**The Hong Kong Polytechnic University**

**Second Semester 2020-2021**

**COMP 1433: Introduction to Data Analytics**

**Final Project**

Data Analysis for House Prices (Option 2)

|  |  |
| --- | --- |
| Group 15 |  |
| 19034824D | Lee Ka Shue |
| 20048925D | Wong Yau Chuen |
| 20037676D | Leung Chun Kit |
| 20084268D | Tang Yalun |

Contents

[Motivation 3](#_Toc71928022)

[Description 3](#_Toc71928023)

[Library 3](#_Toc71928024)

[Regression models 3](#_Toc71928025)

[Linear Regression 3](#_Toc71928026)

[Random Forest Regression 4](#_Toc71928027)

[Implementation 4](#_Toc71928028)

[Data imputation 4](#_Toc71928029)

[Missing values in pool data 4](#_Toc71928030)

[Missing values in garage data 4](#_Toc71928031)

[Missing values in basement data: 5](#_Toc71928032)

[Missing values in Masonry data: 6](#_Toc71928033)

[Other missing values 6](#_Toc71928034)

[Feature engineering 6](#_Toc71928035)

[YrHeld 6](#_Toc71928036)

[Regression 6](#_Toc71928037)

[Data classification 6](#_Toc71928038)

[Key data selection 7](#_Toc71928039)

[Data 8](#_Toc71928040)

[Key Stats 8](#_Toc71928041)

[Overview 8](#_Toc71928042)

[Missing values 8](#_Toc71928043)

[Results and Observation 9](#_Toc71928044)

[Kaggle results 9](#_Toc71928045)

[Observations 9](#_Toc71928046)

[SalePrice correlation 9](#_Toc71928047)

[Skewness of the dependant variable 10](#_Toc71928048)

[Discussions 10](#_Toc71928049)

# Motivation

Shelter is defined as one of human’s most basic and integral needs; However, in this day and age is accommodation more so considered a major financial consideration. If homes are merely for avoiding the elements during resting periods, surely a coffin-sized weather-proof box would suffice; Yet, it is not, quite apparently.

In this paper, we attempt to analyse the housing sales data for Ames, Iowa during the 2000’s to determine the contribution of various creature comforts and other data to the final sales pricing.

The resulting model can be utilized by potential buyers, sellers, and flippers to identify a lot’s general pricing to maximize their profit. The process may also allow us to identify historic and socio-economic factors in house pricing, as well as most importantly, train us in machine learning with the R language.

# Description

## Library

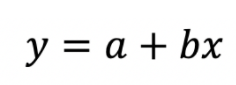
1. Auto directory setting: library(rstudioapi)
2. Data manipulation: library(plyr), library(dyplr)
3. Data visualization: library(corrplot), library(ggplot2), library(lares)
4. Data analysis: library(DataExplorer)
5. Model building: library(caret), library(randomForest)

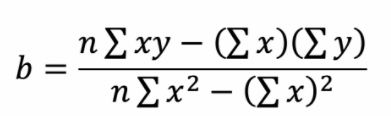
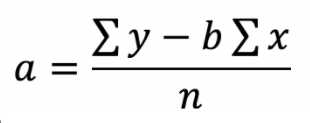
## Regression models

Regression analysis is a technique in statistics. The purpose is to understand the relationship between the dependent variable *(y)* and one or more independent variables *(x1, x2, …, xi)*. In application, given a set of independent variables, the conditional expected value could be estimated. There are several models of regression. Due to the page limit of the report, only two most common models will be discussed.

### Linear Regression

One of the most common models of regression is “Simple linear regression”. It is a way people understand the relationship between one independent variable and one dependent variable *(xi, yi)*. Using the method which is called “least squares method”, a straight line that closely fits the data will be created.

The general form of the line:

Formula of a and b: 

However, in reality, the outcome is probably not only depending on one independent variable, but also multiple variables. Therefore, “multiple linear regression” is needed. A line that fits the data will also be created.



The general form of the line:

Terminologies:

*y* – the value of the dependent variable

*Bn* – the coefficients of the independent variables

*Xn* – the independent variables

In this project, the dependent variable is the house prices of Ames, Lowa. The independent variables are the features, such as “LotArea” and “Street”.

### Random Forest Regression

Random Forest regression is an ensemble of randomly created Decision Trees. Each tree is created at a different position, selecting different samples of features. The forest merges the output of multiple Decision Trees, averages these predictions and generates the final output. This algorithm makes Random Forest Regression generally outperform single decision tree.

# Implementation

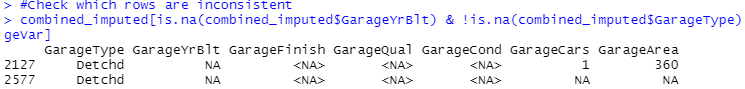
## Data imputation

### Missing values in pool data

Logically, “PoolArea” and “PoolQC” should not be mutually exclusive, since if a pool’s quality is rated, there should be a pool that exists (i.e. non-zero “PoolArea”) and vice versa. However, in row 2421, 2504, and 2600, while the pool exists, there is no “PoolQC” for describing the quality of the pool. To solve this discrepancy, a value based approximated by the overall quality of the house is assigned to the “PoolQC”.

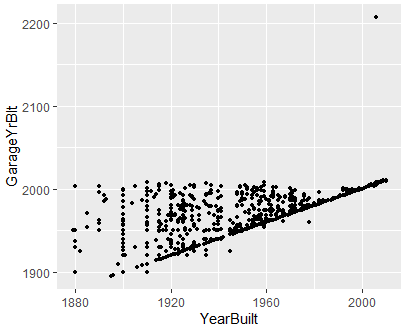
### Missing values in garage data

While most of the missing values of the features surrounding the garage is 159, “GarageType” only counts up to 157. We estimate that there are some inconsistencies in garage.



In row 2577, all variable related to garage are NA except GarageType. Therefore, we treat it as lacking a garage, and the “GarageType” value has been subbed in with “NA”.

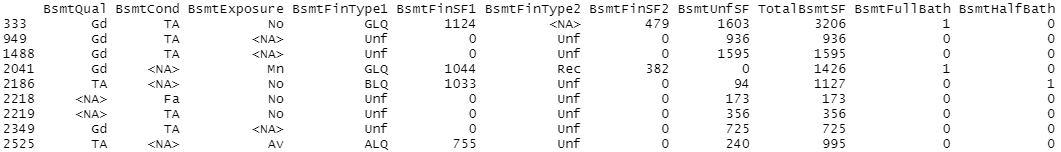
In row 2127, while the “GarageArea” is greater than zero (i.e., the garage exists), other features, like “GarageYrBlt”, are missing. To have a better understanding of the relationship between the built year of the house (YearBuilt) and its garage (GarageYrBlt), a scatter plot is created.



The apparent focused trend within the scatter plot reveals that most of the houses and its garage were built simultaneously. To be precise, we found that there are 2216 out of 2919 rows in the data where their YearBuilt and GarageYrBlt are equal.

Apart from that, a majority of the garages are built after the house is built (i.e., the dots above the line) while only a few of the garages are built before (i.e., the dots below the line). However, a strange dot has appeared on the top-right corner, claiming that the garage has been built after 2200. As this is most likely a typo, we decided to substitute the value with a more suitable value. As we have discovered that most of the garages are built at the same time with the house itself, that particular value from row 2127 was replaced with its “YearBuilt”, 2007. As to the other missing values, they are imputed with the mode of their respective features.

### Missing values in basement data:

Following the same train of though with that of the pool’s inconsistent record values, we find that some features are “NA”s while the “BsmtFinType” is populated with data. Therefore, those particular record’s “NA”s in “BsmtQual”, “BsmtCond”, “BsmtExposure”, and “BsmtFinType2” are replaced with the mode of the corresponding features as well.

Apart from that, we have also found that there are two rows where all the basement variables are “NA”s or 0s, so we similarly estimate that there are no basements.

### Missing values in Masonry data:

For the Masonry variables, row 2611 has “NA” in “MasVnrType”, but not in “MasVnrArea”. Since the value for “MasVnrArea” exists, we assume that there also is a value in “MasVnrType”. Therefore, we impute the “MasVnrType” with the mode of the corresponding feature apart from “None”, which is “BrkFace”.

### Other missing values

Beyond the aforementioned missing values, there are still other meaningful “NA”s; We transform them into processable data by following the same logic from before with reference to the “data\_description.txt”. To quickly list the remaining features without repeating ourselves, for the following columns, we replace “NA” with “None”: “PoolQC”, “MiscFeature”, “Alley”, “Fence”, “GarageType”, “GarageFinish”, “GarageQual”, “GarageCond”, “BsmtCond”, “BsmtQual”, “BsmtExposure”, “BsmtFinType1”, “BsmtFinType2” and “FireplaceQu”; While for the following columns, we replace “NA” with 0: “BsmtFullBath”, “BsmtHalfBath”, “MasVnrArea”, “BsmtUnfSF”, “TotalBsmtSF”, “GarageCars” and “GarageArea”.

And for the following columns where the “NA” should not represent “None” but instead should be merely an error, we replace “NA” with their respective modes of all data, being: “MSZoning”, “Functional”, “Exterior1st”, “Exterior2nd”, “Electrical”, “KitchenQual” and “SaleType”.

## Feature engineering

### YrHeld

As a representator of different year features, we add a new column “YrHeld” to the data set, bearing the age of the house, calculated by the decustion of “YearBuilt” from “YrSold”.

## Regression

### Data classification

Training data is divided into two categories: numeric features and categorical features. Numeric data which identify classification, “MSSubClass” for example, is sorted into the second category as factors. Character variables like quality or condition, which indicate rankings, is converted to numbers by introducing a function called rank2value.

文本

描述已自动生成

In total, there are 36 factor variables and 45 numeric variables including “Id” and “SalePrice”.

### Key data selection

日程表

描述已自动生成To increase the accuracy of linear regression, only the significant variables should be included.

In terms of numeric variables, a correlation coefficient plot is drawn.

Variables with correlation around ±0.6 are selected. It is important to note that “GarageCars” and “GarageArea”, “TotalBsmtSF” and “X1stFlrSF” are highly related, which may influence the predicted value. Hence only one of each is included in the formula.

Factor variables are treated in a different way. A linear model including all factors is built first. Then summary() function is applied to see which predictors are more significant.

At last, predicted prices are calculated based on two linear regression models. The numeric one has a higher coefficient due to its higher reliability.

# Data

## Key Stats

### Overview

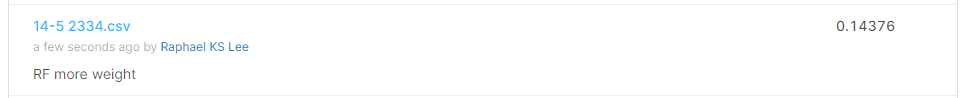
The raw training data holds discrete and continuous features with an almost 50/50 split. While missing only 5.9% of data values, 0% of individual records are fully populated.

### Missing values

There are 19 features in total that has any missing values. It can be seen that the major culprits causing the lack of complete records are namely “Fence”, “Alley”, “MiscFeature”, and “PoolQC”, which are being recommended to be dropped from any consideration. However, alongside with “FireplaceQu” and other “OK” or “Good” features, the record of “NA” may not necessarily mean a failure of data collection; instead, it can signify that a certain feature is not physically present in the household, which is processed accordingly above. Despite this discovery, there are still some NAs that we can determine as a failure in record keeping by identifying the inconsistencies within.

# Results and Observation

## Kaggle results



Our submission for the Kaggle competition results in 0.14376 RMSE.

## Observations

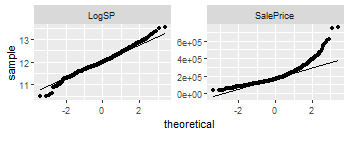
### SalePrice correlation

This chart displays 20 features with the most absolute correlation to the sale price.

Unsurprisingly, the overall quality of the house is the leading indicator of price, with above ground living area closely trailing behind.

However, it is unexpected that the year of remodelling (which defaults to the built year) is less reliable than built year itself, as it gives a fairer representation of the new-ness of the housing.

### Skewness of the dependant variable

As the scalar response, the SalePrice has a slight right skew, as evidenced by the QQ plot on the right. If this data is used without processing, the resultant model generated may have a bias towards a higher price. However, by performing a log transformation, we can see that the price mostly conform to a normal distribution.

# Discussions

In this project, we have encountered many difficulties. The first challenge is the seemingly unsurmountable number of independent variables. Among the 81 variables, both quantitative and qualitative attributes exist. Since we initially decide to implement linear regression, categorial variables would be a problem. Hence the dataset is divided into two class: numeric variables and categorical variables.

The second challenge, also the largest one, is imputation, which is clearly stated in Implementation part. Imputation is defined as the process of replacing missing data with substituted values. When processing data, we found that there are many NAs. Some refer to “None” of certain features, while others are man-made mistakes like omission. If we do not impute on them, there would be a sharp decrease on sample size, thus greatly affecting the accuracy. To replace NAs with meaningful data, we tried various methods. On the basis of understanding the meaning of some features, we can refer to other related variables for help. The value with highest frequency in relevant features will be filled in blank.

In terms of teamwork, our group has drawn up a specific plan at an early stage. With the help of github, we make clear explanation on individual work and maintain effective communication.