Lab 1 assignment

Yan Gao

2022-11-04

Question 1

In this lab, you will be working with two datasets: hsbc_basic.csv and hsbc_health.txt — both originating from a survey dataset, HSBC,containing information about the health of sample of Swedish students in 2014. Your first task is to import them to R.

```
hsbc_basic <- read.csv(file =
"C:/Users/46765/OneDrive/Desktop/statistic 2/lab lecture/hsbc-basic.csv", header = TRUE )
hsbc_health <- read.table(file =
"C:/Users/46765/OneDrive/Desktop/statistic 2/lab lecture/hsbc-health.txt", header = TRUE)
```

Question 2

After importing, your next task is to present the following basic information about the two datasets. a. The number of rows and columns

```
nrow(hsbc_basic)
```

```
## [1] 2000
```

```
ncol(hsbc_basic)
```

```
## [1] 3
```

```
nrow(hsbc_health)
```

```
## [1] 1500
```

```
nrow(hsbc_health)
```

```
## [1] 1500
```

b. The number of numeric, integer, character, and Factor variables.

```
str(hsbc_basic)
```

```
## 'data.frame':    2000 obs. of  3 variables:
##  $ id4   : int  3303 7443 1297 4775 1906 1149 2749 2391 3773 4944 ...
##  $ sex   : chr  "Boy" "Girl" "Girl" "Boy" ...
##  $ AGECAT: int  13 15 11 13 11 11 13 11 13 13 ...
```

```
str(hsbc_health)
```

```
## 'data.frame':    1500 obs. of  4 variables:
##  $ id4         : int  2 27 35 37 38 39 40 41 49 60 ...
##  $ bully_dummy : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ health_index: int  8 8 8 6 6 7 7 6 10 6 ...
##  $ lifesat     : num  6.9 9.84 6.81 9.91 9.62 ...
```

Question 3

Your task is now to perform an inner-join to merge the two datasets, using id4 as they key (Hint: use the merge() function).

```
hsbc <- merge(x = hsbc_basic, y = hsbc_health, by = "id4", all = FALSE)
head(hsbc, n = 5)
```

```
##   id4  sex AGECAT bully_dummy health_index  lifesat
## 1   2  Boy     11           0            8 6.902535
## 2  27  Boy     11           0            8 9.844015
## 3  35  Boy     11           0            8 6.805916
## 4  37 Girl     11           0            6 9.909915
## 5  38 Girl     11           0            6 9.617146
```

Explain why it has the number of rows it has

Answer: Inner-joins only keeps cases that exists in both datasets.After merging hsbc_health and hsbc_basic, the number of rows should be 1500, and the number of columns is 6

Question 4

Next up is data cleaning! Specifically, you are to investigate whether there are any rows of hsbc that contain missing values. If you find such instances, state which column(s) that are affected, and then filter out the rows with missing data.

```
hsbc[!complete.cases(hsbc),]
```

```
##       id4  sex AGECAT bully_dummy health_index lifesat
## 276  1564  Boy     11           1            6      NA
## 509  2847 Girl     13           0            7      NA
## 603  3364  Boy     13           0            9      NA
## 635  3573 Girl     13           1            6      NA
## 724  3964  Boy     13           0            9      NA
## 957  5171 Girl     15           0            8      NA
## 1078 5818 Girl     15           0            5      NA
```

```
## 1123 6049  Boy     15           0            8      NA
## 1290 6857 Girl     15           0            7      NA
## 1305 6906  Boy     15           0            4      NA
```

```
hsbc <- hsbc[complete.cases(hsbc),]
nrow(hsbc)
```

```
## [1] 1490
```

Question 5

Once you have ensured that hsbc does not contain any missing values, your next task is to produce a set of variable-level summaries. Specficially, report: a. The average life satisfaction (lifesat). (Hint: use the mean() function) b. The total number of observations in each age-category (AGECAT). (Hint: use the table() function).Which age-category have the most observations?

```
mean(hsbc$lifesat)
```

```
## [1] 7.344637
```

```
table(hsbc$AGECAT)
```

```
##
##  11  13  15
## 473 443 574
```

Answer: 15

Question 6

Building on 5b, examine which age-category (AGECAT) that have the highest recorded number of bullied kids (bully_dummy==1). (Hint: you may again use the table() function).

```
table(hsbc$AGECAT, hsbc$bully_dummy ==1)
```

```
##
##      FALSE TRUE
##   11   397   76
##   13   390   53
##   15   529   45
```

Answer: 11 age-category

Question 7

Next, you are to perform a counting exercise that involves both continuous and categorical variables simultaneously. Use conditional subsetting to report the following

a. How many bullied kids (bully_dummy==1) there are with a lifesat score lower than 7

```
nrow(hsbc[hsbc$lifesat < 7 & hsbc$bully_dummy == 1,])
```

```
## [1] 95
```

b. How many girls (sex==Girl) there are in age-category 13 (AGECAT==13) that have a lifesat score greater than 8.

```
nrow(hsbc[hsbc$lifesat > 8 & hsbc$AGECAT == 13 & hsbc$sex == "Girl",])
```

```
## [1] 77
```

##Question 8 Create a new column in hsbc that is set to 1 if health_index is greater than or equal to 7, and set to 0 otherwise. Call the new column health_index_binary. (Hint: use ifelse())

```
hsbc$health_index_binary <- ifelse(test = hsbc$health_index >= 7, yes =1, no = 0)
```

for check

```
head(hsbc, n=5)
```

```
##   id4  sex AGECAT bully_dummy health_index  lifesat health_index_binary
## 1   2  Boy     11           0            8 6.902535                   1
## 2  27  Boy     11           0            8 9.844015                   1
## 3  35  Boy     11           0            8 6.805916                   1
## 4  37 Girl     11           0            6 9.909915                   0
## 5  38 Girl     11           0            6 9.617146                   0
```

Question 9

Compute the conditional mean of lifesat given the two different statuses of health_index_binary (0/1). For which out of the two do you find the highest average life satisfaction?

```
aggregate(x=hsbc$lifesat,by = list(hsbc$health_index_binary), FUN = mean)
```

```
##   Group.1        x
## 1       0 6.264939
## 2       1 7.817786
```

#ansewr: #the health_index greater and equal than 7 shows higher average life satisfaction.

Question 10

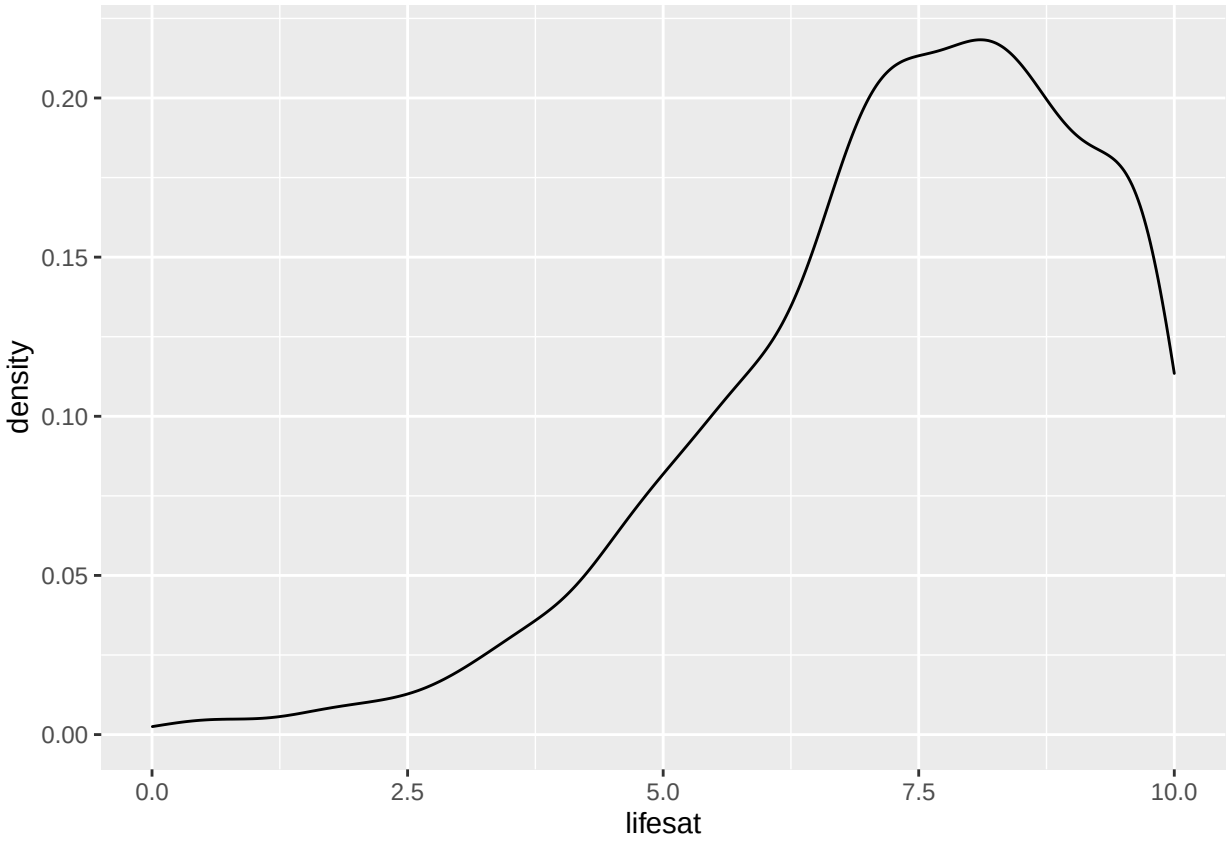
Next up is plotting! As preliminaries, first, load the ggplot2 package. Second, format the variable health_index_binary as a Factor (Hint: using the factor() function). The latter step is performed to make ggplot2 aware that health_index_binary is a discrete variable, and not a continuous one.

```
library(ggplot2)
hsbc$health_index_binary <- factor(hsbc$health_index_binary,levels = c(1,2))
```

Question 11

Construct a density plot of lifesat (Hint: use geom_density()). How would you characterize its distribution?

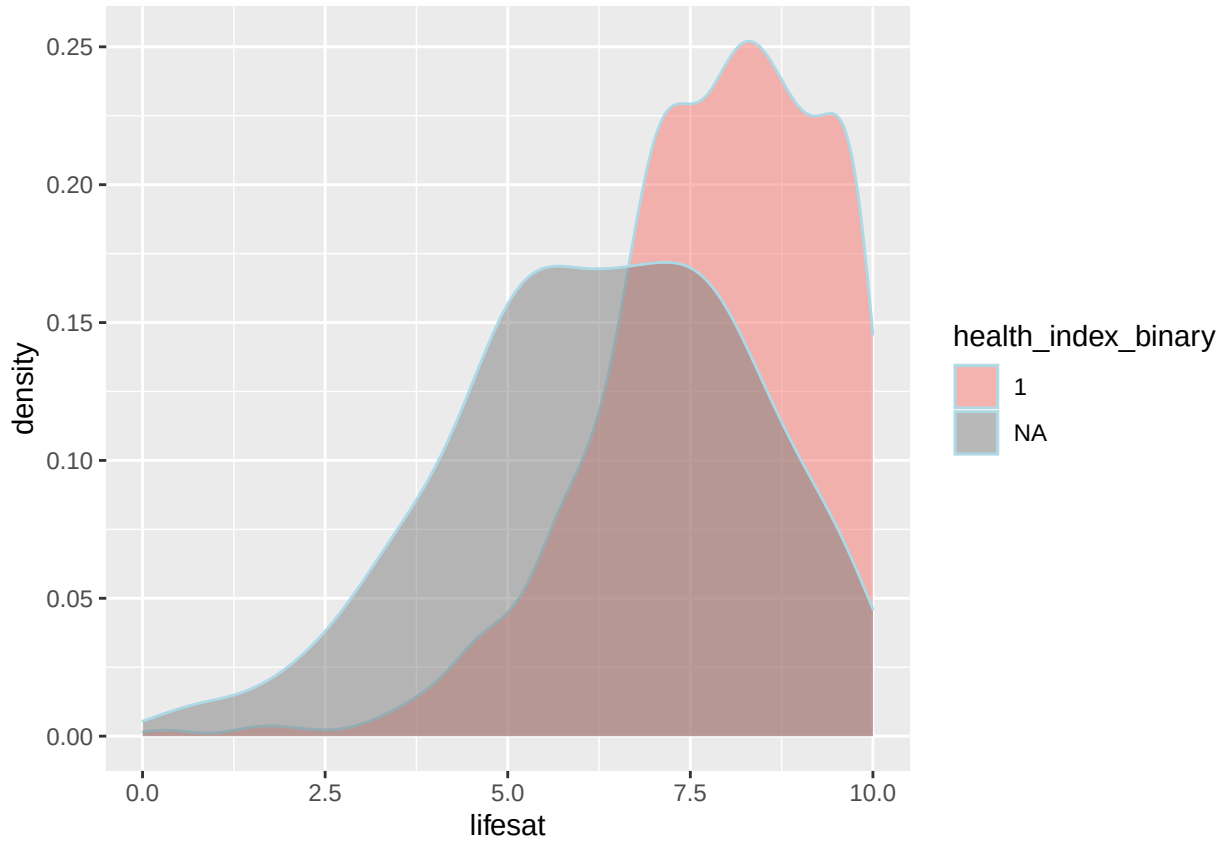
```
ggplot(hsbc,aes(x= lifesat)) + geom_density()
```



Question 12

Extend the plot in 11 by colouring the distribution based on the membership to either of the health_index_binary categories (0/1).

```
ggplot(hsbc,aes(x= lifesat,fill = health_index_binary)) +
        geom_density(color = "lightblue", alpha = 0.5)
```



Question 13

As a final task, export hsbc to your hard-drive (where exactly, you decide). You may export it either as .txt or .csv.

```
write.csv(x= hsbc, file =
"C:/Users/46765/OneDrive/Desktop/statistic 2/lab lecture/hsbc.csv",row.names= FALSE)
write.table(x= hsbc, file =
"C:/Users/46765/OneDrive/Desktop/statistic 2/lab lecture/hsbc.txt",row.names= FALSE)
```