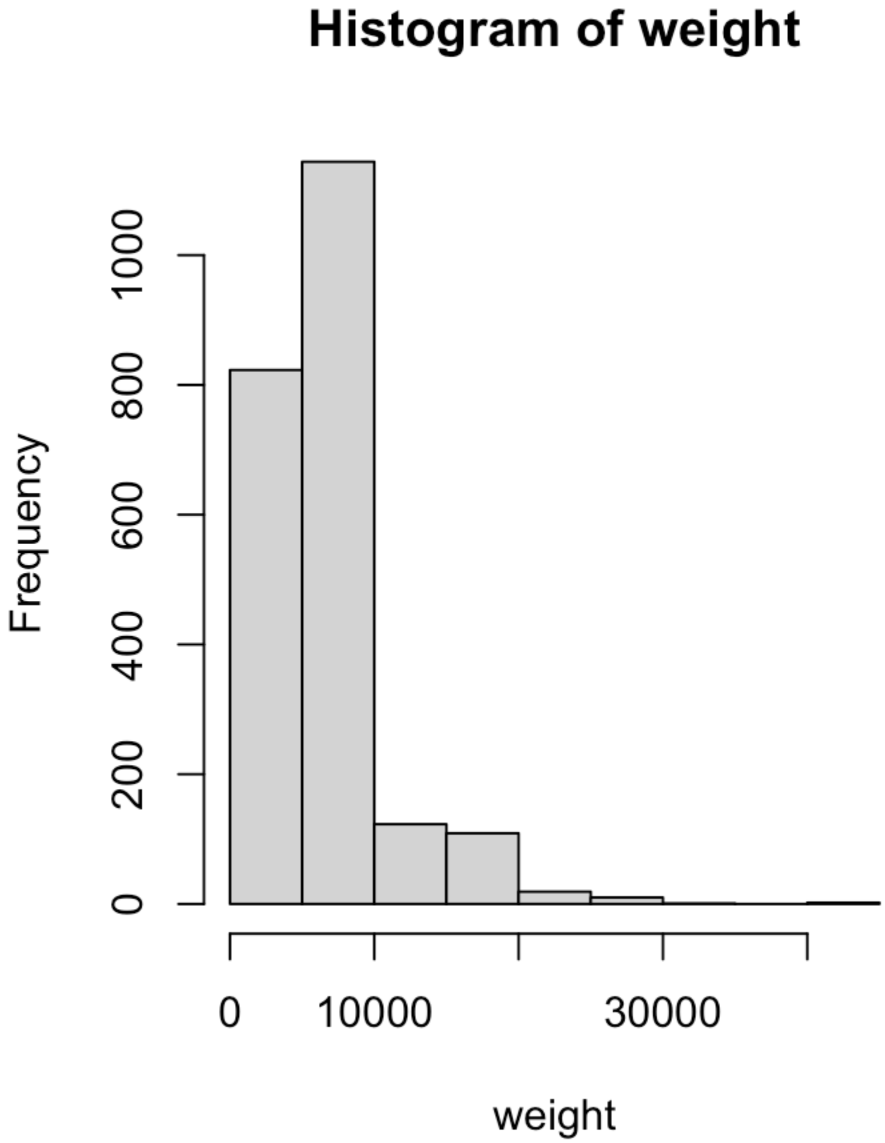
Homework 2 **Name:** Ivan Wang

(Parts 4-5; 40 pts)

1. A survey was taken in 2011 to assess the Canadian electorate's opinions on abortion. A subset of the data can be accessed by loading the **car** library, then using the command data(CES11). (15 pts)

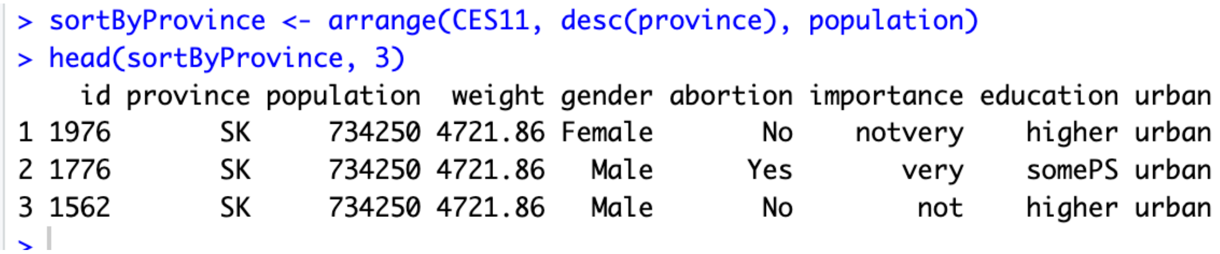
a) Extract the **weight** variable into its own vector object. Use this vector to build a histogram.

```
weight <- c(CES11$weight)
```



Store weights into vector and then create histogram.

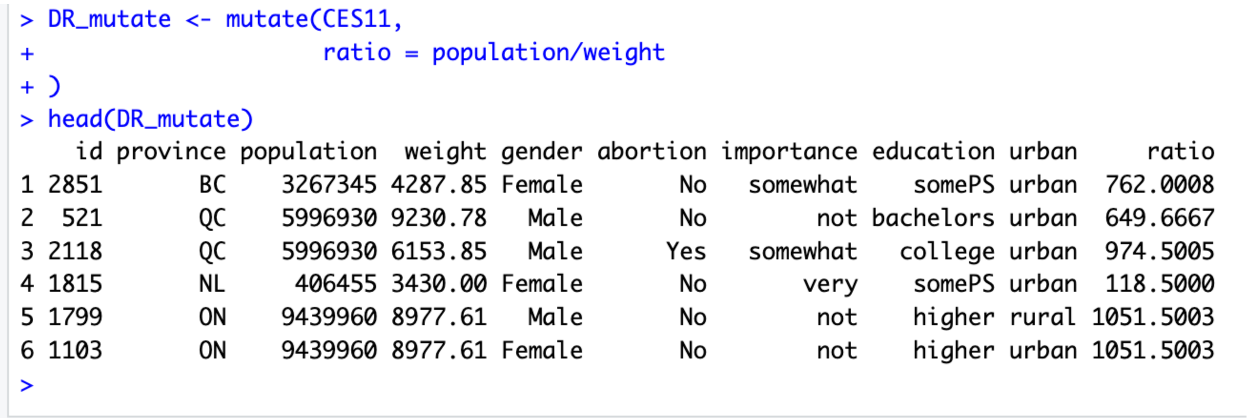
b) Sort the data set by both the **province** variable in reverse alphabetical order and then by **population** (ascending). This should be done as a single sort. Print the first 3 observations.



```
> sortByProvince <- arrange(CES11, desc(province), population)
> head(sortByProvince, 3)
    id province population  weight gender abortion importance education urban
1 1976       SK     734250 4721.86 Female       No    notvery    higher urban
2 1776       SK     734250 4721.86   Male      Yes       very    somePS urban
3 1562       SK     734250 4721.86   Male       No        not    higher urban
>
```

Store dataset by province by reverse alphabetical order and by population in ascending order

c) Add a new column called **ratio**, calculated as population divided by weight.



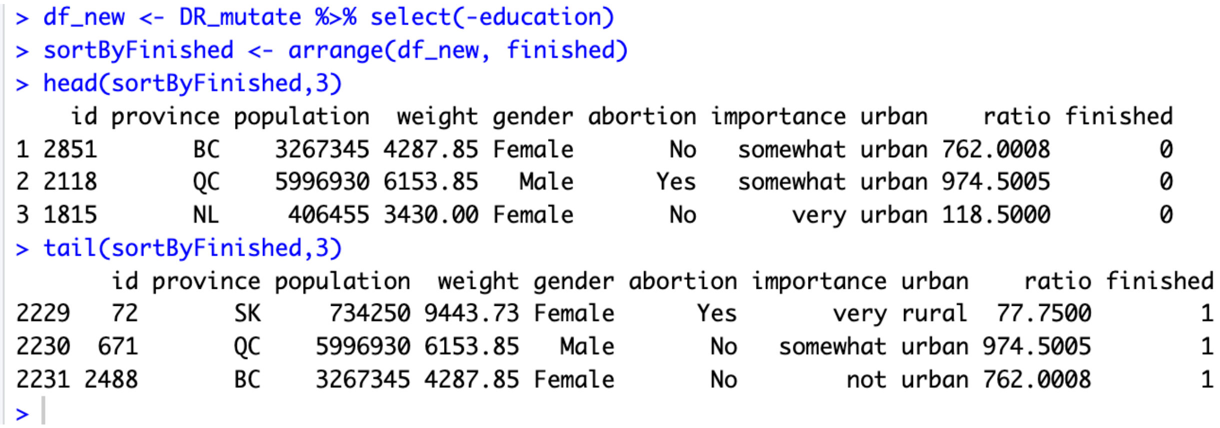
```
> DR_mutate <- mutate(CES11,
+                     ratio = population/weight
+ )
> head(DR_mutate)
    id province population  weight gender abortion importance education urban     ratio
1 2851       BC    3267345 4287.85 Female       No   somewhat    somePS urban  762.0008
2  521       QC    5996930 9230.78   Male       No        not bachelors urban  649.6667
3 2118       QC    5996930 6153.85   Male      Yes   somewhat   college urban  974.5005
4 1815       NL     406455 3430.00 Female       No       very    somePS urban  118.5000
5 1799       ON    9439960 8977.61   Male       No        not    higher rural 1051.5003
6 1103       ON    9439960 8977.61 Female       No        not    higher urban 1051.5003
>
```

Make new column ratio which is the population divided by weight.

d) The column **education** is a categorical variable about the subject's highest level of education. Create a simplified column called **finished** that records whether a subject has received a traditional 4-year college degree using the following definition:

$$\text{finished} = \begin{cases} 1 \text{ if education} \in \{\text{bachelors, higher}\} \\ 0 \text{ if education} \in \{\text{college, HS, lessHS, somePS}\}) \end{cases}$$

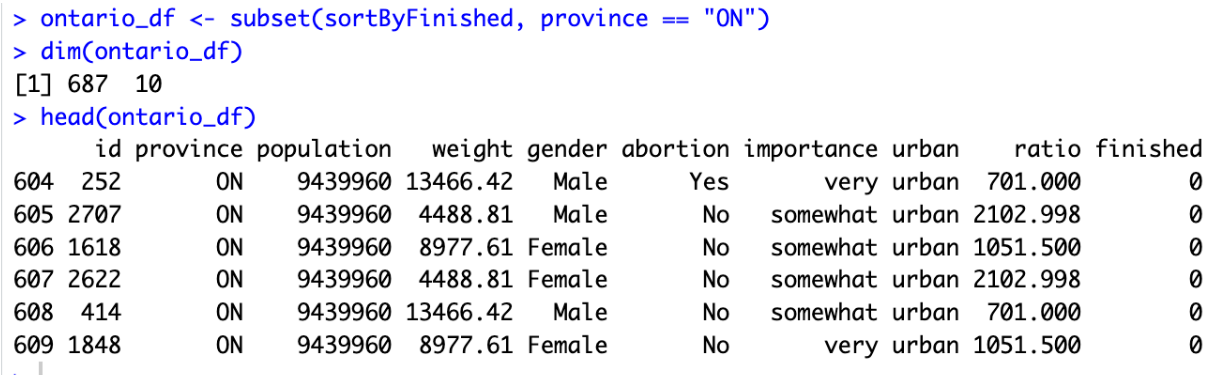
After doing this, remove the original **education** variable. Sort the data by your new variable, and print the first 3 observations of the resulting data frame.



```
> df_new <- DR_mutate %>% select(-education)
> sortByFinished <- arrange(df_new, finished)
> head(sortByFinished,3)
    id province population  weight gender abortion importance urban    ratio finished
1 2851       BC    3267345 4287.85 Female       No   somewhat urban 762.0008        0
2 2118       QC    5996930 6153.85   Male      Yes   somewhat urban 974.5005        0
3 1815       NL     406455 3430.00 Female       No       very urban 118.5000        0
> tail(sortByFinished,3)
       id province population  weight gender abortion importance urban    ratio finished
2229   72       SK     734250 9443.73 Female      Yes       very rural  77.7500        1
2230  671       QC    5996930 6153.85   Male       No   somewhat urban 974.5005        1
2231 2488       BC    3267345 4287.85 Female       No        not urban 762.0008        1
>
```

- Create finished column where the value is 1 if the subject has bachelors or higher and 0 elsewise.
- Removed the education column afterwards and obtained the head.

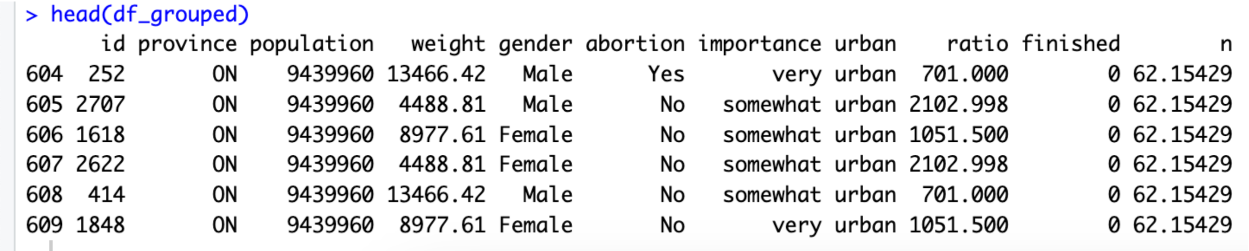
e) A researcher is interested in attitudes on abortion only in the province of Ontario. Create the appropriate subset data frame, and print its dimensions.



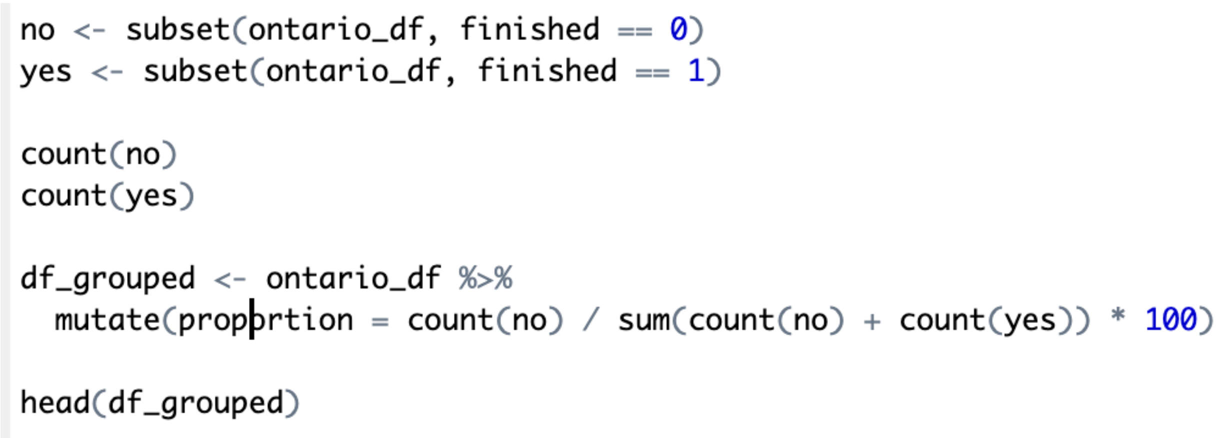
```
> ontario_df <- subset(sortByFinished, province == "ON")
> dim(ontario_df)
[1] 687  10
> head(ontario_df)
      id province population   weight gender abortion importance urban    ratio finished
604  252       ON    9439960 13466.42   Male      Yes       very urban  701.000        0
605 2707       ON    9439960  4488.81   Male       No   somewhat urban 2102.998        0
606 1618       ON    9439960  8977.61 Female       No   somewhat urban 1051.500        0
607 2622       ON    9439960  4488.81 Female       No   somewhat urban 2102.998        0
608  414       ON    9439960 13466.42   Male       No   somewhat urban  701.000        0
609 1848       ON    9439960  8977.61 Female       No       very urban 1051.500        0
```

Creates subset where the province is in Ontario.

f) The **abortion** variable contains responses to the question "Should abortion be banned?" Create a grouped data frame, and obtain the proportion (a percentage) of Ontario survey respondents who were against an abortion ban in 2011.



```
> head(df_grouped)
      id province population   weight gender abortion importance urban    ratio finished        n
604  252       ON    9439960 13466.42   Male      Yes       very urban  701.000        0 62.15429
605 2707       ON    9439960  4488.81   Male       No   somewhat urban 2102.998        0 62.15429
606 1618       ON    9439960  8977.61 Female       No   somewhat urban 1051.500        0 62.15429
607 2622       ON    9439960  4488.81 Female       No   somewhat urban 2102.998        0 62.15429
608  414       ON    9439960 13466.42   Male       No   somewhat urban  701.000        0 62.15429
609 1848       ON    9439960  8977.61 Female       No       very urban 1051.500        0 62.15429
```

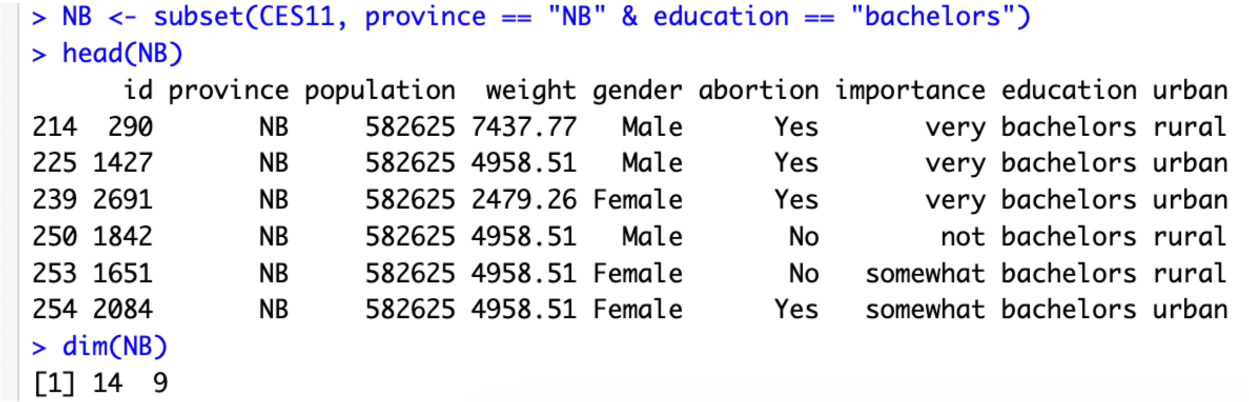


```
no <- subset(ontario_df, finished == 0)
yes <- subset(ontario_df, finished == 1)

count(no)
count(yes)

df_grouped <- ontario_df %>%
  mutate(proportion = count(no) / sum(count(no) + count(yes)) * 100)

head(df_grouped)
```

Obtain the count of subjects in Ontario that said yes to abortion(value 1) and obtain subjects who didn't(value 0). Then calculated the proportion.

2. Return to the original CES11 data set. Obtain the number of rows contained in each of the following subset data frames. (8 pts)

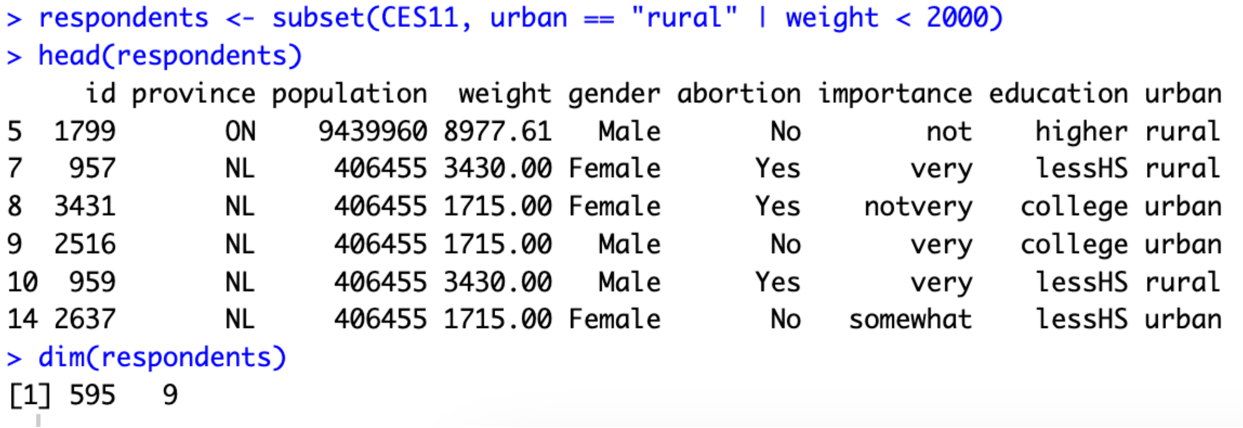
a) Male respondents from the New Brunswick (NB) province who have a bachelors degree.



```
> NB <- subset(CES11, province == "NB" & education == "bachelors")
> head(NB)
      id province population  weight gender abortion importance education urban
214  290       NB     582625 7437.77   Male      Yes       very bachelors rural
225 1427       NB     582625 4958.51   Male      Yes       very bachelors urban
239 2691       NB     582625 2479.26 Female      Yes       very bachelors urban
250 1842       NB     582625 4958.51   Male       No        not bachelors rural
253 1651       NB     582625 4958.51 Female       No   somewhat bachelors rural
254 2084       NB     582625 4958.51 Female      Yes   somewhat bachelors urban
> dim(NB)
[1] 14  9
```

Use subset to see males from NB who has a bachelors degree.

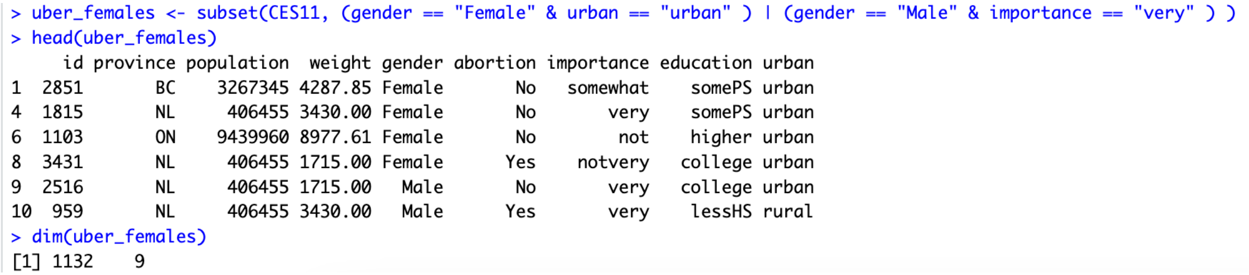
b) Respondents who are either from a rural area, or who have a value of **weight** that is smaller than 2000.



```
> respondents <- subset(CES11, urban == "rural" | weight < 2000)
> head(respondents)
     id province population  weight gender abortion importance education urban
5  1799       ON    9439960 8977.61   Male       No        not    higher rural
7   957       NL     406455 3430.00 Female      Yes       very    lessHS rural
8  3431       NL     406455 1715.00 Female      Yes    notvery   college urban
9  2516       NL     406455 1715.00   Male       No       very   college urban
10  959       NL     406455 3430.00   Male      Yes       very    lessHS rural
14 2637       NL     406455 1715.00 Female       No   somewhat    lessHS urban
> dim(respondents)
[1] 595   9
```

Use subset to people from rural area or who's weight is less then 2000.

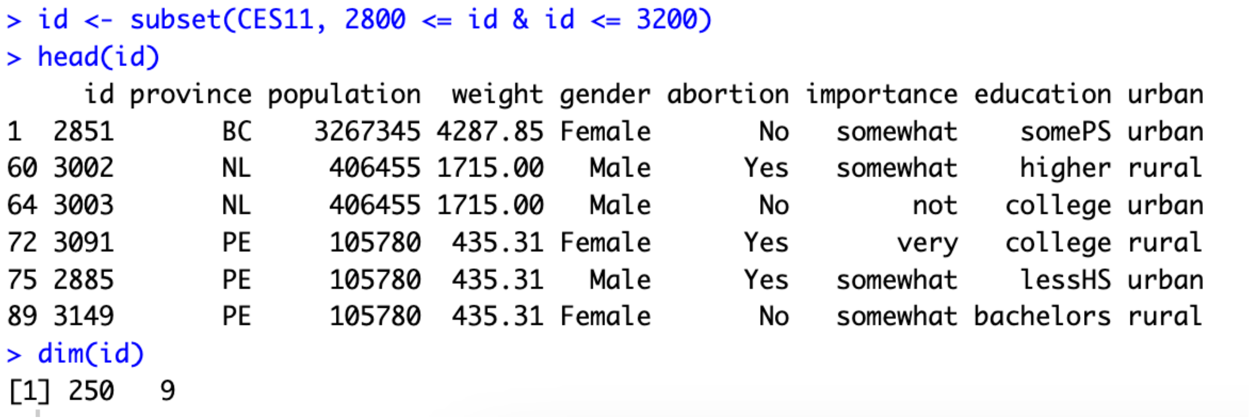
c) Respondents who are urban females, or who are males with the value "very" for the **importance** variable.



```
> uber_females <- subset(CES11, (gender == "Female" & urban == "urban" ) | (gender == "Male" & importance == "very" ) )
> head(uber_females)
     id province population  weight gender abortion importance education urban
1  2851       BC    3267345 4287.85 Female       No   somewhat    somePS urban
4  1815       NL     406455 3430.00 Female       No       very    somePS urban
6  1103       ON    9439960 8977.61 Female       No        not    higher urban
8  3431       NL     406455 1715.00 Female      Yes    notvery   college urban
9  2516       NL     406455 1715.00   Male       No       very   college urban
10  959       NL     406455 3430.00   Male      Yes       very    lessHS rural
> dim(uber_females)
[1] 1132    9
```

Use subset to get urban females or males with value “very”

d) Respondents whose **id** is between 2800 and 3200 (inclusive).

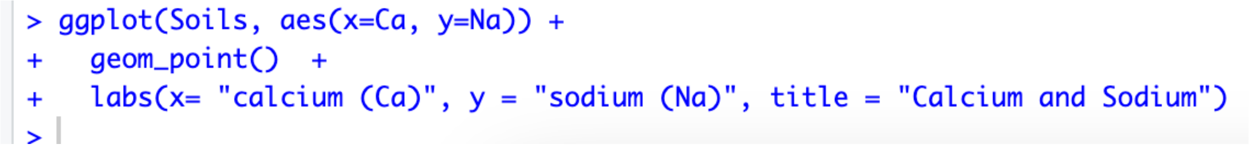


```
> id <- subset(CES11, 2800 <= id & id <= 3200)
> head(id)
     id province population  weight gender abortion importance education urban
1  2851       BC    3267345 4287.85 Female       No   somewhat    somePS urban
60 3002       NL     406455 1715.00   Male      Yes   somewhat    higher rural
64 3003       NL     406455 1715.00   Male       No        not   college urban
72 3091       PE     105780  435.31 Female      Yes       very   college rural
75 2885       PE     105780  435.31   Male      Yes   somewhat    lessHS urban
89 3149       PE     105780  435.31 Female       No   somewhat bachelors rural
> dim(id)
[1] 250   9
```

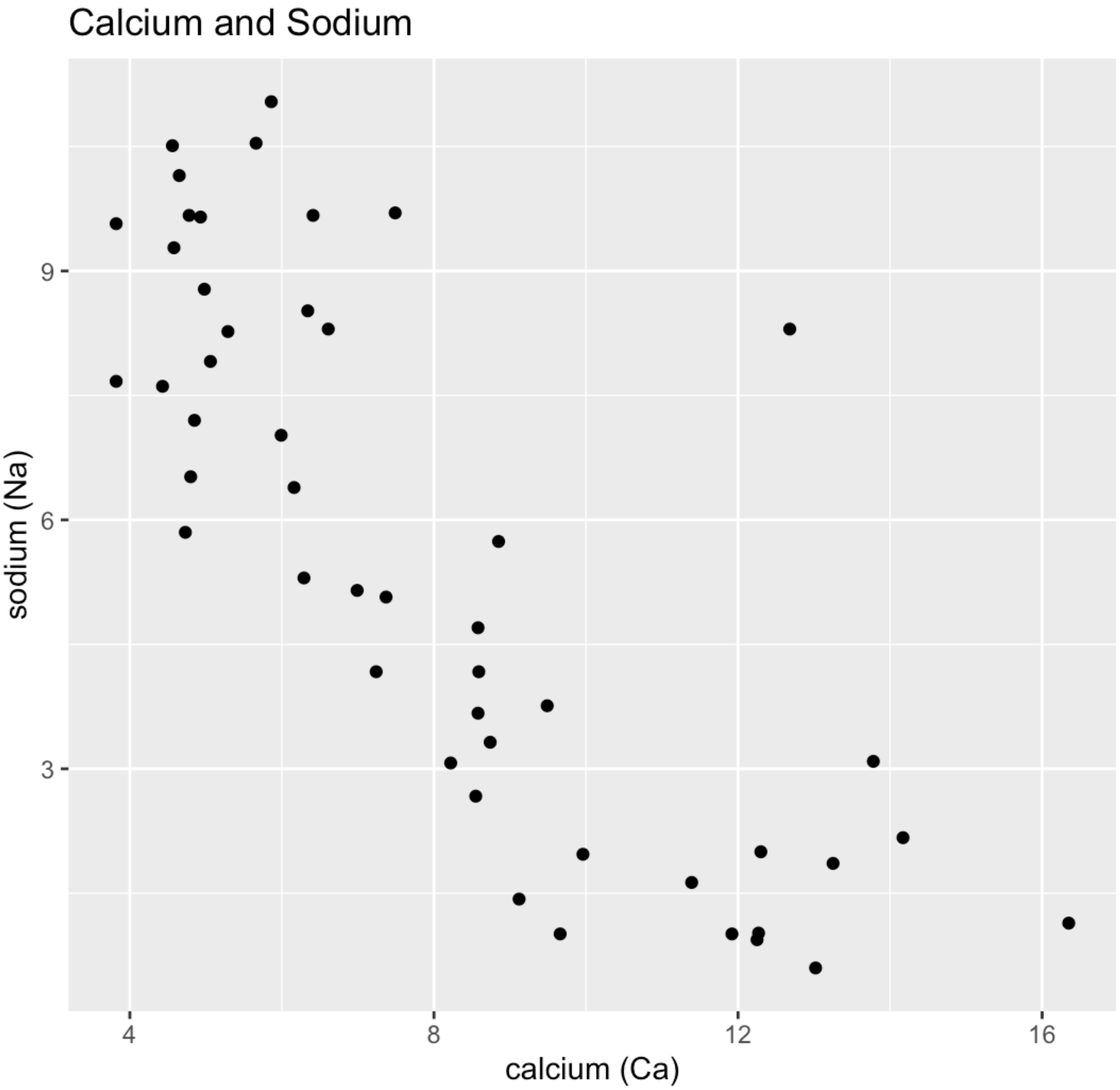
Use subset to see id that is between 2800 and 3200

3. Load the Soils data set, also from the **car** library. (13 pts)

a) Use ggplot(.) to produce a properly labeled scatter plot with calcium (Ca) on the x-axis and sodium (Na) on the y-axis.

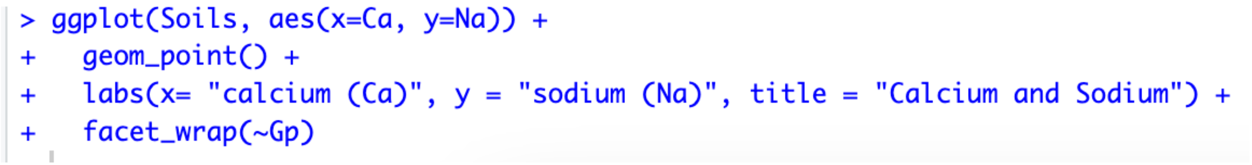


```
> ggplot(Soils, aes(x=Ca, y=Na)) +
+   geom_point()  +
+   labs(x= "calcium (Ca)", y = "sodium (Na)", title = "Calcium and Sodium")
> |
```

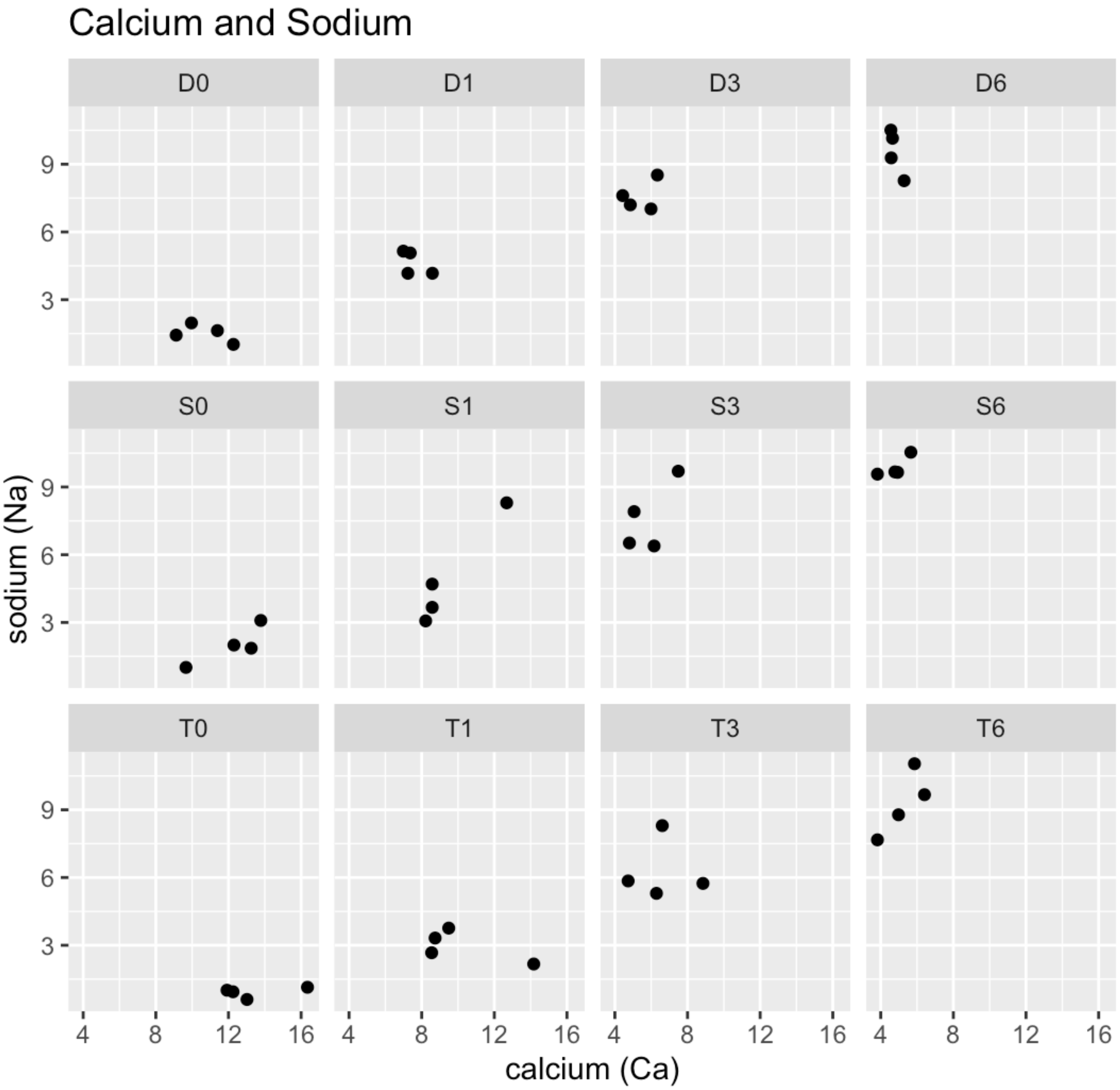


Created scatterplot with Ca and Na.

b) Break up your scatter plot of the overall data into 12 sub-plots using the facet_wrap(.) overlay, using the variable **Gp** to define the 12 facets.



```
> ggplot(Soils, aes(x=Ca, y=Na)) +
+   geom_point() +
+   labs(x= "calcium (Ca)", y = "sodium (Na)", title = "Calcium and Sodium") +
+   facet_wrap(~Gp)
```

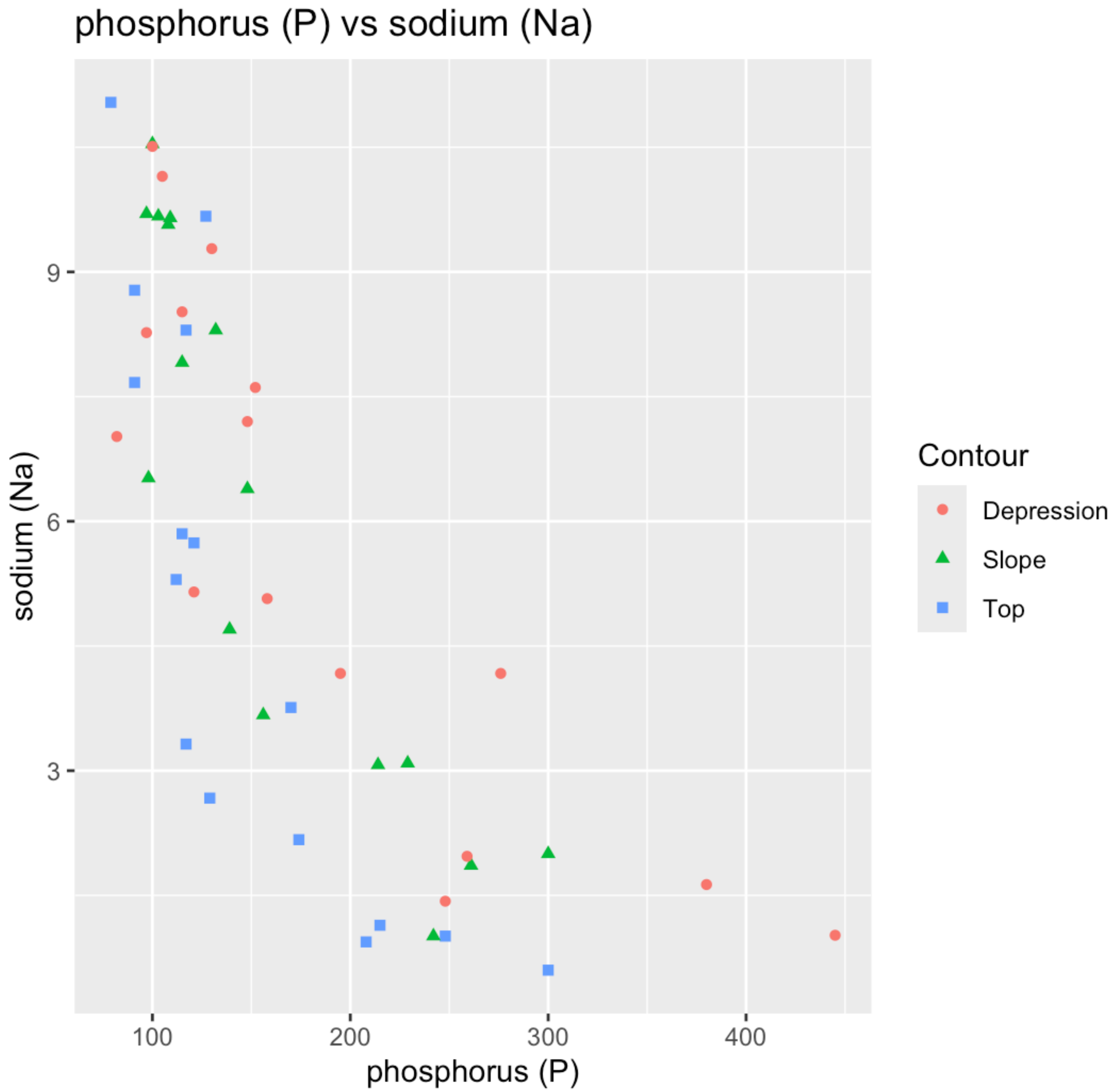


Get 12 subplots of Ca vs Na

c) Judging by the faceted plot in part (b), which soil depth appears to have the lowest sodium content? (Your choices are 0-10 feet, 10-30, 30-60, and 60-90.)

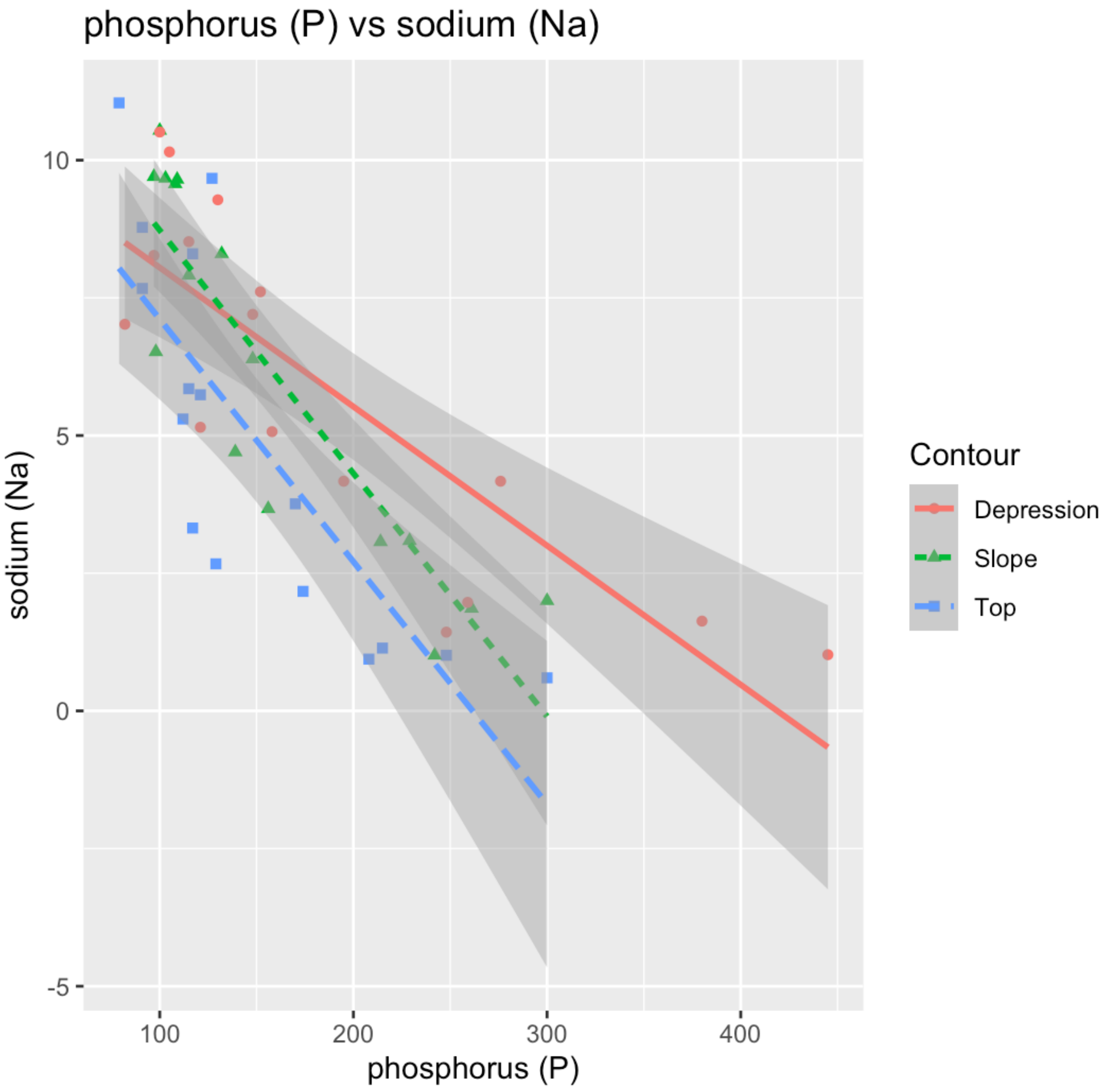
The lowest sodium content seems to be on T0(0-10 feet)

d) Use ggplot(.) to produce a scatter plot of phosphorus (P) vs sodium (Na). Assign the plotted symbols to have different shapes according to the **Contour** variable.

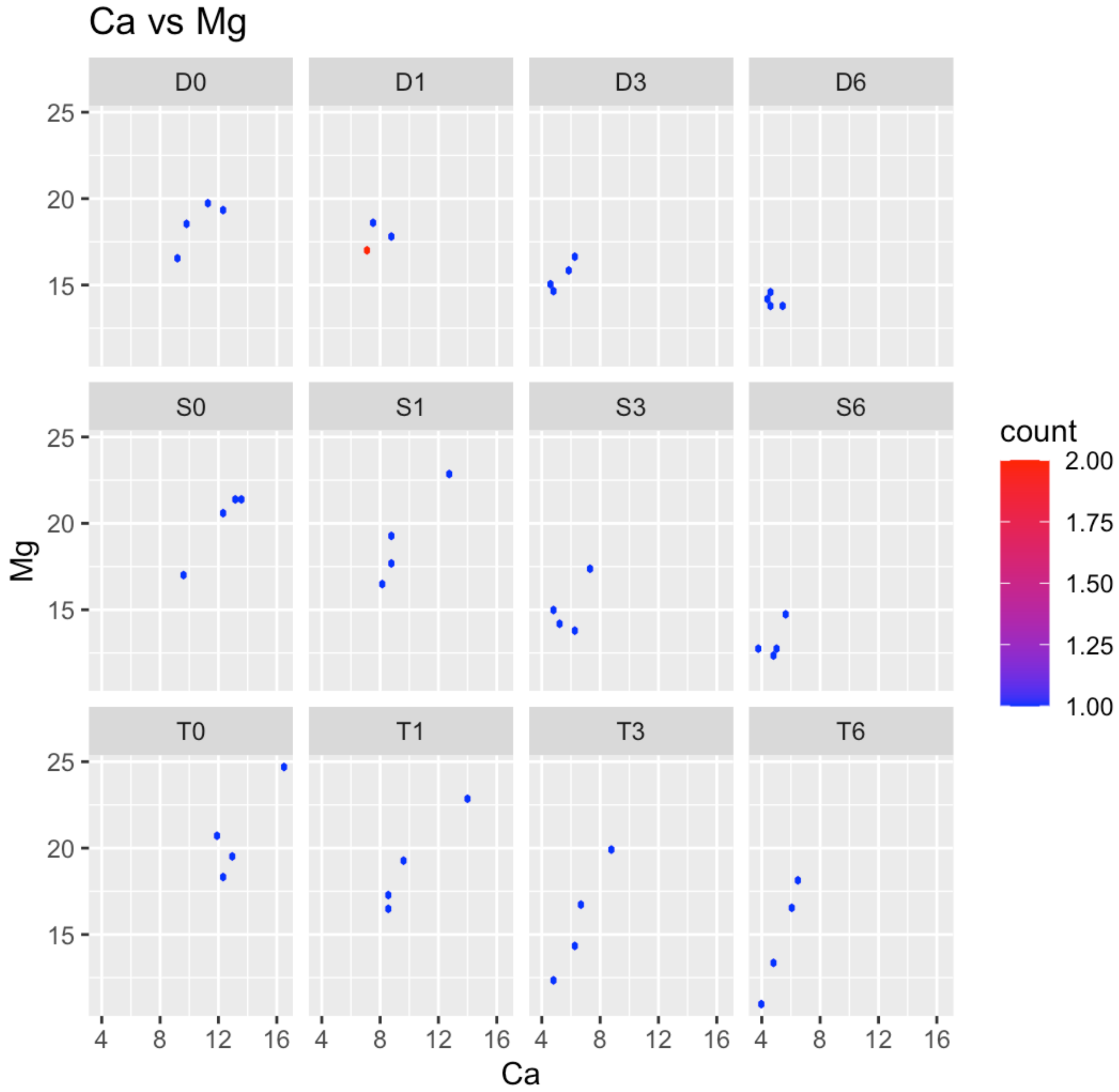


Get scatterplot of P vs Na using Contour to get different shapes

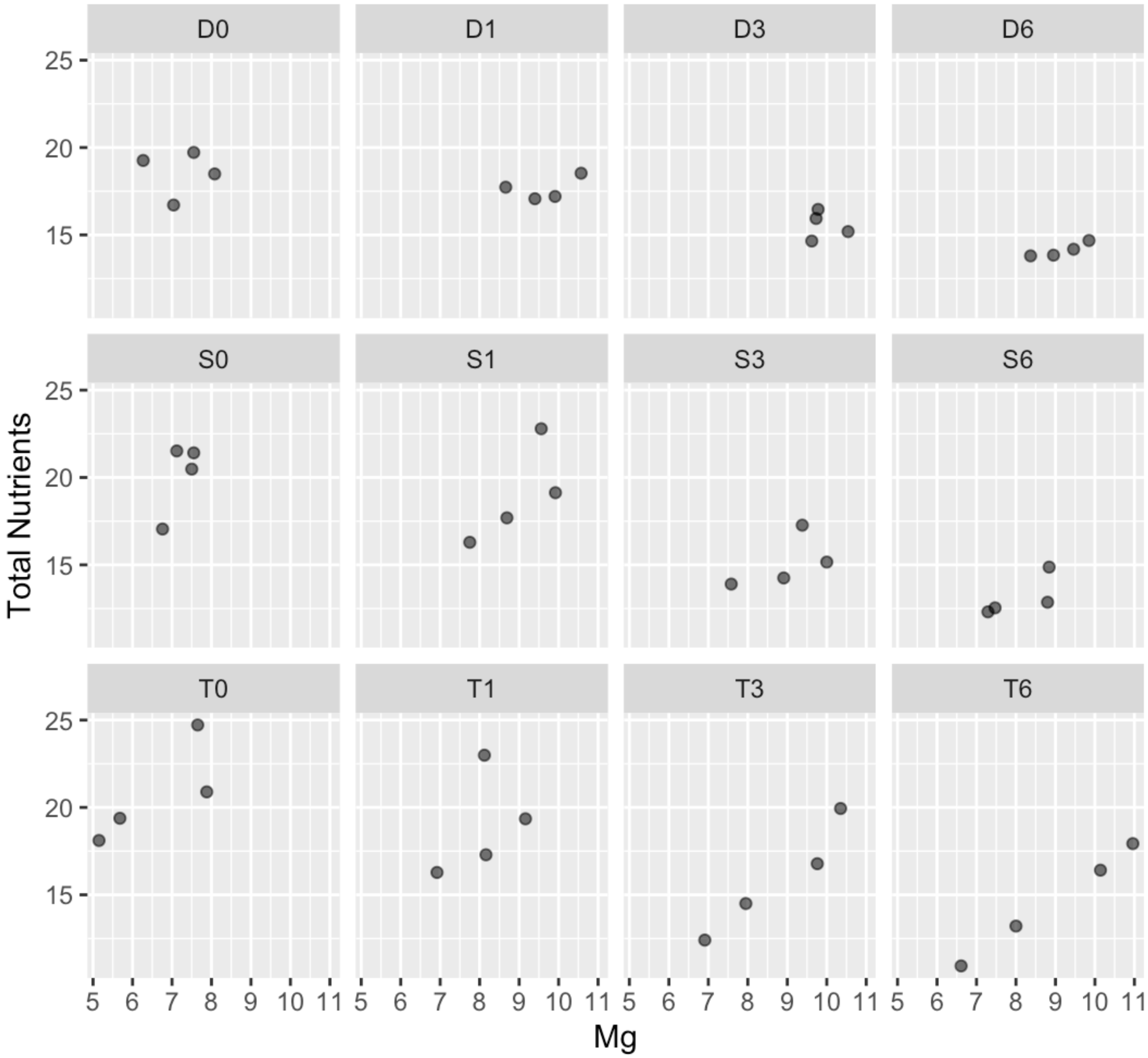
e) Use the geom_smooth(.) overlay to draw three curves through the data points in your part (d) plot, one for each level of **Contour**. Use different line types so that the curves can be distinguished.

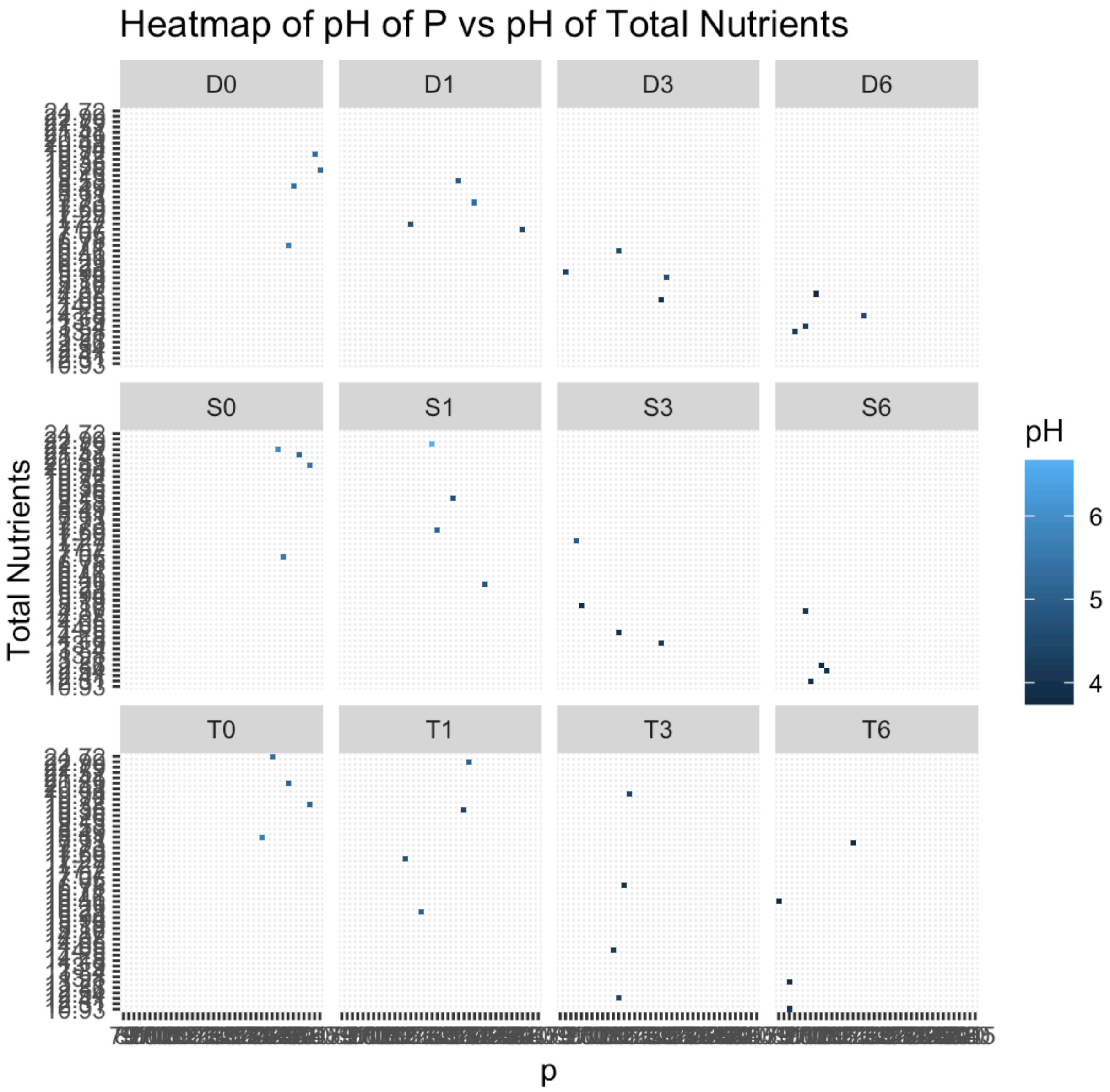

phosphorus (P) vs sodium (Na)
sodium (Na)
10
5
0
-5
100
200
300
400
phosphorus (P)
Contour
Depression
Slope
Top

4. Use the Soils data and options from the ggplot2 cheat sheet to show me a variety of plot I have never seen before. (4 pts)

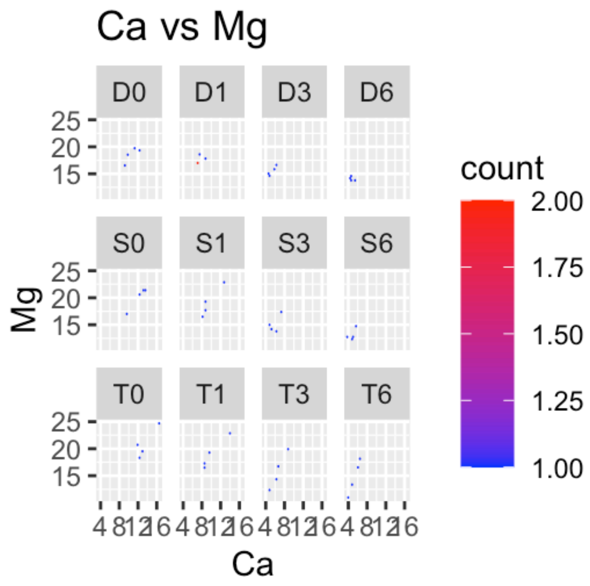


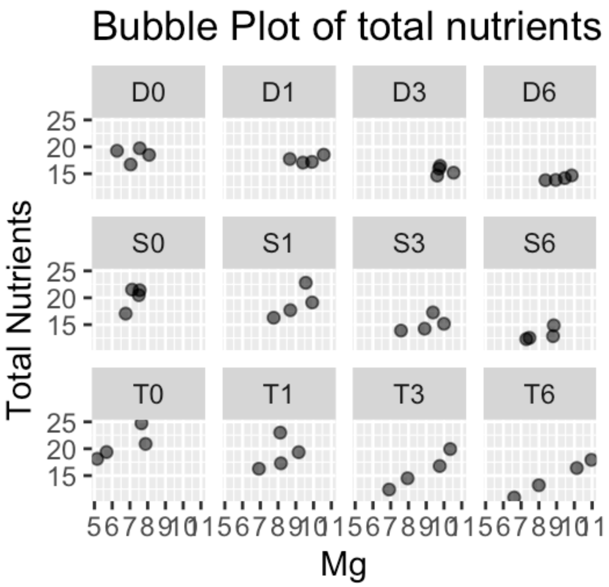
Bubble Plot of total nutrients by Mg

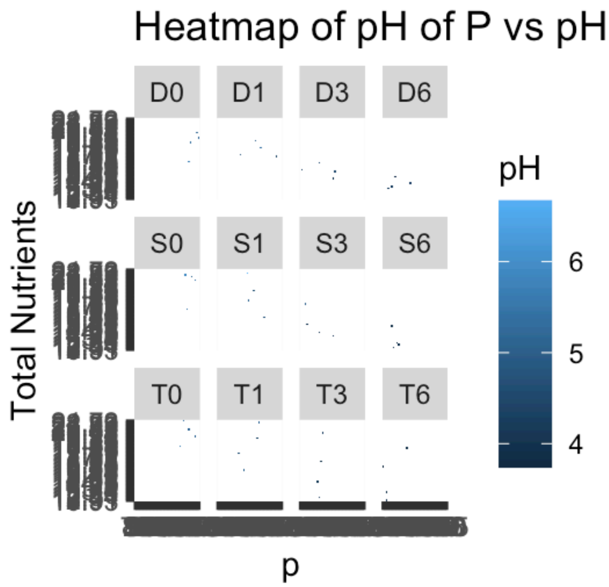




Heatmap of pH of P vs pH of Total Nutrients


Ca vs Mg


Bubble Plot of total nutrients


Heatmap of pH of P vs pH of Total Nutrients

Complete R code:

```
"""
1. A survey was taken in 2011 to assess the Canadian electorate’s opinions on abortion.
A subset of the data can be accessed by loading the car library, then using the command data(CES11). (15 pts)
"""
install.packages("dplyr")
install.packages("car")
install.packages("ggplot2")
install.packages("gganimate")

library(car)
search()

data(CES11)
CES11

#a)        Extract the weight variable into its own vector object. Use this vector to build a histogram.
weight <- c(CES11$weight)
weight

hist(weight)

"""
b)         Sort the data set by both the province variable in reverse alphabetical order and
then by population (ascending). This should be done as a single sort. Print the first 3 observations.
"""
library(dplyr)

sortByProvince <- arrange(CES11, desc(province), population)
head(sortByProvince, 3)



#c)        Add a new column called ratio, calculated as population divided by weight.
DR_mutate <- mutate(CES11,
           ratio = population/weight
)
head(DR_mutate)
```

```
"""
d)        The column education is a categorical variable about the subject’s highest level of education.
Create a simplified column called finished that records whether a subject has received a traditional
4-year college degree using the following definition:
"""

DR_mutate$finished <- ifelse(DR_mutate$education %in% c("bachelors", "higher"), 1, 0)
DR_mutate

df_new <- DR_mutate %>% select(-education)
sortByFinished <- arrange(df_new, finished)
head(sortByFinished,3)
tail(sortByFinished,3)



"""
e)        A researcher is interested in attitudes on abortion only in the province of Ontario.
Create the appropriate subset data frame, and print its dimensions.
"""
ontario_df <- subset(sortByFinished, province == "ON")
dim(ontario_df)
head(ontario_df)



"""
f)        The abortion variable contains responses to the question “Should abortion be banned?”
Create a grouped data frame, and obtain the proportion (a percentage) of Ontario survey
respondents who were against an abortion ban in 2011.
"""
no <- subset(ontario_df, finished == 0)
yes <- subset(ontario_df, finished == 1)

count(no)
count(yes)

df_grouped <- ontario_df %>%
  mutate(proportion = count(no) / sum(count(no) + count(yes)) * 100)

head(df_grouped)



#2. Return to the original CES11 data set. Obtain the number of rows contained in each of the following subset data frames. (8 pts)

#a)          Male respondents from the New Brunswick (NB) province who have a bachelors degree.

NB <- subset(CES11, province == "NB" & education == "bachelors")
head(NB)
dim(NB)

#b)          Respondents who are either from a rural area, or who have a value of weight that is smaller than 2000.

respondents <- subset(CES11, urban == "rural" | weight < 2000)
head(respondents)
dim(respondents)


#c)          Respondents who are urban females, or who are males with the value "very" for the importance variable.

uber_females <- subset(CES11, (gender == "Female" & urban == "urban" ) | (gender == "Male" & importance == "very" ) )
head(uber_females)
dim(uber_females)


#d)          Respondents whose id is between 2800 and 3200 (inclusive).

id <- subset(CES11, 2800 <= id & id <= 3200)
head(id)
dim(id)



#3. Load the Soils data set, also from the car library. (13 pts)
library(ggplot2)
```

```
search()

data(Soils)
head(Soils)

#a)	Use ggplot(.) to produce a properly labeled scatter plot with calcium (Ca) on the x-axis and sodium (Na) on the y-axis.

ggplot(Soils, aes(x=Ca, y=Na)) +
  geom_point()  +
  labs(x= "calcium (Ca)", y = "sodium (Na)", title = "Calcium and Sodium")



#b)	Break up your scatter plot of the overall data into 12 sub-plots using the facet_wrap(.) overlay, using the variable Gp to define the 12 facets.

ggplot(Soils, aes(x=Ca, y=Na)) +
  geom_point() +
  labs(x= "calcium (Ca)", y = "sodium (Na)", title = "Calcium and Sodium") +
  facet_wrap(~Gp)



#d)	Use ggplot(.) to produce a scatter plot of phosphorus (P) vs sodium (Na). Assign the plotted symbols to have different shapes according to the Contour variable.

ggplot(Soils, aes(x=P, y=Na, shape = Contour, color = Contour)) +
  geom_point() +
  labs(x= "phosphorus (P)", y = "sodium (Na)", title = "phosphorus (P) vs sodium (Na)")



"""
e)	Use the geom_smooth(.) overlay to draw three curves through the data points in your part (d) plot, one for each level of Contour.
Use different line types so that the curves can be distinguished.
"""
ggplot(Soils, aes(x=P, y=Na, shape = Contour, color = Contour)) +
  geom_point() +
  labs(x= "phosphorus (P)", y = "sodium (Na)", title = "phosphorus (P) vs sodium (Na)") +
  geom_smooth(method = "lm", aes(linetype = Contour))
```

```
"""
4. Use the Soils data and options from the ggplot2 cheat sheet to show me a variety of plot I have never seen before. (4 pts)
"""


"""
Chemicals can form compounds when two or more elements combine chemically. These compounds are the result of chemcial reactions
and the elements are held together through a chemcial bond(a force of attraction between atoms or ions through the share of valence electrons)

I will now make new columns of new compounds by forming compounds from the aviable elements

"""

Soils$Calcium_Phosphate <- (Soils$Ca * 2/3) * (Soils$P * 3/2)
Soils$Magnesium_Potassium_Salt <- Soils$Mg * 2 * Soils$K
Soils$Sodium_Chloride <- Soils$Ca * Soils$P
head(Soils)

#Upon doing this I realize that this might not be mathematically correct welp, it's already done ¯\_(ツ)_/¯ maybe it'll be useful later.
library(gridExtra)
library(gganimate)



Soils$total_nutrients <- Soils$Ca + Soils$Mg + Soils$K
Soils

plot1 <- ggplot(Soils, aes(x = Ca, y = total_nutrients)) +
  geom_hex(bins = 30) +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Ca vs Mg", x = "Ca", y = "Mg")  +
  facet_wrap(~Gp)



plot1
```

```
plot2 <- ggplot(Soils, aes(x = Mg, y = total_nutrients), size = pH, color = Contour) +
  geom_point(alpha = 0.6) +
  labs(title = "Bubble Plot of total nutrients by Mg", x = "Mg", y = "Total Nutrients") +
  facet_wrap(~Gp)

plot2
```

```
plot3 <- ggplot(Soils, aes(x = factor(P), y = factor(total_nutrients), fill = pH)) +
  geom_tile() +
  labs(title = "Heatmap of pH of P vs pH of Total Nutrients", x = "p", y = "Total Nutrients") +
  facet_wrap(~Gp)

plot3
```

```
grid.arrange(plot1, plot2, plot3, ncol = 2)
```