PCA Prediction

1. Load dataset with imputed missing values

```
library(tidyverse)
library(gridExtra)
library(gplots)
library(glmnet)
library(caret)

dir = getwd()
data_dir <- paste('df_imp.RData', sep = '')
load(data_dir)

response_vars <- c('TARGET_deathRate')
y <- df_imp3 %>% select(response_vars)
y <- unname(unlist(y))


do_not_include <- c('ID', 'TARGET_deathRate', 'incidenceRate', 'popEst2015_log', 'medIncome', 'avgAnnCou
do_not_include <- c('ID', 'TARGET_deathRate')

df <- df_imp3 %>% select(- do_not_include)



vars_1 <- colnames(df)
df <- data.frame(sapply(df, as.numeric))
```

2. Looking at PCA

```
S <- cov(df)
eig <- eigen(S)
eig_vals <- eig$values
eig_vecs <- eig$vectors


cum_var_explained <- cumsum(eig_vals/(sum(eig_vals)))

cum_var_explained
```

```
##  [1] 0.9999667 0.9999848 0.9999902 0.9999936 0.9999957 0.9999973 0.9999979
```

```
##  [8] 0.9999984 0.9999988 0.9999991 0.9999993 0.9999995 0.9999996 0.9999997
## [15] 0.9999997 0.9999998 0.9999998 0.9999998 0.9999999 0.9999999 0.9999999
## [22] 0.9999999 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [29] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

```
prcomp(df)
```

```
## Standard deviations (1, .., p=33):
##  [1] 1.210862e+04 5.160938e+01 2.816341e+01 2.222428e+01 1.745490e+01
##  [6] 1.516215e+01 9.634074e+00 8.913984e+00 7.785428e+00 5.989203e+00
## [11] 5.759091e+00 5.075988e+00 3.809827e+00 3.562453e+00 2.850433e+00
## [16] 2.540515e+00 2.451574e+00 2.064991e+00 2.051304e+00 1.990399e+00
## [21] 1.743992e+00 1.554702e+00 1.425568e+00 1.256612e+00 1.226344e+00
## [26] 1.195367e+00 1.036691e+00 9.267616e-01 6.373798e-01 5.130452e-01
## [31] 2.628119e-01 1.088250e-01 9.137068e-02
##
## Rotation (n x k) = (33 x 33):
##                                          PC1           PC2           PC3
## incidenceRate                   5.350731e-06  9.970764e-01  0.0214996082
## medIncome                       9.999988e-01 -2.355195e-05  0.0003694344
## povertyPercent                 -4.187646e-04  6.799470e-04  0.0083969581
## MedianAge                      -5.063680e-05  2.047124e-04  0.0021448650
## MedianAgeMale                  -3.930669e-05 -1.272339e-03  0.0017932275
## MedianAgeFemale                -6.724856e-05  5.216163e-04  0.0034454542
## AvgHouseholdSize                3.162270e-06 -7.379285e-04  0.0002191423
## PercentMarried                  2.011648e-04 -1.436542e-02 -0.0068850117
## PctNoHS18_24                   -1.912063e-04 -2.876606e-02  0.0148101163
## PctHS18_24                     -1.423392e-04  4.313968e-03  0.0061218059
## PctSomeCol18_24                 1.481195e-04  2.007625e-02 -0.0200666690
## PctHS25_Over                   -2.762837e-04  1.828695e-02 -0.0136355767
## PctBachDeg25_Over               3.156582e-04 -4.413517e-03  0.0008186604
## PctEmployed16_Over              4.294907e-04  1.424553e-03 -0.0519593635
## PctUnemployed16_Over           -1.282863e-04  6.924676e-03  0.0071440127
## PctPrivateCoverage              6.352764e-04  2.187217e-02 -0.0278658852
## PctPrivateCoverageAlone         5.211204e-04  2.385125e-02 -0.7185256777
## PctEmpPrivCoverage              5.823078e-04  2.894801e-02 -0.0269039707
## PctPublicCoverage              -4.908479e-04  8.409242e-03  0.0166580613
## PctPublicCoverageAlone         -3.643485e-04  5.605583e-03  0.0130754171
## PctWhite                        2.189410e-04 -5.840022e-03 -0.0288874453
## PctBlack                       -3.193394e-04  3.570700e-02  0.0134991696
## PctMarriedHouseholds            2.415535e-04 -1.978022e-02 -0.0042781274
## BirthRate                      -2.004765e-06 -4.484274e-03 -0.0011859674
## Imputed_PctEmployed16_Over      4.717393e-05 -1.607625e-03  0.0343920294
## Imputed_PctPrivateCoverageAlone 1.338236e-04 -4.112146e-03  0.6894545862
## avgAnnCount_log                 4.126025e-05  7.905199e-03  0.0002265220
## avgDeathsPerYear_log            2.985699e-05  6.517124e-03  0.0008812325
## popEst2015_log                  4.191961e-05  5.080829e-03  0.0007807718
## studyPerCap_log                 5.259145e-05  8.331199e-03 -0.0001474517
## PctBachDeg18_24_log             4.019505e-05  3.291052e-03 -0.0005208502
## PctAsian_log                    5.276969e-05  2.959141e-03  0.0014226014
## PctOtherRace_log                2.658387e-05 -5.396532e-03  0.0021289257
##                                          PC4           PC5          PC6
## incidenceRate                  -0.0239603406 -0.0102743529 -0.028830246
## medIncome                       0.0006415596 -0.0004246128 -0.000703812
```

```
## povertyPercent                   0.1222075166 -0.0036393704 -0.010854448
## MedianAge                       -0.1027775191 -0.0439291367 -0.117744154
## MedianAgeMale                   -0.1081293099 -0.0437031553 -0.118109239
## MedianAgeFemale                 -0.0957061678 -0.0448537792 -0.119677292
## AvgHouseholdSize                 0.0041504618 -0.0004045594 -0.001868737
## PercentMarried                  -0.2144672294 -0.0289236050 -0.092692774
## PctNoHS18_24                     0.0364289054 -0.0660849794 -0.288900467
## PctHS18_24                      -0.0507133047 -0.0964908229 -0.428255982
## PctSomeCol18_24                  0.0086971784  0.1318678355  0.648216925
## PctHS25_Over                    -0.0956405109 -0.0463118782 -0.149313791
## PctBachDeg25_Over                0.0199431166  0.0311717432  0.117950793
## PctEmployed16_Over              -0.1018027385  0.7261261118 -0.096579401
## PctUnemployed16_Over             0.0720339998 -0.0155780597 -0.027895816
## PctPrivateCoverage              -0.1565847308  0.0502228223  0.215154407
## PctPrivateCoverageAlone         -0.0072716660 -0.0126121411  0.099194820
## PctEmpPrivCoverage              -0.0578692746  0.0562668694  0.188769073
## PctPublicCoverage                0.0045781455 -0.0597782877 -0.153814921
## PctPublicCoverageAlone           0.0732873354 -0.0302212731 -0.102106908
## PctWhite                        -0.7026413873 -0.0690113259  0.070364678
## PctBlack                         0.5608034305  0.0519394699 -0.012122339
## PctMarriedHouseholds            -0.1692330054 -0.0416347675 -0.082257686
## BirthRate                       -0.0020807674  0.0004699628 -0.011029705
## Imputed_PctEmployed16_Over       0.0298021504 -0.6363941558  0.232843011
## Imputed_PctPrivateCoverageAlone -0.0729859636  0.0812418956  0.136212135
## avgAnnCount_log                  0.0097689333  0.0059957830  0.021446123
## avgDeathsPerYear_log             0.0155279039  0.0033173886  0.017486262
## popEst2015_log                   0.0196964230  0.0047923293  0.021979287
## studyPerCap_log                  0.0136713880  0.0118067620  0.043800193
## PctBachDeg18_24_log             -0.0019462766  0.0098677742  0.022167630
## PctAsian_log                     0.0167310672  0.0063386717  0.026076470
## PctOtherRace_log                 0.0204498343  0.0059043031  0.003210755
##                                          PC7           PC8           PC9
## incidenceRate                    0.0390234580 -0.0093662323  0.0274063918
## medIncome                        0.0006339088  0.0003648066 -0.0002398129
## povertyPercent                   0.1467149657 -0.0617366210 -0.0463726552
## MedianAge                       -0.1001072511  0.4386110909 -0.0162655092
## MedianAgeMale                   -0.0901143331  0.4270135556 -0.0184496972
## MedianAgeFemale                 -0.1106408661  0.4446674587 -0.0046186308
## AvgHouseholdSize                 0.0080525471 -0.0145421709 -0.0001485935
## PercentMarried                  -0.0974371630  0.1051191261  0.1221936255
## PctNoHS18_24                     0.1203499449 -0.0565390021  0.7283464021
## PctHS18_24                      -0.2677070314 -0.2423834562 -0.5769585262
## PctSomeCol18_24                  0.1655048131  0.2458258186 -0.1654418168
## PctHS25_Over                    -0.2644968394 -0.0259693958 -0.0154496403
## PctBachDeg25_Over                0.0386782447  0.0547909704 -0.0077330632
## PctEmployed16_Over              -0.0427329057 -0.0688405601  0.0454625679
## PctUnemployed16_Over             0.0921430013  0.0119285085 -0.0789660623
## PctPrivateCoverage              -0.4717355298  0.0771847432  0.1052004813
## PctPrivateCoverageAlone         -0.1687166426 -0.0815586586  0.0491063355
## PctEmpPrivCoverage              -0.3293569742 -0.2387066363  0.0443193595
## PctPublicCoverage                0.2029002501  0.3321865425 -0.1423984507
## PctPublicCoverageAlone           0.2513417065  0.0396999385 -0.1280724991
## PctWhite                         0.0070228497  0.0323074009  0.0094055571
## PctBlack                        -0.4630024395  0.2664729613  0.0759619287
```

```
## PctMarriedHouseholds            -0.0639747084 -0.0290812177  0.1158529233
## BirthRate                        0.0079813261 -0.0325854131  0.0205518403
## Imputed_PctEmployed16_Over      -0.0976464404 -0.1157297889  0.0622880928
## Imputed_PctPrivateCoverageAlone -0.2115646987 -0.0960478738  0.0493637597
## avgAnnCount_log                  0.0163218324 -0.0050783496 -0.0174774005
## avgDeathsPerYear_log             0.0208906574 -0.0084339018 -0.0336987927
## popEst2015_log                   0.0257742305 -0.0242340600 -0.0332690926
## studyPerCap_log                  0.0216654308 -0.0132046275 -0.0291105002
## PctBachDeg18_24_log             -0.0141670329 -0.0024815535 -0.0051169532
## PctAsian_log                     0.0284010386 -0.0220452183 -0.0250619831
## PctOtherRace_log                 0.0465072366 -0.0448634729  0.0027215219
##                                          PC10          PC11          PC12
## incidenceRate                   -0.0134283158  0.0318558371 -0.0002435246
## medIncome                        0.0004781437 -0.0002909277  0.0001037265
## povertyPercent                   0.0798086524 -0.1100247401  0.1375426716
## MedianAge                       -0.1216065205 -0.0346171876 -0.1123221736
## MedianAgeMale                   -0.1219911417 -0.0286851693 -0.1204149545
## MedianAgeFemale                 -0.1193740151 -0.0632119127 -0.1003747269
## AvgHouseholdSize                 0.0126372375  0.0069478247  0.0010748465
## PercentMarried                   0.1497050006  0.5467756147 -0.0044059454
## PctNoHS18_24                     0.0042516891 -0.1365761090  0.0036608045
## PctHS18_24                      -0.0861940016  0.1131876722  0.0730517320
## PctSomeCol18_24                  0.3077621889  0.1258009006 -0.0806383948
## PctHS25_Over                     0.5506744590 -0.2551657622 -0.4167709327
## PctBachDeg25_Over               -0.4205312315  0.0651165745  0.1292239837
## PctEmployed16_Over              -0.1015444747  0.0952517303 -0.0698727307
## PctUnemployed16_Over             0.0870891778 -0.0902639003  0.0639455651
## PctPrivateCoverage              -0.2192365885 -0.0376845667 -0.1465524806
## PctPrivateCoverageAlone         -0.0557410479 -0.0364206103 -0.0156928336
## PctEmpPrivCoverage               0.0945517907 -0.3656841983 -0.0803425237
## PctPublicCoverage                0.1024033915 -0.2107095650 -0.0547871552
## PctPublicCoverageAlone           0.1768203958 -0.1634147082  0.0326913356
## PctWhite                         0.1077953804 -0.2638357656  0.6089161421
## PctBlack                         0.2152541438 -0.0042741212  0.5538391173
## PctMarriedHouseholds             0.3603538823  0.4829638008  0.0917094075
## BirthRate                        0.0155919185  0.0763388597 -0.0327897793
## Imputed_PctEmployed16_Over      -0.1370397299  0.1208462424 -0.0936292170
## Imputed_PctPrivateCoverageAlone -0.0447818320 -0.0444243839 -0.0229374088
## avgAnnCount_log                 -0.0326130117 -0.0385788565  0.0162281339
## avgDeathsPerYear_log            -0.0168318715 -0.0572275247  0.0284497035
## popEst2015_log                  -0.0174058494 -0.0523822286  0.0364669969
## studyPerCap_log                 -0.0988681947 -0.0769302489  0.0353303896
## PctBachDeg18_24_log             -0.0308920487 -0.0229781623 -0.0111188069
## PctAsian_log                    -0.0526703988 -0.0300491736 -0.0026282363
## PctOtherRace_log                -0.0308928426  0.0165209252  0.0180046886
##                                          PC13          PC14          PC15
## incidenceRate                    0.0022913293 -0.0109943203 -9.107062e-03
## medIncome                        0.0002330587 -0.0002213162  8.576295e-05
## povertyPercent                  -0.1809780740 -0.0888201479 -7.856113e-02
## MedianAge                       -0.0094931597  0.0479628229 -2.342462e-01
## MedianAgeMale                   -0.0003928275  0.0584292183 -2.663970e-01
## MedianAgeFemale                 -0.0489988760 -0.0071899876 -1.775884e-01
## AvgHouseholdSize                -0.0082423078  0.0005946404 -8.338760e-03
## PercentMarried                  -0.0325942982  0.3259190336  6.957641e-02
```

```
## PctNoHS18_24                    -0.0587684814 -0.0327221060 -2.480096e-02
## PctHS18_24                      -0.0474911583 -0.0438406742 -4.859026e-02
## PctSomeCol18_24                  0.0164923048 -0.0748348143 -1.071211e-01
## PctHS25_Over                     0.3026992276 -0.1464932094  6.591263e-02
## PctBachDeg25_Over               -0.0312368767  0.0998986184  1.308804e-02
## PctEmployed16_Over               0.3933961981  0.3119742816  3.381187e-02
## PctUnemployed16_Over            -0.2448413815 -0.0233861057 -3.814255e-02
## PctPrivateCoverage              -0.2654314579 -0.2045500276  5.469890e-01
## PctPrivateCoverageAlone         -0.0881385526  0.0405710408  6.309272e-02
## PctEmpPrivCoverage              -0.3358106233  0.4918708065 -3.763295e-01
## PctPublicCoverage               -0.1531935262  0.2860992586  4.986989e-01
## PctPublicCoverageAlone          -0.0369973470  0.4296238412  2.342164e-01
## PctWhite                         0.1909407909 -0.0530983373  2.784330e-02
## PctBlack                         0.1644015512  0.0547381644  4.830649e-02
## PctMarriedHouseholds            -0.3942477735  0.0613896132 -3.669767e-02
## BirthRate                        0.0687144591 -0.0064538018  1.923867e-01
## Imputed_PctEmployed16_Over       0.4258829352  0.3432503089  3.252238e-02
## Imputed_PctPrivateCoverageAlone -0.0872295372  0.0418815617  6.179280e-02
## avgAnnCount_log                 -0.0515945807  0.0992913201  3.891223e-02
## avgDeathsPerYear_log            -0.0622871390  0.0905636412  6.654632e-03
## popEst2015_log                  -0.0671386397  0.0887495854  5.989626e-03
## studyPerCap_log                 -0.0781369833  0.1721533661  5.974683e-02
## PctBachDeg18_24_log              0.0141599248  0.0649513220  8.869358e-02
## PctAsian_log                    -0.0589216377  0.0297040416  3.362946e-02
## PctOtherRace_log                -0.0323886872  0.0190736497 -3.207022e-02
##                                          PC16          PC17          PC18
## incidenceRate                    4.842526e-03  2.404590e-03  0.0031494907
## medIncome                        1.127417e-04  2.215584e-05 -0.0001008893
## povertyPercent                   5.826080e-01 -1.534634e-01 -0.3759198964
## MedianAge                        9.574208e-03  7.242179e-03 -0.0505813683
## MedianAgeMale                    4.622837e-02 -1.524005e-02 -0.0735038438
## MedianAgeFemale                 -1.473951e-02  3.429359e-02  0.0271016584
## AvgHouseholdSize                -9.379081e-03 -3.992482e-03  0.0068239123
## PercentMarried                   6.426948e-02 -4.951017e-02 -0.1647757453
## PctNoHS18_24                    -1.414768e-01  2.380510e-01 -0.2592988152
## PctHS18_24                      -1.068411e-01  2.683803e-01 -0.2327955211
## PctSomeCol18_24                 -1.278886e-01  2.494910e-01 -0.2418215699
## PctHS25_Over                     2.491074e-02 -4.103113e-01 -0.1126604112
## PctBachDeg25_Over                2.420214e-01 -4.843254e-01 -0.1676535053
## PctEmployed16_Over               1.732565e-02  5.966469e-02 -0.1175238480
## PctUnemployed16_Over            -2.389863e-01 -6.555231e-02 -0.4186221950
## PctPrivateCoverage               1.678257e-02  8.801689e-02 -0.0621664815
## PctPrivateCoverageAlone          9.786695e-02  8.960786e-04 -0.0350853824
## PctEmpPrivCoverage               8.084677e-02  1.507130e-01  0.0367902040
## PctPublicCoverage                5.233973e-02  1.490167e-01  0.0491125032
## PctPublicCoverageAlone           1.485381e-01  3.317013e-02  0.0544425366
## PctWhite                        -1.612145e-02  7.578345e-03 -0.0199503112
## PctBlack                        -3.635114e-02 -1.544925e-03  0.0141642083
## PctMarriedHouseholds             6.671775e-05 -1.728284e-01  0.1514622279
## BirthRate                        2.604198e-02  1.985894e-01 -0.3558582236
## Imputed_PctEmployed16_Over       1.993923e-02  6.558930e-02 -0.1142300727
## Imputed_PctPrivateCoverageAlone  9.977192e-02  1.125888e-03 -0.0385752013
## avgAnnCount_log                 -2.014601e-01 -8.857321e-02  0.0554205343
## avgDeathsPerYear_log            -2.115649e-01 -1.323618e-01  0.0546173157
```

```
## popEst2015_log                  -2.114756e-01 -1.393694e-01  0.0542491678
## studyPerCap_log                 -4.970308e-01 -3.830330e-01 -0.3473247709
## PctBachDeg18_24_log             -1.458578e-03 -2.051209e-01  0.2642297442
## PctAsian_log                    -1.474471e-01 -7.760914e-02  0.0712179198
## PctOtherRace_log                -1.918948e-01  6.123876e-02  0.1787723425
##                                          PC19          PC20          PC21
## incidenceRate                    2.343490e-03 -0.0019581646 -2.971202e-03
## medIncome                        7.617127e-05  0.0001193455  8.541024e-05
## povertyPercent                   1.861503e-01  0.3662134079  2.224161e-01
## MedianAge                       -3.427009e-03  0.0473032781 -1.517102e-02
## MedianAgeMale                   -8.427000e-02  0.0526960484  3.864811e-02
## MedianAgeFemale                  1.368263e-01  0.0413745596 -8.883373e-02
## AvgHouseholdSize                -5.020394e-03 -0.0077768259 -1.124884e-03
## PercentMarried                  -1.159348e-01  0.1876000684  6.879145e-02
## PctNoHS18_24                     1.405674e-01 -0.2805561349  1.046970e-01
## PctHS18_24                       1.657761e-01 -0.2473238771  9.775142e-02
## PctSomeCol18_24                  1.682398e-01 -0.2544363043  8.091567e-02
## PctHS25_Over                     1.141145e-02 -0.1709723568 -4.878543e-02
## PctBachDeg25_Over                9.102881e-03 -0.6004565091 -2.113457e-01
## PctEmployed16_Over              -6.581692e-02  0.0059651170  1.557805e-01
## PctUnemployed16_Over            -7.691970e-01 -0.0125568302  8.485677e-02
## PctPrivateCoverage              -2.553352e-02  0.0324599559  1.581896e-01
## PctPrivateCoverageAlone         -6.312931e-03 -0.0217575575  5.614314e-02
## PctEmpPrivCoverage               1.195233e-02 -0.0114387693 -1.557068e-01
## PctPublicCoverage                5.353455e-02 -0.0938629053  5.882485e-02
## PctPublicCoverageAlone           2.707960e-02 -0.1405716709  1.619526e-02
## PctWhite                        -1.856930e-02  0.0001360192 -1.260987e-02
## PctBlack                         5.258384e-03 -0.0207883044 -2.082361e-02
## PctMarriedHouseholds             1.145950e-01 -0.1623547430 -4.604972e-02
## BirthRate                        4.164097e-02  0.2055110927 -8.429651e-01
## Imputed_PctEmployed16_Over      -6.775563e-02  0.0166111248  1.598747e-01
## Imputed_PctPrivateCoverageAlone -6.138679e-03 -0.0221287825  5.533511e-02
## avgAnnCount_log                 -1.073300e-02  0.0415127124 -5.527325e-02
## avgDeathsPerYear_log            -4.221046e-03  0.0137451686 -4.368577e-02
## popEst2015_log                  -9.680990e-03  0.0054499553 -4.706367e-02
## studyPerCap_log                  4.498217e-01  0.2781410373  1.205559e-01
## PctBachDeg18_24_log             -1.510337e-01  0.2186244330 -8.482289e-02
## PctAsian_log                    -2.282620e-02 -0.0212681882 -6.010799e-02
## PctOtherRace_log                 6.393679e-02  0.0416924982 -2.318814e-03
##                                          PC22          PC23          PC24
## incidenceRate                   -2.438766e-03 -3.076443e-03 -0.0033661578
## medIncome                        5.278137e-05  4.918967e-05 -0.0000011818
## povertyPercent                   3.407319e-01 -1.343389e-01 -0.0099787888
## MedianAge                        2.232782e-02 -1.165647e-01  0.0437787094
## MedianAgeMale                   -5.407385e-02 -9.659222e-02  0.3760603524
## MedianAgeFemale                  1.034720e-01 -1.597153e-01 -0.3672683880
## AvgHouseholdSize                -1.119058e-02 -4.151863e-02  0.0104368621
## PercentMarried                   1.711076e-01  5.563802e-01 -0.0284986835
## PctNoHS18_24                     1.252500e-01  1.760215e-02  0.1094831378
## PctHS18_24                       1.121864e-01  1.252626e-02  0.0864669738
## PctSomeCol18_24                  1.172149e-01  2.411130e-02  0.0819206852
## PctHS25_Over                     1.131992e-01  4.190729e-02 -0.0667675063
## PctBachDeg25_Over                5.148637e-02  1.235691e-01 -0.1231824253
## PctEmployed16_Over               5.910006e-02 -2.772836e-01 -0.0479659014
```

```
## PctUnemployed16_Over            -6.210614e-02 -1.415629e-01 -0.1799116754
## PctPrivateCoverage               6.949714e-02 -4.925680e-02 -0.0026384939
## PctPrivateCoverageAlone         -3.352174e-02 -1.553461e-01  0.0770787419
## PctEmpPrivCoverage              -3.677736e-02  1.765569e-01 -0.0776458865
## PctPublicCoverage                2.777508e-03  1.030600e-01 -0.0929324078
## PctPublicCoverageAlone          -1.037752e-01 -1.908304e-01  0.0898321366
## PctWhite                         1.923575e-02 -8.583755e-03 -0.0065560134
## PctBlack                         7.029140e-03  1.243070e-02  0.0009487934
## PctMarriedHouseholds            -1.112448e-01 -4.996091e-01 -0.0045593719
## BirthRate                        1.254114e-02 -1.386528e-01  0.0237029728
## Imputed_PctEmployed16_Over       6.077304e-02 -2.812761e-01 -0.0513270154
## Imputed_PctPrivateCoverageAlone -3.348626e-02 -1.550236e-01  0.0778570418
## avgAnnCount_log                  3.530078e-01 -2.074385e-02  0.1794145337
## avgDeathsPerYear_log             3.024800e-01  1.038296e-02  0.1756722275
## popEst2015_log                   2.983821e-01 -8.034654e-03  0.1874033443
## studyPerCap_log                 -3.338036e-01  3.347111e-03 -0.0673084274
## PctBachDeg18_24_log              2.261081e-02 -4.071483e-02  0.1373389608
## PctAsian_log                     3.739472e-01 -1.218776e-01  0.3168886023
## PctOtherRace_log                 4.275198e-01 -1.145852e-01 -0.6166701625
##                                         PC25          PC26          PC27
## incidenceRate                    1.890466e-03  2.177616e-03  1.542202e-03
## medIncome                       -1.451296e-05  9.786155e-06 -4.127057e-06
## povertyPercent                  -5.688369e-02  9.294539e-02 -4.029734e-02
## MedianAge                        4.584543e-02 -3.385821e-02 -1.344172e-02
## MedianAgeMale                    4.709777e-01  2.842810e-01 -2.243901e-02
## MedianAgeFemale                 -4.382337e-01 -4.031140e-01  2.827926e-02
## AvgHouseholdSize                 3.402472e-03  2.412670e-02 -1.561224e-02
## PercentMarried                   6.246741e-02 -1.849050e-01  9.543409e-02
## PctNoHS18_24                     5.443470e-03 -2.147281e-02 -1.998980e-02
## PctHS18_24                      -4.204039e-03 -1.104503e-02 -2.190619e-02
## PctSomeCol18_24                  7.493719e-03 -2.412060e-02 -1.649195e-02
## PctHS25_Over                     4.758954e-02  7.932542e-03  1.842702e-02
## PctBachDeg25_Over                2.589536e-02  7.238105e-02 -3.240734e-03
## PctEmployed16_Over              -1.280276e-01  1.286556e-01 -3.537748e-02
## PctUnemployed16_Over            -5.202776e-02 -2.496856e-02  2.289118e-02
## PctPrivateCoverage               2.312755e-02  8.310920e-02 -3.793369e-02
## PctPrivateCoverageAlone          1.838653e-01 -3.481945e-01  2.131585e-02
## PctEmpPrivCoverage              -9.556382e-02  1.816999e-01  2.988850e-02
## PctPublicCoverage               -1.483823e-01  3.487092e-01 -1.172524e-02
## PctPublicCoverageAlone           2.729517e-01 -4.429996e-01  4.733842e-02
## PctWhite                         9.020055e-03  1.258218e-02  1.914694e-02
## PctBlack                         1.551806e-02  3.018143e-02  2.649023e-02
## PctMarriedHouseholds            -9.689683e-02  1.988355e-01 -6.305398e-02
## BirthRate                        7.268002e-02  3.462270e-02 -1.911346e-02
## Imputed_PctEmployed16_Over      -1.286713e-01  1.303191e-01 -3.603102e-02
## Imputed_PctPrivateCoverageAlone  1.859945e-01 -3.497499e-01  2.066399e-02
## avgAnnCount_log                 -6.851118e-02 -7.981464e-02 -4.293381e-01
## avgDeathsPerYear_log            -5.700081e-02 -5.820739e-02 -2.495731e-01
## popEst2015_log                  -3.830927e-02 -5.248019e-02 -2.467692e-01
## studyPerCap_log                  5.440314e-02  1.677536e-02  9.531652e-02
## PctBachDeg18_24_log             -4.058248e-02  3.025371e-02 -1.766642e-02
## PctAsian_log                    -1.579263e-01  2.985205e-02  8.041955e-01
## PctOtherRace_log                 5.564569e-01  1.172975e-01  9.121939e-02
##                                         PC28          PC29          PC30
```

```
## incidenceRate                      6.956136e-04  1.867424e-03  3.975707e-04
## medIncome                          2.959991e-05 -9.395454e-06  7.813745e-06
## povertyPercent                     4.337628e-02  1.403595e-02 -2.843450e-02
## MedianAge                          1.913811e-02  2.860484e-02 -2.232121e-02
## MedianAgeMale                     -4.186803e-02 -1.896173e-02 -6.969166e-03
## MedianAgeFemale                    5.880362e-02  1.100830e-02 -3.328481e-02
## AvgHouseholdSize                  -1.850208e-04  2.591021e-03  7.841629e-04
## PercentMarried                     5.832797e-03  2.176398e-03  9.118064e-04
## PctNoHS18_24                       2.355707e-01 -3.198325e-03 -7.775657e-03
## PctHS18_24                         2.286721e-01 -1.424391e-02 -3.267847e-03
## PctSomeCol18_24                    2.304447e-01 -1.096886e-02 -6.035294e-03
## PctHS25_Over                      -5.165768e-03 -1.889471e-02  5.858457e-03
## PctBachDeg25_Over                  6.699634e-02 -3.106110e-02 -1.112187e-02
## PctEmployed16_Over                -6.373020e-03  2.735757e-02 -1.963129e-03
## PctUnemployed16_Over               2.333330e-02 -1.647361e-02  4.086851e-03
## PctPrivateCoverage                -1.573276e-02  3.876843e-02 -3.830964e-01
## PctPrivateCoverageAlone           -5.050807e-04 -2.462160e-03  4.650800e-01
## PctEmpPrivCoverage                 1.396248e-02 -1.724605e-02 -7.875053e-02
## PctPublicCoverage                 -1.142915e-02  3.405467e-03  4.276591e-01
## PctPublicCoverageAlone             2.756674e-03  2.817882e-03 -4.763090e-01
## PctWhite                           3.852901e-03 -7.846141e-03 -7.470709e-03
## PctBlack                          -3.173737e-03 -9.405692e-03 -3.465541e-03
## PctMarriedHouseholds              -2.881219e-02  6.888704e-04  3.935154e-03
## BirthRate                          3.317668e-03  3.045288e-02  1.811025e-02
## Imputed_PctEmployed16_Over        -5.534993e-03  2.719234e-02 -2.374272e-03
## Imputed_PctPrivateCoverageAlone   -1.143363e-03 -2.479579e-03  4.652123e-01
## avgAnnCount_log                   -1.567336e-01 -7.403448e-01 -7.041130e-04
## avgDeathsPerYear_log              -8.021171e-02  4.636242e-01  2.337738e-02
## popEst2015_log                    -8.116699e-02  4.637416e-01  1.278981e-02
## studyPerCap_log                    3.091990e-02 -4.538517e-02  2.922891e-03
## PctBachDeg18_24_log                8.769297e-01 -4.909285e-02  1.320378e-02
## PctAsian_log                      -1.145247e-01 -9.757412e-02  2.320305e-03
## PctOtherRace_log                   6.613138e-02 -7.737386e-03  2.200634e-02
##                                           PC31          PC32          PC33
## incidenceRate                      3.820231e-04  3.362065e-04  6.874338e-04
## medIncome                         -1.651414e-06 -3.613207e-06 -1.252251e-06
## povertyPercent                    -1.784332e-03  1.649722e-03 -1.034232e-03
## MedianAge                         -8.092936e-01 -4.588625e-02  4.183968e-02
## MedianAgeMale                      4.450790e-01  1.685803e-02 -2.359444e-02
## MedianAgeFemale                    3.721076e-01  4.548965e-02 -8.175986e-03
## AvgHouseholdSize                  -7.320946e-02  9.516275e-01 -2.921652e-01
## PercentMarried                     1.488260e-02  3.338421e-02 -1.057383e-03
## PctNoHS18_24                       3.282622e-03 -7.282185e-04  9.557706e-06
## PctHS18_24                         7.608054e-03 -6.374485e-04 -3.628922e-04
## PctSomeCol18_24                    9.213455e-03  1.088432e-03  1.681402e-04
## PctHS25_Over                      -8.939980e-04  2.848658e-03  2.734216e-03
## PctBachDeg25_Over                  2.052391e-03  5.230026e-03 -7.348449e-03
## PctEmployed16_Over                 1.355429e-02 -7.543061e-03  1.519224e-03
## PctUnemployed16_Over               1.757001e-02 -1.051990e-02 -1.623173e-03
## PctPrivateCoverage                -6.403056e-03  8.224594e-03 -3.207421e-03
## PctPrivateCoverageAlone           -3.554374e-03  1.830723e-03 -1.561451e-03
## PctEmpPrivCoverage                 1.832536e-03  9.462857e-04  1.886578e-03
## PctPublicCoverage                 -3.769136e-03  5.930007e-03  9.991116e-03
## PctPublicCoverageAlone            -2.062723e-05 -3.206900e-03 -8.198716e-03
```

```
## PctWhite                          -1.541676e-03  4.628864e-03 -1.370241e-03
## PctBlack                           1.143393e-03  3.371158e-03 -9.087775e-04
## PctMarriedHouseholds              -5.585301e-03 -4.257668e-02  2.494304e-03
## BirthRate                          2.094003e-03 -2.678757e-03 -2.129334e-03
## Imputed_PctEmployed16_Over         1.433571e-02 -7.771756e-03  1.472999e-03
## Imputed_PctPrivateCoverageAlone -3.975178e-03  1.994497e-03 -1.508277e-03
## avgAnnCount_log                   -1.319252e-02 -4.010485e-03  2.392047e-03
## avgDeathsPerYear_log              -7.453614e-04 -2.121665e-01 -6.749481e-01
## popEst2015_log                     3.774315e-02  2.037866e-01  6.755758e-01
## studyPerCap_log                    7.186038e-05  3.130824e-05  3.932616e-07
## PctBachDeg18_24_log                8.289912e-03 -3.262847e-03  8.092976e-05
## PctAsian_log                      -1.108609e-02  3.859229e-03 -5.113184e-03
## PctOtherRace_log                  -1.124529e-02 -1.950202e-03 -2.333589e-03
```

```
PCA_model <- prcomp(df)
PCA_loadings <- PCA_model$rotation
```

```
PCA_summary <- summary(PCA_model)

PCA_summary
```

```
## Importance of components:
##                           PC1      PC2      PC3   PC4   PC5   PC6   PC7   PC8
## Standard deviation     12109 51.60938 28.16341 22.22 17.45 15.16 9.634 8.914
## Proportion of Variance     1  0.00002  0.00001  0.00  0.00  0.00 0.000 0.000
## Cumulative Proportion      1  0.99998  0.99999  1.00  1.00  1.00 1.000 1.000
##                           PC9  PC10  PC11  PC12 PC13  PC14 PC15  PC16  PC17
## Standard deviation     7.785 5.989 5.759 5.076 3.81 3.562 2.85 2.541 2.452
## Proportion of Variance 0.000 0.000 0.000 0.000 0.00 0.000 0.00 0.000 0.000
## Cumulative Proportion  1.000 1.000 1.000 1.000 1.00 1.000 1.00 1.000 1.000
##                           PC18  PC19 PC20  PC21  PC22  PC23  PC24  PC25  PC26
## Standard deviation     2.065 2.051 1.99 1.744 1.555 1.426 1.257 1.226 1.195
## Proportion of Variance 0.000 0.000 0.00 0.000 0.000 0.000 0.000 0.000 0.000
## Cumulative Proportion  1.000 1.000 1.00 1.000 1.000 1.000 1.000 1.000 1.000
##                           PC27   PC28   PC29  PC30   PC31   PC32    PC33
## Standard deviation     1.037 0.9268 0.6374 0.513 0.2628 0.1088 0.09137
## Proportion of Variance 0.000 0.0000 0.0000 0.000 0.0000 0.0000 0.00000
## Cumulative Proportion  1.000 1.0000 1.0000 1.000 1.0000 1.0000 1.00000
```

3 Splitting Test and Training Set

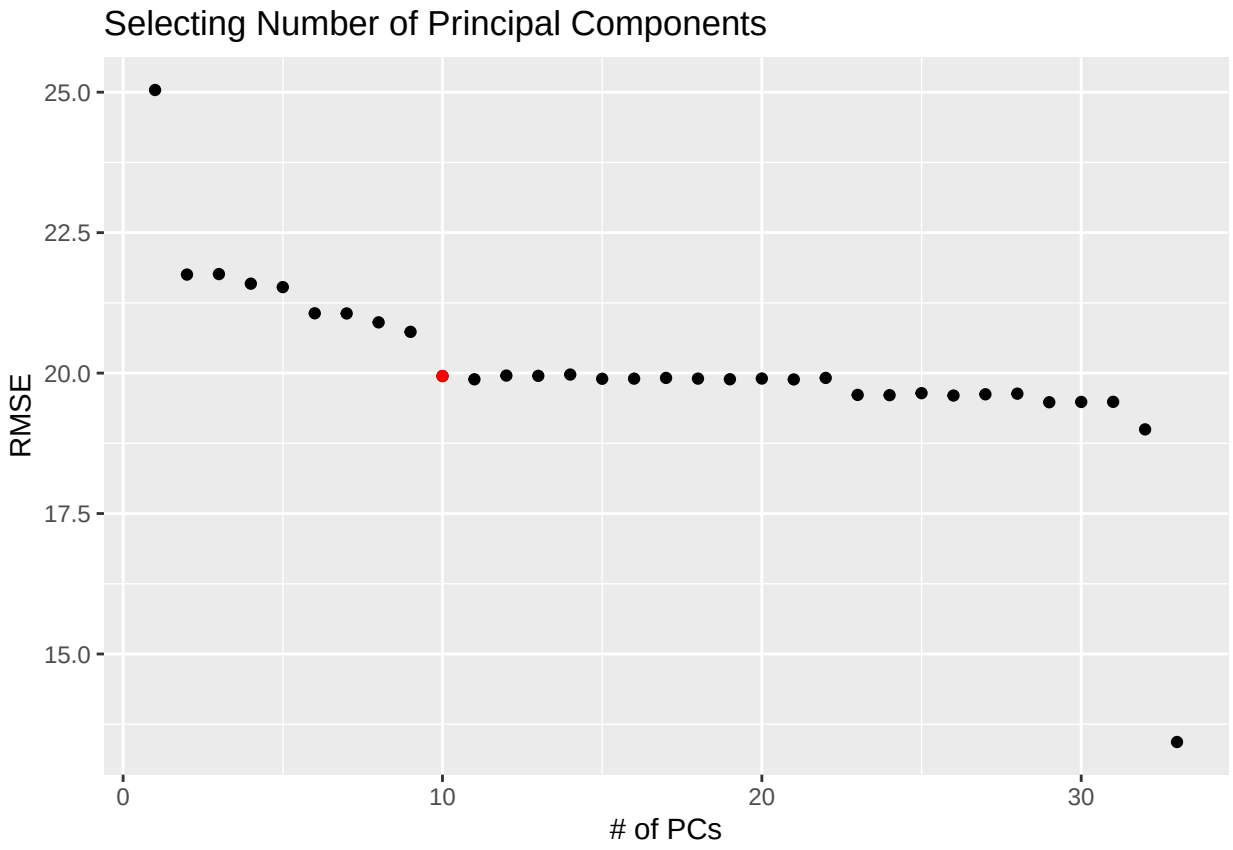
Arbitrarily chose 10% to be test set

```
set.seed(221)
test_set_index <- sample(1:nrow(df), floor(nrow(df))/10)
train_set_index <- setdiff(1:nrow(df), test_set_index)

test <- df[test_set_index,]
train <- df[train_set_index,]
```

```
test_y <- y[test_set_index]
test_x <- test %>% select(vars_1)
train_y <- y[train_set_index]
train_x <- train %>% select(vars_1)
```

4. 10 fold cross validation

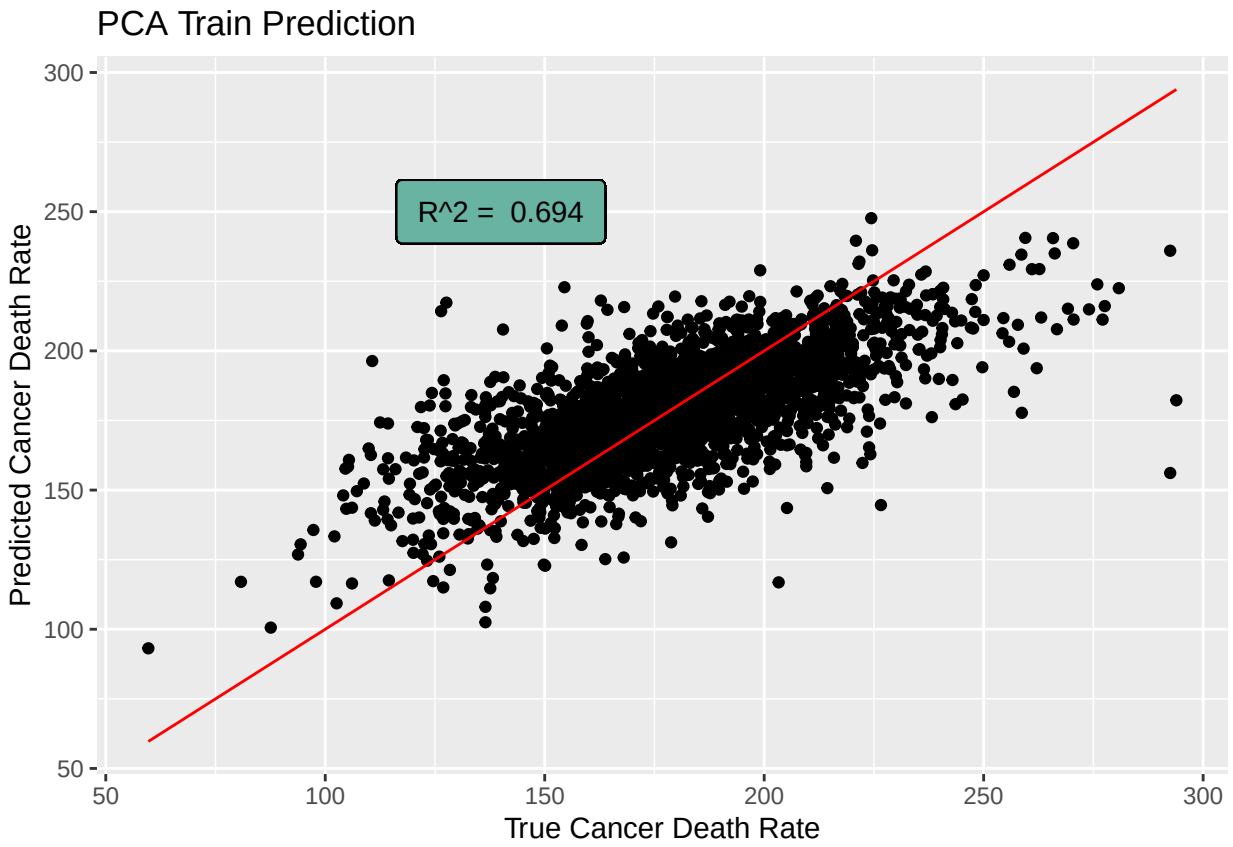
```
num_comp <- 1:ncol(train_x)
mses <- integer(ncol(train_x))
results <- data.frame()

PCA_train_model <- prcomp(train_x)
component_matrix <- data.frame(as.matrix(train_x) %*% PCA_train_model$rotation)

for(i in num_comp){

  pca_df <- data.frame(component_matrix[,1:i])


  eq <- paste(colnames(pca_df), collapse = ' + ')
  eq <- paste('deathRate', eq, sep = ' ~ ')

  pca_df <- pca_df %>% mutate(deathRate = train_y)

  # Train the model

  train.control <- trainControl(method = "cv", number = 10)

  model <- train(deathRate ~., data = pca_df, method = "lm",
                 trControl = train.control)
  if(i == 1){
    results <- model$results
  } else{
    results <- rbind(results, model$results)
  }

}

results <- results %>% mutate(ID = as.numeric(rownames(results)))
results_best <- results %>% filter(ID == 10)

g <- ggplot() + geom_point(data = results, aes(x = ID, y = RMSE)) +
  geom_point(data = results_best, aes(x = ID, y= RMSE), color = 'red') +
  xlab('# of PCs') +
  ylab('RMSE') +
  ggtitle('Selecting Number of Principal Components')

g
```



4.5 R^2 Plot of the Training Set

```
n <- 10

train_component_matrix <- data.frame(as.matrix(train_x) %*% PCA_train_model$rotation)
pca_df <- data.frame(train_component_matrix[,1:n])

eq <- paste(colnames(pca_df), collapse = ' + ')
eq <- paste('deathRate', eq, sep = ' ~ ')

pca_df <- pca_df %>% mutate(deathRate = train_y)

pca_train_model <- lm(eq, data = pca_df)
RMSE_PCA_train <- sqrt(sum(residuals(pca_train_model)^2)/nrow(pca_df))

y_pred <- unname(predict(pca_train_model))

pca_df <- pca_df %>% mutate(y_pred = y_pred)

R_squared <- as.numeric(unname(cor(y_pred, train_y)))
R_squared <- sprintf("%.3f", round(R_squared,3))
R_squared_label <- paste('R^2 = ', R_squared)

R_squared_PCA_train <- R_squared
```

```
g_pca_train <- ggplot(data = pca_df) +
  geom_point(aes(x = deathRate, y = y_pred)) +
  geom_line(aes(x = deathRate, y = deathRate), color = 'red') +
  geom_label(label = R_squared_label, x = 140, y = 250, label.padding = unit(0.55, "lines"),
             label.size = 0.35,
             color = "black",
             fill="#69b3a2") +
  xlab('True Cancer Death Rate') +
  ylab('Predicted Cancer Death Rate') +
  ggtitle('PCA Train Prediction')

g_pca_train
```

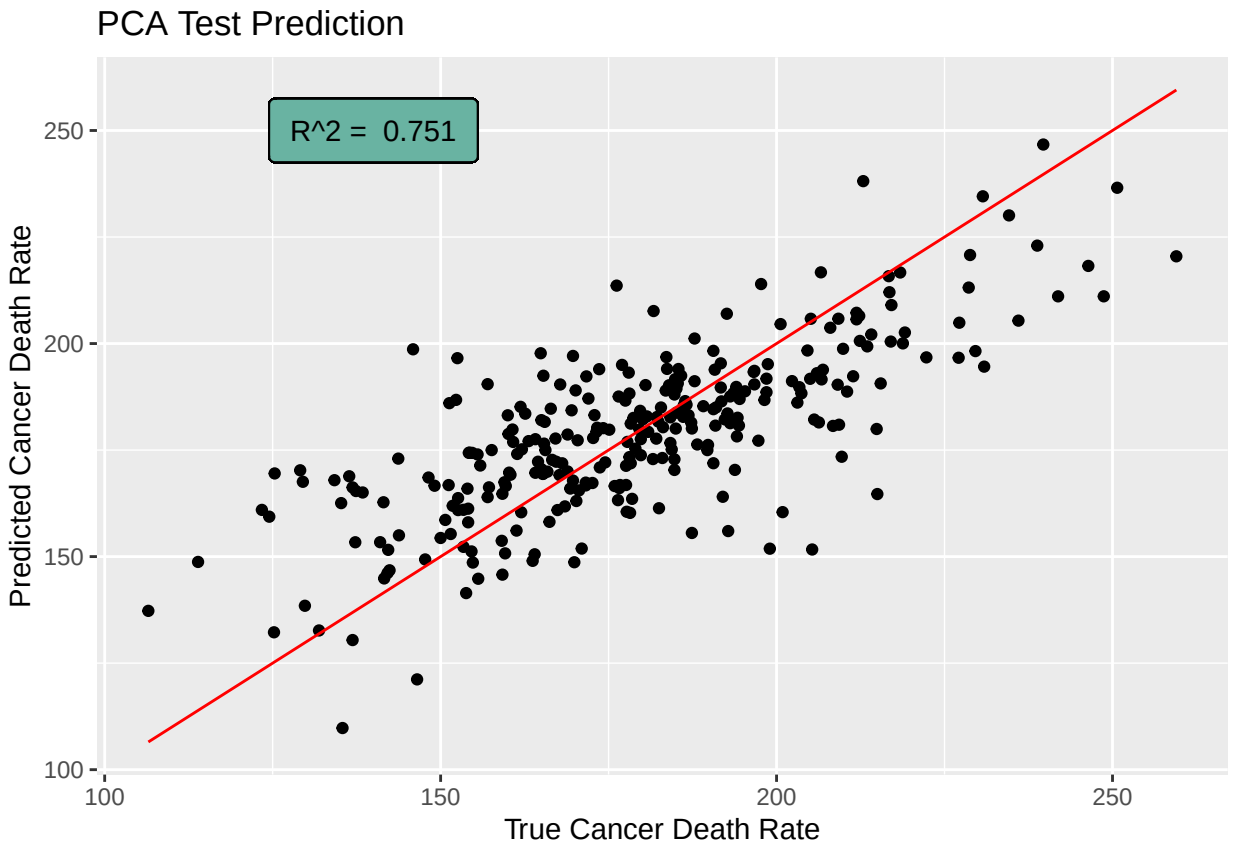


5 Predicting Test Set

Seems like 10 PC's is best

```
n <- 10

# use the train_PCA_loadings

test_component_matrix <- data.frame(as.matrix(test_x) %*% PCA_train_model$rotation)
pca_df <- data.frame(test_component_matrix[,1:n])
```

```r
eq <- paste(colnames(pca_df), collapse = ' + ')
eq <- paste('deathRate', eq, sep = ' ~ ')

pca_df <- pca_df %>% mutate(deathRate = test_y)

pca_test_model <- lm(eq, data = pca_df)
RMSE_PCA_test <- sqrt(sum(residuals(pca_test_model)^2)/nrow(pca_df))

y_pred <- unname(predict(pca_test_model))

pca_df <- pca_df %>% mutate(y_pred = y_pred)

R_squared <- as.numeric(unname(cor(y_pred, test_y)))
R_squared <- sprintf("%.3f", round(R_squared,3))
R_squared_label <- paste('R^2 = ', R_squared)

R_squared_PCA_test <- R_squared

g_pca_test <- ggplot(data = pca_df) +
  geom_point(aes(x = deathRate, y = y_pred)) +
  geom_line(aes(x = deathRate, y = deathRate), color = 'red') +
  geom_label(label = R_squared_label, x = 140, y = 250, label.padding = unit(0.55, "lines"),
             label.size = 0.35,
             color = "black",
             fill="#69b3a2") +
  xlab('True Cancer Death Rate') +
  ylab('Predicted Cancer Death Rate') +
  ggtitle('PCA Test Prediction')

g_pca_test
```




```
R_squared_PCA_train
```

```
## [1] "0.694"
```

```
R_squared_PCA_test
```

```
## [1] "0.751"
```

```
RMSE_PCA_train
```

```
## [1] 19.88739
```

```
RMSE_PCA_test
```

```
## [1] 17.29797
```