Alp 301: Stones 2 Milestones, Data Descriptives

April 2021

Contents

Learning Objective

The team will review data from Stones2Milestones in order to gain an understanding of the characteristics of users and stories on the Freadom App.[1]

In this tutorial, you will analyze the users of the app and their observable characteristics, and look for heterogeneities in the way that the users interact with stories. The users data ew are giving to you are filtered from a bigger user base subject to a minimum activity criteria, so keep in mind that these are relatively 'active' users.

User Observables

First, let's have a look at the user_info dataset, which has user observables.

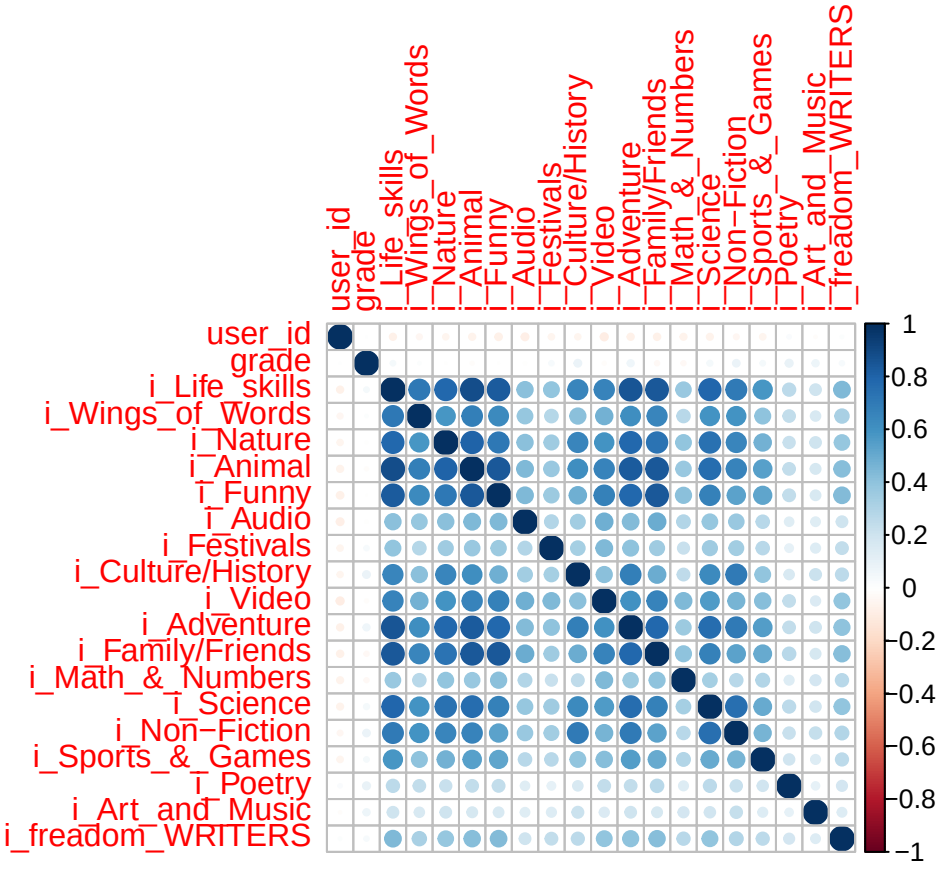
From the Freadom app, we have activity tracking data between 12/23/2019-01/27/2020, and the observable characteristics for these children. The data includes:

The grade: 1 through 7, 1: nursery, 2: junior kindergarden, 3: senior KG, 4-9: grades 1-6 Freadom Point: "Score" of each user on the app (from the quizzes at the end of stories) Total Session Time Story reading time Interest variables: A set of binary variables describing the interests of the child.

```
columns <- c("user_id","grade","i_Life_skills","i_Wings_of_Words","i_Nature","i_Animal",
             "i_Funny","i_Audio","i_Festivals","i_Culture/History","i_Video","i_Adventure",
             "i_Family/Friends","i_Math_&_Numbers","i_Science","i_Non-Fiction",
             "i_Sports_&_Games","i_Poetry","i_Art_and_Music","i_freadom_WRITERS")

userdf_cols <- columns
user_info <- user_df %>% select(all_of(userdf_cols))
```

[1] This tutorial was originally developed by Eray Turkel, Niall Keleher, and Susan Athey for the Spring 2020 course, ALP301 Data-Driven Impact.

```
rm(user_df)
colnames(user_info)<-columns

user_info$grade<-as.integer(str_extract(user_info$grade,"[1-9]"))

#stargazer(as.matrix(user_info),summary=TRUE)

corrplot(cor(user_info))
```



Utility Matrix

We also have a large utility matrix: a user/item matrix representing the level of interaction for every user-story pair. This is tracked user activity from the app between the dates 05/01/2020-01/31/2021. The app has different ways of presenting stories ('story of the day', 'most popular', etc.) so the activity of the users come from different 'pathways'. You can have a look here:

https://play.google.com/store/apps/details?id=com.application.freadom&hl=en_US

For the purposes of these tutorials, we will abstract away from different pathways and aggregate user activity to use the entire dataset.

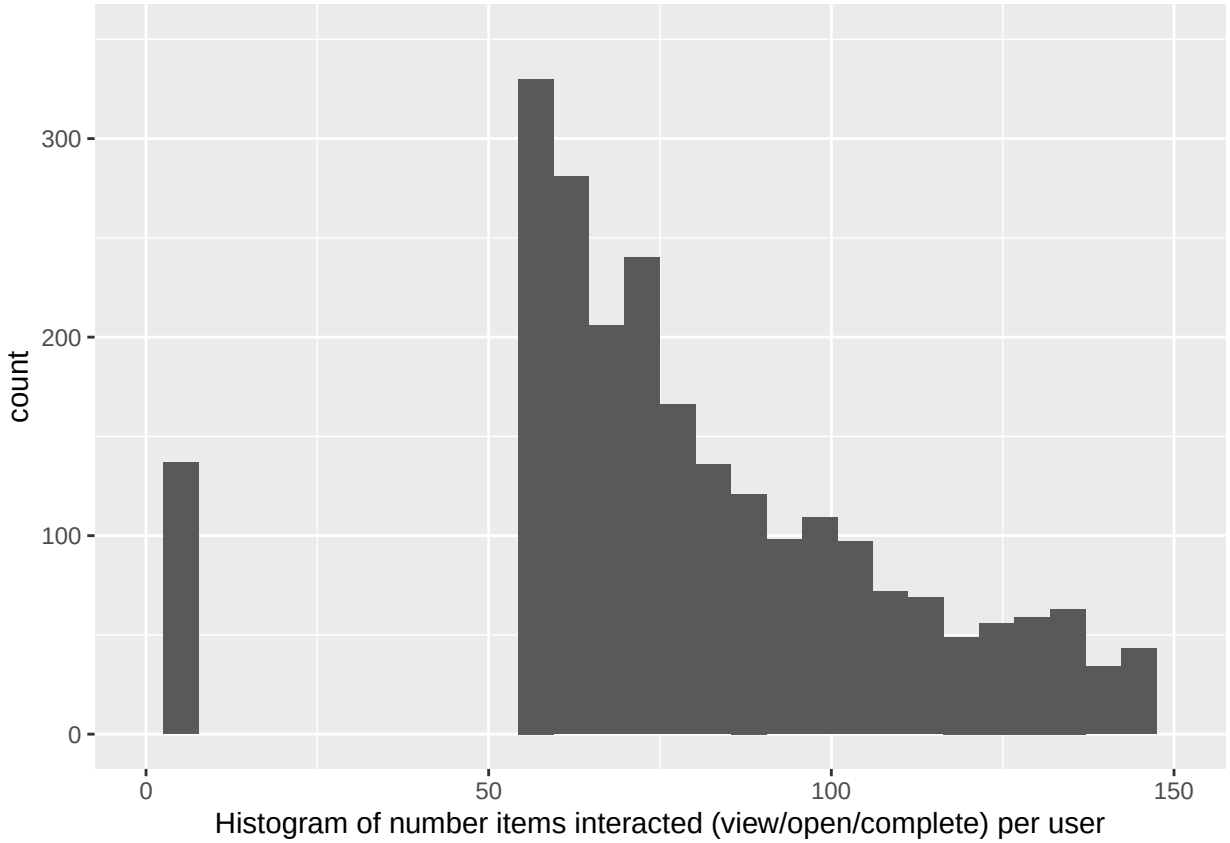
If an entry in this matrix is 0.3, the user "i"" saw story "j"" on the app, if an entry is 0.5, the user "i" opened story "j", and if an entry is 1, user "i" completed story "j". Zeroes represent no interaction. These weights are set by our research team and are by no means set in stone. You can experiment with different weights but you should be thinking about what the weights represent and how to justify them. Intuitively, these weights

represent the relative value of different types of interactions for the organization, and the assumptions we make on user utility to justify these relative values. Here, opening a story is worth one half of the utility for completing a story, and viewing a story is worth one third.

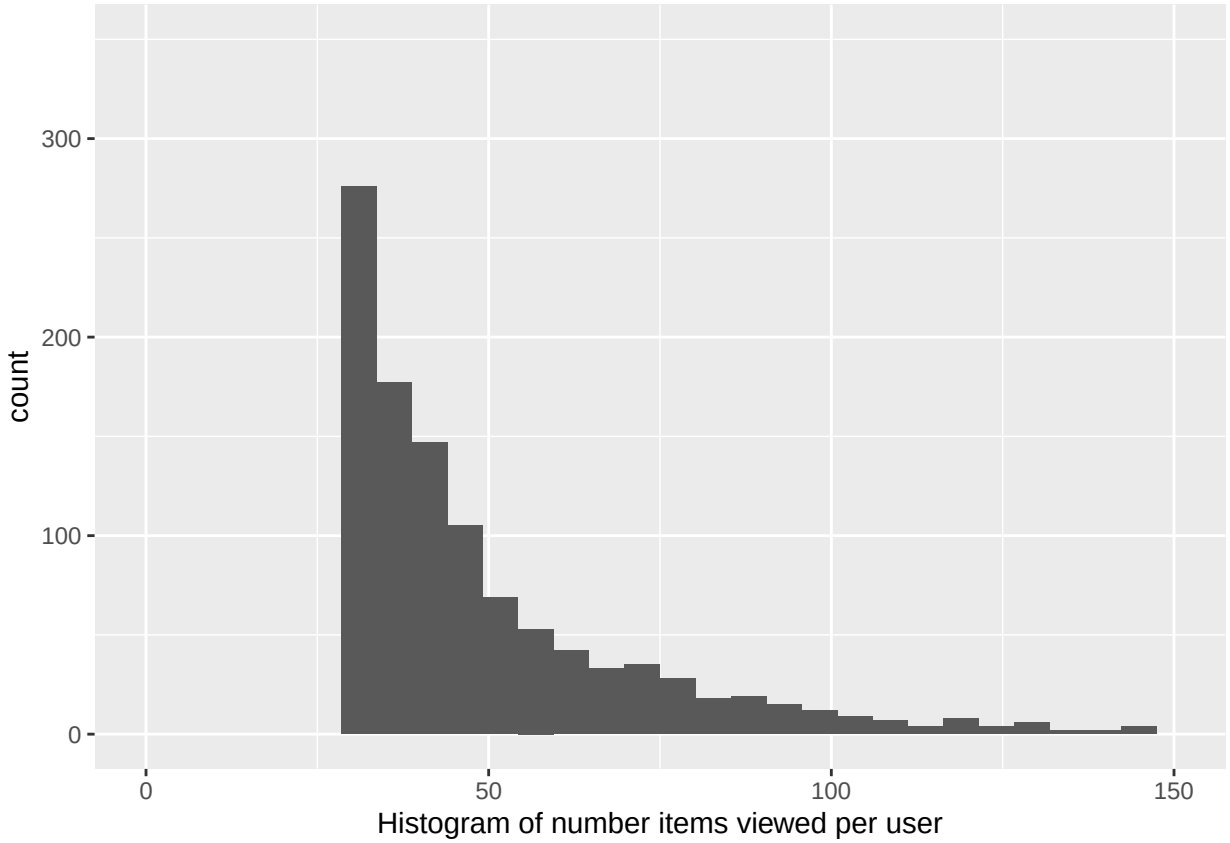
For the 2021 data, in total, we have 92482 users and 2529 stories; after filtering out users with fewer than 60 interactions, we are left with 11202 users in the utility matrix.

Let us check the average activity levels : how many unique stories do users complete, on average?

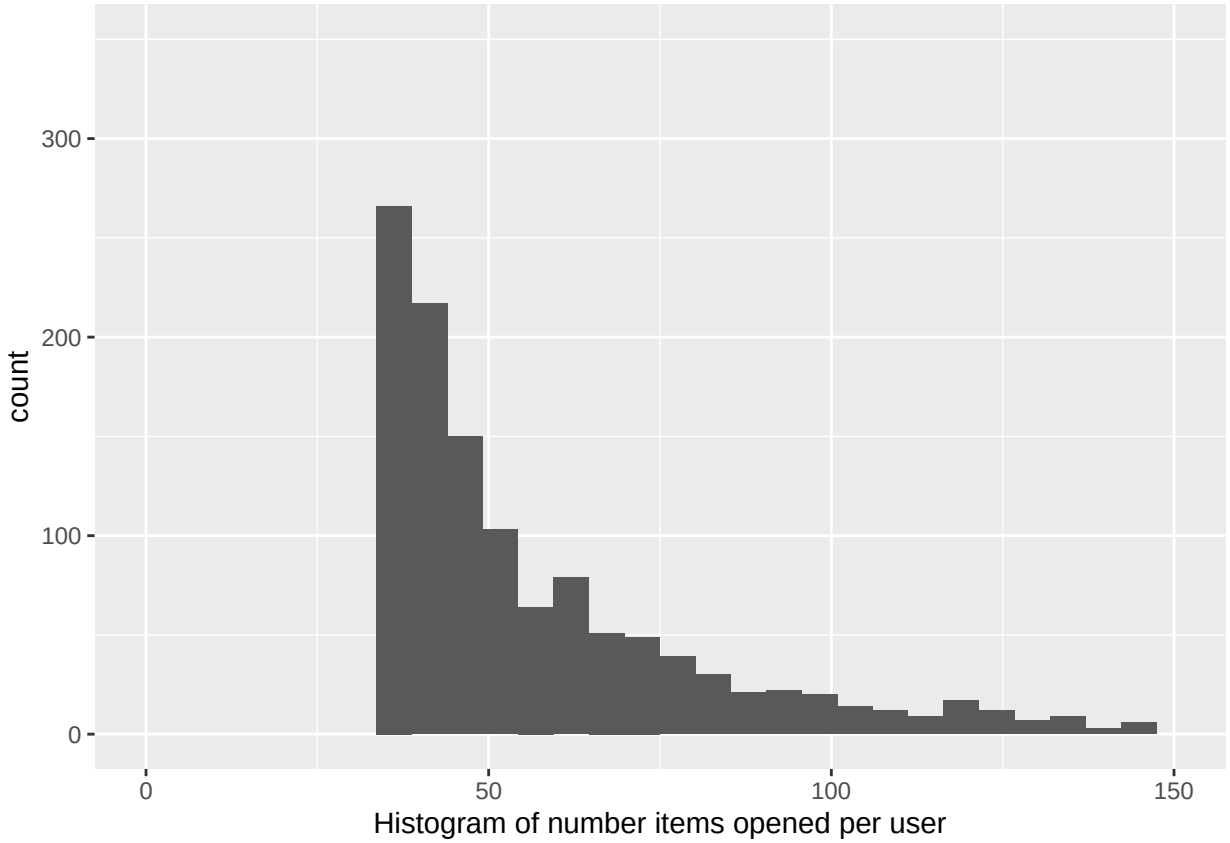
```
util_file <- "data/utils_mat_filtered.csv"
# util_file <- "data/utility_matrix_processed.csv"
num_cols <- 2527  # 1211 for old data, 2527 for new data

utility_mat<- read_csv(util_file,col_types = cols(.default = col_double()))
child_ids<-as.integer(utility_mat$user_id)
story_ids<-as.integer(colnames(utility_mat[,3:num_cols]))
utility_mat[is.na(utility_mat)]<-0.0

utility_matrix<-as.matrix(utility_mat[,3:num_cols])

items_interacted<-rowSums(utility_mat[,3:num_cols]>0)
items_viewed<-rowSums(utility_mat[,3:num_cols]==0.3)
items_opened<-rowSums(utility_mat[,3:num_cols]==0.5)
items_completed<-rowSums(utility_mat[,3:num_cols]==1)

ggplot(data=data.frame(items_interacted),aes(x=items_interacted))+
  geom_histogram()+
  xlab("Histogram of number items interacted (view/open/complete) per user")+
  xlim(c(0,150))+ylim(c(0,350))
```



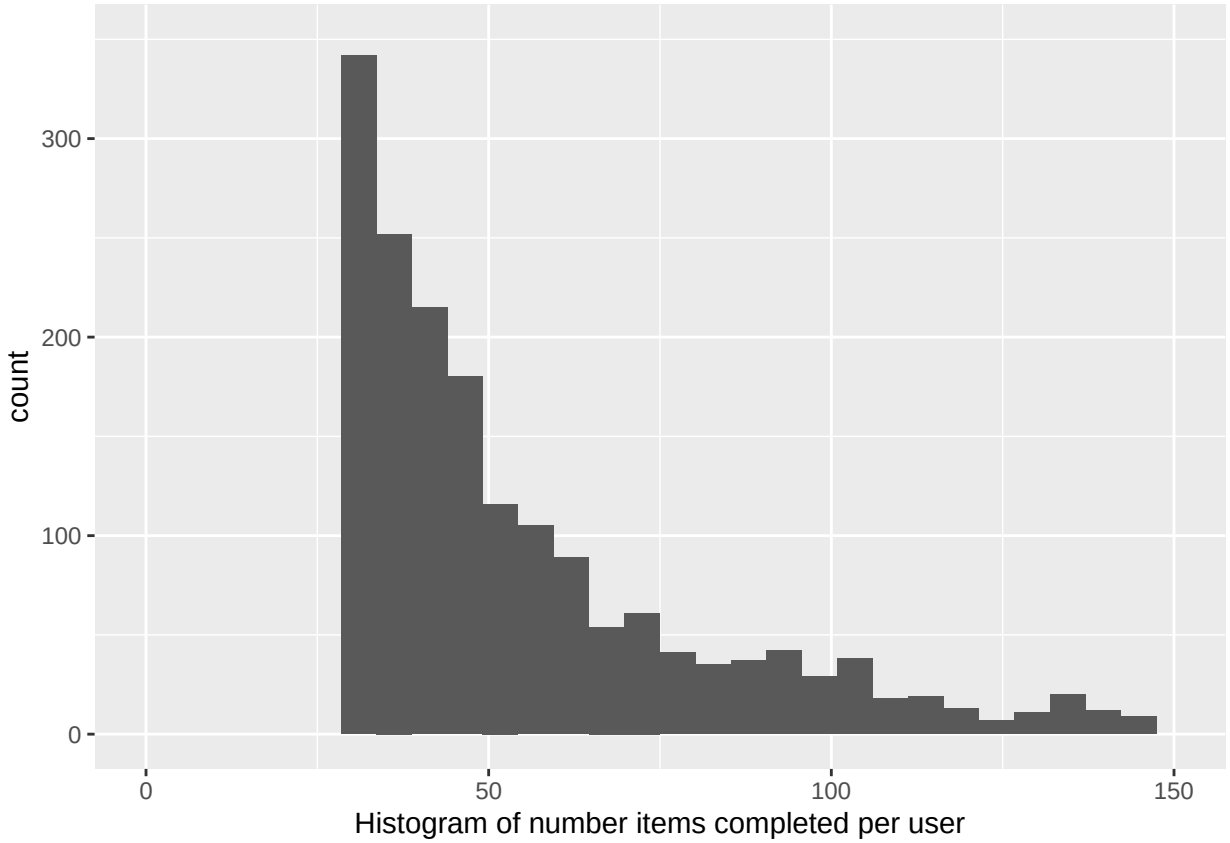

```
ggplot(data=data.frame(items_viewed),aes(x=items_viewed))+
  geom_histogram()+
  xlab("Histogram of number items viewed per user")+
  xlim(c(0,150))+ylim(c(0,350))
```




```
ggplot(data=data.frame(items_opened),aes(x=items_opened))+
  geom_histogram()+
  xlab("Histogram of number items opened per user")+
  xlim(c(0,150))+ylim(c(0,350))
```

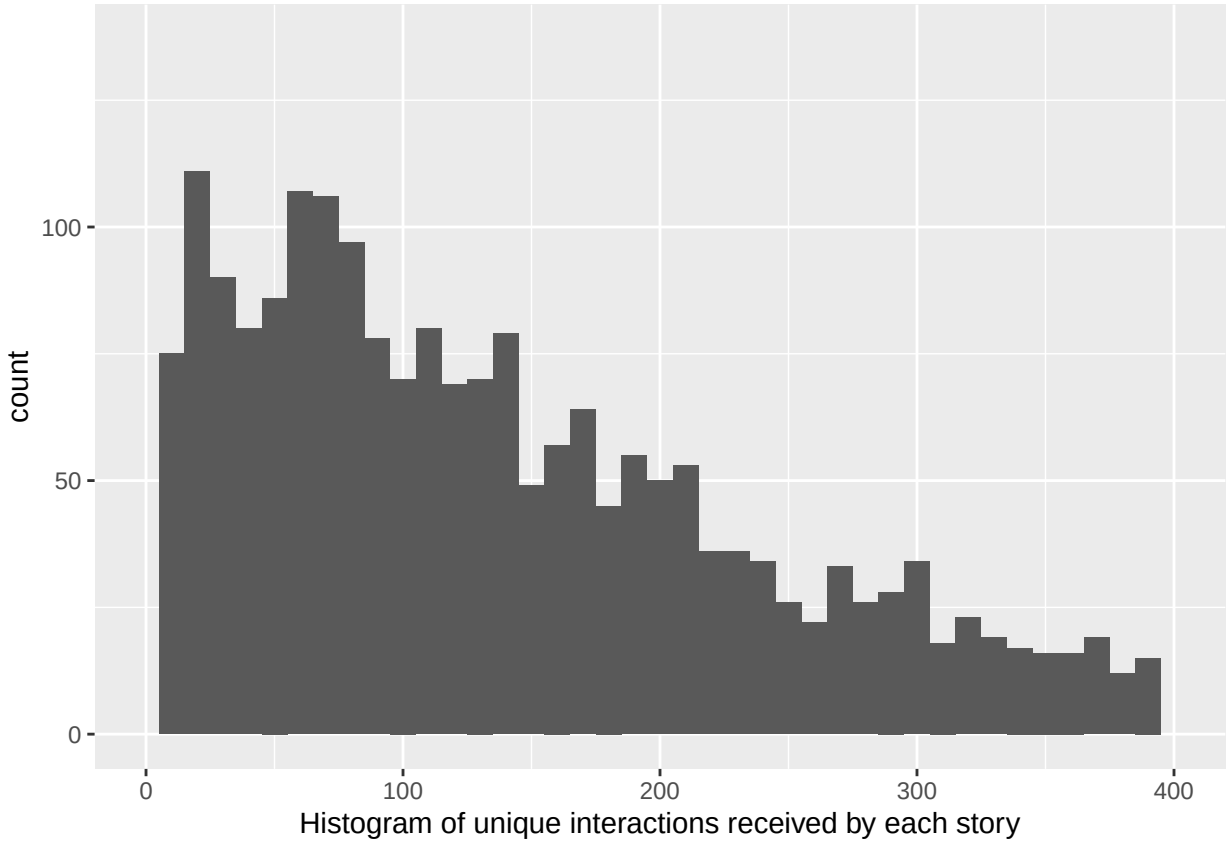



```
ggplot(data=data.frame(items_completed),aes(x=items_completed))+
  geom_histogram()+
  xlab("Histogram of number items completed per user")+
  xlim(c(0,150))+ylim(c(0,350))
```

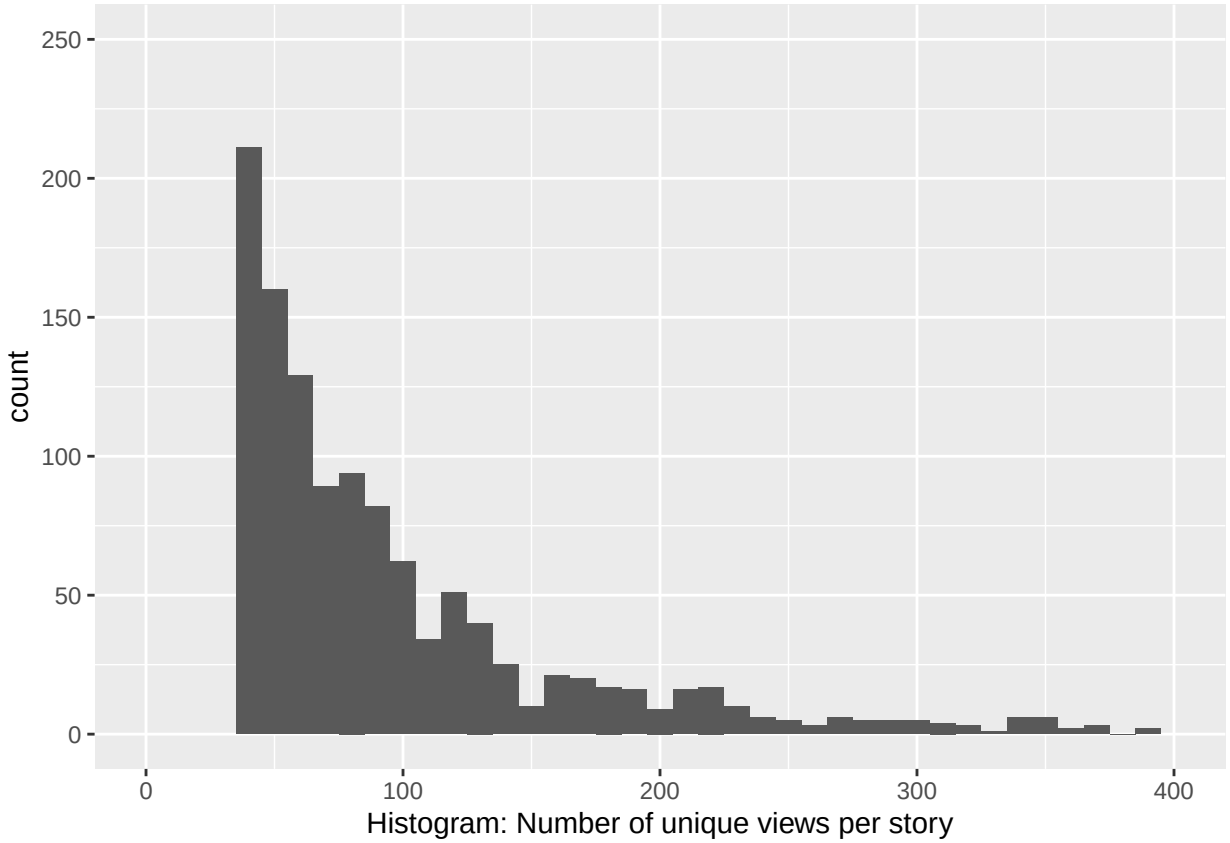


Now let's have a look at what the popularity of stories look like. How many unique interactions does a story receive, on average? How many unique starters and finishers?

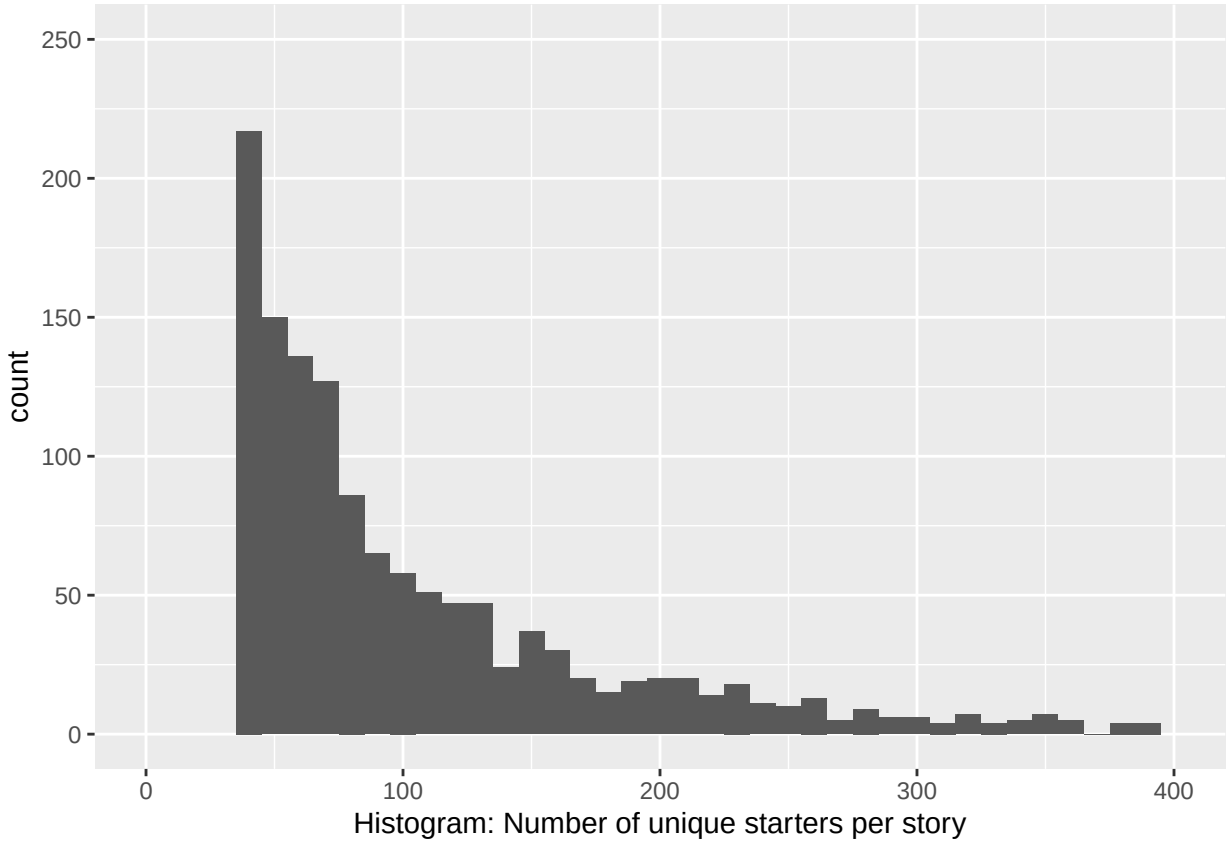
```
item_actions<-colSums(utility_mat[,3:num_cols]>0)
ggplot(data=data.frame(item_actions),aes(x=item_actions))+
  geom_histogram(binwidth = 10)+
  xlim(c(0,400))+
  xlab("Histogram of unique interactions received by each story")
```



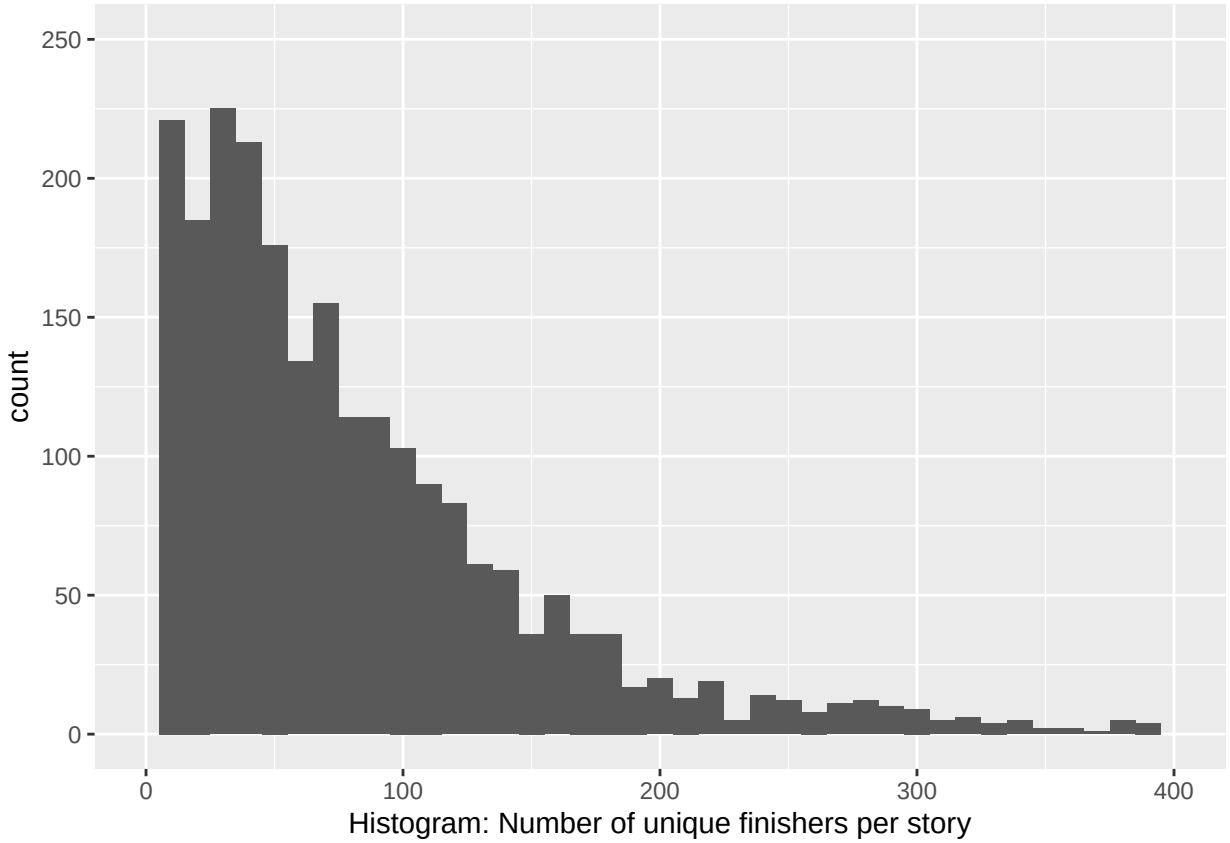

```
item_views<-colSums(utility_mat[,3:num_cols]==0.3)
ggplot(data=data.frame(item_views),aes(x=item_views))+
  geom_histogram(binwidth = 10)+
  xlim(c(0,400))+ylim(c(0,250))+
  xlab("Histogram: Number of unique views per story")+
  ylim(c(0,250))
```




```
item_opens<-colSums(utility_mat[,3:num_cols]==0.5)
ggplot(data=data.frame(item_opens),aes(x=item_opens))+
  geom_histogram(binwidth = 10)+
  xlim(c(0,400))+ylim(c(0,250))+
  xlab("Histogram: Number of unique starters per story")
```




```
item_completes<-colSums(utility_mat[,3:num_cols]==1)
ggplot(data=data.frame(item_completes),aes(x=item_completes))+
  geom_histogram(binwidth = 10)+
  xlim(c(0,400))+ylim(c(0,250))+
  xlab("Histogram: Number of unique finishers per story")
```



Now let's examine the behavior of the users based on their grade level. Filter the utility matrix to keep only the users we have characteristics for.

```
user_info%>%filter(grade>5)->old_users
user_info%>%filter(grade<3)->young_users
```

Questions

How are the activities of old and young users different?

Below, we plot the histogram of the number of interactions with stories per user for young vs old users. We define an interaction as one of these three things: a view on the app, opening a story, or finishing a story.

Can you find any other interesting heterogeneities in the activities of young versus old users?

-Here's a question to get you started: Are older users more likely to complete the stories they started?

-Are there any other meaningful differences between any other subgroups?

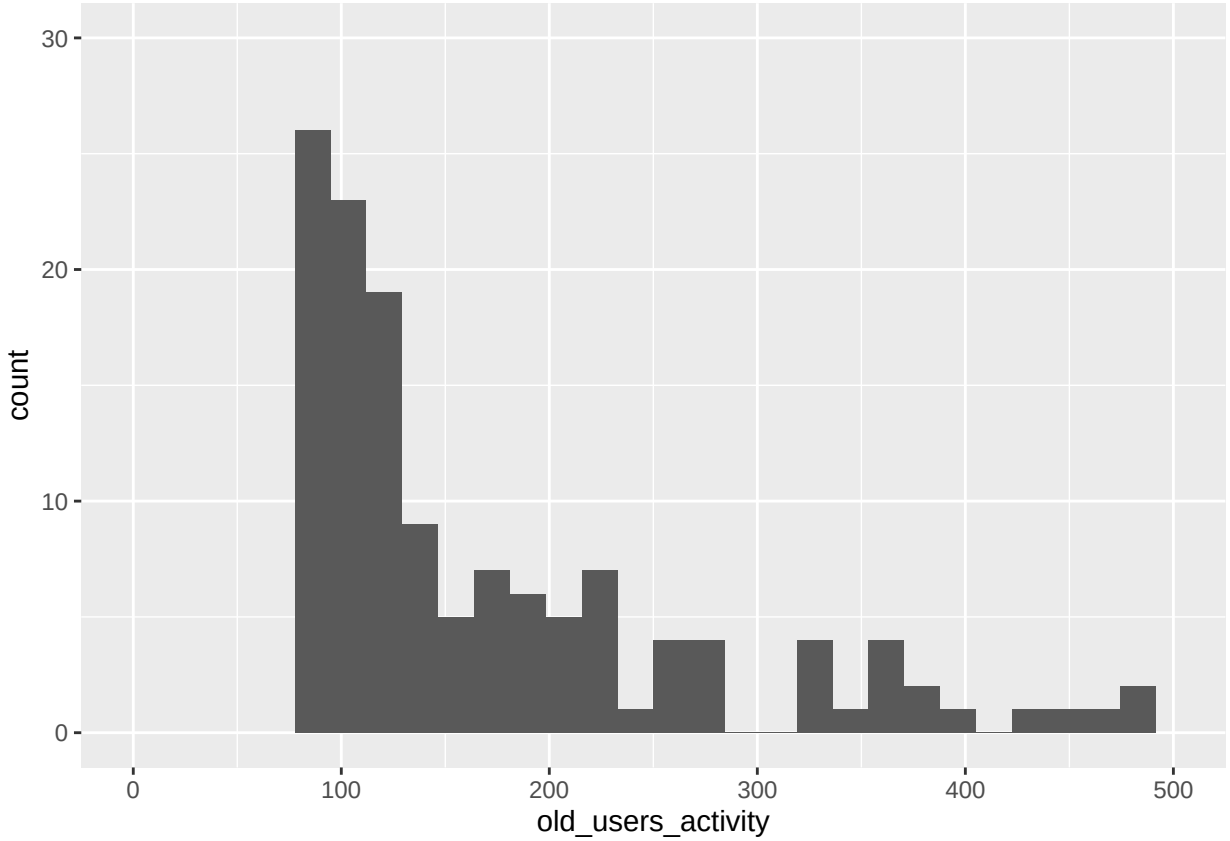
```
user_id <- utility_mat$child_id_code

utility_mat%>%filter(user_id %in% old_users$user_id)->old_users_matrix
utility_mat%>%filter(user_id %in% young_users$user_id)->young_users_matrix

old_users_matrix<-as.matrix(old_users_matrix[,3:num_cols])
young_users_matrix<-as.matrix(young_users_matrix[,3:num_cols])
```

```
old_users_activity<-rowSums(old_users_matrix>0)
young_users_activity<-rowSums(young_users_matrix>0)

ggplot(data=data.frame(old_users_activity),aes(x=old_users_activity))+
  geom_histogram()+
  xlim(c(0,500))+ylim(c(0,30))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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

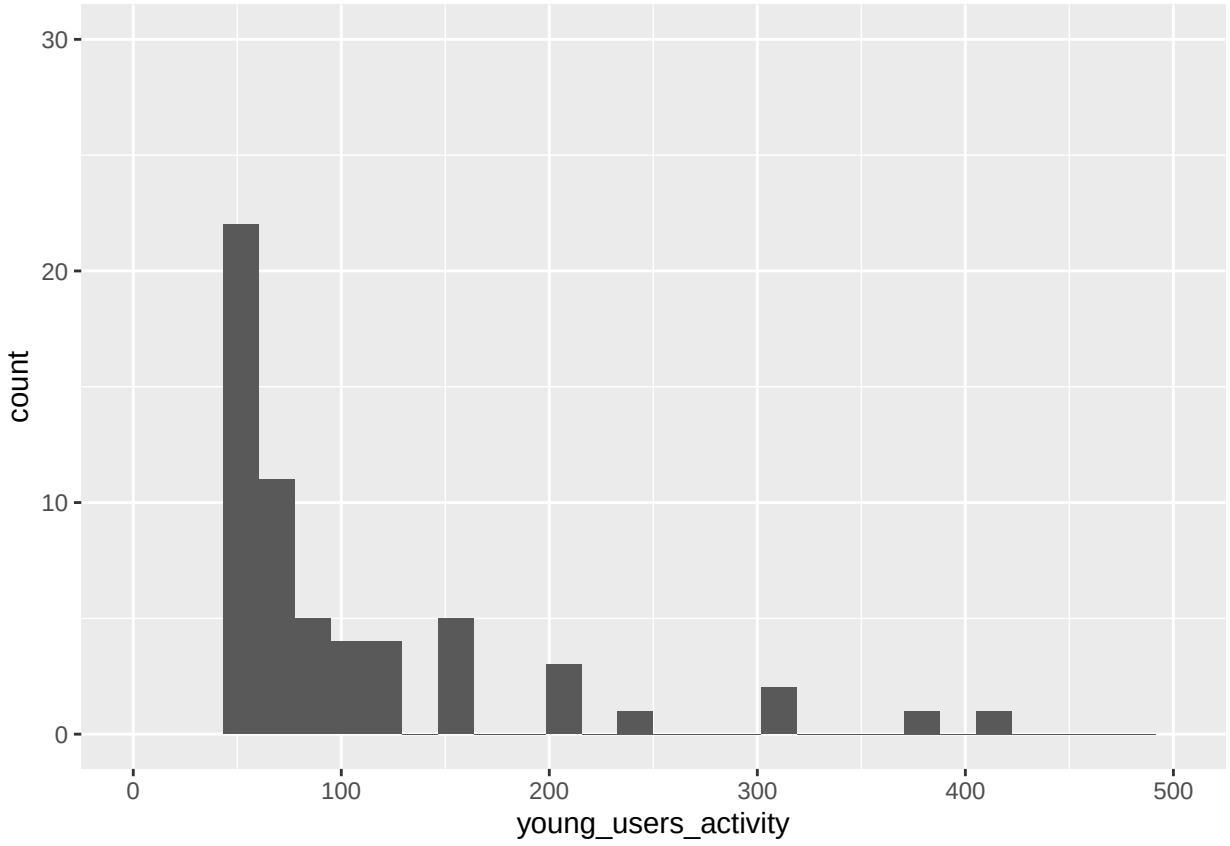
```
## Warning: Removed 6 rows containing missing values (geom_bar).
```




```
ggplot(data=data.frame(young_users_activity),aes(x=young_users_activity))+
  geom_histogram()+
  xlim(c(0,500))+ylim(c(0,30))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```


30
20
10
0
count
0
100
200
300
400
500
young_users_activity