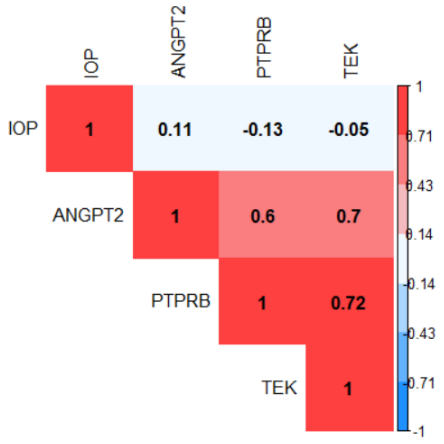
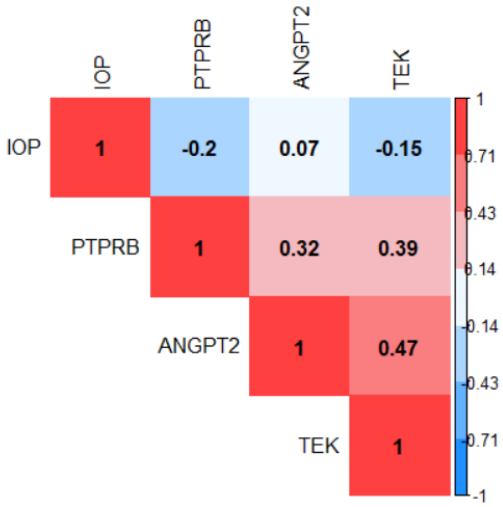
Normalization of Genes: Real counts

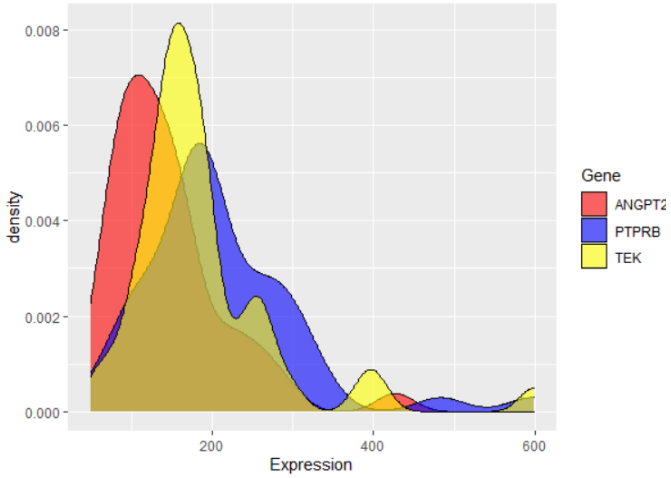
Correlation method: Pearson



Correlation method: Spearman

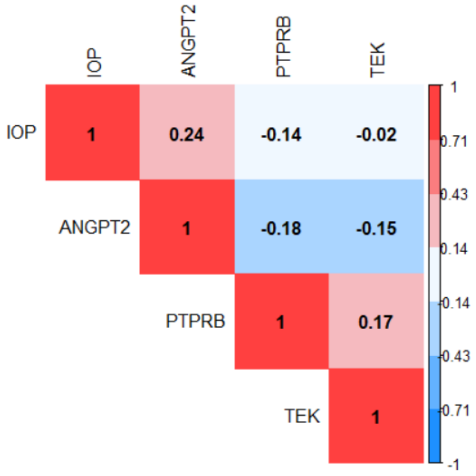


Distribution

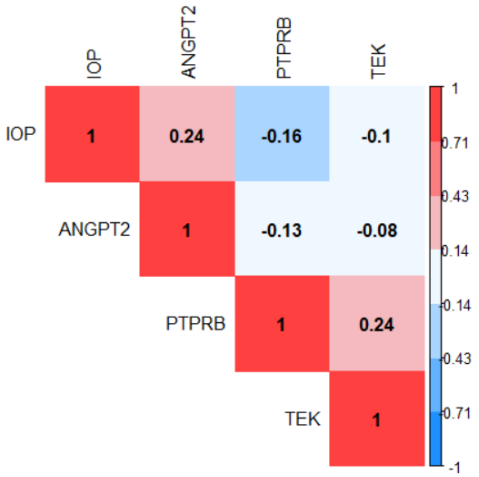


Normalization of Genes: Log2

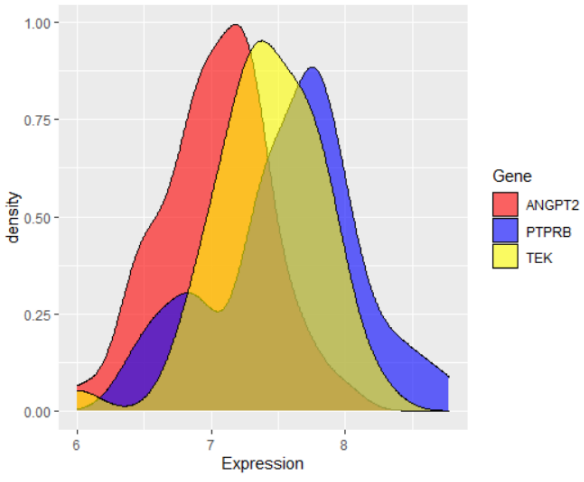
Correlation method: Pearson



Correlation method: Spearman

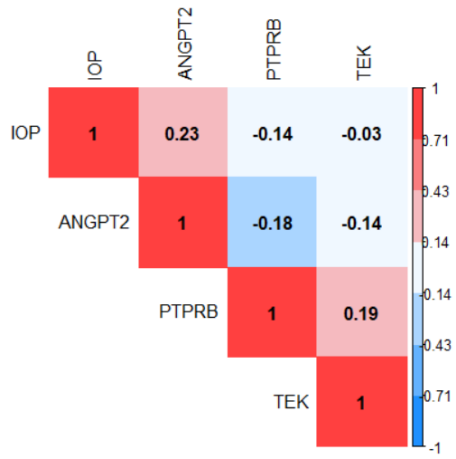


Distribution



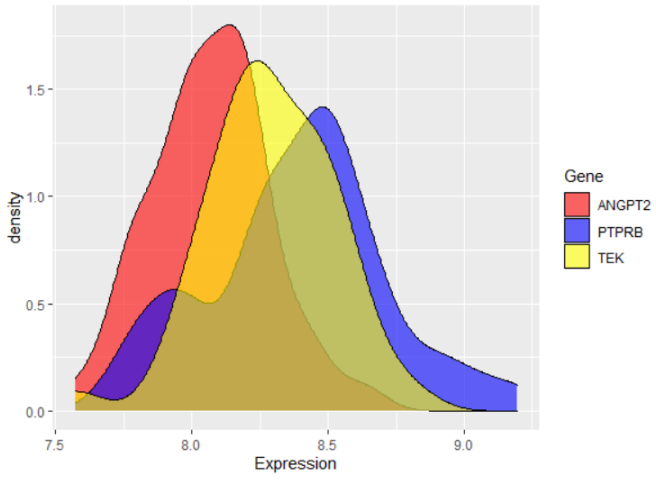
Normalization of Genes: vst

Correlation method: Pearson



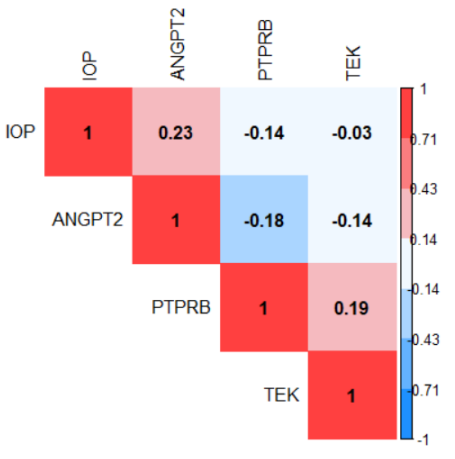
Correlation method: Spearman

Distribution

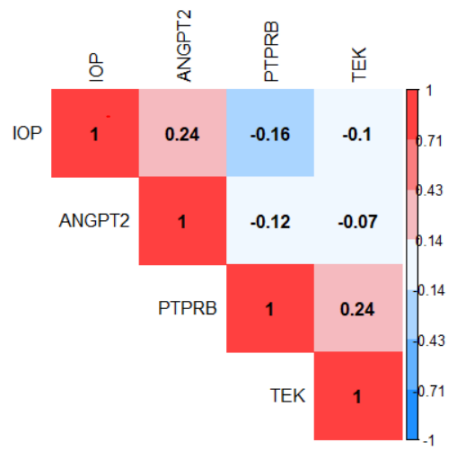


Normalization of Genes: rlog

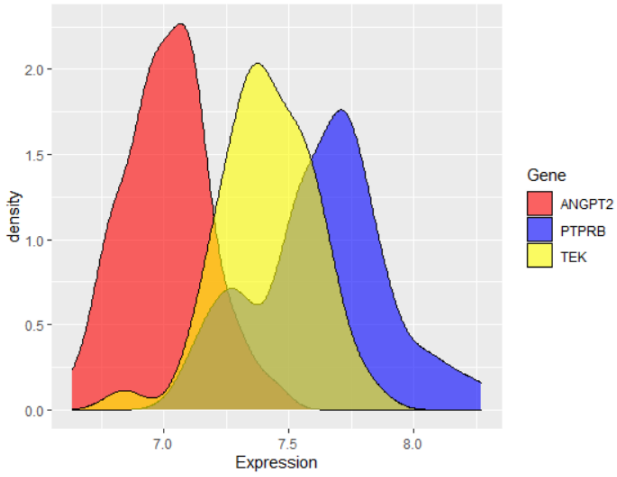
Correlation method: Pearson



Correlation method: Spearman



Distribution



Heatmap figures results

- Based on the heatmap figures above, I found the two genes **PTPRB** and **TEK** act similarly. They have a negative correlation with **IOP** and **ANGPT2**. Also, **IOP** has a weak positive linear 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.

```
dds <- DESeqDataSetFromTximport(txi.rsem, samples, design=~Sex+Batch+Age.scaled)
rld <- rlog(dds, blind=FALSE)
```

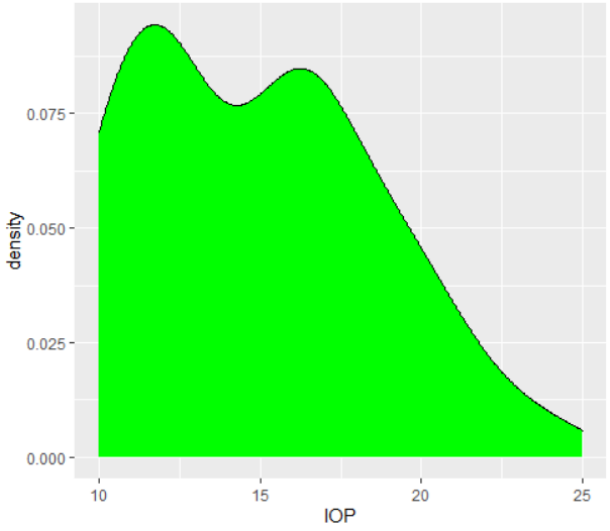
- Also, between Pearson (linear correlation) method and Spearman (linear-skewed correlation) method, I found that the probability density function of TEK is between ANGPT2 and PTPRB. Also, there is a high overlap between TEK and PTPRB. It means that there is a strong relation between PTPRB and TEK. I think the correlation of Spearman is better than Pearson. I believe that ANGPT2 and PTPRB have a negative relationship, and PTPRB and TEK have a positive correlation. So, I choose the **Spearman** method for the subsequent analysis.
- Based on the density function of IOP, in the figure below, I found that maybe its distribution is not normal, and I need to normalize it first, then compare it with the other three genes. There are several ways to check the normality of distribution, like plotting the probability density function (like the figure below), plotting the quartile-quartile plot (Q-Q plot), and do Shapiro-Wilk normality test. Among them, I preferred the reliable Shapiro-Wilk normality test. In the output of this test, the p-value > 0.05, implying that the distribution of the data is not significantly different from the normal distribution. In other words, I can assume the normality for p-value > 0.05 in this test. After running this test for three normalized genes and IOP data, I found that for sure, the IOP data is not a normal distribution; it is right-skewed data. There are some common transformations for normalization of this type of data like **square root**, **cube root**, and **log** functions.

ANGPT2
W = 0.9904, p-value = 0.9686
p-value > 0.05 then ANGPT2 is normal

PTPRB
W = 0.97918, p-value = 0.587
p-value > 0.05 then PTPRB is normal

TEK
W = 0.98158, p-value = 0.6841
p-value > 0.05 then TEK is normal

IOP
W = 0.93549, p-value = 0.01474
p-value < 0.05 then IOP is not normal



Normalization of IOP

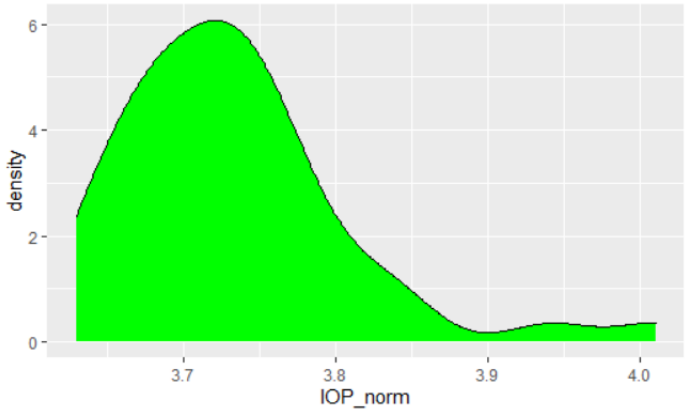
In this section, I am going to normalize the IOP with different methods:

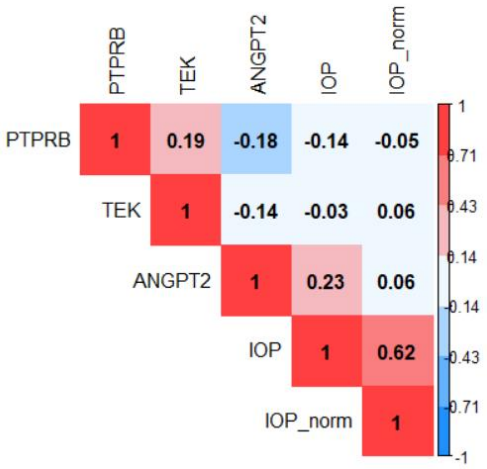
1) rlog

First, I tried to normalize the IOP with other 32883 genes and used the logarithmic geometry mean method. But, after checking the Shapiro-Wilk normality test, I found that the output data is not normal distribution again. It has some skewness in the right side. So, as I thought, this is not a suitable method for normalization, because of mixing IOP data with gene expression data. If you look at the correlation table, the correlation between IOP and IOP_normal is 0.62. It means that the nature of original IOP data is changed during the normalization.

`W = 0.88815, p-value = 0.0004137`

`p-value < 0.05 then IOP_norm is not normal`





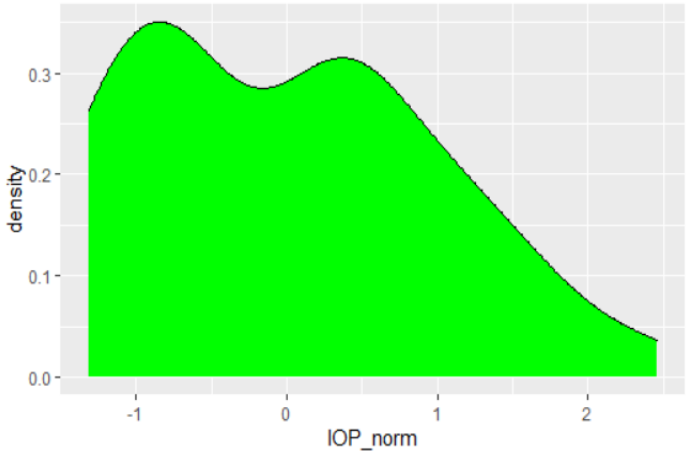
2) Standard method (x - mean(x)) / std(x)

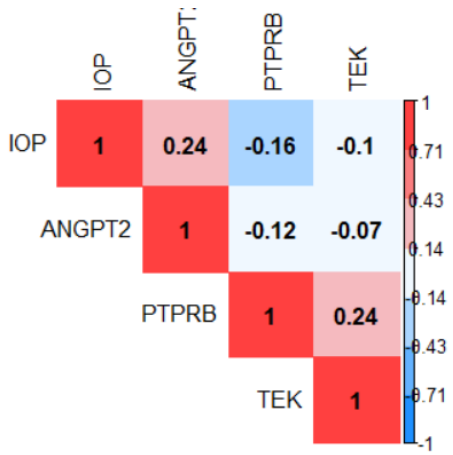
Then uses the standard method for normalization of IOP data to convert its distribution into the normal distribution.

`W = 0.93549, p-value = 0.01474`

`p-value < 0.05 then IOP is not normal`

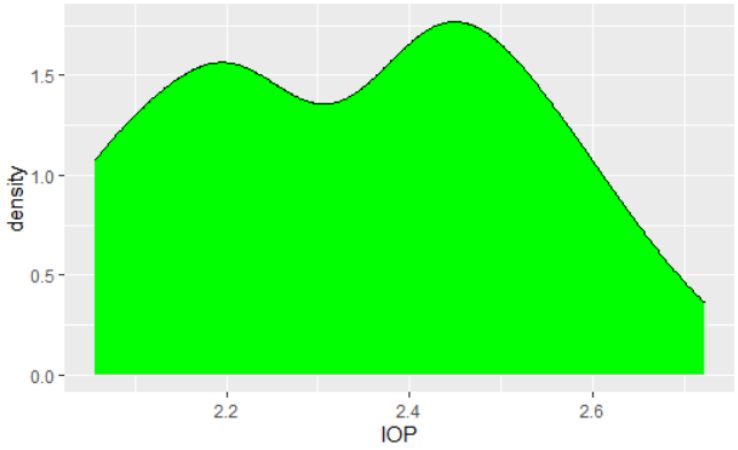
`skewness(corrTable$IOP) = 0.4156063`

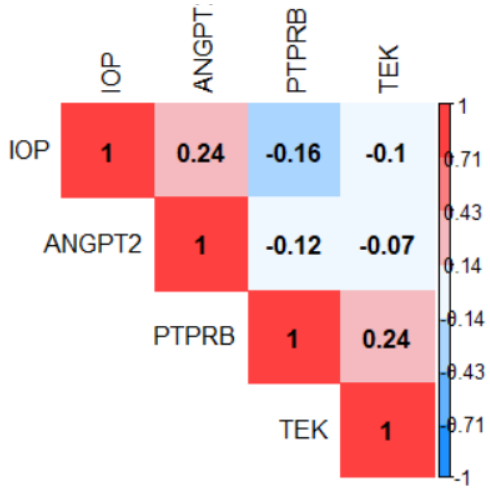




3) Box-Cox power transformation **Lambda = -0.1**

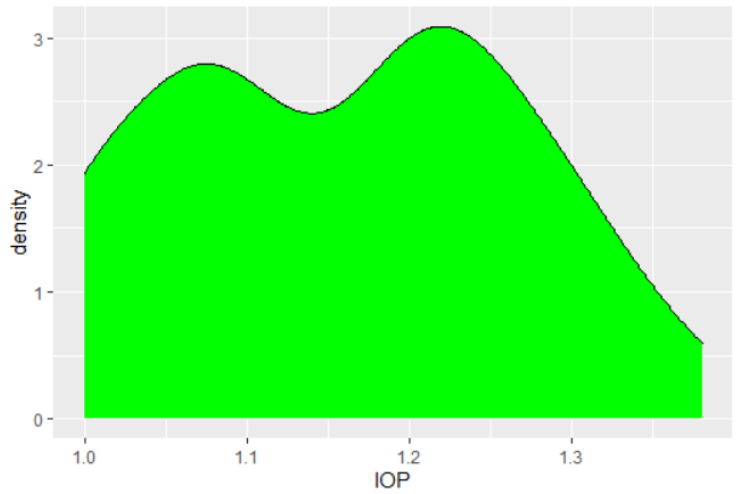
W = 0.93785, p-value = 0.01795
p-value < 0.05 then IOP is not normal
skewness(corrTable$IOP) = 0.01833074

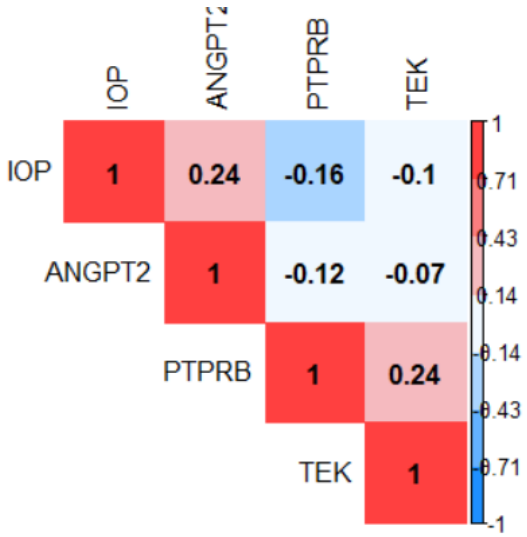




4) Log transformation **Log(x), Log2(x), Log10(x)**

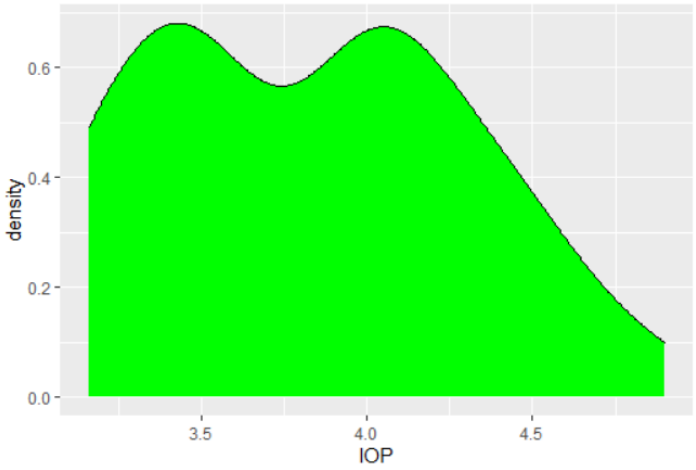
W = 0.93874, p-value = 0.01934
p-value < 0.05 then IOP is not normal
skewness(corrTable$IOP) = 0.05192562

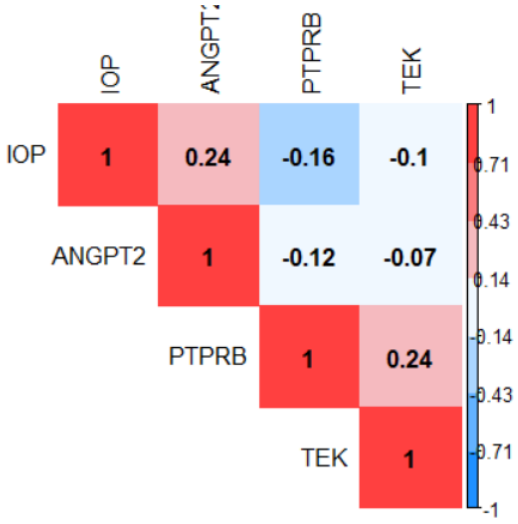




5) Square root transformation sqrt(x)

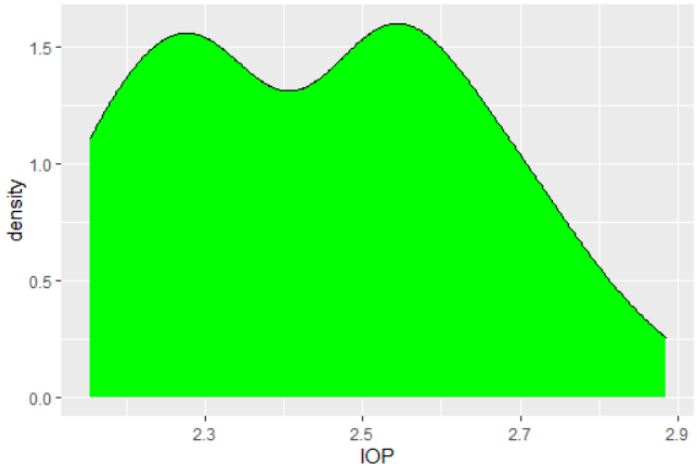
W = 0.93998, p-value = 0.02147
p-value < 0.05 then IOP is not normal
skewness(corrTable$IOP) = 0.2272393

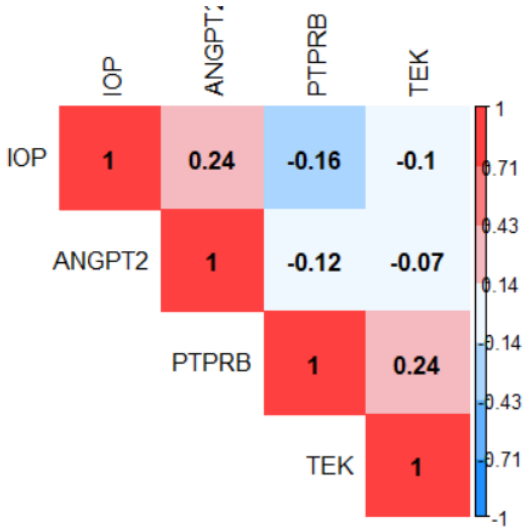




6) Cube root transformation cuberoot(x)

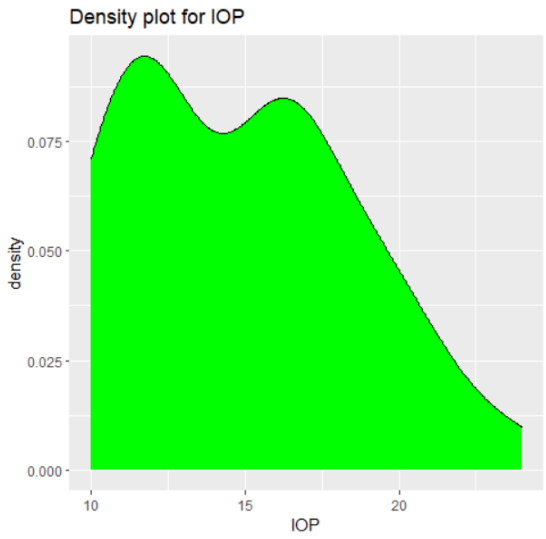
W = 0.94018, p-value = 0.02184
p-value < 0.05 then IOP is not normal
skewness(corrTable$IOP) = 0.1674028

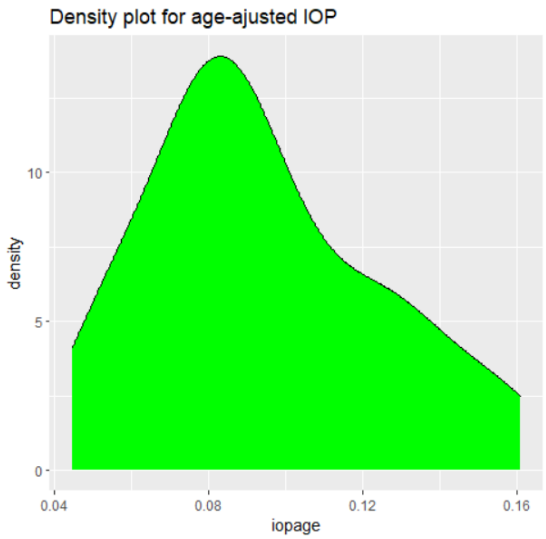




7) Adjusting based on Age (Dr. Chen) IOP(x) / Age(x)

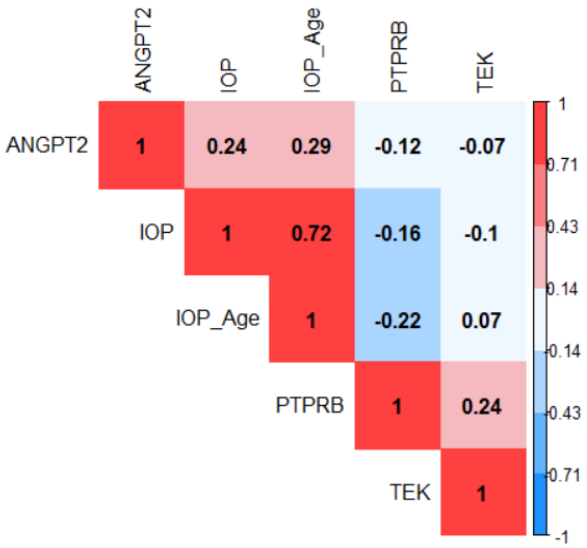
Dr. Chen suggested this method, but in my opinion, this is not a fair method. All of the previous 6 normalization methods are in the transformation class. They are completely reasonable except the first one (using rlog for IOP) that I rejected it by myself. In this method, IOP data are divided into Age of samples. If we look at the IOP density distribution, before and after adjusting, we found that the IOP for samples with around IOP=17 is suppressed. IOP=17 is in the middle of the range (10-24), and I think we are not allowed to do that. If you look at the correlation table below, although the correlation of ANGPT2 with new IOP_Age is increased a bit (from 0.24 to 0.29), however, the correlation between IOP and IOP_Age is 0.72. It means that the nature of original IOP data is changed during the adjustment. Also, in general, correlations between 0.2 and 0.4 are moderate, and those below 0.2 are considered weak. Therefore, 0.24 and 0.29 is not significantly different. However, the correlation 0.29 is almost (semi-) significant, too, with p-value 0.058 compare to the original IOP with a correlation of 0.24 with p-value 0.11. In my opinion, this is not a wise method, and I do not prefer to use this adjustment.

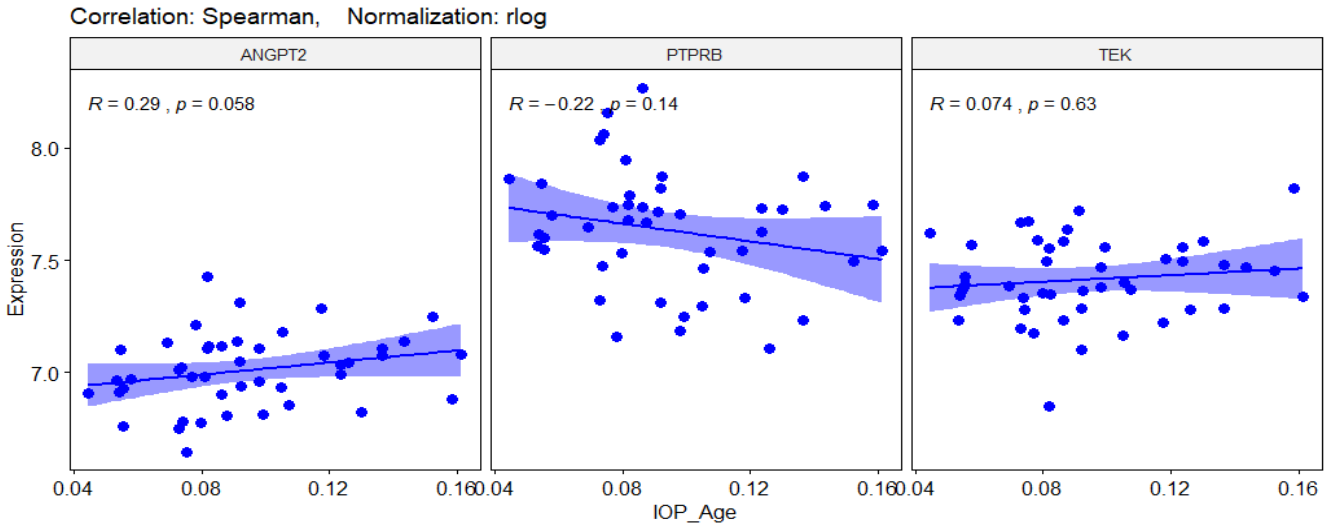





```
W = 0.93549, p-value = 0.01474
p-value < 0.05 then IOP is not normal
skewness(df1$IOP) = 0.4156063
```

```
W = 0.95304, p-value = 0.06623
p-value > 0.05 then iopavg is normal
skewness(df1$iopage) = 0.5328915
```



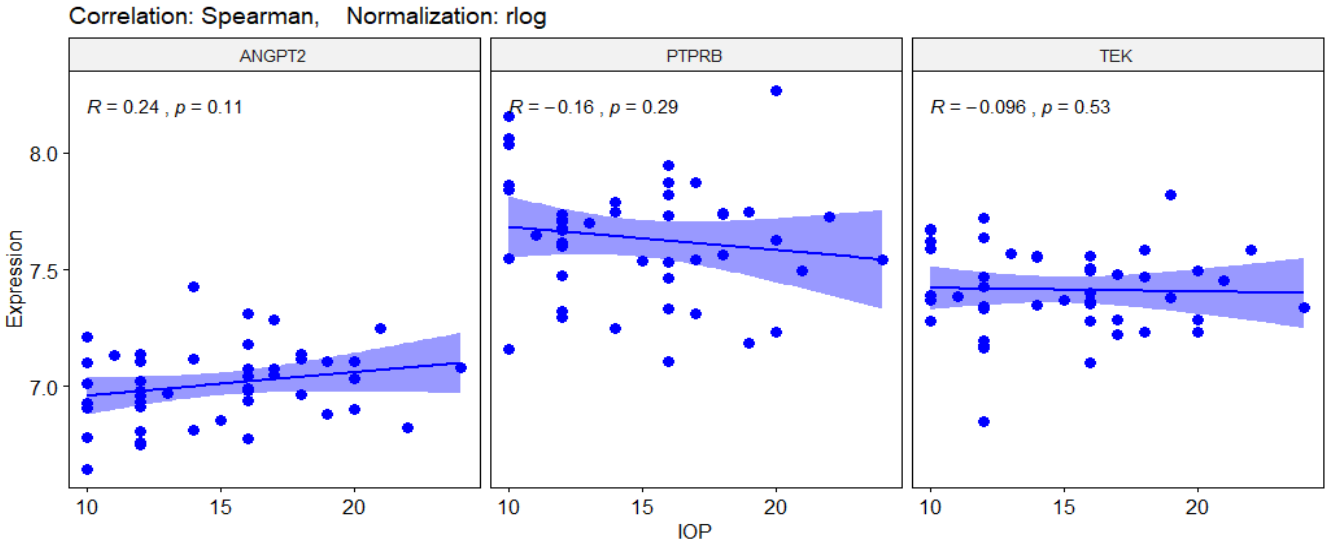


Compare IOP normalization methods

Among these seven normalization methods, just the last one can convert the distribution of IOP to a normal distribution; however, I have doubt it is a reasonable method. Except for rlog method (first method) and adjusted IOP (last one), the correlation matrix is not changed during normalization. Therefore, I cannot accept these first and last ones as a good normalization method. Almost all of them skewed right. I have chosen the **Log10 transformation normalization** as the best normalization method. Its density is semi-normal compare to others and has relatively low skewness. However, the goal is a correlation calculation between ANGPT2 and IOP. None of the transformation normalization methods does affect the correlation analysis significantly. Therefore, I have considered **rlog** as a normalization method for genes and **No-normalization** for IOP. Also, I considered **Spearman** as a correlation method for further analysis.

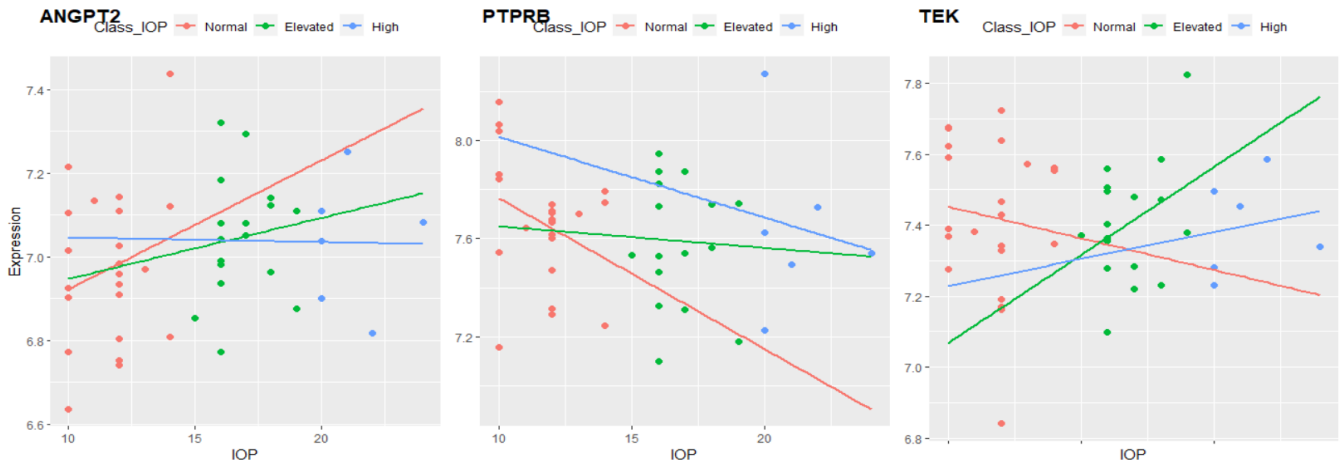
Correlation between IOP and ANGPT2

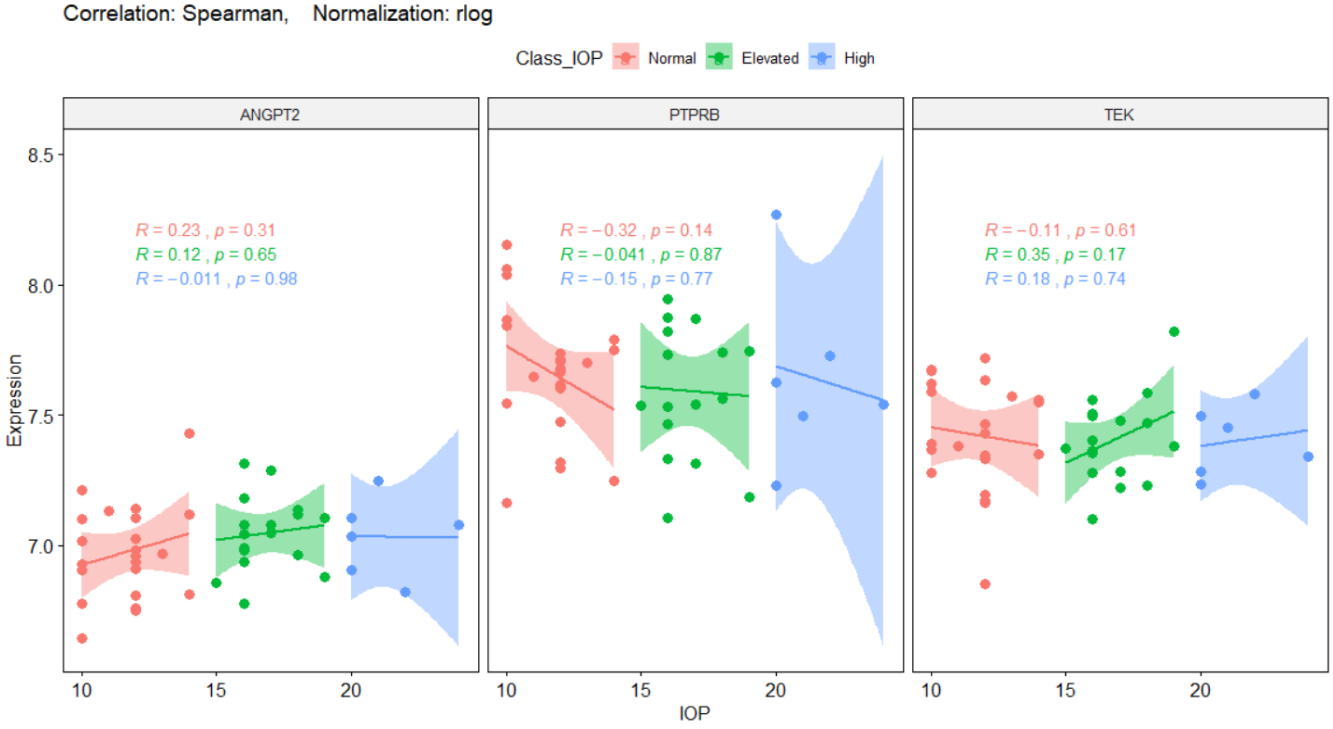
In this section, I want to calculate the **Spearman** correlation of real IOP (without normalization) with ANGPT2 gene expression (normalized using the **rlog** method). Based on the correlation coefficient R and P-values in the figure below, I found that IOP has a positive relationship with ANGPT2 and 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.



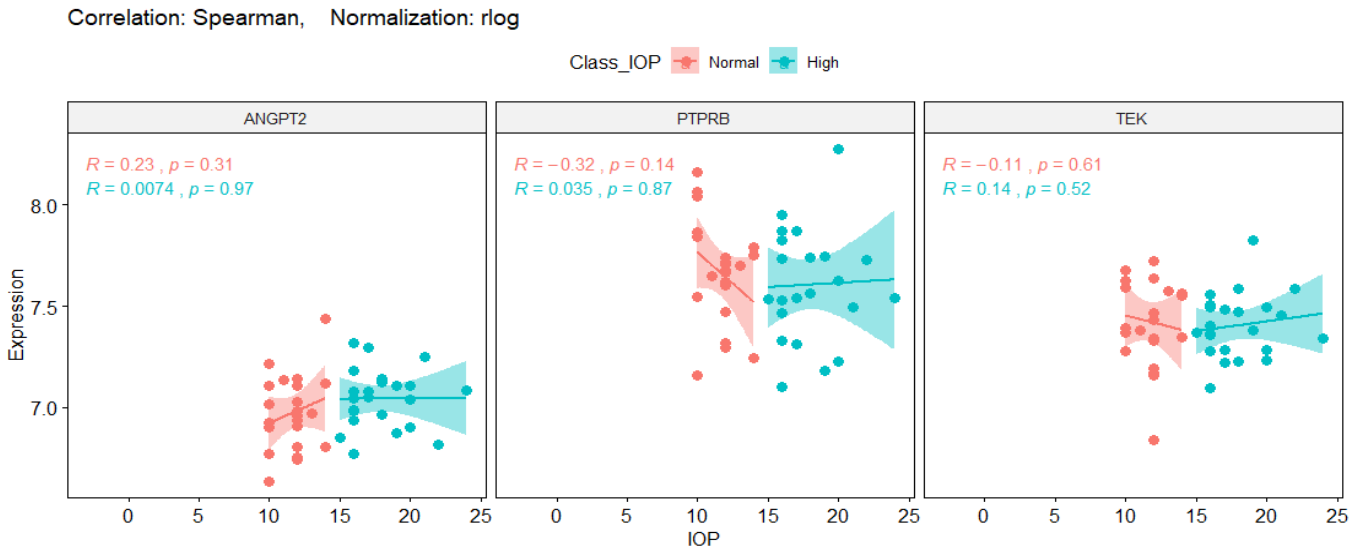
Partial-Correlation between IOP and ANGPT2 based on IOP subgroups

In the two figures below, I plot the correlation of three genes with three subgroups of IOP (Normal, Elevated, and High). 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_IOP subgroup has a major role in a part of the IOP dataset in the association with ANGPT2 and PTPRB. I mean, lower IOP related to the low expression of ANGPT2 and relatively high expression of PTPRB.





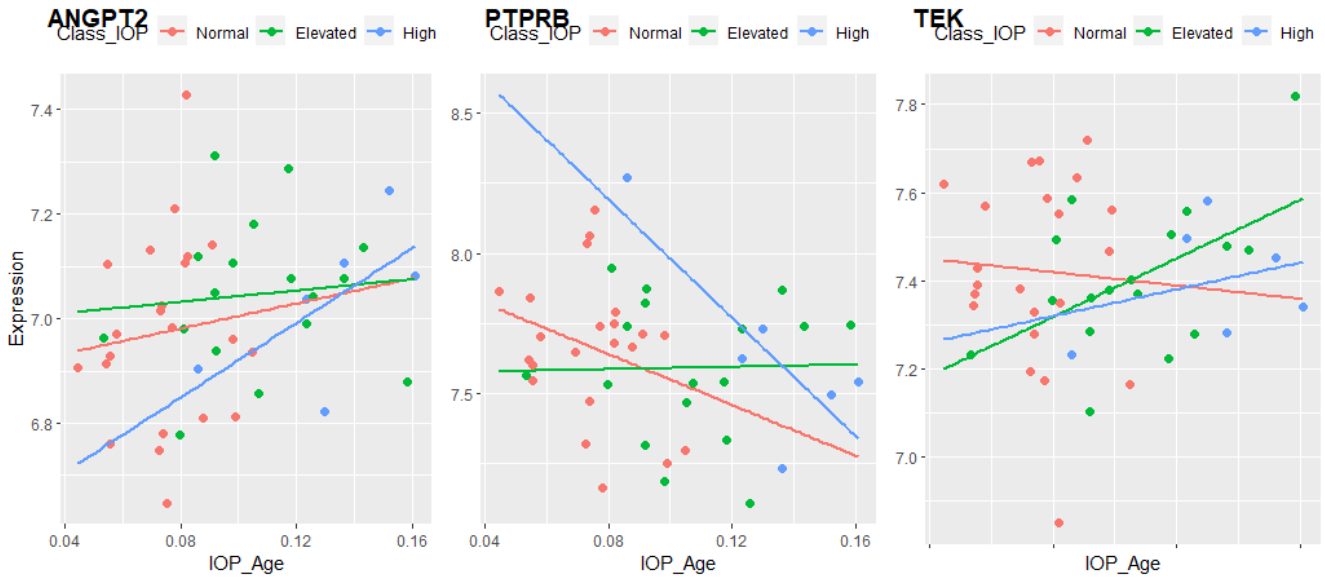
Although I am looking for a high correlation between the High_IOP subgroup and ANGPT2, however, as there are low samples in the high_IOP subgroup, it is not a significant relationship. Therefore, I decided to classify the IOP feature into two subgroups: Normal and High. In the figure below, I look for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

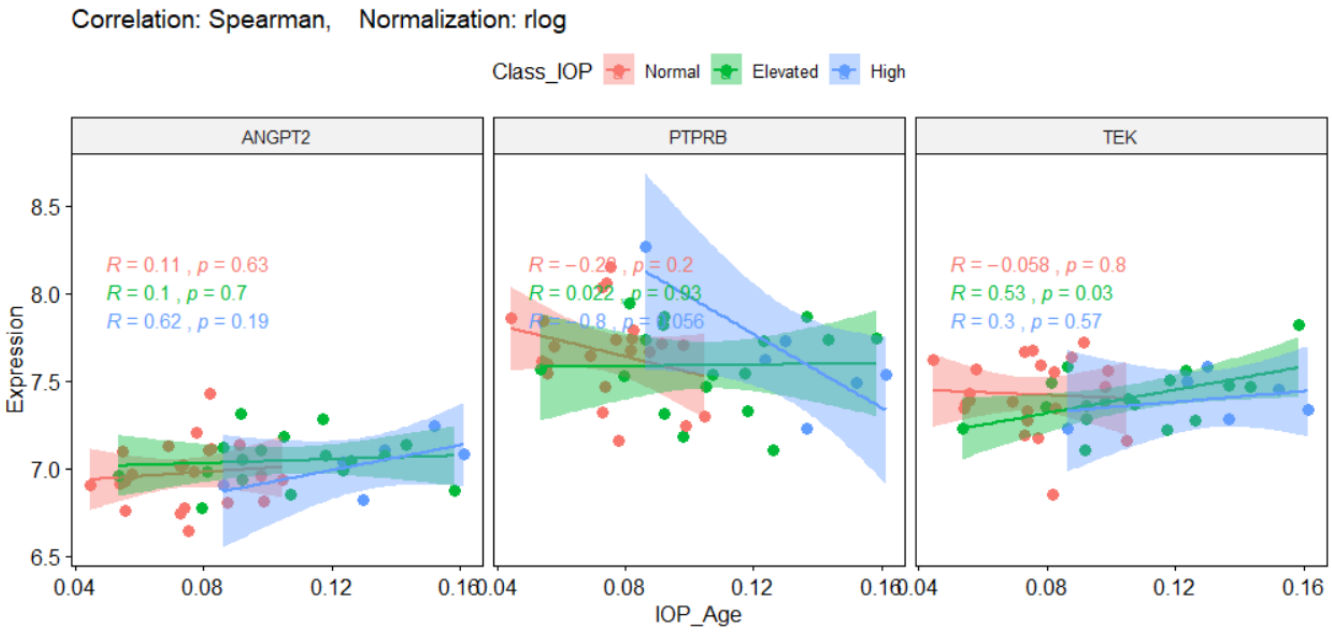


As a conclusion and based on figures above, ANGPT2 has a week-correlation, but not significant, with whole IOP data (R=0.24, P-value=0.11), which related to the Normal_IOP (R=0.23, P-value=0.31). However, the ANGPT2 gene has the maximum week-correlation with IOP, among other genes. Therefore, we need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

Partial-Correlation between IOP_Age and ANGPT2 based on IOP subgroups

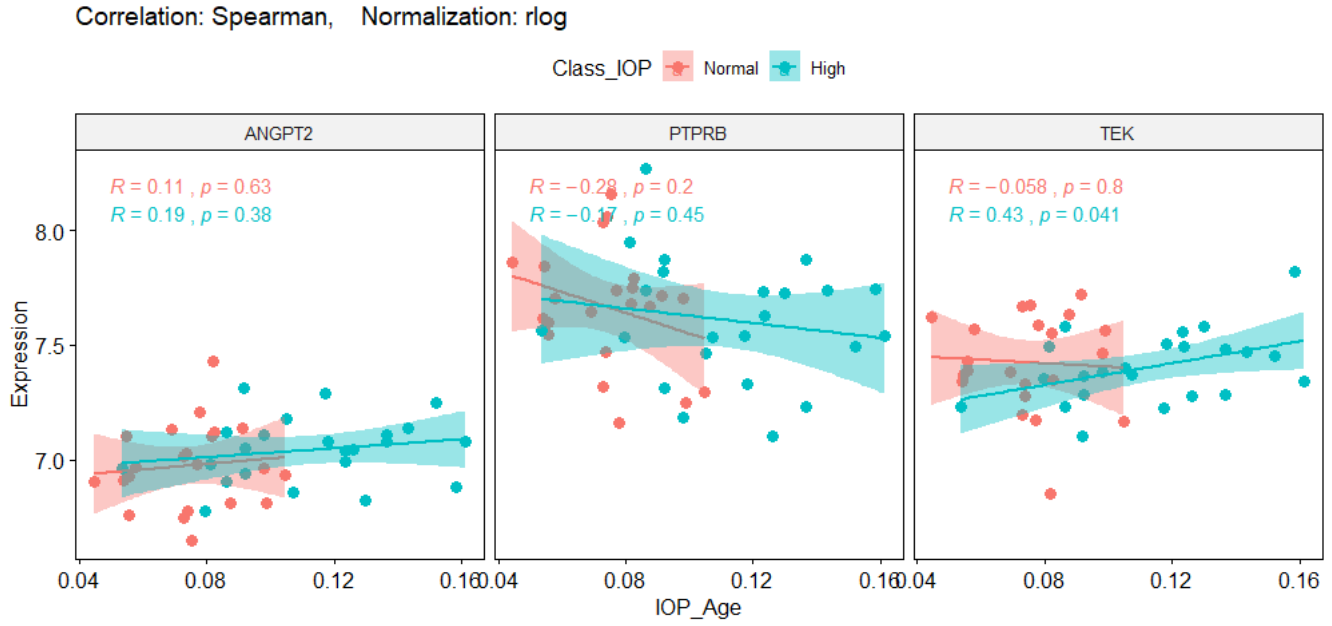
As I am curious about the adjusted IOP based on Age, Dr. Chen's method, I have repeated the partial-correlation analysis for IOP_Age data, too.





There is high positive correlation between **High_IOP_Age** and **ANGPT2** (R=+0.62, P-value=0.19) and high negative correlation between **High_IOP_Age** and **PTPRB** (R=-0.8, P-value=0.056). However, if you look at this figure, all three colores Blue, Green, and Red have overlaps, which is the results of adjustment. Adjusting with Age, caused Normal_IOP (Red) considers as High_IOP (Blue), and it is not correct.

Then, I classified the IOP-Age feature into two subgroups: Normal and High. In this scenario, half of the samples are High_IOP_Age; around 50% of samples (23 samples). In this condition, there is no significant feature to talk about.

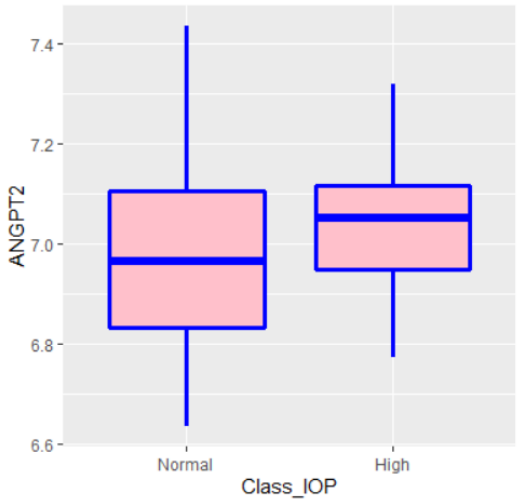


Analysis of t-test for IOP subgroups

To discuss the correlation between ANGPT2 and IOP subgroups statistically, I did a t-test and ANOVA analysis too. In the t-test analysis, I want to test that the mean expression level of ANGPT2 is equal in two separate groups of Normal_IOP and High_IOP (with almost the same sample sizes). If not, then the expression level of ANGPT2 in each subgroup of IOP are not similar together, and the expression levels are varied in different subgroups of IOP.

H0: The mean of ANGPT2 expression level in the Normal_IOP group is equal to the mean of ANGPT2 expression level in the 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)
```

```
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 higher than the significant level, 0.05. Therefore, I cannot reject the Null-Hypothesis to accept the alternative hypothesis. I conclude that the average expression level of the ANGPT2 gene in samples with Normal_IOP is not significantly different from the samples with High_IOP; the expression level of the ANGPT2 gene is almost the same in two groups (6.97 ≈ 7.04).

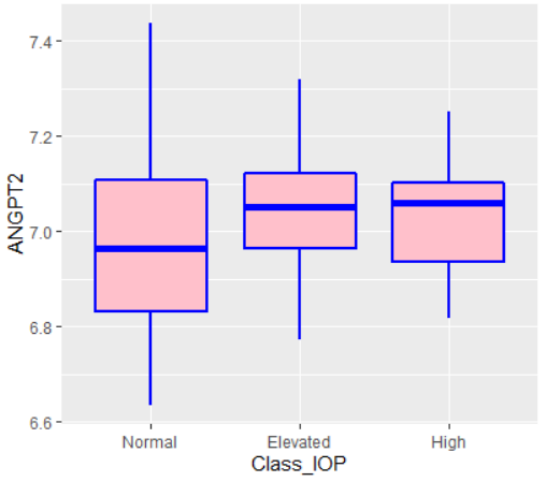
It means that different IOP subgroups have not an essential role in the expression level of ANGPT2.

Analysis of ANOVA for IOP subgroups

As in the real state, we have three subgroups of IOP, so we need to run the ANOVA analysis instead of a 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.



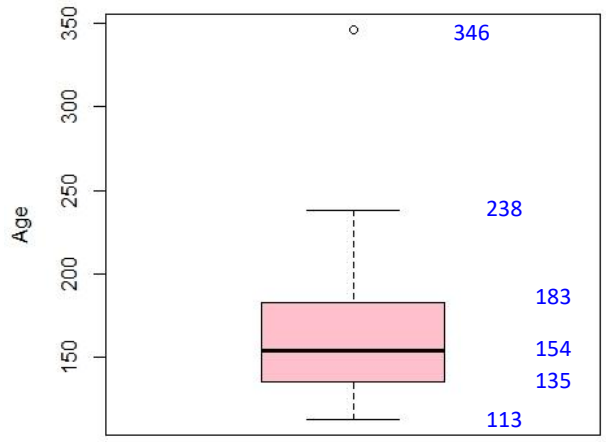

```
Anova_results <- aov(ANGPT2 ~ Class_IOP, data=corrTable)
summary(Anova_results)
```

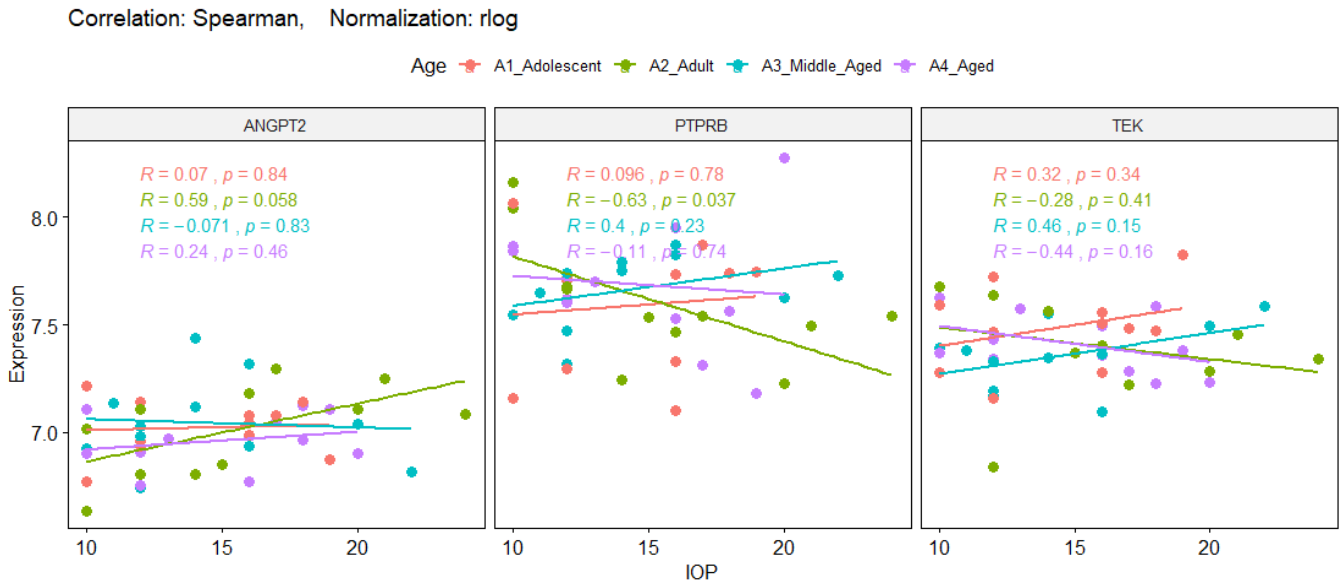
```
             Df Sum Sq Mean Sq F value Pr(>F)
Class_IOP     2 0.0561 0.02806       1  0.377
Residuals    42 1.1789 0.02807
F(2,42) = 1,    P-value > 0.05
```

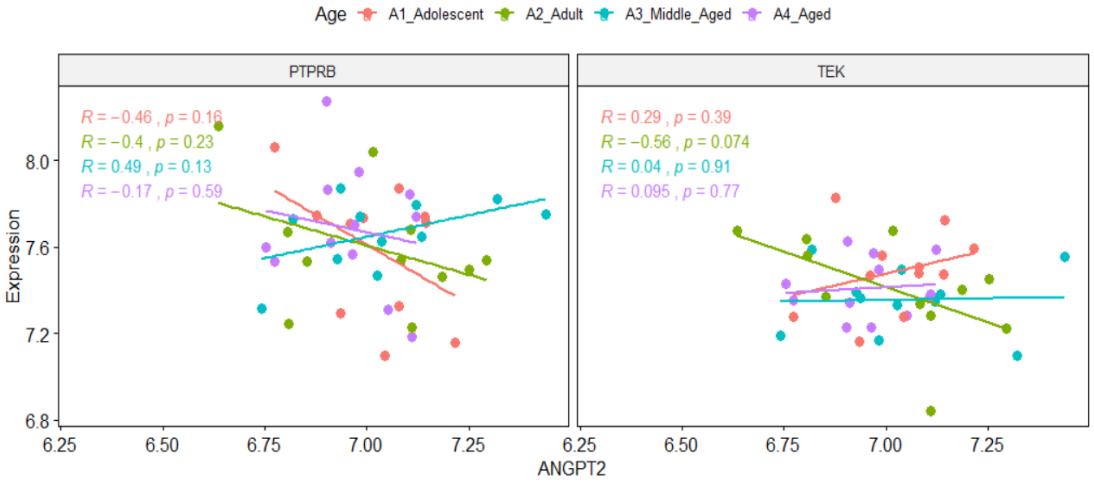
The P-value is higher than 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claim that the expression level of ANGPT2 is significantly varied in three different IOP groups.

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 subgroups. 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 subgroup.







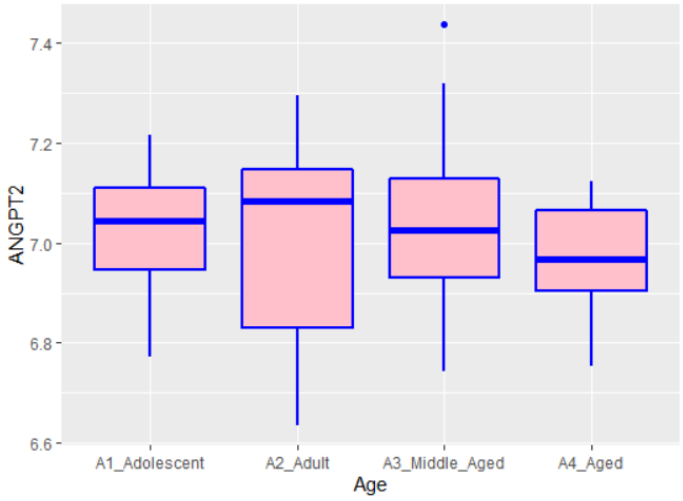
Good news: Based on the results, I can see there is a **semi-significant relationship** in the **Adult** subgroup between IOP and ANGPT2 (R=+0.59, P-value=0.058), and PTPRB (R=-0.63, P-value=0.037). Also, there is a semi-significant relationship between ANGPT2 and TEK genes in this subgroup (Adult); figure below. Therefore, I am curious about **Adult** subgroup. To test these observations, I am going to do a t-test and ANOVA analysis.

Analysis of ANOVA and lm considering Age

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

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

Ha: They are not equal.

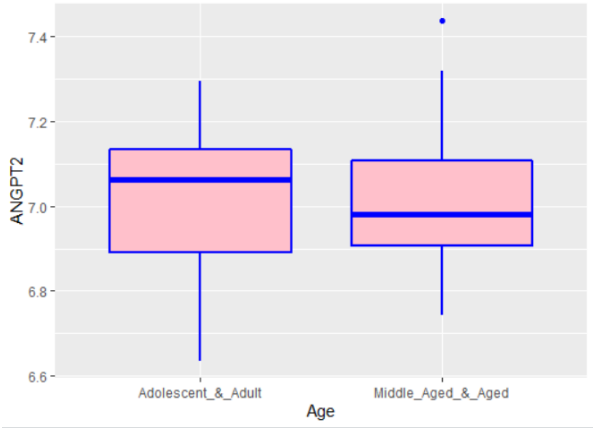



```
Anova_results <- aov(ANGPT2 ~ Age, corrTable); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
Age           3 0.0415 0.01384   0.476  0.701
Residuals    41 1.1935 0.02911

Anova_results <- aov(ANGPT2 ~ IOP + Age, corrTable); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.290  0.138
Age           3 0.0494 0.01647   0.587  0.627
Residuals    40 1.1214 0.02804

Linear_model <- lm(ANGPT2 ~ IOP + Age, corrTable); summary(Linear_model)

Call:
lm(formula = ANGPT2 ~ IOP + Age, data = corrTable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33588 -0.07945  0.00237  0.10007  0.44388

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.9309303  0.1335659  51.891   <2e-16 ***
IOP          0.0109180  0.0067533   1.617    0.113
Age         -0.0005102  0.0005813  -0.878    0.385

Residual standard error: 0.1655 on 42 degrees of freedom
Multiple R-squared:  0.06906,  Adjusted R-squared:  0.02473
F-statistic: 1.558 on 2 and 42 DF,  p-value: 0.2225
```

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

Then, I decided to divide the Age into just two subgroups as below and run the t-test. In each division, I followed a logical rule: 1) split ages half by half (Adolescent and Adults in the first group and Middle-Aged and Aged in the second group). 2) separate Adults from others. 3) separate Aged (as the oldest samples) from others. 4) separate Adolocent (as the youngest samples) from others.

Division 1): Divided ages half by half (Adolescent and Adults in the first group and Middle-Aged and Aged in the second group)

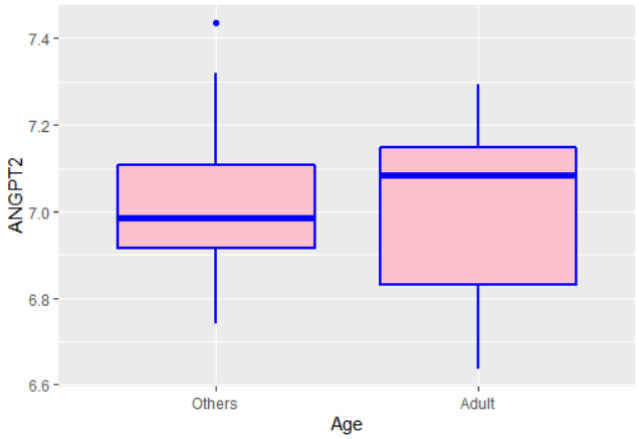



```
Anova_results <- aov(ANGPT2 ~ Age, corrTable_Age2); summary(Anova_results)
             Df Sum Sq  Mean Sq F value Pr(>F)
Age           1  0.003 0.002999   0.105  0.748
Residuals    43  1.232 0.028651
```

```
Anova_results <- aov(ANGPT2 ~ IOP + Age, corrTable_Age2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.308  0.136
Age           1 0.0024 0.00239   0.086  0.771
Residuals    42 1.1684 0.02782
t.test(ANGPT2 ~ Age, corrTable_Age2, mu=0, alt="two.sided", conf=0.95, var.eq=F, p
aired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Age
t = 0.32342, df = 42.84, p-value = 0.748
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08551491  0.11817765
sample estimates:
mean in group Adolescent_&_Adult:      7.017332
mean in group Middle_Aged_&_Aged:      7.001000
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

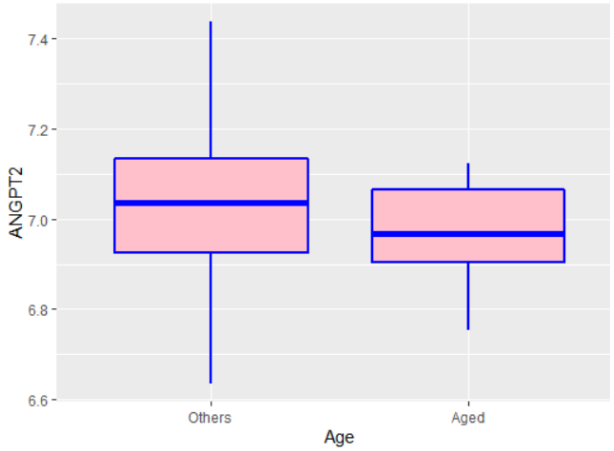
Division 2): separate Adults from others




```
Anova_results <- aov(ANGPT2 ~ Age, corrTable_Age2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
Age           1 0.0003 0.00027   0.009  0.923
Residuals    43 1.2347 0.02872

Anova_results <- aov(ANGPT2 ~ IOP + Age, corrTable_Age2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.303  0.137
Age           1 0.0001 0.00011   0.004  0.951
Residuals    42 1.1707 0.02787

t.test(ANGPT2 ~ Age, corrTable_Age2, mu=0, alt="two.sided", conf=0.95, var.eq=F, p
aired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Age
t = -0.082964, df = 13.696, p-value = 0.9351
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1534463  0.1420401
sample estimates:
mean in group Others  mean in group Adult
            7.007590             7.013293
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

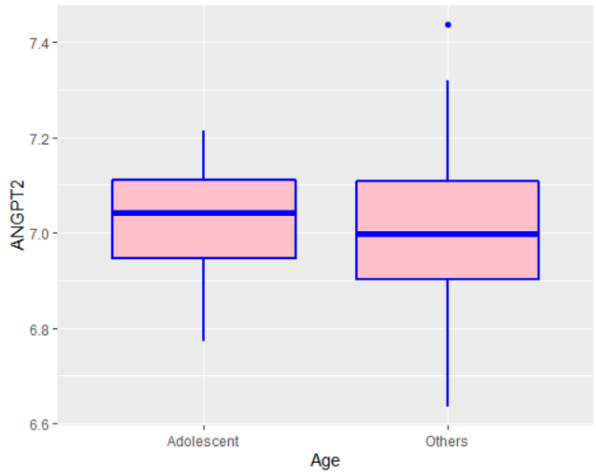
Division 3): separate Aged (as the oldest samples) from others



```
Anova_results <- aov(ANGPT2 ~ Age, 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

Anova_results <- aov(ANGPT2 ~ IOP + Age, corrTable_Age2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.384  0.130
Age           1 0.0397 0.03969   1.474  0.232
Residuals    42 1.1311 0.02693

t.test(ANGPT2 ~ Age, corrTable_Age2, mu=0, alt="two.sided", conf=0.95, var.eq=F, p
aired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Age
t = 1.3563, df = 28.837, p-value = 0.1855
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03257421  0.16072912
sample estimates:
mean in group Others   mean in group Aged
             7.026072             6.961994
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

Division 4): separate Adolocent (as the youngest samples) from others




```
Anova_results <- aov(ANGPT2 ~ Age, data=corrTable_Age2)
summary(Anova_results)
             Df Sum Sq  Mean Sq F value Pr(>F)
Age           1 0.0022 0.002233   0.078  0.782
Residuals    43 1.2328 0.028669

Anova_results <- aov(ANGPT2 ~ IOP + Age, data=corrTable_Age2)
summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.312  0.136
Age           1 0.0045 0.00452   0.163  0.689
Residuals    42 1.1663 0.02777

t.test(ANGPT2 ~ Age, data=corrTable_Age2, mu=0, alt="two.sided", conf=0.95, var.eq
=F, paired=F)

        Welch Two Sample t-test
```

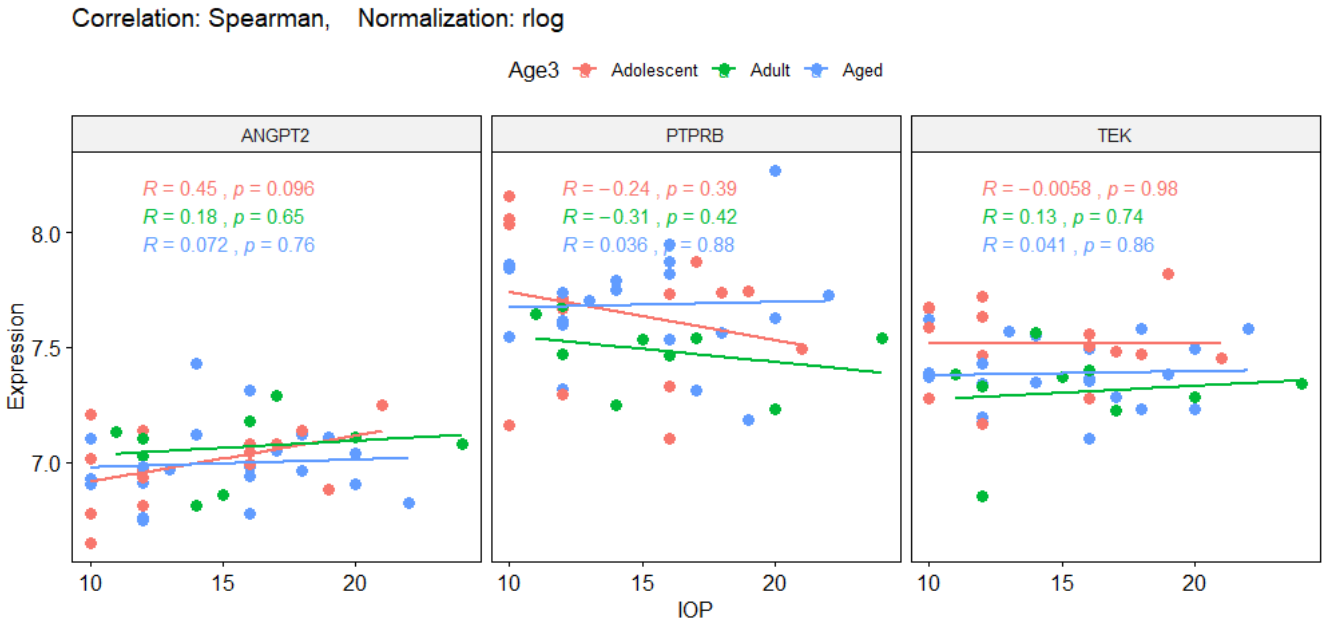
```
data:  ANGPT2 by Age
t = 0.32893, df = 23.455, p-value = 0.7451
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08658991  0.11937447
sample estimates:
mean in group Adolescent     mean in group Others
                7.021370                 7.004977
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

All analysis above indicates that there is no significant relation between IOP and ANGPT2 based on different subgroups of Age feature.

Partial-Correlation between IOP and ANGPT2 based on three subgroups of Age feature

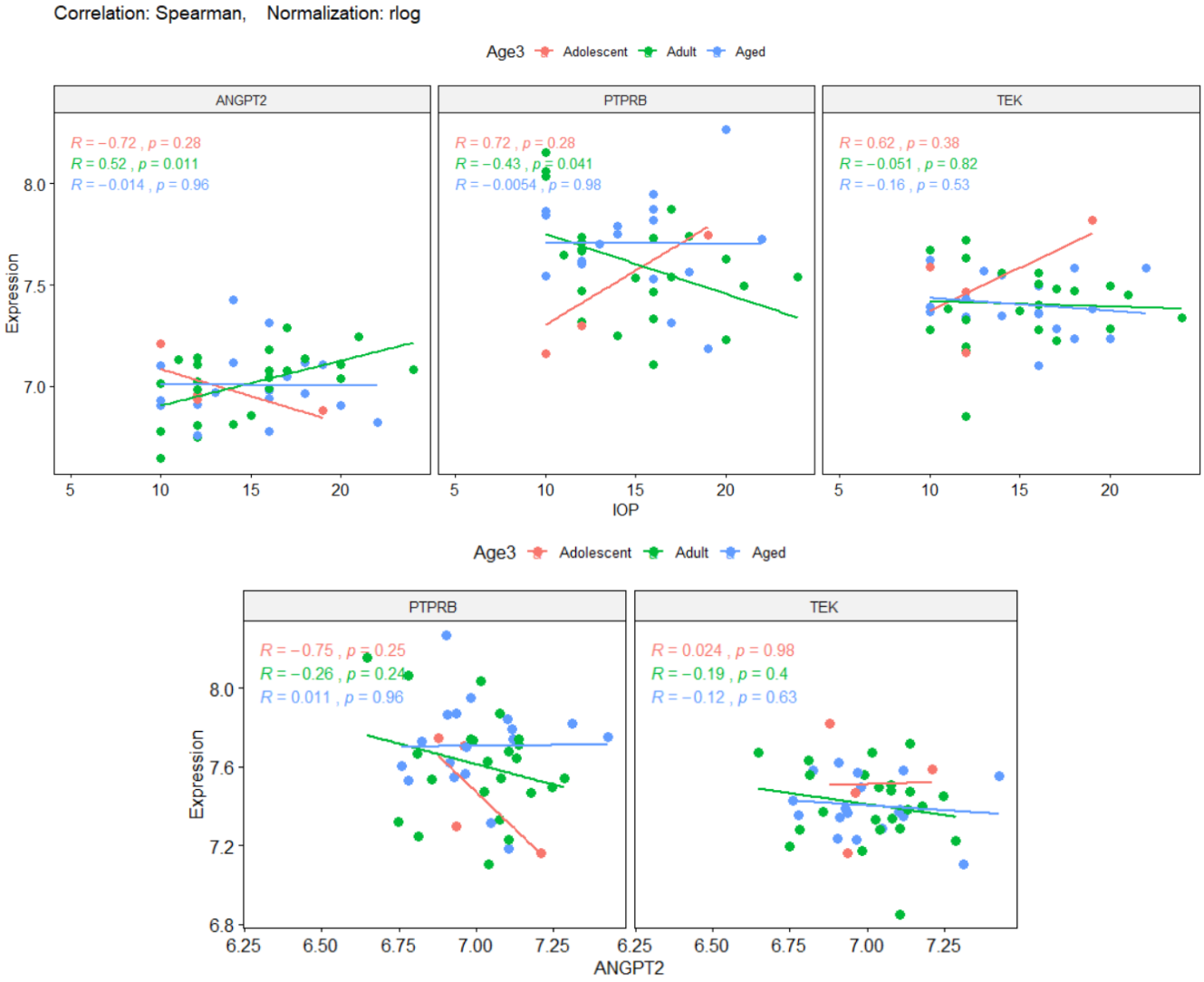
Based on the previous analysis, I cannot find any significant t-test or ANOVA for previous subgroups. However, I found that there is a good relation between IOP and ANGPT2 and PTPRB just in the **Adult** subgroups. Now, I am going to divide samples to three equal subgroups: **Adolescent** (113<=Age<138), **Adult** (138<=Age<178), and **Aged** (178<=Age<=346). Then, plot the correlation based on each subgroup.



Good news: Based on this division, again, there is a **semi-significant relationship** in the youngest subgroup, **Adolescent**, between IOP and ANGPT2 (R=+0.45, P-value=0.096). It means that the relationship of IOP with ANGPT2 is related to the particular subset of Age.

Partial-Correlation between IOP and ANGPT2 based on the best subgroup of Age feature

After previous analysis, I found that there are a best subgroups of Age, which IOP measure of these samples are significantly correlated to the ANGPT2 and PTPRB genes. After several trying, finally, I found this subset. Now, I am going to divide samples into three variant subgroups: **Adolescent** (113<=Age<127) with 4 samples, **Adult** (127<=Age<167) with 23 samples, and **Aged** (167<=Age<=346) with 18 samples. I considered the **Adult** subgroup with an almost sufficient sample size (23) as the best subgroup, which I looked for. Then, I plot the correlation based on each subgroup.



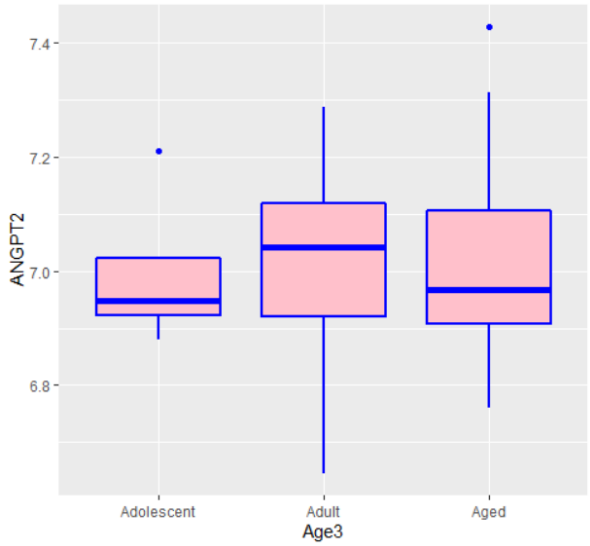
As you can see, there are **significant and strong relationships** in the **Adult** subgroup, between IOP and ANGPT2 (R=+0.52, P-value=0.011), and PTPRB (R=-0.43, P-value=0.041).

Analysis of ANOVA and lm considering Age3

There are three subgroups of Age3, and I run ANOVA analysis in for ANGPT2 expression along with these subgroups.

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

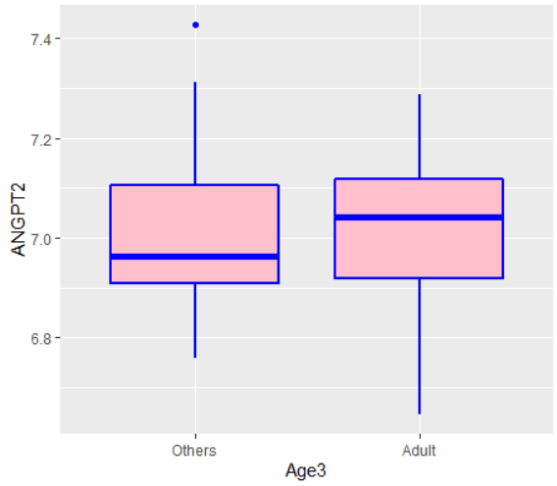
Ha: They are not equal.




```
Anova_results <- aov(ANGPT2 ~ Age3, data=corrTable); summary(Anova_results)
            Df Sum Sq  Mean Sq F value Pr(>F)
Age3         2 0.0013 0.000644   0.023  0.977
Residuals   42 1.1738 0.027948

Anova_results <- aov(ANGPT2 ~ IOP + Age3, data=corrTable); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0609 0.06092   2.244  0.142
Age3         2 0.0012 0.00059   0.022  0.979
Residuals   41 1.1130 0.02715

Linear_model <- lm(ANGPT2 ~ IOP + Age3, corrTable); summary(Linear_model)
Call:
lm(formula = ANGPT2 ~ IOP + Age3, data = corrTable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31830 -0.07672 -0.01378  0.09230  0.43428

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.8629166  0.1216377  56.421   <2e-16 ***
IOP          0.0101084  0.0067542   1.497    0.142
Age3Adult    0.0001003  0.0899622   0.001    0.999
Age3Aged    -0.0103533  0.0919917  -0.113    0.911
---

Residual standard error: 0.1648 on 41 degrees of freedom
Multiple R-squared:  0.05284,  Adjusted R-squared:  -0.01646
F-statistic: 0.7624 on 3 and 41 DF,  p-value: 0.5217
```

The P-value is higher than 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claim that the expression level of ANGPT2 is significantly varied in three different Age3 groups. Then, I decided to divide the Age3 into just two subgroups and run the t-test. I kept the **Adult** subgroups and mixed the **Adolocent** and **Aged** group as another subgroup. Therefore, the sample size of each new subgroups is almost equal.

```
summary(corrTable_Age2$Age3)
Others     Adult
    22     23
```

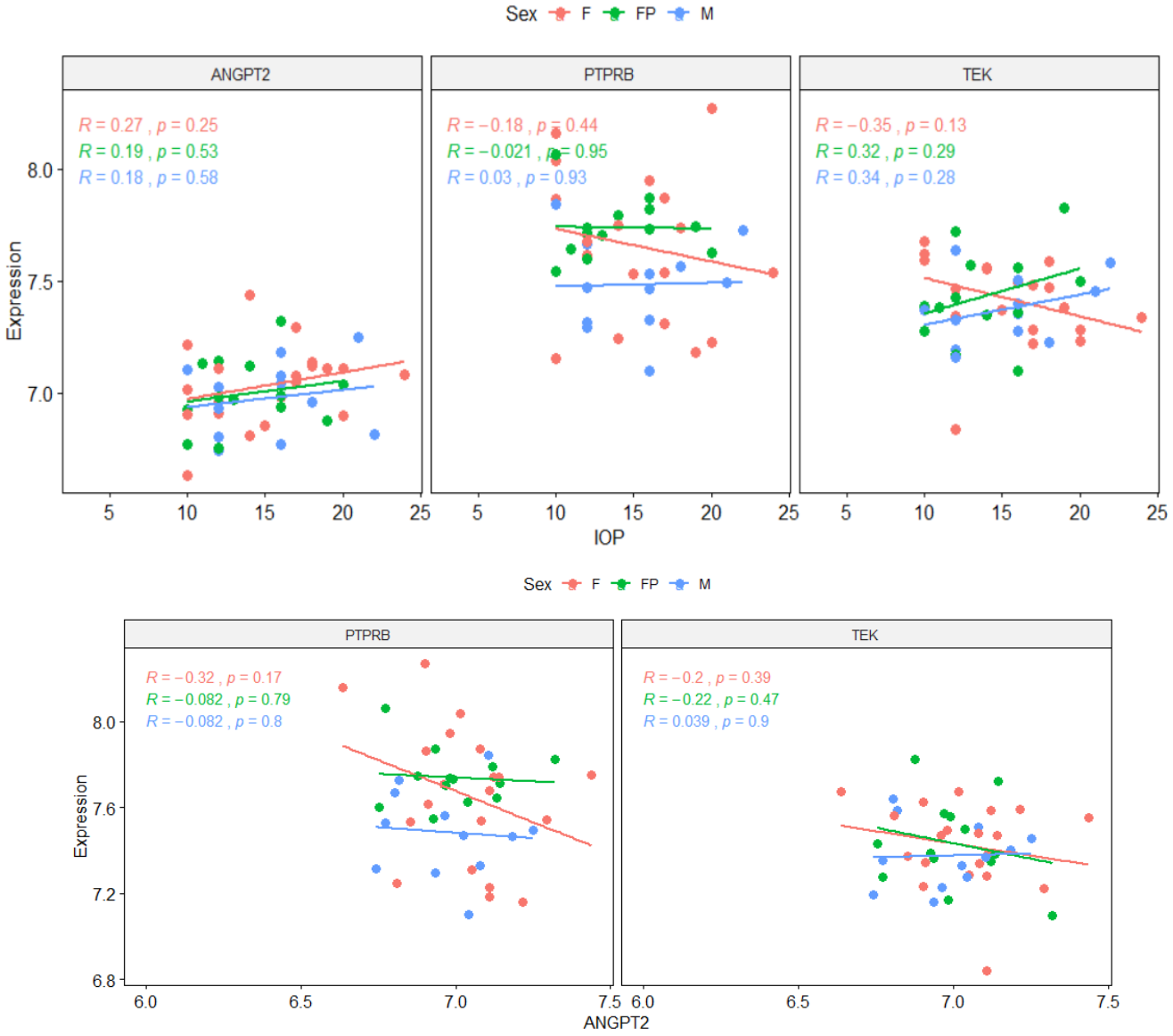



```
Anova_results <- aov(ANGPT2 ~ Age3, corrTable_Age2); summary(Anova_results)
             Df Sum Sq  Mean Sq F value Pr(>F)
Age3          1  0.001 0.001021   0.037  0.848
Residuals    43  1.174 0.027304

Anova_results <- aov(ANGPT2 ~ IOP + Age3, corrTable_Age2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0609 0.06092   2.298  0.137
Age3          1 0.0008 0.00083   0.031  0.861
Residuals    42 1.1134 0.02651

t.test(ANGPT2 ~ Age3, corrTable_Age2, mu=0, alt="two.sided", conf=0.95, var.eq=F,
paired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Age3
t = -0.1934, df = 42.914, p-value = 0.8476
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1089114  0.0898513
sample estimates:
mean in group Others  mean in group Adult
             7.004234              7.013764
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

Partial-Correlation between IOP and ANGPT2 based on Sex feature

In this section, I am going to calculate the correlation between IOP and ANGPT2 based on different Sex subgroups. Among 45 samples, 12 (26%) of them are male, and 33 (73%) are Female. Also, there are 13 (29%)Pregnant samples.



There is no significant correlation between Sex subgroups and genes. Almost all three Sex subgroups have a similar role in the relationship between IOP and ANGPT2.

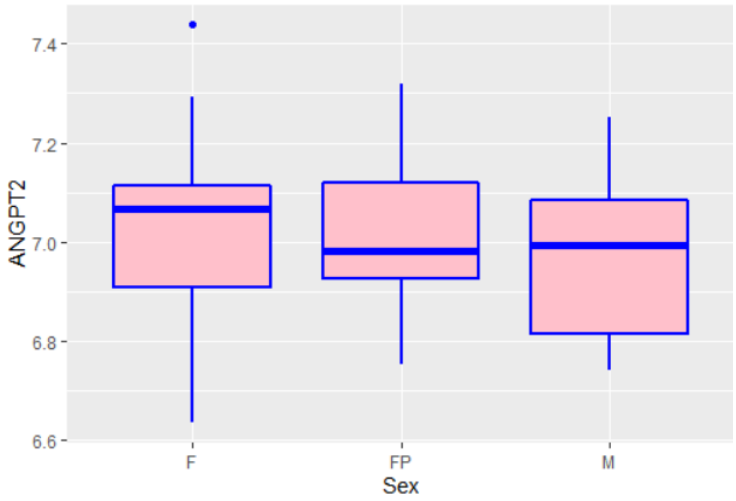
Now, I am going to do a t-test and ANOVA analysis.

Analysis of ANOVA and lm considering Sex

There are three subgroups of Sex, and I run ANOVA analysis in for ANGPT2 expression along with these subgroups.

H0: The mean of ANGPT2 expression level in all three subgroups of Sex is equal together.

Ha: They are not equal.




```
Anova_results <- aov(ANGPT2 ~ Sex, corrTable); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
Sex          2 0.0286 0.01430   0.498  0.611
Residuals   42 1.2064 0.02872
```

```
Anova_results <- aov(ANGPT2 ~ IOP + Sex, corrTable); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.300  0.137
Sex          2 0.0261 0.01306   0.468  0.630
Residuals   41 1.1447 0.02792
```

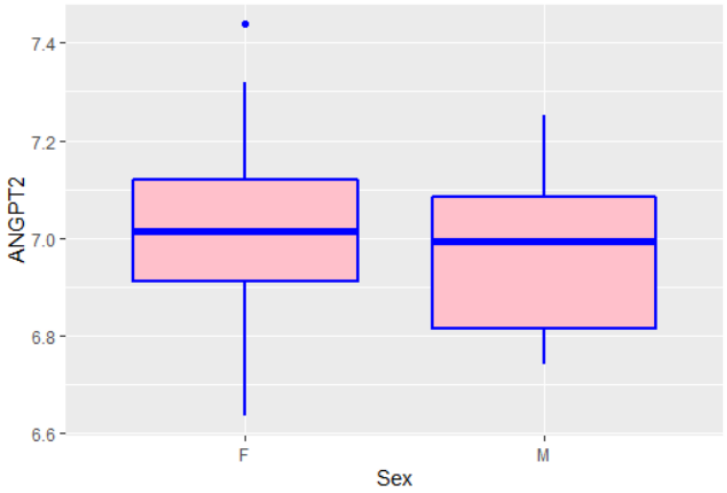
```
Linear_model <- lm(ANGPT2 ~ IOP + Sex, corrTable); summary(Linear_model)

Call:
lm(formula = ANGPT2 ~ IOP + Sex, data = corrTable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34689 -0.09229 -0.00910  0.09460  0.41424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.880054   0.111283  61.825   <2e-16 ***
IOP          0.010219   0.006874   1.487    0.145
SexFP       -0.025231   0.060223  -0.419    0.677
SexM        -0.058909   0.061013  -0.966    0.340
---

Residual standard error: 0.1671 on 41 degrees of freedom
Multiple R-squared:  0.07313,  Adjusted R-squared:  0.005313
F-statistic: 1.078 on 3 and 41 DF,  p-value: 0.3689
```

The P-value is higher than 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claim that the expression level of ANGPT2 is significantly varied in three different Sex subgroups.

Then, I decided to divide the Sex group into just two subgroups as below and run the t-test. Division 1) convert FP to F. Division 2) convert F and M to NonP.

Division 1): Convert FP to F. So, we have just F and M

```
summary(corrTable_Sex2$Sex)
F  M
33 12
```

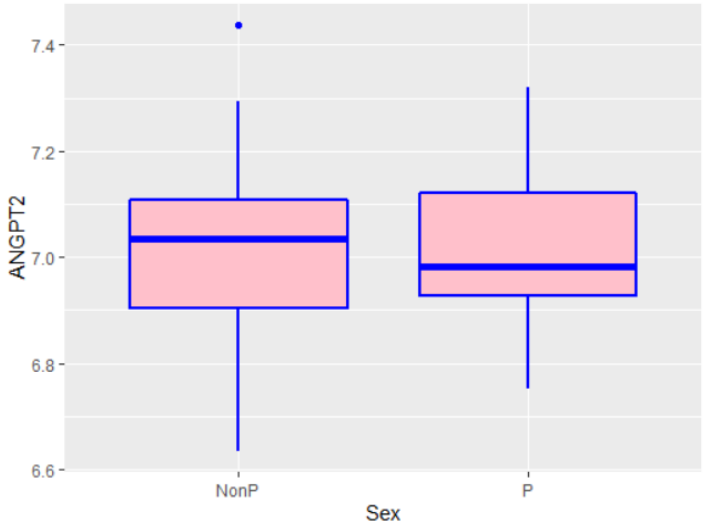



```
Anova_results <- aov(ANGPT2 ~ Sex, corrTable_Sex2); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
Sex          1 0.0167 0.01675   0.591  0.446
Residuals   43 1.2183 0.02833
Anova_results <- aov(ANGPT2 ~ IOP + Sex, corrTable_Sex2); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.346  0.133
Sex          1 0.0212 0.02122   0.775  0.384
Residuals   42 1.1496 0.02737

Anova_results <- aov(ANGPT2 ~ IOP + Sex + Age, corrTable_Sex2); summary(Anova_resu
lts)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.279  0.139
Sex          1 0.0212 0.02122   0.753  0.391
Age          3 0.0511 0.01703   0.605  0.616
Residuals   39 1.0985 0.02817

t.test(ANGPT2 ~ Sex, corrTable_Sex2, mu=0, alt="two.sided", conf=0.95, var.eq=F, p
aired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Sex
t = 0.77303, df = 19.763, p-value = 0.4487
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07418849  0.16144413
sample estimates:
mean in group F        mean in group M
       7.020618        6.976991
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

Division 2): Convert F and M to NonP. So, we have just Pregnant (P) and Non-Pregnant (NonP)

```
summary(corrTable_Sex2$Sex)
NonP     P
  32    13
```




```
Anova_results <- aov(ANGPT2 ~ Sex, corrTable_Sex2); summary(Anova_results)
             Df Sum Sq  Mean Sq F value Pr(>F)
Sex           1 0.0026 0.002578    0.09  0.766
Residuals    43 1.2324 0.028661
```

```
Anova_results <- aov(ANGPT2 ~ IOP + Sex, corrTable_Sex2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.303  0.137
Sex           1 0.0001 0.00009   0.003  0.955
Residuals    42 1.1707 0.02787
```

```
Anova_results <- aov(ANGPT2 ~ IOP + Sex + Age, corrTable_Sex2); summary(Anova_resu
lts)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.254  0.141
Sex           1 0.0001 0.00009   0.003  0.956
Age           3 0.0600 0.01999   0.702  0.557
Residuals    39 1.1108 0.02848
```

```
t.test(ANGPT2 ~ Sex, corrTable_Sex2, mu=0, alt="two.sided", conf=0.95, var.eq=F, p
aired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Sex
t = 0.31355, df = 24.603, p-value = 0.7565
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09308437  0.12648488
sample estimates:
mean in group NonP    mean in group P
          7.013809           6.997109
```

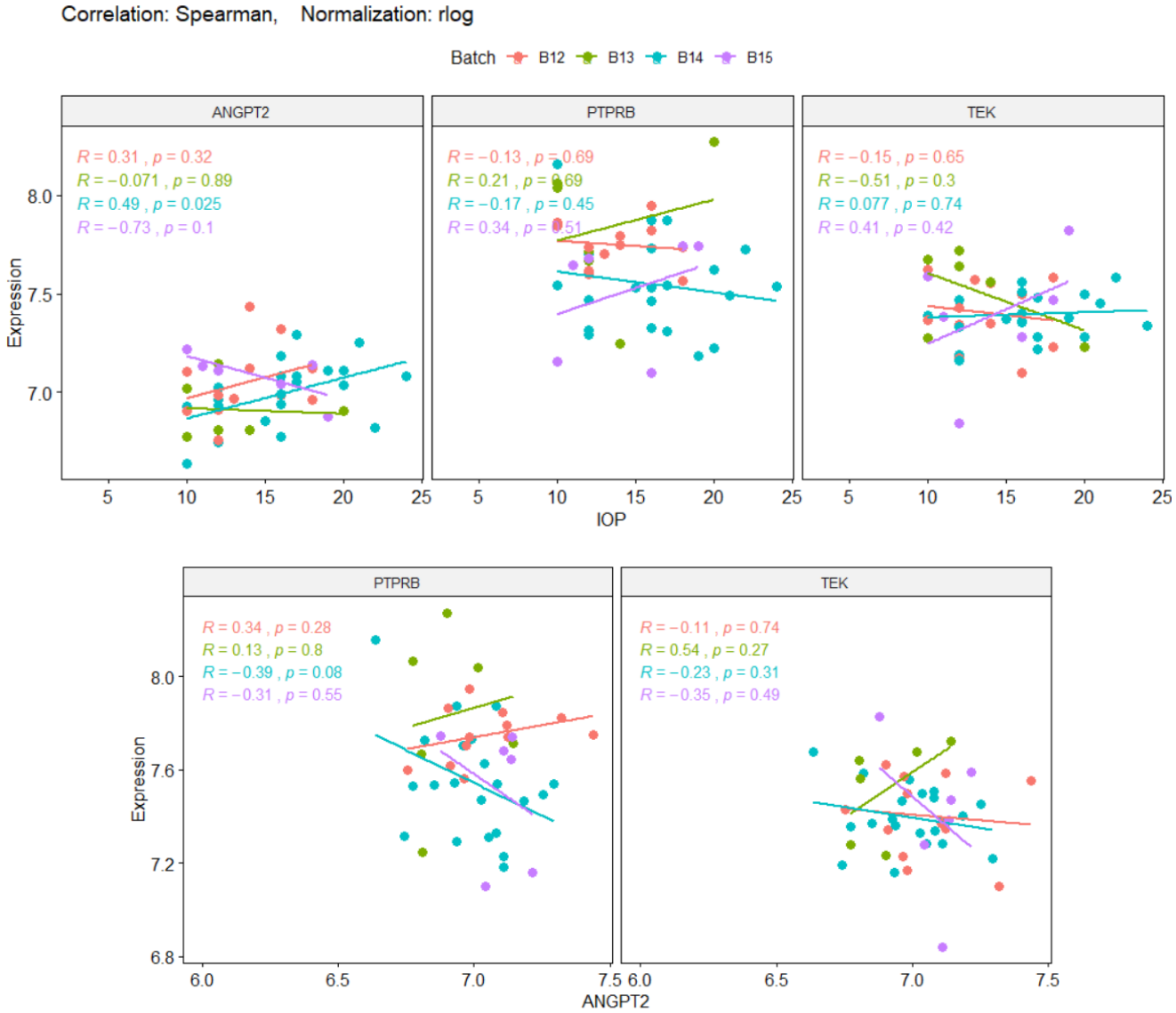
In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

All analysis above indicates that there is no significant relation between IOP and ANGPT2 based on different subgroups of Sex feature.

Partial-Correlation between IOP and ANGPT2 based on Batch feature

In this section, I am going to calculate the correlation between IOP and ANGPT2 based on different Batch subgroups. Our 45 samples have four batch numbers: B12 (12 samples), B13 (6 samples), B14 (21 samples), and B15 (6 samples).

```
> summary(corrTable$Batch)
B12 B13 B14 B15
 12   6  21   6
```



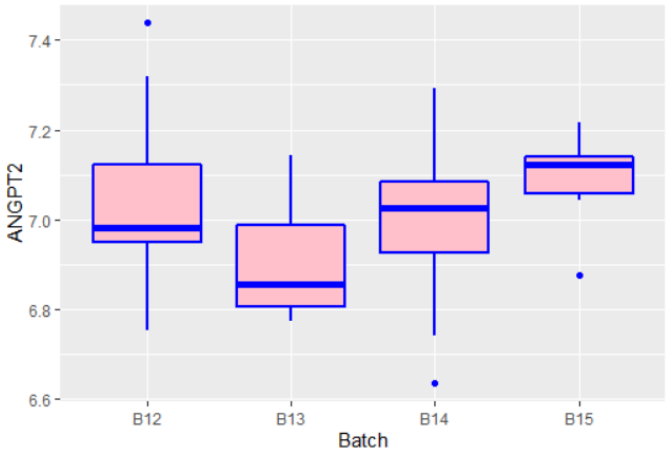
Good news: Based on the results, I can see there is a **significant relationship** in the **B14** subgroup between ANGPT2 and IOP (R=+0.49, P-value=0.025), and PTPRB (R=-0.39, P-value=0.08). Now, I am going to do a t-test and ANOVA analysis.

Analysis of ANOVA and lm considering Batch

There are four subgroups of Batch, and I run ANOVA analysis in for ANGPT2 expression along with these subgroups.

H0: The mean of ANGPT2 expression level in all three subgroups of Batch is equal together.

Ha: They are not equal.



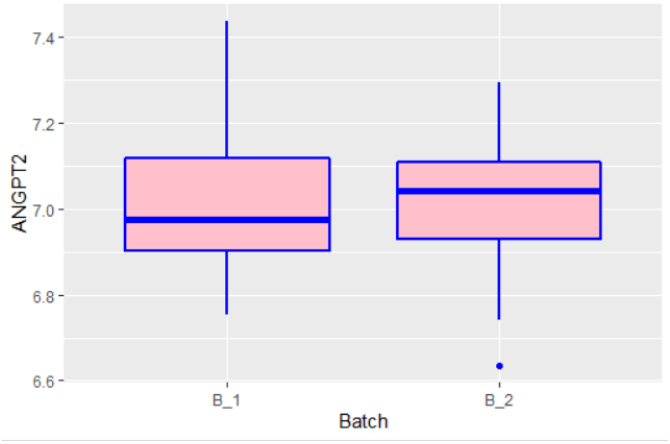

```
Anova_results <- aov(ANGPT2 ~ Batch, corrTable); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
Batch        3  0.120  0.0400   1.471  0.237
Residuals   41  1.115  0.0272

Anova_results <- aov(ANGPT2 ~ IOP + Batch, corrTable); summary(Anova_results)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.468  0.124
Batch        3 0.1301 0.04336   1.666  0.190
Residuals   40 1.0407 0.02602
Anova_results <- aov(ANGPT2 ~ IOP + Batch + Sex + Age, corrTable); summary(Anova_r
esults)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.587  0.117
Batch        3 0.1301 0.04336   1.747  0.175
Sex          2 0.0200 0.00999   0.403  0.672
Age          3 0.1522 0.05073   2.044  0.125
Residuals   35 0.8685 0.02482

Linear_model <- lm(ANGPT2 ~ IOP + Batch, corrTable); summary(Linear_model)

lm(formula = ANGPT2 ~ IOP + Batch, data = corrTable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28542 -0.09903 -0.00121  0.08289  0.38699

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.885063   0.106772  64.484   <2e-16 ***
IOP          0.011808   0.006988   1.690   0.0989 .
BatchB13    -0.130901   0.080821  -1.620   0.1132
BatchB14    -0.082366   0.060811  -1.354   0.1832
BatchB15     0.031996   0.080754   0.396   0.6940

Residual standard error: 0.1613 on 40 degrees of freedom
Multiple R-squared:  0.1573,   Adjusted R-squared:  0.07304
F-statistic: 1.867 on 4 and 40 DF,  p-value: 0.1352
```

The P-value is greater than 0.05, so the null hypothesis is not rejected again. Therefore, I cannot claim that the expression level of ANGPT2 is significantly varied in four different Batch subgroups.

Then, I decided to divide the Batch group into just two subgroups as below and run the t-test. Division 1) divided Batch half by half (B12 and B13 in the first group and B14 and B15 in the second group). 2) separate B14 from others.

Division 1): Devided Batch half by half (B12 and B13 in first group and B14 and B15 in the second group)

```
summary(corrTable_Batch2$Batch)
B_1    B_2
 18     27
```

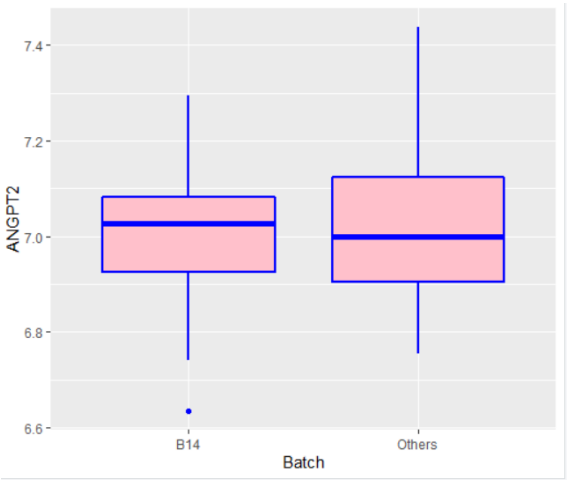



```
Anova_results <- aov(ANGPT2 ~ Batch, corrTable_Batch2); summary(Anova_results)
            Df Sum Sq  Mean Sq F value Pr(>F)
Batch        1  0.002 0.001991   0.069  0.793
Residuals   43  1.233 0.028675

Anova_results <- aov(ANGPT2 ~ IOP + Batch, corrTable_Batch2); summary(Anova_result
s)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.305  0.136
Batch        1 0.0012 0.00115   0.041  0.840
Residuals   42 1.1697 0.02785

Anova_results <- aov(ANGPT2 ~ IOP + Batch + Age, corrTable_Batch2); summary(Anova_
results)
            Df Sum Sq Mean Sq F value Pr(>F)
IOP          1 0.0642 0.06420   2.273  0.140
Batch        1 0.0012 0.00115   0.041  0.841
Age          3 0.0681 0.02269   0.803  0.500
Residuals   39 1.1016 0.02825

t.test(ANGPT2 ~ Batch, corrTable_Batch2, mu=0, alt="two.sided", conf=0.95, var.eq=
F, paired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Batch
t = -0.25607, df = 32.962, p-value = 0.7995
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.12145519  0.09430056
sample estimates:
mean in group B_1 mean in group B_2
         7.000838          7.014415
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

Division 2): Separate B14 from others

```
summary(corrTable_Batch2$Batch)
   B14     Others
    21         24
```




```
Anova_results <- aov(ANGPT2 ~ Batch, corrTable_Batch2); summary(Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
Batch         1  0.009 0.00899   0.315  0.577
Residuals    43  1.226 0.02851

Anova_results <- aov(ANGPT2 ~ IOP + Batch, corrTable_Batch2); summary(Anova_result
s)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.378  0.131
Batch         1 0.0366 0.03665   1.357  0.251
Residuals    42 1.1342 0.02700

Anova_results <- aov(ANGPT2 ~ IOP + Batch + Age + Sex, corrTable_Batch2); summary(
Anova_results)
             Df Sum Sq Mean Sq F value Pr(>F)
IOP           1 0.0642 0.06420   2.388  0.131
Batch         1 0.0366 0.03665   1.363  0.250
Age           3 0.0798 0.02661   0.990  0.408
Sex           2 0.0597 0.02986   1.111  0.340
Residuals    37 0.9946 0.02688

t.test(ANGPT2 ~ Batch, corrTable_Batch2, mu=0, alt="two.sided", conf=0.95, var.eq=
F, paired=F)

        Welch Two Sample t-test

data:  ANGPT2 by Batch
t = -0.56261, df = 42.503, p-value = 0.5767
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1299189  0.0732570
sample estimates:
   mean in group B14 mean in group Others
            6.993875             7.022205
```

In all tests and analyses, P-values are higher than 0.05, and I cannot reject the null hypothesis.

All analysis above indicates that there is no significant relation between IOP and ANGPT2 based on different subgroups of the Batch feature.

Nonlinear Correlation Analysis

In this section, I focused on nonlinear correlation analysis, which I did on the data. I used two NonLinear Correlation (**nlcor**) and Nonlinear Nonparametric Statistics (**NNS**) tests. The only thing I found from nlcor analysis is: "When PTPRB highly expressed, ANGPT2 expressed low".

```
nlcor(corrTable$PTPRB, corrTable$ANGPT2, plt=T)
```

x = corrTable$PTPRB y= corrTable$ANGPT2