Exploratory Data Analysis

Lissa Harrop, Katrina Watkins, Ricky Loo and Max Tan

16 August 2022

Contents

1 Exploratory Data Analysis of our overall data

We (group 10) decided to use the Dating App User Profiles' stats data set. The data set is available on kaggle <> and the license to use the data ser is available on creativecommons <>.

(https://www.kaggle.com/datasets/jmmvutu/dating-app-lovoo-user-profiles?resource=download). The license to use the data set is available at https://creativecommons.org/licenses/by/4.0/.

After some basic exploration of the variables available in the Lovoo v3 data set we decided to explore the variables age, counts_pictures, counts_profileVisits, counts_kisses, distance, country and isVip.

Age is the users age, **counts_pictures** is the number of pictures on the user's profile, **counts_profileVisits** is the number of clicks on this user (to see his/her full profile) from other user accounts, **counts_kisses** is the number of unique user accounts that "liked" (called "kiss" on the platform) this user account, **distance** is the distance between this user's city/location and the location of the user account that was used to fetch the data of this user, **country** is the user's country, **isVip** is a 1 if the user is VIP. [It was possible to buy a VIP status with real money. This status came with benefits.].

It was discovered that there were 46 missing values in the variable distance. These have been replaced by the mean of the distance column, 207.23.After replacing the 46 missing distance variables to ensure we have a full data set, we have a sample size of 3992 for all seven variables.

The ages of the user's of the lovoo app range from 18 years to 28 years with the median age being 22 year. The minimum number of pictures on a user's profile is 0 with the maximum being 30 pictures and the median being 4.The number of clicks on a user's profile to see his/her full profile (from another users account) ranges from 0 to 164425 clicks, with the median being 1222 clicks. The number of unique user accounts that "liked" a users account ranges from 0 to 9288 likes, with the median being 44 likes. The distance between this user's city/location and the location of the user account that was used to fetch the data of the user ranges from 0 to 6918, with the median being 173. These and other summary statistics can be seen in table 1.

The summary of the countries and their counts can be found in table 2 and a visualisation can be seen in figure 3. There are 32 different countries with varying numbers of users. Table 3 shows that 3901 users are not Vip's while only 91 are Vip's.

There appears to be strong positive correlation between the number of profiles visits and the number of likes that a user receives. There is also positive correlation between the number of pictures a user has and the number of profile visits they receive, as well as the number of likes the user has. There is slight positive correlation between the age of the user and the distance between this user's city/location and the location of the user account that was used to fetch the data of the user. There appears to be no correlation between age and likes, profile visits and pictures, nor distance and likes, profile visits and pictures. This can been seen in figure 1 and supported by the pairs plots in figure 2.

Table 1: Summary Statistics - Numerical Variables

	age	counts_pictures	counts_profileVisits	counts_kisses	distance
sample size	3992.00	3992.00	3992.00	3992.00	3992.00
minimum	18.00	0.00	0.00	0.00	0.00
first quartile	20.00	2.00	383.00	11.00	85.27
median	22.00	4.00	1222.00	44.00	173.00
third quartile	24.00	6.00	4063.25	141.00	317.00
maximum	28.00	30.00	164425.00	9288.00	6918.00
IQR	4.00	4.00	3680.25	130.00	231.73
standard deviation	1.96	4.42	6845.04	377.65	195.46
mean	21.99	4.79	3705.47	156.60	207.23

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

2 Age, country and pics or distance....

```
## Warning: Ignoring unknown parameters: binwidth
```

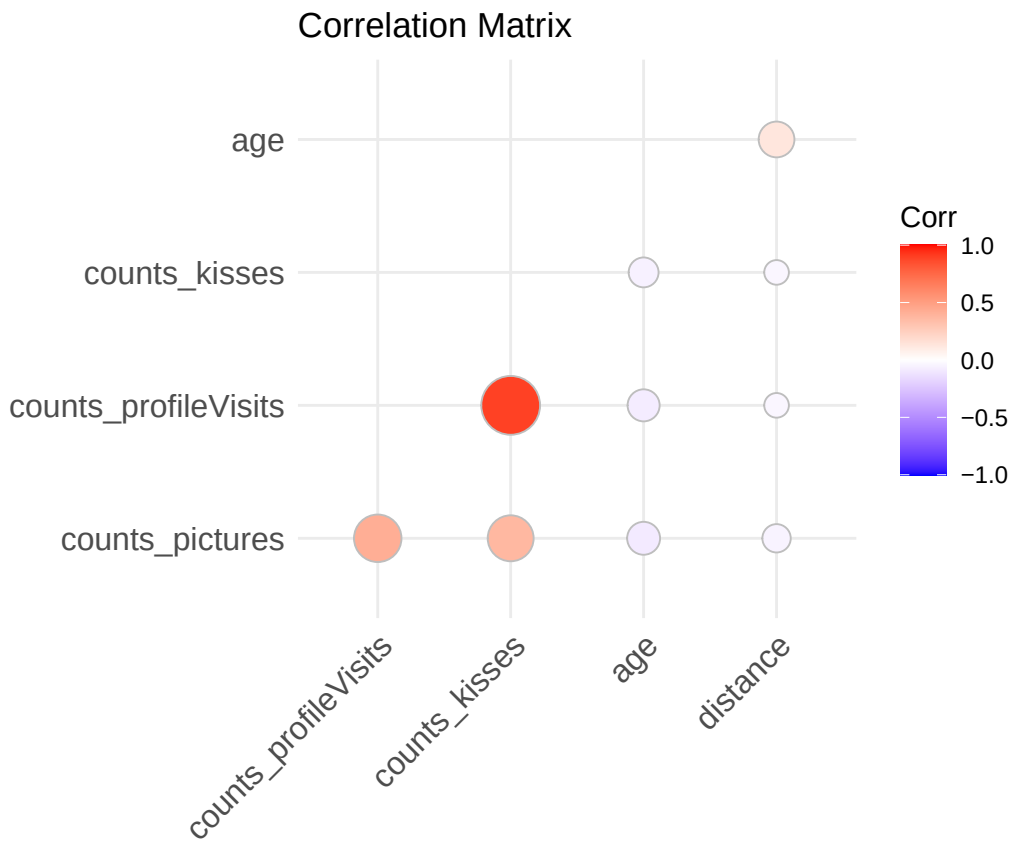


Figure 1: Correlation Matrix

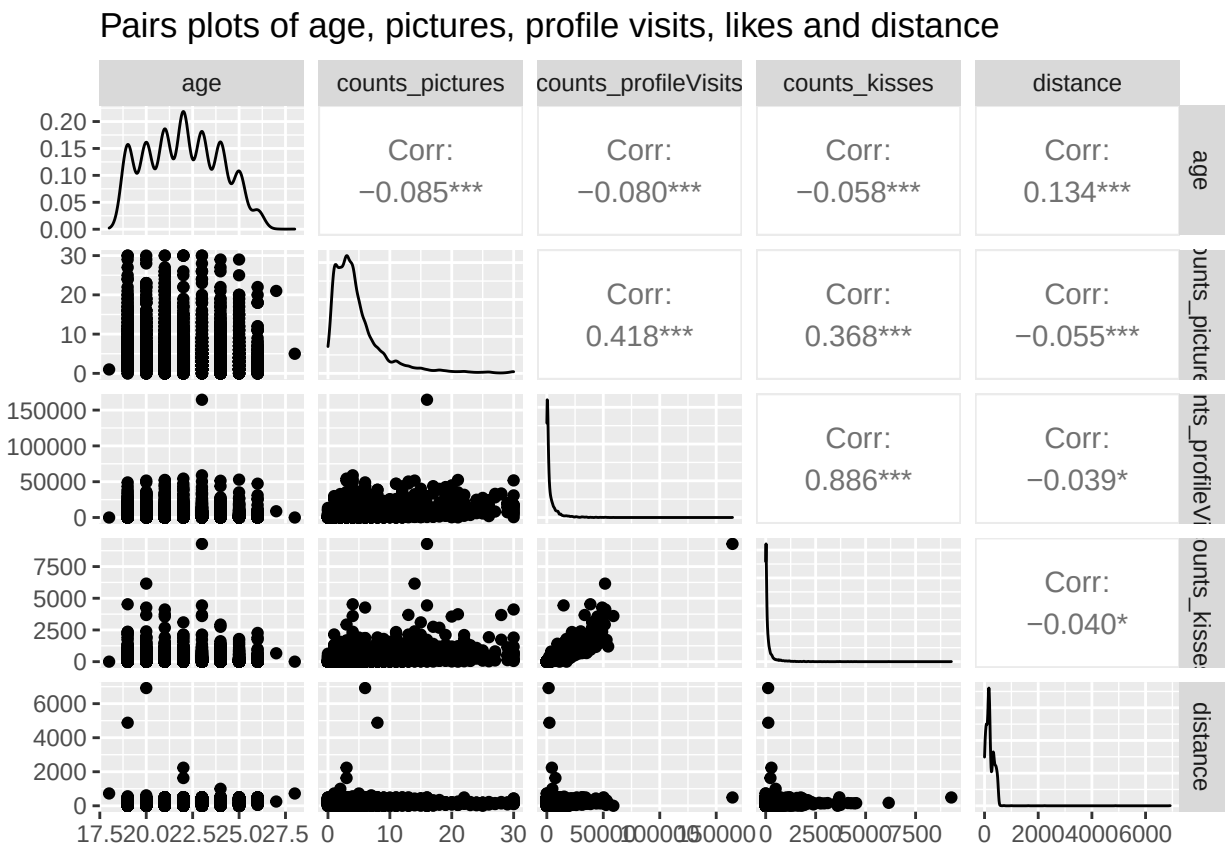


Figure 2: Pairs plot

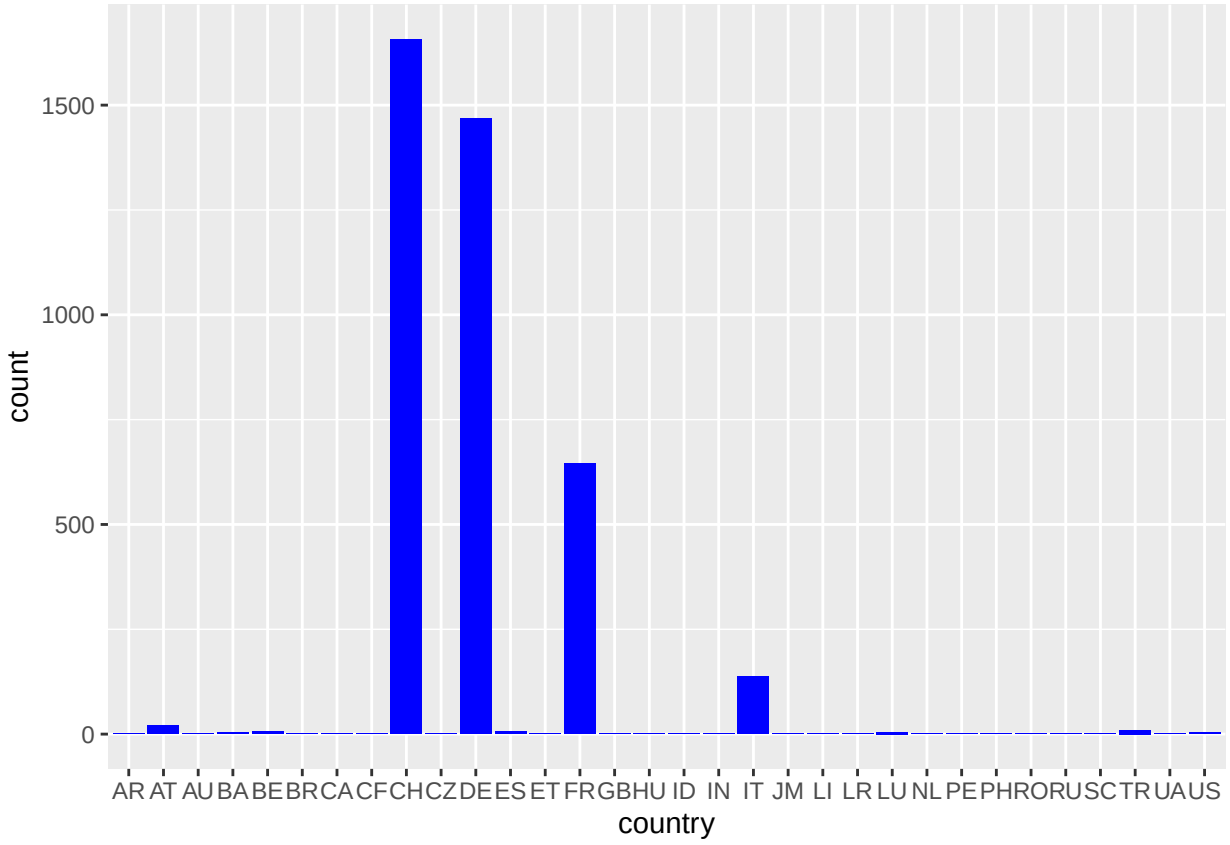
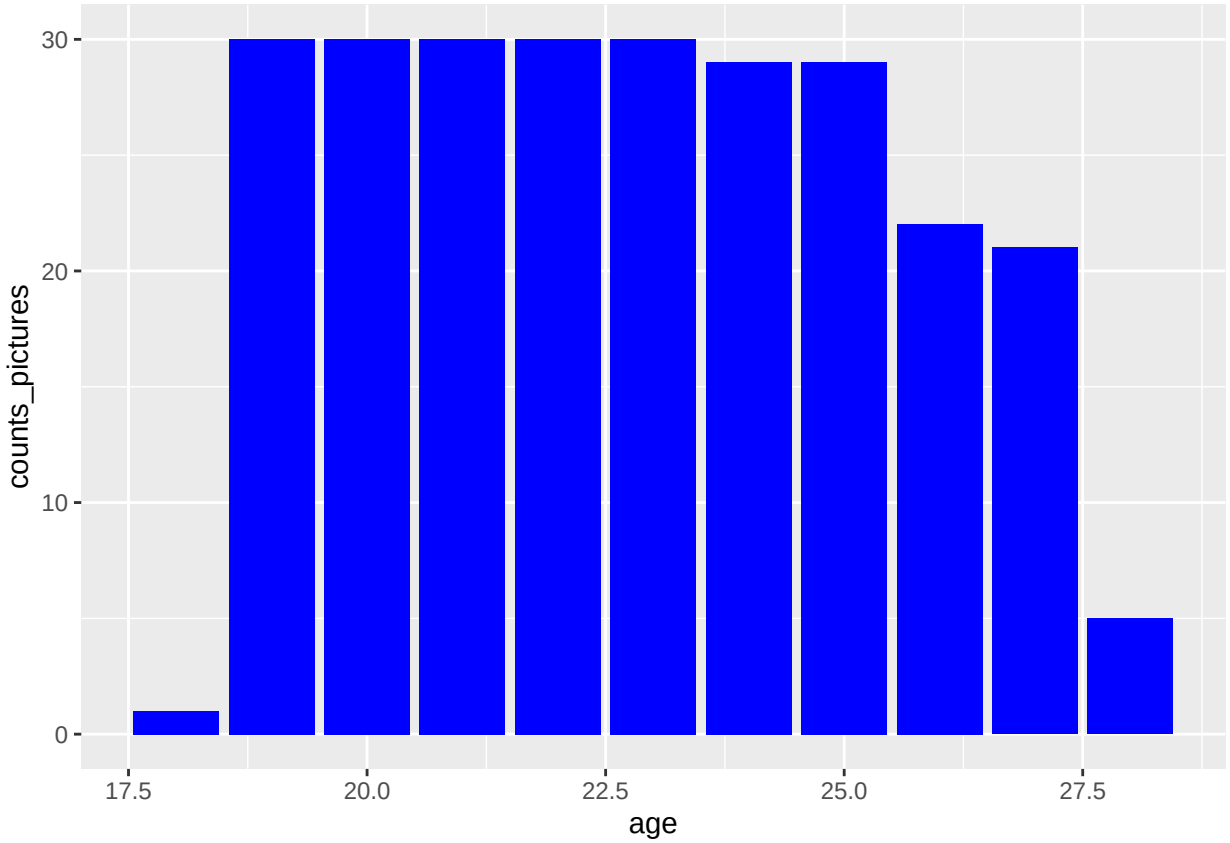


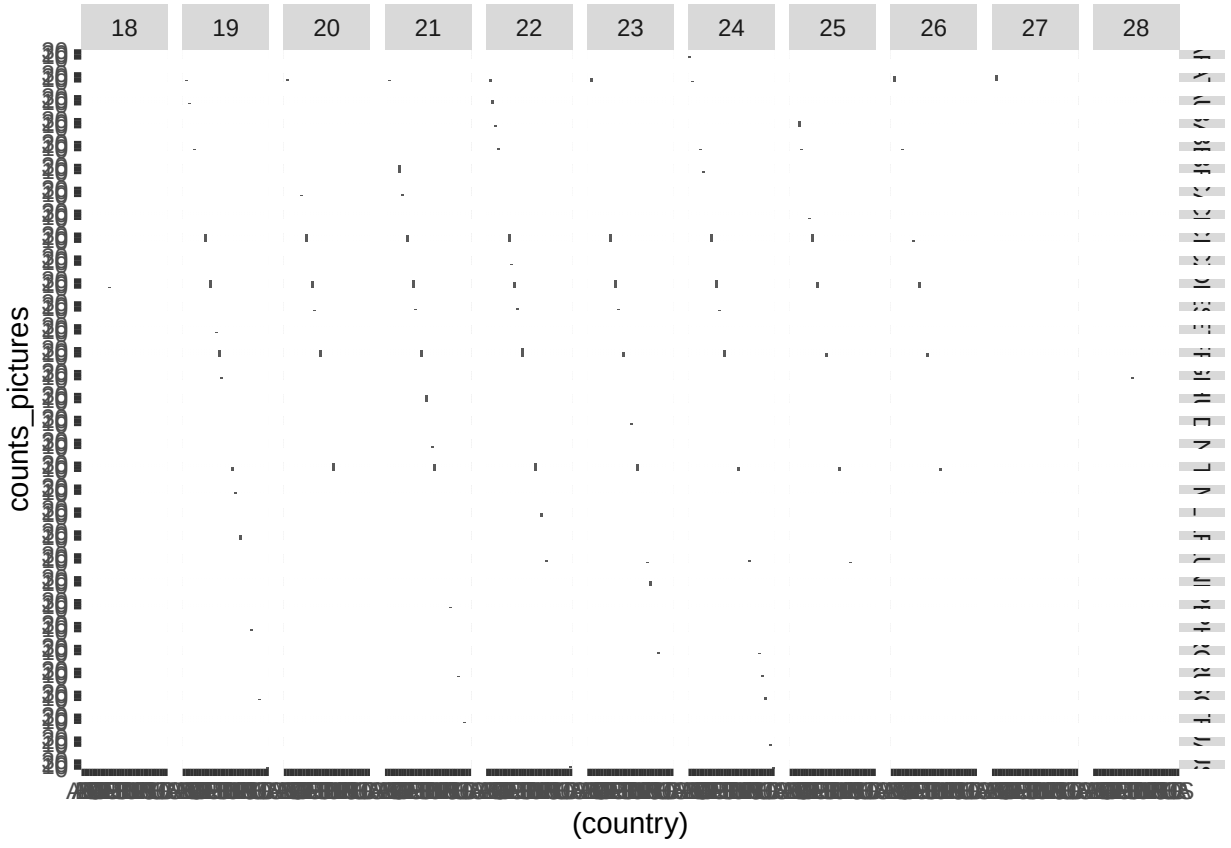
Figure 3: Number of user's by Country

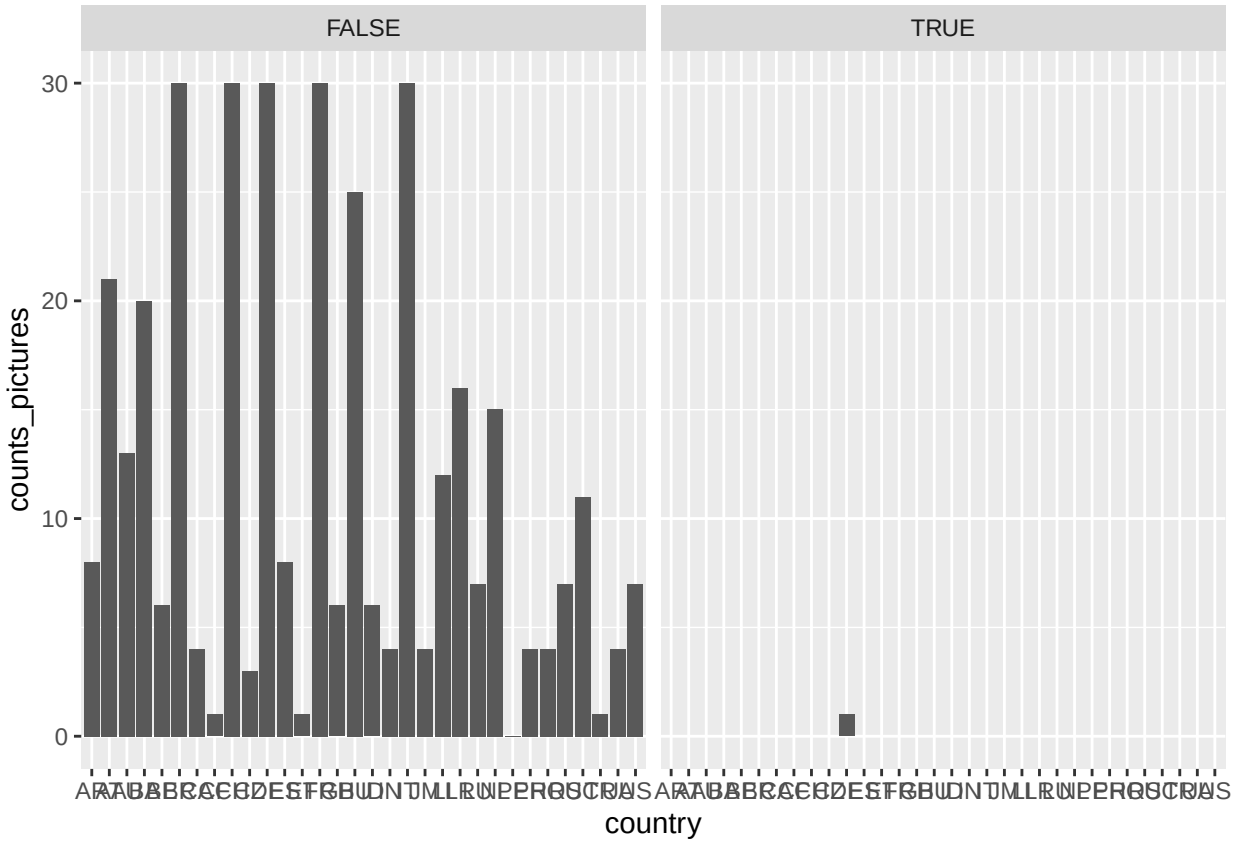


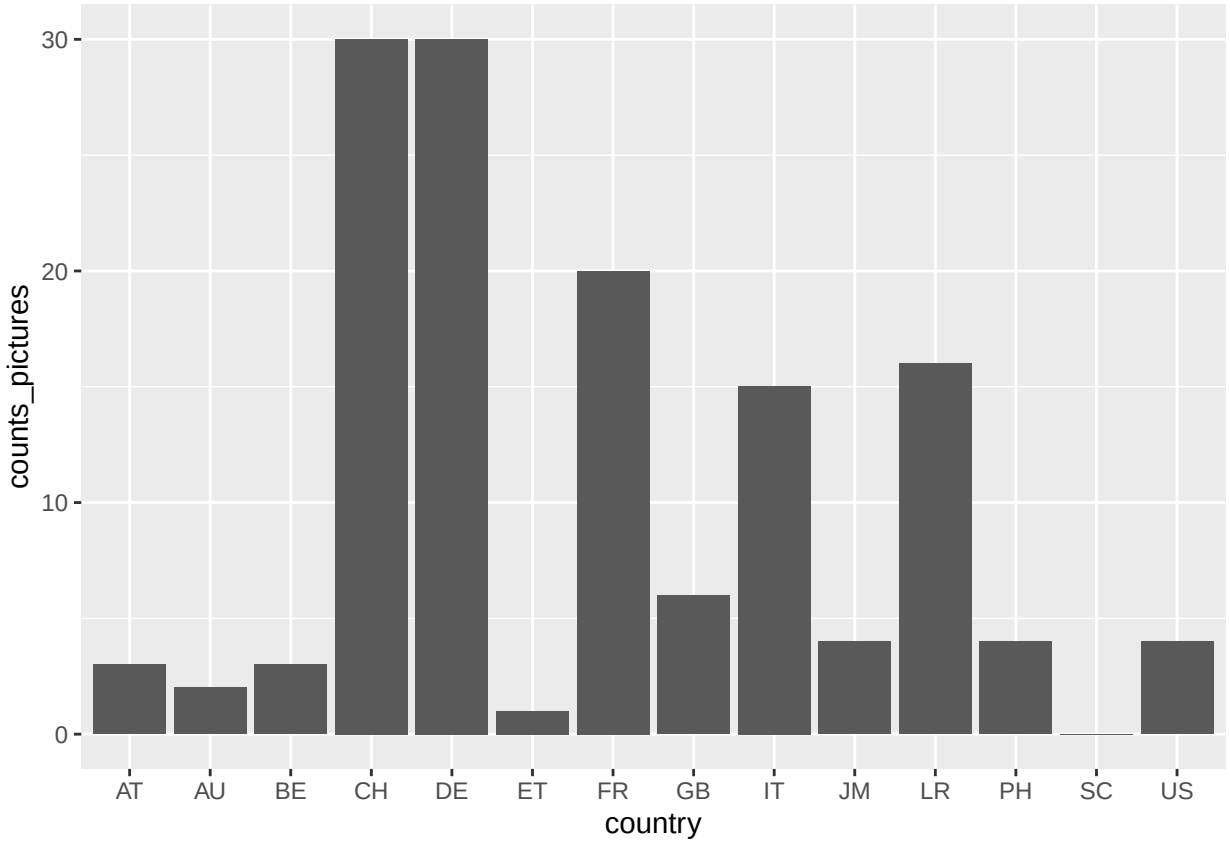

```
table(lovoo$country, lovoo$age) # keep this
```

```
##
##       18  19  20  21  22  23  24  25  26  27  28
##    AR  0   0   0   0   0   0   1   0   0   0   0
##    AT  0   1   1   2   4   4   1   0   6   1   0
##    AU  0   1   0   0   1   0   0   0   0   0   0
##    BA  0   0   0   0   2   0   0   1   0   0   0
##    BE  0   1   0   0   3   0   1   1   1   0   0
##    BR  0   0   0   1   0   0   1   0   0   0   0
##    CA  0   0   1   1   0   0   0   0   0   0   0
##    CF  0   0   0   0   0   0   0   1   0   0   0
##    CH  0 279 264 260 261 242 198 143  10   0   0
##    CZ  0   0   0   0   1   0   0   0   0   0   0
##    DE  1  93 142 212 327 251 221 140  81   0   0
##    ES  0   0   1   1   2   1   1   0   0   0   0
##    ET  0   1   0   0   0   0   0   0   0   0   0
##    FR  0 122 107 102  97  64  86  55  13   0   0
##    GB  0   1   0   0   0   0   0   0   0   0   1
##    HU  0   0   0   1   0   0   0   0   0   0   0
##    ID  0   0   0   0   0   1   0   0   0   0   0
```

```
##    IN   0   0   0   1   0   0   0   0   0   0   0
##    IT   0  19  14  19  20  28  18  14   6   0   0
##    JM   0   1   0   0   0   0   0   0   0   0   0
##    LI   0   0   0   0   1   0   0   0   0   0   0
##    LR   0   1   0   0   0   0   0   0   0   0   0
##    LU   0   0   0   0   1   2   1   1   0   0   0
##    NL   0   0   0   0   0   2   0   0   0   0   0
##    PE   0   0   0   1   0   0   0   0   0   0   0
##    PH   0   1   0   0   0   0   0   0   0   0   0
##    RO   0   0   0   0   0   1   1   0   0   0   0
##    RU   0   0   0   1   0   0   1   0   0   0   0
##    SC   0   1   0   0   0   0   1   0   0   0   0
##    TR   0   0   0  10   0   0   0   0   0   0   0
##    UA   0   0   0   0   0   0   1   0   0   0   0
##    US   0   1   0   0   1   0   1   0   0   0   0
```




FALSE
TRUE
30
20
10
0
counts_pictures
country


30
20
10
0
counts_pictures
AT
AU
BE
CH
DE
ET
FR
GB
IT
JM
LR
PH
SC
US
country

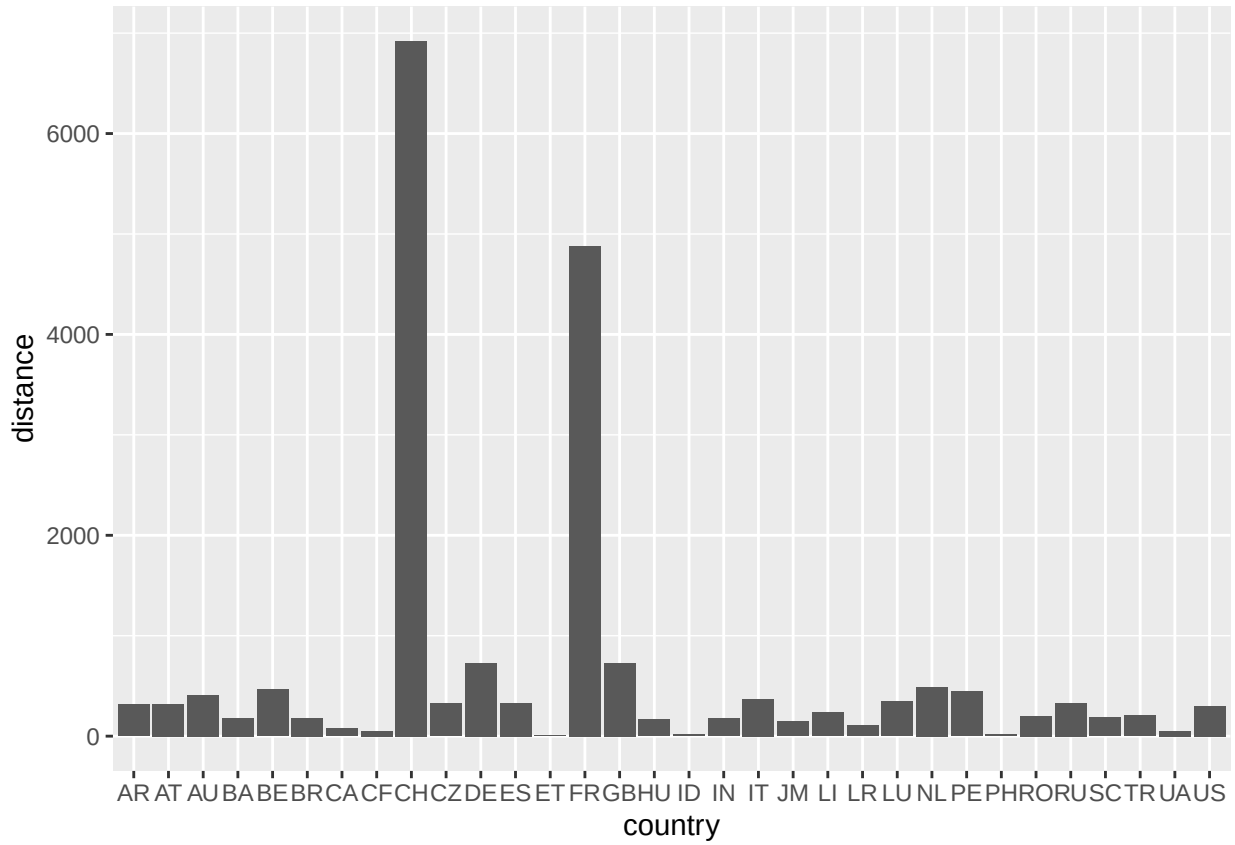

distance
6000
4000
2000
0
AR AT AU BA BE BR CA CF CH CZ DE ES ET FR GB HU ID IN IT JM LI LR LU NL PE PH RO RU SC TR UA US
country

Table 2: Summary Statistics - Countires

	Country	Count
1	AR	1
2	AT	20
3	AU	2
4	BA	3
5	BE	7
6	BR	2
7	CA	2
8	CF	1
9	CH	1657
10	CZ	1
11	DE	1468
12	ES	6
13	ET	1
14	FR	646
15	GB	2
16	HU	1
17	ID	1
18	IN	1
19	IT	138
20	JM	1
21	LI	1
22	LR	1
23	LU	5
24	NL	2
25	PE	1
26	PH	1
27	RO	2
28	RU	2
29	SC	2
30	TR	10
31	UA	1
32	US	3

Table 3: Summary Statistics - isVip

	isVip	Count
No	0	3901
Yes	1	91