CKMT 105
Homework 1
Deadline: 22 Feb 2015

PROBLEM DEFINITION

You are presented with Manhattan real-estate sales data between August 2012 and August 2013 (data is provided as "rollingsales_manhattan.xls"). The goal is to build a prediction model that predicts Manhattan real estate sales price given this historical data.

TASK

1. Import and Clean the data in R. *(10 points)*
2. Identify missing data and propose a method to handle the missing data problem. *(10 points)*
3. Use Visualization techniques to illustrate 5 interesting trends. *(20 points)*
4. Build a univariate linear regression model with only "land square feet" as the input variable. *(20 points)*
5. Build a multivariate linear regression model, using variables of your choice. *(20 points)*
6. Design an experiment, using multivariate linear regression model and k-folds cross validation. *(20 points)*
7. Bonus: Build a stepwise regression model and identify top 5 attributes in terms of information content. *(Extra 20 points)*

OUTPUT

You should upload a zipped file on blackboard containing the following:

1. Cleaned data
2. Report that contains:
    a. The description of data import method
    b. Visualizations
    c. Summary of Linear regression results
    d. Conclusions
    e. Bonus results
3. All the source codes