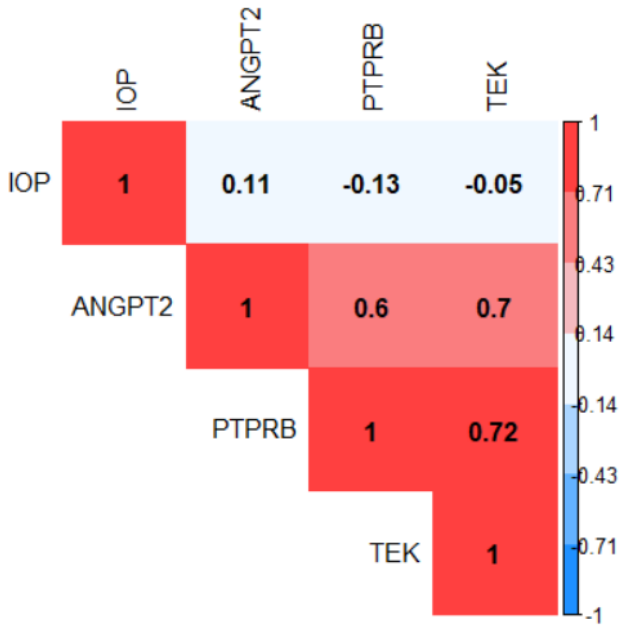
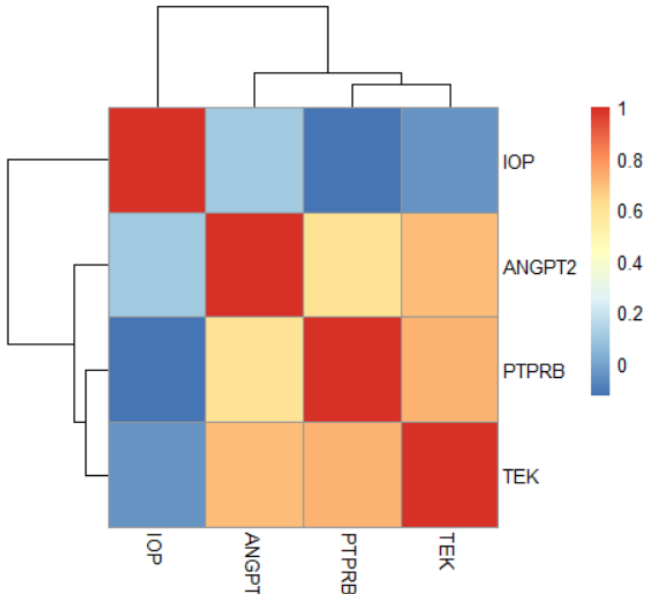
Report date: March 20	Heatmap and Distribution	No Normalization (Real Counts)

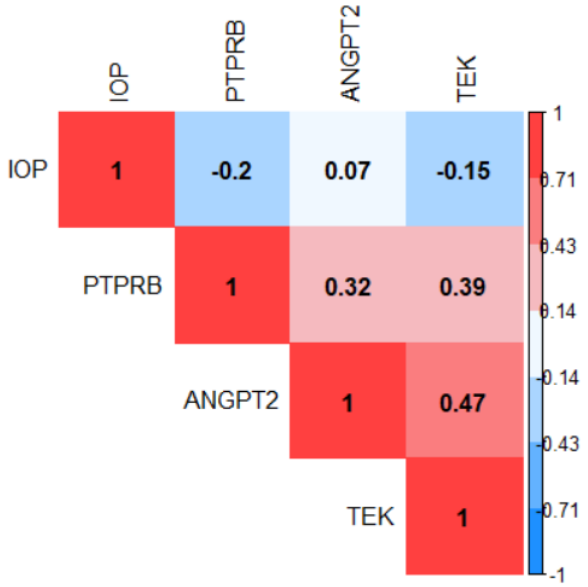
Correlation method: Pearson



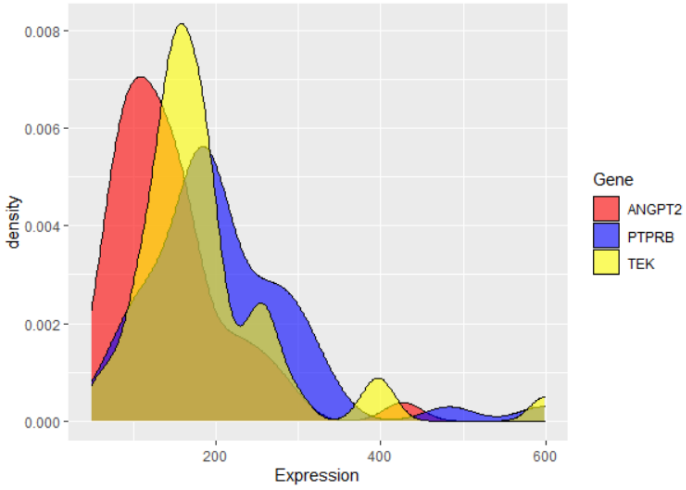
Clustering



Correlation method: Spearman

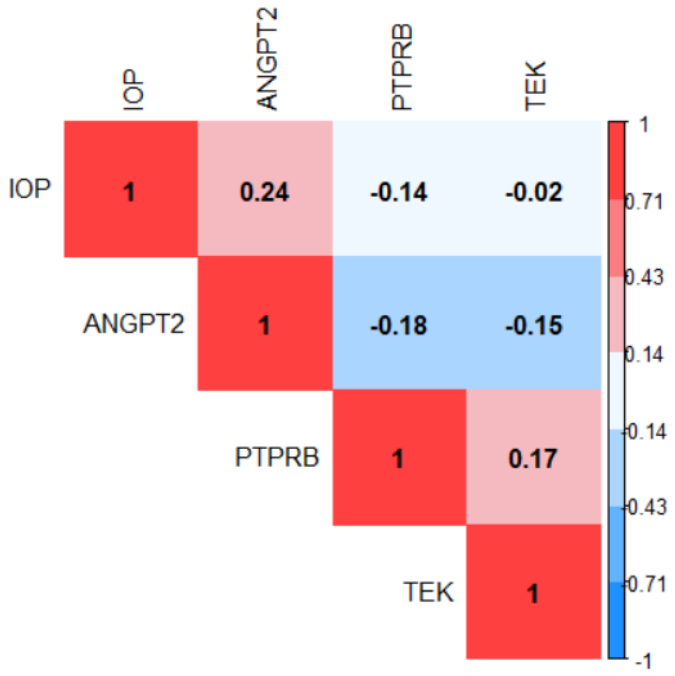


Distribution

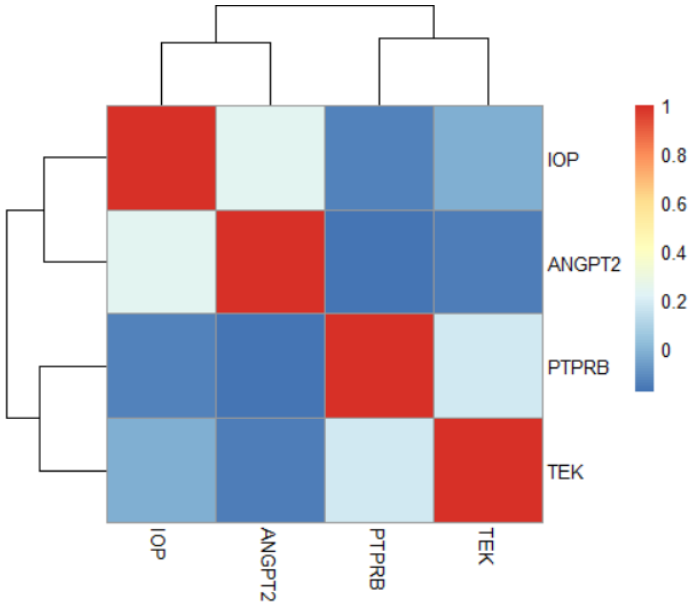


Report date: March 20	Heatmap and Distribution	Normalization method: Log2

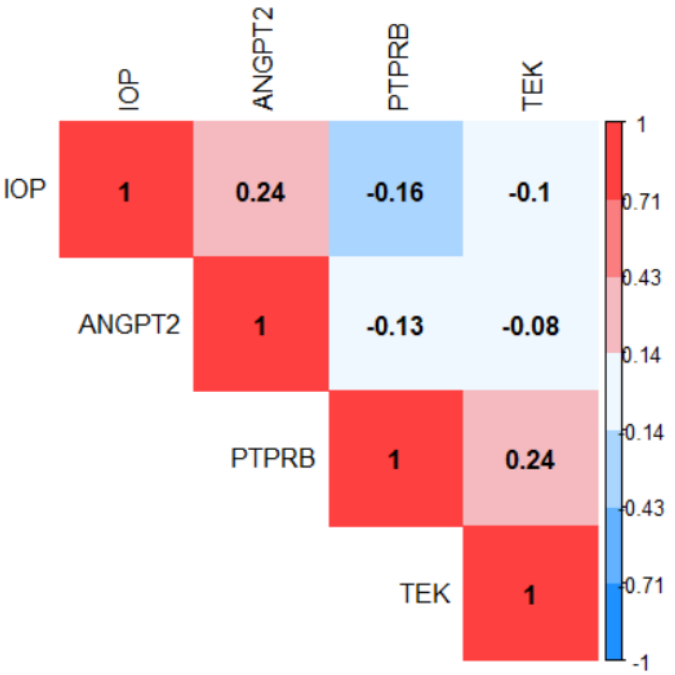
Correlation method: Pearson



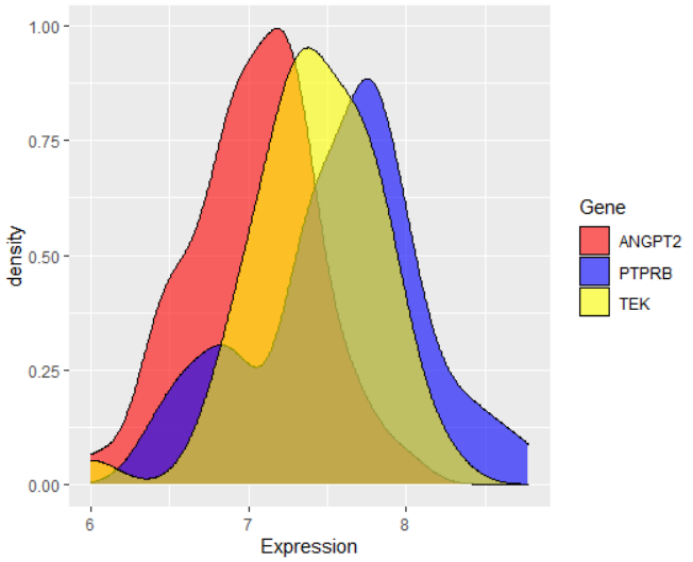
Clustering



Correlation method: Spearman

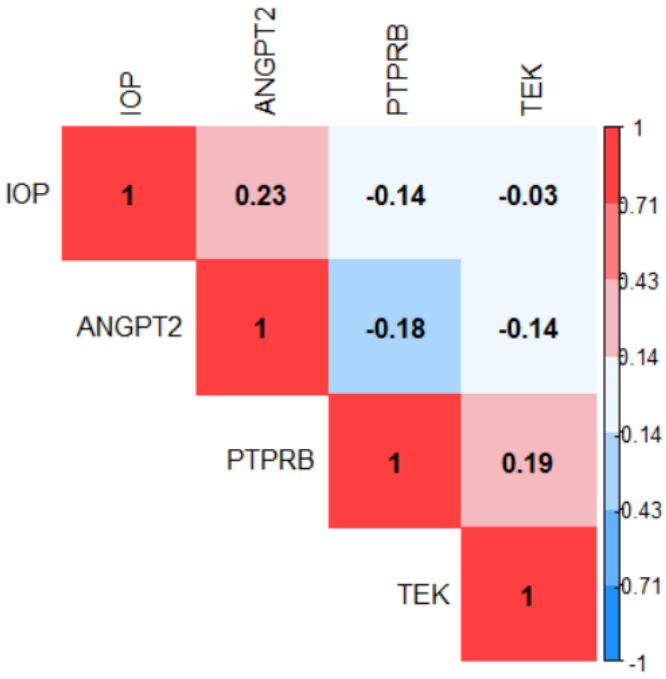


Distribution

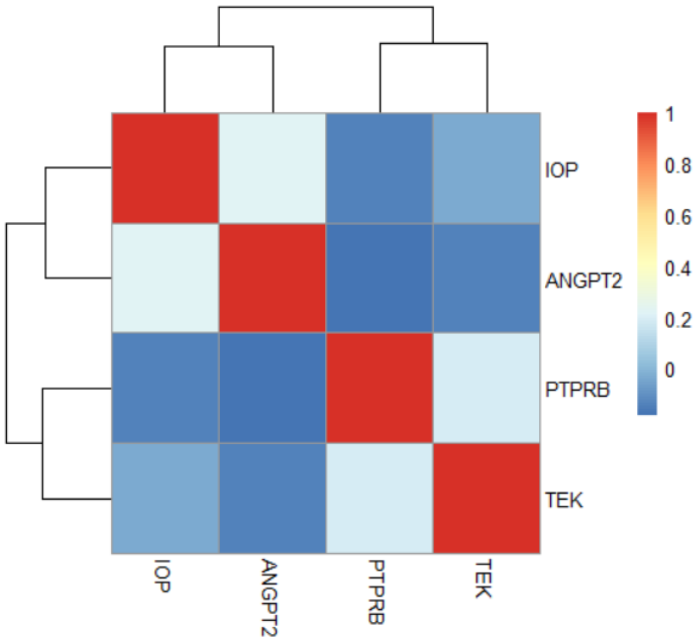


Report date: March 20	Heatmap and Distribution	Normalization method: vst

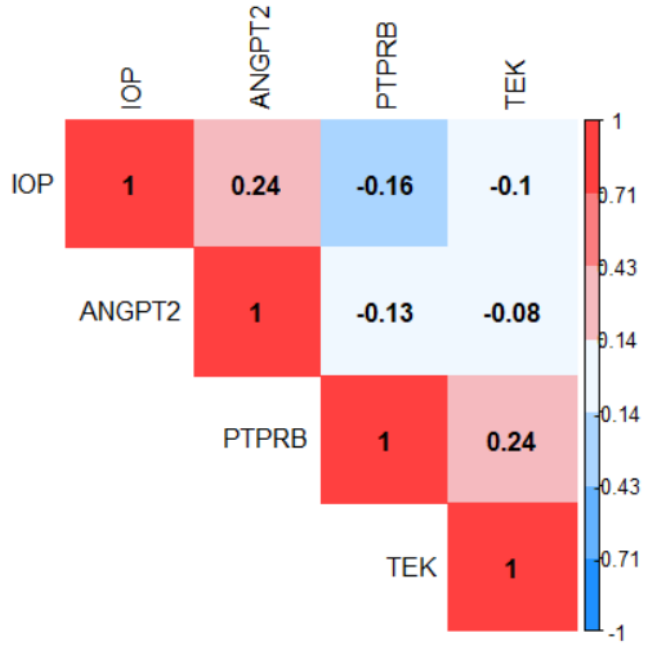
Correlation method: Pearson



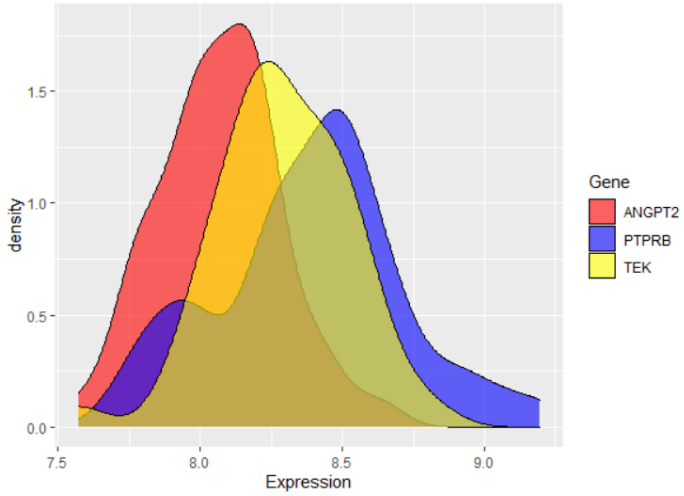
Clustering



Correlation method: Spearman

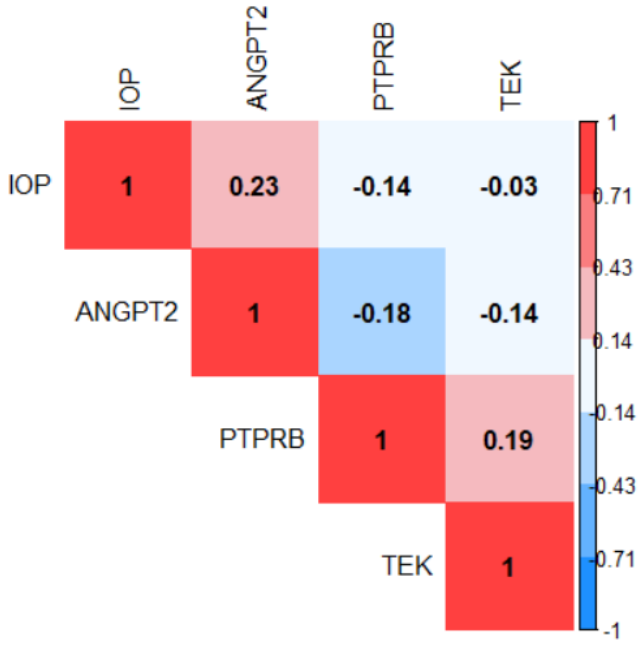


Distribution

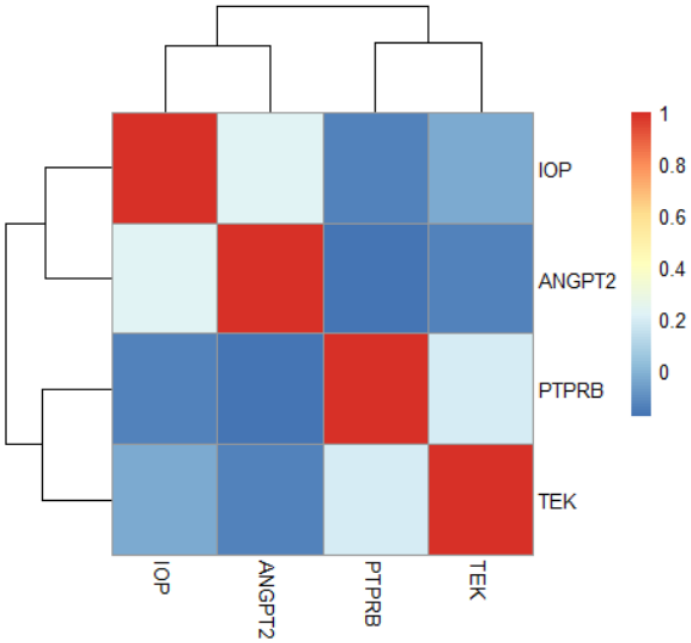


Report date: March 20	Heatmap and Distribution	Normalization method: rlog

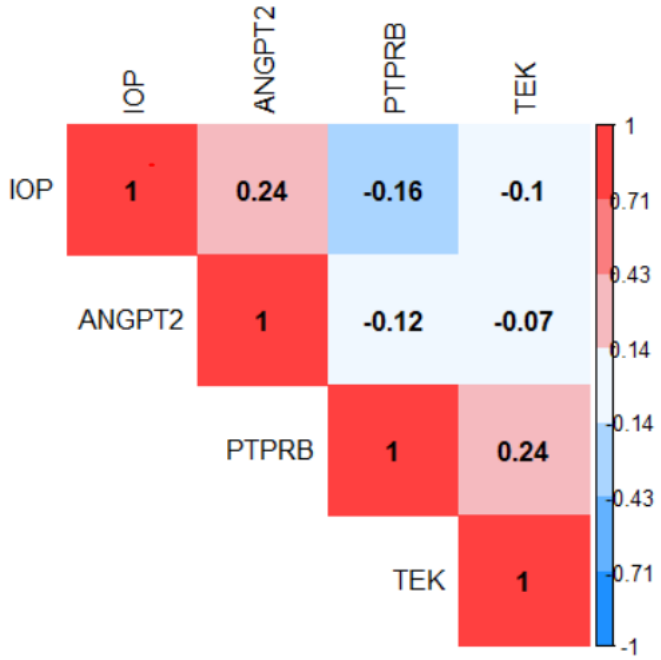
Correlation method: Pearson



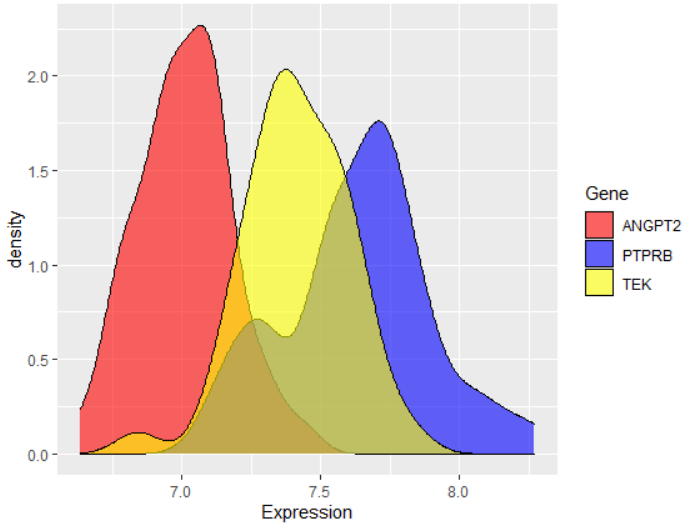
Clustering



Correlation method: Spearman



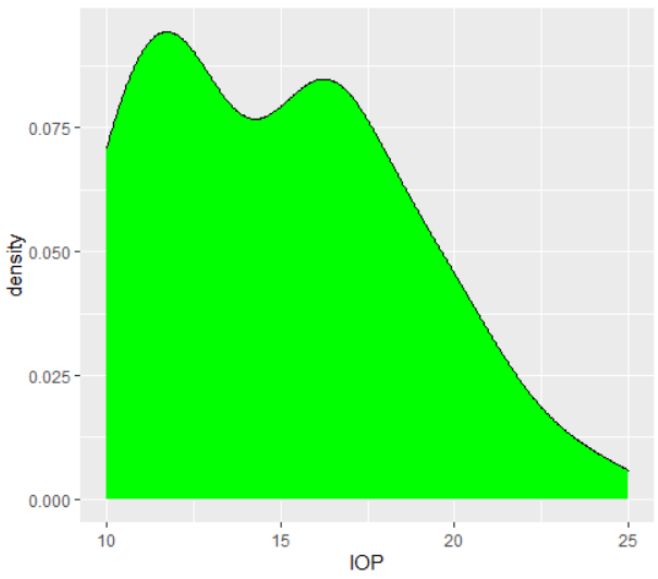
Distribution

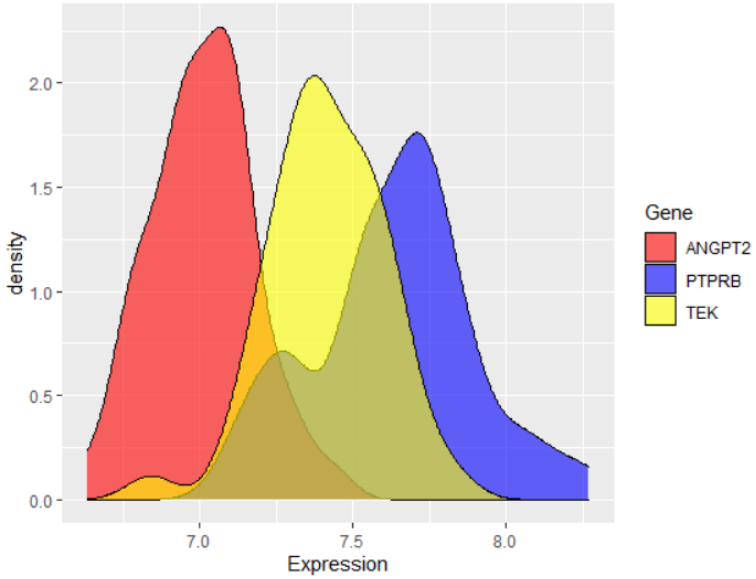


Heatmap figures results

Based on the Clustering and Heatmap figures, I found that two genes **PTPRB** and **TEK** act similarly. They have a negative correlation with **IOP** and **ANGPT2**. Also, **IOP** has a weak positive relationship with **ANGPT2**. Compare three normalization methods: **Log2** (Log2 Normalized Counts Transformation), **rlog** (Regularized Log Transformation), and **vst** (Variance Stabilizing Transformation), and based on the distribution plots of three genes, it shows that **rlog** normalization gives good performance for our analysis in this step. Therefore, I consider the **rlog** function as a normalization method. Also, between Pearson (linear correlation) method and Spearman (linear-skewed correlation) method, I found that the probability density function of TEK (yellow) is between ANGPT2 (red) and PTPRB (blue). It means that there is an important role or relation between ANGPT2 and TEK. I think Spearman correlation is better than Pearson and states that ANGPT2 and TEK are almost uncorrelated. I mean, I believe that ANGPT2 and PTPRB have a negative relationship, PTPRB and TEK have a positive correlation, and ANGPT2 and TEK are uncorrelated. Moreover, based on the density function of IOP, in the figure below, there is an almost negative correlation between the two density functions of IOP and PTPRB.

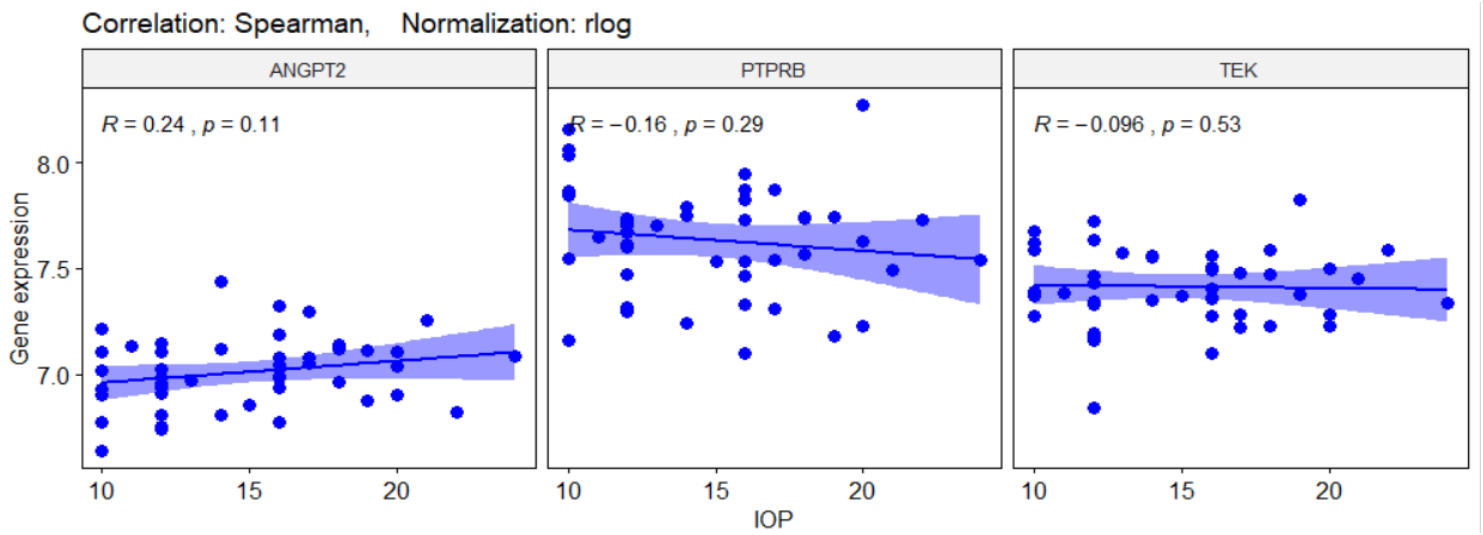
Therefore, I consider **rlog** as a normalization method and **Spearman** as a correlation method for further analysis.



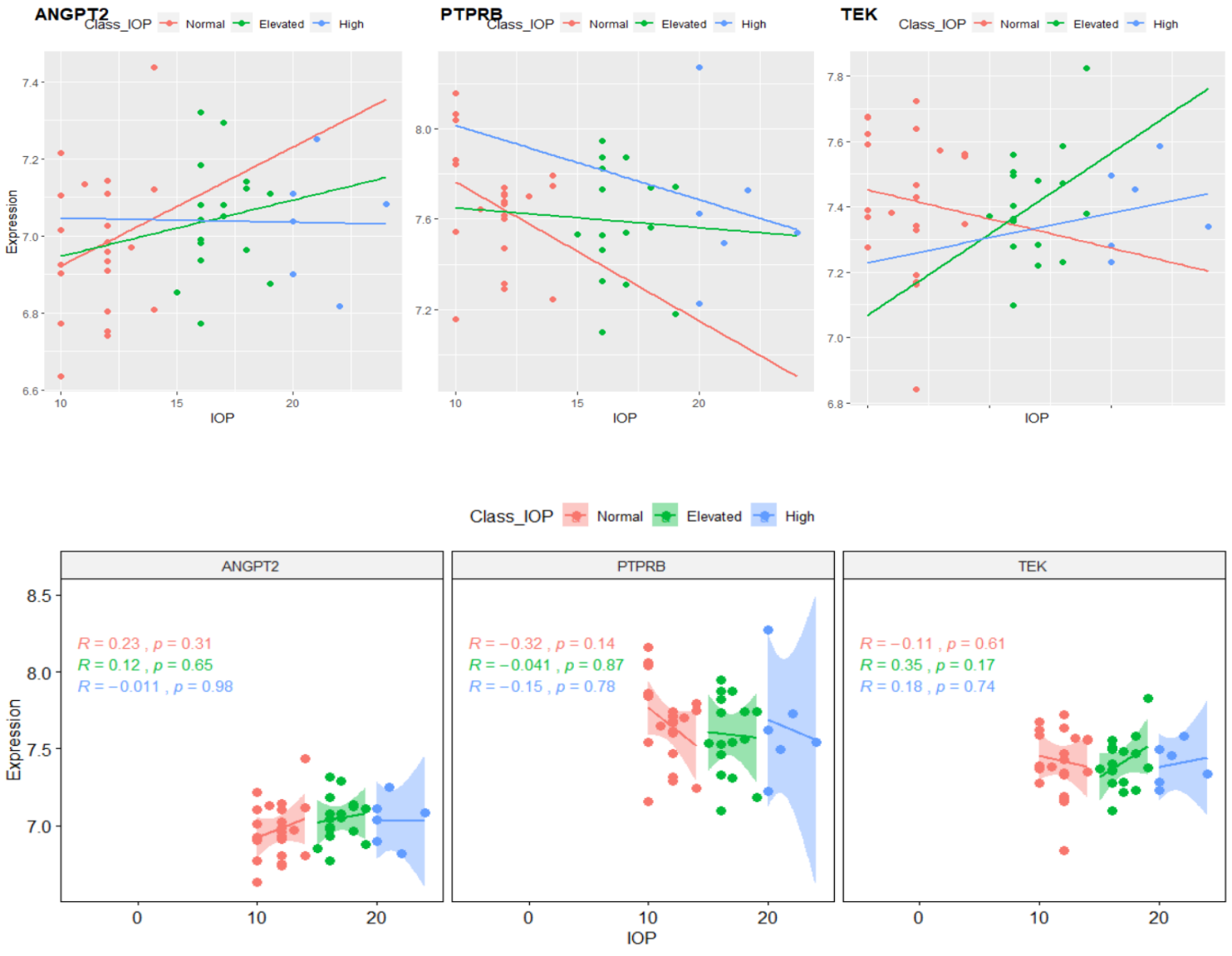


Correlation between IOP and ANGPT2

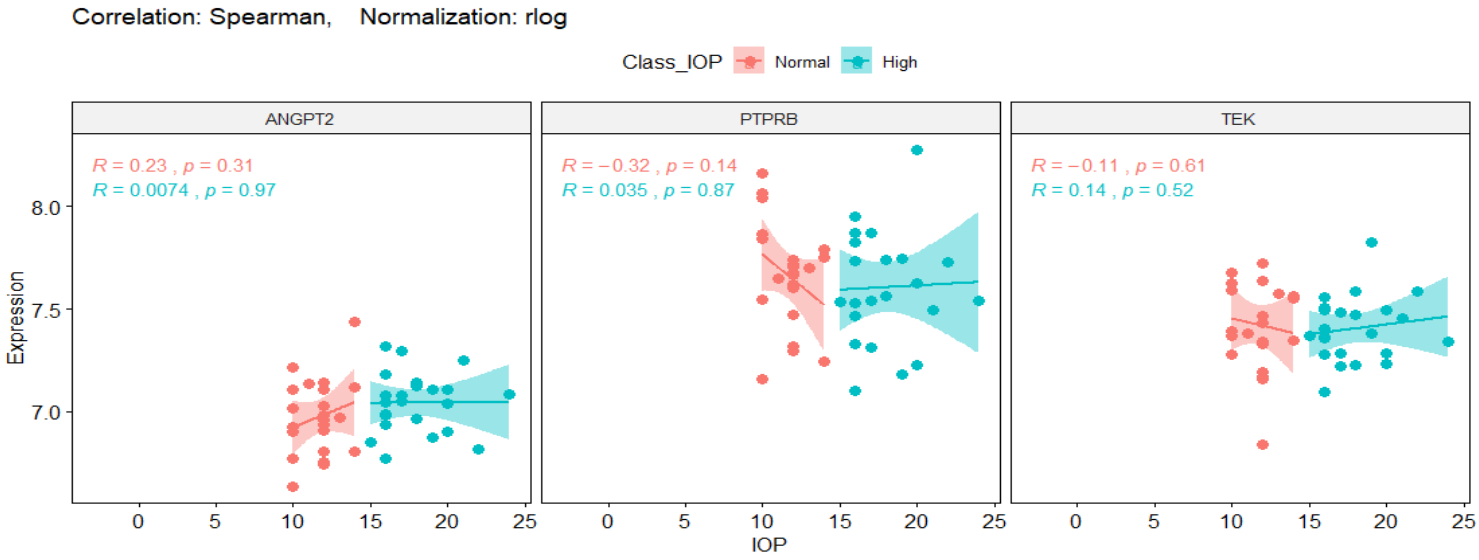
In this section, I want to calculate the **Spearman** correlation of IOP (and all subgroups of IOP) with ANGPT2 gene expression (based on the **rlog** normalization method). Based on the correlation coefficient R and P values in the figure below, I found that IOP has a positive relationship with ANGPT2 and a negative relationship with PTPRB and almost not correlated with TEK. However, none of them are significantly correlated. Based on the best R and P-values, I focused on the week relationship of IOP and ANGPT2.



In the two figures below, I plot the correlation of three genes with three subgroups of IOP. As I can see, ANGPT2 has the highest positive correlation with Normal_IOP (R = 0.23) subgroup. Also, PTPRB has the highest negative correlation with Normal_IOP (R = -0.32) subgroup. It means that the Normal subgroup has a major role in the IOP feature.



Although I am looking for a high correlation between the High_IOP subgroup and ANGPT2, however, as there are low samples in the high_IOP subgroup, it is not a significant relationship. Therefore, I decided to classified the IOP feature into two subgroups: Normal and High. In the figure below, I look for the correlation of IOP subgroups with three genes. Again, here there are more samples with High_IOP; around 50% of samples (23 samples). However, the High_IOP subgroup is not correlated with ANGPT2 and PTPRB. I mean, Normal_IOP has important role in the correlation of IOP with ANGPT2 or PTPRB.



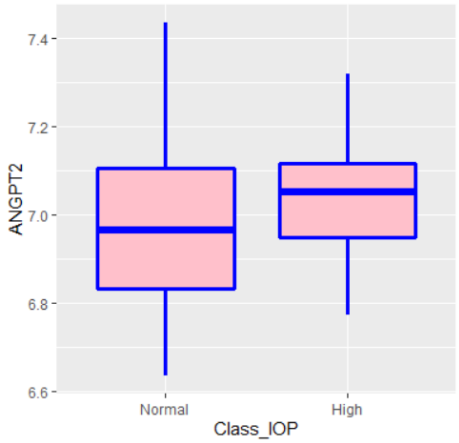
As a conclusion and based on figures above, ANGPT2 has week correlation, but not significant, with whole IOP data (R=0.24, P-value=0.11), which related to the Normal_IOP (R=0.23, P-value=0.31). However, the ANGPT2 gene has the maximum (week) correlation with IOP among other genes. Therefore, we need to look at this gene deeply. Maybe other features like Age, Sex, or Batch give some clues. To discuss statistically about the correlation between ANGPT2 and IOP subgroups, I did t-test and ANOVA analysis too.

Analysis of t-test for IOP subgroups

In this analysis, I want to test that the mean of expression level of ANGPT2 is equal in two separated groups of Normal_IOP and High_IOP? (with almost same sample sizes). If not, then the expression level of ANGPT2 in each subgroup of IOP are not similar together, and the expression level are varied in different subgroupe of IOP.

H0: The mean of ANGPT2 expression level in Normal_IOP group is equal to the mean of ANGPT2 expression level in High_IOP group.

Ha: Both means are not equal.



The results of two sided t-test with 95% confidential is:

```
t.test(ANGPT2 ~ Class_IOP, data=corrTable_IOP2, mu=0, alt="two.sided", conf=0.95, var.eq=F, paired=F)
```

```
        Welch Two Sample t-test

data:  ANGPT2 by Class_IOP

t = -1.4124, df = 39.885, p-value = 0.1656
alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:
 -0.1705142  0.0302361

sample estimates:
mean in group Normal   mean in group High
            6.973136             7.043275
```

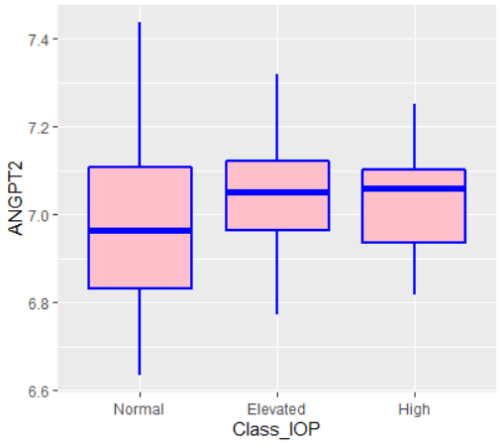
The P-value of the test is 0.1656, which is greater than the significant level alpha = 0.05. Therefore, I cannot reject the Null-Hypothesis and accept the alternative hypothesis. I conclude that the average expression level of ANGPT2 gene in samples with Normal_IOP is not significantly different with the samples with High_IOP; the expression level of ANGPT2 gene are almost same in two gropus. It means that different IOP subgroups have not important role in the expression level of ANGPT2.

Analysis of ANOVA for IOP subgroups

As in the real state, we have three subgroups of IOP, so we need to run ANOVA analysis insteed of t-test for them.

H0: The mean of ANGPT2 expression level in all three groups of IOP (Normal, Elevated, and High) is equal together.

Ha: They are not equal.



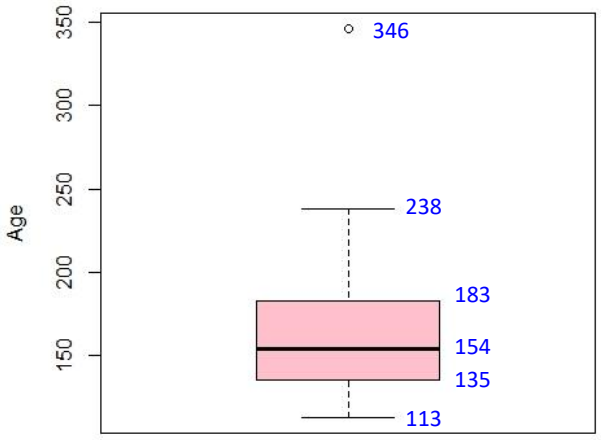

```
Anova_results <- aov(ANGPT2 ~ Class_IOP, data=corrTable)
summary(Anova_results)
```

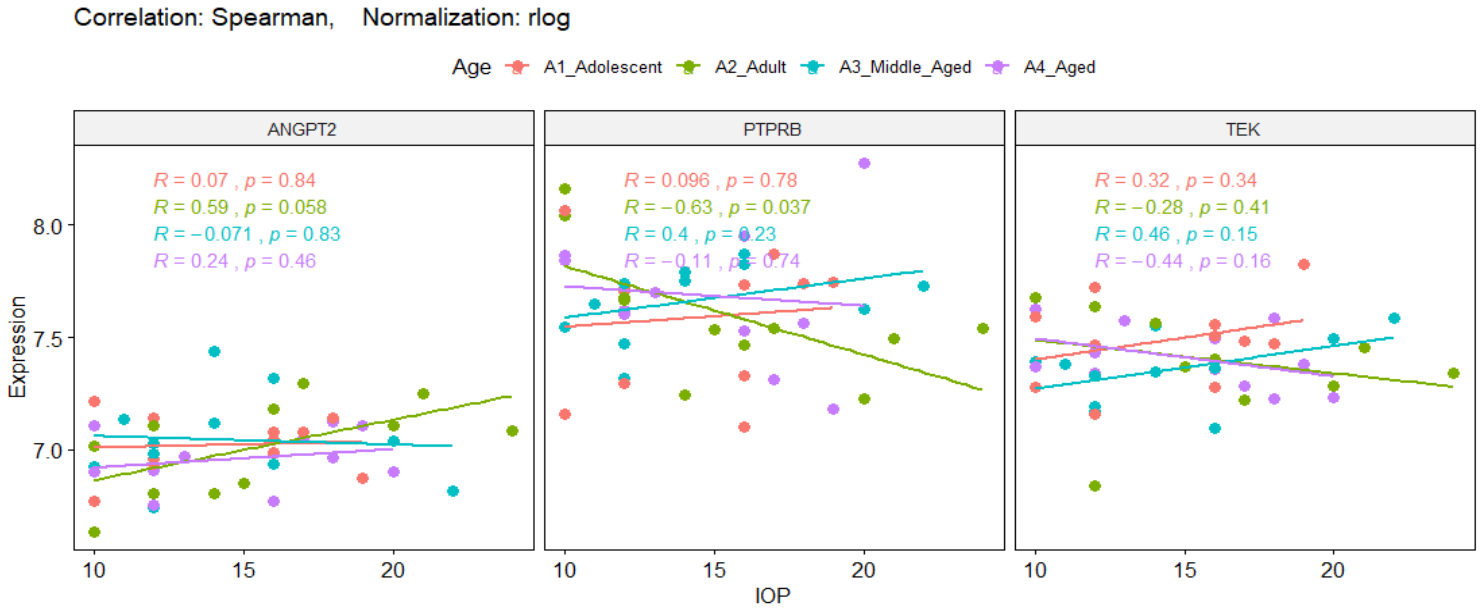
```
             Df Sum Sq Mean Sq F value Pr(>F)
Class_IOP     2 0.0561 0.02806       1  0.377
Residuals    42 1.1789 0.02807

F(2,42) = 1,    P-value>0.05
```

The P-value is greater thatn 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claime that the expression level of ANGPT2 is significantly varied in three different IOP groups.

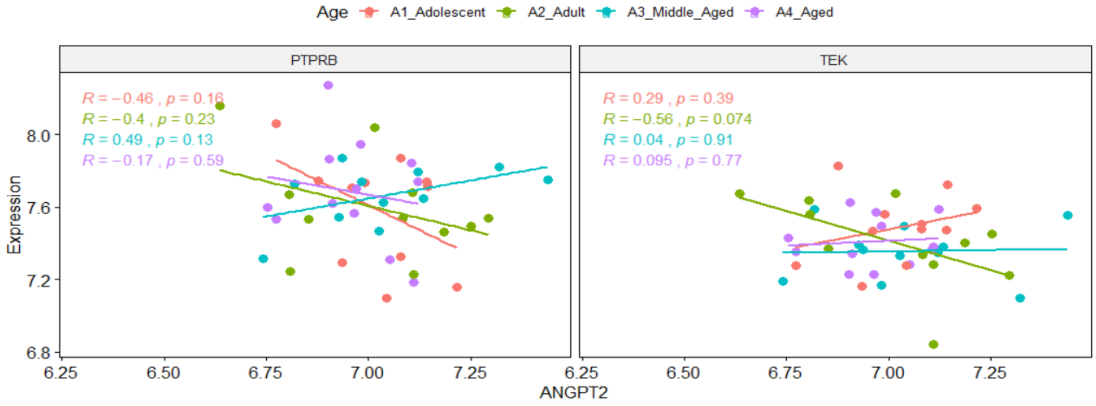
Partial-Correlation between IOP and ANGPT2 based on Age feature

In this section, I am going to calculate the correlation between IOP and ANGPT2 based on different Age groups. I classified samples into four subgroups: **Adolescent** (113<=Age<135), **Adult** (135<=Age<154), **Middle_Aged**(154<=Age<183), and **Aged** (183<=Age<=346). Then, plot the correlation based on each subgroups:





Good news: Based on the results, I can see there is a significant relationship in the Adult subgroup between IOP and ANGPT2 (R=+0.59, P-value=0.05), and PTPRB (R=-0.63, P-value=0.03). Also, there is semi-significant relationship between ANGPT2 and TEK genes in this subgroup (Adult); figure below.

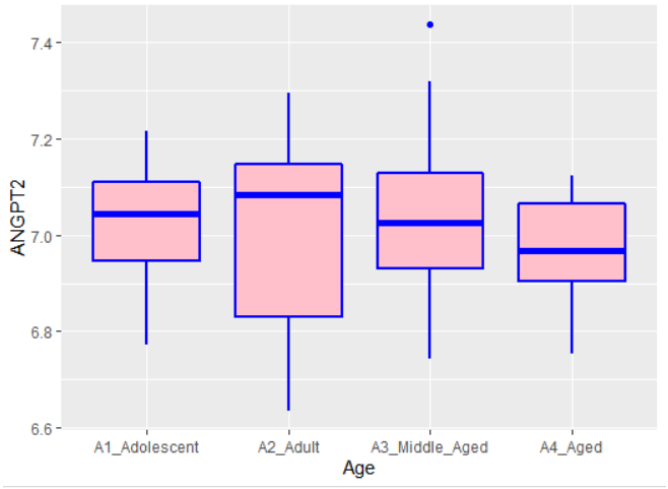


Analysis of ANOVA for Age subgroups

There is four subgroups of Age, and I run ANOVA analysis in for ANGPT2 expression along these subgroups.

H0: The mean of ANGPT2 expression level in all four subgroups of Age is equal together.

Ha: They are not equal.

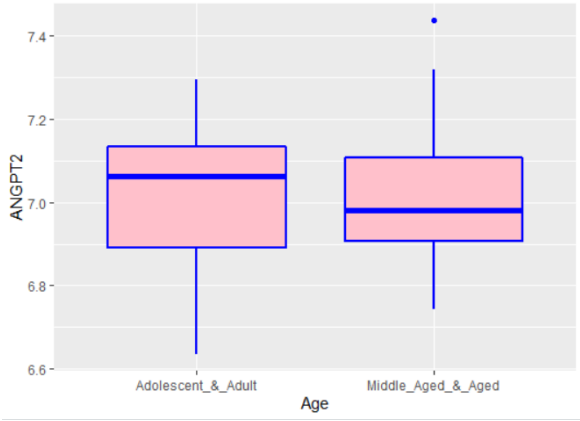


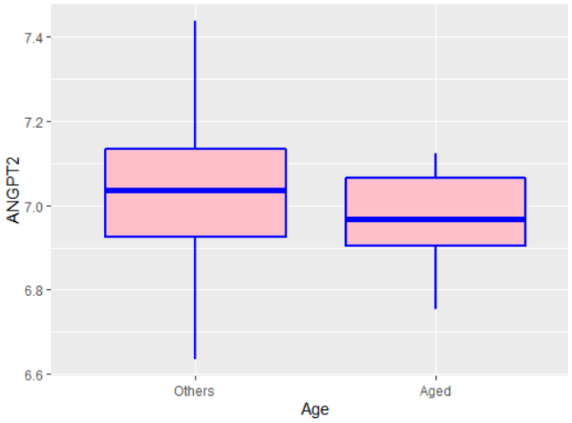

```
Anova_results <- aov(ANGPT2 ~ Age, data=corrTable)
summary(Anova_results)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Age           3 0.0415 0.01384   0.476  0.701
Residuals    41 1.1935 0.02911

F(3,41) = 0.476,    P-value > 0.05
```

The P-value is greater thatn 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claime that the expression level of ANGPT2 is significantly varied in four different Age groups.

Then, I decided to devide the Age into just two subgroups as belows. However, in none of them I did not get significant P-value in ANOVA analysis.






```
Anova_results <- aov(ANGPT2 ~ Age, data=corrTable_Age2)
summary(Anova_results)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1  0.003 0.002999   0.105  0.748
Residuals    43  1.232 0.028651

F(1,43) = 0.105,    P-value > 0.05
```

```
Anova_results <- aov(ANGPT2 ~ Age, data=corrTable_Age2)
summary(Anova_results)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1 0.0361 0.03613   1.296  0.261
Residuals    43 1.1989 0.02788

F(1,43) = 1.296,    P-value > 0.05
```

Partial-Correlation between IOP and ANGPT2 based on Sex feature