HW Lecture 3

Daniel Briseno Servin

9/21/2020

Problem 1

Merge the data frame from Task 5 with coords via three different techniques: inner, right, and left merge. Compare the dimensions of each of the resulting data frame and explain any differences you see.

```
#create dataframe from task 5
ccA = read.csv("./Data/countrycharsA.csv")
ccB = read.csv("./Data/countrycharsB.csv")
gdp = read.csv("./Data/gdp.csv")
ccAB = rbind(ccA,ccB)
ccAB_gdp = cbind(ccAB, gdp)

#load in coords dataframe
load("./Data/map.coords.RData")

#do inner, outer, left and right joins
countryData_inner = merge(ccAB_gdp, coords, by = "country")
countryData_outer = merge(ccAB_gdp, coords, by = "country", all = T)
countryData_left = merge(ccAB_gdp, coords, by = "country", all.x=T)
countryData_right = merge(ccAB_gdp, coords, by = "country", all.y=T)

#compare dimensions

dim(countryData_inner)
```

```
## [1] 1535    8
```

```
dim(countryData_outer)
```

```
## [1] 1799    8
```

```
dim(countryData_left)
```

```
## [1] 1677    8
```

```
dim(countryData_right)
```

```
## [1] 1657    8
```

Here we can see that the outer join had the most rows, the inner had the least, and the left and right joins have a very close number of rows, with the left being greater by 20 rows.

This is what one would expect, since:

- The outer join must include all entries from both tables, including the ones with no matching entry in the "countries" column. This means we can expect this one to have the most entries since it is the least exclusive

- The inner join must include only the entries from each table which can be matched by the "country" attribute. We can expect this join to have the least entries since it is the most exclusive
- The left join must include only the entries from `coords` which can be matched by "country" attribute to some entry in `ccAB_gdp`, and all entries of `ccAB_gdp`.
- The right join must include only the entries from `ccAB_gdp` which can be matched by "country" attribute to some entry in `coords`, and all entries of `coords`.

The small difference between the left join and the right join tells us that there are more entries in `ccAB_gdp` which cannot by matched by country than there are in `coords`.

Problem 2

Suppose a researcher has a hypothesis that there is a relationship between gdp and distance from the equator, i .e. latitude. To explore this, the researcher would like to break the gdp into quartiles and then look at the mean latitude (in absolute value) for each quartile. To do this, perform the following tasks:

A)

Using the data from the inner merge, use the quantile() function to determine the quartiles of gdp

```
q = quantile(countryData_inner$gdp)
q
```

```
##          0%         25%         50%         75%        100%
##    241.1659   1191.0260   3614.1013   9341.5210 113523.1329
```

B)

Using the findInterval() function and your result from above, create a new factor variable in the data set called gdp. q based on the quartile that the a given observation's gdp value lies in. Print a table of the levels of this variable. Does your function seem to be working? How can you tell?

```
interval = findInterval(countryData_inner$gdp, q, all.inside = T)
gdp.q = factor(interval)
gdp.q_levels = data.frame(levels(gdp.q))
colnames(gdp.q_levels) = c("Quartile")
kable(gdp.q_levels)
```

Quartile
1
2
3
4

The function does seem to be working, since we expect that with 5 quantiles the data will be partitioned into 4 quartiles, which we enforced by specifying `all.inside = T`. I believe that this worked, since we see 4 levels in the data which likely correspond to 4 quartiles.

C)

Reverse the order of the factor levels for gdp.q and reprint the table.

```r
gdp.q = factor(gdp.q, levels = rev(levels(gdp.q)))
gdp.q_levels = data.frame(levels(gdp.q))
colnames(gdp.q_levels) = c("Quartile")
kable(gdp.q_levels)
```

Quartile
4
3
2
1

D)

Find the mean latitude (in absolute value) for each quartile. Does there seem to be a difference?

```r
countryData_q = cbind(countryData_inner,"gdpQuartile" = gdp.q)

q1_latitude = mean(abs(countryData_q[countryData_q$gdpQuartile == 1, "lat"]))
q2_latitude = mean(abs(countryData_q[countryData_q$gdpQuartile == 2, "lat"]))
q3_latitude = mean(abs(countryData_q[countryData_q$gdpQuartile == 3, "lat"]))
q4_latitude = mean(abs(countryData_q[countryData_q$gdpQuartile == 4, "lat"]))

mean_latitudes = c("Q1" = q1_latitude, "Q2" = q2_latitude,"Q3" = q3_latitude, "Q4" = q4_latitude)
mean_latitudes
```

```
##       Q1       Q2       Q3       Q4
## 21.93251 27.35796 27.70101 23.33047
```

The mean latitudes do not show a significant difference between the latitudes of countries in different quartiles. Additionally, there is no clear correlation between gdp quartile and mean latitude.