MA 584 Final Project

This final project can be done individually or by a group (at most two members).
The project will be graded based on two activities: in-class presentation and the report. In-class presentation 10 mins during the last week of the semester. The report will be in the format of a scientific report and is due by May 2nd. You need to answer the following questions based on the dataset in the blackboard.

Description: Ames housing data
http://jse.amstat.org/v19n3/decock.pdf

a. **Numeric variables (Continuous):**
LotFrontage, LotArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, X1stFlrSF, X2ndFlrSF,LowQualFinSF, GrLivArea, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, PoolArea, SalePrice.

b. **Numerical variables (Discrete):**
OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, TotalRmsAbvGrd, Fireplaces, GarageYrBlt, GarageCars, MoSold, YrSold.

c. **Categorical (Ordered):**
ExterQual,ExterCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, HeatingQC, CentralAir KitchenQual, Functional, FirepaceQu, GarageFinish, GarageQual,GarageCond, PavedDrive, PoolQC, Fence.

d. **Categorial(Unordered):** MSSubClass, MSZoning, Condition1, Condition2, Street, Neighborhood, BldgType, HouseStype, RoofStype, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, Foundation, Heating, Electrical, GarageType, MiscFeature,MiscVal, SaleType, SaleCondition, Utilities.

Q1) Perform a regression analysis relating SalePrice to other variables related to the sizes. What is the most important feature in explaining the SalePrice?

Q2) Compute the correlation on these continuous features (except for the SalePrice). Which variables are clustered together? Can we find a lower dimensional representation of continuous features?

Q3) Assess the importance of each categorical features to the SalePrice after adjusting the size effect.

Q4) Are the means of house size and sale price different across the neighborhood?

Q5) Cluster houses based on features except for MSSubClass, MSZoning, Neighborhood. Is the clustering result can be annotated with any of information from MSSubClass, MSZoning, or Neighborhood?

Q6) Identify the neighborhood where the houses are over/underpriced.