### **Introduction:**

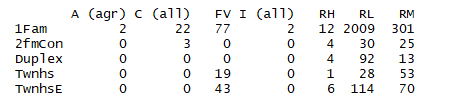
In Assignment #3, we will build off of Assignment #2 and develop predictive models. We will study models using categorical variables, continuous variables and a combination of both. We will also look at the effect of dummy variables and log transformations.

### **Section 1: Sample Definition**

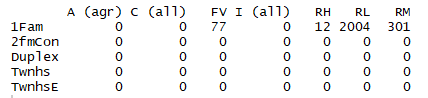
The Ames Housing dataset is comprised of 82 columns including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. There are also 2 columns (SID and PID) used as observation identifiers. These are not relevant for the purposes of this assignment and were removed. There were also 5 additional variables created. This gives a total of 85 variables used for 2930 residential properties.

It was observed that multiple building types were included in the original data set. It does not make sense to compare one building type to another as each type has a different perception of value in the eyes of a buyer and this is reflected in the different zoning types associated with each building.

The original dataset consists of 5 building types and 8 zoning classifications :



As the purpose of this assignment is to predict the sale price of a “typical” home, it does not make sense to include buildings falling under the zoning classifications industrial (I), agriculture (A) or commercial (C). Residential zoning types including residential high density (RH), residential low density (RL), residential low density park (RP), residential medium density (RM) and floating village residential will be used for further analysis. For building type, we are looking to analyze houses that represent the typical “house” that a family would buy. Duplexes (Duplx) and townhouses (TwnhsE and TwnhsI) are not good representations and are removed leaving single family detached (1Fam) and two family conversion houses (2FmCon) left for analysis. After these conditions are met, there are 2399 observations remaining. Here is a table showing the distribution of homes in the four zoning classifications by building type.



It is evident that none of the properties fall under the two family conversion building type so this can be removed from consideration. While investigating the data further, it was also observed that the majority of properties ranged from 250 total sq ft. to 4000 sq ft. There were 5 properties with over 4000 total sq ft. which were identified as outliers and removed from the sample. This leaves us with a dataset of 85 variables and 2395 properties.

### **Section 2: Predictive Modeling Framework**

### **Section 3: Model Identification by Automated Variable Selection**

### **Section 4: Predictive Accuracy**

### **Section 5: Operational Validation**

### **Section 6: Final Model**

### **Section 7: Reflection/Conclusion**