Wang_FinalProject

Yixin Wang

2022-12-06

- 1 **Yixin Wang's Final Project**
  - 1.1 **Introduction**
  - 1.2 **Caution!! I understand my sample size is around 210, which is smaller than 300. I have asked Dr. Crossley and he has approved me to keep using this dataset to conduct the PCA.**
  - 1.3 **Research Question**
    - 1.3.1 Hypothesis
  - 1.4 **Data**
    - 1.4.1 Variables Discussion
    - 1.4.2 **Considering the privacy and confidential issue of the dataset, you can contact me to see the data for further questions. Dr. Crossley has approved my special request on this.**
  - 1.5 **Data Wrangling**
    - 1.5.1 In order to organize the data successfully the following steps were completed.
    - 1.5.2 Visualize the continuous Predictors via Scatterplot by the outcome variable
  - 1.6 **Conduct PCA to reduce dimension**
    - 1.6.1 Road Map for PCA
    - 1.6.2 STEP 1: Correlations for Strong Multicollinearity
    - 1.6.3 STEP 2: Scale all the variables
    - 1.6.4 STEP 3: Visualizing PCA
    - 1.6.5 STEP 4: Bartlett's test
    - 1.6.6 STEP 5: KMO
    - 1.6.7 STEP 6: Baseline PCA
    - 1.6.8 STEP 7: Check that residuals are normally distributed
    - 1.6.9 STEP 8: Informed PCA with specific number of components
    - 1.6.10 STEP 9: Collect factor scores
  - 1.7 **Build Up the Regression Model via K-Fold Cross Validation**
    - 1.7.1 Set up repeated 10-fold CV, define training control parameters
    - 1.7.2 Run a cross validation model and check for suppression effects
    - 1.7.3 Let's check if the linear regression model we build violates any assumptions!
    - 1.7.4 Visualization
  - 1.8 **Disscussion**
    - 1.8.1 Overview
    - 1.8.2 Discussion on PCA outcomes
    - 1.8.3 Discussion on linear regression model
    - 1.8.4 Discussion on hypothesis
    - 1.8.5 Discussion on application of the findings
    - 1.8.6 Discussion on the limitations
    - 1.8.7 Future Directions

1 Yixin Wang's Final Project

1.1 Introduction

This paper is amied at understanding the relationship between social media usage motivation (continuous) and phyiscal health (continuous) among young adults when the school resumed in-person class after the pandemic. I'll use PCA and linear regression to conduct this analysis.

1.2 Caution!! I understand my sample size is around 210, which is smaller than 300. I have asked Dr. Crossley and he has approved me to keep using this dataset to conduct the PCA.

1.3 Research Question

How do different motivations using the top 4 most frequently used social media platforms, relate to young adults' physical health when students resumed in-person class after the pandemic? Is there a linear regression model between the dependent variable and independent variables?

1.3.1 Hypothesis

y - continuous dependent variable, which is the Physical Health of young adults when they resumed in-person class after the pandemic. x - continuous independent variables, which are 18 motivations for using the top 4 social media platforms; thus there should be 18 seperate hypothesis here (before conducting PCA). My below hypothesis is a general hypothesis for y and each x.

$$H_0 : \rho_{yx} = 0$$

$$H_1 : \rho_{yx} \neq 0$$

Furthermore, I assume the slope for significant independent variables and dependent variables should be negative.

1.4 Data

The data used in this assignment was collected by myself in November, 2021 at Vanderbilt University, which is primarily used as my honors independent project. Participants were first asked to recall what they were doing and what they were like during September 2020 (during the Fall semester a year ago from when they answered the questions). They were first asked some questions about where and how they took classes. Then, they were asked about SMU intensity, SMU motivations, SMU addictions, well-being, mental health, and perceived stress respectively. All the participants complete the measures in the same order. After they finished the first part of the survey, they were led to focus on the past month of the time they answered the survey, which was October or November 2021. Then, they answered the same group of questions. At the end of the study, participants answered a series of demographic questions. For this assignment 3, I will only focus on the data collected at Time 1, which intends to show what participants were doing during September 2020.

1.4.1 Variables Discussion

DV: Young Adults' Physical Health - This outcome variable is important because a lot of research evidence has shown that young adults have higher rate of diabetes, cardiovascular problems, muscular endurance, and etc. Thus, it is essential to understand if social media use motivation influence young adults' physical health.

IV: Social Media Usage Motivation: What specific motivations do users have for using Top 4 social media platforms among young adults (TikTok, Snapchat, YouTube, and Instagram) at Time Stamp 2? social media usage motivation always link tightly to students personal mental and physical status. It is interesting to see how the patterns of various SMU motivations combined together to influence physical health. IV may negatively relate to DV.

Motivation list: 1) to eat 2) friends 3) family 4) meet 5) romance 6) hookup 7) compare 8) entertain 9) do 10) wear 11) stay_informed_platform 12) forget 13) relax 14) buy 15) comments 16) support 17) opinion 18) academic purpose

1.4.2 Considering the privacy and confidential issue of the dataset, you can contact me to see the data for further questions. Dr. Crossley has approved my special request on this.

1.5 Data Wrangling

1.5.1 In order to organize the data successfully the following steps were completed.

1. Called in dataframes as tibble
2. Filtered dataframes to narrow scope to answer research question
3. clean the data

```
rm(list=ls(all=TRUE))
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0     ✔ purrr   0.3.4
## ✔ tibble  3.1.8     ✔ dplyr   1.0.9
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(ggplot2)
library(dplyr)
```

```
### let's call in our data as tibbles and examine the data we just called in first

smu <- read.csv("social_media_use_T2.csv", header = TRUE) %>%
  na_if("") %>% #convert empty cells to NA
  select(1:19, 572) %>% #select variables we want to focus this time: ID, IVs (motivatio
n), and DV (physical health)
  na.omit() #get rid of NA variables

str(smu)
```

```
## 'data.frame':    222 obs. of  20 variables:
##  $ record_id                  : int  1 2 3 4 5 7 8 9 10 11 ...
##  $ T4_friends_2               : num  1 2 1 2 2 2 3 2 2 1 ...
##  $ T4_family_2                : num  0 0 0 0 0 1 2 0 2 0 ...
##  $ T4_meet_2                  : num  0 0 0 1 1 2 1 0 0 0 ...
##  $ T4_romance_2               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ T4_hookup_2                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ T4_compare_2               : num  1 0 0 1 0 0 1 2 0 0 ...
##  $ T4_entertain_2             : num  2 4 1 1 2 2 3 1 0 2 ...
##  $ T4_do_2                    : num  0 2 0 0 0 2 1 0 0 0 ...
##  $ T4_eat_2                   : num  1 3 0 0 0 1 1 0 0 0 ...
##  $ T4_wear_2                  : num  0 2 0 0 0 0 0 1 0 0 ...
##  $ T4_stay_informed_platform_2: num  1 1 0 0 0 0 0 1 0 0 ...
##  $ T4_forget_2                : num  1 0 2 1 0 0 0 0 0 0 ...
##  $ T4_relax_platform_2        : num  1 2 1 1 2 3 3 1 0 0 ...
##  $ T4_buy_platform_2          : num  0 3 0 0 0 0 0 1 0 0 ...
##  $ T4_comments_2              : num  0 2 0 0 0 0 1 1 0 0 ...
##  $ T4_support_2               : num  2 0 0 0 0 0 1 0 0 0 ...
##  $ T4_opinion_2               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ T4_academic_platform_2     : num  0 0 0 1 0 0 0 1 0 0 ...
##  $ Full_Physical_Health_H_2   : num  2.5 4.5 3.75 4 3.75 3.5 3.5 4.25 3.25 4 ...
##  - attr(*, "na.action")= 'omit' Named int [1:10] 6 40 54 55 73 98 109 130 216 229
##   ..- attr(*, "names")= chr [1:10] "6" "40" "54" "55" ...
```

1.5.2 Visualize the continuous Predictors via Scatterplot by the outcome variable

Let's take a comprehensive look at the correlation among outcome and predictor variables

```
library(PerformanceAnalytics) #for chart.Correlation
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

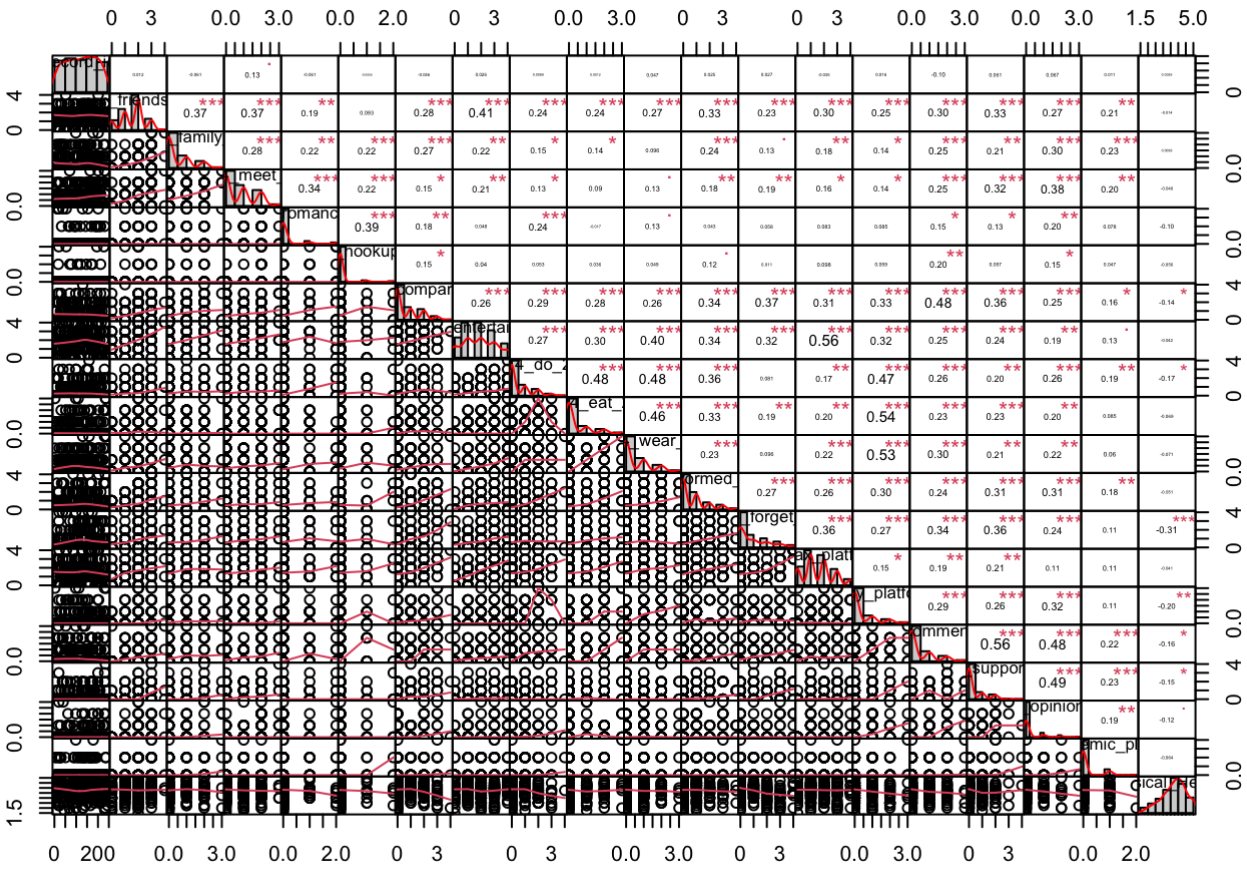
```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
chart.Correlation(smu, histogram = TRUE, method = "pearson")
```

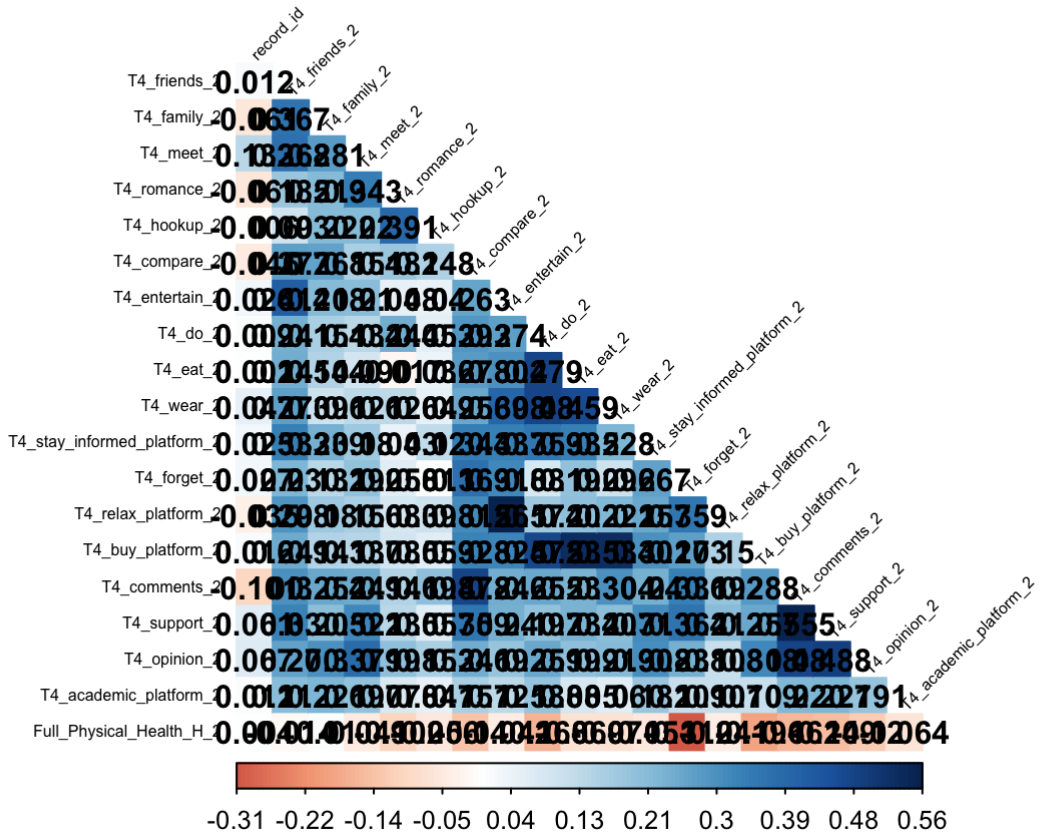


It seems like the comprehensive correlation chart is too packed to clearly observe due to a large number of IV. Thus, let's try to use colorful correlation graphs to see the strength of the correlation.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(smu),
         type="lower", #put color strength on bottom
         tl.pos = "ld", #Character or logical, position of text labels, 'ld'(default if
type=='lower') means left and diagonal,
         tl.cex = 0.5, #Numeric, for the size of text label (variable names).
         method="color",
         addCoef.col="black",
         diag=FALSE,
         tl.col="black", #The color of text label.
         tl.srt=45, #Numeric, for text label string rotation in degrees, see text
         is.corr = FALSE,
         number.digits = 3) #number of digits after decimal
```



Unfortunately, the colorful correlation graph is still too packed to observe. Thus, I choose to observe and interpret the correlation via correlation matrix. This is also the first step of PCA.

1.6 Conduct PCA to reduce dimension

1.6.1 Road Map for PCA

1. Check for multicollinearity between variables (r > .899)
2. Scale variables
3. Visualize the data
4. Bartlett's test including sample size
5. KMO on the data (look for variables below .5 and remove)
6. Baseline PCA to check scree plot, SS loadings above 1, and normal distribution of variables
7. Check that residuals are normally distributed
8. PCA with selected number of components based on interpretation of scree plot and SS loadings
9. Send factor scores csv

1.6.2 STEP 1: Correlations for Strong Multicollinearity

```
corr_mat_pca <- cor(smu %>% select(2:19))

corr_mat_pca
```

```
##                             T4_friends_2 T4_family_2  T4_meet_2 T4_romance_2
## T4_friends_2                  1.00000000  0.36656883 0.36796707   0.18543973
## T4_family_2                   0.36656883  1.00000000 0.28081204   0.21916336
## T4_meet_2                     0.36796707  0.28081204 1.00000000   0.34322583
## T4_romance_2                  0.18543973  0.21916336 0.34322583   1.00000000
## T4_hookup_2                   0.09298218  0.22017278 0.21961944   0.39137107
## T4_compare_2                  0.27659703  0.26789777 0.15395448   0.18237404
## T4_entertain_2                0.41400040  0.21751018 0.20962615   0.04795846
## T4_do_2                       0.24138491  0.15392095 0.13396620   0.24371843
## T4_eat_2                      0.24468048  0.14370537 0.08983147  -0.01706693
## T4_wear_2                     0.27284458  0.09636149 0.12647992   0.12578725
## T4_stay_informed_platform_2   0.33339557  0.23939340 0.17977927   0.04343800
## T4_forget_2                   0.23024938  0.13165478 0.19219402   0.05786354
## T4_relax_platform_2           0.29798936  0.18007706 0.15633437   0.08305949
## T4_buy_platform_2             0.24868662  0.14281891 0.13658155   0.08505894
## T4_comments_2                 0.29972012  0.25413154 0.24895085   0.14587081
## T4_support_2                  0.32999188  0.20533169 0.32164418   0.13480696
## T4_opinion_2                  0.27038913  0.30025532 0.37881458   0.19819504
## T4_academic_platform_2        0.21213083  0.22587750 0.19681632   0.07752888
##                             T4_hookup_2 T4_compare_2 T4_entertain_2    T4_do_2
## T4_friends_2                 0.09298218    0.2765970     0.41400040 0.24138491
## T4_family_2                  0.22017278    0.2678978     0.21751018 0.15392095
## T4_meet_2                    0.21961944    0.1539545     0.20962615 0.13396620
## T4_romance_2                 0.39137107    0.1823740     0.04795846 0.24371843
## T4_hookup_2                  1.00000000    0.1484188     0.03975761 0.05303710
## T4_compare_2                 0.14841877    1.0000000     0.26343852 0.29276926
## T4_entertain_2               0.03975761    0.2634385     1.00000000 0.27364679
## T4_do_2                      0.05303710    0.2927693     0.27364679 1.00000000
## T4_eat_2                     0.03579409    0.2777656     0.30165825 0.47918606
## T4_wear_2                    0.04930598    0.2561618     0.39754319 0.47953721
## T4_stay_informed_platform_2  0.12028831    0.3443259     0.33707872 0.35939535
## T4_forget_2                  0.01084154    0.3693052     0.31761832 0.08087914
## T4_relax_platform_2          0.09822695    0.3123655     0.56464858 0.17405252
## T4_buy_platform_2            0.05939404    0.3282654     0.32497414 0.47202870
## T4_comments_2                0.19785970    0.4781399     0.24645420 0.25532537
## T4_support_2                 0.05654059    0.3585619     0.23995390 0.19729742
## T4_opinion_2                 0.15319502    0.2461543     0.19181171 0.25922619
## T4_academic_platform_2       0.04685451    0.1571856     0.12520635 0.18819663
##                               T4_eat_2  T4_wear_2 T4_stay_informed_platform_2
## T4_friends_2                 0.24468048 0.27284458                  0.3333956
## T4_family_2                  0.14370537 0.09636149                  0.2393934
## T4_meet_2                    0.08983147 0.12647992                  0.1797793
## T4_romance_2                -0.01706693 0.12578725                  0.0434380
## T4_hookup_2                  0.03579409 0.04930598                  0.1202883
## T4_compare_2                 0.27776556 0.25616177                  0.3443259
## T4_entertain_2               0.30165825 0.39754319                  0.3370787
## T4_do_2                      0.47918606 0.47953721                  0.3593954
## T4_eat_2                     1.00000000 0.45873404                  0.3348597
## T4_wear_2                    0.45873404 1.00000000                  0.2277886
## T4_stay_informed_platform_2  0.33485969 0.22778861                  1.0000000
## T4_forget_2                  0.19219806 0.09617603                  0.2666375
## T4_relax_platform_2          0.20170744 0.22131668                  0.2568537
```

```
## T4_buy_platform_2            0.53503258 0.53410650                   0.3011534
## T4_comments_2                0.23260464 0.30395903                   0.2404137
## T4_support_2                 0.23444991 0.20683609                   0.3125553
## T4_opinion_2                 0.19888689 0.21928847                   0.3081205
## T4_academic_platform_2       0.08465127 0.06022756                   0.1822808
##                             T4_forget_2 T4_relax_platform_2 T4_buy_platform_2
## T4_friends_2                 0.23024938          0.29798936         0.24868662
## T4_family_2                  0.13165478          0.18007706         0.14281891
## T4_meet_2                    0.19219402          0.15633437         0.13658155
## T4_romance_2                 0.05786354          0.08305949         0.08505894
## T4_hookup_2                  0.01084154          0.09822695         0.05939404
## T4_compare_2                 0.36930521          0.31236546         0.32826545
## T4_entertain_2               0.31761832          0.56464858         0.32497414
## T4_do_2                      0.08087914          0.17405252         0.47202870
## T4_eat_2                     0.19219806          0.20170744         0.53503258
## T4_wear_2                    0.09617603          0.22131668         0.53410650
## T4_stay_informed_platform_2  0.26663751          0.25685369         0.30115342
## T4_forget_2                  1.00000000          0.35934924         0.27260340
## T4_relax_platform_2          0.35934924          1.00000000         0.14972545
## T4_buy_platform_2            0.27260340          0.14972545         1.00000000
## T4_comments_2                0.33605451          0.19184910         0.28768558
## T4_support_2                 0.36358988          0.21199226         0.25662809
## T4_opinion_2                 0.23781016          0.10805739         0.31813225
## T4_academic_platform_2       0.10913767          0.10730400         0.10933668
##                             T4_comments_2 T4_support_2 T4_opinion_2
## T4_friends_2                    0.2997201   0.32999188    0.2703891
## T4_family_2                     0.2541315   0.20533169    0.3002553
## T4_meet_2                       0.2489508   0.32164418    0.3788146
## T4_romance_2                    0.1458708   0.13480696    0.1981950
## T4_hookup_2                     0.1978597   0.05654059    0.1531950
## T4_compare_2                    0.4781399   0.35856187    0.2461543
## T4_entertain_2                  0.2464542   0.23995390    0.1918117
## T4_do_2                         0.2553254   0.19729742    0.2592262
## T4_eat_2                        0.2326046   0.23444991    0.1988869
## T4_wear_2                       0.3039590   0.20683609    0.2192885
## T4_stay_informed_platform_2     0.2404137   0.31255527    0.3081205
## T4_forget_2                     0.3360545   0.36358988    0.2378102
## T4_relax_platform_2             0.1918491   0.21199226    0.1080574
## T4_buy_platform_2               0.2876856   0.25662809    0.3181322
## T4_comments_2                   1.0000000   0.55523528    0.4796933
## T4_support_2                    0.5552353   1.00000000    0.4876773
## T4_opinion_2                    0.4796933   0.48767727    1.0000000
## T4_academic_platform_2          0.2202705   0.22667058    0.1913272
##                             T4_academic_platform_2
## T4_friends_2                            0.21213083
## T4_family_2                             0.22587750
## T4_meet_2                               0.19681632
## T4_romance_2                            0.07752888
## T4_hookup_2                             0.04685451
## T4_compare_2                            0.15718561
## T4_entertain_2                          0.12520635
## T4_do_2                                 0.18819663
```

```
## T4_eat_2                                0.08465127
## T4_wear_2                               0.06022756
## T4_stay_informed_platform_2             0.18228076
## T4_forget_2                             0.10913767
## T4_relax_platform_2                     0.10730400
## T4_buy_platform_2                       0.10933668
## T4_comments_2                           0.22027051
## T4_support_2                            0.22667058
## T4_opinion_2                            0.19132716
## T4_academic_platform_2                  1.00000000
```

```
#send to a csv file to check out

write.csv(corr_mat_pca, "corr_matrix_for_pca.csv")
#there is no variables having strong correlation (r>0.899), so we don't have to drop any variables now
```

1.6.3 STEP 2: Scale all the variables

```
library(psych)

scaled_data_pca <- smu %>%
  select(2:19)%>%
  mutate_at(c(1:18),~(scale(.) %>% as.vector))

str(scaled_data_pca)
```

```
## 'data.frame':    222 obs. of  18 variables:
##  $ T4_friends_2               : num  -0.703 0.396 -0.703 0.396 0.396 ...
##  $ T4_family_2                : num  -0.677 -0.677 -0.677 -0.677 -0.677 ...
##  $ T4_meet_2                  : num  -0.911 -0.911 -0.911 0.251 0.251 ...
##  $ T4_romance_2               : num  -0.325 -0.325 -0.325 -0.325 -0.325 ...
##  $ T4_hookup_2                : num  -0.234 -0.234 -0.234 -0.234 -0.234 ...
##  $ T4_compare_2               : num  0.269 -0.76 -0.76 0.269 -0.76 ...
##  $ T4_entertain_2             : num  0.121 1.753 -0.695 -0.695 0.121 ...
##  $ T4_do_2                    : num  -0.65 1.64 -0.65 -0.65 -0.65 ...
##  $ T4_eat_2                   : num  0.776 3.426 -0.549 -0.549 -0.549 ...
##  $ T4_wear_2                  : num  -0.686 1.853 -0.686 -0.686 -0.686 ...
##  $ T4_stay_informed_platform_2: num  0.225 0.225 -0.753 -0.753 -0.753 ...
##  $ T4_forget_2                : num  0.0966 -0.7614 0.9547 0.0966 -0.7614 ...
##  $ T4_relax_platform_2        : num  -0.526 0.366 -0.526 -0.526 0.366 ...
##  $ T4_buy_platform_2          : num  -0.541 3.508 -0.541 -0.541 -0.541 ...
##  $ T4_comments_2              : num  -0.608 1.987 -0.608 -0.608 -0.608 ...
##  $ T4_support_2               : num  2.099 -0.534 -0.534 -0.534 -0.534 ...
##  $ T4_opinion_2               : num  -0.392 -0.392 -0.392 -0.392 -0.392 ...
##  $ T4_academic_platform_2     : num  -0.405 -0.405 -0.405 2.164 -0.405 ...
##  - attr(*, "na.action")= 'omit' Named int [1:10] 6 40 54 55 73 98 109 130 216 229
##   ..- attr(*, "names")= chr [1:10] "6" "40" "54" "55" ...
```

```
psych::describe(scaled_data_pca) #gives you a lot of descriptives quickly
```

```
##                              vars   n mean sd median trimmed  mad   min  max
## T4_friends_2                    1 222    0  1   0.40    0.03 1.63 -1.80 2.59
## T4_family_2                     2 222    0  1  -0.68   -0.18 0.00 -0.68 3.24
## T4_meet_2                       3 222    0  1   0.25   -0.09 1.72 -0.91 2.58
## T4_romance_2                    4 222    0  1  -0.33   -0.30 0.00 -0.33 4.19
## T4_hookup_2                     5 222    0  1  -0.23   -0.23 0.00 -0.23 5.87
## T4_compare_2                    6 222    0  1  -0.76   -0.15 0.00 -0.76 3.35
## T4_entertain_2                  7 222    0  1   0.12   -0.03 1.21 -1.51 1.75
## T4_do_2                         8 222    0  1  -0.65   -0.19 0.00 -0.65 3.93
## T4_eat_2                        9 222    0  1  -0.55   -0.23 0.00 -0.55 3.43
## T4_wear_2                      10 222    0  1  -0.69   -0.18 0.00 -0.69 3.12
## T4_stay_informed_platform_2    11 222    0  1  -0.75   -0.19 0.00 -0.75 3.16
## T4_forget_2                    12 222    0  1  -0.76   -0.17 0.00 -0.76 2.67
## T4_relax_platform_2            13 222    0  1  -0.53   -0.05 1.32 -1.42 2.15
## T4_buy_platform_2              14 222    0  1  -0.54   -0.24 0.00 -0.54 3.51
## T4_comments_2                  15 222    0  1  -0.61   -0.21 0.00 -0.61 3.28
## T4_support_2                   16 222    0  1  -0.53   -0.23 0.00 -0.53 4.73
## T4_opinion_2                   17 222    0  1  -0.39   -0.26 0.00 -0.39 4.93
## T4_academic_platform_2         18 222    0  1  -0.41   -0.25 0.00 -0.41 4.73
##                              range  skew kurtosis   se
## T4_friends_2                  4.39 -0.16    -0.47 0.07
## T4_family_2                   3.92  1.24     0.44 0.07
## T4_meet_2                     3.49  0.64    -0.82 0.07
## T4_romance_2                  4.51  3.14     9.10 0.07
## T4_hookup_2                   6.11  4.55    21.01 0.07
## T4_compare_2                  4.11  1.04    -0.04 0.07
## T4_entertain_2                3.26  0.14    -0.96 0.07
## T4_do_2                       4.58  1.44     1.25 0.07
## T4_eat_2                      3.98  1.74     2.07 0.07
## T4_wear_2                     3.81  1.28     0.69 0.07
## T4_stay_informed_platform_2   3.91  1.28     0.82 0.07
## T4_forget_2                   3.43  1.11     0.10 0.07
## T4_relax_platform_2           3.57  0.35    -0.62 0.07
## T4_buy_platform_2             4.05  1.80     2.40 0.07
## T4_comments_2                 3.89  1.52     1.37 0.07
## T4_support_2                  5.27  1.89     3.14 0.07
## T4_opinion_2                  5.32  2.72     7.20 0.07
## T4_academic_platform_2        5.14  2.31     4.61 0.07
```

```
#note that kurtosis is a bit high on some of these
```

1.6.4 STEP 3: Visualizing PCA

This is just to understand the underlying data. This is not the final PCA

```
library(factoextra) #extract and visualize the output of multivariate data analyses, inc
luding 'PCA'
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

```
#line below runs a simple PCA with a component for each variable.
#the most variance will be explained in component 1 and 2
viz_pca <- prcomp(scaled_data_pca, center = TRUE,scale. = TRUE)

summary(viz_pca) #show the proportion of variance explained by all possible components a
long with cumulative variance
```

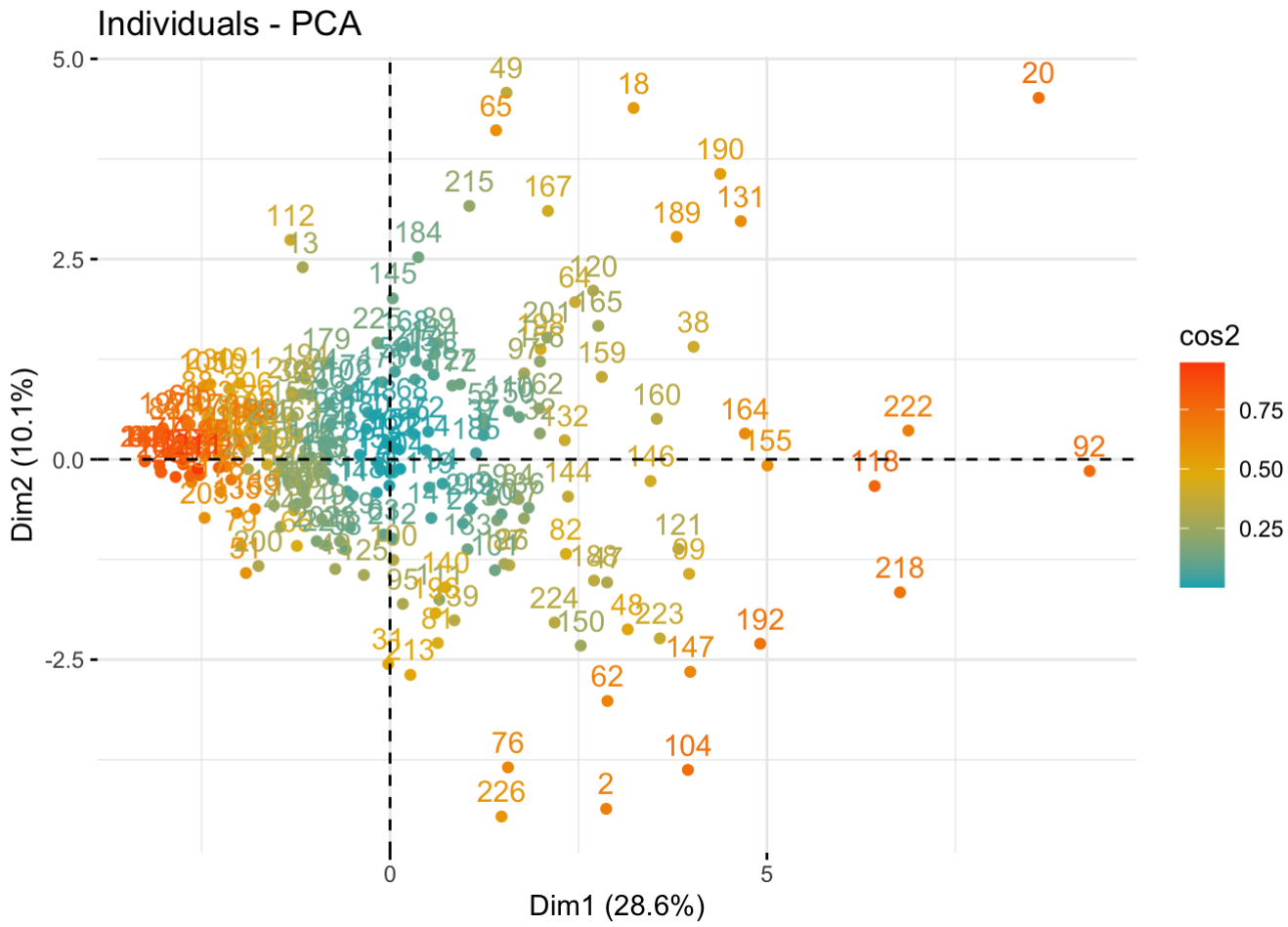
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2677 1.3466 1.18901 1.11964 1.02316 0.95299 0.89829
## Proportion of Variance 0.2857 0.1007 0.07854 0.06964 0.05816 0.05046 0.04483
## Cumulative Proportion  0.2857 0.3864 0.46497 0.53462 0.59277 0.64323 0.68806
##                           PC8     PC9   PC10   PC11    PC12    PC13   PC14
## Standard deviation     0.86206 0.83609 0.8161 0.7754 0.72454 0.70974 0.6546
## Proportion of Variance 0.04129 0.03884 0.0370 0.0334 0.02916 0.02799 0.0238
## Cumulative Proportion  0.72934 0.76818 0.8052 0.8386 0.86775 0.89574 0.9195
##                          PC15    PC16    PC17   PC18
## Standard deviation     0.6462 0.60955 0.59582 0.5515
## Proportion of Variance 0.0232 0.02064 0.01972 0.0169
## Cumulative Proportion  0.9427 0.96338 0.98310 1.0000
```

```
viz_pca$rotation #show the loadings for each component by variable
```

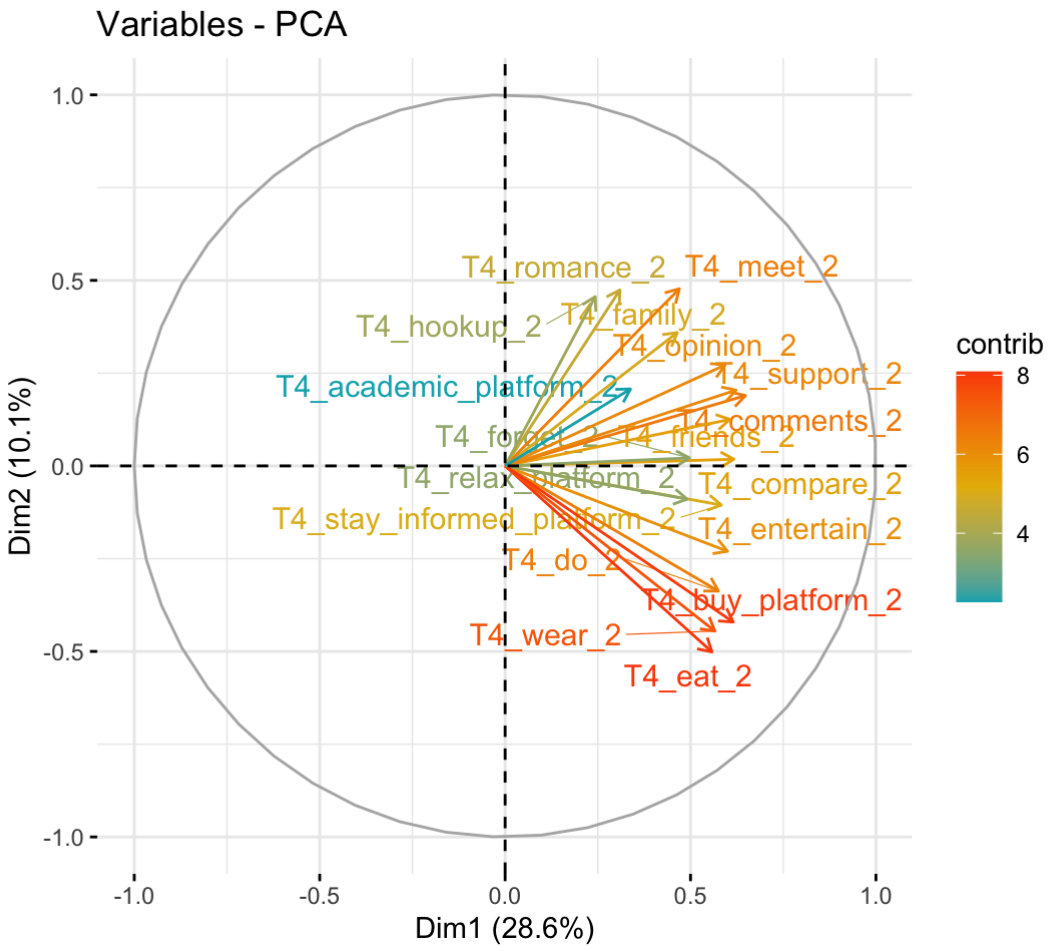
```
##                                    PC1         PC2         PC3         PC4
## T4_friends_2                 0.2673934  0.09307245 -0.09437472  0.20929349
## T4_family_2                  0.2047085  0.26675631  0.03319076  0.15655738
## T4_meet_2                    0.2072303  0.35451376  0.07196707  0.07969627
## T4_romance_2                 0.1366027  0.35202425  0.39571195  0.25283845
## T4_hookup_2                  0.1072248  0.33877250  0.31762927  0.26352860
## T4_compare_2                 0.2720249  0.01340764 -0.10488122 -0.07239000
## T4_entertain_2               0.2645374 -0.17084949 -0.25731425  0.38835011
## T4_do_2                      0.2537602 -0.25057903  0.36305390  0.01060218
## T4_eat_2                     0.2459650 -0.37231353  0.17168364 -0.04371057
## T4_wear_2                    0.2494154 -0.33053465  0.26496941  0.06284165
## T4_stay_informed_platform_2 0.2572165 -0.07840576 -0.07303659  0.01613037
## T4_forget_2                  0.2192775  0.01601849 -0.42783849 -0.07286382
## T4_relax_platform_2          0.2165155 -0.06655984 -0.37388902  0.45592721
## T4_buy_platform_2            0.2712675 -0.31252403  0.21871425 -0.11433294
## T4_comments_2                0.2857793  0.14069237 -0.06190373 -0.34799998
## T4_support_2                 0.2745686  0.15166227 -0.17735125 -0.38248793
## T4_opinion_2                 0.2615508  0.20192192  0.07254501 -0.34707942
## T4_academic_platform_2       0.1491313  0.15397158 -0.04556282 -0.12472211
##                                    PC5          PC6         PC7          PC8
## T4_friends_2                -0.31507459  0.187441019 -0.15325856  0.179734618
## T4_family_2                 -0.28167957 -0.259942521 -0.39342954  0.434035110
## T4_meet_2                   -0.16641826  0.460190032  0.00878896 -0.261157887
## T4_romance_2                 0.22275403  0.072261586  0.28949503 -0.178560450
## T4_hookup_2                  0.34936620 -0.262208140 -0.13848420 -0.029300473
## T4_compare_2                 0.31769796 -0.415922029  0.06669741  0.153579700
## T4_entertain_2              -0.05741585  0.187134464  0.05991164  0.100725366
## T4_do_2                     -0.10776257 -0.115949045  0.17995154 -0.182303179
## T4_eat_2                    -0.02972713 -0.062746286 -0.17621351 -0.058498576
## T4_wear_2                    0.04614955  0.227358106  0.16445497  0.287345501
## T4_stay_informed_platform_2 -0.14758373 -0.304546432 -0.42587243 -0.522803060
## T4_forget_2                  0.30729189  0.008476037  0.06180876 -0.321294509
## T4_relax_platform_2          0.13548756 -0.011932891  0.19068534  0.004468535
## T4_buy_platform_2            0.06992837  0.065598779 -0.02994616 -0.037019353
## T4_comments_2                0.22077639 -0.040151428  0.11377362  0.370966999
## T4_support_2                 0.05272838  0.168117037  0.04308613 -0.002193727
## T4_opinion_2                -0.04520302  0.237933945 -0.19637184 -0.051471753
## T4_academic_platform_2      -0.55689921 -0.388212066  0.58680262 -0.084809397
##                                    PC9         PC10         PC11         PC12
## T4_friends_2                -0.2194757  0.19712200 -0.52248175  0.238916954
## T4_family_2                 -0.2755732 -0.20058198  0.38086673 -0.084965935
## T4_meet_2                   -0.1011038 -0.15840918 -0.17943841 -0.308274869
## T4_romance_2                -0.3092891  0.20088375  0.15456849  0.080828312
## T4_hookup_2                  0.4824999 -0.28583720 -0.26598731  0.056700573
## T4_compare_2                -0.3345802  0.22292010 -0.13987957 -0.070535136
## T4_entertain_2               0.2405392  0.01045938  0.13588461  0.160919645
## T4_do_2                     -0.1004812  0.29752079  0.23751128 -0.155250242
## T4_eat_2                    -0.0278421 -0.29261707 -0.19546085 -0.584826578
## T4_wear_2                    0.1240832  0.04039323 -0.05666839  0.231424899
## T4_stay_informed_platform_2  0.1836727  0.33679048 -0.05617022  0.208823808
## T4_forget_2                 -0.3363286 -0.39042949  0.04705083  0.219417022
## T4_relax_platform_2          0.2080387  0.04538686  0.24708622 -0.310141399
```

```
## T4_buy_platform_2           -0.1356557 -0.40381533  0.01550120  0.321798382
## T4_comments_2                0.1911084  0.11257061 -0.10478646  0.005308328
## T4_support_2                 0.1209575  0.19953186 -0.10009187 -0.251556824
## T4_opinion_2                 0.2393596 -0.01139112  0.47623566  0.132377646
## T4_academic_platform_2       0.1554005 -0.25411949 -0.06925331  0.089662563
##                                     PC13         PC14        PC15         PC16
## T4_friends_2                -0.302179717 -0.367918589 -0.10954160  0.123715737
## T4_family_2                 -0.046969693  0.260843464  0.13612076  0.002917819
## T4_meet_2                    0.550385940  0.039061226  0.16258392 -0.048380131
## T4_romance_2                -0.278530277  0.174847112 -0.13460315 -0.004019637
## T4_hookup_2                 -0.088527516 -0.074425842  0.02495018 -0.040459659
## T4_compare_2                 0.488697323 -0.095750438 -0.28815134 -0.038697249
## T4_entertain_2               0.055154875  0.073182051  0.06478125 -0.604880309
## T4_do_2                     -0.107045559 -0.300269226  0.35049411 -0.188814251
## T4_eat_2                    -0.224952793 -0.056504260 -0.02317287  0.065899662
## T4_wear_2                    0.172056051  0.392014946  0.16939491  0.515449758
## T4_stay_informed_platform_2  0.115414126  0.249012767  0.07968258  0.102261144
## T4_forget_2                 -0.207245003 -0.007130529  0.37434089  0.156753888
## T4_relax_platform_2         -0.024061855 -0.136983938 -0.23487564  0.342110360
## T4_buy_platform_2            0.074097074  0.006014306 -0.41012409 -0.255493232
## T4_comments_2                0.004856409 -0.160154711  0.43112933 -0.109656940
## T4_support_2                -0.347576572  0.478866582 -0.25729944 -0.165480131
## T4_opinion_2                 0.037035436 -0.400769876 -0.23984879  0.215717720
## T4_academic_platform_2       0.007904088  0.031114037 -0.07443443  0.065913221
##                                     PC17         PC18
## T4_friends_2                -0.0614725005  0.045534659
## T4_family_2                 -0.1182962697 -0.002771088
## T4_meet_2                   -0.1161364157 -0.053244419
## T4_romance_2                 0.3351131086 -0.247682901
## T4_hookup_2                 -0.1588568061  0.251415152
## T4_compare_2                 0.1326510199  0.267071033
## T4_entertain_2               0.3490239198  0.170683022
## T4_do_2                     -0.3869538888  0.247470069
## T4_eat_2                     0.4492261027 -0.055646782
## T4_wear_2                    0.0345384903  0.195689839
## T4_stay_informed_platform_2  0.0615354833 -0.251059345
## T4_forget_2                  0.0003479448  0.195297596
## T4_relax_platform_2         -0.2824472818 -0.268779779
## T4_buy_platform_2           -0.3390164709 -0.342305844
## T4_comments_2                0.0527064563 -0.534914311
## T4_support_2                -0.2584379962  0.229592890
## T4_opinion_2                 0.2443065472  0.194273811
## T4_academic_platform_2       0.1063468272  0.016095698
```

```
#Graph of observations (i.e., schools). Schools with a similar profile are grouped toget
her.

fviz_pca_ind(viz_pca,
             c = "point", #point
             col.ind = "cos2", # Color by the quality of representation,
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), #color gradient
             repel = FALSE     # Avoid overlapping numbers, which is not important, so s
et as false
             )
```

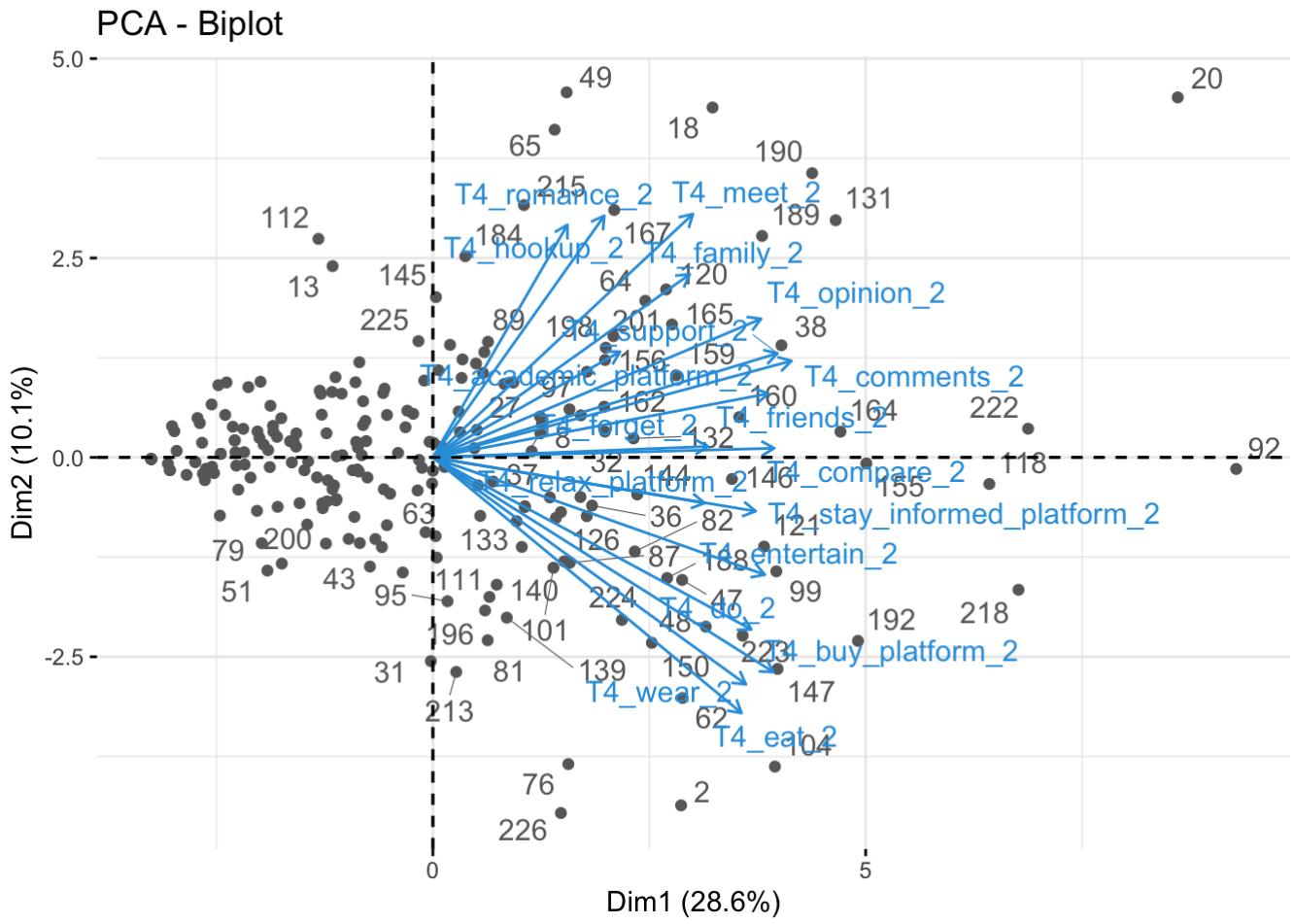



```
#Graph of variables. Positive correlated variables point to the same side of the plot. N
egative correlated variables point to opposite sides of the graph.

fviz_pca_var(viz_pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE #Avoid overlapping text if possible
             )
```




```
#Biplot of schools and variables together.

fviz_pca_biplot(viz_pca, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
                )
```

```
## Warning: ggrepel: 148 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



1.6.5 STEP 4: Bartlett's test

```
cortest.bartlett(scaled_data_pca, 232) #232 equals sample size
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 1134.457
##
## $p.value
## [1] 1.959667e-149
##
## $df
## [1] 153
```

Interpretation: p-value is smaller than 0.05, which is significant and means that the R-matrix is not an identity matrix. There are some relationships between the variables in the analysis and we can conduct PCA.

1.6.6 STEP 5: KMO

```
KMO(scaled_data_pca)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = scaled_data_pca)
## Overall MSA =  0.83
## MSA for each item =
##                 T4_friends_2                  T4_family_2
##                         0.90                         0.87
##                    T4_meet_2                 T4_romance_2
##                         0.85                         0.63
##                  T4_hookup_2                 T4_compare_2
##                         0.62                         0.87
##               T4_entertain_2                      T4_do_2
##                         0.82                         0.82
##                     T4_eat_2                    T4_wear_2
##                         0.86                         0.84
## T4_stay_informed_platform_2                  T4_forget_2
##                         0.87                         0.83
##          T4_relax_platform_2           T4_buy_platform_2
##                         0.77                         0.85
##                T4_comments_2                T4_support_2
##                         0.82                         0.87
##                 T4_opinion_2      T4_academic_platform_2
##                         0.86                         0.88
```

```
#all data above .50 and overal MSA is strong
```

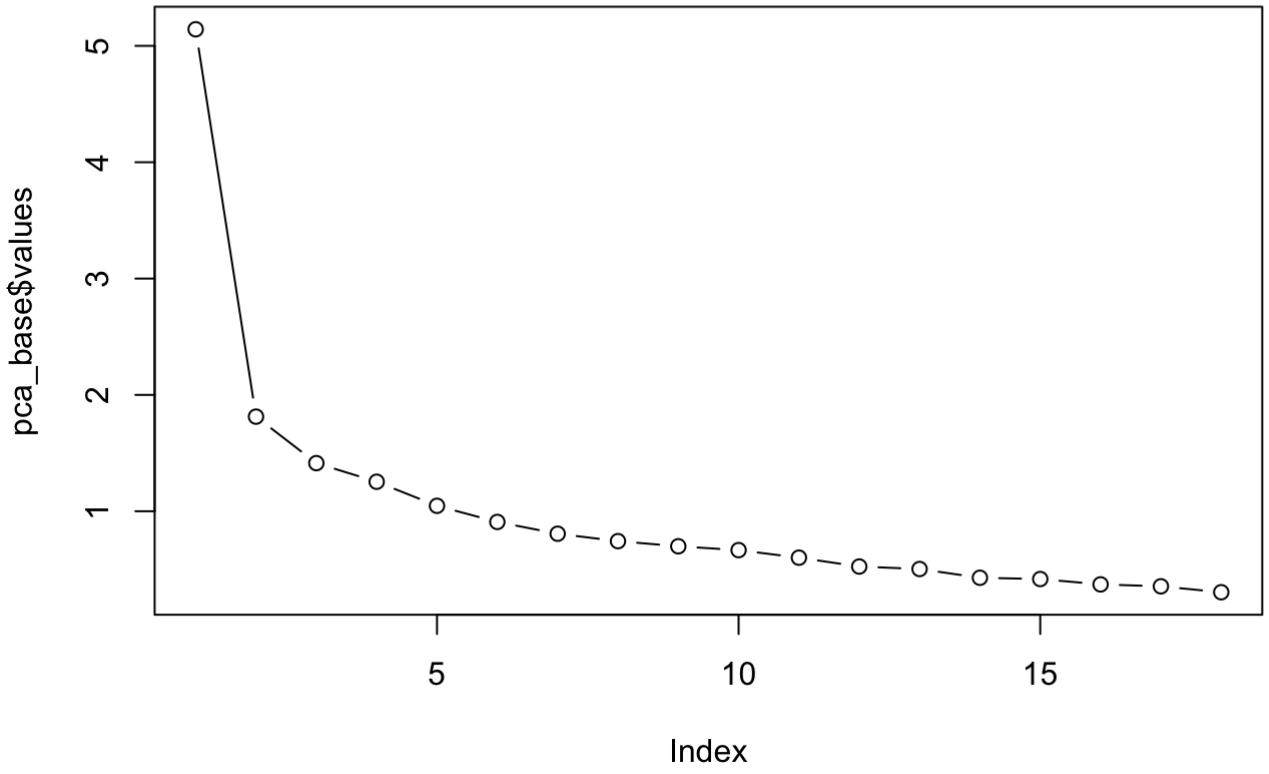
1.6.7 STEP 6: Baseline PCA

```
pca_base <- principal(scaled_data_pca, nfactors = 18, rotate = "none")

pca_base #results
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 18, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                              PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## T4_friends_2                0.61  0.13 -0.11  0.23 -0.32 -0.18 -0.14 -0.15
## T4_family_2                 0.46  0.36  0.04  0.18 -0.29  0.25 -0.35 -0.37
## T4_meet_2                   0.47  0.48  0.09  0.09 -0.17 -0.44  0.01  0.23
## T4_romance_2                0.31  0.47  0.47  0.28  0.23 -0.07  0.26  0.15
## T4_hookup_2                 0.24  0.46  0.38  0.30  0.36  0.25 -0.12  0.03
## T4_compare_2                0.62  0.02 -0.12 -0.08  0.33  0.40  0.06 -0.13
## T4_entertain_2              0.60 -0.23 -0.31  0.43 -0.06 -0.18  0.05 -0.09
## T4_do_2                     0.58 -0.34  0.43  0.01 -0.11  0.11  0.16  0.16
## T4_eat_2                    0.56 -0.50  0.20 -0.05 -0.03  0.06 -0.16  0.05
## T4_wear_2                   0.57 -0.45  0.32  0.07  0.05 -0.22  0.15 -0.25
## T4_stay_informed_platform_2 0.58 -0.11 -0.09  0.02 -0.15  0.29 -0.38  0.45
## T4_forget_2                 0.50  0.02 -0.51 -0.08  0.31 -0.01  0.06  0.28
## T4_relax_platform_2         0.49 -0.09 -0.44  0.51  0.14  0.01  0.17  0.00
## T4_buy_platform_2           0.62 -0.42  0.26 -0.13  0.07 -0.06 -0.03  0.03
## T4_comments_2               0.65  0.19 -0.07 -0.39  0.23  0.04  0.10 -0.32
## T4_support_2                0.62  0.20 -0.21 -0.43  0.05 -0.16  0.04  0.00
## T4_opinion_2                0.59  0.27  0.09 -0.39 -0.05 -0.23 -0.18  0.04
## T4_academic_platform_2      0.34  0.21 -0.05 -0.14 -0.57  0.37  0.53  0.07
##                               PC9  PC10  PC11  PC12  PC13  PC14  PC15  PC16
## T4_friends_2                -0.18 -0.16 -0.41 -0.17  0.21 -0.24 -0.07  0.08
## T4_family_2                 -0.23  0.16  0.30  0.06  0.03  0.17  0.09  0.00
## T4_meet_2                   -0.08  0.13 -0.14  0.22 -0.39  0.03  0.11 -0.03
## T4_romance_2                -0.26 -0.16  0.12 -0.06  0.20  0.11 -0.09  0.00
## T4_hookup_2                  0.40  0.23 -0.21 -0.04  0.06 -0.05  0.02 -0.02
## T4_compare_2                -0.28 -0.18 -0.11  0.05 -0.35 -0.06 -0.19 -0.02
## T4_entertain_2               0.20 -0.01  0.11 -0.12 -0.04  0.05  0.04 -0.37
## T4_do_2                     -0.08 -0.24  0.18  0.11  0.08 -0.20  0.23 -0.12
## T4_eat_2                    -0.02  0.24 -0.15  0.42  0.16 -0.04 -0.01  0.04
## T4_wear_2                    0.10 -0.03 -0.04 -0.17 -0.12  0.26  0.11  0.31
## T4_stay_informed_platform_2  0.15 -0.27 -0.04 -0.15 -0.08  0.16  0.05  0.06
## T4_forget_2                 -0.28  0.32  0.04 -0.16  0.15  0.00  0.24  0.10
## T4_relax_platform_2          0.17 -0.04  0.19  0.22  0.02 -0.09 -0.15  0.21
## T4_buy_platform_2           -0.11  0.33  0.01 -0.23 -0.05  0.00 -0.27 -0.16
## T4_comments_2                0.16 -0.09 -0.08  0.00  0.00 -0.10  0.28 -0.07
## T4_support_2                 0.10 -0.16 -0.08  0.18  0.25  0.31 -0.17 -0.10
## T4_opinion_2                 0.20  0.01  0.37 -0.10 -0.03 -0.26 -0.15  0.13
## T4_academic_platform_2       0.13  0.21 -0.05 -0.06 -0.01  0.02 -0.05  0.04
##                              PC17  PC18 h2       u2 com
## T4_friends_2                -0.04  0.03  1 -4.4e-16 5.4
## T4_family_2                 -0.07  0.00  1 -1.3e-15 8.2
## T4_meet_2                   -0.07 -0.03  1  1.8e-15 6.0
## T4_romance_2                 0.20 -0.14  1 -8.9e-16 7.6
## T4_hookup_2                 -0.09  0.14  1 -4.4e-16 7.9
## T4_compare_2                 0.08  0.15  1 -4.4e-16 4.9
## T4_entertain_2               0.21  0.09  1  3.3e-16 5.0
## T4_do_2                     -0.23  0.14  1  0.0e+00 5.8
## T4_eat_2                     0.27 -0.03  1  6.7e-16 4.9
## T4_wear_2                    0.02  0.11  1  0.0e+00 5.8
```

```
## T4_stay_informed_platform_2  0.04 -0.14  1  1.1e-16 5.2
## T4_forget_2                  0.00  0.11  1  1.3e-15 6.1
## T4_relax_platform_2         -0.17 -0.15  1 -6.7e-16 5.7
## T4_buy_platform_2           -0.20 -0.19  1  0.0e+00 4.9
## T4_comments_2                0.03 -0.29  1 -2.2e-16 4.4
## T4_support_2                -0.15  0.13  1  0.0e+00 4.9
## T4_opinion_2                 0.15  0.11  1  0.0e+00 5.5
## T4_academic_platform_2       0.06  0.01  1 -2.2e-16 4.6
##
##                        PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10 PC11
## SS loadings           5.14 1.81 1.41 1.25 1.05 0.91 0.81 0.74 0.70 0.67 0.60
## Proportion Var        0.29 0.10 0.08 0.07 0.06 0.05 0.04 0.04 0.04 0.04 0.03
## Cumulative Var        0.29 0.39 0.46 0.53 0.59 0.64 0.69 0.73 0.77 0.81 0.84
## Proportion Explained  0.29 0.10 0.08 0.07 0.06 0.05 0.04 0.04 0.04 0.04 0.03
## Cumulative Proportion 0.29 0.39 0.46 0.53 0.59 0.64 0.69 0.73 0.77 0.81 0.84
##                       PC12 PC13 PC14 PC15 PC16 PC17 PC18
## SS loadings           0.52 0.50 0.43 0.42 0.37 0.36 0.30
## Proportion Var        0.03 0.03 0.02 0.02 0.02 0.02 0.02
## Cumulative Var        0.87 0.90 0.92 0.94 0.96 0.98 1.00
## Proportion Explained  0.03 0.03 0.02 0.02 0.02 0.02 0.02
## Cumulative Proportion 0.87 0.90 0.92 0.94 0.96 0.98 1.00
##
## Mean item complexity =  5.7
## Test of the hypothesis that 18 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

```
#SS The eigenvalues associated with each factor represent the variance explained by that particular linear component.
#R calls these SS loadings (sums of squared loadings), because they are the sum of the squared loadings.

#Proportion of var (ss loading divided by sample size)

#How many components to extract? The number of SS loadings greater than 1 (Kaiser's criterion).

#Potentialy 4 here.

#scree plot using eigen values stored in pca_1$values
plot(pca_base$values, type = "b")
```



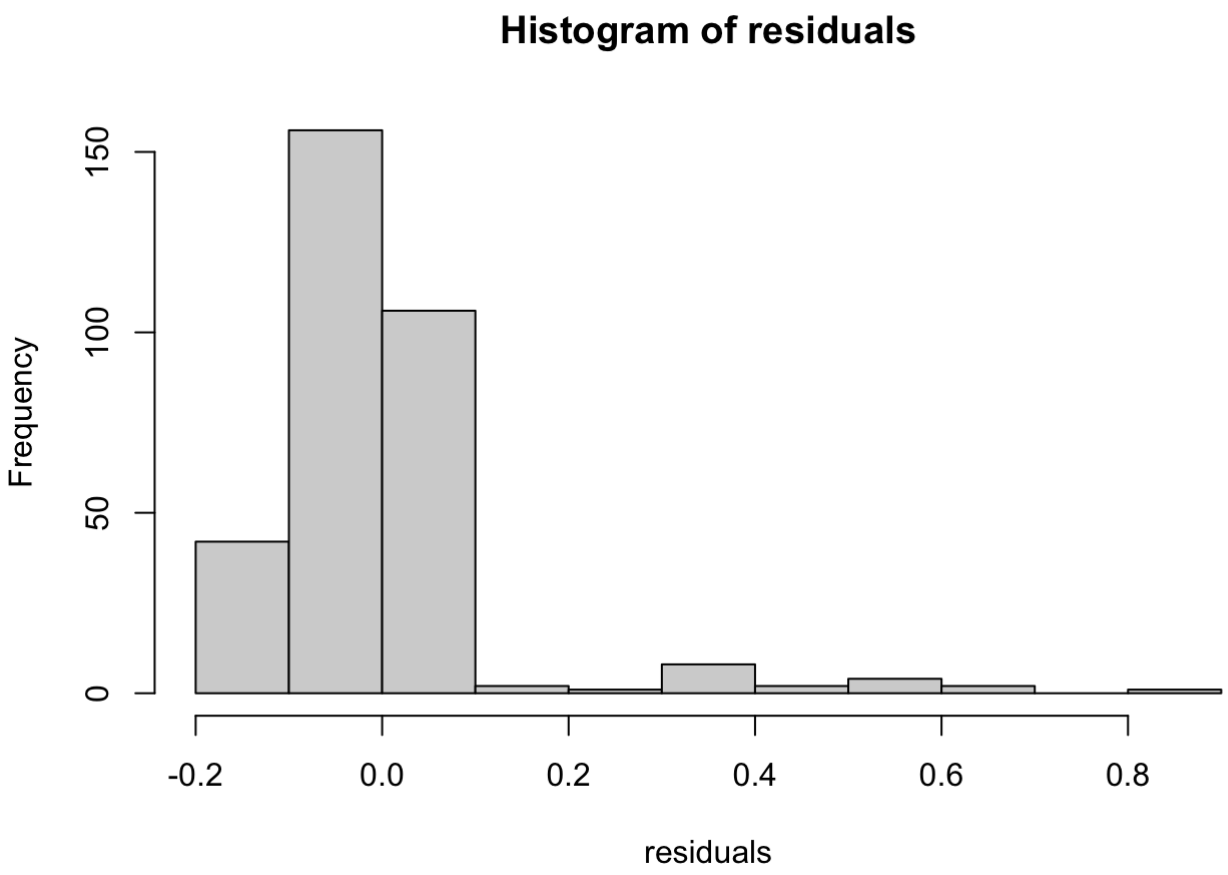

```
#plots the eigenvalues (y) against the factor number (x)
#type = 'b' both gives you a line and points on the same graph

#indicates 3-5 variables here

#Let's pick 4
```

1.6.8 STEP 7: Check that residuals are normally distributed

```
pca_resid <- principal(scaled_data_pca, nfactors = 4, rotate = "none")
pca_resid #results. 5 looks good
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                              PC1   PC2   PC3   PC4   h2   u2 com
## T4_friends_2                0.61  0.13 -0.11  0.23 0.45 0.55 1.5
## T4_family_2                 0.46  0.36  0.04  0.18 0.38 0.62 2.2
## T4_meet_2                   0.47  0.48  0.09  0.09 0.46 0.54 2.1
## T4_romance_2                0.31  0.47  0.47  0.28 0.62 0.38 3.4
## T4_hookup_2                 0.24  0.46  0.38  0.30 0.50 0.50 3.3
## T4_compare_2                0.62  0.02 -0.12 -0.08 0.40 0.60 1.1
## T4_entertain_2              0.60 -0.23 -0.31  0.43 0.70 0.30 2.7
## T4_do_2                     0.58 -0.34  0.43  0.01 0.63 0.37 2.5
## T4_eat_2                    0.56 -0.50  0.20 -0.05 0.61 0.39 2.3
## T4_wear_2                   0.57 -0.45  0.32  0.07 0.62 0.38 2.6
## T4_stay_informed_platform_2 0.58 -0.11 -0.09  0.02 0.36 0.64 1.1
## T4_forget_2                 0.50  0.02 -0.51 -0.08 0.51 0.49 2.1
## T4_relax_platform_2         0.49 -0.09 -0.44  0.51 0.71 0.29 3.0
## T4_buy_platform_2           0.62 -0.42  0.26 -0.13 0.64 0.36 2.3
## T4_comments_2               0.65  0.19 -0.07 -0.39 0.61 0.39 1.9
## T4_support_2                0.62  0.20 -0.21 -0.43 0.66 0.34 2.3
## T4_opinion_2                0.59  0.27  0.09 -0.39 0.58 0.42 2.2
## T4_academic_platform_2      0.34  0.21 -0.05 -0.14 0.18 0.82 2.1
##
##                        PC1  PC2  PC3  PC4
## SS loadings           5.14 1.81 1.41 1.25
## Proportion Var        0.29 0.10 0.08 0.07
## Cumulative Var        0.29 0.39 0.46 0.53
## Proportion Explained  0.53 0.19 0.15 0.13
## Cumulative Proportion 0.53 0.72 0.87 1.00
##
## Mean item complexity =  2.3
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  317.52  with prob <  5.7e-28
##
## Fit based upon off diagonal values = 0.93
```

```
#residuals
#require correlation matrix for final data
corMatrix<-cor(scaled_data_pca)
#corMatrix

#next,create an object from the correlation matrix and the pca loading. Call it residual
s. It will contain the factor residuals
residuals<-factor.residuals(corMatrix, pca_resid$loadings)

#call a histogram to check residuals
hist(residuals) #are the residuals normally distributed? They look okay. That is good
```



Due to the limited number of subjects, the residuals are comparatively normal distributed

1.6.9 STEP 8: Informed PCA with specific number of components

Let's try 4 components.

We are also going to rotate the data

A factor is a classification axis along which variables can be plotted

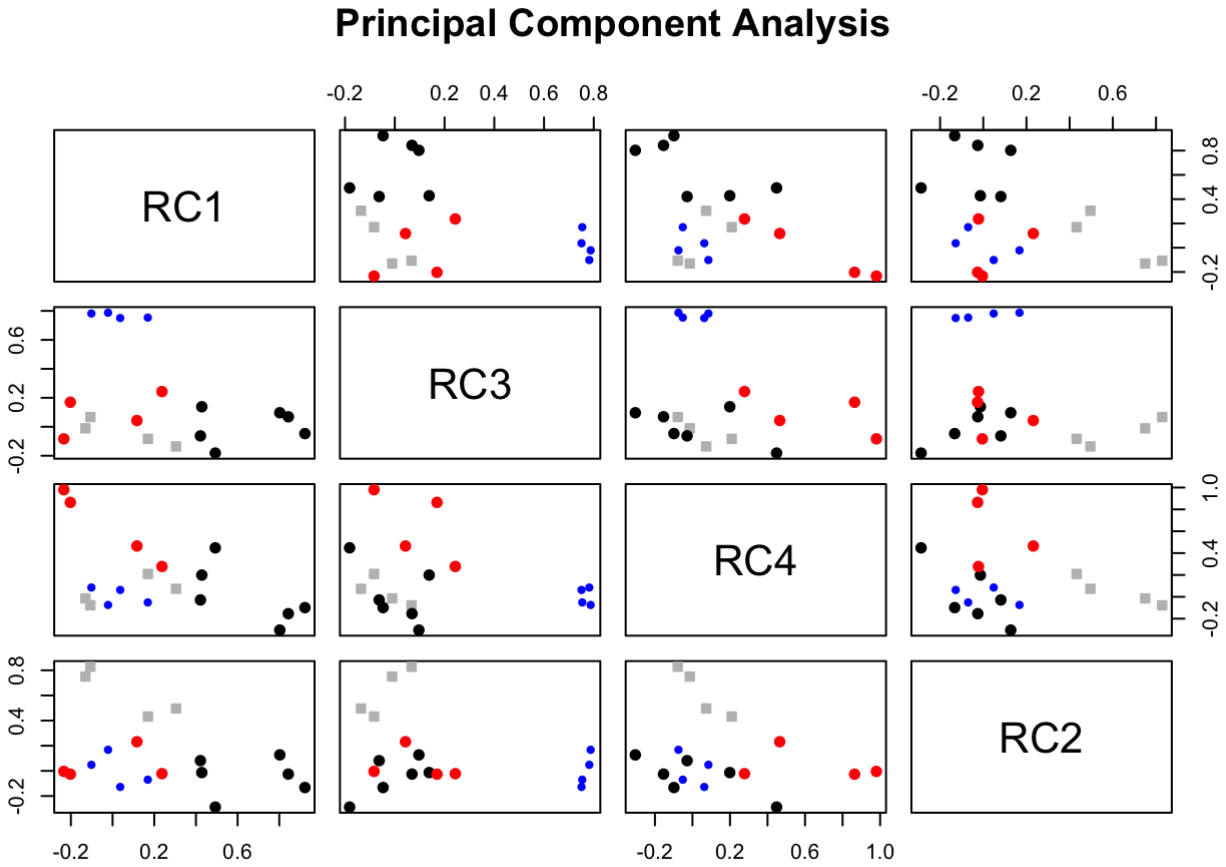
```
#rotation. Since factors should be related, use oblique technique (promax), if unrelate
d, use varimax
pca_final <- principal(scaled_data_pca, nfactors = 4, rotate = "promax")
pca_final #results.
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                               RC1   RC3   RC4   RC2   h2   u2 com
## T4_friends_2                 0.12  0.04  0.47  0.23 0.45 0.55 1.6
## T4_family_2                  0.17 -0.08  0.21  0.43 0.38 0.62 1.9
## T4_meet_2                    0.30 -0.14  0.07  0.50 0.46 0.54 1.9
## T4_romance_2                -0.11  0.07 -0.08  0.83 0.62 0.38 1.1
## T4_hookup_2                 -0.13 -0.01 -0.01  0.75 0.50 0.50 1.1
## T4_compare_2                 0.43  0.14  0.20 -0.01 0.40 0.60 1.6
## T4_entertain_2              -0.20  0.17  0.86 -0.03 0.70 0.30 1.2
## T4_do_2                     -0.02  0.79 -0.07  0.17 0.63 0.37 1.1
## T4_eat_2                     0.04  0.75  0.06 -0.13 0.61 0.39 1.1
## T4_wear_2                   -0.10  0.78  0.08  0.05 0.62 0.38 1.1
## T4_stay_informed_platform_2  0.24  0.24  0.28 -0.02 0.36 0.64 3.0
## T4_forget_2                  0.49 -0.18  0.45 -0.29 0.51 0.49 2.9
## T4_relax_platform_2         -0.23 -0.08  0.98  0.00 0.71 0.29 1.1
## T4_buy_platform_2            0.17  0.75 -0.05 -0.07 0.64 0.36 1.1
## T4_comments_2                0.84  0.07 -0.15 -0.03 0.61 0.39 1.1
## T4_support_2                 0.92 -0.05 -0.10 -0.13 0.66 0.34 1.1
## T4_opinion_2                 0.80  0.10 -0.30  0.13 0.58 0.42 1.4
## T4_academic_platform_2       0.42 -0.06 -0.03  0.08 0.18 0.82 1.1
##
##                        RC1  RC3  RC4  RC2
## SS loadings           2.92 2.62 2.22 1.86
## Proportion Var        0.16 0.15 0.12 0.10
## Cumulative Var        0.16 0.31 0.43 0.53
## Proportion Explained  0.30 0.27 0.23 0.19
## Cumulative Proportion 0.30 0.58 0.81 1.00
##
##  With component correlations of
##      RC1  RC3  RC4  RC2
## RC1 1.00 0.40 0.58 0.40
## RC3 0.40 1.00 0.40 0.17
## RC4 0.58 0.40 1.00 0.27
## RC2 0.40 0.17 0.27 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  317.52  with prob <  5.7e-28
##
## Fit based upon off diagonal values = 0.93
```

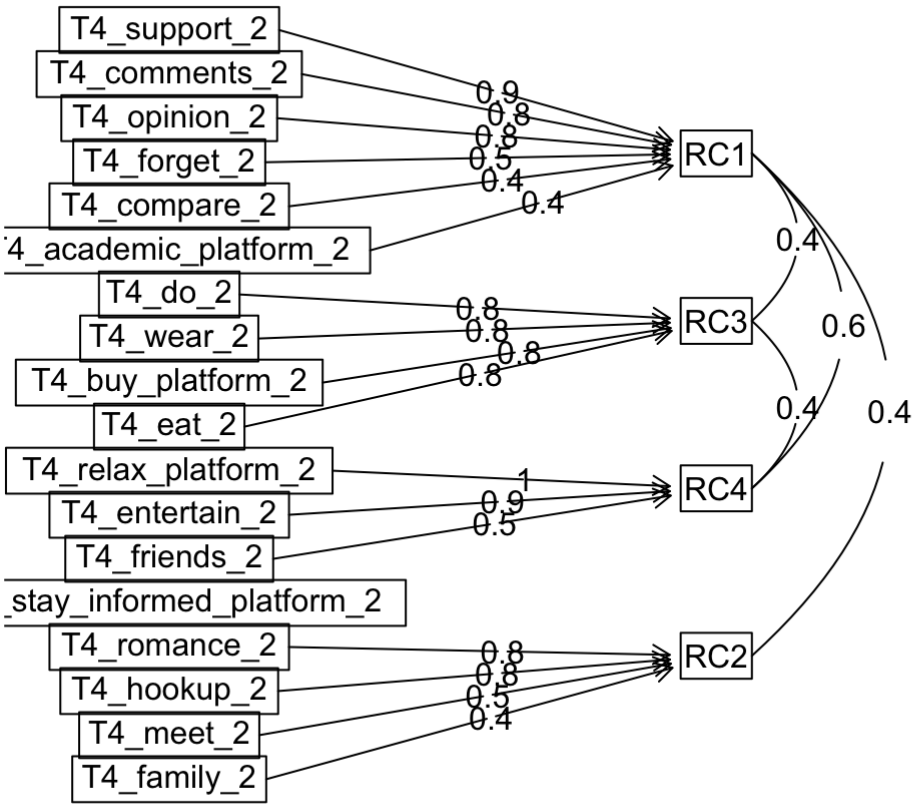
```
#let's make the results easier to read. Include loadings over 3 and sort them

print.psych(pca_final, cut = 0.3, sort = TRUE)
```

```
## Principal Components Analysis
## Call: principal(r = scaled_data_pca, nfactors = 4, rotate = "promax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                             item   RC1   RC3   RC4   RC2   h2   u2 com
## T4_support_2                  16  0.92                   0.66 0.34 1.1
## T4_comments_2                 15  0.84                   0.61 0.39 1.1
## T4_opinion_2                  17  0.80       -0.30       0.58 0.42 1.4
## T4_forget_2                   12  0.49        0.45       0.51 0.49 2.9
## T4_compare_2                   6  0.43                   0.40 0.60 1.6
## T4_academic_platform_2        18  0.42                   0.18 0.82 1.1
## T4_do_2                        8        0.79             0.63 0.37 1.1
## T4_wear_2                     10        0.78             0.62 0.38 1.1
## T4_buy_platform_2             14        0.75             0.64 0.36 1.1
## T4_eat_2                       9        0.75             0.61 0.39 1.1
## T4_relax_platform_2           13              0.98       0.71 0.29 1.1
## T4_entertain_2                 7              0.86       0.70 0.30 1.2
## T4_friends_2                   1              0.47       0.45 0.55 1.6
## T4_stay_informed_platform_2   11                         0.36 0.64 3.0
## T4_romance_2                   4                   0.83 0.62 0.38 1.1
## T4_hookup_2                    5                   0.75 0.50 0.50 1.1
## T4_meet_2                      3  0.30             0.50 0.46 0.54 1.9
## T4_family_2                    2                   0.43 0.38 0.62 1.9
##
##                        RC1  RC3  RC4  RC2
## SS loadings           2.92 2.62 2.22 1.86
## Proportion Var        0.16 0.15 0.12 0.10
## Cumulative Var        0.16 0.31 0.43 0.53
## Proportion Explained  0.30 0.27 0.23 0.19
## Cumulative Proportion 0.30 0.58 0.81 1.00
##
##  With component correlations of
##      RC1  RC3  RC4  RC2
## RC1 1.00 0.40 0.58 0.40
## RC3 0.40 1.00 0.40 0.17
## RC4 0.58 0.40 1.00 0.27
## RC2 0.40 0.17 0.27 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  317.52  with prob <  5.7e-28
##
## Fit based upon off diagonal values = 0.93
```

```
plot(pca_final)
```




```
fa.diagram(pca_final)
```



1.6.10 STEP 9: Collect factor scores

This will give you the factor scores for each observation

```
#we need the pca scores
pca_final_scores <- as.data.frame(pca_final$scores) #scores for each text on each facto
r. You can use these in subsequent analyses. Lot's of them though


write.csv(pca_final_scores,"pca_scores_rc2_is_achieve.csv", row.names=FALSE)
```

```
df <- data.frame(pca_final_scores,smu %>% select(1,20) )
cor(df %>% select(-5))
```

```
##                                RC1        RC3        RC4         RC2
## RC1                      1.0000000  0.4006144  0.5758634  0.39806485
## RC3                      0.4006144  1.0000000  0.3996397  0.17231145
## RC4                      0.5758634  0.3996397  1.0000000  0.26849084
## RC2                      0.3980649  0.1723115  0.2684908  1.00000000
## Full_Physical_Health_H_2 -0.2086203 -0.1388306 -0.1009150 -0.04448617
##                          Full_Physical_Health_H_2
## RC1                                   -0.20862030
## RC3                                   -0.13883063
## RC4                                   -0.10091501
## RC2                                   -0.04448617
## Full_Physical_Health_H_2               1.00000000
```

1.7 Build Up the Regression Model via K-Fold Cross Validation

Call in all the necessary packages

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

1.7.1 Set up repeated 10-fold CV, define training control parameters

```
#method = cross validation, number = ten times (10 fold cross-validation)
train.control <- trainControl(method = "cv", number = 10)
#Set up repeated k-fold cross-validation
```

1.7.2 Run a cross validation model and check for suppression

effects

```
cv01 <- train(
  form = Full_Physical_Health_H_2 ~ .,
  # need to use as.factor() for response to make it as categorical
  # this line of code will be given to student
  data = df %>% select(-5),
  trControl = train.control,
  method = "lm"
  # generalized linear model (glm) is used to build logistic model
  #specifies the distribution of the response variable
)

summary(cv01)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3105 -0.5348  0.1167  0.5804  1.4840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.73198    0.05236  71.279   <2e-16 ***
## RC1         -0.17628    0.06912  -2.550   0.0115 *
## RC3         -0.05997    0.05879  -1.020   0.3088
## RC4          0.03588    0.06597   0.544   0.5871
## RC2          0.03559    0.05728   0.621   0.5351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7801 on 217 degrees of freedom
## Multiple R-squared:  0.0503, Adjusted R-squared:  0.03279
## F-statistic: 2.873 on 4 and 217 DF,  p-value: 0.02389
```

There is only one variable (RC1) is significant, so we have to deduct the other non-signifianct variables one by one based on their p-values; let;s deduct RC2 first because it has the largest t-value

```
cv02 <- train(
  form = Full_Physical_Health_H_2 ~ .,
  # need to use as.factor() for response to make it as categorical
  # this line of code will be given to student
  data = df %>% select(1,2,3,6),
  trControl = train.control,
  method = "lm"
  # generalized linear model (glm) is used to build logistic model
  #specifies the distribution of the response variable
)

summary(cv02)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3278 -0.5101  0.1137  0.5743  1.4690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.73198    0.05228  71.380   <2e-16 ***
## RC1         -0.16336    0.06582  -2.482   0.0138 *
## RC3         -0.05984    0.05870  -1.019   0.3091
## RC4          0.03794    0.06579   0.577   0.5648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.779 on 218 degrees of freedom
## Multiple R-squared:  0.04861,    Adjusted R-squared:  0.03552
## F-statistic: 3.713 on 3 and 218 DF,  p-value: 0.01235
```

let's deduct RC4 then, because it has the largest t-value now

```
cv03 <- train(
  form = Full_Physical_Health_H_2 ~ .,
  # need to use as.factor() for response to make it as categorical
  # this line of code will be given to student
  data = df %>% select(1,2,6),
  trControl = train.control,
  method = "lm"
  # generalized linear model (glm) is used to build logistic model
  #specifies the distribution of the response variable
)

summary(cv03)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3513 -0.5087  0.1000  0.5699  1.4386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.73198    0.05220  71.489   <2e-16 ***
## RC1         -0.14457    0.05710  -2.532   0.0121 *
## RC3         -0.05221    0.05710  -0.914   0.3616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7778 on 219 degrees of freedom
## Multiple R-squared:  0.04716,    Adjusted R-squared:  0.03846
## F-statistic:  5.42 on 2 and 219 DF,  p-value: 0.005043
```

finally, let's deduct the RC3

```
cv04 <- train(
  form = Full_Physical_Health_H_2 ~ .,
  # need to use as.factor() for response to make it as categorical
  # this line of code will be given to student
  data = df %>% select(1,6),
  trControl = train.control,
  method = "lm"
  # generalized linear model (glm) is used to build logistic model
  #specifies the distribution of the response variable
)

summary(cv04)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3171 -0.5129  0.1090  0.5661  1.4425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.73198    0.05218  71.515  < 2e-16 ***
## RC1         -0.16548    0.05230  -3.164  0.00178 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7775 on 220 degrees of freedom
## Multiple R-squared:  0.04352,    Adjusted R-squared:  0.03917
## F-statistic: 10.01 on 1 and 220 DF,  p-value: 0.001776
```

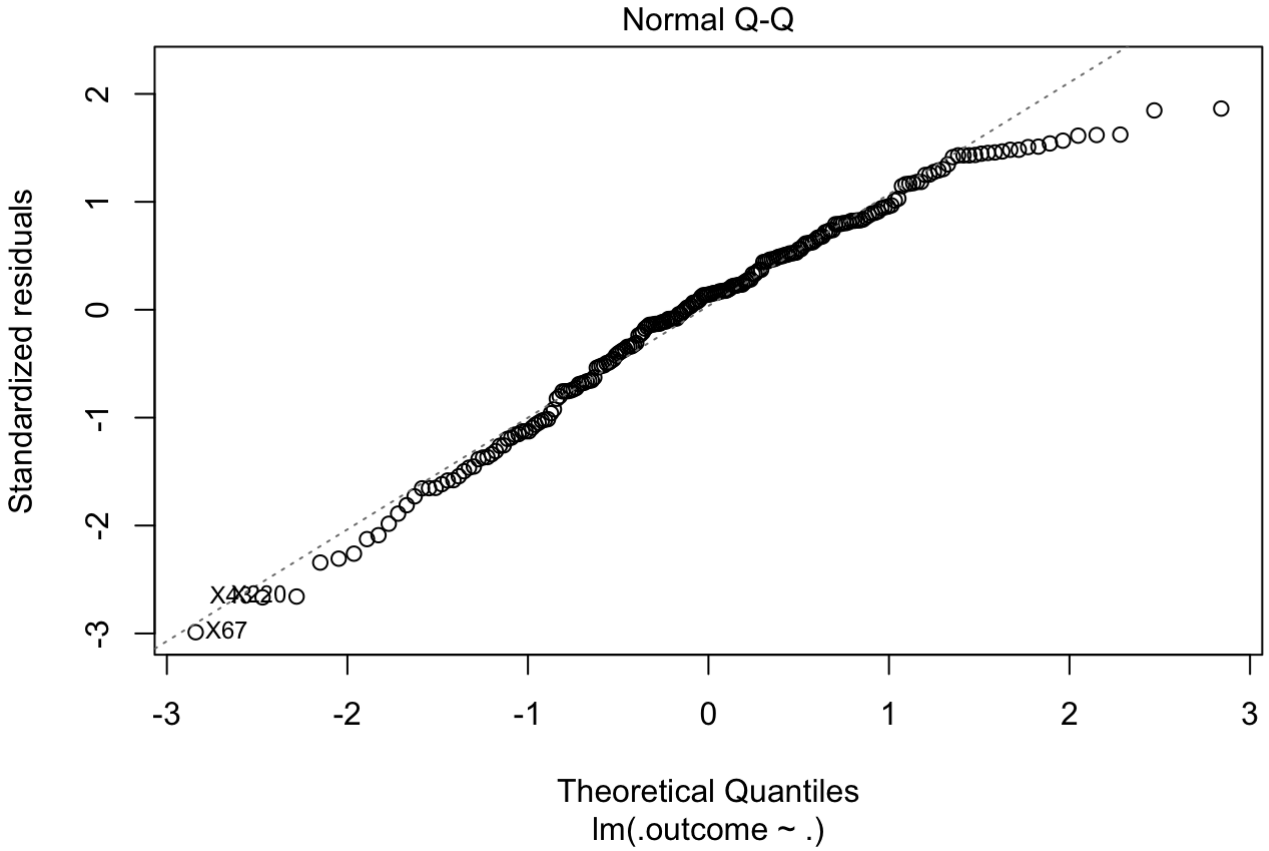
Because I have conducted PCA and I only left one indpeendent variable after pruning, I won't have problems on multicollinearity and supression effects.

1.7.3 Let's check if the linear regression model we build violates any assumptions!

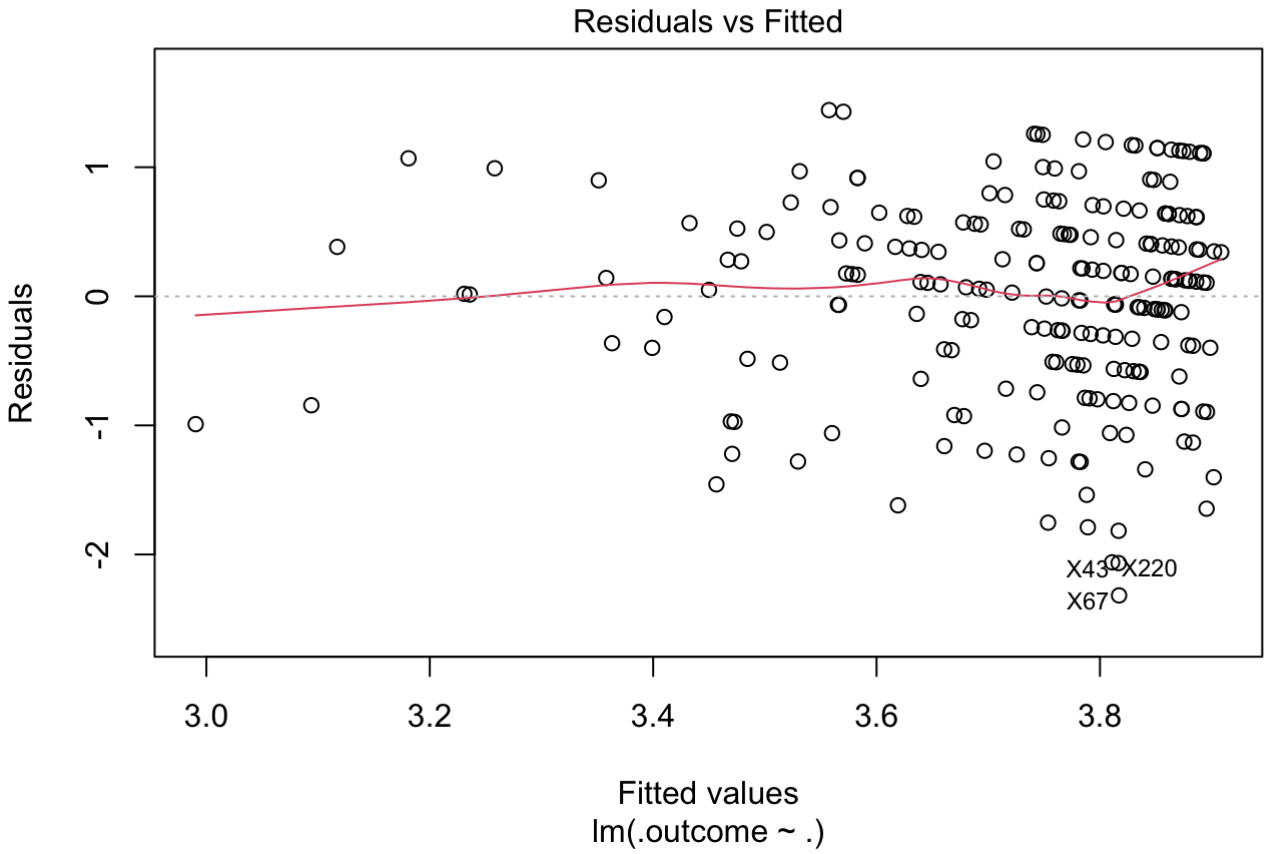
```
#HOMOSCEDASTICITY

#Are residuals normally distributed? Yes, p-value is really small, which indicates a nor
mal distribution.
shapiro.test(residuals(cv04))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(cv04)
## W = 0.97538, p-value = 0.0006421
```

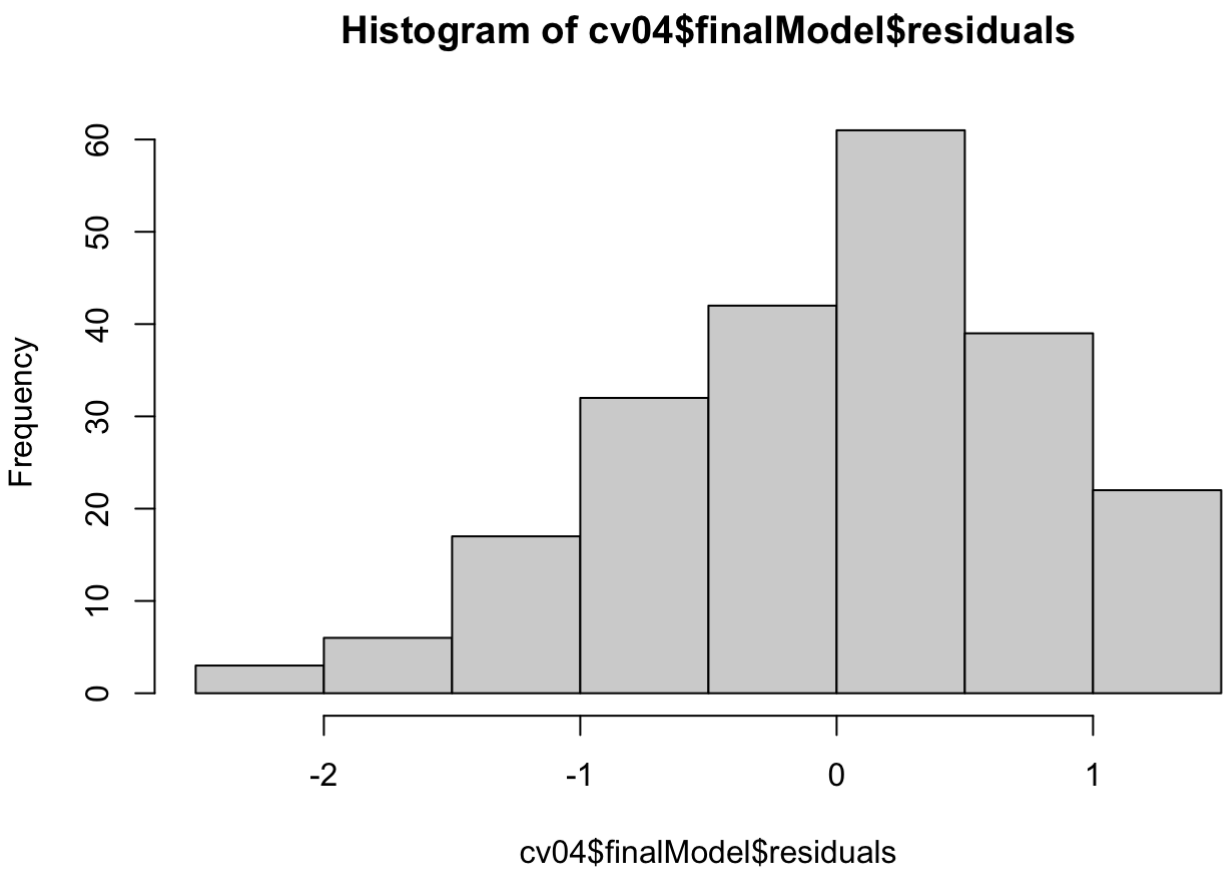
```
#let's plot the qq plot! That's really beautifu;, alomost all the scatterplots are on th
e line, which is good
plot(cv04$finalModel, which = 2)
```




```
#Plot the residuals vs. the fitted values.

plot(cv04$finalModel, which = 1)
```



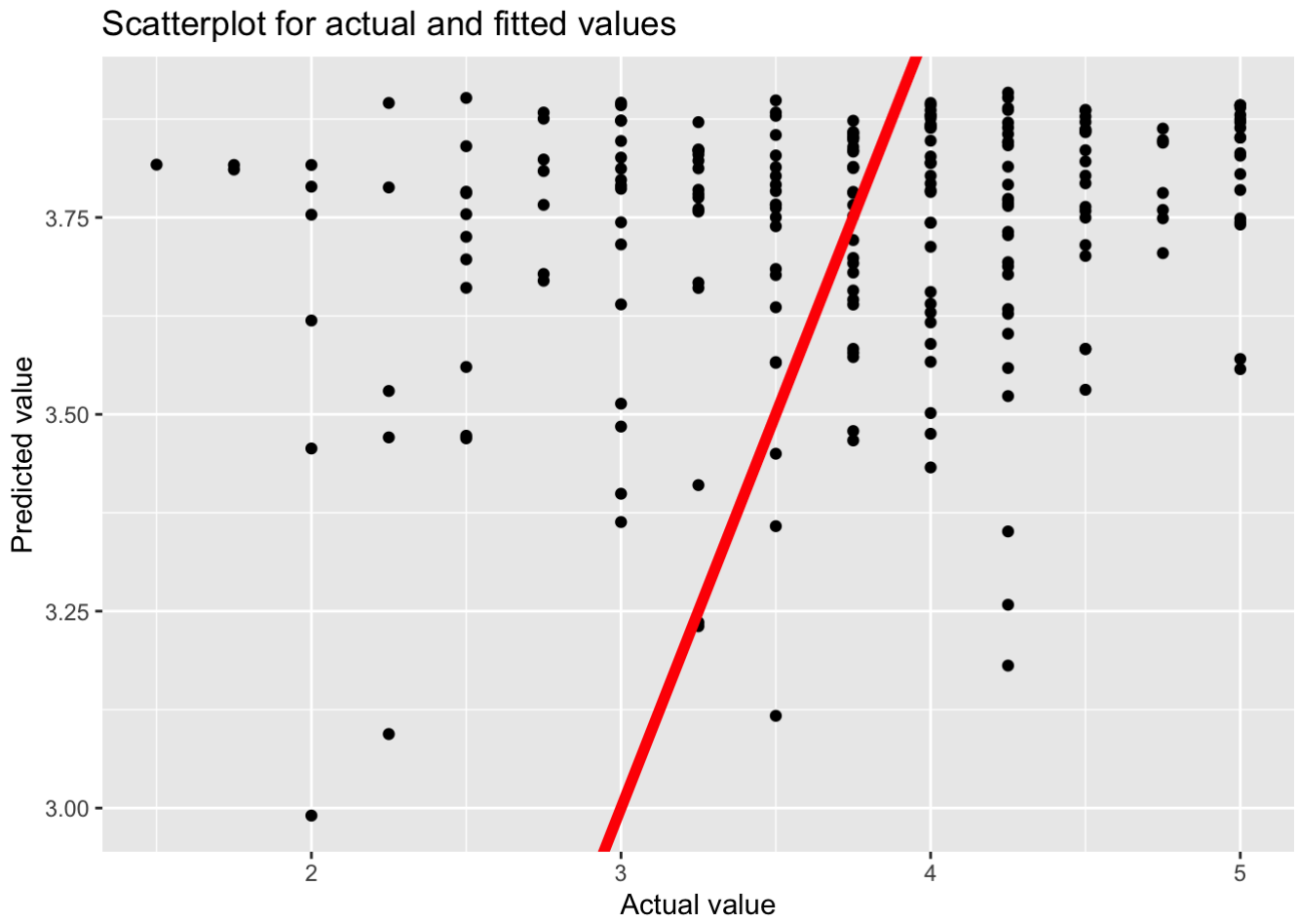

```
#plot the histogram of residuals. The histogram should look like a symmetric bell shaped curve, centered at 0.
hist(cv04$finalModel$residuals)
```



1.7.4 Visualization

```
actual <- df$Full_Physical_Health_H_2
fitted <- unname(cv04$finalModel$fitted.values) #would have been a named number vector i
f unname not used

act_fit <- cbind.data.frame(actual, fitted) #cbind binds the two vectors into a datafram
e

#let's draw the graph!
ggplot(act_fit, aes(x = actual, y = fitted)) +
  geom_point() +
  xlab("Actual value") +
  ylab("Predicted value") +
  ggtitle("Scatterplot for actual and fitted values") +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              size = 2)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
```



1.8 Disscussion

1.8.1 Overview

This paper I investigated the relationship between young adults' social media usage motivation and phyiscla health. I conducted PCA first to reduce the variable dimension and then conducted linear regression model to see what specific type of motivation relates to young adults physical health. In general, "social activity" SMU motivation significantly negatively influenced young adults' phyiscal health when the school resumed in-person class.

1.8.2 Discussion on PCA outcomes

Based on PCA results, we can see that PCA aggregated my 18 variables into 4 pincipal components.

RC1 aggregated by support (loading = 0.9), opinion (loading = 0.8), comments (loading = 0.8),compare (loading = 0.4), academic platform (loading = 0.4), and forget (loading = 0.5), which could be interpreted as "social_activity".

RC2 is aggregated by hookup (loading = 0.8), romance (loading = 0.8), family (loading = 0.4), and meet (loading = 0.5), which could be interpreted as "romance".

RC3 is aggregated by eat (loading = 0.8), buy (loading = 0.8), wear (loading = 0.8), and do (loading = 0.8),which could be interpreted as "social_connection". Among these 6 variables, compare, stay)inforned and compare have lower loading than the previous three.

RC4 is aggregated by relax (loading = 1), entertain (loading = 0.9), and friends (loading = 0.5), which could be interpreted as “activity”. Do is comparatively smaller loading than the previous three motivations.

Stay_informed is an individual variable which is not belong to any above principal components. We will leave it alone.

Meanwhile, based on the results, we know that the SS loading of RC1 is 2.92, which explains 0.16 proportion variance; SS loading of RC 2 is 1.86, which explains 0.10 proportion variance; SS loading of RC3 is 2.62, which explains 0.15 proportion variance; SS loading of RC 4 is 2.22, which explains 0.12 proportion variance. Thus, these four principal components explained 0.53 proportion variance in total matrix correlation.

Among these four principal components, RC1 explained 0.3 propotion, RC2 explained 0.19 proportion, RC3 explained 0.27 proportion, and RC4 explained 0.23 proportion; all of which have 1.0 cumulative proportion of the four principal components.

1.8.3 Discussion on linear regression model

The estimated linear regression model function should be:

$$y = 3.73 - 0.17SocialActivity$$

This function indicated that for every 1 unit increase in social activity, the physical health decrease by 0.17; If the social activity is 0, then the base line phyiscal health should be 3.73.

Because I have conducted PCA and I only left one indpeendent variable after pruning, I won’t have problems on multicollinearity and supression effects.

Meanwhile, based on our test on assumption violations, we can see that my model didn’t violate any assumptions. 1) The Shapiro-Wilk normality test is a statistical test used to determine whether a sample of data comes from a normally distributed population. The test calculates a statistic called the W value, which is used to evaluate whether the sample data are likely to come from a normally distributed population. From my test result, we can see that the W value is 0.97538, which is close to 1, this indicates that the sample is likely to come from a normally distributed population.

2. The QQ plot is a useful tool for evaluating whether a sample of data is likely to come from a specific distribution. From the above graph, we can see that the plot is approximately straight, this indicates that the sample data is likely to come from the specified distribution. Meanwhile, most of the points lie on the straight line, this indicates that the sample data is well-behaved and follows the specified distribution.
3. The residuals vs fitted values plot is a useful tool for evaluating the fit of a regression model. Based on the above graph, the points are randomly dispersed around the horizontal line at 0, this indicates that the regression model is a good fit for the data.
4. As we can see from the histogram of residuals, the graph is comparatively normal distributed.
5. From the scatter plot for actual and fitted values, we can see the red line is comparatively in the middle of the graph, which represents a great sign of my model prediction.

1.8.4 Discussion on hypothesis

We can see that physical health has significant negative relationship with social activity, with p-value smaller than 0.05. Thus, I reject the null hypothesis for factorized factor (social activity) and fail to reject other variables’ null hypothesis.

We lose other variables during the variable pruning process. Nevertheless, when we go back to the full regression analysis, the slope for RC4 (named as activity) and RC2 (named as romance) were positive and the slope for RC3 (social_connection) were negative. These predictors are all not significant.

1.8.4.1 Discussion on R-squared value

As we keep looking at the regression results of RC1, we can see that the multiple r-squared value is 0.04352, which shows how well terms (data points) fit a curve or line. This model predicts 4.352% of the data. The adjusted r-squared value is 0.0392, which also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. This model predicts 3.92% of the data.

Furthermore, the residual standard error is 0.78 and F-test is significant with p-value < 0.002.

1.8.5 Discussion on application of the findings

As we know, as the technology developing quickly, increasing number of young adults get used to use social meida platforms to get connected with their loved ones and be updated with the news and events they care about. However, sometimes young adults overlook how their motivation to use social media may influence their health condition. Based on our results, we can clearly see that when young adults want to use social media platforms to show support, to comment, to express opinion, to forget, to compare with others and to used for academic purpose, these motivations work togther to negatively influence young adults' physical health. Thus, when young adults have hyiscal health issues after using the social media, they may reduce using social media for these specific motivations. At the same time, doctors should also pay attention to these specific fields of motivation.

1.8.6 Discussion on the limitations

Considering the limited number of sample size, our PCA may be not accurate enough. And when I build up the model, only one factor remained, this situation may also due to my linited number of participants. Meanwhile, our data collection is based on participants' self-report, so we cannot make sure that's validate enough.

1.8.7 Future Directions

In the future, I would recommend researchers to recruit more participants to make sure they can condict validate and reliable PCA. Meanwhile, because I only focus on phyiscal health this time, I hope researchers can build up regression model between motivations and mental health or general health problems. It will also be interesting to see if specific social media functions could bring negative health influence to young adults.