Study on 1,000 births in North Carolina

Ling Lee

Dataset obtained from the R Openintro package. Dataset contains information on births recorded in the state of North Carolina in 2004. There are 1000 observations and 13 variables Variables: father's age, mother's age, maturity status of mother, length of pregnancy, premature or full-term birth, number of hospital visits, mother's weight gain, baby's weight, low or normal birth weight, baby's gender, smoking habit, married or non-married, white or non-white

Here are the questions that we will be investigating in our study. 1. What are some factors that affect a baby's weight? 2. Do demographics reveal patterns on how often an expecting mother visits the hospital? 3. Do expecting mothers who smoke more likely to give birth prematurely?

```r
# removing all missing values for the variables that we will be using in our data
births_filter <- ncbirths %>%
  select(weight, whitemom, mature, habit, gender, premie, mage, gained, marital, visits, mage) %>%
  drop_na()
```

As there are some missing data in the dataset, I have decided to drop the empty values since our dataset is quite huge.

```r
# filter to include only full-term births
births_full_term <- births_filter %>%
  filter(premie == "full term")

# preparing data to perform dummy regression
habit_nonsmoker <- as.numeric(births_full_term$habit == "nonsmoker")
mature_young <- as.numeric(births_full_term$mature == "younger mom")
whitemom_white <- as.numeric(births_full_term$whitemom == "white")
gender_male <- as.numeric(births_full_term$gender == "male")
premie_full <- as.numeric(births_full_term$premie == "full term")

# regress baby's weight
weight_habit_mature <- linear_reg() %>%
  set_engine("lm") %>%
  fit(weight ~ habit_nonsmoker + whitemom_white + gender_male + gained + mage, data = births_full_term)

# adjusted r squared
glance(weight_habit_mature)$adj.r.squared
```

```
## [1] 0.09671251
```

```r
# generate values
weight_habit_mature %>% tidy()
```

```
## # A tibble: 6 × 5
##   term            estimate std.error statistic   p.value
##   <chr>              <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      6.16      0.206       29.9  4.51e-133
## 2 habit_nonsmoker  0.327     0.109        2.99 2.84e-  3
## 3 whitemom_white   0.325     0.0815       3.99 7.17e-  5
## 4 gender_male      0.520     0.0715       7.27 8.18e- 13
## 5 gained           0.00655   0.00253      2.59 9.72e-  3
## 6 mage             0.0118    0.00597      1.98 4.79e-  2
```

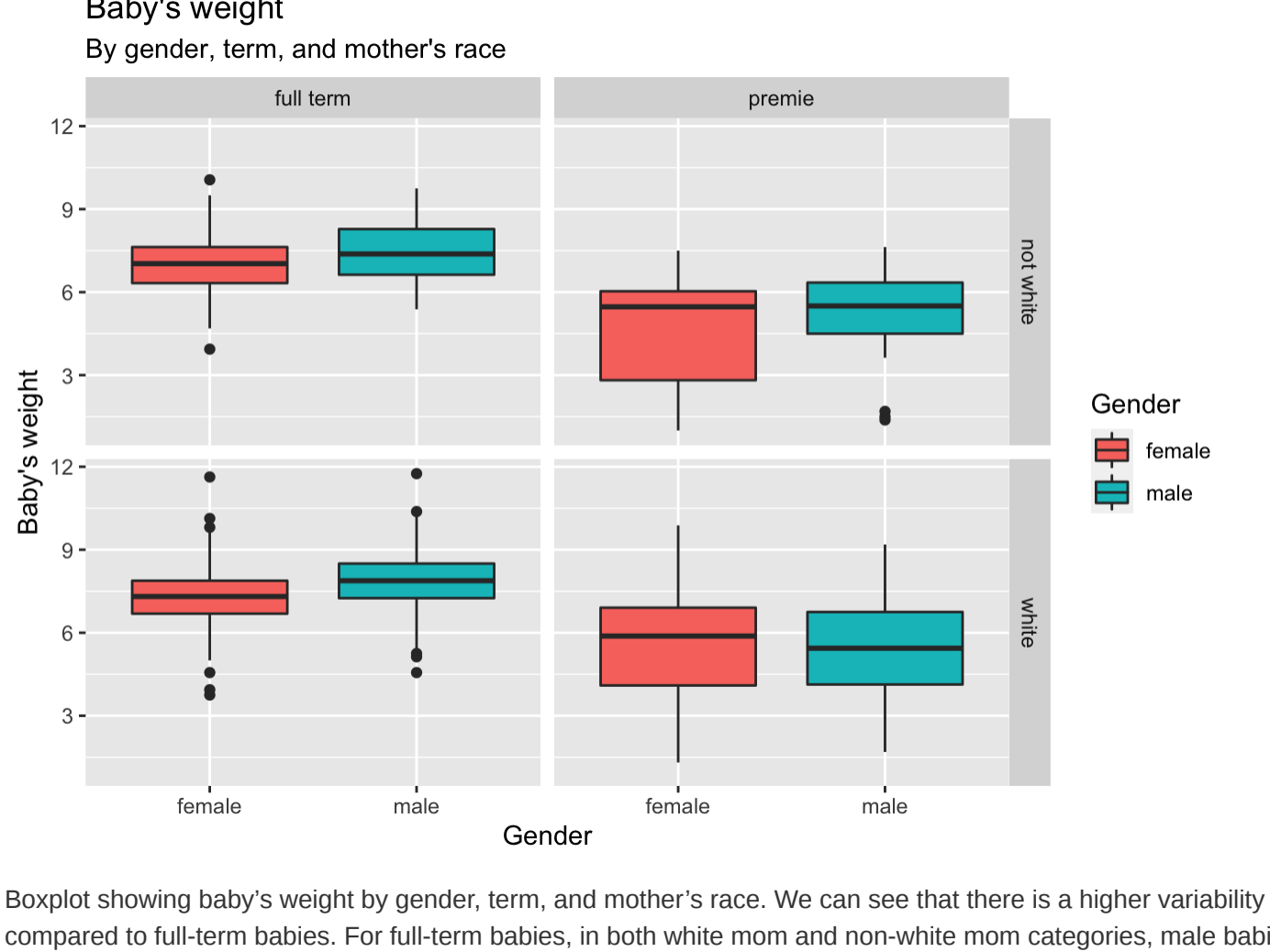
```r
weight_habit_mature
```

```
## parsnip model object
##
## Fit time:  4ms
##
## Call:
## stats::lm(formula = weight ~ habit_nonsmoker + whitemom_white +
##     gender_male + gained + mage, data = data)
##
## Coefficients:
##     (Intercept)  habit_nonsmoker   whitemom_white      gender_male
##        6.158083         0.326631         0.325295         0.519948
##          gained             mage
##        0.006547         0.011822
```

1. What are some factors that affect a baby's weight?

Only including full-term births, multiple regression shows that smoking habits, mother's race, weight gain, age and baby's gender predict a baby weight. All else held constant… 1. Baby boys are predicted, on average, to weigh 0.52 pounds heavier than baby girls. 2. Mothers who smoke are predicted, on average, to have babies weighing 0.33 pounds lighter than non-smoking mothers. 3. Mothers who are white are predicted, on average, to have babies weighing 0.33 pounds heavier than non-white mothers 4. For each additional increase of one pound gained by mother during pregnancy, we would expect the baby's weight to be higher, on average, by 0.007 pounds. 5. For each additional increase of one year in mother's age, we would expect the baby's weight to be higher, on average, by 0.01 pounds.

```r
births_filter %>%
ggplot(aes(x = factor(gender), y = weight, group = factor(gender), fill = factor(gender))) +
  facet_grid(whitemom~premie) +
  geom_boxplot() +
  labs(
  x = "Gender",
  y = "Baby's weight",
  title = "Baby's weight",
  subtitle = "By gender, term, and mother's race",
  fill = "Gender")
```



Boxplot showing baby's weight by gender, term, and mother's race. We can see that there is a higher variability in premature babies weight compared to full-term babies. For full-term babies, in both white mom and non-white mom categories, male babies weight heavier on average. However, we do not see the same pattern in premature babies. Perhaps, premature babies' weight vary depending on the length of pregnancy.

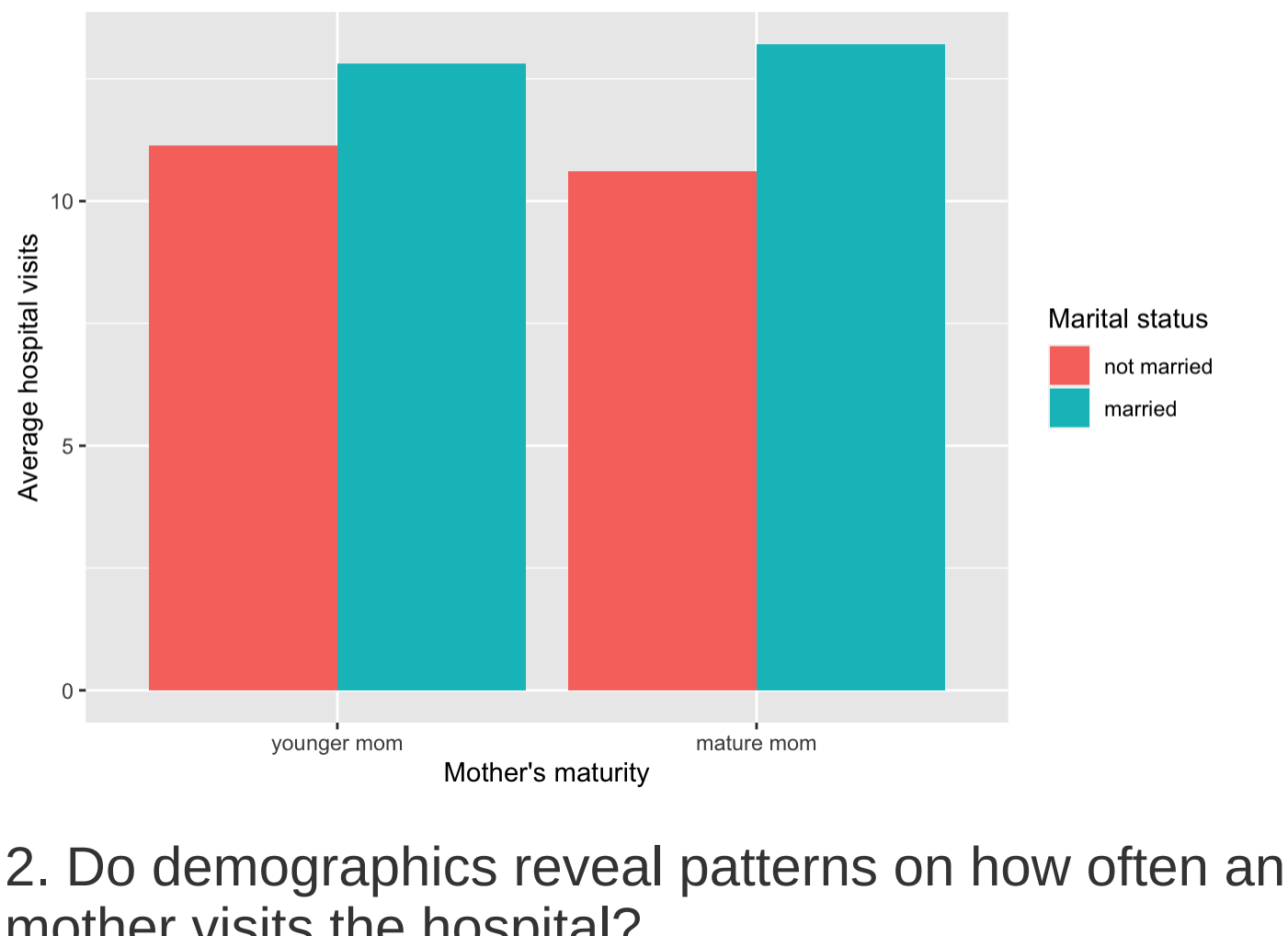
```r
# generate table group by marital status and maturity
births_marital_maturity <- births_filter %>%
  group_by(marital, mature) %>%
  summarise(average_visits = mean(visits), count = n())
```

```
## `summarise()` has grouped output by 'marital'. You can override using the `.groups` argument.
```

```r
births
```

```
## # A tibble: 150 × 9
##    f_age m_age weeks premature visits gained weight sex_baby smoke
##    <int> <int> <int> <fct>      <int>  <int>  <dbl> <fct>    <fct>
##  1    31    30    39 full term     13      1   6.88 male     smoker
##  2    34    36    39 full term      5     35   7.69 male     nonsmoker
##  3    36    35    40 full term     12     29   8.88 male     nonsmoker
##  4    41    40    40 full term     13     30   9    female   nonsmoker
##  5    42    37    40 full term     NA     10   7.94 male     nonsmoker
##  6    37    28    40 full term     12     35   8.25 male     smoker
##  7    35    35    28 premie         6     29   1.63 female   nonsmoker
##  8    28    21    35 premie         9     15   5.5  female   smoker
##  9    22    20    32 premie         5     40   2.69 male     smoker
## 10    36    25    40 full term     13     34   8.75 female   nonsmoker
## # … with 140 more rows
```

```r
# generate bar graph
ggplot(data = births_marital_maturity, aes(x = mature, y = average_visits, fill = marital)) +
  geom_bar(stat="identity", position=position_dodge()) +
  labs(
  x = "Mother's maturity",
  y = "Average hospital visits",
  title = "Expectant mothers' average hospital visits",
  subtitle = "By marital status and maturity",
  fill = "Marital status") +
  scale_x_discrete(limits = c("younger mom", "mature mom"))
```



2. Do demographics reveal patterns on how often an expecting mother visits the hospital?

In this dataset, they have divided mothers into two groups based on their age. Younger moms are those under 35 years old and mature moms are 35 or above. Regardless of the age, married mothers are more likely to visit the hospital. We can also find an interesting pattern. On average, mature married expecting mothers visit the hospital more than younger married expecting mothers. However, it is the reverse in unmarried mothers. Younger unmarried expecting mothers visit the hospital more than older unmarried mothers.

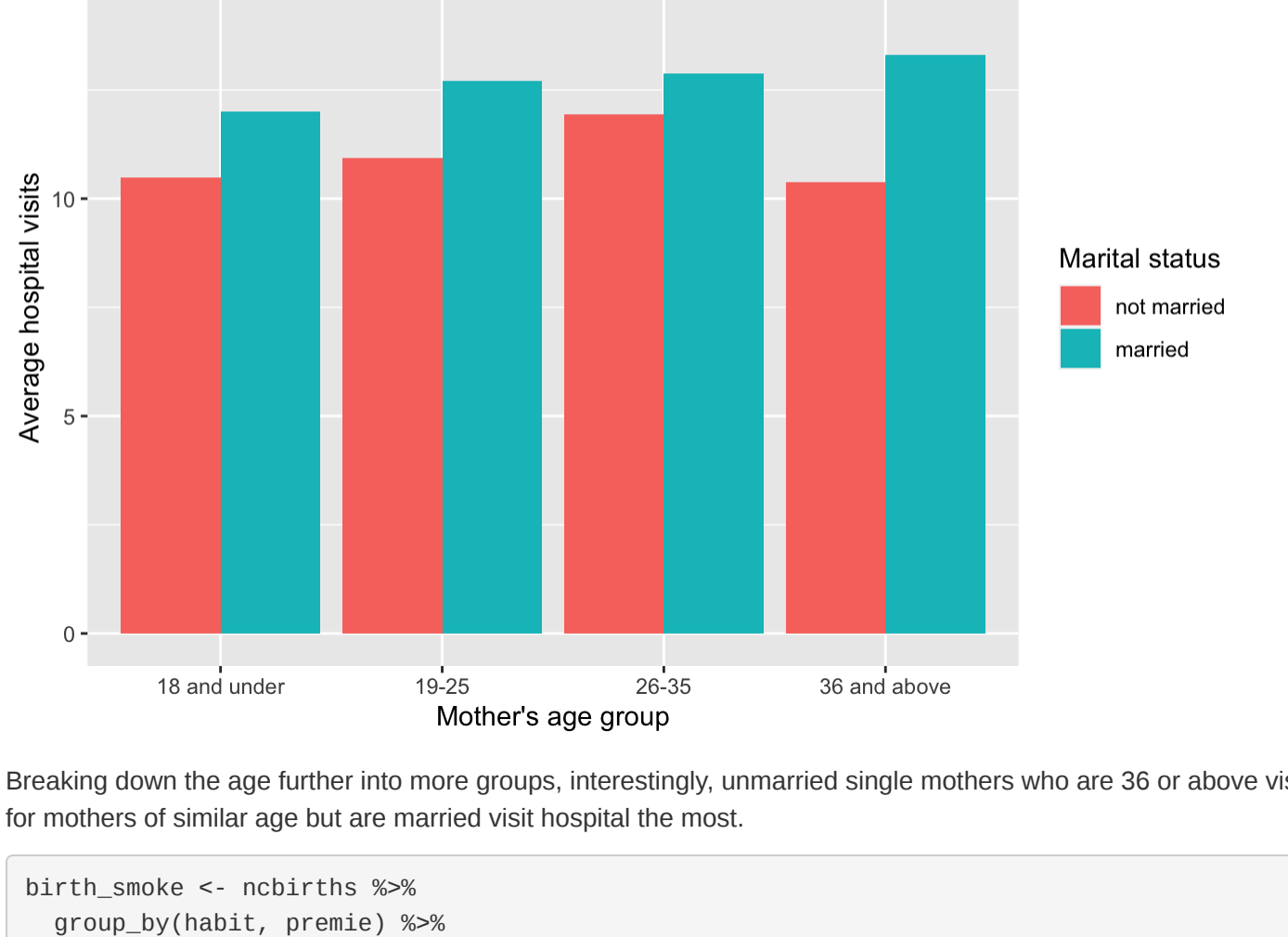
```r
# create a new age group
births_mage_grouped <- births_filter %>%
    mutate(
    age_group = case_when(
      mage <= 18 ~ "18 and under",
      mage >= 19 & mage <= 25 ~ "19-25",
      mage >= 26 & mage <= 35 ~ "26-35",
      mage >= 36 & mage <= 50 ~ "36 and above"))

# generate table group by new age group and maturity
births_marital_age_premie <- births_mage_grouped %>%
    group_by(marital, age_group) %>%
    summarise(average_visits = mean(visits), count = n())
```

```
## `summarise()` has grouped output by 'marital'. You can override using the `.groups` argument.
```

```r
births_marital_age_premie
```

```
## # A tibble: 8 × 4
## # Groups:   marital [2]
##   marital     age_group    average_visits count
##   <fct>       <chr>                 <dbl> <int>
## 1 not married 18 and under           10.5    63
## 2 not married 19-25                  10.9   185
## 3 not married 26-35                  11.9    97
## 4 not married 36 and above           10.4    18
## 5 married     18 and under           12       6
## 6 married     19-25                  12.7   172
## 7 married     26-35                  12.9   347
## 8 married     36 and above           13.3    74
```

```r
# generate bar graph
ggplot(data = births_marital_age_premie, aes(x = age_group, y = average_visits, fill = marital)) +
  geom_bar(stat="identity", position=position_dodge()) +
  ylim(0, 15) +
    labs(
  x = "Mother's age group",
  y = "Average hospital visits",
  title = "Expectant mothers' average hospital visits",
  subtitle = "By marital status and age group",
  fill = "Marital status")
```



Breaking down the age further into more groups, interestingly, unmarried single mothers who are 36 or above visits the hospital the least, whereas for mothers of similar age but are married visit hospital the most.

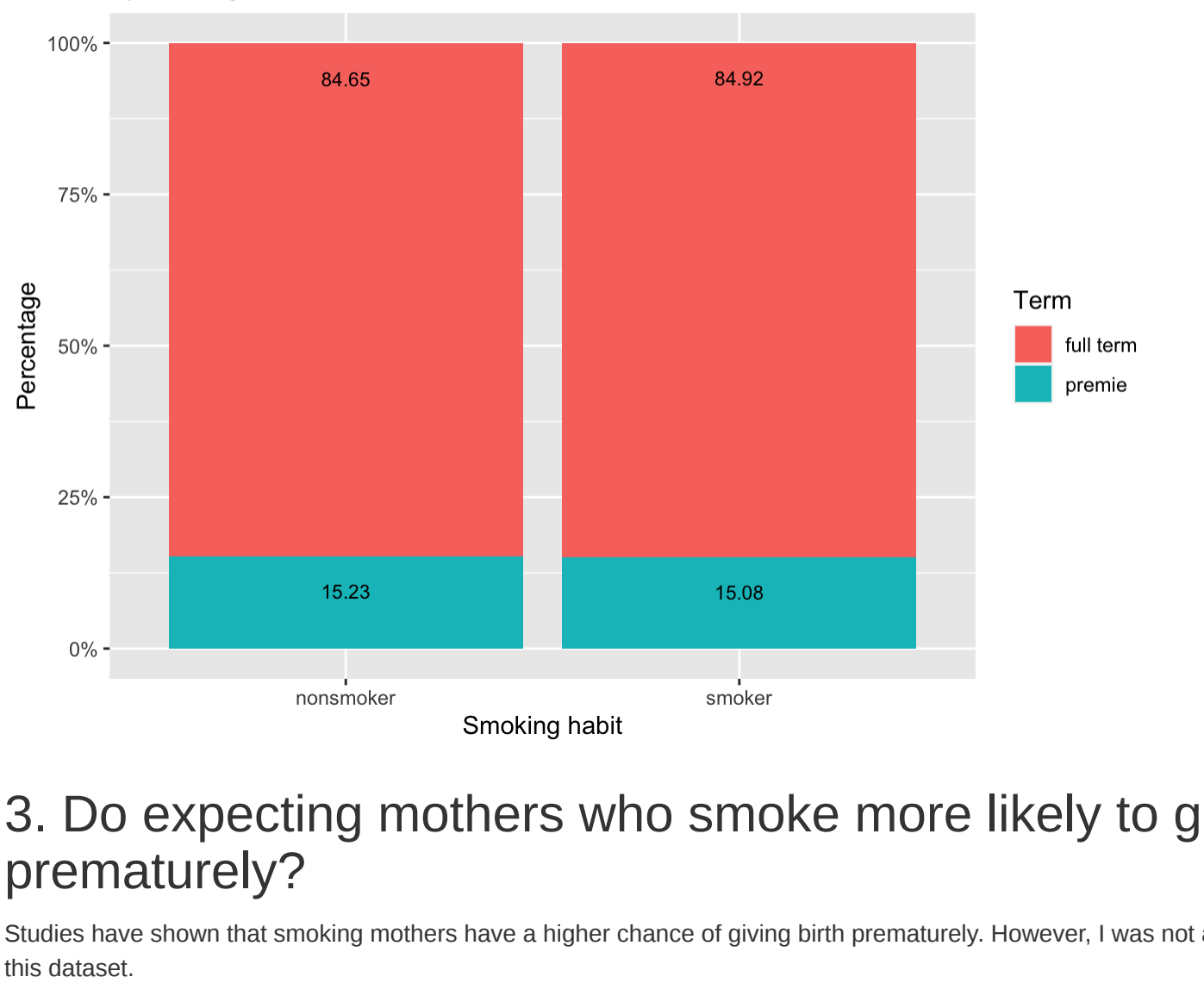
```r
birth_smoke <- ncbirths %>%
  group_by(habit, premie) %>%
  summarise(n = n()) %>%
  mutate(percentage = n/sum(n)) %>%
  drop_na()
```

```
## `summarise()` has grouped output by 'habit'. You can override using the `.groups` argument.
```

```r
birth_smoke
```

```
## # A tibble: 4 × 4
## # Groups:   habit [2]
##   habit     premie        n percentage
##   <fct>     <fct>     <int>      <dbl>
## 1 nonsmoker full term   739      0.847
## 2 nonsmoker premie      133      0.152
## 3 smoker    full term   107      0.849
## 4 smoker    premie       19      0.151
```

```r
# generate bar graph
ggplot(data = birth_smoke, aes(x = habit, y = percentage, fill = premie)) +
  geom_bar(position = "fill", stat="identity") +
  scale_y_continuous(labels = scales::percent) +
      labs(
  x = "Smoking habit",
  y = "Percentage",
  title = "Percentage of preemie and full-term babies (%)",
  subtitle = "By smoking habit",
  fill = "Term") +
  geom_text(aes(label = sprintf("%0.2f", round(percentage*100, digits = 2))), size = 3, vjust = 3, position = "st
ack")
```



3. Do expecting mothers who smoke more likely to give birth prematurely?

Studies have shown that smoking mothers have a higher chance of giving birth prematurely. However, I was not able to find the same result from this dataset.

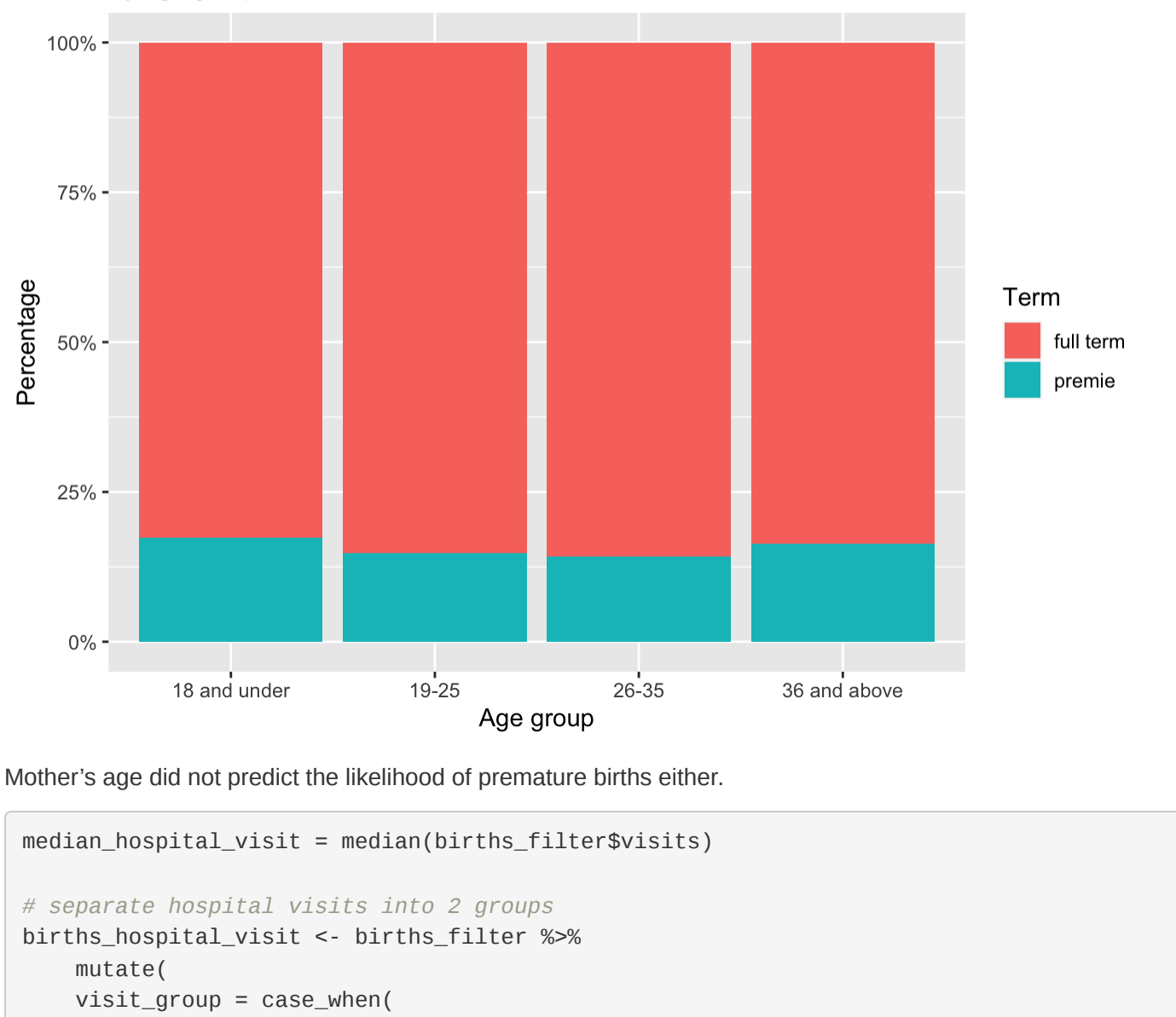
```r
# generate table based on term and age group
births_smoke_age <- births_mage_grouped %>%
  group_by(premie, age_group) %>%
  summarise(n = n()) %>%
  mutate(percentage = n/sum(n)) %>%
  drop_na()
```

```
## `summarise()` has grouped output by 'premie'. You can override using the `.groups` argument.
```

```r
births_smoke_age
```

```
## # A tibble: 8 × 4
## # Groups:   premie [2]
##   premie    age_group        n percentage
##   <fct>     <chr>        <int>      <dbl>
## 1 full term 18 and under    57     0.0696
## 2 full term 19-25          304     0.371
## 3 full term 26-35          381     0.465
## 4 full term 36 and above    77     0.0940
## 5 premie    18 and under    12     0.0839
## 6 premie    19-25           53     0.371
## 7 premie    26-35           63     0.441
## 8 premie    36 and above    15     0.105
```

```r
# generate bar graph
ggplot(data = births_smoke_age, aes(x = age_group, y = n, fill = premie)) +
  geom_bar(stat="identity", position="fill") +
  scale_y_continuous(labels = scales::percent) +
    labs(
  x = "Age group",
  y = "Percentage",
  title = "Percentage of preemie and full-term babies (%)",
  subtitle = "By age group",
  fill = "Term")
```



Mother's age did not predict the likelihood of premature births either.

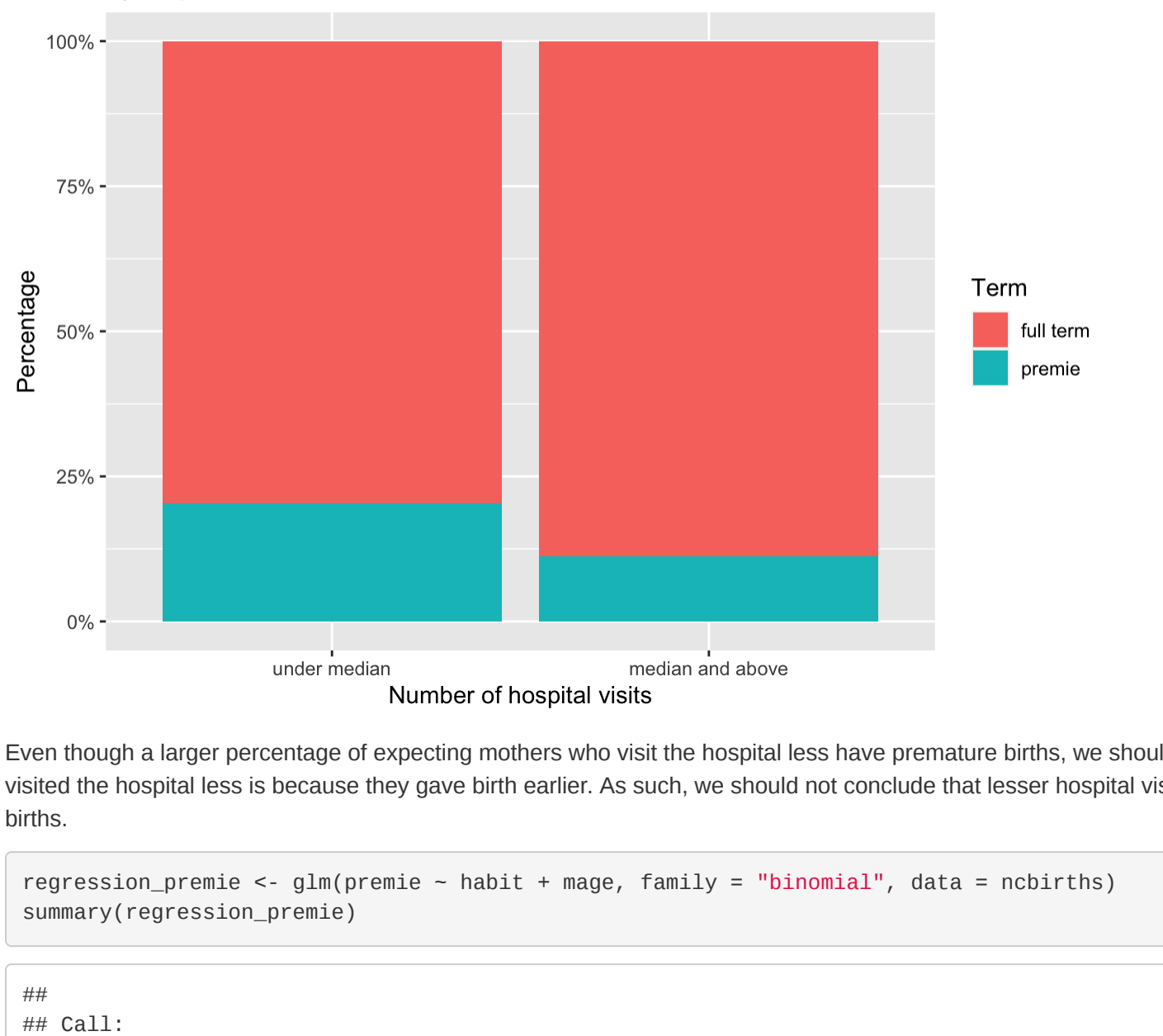
```r
median_hospital_visit = median(births_filter$visits)

# separate hospital visits into 2 groups
births_hospital_visit <- births_filter %>%
    mutate(
    visit_group = case_when(
      visits < median_hospital_visit ~ "under median",
      visits >= median_hospital_visit ~ "median and above"))

# generate table based on term and hospital visits
births_premie_hospital <- births_hospital_visit %>%
  group_by(premie, visit_group) %>%
  summarise(n = n()) %>%
  mutate(percentage = n/sum(n)) %>%
  drop_na()
```

```
## `summarise()` has grouped output by 'premie'. You can override using the `.groups` argument.
```

```r
births_premie_hospital
```

```
## # A tibble: 4 × 4
## # Groups:   premie [2]
##   premie    visit_group          n percentage
##   <fct>     <chr>            <int>      <dbl>
## 1 full term median and above   515      0.629
## 2 full term under median       304      0.371
## 3 premie    median and above    65      0.455
## 4 premie    under median        78      0.545
```

```r
# generate bar graph
ggplot(data = births_premie_hospital, aes(x = visit_group, y = n, fill = premie)) +
  geom_bar(stat="identity", position="fill") +
  scale_y_continuous(labels = scales::percent) +
    labs(
  x = "Number of hospital visits",
  y = "Percentage",
  title = "Percentage of preemie and full-term babies (%)",
  subtitle = "By hospital visits",
  fill = "Term") +
  scale_x_discrete(limits = c("under median", "median and above"))
```



Even though a larger percentage of expecting mothers who visit the hospital less have premature births, we should find out if the reason they visited the hospital less is because they gave birth earlier. As such, we should not conclude that lesser hospital visits could lead to premature births.

```r
regression_premie <- glm(premie ~ habit + mage, family = "binomial", data = ncbirths)
summary(regression_premie)
```

```
##
## Call:
## glm(formula = premie ~ habit + mage, family = "binomial", data = ncbirths)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5881  -0.5782  -0.5737  -0.5684   1.9670
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.623255   0.399838  -4.060 4.91e-05 ***
## habitsmoker -0.020102   0.267663  -0.075    0.940
## mage        -0.003373   0.014318  -0.236    0.814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 851.67  on 997  degrees of freedom
## Residual deviance: 851.61  on 995  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 857.61
##
## Number of Fisher Scoring iterations: 4
```

Through logistic regression, I regressed term by smoking habit and mother's age. The p-values were not significant. Therefore, smoking habit and mother's age did not predict premature births in our dataset.

```r
births_visits <- births_filter %>%
  specify(response = premie, success = "premie") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop")

births_visits %>%
  summarize(lower = quantile(stat, 0.025),
            upper = quantile(stat, 0.975))
```

```
## # A tibble: 1 × 2
##   lower upper
##   <dbl> <dbl>
## 1 0.126 0.172
```

We are 95% confident that the proportion of the population who will give birth prematurely is between 0.13 and 0.17.