R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
tik_df <- read_excel("../data_cleaned_inclusions_19-oct.xlsx")

tik_df$author_followers = as.double(tik_df$author_followers)
```

```
## Warning: NAs introduced by coercion
```

Note: there's one author here whose account was deleted but tiktoks weren't... the API can't pull his followers but I can find a manual count. Should I put those in? (which(is.na(tik_df$author_followers)) is row 52)

Let's start with describing the data with some simple tables: (this first one will be ugly)

```
summary(tik_df)
```

```
##    tiktok_id        included            n_ppl              gender
##  Min.   :  1.00   Length:103         Length:103         Length:103
##  1st Qu.: 31.50   Class :character   Class :character   Class :character
##  Median : 64.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 63.56
##  3rd Qu.: 95.50
##  Max.   :124.00
##
##    gender_2             lang             country          support_vaccine
##  Length:103         Length:103         Length:103         Min.   :0.0000
```

```
##  Class :character   Class :character   Class :character   1st Qu.:1.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :1.0000
##                                                          Mean   :0.8155
##                                                          3rd Qu.:1.0000
##                                                          Max.   :1.0000
##
##       hcp          verified_hcp          genre            satire_who
##  Min.   :0.0000   Length:103         Length:103         Length:103
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.1359
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##    correct          satire_comment     why_incorrect      myth_referenced
##  Length:103         Length:103         Length:103         Length:103
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  caption_helpful    infectious_disease_review general_comments
##  Length:103         Length:103                Length:103
##  Class :character   Class :character          Class :character
##  Mode  :character   Mode  :character          Mode  :character
##
##
##
##
##     author              url                likes             shares
##  Length:103         Length:103         Min.   : 141500   Min.   :   138
##  Class :character   Class :character   1st Qu.: 171900   1st Qu.:  3050
##  Mode  :character   Mode  :character   Median : 227200   Median :  9028
##                                        Mean   : 428412   Mean   : 30868
##                                        3rd Qu.: 443550   3rd Qu.: 20000
##                                        Max.   :2800000   Max.   :807500
##
##     comments     author_followers       plays          description
##  Min.   :    0   Min.   :     427   Min.   :  515000   Length:103
##  1st Qu.: 1470   1st Qu.:    8263   1st Qu.: 1100000   Class :character
##  Median : 3721   Median :   62200   Median : 1600000   Mode  :character
##  Mean   : 6202   Mean   :  485681   Mean   : 2860666
##  3rd Qu.: 6425   3rd Qu.:  252550   3rd Qu.: 3100000
##  Max.   :62600   Max.   :13300000   Max.   :24800000
##                  NA's   :1
```

```
tik_df %>% group_by(support_vaccine) %>% summarize(n=n(), median_plays=median(plays, na.rm=TRUE), media
```

```
## # A tibble: 2 x 7
##   support_vaccine     n median_plays median_comments median_shares
##             <dbl> <int>        <dbl>           <dbl>         <dbl>
## 1               0    19      2900000            6253         17700
## 2               1    84      1600000            2999          9025
```

```
## # ... with 2 more variables: median_followers <dbl>, percent_female <dbl>
```

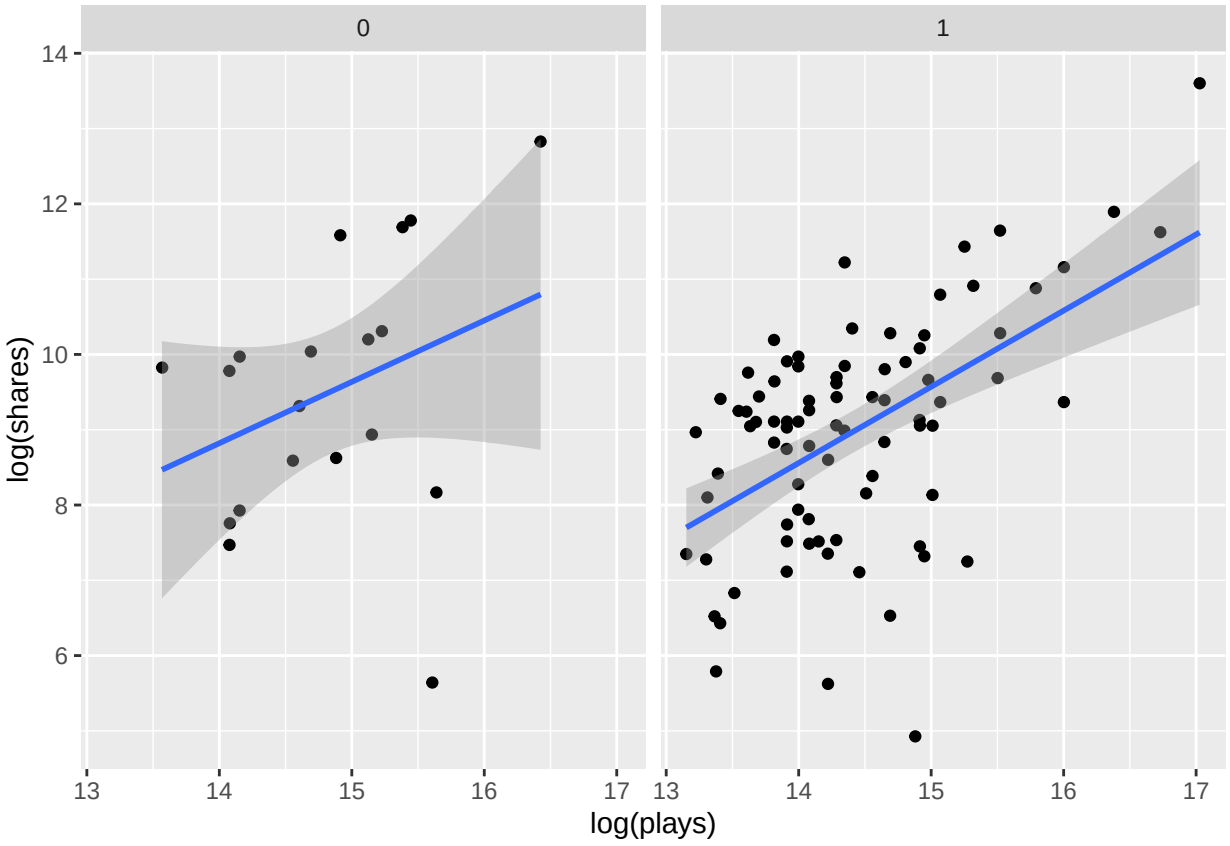
These stats look very different between vaccine supporters and not. Let's make some graphs to see what's going on.

First some sanity checks - we expect to see a positive correlation b/w plays and shares, for example. This looks more or less legit in those who support and don't support vaccine. This isn't very important imo, don't put too much stock into it.

```
p <- ggplot(tik_df, aes(log(plays), log(shares))) #log-ing them both so i can see trends.

p+geom_jitter() + facet_grid(. ~ support_vaccine) + geom_smooth(method=lm)
```

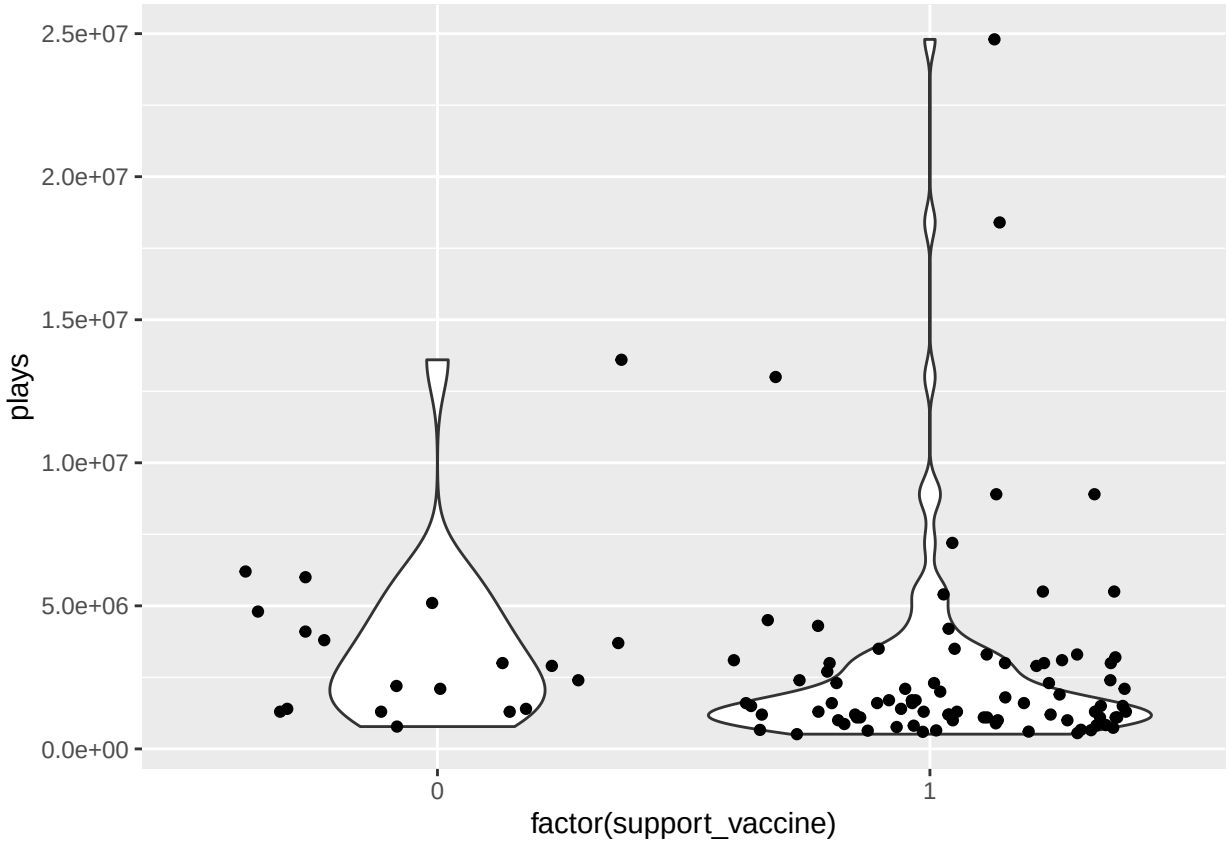
```
## `geom_smooth()` using formula 'y ~ x'
```



Plays

```
p <- ggplot(tik_df, aes(factor(support_vaccine), plays))

p+geom_violin()+geom_jitter()
```



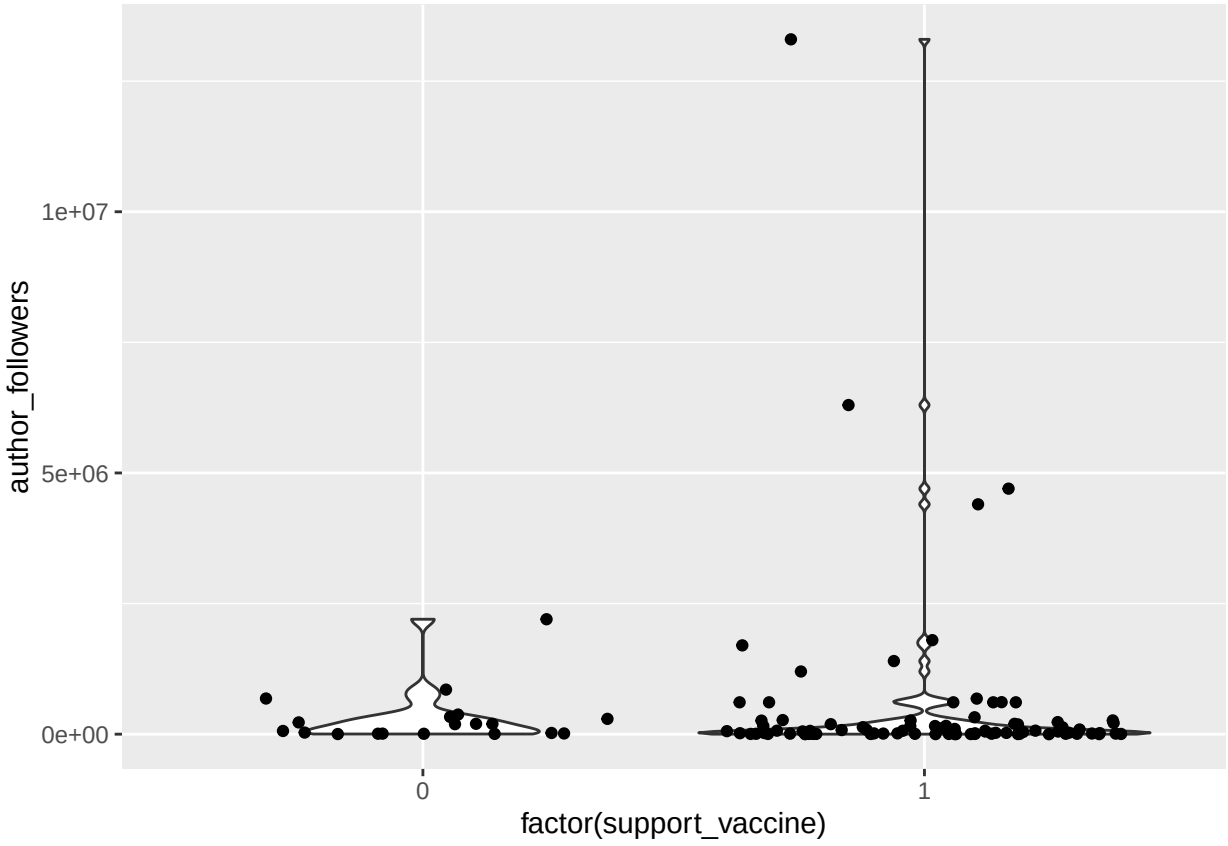
Followers

```
p <- ggplot(tik_df, aes(factor(support_vaccine), author_followers))

p+geom_violin()+geom_jitter()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_ydensity).

## Warning: Removed 1 rows containing missing values (geom_point).
```



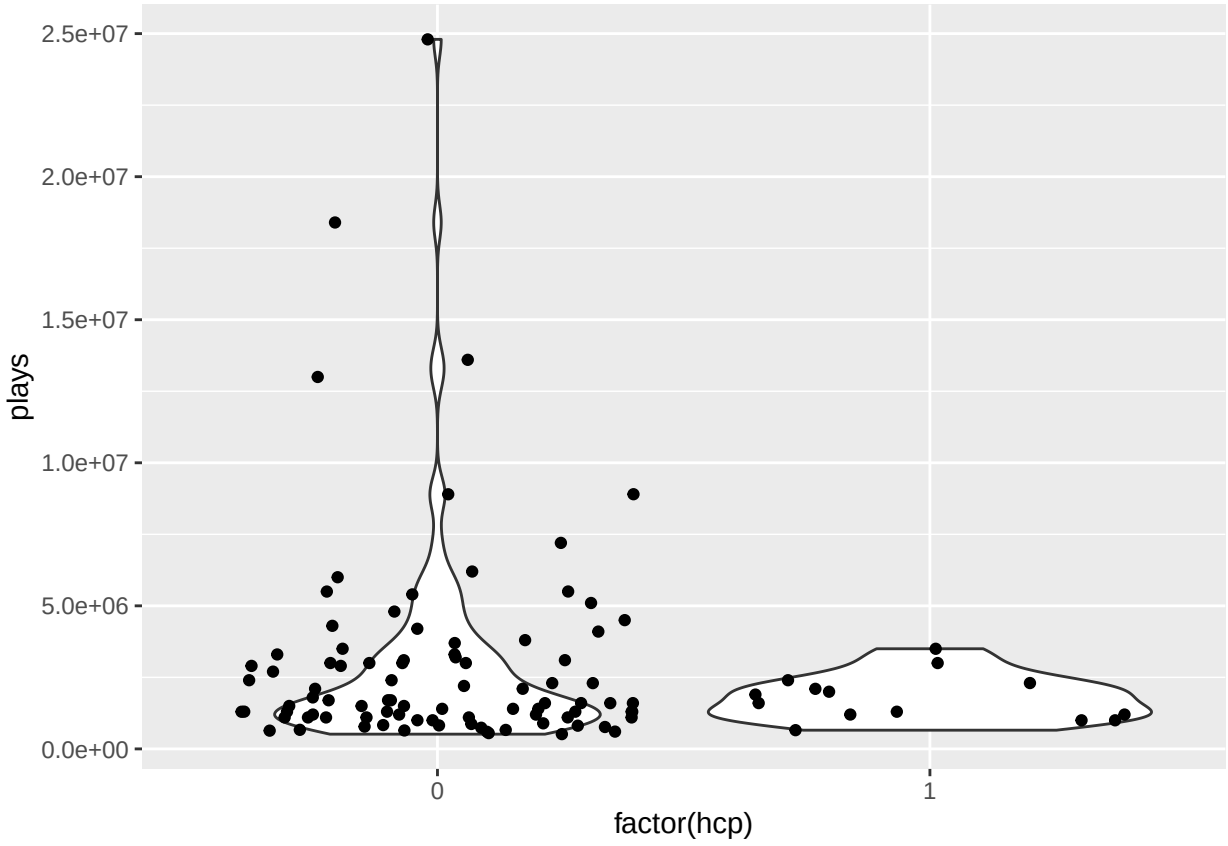
My N is fairly small but I'm still curious to compare health care expert content to layperson produced content

```
tik_df %>% group_by(hcp) %>% summarize(n=n(), median_plays=median(plays, na.rm=TRUE), median_comments=m
```

```
## # A tibble: 2 x 7
##     hcp     n median_plays median_comments median_shares median_followers
##   <dbl> <int>        <dbl>           <dbl>         <dbl>            <dbl>
## 1     0    89      1600000            3545          9224            58100
## 2     1    14      1750000            4894.         7956.          155900
## # ... with 1 more variable: percent_female <dbl>
```

```
p <- ggplot(tik_df, aes(factor(hcp), plays))

p+geom_violin()+geom_jitter()
```

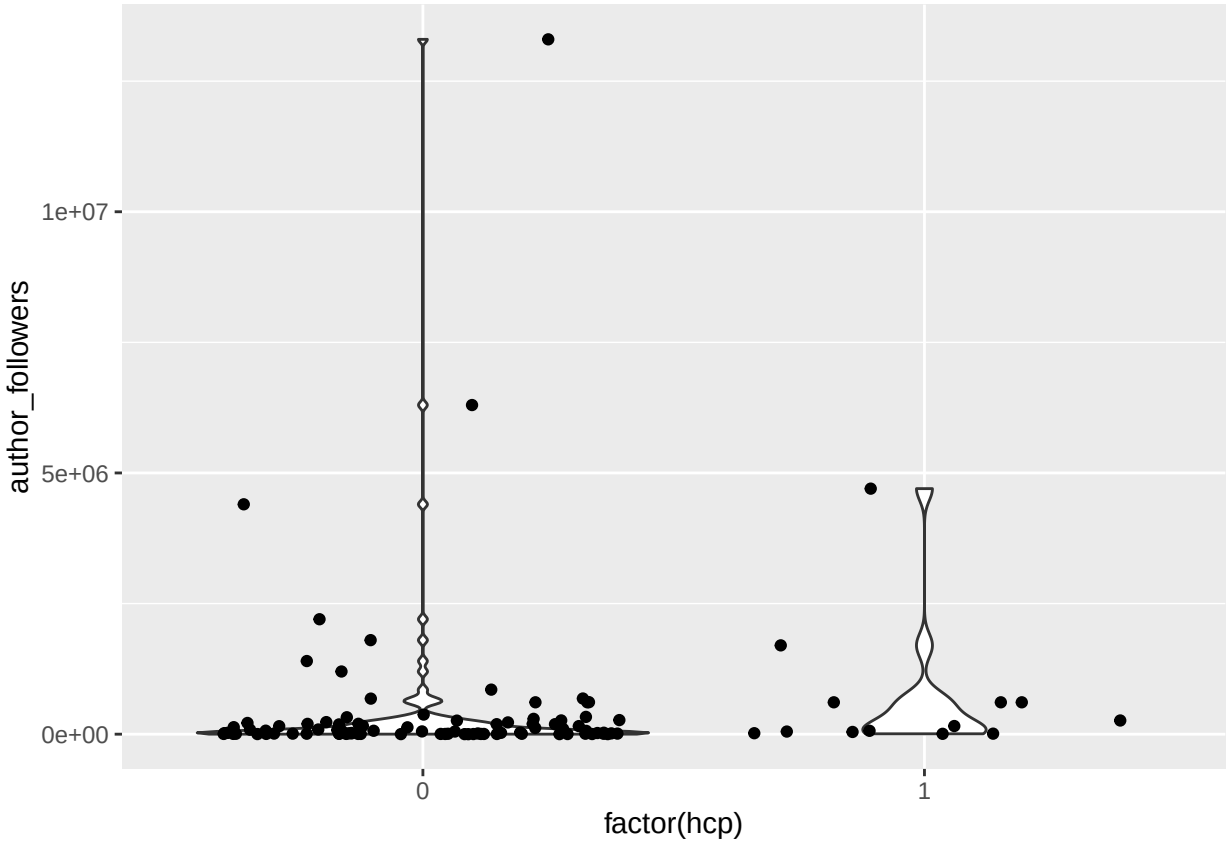



```
p <- ggplot(tik_df, aes(factor(hcp), author_followers))

p+geom_violin()+geom_jitter()
```

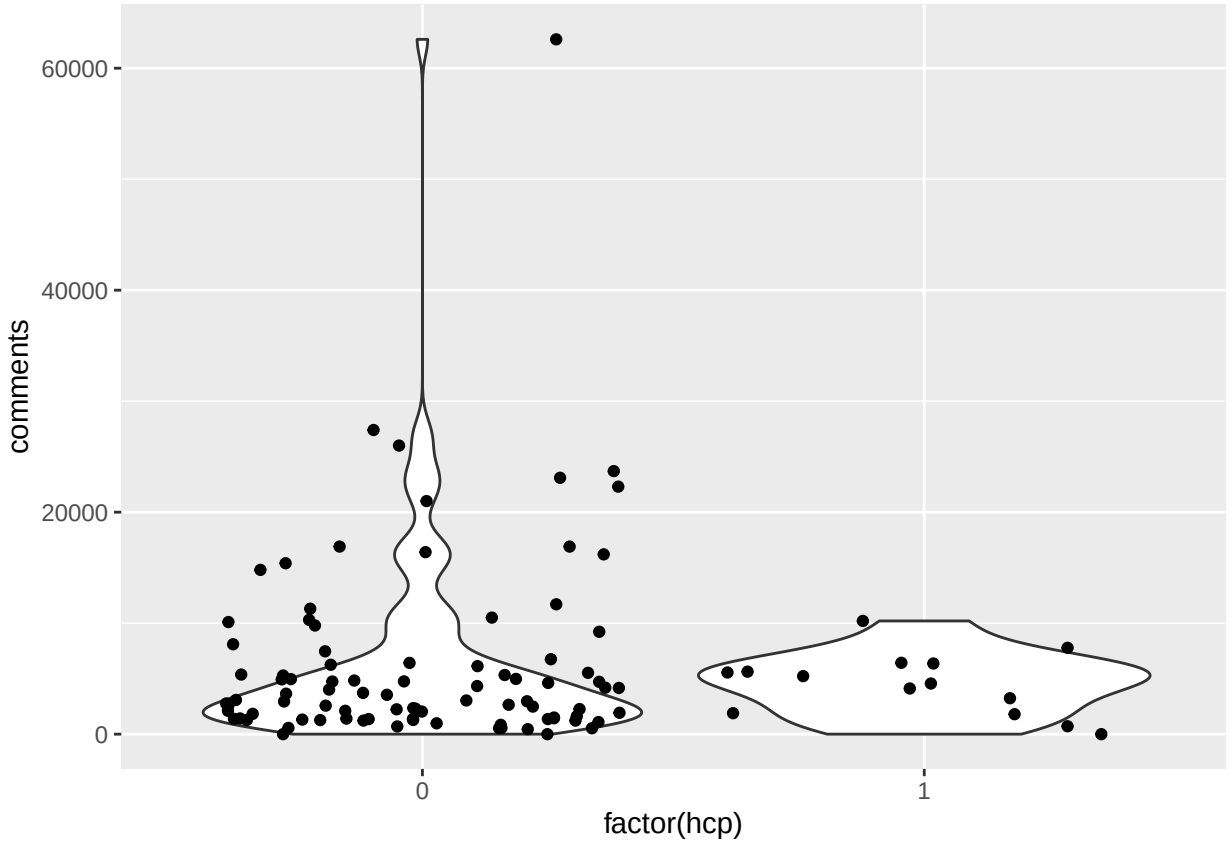
```
## Warning: Removed 1 rows containing non-finite values (stat_ydensity).

## Warning: Removed 1 rows containing missing values (geom_point).
```




```
p <- ggplot(tik_df, aes(factor(hcp), comments))

p+geom_violin()+geom_jitter()
```


60000
40000
20000
0
comments
0
1
factor(hcp)