Homework_2

Seper Abdi

2022-09-19

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ccA <- read.csv("countrycharsA-1.csv")
ccB <- read.csv("countrycharsB-1.csv")
gdb_1 <- read.csv("gdp-1.csv")
```

1. Use the *str()*command to examine the characteristics of each of the four databases. Then append ccA to ccB to make one large data frame and reexamine the output with *str()*.

```
str(ccB)
```

```
## 'data.frame':    921 obs. of  5 variables:
##  $ country  : chr  "Algeria" "Algeria" "Algeria" "Algeria" ...
##  $ continent: chr  "Africa" "Africa" "Africa" "Africa" ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  43.1 45.7 48.3 51.4 54.5 ...
##  $ pop      : int  9279525 10270856 11000948 12760499 14760787 17152804 20033753 23254956 26298373 29
```

– append ccA to ccB to make one large data frame

```
combineddataset <- rbind(ccA, ccB)
str(combineddataset)
```

```
## 'data.frame':    1677 obs. of  5 variables:
##  $ country  : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ continent: chr  "Asia" "Asia" "Asia" "Asia" ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
```

2. Append the *gdp* data to the data frame you created in Number 1. Be careful here and make sure you examine the structure of the resulting data frame to ensure things are correct. Examine the output with *str()*.

```
combineddataset2 <- cbind(combineddataset, gdb_1)
str(combineddataset2)
```

```
## 'data.frame':    1677 obs. of  6 variables:
##  $ country  : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ continent: chr  "Asia" "Asia" "Asia" "Asia" ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdp      : num  779 821 853 836 740 ...
```

3. How many times after 1980 did a country have gdp<20000?

– it will work select gdp more thatn 20k

```
srot_by_gdp <- (combineddataset2[order(combineddataset2$gdp, decreasing = FALSE), ])
identify_rows <-(combineddataset2 $ gdp >= 20000)
gdp_over_20k <- combineddataset2[identify_rows, ]
print("Count of repeated values")
```

```
## [1] "Count of repeated values"
```

```
length(which(gdp_over_20k$gdp >= 20000))
```

```
## [1] 160
```

4. Merge your data frame from 1 &2 above with the *coords* data 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.

```
left_merg = merge(x=combineddataset,y=combineddataset2,by = "continent",all.x=TRUE)

inner_right= combineddataset %>% inner_join(combineddataset2,by="lifeExp")
```

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

```
gdp.q <- quantile(inner_right$gdp)
gdp.q
```

```
##          0%         25%         50%         75%        100%
##    241.1659   1211.0166   3528.4813   9311.1788 113523.1329
```

5 b.) Create a new variable in the data set called gdp.q based on the quartile that the a given observation's gdp value lies in. Print a frequency table of this new variable and explain how you can use it to demonstrate that your procedure is working properly. (The *findInterval()* function might be helpful).

- I have problem with this part I cannot understanding whats the issues ? Could you please help me.

```
#data1 <- data.frame(combineddataset, lower = findInterval(combineddataset2, vec = NULL))
#data1
```

5 c.) Using dplyr's group_by and summarize functioins, find the mean absolute value of the latitude for each quartile. Based on your findings, draw a conclusion about the relationship between GDP

```
mean_1 <- mean(combineddataset2$ gdp, na.rm = TRUE)
sd_1 <-sd(combineddataset2$ gdp, na.rm = TRUE)
sd_1
```

```
## [1] 9884.326
```

```
mean_1
```

```
## [1] 7216.043
```

6. (Bonus) Create a new data set that, for each year, lists the median gdp and also identifies the country whose gdp is nearest to the median.

for the assignment no.6 I just could'nt add the country to the data frame, I have to get some help for this part

```
# Create an Empty DataFrame
df = data.frame(combineddataset2[, c('year', 'gdp')])
summary(df)
```

```
##       year           gdp
##  Min.   :1952   Min.   :   241.2
##  1st Qu.:1962   1st Qu.:  1202.2
##  Median :1977   Median :  3528.5
##  Mean   :1979   Mean   :  7216.0
##  3rd Qu.:1997   3rd Qu.:  9313.9
##  Max.   :2007   Max.   :113523.1
```

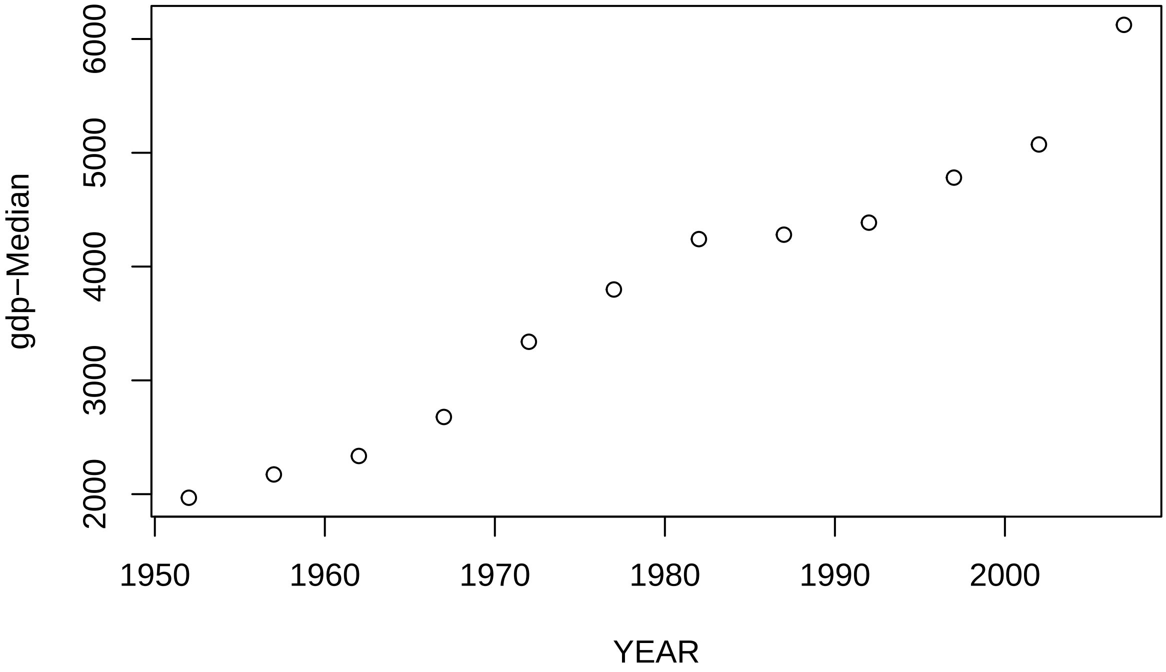
```
# Median of the column by group
df_median <- aggregate(x=df$gdp,by = list(df$year),FUN=median)
colnames(df_median)
```

```
## [1] "Group.1" "x"
```

```
names(df_median)[names(df_median) == "Group.1"] <- "YEAR"
names(df_median)[names(df_median) == "x"] <- "gdp-Median"
df_median
```

```
##    YEAR gdp-Median
## 1  1952   1968.528
## 2  1957   2173.220
## 3  1962   2335.440
## 4  1967   2678.335
## 5  1972   3339.129
## 6  1977   3798.609
## 7  1982   4241.356
## 8  1987   4280.300
## 9  1992   4386.086
## 10 1997   4781.825
## 11 2002   5073.194
## 12 2007   6124.371
```

```
plot(df_median)
```


gdp−Median
2000
3000
4000
5000
6000
1950
1960
1970
1980
1990
2000
YEAR