1. The datafile Real estate valuation data set.csv shows values of

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
X2=the house age (unit: year)
X3=Dist.MRT = the distance to the nearest MRT station (unit: meter)
X4=N.CStores = the number of convenience stores in the living circle on foot (integer)
X5=the geographic coordinate, latitude. (unit: degree)
X6=the geographic coordinate, longitude. (unit: degree)
Dist = approximate dist between (latitude, longitude) and downlown (miles), and the response variable, calculated using latitude and longitude in the data file and the lat-long of downtown `(lat0 = 25.105497, long0 = 121.597366).`

Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

Reference:
Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

For the following problems, use a 75% training – 25% test split, and also set seed in R as follows:
set.seed(1197317)

(a) Fit MLR model to Y as a function of the potential predictors Age, Dist.MRT, N.Cstores, Dist. Verify assumptions, using the training set. Compute RMSE and R^2 for both training and test sets.

(b) Fit SVM model to Y as a function of the potential predictors Age, Dist.MRT, N.Cstores, Dist. Verify assumptions, using the training set. Compute RMSE and R^2 for both training and test sets.

(c) Compare the results for MLR and SVM for both training and test sets.

2. The datafile bank.1.csv has values of:
   1 - age (numeric)
   2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
   3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
   4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
   5 - default: has credit in default? (categorical: 'no','yes','unknown')
   6 - balance (continuous)
   7 - housing: has housing loan? (categorical: 'no','yes','unknown')
   8 - loan: has personal loan? (categorical: 'no','yes','unknown')
   # related with the last contact of the current campaign:
   9 - contact: contact communication type (categorical: 'cellular','telephone')
   10 – day (1-31; do not use this predictor)
   11 – month - discard
   12 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
   # other attributes:
   13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
   14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
   15 - previous: number of contacts performed before this campaign and for this client (numeric)
   16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
   # social and economic context attributes

   Output variable (desired target):
   17 - y - has the client subscribed a term deposit? (binary: 'yes','no')

   (a) Fit a logistic regression model to the response y=yes, and compute its PRF1 values. (Note: The LR model must have all VIF < 5, and all predictors must be significant at test size 0.05).
   (b) Fit the default random forest model to the response, and compare the PRF1 values of the LR and the default random forest model.
   (c) Extra Credit: Fit the default xgboost model to the response, and compare the PRF1 values of the three fitted models.