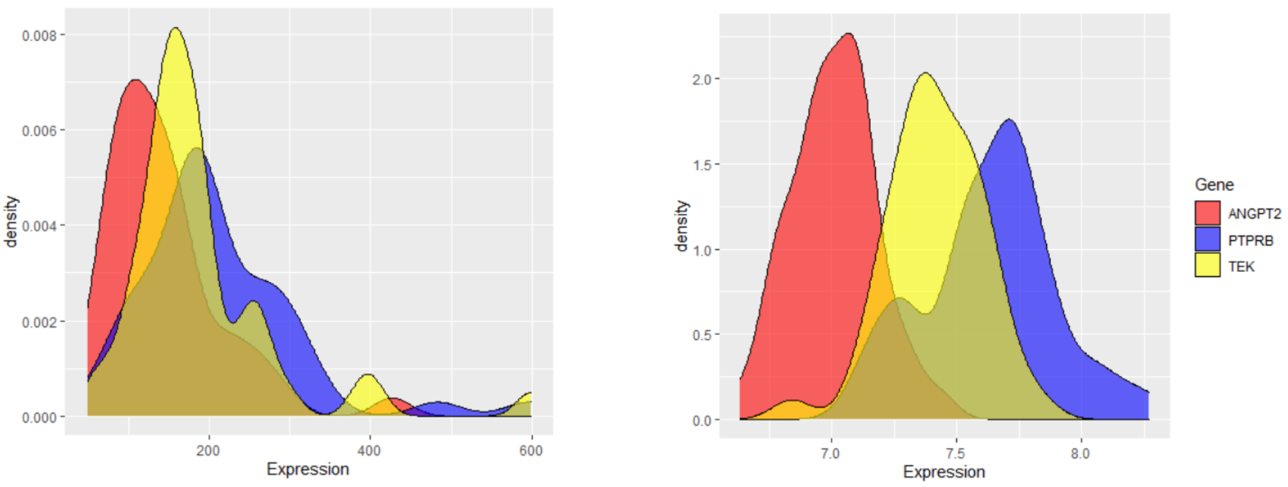
Final report of RNAseq correlation analysis

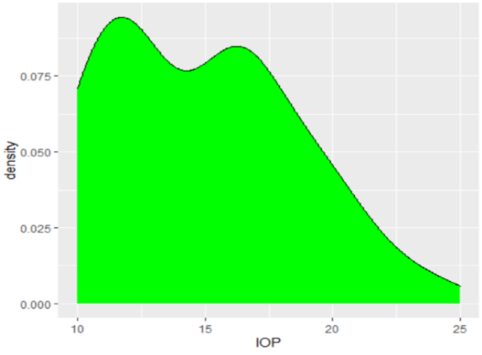
This is the final report of analyzing the RNAseq data and looking for the correlation of IOP data with three genes of 45 samples. First of all, I normalized the expression levels of 32883 genes. I had some options and finally chose the **rlog** (regularized log transformation) method, among others (Log2, Log10, Ln, vst). Then I considered three genes among them.



Real counts

Normalized with rlog

Also, I found that the IOP dataset for 45 samples is not normally distributed. I have tried using 7 different methods to normalize it, but unfortunately, I could not; almost all of them are skewed right. Although, the **Log10 transformation normalization** was the best normalization method for IOP, however, none of them does affect the correlation analysis significantly. Therefore, I prefer to use real IOP data without any normalization.

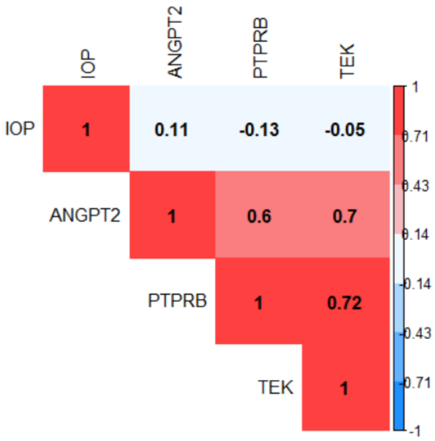


Then, I tried to find a significant correlation between IOP and three genes. There are two primary correlation calculation methods: Pearson (looking for a linear relationship) and Spearman (looking for linear and end-skewed correlation). Between them, I choose the **Spearman** method for the subsequent analysis based on the review of three genes density distribution.

I tried to find a strong correlation, but I could not find any significant relationship, just a non-significant and weak correlation of IOP with ANGPT2 gene.

Real counts

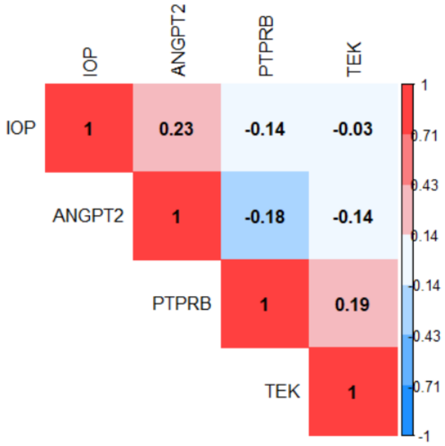
Correlation method: Pearson



	IOP	ANGPT2	PTPRB	TEK
IOP	1	0.11	-0.13	-0.05
ANGPT2		1	0.6	0.7
PTPRB			1	0.72
TEK				1

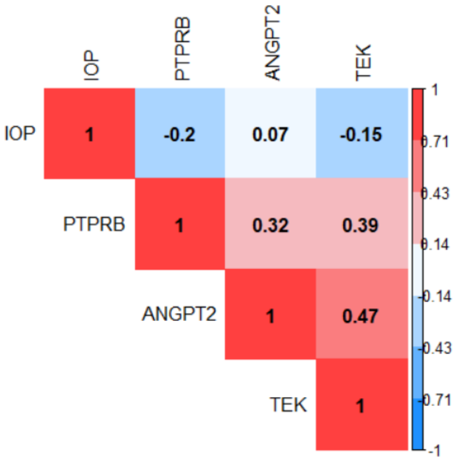
rlog Normalization

Correlation method: Pearson



	IOP	ANGPT2	PTPRB	TEK
IOP	1	0.23	-0.14	-0.03
ANGPT2		1	-0.18	-0.14
PTPRB			1	0.19
TEK				1

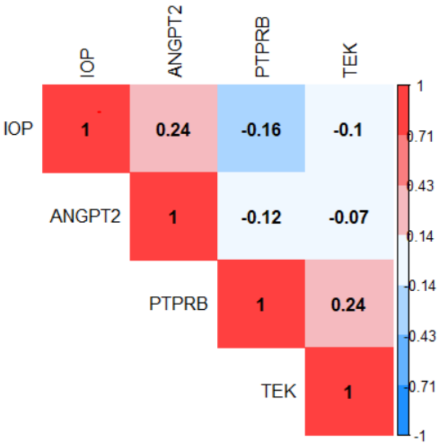
Correlation method: Spearman



	IOP	PTPRB	ANGPT2	TEK
IOP	1	-0.2	0.07	-0.15
PTPRB		1	0.32	0.39
ANGPT2			1	0.47
TEK				1

Selected method

Correlation method: Spearman

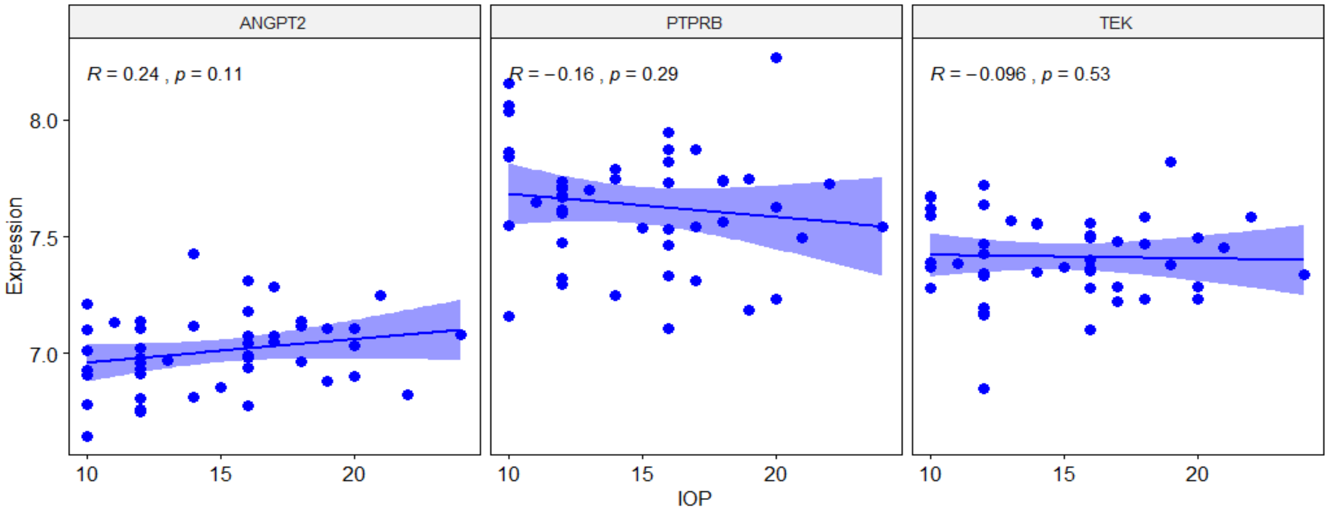


	IOP	ANGPT2	PTPRB	TEK
IOP	1	0.24	-0.16	-0.1
ANGPT2		1	-0.12	-0.07
PTPRB			1	0.24
TEK				1

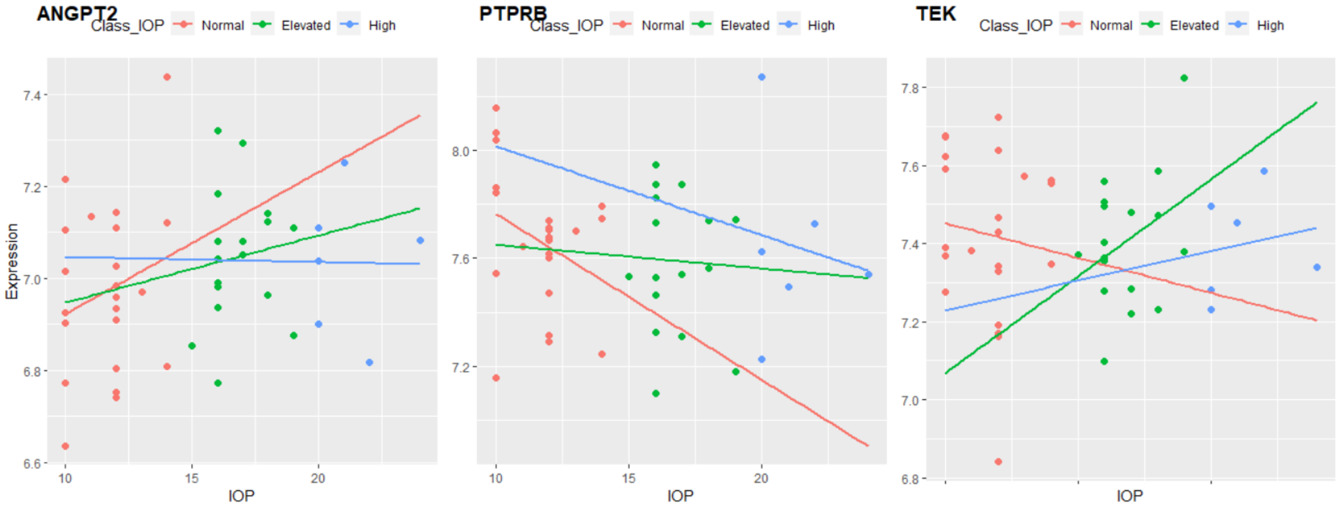
Based on the heatmap figures above, I found the two genes **PTPRB** and **TEK** act similarly. They have a negative correlation with **IOP** and **ANGPT2**. Also, **IOP** has a weak positive linear relationship with **ANGPT2** and a weak negative linear relationship with **PTPRB**.

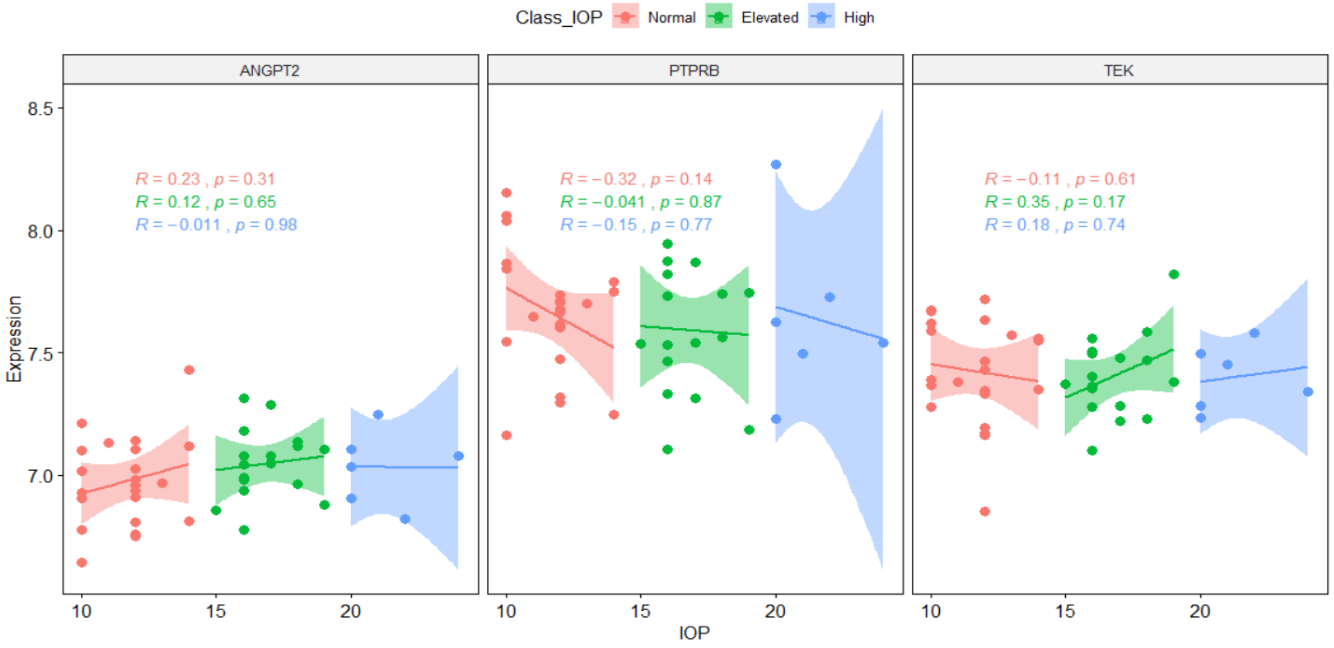
Correlation between IOP and Genes

Considering the **Spearman** correlation method and real IOP (without normalization) and three genes' expression level (normalized using the **rlog** method) calculated the correlation. 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.

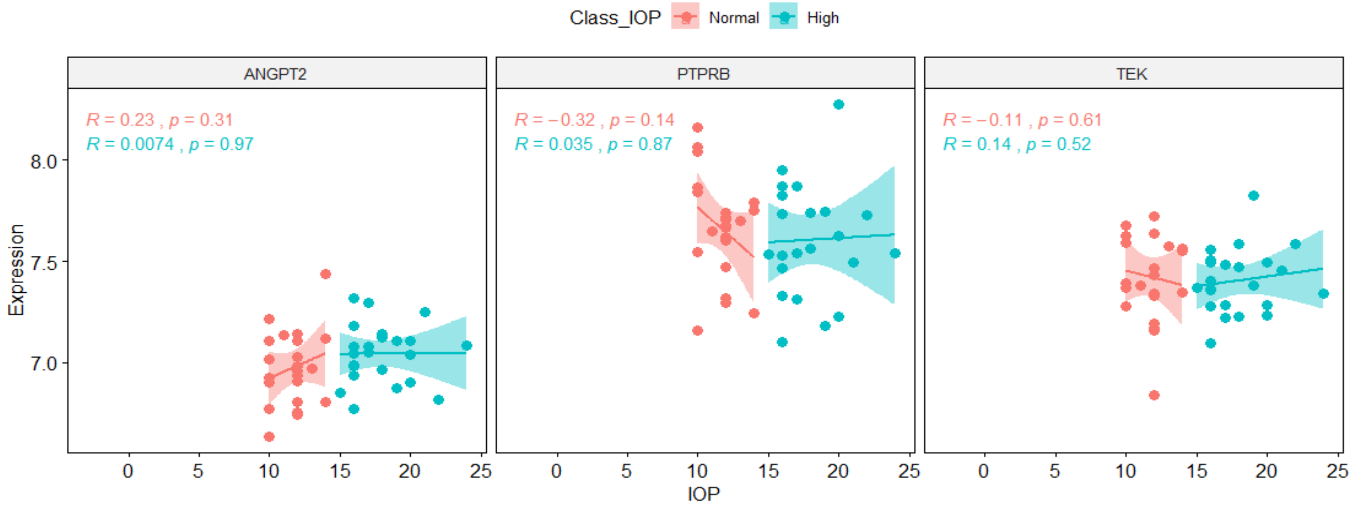


Then, I focused on the three subgroups of IOP (Normal, Elevated, and High) and calculated the correlation based on those data. As I can see, ANGPT2 has the highest positive correlation with Normal_IOP (R = 0.23) subgroup. Also, PTPRB has the highest negative correlation with Normal_IOP (R = -0.32) subgroup. 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. But, none of them are a significant correlation.



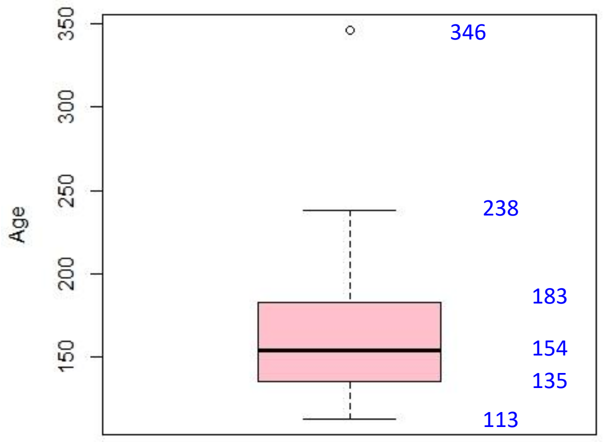


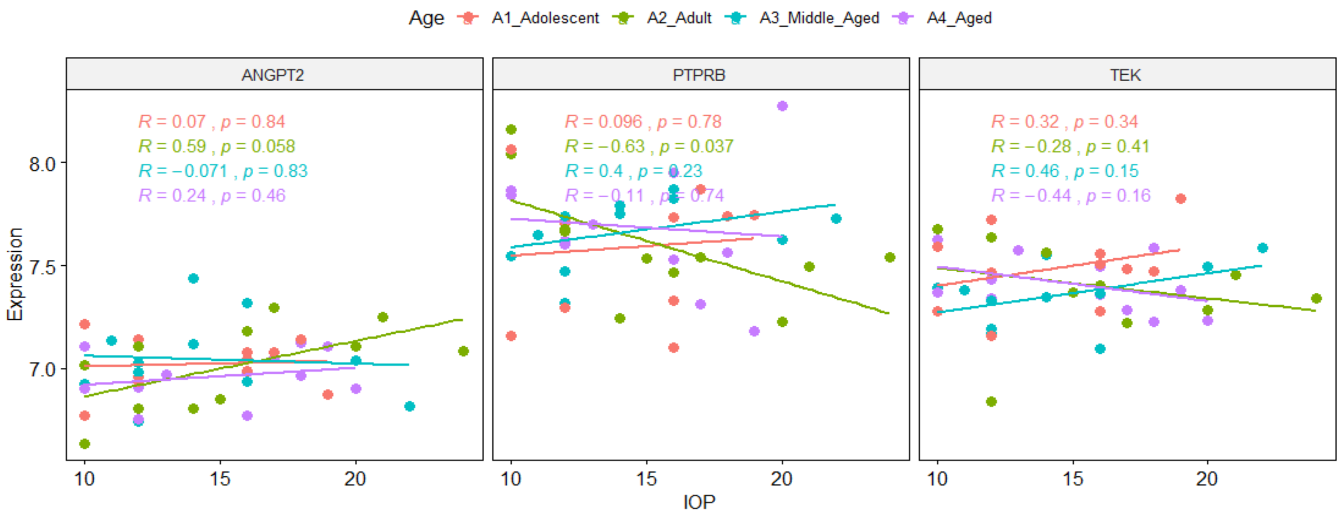
There are small samples in the high_IOP subgroup. Therefore, I decided to classify the IOP feature into two subgroups: Normal and High. In this scenario, half of the samples are High_IOP, around 50% of the samples (23 samples). However, the High_IOP subgroup is not correlated with ANGPT2 and PTPRB again. I mean, Normal_IOP has a vital role in the correlation of IOP with ANGPT2 or PTPRB still. But, none of them are a significant correlation again.

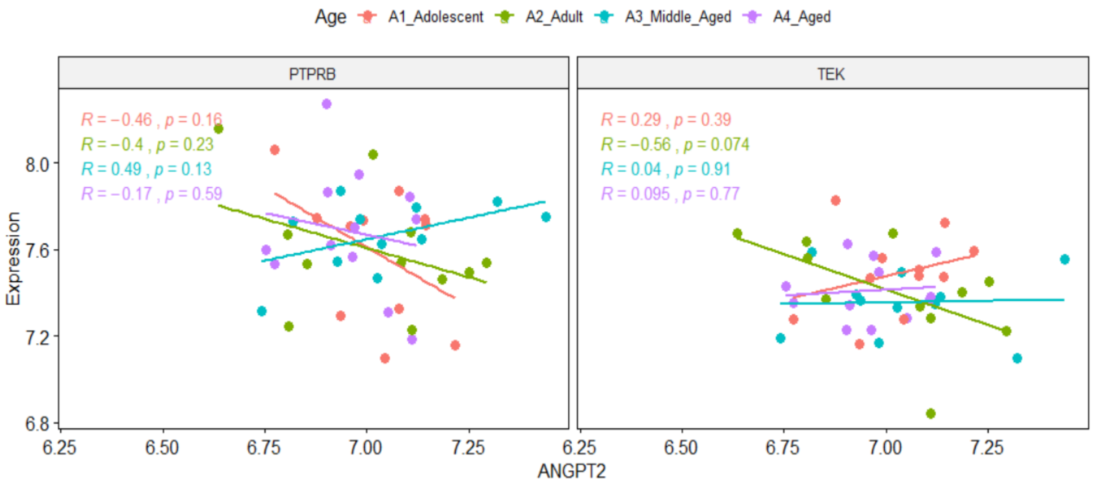


Correlation between IOP and Genes based on Age

In this section, I am going to calculate the correlation between IOP and Genes based on different Age subgroups. I classified samples into four subgroups (quartiles): 1) **Adolescent** (113<=Age<135), 2) **Adult** (135<=Age<154), 3) **Middle_Aged**(154<=Age<183), and 4) **Aged** (183<=Age<=346). Then, plot the correlation based on each subgroup.

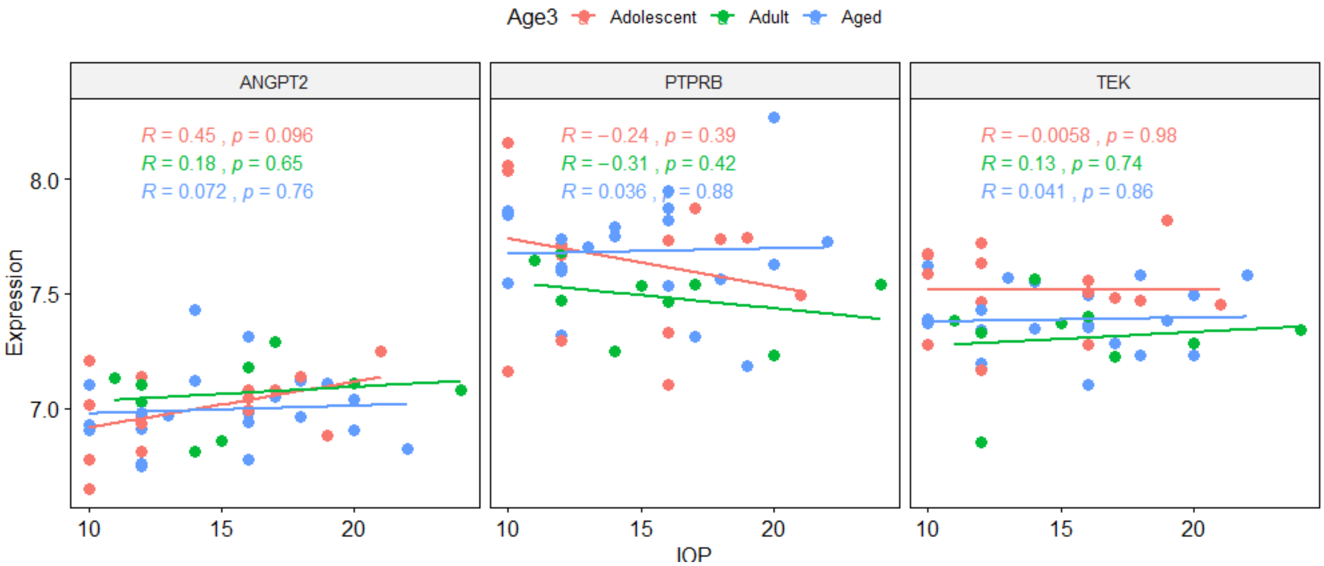




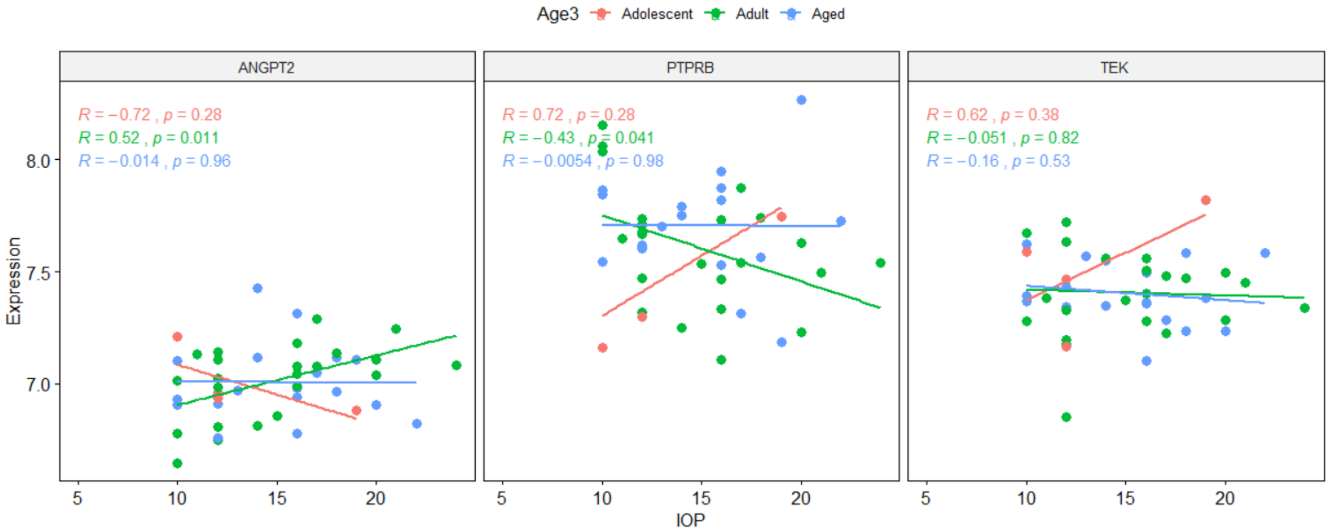


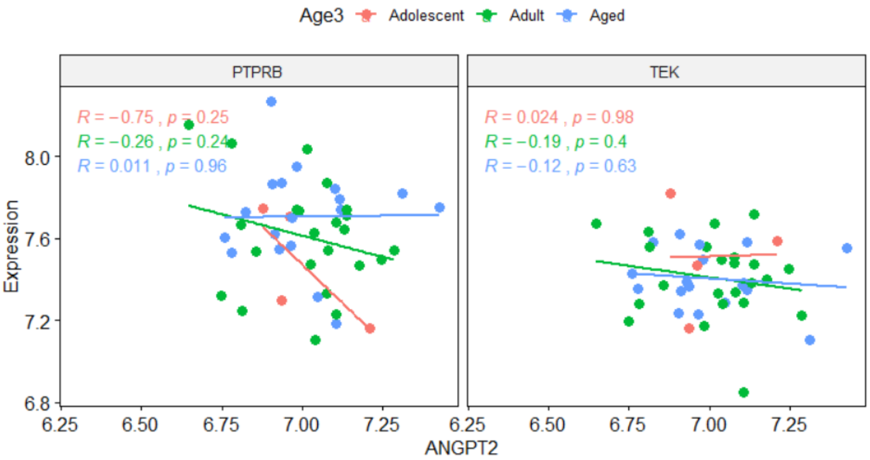
Good news: Based on the results, I can see there is a **semi-significant relationship** in the **Adult** subgroup between IOP and ANGPT2 (R=+0.59, P-value=0.058), and **significant relationship** PTPRB (R=-0.63, P-value=0.037). Also, there is a semi-significant relationship between ANGPT2 and TEK genes in this subgroup (Adult).

I am curious about the **Adult** subgroup. Then, I 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.



As I did not get consistent results, 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 almost sufficient sample-size (23) as the best subgroup, which I looked for. Then, I plot the correlation based on each subgroup.

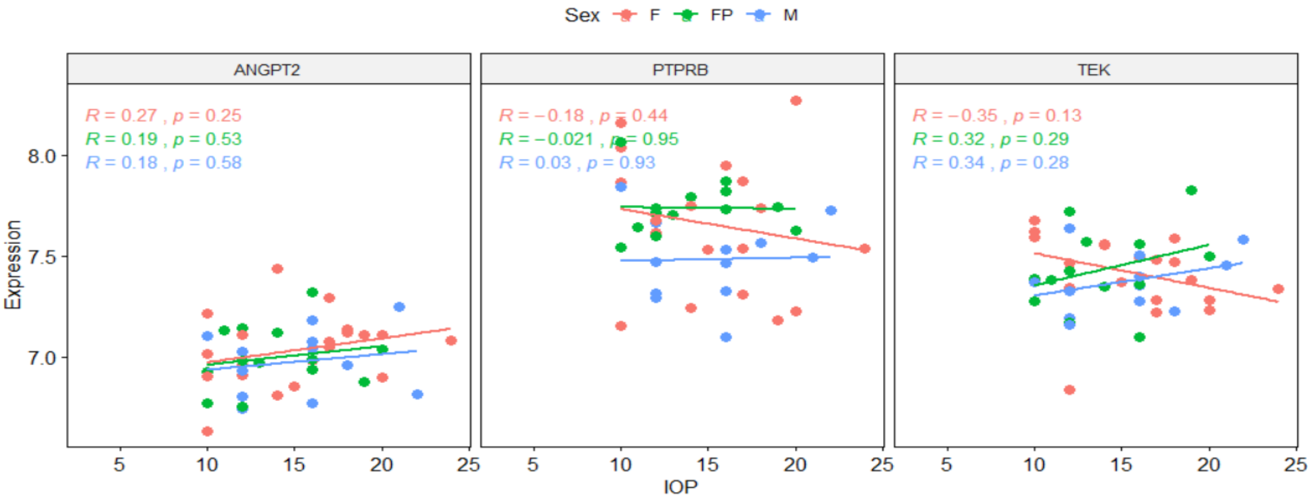


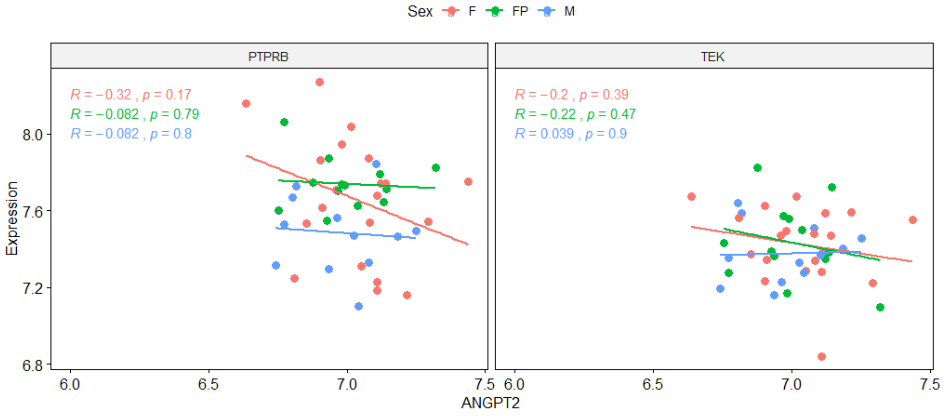


As you can see, there are **significant and reasonable relationship** in the **Adult** subgroup, between IOP and ANGPT2 (R=+0.52, P-value=0.011), and PTPRB (R=-0.43, P-value=0.041). In the next step, I tried to do the same analysis based on other features: Sex and Batch, as below.

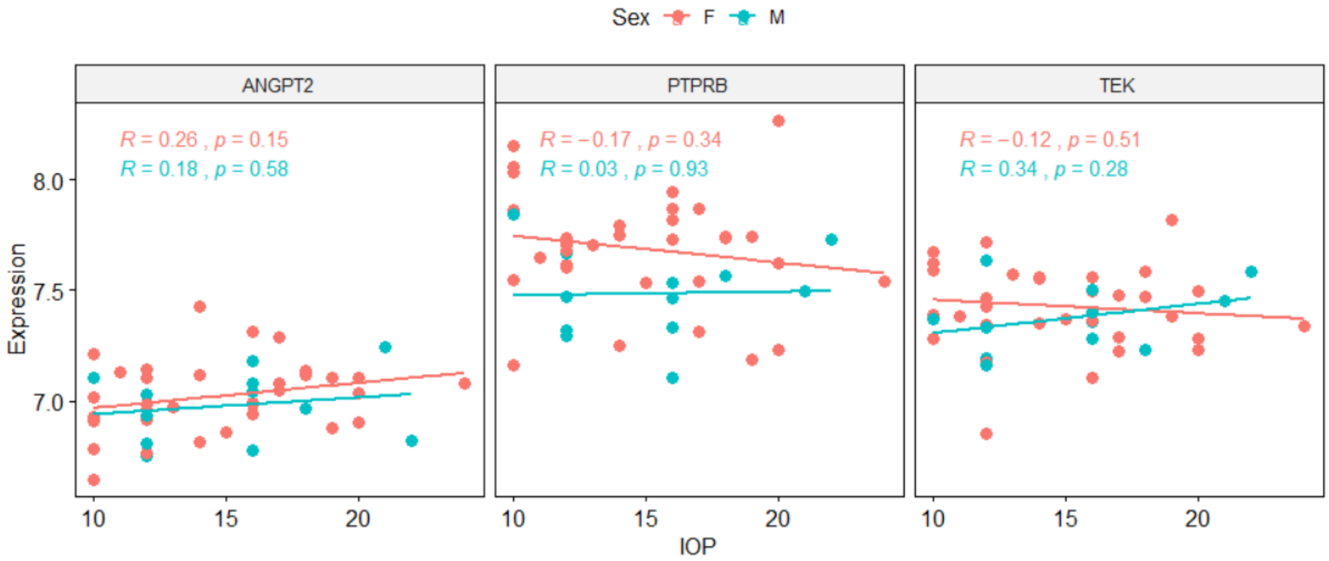
Correlation between IOP and Genes based on Sex

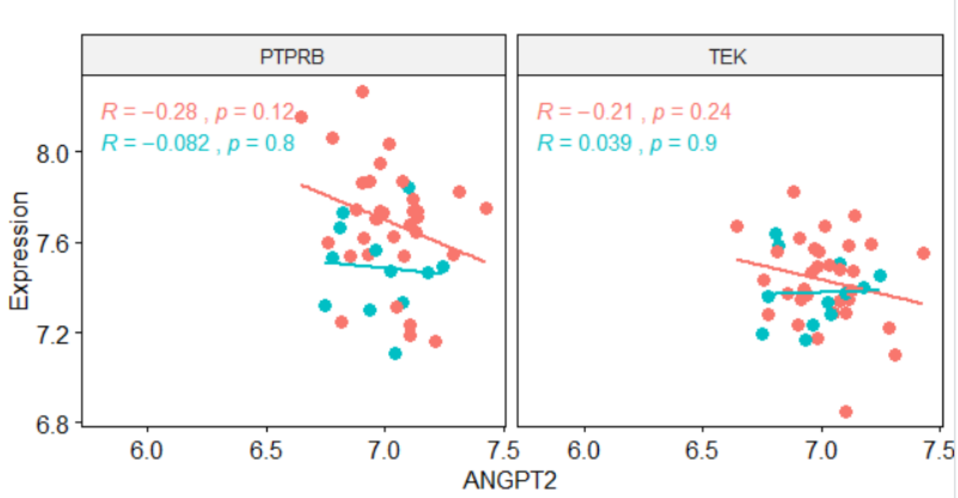
In this section, I am going to calculate the relationship between IOP and genes based on different Sex subgroups. Among 45 samples, 12 (26%) of them are male, and 33 (73%) are Female. Also, there are 13 (29%)Pregnant samples.





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 genes. Then, I divide samples into two subgroups: **Male** with 12 samples, and **Female** with 33 samples. Then, plot the correlation based on each subgroup. Again, there is no significant correlation between these Sex subgroups and genes.

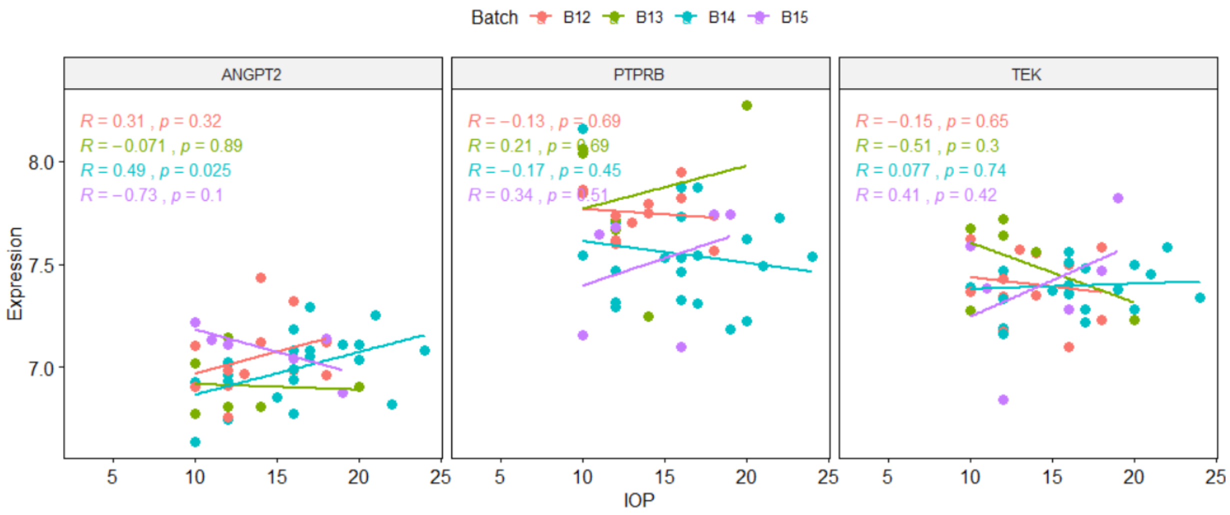


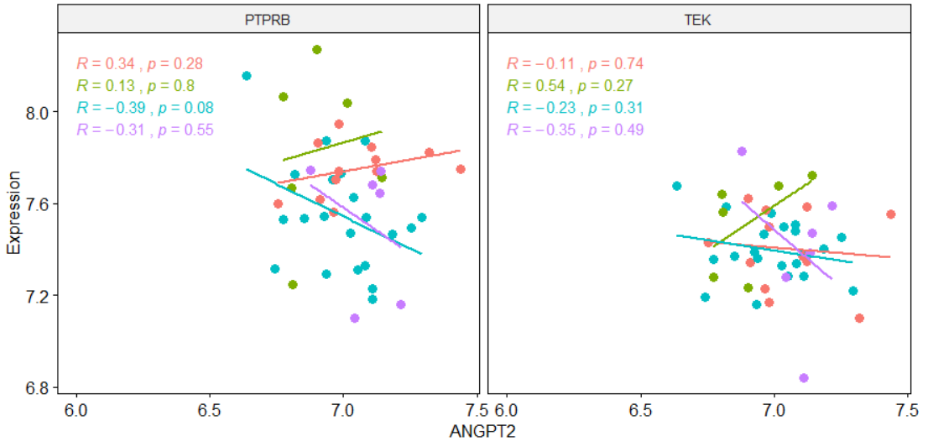


Partial-Correlation between IOP and ANGPT2 based on Batch feature

In this section, I am going to calculate the correlation between IOP and genes 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).

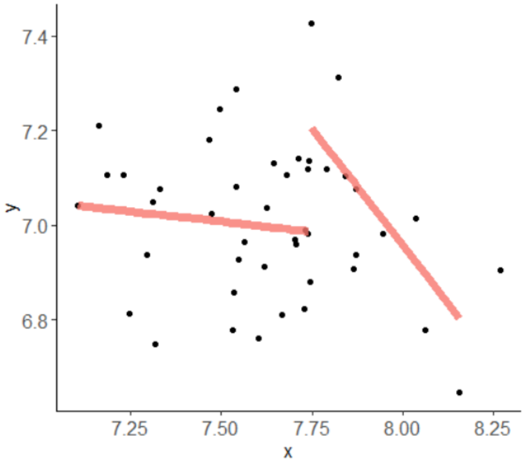
Good news: Based on the results, I can see there is a **significant relationship** in the **B14** subgroup between ANGPT2 and IOP (R=+0.49, P-value=0.025), and also ANGPT2 and PTPRB genes (R=-0.39, P-value=0.08). As these subgroups are classified experimentally, so I did not consider to maneuver on the reclassification of them and look for the best subgroups.





Nonlinear Correlation Analysis

Also, I tried to do nonlinear correlation analysis too. 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". In the figure below, you can see the results: x=PTPRB and y=ANGPT2.

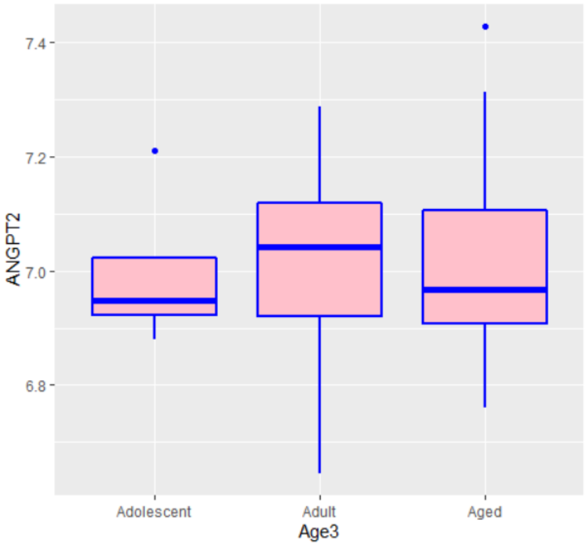


ANOVA Analysis

In this section, I tried to do ANOVA analysis for three subgroups of Age (best fitted one).

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

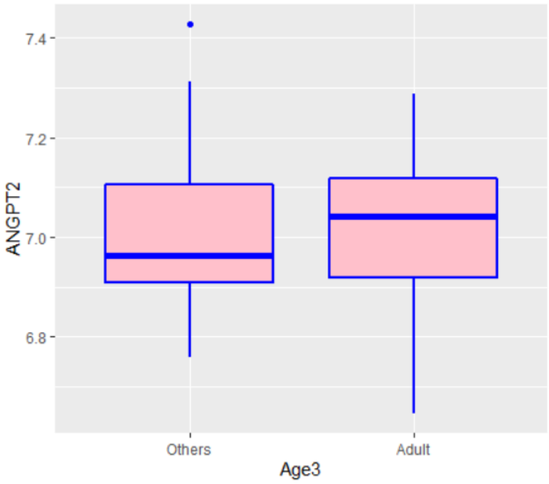
Ha: They are not equal.



```
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. Therefore, I cannot claim that the expression level of ANGPT2 is significantly varied in three different Age3 groups. It means that these three subgroups have similarities and hidden relationships, and we cannot differentiate them in the ANGPT2 expression level.

Then, I decided to divide the Age3 into just two subgroups and run the t-test for a similar hypothesis. I kept the **Adult** subgroup and mixed the **Adolocent** and **Aged** subgroups as the same subgroup. Therefore, the samples size of each new subgroups are almost equal; **Adult** with 23 samples and new **Others** subgroup with 22 samples.



```
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
```

Again, P-values are higher than 0.05, and I cannot reject the null hypothesis, the same conclusion.