Lab 02 R Problems (Part 1)

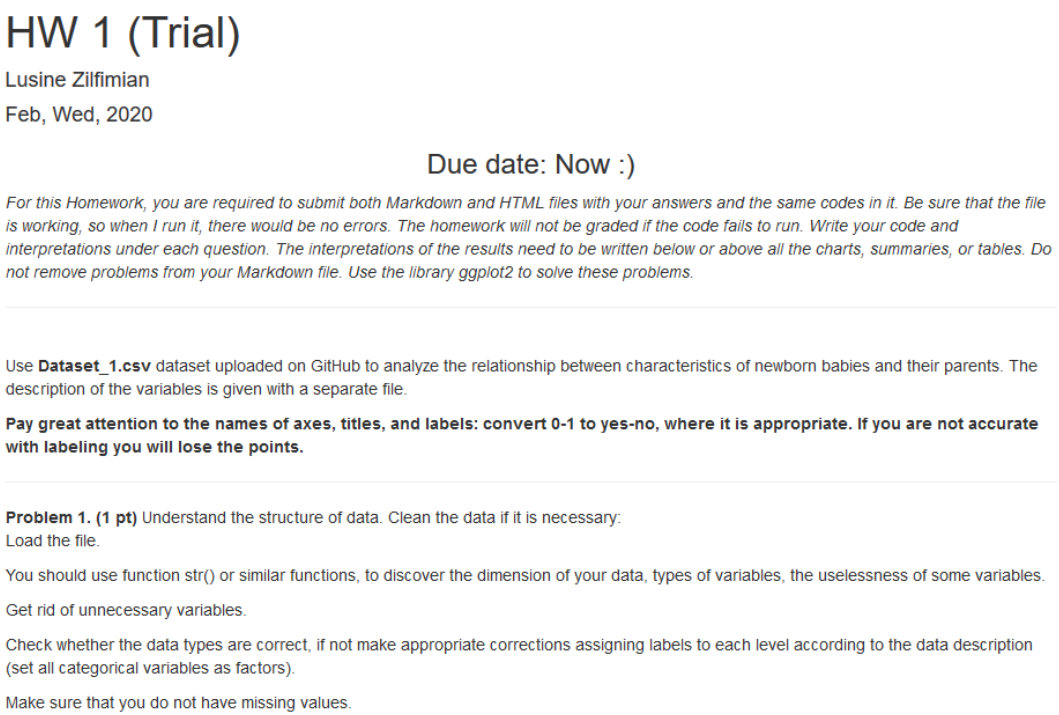
Lusine Zilfimian

February 26 (Wednesday), 2020

Task

- The task is to solve the first problem of Trial HW (link)
- For teaching purpose, all R codes are provided:

```
knitr::include_graphics("HW1trial.PNG")
```



HW 1 (Trial)

Lusine Zilfimian

Feb, Wed, 2020

Due date: Now :)

For this Homework, you are required to submit both Markdown and HTML files with your answers and the same codes in it. Be sure that the file is working, so when I run it, there would be no errors. The homework will not be graded if the code fails to run. Write your code and interpretations under each question. The interpretations of the results need to be written below or above all the charts, summaries, or tables. Do not remove problems from your Markdown file. Use the library ggplot2 to solve these problems.

Use **Dataset_1.csv** dataset uploaded on GitHub to analyze the relationship between characteristics of newborn babies and their parents. The description of the variables is given with a separate file.

Pay great attention to the names of axes, titles, and labels: convert 0-1 to yes-no, where it is appropriate. If you are not accurate with labeling you will lose the points.

Problem 1. (1 pt) Understand the structure of data. Clean the data if it is necessary:
Load the file.

You should use function str() or similar functions, to discover the dimension of your data, types of variables, the uselessness of some variables.

Get rid of unnecessary variables.

Check whether the data types are correct, if not make appropriate corrections assigning labels to each level according to the data description (set all categorical variables as factors).

Make sure that you do not have missing values.

First steps

- Create R Markdown file (see previous slide: Lab 01)
- Set global option for chunks and load needed packages

```
knitr::opts_chunk$set(echo = TRUE, warning = F,message = F)
```

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(ggplot2, dplyr, knitr, kableExtra, ggthemes)
```

- Read the instructions and solve the tasks sequentially
- Open Metadata and try to understand the labels of variable:

This dataset contains information on newborn babies and their parents. It contains both continuous and discrete variables.

ID	Baby number	
length	Length of baby (inches)	Scale
bweight	Weight of baby (lbs)	Scale
headcir	Head Circumference	Scale
gestation	Gestation (weeks)	Scale
smoker	Mother smokes 1 = smoker 0 = non-smoker	Binary
motherage	Maternal age	Scale
mnocig	Number of cigarettes smoked per day by mother	Scale
mheight	Mothers height (inches)	Scale
mppwt	Mothers pre-pregnancy weight (lbs)	Scale
fage	Father's age	Scale
fedyrs	Father's years in education	Scale
fnocig	Number of cigarettes smoked per day by father	Scale
bined_weight	The variable bweight was grouped: Bweight<5.5, bined_weight =5 5.5<=Bweight<6.5, bined_weight=6 6.5<=Bweight<7.5, bined_weight=7 7.5<=Bweight<8.5, bined_weight=8 8.5<=Bweight, bined_weight=9	Nominal

Solving Problem 1.

Load the file. Understand the structure of data.

```
dt <- read.csv("Dataset_1.csv")
str(dt)
```

```
## 'data.frame':    48 obs. of  17 variables:
##  $ id            : int  431 300 15648 15650 15652 25656 156
##  $ headcir       : int  12 12 12 12 12 13 12 13 13 12 ...
##  $ length        : int  19 18 16 17 16 17 17 17 17 19 ...
##  $ bweight       : num  4.2 4.5 4.7 4.8 4.9 4.9 5 5.1 5.1 5
##  $ gestation     : int  33 35 33 33 34 36 36 37 35 37 ...
##  $ smoker        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ motherage     : int  20 41 36 37 28 36 25 36 23 37 ...
##  $ mnocig        : int  7 7 18 17 15 10 8 30 9 7 ...
##  $ mheight       : int  63 65 66 65 64 64 64 63 65 64 ...
##  $ mppwt         : int  109 125 131 124 139 124 134 119 139
##  $ fage          : int  20 37 19 22 24 34 26 24 37 20 ...
```

Load the file. Understand the structure of data.

- The data set consists of 48 observations and 17 attributes.
- Four of the attributes are categorical variables (the last variable LowBirthWeight provides the same information as lowbwt, so we can exclude it from analysis).
- Categorical variables should be defined as factors during computations.
- First variable: id, has no sense, as ID gives us information only about number of babies and is unique for each of the observations, we can drop it from the data frame.

Get rid of unnecessary variables.

```
# Select function is from package dplyr
(dt <- dt %>%
  select(-c("id", "lowbwt")))
```

```
##    headcir length bweight gestation smoker motherage mnocig
## 1       12     19     4.2        33      1        20      7
## 2       12     18     4.5        35      1        41      7
## 3       12     16     4.7        33      1        36     18
## 4       12     17     4.8        33      1        37     17
## 5       12     16     4.9        34      1        28     15
## 6       13     17     4.9        36      1        36     10
## 7       12     17     5.0        36      1        25      8
## 8       13     17     5.1        37      1        36     30
## 9       13     17     5.1        35      1        23      9
## 10      12     19     5.2        37      1        37      7
## 11      12     17     5.4        36      1        38     10
## 12      13     19     5.5        39      1        34      7
```

Check whether the data types are correct, if not make appropriate corrections assigning labels to each level according to the data description.

```
dt$smoker <- factor(dt$smoker, levels = c(0, 1),
                            labels = c("Non-smoker", "Smoker"))
dt$mage35 <- factor(dt$mage35, levels = c(0, 1),
                            labels = c("Under", "Over"))
dt$bined_weight <- factor(dt$bined_weight, ordered = T)
```

Make sure that you do not have missing values.

```
any(is.na(dt))
```

```
## [1] FALSE
```

- There are no missing values in dataset.
- The final data is the following

```
knitr::kable(head(dt[, 1:6], 2)) %>%
 kable_styling(bootstrap_options = "striped",
   full_width = F, font_size = 14)
```

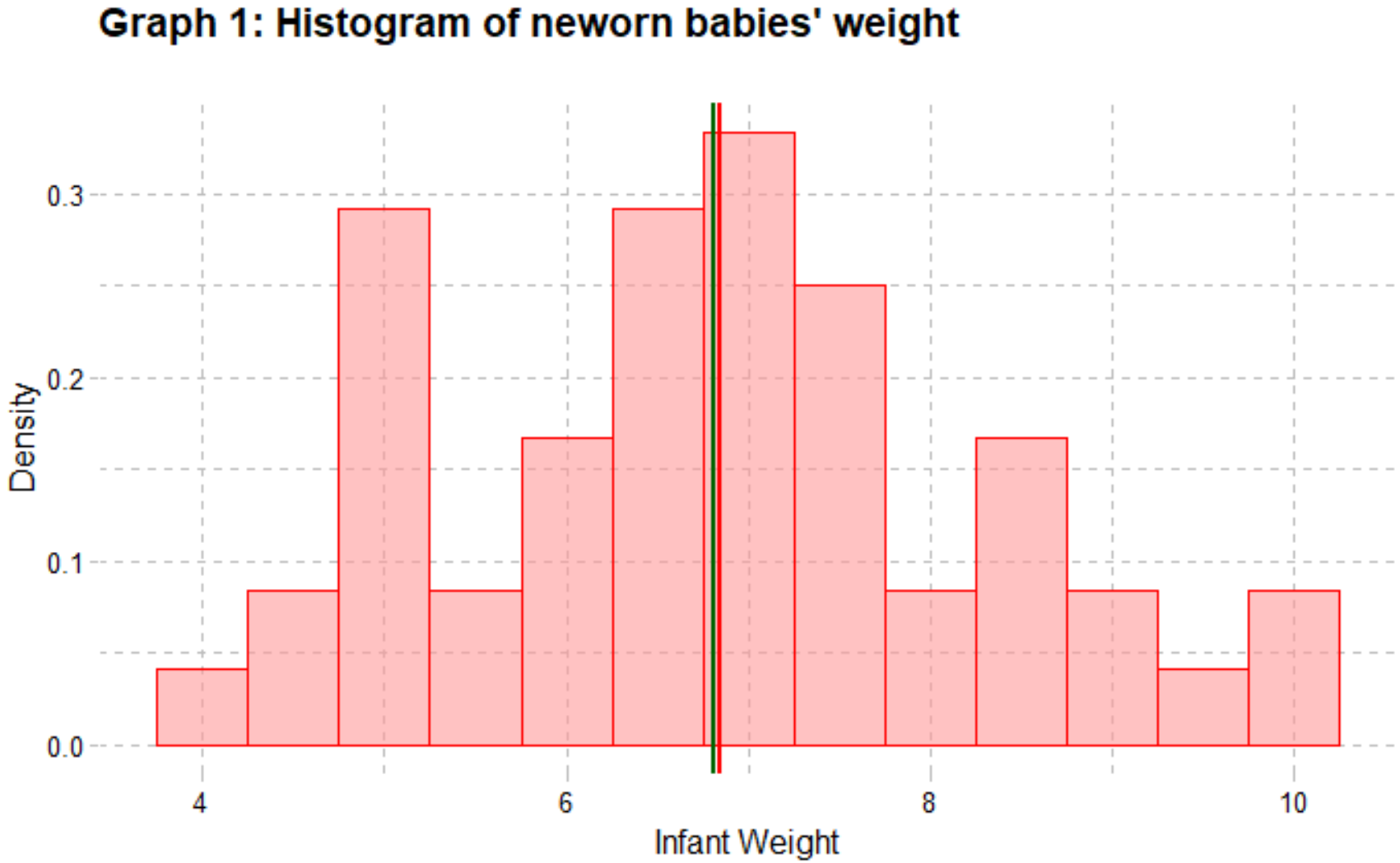
headcir	length	bweight	gestation	smoker	motherage
12	19	4.2	33	Smoker	20
12	18	4.5	35	Smoker	41

Solving Problem 2.

Describe one numeric variable. Use data visualization techniques.[a]

[a] **Note:** you need to pay great attention to the names of axes, titles, and labels

```
g1 <- dt %>%
  ggplot(mapping = aes(x = bweight)) +
  geom_histogram(binwidth = 0.5, alpha=0.6, fill="#FF9999",
     col = "red", aes(y=..density..)) +
  ggtitle("Graph 1: Histogram of neworn babies' weight") +
  xlab("Infant Weight") + ylab("Density") +
  ggthemes::theme_pander() + # optional
  geom_vline(aes(xintercept = mean(bweight)),
    col = 'darkgreen', size = 1) + # optional
  geom_vline(aes(xintercept = median(bweight)),
    col = 'red', size = 1) # optional
```

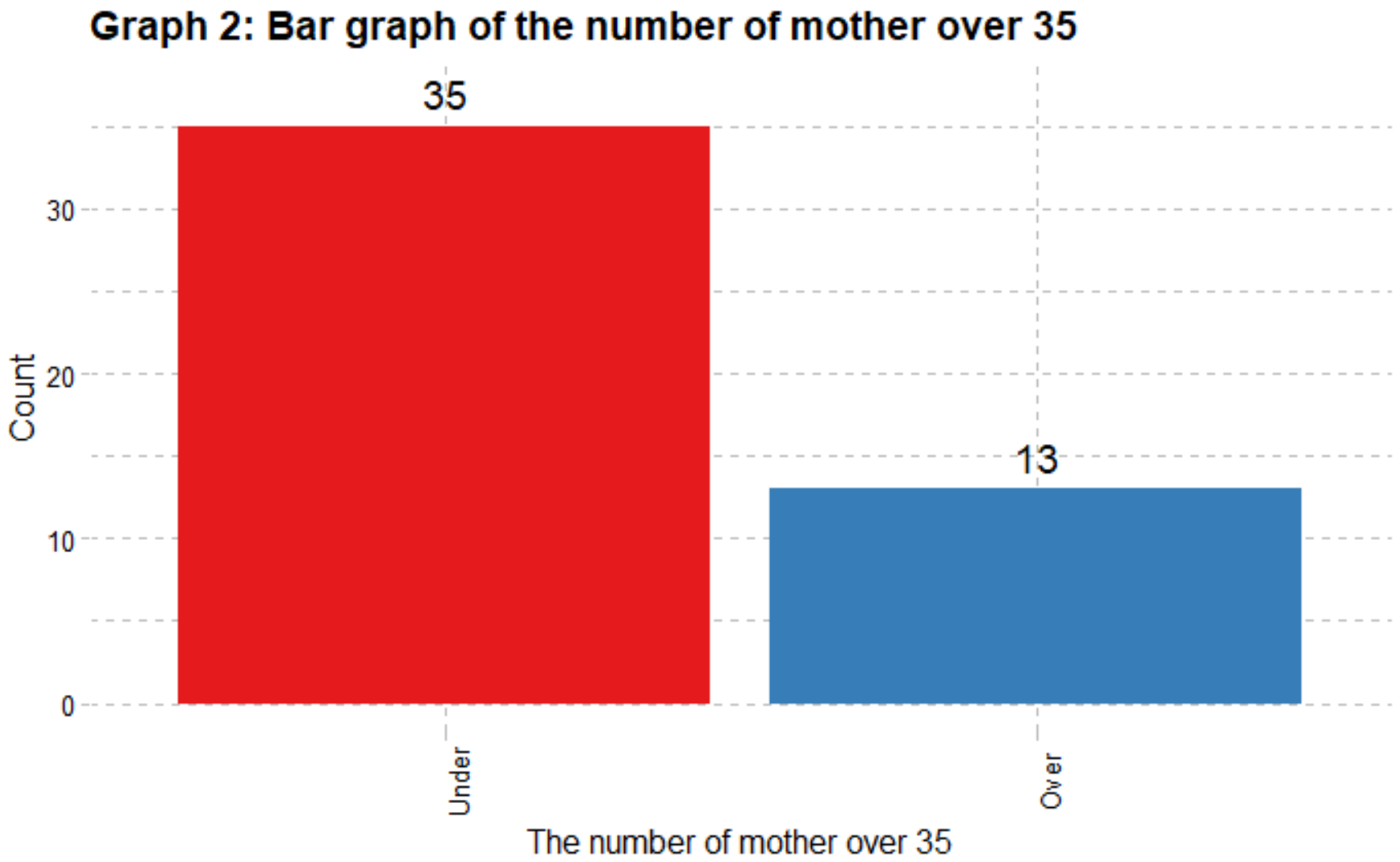

Graph 1: Histogram of neworn babies' weight
Density
0.0
0.1
0.2
0.3
4
6
8
10
Infant Weight

Describe one numeric variable. Use data visualization techniques.

- The shape of the distribution is **approximately** normal.
- As the histogram shows infant weight varies from 3.75 to 10.25 lbs, while the most common values are mostly between 6.75 and 7.25 lbs (by setting the width of bins 0.5).
- The number of extreme values is not higher, however, weights around 5 lbs are relatively frequent.
- Speak about mean, median, spread of data.

Describe one categorical variable. Use data visualization techniques.

```
g2 <- dt %>%
  ggplot(mapping = aes(x = mage35, fill = mage35)) +
  geom_bar() +
  xlab("The number of mother over 35") + ylab("Count") +
  ggtitle("Graph 2: Bar graph of the
    number of mother over 35")+
  scale_fill_brewer(palette = "Set1") + # optional
  ggthemes::theme_pander() + # optional
  theme(legend.position = "None")+
  theme(axis.text.x = element_text(angle = 90)) + # optional
  geom_text(data = tablecat(mage35), aes(y = n + 2,
    label = paste(n)), size = 5) # optional
```


Graph 2: Bar graph of the number of mother over 35
35
13
30
20
10
0
Count
Under
Over
The number of mother over 35

Describe one categorical variable. Use data visualization techniques.

- According to the above bar chart, the number of mothers under the age of 35 in our dataset is nearly 3 times more than the number of mothers above 35 (35 and 13, respectively).
- Note that you need to create the **tablecat()** function by yourself.

Problem 2 and 3

To be continued.