Project 1

Neil Klenk (nlk322)

Instructions

This knitted R Markdown document (as a PDF) *and* the raw R Markdown file (as .Rmd) should both be submitted to Canvas by 11:59pm on **Feb 23rd, 2015**. These two documents will be graded jointly, so they must be consistent (as in, don't change the R Markdown file without also updating the knitted document!).

All results presented *must* have corresponding code. **Any answers/results given without the corresponding R code that generated the result will be considered absent.** To be clear: if you do calculations by hand instead of using R and then report the results from the calculations, **you will not receive credit** for those calculations. All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean.)

For this project, you will be using the gapminder data set. You should be familiar with the gapminder data set from Homework 3.

```
library(gapminder)
head(gapminder)
```

```
## Source: local data frame [6 x 6]
##
##       country continent  year lifeExp      pop gdpPercap
##        (fctr)    (fctr) (int)   (dbl)    (int)     (dbl)
## 1 Afghanistan      Asia  1952  28.801  8425333  779.4453
## 2 Afghanistan      Asia  1957  30.332  9240934  820.8530
## 3 Afghanistan      Asia  1962  31.997 10267083  853.1007
## 4 Afghanistan      Asia  1967  34.020 11537966  836.1971
## 5 Afghanistan      Asia  1972  36.088 13079460  739.9811
## 6 Afghanistan      Asia  1977  38.438 14880372  786.1134
```

This data set contains life expectancies, population counts, and GDP per capita for 192 countries. Data are provided in five-year increments from 1952 to 2007. These data were compiled by the Gapminder non-profit organization as part of the Ignorance Project. You can learn more about the Ignorance Project here (http://www.gapminder.org/ignorance/).

Questions

Question 1: (5 pts) Is this data set tidy? Explain why or why not. If you conclude that the data set is not tidy, suggest a different way to represent this data set which *would* be tidy.

The data present in the gapminder dataset is considered tidy because it follows the three reules for a tidy data set. Each of the variables forms its own column, each observation forms its own row, and each type of observational unit forms a table.

Question 2: (25 pts) Select a year between 1952 and 2007 (remember that the gapminder data set only has data in five-year increments). In the year that you chose, group all countries in the data set into quartiles (i.e. four evenly-sized groups) based on population size. Again, *in just the year that you chose*, compute a Pearson correlation coefficient between life exptectancy and GDP per capita *for each quartile*. Display your data-frame with the correlation coefficients *and p-values* below.

HINTS: You can break data into quartiles using the function `ntile()` provided by the dplyr package. You can calculate Pearson correlation coefficients and p-values using the function `cor.test()`.

```
 Q1 = gapminder %>% filter(year == 1952) %>% mutate(Quartile = ntile(pop,4)) %>% filter(Q
uartile == '1')

head(Q1)
```

```
## Source: local data frame [6 x 7]
##
##                     country continent  year lifeExp     pop gdpPercap
##                      (fctr)    (fctr) (int)   (dbl)   (int)     (dbl)
## 1                  Albania    Europe  1952  55.230 1282697 1601.0561
## 2                  Bahrain      Asia  1952  50.939  120447 9867.0848
## 3                 Botswana    Africa  1952  47.622  442308  851.2411
## 4 Central African Republic    Africa  1952  35.463 1291695 1071.3107
## 5                  Comoros    Africa  1952  40.715  153936 1102.9909
## 6              Congo, Rep.    Africa  1952  42.111  854885 2125.6214
## Variables not shown: Quartile (int)
```

```
cor.test(Q1$lifeExp, Q1$gdpPercap, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Q1$lifeExp and Q1$gdpPercap
## t = 1.3582, df = 34, p-value = 0.1833
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1098655  0.5168714
## sample estimates:
##       cor
## 0.2268586
```

```
Q2 = gapminder %>% filter(year == 1952) %>% mutate(Quartile = ntile(pop,4)) %>% filter(Qu
artile == '2')

head(Q2)
```

```
## Source: local data frame [6 x 7]
##
##                  country continent  year lifeExp     pop gdpPercap
##                   (fctr)    (fctr) (int)   (dbl)   (int)     (dbl)
## 1                  Benin    Africa  1952  38.223 1738315 1062.7522
## 2                Bolivia  Americas  1952  40.414 2883315 2677.3263
## 3 Bosnia and Herzegovina    Europe  1952  53.820 2791000  973.5332
## 4                Burundi    Africa  1952  39.031 2445618  339.2965
## 5                   Chad    Africa  1952  38.092 2682462 1178.6659
## 6          Cote d'Ivoire    Africa  1952  40.477 2977019 1388.5947
## Variables not shown: Quartile (int)
```

```
cor.test(Q2$lifeExp, Q2$gdpPercap, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Q2$lifeExp and Q2$gdpPercap
## t = 7.3544, df = 33, p-value = 1.913e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6168270 0.8880913
## sample estimates:
##       cor
## 0.7880795
```

```
Q3 = gapminder %>% filter(year == 1952) %>% mutate(Quartile = ntile(pop,4)) %>% filter(Qu
artile == '3')

head(Q3)
```

```
## Source: local data frame [6 x 7]
##
##       country continent  year lifeExp     pop  gdpPercap Quartile
##        (fctr)    (fctr) (int)   (dbl)   (int)      (dbl)    (int)
## 1 Afghanistan      Asia  1952  28.801 8425333   779.4453        3
## 2      Angola    Africa  1952  30.015 4232095  3520.6103        3
## 3   Australia   Oceania  1952  69.120 8691212 10039.5956        3
## 4     Austria    Europe  1952  66.800 6927772  6137.0765        3
## 5     Belgium    Europe  1952  68.000 8730405  8343.1051        3
## 6    Bulgaria    Europe  1952  59.600 7274900  2444.2866        3
```

```
cor.test(Q3$lifeExp, Q3$gdpPercap, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Q3$lifeExp and Q3$gdpPercap
## t = 6.2902, df = 34, p-value = 3.626e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5333348 0.8557280
## sample estimates:
##       cor
## 0.7333698
```

```
Q4 = gapminder %>% filter(year == 1952) %>% mutate(Quartile = ntile(pop,4)) %>% filter(Qu
artile == '4')

head(Q3)
```

```
## Source: local data frame [6 x 7]
##
##       country continent  year lifeExp     pop  gdpPercap Quartile
##        (fctr)    (fctr) (int)   (dbl)   (int)      (dbl)    (int)
## 1 Afghanistan      Asia  1952  28.801 8425333   779.4453        3
## 2      Angola    Africa  1952  30.015 4232095  3520.6103        3
## 3   Australia   Oceania  1952  69.120 8691212 10039.5956        3
## 4     Austria    Europe  1952  66.800 6927772  6137.0765        3
## 5     Belgium    Europe  1952  68.000 8730405  8343.1051        3
## 6    Bulgaria    Europe  1952  59.600 7274900  2444.2866        3
```

```
cor.test(Q4$lifeExp, Q4$gdpPercap, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Q4$lifeExp and Q4$gdpPercap
## t = 8.5503, df = 33, p-value = 7.006e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6867865 0.9112424
## sample estimates:
##       cor
## 0.8300583
```

Are these correlations statistically significant? What conclusions, if any, can you draw from your analyses?

For quartiles 2, 3, and 4, the correlations are statistically significant. This is backed up by their respective p-values of 1.91e-08, 3.63e-07, and 7.10e-10 all being significantly below the 0.05 cutoff allowing us to reject that the correlations are due to random sampling. Quartile 1 has a p-value of 0.18, means that we fail to reject that the correlation is due to randome sampling. Thus we are able to come to the conclusion that those in the population quartiles of 2, 3, and 4 experience a strong correlation between life expectancy and GDP per capita.

Question 3: (40 pts)

a. (30 points) Use the ggplot2 library to create a plot displaying life expectancy over time for **three** countries of your choice. Your plot should display the points for each country in different colors, and the size of your points should reflect GDP per capita. Your code should be well-commented and describe the various steps you take to create this figure.

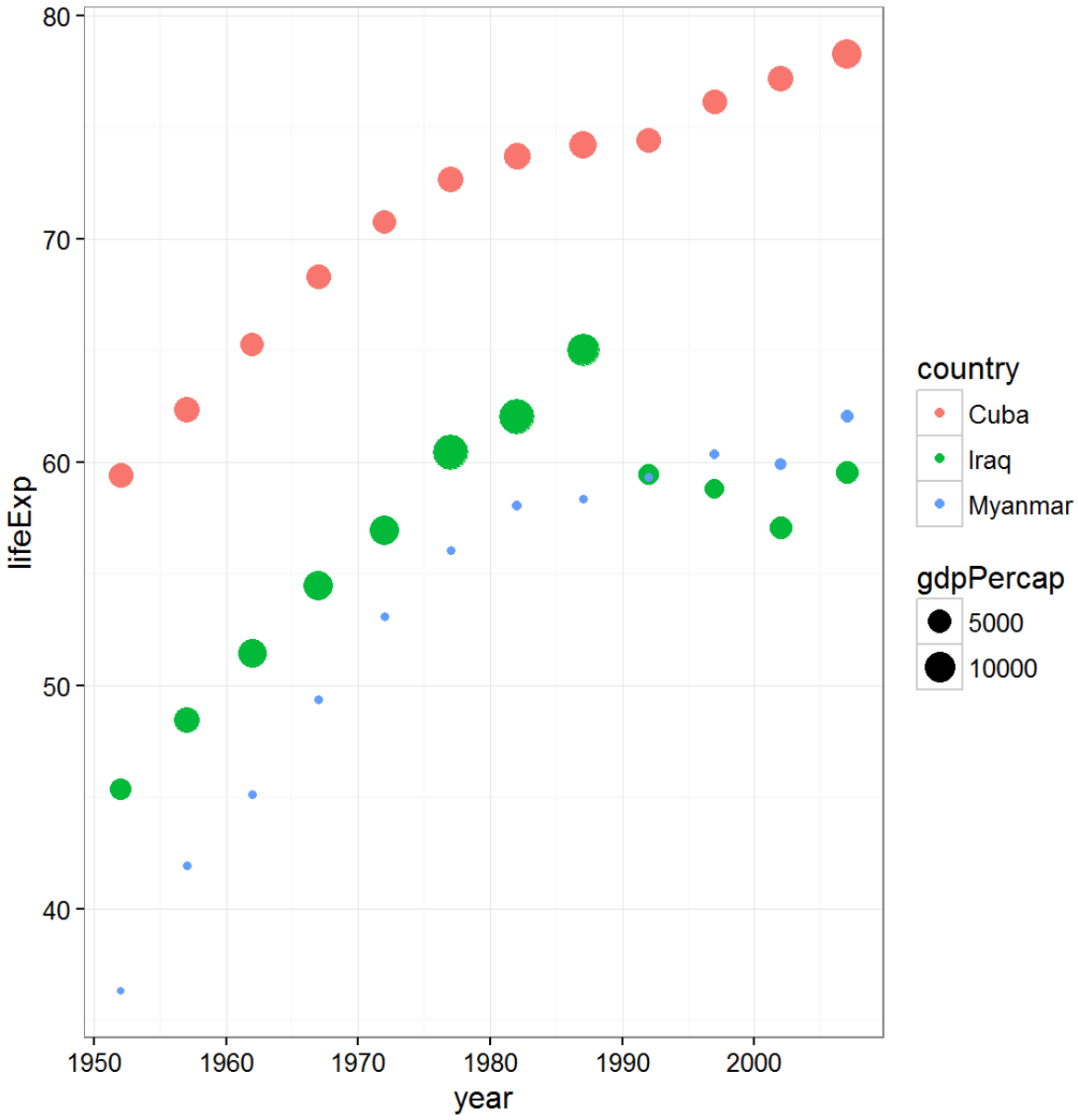
```
#filtering out the desired countries from the gapminder dataset
Cuba  = gapminder %>% filter(country == 'Cuba')
Iraq  = gapminder %>% filter(country == 'Iraq')
Myanmar = gapminder %>% filter(country == 'Myanmar')

#Combining the tables that only contai one country into a table that contains all of the
desired countries
tab = full_join(Cuba, Iraq)
```

```
## Joining by: c("country", "continent", "year", "lifeExp", "pop", "gdpPercap")
```

```
My_countries = full_join(tab, Myanmar)
```

```
## Joining by: c("country", "continent", "year", "lifeExp", "pop", "gdpPercap")
```

```
#Creating the life expectancy vs time plot with points colored by country and with points
scaled proportionally to the GDP at that time
My_countries %>% ggplot(aes(x = year, y = lifeExp, color = country, cex = gdpPercap)) + g
eom_point()
```



b. (10 points) Discuss the information (overarching trends, patterns, etc.) your final plot reveals. Be sure to include in your discussion the similarities/differences among countries and a clear, logical justification for why you selected the particular geom(s) used to represent this data. Please limit your full response to a maximum of 6 sentences.

This plot has revealed many trends in the data for these three countries. The most obvious is the general positive slope of the life expectancy vs year plot , with the exception of Iraq which peaks around 1987. That can be atributed to the instability in the region during that time. The GDP per capita also demonstrates Cuba's economic isolatio as the size of its points are consistantand, the effects of war on the GDP of Iraq. I chose to only use the geom_point() funciton because it gave me enough information to identify trends in the data, while avoiding filling the plot with so much ontent that it was difficult to look at.

Question 4: (30 pts) Think of **two** (and only two!) questions to ask about the gapminder data set. Clearly state each question in the spaces provided. For each question, use the ggplot2 library to create a plot that can help you find an answer to the question. For each plot, provide a clear explanation as to why this type of plot (e.g. boxplot, barplot, histogram, etc.) is best for providing the information you are

asking about. Answer your questions by interpreting your plot and any trends it reveals, or does not reveal, as the case may be. Your two plots *must* use different primary geoms. Please limit the discussion for each question-plot pair to 4-6 sentences.

Question 1

What was the effect of the Sierra Leone Civil war on its per capita GDP in realtion to other countries that were close in proximity?

```
SL <- gapminder %>% filter(country == 'Sierra Leone')
LI <- gapminder %>% filter(country == 'Liberia')
SLLI <- full_join(SL, LI)
```

```
## Joining by: c("country", "continent", "year", "lifeExp", "pop", "gdpPercap")
```

```
SLLI %>% ggplot(aes(x = year, y = gdpPercap, color = country)) +geom_line()
```

The civil war in Sierra Leone can be seen to have absolutely devestated their GDP. Liberia can be seen to also be facing problems with their GDP, but since the two plots show different patterns, it is likely that the drop in Sierra Leone's GDP wa due to war, and not its geographical location. This is further supported by the upturn in Sierra Leones GDP with the conclusion of their war, while Liberia's continues to fall.

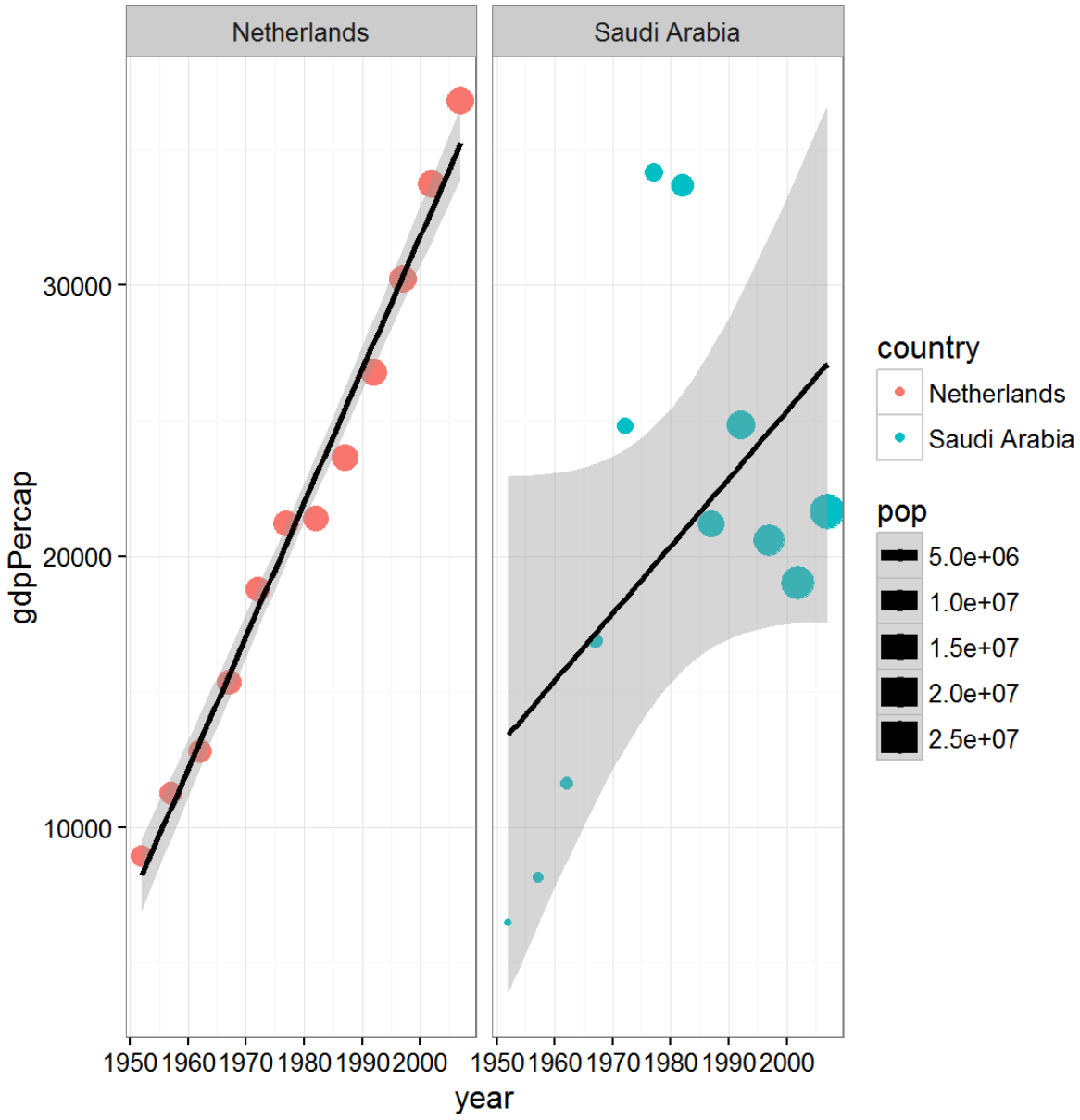
Question 2

Between Saudi Arabia and the Netherlands, Which has had greater change in per capita GDP over the known timeframe, and has this had any effect on population?

```
NL <- gapminder %>% filter(country == 'Netherlands')
SA <- gapminder %>% filter(country == 'Saudi Arabia')
NLSA <- full_join(SA, NL)
```

```
## Joining by: c("country", "continent", "year", "lifeExp", "pop", "gdpPercap")
```

```
NLSA %>% ggplot(aes(x = year, y = gdpPercap, color = country, size = pop)) + geom_point()
+ geom_smooth(method = 'lm', color = 'black') + facet_wrap(~country)
```



The Netherlands and Saudi Arabia have had very smilar changes in overall GDP in the given time frame when just taking into acount the maximium and minimum points. This is not the whole story though as the Netherlands have been consistently increasing their GDP year after year, While Saudi Arabia's is incredibly volatile, likely in response to the oil market. It is likely that when oil prices were high in the 80's many saudi families decided that they were fiscaly secure enough to have children, not anticipating the subsaquent drop in GDP.