Data Mining

Exercise 1 Missing values : Social Survey

A social survey has been performed on 10 persons. The four first characteristics are presented below.

	gender	age	residence	earning	number of children
1	M	22	urban	54	0
2	M	53	urban	67	0
3	M	36	suburban	MV	1
4	M	59	rural	46	5
5	M	53	rural	40	3
6	M	49	urban	78	2
7	F	25	suburban	49	1
8	F	22	urban	37	0
9	F	35	MV	58	1
10	F	45	MV	MV	2

1. Specify the type of each attribute.
2. Compute the percentage of missing values for each attribute.
3. Compute the mean, the standard deviation and the median for the *earning* attribute. Compute also the mode for the *residence* attribute.
4. Complete the table with mean/mode methods following the type of data. Calculate the new mean, standard deviation, median and mode.
5. Complete the *earning* attribute by first separating with the gender, then by applying a 1-NN by taking in account the age and the number of children. Calculate the new mean, standard deviation and median.

Exercise 2 Missing values : Blood pressure

The following table gives is a sample giving the age and the mean blood pressure.

age x_i	36	42	48	50	54	60
blood pressure y_i	12	13.5	13.6	/	14.3	15.4

We want to use a single imputation method to deal with the missing value.

1. Supposing that there exists a linear dependance between the two variables, what single imputation method would you use ?
2. Check the linear dependance hypothesis by filling the following table.

x_i	36	42	48	50	54	60
y_i	12	13.5	13.6	/	14.3	15.4
y_i^*						
e_i						

   First apply a listwise deletion. If there exists a linear dependance, then $y = ax + b$ such that $a = \frac{\sigma_{xy}}{\sigma_x^2}$ and $b = \bar{y} - b\bar{x}$, where $\bar{x}$ and $\bar{y}$ are the mean of the variables, $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and σ_x the standard deviation. We define the error e_i as the difference between the theoretical values y_i^* and the observed value y_i.
3. Conclude about the relevance of the linear dependance.
4. Fill the missing value with the method chosen.