Richard Gresham

CPSC 375-01

09/14/22

Homework #3

Prepare your answers as a **single PDF file**.

Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group should submit to Canvas.

Due: check on Canvas.

The main purpose of this assignment is to test your understanding of how to choose the appropriate visualization. Use the in-built dataset, esoph, for this problem ("Data from a case-control study of (o)esophageal cancer in Ille-et-Vilaine, France."). All plots should use ggplot. For each question, give the code and include the plot.

a. Does the dataset contain any NAs? If so, which variables have NAs? What is the type of variable tobgp? [Hint: use str() and summary()]

Input: Which(is.na(esoph)) #gets the row at which data is NA

Output: integer(0)

So, the dataset does not have any NAs

What is the variable tobgp? Variable tobgp is an Ordered factor with 4 levels.

Input: str(esoph$tobgp)

Output: Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 3 4 1 2 3 4 1 2 ...

Input: summary(esoph$tobgp)

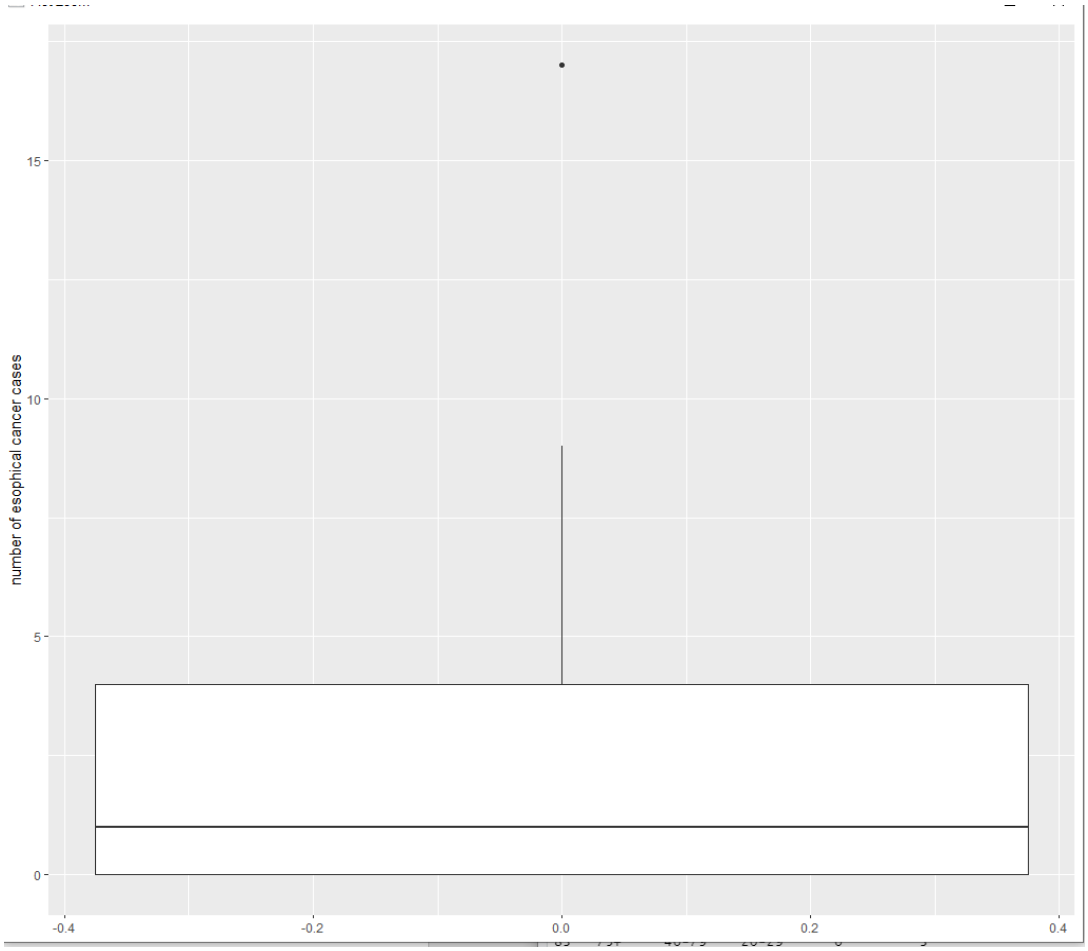
Output:

0-9g/day	10-19	20-29	30+
24	24	20	20

b. Visualize variable ncases. Give a more descriptive name to the axis (Hint: help(esoph) to see a description of the dataset). Does this variable contain outliers? Do you think these values are really outliers or legitimate values?

Input: ggplot(data = esoph) + geom_boxplot(mapping = aes(,y = ncases)) + ylab("number of esophical cancer cases")

Output:



I chose a boxplot visualization for ncases or number of cases as it is a Numerical variable and upon reading the data, I saw a possible outlier in the group. This age group is on row 67 and the reason why I believe this data set is an outlier is because while they are in an older age category of 65-74 years old, their tobacco consumption is little to none, and alcohol consumption is only at a low moderate level compared to some of the other data set to warrant such a high number of esophageal cancer cases.

Input: esoph[67,]

Ouput:

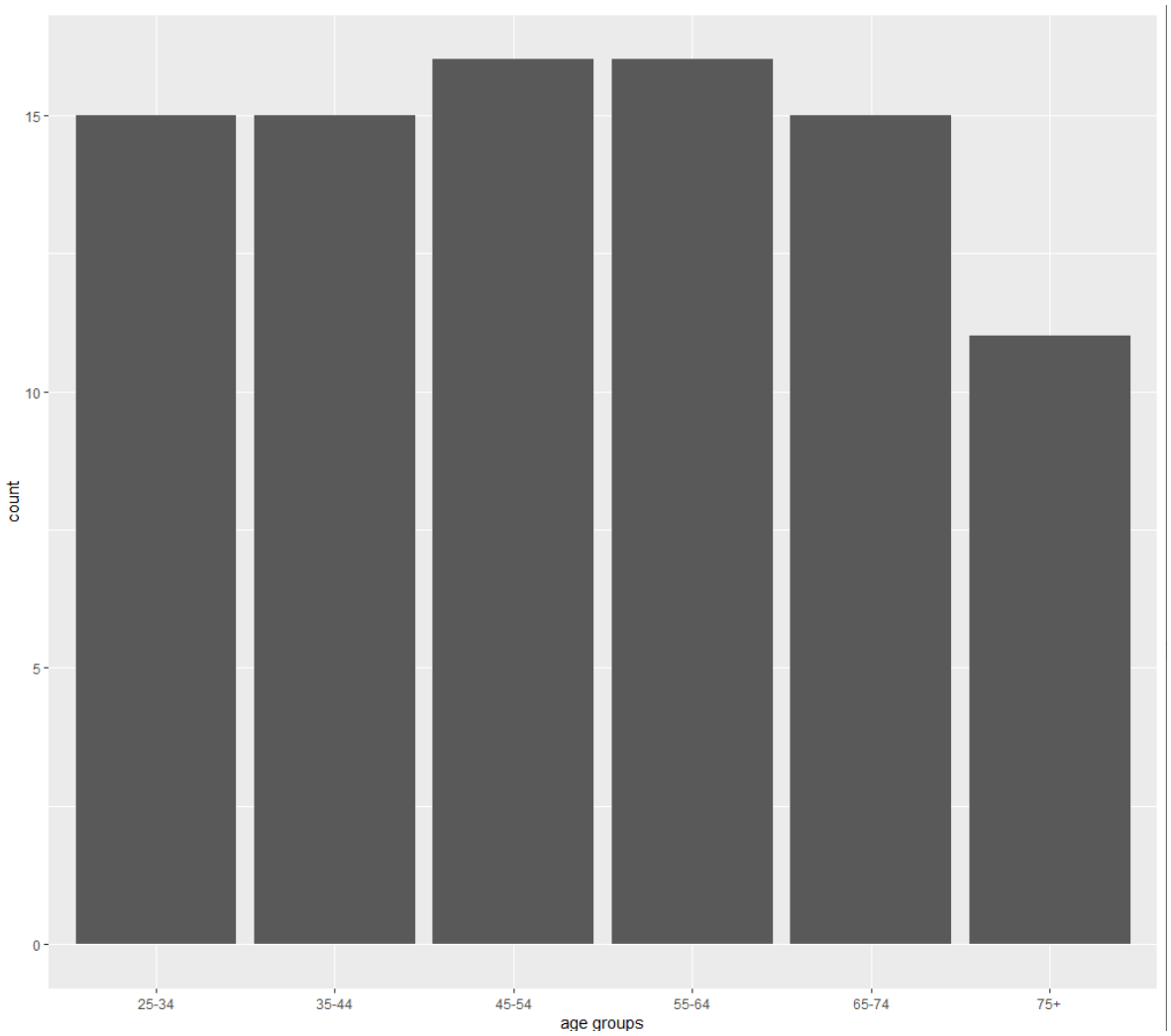
agegp	alcgp	tobgp	ncases	ncontrols
67	65-74	40-79	0-9g/day	17	34

c. Visualize variable agegp. Give a more descriptive name to the axis. (Hint: use geom_bar() for discrete variables.)

Since age-group is groups of ages it can be considered discrete thus I am going to use geom_bar().

Input: ggplot(data = esoph) + geom_bar(mapping = aes(x = agegp,)) + xlab("age groups")

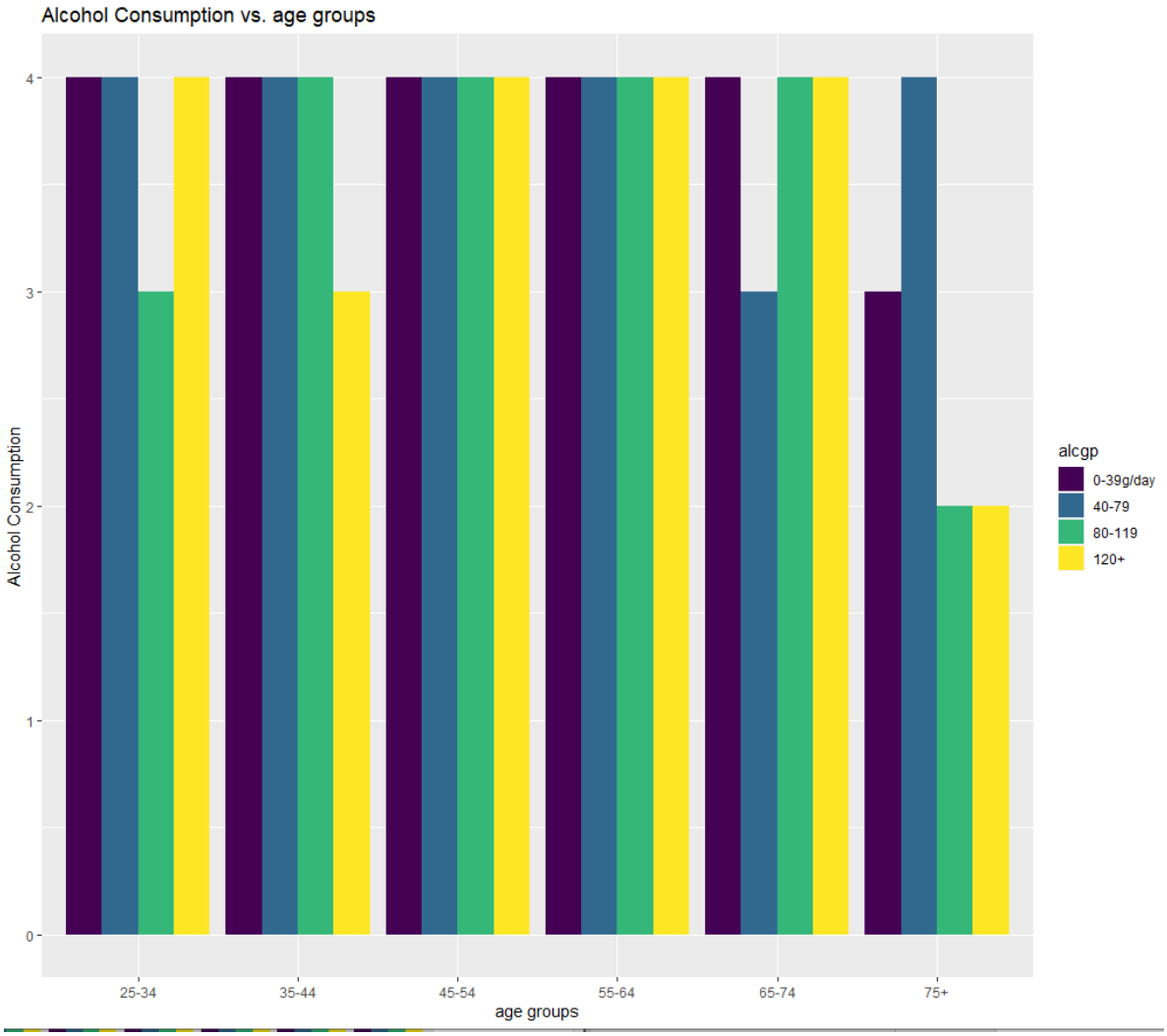
Output:



d. Visualize variables agegp and alcgp. 2 categorical variables so we use a bar graph.

Input: ggplot(data = esoph) + geom_bar(mapping = aes(x = agegp, fill = alcgp), position = "dodge") + xlab("age groups") + ylab("Alcohol Consumption") + ggtitle("Alcohol Consumption vs. age groups")

Output:

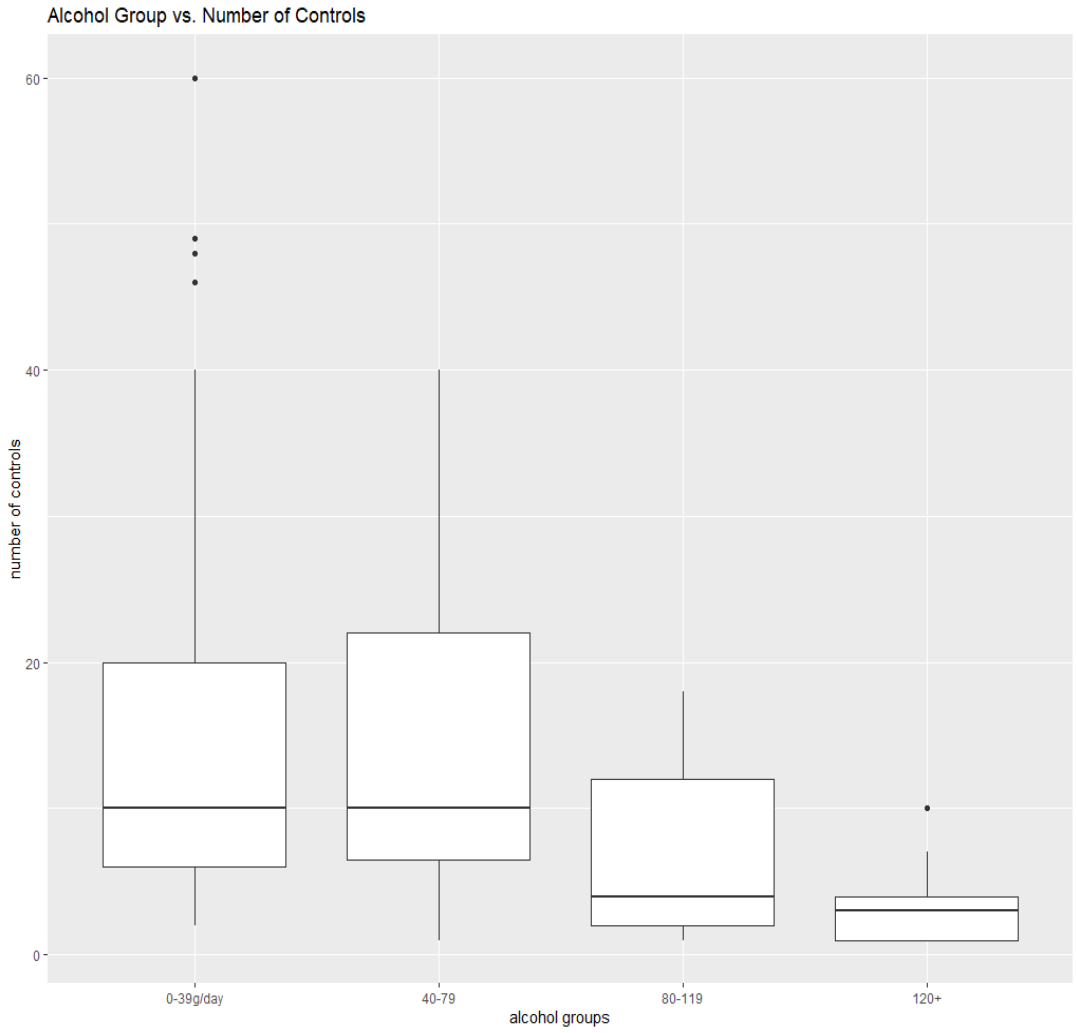


e. Visualize variables alcgp and ncontrols.

1 categorical variable + 1 numerical variable use boxplot.

Input: ggplot(data = esoph) + geom_boxplot(mapping = aes(x = alcgp, y = ncontrols)) + xlab("alcohol groups") + ylab("number of controls") + ggtitle("Alcohol Group vs. Number of Controls")

Output:

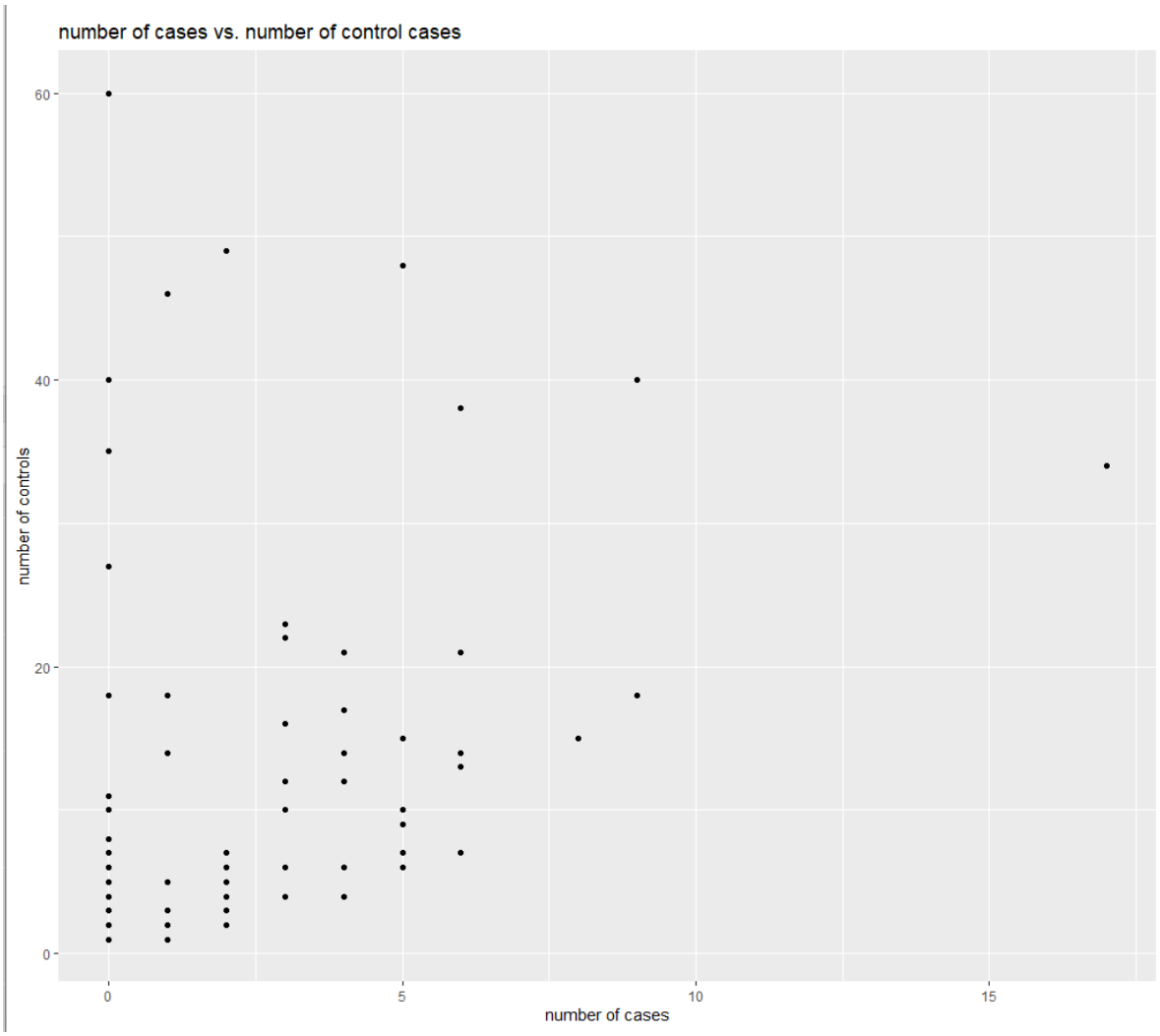


f. Visualize variables ncases and ncontrols.

2 numerical variables so I use scatterplot.

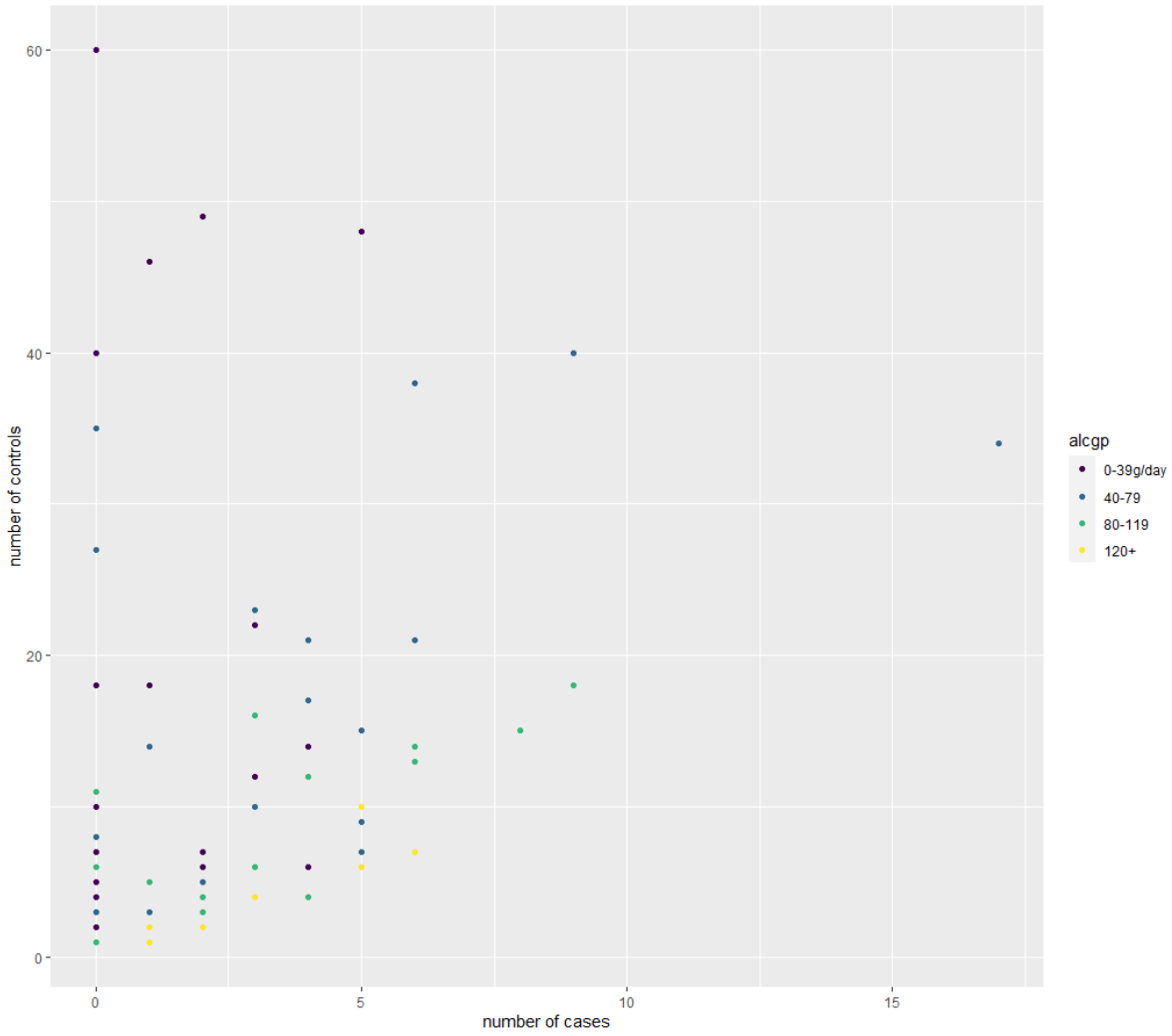
Input: ggplot(data = esoph) + geom_point(mapping = aes(x = ncases, y = ncontrols)) + xlab("number of cases") + ylab("number of controls") + ggtitle("number of cases vs. number of control cases")

Output:



g. Visualize variables ncases, ncontrols, and alcgp.

Input: ggplot(data = esoph) + geom_point(mapping = aes(x = ncases, y = ncontrols, color = alcgp)) + xlab("number of cases") + ylab("number of controls")


60
40
20
0
number of controls
0
5
10
15
number of cases
alcgp
0-39g/day
40-79
80-119
120+