Name: Tam Nguyen

Date: October 18, 2018

Final Project Proposal

1. **Identify the data source. Include a brief summary of how, from whom, and by whom, the data were collected.**

* The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for data science education. It's a great alternative to a modernized and expanded version of the often cited Boston Housing dataset for data scientists to analyze.
* The Ames Housing dataset describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of categories variables (23 nominal, 23 ordinals, 14 discrete, and 20 continuous) involved in home value assessments.

1. Identify the **important research question(s**) which will guide your project (e.g. “Do youth who participate in physical exercise class have lower BMI?”,  “Are males more likely to drink and drive after adjusting for confounding variables?”) and **describe why your chosen project is interesting to you**.

* What are the most important factors to predict housing prices?
* Is there any surprising factor that doesn’t have any linear relationship with housing prices? For example, housing age can possibly have a linear relationship with housing prices: the older the house is, the more expensive it will be and vice versa.
* Can a multiple linear regression model predict precisely the housing price?

1. **Provide a list of variables of interest and their definitions (including units).**

* Since there are almost 80 factors in this dataset, I will provide factors that I believe will be significant in my model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable name** | **Original definition** | **Units** | **Range or Levels** | **Possible recoding** | **Rationale** |
| SalePrice | The Housing price | $ | 3400 - 755000 |  | Response variable |
| MSSubClass | Identifies the type of dwelling involved in the sale. |  | 20 – 190, interger | Currently numeric var. Possibly remain unchanged | Possible confounding var |
| MSZoning | Identifies the general zoning classification of the sale. |  |  | Currently categorical var. Recode to same values but numeric | Possible confounding var |
| LotArea | Lot size in square feet | Sq/ft | 1300 -215245 |  | Main explanatory var of interest |
| Neighborhood | Physical locations within Ames city limits |  | 25 | Currently categorical var. Recode to same values but numeric | Main explanatory var of interest |
| Condition1 | Proximity to various conditions |  |  | Currently categorical var. Recode to same values but numeric | Main explanatory var of interest |
| OverallQual | Rates the overall material and finish of the house |  | 1-10 |  | Possible confounding var |
| OverallCond | Rates the overall condition of the house |  | 1-10 |  | Possible confounding var |
| YearBuilt | Original construction date | year | 1972 - 2010 |  | Possible confounding var |
|  |  |  |  |  |  |
| SaleCondition: | Condition of sale |  | 1-4 |  | Possible confounding var |
| Garage | Garage location |  | 1-7 | Currently categorical var. Recode to binomial value (Have garage or not) | Possible confounding var |
| Utilities | Type of utilities available |  | 1-4 | Currently categorical var. Recode to same values but numeric | Possible confounding var |
| SaleType | Type of sale |  | 1-10 | Currently categorical var. Recode to same values but numeric | Possible confounding var |
| SaleCondition | Condition of sale |  | 1-7 | Currently categorical var. Recode to same values but numeric | Possible confounding var |

1. **Find references for at least two articles**
   1. De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” *Journal of Statistics Education* 19, no. 3 (November 2011). <https://doi.org/10.1080/10691898.2011.11889627>.

I used this resource to understand the origin of dataset, and learn from author’s previous approaches to the dataset.

* 1. Kuhn, Max, and Kjell Johnson. “Data Pre-Processing.” In *Applied Predictive Modeling*, edited by Max Kuhn and Kjell Johnson, 27–59. New York, NY: Springer New York, 2013. <https://doi.org/10.1007/978-1-4614-6849-3_3>.

This dataset has many missing values, which means I need to choose which is the best way to pre-processing them. This book will give me an idea how to solve this problem.

* 1. “Information About Factors That Determine Property Prices - HomeGuru.” Accessed October 18, 2018. <http://www.homeguru.com.au/house-prices/>.

This non-specialist paper mentions what are the factors that can predict housing price. I can gather information from this paper to modify factors in my future model.

1. Briefly **outline how you plan to address your research question(s) with your data**

* Firstly, I will run the correlation matrix to see the interaction between 23 nominal, 23 ordinals, 14 discrete, and 20 continuous to the housing sale prices.
* Then, I plan to build multiple linear regression models with housing sale prices as response and different groups of variables (low variances or high variances) as explanatory variables.