1a.
https://catalog.data.gov/dataset/school-attendance-by-student-group-and-district-2021-2022
1b.
The dataset collects data regarding the attendance rate for public school children (k-12) in different student groups and districts. There is no official reason given for why this data is collected, but I believe that it was collected to see how a student's circumstances affect their ability to attend school.
1c. It is available in CSV, RDF, JSON, and XML formats. It contains 1 table, 12 columns, and 2020 rows. This data is from the 2021-2022 school year.
1d. This data can be used to show which children are most affected by their circumstances. It can show which districts have the highest or lowest rate of attendance. This data can be used by school districts and the Government to determine which groups can be helped.

2a.

```
> 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 ...
> class(esoph$tobgp)
[1] "ordered" "factor"
```

2b. Ncases and ncontrols are numeric

2c.

```
> esoph[1:5, c("agegp", "ncases", "ncontrols")]
  agegp ncases ncontrols
1 25-34      0        40
2 25-34      0        10
3 25-34      0         6
4 25-34      0         5
5 25-34      0        27
```

2d and 2e.

```
a <- esoph[esoph$ncases > 5, ]
> a
   agegp  alcgp    tobgp ncases ncontrols
35 45-54  40-79 0-9g/day      6        32
40 45-54 80-119    10-19      6         8
51 55-64  40-79 0-9g/day      9        31
52 55-64  40-79    10-19      6        15
55 55-64 80-119 0-9g/day      9         9
```

```
56 55-64 80-119    10-19      8         7
60 55-64   120+    10-19      6         1
67 65-74  40-79 0-9g/day     17        17
70 65-74 80-119 0-9g/day      6         7
> str(a)
'data.frame':    9 obs. of  5 variables:
 $ agegp    : Ord.factor w/ 6 levels "25-34"<"35-44"<..: 3 3 4 4 4 4 4 5 5
 $ alcgp    : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<..: 2 3 2 2 3 3 4 2 3
 $ tobgp    : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<..: 1 2 1 2 1 2 2 1 1
 $ ncases   : num  6 6 9 6 9 8 6 17 6
 $ ncontrols: num  32 8 31 15 9 7 1 17 7
```

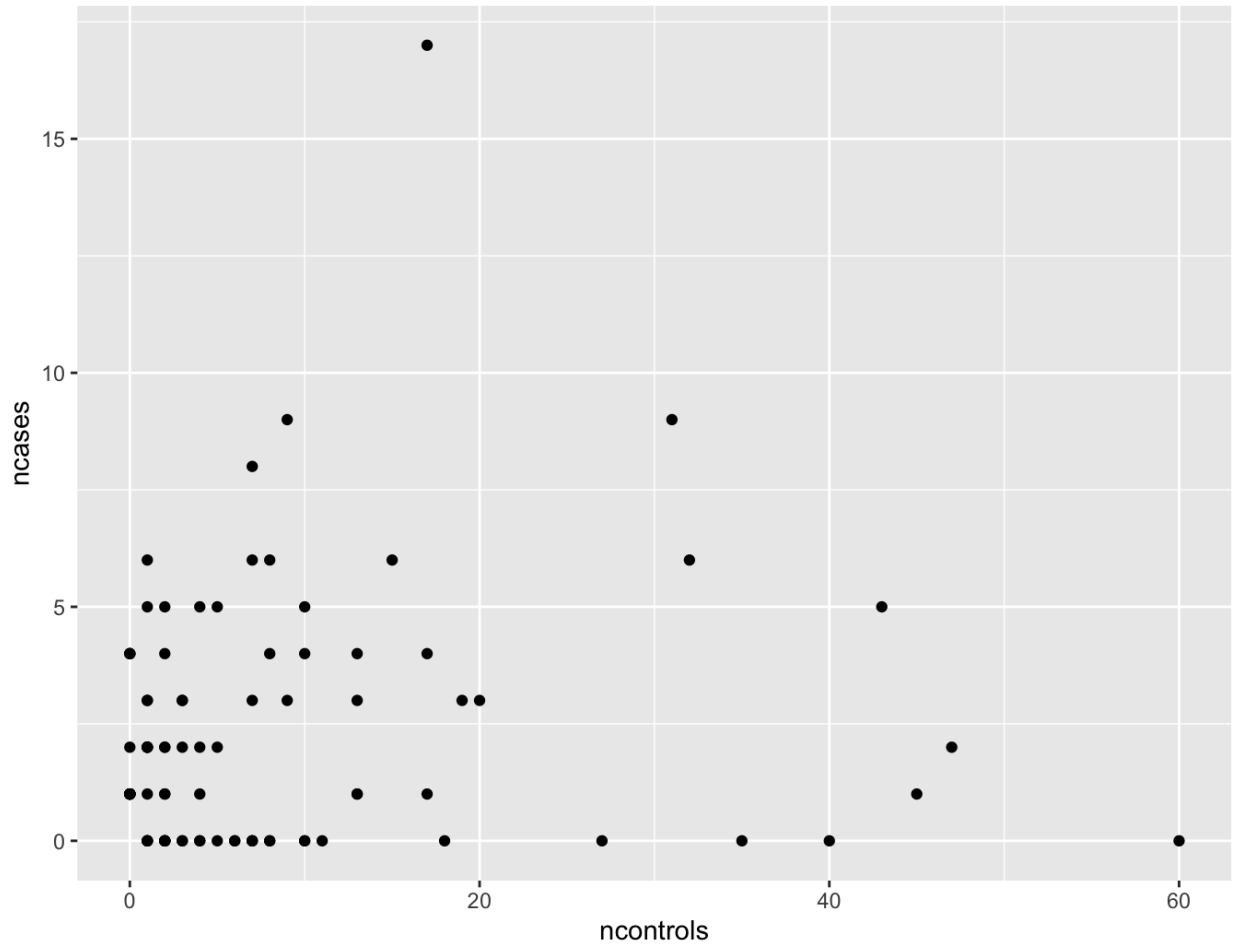
There are 9 rows with their ncases value greater than 5.

2f.

```
> mean(esoph$ncontrols)
[1] 8.806818
```

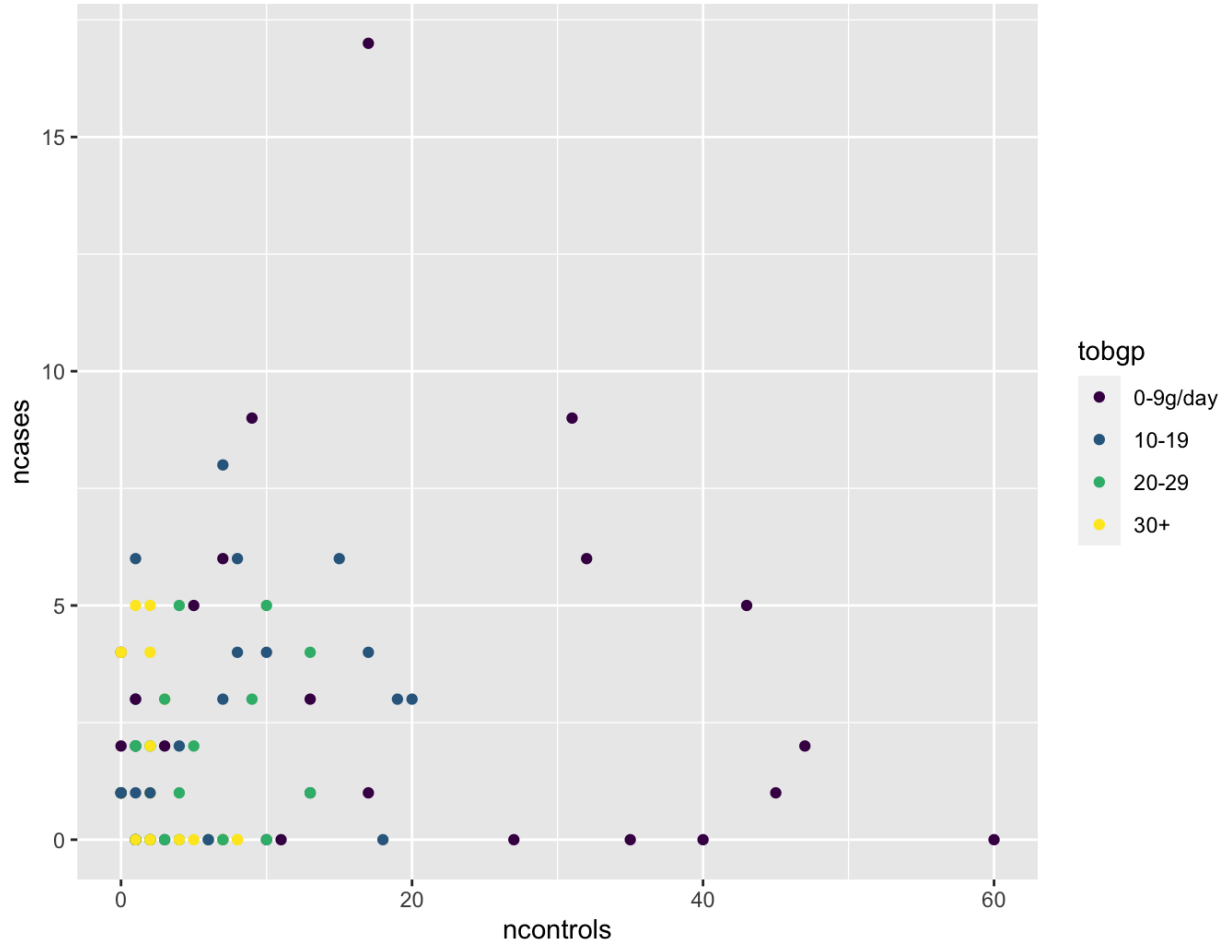
2g.

```
> library(ggplot2)
> ggplot(data = esoph) + geom_point(mapping = aes(x = ncontrols, y = ncases))
```


15
10
5
0
ncases
0
20
40
60
ncontrols

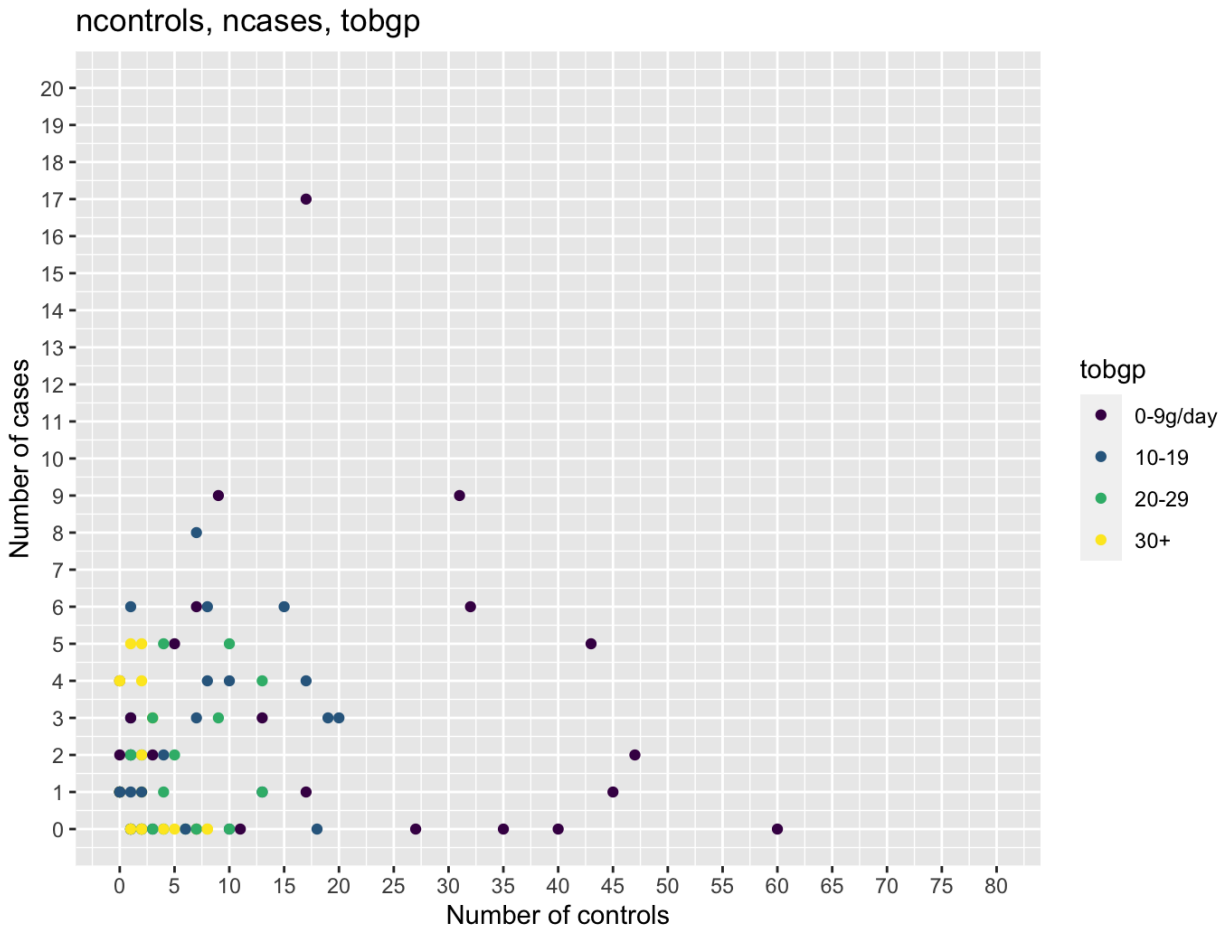
2h.

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



2i.

> ggplot(data = esoph) + geom_point(mapping = aes(x = ncontrols, y = ncases, color = tobgp)) + labs(x="Number of controls", y="Number of cases", title="ncontrols, ncases, tobgp") + scale_x_continuous(limits = c(0, 80), breaks = seq(0, 80, 5)) + scale_y_continuous(limits = c(0, 20), breaks = seq(0, 20, 1))



3a. The 5 parts of the R ecosystem are the RStudio IDE, the R language, tidyverse, various packages and the R community. One community that I found that revolves around R is a YouTube channel R programming 101. This channel teaches the basics to advanced things you can do in R. People who are beginning to learn and experts can both learn from this channel and I will check out his videos since he has videos focusing on some topics from statistics that I have forgotten about like t-test. Link: https://www.youtube.com/@RProgramming101/featured