R Commands for Quantitative Variables

September 17, 2018

Birth Weights and Tobacco Use During Pregnancy

Are babies' birth weights affected by whether or not the mother used tobacco during pregnancy? Low birth weights are a risk factor for health problems later in life, so this is important!

Here is a data set with data on a sample of randomly selected babies who were born in North Carolina in 2004, with some information about the mother and the babies' weights in grams:

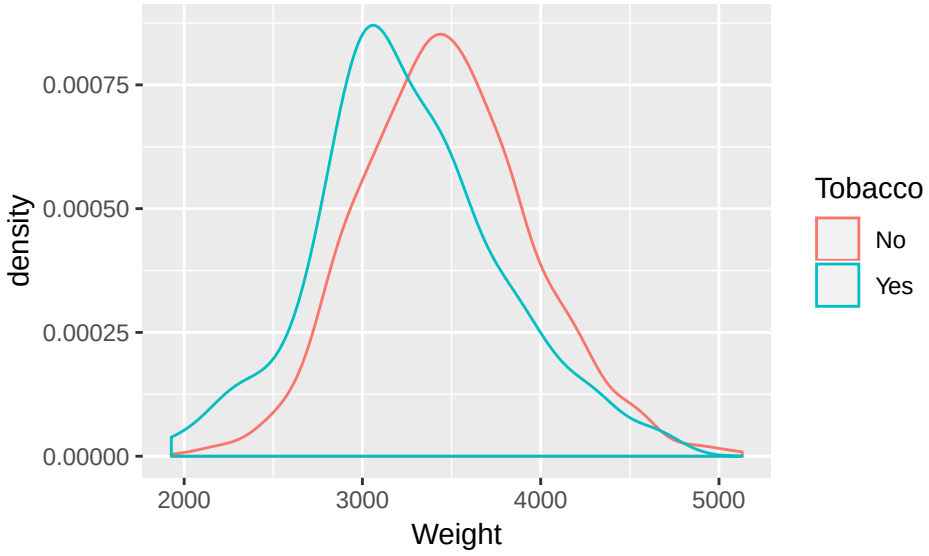
```
dim(NCBirths2004)
```

```
## [1] 1009    7
```

```
head(NCBirths2004)
```

```
##   ID MothersAge Tobacco Alcohol Gender Weight Gestation
## 1  1      30-34      No      No   Male   3827        40
## 2  2      30-34      No      No   Male   3629        38
## 3  3      35-39      No      No Female   3062        37
## 4  4      20-24      No      No Female   3430        39
## 5  5      25-29      No      No   Male   3827        38
## 6  6      35-39      No      No Female   3119        39
```

Here's a plot of the data:



Warm Up #1: What did the code to make that plot look like? Fill in the blanks below.

There are 4 blanks: what was used for data? For the aesthetic mapping to x? For the aesthetic mapping to color? For the geometry?

```
ggplot(data =                    ,
    mapping = aes(x =                , color =                )) +
  geom_                  ()
```

Warm Up #2: What statistics would you use to summarize the center and spread of the distribution of birth weights within each group?

Calculating Summaries of Quantitative Variables: `summarize` and `group_by`

We saw from the plot that there seems to be a difference in birth weights. What are "typical" birth weights for the two groups?

```
NCBirths2004 %>%
  group_by(Tobacco) %>%
  summarize(
    mean_wt = mean(Weight),
    median_wt = median(Weight),
    q1_wt = quantile(Weight, probs = 0.25),
    q3_wt = quantile(Weight, probs = 0.75),
    iqr_wt = IQR(Weight),
    var_wt = var(Weight),
    sd_wt = sd(Weight)
  )
```

```
## # A tibble: 2 x 8
##   Tobacco mean_wt median_wt q1_wt q3_wt iqr_wt  var_wt sd_wt
##   <fct>     <dbl>     <dbl> <dbl> <dbl>  <dbl>   <dbl> <dbl>
## 1 No        3472.      3459  3147 3771    624  229012.  479.
## 2 Yes       3257.      3204  2948 3530.   582. 270898.  520.
```

Note: if we wanted to just calculate these summaries for all babies combined (across both groups), we would eliminate the `group_by` command:

```
NCBirths2004 %>%
  summarize(
    mean_wt = mean(Weight),
    median_wt = median(Weight),
    q1_wt = quantile(Weight, probs = 0.25),
    q3_wt = quantile(Weight, probs = 0.75),
    iqr_wt = IQR(Weight),
    var_wt = var(Weight),
    sd_wt = sd(Weight)
  )
```

```
##   mean_wt median_wt q1_wt q3_wt iqr_wt   var_wt   sd_wt
## 1 3448.26      3430  3119  3771    652 237886.4 487.736
```

Sorting the data: `arrange`

What was the shortest gestation time?

To answer this question, we will `arrange` the babies in increasing order of gestation time. Then, the babies with the shortest gestation times will be near the `head` of the data frame.

```
NCBirths_by_gestation <- NCBirths2004 %>%
  arrange(Gestation)

head(NCBirths_by_gestation)
```

```
##   ID MothersAge Tobacco Alcohol Gender Weight Gestation
## 1  3      35-39      No      No Female   3062        37
## 2 15      30-34      No      No   Male   3232        37
## 3 20      25-29      No      No   Male   3005        37
## 4 29      15-19      No      No Female   2863        37
## 5 31      20-24      No      No   Male   2155        37
## 6 32      20-24     Yes      No Female   3062        37
```

What was the longest gestation time?

Now we want to `arrange` the babies in `desc`ending order of gestation time, so the longest gestation times will be at the `head` of the data frame:

```
NCBirths_by_gestation_descending <- NCBirths2004 %>%
  arrange(desc(Gestation))

head(NCBirths_by_gestation_descending)
```

```
##    ID MothersAge Tobacco Alcohol Gender Weight Gestation
## 1  79      30-34      No      No   Male   4139        42
## 2 165      20-24      No      No   Male   3799        42
## 3 359      20-24      No      No Female   4224        42
## 4 384      30-34      No      No Female   3572        42
## 5 521      20-24      No      No Female   3430        42
## 6 585      35-39      No      No Female   3629        42
```

Modifying or Adding a Variable: `mutate`

The babies' birth `Weights` are currently recorded in grams. Let's add a new variable to the data frame called `Weight_lbs` with the babies' birth weights in pounds.

There are about 453.6 grams in a pound. We can divide the weight in grams by 453.6 to get the weight in pounds.

```
NCBirths2004 <- NCBirths2004 %>%
  mutate(
    Weight_lbs = Weight / 453.6
  )
```

As usual, we can take a look at the results by looking at the output from `head` or `str`.

```
head(NCBirths2004)
```

```
##   ID MothersAge Tobacco Alcohol Gender Weight Gestation Weight_lbs
## 1  1      30-34      No      No   Male   3827        40   8.436949
## 2  2      30-34      No      No   Male   3629        38   8.000441
## 3  3      35-39      No      No Female   3062        37   6.750441
## 4  4      20-24      No      No Female   3430        39   7.561728
## 5  5      25-29      No      No   Male   3827        38   8.436949
## 6  6      35-39      No      No Female   3119        39   6.876102
```