**COSC 3337 Problem Set 1**

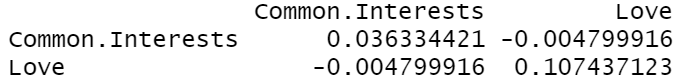
I included a couple of libraries in my program so I have the respective install.packages() function calls at the beginning of the source code file so that you can download the packages as well. I also used the Marriage\_Divorce\_DB.csv file provided by Raunak, I have the .csv attached but you may have to edit the read.csv function that I have at the beginning of the source code and change the path of where the database is. You may have to use double forward slashes (\\) instead of single slashes if you copy the path of the directory. If that is a problem, please let me know!

1. I first used a for loop that goes through every row in our dataset and took the Divorce.Probability of each tuple and assigned a new column named Recommendation that will have the value “Divorce” or “Marry” depending on the probability score.
2. To start task 1, I created a new data frame that contains the following columns; Age.Gap, Economic.Similarity, Common.Interests, and Divorce.Probability. I then normalized that data by applying log transformation to scale them to an easier value to work with. Below are the matrices for both the covariance and the correlation of the attributes.

Table

Description automatically generated

Here, we can see that all relationships are mostly weak relationships since the values never really go above .15. We can also see that 4 of the 6 observed values of correlation and covariance are negative which indicate that most of these relationships exhibit an inverse relationship meaning that when one increase, the other decreases and vice versa. All covariance values of the pairs are relatively close to 0 which tells me that the correlation of the pairs tend to be weak in nature. We can also conclude the previous statement by further examining the correlation values of the pairs. We know that correlation values span from [-1,1] so to have values that never go above +-.17 means that these attributes are weakly related to each other and serve little purpose.

1. Chart, scatter chart

   Description automatically generatedThe scatter plot below shows the visual relationship between the Common Interest and Love attribute. As we can see, there is not much correlation between the two and to further support my previous task, I went ahead and calculated the covariance of these two attributes and the values turned out to be

As we can see, the value is extremely close to 0 which we can visually see by looking at the graph which appears to have no correlation or direction. The values are negative which is slightly visualized in the graph with there being an small empty space in the top right corner.

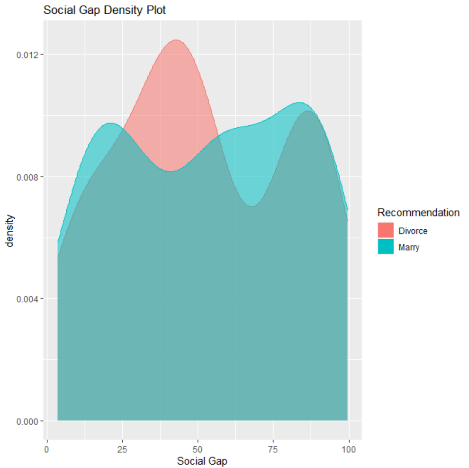
1. For this task (and a couple of others), I used the library multipanelfigure that is a tool to plot several different graphs into the same figure. I created two separate vectors that hold the indexes of the respective Marry and Divorce values of the data frame and then made a histogram using these indexes to distinguish Marry and Divorce rows. The divorce Probability value is A picture containing text, writing implement, stationary, pencil

   Description automatically generatedsurprisingly evenly distributed despite both histograms are for their respective recommendation value but we do see a heavy frequency in the middle of both histograms. There is a little steep in both the Desire to marry histograms that is more directly correlated to the recommendation value. There is a positive slope of frequency as the desire of marry value increases in the marry indexes and there is a negative slope of frequency as the desire of marry value decreases in the divorce indexes. Common interests values of the marry indexes seem more evenly distributed which provides an aspect of consistency whereas the divorce indexes common interest values are more extreme in nature. Both histograms are densely populated in the 70-75 and 80-85 range which indicate that there is a high frequency of records in those ranges in the data set.
2. Chart

   Description automatically generatedHere, I used the multipanelfigure library once more to create two plots in one figure, one plot contains the boxplots for both Marry and Divorce Recommendations with respect to the self confidence attribute and the other plot is a boxplot of all the Recommendation value records with respect to the self confidence attribute. First thing I notice is that none of these boxplots contain outliers therefore all values fall into the range of [25th percentile – 1.5\*(interquartile range), 75th percentile + 1.5\*(interquartile range)]. The interquartile ranges of both Marry and Divorce indexes are of decent sizes and are relatively similar to each other. Although the Interquartile range is larger for Marry, the median self confidence value for Marry is slightly larger than the Divorce index which would mean that the self confidence of Marry index records are concentrated in the range from the 50th percentile to the 75th percentile. When we combine these two boxplots together, it looks like the median and percentile values are the middle of the two indexes from the previous plot. The median in the combined plot seems to be perfectly in the middle of the Marry median and the Divorce median, same goes with the interquartile range of the combined plot.
3. Chart, scatter chart

   Description automatically generatedIn task 5, I used the multipanelfigure library to plot the scatter plots for the following

In these plots, I used a supervised scatter plot that used the Recommendation attribute as the color aesthetic in ggplot and added a liner regression model of the respective Recommendation value with standard error density around the regression line. Here we can visually observe the relationship of values in each attribute based on the Recommendation class. We can see that both classes have more of a direct relationship in the Economic Similarity compared to Common Interests plot. This implies that these two values are directly related to each other not matter what the Recommendation value is so it is true across all tuple that when the Common Interest value increases, Economic Similarity tends to increase as well. Same goes for the Common Interest compared with Loyalty plot, we see that the Marry relation is more inversed than the Divorce relation which means higher Common Interest score tends to lower overall loyalty in most married couple. Finally, we have the Loyalty compared with Economic Similarity plot which suggests the same as the previous plot but in a direct relation. The Marry relation is steeper than the Divorce relation which would mean that Loyalty increases when Economic Similarity increases more often in married couple.

1. Chart

   Description automatically generated To the left, you can observe the density plots that I have made for both Age Gap attribute and Social Gap attribute with respect to their value of Recommendation. First thing I want to point out in the differences is the Y limit. We can see that in the Age Gap density plot, the maximum value is about .175 but in the Social Gap density plot we can see that value only reaches a maximum height of about .012. This evaluation tells me that the data is more centric in the Age Gap plot than the Social Gap plot, by centric I mean that most values in Age Gap are more closely packed in the median interquartile range of the Age Gap attribute. This analysis applies to both Marry and Divorce indexes as well since we see multiple peaks more in the middle of the dataset. This would mean that the Social Gap variable is more evenly distributed which shows in plot by illustrating a plateau-like structure to both Recommendation attributes.
2. Table

   Description automatically generatedFor this task, I made an empty data frame that has 100 rows and 0 columns called zscores. I then made a for loop that goes through all the columns of the original data, calculated the zscores for every value in the column (or vector in my case) and used cbind to combine all vectors into respective columns. I transferred the column names from the main data frame to the zscorez data frame to add some clarity to the data set as well. After the data set has been successfully made, I used the linear regression function and use Divorce Score attribute as the dependent variable and used every column in the data set zscores as independent variables. Below is the summary of my linear regression of the zscores data set. Here we can see the that each attribute has coefficients that are directly related to the Divorce Probability score. We can use t-value to determine the influence that this specific attribute has on the Divroce Probability score, the higher the number, the more influence this attribute has on the probability score. We can see that an obvious attribute that would highly impact the divroce prbability is the Loyalty attribute which would make sense in terms of a real life situation. We can also infer that Education is an attribute that does directly affect the divroce probability but in an inverse relation meaning we will have a lareger divroce probability value if the education value is smaller. R^2 value is .2983 which is relatively low since then value is in the range of [-1,1]. This value means that the regression model only can effectively explain about ~29% of the data set in terms of these coefficients.
3. Diagram

   Description automatically generatedTimeline

   Description automatically generatedTimeline

   Description automatically generatedFor this task, I used rpart and rpart.tool libraris to effectively plot and calculate the accuracy of the training and testing sets. Although we used all 28 attributes to define this decision tree model, only a few were used which indicates that these values had a higher impact on predicting the Recommendation value. When a path has less nodes, that means that these variables are more defined in terms of separating the data between the Marry and Divorce values. As we can see, the relationship with the spouse family attribute has a low impact on the Recommendation value because there are multiple instances that include each Marry and Divorce attributes on either side of the first node in the decision tree. On the next model I used 9 attributes of the data set starting from trading.in attribute to the proportion.of.common.genes attribute. This plot shows the same situation from the prvious decision tree but in a smaller and more confined set of attributes. We should also pay attention to the GINI index of the given Recommendation values provided in the tree. The higher the GINI index, the more divided the values are in terms of the Recommendation attribute value in that specific partition of the data. We can see from the first decision tree model, the middle bottom divroce value is 0 which means that all values that land in the partition Relationship.with.the.spouse.family > 52 -> Social.Gap < 53 -> Economic.Similarity < 46 should all have a value of divorce for the Recommendation attribute. We will also take a look at the second plot that has only one node on the right side of Engagement.Time that has the value of Marry with a GINI index of .78. Since it is a broad term and only one variable constraint is introduced, there will be more variability in the values of Recommendation that land in the Engagement.Time constraint of Engagement.Time < 3.3. For the thrid plot, I used 9 attributes that use the addiction attribute to the start.socializing.with.the.opposite.sex attribute. We can see the same pattern mentioned before with there only being one right child in the first given contraint. Since only one constraint is defined (Addiction < 1.4), there is a high variablitiy with the Recommendation Marry (GINI index = .88)
4. Overall, it seems like there are a lot of attributes in here that don’t have a strong relationship with the divroce probability score. We can easily visualize this by examining the regression model that we observed in task 7, the t-values hold importance as to how much relevance that spefici coefficient (or attribute) has on the divorce probability score. We concluded that Education is an attribute the will effect the divorce probability score directly but in an inverse relation which suggest that the divorce probability score increases as the education score decreases. We can also relate some of these tasks to each other in order to get a deeper understanding of the data set and the results of the attributes. For instance, from task 1, we concluded that the two attributes that are most correlated are the Common.Interest attribute and the Age.Gap attribute. We can then reference the linear regression model that we obtained in task 7 by looking at the t values of the two attributes. Theoretically, if the two values are slightly correlated, then they should have a similar t-value on the linear regression model which is proved since the t-value for Age.Gap is .963 and the t-value for the Common.Interest is -.797. Although they have differing signs, they still have similar impacts on the divorce probability score and recommendation value. I believe that I have learned a lot of useful functions while doing this task set. I learned how to create a model that creates training and testing sets to successfully predict a specific value in the dataset (in this case, the recommendation value) and how to interpret the results. I also have a deeper understanding of certain data science terms and what hypothesis they suggest for certain attributes in the data set such as correlation, covariance, R^2 value, t-value, etc…