**CS573 - HW2**

Wen-Ling Chi

**1. Preprocessing**

After the preprocessing, it will output a new file **dating.csv**.

1. Remove the quotation marks from 8316 cells.
2. Standardized 5707 cells to lower case.
3. Use label encoding to convert the categorical values in columns gender, race, race o and  
   field to numeric values start from 0.
4. The mean values of each column of “pref\_o\_[six attributes]” and “[six attributes]\_important”.

This image is the result of running **preprocessing.py** Text

Description automatically generated

**2. Visualizing interesting trend in data**

1. Contrast how females and males value the six attributes in their romantic partners differently.

Calculate the mean of the six characteristics of the subset of male and female.

Show a single barplot. The visualization code is in **2\_1.py**.

**Chart, bar chart

Description automatically generated**

**a. What do you observe from this visualization?**

Female’s ratings for each attribute are relatively average.

But men pay more attention to "attractive", and less attention to "ambition" and "shared\_interests".

But between female and male, "sincere", "intelligence", "funny" are given high scores.

b. **What characteristics do males favor in their romantic partners?**

Male thinks "attractive" is more important, but they also give high scores to "sincere", "intelligence", and "funny".

1. **How does this differ from what females prefer?**

Female thinks "intelligence" is the most important, but they also give high scores to "attractive", "sincere", and "funny".

1. **How a participant’s rating to their partner on each of the six attributes relate to how likely he/she will decide to give the partner a second date.**

Compute the fraction of participants who decide to give the partner a second date among all participants whose rating of the partner on the chosen attribute (e.g., attractive partner) is the given value (e.g., 10).

Show 6 scatter plots. The visualization code is in **2\_2.py**.

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated** **Chart, scatter chart

Description automatically generated** Chart, scatter chart

Description automatically generated

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**a. What do you observe from these scatter plots?**

Basically, every attribute is positively correlated, which means that the higher the score, the more successful the second date will be.

Attractive, funny, shared\_interests: The slopes of these three categories are relatively high, which means that the higher the score, the easier it is to succeed. But there are a few exceptions

Intelligence, ambition, sincere: The slopes of these three categories are relatively low, which means that higher scores may not necessarily lead to success.

**3. Convert continuous attributes to categorical attributes**

Discretize all columns in *continuous valued columns* by splitting them into **5** bins of equal-width in the range of values for that column.

According **fieldmeaning.pdf**, there are 47 columns needing to discretize.

The script reads **dating.csv** as input and produces **dating-binned.csv** as output.

This image is the result of running **discretize.py**

Text

Description automatically generatedText

Description automatically generated

**4. Training-Test Split**

Use the ***sample*** function from ***pandas*** with the parameters initialized as random state = 47,  
frac = 0.2 to take a random 20% sample from the entire dataset.

Create a new script called **split.py** that takes **dating-binned.csv** as input and outputs **trainingSet.csv** and **testSet.csv**.

* Total has 6744 rows
* trainingSet has 5395 rows (80%)
* testSet has 1349 rows (20%)