Sampling, Normalization and Principal Component Analysis

What's the best possible thing that could happen to house prices? Your immediate answer will depend upon your own situation. If you have recently bought with a whopping great mortgage, you will want prices to rise. If you are planning to downsize, you will also want them to rise (you are interested in the cash left over after the move). If you have a good slug of equity and plan to trade up, you'll want prices to fall. The percentage fall of the cheaper house you occupy will cost you less than you will save from the same percentage fall in the more expensive house you want. Whereas if you want to trade up but have very little equity, you'll want house prices to rise.

Source: https://www.ft.com/

The dataset 'kc_house_data' includes 21 attributes which explain the different aspects of a house, each identified by a unique id. These attributes help determine the price of the house. Analysis of such data will help predict the price of houses, which is very important in today's real estate world.

Problem Statement 1

Building any model involves dividing the dataset into training data and test data. Choosing the training data involves sampling.

a) There are several techniques for sampling data and there is no perfect method of sampling. It depends on the problem you are trying to solve and the dataset you are using. List some of the factors that you should consider in choosing the method of sampling to be considered.

b) For the given dataset, obtain the following samples:

1. Simple Random Sample, with 75% of the data
2. Systematic Sample, with every 4^{th} data point (1^{st}, 5^{th}, 9^{th} etc)
3. Clustered Sample, based on the number of bedrooms in the house (choose 60% of the bedrooms)
4. Stratified sample, based on the number of floors (choose 70% of the floors)

Plot the different samples (the *price* attribute).

Which sampling technique above is likely to produce inaccurate information about the dataset?

c) What is Undersampling and Oversampling? When are they used?

Examine the dataset, and state which attribute would be used to perform undersampling or oversampling. Why do you think so?

Problem Statement 2

While building a model, one cannot use all the 20 attributes. This will lead to overfitting of the data or will require the unnecessary usage of a lot of computer resources. So, you must analyse the data, and figure out which attributes will actually contribute to the model and which attributes are redundant or do not convey much information. For this dataset:

a) Get the correlation matrix for the numeric attributes of the dataset and visualize it. According to this, which attribute would you drop and why?

b) "If the Variance of the attribute is very low, you can drop that attribute as it does not really help discriminate". Comment on this statement and support your view with an example.

Analyse the variance/standard deviation of the attributes in the dataset. Which are the 5 attributes with the lowest standard deviations? Out of these, which attributes can you eliminate? Support your answer with a suitable explanation.

c) i) Perform Principal Component Analysis on this dataset.
Find out the variance of the 2 principal components that are capable of representing the whole dataset.
To what extent do these principal components explain the dataset?
(Score for each principal component P_i is the percentage of $(P_i / \sum P_i)$)

iii) Why is normalization and scaling necessary in PCA?

iv) Write two inferences that you can make about the dataset based on PCA?