Part 1- Retention Rate in College of Engineering

This part focus on exploring the relationship between gender and retention rate as well as the relationship between race and retention rate. The first data set contains the information we need. There are 4 sheets in this data set and according to our definition of *retention rate*(Number of students who graduate in the same major/Number of students who graduate from Engineering) we mainly use the second sheet and third sheet. The first one contains the information of students who graduate from College of Engineering and the second sheet contains the information of students who graduate from the same major.

For this part, we have the following questions:

- Are there significant differences in Retention Rate between female and male students within College of Engineering?
- Does the proportion of female students in each major have significant effect on their *Retention Rate*?
- Does GPA have significant effect on *Retention Rate* of female students?
- Are there significant differences in Retention Rate between URM and NonURM students within College of Engineering?
- Does GPA significantly affect the *Retention Rate* of URM students?

To answer these questions, first we will focus on the relationship between gender and retention rate. Then we will explore the effect of URM vs. Non URM students.

1. Gender and Retention Rate

- Data Description

Based on the questions and the given data set, I create a new data set called *subdata*, which contains 148 obs and 7 variables. Here is the data description table:

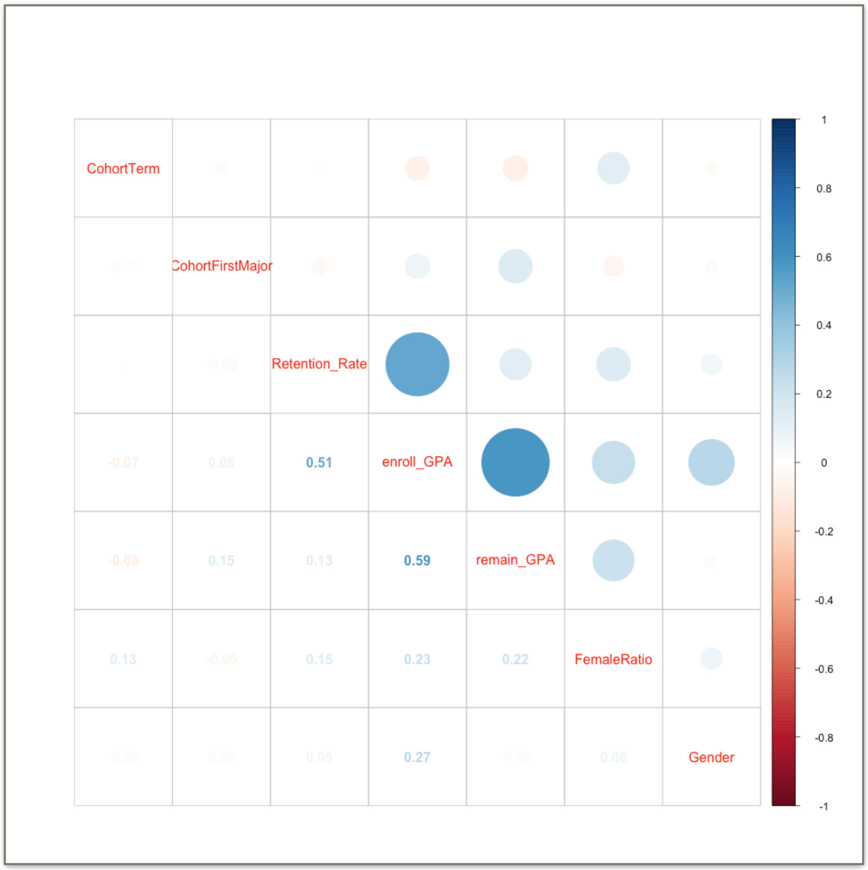
	Type	Interpretation
CohortTerm	Factor	There are 5 cohort terms from 2003-2007
CohortFirstMajor	Factor	There are 15 majors in college of engineering. These are the majors students enroll in.
Enroll GPA	Numeric	It's the 4 year GPA of students who enroll in college of engineering
Remain GPA	Numeric	It's the 4 year GPA of students who graduate in the original major

	Type	Interpretation
Female Ratio	Numeric	This is the proportion of female students in the specific major. Number of female students/Number of all students
Retention Rate	Numeric	Number of students who graduate in the same major/Number of students who graduate from Engineering
Gender	Factor	Students gender

Table 1. description for *subdata*

- Check Correlation

Pearson r's Correlation can be used to check the correlation between each pair of the predictors.



Plot 1. Ellipse Correlation Plot

Here is the ellipse correlation plot. The shade of the color represents the strength of the correlation. The deeper the color, the stronger the correlation.

This plot indicates that there is strong correlation between enroll_GPA and remain_GPA as well as enroll_GPA and Retention Rate.

- **Chisq Test and Likelihood Test**

Retention Or Not	Male	Female	Sum
Yes	3240	2318	5558
NO	638	494	1132
Sum	3878	2812	6690

Table 2. Contingency Table For Gender&Retention

Now let’s take a look at the contingency table, Chisq.test and likelihood test can help us test the correlation between Gender and Retention, here is the result in R:

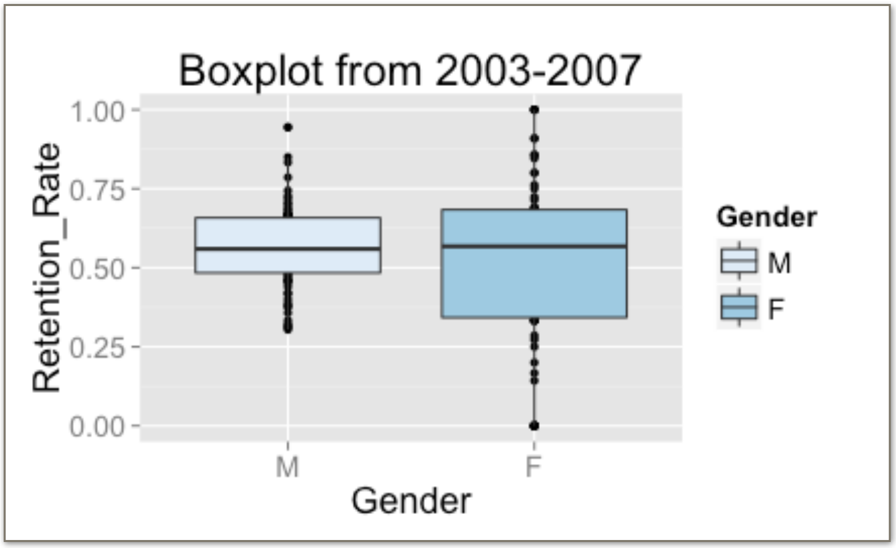
```
> chisq.test(ConTable)

	Pearson's Chi-squared test with Yates' continuity correction

data:  ConTable
X-squared = 1.3653, df = 1, p-value = 0.2426

> likelihood.test(ConTable)

	Log likelihood ratio (G-test) test of independence without correction

data:  ConTable
Log likelihood ratio statistic (G) = 1.4398, X-squared df = 1, p-value = 0.2302
```

Both of the p-value > 0.05 so we can conclude that in overall Gender and Retention are independent.

- **Question 1: Does Gender have significant effect on retention rate in college of engineering?**

First we carry on ANOVA analysis for Gender and Retention rate in all cohort majors and terms. The formula is: Retention_Rate ~ Gender.   Here is the result and box plot.

```
> summary(mod.1)
Coefficients:
            Estimate Std. Error t value Pr(>ltl)
(Intercept)  0.56207    0.02445  22.986   <2e-16 ***
GenderF     -0.04419    0.03458  -1.278    0.203
Residual standard error: 0.2104 on 146 degrees of freedom
Multiple R-squared:  0.01106,Adjusted R-squared:  0.004287
F-statistic: 1.633 on 1 and 146 DF,  p-value: 0.2033
```



Plot 2. Boxplot for Gender and Retention Rate

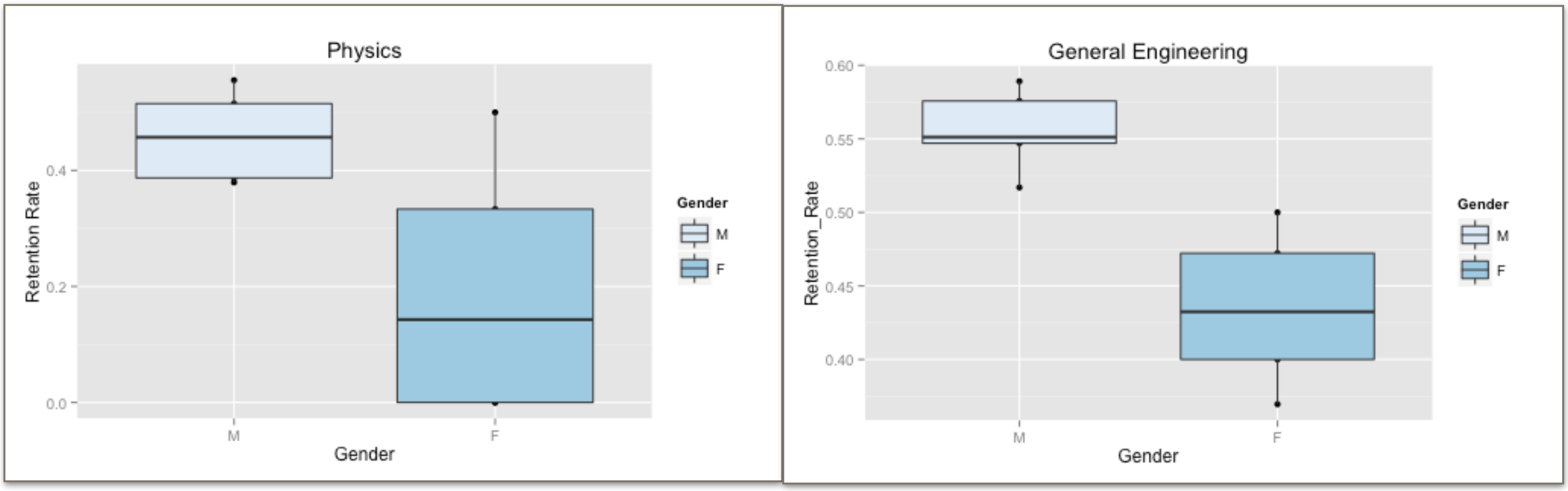
We can notice that the p-value in ANOVA analysis is 0.203>0.05, so in overall there is no significant different in Retention Rate between Female and Male students. In the box plot we can see that the mean values of Male and Female are very close but Female students have larger variance.

Next we'll explore the relationship between gender and retention rate in each major. After carried ANOVA analysis in each major, we found that in Engineering Physics and General Engineering, there is significant difference in Retention Rate between female and male students. Male students have significantly higher Retention Rate than female students. Here are the analysis results for different majors and the box plots in Physics and General Engineering.

Major	Estimate	Std Error	T.value	P-value
ABE	-0.2074	0.1272	-1.6302	0.1471
Aero	-0.0639	0.0596	-1.0726	0.3190
CS	-0.0177	0.0604	-0.2928	0.7782
Chem E	0.0520	0.0627	0.8289	0.4345
Civil	-0.0090	0.0459	-0.1970	0.8494
Comp E	0.0024	0.1037	0.0232	0.9821
E Mech	0.0288	0.2375	0.1211	0.9070

Major	Estimate	Std Error	T.value	P-value
E Phys	-0.2636	0.0814	-3.2368	0.0143
EE	0.0283	0.0354	0.7996	0.4502
Gen Eng	-0.1212	0.0259	-4.6763	0.0023
Ind Eng	0.0583	0.1523	0.3827	0.7133
MatSE	0.0797	0.1204	0.6619	0.5292
Mech E	-0.0934	0.0739	-1.2632	0.2470
NPRE	-0.1150	0.1417	-0.8113	0.4439
Bioen	-0.0154	0.0443	-0.3474	0.7424

Table 2. ANOVA analysis for each major



Plot 3. Box plot for Physics and General Engineering

Take a look at the ANOVA analysis table, we notice that the p-value for Physics is 0.0143 < 0.05 and p-value for General Engineering is 0.023 < 0.05. Both of them are significant. But in other majors, there is no significant difference between each gender.

- **Question 2: Does proportion of female students has significant effect on female's retention rate?**

In this part, we are interested in female students. So this time we subset *subdata* by gender=female and obtain new data set called *subfemale*.

To answer this question, we can fit an ANCOVA model for retention rate with female ratio, enroll_GPA, remain_GPA, CohortFirstMajor and CohortTerm as predictors. We need to check the correlation between female ratio and other 3 predictors since this question focus on the effect of female ratio.

Here is the result of the associations:

variables	Method	Coefficients	Interpretation
enroll_GPA	correlation analysis & linear regression	correlation coefficient r=0.263 p-value of linear regression is 0.0184 < 0.05	correlation analysis shows a weak positive correlation while linear regression shows association between enroll_GPA and female ratio.
remain_GPA	correlation analysis & linear regression	correlation coefficient r=0.269 p-value of linear regression is 0.029 < 0.05	correlation analysis shows a weak positive correlation while linear regression shows association between remain_GPA and female ratio.
CohortFirstMajor	ANOVA	p-value = 2.27E-09	CohortFirstMajor is associated with female ratio.
CohortTerm	ANOVA	p-value = 0.309	CohortTerm has no correlation with female ratio.

Table 3. Association with Female Ratio in *Subfemale*

So I kick out enroll_GPA, remain_GPA and CohortFirstMajor out. The final model is:

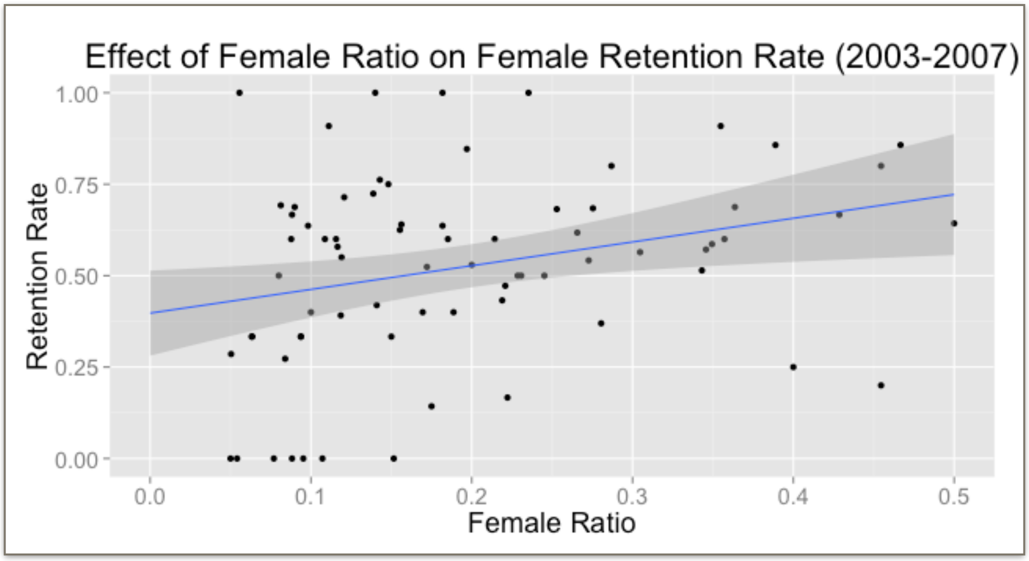
Retention_Rate ~ Female_Ratio + CohortTerm

Now let's take a look at the result of this ANCOVA analysis.

```
> ratio.4 <- lm(Retention_Rate ~ FemaleRatio+CohortTerm, data=subfemale)
> summary(ratio.4)

Call:
lm(formula = Retention_Rate ~ FemaleRatio + CohortTerm, data = subfemale)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51606 -0.12649 -0.01489  0.17392  0.52580 
```

```
Coefficients:
              Estimate Std. Error t value Pr(>ltl)
(Intercept)      0.45987    0.08320   5.527 5.77e-07 ***
FemaleRatio      0.66694    0.26524   2.514   0.0143 *
CohortTerm120048 -0.04697    0.09525  -0.493   0.6236
CohortTerm120058 -0.07904    0.09684  -0.816   0.4173
CohortTerm120068 -0.05782    0.09586  -0.603   0.5485
```

Estimate of Female Ratio is significant. When there is 1% unit increase in female ratio, we expect female's retention rate will increase 0.667%, which is not a quiet obvious positive effect.



Plot 4. Linear Regression Line for Female Ratio and Retention Rate.

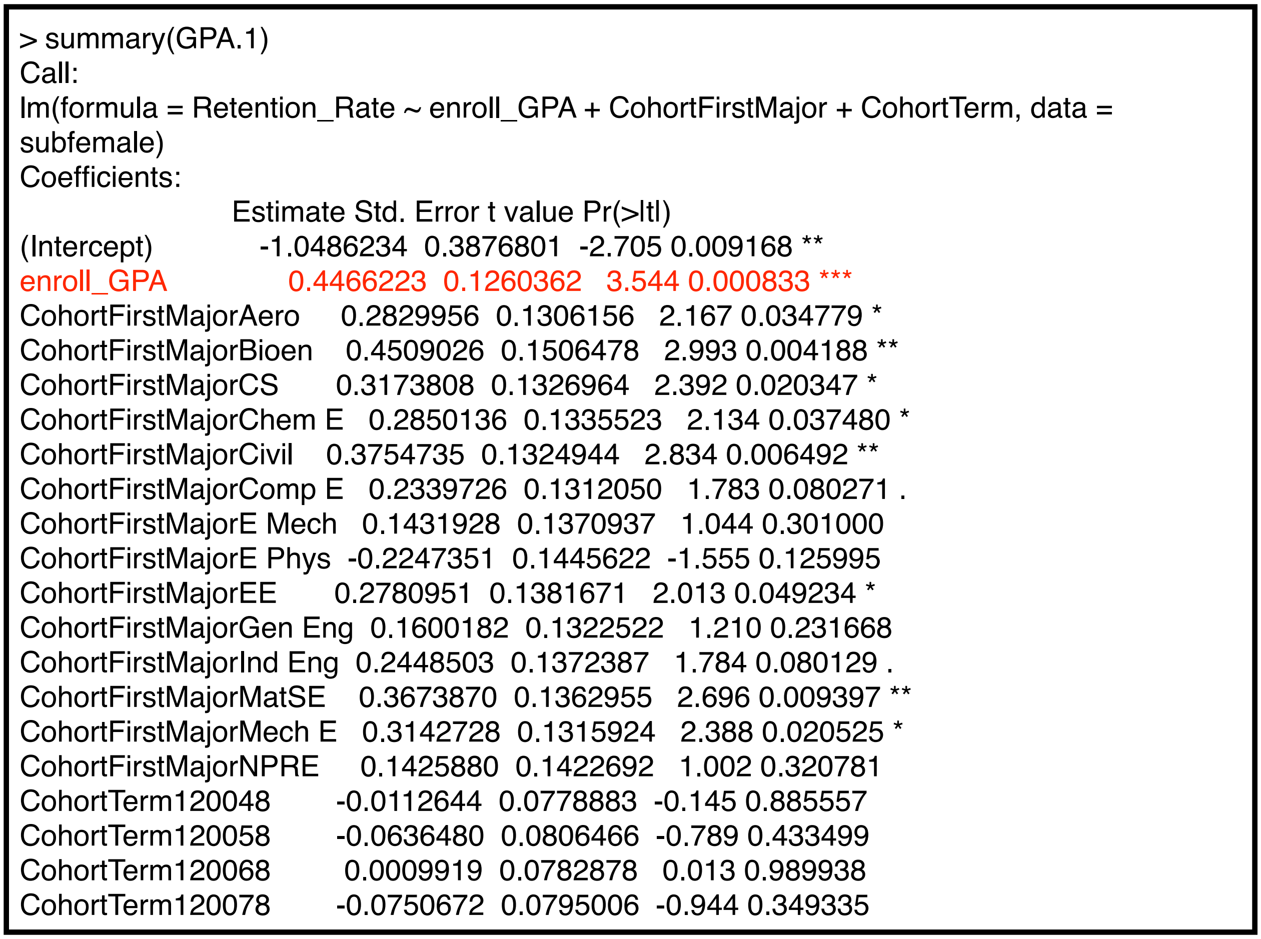
- **Question 3 Does GPA have significant effect on *Retention Rate* of female students?**

First we'll explore effect of enroll_GPA in *subfemale*.

After checking the correlation between enroll_GPA and other predictors, I built model with enroll_GPA, CohortFirstMajor, CohortTerm. The model is:

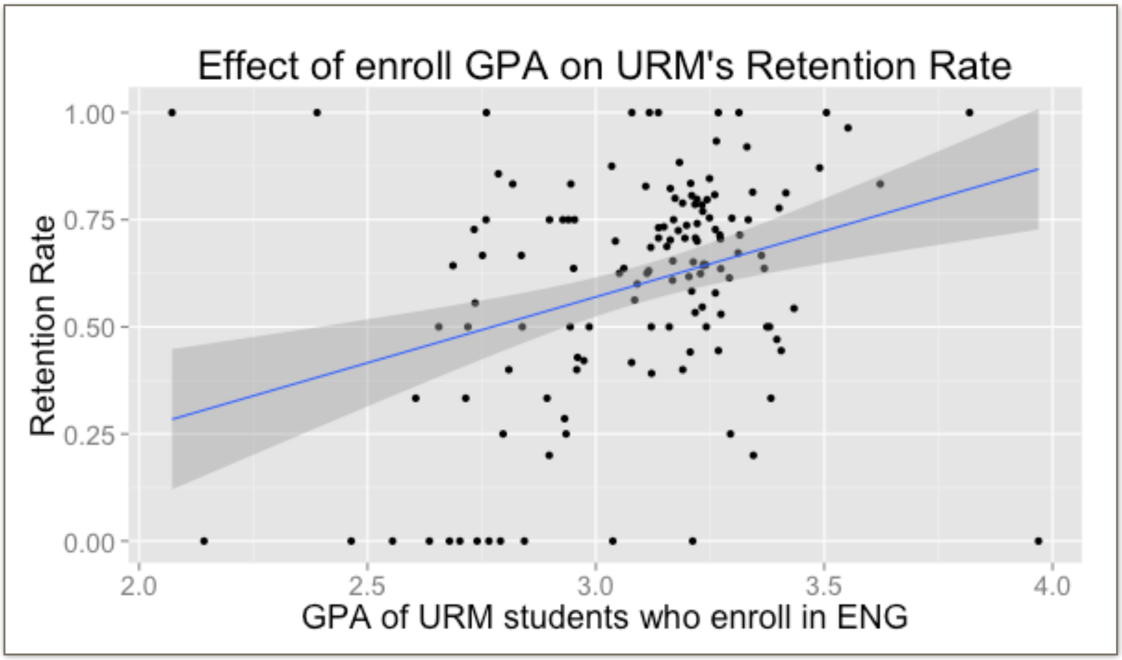
Retention Rate ~ enroll_GPA + CohortFirstMajor + CohortTerm

The result of ANCOVA analysis is here:



```
> summary(GPA.1)
Call:
lm(formula = Retention_Rate ~ enroll_GPA + CohortFirstMajor + CohortTerm, data =
subfemale)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.0486234  0.3876801  -2.705 0.009168 **
enroll_GPA            0.4466223  0.1260362   3.544 0.000833 ***
CohortFirstMajorAero     0.2829956  0.1306156   2.167 0.034779 *
CohortFirstMajorBioen    0.4509026  0.1506478   2.993 0.004188 **
CohortFirstMajorCS       0.3173808  0.1326964   2.392 0.020347 *
CohortFirstMajorChem E   0.2850136  0.1335523   2.134 0.037480 *
CohortFirstMajorCivil    0.3754735  0.1324944   2.834 0.006492 **
CohortFirstMajorComp E   0.2339726  0.1312050   1.783 0.080271 .
CohortFirstMajorE Mech   0.1431928  0.1370937   1.044 0.301000
CohortFirstMajorE Phys  -0.2247351  0.1445622  -1.555 0.125995
CohortFirstMajorEE       0.2780951  0.1381671   2.013 0.049234 *
CohortFirstMajorGen Eng  0.1600182  0.1322522   1.210 0.231668
CohortFirstMajorInd Eng  0.2448503  0.1372387   1.784 0.080129 .
CohortFirstMajorMatSE    0.3673870  0.1362955   2.696 0.009397 **
CohortFirstMajorMech E   0.3142728  0.1315924   2.388 0.020525 *
CohortFirstMajorNPRE     0.1425880  0.1422692   1.002 0.320781
CohortTerm120048        -0.0112644  0.0778883  -0.145 0.885557
CohortTerm120058        -0.0636480  0.0806466  -0.789 0.433499
CohortTerm120068         0.0009919  0.0782878   0.013 0.989938
CohortTerm120078        -0.0750672  0.0795006  -0.944 0.349335
```

Estimate of enroll_GPA is significant(with p-value = 0.0008). When there is 0.1 increase in enroll_GPA, we expect female's retention rate will increase 4.47%, which is a dramatic positive effect.



Next let’s take a look at the effect of remain_GPA.

After checking the correlation between remain_GPA and other predictors, I built model with remain_GPA, CohortFirstMajor, CohortTerm. The model is:

Retention Rate ~ remain_GPA + CohortFirstMajor + CohortTerm

The result of ANCOVA analysis is here:

```
> GPA.1 <- lm(Retention_Rate ~ remain_GPA+CohortFirstMajor+CohortTerm,
data=subfemale)
> summary(GPA.1)

Call:
lm(formula = Retention_Rate ~ remain_GPA + CohortFirstMajor +
   CohortTerm, data = subfemale)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.893457   0.456465   1.957   0.0564 .
remain_GPA            -0.126668   0.143756  -0.881   0.3828
CohortFirstMajorAero    -0.005385   0.165134  -0.033   0.9741
CohortFirstMajorBioen    0.405846   0.182073   2.229   0.0307 *
CohortFirstMajorCS       0.144311   0.168531   0.856   0.3963
CohortFirstMajorChem E   0.124527   0.167783   0.742   0.4617
CohortFirstMajorCivil    0.182434   0.165313   1.104   0.2755
CohortFirstMajorComp E   0.011685   0.166047   0.070   0.9442
CohortFirstMajorE Mech   0.374086   0.187516   1.995   0.0520 .
CohortFirstMajorE Phys  -0.098897   0.193407  -0.511   0.6116
CohortFirstMajorEE       0.191248   0.171529   1.115   0.2707
CohortFirstMajorGen Eng -0.030859   0.166571  -0.185   0.8538
CohortFirstMajorInd Eng  0.155965   0.174527   0.894   0.3762
CohortFirstMajorMatSE    0.264128   0.173337   1.524   0.1344
CohortFirstMajorMech E   0.106681   0.166510   0.641   0.5249
CohortFirstMajorNPRE     0.021007   0.176866   0.119   0.9060
CohortTerm120048        -0.023359   0.074594  -0.313   0.7556
CohortTerm120058         0.007196   0.081072   0.089   0.9297
CohortTerm120068        -0.032640   0.074623  -0.437   0.6639
CohortTerm120078        -0.066899   0.076390  -0.876   0.3857
```

Estimate of remain_GPA is insignificant(with p-value = 0.3828). So remain_GPA may not have effect on retention rate. It’s reasonable since enroll_GPA includes students who transform to other majors so it may have significant effect on retention rate. Remain_GPA is the GPA of students who stay in the same major so it doesn’t affect the retention rate.

2. Race and Retention Rate

- Data Description

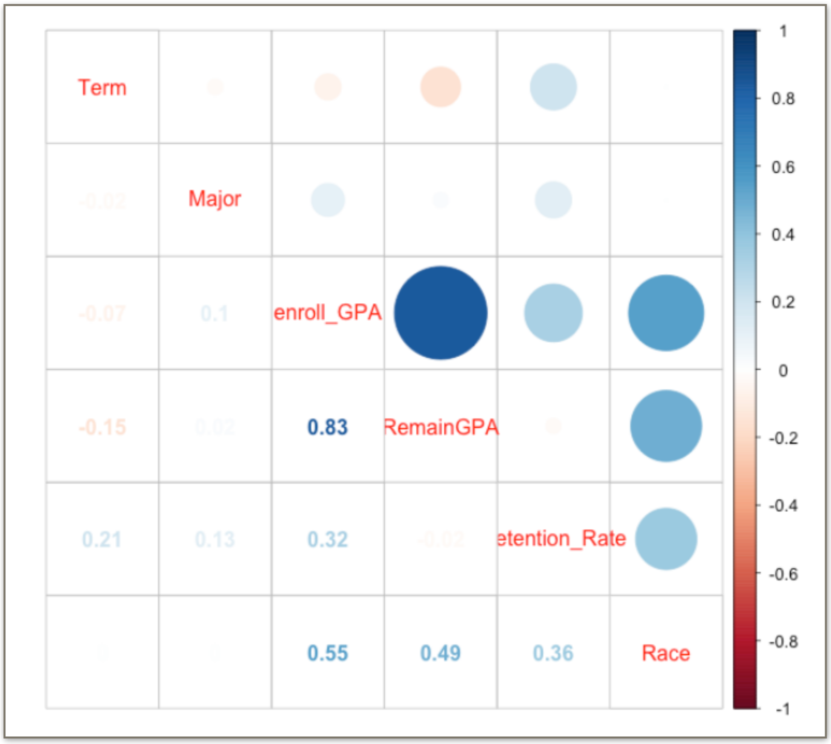
Based on the questions and the given data set, I create a new data set called *URMdata*, which contains 148 obs and 6 variables. Here is the data description table:

	Type	Interpretation
CohortTerm	Factor	There are 5 cohort terms from 2003-2007
CohortFirstMajor	Factor	There are 15 majors in college of engineering. These are the majors students enroll in.
Enroll GPA	Numeric	It's the 4 year GPA of students who enroll in college of engineering
Remain GPA	Numeric	It's the 4 year GPA of students who graduate in the original major
Retention Rate	Numeric	Number of students who graduate in the same major/Number of students who graduate from Engineering
Race	Factor	URM or Non URM students

Table 4. description for *URMdata*

- Check Correlation

Pearson r's Correlation can be used to check the correlation between each pair of the predictors.



Plot 6. Ellipse Correlation Plot

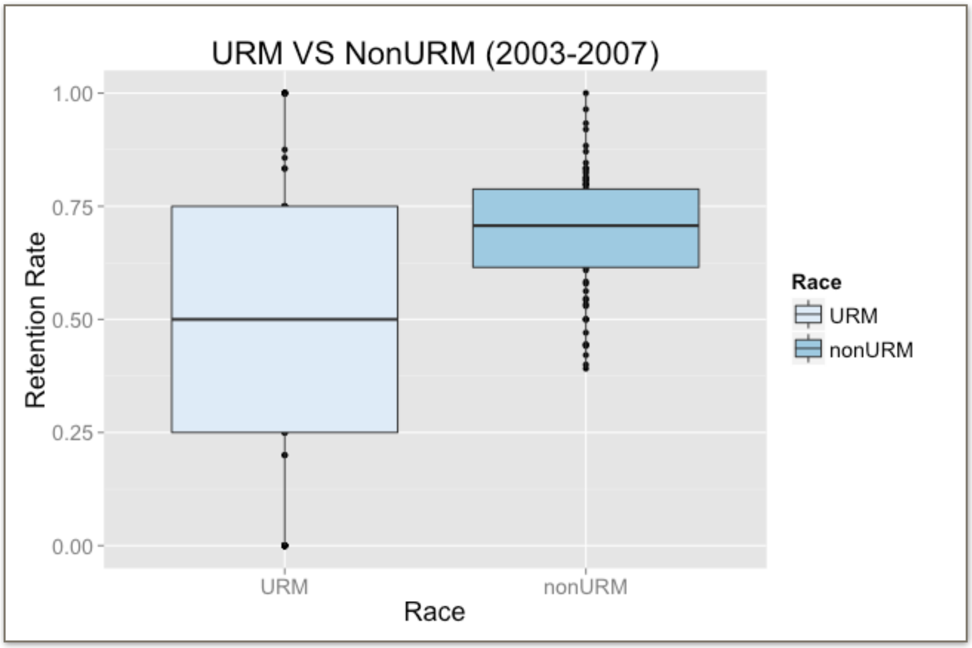
Here is the ellipse correlation plot. The shade of the color represents the strength of the correlation. The deeper the color, the stronger the correlation.

This plot indicates that there is strong correlation between enroll_GPA and remain_GPA as well as enroll_GPA and Retention Rate. There is moderate correlation between enroll_GPA and Race as well as Remain_GPA and Race.

- **Question 4: Are there significant differences in Retention Rate between URM and NonURM students within College of Engineering?**

First we carry on ANOVA analysis for Race and Retention rate in all cohort majors and terms. The formula is: Retention_Rate ~ Race   Here is the result and box plot.

```
> summary(URMmod_1)

Call:
lm(formula = Retention_Rate ~ Race, data = URMdata)

Coefficients:
            Estimate Std. Error t value Pr(>ltl)
(Intercept)  0.49484    0.03161  15.656  < 2e-16 ***
RacenonURM   0.19457    0.04301   4.524 1.32e-05 ***
```



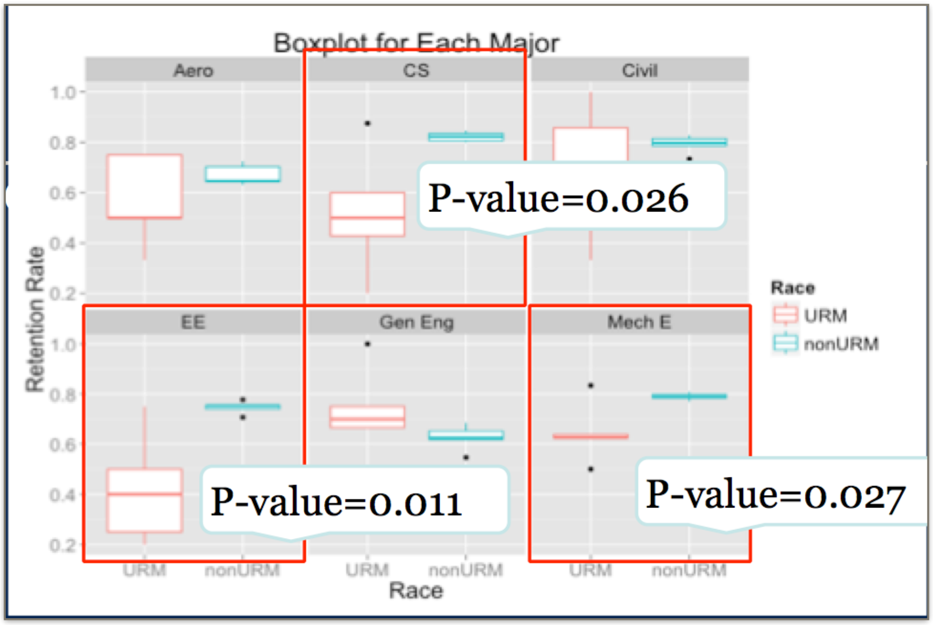
Plot 7. Boxplot for Race and Retention Rate

We can notice that the p-value in ANOVA analysis is much smaller than 0.05, so in overall there is significant difference in Retention Rate between URM and NonURM students. URM students have lower retention rate. In the box plot we can see that the mean value of URM students is smaller than Non URM students. URM students also have larger variance.

Next we'll explore the relationship between Race and Retention Rate in each major. In several majors, there is often no URM students enroll in. So we don't need to consider these majors. In this part, we only build ANOVA model for CS, Aero, Civil, Electronic Engineering, General Engineering, and Mechanical Engineering. Here is the analysis results in these majors.

	Estimate	Std..Error	t-value	P-value
Aero	0.1031	0.0829	1.2445	0.2485
CS	0.3013	0.1107	2.7218	0.0262
Civil	0.0918	0.1138	0.8071	0.4430
EE	0.3266	0.0989	3.3025	0.0108
Gen Eng	-0.0819	0.0960	-0.8528	0.4186
Mech E	0.1457	0.0539	2.7019	0.0270

Table 5. ANOVA analysis for each major



Plot 8. Box plot for each major

Take a look at the ANOVA analysis table, we notice that p-value for CS is 0.026 < 0.05, p-value for EE is 0.011 < 0.05 and p-value for Mech E is 0.027<0.05. All of them are significant.
In these majors, URM students have significantly lower retention rate. But in other majors, there is no significant difference between URM and Non URM students.

- **Question 5: Does GPA significantly affect the *Retention Rate* of URM students?**

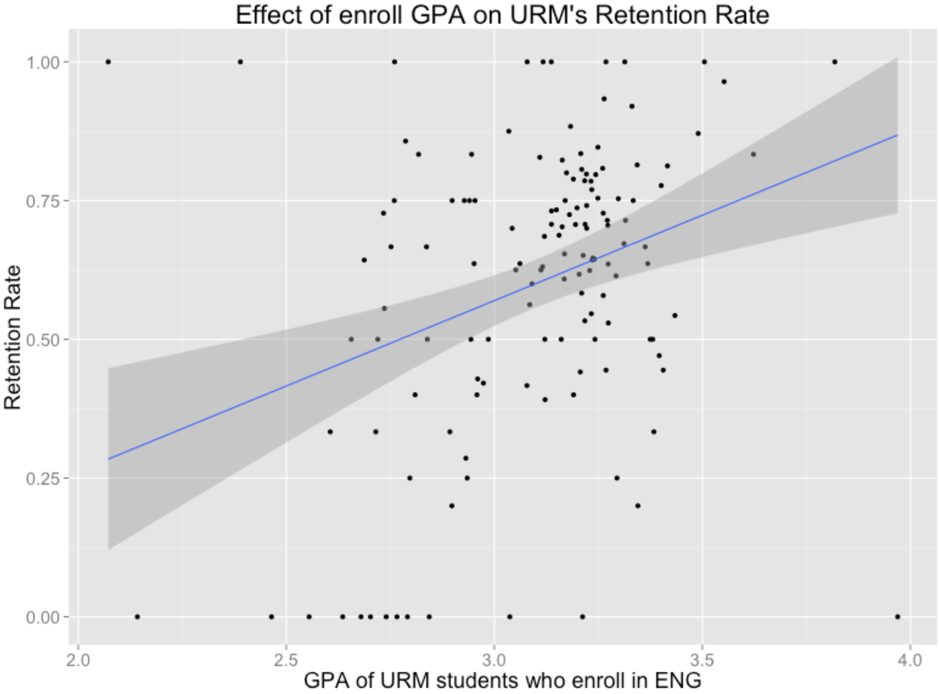
In this part, we are interested in URM students. So this time we subset *URMdata* by Race=URM and obtain new data set called *subURM.*
First we'll explore effect of enroll_GPA in *subURM.*
After checking the correlation between enroll_GPA and other predictors, I built model with enroll_GPA, CohortFirstMajor, CohortTerm. The model is:

Retention Rate ~ enroll_GPA + CohortFirstMajor + CohortTerm

The result of ANCOVA analysis is here:

```
> summary(URMmod_3)

Call:
lm(formula = Retention_Rate ~ enroll_GPA + Term + Major, data = subURM)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.62833    0.24974  -2.516   0.0132 *
enroll_GPA    0.33983    0.08160   4.164    6e-05 ***
Term120048   -0.06377    0.06852  -0.931   0.3540
Term120058    0.04756    0.06724   0.707   0.4808
Term120068    0.06677    0.06664   1.002   0.3184
Term120078    0.16587    0.06694   2.478   0.0146 *
MajorAero     0.13591    0.11509   1.181   0.2400
MajorBioen    0.05294    0.13551   0.391   0.6968
MajorCS       0.18872    0.11510   1.640   0.1038
MajorChem E   0.09897    0.11338   0.873   0.3845
MajorCivil    0.30144    0.11347   2.656   0.0090 **
MajorComp E   0.09776    0.11314   0.864   0.3893
MajorE Mech   0.06403    0.13265   0.483   0.6302
MajorE Phys  -0.03777    0.11976  -0.315   0.7530
MajorEE       0.11679    0.11433   1.022   0.3091
MajorGen Eng  0.19768    0.11440   1.728   0.0866 .
MajorInd Eng  0.04257    0.12495   0.341   0.7339
MajorMatSE    0.18044    0.11452   1.576   0.1178
MajorMech E   0.25043    0.11432   2.191   0.0305 *
MajorNPRE     0.14483    0.11618   1.247   0.2151
```

Estimate of enroll_GPA is significant(with p-value = 6e-15). When there is 0.1 increase in enroll_GPA, we expect female's retention rate will increase 3.40%, which is a dramatic positive effect.



Plot 9. Linear Regression Plot for enroll_GPA and Retention Rate

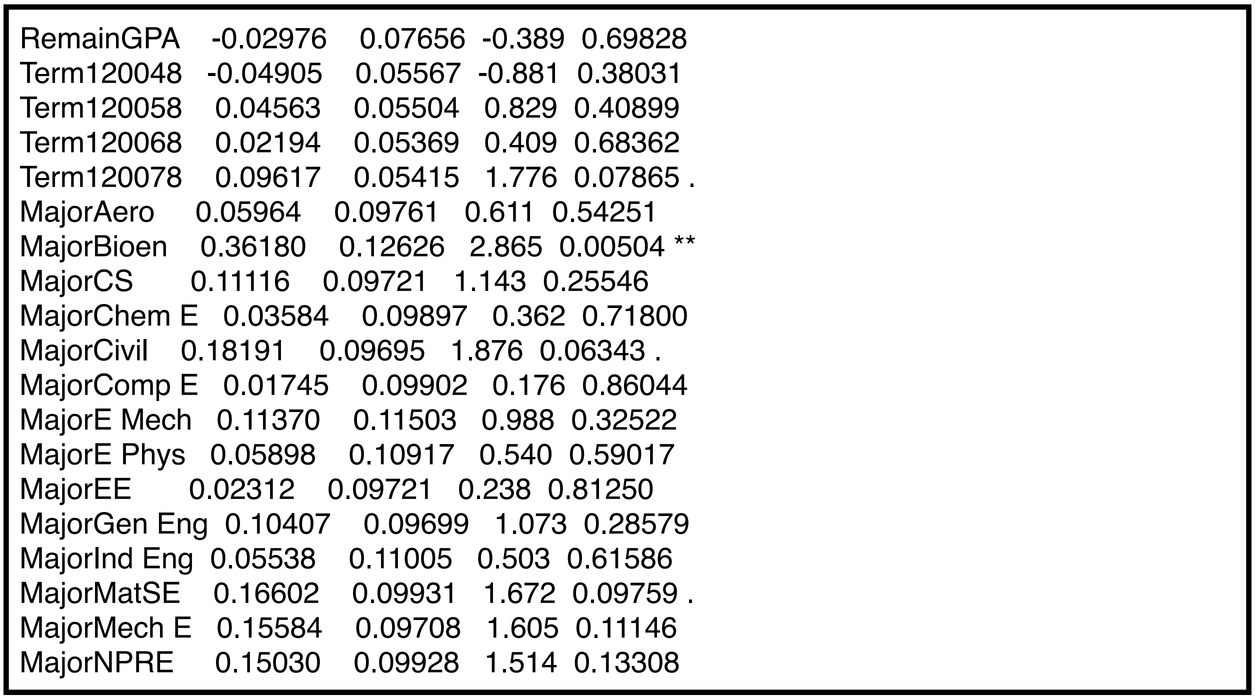
Next let's take a look at the effect of remain_GPA.

After checking the correlation between remain_GPA and other predictors, I built model with remain_GPA, CohortFirstMajor, CohortTerm. The model is:

Retention Rate ~ remain_GPA + CohortFirstMajor + CohortTerm

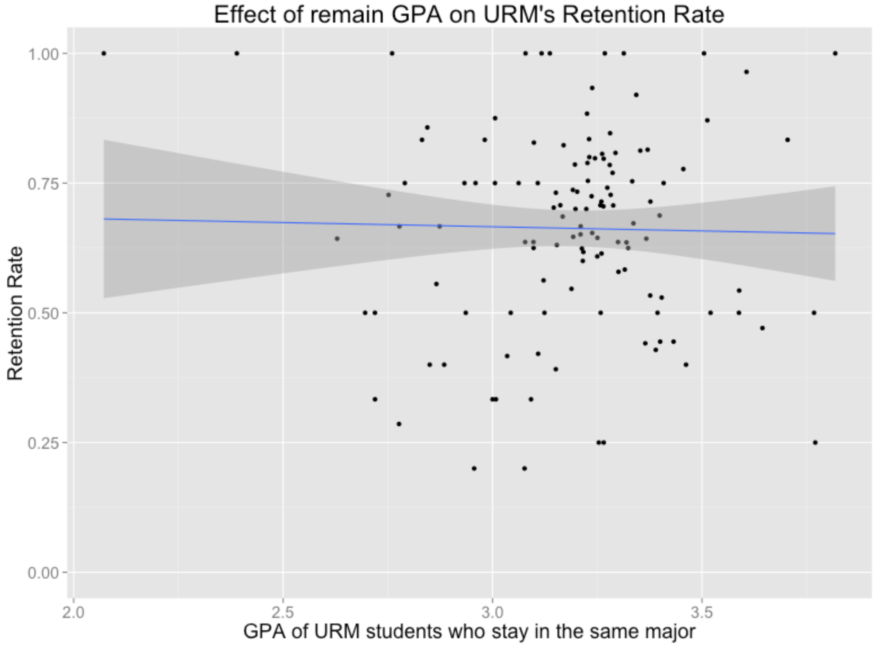
The result of ANCOVA analysis is here:

```
> summary(URMmod_4)

Call:
lm(formula = Retention_Rate ~ RemainGPA + Term + Major, data = subURM)

Coefficients:
            Estimate Std. Error t value Pr(>ltl)
(Intercept)   0.63185    0.25219   2.505  0.01378 *
```



```
RemainGPA    -0.02976    0.07656  -0.389  0.69828
Term120048   -0.04905    0.05567  -0.881  0.38031
Term120058    0.04563    0.05504   0.829  0.40899
Term120068    0.02194    0.05369   0.409  0.68362
Term120078    0.09617    0.05415   1.776  0.07865 .
MajorAero     0.05964    0.09761   0.611  0.54251
MajorBioen    0.36180    0.12626   2.865  0.00504 **
MajorCS       0.11116    0.09721   1.143  0.25546
MajorChem E   0.03584    0.09897   0.362  0.71800
MajorCivil    0.18191    0.09695   1.876  0.06343 .
MajorComp E   0.01745    0.09902   0.176  0.86044
MajorE Mech   0.11370    0.11503   0.988  0.32522
MajorE Phys   0.05898    0.10917   0.540  0.59017
MajorEE       0.02312    0.09721   0.238  0.81250
MajorGen Eng  0.10407    0.09699   1.073  0.28579
MajorInd Eng  0.05538    0.11005   0.503  0.61586
MajorMatSE    0.16602    0.09931   1.672  0.09759 .
MajorMech E   0.15584    0.09708   1.605  0.11146
MajorNPRE     0.15030    0.09928   1.514  0.13308
```

Estimate of remain_GPA is insignificant(with p-value = 0.6982). So remain_GPA may not have effect on URM's retention rate.



Plot 10. Linear Regression Plot for remain_GPA and Retention Rate

3. Conclusions for Part 1

- In overall, there is no significant difference in *Retention Rate* between female students and male students
- But in E Physics and General Engineering, female students have significantly lower *Retention Rate*.
- The proportion of female students has significantly positive effect on female's *Retention Rate*
- GPA(especially *Enroll GPA*) has significantly positive effect on female's *Retention Rate*
- In overall, NonURM students have significantly higher *Retention Rate* than URM students.
- So far, we can conclude that in CS, EE, Mechanical Engineering, NonURM students have significantly higher *Retention Rate* than URM students.
- GPA(especially *Enroll GPA*) has significantly positive effect on URM's *Retention Rate*