The data leukemiasEset

The dataset is from 60 bone marrow samples of patients with one of the four main types of leukemia

(ALL, AML, CLL, and CML) and non-leukemia controls

- Acute Lymphoblastic Leukemia (ALL). Subtype: c-ALL / pre-B-ALL withoutt(9;22)
- Acute Myeloid Leukemia (AML). Subtype: Normal karyotype
- Chronic Lymphocytic Leukemia (CLL)
- Chronic Myeloid Leukemia (CML)
- Non-leukemia and healthy bone marrow (NoL)- this one is the 'heatlhyreference group'

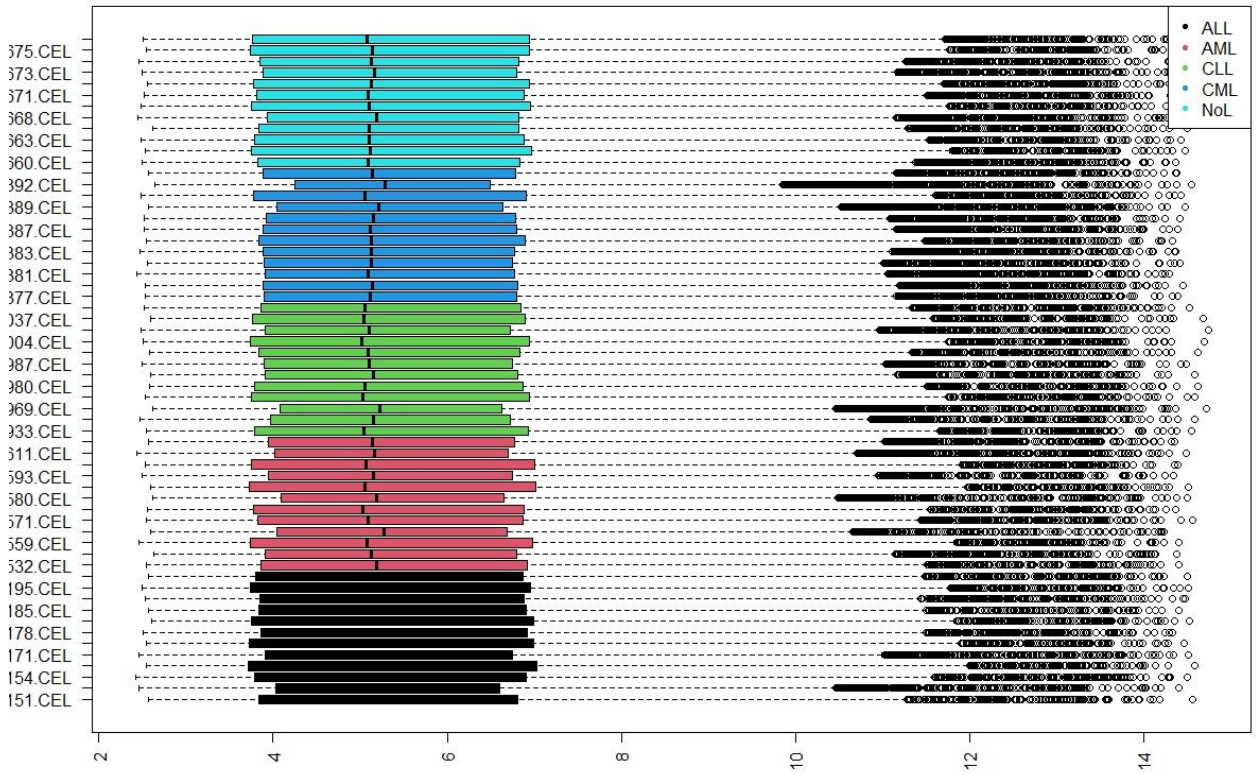
	Subject 1-Subject 12	Subject 13-Subject 24	Subject 25-Subject 36	Subject 37-Subject 48	Subject 49-Subject 60
Gene 1					
Gene 2					
:					
:					
Gene M					

In total there are 20172 genes, 60 samples in this dataset

Let's check the dimension of the data. In this data we have 20172 rows and 60 columns

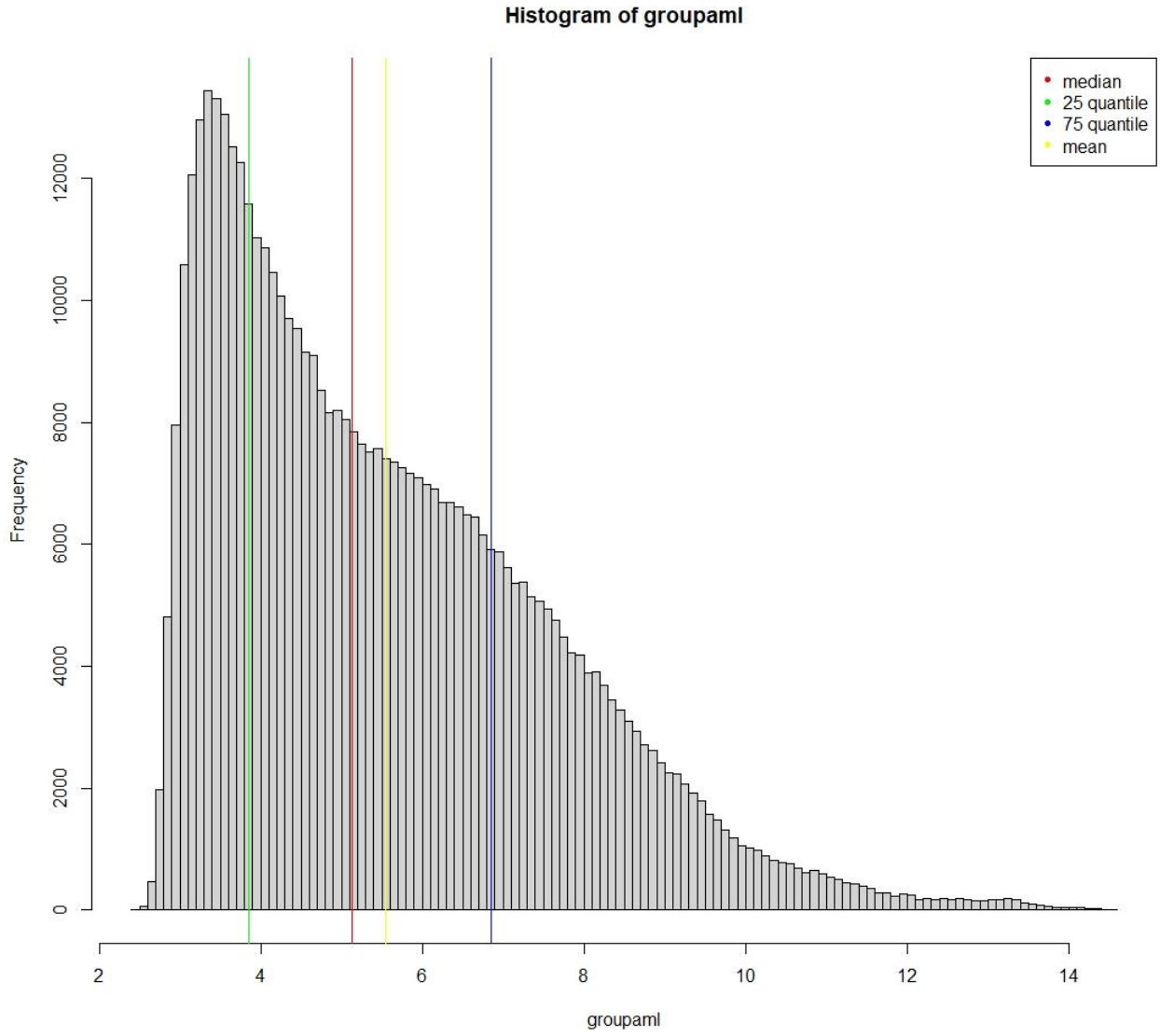
We going to choose to compare the AML and NOL group based on the last digit of my id

Lets explore our whole dataset first before we start



<u>Lets Examine data distributions for all individual patients using boxplot.The graph uses different colors to represent cancer types or normal tissue</u>

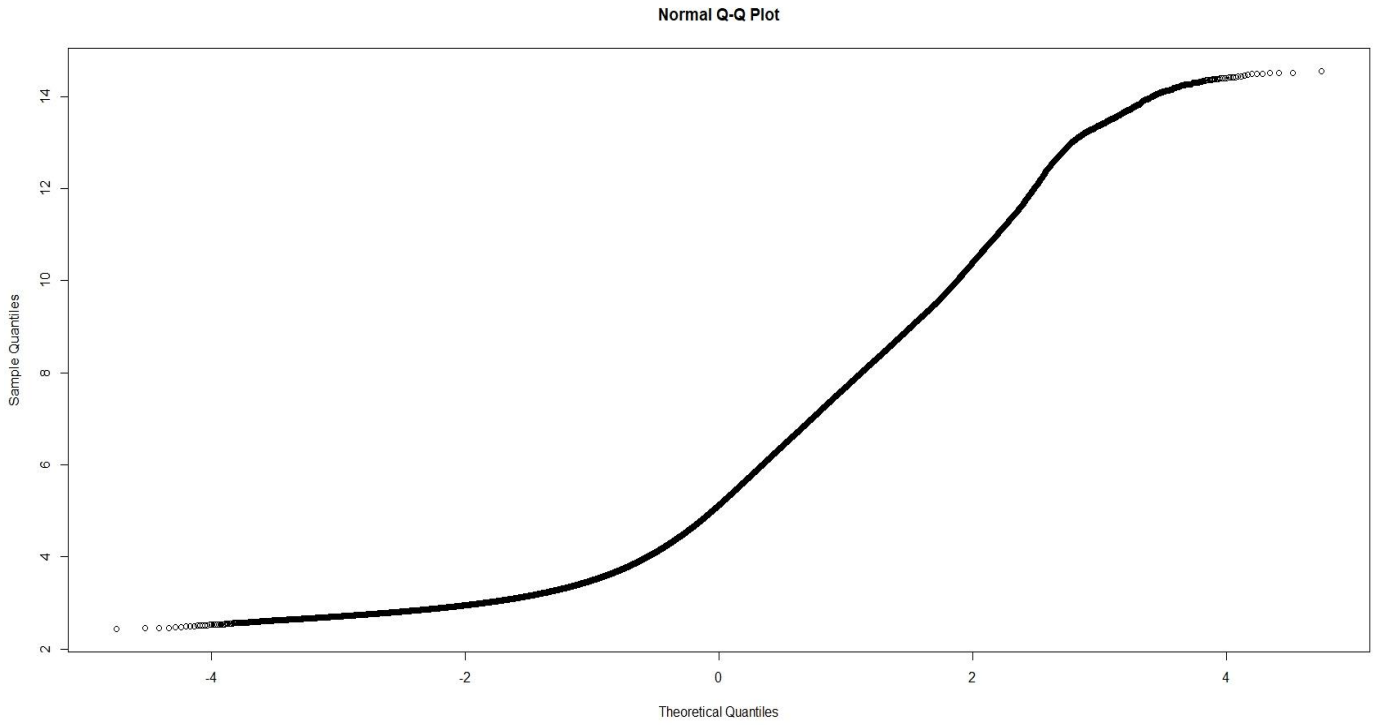
1.1 **Explore and visualize the data. Focus on the research question and try to visually describe the variability of the average gene expression between the two groups.Produce some meaningful summaries and descriptive statistics for your dataset.**



In terms of distribution it seems that our data are not following normal distribution

Most of the observations are in the first quintile

An alternative way to check for distribution is with qqplot although because of the amount of data it may take some time for the creation depending of the machine of course



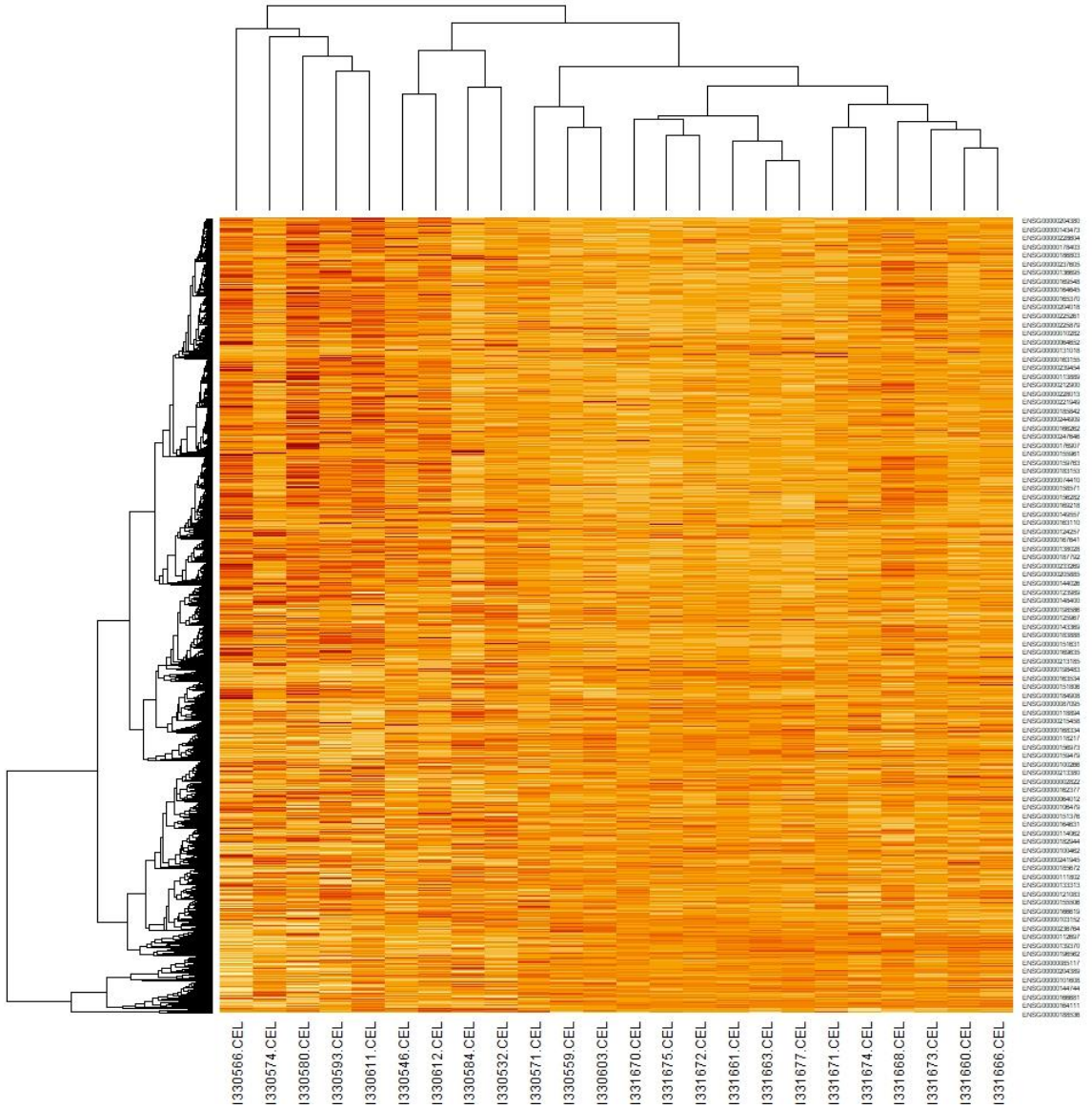
Still no normal distribution

Some of our descriptive statistics

```
> psych::describe(groupaml)
              vars     n mean   sd median trimmed  mad  min   max range skew kurtosis   se
GSM330532.CEL    1 20172 5.56 2.01   5.18    5.37 2.18 2.55 14.41 11.86 0.81     0.33 0.01
GSM330546.CEL    2 20172 5.53 1.97   5.13    5.33 2.06 2.62 14.38 11.76 0.90     0.60 0.01
GSM330559.CEL    3 20172 5.55 2.13   5.08    5.33 2.25 2.46 14.40 11.94 0.85     0.19 0.02
GSM330566.CEL    4 20172 5.53 1.81   5.27    5.37 1.93 2.59 14.23 11.64 0.84     0.71 0.01
GSM330571.CEL    5 20172 5.55 2.09   5.09    5.33 2.13 2.54 14.56 12.02 0.91     0.43 0.01
GSM330574.CEL    6 20172 5.51 2.07   5.04    5.29 2.14 2.56 14.36 11.80 0.88     0.32 0.01
GSM330580.CEL    7 20172 5.57 1.86   5.19    5.37 1.81 2.62 14.50 11.88 1.02     1.04 0.01
GSM330584.CEL    8 20172 5.53 2.12   5.06    5.32 2.26 2.59 14.51 11.92 0.83     0.17 0.01
GSM330593.CEL    9 20172 5.54 1.95   5.15    5.34 1.99 2.50 14.35 11.85 0.90     0.55 0.01
GSM330603.CEL   10 20172 5.56 2.14   5.07    5.34 2.24 2.53 14.39 11.86 0.85     0.21 0.02
GSM330611.CEL   11 20172 5.58 1.97   5.16    5.36 1.88 2.43 14.49 12.06 1.05     0.99 0.01
GSM330612.CEL   12 20172 5.55 1.97   5.14    5.35 2.00 2.57 14.35 11.78 0.94     0.70 0.01
GSM331660.CEL   13 20172 5.54 2.08   5.09    5.32 2.13 2.49 14.36 11.87 0.91     0.47 0.01
GSM331661.CEL   14 20172 5.55 2.10   5.12    5.34 2.27 2.53 14.47 11.95 0.84     0.27 0.01
GSM331663.CEL   15 20172 5.54 2.09   5.11    5.33 2.19 2.48 14.42 11.94 0.88     0.38 0.01
GSM331666.CEL   16 20172 5.54 2.04   5.11    5.32 2.10 2.62 14.50 11.88 0.91     0.51 0.01
GSM331668.CEL   17 20172 5.59 2.03   5.19    5.37 2.06 2.45 14.49 12.04 0.93     0.61 0.01
GSM331670.CEL   18 20172 5.53 2.09   5.11    5.33 2.26 2.49 14.39 11.90 0.82     0.24 0.01
GSM331671.CEL   19 20172 5.55 2.10   5.09    5.33 2.16 2.52 14.41 11.88 0.91     0.46 0.01
GSM331672.CEL   20 20172 5.55 2.10   5.13    5.34 2.23 2.56 14.39 11.83 0.86     0.33 0.01
GSM331673.CEL   21 20172 5.54 2.01   5.16    5.33 2.09 2.49 14.48 11.99 0.92     0.64 0.01
GSM331674.CEL   22 20172 5.55 2.06   5.12    5.33 2.10 2.46 14.46 12.00 0.93     0.53 0.01
GSM331675.CEL   23 20172 5.53 2.10   5.14    5.33 2.27 2.54 14.38 11.85 0.83     0.26 0.01
GSM331677.CEL   24 20172 5.54 2.12   5.09    5.33 2.23 2.51 14.42 11.91 0.86     0.31 0.01
```

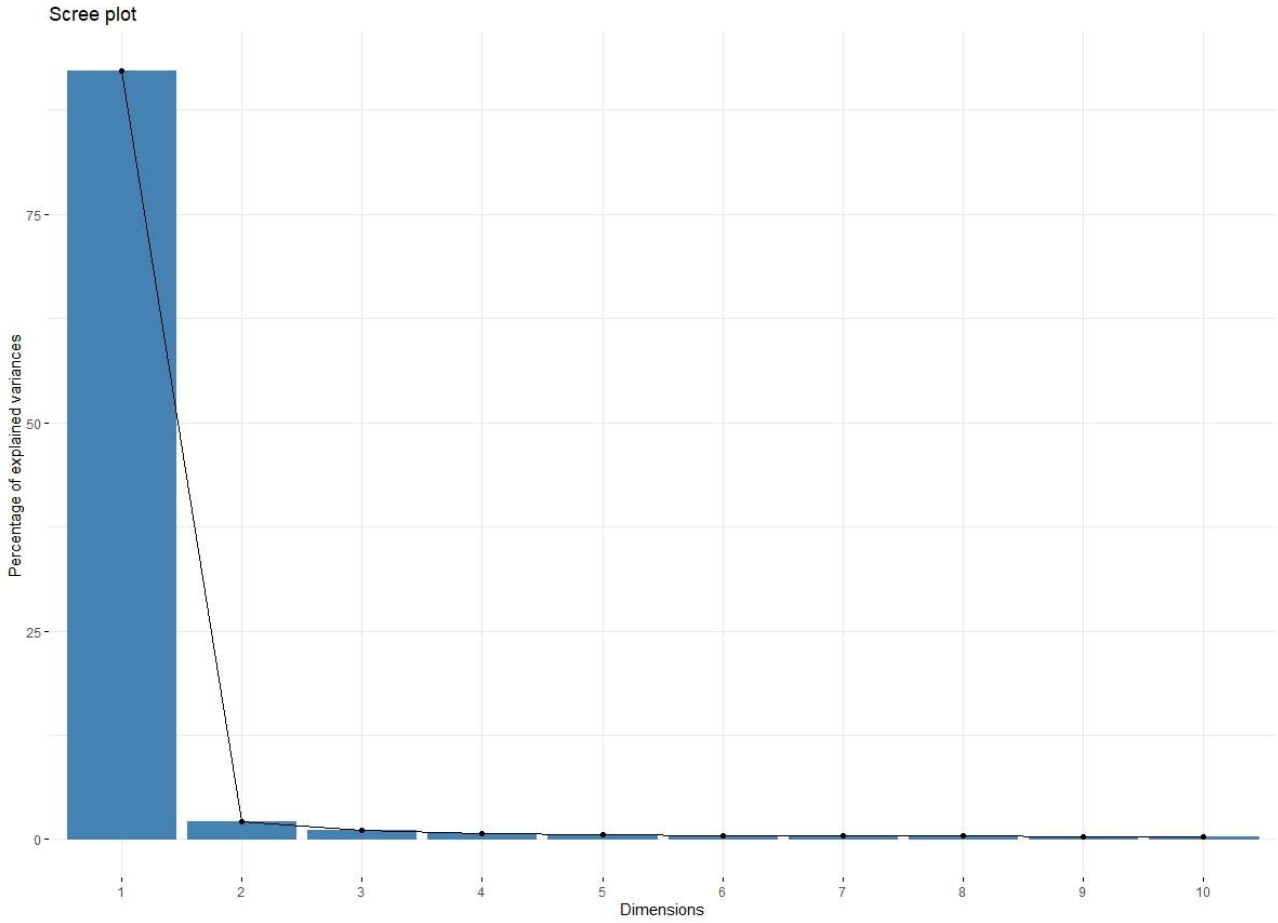
The first 12 observations are from the aml group and the last 12 from the nol group
We can notice pretty similar mean,sd,skewness in all of our groups

We can get a clear picture by our heatmap below



We can see some distinction between of our groups but by all means no clear distinction between our groups

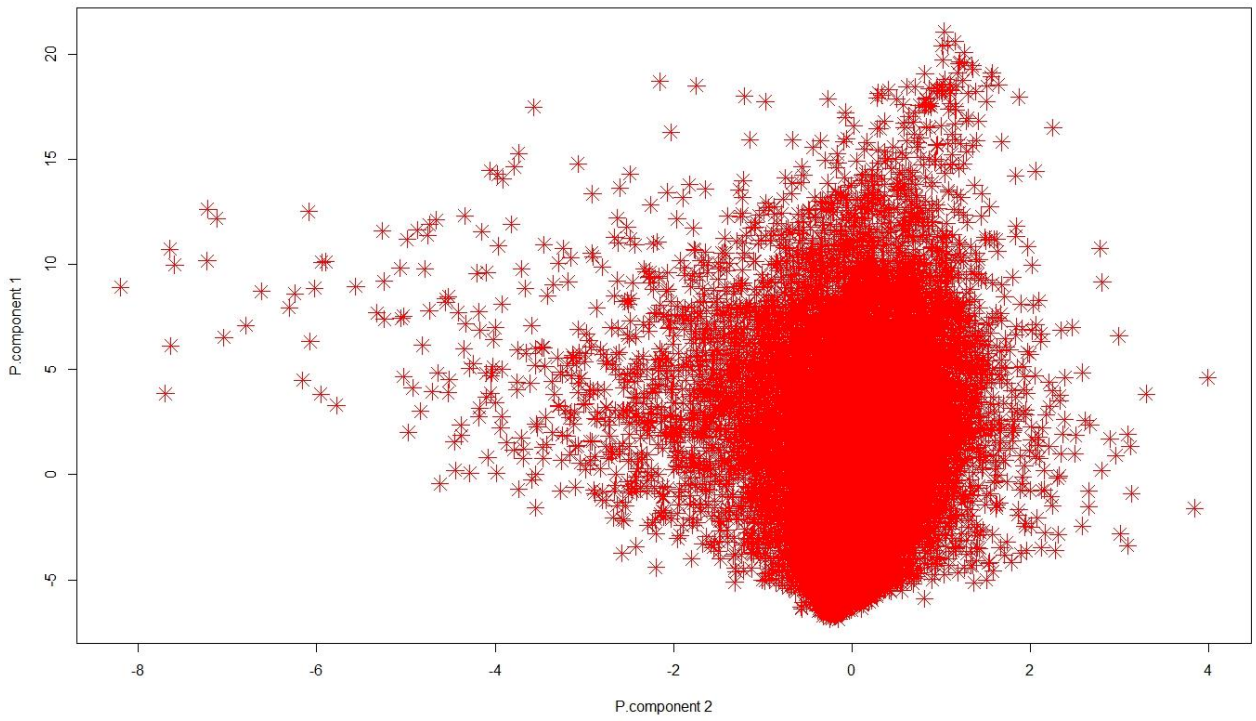
1.2 **Use PCA in order to visualize the dataset (20172 × 24). Project the data on thefirst few principal components and explain your findings. Do the same when considering the transposed input data (24 × 20172). Describe what you see.**



As we can see from the scree plot above we can choose the first, second and maybe 3rd principal components

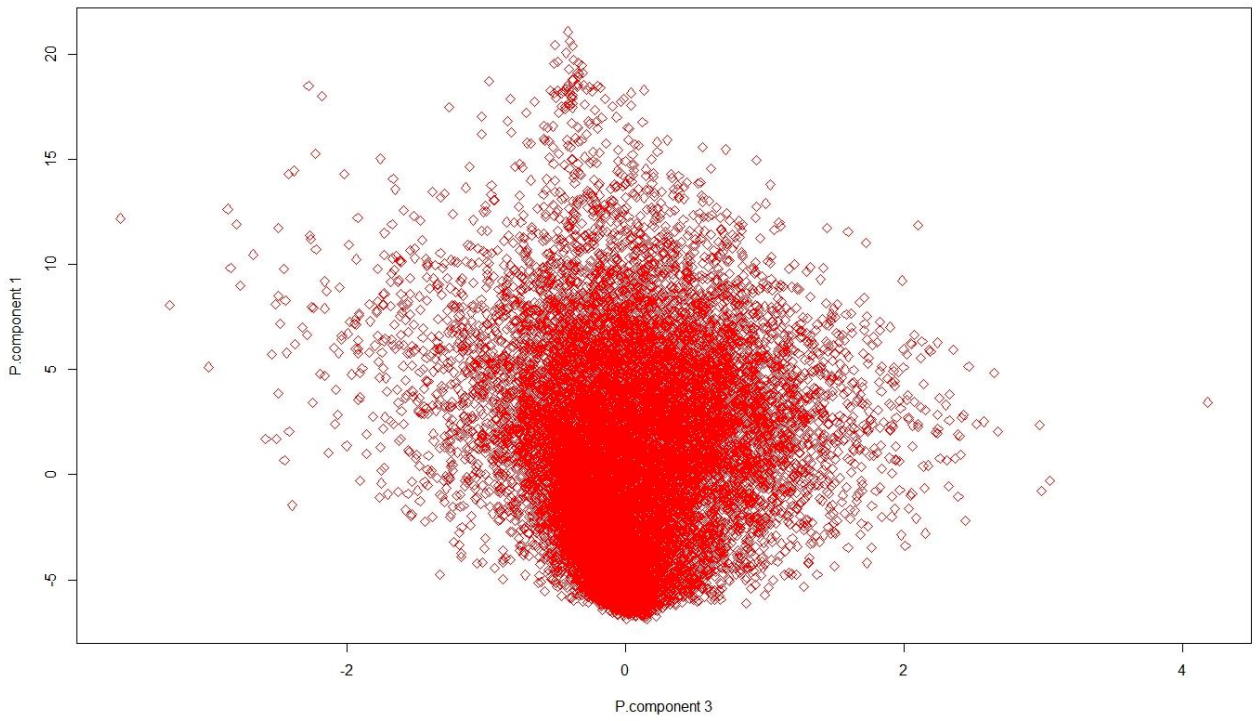
1st PC: corresponds to the largest eigenvalue (and the
Corresponding eigenvector)
2nd PC: corresponds to the second largest eigenvalue (and the
Corresponding eigenvector), and so on...

The projection of our first two principal components:



There is zero separation between our PCA1 and PCA2

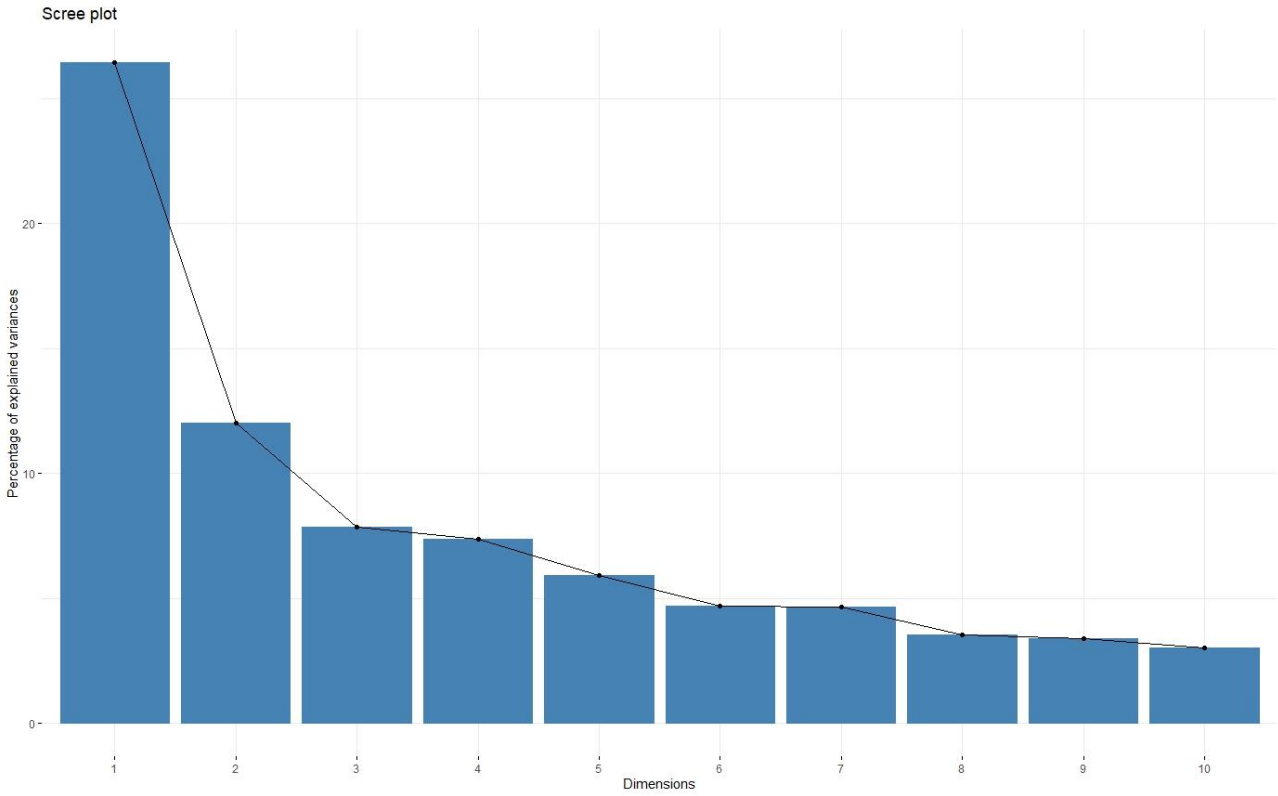
Ploting the PCA1 and PCA3 we get the same results lets check



As we can see it’s the same results

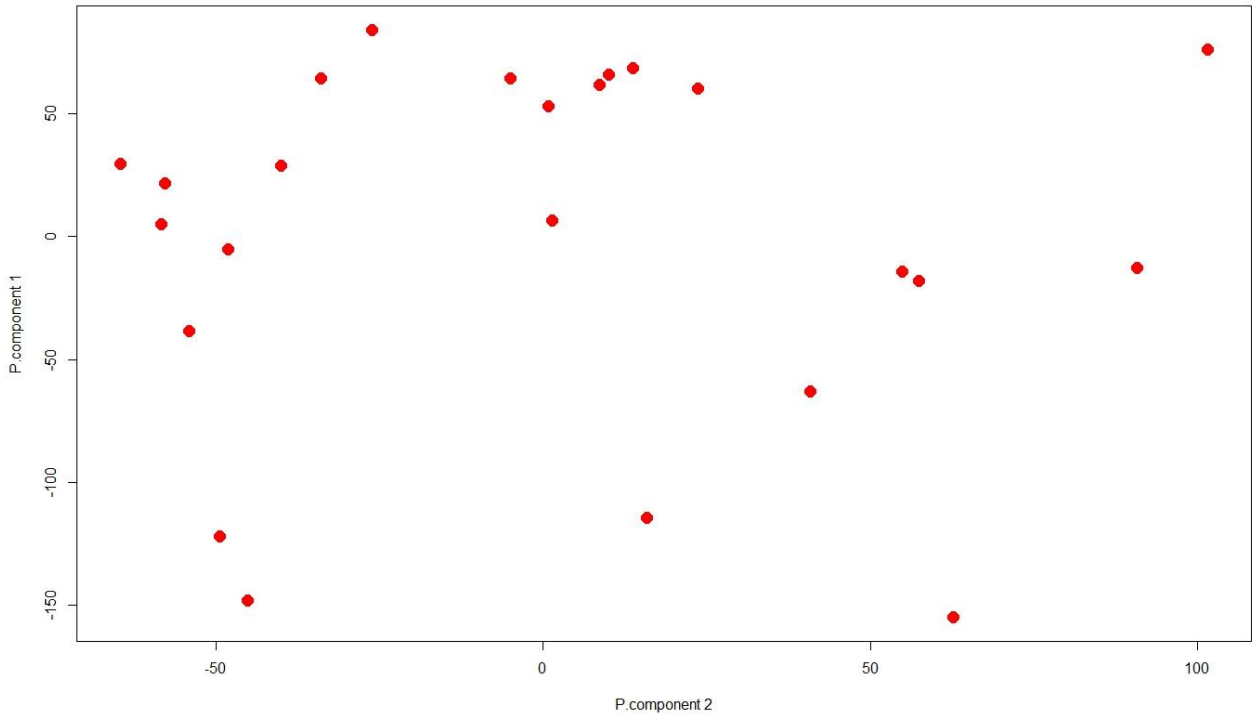
In fact every combination of groups we get no clear separation as we expect.

Lets see the scree plot with our transpose input data



More promising results.
lets project our PCA’S and it seems we can choose every pca to pair this time

1st pca with 3rd pca



We can see a much clearer separation in our transpose data

1.3 **Use two independent samples t-tests (you may assume that the variance is equalbetween groups) in order to test the null hypothesis per gene. State the null and alternative hypothesis per gene, as well as the assumptions you use to model the data. Plot a histogram (relative frequencies) of the p-values.**

In a multiple hypothesis testing problem, m
H_{0i} versus H_{1i},
for $i = 1,2, \dots, m$, simultaneously. Assume that the m tests are constructed based on observed p values, $p_1, \dots, p_m$, respectively. The unknown quantity π_0 to be estimated the proportion of the true null hypotheses.So the hypothesis is:

$H_{0(i)}$: m:aml = m:nol

$H_{1(i)}$: m:aml ≠ m:nol

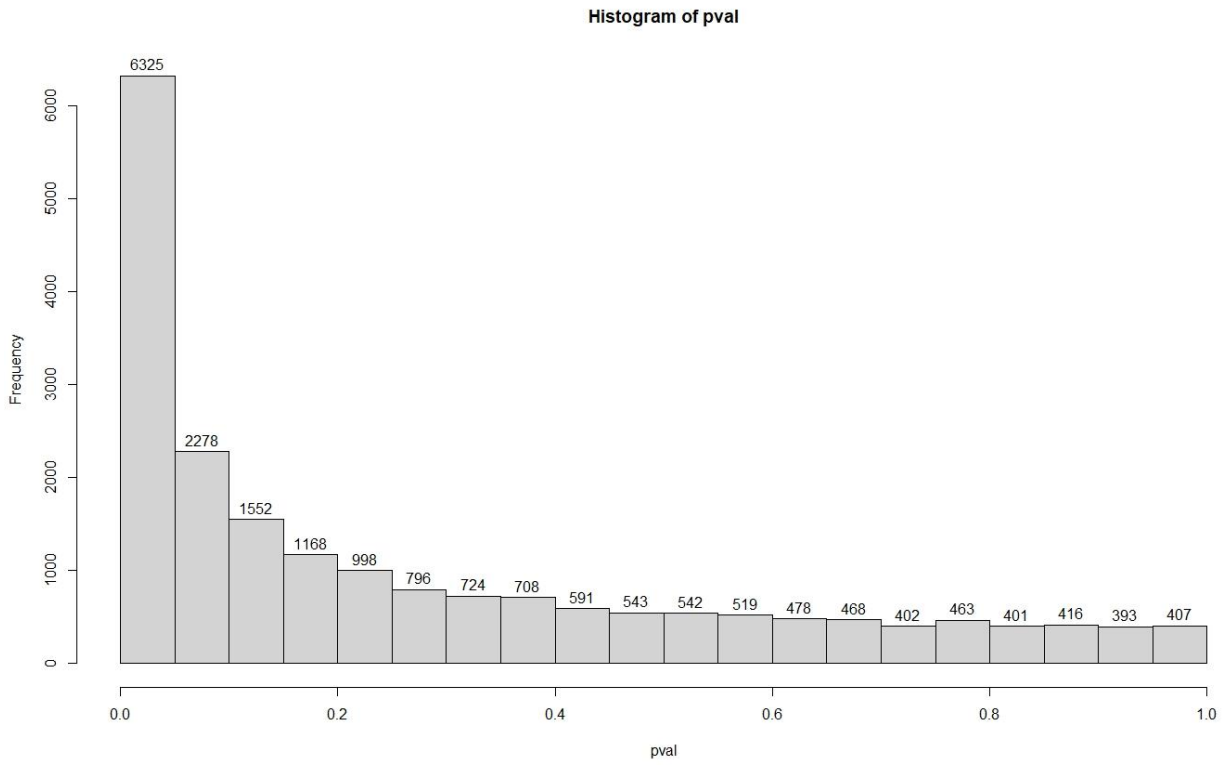
and

$$H_{01}, \dots, H_{0m}.\ H_{(i)} = \begin{cases} 0, & \text{if } H_{0(i)} \text{ is true} \\ 1, & \text{otherwise} \end{cases}$$

And the assumptions are:

1. The data follows continues scale
2. homogeneity of variance.Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal
3. a reasonably large sample size is used. A larger sample size means the distribution of results should approach a normal bell-shaped curve which is true in our case
4. the two samples are independent
5. he data is collected from a representative, randomly selected portion of the total population

The histogram for the p-values is the following



<u>As we expect most of the pvalues are gathered at the left of the graph</u>

1.4 **Can you give a rough estimate of the proportion of true null hypotheses?**

For a rough estimate we can do the following procedure.I notice for the rough only!

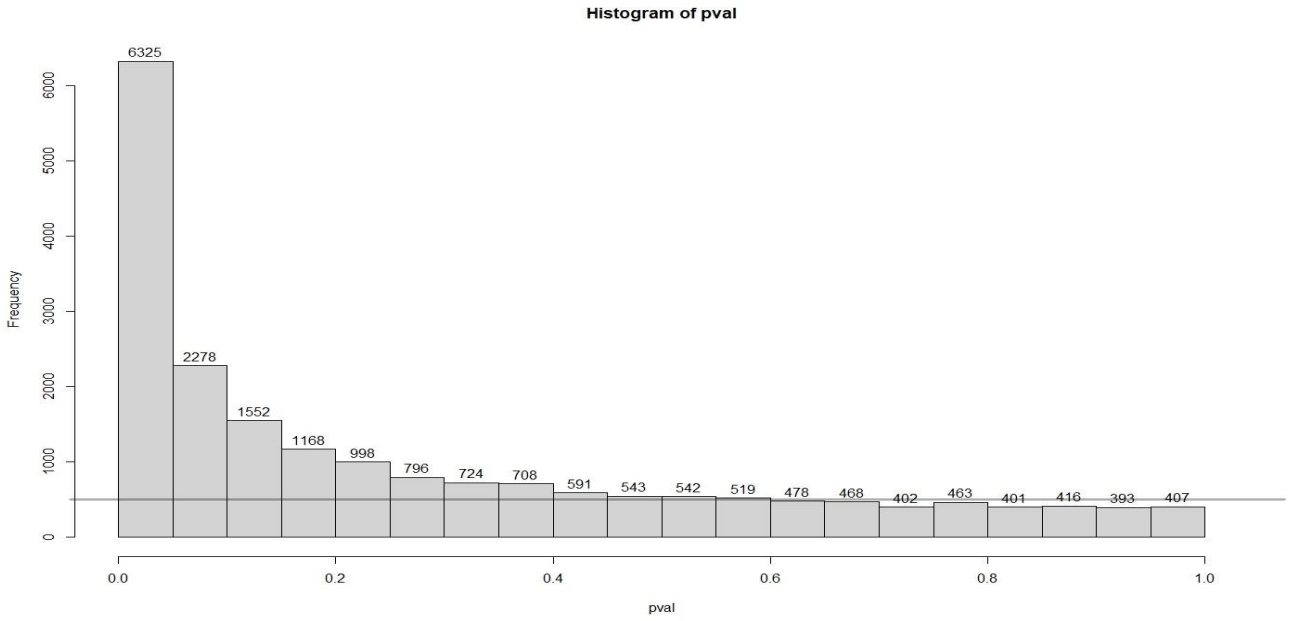
It is difficult to estimate π0 without specifying the distribution of the truly alternative p values but p-values of truly alternative features will tend to be close to zero p-values of null features will be uniformly distributed among [0, 1]. Most of the p-values we observe near 1 will be null then.

We can see our histogram of the 20170 p values from pur aml-nol data from 1.3.

We notice that the bars will go fairly flat at arounf pval of 0.4

this indicates that there are mostly null p values in this region

So a rough estimate of Po≈0.4=40%



1.5: **Report how many genes are differentially expressed when controlling theFWER, FDR and pFDR at α = 0.01.**

FWER

```
fwer1
FALSE  TRUE
20080    92
```

With the tag TRUE is the differentially expressed which are 92 and with the tag FALSE are the ones that are not 20080

FDR

```
fdr1
FALSE  TRUE
19219    953
```

With the tag TRUE is the differentially expressed which are 953 and with the tag FALSE are the ones that are not 19219

pFDR

```
q.val
FALSE  TRUE
18485  1687
```

With the tag TRUE is the differentially expressed which are 1687 and with the tag FALSE are the ones that are not 18485

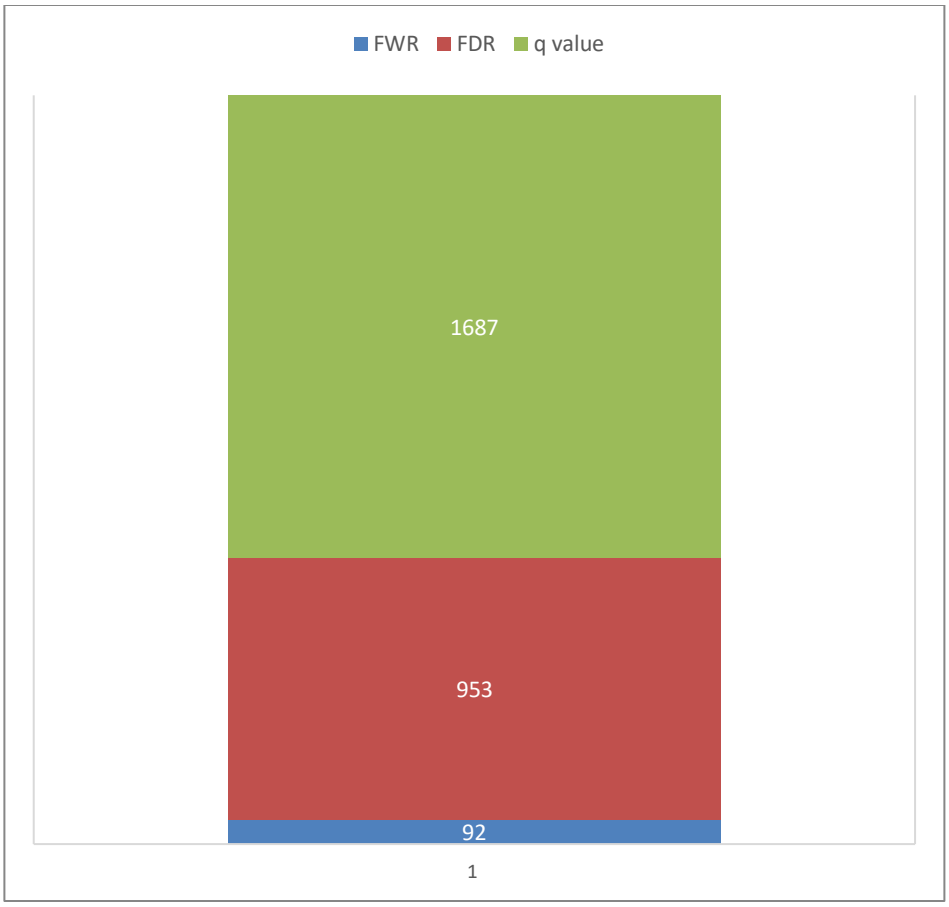
```
> table(fwer1,fdr1,pfdr)
, , pfdr = FALSE

       fdr1
fwer1   FALSE  TRUE
  FALSE 18485     0
  TRUE      0     0

, , pfdr = TRUE

       fdr1
fwer1   FALSE  TRUE
  FALSE   734   861
  TRUE      0    92
```

As we can see from our graph below



This mean that the genes that are differentially expressed with FWR are 92.We would choose this method even if the observations are small in quantity when we want the absolute best with no risk answer for our problem.An example is when

FDR are 953 this means that we can see more genes that deffreriaence baut with higher risk

Q-value and ginally with this method we have a total of 1687 genes that deffreciante but with the higher risk out of the other two

So lets say we want to produce a medicine with 100% certaintly that is safe and ahs no side effects we gonna use the FWR method

But if we want to see which medicines that meet a specific criteria compare to others(which medicine is better than other) use lower risk method because we want higher number of observations like FDR and qvalue methods

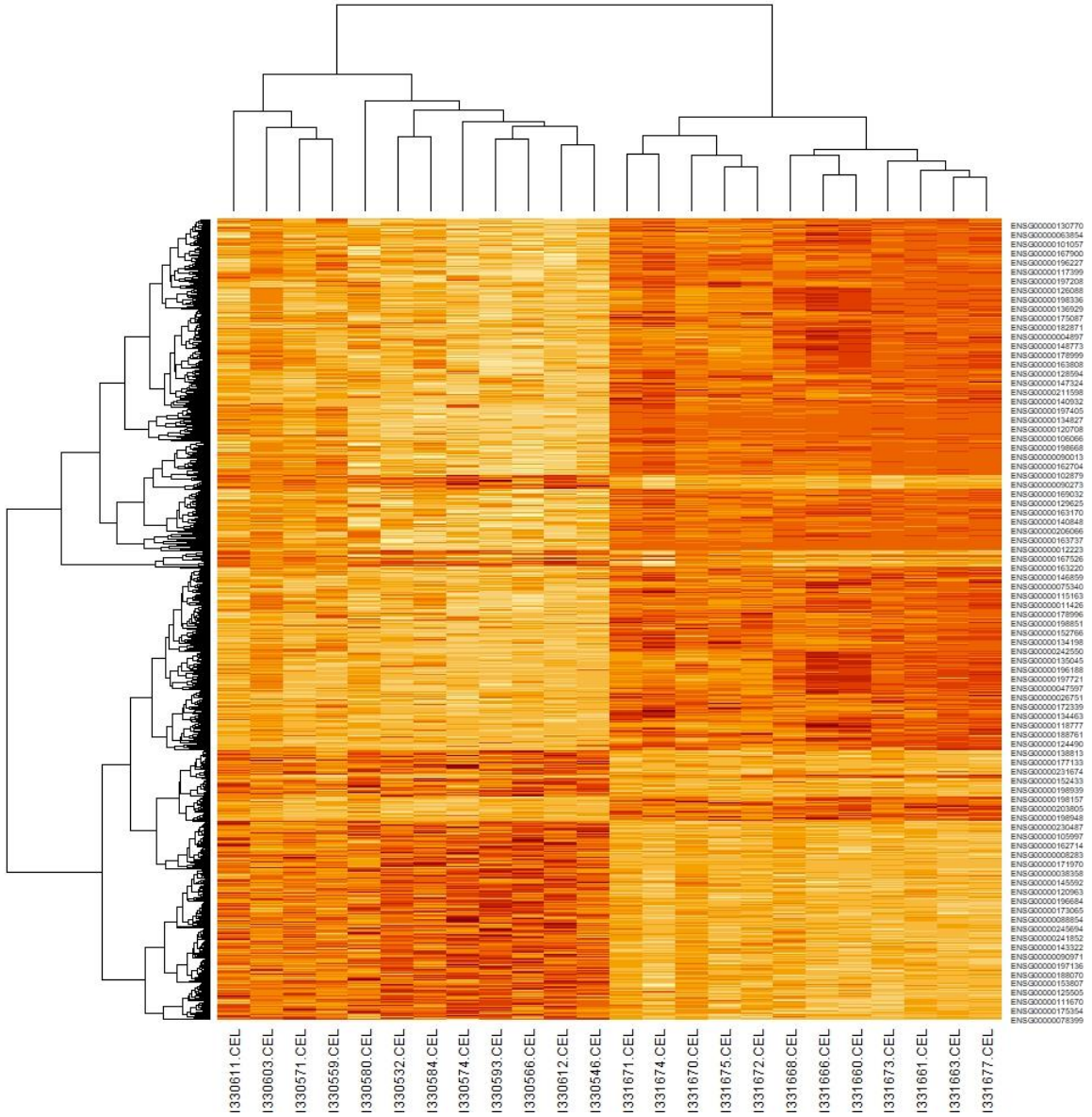
1.6.a: **Visualize the results obtained in question 5 according to whether the corresponding hypothesis is rejected or not when controlling the FDR at 0.01:**

(a) Plot a meaningful summary of the data and colour the genes depending on the result of the test (Differentially Expressed or not Differentially Expressed when controlling the FDR at the given level). Try to take into account both the mean difference as well the standard deviation per gene. Be creative.

(b) using Principal Components projections.

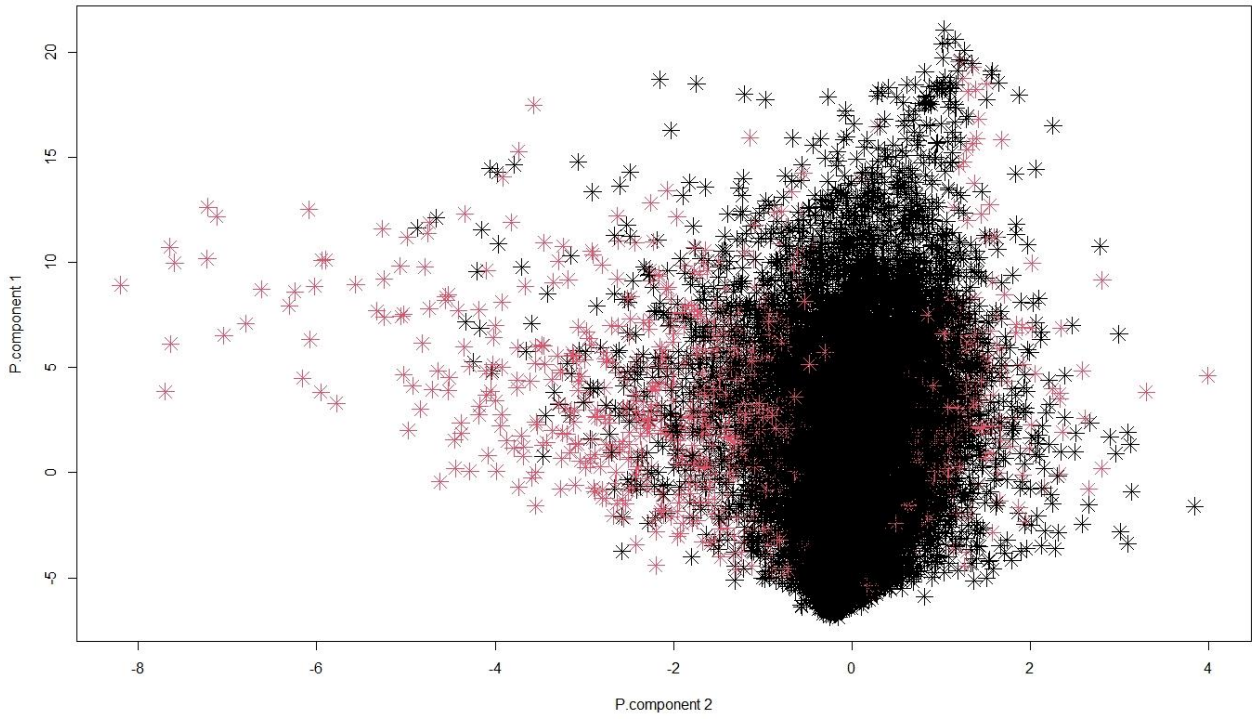
and explain your findings.

If we compare it to our 1st heatmap we can see a more distinctive look but no clear distriction still because only around our 5%(953) of genes were differentially expressed



1.6.b

Lets check our PCA now that we applied the FDR method



Aw before we can notice patterns of distinction but not in a satisfactory level

Exercise 3 (Social network). Consider a sample (minimum: 20 persons) from your personal social environment (e.g. colleagues, family friends, college friends, etc…). Construct the corresponding friendship network (that is: the adjacency matrix takes the value 1 whenever two persons are friends and 0 otherwise). Visualize, analyze and cluster the data using appropriate method(s) and discuss the results.

These are 20 friends from my personal environment from different timestamps of my life
From -> To is when the person A is friend with person B (The opposite may not be always true)

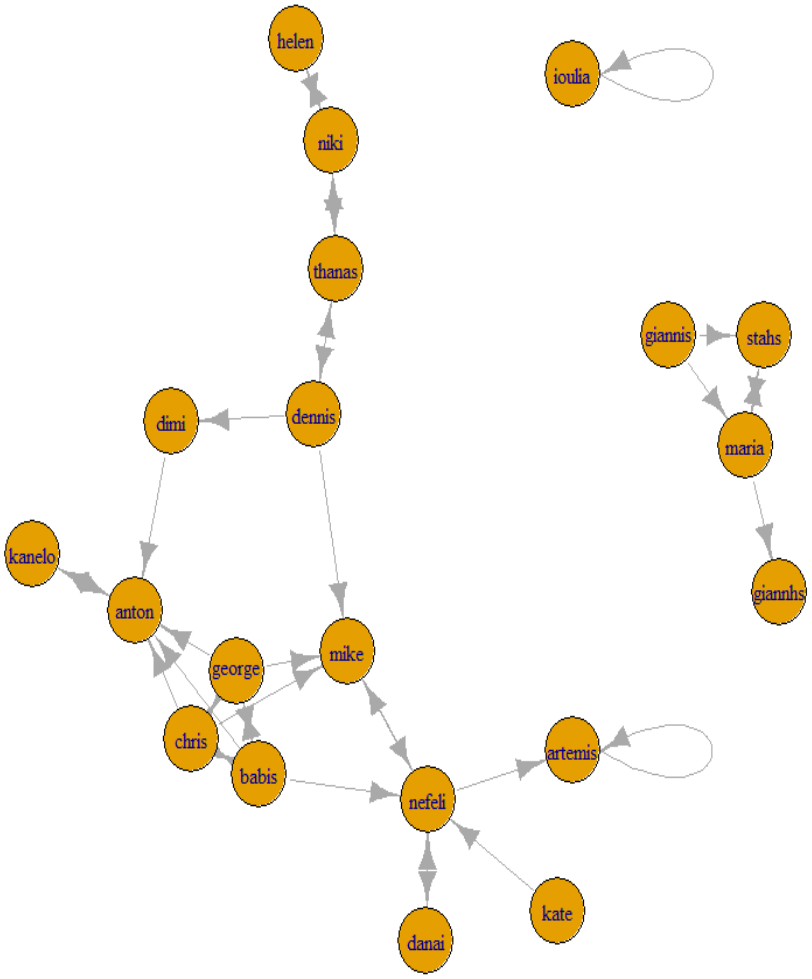
```
from <-
c("chris","chris","chris","chris","mike","thanas","dennis","dennis",
"niki","helen","babis","babis","babis","george","george","george","g
eorge","kanelo","anton","ioulia","dimi","giannis","giannis","stahs",
"maria","maria","nefeli","nefeli","danai","artemis","kate","babis","
dennis","nefeli","niki","thanas")
```

```
to <-
c("mike","babis","george","anton","nefeli","dennis","mike","thanas",
"helen","niki","chris","george","anton","chris","babis","anton","mik
e","anton","kanelo","ioulia","anton","maria","stahs","maria","stahs"
,"giannhs","danai","artemis","nefeli","artemis","nefeli","nefeli","d
imi","mike","thanas","niki")
```

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

Let's see the plot.

- As we can see when someone is friend with someone and the other is not friend with this person it is symbolized with only one arrow mark, otherwise with two(both think friend of each other).
- When someone does not have any friends there will be no arrow marks



This is the adj.matrix.

- The number **1 = friend** with the person
- The symbol **- = no friend** with this person

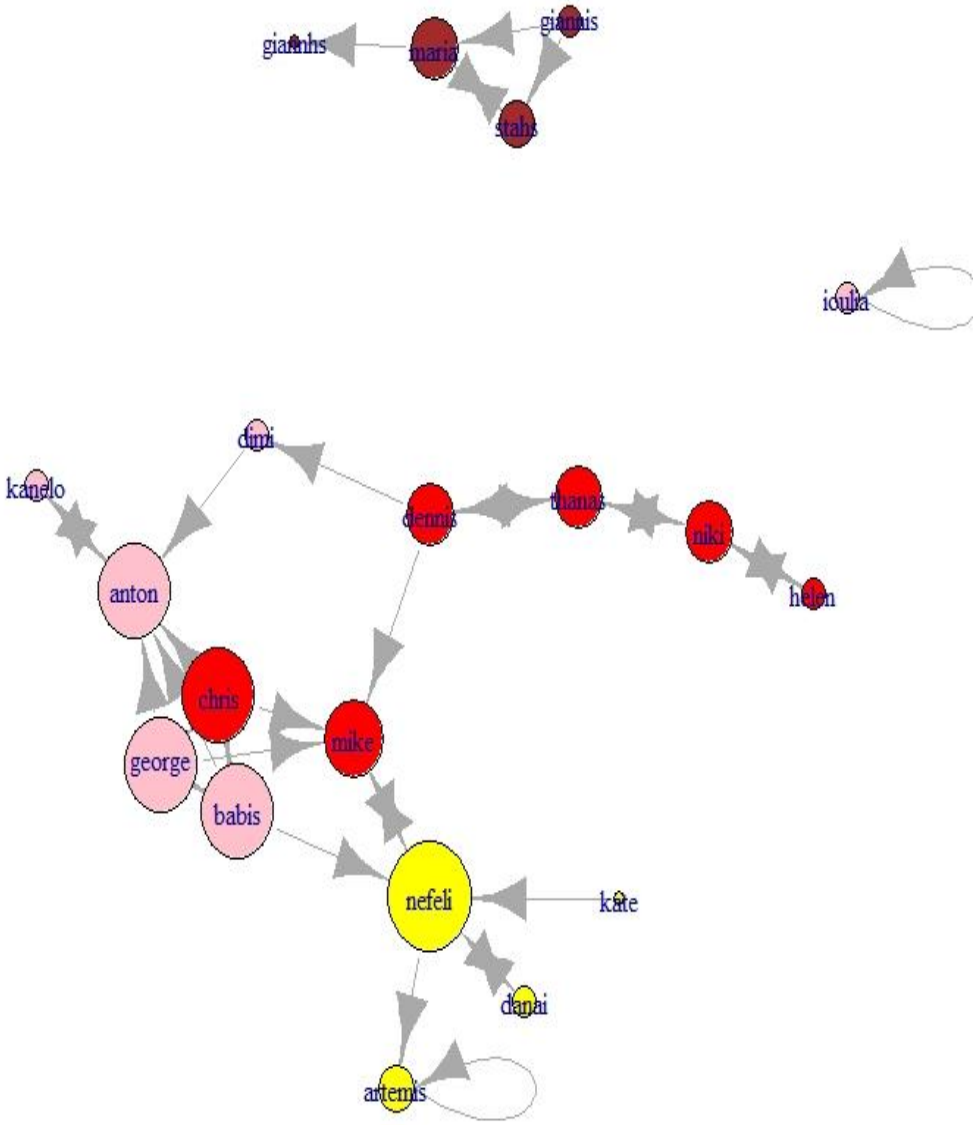
```
> adjmatrix
20 x 20 sparse Matrix of class "dgCMatrix"
   [[ suppressing 20 column names 'chris', 'mike', 'thanas' ... ]]

chris   . 1 . . . . 1 1 . 1 . . . . . . . . . .
mike    . . . . . . . . . . . . . . . 1 . . . .
thanas  . . . 1 1 . . . . . . . . . . . . . . .
dennis  . 1 1 . . . . . . . . 1 . . . . . . . .
niki    . . 1 . . 1 . . . . . . . . . . . . . .
helen   . . . . 1 . . . . . . . . . . . . . . .
babis   1 . . . . . . 1 . 1 . . . . . 1 . . . .
george  1 1 . . . . 1 . . 1 . . . . . . . . . .
kanelo  . . . . . . . . . 1 . . . . . . . . . .
anton   . . . . . . . . 1 . . . . . . . . . . .
ioulia  . . . . . . . . . . 1 . . . . . . . . .
dimi    . . . . . . . . . 1 . . . . . . . . . .
giannis . . . . . . . . . . . . . . 1 1 . . . .
stahs   . . . . . . . . . . . . . . . 1 . . . .
maria   . . . . . . . . . . . . . . 1 . . . . . 1
nefeli  . 1 . . . . . . . . . . . . . . . 1 1 . .
danai   . . . . . . . . . . . . . . . 1 . . . .
artemis . . . . . . . . . . . . . . . . . 1 . .
kate    . . . . . . . . . . . . . . . 1 . . . .
giannhs . . . . . . . . . . . . . . . . . . . .
```

```
ecount(mynetwork)
[1] 36.
```

These are the total connections(arrow marks)

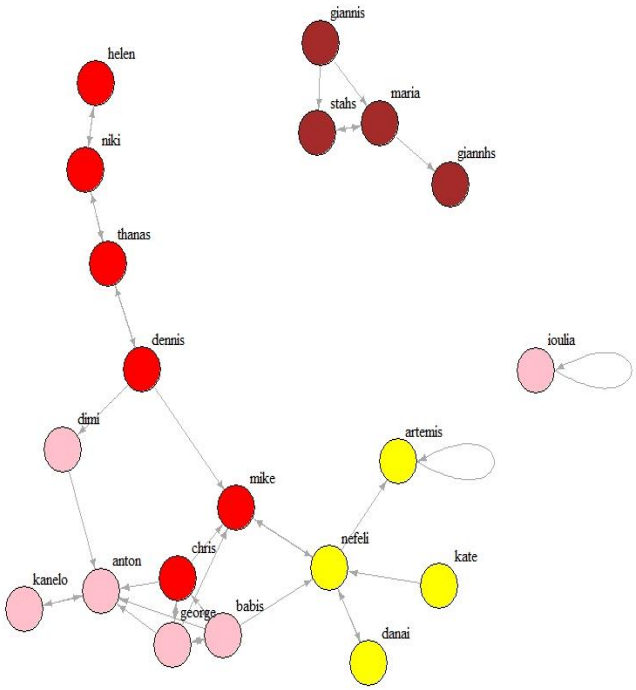
This graph shows the persons likeness(bigger cyrcle more connections=likeness)



Will the code come close to split the teams by the area that I met each one? Or at least come close

First we must assign colors for each person to symbolize the area of which I met them

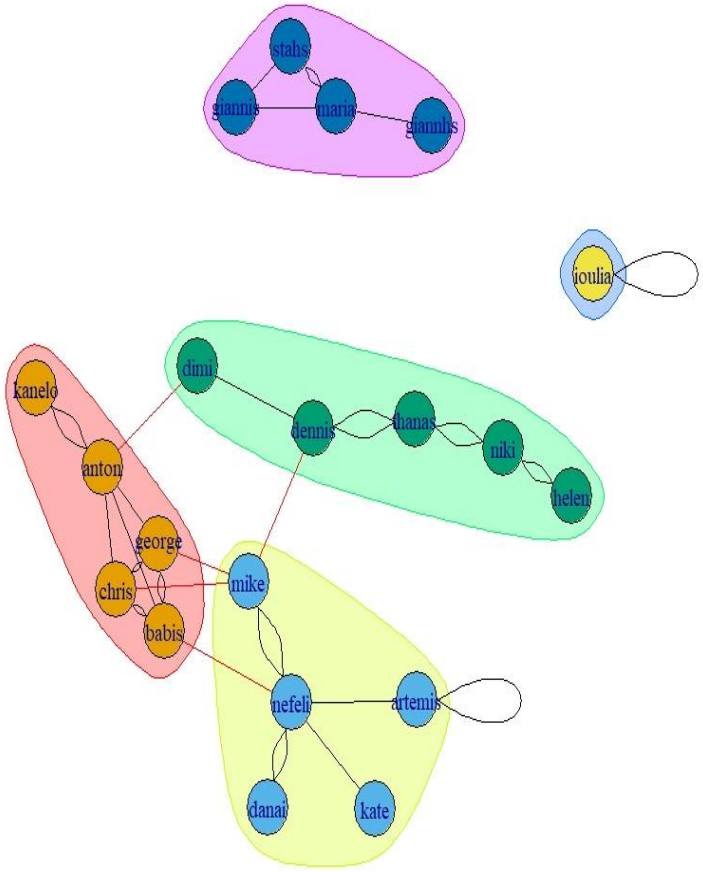
1. **School=red**
2. **Random=pink**
3. **University=Brown**
4. **Family cycle=Yellow**



And this is how likely to be connected to someone in this social network

```
edge_density(mynetwork)
[1] 0.09473684
```

Now that we now the actual real areas of origin of each person, let's see how well the code will identify them



Points to notice

1.Ioylia is in a unique lonesome team which is normal because even if she was in the **random named group**, without any friends is not possible to be assigned to a populated people group, instead it has her own(logical)

2. Mike was to be assigned to the school group but he assigned to the family group because as it seems he has more connection to the family group
*Mike actually is my best friend so it normal to have met my inner family cycle and be assigned to later one

3. Other than that the split was pretty accurate

Σας ευχαριστώ!

Exercise 2 (Big Data Regression: airlines dataset). The data from (datasethttp://stat-computing.org/dataexpo/2009/the-data.htmlThe) provides arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. We will create a model explaining the arrival delay from the year 1993.

- The month
- The weekday
- The distance
- The departure delay
- The departure time

Describe the model you estimated and write a short report explaining what you see.

Based of my AM we will choose the 1993 year

(appendix for the full code)
We will keep the rows with no missing values with this command

```
View(air1993.2)
air1993.3 <- air1993.2[complete.cases(air1993.2), ]
```

Before we create our model we will make factors the month and the day of the week

```
air1993.3$air1993.Month<-as.factor(air1993.3$air1993.Month)
air1993.3$air1993.DayOfWeek<-as.factor(air1993.3$air1993.DayOfWeek)
```

Next we will create our model with the glm command. We could also use the biglm.ffdf to create our model but the amount of data is not very large so glm won't create any problems

Here is our summary

```
> summary(model2.1)

Call:
glm(formula = air1993.3$air1993.ArrDelay ~ ., family = gaussian,
    data = air1993.3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1341.60     -6.69     -1.43      4.79    492.91

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)        -1.203e+00  3.456e-02  -34.819  < 2e-16 ***
air1993.Month2      2.105e-01  3.275e-02    6.428 1.29e-10 ***
air1993.Month3      9.154e-01  3.193e-02   28.669  < 2e-16 ***
air1993.Month4      3.415e-02  3.188e-02    1.071  0.28404
air1993.Month5     -7.919e-01  3.167e-02  -25.000  < 2e-16 ***
air1993.Month6      1.093e-01  3.181e-02    3.436  0.00059 ***
air1993.Month7     -1.032e+00  3.157e-02  -32.684  < 2e-16 ***
air1993.Month8     -4.120e-01  3.148e-02  -13.089  < 2e-16 ***
air1993.Month9     -4.262e-01  3.196e-02  -13.334  < 2e-16 ***
air1993.Month10     8.690e-02  3.170e-02    2.741  0.00612 **
air1993.Month11    -3.621e-01  3.216e-02  -11.259  < 2e-16 ***
air1993.Month12    -1.005e-01  3.180e-02   -3.159  0.00158 **
air1993.DayOfWeek2  2.968e-02  2.399e-02    1.237  0.21598
air1993.DayOfWeek3  6.680e-01  2.401e-02   27.825  < 2e-16 ***
air1993.DayOfWeek4  6.638e-01  2.406e-02   27.591  < 2e-16 ***
air1993.DayOfWeek5  5.667e-01  2.398e-02   23.633  < 2e-16 ***
air1993.DayOfWeek6 -1.330e+00  2.475e-02  -53.727  < 2e-16 ***
air1993.DayOfWeek7 -7.628e-01  2.440e-02  -31.259  < 2e-16 ***
air1993.Distance   -7.501e-05  1.245e-05   -6.027 1.67e-09 ***
air1993.DepTime     1.242e-03  1.398e-05   88.859  < 2e-16 ***
air1993.DepDelay    8.372e-01  3.112e-04 2690.153  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 210.2073)

    Null deviance: 2641218981  on 4993586  degrees of freedom
Residual deviance: 1049683833  on 4993566  degrees of freedom
AIC: 40877382

Number of Fisher Scoring iterations: 2
```

Coefficients explanation

- Intercept if we are on the first month in the 1st day of the week with distance , Depttime,Dep ,Delay=0 we will had delay equal to 1.2 mi. This means that the airplane will arrive earlier by 1.2 min.
- For the second month of the first day of the week and everything set to 0 we will had -1.2+2.1=+0.9,this means that the airplane will arrive by roughly 1 min later
- 1 unit increase in distance the delay will decrease by almost 5% of a minute
- For 1 unit increase in DepTime the odds for delay will increase by almost 6% of a minute
- For 1 unit increase in DepDelay the the delay will increase by 3 min

All our variables are statistical significance and good for the explanation of the arrival delay expect for the 4th month and the 2nd day of the week

What will BIC,AIC say?

According to our BIC criteria, the best model to explain the ArrDelay is the full model

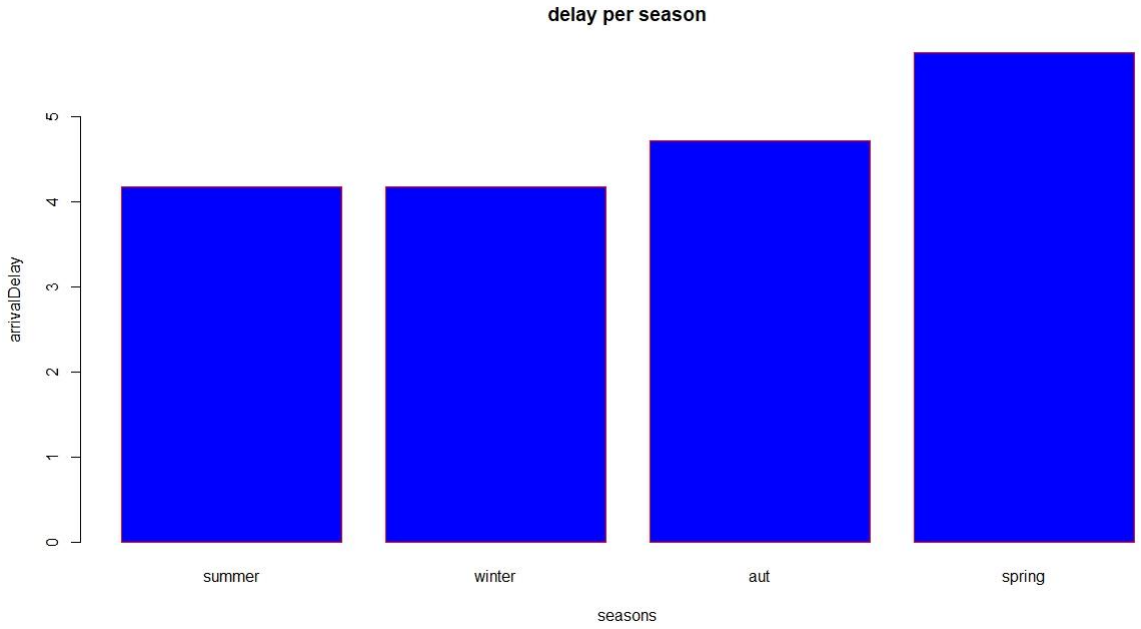
```
> n1<-step(model2.1, direction="both")
Start:  AIC=40877382
air1993.3$air1993.ArrDelay ~ air1993.Month + air1993.DayOfWeek +
    air1993.Distance + air1993.DepTime + air1993.DepDelay

                    Df   Deviance      AIC
<none>                  1049683833 40877382
- air1993.Distance   1 1049691470 40877416
- air1993.Month     11 1050864941 40882975
- air1993.DepTime    1 1051343601 40885269
- air1993.DayOfWeek  6 1052108145 40888889
- air1993.DepDelay   1 2570937712 45350544
```

In the months 1,2,3,4,6,10 there will be increase chance for delay. Also the higher the Dep time and The Dep Delay there is the higher the chance for ta arrival delay will be

On the positive note Distance seems to not affect at all the delay of the plane, which seems logical because the plane travel at a steady speed each travel will standard speed

Let’s compare our 4 season to see which one has the average higher delay. Is the cold weather(snow) of the winter play any role?



It seems that the highest delay for the year 1993 had the spring months