Midterm: Data Analysis Assignment

1. Due Date: November 9th at 1:00 pm. No late work accepted.
2. Your work should be handed electronically in pdf format. You should also provide the R-markdown document you used to generate it.
3. We expect clear and concise explanations and conclusions for all your modeling and statistical investigations. Paper should not be longer than 8 pages (including tables and plots).

Predicting house prices. People interested in buying real estate would like to properly assess the current and sometimes future value of a property before making an investment. We expect the price of a house depends on its physical characteristics: size, age, particular features, quality of construction but also on some external factors like location (the attractiveness of a neighborhood can depend on safety, quality of schools and the overall environment) and current state of the economy.

The data `realestate.csv` is a sample of residential sales in a midwestern city; We would like to fit a multiple linear regression, with sales price as the response and the other variables as predictors. We may have some ideas about the price should behave w.r.t. to the predictors. We are going to use the several statistical tools that we reviewed in class and PC to build this model and study some specific hypothesis.

Your final paper should have the following structure:

- Introduction. Motivate the problem. Describe in a few sentences (say a maximum of 5 sentences) the research hypothesis you are going to study.
- Exploratory Data Analysis. Describe the data set: number of observations and covariates, their nature (quantitative, qualitative,etc.). Use visual plots for your initial investigation to examine the response variable and the predictor variables. Make sure to clearly state the model you are going to study based on this EDA. Note that you may (or not) at this stage want to take transformations of the predictor variables.
- Run the least squares estimator for the initial model you study. Run your diagnostics, detect potential outliers, high leverage or influential points. Discuss whether to keep or discard them from the study.
- You may want to perform some transformations of the predictors to improve the fit of the data or include interactions terms like the size of lot times size of house, or age of house times Adjacency to Highway, etc. Explain whether and why you decide to keep or not the new variables. You can for instance build a statistical test to decide whether the interaction term has an impact on the response variable, etc.
- We may want to study some hypothesis such as
  - Older houses tend to have lower prices.

continued on page 2

  - House with higher bathroom/bedroom ratio should have higher price.
  - School quality impacts the price positively
  - What is the trend of a property value?

  Feel free to formulate your own hypothesis that you will analyze with statistical tools. Originality will be awarded (if supported by a coherent and reasonable analysis).

- Run a statistical analysis of your final model and deduce which predictors have an impact on house price.
- Add plots, tables, formulas as you see fit to build a clear and convincing report.