Mariia Nikitash

Homework 2

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. You may work with students in both sections (375-01, -02). 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, if created.

a. Does the dataset contain any NAs? If so, which variables have NAs?

Esoph data set does not contain any NAs

```
> is.na(esoph)
      agegp alcgp tobgp ncases ncontrols
 [1,] FALSE FALSE FALSE  FALSE     FALSE
 [2,] FALSE FALSE FALSE  FALSE     FALSE
 [3,] FALSE FALSE FALSE  FALSE     FALSE
 [4,] FALSE FALSE FALSE  FALSE     FALSE
 . . . . . . . . . . . . . . . . . . . .
[88,] FALSE FALSE FALSE  FALSE     FALSE
```

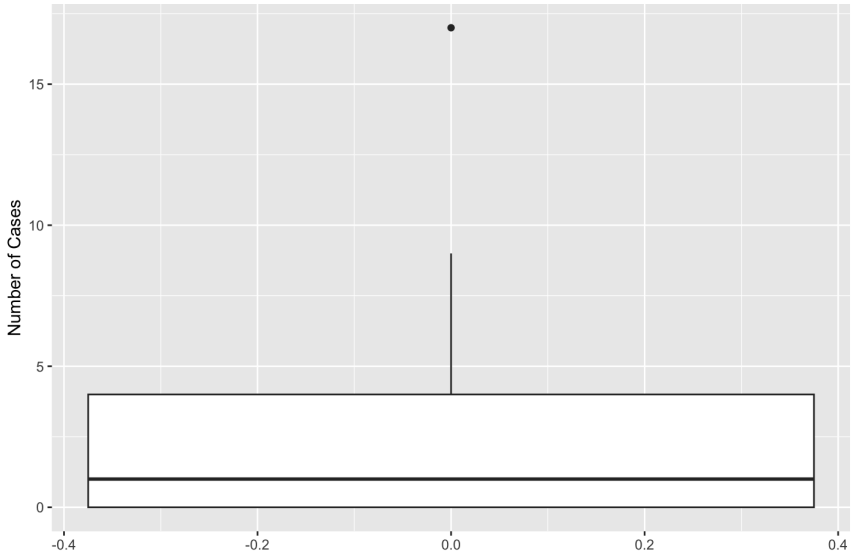
What is the type of variable `tobgp`? [Hint: use `str()` and `summary()`]

The type of variable `tobgp` is `Ord.factor`

```
> str(esoph)
'data.frame':   88 obs. of  5 variables:
 $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 1 1 1
1 1 1 1 1 1 1 ...
 $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 1
1 1 1 2 2 2 2 3 3 ...
 $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2
3 4 1 2 3 4 1 2 ...
 $ ncases   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ncontrols: num  40 10 6 5 27 7 4 7 2 1 …
```

b. Visualize variable `ncases`. Give a more descriptive name to the axis (Hint: `help(esoph)` to see a description of the dataset).

```
ggplot(data=esoph) + geom_boxplot(mapping = aes(y=ncases)) +
labs(y = "Number of Cases")
```



Does this variable contain outliers?

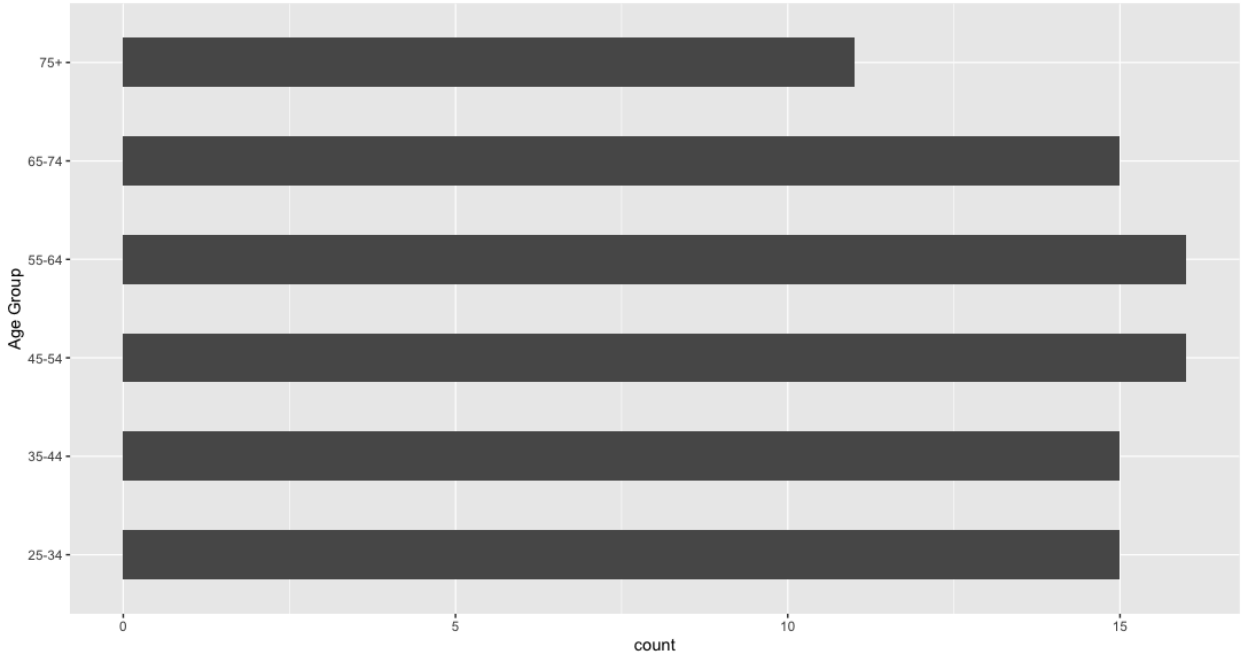
```
> IQR(esoph$ncases)
[1] 4
> quantile(esoph$ncases)
  0%  25%  50%  75% 100%
   0    0    1    4   17
> iqr <-IQR(esoph$ncases)
> iqr
[1] 4
> 0-iqr*1.5
[1] -6
> 4+iqr*1.5
[1] 10
```

Are these outliers errors or legitimate values?

There are no outliers, since all the numbers are in the range of -6 and 10. So these values are legitimate.

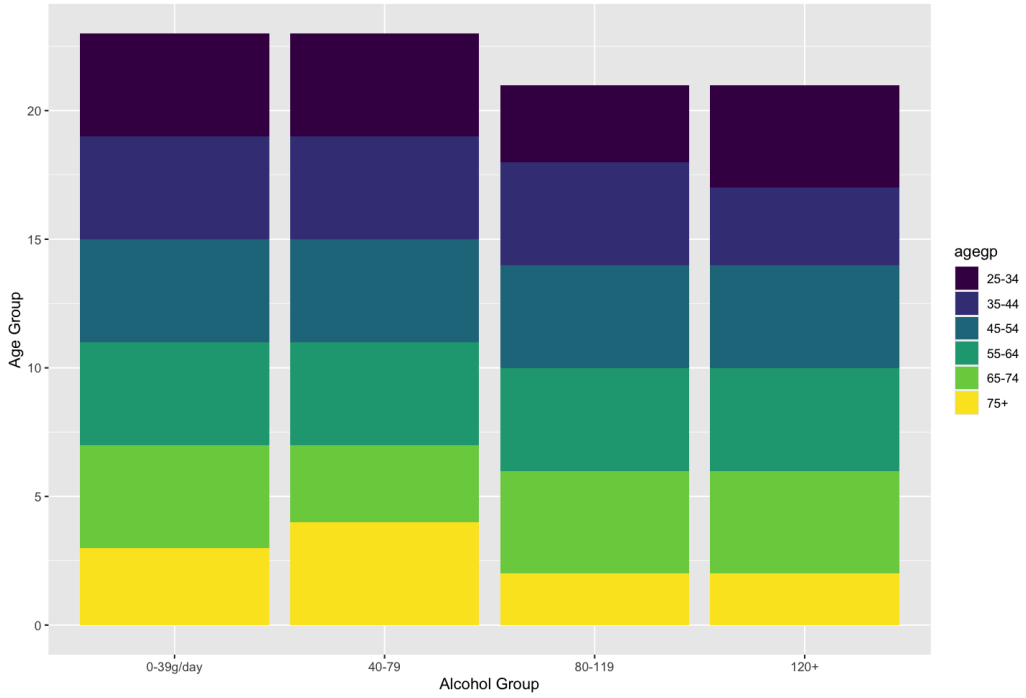
c. Visualize variable `agegp`. Give a more descriptive name to the axis. (Hint: use `geom_bar()` for discrete variables.)

```
> ggplot(data=esoph) + geom_bar(mapping = aes(y=agegp), width = 0.5) + labs(y = "Age Group")
```



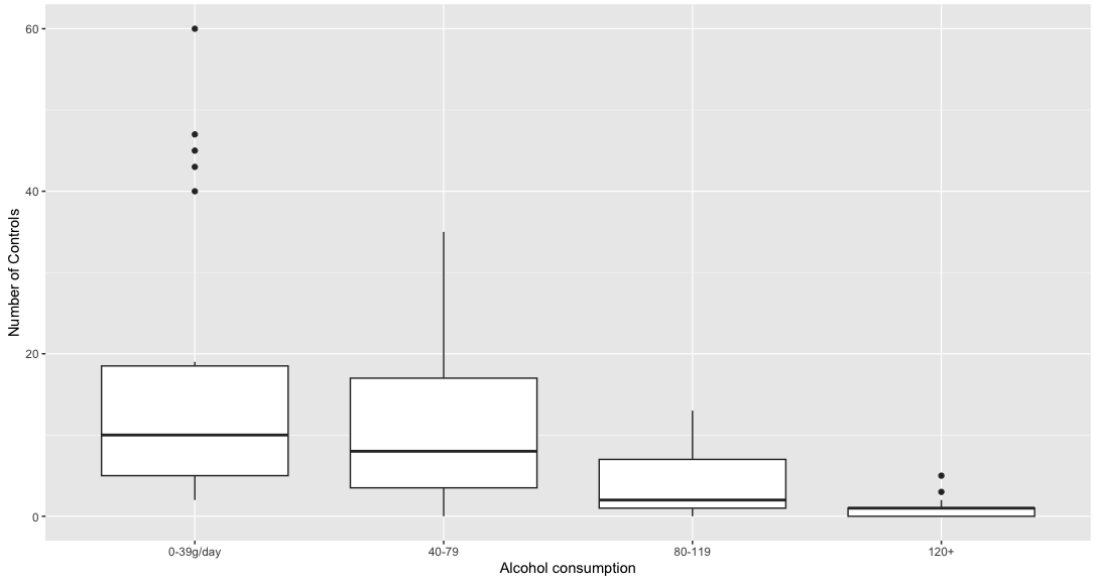
d. Visualize variables `agegp` and `alcgp`.

ggplot(data=esoph) + geom_bar(mapping = aes(x=alcgp,fill=agegp)) + labs(x = "Alcohol Group",y = "Age Group")



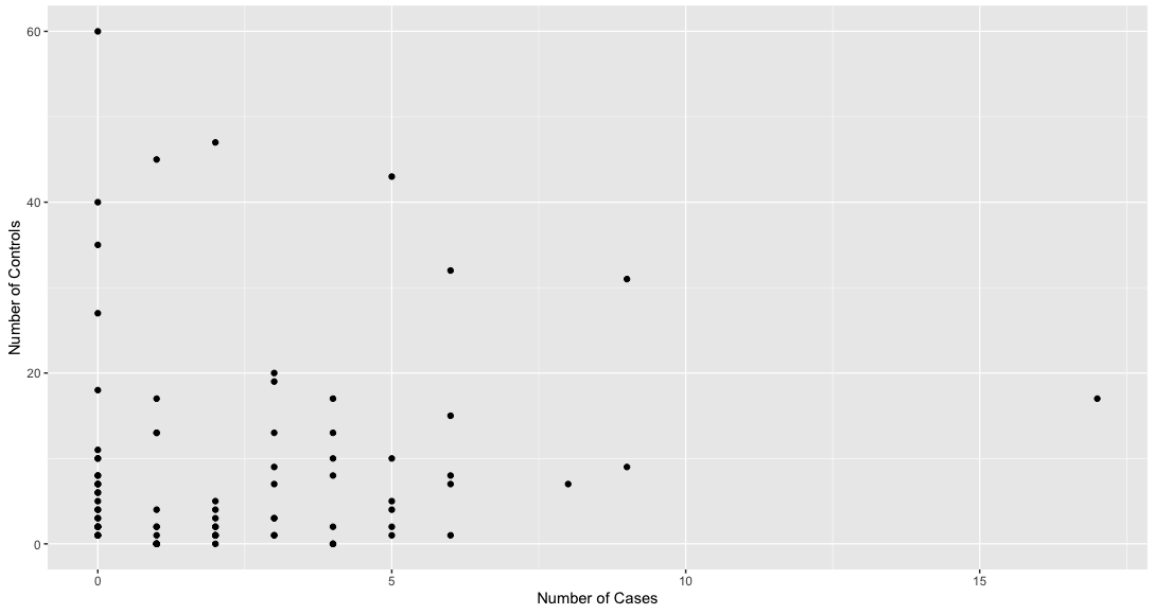
e. Visualize variables `alcgp` and `ncontrols`.

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



f. Visualize variables `ncases` and `ncontrols`.

```
> ggplot(data=esoph) + geom_point(mapping =
aes(x=ncases,y=ncontrols)) + labs(x = "Number of Cases",y = "Number
of Controls")
```



g. Visualize variables `ncases`, `ncontrols`, and `alcgp`.

```
> ggplot(data=esoph) + geom_point(mapping =
aes(x=ncases,y=ncontrols, color = alcgp)) + labs(x = "Number of
Cases",y = "Number of Controls", color="blue")
```