Task1: Exploratory Data Analysis for “Marriage and Divorce Dataset”

Han Hoang

Fall 2022

0. Create the Processed Marriage dataset

```r
Marriage <- read.csv("C:/Users/katie/Downloads/Processed_Marriage_Divorce_DB.csv")
```

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

Covariance matrix

Table 1: Table 1.1: Covariance matrix

	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

Correlation matrix

Table 2: Table 1.2: Correlation matrix

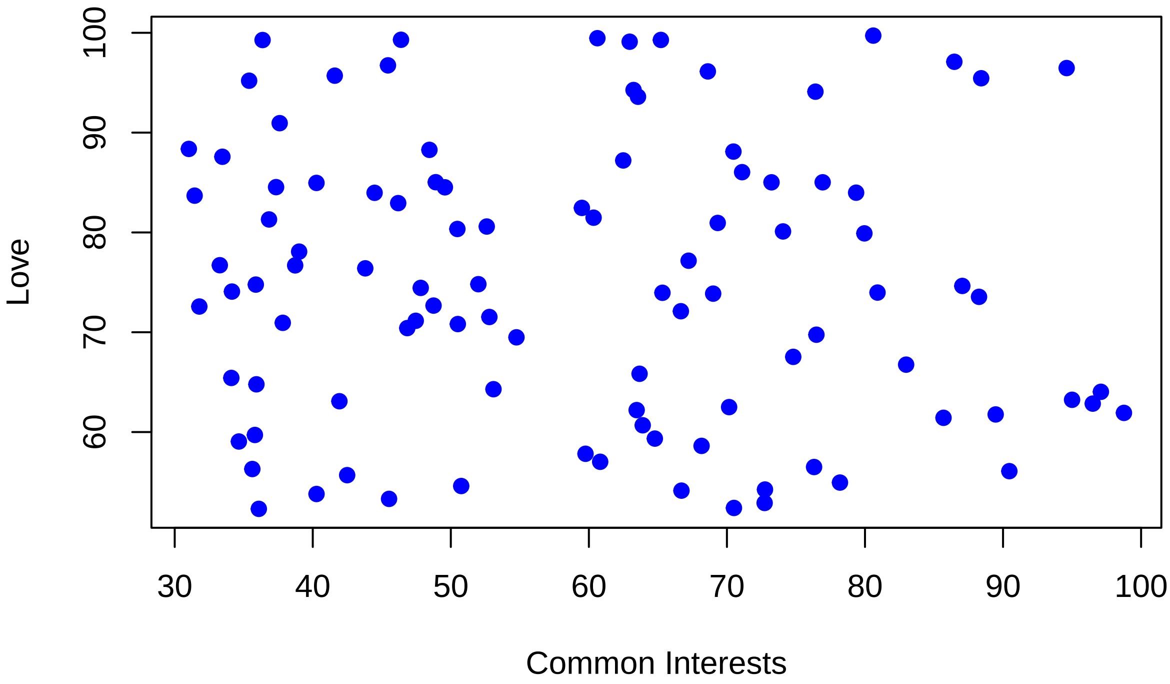
	Age.Gap	Economic.Similarity	Common.Interests	Divorce.Score
Age.Gap	1.0000000	-0.0692849	-0.2400784	0.0970502
Economic.Similarity	-0.0692849	1.0000000	0.1256565	-0.0229647
Common.Interests	-0.2400784	0.1256565	1.0000000	-0.0455697
Divorce.Score	0.0970502	-0.0229647	-0.0455697	1.0000000

Interpretation

- Age Gap and Economic Similarity with negative correlation indicates that people with wider age gap have less economic similarity.
- Age Gap and Common Interests with negative correlation indicates that the wider the age gap, the less common interests they share.
- Economy Similarity and Divorce Score with negative correlation indicates that people with more economic similarity have lower divorce score.
- Common Interests and Divorce Score with negative correlation indicates that people sharing more common interests have lower divorce score

- Age Gap and Divorce Score with positive correlation indicates that the wider the age gap, the higher the divorce score
- Economic Similarity and Common Interests with positive correlation indicates that people with more economic similarity share more common interests.

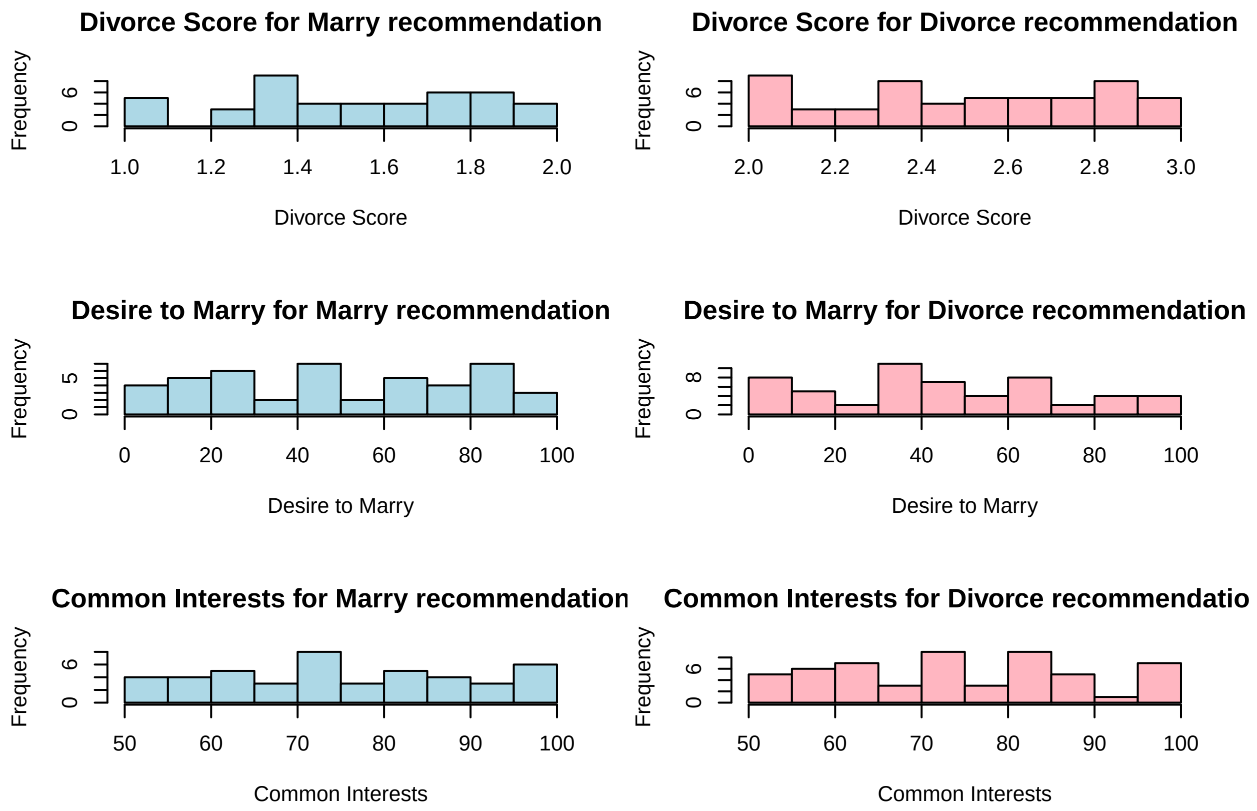
2. Create a scatter plot for the attributes Common Interests and Love. Interpret the scatter plot!



```
## [1] -0.07653391
```

Interpretation There is no correlation between Common Interests and Love. This can be explained by a very small $r = -0.076$ between 2 attributes.

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.


Divorce Score for Marry recommendation

Divorce Score for Divorce recommendation

Desire to Marry for Marry recommendation

Desire to Marry for Divorce recommendation

Common Interests for Marry recommendation

Common Interests for Divorce recommendatio

Histogram 1 - Divorce Score for Marry recommendation: This is a unimodal histogram with most of divorce score is in the range of 1.3-1.4, which is to the left of the histogram. Although the histogram is asymetrical, the divorce score mean and median, whose values are both 1.5, lie in the middle of the histogram

Histogram 2 - Divorce Score for Divorce recommendation: This is a multimodal histogram in which the 3 divorce score peaks are 2.0-2.1, 2.3-2.4 and 2.9-3.0. However, the divorce score mean and median, whose values are both 2.5, lies in the middle of the histogram.

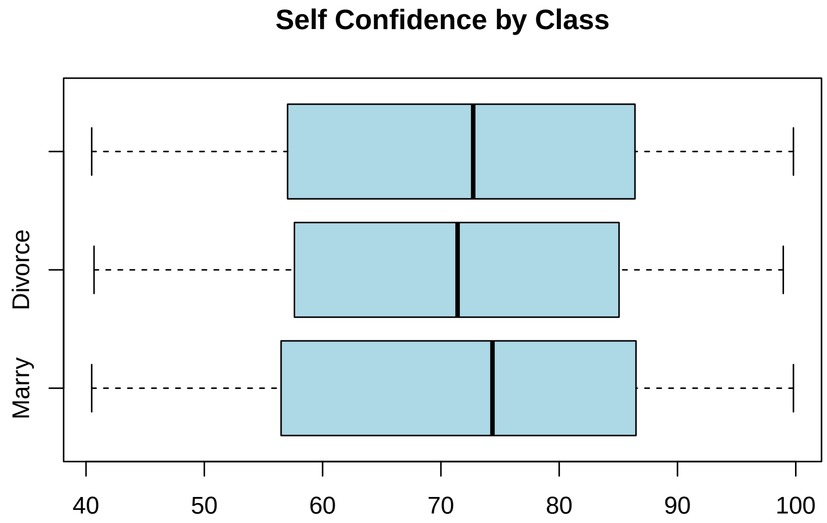
Histogram 3 - Desire to Marry for Marry recommendation: This is a bimodal histogram with 2 peaks at 40-50 and 80-90. Here, the desire to marry mean (approx. 50) is greater than the median (approx. 46), which means there are more values on the right of the histogram

Histogram 4 - Desire to Marry for Divorce recommendation: This is a unimodal histogram with most frequency at 30-40. The histogram also skews to the right, in which the desire to marry mean (approx. 45) is greater than the median (approx. 43).

Histogram 5 - Common Interests for Marry recommendation: This is a symmetrical and unimodal histogram with most frequency at 70-75, which is in the middle. The common interest mean and median also lie in this range, which is both equal to approximately 75.

Histogram 6 - Common Interests for Divorce recommendation: This is a bimodel histogram with 2 peaks at 70-75 and 80-85. In which the mean and median both lie in the middle and is respectively 75 and 74.

4. Create box plots for the Self Confidence attribute for the instances of each 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

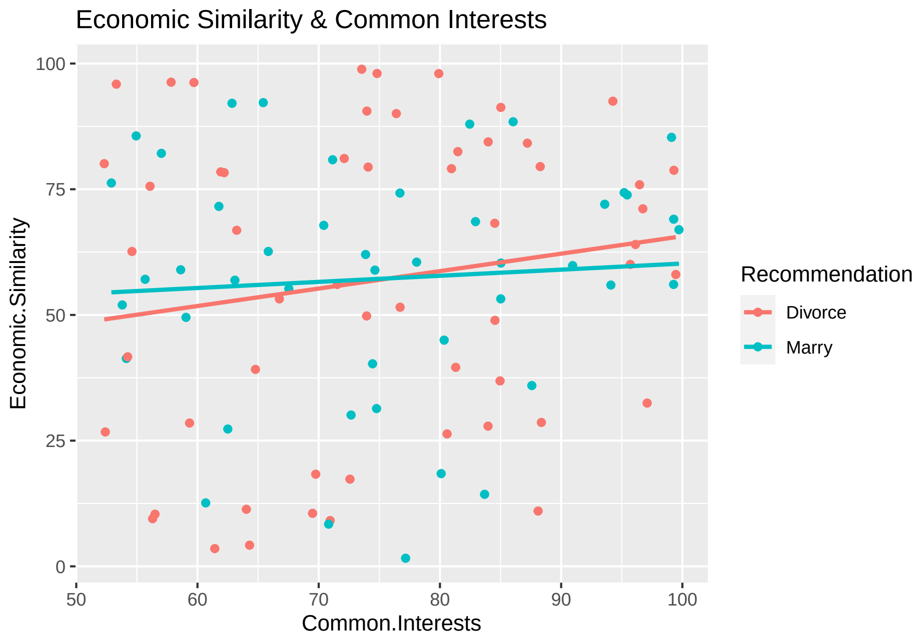


Interpretation

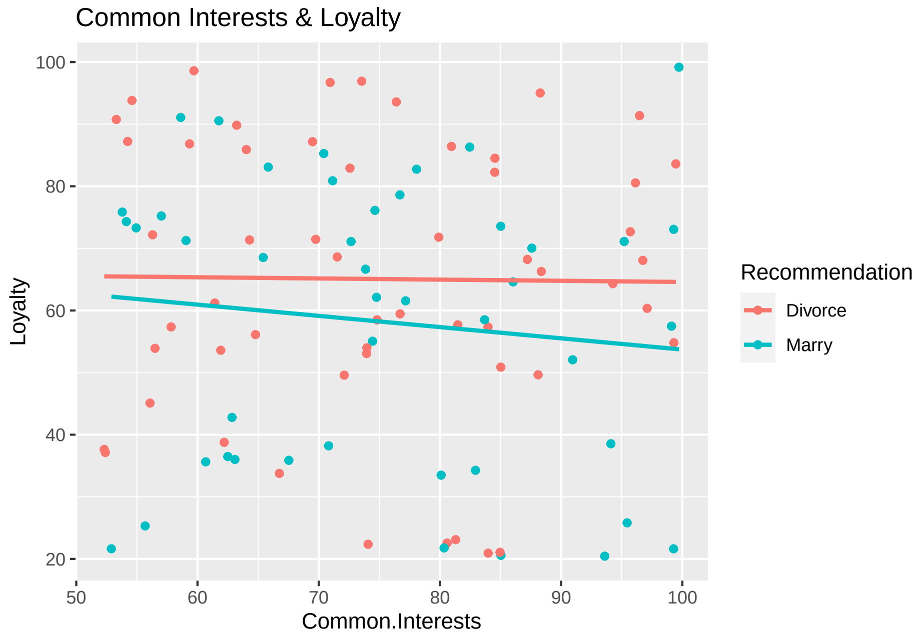
- All 3 boxplots are slightly left-skewed in which the whisker is shorter on the upper end of the boxes.
- The medians of all boxplots are more or less the same, ranging from 70 to 75. To be more specifically, the median of self confidence of class M is the highest (74.36), followed by that of self confidence of both classes (72.73) and lastly the median of self confidence of class D (71.41).
- The 3 whiskers of 3 boxplots are almost the same (99% overlap) with the range from 40 to 100.
- In terms of IQR, the IQR of class "Marry" and of both classes overlap approximately 98%. Class "Divorce" has the narrowest IQR, yet very similar to those of other two boxplots with 90% overlap.
- There are no outliers detected in three boxplots.

5. Create supervised scatter plots/supervised density plots for the following 3 pairs of attributes 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.

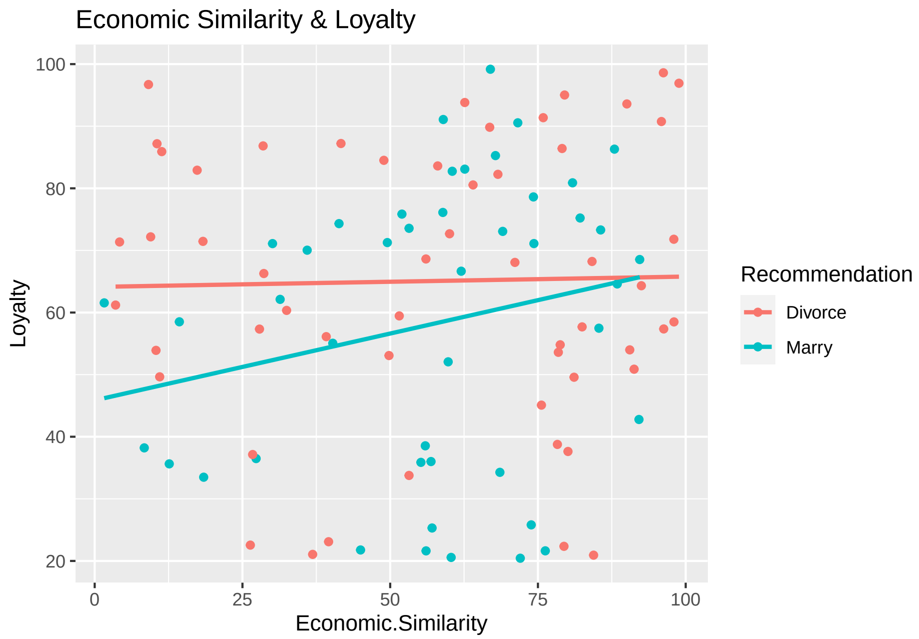
```
## 'geom_smooth()' using formula 'y ~ x'
```



```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

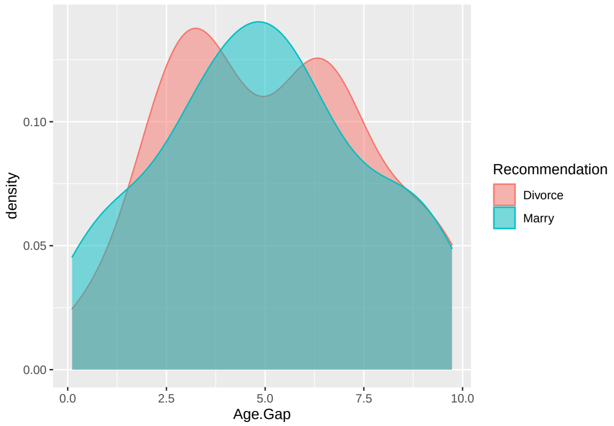


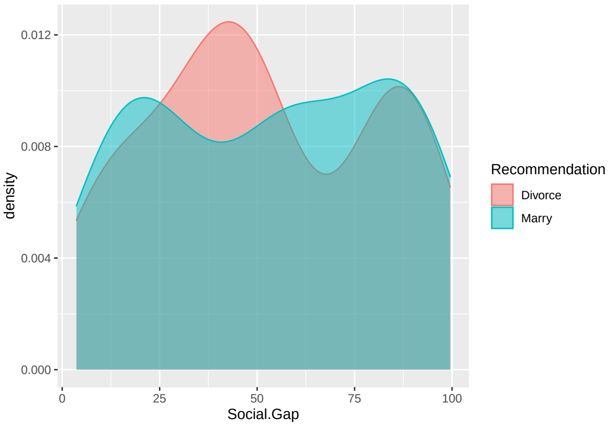
Interpretation

From the three scatterplots, there is a slightly positive correlation between Economic Similarity & Loyalty in class Marry, which can be visualized by the red regression line with a visible upward slope. Except that, there is nearly no correlation shown between the remaining pairs: Economic Similarity & Common Interests and Common Interests & Loyalty.

In addition, the data points in both red and blue are distributed on every region of the graph. This explains why almost all the obtained regression lines have relatively small slopes. Few data points from class Marry and Divorce are very nearby or also overlapped. Hence, the decision boundary between two classes cannot be easily defined and predicting the `Recommendation` attribute can be challenging.

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





In terms of similarity, all four density plots have continuous curves with no gaps or outliers. Both density plots of Class Divorce (red plots) are bimodal with two well separated maxima. However, the plot for Social Gap is not as symmetric as that for Age Gap.

Moreover, density plots of Class Marry have two different distributions. The density plot for Age Gap (blue plot) is symmetric with a bell-curved shape. This density plot is visibly unimodal while the density plot for Social Gap is more likely to be bimodal with quite similar density (approximately 0.01).

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?

```r
#Z-scoring Marriage dataset
Marriage_standardized <- data.frame(scale(Marriage[1:30]))
Marriage_new <- data.frame(Marriage_standardized, Marriage[,31:32])
```

After fitting a linear model to the z-scored dataset, we obtained a regression function with an R-squared of 0.298 and the following coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0672222	0.0564860	36.5970610	0.0000000
Age.Gap	0.0615535	0.0639246	0.9629085	0.3389560
Education	-0.2025734	0.0627008	-3.2307951	0.0018928
Economic.Similarity	-0.0008013	0.0691513	-0.0115877	0.9907880
Social.Similarities	-0.0402279	0.0687034	-0.5855300	0.5601016
Cultural.Similarities	0.0306097	0.0662105	0.4623080	0.6453150
Social.Gap	0.0654290	0.0675590	0.9684726	0.3361909
Common.Interests	-0.0556884	0.0698679	-0.7970516	0.4281552
Religion.Compatibility	-0.0119495	0.0731620	-0.1633295	0.8707364
No.of.Children.from.Previous.Marriage	0.0096192	0.0659522	0.1458504	0.8844648
Desire.to.Marry	-0.1048184	0.0640371	-1.6368375	0.1062178
Independency	0.0282021	0.0682166	0.4134199	0.6805811
Relationship.with.the.Spouse.Family	-0.0405931	0.0683914	-0.5935411	0.5547599
Trading.in	0.0106853	0.0700139	0.1526170	0.8791457
Engagement.Time	0.0288569	0.0666807	0.4327618	0.6665375
Love	0.0870294	0.0633108	1.3746362	0.1736922
Commitment	0.0371127	0.0673503	0.5510407	0.5833857
Mental.Health	0.0526709	0.0653366	0.8061468	0.4229280
The.Sense.of.Having.Children	-0.0104987	0.0672705	-0.1560676	0.8764355
Previous.Trading	-0.0236063	0.0663165	-0.3559646	0.7229531
Previous.Marriage	0.0338585	0.0663047	0.5106492	0.6112267
The.Proportion.of.Common.Genes	-0.0142903	0.0676331	-0.2112919	0.8332823
Addiction	0.0348105	0.0668211	0.5209502	0.6040694
Loyalty	0.0827540	0.0706585	1.1711828	0.2455540
Height.Ratio	0.0080380	0.0651370	0.1234011	0.9021483
Good.Income	-0.1014593	0.0647012	-1.5681212	0.1214278
Self.Confidence	-0.0110093	0.0728094	-0.1512072	0.8802535
Relation.with.Non.spouse.Before.Marriage	0.0629088	0.0674263	0.9330000	0.3540734
Spouse.Confirmed.by.Family	0.0131109	0.0654809	0.2002248	0.8418937
Divorce.in.the.Family.of.Grade.1	0.0275265	0.0720250	0.3821797	0.7035027
Start.Socializing.with.the.Opposite.Sex.Age	-0.0978082	0.0680233	-1.4378629	0.1549930

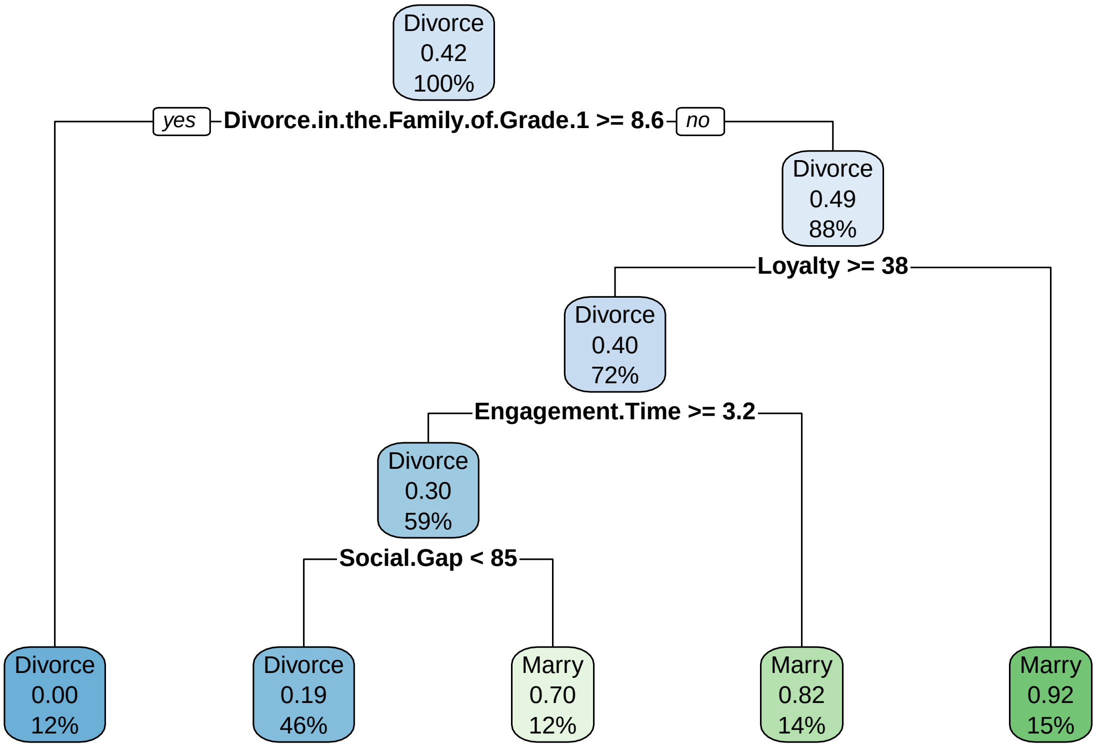
```
[1] "Obtained R2: 0.298348"
```

The obtained coefficients show that `Education` is the most important predictor with $\beta = -0.2$, followed by `Desire to Marry` and `Good Income` both with $\beta = 0.1$. The remaining 27 attributes are visibly not as important, all of which have $\beta < 0.1$. Therefore, the regression model with 30 attributes results in a quite small R2 (0.298) which indicates the model is not predicting a successful marriage very well.

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 Second (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?

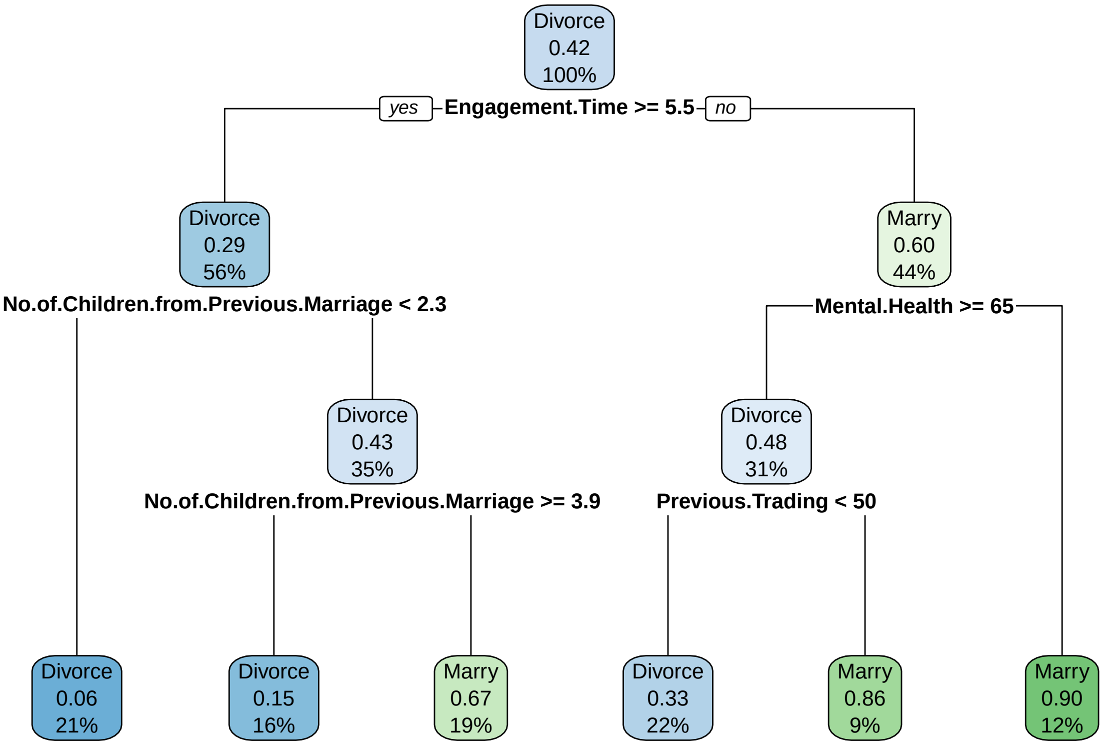
```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
## Loading required package: lattice
```



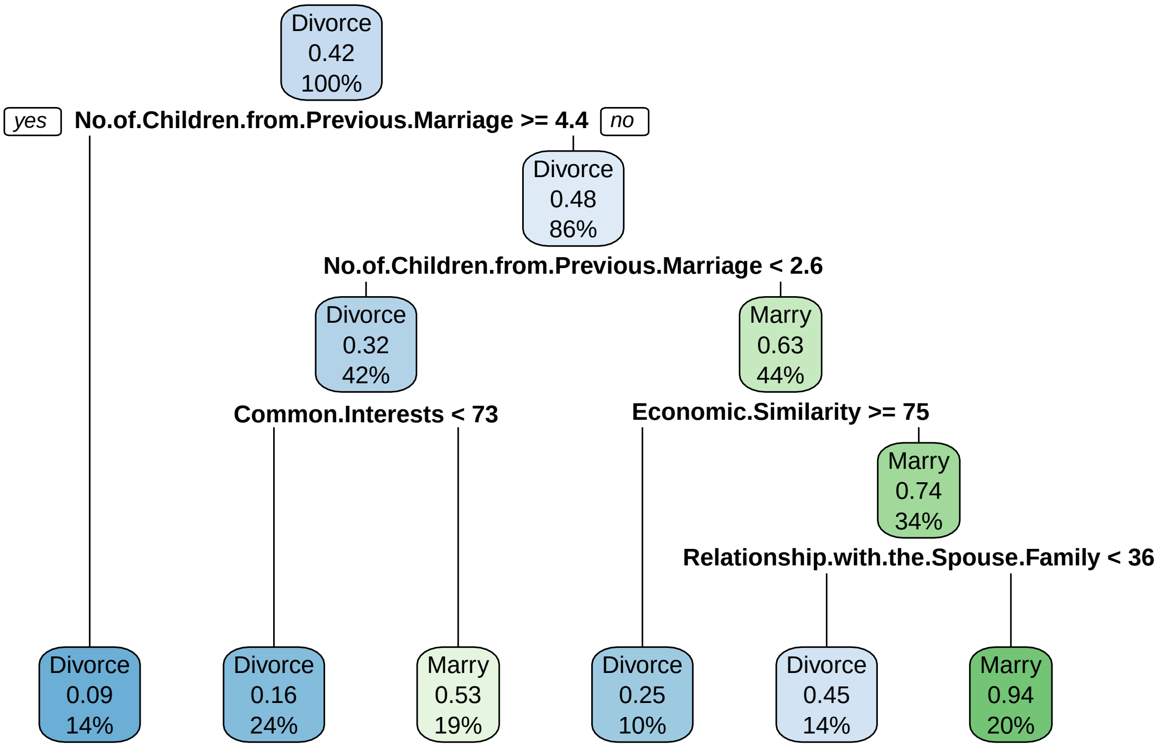
```
##
## tree1.predicted Divorce Marry
##         Divorce       7     6
##         Marry         2     5
```

```
## [1] "Model 1 test accuracy: 0.600000"
```



```
##
## tree2.predicted Divorce Marry
##         Divorce       6     6
##         Marry         3     5
```

```
## [1] "Model 2 test accuracy: 0.550000"
```



```
##
## tree3.predicted Divorce Marry
##         Divorce       5     8
##         Marry         4     3
```

```
## [1] "Model 3 test accuracy: 0.400000"
```

Before fitting the tree models, I divided the dataset using 80/20 split to train the model on the 80% training data, and record the test accuracy on left out 20% testing data.

For the first model, I fitted the tree with all 28 attributes in the dataset, excluding Education and Independency and ended up with a tree of 9 nodes. In this tree, Divorce in the Family of Grade 1 is the most important attribute, followed by Loyalty, Engagement time, and Social Gap. Also, the first tree model obtained a test accuracy of 0.6 (60%).

Proceeding to the second model, the tree is fitted with the first 20 attributes of the dataset, excluding Education and Independency and has a total of 11 nodes. Here, Engagement Time is most important attribute, followed by Number of Children from previous Marriage, Mental Health and Previous Trading. However, this model obtained a lower test accuracy of the first one, which is 0.55 (55%).

Finally, for the third model, I fitted the tree with the first 10 attributes in the dataset, excluding Education and Independency and obtained a tree with 11 nodes. The attributes used in the tree are Number of Children from previous Marriage, Common Interests, Economic Similarity and Relationship with the Spouse Family, which are also the important ones. However, we observed the lowest test accuracy of 40% for this model.

In conclusion, the first tree models constructed with 28 attributes is still the best model to predict `Recommendation`. The most significant attributes should be those used in the first tree as well.

9. Write a conclusion (<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!

Firstly, by observing the correlation matrix, there are very little correlation between the attributes in the dataset. Second, from the scatterplot, we can see that all the data points of two classes scatter on every region of the plots and sometimes overlap. Therefore, it is difficult to define the decision boundary and predict `Recommendation` attribute for this problem. Moreover, if we fit a linear model that predicts the Divorce Score attribute using the 30 z-scored attributes, we obtained a relatively low R2, which indicates that this is not a good model to predict a successful marriage. According to the coefficients in regression function, the most important attributes are `Education`, `Desire to Marry` and `Good Income` with significantly higher slopes than the remaining attributes. Meanwhile, when constructing the decision trees, we observed something different. Here, the most important attributes are Divorce in the Family of Grade 1, Loyalty, Engagement time, and Social Gap, which is obtained by the tree model with the largest test accuracy. Despite the difference, all the attributes I have mentioned are very likely to contribute to a successful or unsuccessful marriage.