Final Project – Methods in Data Science

BY: Maryam Wahbi & Amal Khatib

Datasets:

We chose to work on the liver data, we worked with two datasets:

1) Gene expression data :

R object: Dataframe name

mat.f.coding.RData: mat.f.coding

2 rows – represent the genes

columns – represent the samples

2) Samples Characteristics :

R object: Dataframe name

pheno.f.RData: pheno.f

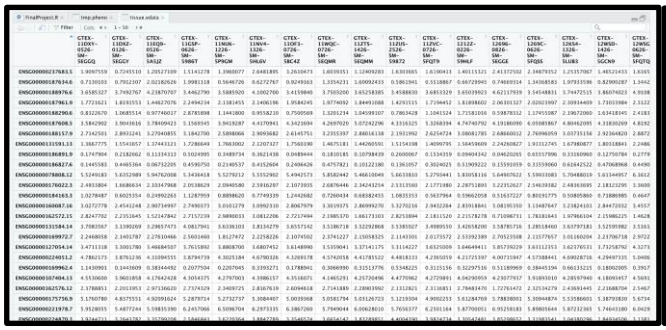


Fig 1 : Gene expression data

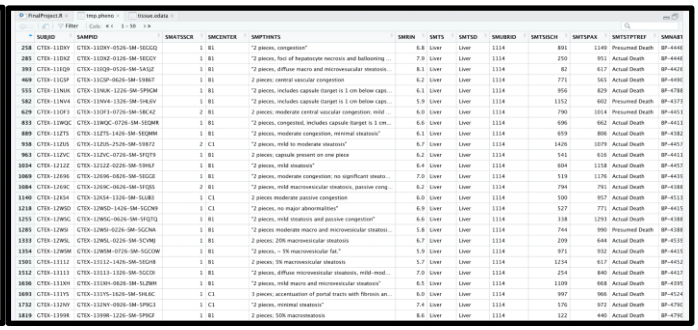


Fig 2: Samples Characteristics data

Ideas and designs:

- Research the data and the effects on it.
- Finding which features affect the level of gene expression so we can find which features need to be fixed or corrected in the data.
- Research relations between features.

Explaining the work:

- **Initial processing of the data**

We started by downloading and loading; mat.f.coding.RData and pheno.f.RData, we picked the data on the liverto research, which is a major organ only found in vertebrates which performs many essential biological functions.
We then wanted to filter out genes that had a low expression value or genes that had a low variance.
In the beginning we had 2,357,866 genes. After filtering there was a deletion of 792,540 genes so we had 1,565,326 remaining genes.
Finding and deleting outliers, we used two different methods to identify outliers; the variance method for analyzation and the hclust function to graphically examine the samples we divided by height = 91 into 3 clusters (we colored them; pink, green and blue) see Fig 3. and Fig 4 below.
We ended up deleting 2 outliers respectively ("GTEX-12ZZZ-1326-SM-59HKW", "GTEX-131XE-0326-SM-5LZVO").

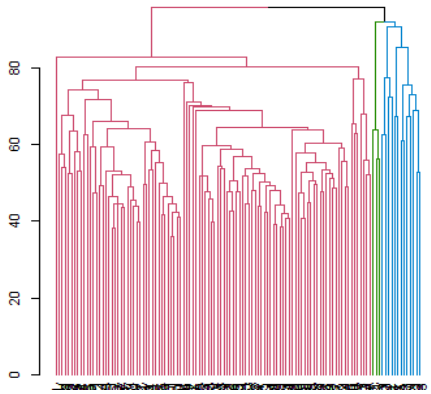


Fig 3: Sample clustering after first outlier removal

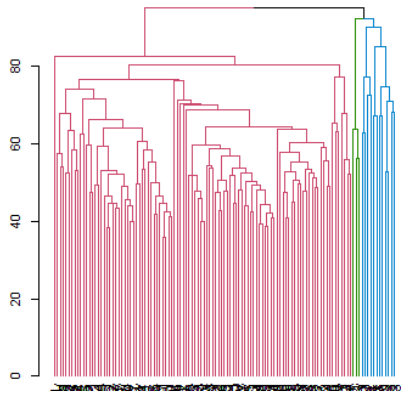


Fig 4: Sample clustering after second outlier removal

The methods we used to identify outliers:

1. The variance method for analyzation - It checked the distance between the samples to find the outlier (the samples that are the furthest away from the other samples).
2. hclust: Hierarchical Clustering -Clustering the data into k clusters and finds the majority of patterns in the data

set and organized the data accordingly .
We found that these samples were further from all of other outliers e.g. ”GTEX-12ZZZ-1326-SM-59HKW” had the smallest avg in gene expression.
There are three causes for outliers-We assume that there could have been a data entry error, experiment measurement errors, sampling problems or natural variation.
Outliers are problematic and should be removed because they represent measurement errors, data entry or processing errors, or poor sampling.

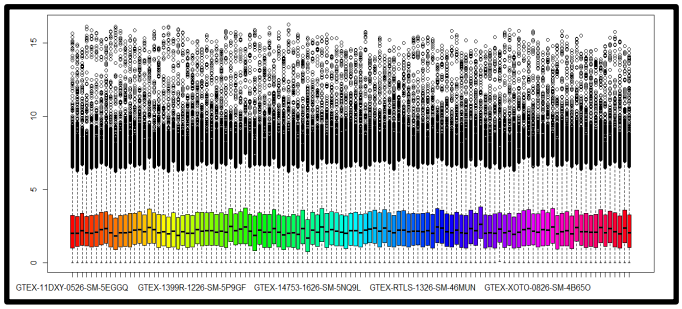


Fig 5: Observations boxplot before quantile normalization

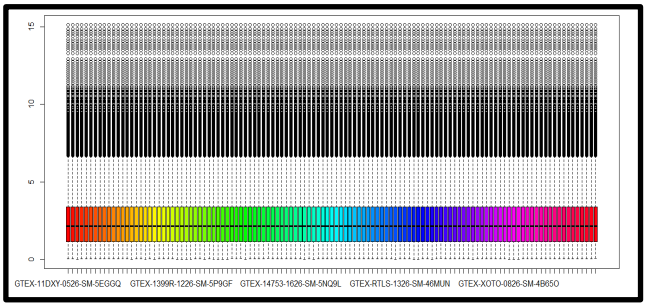


Fig 6: Observations boxplot after quantile normalization

Later on we performed quantile normalization on the data to improve its accuracy and integrity.

Now our data is ready for the next step.

- **Correction of biases - identification of factors (features) that add “noises” to the data**
-

Finding features that affect the level of gene expression :

1) PCA:
We used PCA to research the gene expression with the 5 features (SMRIN, SMTSISCH, SMGEBTCH, AGE, DTHHRDY).
Sample Characteristics:

- SMRIN - The RNA Integrity Number, a measure of the quality of the RNA.
- SMTSISCH - Interval in minutes between actual death and final tissue stabilization.
- SMGEBTCH - Expression Batch ID.
- AGE – the age of the individual in years.
- DTHHRDY - Death Circumstances. Death classification based on the 4-point Hardy Scale.

We used the built-in R functions prcomp() and fviz_pca_ind().We also used `factoextra` package.
Now we will present 5 PCA plots that we came up with and explain the results.

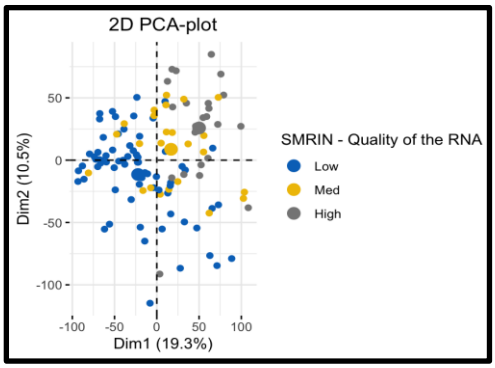


Fig 7: SMRIN PCA

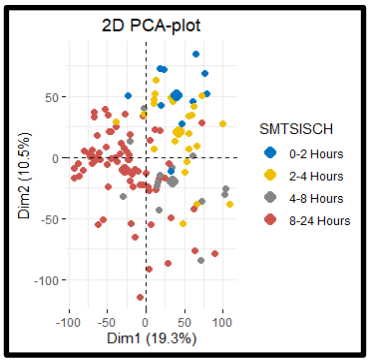


Fig 8: SMTSISCH PCA

Fig.8: We divided the data into 3 clusters according to SMRIN(Low, Med, High), we colored the samples (Blue,Yellow and Gray) respectively. In Fig.8 we can see a concentration of the samples of low quality, med quality

and high quality together so we conclude that SMRIN does affect the samples therefore this feature does need correction.

Fig.9: We divided the data into 4 clusters according to SMTSISCH(0-2, 2-4, 4-8, 8-24 hours), and we colored the samples (Blue, Yellow, Gray and Red) respectively. In Fig.9 we can see a concentration of the samples of 0-2, 2-4, 4-8, 8-24 hours together so we conclude that SMTSISCH does affect the samples therefore this feature does need correction.

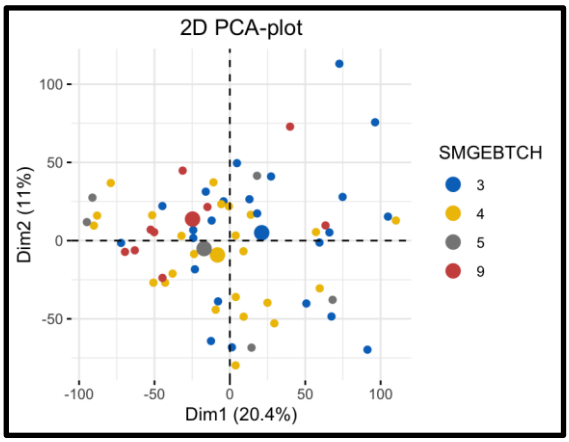


Fig 9: SMGEBTCH PCA

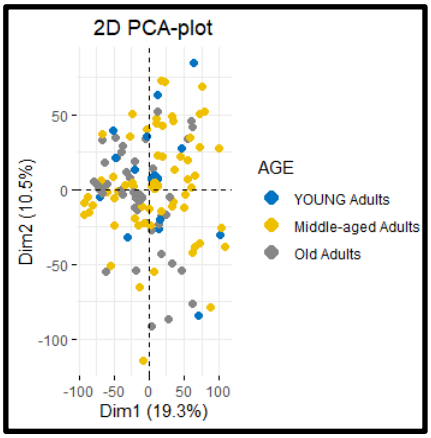


Fig 10: AGE PCA

Fig.9: We divided the data into 4 clusters according to SMGEBTCH how many people were tested in that experiment(3, 4, 5, 9), and we colored the samples (Blue, Yellow, Gray and Red) respectively. In Fig.10 we can not see a concentration of the samples of the frequency together so we conclude that SMGEBTCH does not affect the samples.

Fig.10: We divided the data into 3 clusters according to AGE(Young adults, Middle-aged adults and Old adults), we colored the samples (Blue, Yellow and Gray) respectively. In Fig.11 we can see a concentration of the samples of Old adults together(on the left side) but the samples of Middle-aged adults are not concentrated together; therefore we conclude that the AGE does affect the samples in a slight way.

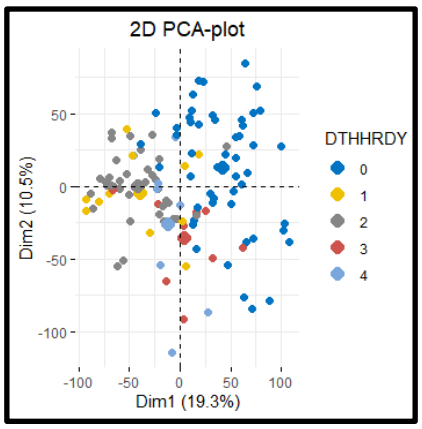
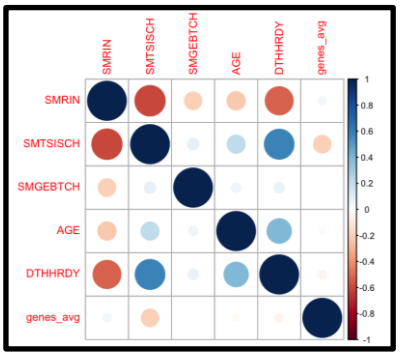


Fig 11: DTHHRDY PCA

Fig.11: We divided the data into 5 clusters according to DTHHRDY(0, 1, 2, 3, 4), and we colored the samples (Blue, Yellow, Gray, Red and Light Blue) respectively. In Fig.12 we can see a concentration of the samples 0, 1, 2, 3, 4 together so we conclude that DTHHRDY does affect the samples therefore this feature does need correction.

2) Correlation between features & avg of gene expression




```
> cor(genes_avg_df)
                SMRIN     SMTSISCH     SMGEBTCH          AGE      DTHHRDY   genes_avg
SMRIN       1.00000000 -0.60901707 -0.20437313 -0.22230382 -0.52036292  0.04361101
SMTSISCH   -0.60901707  1.00000000  0.08433414  0.21617901  0.58527599 -0.20378717
SMGEBTCH   -0.20437313  0.08433414  1.00000000  0.05818016  0.07267397  0.00691621
AGE        -0.22230382  0.21617901  0.05818016  1.00000000  0.39604742 -0.02161075
DTHHRDY    -0.52036292  0.58527599  0.07267397  0.39604742  1.00000000 -0.04787294
genes_avg   0.04361101 -0.20378717  0.00691621 -0.02161075 -0.04787294  1.00000000
```

Fig 12 & 13: Correlation between features & avg of gene expression

Fig 12 & 13: We used the built-in R functions cor() and corrplot(). We also used corrplot,and the ggcorrplot package.

First we added genes average, and we found inverse correlations(negative) and positive correlations.

The most prominent inverse correlations were found between SMRIN and SMTSISCH (-0.60901707) and between SMRIN and DTHHRDY (-0.52036292).

And the most prominent positive correlations were found between DTHHRDY and SMTSISCH(0.58527599) and between DTHHRDY and AGE(0.39604742).

There are more relations(obviously) but they are less prominent(smaller and closer to zero) e.g. between SMGEBTCH and genes_avg(0.00691621)

3) Machine learning

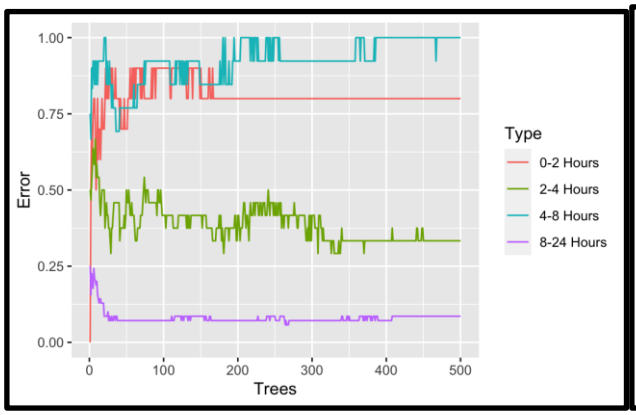


Fig 14: OOB with ntree = 500

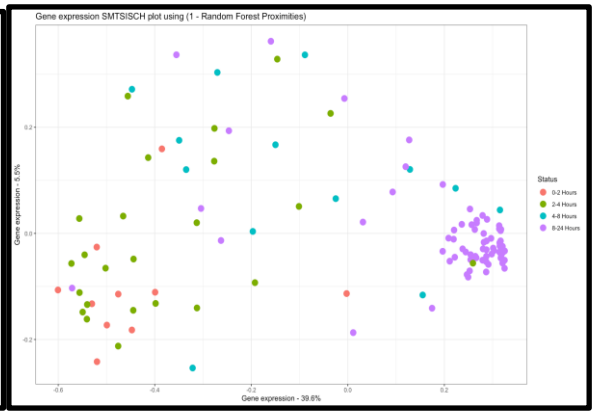


Fig 15: Gene expression SMTSISCH plot using Random Forest

We used the built-in R functions randomForest(), as.dist(), round(), cmdscale(), and ggplot().We also used e1071, cowplot, and randomForset packages.

Fig14: We used the Out-of-bag (OOB) error method which is used for measuring the prediction error of random forests.

First we ran the method with the default ntree value(ntree = 500), then we wanted to find how a different ntree value would affect the result so we ran the method with ntree = 1000, but we found that it didn't affect the result.

Also to try to control how many variables are considered at each step we ran the random forest with different values for mtry

The highest error rate was with 4-8 hours, and the lowest error rate with 8-24 hours.

Fig15: We started building a Gene Expression-plot to show how the samples are related to SMTSISCH.

We converted the proximity matrix into a distance matrix, then calculated the percentage of variation that each Gene Expression axis accounts for, making a fancy looking plot that shows the Gene Expression axes and the variation. It is clear to see in the plot that the samples were divided into groups(colored accordingly), we saw a concentration of the samples meaning that this feature does affect the samples therefore this feature does need correction.

- **Research relations between features**

We used the built-in R functions lm(),predict(),par(),plot(),lm(),ggplot.We also used ggpubr,ggplot,packages.
From the correlation matrix above we could see the relation between the different features, but we wanted to go deeply into the relations, we found some interesting relations that we wanted to see more of; (to see all of the other relations you find them in the github link).

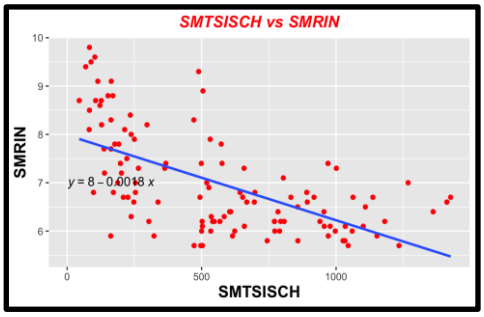


Fig 16: SMTSISCH VS SMRIN linear regression

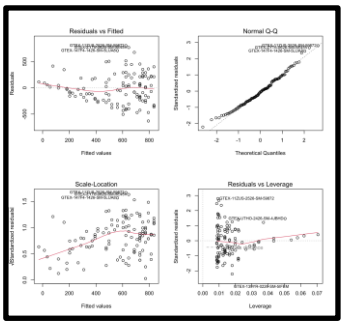


Fig 17: residual plots

Fig 16: We started by using the lm() function to fit a simple linear regression model, with SMRIN as the response and SMTSISCHt as the predictor.
We observed an inverse relationship between SMTSISCH and SMRIN,so we can predict the value of a SMTSISCH based on the value of SMRIN.
Fig 17: Next we examine some diagnostic plots. Four diagnostic plots are automatically produced by applying the plot() function directly to the output from lm()

- In the Normal Q-Qplot in the top right, we can see that the residuals follow close to a straight line on this plot, it is a good indication they are normally distributed.
- In the Residuals vs Fitted plot in the top left, we can see that linearity seems to hold, as the red line is close to the dashed line. We notice that points 'GTEX-11ZUS-2526-SM-59872', 'GTEX-ZYT6-0626-SM-5E45V',, and 'GTES-147F4-1426-SM-5LUA8' they far from the data may be outliers, as they has large residual values.

- In the Residuals vs Leverage plot in the bottom right , we can see that observation 'GTEX-11ZUS-2526-SM-59872' in the top left right corner falls outside of the red dashed lines. This indicates that it is an influential point.This means that if we removed this observation from our dataset and fit the regression model again, the coefficients of the model would change significantly.
- In the Scale-Location plot in the bottom left , we can see that it is not a horizontal line with equally spread points, it has trend, so the error terms do not have a constant variance.

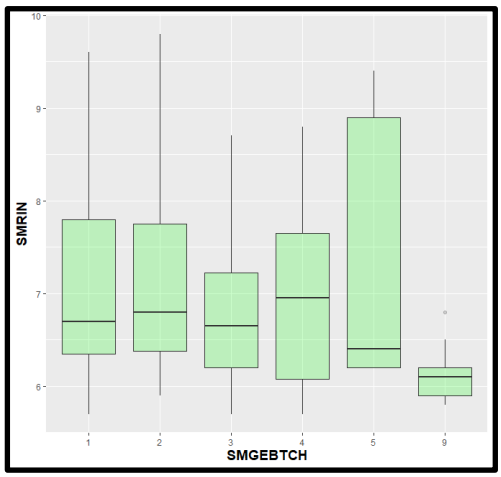


Fig 18: Boxplot of SMGEBTCH VS SMRIN

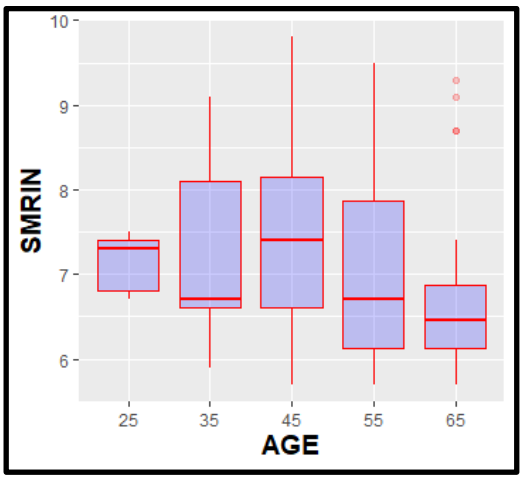


Fig 19: Boxplot of AGE VS SMRIN

Fig18: We see an obvious sharp drop in SMRIN after SMGEBTCH 4, so in a way an increased number of people in the same experiment (more than 4) affect SMRIN.
Fig19: We see an obvious sharp drop in SMRIN after the AGE of 45, so in a way old age affects SMRIN.

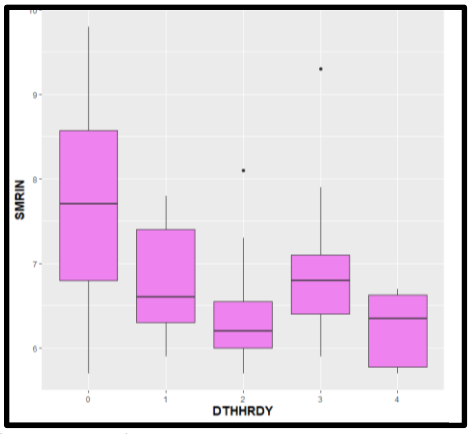


Fig 20: Boxplot DTHHEDY VS SMRIN

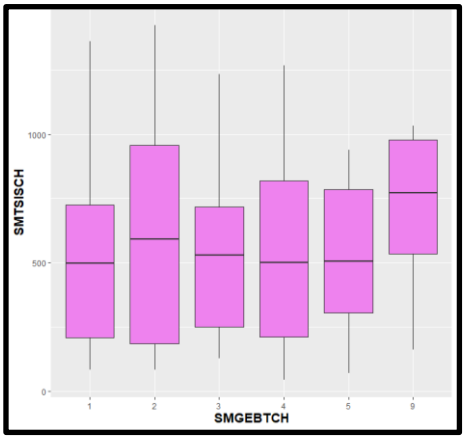


Fig 21: Boxplot of SMGEBTCH VS SMTSISCH

Fig20: We see how the highest SMRIN was recorded with DTHHRDY 0(could be because these patients were ventilated -they were at the hospital at the time of death and maybe the parents had already given the permission for organs donation), and the lowest SMRIN was recorded with DTHHRDY 2(could be because these patients died of natural causes -took some time to get them to a hospital resulting in DNA disintegration),so in a way DTHHEDY affects SMRIN.

Fig21: We see that SMGETCH Frequency does not affect SMTSISCH,this is because the avg is almost equal.

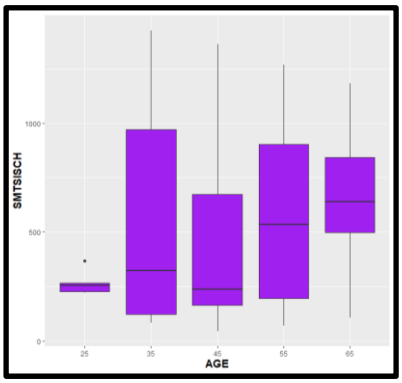


Fig 22: Boxplot of AGE VS SMTSISCH

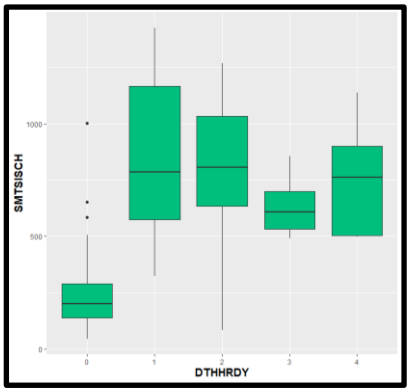


Fig 23: Boxplot of DTHHRDY VS SMTSISCH

Fig 22: An obvious and low SMTSISCH for young patients , and We can see a positive relationship between age and SMTSISCH.

Fig23: An obvious and low SMTSISCH for patients with the classification 0, because they were ventilated and already at the hospital. But an interesting high for patients with the classification 4 who had a long illnes could be due to them being at home with their loved ones because of their terminal phase.

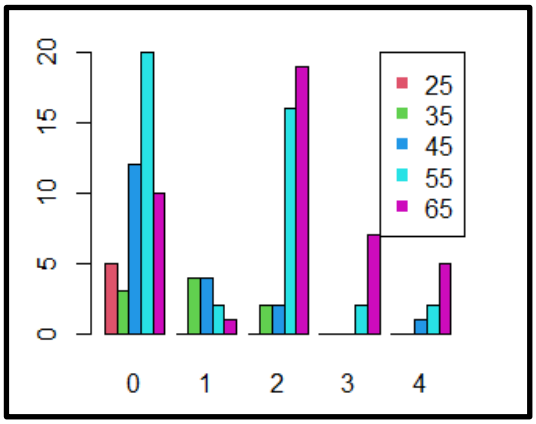


Fig 24: DTHHRDY VS AGE

Fig 24: We see that the first classification includes all generations, but we note that with the increase in the classification number, it includes the categories of people of old age.

Solution and results:

- We found that apart from SMGEBTCH all of the other features affect the integrity of the RNA and the gene expression.
- The feature that affected the integrity of the RNA and the gene expression the most is SMTSISCH.
- We found a correlation between most of the features:
  The most prominent inverse correlations were found between SMRIN and SMTSISCH (-0.60901707) and between SMRIN and DTHHRDY (-0.52036292).
  And the most prominent positive correlations were found between DTHHRDY and SMTSISCH(0.58527599) and between DTHHRDY and AGE(0.39604742).
- Through using the RandomForest and clustering by SMTSISCH, we proved how SMTSISCH does in fact affect the integrity of the RNA and the gene expression therefore, it does need correction.