Project 1

Evan Yacek ety78

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)
```

```
## Warning: package 'gapminder' was built under R version 3.1.3
```

```
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.

Gapminder is a tidy data set. Each variable forms a column. Each observation forms a row. Finally, each 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()`.

```
gapminderfilter <- filter(gapminder, year == 1992)
gapminderfilter %>% mutate(quintile = ntile(pop, 4)) -> gap_four
gap_four %>% group_by(quintile) %>% summarize(corval_Life_GDP = cor.test(lifeExp, gdpPercap)$estimat
e, pval = cor.test(lifeExp, gdpPercap)$p.value) -> CorTable
CorTable
```

```
## Source: local data frame [4 x 3]
##
##   quintile corval_Life_GDP         pval
##      (int)           (dbl)        (dbl)
## 1        1       0.6635218 1.029610e-05
## 2        2       0.6762837 8.222702e-06
## 3        3       0.7367386 3.006821e-07
## 4        4       0.7892064 1.768142e-08
```

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

With a p-value far below .01 we can say our results are statistically significant. From our results we can say that their is a statisticaly significant positive correalation for each quintile(grouped by pop.) between life expectancy and GDP.

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.

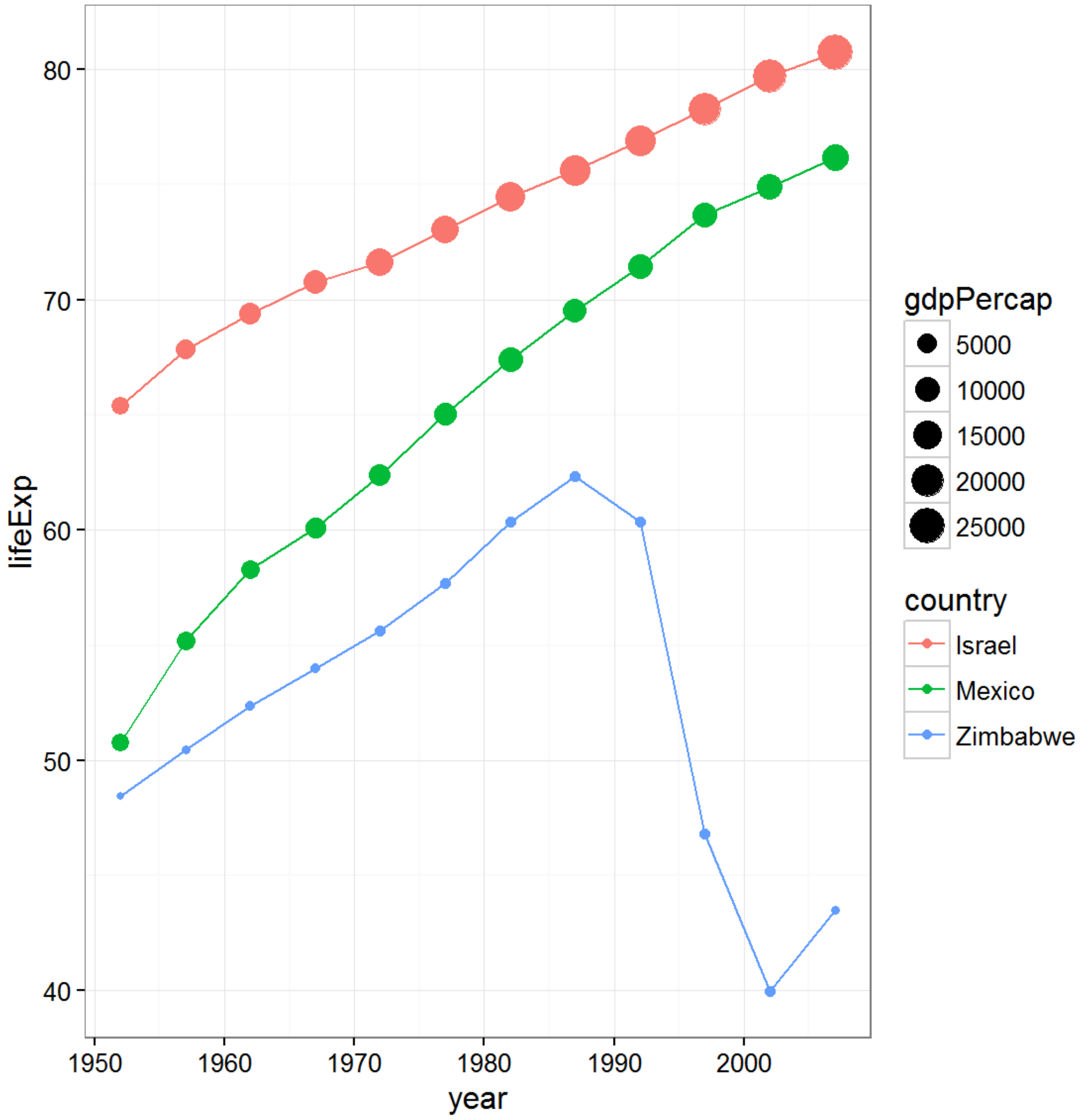
```
#Creates Dataset containing all elements in range of years described

gapTime <- filter(gapminder, year >= 1952 & year <= 2007)

#Filters dataset gapTime so that only three countries are left

threeCountryData <- filter(gapTime, country == "Zimbabwe" | country == "Israel" | country == "Mexico"
)

#Creates a plot of life Expectancy over time, grouped and colored by country
#Geom_point is altered to reflect the GDP at the time
ggplot(data = threeCountryData, aes(x = year, y= lifeExp, group = country, color = country))+geom_lin
e() + geom_point(aes(size = gdpPercap))
```



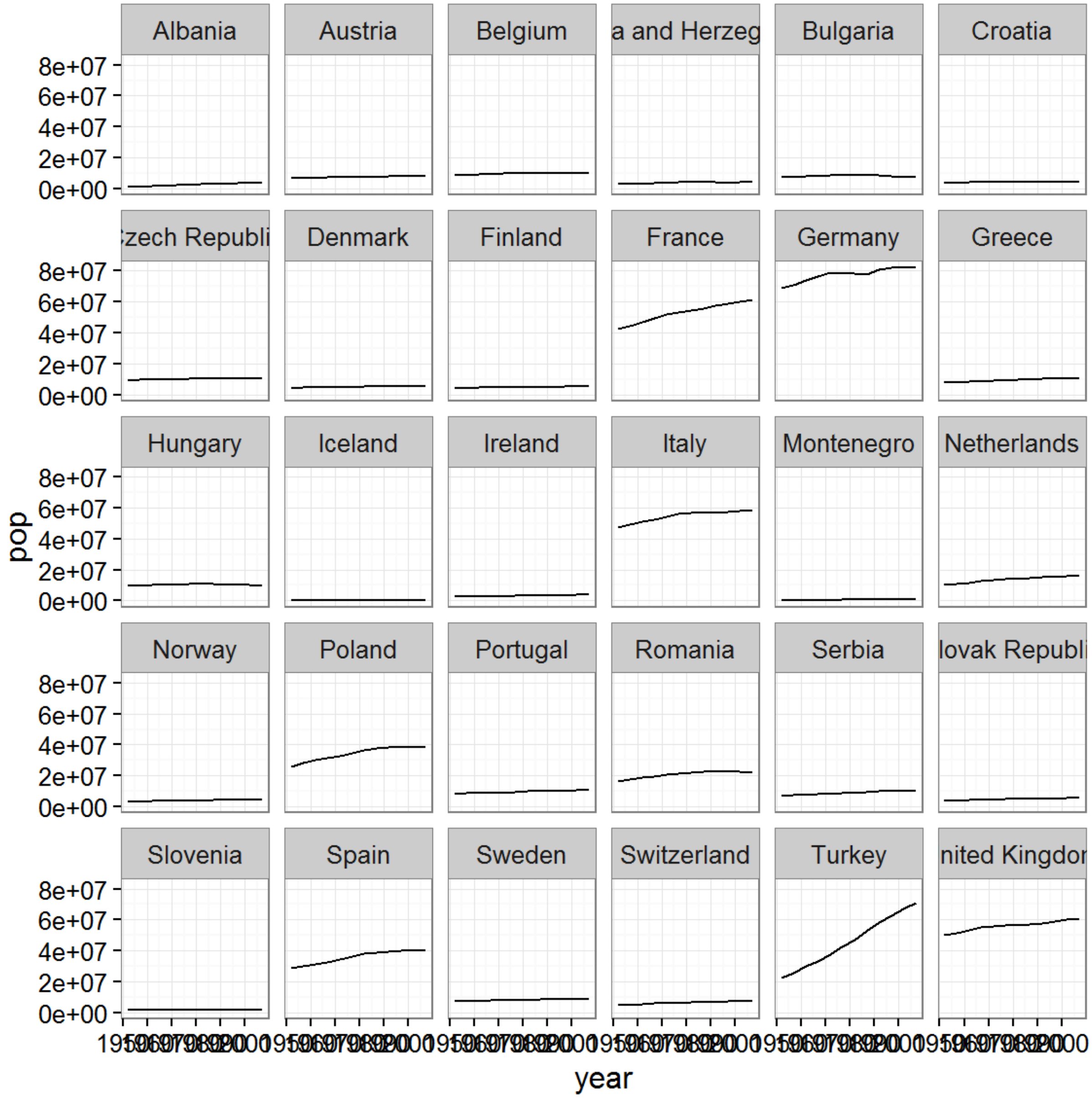
b. (10 points) Discuss the information (overarching trends, patterns, etc.) your final plot reveals. Be sure to include in your discussion the similarities/differences among countries and a clear, logical justification for why you selected the particular geom(s) used to represent this data. Please limit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

Both Israel and Mexico experienced a steady increase in life expectancy over time, as well as their overall GDP increasing. Zimbabwe on the other hand was steadily increasing then experienced a sudden drop in life expectancy in the early 90s, possibly due to the AIDS crisis. Zimbabwe’s GDP has seemed to stay fairly consistent over time. I used geom_line to connect my geom_points which I sized according to gdpPercap.

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 Which European country had the most dramatic change in population over the time period 1952-2007.

```
gapTime2 <- filter(gapminder, year >= 1952 & year <= 2007 & continent== "Europe")
ggplot( data = gapTime2, aes(x = year, y = pop))+facet_wrap(~country)+geom_line()
```



```
#55 is used as the difference of years
gapTime2 %>% group_by(country) %>% summarize(popRate = (max(pop)-min(pop))/(55*min(pop))) %>% arrange
(desc(popRate)) -> out
out
```

```
## Source: local data frame [30 x 2]
##
##                   country     popRate
##                    (fctr)       (dbl)
## 1                  Turkey 0.040003664
## 2                 Albania 0.032854439
## 3                 Iceland 0.018919968
## 4              Montenegro 0.013461524
## 5  Bosnia and Herzegovina 0.011473229
## 6             Netherlands 0.010838045
## 7             Switzerland 0.010345175
## 8         Slovak Republic 0.009654516
## 9                  Serbia 0.009213815
## 10                 Poland 0.009132692
## ..                    ...         ...
```

I used geom_line with a facet wrap on country to determine which country had the highest growth rate over time. Just looking at the ggplot it was easy to tell that Turkey had a dramatic increase due to its slope. However, in order to confirm I calculated the population rate using max and min population totals, so that dramatic variation could be accounted for. After doing so I created a table with this info showing that in fact Turkey had the highest growth rate, however countries with lower populations were not represented as well on my plot, but appeared higher than expected on my table than the plot would indicate.

Question 2

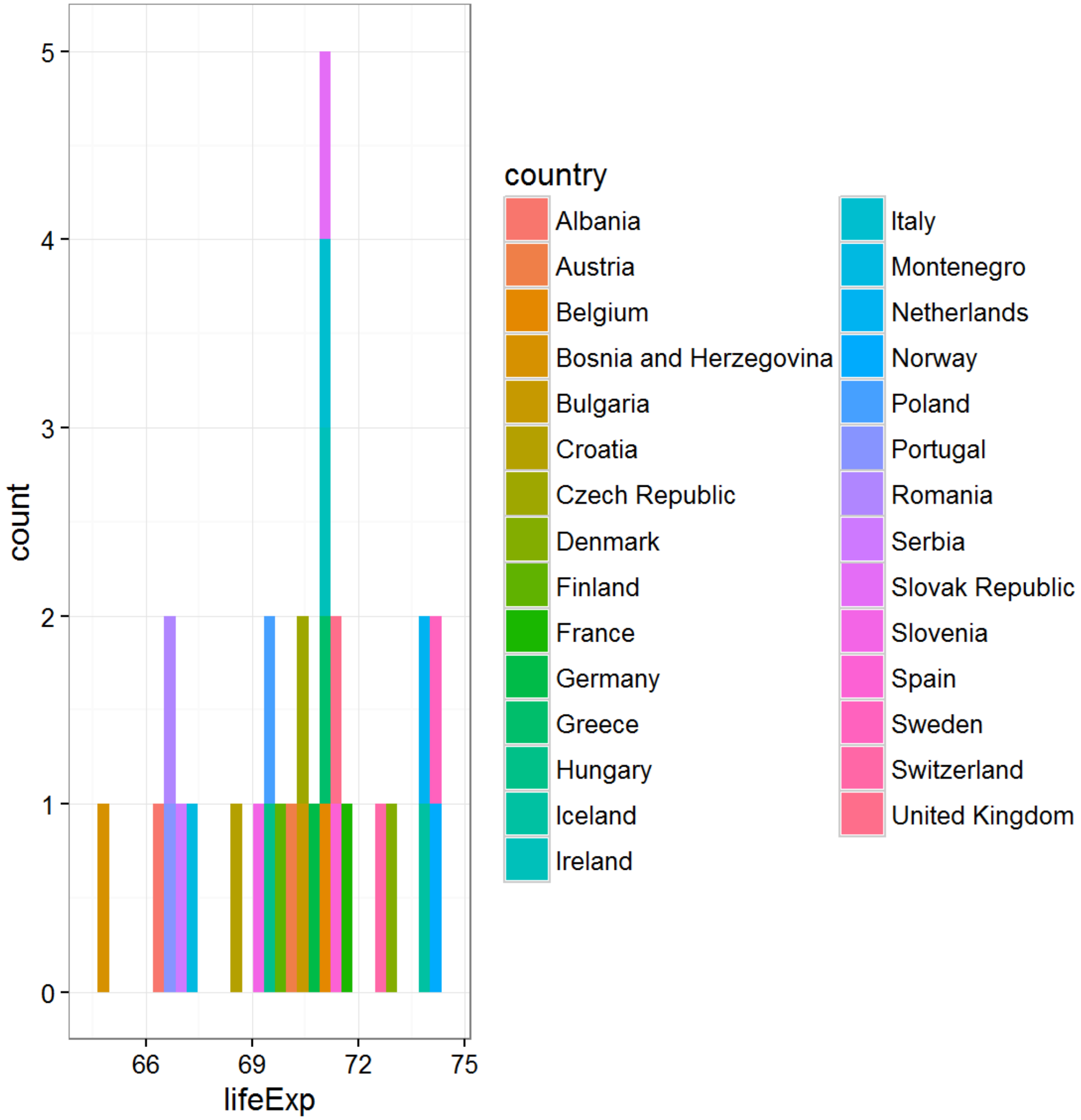
Determine if 72 could be the true mean life expectancy for European countries in 1967?

```
#Filter year, country, and outlier
gapTime3 <- filter(gapminder, year == 1967 & continent == "Europe" & lifeExp >= 55)




ggplot(gapTime3, aes(lifeExp, fill = country))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
t.test(gapTime3$lifeExp, mu =69)
```

```
##
##  One Sample t-test
##
## data:  gapTime3$lifeExp
## t = 2.7461, df = 28, p-value = 0.01042
## alternative hypothesis: true mean is not equal to 69
## 95 percent confidence interval:
##  69.32234 71.21504
## sample estimates:
## mean of x
##  70.26869
```

Using the geom_histogram function we can see that most countries life expectancies are between 65 and 75 for European countries in 1967. The histogram shows a somewhat normal distribution . In order to determine if their is a statistical difference in life expectency, first I filtered out the one outlier, and next ran a one sample t.test to see if the mean was equal to 72. The results confirmed what the histogram showed. With a p-value < .01 we reject the null hypothesis, thus the true mean is not 72, rather the true mean lies somewhere between 70-71(the area with the highest count), the values that do not reject the null in the t.test.