Quantiles and facators in R

Ramses

Introduction

Let's revisit several functions or concepts we've learnt from lectures and QSS tutorials that help you succeed in finishing Problem Set 1. We'll cover:

- `tapply()`
- `quantile()`
- `ifelse()` or `case_when`
- class `factor`

Load the gapminder data again

```
# load data

data <- read.csv("data/gapminder.csv")
```

`tapply` for group means

Using the `tapply` function, please find:

1. The average `gdpPercap` for each `continent`
2. The average `gdpPercap` for each `year`
3. The average `gdpPercap` for each `continent` over every `year`

```
# 1.

tapply(data$gdpPercap, data$continent, mean)
```

```
##    Africa  Americas      Asia    Europe   Oceania
##  2193.755  7136.110  7902.150 14469.476 18621.609
```

```
# 2.

tapply(data$gdpPercap, data$year, mean)
```

```
##      1952      1957      1962      1967      1972      1977      1982      1987
##  3725.276  4299.408  4725.812  5483.653  6770.083  7313.166  7518.902  7900.920
##      1992      1997      2002      2007
##  8158.609  9090.175  9917.848 11680.072
```

```
# 3. in here you will need to use list()

tapply(data$gdpPercap, list(data$year,
                            data$continent), mean)
```

```
##        Africa  Americas      Asia    Europe  Oceania
## 1952 1252.572  4079.063  5195.484  5661.057 10298.09
## 1957 1385.236  4616.044  5787.733  6963.013 11598.52
## 1962 1598.079  4901.542  5729.370  8365.487 12696.45
## 1967 2050.364  5668.253  5971.173 10143.824 14495.02
## 1972 2339.616  6491.334  8187.469 12479.575 16417.33
## 1977 2585.939  7352.007  7791.314 14283.979 17283.96
## 1982 2481.593  7506.737  7434.135 15617.897 18554.71
## 1987 2282.669  7793.400  7608.227 17214.311 20448.04
## 1992 2281.810  8044.934  8639.690 17061.568 20894.05
## 1997 2378.760  8889.301  9834.093 19076.782 24024.18
## 2002 2599.385  9287.677 10174.090 21711.732 26938.78
## 2007 3089.033 11003.032 12473.027 25054.482 29810.19
```

quantile and ifelse

Using `quantile` and `ifelse` function, please create:

1. An object `gdp_qt` that records the lower quartile, median, and upper quartile of `gdpPercap` variable
2. A new variable `poverty` that takes the value of 1 if `gdpPercap` is lower than or equal to the lower quartile; 0 otherwise. What is the `sum` of countries in poverty? And their proportion?
3. A new variable `gdpPercap_cat` that converts `gdpPercap` into four categories: `poor`, `middle`, `wealthy`, and `very wealthy` based on quartiles in `gdp_qt`
4. Use `tapply` to find the mean of `lifeExp` for each income group, based on `gdpPercap_cat`

```
# 1.
gdp_qt <- quantile(data$gdpPercap)

gdp_qt
```

```
##          0%         25%         50%         75%        100%
##    241.1659   1202.0603   3531.8470   9325.4623 113523.1329
```

```
# 2.
data$poverty <- ifelse(data$gdpPercap <= gdp_qt[2], 1, 0)

table(data$poverty)
```

```
##
##    0    1
## 1278  426
```

```
sum(data$poverty)
```

```
## [1] 426
```

```
mean(data$poverty)
```

```
## [1] 0.25
```

```
# 3. with nested ifelse

data$gdpPercap_cat <-
ifelse(data$gdpPercap <= gdp_qt[2], "poor",
       ifelse(data$gdpPercap > gdp_qt[2] & data$gdpPercap <= gdp_qt[3], "middle",
              ifelse(data$gdpPercap > gdp_qt[3] & data$gdpPercap <= gdp_qt[4], "wealthy",
                     ifelse(data$gdpPercap > gdp_qt[4], "very wealthy", NA
                            )
                     )
              )
       )

# 3. with nested ifelse

data$gdpPercap_cat <- case_when(data$gdpPercap <= gdp_qt[2] ~ "poor",

                                data$gdpPercap > gdp_qt[2] &
                                  data$gdpPercap <= gdp_qt[3] ~ "middle",

                                data$gdpPercap > gdp_qt[3] &
                                  data$gdpPercap <= gdp_qt[4] ~ "wealthy",

                                data$gdpPercap > gdp_qt[4] ~"very wealthy")


# 4.

tapply(data$lifeExp, data$gdpPercap_cat, mean)
```

```
##       middle         poor very wealthy      wealthy
##     54.04259     45.99939     72.67556     65.18023
```

Factor

How to inform `R` that `gdpPercap_cat` has an inherent order?

1. Check out the class of `gdpPercap_cat`

2. Use `factor()` to convert `gdpPercap_cat` into factor, and specify the `levels = c(...)` argument. In the levels argument you will concatenate the four categories `poor`, `middle`, `wealthy`, and `very wealthy` in this order.
3. Check out the class of `gdpPercap_cat` again
4. Use `tapply` to find the mean of `lifeExp` for each income group, based on `gdpPercap_cat`

```r
# look at the class of gdpPercap_cat
class(data$gdpPercap_cat)
```

```
## [1] "character"
```

```r
# Turn it into a factor with ordered levels
data$gdpPercap_cat <- factor(data$gdpPercap_cat,
                             levels = c("poor", "middle", "wealthy", "very wealthy"))

class(data$gdpPercap_cat)
```

```
## [1] "factor"
```

```r
# Look at the conditional mean of life expectancy by income group
tapply(data$lifeExp, data$gdpPercap_cat, mean)
```

```
##         poor       middle      wealthy very wealthy
##     45.99939     54.04259     65.18023     72.67556
```

```r
# Look at the conditional standard deviation of life expectancy by income group
tapply(data$lifeExp, data$gdpPercap_cat, sd)
```

```
##         poor       middle      wealthy very wealthy
##     7.681382     9.121956     8.227916     6.307244
```

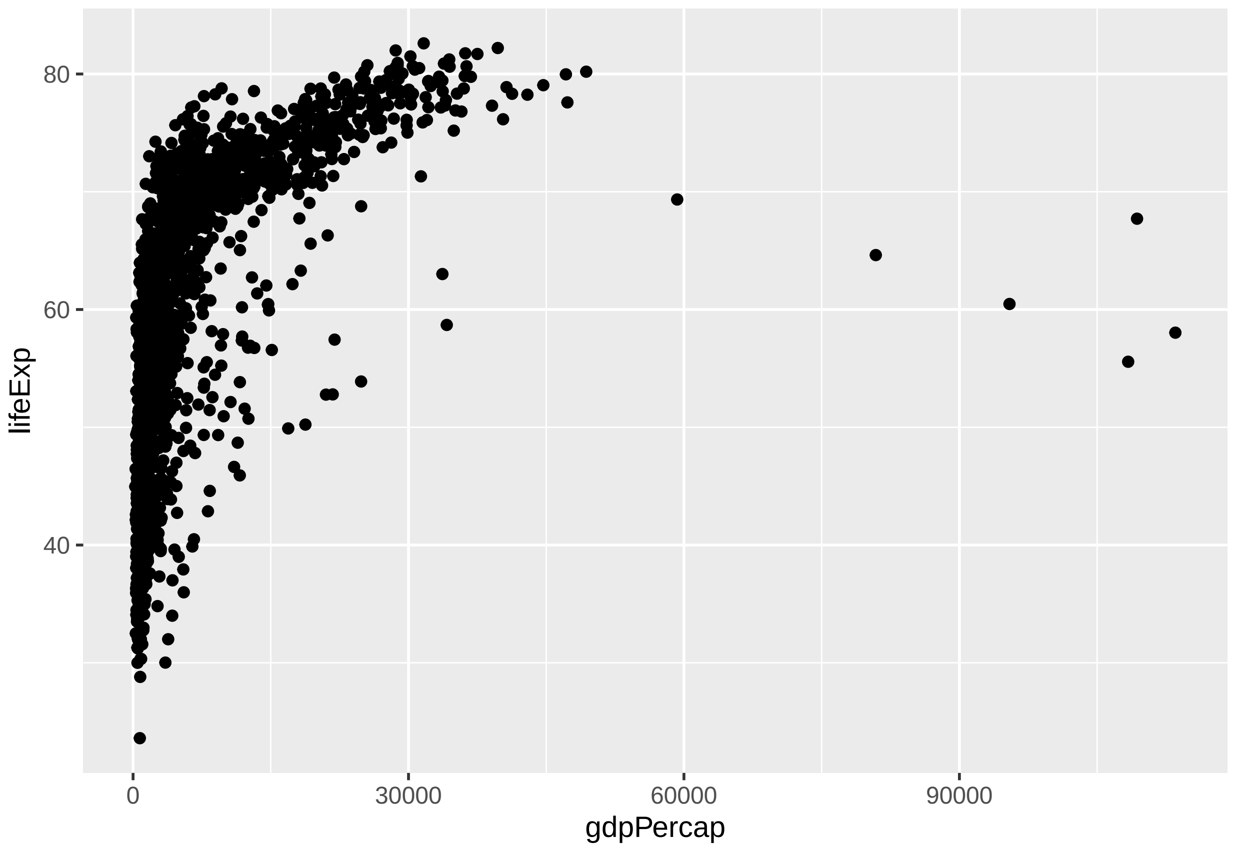
Intro to ggplot

Using `ggplot`, create two visualizations:

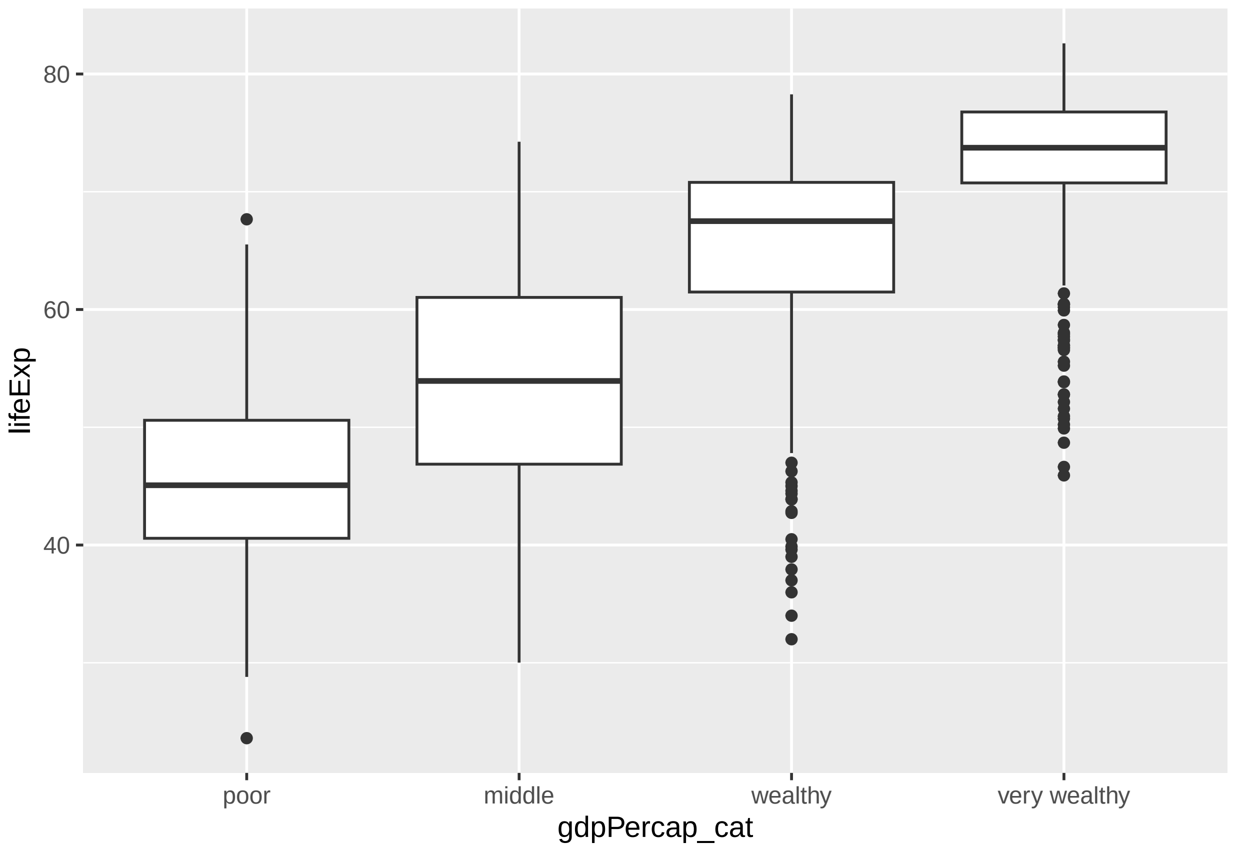
- A scatter plot between life expectancy (`lifeExp`) and income (`gdpPercap_cat`).
- A boxplot between life expectancy (`lifeExp`) and the categorical variable of income (`gdpPercap_cat`).

What do you observe in terms of associations and dispersion of the distributions? Remember that you will need to load either the library of `ggplot2` or `tidyverse`.

```r
# create a scatter plot
ggplot(data=data,
       aes(x=gdpPercap,
           y=lifeExp)) +
  geom_point()
```



```
# create a boxplot
ggplot(data=data,
       aes(x=gdpPercap_cat,
           y=lifeExp)) +
  geom_boxplot()
```


80
60
40
lifeExp
poor
middle
wealthy
very wealthy
gdpPercap_cat