Pre_processing

COLLINS M NJAGI

2025-04-21

```
#libraries

library(readxl)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

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

```
library(tidyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```r
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.3.3
```

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```r
library(pls)
```

```
## Warning: package 'pls' was built under R version 4.3.3
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:corrplot':
##
##     corrplot
```

```
## The following object is masked from 'package:caret':
##
##     R2
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```r
#reading data

# data_GTrends <- read_excel("C:/data_science/DSE6311OM_/week3_Exploratory Data Analysis/googleTrendsMH
# acs_data <- load("C:\\data_science\\DSE6311OM_\\Week4_Pre-processing and Feature Engineering\\ACS_for
# Replace with this code to allow quick access regardless of download location
data_GTrends <- read_excel("~/GitHub/DSE6311OM_SP2025R2_Data-Science-Capstone/Data/googleTrendsMH.xlsx"
    sheet = "googleTrendsMH")
acs_data <- load("~/GitHub/DSE6311OM_SP2025R2_Data-Science-Capstone/Data/ACS_for_MHGoogleTrends.Rdata")

acs_data <- ACS_data
ACS_data <- NULL
```

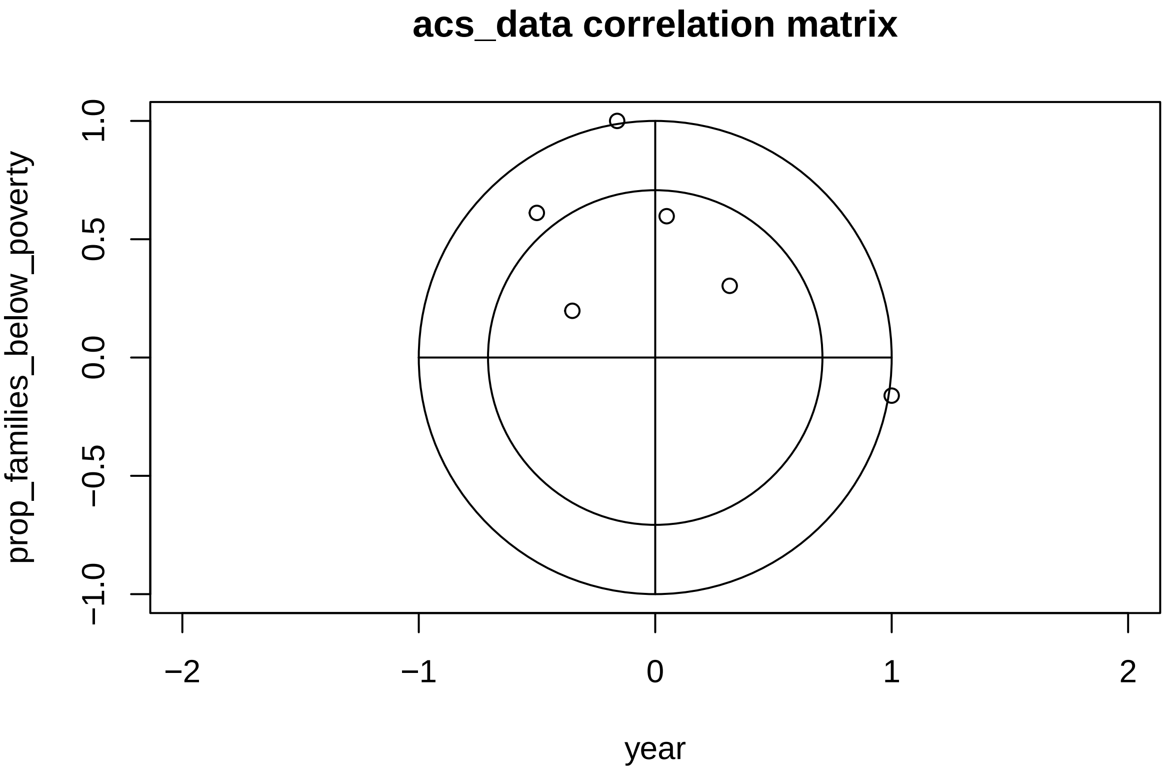
```
##CORRELATION MATRIX FOR acs_data

acs_correlation_matrix <- acs_data %>%
  select_if(is.numeric) %>%
  select(-prop_persons_below_poverty_threshold, -prop_veterans_disability) %>%
  cor()

print(acs_correlation_matrix)
```

```
##                                             year prop_families_below_poverty
## year                                  1.00000000                  -0.1610309
## prop_families_below_poverty          -0.16103094                   1.0000000
## prop_adults_without_health_insurance -0.35051348                   0.1974453
## prop_unemployed_in_labor_force       -0.50071692                   0.6113240
## prop_without_internet_access          0.31496819                   0.3030755
## prop_adult_disability                 0.04834553                   0.5972604
##                                      prop_adults_without_health_insurance
## year                                                           -0.3505135
## prop_families_below_poverty                                     0.1974453
## prop_adults_without_health_insurance                            1.0000000
## prop_unemployed_in_labor_force                                  0.2889701
## prop_without_internet_access                                   -0.1226758
## prop_adult_disability                                           0.1945398
##                                      prop_unemployed_in_labor_force
## year                                                     -0.5007169
## prop_families_below_poverty                               0.6113240
## prop_adults_without_health_insurance                      0.2889701
## prop_unemployed_in_labor_force                            1.0000000
## prop_without_internet_access                             -0.1705119
## prop_adult_disability                                     0.1723363
##                                      prop_without_internet_access
## year                                                    0.3149682
## prop_families_below_poverty                             0.3030755
## prop_adults_without_health_insurance                   -0.1226758
## prop_unemployed_in_labor_force                         -0.1705119
## prop_without_internet_access                            1.0000000
## prop_adult_disability                                   0.3494365
##                                      prop_adult_disability
## year                                            0.04834553
## prop_families_below_poverty                     0.59726036
## prop_adults_without_health_insurance            0.19453980
## prop_unemployed_in_labor_force                  0.17233629
## prop_without_internet_access                    0.34943653
## prop_adult_disability                           1.00000000
```

updated code

```
#presenting correlation matrix in graphic format

acs_correlation_matrix <- acs_data %>%
  select_if(is.numeric) %>%
  select(-prop_persons_below_poverty_threshold, -prop_veterans_disability) %>%
  cor() %>%
  corrplot( diag = F,
           tl.cex = 0.7,
           tl.col = "black",
           main = "acs_data correlation matrix",
           mar = c(0,0,1,0))
```

```
## Warning in plot.window(...): "diag" is not a graphical parameter

## Warning in plot.window(...): "tl.cex" is not a graphical parameter

## Warning in plot.window(...): "tl.col" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "diag" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.cex" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.col" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "diag" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "diag" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter

## Warning in box(...): "diag" is not a graphical parameter

## Warning in box(...): "tl.cex" is not a graphical parameter

## Warning in box(...): "tl.col" is not a graphical parameter

## Warning in title(...): "diag" is not a graphical parameter

## Warning in title(...): "tl.cex" is not a graphical parameter

## Warning in title(...): "tl.col" is not a graphical parameter
```



```
#removing correlated features


acs_data_clean <- acs_data %>%
  select(-prop_persons_below_poverty_threshold, -prop_veterans_disability)


# convert state names into abbreviation to match state in data_GTrends

acs_data_clean$state <- toupper(state.abb[match(tolower(acs_data_clean$state), tolower(state.name))])
```

testing GitHub

```
#data transformations ct variables
#creating response variable => state_mentalhealth_utili = state_psych_care / population_est

#state_mentalhealth_utili <- data_GTrends$state_psych_care / data_GTrends$population_est

data_GTrends <- data_GTrends %>%
  mutate(state_mentalhealth_util = state_psych_care/population_est,
         anxiety_prop = anxiety_ct/ population_est,
         trauma_stress_prop = trauma_stress_ct/population_est,
         adhd_prop = adhd_ct/population_est,
         bipolar_prop = bipolar_ct/population_est,
         depression_prop = depression_ct/population_est)

#data_GTrends <- data_GTrends %>%
  #select(-state_psych_care, -anxiety_ct, -trauma_stress_ct, -adhd_ct, -bipolar_ct, -depression_ct) all
```

```r
#joining both datasets acs_data and data_GTrends

GTrends_acs_joined <- inner_join(data_GTrends, acs_data_clean, by = c("year", "state"))
```

```r
#testing correlation


correlation_matrix <- GTrends_acs_joined %>%
  select_if(is.numeric) %>%
  select(-fips, -population_est,-private_psych_care, -total_util, -outpatient_util, -mean_anxiety, -resi
         -total_util) %>%
  cor()

print(correlation_matrix)
```

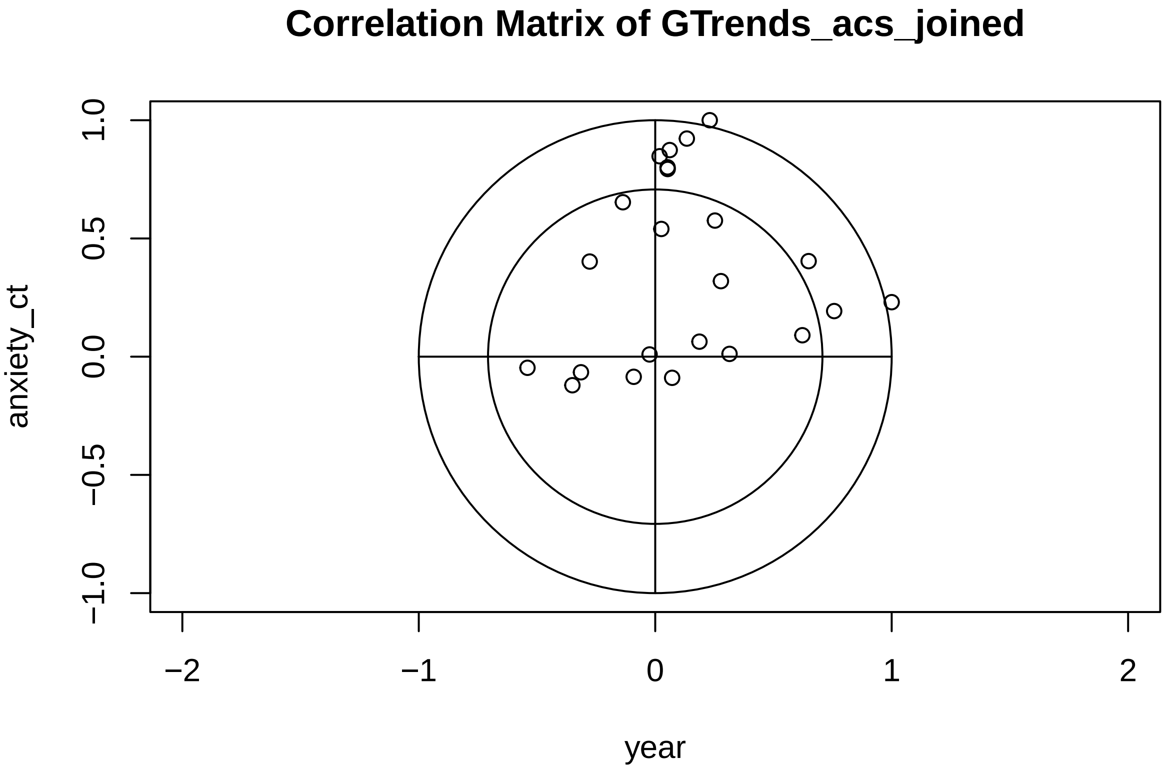
```
##                                             year   anxiety_ct trauma_stress_ct
## year                                  1.00000000  0.230563501       0.13366856
## anxiety_ct                            0.23056350  1.000000000       0.92240079
## trauma_stress_ct                      0.13366856  0.922400795       1.00000000
## adhd_ct                               0.01851770  0.847645702       0.87161036
## bipolar_ct                           -0.13690754  0.653131435       0.75571956
## depression_ct                         0.06120702  0.873780027       0.94087338
## comm_psych_care                       0.05264059  0.793626073       0.89977194
## state_psych_care                      0.05220254  0.800842275       0.90248691
## mean_adhd                             0.75682637  0.192811841       0.08958471
## mean_ptsd                             0.62228218  0.090669189       0.04475684
## mean_bipolar                         -0.09097469 -0.085128361      -0.08423315
## mean_depression                      -0.02390143  0.009319898      -0.02136263
## mean_mental_hospital                  0.27777930  0.319455125       0.28112091
## mean_psychiatrists_near_me            0.18697534  0.063526502       0.09919989
## mean_psychologist_near_me             0.64878930  0.404062943       0.38356349
## anxiety_prop                          0.25256530  0.575638687       0.40794338
## adhd_prop                             0.02582844  0.540119606       0.44884626
## bipolar_prop                         -0.27713846  0.402247684       0.39406527
## prop_families_below_poverty          -0.31411265 -0.065951520      -0.02266406
## prop_adults_without_health_insurance -0.35036488 -0.120820100      -0.08943951
## prop_unemployed_in_labor_force       -0.54031845 -0.047006409       0.07676369
## prop_without_internet_access          0.31423583  0.011777977      -0.03506000
## prop_adult_disability                 0.07154859 -0.089418168      -0.12802032
##                                           adhd_ct  bipolar_ct depression_ct
## year                                  0.018517704 -0.13690754    0.06120702
## anxiety_ct                            0.847645702  0.65313144    0.87378003
## trauma_stress_ct                      0.871610355  0.75571956    0.94087338
## adhd_ct                               1.000000000  0.83440163    0.90823233
## bipolar_ct                            0.834401629  1.00000000    0.88673220
## depression_ct                         0.908232333  0.88673220    1.00000000
## comm_psych_care                       0.874225711  0.87090215    0.95667411
## state_psych_care                      0.884006979  0.87166405    0.95701158
## mean_adhd                            -0.007745775 -0.10866030    0.02253769
## mean_ptsd                            -0.124707857 -0.22821302   -0.08131642
## mean_bipolar                         -0.082850695 -0.03030126   -0.08659302
## mean_depression                      -0.026389005 -0.09361394   -0.02884011
## mean_mental_hospital                  0.220054198  0.21655455    0.28147786
## mean_psychiatrists_near_me            0.086212620  0.06521304    0.09221333
## mean_psychologist_near_me             0.316683082  0.20732437    0.35169600
## anxiety_prop                          0.306023903  0.03211950    0.27306557
## adhd_prop                             0.557691198  0.19368296    0.36224924
## bipolar_prop                          0.458390120  0.36562312    0.36378200
## prop_families_below_poverty           0.091452450  0.21421452    0.06093810
## prop_adults_without_health_insurance  0.001121328  0.24369742    0.03448441
## prop_unemployed_in_labor_force        0.124358517  0.28278587    0.13217179
## prop_without_internet_access          0.010097643 -0.11859483   -0.03027184
## prop_adult_disability                -0.041620397 -0.11618594   -0.11834226
##                                      comm_psych_care state_psych_care
## year                                      0.05264059       0.05220254
## anxiety_ct                                0.79362607       0.80084228
## trauma_stress_ct                          0.89977194       0.90248691
## adhd_ct                                   0.87422571       0.88400698
## bipolar_ct                                0.87090215       0.87166405
## depression_ct                             0.95667411       0.95701158
## comm_psych_care                           1.00000000       0.99936080
## state_psych_care                          0.99936080       1.00000000
## mean_adhd                                 0.01154550       0.01301038
## mean_ptsd                                -0.09505592      -0.09409334
## mean_bipolar                             -0.06243299      -0.06269307
## mean_depression                          -0.04094749      -0.04237320
## mean_mental_hospital                      0.24373032       0.24415647
## mean_psychiatrists_near_me                0.13571311       0.13354197
## mean_psychologist_near_me                 0.36100819       0.35825438
## anxiety_prop                              0.18813746       0.20049138
## adhd_prop                                 0.28982510       0.30527377
## bipolar_prop                              0.30483675       0.31814831
## prop_families_below_poverty               0.06341390       0.06303851
## prop_adults_without_health_insurance      0.02920460       0.02820942
## prop_unemployed_in_labor_force            0.16815934       0.16554652
## prop_without_internet_access             -0.03609294      -0.03484673
## prop_adult_disability                    -0.15530682      -0.14673191
##                                         mean_adhd   mean_ptsd mean_bipolar
## year                                  0.756826372  0.62228218 -0.090974692
## anxiety_ct                            0.192811841  0.09066919 -0.085128361
## trauma_stress_ct                      0.089584712  0.04475684 -0.084233146
## adhd_ct                              -0.007745775 -0.12470786 -0.082850695
## bipolar_ct                           -0.108660303 -0.22821302 -0.030301260
## depression_ct                         0.022537693 -0.08131642 -0.086593022
## comm_psych_care                       0.011545502 -0.09505592 -0.062432992
## state_psych_care                      0.013010379 -0.09409334 -0.062693072
## mean_adhd                             1.000000000  0.42495384  0.179510680
## mean_ptsd                             0.424953840  1.00000000  0.193509244
## mean_bipolar                          0.179510680  0.19350924  1.000000000
## mean_depression                      -0.245750075  0.41128942  0.308755245
## mean_mental_hospital                  0.287677009  0.09702821  0.232486981
## mean_psychiatrists_near_me            0.042769431  0.05674090 -0.005280538
## mean_psychologist_near_me             0.415735545  0.23433255 -0.080183845
## anxiety_prop                          0.222753634  0.30520691 -0.005956554
## adhd_prop                             0.028590323  0.09085592 -0.010212770
## bipolar_prop                         -0.159049076 -0.04663275  0.157398435
## prop_families_below_poverty          -0.208577621 -0.20391856  0.293106346
## prop_adults_without_health_insurance -0.186412427 -0.24473889  0.233057761
## prop_unemployed_in_labor_force       -0.327758496 -0.43653037  0.157300589
## prop_without_internet_access         -0.126520915  0.33393361 -0.090016482
## prop_adult_disability                 0.109982033  0.10629585  0.222236769
##                                      mean_depression mean_mental_hospital
## year                                    -0.023901425           0.27777930
## anxiety_ct                               0.009319898           0.31945513
## trauma_stress_ct                        -0.021362629           0.28112091
## adhd_ct                                 -0.026389005           0.22005420
## bipolar_ct                              -0.093613944           0.21655455
## depression_ct                           -0.028840113           0.28147786
## comm_psych_care                         -0.040947486           0.24373032
## state_psych_care                        -0.042373199           0.24415647
## mean_adhd                               -0.245750075           0.28767701
## mean_ptsd                                0.411289416           0.09702821
## mean_bipolar                             0.308755245           0.23248698
## mean_depression                          1.000000000          -0.10548867
## mean_mental_hospital                    -0.105488666           1.00000000
## mean_psychiatrists_near_me               0.001374564           0.15614239
## mean_psychologist_near_me               -0.098056483           0.41633384
## anxiety_prop                             0.050429764           0.02664347
## adhd_prop                                0.069487449          -0.06288825
## bipolar_prop                             0.026384149          -0.09485722
## prop_families_below_poverty             -0.077146712           0.21535926
## prop_adults_without_health_insurance    -0.062380502          -0.02688604
## prop_unemployed_in_labor_force          -0.348426242           0.10886182
## prop_without_internet_access             0.385215253           0.07508085
## prop_adult_disability                   -0.081676556           0.16483923
##                                      mean_psychiatrists_near_me
## year                                                0.186975337
## anxiety_ct                                          0.063526502
## trauma_stress_ct                                    0.099199887
## adhd_ct                                             0.086212620
## bipolar_ct                                          0.065213036
## depression_ct                                       0.092213328
## comm_psych_care                                     0.135713106
## state_psych_care                                    0.133541968
## mean_adhd                                           0.042769431
## mean_ptsd                                           0.056740904
## mean_bipolar                                       -0.005280538
## mean_depression                                     0.001374564
## mean_mental_hospital                                0.156142388
## mean_psychiatrists_near_me                          1.000000000
## mean_psychologist_near_me                           0.466711912
## anxiety_prop                                       -0.104990533
## adhd_prop                                          -0.105489672
## bipolar_prop                                       -0.156142069
## prop_families_below_poverty                        -0.185544042
## prop_adults_without_health_insurance               -0.257450224
## prop_unemployed_in_labor_force                     -0.020698183
## prop_without_internet_access                        0.051130358
## prop_adult_disability                              -0.239770625
##                                      mean_psychologist_near_me anxiety_prop
## year                                                0.64878930  0.252565296
## anxiety_ct                                          0.40406294  0.575638687
## trauma_stress_ct                                    0.38356349  0.407943378
## adhd_ct                                             0.31668308  0.306023903
## bipolar_ct                                          0.20732437  0.032119498
## depression_ct                                       0.35169600  0.273065574
## comm_psych_care                                     0.36100819  0.188137462
## state_psych_care                                    0.35825438  0.200491380
## mean_adhd                                           0.41573555  0.222753634
## mean_ptsd                                           0.23433255  0.305206913
## mean_bipolar                                       -0.08018385 -0.005956554
## mean_depression                                    -0.09805648  0.050429764
## mean_mental_hospital                                0.41633384  0.026643466
## mean_psychiatrists_near_me                          0.46671191 -0.104990533
## mean_psychologist_near_me                           1.00000000  0.018713136
## anxiety_prop                                        0.01871314  1.000000000
## adhd_prop                                          -0.02192663  0.772593545
## bipolar_prop                                       -0.20102389  0.592973858
## prop_families_below_poverty                        -0.16397365 -0.139411004
## prop_adults_without_health_insurance               -0.20618180 -0.202330161
## prop_unemployed_in_labor_force                     -0.18536934 -0.244392365
## prop_without_internet_access                        0.15990322  0.090420463
## prop_adult_disability                              -0.08569762  0.099264075
##                                        adhd_prop bipolar_prop
## year                                  0.02582844  -0.27713846
## anxiety_ct                            0.54011961   0.40224768
## trauma_stress_ct                      0.44884626   0.39406527
## adhd_ct                               0.55769120   0.45839012
## bipolar_ct                            0.19368296   0.36562312
## depression_ct                         0.36224924   0.36378200
## comm_psych_care                       0.28982510   0.30483675
## state_psych_care                      0.30527377   0.31814831
## mean_adhd                             0.02859032  -0.15904908
## mean_ptsd                             0.09085592  -0.04663275
## mean_bipolar                         -0.01021277   0.15739843
## mean_depression                       0.06948745   0.02638415
## mean_mental_hospital                 -0.06288825  -0.09485722
## mean_psychiatrists_near_me           -0.10548967  -0.15614207
## mean_psychologist_near_me            -0.02192663  -0.20102389
## anxiety_prop                          0.77259354   0.59297386
## adhd_prop                             1.00000000   0.73676449
## bipolar_prop                          0.73676449   1.00000000
## prop_families_below_poverty           0.06474605   0.24288704
## prop_adults_without_health_insurance -0.10333794   0.15947980
## prop_unemployed_in_labor_force       -0.06381305   0.17824936
## prop_without_internet_access          0.10675502  -0.09079816
## prop_adult_disability                 0.20587109   0.24830497
##                                      prop_families_below_poverty
## year                                                 -0.31411265
## anxiety_ct                                           -0.06595152
## trauma_stress_ct                                     -0.02266406
## adhd_ct                                               0.09145245
## bipolar_ct                                            0.21421452
## depression_ct                                         0.06093810
## comm_psych_care                                       0.06341390
## state_psych_care                                      0.06303851
## mean_adhd                                            -0.20857762
## mean_ptsd                                            -0.20391856
## mean_bipolar                                          0.29310635
## mean_depression                                      -0.07714671
## mean_mental_hospital                                  0.21535926
## mean_psychiatrists_near_me                           -0.18554404
## mean_psychologist_near_me                            -0.16397365
## anxiety_prop                                         -0.13941100
## adhd_prop                                             0.06474605
## bipolar_prop                                          0.24288704
## prop_families_below_poverty                           1.00000000
## prop_adults_without_health_insurance                  0.60329043
## prop_unemployed_in_labor_force                        0.52364772
## prop_without_internet_access                          0.12312374
## prop_adult_disability                                 0.65543780
##                                      prop_adults_without_health_insurance
## year                                                         -0.350364883
## anxiety_ct                                                   -0.120820100
## trauma_stress_ct                                             -0.089439512
## adhd_ct                                                       0.001121328
## bipolar_ct                                                    0.243697423
## depression_ct                                                 0.034484408
## comm_psych_care                                               0.029204600
## state_psych_care                                              0.028209419
## mean_adhd                                                    -0.186412427
## mean_ptsd                                                    -0.244738889
## mean_bipolar                                                  0.233057761
## mean_depression                                              -0.062380502
## mean_mental_hospital                                         -0.026886042
## mean_psychiatrists_near_me                                   -0.257450224
## mean_psychologist_near_me                                    -0.206181798
## anxiety_prop                                                 -0.202330161
## adhd_prop                                                    -0.103337943
## bipolar_prop                                                  0.159479797
## prop_families_below_poverty                                   0.603290434
## prop_adults_without_health_insurance                          1.000000000
## prop_unemployed_in_labor_force                                0.409465887
## prop_without_internet_access                                 -0.106556672
## prop_adult_disability                                         0.289928013
##                                      prop_unemployed_in_labor_force
## year                                                    -0.54031845
## anxiety_ct                                              -0.04700641
## trauma_stress_ct                                         0.07676369
## adhd_ct                                                  0.12435852
## bipolar_ct                                               0.28278587
## depression_ct                                            0.13217179
## comm_psych_care                                          0.16815934
## state_psych_care                                         0.16554652
## mean_adhd                                               -0.32775850
## mean_ptsd                                               -0.43653037
## mean_bipolar                                             0.15730059
## mean_depression                                         -0.34842624
## mean_mental_hospital                                     0.10886182
## mean_psychiatrists_near_me                              -0.02069818
## mean_psychologist_near_me                               -0.18536934
## anxiety_prop                                            -0.24439237
## adhd_prop                                               -0.06381305
## bipolar_prop                                             0.17824936
## prop_families_below_poverty                              0.52364772
## prop_adults_without_health_insurance                     0.40946589
## prop_unemployed_in_labor_force                           1.00000000
## prop_without_internet_access                            -0.34452758
## prop_adult_disability                                    0.06756309
##                                      prop_without_internet_access
## year                                                   0.31423583
## anxiety_ct                                             0.01177798
## trauma_stress_ct                                      -0.03506000
## adhd_ct                                                0.01009764
## bipolar_ct                                            -0.11859483
## depression_ct                                         -0.03027184
## comm_psych_care                                       -0.03609294
## state_psych_care                                      -0.03484673
## mean_adhd                                             -0.12652092
## mean_ptsd                                              0.33393361
## mean_bipolar                                          -0.09001648
## mean_depression                                        0.38521525
## mean_mental_hospital                                   0.07508085
## mean_psychiatrists_near_me                             0.05113036
## mean_psychologist_near_me                              0.15990322
## anxiety_prop                                           0.09042046
## adhd_prop                                              0.10675502
## bipolar_prop                                          -0.09079816
## prop_families_below_poverty                            0.12312374
## prop_adults_without_health_insurance                  -0.10655667
## prop_unemployed_in_labor_force                        -0.34452758
## prop_without_internet_access                           1.00000000
## prop_adult_disability                                  0.30396009
##                                      prop_adult_disability
## year                                            0.07154859
## anxiety_ct                                     -0.08941817
## trauma_stress_ct                               -0.12802032
## adhd_ct                                        -0.04162040
## bipolar_ct                                     -0.11618594
## depression_ct                                  -0.11834226
## comm_psych_care                                -0.15530682
## state_psych_care                               -0.14673191
## mean_adhd                                       0.10998203
## mean_ptsd                                       0.10629585
## mean_bipolar                                    0.22223677
## mean_depression                                -0.08167656
## mean_mental_hospital                            0.16483923
## mean_psychiatrists_near_me                     -0.23977062
## mean_psychologist_near_me                      -0.08569762
## anxiety_prop                                    0.09926407
## adhd_prop                                       0.20587109
## bipolar_prop                                    0.24830497
## prop_families_below_poverty                     0.65543780
## prop_adults_without_health_insurance           0.28992801
## prop_unemployed_in_labor_force                 0.06756309
## prop_without_internet_access                   0.30396009
## prop_adult_disability                          1.00000000
```

high correlation variables

1. private, reside and comm_psych_care, 2.inpatient_util vs outpatient_util (i already have state_mentalhealth_util) 3.mean_therapist near_me vs mean_psychiatrist and mean_psychologist 4.mean_alltrend vs mean_adhd, mean_ptsd, mean_anxiety, mean_mentalhospital.
2. mean_anxiety vs year, mean_adhd & ptsd 6.outpatient_util vs total_util, adhd, bipolar & depression 7.total_util 8.depression prob vs adhd. ptsd, bipolar and trauma_stress_prop 9.trauma_stress_prop vs adhd, anxiety_prop and state_mentalhealth_util 10.state_mentalhealth_util vs adhd, ptsd, bipolar

```
#correlation matrix

GTrends_acs_joined %>%
  select_if(is.numeric) %>%
  select(-fips, -population_est,-private_psych_care, -total_util, -outpatient_util, -mean_anxiety, -resi
         -total_util) %>%
  cor() %>%

corrplot(diag = F,
         tl.cex = 0.7,
         tl.col = "black",
         main = "Correlation Matrix of GTrends_acs_joined",
         mar = c(0, 0, 1, 0))
```

```
## Warning in plot.window(...): "diag" is not a graphical parameter

## Warning in plot.window(...): "tl.cex" is not a graphical parameter

## Warning in plot.window(...): "tl.col" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "diag" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.cex" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "tl.col" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "diag" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "diag" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.cex" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "tl.col" is not a
## graphical parameter

## Warning in box(...): "diag" is not a graphical parameter

## Warning in box(...): "tl.cex" is not a graphical parameter

## Warning in box(...): "tl.col" is not a graphical parameter

## Warning in title(...): "diag" is not a graphical parameter

## Warning in title(...): "tl.cex" is not a graphical parameter

## Warning in title(...): "tl.col" is not a graphical parameter
```



```
#data split: train and test dataset

clean_GTrends_acs_joined <- GTrends_acs_joined %>%
  select(-fips, -population_est,-private_psych_care, -total_util, -outpatient_util, -region, -mean_anxi
test_n <- (1/sqrt(19))*nrow(clean_GTrends_acs_joined)
  test_prop <- round((1/sqrt(19))*nrow(clean_GTrends_acs_joined)/nrow(clean_GTrends_acs_joined), 2)
  train_prop <- 1-test_prop

  print(paste0("the ideal split ratio is", train_prop, ":", test_prop, "training:testing"))
```

```
## [1] "the ideal split ratio is0.77:0.23training:testing"
```

```
# Show the dimensions of the dataframe and the column names.
dim(clean_GTrends_acs_joined)
```

```
## [1] 433  25
```

```
names(clean_GTrends_acs_joined)
```

```
##  [1] "year"
##  [2] "state"
##  [3] "anxiety_ct"
##  [4] "trauma_stress_ct"
##  [5] "adhd_ct"
##  [6] "bipolar_ct"
##  [7] "depression_ct"
##  [8] "comm_psych_care"
##  [9] "state_psych_care"
## [10] "mean_adhd"
## [11] "mean_ptsd"
## [12] "mean_bipolar"
## [13] "mean_depression"
## [14] "mean_mental_hospital"
## [15] "mean_psychiatrists_near_me"
## [16] "mean_psychologist_near_me"
## [17] "state_mentalhealth_util"
## [18] "anxiety_prop"
## [19] "adhd_prop"
## [20] "bipolar_prop"
## [21] "prop_families_below_poverty"
## [22] "prop_adults_without_health_insurance"
## [23] "prop_unemployed_in_labor_force"
## [24] "prop_without_internet_access"
## [25] "prop_adult_disability"
```

```
#write the merged dataframe to a CSV file with a time stamp in the  name.
# This way we don't overwrite the file in case someone else is working on the file.
# TimeStamp <- format(Sys.time(), "%Y%m%d_%H%M%S")
# file_name <- paste("~/GitHub/DSE6311OM_SP2025R2_Data-Science-Capstone/Data/clean_GTrends_acs_joined_"
# write.csv(clean_GTrends_acs_joined, file_name, row.names = FALSE)
```

```
train <- createDataPartition(clean_GTrends_acs_joined$state_mentalhealth_util,
                             p = 0.77,
                             list = FALSE,
                             times = 1)
```

```
GTrend_training_set <- clean_GTrends_acs_joined[train, ]


test_set <- clean_GTrends_acs_joined[-train, ]

dim(GTrend_training_set)
```

```
## [1] 336  25
```

```
dim(test_set)
```

```
## [1] 97 25
```

TARGET ENCODING OF STATE BY Njagi

```
unique(clean_GTrends_acs_joined$state)
```

```
##  [1] "AL" "AZ" "AR" "CA" "CO" "CT" "DE" "FL" "HI" "ID" "IL" "IN" "IA" "KS" "KY"
## [16] "LA" "MA" "MS" "MO" "MT" "NE" "NV" "NJ" "NM" "NY" "NC" "ND" "OH" "OK" "OR"
## [31] "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VT" "VA" "WA" "WI" "WY" "MN" "MI" "AK"
## [46] "GA"
```

```
is.factor(clean_GTrends_acs_joined$state) #checking whether region is a factor = false
```

```
## [1] FALSE
```

```
GTrend_training_set$state <- factor(GTrend_training_set$state)

     class(GTrend_training_set$state)
```

```
## [1] "factor"
```

```
     levels(GTrend_training_set$state)
```

```
##  [1] "AK" "AL" "AR" "AZ" "CA" "CO" "CT" "DE" "FL" "GA" "HI" "IA" "ID" "IL" "IN"
## [16] "KS" "KY" "LA" "MA" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE" "NJ" "NM" "NV"
## [31] "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VA" "VT" "WA" "WI"
## [46] "WY"
```

```
# we are going to apply target encoding (state_mentalhealth_util). To avoid overfitting we are going to
#smoothed version of target encoding

main_mean <- mean( GTrend_training_set$state_mentalhealth_util)

  smoothing_factor <- 10
  #calculating the smoothed state means from the training set
   state_encoded_by_smoothedmean <- GTrend_training_set %>%
     group_by(state) %>%
     summarise(state_encoded = (mean(state_mentalhealth_util) * n() + main_mean * smoothing_factor) / (

   #merging the smoothed encoded state means with the training set

          GTrend_training_set_f <- GTrend_training_set %>%
            left_join(state_encoded_by_smoothedmean, by = "state") %>%
            select(-state)


    #merging smoothed encoded state means with the test_set



              test_set$state <- factor(test_set$state)

                 test_set_f <- test_set%>%
                   left_join(state_encoded_by_smoothedmean, by = "state") %>%
                   select(-state)
```

```
dim(test_set_f)
```

```
## [1] 97 25
```

```
#center and scale

test_set_f[, c(-10)] <- scale(test_set_f[, c(-10)],
                              center = apply(GTrend_training_set_f[, c(-10)], 2, mean),
                              scale = apply(GTrend_training_set_f[, c(-10)], 2, sd))

#(-10) is the state_mentalhealth_util, i want to exclude it from center and scale since its already a p


GTrend_training_set_f[, -10] <- scale(GTrend_training_set_f[, -10])

head(GTrend_training_set_f)
```

```
## # A tibble: 6 x 25
##    year anxiety_ct trauma_stress_ct adhd_ct bipolar_ct depression_ct
##   <dbl>      <dbl>            <dbl>   <dbl>      <dbl>         <dbl>
## 1 -1.60     -0.348           -0.507   0.128     0.0283        -0.150
## 2 -1.60      0.570            0.734   1.55      1.64           0.366
## 3 -1.60     -0.610           -0.557  -0.430    -0.199         -0.468
## 4 -1.60      1.46             2.84    2.45      4.57           3.80
## 5 -1.60     -0.255            0.135  -0.235     0.321         -0.164
## 6 -1.60     -0.409           -0.201  -0.241    -0.101         -0.344
## # i 19 more variables: comm_psych_care <dbl>, state_psych_care <dbl>,
## #   mean_adhd <dbl>, mean_ptsd <dbl>, mean_bipolar <dbl>,
## #   mean_depression <dbl>, mean_mental_hospital <dbl>,
## #   mean_psychiatrists_near_me <dbl>, mean_psychologist_near_me <dbl>,
## #   state_mentalhealth_util <dbl>, anxiety_prop <dbl>, adhd_prop <dbl>,
## #   bipolar_prop <dbl>, prop_families_below_poverty <dbl>,
## #   prop_adults_without_health_insurance <dbl>, ...
```

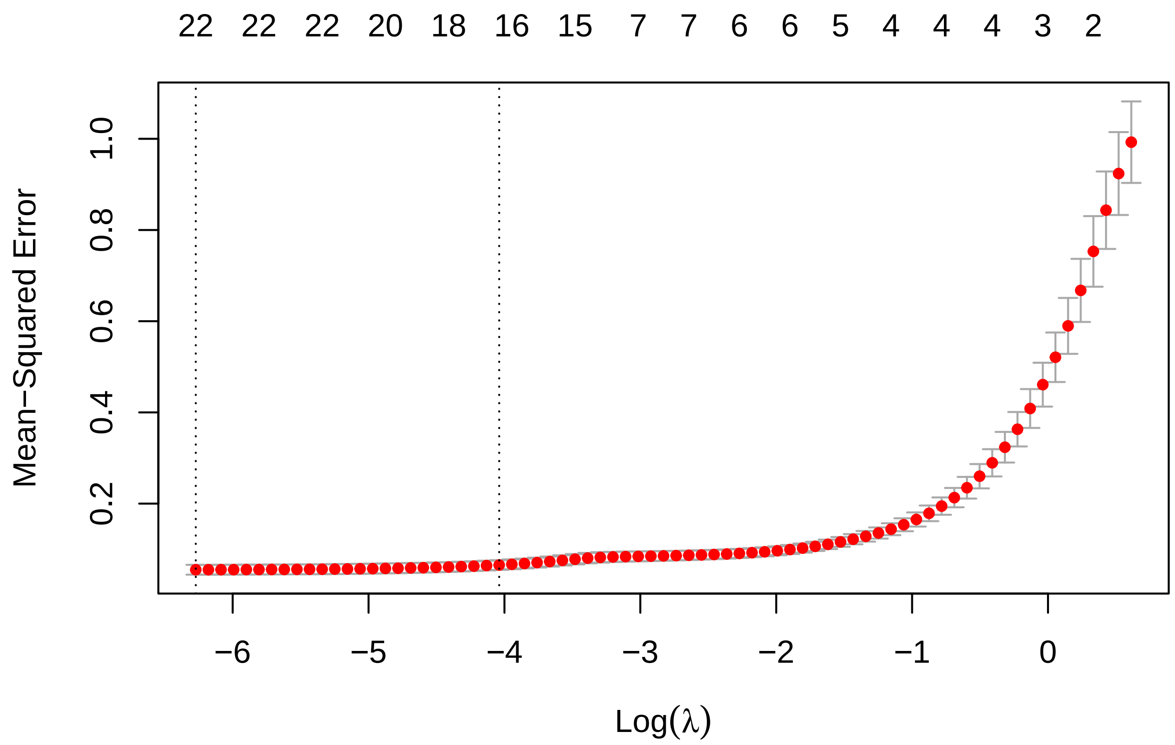
```
#generating codebook

library(tibble)

codebook <- tibble(
  variable = names(clean_GTrends_acs_joined),
  class = sapply(clean_GTrends_acs_joined, class),
  "Number of Missing Values" = sapply(clean_GTrends_acs_joined, function(x) sum(is.na(x))),
  "Number of Unique Values" = sapply(clean_GTrends_acs_joined, function(x) length(unique(x)))
)

print(codebook)
```

```
## # A tibble: 25 x 4
##    variable         class     `Number of Missing Values` Number of Unique Valu~1
##    <chr>            <chr>                          <int>                   <int>
##  1 year             numeric                            0                      10
##  2 state            character                          0                      46
##  3 anxiety_ct       numeric                            0                     433
##  4 trauma_stress_ct numeric                            0                     431
##  5 adhd_ct          numeric                            0                     423
##  6 bipolar_ct       numeric                            0                     430
##  7 depression_ct    numeric                            0                     432
##  8 comm_psych_care  numeric                            0                     432
##  9 state_psych_care numeric                            0                     433
## 10 mean_adhd        numeric                            0                     205
## # i 15 more rows
## # i abbreviated name: 1: `Number of Unique Values`
```

INITIAL MODELS BY Njagi

1. LINEAR REGRESSION (ELASTIC NET REGULARIZATION)

```
# DEVELOPING THE MODEL (LR. ENR)

x <- model.matrix(state_mentalhealth_util ~ ., data = GTrend_training_set_f, intercept = FALSE)

y <- GTrend_training_set_f$state_mentalhealth_util

    #Performing cross_validation to find the best lambda

   set.seed(123) # for consistent and replicable results

   cv_model <- cv.glmnet(x, y, alpha = 0.5, family = "gaussian", nfolds = 5)

   plot(cv_model) #plotting cross-validation curve
```



```
#getting the best/ optimal lambda

best_lambda <- cv_model$lambda.min
best_lambda_1se <- cv_model$lambda.1se


  #developing the model using the best lambda

model_min <- glmnet(x, y, alpha = 0.5, lambda = best_lambda, family = "gaussian")
model_lambda_1se <- glmnet(x, y, alpha = 0.5, lambda = best_lambda_1se, family = "gaussian")


#preparing the test set into matrx

x_test <- model.matrix(state_mentalhealth_util ~ ., data = test_set_f, intercept = FALSE)
y_test <- test_set_f$state_mentalhealth_util

#ensure x and x_test have the same number of columns. its a good practise after using model.matrix

common_columns <- intersect(colnames(x), colnames(x_test))
x <- x[, common_columns]
x_test <- x_test[, common_columns]

# use test set to make predictions, use lambda min and lambda_1se

y_pred_min <- predict(model_min, newx = x_test)
y_pred_1se <- predict(model_lambda_1se, newx = x_test)

#calculate the mean squared error
```

```
mse_min <- mean((y_test - y_pred_min)^2)
mse_1se <- mean((y_test - y_pred_1se)^2)


print(paste("MSE (MIN):", mse_min))
```

```
## [1] "MSE (MIN): 0.061937152516409"
```

```
print(paste("MSE (1SE):", mse_1se))
```

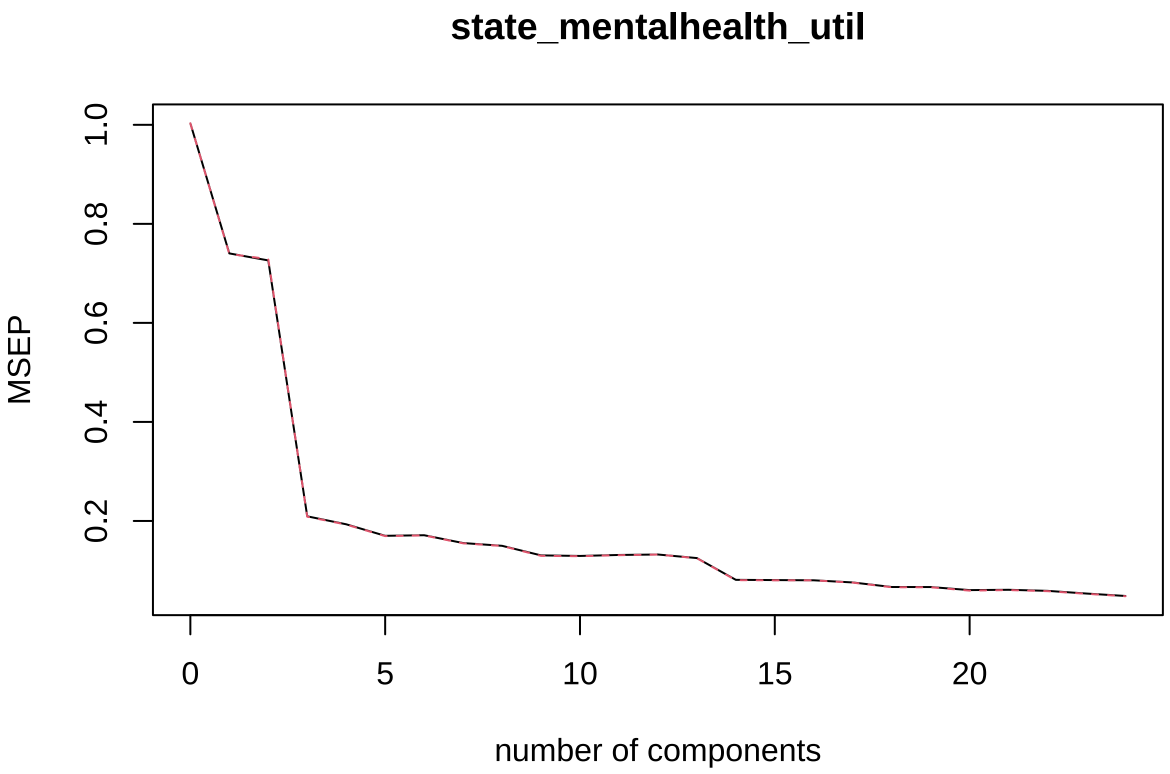
```
## [1] "MSE (1SE): 0.0833464708117956"
```

Principle Component Regression (PCR)

```
pcr_m_selected <- 1

# Get the PCR fit for the training data set
pcr_fit <- pcr(state_mentalhealth_util ~ ., data =GTrend_training_set_f ,
               scale=TRUE, validation="CV")

# Show the summary of the PCR fit.
summary(pcr_fit)
```

```
## Data:    X dimension: 336 24
##  Y dimension: 336 1
## Fit method: svdpc
## Number of components considered: 24
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.001   0.8604   0.8521   0.4576   0.4396   0.4123   0.4137
## adjCV        1.001   0.8598   0.8537   0.4567   0.4389   0.4116   0.4132
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.3942   0.3870   0.3612    0.3597    0.3624    0.3636    0.3536
## adjCV   0.3935   0.3865   0.3599    0.3589    0.3617    0.3635    0.3532
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## CV       0.2848    0.2839    0.2832    0.2753    0.2580    0.2580    0.2457
## adjCV    0.2836    0.2830    0.2830    0.2746    0.2571    0.2571    0.2434
##        21 comps  22 comps  23 comps  24 comps
## CV       0.2469    0.2426    0.2311    0.2202
## adjCV    0.2458    0.2412    0.2299    0.2190
##
## TRAINING: % variance explained
##                          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                          30.12    46.20    59.52    69.10    75.86    80.92
## state_mentalhealth_util    27.18    30.34    79.83    81.47    83.80    83.80
##                          7 comps  8 comps  9 comps  10 comps  11 comps
## X                          85.25    88.21    90.45     92.34     93.72
## state_mentalhealth_util    85.34    85.92    87.89     88.15     88.18
##                          12 comps  13 comps  14 comps  15 comps  16 comps
## X                           94.99    96.11    97.09    97.88    98.44
## state_mentalhealth_util     88.27    89.28    92.79    92.82    92.82
##                          17 comps  18 comps  19 comps  20 comps  21 comps
## X                           98.92    99.29    99.52    99.70    99.87
## state_mentalhealth_util     93.37    94.23    94.31    94.83    94.95
##                          22 comps  23 comps  24 comps
## X                           99.95   100.00   100.00
## state_mentalhealth_util     95.23    95.68    96.06
```

```
# Show the validation plot.
validationplot(pcr_fit, val.type="MSEP")
```



```
# Get the predictions
pcr_preds <- predict(pcr_fit, data=test_set, ncomp=pcr_m_selected)

# Store and print the pcr mean square error for M_selected.
pcr_mse <- mean((pcr_preds-test_set$state_mentalhealth_util)^2)
```

```
## Warning in pcr_preds - test_set$state_mentalhealth_util: longer object length
## is not a multiple of shorter object length
```

```
paste("PCR MSE for M Selected:",pcr_m_selected,"is", pcr_mse)
```

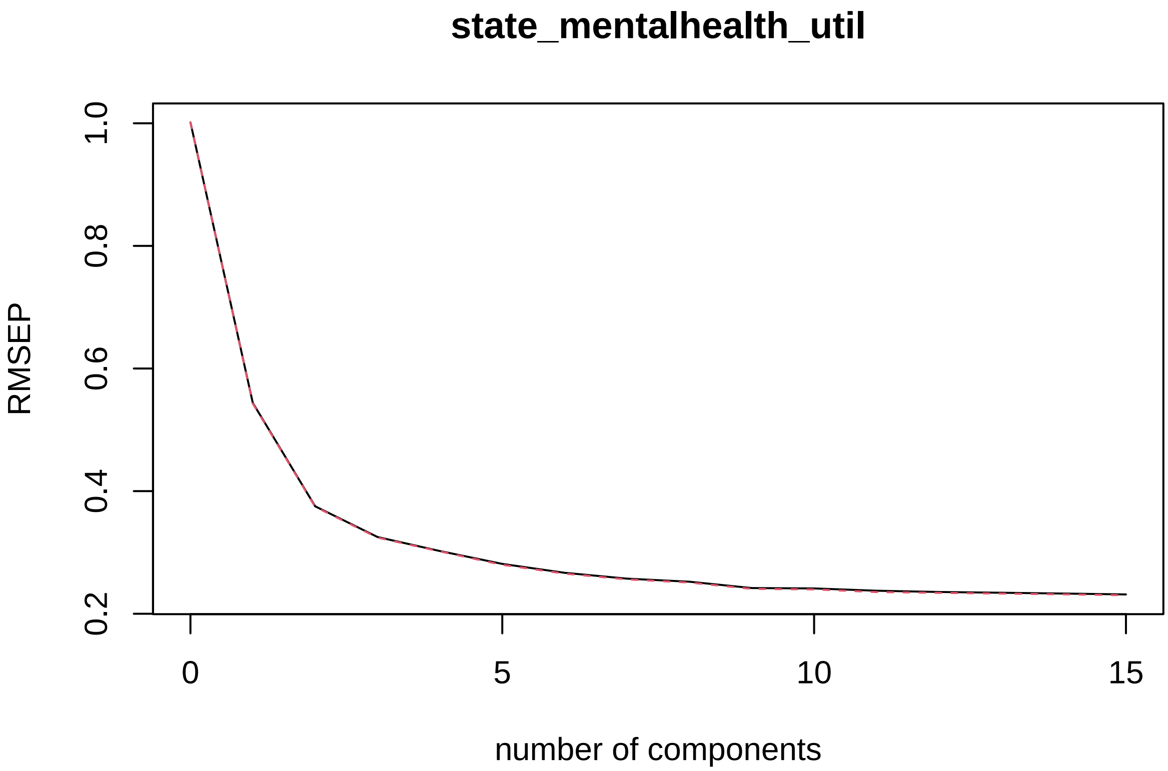
```
## [1] "PCR MSE for M Selected: 1 is 0.272348731197608"
```

Partial Least Squares Regression (PLSR)

```r
# Set the PLS M selected value.
plsr_M_selected <- 15

# Get the PCR fit for the training data set
plsr_fit <- plsr(state_mentalhealth_util ~ ., data=GTrend_training_set_f ,
                 scale=TRUE, validation="CV", ncomp=plsr_M_selected)

# print the summary of the partial least square regression fit.
summary(plsr_fit)
```

```
## Data:    X dimension: 336 24
##  Y dimension: 336 1
## Fit method: kernelpls
## Number of components considered: 15
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.001   0.5440   0.3753   0.3252   0.3023   0.2813   0.2668
## adjCV        1.001   0.5428   0.3744   0.3241   0.3016   0.2798   0.2655
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.2573   0.2523   0.2419    0.2413    0.2376    0.2355    0.2342
## adjCV   0.2560   0.2509   0.2406    0.2397    0.2354    0.2341    0.2329
##        14 comps  15 comps
## CV       0.2330    0.2315
## adjCV    0.2317    0.2302
##
## TRAINING: % variance explained
##                           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                           25.32     43.2    50.31    61.03    68.44    74.50
## state_mentalhealth_util    72.14     86.8    90.43    92.13    93.44    94.11
##                           7 comps  8 comps  9 comps  10 comps  11 comps
## X                           79.72    83.04    84.44     86.64     89.01
## state_mentalhealth_util    94.50    94.79    95.18     95.31     95.42
##                           12 comps  13 comps  14 comps  15 comps
## X                            92.25     93.67     94.85     95.40
## state_mentalhealth_util     95.51     95.58     95.62     95.67
```

```r
# Show the validation plot
validationplot(plsr_fit)
```



```
# Get the predictions
plsr_preds <- predict(plsr_fit, data=test_set, ncomp=plsr_M_selected)

# Store and print the MSE value for the PLSR
plsr_mse <- mean((plsr_preds-test_set$state_mentalhealth_util)^2)
```

```
## Warning in plsr_preds - test_set$state_mentalhealth_util: longer object length
## is not a multiple of shorter object length
```

```
paste("PLSR MSE for M Selected:",plsr_M_selected,"is", plsr_mse)
```

```
## [1] "PLSR MSE for M Selected: 15 is 0.954239539174353"
```