Case study

Introduction to R

EBdata\_Hungary

Mengfei Cai

P282929

[m.cai@umcg.nl](mailto:m.cai@umcg.nl)

Formal evaluation is not needed

**1. Importing and Cleaning Your Dataset.**

**1 a,b,c : dataset processing**

* Set working directory and load Haven
* Select the subset, country of interest: **Hungary**
* Rename variables as required

**1d : describle rough data**

Observations: 1044,

Variables: 13

**Table 1: Table of NA in each variavle**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| alc\_12m | alcfreq\_5dr | alc\_30da | alcfreq\_30d | alc\_am\_dd | pa7d\_work | pa7d\_mov | pa7d\_recr |
| 7 | 397 | 397 | 515 | 512 | 89 | 7 | 1 |

**1e & 1f : coverstion and droplevels**

use the ***unique( )*** function to check categorical values

the function ***data\_converter*** is created to convert *categorical ones into factors and drop the levels*

**1g: descriptive statistics of alc\_12m and pa7d\_work**

***Table 2: Descriptive statistics of the variable alc\_12***

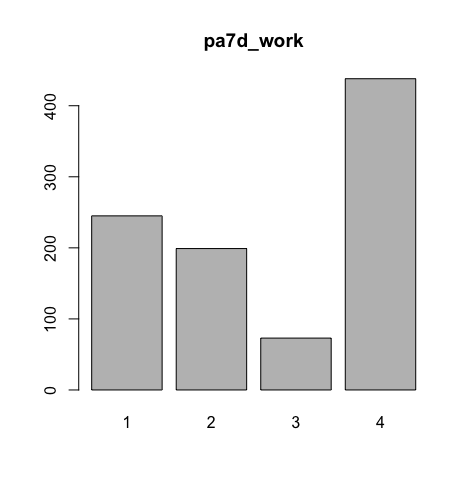
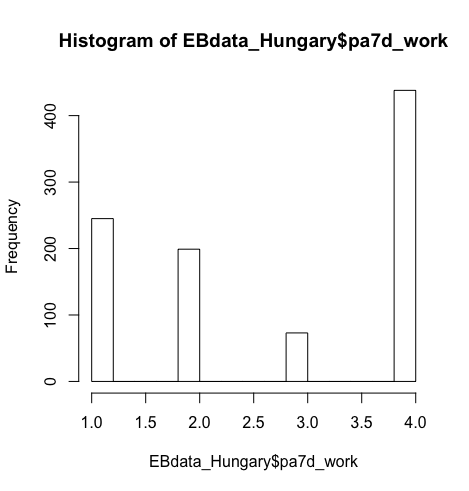
|  |  |  |
| --- | --- | --- |
| Yes | No | NA |
| 647 | 390 | 7 |

***Table 3: Descriptive statistics of the variable pa7d\_work***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A lot | Some | Little | None | NA |
| 245 | 199 | 73 | 438 | 89 |

**#1h: histogram and bar chart of pa7d\_work**

Apparently, bar chart is more suitable to present these data.



**2. Problematic Alcohol Use**

**2a:** **number of drinking days/week**

In terms of drinking days per week,

assumption are as follows:

"once"=0.25,"once a week"=1,

"2 - 3 times a month"=0.5,

"Daily"=7, "4 - 5 times a week"=4,

"2 - 3 times a week"=2

**2b: quantity of alcohol consumed on a drinking day**

If one drinks equal 12g alcohol, accordingly, the assumptions are as follows:

"Less than 1 drink"=6,"1 - 2 drinks"=18,

"3 - 4 drinks"=42,"5 - 6 drinks"=66,

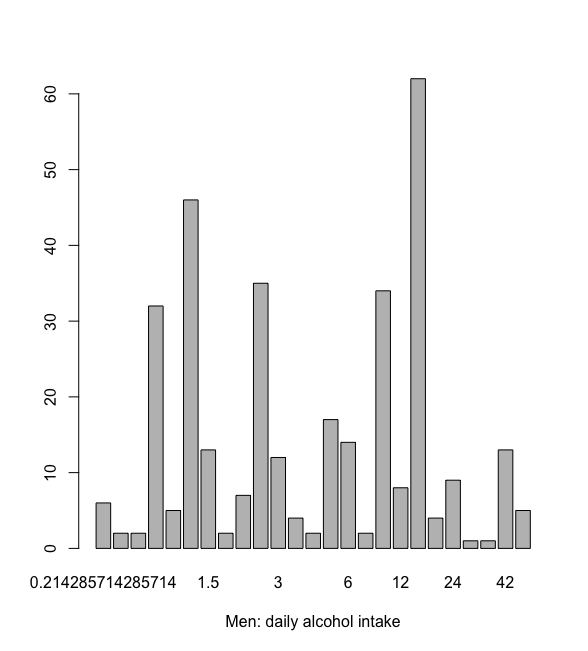
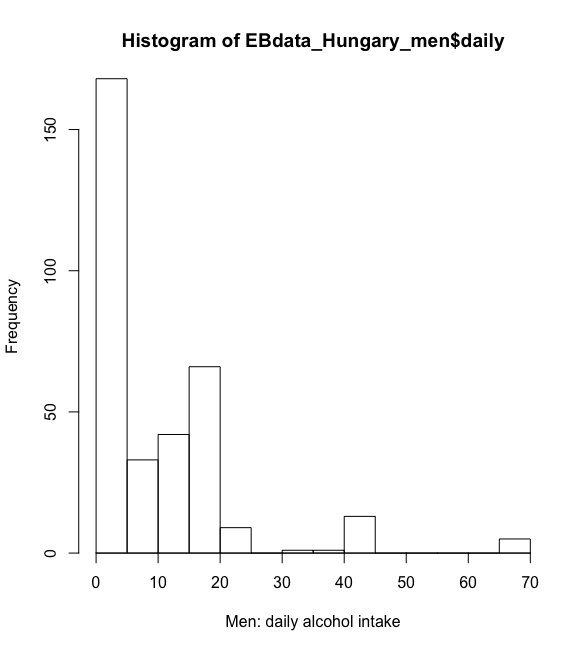
"7 - 9 drinks"= 8," 10 drinks or more"=120,

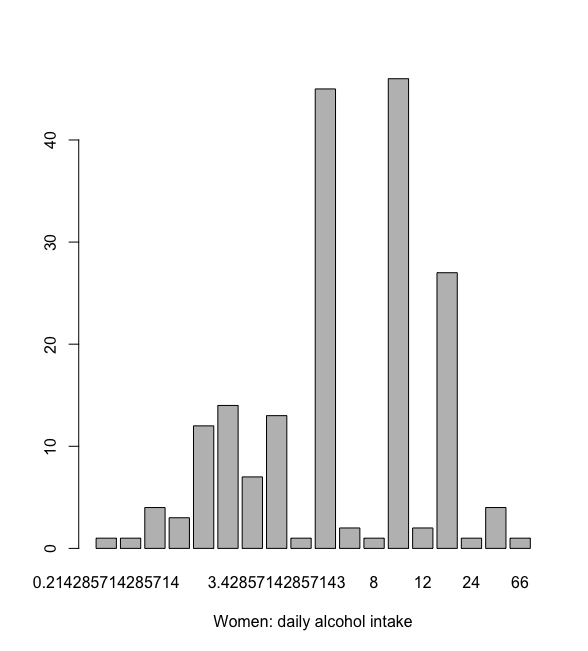
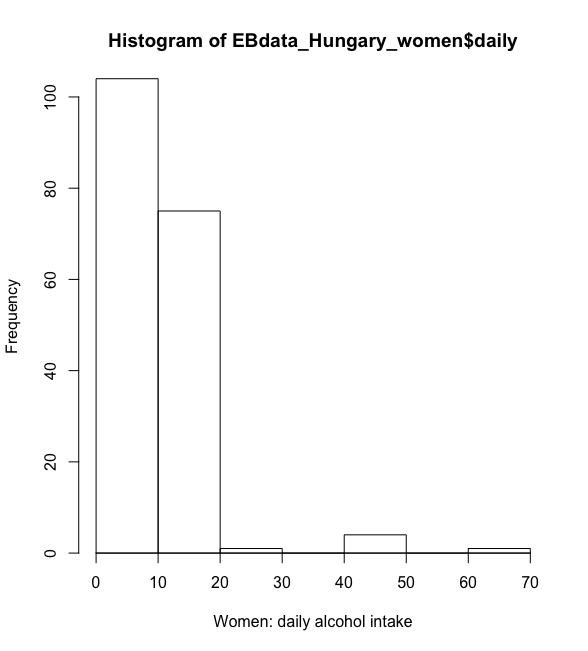
**2c: average quantity of alcohol per week**

*Multiplying the two variables above: i.e****.***

*(quantity of alcohol per day) \* (*number of drinking days per week)

**2d: histogram / bar plot of daily intake.**





**2e: heavy drinker” for men and women and set cut-off**

alcohol daily intake = quant per week / 7

**2f: Define a new variable “binge”**

**2g: problem drinker**

Choose the logistic operators |, meaning “OR”.

**2h: Check the contents of this new variable: problem\_drinker**

Yes, the variable makes sense due to it meets the given requirements.

**2i: define a new variable total physical activity and categorization of physical activity**

***as.numeric*** was used to convert these factors (pa7d\_\*) to numeric values before doing the calculation.

***Ifelse*** was used to categorize the people as “inactive, active, moderately active”.

**3. Analyzing the data**

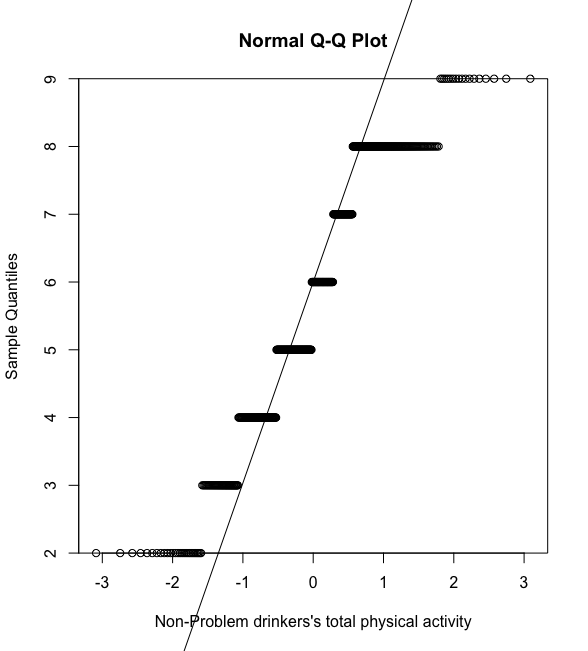
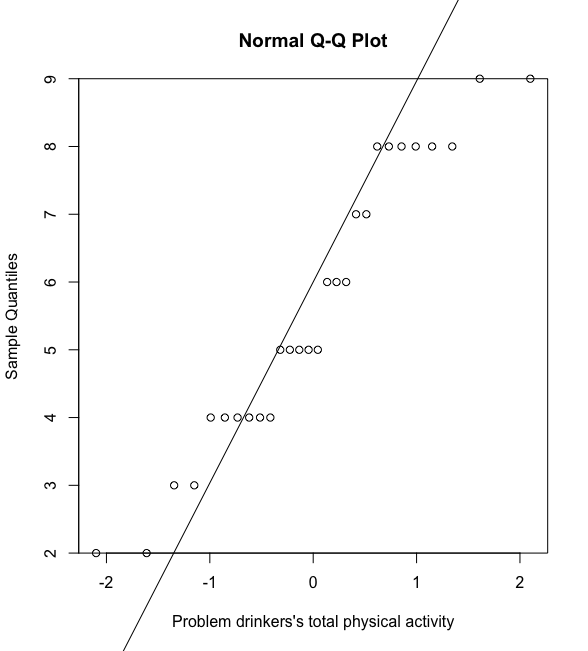
3a: Print a table contrasting the variable on physical activity VS problem drinker and non-problem drinker

Table 4: Physical activity in problem drinker and non-problem drinker

|  |  |  |
| --- | --- | --- |
| Problem\_drinker | NO | YES |
| Active | 28 | 2 |
| Inactive | 193 | 10 |
| Mod. active | 276 | 16 |

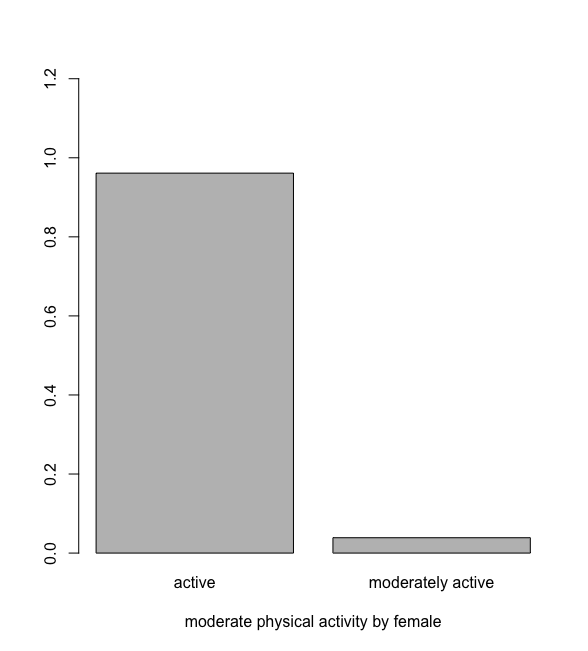
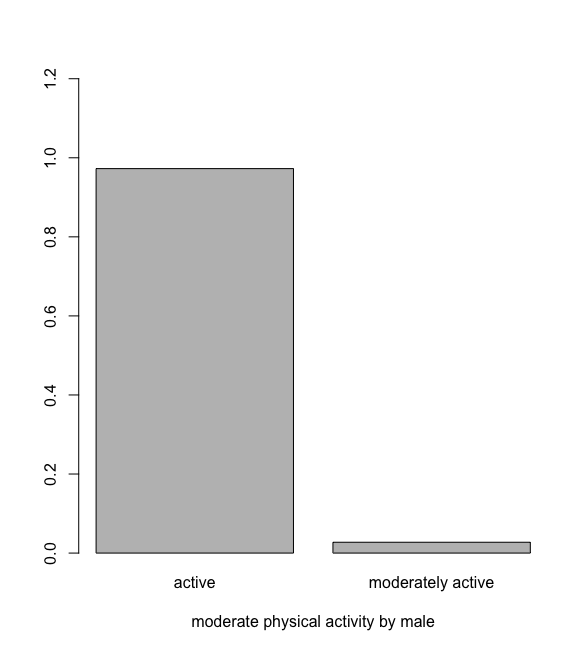
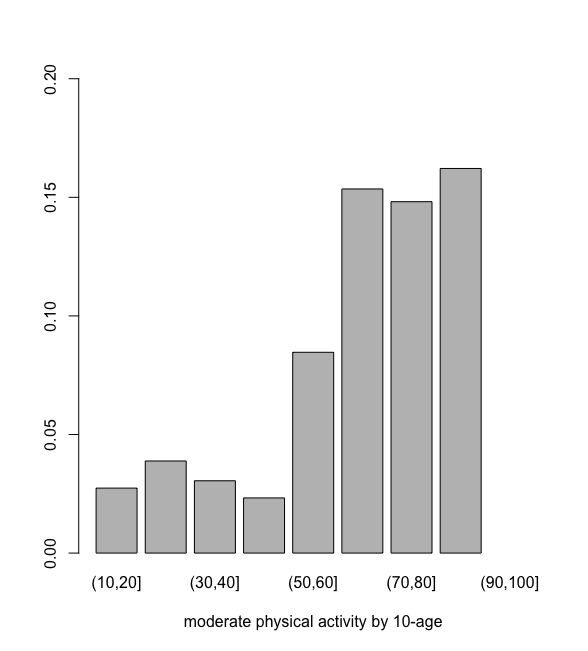
Before choosing some tests to check the difference, it is of great significance to check if these data are normally distributed. Therefore, the ***qqnorm()*** and ***qqline()*** was used. These data are not distributed along the given line and the

non-parametrical approach ***wilcox.test()*** was chosen to compare the difference of total physical acitivities between problem drinkers and non-problems drinkers.



After using ***wilcox.test( ),*** the p-value is 0.7619, indicating that there is no significant difference between these two groups of alcohol drinkers.

**3b: Make a graph of percentages with at least moderate physical activity**



**3c: daily alcohol use in grams and check if normal distribution**

By using QQplot and QQline, daily alcohol use is not normally distributed.



**3d: regress total daily alcohol use on age, gender, and physical activity**

By using the logistic regression ***lm( ),*** the P-value for age, gender, total physical activity is 1.406e-09, 0.3236, 1.317e-05 respectively, indicating that there is no significant correlation between gender and daily alcohol use, but the significant correlation can be found in terms of age and physical activity.

**3e+3f: regression analysis**

To be honest, it is difficult for me to interpret the results of regression analysis. To better understand the regression of daily alcohol intake to the addition of age plus total physical alcohol, or gender plus total physical alcohol, the further learning of statistical background is greatly needed.

**4. The uncounted problems**

* The unfamiliarity with the specific function： The first eight-day exercises are greatly helpful for me get a better understanding of the detail and the meaning of some important functions. Although I can understand most of the codes, I cannot make it alone. Therefore, I have searched for some functions and examples in these previous exercises.
* The coding of specific function: For the case study, some new functions, like ***switch( )*** are advised to use to achieve the main aim. Difficult as it is at first sight, I finally made it by browsing through the webpages on ***stackoverflow***, after some struggles.