R Practical Exam

Marwa Tawfik Badawy

May 14th, 2019

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyvers
e 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0       ✔ purrr   0.3.2
## ✔ tibble  2.0.1       ✔ dplyr   0.8.0.1
## ✔ tidyr   0.8.3       ✔ stringr 1.3.1
## ✔ readr   1.3.1       ✔ forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conf
licts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Loading Babies dataset

```
babies_original = read_csv("http://bit.ly/babies-weight-smoking-csv")
```

```
## Parsed with column specification:
## cols(
##   weight = col_double(),
##   gestation = col_double(),
##   parity = col_double(),
##   mom.race = col_character(),
##   mom.age = col_double(),
##   mom.edu = col_double(),
##   mom.height = col_double(),
##   mom.weight = col_double(),
##   dad.race = col_character(),
##   dad.age = col_double(),
##   dad.edu = col_double(),
##   dad.height = col_double(),
##   dad.weight = col_double(),
##   marital = col_double(),
##   income = col_double(),
##   smoke = col_character(),
##   quit.time = col_double(),
##   cigs = col_double()
## )
```

Summary of the data

```
glimpse(babies_original)
```

```
## Observations: 610
## Variables: 18
## $ weight     <dbl> 120, 113, 136, 132, 120, 144, 115, 115, 119, 115, 137…
## $ gestation  <dbl> 284, 282, 286, 245, 289, 282, 285, 261, 288, 274, 287…
## $ parity     <dbl> 1, 2, 4, 2, 3, 4, 4, 3, 3, 1, 1, 6, 1, 3, 2, 3, 2, 1,…
## $ mom.race   <chr> "asian", "white", "white", "black", "white", "white",…
## $ mom.age    <dbl> 27, 33, 25, 23, 25, 32, 38, 33, 43, 27, 25, 30, 26, 3…
## $ mom.edu    <dbl> 5, 5, 2, 1, 4, 2, 2, 2, 2, 4, 4, 1, 0, 2, 5, 1, 2, 2,…
## $ mom.height <dbl> 62, 64, 62, 65, 62, 64, 63, 60, 66, 67, 66, 68, 58, 6…
## $ mom.weight <dbl> 100, 135, 93, 140, 125, 124, 130, 125, 142, 175, 145,…
## $ dad.race   <chr> "asian", "white", "white", "black", "white", "white",…
## $ dad.age    <dbl> 31, 38, 28, 23, 26, 36, 37, 33, 45, 26, 25, 38, 29, 2…
## $ dad.edu    <dbl> 5, 5, 2, 4, 1, 1, 0, 2, 2, 4, 5, 1, 2, 5, 5, 2, 2, 2,…
## $ dad.height <dbl> 65, 70, 64, 71, 70, 74, 71, 70, 73, 73, 70, 73, 68, 6…
## $ dad.weight <dbl> 110, 148, 130, 192, 180, 185, 205, 140, 195, 180, 150…
## $ marital    <dbl> 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ income     <dbl> 1, 4, 4, 2, 2, 2, 1, 4, 5, 3, 2, 2, 2, 1, 6, 4, 7, 2,…
## $ smoke      <chr> "never", "never", "until_pregnancy", "never", "never"…
## $ quit.time  <dbl> 0, 0, 2, 0, 0, 1, 0, 1, 1, 1, 2, 0, 2, 0, 1, 2, 4, 0,…
## $ cigs       <dbl> 0, 0, 2, 0, 0, 1, 0, 5, 6, 9, 5, 0, 1, 0, 5, 1, 1, 0,…
```

Adjust the data (convert categorical variables into factors)

```
babies = babies_original %>%
  mutate(parity = factor(parity)) %>%
  mutate(mom.race = factor(mom.race)) %>%
  mutate(mom.edu = factor(mom.edu)) %>%
  mutate(dad.race = factor(dad.race)) %>%
  mutate(dad.edu = factor(dad.edu)) %>%
  mutate(marital = factor(marital)) %>%
  mutate(mom.edu = factor(mom.edu)) %>%
  mutate(income = factor(income)) %>%
  mutate(cigs = factor(cigs)) %>%
  mutate(quit.time  = factor(quit.time )) %>%
  mutate(smoke = factor(smoke, levels = c("never",
                                          "once_not_now",
                                          "until_pregnancy",
                                          "now")))

# Now the dataset is called "babies"
```

Summary of the adjusted (converted) data

```
glimpse(babies)
```

```
## Observations: 610
## Variables: 18
## $ weight     <dbl> 120, 113, 136, 132, 120, 144, 115, 115, 119, 115, 137…
## $ gestation  <dbl> 284, 282, 286, 245, 289, 282, 285, 261, 288, 274, 287…
## $ parity     <fct> 1, 2, 4, 2, 3, 4, 4, 3, 3, 1, 1, 6, 1, 3, 2, 3, 2, 1,…
## $ mom.race   <fct> asian, white, white, black, white, white, black, whit…
## $ mom.age    <dbl> 27, 33, 25, 23, 25, 32, 38, 33, 43, 27, 25, 30, 26, 3…
## $ mom.edu    <fct> 5, 5, 2, 1, 4, 2, 2, 2, 2, 4, 4, 1, 0, 2, 5, 1, 2, 2,…
## $ mom.height <dbl> 62, 64, 62, 65, 62, 64, 63, 60, 66, 67, 66, 68, 58, 6…
## $ mom.weight <dbl> 100, 135, 93, 140, 125, 124, 130, 125, 142, 175, 145,…
## $ dad.race   <fct> asian, white, white, black, white, white, black, whit…
## $ dad.age    <dbl> 31, 38, 28, 23, 26, 36, 37, 33, 45, 26, 25, 38, 29, 2…
## $ dad.edu    <fct> 5, 5, 2, 4, 1, 1, 0, 2, 2, 4, 5, 1, 2, 5, 5, 2, 2, 2,…
## $ dad.height <dbl> 65, 70, 64, 71, 70, 74, 71, 70, 73, 73, 70, 73, 68, 6…
## $ dad.weight <dbl> 110, 148, 130, 192, 180, 185, 205, 140, 195, 180, 150…
## $ marital    <fct> 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
## $ income     <fct> 1, 4, 4, 2, 2, 2, 1, 4, 5, 3, 2, 2, 2, 1, 6, 4, 7, 2,…
## $ smoke      <fct> never, never, until_pregnancy, never, never, now, nev…
## $ quit.time  <fct> 0, 0, 2, 0, 0, 1, 0, 1, 1, 1, 2, 0, 2, 0, 1, 2, 4, 0,…
## $ cigs       <fct> 0, 0, 2, 0, 0, 1, 0, 5, 6, 9, 5, 0, 1, 0, 5, 1, 1, 0,…
```

Q0 - The mean weight

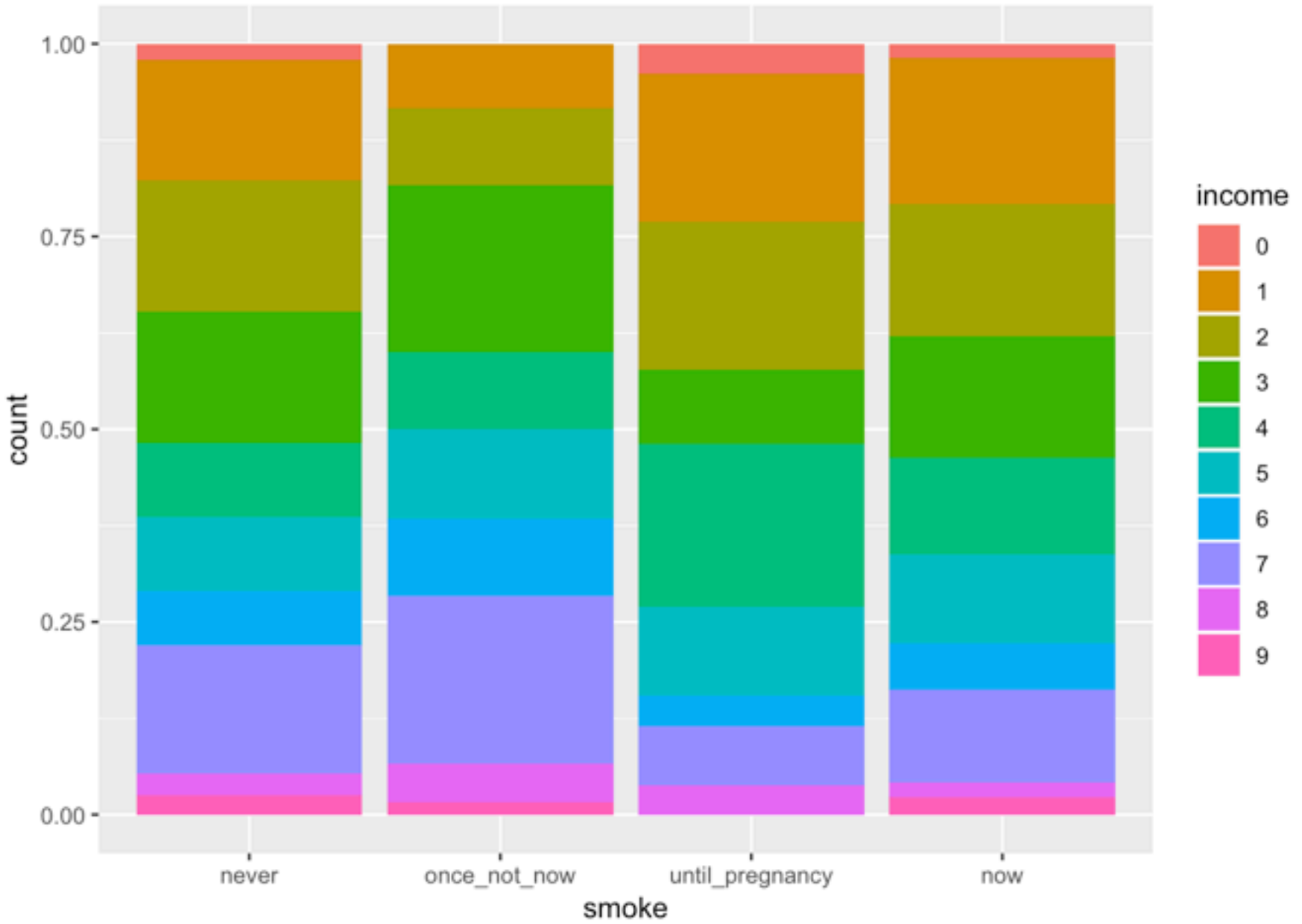
```
mean(babies$weight)
```

```
## [1] 119.2902
```

The mean of babies in the study is 119

Q1. [10 points] Write the R to produce the following figure. What insights might be concluded from the figure?

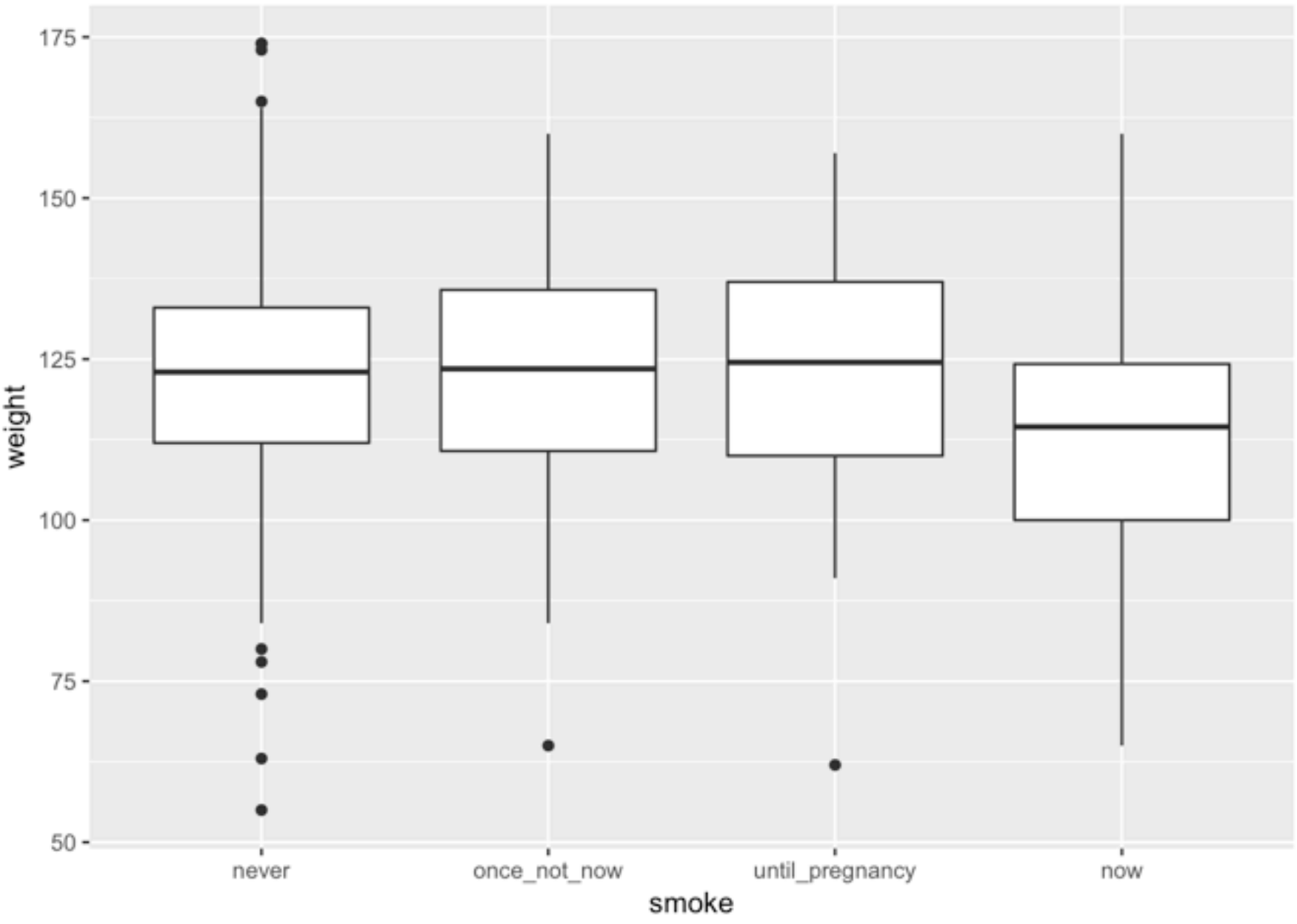
```
ggplot(babies) +
  geom_bar(aes(x = smoke, fill = income), position = "fill")
```



We can conclude that, the income is affecting the percetage of the somker moms.
The moms with low income are with high precetages in never smoke.
For once not now, we can see the middle income families with high representations.
Fro Until pregnancy, the low income families accounts for more percantages.
For now,the moms with high income accounts for alot of the populations.

Q2. [20 points] There are four smoking maternal categories, “never”, “once_not_now”, “until_pregnancy”, and “now”. Summeraize the differences between the weights of the newborn babies from the four groups, numerically and visually. Discuss the results.

```
ggplot(babies) + geom_boxplot(aes(x=smoke, y=weight))
```




```
babies %>% filter(smoke=="never") %>% summarise(mean(weight))
```

```
## # A tibble: 1 x 1
##   `mean(weight)`
##            <dbl>
## 1           122.
```

```
babies %>% filter(smoke=="once_not_now") %>% summarise(mean(weight))
```

```
## # A tibble: 1 x 1
##   `mean(weight)`
##            <dbl>
## 1           123.
```

```
babies %>% filter(smoke=="until_pregnancy") %>% summarise(mean(weight))
```

```
## # A tibble: 1 x 1
##   `mean(weight)`
##            <dbl>
## 1           123.
```

```
babies %>% filter(smoke=="now") %>% summarise(mean(weight))
```

```
## # A tibble: 1 x 1
##   `mean(weight)`
##            <dbl>
## 1           113.
```

Here we can see that the weights of babies for the moms who were smoking within the pregrnancy(now) were the most lowest weight among all of the other caterigories. The three other cateogries we can see they are among, approximately, the same weights of newborn babies.
For the never somked and once not now, we can see that they have equal means.
For until pregnancy, the mean is slightly above.
For now, the mean is seems to be diiferent, so it might has a significant difference.

Q3. [20 points] Are the differences between the groups important (statistically significant)? If so, which are groups are statistically significant?

```
anova=aov(data=babies, formula = weight ~ smoke)
broom::tidy(anova)
```

```
## # A tibble: 2 x 6
##   term         df   sumsq meansq statistic       p.value
##   <chr>     <dbl>   <dbl>  <dbl>     <dbl>         <dbl>
## 1 smoke         3  11843.  3948.      12.4  0.0000000679
## 2 Residuals   606 192552.   318.      NA   NA
```

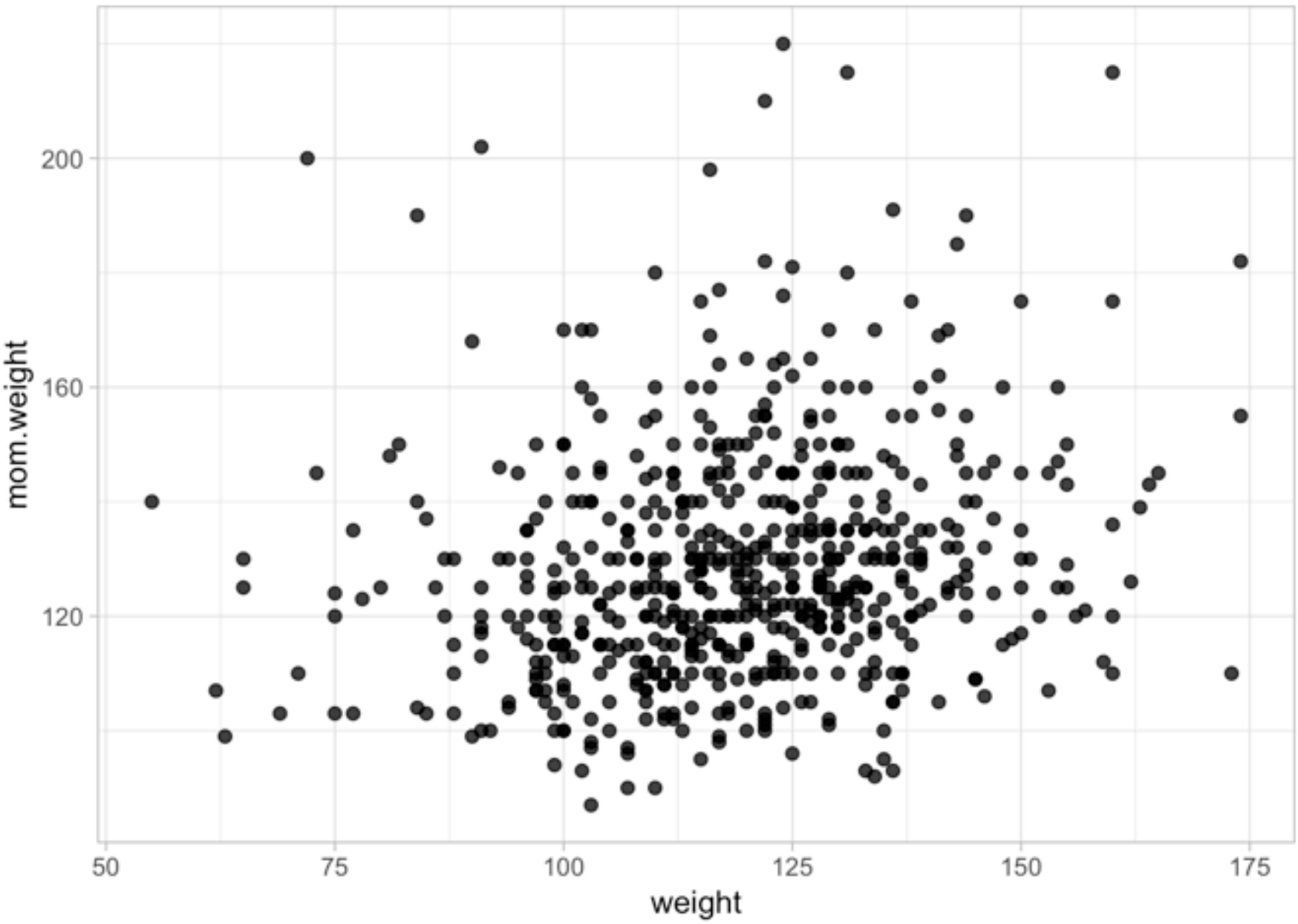
```
summary(anova)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## smoke          3  11843    3948   12.42 6.79e-08 ***
## Residuals    606 192552     318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

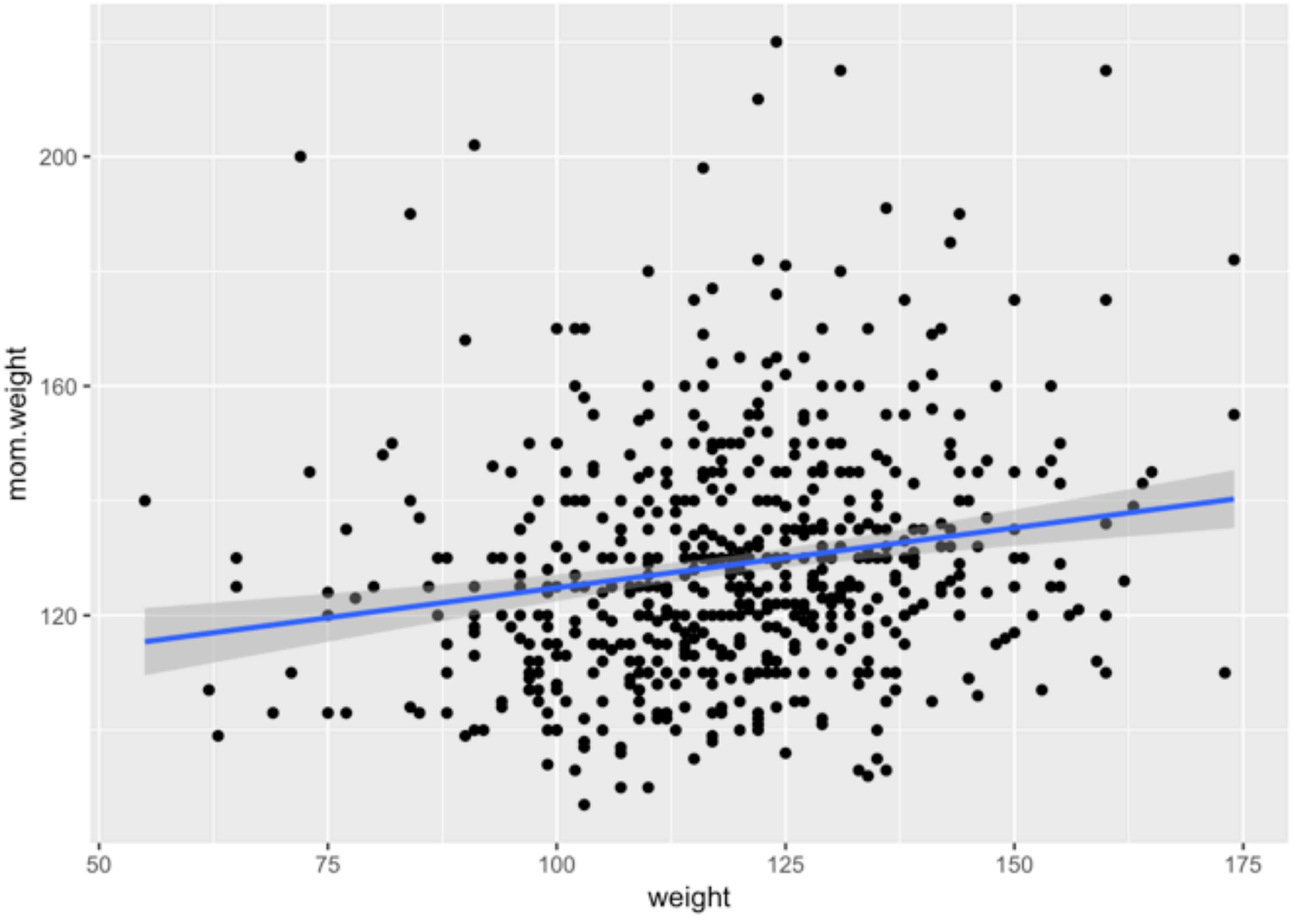
Yes, they are statstically significane

Q4. [10 points] Is the newborn weight associated with the mom's weight? Use analytical and visual methods. Discuss the results.

```
p = ggplot(babies)
p = p + geom_point(aes(x = weight , y = mom.weight ), alpha = 0.8, size = 2)
p = p + labs (x = "weight", y = "mom.weight")
p = p + theme_light(base_size = 12)
print(p)
```




```
ggplot(data = babies) + geom_point(mapping = aes(x = weight, y = mom.weight)) + ge
om_smooth(aes(x = weight, y = mom.weight), method = "lm")
```



The slope here indicaties a positive correlation between the mom's weight to the newborn wieights.

```
model = lm(data = babies, formula = weight ~ mom.weight)
broom::tidy(model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   97.7      4.65       21.0  5.20e-74
## 2 mom.weight     0.168    0.0357      4.70 3.26e- 6
```

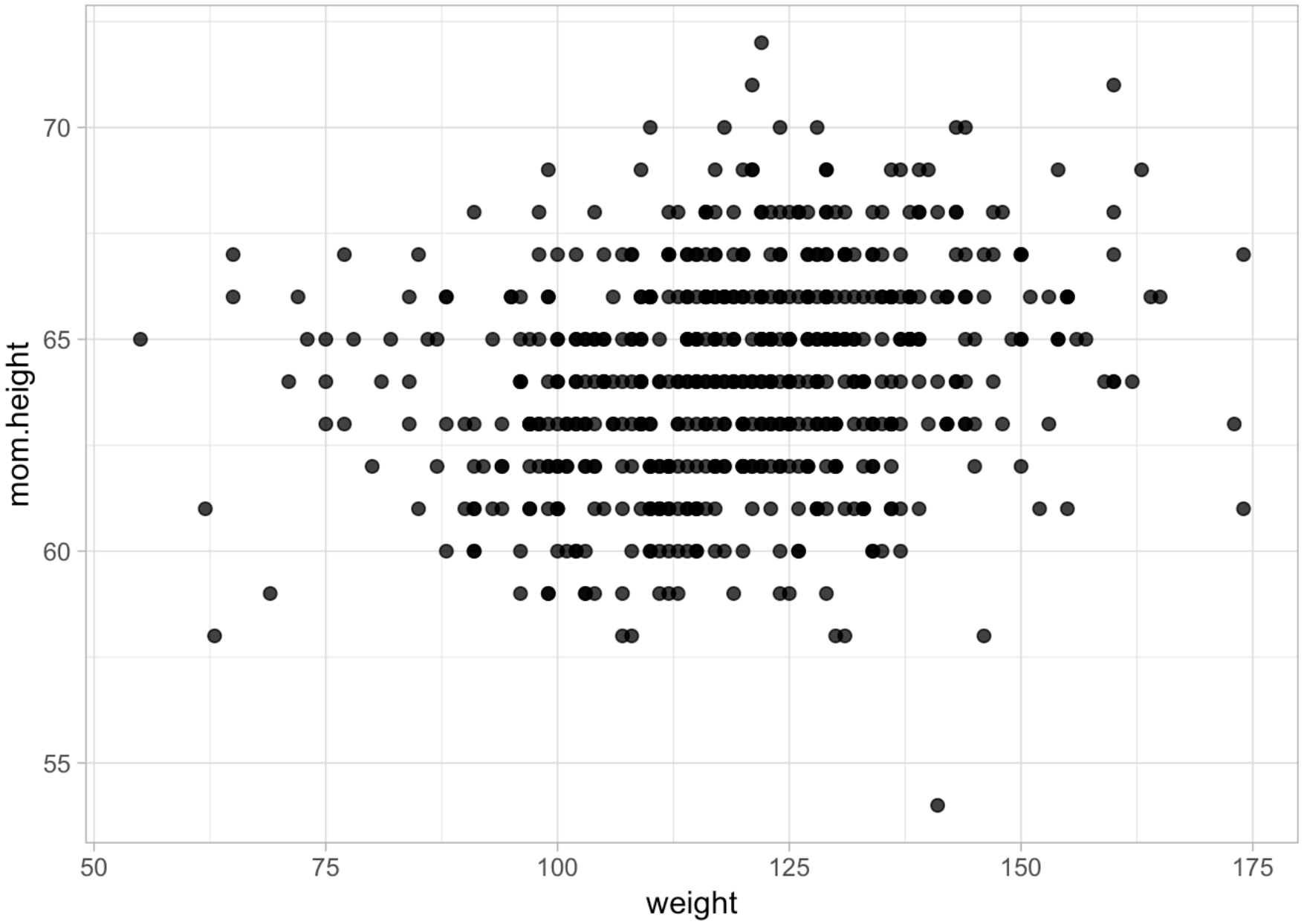
B1 = 0.61 This means that the newborn weight is estimated to differ by 0.16 grams for each one Kg differ in mom's weight.

```
summary(model)
```

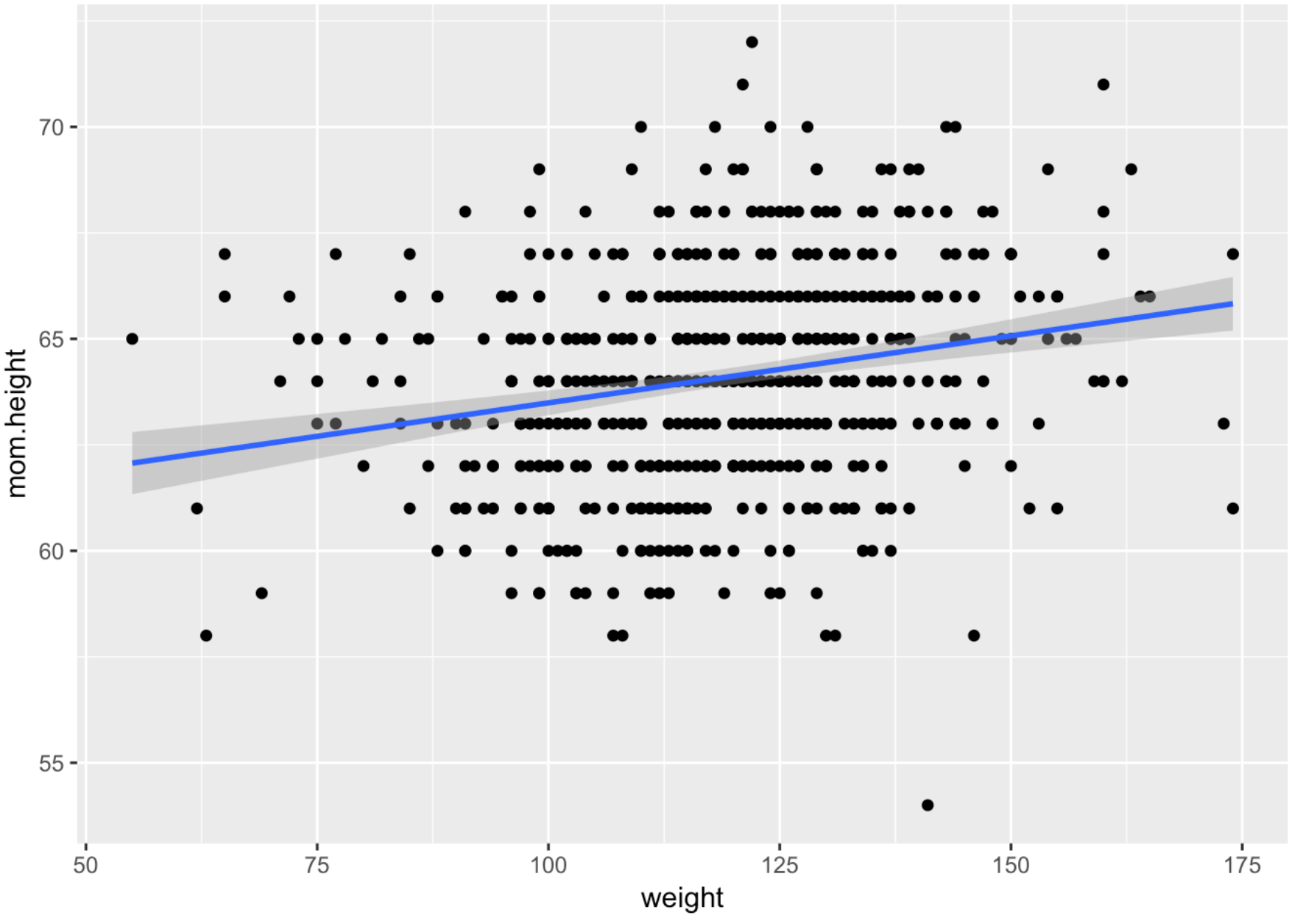
```
##
## Call:
## lm(formula = weight ~ mom.weight, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.158 -10.624   0.093  11.138  56.868
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.70039    4.65358  20.995  < 2e-16 ***
## mom.weight   0.16756    0.03567   4.697 3.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.01 on 608 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03343
## F-statistic: 22.07 on 1 and 608 DF,  p-value: 3.26e-06
```

Q5. [10 points] Is the newborn weight associated with the mom's height? Use analytical and visual methods. Discuss the results.

```
p = ggplot(babies)
p = p + geom_point(aes(x = weight , y = mom.height ), alpha = 0.8, size = 2)
p = p + labs (x = "weight", y = "mom.height")
p = p + theme_light(base_size = 12)
print(p)
```




```
ggplot(data = babies) + geom_point(mapping = aes(x = weight, y = mom.height)) + ge
om_smooth(aes(x = weight, y = mom.height), method = "lm")
```



We can interpret that there's a postive slope here as well which indicates that the mom's height is associated with the newborn weight.

```
model = lm(data = babies, formula = weight ~ mom.height)
broom::tidy(model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic      p.value
##   <chr>          <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)    17.4      18.0       0.968 0.333
## 2 mom.height      1.59      0.280     5.67  0.0000000222
```

ß1 = 1.5 This means that the mean weight of newborn is estimated to differ by 1.5 grams for each one cm difference in mom's height.

```
summary(model)
```

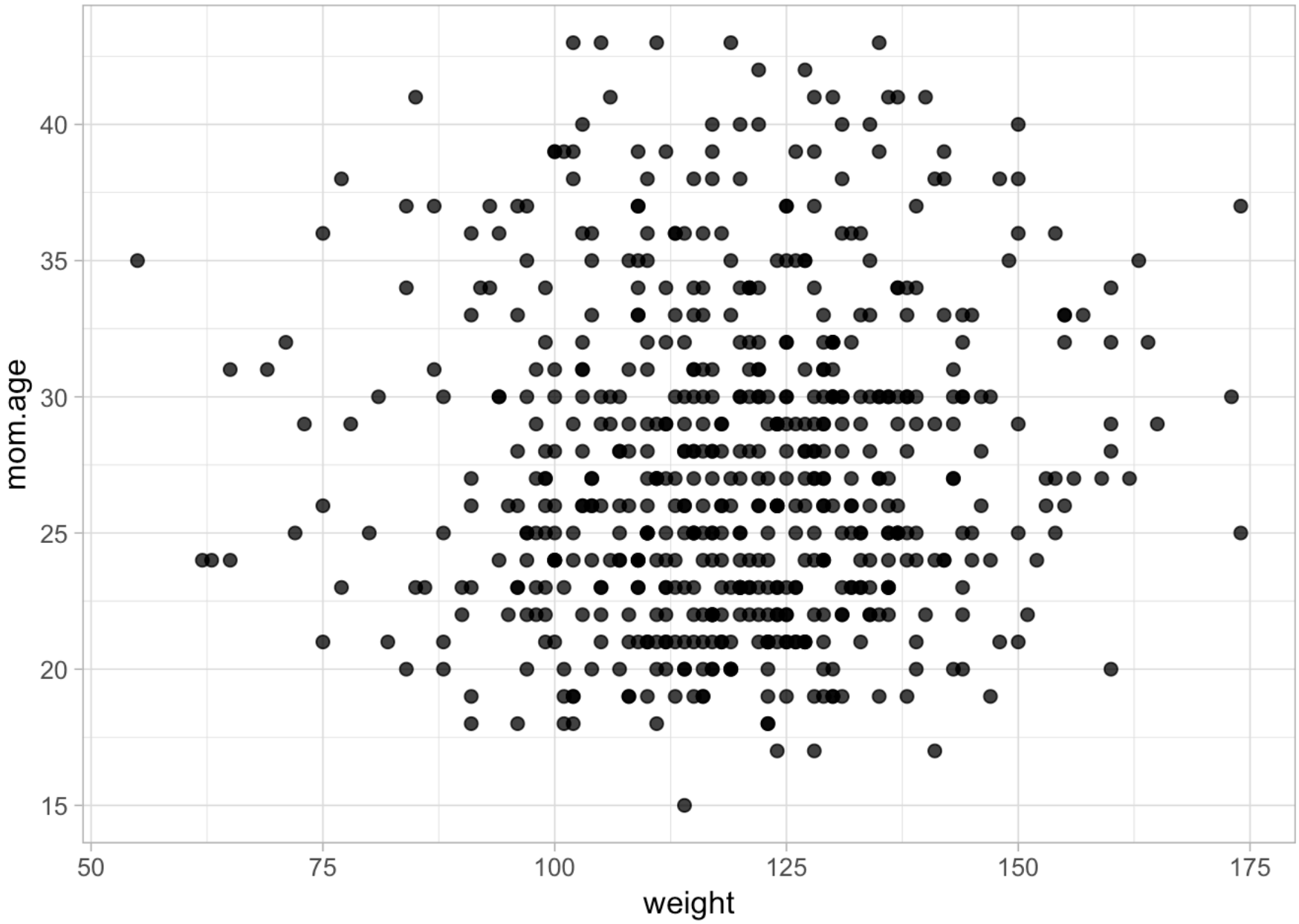
```
##
## Call:
## lm(formula = weight ~ mom.height, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.721 -10.774   0.369  11.413  59.637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.4139    17.9841   0.968    0.333
## mom.height    1.5893     0.2803   5.669 2.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.87 on 608 degrees of freedom
## Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04865
## F-statistic: 32.14 on 1 and 608 DF,  p-value: 2.216e-08
```

Q6. [10 points] Compare the results from Q3 and Q4.

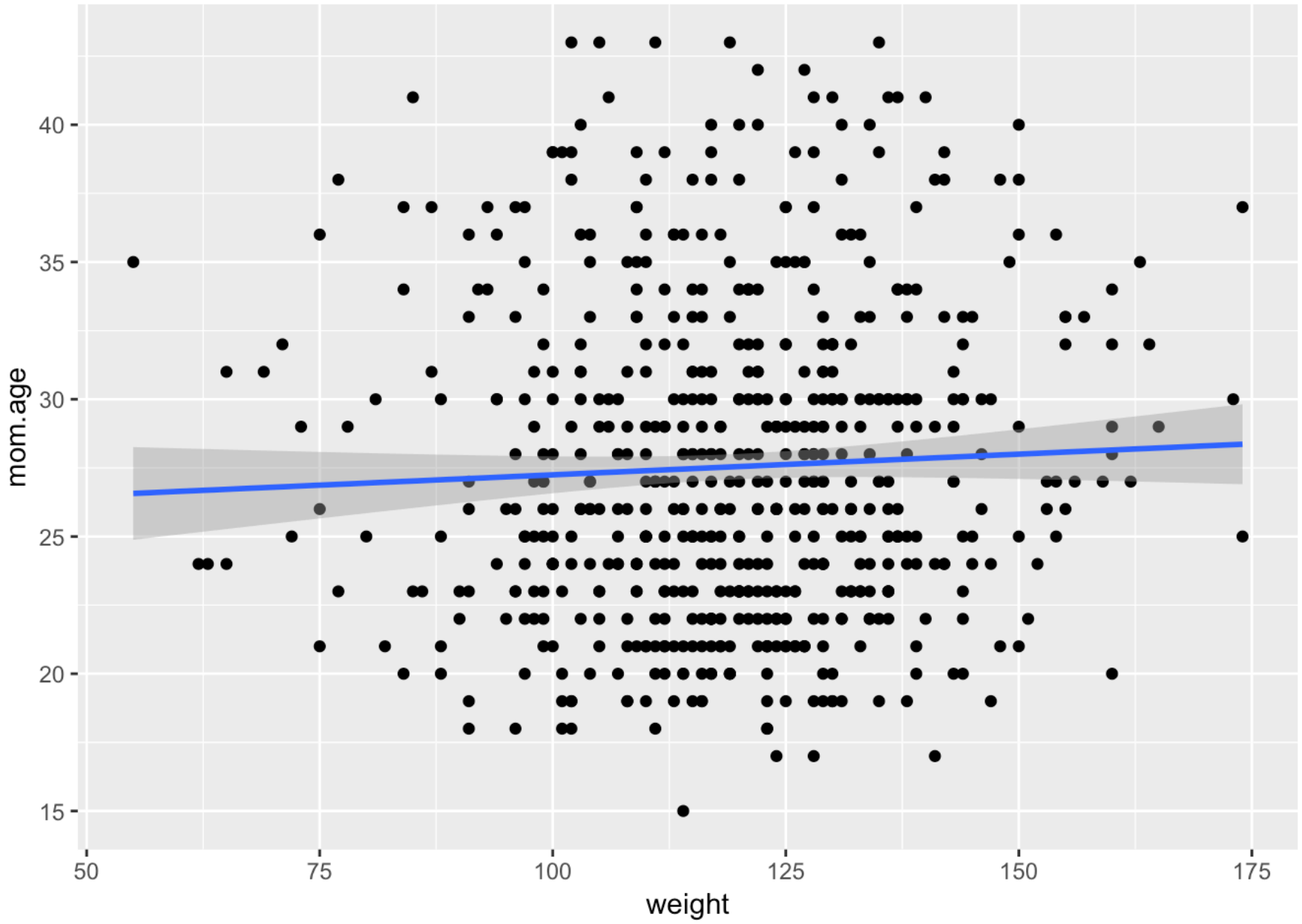
newborn weight ~ mom's weight model to newborn weight ~ mom's height model both are associated significantly

[20 points] In the dataset, there are different variables that might contribute to the newborn's weight. Determine which variables (you may choose to include all or only a subset of those variables) are associated with the newborn's weight. To learn more about the different variables, here is a link to a readme http://bit.ly/babies-weight- (http://bit.ly/babies-weight-) smoking-readme . Discuss the results.

```
p = ggplot(babies)
p = p + geom_point(aes(x = weight , y = mom.age ), alpha = 0.8, size = 2)
p = p + labs (x = "weight", y = "mom.age")
p = p + theme_light(base_size = 12)
print(p)
```




```
ggplot(data = babies) + geom_point(mapping = aes(x = weight, y = mom.age)) + geom_
smooth(aes(x = weight, y = mom.age), method = "lm")
```




```
model = lm(data = babies, formula = weight ~ mom.age)
broom::tidy(model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  115.        3.59      32.0  1.05e-132
## 2 mom.age        0.150     0.128      1.17 2.41e-  1
```

ß1 = 1.5 This means that the mean weight of newborn is estimated to differ by 0.14 grams for each one year difference in mom's age.

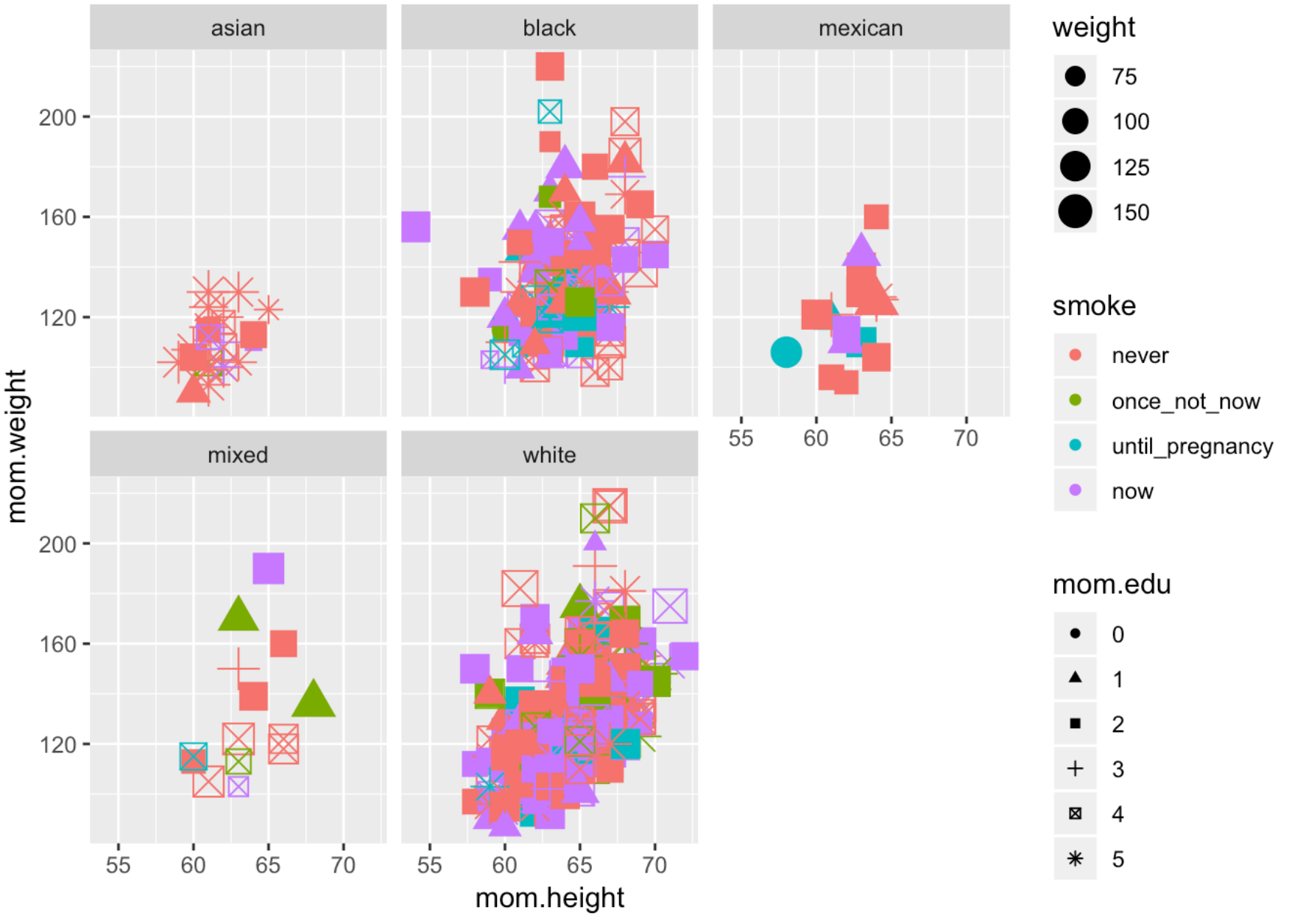
```
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ mom.age, data = babies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.408 -11.148   0.665  11.341  55.090
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 115.1666     3.5940  32.044   <2e-16 ***
## mom.age       0.1497     0.1277   1.173    0.241
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.31 on 608 degrees of freedom
## Multiple R-squared:  0.002256,   Adjusted R-squared:  0.0006153
## F-statistic: 1.375 on 1 and 608 DF,  p-value: 0.2414
```

We see that there's no significance between the mom's age and newborn weight

Bonus 2

```
ggplot(data = babies) +
  geom_point(mapping = aes(x = mom.height, y = mom.weight, shape = mom.edu, size=
weight, color= smoke)) + facet_wrap(~ mom.race, nrow = 2)
```



Smoking and education are the most affecting catergories in the black and white moms.