Exploratory Data Analysis

Lissa Harrop, Katrina Watkins, Ricky Loo and Max Tan

September 2022

Contents

1 Exploratory Data Analysis of our overall data

2 Univariate

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 set is avail-

able on creativecommons (Commons (n.d.)). 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 These have been replaced by the mean of the distance column (207.23).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.

As identified in the summary statistics the median age is 22 and this is supported by the boxplot in figure 1. We can also see that the mean age is equal to 22 also. The green line shows the estimated normal density for the age. The age of the users appears to be normally distributed. The blue line shows the smoothed density histogram for age, this appears to go down between each value of age. This is likely to the fact that age has been measured discretely and the smoothed histogram is continuous. *We are unsure as to why we are getting a warning message for figure 1 as there is no missing data for age.*

The bulk of users tend to have less than 10 profile pictures on their account. We can see from the boxplot in figure 2 that there are a number of users that have a higher number of profile pictures on their account. This is also visible in the histogram which shows a long tail to the right.

The original graph for the number of profile visits a user has is very hard to read, so figure 3 shows the squre root transformed data. This is still hard to read due to an extreme outlier (+100,000 visits counted). We removed the extreme outlier in order to better read the graph. We can see in figure 4 that mean is much closer to the 75% percentile than it is the median. User's tend to have between 0 and 4000 profile visits, however, about 25% of user's have a much higher number of visits and this is skewing the data.

The number of 'kisses' a user receives appears to be heavily skewed to the right. There are a large number of outliers or large values.The mean appears to be larger than the 75% percentile as seen in figure 5.

There are 32 different countries that users come from. The summary of the countries and their counts can be found in table 3 and a visualisation can be seen in figure 8. Table 2 shows that 3901 users are not Vip's while only 91 are Vip's and figure 7 shows a visualisation.

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: Removed 2 rows containing missing values (geom_bar).
```

Table 2: Summary Statistics - isVip

	isVip	Count
No	0	3901
Yes	1	91

After the initial analysis of our variables we noticed that four of the countries (Switzerland (CH), Germany (DE), France (FR) and Italy (IT)) have significantly larger number of users compared to the other 28 countries. The other countries have between 1 and 20 users, we decided that we would group all other countries into one category called 'other' and explore the data further. Even with grouping the 28 countries they still have the lowest number of users out of the five new groups. The 'other' group has 83 users as seen in table 4. A visualisation of the users by country can be seen in figure 9.

3 Bivariate and Multivariate

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

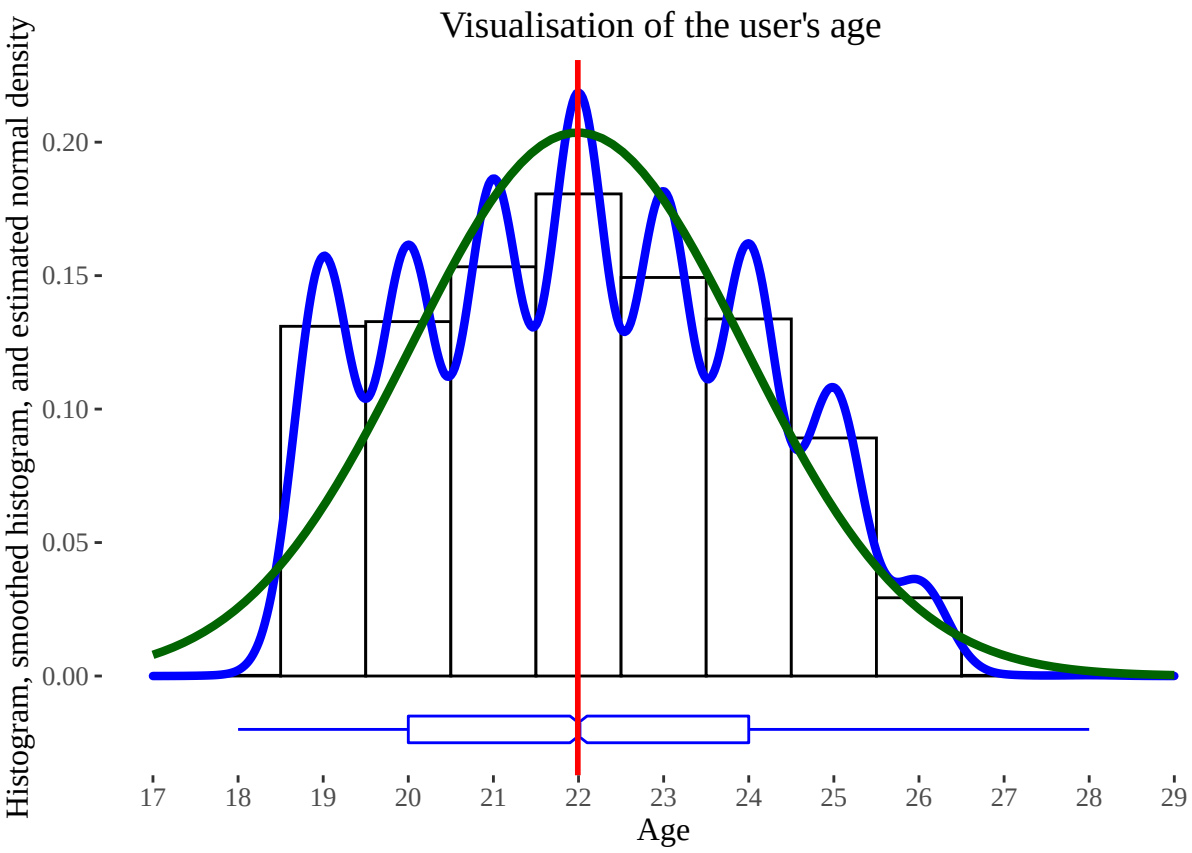

Visualisation of the user's age
Histogram, smoothed histogram, and estimated normal density
0.20
0.15
0.10
0.05
0.00
17
18
19
20
21
22
23
24
25
26
27
28
29
Age

Figure 1: Age

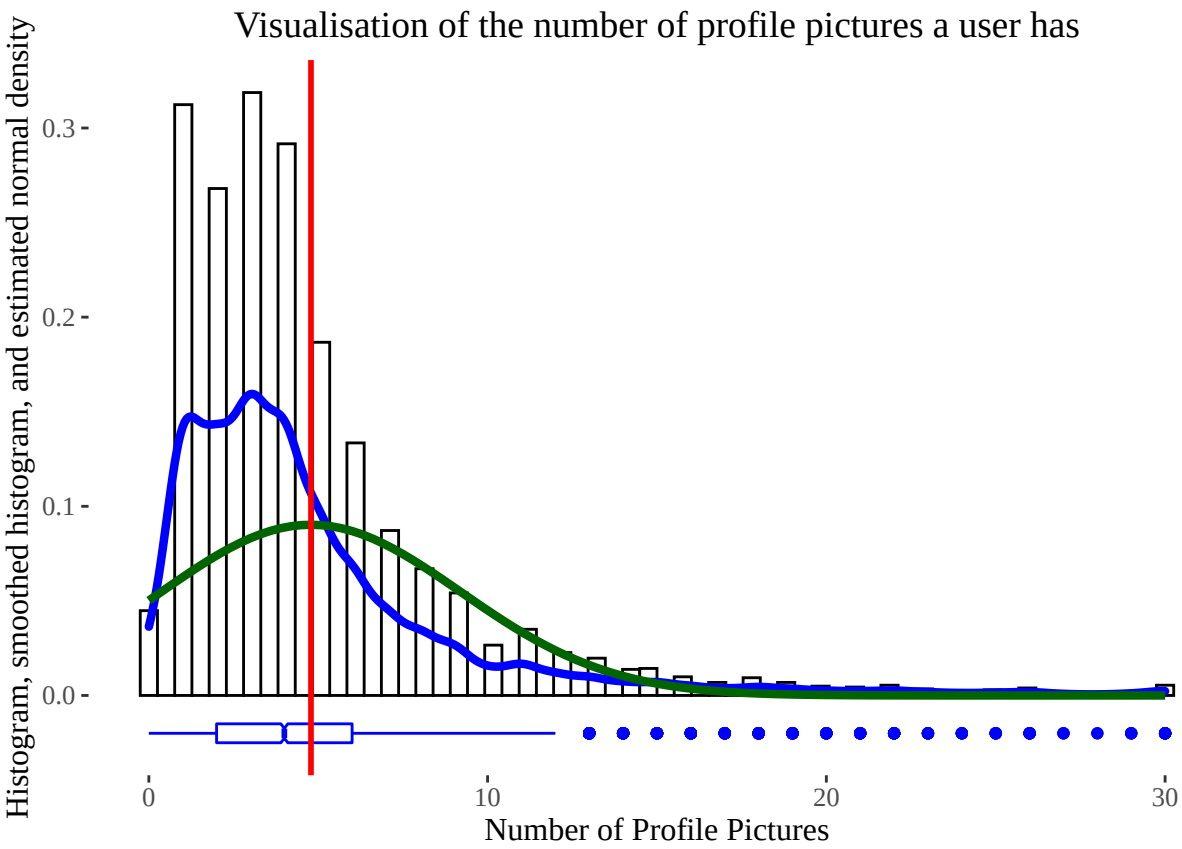


Figure 2: Profile pictures

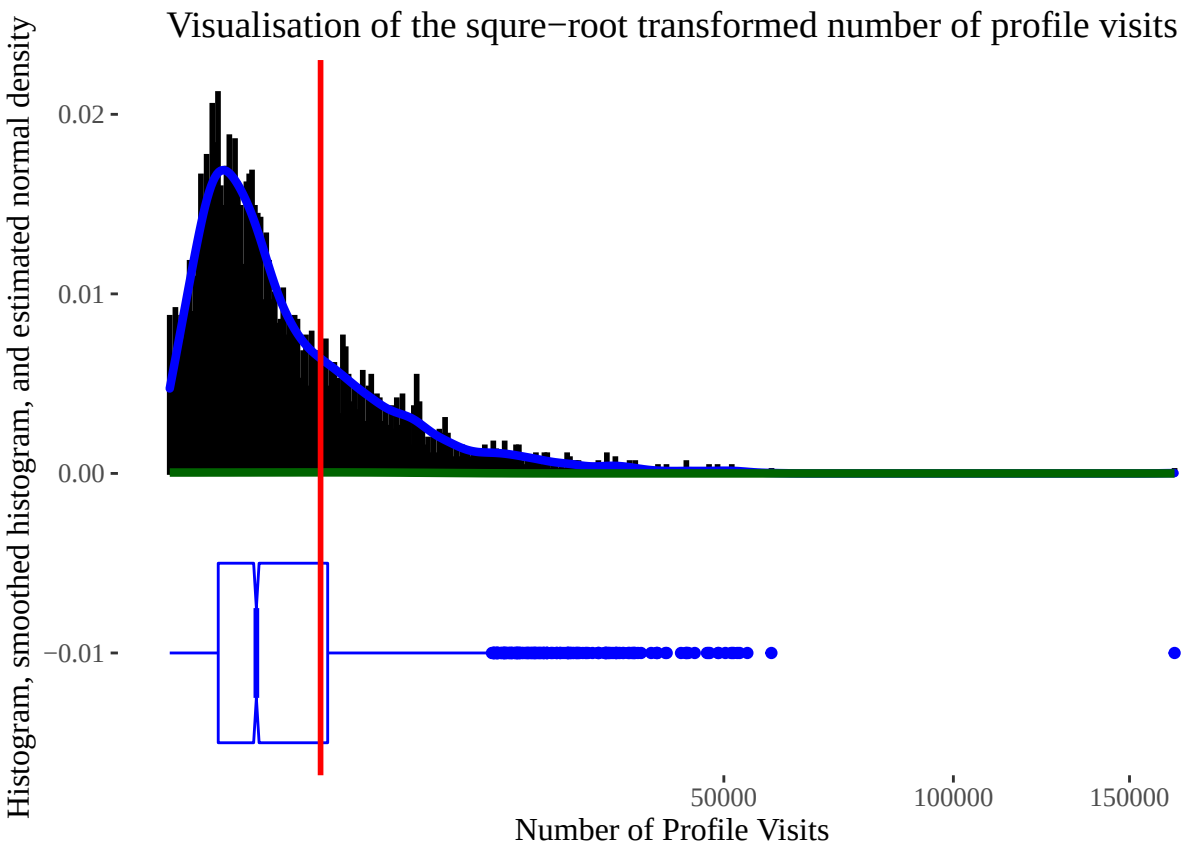


Figure 3: Square-root transformed Profile visits

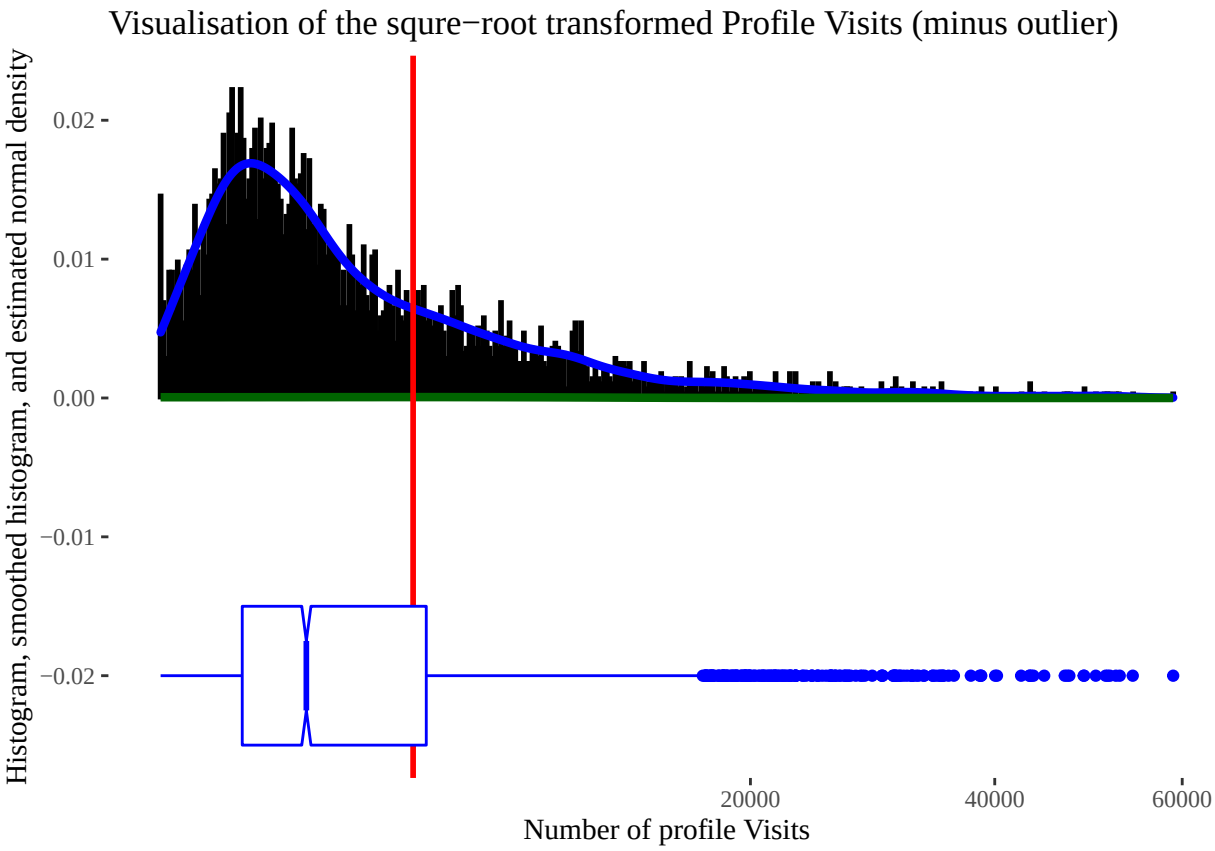


Figure 4: Square-root transformed Profile visits - minus the outlier

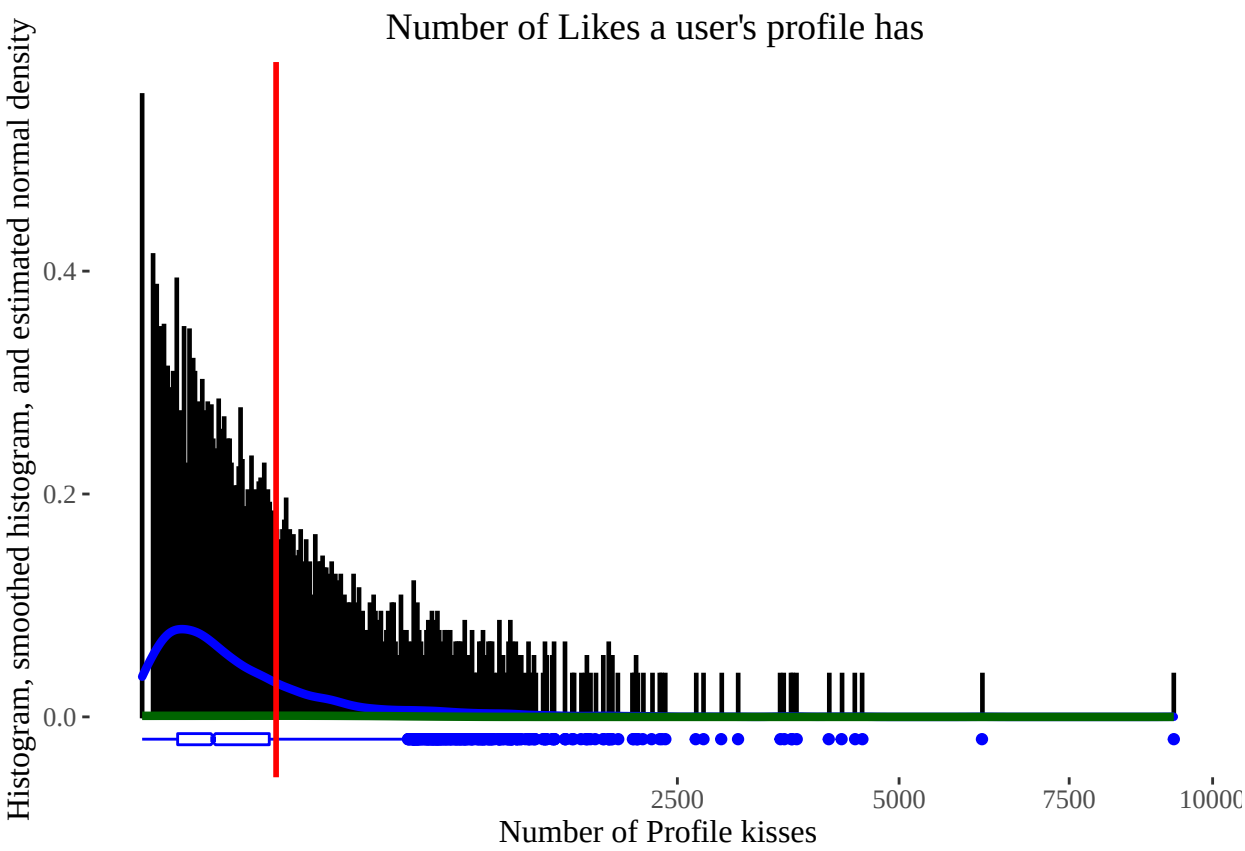


Figure 5: Profile Likes

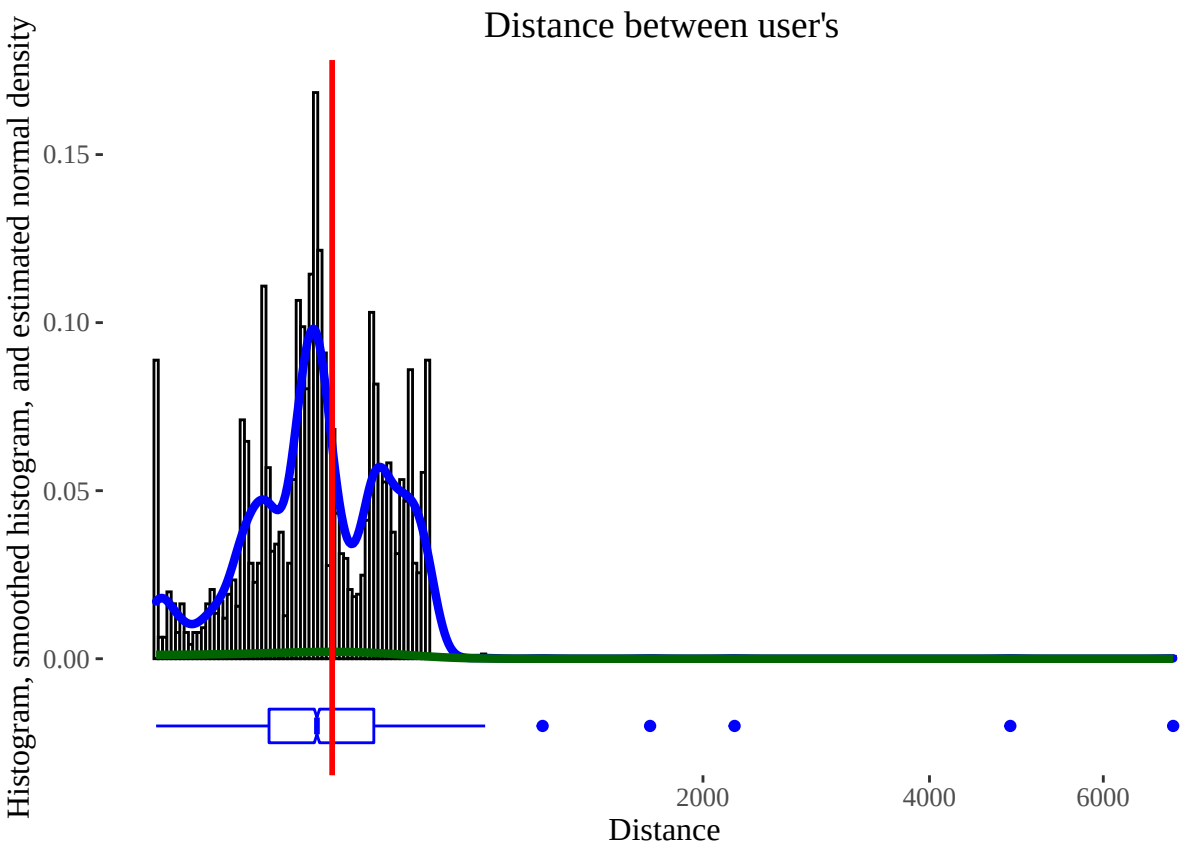


Figure 6: Distance

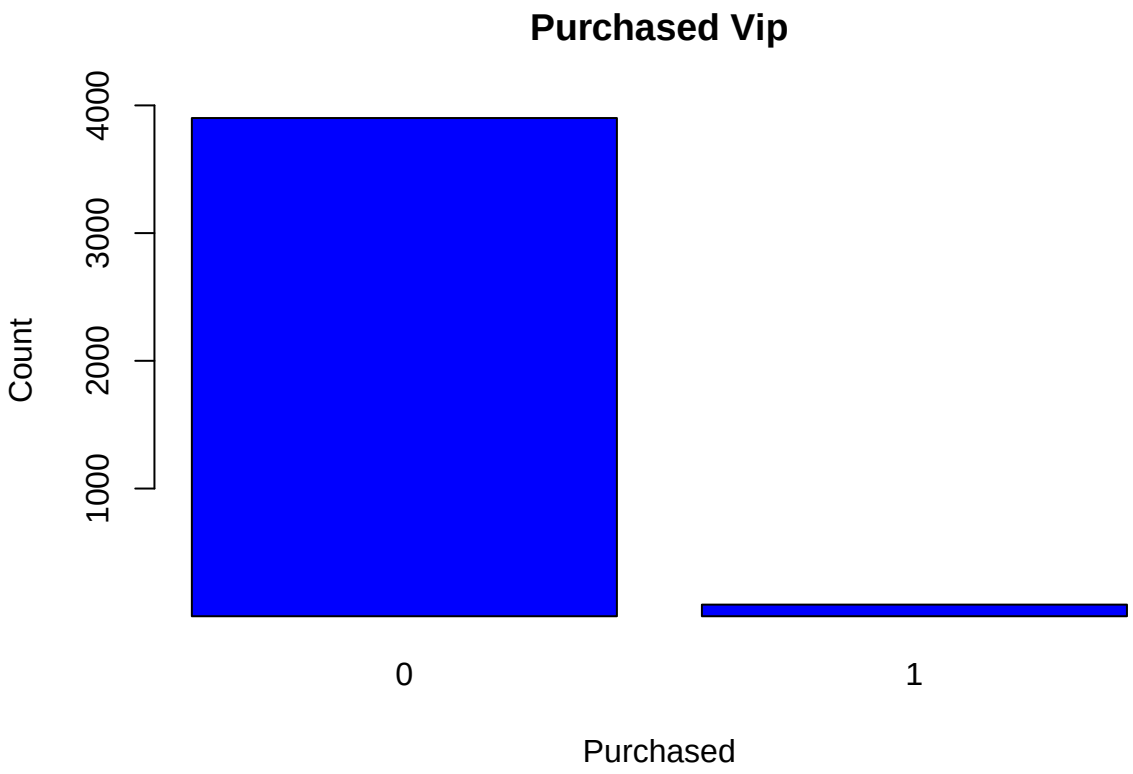


Figure 7: Count of whether a user is a VIP (nor not)

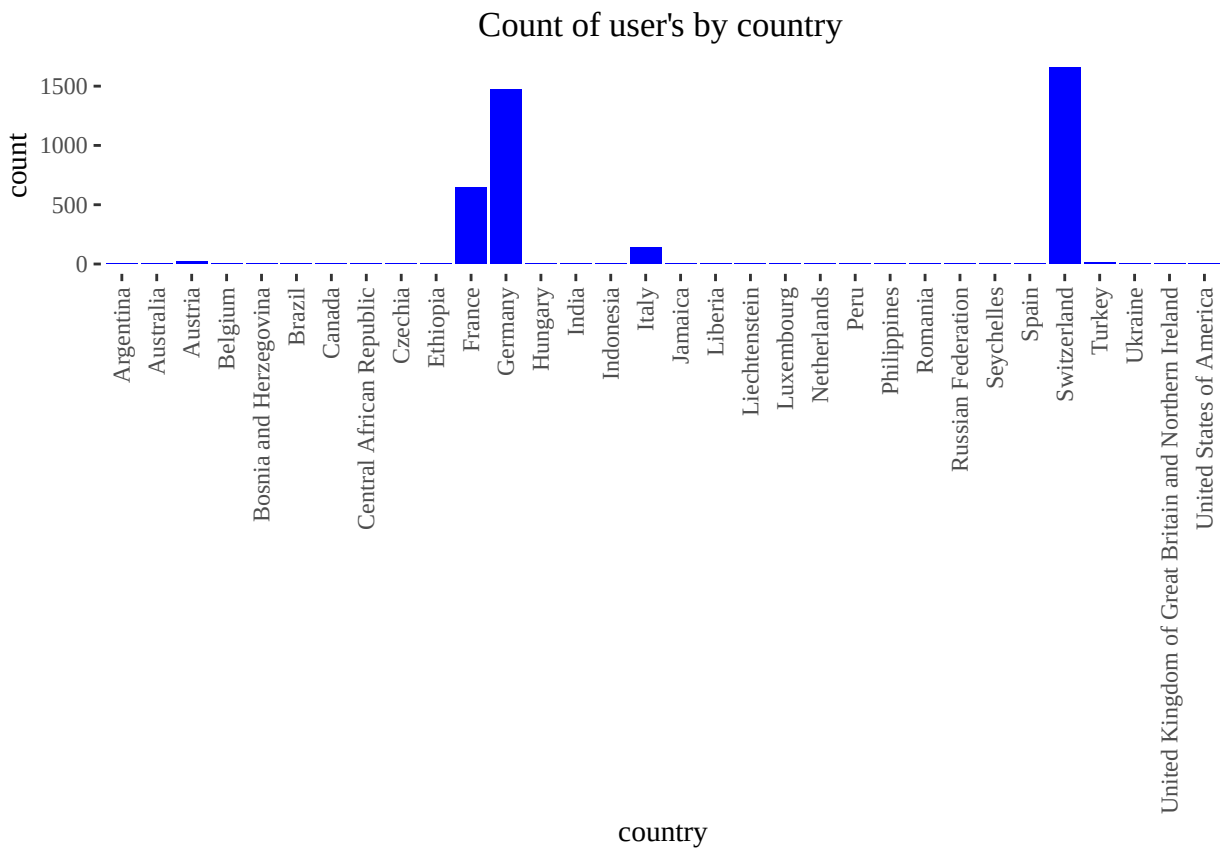


Figure 8: Number of user's by Country

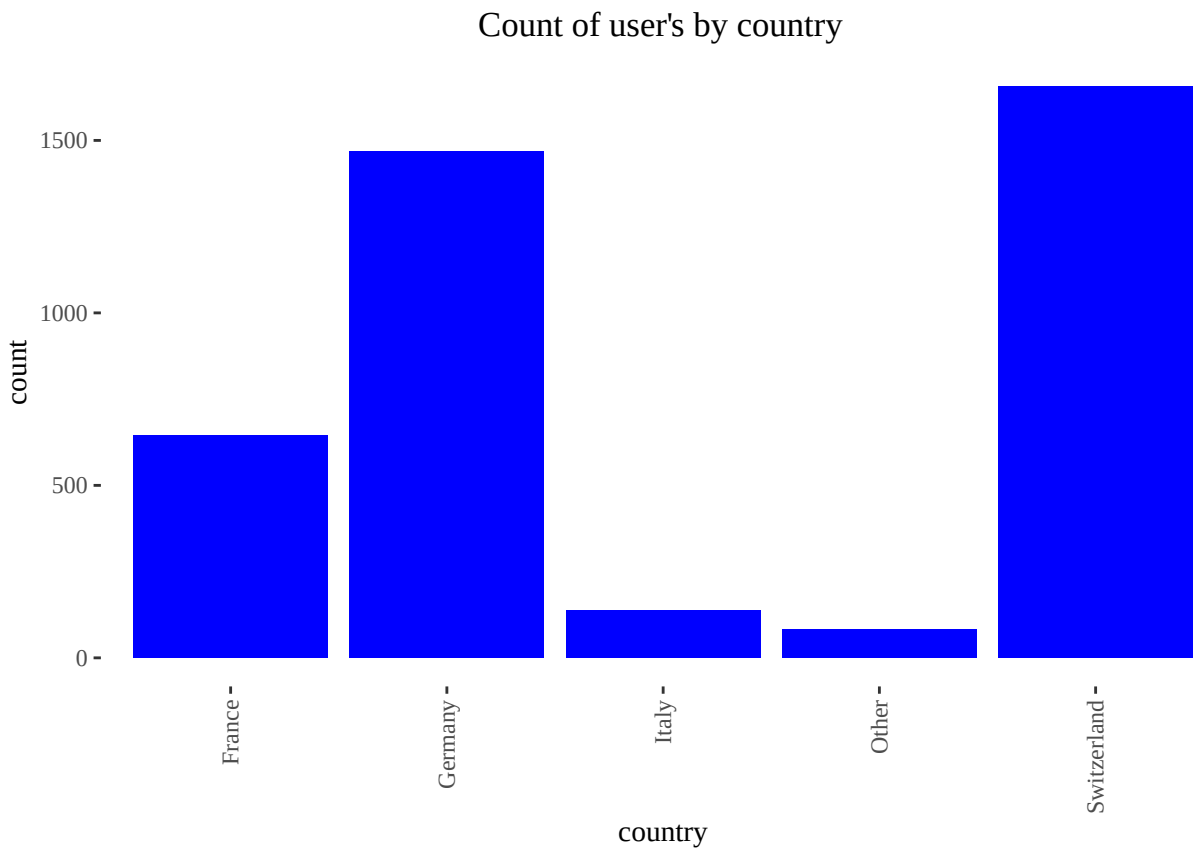


Figure 9: Number of user's by Country

between age and likes, profile visits and pictures, nor distance and likes, profile visits and pictures. This can been seen in figure 10 and numerically in table 5, and supported by the pairs plots in figure 11.

Earlier we saw that there was significantly more non-Vip user's than Vip users (table 2). In shows that 3901 users are not Vip's while only 91 are Vip's. We can see in table 6 that Switzerland has the most Vip's, this is supported in figure 12.

We see that there is a high correlation (0.89) between profile visits and profile likes (table 5). We see from figure 14 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 14 we see the same positive relationship but the outlieing profiles are gone (figure 15).

We see that the number of users by age tends to increase in Germany and peak at 22 years old before decreasing. Where as in France and Switzerland they tend to start with a larger number of users at younger age and decrease as age increases. This can be seen in figure 13.

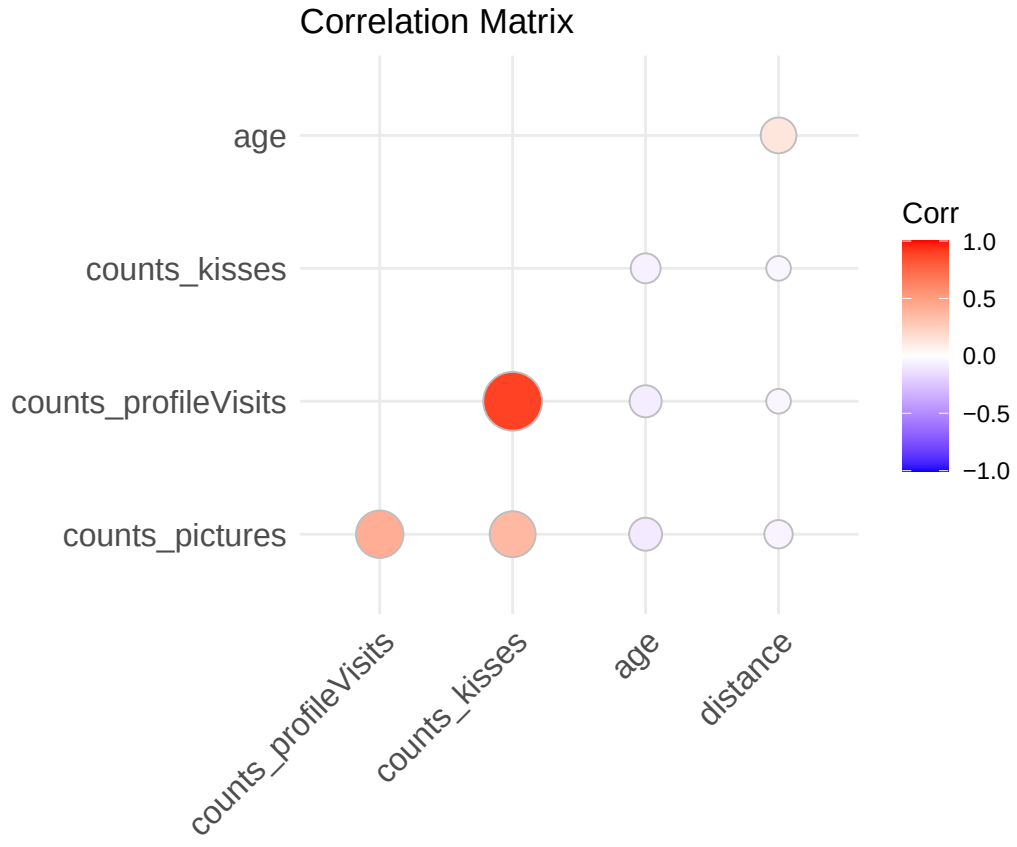


Figure 10: Correlation Matrix

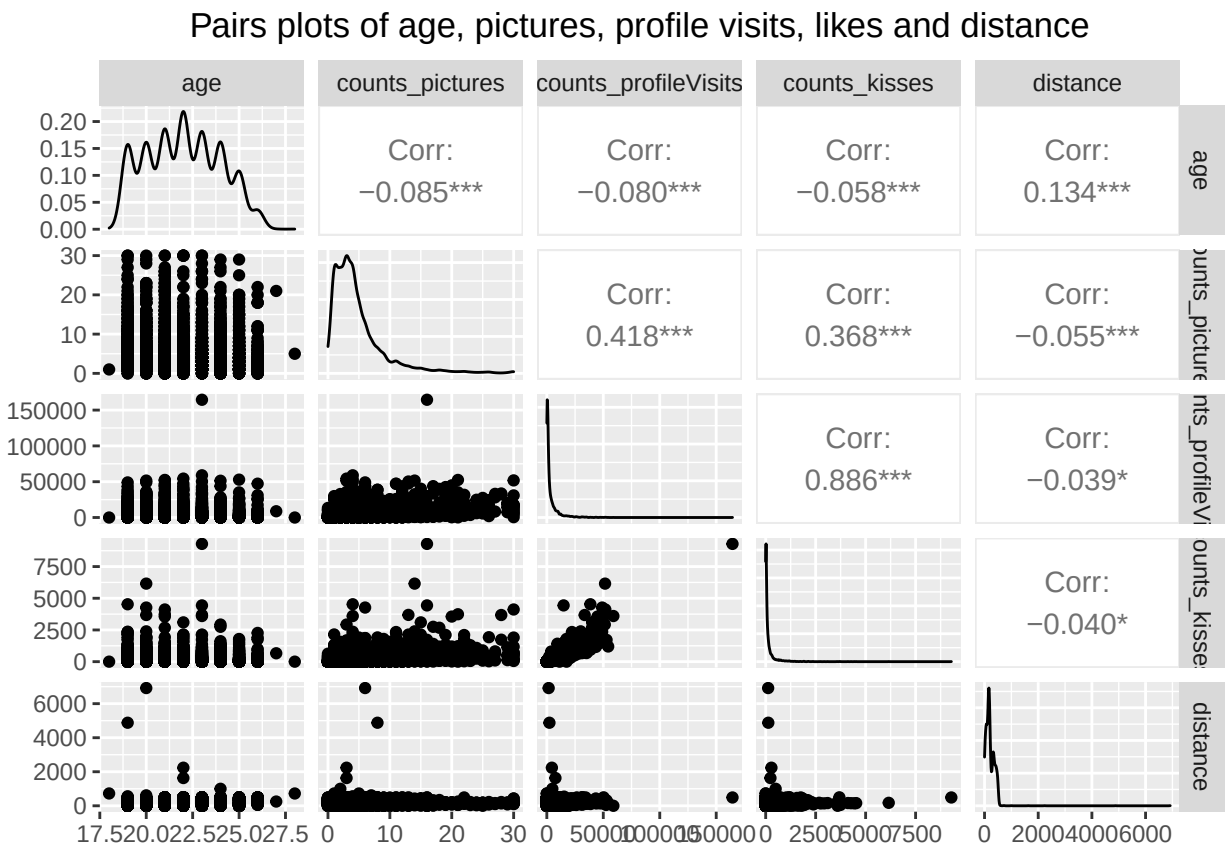


Figure 11: Pairs plot

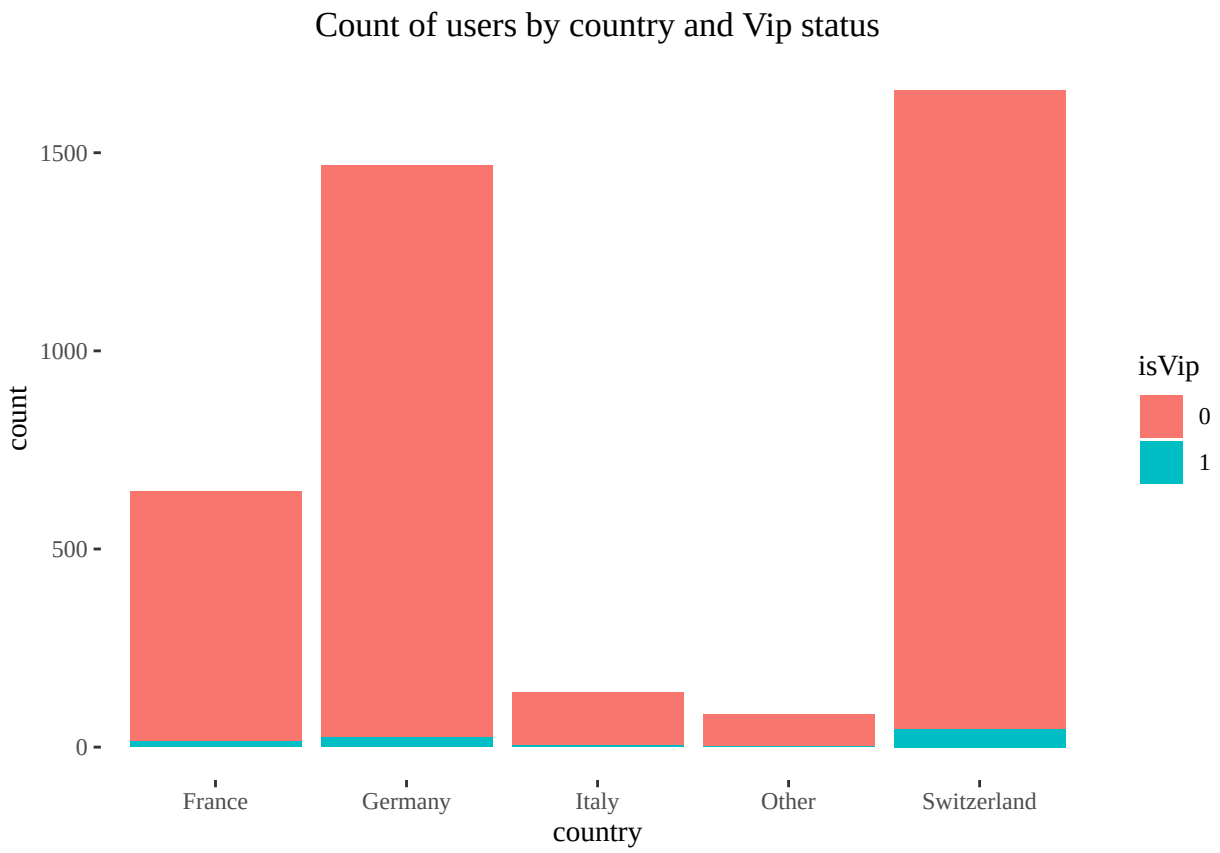


Figure 12: Count of users by country and Vip status

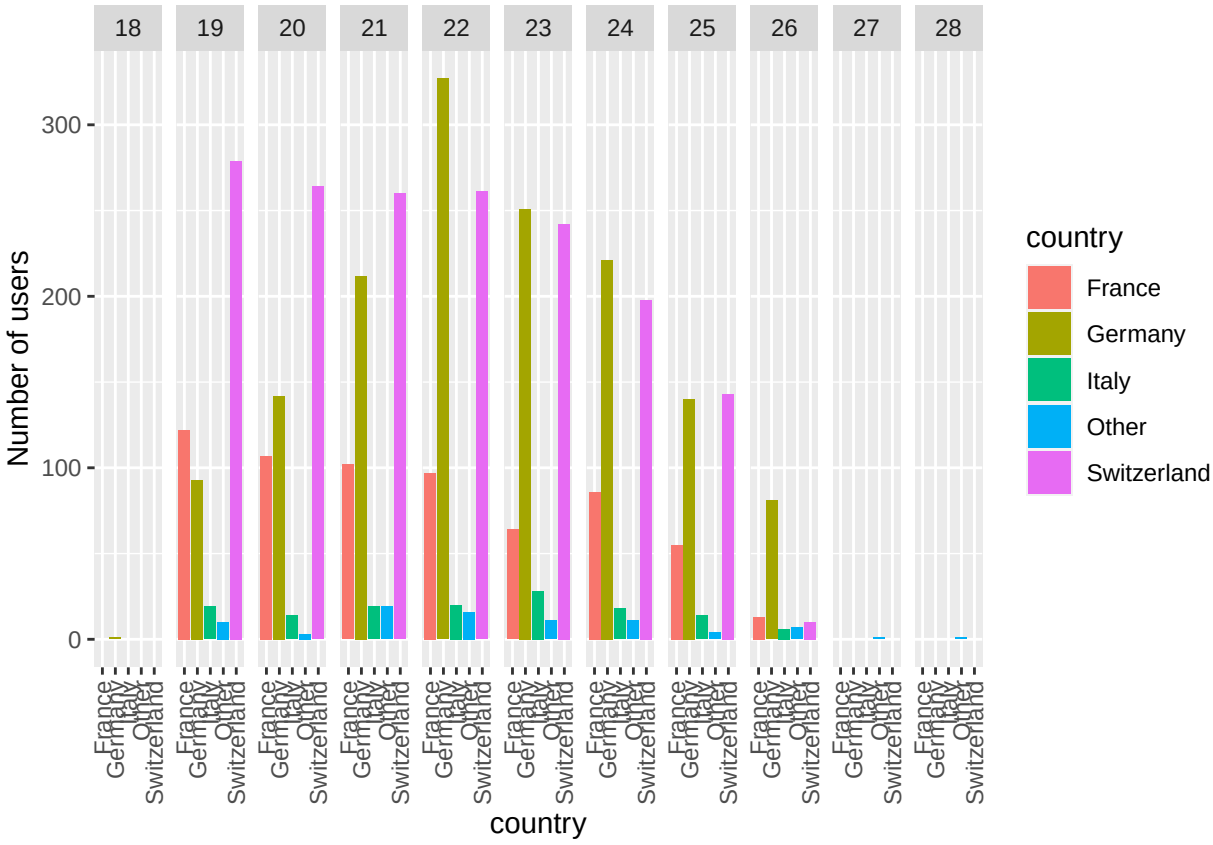


Figure 13: Number of users by age and country

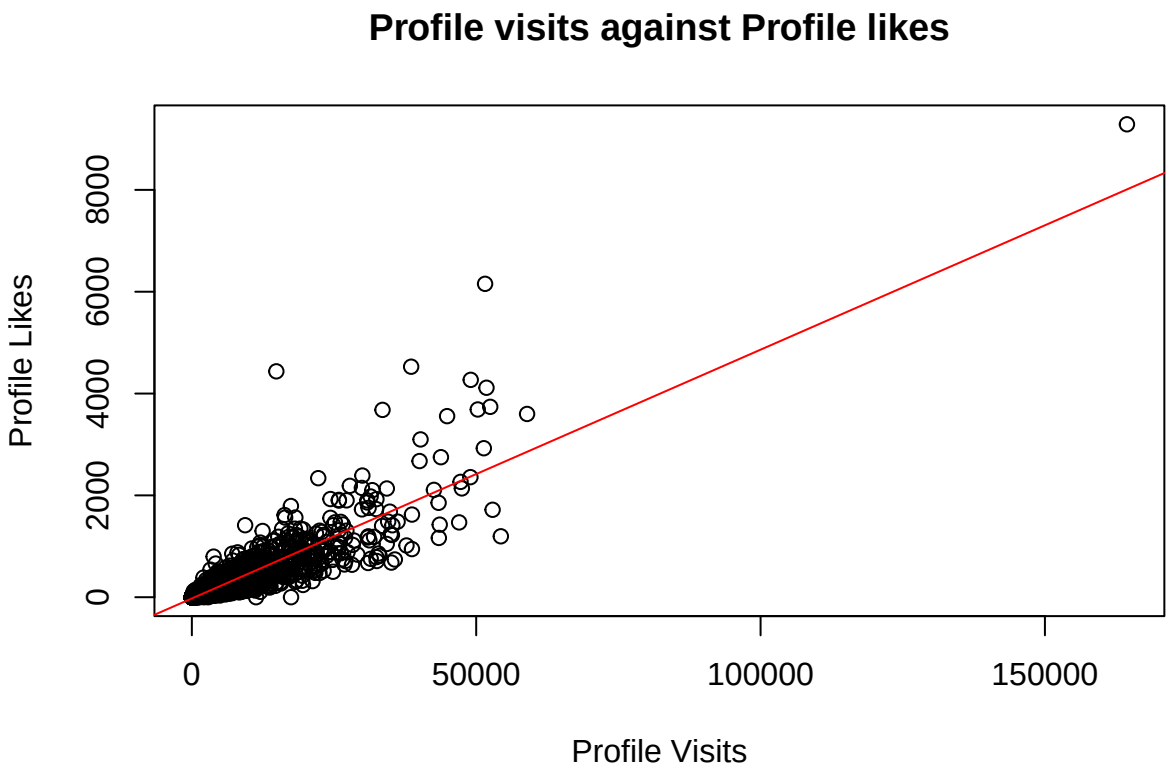


Figure 14: Scatterplot of profile visits vs profile likes

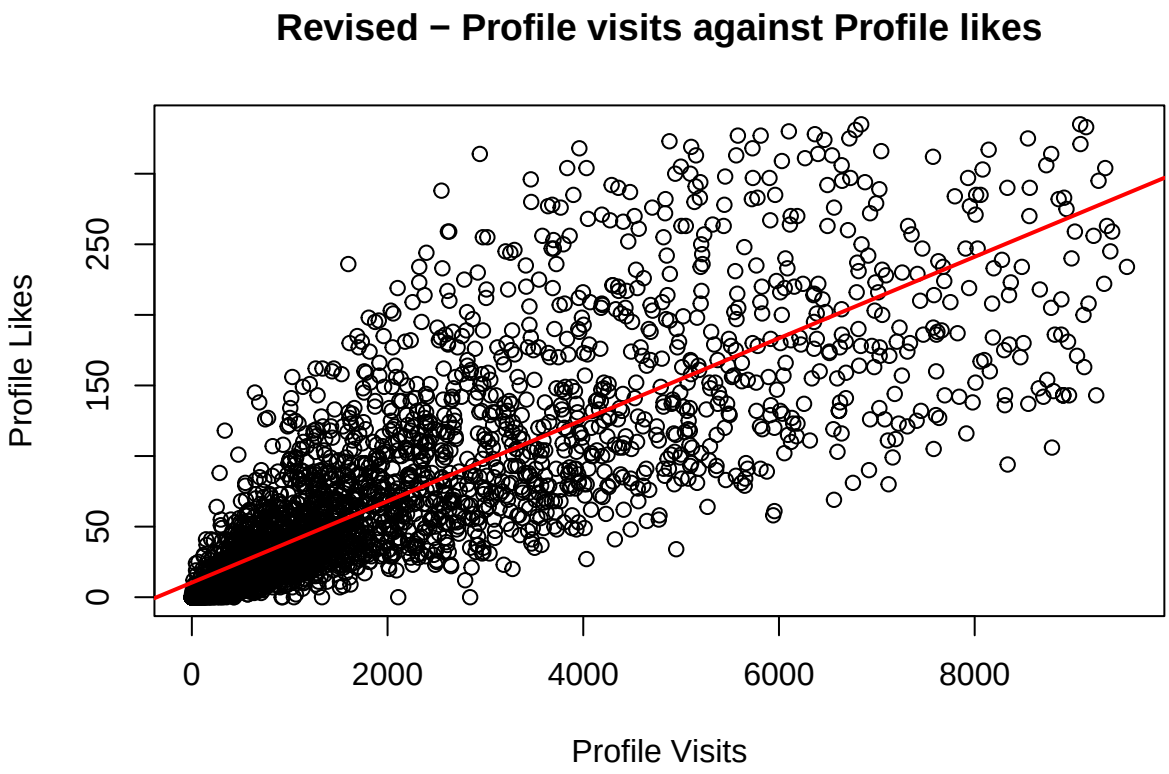


Figure 15: Revised scatterplot of profile visits vs profile likes

4 Burning questions

4.1 Profile Visits, Like, Country and Vip

After our initial analysis we know that profile visits and profile likes are highly correlated. We want to know if the number of profile visits and likes varies from country. We are also interested to know if having Vip status in each country affects the number of profile visits and likes.

There appears to be a positive correlation between the number of profile pictures a user has and the number of profile visits they receive, as well as the number of likes the profile receives. It appears that having Vip status has a higher correlation in these cases. The number of likes on a profile and the number of profile visits is strongly positively correlated, this is only slightly higher in Vip's compared to non-Vip's as seen in figure 16. The correlation between countries for these variables varies a little. The biggest thing to note is that Italy appears to have a correlation of 0.947 between the number of likes a user has and the number of visits (figure 17).

The boxplot of the number of profile visits and likes by country is very hard to read (figure 18), we have transformed this by taking the log of profile likes. We can see from th revised boxplot (figure 18) that Germany has the largest range of profile visits while France has the least. France, Germany and Italy have very similar ranges of the number of profile likes. Our group of 'other' countries appears to have the largest median for the number of profile likes. We can conclude that there is a difference in the number of profile visits and the number of profile likes for each country.

Looking at figure 20 we can see that having Vip status tends to improve the number of profile likes a user receives in our four main countries. However this is not the case in the 'other' countries as only 1 person from 'other' has Vip status. The biggest difference can be seen in Germany.

To conclude if you live in Germany and you want more profile likes then you should like to purchase Vip status.

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

```
## Warning: Removed 212 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 212 rows containing non-finite values (stat_boxplot).
```

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

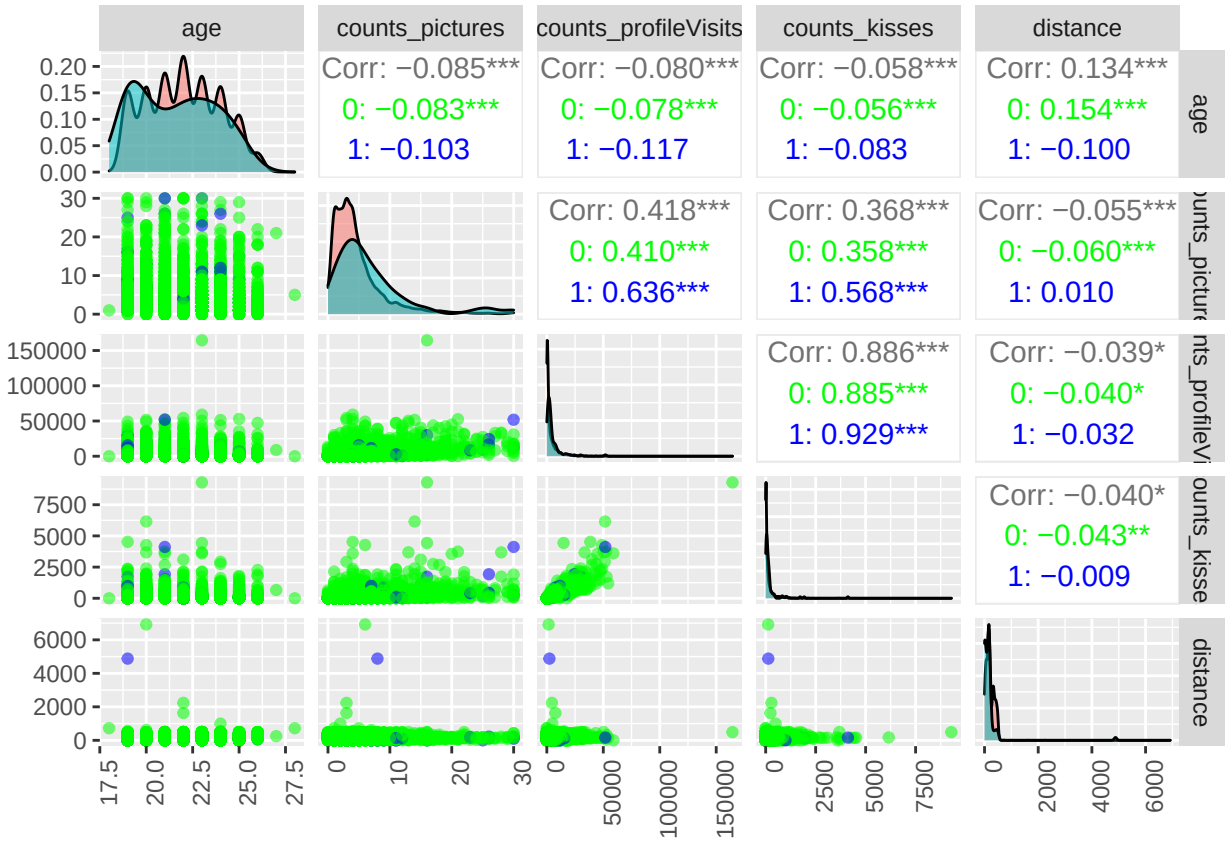


Figure 16: Pairs plot by isVip

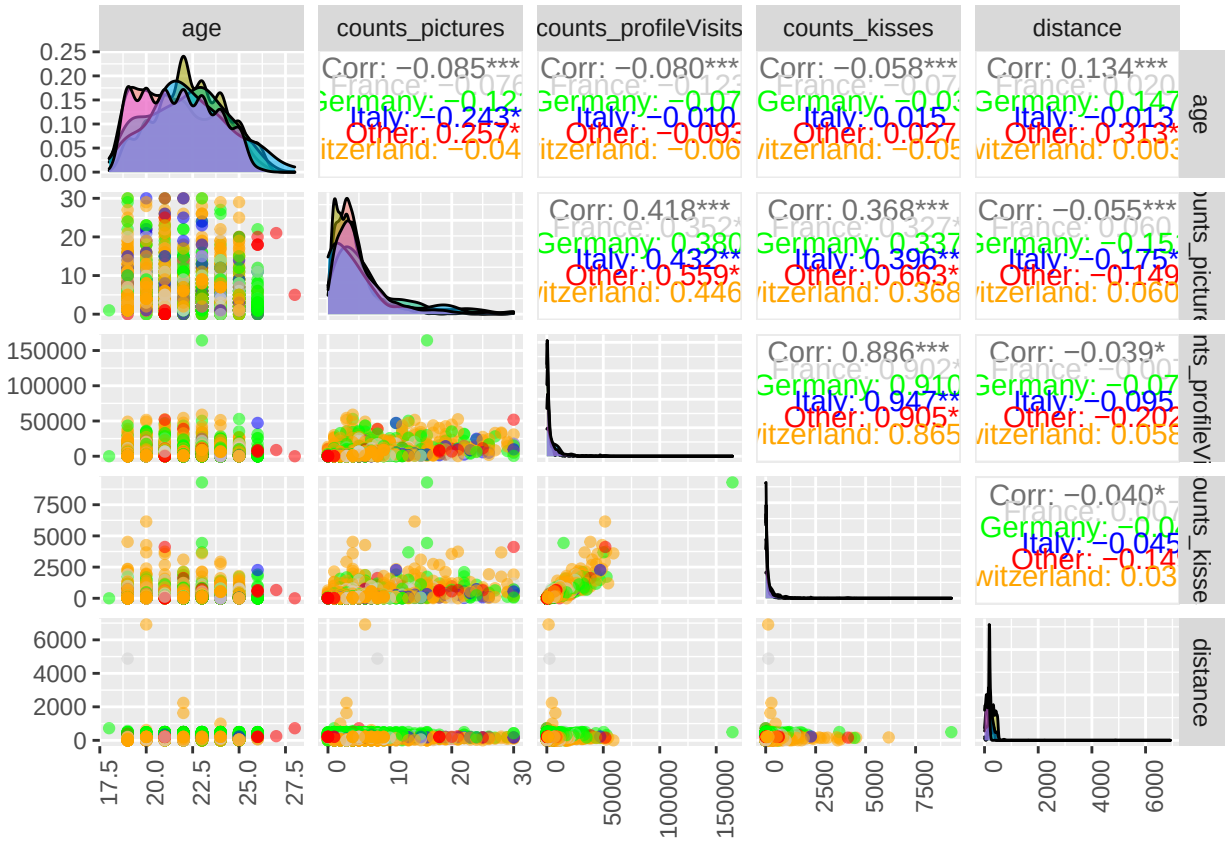


Figure 17: Pairs plot by country

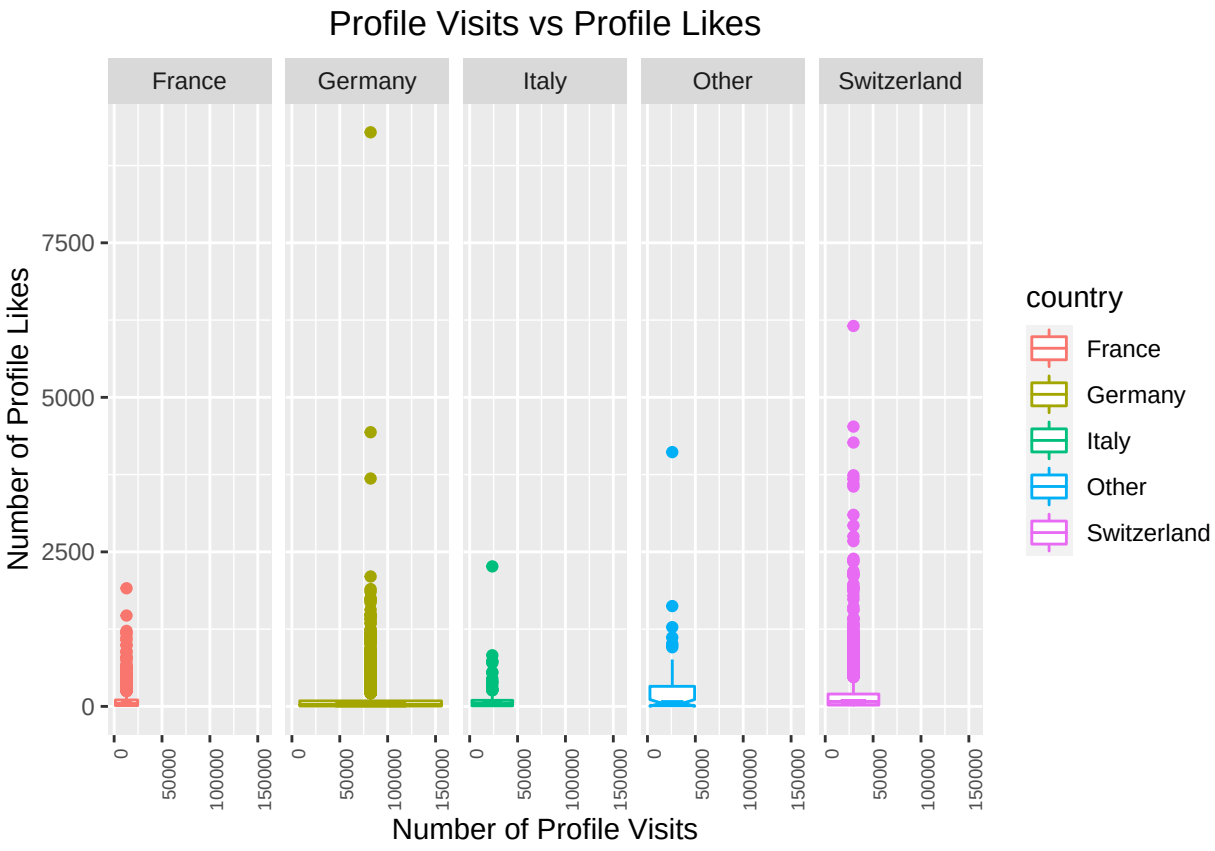


Figure 18: Boxplot of profile visits and likes by country

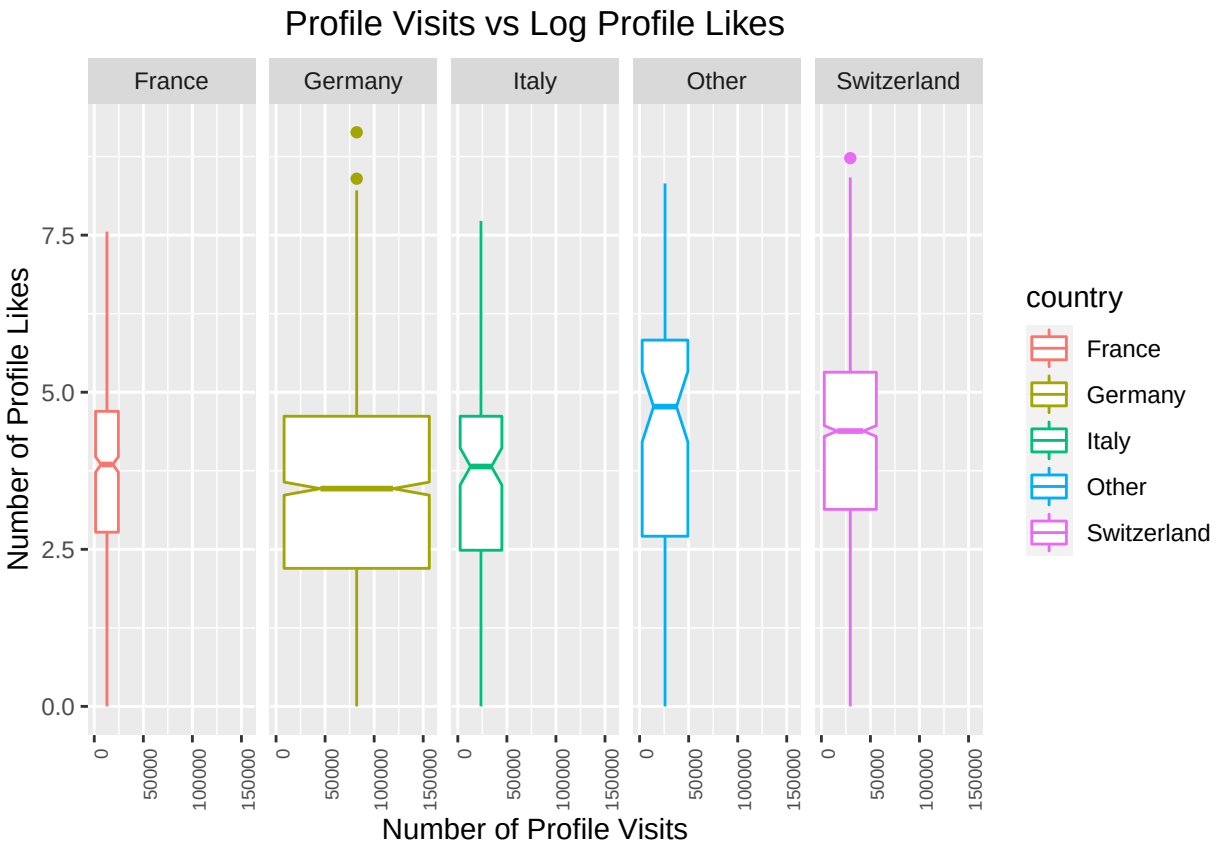


Figure 19: Revised boxplot of profile visits and likes by country

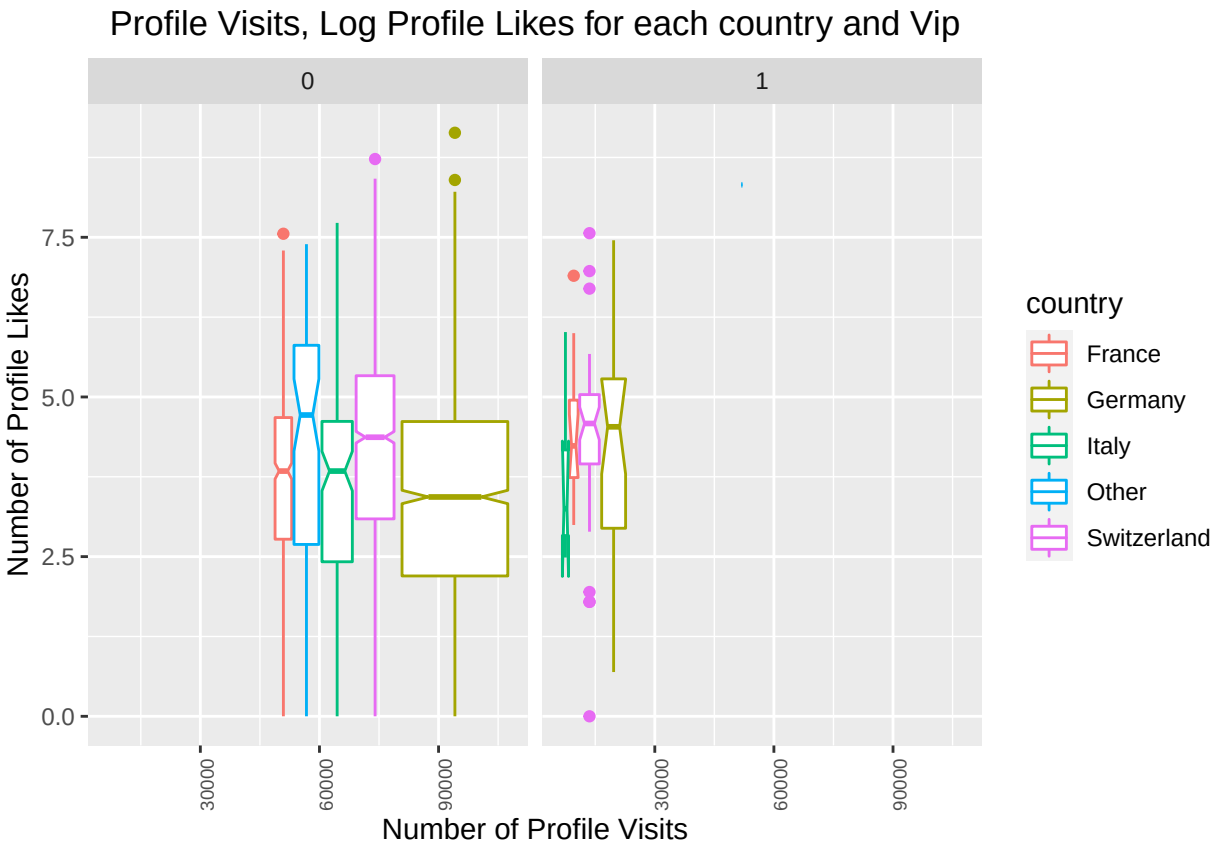


Figure 20: Revised boxplot of profile visits and likes by country and vip

5 Normality Test

5.1 Shapiro-Wilks test of normality

```
##
##  Shapiro-Wilk normality test
##
## data:  isavip$age
## W = 0.90264, p-value = 4.786e-06

##
##  Shapiro-Wilk normality test
##
## data:  isavip$counts_pictures
## W = 0.77427, p-value = 1.668e-10

##
##  Shapiro-Wilk normality test
##
## data:  isavip$counts_profileVisits
## W = 0.58171, p-value = 1.056e-14

##
##  Shapiro-Wilk normality test
##
## data:  isavip$counts_kisses
## W = 0.38621, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  isavip$distance
## W = 0.22572, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  isnotavip$age
## W = 0.94802, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  isnotavip$counts_pictures
## W = 0.76462, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  isnotavip$counts_profileVisits
## W = 0.52133, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  isnotavip$counts_kisses
## W = 0.39004, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  isnotavip$distance
## W = 0.64216, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Switzerland$age
## W = 0.92955, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Switzerland$counts_pictures
## W = 0.77832, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Switzerland$counts_profileVisits
## W = 0.62944, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Switzerland$counts_kisses
## W = 0.46077, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Switzerland$distance
## W = 0.19327, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Germany$age
## W = 0.96006, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Germany$counts_pictures
## W = 0.75416, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Germany$counts_profileVisits
## W = 0.40898, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Germany$counts_kisses
## W = 0.28469, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Germany$distance
## W = 0.94874, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  France$age
## W = 0.92135, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  France$counts_pictures
## W = 0.79753, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  France$counts_profileVisits
## W = 0.62582, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  France$counts_kisses
## W = 0.53568, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  France$distance
## W = 0.3446, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Italy$age
## W = 0.9435, p-value = 2.16e-05

##
##  Shapiro-Wilk normality test
##
## data:  Italy$counts_pictures
## W = 0.80873, p-value = 3.904e-12

##
##  Shapiro-Wilk normality test
##
## data:  Italy$counts_profileVisits
## W = 0.49705, p-value < 2.2e-16
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Italy$counts_kisses
## W = 0.43292, p-value < 2.2e-16

##
##  Shapiro-Wilk normality test
##
## data:  Italy$distance
## W = 0.85466, p-value = 2.45e-10

##
##  Shapiro-Wilk normality test
##
## data:  Other$age
## W = 0.9501, p-value = 0.002806

##
##  Shapiro-Wilk normality test
##
## data:  Other$counts_pictures
## W = 0.77225, p-value = 5.205e-10

##
##  Shapiro-Wilk normality test
##
## data:  Other$counts_profileVisits
## W = 0.65371, p-value = 1.086e-12

##
##  Shapiro-Wilk normality test
##
## data:  Other$counts_kisses
## W = 0.51699, p-value = 4.466e-15

##
##  Shapiro-Wilk normality test
##
## data:  Other$distance
## W = 0.95579, p-value = 0.006141
```

all tests of normality's p-values support rejecting the null hypothesis (H_0: the population follows a normal distribution.) So, we believe that none of the variables factors follow a normal distribution.

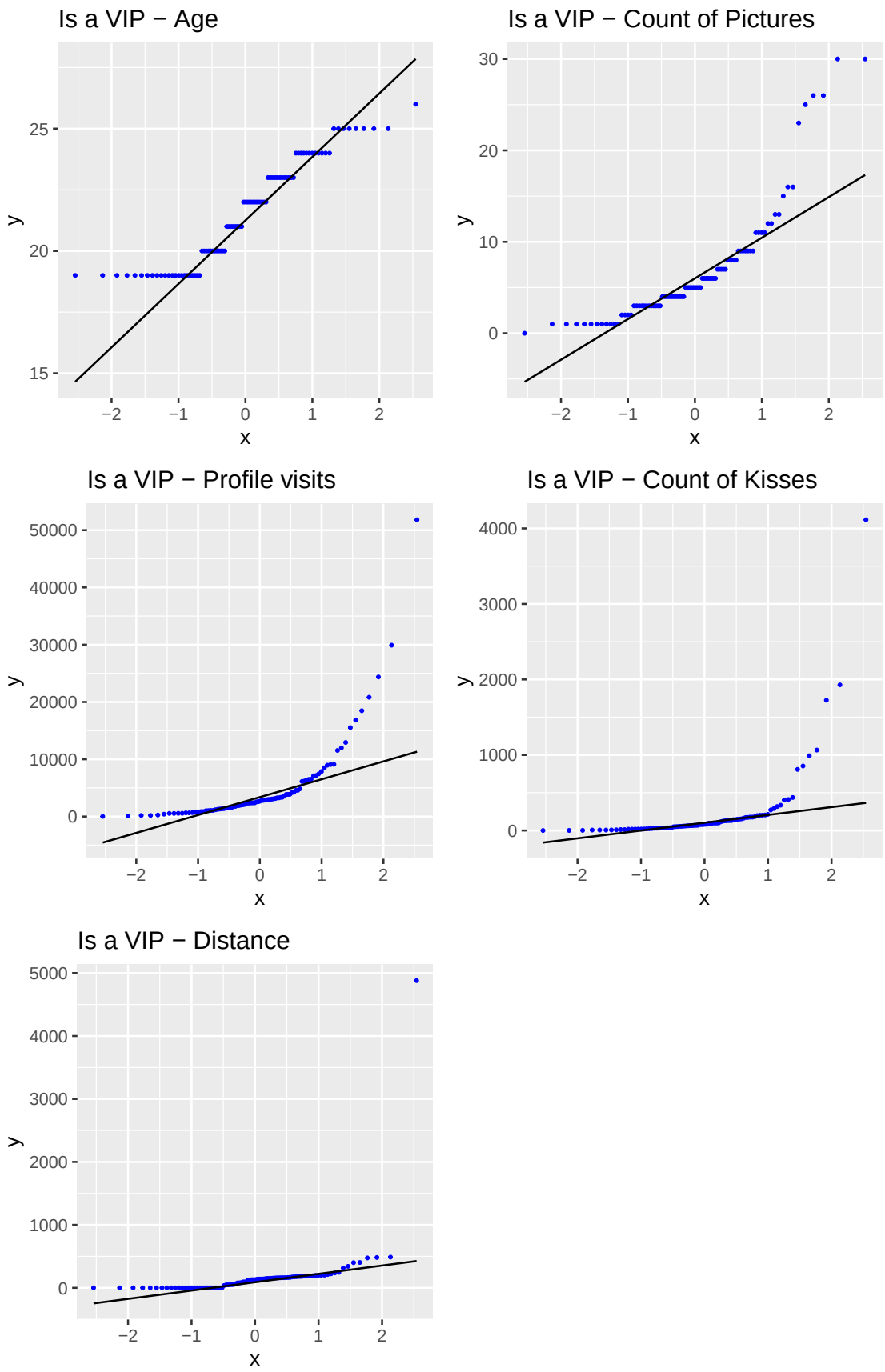


Figure 21: factor: isavip

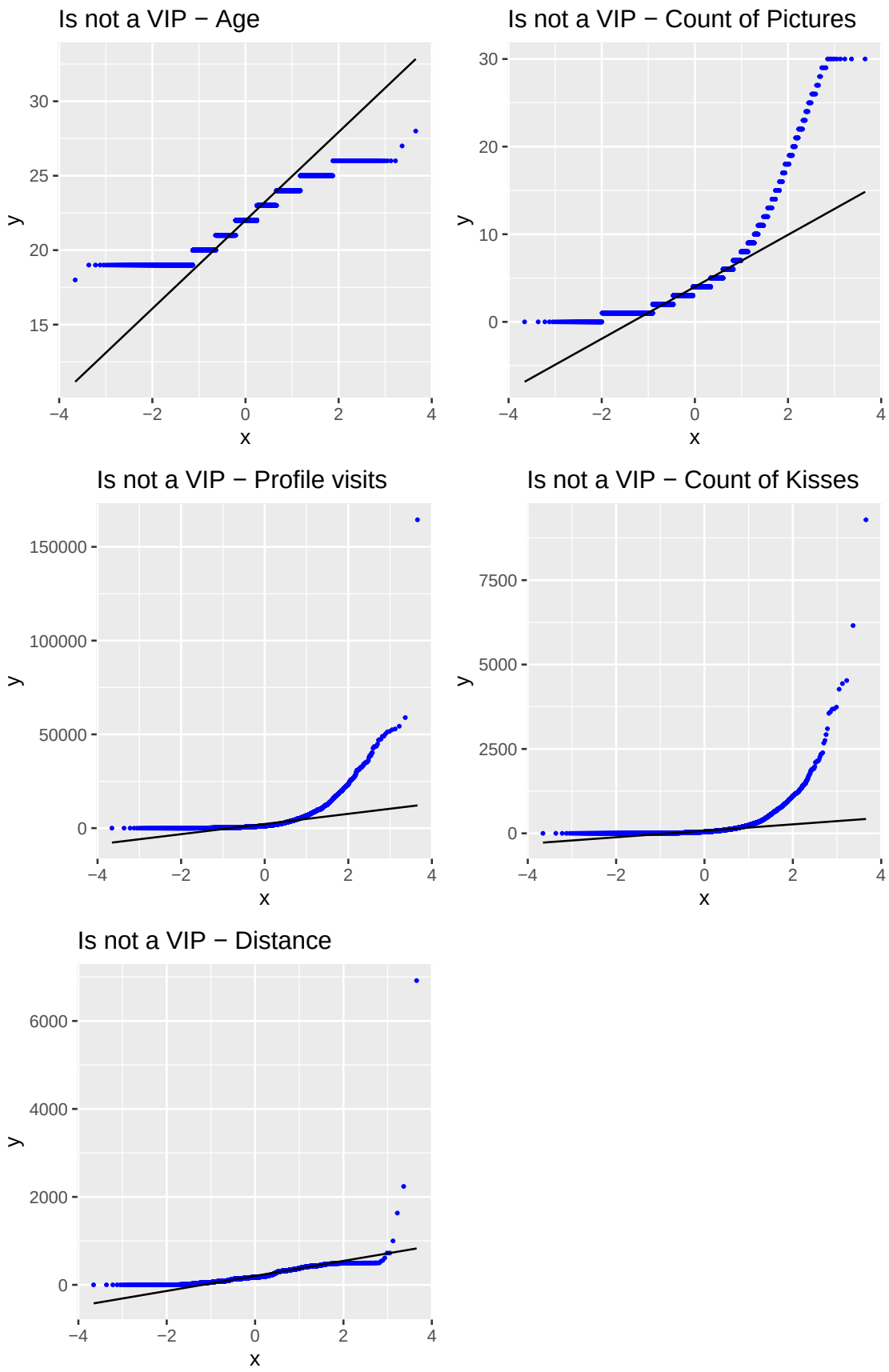


Figure 22: factor: isnotavip

Figure 23: factor: Switzerland

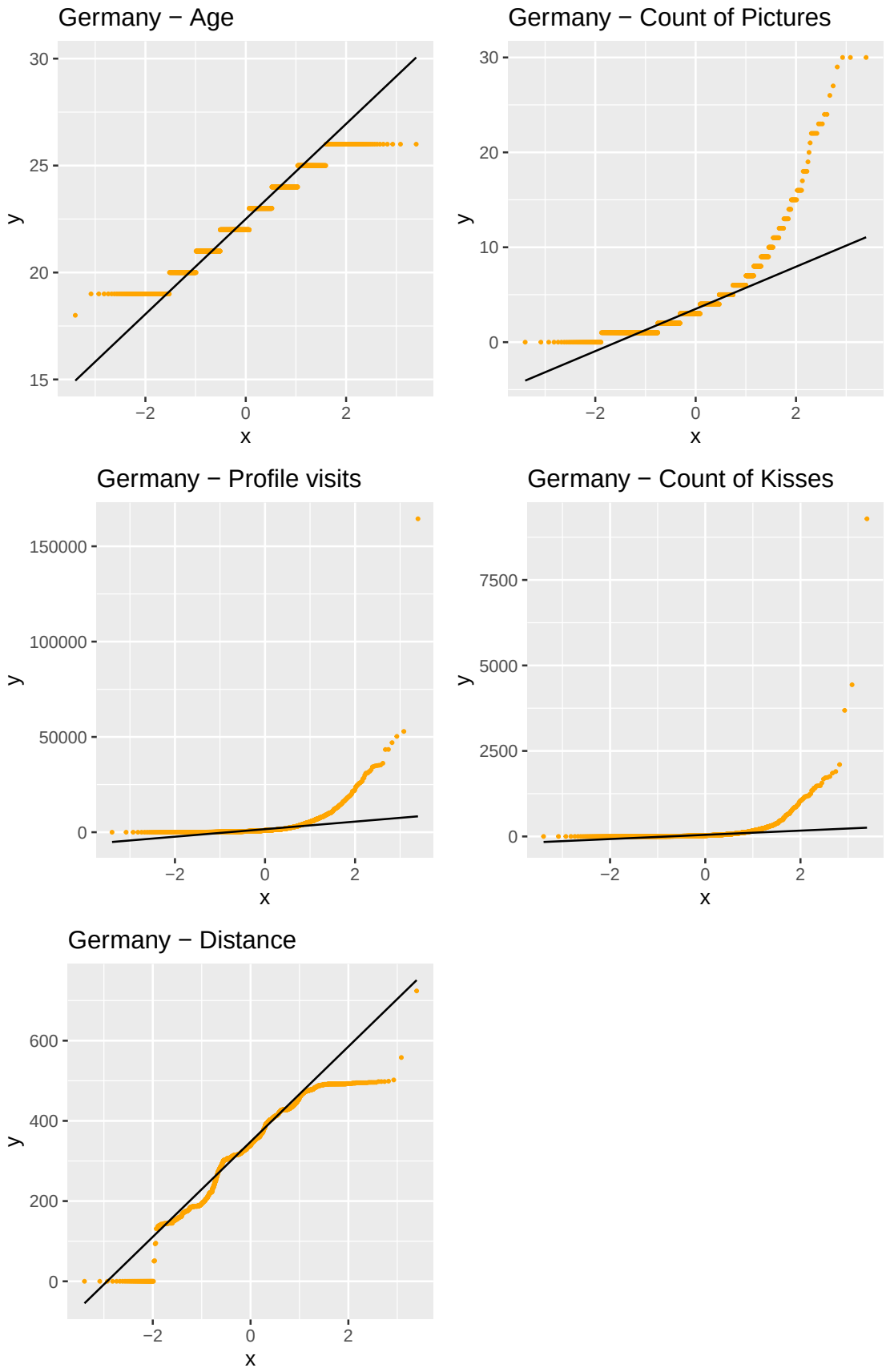


Figure 24: factor: Germany

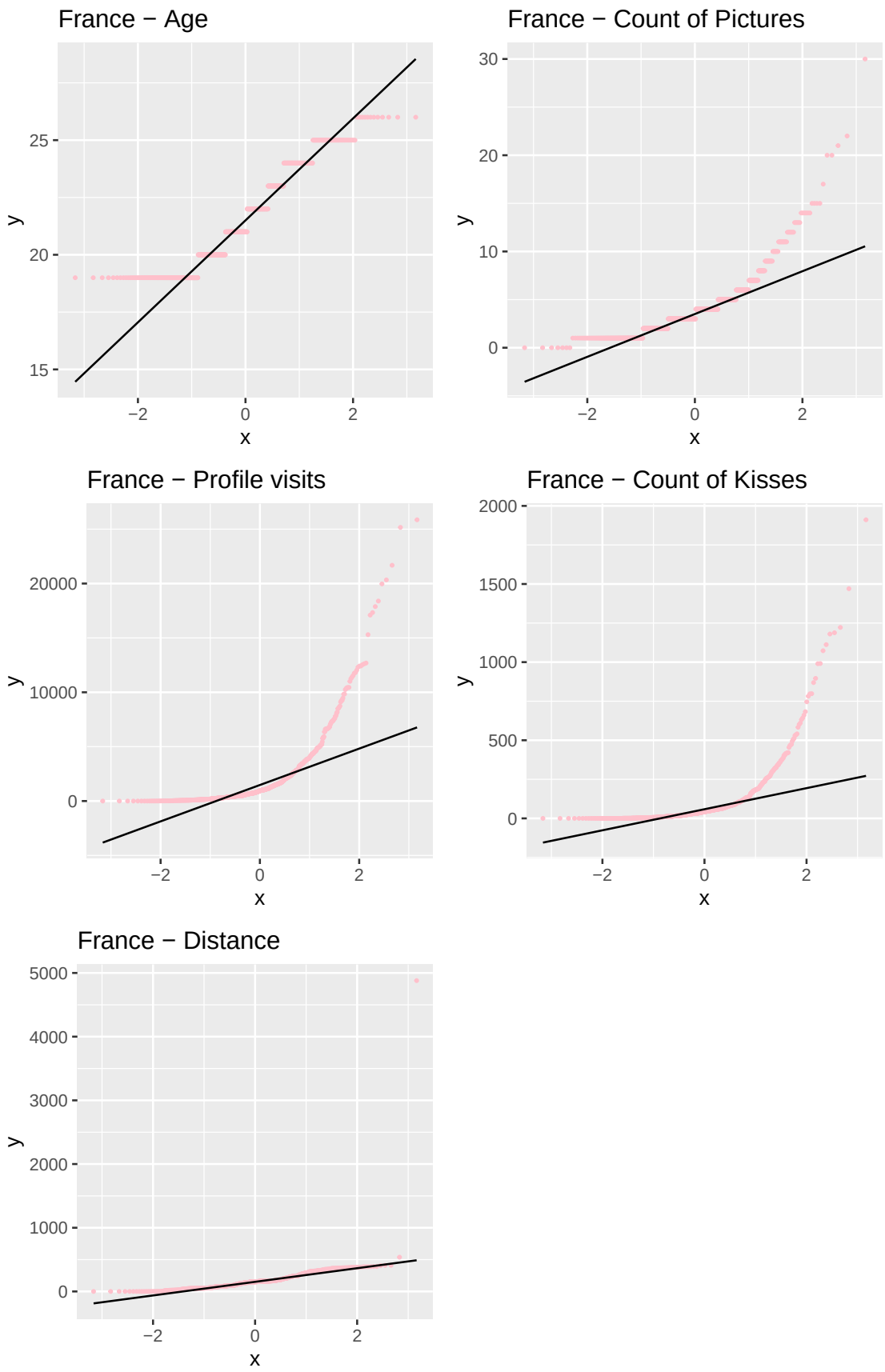


Figure 25: factor: France

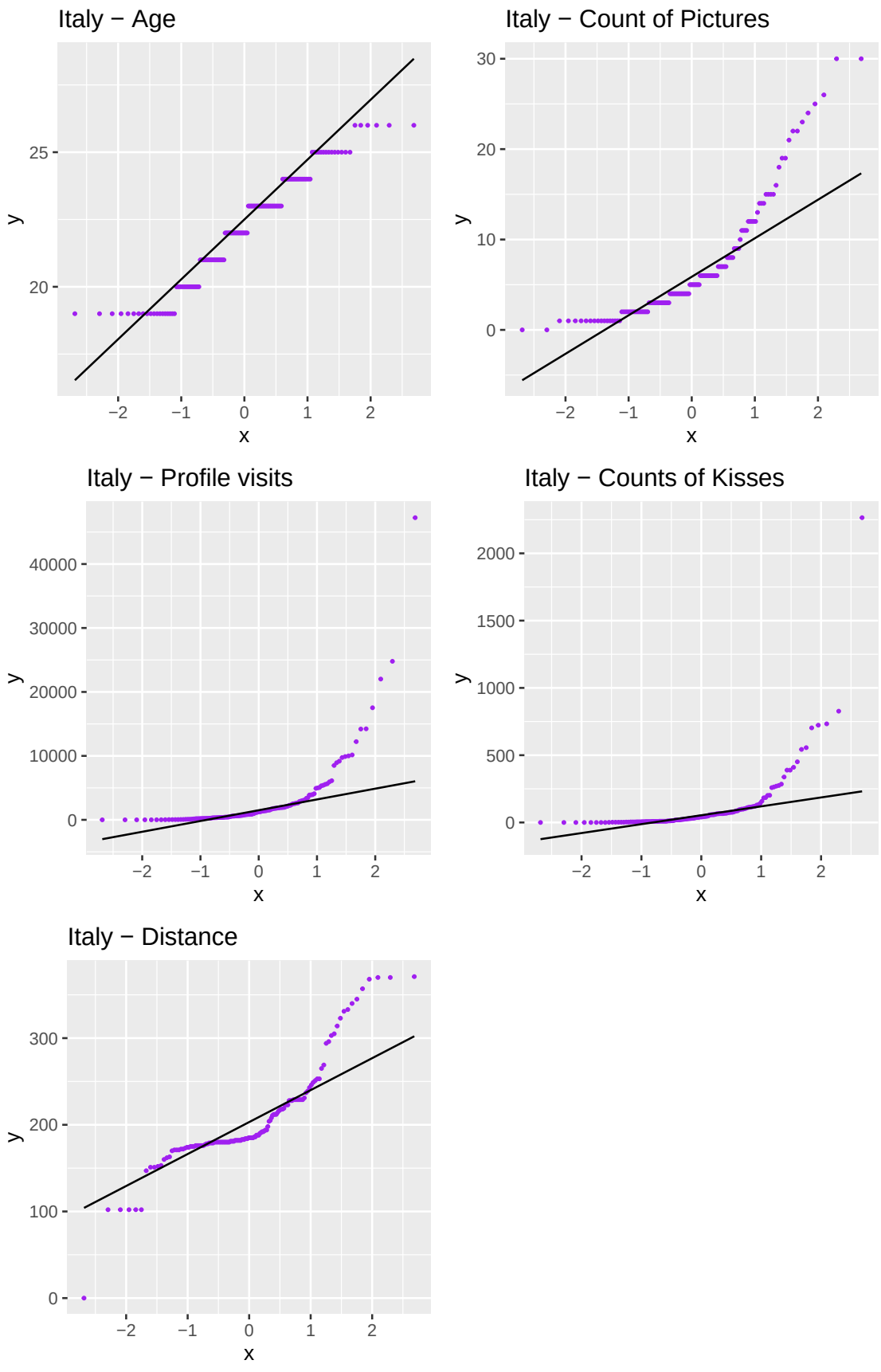
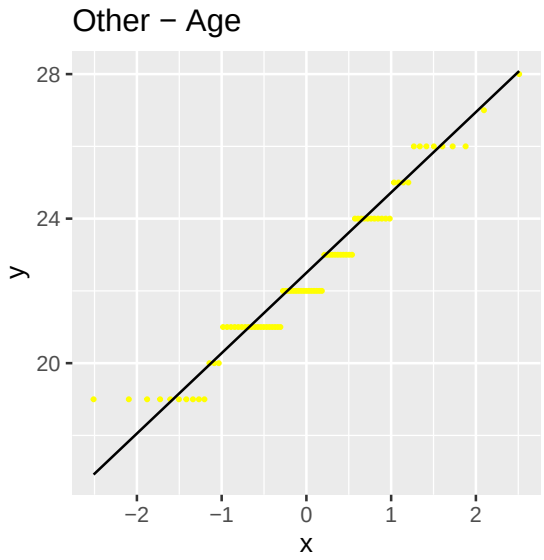
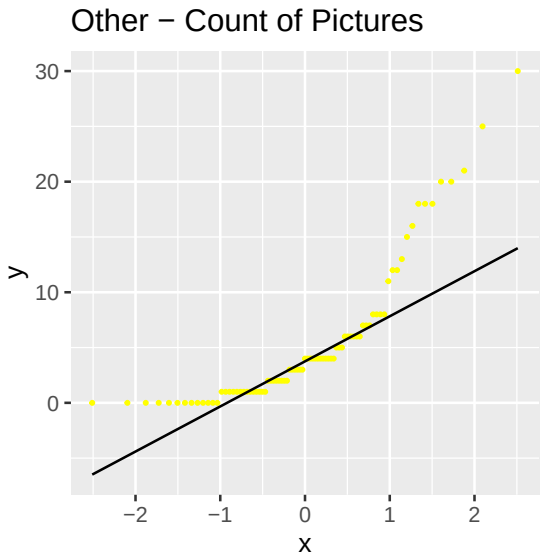
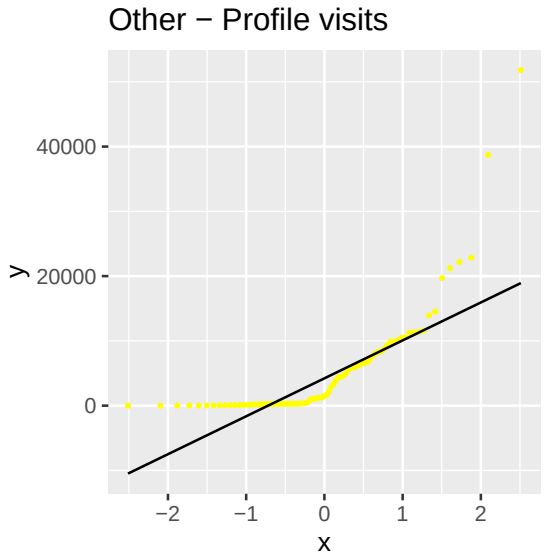


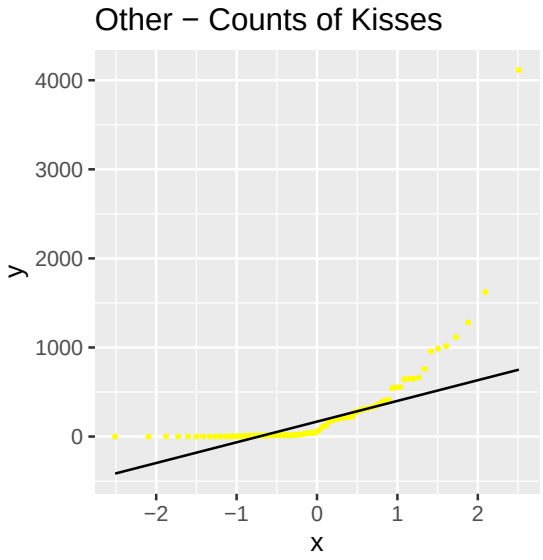
Figure 26: factor: Italy

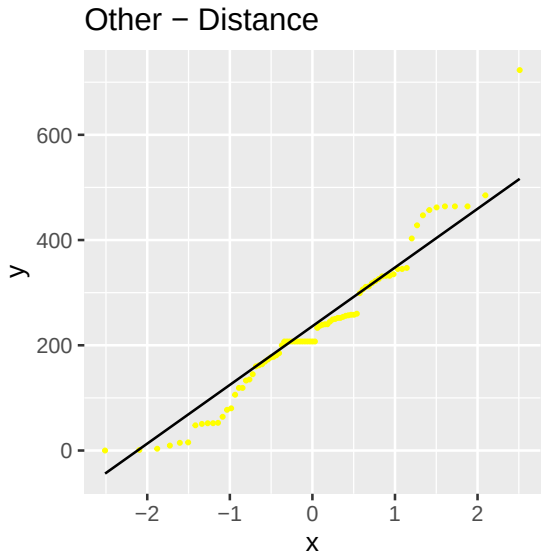
5.2 Quantile-Quantile Plots











All of the Q-Q plots support the quantitative analysis, the QQ points have points off the line in its tail which violate the assumption of normality.

6 Testing for class membership using Mahalanobis distance (isVip)

Firstly, we will subset the data based on whether or not the user is a VIP, then iterate through each observation and check it against the data as if it were a new sample. We will take it's measures (pictures, visits, and kisses) and delete it from the data set.

From this we will estimate the mean and covariance vectors of the measures to find the distance between the observation and the measurement model (using df = 3). The p-value of each distance will be recorded as that shows how significant the distance is and helps us to reject or accept the null hypothesis. Our alpha level is 0.05.

Hypothesis: + H0: x_{obs} is not consistent with the ideal distance + H1: x_{obs} is consistent with the ideal distance

From figure 27 we can see that for about 170 observations of non VIPs have p-values lower than our alpha level, so we reject the null hypothesis and conclude they are not consistent with the ideal distance. For the VIPs, 8 observations will reject the null hypothesis.

This is a good result for our data, meaning only a small proportion of our data needs to be looked into for outlier removal.

7 Testing for class membership using Mahalanobis distance (Country)

We will subset the data based on the country the profile is from, we will then go through each observation and check it against the data as if it were a new sample.

In general, the p-value reflects the probability of seeing a Mahalanobis value as large or larger than the actual Mahalanobis value, assuming the vector of predictor values that produced that Mahalanobis value was sampled from a population with an ideal mean (i.e. equal to the vector of mean predictor variable values used to generate the Mahalanobis value). P-values close to 0 reflect high Mahalanobis distance values and are therefore very dissimilar to the ideal combination of predictor variables. P-values close to 1 reflect low Mahalanobis distances and are therefore very similar to the ideal combination of predictor variables. The closer the p-value is to 1, the more similar that combination of predictor values is to the ideal combination. There are 75 observations that have a p-value of less than our chosen alpha level of 0.05. Those observations can be considered outliers as they are too far from our ideal combination of predictor variables.

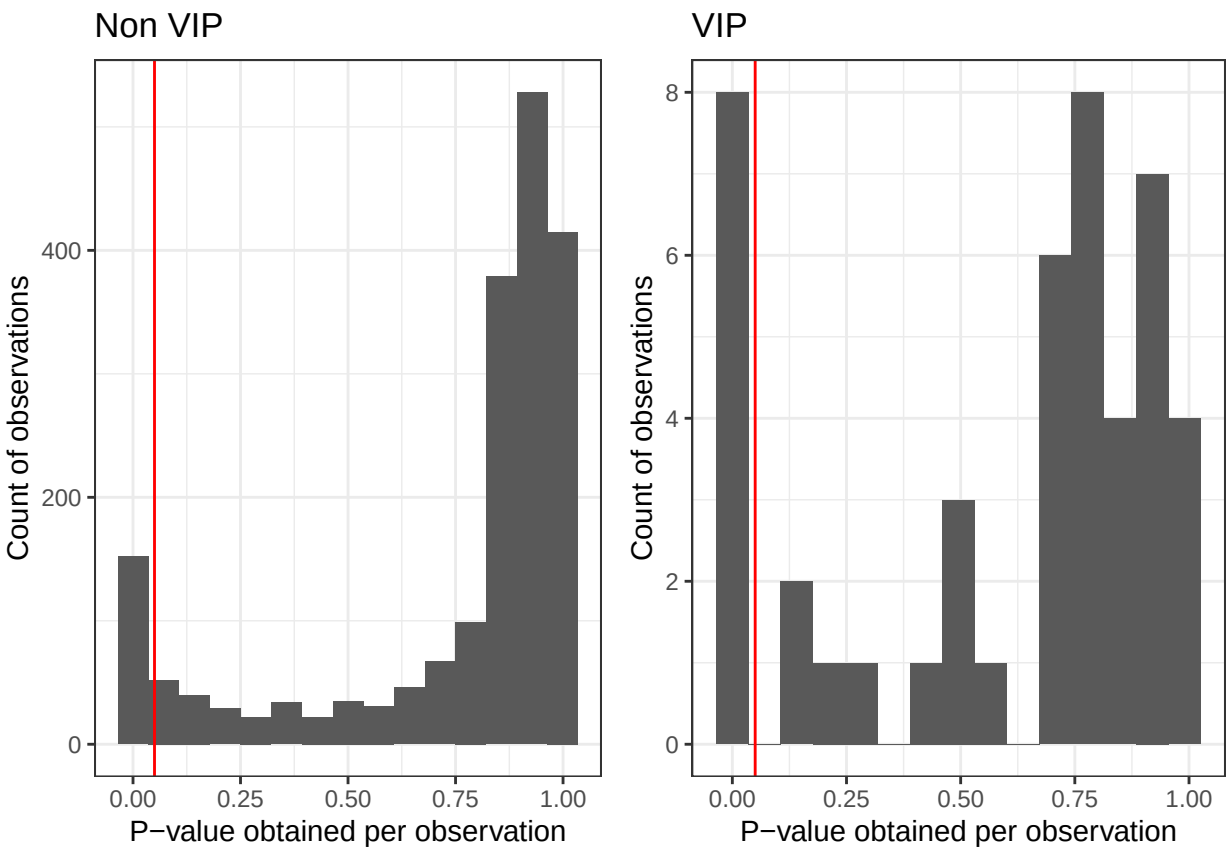


Figure 27: Histograms of Mahalanobis Distance p-values

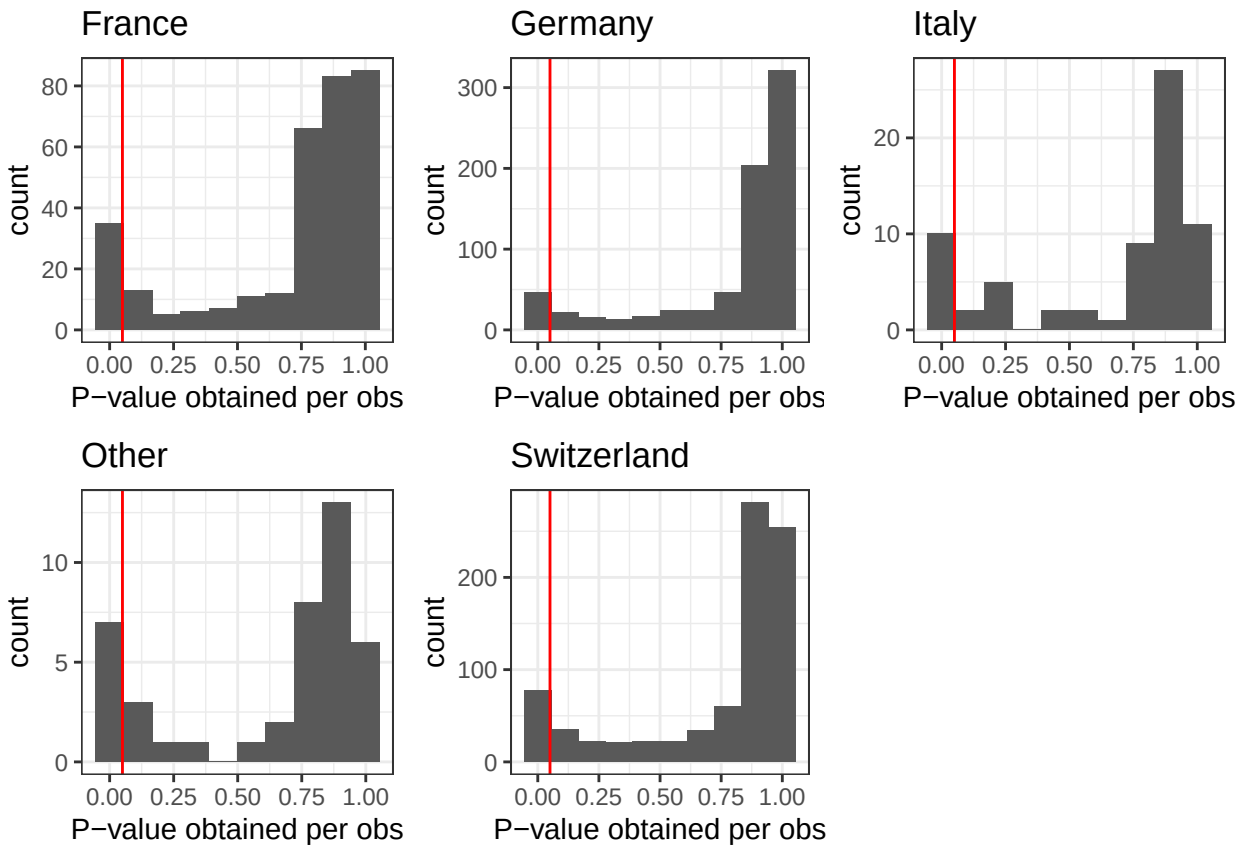


Figure 28: Histograms of Mahalanobis Distance p-values

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 3: Summary Statistics - Countires

	Country	Count
1	Argentina	1
2	Australia	2
3	Austria	20
4	Belgium	7
5	Bosnia and Herzegovina	3
6	Brazil	2
7	Canada	2
8	Central African Republic	1
9	Czechia	1
10	Ethiopia	1
11	France	646
12	Germany	1468
13	Hungary	1
14	India	1
15	Indonesia	1
16	Italy	138
17	Jamaica	1
18	Liberia	1
19	Liechtenstein	1
20	Luxembourg	5
21	Netherlands	2
22	Peru	1
23	Philippines	1
24	Romania	2
25	Russian Federation	2
26	Seychelles	2
27	Spain	6
28	Switzerland	1657
29	Turkey	10
30	Ukraine	1
31	United Kingdom of Great Britain and Northern Ireland	2
32	United States of America	3

Table 4: Summary Statistics - with grouped countries

	Country	Count
1	France	646
2	Germany	1468
3	Italy	138
4	Other	83
5	Switzerland	1657

Table 5: Correlation matrix

	age	counts_pictures	counts_profileVisits	counts_kisses	distance
age	1.00	-0.09	-0.08	-0.06	0.13
counts_pictures	-0.09	1.00	0.42	0.37	-0.05
counts_profileVisits	-0.08	0.42	1.00	0.89	-0.04
counts_kisses	-0.06	0.37	0.89	1.00	-0.04
distance	0.13	-0.05	-0.04	-0.04	1.00

Table 6: Summary Statistics - isVip (or not) by country

	isVip - No	isVip - Yes
France	631	15
Germany	1443	25
Italy	134	4
Other	82	1
Switzerland	1611	46