Data Analysis using R

Data Cleaning

Dimension Reduction Techniques:

1. Missing Value Filter

2. Low Variance Filter

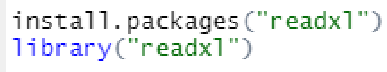
3. High Correlation Filter

Dataset context:

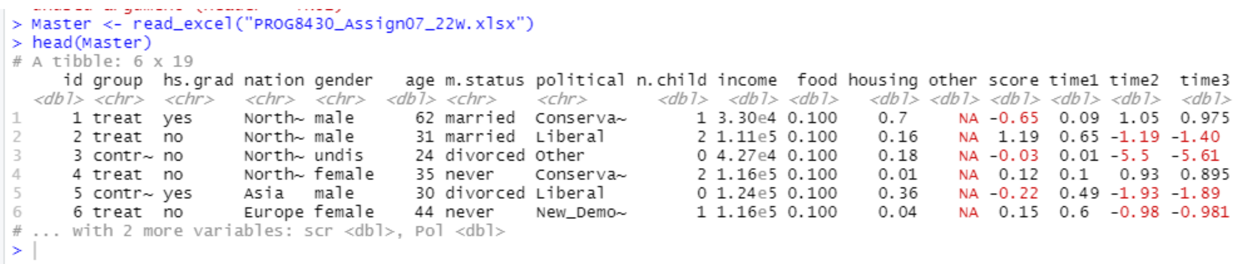
Key factors associated with political engagement for 3312 residents in Canada.

Data Transformation and Preparation:

Import data:



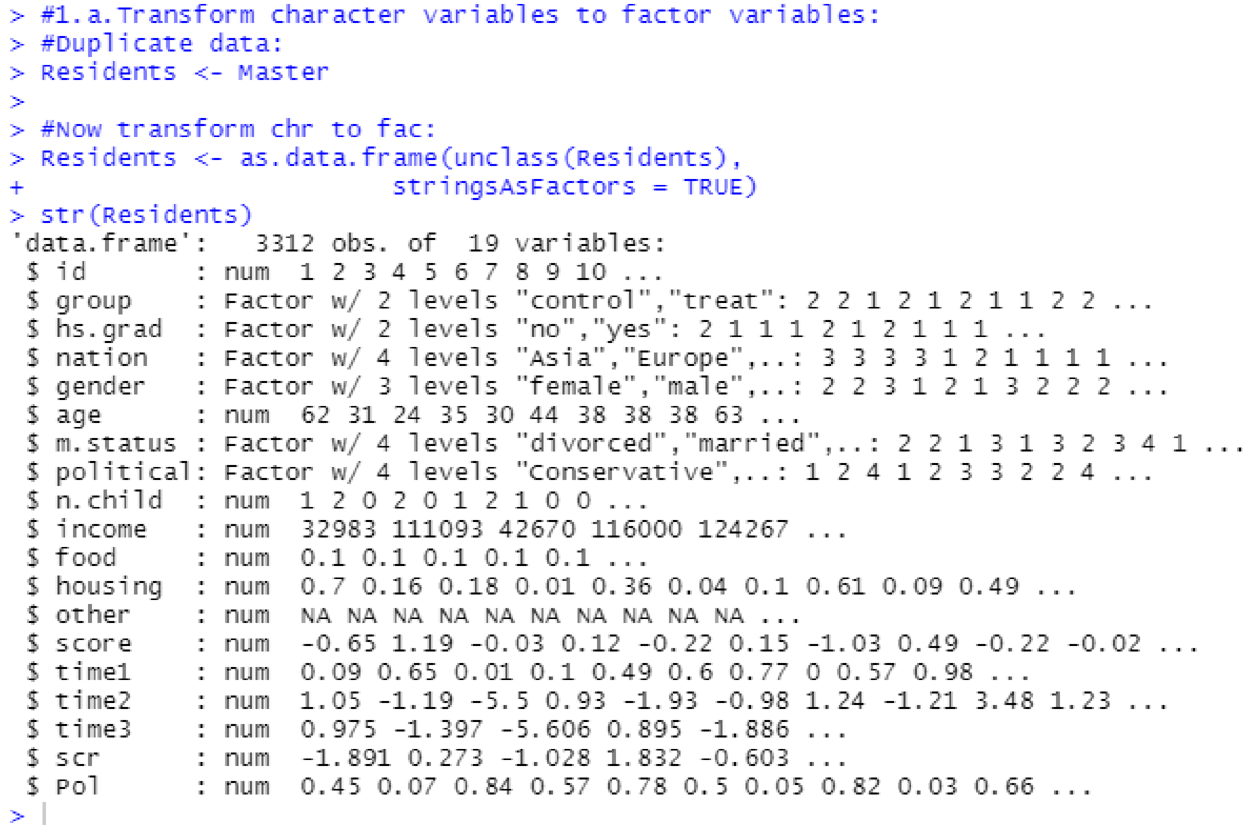
```
install.packages("readxl")
library("readxl")
```



```
> Master <- read_excel("PROG8430_Assign07_22W.xlsx")
> head(Master)
# A tibble: 6 x 19
     id group  hs.grad nation gender   age m.status political n.child income  food housing other score time1 time2  time3
  <dbl> <chr>  <chr>   <chr>  <chr>  <dbl> <chr>    <chr>       <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
1     1 treat  yes     North~ male      62 married  Conserva~       1 3.30e4 0.100    0.7     NA -0.65  0.09  1.05  0.975
2     2 treat  no      North~ male      31 married  Liberal         2 1.11e5 0.100    0.16    NA  1.19  0.65 -1.19 -1.40
3     3 contr~ no      North~ undis     24 divorced Other           0 4.27e4 0.100    0.18    NA -0.03  0.01 -5.5  -5.61
4     4 treat  no      North~ female    35 never    Conserva~       2 1.16e5 0.100    0.01    NA  0.12  0.1   0.93  0.895
5     5 contr~ yes     Asia   male      30 divorced Liberal         0 1.24e5 0.100    0.36    NA -0.22  0.49 -1.93 -1.89
6     6 treat  no      Europe female    44 never    New_Demo~       1 1.16e5 0.100    0.04    NA  0.15  0.6  -0.98 -0.981
# ... with 2 more variables: scr <dbl>, Pol <dbl>
> |
```

Note: we see that we have both character and numerical variables.

a. Transform character variables to factor variables:



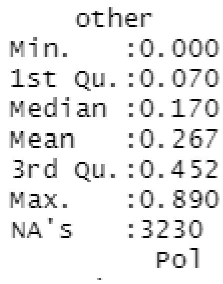
```
> #1.a.Transform character variables to factor variables:
> #Duplicate data:
> Residents <- Master
>
> #Now transform chr to fac:
> Residents <- as.data.frame(unclass(Residents),
+                       stringsAsFactors = TRUE)
> str(Residents)
'data.frame':	3312 obs. of  19 variables:
 $ id       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ group    : Factor w/ 2 levels "control","treat": 2 2 1 2 1 2 1 1 2 2 ...
 $ hs.grad  : Factor w/ 2 levels "no","yes": 2 1 1 1 2 1 2 1 1 1 ...
 $ nation   : Factor w/ 4 levels "Asia","Europe",..: 3 3 3 3 1 2 1 1 1 1 ...
 $ gender   : Factor w/ 3 levels "female","male",..: 2 2 3 1 2 1 3 2 2 2 ...
 $ age      : num  62 31 24 35 30 44 38 38 38 63 ...
 $ m.status : Factor w/ 4 levels "divorced","married",..: 2 2 1 3 1 3 2 3 4 1 ...
 $ political: Factor w/ 4 levels "Conservative",..: 1 2 4 1 2 3 3 2 2 4 ...
 $ n.child  : num  1 2 0 2 0 1 2 1 0 0 ...
 $ income   : num  32983 111093 42670 116000 124267 ...
 $ food     : num  0.1 0.1 0.1 0.1 0.1 ...
 $ housing  : num  0.7 0.16 0.18 0.01 0.36 0.04 0.1 0.61 0.09 0.49 ...
 $ other    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ score    : num  -0.65 1.19 -0.03 0.12 -0.22 0.15 -1.03 0.49 -0.22 -0.02 ...
 $ time1    : num  0.09 0.65 0.01 0.1 0.49 0.6 0.77 0 0.57 0.98 ...
 $ time2    : num  1.05 -1.19 -5.5 0.93 -1.93 -0.98 1.24 -1.21 3.48 1.23 ...
 $ time3    : num  0.975 -1.397 -5.606 0.895 -1.886 ...
 $ scr      : num  -1.891 0.273 -1.028 1.832 -0.603 ...
 $ Pol      : num  0.45 0.07 0.84 0.57 0.78 0.5 0.05 0.82 0.03 0.66 ...
> |
```

Result: all character variables have been changed to factor variables for easier statistical analysis.

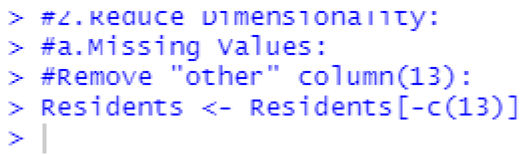
Reduce Dimensionality:

a. Apply Missing Values Filter:

-col13: “other” has 3230 Na’s (>97%) and the “other expense” has no significant meaning in the context, since what is considered “other expense” is not standardized for every resident, therefore we remove it.



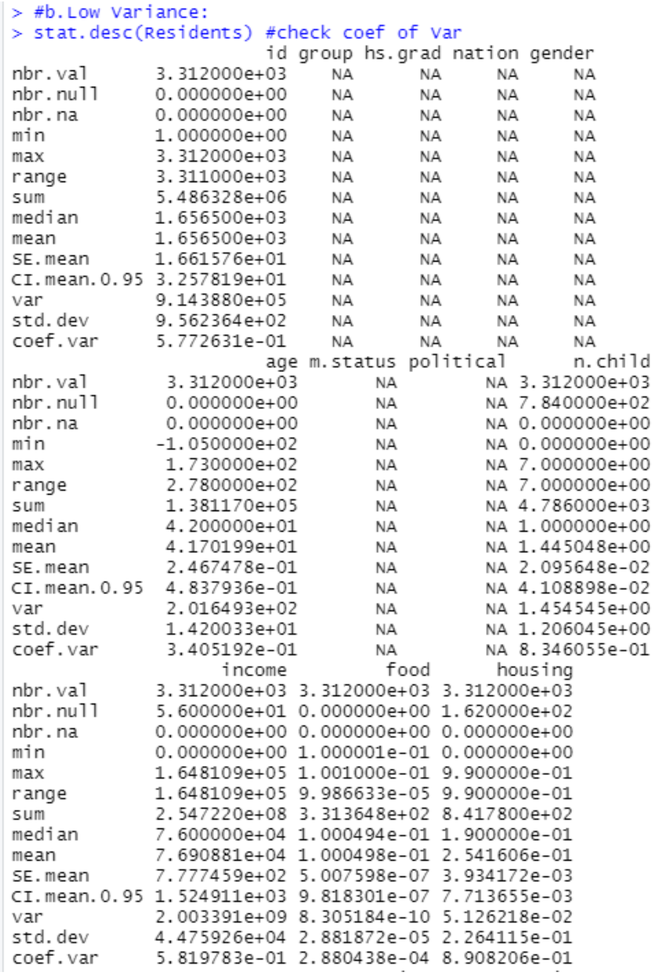
```
     other
 Min.   :0.000
 1st Qu.:0.070
 Median :0.170
 Mean   :0.267
 3rd Qu.:0.452
 Max.   :0.890
 NA's   :3230
          Pol
```



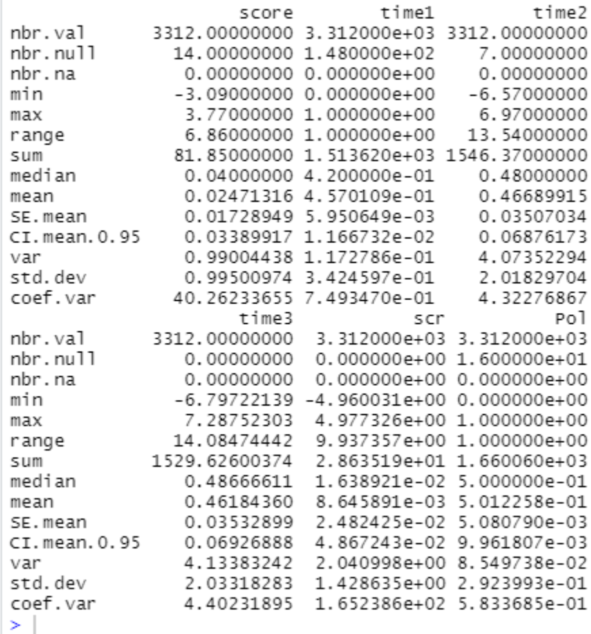
```
> #2.Reduce Dimensionality:
> #a.Missing Values:
> #Remove "other" column(13):
> Residents <- Residents[-c(13)]
> 
```

b. Apply Low Variance filter:

-Low Variance(coef of var): means data not significant, has no or minimal effect on dataset if removed.



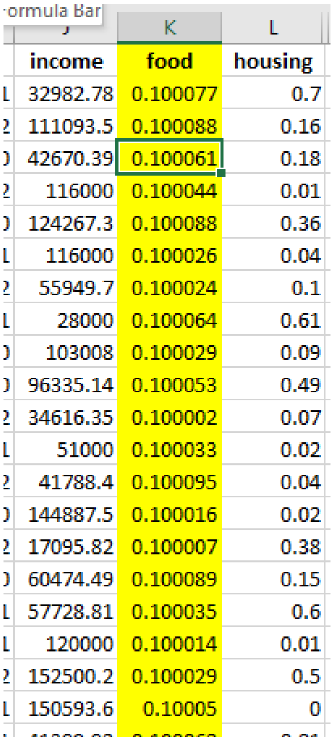
```
> #b.Low Variance:
> stat.desc(Residents) #check coef of Var
                        id group hs.grad nation gender
nbr.val       3.312000e+03    NA      NA     NA     NA
nbr.null      0.000000e+00    NA      NA     NA     NA
nbr.na        0.000000e+00    NA      NA     NA     NA
min           1.000000e+00    NA      NA     NA     NA
max           3.312000e+03    NA      NA     NA     NA
range         3.311000e+03    NA      NA     NA     NA
sum           5.486328e+06    NA      NA     NA     NA
median        1.656500e+03    NA      NA     NA     NA
mean          1.656500e+03    NA      NA     NA     NA
SE.mean       1.661576e+01    NA      NA     NA     NA
CI.mean.0.95  3.257819e+01    NA      NA     NA     NA
var           9.143880e+05    NA      NA     NA     NA
std.dev       9.562364e+02    NA      NA     NA     NA
coef.var      5.772631e-01    NA      NA     NA     NA
                        age m.status political      n.child
nbr.val        3.312000e+03       NA        NA 3.312000e+03
nbr.null       0.000000e+00       NA        NA 7.840000e+02
nbr.na         0.000000e+00       NA        NA 0.000000e+00
min           -1.050000e+02       NA        NA 0.000000e+00
max            1.730000e+02       NA        NA 7.000000e+00
range          2.780000e+02       NA        NA 7.000000e+00
sum            1.381170e+05       NA        NA 4.786000e+03
median         4.200000e+01       NA        NA 1.000000e+00
mean           4.170199e+01       NA        NA 1.445048e+00
SE.mean        2.467478e-01       NA        NA 2.095648e-02
CI.mean.0.95   4.837936e-01       NA        NA 4.108898e-02
var            2.016493e+02       NA        NA 1.454545e+00
std.dev        1.420033e+01       NA        NA 1.206045e+00
coef.var       3.405192e-01       NA        NA 8.346055e-01
                    income         food      housing
nbr.val       3.312000e+03 3.312000e+03 3.312000e+03
nbr.null      5.600000e+01 0.000000e+00 1.620000e+02
nbr.na        0.000000e+00 0.000000e+00 0.000000e+00
min           0.000000e+00 1.000001e-01 0.000000e+00
max           1.648109e+05 1.001000e-01 9.900000e-01
range         1.648109e+05 9.986633e-05 9.900000e-01
sum           2.547220e+08 3.313648e+02 8.417800e+02
median        7.600000e+04 1.000494e-01 1.900000e-01
mean          7.690881e+04 1.000498e-01 2.541606e-01
SE.mean       7.777459e+02 5.007598e-07 3.934172e-03
CI.mean.0.95  1.524911e+03 9.818301e-07 7.713655e-03
var           2.003391e+09 8.305184e-10 5.126218e-02
std.dev       4.475926e+04 2.881872e-05 2.264115e-01
coef.var      5.819783e-01 2.880438e-04 8.908206e-01
```



```
                    score        time1         time2
nbr.val       3312.00000000 3.312000e+03 3312.00000000
nbr.null        14.00000000 1.480000e+02    7.00000000
nbr.na           0.00000000 0.000000e+00    0.00000000
min             -3.09000000 0.000000e+00   -6.57000000
max              3.77000000 1.000000e+00    6.97000000
range            6.86000000 1.000000e+00   13.54000000
sum             81.85000000 1.513620e+03 1546.37000000
median           0.04000000 4.200000e-01    0.48000000
mean             0.02471316 4.570109e-01    0.46689915
SE.mean          0.01728949 5.950649e-03    0.03507034
CI.mean.0.95     0.03389917 1.166732e-02    0.06876173
var              0.99004438 1.172786e-01    4.07352294
std.dev          0.99500974 3.424597e-01    2.01829704
coef.var        40.26233655 7.493470e-01    4.32276867
                    time3           scr          Pol
nbr.val       3312.00000000  3.312000e+03 3.312000e+03
nbr.null         0.00000000  0.000000e+00 1.600000e+01
nbr.na           0.00000000  0.000000e+00 0.000000e+00
min             -6.79722139 -4.960031e+00 0.000000e+00
max              7.28752303  4.977326e+00 1.000000e+00
range           14.08474442  9.937357e+00 1.000000e+00
sum           1529.62600374  2.863519e+01 1.660060e+03
median           0.48666611  1.638921e-02 5.000000e-01
mean             0.46184360  8.645891e-03 5.012258e-01
SE.mean          0.03532899  2.482425e-02 5.080790e-03
CI.mean.0.95     0.06926888  4.867243e-02 9.961807e-03
var              4.13383242  2.040998e+00 8.549738e-02
std.dev          2.03318283  1.428635e+00 2.923993e-01
coef.var         4.40231895  1.652386e+02 5.833685e-01
> 
```

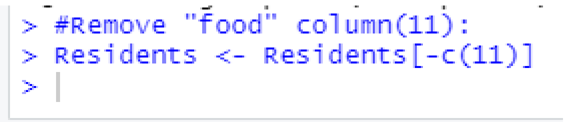
Based on the result, the column which has the lowest coef.var is "food".
Col11. **food**. Coef.Var = 0.000288

Double check the column, values indeed look alike:



income	food	housing
32982.78	0.100077	0.7
111093.5	0.100088	0.16
42670.39	0.100061	0.18
116000	0.100044	0.01
124267.3	0.100088	0.36
116000	0.100026	0.04
55949.7	0.100024	0.1
28000	0.100064	0.61
103008	0.100029	0.09
96335.14	0.100053	0.49
34616.35	0.100002	0.07
51000	0.100033	0.02
41788.4	0.100095	0.04
144887.5	0.100016	0.02
17095.82	0.100007	0.38
60474.49	0.100089	0.15
57728.81	0.100035	0.6
120000	0.100014	0.01
152500.2	0.100029	0.5
150593.6	0.10005	0

Therefore, we remove the column “food” since there is not much difference in residents’ income percentage spending on food :

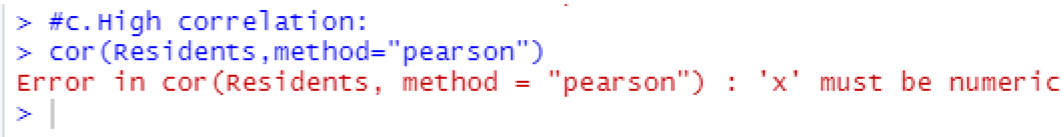


```
> #Remove "food" column(11):
> Residents <- Residents[-c(11)]
> 
```

c. Apply High Correlation Filter:

- High correlation btw variables means that data might be “overly described”, number of variables that provide similar information on the dataset may be reduced.

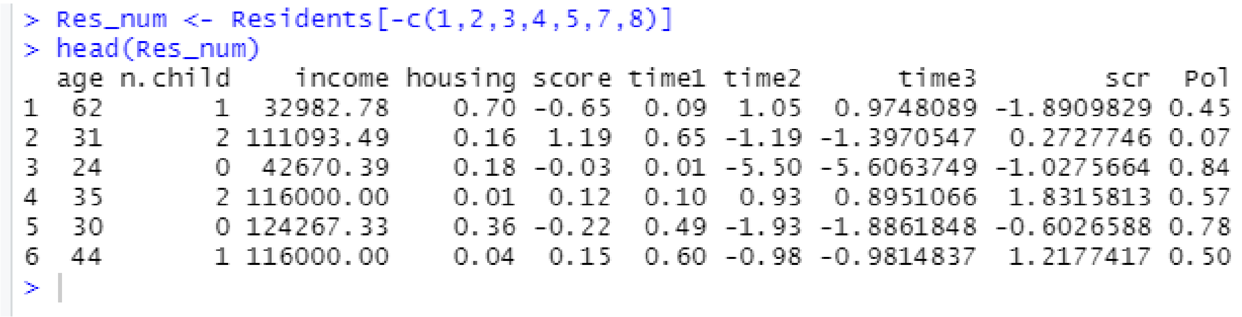
step1: run correlation function on the dataframe:



```
> #c.High correlation:
> cor(Residents,method="pearson")
Error in cor(Residents, method = "pearson") : 'x' must be numeric
> 
```

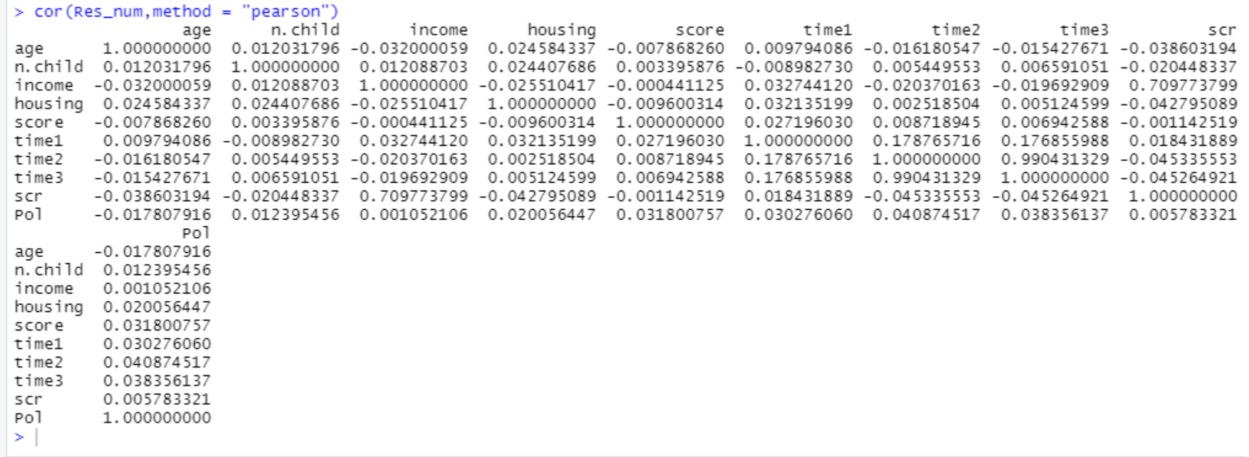
Note: we got an error since the Pearson correlation examines whether a statistically significant linear relationship exists between two continuous variables, we might need to remove the factor variables to apply such method. Also, “id” column should be removed since id does not provide significant statistical importance.

=>We temporarily remove columns 1, 2, 3, 4, 5, 7, 8 to perform the Pearson correlation test.



```
> Res_num <- Residents[-c(1,2,3,4,5,7,8)]
> head(Res_num)
  age n.child    income housing score time1 time2      time3        scr  Pol
1  62       1  32982.78    0.70 -0.65  0.09  1.05  0.9748089 -1.8909829 0.45
2  31       2 111093.49    0.16  1.19  0.65 -1.19 -1.3970547  0.2727746 0.07
3  24       0  42670.39    0.18 -0.03  0.01 -5.50 -5.6063749 -1.0275664 0.84
4  35       2 116000.00    0.01  0.12  0.10  0.93  0.8951066  1.8315813 0.57
5  30       0 124267.33    0.36 -0.22  0.49 -1.93 -1.8861848 -0.6026588 0.78
6  44       1 116000.00    0.04  0.15  0.60 -0.98 -0.9814837  1.2177417 0.50
> 
```

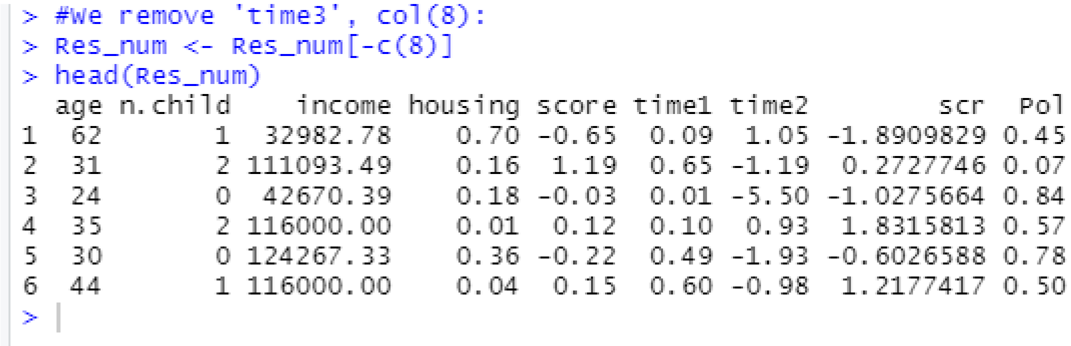
Now apply the Pearson test again:



```
> cor(Res_num,method = "pearson")
                 age      n.child       income      housing        score        time1        time2        time3          scr
age      1.000000000  0.012031796 -0.032000059  0.024584337 -0.007868260  0.009794086 -0.016180547 -0.015427671 -0.038603194
n.child  0.012031796  1.000000000  0.012088703  0.024407686  0.003395876 -0.008982730  0.005449553  0.006591051 -0.020448337
income  -0.032000059  0.012088703  1.000000000 -0.025510417 -0.000441125  0.032744120 -0.020370163 -0.019692909  0.709773799
housing  0.024584337  0.024407686 -0.025510417  1.000000000 -0.009600314  0.032135199  0.002518504  0.005124599 -0.042795089
score   -0.007868260  0.003395876 -0.000441125 -0.009600314  1.000000000  0.027196030  0.008718945  0.006942588 -0.001142519
time1    0.009794086 -0.008982730  0.032744120  0.032135199  0.027196030  1.000000000  0.178765716  0.176855988  0.018431889
time2   -0.016180547  0.005449553 -0.020370163  0.002518504  0.008718945  0.178765716  1.000000000  0.990431329 -0.045335553
time3   -0.015427671  0.006591051 -0.019692909  0.005124599  0.006942588  0.176855988  0.990431329  1.000000000 -0.045264921
scr     -0.038603194 -0.020448337  0.709773799 -0.042795089 -0.001142519  0.018431889 -0.045335553 -0.045264921  1.000000000
Pol     -0.017807916  0.012395456  0.001052106  0.020056447  0.031800757  0.030276060  0.040874517  0.038356137  0.005783321
                 Pol
age     -0.017807916
n.child  0.012395456
income   0.001052106
housing  0.020056447
score    0.031800757
time1    0.030276060
time2    0.040874517
time3    0.038356137
scr      0.005783321
Pol      1.000000000
> 
```

Note: The pairs of variables that have a relatively higher correlations are:
1. “income” and “scr”: 0.70977 (significant but not very strong positive relationship)
2. “time2” and “time3”: 0.99 (very strong positive relationship)

Decision: (based on the context)
1. Keep both “income” and “scr” variables since a higher income level does not intuitively imply that the standardized score test will be higher in this context, plus the correlation is not a strong one.
2. Remove column “time2” or “time3” as they represent similar information; in this context, the amount of time each resident spend on Section1 is likely to be similar to Section2.



```
> #We remove 'time3', col(8):
> Res_num <- Res_num[-c(8)]
> head(Res_num)
  age n.child    income housing score time1 time2        scr  Pol
1  62       1  32982.78    0.70 -0.65  0.09  1.05 -1.8909829 0.45
2  31       2 111093.49    0.16  1.19  0.65 -1.19  0.2727746 0.07
3  24       0  42670.39    0.18 -0.03  0.01 -5.50 -1.0275664 0.84
4  35       2 116000.00    0.01  0.12  0.10  0.93  1.8315813 0.57
5  30       0 124267.33    0.36 -0.22  0.49 -1.93 -0.6026588 0.78
6  44       1 116000.00    0.04  0.15  0.60 -0.98  1.2177417 0.50
> 
```