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 (Mabilama (2020)) and the license to use the data ser is available on creativecommons (Commons (n.d.)).

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 These have been replaced by the mean of the distance column, 207.23.After replacing the 46 missing distance variables to ensure we have a full data set, we have a sample size of 3992 for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

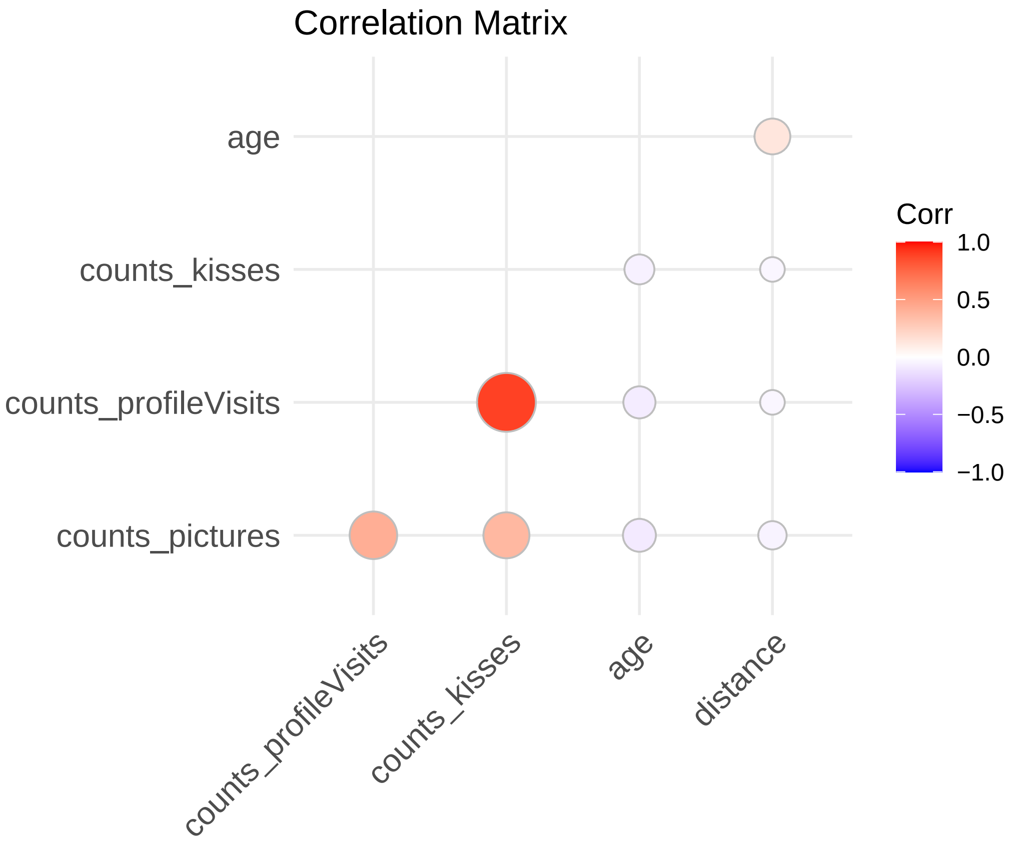


Figure 1: Correlation Matrix

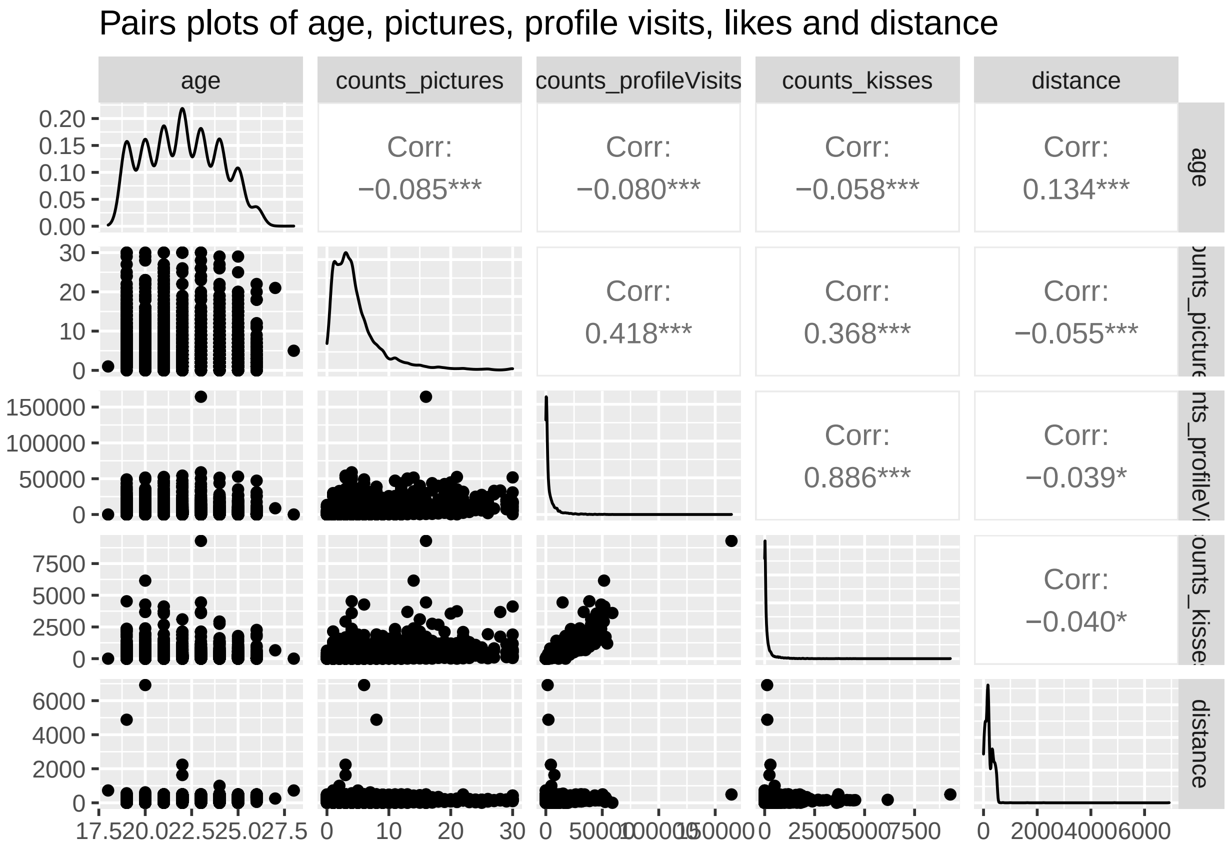


Figure 2: Pairs plot

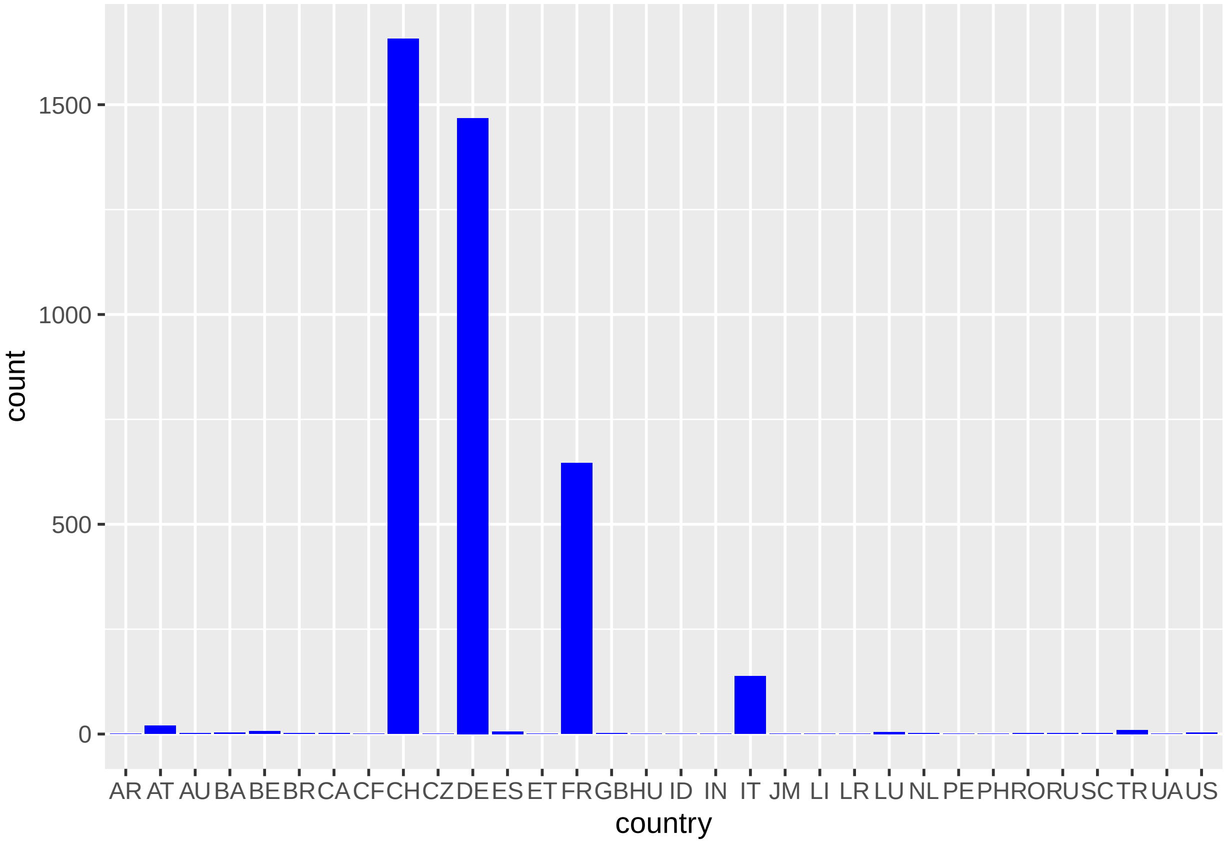


Figure 3: Number of user's by Country

2 Counts_profileVisits, counts_kisses and isVip

How does profile visits affect profile likes?

We see that there is a high correlation (0.89) between profile visits and profile likes (figure 4). We see from figure 5 that there is a positive relationship between profile visits and profile likes. We can also see there are several outliers that we should remove as they can be very influential to our data. After removing the outliers from figure 5 we see the same positive relationship but the outlieing profiles are gone (figure 6).

Does having "VIP" mean you get more profile visits and likes?

The distribution of profile visits and profile likes is right skewed and definitely does not follow a normal distribution as seen in figure 7. The density graphs (figure **??**) for both profile visits and profile likes is right skewed. This means that for both profile likes and visits, the mean is greater than the median. The mean of the boxplot for profiles with "VIP" is less than the mean of the boxplot for profiles without "VIP", this means on average profiles with "VIP" get less profile visits and likes than profiles without "VIP" as shown in figure 8

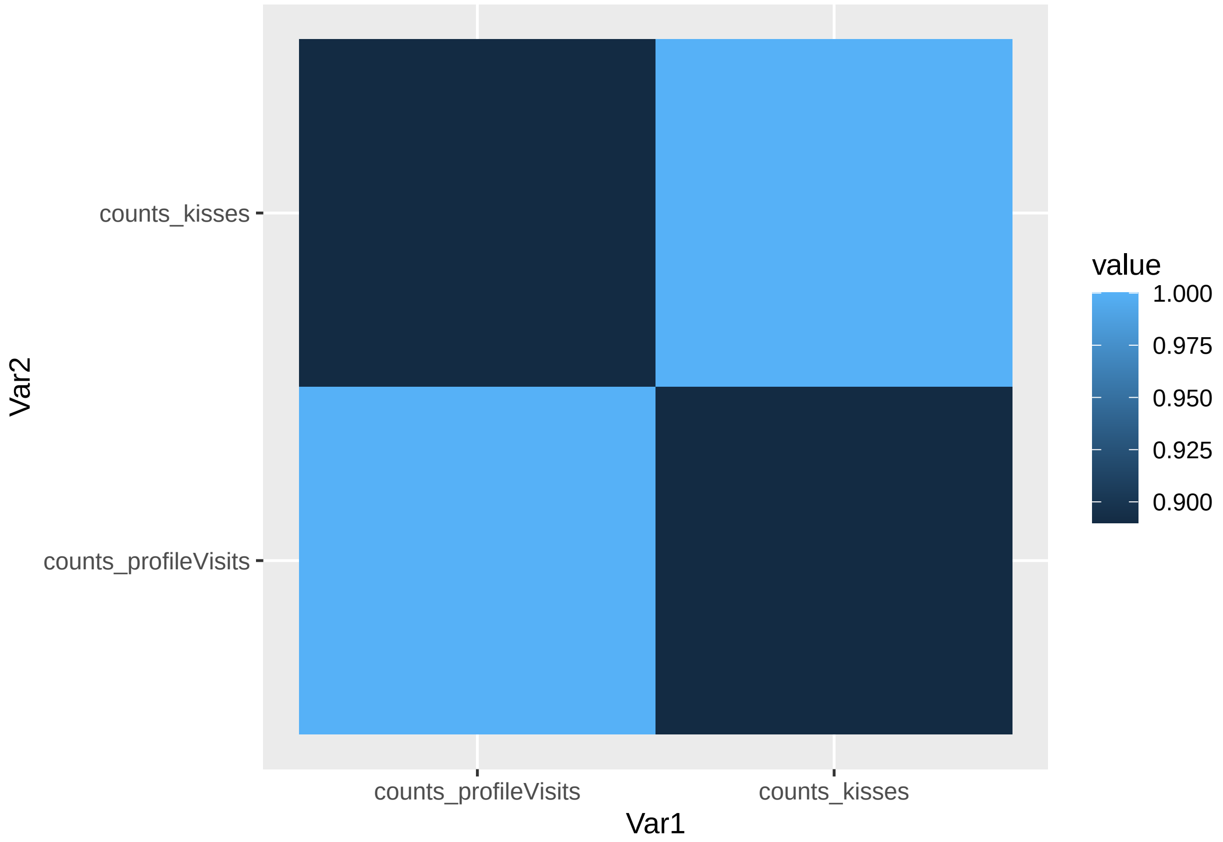


Figure 4: Correlation of profile visits and kisses

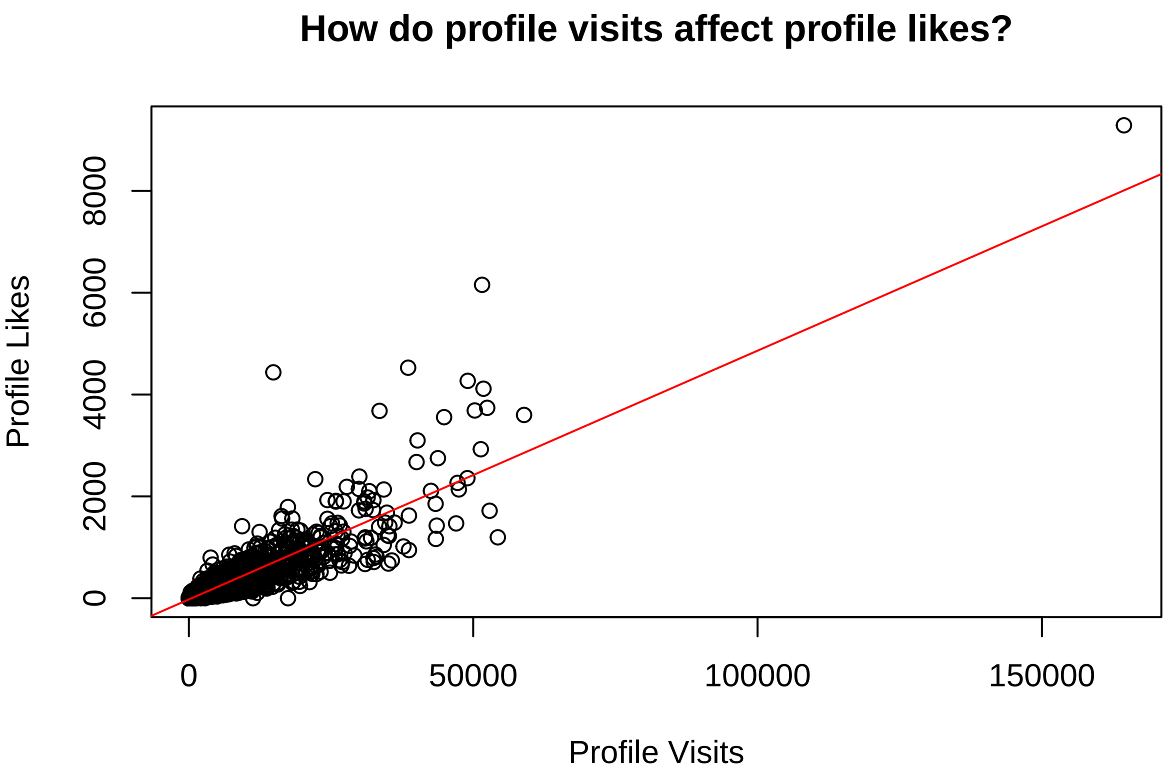


Figure 5: Scatterplot of profile visits vs profile likes

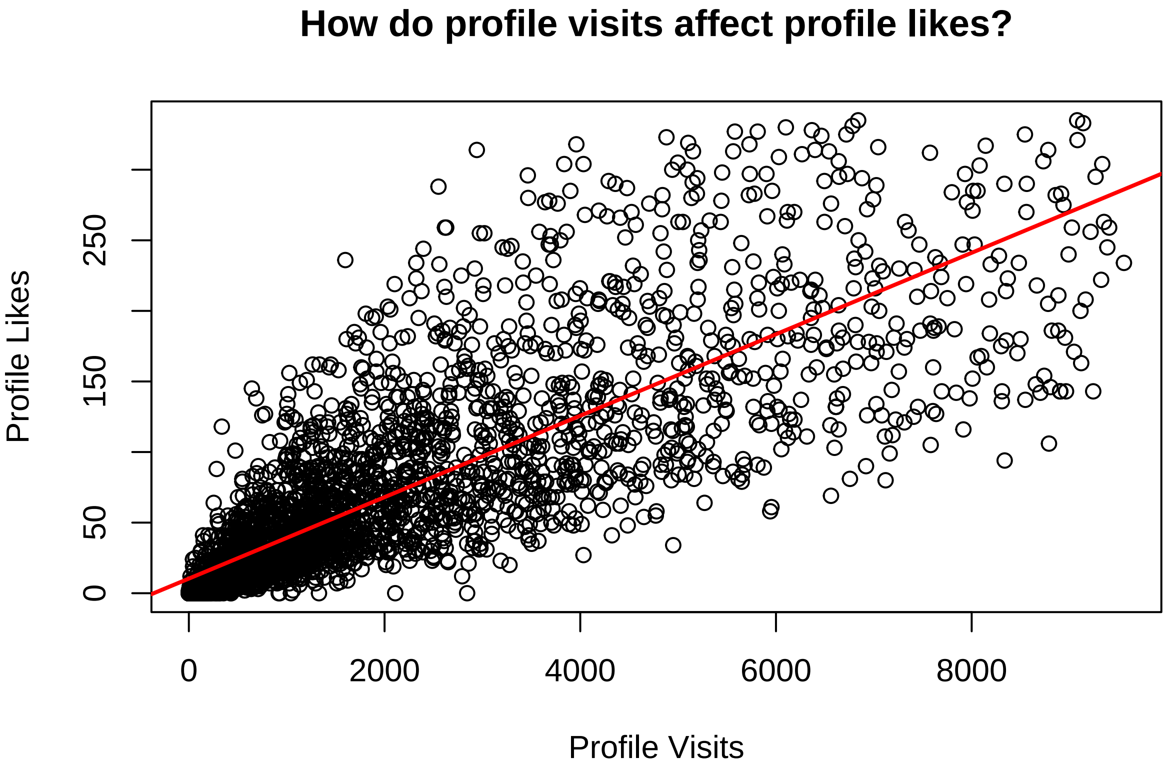


Figure 6: Revised scatterplot of profile visits vs profile likes

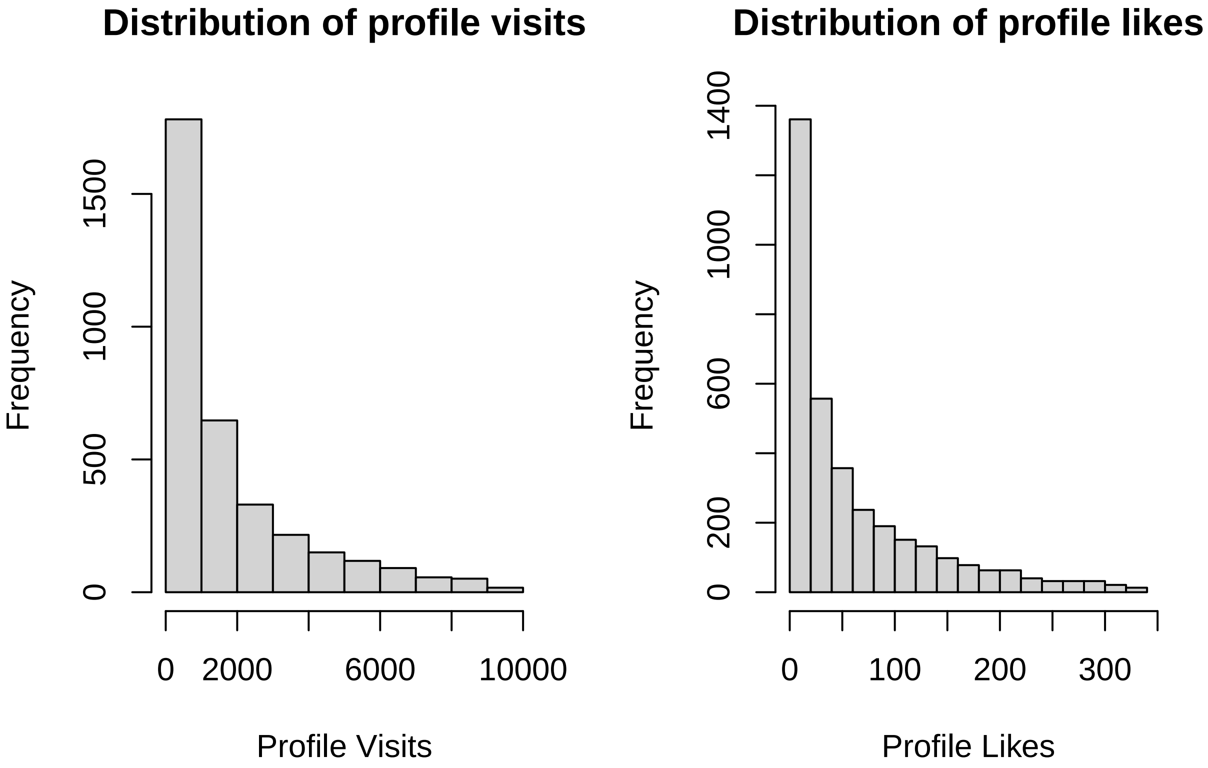
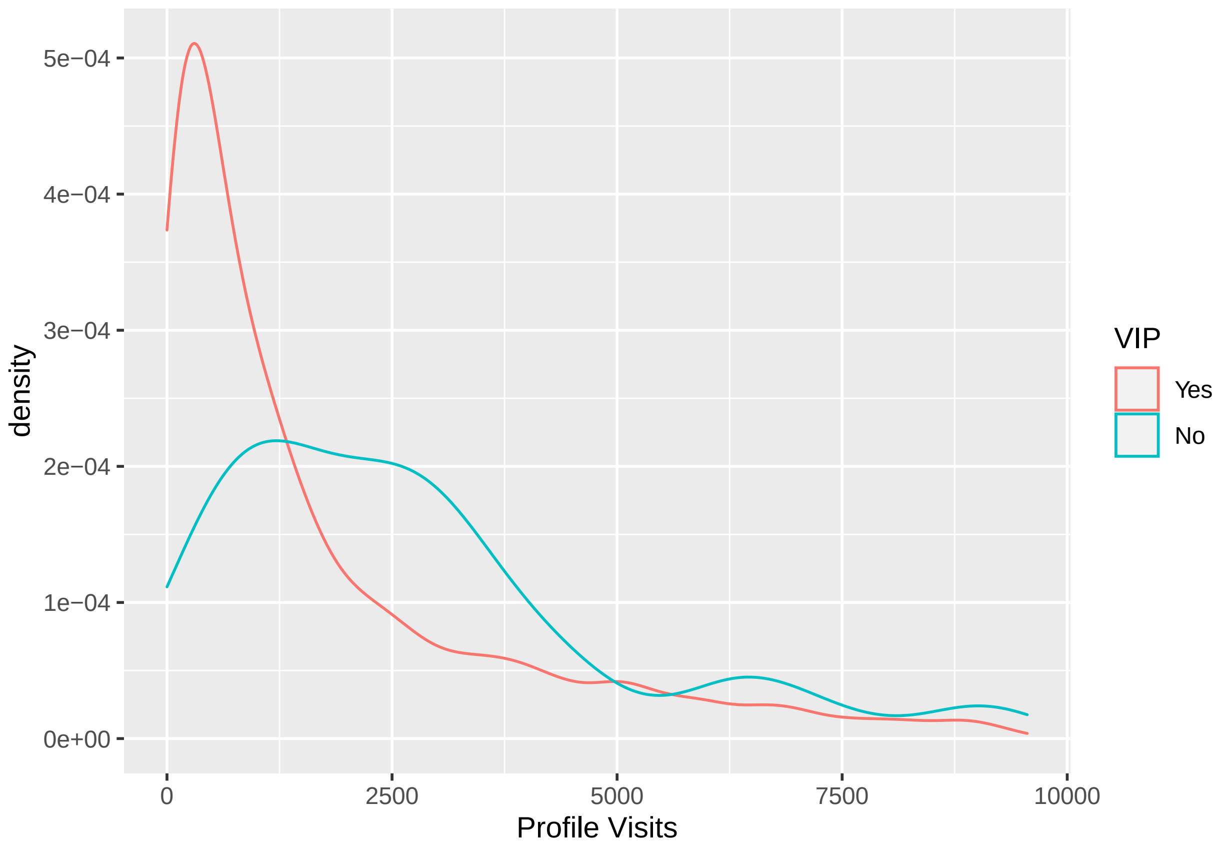
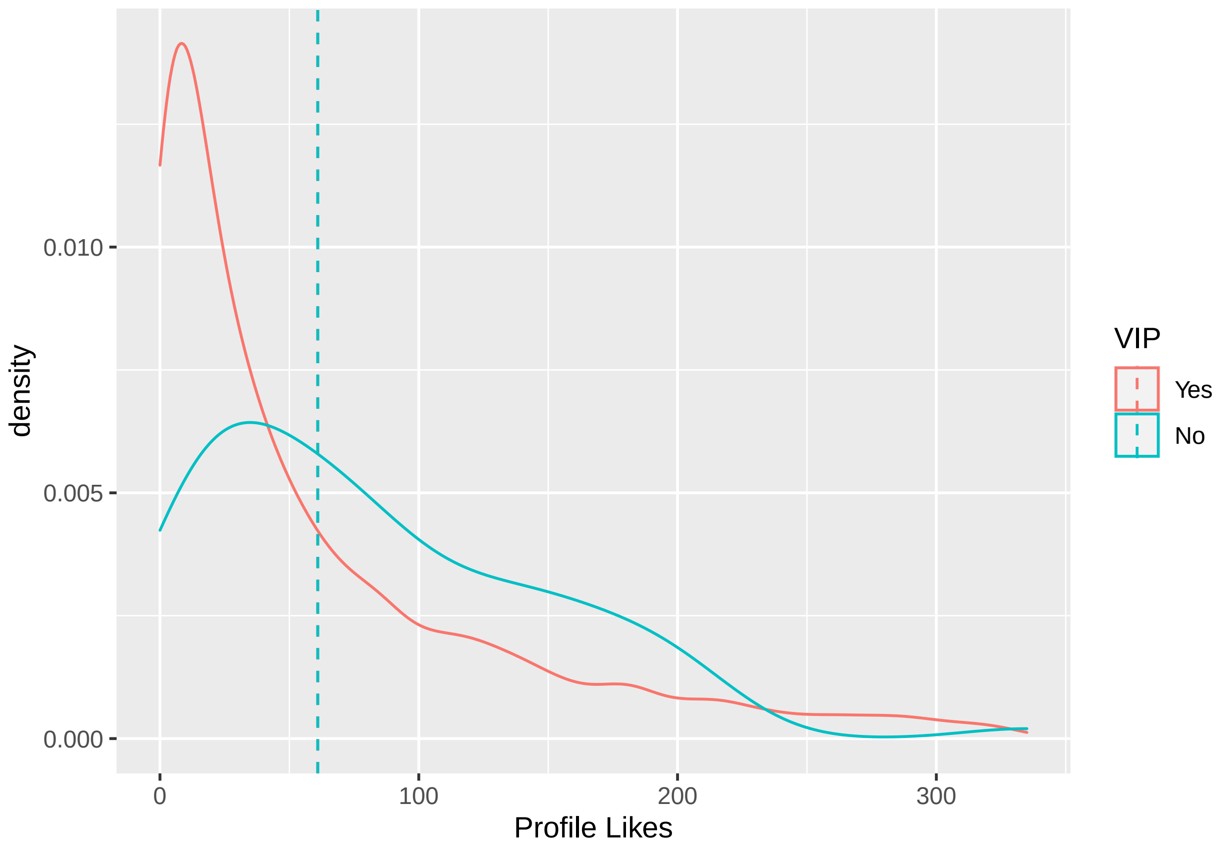


Figure 7: Histograms of Frequency of Profile Visits and Frequency of Profile Likes


5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
density
0
2500
5000
7500
10000
Profile Visits
VIP
Yes
No



3 Country, isVip and Distance

Looking at the variables - country, isVIp and distance individually. It can be observed that the greatest number of users come from: Switzerland (CH), Germany (DE), France (FR) and Italy (IT) (figure 9. From figure 10 we can see that the majority of users haven't purchased VIP status. It can be observed that the distance data is left skewed. The majority of distance between this user's city/location and the location of the user account that was used to fetch the data is higher than the median (figures 11 and 12).

13 shows that the median of Vip and non-Vip are approximately equal, however the distance of the not a Vip's is much larger than that of the Vip's which is to be expected given that a there is a larger number of non-Vip's compared to Vip's. This can also be seen in 4.

14 shows it can be observed from this graph that Switzerland (CH), Germany (DE), France (FR) and Italy (IT) have the greatest number of users who have purchased VIP status.

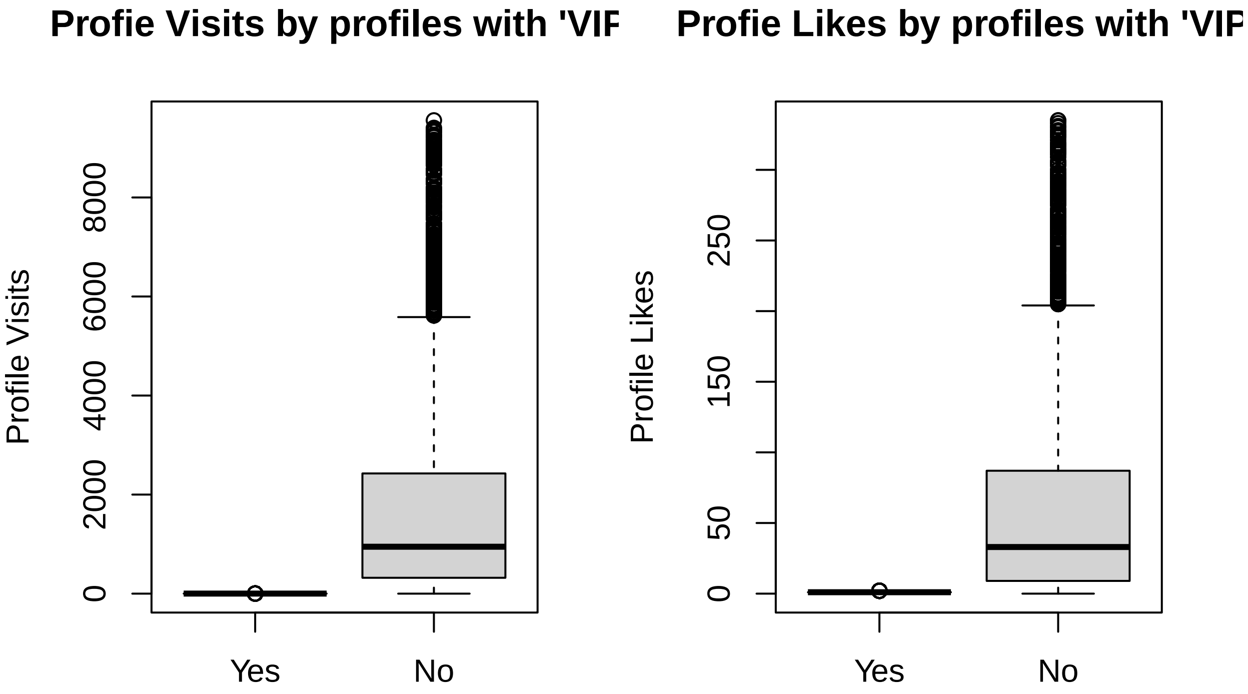


Figure 8: side by side boxplots of Profile Visits and Profile Likes against VIP status

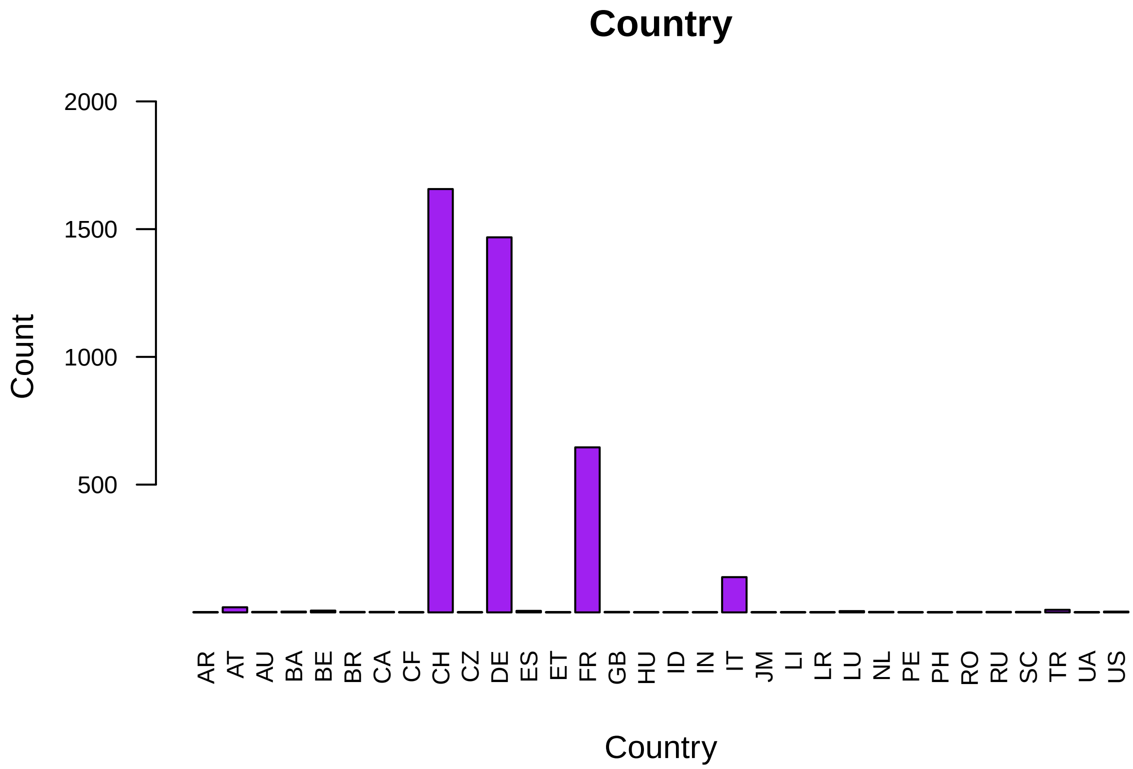


Figure 9: Number of users by country

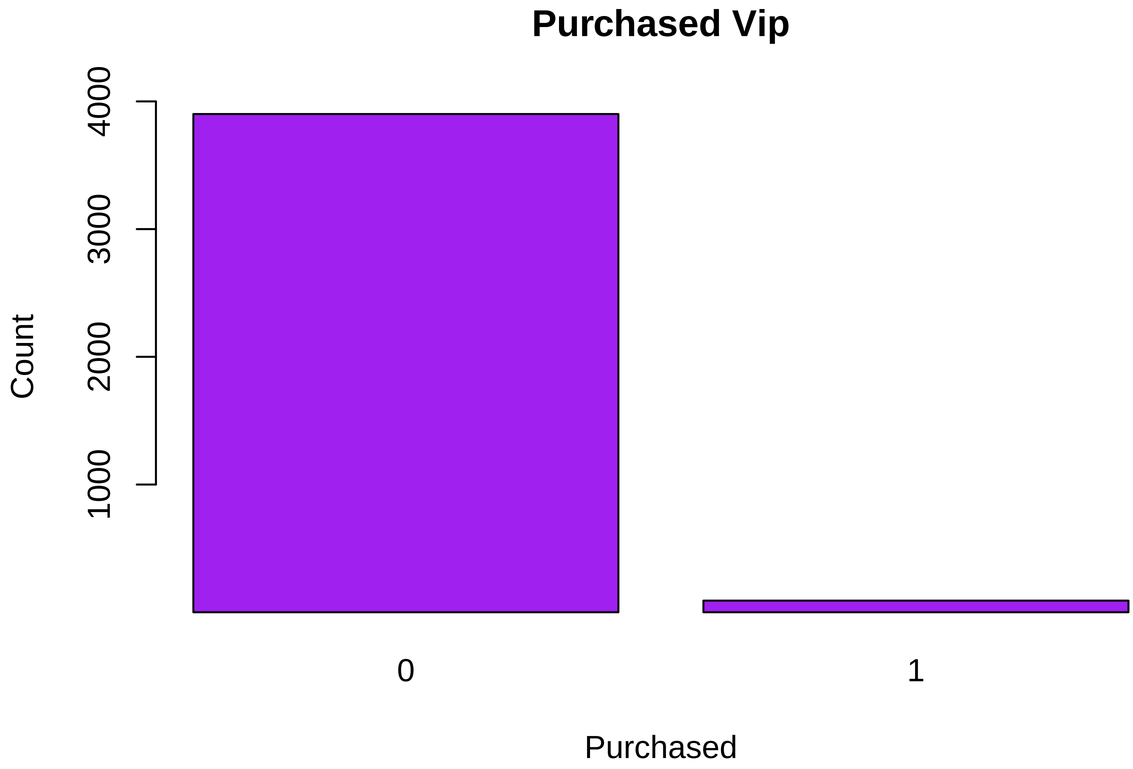


Figure 10: Count of whether a user is a VIP (nor not)

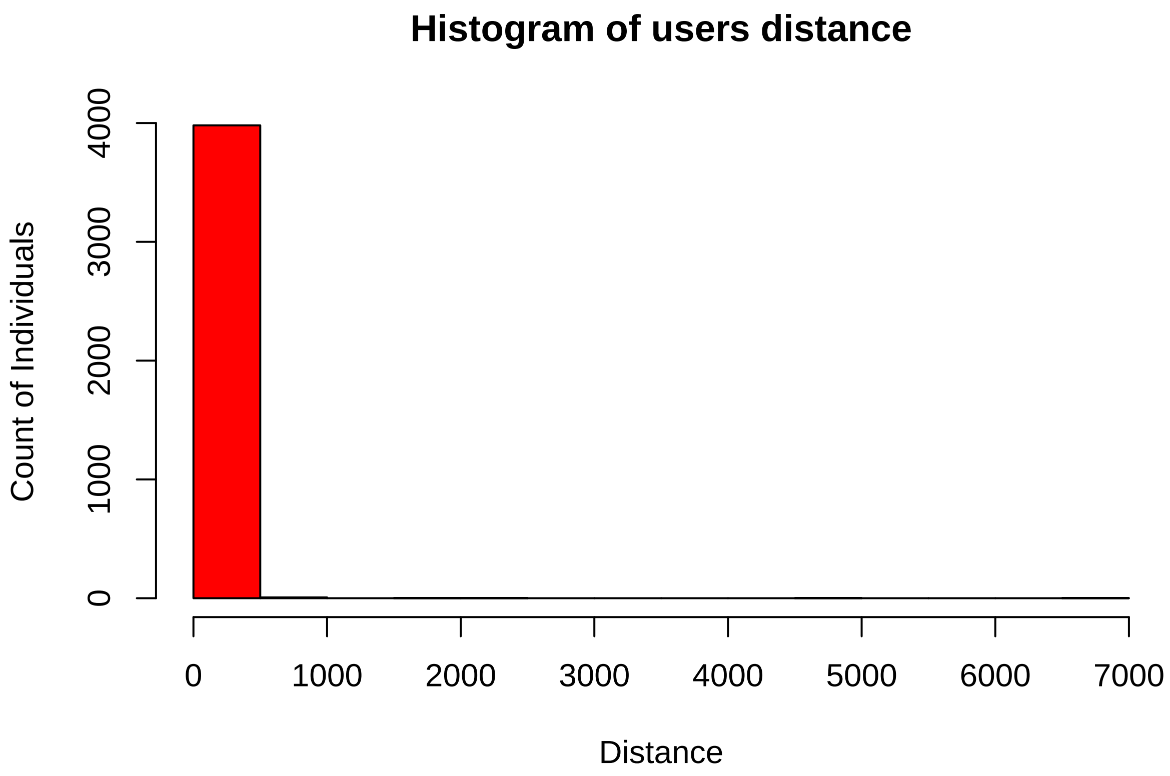


Figure 11: Histogram of individuals distance

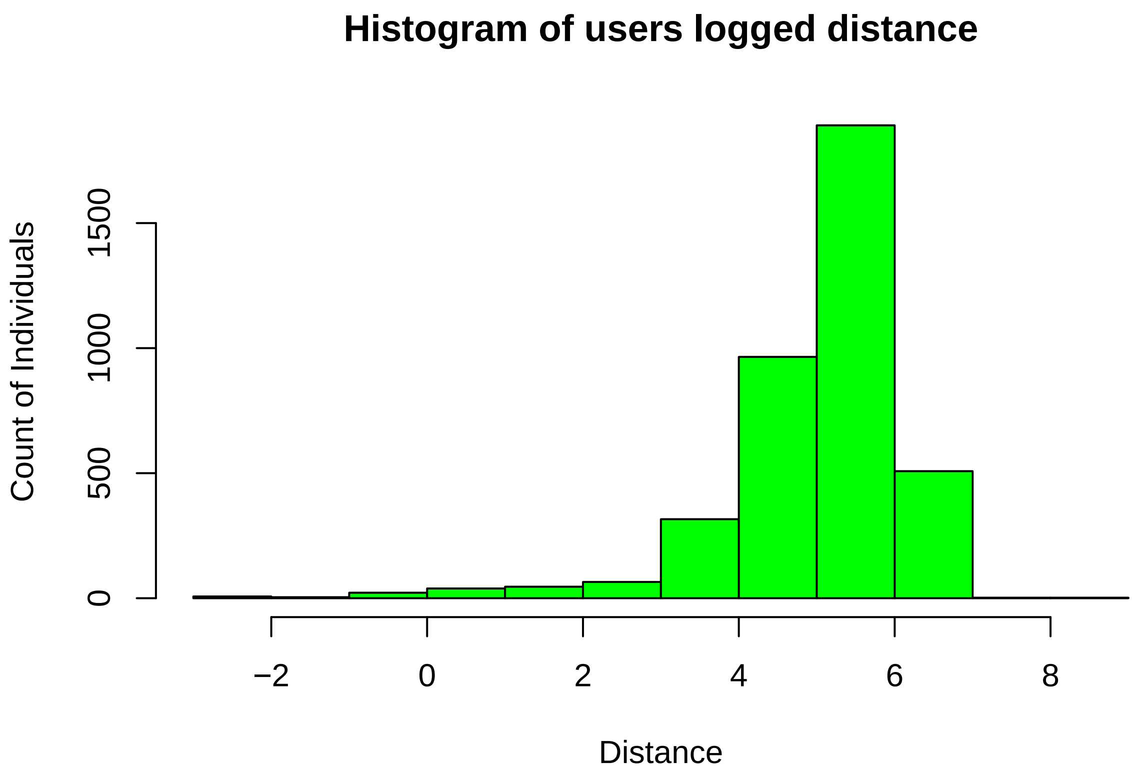


Figure 12: Histogram of users logged distance

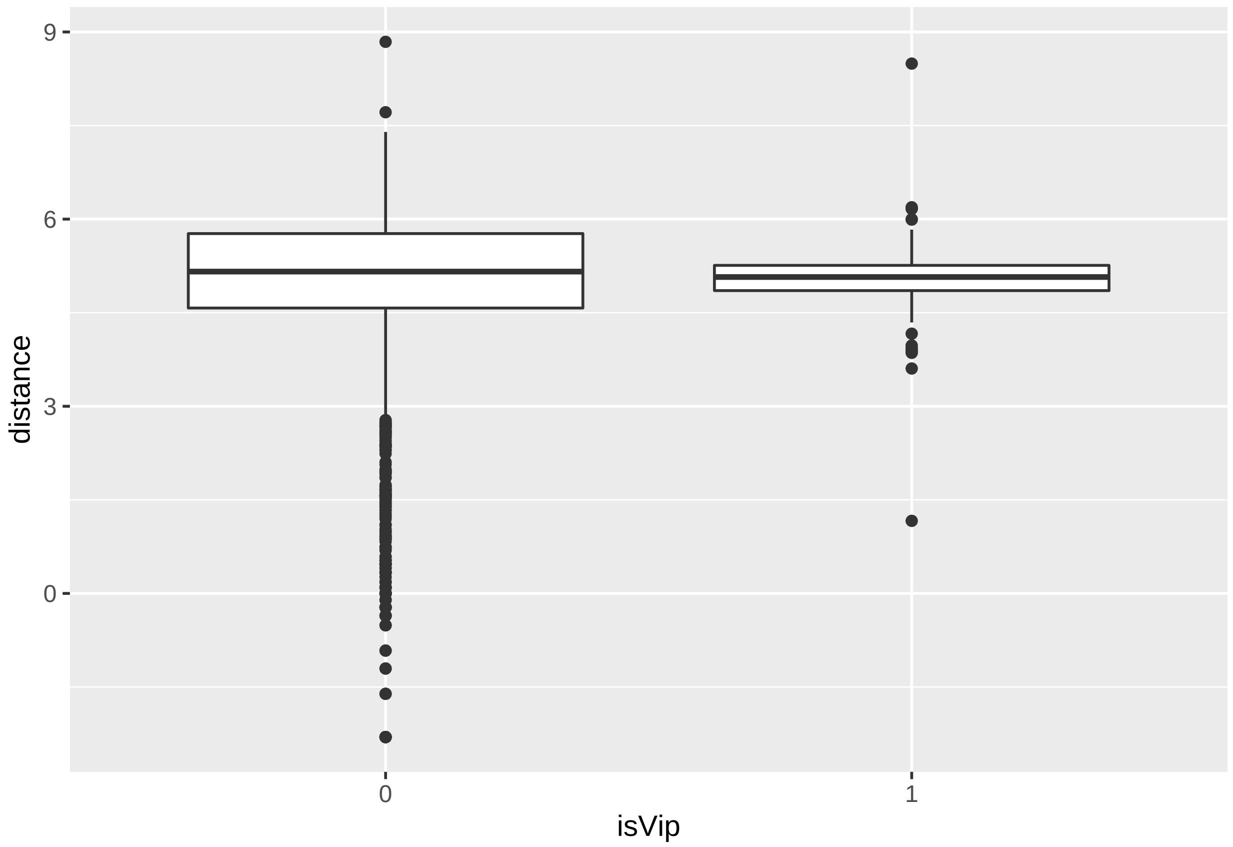


Figure 13: Boxplot of distance and isVip

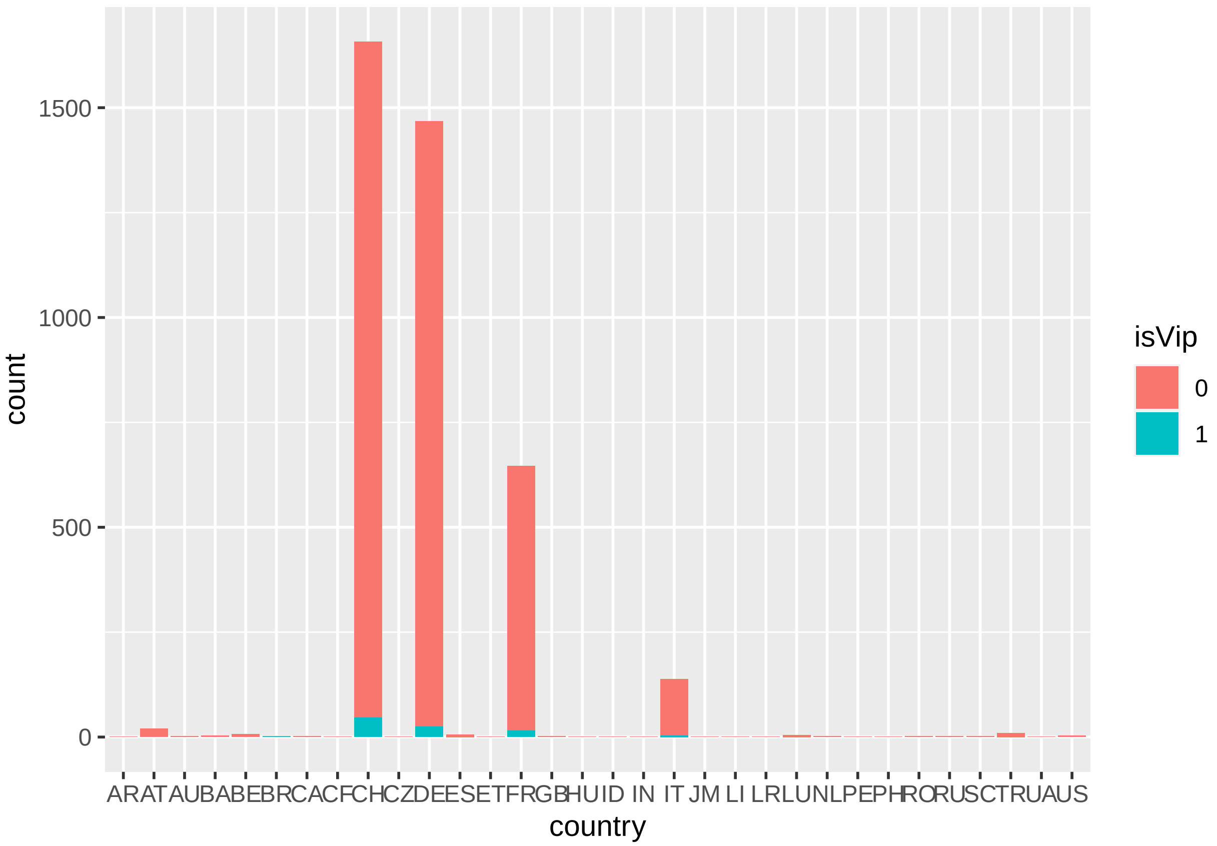


Figure 14: Count of users by country and Vip status

4 Counts_pictures, counts_profiles and counts_kisses

From table 1 we can see that the number of pictures a user profile has, has a heavy right tail, showing most women have about 4 photos attached to their accounts and the maximum the can choose to have is 30. All values are positive. Profile visits is slightly more symmetric than the pictures distribution but still has a heavy skew with some extreme values and large range. This variable would benefit from the analysis of distances. All values are positive. The number of kisses (likes) has a heavy right tail with large range. But all values are positive.

Figure 15 shows the Square-root transformed (for readability) distribution of counts_pictures. As we can see, is it pretty normal to have between 0 & 12 photos as this is where the boxplot spans. It is heavily centered around the lower end of the count. Figure 16 shows the Square-root transformed distribution of counts_profileVisits. There has been an extreme value (100,000 visits counted) removed as it made the graph unreadable. It is heavily centered around the lower end of the count. Figure 17 shows the Square-root transformed distribution of counts_kisses. It is heavily centered around the lower end of the count.

The variance and covariance matrices can be seen in tables (5 and 6). All variables are positively related

All variables are positively correlated, this is what we expect as with more visits there will be more likes and so on. Figure 18 shows that there is a correlation of interest between profile visits and kisses.

Looking at Mahalanobis distance we can see there are some very surprising values (108) from the table and Figure 19, it could be worth while examining these and possibly removing them as outliers. These might be famous people that receive a lot of attention online or other cases that are not simply a single woman looking for love online.

Figure 20 further shows where the very surprising values are - they are mainly on the fringes of the data. although in the univariate distributions we can see that most very suprising values do make up the lower end of the photos which is what we would assume was normal. There will have to be more discussion around the exclusion of points from the data.

The contour plots are giving a warning message, I think due to scale

All contour plots (Figure 21, Figure 22, Figure 23) are centered around zero and spread with a positive relationship to each other.

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

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

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

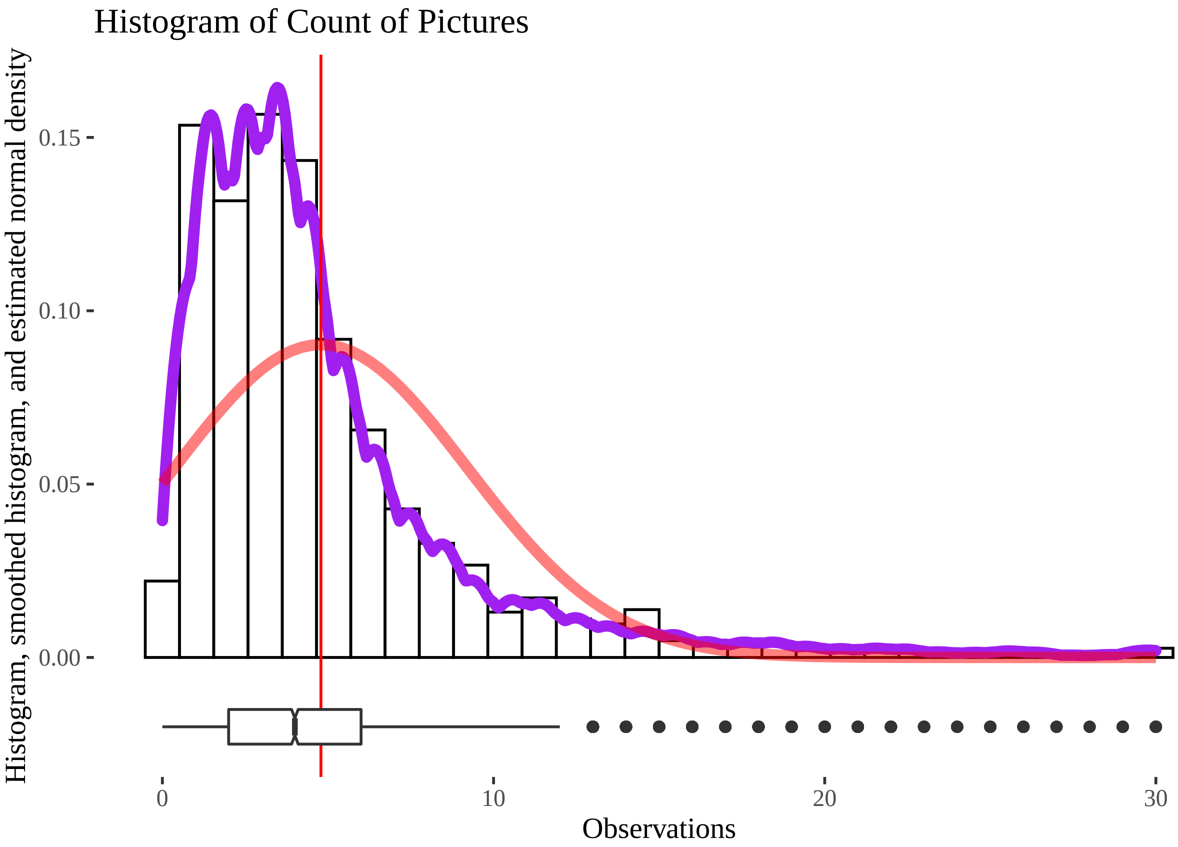


Figure 15: Square-root transformed histogram of counts_pictures

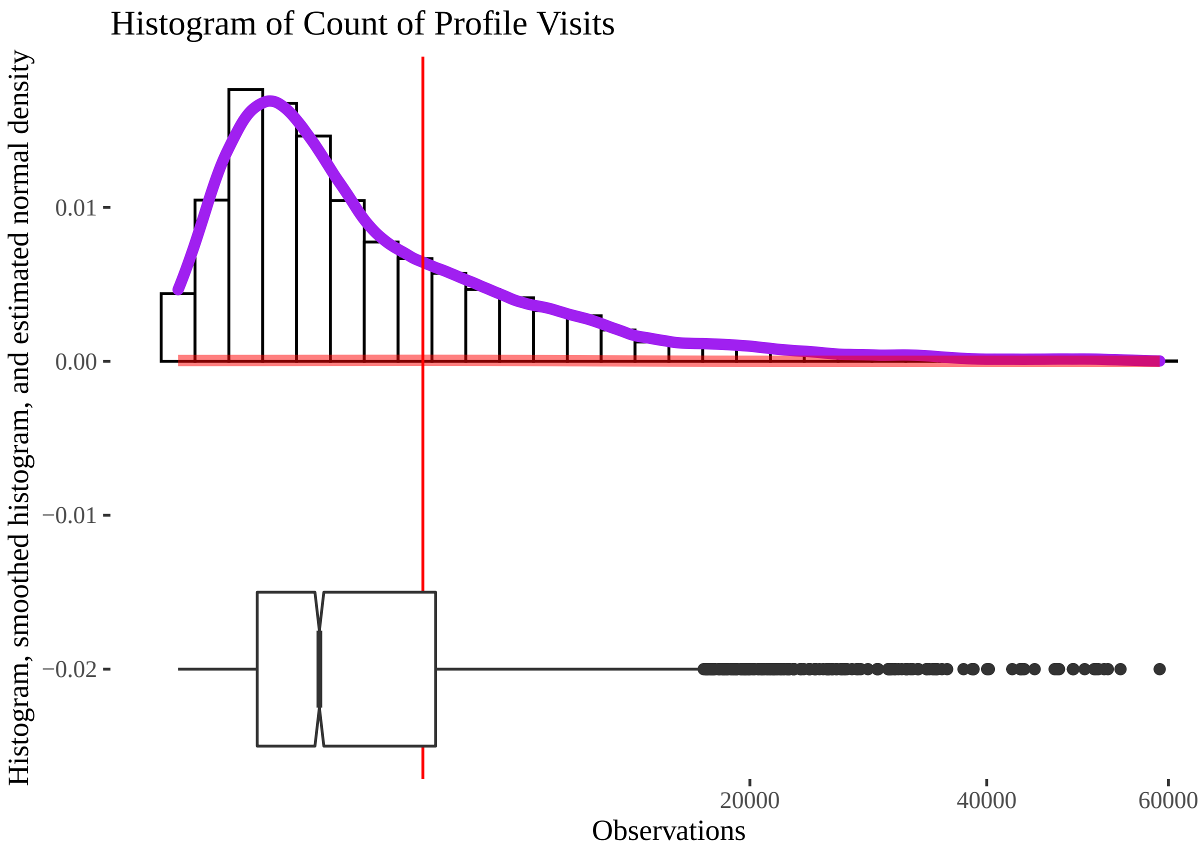


Figure 16: Square-root transformed histogram of counts_profileVisits

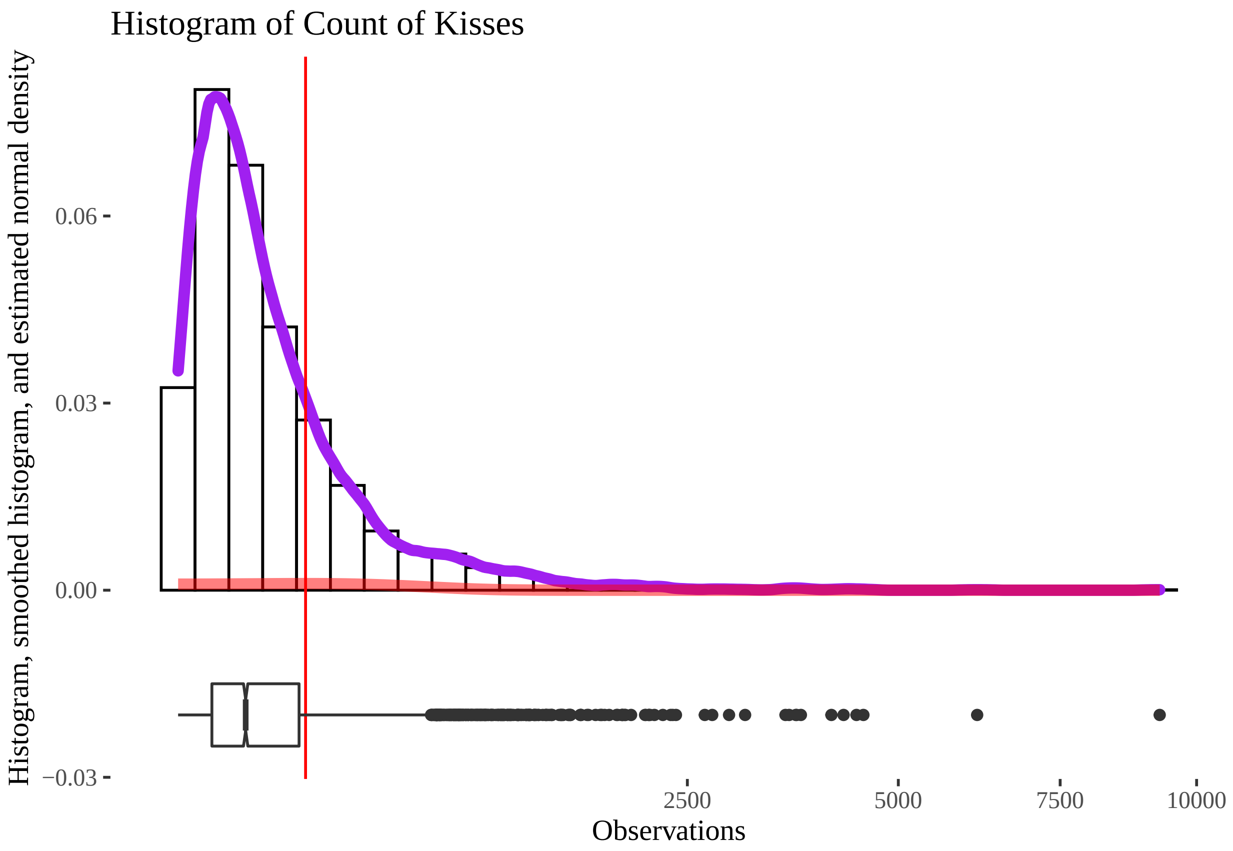


Figure 17: Square-root transformed histogram of counts_kisses

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

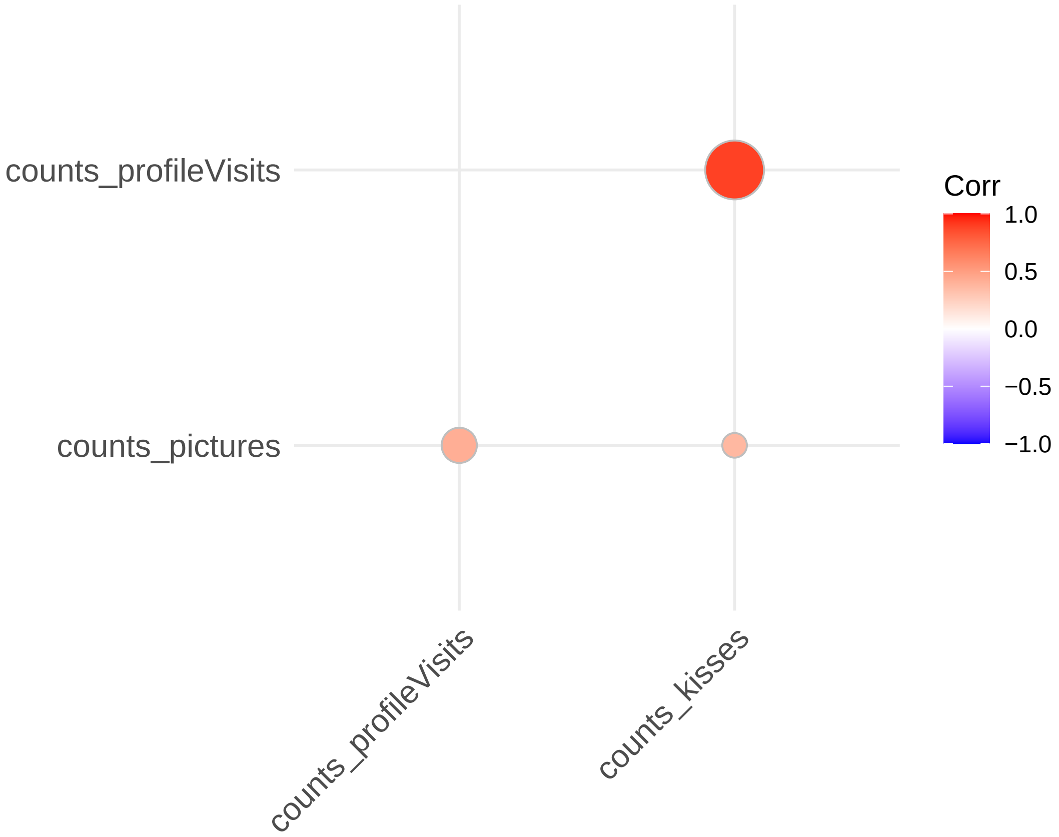


Figure 18: Correlation plot for counts pictures, profile visits, and kisses visualised

5 Age, country and number of pictures

Figure 24 shows the number of users by age. The number of users is heavier in the age range 19-22 (inclusive). It appears that a large number of countries have a very small number of users, while a few countries such as Switzerland, Germany, France and Italy have larger numbers of users. This can be seen in 9. Table 8 shows us a breakdown of the number of users per country by age. Again this supports higher numbers in Switzerland, Germany, France and Italy. Taking a closer look at the countries - Switzerland, Germany, France and Italy we can look at the number of users per country by age. Figure 25 shows that Switzerland has the larger number of users with ages 19-21 and Germany has the larger number of users for ages 22-24. Looking at the boxplot in figure 26 we can see that there are a few very large outliers in the data and it is very hard to read the boxplots. By taking the log of the data we can see the median more clearly. Figure 27 shows that the

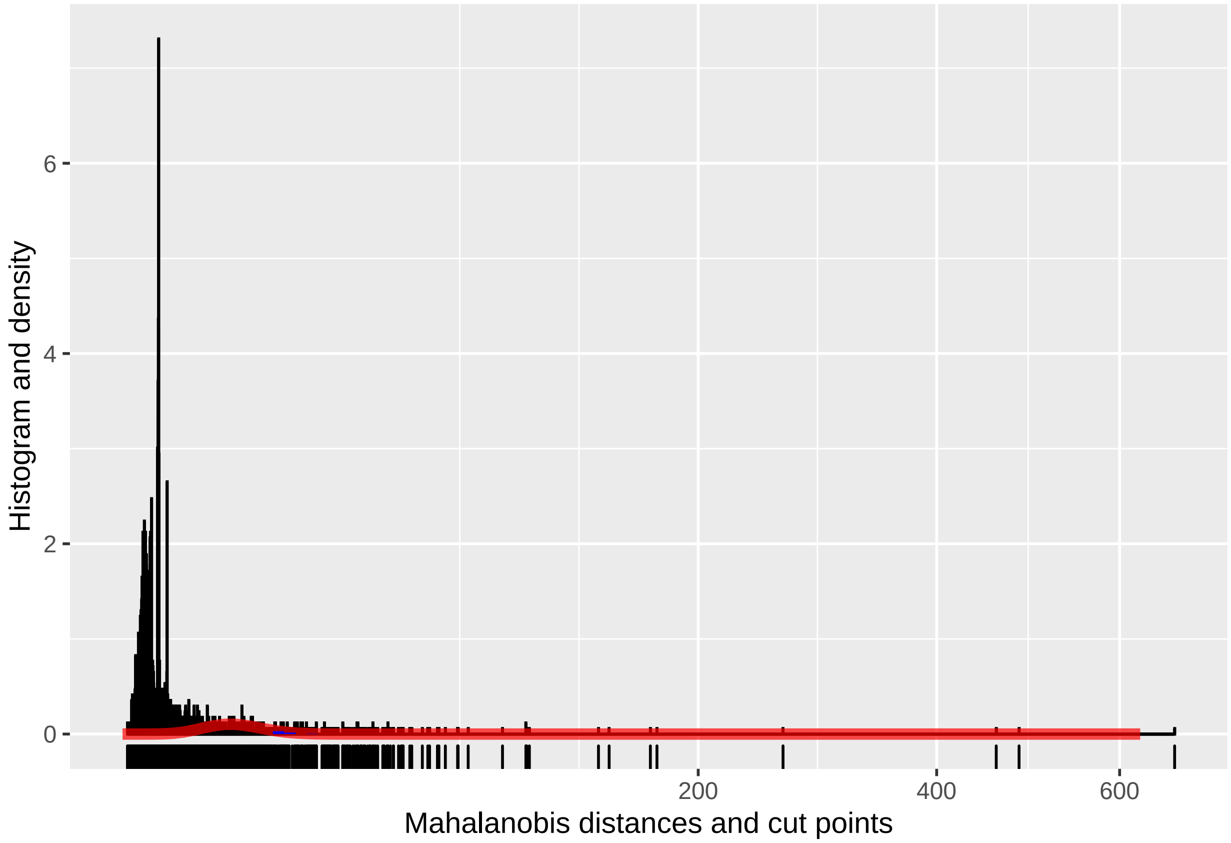


Figure 19: Mahalanobis distince plotted

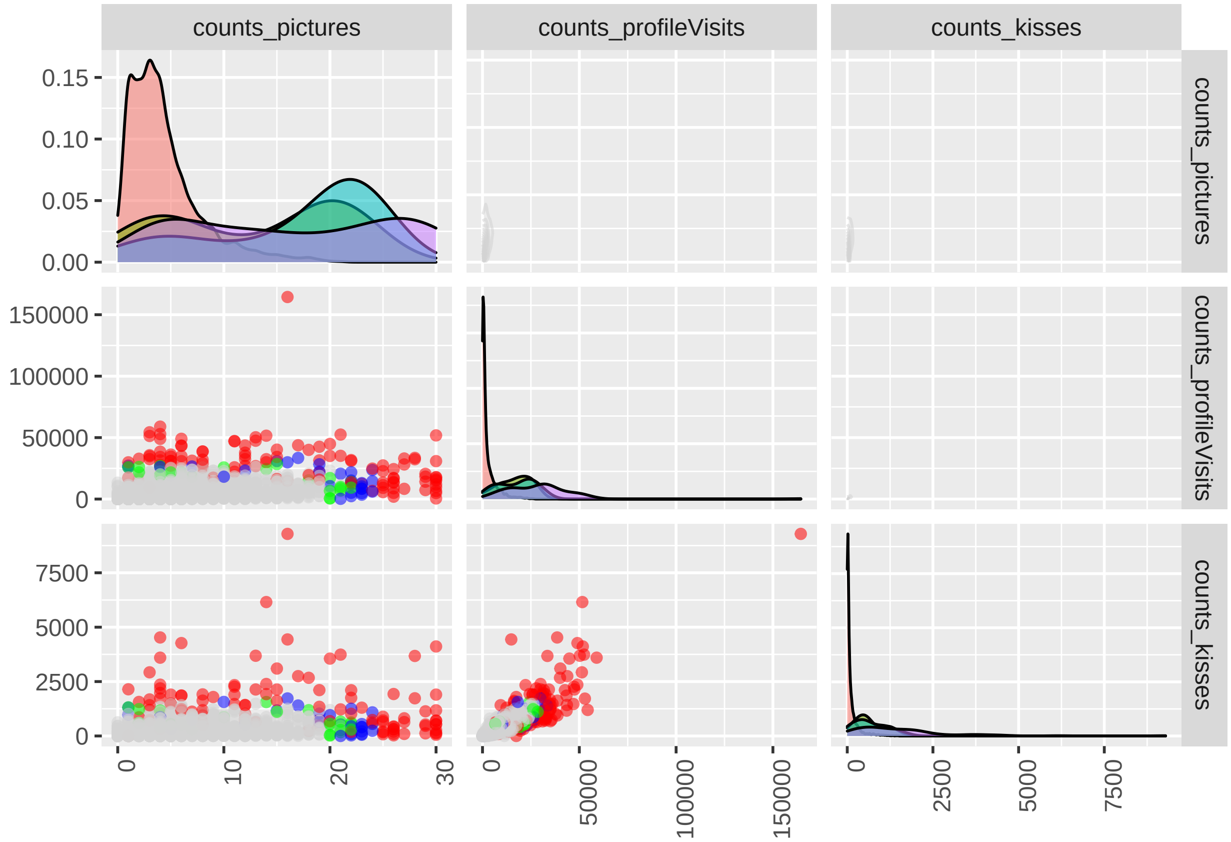


Figure 20: Pairs plot of count pictures, profile visits, and kisses

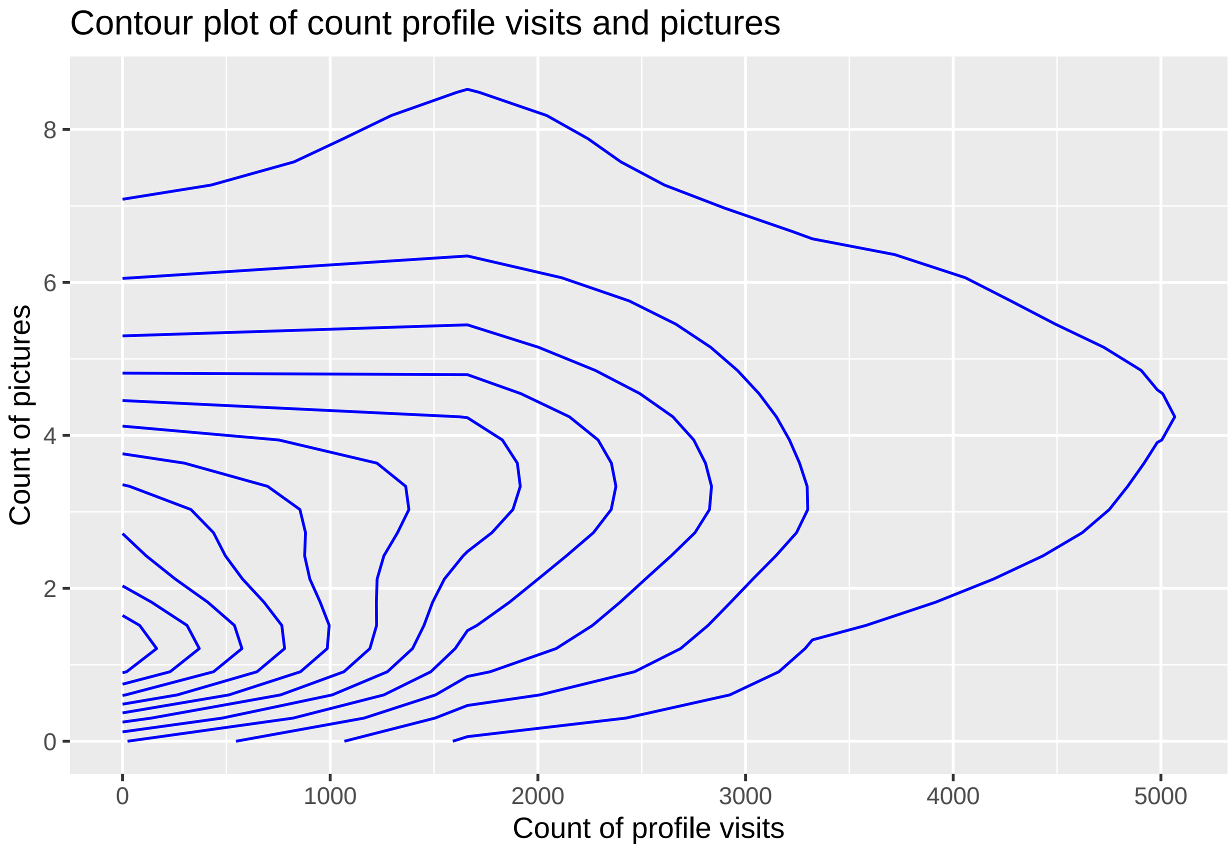


Figure 21: Contour plot of count profile visits and pictures

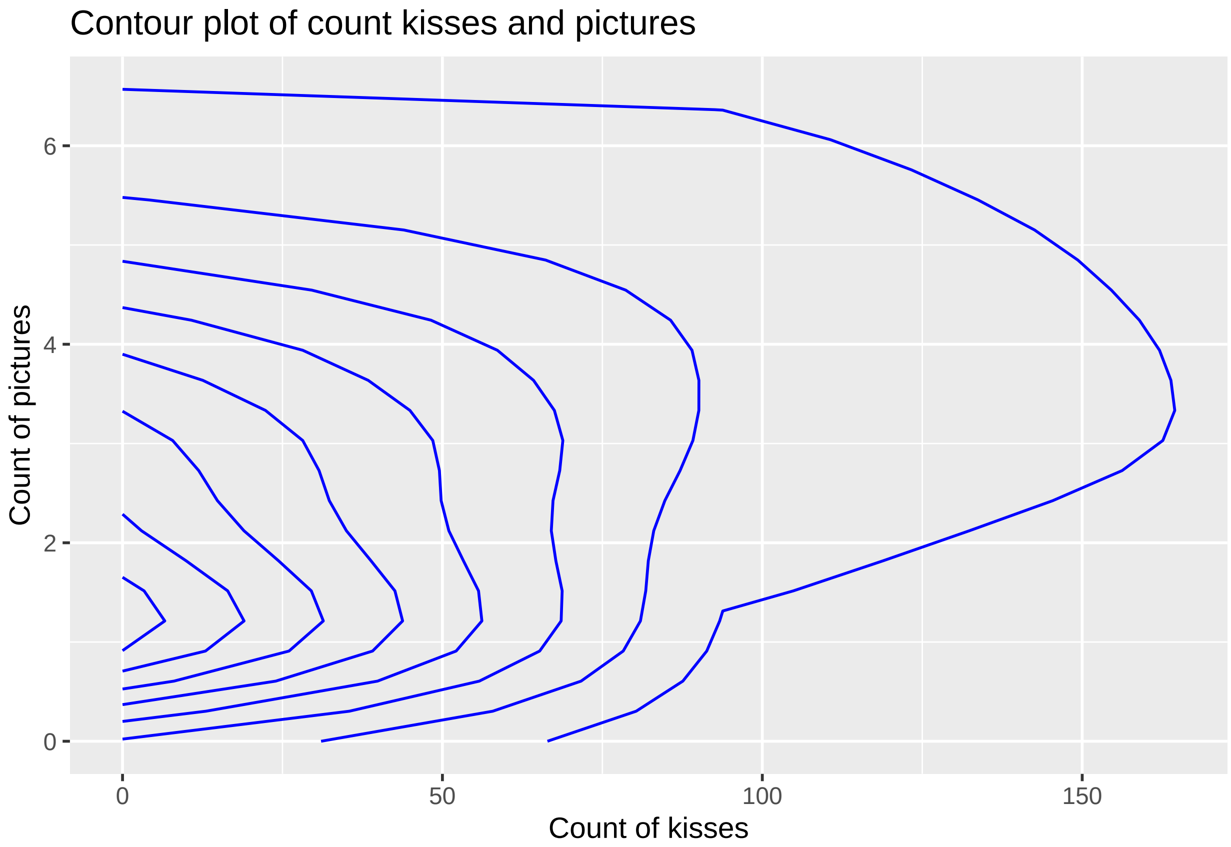


Figure 22: Contour plot of count kisses and pictures

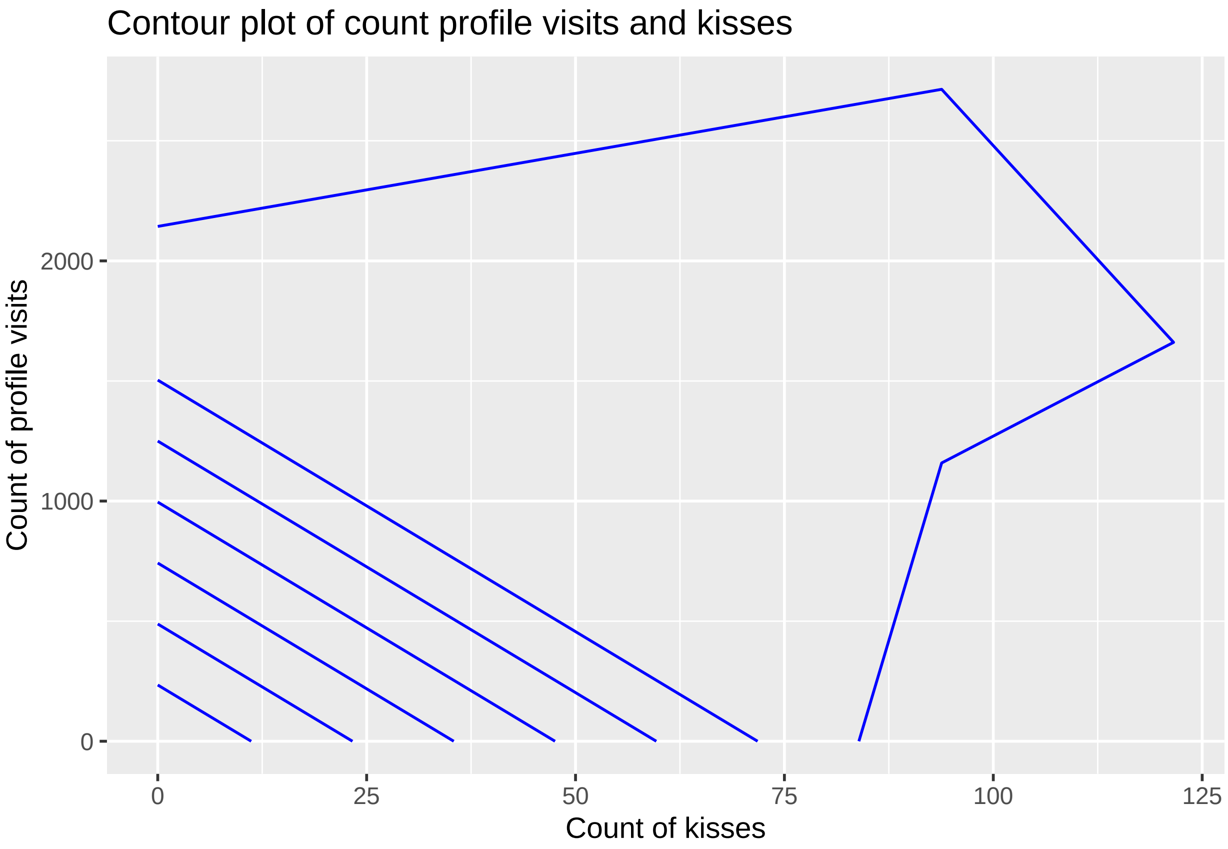


Figure 23: Contour plot of count profile visits and kisses

medians vary, however there is not much variability within a country across the different ages.

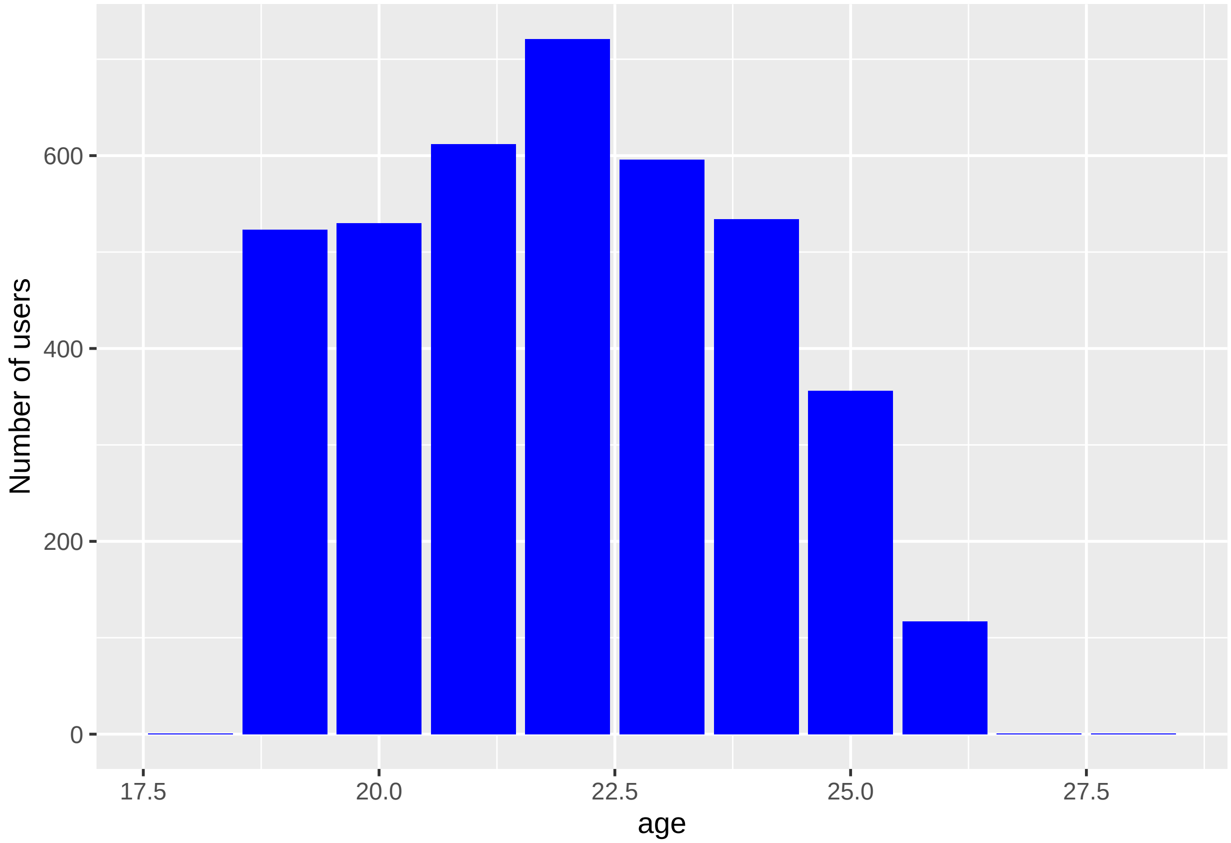


Figure 24: Number of users by age

```
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
```

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

```
## Warning: Removed 124 rows containing non-finite values (stat_boxplot).
```

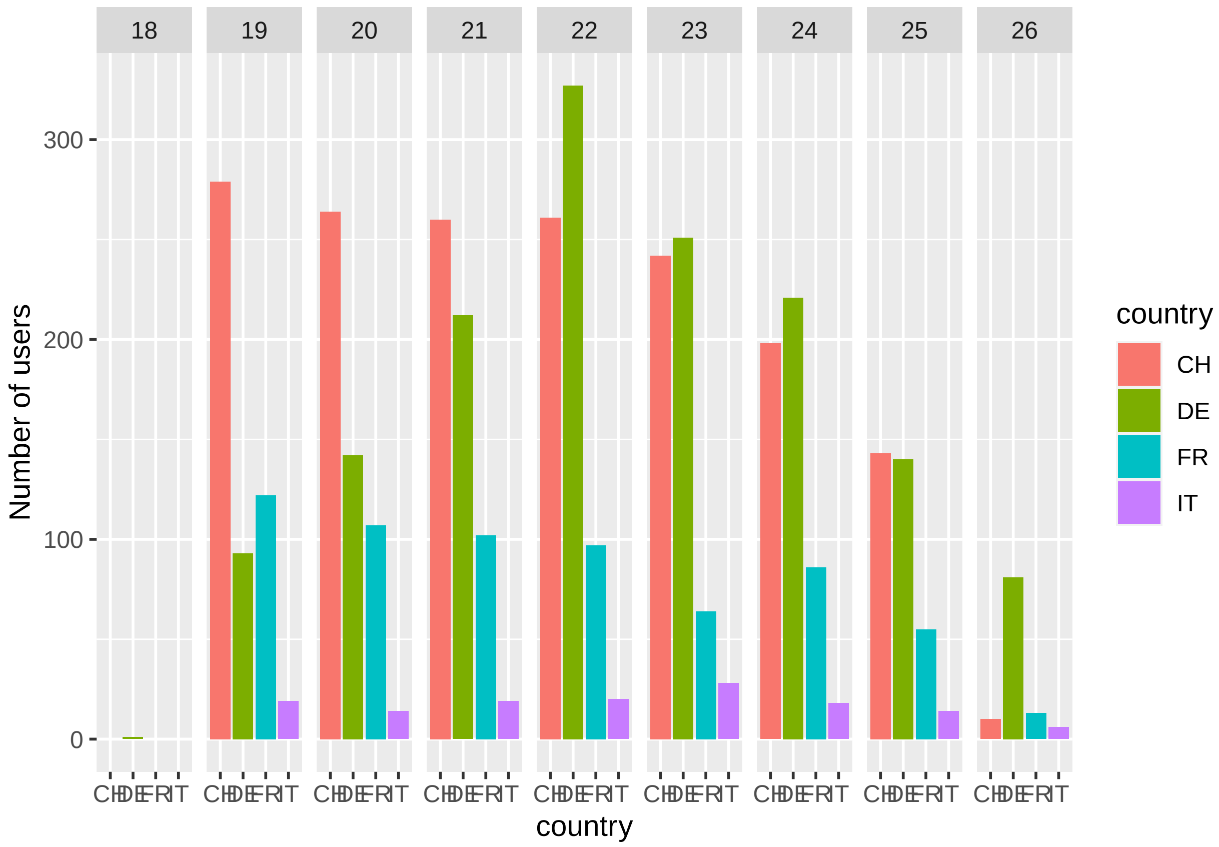


Figure 25: Number of users by age and country

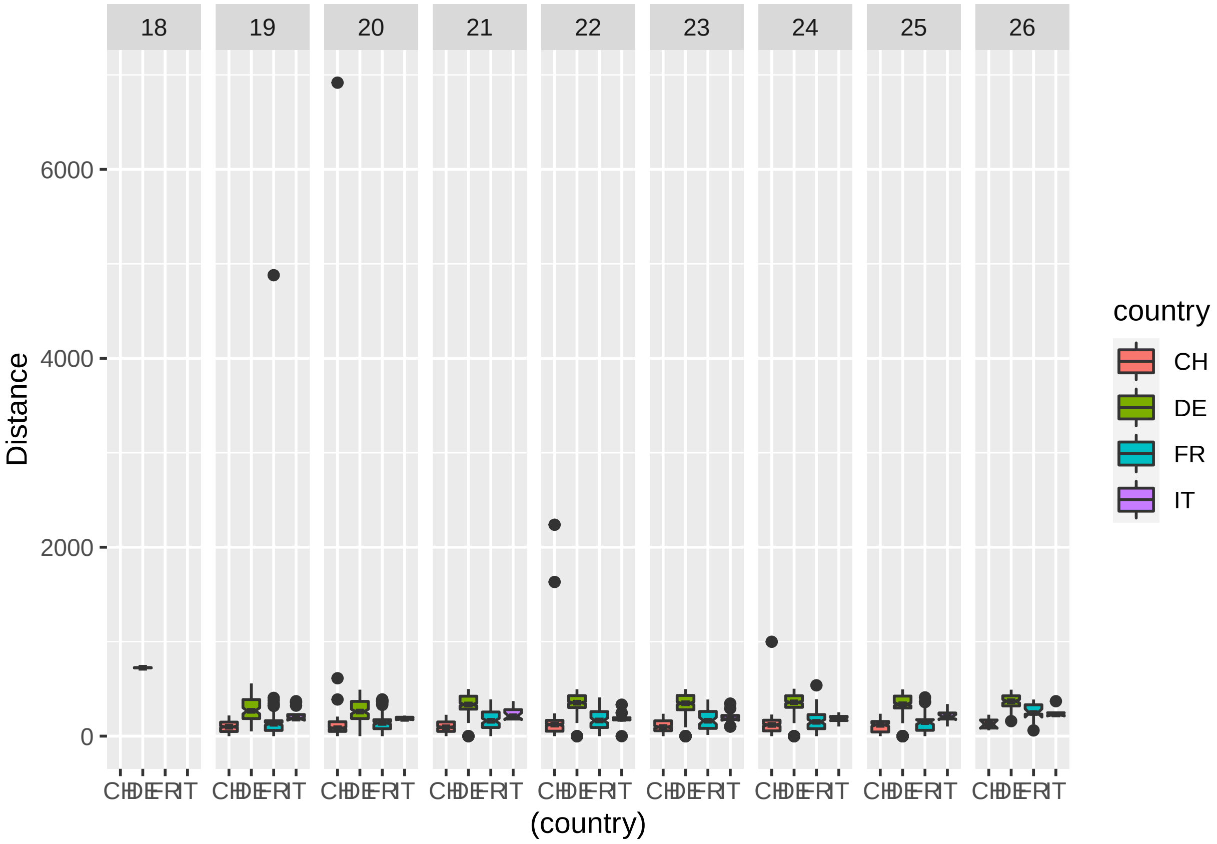


Figure 26: Side by side boxplots of age, country and distance

```
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
```

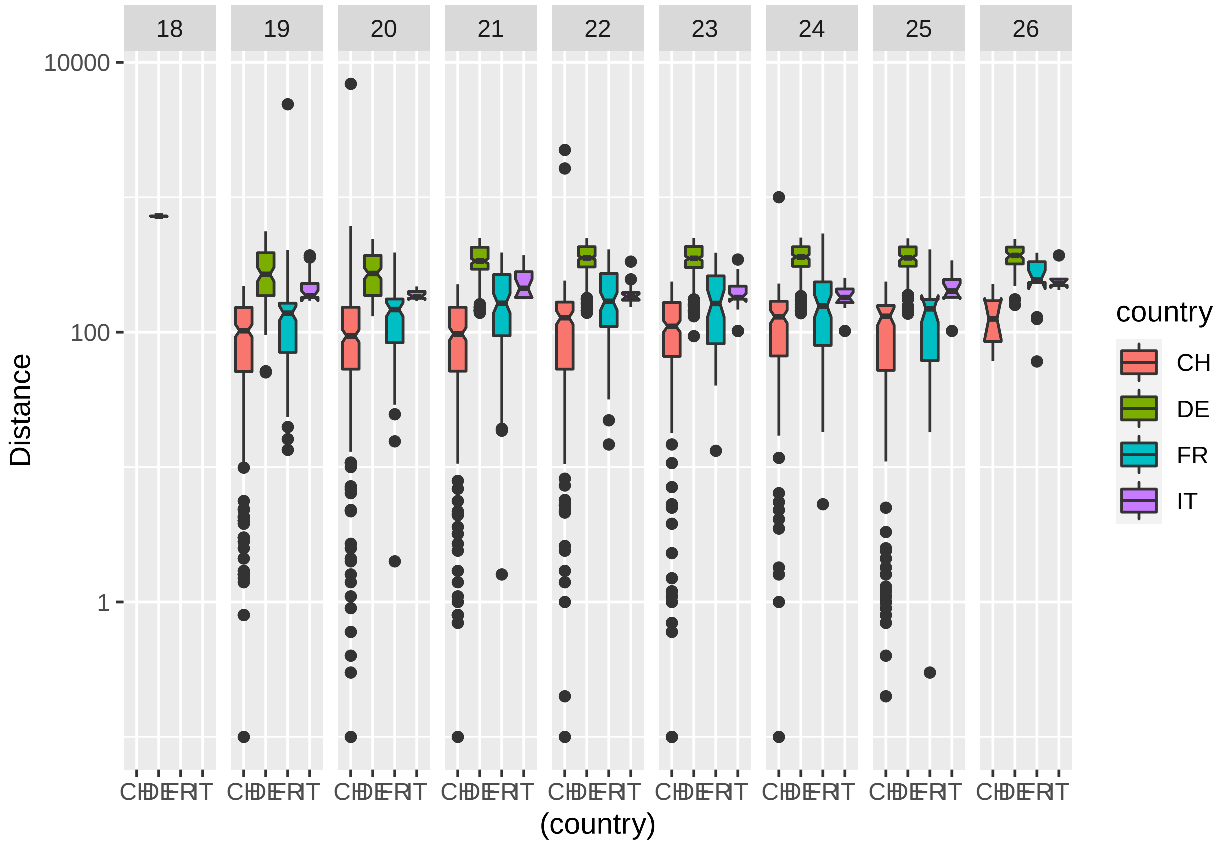


Figure 27: Side by side boxplots of age, country and distance (log scale)

References

Commons, Creative. n.d. "Attribution 4.0 International (CC BY 4.0)." https://www.kaggle.com/datasets/sandragracenelson/suicide-rate-of-countries-per-every-year.

Mabilama, Jeffrey Mvutu. 2020. "Dating App User Profiles' Stats - Lovoo V3." https://www.kaggle.com/datasets/jmmvutu/dating-app-lovoo-user-profiles?resource=download.

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

Table 4: Summary Statistics - isVip (or not) by country

	isVip - No	isVip - No
AR	1	0
AT	20	0
AU	2	0
BA	3	0
BE	7	0
BR	1	1
CA	2	0
CF	1	0
CH	1611	46
CZ	1	0
DE	1443	25
ES	6	0
ET	1	0
FR	631	15
GB	2	0
HU	1	0
ID	1	0
IN	1	0
IT	134	4
JM	1	0
LI	1	0
LR	1	0
LU	5	0
NL	2	0
PE	1	0
PH	1	0
RO	2	0
RU	2	0
SC	2	0
TR	10	0
UA	1	0
US	3	0

Table 5: Variance matrix

	counts_pictures	counts_profileVisits	counts_kisses
counts_pictures	19.54	12647.04	614.40
counts_profileVisits	12647.04	46854549.74	2289351.68
counts_kisses	614.40	2289351.68	142620.04

Table 6: Correlation matrix

	counts_pictures	counts_profileVisits	counts_kisses
counts_pictures	1.00	0.42	0.37
counts_profileVisits	0.42	1.00	0.89
counts_kisses	0.37	0.89	1.00

Table 7: Surprising table

	V1
Typical	3832
Somewhat	23
Surprising	29
Very	108

Table 8: Country vs age

	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