100634271\_2.R

USER

2022-11-27

# ANALYSIS OF MELBOURNE DATASET  
  
# load the necessary libraries  
library(GGally)  
library(car)  
library(tidyverse)  
  
# Force R not to use scientific notations  
options(scipen = 99)  
  
# Load the Melbourne Housing dataset  
# This is the second data statistical analyses will be performed on.  
  
melbourne\_raw <- read.csv(  
 "C:\\Users\\USER\\Desktop\\100634271\\Melbourne\_housing.csv")  
  
str(melbourne\_raw)

## 'data.frame': 34857 obs. of 21 variables:  
## $ Suburb : chr "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford" ...  
## $ Address : chr "68 Studley St" "85 Turner St" "25 Bloomburg St" "18/659 Victoria St" ...  
## $ Rooms : int 2 2 2 3 3 3 4 4 2 2 ...  
## $ Type : chr "h" "h" "h" "u" ...  
## $ Price : int NA 1480000 1035000 NA 1465000 850000 1600000 NA NA NA ...  
## $ Method : chr "SS" "S" "S" "VB" ...  
## $ SellerG : chr "Jellis" "Biggin" "Biggin" "Rounds" ...  
## $ Date : chr "3/09/2016" "3/12/2016" "4/02/2016" "4/02/2016" ...  
## $ Distance : chr "2.5" "2.5" "2.5" "2.5" ...  
## $ Postcode : chr "3067" "3067" "3067" "3067" ...  
## $ Bedroom2 : int 2 2 2 3 3 3 3 3 4 3 ...  
## $ Bathroom : int 1 1 1 2 2 2 1 2 1 2 ...  
## $ Car : int 1 1 0 1 0 1 2 2 2 1 ...  
## $ Landsize : int 126 202 156 0 134 94 120 400 201 202 ...  
## $ BuildingArea : num NA NA 79 NA 150 NA 142 220 NA NA ...  
## $ YearBuilt : int NA NA 1900 NA 1900 NA 2014 2006 1900 1900 ...  
## $ CouncilArea : chr "Yarra City Council" "Yarra City Council" "Yarra City Council" "Yarra City Council" ...  
## $ Lattitude : num -37.8 -37.8 -37.8 -37.8 -37.8 ...  
## $ Longtitude : num 145 145 145 145 145 ...  
## $ Regionname : chr "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" ...  
## $ Propertycount: chr "4019" "4019" "4019" "4019" ...

# Dataset Description  
# The wine quality dataset consists of 21 variables and 34857 observations  
# The data consist of variables of different data type  
# Some of the variables are of number data type, while some others are of character data type.  
# The dependent variable is Price.  
# Some of the other variables are useful as independent variables for the statistical analysis   
# that will be performed in this section.  
  
# select the columns needed for analysis  
# Also rename the Bedroom2 column properly  
  
melbourne <- melbourne\_raw%>%  
 select(c(Rooms,Price,Distance,Bedroom2, Bathroom,Car,Landsize,  
 BuildingArea,Propertycount))%>%  
 rename("Bedroom" = "Bedroom2")  
  
# Check missing values in each column  
sum(is.na(melbourne$Rooms))

## [1] 0

sum(is.na(melbourne$Price))

## [1] 7610

sum(is.na(melbourne$Distance))

## [1] 0

sum(is.na(melbourne$Bedroom))

## [1] 8217

sum(is.na(melbourne$Bathroom))

## [1] 8226

sum(is.na(melbourne$Car))

## [1] 8728

sum(is.na(melbourne$Landsize))

## [1] 11810

sum(is.na(melbourne$BuildingArea))

## [1] 21115

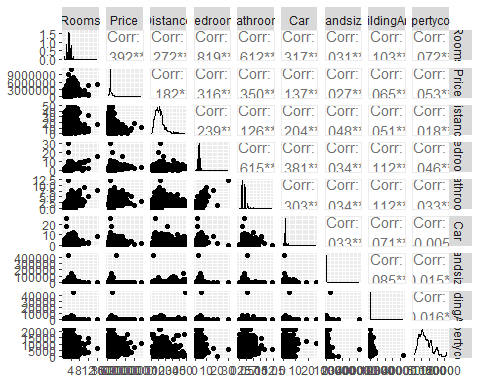
sum(is.na(melbourne$Propertycount))

## [1] 0

# There are missing values in some of the columns  
  
# Remove unwanted characters from the Distance and Propertycount columns  
# as this makes them variables of character data type when they are not.  
melbourne$Distance <- gsub("#N/A", "",as.character(melbourne$Distance))  
melbourne$Propertycount <- gsub("#N/A","",as.character(melbourne$Propertycount))  
  
# Convert the two columns back to numeric   
melbourne$Distance <- as.numeric(as.character(melbourne$Distance))  
melbourne$Propertycount <- as.numeric(as.character(melbourne$Propertycount))  
  
# Impute missing values with the median   
melbourne$Price[is.na(melbourne$Price)]<- median(melbourne$Price,na.rm = TRUE)  
melbourne$Distance[is.na(melbourne$Distance)]<- median(melbourne$Distance,na.rm = TRUE)  
melbourne$Bedroom[is.na(melbourne$Bedroom)]<- mean(melbourne$Bedroom,na.rm = TRUE)  
melbourne$Car[is.na(melbourne$Car)]<- median(melbourne$Car,na.rm = TRUE)  
melbourne$Landsize[is.na(melbourne$Landsize)]<- median(melbourne$Landsize,na.rm = TRUE)  
melbourne$BuildingArea[is.na(melbourne$BuildingArea)]<- median(melbourne$BuildingArea,na.rm = TRUE)  
melbourne$Propertycount[is.na(melbourne$Propertycount)]<- median(melbourne$Propertycount,na.rm = TRUE)  
  
# Check the structure of the new data frame.  
str(melbourne)

## 'data.frame': 34857 obs. of 9 variables:  
## $ Rooms : int 2 2 2 3 3 3 4 4 2 2 ...  
## $ Price : int 870000 1480000 1035000 870000 1465000 850000 1600000 870000 870000 870000 ...  
## $ Distance : num 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...  
## $ Bedroom : num 2 2 2 3 3 3 3 3 4 3 ...  
## $ Bathroom : int 1 1 1 2 2 2 1 2 1 2 ...  
## $ Car : int 1 1 0 1 0 1 2 2 2 1 ...  
## $ Landsize : int 126 202 156 0 134 94 120 400 201 202 ...  
## $ BuildingArea : num 136 136 79 136 150 136 142 220 136 136 ...  
## $ Propertycount: num 4019 4019 4019 4019 4019 ...

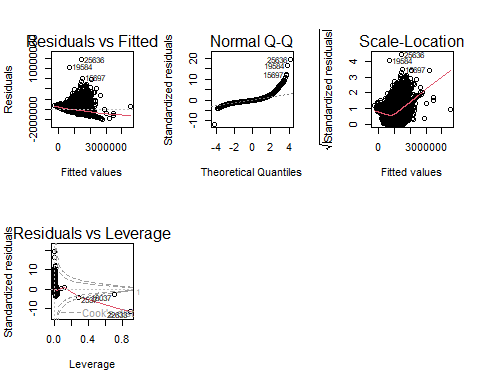
# There are no more missing values.  
  
  
# Pairplots of the variables  
# Thus will enable us see the distribution of the variables  
# as well as the correlations among them  
ggpairs(melbourne)



# Build the linear model  
melbourne\_model<- lm(Price ~ ., data = melbourne)  
summary(melbourne\_model)

##   
## Call:  
## lm(formula = Price ~ ., data = melbourne)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2046604 -300593 -91505 180831 9767680   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 466606.3352 12301.5788 37.931 < 0.0000000000000002 \*\*\*  
## Rooms 186282.7509 10065.1762 18.508 < 0.0000000000000002 \*\*\*  
## Distance -27907.4711 471.2114 -59.225 < 0.0000000000000002 \*\*\*  
## Bedroom 11988.5451 9866.7756 1.215 0.224   
## Bathroom 134694.2138 5466.3753 24.640 < 0.0000000000000002 \*\*\*  
## Car 38600.6742 3391.5735 11.381 < 0.0000000000000002 \*\*\*  
## Landsize 4.7233 0.9785 4.827 0.00000139 \*\*\*  
## BuildingArea 51.8883 10.7981 4.805 0.00000155 \*\*\*  
## Propertycount -2.7693 0.7107 -3.896 0.00009785 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 502300 on 26622 degrees of freedom  
## (8226 observations deleted due to missingness)  
## Multiple R-squared: 0.2602, Adjusted R-squared: 0.26   
## F-statistic: 1170 on 8 and 26622 DF, p-value: < 0.00000000000000022

# Examine the diagnostics plots to see that the assumptions are not violated.   
par(mfrow = c(2, 3))  
plot(melbourne\_model)  
par(mfrow = c(1, 1))



# MULTICOLINEARITY TEST  
# The vif function from the car package is used.  
  
vif(melbourne\_model)

## Rooms Distance Bedroom Bathroom Car   
## 9.850890 1.112333 9.881315 1.654258 1.218722   
## Landsize BuildingArea Propertycount   
## 1.010483 1.024355 1.003798

# the vif for each of the independent variable is less than 10.  
# Therefore, there is no serious collinearity among them.