**Hugh Hoang 1833106**

**0.** Create the *Processed Marriage* dataset or use the one created by the TA!

1. Compute the covariance matrix for each pair of the following attributes: Age Gap (1), Economic Similarity (3), Common Interests (7) and Divorce Score; next, compute the correlations for each of the 10 pairs of the 5 attributes. Interpret the statistical findings! **3 points**

To compute the covariance matrix, I read the database and got the relevant columns and ran them all through both a covariance and correlation function to get the following graphs

> cov(covariance)

Age.Gap Economic.Similarity Common.Interests Divorce.Score

Age.Gap 6.5870985 -4.8473702 -8.6726772 0.1402266

Economic.Similarity -4.8473702 743.0897882 48.2123964 -0.3524262

Common.Interests -8.6726772 48.2123964 198.1098969 -0.3610909

Divorce.Score 0.1402266 -0.3524262 -0.3610909 0.3169377

> cor(covariance)

Age.Gap Economic.Similarity Common.Interests Divorce.Score

Age.Gap 1.00000000 -0.06928488 -0.24007836 0.09705019

Economic.Similarity -0.06928488 1.00000000 0.12565651 -0.02296469

Common.Interests -0.24007836 0.12565651 1.00000000 -0.04556973

Divorce.Score 0.09705019 -0.02296469 -0.04556973 1.00000000

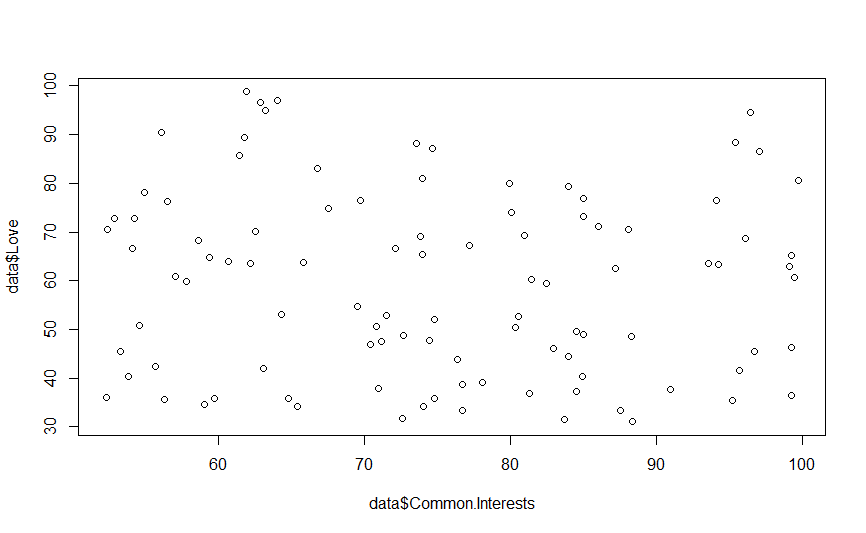
Since the covariance variable indicates the relationship of two variables whenever one variable changes, it means that for example if there is an increase in one variable results then there will most likely be an increase in the other variable then we can say that both variables are said to have a positive covariance.

So the strong positive pairs that we have excluding pairs of oneself is: economic similarity/common interests meaning that there is a positive relationship between the two . So the strong negative pairs that we have excluding pairs of oneself is: economic similarity/age gap and common interest/age gap meaning that there is a negative relationship between the two. The neutral pairs: divorce score/age gap, divorce score/economic similarity,divorce score/common interest means that there is not much of a relationship between the variables where changes in one wont really affect the other. But the pairs for divorce score are very low because we are comparing small numbers(divorce score) with big numbers(other variables) so while increases in something like age gap wont increase divorce score it by a lot, it shows a positive relationship which is similar with divorce pairs with economic similarity and common interest except its negative.

Now we move onto correlation to see how strong these positive and negative relationships are which basically see how related the points are to each other.

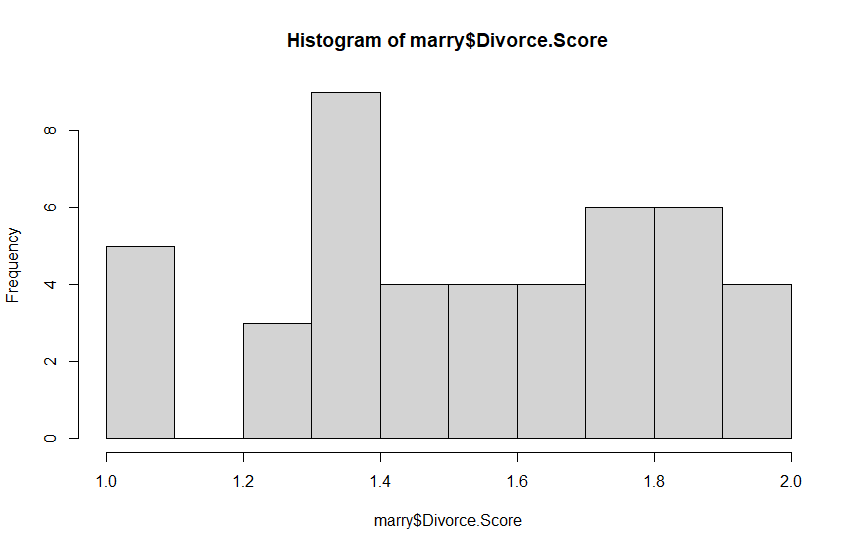
So the pairs closest to one that we have excluding pairs of oneself is: divorce score/age gap meaning that the points are closely correlated and negative. So the closest pairs negative one that we have excluding pairs of oneself is: common interest/age gap meaning that there is a negative relationship between the two. The pairs closest to 0: economic similarity/age gap, divorce score/economic similarity,divorce score/common interest means that the points are mostly scattered and there is not much of a relationship between the variables.

**2.** Create a scatter plot for the attributes Common Interests and Love. Interpret the scatter plot**! 3 points**

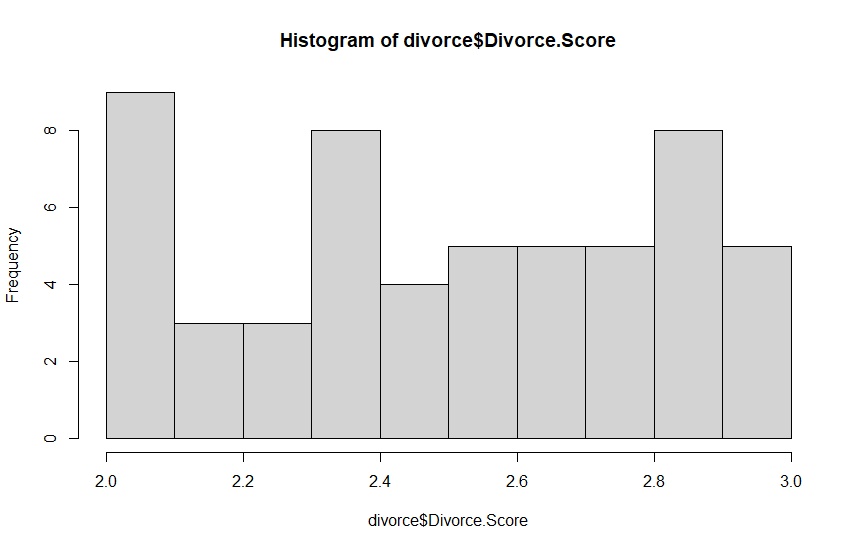
****

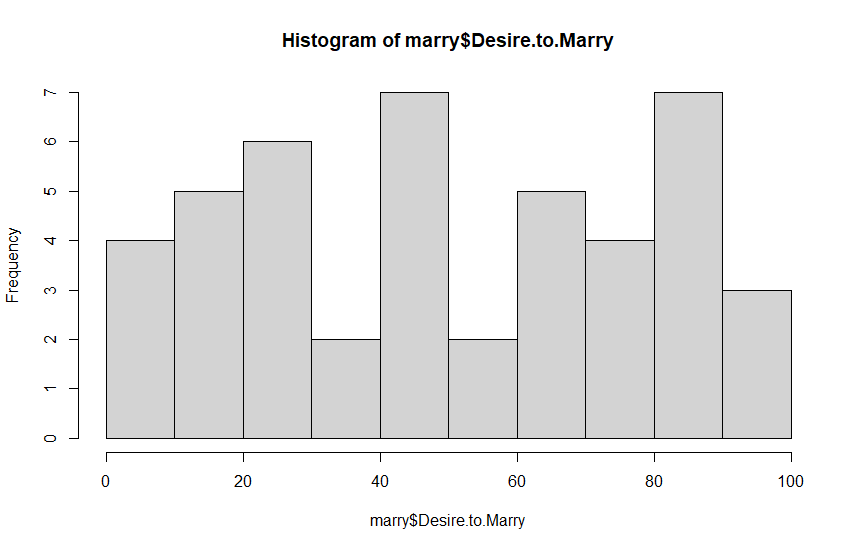
Since the scatterplot looks very scattered with the points having little to no relation with each other and no apparent trend then I would say that the covariance and the correlation are close to 0. You can see this as common interests increase the other variable love doesn’t seem to have a net increase or decrease. The points also don't seem to have much relation to each other or have an overall trend as it is hard to place a line of regression except as a horizontal line. This would indicate that common interests and love have a very weak relationship and would not be a good way to help find out one variable with another.

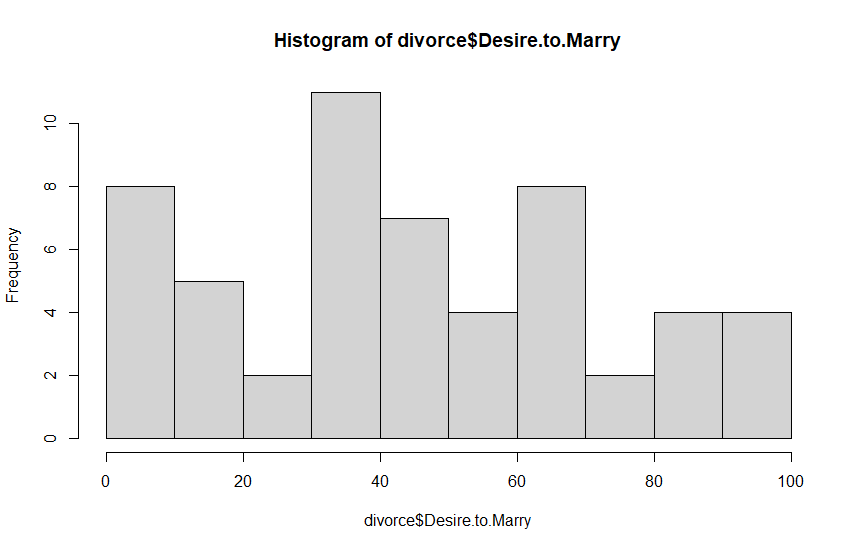
**3.** Create histograms for Divorce, Desire to Marry, and Common Interests attributes for both the Marry and the Divorce recommendations; interpret the obtained 6 histograms. **6 points**

****

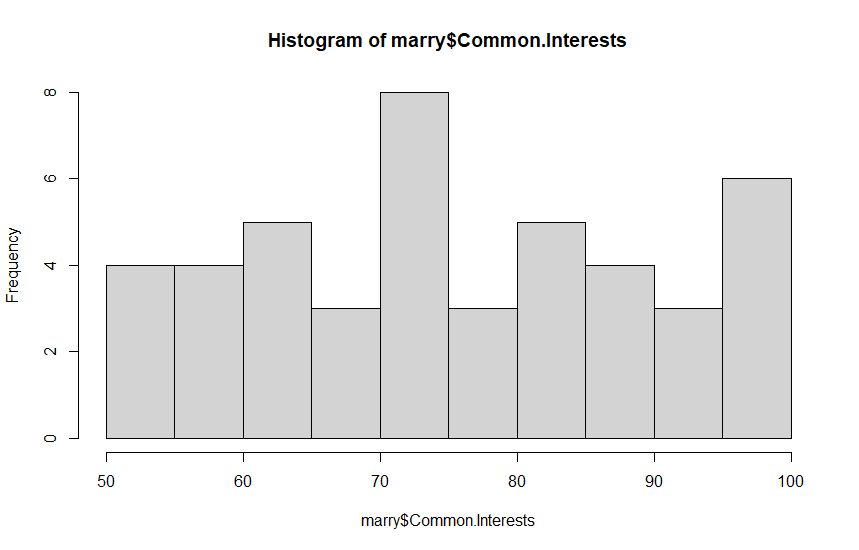
Divorce Score should have an equal frequency for numbers 1-2 for there to be a strong correlation to show that divorce score can accurately predict and recommend if they will get married. I believe that there is a pretty equal amount of frequencies excluding some small variations and outliers so I can say that there is a somewhat strong correlation and divorce score can predict marriage.

****Divorce Score should have an equal frequency for numbers 2-3 for there to be a strong correlation to show that divorce score can accurately predict and recommend if they will get divorced. Similar to marriage, I believe that there is a pretty equal amount of frequencies excluding some small variations and outliers so I can say that there is a somewhat strong correlation and divorce score can predict marriage.

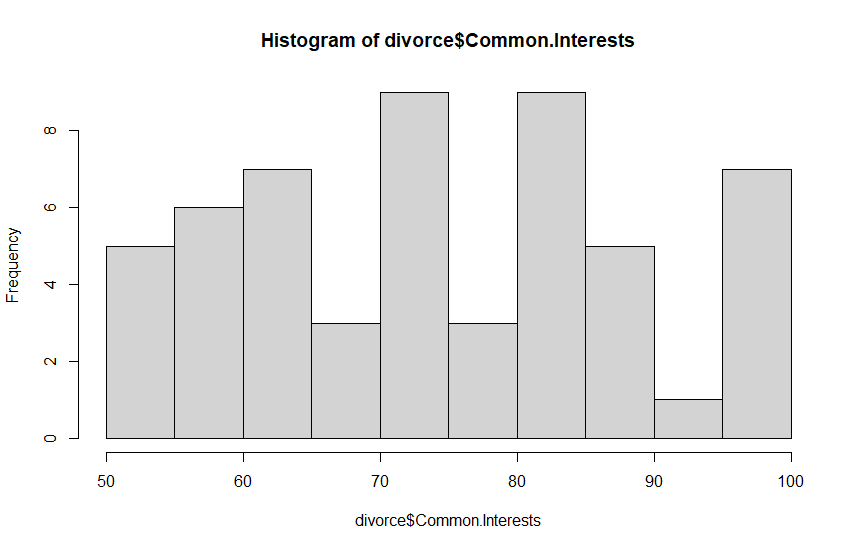
****desire to marry should have an upward or downward trend for there to be a strong correlation to show that changes in the desire to marry directly affects the frequency of marriage so that we can accurately predict and recommend if they will get married. I believe that there is not a stable enough trend so I can say that there is not a strong correlation to use desire to Marry to predict marriage.

****

Desire to marry should have an upward or downward trend for there to be a strong correlation to show that desire to marry can accurately predict and recommend if they will get divorced. I believe that there is not a stable enough trend so I can say that there is not a strong correlation to use desire to Marry to predict divorce.

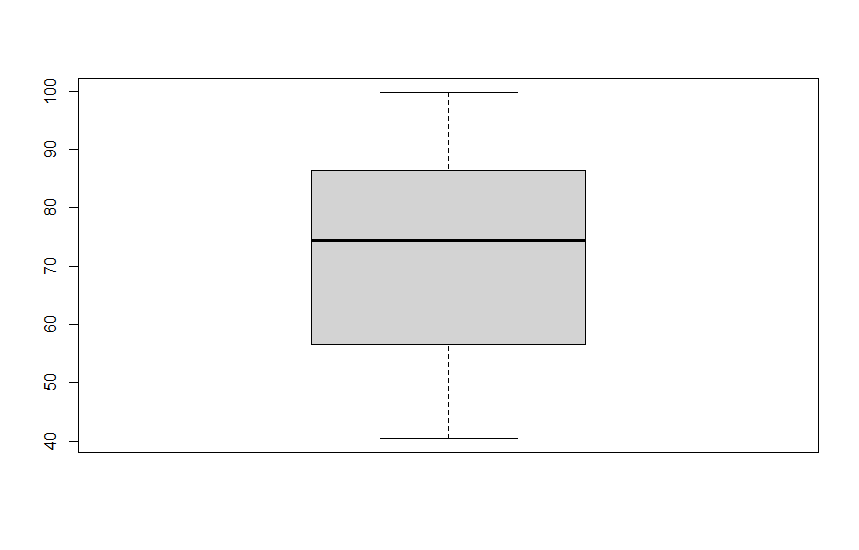
****

Common Interest should have an upward or downward trend for there to be a strong correlation to show that common interests can accurately predict and recommend if they will get married. I believe that there is not a stable enough trend so I can say that there is not a strong correlation to use common interests to predict marriage.

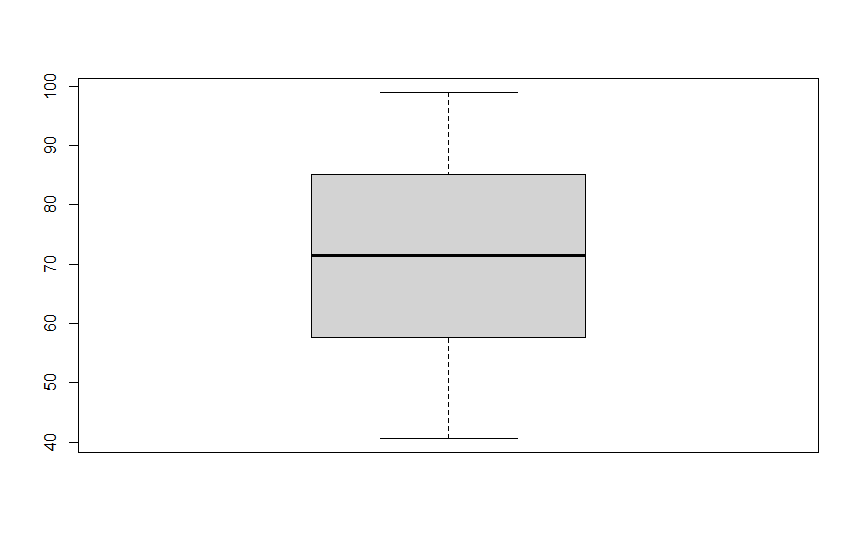
****

Common Interest should have an upward or downward trend for there to be a strong correlation to show that common interests can accurately predict and recommend if they will get divorced. I believe that there is not a stable enough trend so I can say that there is not a strong correlation to use common interests to predict divorce.

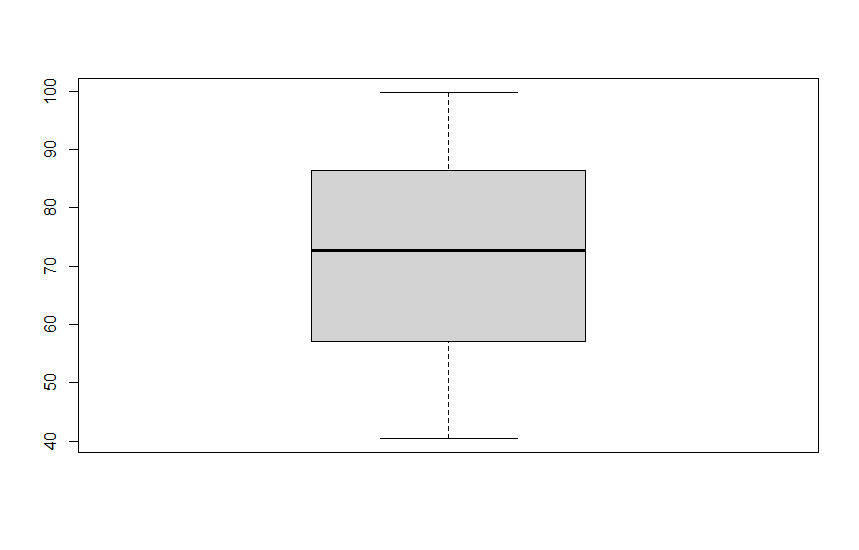
4. Create box plots for the Self Confidence attribute for the instances of each age class—one for M and D — and a third box plot for all instances in the dataset. Interpret and compare the 3 box plots for each attribute! **4 points**

****

For the self confidence attribute with the marriage boxplot, we can tell that the median is close to the center with the first quartile box being slightly larger but overall equal with the third quartile. The box plot is for the most part skewed normally with a slight positive skew which means that the data is close to symmetric with a leaning to smaller numbers. This means that as self confidence increases, the chances of marriage increases then peaks at around the middle then starts to decrease.

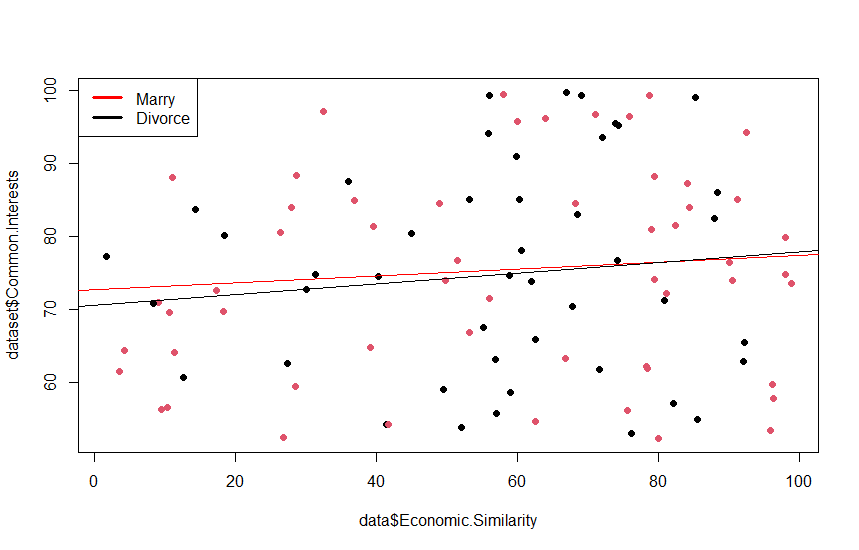
****

For the self confidence attribute with the divorce boxplot, we can tell that the median is close to the center with an overall equal with the third quartile. The box plot is basically skewed normally which means that the data is symmetric. This means that as self confidence increases, the chances of marriage increases then peaks at the middle then starts to decrease.

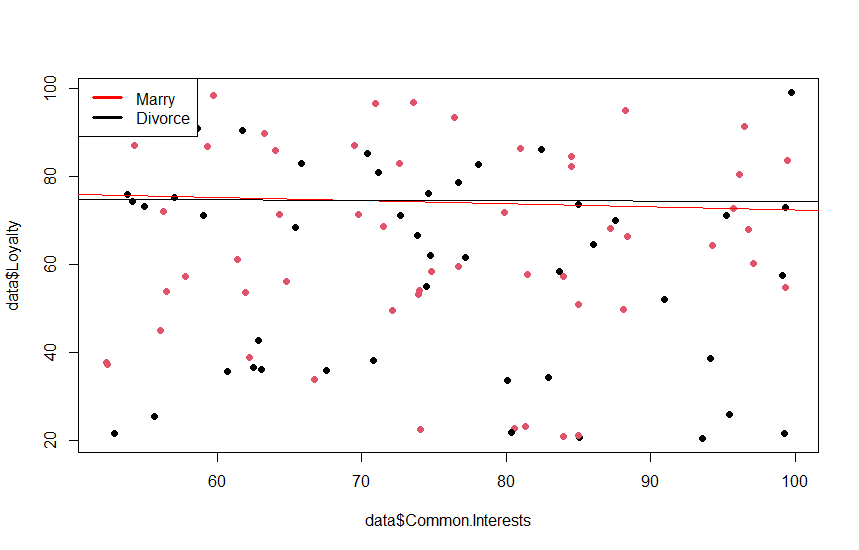
****

For the self confidence attribute with both the marriage and divorce boxplot, we can tell that the median is close to the center. The box plot is basically skewed normally which means that the data is symmetric. This means that as self confidence increases, the chances of marriage increases then peaks at around the middle then starts to decrease. We can see a very slight positive skew as the skew from the marriage skew affects the normal skew from the divorce skew. In general this should not mean much except that marriage has more outliers, but we can tell by the normal skew that there is not much of a positive or negative trend. Because there is no positive or negative skew that tends to show that self confidence is not a good way to predict marriage or divorce.

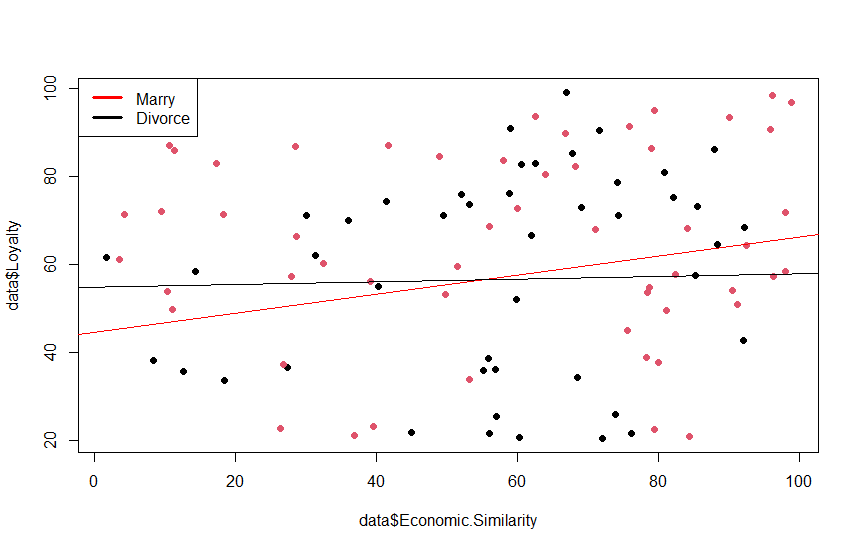
**5.** Create supervised scatter plots/supervised density plots for the following 3 pairs ofattributes using the Class attribute as a class variable: Economic Similarity & Common Interests, Common Interests & Loyalty and Economic Similarity & Loyalty. Use different colors for the class variable. Interpret the obtained plots; in particular, address what can be said about the difficulty in predicting the Recommendation and the distribution of the instances of the two classes. **6 points**

****

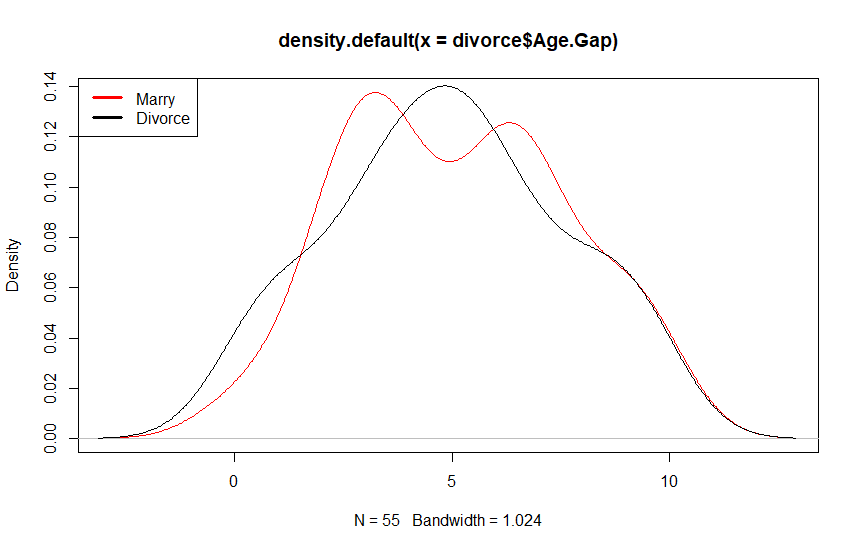
Since the scatterplot looks very scattered with the points having little to no relation with each other and no apparent trend then I would say that the covariance is slightly positive but close to 0 and the correlation is close to 0. You can see this as common interests increase the other variable love has an almost negligible net increase. The points also don't seem to have much relation to each other or have an overall trend as the line of regression is almost a horizontal line. This would indicate that common interests and love have a very weak and slightly positive trending relationship which means that these variables would not be a good way to help find out one variable with another. It also means that divorce would be slightly easier to predict than marriage using these variables.

****

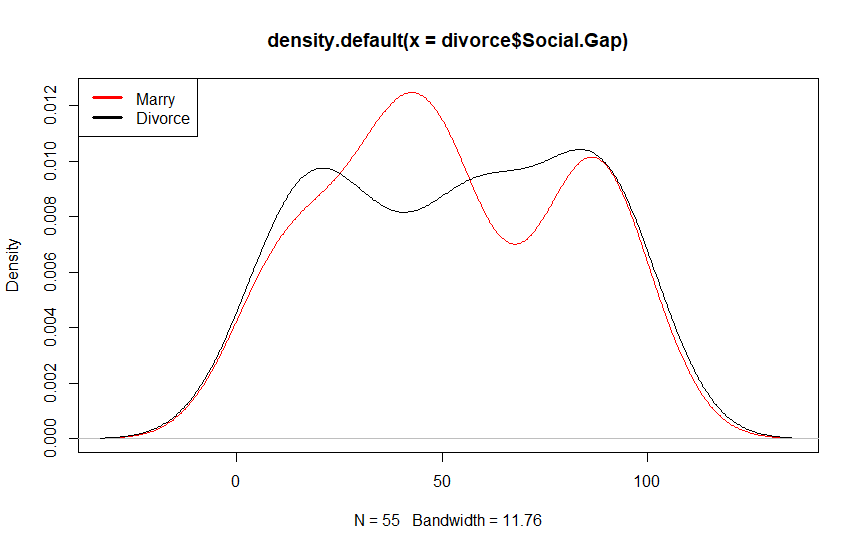
Since the scatterplot looks very scattered with the points having little to no relation with each other and no apparent trend then I would say that the covariance and the correlation is close to 0. You can see this as common interests increase the other variable love has an almost negligible net increase. The points also don't seem to have much relation to each other or have an overall trend as the line of regression is almost a horizontal line. This would indicate that common interests and Loyalty have a very weak relationship which means that these variables would not be a good way to help find out one variable with another.

****Since the scatterplot looks very scattered with the points having little to no relation with each other and no apparent trend then I would say that for divorce the covariance and the correlation is close to 0. But with marriage, there is a noticeable positive trend so we can see that covariance is further from 0, but correlation is still relatively low. You can see this as common interests increase the other variable love has a slight net increase with chances for outliers. The points also don't seem to have much relation to each other, but marriage has an overall trend as the line of regression is not a horizontal line. This would indicate that for divorce, economic interests and Loyalty have a very weak and no trending relationship but for marriage there is a somewhat weak and slightly positive trending relationship which means that these variables could be used to find out one variable with another. It also means that marriage would be slightly easier to predict than divorce using these variables.

**6.** Create 2 density plots for the instances of the 2 classes in the Age Gap/Social Gap space. Compare the 2 density plots! **6 points**

****

A density curve is a curve on a graph that represents the distribution of values in a dataset which we can tell by the general shape of the curve. We can see from this graph that both marriage and divorce follow a general bell shape with marriage having two peaks instead of one which just means that its thinner on the sides and flatter on the top. This bell shape means that we have a normal distribution of values with points keeping to the center which lessens as it goes further from the center. This goes to show that there isn’t an overall trend and this relationship would not be that effective in predicting either marriage or divorce.

****A density curve is a curve on a graph that represents the distribution of values in a dataset which we can tell by the general shape of the curve. We can see from this graph that both marriage and divorce still follow a general bell shape both now this time it plateaus near the center and still keeping the attribute of thinning out near the sides. Just like the first one, it's still symmetrical on both sides when coming from the center for both marriage and divorce. This bell shape means that we have a normal distribution of values with points keeping to the center which lessens as it goes further from the center. Neither graphs are skewed to either left or right which would indicate an overall trend but since it doesn’t then that means this relationship would not be that effective in predicting either marriage or divorce.

**7.** Create a new dataset *Z*-*Processed Marriage* from the *Processed Marriage* dataset by transforming the first 30 continuous attributes into z-scores. Fit a linear model that predicts the Divorce Score attribute using the 30 z-scored, continuous attributes as the independent variables. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting a successful marriage? **8 points**

Residuals:

Min 1Q Median 3Q Max

-0.56735 -0.15213 -0.01432 0.17763 0.60950

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.550000 0.026504 20.752 <2e-16 \*\*\*

s$Age.Gap -0.002989 0.030135 -0.099 0.9213

s$Social.Similarities 0.033486 0.032295 1.037 0.3034

s$Economic.Similarity 0.020756 0.032338 0.642 0.5231

s$Social.Gap -0.006283 0.031740 -0.198 0.8437

s$Common.Interests -0.003155 0.032735 -0.096 0.9235

s$Religion.Compatibility 0.045854 0.033322 1.376 0.1732

s$No.of.Children.from.Previous.Marriage 0.015344 0.030950 0.496 0.6216

s$Desire.to.Marry 0.004849 0.030297 0.160 0.8733

s$Relationship.with.the.Spouse.Family -0.063422 0.032159 -1.972 0.0526 .

s$Trading.in 0.031976 0.032749 0.976 0.3323

s$Engagement.Time 0.077190 0.031203 2.474 0.0158 \*

s$Love -0.016262 0.029704 -0.547 0.5858

s$Commitment -0.013330 0.031581 -0.422 0.6743

s$Mental.Health 0.007565 0.030292 0.250 0.8035

s$The.Sense.of.Having.Children -0.052488 0.031568 -1.663 0.1009

s$Previous.Trading -0.018044 0.031142 -0.579 0.5642

s$Previous.Marriage 0.013764 0.031152 0.442 0.6600

s$The.Proportion.of.Common.Genes 0.004014 0.031419 0.128 0.8987

s$Addiction -0.002328 0.031409 -0.074 0.9411

s$Loyalty 0.045885 0.033426 1.373 0.1743

s$Height.Ratio 0.027325 0.030401 0.899 0.3719

s$Good.Income 0.003310 0.030872 0.107 0.9149

s$Self.Confidence -0.010854 0.033800 -0.321 0.7491

s$Relation.with.Non.spouse.Before.Marriage -0.003717 0.031832 -0.117 0.9074

s$Spouse.Confirmed.by.Family -0.036224 0.030575 -1.185 0.2402

s$Divorce.in.the.Family.of.Grade.1 -0.009243 0.033787 -0.274 0.7852

s$Start.Socializing.with.the.Opposite.Sex.Age 0.023038 0.032124 0.717 0.4757

s$Education -0.024360 0.031255 -0.779 0.4384

s$Independency -0.022785 0.032004 -0.712 0.4789

s$Divorce.Score 0.417549 0.031751 13.151 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.265 on 69 degrees of freedom

Multiple R-squared: 0.8042, Adjusted R-squared: 0.719

F-statistic: 9.445 on 30 and 69 DF, p-value: 1.109e-14

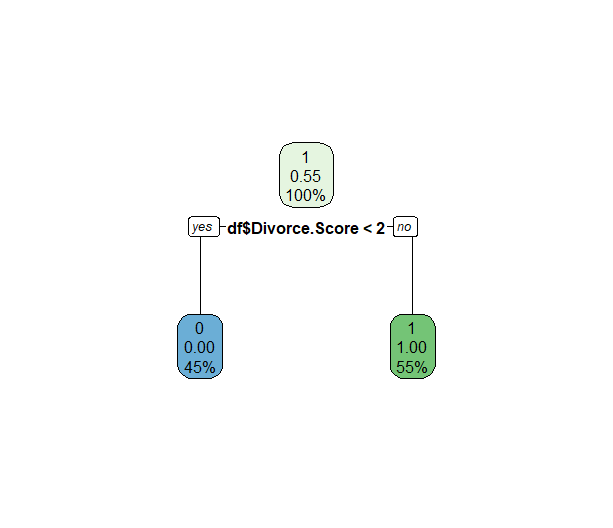
> summary(mod)$r.squared

[1] 0.8041693

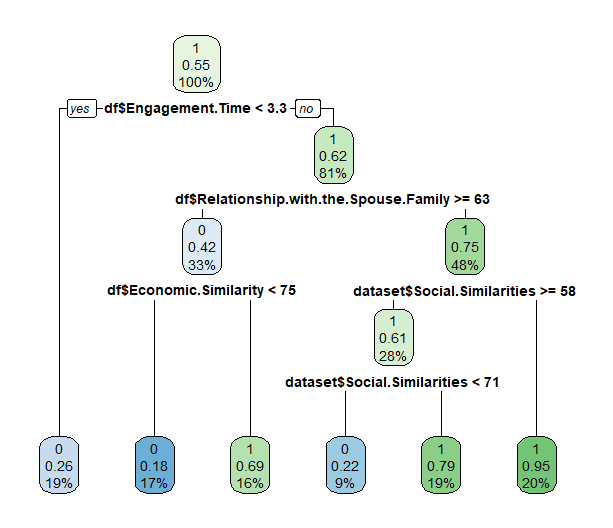
In order to get the coefficients and r^2, we had to first get the proper subset for the dataset, scale it in order to get the z-scores for all the attributes, run a linear model for divorce score and the other attributes, run the coefficient, summary and r^2 functions.

After getting the linear model we are now able to see the correlation coefficient between divorce score and all other attributes to accurately see both the strength of the correlation and which direction it goes with both positive and negative. The importance of each attribute for predicting a successful marriage is tied to the net value of the number which basically represents how much the mean of the divorce score changes as an attribute increases. We can see exactly which attributes are strongly related as well as other helpful information such as standard error to see how much it can vary, t-value which measures the size of the difference relative to the variation in your sample data .The obtained r^2 can be used to show if the linear model fits with values being closer to 1 indicating a good fit. Since our value is relatively high we can say that the linear model is a good fit.

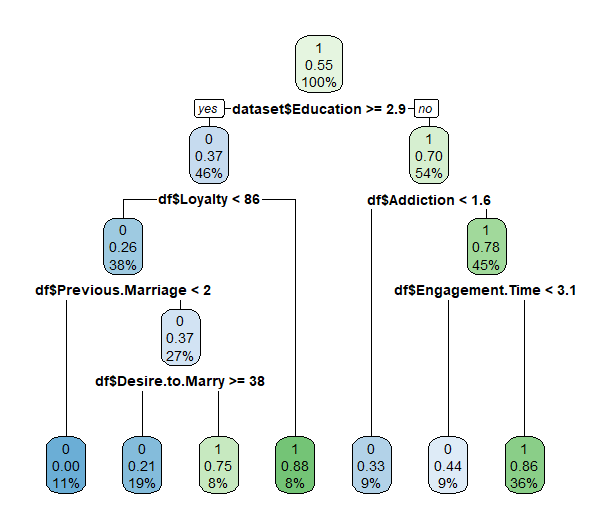
8. Create 3 decision tree models with 20 or less nodes for the dataset (leaf nodes count; do not submit models with more than 20 nodes!); use the Recommendation attribute as the class variable, and use 28 of the continuous attributes of the dataset, excluding the Seco`nd (Education) and Eleventh (Independency) attribute when building the decision tree model. Explain how the 3 decision tree models were obtained! Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the 28 chosen attributes for the classification problem? **9 points**

****

For this decision tree I used all 31 attributes excluding only recommendation because theres no useful data in a decision tree if you compare itself. From this data we can see that divorce score is by far the best way to predict marriage when compared to all the other attributes where here divorce is labeled with 0 being marriage and 1 being divorce. We see here that since there are no other leaf nodes and divorce score being at the head node and that divorce score can accurately predict divorce or marriage by itself.

****

For this decision tree I used 28 attributes excluding education,independency because of what was asked in the prompt recommendation because theres no useful data in a decision tree if you compare itself and divorce score because from the previous graph we were able to see that it would take over the graph. From here we can see that there is now more of a tree than the first one with engagement time being the number one deciding factor then we can predict with relationship with spouse’s family best for divorce, then economic similarity for deciding marriage and social similarities for deciding divorce. Using this decision tree we can visualize which would be the best variables to help predict marriage or divorce and which ones are better than others for deciding each recommendation

****

For this decision tree I used 30 attributes reincluding education and independency, but still excluding recommendation because there's no useful data in a decision tree if you compare itself and divorce score because from the first graph we were able to see that it would take over the graph. Now with the addition of the two attributes education and independency, we can see a change in the number one deciding factor and the resulting nodes in the tree. Now education is the number one deciding factor with addiction for divorce and with loyalty for marriage. These split off into smaller nodes, but it shows that education now being the head node has changed the decision tree by quite a bit and we can visualize this change. We now can conclude that aside from divorce score, education and its subsequent variables in the tree are the best variables to help predict marriage or divorce.

Since there is a limit set of 20 nodes on a decision tree, it will not give me 100% accuracy on the training data set because we had to limit some of the data which means that this affect the accuracy when predicting samples that are not part of the training set.

9. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular address the findings obtained related to predicting a successful marriage (the values of attributes 31 and 32) using attributes 1-30. If possible, write about which attributes seem useful for predicting successful marriages and what you as an individual can learn from this dataset! **6 points (and up to 4 extra points)**

Going in order, problem 1 uses covariance and correlation to try to find both the overall trend/pattern and how strong the relationship between two attributes are and helped get numbers for that which ended up showing that none of the pairs had a significant correlation and can’t be used as such. For problem 2, it uses scatterplot to try to visualize a pattern or trend between the relationship of all the points, this helped to see the points but any trends were either nonexistent or hard to tell which let me conclude there probably was not a significant relationship. For problem 3, it uses histograms to help show the frequency of certain relationships which from the first two we can see that divorce score has a relatively flat curve as it increases which is good as it shows that the numbers in the range 1-2 for divorce and 2-3 don't change as it increases showing a stable relationship which is different than the other 4 graphs as it isn’t put on the range like divorce score and still tend to be more erratic and not show a positive or negative trend meaning it isn’t as good for predicting. For problem 4, we use boxplots to try to visualize where data is grouped and see any skews as well as the ability to see things like first quartile, mean, third,quartile and its relation with the whiskers which did not end up showing much findings as there was not really much of a skew in either directions and the mean was mostly centered with equal whisker size. For problem 5, we used supervised scatterplots which use a line of regression, this helps us visualize the trend more clearly than in problem 2 but mostly showed a relatively flat line of regression with slight upwards or downwards trend meaning that theres not much of a correlation for all of them except maybe for the last graph which showed a slight upwards trend for marriage. For problem 6, we used density curves to visualize the distribution of values in this dataset which we can tell by the general shape of the curve but the graphs ended up showing a bell curve and a flatter bell curve for the second one which means that we have a normal distribution of values so not much of a correlation meaning it wouldn’t be useful in predicting. For problem 7, we used a linear regressionFit from z-scored attributes to find both the coefficients and the r^2 score, which enables us to see the correlation between divorce score and all other attributes. This information is actually very useful as it both gives a positive or negative correlation and a value which basically signifies how much the mean of the divorce score changes as an attribute increase which from this we can see the strongest correlation which is a negative coefficient for education with 0.0220347, this means that education is our best variable to use for predicting divorce score. From problem 8, we use decision tree models to see what attributes have the highest chance of predicting a Recommendation with various attributes for the dataset which I believe was the most useful data. The results of the graphs were able to show us that divorce score indeed was the best variable to help predict recommendations by itself, but the other graphs also showed us that there were other ways to predict without the divorce score with the alternative being the use of education and its other nodes to help give a more accurate recommendation based on all the attributes for a certain marriage to help give a more informed prediction.