Project 1 Example Solution

This is the dataset you will be working with:

```
NCbirths <- read_csv("https://wilkelab.org/classes/SDS348/data_sets/NCbirths.
csv")

NCbirths
```

```
# A tibble: 1,409 × 10
   Plural   Sex MomAge Weeks Gained Smoke BirthWeightGm   Low Premie Marital
    <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>         <dbl> <dbl>  <dbl>   <dbl>
 1      1     1     32    40     38     0         3147.     0      0       0
 2      1     2     32    37     34     0         3289.     0      0       0
 3      1     1     27    39     12     0         3912.     0      0       0
 4      1     1     27    39     15     0         3856.     0      0       0
 5      1     1     25    39     32     0         3430.     0      0       0
 6      1     1     28    43     32     0         3317.     0      0       0
 7      1     2     25    39     75     0         4054.     0      0       0
 8      1     2     15    42     25     0         3204.     0      0       1
 9      1     2     21    39     28     0         3402      0      0       0
10      1     2     27    40     37     0         3515.     0      0       1
# ℹ 1,399 more rows
```

Questions:

1. Is there a relationship between whether a mother smokes or not and her baby's weight at birth?

2. How many mothers are smokers or non-smokers?

3. What are the age distributions of mothers of twins or triplets?

Introduction: We are working with the `NCbirths` dataset, which contains 1409 birth records from North Carolina in 2001. In this dataset, each row corresponds to one birth, and there are ten columns providing information about the birth, the mother, and the baby. Information about the birth includes whether it is a single, twin, or triplet birth, the number of completed weeks of gestation, and whether the birth is premature. Information about the baby includes the sex, the weight at birth, and whether the birth weight should be considered low. Information about the mother includes her age, the weight gained during pregnancy, whether she is a smoker, and whether she is married.

To answer the three questions, we will work with five variables, the baby's birthweight (column `BirthWeightGm`), whether the baby was born prematurely (column `Premie`), whether it was a singleton, twin, or triplet birth (column `Plural`), whether the mother is a smoker or not (column `Smoke`), and the mother's age (column `MomAge`). The birthweight is provided as a numeric value, in grams. The premature birth status is encoded as 0/1, where 0 means regular and 1 means premature (36 weeks or sooner). The number of births is encoded as 1/2/3 representing singleton, twins, and triplets, respectively. The smoking status is encoded as 0/1, where 0 means the mother is not a smoker and 1 means she is a smoker. The mother's age is provided in years.

Approach: To show the distributions of birthweights versus the mothers' smoking status we will be using violin plots (`geom_violin()`). We also separate out regular and premature births, because babies born prematurely have much lower birthweight and therefore must be considered separately. Violins make it easy to compare multiple distributions side-by-side.

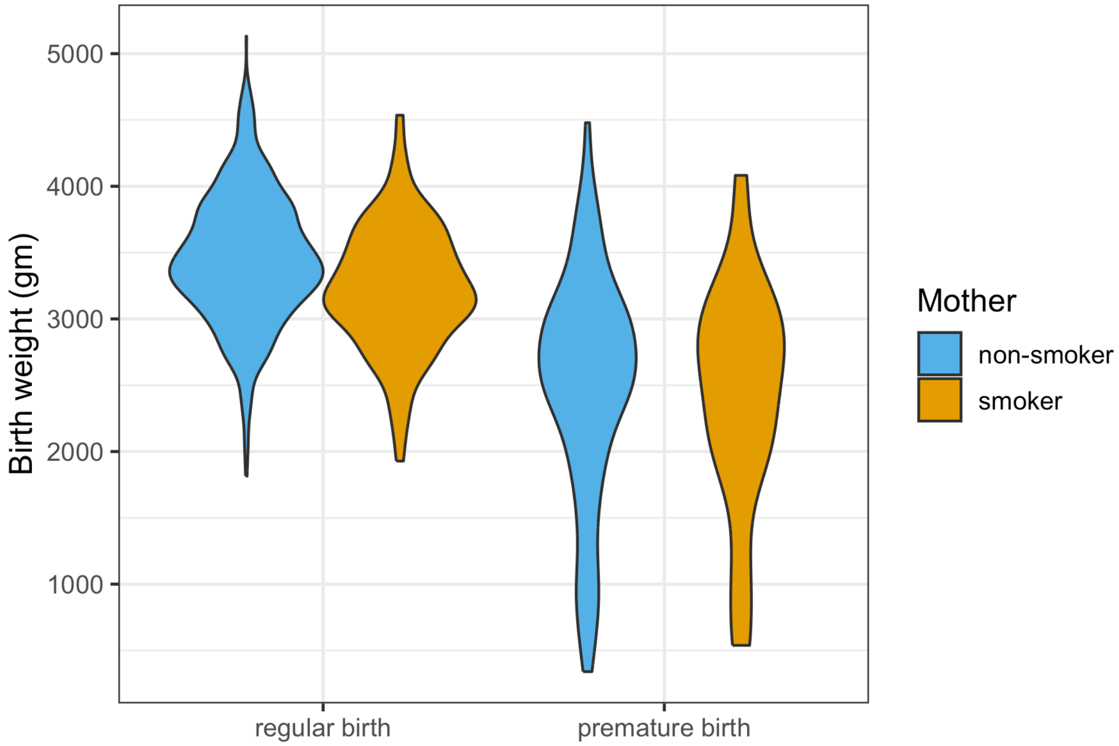
To show the number of mothers that are smokers or non-smokers we will use a simple bar plot (`geom_bar()`). Finally, to show the distribution of mothers' ages we will make a strip chart. The number of twin and triplet births in the dataset is not that large, so a strip chart is a good option here.

Analysis:

Question 1: Is there a relationship between whether a mother smokes or not and her baby's weight at birth?

To answer this question, we plot the birthweight distributions as violins, separated by both smoking status and by whether the birth was regular or premature.

```r
# The columns `Premie` and `Smoke` are numerical but contain
# categorical data, so we convert to factors to ensure ggplot
# treats them correctly
ggplot(NCbirths, aes(factor(Premie), BirthWeightGm)) +
  geom_violin(aes(fill = factor(Smoke))) +
  scale_x_discrete(
    name = NULL, # remove axis title entirely
    labels = c("regular birth", "premature birth")
  ) +
  scale_y_continuous(
    name = "Birth weight (gm)"
  ) +
  scale_fill_manual(
    name = "Mother",
    labels = c("non-smoker", "smoker"),
    # explicitly assign colors to specific data values
    values = c(`0` = "#56B4E9", `1` = "#E69F00")
  ) +
  theme_bw(12)
```

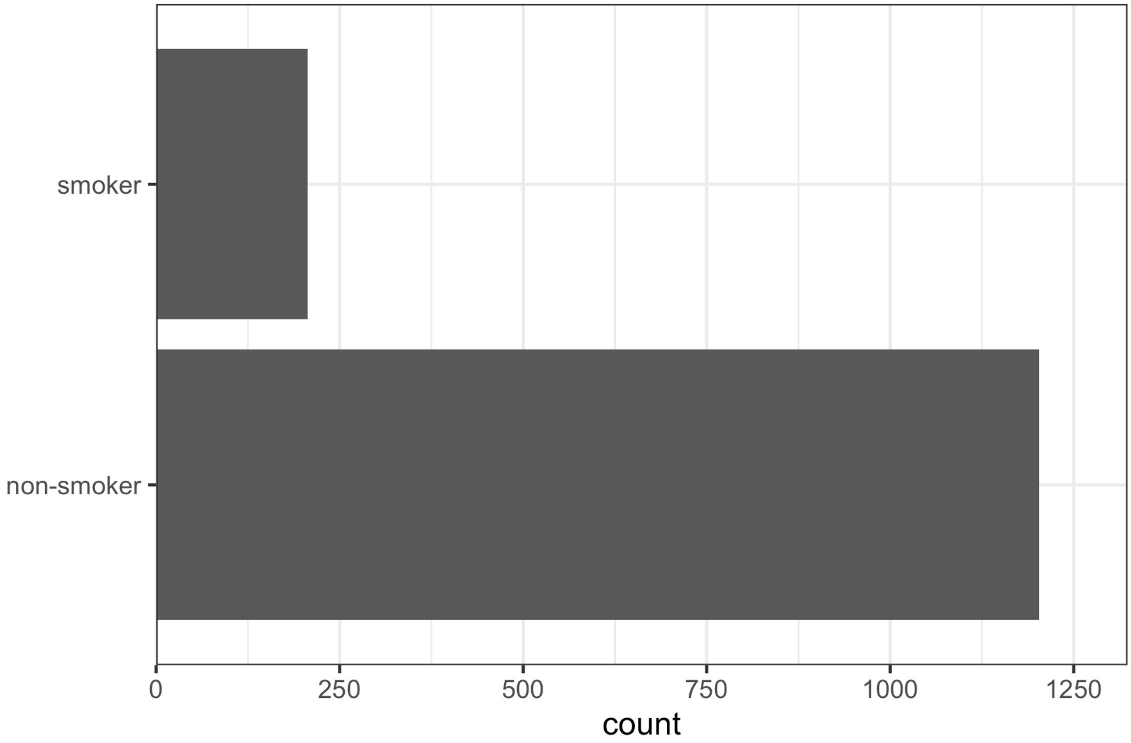


There is a clear difference between birthweight for regular and premature births, and for regular births the birthweight also seems to be lower when the mother smokes.

Question 2: How many mothers are smokers or non-smokers?

To answer this question, we make a simple bar plot of the number of mothers by smoking status.

```
# again, convert `Smoke` into factor so it's categorical
ggplot(NCbirths, aes(y = factor(Smoke))) +
  geom_bar() +
  scale_y_discrete(
    name = NULL,
    labels = c("non-smoker", "smoker")
  ) +
  scale_x_continuous(
    # ensure there's no gap between the beginning of the bar
    # and the edge of the plot panel
    expand = expansion(mult = c(0, 0.1))
  ) +
  theme_bw(12)
```

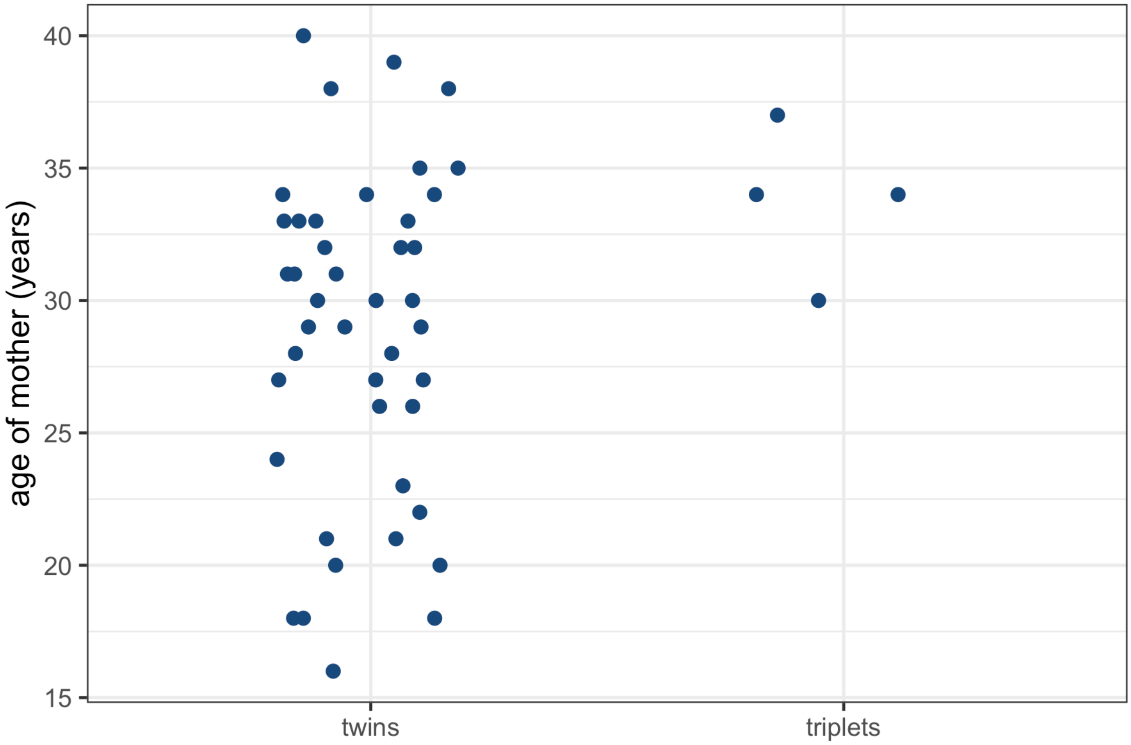


The vast majority of mothers in the dataset are non-smokers (almost 1250). Fewer than 250 are smokers.

Question 3. What are the age distributions of mothers of twins or triplets?

To answer this question, we first remove singleton births from the dataset and then show age distributions as a strip chart.

```r
NCbirths %>%
  filter(Plural > 1) %>% # remove singlet births
  ggplot(aes(x = factor(Plural), y = MomAge)) +
  geom_point(
    # jitter horizontally so points don't overlap
    position = position_jitter(
      width = 0.2,
      height = 0
    ),
    # it's nice to make points a little bigger and give them some color
    size = 2,
    color = "#1E4A7F"
  ) +
  scale_x_discrete(
    name = NULL,
    labels = c("twins", "triplets")
  ) +
  scale_y_continuous(
    name = "age of mother (years)"
  ) +
  theme_bw(12)
```



Mothers of twins span the entire childbearing range, from 15 years to approximately 40 years old. By contrast, mothers of triplets tend to be in their thirties.

Discussion: The smoking status of the mother appears to have a small effect on the average birth weight for regular births. We can see this by comparing the two left-most violins in the first plot, where we see that they are slightly vertically shifted relative to each other but have otherwise a comparable shape. However, a much bigger effect comes from whether the baby is born prematurely or not. Premature births have on average a much lower birthweight than regular births, and the variance is also bigger (the two right-most violins are taller than the two left-most violins). Interestingly, smoking status does not seem to affect the distribution of birthweights for premature births much. We can see this from the fact that the two right-most violins look approximately the same. We would have to run a multivariate statistical analysis to determine whether any of these observed patterns are statistically significant.

There are many more births to non-smoking mothers than to smoking mothers in the dataset. This is important because it means we have more complete data for non-smoking mothers. Some of the differences we saw in the first graph, such as the slightly lower variance in birthweight for premature births to smoking mothers—as compared to premature births to non-smoking mothers —may simply be due to a smaller data set.

When comparing age distributions of mothers of twins or of triplets we see an unexpected difference. It appears that mothers of all ages, from teenage moms to moms in their early fourties, all can have twins. By contrast, only mothers in their thirties appear to have triplets. We can think of a possible explanation. Twin births happen due to natural causes and therefore can occur in mothers of all ages. Triplet births, however, are extremely unlikely to occur naturally, and most commonly are caused by fertility treatments that cause multiple eggs to mature at once. It is unlikely that women in their late teens or twenties will undergo fertility treatment, whereas women in their thirties do so frequently. We also note, however, that there are only four triplet births in the dataset, so the lack of younger mothers could be due to random chance. We would have to perform further analysis or run statistical tests develop a clearer picture of what mechanisms may have caused the observed patterns in the data.