Note : Answers are in **bold**.

2.1  Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

1. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers). **[Supervised]**
2. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.  **[Supervised]**
3. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.  **[Supervised]**
4. Identifying segments of similar customers.  **[Unsupervised]**
5. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.  **[Supervised]**
6. Estimating the repair time required for an aircraft based on a trouble ticket.  **[Supervised]**
7. Automated sorting of mail by zip code scanning.  **[Supervised]**
8. Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.  **[Supervised]**
   1. Describe the difference in roles assumed by the validation partition and the test partition.

**Both validation partition and test partitions are the set of data for which the outcome is known. Difference is – If multiple models are used, validation data is used to compare the different models. Test data is used with the finally selected model to predict how well that model with do.**

* 1. Consider the sample from a database of credit applicants in Table 2.15. Comment on the likelihood that it was sampled randomly, and whether it is likely to be a useful sample.

**This does NOT look like randomly selected data. OBS# is a multiple of 8 which means data was selected with a predefined calculation. This data may have inherent biases and therefore not a useful sample.**

* 1. Consider the sample from a bank database shown in Table 2.16; it was selected randomly from a larger database to be the training set. Personal Loan indicates whether a solicitation for a personal loan was accepted and is the response variable. A campaign is planned for a similar solicitation in the future and the bank is looking for a model that will identify likely responders. Examine the data carefully and indicate what your next step would be.

**We already know that what we are trying to do and it’s clear that we need to predict the variable ‘Personal loan’. Also, since the sample data is known it would be supervised learning. Next natural steps would be to explore and clean the data, followed by applying the DM methods to see what would give us the best results.**

* 1. Using the concept of overfitting, explain why when a model is fit to training data, zero error with those data is not necessarily good.

**We use a model essentially to predict the future values. The goal is not to find the best model that works for training data. The training data is just a sample that may contain errors and outliers. When we overfit the mode with training data, we incorporate the errors and outliers of that data too which would lead to bad predictions. This would also mean that model would not adjust with the new data.**

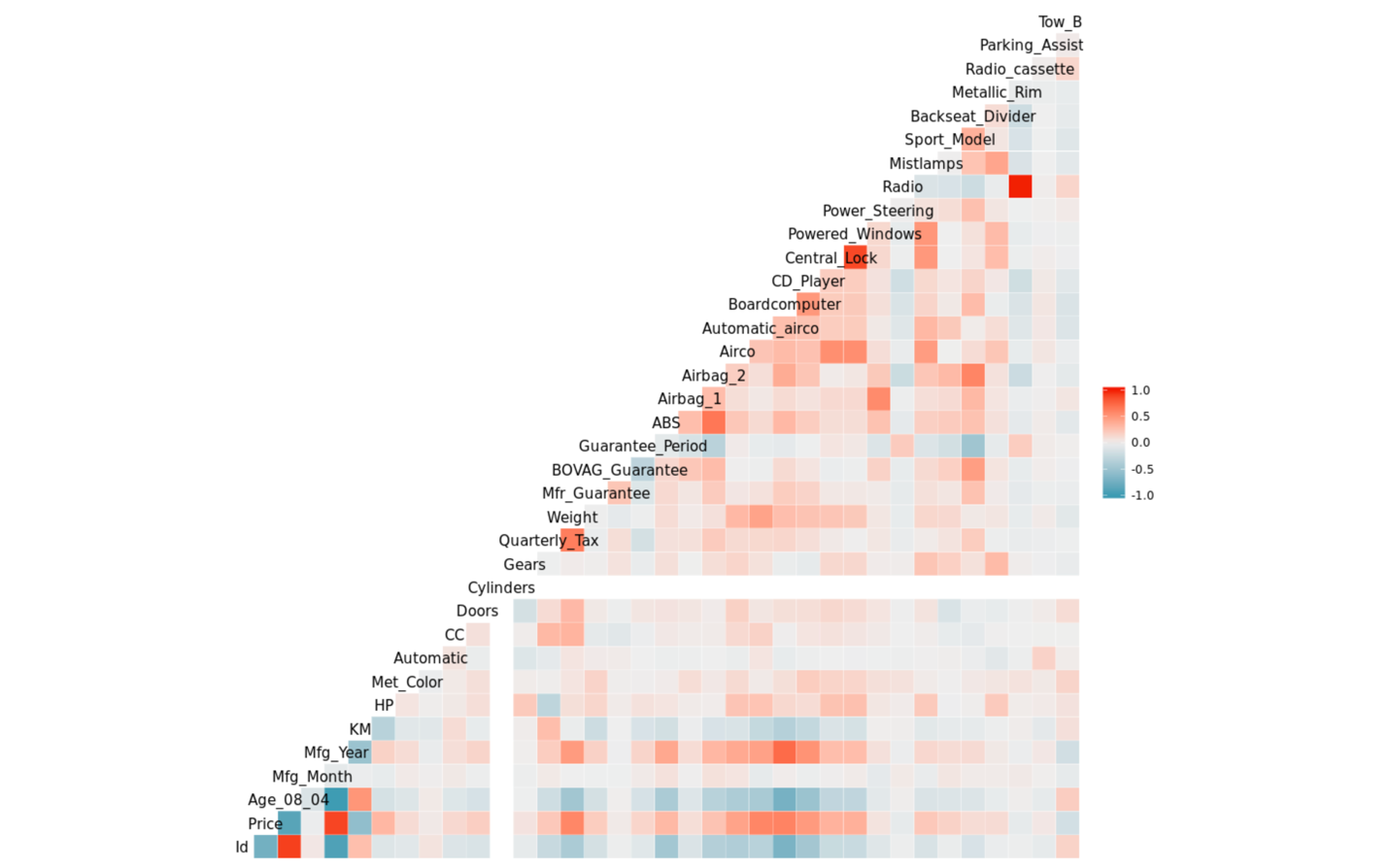
* 1. In fitting a model to classify prospects as purchasers or nonpurchasers, a certain company drew the training data from internal data that include demographic and purchase information. Future data to be classified will be lists purchased from other sources, with demographic (but not purchase) data included. It was found that “refund issued” was a useful predictor in the training data. Why is this not an appropriate variable to include in the model?

**In this case our outcome variable is Purchase. Refund issued is not an appropriate predictor variable because it depends on the purchase variable. We are trying to predict purchasers and refund means there was a successful purchase.**

* 1. NA
  2. NA
  3. NA
  4. NA
  5. The dataset ToyotaCorolla.csv contains data on used cars on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications.

1. Explore the data using the data visualization capabilities of R. Which of the pairs among the variables seem to be correlated?

Cor() is used to demonstrate coreelation in 2 variables age and price. Older vehicles are less expensive , os therei s -ve correlations. I found [here](https://www.datacamp.com/community/blog/r-correlation-tutorial) that ggcorr() function in GGally package may be used to find all corrwelations of all variables, used that to build following graph.



Where the color box represents the correlation of variables.