Lab 03 R Problems (Part 2)

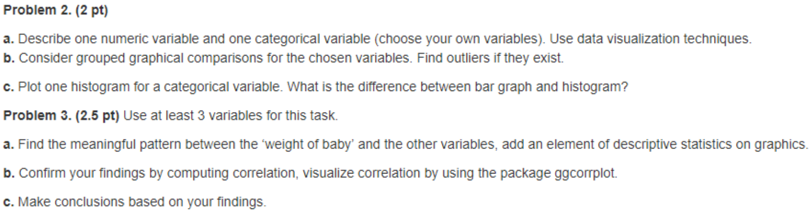
Lusine Zilfimian

March 04 (Wednesday), 2020

Task

- The task is to solve the second and third problem of Trial HW (**link**)

```r
knitr::include_graphics("HW2trial.PNG")
```



Problem 2. (2 pt)

a. Describe one numeric variable and one categorical variable (choose your own variables). Use data visualization techniques.
b. Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

c. Plot one histogram for a categorical variable. What is the difference between bar graph and histogram?

Problem 3. (2.5 pt) Use at least 3 variables for this task.

a. Find the meaningful pattern between the 'weight of baby' and the other variables, add an element of descriptive statistics on graphics.

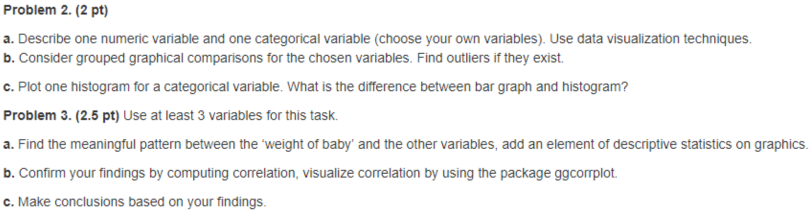
b. Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

c. Make conclusions based on your findings.

Task

- The task is to solve the second and third problem of Trial HW (**link**)

```r
knitr::include_graphics("HW2trial.PNG")
```



Problem 2. (2 pt)

a. Describe one numeric variable and one categorical variable (choose your own variables). Use data visualization techniques.
b. Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

c. Plot one histogram for a categorical variable. What is the difference between bar graph and histogram?

Problem 3. (2.5 pt) Use at least 3 variables for this task.

a. Find the meaningful pattern between the ‘weight of baby’ and the other variables, add an element of descriptive statistics on graphics.

b. Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

c. Make conclusions based on your findings.

- Note, that the previous codes were executed to have correct data for P2 and P3.

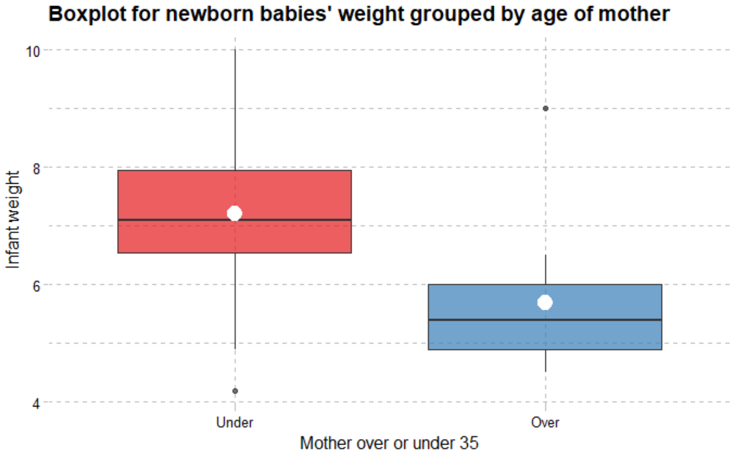
Solving Problem 2 (b)

Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

```r
g3 <- dt %>%
ggplot(mapping = aes(x = mage35, y = bweight, fill = mage35))
  geom_boxplot(alpha = 0.7) +
  ggtitle("Boxplot for newborn babies' weight grouped
    by age of mother") +
  ylab("Infant weight") +
  xlab("Mother over or under 35") +
  stat_summary(fun.y = mean, geom = "point", shape = 20,
    size = 7,  color = "white") + # optional
  scale_fill_brewer(palette = "Set1") + # optional
  ggthemes::theme_pander() + # optional
  theme(legend.position = "None")
```

Solving Problem 2 (b)

Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.



Solving Problem 2 (b)

Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

- It can be seen that the middle quantile(median) of infant weight in the group of the mothers under 35 years old is higher (7 lbs) than others, while in the second group (above 35) median is nearly 5.5 lbs. Upper and lower quantiles also show a noticeable difference between babies' weight in groups. It means that being pregnant in years above 35 can reduce the weight of newborn babies.

Solving Problem 2 (b)

Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

- It can be seen that the middle quantile(median) of infant weight in the group of the mothers under 35 years old is higher (7 lbs) than others, while in the second group (above 35) median is nearly 5.5 lbs. Upper and lower quantiles also show a noticeable difference between babies' weight in groups. It means that being pregnant in years above 35 can reduce the weight of newborn babies.
- The dispersion of the first group is higher.

Solving Problem 2 (b)

Consider grouped graphical comparisons for the chosen variables. Find outliers if they exist.

- It can be seen that the middle quantile(median) of infant weight in the group of the mothers under 35 years old is higher (7 lbs) than others, while in the second group (above 35) median is nearly 5.5 lbs. Upper and lower quantiles also show a noticeable difference between babies' weight in groups. It means that being pregnant in years above 35 can reduce the weight of newborn babies.
- The dispersion of the first group is higher.
- Also, boxplot shows one outlier in each group (nearly 4.2 and 9 lbs, respectively).

Solving Problem 2 (c)

Plot one histogram for a categorical variable. What is the difference between bar graph and histogram?

- Error: StatBin requires a continuous x variable: the x variable is discrete. Perhaps you want stat="count"?

```
ggplot(dt, aes(x=mage35)) +
  geom_histogram() +
  xlab("Mother over 35")
```

```
## Error: StatBin requires a continuous x vari
```

Solving Problem 2 (c)

Plot one histogram for a categorical variable. What is the difference between bar graph and histogram?

- Error: StatBin requires a continuous x variable: the x variable is discrete. Perhaps you want stat=“count”?
- Histograms are used to visualize the continuous data using binning, it shows distributions of variable(s). There is no space between bins.

```r
ggplot(dt, aes(x=mage35)) +
  geom_histogram() +
  xlab("Mother over 35")
```

```
## Error: StatBin requires a continuous x vari
```

Solving Problem 2 (c)

Plot one histogram for a categorical variable. What is the difference between bar graph and histogram?

- Error: StatBin requires a continuous x variable: the x variable is discrete. Perhaps you want stat=“count”?
- Histograms are used to visualize the continuous data using binning, it shows distributions of variable(s). There is no space between bins.
- Bar charts are used for comparing categorical data.

```
ggplot(dt, aes(x=mage35)) +
  geom_histogram() +
  xlab("Mother over 35")
```

```
## Error: StatBin requires a continuous x variable: the x vari
```

Solving Problem 3 (a)

Use at least 3 variables for this task. Find the meaningful pattern between the ‘weight of baby’ and the other variables, add an element of descriptive statistics on graphic.

- First, to add an element of descriptive statistics on a graphic, we need to calculate the means of chosen variables.

```
(dtSum <- dt %>%
     group_by(LowBirthWeight) %>%
     summarise( mnocig = mean(mnocig), bweight = mean(bweight)
```

```
## # A tibble: 2 x 3
##   LowBirthWeight mnocig bweight
##   <fct>           <dbl>   <dbl>
## 1 Low            12.1      4.94
## 2 Normal          2.58     7.43
```

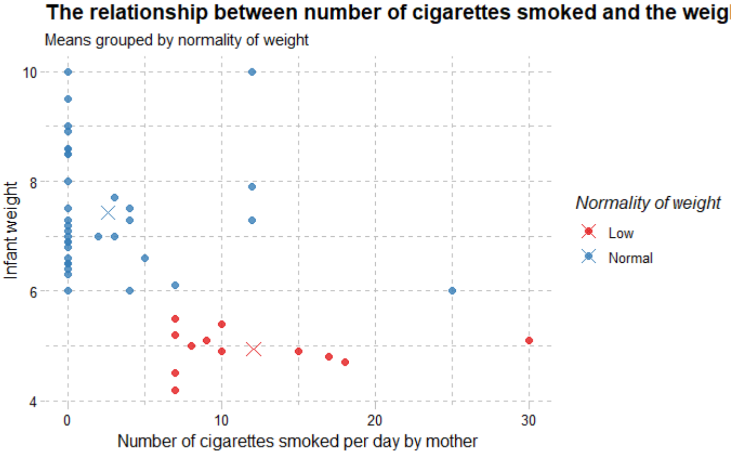
Solving Problem 3 (a)

Use at least 3 variables for this task. Find the meaningful pattern between the 'weight of baby' and the other variables, add an element of descriptive statistics on graphics.

```
g4 <- ggplot(dt, mapping = aes(x = mnocig, y = bweight, color
  geom_point(alpha = 0.8, size = 2) +
  ggtitle( "The relationship between number of cigarettes smok
  ylab("Infant weight") +
  xlab("Number of cigarettes smoked per day by mother") +
  geom_point(data = dtSum, shape = 4, size = 4) +
  scale_color_brewer(palette = "Set1") +
  labs(color="Normality of weight") +
  ggthemes::theme_pander() # otpional
```

Solving Problem 3 (a)

Use at least 3 variables for this task. Find the meaningful pattern between the 'weight of baby' and the other variables, add an element of descriptive statistics on graphics.



Solving Problem 3 (a)

Use at least 3 variables for this task. Find the meaningful pattern between the 'weight of baby' and the other variables, add an element of descriptive statistics on graphics.

One of the most useful plots for visualization of relationship is scatterplot. This scatter plot (plots) shows the relationship between the number of cigarettes smoked per day by mother and infant weight. From the first glance, the negative correlation between these two variables can be seen. 'Low birth weight' was considered as a third variable, which is binary. If infant weight is less than 6 lbs the third variable equals 1(Low weight). Here, the graph illustrates that the weight of babies whose mothers do not smoke is higher than the non-smoker ones. This can be confirmed by computing group mean. The mean weight of the first group is 7.5 lbs, while the second is nearly 5 lbs.

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

- First, the numeric variables should be selected

```
correl <- round(cor(dt[, sapply(dt, is.numeric)]),1)
g5 <- ggcorrplot::ggcorrplot(correl,
  method = "circle", lab = TRUE)
```

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

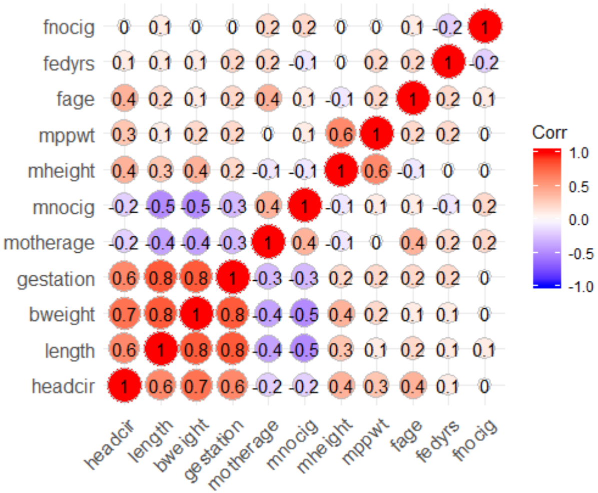
- First, the numeric variables should be selected

```
correl <- round(cor(dt[, sapply(dt, is.numeric)]),1)
g5 <- ggcorrplot::ggcorrplot(correl,
  method = "circle", lab = TRUE)
```

- To plot correlation plot you need to install and call the package `ggcorrplot`

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.



	headcir	length	bweight	gestation	motherage	mnocig	mheight	mppwt	fage	fedyrs	fnocig
fnocig	0	0.1	0	0	0.2	0.2	0	0	0.1	-0.2	1
fedyrs	0.1	0.1	0.1	0.2	0.2	-0.1	0	0.2	0.2	1	-0.2
fage	0.4	0.2	0.1	0.2	0.4	0.1	-0.1	0.2	1	0.2	0.1
mppwt	0.3	0.1	0.2	0.2	0	0.1	0.6	1	0.2	0.2	0
mheight	0.4	0.3	0.4	0.2	-0.1	-0.1	1	0.6	-0.1	0	0
mnocig	-0.2	-0.5	-0.5	-0.3	0.4	1	-0.1	0.1	0.1	-0.1	0.2
motherage	-0.2	-0.4	-0.4	-0.3	1	0.4	-0.1	0	0.4	0.2	0.2
gestation	0.6	0.8	0.8	1	-0.3	-0.3	0.2	0.2	0.2	0.2	0
bweight	0.7	0.8	1	0.8	-0.4	-0.5	0.4	0.2	0.1	0.1	0
length	0.6	1	0.8	0.8	-0.4	-0.5	0.3	0.1	0.2	0.1	0.1
headcir	1	0.6	0.7	0.6	-0.2	-0.2	0.4	0.3	0.4	0.1	0

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

- There is a negative correlation between the number of cigarettes smoked by the mother and babies' weight.

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

- There is a negative correlation between the number of cigarettes smoked by the mother and babies' weight.
- Pearson correlation coefficient shows only a linear relationship between 2 variables.

Solving Problem 3 (b)

Confirm your findings by computing correlation, visualize correlation by using the package ggcorrplot.

- There is a negative correlation between the number of cigarettes smoked by the mother and babies' weight.
- Pearson correlation coefficient shows only a linear relationship between 2 variables.
- Main conclusion: the mothers' smoking increases the risk of having low weighted babies.

Solving Problem 3 (c)

Make conclusions based on your findings.

- See 3 a solution's conlusion.

Done

- **Note:** this is just a sample solution.

Done

- **Note:** this is just a sample solution.
- I paid attention to codes rather than conclusions. So do not copy or rely only on these texts.

Done

- **Note:** this is just a sample solution.
- I paid attention to codes rather than conclusions. So do not copy or rely only on these texts.
- Enjoy doing HW1!

Done

- **Note:** this is just a sample solution.
- I paid attention to codes rather than conclusions. So do not copy or rely only on these texts.
- Enjoy doing HW1!
- HW2 will be uploaded at the end of this weekend.

Done

- **Note:** this is just a sample solution.
- I paid attention to codes rather than conclusions. So do not copy or rely only on these texts.
- Enjoy doing HW1!
- HW2 will be uploaded at the end of this weekend.
- Stay Healthy! $\ddot\smile$