CPSC 375 Homework 3

Kenn Son, Hamid Suda, Vivian Truong

2/16/2022

```
library(ggplot2)
```

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, if created.

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()]

   There is no NAs in the dataset.

```
esoph[is.na(esoph),]
```

```
## [1] agegp     alcgp     tobgp     ncases    ncontrols
## <0 rows> (or 0-length row.names)
```

   topgp variable is a factor.

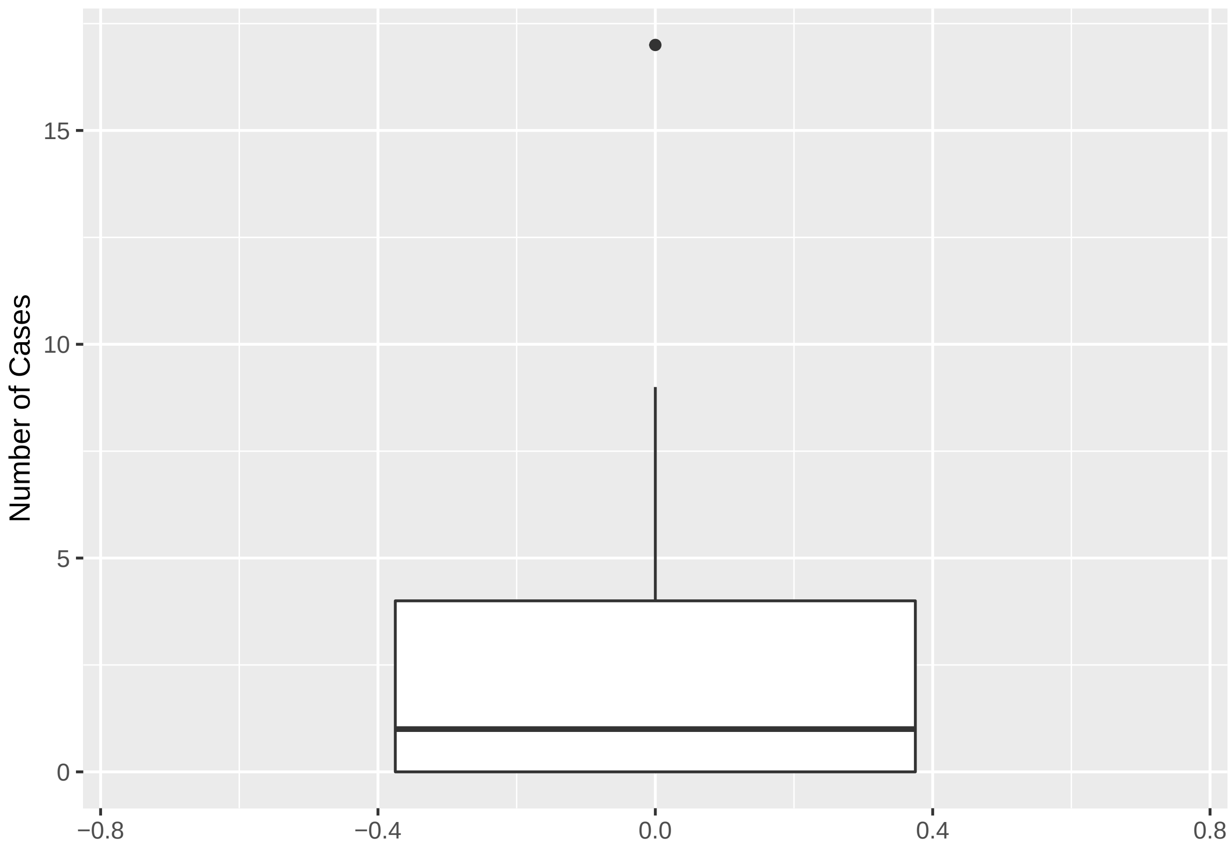
```
str(esoph$tobgp)
```

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

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?

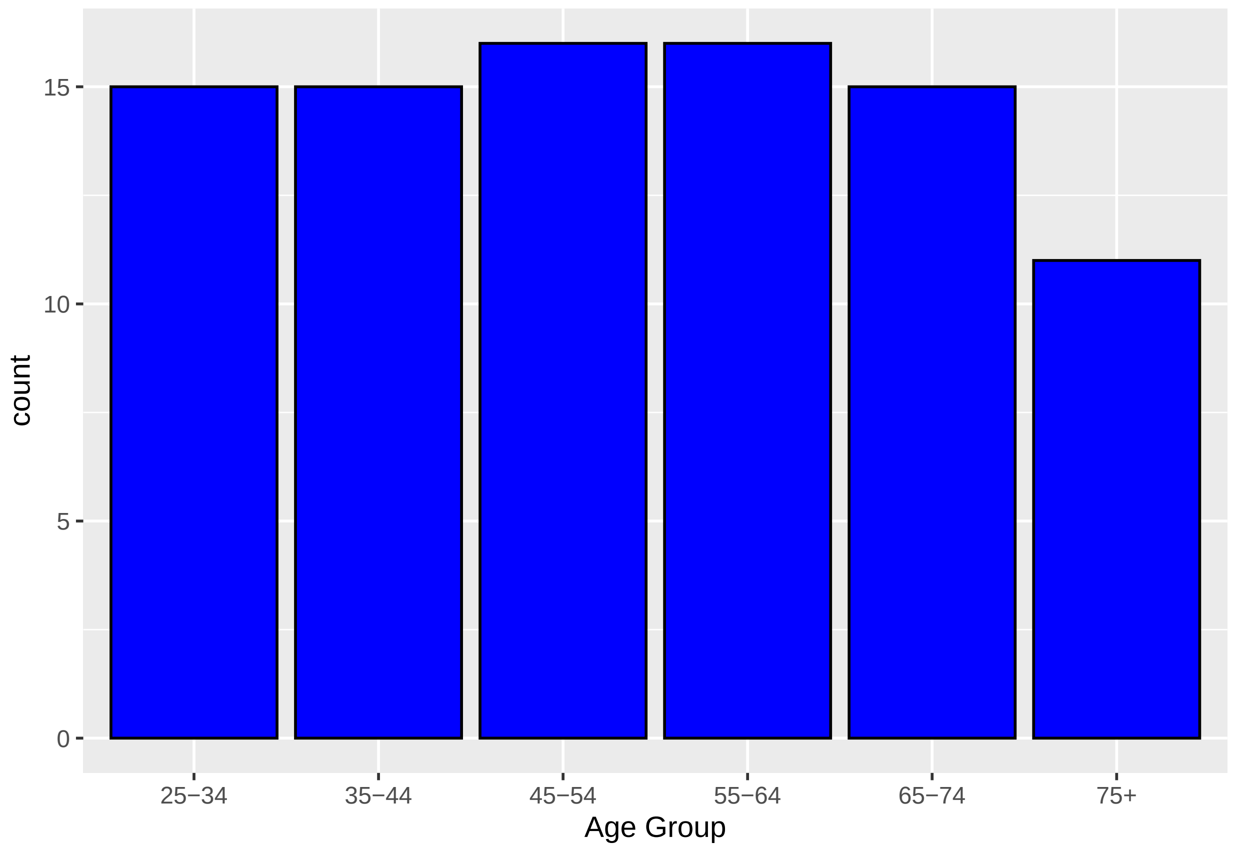
   Yes there is an outlier if you look at the corresponding boxplot

```
ggplot(data=esoph) + geom_boxplot(mapping = aes(y = ncases)) +
  labs(y = "Number of Cases") + xlim(c(-.75,.75))
```



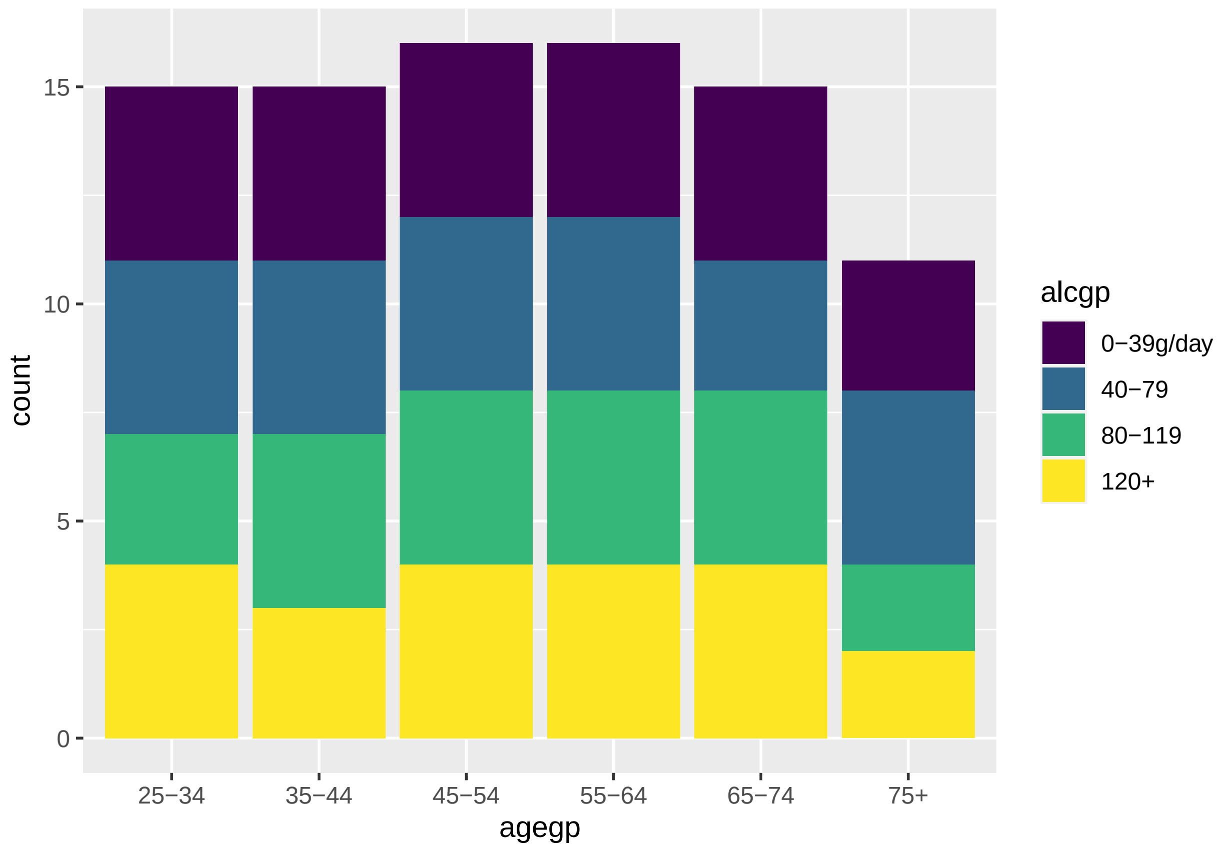
c) Visualize variable agegp. Give a more descriptive name to the axis. (Hint: use geom_bar() for discrete variables.)

```
ggplot(data=esoph) +
  geom_bar(mapping = aes(x = agegp), fill = "blue", color = "black") +
  labs(x = "Age Group")
```



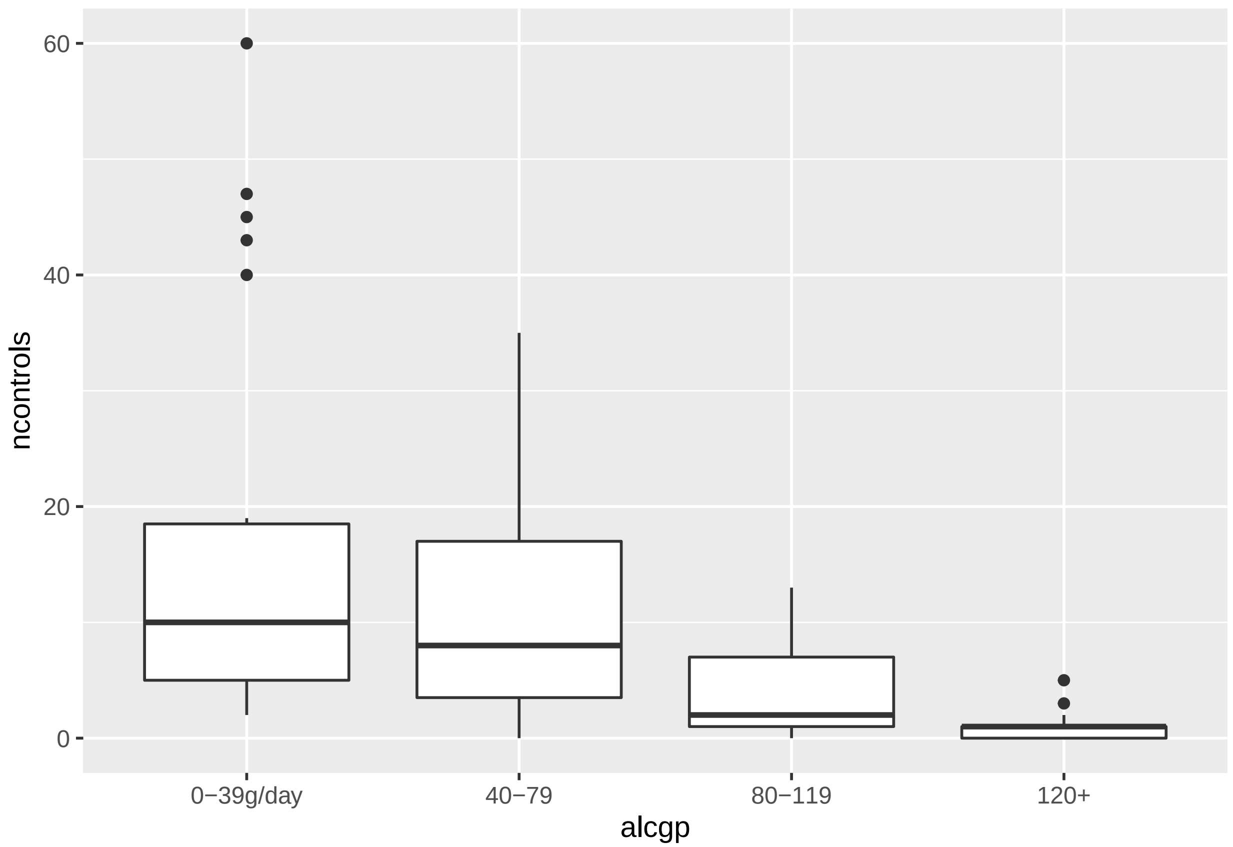
d) Visualize variables *agegp* and *alcgp*.

```
ggplot(data=esoph) + geom_bar(mapping = aes(x=agegp, fill=alcgp))
```



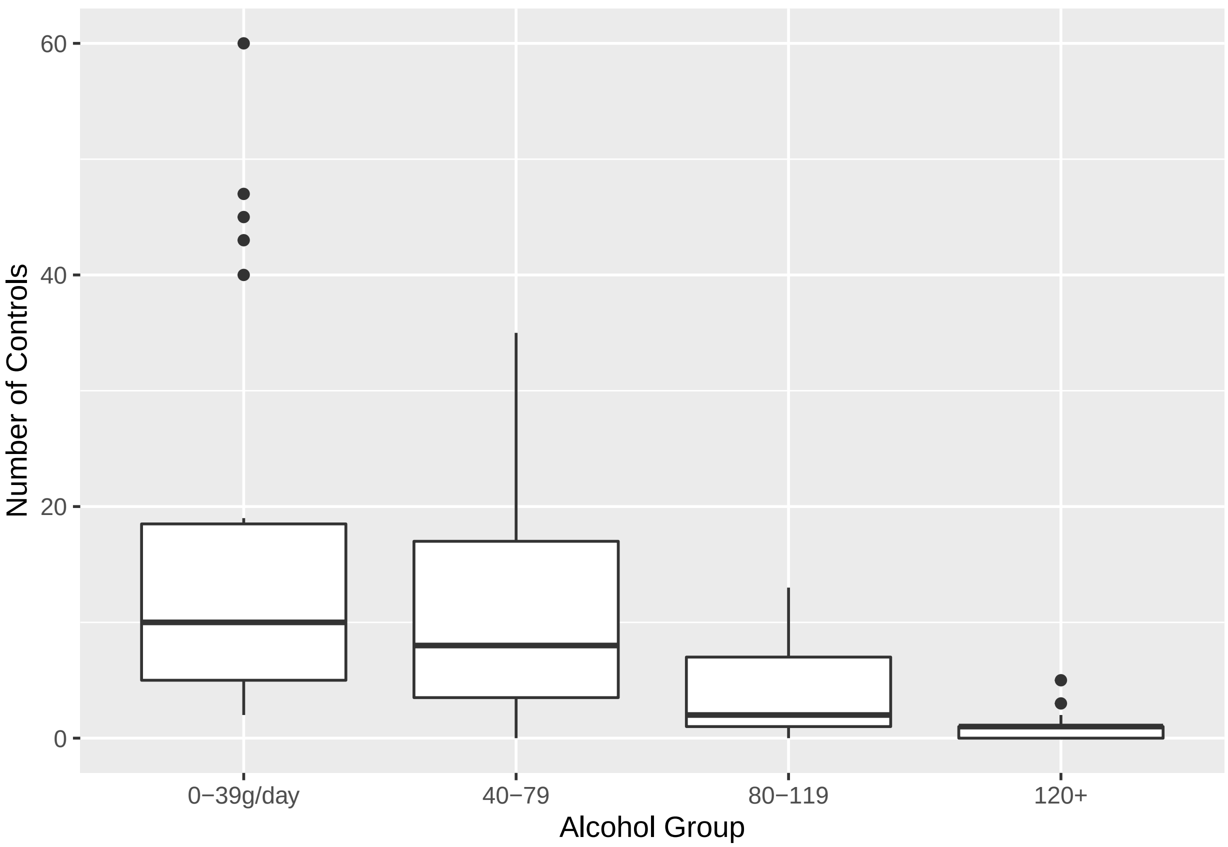
e) Visualize variables *alcgp* and *ncontrols*.

```
ggplot(data=esoph) + geom_boxplot(mapping = aes(x=alcgp, y=ncontrols))
```



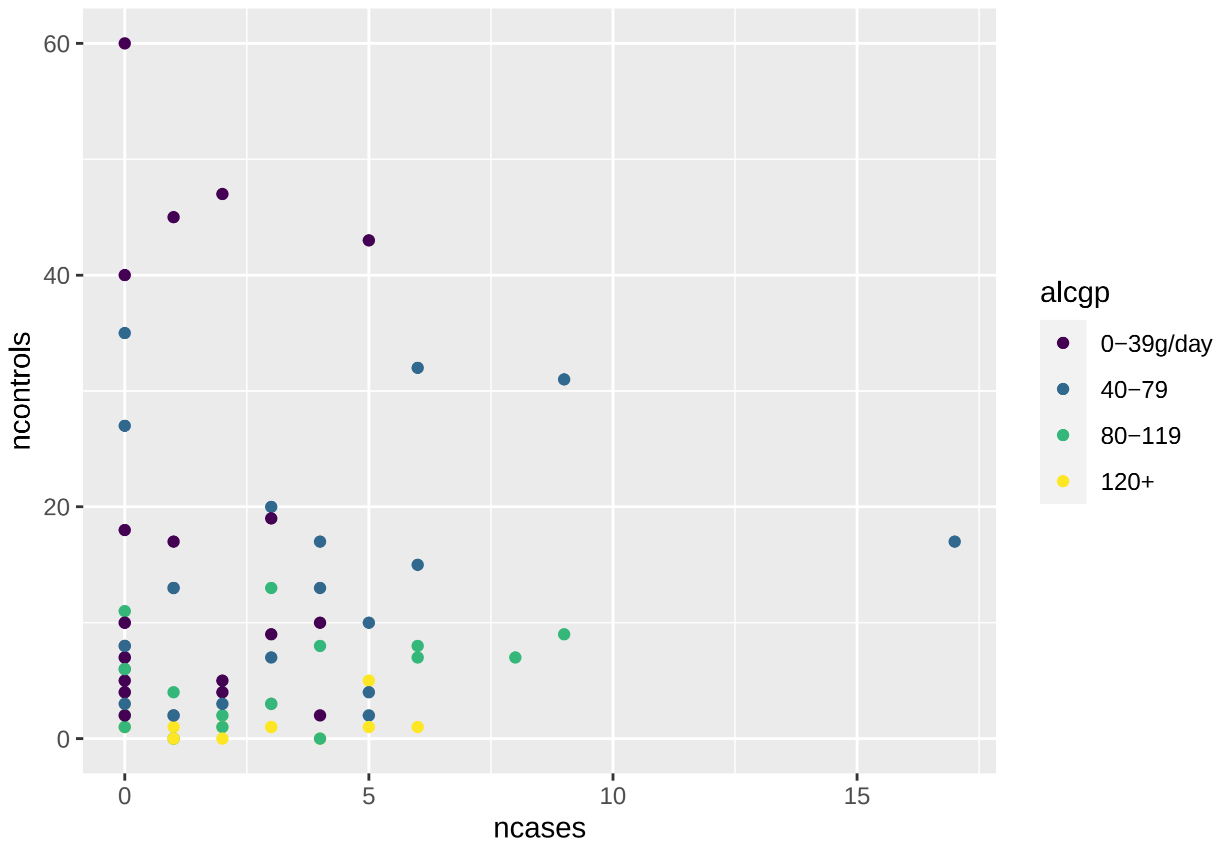
f) Visualize variables *ncases* and *ncontrols*.

```
ggplot(data=esoph) + geom_boxplot(mapping = aes(x=alcgp, y=ncontrols)) +
  labs(x = "Alcohol Group", y = "Number of Controls")
```



g) Visualize variables *ncases*, *ncontrols*, and *alcgp*.

```
ggplot(data=esoph) + geom_point(mapping = aes(x=ncases,y=ncontrols, color = alcgp))
```


60
40
20
0
ncontrols
0
5
10
15
ncases
alcgp
0−39g/day
40−79
80−119
120+