Data Mining Homework 2

Exercise 1

Task 1: What is the Data about?

The data measures the job satisfaction level and capital gain/loss of different individuals in the working class ages. It shows details about each individual, including their Age, Working Class, Highest level of Education, Native Country, Gender and Salary.

Task 2: What are different features and their type?

The different features in the data frame are as follows:

Name	Data Type	Feature Type	Comment
Age	Factor (Integer)	Discrete	Age of each individual
Workclass	Factor (Character)	Categorical (Nominal without Order)	This feature represents the current working status of the individual. It contains a set of predefined values
Education	Factor (Character)	Categorical (Nominal with Order)	This feature indicates the highest level of education of the individual. Although it has predefined possible values, some values indicate a category e.g. 1st-4th, 5th-6th etc. These values indicate a possible mid-point
Occupation	Factor (Character)	Categorical	This feature indicates the category of the occupation of the individual. It has a predefined set of groups and each individual is placed in one of these groups
Capital.gain	Integer	Continuous	This feature indicates the gain in capital for the individual. I consider it continuous because the gain for each individual is different and it is different for each individual
Capital.loss	Integer	Continuous	This feature indicates the loss in capital. It is the converse of the Capital.gain. Same reason as above
Native.country	Factor (Character)	Categorical	This feature indicates the nationality of the individual. It is categorical because it groups the individual by their nationality, it does not

			show the state the individual comes from
Salaries	Numeric	Continuous	This feature indicates the salary earning of the individual. It is continuous because the possible values cannot be totally covered
JobSatisfaction	Factor (Integer)	Categorical (Ordinal)	This feature measures the satisfaction level of each individual on his or her job. It is categorical because it groups the satisfaction level on a scale between 0 and 15. It can also be considered ordinal because it indicates a numerical scale from low to high
Male	Integer	Discrete (Dichotomous)	This feature indicates if the individual is male or not. It is dichotomous because it is only two possible values, 1 or NA. It is discrete because its value is numeric. It should be merged with the *female* feature
Female	Integer	Discrete (Dichotomous)	This feature indicates if the individual is female or not. The reason for Feature classification is the same as for *male*. It should also be merged with the *male* feature

Task 3: How many rows are in the dataset?

32561 rows (excluding the headers)

Task 4: Identify and fix the problems in the initial dataset. Describe all changes you made and why

Steps taken to clean up the data

1. I changed all instances of '?' to NA in the following fields: age, workclass, occupation, native.country
2. I replace "Very Good" in jobstatisfaction field with NA
3. I replaced -57 with 57 and 320 with 32 in the age field
4. I renamed "privat" to "Private"
5. I renamed "Unitedstates", "United-states" and "United States" with "United-States" in the native.country field

6. I merged the male and female columns. If the value of the male column is 1, I put “Male” in the gender column while if the value of the female column is 1, I replaced it with “Female” in the gender column
7. I also merged the capital gain and capital loss into one column. I used capital. Change to represent the data. If the capital gain was available I represented it as positive, while if the capital loss was available, I represented it as negative
8. I trimmed out white space in the text in the character fields.

Exercise 2

Categorize the data

Numerical features:

1) **Age**:
    a. Mean: 38.56071
    b. Median: 37
    c. Max: 90
    d. Min: 17
    e. SD: 13.64282
    f. Missing Records: 97
2) **Capital.change**:
    a. Mean: 990.345
    b. Median: 0
    c. Max: 99999
    d. Min: -4356
    e. SD: 7408.987
    f. Missing: 0
3) **Salaries**:
    a. Mean: 39306.46
    b. Median: 33927
    c. Max: 140000
    d. Min: 38.34125
    e. SD: 17356.39
    f. Missing: 0
4) **Job Satisfaction**:
    a. Mean: 7.548771
    b. Median: 7
    c. Max: 15
    d. Min: 0
    e. SD: 4.457437
    f. Missing: 1

Categorical features

1) WorkClass
    a. Missing: 1836
2) Education

    a. Missing: 0

3) Occupation:
    a. Missing: 1843
4) Native Country:
    a. Missing: 583
5) Gender:
    a. Missing: 0

Comment on the implications of the outcome (is the distribution skewed - more small/large values, any unexpectancies).

The distribution is skewed correctly