Title: High School Grades vs College GPAs

Data Summary:

The *test* dataset contains the percentile of high school grade and college GPA of 20 individuals. The first set of variables is the percentile of high school grades in mathematics (MATHHS), verbal (VERBHS), and creativity writing (CREHS). The second set of variables is college GPA in mathematics (MATHCOLL), English (ENGCOLL), science (SCICOLL), history (HISCOLL) and humanity (HUMCOLL).

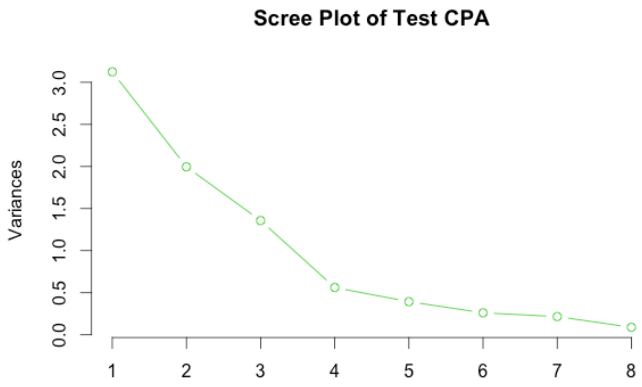
Statement of Problem:

Frist we want to find relationship of variables in the data using Principle Component Analysis (PCA). Then, we want to find the relationship between two sets of variable (grades in high school and GPA in college) using canonical correlation.

Analysis:

PCA

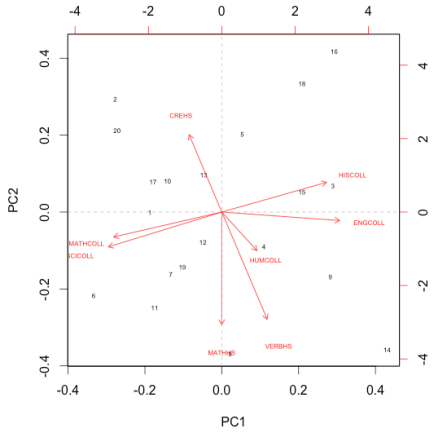
```
pca = prcomp(test, scale. =T)
summary(pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     1.7676 1.4125 1.1652 0.74949 0.62765 0.51093 0.46554 0.29853
Proportion of Variance 0.3906 0.2494 0.1697 0.07022 0.04924 0.03263 0.02709 0.01114
Cumulative Proportion  0.3906 0.6400 0.8097 0.87990 0.92914 0.96177 0.98886 1.00000

screeplot(pca, type="lines",col=3, ,main = "Scree Plot of Test CPA")

> round(pca$rotation[,1:4],3)
            PC1    PC2    PC3    PC4
MATHHS    0.000 -0.607 -0.078  0.600
VERBHS    0.196 -0.578  0.185 -0.366
CREHS    -0.141  0.417  0.565  0.333
MATHCOLL -0.466 -0.134 -0.195  0.376
ENGCOLL   0.508 -0.046 -0.049  0.124
SCICOLL  -0.489 -0.187  0.223 -0.010
HISCOLL   0.452  0.160 -0.058  0.485
HUMCOLL   0.152 -0.208  0.740  0.049
```

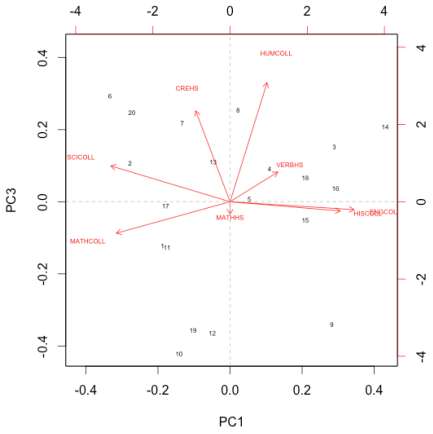


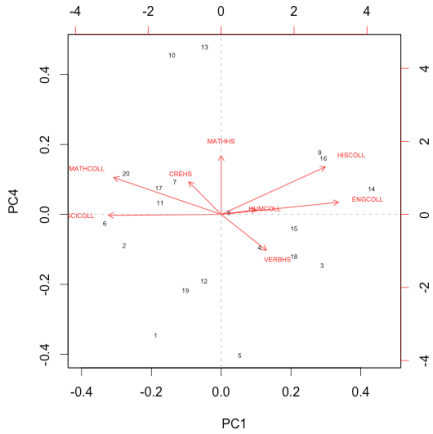
For this analysis, we would use 4 principal components because it explains about 89% of the variance (which is sufficient in this case) and the variance is relative low on the scree plot.
The formulas for component scores are using the eigenvectors (in this case shown in rotation, pca`$rotation`).

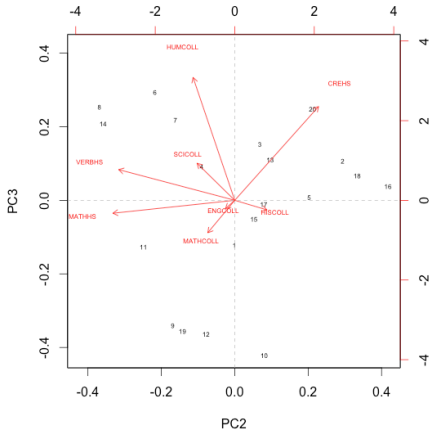
The component scores and biplot can see the relationship of variables in the data.

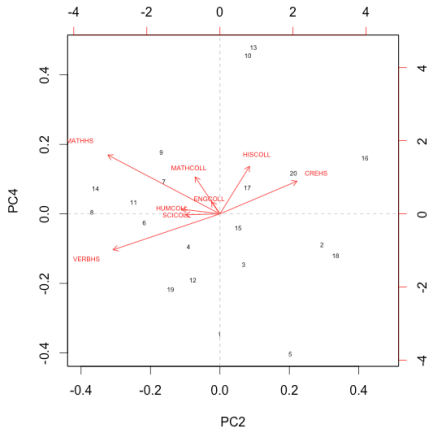
The biplots of the components (PC1 vs PC2), (PC1 vs PC3), (PC1 vs PC4), (PC2 vs PC3), (PC2 vs PC4), and (PC3 vs PC4):

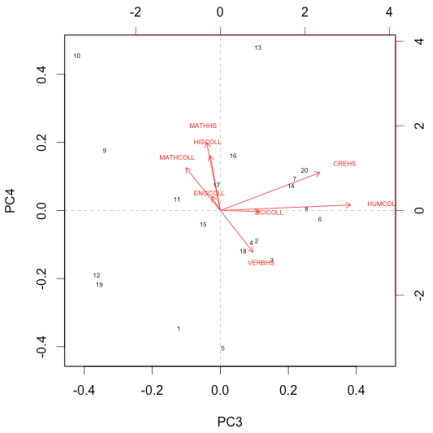












PCA Conclusion :

Frist principal component mainly is a measure the GPA in college. English & history and math & science are positively correlated. As English and history grades increase the grades for math and science tends to decrease.

Second principal component is a measure of mathematics and verbal grades of high school. As mathematics grade increase the verbal grades also increases.

Third component is a measure of creative writing in high school and humanity GPA in college. As the grade of creative writing in high school increases humanity GPA in college also increase.

Fourth component is a measure of mainly measurement of mathematics grades in high school and history GPA in college. If the grade in mathematics is high, then the GPA of history in college also tends to be high.

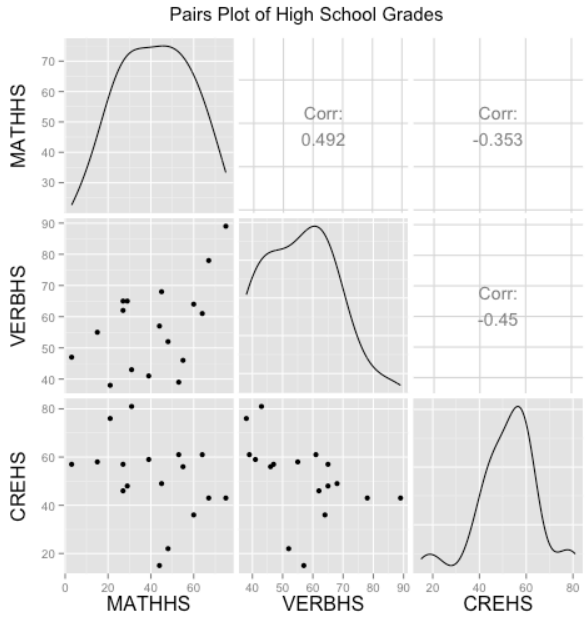
Canonical Correlation

There are 3 variables in the first set relating to the grades of high school and 5 variables in the second set relating to the GPA of college. First, we want to check if the data consist any unusual observations and if there any correlation between variables within set and between set.

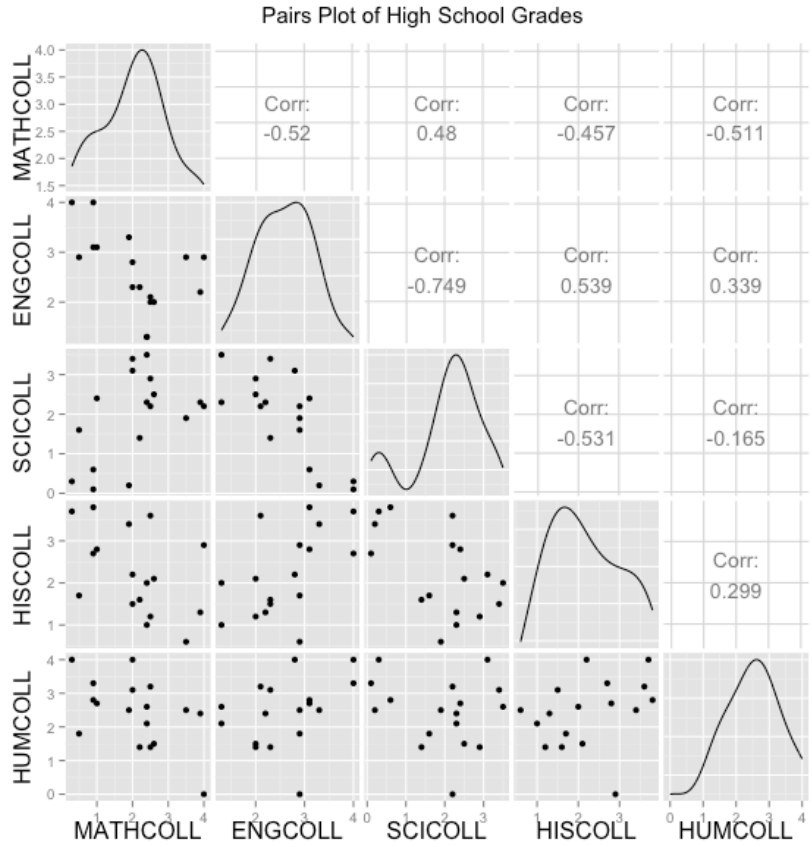
```
> summary(test)
     MATHHS          VERBHS          CREHS         MATHCOLL
 Min.   : 3.00   Min.   :36.00   Min.   :15.0   Min.   :0.300
 1st Qu.:27.00   1st Qu.:45.25   1st Qu.:43.0   1st Qu.:0.975
 Median :44.50   Median :58.00   Median :56.5   Median :2.100
 Mean   :42.50   Mean   :56.55   Mean   :51.5   Mean   :2.040
 3rd Qu.:56.25   3rd Qu.:65.00   3rd Qu.:61.0   3rd Qu.:2.525
 Max.   :77.00   Max.   :89.00   Max.   :81.0   Max.   :4.000
    ENGCOLL         SCICOLL         HISCOLL         HUMCOLL
 Min.   :1.30   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:2.10   1st Qu.:0.525   1st Qu.:1.450   1st Qu.:1.750
 Median :2.85   Median :2.200   Median :2.150   Median :2.500
 Mean   :2.72   Mean   :1.845   Mean   :2.255   Mean   :2.440
 3rd Qu.:3.15   3rd Qu.:2.600   3rd Qu.:3.100   3rd Qu.:3.125
 Max.   :4.00   Max.   :4.000   Max.   :4.000   Max.   :4.000
```

From the summary, there are few points with 0 GPA, which meant they have never take courses in that subject. Therefore, we want to remove than from the data to get a better relationship between high school grades and college GPA.

```
Remove that College grade contains 0
r0hs = test[-c(6,9,16),1:3]
r0col = test[-c(6,9,16),4:8]
ggpairs(r0hs)
ggpairs(r0col)
```



In high school, there is positive correlation between grades in mathematics & verbal and negative correlation between mathematics & creative writing and verbal & creative writing.



In college, there is negative relationship between math & English, math & history, math & humanity, English & Science, and science & history. There is positive relationship between, math & science, and English & history.

```
round(cor(r0hs, r0col),2)
       MATHCOLL ENGCOLL SCICOLL HISCOLL HUMCOLL
MATHHS     0.25    0.11    0.26    0.05    0.26
VERBHS    -0.46    0.58   -0.23    0.27    0.48
CREHS     -0.06   -0.18    0.16   -0.02    0.17
```

For between sets, there is strong negative correlation between high school verbal & college math. And there is strong positive correlation between high school verbal and college English and humanity. Now, we use canonical correlation analysis:

```
> cc1 = cc(r0hs, r0col)
Display the canonical correlations
>cc1$cor
[1] 0.8095189 0.7492662 0.3147305

Compute canonical loadings using cc1
> r.cc2 = comput(r0hs, r0col, cc1)
> r.cc2[3:6]
> r.cc2[3:6]

$corr.X.xscores
             [,1]       [,2]       [,3]
MATHHS -0.5006461  0.6331889 -0.5041742
VERBHS -0.7221170 -0.3495590 -0.6364492
CREHS  -0.2404995  0.1515167  0.9703801

$corr.Y.xscores
                [,1]        [,2]        [,3]
MATHCOLL  0.380611030  0.62996506  0.01878566
ENGCOLL  -0.423975873 -0.38894193 -0.28632912
SCICOLL   0.004842206  0.50384486  0.15194635
HISCOLL  -0.238012179 -0.16381983 -0.08030137
HUMCOLL  -0.647305032 -0.07281052  0.01593756

$corr.X.yscores
              [,1]       [,2]       [,3]
MATHHS -0.47757990  0.4231053 -0.1295197
VERBHS -0.62171758 -0.2713919 -0.3048909
CREHS  -0.08716687 -0.0653769  0.3015266

$corr.Y.yscores
               [,1]         [,2]        [,3]
MATHCOLL  0.3807754  0.856101688 -0.04720310
ENGCOLL  -0.4398514 -0.276596571 -0.71759930
SCICOLL  -0.1063103  0.412250217  0.41551362
HISCOLL  -0.1705005  0.004180802 -0.07449178
HUMCOLL  -0.8058019 -0.357807472  0.29196586
```

```
Tests of canonical dimensions
        WilksL         F df1      df2          p
[1,] 0.1362015 2.3668973  15 33.52812 0.01886443
[2,] 0.3951545 1.9201115   8 26.00000 0.09976607
[3,] 0.9009447 0.5130816   3 14.00000 0.67983399


Standardized HS canonical coefficients diagonal matrix of HS sd's
> s1 <- diag(sqrt(diag(cov(r0hs))))
> s1 %*% cc1$xcoef
           [,1]       [,2]       [,3]
[1,] -0.3218491  1.1053865 -0.1045079
[2,] -0.9005273 -0.8190219 -0.2010432
[3,] -0.7578899  0.1743547  0.8213364

Standardized college canonical coefficients diagonals matrix of college
sd's
> s2 <- diag(sqrt(diag(cov(r0col))))
> s2 %*% cc1$ycoef
            [,1]        [,2]       [,3]
[1,]  0.02772908  1.07156792 -0.2518086
[2,] -0.86083482  0.29171244 -1.3323733
[3,] -0.84044502  0.42008083 -0.2529961
[4,]  0.05499597  0.56435888  0.2467895
[5,] -0.64603957 -0.01495204  0.4970231
```

Conclusion for Canonical Correlation

The first row test of canonical dimensions (with p = 0.02) suggests that the there is at least one correlated vairate pairs. Since Wilk's lambda is significant, and since the canonical correlations are ordered from largest to smallest, we can conclude that at least the first variate pairs are significant. The second row of the test (with p = 0.10) does not firmly reject the null hypotheses and with strong canonical correlation (0.75), thus we would considered the second variate pairs to be significant. And we can tell that the third variate pairs are not significant with (p = 0.68).

According to the canonical loading, the first dimensions with correlation 0.809 are mainly due to verbal grade (loading = -0.72 with coef. = -0.90) in high school and humanity (loading = -0.81 with coef. = -0.65) and English (loading = -0.44 with coef. = -0.86) GPA in college.

The second dimensions with correlation 0.75 are mainly due to math grade (loading = 0.63 with coef. = 1.11) in high school and math (loading = 0.86 with coef. = 1.07) and science (loading = 0.41 with coef. = 0.56) GPA in college.

Since, the given data is very small we need more observations to draw more precise and accurate relationship between high school grades and college GPA.