Homework-3

Madhu Peduri

6/17/2021

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.

We use rm() function to remove objects to control the usage of memory. We use data() to load the 'longley' dataset and head() to see the sample data.

```
rm(list = ls(all=TRUE))
ldata = longley
head(ldata)
```

```
##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
## 1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
## 1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
## 1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
## 1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
## 1951         96.2 328.975      209.9        309.9    112.075 1951   63.221
## 1952         98.1 346.999      193.2        359.4    113.270 1952   63.639
```

```
class(ldata)
```

```
## [1] "data.frame"
```

```
names(ldata)
```

```
## [1] "GNP.deflator" "GNP"          "Unemployed"   "Armed.Forces" "Population"
## [6] "Year"         "Employed"
```

1.(b) Print only the records in the 'longley' dataset that are from the years 1947-1950

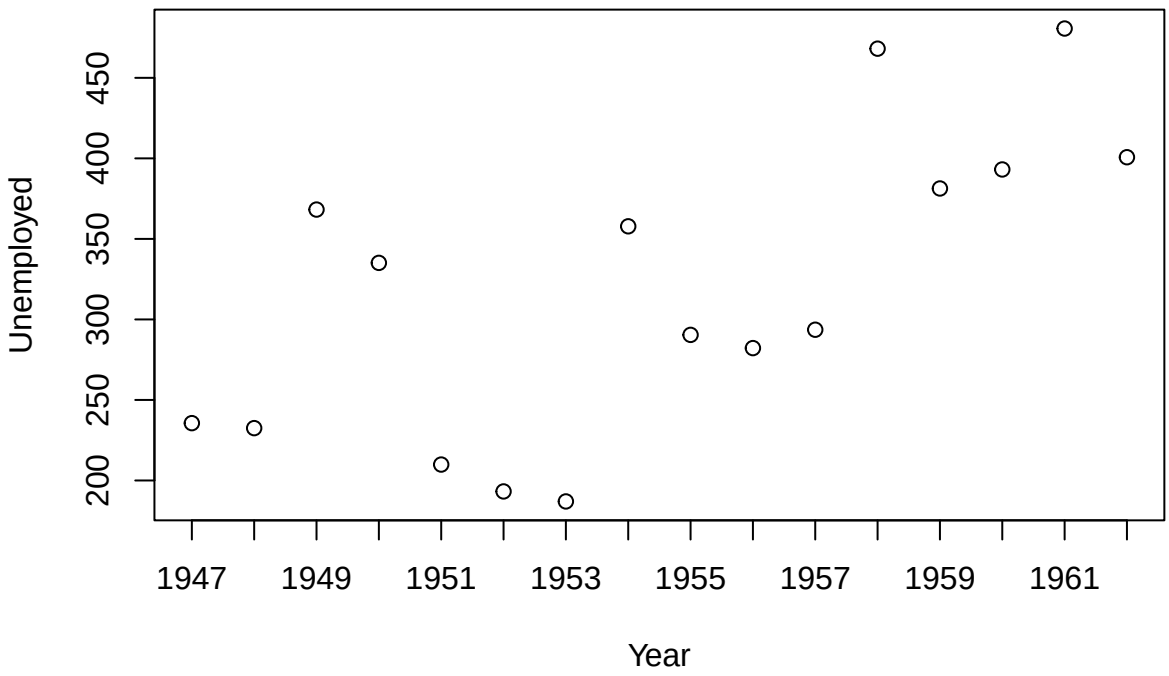
```
longley[longley$Year == 1947:1950,]
```

```
##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
## 1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
## 1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
## 1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
## 1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
```

1.(c) Plot (Unemployed ~ Year)

```
xmin = min(longley$Year)
xmax = max(longley$Year)
```

```
plot(longley$Year,longley$Unemployed,xlab = 'Year',ylab='Unemployed',xaxt="n",xlim=c(xmin,xmax))
axis(1, at = xmin:xmax)
```



1.(d) Change the type of plot to a line

```
xmin = min(longley$Year)
xmax = max(longley$Year)
plot(longley$Year,longley$Unemployed,type='l',xlab = 'Year',ylab='Unemployed',xaxt="n",xlim=c(xmin,xmax)
axis(1, at = xmin:xmax)
```

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.

```
ctime <- c(17, 16, 20, 24, 22, 15, 21, 15, 17, 22)
summary(ctime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   16.25   18.50   18.90   21.75   24.00
```

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.

```
ctime <- c(17, 16, 20, 24, 22, 15, 21, 15, 17, 22)
ctime[ctime == 24] <- 18
print(ctime)
```

```
##  [1] 17 16 20 18 22 15 21 15 17 22
```

```
summary(ctime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   16.25   17.50   18.30   20.75   22.00
```

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

```
lt <- length(ctime[ctime <= 20])
sprintf("Number of times commute was atleast 20 min: %i", lt)
```

```
## [1] "Number of times commute was atleast 20 min: 7"
```

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

```
lt <- length(ctime[ctime < 17])
sprintf("Number of times commute was less than 17 min: %i", lt)
```

```
## [1] "Number of times commute was less than 17 min: 3"
```

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.

```
library("readstata13")
library("epiDisplay")
```

```
## Loading required package: foreign
```

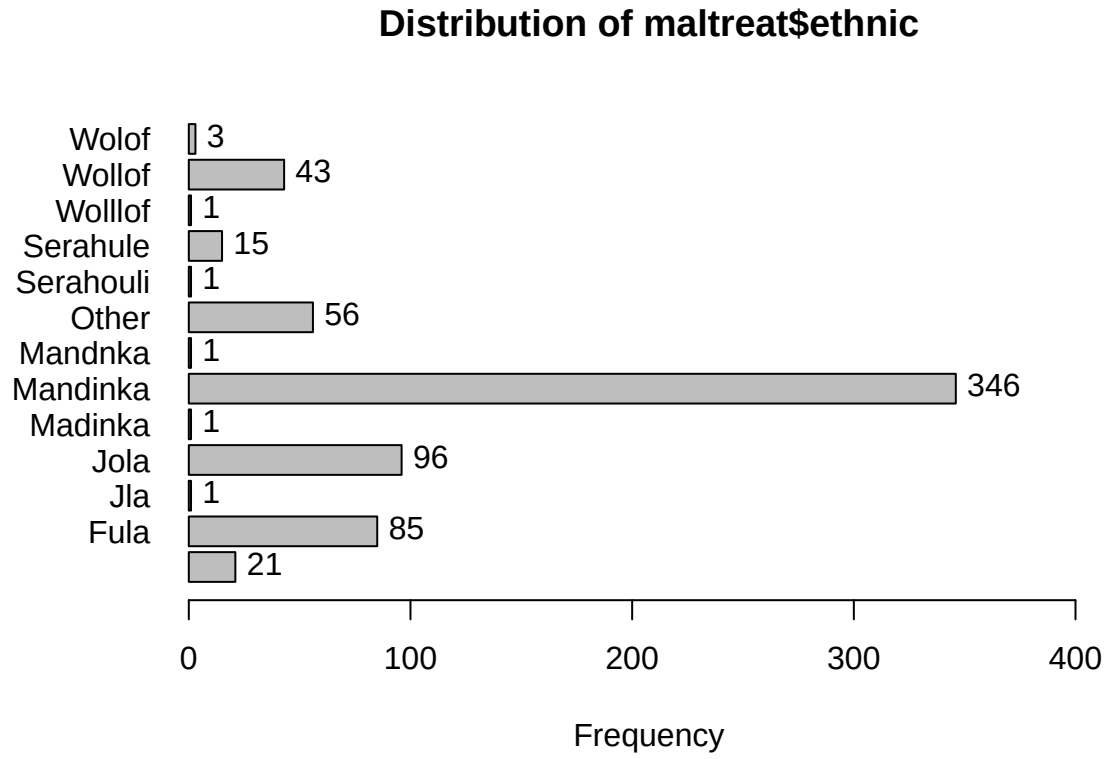
```
## Loading required package: survival
```

```
## Loading required package: MASS
```

```
## Loading required package: nnet
```

```
#Read the maltread data
maltreat <- read.dta13("data/maltreat.dta")

#Display the frequency
tab1(maltreat$ethnic, col = "grey")
```




```
## maltreat$ethnic :
##           Frequency Percent Cum. percent
##                  21     3.1          3.1
## Fula             85    12.7         15.8
## Jla               1     0.1         16.0
## Jola             96    14.3         30.3
## Madinka           1     0.1         30.4
## Mandinka        346    51.6         82.1
## Mandnka           1     0.1         82.2
## Other            56     8.4         90.6
## Serahouli         1     0.1         90.7
## Serahule         15     2.2         93.0
## Wolllof           1     0.1         93.1
## Wollof           43     6.4         99.6
## Wolof             3     0.4        100.0
##   Total         670   100.0        100.0
```

```
#create the categorical feature enthnicity from enthnic
maltreat$ethnicity <- as.factor(maltreat$ethnic)
levels(maltreat$ethnicity)
```

```
##  [1] ""          "Fula"      "Jla"       "Jola"      "Madinka"   "Mandinka"
##  [7] "Mandnka"   "Other"     "Serahouli" "Serahule"  "Wolllof"   "Wollof"
## [13] "Wolof"
```

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

```
#Correct Jola
levels(maltreat$ethnicity)[startsWith(levels(maltreat$ethnicity),
"J")] <- "Jola"
maltreat$ethnicity[startsWith(as.character(maltreat$ethnicity),"J")] <- "Jola"
```

(ii). Replace ethnic = "Mandinka" if ethnic value starts with an "M"

```
#Correct Jola
levels(maltreat$ethnicity)[startsWith(levels(maltreat$ethnicity),
"M")] <- "Mandinka"
maltreat$ethnicity[startsWith(as.character(maltreat$ethnicity),"M")] <- "Mandinka"
```

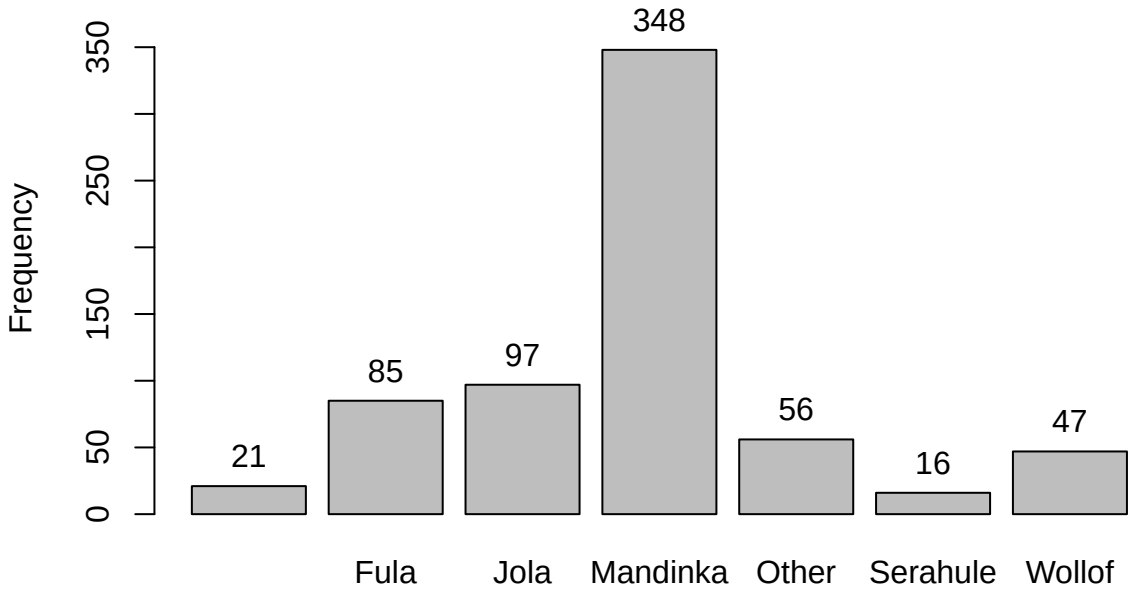
(iii). Replace ethnic = "Serahule" if ethnic value starts with an "S"

```
#Correct Jola
levels(maltreat$ethnicity)[startsWith(levels(maltreat$ethnicity),
"S")] <- "Serahule"
maltreat$ethnicity[startsWith(as.character(maltreat$ethnicity),"S")] <- "Serahule"
```

(iv). Replace ethnic = "Wollof" if ethnic value starts with a "W"

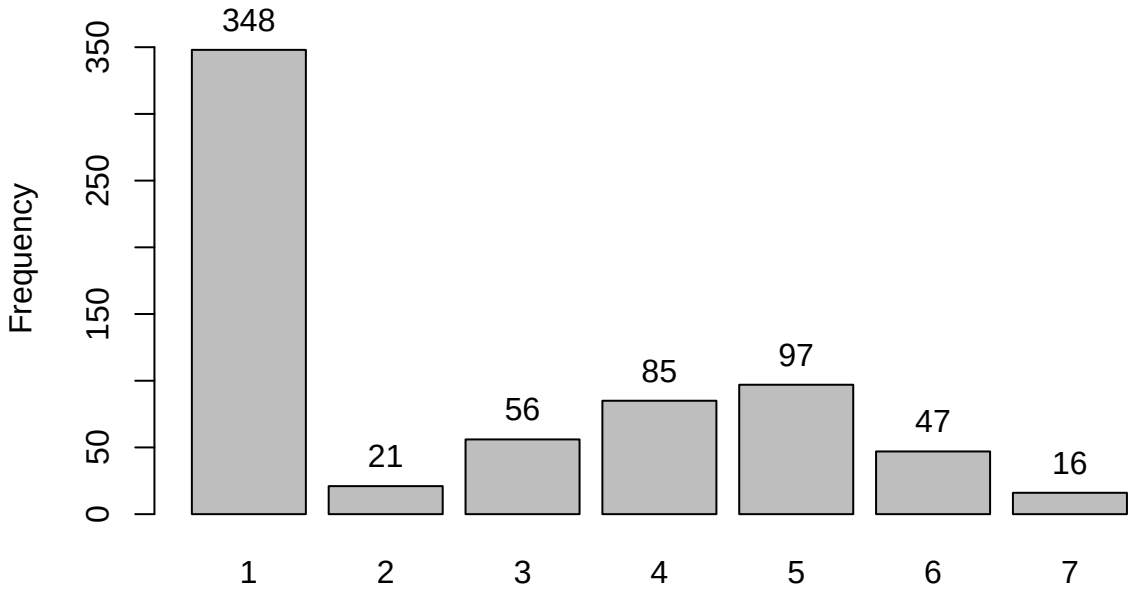
```
#Correct Jola
levels(maltreat$ethnicity)[startsWith(levels(maltreat$ethnicity),
"W")] <- "Wollof"
maltreat$ethnicity[startsWith(as.character(maltreat$ethnicity),"W")] <- "Wollof"

# Correct the feature enthic from corrected enthnicity
maltreat$ethnic <- maltreat$ethnicity
levels(maltreat$ethnic) <- levels(maltreat$ethnicity)
tab1(maltreat$ethnic, col = "grey")
```




```
## maltreat$ethnic :
##          Frequency Percent Cum. percent
##                 21     3.1          3.1
## Fula            85    12.7         15.8
## Jola            97    14.5         30.3
## Mandinka       348    51.9         82.2
## Other           56     8.4         90.6
## Serahule        16     2.4         93.0
## Wollof          47     7.0        100.0
##   Total        670   100.0        100.0
```

```
# Change the feature enthnicity to numeric and categorical
maltreat$ethnicity <- as.numeric(factor(maltreat$ethnicity,levels=unique(maltreat$ethnic),exclude = NULL
tab1(maltreat$ethnicity,col = "grey")
```




```
## maltreat$ethnicity :
##          Frequency Percent Cum. percent
## 1              348    51.9         51.9
## 2               21     3.1         55.1
## 3               56     8.4         63.4
## 4               85    12.7         76.1
## 5               97    14.5         90.6
## 6               47     7.0         97.6
## 7               16     2.4        100.0
##   Total        670   100.0        100.0
```