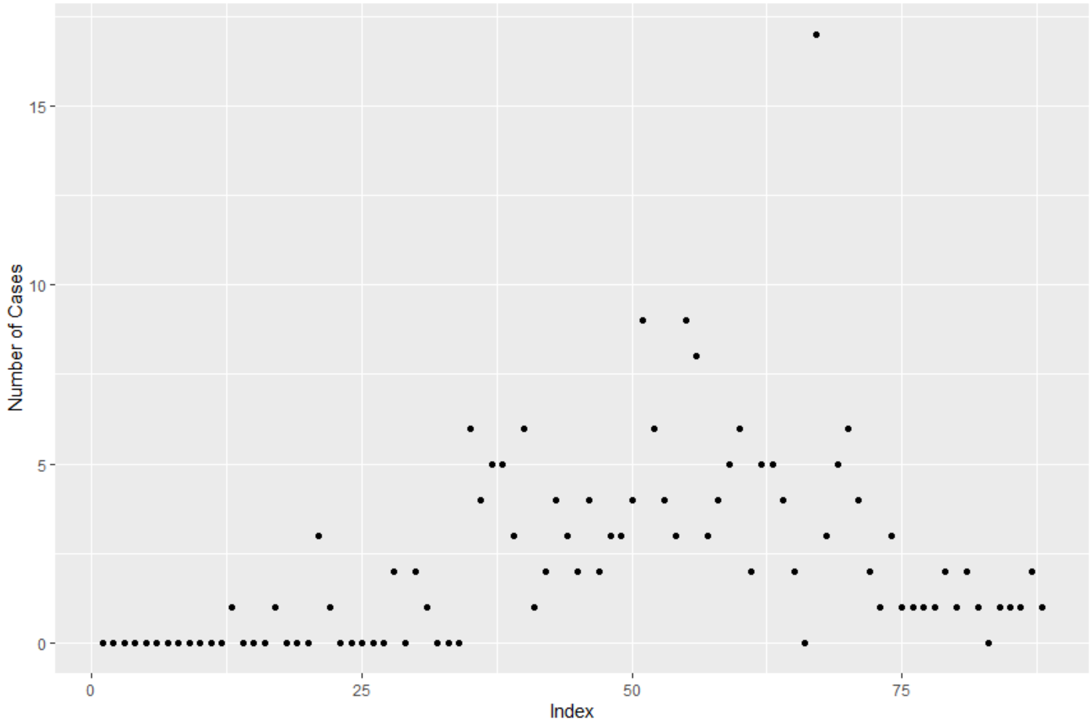
Homework 2

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? What is the type of variable `tobgp`? [Hint: use `str()` and `summary()`]
   i. **Esoph dataset does not contain any NA's. The type of variable tobgp is integer.**
   ii. `summary(esoph)`
   iii. `typeof(esoph$tobgp)`
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? Are these outliers errors or legitimate values?



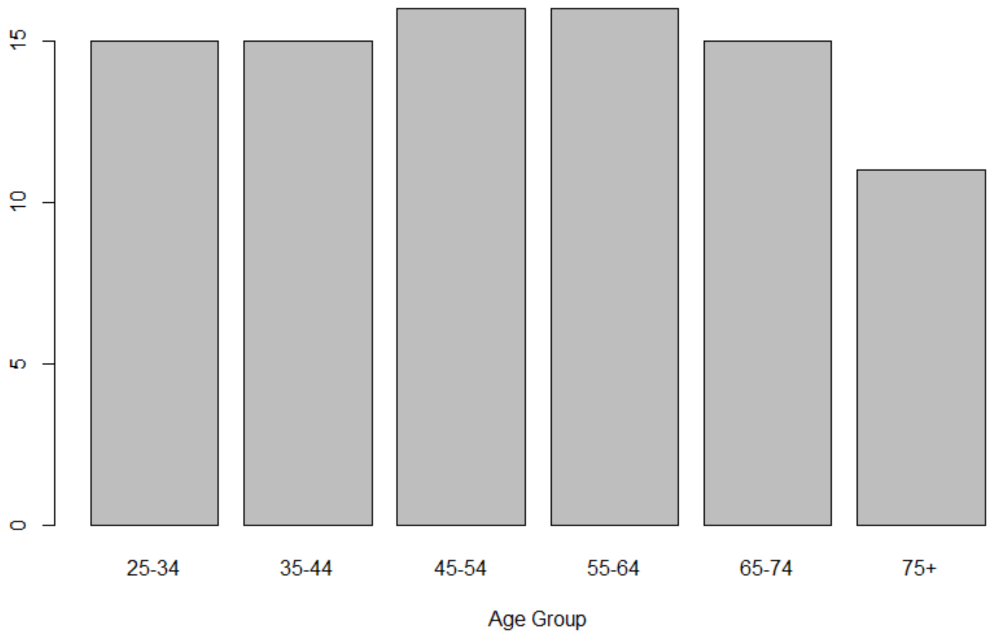
   i.

ii. `ggplot(data=esoph) + geom_point(mapping=aes(x = seq(1, 88), y = ncases)) + labs(y = "Number of Cases", x = "Index")`

iii. **Yes it contains outliers. There is an outlier at 17 cases shown in the upper right corner of the graph. This outlier is an error.**

iv. **Descriptive name: Number of Cases**

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



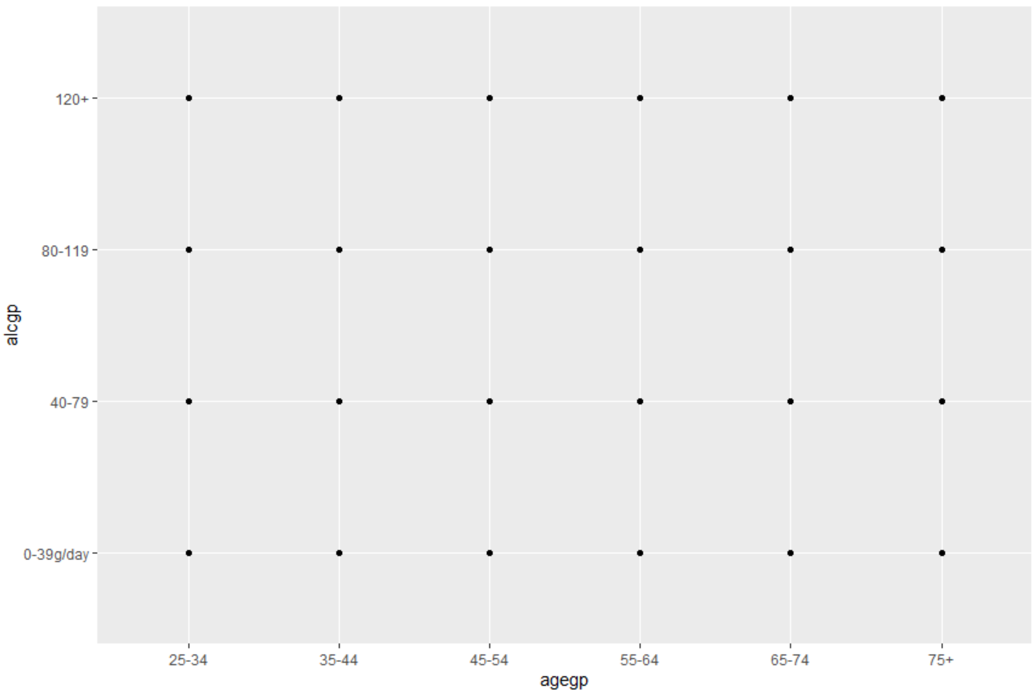
i.

ii. `age <- table(esoph$agegp)`

iii. `barplot(age, xlab="Age Group")`

iv. **Descriptive name: Age Group**

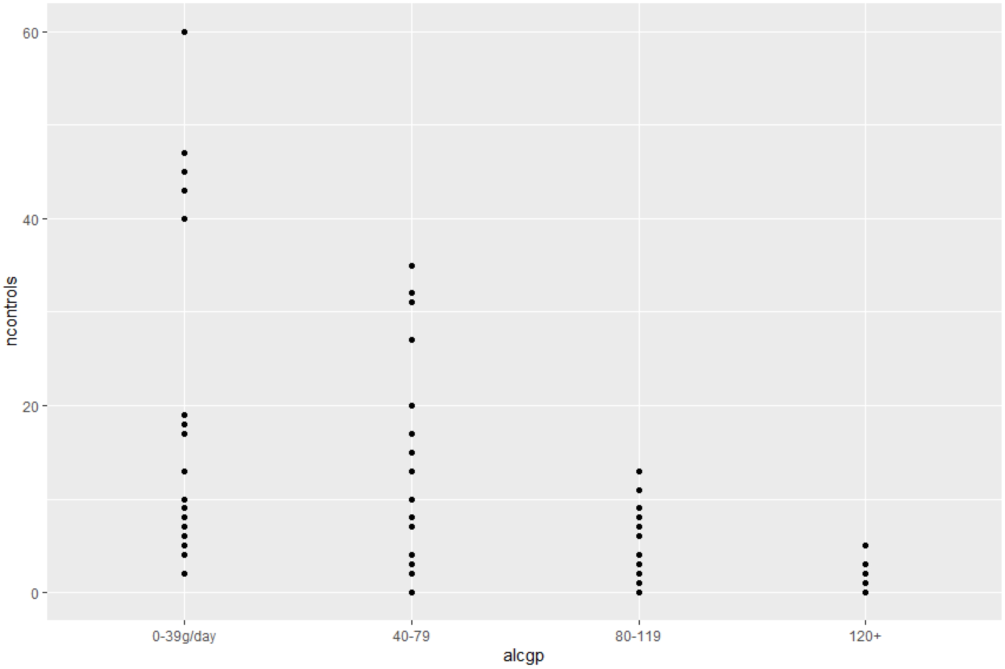
d. Visualize variables `agegp` and `alcgp`.



i.

ii.
```
ggplot(data=esoph) + geom_point(mapping=aes(x = agegp, y = alcgp))
```

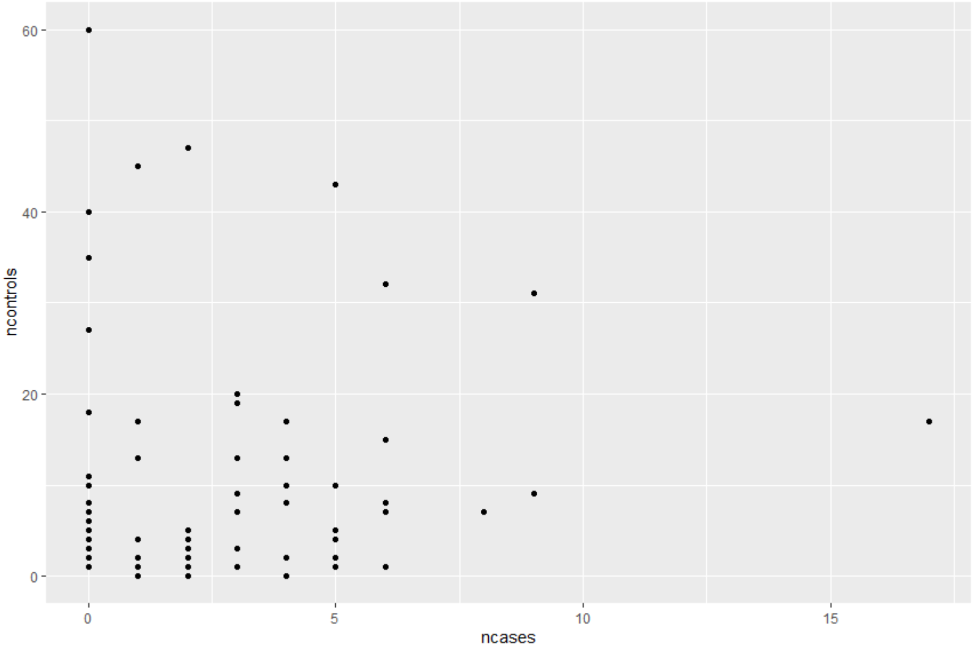
e. Visualize variables `alcgp` and `ncontrols`.



i.

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

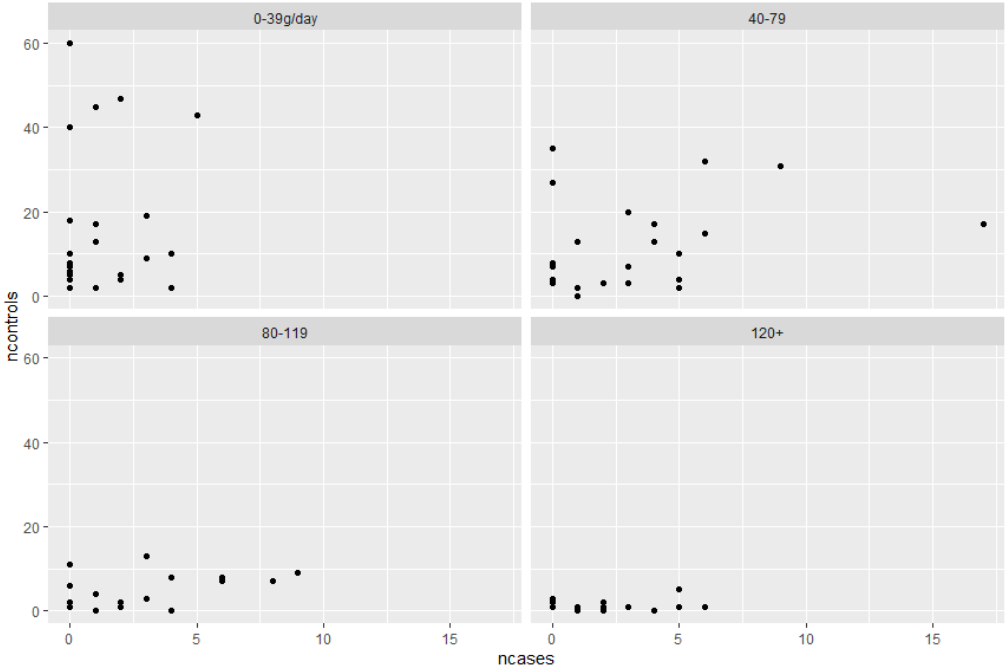
f. Visualize variables `ncases` and `ncontrols`.



i.

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

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



i.

ii. `ggplot(data = esoph) + geom_point(mapping=aes(x=ncases, y=ncontrols)) + facet_wrap(esoph$alcgp)`