

ModelosDeRegresionLineal

Cristopher Barrios, Carlos Daniel Estrada

2023-03-06

```
library(haven)
library(rpart)
library(stats)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
library(cluster)
library(rpart.plot)
library(fpc)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.2
```

```
## --
```

```
## v tibble  3.1.8      v purrr  1.0.1
```

```
## v tidyr   1.3.0      v stringr 1.5.0
```

```
## v readr   2.1.3      v forcats 1.0.0
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
```

```
## x purrr::lift()   masks caret::lift()
```

```
library(e1071)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dummy)
```

```
## dummy 0.1.3
## dummyNews()
```

```
#train <- read.csv("train.csv")
#test <- read.csv("test.csv")
```

1. Descargue los conjuntos de datos de la plataforma kaggle.

librerías

```
datos = read.csv("./train.csv")
test<- read.csv("./test.csv", stringsAsFactors = FALSE)

#Columnas
house <-select(datos, LotFrontage, LotArea, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1,BsmtFinSF2,

#Data
house <- na.omit(house)

# Resumen
summary(house)
```

```
##   LotFrontage      LotArea      YearBuilt      YearRemodAdd
##   Min.   : 21.00   Min.   : 1300   Min.   :1880   Min.   :1950
##   1st Qu.: 60.00   1st Qu.: 7590   1st Qu.:1953   1st Qu.:1966
##   Median : 70.00   Median : 9416   Median :1974   Median :1995
##   Mean   : 70.67   Mean   :10123   Mean   :1972   Mean   :1986
##   3rd Qu.: 80.00   3rd Qu.:11361   3rd Qu.:2003   3rd Qu.:2005
##   Max.   :313.00   Max.   :215245   Max.   :2010   Max.   :2010
##   MasVnrArea      BsmtFinSF1      BsmtFinSF2      BsmtUnfSF
##   Min.   : 0.0     Min.   : 0.0     Min.   : 0.00   Min.   : 0.0
##   1st Qu.: 0.0     1st Qu.: 0.0     1st Qu.: 0.00   1st Qu.: 250.0
##   Median : 0.0     Median : 374.0   Median : 0.00   Median : 506.0
##   Mean   :108.5     Mean   : 438.4   Mean   : 44.59   Mean   : 594.1
##   3rd Qu.:170.0     3rd Qu.: 702.0   3rd Qu.: 0.00   3rd Qu.: 840.0
##   Max.   :1600.0    Max.   :5644.0   Max.   :1474.00   Max.   :2336.0
##   TotalBsmtSF      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##   Min.   : 0       Min.   : 438     Min.   : 0.0     Min.   : 0.000
##   1st Qu.: 803     1st Qu.: 894     1st Qu.: 0.0     1st Qu.: 0.000
##   Median :1008     Median :1097     Median : 0.0     Median : 0.000
##   Mean   :1077     Mean   :1174     Mean   : 353.3   Mean   : 4.568
##   3rd Qu.:1324     3rd Qu.:1411     3rd Qu.: 728.0   3rd Qu.: 0.000
##   Max.   :6110     Max.   :4692     Max.   :2065.0   Max.   :572.000
```

```
##      GrLivArea      TotRmsAbvGrd      Fireplaces      GarageYrBlt
## Min.      : 438      Min.      : 3.000      Min.      :0.0000      Min.      :1900
## 1st Qu.:1155      1st Qu.: 5.000      1st Qu.:0.0000      1st Qu.:1959
## Median :1479      Median : 6.000      Median :1.0000      Median :1981
## Mean      :1531      Mean      : 6.576      Mean      :0.6039      Mean      :1978
## 3rd Qu.:1776      3rd Qu.: 7.000      3rd Qu.:1.0000      3rd Qu.:2003
## Max.      :5642      Max.      :12.000      Max.      :3.0000      Max.      :2010
##      GarageCars      GarageArea      WoodDeckSF      OpenPorchSF
## Min.      :1.000      Min.      : 160      Min.      : 0.00      Min.      : 0.00
## 1st Qu.:1.000      1st Qu.: 360      1st Qu.: 0.00      1st Qu.: 0.00
## Median :2.000      Median : 484      Median : 0.00      Median : 27.00
## Mean      :1.879      Mean      : 503      Mean      : 92.61      Mean      : 46.13
## 3rd Qu.:2.000      3rd Qu.: 600      3rd Qu.:168.00      3rd Qu.: 68.00
## Max.      :4.000      Max.      :1418      Max.      :857.00      Max.      :547.00
##      EnclosedPorch      ScreenPorch      PoolArea      MoSold
## Min.      : 0.00      Min.      : 0.0      Min.      : 0.000      Min.      : 1.00
## 1st Qu.: 0.00      1st Qu.: 0.0      1st Qu.: 0.000      1st Qu.: 5.00
## Median : 0.00      Median : 0.0      Median : 0.000      Median : 6.00
## Mean      : 21.84      Mean      : 16.1      Mean      : 2.935      Mean      : 6.34
## 3rd Qu.: 0.00      3rd Qu.: 0.0      3rd Qu.: 0.000      3rd Qu.: 8.00
## Max.      :552.00      Max.      :480.0      Max.      :648.000      Max.      :12.00
##      YrSold      SalePrice
## Min.      :2006      Min.      : 35311
## 1st Qu.:2007      1st Qu.:131000
## Median :2008      Median :164900
## Mean      :2008      Mean      :185506
## 3rd Qu.:2009      3rd Qu.:219500
## Max.      :2010      Max.      :755000
```

2. Haga un análisis exploratorio extenso de los datos. Explique bien todos los hallazgos. No ponga solo gráficas y código. Debe llegar a conclusiones interesantes para poder predecir. Explique el preprocesamiento que necesitó hacer.

Analisis exploratorio

—Exploración rápida de datos— train

```
summary(datos)
```

```
##      Id      MSSubClass      MSZoning      LotFrontage
## Min.      : 1.0      Min.      : 20.0      Length:1460      Min.      : 21.00
## 1st Qu.: 365.8      1st Qu.: 20.0      Class :character      1st Qu.: 59.00
## Median : 730.5      Median : 50.0      Mode  :character      Median : 69.00
## Mean      : 730.5      Mean      : 56.9                        Mean      : 70.05
## 3rd Qu.:1095.2      3rd Qu.: 70.0                        3rd Qu.: 80.00
## Max.      :1460.0      Max.      :190.0                        Max.      :313.00
##                                     NA's      :259
##      LotArea      Street      Alley      LotShape
## Min.      : 1300      Length:1460      Length:1460      Length:1460
## 1st Qu.: 7554      Class :character      Class :character      Class :character
## Median : 9478      Mode  :character      Mode  :character      Mode  :character
## Mean      : 10517
## 3rd Qu.: 11602
```

```

## Max.      :215245
##
## LandContour      Utilities      LotConfig      LandSlope
## Length:1460      Length:1460      Length:1460      Length:1460
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## Neighborhood      Condition1      Condition2      BldgType
## Length:1460      Length:1460      Length:1460      Length:1460
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## HouseStyle      OverallQual      OverallCond      YearBuilt
## Length:1460      Min.      : 1.000      Min.      :1.000      Min.      :1872
## Class :character  1st Qu.: 5.000      1st Qu.:5.000      1st Qu.:1954
## Mode  :character  Median : 6.000      Median :5.000      Median :1973
##                      Mean  : 6.099      Mean  :5.575      Mean  :1971
##                      3rd Qu.: 7.000      3rd Qu.:6.000      3rd Qu.:2000
##                      Max.   :10.000      Max.   :9.000      Max.   :2010
##
## YearRemodAdd      RoofStyle      RoofMatl      Exterior1st
## Min.      :1950      Length:1460      Length:1460      Length:1460
## 1st Qu. :1967      Class :character  Class :character  Class :character
## Median :1994      Mode  :character  Mode  :character  Mode  :character
## Mean    :1985
## 3rd Qu. :2004
## Max.    :2010
##
## Exterior2nd      MasVnrType      MasVnrArea      ExterQual
## Length:1460      Length:1460      Min.      : 0.0      Length:1460
## Class :character  Class :character  1st Qu. : 0.0      Class :character
## Mode  :character  Mode  :character  Median : 0.0      Mode  :character
##                      Mean  : 103.7
##                      3rd Qu.: 166.0
##                      Max.   :1600.0
##                      NA's   :8
## ExterCond      Foundation      BsmtQual      BsmtCond
## Length:1460      Length:1460      Length:1460      Length:1460
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## BsmtExposure      BsmtFinType1      BsmtFinSF1      BsmtFinType2
## Length:1460      Length:1460      Min.      : 0.0      Length:1460
## Class :character  Class :character  1st Qu. : 0.0      Class :character
## Mode  :character  Mode  :character  Median : 383.5      Mode  :character

```

```

##                               Mean    : 443.6
##                               3rd Qu.: 712.2
##                               Max.    :5644.0
##
##      BsmtFinSF2      BsmtUnfSF      TotalBsmtSF      Heating
##  Min.    : 0.00  Min.    : 0.0  Min.    : 0.0  Length:1460
## 1st Qu.: 0.00 1st Qu.: 223.0 1st Qu.: 795.8 Class :character
## Median : 0.00 Median : 477.5 Median : 991.5 Mode  :character
## Mean    : 46.55 Mean    : 567.2 Mean    :1057.4
## 3rd Qu.: 0.00 3rd Qu.: 808.0 3rd Qu.:1298.2
## Max.    :1474.00 Max.    :2336.0 Max.    :6110.0
##
##      HeatingQC      CentralAir      Electrical      X1stFlrSF
## Length:1460      Length:1460      Length:1460      Min.    : 334
## Class :character Class :character Class :character 1st Qu.: 882
## Mode  :character Mode  :character Mode  :character Median :1087
##                                     Mean    :1163
##                                     3rd Qu.:1391
##                                     Max.    :4692
##
##      X2ndFlrSF      LowQualFinSF      GrLivArea      BsmtFullBath
##  Min.    : 0  Min.    : 0.000  Min.    : 334  Min.    :0.0000
## 1st Qu.: 0 1st Qu.: 0.000 1st Qu.:1130 1st Qu.:0.0000
## Median : 0 Median : 0.000 Median :1464 Median :0.0000
## Mean    : 347 Mean    : 5.845 Mean    :1515 Mean    :0.4253
## 3rd Qu.: 728 3rd Qu.: 0.000 3rd Qu.:1777 3rd Qu.:1.0000
## Max.    :2065 Max.    :572.000 Max.    :5642 Max.    :3.0000
##
##      BsmtHalfBath      FullBath      HalfBath      BedroomAbvGr
##  Min.    :0.00000  Min.    :0.000  Min.    :0.0000  Min.    :0.000
## 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000
## Median :0.00000 Median :2.000 Median :0.0000 Median :3.000
## Mean    :0.05753 Mean    :1.565 Mean    :0.3829 Mean    :2.866
## 3rd Qu.:0.00000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000
## Max.    :2.00000 Max.    :3.000 Max.    :2.0000 Max.    :8.000
##
##      KitchenAbvGr      KitchenQual      TotRmsAbvGrd      Functional
##  Min.    :0.000  Length:1460  Min.    : 2.000  Length:1460
## 1st Qu.:1.000  Class :character 1st Qu.: 5.000  Class :character
## Median :1.000  Mode  :character Median : 6.000  Mode  :character
## Mean    :1.047  Mean    : 6.518
## 3rd Qu.:1.000  3rd Qu.: 7.000
## Max.    :3.000  Max.    :14.000
##
##      Fireplaces      FireplaceQu      GarageType      GarageYrBlt
##  Min.    :0.000  Length:1460  Length:1460  Min.    :1900
## 1st Qu.:0.000  Class :character  Class :character 1st Qu.:1961
## Median :1.000  Mode  :character  Mode  :character Median :1980
## Mean    :0.613  Mean    :1979
## 3rd Qu.:1.000  3rd Qu.:2002
## Max.    :3.000  Max.    :2010
##                                     NA's    :81
##      GarageFinish      GarageCars      GarageArea      GarageQual
## Length:1460      Min.    :0.000  Min.    : 0.0  Length:1460

```

```

## Class :character 1st Qu.:1.000 1st Qu.: 334.5 Class :character
## Mode :character Median :2.000 Median : 480.0 Mode :character
## Mean :1.767 Mean : 473.0
## 3rd Qu.:2.000 3rd Qu.: 576.0
## Max. :4.000 Max. :1418.0
##
## GarageCond PavedDrive WoodDeckSF OpenPorchSF
## Length:1460 Length:1460 Min. : 0.00 Min. : 0.00
## Class :character Class :character 1st Qu.: 0.00 1st Qu.: 0.00
## Mode :character Mode :character Median : 0.00 Median : 25.00
## Mean : 94.24 Mean : 46.66
## 3rd Qu.:168.00 3rd Qu.: 68.00
## Max. :857.00 Max. :547.00
##
## EnclosedPorch X3SsnPorch ScreenPorch PoolArea
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.000
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000
## Median : 0.00 Median : 0.00 Median : 0.00 Median : 0.000
## Mean : 21.95 Mean : 3.41 Mean : 15.06 Mean : 2.759
## 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.000
## Max. :552.00 Max. :508.00 Max. :480.00 Max. :738.000
##
## PoolQC Fence MiscFeature MiscVal
## Length:1460 Length:1460 Length:1460 Min. : 0.00
## Class :character Class :character Class :character 1st Qu.: 0.00
## Mode :character Mode :character Mode :character Median : 0.00
## Mean : 43.49
## 3rd Qu.: 0.00
## Max. :15500.00
##
## MoSold YrSold SaleType SaleCondition
## Min. : 1.000 Min. :2006 Length:1460 Length:1460
## 1st Qu.: 5.000 1st Qu.:2007 Class :character Class :character
## Median : 6.000 Median :2008 Mode :character Mode :character
## Mean : 6.322 Mean :2008
## 3rd Qu.: 8.000 3rd Qu.:2009
## Max. :12.000 Max. :2010
##
## SalePrice
## Min. : 34900
## 1st Qu.:129975
## Median :163000
## Mean :180921
## 3rd Qu.:214000
## Max. :755000
##

```

test

```
summary(test)
```

```

##      Id      MSSubClass      MSZoning      LotFrontage
## Min.   :1461   Min.     : 20.00   Length:1459   Min.     : 21.00

```

```

## 1st Qu.:1826    1st Qu.: 20.00    Class :character    1st Qu.: 58.00
## Median :2190    Median : 50.00    Mode  :character    Median : 67.00
## Mean   :2190    Mean   : 57.38                      Mean   : 68.58
## 3rd Qu.:2554    3rd Qu.: 70.00                      3rd Qu.: 80.00
## Max.   :2919    Max.   :190.00                      Max.   :200.00
##                                     NA's   :227
##      LotArea      Street      Alley      LotShape
## Min.   : 1470    Length:1459    Length:1459    Length:1459
## 1st Qu.: 7391    Class :character    Class :character    Class :character
## Median : 9399    Mode  :character    Mode  :character    Mode  :character
## Mean   : 9819
## 3rd Qu.:11518
## Max.   :56600
##
## LandContour      Utilities      LotConfig      LandSlope
## Length:1459      Length:1459      Length:1459      Length:1459
## Class :character    Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character    Mode  :character
##
##
##
## Neighborhood      Condition1      Condition2      BldgType
## Length:1459      Length:1459      Length:1459      Length:1459
## Class :character    Class :character    Class :character    Class :character
## Mode  :character    Mode  :character    Mode  :character    Mode  :character
##
##
##
## HouseStyle      OverallQual      OverallCond      YearBuilt
## Length:1459      Min.   : 1.000    Min.   :1.000    Min.   :1879
## Class :character    1st Qu.: 5.000    1st Qu.:5.000    1st Qu.:1953
## Mode  :character    Median : 6.000    Median :5.000    Median :1973
##                                     Mean   : 6.079    Mean   :5.554    Mean   :1971
##                                     3rd Qu.: 7.000    3rd Qu.:6.000    3rd Qu.:2001
##                                     Max.   :10.000    Max.   :9.000    Max.   :2010
##
## YearRemodAdd      RoofStyle      RoofMatl      Exterior1st
## Min.   :1950      Length:1459      Length:1459      Length:1459
## 1st Qu.:1963      Class :character    Class :character    Class :character
## Median :1992      Mode  :character    Mode  :character    Mode  :character
## Mean   :1984
## 3rd Qu.:2004
## Max.   :2010
##
## Exterior2nd      MasVnrType      MasVnrArea      ExterQual
## Length:1459      Length:1459      Min.   : 0.0    Length:1459
## Class :character    Class :character    1st Qu.: 0.0    Class :character
## Mode  :character    Mode  :character    Median : 0.0    Mode  :character
##                                     Mean   : 100.7
##                                     3rd Qu.: 164.0
##                                     Max.   :1290.0
##                                     NA's   :15

```

```

##      ExterCond      Foundation      BsmtQual      BsmtCond
## Length:1459      Length:1459      Length:1459      Length:1459
## Class :character      Class :character      Class :character      Class :character
## Mode  :character      Mode  :character      Mode  :character      Mode  :character
##
##
##
##      BsmtExposure      BsmtFinType1      BsmtFinSF1      BsmtFinType2
## Length:1459      Length:1459      Min.   : 0.0      Length:1459
## Class :character      Class :character      1st Qu.: 0.0      Class :character
## Mode  :character      Mode  :character      Median : 350.5      Mode  :character
##                                     Mean  : 439.2
##                                     3rd Qu.: 753.5
##                                     Max.   :4010.0
##                                     NA's    :1
##      BsmtFinSF2      BsmtUnfSF      TotalBsmtSF      Heating
## Min.   : 0.00      Min.   : 0.0      Min.   : 0      Length:1459
## 1st Qu.: 0.00      1st Qu.: 219.2      1st Qu.: 784      Class :character
## Median : 0.00      Median : 460.0      Median : 988      Mode  :character
## Mean   : 52.62      Mean   : 554.3      Mean   :1046
## 3rd Qu.: 0.00      3rd Qu.: 797.8      3rd Qu.:1305
## Max.   :1526.00      Max.   :2140.0      Max.   :5095
## NA's    :1      NA's    :1      NA's    :1
##      HeatingQC      CentralAir      Electrical      X1stFlrSF
## Length:1459      Length:1459      Length:1459      Min.   : 407.0
## Class :character      Class :character      Class :character      1st Qu.: 873.5
## Mode  :character      Mode  :character      Mode  :character      Median :1079.0
##                                     Mean   :1156.5
##                                     3rd Qu.:1382.5
##                                     Max.   :5095.0
##
##      X2ndFlrSF      LowQualFinSF      GrLivArea      BsmtFullBath
## Min.   : 0      Min.   : 0.000      Min.   : 407      Min.   :0.0000
## 1st Qu.: 0      1st Qu.: 0.000      1st Qu.:1118      1st Qu.:0.0000
## Median : 0      Median : 0.000      Median :1432      Median :0.0000
## Mean   : 326      Mean   : 3.543      Mean   :1486      Mean   :0.4345
## 3rd Qu.: 676      3rd Qu.: 0.000      3rd Qu.:1721      3rd Qu.:1.0000
## Max.   :1862      Max.   :1064.000      Max.   :5095      Max.   :3.0000
##                                     NA's    :2
##      BsmtHalfBath      FullBath      HalfBath      BedroomAbvGr
## Min.   :0.0000      Min.   :0.000      Min.   :0.0000      Min.   :0.000
## 1st Qu.:0.0000      1st Qu.:1.000      1st Qu.:0.0000      1st Qu.:2.000
## Median :0.0000      Median :2.000      Median :0.0000      Median :3.000
## Mean   :0.0652      Mean   :1.571      Mean   :0.3777      Mean   :2.854
## 3rd Qu.:0.0000      3rd Qu.:2.000      3rd Qu.:1.0000      3rd Qu.:3.000
## Max.   :2.0000      Max.   :4.000      Max.   :2.0000      Max.   :6.000
## NA's    :2
##      KitchenAbvGr      KitchenQual      TotRmsAbvGrd      Functional
## Min.   :0.000      Length:1459      Min.   : 3.000      Length:1459
## 1st Qu.:1.000      Class :character      1st Qu.: 5.000      Class :character
## Median :1.000      Mode  :character      Median : 6.000      Mode  :character
## Mean   :1.042                                     Mean   : 6.385
## 3rd Qu.:1.000                                     3rd Qu.: 7.000

```



```

NombreDeLaVariables <- c("SalePrice", "MSSubClass", "MSZoning", "LotFrontage", "LotArea", "Street", "Al
Cualitativa <- c("", "", "x", "", "", "x", "x", "x", "x", "x", "x", "x", "x", "x", "x", "x", "x", "", "

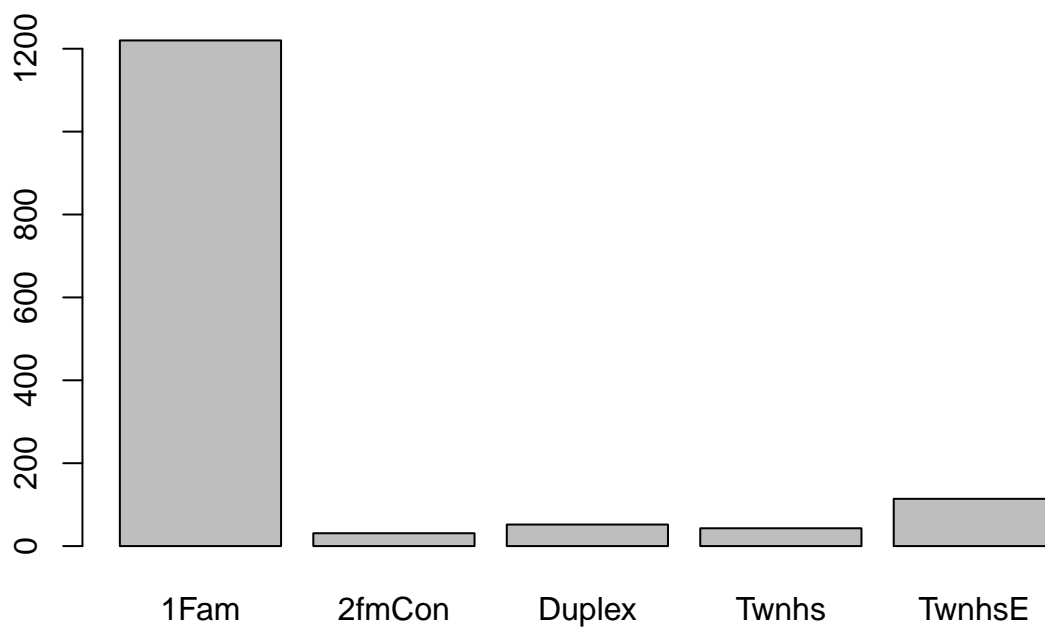
```

```
Cuantitativa <- c("x", "x", "", "x", "x", "", "", "", "", "", "", "", "", "", "", "", "x", "x", "x",
Discreta <- c("", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "", "",
Continua <- c("x", "x", "", "x", "x", "", "", "", "", "", "", "", "", "", "", "", "", "", "x", "x", "x", "x",
DataFrame.NombreDeLaVariables <- data.frame(NombreDeLaVariables, Cualitativa,Cuantitativa,Discreta,Cont.
print(DataFrame.NombreDeLaVariables)
```

## 42	CentralAir	x			
## 43	Electrical	x			
## 44	1stFlrSF		x		x
## 45	2ndFlrSF		x		x
## 46	LowQualFinSF		x		x
## 47	GrLivArea		x		x
## 48	BsmtFullBath		x	x	
## 49	BsmtHalfBath		x	x	
## 50	FullBath	f	xf	xf	f
## 51	HalfBath		x	x	
## 52	Bedroom		x	x	
## 53	Kitchen		x	x	
## 54	KitchenQual	x			
## 55	TotRmsAbvGrd		x	x	
## 56	Functional	x			
## 57	Fireplaces		x	x	
## 58	FireplaceQu	x			
## 59	GarageType	x			
## 60	GarageYrBlt		x		x
## 61	GarageFinish	x			
## 62	GarageCars		x	x	
## 63	GarageArea		x		x
## 64	GarageQual	x			
## 65	GarageCond	x			
## 66	PavedDrive	x			
## 67	WoodDeckSF		x		x
## 68	OpenPorchSF		x		x
## 69	EnclosedPorch		x		x
## 70	3SsnPorch		x		x
## 71	ScreenPorch		x		x
## 72	PoolArea		x		x
## 73	PoolQC	x			
## 74	Fence	x			
## 75	MiscFeature	x			
## 76	MiscVal		x		x
## 77	MoSold		x		x
## 78	YrSold		x		x
## 79	SaleType	x			
## 80	SaleCondition	x			

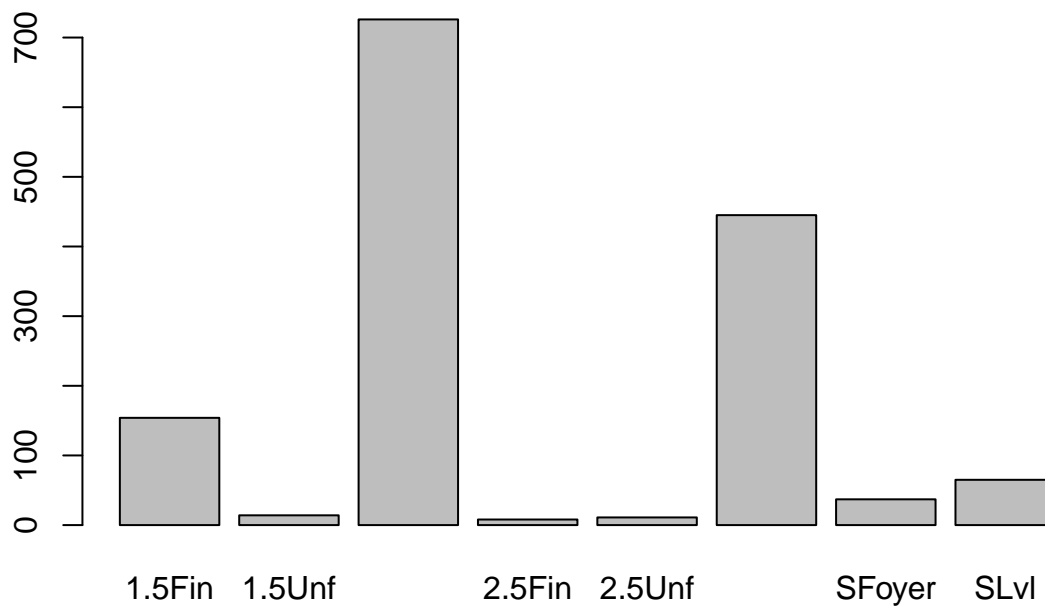
Estilo de vivienda

```
totalGenres <- unlist(strsplit(as.character(datos$BldgType), "\\|"))
barplot(table(totalGenres))
```



La mayoría de las casas son para una familia

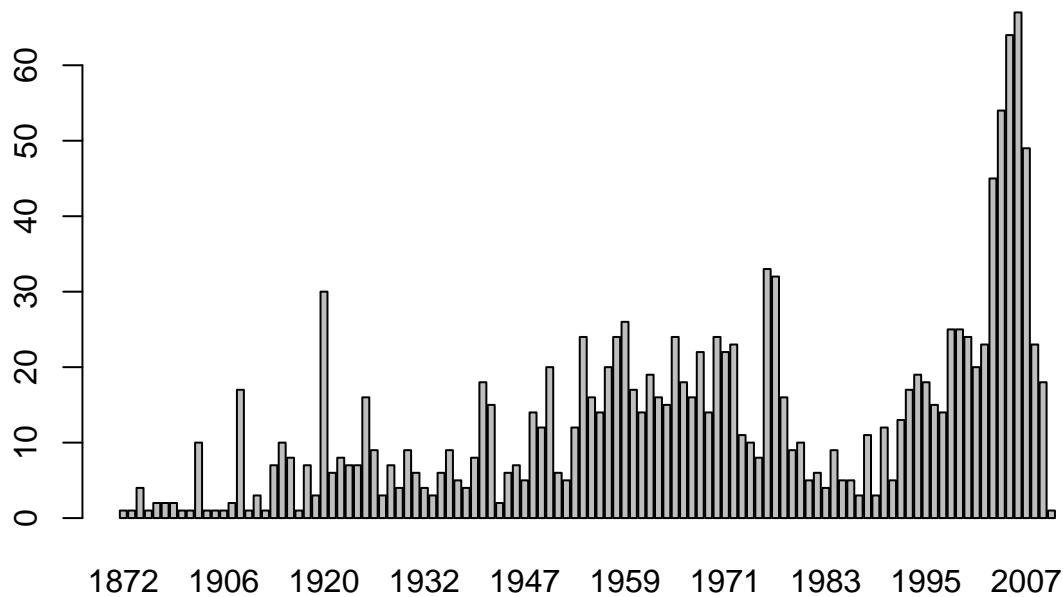
```
totalGenres <- unlist(strsplit(as.character(datos$HouseStyle), "\\|"))  
barplot(table(totalGenres))
```



El estilo más común de casa es de 1 solo nivel, por lo que podemos deducir que las casas de 2 niveles son más cotizadas

Año de construcción

```
totalGenres <- unlist(strsplit(as.character(datos$YearBuilt), "\\|"))
barplot(table(totalGenres))
```



La mayoría de las casas fueron construidas en los años 2000

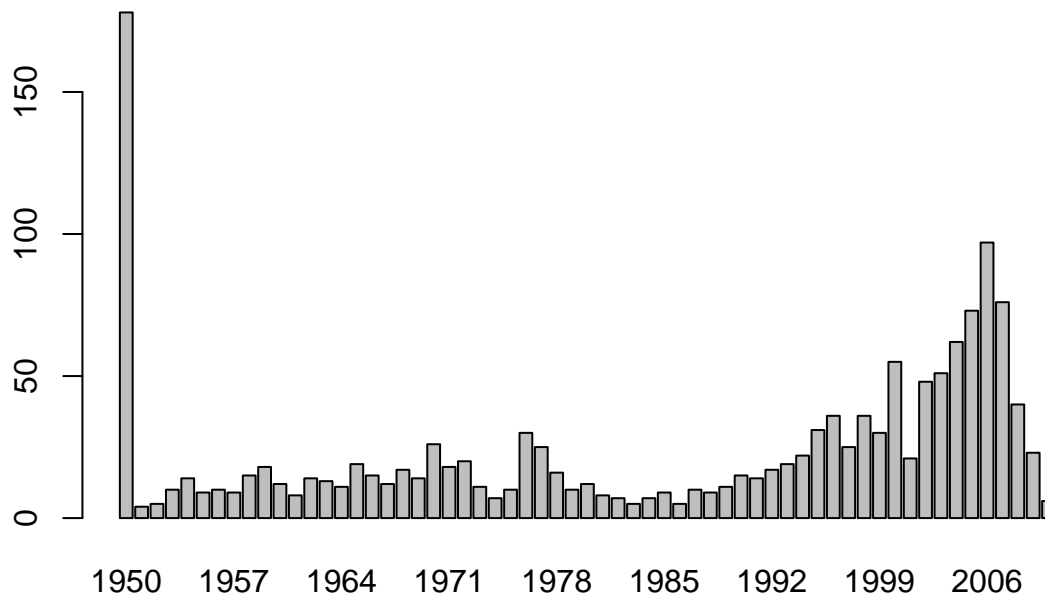
Casas mas nuevas:

```
newhouses <- head(order(datos$YearBuilt, decreasing = T), n=10)
datos[newhouses,c("Id","YearBuilt","SalePrice")]
```

```
##      Id YearBuilt SalePrice
## 379 379      2010    394432
##  88  88      2009    164500
## 104 104      2009    198900
## 158 158      2009    269500
## 212 212      2009    186000
## 213 213      2009    252678
## 413 413      2009    222000
## 461 461      2009    263435
## 508 508      2009    208300
## 516 516      2009    402861
```

Año de remodelación

```
totalGenres <- unlist(strsplit(as.character(datos$YearRemodAdd), "\\|"))
barplot(table(totalGenres))
```



La mayoría de las casas no han sido remodeladas desde la década de los 50s

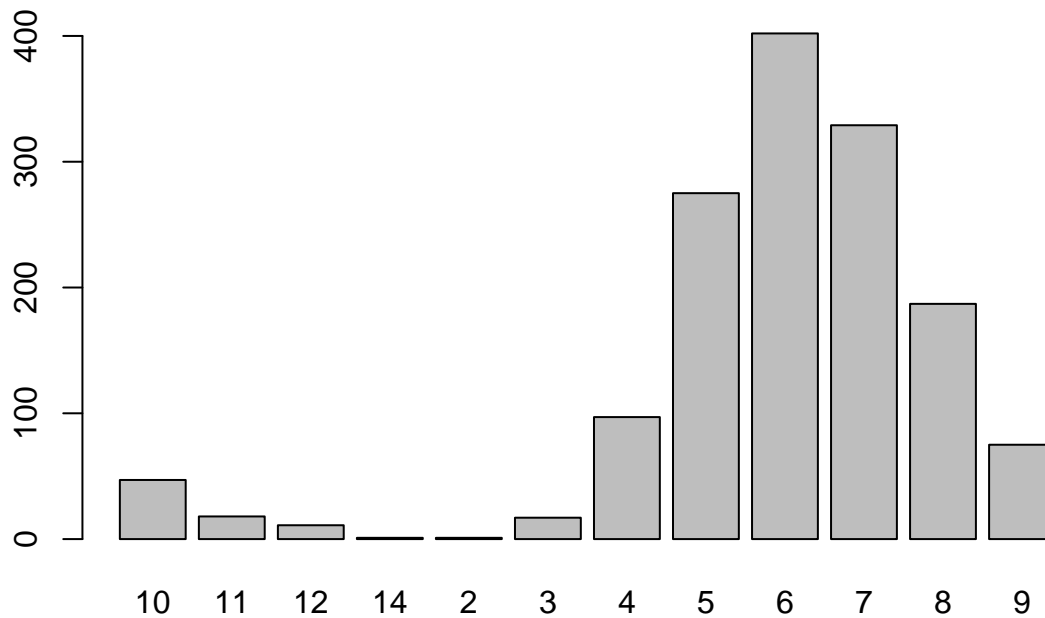
Casas recientemente remodeladas:

```
newhouses <- head(order(datos$YearRemodAdd, decreasing = T), n=10)
datos[newhouses,c("Id","YearRemodAdd", "SalePrice")]
```

##	Id	YearRemodAdd	SalePrice
## 158	158	2010	269500
## 379	379	2010	394432
## 820	820	2010	224000
## 856	856	2010	127000
## 899	899	2010	611657
## 988	988	2010	395192
## 88	88	2009	164500
## 104	104	2009	198900
## 179	179	2009	501837
## 212	212	2009	186000

Cantidad de habitaciones

```
totalGenres <- unlist(strsplit(as.character(datos$TotRmsAbvGrd), "\\|"))
barplot(table(totalGenres))
```



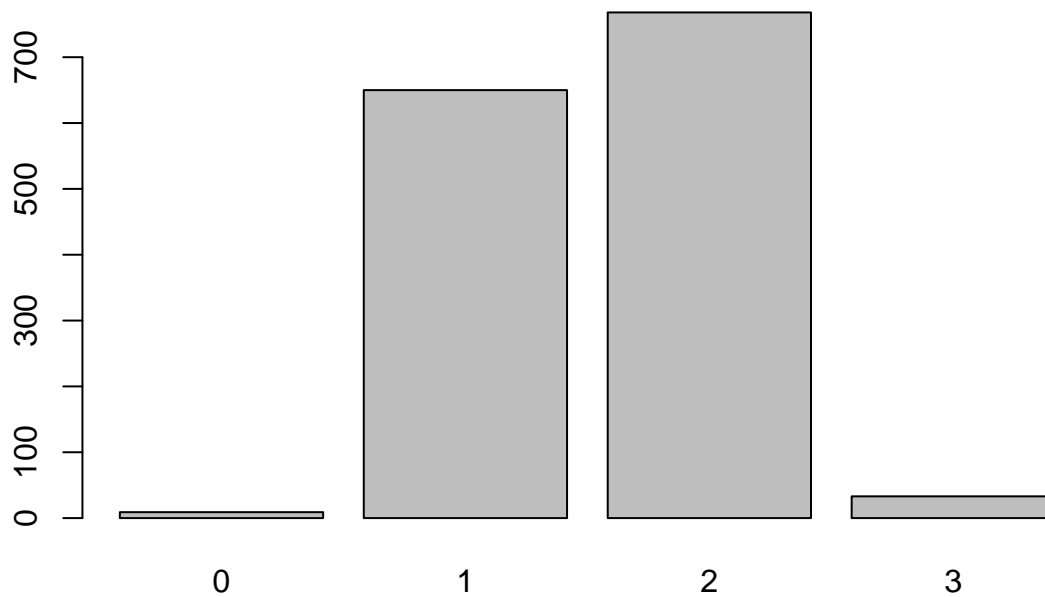
La mayoría de las casast tienen solo 6 habitaciones

```
newhouses <- head(order(datos$TotRmsAbvGrd, decreasing = T), n=10)
datos[newhouses,c("Id","TotRmsAbvGrd","SalePrice")]
```

```
##      Id TotRmsAbvGrd SalePrice
## 636  636          14   200000
## 186  186          12   475000
## 770  770          12   538000
## 804  804          12   582933
## 898  898          12   142953
## 911  911          12   154300
## 1032 1032          12   197000
## 1174 1174          12   200500
## 1231 1231          12   190000
## 1299 1299          12   160000
```

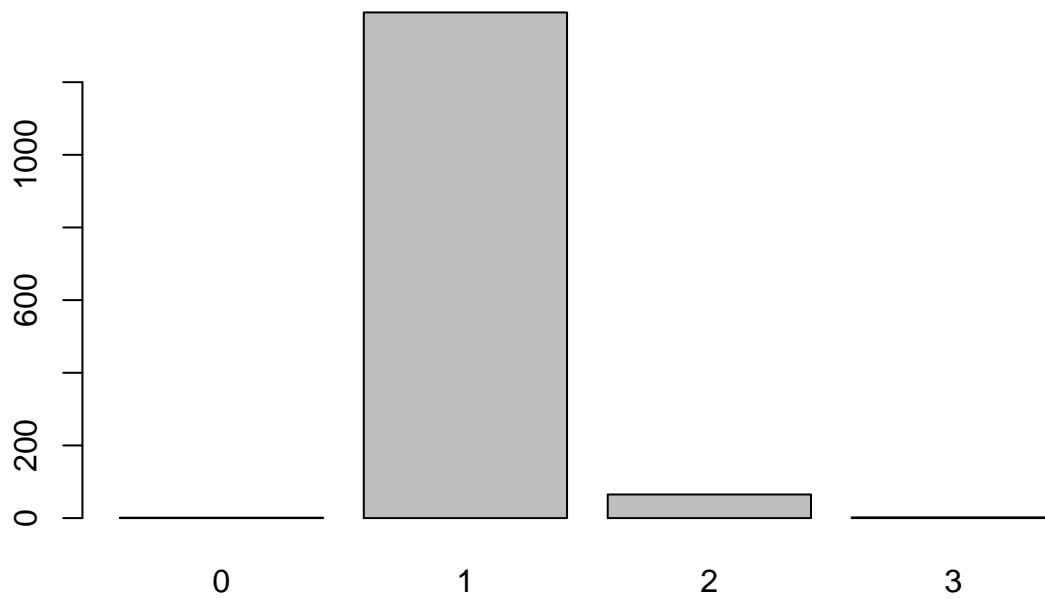
cantidad de baños

```
totalGenres <- unlist(strsplit(as.character(datos$FullBath), "\\|"))
barplot(table(totalGenres))
```

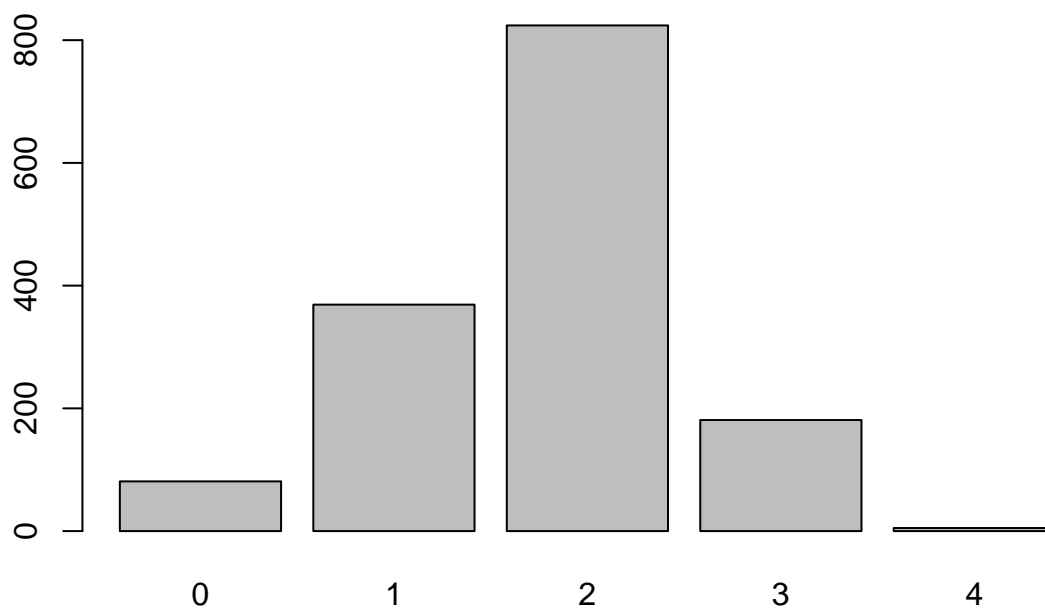
Las casas familiares tienen entre 1 y 2 baños
cantidad de cocinas

```
totalGenres <- unlist(strsplit(as.character(datos$KitchenAbvGr), "\\|"))  
barplot(table(totalGenres))
```



LAs casas por lo general tienen solo una cocina
capacidad de garages

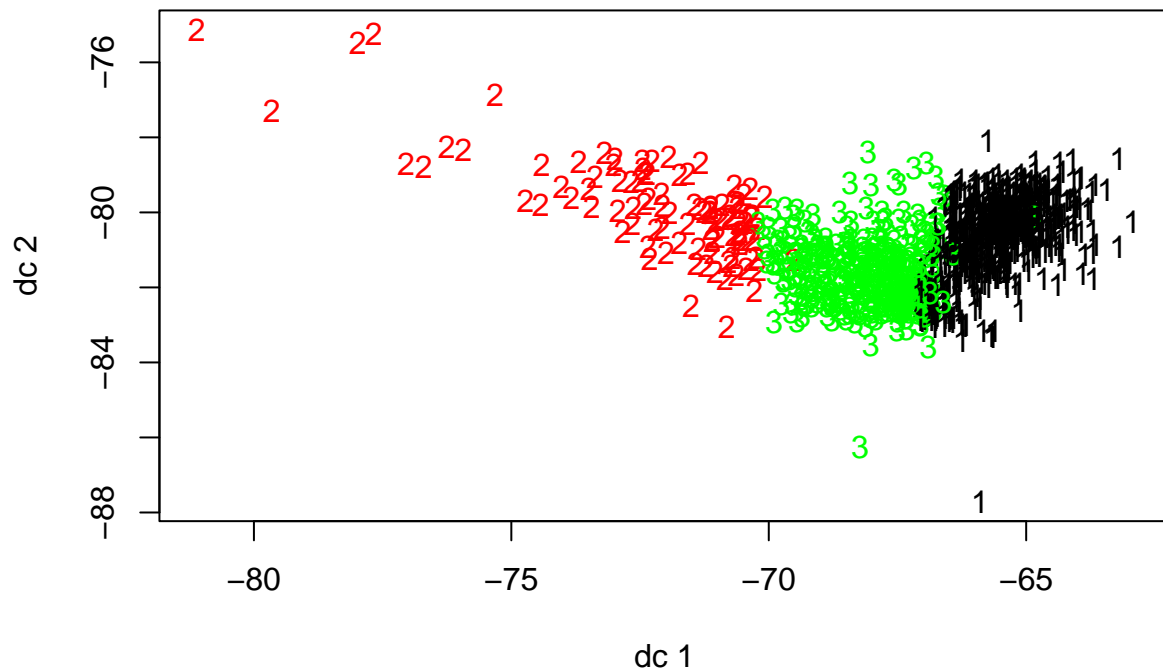
```
totalGenres <- unlist(strsplit(as.character(datos$GarageCars), "\\|"))  
barplot(table(totalGenres))
```



La mayoría de las casas estan adecuadas para estacionar 2 autos

3. Incluya un análisis de grupos en el análisis exploratorio. Explique las características de los grupos.

```
# con k-medias
cluster <- house
km<-kmeans(house,3)
house$grupo<-km$cluster
plotcluster(cluster,km$cluster) # los cluster
```



```
#método de la silueta
silkm<-silhouette(km$cluster,dist(house))
mean(silkm[,3]) #Silueta
```

```
## [1] 0.561677
```

```
g1<- house[house$grupo==1,]
prop.table(table(g1$Species))*100
```

```
## numeric(0)
```

```
g2<- house[house$grupo==2,]
prop.table(table(g2$Species))*100
```

```
## numeric(0)
```

```
g3<- house[house$grupo==3,]
prop.table(table(g3$Species))*100
```

```
## numeric(0)
```

```
summary(g1)
```

```
##      LotFrontage      LotArea      YearBuilt      YearRemodAdd
##  Min.   : 21.00    Min.   : 1300    Min.   :1880    Min.   :1950
## 1st Qu.: 55.00    1st Qu.: 7000    1st Qu.:1940    1st Qu.:1956
## Median : 64.00    Median : 8593    Median :1959    Median :1972
## Mean   : 65.45    Mean   : 8680    Mean   :1957    Mean   :1976
## 3rd Qu.: 75.00    3rd Qu.:10172    3rd Qu.:1973    3rd Qu.:1998
## Max.   :313.00    Max.   :63887    Max.   :2009    Max.   :2009
##      MasVnrArea      BsmtFinSF1      BsmtFinSF2      BsmtUnfSF
##  Min.   : 0.00    Min.   : 0.0    Min.   : 0.00    Min.   : 0.0
## 1st Qu.: 0.00    1st Qu.: 0.0    1st Qu.: 0.00    1st Qu.:192.0
## Median : 0.00    Median : 319.0    Median : 0.00    Median : 450.0
## Mean   : 60.21    Mean   : 351.2    Mean   : 51.82    Mean   : 491.1
## 3rd Qu.: 66.50    3rd Qu.: 600.0    3rd Qu.: 0.00    3rd Qu.: 719.0
## Max.   :1129.00    Max.   :5644.0    Max.   :1085.00    Max.   :1907.0
##      TotalBsmtSF      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##  Min.   : 0.0    Min.   : 438    Min.   : 0.0    Min.   : 0.000
## 1st Qu.: 728.0    1st Qu.: 834    1st Qu.: 0.0    1st Qu.: 0.000
## Median : 894.0    Median : 981    Median : 0.0    Median : 0.000
## Mean   : 894.1    Mean   :1020    Mean   : 256.5    Mean   : 4.513
## 3rd Qu.:1060.5    3rd Qu.:1150    3rd Qu.: 588.0    3rd Qu.: 0.000
## Max.   :6110.0    Max.   :4692    Max.   :1230.0    Max.   :481.000
##      GrLivArea      TotRmsAbvGrd      Fireplaces      GarageYrBlt      GarageCars
##  Min.   : 438    Min.   : 3    Min.   :0.0000    Min.   :1900    Min.   :1.000
## 1st Qu.:1032    1st Qu.: 5    1st Qu.:0.0000    1st Qu.:1953    1st Qu.:1.000
## Median :1224    Median : 6    Median :0.0000    Median :1966    Median :2.000
## Mean   :1281    Mean   : 6    Mean   :0.4137    Mean   :1967    Mean   :1.568
## 3rd Qu.:1484    3rd Qu.: 7    3rd Qu.:1.0000    3rd Qu.:1983    3rd Qu.:2.000
## Max.   :5642    Max.   :12    Max.   :3.0000    Max.   :2009    Max.   :4.000
##      GarageArea      WoodDeckSF      OpenPorchSF      EnclosedPorch
##  Min.   : 160.0    Min.   : 0.00    Min.   : 0.00    Min.   : 0.00
## 1st Qu.: 288.0    1st Qu.: 0.00    1st Qu.: 0.00    1st Qu.: 0.00
## Median : 416.0    Median : 0.00    Median : 0.00    Median : 0.00
## Mean   : 417.4    Mean   : 63.95    Mean   : 26.44    Mean   : 30.07
## 3rd Qu.: 506.0    3rd Qu.:106.00    3rd Qu.: 35.00    3rd Qu.: 0.00
## Max.   :1418.0    Max.   :736.00    Max.   :418.00    Max.   :330.00
##      ScreenPorch      PoolArea      MoSold      YrSold
##  Min.   : 0.00    Min.   : 0.000    Min.   : 1.000    Min.   :2006
## 1st Qu.: 0.00    1st Qu.: 0.000    1st Qu.: 4.000    1st Qu.:2007
## Median : 0.00    Median : 0.000    Median : 6.000    Median :2008
## Mean   : 13.26    Mean   : 1.612    Mean   : 6.157    Mean   :2008
## 3rd Qu.: 0.00    3rd Qu.: 0.000    3rd Qu.: 7.000    3rd Qu.:2009
## Max.   :385.00    Max.   :576.000    Max.   :12.000    Max.   :2010
##      SalePrice      grupo
##  Min.   : 35311    Min.   :1
## 1st Qu.:118450    1st Qu.:1
## Median :135000    Median :1
## Mean   :134136    Mean   :1
## 3rd Qu.:154400    3rd Qu.:1
## Max.   :178000    Max.   :1
```

```
summary(g2)
```

```
##      LotFrontage      LotArea      YearBuilt      YearRemodAdd
##  Min.   : 32.00    Min.   : 5119    Min.   :1892    Min.   :1965
## 1st Qu.: 76.00    1st Qu.: 11003    1st Qu.:1998    1st Qu.:2000
## Median : 86.00    Median : 12444    Median :2005    Median :2006
## Mean   : 87.82    Mean   : 16048    Mean   :1999    Mean   :2003
## 3rd Qu.: 99.00    3rd Qu.: 14601    3rd Qu.:2007    3rd Qu.:2007
## Max.   :174.00    Max.   :215245    Max.   :2010    Max.   :2010
##      MasVnrArea      BsmtFinSF1      BsmtFinSF2      BsmtUnfSF
##  Min.   : 0.0      Min.   : 0.0      Min.   : 0.0      Min.   : 0.0
## 1st Qu.: 160.0     1st Qu.: 240.0     1st Qu.: 0.0      1st Qu.: 294.0
## Median : 302.0     Median :1163.0     Median : 0.0      Median : 528.0
## Mean   : 345.6     Mean   : 939.3     Mean   : 28.7      Mean   : 705.9
## 3rd Qu.: 466.0     3rd Qu.:1406.0     3rd Qu.: 0.0      3rd Qu.:1117.0
## Max.   :1378.0     Max.   :2188.0     Max.   :1474.0     Max.   :2336.0
##      TotalBsmtSF      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##  Min.   : 853      Min.   :1026      Min.   : 0.0      Min.   : 0.000
## 1st Qu.:1410      1st Qu.:1470      1st Qu.: 0.0      1st Qu.: 0.000
## Median :1702      Median :1718      Median : 568.0     Median : 0.000
## Mean   :1674      Mean   :1699      Mean   : 593.6     Mean   : 5.448
## 3rd Qu.:1926      3rd Qu.:1940      3rd Qu.:1177.0     3rd Qu.: 0.000
## Max.   :3094      Max.   :2444      Max.   :2065.0     Max.   :572.000
##      GrLivArea      TotRmsAbvGrd      Fireplaces      GarageYrBlt      GarageCars
##  Min.   :1419      Min.   : 5.000      Min.   :0.0      Min.   :1932      Min.   :2.000
## 1st Qu.:1869      1st Qu.: 7.000      1st Qu.:1.0      1st Qu.:1998      1st Qu.:3.000
## Median :2224      Median : 8.000      Median :1.0      Median :2005      Median :3.000
## Mean   :2298      Mean   : 8.457      Mean   :1.2      Mean   :2001      Mean   :2.781
## 3rd Qu.:2610      3rd Qu.:10.000      3rd Qu.:1.0      3rd Qu.:2007      3rd Qu.:3.000
## Max.   :4476      Max.   :12.000      Max.   :3.0      Max.   :2010      Max.   :3.000
##      GarageArea      WoodDeckSF      OpenPorchSF      EnclosedPorch
##  Min.   : 380.0      Min.   : 0      Min.   : 0.00      Min.   : 0.000
## 1st Qu.: 672.0      1st Qu.:113      1st Qu.: 45.00      1st Qu.: 0.000
## Median : 774.0      Median :186      Median : 72.00      Median : 0.000
## Mean   : 762.2      Mean   :192      Mean   : 88.22      Mean   : 5.876
## 3rd Qu.: 842.0      3rd Qu.:250      3rd Qu.:111.00      3rd Qu.: 0.000
## Max.   :1220.0      Max.   :857      Max.   :502.00      Max.   :216.000
##      ScreenPorch      PoolArea      MoSold      YrSold
##  Min.   : 0.0      Min.   : 0.000      Min.   : 1.000      Min.   :2006
## 1st Qu.: 0.0      1st Qu.: 0.000      1st Qu.: 5.000      1st Qu.:2007
## Median : 0.0      Median : 0.000      Median : 7.000      Median :2008
## Mean   : 22.3      Mean   : 5.286      Mean   : 6.743      Mean   :2008
## 3rd Qu.: 0.0      3rd Qu.: 0.000      3rd Qu.: 9.000      3rd Qu.:2009
## Max.   :410.0      Max.   :555.000      Max.   :12.000      Max.   :2010
##      SalePrice      grupo
##  Min.   :301000      Min.   :2
## 1st Qu.:320000      1st Qu.:2
## Median :348000      Median :2
## Mean   :377265      Mean   :2
## 3rd Qu.:395000      3rd Qu.:2
## Max.   :755000      Max.   :2
```

```
summary(g3)
```

```
##      LotFrontage      LotArea      YearBuilt      YearRemodAdd
##  Min.   : 24.00    Min.   : 2280    Min.   :1880    Min.   :1950
## 1st Qu.: 63.00    1st Qu.: 8737    1st Qu.:1990    1st Qu.:1996
## Median : 74.00    Median :10206    Median :2002    Median :2003
## Mean   : 75.13    Mean   :11018    Mean   :1991    Mean   :1998
## 3rd Qu.: 85.00    3rd Qu.:12011    3rd Qu.:2005    3rd Qu.:2006
## Max.   :313.00    Max.   :70761    Max.   :2009    Max.   :2010
##      MasVnrArea      BsmtFinSF1      BsmtFinSF2      BsmtUnfSF
##  Min.   : 0.0      Min.   : 0      Min.   : 0.0      Min.   : 0.0
## 1st Qu.: 0.0      1st Qu.: 0      1st Qu.: 0.0      1st Qu.: 314.0
## Median : 40.0      Median : 410      Median : 0.0      Median : 725.0
## Mean   :127.1      Mean   : 451      Mean   : 36.1      Mean   : 748.5
## 3rd Qu.:210.0      3rd Qu.: 767      3rd Qu.: 0.0      3rd Qu.:1141.0
## Max.   :1600.0      Max.   :2260      Max.   :1127.0      Max.   :2042.0
##      TotalBsmtSF      X1stFlrSF      X2ndFlrSF      LowQualFinSF
##  Min.   : 0      Min.   : 495      Min.   : 0.0      Min.   : 0.000
## 1st Qu.: 915      1st Qu.:1020      1st Qu.: 0.0      1st Qu.: 0.000
## Median :1240      Median :1307      Median : 0.0      Median : 0.000
## Mean   :1236      Mean   :1299      Mean   : 458.9      Mean   : 4.413
## 3rd Qu.:1496      3rd Qu.:1552      3rd Qu.: 872.0      3rd Qu.: 0.000
## Max.   :3206      Max.   :3138      Max.   :1818.0      Max.   :420.000
##      GrLivArea      TotRmsAbvGrd      Fireplaces      GarageYrBlt      GarageCars
##  Min.   :1146      Min.   : 4.000      Min.   :0.0000      Min.   :1908      Min.   :1.00
## 1st Qu.:1502      1st Qu.: 6.000      1st Qu.:0.0000      1st Qu.:1991      1st Qu.:2.00
## Median :1684      Median : 7.000      Median :1.0000      Median :2002      Median :2.00
## Mean   :1763      Mean   : 7.075      Mean   :0.7756      Mean   :1993      Mean   :2.18
## 3rd Qu.:1947      3rd Qu.: 8.000      3rd Qu.:1.0000      3rd Qu.:2005      3rd Qu.:2.00
## Max.   :4676      Max.   :12.000      Max.   :3.0000      Max.   :2010      Max.   :4.00
##      GarageArea      WoodDeckSF      OpenPorchSF      EnclosedPorch
##  Min.   : 180.0      Min.   : 0.0      Min.   : 0.00      Min.   : 0.00
## 1st Qu.: 484.0      1st Qu.: 0.0      1st Qu.: 30.00      1st Qu.: 0.00
## Median : 552.0      Median :120.0      Median : 54.00      Median : 0.00
## Mean   : 582.9      Mean   :115.7      Mean   : 69.62      Mean   : 11.56
## 3rd Qu.: 662.0      3rd Qu.:192.0      3rd Qu.: 96.00      3rd Qu.: 0.00
## Max.   :1390.0      Max.   :635.0      Max.   :547.00      Max.   :552.00
##      ScreenPorch      PoolArea      MoSold      YrSold
##  Min.   : 0.00      Min.   : 0.000      Min.   : 1.000      Min.   :2006
## 1st Qu.: 0.00      1st Qu.: 0.000      1st Qu.: 5.000      1st Qu.:2007
## Median : 0.00      Median : 0.000      Median : 6.000      Median :2008
## Mean   : 19.45      Mean   : 4.651      Mean   : 6.554      Mean   :2008
## 3rd Qu.: 0.00      3rd Qu.: 0.000      3rd Qu.: 8.000      3rd Qu.:2009
## Max.   :480.00      Max.   :648.000      Max.   :12.000      Max.   :2010
##      SalePrice      grupo
##  Min.   :178900      Min.   :3
## 1st Qu.:193000      1st Qu.:3
## Median :216837      Median :3
## Mean   :222938      Mean   :3
## 3rd Qu.:248900      3rd Qu.:3
## Max.   :297000      Max.   :3
```

— prueba 1 —

```

##Similitud en las variables independientes y los precios de venta:
cor(house$YearBuilt, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.5253936

cor(house$YearRemodAdd, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.5212533

cor(house$TotalBsmtSF, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.6156122

cor(house$X1stFlrSF, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.6079691

cor(house$GrLivArea, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.7051536

cor(house$TotRmsAbvGrd, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.5470674

cor(house$GarageCars, house$SalePrice, method = c("pearson", "kendall", "spearman"))

## [1] 0.6470336

cor(house$GarageArea, house$SalePrice, method = c("pearson", "kendall", "spearman"))

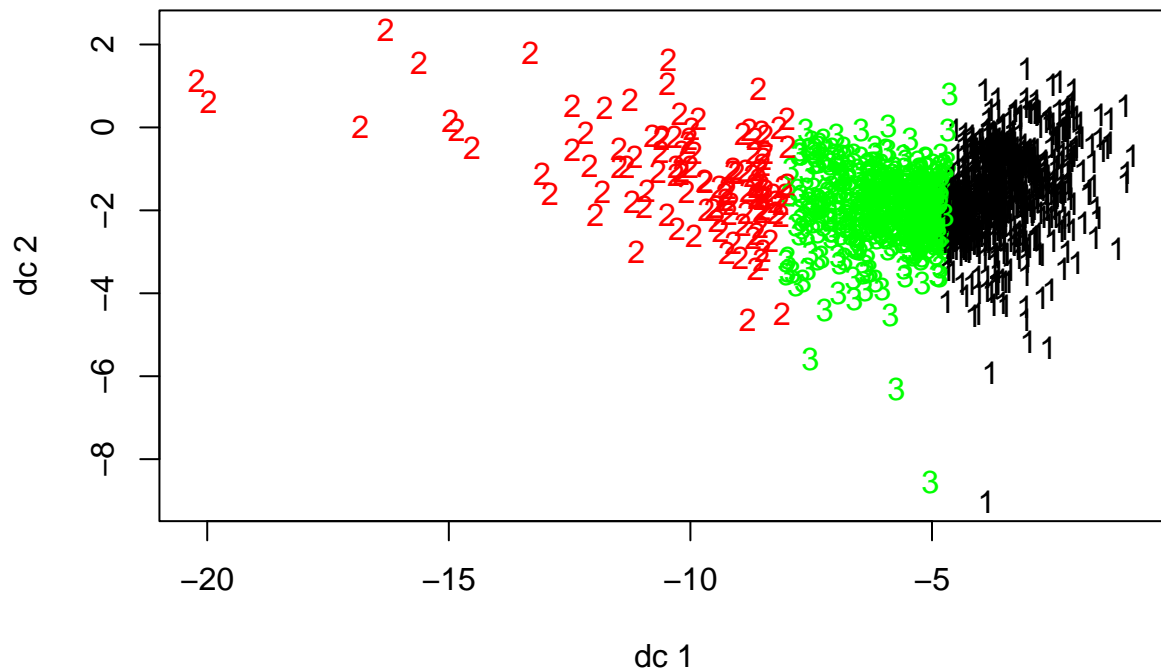
## [1] 0.6193296

#Columnas
house <-select(datos,TotalBsmtSF,X1stFlrSF,GrLivArea,GarageCars,GarageArea,SalePrice)

#limpiamos
house <- na.omit(house)

#k-medias
cluster <- house
km<-kmeans(house,3)
house$grupo<-km$cluster
plotcluster(cluster,km$cluster) #graficamos ubicacion de cluster

```

```
#Silueta
silkm<-silhouette(km$cluster,dist(house))
mean(silkm[,3])
```

```
## [1] 0.5612983
```

— prueba 2 —

```
cor(house$TotalBsmtSF, house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6135806
```

```
cor(house$X1stFlrSF, house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6058522
```

```
cor(house$GrLivArea, house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.7086245
```

```
cor(house$GarageCars, house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6404092
```

```
cor(house$GarageArea, house$SalePrice, method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.6234314
```

```
#Columnas
```

```
house <-select(datos,TotalBsmtSF,X1stFlrSF,GrLivArea,GarageCars,GarageArea,SalePrice)
```

```
#limpiamos
```

```
house <- na.omit(house)
```

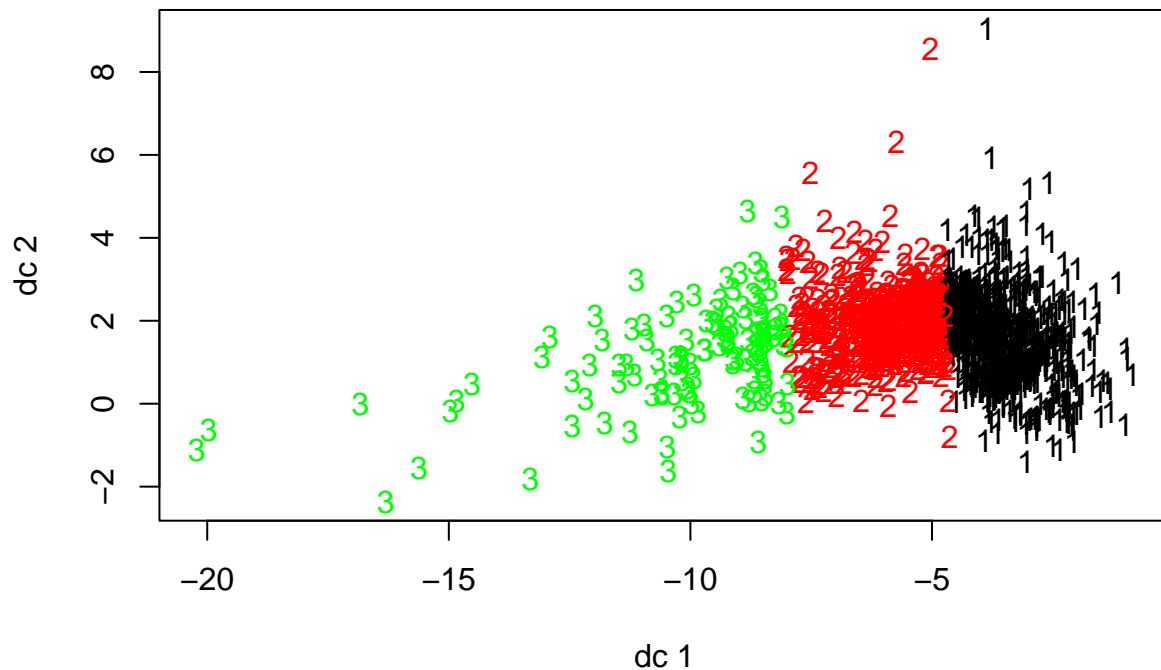
```
#k-medias
```

```
cluster <- house
```

```
km<-kmeans(house,3)
```

```
house$grupo<-km$cluster
```

```
plotcluster(cluster,km$cluster) #Graficamos clusters
```



```
#silueta
```

```
silkkm<-silhouette(km$cluster,dist(house))
```

```
mean(silkkm[,3])
```

```
## [1] 0.5612983
```

En el análisis de grupos, se aplicó el algoritmo de K-Means a cada una de las variables que se habían analizado previamente. El objetivo era agrupar las casas según su similitud en cada una de esas variables. Además,

se identificaron las variables que se consideraron importantes para determinar el precio de una casa y se calcularon medidas de tendencia central para cada una de ellas, como la media y la mediana.

Por otro lado, se identificaron las variables que se consideraron relevantes para el precio de una casa, lo que permitió centrar el análisis en aquellas variables que podrían ser más influyentes en el precio de una casa. Se calcularon medidas de tendencia central para cada una de estas variables.

La silueta no muestra mejoría ya que los clusters disminuyen su valor.

4. Divida el set de datos preprocesados en dos conjuntos: Entrenamiento y prueba. Describa el criterio que usó para crear los conjuntos: número de filas de cada uno, estratificado o no, balanceado o no, etc. Si le proveen un conjunto de datos de prueba y tiene suficientes datos, tómelo como de validación, pero haga sus propios conjuntos de prueba.

Division de sets: porcentajes de 70% entrenamiento 30% de prueba

```
set_entrenamiento <- sample_frac(datos, .7)
set_prueba <- setdiff(datos, set_entrenamiento)

drop <- c("LotFrontage", "Alley", "MasVnrType", "MasVnrArea", "BsmtQual", "BsmtCond", "BsmtExposure", "
set_entrenamiento <- set_entrenamiento[, !(names(set_entrenamiento) %in% drop)]
set_prueba <- set_prueba[, !(names(set_prueba) %in% drop)]
```

5. Haga ingeniería de características, ¿qué variables cree que puedan ser mejores predictores para el precio de las casas? Explique en que basó la selección o no de las variables.

LotArea: Tamaño del terreno de la casa Neighborhood: Vecindario donde esta ubicada la casa BldgType: Tipo/estilo de casa OverallQual: Material de la casa FullBath: Cantidad de baños TotRmsAbvGrd: Cantidad de habitaciones Fireplaces: Si tiene chimenea GarageCars: tamaño del parqueo en capacidad de autos

Nos basamos en las características mas distintivas de las casas, y las que considera el mercado actual para darle valor a un inmueble La ubicación y tamaño de la casa son una de las más importantes, así como la cantidad de habitaciones y parqueos que tiene. Loas materiales de construcción nos permite deducir cuanto dinero se invirtió durante su construcción y el hecho de tener chimenea hace que aumenta el valor de un inmueble considerablemente

6. Todos los resultados deben ser reproducibles por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.

```
set.seed(123) # Establecer semilla aleatoria
indices_entrenamiento <- sample(1:nrow(datos), round(0.7*nrow(datos)), replace = FALSE)
datos_entrenamiento <- datos[indices_entrenamiento, ]
datos_prueba <- datos[-indices_entrenamiento, ]
```

se pueden establecer semillas aleatorias en el proceso de división del conjunto de datos en entrenamiento y prueba. De esta manera, se asegura que siempre se obtengan los mismos conjuntos de datos.

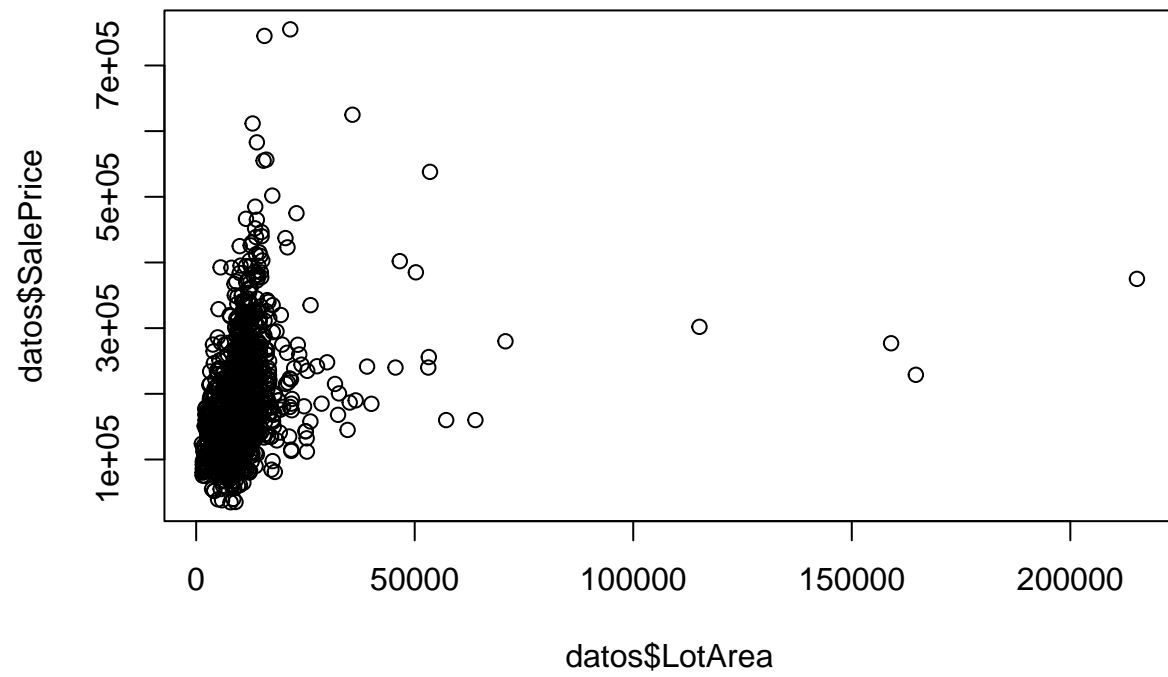
7. Seleccione una de las variables y haga un modelo univariado de regresión lineal para predecir el precio de las casas. Analice el modelo

```
modelo_univariado <- lm(SalePrice ~ LotArea, data = datos_entrenamiento)
summary(modelo_univariado)
```

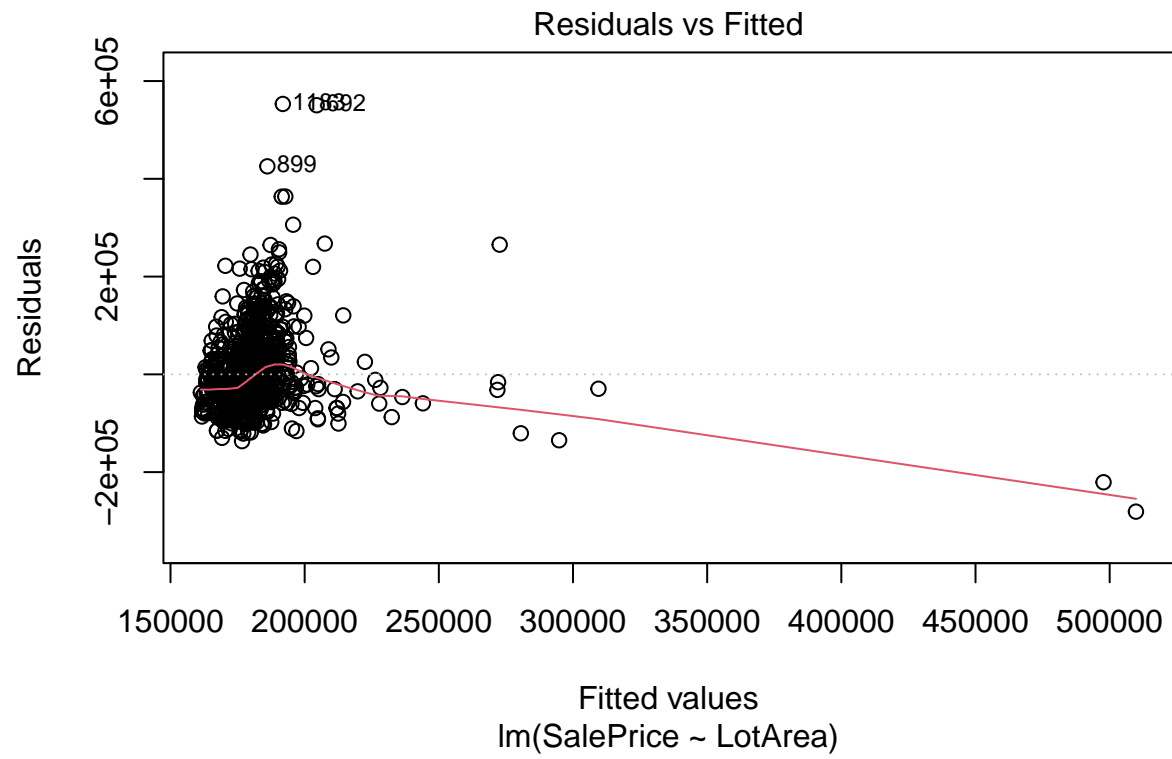
```
##
## Call:
## lm(formula = SalePrice ~ LotArea, data = datos_entrenamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -280843  -48715  -19730   31293  553138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.585e+05  3.742e+03  42.368  < 2e-16 ***
## LotArea      2.133e+00  2.738e-01   7.792 1.61e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77310 on 1020 degrees of freedom
## Multiple R-squared:  0.05619,    Adjusted R-squared:  0.05526
## F-statistic: 60.72 on 1 and 1020 DF,  p-value: 1.614e-14
```

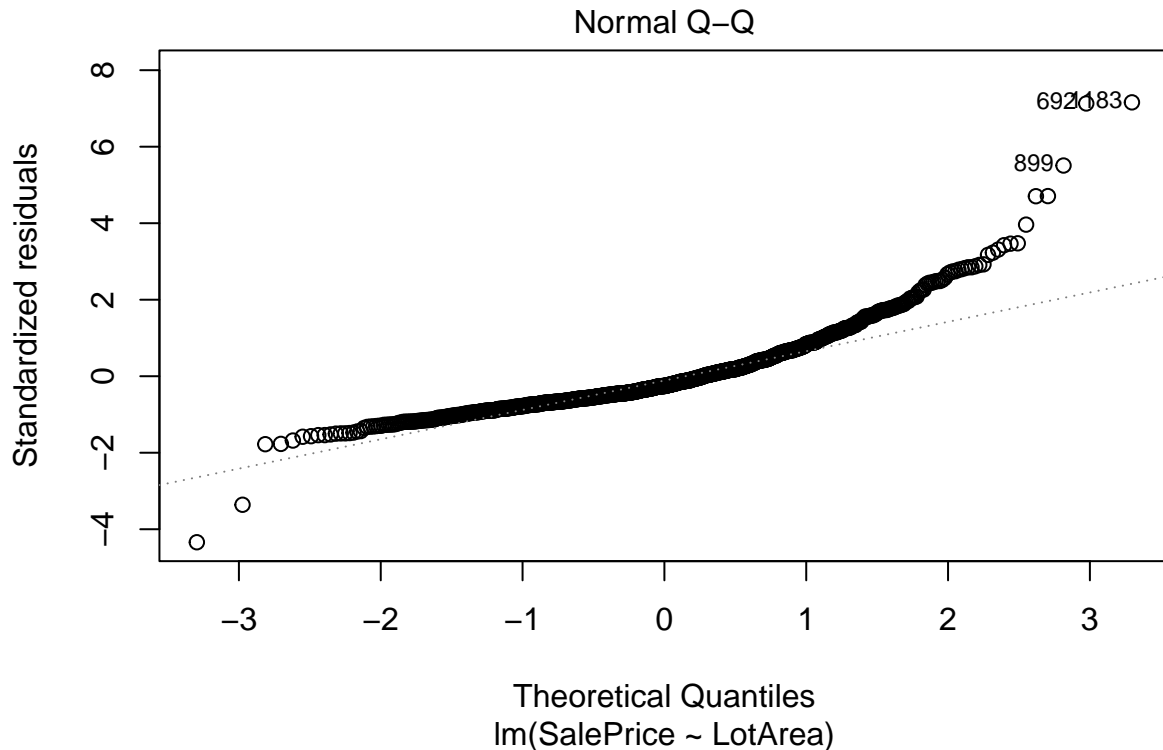
El resultado del modelo muestra el coeficiente de la variable “LotArea” y el coeficiente de determinación (R^2), que indica la proporción de la variación en la variable dependiente que se explica por la variable independiente en el modelo.

```
plot(datos$LotArea,datos$SalePrice)
```



```
plot(modelo_univariado, which = c(1,2))
```





Si los residuos están distribuidos aleatoriamente alrededor de cero y no muestran patrones claros, se puede asumir que el modelo es adecuado.

8. Haga un modelo de regresión lineal con todas las variables numéricas para predecir el precio de las casas. Analice el modelo (resumen, residuos, resultados de la predicción). Muestre el modelo gráficamente.

```
porciento <- 70/100
datos$clasificacion <- ifelse(datos$SalePrice <= 251000, "Economicas", ifelse(datos$SalePrice <= 538000, "Medias", "Altas"))

datos$y <- as.numeric(factor(datos$clasificacion))
datosCC <- datos[,c(2,4,18,19,20,21,27,35,37,38,39,44,45,46,47,48,49,50,51,52,53,55,57,60,62,63,67,68,69)]
datosCC <- datosCC[,colSums(is.na(datosCC))==0]
set.seed(123)
trainRowsNumber<-sample(nrow(datosCC),porciento*nrow(datosCC))
train<-datosCC[trainRowsNumber,]
test<-datosCC[-trainRowsNumber,]

fitLM<-lm(SalePrice~., data = train)
summary(fitLM)

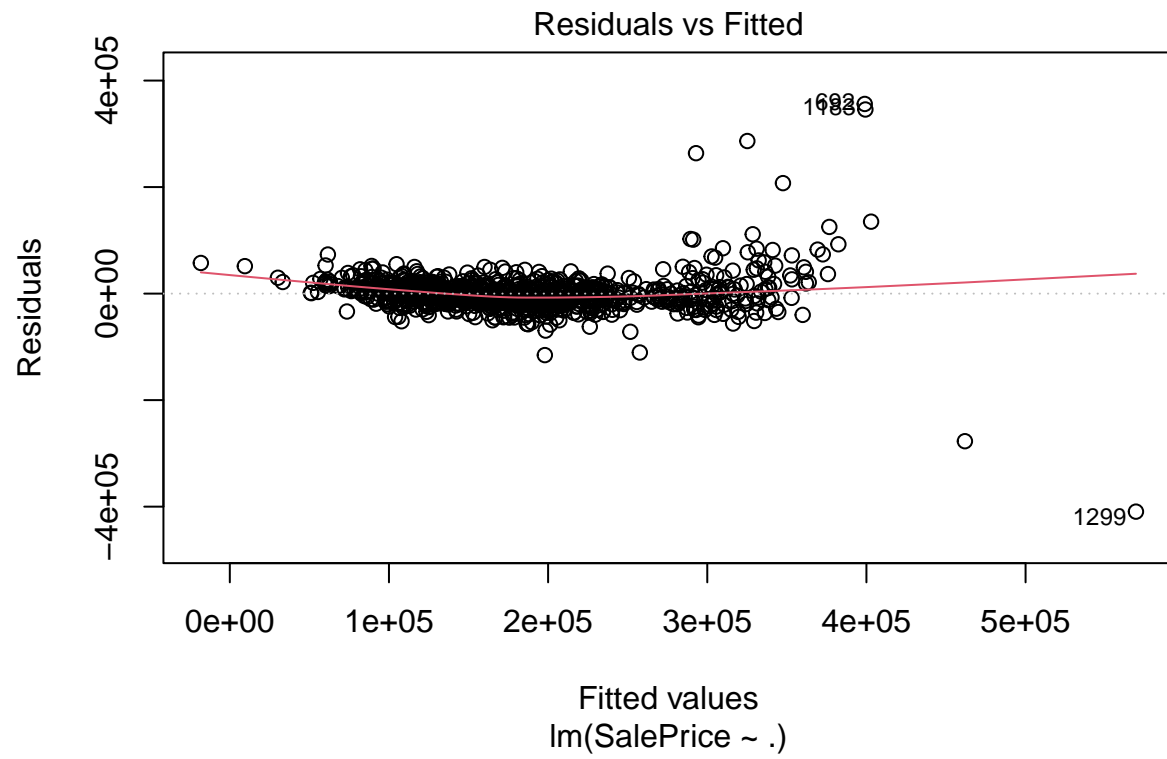
##
## Call:
## lm(formula = SalePrice ~ ., data = train)
##
```

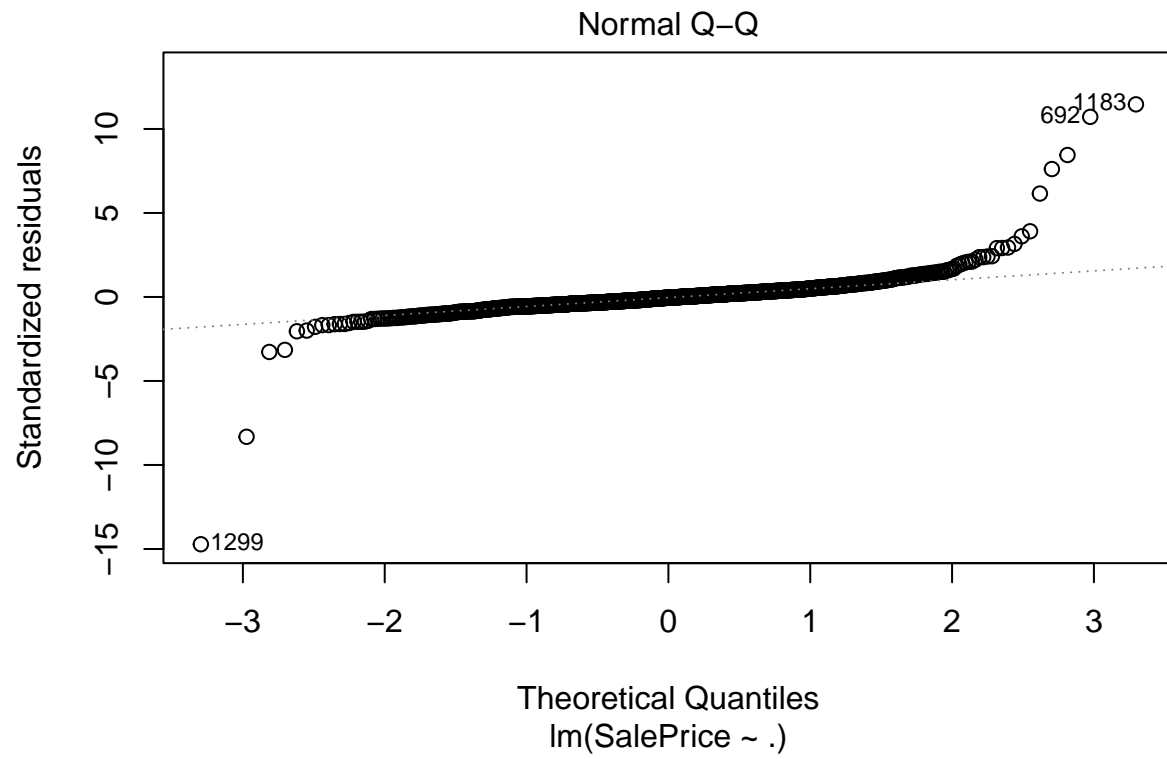
```

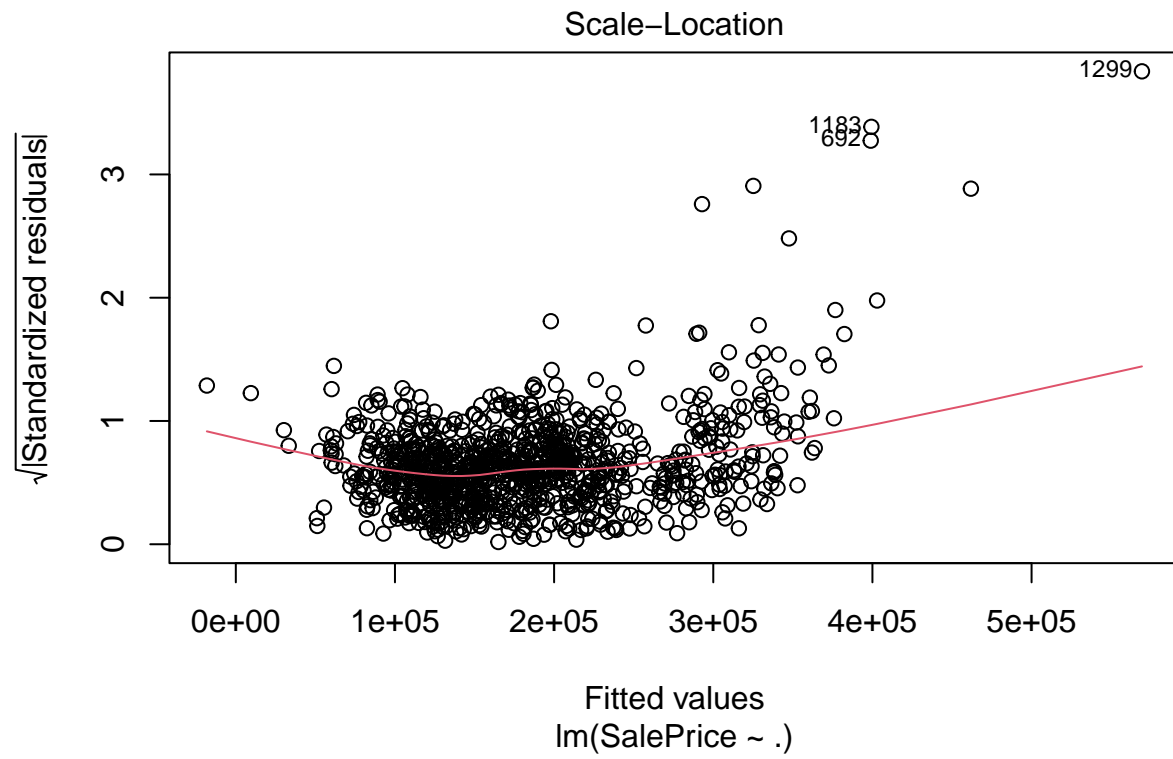
## Residuals:
##      Min       1Q   Median       3Q      Max
## -409357  -13924   -1521   11097  356054
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.064e+05  1.717e+06  -0.237  0.812984
## MSSubClass    -1.520e+02  3.248e+01  -4.680  3.26e-06 ***
## OverallQual    1.523e+04  1.494e+03  10.196  < 2e-16 ***
## OverallCond    4.302e+03  1.271e+03   3.386  0.000737 ***
## YearBuilt      3.472e+02  7.624e+01   4.553  5.94e-06 ***
## YearRemodAdd   1.128e+02  8.110e+01   1.391  0.164565
## BsmtFinSF1     1.868e+01  5.805e+00   3.218  0.001331 **
## BsmtFinSF2     8.863e+00  8.865e+00   1.000  0.317656
## BsmtUnfSF      7.719e+00  5.330e+00   1.448  0.147847
## TotalBsmtSF      NA         NA         NA      NA
## X1stFlrSF      4.513e+01  7.292e+00   6.188  8.89e-10 ***
## X2ndFlrSF      4.172e+01  6.006e+00   6.946  6.81e-12 ***
## LowQualFinSF   3.158e+01  2.310e+01   1.368  0.171753
## GrLivArea      NA         NA         NA      NA
## BsmtFullBath    6.811e+03  3.198e+03   2.130  0.033430 *
## BsmtHalfBath    9.361e+03  5.365e+03   1.745  0.081329 .
## FullBath        5.713e+03  3.515e+03   1.625  0.104417
## HalfBath       -1.497e+02  3.257e+03  -0.046  0.963340
## BedroomAbvGr   -9.937e+03  2.087e+03  -4.762  2.20e-06 ***
## KitchenAbvGr   -1.536e+04  6.372e+03  -2.410  0.016134 *
## TotRmsAbvGrd    5.698e+03  1.528e+03   3.730  0.000203 ***
## Fireplaces      4.313e+03  2.144e+03   2.012  0.044530 *
## GarageCars      1.217e+04  3.478e+03   3.498  0.000490 ***
## GarageArea     -1.132e+01  1.166e+01  -0.971  0.331910
## WoodDeckSF      2.184e+01  9.935e+00   2.198  0.028178 *
## OpenPorchSF    -2.412e+01  1.853e+01  -1.302  0.193369
## EnclosedPorch   1.174e+01  2.133e+01   0.550  0.582223
## X3SsnPorch      3.492e+01  3.496e+01   0.999  0.318074
## ScreenPorch     4.727e+01  2.126e+01   2.224  0.026395 *
## PoolArea       -6.411e-01  3.308e+01  -0.019  0.984540
## MiscVal         2.427e-02  2.144e+00   0.011  0.990970
## MoSold         -2.243e+02  4.264e+02  -0.526  0.599008
## YrSold         -3.044e+02  8.539e+02  -0.356  0.721578
## y               3.591e+04  3.874e+03   9.270  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35520 on 989 degrees of freedom
## Multiple R-squared:  0.8068, Adjusted R-squared:  0.8007
## F-statistic: 133.2 on 31 and 989 DF, p-value: < 2.2e-16

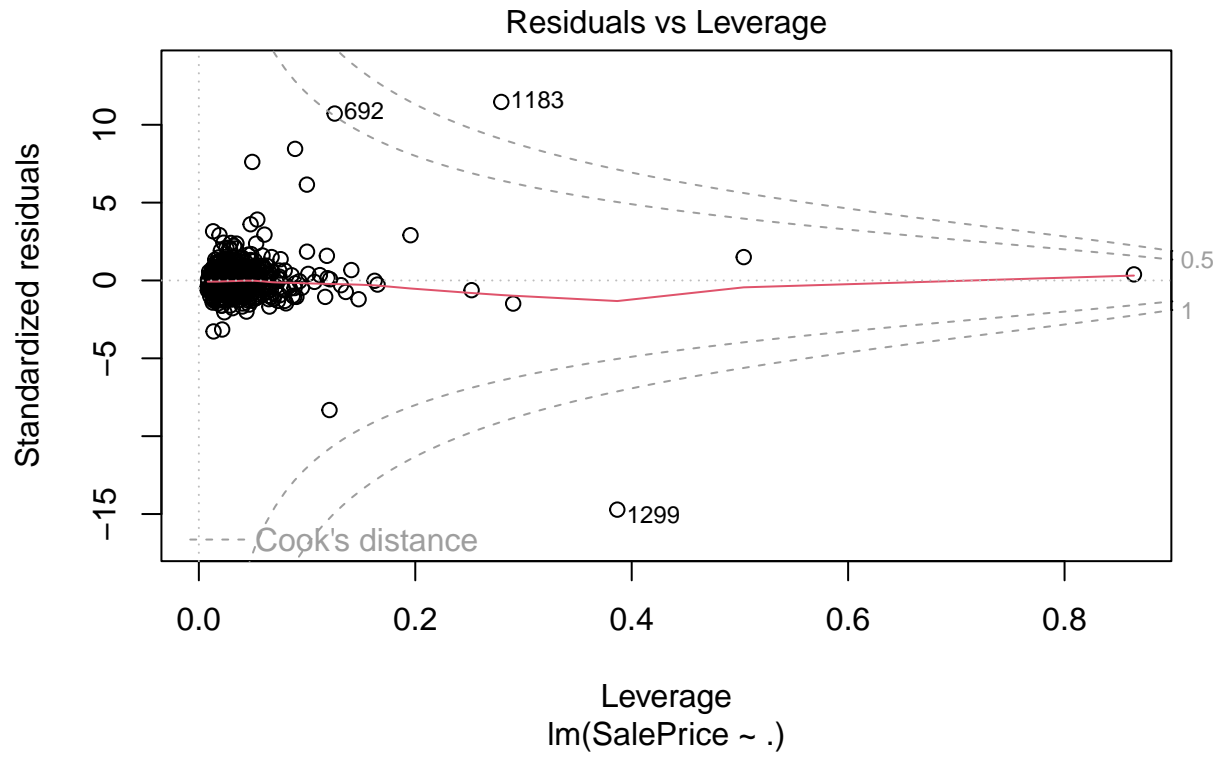
```

```
plot(fitLM)
```

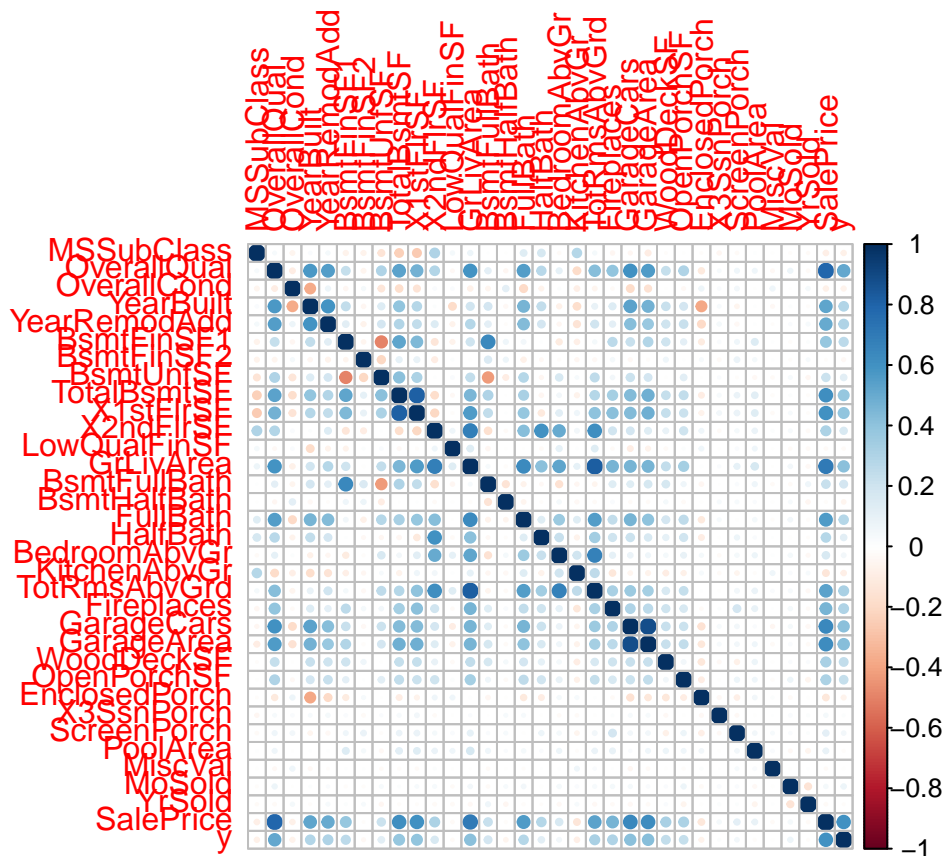






9. Analice el modelo. Determine si hay multicolinealidad entre las variables, y cuáles son las que aportan al modelo, por su valor de significación. Haga un análisis de correlación de las características del modelo y especifique si el modelo se adapta bien a los datos. Explique si hay sobreajuste (overfitting) o no. En caso de existir sobreajuste, haga otro modelo que lo corrija.

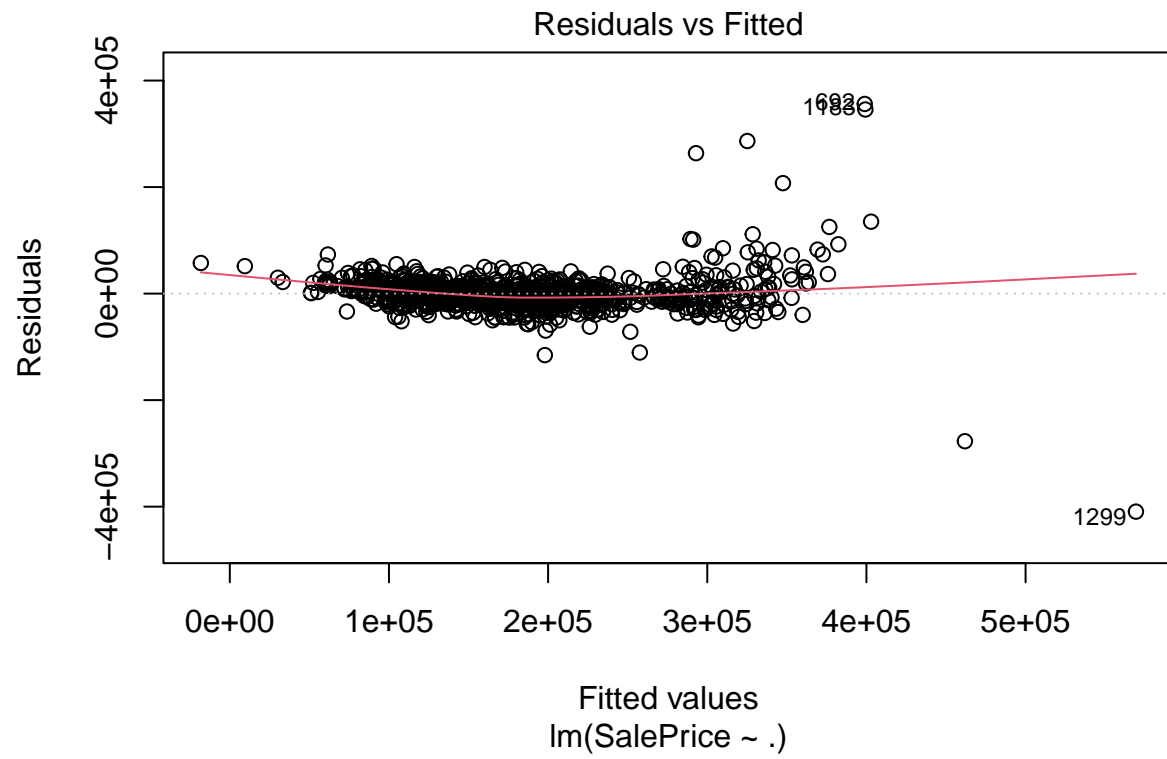
```
corrplot(cor(datosCC), method = "circle")
```

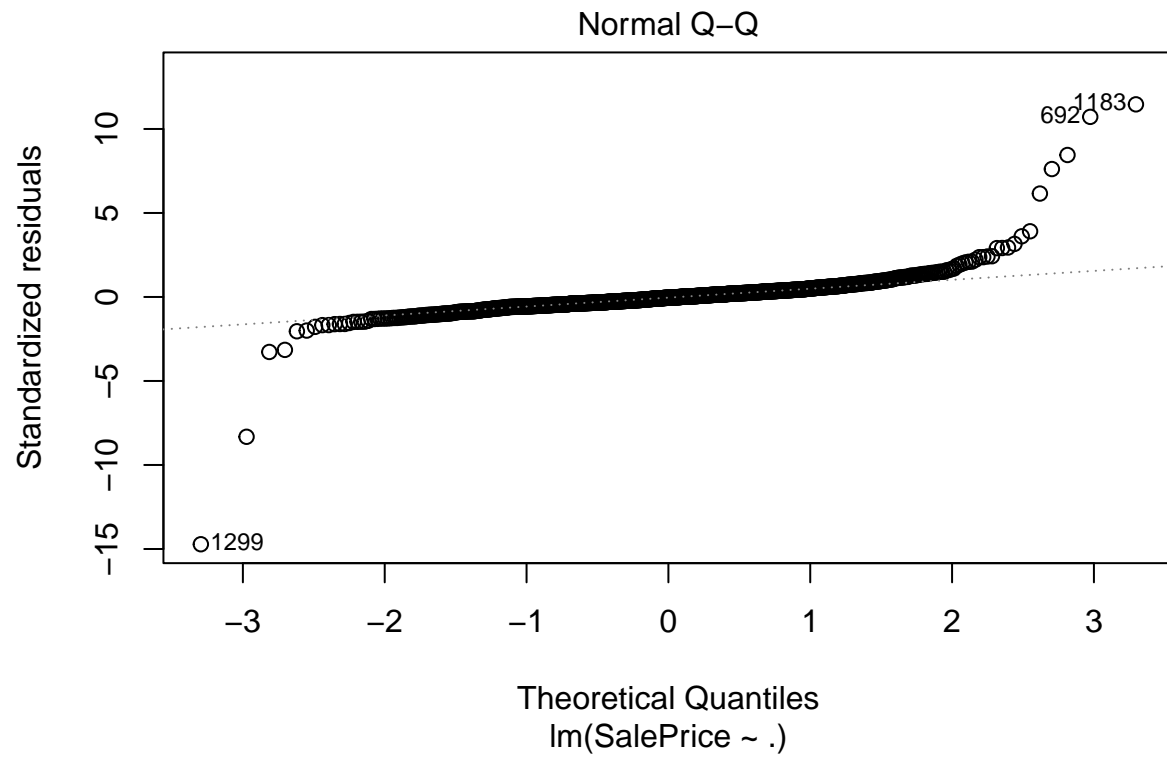


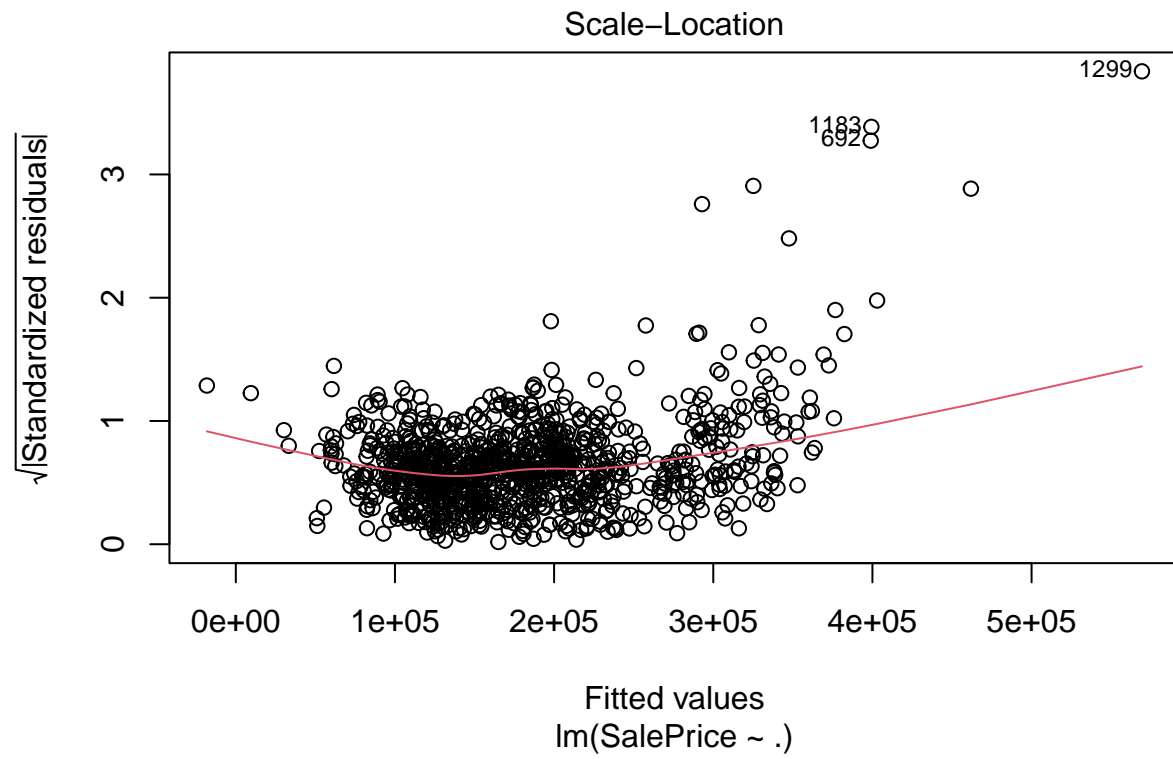
En general, hay poca correlación lineal entre las variables, aunque en algunas situaciones específicas se observan valores altos. Por lo tanto, se puede afirmar que el modelo se ajusta bien a los datos y algunas variables están altamente correlacionadas entre sí. Debido a la gran cantidad de correlaciones obtenidas en el análisis, se puede concluir que el modelo no está sobreajustado (overfitting).

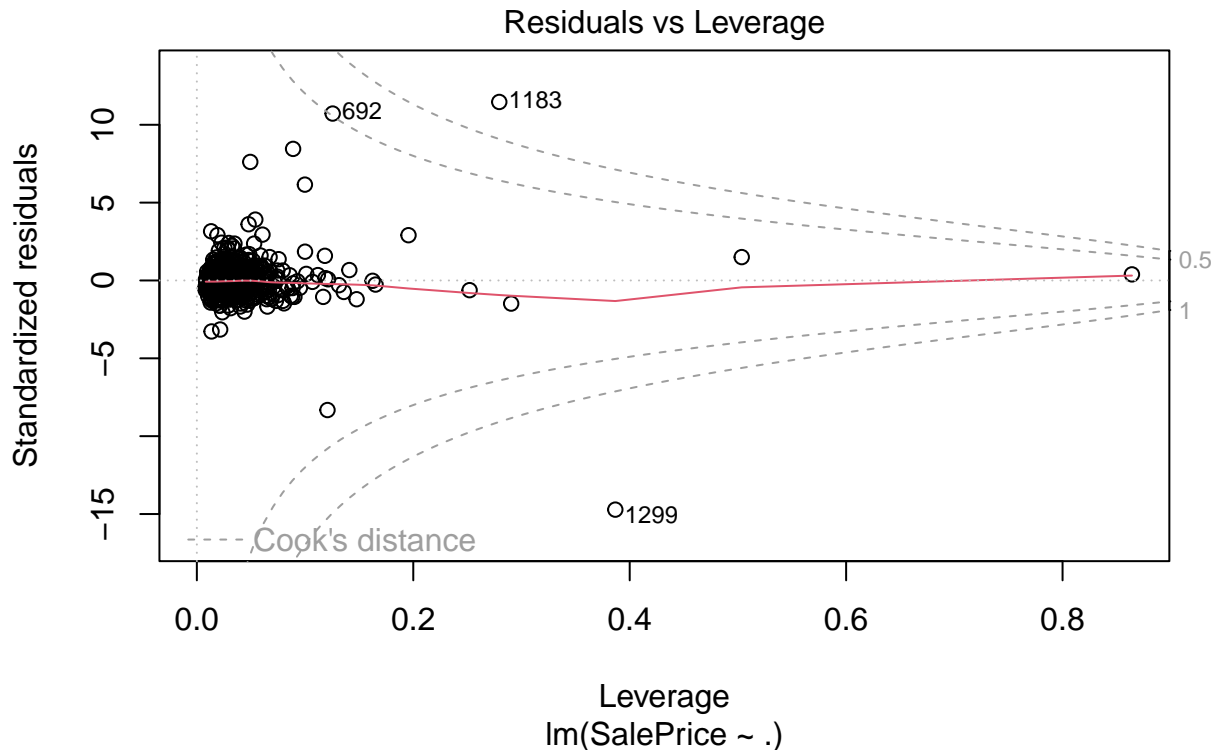
10. Si tiene multicolinealidad o sobreajuste, haga un modelo con las variables que sean mejores predictoras del precio de las casas. Determine la calidad del modelo realizando un análisis de los residuos. Muéstrelo gráficamente.

```
plot(fitLM)
```









Residuals vs. fitted La gráfica se utiliza para detectar patrones no lineales en los residuos. Sin embargo, al observarla visualmente, no se puede identificar ningún patrón que sugiera la presencia de relaciones no lineales. Los residuos se distribuyen uniformemente en toda la gráfica.

Normal Q-Q La gráfica se utiliza para verificar si los residuos tienen una distribución normal, donde se considera mejor aquellos con más puntos alineados cerca de la diagonal. En la gráfica específica que se ha generado, se observa que la mayoría de los puntos o cuantiles están cercanos o alineados con la diagonal, lo que indica que la distribución de los residuos es normal.

Scale-Location la gráfica de residuos versus ajustes permite identificar la presencia de homocedasticidad o heterocedasticidad. En este caso, como la mayoría de los residuos se encuentran alrededor de cero, y no hay una tendencia clara en la dispersión de los residuos en función del ajuste, se puede concluir que no hay una violación importante de la suposición de homocedasticidad. En este caso, se observa que la mayoría de los datos están distribuidos en un rango aceptable entre 100,000 y 400,000, aunque hay algunos datos en el extremo derecho que se comportan de manera diferente.

Residuals vs Leverage es una herramienta importante en el análisis de regresión lineal, ya que permite identificar observaciones que pueden tener un impacto significativo en la línea de regresión. Si un punto está ubicado más allá de esta línea, se considera que tiene un alto leverage y puede tener un impacto significativo en la línea de regresión. En el caso específico del gráfico realizado, se identificaron dos puntos (1183 y 1299) que se encuentran más allá de la línea de Cook y que podrían tener un impacto significativo en la línea de regresión. Por lo tanto, se debe analizar la influencia de estos puntos en el modelo y considerar si deben ser eliminados o no. Es importante recordar que no siempre es necesario eliminar los puntos de alto leverage, ya que pueden ser puntos importantes en el conjunto de datos.

11. Utilice cada modelo con el conjunto de prueba y determine la eficiencia del algoritmo para predecir el precio de las casas. ¿Qué tan bien lo hizo?

```
prediction <- fitLM %>% predict(test)
```

```
## Warning in predict.lm(., test): prediction from a rank-deficient fit may be
## misleading
```

```
data.frame(RMSE = RMSE(prediction, test$SalePrice), R2 = R2(prediction, test$SalePrice))
```

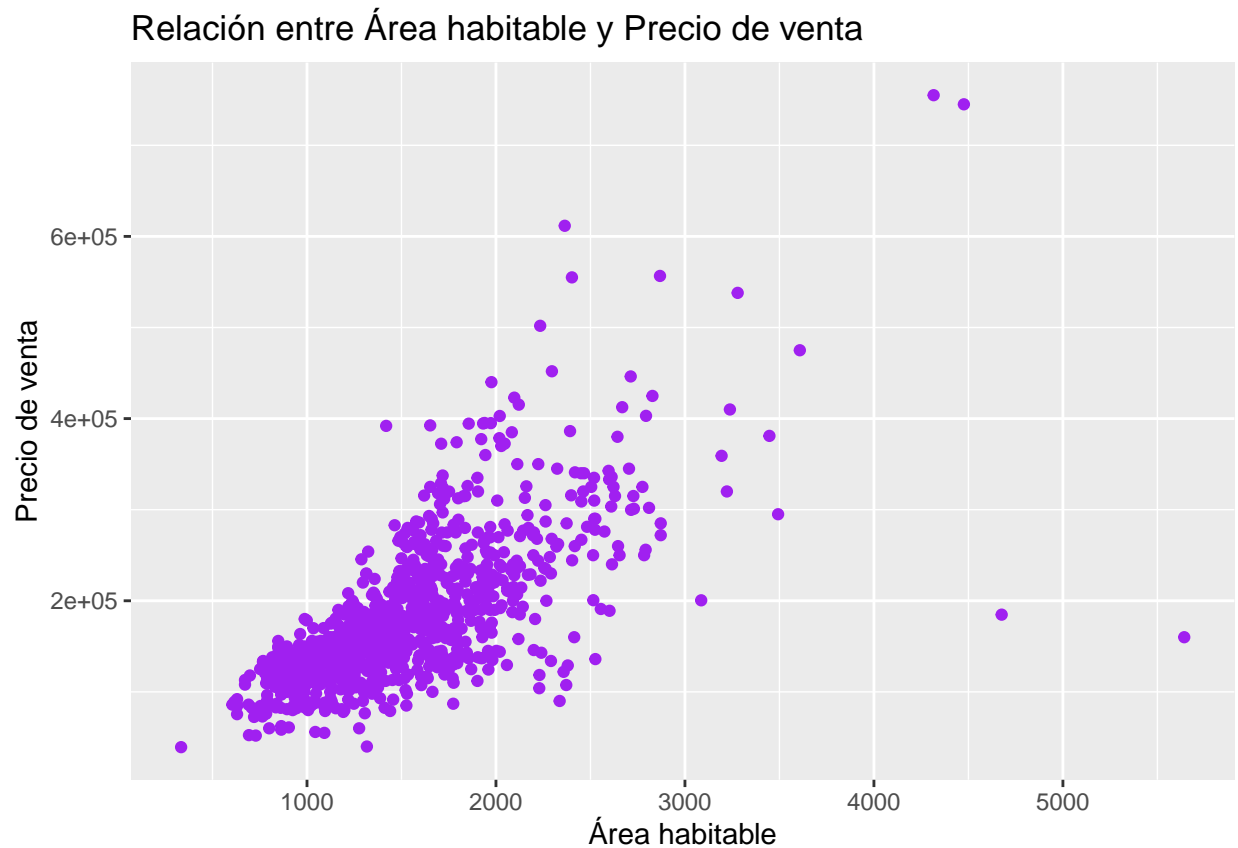
```
##      RMSE      R2
## 1 30521.33 0.8535235
```

```
anova(fitLM)
```

```
## Analysis of Variance Table
##
## Response: SalePrice
##      Df      Sum Sq    Mean Sq    F value    Pr(>F)
## MSSubClass      1 3.5675e+10 3.5675e+10    28.2812 1.297e-07 ***
## OverallQual      1 4.0112e+12 4.0112e+12   3179.7886 < 2.2e-16 ***
## OverallCond      1 6.3890e+08 6.3890e+08     0.5065 0.4768347
## YearBuilt        1 5.4289e+10 5.4289e+10    43.0366 8.649e-11 ***
## YearRemodAdd      1 2.2356e+10 2.2356e+10    17.7224 2.789e-05 ***
## BsmtFinSF1        1 2.3431e+11 2.3431e+11   185.7483 < 2.2e-16 ***
## BsmtFinSF2        1 1.4666e+10 1.4666e+10    11.6266 0.0006764 ***
## BsmtUnfSF         1 6.3079e+10 6.3079e+10    50.0051 2.897e-12 ***
## X1stFlrSF         1 1.3090e+11 1.3090e+11   103.7690 < 2.2e-16 ***
## X2ndFlrSF         1 3.7405e+11 3.7405e+11   296.5224 < 2.2e-16 ***
## LowQualFinSF      1 5.8249e+09 5.8249e+09     4.6176 0.0318875 *
## BsmtFullBath      1 1.2513e+10 1.2513e+10     9.9194 0.0016845 **
## BsmtHalfBath      1 3.0155e+09 3.0155e+09     2.3905 0.1223936
## FullBath          1 2.4627e+08 2.4627e+08     0.1952 0.6587004
## HalfBath          1 2.2357e+07 2.2357e+07     0.0177 0.8941197
## BedroomAbvGr      1 2.9899e+10 2.9899e+10    23.7023 1.308e-06 ***
## KitchenAbvGr      1 6.5956e+09 6.5956e+09     5.2286 0.0224293 *
## TotRmsAbvGrd      1 3.0906e+10 3.0906e+10    24.5007 8.729e-07 ***
## Fireplaces        1 1.0430e+10 1.0430e+10     8.2681 0.0041213 **
## GarageCars        1 3.7520e+10 3.7520e+10    29.7436 6.231e-08 ***
## GarageArea        1 6.7160e+07 6.7160e+07     0.0532 0.8175663
## WoodDeckSF        1 1.2077e+10 1.2077e+10     9.5739 0.0020291 **
## OpenPorchSF       1 4.8306e+08 4.8306e+08     0.3829 0.5361783
## EnclosedPorch     1 1.2265e+08 1.2265e+08     0.0972 0.7552503
## X3SsnPorch        1 7.6378e+08 7.6378e+08     0.6055 0.4366835
## ScreenPorch       1 7.1385e+09 7.1385e+09     5.6590 0.0175554 *
## PoolArea          1 1.9880e+09 1.9880e+09     1.5759 0.2096452
## MiscVal           1 2.6493e+07 2.6493e+07     0.0210 0.8848033
## MoSold            1 7.4047e+06 7.4047e+06     0.0059 0.9389448
## YrSold            1 1.1687e+08 1.1687e+08     0.0926 0.7609026
## y                 1 1.0839e+11 1.0839e+11   85.9264 < 2.2e-16 ***
## Residuals       989 1.2476e+12 1.2615e+09
```

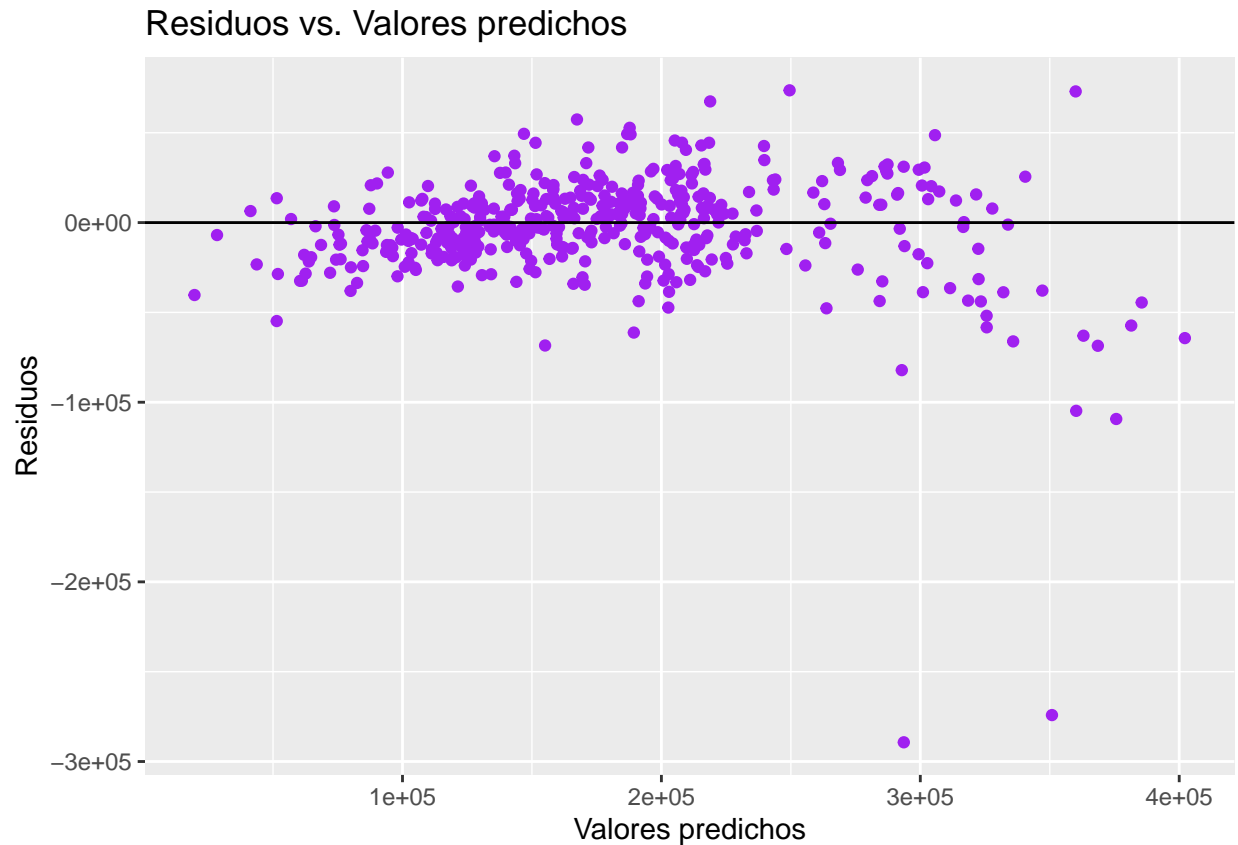
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(train, aes(x=GrLivArea, y=SalePrice)) +
  geom_point(color="purple") +
  labs(x="Área habitable", y="Precio de venta", title="Relación entre Área habitable y Precio de venta")
```



En la gráfica anterior, se puede ver que hay una clara relación positiva entre la variable Área habitable y el Precio de venta, lo cual es esperado.

```
ggplot(data.frame(predicted=prediction, actual=test$SalePrice, residuales=prediction-test$SalePrice), aes(x=predicted, y=residuales)) +
  geom_point(color="purple") +
  labs(x="Valores predichos", y="Residuos", title="Residuos vs. Valores predichos") +
  geom_hline(yintercept=0, color="black")
```

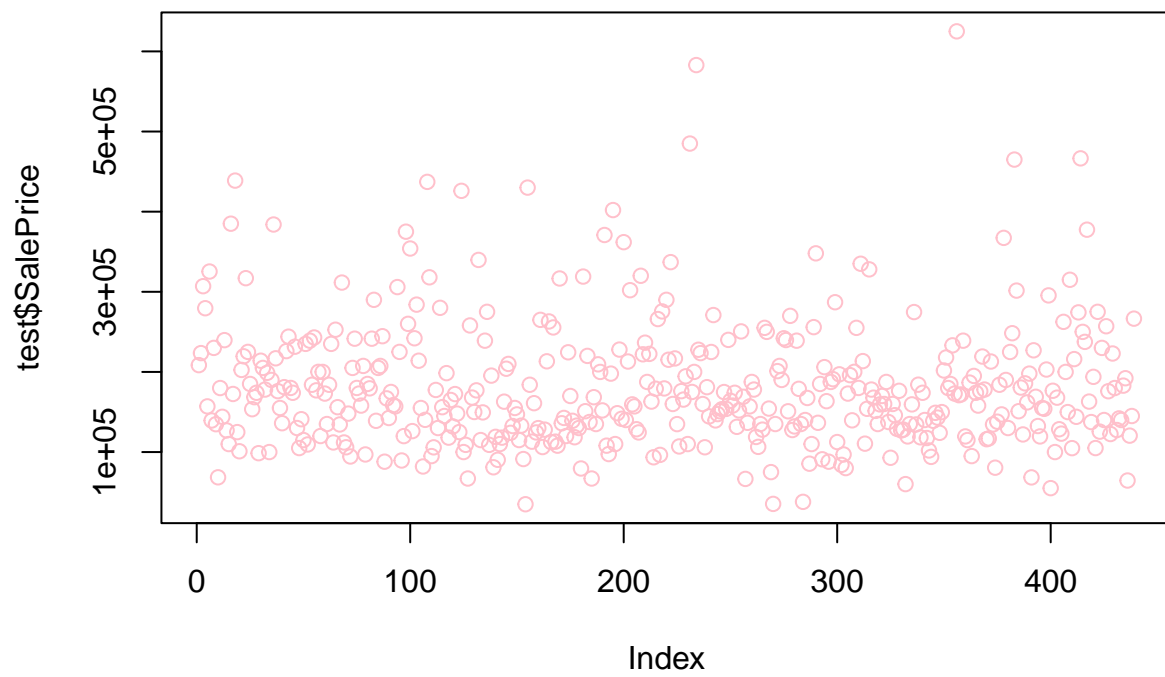


En la gráfica anterior, se puede ver que los residuos están centrados alrededor de 0, lo cual es indicativo de que el supuesto de se cumple. Sin embargo, se observa una cierta asimetría en los residuos, lo cual podría indicar que el supuesto de normalidad no se cumple completamente.

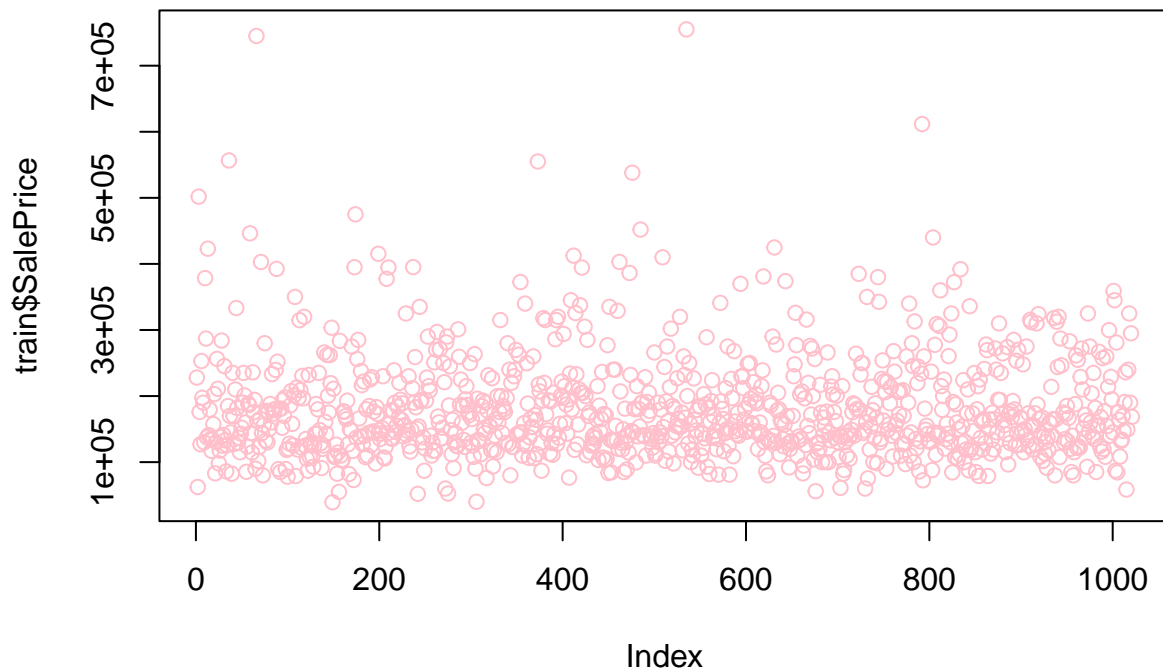
El modelo de regresión lineal parece ser adecuado para predecir el precio de las casas, aunque hay algunas variables independientes que podrían mejorarse para aumentar la precisión del modelo. La visualización de los datos y los residuos es útil para entender mejor la relación entre las variables y verificar si se cumplen los supuestos del modelo.

12. Discuta sobre la efectividad de los modelos. ¿Cuál lo hizo mejor? ¿Cuál es el mejor modelo para predecir el precio de las casas? Haga los gráficos que crea que le pueden ayudar en la discusión.

```
plot(test$SalePrice, col="pink")
```



```
plot(train$SalePrice, col="pink")
```



Vamos a volver a utilizar el resumen que se uso para el inciso 8, en este se puede encontrar el coeficiente de determinación (Multiple R-squared), el cual indica la proporción de varianza en la variable objetivo que es explicada por las variables independientes en el modelo. En este caso, el valor obtenido de 0.8 es considerado como un valor razonablemente bueno se puede concluir que el modelo tiene una efectividad aceptable, aunque se podría trabajar en la mejora de algunas variables independientes para aumentar la precisión del modelo.