Dr. Eick

COSC 3337 *“Data Science I”* Fall 2022

Problem Set1

Third Draft

Last Updated: September 5, 6p

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



Task1 Due: Friday, Sept. 23, 11:59p (electronic Submission)

Responsible TA: Raunak

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**The Marriage and Divorce Dataset is a collection of most important factors leading up to marriage, as determined by a human expert (family counselor). The data was collected over 5 years from 115 subjects.** The goal of this project is to perform exploratory data analysis for the *Processed Marriage and Divorce Dataset* which is a modification of the Marriage Dataset (<https://www.kaggle.com/datasets/hosseinmousavi/marriage-and-divorce-dataset>). The original *Marriage* dataset is a (30+1)D dataset and *Processed Marriage and Divorce Dataset* is a (30+2)D dataset with an nominal Recommendation attribute added; the continous attributes of this dataset are listed below; their range in square brackets:

1 Age Gap [0 10]

2 Education [1 5]

3 Economic Similarity [1 100]

4 Social Similarities [1 100]

5 Cultural Similarities [1 100]

6 Social Gap [1 100]

7 Common Interests [50 100]

8 Religion Compatibility [1 100]

9 No of Children from Previous Marriage [1 5]

10 Desire to Marry [1 100]

11 Independency [1 2]

12 Relationship with the Spouse Family [1 100]

13 Trading in [0 100]

14 Engagement Time [1 10]

15 Love [30 100]

16 Commitment [40 100]

17 Mental Health [50 100]

18 The Sense of Having Children [10 100]

19 Previous Trading [1 80]

20 Previous Marriage [1 5]

21 The Proportion of Common Genes [1 50]

22 Addiction [1 5]

23 Loyalty [20 100]

24 Height Ratio [1 100]

25 Good Income [1 100]

26 Self Confidence [40 100]

27 Relation with Non-spouse Before Marriage [1 10]

28 Spouse Confirmed by Family [1 10]

29 Divorce in the Family of Grade 1 [1 10]

30 Start Socializing with the Opposite Sex Age [15 40]

31. (original target variable) Divorce Score [1 3]

32. (processed target variable) Recommendation (‘Marry’ and ‘Divorce’; computed from attribute Divorce Scoer)

The first 3 examples of the dataset are listed below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age Gap | Education | Economic Similarity | Social Similarities | Cultural Similarities | Social Gap | Common Interests | Religion Compatibility | No of Children from Previous Marriage | Desire to Marry |
| 0.111633 | 1.915111 | 10.99868 | 76.45606 | 47.84746 | 50.31766 | 88.0999 | 83.73807 | 4.402822 | 22.86802 |
| 3.355384 | 2.957842 | 82.13812 | 48.65603 | 30.18852 | 54.11461 | 57.02097 | 98.40813 | 4.367024 | 40.33684 |
| 6.527365 | 2.772463 | 26.33783 | 59.35624 | 10.34025 | 76.59538 | 80.59099 | 41.74346 | 1.19712 | 45.94185 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Independency | Relationship with the Spouse Family | Trading in | Engagement Time | Love | Commitment | Mental Health | The Sense of Having Children | Previous Trading | Previous Marriage |
| 1.269738 | 73.20695 | 79.26236 | 5.424734 | 70.47223 | 76.10683 | 70.2418 | 86.13846 | 39.43739 | 2.766927 |
| 1.658179 | 91.66609 | 63.76329 | 4.371315 | 60.81839 | 70.94377 | 80.029 | 60.66823 | 26.25123 | 1.308314 |
| 1.766594 | 17.9265 | 65.0375 | 6.514788 | 52.60183 | 96.09261 | 80.38394 | 28.22565 | 7.228832 | 4.203533 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The Proportion of Common Genes | Addiction | Loyalty | Height Ratio | Good Income | Self Confidence | Relation with Non-spouse Before Marriage | Spouse Confirmed by Family | Divorce in the Family of Grade 1 | Start Socializing with the Opposite Sex Age | Divorce Score | Recommendation |
| 21.03288 | 3.134119 | 49.64848 | 30.82295 | 94.49916 | 45.96482 | 2.03261 | 1.719332 | 2.262242 | 24.35677 | 2.76019 | Divorce |
| 41.25763 | 2.067377 | 75.2207 | 68.26822 | 41.1026 | 65.38771 | 1.053402 | 1.456192 | 9.795998 | 19.66715 | 1.962979 | Marry |
| 23.91732 | 3.599095 | 22.55187 | 59.13487 | 23.05358 | 84.2719 | 8.268308 | 7.095241 | 9.986173 | 15.52252 | 2.858803 | Divorce |

The values of the class attribute Recommendation have been computed from the Divorce Probability attribute as follows: [1.0-2.0)🡪Marry, [2.0-3.0}🡪Divorce; here *Divorce* represents both ‘*do not get married*’ and ‘*high possibility of divorce if married*.’ In general, we are interested in predicting Attributes 31 and 32 using the other attributes; that is, we like to predict whether or not the couple should marry based on the factors described by the values of the first 30 attributes in the dataset. Another subject we are interested in is finding relationships between the continous attributes in the dataset, and to understand what factors influences successful marriages the most.

Task1 Subtasks:

Apply the following exploratory data analysis techniques **using R** or other tools of your liking to your dataset:

1. Create the *Processed Marriage* dataset or use the one created by the TA!
2. 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! **3 points**
3. Create a scatter plot for the attributes Common Interests and Love. Interpret the scatter plot**! 3 points**
4. 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**
5. 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**
6. 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. **6 points**
7. 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**
8. 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**
9. 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? **9 points**
10. 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)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1[[1]](#footnote-1)**: When you submit your task 1 for problem set 1, you should submit a compressed (zipped) folder that contains a word file that displays your graphs and your interpretations. Each interpretation should use complete sentences to describe your findings. Also in the folder, you should include all files used to complete your tasks, such as your R or python files. If you have doubts what to submit send Raunak an e-mail.

1. More detailed submission instructions for Task1 will be added to this specification by Sept. 20, 2022 the latest. [↑](#footnote-ref-1)