

CMTH 642

Homework 1

Deadline: 30 October 2015

PROBLEM DEFINITION

Your goal is building a prediction model for real estate sale prices in Manhattan.

DATA

You have Manhattan real-estate sales data for real estate data between August 2012 - August 2013.

TASK

1. Clean the data and import to R.
2. Identify missing data and propose a method to handle the problem of missing data.
3. Visualize 5 interesting trends.
4. Build a uni-variate linear regression model with only land square feet as the input variable.
5. Build a multivariate linear regression model.
6. Design an experiment using multivariate linear regression model and k-folds cross validation.
7. Bonus (extra 20 points): Build a stepwise regression model and identify 5 top attributes in terms of information content.

OUTPUT

A zip file containing the following

- Cleaned data
- Report that contains:
 - The data import method
 - Visualizations
 - Linear regression results summary
 - Conclusions
 - Bonus:
- All the source code