Homework 4

Cody

October 5, 2017

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(gapminder)
```

Choose your own adventure Pick one of the data reshaping prompts and do it.

Pick a join prompt and do it.

It is fine to work with a new dataset and/or create variations on these problem themes.

General data reshaping and relationship to aggregation Problem: You have data in one "shape" but you wish it were in another. Usually this is because the alternative shape is superior for presenting a table, making a figure, or doing aggregation and statistical analysis.

Solution: Reshape your data. For simple reshaping, gather() and spread() from tidyr will suffice. Do the thing that it possible / easier now that your data has a new shape.

Prompts:

Activity #2

Make a tibble with one row per year and columns for life expectancy for two or more countries. Use knitr::kable() to make this table look pretty in your rendered homework. Take advantage of this new data shape to scatterplot life expectancy for one country against that of another.

```
bothcountries <- gapminder %>%
                   filter(country=="Canada"|country=="United States")
country1 <- gapminder %>%
              filter(country=="Canada") %>%
              mutate(CanlifeExp=lifeExp) %>%
              select(year,CanlifeExp)
country2 <- gapminder %>%
              filter(country=="United States") %>%
              mutate(USlifeExp=lifeExp) %>%
              select(year,USlifeExp)
newtibble <- data.frame(year=bothcountries$year)
newtibble <- left_join(newtibble,country1)
```

```
## Joining, by = "year"
```

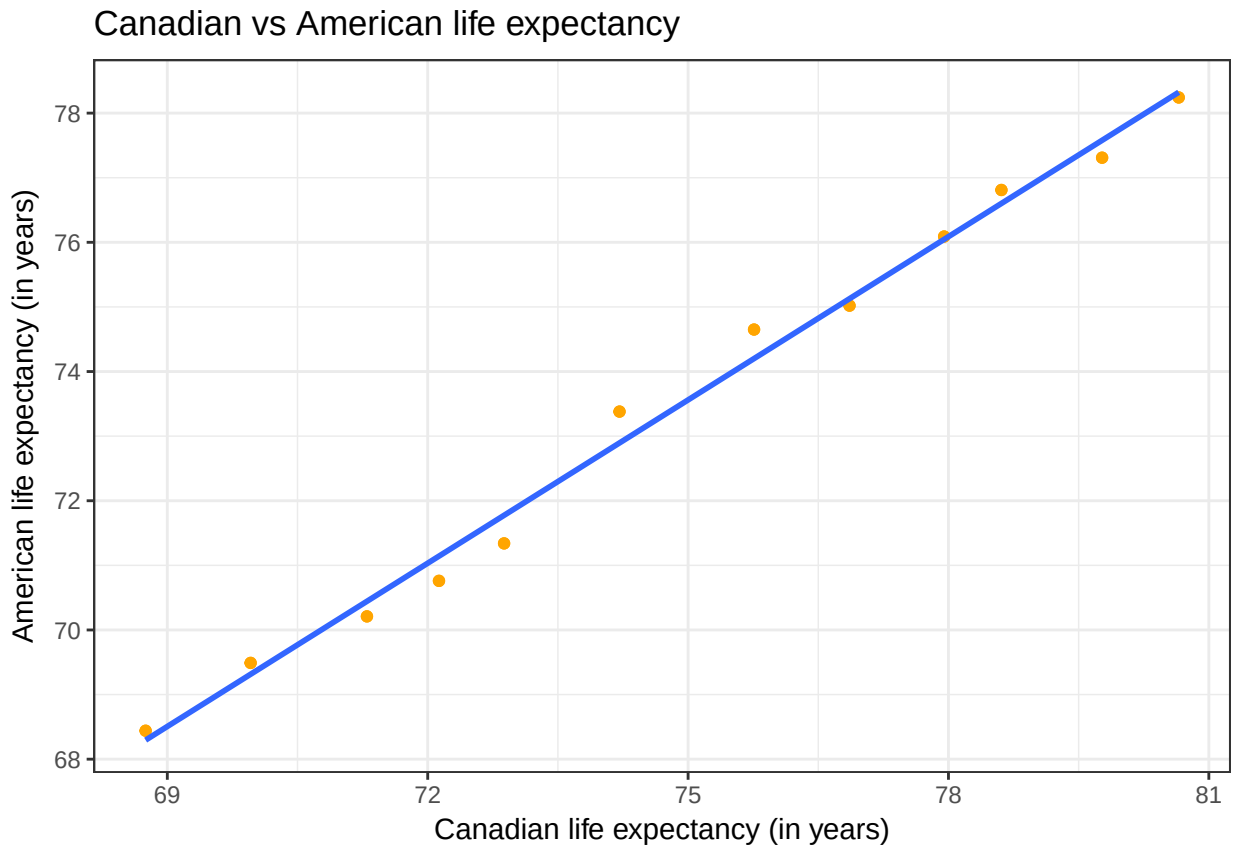
```
newtibble <- left_join(newtibble,country2)
```

```
## Joining, by = "year"
```

```
knitr::kable(newtibble)
```

year	CanlifeExp	USlifeExp
1952	68.750	68.440
1957	69.960	69.490
1962	71.300	70.210
1967	72.130	70.760
1972	72.880	71.340
1977	74.210	73.380
1982	75.760	74.650
1987	76.860	75.020
1992	77.950	76.090
1997	78.610	76.810
2002	79.770	77.310
2007	80.653	78.242
1952	68.750	68.440
1957	69.960	69.490
1962	71.300	70.210
1967	72.130	70.760
1972	72.880	71.340
1977	74.210	73.380
1982	75.760	74.650
1987	76.860	75.020
1992	77.950	76.090
1997	78.610	76.810
2002	79.770	77.310
2007	80.653	78.242

```
newtibble %>%
  ggplot(aes(x=CanlifeExp,y=USlifeExp))+
  geom_point(color="orange")+
  geom_smooth(method=lm,se=FALSE)+
  theme_bw()+
  ggtitle("Canadian vs American life expectancy")+
  xlab("Canadian life expectancy (in years)")+
  ylab("American life expectancy (in years)")
```




```
linearFit <- lm(USlifeExp~CanlifeExp,data=newtibble)
summary(linearFit)
```

```
##
## Call:
## lm(formula = USlifeExp ~ CanlifeExp, data = newtibble)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43450 -0.24230 -0.01808  0.18363  0.48508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.37918    1.22834    8.45 2.35e-08 ***
## CanlifeExp   0.84242    0.01638   51.44  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3037 on 22 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9914
## F-statistic:  2646 on 1 and 22 DF,  p-value: < 2.2e-16
```

So above we have put together a data set called "newtibble" that has a column for years, for Canadian life expectancy and for American life expectancy. We then plot the two together as a scatterplot and perform a linear fit to see if there is some relationship between the two life expectancies.

From the summary of the linear fit, we have that since $r^2 = .992$, more than 99% of the variation in the

data is explained by the relationship between Canadian and American life expectancy. We also see a strong degree of positive correlation as $r = 0.996$, namely the higher Canadian life expectancy goes, the higher American life expectancy goes. From the linear model, we can say that for every year increase in Canadian life expectancy, there is about a 0.84 year increase in American life expectancy.

Join, merge, look up Problem: You have two data sources and you need info from both in one new data object.

Solution: Perform a join, which borrows terminology from the database world, specifically SQL.

Prompts:

Activity #1

Create a second data frame, complementary to Gapminder. Join this with (part of) Gapminder using a dplyr join function and make some observations about the process and result. Explore the different types of joins. Examples of a second data frame you could build: One row per country, a country variable and one or more variables with extra info, such as language spoken, NATO membership, national animal, or capitol city. If you really want to be helpful, you could attempt to make a pull request to resolve this issue, where I would like to bring ISO country codes into the gapminder package. One row per continent, a continent variable and one or more variables with extra info, such as northern versus southern hemisphere.

```
hemisphere <- data.frame(continent=c("Europe","Americas","Asia","Africa","Oceania"),
                         hemisphere=c("Northern","Northern","Northern","Northern","Southern"))

specialcasesAfrica <- data.frame(country=c("Botswana","Angola","Burundi","Comoros","Lesotho",
                          "Madagascar","Malawi","Mauritius","Mayotte","Mozambique","Namibia",
                          "Reunion","Rwanda","Seychelles","South Africa",
                          "Swaziland","Tanzania","Zambia","Zimbabwe"),
                        hemisphere=rep("Southern",19))

specialcasesAmericas <- data.frame(country=c("Argentina","Bolivia","Chile","Paraguay","Peru",
                                             "Uruguay"),
                                   hemisphere=rep("Southern",6))

newdataset <- left_join(gapminder,hemisphere,by="continent")
newdataset <- left_join(newdataset,specialcasesAmericas,by="country")
```

```
## Warning: Column `country` joining factors with different levels, coercing
## to character vector
```

```
newdataset <- mutate(newdataset,hemisphere= ifelse(is.na(hemisphere.y),"Northern","Southern"))
newdataset <- select(newdataset,-hemisphere.x,-hemisphere.y)
newdataset <- left_join(newdataset,specialcasesAfrica,by="country")
```

```
## Warning: Column `country` joining character vector and factor, coercing
## into character vector
```

```
newdataset <- mutate(newdataset,hemisphere= ifelse(is.na(hemisphere.y),"Northern","Southern"))
newdataset <- select(newdataset,-hemisphere.x,-hemisphere.y)

tail(newdataset)
```

```
## # A tibble: 6 x 7
##    country continent  year lifeExp      pop gdpPercap hemisphere
##      <chr>    <fctr> <int>   <dbl>    <int>     <dbl>      <chr>
## 1 Zimbabwe    Africa  1982  60.363  7636524  788.8550   Southern
```

```
## 2 Zimbabwe    Africa  1987  62.351  9216418  706.1573   Southern
## 3 Zimbabwe    Africa  1992  60.377 10704340  693.4208   Southern
## 4 Zimbabwe    Africa  1997  46.809 11404948  792.4500   Southern
## 5 Zimbabwe    Africa  2002  39.989 11926563  672.0386   Southern
## 6 Zimbabwe    Africa  2007  43.487 12311143  469.7093   Southern
```

I have had quite a bit of trouble trying to overwrite while merging data sets together. My goal here was to make an initial pass through by labelling the continents by their majority hemisphere and there then make corrections by continent by overwriting with a new data set. This has proved to be tedious as I can merge and mutate the data set by adding the corrections but I cannot seem to overwrite the inital pass through. I have tried:

- An initial left join to get the majority hemisphere
- Attemped to left join the corrections to this but the values do not change, no effect elsewhere.
- Tried to do an anti join to remove values that I want to overwrite then re- left join to actually change them. I lose information and can't get it back.
- Will try basic merge and see if this allows overwritting - Result, unsuccessful
- My final attempt has been more of a jerryrigged solution. I created the extra column and then just did a ifelse statement to replace any values that I wanted. This only works in the binary case that I have here and would not work if there were more possibilities.

The final output is proof that this works as my intial pass would have labelled Zimbabwe, which is in Africa, as a northern hemisphere country. Here it is labelled at southern.