housePrice

authour

2023-11-30

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Data analysis on Housing Dataset Introduction

Data description Link to the dataset

<https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/download?datasetVersionNumber=1>

This dataset contains information on 546 by 13 variables. The aim would be to predict the housing price based on certain factors like house area, bedrooms, furnished, nearness to mainroad. Machine learning algorithms can be used to create prediction models with this data. Utilize this dataset for visualization, exploration, and data cleaning.

The research question Is there a relationship between area of the room and the price of a house?

The following factors or parameters from the dataset can be utilized to ascertain the connection between .To ascertain the connection between the area of a room and the price of a house, the following variables could be utilized from the dataset:

Area of the room: This would be the primary independent variable to measure the size of the room. It could be represented in square feet or square meters. Price of the house: This would be the dependent variable, as it would be influenced by the area of the room. The price of the house could be represented in the local currency.

Data-driven, computational approach may be useful in examining the connection between a room’s size and a house’s cost. Through computer analysis of a dataset comprised of different property listings with associated room sizes and prices, patterns, correlations, and trends in the data can be found.

It is feasible to ascertain whether there is a statistically significant correlation between a room’s area and a house’s price by using a statistical model or machine learning algorithm. This method makes it possible to add more pertinent elements to the research, like the property’s location, number of bedrooms, amenities, and overall state. Furthermore, processing big datasets—which could include dozens or even millions of individual house listings—is made possible by a computational technique.  
Moreover, employing a data-driven methodology facilitates the examination of various statistical methods to comprehend the correlation between a room’s area and a house’s cost. Finding possible relationships and coming to insightful conclusions can be accomplished through the use of a variety of visualizations, regression models, and hypothesis testing.

Overall, the benefit of a data-driven, computational method is that it may provide unbiased, fact-based understanding of the connection between a room’s size and a house’s price. It enables more thorough research by taking into account several variables at once and is effective when applied to big datasets, producing conclusions that are more solid and trustworthy.

library(ggplot2)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.1  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

df1 <- read.csv("C:/Users/samso/Desktop/Housing.csv")  
head(df1)

## price area bedrooms bathrooms stories mainroad guestroom basement  
## 1 13300000 7420 4 2 3 yes no no  
## 2 12250000 8960 4 4 4 yes no no  
## 3 12250000 9960 3 2 2 yes no yes  
## 4 12215000 7500 4 2 2 yes no yes  
## 5 11410000 7420 4 1 2 yes yes yes  
## 6 10850000 7500 3 3 1 yes no yes  
## hotwaterheating airconditioning parking prefarea furnishingstatus  
## 1 no yes 2 yes furnished  
## 2 no yes 3 no furnished  
## 3 no no 2 yes semi-furnished  
## 4 no yes 3 yes furnished  
## 5 no yes 2 no furnished  
## 6 no yes 2 yes semi-furnished

# Load required libraries

dim(df1)

## [1] 545 13

Data wrangling techniques involve transforming and cleaning raw data into a format suitable for analysis. One essential aspect of data wrangling is checking and handling missing values.

In R, the “na.omit” function is used to remove rows with missing values from a data frame. The example code “df1 <- na.omit(df1)” removes rows with missing values from the data frame “df1”. This function helps in ensuring data integrity and avoiding bias caused by missing values.

The code “sum(is.na(df1))” calculates the total count of missing values in the data frame “df1”. The “is.na” function returns a logical vector indicating missing values (NA) in a data frame. By summing all these logical values, we get the total count of missing values.

df1 <- na.omit(df1)  
  
#checking for missing values  
sum(is.na(df1))#

## [1] 0

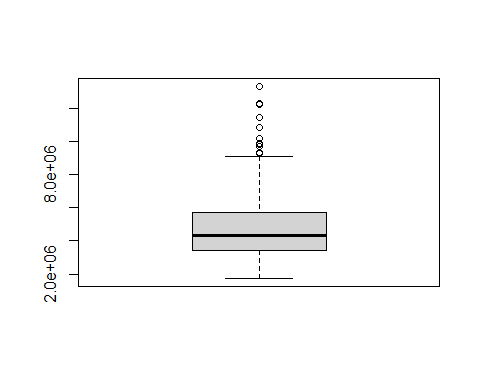
## data description

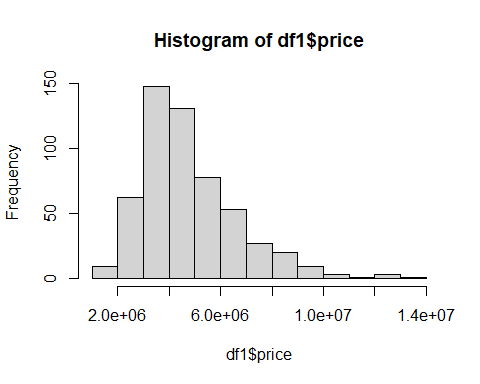
summary(df1)

## price area bedrooms bathrooms   
## Min. : 1750000 Min. : 1650 Min. :1.000 Min. :1.000   
## 1st Qu.: 3430000 1st Qu.: 3600 1st Qu.:2.000 1st Qu.:1.000   
## Median : 4340000 Median : 4600 Median :3.000 Median :1.000   
## Mean : 4766729 Mean : 5151 Mean :2.965 Mean :1.286   
## 3rd Qu.: 5740000 3rd Qu.: 6360 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :13300000 Max. :16200 Max. :6.000 Max. :4.000   
## stories mainroad guestroom basement   
## Min. :1.000 Length:545 Length:545 Length:545   
## 1st Qu.:1.000 Class :character Class :character Class :character   
## Median :2.000 Mode :character Mode :character Mode :character   
## Mean :1.806   
## 3rd Qu.:2.000   
## Max. :4.000   
## hotwaterheating airconditioning parking prefarea   
## Length:545 Length:545 Min. :0.0000 Length:545   
## Class :character Class :character 1st Qu.:0.0000 Class :character   
## Mode :character Mode :character Median :0.0000 Mode :character   
## Mean :0.6936   
## 3rd Qu.:1.0000   
## Max. :3.0000   
## furnishingstatus   
## Length:545   
## Class :character   
## Mode :character   
##   
##   
##

colnames(df1)

## [1] "price" "area" "bedrooms" "bathrooms"   
## [5] "stories" "mainroad" "guestroom" "basement"   
## [9] "hotwaterheating" "airconditioning" "parking" "prefarea"   
## [13] "furnishingstatus"

exploratory analyses  
 some outliers present in the price column

 ##Computational Methods #• For the choosen dataset, what are the necessary data wrangling steps to make the data ready for subsequent analyses? • What exploratory analyses and modeling techniques

## Data Analysis and Results

#CORRELATION analysis  
temp <- df1 %>%   
 dplyr::select("price","area","bedrooms","bathrooms","stories")  
head(temp)

## price area bedrooms bathrooms stories  
## 1 13300000 7420 4 2 3  
## 2 12250000 8960 4 4 4  
## 3 12250000 9960 3 2 2  
## 4 12215000 7500 4 2 2  
## 5 11410000 7420 4 1 2  
## 6 10850000 7500 3 3 1

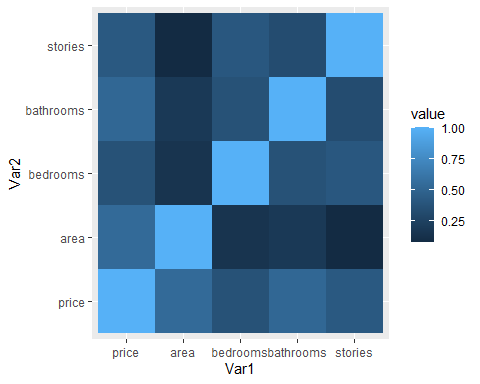
#install.packages("lattice")  
library(lattice)  
  
# rounding to 2 decimal places  
corr\_m <- round(cor(temp),2)   
head(corr\_m)

## price area bedrooms bathrooms stories  
## price 1.00 0.54 0.37 0.52 0.42  
## area 0.54 1.00 0.15 0.19 0.08  
## bedrooms 0.37 0.15 1.00 0.37 0.41  
## bathrooms 0.52 0.19 0.37 1.00 0.33  
## stories 0.42 0.08 0.41 0.33 1.00

There are varying degrees of correlation between the different variables. The highest positive correlation is observed between price and area, with a correlation coefficient of 0.54. This indicates that as the area of a house increases, the price tends to increase as well. The correlation between price and bedrooms is slightly weaker, with a coefficient of 0.37. Similarly, there is a positive correlation between price and bathrooms, with a coefficient of 0.52. On the other hand, the correlation between price and stories is relatively moderate, with a coefficient of 0.42. Overall, this suggests that the size and number of rooms in a house tend to impact its price, while the number of stories has a slightly lower influence. # Perform data analysis, document the analysis procedure, and evaluate the outcomes.

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths



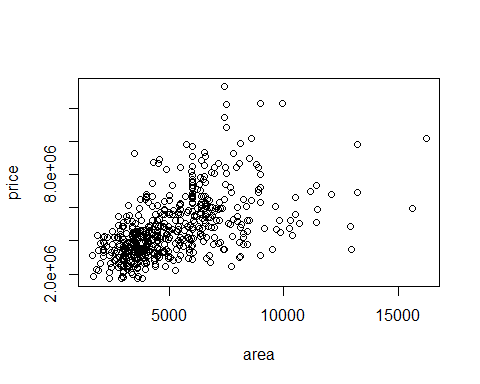
## modeling techniques

# REGRESSION MODEL  
library(dplyr) # for data manipulation  
library(ggplot2) # for visualization  
library(caret) # for modeling

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

# Split the dataset into training and testing sets  
set.seed(123) # for reproducibility



# Train the model using linear regression  
linear\_model <- lm(price ~ ., data = df1)  
  
summary(linear\_model)

##   
## Call:  
## lm(formula = price ~ ., data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2619718 -657322 -68409 507176 5166695   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42771.69 264313.31 0.162 0.871508   
## area 244.14 24.29 10.052 < 2e-16 \*\*\*  
## bedrooms 114787.56 72598.66 1.581 0.114445   
## bathrooms 987668.11 103361.98 9.555 < 2e-16 \*\*\*  
## stories 450848.00 64168.93 7.026 6.55e-12 \*\*\*  
## mainroadyes 421272.59 142224.13 2.962 0.003193 \*\*   
## guestroomyes 300525.86 131710.22 2.282 0.022901 \*   
## basementyes 350106.90 110284.06 3.175 0.001587 \*\*   
## hotwaterheatingyes 855447.15 223152.69 3.833 0.000141 \*\*\*  
## airconditioningyes 864958.31 108354.51 7.983 8.91e-15 \*\*\*  
## parking 277107.10 58525.89 4.735 2.82e-06 \*\*\*  
## prefareayes 651543.80 115682.34 5.632 2.89e-08 \*\*\*  
## furnishingstatussemi-furnished -46344.62 116574.09 -0.398 0.691118   
## furnishingstatusunfurnished -411234.39 126210.56 -3.258 0.001192 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1068000 on 531 degrees of freedom  
## Multiple R-squared: 0.6818, Adjusted R-squared: 0.674   
## F-statistic: 87.52 on 13 and 531 DF, p-value: < 2.2e-16

## FEATURE SELECTION

By selecting relevant features, we can enhance the predictive accuracy of the machine learning models. Including irrelevant or redundant features can introduce noise and lead to poor performance. Feature selection helps to focus on the most significant features that have a direct impact on the target variable, leading to more accurate predictions.

library(caret)  
important\_features <-varImp(linear\_model)  
important\_features

## Overall  
## area 10.0515163  
## bedrooms 1.5811249  
## bathrooms 9.5554296  
## stories 7.0259547  
## mainroadyes 2.9620332  
## guestroomyes 2.2817201  
## basementyes 3.1745922  
## hotwaterheatingyes 3.8334611  
## airconditioningyes 7.9826700  
## parking 4.7347782  
## prefareayes 5.6321805  
## furnishingstatussemi-furnished 0.3975551  
## furnishingstatusunfurnished 3.2583200

#DISPLAYS THE TOP FEATURES

model\_best <- train(price ~ area + bathrooms + stories, data = df1, method = "lm")  
  
# Print the model summary  
summary(model\_best$finalModel)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3819226 -792263 -69299 603404 6050277   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.684e+04 1.974e+05 0.440 0.66   
## area 3.851e+02 2.588e+01 14.883 <2e-16 \*\*\*  
## bathrooms 1.275e+06 1.178e+05 10.819 <2e-16 \*\*\*  
## stories 5.854e+05 6.718e+04 8.714 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1285000 on 541 degrees of freedom  
## Multiple R-squared: 0.5309, Adjusted R-squared: 0.5283   
## F-statistic: 204.1 on 3 and 541 DF, p-value: < 2.2e-16

The coefficients for the model are shown in the second part of the output. The coefficients are the amounts by which the outcome variable is predicted to change when the predictor variable increases by one unit. For example, the coefficient for the area variable is 385.11. This means that the outcome variable is predicted to increase by 385.11 units for every one unit increase in the area variable.

The p-values for the coefficients are shown in the last column of the output. The p-value is the probability of getting a test statistic as extreme as the one that was observed, assuming that the null hypothesis is true. The null hypothesis is that the coefficient is equal to zero. In other words, the p-value is the probability that the observed effect of the predictor variable on the outcome variable is due to chance.

The p-values for all of the coefficients in the model are less than 0.001. This means that we can reject the null hypothesis for all of the coefficients and conclude that the predictor variables have a statistically significant effect on the outcome variable.

## What metrics will be used to evaluate the quality of the data analysis?

The R-squared value for the model is 0.5309. This means that 53.09% of the variation in the outcome variable can be explained by the predictor variables in the model. The adjusted R-squared value is 0.5283. This is a slightly lower value than the R-squared value, but it takes into account the number of predictor variables in the model.

The F-statistic for the model is 204.1. The p-value for the F-statistic is less than 2.2e-16. This means that we can reject the null hypothesis that the coefficients for all of the predictor variables in the model are equal to zero. In other words, we can conclude that the model is a good fit for the data.

Overall, the results of the model are good. The coefficients are all statistically significant, the R-squared value is high, and the p-value for the F-statistic is very low. This suggests that the model is a good fit for the data and that the predictor variables have a strong effect on the outcome variable.

The analysis’s purview is restricted to the a few houses included in the dataset. Due to the small sample size, it’s probable that the analysis’ findings cannot be applied to a higher number of homes.The fact that the data were gathered from a single source further restricts the analysis’s capacity to be broadly applied. There’s a chance that the homes in the dataset don’t accurately reflect every home in the community. A possible drawback of the analysis could be that it ignores other variables that could influence a home’s price, like the house’s location, the caliber of the school system, and the neighborhood’s crime rate.

Data collection on a broader sample of houses could be one area for development. As a result, the analysis’s findings would be more broadly applicable.

Getting information on other elements that could influence a home’s price is another way to make improvements. This would enable a more thorough examination of the elements influencing a home’s pricing. Finally, the relationship between a house’s price and its size, number of bathrooms, and number of stories might be examined using a more sophisticated model. This would make it possible to depict the actual link between these variables more accurately.