My First Exercise

Warren Liu

May 7th 2020

Contents

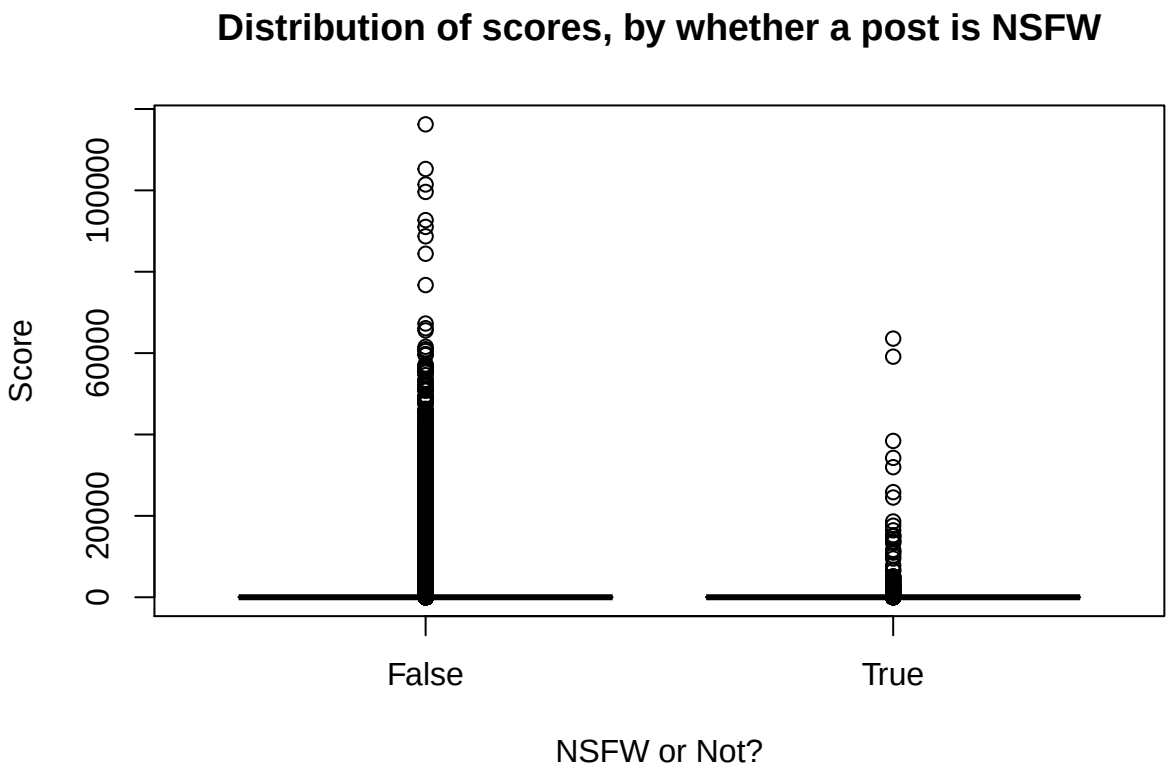
NSFW Posts? Ooh, getting saucy here aren't we?

Firstly, let's explore the relation between Whether a Post is marked Over 18 and the amount of points it receives

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

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

```
library(ggplot2)

data_df <- read.csv("r_dataisbeautiful_posts.csv")

boxplot(score ~ over_18 , data= data_df, main="Distribution of scores, by whether a post is NSFW",
   xlab="NSFW or Not?", ylab="Score")
```



As we see, the IQR is quite similar, and there are a lot of outliers. However, it seems that non NSFW posts generally garnish more points.

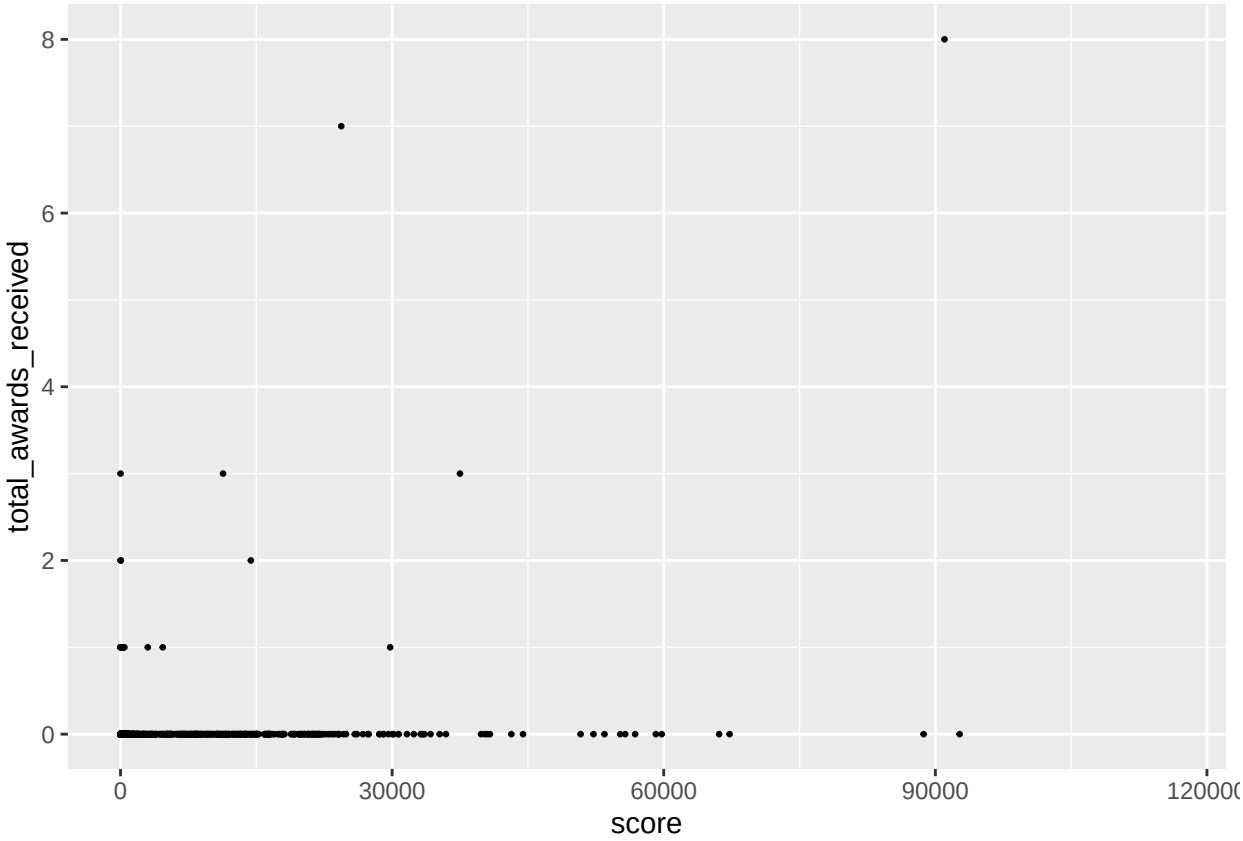
Well, data isn't always beautiful;

Scores vs Awards Received

Why don't we see how generous our community is?

```
library(ggplot2)
ggplot(data_df, aes(x=score, y=total_awards_received)) + geom_point(size=0.5)
```

```
## Warning: Removed 140006 rows containing missing values (geom_point).
```

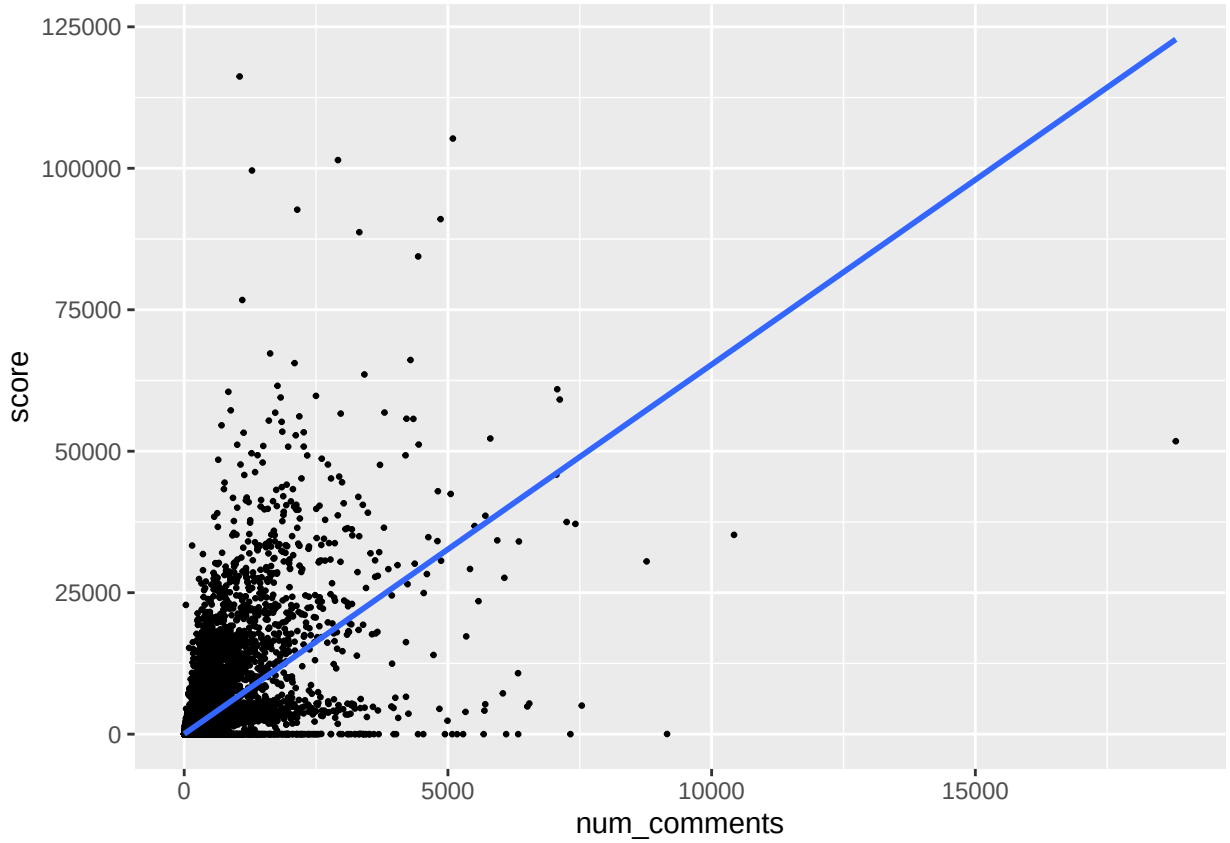


Well crap. It seems like people generally don't give many awards on reddit.

What is another thing we could potentially explore?

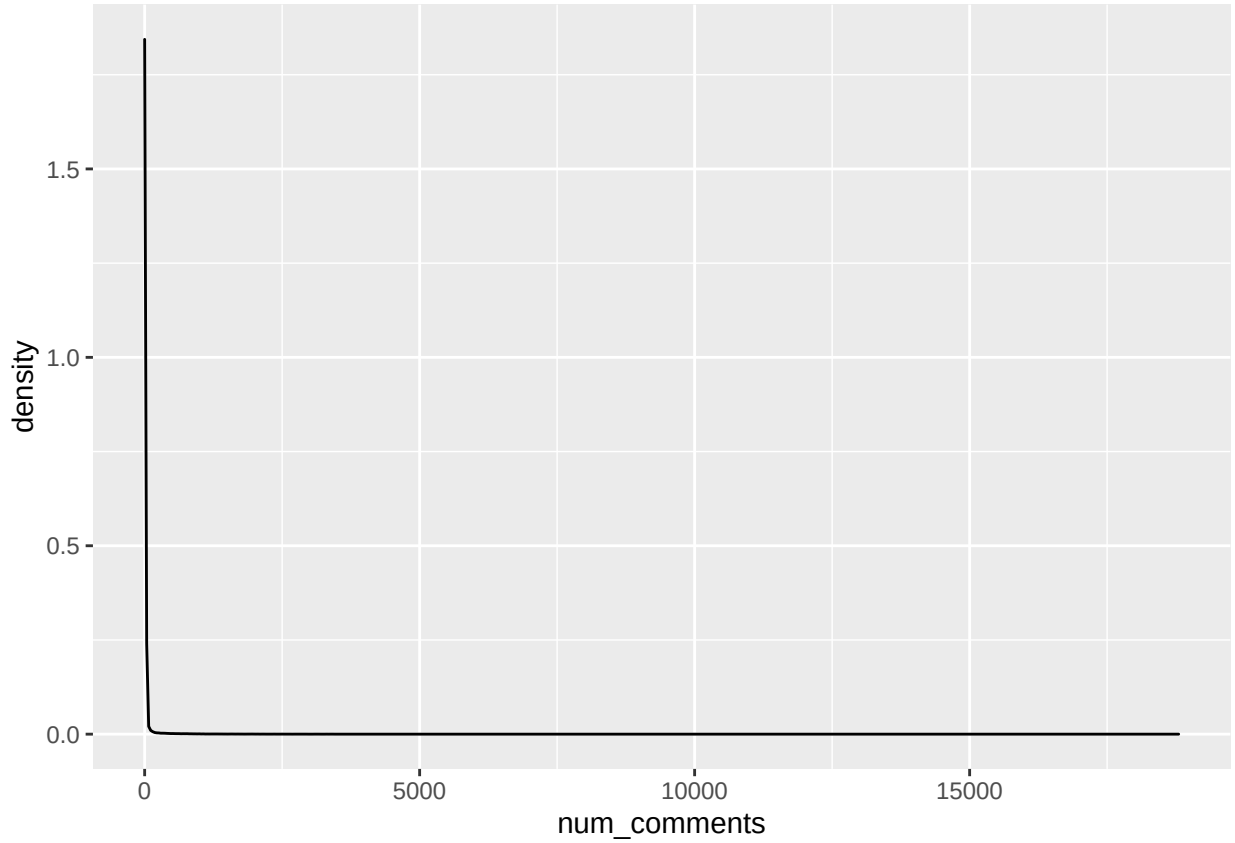
```
library(ggplot2)
ggplot(data_df, aes(x=num_comments, y=score)) + geom_point(size=0.5) + geom_smooth(method=lm, se=FALSE)
```

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



Bah, still Inconclusive (at least the relationship isn't independent). Why don't we try another type of graph?

```
p <- ggplot(data_df, aes(x=num_comments)) +
  geom_density()
print(p)
```



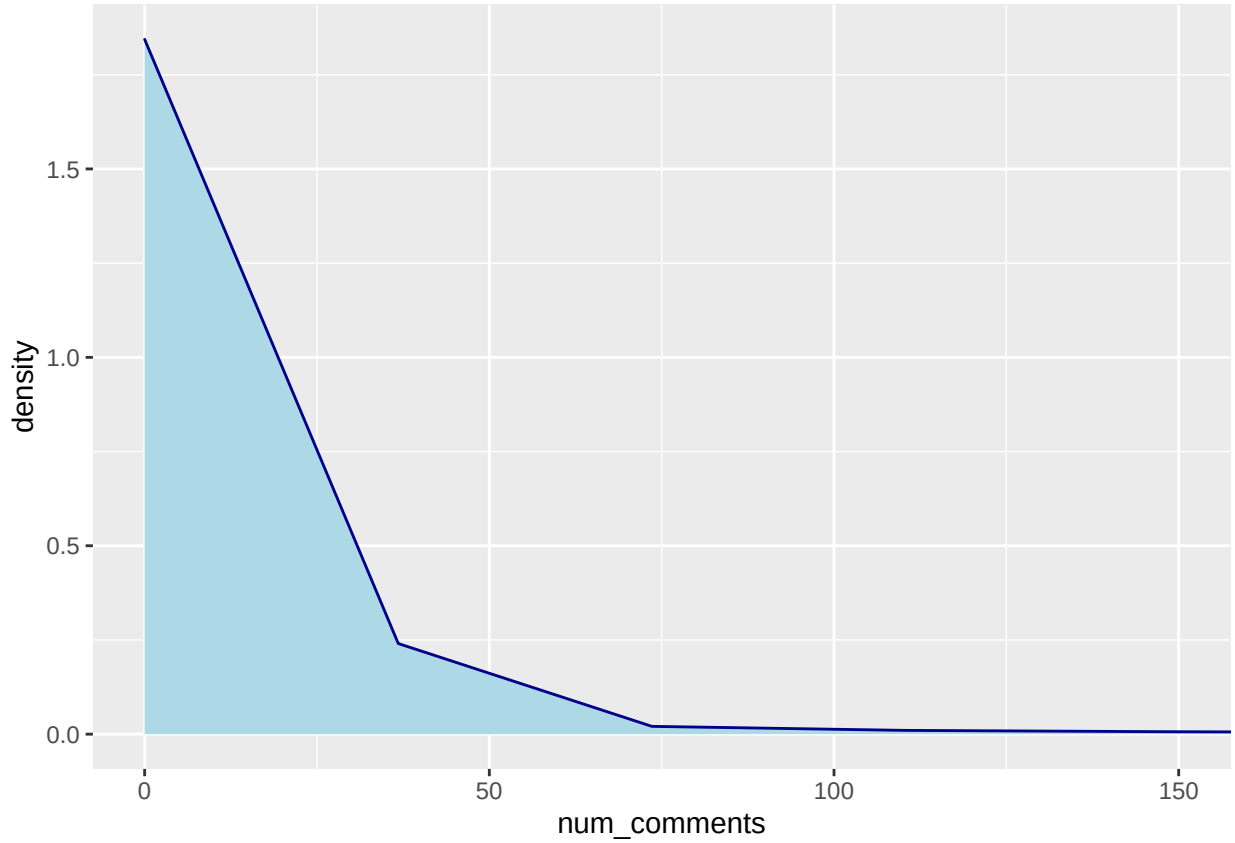
Well damn, that was way too big. It almost looks like the axis. Thus, lets perform some calculations and eliminate all the outliers.

```
data_num_comm = data_df$num_comments

summary(data_num_comm)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00     1.00     1.00    25.28     4.00 18801.00
```

Okay, this means that the limit for outliers is Q3 + 1.5IQR should be 8.5 or 8 comments. This suggests that most posts tend to generally Die in NEW, and Reddit's algorithm tends to promote only a select few posts to be trending. For the sake of data, i'll expand it to 150 comments. This shows a distinct left skew.

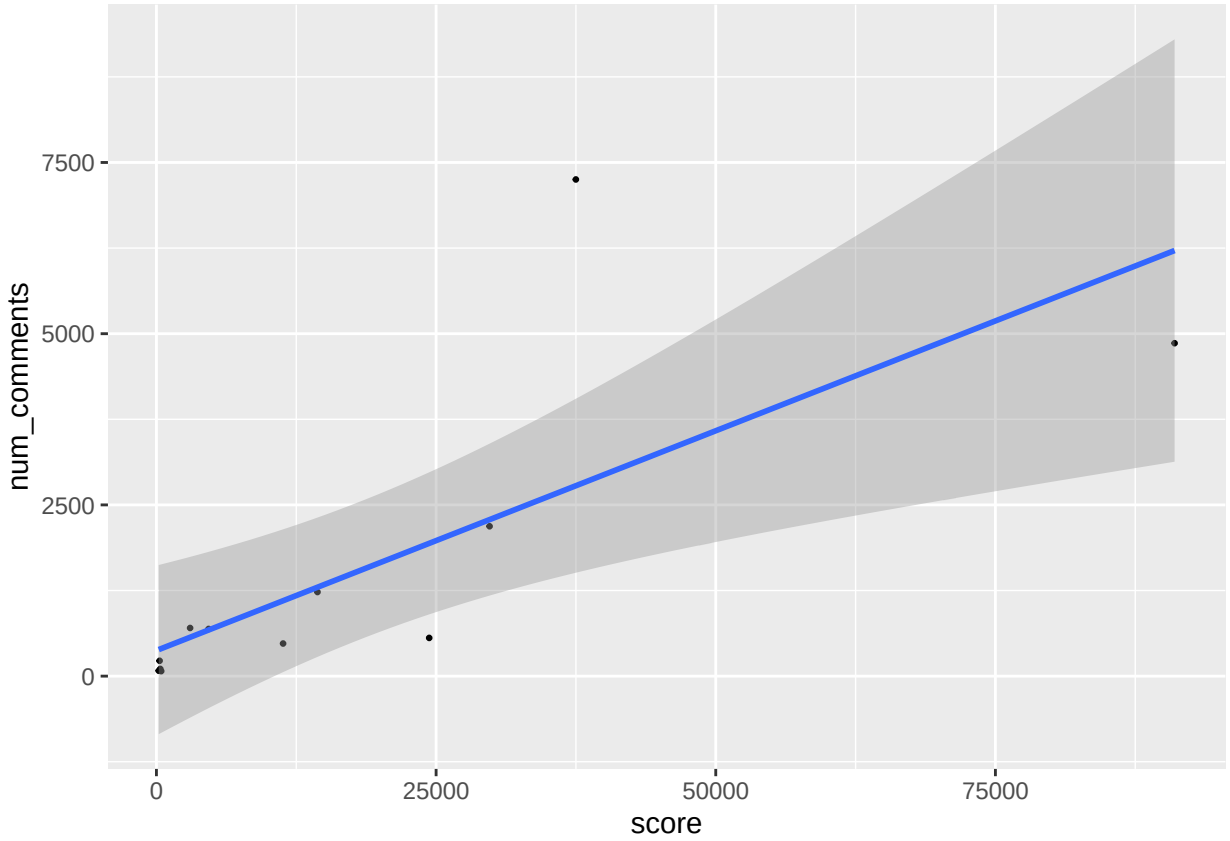
```
ggplot(data_df, aes(x=num_comments)) +
  geom_density(color="darkblue", fill="lightblue") + coord_cartesian(xlim = c(0,   150))
```



Now Let us Filter out some undesirable Variables. We'll filter out the no awards, and ensure that the score is at least 100.

```
library(ggplot2)

filt_df <- data_df %>% filter(
  100 < score,  total_awards_received > 0
  )

ggplot(filt_df, aes(x=score, y=num_comments)) + geom_point(size=0.5) + geom_smooth(method=lm)
```

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



Welp, Guess I'll stick to matplotlib and python.