**HOUSING PRICE: USING ADVANCED REGRESSION TECHNIQUES**

**PRESENTED BY JIANWEN WU, NITASHA TULI AND THOMAS KWOK**

**M.A STATISTICS, HUNTER COLLEGE**

**Table of contents**

**1. Introduction**

**2. Data Analysis**

1. **Data Analysis**
2. **Data Integration**
3. **Variable Description**

**3. Model I – Regression Analysis**

1. **Variable Selection** 
   1. **Correlation Matrix**
   2. **VIF**
2. **Model output**
3. **Model Diagnostic**
4. **Model Prediction**
5. **Model Forecasting**

**4. Discussion**

**INTRODUCTION:**

Housing has always been seen as one of the major stepping stones in adulthood; people graduate college, get a job, start a family, and save enough money to buy a house. The question though has always been how much does it cost to buy a house and what factors into the price of a house. In this project, we looked into dataset from 1460 houses bought in Ames Iowa to create a regression model that best predicts the price of a house in Ames, Iowa. The variable of interest that we study is Sales Price and in our dataset we have eighty predictor variables that are utilized to predict the cost of a house. Of the eighty predictors, we have 23 nominal, 23 ordinal, 14 discrete, and 20 continuous and these range from everything from Total Plot Area to Fire Place and Pool. Our interest in this project is to select the most important variables to study and create a model that best predicts the Sales Price of a house. In evaluating our model, we will look into the bias of our model to the actual sales price, the maximal deviation, mean iabsolute deviation, and mean square error to conclude how accurately we created a model to predict the Sales Price. We will use 1060 observations in our training data set and compare our model to the 400 observations in the test data set to see how our modeled Sales Price compares to the actual Sales Price of the 400 observations.

From our dataset, we will look into normalizing the Sales Price by using the log function, and we will convert some of our predictors into dummy variables. From the list of variables, we will focus on several that we believe are highly correlated with SalesPrice, and then variable selection techniques also to see the maximum number of variables for predicting a highly accurate model without overfitting the data. The purpose is to create a model that can best predict the Sales Price from our validation data set, but also that could be used on other data sets and still get a good prediction.

**DATA PREPROCESSING**

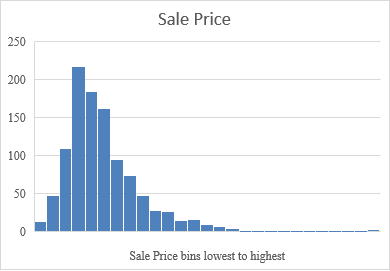
Before we start to apply our regression model to the data set, it is essential for us to have a clean and tide training data, the better training data we can get, the better performance of our regression model could behave. In data preprocessing we are mostly detecting and dealing with missing values, handling outliers and replacing the values with the median or most frequent value Treatment given to each variable is described below.

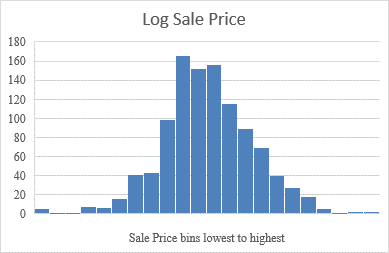
* + Missing values in variables like PoolQC, MiscFeature, Alley, Fence, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, MasVnrType, MSSubClass are replaced by None based on data description.
  + Missing values in LotFrontage is replace by median of the data series.
  + Missing values in GarageYrBlt, GarageArea, GarageCars, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, MasVnrArea are replaced with zero.
  + Electrical which denotes type of electrical sytem is replace by “SBrkr” because it was most frequent data value in that series.
  + Rest all categorical variables are converted into factors.

**DATA EXPLORATION**

After data cleaning and transformation we began with some exploratory analysis. A histogram plot shows the distribution of the target variable ‘SalePrice’ as being was right-skewed. So we decided to take log of sale price in a way to obtain normal distribution for sale price.

Transformation of Sale Price





We also noticed that most of the variables are positively skewed so as the sale price like 1stfloor surface area, second floor surface area, BedroomabvGr (bedrooms above basement level), LotFrontage, OpenporchSF etc.

While data exploration we noticed that some features had outliers like sale price was very huge for medium garage area and above the ground living area.

We have removed those outliers from the data.

The heat map below shows the correlation between each numeric variable with the Sale price.

