VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF SCIENCE

ADVANCED PROGRAM IN COMPUTER SCIENCE

FINAL PROJECT

Course: Applied Statistics II (STAT452)

Instructor: Đinh Ngọc Thanh

TA: Nguyễn Hữu Toàn

Full Name: Đinh Mỹ Kỳ

Student ID: 20125060

Class: 20CTT

Contents

1. Data information

1.1 Data description

- The data file is *bwght2.csv* with the description file *bwght2._description.txt*
- The data is about baby's health measured by birthweight together with other factors from their parents and prenatal care (baby healthcare before birth)

1.2 Some important baby's healthcare terms

Since this data is related to baby's health, I researched on some **important terms** to fully understand the data:

+ Prenatal: before birth; during or relating to pregnancy.

+ APGAR score:

- The APGAR is a quick, overall assessment of newborn well-being. The test is used immediately after the delivery of a baby. APGAR scores are recorded at one minute **(omaps)** and five minutes **(fmaps)** from the time of birth.
- APGAR measures the baby's color, heart rate, reflexes, muscle tone, and respiratory effort.
- APGAR scores range from zero to two for each condition with a maximum final total score of ten
- **omaps** and **fmaps** range help us to know the baby's health and suitable healthcare needed to provide.
- **omaps** provides information about the baby's physical health and helps the physician determine if an immediate or future medical treatment will be required.
- **fmaps** measures how the baby has responded to previous *resuscitation(*)* attempts if such attempts were made.

(*) resuscitation: cause someone who has stopped breathing to start breathing again

Source:

https://americanpregnancy.org/healthy-pregnancy/labor-and-birth/apgar-test/

https://www.acog.org/clinical/clinical-guidance/committee-opinion/articles/2015/10/the-apgar-score?utm_source=redirect&utm_medium=web&utm_campaign=otn

1.3 Data format

- A data frame with 1832 observations on 23 variables:

Column ID	Variable	Explain
1	mage	mother's age, years
2	meduc	mother's educ, years
3	monpre	month prenatal care began:
4	npvis	total number of prenatal visits
5	fage	father's age, years
6	feduc	father's educ, years
7	bwght	birth weight, grams
8	**omaps**	one minute apgar score
9	**fmaps**	five minute apgar score
10	cigs	avg cigarettes per day
11	drink	avg drinks per week
12	lbw	=1 if bwght <= 2000
13	vlbw	=1 if bwght <= 1500
14	male	=1 if baby male
15	mwhte	=1 if mother white
16	mblck	=1 if mother black
17	moth	=1 if mother is other
18	fwhte	=1 if father white
19	fblck	=1 if father black
20	foth	=1 if father is other
21	lbwght	log(bwght)
22	magesq	mage^2
23	npvissq	npvis^2

2. Data pre-processing

2.1 Read data

- Code:

```
setwd("D:/2.Year2/Semester3/STAT452/Dataforfinal")
df<-read.csv("bwght2.csv",header=TRUE)
attach(df)
str(df)
```

- Output (already eliminate some rows of output for easier observation):

```
> str(df)
'data.frame':   1832 obs. of  23 variables:
 $ mage   : int  26 29 33 28 23 28 27 41 32 16 ...
 $ meduc  : chr  "12" "12" "12" "17" ...
 $ monpre : chr  "2" "2" "1" "5" ...
 ...
 $ bwght  : int  3060 3730 2530 3289 3590 3420 3355 3459 3590 4410 ...
 ...
 $ cigs   : chr  "0" "." "0" "0" ...
 $ drink  : chr  "0" "." "0" "0" ...
 ...
 $ foth   : int  1 0 0 0 0 0 0 1 0 0 ...
 $ lbwght : num  8.03 8.22 7.84 8.1 8.19 ...
 $ magesq : int  676 841 1089 784 529 784 729 1681 1024 256 ...
 $ npvissq: chr  "144" "144" "144" "64" ...
>
```

→ Observation:

- From the output, we see that many columns are character while they're supposed to be integer.

- Also, some of them contain **"."**

- Therefore, we need to convert all these character columns to integer type.

2.2 Convert character data to integer type

- Code:

+ isIntChar: check whether the columns contain characters that are integer

+ mutate_if: convert characters that satisfy isIntChar to integer

```
isIntChar<-function(x)
{
  return(all(is.character(x)))
}
isIntChar<-function(x)
```

- Output:

```
> df <- df %>% mutate_if(isIntChar, as.integer)
Warning messages:
1: Problem while computing `meduc = .Primitive("as.integer")(meduc)`.
ℹ NAs introduced by coercion
...
10: Problem while computing `npvissq =.Primitive("as.integer")(npvissq)`.
ℹ NAs introduced by coercion
```

→ The data has NAs (missing values) because of the **“.”** before (when the data still had character values). Thus, we fill these missing values with 0 for calculation.

2.3 Filling missing values

- Code:

```
df[is.na(df)]<-as.integer(0)

str(df)
```

- Output:

```
> df[is.na(df)]<-as.integer(0)
> str(df)
'data.frame':   1832 obs. of  23 variables:
 $ mage   : int  26 29 33 28 23 28 27 41 32 16 ...
 $ meduc  : int  12 12 12 17 13 12 16 17 12 11 ...
 ...
 $ foth   : int  1 0 0 0 0 0 0 1 0 0 ...
 $ lbwght : num  8.03 8.22 7.84 8.1 8.19 ...
 $ magesq : int  676 841 1089 784 529 784 729 1681 1024 256 ...
 $ npvissq: int  144 144 144 64 36 144 121 64 121 100 ...
```

Now we don’t have any NAs anymore.

3. Descriptive statistics

3.1 General

- Use ggplot2 library to visualize data

- Use custom color and theme to plot

- Level qualitative variables to explicit name instead of 0, 1

- Since some quantitative variables can be divided into sub-categories to analyze, we treat some quantitative as qualitative and visualize them to get some insights.

3.2 Qualitative variables

3.2.1 male attribute

- Create type.data with male attribute + change the level 0, 1 of male attribute to Female and Male

```
type.data <- data.frame(c=1:1832)
type.male = 1:length(male);
for (i in 1:length(male))
{
  if (male[i] == 0)
    type.male[i] = "Female"
  else
    type.male[i] = "Male"
}
type.data$type.male<-type.male
```

- Create typemaledf to show the frequency and percentage of male and female babies
+ Code:

```
typemaledf<-type.data%>%
  group_by(type.male)%>%
  summarise(count = n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal))
typemaledf
```

+ Output:

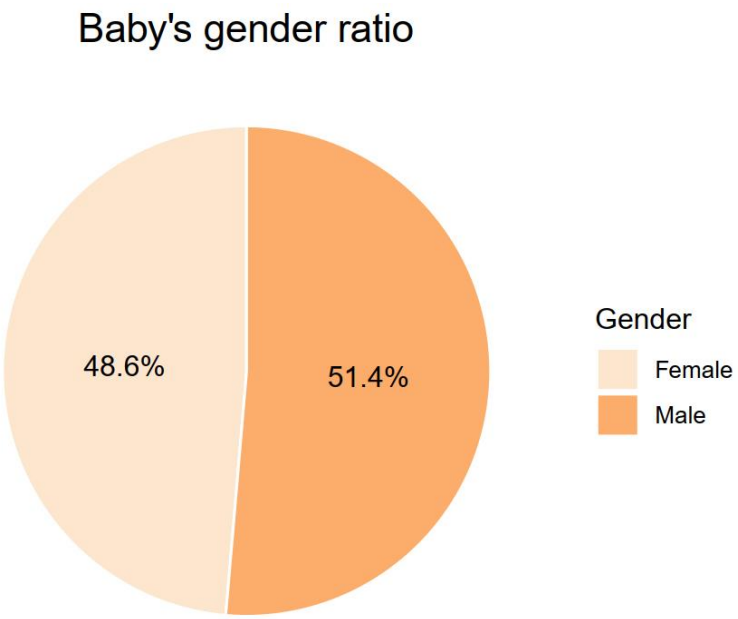
```
> typemaledf
# A tibble: 2 × 4
  type.male count ratioVal perc
  <chr>     <int>    <dbl> <chr>
1 Female      891    0.486 48.6%
2 Male        941    0.514 51.4%
```

- Draw pie chart of male attribute

+ Code:

```
malepie<-ggplot(typemaledf, aes(x="",y=ratioVal,fill=type.male))+
  theme_bw()+
  geom_bar(width = 2, stat = "identity", color="white") +
  coord_polar("y", start=0)+
  ggtitle("Baby's gender ratio")+
   scale_fill_brewer(name = "Gender", labels = c("Female","Male"),
palette="Oranges")+
  theme(plot.title=element_text(hjust=0.5, size=15),
        axis.title=element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.border = element_blank())+
  geom_text(aes(label=perc),position = position_stack(vjust = 0.5))
malepie
```

+ Output:



→ Analysis on male:

- Female: 51.4% - Male: 48.6%

- The number of female babies is slightly higher than that of male babies.

3.2.2 mblck attribute (mother is black)

- mblck=1 means mother is black

- Code:

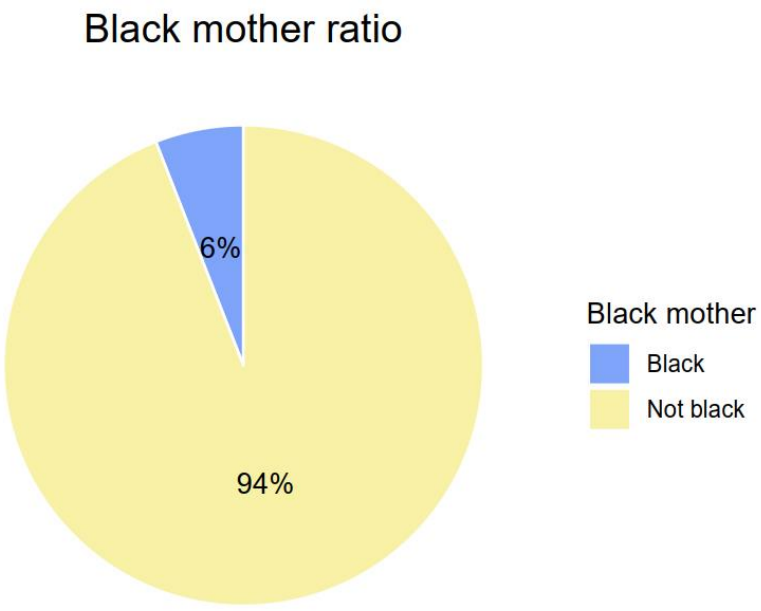
Create type.mblck, change levels, create typemblckdf, draw pie chart of mblck attribute

```
type.mblck = 1:length(mblck);
for (i in 1:length(mblck)){
  if (mblck[i] == 0)
    type.mblck[i] = "Not black"
  else
    type.mblck[i] = "Black"
}
```

```
type.data$type.mblck<-type.mblck
type.data
typemblckdf<-type.data%>%
  group_by(type.mblck)%>%
  summarise(count = n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal))
typemblckdf
mblckpie<-ggplot(typemblckdf, aes(x="",y=ratioVal,fill=type.mblck))+
  theme_bw()+
  geom_bar(width = 2, stat = "identity", color="white") +
  coord_polar("y", start=0)+
  ggtitle("Black mother ratio")+
  theme(plot.title=element_text(hjust=0.5, size=15),
        axis.title=element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.border = element_blank())+
   scale_fill_manual(name = "Black mother", labels = c("Black","Not
black"), values=c("#7fa5fb","#f8f2a6"))+
  geom_text(aes(label=perc),position = position_stack(vjust = 0.5))
mblckpie
```

- Output:

```
> typemblckdf
# A tibble: 2 × 4
  type.mblck count ratioVal perc
  <chr>      <int>    <dbl> <chr>
1 Black        109   0.0595 6%
2 Not black   1723   0.941  94%
>
```



→ Analysis on mblck

- 94% of mothers are not black, only 6% of mothers are black.

- Most of the babies' mothers observed are not black.

3.2.3 lbw attribute (low birth weight)

- Low birth weight babies are those who have weight less than or equal to 2000 g.

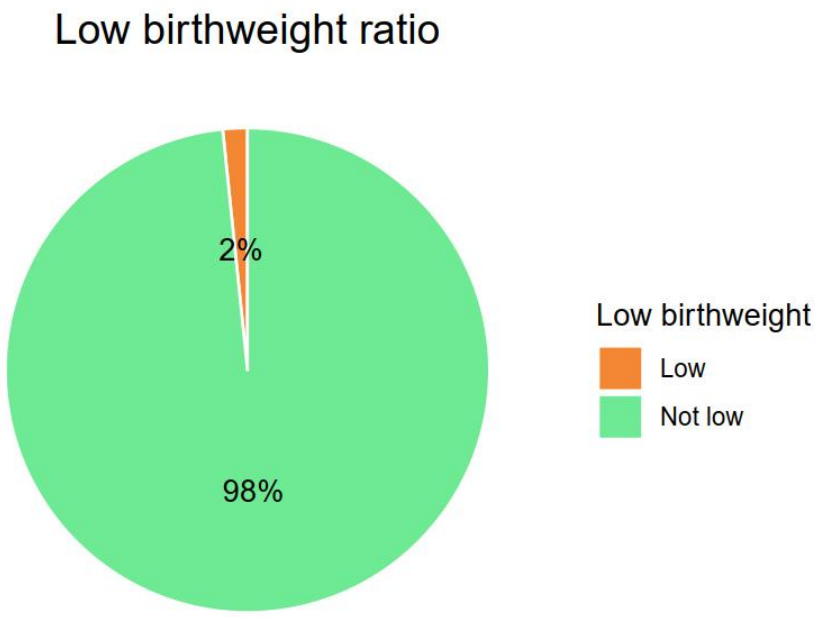
- Code:

Create type.lbw, change levels, create typelbwdf, draw pie chart of lbw attribute

```
lbwpie<-ggplot(typelbwdf, aes(x="",y=ratioVal,fill=type.lbw))+
  theme_bw()+
  geom_bar(width = 2, stat = "identity", color="white") +
  coord_polar("y", start=0)+
  ggtitle("Low birthweight ratio")+
  theme(plot.title=element_text(hjust=0.5, size=15),
        axis.title=element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
```

```
        panel.grid = element_blank(),
        panel.border = element_blank())+
  scale_fill_manual(name = "Low birthweight", labels = c("Low","Not low"),
values=c("#f48833","#6fea95"))+
  geom_text(aes(label=perc),position = position_stack(vjust = 0.5))
lbwpie
```

- Output:



→ ***Analysis on lbw:***

- 98% of babies don't have low birth weight, only 2% of babies has low birth weight (<=2000g).

- Most of the babies observed don't have low birth weight.

3.2.4 vlbw attribute (very low birth weight)

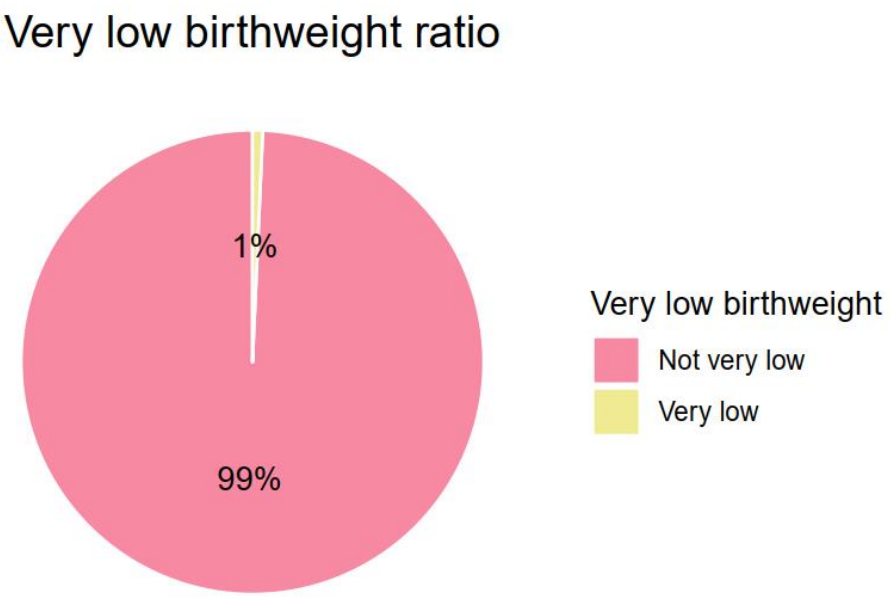
- Very low birth weight babies are those who have weight less than or equal to 1500 g.
- Code:

Create type.vlbw, change levels, create typevlbwdf, draw pie chart of vlbw attribute

```
typevlbwdf
vlbwpie<-ggplot(typevlbwdf, aes(x="",y=ratioVal,fill=type.vlbw))+
  theme_bw()+
  geom_bar(width = 2, stat = "identity", color="white") +
  coord_polar("y", start=0)+
  ggtitle("Very low birthweight ratio")+
  theme(plot.title=element_text(hjust=0.5, size=15),
        axis.title=element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.border = element_blank())+
  scale_fill_manual(name = "Very low birthweight", labels = c("Not very
low","Very low"), values=c("#f98aa4", "#f1eb92"))+
  geom_text(aes(label=perc),position = position_stack(vjust = 0.5))
```

- Output:

```
> typevlbwdf
# A tibble: 2 × 4
  type.vlbw    count ratioVal perc
  <chr>        <int>    <dbl> <chr>
1 Not very low  1819  0.993   99%
2 Very low        13  0.00710 1%
>
```



→ ***Analysis on vlbw:***

- 99% of babies don't have very low birth weight, only 1% of babies has very low birth weight (<=1500g).

- Most of the babies observed don't have very low birth weight.

=> Analysis on bwght (birth weight)

Most of the babies observed in dataset have normal birth weight. This may be hard for us to predict if the bad factors cause low birth weight in baby. (This is just prediction because we haven't analyzed all factors up to this point)

3.3 Qualitative – Qualitative

3.3.1 omaps – lbw (1-minute APGAR – low birth weight)

a. Qualitate omaps

- Research has shown that APGAR score has strong association with birth weight, which means low APGAR score leads to low birth weight.

(Source: https://bmcpediatr.biomedcentral.com/articles/10.1186/s12887-021-02745-6)

- We need to know if omaps or fmaps has higher impact on birth weight
- As mentioned in section 1.2, omaps and fmaps range helps us to measure baby's health.
- omaps can be divided into 3 ranges of values:
  - 0 - 3: Lifesaving measures
  - 4 - 6: Assistance for breathing
  - 7 - 10: routine post-delivery care
- Code:

```
type.omaps = 1:length(omaps);
for (i in 1:length(omaps)){
  if (omaps[i]>=0 && omaps[i]<=3)
    type.omaps[i] = "1. Lifesaving"
  else if (omaps[i]>=4 && omaps[i]<=6)
    type.omaps[i] = "2. Breathing assistance"
  else type.omaps[i] = "3. Routine care"
}
type.data$type.omaps<-type.omaps
typeomapsdf<-type.data%>%
  group_by(type.omaps)%>%
  summarise(count = n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal))
typeomapsdf
```

- Output:

```
# A tibble: 3 × 4
  type.omaps              count ratioVal perc
  <chr>                   <int>    <dbl> <chr>
1 1. Lifesaving              32   0.0175 1.7%
2 2. Breathing assistance    58   0.0317 3.2%
3 3. Routine care          1742   0.951  95.1%
>
```

b. omaps vs lbw

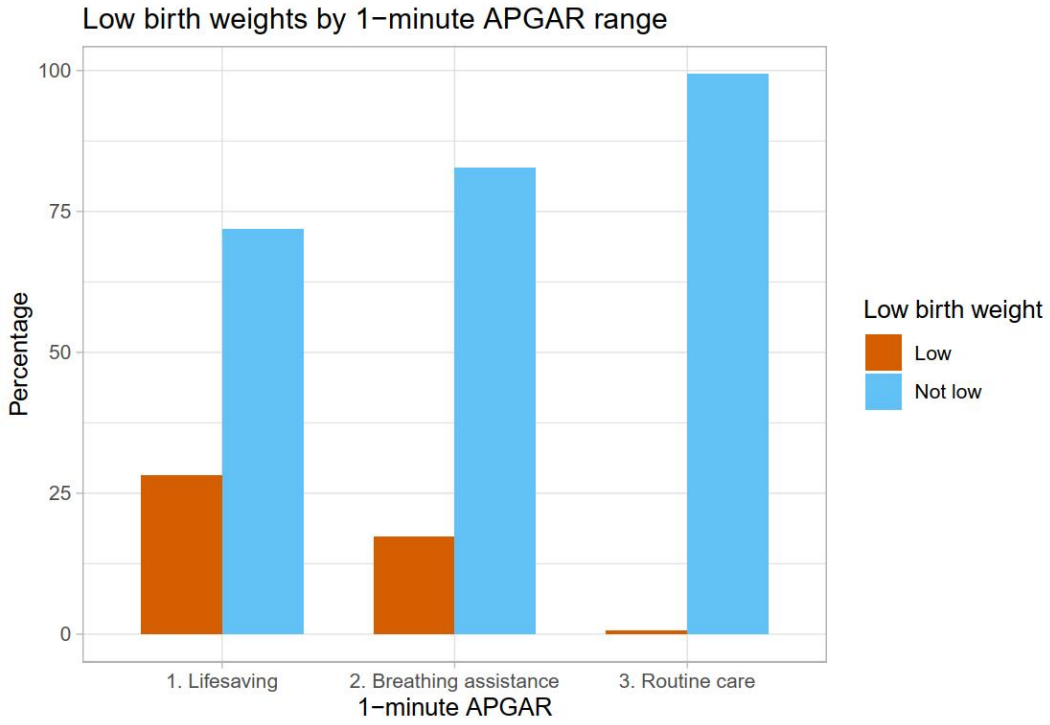
- Code:

```
#Group by omaps and lbw
omaps_lbw<-type.data %>%
  group_by(type.omaps, type.lbw) %>%
  summarise(count=n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal)) #%>%
omaps_lbw
omaps_lbw_bar<-ggplot(omaps_lbw,  aes(x=type.omaps,  y=ratioVal*100,
fill=type.lbw))+
  theme_light()+
  geom_col(width=0.7,
           position=position_dodge(0.7))+
  xlab("1-minute APGAR")+
  ylab("Percentage")+
  ggtitle("Low birth weights by 1-minute APGAR range")+
   scale_fill_manual(name="Low  birth  weight",  labels=c("Low","Not
low"),values=c("#D55E00","#63c1f5"))
omaps_lbw_bar
```

- Output:

```
  type.omaps              type.lbw count ratio…¹ perc
    <chr>                   <chr>    <int>   <dbl> <chr>
  1 1. Lifesaving           Low          9 0.281   28%
  2 1. Lifesaving           Not low     23 0.719   72%
  3 2. Breathing assistance Low         10 0.172   17%
  4 2. Breathing assistance Not low     48 0.828   83%
  5 3. Routine care         Low         11 0.00631 1%
  6 3. Routine care         Not low   1731 0.994   99%
  # … with abbreviated variable name ¹ratioVal
```

Bar chart that represents the relationship between omaps and lbw ***(use percentage)***



→ Analysis on omaps and lbw

- The lower the 1-minute APGAR, the more percentage of babies have low weight:

+ Lifesaving: 28% is low

+ Breathing assistance: 17% is low

+ Routine care: 1% is low

- Thus, low 1-minute APGAR score causes higher percentage of babies to have low birth weight, but we can't guarantee that the number of low birth weight is bigger than that of normal birth weight at each level.

- Since most of the observed babies are at routine care level (1742/1832), the number of low and normal birth weight at this omaps level is the highest among 3 levels (refer to omaps_lbw)

3.3.2 fmaps – lbw (5-minute APGAR – low birth weight)

a. Qualitate fmaps

- fmaps can be divided into 3 ranges of values:
  - 0 - 3: Low
  - 4 - 6: Moderately abnormal
  - 7 - 10: Normal

- Code:

```
type.fmaps = 1:length(fmaps);
for (i in 1:length(fmaps)){
  if (omaps[i]>=0 && fmaps[i]<=3)
    type.fmaps[i] = "1. Low"
  else if (fmaps[i]>=4 && fmaps[i]<=6)
    type.fmaps[i] = "2. Moderately abnormal"
  else type.fmaps[i] = "3. Normal"
}
type.data$type.fmaps<-type.fmaps
type.data
typefmapsdf<-type.data%>%
  group_by(type.fmaps)%>%
  summarise(count = n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal))
typefmapsdf
```

- Output:

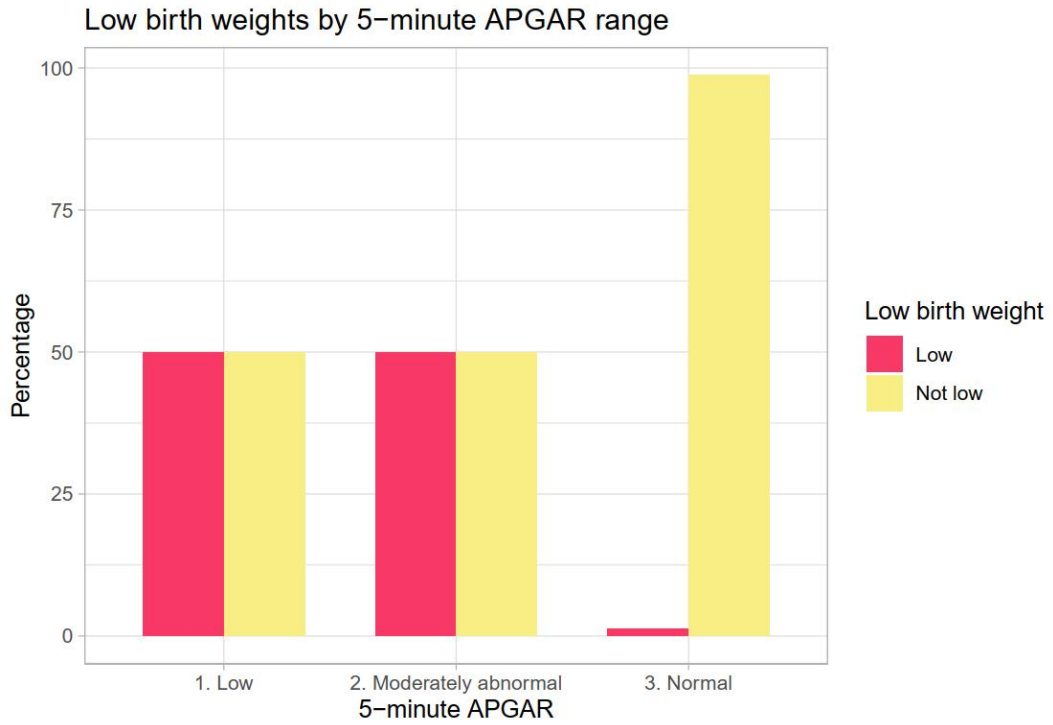
```
> typefmapsdf
# A tibble: 3 × 4
  type.fmaps             count ratioVal perc
  <chr>                  <int>    <dbl> <chr>
1 1. Low                     4  0.00218 0.22%
2 2. Moderately abnormal    10  0.00546 0.55%
3 3. Normal               1818  0.992   99.24%
>
```

b. fmaps vs lbw

Similar as omaps vs lbw, we create fmaps_lbw to count number of babies have low weight or normal weight according to their 5-minute APGAR score type. Here is the output:

type.fmaps	type.lbw	count	ratioVal	perc
<chr>	<chr>	<int>	<dbl>	<chr>
1 1. Low	Low	2	0.5	50%
2 1. Low	Not low	2	0.5	50%
3 2. Moderately abnormal	Low	5	0.5	50%
4 2. Moderately abnormal	Not low	5	0.5	50%
5 3. Normal	Low	23	0.0127	1%
6 3. Normal	Not low	1795	0.987	99%

Bar chart that represents the relationship between fmaps and lbw ***(use percentage)***



→ Analysis on fmaps and lbw

- The percentage of low and normal birth weight at low and moderately abnormal 5-minute APGAR level are the same (50% low birthweight – 50% normal birthweight).

- Only 1% of babies at normal 5-minute APGAR level.

- The number of normal birth weights at normal fmaps level is the highest among 3 fmaps levels (1795/1832), so the number of low and normal birth weight at this fmaps level is the highest among 3 levels (refer to fmaps_lbw)

=> Conclusion:

- The percentage of low birth weight at normal 5-minute APGAR level

- We can conclude that 5-minute APGAR score affects baby's birth weight as our initial prediction. However, there is unbalance in data among all fmaps level:

- The number of babies at normal fmaps level is much higher than that of babies at low and moderately abnormal level
- The percentage of babies at low and moderately abnormal level are small and quite close to each other (refer to the perc column typefmapsdf: low 0.22% - moderately abnormal level 0.55% compared to 99.23% of normal level)
- Based on the percentage (barchart) and frequency (fmaps_lbw) of low birth weight at each level, we can only conclude that the higher the fmaps level, the more normal birth weight babies.

3.3.3 male – lbw (male – low birth weight)

- Code:

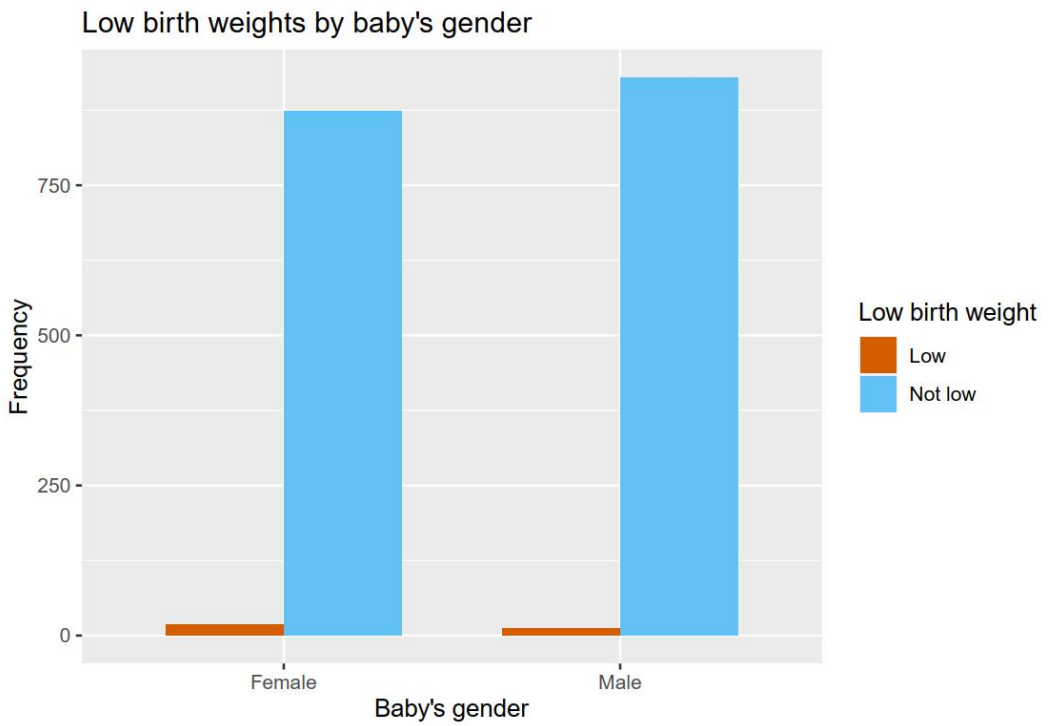
```
male_lbw<-type.data %>%
  group_by(type.male, type.lbw) %>%
  summarise(count=n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
  mutate(perc=scales::percent(ratioVal))
male_lbw
male_lbw_bar<-ggplot(male_lbw, aes(x=type.male, y=count,
fill=type.lbw))+
  geom_col(width=0.7,
           position=position_dodge(0.7))+
  xlab("Baby's gender")+
  ylab("Frequency")+
```

```
  ggtitle("Low birth weights by baby's gender")+
  scale_fill_manual(name="Gender",
labels=c("Male","Female"),values=c("#D55E00","#63c1f5"))
male_lbw_bar
```

- Output:

```
> male_lbw
# A tibble: 4 × 5
# Groups:   type.male [2]
  type.male type.lbw count ratioVal perc
  <chr>     <chr>    <int>    <dbl> <chr>
1 Female    Low         18   0.0202 2%
2 Female    Not low    873   0.980  98%
3 Male      Low         12   0.0128 1%
4 Male      Not low    929   0.987  99%
```

Bar chart that represents the relationship between male and lbw ***(use frequency)***



→ Analysis on male and lbw

- The number of male and female low birth weight are nearly the same.

- The number of female normal birth weight seems to be fewer than that of male normal birth weight, and ***the proportion of female normal birth weight are slightly lower than that of normal birth weight*** (output of male_lbw: Female + Not low: 98%, Male + Not low: 99%)

- Therefore, we can conclude that babies' gender does not have much impact on their weights (has slightly more impact on male in this data). This is reasonable since the number of female babies is slightly lower than that of male babies.

3.3.4 mbck – lbw (black mother – low birth weight)

- African-American mothers tend to have low birth weighted babies. This is shown by Stanford Children's Health and UNICEF.

Source:

https://www.stanfordchildrens.org/en/topic/default?id=low-birthweight-90-P02382

https://data.unicef.org/topic/nutrition/low-birthweight/

- Hence, our prediction is black mother having low weighted babies. We need to analyze the relationship between mblck and lbw.

- Code:

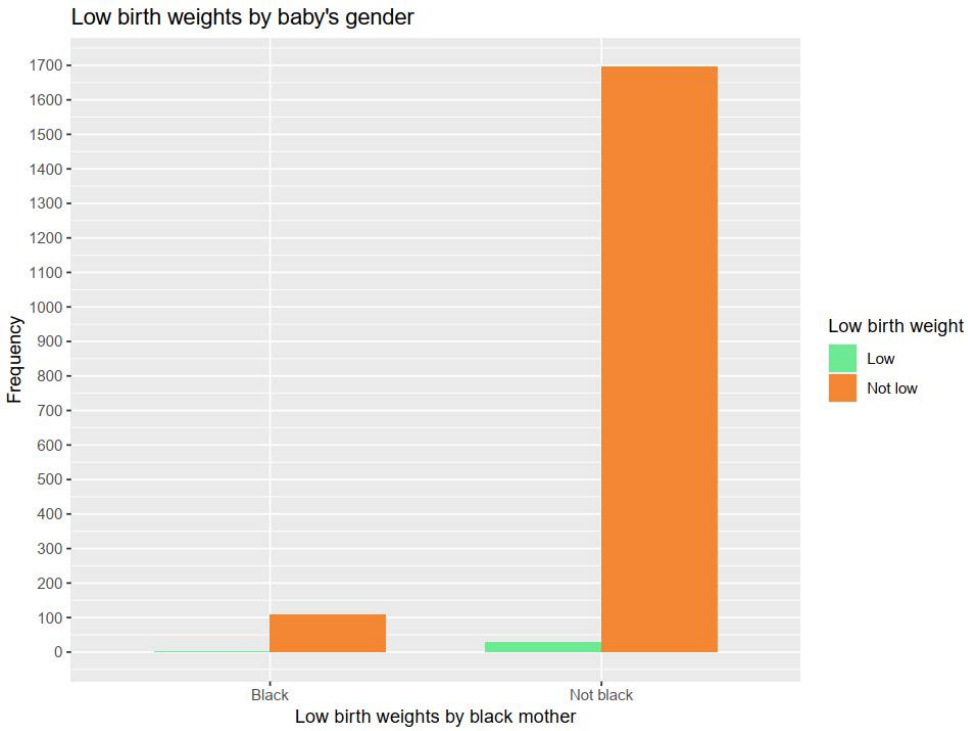
mblck_lbw is found similar as male_lbw. The code below is plotting mblck_lbw:
Since some values of mblck_lbw is quite small, we ***break the y axis into smaller units*** to easily observe.

```
mblck_lbw
mblck_lbw_bar<-ggplot(mblck_lbw, aes(x=type.mblck, y=count,
fill=type.lbw))+
  theme_grey()+
  geom_col(width=0.7,
           position=position_dodge(0.7))+
```

```
  xlab("Low birth weights by black mother")+
  ylab("Frequency")+
  ggtitle("Low birth weights by baby's gender")+
  scale_fill_manual(name="Low birth weight", labels=c("Low","Not
low"),values=c("#6fea95","#f48833"))+
  #break into smaller units to easily observe the small value
  scale_y_continuous(breaks = seq(0, 1750, by = 100))
mblck_lbw_bar
```

- Output:

```
> mblck_lbw
# A tibble: 4 × 5
# Groups:   type.mblck [2]
  type.mblck type.lbw count ratioVal perc
  <chr>      <chr>    <int>    <dbl> <chr>
1 Black      Low          2   0.0183 2%
2 Black      Not low    107   0.982  98%
3 Not black  Low         28   0.0163 2%
4 Not black  Not low   1695   0.984  98%
>
```

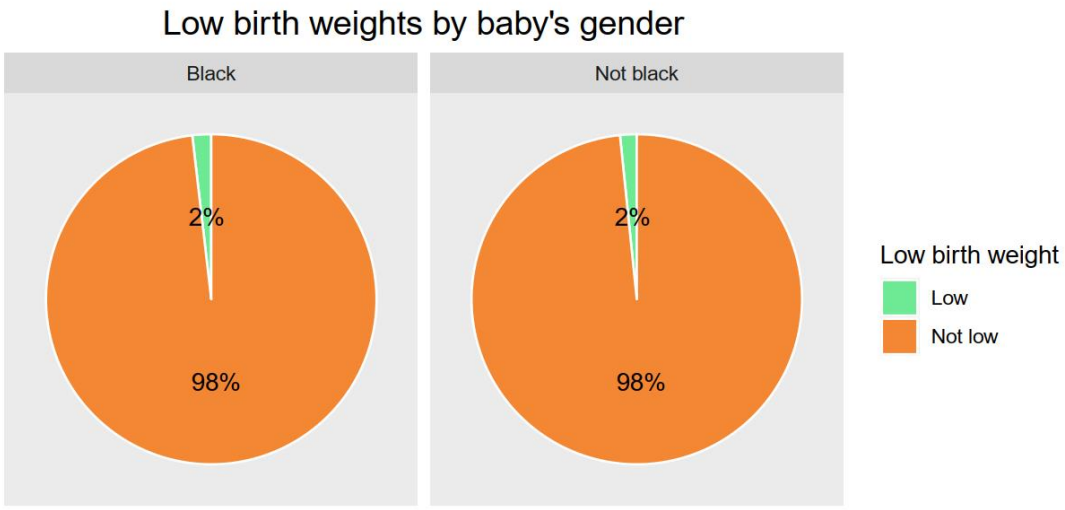


→ Analysis on mblck and lbw

- By merely looking at this bar chart, we may think that non-black mothers tend to have low weighted babies.

- In fact, it's because there are more non-black mothers than black mothers in our data (94% non-black, 6% black)

- By visualizing bar plot using 2 pie charts, we can see clearly that the proportion of low weighted babies are **nearly the same** for both black and non-black mothers. Because the scales::percent rounds to 1 digit after decimal point so we have 2% low - 98% not low.



- Therefore, we can only conclude that black mother has nearly the same percentage of low weighted babies as non-black mothers.

3.4 Quantitative variables

- We use histogram, boxplot and summarize analyze each of the following attributes

- Libraries used for analyzing quantitative variables:

+ ggplot2: draw chart

+ boxplot.stats: indentify outliers

+ DescTools: find all modes and their frequency

+ moments: find skewness of data

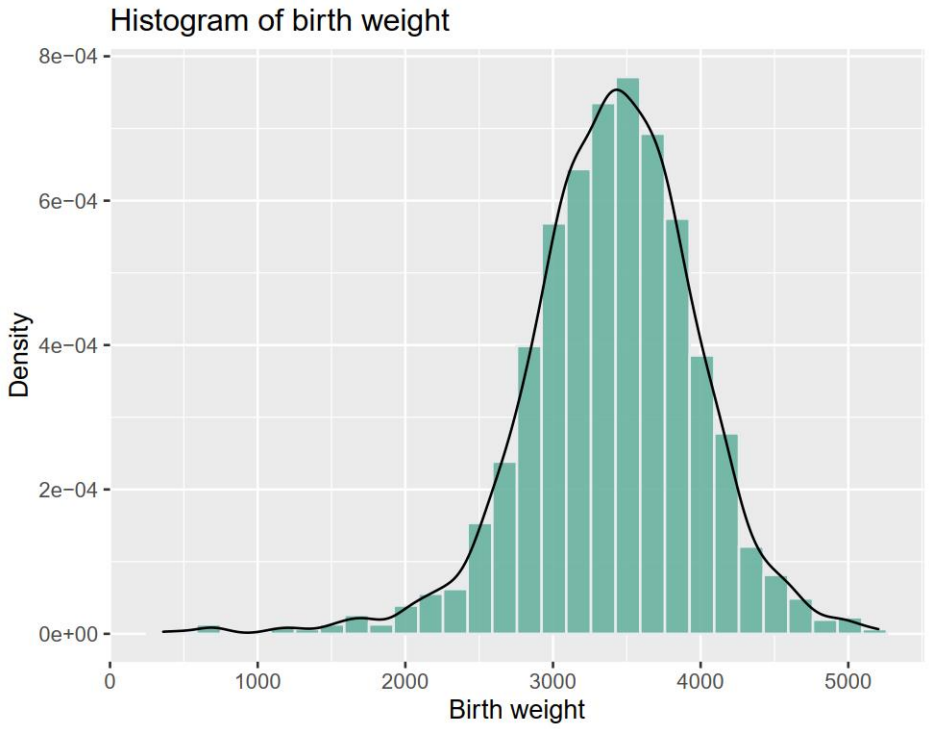
3.4.1 bwght attribute (birth weight)

- Histogram of birth weight

+ Code:

```
bwghtHist<-ggplot(df,aes(x=bwght)) +
geom_histogram(aes(y=..density..),fill="#69b3a2", color="#e9ecef",
alpha=0.9) +
geom_density(alpha=0.9) +
xlab("Birth weight")+
ylab("Density")+
ggtitle("Histogram of birth weight")
bwghtHist
```

+ Output:



- Find skewness of bwght

Use moments library

+ Code:

```
library(moments)
skewness(bwght)
```

+ Output:

```
> skewness(bwght)
[1] -0.600648
```

→ Analysis on histogram of bwght and the skewness of data:

- The birth weight is normally distributed.
- We can see clearly some outliers that are smaller than min. This is reasonable since the histogram is negatively skewed.

- The skewness is approximately -0.6. It is acceptable if we ignore the outliers. However, we still need to analyze more on the outliers of bwght.

- Summarize birth weight:

+ Code:

```
summary(bwght)
```

+ Output:

```
> summary(df$bwght)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    360    3076    3425    3401    3770    5204
>
```

→ Analysis on statistical summary of bwght

- Birth weight ranges from 360 g to 5204 g
- The mean birth weight is 3401 g

- Find mode of birth weight:

+ Code:

```
library ("DescTools")
modbwght <- Mode(df$bwght)
print(modbwght)
```

+ Output:

```
> print(modbwght)
[1] 3600
attr(,"freq")
[1] 24
>
```

→ Analysis on mode of bwght

- The mode of bwght is 3600 g, with 24 occurrences.

- Therefore, the peak of birth weight density curve (normal curve) is at 3600 g

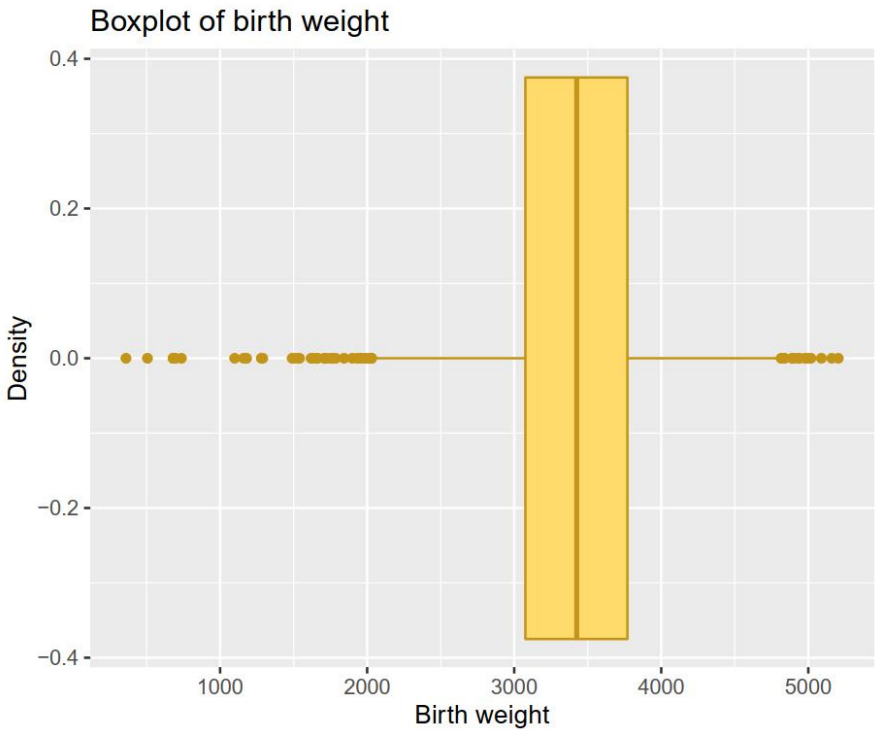
- Boxplot of birth weight:

We use boxplot to identify the outliers much easier.

+ Code:

```
bwghtBox<-ggplot(df,aes(x=bwght)) +
  geom_boxplot(fill = "#FFDB6D", color = "#C4961A") +
  xlab("Birth weight")+
  ylab("Density")+
  ggtitle("Boxplot of birth weight")
bwghtBox
```

+ Output:



→ Analysis on boxplot of birth weight

- There are outliers from both cases: smaller than min and larger than max.

- Most of the outliers are smaller than min.

- Find the exact outliers and total number of outliers

Use boxplot.stats to find all outliers.

+ Code:

```
outlierBwght<-boxplot.stats(bwght)$out
outlierBwght
length(outlierBwght)
```

+ Output:

```
> outlierBwght
 [1] 4990  697 4900 1720 4940  506 4890 1710 4933  680 4840 1766 1180 1710
[15] 1160 2010 1280 4980 1984 5204  681  360 4815 1956 5160 1660 5018 1531
```

```
[29] 1843 2030 1290  737 1540 1099 2030 4987 1500 5089 1934 2023 1899 1660
[43] 1786 1630 1490 1750 1619
> length(outlierBwght)
[1] 47
>
```

→ Analysis on detailed outliers of bwght:

- Most outliers are smaller than min (<=2030 g).

- There are 47 outliers in total.

3.4.2 lbwght attribute (log of birth weight)

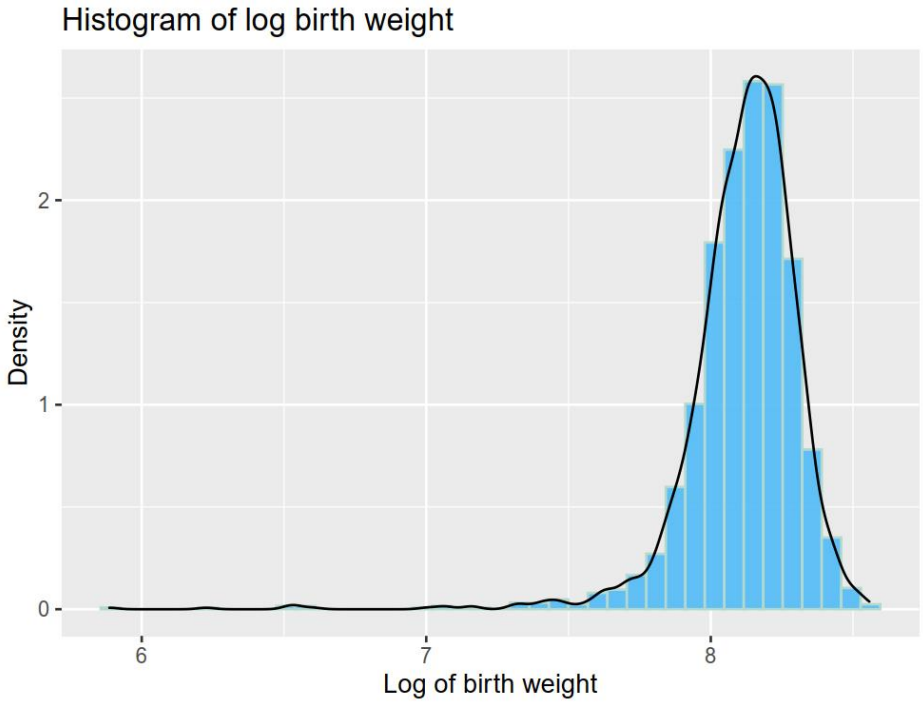
Log of birth weight (lbwght) is provided in the data. We will plot histogram of lbwght to know if it is “better” than (birth weight) bwght

- Histogram of log birth weight

+ Code:

```
lbwghtHist<-ggplot(df,aes(x=lbwght)) +
  geom_histogram(bins=40, aes(y=..density..),fill="#51bdf8",
color="#b4d9d1", alpha=0.9) +
  geom_density(alpha=0.9) +
  xlab("Log of birth weight")+
  ylab("Density")+
  ggtitle("Histogram of birth weight")
lbwghtHist
```

+ Output:



- Find skewness of lbwght

Similar to finding skewness of bwght, we get the result

```
> skewness(lbwght)
[1] -2.951033
>
```

→ Analysis on histogram of lbwght and the skewness of data

- At the first glance, there are many outliers at the left-hand side.

- Log birth weight is normally distributed

- The density curve is negatively skewed

=> bwght in the previous part is "better" than lbwght since bwght has smaller skewness (Here we have skewness of -2.95)

- Summarize log birth weight:

+ Code:

```
summary(lbwght)
```

+ Output:

```
> summary(df$lbwght)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.886   8.031   8.139   8.114   8.235   8.557
>
```

➔ Analysis on statistical summary of lbwght

- Log birth weight ranges from 5.886 g to 8.557 g

- The mean log birth weight is 8.114 g

- Find mode of log birth weight:

+ Code:

```
library ("DescTools")
modlbwght <- Mode(df$lbwght)
print(modlbwght)
```

+ Output:

```
> print(modlbwght)
[1] 8.188689
attr(,"freq")
[1] 24
>
```

➔ Analysis on mode of lbwght

- The mode of lbwght is approximately 8.19, with 24 occurrences.

- Therefore, the peak of birth weight density curve (normal curve) is at 8.19

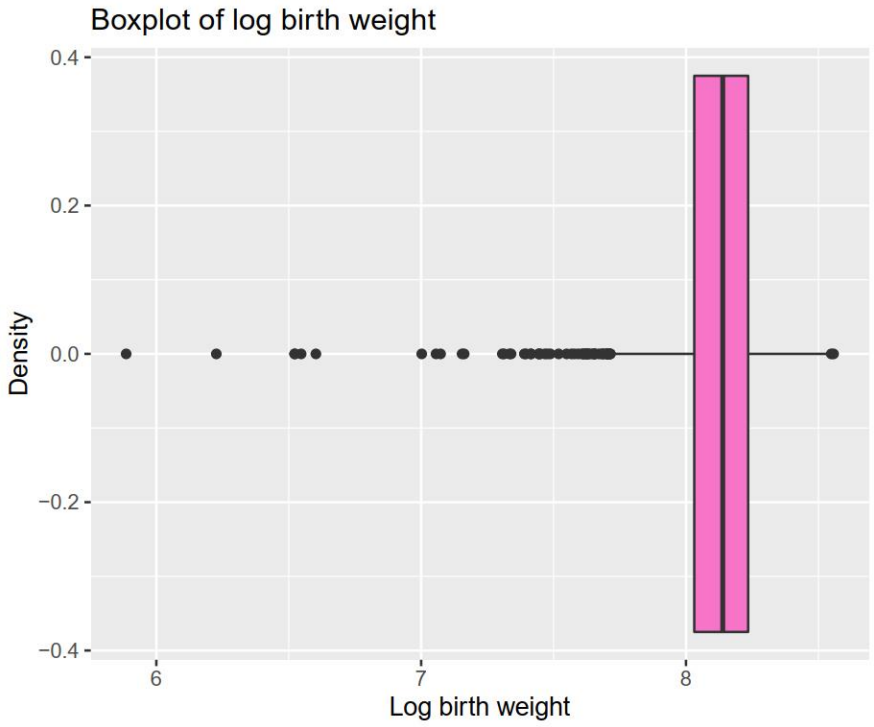
- Boxplot of log birth weight:

We use boxplot to identify the outliers much easier.

+ Code:

```
lbwghtBox<-ggplot(df,aes(x=lbwght)) +
  geom_boxplot(fill = c("#f774ca")) +
  xlab("Log birth weight")+
  ylab("Density")+
  ggtitle("Boxplot of log birth weight")
lbwghtBox
```

+ Output:



→ Analysis on boxplot of log birth weight

- There are outliers from both cases: smaller than min and larger than max.
- All the outliers are smaller than min except one is larger than max
- The range of data is small

- Find the exact outliers and total number of outliers

Use boxplot.stats to find all outliers.

+ Code:

```
outlierLbwght<-boxplot.stats(lbwght)$out
length(outlierLbwght)
```

+ Output:

```
> length(outlierLbwght)
[1] 58
>
```

→ Analysis on detailed outliers of lbwght:

- There are 58 outliers in total.

- Therefore, bwght is “better” than lbwght in terms of **total number of outliers**

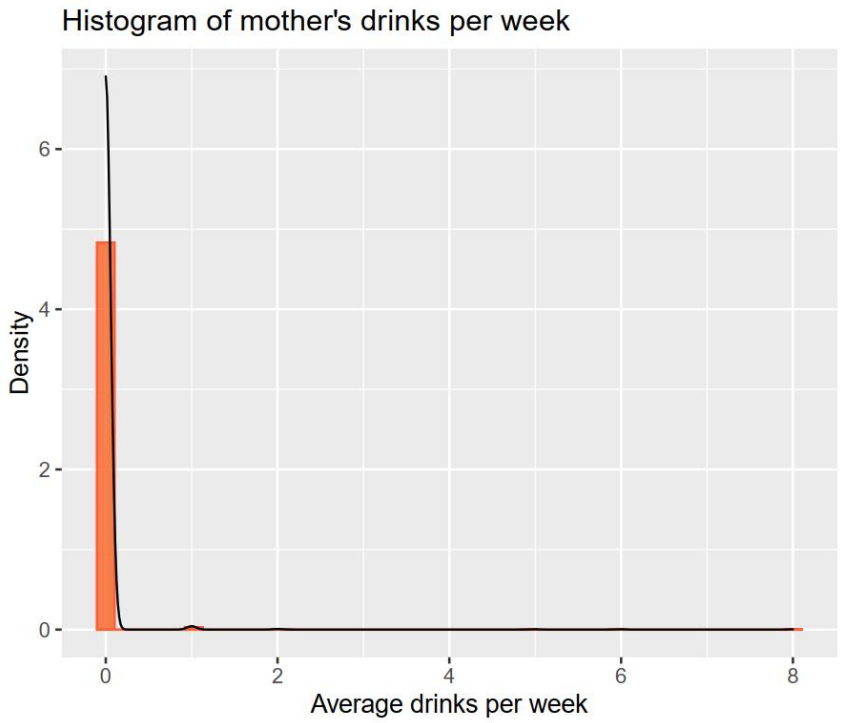
3.4.3 drink attribute (average mother’s drinks per week)

- Histogram of average mother’s drinks per week.

+ Code:

```
drinkHist<-ggplot(df,aes(x=drink)) +
  geom_histogram(bins=40, aes(y=..density..),fill="#f86423",
color="#ff5e39", alpha=0.8) +
  geom_density(size=0.46)+
  xlab("Average drinks per week")+
  ylab("Density")+
  ggtitle("Histogram of mother's drinks per week")
drinkHist
```

+ Output:



→ *Analysis on histogram of drink*

- The data is unbalanced, so it isn't normally distributed.
- Most of the observed mothers do not use alcohol. A few of them drink a bit each week.
- All values greater than 0 are considered outliers

- Summarize average mother's drinks per week:

+ Code:

```
summary(drink)
```

+ Output:

```
> summary(df$drink)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.01856 0.00000 8.00000
>
```

→ *Analysis on statistical summary of lbwght*

- Average drinks per week ranges from 0 to 8

- The mean average drinks per week is 0.01856

- Based on this summary, we see that drink is unbalanced since average is very close to min.

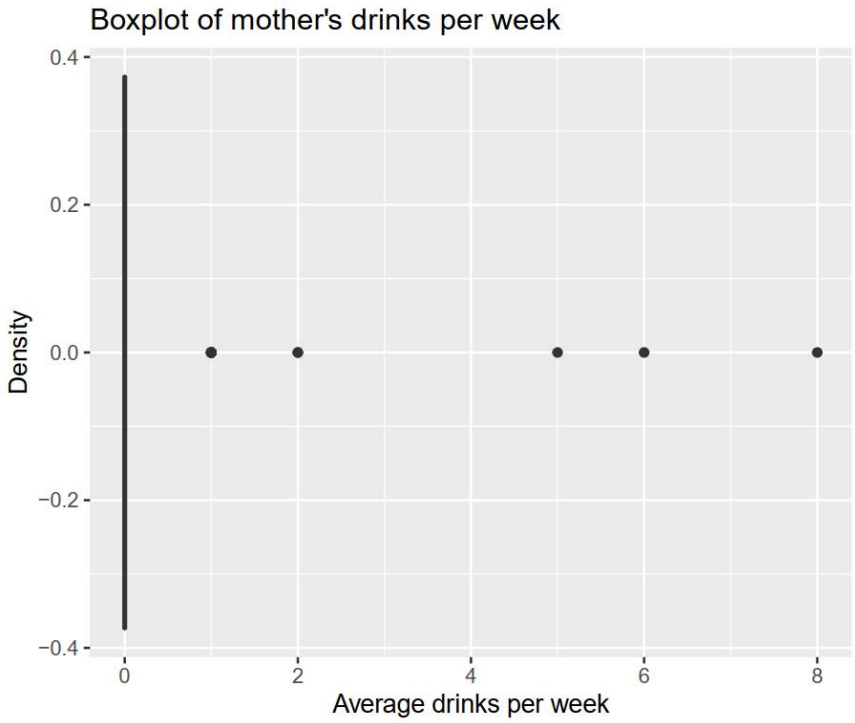
- Boxplot of average drinks per week:

We use boxplot to identify the outliers much easier.

+ Code:

```
drinkBox<-ggplot(df,aes(x=drink)) +
  geom_boxplot(fill = c("red")) +
  xlab("Average drinks per week")+
  ylab("Density")+
  ggtitle("Boxplot of mother's drinks per week ")
drinkBox
```

+ Output:



→ Analysis on boxplot of average drinks per week

- There are outliers from only one case: larger than max.

- The range of data is so small that boxplot cannot visualize median, min, max clearly.

- Find the exact outliers and total number of outliers

Use table to know the frequency of each value of drink, and boxplot.stats to find all outliers.

+ Code

```
table(drink)
outlierdrink<-boxplot.stats(drink)$out
outlierdrink
length(outlierdrink)
```

+ Output:

```
> table(drink)
drink
   0    1    2    5    6    8
1816   11    2    1    1    1
> outlierdrink<-boxplot.stats(drink)$out
> outlierdrink
 [1] 5 8 2 1 2 1 1 1 1 1 6 1 1 1 1 1
> length(outlierdrink)
[1] 16
>
```

→ Analysis on detailed outliers of drink:

- There are 16 outliers in total.

- All of the outliers are greater than 0.

- Because value 0 has most frequency (1816/1832); therefore, all the remaining values, which are greater than 0, are all outliers

3.5 Quantitative - Qualitative

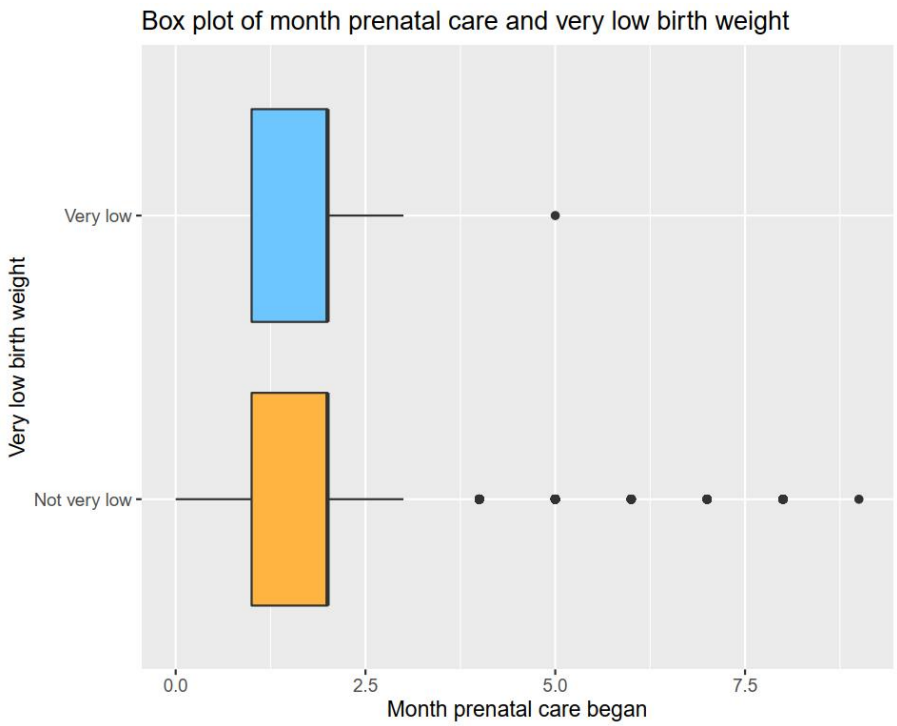
3.5.1 monpre – vlbw (month prenatal care began – very low birth weight)

Since month prenatal care began can affect baby's health, we will find the relationship between monpre and vlbw (very low birth weight) by visualizing boxplots

- Code:

```
monpre_vlbw_Box<-ggplot(type.data,aes(x=monpre,y = type.vlbw, fill =
type.vlbw)) +
  geom_boxplot(fill=c("#ffb441","#6ec7ff")) +
  xlab("Month prenatal care began") +
  ylab("Very low birth weight") +
  ggtitle("Box plot of month prenatal care and very low birth
weight") +
  theme(legend.position = "none")
monpre_vlbw_Box
```

- Output:



→ Analysis on boxplot:

- 1st quartile, 3rd quartile, and max of month prenatal care began are the same for both very low and not very low birth weight

- Min of monpre of very low birth weight seems to be close to 1st quartile since it's hard to see on boxplot.

- There is only 1 outlier of very low birth weight, while there are 6 outliers of not very low birth weight.

- If we refer to typevlbwdf, we see that 99% of babies have not very low birth weight. This may be the reason why monpre of not very low birth weight spreads out more than very low birth weight.

3.5.1 mage – male (mother's age – male babies)

Research has shown that older parents are more likely to have female babies.

(Source:

https://www.psychologytoday.com/us/blog/the-scientific-fundamentalist/201104/why-are-older-parents-more-likely-have-daughters

https://pubmed.ncbi.nlm.nih.gov/22025225/)

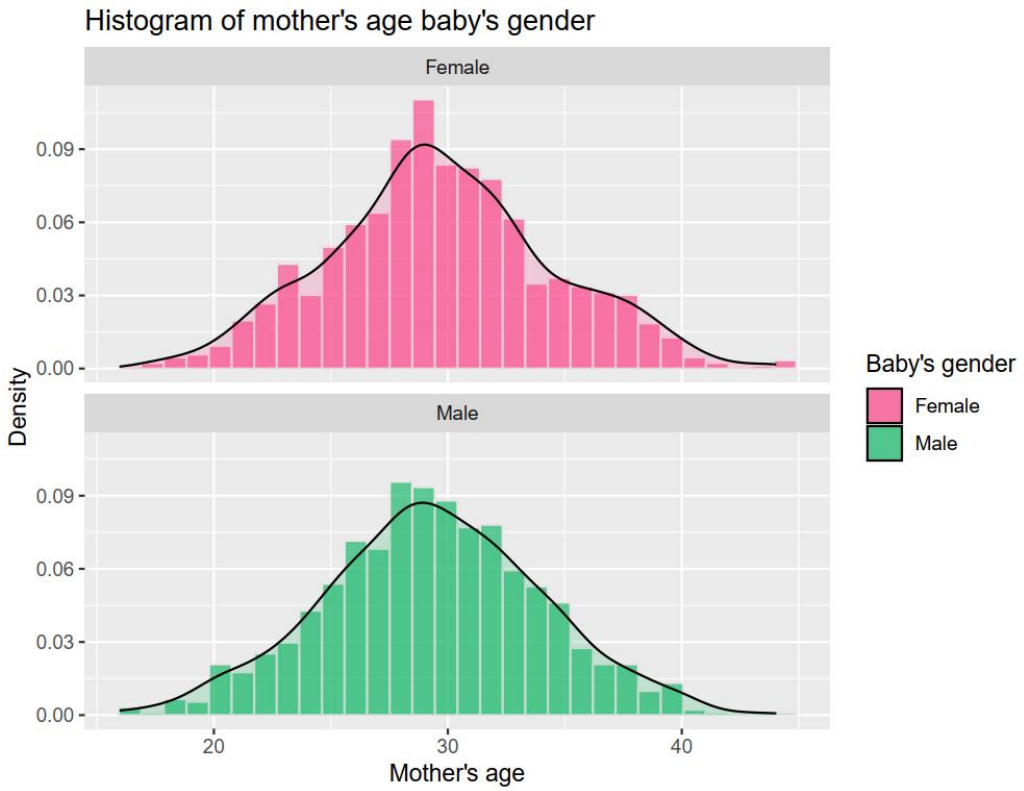
Therefore, we will visualize histogram to show the relationship between mother's age and male babies.

- Code:

```
mage_male_Hist<-ggplot(type.data,aes(x=mage, fill = type.male)) +
  geom_histogram(aes(y = ..density..), color="#eae6e9",alpha=0.7) +
  geom_density(alpha = 0.2) +
  facet_wrap(~type.male,ncol = 1,scale = "fixed")+
  xlab("Mother's age") +
  ylab("Density") +
  ggtitle("Histogram of mother's age baby's gender") +
```

```
  scale_fill_manual(name="Baby's gender", values =
c("#fd4d8f","#1ebb6e"))
mage_male_Hist
```

- Output:



→ Analysis on histogram

- Both mother's age of male and female babies are normally distributed

- Mean of mother's age of male and female babies are the same.

- This is reasonable since the proportion of male and female are close to each other (female 48.6% - male 51.4%)

4. Inferential statistics

4.1 Inferential statistics on quantitative variables

4.1.1 npvis attribute (total number of prenatal visits)

npvis is total number of prenatal visits, which means the number of prenatal care appointments for doctors, nurses and midwifes to take care of pregnant women.

- Comment on μ of npvis:

+ Code:

```
summary(npvis)
```

+ Output:

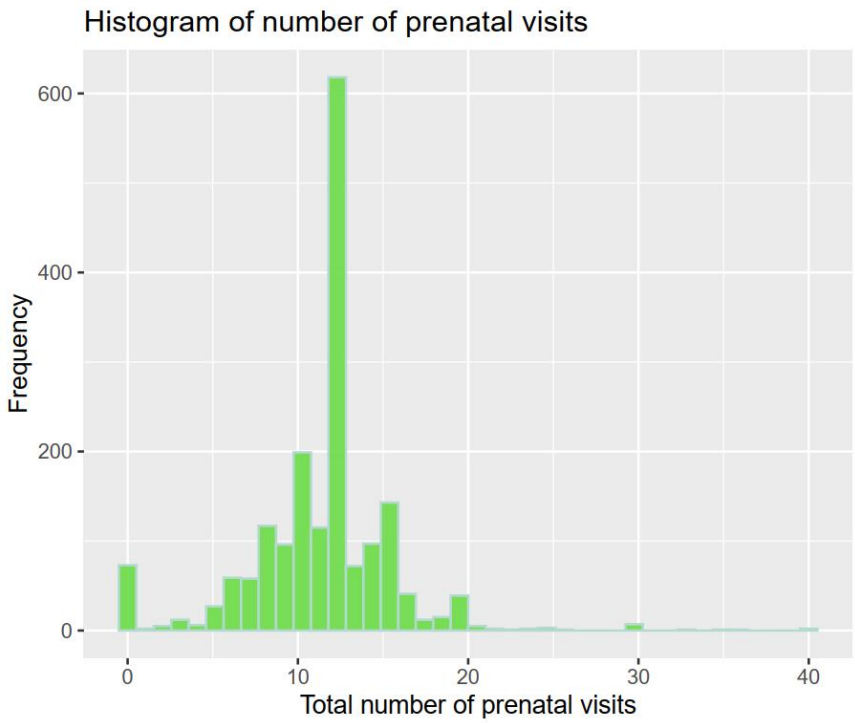
```
> summary(npvis)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   10.00   12.00   11.19   12.00   40.00
>
```

→ Mean of total number of prenatal visits is 11.19 visits

Also, by visualizing the histogram of npvis, we see that

+ mean μ is quite close to mode.

+ Since the mode is approximately 12, and its frequency is much higher than other values, it affects the mean. Hence, mean is close to 11.



- Hypothesis test for μ of npvis

+ Our prediction: mean of npvis is 20 visits (μ = 20)

+ Hypothesis:

H_0: μ = 20

H_1: μ < 20

+ We conduct a left-tailed t-test on mean of npvis at α = 0.05, df = N – 1 = 1832 – 1 = 1831 by using t.test function

- Code:

```
t.test(npvis, mu=20, alternative="less")
```

- Output

```
> t.test(npvis, mu=20, alternative="less")
    One Sample t-test
data:  npvis
t = -89.205, df = 1831, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is less than 20
95 percent confidence interval:
     -Inf 11.35142
sample estimates:
mean of x
 11.18886
>
```

Since the p-value < 2.2e-16 < 0.05, we **reject the null hypothesis**, which means the true mean of total number of prenatal visits is less than 20 visits at significance level α = 0.05.

4.1.2 mage attribute (mother's age)

- Comment on μ of mage:

+ Code:

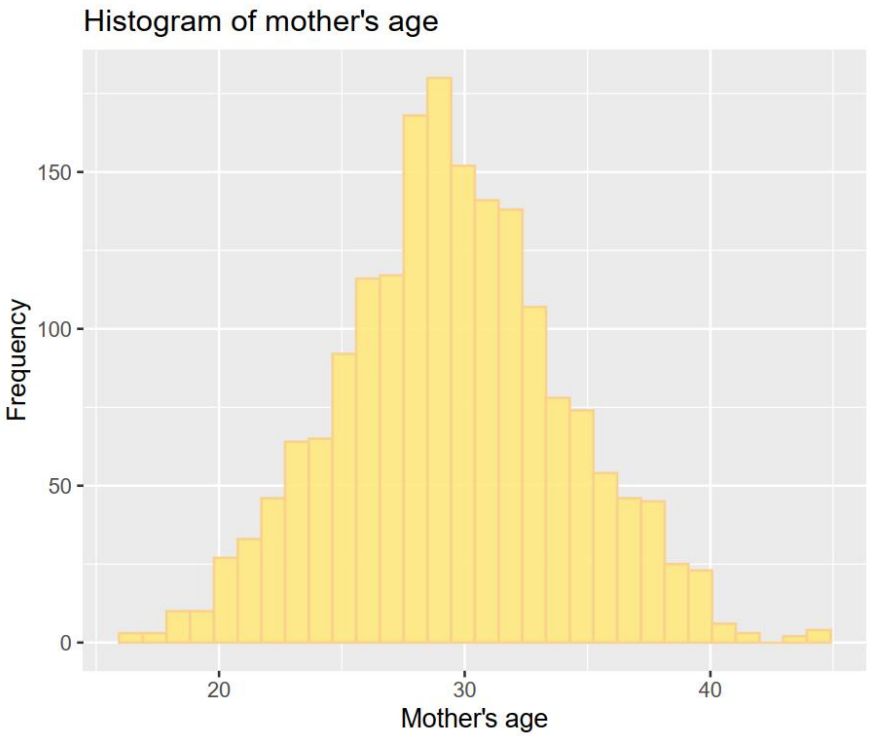
```
summary(mage)
```

+ Output:

```
> summary(mage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   26.00   29.00   29.56   33.00   44.00
```

→ Mean of mother's age is 29.56 years old

Also, by visualizing the histogram of mage, we see that the data is normally distributed. Therefore, we can know that μ is somewhere between 29 and 30 just by observing this histogram.



- Hypothesis test for μ of mage

+ Our prediction: mean of mother's age is 24 years old (μ = 24)

+ Hypothesis:

$$H_0: \mu = 24$$

$$H_1: \mu > 24$$

- We conduct a right-tailed t-test on mean of mage at α = 0.05, df = N – 1 = 1832 – 1 = 1831 by using t.test function

+ Code:

```
t.test(mage, mu=24, alternative="greater")
```

+ Output

```
> t.test(mage, mu=24, alternative="greater")
    One Sample t-test
data:  mage
```

```
t = 49.861, df = 1831, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 24
95 percent confidence interval:
 29.37442      Inf
sample estimates:
mean of x
 29.55786
>
```

Since the p-value < 2.2e-16 < 0.05, we **reject the null hypothesis**, which means the true mean of mother's age is greater than 24 years old at significance level α = 0.05.

4.2 Inferential statistics on qualitative variables

4.2.1 mwhte attribute (white mother)

mwhte = 1 means mother is white. We will test the proportion of white mother.

- Comment on proportion p of white mother:

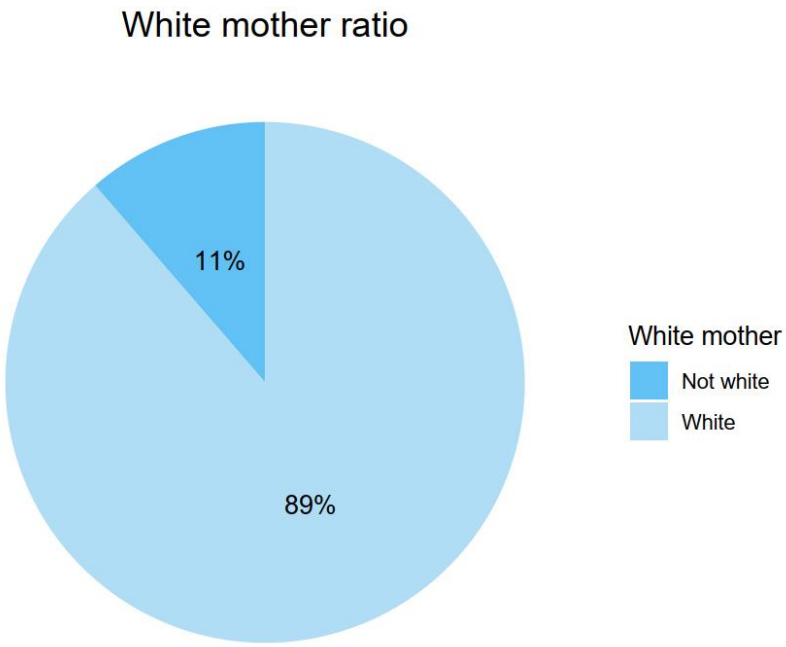
Creating type.mwhte, change levels, create typemwhtedf, visualize pie chart of mwhte attribute

+ Code:

```
type.mwhte = 1:length(mwhte);
for (i in 1:length(mwhte)){
  if (mwhte[i] == 0)
    type.mwhte[i] = "Not white"
  else
    type.mwhte[i] = "White"
}
type.data$type.mwhte<-type.mwhte
typemwhtedf<-type.data%>%
  group_by(type.mwhte)%>%
  summarise(count = n()) %>%
  mutate(ratioVal=count/sum(count)) %>%
```

```
  mutate(perc=scales::percent(ratioVal))
mwhtepie<-ggplot(typemwhtedf, aes(x="",y=ratioVal,fill=type.mwhte))+
  theme_light()+
  geom_bar(width = 2, stat = "identity") +
  coord_polar("y", start=0)+
  ggtitle("White mother ratio")+
  theme(plot.title=element_text(hjust=0.5, size=15),
        axis.title=element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        panel.grid = element_blank(),
        panel.border = element_blank())+
  scale_fill_manual(name = "White mother", labels = c("Not
white","White"), values=c("#63c1f5","#b1ddf5"))+
  geom_text(aes(label=perc),position = position_stack(vjust = 0.5))
mwhtepie
```

+ Output:



→ Most of the mothers are white. The proportion p of white mothers is 0.89.

- Hypothesis test for proportion p of white mother

+ Our prediction: proportion of white mother is 0.6 (p = 0.6)

+ Hypothesis:

H_0: p = 0.6

H_1: p > 0.6

+ Print out details of typemwhtedf to determine x and n:

- Code: `typemwhtedf`
- Output:

```
> typemwhtedf
# A tibble: 2 × 4
  type.mwhte count ratioVal perc
  <chr>      <int>    <dbl> <chr>
1 Not white    208    0.114 11%
2 White       1624    0.886 89%
>
```

→ x = 1624 (number of white mothers), n = 1832 (total number of observations)

+ We conduct a right-tailed test on proportion p at α = 0.05, x = 1624, n=1832 by using prop.test function

+ Code:

```
prop.test(x=1624, n=1832, p=0.6, alternative = "greater")
```

+ Output

```
> prop.test(x=1624, n=1832, p=0.6, alternative = "greater")
    1-sample proportions test with continuity correction
data:  1624 out of 1832, null probability 0.6
X-squared = 625.21, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.6
95 percent confidence interval:
 0.8734117 1.0000000
sample estimates:
        p
0.8864629
```

Since the p-value < 2.2e-16 < 0.05, we **reject the null hypothesis**, which means the true proportion of white mothers is greater than 0.6 at significance level α = 0.05.

4.2.2 vlbw attribute (very low birth weight)

vlbw = 1 means the baby has very low birth weight (weight ≤ 1500 g). We will test the proportion of very low birth weight.

- Comment on proportion p of very low birth weight:

Creating type.vlbw, change levels, create typevlbwdf, visualize pie chart of vlbw attribute

+ Code:

```
typevlbwdf
vlbwpie
```

+ Output:

```
> typevlbwdf
# A tibble: 2 × 4
  type.vlbw    count ratioVal perc
  <chr>        <int>    <dbl> <chr>
1 Not very low  1819  0.993   99%
2 Very low        13  0.00710 1%
>
```

→ Most of the babies do not have very low birth weight. The proportion p of very low birth weight is 0.01

- Hypothesis test for proportion p of white mother

+ Our prediction: proportion of very low birth weight is 0.08 (p = 0.08)

+ Hypothesis:

H_0: p = 0.08

H_1: p < 0.08

+ From data frame typevlbwdf, x = 13 (number of very low birth weight), n = 1832 (total number of observations)

+ We conduct a right-tailed test on proportion p at α = 0.05, x = 1624, n=1832 by using prop.test function

+ Code:

```
prop.test(x=13, n=1832, p=0.08, alternative="less")
```

+ Output

```
> prop.test(x=13, n=1832, p=0.08, alternative="less")
    1-sample proportions test with continuity correction
data:  13 out of 1832, null probability 0.08
X-squared = 131.31, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is less than 0.08
95 percent confidence interval:
 0.00000000 0.01145913
sample estimates:
         p
0.00709607
```

Since the p-value < 2.2e-16 < 0.05, we **reject the null hypothesis**, which means the true proportion of very low birth weight is less than 0.08 at significance level α = 0.05.

5. Linear regression

5.1 Split data into train and test set

We split into: Train 70% - Test 30% (set seed to reproduce later since this is random splitting)

```
set.seed(1)
sample<-sample(c(TRUE, FALSE), nrow(df), replace=TRUE,
prob=c(0.7,0.3))
train<-df[sample, ]
test<-df[!sample, ]
```

5.2 Simple linear regression model

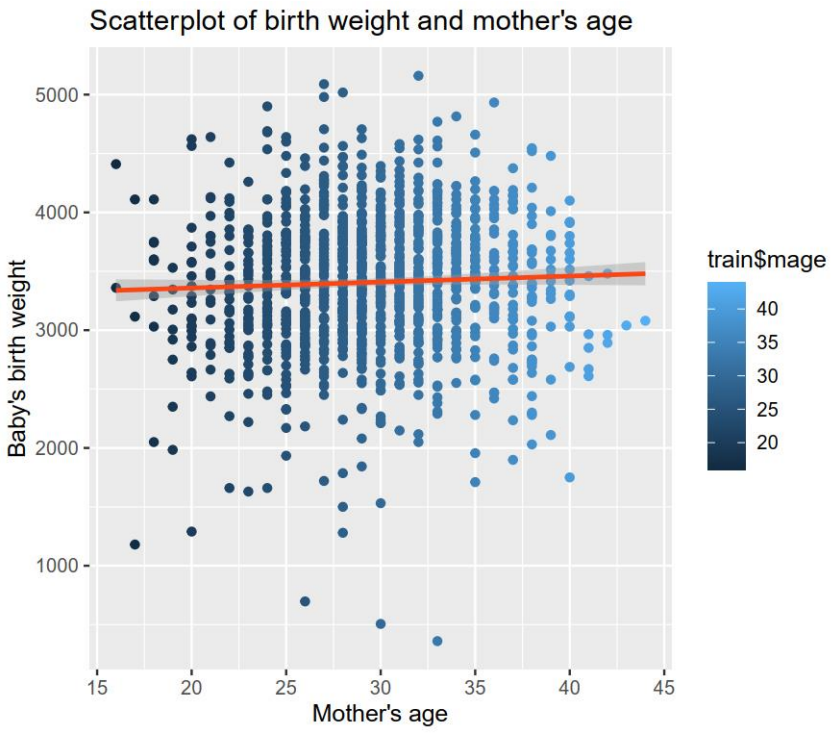
5.2.1 bwght and mage

- Visualize scatter plot + linear model

+ Code:

```
plotMage<-ggplot(train,aes(train$mage,train$bwght,color = train$mage))+
  geom_point(alpha = 1)+
  geom_smooth(method = "lm", color="#fd4716") +
  xlab("Mother's age") +
  ylab("Baby's birth weight")+
  ggtitle("Scatterplot of birth weight and mother's age")
plotMage
```

+ Output:



Note: red line is the linear model visualization, dark gray shaded is confidence interval

→ The range of confidence interval is large when the data is parse and the actual values are too small / too big

- Find linear model equation

$$bwght = \beta_1 + \beta_2 \times mage$$

+ Code:

```
train
model1<-lm(train$bwght~train$mage)
summary(model1)
```

+ Output:

```
> summary(model1)

Call:
lm(formula = train$bwght ~ train$mage)

Residuals:
     Min       1Q   Median       3Q      Max
-3063.89  -338.15     8.43   373.74  1741.16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3257.152     99.212  32.830   <2e-16 ***
train$mage     5.053      3.312   1.526    0.127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.1 on 1277 degrees of freedom
Multiple R-squared:  0.00182,   Adjusted R-squared:  0.001038
F-statistic: 2.328 on 1 and 1277 DF,  p-value: 0.1273

>
```

+ Model:

$bwght = 3257.152 + 5.053 \times mage + \varepsilon$

+ Meaning of coefficients:

$\beta_1 = 3257.152$ means when mother’s age is 0, birth weight is 3257.152 g

β_2 = 5.053 means when mother gets 1 age older, birth weight increases by 5.053 g

(R^2 = 0.00182)

- Confident interval of linear model

+ Code:

```
confint(model1)
```

+ Output:

```
> confint(model1)
                   2.5 %     97.5 %
(Intercept) 3062.515124 3451.78882
train$mage    -1.444024   11.54923
>
```

which means:

95% confident interval of

β_1 is (3062.515124, 3451.78882)

β_2 is (-1.444024, 11.54923)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction1<-predict(model1,data.frame(test$mage), interval =
"confidence")
prediction1
```

- Output:

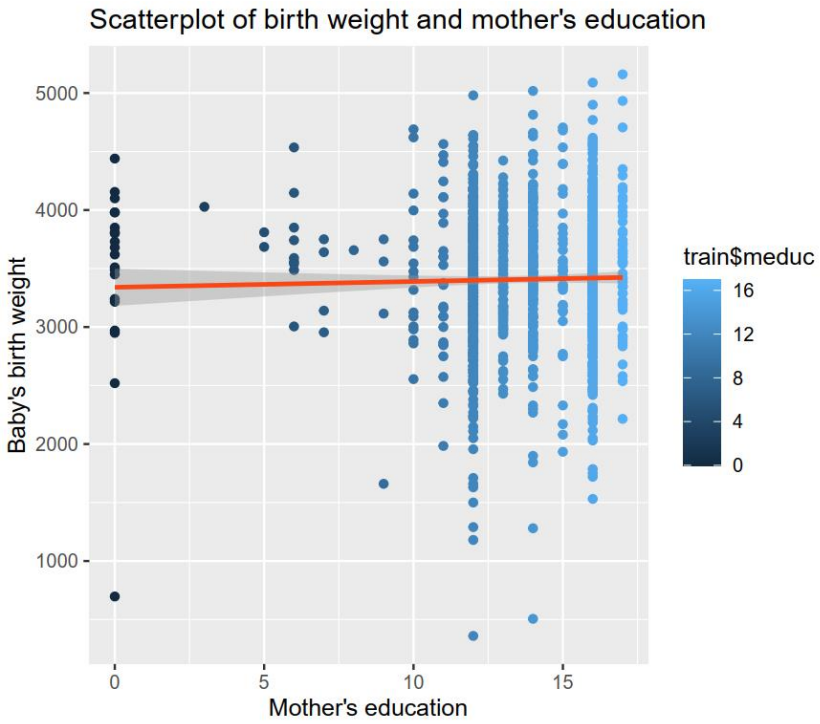
```
> prediction1
          fit      lwr      upr
1    3388.520 3349.659 3427.381
2    3403.677 3372.290 3435.065
```

```
3    3423.888 3385.587 3462.189
...
```

5.2.2 bwght and meduc

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The range of confidence interval is large when the actual values of data are small and shrinks when the actual values are big.

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times meduc$$

+ Code:

```
model2<-lm(train$bwght~train$meduc)
summary(model2)
```

+ Output:

```
> summary(model2)

Call:
lm(formula = train$bwght ~ train$meduc)

Residuals:
     Min       1Q   Median       3Q      Max
-3039.07  -340.54    18.15   371.93  1736.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3339.715     80.968  41.247   <2e-16 ***
train$meduc    4.946      5.873   0.842      0.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.5 on 1277 degrees of freedom
Multiple R-squared:  0.000555,  Adjusted R-squared:  -0.0002276
F-statistic: 0.7092 on 1 and 1277 DF,  p-value: 0.3999

>
```

+ Model:

$bwght = 3339.715 + 4.946 \times meduc + \varepsilon$

($R^2 = 0.000555$)

+ Meaning:

$\beta_1 = 3339.715$ means when mother’s education is 0, birth weight is 3339.715 g

$\beta_2 = 4.946$ means when mother’s education increases by 1 year, birth weight increases by 4.946 g

- Confident interval of each coefficient:

+ Code:

```
confint(model2)
```

+ Output

```
> confint(model2)
                  2.5 %     97.5 %
(Intercept) 3180.870153 3498.56067
train$meduc   -6.576329   16.46828
```

which means:

95% confident interval of

β_1 is (3180.870153, 3498.56067)

β_2 is (-6.576329, 16.46828)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction2<-predict(model2,data.frame(test$meduc), interval =
"confidence")
prediction2
```

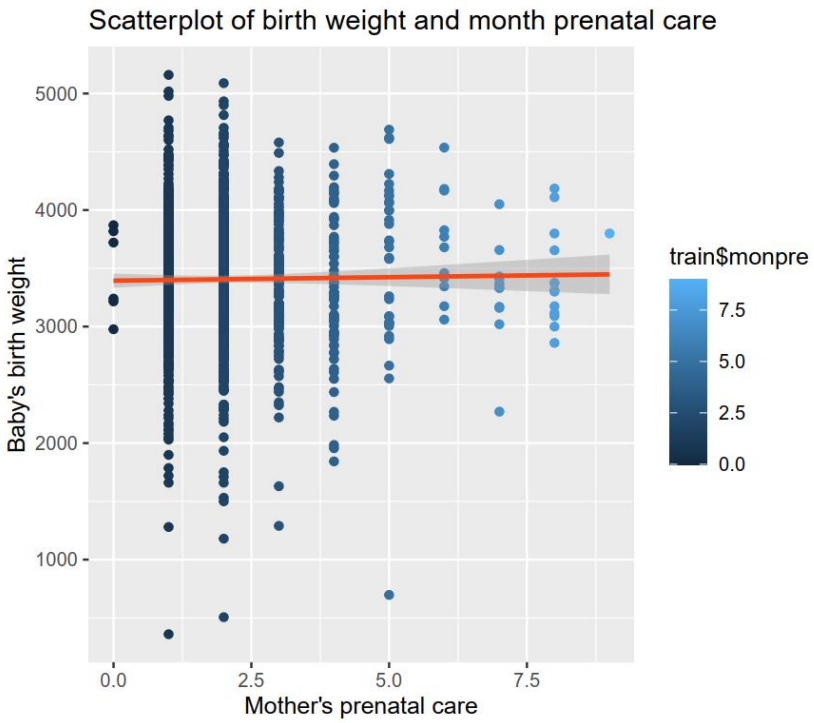
- Output:

```
> prediction2
         fit      lwr      upr
1    3399.067 3363.313 3434.821
2    3399.067 3363.313 3434.821
3    3399.067 3363.313 3434.821
...
```

5.2.3 bwght and monpre

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The range of confidence interval is large when the data is parse and the actual values are big.

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times monpre$$

+ Code:

```
model3<-lm(train$bwght~train$monpre)
summary(model3)
```

+ Output:

```
> summary(model3)

Call:
lm(formula = train$bwght ~ train$monpre)

Residuals:
     Min       1Q   Median       3Q      Max
-3039.71  -352.73    20.29   370.29  1760.29
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3393.702     31.000 109.476   <2e-16 ***
train$monpre    6.006     12.418   0.484    0.629
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.6 on 1277 degrees of freedom
Multiple R-squared:  0.0001831, Adjusted R-squared:  -0.0005998
F-statistic: 0.2339 on 1 and 1277 DF,  p-value: 0.6287

>
```

+ Model:

bwght = 3393.702 + 6.006 × *monpre* + **ε**

(R^2 = 0.0001831)

+ Meaning:

β_1 = 3393.702 means when month prenatal care is 0, birth weight is 3393.702 g

β_2 = 6.006 means when month prenatal care increases by 1 month, birth weight increases by 6.006 g

- Confident interval:

+ Code:

```
confint(model3)
```

+ Output

```
> confint(model3)
                  2.5 %     97.5 %
(Intercept)  3332.88639 3454.51786
train$monpre  -18.35515   30.36681
>
```

---

which means:

95% confident interval of

β_1 is (3332.88639, 3454.51786)

β_2 is (-18.35515, 30.36681)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction3<-predict(model3,data.frame(test$monpre), interval =
"confidence")
prediction3
```

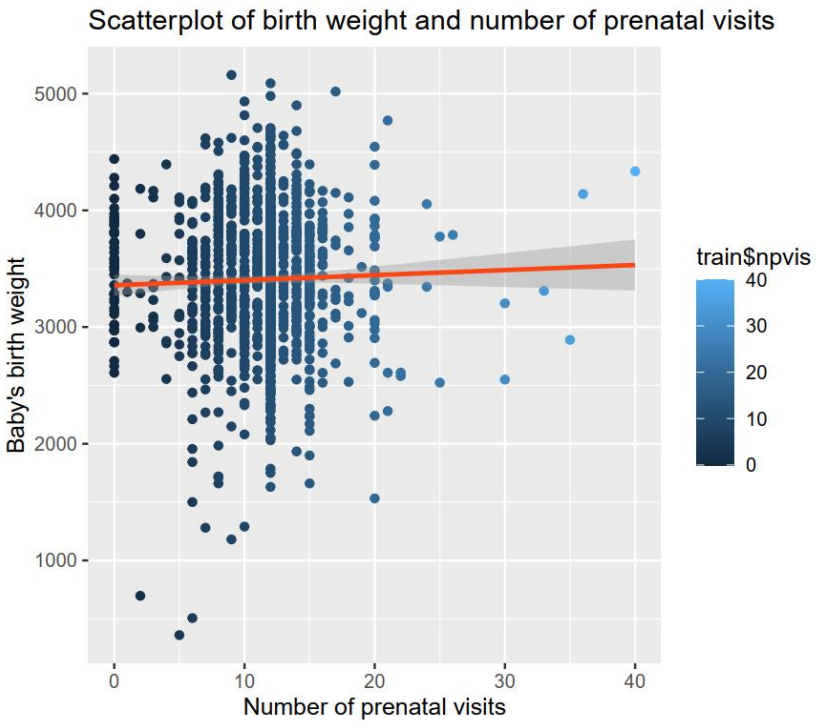
- Output:

```
> prediction3
          fit      lwr      upr
1    3405.714 3374.328 3437.100
2    3405.714 3374.328 3437.100
3    3399.708 3357.895 3441.521
...
```

5.2.4 bwght and npvis

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The range of confidence interval is large when data is sparse and the actual values are big

→ The range of confidence interval shrinks when the data is dense

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times npvis$$

+ Code:

```
model4<-lm(train$bwght~train$npvis)
summary(model4)
```

+ Output:

```
> summary(model4)

Call:
lm(formula = train$bwght ~ train$npvis)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-3019.66  -338.66    15.63   361.50  1762.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3357.971     45.882  73.188   <2e-16 ***
train$npvis    4.339      3.842   1.129    0.259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.4 on 1277 degrees of freedom
Multiple R-squared:  0.0009975, Adjusted R-squared:  0.0002152
F-statistic: 1.275 on 1 and 1277 DF,  p-value: 0.259

>
```

+ Model:

$bwght = 3357.971 + 4.339 \times npvis + \varepsilon$

($R^2 = 0.0009975$)

+ Meaning:

$\beta_1 = 3357.971$ means when the number of prenatal visits is 0, birth weight is 3357.971 g

$\beta_2 = 4.339$ means when the number of prenatal visits increases by 1, birth weight increases by 4.339 g

- Confident interval:

+ Code:

```
confint(model4)
```

+ Output

```
> confint(model4)
                  2.5 %     97.5 %
(Intercept) 3267.959071 3447.98258
train$npvis   -3.199013   11.87605
```

which means:

95% confident interval of

β_1 is (3332.88639, 3454.51786)

β_2 is (-18.35515, 30.36681)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction4<-predict(model4,data.frame(test$npvis), interval =
"confidence")
prediction4
```

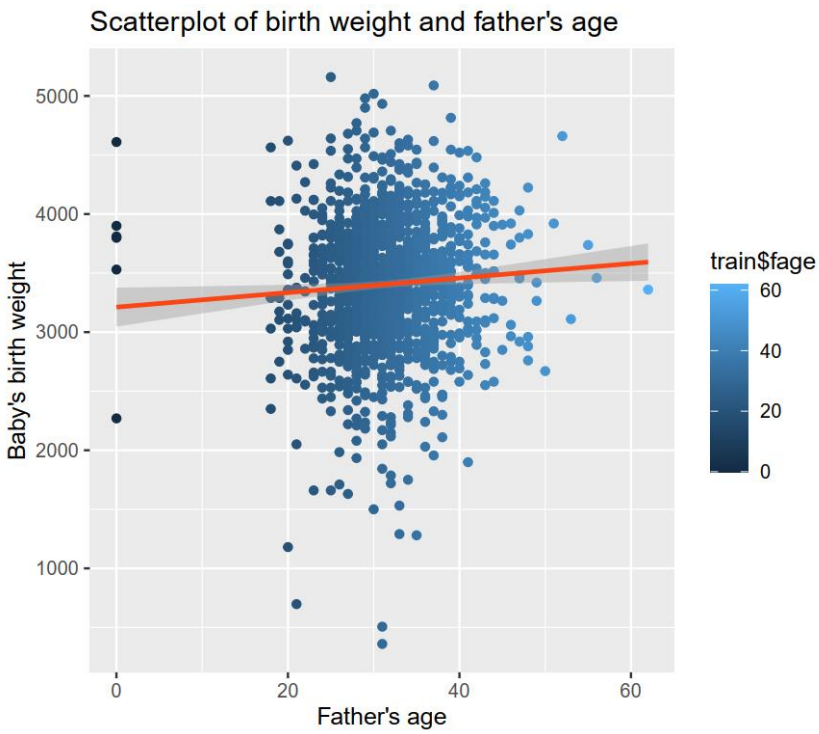
- Output:

```
> prediction4
          fit      lwr      upr
1    3410.033 3378.280 3441.786
2    3410.033 3378.280 3441.786
3    3410.033 3378.280 3441.786
...
```

5.2.5 bwght and fage

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The range of confidence interval is large when data is sparse and the actual values are too small / too big

→ The range of confidence interval shrinks when the data is dense

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times fage$$

+ Code:

```
model5<-lm(train$bwght~train$fage)
summary(model5)
```

+ Output:

```
> summary(model5)

Call:
lm(formula = train$bwght ~ train$fage)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-3042.48  -339.83    15.21   368.41  1794.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3211.718     83.842  38.307   <2e-16 ***
train$fage     6.154      2.600   2.367   0.0181 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 567.4 on 1277 degrees of freedom
Multiple R-squared:  0.004368,  Adjusted R-squared:  0.003588
F-statistic: 5.602 on 1 and 1277 DF,  p-value: 0.01809

>
```

+ Model:

$bwght = 3211.718 + 6.154 \times fage + \varepsilon$

($R^2 = 0.004368$)

+ Meaning:

$\beta_1 = 3211.718$ means when father age is 0, birth weight is 3211.718 g

$\beta_2 = 6.154$ means when the father gets 1 year older, birth weight increases by 6.154 g

- Confident interval:

+ Code:

```
confint(model5)
```

+ Output

```
> confint(model5)
                  2.5 %     97.5 %
(Intercept) 3047.234058 3376.20212
train$fage     1.052979   11.25416
>
```

---

which means:

95% confident interval of

β_1 is (3047.234058, 3376.20212)

β_2 is (1.052979, 11.25416)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction5<-predict(model5,data.frame(test$fage), interval =
"confidence")
prediction5
```

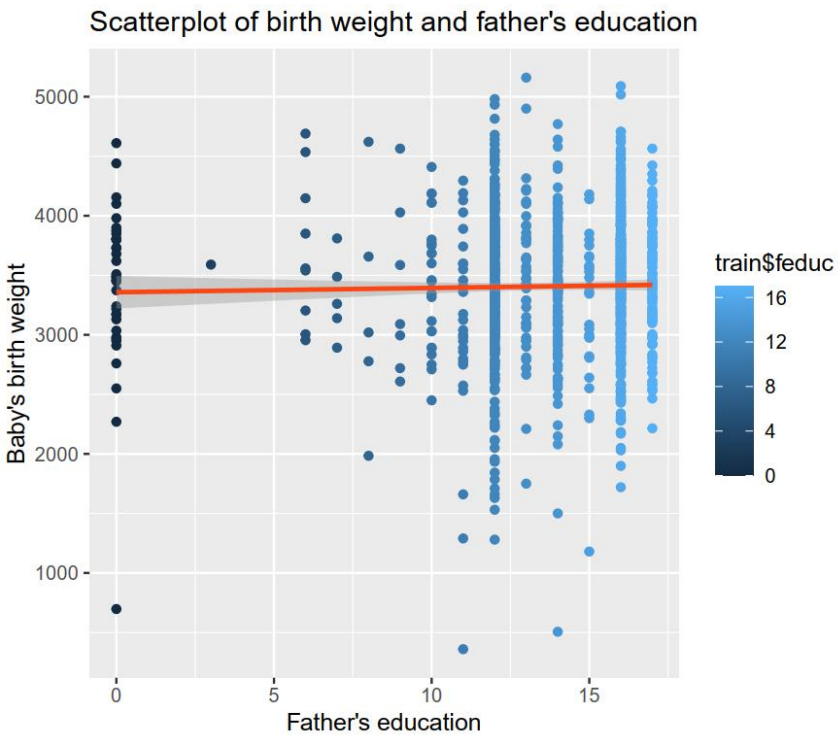
- Output:

```
> prediction5
          fit      lwr      upr
1    3420.939 3387.613 3454.266
2    3408.632 3377.459 3439.805
3    3433.247 3395.067 3471.426
...
```

5.2.6 bwght and feduc

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The smaller the value, the larger the range of confidence interval

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times feduc$$

+ Code:

```
model6<-lm(train$bwght~train$feduc)
summary(model6)
```

+ Output:

```
> summary(model6)

Call:
lm(formula = train$bwght ~ train$feduc)

Residuals:
     Min       1Q   Median       3Q      Max
-3037.37  -340.98    17.98   367.22  1755.42
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3357.675     69.584  48.253   <2e-16 ***
train$feduc    3.608      4.999   0.722    0.471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.5 on 1277 degrees of freedom
Multiple R-squared:  0.0004078, Adjusted R-squared:  -0.0003749
F-statistic: 0.521 on 1 and 1277 DF,  p-value: 0.4705

>
```

+ Model:

bwght = 3357.675 + 3.608 × *feduc* + *ε*

(R^2 = 0.0004078)

+ Meaning:

β_1 = 3357.675 means when father’s education is 0, birth weight is 3357.675 g

β_2 = 3.608 means when the father’s education gets increases by 1, birth weight increases by 3.608 g

- Confident interval:

+ Code:

```
confint(model6)
```

+ Output

```
> confint(model6)
                  2.5 %     97.5 %
(Intercept) 3221.163076 3494.18782
train$feduc   -6.199022   13.41574
>
```

which means:

95% confident interval of

β_1 is (3221.163076, 3494.18782)

β_2 is (-6.199022, 13.41574)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction6<-predict(model6,data.frame(test$feduc), interval =
"confidence")
prediction6
```

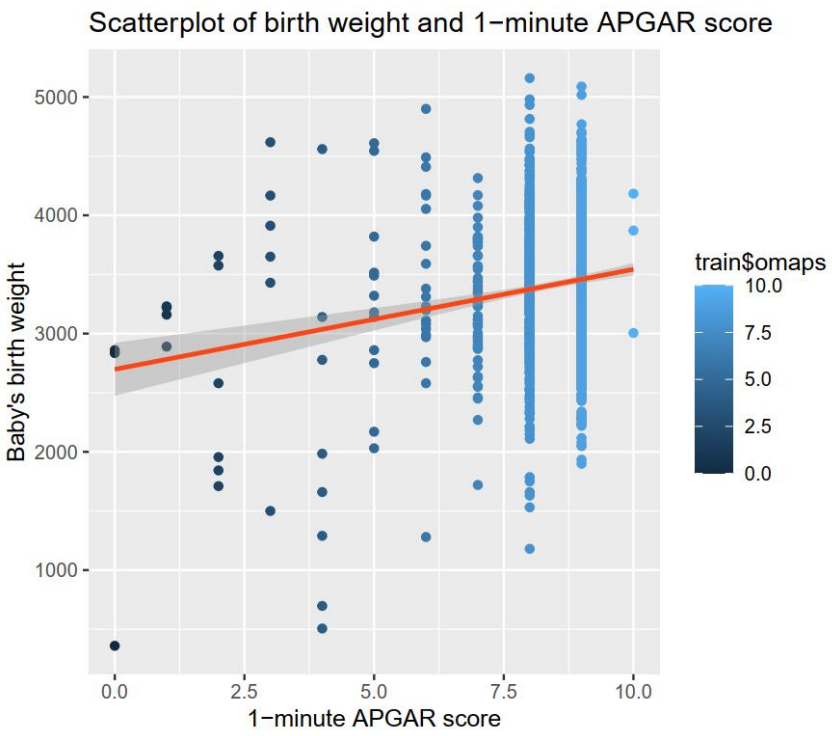
- Output:

```
> prediction6
         fit      lwr      upr
1    3415.409 3376.046 3454.773
2    3400.976 3366.275 3435.677
...
```

5.2.7 bwght and omaps

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The smaller the value, the larger the range of confidence interval

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times omaps$$

+ Code:

```
model7<-lm(train$bwght~train$omaps)
summary(model7)
```

+ Output:

```
> summary(model7)

Call:
lm(formula = train$bwght ~ train$omaps)

Residuals:
     Min       1Q   Median       3Q      Max
-2529.98  -358.88     1.12   362.41  1785.70
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2697.66     114.22  23.619  < 2e-16 ***
train$omaps    84.58      13.50   6.266 5.06e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560.1 on 1277 degrees of freedom
Multiple R-squared:  0.02983,   Adjusted R-squared:  0.02907
F-statistic: 39.26 on 1 and 1277 DF,  p-value: 5.059e-10

>
```

+ Model:

$bwght = 2697.66 + 84.58 \times omaps + \varepsilon$

($R^2 = 0.02983$)

+ Meaning:

$\beta_1 = 2697.66$ means when 1-minute APGAR score is 0, birth weight is 2697.66 g

$\beta_2 = 84.58$ means when the 1-minute APGAR score increases by 1, birth weight increases by 84.58 g

- Confident interval:

+ Code:

```
confint(model7)
```

+ Output

```
> confint(model7)
                  2.5 %    97.5 %
(Intercept) 2473.58631 2921.7363
train$omaps   58.09827  111.0618
>
```

which means:

95% confident interval of

β_1 is (2473.58631, 2921.7363)

β_2 is (58.09827, 111.0618)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction7<-predict(model7,data.frame(test$omaps), interval =
"confidence")
prediction7
```

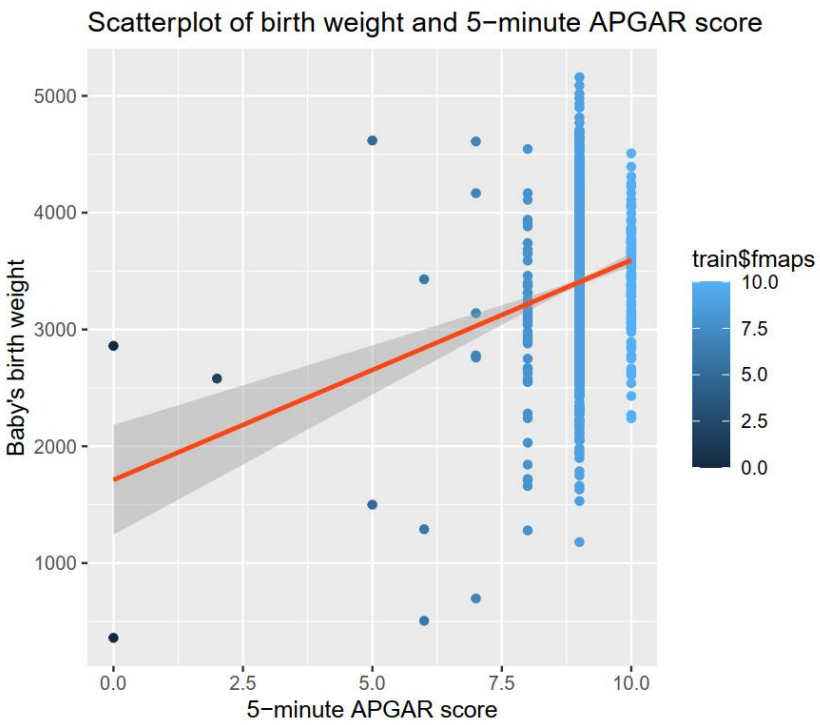
- Output:

```
> prediction7
         fit      lwr      upr
1    3458.882 3424.064 3493.700
2    3374.302 3341.957 3406.646
...
```

5.2.8 bwght and fmaps

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The smaller the value, the larger the range of confidence interval

→ The range of confidence interval shrinks significantly when the actual values of birth weight increase

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times fmaps$$

+ Code:

```
model8<-lm(train$bwght~train$fmaps)
summary(model8)
```

+ Output:

```
> summary(model8)

Call:
lm(formula = train$bwght ~ train$fmaps)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2335.72  -344.43    13.57   363.57  1964.51

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1712.3      239.0   7.164 1.32e-12 ***
train$fmaps    188.2       26.5   7.103 2.02e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 557.8 on 1277 degrees of freedom
Multiple R-squared:  0.03801,   Adjusted R-squared:  0.03726
F-statistic: 50.46 on 1 and 1277 DF,  p-value: 2.018e-12

>
```

+ Model:

bwght = 1712.3 + 188.2 × *fmaps* + *ε*

(R^2 = 0.03801)

+ Meaning:

β_1 = 1712.3 means when 5-minute APGAR score is 0, birth weight is 1712.3 g

β_2 = 188.2 means when the 5-minute APGAR score increases by 1, birth weight increases by 188.2 g

- Confident interval:

+ Code:

```
confint(model8)
```

+ Output

```
> confint(model8)
                 2.5 %    97.5 %
(Intercept) 1243.3947 2181.2487
train$fmaps  136.2463  240.2213
```

which means:

95% confident interval of

β_1 is (1243.3947, 2181.2487)

β_2 is (136.2463, 240.2213)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction8<-predict(model8,data.frame(test$fmaps), interval =
"confidence")
prediction8
```

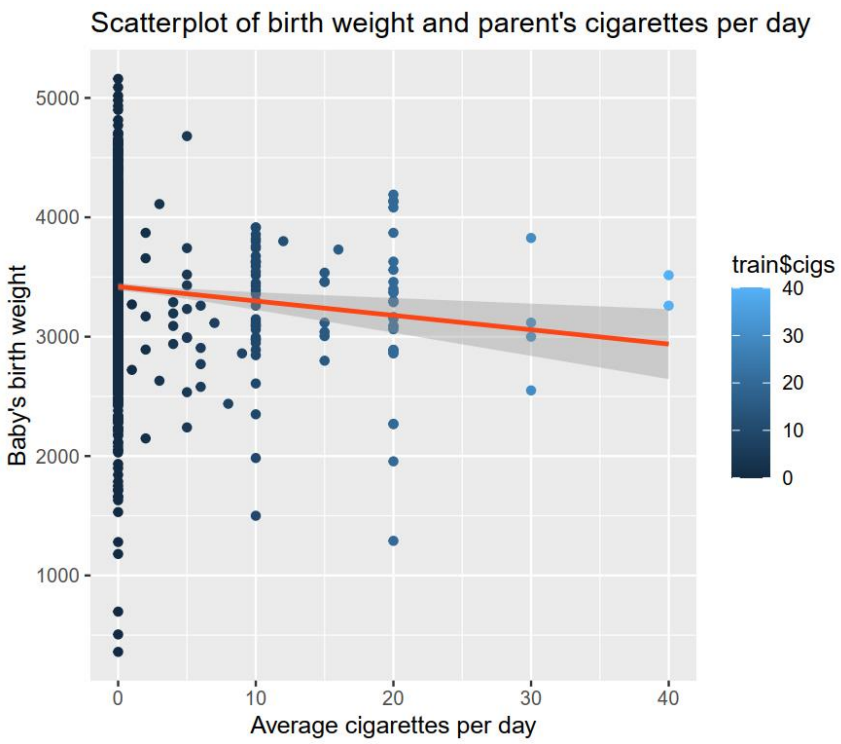
- Output:

```
> prediction8
          fit      lwr      upr
1    3406.426 3375.830 3437.022
2    3406.426 3375.830 3437.022
...
```

5.2.9 bwght and cigs

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The bigger the value, the larger the range of confidence interval

→ The range of confidence interval expands significantly when the actual values of birth weight increase

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times cigs$$

+ Code:

```
model9<-lm(train$bwght~train$cigs)
summary(model9)
```

+ Output:

```
> summary(model9)

Call:
lm(formula = train$bwght ~ train$cigs)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-3059.18  -339.18    10.82   369.82  1740.82

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3419.175     16.336 209.303  < 2e-16 ***
train$cigs   -12.028      3.816  -3.152  0.00166 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 566.5 on 1277 degrees of freedom
Multiple R-squared:  0.007721,  Adjusted R-squared:  0.006944
F-statistic: 9.937 on 1 and 1277 DF,  p-value: 0.001658

>
```

+ Model:

bwght = 3419.18 - 12.03 × *cigs* + *ε*

(R^2 = 0.007721)

+ Meaning:

β_1 = 3419.18 means when the number of cigarettes per day is 0, birth weight is 3419.18g

β_2 = - 12.03 means when the number of cigarettes per day increases by 1, birth weight decreases by 12.03 g

- Confident interval:

+ Code:

```
confint(model9)
```

+ Output

```
> confint(model9)
                 2.5 %      97.5 %
(Intercept) 3387.12680 3451.223651
train$cigs   -19.51436   -4.542535
```

which means:

95% confident interval of

β_1 is (3387.12680, 3451.223651)

β_2 is (-19.51436, -4.542535)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction9<-predict(model9,data.frame(test$cigs), interval =
"confidence")
prediction9
```

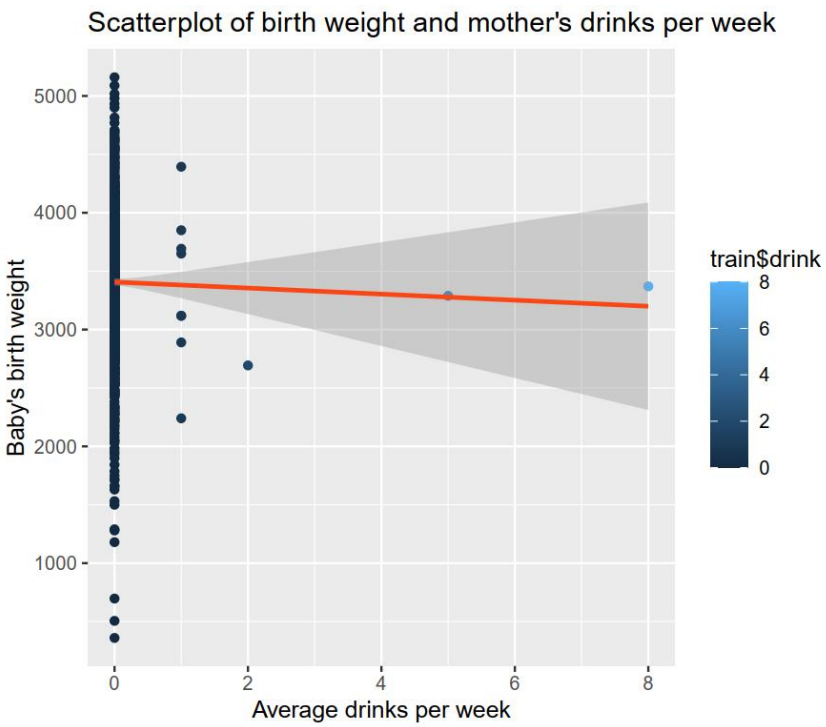
- Output:

```
> prediction9
          fit      lwr      upr
1    3419.175 3387.127 3451.224
2    3419.175 3387.127 3451.224
...
```

5.2.10 bwght and drink

- Visualize scatter plot + linear model

Code is similar as above. It gives us the scatter plot + linear model visualization:



→ The bigger the value, the larger the range of confidence interval

→ The range of confidence interval expands constantly when the actual values of birth weight increase. Perhaps the unbalance in data (most of the observed mothers do not drink, which means drink=0) cause the range of confidence interval to expand in such way.

- Find linear model:

$$bwght = \beta_1 + \beta_2 \times drink$$

+ Code:

```
model10<-lm(train$bwght~train$drink)
summary(model10)
```

+ Output:

```
> summary(model10)

Call:
```

```
lm(formula = train$bwght ~ train$drink)

Residuals:
     Min       1Q   Median       3Q      Max
-3047.04  -347.04    22.96   363.96  1752.96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3407.04      15.93 213.846   <2e-16 ***
train$drink   -25.93      56.70  -0.457    0.647
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.6 on 1277 degrees of freedom
Multiple R-squared:  0.0001638, Adjusted R-squared:  -0.0006192
F-statistic: 0.2092 on 1 and 1277 DF,  p-value: 0.6475

>
```

+ Model:

bwght = 3407.04 - 25.93 × *drink* + *ε*

(R^2 = 0.0001638)

+ Meaning:

β_1 = 3407.04 means when the number of drinks per week is 0, birth weight is 3407.04g

β_2 = - 25.93 means when the number of drinks per week increases by 1, birth weight decreases by 25.93 g

- Confident interval:

+ Code:

```
confint(model10)
```

+ Output

```
> confint(model10)
```

```
                  2.5 %     97.5 %
(Intercept) 3375.7833 3438.29551
train$drink -137.1566   85.29707
>
```

which means:

95% confident interval of

β_1 is (3375.7833, 3438.29551)

β_2 is (-137.1566, 85.29707)

- Predict test data using linear model built from train data

Predict test$bwght using predict function

- Code:

```
prediction10<-predict(model10,data.frame(test$drink), interval =
"confidence")
prediction10
```

- Output:

```
> prediction10
          fit      lwr      upr
1    3407.039 3375.783 3438.296
2    3407.039 3375.783 3438.296
...
```

5.2.11 Summary on simple linear regression model

- Most of the quantitative attributes have positive correlation with birth weight (bwght), except cigs and drink

- Simple linear regression model of fmaps and bwght gives the highest R^2 value (0.03801), so the regression line fits the actual birth weight more than other regression lines. However, there are so many big values of birth weight associate with high fmaps, while there are just few small values on the left-hand side (actual data presented via points on scatter plot). Therefore, it may be a bias for regression line to predict better.

5.3 Multiple linear regression model

5.3.1 Find multiple linear regression model

- We have list of R^2 from the previous part

$R_1^2 = 0.00182$, $R_2^2 = 0.000555$, $R_3^2 = 0.0001831$, $R_4^2 = 0.0009975$, $R_5^2 = 0.004368$,

$R_6^2 = 0.0004078$, $R_7^2 = 0.02983$, $R_8^2 = 0.03801$, $R_9^2 = 0.007721$, $R_{10}^2 = 0.0001638$

→ fmaps has the strongest correlation with bwght ($R_8^2 = 0.03801$)

$$bwght = \beta_1 + \beta_2 \times fmaps + \beta_3 \times omaps + \beta_4 \times cigs + \beta_5 \times fage + \beta_6 \times mage + \beta_7 \times npvis + \beta_8 \times meduc + \beta_9 \times feduc + \beta_{10} \times monpre + \beta_{11} \times drink$$

- Code:

```
modelMul<-
lm(bwght~fmaps+omaps+cigs+fage+mage+npvis+meduc+feduc+monpre+drink)
summary(modelMul)
```

- Output:

```
> summary(modelMul)

Call:
lm(formula = bwght ~ fmaps + omaps + cigs + fage + mage + npvis +
    meduc + feduc + monpre + drink)

Residuals:
     Min       1Q   Median       3Q      Max
-2152.10  -334.38     4.28   345.15  2160.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1228.461    222.010   5.533 3.60e-08 ***
fmaps        165.158     27.039   6.108 1.23e-09 ***
omaps         51.958     13.934   3.729 0.000198 ***
cigs          -9.823      3.252  -3.020 0.002561 **
fage           5.882      2.918   2.016 0.043969 *
mage          -1.973      3.728  -0.529 0.596741
```

```
npvis          9.302      3.243   2.869 0.004171 **
meduc         -1.976      6.265  -0.315 0.752546
feduc          1.497      5.380   0.278 0.780867
monpre        17.498     11.070   1.581 0.114119
drink        -24.781     47.178  -0.525 0.599459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 554.7 on 1821 degrees of freedom
Multiple R-squared:  0.07924,   Adjusted R-squared:  0.07418
F-statistic: 15.67 on 10 and 1821 DF,  p-value: < 2.2e-16

>
```

→ Model is:

bwght = 1228.461 + 165.158 × fmaps + 51.958 × omaps - 9.823 × cigs + 5.882 × fage

– 1.973 × mage + 9.302 × npvis – 1.976 × meduc + 1.497 × feduc + 17.498 × monpre

– 24.781 × drink + ε

=> R^2 = 0.07924 is **higher than all R^2 of simple linear regression models** in the previous part)

5.3.2 Meaning of coefficients

+ β_1 = 1228.461 means when 5-minute APGAR, 1-minute APGAR, average cigarettes per day, father’s age, mother’s age, number of prenatal visits, mother’s education, father’s education, month prenatal care began and average drinks per week are all equal to 0, birth weight is 1228.461 g

+ β_2 = 165.158 means when fmaps increases by 1 grade, and all other variables are fixed, birth weight increases by 165.158 g

+ β_3 = 51.958 means when omaps increases by 1 grade, and all other variables are fixed, birth weight increases by 51.958 g

+ β_4 = - 9.823 means when parental average cigarettes per day increases by 1, and all other variables are fixed, birth weight decreases by 9.823 g

+ β_5 = 5.882 means when father's age increases by 1 year-old, and all other variables are fixed, birth weight increases by 5.882 g

+ β_6 = – 1.973 means when mother's age increases by 1 year-old, and all other variables are fixed, birth weight decreases by 1.973 g

+ β_7 = 9.302 means when total number of prenatal visits increases by 1 visit, and all other variables are fixed, birth weight increases by 9.302 g

+ β_8 = – 1.976 means when mother's education increases by 1 year, and all other variables are fixed, birth weight decreases by 1.976 g

+ β_9 = 1.497 means when father's education increases by 1 year, and all other variables are fixed, birth weight increases by 1.497 g

+ β_{10} = 17.498 means when month prenatal care began increases by 1 month, and all other variables are fixed, birth weight increases by 17.498 g

+ β_{11} = – 24.781 means when mother's average drinks per week increases by 1, and all other variables are fixed, birth weight decreases by 24.781 g

5.3.3 Confident interval of each coefficients

+ Code: `confint(modelMul)`

+ Output:

```
> confint(modelMul)
                   2.5 %      97.5 %
(Intercept)  793.0409088 1663.881467
fmaps        112.1262388  218.188790
omaps         24.6306349   79.286086
cigs         -16.2016945   -3.444115
fage           0.1590695   11.604136
mage          -9.2841307    5.338641
npvis          2.9420381   15.661502
```

```
meduc         -14.2631442   10.311970
feduc          -9.0552615   12.049201
monpre         -4.2127442   39.209165
drink        -117.3087282   67.746787
>
```

which means:

95% confident interval of

β_1 is (793.0409088, 1663.881467)

β_2 is (112.1262388, 218.188790)

β_3 is (24.6306349, 79.286086)

β_4 is (-16.2016945, -3.444115)

β_5 is (0.1590695, 11.604136)

β_6 is (-9.2841307, 5.338641)

β_7 is (2.9420381, 15.661502)

β_8 is (-14.2631442, 10.311970)

β_9 is (-9.0552615, 12.049201)

β_{10} is (-4.2127442, 39.209165)

β_{11} is (-117.3087282, 67.746787)

6. Goodness of fit test

6.1 fblck – lbw (black father – low birth weight)

- We want to know whether the proportion of low birth weight and normal birth weight are the same when their father are blacks or non-blacks. Research has shown that parents who are African-American tend to to have low birth weight babies. We will perform goodness of fit test for fblck and lbw.

- Create table of fblck and lbw

+ Code:

```
table_fblck_lbw<-table(fblck, lbw)
table_fblck_lbw
```

+ Output

```
> table_fblck_lbw
      lbw
fblck    0    1
```

```
  0 1697   28
  1  105    2
```

- Goodness of fit test by using chi-square test on table_fblck_lbw

Our hypothesis is:

H_0: $p_{1j} = p_{2j} = p_{ij}$ (j=1, 2) (distributions are the same for babies whose father is black and non-black)

H_1: H_0 is not true

+ Code:

```
chisq.test(table_fblck_lbw)
```

+ Output:

```
> chisq.test(table_fblck_lbw)
    Pearson's Chi-squared test with Yates'
    continuity correction
data:  table_fblck_lbw
X-squared = 4.0174e-29, df = 1, p-value = 1

Warning message:
In chisq.test(table_fblck_lbw) : Chi-squared approximation may be
incorrect
>
```

→ Analysis:

- Up to this point, the p-value is 1 > α = 0.05. Therefore, we accept the null hypothesis, which means the proportion of low birth weight and normal birth weight are the same for babies whose father is black or non-black.

- However, the code above gives us **warning!** To make sure that we get the correct answer, let's try to use simulate to find p value

+ Code:

```
chisq.test(table_fblck_lbw, simulate.p.value = TRUE)
```

+ Output:

```
> chisq.test(table_fblck_lbw, simulate.p.value = TRUE)
    Pearson's Chi-squared test with simulated
    p-value (based on 2000 replicates)
data:  table_fblck_lbw
X-squared = 0.037843, df = NA, p-value = 1
>
```

→ This still gives us the same p-value

6.2 male – vlbw (male babies – very low birth weight)

- We want to know whether the proportion of very low birth weight and not very low birth weight are the same for male and female babies.

- Create table of male and vlbw

+ Code:

```
table_male_vlbw<-table(male, vlbw)
table_male_vlbw
```

+ Output

```
> table_male_vlbw
        vlbw
male       0   1
  Female 885   6
  Male   934   7
>
```

- Goodness of fit test by using chi-square test on table_male_vlbw

Our hypothesis is:

H_0: $p_{1j} = p_{2j} = p_{ij}$ (j=1, 2) (distributions are the same for male and female babies)

H_1: H_0 is not true

+ Code:

```
chisq.test(table_male_vlbw)
```

+ Output:

```
> chisq.test(table_male_vlbw)
    Pearson's Chi-squared test with Yates'
    continuity correction
data:  table_male_vlbw
X-squared = 2.9182e-30, df = 1, p-value = 1
>
```

→ Since p-value is 1 > α = 0.05. Therefore, we accept the null hypothesis, which means the proportion of very low birth weight and not very low birth weight are the same for male and female babies.

7. Summary on data

- Data need to be pre-processed in order to work on since there are inconsistency in data types and many missing values (“.” instead of #NA)

- The data is unbalance in many attributes:

+ Some typical qualitative variables such as mblck, fblck, lbw, vlbw has very small proportion of value = 1, while variables such as mwhte and fwhte has very high proportion of values = 1

+ Quantitative variables:

- mage, fage, bwght, lbwght are normally distributed

- Other variables such as omaps, fmaps, cigs, drinks are unbalance: omaps and fmaps have big modes while cigs and drinks have small modes (mode = 0)
- All the remainings are not normally distributed and has many peaks.

+ Although some pairs of qualitatives give us the same proportion on each, the frequencies of them deviate much from each other.

- There are two variables don't show much meaning in relation with baby's birth weight: meduc (mother's education) and feduc (father's education). Since this data is related to baby's health care, I researched a lot on the conditions that affect baby's birth weight, but I haven't read any papers regarding parental education and baby's birth weight. Also, the regression models of these 2 and bwght have very small R^2

- Attributes such as monpre (month prenatal care began), mage (mother's age) and npvis (number of prenatal visits) are known to have strong correlation with birth weight according to research. However, fmaps and omaps has stronger correlation with birth weight than those attributes in this data.

- Cigs and drink have negative correlation with bwght as expected.

- The multiple linear regression model gives better R^2 since it identify birth weight based on all quantitative variables.