**Predictive Analysis of House Prices Using Regression Techniques**

Higher National Diploma in Software Engineering

Statistics for Computing Project Documentation

2023.3F

COHNDSE 23.3F – 018 Diluksha Malan

COHNDSE 23.3F – 020 Sohan Weerasinghe

COHNDSE 23.3F – 019 Oshadha Samarasinghe

COHNDSE 23.3F – 021 Avinda Udeshitha

A blue and black logo  Description automatically generated

School of Computing and Engineering

National Institute of Business Management

Colombo-7

Table of Contents

[1. Introduction 4](#_Toc170367002)

[Background of the Research 4](#_Toc170367003)

[Research Questions 4](#_Toc170367004)

[Objectives 4](#_Toc170367005)

[Variable Description 5](#_Toc170367006)

[2. Methodology 5](#_Toc170367007)

[Hypothesis 5](#_Toc170367008)

[Hypothesis 01 5](#_Toc170367009)

[Hypothesis 02 5](#_Toc170367010)

[Hypothesis 03 5](#_Toc170367011)

[Hypothesis 04 6](#_Toc170367012)

[Data Analysis Plan 6](#_Toc170367013)

[3. Data Analysis 6](#_Toc170367014)

[Exploratory Data Analysis (EDA) 6](#_Toc170367015)

[Dataset Description 6](#_Toc170367016)

[Data Structure 7](#_Toc170367017)

[Summary Statistics 8](#_Toc170367018)

[Data Visualization 9](#_Toc170367019)

[Scatter plot of Masonry veneer area vs Sale price 9](#_Toc170367020)

[Box Plot of Sale Price 10](#_Toc170367021)

[Scatter plot of Basement Area vs Sale price 11](#_Toc170367022)

[Scatter Plot of Year Built vs Sale price 12](#_Toc170367023)

[Scatter Plot of Living Area vs Sale Price 13](#_Toc170367024)

[Scatter plot of Actual vs Predicted Sale price 14](#_Toc170367025)

[Handling Outliers 15](#_Toc170367026)

[Scatter plot of Masonry veneer area vs Sale price 15](#_Toc170367027)

[Box Plot of Sale Price 16](#_Toc170367028)

[Scatter plot of Basement Area vs Sale price 17](#_Toc170367029)

[Scatter Plot of Year Built vs Sale price 18](#_Toc170367030)

[Scatter Plot of Living Area vs Sale Price 19](#_Toc170367031)

[Handling Missing Values 20](#_Toc170367032)

[Implementing Linear Regression Model 21](#_Toc170367033)

[Evaluation of Model Performance 22](#_Toc170367034)

[4. Discussion and Recommendations 23](#_Toc170367035)

[5. Conclusion 24](#_Toc170367036)

[6. References 26](#_Toc170367037)

[Predictive Analysis of House Prices Using Regression Techniques 27](#_Toc170367038)

# 1. Introduction

## Background of the Research

When considering real estate market and acknowledging the factors that affect on house prices is important for many kinds of stakeholders, buyers and homeowners. In general economy real estate industry is a important component. This research used a dataset from Kaggle Competition “House Prices: Advance Regression Techniques” to study the main factors that affect house prices.

## Research Questions

1. What is the impact of masonry veneer area and house prices, and is there a significant relationship between these variables?
2. Is there any relationship between the year built of the house and sale price?
3. What is the relationship between basement area and house prices, and to what extent basement area influence the sale price of the house?
4. What is the relationship between various features(ground living area, year built, total basement area, masonry veneer area, and garage year built) and house prices, and how this relationship be leveraged to develop predictive model for estimate house price?

## Objectives

1. Investigate the impact of masonry veneer area and house price to identify any relationship between two variables.
2. Examine the relationship between year built and the sale price of house, identifying trends and outliers, and assess the impact on sale price.
3. Examine the impact of basement area on house price, and provide reasonable insights for real estate stakeholders.
4. Develop multiple linear regression model

## Variable Description

* + YearBuilt: Original construction date
  + GroundLivArea : Ground living area a single house have
  + MasVnrArea: Masonry veneer area in square feet
  + TotalBsmtSF: Total square feet of basement area
  + SalePrice: Sold price
  + GarageYrBlt: Year garage was built

# 2. Methodology

## Hypothesis

### Hypothesis 01

H0 – There is no significant relationship between masonry veneer area and house prices.

H1 – There is significant positive relationship between masonry veneer area house prices.

### Hypothesis 02

H0 – The year built does not affect the house prices

H1 – The year built significantly affects house prices, with newer houses tending to have higher prices

### Hypothesis 03

H0- Theres no relationship between basement area and sale price.

H1- Theres a significant relationship between basement area and sale price

### Hypothesis 04

H0- There is no significant relationship between ground living area, year built, total basement area, masonry veneer area, garage year built, and house prices.

H1- There is a significant relationship between at least one of the predictor variables (ground living area, year built, total basement area, masonry veneer area garage year built) and house prices, indicate that these variables collectively influence house price in the multiple linear regression model.

## Data Analysis Plan

* Exploratory Data Analysis (EDA)
* Handling missing values
* Implementing linear Regression Model
* Evaluation of model performance

# 3. Data Analysis

This section involved several steps aim for understand the dataset, handling missing values, develop a predictive model, and evaluating own performance.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis used to identify the patterns and relationships inside the dataset.

### Dataset Description

The dataset is consist with 1460 observation and 81 variables. It include both numerical and categorical data regarding residential properties

### Data Structure

* Number of variables : 81
* Number of observations : 1460
* Types of variables : Numerical and Categorical

### Summary Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Mean | Median | Standard Deviation | Mode | 1QR | 3QR |
| Ground Living Area | 1515 | 1464 | 525.45 | 864 | 1130 | 1777 |
| Year Built | 1971 | 1973 | 30.202 | 2006 | 1954 | 2000 |
| Total Basement Area | 1057.4 | 991.5 | 438.7 | 0 | 795.8 | 1298.2 |
| Masonry Veneer Area | 103.7 | 0.0 | 181.06 | 0 | 0.0 | 166 |
| Garage Year Built | 1979 | 1980 | 24.689 | 2005 | 1961 | 2002 |

summary(data4)

### Data Visualization

#### Scatter plot of Masonry veneer area vs Sale price

A graph showing a line and a line

Description automatically generated with medium confidence

The blue line goes between the scatterplot shows the trend, indicating a positive relationship between two variables

Masonry veneer increases the sale price tends to increase.

A major group of data clustered towards the lower end of masonry veneer data axis, shows that most of the houses have small masonry veneer areas.

There are some data lies further away from the main group of data. These can be configured as outliers, shows that exceptionally larger areas tends to have high sale price.

This indicates the masonry veneer area relationship with sale price, this is not perfect linear relationship because of have significant amount of outliers.

ggplot(data4, aes(x = MasVnrArea, y = SalePrice)) +

geom\_point(color = "purple") +

geom\_smooth(method = "lm", color = "blue", se = FALSE) +

labs(title = "Scatter Plot of MasVnrArea vs Sale Price",x = "MasVnrArea",y = "Sale Price")

#### Box Plot of Sale Price

A graph with a line and a line

Description automatically generated

The box shows the Interquartile Range (IQR), containing the middle of the data.

The line inside the box is to represent the median sale price, that gives the central distribution, which 50% of the sale prices lie.

These extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles.

Dots outside the whiskers are considered as outliers, there are number of high value outliers that shows positive strew.

This box plot illustrate that how the house prices vary throughout the dataset.

ggplot(data4, aes(y = SalePrice)) +

geom\_boxplot(fill = "orange") +

labs(title = "Box Plot of Sale Price",

y = "Sale Price")

#### Scatter plot of Basement Area vs Sale price

A graph with red dots

Description automatically generated

This diagram shows that the relationship between Basement Area and the Sale price. The basement area is on the x- axis and the Sale price is on the y- axis.

There is a a positive correlation between basement area and the sale price.

The larger basement area means the higher sale price.

There is some changes in the trend but overall trend is clear.

This relationship suggest that total basement area is significantly affect the selling price of a house.

ggplot(data4, aes(x = TotalBsmtSF, y = SalePrice)) +

geom\_point(color = "red") +

labs(title = "Scatter Plot of Basement Area vs Sale Price",

x = "Total Basement Area",

y = "Sale Price")

#### Scatter Plot of Year Built vs Sale price

A graph showing a plot of a built vs sale price

Description automatically generated

There is a very noticeable tend that newly built houses have higher market value. This indicates that newly built houses sold more than old houses.

Overall tend shows newer house price is high. This means that not all the higher prices houses are newer houses, older houses can sell high prices when considering renovation, and location.

The plot indicate a many homes that build around mid 1900s which may be a period of rapid development.

Finally there are some outliers in this plot, which shows that houses with unique features which could be affect on selling price of the house.

ggplot(data4, aes(x = YearBuilt, y = SalePrice)) +

geom\_point(color = "green") +

labs(title = "Scatter Plot of Year Built vs Sale Price",

x = "Year Built", y = "Sale Price")

#### Scatter Plot of Living Area vs Sale Price

A graph showing a number of blue dots

Description automatically generated

There is a positive correlation between the living area and the sale price. If the living area increases the sale price also increases. Which conclude that larger houses sold for more value.

This appears to be a linear relationship, which is each unit of living area equals to consistence amount of sale price, but it may not work when higher values.

Most of the data illustrate that the most homes contains similar range of living areas and its sale price.

This plot also have some outliers, where the sale price does not follow the general flow. Because of the extremely large houses like mansions do not fit in the general flow of house prices which shows some outliers.

Living area is represented on log scale which helps to visualize homes with small living area.

(Codes included in the appendices)

#### Scatter plot of Actual vs Predicted Sale price

A graph with black dots

Description automatically generated

The blue diagonal dashed line represents where actual sale price would match the predicted sale price.

Dots that close to blue line indicate more accurate predictions of the sale price, while further points represents the errors in the prediction.

This show the cluster of dots around the center, suggest that many of the predictions are likely to be close to the actual value, but there is also noticeable deviation.

By calculating the Root Mean Squared Error (RMSE) it would be major help to quantitatively accessed this.

(Codes included in the appendices)

### Handling Outliers

#### Scatter plot of Masonry veneer area vs Sale price

Before Handle Outliers

A graph showing a line and a line

Description automatically generated with medium confidence

After Handle Outliers

A graph with purple dots and a line

Description automatically generated

#### Box Plot of Sale Price

Before Handle Outliers

A graph with a line and a line

Description automatically generated

After Handle Outliers

A graph with a line

Description automatically generated

#### Scatter plot of Basement Area vs Sale price

Before Handle Outliers

A graph with red dots

Description automatically generated

After Handle Outliers

A diagram of a basement area

Description automatically generated

#### Scatter Plot of Year Built vs Sale price

Before Handle Outliers

A graph showing a plot of a built vs sale price

Description automatically generated

After Handle Outliers

A chart showing a number of green dots

Description automatically generated

#### Scatter Plot of Living Area vs Sale Price

Before Handle Outliers

A graph showing a number of blue dots

Description automatically generated

After Handle Outliers

A graph showing a number of blue dots

Description automatically generated

### Handling Missing Values

* Identify Variables with missing values.

This diagram 1.1 is a main part of the initial data exploration process. It help to identify columns with missing values, which guide the next step in data purification and pre processing columns with the large number of missing values like FireplaceQu, PoolQu, Fence and MiscFeature will be required an extra focus. If these values are not making a major impact on the analysis, it can be removed from the dataset.

A group of words on a white background

Description automatically generated

Diagram 1.1

The diagram 1.2 shows the repair dataset, indicating through preprocessing actions to address the missing values. This ensure that dataset is ready for analysis to get solid outcomes.

A close-up of a white background

Description automatically generated

Diagram 1.2

The following diagram shows that how the above mentioned missing value repairing completed.

A screenshot of a computer code

Description automatically generated

Diagram 1.3

### Implementing Linear Regression Model

We have implemented the multiple linear regression model to predict house prices based on the identified key features.

The equation: -

**Sale Price= β0 ​+ β1​(Ground Living Area) + β2​(Year Built) + β3​(Total Basement Area) + β4​(Masonry Veneer Area) + β5​(Garage Year Built)**

**Model Parameters**

* β0: Intercept
* β1: Coefficient for Ground Living Area
* β2: Coefficient for Year Built
* β3: Coefficient for Total Basement Area
* β4: Coefficient for Masonry Veneer Area
* β5: Coefficient for Garage Year Built

### Evaluation of Model Performance

The outcome of the regression model was evaluated using two major indicators.

* Mean Squared Error (MSE):
  + This indicator shows the average of the squares of the mistakes, or the average squared difference between the noted actual results and the predicted result by the model.
  + The lower MSE shows that data is perfectly fitted to the model. The calculated MSE value is 0.1397524, shows that the relative squared error in this model is noticeably small.
* Root Mean Squared Error (RMSE):
  + This is the root of the MSE and shows the standard deviation of prediction errors. It is less difficult to understand because it is in similar units as the response parameter or the Sale Price.
  + The RMSE value for this model is 0.3738348, means that the prediction error is approximately 0.37 of the sale price of houses. This indicates that the prediction values are closer to actual sale price values.
* R – Squared value:
  + The R-squared value is the statistical measurement that shows the proportion of the variance of dependent variables (GrLivArea, YearBuilt, TotalBsmtSF, MasVnrArea, GarageYrBlt) in the model.
  + The value of the r-squared is 0.7047986, shows that approximately 70.48% of the differences of the sale price in the model is explained by the predictors. This shows that the predictors having strong relationship with the sale prices, indicate that model is fits well with the dataset.
* Scatter Plot of Actual and Predicted Sale Prices
  + The scatter plot of actual vs predicted sale prices shows that most of the predictions were close to the actual values, indicating the model was accurate. Dots that near the blue dashed line indicate accurate predictions. Dots that far away from the blue dashed line shows prediction mistakes, that regions where the model needs improvements.

# 4. Discussion and Recommendations

* Discussion
  + Feature Importance
    - The research proved that the ground living are, year built, total basement area, masonry veneer area, and garage year built are major predictors of house prices.
    - The positive correlation between basement area and sale price, as well as ground living area and sale price, indicates the importance of these features in real state market.
    - The regression model provided reasonable and accurate predictions, those predictions were close to the actual sale prices.
  + Model Robustness
    - Low value of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) indicating that the features are strong predictors of house prices and also model robustness.
    - The scatter plot of actual vs predicted sale price indicate a tight group around the diagonal line, indicating the model captures patterns in the dataset well.
  + Limitations
    - The research was limited to the features available in the dataset, other factors such as neighborhood quality, proximity to amenities, and economic conditions, were not stated.
    - The model only assume linear relationships between the variables, it may affect on when considering complex interactions.
* Recommendations
  + For buyers:
    - Focus on the properties with larger living basement areas because these factor affect mainly affect higher sale prices.
  + For sellers:
    - Highlight newer construction dates and well maintained masonry veneer areas to increase real state value
  + Future Research
    - Explore additional factors like energy efficiency, and smart technologies.
    - Investigate the affect of location specific factors such as neighborhood safety, education quality, also public transports on house prices.

# 5. Conclusion

This research tried to predict house prices using regression techniques by using various factors and their impact on real state prices. The study main aim was to identify the key factors of that affect mainly on house prices, relationships between factors that affect on house prices, and develop a predictive model for house prices.

1. The key factors that affect on house prices are mentioned below:
   1. Ground Living Area: The size of living space that available above the ground level has a major relationship with the house prices. (Larger living areas increase house prices)
   2. Year Built: The newer houses are likely to be more expensive than older ones.
   3. Total basement Area: Houses that have larger basement areas are likely to be more expensive than other ones. (Larger Basement Area increase house prices)
   4. Masonry Veneer Area: The availability and the size of the masonry veneer bring benefits to the house prices.
   5. Garage Year Built: The year that garage was built has great impact on the house prices, which indicate its condition.
2. Examine the relationships between features and house prices
   1. Larger Living areas and Basement areas: Both of these factors have lead to tend higher house prices
   2. Newer Construction: Newer houses and well masonry veneer areas are associated with house prices.
3. Create predictive model for house pricing

**Sale Price= β0 ​+ β1​(Ground Living Area) + β2​(Year Built) + β3​(Total Basement Area) + β4​(Masonry Veneer Area) + β5​(Garage Year Built)**

* 1. The model was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), indicates that the model predictions were close to the actual sale prices of the houses. This validates that the chosen features are main predictions of house prices.

1. Analyze patterns in house prices over different time periods
   1. Pattern over time:

The research of house prices over years shows that it stays overall steady market with some variance. Specific findings:

* + - * + **Steady Growth**: Considering overall, house prices showed a steady growth against the time which reflect on economic growth.
        + **Market Fluctuations**: Certain years experienced due to economic conditions, policy changes that affect the house prices.

This research successfully identified the key factors that impact house prices, developed a predictive model with significant accuracy, and provided insights into patterns of house prices over the time. These finds can be valuable for buyers, sellers in the real estate market to made their decisions.

# 6. References

* Kaggle. (n.d). House Prices – Advanced Regression Techniques. Retrieved from [Kaggle]

( <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>)

* R Programming 101. (2020) ‘**ggplot for plots and graphs. An introduction to data visualization using R programming’**.[Online video].Available at <https://www.youtube.com/watch?v=HPJn1CMvtmI&list=PLtL57Fdbwb_C6RS0JtBojTNOMVlgpeJkS>
* Stack Overflow. (n.d.) *Home page.* Retrieved 09 June 2024, form <https://stackoverflow.com/>

**Appendices**

# Predictive Analysis of House Prices Using Regression Techniques

Oshadha Kaveen Samarasinghe

2024-06-14

{r setup, include=FALSE} knitr::opts\_chunk$set(echo = TRUE)

library(ggplot2)

library(caTools)

data4=read.csv(file.choose(),header = TRUE)

str(data4)

summary(data4)

View(data4)

missing\_values <- colSums(is.na(data4))

missing\_values

mode\_garageFinish <- names(sort(table(data4$GarageFinish), decreasing = TRUE))[1]

data4$GarageFinish[is.na(data4$GarageFinish)] <- mode\_garageFinish

mode\_garageQual <- names(sort(table(data4$GarageQual), decreasing = TRUE))[1]

data4$GarageQual[is.na(data4$GarageQual)] <- mode\_garageQual

mode\_garageCond <- names(sort(table(data4$GarageCond), decreasing = TRUE))[1]

data4$GarageCond[is.na(data4$GarageCond)] <- mode\_garageCond

mode\_bsmtQual <- names(sort(table(data4$BsmtQual), decreasing = TRUE))[1]

data4$BsmtQual[is.na(data4$BsmtQual)] <- mode\_bsmtQual

mode\_bsmtCond <- names(sort(table(data4$BsmtCond), decreasing = TRUE))[1]

data4$BsmtCond[is.na(data4$BsmtCond)] <- mode\_bsmtCond

mode\_bsmtExposure <- names(sort(table(data4$BsmtExposure), decreasing = TRUE))[1]

data4$BsmtExposure[is.na(data4$BsmtExposure)] <- mode\_bsmtExposure

mode\_bsmtFinType1 <- names(sort(table(data4$BsmtFinType1), decreasing = TRUE))[1]

data4$BsmtFinType1[is.na(data4$BsmtFinType1)] <- mode\_bsmtFinType1

mode\_bsmtFinType2 <- names(sort(table(data4$BsmtFinType2), decreasing = TRUE))[1]

data4$BsmtFinType2[is.na(data4$BsmtFinType2)] <- mode\_bsmtFinType2

data4 <- data4[ , !(names(data4) %in% c("PoolQC", "MiscFeature", "Alley", "Fence", "FireplaceQu", "BsmtCond", "BsmtExposure"))]

data4$GarageYrBlt[is.na(data4$GarageYrBlt)] <- median(data4$GarageYrBlt, na.rm = TRUE)

mode\_electrical <- names(sort(table(data4$Electrical), decreasing = TRUE))[1]

data4$Electrical[is.na(data4$Electrical)] <- mode\_electricalsummary(data4)

ggplot(data4, aes(x = GrLivArea, y = SalePrice)) +

geom\_point(color = "blue") +

labs(title = "Scatter Plot of Living Area vs Sale Price",

x = "Ground Living Area",

y = "Sale Price")

ggplot(data4, aes(x = YearBuilt, y = SalePrice)) +

geom\_point(color = "green") +

labs(title = "Scatter Plot of Year Built vs Sale Price",

x = "Year Built",

y = "Sale Price")

ggplot(data4, aes(x = TotalBsmtSF, y = SalePrice)) +

geom\_point(color = "red") +

labs(title = "Scatter Plot of Basement Area vs Sale Price",

x = "Total Basement Area",

y = "Sale Price")

ggplot(data4, aes(y = SalePrice)) +

geom\_boxplot(fill = "orange") +

labs(title = "Box Plot of Sale Price",

y = "Sale Price")

ggplot(data4, aes(x = MasVnrArea, y = SalePrice)) +

geom\_point(color = "purple") +

geom\_smooth(method = "lm", color = "blue", se = FALSE) +

labs(title = "Scatter Plot of MasVnrArea vs Sale Price",

x = "MasVnrArea",

y = "Sale Price")

}

set.seed(123)

split <- sample.split(data4$SalePrice, SplitRatio = 0.8)

train\_data <- subset(data4, split == TRUE)

test\_data <- subset(data4, split == FALSE)

Multiple\_linear\_mod <- lm(SalePrice ~ GrLivArea + YearBuilt + TotalBsmtSF + MasVnrArea + GarageYrBlt, data = train\_data)

summary(Multiple\_linear\_mod)

predictions <- predict(Multiple\_linear\_mod, newdata = test\_data)

mse <- mean((test\_data$SalePrice - predictions)^2)

rmse <- sqrt(mse)

r\_squared <- summary(Multiple\_linear\_mod)$r.squared

cat("Mean Squared Error (MSE):", mse, "\n")

cat("Root Mean Squared Error (RMSE):", rmse, "\n")

cat("R-Squared:", r\_squared, "\n")

comparison <- data.frame(Actual = test\_data$SalePrice, Predicted = predictions)

print(comparison)

summary(comparison)

ggplot(comparison, aes(x = Actual, y = Predicted)) +

geom\_point() +

geom\_abline(intercept = 0, slope = 1, linetype = "dashed", color = "blue") +

labs(title = "Actual vs Predicted Sale Price",

x = "Actual Sale Price",

y = "Predicted Sale Price")