STAT 448

Final Project Group 5

Professor Darren Glosemeyer

Project Proposal

Group Members: Siddharth Ahuja, Ari Ben-Zeev, Rongqi Pan, Sooeun Kim

Title of the project: Ames, Iowa Housing Dataset Analysis

Basic description of data: The dataset we are using is sourced from: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data. This data describes the features of several houses. There are 83 variables (79 explanatory variables) and 2930 observations. The variables each provide some description about the house Each observation represents one house. Some of the variables in the dataset include:

- SalePrice - the property's sale price in dollars.
- LotShape: What is the general shape of property
- GrLivArea: Above grade (ground) living area square feet
- CentralAir: Is there central air conditioning?

Description of the questions each member intends to answer and analyses and technique(s) they intend to use

- Question 1: Which variables are the strongest predictors for above grade (ground) living area square feet (GrLivArea, continuous variable)? Since GrLivArea equals the sum of first and second floor square feet, I will leave out 1stFlrSF and 2ndFlrSF variables when running tests. This could be done by using linear or generalized linear model. - Sooeun Kim
- Question 2: Which variables are the strongest predictors for the sale price of the house (continuous variable)? (This will be done with a generalized linear model) - Ari Ben-Zeev
- Question 3: Which variables are the strongest predictors for whether or not there is central air in the house (binary variable) (this will be done using Logistic Regression and Discriminant analysis)? - Siddharth Ahuja
- Question 4: Which variables has the strongest relationship with general shape of the house(variable: Lotshape)(categorical and ordinal variable)(done using contingency table for descriptive analysis and using discriminant analysis with all the continuous variables as predictors to classify lotshape) –Rongqi Pan