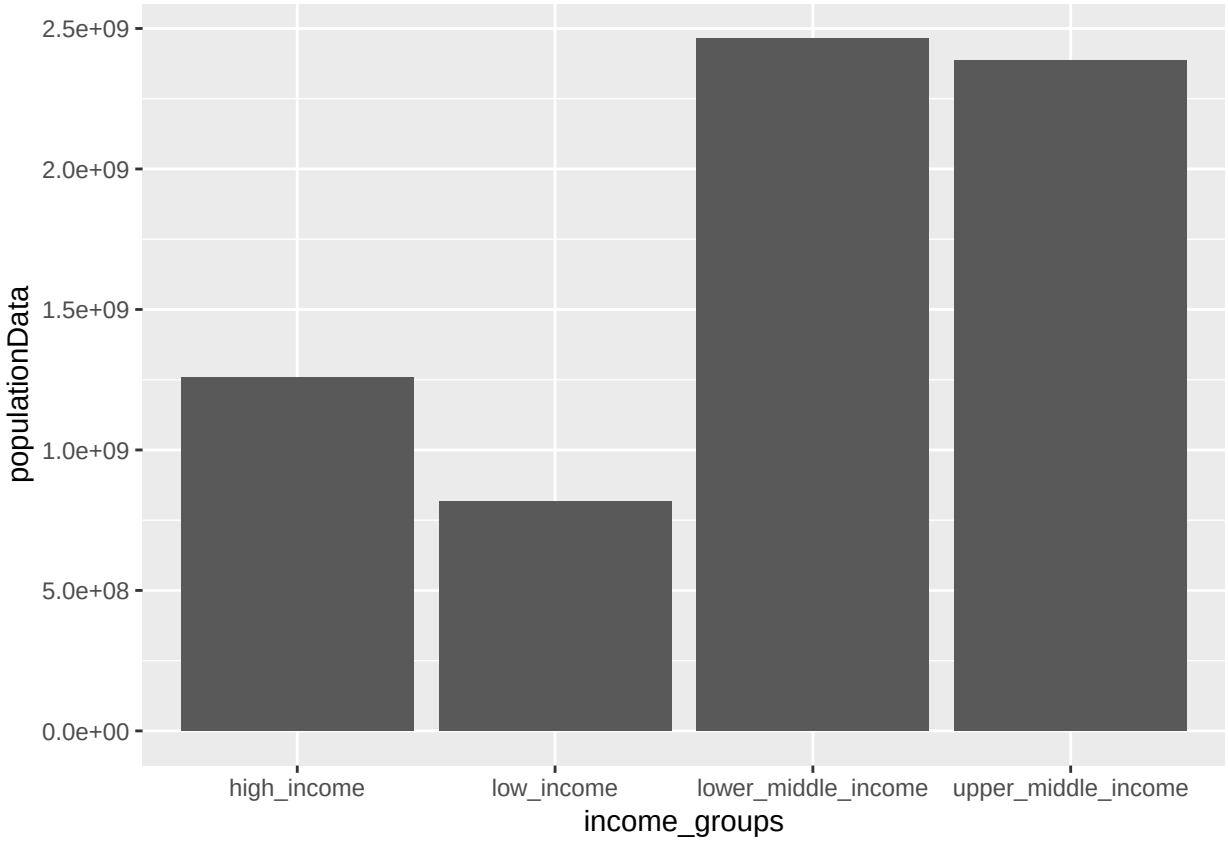
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `is--country` = col_logical(),
##   iso3166_1_numeric = col_integer(),
##   latitude = col_double(),
##   longitude = col_double(),
##   un_state = col_logical()
## )

## See spec(...) for full column specifications.
```

```
## Parsed with column specification:
## cols(
##   world_4region = col_character(),
##   color = col_character(),
##   description = col_character(),
##   `is--world_4region` = col_logical(),
##   latitude = col_double(),
##   longitude = col_double(),
##   name = col_character(),
##   name_long = col_character(),
##   name_short = col_character(),
##   rank = col_integer(),
##   shape_lores_svg = col_character()
## )
```

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   population_total = col_integer()
## )
```

```
## Warning: Ignoring unknown parameters: stat
```

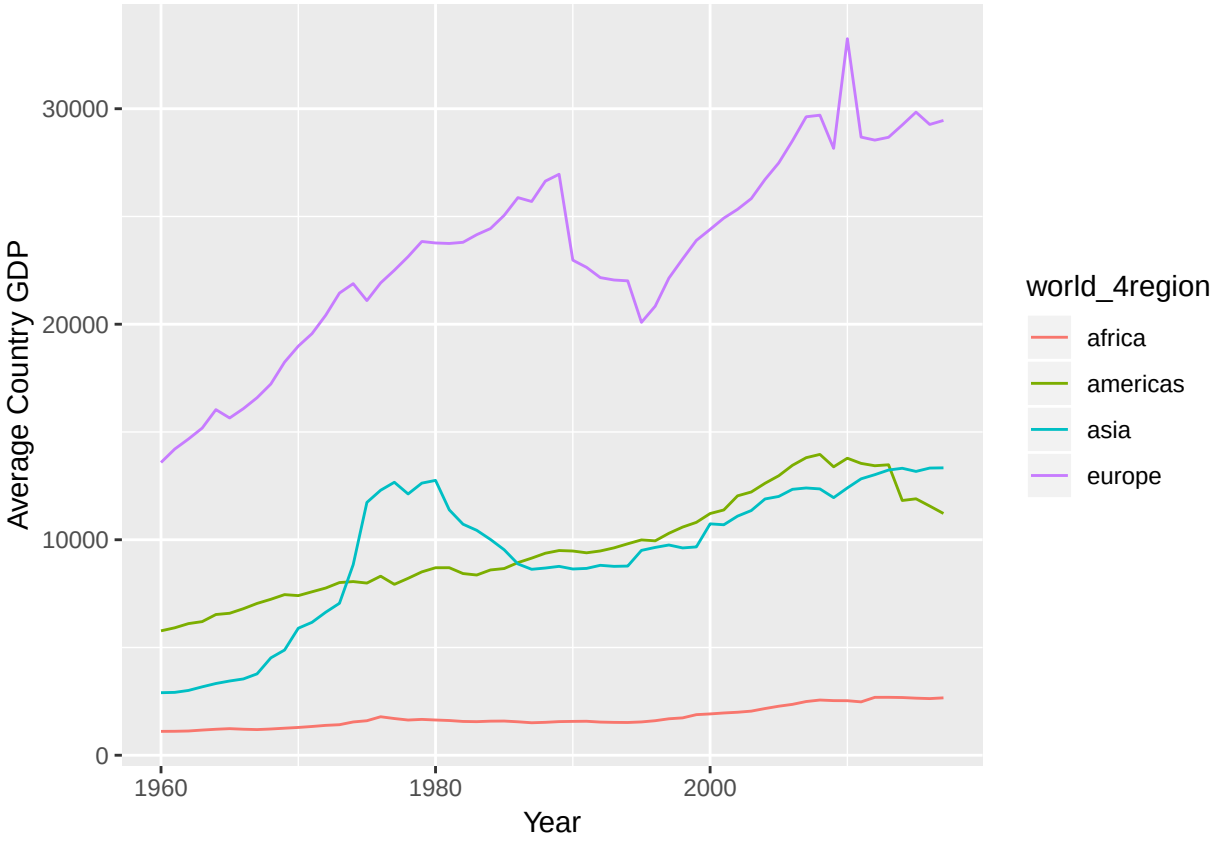


– – – –

Problem 2: Visualize the distribution of income (GDP / capita) across countries and continents, and how the distribution of income changes over time. Interpret the visualization and what you notice. Are they any notable trends and/or deviations from that trend? What caveats apply to your conclusions?

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   gdppercapita_us_inflation_adjusted = col_double()
## )
```

```
## Selecting by gdp_per_cap
```



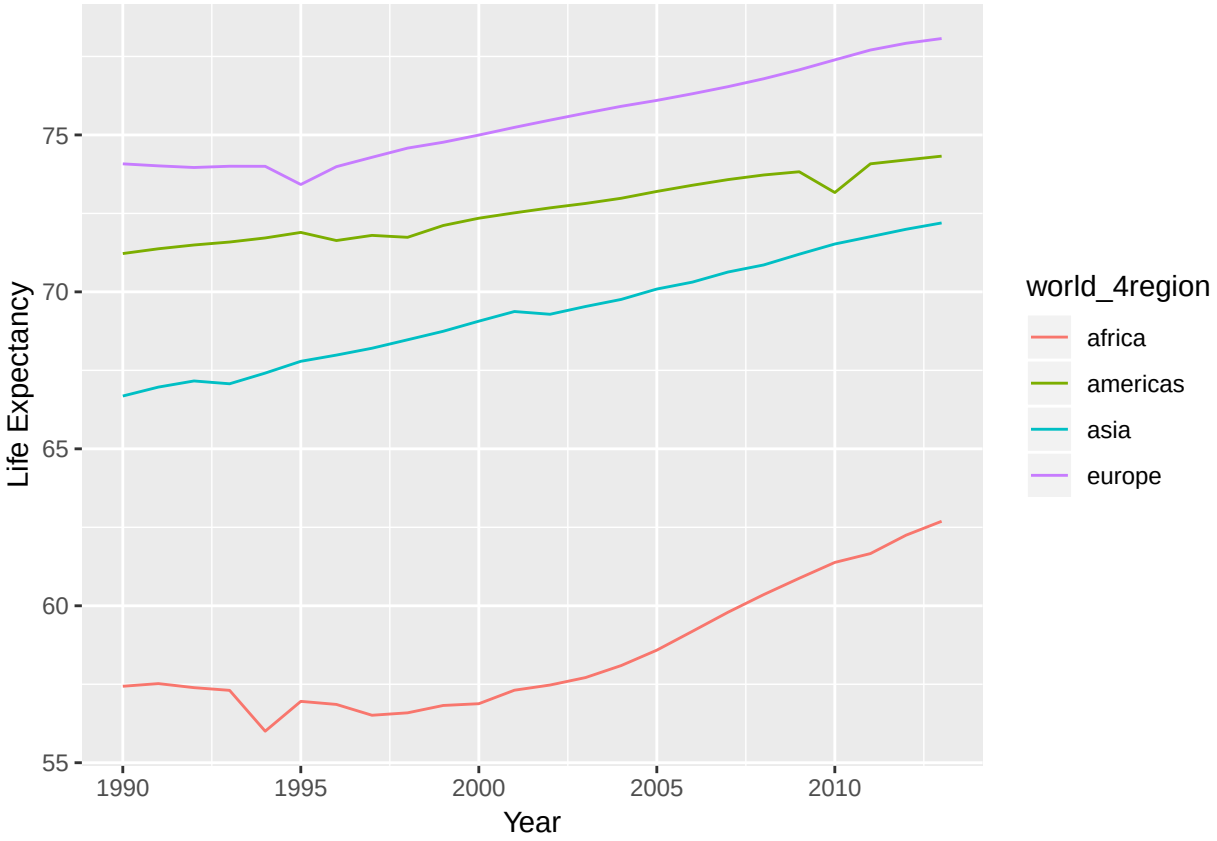
Interpretation: From the figure produced below, it is clear that the Average GDP for Europe is much higher than that of any other country and that of Africa is consistently lower than any other country, with Average GDP gradually increasing for America and Asia. What's interesting is the growth in GDP for America and Asia, there seems to be a sudden jump in GDP for Asia between the 1970s and 1980s followed by a sudden dip. And for America there is a sudden dip around the 2010-2015.

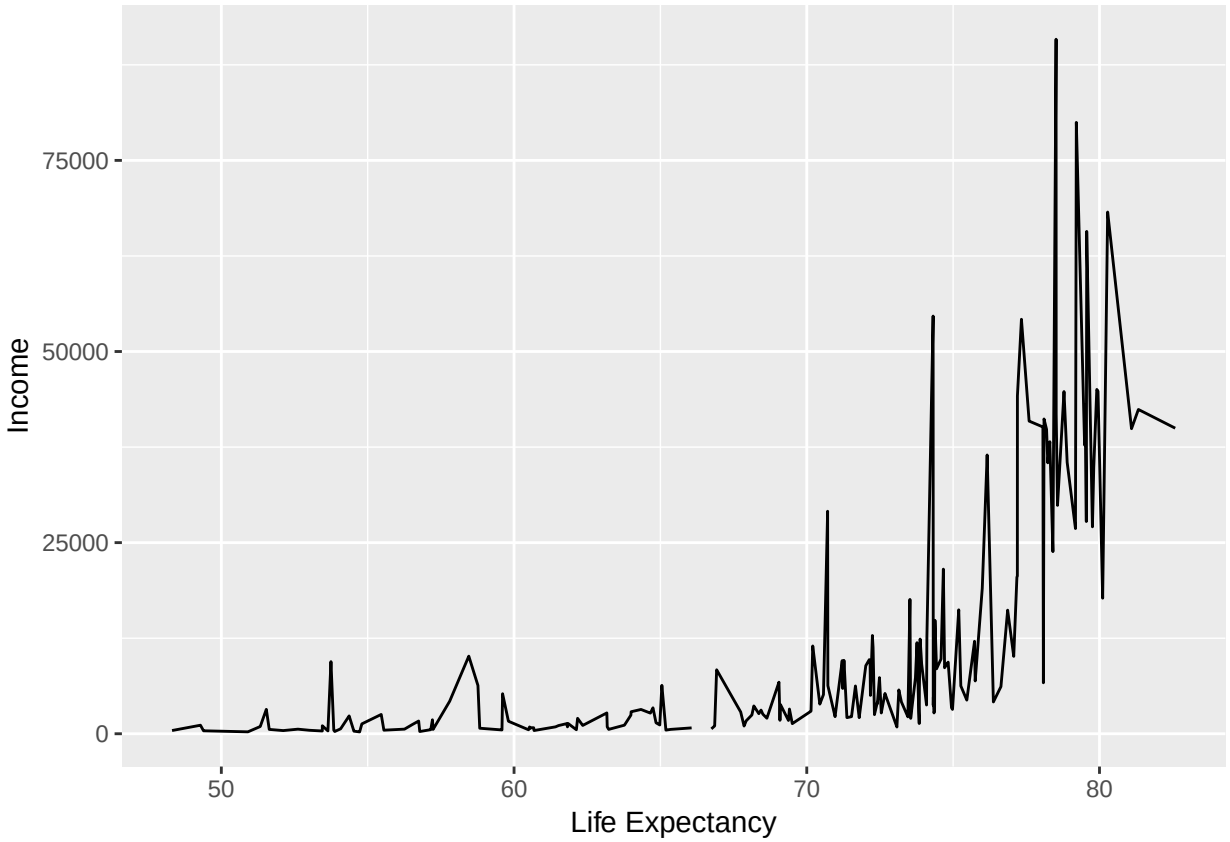
– – – – –

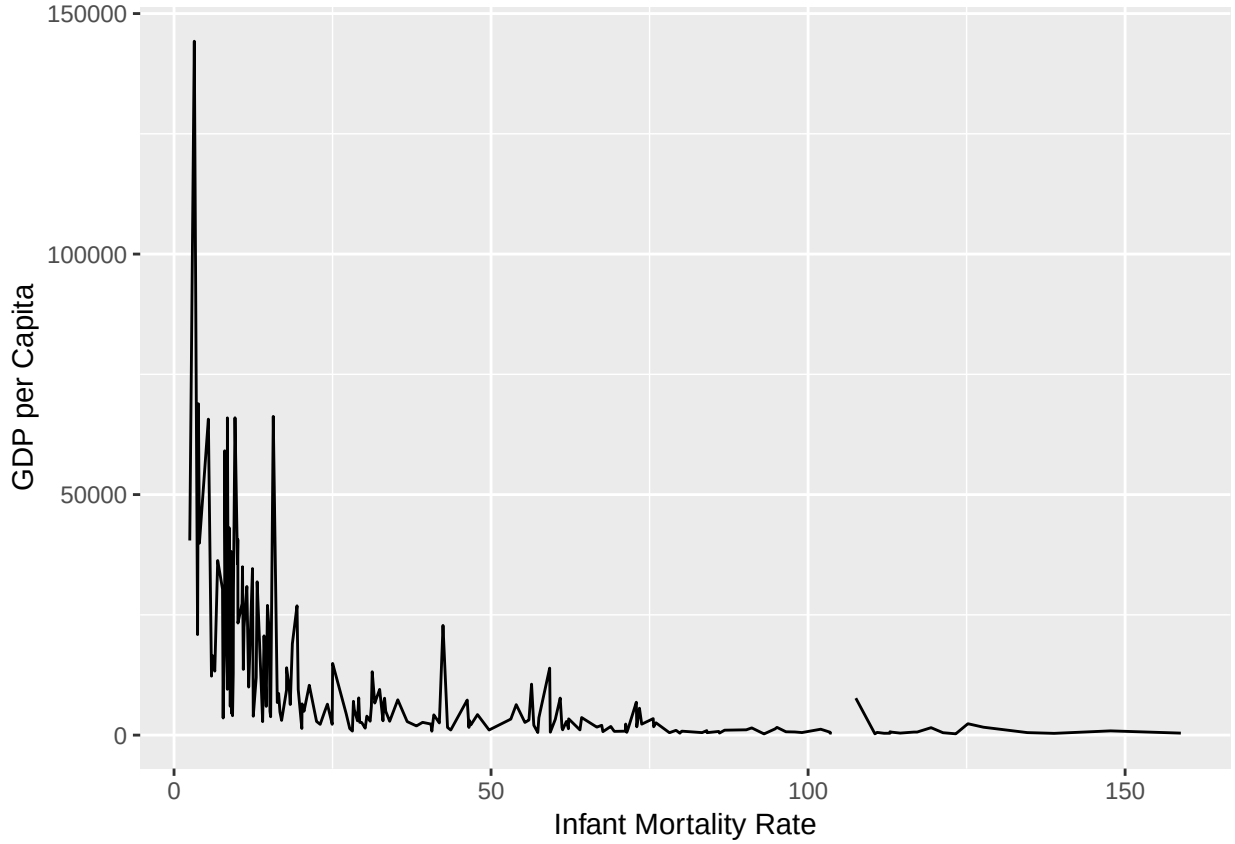
Problem 3: Use visualization to investigate the relationship between income (GDP / capita), life expectancy, and child mortality over time. How does each measure change over time within each continent? Interpret your visualizations, noting any trends and/or outliers.

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   infant_mortality_rate_per_1000_births = col_double()
## )
```

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   life_expectancy_at_birth_data_from_ihme = col_double()
## )
```




75000
50000
25000
0
Income
50
60
70
80
Life Expectancy

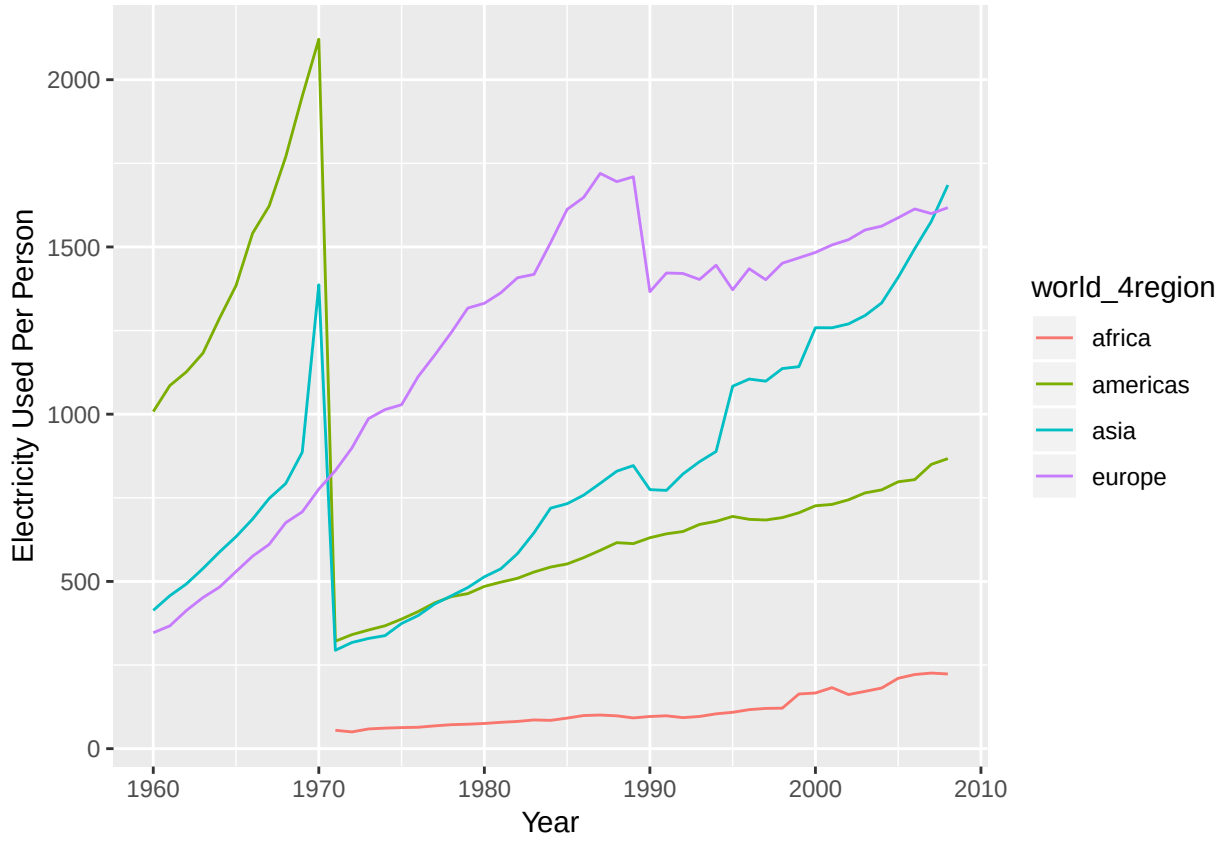


Interpretation: On observing the following plots, we can say that the relationship between Income and Life Expectancy is definitely positive as Life Expectancy seemes to have a positive effect on Income. It's like the more you live, the higher your income would be. From another graph, we can observe that Life Expectancy seems to be growing over the years for every continent with Europe having the highest average Life Expectancy and Africa having the lowest. There is also a drastic gap between the average Life Expectancy of Asia and Africa. And lastly from the graph between Infant Mortality Rate and Income, we see that there is a direct negative relationship between the two.

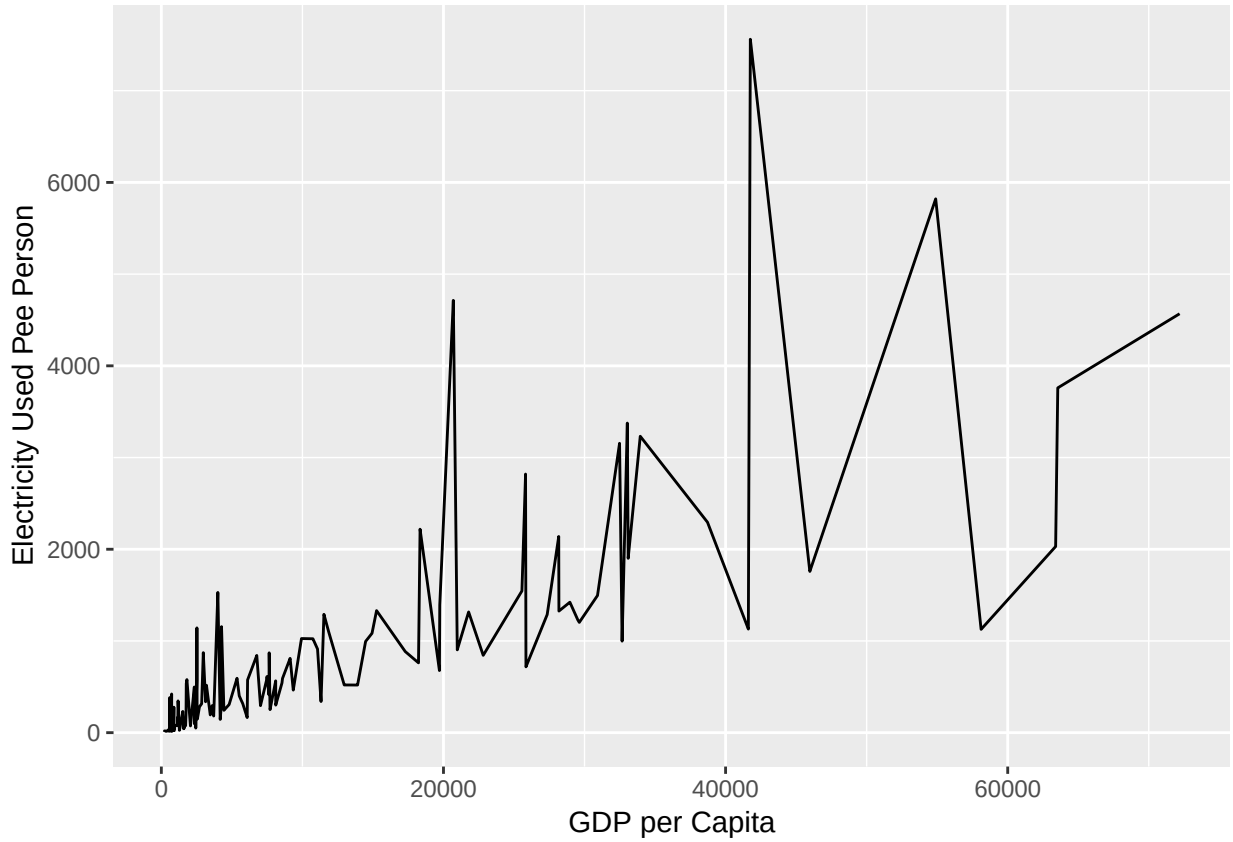
– – – –

Probelm 4: Choose two variables you have not investigated yet, and visualize their distributions, their relationship with each other, and how these change over time. Interpret your visualizations, noting any trends and/or outliers.

```
## Parsed with column specification:
## cols(
##   geo = col_character(),
##   time = col_integer(),
##   residential_electricity_use_per_person = col_double()
## )
```




```
## Warning: Removed 1 rows containing missing values (geom_path).
```



Interpretation: For this question I picked Electricity used per person and plotted that quantity against time to see how it varied over the years accross all continents. Initially America had the highest consumption of electiricity until it dipped around the 1970s only to rise slowly as time passed by. Electricity consumption for Europe seems to be the highest throughout the time, even after a small dip that they took around 1990. Even for Asia, there's a dip around 1970s right at the same time as America which could hint at some relationship or practice that both the countries adapted together. Africa stays the conuntry with lowest and most consistent comsupmtion of electricity per person. For fun I wanted to see if there was any relationship between Electricity consumption and GDP per Capita, only to observe that there isn't any strong relationship when plotted together, but there should be some correlation because individually both these variables grow, showing some correlation between the two signifying the growth of the given countries.

– – – –

Problem 5: Did you use static or interactive plots to answer the previous problems? Explore the data using the interactive visualization tools at https://www.gapminder.org/tools, and watch the TED talk "The best stats you've ever seen" at https://www.youtube.com/watch?v=hVimVzgtD6w. Discuss the advantages, disadvantages, and relative usefulness of using interactive/dynamic visualizations versus static visualizations.

Answer: For my results, I have used Static plots over Dynamic plots. Dynamic plots are definitely the better representation of flow of data, but like explained in the guest lecture

today, it is cahllenging to understand dynamic plots if you do not have enough practice. No such situation arises in Static plots, they are quite easy to understand. Like discussed in the lecture today, it's easy to execute Dynamic plots for high dimmensional data, giving dynamic the advantage over static. Another drawback for dynamic plots is that they are suitable only for a few number of audience, not everyone enjoys or understands them, making it tough for one to explain.