Exploration of IMDB Quotes

Tobias Machnitzki, Finn Burgemeister

5 Dezember 2017

R Markdown

```
library(ggplot2)
DATAPATH = "C:/Users/darkl/Dropbox/Transfer1/"
DATAPATH = paste(DATAPATH, "imdb-quotes.csv", sep="")
d = read.csv(DATAPATH, header=F, stringsAsFactors=F, sep="|", quote="\"")
colnames(d) = c("movie", "year", "episode", "actors", "quote")

# data cleaning
d$year = as.numeric(d$year)
```

```
## Warning: NAs durch Umwandlung erzeugt
```

```
d = d[d$year > 1800 & d$year < 2020 & ! is.na(d$year),]
```

First have a look at what we've got:

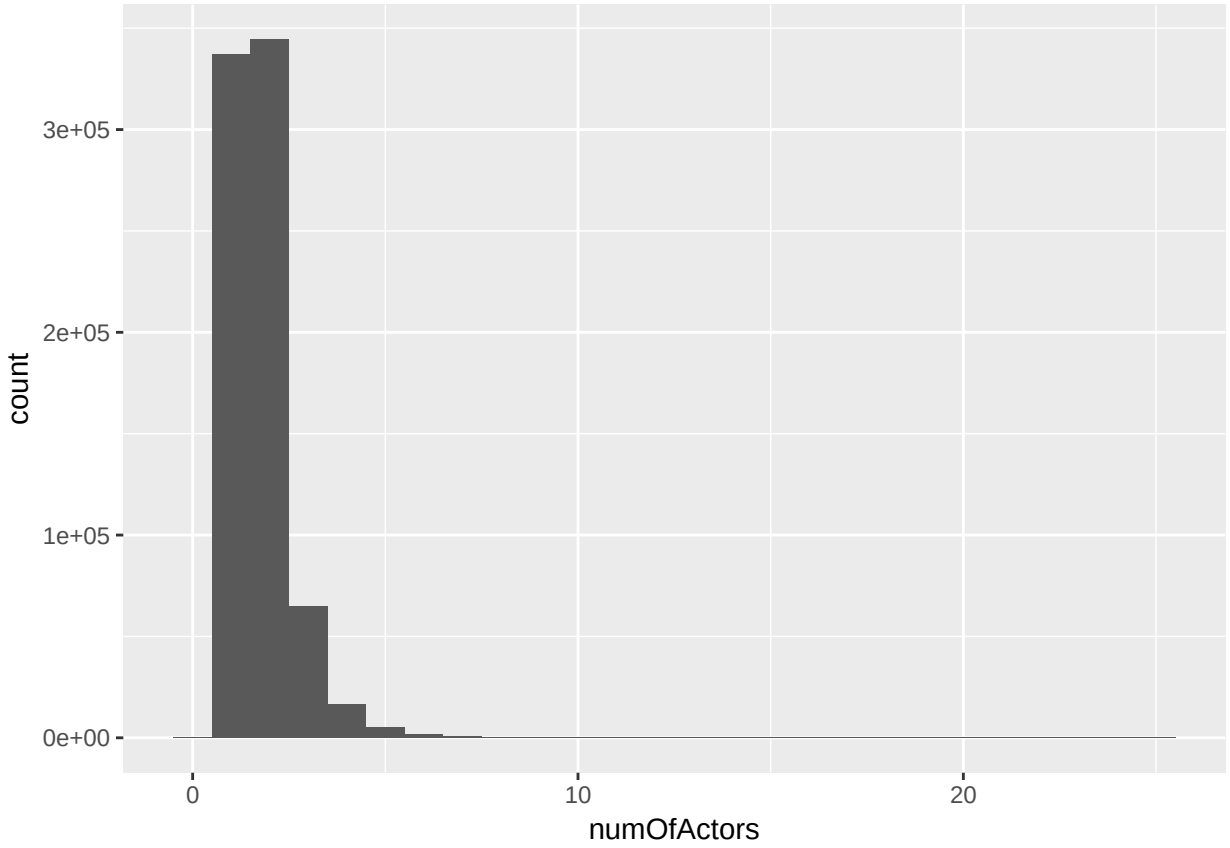
```
summary(d)
```

```
##     movie                year        episode             actors
##  Length:771795      Min.   :1894   Length:771795      Length:771795
##  Class :character   1st Qu.:1986   Class :character   Class :character
##  Mode  :character   Median :1999   Mode  :character   Mode  :character
##                     Mean   :1993
##                     3rd Qu.:2006
##                     Max.   :2019
##     quote
##  Length:771795
##  Class :character
##  Mode  :character
##
##
##
```

Calculate the number of acters for each movie:

```
numOfActors = rep(777,length(d$actors))
for (i in c(0:length(d$actors))){
numOfActors[i] = length(unlist(strsplit(d$actors[i],",")))
}

qplot(numOfActors,geom="histogram",binwidth=1)
```

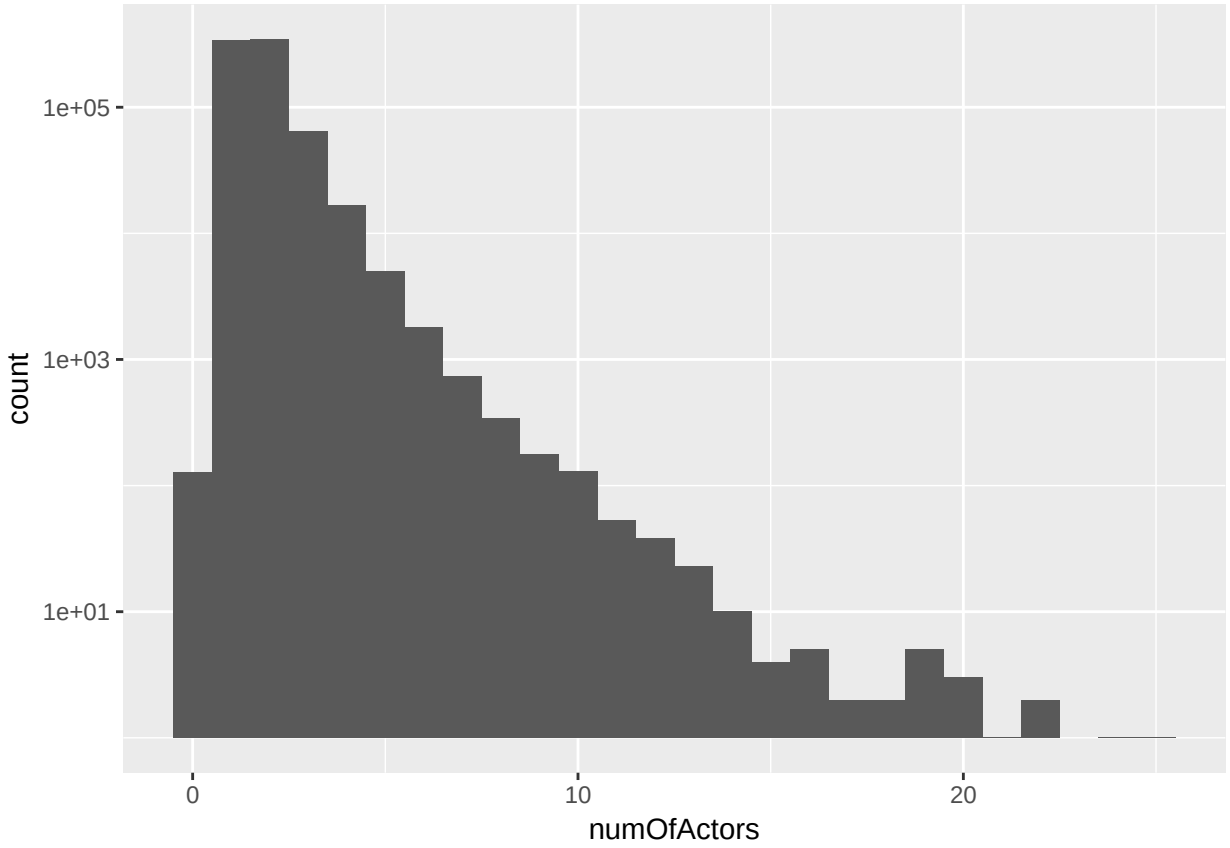


There seem to be some outliers on the very right side, lets try to visualize them:

```
qplot(numOfActors,geom="histogram", log="y",binwidth=1)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

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



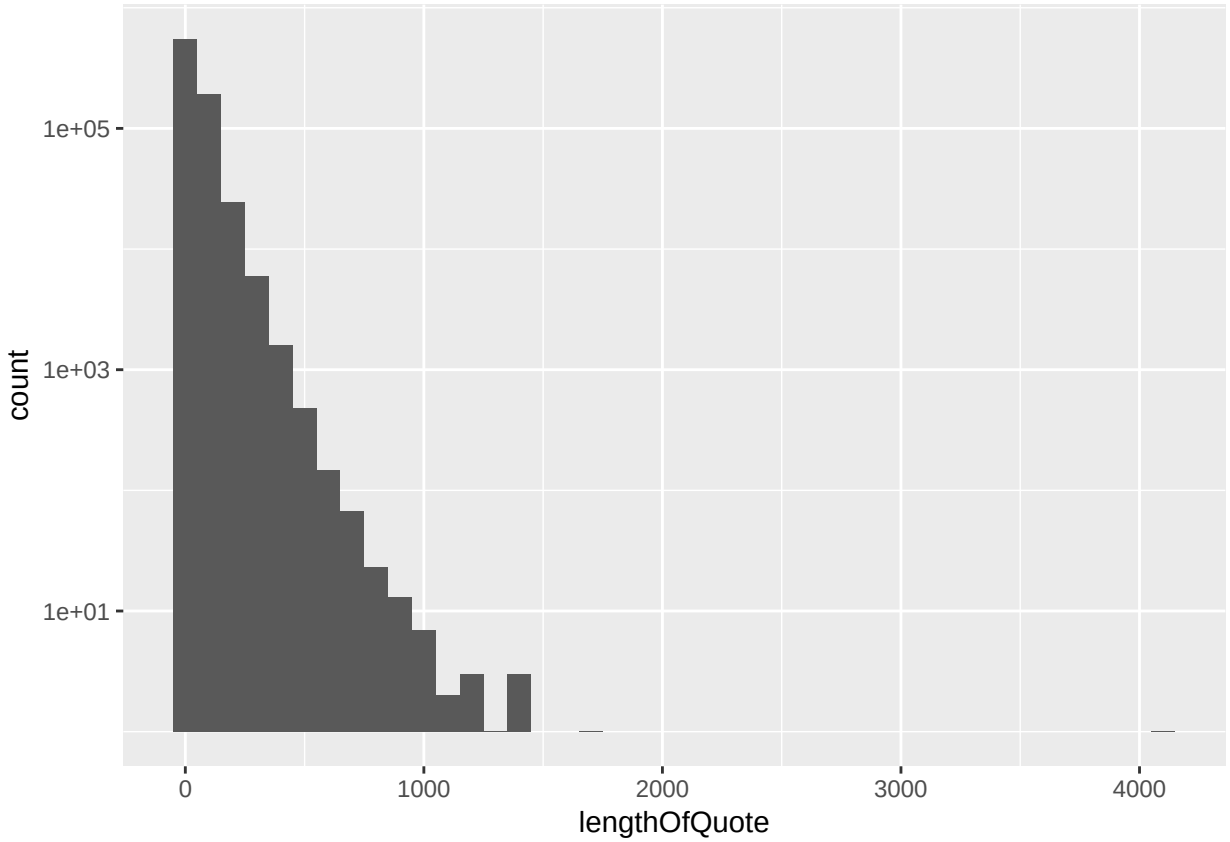
So the huge majority of quotes have 2 to 3 actors involved.

Lets Calculate the number of words in each quote:

```
lengthOfQuote = rep(777,length(d$quote))
for (i in c(0:length(d$quote))){
lengthOfQuote[i] = unlist(sapply(gregexpr("\\w+", d$quote[i]), length))
}
lengthOfQuote = data.frame(lengthOfQuote)

qplot(lengthOfQuote,geom="histogram",log="y",binwidth=100)
```

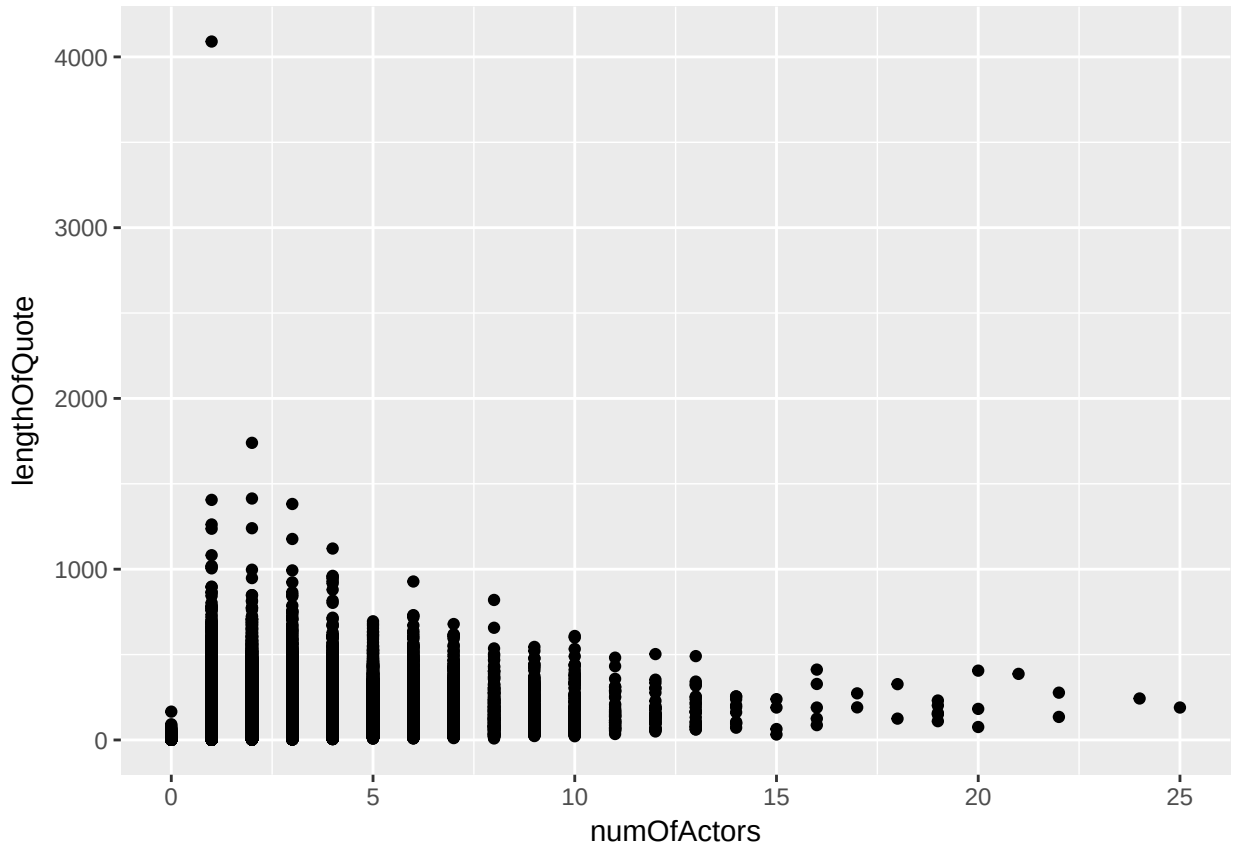
```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 25 rows containing missing values (geom_bar).
```



So how do the length correlate with the number of actors?

```
r = cor(numOfActors,lengthOfQuote)
qplot(numOfActors,lengthOfQuote,geom="point")
```

```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
```



So actually the length of the article has nothing to do with the number of actors taking part.

```
summary(lengthOfQuote)
```

```
##  lengthOfQuote
##  Min.   :   1.00
##  1st Qu.:  18.00
##  Median :  31.00
##  Mean   :  46.87
##  3rd Qu.:  56.00
##  Max.   :4090.00
```

So most quotes have about 31 words but the longest quote seems to be a monologue.

```
quoteWord = character(length(d$quote))
wordOccurences = rep(777,length(d$quote))
for (i in c(0:length(d))){
  quote_length = 0
  for (word in strsplit(d$movie[i], " ")){
    act_length = length(grep(word,strsplit(d$quote[i], " ")))
    if (act_length > quote_length) {
      quote_length = act_length
      act_word = word
    }
    quoteWord[i] = word
    wordOccurences[i] = quote_length
  }
}
```

```
## Warning in grep(word, strsplit(d$quote[i], " ")): Argument 'pattern' hat
## einen Länge > 1 und nur das erste Element wird benutzt

## Warning in quoteWord[i] <- word: Anzahl der zu ersetzenden Elemente ist
## kein Vielfaches der Ersetzungslänge
```