Homework 3: Data Cleaning and Management

Directions

Launch RStudio and open a new RMarkdown file or use the class RMarkdown template provided and save it on your working directory as a .Rmd file. At the end of the activity, save your **pdf** generated from `RMarkdown+Knitr` and submit it in the Blackboard.

Show all your work. Late submission will attract a penalty of 10 points per day after the due date.

If you have questions, please post them on the lesson discussion board.

1. (a) Clean up the workspace using the `rm()` function. Use the `data()` function to display the built-in datasets you can access. Use the R help to learn more about the `‘longley’` dataset: `?longley`.

   (b) Print only the records in the `‘longley’` dataset that are from the years `1947-1950`: longley[longley$Year==1947:1950,]. `attach(longley)`.

   (c) `plot(Unemployed ∼ Year)`.

   (d) Change the type of plot to a line: `plot(Unemployed ∼ Year, type ="l")`

2. You track your commute times for two weeks and record the following (in minutes):`17 16 20 24 22 15 21 15 17 22`.

   (a) Enter these numbers into R and find the 5-number summary.

   (b) You find a data entry error, the number 24 should have been 18. Using R, replace the incorrect value without reentering the entire set of data and find the new 5-number summary.

   (c) Use R to count the number of times your commute was at least 20 minutes.

   (d) Use R to calculate the percent of your commutes that were less than 17 minutes.

3. Using the `maltreat.dta` dataset, explore the variable `ethnic` using `tab1(ethnic)`. There are spelling mistakes that need to be corrected. Correct mis-spelt names, and create a numeric, categorical variable `ethncity`. The “Jola” cleaning code for part (i) has been provided. Finish the remaining part of the code and produce the final (clean) bar chart.

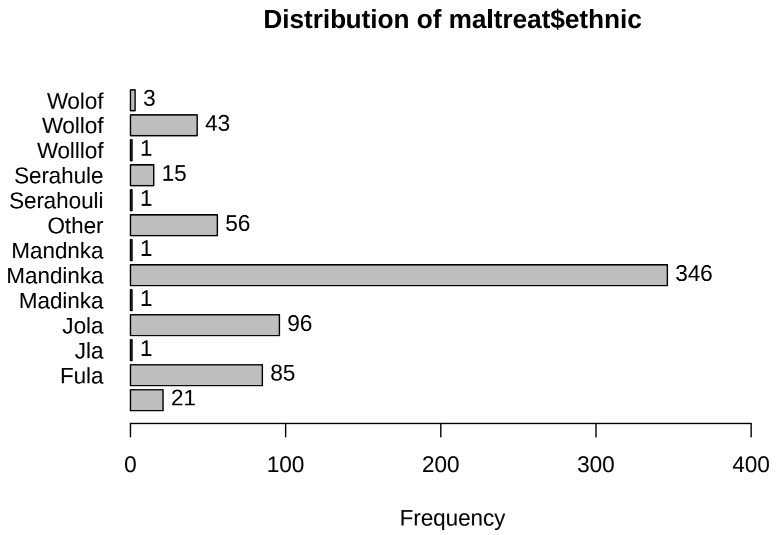
   (i) Replace ethnic = “Jola” if ethnic value starts with a “J”.

   (ii) Replace ethnic = “Mandinka” if ethnic value starts with an “M”

   (iii) Replace ethnic = “Serahule” if ethnic value starts with an “S”

(iv) Replace ethnic = “Wollof” if ethnic value starts with a “W”

```
library("readstata13")
maltreat <- read.dta13("data/maltreat.dta")
# Original ethnic (string) variable
tab1(maltreat$ethnic, col = "grey")
# convert it to a new factor variable ethnicity
maltreat$ethnicity <- as.factor(maltreat$ethnic)
# explore the levels (unclean)
levels(maltreat$ethnicity)
# clean up for Jola
levels(maltreat$ethnicity)[startsWith(levels(maltreat$ethnicity),
    "J")] <- "Jola"
```


Distribution of maltreat$ethnic
Wolof
Wollof
Wolllof
Serahule
Serahouli
Other
Mandnka
Mandinka
Madinka
Jola
Jla
Fula
Frequency