Exam: STA 380: Introduction to Machine Learning

Snehal Naravane (sn27429)

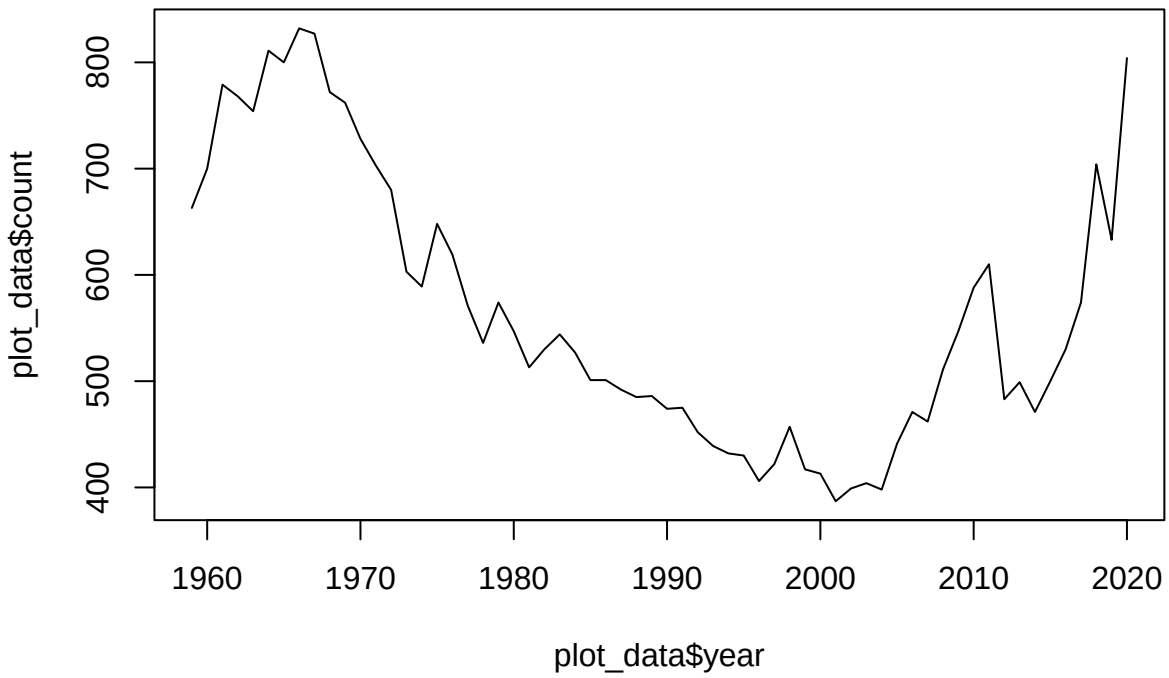
8/15/2022

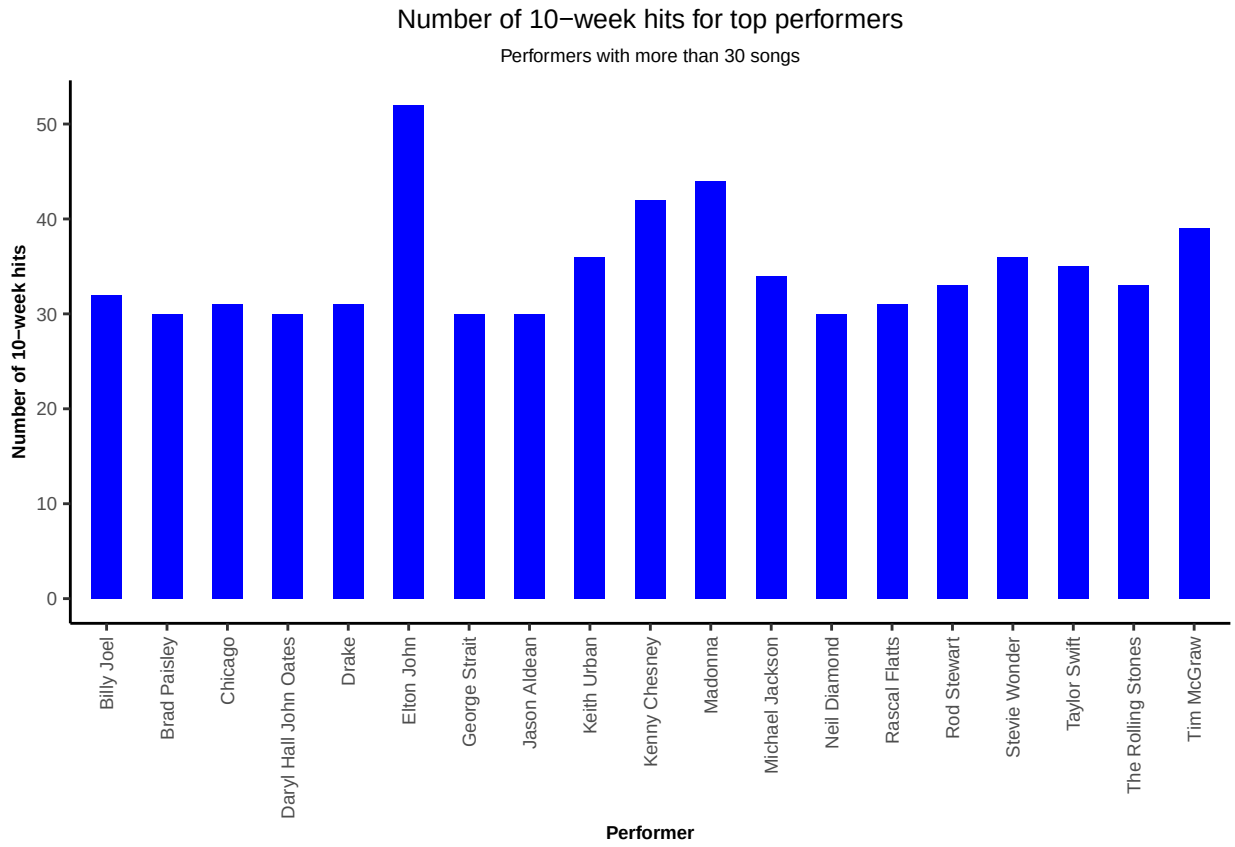
```
## 71.43 % of people who are truthful clickers answered yes.
```

```
## The probability that they have the disease given that they test postitive is 19.89%
```

```
## The top 10 most popular songs since 1958 are as below:
```

```
##                                     performer
## 1                             Imagine Dragons
## 2                                  AWOLNATION
## 3                                  The Weeknd
## 4                                  Jason Mraz
## 5                                 LeAnn Rimes
## 6                                  OneRepublic
## 7  LMFAO Featuring Lauren Bennett & GoonRock
## 8                                       Adele
## 9                                       Jewel
## 10                          Carrie Underwood
##                                     song count
## 1                            Radioactive    87
## 2                                   Sail    79
## 3                          Blinding Lights    76
## 4                                I'm Yours    76
## 5                           How Do I Live    69
## 6                           Counting Stars    68
## 7                       Party Rock Anthem    68
## 8                     Rolling In The Deep    65
## 9  Foolish Games/You Were Meant For Me    65
## 10                       Before He Cheats    64
```


plot_data$count
400
500
600
700
800
1960
1970
1980
1990
2000
2010
2020
plot_data$year



The summary statitics of the green buildings data is as below:

```
##  CS_PropertyID        cluster           size            empl_gr
##  Min.   :      1   Min.   :   1.0   Min.   :   1624   Min.   :-24.950
##  1st Qu.: 157452   1st Qu.: 272.0   1st Qu.:  50891   1st Qu.:  1.740
##  Median : 313253   Median : 476.0   Median : 128838   Median :  1.970
##  Mean   : 453003   Mean   : 588.6   Mean   : 234638   Mean   :  3.207
##  3rd Qu.: 441188   3rd Qu.:1044.0   3rd Qu.: 294212   3rd Qu.:  2.380
##  Max.   :6208103   Max.   :1230.0   Max.   :3781045   Max.   : 67.780
##                                                       NA's   :74
##       Rent         leasing_rate       stories            age
##  Min.   :  2.98   Min.   :  0.00   Min.   :  1.00   Min.   :  0.00
##  1st Qu.: 19.50   1st Qu.: 77.85   1st Qu.:  4.00   1st Qu.: 23.00
##  Median : 25.16   Median : 89.53   Median : 10.00   Median : 34.00
##  Mean   : 28.42   Mean   : 82.61   Mean   : 13.58   Mean   : 47.24
##  3rd Qu.: 34.18   3rd Qu.: 96.44   3rd Qu.: 19.00   3rd Qu.: 79.00
##  Max.   :250.00   Max.   :100.00   Max.   :110.00   Max.   :187.00
##
##    renovated          class_a          class_b           LEED
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.3795   Mean   :0.3999   Mean   :0.4595   Mean   :0.006841
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
```

```
##
##    Energystar       green_rating          net            amenities
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :1.0000
##  Mean   :0.08082   Mean   :0.08677   Mean   :0.03471   Mean   :0.5266
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##   cd_total_07     hd_total07    total_dd_07   Precipitation
##  Min.   :  39   Min.   :   0   Min.   :2103   Min.   :10.46
##  1st Qu.: 684   1st Qu.:1419   1st Qu.:2869   1st Qu.:22.71
##  Median : 966   Median :2739   Median :4979   Median :23.16
##  Mean   :1229   Mean   :3432   Mean   :4661   Mean   :31.08
##  3rd Qu.:1620   3rd Qu.:4796   3rd Qu.:6413   3rd Qu.:43.89
##  Max.   :5240   Max.   :7200   Max.   :8244   Max.   :58.02
##
##    Gas_Costs        Electricity_Costs  cluster_rent
##  Min.   :0.009487   Min.   :0.01780   Min.   : 9.00
##  1st Qu.:0.010296   1st Qu.:0.02330   1st Qu.:20.00
##  Median :0.010296   Median :0.03274   Median :25.14
##  Mean   :0.011336   Mean   :0.03096   Mean   :27.50
##  3rd Qu.:0.011816   3rd Qu.:0.03781   3rd Qu.:34.00
##  Max.   :0.028914   Max.   :0.06280   Max.   :71.44
##
```

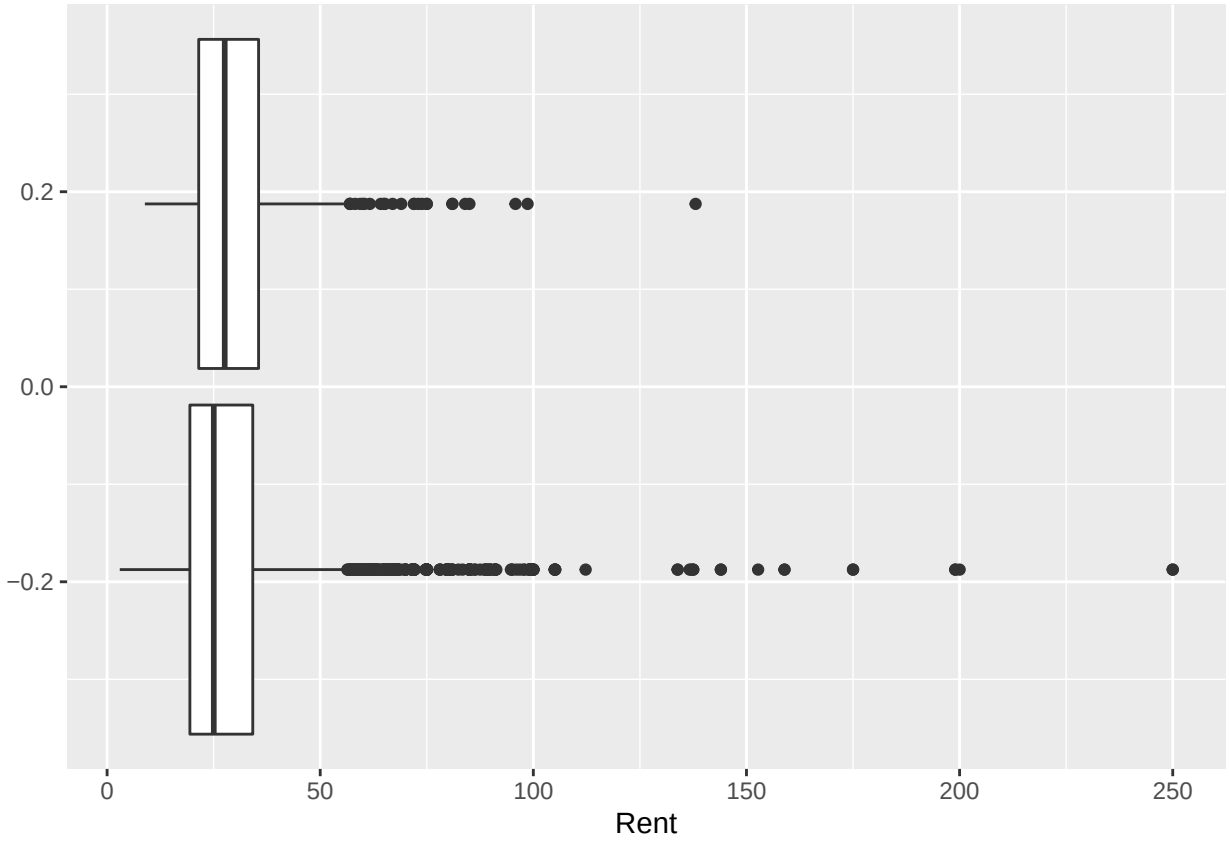
The summary statitics of the leasing rate (occupancy) of the green buildings data is as below:

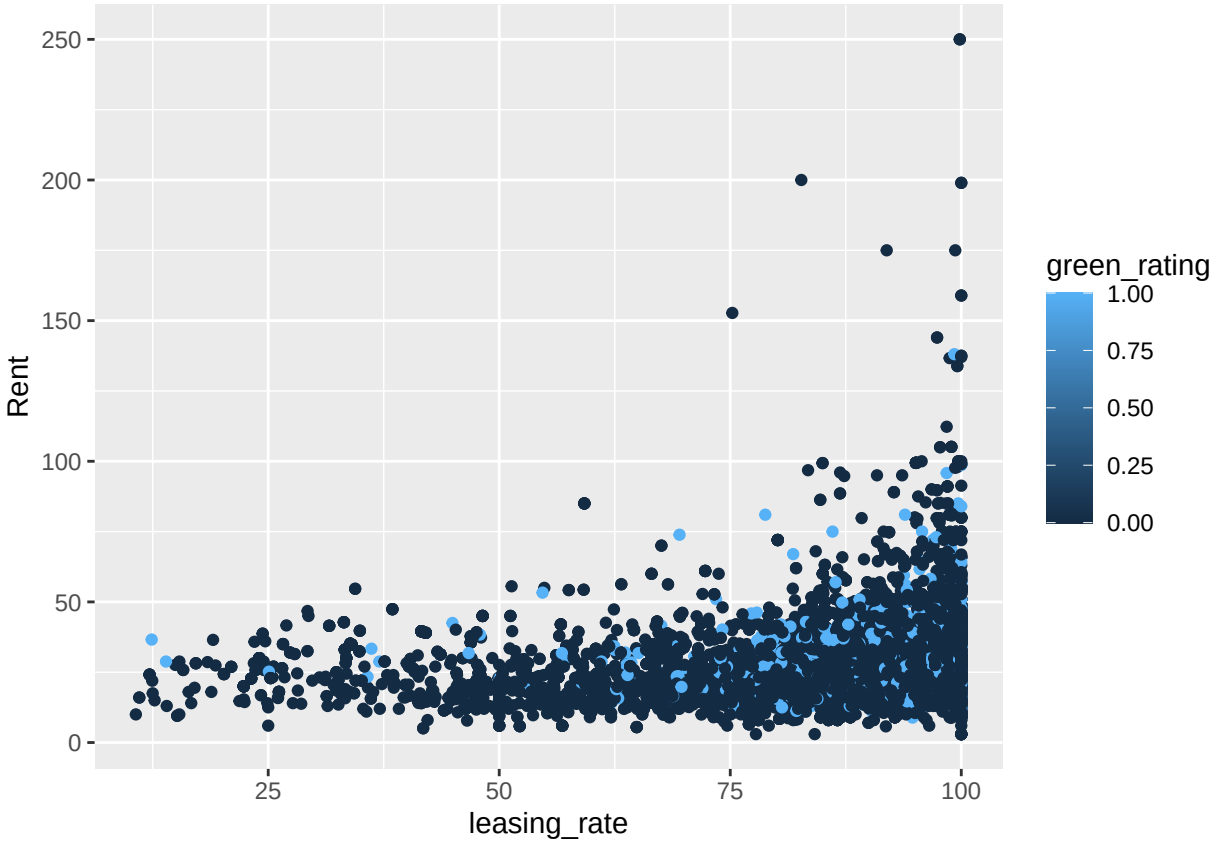
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   77.85   89.53   82.61   96.44  100.00
```

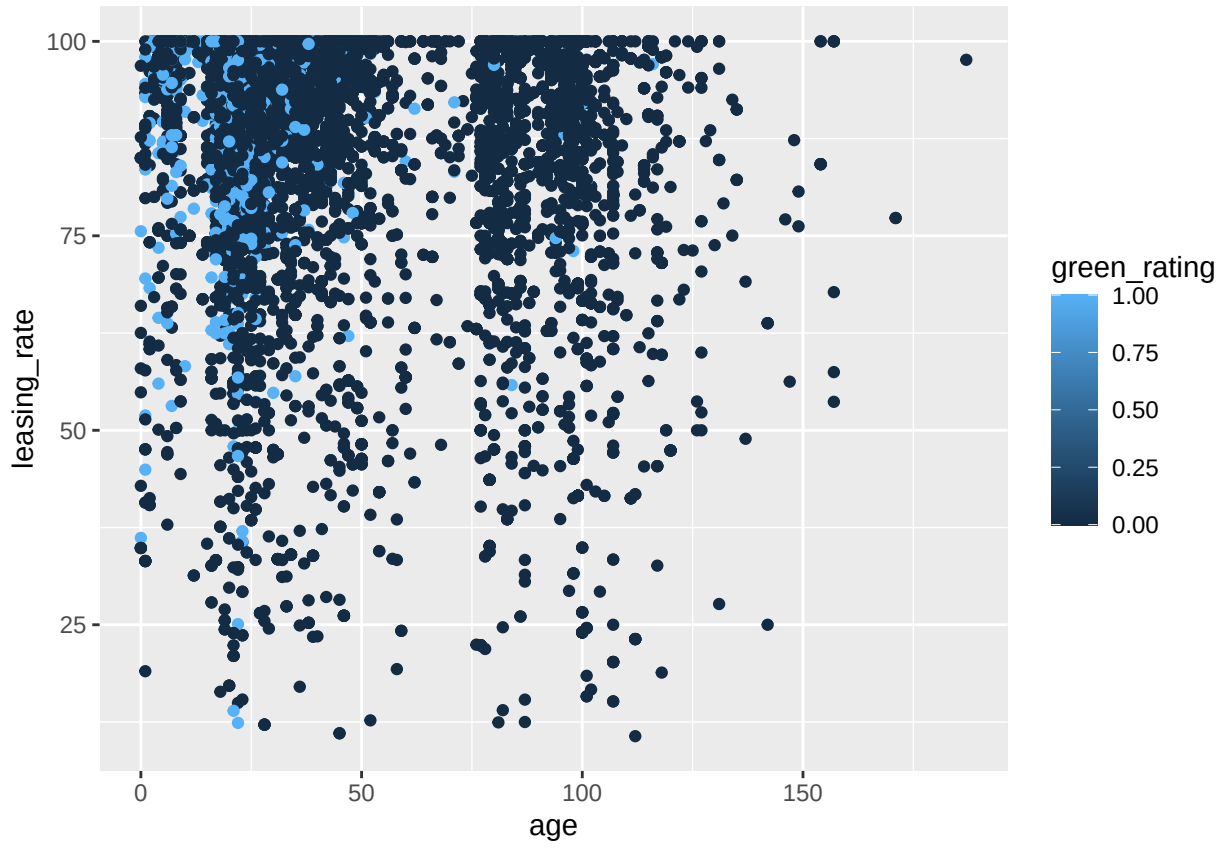
The 2.72% quantile of the leasing rate (occupancy) of the green buildings data is:

```
##    2.72%
## 10.40064
```

```
## # A tibble: 2 x 3
##   green_rating med_rent count
##          <int>    <dbl> <int>
## 1            0     25.0  6995
## 2            1     27.6   684
```

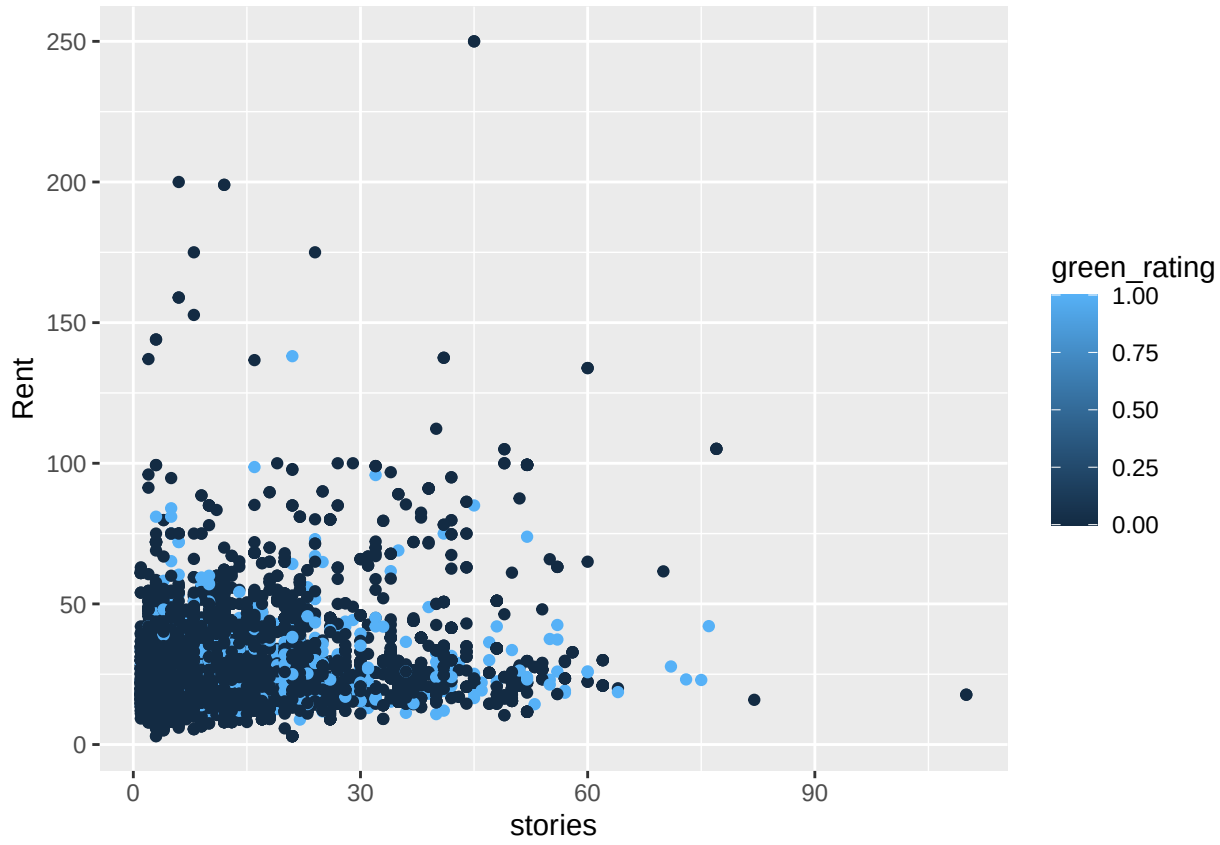

0.2
0.0
−0.2
0
50
100
150
200
250
Rent


250
200
150
100
50
0
Rent
25
50
75
100
leasing_rate
green_rating
1.00
0.75
0.50
0.25
0.00

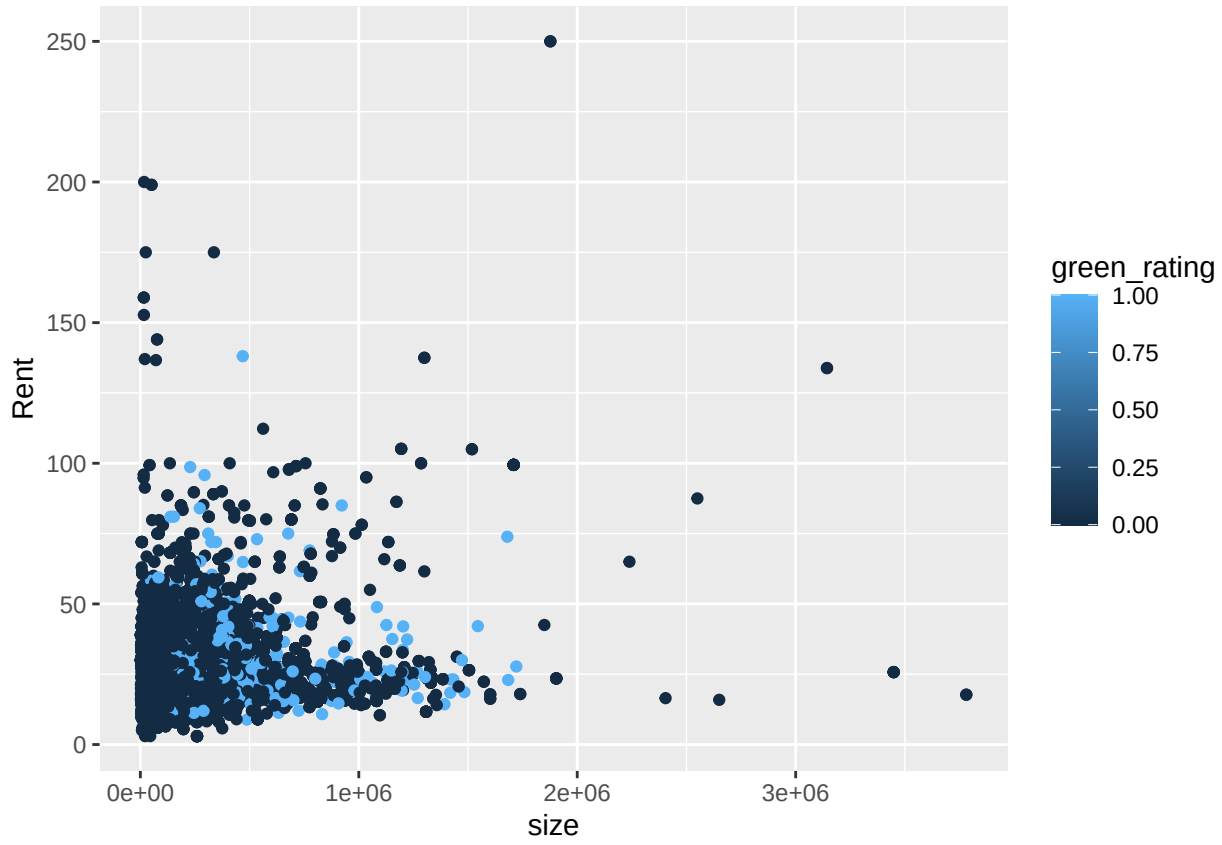



```
## `summarise()` has grouped output by 'green_rating'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 4
## # Groups:   green_rating [2]
##   green_rating amenities med_rent count
##          <int>     <int>    <dbl> <int>
## 1            0         0     25.1  3362
## 2            0         1     25    3633
## 3            1         0     27.0   186
## 4            1         1     27.8   498
```

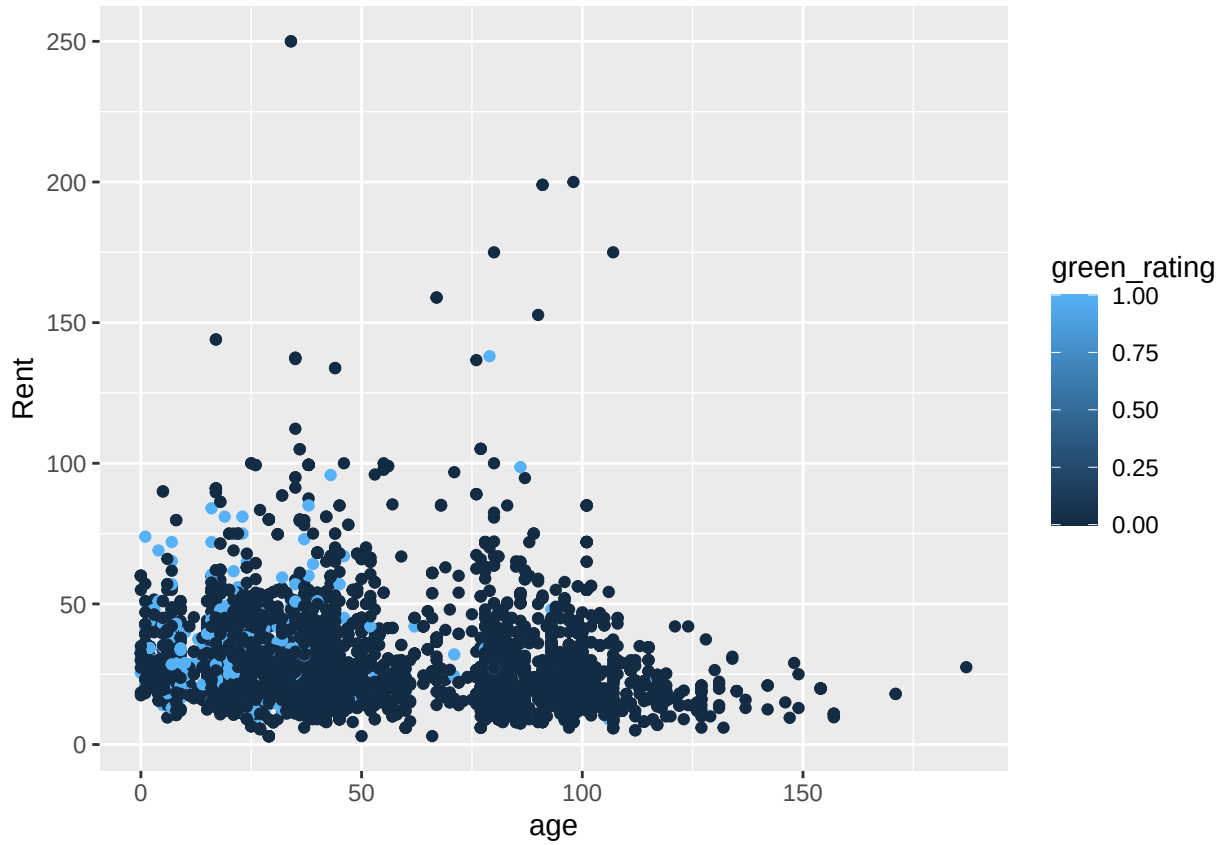



```
## # A tibble: 2 x 2
##   green_rating med_stories
##          <int>       <dbl>
## 1            0          10
## 2            1          11
```

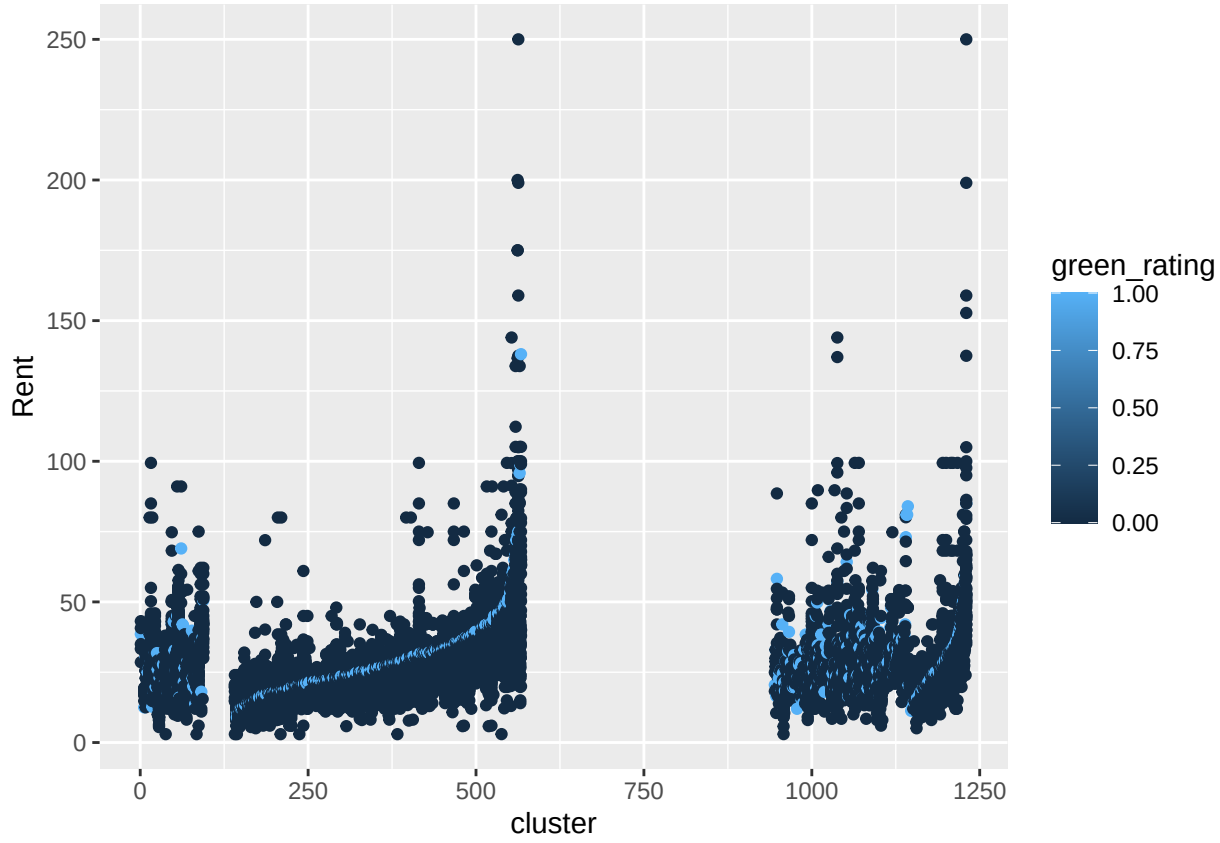



```
## # A tibble: 2 x 3
##   green_rating med_size mean_size
##          <int>    <dbl>     <dbl>
## 1            0   123250   231007.
## 2            1   241199   325965.
```

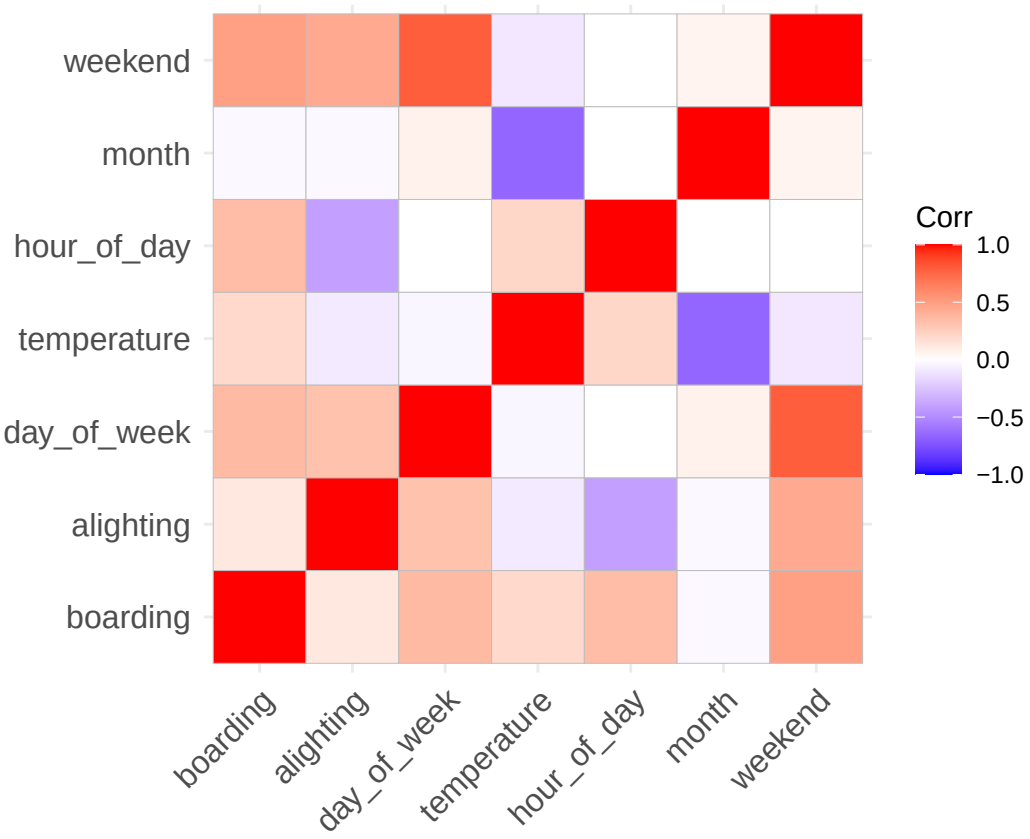
```
## # A tibble: 2 x 3
##   green_rating med_age count
##          <int>   <dbl> <int>
## 1            0      36  6995
## 2            1      22   684
```

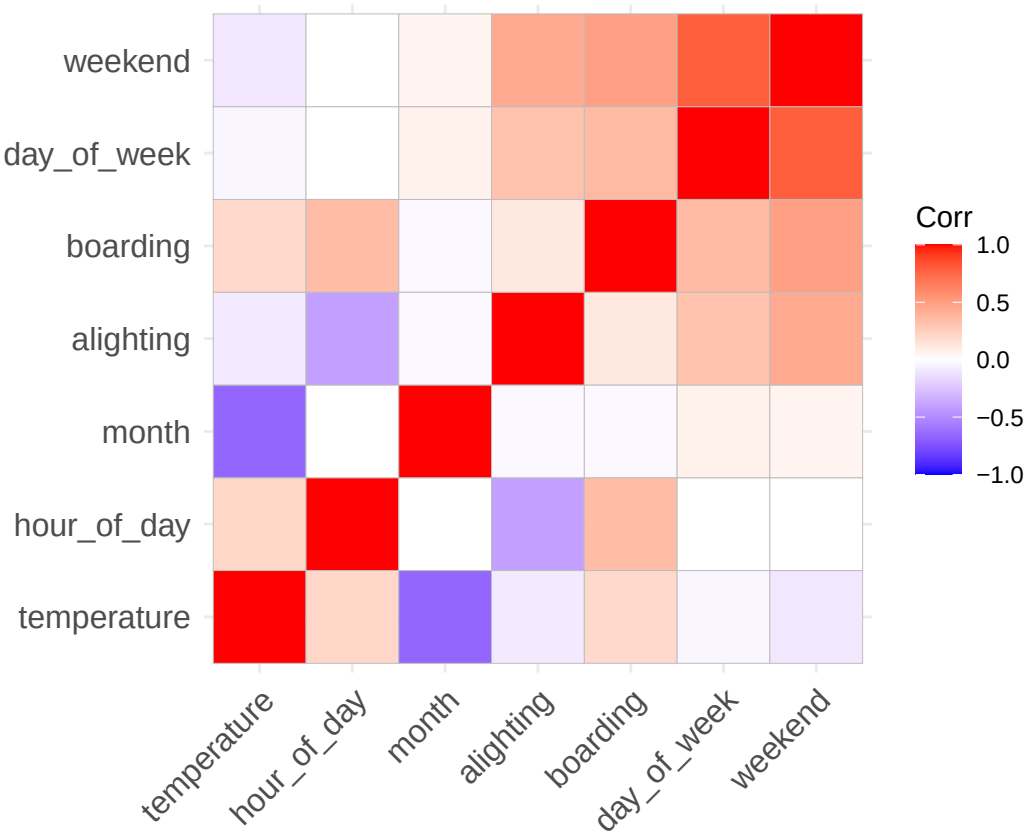


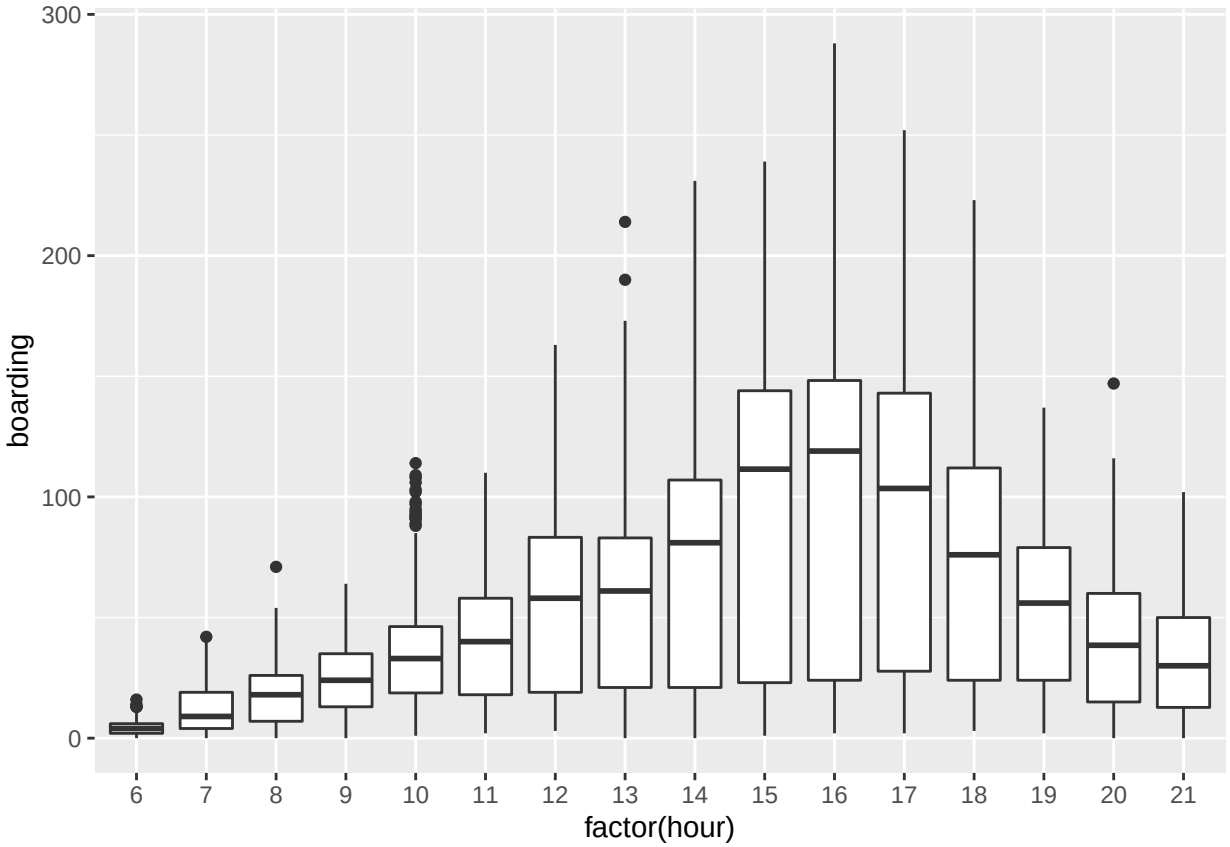

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      7998    987622   2937571   6718444   7957758 468492211
```

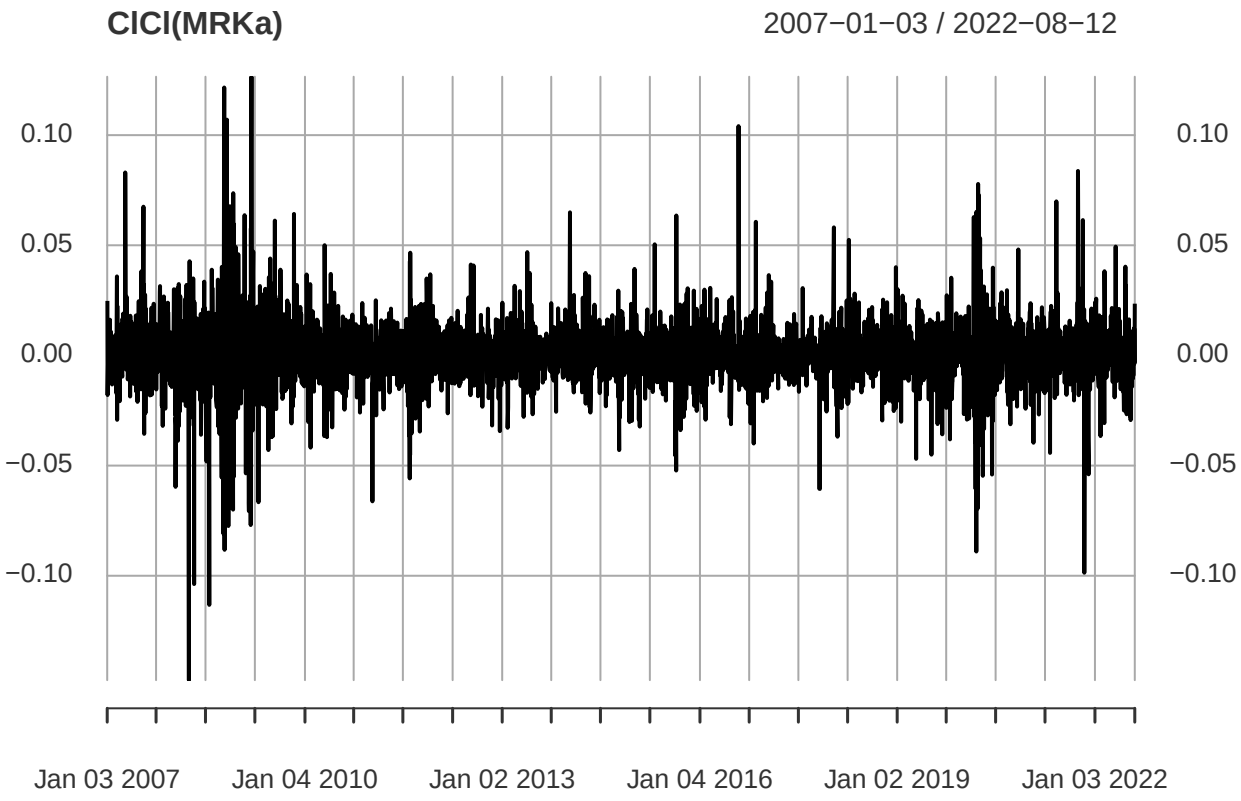


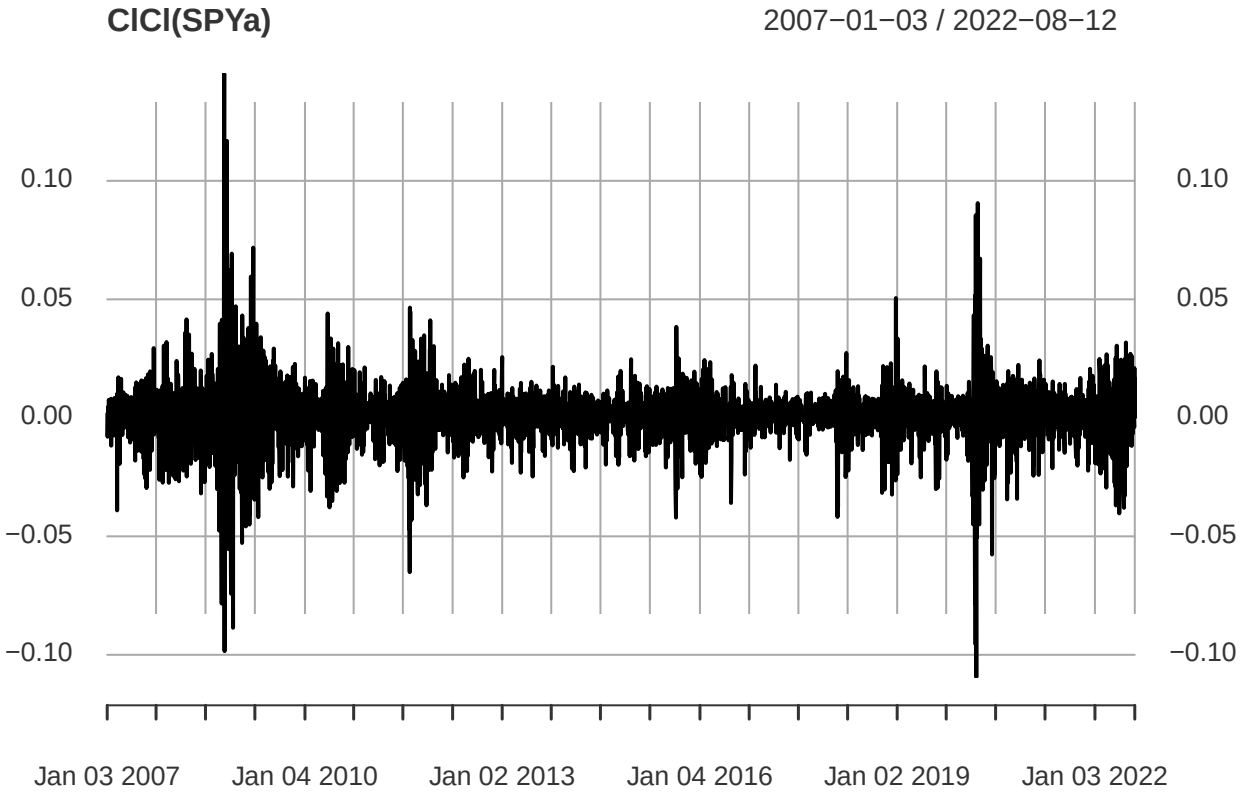

```
## # A tibble: 2 x 3
##   green_rating   med count
##          <int> <dbl> <int>
## 1            0  34.2  1466
## 2            1  39.4   131

## The median market rent for green buildings is $27.6 per square foot per year while the median market

## Hence, green buildings have median rents of $2.57 per square foot per year more.

##                boarding   alighting day_of_week temperature hour_of_day
## boarding     1.00000000  0.12022500  0.35841096  0.19758469   0.3519073
## alighting    0.12022500  1.00000000  0.32312023 -0.08640064  -0.4109057
## day_of_week  0.35841096  0.32312023  1.00000000 -0.04269766   0.0000000
## temperature  0.19758469 -0.08640064 -0.04269766  1.00000000   0.2100375
## hour_of_day  0.35190730 -0.41090568  0.00000000  0.21003748   1.0000000
## month       -0.02955535 -0.03449006  0.06766650 -0.66049069   0.0000000
## weekend      0.49110445  0.43798209  0.79056942 -0.09739292   0.0000000
##                   month     weekend
## boarding    -0.02955535  0.49110445
## alighting   -0.03449006  0.43798209
## day_of_week  0.06766650  0.79056942
## temperature -0.66049069 -0.09739292
## hour_of_day  0.00000000  0.00000000
## month        1.00000000  0.05991447
## weekend      0.05991447  1.00000000
```

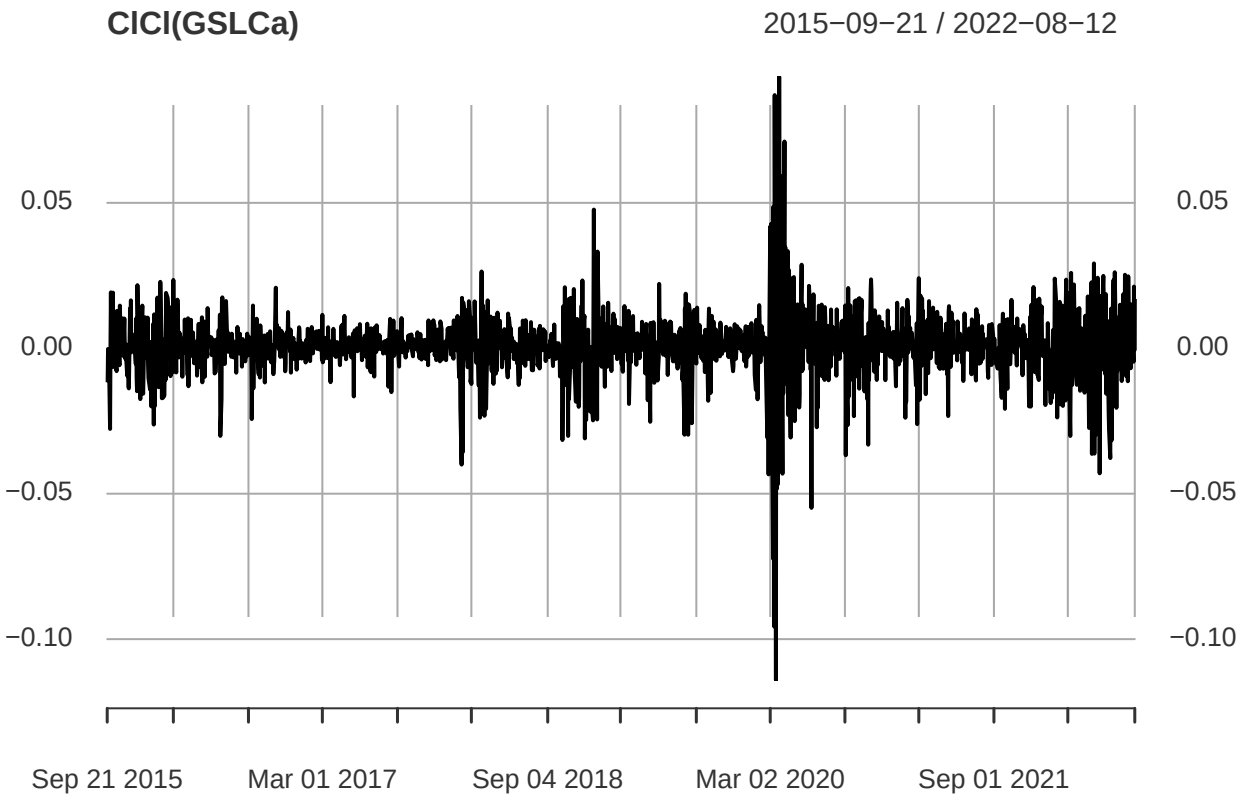

weekend
month
hour_of_day
temperature
day_of_week
alighting
boarding
boarding
alighting
day_of_week
temperature
hour_of_day
month
weekend
Corr
1.0
0.5
0.0
−0.5
−1.0

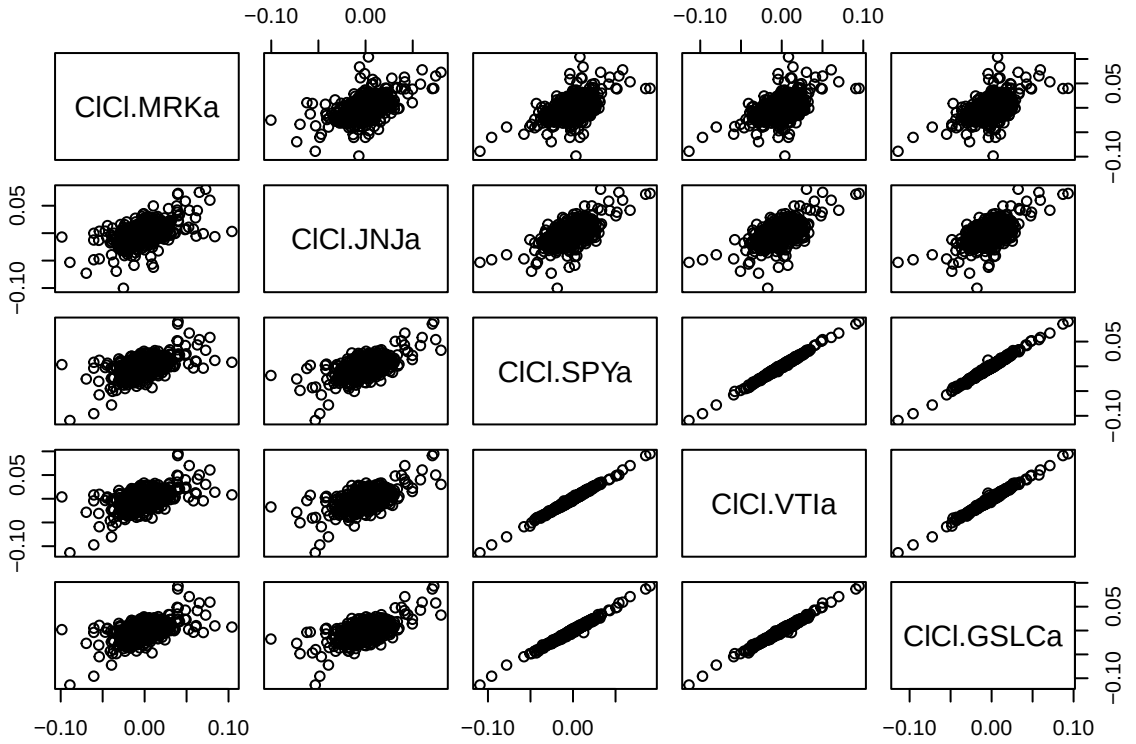

weekend
day_of_week
boarding
alighting
month
hour_of_day
temperature
temperature
hour_of_day
month
alighting
boarding
day_of_week
weekend
Corr
1.0
0.5
0.0
−0.5
−1.0

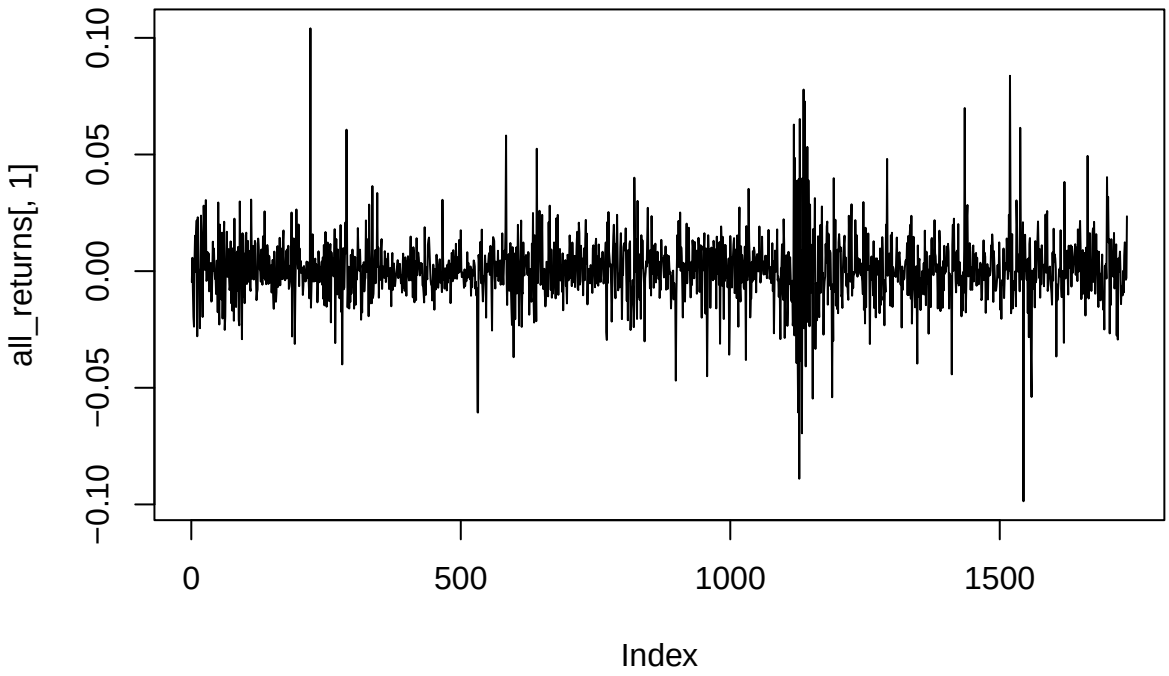

300
200
100
0
boarding
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
factor(hour)

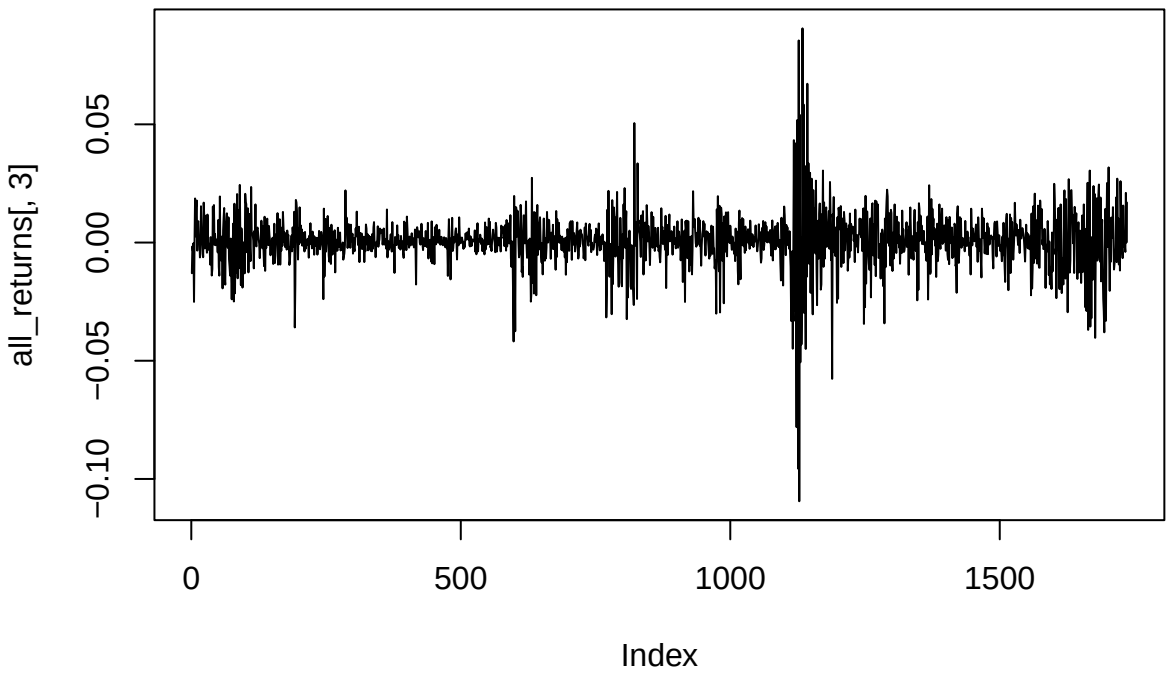

ClCl(MRKa)
2007−01−03 / 2022−08−12
0.10
0.05
0.00
−0.05
−0.10
Jan 03 2007
Jan 04 2010
Jan 02 2013
Jan 04 2016
Jan 02 2019
Jan 03 2022

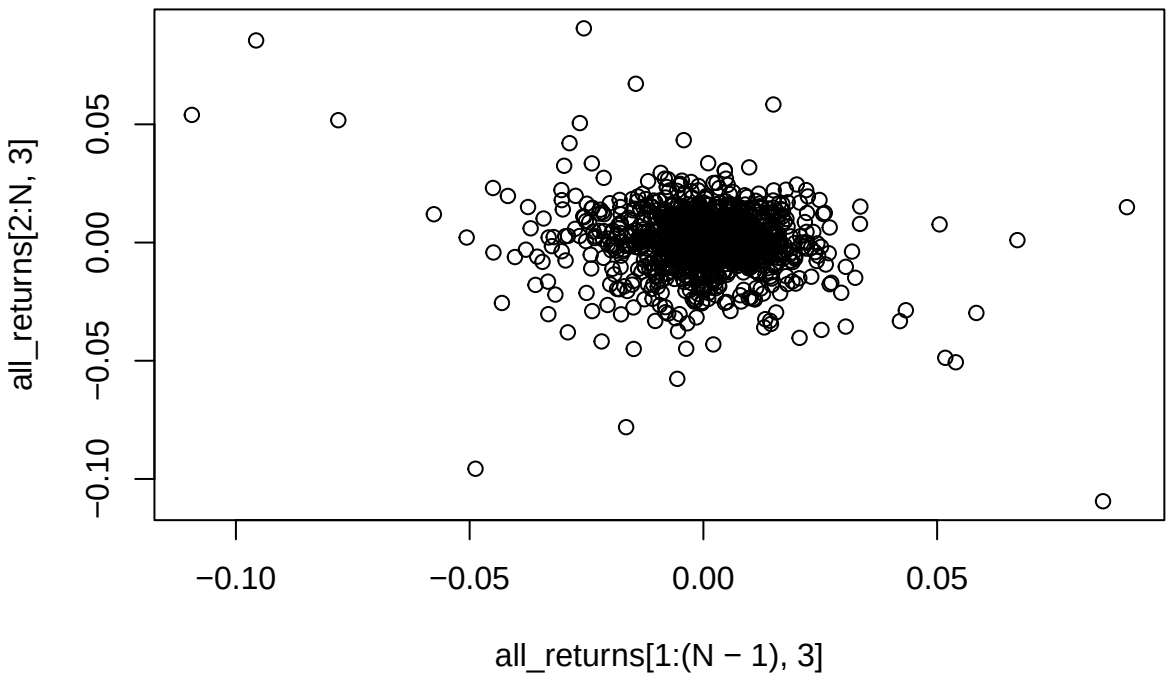

ClCl(SPYa)
2007−01−03 / 2022−08−12
0.10
0.05
0.00
−0.05
−0.10
Jan 03 2007
Jan 04 2010
Jan 02 2013
Jan 04 2016
Jan 02 2019
Jan 03 2022

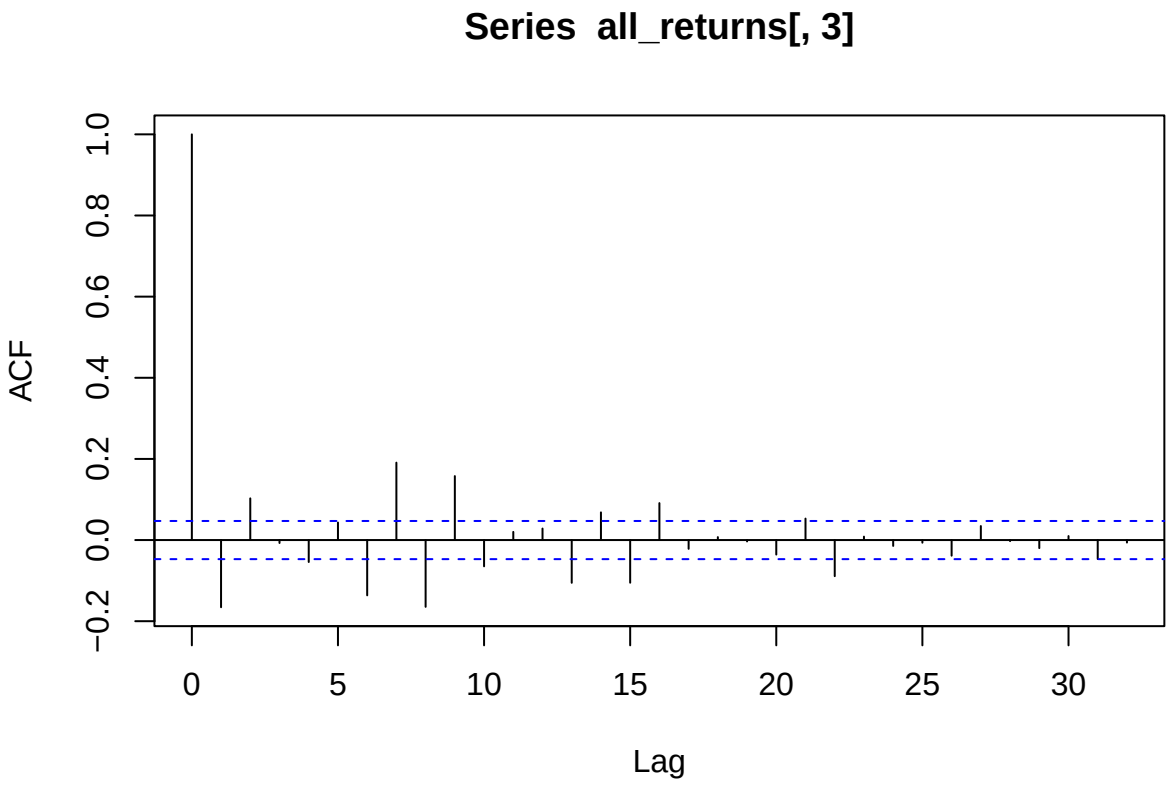

ClCl(GSLCa)
2015−09−21 / 2022−08−12
0.05
0.00
−0.05
−0.10
Sep 21 2015
Mar 01 2017
Sep 04 2018
Mar 02 2020
Sep 01 2021

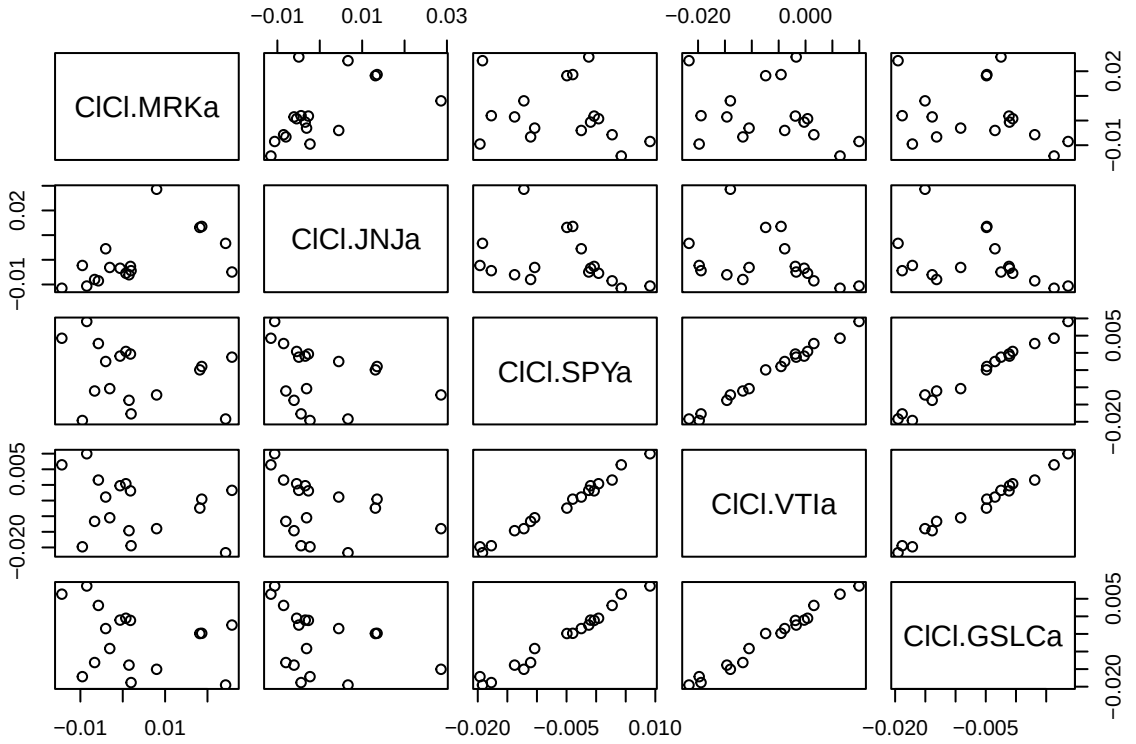

ClCl.MRKa
ClCl.JNJa
ClCl.SPYa
ClCl.VTIa
ClCl.GSLCa
−0.10
0.00
0.10
0.05


all_returns[, 1]
0.10
0.05
0.00
−0.05
−0.10
0
500
1000
1500
Index

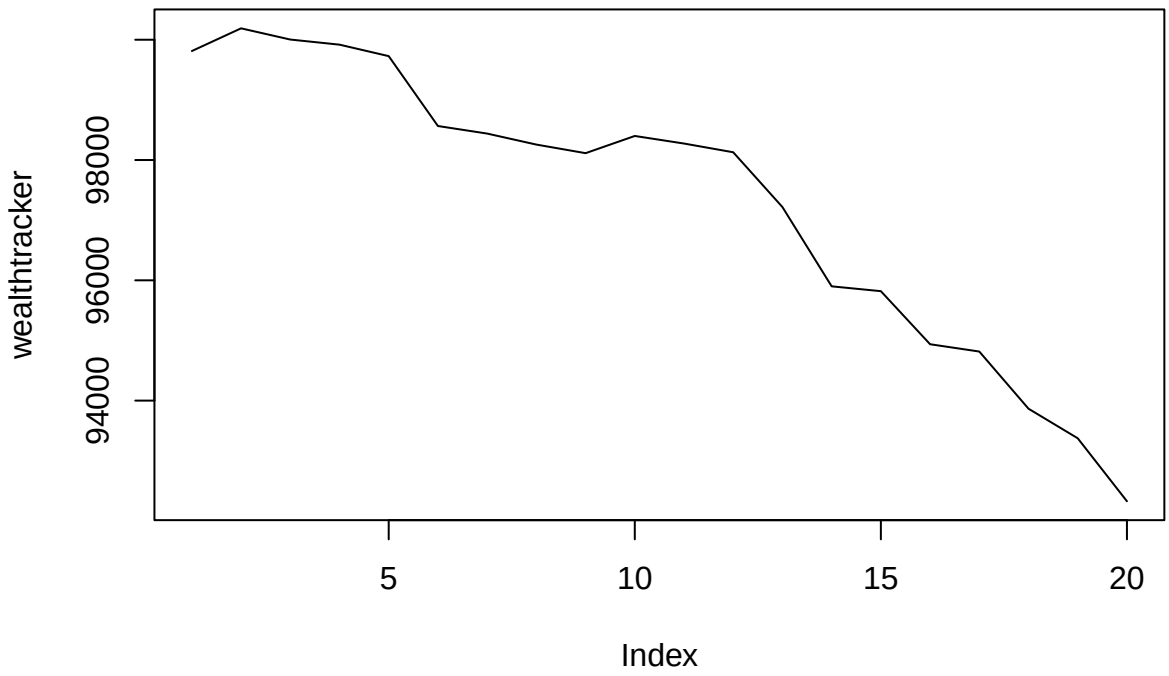

all_returns[, 3]
0.05
0.00
−0.05
−0.10
0
500
1000
1500
Index


all_returns[2:N, 3]
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
all_returns[1:(N − 1), 3]

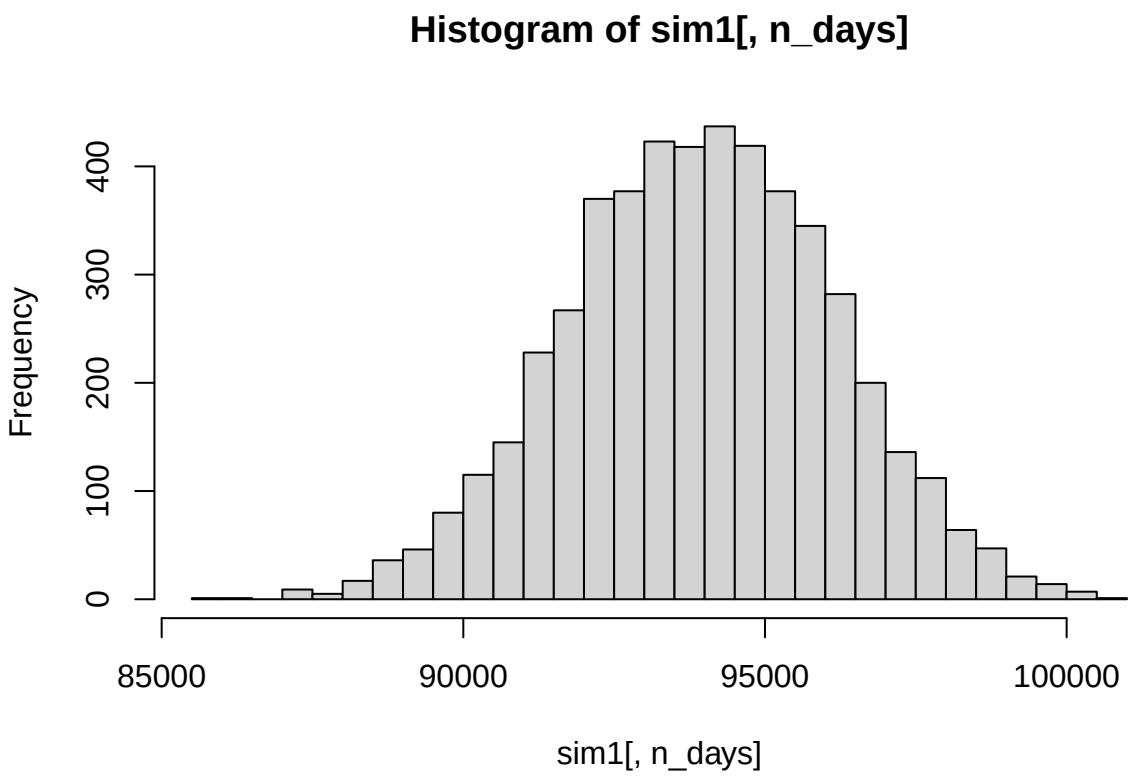

Series all_returns[, 3]
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
5
10
15
20
25
30
Lag

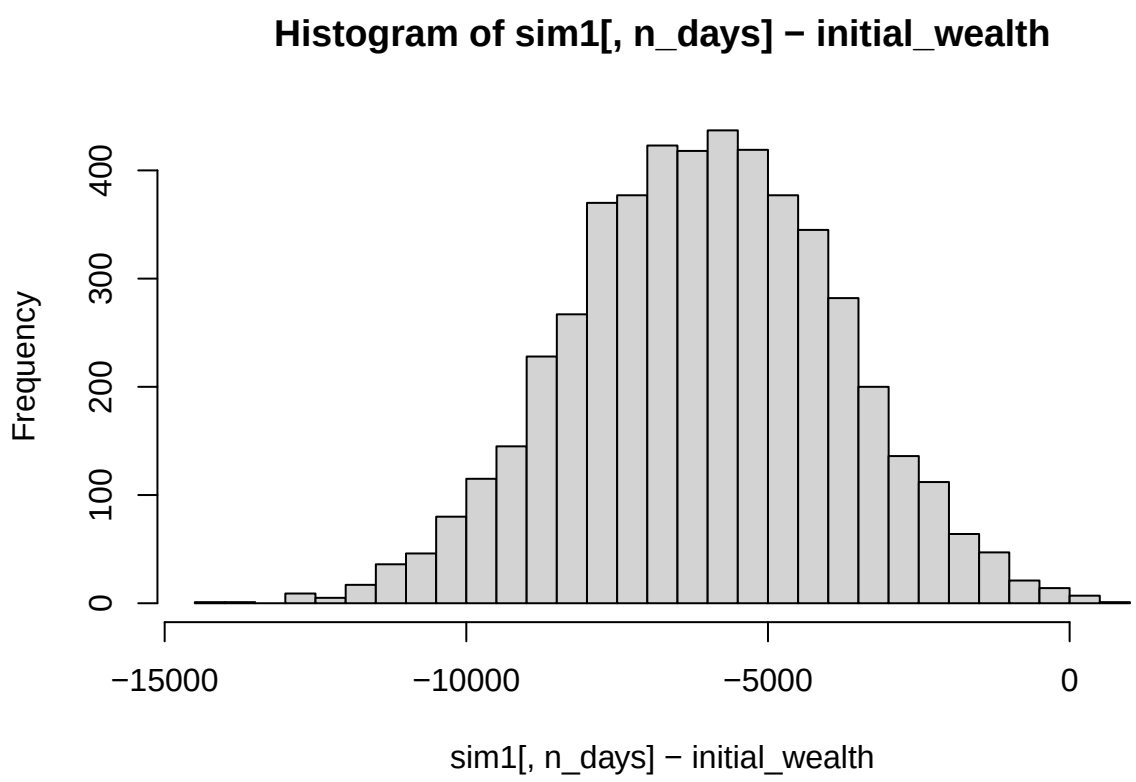

ClCl.MRKa
ClCl.JNJa
ClCl.SPYa
ClCl.VTIa
ClCl.GSLCa
−0.01
0.01
0.03
−0.020
0.000
0.02
−0.01
0.005
−0.020
−0.01
0.01
−0.020
−0.005
0.010
−0.020
−0.005

```
## [1] 92327.11
```

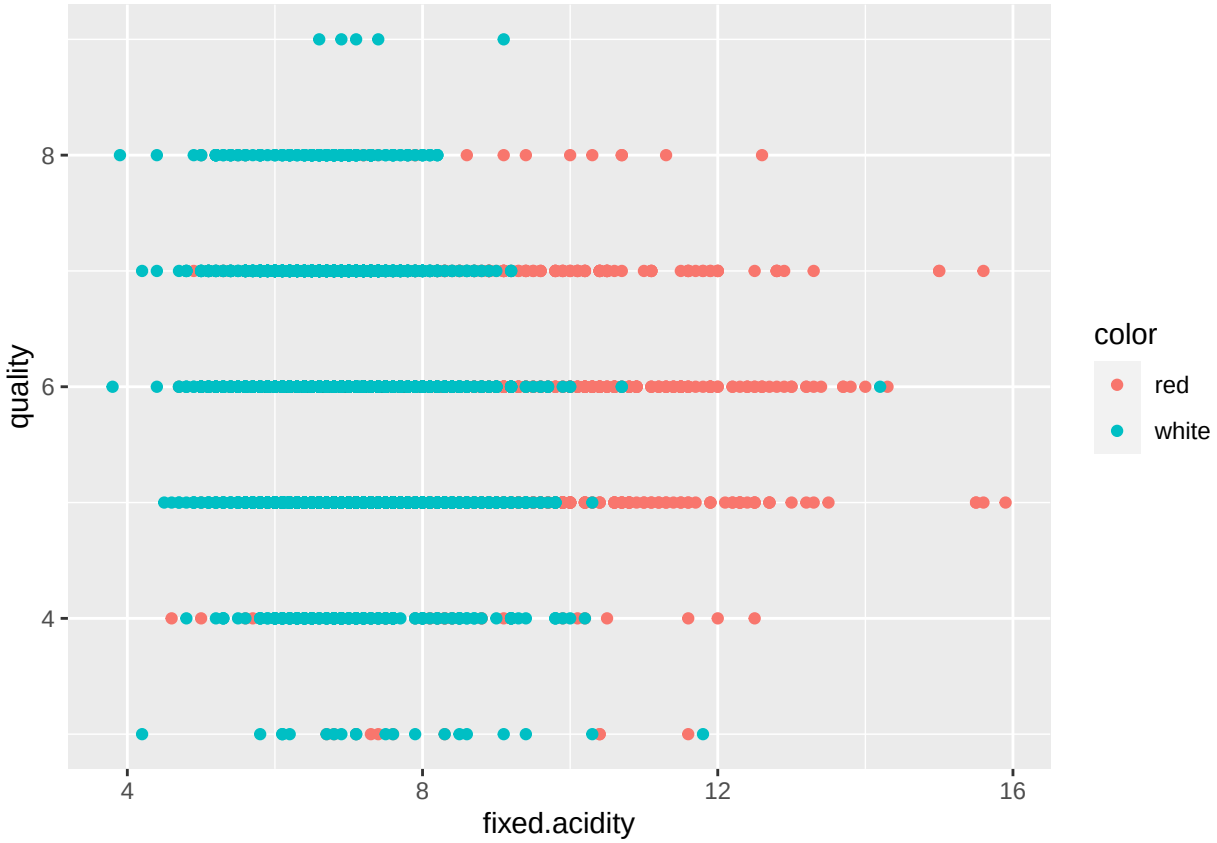


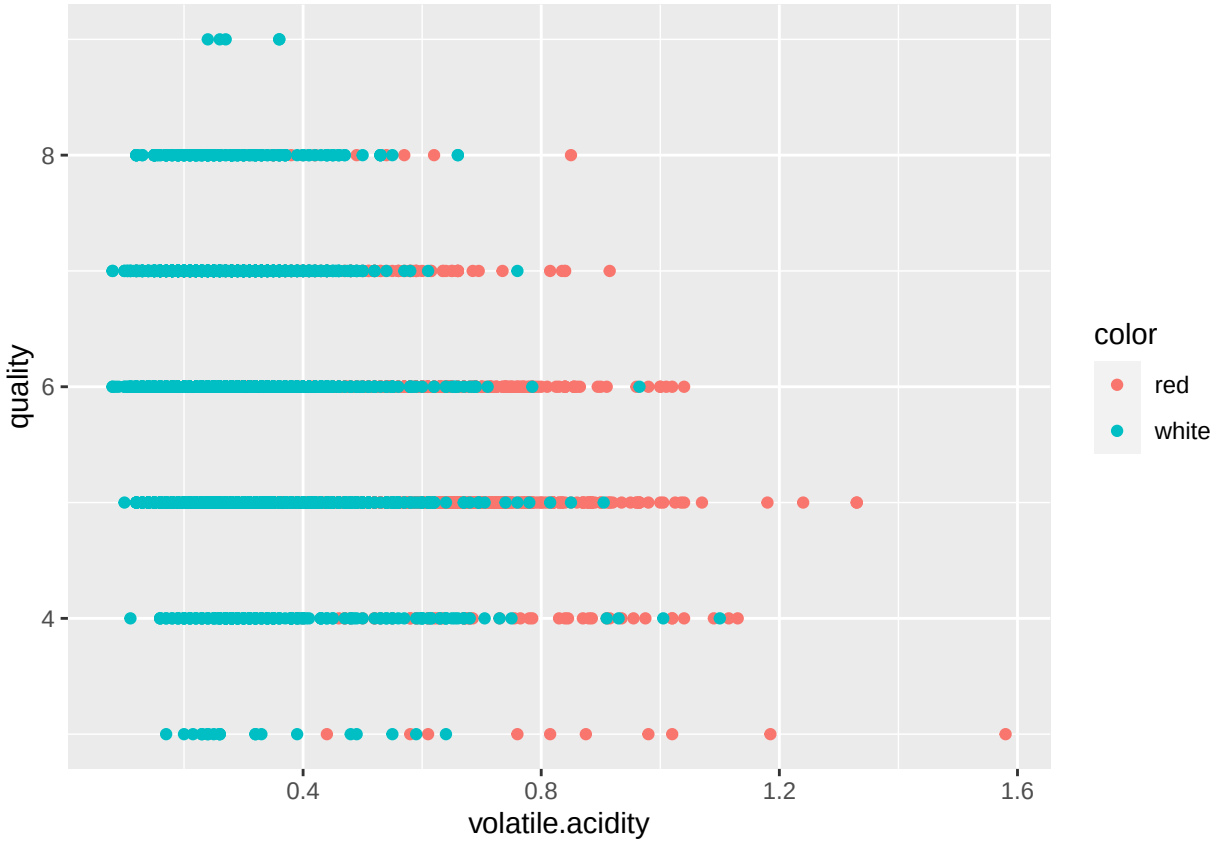

```
##               [,1]      [,2]      [,3]     [,4]      [,5]      [,6]      [,7]
## result.1 100376.73 100292.17  99129.68 97988.47  97287.89  96395.17  95096.96
## result.2 100281.66 100138.14 100011.34 99923.81 100320.71 100276.85 100405.18
## result.3  99821.70  98450.53  97863.90 97313.74  97189.36  97515.22  96411.06
## result.4  99275.62  99582.97  98429.20 97722.26  97538.79  96420.18  96297.27
## result.5  98990.05  98063.46  98224.67 98089.52  98360.56  98222.85  98500.68
## result.6  99821.70  98880.41  97525.28 97342.38  96638.90  95664.45  95542.57
##               [,8]      [,9]    [,10]    [,11]    [,12]    [,13]    [,14]
## result.1  94146.24  94025.10 94336.16 94667.81 94443.34 94764.91 94643.07
## result.2 100726.31 100013.05 99929.25 99799.21 99574.75 98446.01 98351.75
## result.3  96754.65  96668.15 95798.81 95914.42 94840.63 94813.39 94813.49
## result.4  95332.10  95466.99 95381.53 94698.09 94838.40 94652.78 93969.75
## result.5  98660.82  98941.58 98853.96 98772.40 99096.31 99397.25 99317.78
## result.6  95363.80  94673.30 94943.66 93987.73 92712.39 91458.74 90805.65
##              [,15]    [,16]    [,17]    [,18]    [,19]    [,20]
## result.1 93381.18 92954.90 92882.55 92065.46 91690.30 91570.90
## result.2 97037.06 96158.24 95932.31 95254.55 95097.67 94234.41
## result.3 94155.33 93502.88 93122.90 92484.51 92544.93 91634.46
## result.4 93892.44 93713.31 94071.30 94351.54 94672.31 94530.58
## result.5 99238.61 99090.78 97959.07 98099.92 98022.32 97878.99
## result.6 90665.43 90796.77 90145.17 90430.87 90798.43 89768.86
```

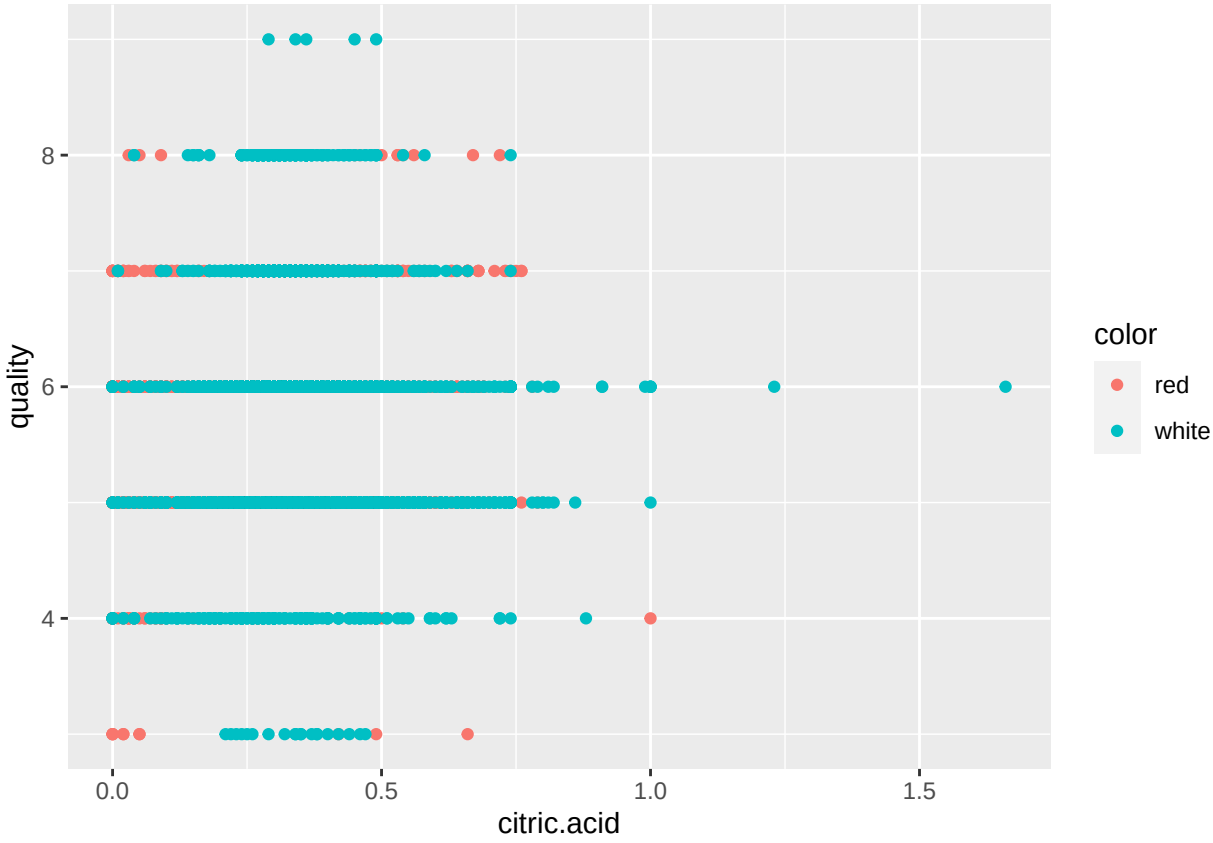

Histogram of sim1[, n_days]
Frequency
0
100
200
300
400
85000
90000
95000
100000
sim1[, n_days]

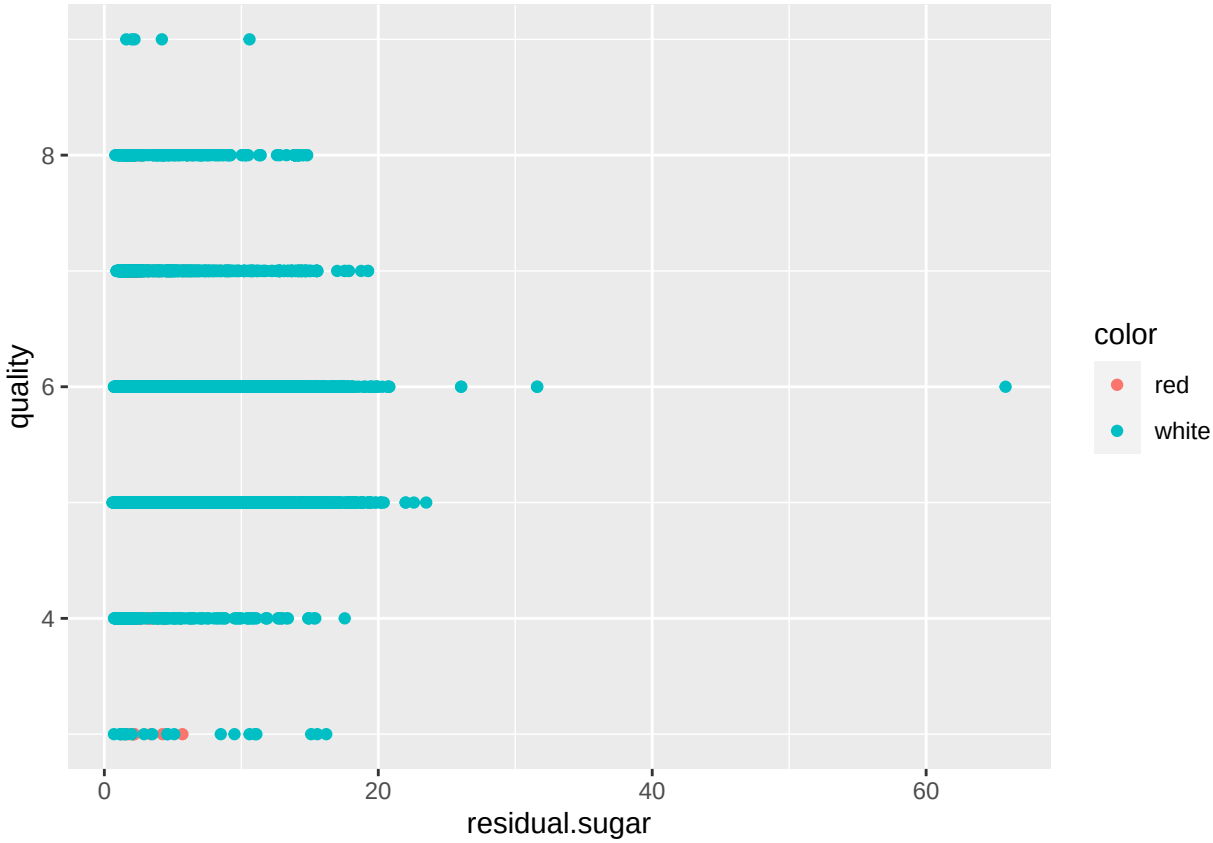

Histogram of sim1[, n_days] − initial_wealth
Frequency
0
100
200
300
400
−15000
−10000
−5000
0
sim1[, n_days] − initial_wealth

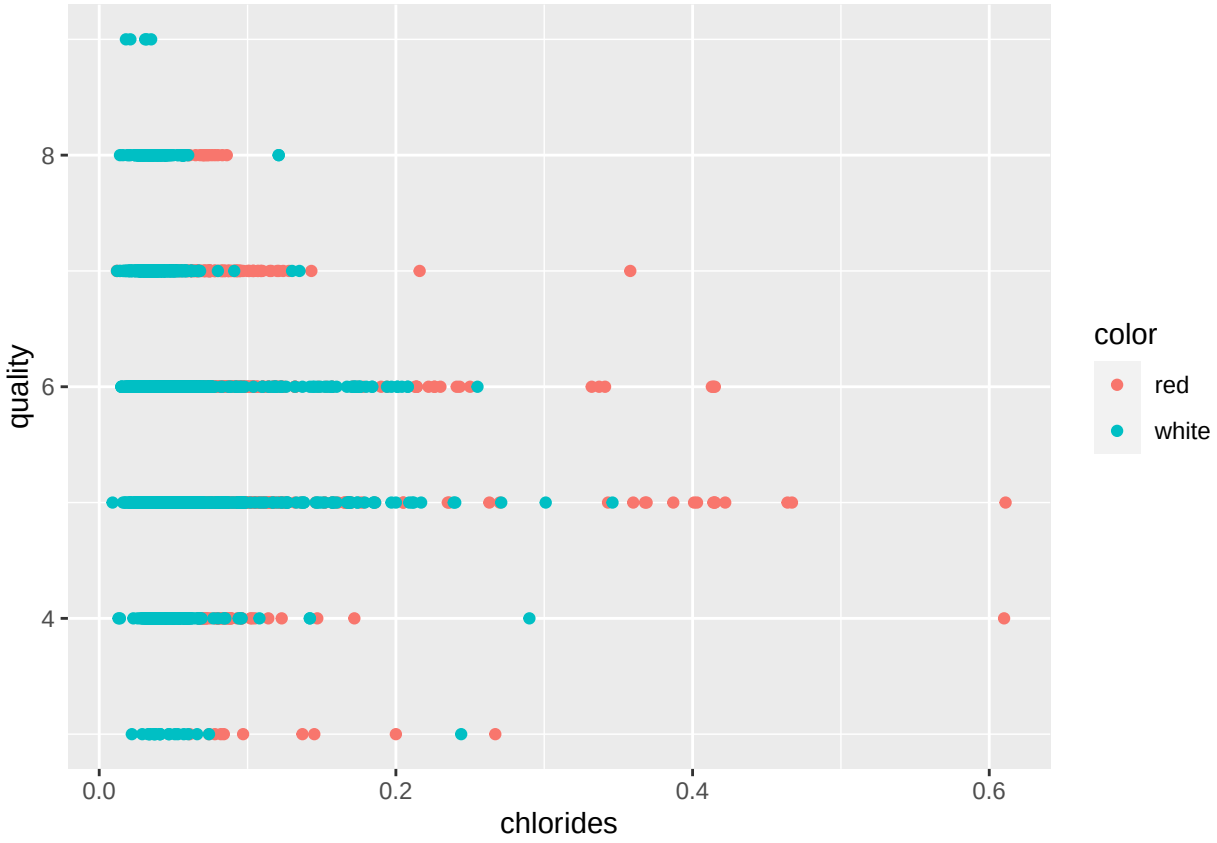
```
##       5%
## -9739.07
```

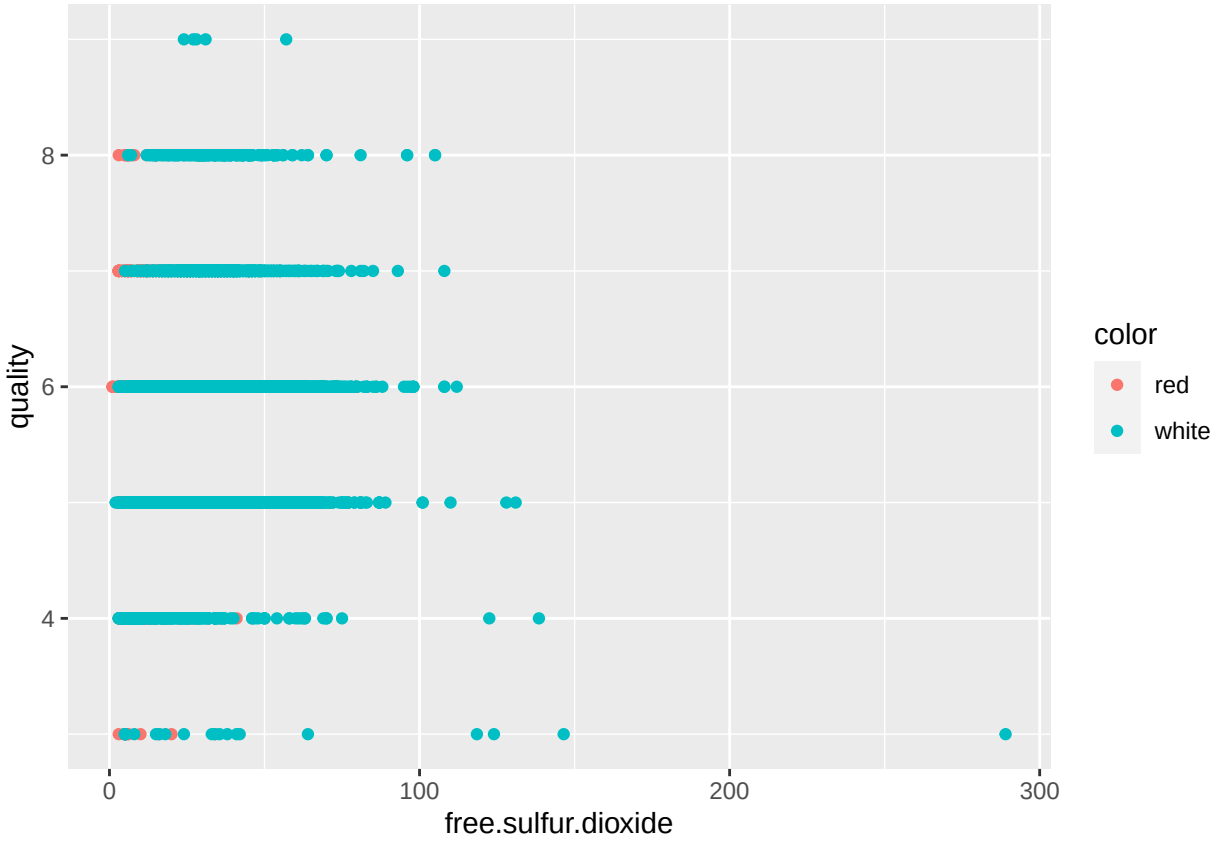

quality
8
6
4
4
8
12
16
fixed.acidity
color
red
white

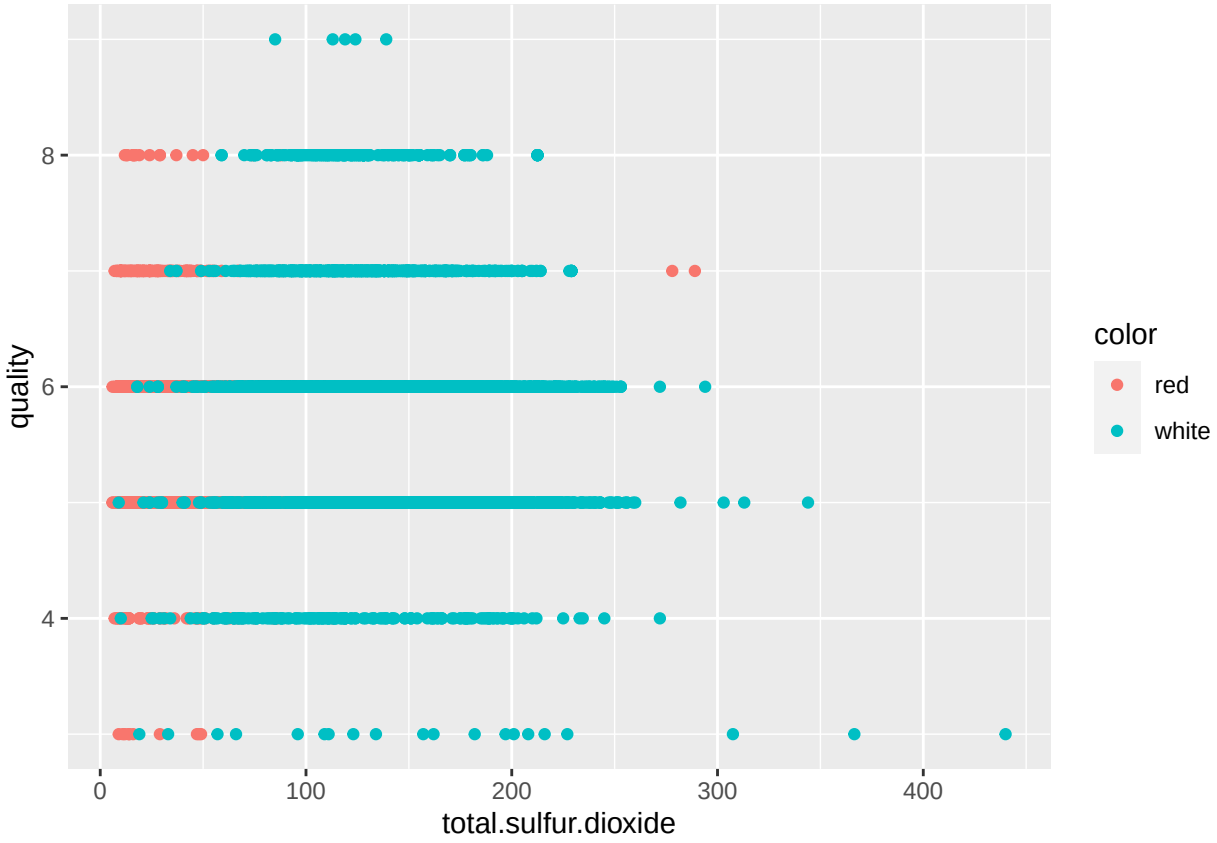

quality
8
6
4
0.4
0.8
1.2
1.6
volatile.acidity
color
red
white

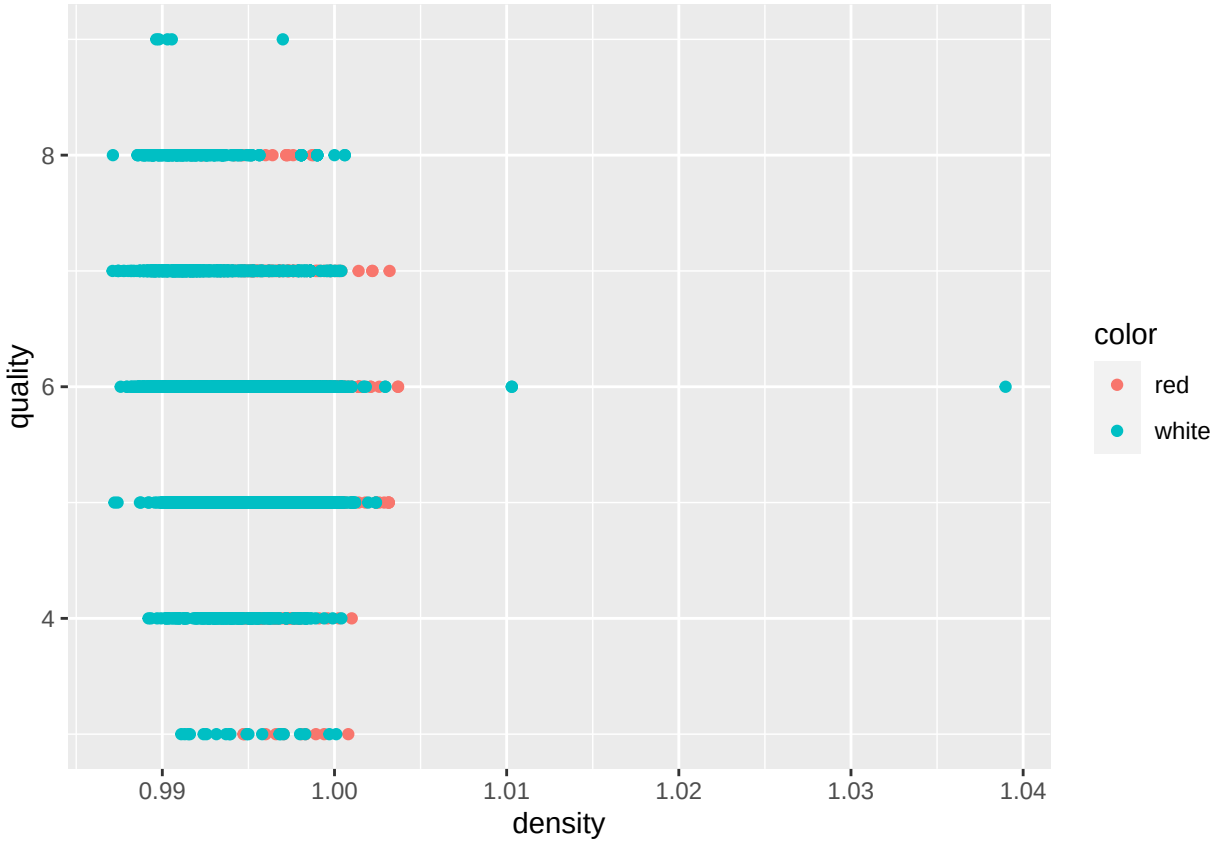

quality
8
6
4
0.0
0.5
1.0
1.5
citric.acid
color
red
white

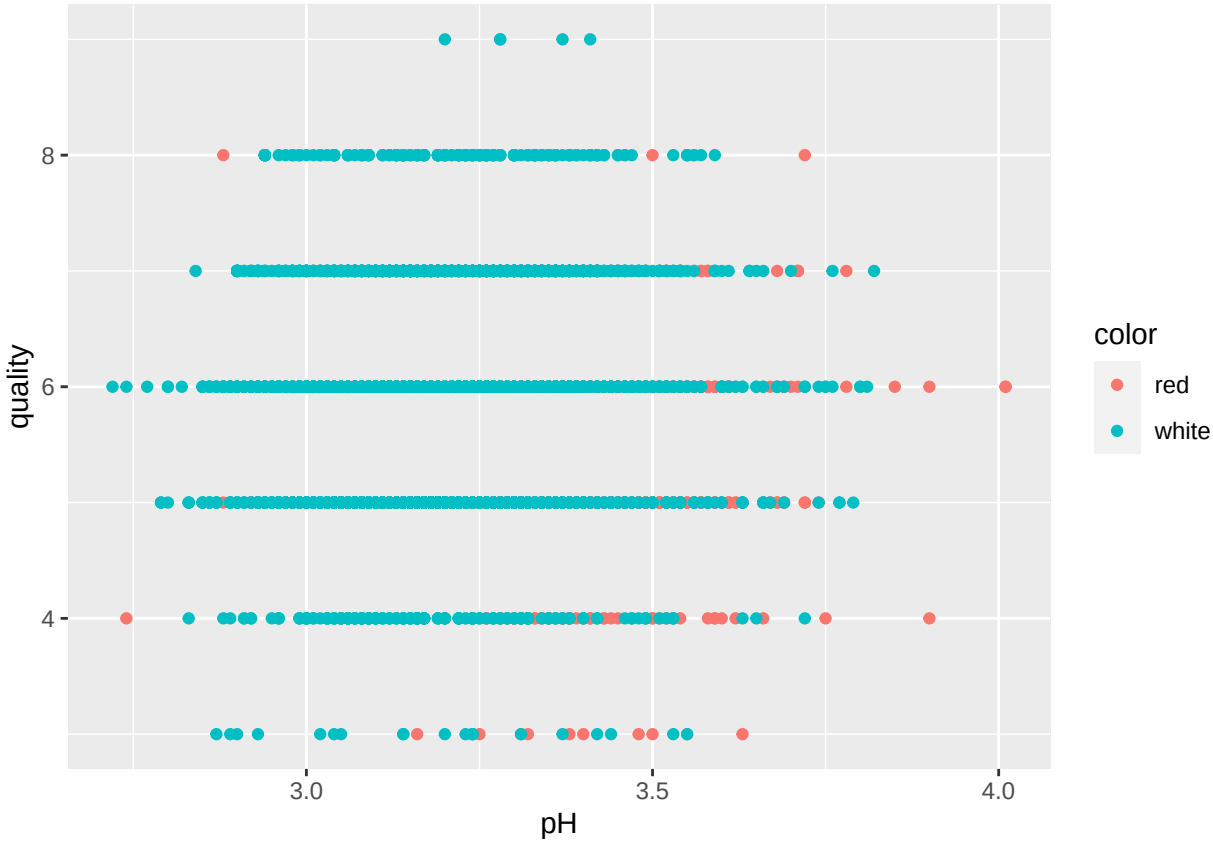

quality
8
6
4
0
20
40
60
residual.sugar
color
red
white

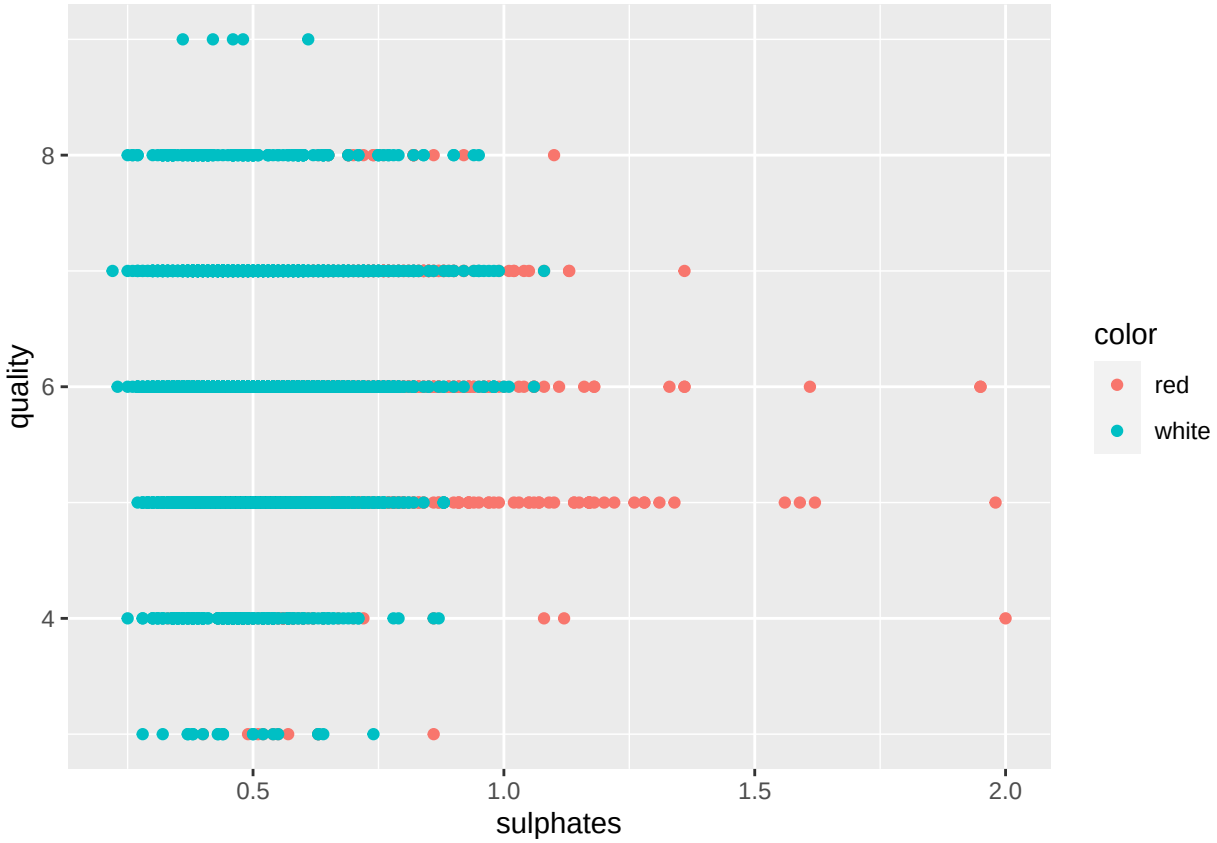

quality
8
6
4
0.0
0.2
0.4
0.6
chlorides
color
red
white

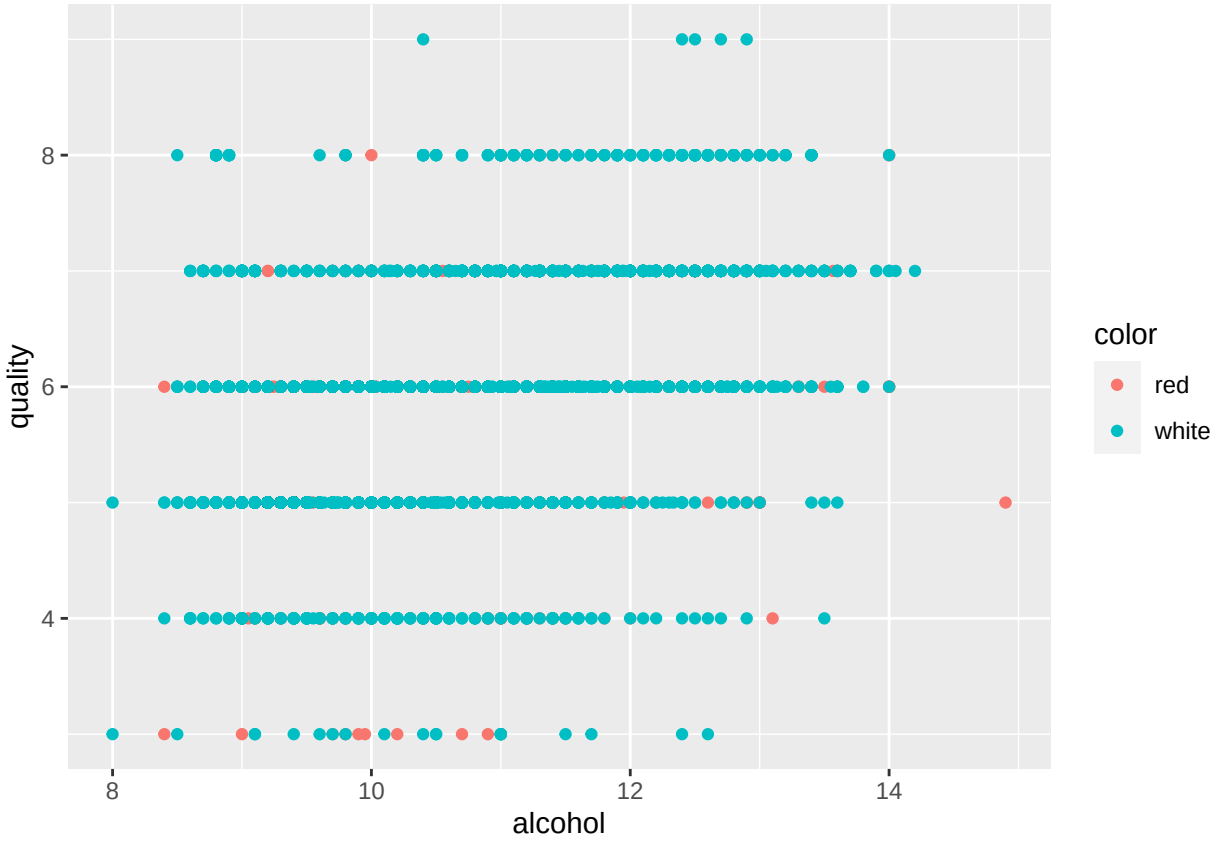

quality
8
6
4
0
100
200
300
free.sulfur.dioxide
color
red
white

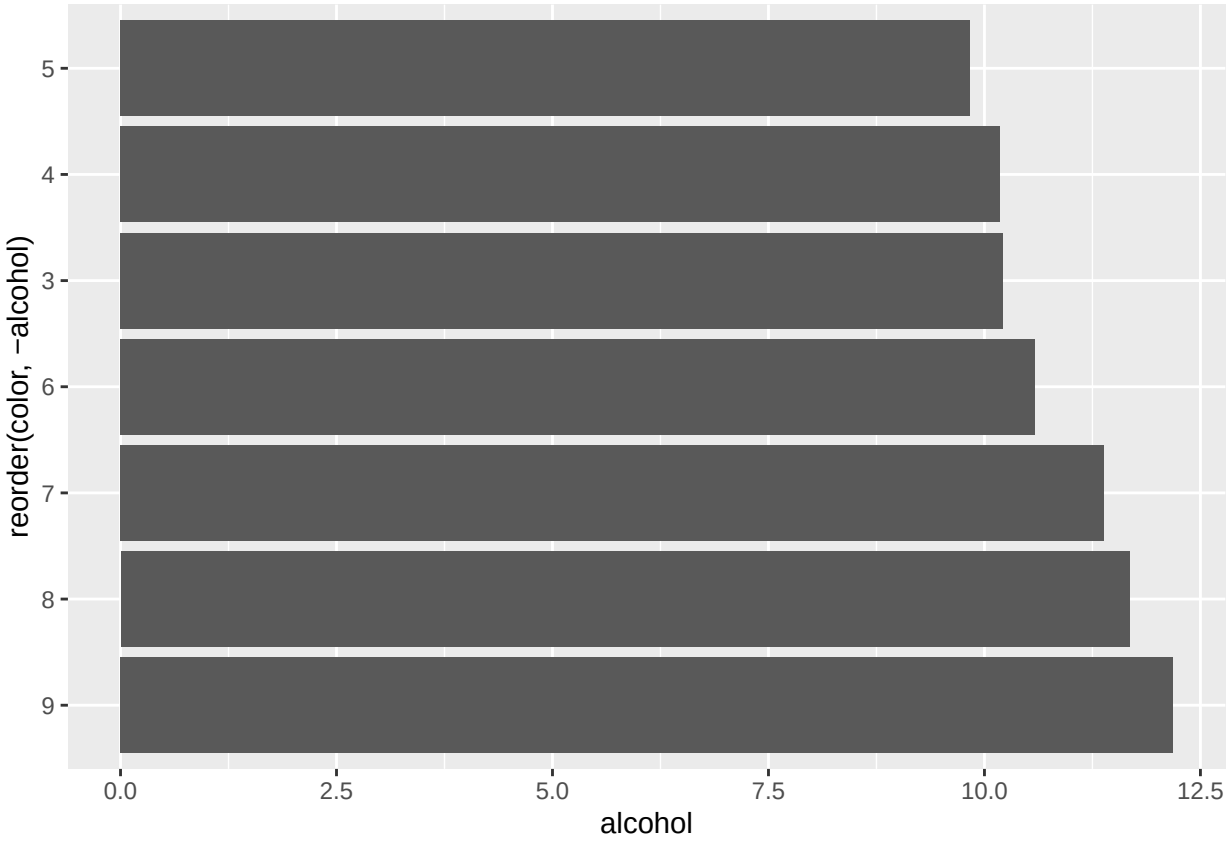

quality
8
6
4
0
100
200
300
400
total.sulfur.dioxide
color
red
white

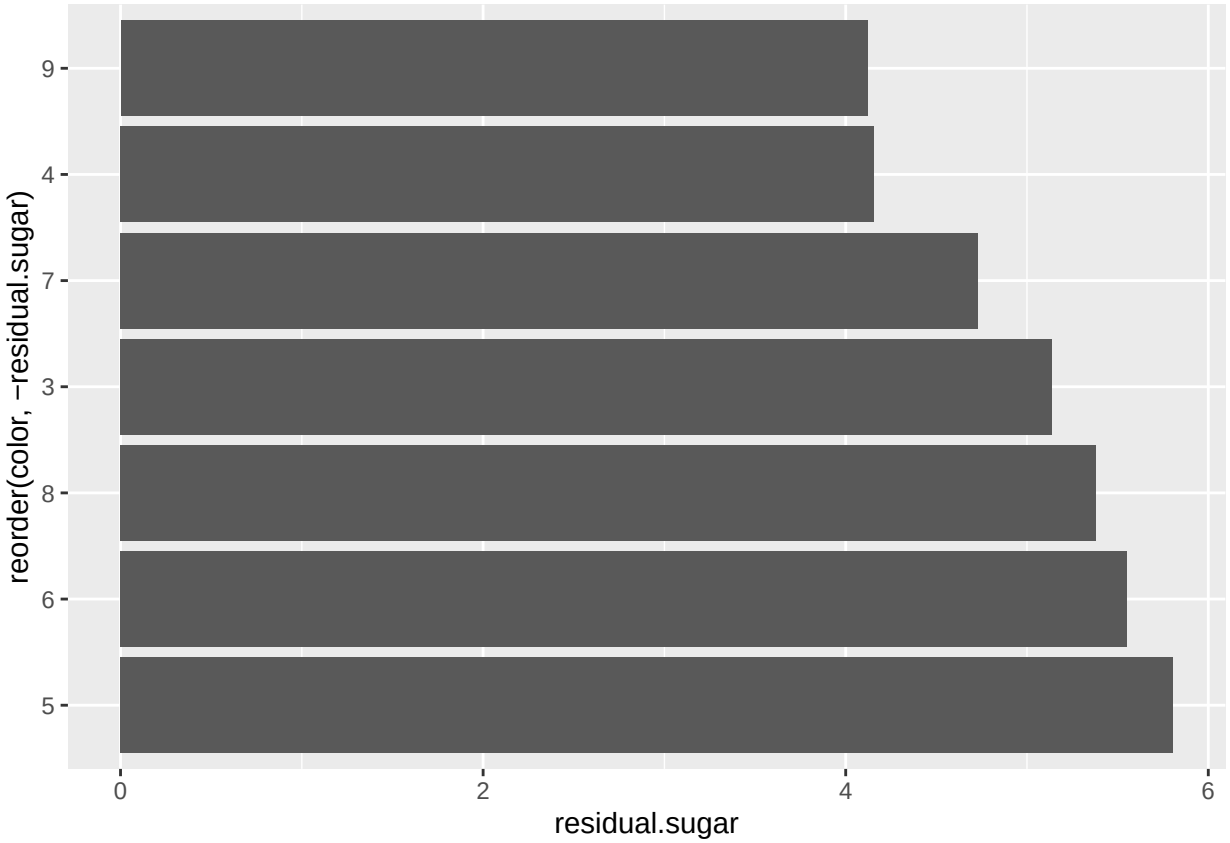

quality
8
6
4
0.99
1.00
1.01
1.02
1.03
1.04
density
color
red
white


quality
8
6
4
3.0
3.5
4.0
pH
color
red
white


quality
8
6
4
0.5
1.0
1.5
2.0
sulphates
color
red
white

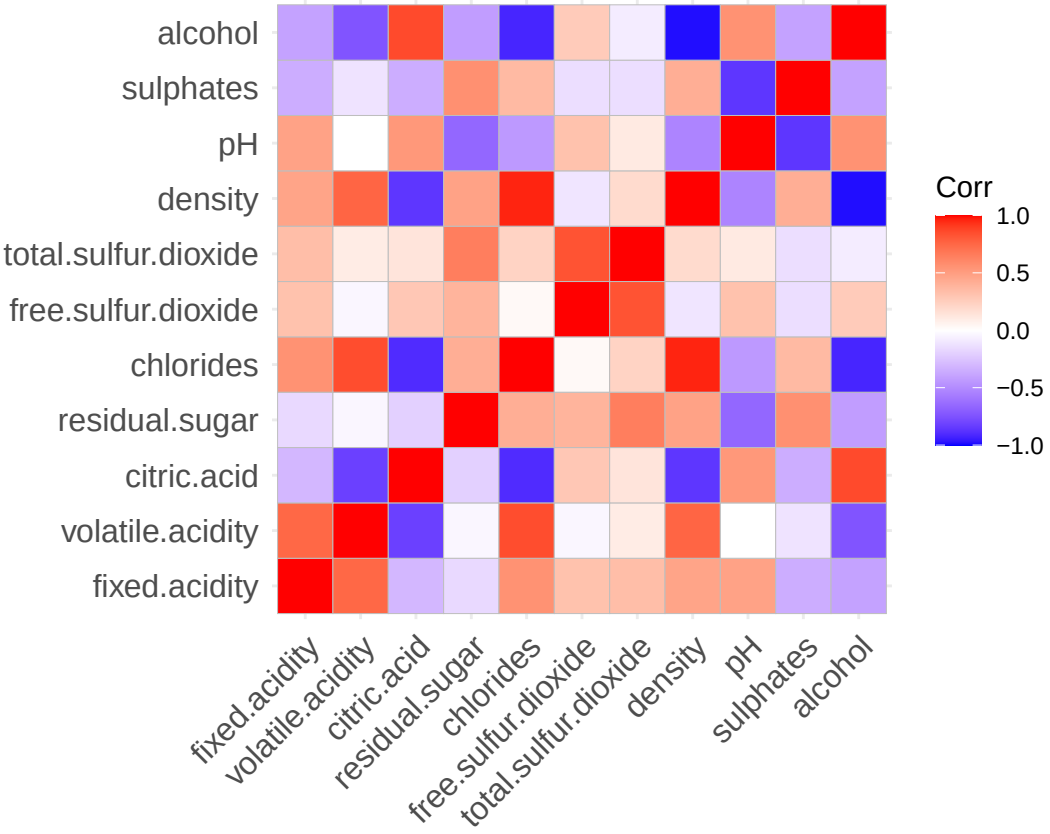

quality
8
6
4
8
10
12
14
alcohol
color
red
white

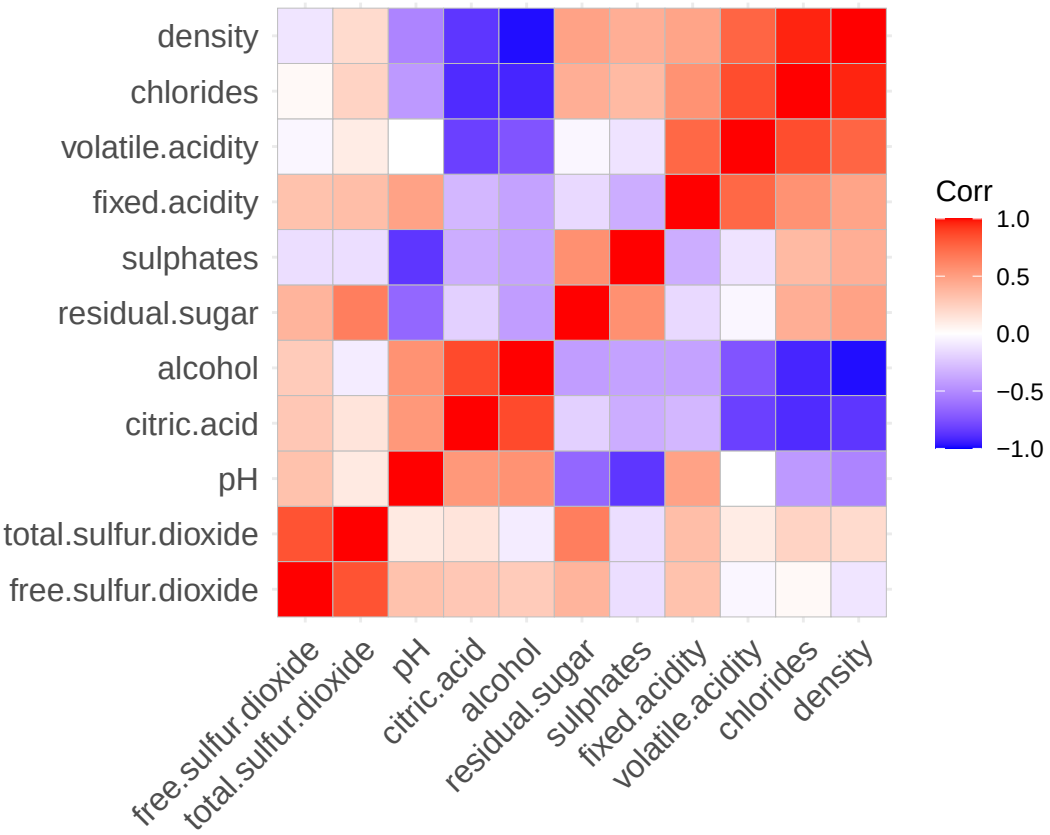

reorder(color, −alcohol)
5
4
3
6
7
8
9
0.0
2.5
5.0
7.5
10.0
12.5
alcohol

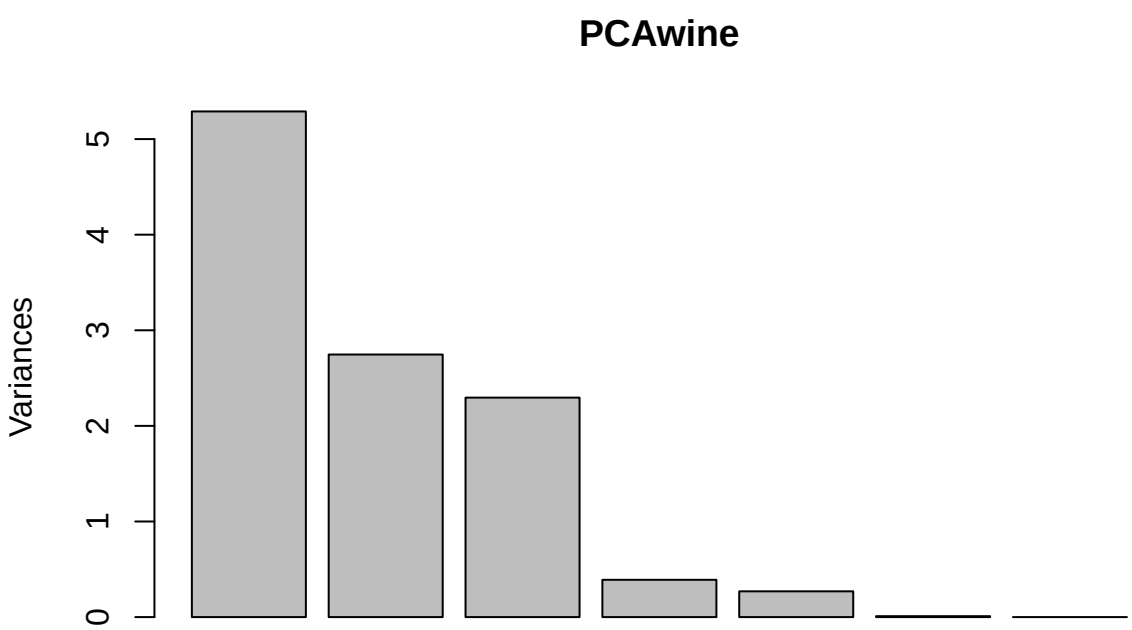



```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity             1.0000000      0.753660964  -0.3134524    -0.15871408
## volatile.acidity          0.7536610      1.000000000  -0.8249347    -0.03967801
## citric.acid              -0.3134524     -0.824934727   1.0000000    -0.19860492
## residual.sugar           -0.1587141     -0.039678008  -0.1986049     1.00000000
## chlorides                 0.5587635      0.851335030  -0.8969220     0.41674150
## free.sulfur.dioxide       0.3201482     -0.038456840   0.2922908     0.38880588
## total.sulfur.dioxide      0.3352811      0.097497567   0.1367631     0.65493123
## density                   0.4702587      0.762888881  -0.8551123     0.48125596
## pH                        0.4783328     -0.002380562   0.5338545    -0.65828568
## sulphates                -0.3477764     -0.124121058  -0.3532299     0.56808718
## alcohol                  -0.3968891     -0.741513840   0.8638209    -0.41656504
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity          0.55876352          0.32014817           0.33528110
## volatile.acidity       0.85133503         -0.03845684           0.09749757
## citric.acid           -0.89692196          0.29229080           0.13676312
## residual.sugar         0.41674150          0.38880588           0.65493123
## chlorides              1.00000000          0.03072666           0.22885312
## free.sulfur.dioxide    0.03072666          1.00000000           0.83159945
## total.sulfur.dioxide   0.22885312          0.83159945           1.00000000
## density                0.96376636         -0.11366679           0.19474383
## pH                    -0.43527453          0.31724955           0.11414665
## sulphates              0.35804899         -0.13828072          -0.13903326
## alcohol               -0.91920515          0.27327155          -0.07867315
##                          density           pH  sulphates     alcohol
## fixed.acidity          0.4702587  0.478332849 -0.3477764 -0.39688913
```

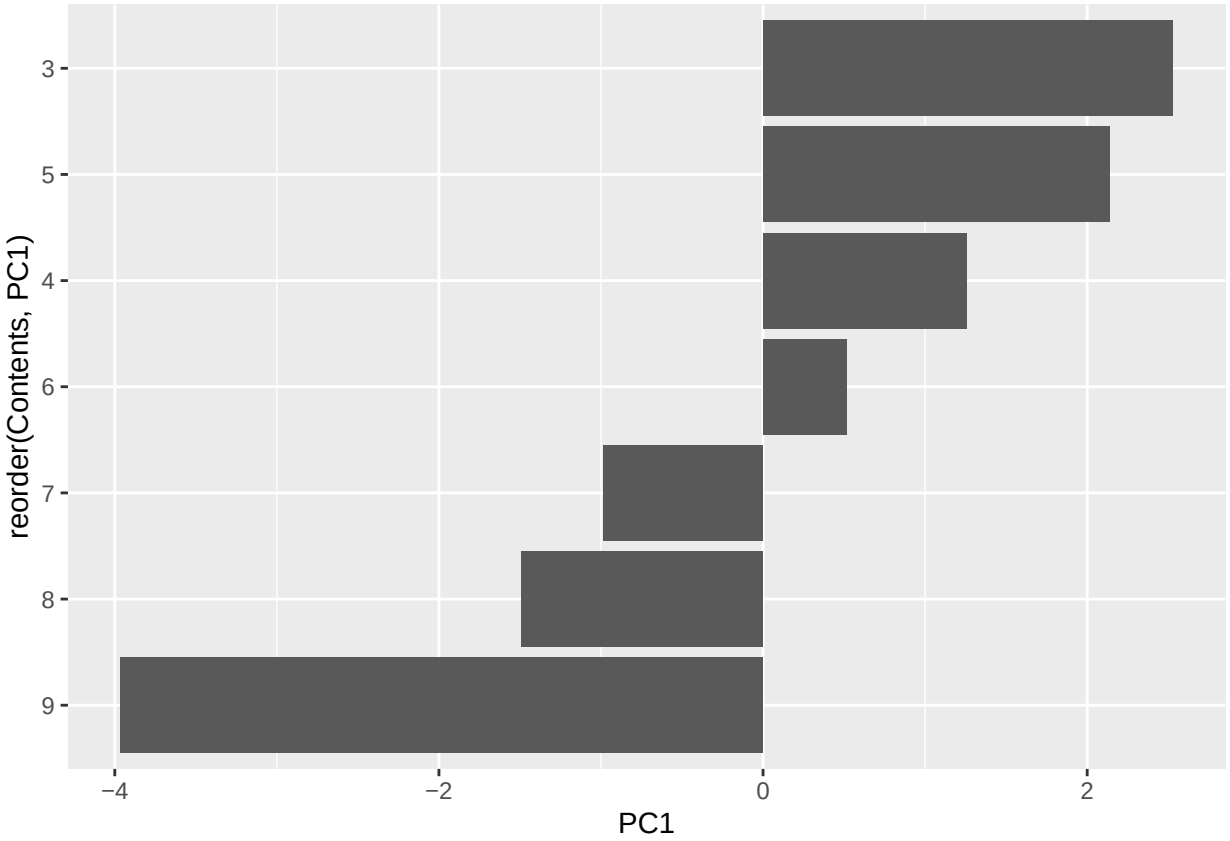
```
## volatile.acidity      0.7628889 -0.002380562 -0.1241211 -0.74151384
## citric.acid          -0.8551123  0.533854467 -0.3532299  0.86382094
## residual.sugar        0.4812560 -0.658285683  0.5680872 -0.41656504
## chlorides             0.9637664 -0.435274532  0.3580490 -0.91920515
## free.sulfur.dioxide  -0.1136668  0.317249554 -0.1382807  0.27327155
## total.sulfur.dioxide  0.1947438  0.114146654 -0.1390333 -0.07867315
## density               1.0000000 -0.527214901  0.4154190 -0.98499040
## pH                   -0.5272149  1.000000000 -0.8583849  0.55616395
## sulphates             0.4154190 -0.858384870  1.0000000 -0.39564324
## alcohol              -0.9849904  0.556163953 -0.3956432  1.00000000
```

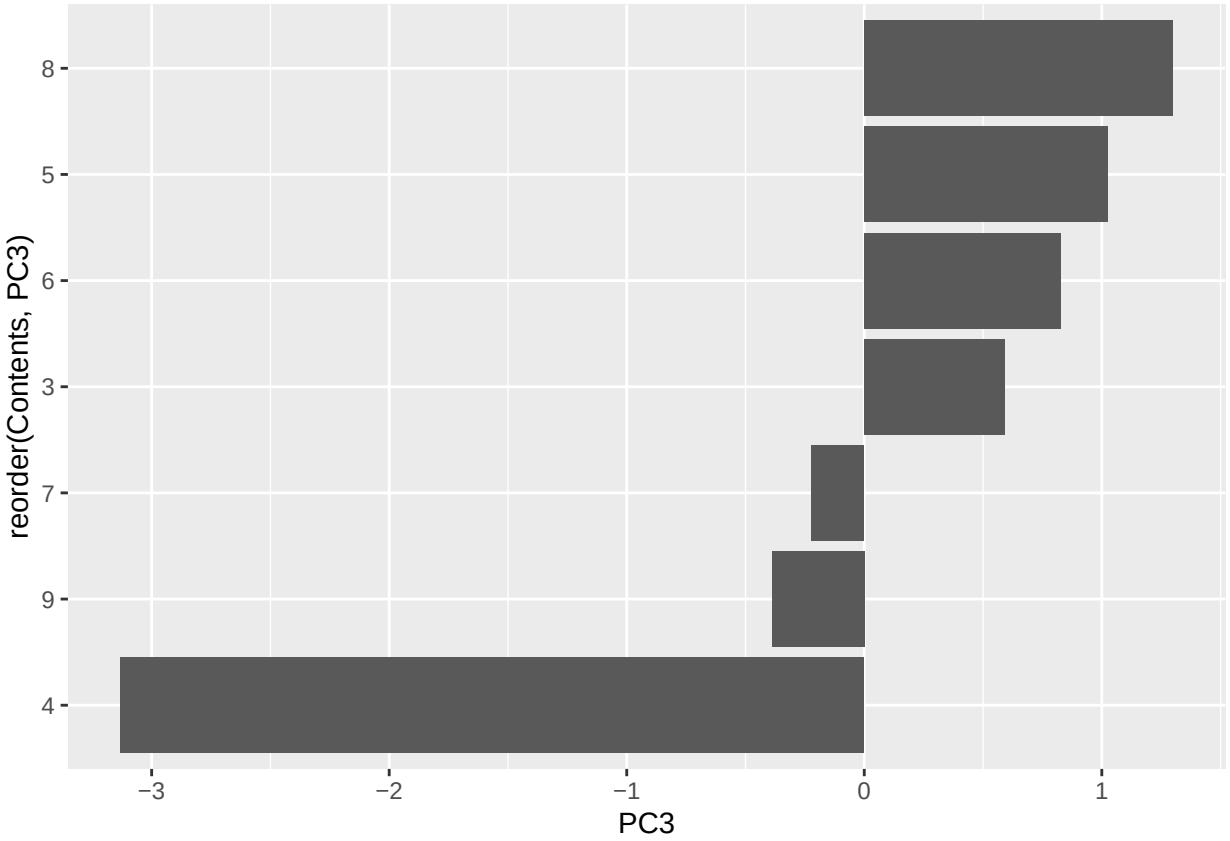



density
chlorides
volatile.acidity
fixed.acidity
sulphates
residual.sugar
alcohol
citric.acid
pH
total.sulfur.dioxide
free.sulfur.dioxide
free.sulfur.dioxide
total.sulfur.dioxide
pH
citric.acid
alcohol
residual.sugar
sulphates
fixed.acidity
volatile.acidity
chlorides
density
Corr
1.0
0.5
0.0
−0.5
−1.0

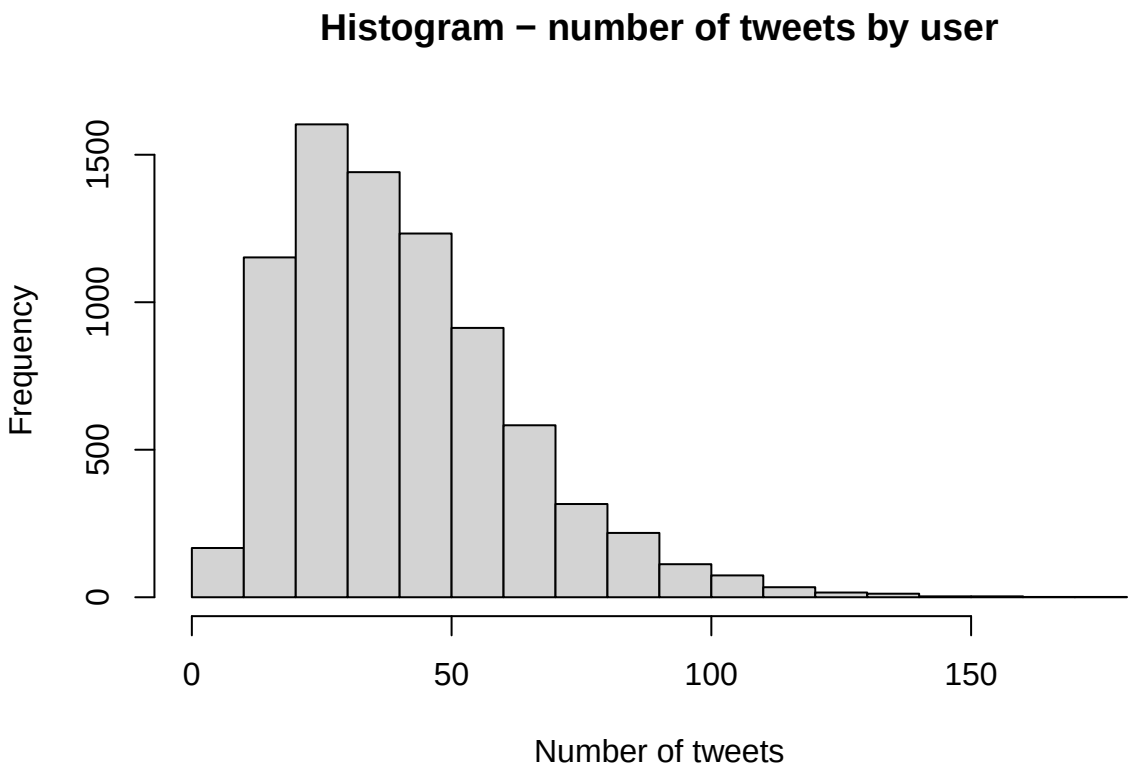



```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6       PC7
## Standard deviation     2.2998 1.6572 1.5151 0.62456 0.51901 0.09746 1.005e-14
## Proportion of Variance 0.4808 0.2497 0.2087 0.03546 0.02449 0.00086 0.000e+00
## Cumulative Proportion  0.4808 0.7305 0.9392 0.97465 0.99914 1.00000 1.000e+00
```


reorder(Contents, PC1)
3
5
4
6
7
8
9
−4
−2
0
2
PC1



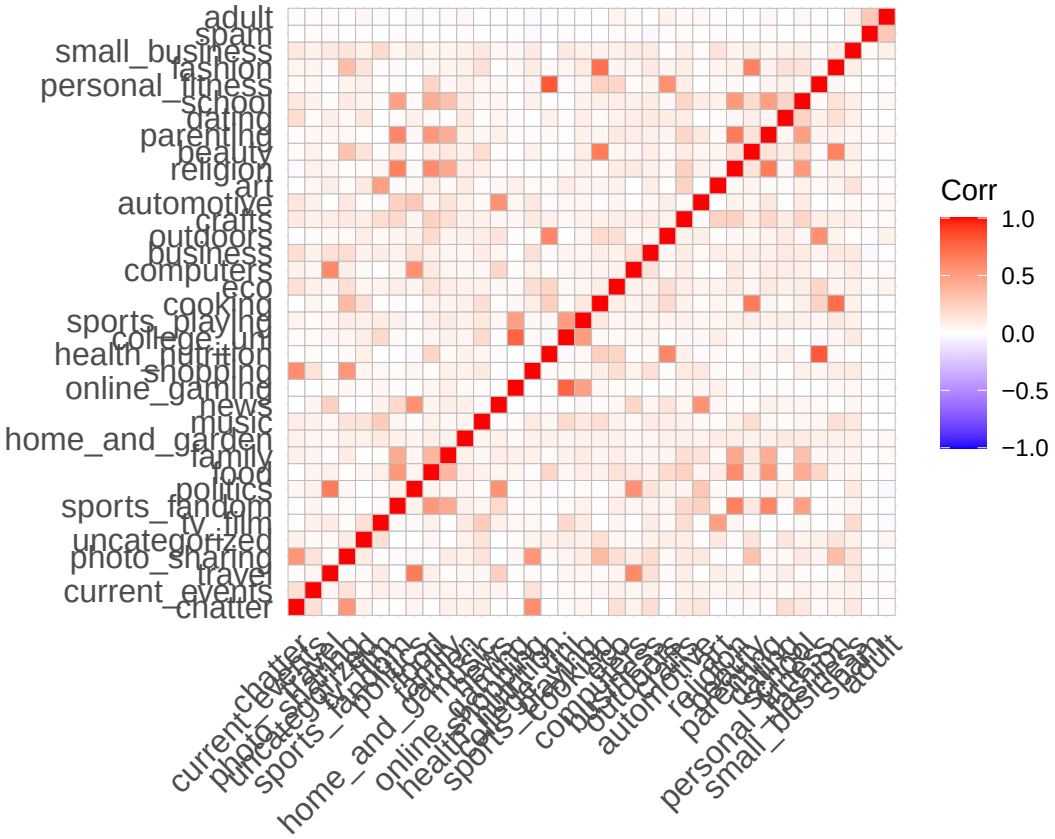

```
##
## Call:
## lm(formula = quality ~ PC1 + PC2 + PC3, data = wine_data)
##
## Residuals:
##       1       2       3       4       5       6       7
## -0.09449  0.40959  0.06442 -0.39356 -0.91554  1.04044 -0.11086
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.71429    0.32869  17.385 0.000415 ***
## PC1         -0.30428    0.15437  -1.971 0.143290
## PC2          0.07674    0.21423   0.358 0.743924
## PC3         -0.09508    0.23432  -0.406 0.712111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8696 on 3 degrees of freedom
## Multiple R-squared:  0.5821, Adjusted R-squared:  0.1641
## F-statistic: 1.393 on 3 and 3 DF,  p-value: 0.396
```

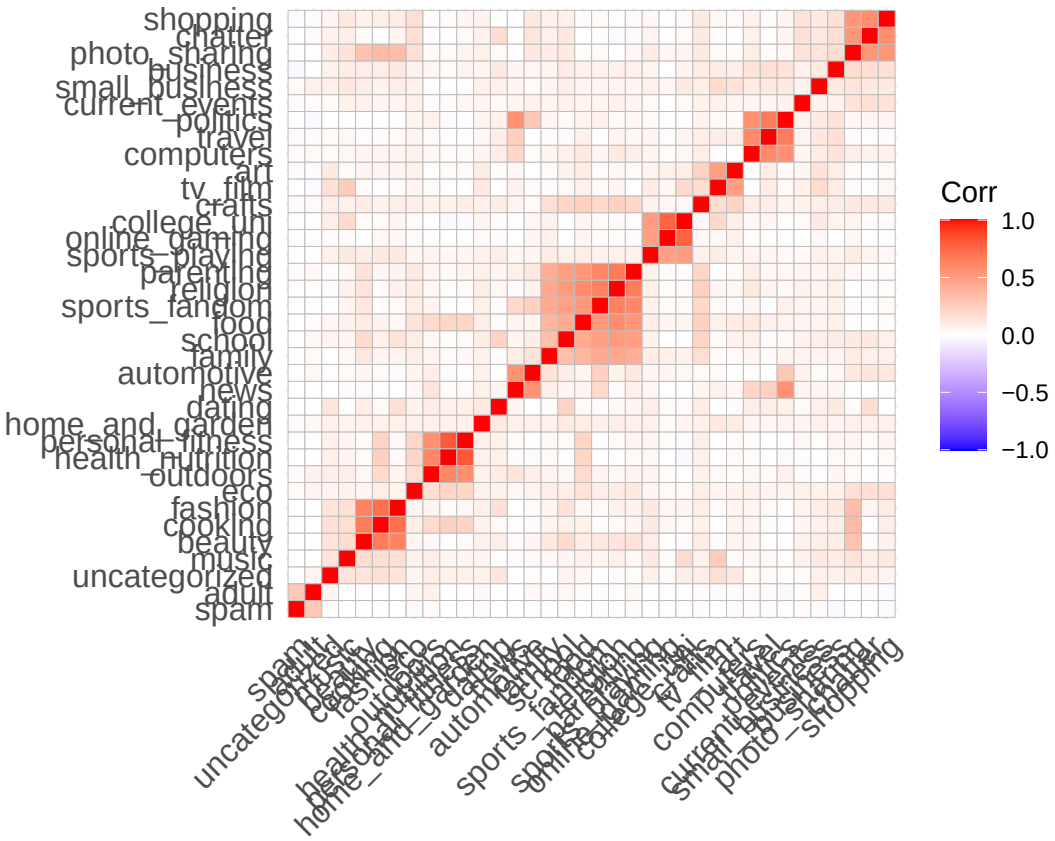



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   4.399   6.000  26.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.00000 0.00647 0.00000 2.00000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4033  0.0000 26.0000
```


adult
spam
small_business
fashion
personal_fitness
school
dating
parenting
beauty
religion
art
automotive
crafts
outdoors
business
computers
eco
cooking
sports_playing
college_uni
health_nutrition
shopping
online_gaming
news
music
home_and_garden
family
food
politics
sports_fandom
tv_film
uncategorized
photo_sharing
travel
current_events
chatter
Corr
1.0
0.5
0.0
−0.5
−1.0



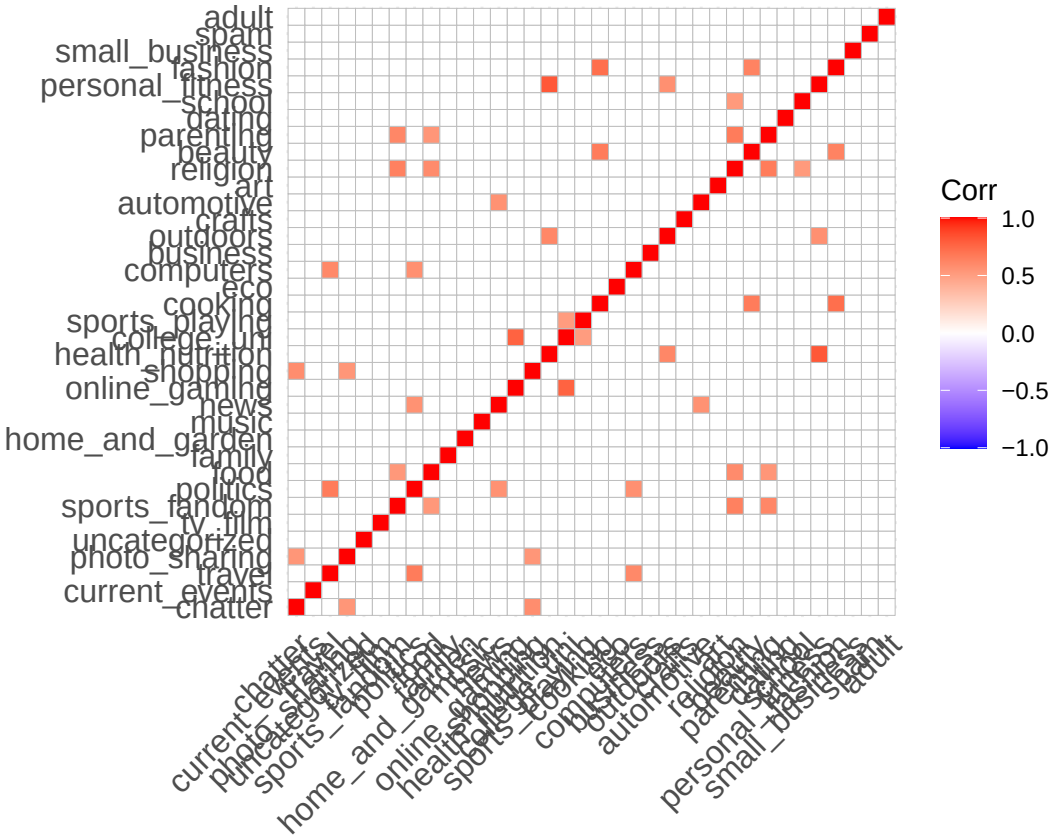
	chatter	current_events	travel	photo_sharing	uncategorized
chatter	1.00	0	0.00	0.54	0
current_events	0.00	1	0.00	0.00	0
travel	0.00	0	1.00	0.00	0
photo_sharing	0.54	0	0.00	1.00	0
uncategorized	0.00	0	0.00	0.00	1
tv_film	0.00	0	0.00	0.00	0
sports_fandom	0.00	0	0.00	0.00	0
politics	0.00	0	0.66	0.00	0
food	0.00	0	0.00	0.00	0
family	0.00	0	0.00	0.00	0
home_and_garden	0.00	0	0.00	0.00	0
music	0.00	0	0.00	0.00	0
news	0.00	0	0.00	0.00	0
online_gaming	0.00	0	0.00	0.00	0
shopping	0.58	0	0.00	0.54	0
health_nutrition	0.00	0	0.00	0.00	0
college_uni	0.00	0	0.00	0.00	0
sports_playing	0.00	0	0.00	0.00	0
cooking	0.00	0	0.00	0.00	0
eco	0.00	0	0.00	0.00	0
computers	0.00	0	0.60	0.00	0
business	0.00	0	0.00	0.00	0
outdoors	0.00	0	0.00	0.00	0
crafts	0.00	0	0.00	0.00	0
automotive	0.00	0	0.00	0.00	0

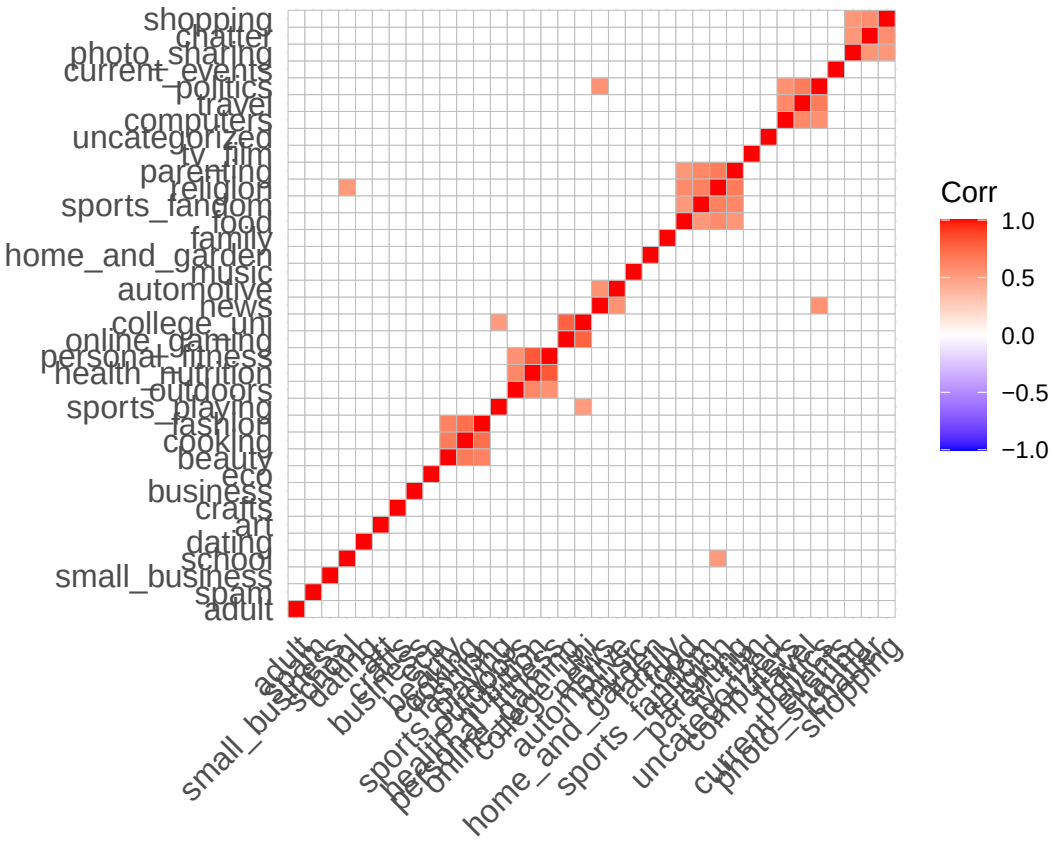
```
## art                0.00             0   0.00          0.00             0
## religion           0.00             0   0.00          0.00             0
## beauty             0.00             0   0.00          0.00             0
## parenting          0.00             0   0.00          0.00             0
## dating             0.00             0   0.00          0.00             0
## school             0.00             0   0.00          0.00             0
## personal_fitness   0.00             0   0.00          0.00             0
## fashion            0.00             0   0.00          0.00             0
## small_business     0.00             0   0.00          0.00             0
## spam               0.00             0   0.00          0.00             0
## adult              0.00             0   0.00          0.00             0
##                  tv_film sports_fandom politics food family home_and_garden
## chatter                0          0.00     0.00 0.00      0               0
## current_events         0          0.00     0.00 0.00      0               0
## travel                 0          0.00     0.66 0.00      0               0
## photo_sharing          0          0.00     0.00 0.00      0               0
## uncategorized          0          0.00     0.00 0.00      0               0
## tv_film                1          0.00     0.00 0.00      0               0
## sports_fandom          0          1.00     0.00 0.53      0               0
## politics               0          0.00     1.00 0.00      0               0
## food                   0          0.53     0.00 1.00      0               0
## family                 0          0.00     0.00 0.00      1               0
## home_and_garden        0          0.00     0.00 0.00      0               1
## music                  0          0.00     0.00 0.00      0               0
## news                   0          0.00     0.56 0.00      0               0
## online_gaming          0          0.00     0.00 0.00      0               0
## shopping               0          0.00     0.00 0.00      0               0
## health_nutrition       0          0.00     0.00 0.00      0               0
## college_uni            0          0.00     0.00 0.00      0               0
## sports_playing         0          0.00     0.00 0.00      0               0
## cooking                0          0.00     0.00 0.00      0               0
## eco                    0          0.00     0.00 0.00      0               0
## computers              0          0.00     0.57 0.00      0               0
## business               0          0.00     0.00 0.00      0               0
## outdoors               0          0.00     0.00 0.00      0               0
## crafts                 0          0.00     0.00 0.00      0               0
## automotive             0          0.00     0.00 0.00      0               0
## art                    0          0.00     0.00 0.00      0               0
## religion               0          0.64     0.00 0.59      0               0
## beauty                 0          0.00     0.00 0.00      0               0
## parenting              0          0.61     0.00 0.54      0               0
## dating                 0          0.00     0.00 0.00      0               0
## school                 0          0.00     0.00 0.00      0               0
## personal_fitness       0          0.00     0.00 0.00      0               0
## fashion                0          0.00     0.00 0.00      0               0
## small_business         0          0.00     0.00 0.00      0               0
## spam                   0          0.00     0.00 0.00      0               0
## adult                  0          0.00     0.00 0.00      0               0
##                  music news online_gaming shopping health_nutrition college_uni
## chatter              0 0.00          0.00     0.58             0.00        0.00
## current_events       0 0.00          0.00     0.00             0.00        0.00
## travel               0 0.00          0.00     0.00             0.00        0.00
## photo_sharing        0 0.00          0.00     0.54             0.00        0.00
## uncategorized        0 0.00          0.00     0.00             0.00        0.00
```

```
## tv_film              0 0.00          0.00     0.00             0.00         0.00
## sports_fandom        0 0.00          0.00     0.00             0.00         0.00
## politics             0 0.56          0.00     0.00             0.00         0.00
## food                 0 0.00          0.00     0.00             0.00         0.00
## family               0 0.00          0.00     0.00             0.00         0.00
## home_and_garden      0 0.00          0.00     0.00             0.00         0.00
## music                1 0.00          0.00     0.00             0.00         0.00
## news                 0 1.00          0.00     0.00             0.00         0.00
## online_gaming        0 0.00          1.00     0.00             0.00         0.77
## shopping             0 0.00          0.00     1.00             0.00         0.00
## health_nutrition     0 0.00          0.00     0.00             1.00         0.00
## college_uni          0 0.00          0.77     0.00             0.00         1.00
## sports_playing       0 0.00          0.00     0.00             0.00         0.51
## cooking              0 0.00          0.00     0.00             0.00         0.00
## eco                  0 0.00          0.00     0.00             0.00         0.00
## computers            0 0.00          0.00     0.00             0.00         0.00
## business             0 0.00          0.00     0.00             0.00         0.00
## outdoors             0 0.00          0.00     0.00             0.61         0.00
## crafts               0 0.00          0.00     0.00             0.00         0.00
## automotive           0 0.56          0.00     0.00             0.00         0.00
## art                  0 0.00          0.00     0.00             0.00         0.00
## religion             0 0.00          0.00     0.00             0.00         0.00
## beauty               0 0.00          0.00     0.00             0.00         0.00
## parenting            0 0.00          0.00     0.00             0.00         0.00
## dating               0 0.00          0.00     0.00             0.00         0.00
## school               0 0.00          0.00     0.00             0.00         0.00
## personal_fitness     0 0.00          0.00     0.00             0.81         0.00
## fashion              0 0.00          0.00     0.00             0.00         0.00
## small_business       0 0.00          0.00     0.00             0.00         0.00
## spam                 0 0.00          0.00     0.00             0.00         0.00
## adult                0 0.00          0.00     0.00             0.00         0.00
##                  sports_playing cooking eco computers business outdoors crafts
## chatter                    0.00    0.00   0      0.00        0     0.00      0
## current_events             0.00    0.00   0      0.00        0     0.00      0
## travel                     0.00    0.00   0      0.60        0     0.00      0
## photo_sharing              0.00    0.00   0      0.00        0     0.00      0
## uncategorized              0.00    0.00   0      0.00        0     0.00      0
## tv_film                    0.00    0.00   0      0.00        0     0.00      0
## sports_fandom              0.00    0.00   0      0.00        0     0.00      0
## politics                   0.00    0.00   0      0.57        0     0.00      0
## food                       0.00    0.00   0      0.00        0     0.00      0
## family                     0.00    0.00   0      0.00        0     0.00      0
## home_and_garden            0.00    0.00   0      0.00        0     0.00      0
## music                      0.00    0.00   0      0.00        0     0.00      0
## news                       0.00    0.00   0      0.00        0     0.00      0
## online_gaming              0.00    0.00   0      0.00        0     0.00      0
## shopping                   0.00    0.00   0      0.00        0     0.00      0
## health_nutrition           0.00    0.00   0      0.00        0     0.61      0
## college_uni                0.51    0.00   0      0.00        0     0.00      0
## sports_playing             1.00    0.00   0      0.00        0     0.00      0
## cooking                    0.00    1.00   0      0.00        0     0.00      0
## eco                        0.00    0.00   1      0.00        0     0.00      0
## computers                  0.00    0.00   0      1.00        0     0.00      0
## business                   0.00    0.00   0      0.00        1     0.00      0
```

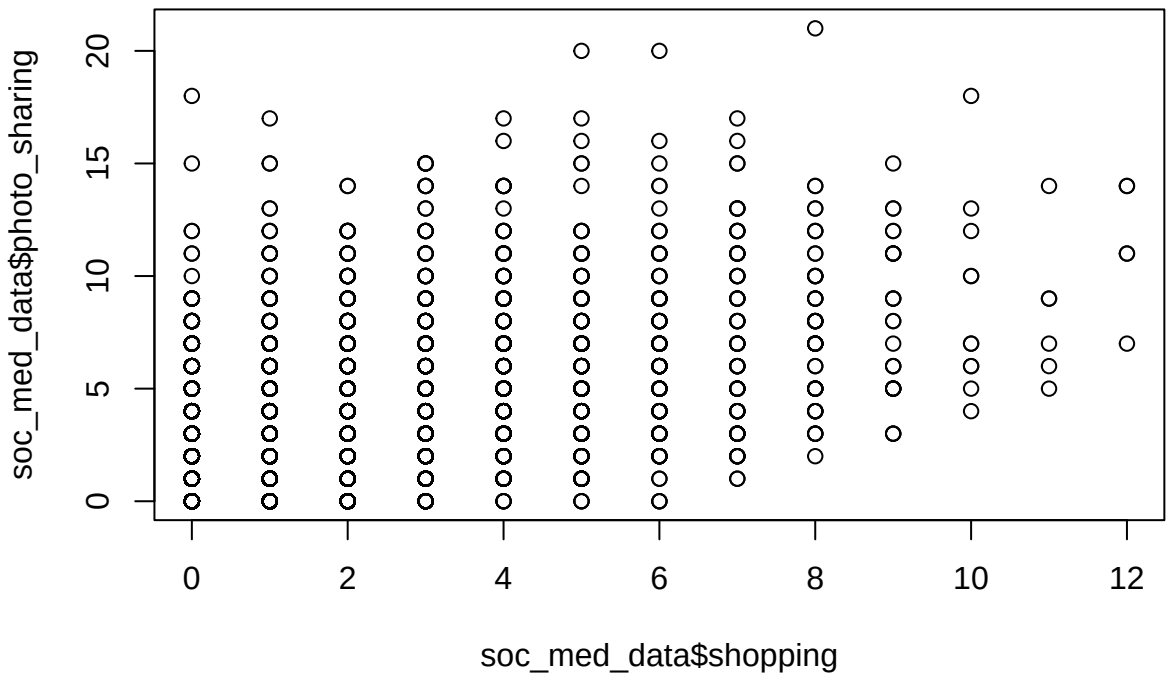
```
## outdoors                  0.00    0.00   0      0.00        0     1.00      0
## crafts                    0.00    0.00   0      0.00        0     0.00      1
## automotive                0.00    0.00   0      0.00        0     0.00      0
## art                       0.00    0.00   0      0.00        0     0.00      0
## religion                  0.00    0.00   0      0.00        0     0.00      0
## beauty                    0.00    0.66   0      0.00        0     0.00      0
## parenting                 0.00    0.00   0      0.00        0     0.00      0
## dating                    0.00    0.00   0      0.00        0     0.00      0
## school                    0.00    0.00   0      0.00        0     0.00      0
## personal_fitness          0.00    0.00   0      0.00        0     0.57      0
## fashion                   0.00    0.72   0      0.00        0     0.00      0
## small_business            0.00    0.00   0      0.00        0     0.00      0
## spam                      0.00    0.00   0      0.00        0     0.00      0
## adult                     0.00    0.00   0      0.00        0     0.00      0
##                  automotive art religion beauty parenting dating school
## chatter                0.00   0     0.00   0.00      0.00      0   0.00
## current_events         0.00   0     0.00   0.00      0.00      0   0.00
## travel                 0.00   0     0.00   0.00      0.00      0   0.00
## photo_sharing          0.00   0     0.00   0.00      0.00      0   0.00
## uncategorized          0.00   0     0.00   0.00      0.00      0   0.00
## tv_film                0.00   0     0.00   0.00      0.00      0   0.00
## sports_fandom          0.00   0     0.64   0.00      0.61      0   0.00
## politics               0.00   0     0.00   0.00      0.00      0   0.00
## food                   0.00   0     0.59   0.00      0.54      0   0.00
## family                 0.00   0     0.00   0.00      0.00      0   0.00
## home_and_garden        0.00   0     0.00   0.00      0.00      0   0.00
## music                  0.00   0     0.00   0.00      0.00      0   0.00
## news                   0.56   0     0.00   0.00      0.00      0   0.00
## online_gaming          0.00   0     0.00   0.00      0.00      0   0.00
## shopping               0.00   0     0.00   0.00      0.00      0   0.00
## health_nutrition       0.00   0     0.00   0.00      0.00      0   0.00
## college_uni            0.00   0     0.00   0.00      0.00      0   0.00
## sports_playing         0.00   0     0.00   0.00      0.00      0   0.00
## cooking                0.00   0     0.00   0.66      0.00      0   0.00
## eco                    0.00   0     0.00   0.00      0.00      0   0.00
## computers              0.00   0     0.00   0.00      0.00      0   0.00
## business               0.00   0     0.00   0.00      0.00      0   0.00
## outdoors               0.00   0     0.00   0.00      0.00      0   0.00
## crafts                 0.00   0     0.00   0.00      0.00      0   0.00
## automotive             1.00   0     0.00   0.00      0.00      0   0.00
## art                    0.00   1     0.00   0.00      0.00      0   0.00
## religion               0.00   0     1.00   0.00      0.66      0   0.52
## beauty                 0.00   0     0.00   1.00      0.00      0   0.00
## parenting              0.00   0     0.66   0.00      1.00      0   0.00
## dating                 0.00   0     0.00   0.00      0.00      1   0.00
## school                 0.00   0     0.52   0.00      0.00      0   1.00
## personal_fitness       0.00   0     0.00   0.00      0.00      0   0.00
## fashion                0.00   0     0.00   0.63      0.00      0   0.00
## small_business         0.00   0     0.00   0.00      0.00      0   0.00
## spam                   0.00   0     0.00   0.00      0.00      0   0.00
## adult                  0.00   0     0.00   0.00      0.00      0   0.00
##                  personal_fitness fashion small_business spam adult
## chatter                      0.00    0.00              0    0     0
## current_events               0.00    0.00              0    0     0
```

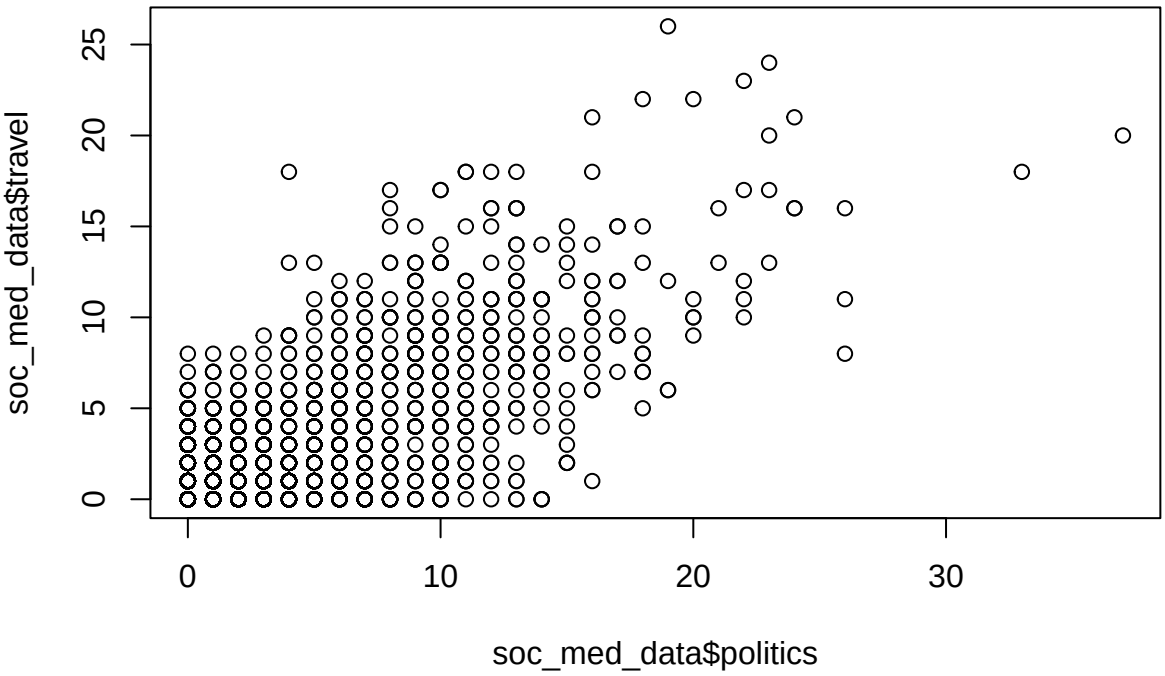
```
## travel                0.00    0.00            0    0     0
## photo_sharing         0.00    0.00            0    0     0
## uncategorized         0.00    0.00            0    0     0
## tv_film               0.00    0.00            0    0     0
## sports_fandom         0.00    0.00            0    0     0
## politics              0.00    0.00            0    0     0
## food                  0.00    0.00            0    0     0
## family                0.00    0.00            0    0     0
## home_and_garden       0.00    0.00            0    0     0
## music                 0.00    0.00            0    0     0
## news                  0.00    0.00            0    0     0
## online_gaming         0.00    0.00            0    0     0
## shopping              0.00    0.00            0    0     0
## health_nutrition      0.81    0.00            0    0     0
## college_uni           0.00    0.00            0    0     0
## sports_playing        0.00    0.00            0    0     0
## cooking               0.00    0.72            0    0     0
## eco                   0.00    0.00            0    0     0
## computers             0.00    0.00            0    0     0
## business              0.00    0.00            0    0     0
## outdoors              0.57    0.00            0    0     0
## crafts                0.00    0.00            0    0     0
## automotive            0.00    0.00            0    0     0
## art                   0.00    0.00            0    0     0
## religion              0.00    0.00            0    0     0
## beauty                0.00    0.63            0    0     0
## parenting             0.00    0.00            0    0     0
## dating                0.00    0.00            0    0     0
## school                0.00    0.00            0    0     0
## personal_fitness      1.00    0.00            0    0     0
## fashion               0.00    1.00            0    0     0
## small_business        0.00    0.00            1    0     0
## spam                  0.00    0.00            0    1     0
## adult                 0.00    0.00            0    0     1
```

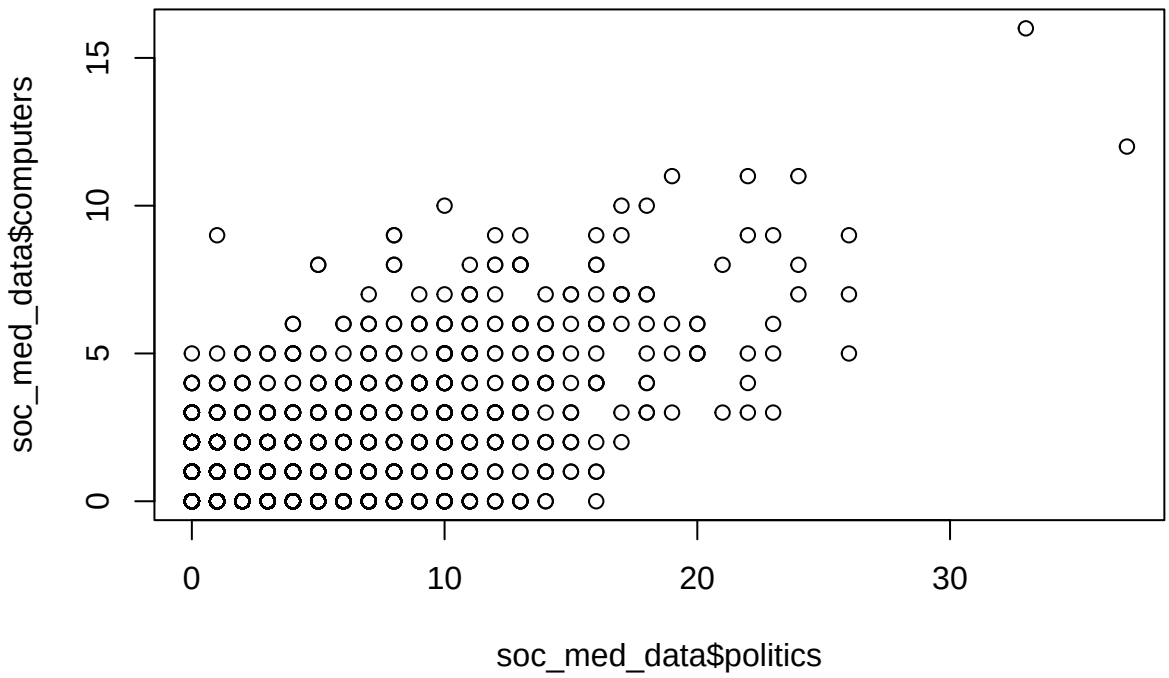

adult
spam
small_business
fashion
personal_fitness
school
dating
parenting
beauty
religion
art
automotive
crafts
outdoors
business
computers
eco
cooking
sports_playing
college_uni
health_nutrition
shopping
online_gaming
news
music
home_and_garden
family
food
politics
sports_fandom
tv_film
uncategorized
photo_sharing
travel
current_events
chatter
Corr
1.0
0.5
0.0
−0.5
−1.0

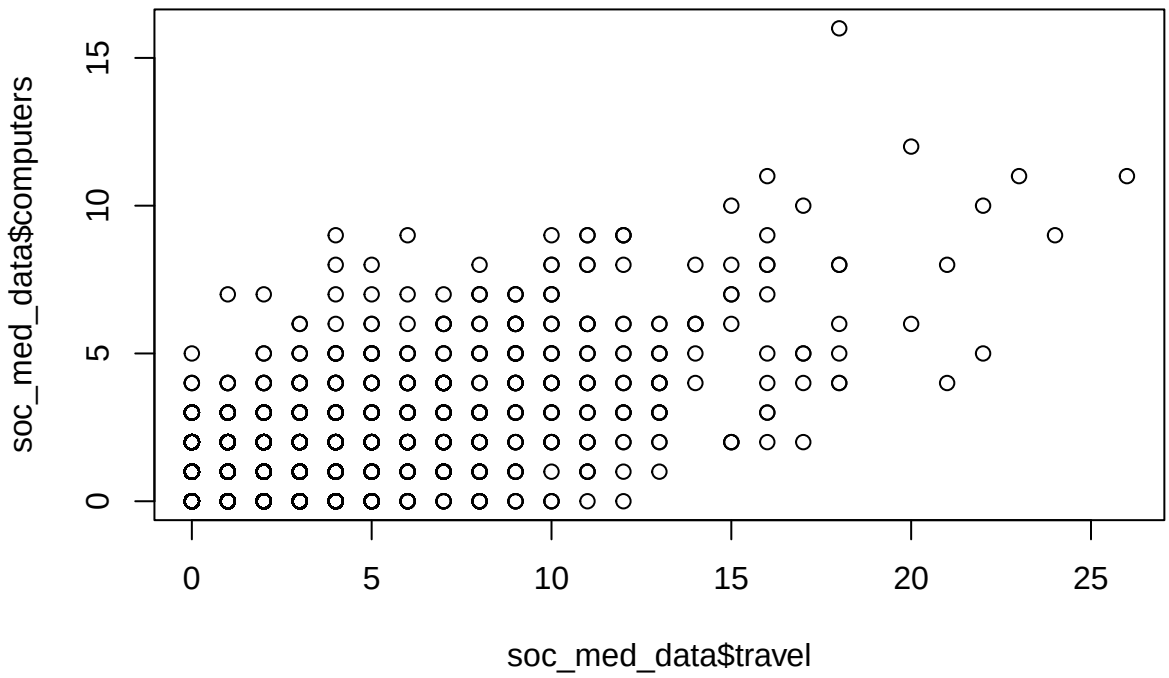


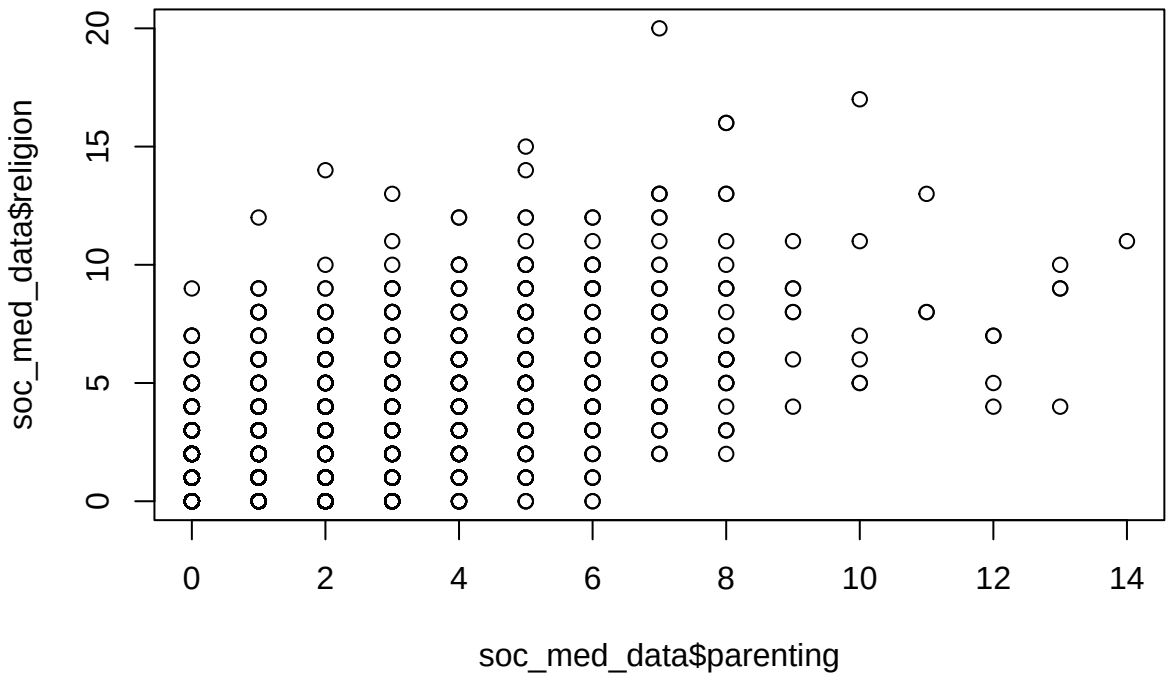

```
## High correlations are found between:
##     (1) shopping, chatter and photo_sharing
##     (2) politics, travel and computers
##     (3) parenting, religion, sports_fandom and food
##     (4) automotive and news
##     (5) college_uni and online_gaming
##     (6) personal_fitness, health_nutrition and outdoors
##     (7) fashion, cooking and beauty
```

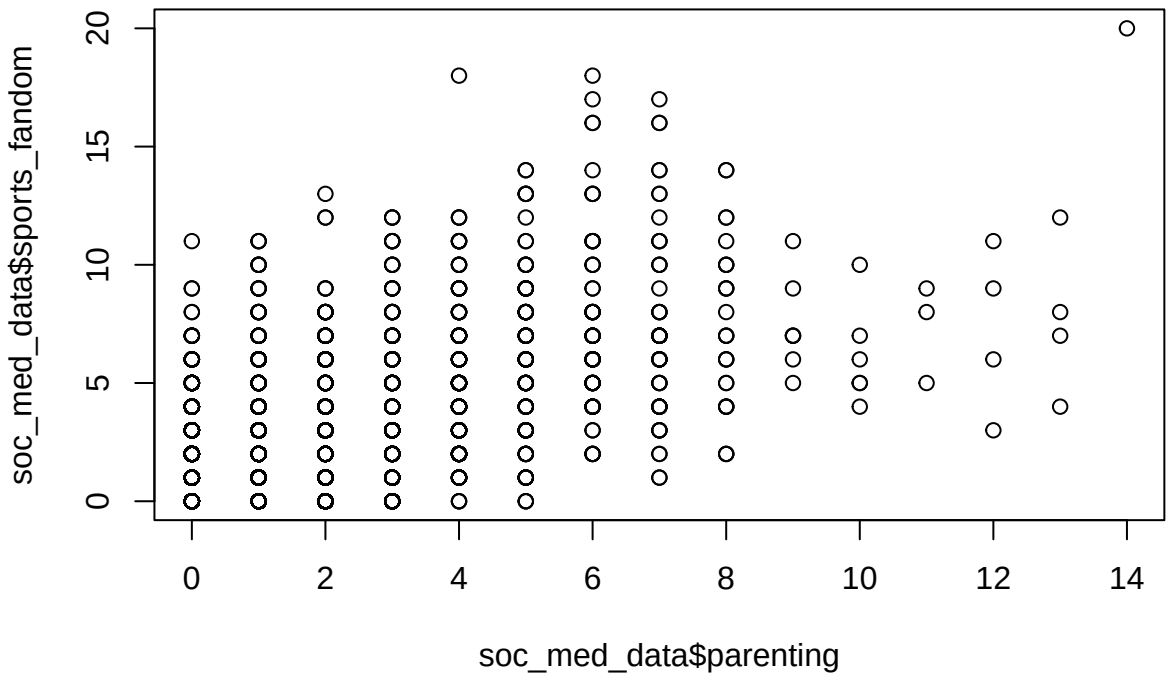

soc_med_data$photo_sharing
0
5
10
15
20
0
2
4
6
8
10
12
soc_med_data$shopping

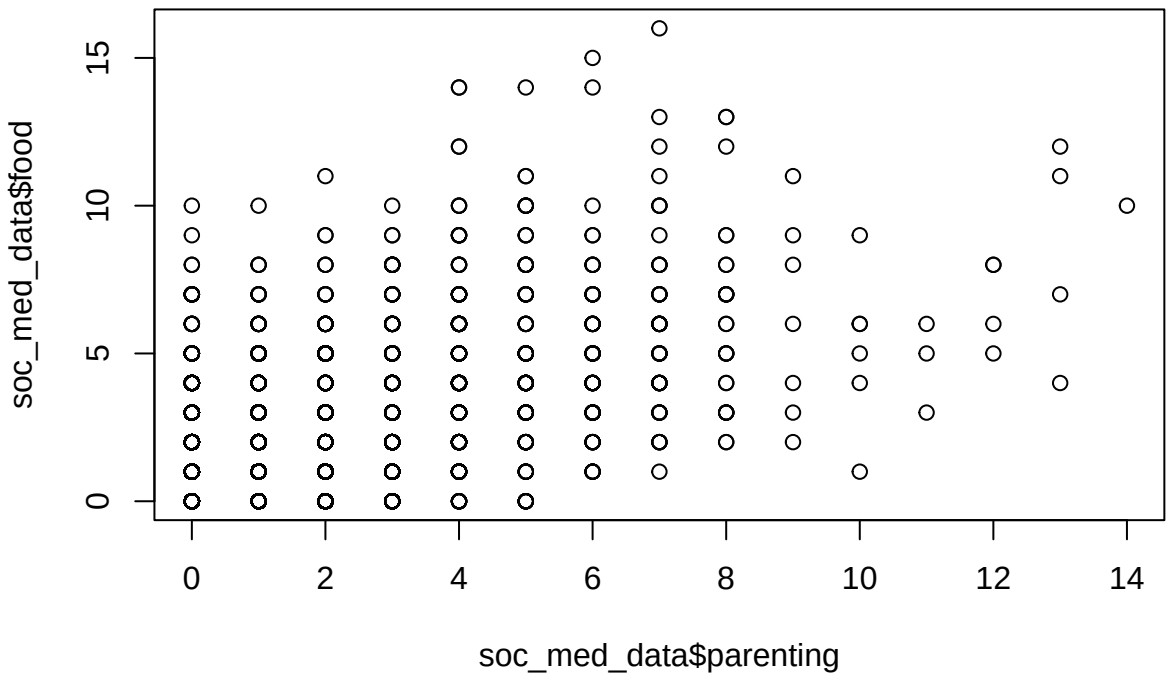

soc_med_data$travel
0
5
10
15
20
25
0
10
20
30
soc_med_data$politics

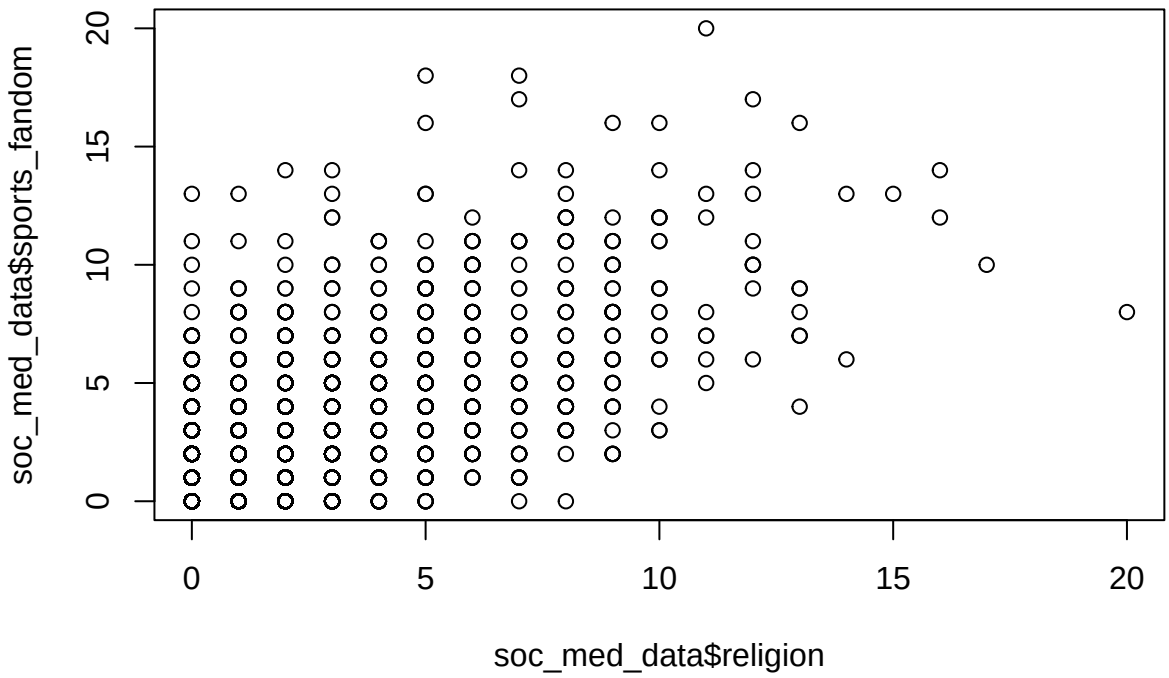

soc_med_data$computers
0
5
10
15
0
10
20
30
soc_med_data$politics

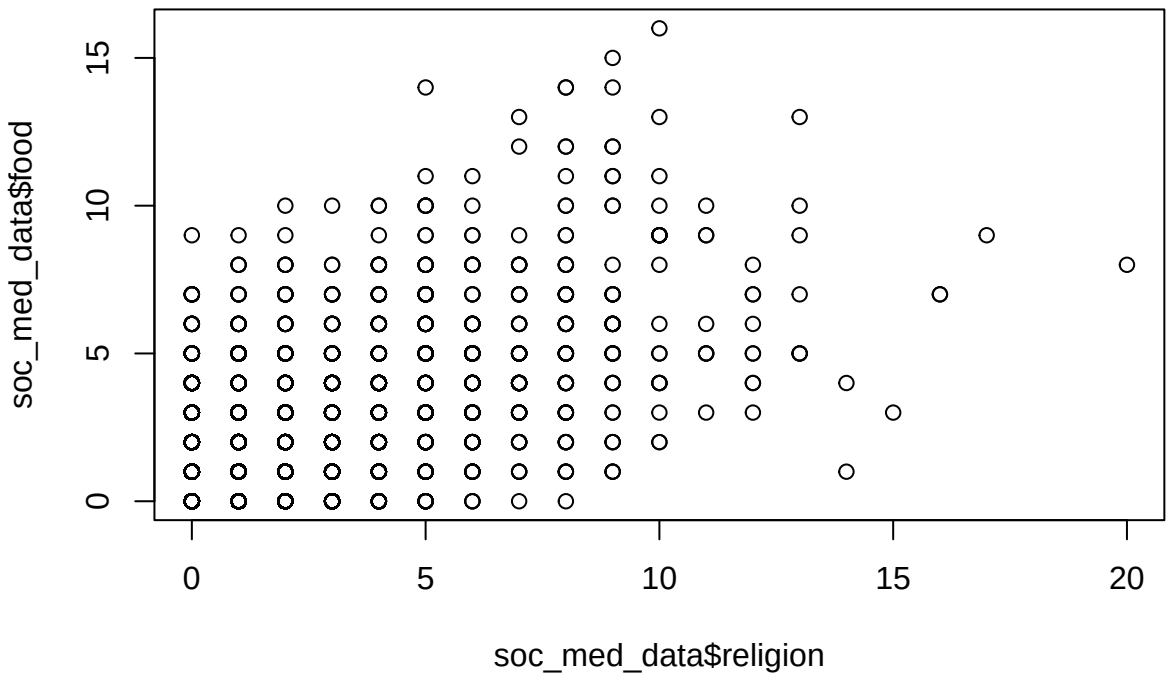

soc_med_data$computers
0
5
10
15
0
5
10
15
20
25
soc_med_data$travel

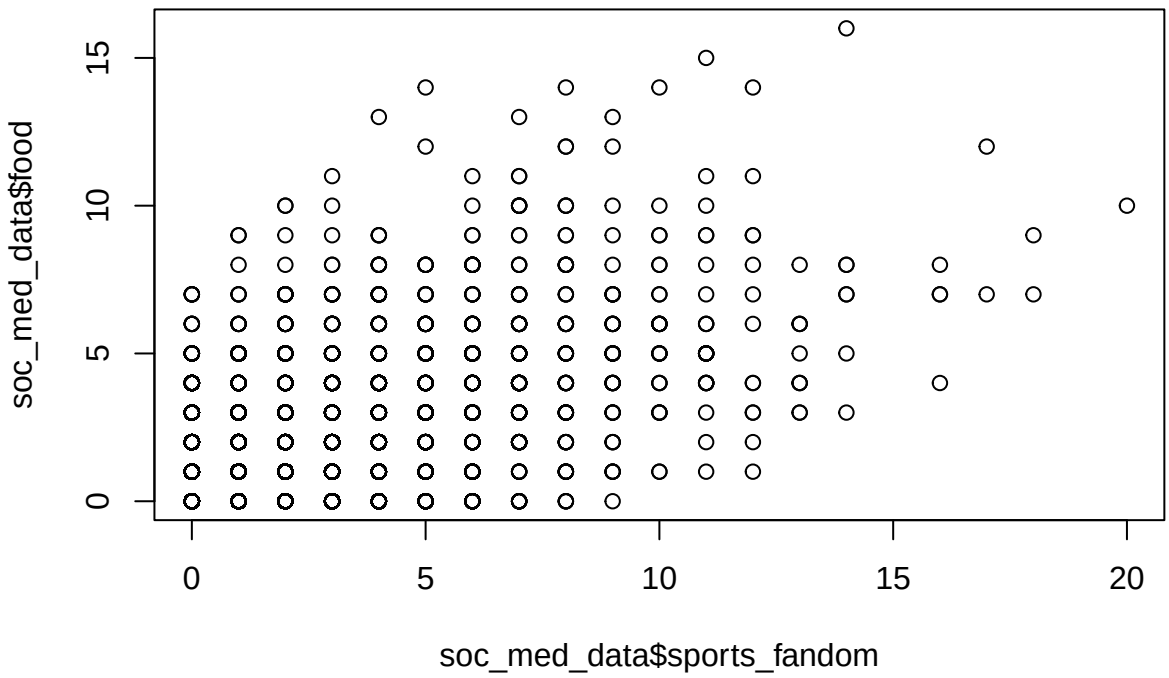

soc_med_data$religion
0
5
10
15
20
0
2
4
6
8
10
12
14
soc_med_data$parenting

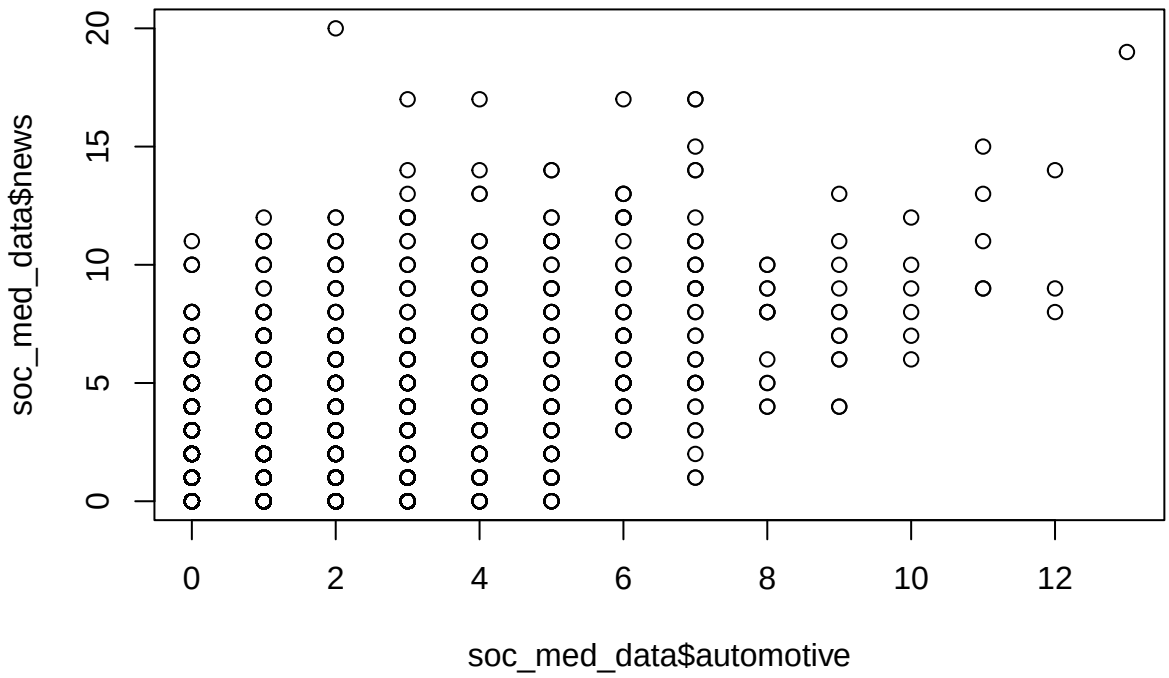

soc_med_data$sports_fandom
0
5
10
15
20
0
2
4
6
8
10
12
14
soc_med_data$parenting

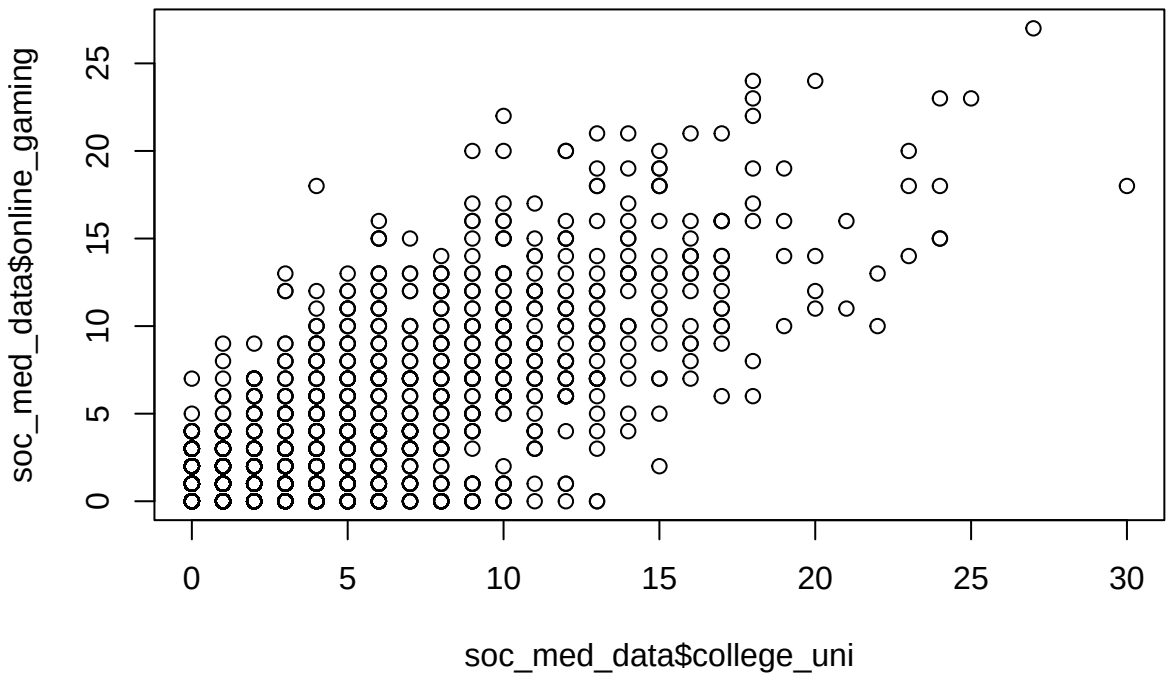

soc_med_data$food
0
5
10
15
0
2
4
6
8
10
12
14
soc_med_data$parenting

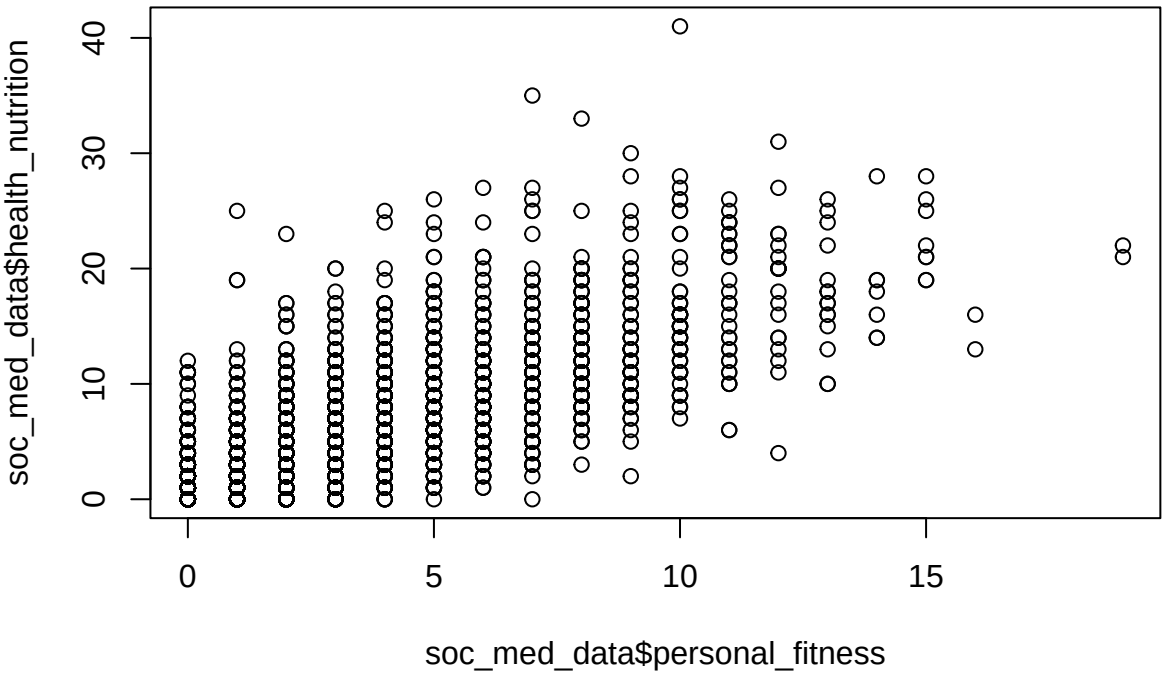

soc_med_data$sports_fandom
0
5
10
15
20
0
5
10
15
20
soc_med_data$religion

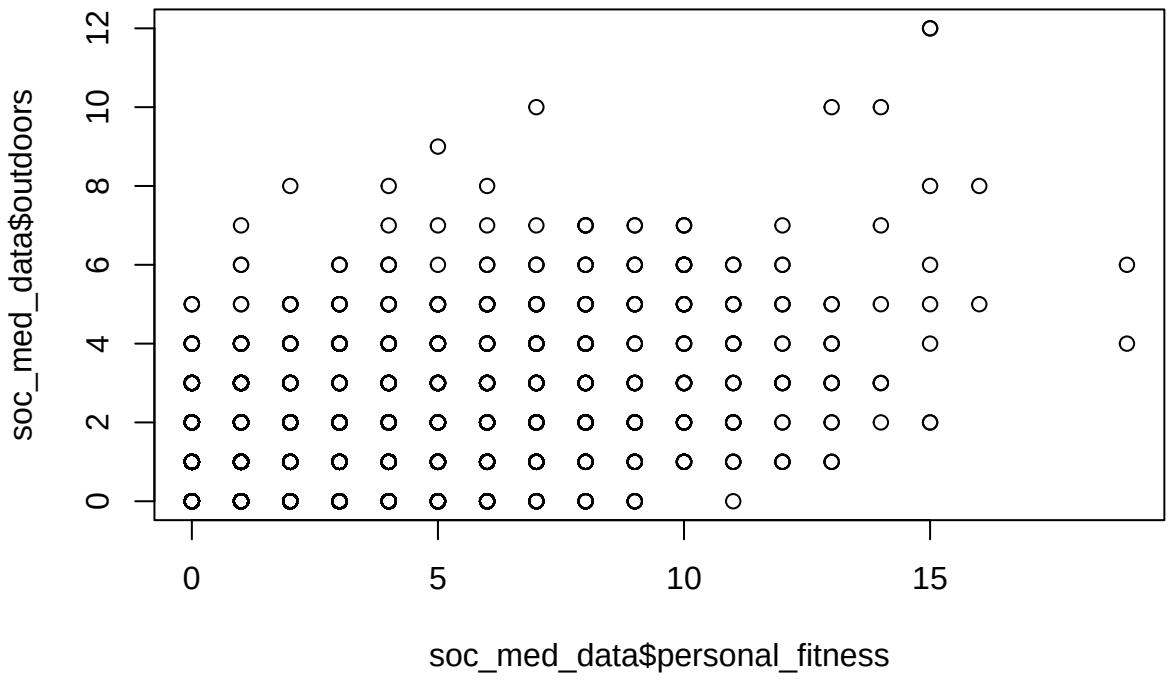

soc_med_data$food
0
5
10
15
0
5
10
15
20
soc_med_data$religion

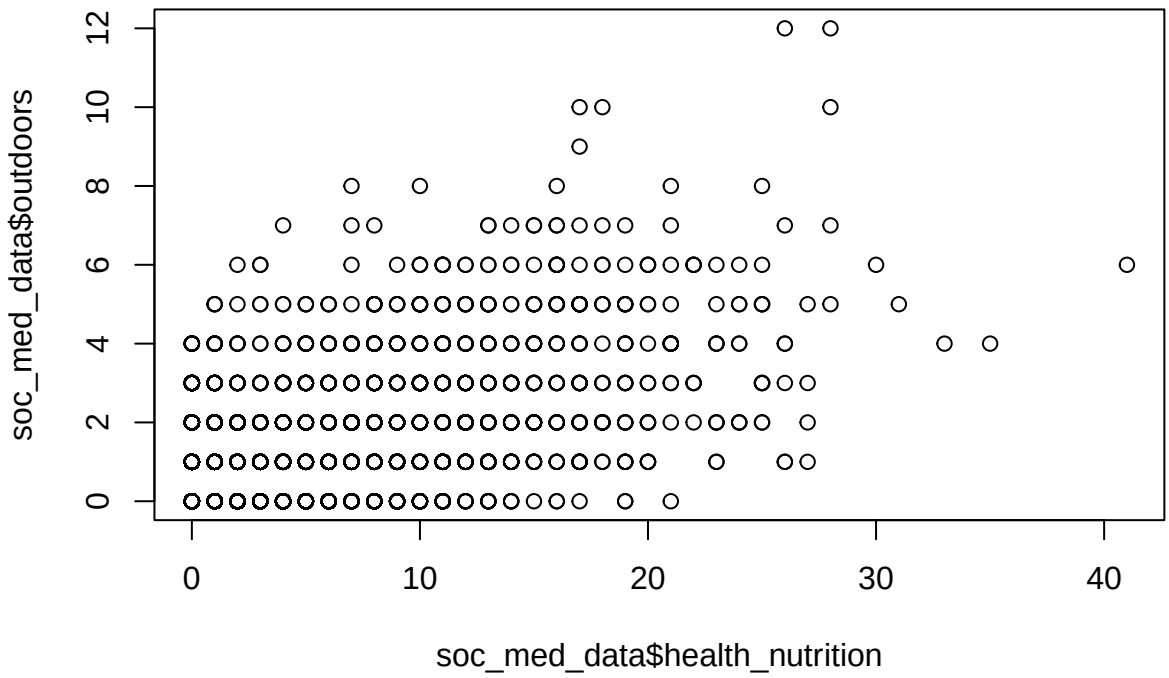

soc_med_data$food
0
5
10
15
0
5
10
15
20
soc_med_data$sports_fandom

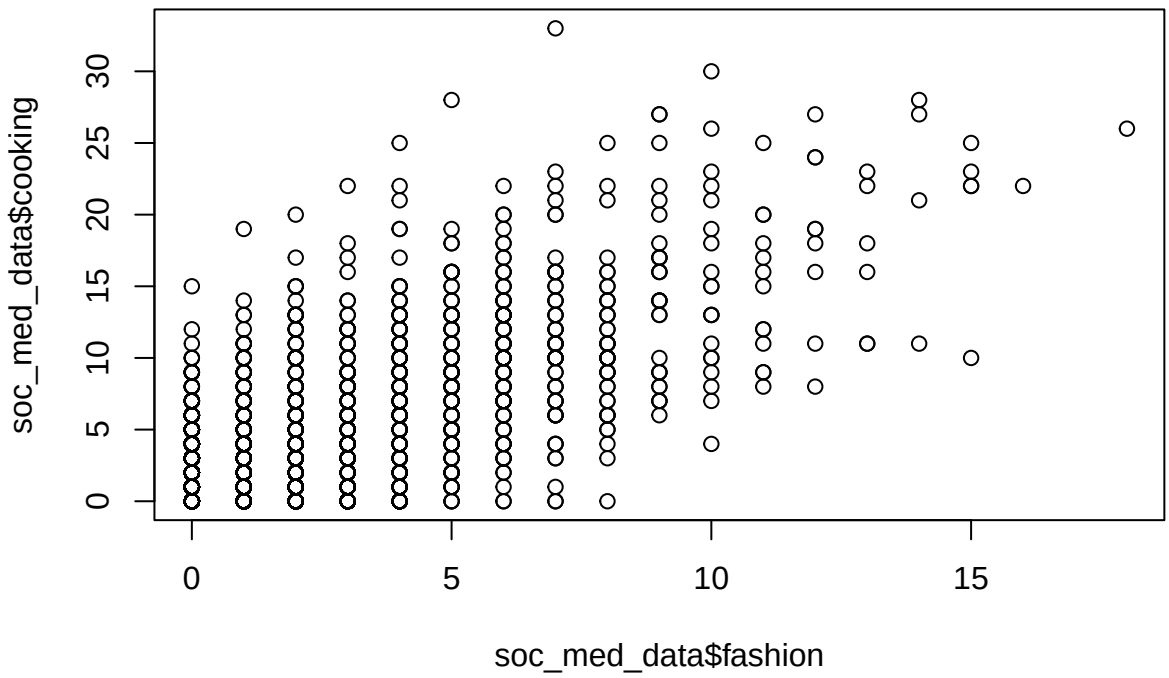

soc_med_data$news
0
5
10
15
20
0
2
4
6
8
10
12
soc_med_data$automotive

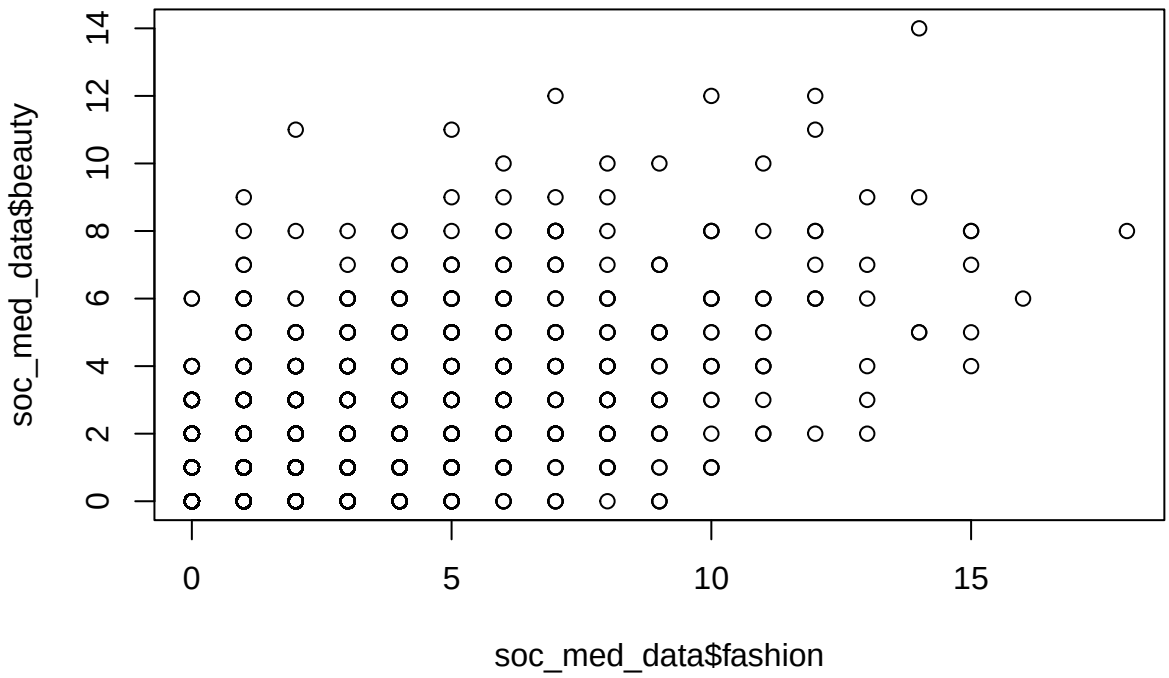

soc_med_data$online_gaming
0
5
10
15
20
25
0
5
10
15
20
25
30
soc_med_data$college_uni

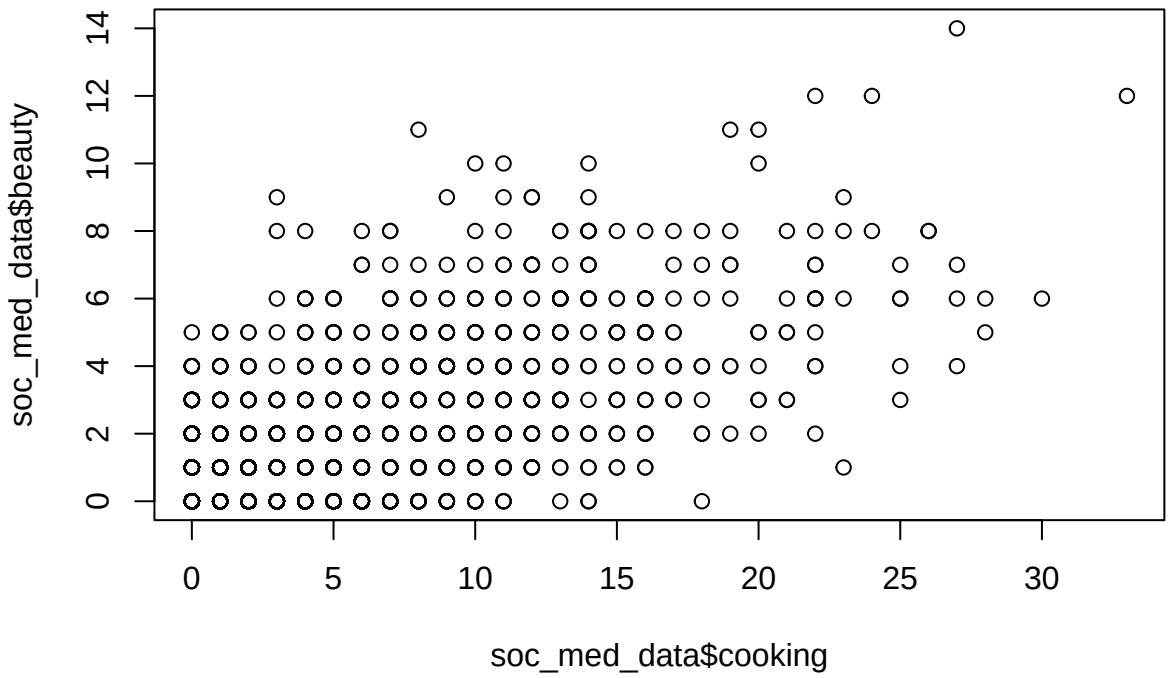

soc_med_data$health_nutrition
0
10
20
30
40
0
5
10
15
soc_med_data$personal_fitness

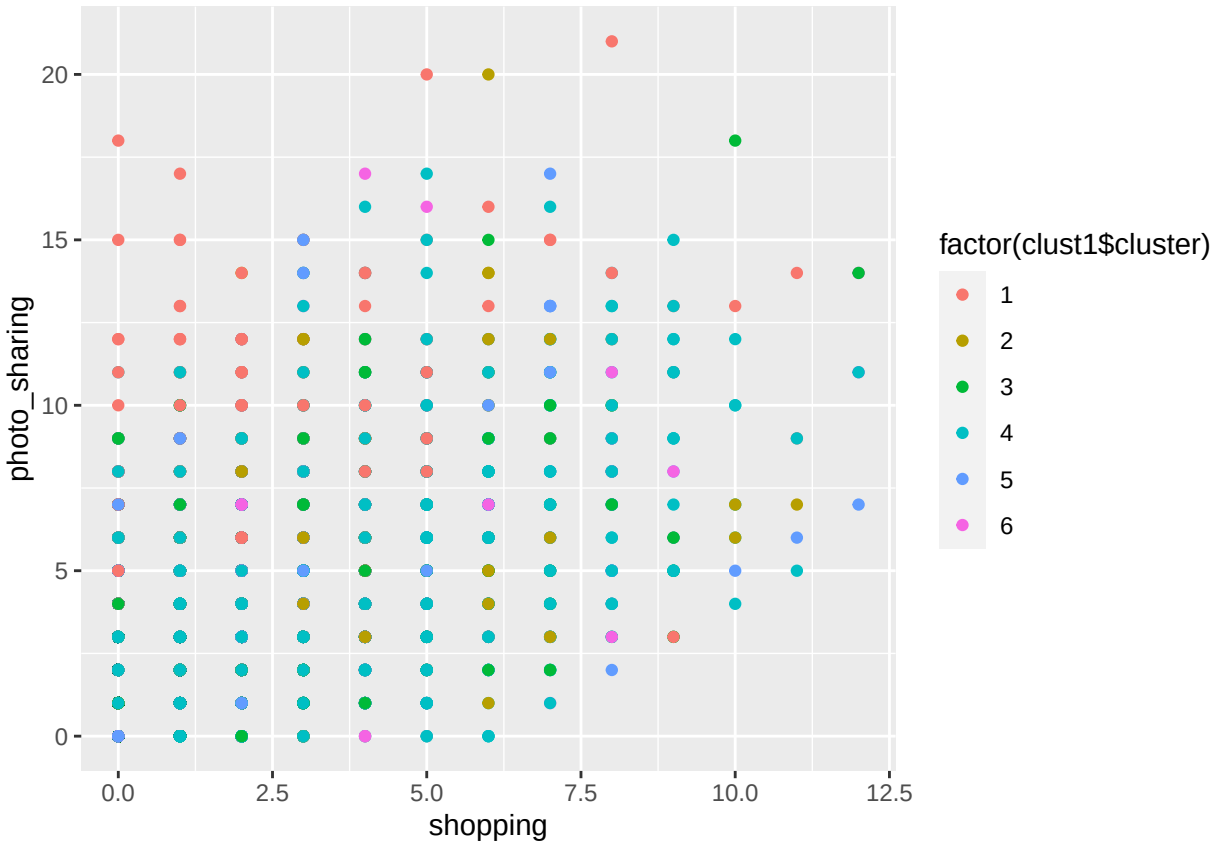

soc_med_data$outdoors
0
2
4
6
8
10
12
0
5
10
15
soc_med_data$personal_fitness

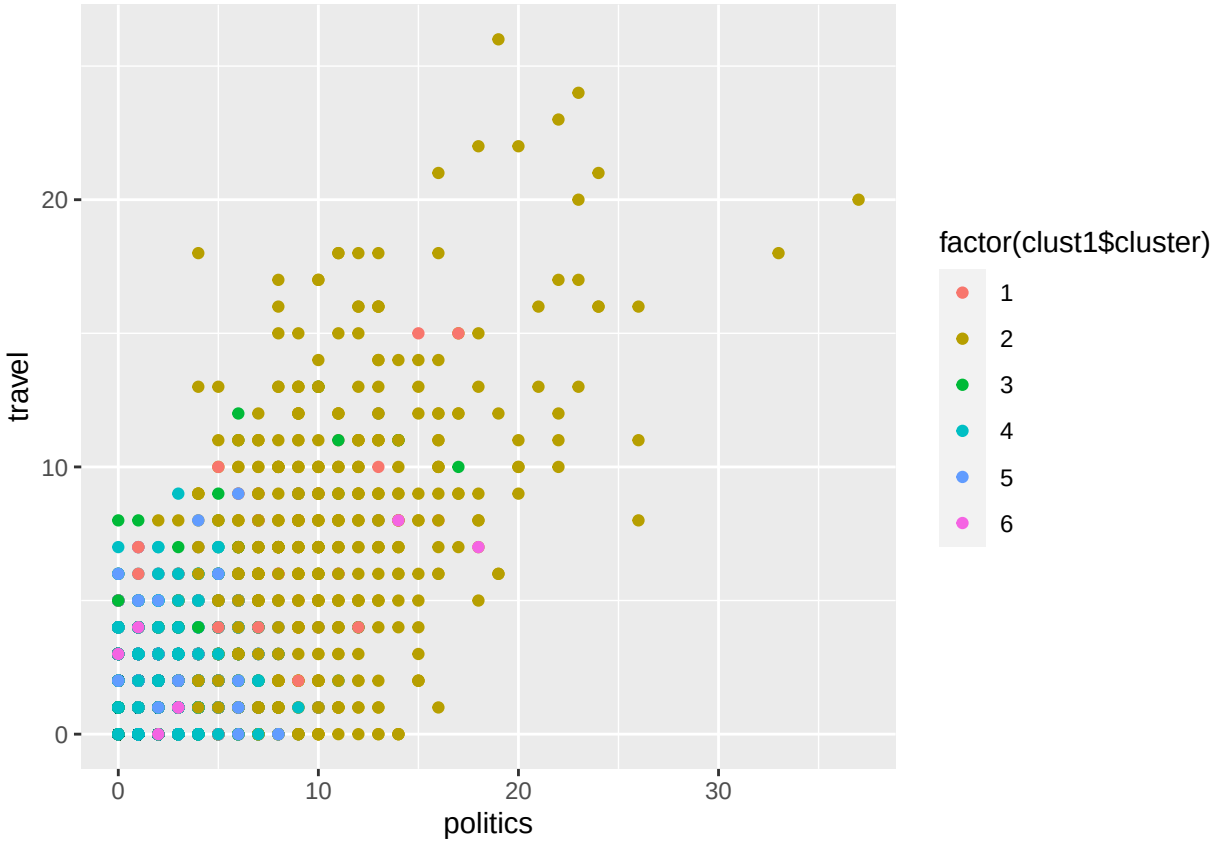

soc_med_data$outdoors
0
2
4
6
8
10
12
0
10
20
30
40
soc_med_data$health_nutrition


soc_med_data$cooking
0
5
10
15
20
25
30
0
5
10
15
soc_med_data$fashion


soc_med_data$beauty
0
2
4
6
8
10
12
14
0
5
10
15
soc_med_data$fashion


soc_med_data$beauty
0
2
4
6
8
10
12
14
0
5
10
15
20
25
30
soc_med_data$cooking


photo_sharing
shopping
0
5
10
15
20
0.0
2.5
5.0
7.5
10.0
12.5
factor(clust1$cluster)
1
2
3
4
5
6



Pre-processing data: *

All the files will be taken to be used for training and store their names in 'file_list'

As the data does not have the column name for the Author name, the file name will be used to extract it

Cleaning up the file name using piping operator from magrittr

Some pre-processing/tokenization steps uses tm_map which maps some function to every document in the corpus

Using this we make everything lowercase, remove numbers, punctuations, excess white spaces and stopwords

```
## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removeNumbers)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(stripWhitespace)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(removeWords), :
## transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(stemDocument)):
## transformation drops documents
```

Creating a doc-term-matrix

Removing rare terms

Removing those terms that have count 0 in >97% of docs.

```
## <<DocumentTermMatrix (documents: 2500, terms: 1264)>>
## Non-/sparse entries: 360611/2799389
## Sparsity           : 89%
## Maximal term length: 13
## Weighting          : term frequency (tf)
```

Constructing TF IDF weights

Repeating the above steps to obtain the test data

```
## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(tolower)):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removeNumbers)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removePunctuation)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(stripWhitespace)): transformation drops documents

## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(removeWords), :
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_documents, content_transformer(stemDocument)):
## transformation drops documents
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 22987)>>
## Non-/sparse entries: 502733/56964767
## Sparsity           : 99%
## Maximal term length: 45
## Weighting          : term frequency (tf)
```

```
## [1] "DocumentTermMatrix"    "simple_triplet_matrix"
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 1296)>>
## Non-/sparse entries: 366936/2873064
## Sparsity           : 89%
## Maximal term length: 13
## Weighting          : term frequency (tf)
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 1296)>>
## Non-/sparse entries: 326936/2913064
## Sparsity           : 90%
## Maximal term length: 13
## Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(xc)‘ instead of ‘xc‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

- Running Random Forest on the Train Data *

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
## Call:
##  randomForest(formula = new_train$V1 ~ ., data = new_train, ntree = 37,      proximity = T)
##                Type of random forest: classification
##                      Number of trees: 37
## No. of variables tried at each split: 34
```

```
##
##         OOB estimate of  error rate: 26.52%
## Confusion matrix:
##                   AaronPressman AlanCrosby AlexanderSmith BenjaminKangLim
## AaronPressman                39          0              0               0
## AlanCrosby                    0         40              0               0
## AlexanderSmith                0          0             29               0
## BenjaminKangLim               1          0              0              31
## BernardHickey                 1          0              0               0
## BradDorfman                   1          1              0               0
## DarrenSchuettler              0          0              0               0
## DavidLawder                   0          0              0               0
## EdnaFernandes                 0          0              0               0
## EricAuchard                   1          1              1               0
## FumikoFujisaki                1          0              0               0
## GrahamEarnshaw                0          0              0               0
## HeatherScoffield              0          0              0               0
## JaneMacartney                 0          0              0               3
## JanLopatka                    0          4              0               0
## JimGilchrist                  0          0              0               0
## JoeOrtiz                      0          0              2               0
## JohnMastrini                  0          4              0               0
## JonathanBirt                  0          0              0               0
## JoWinterbottom                0          0              0               0
## KarlPenhaul                   0          0              1               0
## KeithWeir                     0          0              0               0
## KevinDrawbaugh                0          0              0               0
## KevinMorrison                 0          0              0               0
## KirstinRidley                 0          0              2               0
## KouroshKarimkhany             0          0              0               0
## LydiaZajc                     0          0              0               0
## LynneO'Donnell                0          0              0               0
## LynnleyBrowning               0          0              0               0
## MarcelMichelson               0          0              0               0
## MarkBendeich                  0          0              1               0
## MartinWolk                    0          0              0               0
## MatthewBunce                  0          0              0               0
## MichaelConnor                 0          0              0               0
## MureDickie                    0          0              0              10
## NickLouth                     0          0              0               0
## PatriciaCommins               0          0              0               0
## PeterHumphrey                 0          0              0               1
## PierreTran                    0          0              1               0
## RobinSidel                    1          0              0               0
## RogerFillion                  1          0              0               0
## SamuelPerry                   5          0              0               0
## SarahDavison                  1          0              1               1
## ScottHillis                   0          0              0              11
## SimonCowell                   2          0              1               0
## TanEeLyn                      0          0              0               0
## TheresePoletti                1          0              0               0
## TimFarrand                    0          0              2               0
## ToddNissen                    0          0              0               0
## WilliamKazer                  0          0              0               2
```

```
##                   BernardHickey BradDorfman DarrenSchuettler DavidLawder
## AaronPressman                 0           0                 0           0
## AlanCrosby                    0           0                 0           0
## AlexanderSmith                2           0                 0           0
## BenjaminKangLim               0           0                 0           0
## BernardHickey                34           0                 0           1
## BradDorfman                   0          26                 1           1
## DarrenSchuettler              0           0                44           0
## DavidLawder                   0           0                 0          46
## EdnaFernandes                 1           0                 0           1
## EricAuchard                   0           0                 0           0
## FumikoFujisaki                0           0                 0           0
## GrahamEarnshaw                0           0                 0           0
## HeatherScoffield              0           1                 1           0
## JaneMacartney                 0           0                 0           0
## JanLopatka                    0           0                 0           0
## JimGilchrist                  0           0                 0           0
## JoeOrtiz                      0           0                 0           0
## JohnMastrini                  0           0                 0           0
## JonathanBirt                  0           0                 0           0
## JoWinterbottom                0           0                 0           0
## KarlPenhaul                   0           0                 0           0
## KeithWeir                     0           0                 0           0
## KevinDrawbaugh                0           1                 1           0
## KevinMorrison                 6           0                 0           0
## KirstinRidley                 0           0                 0           0
## KouroshKarimkhany             0           0                 0           0
## LydiaZajc                     0           0                 7           0
## LynneO'Donnell                0           0                 0           0
## LynnleyBrowning               0           0                 0           0
## MarcelMichelson               0           0                 0           0
## MarkBendeich                  6           0                 0           0
## MartinWolk                    1           1                 0           0
## MatthewBunce                  0           0                 0           0
## MichaelConnor                 2           4                 0           0
## MureDickie                    0           0                 0           0
## NickLouth                     1           0                 0           0
## PatriciaCommins               0           8                 0           0
## PeterHumphrey                 0           0                 1           0
## PierreTran                    0           0                 0           1
## RobinSidel                    0           0                 0           0
## RogerFillion                  0           0                 1           0
## SamuelPerry                   0           1                 0           0
## SarahDavison                  0           0                 1           0
## ScottHillis                   0           0                 0           0
## SimonCowell                   0           1                 0           0
## TanEeLyn                      0           0                 0           0
## TheresePoletti                0           1                 0           0
## TimFarrand                    1           0                 0           0
## ToddNissen                    0           1                 1           3
## WilliamKazer                  0           0                 0           1
##                   EdnaFernandes EricAuchard FumikoFujisaki GrahamEarnshaw
## AaronPressman                 0           0               1              0
## AlanCrosby                    0           0               0              0
```

```
## AlexanderSmith                0          0             1             0
## BenjaminKangLim               0          0             0             0
## BernardHickey                 1          1             0             0
## BradDorfman                   0          1             0             0
## DarrenSchuettler              0          0             0             0
## DavidLawder                   0          0             0             0
## EdnaFernandes                32          0             0             0
## EricAuchard                   0         28             0             0
## FumikoFujisaki                0          0            47             0
## GrahamEarnshaw                0          0             0            41
## HeatherScoffield              0          0             0             0
## JaneMacartney                 0          0             0             3
## JanLopatka                    0          0             0             0
## JimGilchrist                  0          0             0             0
## JoeOrtiz                      2          0             0             0
## JohnMastrini                  0          0             0             0
## JonathanBirt                  4          0             0             0
## JoWinterbottom                2          0             0             0
## KarlPenhaul                   0          0             0             0
## KeithWeir                     2          0             0             0
## KevinDrawbaugh                0          1             0             0
## KevinMorrison                 0          0             0             0
## KirstinRidley                 0          0             1             0
## KouroshKarimkhany             0          2             0             0
## LydiaZajc                     0          0             0             0
## LynneO'Donnell                0          0             0             0
## LynnleyBrowning               0          0             0             0
## MarcelMichelson               0          0             1             0
## MarkBendeich                  0          0             0             0
## MartinWolk                    0          0             1             0
## MatthewBunce                  0          0             0             0
## MichaelConnor                 1          0             0             0
## MureDickie                    0          0             0             1
## NickLouth                     0          1             0             0
## PatriciaCommins               1          2             1             0
## PeterHumphrey                 0          0             0             0
## PierreTran                    0          0             1             0
## RobinSidel                    1          4             0             0
## RogerFillion                  0          0             0             0
## SamuelPerry                   0          3             0             0
## SarahDavison                  0          0             2             1
## ScottHillis                   1          0             0             0
## SimonCowell                   1          0             0             0
## TanEeLyn                      1          0             0             1
## TheresePoletti                0          4             0             0
## TimFarrand                    3          0             1             0
## ToddNissen                    0          0             0             0
## WilliamKazer                  0          0             0             3
##                  HeatherScoffield JaneMacartney JanLopatka JimGilchrist
## AaronPressman                   0            0          0            0
## AlanCrosby                      0            0          5            0
## AlexanderSmith                  0            0          0            0
## BenjaminKangLim                 0            4          0            1
## BernardHickey                   1            0          0            1
```

```
## BradDorfman                    0            0          0            0
## DarrenSchuettler               0            0          0            0
## DavidLawder                    0            0          0            0
## EdnaFernandes                  0            0          0            1
## EricAuchard                    0            0          0            0
## FumikoFujisaki                 0            0          0            0
## GrahamEarnshaw                 0            2          0            0
## HeatherScoffield              45            0          0            0
## JaneMacartney                  0           27          0            0
## JanLopatka                     0            0         43            0
## JimGilchrist                   0            0          0           49
## JoeOrtiz                       0            0          0            0
## JohnMastrini                   0            0          6            0
## JonathanBirt                   0            0          0            0
## JoWinterbottom                 0            0          0            0
## KarlPenhaul                    0            0          0            0
## KeithWeir                      0            0          0            0
## KevinDrawbaugh                 0            0          0            0
## KevinMorrison                  1            0          0            2
## KirstinRidley                  0            0          0            0
## KouroshKarimkhany              0            0          0            0
## LydiaZajc                      1            0          0            0
## LynneO'Donnell                 0            0          0            0
## LynnleyBrowning                1            0          0            0
## MarcelMichelson                0            0          0            0
## MarkBendeich                   1            0          0            0
## MartinWolk                     1            0          0            0
## MatthewBunce                   0            0          1            0
## MichaelConnor                  0            0          0            0
## MureDickie                     0            2          0            0
## NickLouth                      0            0          0            0
## PatriciaCommins                0            0          0            0
## PeterHumphrey                  0            0          0            0
## PierreTran                     0            0          0            0
## RobinSidel                     0            0          0            0
## RogerFillion                   0            0          0            0
## SamuelPerry                    0            0          0            0
## SarahDavison                   0            0          0            1
## ScottHillis                    0            5          0            0
## SimonCowell                    0            0          1            0
## TanEeLyn                       0            0          0            2
## TheresePoletti                 0            0          0            0
## TimFarrand                     0            0          0            0
## ToddNissen                     0            0          0            0
## WilliamKazer                   0            6          0            0
##                   JoeOrtiz JohnMastrini JonathanBirt JoWinterbottom KarlPenhaul
## AaronPressman            0           0           0              0          0
## AlanCrosby               0           4           1              0          0
## AlexanderSmith           2           0           0              3          0
## BenjaminKangLim          0           0           0              0          0
## BernardHickey            0           0           0              0          0
## BradDorfman              0           0           0              0          1
## DarrenSchuettler         0           0           0              0          0
## DavidLawder              0           0           0              0          0
```

```
## EdnaFernandes            0           0           0           2           0
## EricAuchard              0           0           0           0           0
## FumikoFujisaki           0           0           0           0           1
## GrahamEarnshaw           0           0           0           0           0
## HeatherScoffield         0           0           0           0           0
## JaneMacartney            0           0           0           0           0
## JanLopatka               0           3           0           0           0
## JimGilchrist             0           0           0           0           0
## JoeOrtiz                35           0           0           4           0
## JohnMastrini             0          40           0           0           0
## JonathanBirt             1           0          43           1           0
## JoWinterbottom           3           0           0          44           0
## KarlPenhaul              0           0           0           0          44
## KeithWeir                0           0           0           0           2
## KevinDrawbaugh           0           0           1           0           1
## KevinMorrison            0           0           0           0           0
## KirstinRidley            1           0           1           2           1
## KouroshKarimkhany        0           0           0           0           0
## LydiaZajc                0           0           0           0           0
## LynneO'Donnell           0           0           0           0           0
## LynnleyBrowning          0           0           0           0           0
## MarcelMichelson          0           0           1           0           0
## MarkBendeich             0           0           0           0           1
## MartinWolk               0           0           1           0           1
## MatthewBunce             0           0           0           0           3
## MichaelConnor            0           0           0           0           0
## MureDickie               0           0           0           0           0
## NickLouth                0           0           0           0           0
## PatriciaCommins          0           0           0           0           0
## PeterHumphrey            0           0           0           0           0
## PierreTran               1           0           0           0           0
## RobinSidel               0           0           0           0           0
## RogerFillion             0           0           0           0           0
## SamuelPerry              0           0           0           0           0
## SarahDavison             0           0           1           0           0
## ScottHillis              0           0           0           0           0
## SimonCowell              2           0           0           0           0
## TanEeLyn                 0           1           0           0           0
## TheresePoletti           0           0           0           0           0
## TimFarrand               2           0           1           2           0
## ToddNissen               0           0           0           0           0
## WilliamKazer             0           0           0           0           0
##                  KeithWeir KevinDrawbaugh KevinMorrison KirstinRidley
## AaronPressman            0              0            0             0
## AlanCrosby               0              0            0             0
## AlexanderSmith           3              0            0             3
## BenjaminKangLim          0              0            0             0
## BernardHickey            0              0            3             0
## BradDorfman              0              5            0             0
## DarrenSchuettler         0              0            0             0
## DavidLawder              0              0            0             0
## EdnaFernandes            0              0            1             2
## EricAuchard              0              0            0             0
## FumikoFujisaki           0              0            0             0
```

```
## GrahamEarnshaw            0             0             0             0
## HeatherScoffield          0             0             0             0
## JaneMacartney             0             0             0             0
## JanLopatka                0             0             0             0
## JimGilchrist              0             0             0             0
## JoeOrtiz                  1             0             0             0
## JohnMastrini              0             0             0             0
## JonathanBirt              0             0             0             0
## JoWinterbottom            0             0             0             0
## KarlPenhaul               0             0             0             0
## KeithWeir                40             0             1             1
## KevinDrawbaugh            1            35             2             0
## KevinMorrison             0             1            37             0
## KirstinRidley             3             0             0            30
## KouroshKarimkhany         0             0             0             0
## LydiaZajc                 0             0             0             0
## LynneO'Donnell            0             0             0             0
## LynnleyBrowning           0             0             0             0
## MarcelMichelson           0             0             0             0
## MarkBendeich              0             0             3             0
## MartinWolk                0             1             0             1
## MatthewBunce              1             0             0             0
## MichaelConnor             0             1             1             0
## MureDickie                0             0             0             0
## NickLouth                 0             0             0             1
## PatriciaCommins           0             3             0             0
## PeterHumphrey             0             0             0             0
## PierreTran                0             0             0             1
## RobinSidel                0             3             0             0
## RogerFillion              0             0             0             0
## SamuelPerry               0             0             0             0
## SarahDavison              0             0             0             0
## ScottHillis               0             0             0             0
## SimonCowell               0             0             0             0
## TanEeLyn                  0             0             0             0
## TheresePoletti            0             0             0             0
## TimFarrand                1             0             0             2
## ToddNissen                0             0             0             0
## WilliamKazer              0             0             0             1
##                   KouroshKarimkhany LydiaZajc LynneO'Donnell LynnleyBrowning
## AaronPressman                     1         0              0              0
## AlanCrosby                        0         0              0              0
## AlexanderSmith                    0         0              0              0
## BenjaminKangLim                   0         0              1              0
## BernardHickey                     0         0              0              0
## BradDorfman                       0         0              0              0
## DarrenSchuettler                  0         4              0              0
## DavidLawder                       0         0              0              0
## EdnaFernandes                     0         0              0              0
## EricAuchard                       0         1              0              0
## FumikoFujisaki                    0         0              0              0
## GrahamEarnshaw                    0         0              0              0
## HeatherScoffield                  0         2              0              0
## JaneMacartney                     0         0              1              0
```

```
## JanLopatka                    0         0              0              0
## JimGilchrist                  0         0              1              0
## JoeOrtiz                      0         0              0              0
## JohnMastrini                  0         0              0              0
## JonathanBirt                  0         0              0              0
## JoWinterbottom                0         0              0              0
## KarlPenhaul                   0         0              0              2
## KeithWeir                     0         0              0              0
## KevinDrawbaugh                1         1              0              0
## KevinMorrison                 0         0              0              0
## KirstinRidley                 1         0              0              0
## KouroshKarimkhany            45         0              0              0
## LydiaZajc                     0        42              0              0
## LynneO'Donnell                0         0             49              0
## LynnleyBrowning               0         0              0             48
## MarcelMichelson               0         0              0              0
## MarkBendeich                  0         0              0              0
## MartinWolk                    3         0              0              0
## MatthewBunce                  0         0              0              2
## MichaelConnor                 0         0              0              2
## MureDickie                    0         0              1              0
## NickLouth                     0         0              0              0
## PatriciaCommins               0         0              0              0
## PeterHumphrey                 0         0              0              0
## PierreTran                    0         0              0              1
## RobinSidel                    0         0              0              0
## RogerFillion                  0         0              0              0
## SamuelPerry                   2         1              0              0
## SarahDavison                  0         0              0              0
## ScottHillis                   0         0              0              0
## SimonCowell                   0         0              0              0
## TanEeLyn                      0         0              3              0
## TheresePoletti                4         0              0              0
## TimFarrand                    0         0              0              0
## ToddNissen                    0         0              0              0
## WilliamKazer                  0         0              3              0
##                  MarcelMichelson MarkBendeich MartinWolk MatthewBunce
## AaronPressman                  0            0          1            0
## AlanCrosby                     0            0          0            0
## AlexanderSmith                 0            0          0            0
## BenjaminKangLim                0            0          0            0
## BernardHickey                  0            4          1            0
## BradDorfman                    0            0          1            0
## DarrenSchuettler               0            0          0            0
## DavidLawder                    0            0          0            0
## EdnaFernandes                  1            1          0            0
## EricAuchard                    0            0          1            0
## FumikoFujisaki                 0            0          0            0
## GrahamEarnshaw                 0            0          0            0
## HeatherScoffield               0            0          0            0
## JaneMacartney                  0            0          0            0
## JanLopatka                     0            0          0            0
## JimGilchrist                   0            0          0            0
## JoeOrtiz                       0            0          0            0
```

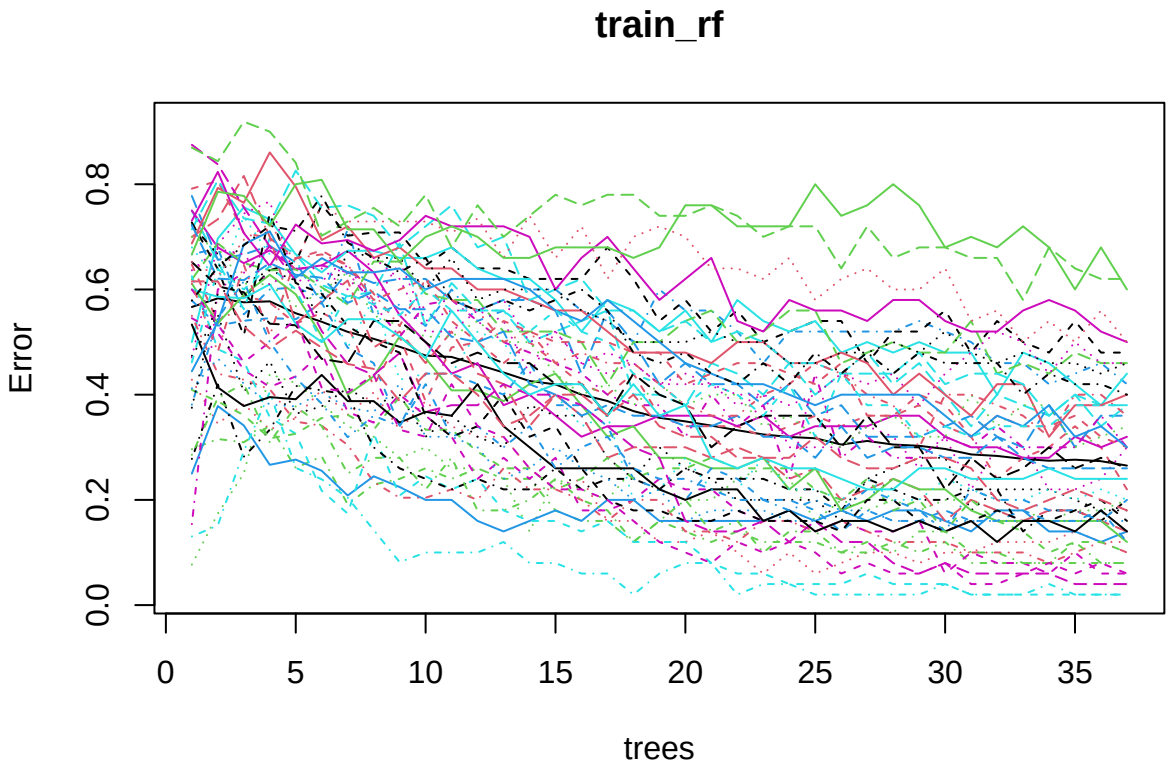
```
## JohnMastrini                    0           0          0           0
## JonathanBirt                    0           0          0           0
## JoWinterbottom                  0           0          0           0
## KarlPenhaul                     0           0          0           1
## KeithWeir                       1           0          0           1
## KevinDrawbaugh                  0           0          0           0
## KevinMorrison                   0           2          0           0
## KirstinRidley                   0           1          1           0
## KouroshKarimkhany               0           0          0           0
## LydiaZajc                       0           0          0           0
## LynneO'Donnell                  0           0          0           0
## LynnleyBrowning                 0           0          0           1
## MarcelMichelson                43           0          1           0
## MarkBendeich                    0          35          0           0
## MartinWolk                      0           1         33           0
## MatthewBunce                    0           0          0          40
## MichaelConnor                   0           1          1           0
## MureDickie                      0           1          0           0
## NickLouth                       0           0          0           0
## PatriciaCommins                 0           0          1           0
## PeterHumphrey                   0           0          0           0
## PierreTran                      6           0          0           1
## RobinSidel                      0           0          1           0
## RogerFillion                    0           0          0           0
## SamuelPerry                     0           0          1           0
## SarahDavison                    0           0          0           0
## ScottHillis                     0           0          0           1
## SimonCowell                     0           1          0           0
## TanEeLyn                        0           1          0           0
## TheresePoletti                  0           0          1           0
## TimFarrand                      0           3          0           0
## ToddNissen                      0           0          0           1
## WilliamKazer                    0           0          0           0
##                  MichaelConnor MureDickie NickLouth PatriciaCommins
## AaronPressman                0          0         1                0
## AlanCrosby                   0          0         0                0
## AlexanderSmith               0          0         0                1
## BenjaminKangLim              1          7         0                0
## BernardHickey                1          0         0                1
## BradDorfman                  3          0         0                2
## DarrenSchuettler             0          0         0                0
## DavidLawder                  0          0         0                0
## EdnaFernandes                0          0         0                0
## EricAuchard                  1          0         4                1
## FumikoFujisaki               0          0         0                0
## GrahamEarnshaw               0          1         0                0
## HeatherScoffield             0          0         0                0
## JaneMacartney                0          1         0                0
## JanLopatka                   0          0         0                0
## JimGilchrist                 0          0         0                0
## JoeOrtiz                     0          0         0                0
## JohnMastrini                 0          0         0                0
## JonathanBirt                 0          0         0                0
## JoWinterbottom               0          0         0                0
```

```
## KarlPenhaul                    0          1          0                0
## KeithWeir                      0          0          1                0
## KevinDrawbaugh                 0          0          0                2
## KevinMorrison                  0          0          0                0
## KirstinRidley                  1          1          1                0
## KouroshKarimkhany              0          0          0                0
## LydiaZajc                      0          0          0                0
## LynneO'Donnell                 0          0          0                0
## LynnleyBrowning                0          0          0                0
## MarcelMichelson                0          0          0                0
## MarkBendeich                   0          1          0                0
## MartinWolk                     1          0          0                0
## MatthewBunce                   0          0          0                0
## MichaelConnor                 31          0          1                1
## MureDickie                     0         25          1                0
## NickLouth                      1          0         42                0
## PatriciaCommins                0          0          0               25
## PeterHumphrey                  0          1          0                0
## PierreTran                     0          0          0                0
## RobinSidel                     0          0          0                1
## RogerFillion                   0          0          0                0
## SamuelPerry                    1          0          1                1
## SarahDavison                   0          1          0                0
## ScottHillis                    1          4          0                0
## SimonCowell                    1          0          0                0
## TanEeLyn                       0          0          0                0
## TheresePoletti                 0          0          3                1
## TimFarrand                     0          0          0                0
## ToddNissen                     0          0          0                1
## WilliamKazer                   0          2          0                0
##                   PeterHumphrey PierreTran RobinSidel RogerFillion SamuelPerry
## AaronPressman                 0          0          1             0           3
## AlanCrosby                    0          0          0             0           0
## AlexanderSmith                0          0          0             0           0
## BenjaminKangLim               1          0          0             0           0
## BernardHickey                 0          0          0             0           0
## BradDorfman                   0          0          1             0           1
## DarrenSchuettler              0          0          1             0           0
## DavidLawder                   0          0          0             0           0
## EdnaFernandes                 0          3          0             0           0
## EricAuchard                   0          0          0             0           3
## FumikoFujisaki                0          0          1             0           0
## GrahamEarnshaw                0          0          0             0           0
## HeatherScoffield              0          0          1             0           0
## JaneMacartney                 0          0          0             0           0
## JanLopatka                    0          0          0             0           0
## JimGilchrist                  0          0          0             0           0
## JoeOrtiz                      0          1          0             0           0
## JohnMastrini                  0          0          0             0           0
## JonathanBirt                  0          0          0             0           0
## JoWinterbottom                0          0          0             0           0
## KarlPenhaul                   0          0          0             1           0
## KeithWeir                     0          0          0             0           0
## KevinDrawbaugh                0          0          1             0           0
```

```
## KevinMorrison                0          0          0          0          0
## KirstinRidley                0          0          0          0          0
## KouroshKarimkhany            0          0          0          0          2
## LydiaZajc                    0          0          0          0          0
## LynneO'Donnell               1          0          0          0          0
## LynnleyBrowning              0          0          0          0          0
## MarcelMichelson              0          4          0          0          0
## MarkBendeich                 0          0          0          0          0
## MartinWolk                   0          1          0          0          2
## MatthewBunce                 0          2          0          0          0
## MichaelConnor                0          0          1          1          0
## MureDickie                   3          0          0          0          1
## NickLouth                    0          0          0          2          2
## PatriciaCommins              0          1          4          1          1
## PeterHumphrey               46          0          0          0          0
## PierreTran                   0         37          0          0          0
## RobinSidel                   0          0         38          0          0
## RogerFillion                 0          0          1         47          0
## SamuelPerry                  0          0          0          0         27
## SarahDavison                 6          0          0          0          0
## ScottHillis                  0          0          0          0          0
## SimonCowell                  0          1          1          0          0
## TanEeLyn                     8          0          0          0          0
## TheresePoletti               0          0          0          0          8
## TimFarrand                   0          0          0          0          0
## ToddNissen                   0          0          1          0          0
## WilliamKazer                 0          0          0          0          0
##                  SarahDavison ScottHillis SimonCowell TanEeLyn TheresePoletti
## AaronPressman                0          0          0        0              2
## AlanCrosby                   0          0          0        0              0
## AlexanderSmith               1          0          3        0              0
## BenjaminKangLim              0          2          0        1              0
## BernardHickey                0          0          0        0              0
## BradDorfman                  0          0          1        0              0
## DarrenSchuettler             0          1          0        0              0
## DavidLawder                  0          0          0        0              0
## EdnaFernandes                0          0          2        0              0
## EricAuchard                  0          0          1        0              7
## FumikoFujisaki               0          0          0        0              0
## GrahamEarnshaw               2          0          0        0              0
## HeatherScoffield             0          0          0        0              0
## JaneMacartney                1          5          0        2              0
## JanLopatka                   0          0          0        0              0
## JimGilchrist                 0          0          0        0              0
## JoeOrtiz                     0          0          3        1              0
## JohnMastrini                 0          0          0        0              0
## JonathanBirt                 0          0          1        0              0
## JoWinterbottom               0          0          0        0              0
## KarlPenhaul                  0          0          0        0              0
## KeithWeir                    0          0          1        0              0
## KevinDrawbaugh               0          0          1        0              0
## KevinMorrison                1          0          0        0              0
## KirstinRidley                0          0          2        0              0
## KouroshKarimkhany            0          0          0        0              1
```

```
## LydiaZajc                  0           0           0        0              0
## LynneO'Donnell             0           0           0        0              0
## LynnleyBrowning            0           0           0        0              0
## MarcelMichelson            0           0           0        0              0
## MarkBendeich               1           0           0        0              0
## MartinWolk                 0           0           0        0              1
## MatthewBunce               0           0           0        1              0
## MichaelConnor              0           0           0        0              0
## MureDickie                 0           2           0        0              0
## NickLouth                  0           0           0        0              0
## PatriciaCommins            0           0           0        0              1
## PeterHumphrey              0           0           0        1              0
## PierreTran                 0           0           0        0              0
## RobinSidel                 0           0           0        0              0
## RogerFillion               0           0           0        0              0
## SamuelPerry                0           0           0        0              5
## SarahDavison              32           1           0        1              0
## ScottHillis                0          19           0        2              0
## SimonCowell                0           0          35        0              0
## TanEeLyn                   0           1           0       31              0
## TheresePoletti             0           0           0        0             27
## TimFarrand                 0           0           2        0              0
## ToddNissen                 0           0           0        0              1
## WilliamKazer               1           9           0        2              0
##                   TimFarrand ToddNissen WilliamKazer class.error
## AaronPressman              0          1            0        0.22
## AlanCrosby                 0          0            0        0.20
## AlexanderSmith             1          1            0        0.42
## BenjaminKangLim            0          0            0        0.38
## BernardHickey              0          0            0        0.32
## BradDorfman                0          4            0        0.48
## DarrenSchuettler           0          0            0        0.12
## DavidLawder                0          4            0        0.08
## EdnaFernandes              3          0            0        0.36
## EricAuchard                0          0            0        0.44
## FumikoFujisaki             0          0            0        0.06
## GrahamEarnshaw             0          0            4        0.18
## HeatherScoffield           0          0            0        0.10
## JaneMacartney              0          1            6        0.46
## JanLopatka                 0          0            0        0.14
## JimGilchrist               0          0            0        0.02
## JoeOrtiz                   0          0            1        0.30
## JohnMastrini               0          0            0        0.20
## JonathanBirt               0          0            0        0.14
## JoWinterbottom             1          0            0        0.12
## KarlPenhaul                0          0            0        0.12
## KeithWeir                  0          0            0        0.20
## KevinDrawbaugh             0          1            0        0.30
## KevinMorrison              0          0            0        0.26
## KirstinRidley              1          0            0        0.40
## KouroshKarimkhany          0          0            0        0.10
## LydiaZajc                  0          0            0        0.16
## LynneO'Donnell             0          0            0        0.02
## LynnleyBrowning            0          0            0        0.04
```

```
## MarcelMichelson          0          0          0        0.14
## MarkBendeich             1          0          0        0.30
## MartinWolk               0          0          0        0.34
## MatthewBunce             0          0          0        0.20
## MichaelConnor            1          1          0        0.38
## MureDickie               0          0          3        0.50
## NickLouth                0          0          0        0.16
## PatriciaCommins          0          1          0        0.50
## PeterHumphrey            0          0          0        0.08
## PierreTran               0          0          0        0.26
## RobinSidel               0          1          0        0.24
## RogerFillion             0          0          0        0.06
## SamuelPerry              2          0          0        0.46
## SarahDavison             0          0          0        0.36
## ScottHillis              0          0          6        0.62
## SimonCowell              3          0          0        0.30
## TanEeLyn                 0          0          1        0.38
## TheresePoletti           0          0          0        0.46
## TimFarrand              30          0          0        0.40
## ToddNissen               0         41          0        0.18
## WilliamKazer             0          0         20        0.60
```



- Running the model on the Test Data *

```
## [1] 0.672
```

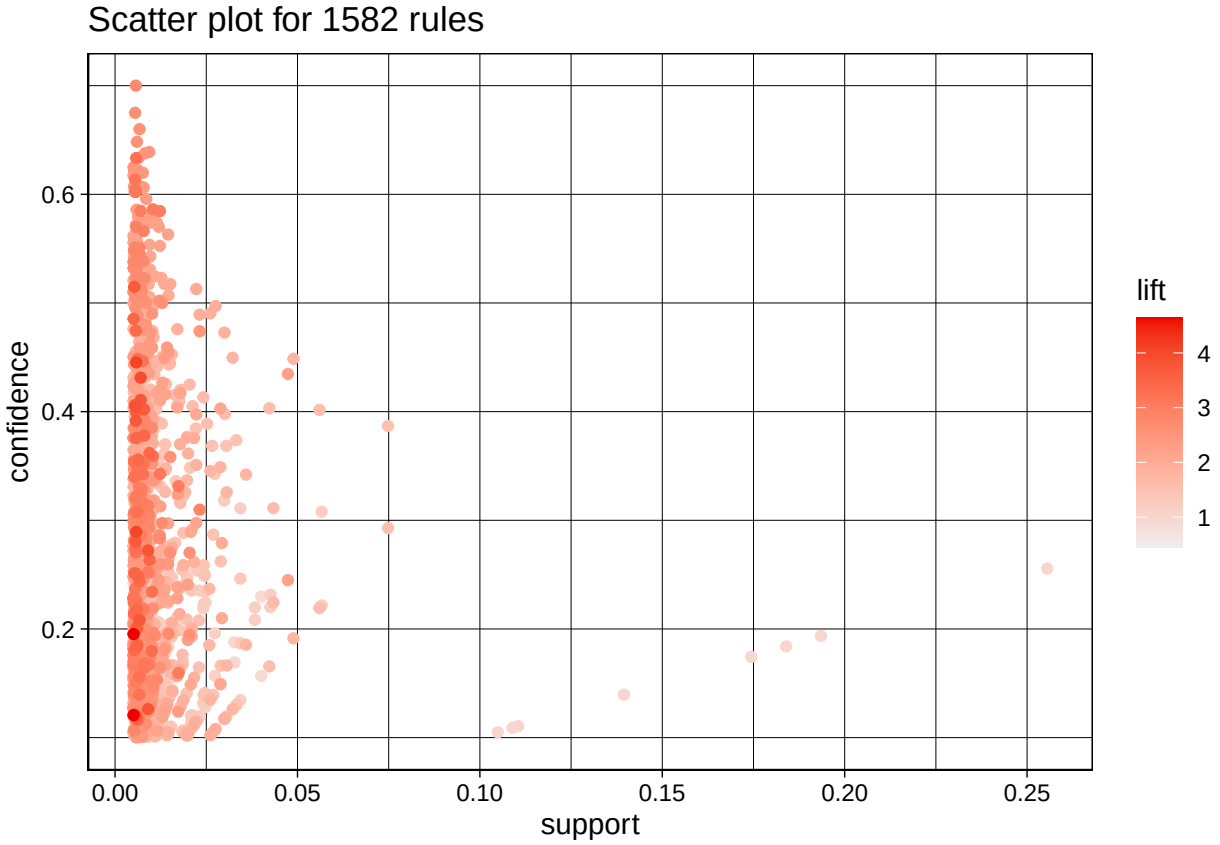
- Running Random Forest on Train data
- Checking model on test data *
- The performance does not improve that much for various values used to subset features.
- Summary of the Process: *
- We first extract the author name from the file paths.
- Next, we clean the file names and pre-process it in the following order:

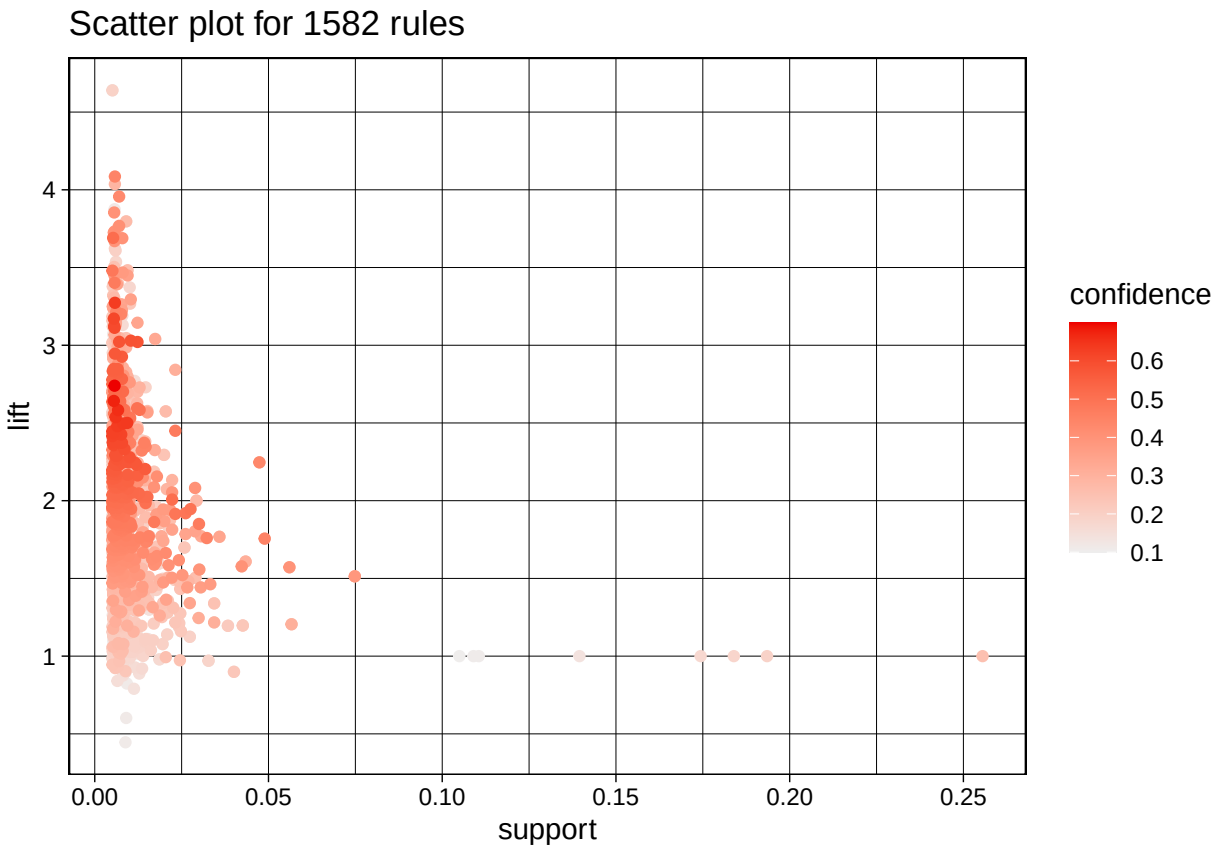
a) Join all files and convert into one corpus. This corpus will have rows as each document.
b) Tokenize the documents(split each document into separate words) and convert to lower case c)Remove numbers, punctuation, extra white spaces and stop words (Words like 'as','the','so' do not add much meaning to the sentence without context. Hence we remove them to reduce any noise in the data) from the document
c) Stemming (Words such as 'run' and 'running' essentially mean the same. So in stemming, we take each word and use it's root value. In this example, our root value will be 'run')

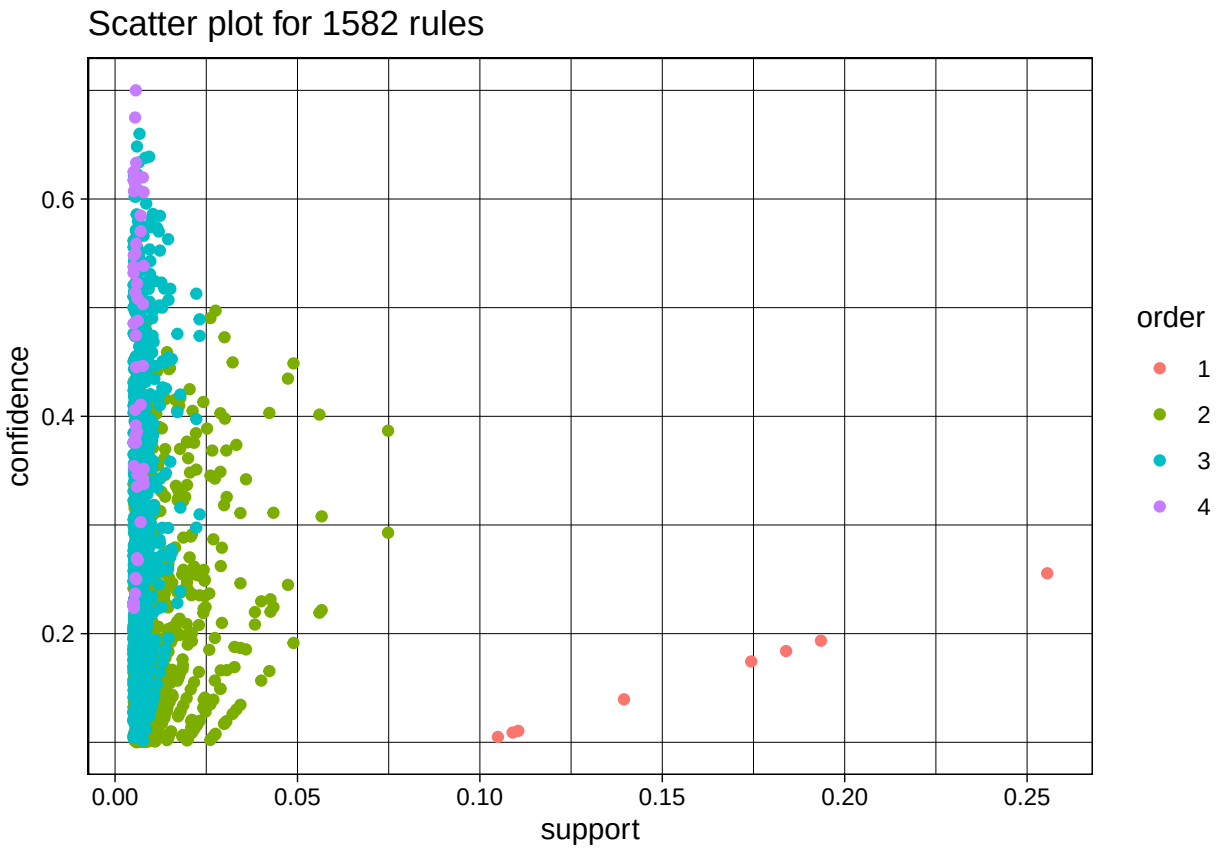
- We convert the output from step 3 to a sparse matrix. Rows in a sparse matrix represent data for each document. The columns in the sparse matrix represent each word identified after Step 2. The values for a particular row, column in the matrix is the number of times the word appears in the particular document.
- We drop terms that may occur only once or twice in the documents. This further removes some noise from the data and reduces number of features.
- Some texts can be small while some can be large. To compare several texts, the frequency of each word relative to the length of the text is more helpful than the count of each word in the text. We thus use the TF IDF values for this Purpose. So we replace values in the sparse matrix to TF IDF scores.
- We then merge the author names with output from 5 to get train data.
- Repeat steps 1 to step 6 using test data.
- We ignore words present in the test set but not in the train set.
- Using an intersection of words between train and test set, We now run Random Forest for classification to get accuracy of 70.8%.
- We try using a number of most important features to reduce dimensionality. However, this does not improve the performance.

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
```

```
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```




Scatter plot for 1582 rules
lift
4
3
2
1
0.00
0.05
0.10
0.15
0.20
0.25
support
confidence
0.6
0.5
0.4
0.3
0.2
0.1


Scatter plot for 1582 rules
confidence
0.6
0.4
0.2
0.00
0.05
0.10
0.15
0.20
0.25
support
order
1
2
3
4