IS4240 Business Intelligence Systems – Assignment 1
Raghavendhra Balaraman – A0123443R

a) Provide the R codes for loading the dataset into a variable heart. The attributes should be given reasonable attribute names based on the description given above. Ensure that all the attributes are of numeric (or integer) type. (Hint: you should be able to easily convert missing values to be of NA type by using an appropriate function argument) (3 marks)

```
heart <- read.table("processed.va.data", sep=",", col.names = c("age",
"sex", "chest_pain_type", "resting_blood_pressure", "serum_cholestoral",
"fasting_blood_sugar_greater_than_120", "resting_electrocardiograph",
"maximum_heart_rate_achieved", "exercise_induced_angina",
"stdepression_induced_by_exercise_related_to_rest",
"slope_of_peak_exercise", "flourosopy_colored_vessels", "thal",
"heart_disease_diagnosis"), na.strings=c("?"), colClasses="numeric")
```

Note :

- The reasonable attributes names are given using 'col.names'
- **colClassess = 'numeric'** ensures that all the attributes are numeric
- **na.strings=c("?")** replaces all the missing values (denoted as '?' in the dataset) with NA type.

b) Provide the R codes for getting the number of missing values for each attribute. Fill in the table below. (5 marks)

The number of missing values for each attribute is identified using the below R code.

```
sum(is.na(df$attribute))
```

Here $attribute denotes the attribute names in the dataset

Attribute	R code	No. of missing values
age	sum(is.na(df$age))	0
sex	sum(is.na(df$sex))	0
cp	sum(is.na(df$chest_pain_type))	0
trestbps	sum(is.na(df$resting_blood_pressure))	56
chol	sum(is.na(df$serum_cholestrol))	7
fbs	sum(is.na(df$fating_blood_sugar_greater_than_120))	7
restecg	sum(is.na(df$resting_electrocardiograph))	0
thalach	sum(is.na(df$maximum_heart_rate_achieved))	53
exang	sum(is.na(df$exercise_induced_angina))	53
oldpeak	sum(is.na(df$stdepression_induced_by_exercise_related_to_rest))	56

slope	sum(is.na(df$slope_of_peak_exercise))	102
ca	sum(is.na(df$flourosopy_colored_vessels))	198
thal	sum(is.na(df$thal))	166
num	sum(is.na(df$heart_disease_diagnosis))	0

c) **Based on the number of missing values for each attribute, discuss one potential issue if we were to remove instances with one or more missing attributes. (4 mark)**

The following are some of the issues in removing instances with one or more missing values.

1) The removal of missing data could affect the statistical analysis of dataset. For example, while calculating the average of an attribute, the analysis would provide a different result with instances containing missing records and without those instances. From the above number of missing values details, we find that the most missing attributes are ‘ca’, ‘thal’, ‘slope’ with their number of missing values 198,166,102 respectively. The below table gives a good representation of the issue in removal of missing attribute records.

Attribute	Average with instances containing missing value	Average without instances containing missing value
Thal	1.07	6.29
Slope(slope_of_peak_exercise)	1.045	2.132

From the above results, we can infer that the average of an attribute varies significantly when missing records are accounted and when not accounted. Depending on the analysis we perform, we have to decide whether the instances of missing attribute are to be accounted or not.

2) Removing instances of missing records not only affects the analysis of that particular attribute but also affects the analysis of other attributes that contains valid data.

d) **Instead of removing instances with one or more missing attributes, propose an alternative approach for handling this problem? (4 mark)**

One traditional method of handling missing attributes in an instance is by replacing the missing data with substituted values. **This substituted values could be either '0' or some default value.** This method in data mining is called as 'Imputation'. For example, in the above case, the missing records for the attribute 'Thal' can be replaced with a default values of '3' which means 'normal'.

The other case would be to **replace the missing values with the sample mean of the data** measured by not accounting the instances containing missing values. For example, in the above case, the missing records for the attribute 'Slope (slope of peak exercise)' can be replaced with the value of '2' as the sample mean of this attribute without accounting the instances of the missing value. This would result in the total sample mean not being biased and tend to remain the same.

e) **Provide the R codes for generating the correlation matrix for the attributes: age, sex, cp, restecg, num. Show the correlation matrix. (4 mark)**

```
x <- data.frame(heart$age, heart$sex, heart$chest_pain_type,
heart$resting_electrocardiograph, heart$heart_disease_diagnosis)

y <- data.frame(heart$age, heart$sex, heart$chest_pain_type,
heart$resting_electrocardiograph, heart$heart_disease_diagnosis)

cor(x,y)
```

Correlation Matrix:

	heart.age	**heart.sex**	**heart.chest_pain_type**	**heart.resting_electrocardiograph**	**heart.heart_disease_diagnosis**
heart.age	1.000000000	0.03423008	-0.03827995	0.002400114	0.28728870
heart.sex	0.034230084	1.00000000	0.03803637	0.060620741	0.14746969
heart.chest_pain_type	-0.038279952	0.03803637	1.00000000	0.034789801	0.16821015
heart.resting_electrocardiograph	0.002400114	0.06062074	0.03478980	1.000000000	-0.03280011
heart.heart_disease_diagnosis	0.287288697	0.14746969	0.16821015	-0.032800115	1.00000000